

***Valeriano Comincioli***

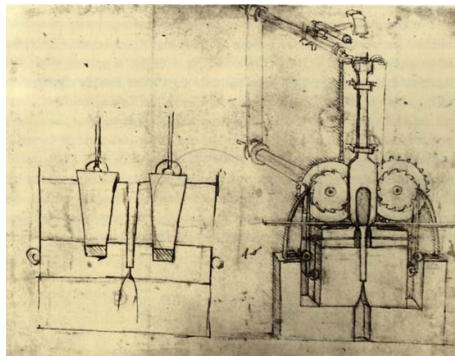
***Metodi Numerici e Statistici  
per le Scienze Applicate***



***Università degli Studi di Pavia***

**Valeriano Comincioli**

*Professore ordinario di Analisi Numerica  
Università degli Studi di Pavia*



seconda edizione © 2004

Metodi e Modelli Numerici e Probabilistici per  
Problemi Differenziali.

F.A.R. Fondo d'Ateneo per la Ricerca.  
Università degli Studi di Pavia

I diritti di traduzione, di riproduzione, di memorizzazione elettronica e di adattamento totale e parziale con qualsiasi mezzo (compresi i microfilm e le copie fotostatiche) sono riservati per tutti i paesi.

Realizzazione a cura dell'Autore mediante  $\text{\LaTeX}$

The purpose of computing  
is insight, not numbers.

(Hamming)

## Introduzione

È in atto da alcuni anni una profonda revisione dei programmi dei Corsi di Laurea ad orientamento scientifico: Biologia, Chimica, Geologia, Ingegneria, Medicina, per citarne qualcuno. Recente è l'attivazione delle Lauree triennali per la preparazione di nuove figure professionali e l'istituzione di Corsi di Laurea a carattere innovativo, quali Biotecnologie e Scienze dei Materiali. Elemento comune ai vari programmi di studio è la presenza, nuova o rinforzata, con denominazioni diverse, di insegnamenti di matematica ad alto contenuto applicativo. Si tratta di una risposta positiva ad un problema creatosi negli ultimi anni in seguito alla rapida diffusione, in tutti i campi del sapere, di mezzi di elaborazione elettronica, strumenti che richiedono la disponibilità e la conoscenza di adeguate tecniche matematiche per l'analisi e la soluzione di problemi sempre più complessi.

Dalla macroeconomia al metabolismo, dalle prospezioni geologiche alla meteorologia, il ricorso a modelli di tipo matematico e statistico è sempre più frequente e si rivela spesso il sistema più idoneo ed efficace per lo studio di situazioni reali. La modellizzazione matematica, sia di tipo deterministico, sia di tipo statistico, è il principale strumento di interpretazione, simulazione e predizione di fenomeni reali. Si tratta di un processo interdisciplinare che, in modo schematico, si articola nei seguenti passi: formulazione del problema, a partire dai dati sperimentali; costruzione di un modello; elaborazione e analisi matematica del modello; calcolo della soluzione; validazione del modello, ossia confronto dei risultati con i dati sperimentali. In tale processo hanno un ruolo fondamentale il *calcolo numerico* e la *statistica*, per tradizione trattati in testi diversi. *Uno degli scopi del presente volume è proprio quello di sviluppare, attorno all'idea di fondo del modello matematico, una trattazione organica e integrata degli strumenti numerici e statistici*, consentendo comunque, a chi fosse interessato in particolare ad uno dei due aspetti, di trovare le nozioni, i metodi e gli algoritmi desiderati.

Nella prima parte del volume (Cap. 1–7) sono introdotti e analizzati i principali metodi numerici riguardanti: l'analisi degli errori, la risoluzione di sistemi lineari, il calcolo di autovalori e di autovettori, l'approssimazione di funzioni, la risoluzione di equazioni non lineari, la programmazione matematica, il calcolo di integrali e la risoluzione di equazioni differenziali. Numerosi esempi, tratti da diversi campi

applicativi, accompagnano l'illustrazione degli algoritmi, agevolandone la comprensione e motivandone l'apprendimento. Alcuni degli esercizi proposti suggeriscono opportune estensioni del materiale presentato. I successivi quattro capitoli (Cap. 8–11) introducono prima le idee base del calcolo delle probabilità e della statistica, e poi tecniche ormai divenute classiche quali: catene di Markov, “cluster analysis”, metodo Monte Carlo, filtraggio dei segnali. Nei capitoli finali (Cap. 12–14) i diversi strumenti e metodi introdotti in precedenza trovano una loro naturale applicazione per la costruzione, soluzione, identificazione e controllo di modelli matematici. La modularità della trattazione e l'ampio spettro degli argomenti presentati permettono di adattare il volume alle esigenze di differenti corsi mediante una scelta conveniente dei capitoli e dei corrispondenti paragrafi. Alcuni risultati di base, spesso richiamati e utilizzati nel testo, sono ordinati e raccolti in Appendici. La bibliografia fornisce indicazioni utili per l'approfondimento di argomenti particolari. Gli algoritmi trattati nel volume sono in gran parte descritti in forma indipendente da un particolare linguaggio di programmazione. A scopo esemplificativo, di alcuni algoritmi è riportata la corrispondente implementazione in FORTRAN e nel linguaggio integrato MATLAB.

La scelta degli argomenti trattati riflette in parte le mie esperienze di ricerca pluridisciplinare. A questo riguardo, un particolare ringraziamento va ai colleghi G. Corona, E. D'Angelo, A. Faucitano, S. Nazari, L. Nespoli, P. Periti, C. Poggesi, C. Reggiani, A. Ugazio, V. Taglietti, F. Trimarchi, e A. Zonta, ai quali devo l'interesse e il gusto alla ricerca in vari campi della Biologia, della Chimica e della Medicina.

Pavia, gennaio 2004

Valeriano Comincioli

Ma le vere scienze son quelle che la sperienza ha fatto penetrare per li sensi,  
e posto silenzio alle lingue de' litiganti, e che non pasce di sogno li suoi investigatori,  
ma sempre sopra li primi veri e noti principii procede successivamente  
e con vere seguenzie insino al fine, come si denota nelle prime matematiche,  
cioè numero e misura, dette arismetica e geometria,  
che trattano con somma verità della quantità discontinua e continua.

**Leonardo da Vinci**

# Indice

<b>1</b>	<b>Analisi degli errori</b>	<b>1</b>
1.1	Sorgenti di errore . . . . .	3
1.2	Rappresentazione dei numeri sul calcolatore . . . . .	4
1.2.1	Rappresentazione dei numeri in differenti basi . . . . .	4
1.2.2	Conversione della rappresentazione di un numero reale . . . . .	6
1.2.3	Numeri macchina; sistema floating-point . . . . .	9
1.2.4	Operazione di arrotondamento . . . . .	13
1.2.5	Aritmetica in virgola mobile . . . . .	15
1.2.6	Propagazione degli errori . . . . .	18
1.2.7	Condizionamento di un problema . . . . .	21
<b>2</b>	<b>Algebra lineare numerica</b>	<b>25</b>
2.1	Metodi diretti . . . . .	31
2.1.1	Sistemi triangolari . . . . .	33
2.1.2	Sistemi generali; metodo di Gauss . . . . .	34
2.1.3	Strategia del pivot . . . . .	40
2.1.4	Pivoting scalato . . . . .	44
2.1.5	Decomposizione <b>LU</b> . . . . .	48
2.1.6	Decomposizione <b>LDM<sup>T</sup></b> . . . . .	49
2.1.7	Metodi di Crout e di Doolittle . . . . .	50
2.1.8	Matrici simmetriche . . . . .	51
2.1.9	Matrici a banda . . . . .	53
2.1.10	Matrici sparse . . . . .	59
2.1.11	Introduzione all'updating . . . . .	63
2.1.12	Fattorizzazione a blocchi . . . . .	64
2.2	Analisi degli errori; condizionamento e stabilità . . . . .	67
2.2.1	Stabilità degli algoritmi . . . . .	78
2.2.2	Fattorizzazione <b>A = QR</b> . . . . .	80
2.3	Metodi iterativi . . . . .	87
2.3.1	Metodi di Jacobi, Gauss-Seidel, rilassamento . . . . .	87
2.3.2	Metodi iterativi a blocchi . . . . .	93

2.3.3	Studio della convergenza . . . . .	95
2.3.4	Metodo del gradiente coniugato . . . . .	100
2.3.5	Precondizionamento . . . . .	105
<b>3</b>	<b>Autovalori e autovettori</b>	<b>110</b>
3.1	Condizionamento del problema degli autovalori . . . . .	111
3.2	Metodo delle potenze . . . . .	112
3.2.1	Iterazione inversa . . . . .	116
3.2.2	Deflazione . . . . .	117
3.2.3	Metodo di Lanczos . . . . .	117
3.3	Metodi di trasformazione per similitudine . . . . .	120
3.3.1	Metodo di Jacobi . . . . .	122
3.3.2	Metodo di Householder . . . . .	128
3.3.3	Metodo di Givens . . . . .	130
3.3.4	Matrici non simmetriche . . . . .	131
3.3.5	Matrici tridiagonali simmetriche . . . . .	131
3.3.6	Metodo <b>QR</b> . . . . .	134
3.4	Problema degli autovalori generalizzato . . . . .	136
3.5	Decomposizione SVD . . . . .	138
<b>4</b>	<b>Approssimazione di funzioni</b>	<b>141</b>
4.1	Interpolazione . . . . .	142
4.1.1	Interpolazione mediante polinomi . . . . .	143
4.1.2	Errore di troncamento nella interpolazione . . . . .	146
4.1.3	Convergenza del polinomio di interpolazione . . . . .	148
4.1.4	Costruzione del polinomio di interpolazione . . . . .	150
4.1.5	Interpolazione mediante spline . . . . .	155
4.1.6	Approssimazione di Bézier . . . . .	163
4.2	Problema generale di approssimazione . . . . .	170
4.2.1	Norma euclidea. Minimi quadrati . . . . .	173
4.2.2	Polinomi ortogonali . . . . .	178
4.2.3	Norma del massimo. Approssimazione di Chebichev . . . . .	183
4.3	Calcolo numerico delle derivate . . . . .	186
4.3.1	Studio dell'errore di troncamento . . . . .	187
4.3.2	Influenza degli errori di arrotondamento . . . . .	188
<b>5</b>	<b>Equazioni non lineari e ottimizzazione</b>	<b>191</b>
5.1	Caso unidimensionale . . . . .	196
5.1.1	Metodo di bisezione . . . . .	197
5.1.2	Metodo regula falsi . . . . .	200
5.1.3	Metodo di Newton . . . . .	202
5.1.4	Metodo di Newton in più dimensioni . . . . .	205

5.1.5	Studio della convergenza del metodo di Newton . . . . .	207
5.1.6	Metodo di Newton e radici multiple . . . . .	211
5.1.7	Alcune applicazioni del metodo di Newton . . . . .	212
5.1.8	Modifiche del metodo di Newton . . . . .	213
5.1.9	Radici di polinomi . . . . .	216
5.1.10	Sensitività delle radici di un polinomio . . . . .	220
5.2	Metodi di punto fisso . . . . .	226
5.2.1	Aspetti computazionali . . . . .	234
5.2.2	Accelerazione della convergenza . . . . .	235
5.3	Sistemi dinamici discreti . . . . .	239
5.4	Programmazione lineare . . . . .	242
5.4.1	Trasformazione di problemi LP nella prima forma primale . . . . .	248
5.4.2	Problema duale . . . . .	249
5.4.3	Seconda forma primale . . . . .	250
5.4.4	Alcuni esempi applicativi . . . . .	251
5.4.5	Metodo del simplesso . . . . .	253
5.4.6	Risoluzione di sistemi lineari inconsistenti . . . . .	256
5.5	Metodi di ottimizzazione . . . . .	258
5.5.1	Ottimizzazione unidimensionale . . . . .	259
5.5.2	Ottimizzazione in più dimensioni . . . . .	268
5.5.3	Metodo SOR . . . . .	281
5.5.4	Minimi quadrati non lineari . . . . .	284
<b>6</b>	<b>Integrazione numerica</b>	<b>288</b>
6.1	Formule di Newton–Cotes . . . . .	296
6.1.1	Convergenza delle formule di quadratura . . . . .	299
6.1.2	Formule composte . . . . .	300
6.2	Formule di Gauss . . . . .	302
6.2.1	Formule di Lobatto . . . . .	305
6.2.2	Formule di quadratura di Gauss-Kronrod . . . . .	305
6.3	Formule adattive . . . . .	307
6.3.1	Formula di Simpson adattiva . . . . .	308
6.4	Formule di estrapolazione . . . . .	310
6.5	Difficoltà nell'integrazione numerica . . . . .	314
6.6	Integrali multipli . . . . .	316
<b>7</b>	<b>Equazioni differenziali</b>	<b>321</b>
7.1	Aspetti introduttivi . . . . .	321
7.1.1	Definizione di soluzione . . . . .	323
7.1.2	Curve soluzioni e campi di direzioni . . . . .	324
7.1.3	Problemi ai valori iniziali . . . . .	325
7.2	Alcuni modelli . . . . .	330

7.3	Metodi numerici . . . . .	343
7.3.1	Metodo di Eulero . . . . .	343
7.3.2	Influenza degli errori di arrotondamento . . . . .	351
7.3.3	Metodi di sviluppo in serie . . . . .	352
7.3.4	Metodi di Runge-Kutta . . . . .	354
7.3.5	Metodo di Eulero implicito e formula dei trapezi . . . . .	358
7.3.6	Metodi di Runge-Kutta-Fehlberg . . . . .	363
7.3.7	Metodi a più passi . . . . .	368
7.3.8	Convergenza dei metodi lineari a più passi . . . . .	372
7.3.9	Stabilità per passo fissato . . . . .	376
7.3.10	Sistemi di equazioni del primo ordine . . . . .	378
7.3.11	Metodo di Cowell-Numerov . . . . .	381
7.4	Equazioni stiff . . . . .	382
7.4.1	Metodi numerici . . . . .	388
7.4.2	Sistemi altamente oscillatori . . . . .	390
7.5	Problemi ai limiti . . . . .	391
7.5.1	Alcuni modelli . . . . .	392
7.5.2	Metodo shooting . . . . .	395
7.5.3	Metodo alle differenze . . . . .	399
7.5.4	Metodo degli elementi finiti . . . . .	404
7.5.5	Problema degli autovalori . . . . .	410
7.6	Equazioni integrali . . . . .	412
7.7	Equazioni con ritardo . . . . .	420
7.7.1	Introduzione ai metodi numerici . . . . .	423
7.8	Equazioni alle derivate parziali . . . . .	426
7.8.1	Propagazione delle onde . . . . .	428
7.8.2	Approssimazione numerica . . . . .	431
7.8.3	Equazioni non lineari . . . . .	435
7.8.4	Equazione delle onde . . . . .	438
7.8.5	Equazione della diffusione . . . . .	446
7.8.6	Equazione di Laplace . . . . .	453
<b>8</b>	<b>Probabilità e statistica</b>	<b>462</b>
8.1	Elementi di calcolo della probabilità . . . . .	463
8.1.1	Probabilità matematica e probabilità statistica . . . . .	463
8.1.2	Elementi di calcolo combinatorio . . . . .	464
8.1.3	Teoria assiomatica della probabilità . . . . .	466
8.1.4	Probabilità condizionata e indipendenza statistica . . . . .	472
8.2	Variabili aleatorie e funzioni di distribuzione . . . . .	480
8.2.1	Parametri di una distribuzione . . . . .	482
8.2.2	Variabili aleatorie multivariate . . . . .	488
8.2.3	Distribuzioni $n$ -dimensionali ( $n > 2$ ) . . . . .	498



8.2.4	Analisi di alcune distribuzioni . . . . .	499
8.3	Densità di Gauss o normale . . . . .	511
8.3.1	Distribuzione normale multivariata . . . . .	513
8.4	Campionamenti . . . . .	516
8.4.1	Introduzione al problema . . . . .	517
8.4.2	Campionamento di distribuzioni normali . . . . .	518
8.4.3	Teorema limite centrale e applicazioni . . . . .	526
8.5	Elementi di statistica inferenziale . . . . .	530
8.5.1	Modello deterministico e modello stocastico . . . . .	531
8.5.2	Stimatori e loro proprietà . . . . .	532
8.5.3	Verifica di ipotesi . . . . .	538
8.6	Software numerico . . . . .	550
8.7	Catene di Markov . . . . .	551
8.7.1	Concetti di base . . . . .	554
8.7.2	Descrizione di un sistema . . . . .	558
8.7.3	Catene di Markov infinite . . . . .	567
8.8	Introduzione alla teoria dell'Informazione . . . . .	570
8.8.1	Quantità di Informazione . . . . .	571
8.8.2	Disuguaglianza di Gibbs . . . . .	575
8.8.3	Codifica e disuguaglianza di Kraft . . . . .	575
8.8.4	Codici ottimali; algoritmo di Huffman . . . . .	578
8.8.5	Codifica di blocchi . . . . .	578
8.8.6	Applicazione alla costruzione di questionari . . . . .	579
<b>9</b>	<b>Algoritmi nella cluster analysis</b>	<b>581</b>
9.1	Rappresentazione dei dati . . . . .	581
9.1.1	Matrice campione . . . . .	582
9.1.2	Matrice di prossimità . . . . .	582
9.1.3	Il problema del clustering . . . . .	583
9.1.4	Tipi di dati e scale . . . . .	584
9.1.5	Indici di prossimità . . . . .	584
9.1.6	Variabili nominali . . . . .	586
9.1.7	Proiezioni lineari . . . . .	587
9.2	Metodi e algoritmi di clustering . . . . .	588
9.2.1	Algoritmi di tipo gerarchico . . . . .	588
9.2.2	Elementi di teoria dei grafi . . . . .	590
9.2.3	Algoritmi Single-Link e Complete-Link . . . . .	594
9.2.4	Algoritmi di clustering di tipo partizione . . . . .	596
9.2.5	Fuzzy clustering . . . . .	606
9.2.6	Clustering software . . . . .	609
9.3	Validazione del clustering . . . . .	610

<b>10 Metodo Monte Carlo</b>	<b>612</b>
10.1 Numeri casuali e pseudo-casuali . . . . .	614
10.1.1 Distribuzioni uniformi . . . . .	614
10.1.2 Numeri casuali secondo una distribuzione assegnata . . . . .	617
10.2 Calcolo di integrali . . . . .	622
10.2.1 Metodo Monte Carlo hit or miss . . . . .	623
10.2.2 Metodo Monte Carlo sample-mean . . . . .	625
10.2.3 Calcolo di integrali multipli . . . . .	627
10.2.4 Tecniche di riduzione della varianza . . . . .	628
10.3 Simulazione . . . . .	631
10.3.1 Problema dei due dadi . . . . .	631
10.3.2 Problema di Buffon . . . . .	632
10.3.3 Simulazione di traiettorie con collisioni . . . . .	633
<b>11 Trattamento dei segnali</b>	<b>637</b>
11.1 Teoria dei sistemi lineari . . . . .	638
11.1.1 Sistemi lineari invarianti nel tempo . . . . .	640
11.1.2 Equazioni alle differenze . . . . .	643
11.1.3 Rappresentazione in dominio di frequenza . . . . .	644
11.1.4 Relazione tra sistemi continui e discreti . . . . .	647
11.1.5 La trasformata $z$ . . . . .	649
11.1.6 La trasformata di Fourier discreta . . . . .	651
11.2 Analisi dei segnali mediante MATLAB . . . . .	656
11.2.1 Segnali come vettori . . . . .	656
11.2.2 Sistemi lineari discreti . . . . .	657
11.2.3 Analisi dei filtri . . . . .	658
11.2.4 FFT e analisi spettrale . . . . .	660
11.3 Introduzione al filtro di Kalman . . . . .	662
11.3.1 Metodo dei minimi quadrati con peso . . . . .	663
11.3.2 Metodo dei minimi quadrati ricorsivo . . . . .	664
11.3.3 Filtro di Kalman . . . . .	666
11.4 Introduzione alle funzioni wavelet . . . . .	669
<b>12 Teoria dei compartimenti</b>	<b>677</b>
12.1 Elementi introduttivi . . . . .	678
12.2 Modello a compartimenti generale . . . . .	688
12.3 Cinetica dei traccianti . . . . .	690
12.3.1 Modelli a compartimenti lineari . . . . .	692
12.3.2 Equazioni della concentrazione del tracciante . . . . .	695
12.3.3 Struttura di un sistema e connettività . . . . .	696
12.3.4 Stabilità . . . . .	698
12.4 Identificazione del modello . . . . .	703

<b>13</b>	<b>Identificazione dei parametri</b>	<b>707</b>
13.1	Formulazione del problema . . . . .	707
13.1.1	Difficoltà nella identificazione . . . . .	709
13.2	Intervalli di confidenza per i parametri . . . . .	710
13.2.1	Equazioni di sensitività . . . . .	711
13.2.2	Problema linearizzato . . . . .	712
13.2.3	Scelta della matrice $W$ . . . . .	716
13.2.4	Formule riassuntive . . . . .	716
13.2.5	Pianificazione degli esperimenti . . . . .	717
<b>14</b>	<b>Teoria del controllo ottimo</b>	<b>721</b>
14.1	Modelli introduttivi . . . . .	722
14.2	Formulazione di un problema di controllo . . . . .	733
14.2.1	Forme diverse di un controllo ottimo . . . . .	734
14.3	Metodo della programmazione dinamica . . . . .	737
14.3.1	Principio di ottimalità . . . . .	737
14.3.2	Programmazione dinamica nel caso continuo . . . . .	745
14.4	Principio del minimo di Pontryagin . . . . .	751
14.4.1	Problemi di controllo discreti . . . . .	758
14.4.2	Problemi di controllo continui . . . . .	761
14.4.3	Metodi numerici . . . . .	767
14.4.4	Programmazione dinamica e principio del minimo . . . . .	769
14.4.5	Legame con il calcolo delle variazioni . . . . .	770
14.4.6	Applicazioni; modelli matematici . . . . .	772
14.5	Identificazione di parametri . . . . .	790
14.5.1	Equazioni di sensitività . . . . .	791
14.5.2	Metodo basato sulla teoria dei controlli . . . . .	792
<b>A</b>	<b>Elementi di algebra lineare</b>	<b>798</b>
A.1	Matrici. Definizioni fondamentali . . . . .	798
A.1.1	Matrici particolari . . . . .	800
A.1.2	Operazioni su matrici . . . . .	800
A.1.3	Matrici partizionate . . . . .	806
A.1.4	Indipendenza lineare, base e dimensione . . . . .	806
A.1.5	Determinante, inversa e rango . . . . .	810
A.1.6	Matrici elementari . . . . .	815
A.2	Sistemi lineari . . . . .	821
A.3	Autovalori e trasformazioni per similitudine . . . . .	822
A.3.1	Trasformazioni per similitudine . . . . .	824
A.3.2	Autovettori a sinistra . . . . .	825
A.3.3	Riduzione delle matrici . . . . .	826
A.3.4	Fattorizzazione unitaria di una matrice . . . . .	827

A.4	Localizzazione degli autovalori . . . . .	830
A.4.1	Norma di vettore e di matrice . . . . .	831
A.5	I valori singolari e la pseudoinversa . . . . .	835
A.5.1	Decomposizione in valori singolari . . . . .	836
A.5.2	Risultati di perturbazione per i valori singolari . . . . .	840
A.5.3	Applicazioni della SVD . . . . .	841
A.5.4	Pseudoinversa . . . . .	844
A.6	Matrici non negative . . . . .	848
A.6.1	Matrici irriducibili . . . . .	849
A.6.2	Matrici con inverse non negative; M-matrici . . . . .	851
<b>B</b>	<b>Equazioni differenziali. Tecniche analitiche</b>	<b>854</b>
B.1	Separazione delle variabili . . . . .	854
B.2	Equazione lineare del primo ordine . . . . .	855
B.3	Equazione di Bernoulli . . . . .	858
B.4	Equazione di Riccati . . . . .	859
B.5	Equazione omogenea . . . . .	860
B.6	Equazione esatta . . . . .	863
B.7	Equazione di Clairaut . . . . .	864
B.8	Equazioni lineari del secondo ordine . . . . .	866
B.8.1	Equazioni lineari non omogenee particolari . . . . .	869
B.8.2	Sistemi differenziali lineari del primo ordine . . . . .	871
B.8.3	Equazioni lineari di ordine n . . . . .	874
B.9	Trasformata di Laplace . . . . .	876
B.9.1	Proprietà della trasformata di Laplace . . . . .	881
B.9.2	Applicazioni della trasformata di Laplace . . . . .	885
B.10	Serie di Fourier . . . . .	890
B.10.1	Equazione della diffusione . . . . .	895
B.10.2	Equazione delle onde . . . . .	897
<b>C</b>	<b>Algoritmi di ricerca e di ordinamento</b>	<b>900</b>
C.1	Algoritmi di ricerca . . . . .	900
C.1.1	Ricerca sequenziale . . . . .	900
C.1.2	Ricerca in strutture dati ordinate . . . . .	902
C.2	Algoritmi di ordinamento . . . . .	904
<b>D</b>	<b>Tabelle statistiche</b>	<b>913</b>
	<b>Bibliografia</b>	<b>931</b>

Errors, like straws, upon the surface flow;  
He who would search for pearls must dive below.  
John Dryden, *All for love*, 1677

## Capitolo 1

# Analisi degli errori

In questo capitolo esamineremo in particolare gli *errori* relativi all'uso di uno *strumento di calcolo*: la loro *origine*, la *propagazione*, e alcune *tecniche* per la loro *valutazione*. Ciò è in analogia a quanto avviene quando si usa in laboratorio uno *strumento di misurazione*: le misure ottenute hanno senso soltanto se è *nota* la *sensibilità*, cioè la capacità di discriminazione, dello strumento.

Un *calcolatore numerico* è in grado di rappresentare soltanto un numero *finito* di cifre; ne consegue la possibilità che un numero reale introdotto nel calcolatore venga *approssimato*; le operazioni elementari eseguite su tali numeri possono, a loro volta, produrre risultati non rappresentabili esattamente nel calcolatore. Pertanto, quando un *algoritmo*<sup>1</sup> viene eseguito su un calcolatore, si ha in generale una successiva *creazione e propagazione* di errori. Tali errori sono comunemente chiamati *errori di arrotondamento*, dal nome di una tecnica usuale per rappresentare i numeri reali sul calcolatore.

Il risultato prodotto dall'algoritmo differisce, quindi, in generale dal *risultato esatto*, cioè da quel risultato *ideale* che si potrebbe ottenere operando con tutte le cifre richieste. Senza un'idea, più precisamente una *maggiorazione* della differenza dei due risultati, il risultato *numerico* può essere del tutto *illusorio*. Infatti, esso può dipendere dal numero di cifre usate e/o dall'*ordine* in cui vengono effettuate le operazioni (cioè dal particolare algoritmo utilizzato).

Nel seguito, per esaminare l'*origine* degli errori si analizzeranno brevemente alcune tecniche usuali di *rappresentazione dei numeri reali su un calcolatore numerico*,

---

<sup>1</sup>Per il seguito, chiameremo *algoritmo* una successione finita di operazioni elementari, quale ad esempio la successione di istruzioni in un particolare linguaggio di programmazione, che indica in maniera precisa, senza ambiguità, come calcolare la soluzione (*output*) di un problema a partire da valori iniziali assegnati (*input*). La parola algoritmo è derivata dal nome del matematico persiano Abu Jafar Mohammed ibn Musa al-Khowarizmi, che operò in Baghdad intorno agli anni 840.

in modo da definire l'insieme dei cosiddetti *numeri macchina*. Successivamente si definiranno per tali numeri le operazioni elementari, cioè le *operazioni macchina* e si studierà il loro comportamento rispetto alla propagazione degli errori. Si analizzerà quindi, più in generale il comportamento, rispetto agli errori, di una successione di operazioni macchina, cioè di un *algoritmo*. Importanti, in questo contesto, saranno i concetti di *stabilità di un algoritmo* e di *condizionamento di un problema*. A queste considerazioni premetteremo alcune notazioni e semplici definizioni ed una *classificazione* più generale degli errori che si possono commettere quando si modella un problema reale.

### Alcune notazioni

Se  $a$  e  $b$  sono due numeri reali assegnati

- $a \gg b$  significa  $a$  “molto più grande di”  $b$ ; il significato preciso di “molto più grande” dipende dal contesto.
- $a \approx b$ , significa  $a$  “approssimativamente uguale a”  $b$ ; anche qui il significato preciso dipende dal contesto e dalla precisione usata.

**Simboli di Landau** Consideriamo due funzioni  $f, g$  definite su un intervallo  $I$  della retta reale  $\mathbb{R}$ , con  $g(x) \neq 0$  per  $x \in I$ .

1. Si dice che  $f$  è di ordine “*O grande*” rispetto a  $g$  per  $x$  che tende a  $x_0 \in I$ , quando esiste una costante  $C > 0$  e un  $\delta > 0$ , tali che

$$\left| \frac{f(x)}{g(x)} \right| \leq C$$

per ogni  $x \in I$  con  $x \neq x_0$  e  $|x - x_0| < \delta$ . Si scrive  $f(x) = O(g(x))$  per  $x \rightarrow x_0$ .

2. Si dice che  $f$  è di ordine “*o piccolo*” rispetto a  $g$  per  $x$  che tende a  $x_0$ , quando per ogni costante  $C > 0$  esiste un  $\delta > 0$  tale che

$$\left| \frac{f(x)}{g(x)} \right| \leq C$$

per ogni  $x \in I$  con  $x \neq x_0$  e  $|x - x_0| < \delta$ . Si scrive  $f(x) = o(g(x))$  per  $x \rightarrow x_0$ .

Ad esempio, se  $f(x)$  è una funzione continua sull'intervallo  $[0, 1]$  con  $f(0) = 0$ , allora  $f(x) = o(1)$  per  $x \rightarrow 0$ . Se  $f(x)$  è derivabile con derivata  $f'(x)$  continua su  $[0, 1]$ , e quindi tale che  $|f'(x)| \leq C$  per una costante  $C$  conveniente, si ha  $f(x) = O(x)$  per  $x \rightarrow 0$ .

Ricordiamo alcune proprietà dei simboli di Landau.

- (1)  $f(x) = KO(g(x)) \Rightarrow f(x) = O(g(x))$ , per ogni  $K \in \mathbb{R}$ ;

- (2)  $f(x) = O(g_1(x))$  e  $g_1(x) = O(g_2(x)) \Rightarrow f(x) = O(g_2(x))$ ;
- (3)  $f_1(x) = O(g_1(x))$ ,  $f_2(x) = O(g_2(x)) \Rightarrow f_1(x) f_2(x) = O(g_1(x) g_2(x))$ ;
- (4)  $f(x) = O(g_1(x) g_2(x)) \Rightarrow f(x) = g_1(x) O(g_2(x))$ .

Analoghe proprietà valgono per il simbolo “o”.

Se  $\tilde{a}$  è un valore *approssimato* del numero reale  $a$ , si definisce

**Errore assoluto** la quantità  $\tilde{a} - a$ .

**Errore relativo** la quantità  $(\tilde{a} - a)/a$ , se  $a \neq 0$ . Usualmente l'errore relativo è dato in percentuale; ad esempio, 5% significa che l'errore relativo è 0.05.

L'errore può essere, naturalmente, sia positivo che negativo. In generale, tuttavia, si è interessati ad una sua *limitazione positiva*. La notazione  $a = \tilde{a} \pm \epsilon$  indica allora che  $|\tilde{a} - a| \leq \epsilon$  e  $\epsilon$  rappresenta una limitazione dell'errore. Ad esempio,  $a = 0.7136 \pm 0.0013$  significa  $0.7123 \leq a \leq 0.7149$ . Ricordiamo, tuttavia, che nella valutazione *statistica* degli errori si utilizza la notazione  $a = \tilde{a} \pm \epsilon$  per indicare l'*errore standard* o altre analoghe misure della deviazione dal valore medio (cfr. Capitolo 8).

Nel caso in cui  $\mathbf{a}$  rappresenti un vettore o una matrice, come misura dell'errore assoluto, e rispettivamente relativo, si può assumere  $\|\tilde{\mathbf{a}} - \mathbf{a}\|$ , e rispettivamente  $\|\tilde{\mathbf{a}} - \mathbf{a}\|/\|\mathbf{a}\|$ , ove  $\|\cdot\|$  rappresenta una opportuna *norma* di vettore o di matrice.

Se la limitazione dell'errore in  $\tilde{a}$  è inferiore a  $\frac{1}{2} 10^{-t}$ , con  $t$  intero fissato, allora si dice che  $\tilde{a}$  ha  $t$  *cifre decimali corrette*. Ad esempio,  $0.001234 \pm 0.000004$  ha *cinque* cifre decimali corrette, mentre  $0.001234 \pm 0.000006$  ne ha *quattro*. Nel primo caso si dice anche che la stima ha tre *cifre significative* (osserviamo che gli zeri iniziali non sono contati), mentre nel secondo caso si hanno due cifre significative.

Il numero delle cifre corrette dà una idea della grandezza dell'errore assoluto, mentre il numero delle cifre significative fornisce una idea della grandezza dell'errore relativo.

## 1.1 Sorgenti di errore

I *risultati numerici* possono essere *influenzati* da diversi tipi di *errori*. Può essere di interesse una loro *classificazione* schematica, anche allo scopo di individuare quali di essi sono di maggior interesse per il seguito.

**1. Semplificazioni introdotte nel modello** Sono gli errori dovuti al fatto che, ad esempio, il modello è supposto di tipo lineare, oppure si suppongono “trascurabili” alcune grandezze fisiche. In questo caso si parla di *adeguatezza* del modello.

**2. Errori nei dati** I dati di un problema sono il risultato, in generale, di misurazioni, che possono essere influenzate da errori *sistematici* e/o da errori *random*. Gli errori sistematici sono gli errori che dipendono dalla sensibilità dello *strumento di*

*misurazione*. Gli errori random, invece, sono dovuti al verificarsi in concomitanza di eventi “imprevedibili”.

La conoscenza o, più realisticamente, una *stima* degli *errori nei dati* è importante per il calcolo numerico. Da essa, infatti dipende, come vedremo, la *scelta* della *precisione* e dei *test di arresto* da utilizzare nella *implementazione* degli algoritmi.

Mentre la stima degli errori sistematici non è di competenza specifica dell’analisi numerica, la stima degli *errori random* può essere ottenuta mediante opportune *tecniche statistiche*, in particolare mediante il metodo dei minimi quadrati. Tali tecniche saranno analizzate nel successivo Capitolo 8.

**3. Errori di arrotondamento nei dati e nei calcoli** Sono gli errori introdotti nella rappresentazione dei numeri sul calcolatore. Essi sono l’oggetto principale di studio del presente capitolo.

**4. Errori di troncamento** Sono gli errori che si introducono quando un procedimento *infinito* è approssimato mediante un procedimento *finito*. Per esempio, quando una derivata viene approssimata mediante un rapporto incrementale o un integrale mediante una formula di quadratura. Lo studio di questo tipo di errori costituirà l’oggetto principale dei capitoli successivi. Nell’elenco precedente non sono stati incluse altre ovvie possibili cause di errore: ad esempio, gli errori *umani* di disattenzione in calcoli manuali, di introduzione dati, nonché gli errori che uno strumento di calcolo, software o hardware, *può* commettere.

## 1.2 Rappresentazione dei numeri sul calcolatore

In questo paragrafo analizzeremo alcune tecniche di rappresentazione dei numeri reali su un calcolatore. Si tratta di un’*analisi schematica*, ma sufficiente per evidenziare gli aspetti essenziali di interesse per il calcolo numerico.

### 1.2.1 Rappresentazione dei numeri in differenti basi

In *notazione posizionale* a base 10 un simbolo come 67041 rappresenta il seguente numero intero

$$67041 = 6 \cdot 10^4 + 7 \cdot 10^3 + 0 \cdot 10^2 + 4 \cdot 10^1 + 1 \cdot 10^0$$

I simboli 0123456789 sono le *cifre arabiche o decimali*. Più in generale, nel *sistema decimale* un *numero intero positivo*  $N$  viene espresso nella seguente forma

$$N = d_n 10^n + d_{n-1} 10^{n-1} + \dots + d_0 10^0$$

e rappresentato con il simbolo

$$N = d_n d_{n-1} \dots d_1 d_0$$



ove le cifre  $d_n, d_{n-1}, \dots, d_1, d_0$  sono numeri interi compresi tra zero e nove.

Non vi è una ragione intrinseca per usare la base 10. In realtà, in altre civiltà si è fatto uso di basi diverse, ad esempio le basi 12, 20, 60. Nei *calcolatori numerici*, nei quali le cifre vengono rappresentate mediante *stati fisici*, è conveniente rappresentare i numeri in un sistema a base 2, detto *sistema binario*<sup>2</sup>, nel quale le cifre si riducono ai due simboli 0, 1, detti anche *bit* (= binary digit).

In un sistema avente come *base* un generico intero  $\beta > 1$  si ha la rappresentazione

$$N = (d_n d_{n-1} \dots d_1 d_0)_\beta = d_n \beta^n + d_{n-1} \beta^{n-1} + \dots + d_1 \beta^1 + d_0 \beta^0 \quad (1.1)$$

ove  $d_i$ , dette *cifre* (o digit), sono interi compresi tra 0 e  $\beta - 1$ . Se  $\beta$  è più piccolo di 10, per indicare le  $d_i$  si usano ancora le cifre decimali. In caso contrario, si devono introdurre ulteriori simboli; ad esempio, in base 16 le cifre sono 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F e sono dette *cifre esadecimali*. È interessante osservare che se  $\beta$  è una potenza  $n$ -ma di 2, per rappresentare le cifre in tale base bastano  $n$  bits; ad esempio, per  $\beta = 2^4$  le *quaterne*  $(0000)_2, (0001)_2, (0010)_2, \dots, (1111)_2$  corrispondono alle *cifre esadecimali*. Questo permette un immediato passaggio dalla base 2 alla base  $2^n$ .

La scelta di una particolare base è legata, in particolare, allo strumento di calcolo, tenendo conto del fatto che *più piccola* è la base e *più lunga* risulta la stringa di cifre per rappresentare lo stesso numero. Al contrario, il *numero di cifre* necessarie è più piccolo quanto più è grande la base. Per un esempio illustrativo si veda la Tabella 1.1.

base		base	
2	100011111	12	1BB
4	10133	16	11F
8	437	32	8V

Tabella 1.1: Rappresentazione in differenti basi del numero  $(287)_{10}$ .

Fissato ora un numero  $\beta$  intero maggiore di 1, rappresentiamo nella base  $\beta$  un *numero reale* qualunque  $x$ .

Ricordiamo il seguente importante risultato.

---

<sup>2</sup>La notazione in base 2 pare essere stata usata per la prima volta nel 1605 da Harriot; tuttavia, la data di origine della numerazione binaria è comunemente fatta risalire all'anno 1703 nel quale Leibnitz pubblicò un lavoro completo sull'aritmetica in base 2. L'utilizzo di una base 12 fu suggerito da Pascal (1658). La *notazione posizionale* oggi utilizzata nella rappresentazione dei numeri è fatta risalire ai babilonesi, che usavano la base 60, ancora oggi utilizzata nelle misure del tempo e degli angoli. L'introduzione in Europa della notazione posizionale in base 10 è accreditata agli arabi (intorno agli anni 1000), che perfezionarono una notazione già diffusa in India. La sua diffusione è merito particolare di Leonardo Pisano (detto Fibonacci) con il suo *Liber Abaci* (1202).

**Teorema 1.1** Dato un intero  $\beta > 1$ , un numero reale  $x$  diverso dallo zero può essere espresso in forma univoca nella seguente forma

$$x = \text{sign}(x)(d_1\beta^{-1} + d_2\beta^{-2} + d_3\beta^{-3} + \dots)\beta^p = \text{sign}(x)m\beta^p \quad (1.2)$$

ove  $\text{sign}(x) = 1$  se  $x > 0$ ,  $\text{sign}(x) = -1$  se  $x < 0$ ,  $p$  è un numero intero e le cifre  $d_1, d_2, d_3, \dots$  verificano le condizioni

(i)  $0 \leq d_i \leq \beta - 1$

(ii)  $d_1 \neq 0$  e le  $d_i$  non sono tutte uguali a  $\beta - 1$  a partire da un indice in poi.

Il numero  $m = d_1\beta^{-1} + d_2\beta^{-2} + d_3\beta^{-3} + \dots$  viene detto comunemente la *mantissa* di  $x$  e  $\beta^p$  la *parte esponente*; il numero  $p$  è detto anche *caratteristica* di  $x$ . Grazie al fatto che  $d_1 \neq 0$ , si ha che il numero reale  $m$  soddisfa alla seguente condizione

$$\frac{1}{\beta} \leq m < 1 \quad (1.3)$$

Utilizzando la notazione posizionale, un numero  $x \in \mathbb{R}$ , con  $x \neq 0$ , può, allora, essere rappresentato nella seguente forma

$$x = \pm (.d_1d_2d_3\dots)\beta^p \quad (1.4)$$

che è anche chiamata *rappresentazione normalizzata*. La rappresentazione più usuale di tipo *misto* corrisponde a porre  $x = \pm(0.00\dots 0d_1d_2d_3\dots)_\beta$ , se  $p \leq 0$  e  $x = \pm(d_1d_2d_3\dots d_p.d_{p+1}d_{p+2}\dots)_\beta$ , se  $p > 0$ . Il numero *zero* è rappresentato dal simbolo 0.

+	0	1
0	0	1
1	1	10

×	0	1
0	0	0
1	0	1

Tabella 1.2: Addizione e moltiplicazione in base 2.

Per le *operazioni aritmetiche* sui numeri rappresentati in una base  $\beta > 1$ , valgono le stesse regole e proprietà formali note per l'aritmetica in base 10; naturalmente, si devono usare nuove tavole dell'*addizione* e della *moltiplicazione*. In Tabella 1.2 sono riportate come esempio quelle relative alla base  $\beta = 2$ .

## 1.2.2 Conversione della rappresentazione di un numero reale

La conversione di un numero in base  $\beta$  alla forma decimale può essere ottenuta direttamente dall'espressione (1.2) operando in aritmetica decimale. La procedura è illustrata dal seguente esempio

$$(257)_8 = 2 \cdot 8^2 + 5 \cdot 8 + 7 = (2 \times 8 + 5) \times 8 + 7 = 175_{10}$$

Sottolineiamo il modo “nidificato” (*nested*) con il quale si è calcolato il valore del polinomio di secondo grado per il valore 8 della variabile indipendente; con tale procedura, nota nel caso di un generico polinomio come *algoritmo di Horner-Ruffini*<sup>3</sup>, sono richieste solo 2 moltiplicazioni.

In modo analogo si procede per un numero frazionario. Si esamini il seguente esempio.

$$\begin{aligned}
 (.257)_8 &= 2 \cdot 8^{-1} + 5 \cdot 8^{-2} + 7 \cdot 8^{-3} = \\
 &= \left(\left(\frac{7}{8} + 5\right)/8 + 2\right)/8 = (5.875/8 + 2)/8 = 2.734375/8 = (0.341796875)_{10}
 \end{aligned}$$

L'operazione inversa, cioè il passaggio dalla *rappresentazione decimale* alla *rappresentazione in una generica base  $\beta > 1$*  è illustrata nei seguenti esempi.

► **Esempio 1.1** Fornire la rappresentazione in base 16 e in base 2 del numero  $a_1 = (0.1)_{10}$   
 La rappresentazione in base 16 del numero  $a_1$  ha la seguente forma

$$a_1 = (0.1)_{10} = \frac{d_1}{16} + \frac{d_2}{16^2} + \frac{d_3}{16^3} + \dots$$

ove  $0 \leq d_i < 16$ . Moltiplicando ambo i membri della relazione precedente per 16, si ottiene

$$16a_1 = (1.6)_{10} = d_1 + \frac{d_2}{16} + \frac{d_3}{16^2} + \dots$$

da cui  $d_1 = 1$  e

$$a_2 = (0.6)_{10} = \frac{d_2}{16} + \frac{d_3}{16^2} + \frac{d_4}{16^3} + \dots$$

Iterando la procedura, si ha

$$16a_2 = (9.6)_{10} = d_2 + \frac{d_3}{16} + \frac{d_4}{16^2} + \dots$$

e quindi  $d_2 = 9$  e  $a_3 = (0.6)_{10}$  e il risultato diventa periodico. Pertanto  $(0.1)_{10} = (0.19999\dots)_{16}$ . Poiché  $(1)_{16} = (0001)_2$  e  $(9)_{16} = (1001)_2$ , si avrà in definitiva

$$\boxed{(0.1)_{10} = (0.19999\dots)_{16} = (0.0001\ 1001\ 1001\dots)_2}$$

L'esempio evidenzia il fatto importante che un numero reale può avere una rappresentazione *finita* in una particolare base e *infinita* in altre. ■

► **Esempio 1.2** Fornire la rappresentazione in base 8 del numero  $(375)_{10}$ .

Partendo dalla uguaglianza  $(375)_{10} = d_0 + d_1 \cdot 8 + d_2 \cdot 8^2 + \dots$  e dividendo ambo i membri per 8, si ha

$$46 + \frac{7}{8} = \frac{d_0}{8} + d_1 + d_2 \cdot 8 + \dots$$

<sup>3</sup>L'analisi di tale metodo verrà ripresa successivamente nel Capitolo 5 nell'ambito della ricerca degli zeri di un polinomio. Il metodo venne introdotto in maniera indipendente da P. Ruffini (1804) e da W. G. Horner (1819); in effetti il metodo era noto cinque secoli prima ai cinesi (Ch'in Chiu-Shao: *Su-shu Chiu-chang* (Nove sezioni di matematica, 1247)), che lo indicarono come *metodo celestiale*.

Uguagliando le parti intere, e rispettivamente le parti frazionarie, si ottiene  $d_0 = 7$  e

$$(46)_{10} = d_1 + d_2 \cdot 8 + d_3 \cdot 8^2 + \dots$$

Dividendo ancora per 8 si ottiene

$$5 + \frac{6}{8} = \frac{d_1}{8} + d_2 + d_3 \cdot 8 + \dots$$

da cui  $d_1 = 6$  e  $5 = d_2 + d_3 \cdot 8 + d_4 \cdot 8^2 + \dots$ . Infine, dividendo per 8, si ha

$$0 + \frac{5}{8} = \frac{d_2}{8} + d_3 + d_4 \cdot 8 + \dots$$

e quindi  $d_2 = 5$  e  $d_3 = d_4 = \dots = 0$ . Pertanto  $(375)_{10} = (d_2 d_1 d_0)_8 = (567)_8$  ■

Le procedure descritte negli esempi precedenti corrispondono ai seguenti algoritmi generali. Se  $x$  è un numero reale con  $0 < x < 1$  e rappresentato nel sistema decimale, le prime  $k$  cifre  $d_i$  nella rappresentazione in una base  $\beta > 1$  possono essere calcolate iterativamente nel seguente modo, ove  $\text{Int}(x)$  significa la parte intera di  $x$ .

**Algoritmo 1.1** *Calcolo delle prime  $k$  cifre nella rappresentazione in una base  $\beta > 1$  di un numero  $x$ , con  $0 < x < 1$ , rappresentato in base 10.*

```

p = 1; (determinazione dell'esponente)
while Int(x) = 0 do
    x = βx;
    p = p - 1
end;
i = 0; (determinazione delle cifre di)
repeat
    i = i + 1;
    di = Int(x);
    x = β(x - di)
until i = k;

```

La rappresentazione in base  $\beta$  di un numero *intero* positivo  $x$ , rappresentato in base decimale, può essere ottenuta col seguente algoritmo, ove  $\text{mod}(x, \beta)$  indica il resto della divisione fra gli interi  $x$  e  $\beta$ .

**Algoritmo 1.2** *Rappresentazione in base  $\beta > 1$  di un numero intero  $x$  positivo rappresentato in base 10.*

```

c0 = x ; i = -1
repeat
    i = i + 1

```

$$d_i = \text{mod}(c_i, \beta)$$

$$c_{i+1} = \text{Int}(c_i/\beta)$$

until  $c_{i+1} = 0$   
 $k = i$

Si ha quindi:  $x = d_k\beta^k + d_{k-1}\beta^{k-1} + \dots + d_1\beta + d_0$

In definitiva, la rappresentazione in base  $\beta$  di un numero reale positivo  $x$  può essere fatta nel modo seguente. Si pone  $x = x_1 + x_2$ , dove  $x_1 = \text{Int}(x)$ ,  $x_2 = x - x_1 < 1$ . La rappresentazione si ottiene, allora, applicando ad  $x_2$  l'Algoritmo 1.1 e ad  $x_1$  l'Algoritmo 1.2. Considerando ad esempio, il numero  $x = 238.75$  si ha  $x_1 = 238$ ,  $x_2 = 0.75$ , da cui  $x_1 = (EE)_{16}$ ,  $x_2 = (0.C)_{16}$  e, quindi  $238.75 = (EE.C)_{16}$ .

◆ **Esercizio 1.1** Rappresentare in base 10 i seguenti numeri  $(11101110)_2$ ,  $(745)_8$ ,  $(E7)_{16}$ ,  $(0.C8)_{16}$ .

◆ **Esercizio 1.2** Eseguire le operazioni elementari sulle seguenti coppie di numeri

- $(1011100)_2$ ,  $(10001010)_2$
- $(1FA)_{16}$ ,  $(43E0)_{16}$

verificando l'esattezza delle operazioni dopo aver fornito la rappresentazione in base 10 dei numeri.

◆ **Esercizio 1.3** Dare la rappresentazione in base 2, 8, 16 dei seguenti numeri:  $487.$ ,  $0.97$ ,  $-2.654$ ,  $543.625$ .

Convertire la rappresentazione  $(C3C.35)_{16}$  di un numero in base 16 nella sua rappresentazione in base 8.

◆ **Esercizio 1.4** Siano  $t_2$  e  $t_{10}$ , rispettivamente la lunghezza della mantissa nella rappresentazione binaria e decimale di un intero  $n$ . Mostrare che

$$\text{Int}(t_{10}/\log_{10} 2) - 3 \leq t_2 \leq \text{Int}(t_{10}/\log_{10} 2) + 1$$

◆ **Esercizio 1.5** Scrivere e testare un sottoprogramma per convertire interi nella forma ottale (base 8) e binaria e un sottoprogramma per convertire frazioni decimali in forma ottale e binaria. Usare i due sottoprogrammi per la costruzione di un programma che legga numeri decimali e stampi la loro rappresentazione ottale e binaria.

### 1.2.3 Numeri macchina; sistema floating-point

A causa della sua capacità *finita*, un calcolatore non è in grado di rappresentare tutto l'insieme dei numeri reali. Si pone pertanto il problema di *definire*, per ogni  $x$  rappresentato nella forma (1.2), una sua *rappresentazione approssimata* nel calcolatore. Si tratta, in sostanza, di un fatto *tecnico*, ma con importanti implicazioni nel calcolo numerico.

A tale scopo possono essere utilizzate differenti idee; nel seguito analizzeremo, in particolare, un metodo divenuto ormai usuale per il calcolo scientifico e noto come sistema *floating-point* o virgola mobile. Esso, in forma schematica, permette la rappresentazione di un *ampio* intervallo della retta reale con una distribuzione uniforme degli *errori relativi*. L'ampiezza effettiva dell'intervallo dipende dal particolare calcolatore su cui la procedura è implementata<sup>4</sup>.

Naturalmente, l'implementazione *reale* su singoli calcolatori del sistema *floating-point* può differire nei dettagli da quella *ideale* qui presentata. La nostra analisi ha il solo scopo di evidenziare gli aspetti che sono importanti per le implicazioni nel calcolo numerico.

**Definizione 1.1** Si definisce insieme dei numeri macchina (*floating-point*) con  $t$  cifre significative, base  $\beta$  e range  $(L, U)$ , l'insieme dei numeri reali definito nel modo seguente

$$\mathbb{F}(\beta, t, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} = \text{sign}(x)\beta^p \sum_{i=1}^t d_i \beta^{-i} \right\} \quad (1.5)$$

ove  $t, \beta$  sono interi positivi con  $\beta \geq 2$ . Si ha inoltre

$$0 \leq d_i \leq \beta - 1, \quad i = 1, 2, \dots \quad (1.6)$$

$$d_1 \neq 0, \quad L \leq p \leq U \quad (1.7)$$

Usualmente  $U$  è positivo e  $L$  negativo.

In rappresentazione posizionale un numero macchina  $x \neq 0$  viene denotato con

$$x = \pm .d_1 d_2 \dots d_t \beta^p$$

La maggior parte dei calcolatori ha la possibilità di operare con lunghezze diverse di  $t$ , a cui corrispondono, ad esempio, la *semplice* e la *doppia precisione*. Nella Tabella 1.3 sono dati alcuni esempi di sistemi *floating-point reali*.

È importante osservare che l'insieme  $\mathbb{F}$  non è un insieme *continuo* e neppure *infinito*. In effetti, lasciando come esercizio la dimostrazione, si ha che la cardinalità di  $\mathbb{F}$  è data dal numero  $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$ .

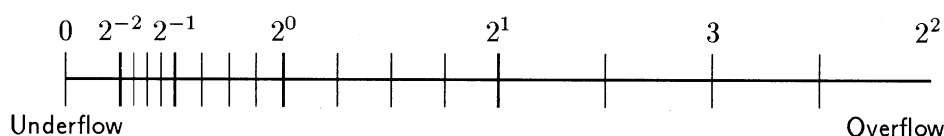
Osserviamo, inoltre, che i numeri dell'insieme  $\mathbb{F}$  sono ugualmente spazati tra le successive potenze di  $\beta$ , ma non su tutto l'intervallo. Per vedere meglio questo fatto importante, analizziamo l'esempio illustrato in Figura 1.1, che corrisponde ai seguenti dati  $\beta = 2$ ,  $t = 3$ ,  $L = -1$ ,  $U = 2$ .

Si tratta di 33 numeri, compreso lo zero. Sono tutti i numeri che possono essere rappresentati nella forma  $0.1d_2d_3 2^p$ , con  $d_i$  cifre binarie, cioè

<sup>4</sup>La possibilità di rappresentare un ampio intervallo della retta reale è importante nelle applicazioni scientifiche, in quanto le costanti fisiche possono differire per vari ordini di grandezza; ad esempio, per la massa di un elettrone si ha  $m_0 \approx 9.11 \cdot 10^{-28}$  g, e per la velocità della luce  $c \approx 2.998 \cdot 10^{10}$  cm/sec.

calcolatore	semplice precisione				doppia precisione		
	$\beta$	$t$	$L$	$U$	$t$	$L$	$U$
DEC 11/780 VAX	2	24	-128	127	56	-128	127
IBM /390	16	6	-64	63	14	-64	63
Cray Y	2	48	-16384	16383	96	-16384	16384
Sperry 2200	2	27	-128	127	60	-1024	1023
IEEE standard chip	2	23	-128	127	53	-1024	1023

Tabella 1.3: Sistemi di numeri macchina per alcuni calcolatori.

Figura 1.1: La parte positiva del sistema *floating point* relativo a  $\beta = 2$ ,  $t = 3$ ,  $L = -1$ ,  $U = 2$ .

.100(-1)	.101(-1)	.110(-1)	.111(-1)
$1/4$	$5/16$	$6/16$	$7/16$
.100(0)	.101(0)	.110(0)	.111(0)
$1/2$	$5/8$	$6/8$	$7/8$
.100(1)	.101(1)	.110(1)	.111(1)
$1$	$5/4$	$6/4$	$7/4$
.100(2)	.101(2)	.110(2)	.111(2)
$2$	$5/2$	$6/2$	$7/2$

Il più piccolo numero positivo dell'insieme è dato da  $0.100 2^{-1} = 1/4$ , mentre il più grande è dato da  $0.111 2^2 = 7/2$ . Con tale sistema si possono pertanto rappresentare numeri reali positivi compresi nell'intervallo  $[1/4, 7/2]$ . Più in generale, per un sistema *floating point* generico si possono rappresentare i numeri reali positivi compresi nell'intervallo  $[\beta^{-1}\beta^L, (1 - \beta^{-t})\beta^U]$ . Naturalmente per i numeri negativi si ragiona in modo simmetrico.

Nella Figura 1.2 è schematizzata una possibile rappresentazione interna ad un calcolatore dell'insieme dei numeri macchina, caratterizzato da  $\beta = 2$ ,  $t = 24$ ,  $U - L + 1 = 256$ . Il primo gruppo di otto bit (*byte*) determina, nella rappresentazione in base 2, il valore della caratteristica; per essa si utilizza, in generale la rappresentazione in traslazione, in modo che la configurazione nulla corrisponda all'esponente  $L$ . Il bit successivo determina il segno (ad esempio positivo se il bit è 0, negativo se il bit è 1). Il gruppo successivo di 23 bit contiene le cifre  $d_2, d_3, \dots, d_{24}$  della

mantissa; nella rappresentazione normalizzata in base 2, si ha  $d_1 = 1$  e quindi non è necessario memorizzare il primo bit della mantissa. La lunghezza della configurazione considerata è di 32 bit (4 byte). Per la precisione doppia tale lunghezza diventa di 8 bytes.

Nelle macchine con aritmetica a base 16 l'insieme dei numeri macchina è dato da  $\mathbb{F}(16, 6, -64, 63)$  per la precisione semplice e  $\mathbb{F}(16, 14, -64, 63)$  per la precisione doppia. Tenendo conto che ogni cifra è rappresentata da *mezzo byte*, si può vedere che, anche in questo caso, occorrono 32 bit per rappresentare un numero macchina, mentre in precisione doppia occorrono 64 bits. Un esempio di calcolatori che utilizzano tale rappresentazione è fornito dalla serie IBM /370.

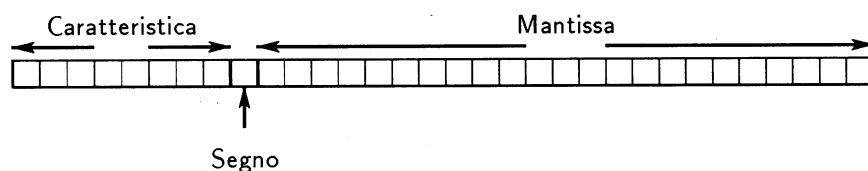


Figura 1.2: Rappresentazione di un numero macchina.

Vediamo, ora, come rappresentare un numero *reale positivo*  $x$ , dato nella forma (1.2) in un sistema di numeri macchina  $\mathbb{F}(\beta, t, L, U)$ . Si possono presentare i seguenti casi

1. Il numero  $x$  è tale che  $L \leq p \leq U$  e  $d_i = 0$  per  $i > t$ ; allora  $x$  è un *numero macchina* ed è rappresentato *esattamente*.
2. La caratteristica  $p$  non appartiene all'intervallo  $[L, U]$ . Il numero non può essere rappresentato esattamente. Se  $p < L$ , come ad esempio  $1/8$  nell'insieme di Figura 1.1, si dice che si verifica un *underflow*; *solitamente* si assume come valore approssimato del numero  $x$  il numero *zero*. Alcuni compilatori, tuttavia, in presenza di questa situazione danno un avvertimento (*warning*); in particolari situazioni, infatti, l'introduzione di questi errori può portare a situazioni inaccettabili.

Se  $p > U$  si verifica un *overflow* e solitamente non si effettua nessuna approssimazione, ma il sistema di calcolo dà un avvertimento più drastico, come ad esempio, l'arresto del calcolo.

3. La caratteristica  $p$  appartiene all'intervallo  $[L, U]$ , ma le cifre  $d_i$ , per  $i > t$ , non sono tutte nulle. In questo caso si pone il problema di scegliere un suo rappresentante in  $\mathbb{F}$ . Tale operazione viene indicata comunemente come operazione di *arrotondamento* (rounding), anche se in realtà possono essere utilizzate tecniche di tipo diverso.



### 1.2.4 Operazione di arrotondamento

Sia  $x$  un numero *reale positivo* dato nella forma (1.2), diverso dallo zero e tale che  $p \in [L, U]$ . Denotiamo con  $\text{fl}(x)$  il numero macchina, cioè  $\text{fl}(x) \in \mathbb{F}$ , che si ottiene in uno dei modi seguenti

**Troncamento:**  $\text{fl}(x) = \text{tronc}(x) = \beta^p \sum_{i=1}^t d_i \beta^{-i}$

**Arrotondamento:**  $\text{fl}(x) = \beta^p \text{tronc} \left( \sum_{i=1}^{t+1} d_i \beta^{-i} + \frac{1}{2} \beta^{-t} \right)$

La definizione si adatta in maniera ovvia quando  $x$  è negativo.

Osserviamo che, se  $d_{t+1} < \beta/2$  allora le due operazioni danno i medesimi risultati, altrimenti essi differiscono di  $\beta^{p-t}$ . Nell'operazione di arrotondamento si può verificare un *overflow*; nell'esempio di Figura 1.3, si ha ad esempio  $\text{fl}(15/4) = 4$ .

Sottolineiamo il fatto che nella sostituzione di  $x$  con  $\text{fl}(x)$  si ha l'*origine degli errori di arrotondamento*. È importante, quindi, stabilire una stima, cioè una *maggiorazione* della quantità  $x - \text{fl}(x)$ . In questa direzione abbiamo il seguente risultato.

**Proposizione 1.1** *Sia  $x$  un numero reale rappresentato in base  $\beta$*

$$x = \left( \sum_{i=1}^{\infty} d_i \beta^{-i} \right) \beta^p, \quad d_1 \neq 0, \quad p \in [L, U]$$

*Allora, se non si verifica una situazione di overflow, si ha la seguente maggiorazione*

$$\boxed{\left| \frac{\text{fl}(x) - x}{x} \right| \leq k \beta^{1-t}} \quad (1.8)$$

*ove si ha  $k = 1$  nel caso del troncamento (chopping) e  $k = 1/2$  nel caso di arrotondamento (rounding).*

**DIMOSTRAZIONE.** Considerando ad esempio il caso dell'arrotondamento, si ha

$$|\text{fl}(x) - x| \leq \frac{1}{2} \beta^{-t} \beta^p$$

Ricordando che, dal fatto che  $d_1 \neq 0$ , si ha  $m \geq \beta^{-1}$ , si ottiene

$$\left| \frac{\text{fl}(x) - x}{x} \right| \leq \frac{\frac{1}{2} \beta^{-t} \beta^p}{|m| \beta^p} \leq \frac{\frac{1}{2} \beta^{-t}}{\beta^{-1}} = \frac{1}{2} \beta^{1-t}$$

■

Si osservi che nella situazione di underflow l'errore relativo, che è del 100% poiché  $\text{fl}(x) = 0$ , non soddisfa alla maggiorazione (1.8).

La quantità  $\boxed{\text{eps} = k\beta^{1-t}}$  è detta *precisione macchina* nel fissato *sistema floating point*. La sua importanza numerica è data dalla seguente caratterizzazione: *eps* è il più piccolo numero macchina positivo tale che

$$\boxed{\text{fl}(1 + \text{eps}) > 1} \quad (1.9)$$

Se, ad esempio,  $\beta = 10$ ,  $t = 8$ , si ha, nel caso di arrotondamento,  $\text{eps} = 5 \cdot 10^{-8}$ , da cui

$$x = 1 + \text{eps} = 1.00000005 = 0.100000005 \cdot 10^1$$

con  $x \notin \mathbb{F}$ ; pertanto

$$\text{fl}(x) = 1.0000001 > 1$$

Al contrario,  $\text{fl}(1.00000004) = 1$ .

In sostanza, su un calcolatore che utilizza una particolare rappresentazione floating point, non ha senso cercare approssimazioni di numeri con precisione relativa inferiore alla corrispondente *precisione macchina eps*.

A partire dalla proprietà (1.9), è possibile calcolare *eps* attraverso software; nel seguente segmento di programma è indicata una procedura in FORTRAN.

```
C  calcolo della precisione macchina
      EPS=1.
1     EPS=0.5*EPS
      EPS1=EPS+1
      IF(EPS1.GT.1) GO TO 1
      EPS=2.*EPS
```

In realtà, il valore calcolato nel precedente programma corrisponde alla precisione macchina se la base è 2, altrimenti ne rappresenta soltanto un valore approssimato, sufficiente, comunque, nelle applicazioni. Si lascia come esercizio l'implementazione di un programma che calcoli per un determinato calcolatore la *base* del sistema floating point usato e che a partire dal valore della base così ricavato determini il valore esatto della precisione macchina. Come suggerimento si osservi, che  $\text{fl}(n + \text{eps}) > n$ , per ogni  $n$  intero compreso tra 1 e  $\beta - 1$ , ma che  $\text{fl}(\beta + \text{eps}) = \beta$ . Operando in maniera analoga è anche possibile scoprire se il calcolatore arrotonda o tronca.

Osserviamo, infine, che un modo equivalente, e più utile nelle considerazioni che seguiranno, di esprimere (1.8) è il seguente

$$\boxed{\text{fl}(x) = x(1 + \epsilon), \quad \text{con } |\epsilon| \leq \text{eps}} \quad (1.10)$$

### 1.2.5 Aritmetica in virgola mobile

Sull'insieme dei *numeri macchina*, che rappresenta una *approssimazione* dell'insieme dei numeri reali, si devono ridefinire le *operazioni aritmetiche*. Si tratta, cioè, di introdurre una *aritmetica di macchina*.

Il risultato di un'operazione aritmetica, anche nel caso in cui gli operandi *sono numeri macchina*, può *non essere un numero macchina* per due motivi: si ha una situazione di underflow o di overflow, oppure il risultato ha un numero di cifre superiore alla precisione  $t$ . Per il seguito consideriamo in particolare questa seconda eventualità.

Siano, pertanto,  $x, y$  due numeri appartenenti ad un particolare sistema floating point  $\mathbb{F}(\beta, t, L, U)$  e il risultato di un'operazione aritmetica con operandi  $x, y$  sia tale che  $p \in [L, U]$ . Introduciamo, allora le seguenti definizioni di *operazione macchina* o operazioni floating point

$$x \oplus y = \text{fl}(x + y), \quad x \ominus y = \text{fl}(x - y), \quad x \otimes y = \text{fl}(x * y), \quad x \oslash y = \text{fl}(x/y)$$

Analogamente si procede per la definizione di altre operazioni, che possono essere considerate “elementari”, come, ad esempio, l'operazione di estrazione di radice quadrata, il calcolo delle funzioni trigonometriche, del logaritmo. La definizione data di operazione macchina è naturalmente semplificata e può non corrispondere a delle situazioni reali, solitamente più sofisticate, ma essa è, tuttavia, sufficiente per la nostra analisi.

L'*errore relativo*, introdotto da una operazione macchina può essere calcolato utilizzando la maggiorazione (1.8). Indicando con il simbolo  $\odot$  una qualunque delle operazioni macchina precedenti e corrispondente all'operazione aritmetica  $\circ$ , si ha

$$\boxed{x \odot y = (x \circ y)(1 + \epsilon)} \quad \text{con} \quad |\epsilon| \leq \text{eps} \quad (1.11)$$

La formula (1.11) può essere riscritta in modi diversi, alcuni dei quali possono suggerire un'utile interpretazione degli errori di arrotondamento. Si consideri, come esempio, la seguente

$$\text{fl}(x + y) = x(1 + \epsilon) + y(1 + \epsilon) \quad (1.12)$$

L'equazione (1.12) dice, in sostanza, che l'operazione di *somma macchina* può essere *interpretata* come la somma *aritmetica* esatta dei due numeri:  $x(1 + \epsilon)$ ,  $y(1 + \epsilon)$ , che rappresentano delle *perturbazioni* dei numeri dati  $x, y$ .

L'osservazione ora fatta rappresenta una prima semplice applicazione di una tecnica più generale per lo studio della *propagazione* degli errori di arrotondamento. Per l'idea su cui si basa, tale tecnica è detta *analisi all'indietro*, o backward analysis, e il suo studio verrà ripreso successivamente.

È importante osservare che per le *operazioni in virgola mobile* gran parte delle proprietà dell'aritmetica nel *campo reale* possono *non essere più valide*. In

particolare, ad esempio si ha

$$x \oplus y = x, \quad \text{se } |y| < \frac{\text{eps}}{\beta} |x|$$

Inoltre, le operazioni macchina possono non verificare la proprietà associativa e la proprietà distributiva.

Analizziamo, come esemplificazione, il seguente calcolo eseguito in aritmetica di macchina definita da  $\beta = 10$ ,  $t = 8$ .

► **Esempio 1.3** Dati i numeri macchina

$$\begin{aligned} a &:= 0.23371258 \cdot 10^{-4} \\ b &:= 0.33678429 \cdot 10^2 \\ c &:= -0.33677811 \cdot 10^2 \end{aligned}$$

Si ha

$$\begin{aligned} \text{(I)} \quad (a \oplus b) \oplus c &= 0.33678452 \cdot 10^2 \ominus 0.33677811 \cdot 10^2 \\ &= 0.64100000 \cdot 10^{-3} \\ \text{(II)} \quad a \oplus (b \oplus c) &= 0.23371258 \cdot 10^{-4} \oplus 0.61800000 \cdot 10^{-3} \\ &= 0.64137126 \cdot 10^{-3} \end{aligned}$$

Il risultato esatto è dato da  $a + b + c = 0.641371258 \cdot 10^{-3}$ .

Il calcolo (II) fornisce come risultato il numero  $\text{fl}(a + b + c)$ , cioè un risultato *ottimale* nell'aritmetica di macchina fissata. Il risultato ottenuto con (I) ha, invece, soltanto *tre cifre significative esatte*, con un *errore relativo*  $\approx 5.78 \cdot 10^{-2}$ . Vediamone le cause.

In ambedue i calcoli si è effettuata una sottrazione di due numeri con lo *stesso segno e vicini*, cioè con un certo numero di cifre significative *uguali*. Nel calcolo (II) ciò è avvenuto nell'operazione  $b \oplus c$ , mentre in (I) si è verificato quando il risultato di  $(a \oplus b)$  è stato sommato a  $c$ . In situazioni di questo genere si verifica una *cancellazione*, cioè le cifre comuni spariscono, con conseguente introduzione di quantità *spurie*<sup>5</sup>, ad esempio zeri. Il risultato dell'operazione, avendo un numero di cifre inferiore rispetto agli addendi, è rappresentabile *esattamente*, cioè è un *numero macchina*. Tuttavia, mentre nel secondo caso l'operazione è stata *innocua*, nel primo ha prodotto un *grosso errore*.

La differenza tra i due casi è la seguente. Nel secondo la cancellazione è avvenuta tra i numeri di partenza, che per ipotesi sono supposti *non affetti da errore*; nel primo caso, invece, uno degli addendi, precisamente il termine  $a \oplus b$  è un *valore arrotondato* di  $a + b$ . Osserviamo, anche, che i due numeri  $a$  e  $b$  hanno ordine di grandezza diverso e nella somma il numero  $a$  contribuisce soltanto con le prime tre cifre, mentre le altre vanno *perse* nell'arrotondamento e, quindi, non compaiono più nel *risultato finale*. Una situazione di questo tipo, in cui si verifica una perdita di cifre, è anche indicata come *effetto smearing*. È praticamente la perdita di cifre che viene *amplificata* dall'operazione di *cancellazione*. ■

Dall'esempio precedente si ricava, in particolare, il fatto che la proprietà *associativa* può non essere valida per l'operazione di *somma numerica*. Inoltre, l'esempio ha

<sup>5</sup>Il termine *spurio* qui sta a significare mancanza di informazione.

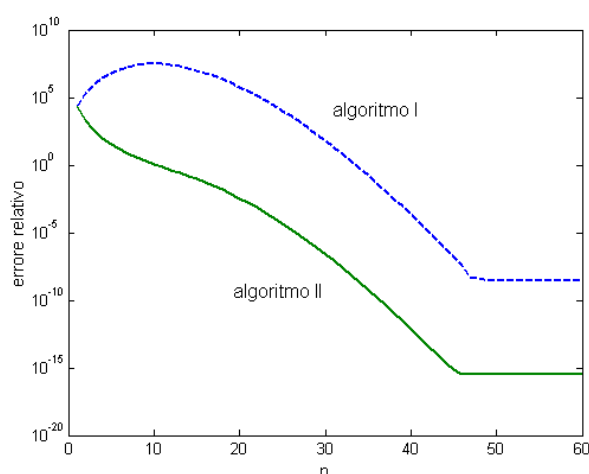


Figura 1.3: Rappresentazione degli errori relativi corrispondenti al calcolo in doppia precisione ( $\text{eps} \approx 2.2 \cdot 10^{-16}$ ) di  $e^{-10}$  mediante due diversi algoritmi.

evidenziato nella *cancellazione* una *situazione delicata*. A questo proposito possiamo riassumere le considerazioni precedenti osservando che l'operazione di *cancellazione* non è una operazione pericolosa in se stessa; lo può diventare quando viene eseguita su *numeri arrotondati* (più in generale *approssimati*), in quanto amplifica gli errori contenuti nei dati. Sottolineiamo il fatto che, in sostanza, è proprio la *cancellazione* la causa principale di *instabilità* (cioè dell'amplificazione eccessiva degli errori) negli algoritmi. Come esemplificazione, in Figura 1.3 sono rappresentati gli errori relativi corrispondenti al calcolo di  $e^{-10}$  mediante due differenti algoritmi basati sull'utilizzo del seguente sviluppo in serie

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$$

Più precisamente, nell'algoritmo I si pone nello sviluppo precedente  $x = -10$  e si calcolano per  $n \geq 0$  le seguenti somme parziali

$$s_n = 1 - 10 + \frac{10^2}{2!} - \dots + (-1)^n \frac{10^n}{n!}$$

Nella figura è rappresentata quindi la funzione  $n \rightarrow |s_n - e^{-10}|/e^{-10}$ . I risultati sono ottenuti in *doppia precisione*, corrispondente alla precisione macchina  $\text{eps} \approx 2.2 \cdot 10^{-16}$ .

Nell'algoritmo II si pone  $x = 10$  e si approssima il valore  $e^{-10}$ , per ogni  $n \geq 0$ , nel seguente modo

$$e^{-10} \approx \frac{1}{e^{10}} = \frac{1}{1 + 10 + \frac{10^2}{2!} + \dots + \frac{10^n}{n!}}$$

Come si vede, i risultati più accurati sono quelli ottenuti mediante l'algoritmo II, nel quale non sono presenti cancellazioni. I risultati ottenuti mediante l'algoritmo I si stabilizzano per  $n \geq 59$  sul valore  $4.539992976248485 \cdot 10^{-5}$ , mentre il valore esatto a 16 cifre è dato da  $4.539992962303128 \cdot 10^{-5}$ .

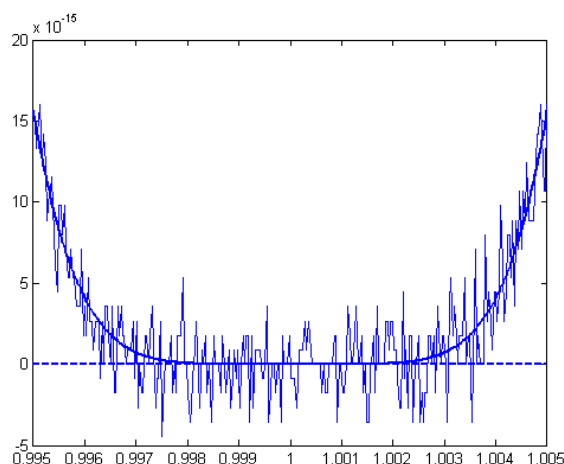


Figura 1.4: Grafico della funzione  $x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1$  calcolata in doppia precisione.

Come ulteriore esemplificazione, la Figura 1.4 rappresenta il grafico della funzione  $(x - 1)^6$ , calcolata in doppia precisione ( $\text{eps} \approx 2.2 \cdot 10^{-16}$ ) mediante lo sviluppo  $x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1$ . Osserviamo che l'errore relativo, ad esempio nel punto  $x = 1.005$ , è del 25%.

◆ **Esercizio 1.6** I seguenti numeri sono dati in un calcolatore a base 10 e precisione  $t = 4$

$$(a) 0.4523 \cdot 10^4, \quad (b) 0.2115 \cdot 10^{-3}, \quad (c) 0.2583 \cdot 10^1$$

Eeguire le seguenti operazioni, e indicare l'errore nel risultato, supponendo di arrotondare

1.  $(a) + (b) + (c)$
2.  $(a)/(c)$
3.  $(a) - (b) - (c)$
4.  $(a)(b)/(c)$

◆ **Esercizio 1.7** Dare esempi che mettano in evidenza il fatto che la maggior parte delle leggi, valide per le operazioni elementari nel campo dei numeri reali, possono non essere più valide nell'aritmetica floating-point.

## 1.2.6 Propagazione degli errori

Conoscere come gli errori si propagano attraverso i calcoli è senza dubbio un elemento importante nella valutazione dei risultati ottenuti, ma ugualmente interessante

è il contributo che tale conoscenza può dare alla pianificazione dei calcoli e degli esperimenti.

Introduciamo le idee e i risultati principali attraverso alcuni esempi particolari.

Se  $x_1 = 4.62 \pm 0.02$  e  $x_2 = 2.84 \pm 0.03$ , calcoliamo una limitazione dell'errore corrispondente al calcolo di  $x_1 - x_2$ .

Poiché il massimo valore di  $x_1$  è dato da 4.64 e il minimo valore di  $x_2$  è dato da 2.81, si ha che il massimo valore di  $x_1 - x_2$  è dato da  $4.64 - 2.81 = 1.83$ . Procedendo in maniera analoga, si trova che il minimo valore di  $x_1 - x_2$  è dato da  $4.60 - 2.87 = 1.73$ . Pertanto si ha  $1.73 \leq x_1 - x_2 \leq 1.83$ , da cui

$$\boxed{x_1 - x_2 = 1.78 \pm 0.05}$$

Più in generale, se  $x_1 = \tilde{x}_1 \pm \epsilon_1$ ,  $x_2 = \tilde{x}_2$ , si ha

$$\begin{aligned} \tilde{x}_1 - \epsilon_1 - (\tilde{x}_2 + \epsilon_2) &\leq x_1 - x_2 \leq \tilde{x}_1 + \epsilon_1 - (\tilde{x}_2 - \epsilon_2) \\ \tilde{x}_1 - \tilde{x}_2 - (\epsilon_1 + \epsilon_2) &\leq x_1 - x_2 \leq \tilde{x}_1 - \tilde{x}_2 + (\epsilon_1 + \epsilon_2) \\ x_1 - x_2 &= \tilde{x}_1 - \tilde{x}_2 \pm (\epsilon_1 + \epsilon_2) \end{aligned}$$

In maniera del tutto analoga si ottiene

$$x_1 + x_2 = \tilde{x}_1 + \tilde{x}_2 \pm (\epsilon_1 + \epsilon_2)$$

Procedendo per induzione, si può dimostrare il seguente risultato per un numero arbitrario di termini.

**Proposizione 1.2** *Nella addizione e nella sottrazione, la limitazione sull'errore assoluto nel risultato è la somma delle limitazioni sugli errori negli operandi.*

Naturalmente (e fortunatamente!) la limitazione ottenuta con il procedimento precedente può essere una *sovrastima* dell'errore attuale, in quanto non si è tenuto conto di eventuali compensazioni negli errori.

Consideriamo ora la propagazione degli *errori relativi* nel calcolo di  $y = x_1 - x_2$ , nell'ipotesi che  $x_i \neq 0$  e  $y \neq 0$ . Se  $r_i$ ,  $i = 1, 2$  sono gli errori relativi sui dati, possiamo scrivere

$$\tilde{x}_i = x_i(1 + r_i) \tag{1.13}$$

e quindi

$$\tilde{x}_1 - \tilde{x}_2 = x_1(1 + r_1) - x_2(1 + r_2) = (x_1 - x_2) + x_1 r_1 - x_2 r_2$$

da cui

$$\frac{\tilde{y} - y}{y} = \frac{x_1}{y} r_1 - \frac{x_2}{y} r_2$$

ove  $\tilde{y} = \tilde{x}_1 - \tilde{x}_2$ . Il risultato ottenuto mostra che gli errori relativi nei dati possono essere *amplificati* nel risultato quando si esegue la differenza tra due numeri "vicini".

È la situazione della *cancellazione* che abbiamo esaminato nel paragrafo precedente. Quando è possibile, è opportuno, quindi, riscrivere le formule in maniera da evitare il verificarsi di cancellazioni. Come semplice illustrazione, si consideri il calcolo della seguente espressione

$$y = \sin(1 + x) - \sin(1)$$

per  $|x| \ll 1$ . Ad esempio, per  $x = 0.0123$  e operando con arrotondamento alla terza cifra, si ottiene il valore  $\tilde{y} = 0.006$ , mentre il valore esatto arrotondato alla terza cifra è dato da 0.00658. Utilizzando note regole trigonometriche, si può riscrivere l'espressione data nel seguente modo, "analiticamente" equivalente

$$y = \sin\left(\frac{x}{2}\right) \cos\left(1 + \frac{x}{2}\right)$$

In questa forma si ottiene come risultato numerico il valore 0.00654.

Esaminiamo, infine, la propagazione degli *errori relativi* nelle operazioni di *prodotto* e *divisione*. Dati due numeri  $x_1, x_2$  come in (1.13), abbiamo

$$\tilde{x}_1 \tilde{x}_2 = x_1(1 + r_1)x_2(1 + r_2) = x_1x_2(1 + r_1)(1 + r_2)$$

L'errore relativo nel prodotto è quindi dato da

$$(1 + r_1)(1 + r_2) - 1 = r_1 + r_2 + r_1r_2 \approx r_1 + r_2, \quad \text{se } |r_1| \ll 1, |r_2| \ll 1$$

Per esempio, se  $r_1 = 0.02$ ,  $r_2 = -0.01$ , allora l'errore relativo nel prodotto è  $0.0098 \approx 0.01$ . Procedendo in maniera analoga, si trova per l'errore relativo nel quoziente  $x_1/x_2$

$$\frac{1 + r_1}{1 + r_2} - 1 = \frac{r_1 - r_2}{1 + r_2} \approx r_1 - r_2, \quad \text{se } |r_1| \ll 1, |r_2| \ll 1$$

Se  $\rho_i$  sono le limitazioni per gli errori relativi nei dati  $x_i$ , si può quindi concludere che  $\rho_1 + \rho_2$  è una limitazione sia di  $|r_1 + r_2|$  che di  $|r_1 - r_2|$  e pertanto si ha il seguente risultato.

**Proposizione 1.3** *Nella moltiplicazione e nella divisione la limitazione dell'errore è maggiorata dalla somma delle limitazioni degli errori negli operandi.*

Vediamo una interessante applicazione del risultato precedente. Supponiamo che per il calcolo dell'espressione  $y = x_1x_2$  la quantità  $x_1$  sia nota con una accuratezza assegnata, mentre  $x_2$  sia da calcolare. Consideriamo, allora, il problema pratico di stimare l'accuratezza con la quale valutare la quantità  $x_2$ , tenendo presente che, in generale, maggiore accuratezza significa maggiore "difficoltà" nel calcolo (ad esempio, la richiesta di strumenti più precisi, eccetera).

Dal risultato precedente abbiamo la seguente risposta al problema. Dal momento che la limitazione sull'errore relativo in  $y$  è uguale alla somma delle limitazioni sugli errori relativi in  $x_1$  e in  $x_2$ , non è *conveniente* cercare una valutazione di  $x_2$  "molto" più accurata di quella di  $x_1$ .



I risultati precedenti possono essere generalizzati nel seguente modo. Supponiamo di avere un generico problema definito da  $y = \phi(x)$ , con

$$\phi : D \rightarrow \mathbb{R}^m, \quad \phi(x) = \begin{bmatrix} \phi_1(x_1, \dots, x_n) \\ \vdots \\ \phi_m(x_1, \dots, x_n) \end{bmatrix}, \quad \phi \in C^1(D)$$

Sia  $\tilde{x}$  un valore approssimato di  $x$ . Indichiamo, allora, con  $\Delta x_i := \tilde{x}_i - x_i$  (risp.  $\Delta x := \tilde{x} - x$ ) l'errore assoluto relativo a  $\tilde{x}_i$  (rispettivamente relativo a  $\tilde{x}$ ). L'errore relativo relativo a  $\tilde{x}_i$  è definito dalla quantità

$$\epsilon_{\tilde{x}_i} := \frac{\tilde{x}_i - x_i}{x_i}, \quad \text{se } x_i \neq 0$$

Se indichiamo con  $\tilde{y} := \phi(\tilde{x})$  il risultato corrispondente a  $\tilde{x}$ , sviluppando in serie con arresto ai termini del primo ordine, si ha

$$\begin{aligned} \Delta y_i := \tilde{y}_i - y_i &= \phi_i(\tilde{x}) - \phi_i(x) \approx \sum_{j=1}^n (\tilde{x}_j - x_j) \frac{\partial \phi_i(x)}{\partial x_j} \\ &= \sum_{j=1}^n \frac{\partial \phi_i(x)}{\partial x_j} \Delta x_j, \quad i = 1, \dots, m \end{aligned} \quad (1.14)$$

Le quantità  $\partial \phi_i(x) / \partial x_j$  indicano come l'errore sulla componente  $x_j$  del dato si *amplifica* sulla componente  $y_i$  del risultato.

Se  $y_i \neq 0$  per  $i = 1, \dots, m$  e  $x_j \neq 0$ , per  $j = 1, \dots, n$ , si ha la seguente formula per gli *errori relativi*

$$\epsilon_{y_i} \approx \sum_{j=1}^n \left[ \frac{x_j}{\phi_i(x)} \frac{\partial \phi_i(x)}{\partial x_j} \right] \epsilon_{x_j} \quad (1.15)$$

I numeri  $(x_j / \phi_i) \partial \phi_i / \partial x_j$  danno una indicazione sulla amplificazione degli errori relativi e sono chiamati *numeri di condizionamento* del problema.

### 1.2.7 Condizionamento di un problema

Dato un *problema matematico* possiamo, in maniera schematica, distinguere, per quanto riguarda la *propagazione degli errori*, il comportamento del *problema* e il comportamento di un *particolare algoritmo* utilizzato per risolvere tale problema. Nel primo caso, facendo l'ipotesi *ideale* di essere in grado di risolvere *esattamente* il problema, si è interessati a vedere come eventuali perturbazioni sui dati del problema si trasmettono sui risultati. Per caratterizzare un problema rispetto a questo tipo di comportamento si utilizza comunemente il termine di *condizionamento*. Più precisamente, si parla di problema *bencondizionato* (o *malcondizionato*) a seconda

che nel particolare contesto le perturbazioni sui dati non influenzino (o influenzino) *eccessivamente* i risultati. Nel caso di un *algoritmo*, per indicare il suo comportamento rispetto alla propazione degli errori è più usuale il termine di *stabilità*. Si dirà quindi *algoritmo stabile* (o *instabile*) un algoritmo nel quale la successione delle operazioni non amplifica (o amplifica) eccessivamente gli errori di arrotondamento.

La distinzione tra il *condizionamento* di un problema e la *stabilità* di un algoritmo è importante perché, mentre per un problema bencondizionato è possibile, in generale, trovare algoritmi stabili, per un problema malcondizionato può essere opportuna una sua riformulazione.

Illustreremo ora i concetti precedenti mediante alcuni esempi, incominciando da un esempio di *algoritmo instabile*.

► **Esempio 1.4** Per il calcolo dei seguenti integrali, per  $n=1,2,\dots$

$$I_n = \int_0^1 \frac{x^n}{x+5} dx$$

si può utilizzare la seguente formula ricorrente

$$I_n + 5I_{n-1} = \frac{1}{n} \quad (1.16)$$

che risulta dalla seguente osservazione

$$I_n + 5I_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} dx = \int_0^1 \frac{x^{n-1}(x+5)}{x+5} dx = \int_0^1 x^{n-1} dx = \frac{1}{n}$$

Osservato che il valore di  $I_0$  può essere calcolato analiticamente mediante la formula

$$I_0 = \int_0^5 \frac{dx}{x+5} = [\ln(x+5)]_0^1 = \ln 6 - \ln 5 \approx 0.1823$$

si possono calcolare i successivi valori di  $I_n$ ,  $n = 1, 2, \dots$  ponendo

$$I_n = \frac{1}{n} - 5I_{n-1}$$

In questo modo, operando in aritmetica a quattro cifre decimali, si ottengono i seguenti risultati

$n$	1	2	3	4	5	6
$I_n \approx$	0.0885	0.0575	0.0458	0.0210	0.0950	-0.3083

I risultati ottenuti non sono accettabili in quanto, come si vede facilmente, i valori esatti della successione  $I_n$  sono decrescenti e positivi. La ragione di quanto ottenuto va ricercata nel fatto che l'errore di arrotondamento  $\epsilon$  commesso nel calcolo di  $I_0$ , che è dell'ordine di  $5 \cdot 10^{-5}$ , è moltiplicato per  $-5$  nel calcolo di  $I_1$ , che ha quindi un errore di  $-5\epsilon$ . Allora l'errore in  $I_2$  è  $25\epsilon$ , e via di seguito. A questi errori vanno aggiunti naturalmente quelli commessi nei vari passi del calcolo.

Possiamo, pertanto, concludere che l'algoritmo precedente non è di interesse pratico, almeno per  $n$  grande. Esso rappresenta un esempio di *instabilità numerica*.

Per fornire un algoritmo alternativo osserviamo che la formula ricorrente (1.16) può essere riscritta nella seguente forma

$$I_{n-1} = \frac{1}{5n} - \frac{I_n}{5} \quad (1.17)$$

con la quale è possibile il calcolo di  $I_{n-1}$  mediante il valore di  $I_n$ . Naturalmente, non è possibile conoscere il valore esatto di  $I_n$  per  $n \geq 1$ ; possiamo, tuttavia, stimare tale valore per  $n$  sufficientemente grande osservando che la successione  $I_n$  tende a zero. Proviamo, ad esempio, a porre  $I_{10} = I_9$ . Dalla relazione ricorrente si ottiene allora

$$I_{10} + 5I_9 \approx \frac{1}{10} \Rightarrow I_9 \approx \frac{1}{60} \approx 0.0167$$

Applicando la (1.17) si ottengono i seguenti risultati

$n$	8	7	6	5	4	3	2	1	0
$I_n \approx$	0.0189	0.0212	0.0243	0.0285	0.0343	0.0431	0.0580	0.0884	0.1823

Nel secondo algoritmo l'errore commesso sul valore  $I_9$  viene successivamente diviso per 5 e il valore calcolato di  $I_0$  è esatto a quattro cifre. Si tratta quindi di un algoritmo *stabile*. ■

Vediamo ora alcuni esempi di *problemi malcondizionati*.

► **Esempio 1.5** Calcolo delle radici dell'equazione

$$(x - 2)^2 = 10^{-6}, \quad \text{ossia: } P(x, \alpha) := x^2 - 4x + \alpha = 0, \quad \alpha = 4 - 10^{-6}$$

Per  $\alpha = 4$ , le radici sono coincidenti:  $x_1 = x_2 = 2$ , mentre per  $\alpha = 4 - 10^{-6}$ , che corrisponde a una perturbazione *relativa* del termine noto di  $10^{-6}$ , si hanno le radici  $2 \pm 10^{-3}$ . La perturbazione *relativa* nei dati si è quindi *amplificata* nei risultati di un fattore  $10^3$ . ■

► **Esempio 1.6** Risoluzione del sistema lineare

$$\begin{cases} 6x_1 + 61.5x_2 = 7.259 \\ 61.5x_1 + 630.55x_2 = 74.4843 \end{cases} \quad (1.18)$$

Il sistema (1.18) rappresenta il *sistema delle equazioni normali*, relativo al seguente problema dei minimi quadrati. Date le coppie di punti

$\tau_i$	10.0	10.1	10.2	10.3	10.4	10.5
$f(\tau_i)$	1.000	1.200	1.25	1.267	1.268	1.274

si cercano  $x_1, x_2$ , che minimizzano la funzione

$$\sum_i (f(\tau_i) - x_1 - x_2\tau_i)^2$$

Ponendo uguali a zero le derivate rispetto a  $x_1, x_2$ , si ottiene il sistema (1.18).

La soluzione esatta è data da  $x_1 = -3.44952372$ ,  $x_2 = 0.45457142$ . Supponendo di rappresentare il sistema in una aritmetica floating-point decimale con  $t = 4$ , si ha:  $\text{fl}(630.55) = 630.6$ ,  $\text{fl}(74.4843) = 74.48$ .

La soluzione *esatta* del sistema così perturbato è la seguente

$$x_1 = -2.218222217, \quad x_2 = 0.33444444$$

Un errore relativo di  $10^{-4}$  nei dati ha prodotto un errore relativo dell'ordine dell'unità nei risultati. Sottolineiamo che ciò è dovuto esclusivamente al problema e non ad un metodo numerico o ad una particolare aritmetica floating-point utilizzata nei calcoli.

Il sistema lineare (1.18) è quindi un esempio di problema malcondizionato; rinviamo alla Appendice A per una discussione più generale sui sistemi lineari relativi all'applicazione del metodo dei minimi quadrati.

È interessante osservare che se come base dei polinomi, anziché 1,  $\tau$ , si utilizza la seguente: 1,  $\tau - 10$ , si ottiene al contrario un sistema *bencondizionato* per lo stesso insieme di dati. In questo caso, quindi, una *riformulazione opportuna* del problema ha migliorato il suo condizionamento. ■

► **Esempio 1.7** Approssimazione mediante *funzioni esponenziali*. Si tratta di un problema importante in diverse applicazioni, in particolare la chimica, la biologia, la fisica, dove si hanno fenomeni corrispondenti a reazioni con costanti di velocità differenti. Il seguente esempio, attribuito a Lanczos, mostra come il problema della determinazione *numerica*, a partire da dati sperimentali, dei parametri di una combinazione di funzioni esponenziali può essere un problema estremamente malcondizionato. Infatti, le seguenti tre funzioni

$$f_1(x) = 0.0951 e^{-x} + 0.8607 e^{-3x} + 1.5576 e^{-5x}$$

$$f_2(x) = 0.305 e^{-1.58x} + 2.202 e^{-4.45x}$$

$$f_3(x) = 0.041 e^{-0.5x} + 0.79 e^{-2.73x} + 1.68 e^{-4.96x}$$

nell'intervallo  $0 \leq x \leq 1.2$  danno risultati che coincidono nelle prime due cifre significative; in effetti, si ha  $\max_{0 \leq x \leq 1.2} |f_1(x) - f_2(x)| = 0.0064$ ,  $\max_{0 \leq x \leq 1.2} |f_1(x) - f_3(x)| = 0.0055$ ,  $\max_{0 \leq x \leq 1.2} |f_2(x) - f_3(x)| = 0.0093$ . Come conseguenza, si ha che se i dati sperimentali sono noti a due cifre, non ha praticamente significato considerare il problema della identificazione dei parametri con questo tipo di funzioni. ■

◆ **Esercizio 1.8** Studiare la propagazione dell'errore nel calcolo delle seguenti espressioni

$$(1) x(x+3) + 3; \quad (2) (x+1)^2(x+1); \quad (3) x^2\sqrt{x}; \quad (4) (\sqrt{x^2+1} - x)x$$

$$(5) \frac{\sqrt{x} - \sqrt{y}}{\sqrt{x} + \sqrt{y}}; \quad (6) \frac{\sqrt{x} + \sqrt{y}}{\sqrt{x} - \sqrt{y}}, \quad x, y > 0$$

◆ **Esercizio 1.9** Dato il sistema lineare

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases}$$

considerare la soluzione  $\mathbf{x} = [x_1, x_2]^T$  come funzione dei coefficienti  $a_{ij} \in \mathbb{R}$  e dei termini noti  $b_i$ ,  $i = 1, 2$ ;  $j = 1, 2$ . Calcolare quindi i numeri di condizionamento di tale problema, cercando condizioni sufficienti affinché il problema sia ben condizionato.

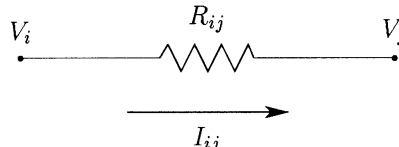


problema verrà trattato nel seguito nell'ambito dello studio del condizionamento del sistema lineare (2.1).

La risoluzione dei sistemi lineari è di fondamentale importanza nella matematica applicata. In effetti, come vedremo nei successivi capitoli del presente volume, i sistemi lineari, da una parte costituiscono lo strumento di base per la formulazione di numerosi modelli matematici e dall'altra rappresentano una tappa fondamentale per la risoluzione iterativa dei modelli non lineari. Rinviando, quindi, agli altri capitoli per ulteriori esemplificazioni significative, esamineremo, ora, alcune situazioni classiche.

► **Esempio 2.1** *Circuiti elettrici*. Consideriamo il circuito elettrico illustrato in Figura 2.1. Le resistenze sono indicate in *ohm* e il voltaggio applicato al generico nodo  $i$  è rappresentato da  $V_i$  volt. L'intensità di corrente che fluisce dal nodo  $i$  al nodo  $j$  è indicata con  $I_{ij}$  ampere. Per ottenere la differenza di potenziale tra due nodi, si applicano le seguenti due leggi.

- (a) La *legge di Ohm*, che mette in relazione l'intensità di corrente che fluisce in una resistenza con la differenza di potenziale agli estremi

$$I_{ij} = \frac{V_i - V_j}{R_{ij}} \quad I_{ij} = \frac{V_i - V_j}{R_{ij}}$$


- (b) La *legge di Kirchhoff*, che stabilisce che la somma algebrica di tutte le correnti che entrano in un nodo deve essere uguale a zero.

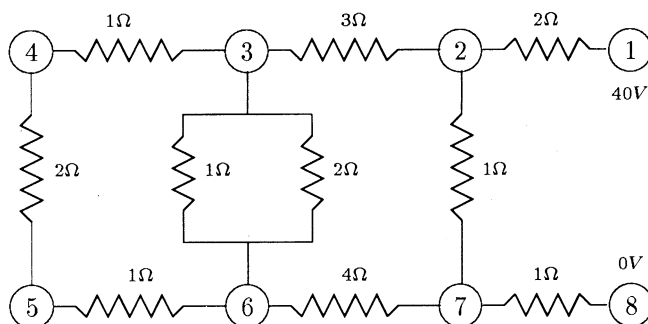


Figura 2.1: Un esempio di rete elettrica.

Osserviamo che la resistenza  $R_{36}$  è calcolata mediante la relazione

$$\frac{1}{R_{36}} = \frac{1}{1} + \frac{1}{2} = \frac{3}{2} \Rightarrow R_{36} = \frac{2}{3}$$

Applicando allora le due leggi precedenti successivamente ai nodi 2, 3, ..., 7 si ottengono le seguenti equazioni

$$\begin{aligned}
 \text{nodo 2} \quad I_{32} + I_{72} + I_{12} &= \frac{V_3 - V_2}{3} + \frac{V_7 - V_2}{1} + \frac{40 - V_2}{2} = 0 \\
 \text{nodo 3} \quad I_{23} + I_{63} + I_{43} &= \frac{V_2 - V_3}{3} + \frac{V_6 - V_3}{2/3} + \frac{V_4 - V_3}{1} = 0 \\
 \text{nodo 4} \quad I_{34} + I_{54} &= \frac{V_3 - V_4}{1} + \frac{V_5 - V_4}{2} = 0 \\
 \text{nodo 5} \quad I_{45} + I_{65} &= \frac{V_4 - V_5}{2} + \frac{V_6 - V_5}{1} = 0 \\
 \text{nodo 6} \quad I_{56} + I_{36} + I_{76} &= \frac{V_5 - V_6}{1} + \frac{V_3 - V_6}{2/3} + \frac{V_7 - V_6}{4} = 0 \\
 \text{nodo 7} \quad I_{67} + I_{27} + I_{87} &= \frac{V_6 - V_7}{4} + \frac{V_2 - V_7}{1} + \frac{0 - V_7}{1} = 0
 \end{aligned}$$

da cui il seguente *sistema lineare*

$$\begin{cases}
 11V_2 - 2V_3 & & -6V_7 = 120 \\
 -2V_2 + 17V_3 - 6V_4 & & -9V_6 = 0 \\
 & -2V_3 + 3V_4 - V_5 & = 0 \\
 & & -V_4 + 3V_5 - 2V_6 = 0 \\
 & -6V_3 & -4V_5 + 11V_6 - V_7 = 0 \\
 -4V_2 & & -V_6 + 9V_7 = 0
 \end{cases}$$

con la seguente matrice dei coefficienti

$$\mathbf{A} = \begin{bmatrix}
 11 & -2 & 0 & 0 & 0 & -6 \\
 -2 & 17 & -6 & 0 & -9 & 0 \\
 0 & -2 & 3 & -1 & 0 & 0 \\
 0 & 0 & -1 & 3 & -2 & 0 \\
 0 & -6 & 0 & -4 & 11 & -1 \\
 -4 & 0 & 0 & 0 & -1 & 9
 \end{bmatrix}$$

Si vede facilmente che è possibile scalare le righe in maniera tale che la matrice che ne risulti sia simmetrica; in altre parole, esiste una opportuna matrice diagonale  $\mathbf{D}$  tale che  $\mathbf{D} \mathbf{A}$  è simmetrica. Rinviando, all'Appendice A per le relative definizioni, si può anche vedere che  $\mathbf{A}$  è *irriducibile* e a *predominanza diagonale* (stretta per almeno una riga) ed è, pertanto, *invertibile*. Osservando, inoltre, che gli elementi fuori dalla diagonale hanno tutti lo stesso segno, che risulta contrario a quello degli elementi sulla diagonale, si ha che  $\mathbf{A}$  è una M-matrice, ossia la matrice inversa  $\mathbf{A}^{-1}$  ha elementi non negativi. Ne segue che la soluzione del sistema è, come fisicamente ragionevole, non negativa. Nel caso particolare che stiamo considerando, si ottiene, con il metodo di eliminazione che esamineremo nel seguito, la seguente soluzione

$$\mathbf{V} = \begin{bmatrix} V_2 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \\ V_7 \end{bmatrix} = \begin{bmatrix} 19.3991 \\ 15.7940 \\ 15.6223 \\ 15.2790 \\ 15.1073 \\ 10.3004 \end{bmatrix}$$

► **Esempio 2.2** *Analisi di strutture.* In Figura 2.2 è mostrato un esempio di struttura statica bidimensionale (un telaio). Le forze  $F$  rappresentano tensioni e compressioni agenti sulle aste che compongono la struttura. Le reazioni esterne  $H_2$ ,  $V_2$  e  $V_3$  sono forze corrispondenti all'interazione della struttura con il supporto. Nell'esempio rappresentato in figura si suppone che il nodo **2** sia vincolato, e quindi possa trasmettere al supporto forze con componenti sia orizzontali che verticali. Si suppone, invece, che il nodo **3** possa scivolare sul supporto, in modo che la struttura possa trasmettere solo forze verticali.

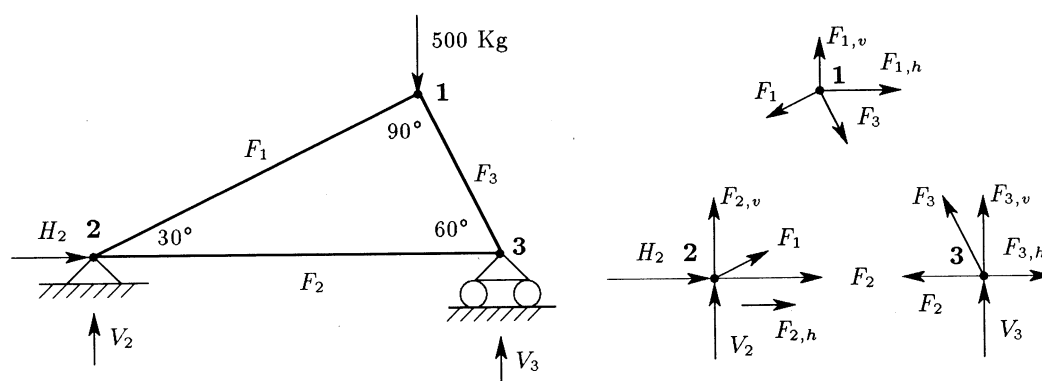


Figura 2.2: Forze in una struttura in condizioni statiche.

Consideriamo, ora, il problema del calcolo delle forze  $F_i$ ,  $i = 1, 2, 3$  e  $H_2$ ,  $V_2$  e  $V_3$ , quando nei nodi della struttura vengono applicate delle forze esterne  $E_i$ ,  $i = 1, 2, 3$ . Indichiamo rispettivamente con  $E_{i,h}$  e  $E_{i,v}$  le componenti orizzontali e verticali delle forze  $E_i$ . Studiando il sistema in condizioni statiche, si ha che in ciascun nodo la somma delle componenti orizzontali delle forze agenti in quel nodo, e rispettivamente delle componenti verticali, deve essere nulla (*legge di equilibrio*). Si hanno allora le seguenti relazioni, nella scrittura delle quali si è seguita la convenzione che le forze dirette verso destra e verso l'alto sono positive.

$$\begin{aligned} \text{nodo 1} & \begin{cases} -F_1 \cos 30^\circ + F_3 \cos 60^\circ + F_{1,h} = 0 \\ -F_1 \sin 30^\circ - F_3 \sin 60^\circ + F_{1,v} = 0 \end{cases} \\ \text{nodo 2} & \begin{cases} F_1 \cos 30^\circ + F_2 + H_2 + F_{2,h} = 0 \\ F_1 \sin 30^\circ + V_2 + F_{2,v} = 0 \end{cases} \\ \text{nodo 3} & \begin{cases} -F_2 - F_3 \cos 60^\circ + F_{3,h} = 0 \\ F_3 \sin 60^\circ + V_3 + F_{3,v} = 0 \end{cases} \end{aligned}$$

Si ha, quindi, il seguente sistema lineare

$$\begin{bmatrix} -\cos 30^\circ & 0 & \cos 60^\circ & 0 & 0 & 0 \\ -\sin 30^\circ & 0 & -\sin 60^\circ & 0 & 0 & 0 \\ \cos 30^\circ & 1 & 0 & 1 & 0 & 0 \\ \sin 30^\circ & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & -\cos 60^\circ & 0 & 0 & 0 \\ 0 & 0 & \sin 60^\circ & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ H_2 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} -F_{1,h} \\ -F_{1,v} \\ -F_{2,h} \\ -F_{2,v} \\ -F_{3,h} \\ -F_{3,v} \end{bmatrix} \quad (2.3)$$



A differenza della matrice relativa all'esempio precedente, la matrice dei coefficienti del sistema (2.3) non è a predominanza diagonale. Facciamo, anzi, osservare che il minore principale di ordine 2 è nullo. Questo fatto comporta, come vedremo nel seguito, l'utilizzo nell'applicazione del metodo di sostituzione di opportune permutazioni delle righe o delle colonne (metodo pivotale). Si può verificare, tuttavia, che il determinante della matrice dei coefficienti è uguale a 1.0 e quindi che la matrice è invertibile, con matrice inversa data da

$$\mathbf{A}^{-1} = \begin{bmatrix} -0.8660 & -0.5000 & 0. & 0. & 0. & 0. \\ -0.2500 & 0.4330 & 0. & 0. & -1.0000 & 0. \\ 0.5000 & -0.8660 & 0. & 0. & 0. & 0. \\ 1.0000 & 0. & 1.0000 & 0. & 1.0000 & 0. \\ 0.4330 & 0.2500 & 0. & 1.0000 & 0. & 0. \\ -0.4330 & 0.7500 & 0. & 0. & 0. & 1.0000 \end{bmatrix}$$

La conoscenza della matrice inversa permette di calcolare le forze in corrispondenza a differenti carichi esterni. Osserviamo, infatti, che la struttura è completamente caratterizzata dalla matrice  $\mathbf{A}$  (detta anche *matrice stiffness*), mentre, il vettore dei termini noti fornisce il carico dovuto alle forze esterne. Ad esempio, nella situazione illustrata in Figura 2.2, nella quale le forze esterne sono tutte nulle, salvo la forza verticale nel nodo  $\mathbf{1}$  rivolta verso il basso e uguale a -500 (il segno meno segue dalla convenzione fatta in precedenza), si ha come vettore dei termini noti  $\mathbf{b} = [0, 500, 0, 0, 0, 0]^T$ , a cui corrisponde la soluzione

$$\begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ H_2 \\ V_2 \\ V_3 \end{bmatrix} = \mathbf{A}^{-1}\mathbf{b} = \begin{bmatrix} -250.0000 \\ 216.5064 \\ -433.0127 \\ 0. \\ 125.0000 \\ 375.0000 \end{bmatrix}$$

Concludiamo, osservando che i singoli elementi della matrice inversa hanno un importante significato applicativo. Ciascuno di essi, infatti, rappresenta la variazione in un'incognita dovuta a una variazione unitaria in uno dei carichi esterni. Ad esempio, l'elemento  $a_{21}^{-1}$  indica il cambiamento della seconda incognita ( $F_2$ ) dovuto a una variazione unitaria nel carico esterno  $F_{1,h}$ . In particolare, la presenza di elementi nulli nella matrice inversa significa che certe incognite non sono influenzate dal carico; ad esempio,  $a_{13}^{-1} = 0$  significa che  $F_1$  non è influenzata da variazioni nel carico  $F_{2,h}$  (come è naturale, dal momento che il nodo  $\mathbf{2}$  è vincolato).

► **Esempio 2.3** *Modello economico di Leontief*. Consideriamo una economia basata su  $n$  industrie, ognuna delle quali produce un singolo prodotto. Per produrre il proprio prodotto, ogni industria ha bisogno, come accade usualmente nella realtà, di certe quantità di prodotti forniti da altre industrie, ed eventualmente anche del proprio prodotto. Si pensi, come esemplificazioni alla industria automobilistica, o a una industria di hardware di calcolatori.

Indichiamo con  $a_{ij}$  l'unità di prodotto dell'industria  $i$  richiesto per la produzione di unità di materiale prodotto dall'industria  $j$ . Le costanti  $a_{ij}$  sono dette i *coefficienti tecnici*. Indichiamo, inoltre, con  $x_1, x_2, \dots, x_n$  le quantità di materiali prodotti dalle  $n$  industrie durante un ciclo fissato (ad esempio, un anno) di produzione.

La quantità totale di materiale prodotto  $i$  viene in parte consumato dalle altre industrie, e in parte va a soddisfare le domande dei consumatori. Si ha, pertanto, il seguente bilancio

$$x_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n + d_i, \quad i = 1, 2, \dots, n$$

ove  $d_i$  è la domanda per il prodotto  $i$ . Introducendo la matrice  $\mathbf{A}=[a_{ij}]$  e i vettori  $\mathbf{x}$ ,  $\mathbf{d}$ , di componenti  $x_i$ , e rispettivamente  $d_i$ ,  $i = 1, 2, \dots, n$ , il modello di produzione è descritto dal seguente sistema lineare

$$(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{d} \quad (2.4)$$

■

I risultati teorici relativi alla *risolubilità* del problema (2.1) sono riportati nell'Appendice A, che contiene, più in generale, le nozioni di algebra lineare che sono alla base della formulazione degli algoritmi numerici.

I metodi numerici per la risoluzione dei sistemi lineari possono essere, in maniera schematica, raggruppati nel modo seguente

- *Metodi diretti*, ossia i metodi che, in assenza di errori di arrotondamento, forniscono la soluzione in un numero finito di operazioni. Tali metodi utilizzano, in sostanza, l'idea della *eliminazione di Gauss*, e più in generale, l'idea della *fattorizzazione* della matrice  $\mathbf{A}$  nel prodotto di due matrici *più semplici* (triangolari o ortogonali).
- *Metodi iterativi*, nei quali la soluzione è ottenuta come *limite* di una successione di soluzioni di *sistemi lineari più semplici*. Nella risposta fornita da un metodo iterativo è, quindi, presente usualmente un *errore di troncamento*.

Nei metodi iterativi, a differenza che nei metodi diretti, la matrice dei coefficienti non viene modificata. In tali metodi, quindi, è possibile sfruttare più opportunamente la presenza di elementi nulli nella matrice dei coefficienti. In effetti, i metodi iterativi hanno un particolare interesse per la risoluzione di sistemi a grandi dimensioni e con matrici *sparse*, ossia matrici per le quali il numero degli elementi che possono essere diversi dallo zero è proporzionale all'ordine  $n$ .

### Calcolo dell'inversa di una matrice

Il calcolo dell'inversa di una matrice  $\mathbf{A}$  di ordine  $n$  e non singolare è un problema strettamente legato a quello della risoluzione di un sistema lineare. In effetti, dalla conoscenza dell'inversa  $\mathbf{A}^{-1}$  si può ottenere la soluzione del sistema lineare corrispondente ad un particolare termine noto  $\mathbf{b}$  mediante semplicemente il prodotto di una matrice per un vettore, con un numero di operazioni proporzionale a  $n^2$

$$\mathbf{A}\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

D'altra parte, le colonne  $\mathbf{X}_j$ ,  $j = 1, 2, \dots, n$  della matrice inversa  $\mathbf{A}^{-1}$  possono essere calcolate mediante la risoluzione dei seguenti  $n$  sistemi lineari, tutti con la stessa matrice dei coefficienti

$$\mathbf{A}\mathbf{X}_j = \mathbf{e}_j \Rightarrow \mathbf{A}^{-1} \equiv [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n] \quad (2.5)$$

ove i vettori  $\mathbf{e}_j$ ,  $j = 1, \dots, n$  sono le colonne della matrice identità  $\mathbf{I}_n$ . In altre parole, un metodo numerico per la risoluzione di un sistema lineare può essere utilizzato anche per il calcolo dell'inversa di una matrice. La risoluzione dei sistemi (2.5) risulta semplificata quando della matrice  $\mathbf{A}$  si costruisce una fattorizzazione.

## 2.1 Metodi diretti

Introduciamo l'idea dell'*eliminazione* attraverso un semplice esempio. L'idea verrà, poi, generalizzata e discussa nei paragrafi successivi.

► **Esempio 2.4** Consideriamo la risoluzione del seguente sistema

$$\begin{cases} 2x_1 + 4x_2 + 2x_3 = 4 \\ 4x_1 + 7x_2 + 7x_3 = 13 \\ -2x_1 - 7x_2 + 5x_3 = 7 \end{cases} \quad (2.6)$$

Ricaviamo l'incognita  $x_1$  dalla prima equazione e la sostituiamo nelle altre due equazioni. L'operazione può essere ottenuta sottraendo la prima equazione moltiplicata per 2 dalla seconda equazione, e analogamente sottraendo la prima equazione moltiplicata per -1 dalla terza equazione. I fattori 2 e -1 utilizzati nella trasformazione sono detti *moltiplicatori*. Si ottiene

$$\begin{cases} 2x_1 + 4x_2 + 2x_3 = 4 \\ -x_2 + 3x_3 = 5 \\ -3x_2 + 7x_3 = 11 \end{cases}$$

A questo punto si elimina  $x_2$  dalla terza equazione sottraendo dalla terza equazione la seconda moltiplicata per 3, e si ottiene la seguente struttura triangolare

$$\begin{cases} 2x_1 + 4x_2 + 2x_3 = 4 \\ -x_2 + 3x_3 = 5 \\ -2x_3 = -4 \end{cases}$$

Nella fase successiva, chiamata *sostituzione all'indietro*, si calcola  $x_3$  dalla terza equazione

$$x_3 = \frac{-4}{-2} = 2$$

e si sostituisce il valore ottenuto nella seconda equazione e si ottiene per  $x_2$  il valore

$$x_2 = \frac{5 - 3x_3}{-1} = \frac{5 - 3 \times 2}{-1} = 1$$

e, infine, dalla prima equazione

$$x_1 = \frac{4 - 4x_2 - 2x_3}{2} = \frac{4 - 4 \times 1 - 2 \times 2}{2} = -2$$

e quindi la soluzione del sistema è data dal vettore  $\mathbf{x} = [-2, 1, 2]^T$ .

Introduciamo le seguenti due matrici triangolari

$$\mathbf{U} = \begin{bmatrix} 2 & 4 & 2 \\ 0 & -1 & 3 \\ 0 & 0 & -2 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix} \quad (2.7)$$

In particolare, la matrice  $\mathbf{L}$  ha tutti gli elementi sulla diagonale principale uguali a 1; per tale motivo viene anche indicata come matrice triangolare unitaria. Osserviamo che gli elementi della matrice  $\mathbf{L}$  che sono al di sotto della diagonale principale corrispondono ai moltiplicatori utilizzati nella operazione di eliminazione. Più precisamente, gli elementi  $l_{21} = 2, l_{31} = -1$  nella prima colonna sono i moltiplicatori utilizzati per eliminare l'incognita  $x_1$ , mentre l'elemento  $l_{32} = 3$  è il moltiplicatore utilizzato per eliminare l'incognita  $x_2$ . Facciamo, ora, vedere che tra la matrice  $\mathbf{A}$  e le matrici  $\mathbf{U}$  e  $\mathbf{L}$  sussiste la seguente relazione

$$\mathbf{A} = \mathbf{L} \mathbf{U} \quad (2.8)$$

Tale identità può essere verificata direttamente eseguendo il prodotto, ma, più significativamente, essa può essere giustificata nel seguente modo. Indichiamo con  $\mathbf{a}_i, i = 1, 2, 3$  le righe di  $\mathbf{A}$  e con  $\mathbf{u}_i$  le righe della matrice  $\mathbf{U}$ . Dal momento che durante l'eliminazione la prima riga non viene mai cambiata, si ha  $\mathbf{u}_1 = \mathbf{a}_1$ . Nella prima eliminazione si ha

$$\mathbf{u}_2 = \mathbf{a}_2 - 2 \mathbf{a}_1 = \mathbf{a}_2 - 2 \mathbf{u}_1$$

La terza riga  $\mathbf{u}_3$  è ottenuta in due passi, ossia

$$\mathbf{u}_3 = \mathbf{a}_3 - (-1 \mathbf{u}_1) - 3 \mathbf{u}_2$$

Risolviendo le equazioni precedenti rispetto ai vettori  $\mathbf{a}_i$ , si ottiene

$$\begin{aligned} \mathbf{a}_1 &= \mathbf{u}_1 \\ \mathbf{a}_2 &= 2 \mathbf{u}_1 + \mathbf{u}_2 \\ \mathbf{a}_3 &= -1 \mathbf{u}_1 + 3 \mathbf{u}_2 + \mathbf{u}_3 \end{aligned}$$

che mostra l'identità (2.8). L'esempio, quindi, mostra che il metodo di eliminazione è equivalente a fattorizzare una matrice nel prodotto tra una matrice triangolare inferiore con elementi diagonali uguali a 1 e una matrice triangolare superiore. Si può mostrare che tale decomposizione è *unica*.

Il metodo di eliminazione esposto in precedenza è *per righe*. Come mostreremo sempre sull'esempio (2.6), è possibile implementare il metodo procedendo *per colonne*, anziché per righe. Tali implementazioni hanno interesse per i linguaggi di programmazione che, come il FORTRAN, memorizzano le matrici per colonne.

Sottraendo dalla seconda colonna la prima moltiplicata per 2 e dalla terza la prima moltiplicata per 1, e successivamente dalla terza colonna la seconda colonna moltiplicata per -3, si ottengono le seguenti trasformazioni

$$\begin{bmatrix} 2 & 0 & 0 \\ 4 & -1 & 3 \\ -2 & -3 & 7 \end{bmatrix} \rightarrow \mathbf{L} = \begin{bmatrix} 2 & 0 & 0 \\ 4 & -1 & 0 \\ -2 & -3 & -2 \end{bmatrix}$$

Usando i moltiplicatori 2, 1 e -3, formiamo la seguente matrice triangolare superiore con elementi uguali a 1 sulla diagonale

$$\mathbf{U}_1 = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix}$$

L'indice 1 indica che la matrice ha elementi sulla diagonale principale uguali a 1. Analogamente a quanto si è fatto in precedenza, si può mostrare che l'eliminazione per colonne produce la seguente decomposizione

$$\mathbf{A} = \mathbf{L}\mathbf{U}_1$$

mentre, con la notazione ora introdotta l'eliminazione per righe produce la decomposizione  $\mathbf{A} = \mathbf{L}_1\mathbf{U}$ . Si può mostrare che introducendo la seguente matrice diagonale

$$\mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$$

si ha

$$\mathbf{A} = \mathbf{L}_1\mathbf{U} = \mathbf{L}_1(\mathbf{D}\mathbf{U}_1) = (\mathbf{L}_1\mathbf{D})\mathbf{U}_1 = \mathbf{L}\mathbf{U}_1 = \mathbf{A}$$

La fattorizzazione  $\mathbf{A} = \mathbf{L}_1(\mathbf{D}\mathbf{U}_1) = (\mathbf{L}_1\mathbf{D})\mathbf{U}_1$  è chiamata la fattorizzazione **LDU** della matrice  $\mathbf{A}$ .

Concludiamo, osservando che a partire dalle decomposizioni precedenti la risoluzione del sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  può essere ricondotta a quella di due sistemi più semplici. Ad esempio, se  $\mathbf{A} = \mathbf{L}\mathbf{U}$ , si ha

$$\mathbf{A}\mathbf{x} = \mathbf{b} \iff \mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{b} \iff \begin{cases} \mathbf{L}\mathbf{y} = \mathbf{b} \\ \mathbf{U}\mathbf{x} = \mathbf{y} \end{cases} \quad (2.9)$$

Pertanto, se le due matrici  $\mathbf{L}$ ,  $\mathbf{U}$  sono triangolari, la risoluzione del sistema è ricondotta alla risoluzione successiva di due sistemi con matrici triangolari. ■

### 2.1.1 Sistemi triangolari

Un sistema viene detto *triangolare*, quando la matrice dei coefficienti ha una struttura triangolare, o più in generale quando tale matrice può essere ricondotta a una forma triangolare mediante opportuni scambi di righe e di colonne. Per tali sistemi l'operazione di eliminazione si applica in maniera molto semplice. Consideriamo, in particolare, i sistemi con matrice dei coefficienti triangolare superiore, o inferiore. Siano, quindi,  $\mathbf{L}$  e  $\mathbf{U}$  due matrici di ordine  $n$  della forma seguente

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & u_{nn} \end{bmatrix}; \quad \mathbf{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$$

Generalizzando la procedura seguita nell'esempio precedente, abbiamo i seguenti algoritmi.

**Algoritmo 2.1** (Eliminazione in avanti) *Data una matrice triangolare inferiore  $\mathbf{L} \in \mathbb{R}^{n \times n}$  non singolare ed un vettore  $\mathbf{b} \in \mathbb{R}^n$ , la soluzione del sistema  $\mathbf{L}\mathbf{y} = \mathbf{b}$  è data da*

```

 $y_1 = b_1/l_{11}$ 
For  $i = 2, \dots, n$ 
   $y_i = b_i$ 
  For  $j = 1, \dots, i - 1$ 
     $y_i = y_i - l_{ij}y_j$ 
  end  $j$ 
   $y_i = y_i/l_{ii}$ 
end  $i$ 

```

L'algoritmo precedente richiede un numero di operazioni, valutato in flops<sup>1</sup>, dato da

$$\sum_{i=1}^n (i-1) = \frac{1}{2}n(n-1) \approx \frac{1}{2}n^2$$

Osserviamo che  $n^2/2$  è anche l'ordine del numero delle moltiplicazioni necessarie per valutare il prodotto della matrice  $\mathbf{L}$  per un vettore.

Analogamente, per i sistemi triangolari superiori si ha il seguente algoritmo.

**Algoritmo 2.2** (Eliminazione all'indietro) Data una matrice triangolare superiore  $\mathbf{U} \in \mathbb{R}^{n \times n}$ , non singolare, ed un vettore  $\mathbf{y} \in \mathbb{R}^n$ , la soluzione del sistema  $\mathbf{U}\mathbf{x} = \mathbf{y}$  è data da

```

 $x_n = y_n/u_{nn}$ 
For  $i = n - 1, \dots, 1$ 
   $x_i = y_i$ 
  For  $j = i + 1, \dots, n$ 
     $x_i = x_i - u_{ij}x_j$ 
  end  $j$ 
   $x_i = x_i/u_{ii}$ 
end  $i$ 

```

Lasciamo come esercizio la modifica degli algoritmi precedenti nel caso in cui, come è stato analizzato nell'esempio precedente, la eliminazione venga effettuata *per colonne*. e nel caso in cui le matrici  $\mathbf{L}$  e  $\mathbf{U}$  siano memorizzate sotto forma di vettori a una dimensione, ossia, ad esempio per la matrice  $\mathbf{U}$  nella forma  $[u_{11}, u_{12}, u_{22}, u_{13}, u_{23} \dots u_{1n}, \dots, u_{nn}]$ .

## 2.1.2 Sistemi generali; metodo di Gauss

Consideriamo il sistema lineare (2.1) nel caso in cui  $m = n$ , e nel quale, per motivi di convenienza di scrittura dell'algoritmo, il termine noto è definito come la colonna

<sup>1</sup>Il flop è una unità di misura delle operazioni richieste in un algoritmo. Secondo una definizione usuale, un flop corrisponde al numero di operazioni nell'istruzione FORTRAN:  $A(I,J)=A(I,J)+T*A(I,K)$ . Essa comprende una moltiplicazione e una addizione in aritmetica floating point, alcuni calcoli di indici e alcuni riferimenti alla memoria. Secondo un'altra definizione un flop corrisponde ad *una* sola operazione floating point. Nel testo viene utilizzata la prima definizione.



ove, per uniformità di scrittura, si è posto

$$a_{ij}^{(1)} := a_{ij}, \quad \begin{array}{l} i = 1, 2, \dots, n \\ j = 1, 2, \dots, n+1 \end{array}$$

Per  $k = n$ , cioè dopo  $n - 1$  trasformazioni, si ottiene un sistema triangolare, che può essere risolto con l'algoritmo esaminato nel paragrafo precedente.

Il determinante della matrice  $\mathbf{A}$  è uguale al determinante della matrice triangolare finale, in quanto le successive trasformazioni della matrice (sottrazioni di righe da altre moltiplicate per opportuni fattori) lasciano invariato il determinante. Si ha, pertanto

$$\det(\mathbf{A}) = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)}$$

Si ha, in definitiva, la seguente procedura.

**Algoritmo 2.3** (Algoritmo di Gauss) *Data una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  tale che  $\det(\mathbf{A}_k) \neq 0$ ,  $k = 1, 2, \dots, n$ , e quindi in particolare non singolare, ed un vettore  $a_{i,n+1}$ ,  $i = 1, 2, \dots, n$ , si ha*

```

For  $k = 1, 2, \dots, n - 1$ 
  For  $i = k + 1, \dots, n$ 
     $m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ 
    For  $j = k + 1, \dots, n + 1$ 
       $a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}$ 
    end  $j$ 
  end  $i$ 
end  $k$ 

```

A solo scopo di esemplificazione, l'algoritmo può essere implementato in FORTRAN nel seguente modo. Lasciamo come esercizio, in particolare, l'introduzione delle opportune modifiche per il controllo in *aritmetica di calcolatore* dell'annullarsi degli elementi  $a_{kk}^{(k)}$ , e quindi della singolarità delle sottomatrici principali.

```

DIMENSION A(20,21),X(20)
PRINT*, 'N= '
READ*,N
PRINT*, 'MATRICE'
C   lettura della matrice aumentata
DO 2 I=1,N
  PRINT*,I,' A(I,J),J=1,...,N+1 '
2  READ*,(A(I,J), J=1,N+1)
CALL GAUSS(N,A,X)
PRINT*,(X(I), I=1,N)
DET=1
DO 4 K=1,N

```



```

4      DET=DET*A(K,K)
      PRINT*, 'DETERMINANTE= ', DET
      STOP
      END

      SUBROUTINE GAUSS(N,A,X)
C      eliminazione
      DIMENSION A(20,21),X(20)
      NP1=N+1
      NM1=N-1
      DO 10 K=1,NM1
          KP=K+1
          IF (ABS(A(K,K)).EQ.0) THEN
              PRINT*, 'MINORE DI ORDINE ',K,' NULLO'
              STOP
          ENDIF
          DO 10 I=KP,N
              QT=A(I,K)/A(K,K)
              DO 10 J=KP,NP1
10         A(I,J)=A(I,J)-QT*A(K,J)
C
          IF (ABS(A(N,N)).EQ.0) THEN
              PRINT*, 'MATRICE SINGOLARE'
              STOP
          ENDIF
          X(N)=A(N,NP1)/A(N,N)
          DO 20 NN=1,NM1
              SUM=0.
              I=N-NN
              IP=I+1
              DO 30 J=IP,N
30         SUM=SUM+A(I,J)*X(J)
              X(I)=(A(I,NP1)-SUM)/A(I,I)
20      CONTINUE
          RETURN
      END

```

Nella forma precedente l'algoritmo è di tipo *riga orientato*, nel senso che il loop più interno si riferisce all'indice di colonna. Come abbiamo già discusso nell'Esempio 2.4, per alcuni linguaggi di programmazione, in particolare per il FORTRAN, può essere più conveniente una implementazione *colonna orientata*. Il seguente segmento di programma esemplifica tale modifica.

```

      DO 20 K=1,N
          KP1=K+1
          NP1=N+1
          T=A(K,K)
*      Calcolo dei moltiplicatori
          DO 10 I=KP1,N
10         A(I,K)=A(I,K)/T

```

```

* Eliminazione per colonne
  DO 20 J=KP1,NP1
    T=A(K,J)
    DO 20 I=KP1,N
20      A(I,J)=A(I,J)-A(I,K)*T

```

Nella implementazione dell'algoritmo di eliminazione, i *moltiplicatori*  $m_{ij}$  possono essere memorizzati nello spazio occupato dagli elementi della matrice che diventano successivamente nulli. In questo modo, a seguito delle operazioni di eliminazione, la matrice originaria  $\mathbf{A}$  è sostituita dalla struttura descritta in Figura 2.3, ove i moltiplicatori sono memorizzati sotto la diagonale principale nella matrice  $\mathbf{L}$ , e gli elementi  $a_{ij}^{(k)}$  formano la matrice triangolare  $\mathbf{U}$ .

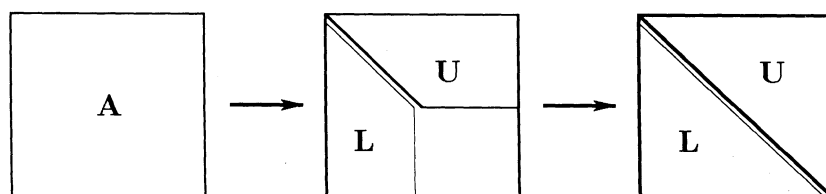


Figura 2.3: Struttura della matrice a seguito delle operazioni di eliminazione.

### Numero delle operazioni

Per valutare il numero delle operazioni richieste dal procedimento di eliminazione, osserviamo che il generico passo  $k$  comporta  $n - k$  divisioni e  $(n - k)(n - k + 1)$  moltiplicazioni e addizioni. In flops il costo totale dell'algoritmo è, quindi, dell'ordine di

$$\sum_{k=1}^{n-1} (n - k)(n - k + 1) = \frac{1}{3}n(n^2 - 1)$$

Per la risoluzione del sistema triangolare finale è richiesto, come abbiamo visto, un numero di flops dell'ordine di  $\frac{1}{2}n^2$ ; in definitiva, quindi, il costo in flops per la risoluzione di un sistema lineare è  $O(\frac{1}{3}n^3 + \frac{1}{2}n^2)$ . Per valori grandi di  $n$ , il termine  $n^3$  è predominante, e quindi si può dire che la riduzione di un sistema lineare con  $n$  incognite alla forma triangolare mediante il metodo di eliminazione di Gauss richiede approssimativamente  $n^3/3$  operazioni. Assumendo, ad esempio, che una operazione su un determinato calcolatore richieda  $3\mu s$ , la Tabella 2.1 fornisce i tempi, per differenti valori di  $n$ , per ottenere l'eliminazione di Gauss e la sostituzione all'indietro.

Si può dimostrare che per risolvere il sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{A}$  matrice *qualsivoglia*, il metodo di Gauss è ottimale, nel senso che è il metodo con l'ordine di

$n$	eliminazione	sostituzione all'indietro
100	1.	0.015
1000	$1 \cdot 10^3$	1.5
10000	$1 \cdot 10^6$ ( $\approx 12$ giorni)	150

Tabella 2.1: Tempo in secondi per il metodo di eliminazione e la sostituzione all'indietro nell'ipotesi che ogni operazione richieda  $3\mu s$ .

operazioni (esponente di  $n$ ) minimo fra tutti i metodi che utilizzano solo combinazioni lineari di righe e di colonne. Recentemente, a partire da un lavoro di Strassen<sup>2</sup>, sono stati proposti metodi basati su tecniche diverse, che permettono il calcolo numerico della soluzione di un sistema lineare con un costo computazionale di ordine inferiore. Tali metodi si basano, in sostanza, sull'equivalenza, per quanto riguarda il costo computazionale, del problema della inversione di una matrice di ordine  $n$ , e quindi della soluzione di un sistema lineare di ordine  $n$ , e del problema della moltiplicazione di due matrici di ordine  $n$ . Più precisamente, si può mostrare che se il numero delle operazioni aritmetiche sufficienti a calcolare il prodotto di due matrici di ordine  $n$  è al più  $kn^\theta$ , con  $k, \theta$  costanti positive e  $2 \leq \theta \leq 3$ , allora per invertire una matrice non singolare di ordine  $n$  sono sufficienti  $hn^\theta$  operazioni aritmetiche, con  $h$  opportuna costante positiva. Con l'algoritmo proposto da Strassen, per il calcolo del prodotto di due matrici si ha (cfr. l'Esempio successivo) un ordine di operazioni  $\theta = \log_2 7 = 2.807\dots$ ; successivamente, sono stati proposti algoritmi con valori di  $\theta$  minori (segnaliamo in particolare un risultato di Coppersmith, Winograd (1987) con  $\theta < 2.38$ ). Sebbene l'esponente  $\theta$  sia almeno 2, il più piccolo valore di  $\theta$  che può essere ottenuto è attualmente incognito. È da osservare che nei metodi ora segnalati la costante di proporzionalità che moltiplica  $n^\theta$  può essere molto grande, e quindi, anche se tali schemi sono asintoticamente più veloci del metodo di Gauss, la loro superiorità è osservabile solo per  $n$  molto grandi.

Per terminare le considerazioni sul costo computazionale del metodo di Gauss, è opportuno osservare che l'ordine è  $O(n^3)$  per una matrice *generica*. In altre parole, l'ordine può essere molto più basso per matrici *particolari*; si pensi, ovviamente, alle matrici diagonali, triangolari, tridiagonali, cioè alle matrici con strutture a *banda* e più in generale alle matrici *sparse*. Riprenderemo questo argomento nei paragrafi successivi.

► **Esempio 2.5** *Algoritmo di Strassen*. Date le matrici  $\mathbf{A}$ ,  $\mathbf{B}$  di ordine  $n = 2^k$ , con  $k$  intero positivo, gli  $n^2$  elementi  $c_{ij}$  della matrice prodotto  $\mathbf{C} = \mathbf{AB}$  possono essere ottenuti dalla formula

$$c_{ij} = \sum_{p=1}^n a_{ip}b_{pj}, \quad i, j = 1, \dots, n \quad (2.12)$$

<sup>2</sup>V. Strassen, *Gaussian Elimination is not Optimal*, Numer. Math. 13, 1969, 354–356.

mediante  $n^3$  moltiplicazioni e  $n^3 - n^2$  addizioni. Mostriamo ora che è possibile ottenere gli  $n^2$  elementi  $c_{ij}$  con non più di  $4.7n^\theta$  operazioni aritmetiche, ove  $\theta = \log_2 7$ .

Per  $n = 2$ , gli elementi  $c_{ij}$  dati da

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21}, & c_{12} &= a_{11}b_{12} + a_{12}b_{22}, \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21}, & c_{22} &= a_{21}b_{12} + a_{22}b_{22} \end{aligned}$$

possono essere calcolati con 7 moltiplicazioni e 18 addizioni mediante le seguenti relazioni

$$\begin{aligned} s_1 &= (a_{11} + a_{22})(b_{11} + b_{22}) & s_2 &= (a_{21} + a_{22})b_{11} \\ s_3 &= a_{11}(b_{12} - b_{22}) & s_4 &= a_{22}(b_{21} - b_{11}) \\ s_5 &= (a_{11} + a_{12})b_{22} & s_6 &= (a_{21} - a_{11})(b_{11} + b_{12}) \\ s_7 &= (a_{12} - a_{22})(b_{21} + b_{22}) & & \\ c_{11} &= s_1 + s_4 - s_5 + s_7 & c_{12} &= s_3 + s_5 \\ c_{21} &= s_2 + s_4 & c_{22} &= s_1 - s_2 + s_3 + s_6 \end{aligned} \quad (2.13)$$

Le formule (2.13) possono essere applicate anche nel caso in cui gli elementi  $a_{ij}$ ,  $b_{ij}$ ,  $c_{ij}$  sono blocchi di matrici  $\mathbf{A}_{ij}$ ,  $\mathbf{B}_{ij}$ ,  $\mathbf{C}_{ij}$ , dal momento che non viene utilizzata la proprietà commutativa della moltiplicazione.

Supponendo, allora,  $k > 1$  e partizionando le matrici  $\mathbf{A}$ ,  $\mathbf{B}$  e  $\mathbf{C}$  in quattro sottomatrici di ordine  $n/2$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$$

le matrici  $\mathbf{C}_{ij}$ ,  $i, j = 1, 2$ , possono essere calcolate mediante le formule (2.13) con 7 moltiplicazioni di matrici di ordine  $n/2$ . Il procedimento può essere iterato partizionando ogni matrice di ordine  $n/2$  in quattro sottomatrici di ordine  $n/4$ . Se indichiamo con  $p_n$  il numero di moltiplicazioni impiegate dal metodo ora descritto per moltiplicare matrici di ordine  $n$ , vale la relazione

$$p_n = 7p_{n/2}$$

da cui, essendo  $n = 2^k$  e  $p_1 = 1$  si ha

$$p_n = 7p_{n/2} = 7^2p_{n/4} = \dots = 7^k p_1 = n^\theta, \quad \theta = \log_2 7$$

### 2.1.3 Strategia del pivot

Il metodo di eliminazione, nella forma che abbiamo presentato nel paragrafo precedente, è applicabile solo quando la matrice dei coefficienti  $\mathbf{A}$  ha i minori principali diversi dallo zero. In particolare, quindi, tale metodo è dal punto di vista *teorico* applicabile ai sistemi con matrici dei coefficienti a *predominanza diagonale*, oppure *simmetriche definite positive*. L'idea che si utilizza per estendere il metodo a matrici generiche, cioè solamente non singolari, consiste nel permutare, durante l'operazione di eliminazione, opportunamente le righe e/o le colonne della matrice. Introduciamo tale idea attraverso un semplice esempio, che illustra anche l'interesse di tale procedura per migliorare la *stabilità* dell'algoritmo di eliminazione, cioè il suo comportamento rispetto alla propagazione degli errori di arrotondamento.

► **Esempio 2.6** Consideriamo il seguente sistema

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases} \quad (2.14)$$

La matrice dei coefficienti è non singolare ed ha come soluzione *esatta* il vettore  $\mathbf{x} = [1, -1, 1]^T$ . Dopo la prima eliminazione si ottiene per la seconda e terza equazione

$$\begin{aligned} x_3 &= 1 \\ x_2 + x_3 &= 0 \end{aligned}$$

e, quindi, essendo  $a_{22}^{(2)} = 0$ , nella forma precedente il metodo si arresta. In questo caso è evidente il *rimedio*; basta scambiare la 2<sup>a</sup> equazione con la 3<sup>a</sup>, oppure scambiare la 2<sup>a</sup> colonna con la 3<sup>a</sup>. La seconda operazione, in particolare, equivale a un cambiamento nell'ordine delle incognite.

Esaminiamo, ora, un altro effetto dell'operazione di scambio di righe e/o di colonne. A tale scopo sostituiamo al sistema (2.14) il seguente

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + 1.0001x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases} \quad (2.15)$$

ottenuto *perturbando* l'elemento  $a_{22}$

$$a_{22} \rightarrow \tilde{a}_{22} := a_{22} + 0.0001$$

La soluzione  $\tilde{\mathbf{x}}$  del sistema (2.15), arrotondata a 4 decimali, è la seguente

$$\tilde{x}_1 = 1.0000, \quad \tilde{x}_2 = -1.0001, \quad \tilde{x}_3 = 1.0001$$

Come si vede, gli errori relativi sui risultati sono dello stesso ordine di grandezza dell'errore relativo introdotto nella matrice dei coefficienti. Questo risultato significa che il sistema dato non è mal condizionato. Applicato al sistema (2.15), il metodo di eliminazione fornisce al primo passaggio il seguente sistema ridotto

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ 0.0001x_2 + x_3 = 1 \\ x_2 + x_3 = 0 \end{cases} \quad (2.16)$$

da cui, eliminando la variabile  $x_2$  dalla seconda equazione, si ottiene il sistema triangolare

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ 0.0001x_2 + x_3 = 1 \\ -9999x_3 = -10000 \end{cases} \quad (2.17)$$

mentre, se scambiamo prima dell'eliminazione della variabile  $x_2$  la seconda riga con la terza si ottiene

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_2 + x_3 = 0 \\ 0.9999x_3 = 1 \end{cases} \quad (2.18)$$

Osserviamo che nella prima procedura i moltiplicatori dipendono dal fattore  $1/0.0001 = 10000$ , mentre dopo lo scambio delle righe tale fattore diventa  $0.0001$ .

Esaminiamo, ora, l'effetto degli errori di arrotondamento nella risoluzione del sistema triangolare (2.17) e rispettivamente del sistema (2.18). Supponiamo, a scopo di esemplificazione, di operare in una *aritmetica floating a 4 cifre*. Operando una sostituzione all'indietro, nel caso del sistema (2.17) si ottiene  $x_3 = 10000/9999 = 1.00010001\dots$ , da cui il valore arrotondato a quattro cifre  $\bar{x}_3 = 1.000$ . Sostituendo nelle equazioni precedenti, si ottiene allora la seguente soluzione numerica

$$\text{sistema (2.17)} \quad \bar{x}_1 = 0, \quad \bar{x}_2 = 0, \quad \bar{x}_3 = 1.000$$

Nel caso del sistema (2.18) si ottiene ancora  $\bar{x}_3 = 1.000$ , ma questa volta la sostituzione nelle equazioni precedenti fornisce la seguente soluzione

$$\text{sistema (2.18)} \quad \bar{x}_1 = 1.000, \quad \bar{x}_2 = -1.000, \quad \bar{x}_3 = 1.000$$

Come si vede dal confronto con la soluzione esatta del problema perturbato, l'operazione di scambio delle due righe ha migliorato la *stabilità numerica* dell'algoritmo. ■

L'esempio precedente *suggerisce* l'opportunità, dal punto di vista della *propagazione degli errori di arrotondamento*, di evitare durante l'eliminazione la presenza di *moltiplicatori grandi*. Questo risultato equivale, come si è visto nell'esempio, ad una opportuna scelta, ad ogni passo  $k$ , dell'elemento  $a_{kk}^{(k)}$ . Poiché tale elemento individua la incognita da eliminare e la riga in cui operare l'eliminazione, esso viene detto usualmente *elemento pivot*, da cui il nome di tecnica, o metodo pivotale (pivoting).

A secondo di come l'elemento pivot viene individuato, si possono avere differenti varianti, tra le quali segnaliamo le seguenti, illustrate in maniera schematica nella Figura 2.4, ove le frecce indicano gli opportuni scambi di righe e/o di colonne.

- *pivoting parziale*: si sceglie come elemento pivot il massimo in modulo degli elementi di una *colonna* (pivoting di colonna)

$$|a_{rk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

e si scambiano la riga  $r$ -ma con la riga  $k$ -ma prima di procedere alla eliminazione. In modo analogo, si definisce il pivoting per riga.

- *pivoting totale*: l'elemento pivot è il massimo in modulo tra tutti gli elementi della matrice che rimane da trasformare

$$|a_{rs}^{(k)}| = \max_{k \leq i, j \leq n} |a_{ij}^{(k)}|$$

Per portare l'elemento  $a_{rs}^{(k)}$  nella posizione  $(k, k)$  del pivot, è necessario uno scambio fra le righe di indice  $r$  e  $k$  e uno scambio fra le colonne di indice  $s$  e  $k$ . Lo scambio di righe non modifica la soluzione del sistema lineare, mentre lo scambio di colonne modifica l'ordinamento delle componenti del vettore soluzione.

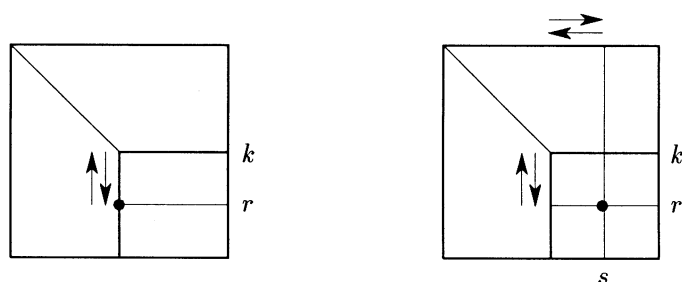


Figura 2.4: Illustrazione della tecnica di pivoting di colonna e di pivoting totale.

È, tuttavia, importante osservare che la tecnica del pivot risulta efficace nel migliorare la stabilità numerica del procedimento di eliminazione soltanto se la matrice dei coefficienti è opportunamente *equilibrata*. Questo aspetto è illustrato nel successivo esempio.

► **Esempio 2.7** Consideriamo il seguente sistema

$$\begin{cases} 0.005x_1 + x_2 = 0.5 \\ x_1 + x_2 = 1 \end{cases} \quad (2.19)$$

la cui soluzione *esatta* è data da

$$x_1 = 5000/9950 = 0.503\dots; \quad x_2 = 4950/9950 = 0.497\dots \quad (2.20)$$

Eliminando la variabile  $x_1$  dalla prima equazione, ossia assumendo come elemento pivot  $a_{11} = 0.005$ , si ottiene

$$\begin{cases} 0.005x_1 + x_2 = 0.5 \\ -199x_2 = -99 \end{cases} \quad (2.21)$$

In questo caso si ha come moltiplicatore il numero  $m_{21} = a_{21}/a_{11} = 1/0.005 \gg 1$  e, come abbiamo visto nell'esempio precedente, possono verificarsi difficoltà numeriche nella risoluzione del sistema triangolare. In effetti, supponendo di usare un'*aritmetica a 2 cifre* e indicando con  $\bar{x} = [\bar{x}_1, \bar{x}_2]$  la soluzione numerica ottenuta risolvendo all'indietro il sistema (2.21), si ottiene

$$\bar{x}_2 = \text{fl}(99/199) = \text{fl}(0.49548\dots) = 0.50, \quad \bar{x}_1 = 0$$

Se, in alternativa, assumiamo come elemento pivot l'elemento  $a_{21} = 1$ , si ha come moltiplicatore il valore  $m_{21} = 0.005 \ll 1$  e il seguente sistema triangolare, nel quale abbiamo scambiato le due righe

$$\begin{cases} x_1 + x_2 = 1 \\ 0.995x_2 = 0.495 \end{cases} \quad (2.22)$$

Si ottiene pertanto come soluzione

$$\bar{x}_2 = \text{fl}(0.495/0.995) = \text{fl}(0.4974\dots) = 0.50, \quad \bar{x}_1 = 0.50$$

che corrisponde all'arrotondamento della soluzione esatta (2.20).

I risultati ora ottenuti concordano con quelli che abbiamo ottenuto nell'esempio precedente, nel senso che la scelta opportuna dell'equazione da cui eliminare l'incognita corrisponde a quella che fornisce il moltiplicatore più piccolo. L'esempio che stiamo considerando ci permette, tuttavia, di evidenziare un altro aspetto importante. A tale scopo, consideriamo il sistema lineare che si ottiene dal sistema (2.19) moltiplicando la prima riga per 200

$$\begin{cases} x_1 + 200x_2 = 100 \\ x_1 + x_2 = 1 \end{cases} \quad (2.23)$$

I sistemi (2.19) e (2.23) sono naturalmente equivalenti, dal punto di vista teorico. Tuttavia, per quanto riguarda la tecnica pivotale di colonna, nel sistema (2.23) non si vede la necessità di scambiare le due righe. Ma, come si verifica facilmente, la scelta della prima equazione, come equazione da cui ricavare  $x_1$ , porta alle stesse difficoltà numeriche esaminate in precedenza. Indicando con  $\mathbf{A}_1$ , e rispettivamente  $\mathbf{A}_2$ , la matrice dei coefficienti del sistema (2.19) e del sistema (2.23), si ha

$$\begin{aligned} \mathbf{A}_1 &= \begin{bmatrix} 0.005 & 1 \\ 1 & 1 \end{bmatrix}, & \mathbf{A}_1^{-1} &\approx \begin{bmatrix} -1.005 & 1.005 \\ 1.005 & -0.005 \end{bmatrix} \\ \mathbf{A}_2 &= \begin{bmatrix} 1 & 200 \\ 1 & 1 \end{bmatrix}, & \mathbf{A}_2^{-1} &\approx \begin{bmatrix} -0.005 & 1.005 \\ 0.005 & -0.005 \end{bmatrix} \end{aligned}$$

La matrice  $\mathbf{A}_1$  è “più equilibrata” della matrice  $\mathbf{A}_2$ , nel senso che le somme dei moduli degli elementi delle righe, e delle colonne hanno un ordine di grandezza tra loro confrontabile, mentre per la matrice  $\mathbf{A}_2$  le somme delle righe differiscono per due ordini di grandezza.

Dalla conoscenza delle matrici inverse ricaviamo un'altra informazione, che sarà approfondita nel successivo paragrafo relativo al condizionamento di un sistema lineare. Si ha

$$\|\mathbf{A}_1\|_1 \|\mathbf{A}_1^{-1}\|_1 \approx 2, \quad \|\mathbf{A}_2\|_1 \|\mathbf{A}_2^{-1}\|_1 \approx 200 \quad (2.24)$$

Come vedremo, il risultato precedente esprime il fatto che la moltiplicazione della prima riga per 200 ha peggiorato il condizionamento della matrice. ■

### 2.1.4 Pivoting scalato

L'Esempio 2.7 ha evidenziato l'importanza, quando si applica il metodo del pivot, dello scaling della matrice dei coefficienti. Nelle applicazioni, suggerimenti per un scaling opportuno possono venire dall'analisi del significato delle singole variabili  $x_i$ , e quindi da una scelta conveniente delle loro unità di misura. In termini algoritmici, è possibile modificare opportunamente il metodo in modo adattivo, ossia in modo da tener conto della “grandezza” delle singole righe del sistema durante l'applicazione stessa del metodo. Più precisamente, si calcola inizialmente la *grandezza*  $d_i$  della riga  $i$  di  $\mathbf{A}$ , per  $i = 1, 2, \dots, n$ , ponendo

$$d_i = \max_{1 \leq j \leq n} |a_{ij}|$$

Al generico passo  $k$ -mo dell'eliminazione, si assume come equazione pivotale quella fra le  $n - k$  rimanenti che ha il coefficiente di  $x_k$  più grande in modulo relativamente



alla grandezza della riga. In altre parole si cerca l'indice  $p$  (usualmente il più piccolo), tra  $k$  e  $n$ , per il quale

$$\frac{|a_{pk}^{(k)}|}{d_p} \geq \frac{|a_{ik}^{(k)}|}{d_i}, \quad \forall i = k, \dots, n$$

Si ha in corrispondenza il seguente algoritmo.

**Algoritmo 2.4** (Algoritmo con pivoting colonna scalato) *Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  non singolare è ridotta ad una matrice di forma triangolare procedendo nel seguente modo*

```

For  $i = 1, 2, \dots, n$ 
   $ipivot(i) = i$ 
   $d_i = \max_{1 \leq j \leq n} |a_{ij}|$ 
end  $i$ 
For  $k = 1, 2, \dots, n - 1$ 
   $p$  intero tale che  $\frac{|a_{pk}^{(k)}|}{d_p} = \max_{k \leq i \leq n} \frac{|a_{ik}^{(k)}|}{d_i}$ 
  Scambio  $a_{kj}^{(k)}$  e  $a_{pj}^{(k)}$ ,  $j = 1, \dots, n$ 
  Scambio  $ipivot(k)$  e  $ipivot(p)$ 
  For  $i = k + 1, \dots, n$ 
     $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$ 
    For  $j = k + 1, \dots, n$ 
       $a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}$ 
    end  $j$ 
  end  $i$ 
end  $k$ 

```

Di seguito, nella routine **FACTOR**, viene presentato un esempio di implementazione dell'algoritmo precedente. La fattorizzazione ottenuta viene, quindi, utilizzata nella routine **SUBST** per la risoluzione di un sistema lineare.

```

c      esempio di utilizzo della fattorizzazione della matrice
PARAMETER (IA=20)
DIMENSION A(IA,IA),X(IA),D(IA),B(IA)
INTEGER IPIVOT(IA)
PRINT*, 'N= '
READ*, N
PRINT*, 'MATRICE'
DO 2 I=1,N
  PRINT*, I, ' A(I,J), J=1, . . . N '
2    READ*, (A(I,J), J=1,N)
CALL FACTOR(N,A,IA,D,IPIVOT,IFLAG)
PRINT*, 'IFLAG= ',IFLAG
PRINT*, (IPIVOT(I), I=1,N)
IF (IFLAG.EQ.0) THEN
  PRINT*, 'SISTEMA SINGOLARE'
  STOP
ELSE

```

```

PRINT*, 'TERMINE NOTO'
READ*, (B(I), I=1, N)
CALL SUBST(A, N, IA, IPIVOT, B, X)
PRINT*, (X(I), I=1, N)
ENDIF
STOP
END

SUBROUTINE FACTOR(N, A, IA, D, IPIVOT, IFLAG)
INTEGER IFLAG, IPIVOT(N), I, IS, J, K
REAL D(N), A(IA, N), AW, COLMAX, ROWMAX, TEMP
c   Algoritmo di eliminazione con pivoting scalato
c   Input
c   A array di dimensione (IA, N)
c   N ordine della matrice
c   IA .GE. N dimensione della array nel main
c   D vettore di lunghezza N, contiene le norme delle righe di A
c   Output
c   A contiene la fattorizzazione LU di P A
c   P permutazione specificata da IPIVOT
c   IPIVOT indica la riga IPIVOT(K) usata per eliminare X(K)
c   IFLAG = 1 numero pari di scambi
c           = -1 numero dispari di scambi
c           = 0 se U ha elementi diagonali nulli
c   Determinante=IFLAG*A(1,1)*...*A(N,N)
c
c   Se IFLAG.NE.0 il sistema AX=B e' risolto mediante
c   CALL SUBST(A, N, IA, IPIVOT, B, X)
c   IFLAG=1
c   Inizializzazione IPIVOT
DO 1 I=1, N
  IPIVOT(I)=I
  ROWMAX=0.
  DO 5 J=1, N
5    ROWMAX=AMAX1(ROWMAX, ABS(A(I, J)))
    IF (ROWMAX. EQ. 0) THEN
      IFLAG=0
      ROWMAX=1.
    END IF
1    D(I)=ROWMAX
  IF (N. LE. 1) RETURN
c   Fattorizzazione
DO 10 K=1, N-1
  COLMAX =ABS(A(K, K))/D(K)
  IS=K
  DO 12 I=K+1, N
    AW=ABS(A(I, K))/D(I)
    IF (AW. GT. COLMAX) THEN
      COLMAX=AW
      IS=I
    END IF

```

```

12  CONTINUE
    IF (COLMAX .EQ. 0.) THEN
        IFLAG=0
    ELSE
        IF (IS .GT. K) THEN
c   Scambio di righe
            IFLAG=-IFLAG
            I=IPIVOT(IS)
            IPIVOT(IS)=IPIVOT(K)
            IPIVOT(K)=I
            TEMP=D(IS)
            D(IS)=D(K)
            D(K)=TEMP
            DO 15 J=1,N
                TEMP=A(IS,J)
                A(IS,J)=A(K,J)
15             A(K,J)=TEMP
            END IF
            DO 20 I=K+1,N
                A(I,K)=A(I,K)/A(K,K)
                DO 20 J=K+1,N
                    A(I,J)=A(I,J)-A(I,K)*A(K,J)
20             CONTINUE
            END IF
10  CONTINUE
    IF(A(N,N) .EQ. 0.) IFLAG=0
    RETURN
    END

    SUBROUTINE SUBST(A,N,IA,IPIVOT,B,X)
    INTEGER IPIVOT(N),I,IP,J
    REAL B(N),A(IA,N),X(N),SUM
c   Input
c   A array di dimensione (NA,N) output di FACTOR
c   N ordine della matrice
c   NA .GE. N dimensione della matrice nel main
c   B vettore termini noti
c   Output
c   X soluzione di AX=B
c   Forward
    IF(N. LE. 1) THEN
        X(1)=B(1)/A(1,1)
        RETURN
    END IF
    IP=IPIVOT(1)
    X(1)=B(IP)
    DO 10 I=2,N
        SUM=0.
        DO 5 J=1,I-1
5           SUM =A(I,J)*X(J)+SUM
        IP=IPIVOT(I)

```

```

10      X(I)=B(IP)-SUM
c  Backward
      X(N)=X(N)/A(N,N)
      DO 30 I=N-1,1,-1
          SUM=0.
          DO 20 J=I+1,N
20              SUM=A(I,J)*X(J)+SUM
30      X(I)=(X(I)-SUM)/A(I,I)
      RETURN
      END

```

### 2.1.5 Decomposizione LU

È interessante sottolineare che il procedimento di eliminazione di Gauss realizza, in sostanza, la seguente fattorizzazione della matrice  $\mathbf{A}$

$$\boxed{\mathbf{PA} = \mathbf{LU}} \quad (2.25)$$

ove  $\mathbf{P}$  è la matrice di permutazione che corrisponde agli eventuali scambi di righe (cfr. Algoritmo 2.4),  $\mathbf{L}$  è la matrice triangolare inferiore che contiene i moltiplicatori, e con elementi sulla diagonale uguali a 1, e  $\mathbf{U}$  è la matrice triangolare superiore che risulta dall'applicazione del procedimento di eliminazione. In effetti, come vedremo nel seguito, l'idea della decomposizione della matrice dei coefficienti nel prodotto di matrici più semplici può essere generalizzata, dando origine a interessanti varianti del metodo di Gauss, note come *metodi compatti*. Osserviamo che dalla decomposizione (2.25) si ottiene la soluzione del sistema lineare  $\mathbf{Ax} = \mathbf{b}$  mediante la successiva risoluzione due sistemi triangolari

$$\boxed{\mathbf{Ax} = \mathbf{b}} \iff \begin{array}{l} \boxed{\mathbf{Ly} = \mathbf{Pb}} \\ \boxed{\mathbf{Ux} = \mathbf{y}} \end{array} \quad (2.26)$$

Una esemplificazione del risultato (2.25) è stata esaminata nell'Esempio 2.4, in corrispondenza al caso in cui  $\mathbf{P} = \mathbf{I}$ , ossia nel caso in cui non vengano effettuati scambi di righe. Riprenderemo, ora, tale caso per matrici di ordine qualunque, lasciando, invece, come esercizio l'estensione al caso in cui  $\mathbf{P}$  sia distinta dalla matrice identità. A tale scopo, osserviamo che il generico passo di eliminazione  $k$ -mo può essere descritto come il risultato del prodotto della matrice  $\mathbf{A}^{(k)}$ , ossia della matrice che risulta dalle sostituzioni precedenti, per una particolare *matrice elementare*. Si consideri, infatti, la matrice (cfr. Appendice A)

$$\mathbf{M}_k := \mathbf{H}(1, \mathbf{m}_k, \mathbf{e}_k) = \mathbf{I} - \mathbf{m}_k \mathbf{e}_k^T$$

ove  $\mathbf{m}_k = [0, \dots, m_{k+1k}, m_{k+2k}, \dots, m_{nk}]^T$ . La premoltiplicazione di  $\mathbf{A}^{(k)}$  per  $\mathbf{M}_k$  ha come effetto la sottrazione da ogni riga  $i$  ( $i = k+1, \dots, n$ ) della riga  $k$ -ma moltiplicata per  $m_{ik}$ . Per avere quindi l'eliminazione richiesta basta assumere  $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$ .

Si ha, pertanto

$$\mathbf{U} = \mathbf{M}_{n-1}\mathbf{M}_{n-2}\cdots\mathbf{M}_2\mathbf{M}_1\mathbf{A}$$

da cui si ricava

$$\mathbf{A} = \mathbf{LU} \quad (2.27)$$

tenendo conto che il prodotto di matrici triangolari e l'inversa di una matrice triangolare sono ancora matrici triangolari dello stesso tipo.

### 2.1.6 Decomposizione $\mathbf{LDM}^T$

La descrizione in termini matriciali del metodo di eliminazione suggerisce opportune modifiche che hanno interesse quando la matrice dei coefficienti presenta una struttura particolare, ad esempio quando è una matrice *simmetrica* e/o a *banda*. A tale scopo, introduciamo dapprima una variante della decomposizione  $\mathbf{LU}$ .

**Proposizione 2.1** (Decomposizione  $\mathbf{LDM}^T$ ) *Se i minori principali di una matrice  $\mathbf{A}$  di ordine  $n$  sono diversi dallo zero, allora esistono due matrici triangolari inferiori  $\mathbf{L}$  e  $\mathbf{M}$  con elementi diagonali uguali a 1 e una matrice diagonale  $\mathbf{D} = \mathbf{diag}(d_1, \dots, d_n)$  tali che  $\mathbf{A} = \mathbf{LDM}^T$ .*

**DIMOSTRAZIONE.** Dall'applicazione del metodo di eliminazione si ha  $\mathbf{A} = \mathbf{LU}$ , con  $\mathbf{L}$  triangolare inferiore e  $l_{ii} = 1$  e  $\mathbf{U}$  matrice triangolare superiore. Posto  $\mathbf{D} = \mathbf{diag}(d_1, \dots, d_n)$ , con  $d_j = u_{jj}$ ,  $j = 1, \dots, n$ , si ha che  $\mathbf{M}^T = \mathbf{D}^{-1}\mathbf{U}$  è una matrice triangolare superiore con elementi uguali ad uno sulla diagonale principale. Pertanto,  $\mathbf{A} = \mathbf{LU} = \mathbf{LD}(\mathbf{D}^{-1}\mathbf{U}) = \mathbf{LDM}^T$ . ■

Una volta che sia nota la decomposizione  $\mathbf{LDM}^T$ , la soluzione del sistema  $\mathbf{Ax} = \mathbf{b}$  può essere ottenuta con un numero  $O(n^2)$  di flops risolvendo i sistemi  $\mathbf{Ly} = \mathbf{b}$  (eliminazione in avanti),  $\mathbf{Dz} = \mathbf{y}$  e  $\mathbf{M}^T\mathbf{x} = \mathbf{z}$  (sostituzione all'indietro).

La decomposizione  $\mathbf{LDM}^T$  può essere ottenuta numericamente, come indicato nella dimostrazione della Proposizione 2.1, mediante il calcolo delle matrici  $\mathbf{L}$  e  $\mathbf{U}$  attraverso l'eliminazione di Gauss.

Un modo alternativo consiste nel considerare la relazione  $\mathbf{LDM}^T = \mathbf{A}$  come un sistema in cui le incognite sono gli elementi delle tre matrici  $\mathbf{L}$ ,  $\mathbf{D}$ ,  $\mathbf{M}$  e le equazioni si ottengono uguagliando gli elementi della matrice a primo e a secondo membro.

Si ottengono in questo modo le seguenti equazioni

$$\begin{aligned} \sum_{p=1}^{k-1} l_{kp}d_p m_{kp} + d_k &= a_{kk}, \quad k = 1, \dots, n \\ \sum_{p=1}^{k-1} l_{ip}d_p m_{kp} + l_{ik}d_k &= a_{ik}, \quad i > k \\ \sum_{p=1}^{k-1} l_{kp}d_p m_{ip} + m_{ik}d_k &= a_{ki}, \quad i > k \end{aligned}$$

che possono essere risolte mediante il seguente algoritmo<sup>3</sup>.

**Algoritmo 2.5** (Decomposizione  $\mathbf{LDM}^T$ ) Per ogni matrice  $\mathbf{A}$  di ordine  $n$ , con minori principali non nulli, la decomposizione  $\mathbf{LDM}^T$  può essere ottenuta mediante il seguente algoritmo nel quale gli elementi  $a_{ij}$  della matrice originaria sono sostituiti da  $l_{ij}$  se  $i > j$  e con  $m_{ji}$  se  $i < j$ .

```

For  $k = 1, \dots, n$ 
  For  $p = 1, 2, \dots, k - 1$ 
     $r_p := d_p a_{pk}$ 
     $w_p := a_{kp} d_p$ 
  end  $p$ 
   $d_k := a_{kk} - \sum_{p=1}^{k-1} a_{kp} r_p$ 
  For  $i = k + 1, \dots, n$ 
     $a_{ik} = (a_{ik} - \sum_{p=1}^{k-1} a_{ip} r_p) / d_k$ 
     $a_{ki} = (a_{ki} - \sum_{p=1}^{k-1} w_p a_{pi}) / d_k$ 
  end  $i$ 
end  $k$ 

```

### 2.1.7 Metodi di Crout e di Doolittle

Ricordiamo due varianti dell'Algoritmo 2.5, note in letteratura, rispettivamente, come *metodo di Doolittle* e *metodo di Crout*<sup>4</sup> e ambedue esempi di *schemi compatti*. Sia il metodo di Doolittle che il metodo di Crout costruiscono una decomposizione di  $\mathbf{A}$  della forma  $\mathbf{A} = \mathbf{LU}$ ; differiscono per il fatto che nel metodo di Doolittle si richiede  $l_{ii} = 1, i = 1, \dots, n$ , mentre nel metodo di Crout si pone  $u_{ii} = 1, i = 1, \dots, n$ . In particolare, quindi, il metodo di Doolittle fornisce la medesima decomposizione del metodo di Gauss. In sostanza, la fattorizzazione di Crout corrisponde alla associazione  $(\mathbf{LD})\mathbf{U}$ , mentre la fattorizzazione di Doolittle corrisponde a  $\mathbf{L}(\mathbf{DU})$ .

**Algoritmo 2.6** (Doolittle) Data una matrice  $\mathbf{A}$  di ordine  $n$ , con i minori principali differenti dallo zero, l'algoritmo fornisce la decomposizione  $\mathbf{A} = \mathbf{LU}$ , con  $l_{ii} = 1$ .

```

For  $k = 1, 2, \dots, n$ 
  For  $i = k, \dots, n$ 
     $u_{ki} = a_{ki} - \sum_{p=1}^{k-1} l_{kp} u_{pi}$ 
  end  $i$ 
  For  $i = k + 1, \dots, n$ 
     $l_{ik} = (a_{ik} - \sum_{p=1}^{k-1} l_{ip} u_{pk}) / u_{kk}$ 
  end  $i$ 
end  $k$ 

```

<sup>3</sup>Nella descrizione dell'algoritmo si conviene, come avviene automaticamente in alcuni linguaggi, che quando nella sommatoria e nel loop di iterazione il secondo indice è minore del primo la sommatoria e il loop non sono eseguiti.

<sup>4</sup>Crout, Prescott D., *A Short Method for Evaluating Determinants and Solving Systems of Linear Equations with Real or Complex Coefficients*, Trans. AIEE, **60**:1235–1240 (1941).

Considerando, come illustrazione, il caso  $n = 3$  si ha

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21} & a_{22} & a_{23} \\ l_{31} & a_{32} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21} & u_{22} & u_{23} \\ l_{31} & l_{32} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21} & u_{22} & u_{23} \\ l_{31} & l_{32} & u_{33} \end{bmatrix}$$

ove, ad esempio  $u_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}$ .

**Algoritmo 2.7** (Crout) *Data una matrice  $\mathbf{A}$  di ordine  $n$ , con i minori principali differenti dallo zero, l'algoritmo fornisce la decomposizione  $\mathbf{A} = \mathbf{L}\mathbf{U}$ , con  $u_{ii} = 1$ .*

```

For  $k = 1, 2, \dots, n$ 
  For  $i = k, \dots, n$ 
     $l_{ik} = a_{ik} - \sum_{p=1}^{k-1} l_{ip}u_{pk}$ 
  end  $i$ 
  For  $i = k + 1, \dots, n$ 
     $u_{ki} = (a_{ki} - \sum_{p=1}^{k-1} l_{kp}u_{pi})/l_{kk}$ 
  end  $i$ 
end  $k$ 

```

Nel caso di  $n = 3$  si ha

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} l_{11} & u_{12} & u_{13} \\ l_{21} & a_{22} & a_{23} \\ l_{31} & a_{32} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21} & l_{22} & u_{23} \\ l_{31} & l_{32} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21} & u_{22} & u_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}$$

ove, ad esempio  $l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}$ .

### 2.1.8 Matrici simmetriche

Quando la matrice  $\mathbf{A}$  è simmetrica vi è una ridondanza nella fattorizzazione  $\mathbf{LDM}^T$ . Si ha, infatti, il seguente risultato.

**Proposizione 2.2** (Decomposizione  $\mathbf{LDL}^T$ ) *Se  $\mathbf{A} = \mathbf{LDM}^T$  è la decomposizione di una matrice  $\mathbf{A}$  simmetrica non singolare, allora  $\mathbf{L} = \mathbf{M}$ .*

**DIMOSTRAZIONE.** La matrice  $\mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-T} = \mathbf{M}^{-1}\mathbf{L}\mathbf{D}$  è simmetrica e triangolare inferiore, e quindi diagonale. Dal momento che  $\mathbf{D}$  è non singolare, questo implica che anche  $\mathbf{M}^{-1}\mathbf{L}$  è diagonale ed avendo gli elementi sulla diagonale principale tutti uguali a 1 si ha  $\mathbf{M}^{-1}\mathbf{L} = \mathbf{I}$ . ■

Ad esempio, si ha

$$\mathbf{A} = \begin{bmatrix} 10 & 20 & 30 \\ 20 & 45 & 80 \\ 30 & 80 & 171 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{bmatrix} \begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.28)$$

A partire da tale risultato, l'Algoritmo 2.5 può essere modificato nel seguente modo.

**Algoritmo 2.8** (Decomposizione  $\mathbf{LDL}^T$ ) *Data una matrice simmetrica  $\mathbf{A}$ , le matrici  $\mathbf{D}$ ,  $\mathbf{L}$  si calcolano nel modo seguente; gli elementi  $a_{ij}$  sono sostituiti da  $l_{ij}$  per  $i > j$ .*

```

For  $k = 1, \dots, n$ 
  For  $p = 1, \dots, k - 1$ 
     $r_p := d_p a_{kp}$ 
  end  $p$ 
   $d_k := a_{kk} - \sum_{p=1}^{k-1} a_{kp} r_p$ 
  If  $d_k = 0$ 
    then stop
  else
    For  $i = k + 1, \dots, n$ 
       $a_{ik} := (a_{ik} - \sum_{p=1}^{k-1} a_{ip} r_p) / d_k$ 
    end  $i$ 
  end if
end  $k$ 

```

Quando applicato alla matrice (2.28), l'algoritmo fornisce la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 10 & 20 & 30 \\ 2 & 5 & 80 \\ 3 & 4 & 1 \end{bmatrix}$$

L'Algoritmo 2.5 diventa particolarmente interessante, quando applicato ad una *matrice definita positiva*. Si può infatti dimostrare, utilizzando le proprietà delle matrici simmetriche definite positive (cfr. Appendice A) che per esse la fattorizzazione  $\mathbf{A} = \mathbf{LDM}^T$  esiste e la matrice diagonale  $\mathbf{D}$  ha elementi positivi. Si può, inoltre, mostrare che l'algoritmo è *stabile* nel senso della propagazione degli errori di arrotondamento.

Nel caso di matrici simmetriche definite positive, una variante della decomposizione  $\mathbf{LDL}^T$  è data dal noto *metodo di Cholesky*, o metodo delle radici quadrate<sup>5</sup>.

**Proposizione 2.3** (Decomposizione di Cholesky) *Se  $\mathbf{A}$  è una matrice simmetrica definita positiva, allora esiste un'unica matrice  $\mathbf{R}$  triangolare inferiore, con elementi positivi sulla diagonale principale, tale che*

$$\mathbf{A} = \mathbf{R}\mathbf{R}^T \quad (2.29)$$

Il risultato segue dal fatto che, come abbiamo visto in precedenza, si ha  $\mathbf{A} = \mathbf{LDL}^T$ . Poiché  $d_k > 0$ , la matrice  $\mathbf{R} = \mathbf{L} \mathbf{diag}(d_1^{1/2}, \dots, d_n^{1/2})$  è reale, triangolare inferiore con elementi positivi sulla diagonale principale e verifica (2.29). L'unicità segue dalla unicità della fattorizzazione  $\mathbf{LDL}^T$ .

<sup>5</sup>A. L. Cholesky (1875-1918), ufficiale francese, ha sviluppato tale metodo allo scopo di risolvere le equazioni normali ottenute applicando il metodo dei minimi quadrati nel fitting di dati geodetici. Il metodo è anche noto come *metodo di Banachiewicz* (1937).



Come esemplificazione, si consideri

$$\begin{bmatrix} 2 & -2 \\ -2 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 \\ -\sqrt{2} & \sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 0 & \sqrt{3} \end{bmatrix}$$

Osserviamo che una volta calcolata la decomposizione di Cholesky la soluzione del sistema  $\mathbf{Ax} = \mathbf{b}$  può essere ottenuta risolvendo successivamente i due sistemi triangolari  $\mathbf{Ry} = \mathbf{b}$  e  $\mathbf{R}^T \mathbf{x} = \mathbf{y}$ .

Un modo efficiente per calcolare la matrice di Cholesky  $\mathbf{R}$  può essere ricavato dal confronto degli elementi nell'equazione  $\mathbf{A} = \mathbf{RR}^T$ . Per  $i \geq k$  si ha

$$a_{ik} = \sum_{p=1}^k r_{ip} r_{kp} \Rightarrow \begin{cases} r_{ik} = \left( a_{ik} - \sum_{p=1}^{k-1} r_{ip} r_{kp} \right) / r_{kk}, & i > k \\ r_{kk} = \left( a_{kk} - \sum_{p=1}^{k-1} r_{kp}^2 \right)^{1/2} \end{cases}$$

Ne consegue il seguente algoritmo.

**Algoritmo 2.9** (Metodo di Cholesky; versione per colonne) *Data una matrice  $\mathbf{A}$  simmetrica definita positiva, la matrice triangolare inferiore  $\mathbf{R}$  tale che  $\mathbf{RR}^T = \mathbf{A}$  si calcola nel seguente modo; gli elementi  $a_{ij}$  sono sostituiti da  $r_{ij}$  ( $i \geq j$ ).*

```

For  $k = 1, 2, \dots, n$ 
   $a_{kk} := \left( a_{kk} - \sum_{p=1}^{k-1} a_{kp}^2 \right)^{1/2}$ 
  For  $i = k + 1, \dots, n$ 
     $a_{ik} := \left( a_{ik} - \sum_{p=1}^{k-1} a_{ip} a_{kp} \right) / a_{kk}$ 
  end  $i$ 
end  $k$ 

```

L'algoritmo richiede  $n^3/6$  flops. È interessante osservare che il calcolo degli elementi  $a_{ik}$ , per ogni  $k$  fissato e  $i = k + 1, \dots, n$ , può essere eseguito in *parallelo*.

La *stabilità* numerica dell'algoritmo segue dal seguente risultato

$$r_{ij}^2 \leq \sum_{p=1}^i r_{ip}^2 = a_{ii}$$

che mostra il fatto importante che gli elementi della matrice di Cholesky sono limitati (con costante moltiplicativa uguale a uno) dagli elementi della matrice  $\mathbf{A}$ .

### 2.1.9 Matrici a banda

Una matrice è chiamata *a banda* quando gli elementi che possono essere diversi dallo zero sono concentrati vicino alla diagonale. Più precisamente, si dice che  $\mathbf{A} = [a_{ij}]$  ha una banda superiore  $q$ , se  $a_{ij} = 0$  per  $j > i + q$  e una banda inferiore  $p$ , se  $a_{ij} = 0$  per  $i > j + p$  (cfr. per una rappresentazione schematica la Figura 2.5).

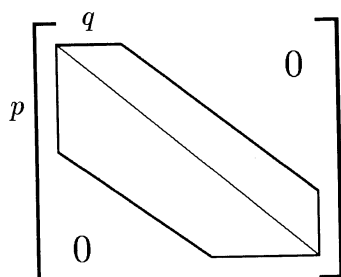


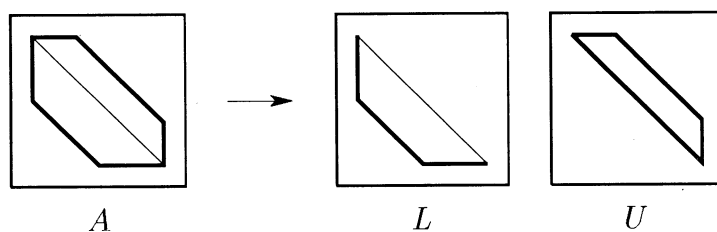
Figura 2.5: Struttura a banda.

Matrici con struttura a banda hanno origine in molteplici applicazioni, in particolare nell'ambito della discretizzazione dei problemi differenziali con il metodo delle differenze finite o degli elementi finiti, o in modelli di reti.

Per la risoluzione di sistemi con matrici a banda risultano particolarmente interessanti i metodi di fattorizzazione, in quanto i fattori triangolari  $\mathbf{L}$ ,  $\mathbf{U}$ , e  $\mathbf{R}$  nel caso simmetrico, possono essere calcolati in maniera da conservare la struttura a banda.

In effetti, con un procedimento di induzione si può dimostrare il seguente risultato, illustrato nella Figura 2.6.

**Proposizione 2.4** *Supponiamo che la matrice  $\mathbf{A}$  di ordine  $n$  abbia la fattorizzazione  $\mathbf{A} = \mathbf{L}\mathbf{U}$ . Allora, se  $\mathbf{A}$  ha una banda superiore  $q$  e una banda inferiore  $p$ , la matrice  $\mathbf{U}$  ha banda superiore  $q$  e  $\mathbf{L}$  una banda inferiore  $p$ .*

Figura 2.6: Decomposizione  $\mathbf{LU}$  per una matrice a banda.

Come illustrazione, si considerino i seguenti passaggi ottenuti mediante il metodo di eliminazione di Gauss in corrispondenza a una matrice a banda, con  $p = 2$  e  $q = 1$ . Osserviamo che ad ogni passo del procedimento di eliminazione, si opera su una sottomatrice  $3 \times 2$ ; in generale, per una matrice a banda  $(p, q)$ , ad ogni passo si

opera una sottomatrice di dimensioni  $(l+1) \times (u+1)$ .

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 4 & 4 & 1 & 0 & 0 \\ 6 & 5 & 3 & 1 & 0 \\ 0 & 6 & 5 & 3 & 1 \\ 0 & 0 & 6 & 5 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 2 & 3 & 1 & 0 \\ 0 & 6 & 5 & 3 & 1 \\ 0 & 0 & 6 & 5 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 3 & 1 \\ 0 & 0 & 6 & 5 & 3 \end{bmatrix} \rightarrow$$

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

e, quindi, introducendo la matrice dei moltiplicatori, la seguente fattorizzazione

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 4 & 4 & 1 & 0 & 0 \\ 6 & 5 & 3 & 1 & 0 \\ 0 & 6 & 5 & 3 & 1 \\ 0 & 0 & 6 & 5 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 0 \\ 0 & 3 & 1 & 1 & 0 \\ 0 & 0 & 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Una matrice di ordine  $n$  e a banda  $(p, q)$  può essere memorizzata in forma compatta in una array rettangolare ABAND di dimensioni  $(p+q+1) \times n$ , con la convenzione che

$$a_{ij} = \text{ABAND}(i-j+q+1, j) \quad (2.30)$$

per tutte le coppie di indici  $(i, j)$  che sono interne alla banda. Come esemplificazione, la matrice precedente può essere memorizzata nella seguente forma

ABAND=	<table style="border-collapse: collapse; text-align: center;"> <tr><td>*</td><td><math>a_{12}</math></td><td><math>a_{23}</math></td><td><math>a_{34}</math></td><td><math>a_{45}</math></td></tr> <tr><td><math>a_{11}</math></td><td><math>a_{22}</math></td><td><math>a_{33}</math></td><td><math>a_{44}</math></td><td><math>a_{55}</math></td></tr> <tr><td><math>a_{21}</math></td><td><math>a_{32}</math></td><td><math>a_{43}</math></td><td><math>a_{54}</math></td><td>*</td></tr> <tr><td><math>a_{31}</math></td><td><math>a_{42}</math></td><td><math>a_{53}</math></td><td>*</td><td>*</td></tr> </table>	*	$a_{12}$	$a_{23}$	$a_{34}$	$a_{45}$	$a_{11}$	$a_{22}$	$a_{33}$	$a_{44}$	$a_{55}$	$a_{21}$	$a_{32}$	$a_{43}$	$a_{54}$	*	$a_{31}$	$a_{42}$	$a_{53}$	*	*	<table style="border-collapse: collapse; text-align: center;"> <tr><td>*</td><td>1</td><td>1</td><td>1</td><td>1</td></tr> <tr><td>2</td><td>4</td><td>3</td><td>3</td><td>3</td></tr> <tr><td>4</td><td>5</td><td>5</td><td>5</td><td>*</td></tr> <tr><td>6</td><td>6</td><td>6</td><td>*</td><td>*</td></tr> </table>	*	1	1	1	1	2	4	3	3	3	4	5	5	5	*	6	6	6	*	*	sopradiagonale diagonale 1 <sup>a</sup> sottodiagonale 2 <sup>a</sup> sottodiagonale
*	$a_{12}$	$a_{23}$	$a_{34}$	$a_{45}$																																							
$a_{11}$	$a_{22}$	$a_{33}$	$a_{44}$	$a_{55}$																																							
$a_{21}$	$a_{32}$	$a_{43}$	$a_{54}$	*																																							
$a_{31}$	$a_{42}$	$a_{53}$	*	*																																							
*	1	1	1	1																																							
2	4	3	3	3																																							
4	5	5	5	*																																							
6	6	6	*	*																																							

ove con \* si è indicato le posizioni di memoria che non sono utilizzate (tali posizioni potrebbero essere evitate memorizzando la banda in un unico vettore). Ad eliminazione terminata si ottiene

<table style="border-collapse: collapse; text-align: center;"> <tr><td>*</td><td>1</td><td>1</td><td>1</td><td>1</td></tr> <tr><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td></tr> <tr><td>2</td><td>1</td><td>1</td><td>1</td><td>*</td></tr> <tr><td>3</td><td>3</td><td>3</td><td>*</td><td>*</td></tr> </table>	*	1	1	1	1	2	2	2	2	2	2	1	1	1	*	3	3	3	*	*	sopradiagonale di <b>U</b> diagonale di <b>U</b> 1 <sup>a</sup> sottodiagonale di <b>L</b> 2 <sup>a</sup> sottodiagonale di <b>L</b>
*	1	1	1	1																	
2	2	2	2	2																	
2	1	1	1	*																	
3	3	3	*	*																	

Come ulteriore approfondimento della tecnica di memorizzazione (2.30), consideriamo il seguente algoritmo mediante il quale la moltiplicazione di una matrice a banda per un vettore può essere ottenuta mediante circa  $2n(p + q + 1)$  flops.

**Algoritmo 2.10** (Prodotto di matrice a banda per vettore) *Data una matrice  $\mathbf{A}$  a banda inferiore  $p$  e banda superiore  $q$  e memorizzata nella forma (2.30), e un vettore  $\mathbf{x} \in \mathbb{R}^n$ , il seguente algoritmo calcola  $\mathbf{z} = \mathbf{Ax}$  con un procedimento per colonne.*

```

 $z_i = 0, i = 1, 2, \dots, n$ 
For  $j = 1, 2, \dots, n$ 
   $itop = \max(1, j - q); ibot = \min(n, j + p)$ 
   $iatop = \max(1, q + 2 - j); iabot = iatop + ibot - itop$ 
  For  $i = itop, \dots, iabot$ 
     $z_i := z_i + x_j \text{ABAND}(iatop - itop + i, j)$ 
  end  $i$ 
end  $j$ 

```

Il metodo di eliminazione, che abbiamo visto applicato su un esempio particolare, può essere generalizzato nel seguente modo.

**Algoritmo 2.11** (Eliminazione di Gauss per matrici a banda) *Per una matrice  $\mathbf{A}$  di ordine  $n$  a banda  $(p, q)$ , la decomposizione  $\mathbf{A} = \mathbf{LU}$ , se esiste, può essere calcolata col seguente algoritmo, nel quale gli elementi  $a_{ij}$  sono sostituiti da  $l_{ij}$  se  $i > j$  e da  $u_{ij}$  se  $i \leq j$ .*

```

For  $k = 1, \dots, n - 1$ 
  For  $i = k + 1, \dots, \min(k + p, n)$ 
     $a_{ik} := a_{ik} / a_{kk}$ 
  end  $i$ 
  For  $i = k + 1, \dots, \min(k + p, n)$ 
    For  $j = k + 1, \dots, \min(k + q, n)$ 
       $a_{ij} := a_{ij} - a_{ik} a_{kj}$ 
    end  $j$ 
  end  $i$ 
end  $k$ 

```

L'algoritmo precedente richiede un numero di operazioni in flops dato da

$$\begin{cases} npq - \frac{1}{2}pq^2 - \frac{1}{6}p^3 + pn & \text{se } p \leq q \\ npq - \frac{1}{2}p^2q - \frac{1}{6}q^3 + qn & \text{se } p > q \end{cases}$$

Estensioni immediate si hanno per le decomposizioni  $\mathbf{LDM}^T$ . Per quanto riguarda, poi, la risoluzione dei sistemi triangolari a banda si hanno i seguenti algoritmi.

**Algoritmo 2.12** (Sistema triangolare a banda inferiore) *Sia  $\mathbf{L}$  una matrice di ordine  $n$  triangolare a banda inferiore  $p$  e tale che  $l_{ii} = 1, i = 1, 2, \dots, n$ . Dato  $\mathbf{b} \in \mathbb{R}^n$ , il seguente algoritmo calcola la soluzione del sistema  $\mathbf{Ly} = \mathbf{b}$ ; il vettore  $\mathbf{b}$  è sostituito dalla soluzione  $\mathbf{y}$ .*

```

For  $i = 1, 2, \dots, n$ 
   $b_i := b_i - \sum_{j=\max(1, i-p)}^{i-1} l_{ij} b_j$ 
end  $i$ 

```

**Algoritmo 2.13** (Sistema triangolare a banda superiore) *Sia  $\mathbf{U}$  una matrice di ordine  $n$  triangolare a banda superiore  $q$ . Dato  $\mathbf{b} \in \mathbb{R}^n$ , il seguente algoritmo calcola la soluzione del sistema  $\mathbf{U}\mathbf{x} = \mathbf{b}$ ; il vettore  $\mathbf{b}$  è sostituito dalla soluzione  $\mathbf{x}$ .*

```

For  $i = n, n-1, \dots, 2, 1$ 
   $b_i := (b_i - \sum_{j=i+1}^{\min(i+q, n)} u_{ij} b_j) / u_{ii}$ 
end  $i$ 

```

Gli algoritmi precedenti richiedono  $np - p^2/2$  (rispettivamente  $n(q+1) - q^2/2$ ) flops.

L'Algoritmo 2.11 non prevede la tecnica pivotale. In effetti, gli scambi delle righe possono allargare la banda della matrice. Per esempio, se  $\mathbf{A}$  è tridiagonale e alla prima eliminazione si scambiano le prime due righe, l'elemento  $u_{13}$  risulta in generale differente dallo zero. L'algoritmo di eliminazione di Gauss per le matrici a banda è pertanto particolarmente interessante nel caso che non sia richiesta l'operazione di pivoting, cioè ad esempio per le matrici a *predominanza diagonale* e le matrici *simmetriche definite positive*. Rinviamo alla bibliografia per una più adeguata trattazione, ci limiteremo ad analizzare alcuni algoritmi per le matrici simmetriche e definite positive.

**Algoritmo 2.14** (Algoritmo di Cholesky per matrici a banda) *Per una matrice  $\mathbf{A}$  simmetrica definita positiva e a banda  $p$ , il seguente algoritmo calcola una matrice triangolare inferiore  $\mathbf{R}$  con banda inferiore  $p$  tale che  $\mathbf{A} = \mathbf{R}\mathbf{R}^T$ . Gli elementi  $r_{ij}$  sostituiscono gli elementi  $a_{ij}$  ( $i \geq j$ ).*

```

For  $i = 1, \dots, n$ 
  For  $j = \max(1, i-p), \dots, i-1$ 
     $a_{ij} := (a_{ij} - \sum_{k=\max(1, i-p)}^{j-1} a_{ik} a_{jk}) / a_{jj}$ 
  end  $j$ 
   $a_{ii} := (a_{ii} - \sum_{k=\max(1, i-p)}^{i-1} a_{ik}^2)^{1/2}$ 
end  $i$ 

```

L'algoritmo richiede  $(np^2/2) - (p^3/3) + (3/2)(np - p^2)$  flops e  $n$  radici quadrate. Quindi per  $p \ll n$ , la decomposizione di Cholesky richiede  $n(p^2 + 3p)/2$  flops e  $n$  radici quadrate. Inoltre, la matrice  $\mathbf{A}$  può essere memorizzata in una array ABAND di dimensioni  $n \times (p+1)$  nella seguente forma (in cui  $p = 2$ )

$$\text{ABAND} = \begin{bmatrix} * & * & a_{11} \\ * & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} \\ a_{42} & a_{43} & a_{44} \\ \vdots & \vdots & \vdots \\ a_{n, n-2} & a_{n, n-1} & a_{nn} \end{bmatrix} \quad \text{con } a_{ij} = \text{ABAND}(i, j - i + p + 1)$$

Per valori piccoli di  $p$  le radici quadrate costituiscono una parte consistente del calcolo. In questo caso si preferisce una decomposizione  $\mathbf{LDL}^T$ , nella quale non è necessario il calcolo delle radici quadrate. Come illustrazione, esplicitiamo il caso di una matrice  $\mathbf{A}$  tridiagonale simmetrica. Indichiamo con  $\mathbf{L}$  la matrice bidiagonale triangolare inferiore e con  $\mathbf{D}$  la matrice diagonale  $\mathbf{diag}(d_1, \dots, d_n)$ . Dall'uguaglianza

$$\begin{bmatrix} 1 & & & & 0 \\ c_1 & 1 & & & \\ & c_2 & 1 & & \\ & & \diagdown & \diagup & \\ 0 & & & c_{n-1} & 1 \end{bmatrix}$$

$\mathbf{A} = \mathbf{LDL}^T$  si ha

$$\begin{aligned} a_{11} &= d_1 \\ a_{k,k-1} &= c_{k-1}d_{k-1} && k = 2, 3, \dots, n \\ a_{k,k} &= d_k + c_{k-1}^2d_{k-1} = d_k + c_{k-1}a_{k,k-1} && k = 2, 3, \dots, n \end{aligned}$$

da cui

$$\begin{aligned} d_1 &= a_{11} \\ c_{k-1} &= a_{k,k-1}/d_{k-1} \\ d_k &= a_{kk} - c_{k-1}a_{k,k-1} \end{aligned}$$

per  $k = 2, \dots, n$ . Completando con la soluzione dei sistemi  $\mathbf{Ly}=\mathbf{b}$ ,  $\mathbf{Dz}=\mathbf{y}$  e  $\mathbf{L}^T\mathbf{x}=\mathbf{z}$ , si ha il seguente algoritmo che richiede  $5n$  flops.

**Algoritmo 2.15** (Algoritmo di Cholesky per matrici tridiagonali) *Data una matrice  $\mathbf{A}$  tridiagonale simmetrica definita positiva, con gli elementi della diagonale memorizzati nel vettore  $[d_1, d_2, \dots, d_n]$  e gli elementi sopra la diagonale nel vettore  $[c_1, c_2, \dots, c_n]$ , il seguente algoritmo calcola la soluzione del sistema  $\mathbf{Ax}=\mathbf{b}$ ; il vettore  $\mathbf{x}$  sostituisce il vettore  $\mathbf{b}$ .*

```

For  $k = 2, \dots, n$ 
   $t := c_{k-1}$ ,  $c_{k-1} := t/d_{k-1}$ ,  $d_k := d_k - tc_{k-1}$ 
end  $k$ 
For  $k = 2, \dots, n$ 
   $b_k := b_k - c_{k-1}b_{k-1}$ 
end  $k$ 
 $b_n := b_n/d_n$ 
end  $k$ 
For  $k = n-1, \dots, 1$ 
   $b_k := b_k/d_k - c_k b_{k+1}$ 
end  $k$ 

```

### 2.1.10 Matrici sparse

In termini qualitativi, una matrice è considerata *sparsa* quando il suo ordine è sufficientemente elevato, ma i suoi elementi sono in “predominanza” nulli. Le matrici a banda che abbiamo considerato nel paragrafo precedente sono, quindi, casi particolari di matrici sparse. Ma, più in generale, nelle applicazioni (cfr. per una esemplificazione l’Esempio 2.1), una matrice sparsa può avere gli elementi diversi dallo zero non necessariamente addensati intorno alla diagonale. Per tali matrici si pone da una parte il problema di una loro conveniente rappresentazione in memoria, e dall’altra l’individuazione di opportune strategie nell’applicazione del metodo di eliminazione, in maniera da limitare il più possibile il fenomeno del *fill-in*. Per *fill-in* si intende il fatto che un elemento della matrice originariamente nullo diventa diverso dallo zero per effetto della sostituzione. Per la risoluzione dei problemi ora prospettati sono state proposte diverse soluzioni. In questo paragrafo forniremo una breve introduzione, che evidenzia, tuttavia, le difficoltà di base. Nel seguito considereremo l’applicazione dei metodi iterativi, che rispetto ai metodi diretti hanno il vantaggio di non modificare la matrice.

► **Esempio 2.8** Analizziamo l’applicazione del metodo di eliminazione ad un sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{A}$  matrice della seguente forma

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ a_{n1} & 0 & \cdots & 0 & a_{nn} \end{bmatrix}$$

La matrice assegnata è una particolare matrice *sparsa*, detta anche matrice *bordata* (bordered). Si tratta di una matrice completamente nulla, salvo eventualmente la diagonale, la prima riga e la prima colonna. Per la *memorizzazione* di una matrice di questo tipo sono quindi sufficienti tre vettori. Quando, tuttavia, si applica il metodo di eliminazione nella forma usuale, il primo passo di eliminazione sostituisce coefficienti nulli della matrice con coefficienti, in generale, diversi dallo zero; si verifica cioè un *fill-in*. Come esemplificazione, si consideri la seguente trasformazione, ottenuta sottraendo la prima equazione dalle successive

$$\begin{bmatrix} 2 & 1 & -1 & -1 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & 1 & -1 & -1 \\ 0 & 1 & 1 & 1 \\ 0 & -1 & 2 & 1 \\ 0 & -1 & 1 & 2 \end{bmatrix}$$

Come si vede, in questo caso tutti gli elementi nulli della matrice originaria dopo la prima eliminazione sono sostituiti da numeri diversi dallo zero. Ricordiamo, anche, che le posizioni degli elementi che si azzerano nella prima colonna sono, nel procedimento di eliminazione, usualmente utilizzati per memorizzare i moltiplicatori, e quindi non sono ulteriormente disponibili. Si conclude, quindi, che procedendo nel modo precedente, anche se per memorizzare la matrice originaria bastano tre vettori, successivamente è necessaria la memorizzazione di una matrice  $n \times n$ . Tuttavia, un’osservazione importante per il trattamento delle matrici

sparsa è che l'effetto fill-in può dipendere dall'*ordinamento* delle righe e delle colonne della matrice. Nel sistema lineare tale ordinamento corrisponde all'ordine in cui si considerano le equazioni e le incognite.

Nel caso delle matrici bordate si può eliminare completamente il fill-in procedendo nel seguente modo. Si scambiano tra loro la prima e l'ultima equazione (la prima e l'ultima riga della matrice) e successivamente l'incognita  $x_1$  con l'incognita  $x_n$  (la prima e l'ultima colonna). Si ottiene in questo modo la seguente matrice

$$\hat{\mathbf{A}} = \begin{bmatrix} a_{nn} & 0 & \cdots & 0 & a_{n1} \\ 0 & a_{22} & 0 & \cdots & a_{21} \\ 0 & 0 & a_{33} & \cdots & a_{31} \\ \cdots & \vdots & \ddots & \ddots & \vdots \\ a_{1n} & a_{12} & a_{13} & \cdots & a_{11} \end{bmatrix} \quad (2.31)$$

In questo caso per avere una matrice triangolare è sufficiente sottrarre dall'ultima equazione le precedenti moltiplicate per opportuni coefficienti. Se memorizziamo la matrice mediante tre vettori  $\mathbf{l}$ ,  $\mathbf{u}$ ,  $\mathbf{d}$  nella forma

$$\begin{bmatrix} d_1 & 0 & 0 & \cdots & u_1 \\ 0 & d_2 & 0 & \cdots & u_2 \\ 0 & \cdot & d_3 & \cdot & u_3 \\ \cdots & & & & \cdots \\ l_1 & l_2 & \cdots & & d_n \end{bmatrix}$$

si può fattorizzare la matrice mediante il seguente algoritmo, che sostituisce gli elementi  $l_i$  con i moltiplicatori  $l_i/d_i$  e sostituisce  $d_n$  con l'elemento diagonale  $n$ -mo della matrice  $\mathbf{U}$ .

```
for  $i = 1 : n - 1$ 
   $l_i \leftarrow l_i/d_i$ 
   $d_n \leftarrow d_n - l_i u_i$ 
end  $i$ 
```

Per matrici *sparse di tipo generale* la ricerca di un ordinamento ottimale per ridurre il fill-in non è un problema di facile soluzione. Per un opportuno approfondimento rinviamo ad esempio a George e Liu [65]. Qui ci limiteremo a considerare alcune tecniche di memorizzazione di una matrice sparsa, senza particolari strutture.

Un procedimento, attribuito a Gustavson, utilizza un vettore reale e due vettori interi. Più precisamente, un vettore  $A$  contiene gli elementi diversi dallo zero riga per riga, un vettore intero  $C$  memorizza l'indice di colonna di ogni elemento diverso dallo zero e infine un vettore intero  $R$  memorizza la posizione nel vettore  $A$  del primo elemento diverso dallo zero in ciascuna riga. Come esemplificazione, la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 \\ 0 & 0 & 0 & 7 & 1 \\ 0 & 4 & 1 & 0 & 2 \end{bmatrix}$$

è memorizzata come segue



I	1	2	3	4	5	6	7	8	9	10	11
A(I)	3	1	1	2	5	1	7	1	4	1	2
C(I)	1	3	1	2	3	4	4	5	2	3	5
R(I)	1	3	5	7	9	12					

Ad ogni elemento diverso dallo zero della matrice  $\mathbf{A} = [a_{ij}]$  corrisponde un elemento nel vettore  $A$  ed un elemento nel vettore  $C$ . In  $R(n+1)$  si pone, per convenienza, il numero degli elementi in  $A$  più uno.

Vediamo ora come localizzare nel vettore  $A$  un generico elemento  $a_{ij}$  della matrice  $\mathbf{A}$ . Consideriamo ad esempio l'elemento  $a_{53}$ . Poiché  $R(5) = 9$ , gli elementi diversi dallo zero della riga 5 della matrice  $\mathbf{A}$  sono memorizzati nel vettore  $A$  a partire dall'indice 9. A partire da tale indice, si avanza nel vettore  $C$  fino a trovare l'indice di colonna 3. Se si raggiunge la fine della memorizzazione della riga 5 vuol dire che il coefficiente è nullo. Nel caso dell'esempio si ha  $3 = C(10)$ , per cui  $a_{53}$  è memorizzato in  $A(10)$ . Più in generale, il seguente programma localizza il generico elemento  $a_{ij}$  e pone il corrispondente valore nella variabile  $T$

```

T=0.
L=R(I)
K=R(I+1)-1
DO 5 M=L,K
5   IF(C(M).EQ.J) GOTO 10
   GOTO 20
10  T=A(M)
20  CONTINUE

```

Il programma precedente permette di adattare in maniera semplice il metodo di eliminazione. Consideriamo ora un'altra procedura di memorizzazione, che può essere conveniente quando gli elementi diversi dallo zero della matrice sparsa sono *vicini* tra loro in ogni riga. La procedura, nota anche come *profile storage*, utilizza come la precedente tre vettori. Un vettore  $A$  memorizza gli elementi compresi, in ogni riga, tra il primo e l'ultimo elemento diversi dallo zero. Il vettore intero  $C$  memorizza la colonna del primo elemento diverso dallo zero in ogni riga e il vettore intero  $R$  memorizza la locazione in  $A$  del primo elemento diverso dallo zero in ogni riga. Per la matrice  $\mathbf{A}$  precedente si ha la seguente struttura

I	1	2	3	4	5	6	7	8	9	10	11	12	13
A(I)	3	0	1	1	2	5	1	7	1	4	1	0	2
C(I)	1	1	3	4	2								
R(I)	1	4	6	8	10	14							

In questa procedura possono essere memorizzati anche alcuni coefficienti nulli, per cui la dimensione del vettore  $A$  risulta più elevata che non nella struttura precedente. Viceversa, la dimensione del vettore  $C$  è più piccola, in quanto si memorizza solo la colonna del primo elemento diverso dallo zero in ciascuna riga. Per localizzare, ad esempio, l'elemento  $a_{53}$  di  $\mathbf{A}$ , si esamina dapprima il vettore  $R$ . Poiché  $R(5) = 10$ , il primo elemento diverso dallo zero nella riga 5 è memorizzato in  $A(10)$ . Poiché  $C(5) = 2$ , tale elemento è nella colonna 2 della matrice  $\mathbf{A}$  e corrisponde a  $a_{52}$ . In definitiva, si ha che  $a_{53}$  è memorizzato in  $A(10+1) = A(11)$ . Più in generale, il seguente segmento di programma individua in  $A$  l'elemento  $a_{ij}$  e pone il suo valore nella variabile  $T$ .

```

T=0.
IF(J.LT.C(I)) GOTO 5
M=J-C(I)+R(I)
IF (M.GE.R(I+1)) GOTO 5
T=A(M)
5   CONTINUE

```

La procedura *profile* può essere opportunamente adattata quando la matrice è *simmetrica*. Si ottiene in tale caso lo schema noto come *Jennings profile*. In esso si memorizzano nel vettore  $A$  gli elementi che in ogni riga sono tra il primo elemento diverso dallo zero a sinistra della diagonale e l'elemento diagonale. Nel vettore intero  $R$  si memorizza la locazione nel vettore  $A$  del primo elemento diverso dallo zero in ciascuna riga. Ad esempio, la seguente matrice simmetrica

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 & 0 & 0 \\ 2 & 6 & 0 & 0 & 0 \\ 4 & 0 & 3 & 1 & 0 \\ 0 & 0 & 1 & 6 & 5 \\ 0 & 0 & 0 & 5 & 1 \end{bmatrix}$$

è memorizzata nella seguente maniera

I	1	2	3	4	5	6	7	8	9	10
$A(I)$	1	2	6	4	0	3	1	6	5	1
$C(I)$	1	2	4	7	9	11				

Per localizzare ad esempio l'elemento  $a_{43}$  nel vettore  $A$ , si osserva che  $R(5) = 9$  e quindi  $a_{44}$ , l'elemento diagonale nella riga 4 è memorizzato in  $A(9 - 1) = A(8)$ . Poiché  $a_{43}$  è nella posizione immediatamente precedente a  $a_{44}$  nella riga 4, si ha che  $a_{43}$  è memorizzato in  $A(8 - 1)$ . Il seguente segmento di programma localizza al solito l'elemento  $a_{ij}$  e lo memorizza in  $T$ .

```

T=0.
IF(I.GT.J) GOTO 5
M=R(J)
L=R(J+1)+I-J-1
GOTO 10
5   M=R(I)
    L=R(I+1)+J-I-1
10  IF(L.LT.M) GOTO 15
    T=A(L)
15  CONTINUE

```

Un aspetto importante della struttura *Jennings profile* è che, nel caso in cui la matrice simmetrica non richieda la procedura pivoting, ad esempio quando essa è definita positiva, la fattorizzazione della matrice può essere ottenuta rimanendo nella medesima struttura di memorizzazione. In altre parole, il fill-in corrisponde ad elementi che già sono memorizzati. Nel caso dell'esempio della matrice simmetrica  $\mathbf{A}$ , nella fattorizzazione lo zero nella posizione  $(3, 2)$  è sostituito da un elemento diverso dallo zero; ma, come abbiamo visto, l'elemento  $a_{32}$  è stato memorizzato nel vettore  $A$ . Al contrario, nelle due procedure viste precedentemente, è necessario *prevedere* a priori sufficiente spazio per il fill-in. ■

### 2.1.11 Introduzione all'updating

Le tecniche per aggiornare (*updating*) la fattorizzazione di una matrice hanno un ruolo centrale nell'algebra lineare moderna e nella ottimizzazione. In questo paragrafo forniremo, mediante opportune esemplificazioni, alcune idee di base, rinviando ad esempio a Gill et al. [68] per una trattazione più adeguata.

Se  $\mathbf{B}$  è una matrice di ordine  $n$  e  $\mathbf{u}$ ,  $\mathbf{v}$  sono due vettori in  $\mathbb{R}^n$  la matrice  $\mathbf{A}$  definita nel seguente modo

$$\mathbf{A} = \mathbf{B} - \mathbf{u}\mathbf{v}^T \quad (2.32)$$

è detta un aggiornamento della matrice  $\mathbf{B}$  ottenuta mediante la matrice di rango 1  $\mathbf{u}\mathbf{v}^T$ . Come esemplificazione, si consideri la seguente decomposizione

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 & 2 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad (2.33)$$

In questo caso la matrice  $\mathbf{B}$  è una matrice triangolare, e  $\mathbf{u} = [0, 1, 1, 1]^T$ ,  $\mathbf{v} = -[1, 0, 0, 0]^T$ . Nelle applicazioni ha interesse ottenere un modo rapido (ossia, con un numero di operazioni di ordine  $n^2$ ) per ottenere una fattorizzazione della matrice  $\mathbf{A}$ , una volta che sia nota una fattorizzazione di  $\mathbf{B}$ . In maniera equivalente, si tratta di costruire l'inversa  $\mathbf{A}^{-1}$ , a partire dall'inversa di  $\mathbf{B}$ . Una risposta teorica a tale problema è fornita dalla seguente formula, nota come *formula di Sherman-Morrison* (cfr. Appendice A)

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \alpha \mathbf{B}^{-1} \mathbf{u}\mathbf{v}^T \mathbf{B}^{-1} \quad (2.34)$$

ove  $\alpha = 1/(1 - \mathbf{v}^T \mathbf{B}^{-1} \mathbf{u})$ . La formula (2.34) può essere utilizzata per risolvere il sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  nel seguente modo

$$\begin{aligned} \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} &= (\mathbf{B}^{-1} + \alpha \mathbf{B}^{-1} \mathbf{u}\mathbf{v}^T \mathbf{B}^{-1})\mathbf{b} \\ &= \mathbf{B}^{-1}\mathbf{b} + \alpha \mathbf{B}^{-1} \mathbf{u}\mathbf{v}^T \mathbf{B}^{-1}\mathbf{b} \\ &= \mathbf{B}^{-1}\mathbf{b} + \alpha (\mathbf{v}^T \mathbf{B}^{-1}\mathbf{b}) \mathbf{B}^{-1}\mathbf{u} \\ &= \mathbf{B}^{-1}\mathbf{b} + \beta \mathbf{B}^{-1}\mathbf{u} \end{aligned}$$

ove  $\beta = \alpha (\mathbf{v}^T \mathbf{B}^{-1}\mathbf{b}) = \mathbf{v}^T \mathbf{B}^{-1}\mathbf{b} / (1 - \mathbf{v}^T \mathbf{B}^{-1}\mathbf{u})$ . Posto  $\mathbf{y} = \mathbf{B}^{-1}\mathbf{b}$  e  $\mathbf{z} = \mathbf{B}^{-1}\mathbf{u}$  si ha

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{B}^{-1}\mathbf{b} + \beta \mathbf{B}^{-1}\mathbf{u} = \mathbf{y} + \frac{\mathbf{v}^T \mathbf{y}}{1 - \mathbf{v}^T \mathbf{z}} \mathbf{z}$$

In definitiva, la soluzione del sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  può essere ottenuta nel seguente modo

1. Si fattorizza  $\mathbf{B}$ .
2. Si risolvono i sistemi  $\mathbf{B}\mathbf{y} = \mathbf{b}$ ,  $\mathbf{B}\mathbf{x} = \mathbf{u}$ .

3. Si calcola lo scalare  $\beta = \mathbf{v}^T \mathbf{y} (1 - \mathbf{v}^T \mathbf{z})$ .
4. Si pone  $\mathbf{x} = \mathbf{y} + \beta \mathbf{z}$ .

► **Esempio 2.9** La seguente matrice

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 0 & 0 & -1 \\ -1 & 4 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 4 & -1 \\ -1 & 0 & 0 & -1 & 3 \end{bmatrix} \quad (2.35)$$

ha origine nell'approssimazione mediante il metodo delle differenze finite dei problemi differenziali del secondo ordine con *condizioni periodiche*. Essa può essere espressa nella somma  $\mathbf{A} = \mathbf{B} - \mathbf{u}\mathbf{v}$ , assumendo

$$\begin{bmatrix} 4 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & 4 \end{bmatrix}; \quad \mathbf{u} = \mathbf{v} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

ove  $\mathbf{B}$  è una matrice *tridiagonale*. Applicando l'algoritmo precedente, si ha

$$\mathbf{b} = \begin{bmatrix} 0 \\ 4 \\ 9 \\ 0 \\ 1 \end{bmatrix}; \quad \mathbf{B}\mathbf{y} = \mathbf{b} \Rightarrow \mathbf{y} = \begin{bmatrix} 0.4615 \\ 1.8462 \\ 2.9231 \\ 0.8462 \\ 0.4615 \end{bmatrix}; \quad \mathbf{B}\mathbf{z} = \mathbf{u} \Rightarrow \mathbf{z} = \begin{bmatrix} 0.2692 \\ 0.0769 \\ 0.0385 \\ 0.0769 \\ 0.2692 \end{bmatrix}$$

da cui

$$\beta = \frac{\mathbf{v}^T \mathbf{y}}{1 - \mathbf{v}^T \mathbf{z}} = 2; \quad \mathbf{x} = \mathbf{y} + \beta \mathbf{z} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 1 \end{bmatrix}$$

▼ **Osservazione 2.1** La procedura basata sulla formula di Sherman-Morrison può presentare, come si può mostrare attraverso opportune esemplificazioni, problemi di stabilità numerica. Procedure più stabili possono essere ottenute utilizzando la decomposizione  $\mathbf{B} = \mathbf{Q}\mathbf{R}$  in una matrice ortogonale e una matrice triangolare; i fattori della decomposizione della matrice modificata  $\mathbf{A}$  possono essere ottenuti mediante la modifica diretta dei fattori della decomposizione di  $\mathbf{B}$ . Il metodo  $\mathbf{QR}$  sarà considerato nei paragrafi successivi. Per la sua applicazione all'aggiornamento di una matrice mediante matrici di rango uno si veda, ad esempio Gill et al. [67]. ■

### 2.1.12 Fattorizzazione a blocchi

Le idee discusse in questo paragrafo possono essere estese alle matrici a blocchi, trattando formalmente i blocchi come gli elementi della matrice. Come esemplificazione,

consideriamo la seguente decomposizione a due blocchi

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad (2.36)$$

ove  $\mathbf{A}_{11}, \mathbf{A}_{22}$  sono matrici quadrate, dette *blocchi diagonali*, non necessariamente diagonali. La fattorizzazione  $\mathbf{LU}$  di  $\mathbf{A}$  può essere scritta nella seguente forma

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{11} & \\ & \mathbf{L}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ & \mathbf{U}_{22} \end{bmatrix} \quad (2.37)$$

ove  $\mathbf{L}_{11}$  e  $\mathbf{L}_{22}$  sono sottomatrici triangolari inferiori e  $\mathbf{U}_{11}$  e  $\mathbf{U}_{22}$  sono sottomatrici triangolari superiori. Dalla definizione (2.37) si ricavano le seguenti relazioni

$$\mathbf{A}_{11} = \mathbf{L}_{11}\mathbf{U}_{11} \quad (2.38)$$

$$\mathbf{L}_{11}\mathbf{U}_{12} = \mathbf{A}_{12} \quad (2.39)$$

$$\mathbf{L}_{21}\mathbf{U}_{11} = \mathbf{A}_{21} \quad (2.40)$$

$$\mathbf{A}_{22} - \mathbf{L}_{21}\mathbf{U}_{12} = \mathbf{L}_{22}\mathbf{U}_{22} \quad (2.41)$$

Pertanto, la fattorizzazione (2.36) può essere costruita calcolando successivamente la fattorizzazione della matrice  $\mathbf{A}_{11}$ , la formazione delle colonne di  $\mathbf{U}_{12}$  e delle righe di  $\mathbf{L}_{21}$  mediante una sostituzione in avanti, e infine la fattorizzazione della matrice  $\mathbf{A}_{22} - \mathbf{L}_{21}\mathbf{U}_{12}$ . Per ognuna delle fattorizzazioni triangolari richieste si utilizzano le tecniche che abbiamo esaminato in precedenza, con eventuali opportune permutazioni di righe e di colonne. Come nel caso della fattorizzazione  $\mathbf{LU}$  per punti, si può scegliere alternativamente che la matrice  $\mathbf{U}$  o la matrice  $\mathbf{L}$  siano triangolari con elementi sulla diagonale principale uguali a 1. Scegliendo ad esempio la matrice  $\mathbf{U}$ , si ha la decomposizione

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \\ & \overline{\mathbf{A}}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \overline{\mathbf{U}}_{12} \\ & \mathbf{I} \end{bmatrix} \quad (2.42)$$

con

$$\mathbf{A}_{11}\overline{\mathbf{U}}_{12} = \mathbf{A}_{12} \quad (2.43)$$

$$\overline{\mathbf{A}}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\overline{\mathbf{U}}_{12} \quad (2.44)$$

Data la decomposizione  $\mathbf{A} = \mathbf{LU}$ , ottenuta nel modo (2.42), si ottiene la soluzione del sistema lineare  $\mathbf{Ax} = \mathbf{b}$  risolvendo successivamente i sistemi lineari

$$\begin{aligned} \mathbf{Ly} = \mathbf{b} &\Rightarrow \begin{cases} \mathbf{A}_{11}\mathbf{y}_1 = \mathbf{b}_1 \\ \overline{\mathbf{A}}_{22}\mathbf{y}_2 = \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{y}_1 \end{cases} \\ \mathbf{Ux} = \mathbf{y} &\Rightarrow \begin{cases} \mathbf{x}_2 = \mathbf{y}_2 \\ \mathbf{x}_1 = \mathbf{y}_1 - \mathbf{U}_{12}\mathbf{x}_2 \end{cases} \end{aligned}$$

Osserviamo che il procedimento precedente richiede che la matrice  $\mathbf{A}_{11}$  (il pivot a blocchi) sia non singolare. Il fatto che la matrice  $\mathbf{A}$  sia non singolare non garantisce la non singolarità della matrice  $\mathbf{A}_{11}$ , come mostra il seguente semplice esempio

$$\mathbf{A} = \left[ \begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 1 & -1 \\ \hline 0 & 1 & 1 \end{array} \right]$$

◆ **Esercizio 2.1** Risolvere il seguente sistema di equazioni

$$\begin{cases} 0.5x_1 & +x_3=1 \\ x_1+2x_2-x_3=0 \\ x_1 & +x_3=0 \end{cases}$$

usando il metodo di eliminazione di Gauss e il pivoting parziale.

◆ **Esercizio 2.2** Determinare una matrice di permutazione  $\mathbf{P}$  tale che la matrice

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ -1 & 2 & 3 & -1 \\ 3 & -1 & -1 & 2 \end{bmatrix}$$

ammetta la fattorizzazione  $\mathbf{PA}=\mathbf{LU}$ .

◆ **Esercizio 2.3** Riscrivere l'Algoritmo 2.2 supponendo che la matrice  $\mathbf{U}$  sia memorizzata sotto forma di vettore di dimensione 1  $[u_{11}, u_{12}, u_{22}, u_{13}, u_{23} \dots]$ .

◆ **Esercizio 2.4** Sia  $\mathbf{A}$  una matrice non singolare. Estendere il metodo di eliminazione con pivoting parziale per risolvere  $p > 1$  sistemi lineari, corrispondenti a  $p$  vettori termini noti diversi. Ricavarne un algoritmo per il calcolo di  $\mathbf{A}^{-1}$ .

◆ **Esercizio 2.5** Metodo di Gauss-Jordan. Si considerino le matrici elementari  $\mathbf{N}(\mathbf{y}, i) = \mathbf{I} + \mathbf{y}\mathbf{e}_i^T$ , con  $\mathbf{y} \in \mathbb{R}^n$ . Dato  $\mathbf{x} \in \mathbb{R}^n$  esaminare il calcolo di  $\mathbf{y}$  in modo che  $\mathbf{N}(\mathbf{y}, i)\mathbf{x} = \mathbf{e}_i$ ; proporre un algoritmo basato sulle trasformazioni precedenti per il calcolo di  $\mathbf{A}^{-1}$  con il risultato sostituito alla matrice  $\mathbf{A}$ .

◆ **Esercizio 2.6** Mostrare che la matrice non singolare

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

non ha una decomposizione  $\mathbf{A} = \mathbf{LU}$ . Trovare una matrice permutazione  $\mathbf{P}$  tale che  $\mathbf{PA}$  abbia una decomposizione  $\mathbf{LU}$ .

◆ **Esercizio 2.7** Modificare l'Algoritmo 2.8, supponendo che la matrice simmetrica  $\mathbf{A}$  sia memorizzata in un vettore di dimensione  $n(n+1)/2$ , nella seguente forma

$$\mathbf{A} = [a_{11}, a_{21}, a_{22}, a_{31}, \dots, a_{n1}, a_{n2}, \dots, a_{nn}]$$

◆ **Esercizio 2.8** Calcolare il numero delle operazioni richieste per il calcolo dell'inversa di una matrice tridiagonale con il metodo di Gauss, senza scambio di righe. Considerare, quindi, il caso in cui la matrice è simmetrica.

◆ **Esercizio 2.9** Studiare opportune strategie per la risoluzione del sistema  $\mathbf{A}^2\mathbf{x} = \mathbf{b}$ , nell'ipotesi di conoscere la matrice  $\mathbf{A}$ .

◆ **Esercizio 2.10** Modificare l'Algoritmo 2.14 supponendo che la matrice  $\mathbf{A}$  sia memorizzata per diagonali (relative alla banda) in un vettore a una dimensione.

◆ **Esercizio 2.11** Esaminare il caso particolare di sistema a banda  $\mathbf{H}\mathbf{x} = \mathbf{b}$  con  $\mathbf{H}$  matrice di Hessenberg.

◆ **Esercizio 2.12** Indicare una strategia per il calcolo del seguente vettore

$$\mathbf{x} = \mathbf{B}^{-1}(2\mathbf{A} + \mathbf{I})(\mathbf{C}^{-1} + \mathbf{A})\mathbf{b}$$

senza il calcolo di matrici inverse.

◆ **Esercizio 2.13** Supponendo di avere calcolata la decomposizione  $\mathbf{LU}$  della matrice  $\mathbf{A}$ , scrivere un programma, colonna orientato, per risolvere  $\mathbf{A}^T\mathbf{x} = \mathbf{b}$ .

## 2.2 Analisi degli errori; condizionamento e stabilità

Quando un sistema lineare viene risolto numericamente, ossia mediante un particolare algoritmo implementato su un calcolatore, la soluzione ottenuta differisce usualmente dalla soluzione esatta, cioè dalla soluzione che si potrebbe ottenere eseguendo le operazioni con una precisione infinita. Dal punto di vista delle applicazioni, è naturalmente importante fornire una stima dell'errore commesso<sup>6</sup>.

Con riferimento all'analisi introdotta più in generale nel capitolo precedente, possiamo separare due tipi di contributi nella formazione dell'errore globale. Il primo si riferisce al *condizionamento* del problema, ed è essenzialmente legato con i dati, ossia la matrice  $\mathbf{A}$  e il termine noto  $\mathbf{b}$ , ed è indipendente dall'algoritmo, mentre il secondo riguarda la stabilità dell'algoritmo. Nei paragrafi precedenti abbiamo visto, in particolare, che la procedura del pivoting diminuisce l'effetto della propagazione degli errori di arrotondamento e quindi fornisce un algoritmo più stabile; ma anche tale procedura può risultare inefficace, quando il sistema non è ben condizionato.

In questo paragrafo, oltre che precisare il senso del condizionamento per il problema della risoluzione di un sistema lineare, e più in particolare definire una misura di condizionamento, analizzeremo alcuni algoritmi numerici più idonei, rispetto a

<sup>6</sup> When a problem in pure or in applied mathematics is "solved" by numerical computation, errors, that is, deviations of the numerical "solution" obtained from the true, rigorous one, are unavoidable. Such a "solution" is therefore meaningless, unless there is an estimate of the total error in the above sense, J. Von Neumann, H. H. Goldstine (1947).

quelli esaminati nei paragrafi precedenti, a risolvere sistemi lineari mediamente mal condizionati. Per quanto riguarda, invece, i sistemi decisamente mal condizionati, lo studio del malcondizionamento può essere utile per individuare le cause di tale comportamento ed avere quindi suggerimenti per modificare opportunamente il modello. Introduciamo l'analisi del condizionamento mediante un semplice esempio di sistema mal condizionato.

► **Esempio 2.10** Consideriamo il seguente sistema lineare

$$\begin{cases} 2x_1 + 3x_2 = 5 \\ 2x_1 + 3.1x_2 = 5.1 \end{cases} \quad (2.45)$$

che ha come soluzione esatta  $x_1 = x_2 = 1$ . Se modifichiamo la matrice dei coefficienti e il termine noto nella seguente maniera

$$\mathbf{A} = \begin{bmatrix} 2.000 & 3.000 \\ 2.000 & 3.100 \end{bmatrix} \Rightarrow \tilde{\mathbf{A}} = \begin{bmatrix} 2.000 & 3.000 \\ 1.999 & 3.000 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 5.000 \\ 5.100 \end{bmatrix} \Rightarrow \tilde{\mathbf{b}} = \begin{bmatrix} 5.000 \\ 4.990 \end{bmatrix}$$

la soluzione esatta del sistema perturbato  $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$  diventa  $\tilde{x}_1 = 10$ ,  $\tilde{x}_2 = -5$ . Posto

$$\delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}, \quad \delta\mathbf{A} = \tilde{\mathbf{A}} - \mathbf{A}, \quad \delta\mathbf{b} = \tilde{\mathbf{b}} - \mathbf{b}$$

si ha

$$\frac{\|\delta\mathbf{A}\|_1}{\|\mathbf{A}\|_1} = 0.0164; \quad \frac{\|\delta\mathbf{b}\|_1}{\|\mathbf{b}\|_1} = 0.0109; \quad \frac{\|\delta\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = 7.5; \quad \frac{\|\delta\mathbf{x}\|_1}{\|\tilde{\mathbf{x}}\|_1} = 1.34$$

Si vede, pertanto, che l'errore relativo sui risultati è di due ordini di grandezza superiore agli errori relativi sui dati. È interessante esaminare l'inversa della matrice  $\mathbf{A}$ ; si ha

$$\mathbf{A}^{-1} = \begin{bmatrix} 15.500 & -15.000 \\ -10.000 & 10.000 \end{bmatrix}$$

con  $\|\mathbf{A}^{-1}\|_1 = 25.5$ , mentre  $\|\mathbf{A}\|_1 = 6.1$ . Come vedremo nel seguito, il numero  $\mu_1(\mathbf{A}) := \|\mathbf{A}\|_1 \|\mathbf{A}^{-1}\|_1$  fornisce una indicazione del fattore di amplificazione degli errori nei dati. In questo caso si ha  $\mu_1(\mathbf{A}) = 155.55$ . ■

Lo studio del condizionamento di un sistema può essere fatto *perturbando* i dati ed esaminando gli effetti prodotti da tali perturbazioni sulla soluzione. Per semplicità, consideriamo separatamente gli effetti prodotti delle variazioni sul termine noto  $\mathbf{b}$  e quelli sulla matrice  $\mathbf{A}$ .

Sia  $\mathbf{A}$  una matrice non singolare e siano  $\mathbf{x}$  e  $\mathbf{x} + \delta\mathbf{x}$  le rispettive soluzioni dei seguenti due sistemi

$$\mathbf{A}\mathbf{x} = \mathbf{b}; \quad \mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$$

Sottraendo le due equazioni, si ricava

$$\mathbf{A}(\delta\mathbf{x}) = \delta\mathbf{b} \Rightarrow \delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b}$$

Indicando con  $\|\cdot\|$  una *norma naturale* di matrice, ad esempio una delle norme  $\|\cdot\|_1$ ,  $\|\cdot\|_2$ ,  $\|\cdot\|_\infty$ , si ha

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{b}\| \leq \|\mathbf{A}^{-1}\| \|\delta\mathbf{b}\|$$



D'altra parte, si ha

$$\|\mathbf{b}\| = \|\mathbf{Ax}\| \leq \|\mathbf{A}\| \|\mathbf{x}\| \Rightarrow \frac{1}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}$$

da cui

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \boxed{\|\mathbf{A}\| \|\mathbf{A}^{-1}\|} \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \quad (2.46)$$

Si può mostrare che per ogni norma naturale fissata esistono dei vettori  $\delta\mathbf{b}$  e  $\mathbf{x}$  tali che

$$\|\mathbf{A}^{-1}\delta\mathbf{b}\| = \|\mathbf{A}^{-1}\| \|\delta\mathbf{b}\|; \quad \|\mathbf{Ax}\| = \|\mathbf{A}\| \|\mathbf{x}\|$$

per cui la maggiorazione (2.46) è *stretta*.

Consideriamo, ora, il caso in cui  $\mathbf{x}$  e  $\mathbf{x} + \delta\mathbf{x}$  sono le rispettive soluzioni dei seguenti due sistemi

$$\mathbf{Ax} = \mathbf{b}, \quad (\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$$

Con semplici passaggi si ha

$$\begin{aligned} \mathbf{Ax} = \mathbf{b} = (\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) &\Rightarrow 0 = \mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) \\ &\Rightarrow \delta\mathbf{x} = -\mathbf{A}^{-1} \delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) \end{aligned}$$

da cui

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\| \leq \|\mathbf{A}^{-1}\| \|\delta\mathbf{A}\| \|\mathbf{x} + \delta\mathbf{x}\|$$

e quindi

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x} + \delta\mathbf{x}\|} \leq \boxed{\|\mathbf{A}\| \|\mathbf{A}^{-1}\|} \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \quad (2.47)$$

Modificando opportunamente la procedura precedente, si può mostrare che nel caso in cui la perturbazione  $\delta\mathbf{A}$  sia sufficientemente piccola, più precisamente si abbia  $\|\mathbf{A}^{-1}\| \|\delta\mathbf{A}\| < 1$ , allora per la soluzione  $\mathbf{x} + \delta\mathbf{x}$  del seguente sistema perturbato

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$$

si ha la maggiorazione

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \boxed{\|\mathbf{A}\| \|\mathbf{A}^{-1}\|} \frac{\|\delta\mathbf{A}\|/\|\mathbf{A}\| + \|\delta\mathbf{b}\|/\|\mathbf{b}\|}{1 - \boxed{\|\mathbf{A}\| \|\mathbf{A}^{-1}\|} \|\delta\mathbf{A}\|/\|\mathbf{A}\|} \quad (2.48)$$

I risultati precedenti portano alla seguente definizione.

**Definizione 2.1** Se  $\|\cdot\|$  indica una norma naturale di matrice, il condizionamento di una matrice  $\mathbf{A}$  non singolare, associato a tale norma e relativo alla risoluzione di un sistema lineare, è dato dal seguente numero<sup>7</sup>

<sup>7</sup>tale definizione è stata introdotta nel caso delle matrici simmetriche definite positive da Turing nel 1948 nella norma di Frobenius, e da Von Neumann (1947) nella norma 2.

$$\mu(\mathbf{A}) := \|\mathbf{A}\| \|\mathbf{A}^{-1}\| \quad (2.49)$$

Se la norma utilizzata è del tipo  $\|\cdot\|_p$ , si utilizzerà la notazione  $\mu_p(A)$ .

Indicando con  $\epsilon_{\mathbf{A}} = \|\delta\mathbf{A}\|/\|\mathbf{A}\|$  e  $\epsilon_{\mathbf{b}} = \|\delta\mathbf{b}\|/\|\mathbf{b}\|$  gli *errori relativi* corrispondenti rispettivamente alla matrice e al termine noto, e con  $\epsilon_{\mathbf{x}} = \|\delta\mathbf{x}\|/\|\mathbf{x}\|$  l'errore relativo indotto sulla soluzione, la maggiorazione (2.48) può essere riscritta nella forma

$$\epsilon_{\mathbf{x}} \leq \mu(\mathbf{A}) \frac{\epsilon_{\mathbf{A}} + \epsilon_{\mathbf{b}}}{1 - \mu(\mathbf{A})\epsilon_{\mathbf{A}}}$$

Ricordiamo le seguenti importanti proprietà del numero di condizionamento.

1.  $\mu(\alpha\mathbf{A}) = \mu(\mathbf{A})$ , per ogni matrice  $\mathbf{A}$  e ogni scalare  $\alpha \neq 0$ .
2.  $\mu(\mathbf{A}) \geq 1$ , se la norma è naturale.
3.  $\mu_2(\mathbf{A}) = \frac{\sigma_{max}}{\sigma_{min}}$ , ove  $\sigma_{max}$  e  $\sigma_{min}$  sono rispettivamente il massimo e il minimo dei valori singolari della matrice  $\mathbf{A}$  (cfr. Appendice A).
4.  $\mu_2(\mathbf{A}) = 1$  se e solo se  $\mathbf{A} = \alpha\mathbf{Q}$ , ove  $\alpha$  è uno scalare e  $\mathbf{Q}$  è una matrice unitaria.

Una matrice è detta *bencondizionata* relativamente alla risoluzione di un sistema lineare, se il numero di condizionamento non è “troppo grande”. Le matrici unitarie sono, allora, le matrici *meglio condizionate*; questo è, in sostanza, il motivo del loro interesse nel calcolo numerico. Il significato di “troppo grande” è, in generale, dipendente dal contesto. Vi sono comunque matrici ormai classicamente ritenute malcondizionate, per le quali l'unico rimedio è la riformulazione del problema. Per le matrici, invece, “mediamente” malcondizionate un rimedio può essere la scelta di un opportuno *metodo stabile*.

Osserviamo che il numero di condizionamento è una proprietà che dipende dalla norma scelta. Comunque, due qualunque numeri di condizionamento  $\mu_{\alpha}(\mathbf{A}), \mu_{\beta}(\mathbf{A})$  sono equivalenti, nel senso che esistono due costanti  $c_1, c_2$  tali che

$$c_1\mu_{\alpha}(\mathbf{A}) \leq \mu_{\beta}(\mathbf{A}) \leq c_2\mu_{\alpha}(\mathbf{A}), \quad \mathbf{A} \in \mathbb{R}^{n \times n}$$

Ad esempio

$$\begin{aligned} \frac{1}{n}\mu_2(\mathbf{A}) &\leq \mu_1(\mathbf{A}) \leq n\mu_2(\mathbf{A}) \\ \frac{1}{n}\mu_{\infty}(\mathbf{A}) &\leq \mu_2(\mathbf{A}) \leq n\mu_{\infty}(\mathbf{A}) \\ \frac{1}{n^2}\mu_1(\mathbf{A}) &\leq \mu_{\infty}(\mathbf{A}) \leq n^2\mu_1(\mathbf{A}) \end{aligned}$$

Allo scopo di chiarire ulteriormente il significato e le proprietà del numero di condizionamento, analizzeremo alcuni esempi classici. Da tali esempi emergerà, in particolare, che il reciproco del numero di condizionamento è uno strumento più adeguato del determinante per misurare quanto un sistema assegnato  $\mathbf{Ax} = \mathbf{b}$  è vicino ad essere singolare, e quindi difficile o impossibile da risolvere.

► **Esempio 2.11** La seguente matrice

$$\mathbf{B}_n = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ 0 & 0 & 1 & \ddots & -1 \\ \vdots & \vdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix}; \quad \mathbf{B}_n^{-1} = \begin{bmatrix} 1 & 1 & 2 & \cdots & 2^{n-2} \\ 0 & 1 & 1 & \cdots & 2^{n-3} \\ 0 & 0 & 1 & \ddots & 2^{n-4} \\ \vdots & \vdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix}$$

ha determinante uguale a 1, ma  $\mu_\infty(\mathbf{B}_n) = n2^{n-1}$ . L'esempio mostra il fatto importante che una matrice può essere estremamente mal condizionata, senza che il determinante sia piccolo. D'altra parte, una matrice ben condizionata può avere un determinante piccolo. Come esempio, basta considerare la matrice

$$\mathbf{D}_n = \text{diag}(10^{-1}, \dots, 10^{-1}) \in \mathbb{R}^{n \times n}$$

per la quale  $\mu_p(\mathbf{D}_n) = 1$  e  $\det(\mathbf{D}) = 10^{-n}$ . ■

► **Esempio 2.12** Per  $\epsilon > 0$ , opportunamente piccolo, consideriamo la matrice

$$\mathbf{A} = \begin{bmatrix} \epsilon & 0 \\ 0 & \frac{1}{\epsilon} \end{bmatrix}$$

Per tale matrice si ha  $\mu_1(\mathbf{A}) = (1/\epsilon)^2$ . Il sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{b}$  vettore assegnato, è equivalente al sistema lineare che si ottiene moltiplicando la seconda equazione per  $\epsilon^2$ . La matrice dei coefficienti di tale sistema è la seguente

$$\mathbf{A}' = \begin{bmatrix} \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}$$

che corrisponde ad un opportuno scaling della matrice  $\mathbf{A}$  e per la quale si ha  $\mu_1(\mathbf{A}') = 1$ .

L'esempio mostra che il condizionamento di una matrice può essere modificato dallo *scaling*. Si pone, allora, in generale il problema della ricerca di due matrici diagonali  $\mathbf{D}_1$ ,  $\mathbf{D}_2$  in maniera da minimizzare la quantità

$$\mu(\mathbf{D}_1 \mathbf{A} \mathbf{D}_2)$$

Si tratta di una operazione di *preprocessing*. Come esempio di applicazione di tale tecnica si consideri il seguente caso particolare.

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

con  $a, b, c, d \in \mathbb{R}_+$  e  $ad \neq cd$ . Mediante le matrici

$$\mathbf{D}_1 = \begin{bmatrix} \sqrt{abcd} & 0 \\ 0 & cd \end{bmatrix}; \quad \mathbf{D}_2 = \begin{bmatrix} \sqrt{abcd} & 0 \\ 0 & ac \end{bmatrix}$$

si ottiene il seguente scaling

$$\mathbf{A}' = \mathbf{D}_1 \mathbf{A} \mathbf{D}_2 = \begin{bmatrix} a & \sqrt{abc/d} \\ \sqrt{abc/d} & a \end{bmatrix}$$

Per i corrispondenti numeri di condizionamento si ottiene

$$\mu_\infty(\mathbf{A}) = \frac{\max(a+b, c+d) \max(b+d, a+c)}{|ad-bc|}$$

$$\mu_\infty(\mathbf{A}') = \frac{ad+bc+2\sqrt{abcd}}{|ad-bc|}$$

Scegliendo ad esempio  $a = 100$ ,  $b = 0.01$ ,  $c = 99$ ,  $d = 0.01$  si ha

$$\mu_\infty(\mathbf{A}) \approx 1.99 \cdot 10^6; \quad \mu_\infty(\mathbf{A}') \approx 3.97 \cdot 10^2$$

► **Esempio 2.13** Consideriamo la matrice  $\mathbf{A}$  e il vettore  $\mathbf{b}$  definiti da

$$\mathbf{A} = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 0.8642 \\ 0.1440 \end{bmatrix}$$

La soluzione *esatta* del sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  è data da  $[2, -2]^T$ . Calcoliamo il residuo  $\mathbf{r} = \mathbf{A}\bar{\mathbf{x}} - \mathbf{b}$  per

$$\bar{\mathbf{x}} = [0.9911, -0.4870]^T \Rightarrow \mathbf{r} = [-10^{-8}, 10^{-8}]^T$$

La matrice inversa è data da

$$\mathbf{A}^{-1} = -10^8 \begin{bmatrix} 0.8648 & -0.1441 \\ -1.2969 & 0.2161 \end{bmatrix}$$

per cui

$$\mu_\infty(\mathbf{A}) \approx 3 \cdot 10^8$$

L'esempio mostra, in sostanza, che per una matrice malcondizionata il *residuo* non è una buona indicazione della precisione della soluzione di un sistema lineare. In effetti si ha

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \mu(\mathbf{A}) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$$

► **Esempio 2.14** Consideriamo la seguente matrice *simmetrica*

$$\mathbf{W} = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix}$$

Operando la perturbazione  $\mathbf{b}' = \mathbf{b} + \boldsymbol{\beta}$ , con  $\boldsymbol{\beta} = [\epsilon, -\epsilon, \epsilon, -\epsilon]^T$ , la soluzione del sistema  $\mathbf{W}\bar{\mathbf{x}} = \mathbf{b} + \boldsymbol{\beta}$  è data da

$$\bar{\mathbf{x}} = \mathbf{W}^{-1}\mathbf{b} + \mathbf{W}^{-1}\boldsymbol{\beta} = \mathbf{e} + \mathbf{W}^{-1}\boldsymbol{\beta}$$

ove  $\mathbf{e} = [1, 1, 1, 1]^T$ . La matrice  $\mathbf{W}^{-1}$  è data da

$$\mathbf{W}^{-1} = \begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 10 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix}$$

per cui  $\bar{\mathbf{x}} = [1 + 82\epsilon, 1 - 136\epsilon, 1 + 35\epsilon, 1 - 21\epsilon]^T$ . Il numero di condizionamento è dato da

$$\mu_1(\mathbf{W}) = \mu_\infty(\mathbf{W}) = 4488; \quad \mu_2(\mathbf{W}) \approx 2984$$

► **Esempio 2.15** (*Equilibrio di forze elastiche*). Consideriamo un sistema costituito da tre molle in serie fissate ai due supporti rigidi  $A$  e  $D$  e collegate tra loro nei punti  $B$  e  $C$  (cfr. Figura 2.7).

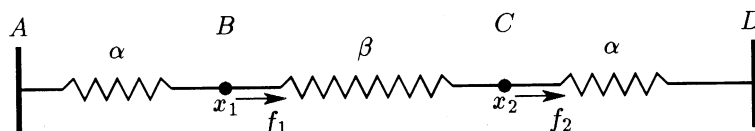


Figura 2.7: Equilibrio di forze elastiche.

Indichiamo con  $f_1$  e  $f_2$  l'intensità di due forze (stress) applicate orizzontalmente nei punti  $B$  e  $C$  e che supporremo *note a priori*. Con  $x_1$  e  $x_2$  indichiamo, poi, gli spostamenti (strain) prodotti dalle due forze. La formulazione del modello matematico è basata sul *principio di equilibrio*, in base al quale, se il sistema è in equilibrio, la somma algebrica delle componenti delle forze deve essere in ogni punto uguale a zero. Nel sistema dato, oltre alle forze esterne  $f_i$ , si deve tenere conto delle forze di reazione di tipo elastico prodotte dalle molle, che hanno una direzione opposta a quella di  $x$  positivo. Se supponiamo che le forze elastiche siano *lineari*, ossia che il materiale sia di tipo Hooke, e indichiamo con  $\alpha$  e  $\beta$  i coefficienti di stiffness, si ha per il sistema illustrato in Figura 2.7 il seguente modello matematico, rappresentato da un sistema lineare nelle incognite  $x_1, x_2$

$$\begin{cases} (\alpha + \beta)x_1 - \beta x_2 = f_1 \\ -\beta x_1 + (\alpha + \beta)x_2 = f_2 \end{cases} \quad (2.50)$$

La matrice  $\mathbf{A}$  dei coefficienti del sistema (2.50) e la sua inversa hanno la seguente rappresentazione

$$\mathbf{A} = \begin{bmatrix} \alpha + \beta & -\beta \\ -\beta & \alpha + \beta \end{bmatrix}; \quad \mathbf{A}^{-1} = \begin{bmatrix} \frac{\alpha + \beta}{\alpha(\alpha + 2\beta)} & \frac{\beta}{\alpha(\alpha + 2\beta)} \\ \frac{\beta}{\alpha(\alpha + 2\beta)} & \frac{\alpha + \beta}{\alpha(\alpha + 2\beta)} \end{bmatrix}$$

Possiamo facilmente calcolare il condizionamento della matrice nella norma 1 (o equivalentemente nella norma  $\infty$ ). Si ha

$$\mu_1(\mathbf{A}) = \frac{\alpha + 2\beta}{\alpha} = 1 + 2\frac{\beta}{\alpha}$$

da cui si vede che il condizionamento del problema dipende dai valori dei parametri  $\alpha, \beta$ , e, in particolare, dal rapporto  $\beta/\alpha$ . Quando tale rapporto è grande, ossia quando la molla centrale è molto più rigida delle adiacenti, il sistema lineare (2.50) è mal condizionato: a piccole variazioni nei parametri  $\alpha, \beta$  e nelle forze  $f_1, f_2$  possono corrispondere grandi variazioni negli spostamenti  $x_1, x_2$ . Come esempio illustrativo, si ponga

$$\alpha = 1, \quad \beta = 1000 \quad \Rightarrow \quad \mu_1(\mathbf{A}) = 1 + 2000 = 2001$$

La soluzione esatta corrispondente alle forze  $f_1 = f_2 = 1$  è data da  $x_1 = x_2 = 1$ . Operando la seguente perturbazione

$$\mathbf{A} = \begin{bmatrix} 1001 & -1000 \\ -1000 & 1001 \end{bmatrix} \Rightarrow \tilde{\mathbf{A}} = \begin{bmatrix} 1000 & -1000 \\ -1000 & 1001 \end{bmatrix}$$

con  $\|\delta\mathbf{A}\|_1/\|\mathbf{A}\|_1 \approx 5 \cdot 10^{-4}$ . La soluzione del sistema  $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \mathbf{f}$ , ove il termine noto  $\mathbf{f} = [f_1, f_2]^T$  è rimasto inalterato, è data da  $\tilde{\mathbf{x}} = [2.001, 2.000]^T$ , con un errore relativo  $\|\delta\mathbf{x}\|_1/\|\tilde{\mathbf{x}}\|_1 = 0.5001$ .

Lasciamo come esercizio l'estensione delle considerazioni precedenti al caso in cui le tre molle del sistema illustrato in Figura 2.7 possano avere coefficienti di stiffness differenti tra loro.

► **Esempio 2.16** (*Matrici di Hilbert*). Le matrici di Hilbert rappresentano un classico esempio di matrici *malcondizionate*, e che hanno interesse nelle applicazioni. In effetti, esse hanno origine nell'applicazione del metodo dei minimi quadrati, come ora ricorderemo brevemente (cfr. anche Appendice A). Se  $f(x)$  è una funzione continua sull'intervallo  $[0,1]$ , cerchiamo il polinomio  $p_{n-1}(x)$  di grado al più  $n-1$  che minimizza il seguente integrale

$$\int_0^1 (f(x) - p_{n-1}(x))^2 dx \quad (2.51)$$

Scrivendo il polinomio  $p_{n-1}(x)$  nella forma

$$p_{n-1}(x) = c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \dots + c_1x + c_0$$

si ha che i coefficienti  $c_i$  del polinomio che minimizza (2.51) sono le soluzioni del seguente sistema lineare

$$\sum_{i=0}^n c_i \int_0^1 x^{j+i} dx = \int_0^1 x^j f(x) dx, \quad j = 0, 1, \dots, n-1$$

La matrice dei coefficienti è data dalla seguente matrice

$$\mathbf{H}_n = (h_{ij}) = \left( \frac{1}{j+i-1} \right), \quad i, j = 1, \dots, n$$

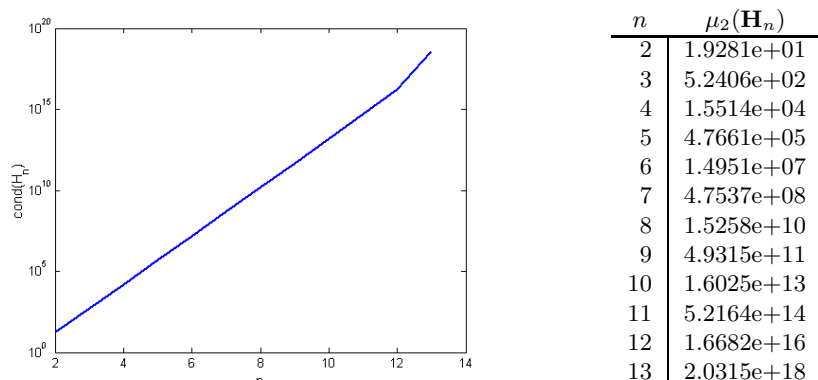


Figura 2.8: Numero di condizionamento della matrice di Hilbert  $\mathbf{H}_n$ , per  $n = 2, \dots, 13$ .

che è nota come *matrice di Hilbert* di ordine  $n$ .

È possibile verificare che la matrice *inversa esatta*  $\mathbf{H}^{-1}(n)$  ha i seguenti elementi

$$h_{ij}^{-1} := \frac{(-1)^{i+j}(n+i-1)(n+j-1)}{(i+j-1)[(i-1)(j-1)]^2(n-i)(n-j)}$$

Esaminiamo il malcondizionamento della matrice di Hilbert, considerando il caso particolare  $n=3$  e usando una aritmetica a tre cifre.

$$\mathbf{H}_3 = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \Rightarrow \tilde{\mathbf{H}}_3 = \begin{bmatrix} 1.00 & .500 & .333 \\ .500 & .333 & .250 \\ .333 & .250 & .200 \end{bmatrix}$$

ove con  $\tilde{\mathbf{H}}_3$  si è indicata la matrice  $\mathbf{H}_3$  arrotondata a tre cifre. Risolvendo, allora, il sistema  $\tilde{\mathbf{H}}_3 \mathbf{x} = \mathbf{b}$ , con  $\mathbf{b} = [1, 0, 0]^T$ , mediante il metodo di eliminazione di Gauss con pivoting parziale e con l'aritmetica a tre cifre, si perviene a risolvere il seguente sistema triangolare

$$\begin{bmatrix} 1.00 & .500 & .333 \\ 0.00 & .084 & .090 \\ 0.00 & .000 & -.004 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1.00 \\ -.333 \\ -.171 \end{bmatrix}$$

Mediante la sostituzione all'indietro, si ricava la soluzione  $\tilde{\mathbf{x}} = [11.6, -49.6, 42.7]^T$ , mentre, come si verifica facilmente, la soluzione esatta è data da  $\mathbf{x} = [9, -36, 30]^T$ . Naturalmente, la difficoltà ora rilevata può essere ridotta aumentando opportunamente il numero di cifre utilizzate. Tuttavia, come si vede in Figura 2.8, il numero di condizionamento della matrice  $\mathbf{H}_n$  aumenta rapidamente al crescere di  $n$ . In effetti, si ha la seguente stima

$$\mu_2(\mathbf{H}(n)) = O(\exp(\alpha n)), \quad \alpha \text{ costante} > 0$$

In conclusione, operando ad esempio in semplice precisione, le matrici di Hilbert diventano intrattabili già a partire da  $n \geq 6$ , e, pertanto, in tali casi il corrispondente problema di minimi quadrati deve essere opportunamente riformulato (cfr. nel paragrafo successivo il metodo **QR**, e, per una discussione più generale, l'Appendice A).

n	2	3	4	5	6	7	8	9
$\mu_2(\mathbf{V}_n)$	7.0e+01	1.1e+03	2.6e+04	7.3e+05	2.4e+07	9.5e+08	4.2e+10	2.1e+12

Tabella 2.2: Numero di condizionamento della matrice di Vandermonde  $\mathbf{V}_n$ , per  $n = 2, \dots, 9$ .

Sempre nell'ambito dell'approssimazione polinomiale, un secondo problema che può dare origine a sistemi lineari mal condizionati riguarda l'*interpolazione polinomiale*. Rinviamo al successivo Capitolo 4 per maggiori dettagli, ricordiamo brevemente il contesto. Dati  $n + 1$  coppie di punti  $(x_i, y_i)$ , per  $i = 0, 1, \dots, n$  e con  $x_i \neq x_j$ , per  $i \neq j$ , si cerca il polinomio di grado al più  $n$  che passa per tali punti, vale a dire che verifica le seguenti condizioni

$$p_n(x_i) = y_i, \quad i = 0, 1, \dots, n \quad \text{ove} \quad p_n(x) = \sum_{k=0}^n c_k x^k$$

Si vede facilmente che il problema precedente è equivalente alla risoluzione di un sistema lineare nelle  $n + 1$  incognite  $\mathbf{c} = [c_0, c_1, \dots, c_n]^T$

$$\mathbf{V}_n \mathbf{c} = \mathbf{y}, \quad \mathbf{y} = [y_0, y_1, \dots, y_n]^T$$

ove la matrice dei coefficienti  $\mathbf{V}_n$ , di dimensione  $(n + 1) \times (n + 1)$  ha la seguente forma

$$\mathbf{V}_n = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \quad (2.52)$$

ed è chiamata *matrice di Vandermonde*<sup>8</sup>. Consideriamo il condizionamento di tale matrice nel caso particolare in cui i punti  $x_i$  sono *equidistanti*, ossia  $x_k = x_0 + kh$ , per  $k = 0, 1, \dots, n$ . Nella Tabella 2.2 sono riportati i numeri di condizionamento delle matrici  $\mathbf{V}_n$ , quando  $x_0 = 1$  e  $h = 1$ . Come si vede, in questo caso il problema dell'interpolazione è mal condizionato già per valori piccoli di  $n$ . Nel successivo Capitolo 4 vedremo altre difficoltà inerenti alla scelta nell'interpolazione di punti equidistanti e riguardanti la convergenza per  $n \rightarrow \infty$ .

► **Esempio 2.17** *Discretizzazione di un problema ai limiti*. Nel successivo Capitolo 7 considereremo più in dettaglio la risoluzione numerica di problemi differenziali del seguente tipo

$$\begin{cases} \frac{d^2 y}{dx^2} = -f(x) & 0 \leq x \leq 1 \\ y(0) = y(1) = 0 & \text{condizioni ai limiti} \end{cases} \quad (2.53)$$

che sono alla base della modellizzazione matematica di numerosi fenomeni, quali ad esempio la distribuzione della temperatura e la diffusione di sostanze. Un modo semplice per approssimare la soluzione del problema (2.53) consiste nel suddividere l'intervallo  $[0, 1]$  in  $n$  intervalli di uguale lunghezza  $h = 1/n$  e nell'approssimare la soluzione  $y(x)$  nei nodi

<sup>8</sup>Alexandre Théophile Vandermonde (1735-1796). Fu tra i primi a dare una organica esposizione della teoria dei determinanti, della quale può essere considerato tra i fondatori.



$x_k = kh, k = 0, 1, \dots, n$  mediante la discretizzazione della derivata seconda  $d^2y/dx^2$  con una differenza finita centrale, ossia

$$\frac{d^2y(x_k)}{dx^2} \approx \frac{y_{k-1} - 2y_k + y_{k+1}}{h^2}$$

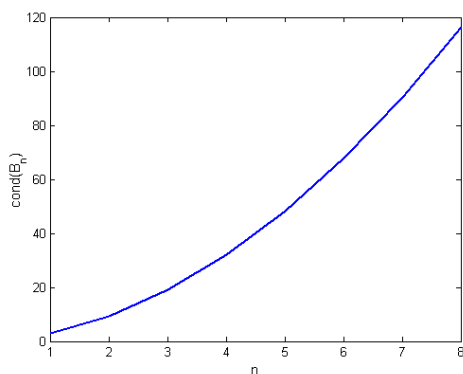
Indicando con  $\bar{\mathbf{y}} = [\bar{y}_0, \bar{y}_1, \dots, \bar{y}_n]^T$  la soluzione approssimata, al problema (2.53) corrisponde il seguente *problema discreto*

$$\begin{cases} \frac{\bar{y}_{k-1} - 2\bar{y}_k + \bar{y}_{k+1}}{h^2} = -f(x_k) & k = 1, 2, \dots, n-1 \\ \bar{y}_0 = \bar{y}_1 = 0 & \text{condizioni ai limiti} \end{cases} \quad (2.54)$$

che corrisponde alla risoluzione del seguente sistema lineare

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \vdots \\ \bar{y}_{n-1} \end{bmatrix} = h^2 \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{n-1}) \end{bmatrix} \quad (2.55)$$

Nella Tabella 2.9 sono riportati i numeri di condizionamento della matrice  $\mathbf{B}_n$  dei coefficienti del sistema (2.55). Si vede, quindi, che il condizionamento del sistema lineare (2.55) peggiora



$n$	$\mu_2(\mathbf{B}_n)$
4	9.4721
6	19.1957
8	32.1634
10	48.3742
12	67.8274
14	90.5231
16	116.4612

Figura 2.9: Numero di condizionamento della matrice tridiagonale  $\mathbf{B}_n$  corrispondente alla risoluzione di un problema ai limiti mediante il metodo delle differenze finite. Si può dimostrare che  $\mu_2(\mathbf{B}_n)$  cresce come  $1/2n^2$ .

al crescere del numero di suddivisioni  $n$ , o equivalentemente al tendere a zero del passo di discretizzazione  $h$ . D'altra parte, come vedremo nel successivo Capitolo 7, per  $h \rightarrow 0$  la soluzione discreta converge alla soluzione continua. In altre parole, e in forma schematica, *più il problema discreto è vicino al problema continuo e più crescono le difficoltà numeriche per la sua soluzione.*

### 2.2.1 Stabilità degli algoritmi

Come abbiamo già osservato, l'errore contenuto nella *soluzione numerica* di un sistema lineare può essere, per comodità, scomposto in due componenti: il cosiddetto *errore inerente*, ossia l'errore legato al condizionamento del problema e indipendente dall'algoritmo utilizzato, e l'*errore algoritmico* legato alla stabilità del particolare algoritmo utilizzato. In questo paragrafo ricorderemo i principali risultati che si riferiscono all'errore algoritmico, cercando, infine, di ottenere, attraverso i risultati ottenuti sul condizionamento nel paragrafo precedente, un'indicazione, ossia una stima, sull'errore totale presente nella soluzione numerica di un sistema lineare.

Per analizzare l'errore algoritmico è particolarmente utile la cosiddetta tecnica dell'*analisi all'indietro* (backward analysis)<sup>9</sup>, nella quale la soluzione calcolata  $\tilde{\mathbf{x}}$  viene considerata come la soluzione *esatta* di un opportuno problema perturbato del tipo

$$(\mathbf{A} + \delta\mathbf{A})\tilde{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$$

ove i termini  $\delta\mathbf{A}$  e  $\delta\mathbf{b}$  sono legati agli errori commessi durante i calcoli e rappresentano, quindi, il contributo del particolare algoritmo utilizzato. La “grandezza” di tali termini misura, in sostanza, la stabilità dell'algoritmo; vale a dire, in maniera per ora formale, un algoritmo risulta più stabile di un altro se i corrispondenti termini  $\delta\mathbf{A}$  e  $\delta\mathbf{b}$  sono più “piccoli”. L'interesse della tecnica di backward analysis consiste nel fatto che, una volta nota una stima dei termini  $\delta\mathbf{A}$  e  $\delta\mathbf{b}$ , per avere una maggiorazione dell'errore  $\delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}$  è sufficiente applicare la maggiorazione (2.48) ottenuta nel paragrafo precedente. Sottolineiamo, comunque, il significato ora diverso dei termini  $\delta\mathbf{A}$  e  $\delta\mathbf{b}$ ; in effetti, mentre nel paragrafo precedenti essi rappresentano perturbazioni reali nei dati del problema, ora essi rappresentano delle perturbazioni virtuali, dovute agli errori commessi durante i calcoli<sup>10</sup>.

Naturalmente, la difficoltà principale nell'applicazione dell'idea consiste nell'ottenere delle indicazioni significative, ossia “non troppo pessimistiche” dei termini  $\delta\mathbf{A}$  e  $\delta\mathbf{b}$ . Trattandosi di dimostrazioni molto tecniche, ci limiteremo all'enunciato e alla discussione di alcuni risultati, rinviando alla bibliografia per maggiori dettagli.

**Teorema 2.1** *Sia  $\mathbf{A}$  una matrice di ordine  $n$  i cui elementi sono numeri macchina, ossia  $\text{fl}(\mathbf{A}) \equiv \mathbf{A}$ , e siano  $\tilde{\mathbf{L}}, \tilde{\mathbf{U}}$  le matrici della fattorizzazione  $\mathbf{LU}$  ottenute*

<sup>9</sup>Introdotta da Givens (1954) e da Wilkinson (1961). *The main object of a priori error analysis is to expose the potential instabilities, if any, of an algorithm so that hopefully from the insight thus obtained one might be led to improved algorithms... Practical error bounds should usually be determined by some form of a posteriori error analysis since this take full advantage of the statistical distribution of rounding errors and of any special features, such as sparseness, in the matrix,* Wilkinson.

<sup>10</sup>Come illustrazione dell'idea, si ricordi la definizione dell'operazione macchina  $\text{fl}(x + y) = (x + y) + \epsilon(x + y) = (x + \epsilon x) + (y + \epsilon y)$ , con  $|\epsilon| \leq \text{eps}$ : *la somma macchina è interpretabile come la somma esatta dei numeri perturbati  $x + \epsilon x$ ,  $y + \epsilon y$ .*

numericamente con il metodo di Gauss. Si ha allora<sup>11</sup>

$$\tilde{\mathbf{L}}\tilde{\mathbf{U}} = \mathbf{A} + \mathbf{E}$$

con

$$|\mathbf{E}| \leq 2n \text{ eps} (|\mathbf{A}| + |\tilde{\mathbf{L}}| |\tilde{\mathbf{U}}|) + O(\text{eps}^2)$$

ove  $\text{eps}$  è la precisione macchina utilizzata e  $O(\text{eps}^2)$  indica una matrice i cui elementi sono funzioni di potenze di  $\text{eps}$  di grado maggiore o uguale al secondo.

**Teorema 2.2** Sia  $\mathbf{A}$  una matrice triangolare inferiore di ordine  $n$  e  $\mathbf{b}$  un vettore di ordine  $n$ , con  $\text{fl}(\mathbf{A}) = \mathbf{A}$  e  $\text{fl}(\mathbf{b}) = \mathbf{b}$ . Indicata con  $\tilde{\mathbf{x}}$  la soluzione numerica del sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  ottenuta mediante l'algoritmo della sostituzione in avanti, esiste una matrice  $\mathbf{E}$  di ordine  $n$  tale che

$$(\mathbf{A} + \mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}, \quad \text{con} \quad |\mathbf{E}| \leq n \text{ eps} |\mathbf{A}| + O(\text{eps}^2)$$

Un analogo risultato si ha quando la matrice  $\mathbf{A}$  è triangolare superiore. Combinando, allora, i risultati precedenti si ottiene la seguente valutazione degli errori relativi al metodo di Gauss nella risoluzione del sistema lineare generico  $\mathbf{A}\mathbf{x} = \mathbf{b}$ .

**Teorema 2.3** Siano  $\mathbf{A}$  e  $\mathbf{b}$  una matrice e un vettore di ordine  $n$ , con elementi numeri macchina, e sia  $\tilde{\mathbf{x}}$  la soluzione del sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  calcolata nel seguente modo

1. calcolo delle matrici  $\tilde{\mathbf{L}}, \tilde{\mathbf{U}}$ ;
2. calcolo del vettore  $\tilde{\mathbf{y}}$ , soluzione numerica del sistema  $\tilde{\mathbf{L}}\mathbf{y} = \mathbf{b}$ ;
3. calcolo del vettore  $\tilde{\mathbf{x}}$ , soluzione numerica del sistema  $\tilde{\mathbf{U}}\mathbf{x} = \tilde{\mathbf{y}}$ ;

Si ha allora

$$(\mathbf{A} + \delta\mathbf{A})\tilde{\mathbf{x}} = \mathbf{b} \quad \text{con} \quad |\delta\mathbf{A}| \leq 4n \text{ eps} (|\mathbf{A}| + |\tilde{\mathbf{L}}| |\tilde{\mathbf{U}}|) + O(\text{eps}^2) \quad (2.56)$$

Un aspetto interessante messo in evidenza dal risultato (2.56) è il fatto che l'errore numerico può essere tanto più elevato quanto più sono grandi gli elementi delle matrici  $|\tilde{\mathbf{L}}|, |\tilde{\mathbf{U}}|$ . L'applicazione delle tecniche del pivoting ha, in effetti, come effetto la riduzione della grandezza di tali elementi, e questo è in sostanza il motivo della loro maggiore stabilità. In questo senso, il *pivoting totale* fornisce l'algoritmo più stabile, ma, come abbiamo visto esso richiede un numero superiore di confronti, e, quindi, il pivoting parziale è in generale un compromesso più opportuno.

Terminiamo, osservando che la stima (2.56) fornisce solo una limitazione superiore all'errore; in altre parole, dal momento che gli errori di segno contrario possono

---

<sup>11</sup>Con  $|\mathbf{A}|$  si intende la matrice di elementi  $|a_{ij}|$  e la relazione  $\mathbf{A} \leq \mathbf{B}$  significa  $a_{ij} \leq b_{ij}$ , per  $i, j = 1, \dots, n$ .

eliminarsi o ridursi in grandezza, l'errore effettivo può essere anche notevolmente inferiore a quello indicato dalla formula. In effetti, il metodo di eliminazione pivotale di Gauss, *quando la matrice è ben condizionata*, è un procedimento stabile. Quando, invece, la matrice non è ben condizionata, possono essere più opportuni i metodi che utilizzano per la fattorizzazione della matrice le trasformazioni ortogonali, le quali, come abbiamo già osservato, hanno la proprietà di non peggiorare il condizionamento della matrice di partenza. Tali metodi saranno introdotti nel prossimo paragrafo. Rinviando, invece, all'Appendice A per l'analisi degli algoritmi basati sulla *decomposizione in valori singolari* (SVD) della matrice.

### 2.2.2 Fattorizzazione $\mathbf{A} = \mathbf{QR}$

Data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , si cerca una matrice ortogonale  $\mathbf{Q}$  e una matrice triangolare  $\mathbf{R}$  tale che  $\mathbf{A} = \mathbf{QR}$ . Tale decomposizione può essere ottenuta con differenti tecniche, in particolare mediante l'utilizzo delle matrici elementari di Householder o di Givens, oppure mediante il metodo di ortogonalizzazione di Gram-Schmidt.

Incominciamo con il seguente risultato di esistenza, di cui forniremo nel seguito una dimostrazione costruttiva.

**Teorema 2.4 (Decomposizione  $\mathbf{QR}$ )** *Data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , con  $m \geq n$  e di rango  $n$ , esiste un'unica matrice  $\mathbf{Q} \in \mathbb{R}^{m \times n}$  con*

$$\mathbf{Q}^T \mathbf{Q} = \mathbf{D}, \quad \mathbf{D} = \text{diag}(d_1, \dots, d_n), \quad d_k > 0, \quad k = 1, \dots, n$$

e una matrice triangolare superiore  $\mathbf{R}$ , con  $r_{kk} = 1$ ,  $k = 1, \dots, n$ , tali che

$$\boxed{\mathbf{A} = \mathbf{QR}} \tag{2.57}$$

La decomposizione (2.57) viene anche chiamata *decomposizione  $\mathbf{QR}$  non normalizzata*. Come *decomposizione normalizzata*, si intende una decomposizione nella quale  $\mathbf{Q}$  è una matrice  $m \times m$  ortogonale (quindi  $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ ) e  $\mathbf{R}$  è una matrice  $m \times n$  della forma

$$\mathbf{R} = \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix}$$

ove  $\mathbf{T}$  è una matrice  $n \times n$  triangolare superiore e  $\mathbf{0}$  è una matrice  $(m - n) \times n$  di zeri.

Mentre la decomposizione (2.57) può essere ottenuta, come vedremo nel seguito, mediante un procedimento di ortogonalizzazione delle colonne di  $\mathbf{A}$ , la decomposizione normalizzata può essere ottenuta mediante successive moltiplicazioni di matrici di Householder (o di Givens) e la matrice  $\mathbf{Q}$  così ottenuta contiene sia una base ortogonale per lo spazio generato dalle colonne di  $\mathbf{A}$  che una base per lo spazio perpendicolare alle colonne di  $\mathbf{A}$ .

La decomposizione (2.57) può essere utilizzata per risolvere il problema dei minimi quadrati, e quindi in particolare per risolvere i sistemi lineari, nel seguente modo. Ricordiamo (cfr. Appendice A) che la soluzione del seguente problema di minimo

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2, \quad \mathbf{b} \in \mathbb{R}^m \quad (2.58)$$

quando la matrice  $\mathbf{A}$  ha rango  $n$ , è la soluzione del seguente sistema lineare (sistema delle equazioni normali)

$$\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b}$$

Utilizzando, allora, la decomposizione (2.57), si ha

$$\mathbf{A}^T (\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{R}^T \mathbf{Q}^T (\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{R}^T (\mathbf{Q}^T \mathbf{b} - \mathbf{Q}^T \mathbf{Q} \mathbf{R} \mathbf{x}) = 0 \Rightarrow \begin{cases} \mathbf{D}\mathbf{y} = \mathbf{Q}^T \mathbf{b} \\ \mathbf{R}\mathbf{x} = \mathbf{y} \end{cases}$$

Sottolineiamo che l'interesse numerico della procedura precedente consiste nel fatto che procedendo in tale modo il sistema delle equazioni normali viene risolto, *senza costruire esplicitamente la matrice  $\mathbf{A}^T \mathbf{A}$* . Ricordiamo, infatti, che il numero di condizionamento della matrice  $\mathbf{A}^T \mathbf{A}$  è il quadrato di quello della matrice  $\mathbf{A}$ .

Lasciamo come esercizio l'estensione delle considerazioni precedenti al caso in cui sia disponibile una decomposizione  $\mathbf{QR}$  in forma normalizzata e al caso in cui il rango della matrice  $\mathbf{A}$  sia minore di  $n$ .

### Metodo modificato di Gram-Schmidt

Sia  $\mathbf{A}$  una matrice  $\in \mathbb{R}^{m \times n}$ , con  $m \geq n$  e di rango  $n$ . Le colonne di  $\mathbf{A}$  sono, quindi, vettori linearmente indipendenti. A partire da tali vettori, si può costruire un insieme di vettori ortogonali utilizzando il procedimento di Gram-Schmidt (cfr. Appendice A). Dal punto di vista numerico, più precisamente per motivi di stabilità, è, tuttavia, preferibile utilizzare la seguente variante del procedimento classico.

Si calcola una successione di matrici  $\mathbf{A}^{(1)} := \mathbf{A}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(n+1)} =: \mathbf{Q}$ . Al passo generico  $k$ -mo la matrice  $\mathbf{A}^{(k)}$  ha la seguente forma

$$\mathbf{A}^{(k)} = [\mathbf{q}_1, \dots, \mathbf{q}_{k-1}, \mathbf{a}_k^{(k)}, \dots, \mathbf{a}_n^{(k)}]$$

ove i vettori  $\mathbf{a}_k^{(k)}, \dots, \mathbf{a}_n^{(k)}$  sono per ipotesi ortogonali ai vettori  $\mathbf{q}_1, \dots, \mathbf{q}_{k-1}$ . Descriviamo, pertanto, il passaggio da  $\mathbf{A}^{(k)}$  a  $\mathbf{A}^{(k+1)}$ .

In tale passaggio si assume  $\mathbf{q}_k = \mathbf{a}_k^{(k)}$  e si trasformano i vettori  $\mathbf{a}_{k+1}^{(k)}, \dots, \mathbf{a}_n^{(k)}$  in maniera da ottenere dei vettori  $\mathbf{a}_j^{(k+1)}$ ,  $j = k+1, \dots, n$  ortogonali al vettore  $\mathbf{q}_k$ . Il risultato è ottenuto mediante le seguenti formule

$$\begin{aligned} \mathbf{q}_k &= \mathbf{a}_k^{(k)}, \quad d_k = \mathbf{q}_k^T \mathbf{q}_k, \quad r_{kk} = 1 \\ \mathbf{a}_j^{(k+1)} &= \mathbf{a}_j^{(k)} - r_{kj} \mathbf{q}_k, \quad r_{kj} = \frac{\mathbf{q}_k^T \mathbf{a}_j^{(k)}}{d_k}, \quad j = k+1, \dots, n \end{aligned}$$

Con riferimento alla risoluzione del problema dei minimi quadrati (2.58), si trasforma il vettore  $\mathbf{b}$  allo stesso modo

$$\mathbf{b}^{(k+1)} = \mathbf{b}^{(k)} - y_k \mathbf{q}_k, \quad y_k = \frac{\mathbf{q}_k^T \mathbf{b}^{(k)}}{d_k}$$

Il vettore  $\mathbf{b}^{(n+1)}$  risulta essere ortogonale allo spazio  $\mathcal{R}(\mathbf{A})$  e quindi coincide con il vettore residuo  $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ . Pertanto, dopo  $n$  passi si ottengono i seguenti risultati

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n], \quad \mathbf{R} = (r_{kj}), \quad \mathbf{y} = [y_1, \dots, y_n]^T$$

con

$$\mathbf{Q}^T \mathbf{Q} = \text{diag}(d_k), \quad \mathbf{A} = \mathbf{Q}\mathbf{R}, \quad \mathbf{b} = \mathbf{Q}\mathbf{y} + \mathbf{r}$$

Il numero delle operazioni è approssimativamente dato da

$$2m \sum_{k=1}^n (n - k + 1) = mn(n + 1)$$

**Algoritmo 2.16** (Gram-Schmidt modificato) *Data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  con rango  $n$  ed un vettore  $\mathbf{b} \in \mathbb{R}^m$ , il seguente algoritmo calcola la fattorizzazione  $\mathbf{A} = \mathbf{Q}\mathbf{R}$ , con  $\mathbf{Q}^T \mathbf{Q} = \mathbf{D}$  ed il vettore  $\mathbf{y} = \mathbf{D}^{-1} \mathbf{Q}^T \mathbf{b}$ . La soluzione del problema:  $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$  è ottenuta risolvendo il sistema triangolare  $\mathbf{R}\mathbf{x} = \mathbf{y}$ . La matrice  $\mathbf{A}$  è sostituita dalla matrice  $\mathbf{Q}$ .*

```

For  $k = 1, \dots, n$ 
   $d_k := \sum_{i=1}^m a_{ik}^2$ 
   $y_k := (\sum_{i=1}^m a_{ik} b_i) / d_k$ ;  $r_{kk} := 1$ 
  For  $j = k + 1, \dots, n$ 
     $r_{kj} := (\sum_{i=1}^m a_{ik} a_{ij}) / d_k$ 
    For  $i = 1, \dots, m$ 
       $a_{ij} := a_{ij} - a_{ik} r_{kj}$ 
    end  $i$ 
  end  $j$ 
  For  $i = 1, \dots, m$ 
     $b_i = b_i - a_{ik} y_k$ 
  end  $i$ 
end  $k$ 

```

Come esemplificazione, riportiamo l'implementazione dell'algoritmo precedente in MATLAB, che mette in evidenza la vettorialità dell'algoritmo.

```

for k=1:n
  d(k)=norm(a(:,k),2)^2;
  y(k)=a(:,k)'*b/d(k);
  r(k,k)=1
  for j=k+1:n
    r(k,j)=a(:,k)'*a(:,j)/d(k);

```

```

    a(:,j)=a(:,j)-a(:,k)*r(k,j);
end
b=b-a(:,k)*y(k);
end
x=r\y' %soluzione del problema dei minimi quadrati

```

► **Esempio 2.18** Come semplice esempio illustrativo, consideriamo il seguente problema dei minimi quadrati

$$\min_{\mathbf{x} \in \mathbb{R}^2} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \quad \text{con} \quad \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ -5 \end{bmatrix}$$

L'algoritmo precedente fornisce i seguenti risultati

$$\mathbf{Q} = \begin{bmatrix} 1.000 & 0.500 \\ 1.000 & -0.500 \\ 0. & 1.000 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1.000 & 0.500 \\ 0. & 1.000 \end{bmatrix}, \quad \mathbf{y}^T = \begin{bmatrix} 0.500 \\ -3.000 \end{bmatrix}$$

$$\mathbf{x}^* = \begin{bmatrix} 2.000 \\ -3.000 \end{bmatrix}, \quad \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}^* = \begin{bmatrix} 2.000 \\ -2.000 \\ -2.000 \end{bmatrix}$$

ove  $\mathbf{x}^*$  è il punto di minimo. Il sistema delle equazioni normali in questo caso è dato da

$$\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b}, \quad \mathbf{A}^T \mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{A}^T \mathbf{b} = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$$

che ammette, come si verifica facilmente, la medesima soluzione trovata con il metodo precedente.

Terminiamo, osservando che mentre il numero di condizionamento  $\mu_2(\mathbf{A})$ , definito come il rapporto tra il massimo valore singolare  $\sigma_{max}$  di  $\mathbf{A}$  e il minimo valore singolare  $\sigma_{min}$ , è dato da 1.7321, il numero di condizionamento della matrice  $\mathbf{A}^T \mathbf{A}$  è dato da 3.000, ossia il quadrato del precedente. In effetti, ricordiamo che  $\mu_2(\mathbf{A}\mathbf{A}^T) = \mu_2(\mathbf{A}) \cdot \mu_2(\mathbf{A}^T)$ , e questo è il motivo dell'interesse della fattorizzazione **QR** quando la matrice  $\mathbf{A}$  è malcondizionata.

Le proprietà di stabilità che caratterizzano il metodo **QR** sono ulteriormente evidenziate dalla seguente semplice esemplificazione. Consideriamo la risoluzione del seguente sistema lineare

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{con} \quad \mathbf{A} = \begin{bmatrix} 600 & 800 \\ 30001 & 40002 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 200 \\ 10001 \end{bmatrix}$$

che risulta estremamente malcondizionato, in quanto  $\mu_2(\mathbf{A}) \approx 2.25 \cdot 10^6$ . Applicando il metodo di eliminazione, con la tecnica del pivot parziale e il metodo **QR**, si ottengono i seguenti risultati (in aritmetica con precisione macchina  $\text{eps} \approx 2.22 \cdot 10^{-16}$ )

$$\mathbf{x}_{\text{pivot}} = \begin{bmatrix} -0.99999999999700 \\ 0.99999999999775 \end{bmatrix}, \quad \mathbf{x}_{\text{QR}} = \begin{bmatrix} -1.00000000000080 \\ 1.00000000000060 \end{bmatrix}$$

### Metodo di Householder

Nel metodo di Householder una matrice  $\mathbf{A}$  di ordine  $m \times n$  viene trasformata in una matrice triangolare superiore mediante l'applicazione successiva di trasformazioni elementari di Householder (cfr. per la definizione e gli algoritmi corrispondenti l'Appendice A). Cerchiamo di spiegare l'essenza del metodo mediante un esempio schematico. Per  $m = 6, n = 5$ , supponiamo di aver già calcolato due matrici di trasformazione  $\mathbf{H}_1, \mathbf{H}_2$  tali che

$$\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & \boxed{x} & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \end{bmatrix}$$

e cerchiamo una matrice di Householder  $\overline{\mathbf{H}}_3$  di ordine  $4 \times 4$  tale che

$$\overline{\mathbf{H}}_3 \begin{bmatrix} x \\ x \\ x \\ x \end{bmatrix} = \begin{bmatrix} x \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Poniamo quindi  $\mathbf{H}_3 = \mathbf{diag}(\mathbf{I}_2, \overline{\mathbf{H}}_3)$ . Si ha allora

$$\mathbf{H}_3\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & x & x \end{bmatrix}$$

Alla fine  $\mathbf{H}_n\mathbf{H}_{n-1} \cdots \mathbf{H}_1\mathbf{A} = \mathbf{R}$  è una matrice triangolare superiore e ponendo  $\mathbf{Q} = \mathbf{H}_1 \cdots \mathbf{H}_n$  si ottiene  $\mathbf{A} = \mathbf{QR}$ .

L'algoritmo richiede  $n^2(m-n/3)$  flops. Per la risoluzione del problema dei minimi quadrati non occorre formare esplicitamente la matrice  $\mathbf{Q}$ , ma basta modificare successivamente il vettore  $\mathbf{b}$  mediante le moltiplicazioni  $\mathbf{H}_k\mathbf{b}$ .

► **Esempio 2.19** Consideriamo la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 4 & 5 \\ 0 & 4 \end{bmatrix}$$

Il primo passo della fattorizzazione  $\mathbf{QR}$  costruisce la matrice di Householder  $\mathbf{H}_1$  che annulla il secondo e il terzo elemento della prima colonna



$$\mathbf{H}_1 = \frac{1}{5} \begin{bmatrix} -3 & -4 & 0 \\ -4 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix} \Rightarrow \mathbf{A}_1 = \mathbf{H}_1 \mathbf{A} = \begin{bmatrix} -5 & -4 \\ 0 & 3 \\ 0 & 4 \end{bmatrix}$$

Nel secondo passo, si costruisce la matrice di Householder  $\mathbf{H}_2$  che annulla il terzo elemento nella seconda colonna della matrice  $\mathbf{A}_1$ , ma lascia invariato il primo elemento

$$\mathbf{H}_2 = \frac{1}{5} \begin{bmatrix} 5 & 0 & 0 \\ 0 & -3 & -4 \\ 0 & -4 & 3 \end{bmatrix} \Rightarrow \mathbf{R} = \mathbf{A}_2 = \mathbf{H}_1 \mathbf{A} = \begin{bmatrix} -5 & -4 \\ 0 & -5 \\ 0 & 0 \end{bmatrix}$$

Si ha, quindi,  $\mathbf{R} = \mathbf{A}_2$  e  $\mathbf{A} = \mathbf{QR}$ , ove  $\mathbf{Q} = \mathbf{H}_1 \mathbf{H}_2$

$$\mathbf{A} = \left( \frac{1}{25} \begin{bmatrix} -15 & 12 & 16 \\ -20 & -9 & -12 \\ 0 & -20 & 15 \end{bmatrix} \right) \begin{bmatrix} -5 & -4 \\ 0 & -5 \\ 0 & 0 \end{bmatrix}$$

### Metodo di Givens

Nel metodo di Givens si utilizzano per la riduzione della matrice  $\mathbf{A}$  alla forma triangolare le matrici di *rotazione* di Givens (cfr. Appendice A). Le trasformazioni di Givens permettono rispetto alle trasformazioni di Householder una maggiore *selettività*. Il loro uso è quindi interessante per trasformare matrici con particolari strutture sparse.

**Algoritmo 2.17** (Metodo di Givens) *Data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , il seguente algoritmo calcola la fattorizzazione  $\mathbf{A} = \mathbf{QR}$  con  $\mathbf{Q}$  ortonormale e  $\mathbf{R}$  triangolare. La matrice  $\mathbf{A}$  è sostituita dalla matrice  $\mathbf{Q}^T \mathbf{A} = \mathbf{R}$ .*

```

For  $q = 2, \dots, m$ 
  For  $p = 1, 2, \dots, \min(q-1, n)$ 
    trovare  $c = \cos\theta$ ,  $s = \sin\theta$  tali che
       $-sa_{pp} + ca_{qp} = 0$ 
     $\mathbf{A} := \mathbf{J}(p, q, \theta)\mathbf{A}$ 
     $\mathbf{b} := \mathbf{J}(p, q, \theta)\mathbf{b}$ 
  end  $p$ 
end  $q$ 

```

L'algoritmo richiede  $2n^2(m - n/3)$  flops, circa il doppio di quelle richieste dal metodo di Householder. Esiste, comunque, una versione del metodo, chiamata *metodo rapido di Givens*, che richiede un numero di operazioni paragonabile a quello delle trasformate di Householder. Per i dettagli di questa variante rinviamo ad esempio a Golub e Van Loan [69].

◆ **Esercizio 2.14** *Sia  $\mathbf{Q}$  una matrice ortogonale. Dimostrare che  $\mu_2(\mathbf{QA}) = \mu_2(\mathbf{A})$ .*

◆ **Esercizio 2.15** Analizzare il condizionamento delle seguenti matrici

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 2 & -1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 8 \end{bmatrix}, \quad \mathbf{A}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

◆ **Esercizio 2.16** La successione di Fibonacci è generata dalla seguente equazione alle differenze

$$f_0 = 0, \quad f_1 = 1, \quad f_j = f_{j-1} + f_{j-2}, \quad j > 1$$

Mostrare che  $f_n f_{n+2} - f_{n+1}^2 = (-1)^{n+1}$ ,  $n = 0, 1, \dots$ . Trovare allora l'unica soluzione del sistema

$$\begin{cases} f_n x_1 + f_{n+1} x_2 = f_{n+2} \\ f_{n+1} x_1 + f_{n+2} x_2 = f_{n+3} \end{cases}$$

Studiare il malcondizionamento di tale sistema al crescere di  $n$ . In particolare, per  $n = 10$  esaminare la variazione della soluzione in corrispondenza alla perturbazione nella seconda equazione  $\tilde{f}_{n+2} = f_{n+2} + \epsilon$ , con  $\epsilon = 0.018$  e rispettivamente  $\epsilon = 0.02$ .

◆ **Esercizio 2.17** Se  $\mathbf{A}$ ,  $\mathbf{B}$  sono due matrici simmetriche definite positive, allora

$$\mu_2(\mathbf{A} + \mathbf{B}) \leq \max[\mu_2(\mathbf{A}), \mu_2(\mathbf{B})]$$

◆ **Esercizio 2.18** Considerare il sistema lineare  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , con

$$\mathbf{A} = \begin{bmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 4.1 \\ 9.7 \end{bmatrix}$$

Calcolare la soluzione esatta e la soluzione perturbata corrispondente a  $\mathbf{b}' = [4.11, 9.70]^T$ . Analizzare quindi il condizionamento della matrice.

◆ **Esercizio 2.19** La matrice  $\mathbf{C} = (\mathbf{A}^T \mathbf{A})^{-1}$ , ove  $\text{rank}(\mathbf{A}) = n$ , è nota in statistica come matrice varianza-covarianza. Sia  $\mathbf{A} = \mathbf{Q}\mathbf{R}$  la decomposizione ortogonale di  $\mathbf{A}$ .

1. Mostrare che  $\mathbf{C} = (\mathbf{R}^T \mathbf{R})^{-1}$ .
2. Dare un algoritmo per calcolare  $c_{11}, \dots, c_{nn}$  che richieda  $n^3/6$  flops.
3. Mostrare che se  $\mathbf{R} = \begin{bmatrix} \alpha & \mathbf{v}^T \\ 0 & \mathbf{S} \end{bmatrix}$  e  $\mathbf{C}_1 = (\mathbf{S}^T \mathbf{S})^{-1}$ , allora

$$\mathbf{C} = (\mathbf{R}^T \mathbf{R})^{-1} = \begin{bmatrix} (1 + \mathbf{v}^T \mathbf{C}_1 \mathbf{v})/\alpha^2 & -\mathbf{v}^T \mathbf{C}_1/\alpha \\ -\mathbf{C}_1 \mathbf{v}/\alpha & \mathbf{C}_1 \end{bmatrix}$$

4. Usando il risultato precedente dare un algoritmo che sostituisce  $\mathbf{R}$  con la parte triangolare superiore di  $\mathbf{C}$ , con  $n^3/3$  flops.

## 2.3 Metodi iterativi

Per le matrici di *ordine elevato* e di tipo *sparso*, una importante alternativa ai metodi diretti è costituita dai *metodi iterativi*, la cui idea di base è la costruzione, a partire da una stima iniziale  $\mathbf{x}^{(0)}$ , di una successione convergente di approssimanti  $\{\mathbf{x}^{(k)}\}$ ,  $k = 1, 2, \dots$ , ove ciascun vettore  $\mathbf{x}^{(k)}$  è la soluzione di problemi computazionalmente *più semplici*. Come vedremo successivamente nel Capitolo 5, tale idea costituisce lo strumento fondamentale per la risoluzione dei sistemi non lineari e dei problemi di ottimizzazione. I risultati che analizzeremo in questo paragrafo potrebbero, pertanto, essere inquadrati in tale contesto più generale, ma l'adattamento della teoria al caso particolare dei sistemi lineari permette di ottenere risultati più significativi.

### 2.3.1 Metodi di Jacobi, Gauss-Seidel, rilassamento

I metodi iterativi corrispondono al seguente schema generale. Data una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , non singolare, si considera una decomposizione (*splitting*) della matrice  $\mathbf{A}$  del seguente tipo

$$\mathbf{A} = \mathbf{M} - \mathbf{N}, \quad \det(\mathbf{M}) \neq 0 \quad (2.59)$$

Per ogni vettore  $\mathbf{b} \in \mathbb{R}^n$  si ha allora

$$\mathbf{A}\mathbf{x} = \mathbf{b} \iff \mathbf{M}\mathbf{x} = \mathbf{N}\mathbf{x} + \mathbf{b}$$

da cui il seguente *procedimento iterativo*

$$\boxed{\mathbf{M}\mathbf{x}^{(k+1)} = \mathbf{N}\mathbf{x}^{(k)} + \mathbf{b}} \quad (2.60)$$

che prevede la risoluzione, ad ogni passo dell'iterazione, di un sistema lineare con matrice dei coefficienti  $\mathbf{M}$ . La seguente matrice

$$\mathbf{B} = \mathbf{M}^{-1}\mathbf{N} = \mathbf{M}^{-1}(\mathbf{M} - \mathbf{A}) = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A}$$

è detta *matrice di iterazione*; essa individua un particolare metodo ed il suo studio è fondamentale per stabilire la *convergenza* e la *rapidità di convergenza* del corrispondente metodo.

Per introdurre i metodi classici di Jacobi e di Gauss-Seidel è utile considerare la

seguinte decomposizione della matrice  $A$

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F} \quad \Rightarrow \quad \mathbf{A} = \left[ \begin{array}{ccc} & & -\mathbf{F} \\ & \mathbf{D} & \\ -\mathbf{E} & & \end{array} \right]$$

### Metodo di Jacobi

Il metodo di Jacobi<sup>12</sup> si ottiene scegliendo

$$\mathbf{M} = \mathbf{D}; \quad \mathbf{N} = \mathbf{E} + \mathbf{F} \quad \Rightarrow \quad \boxed{\mathbf{B}_J = \mathbf{D}^{-1}(\mathbf{E} + \mathbf{F}) = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}}$$

**Algoritmo 2.18** (Metodo di Jacobi) *Data una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , con  $a_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$  e un vettore  $\mathbf{b} \in \mathbb{R}^n$ , a partire da un vettore di tentativo  $\mathbf{x}^{(0)} \in \mathbb{R}^n$  si costruisce la successione  $\{\mathbf{x}^{(k)}\}$  mediante le seguenti formule*

$$\begin{array}{l} \text{For } i = 1, \dots, n \\ \quad x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)})/a_{ii} \\ \text{end } i \end{array}$$

L'implementazione del metodo richiede l'utilizzo di *due* vettori, diciamo  $x_{\text{old}}, x_{\text{new}}$ ; alla fine di ogni ciclo si pone  $x_{\text{new}} \rightarrow x_{\text{old}}$ . Le singole componenti del vettore  $x_{\text{new}}$  sono costruite a partire dal vettore  $x_{\text{old}}$  in maniera indipendente fra loro; l'algoritmo può essere quindi implementato facilmente su calcolatori ad architettura parallela. In linguaggio MATLAB il metodo può essere implementato nel seguente modo

$$\mathbf{M} = \text{diag}(\text{diag}(\mathbf{A})); \quad \mathbf{N} = \mathbf{M} - \mathbf{A}; \quad \mathbf{x} = \mathbf{M} \setminus (\mathbf{N} * \mathbf{x} + \mathbf{b});$$

Osserviamo che, nella forma precedente, l'algoritmo di Jacobi è definito solo se  $a_{ii} \neq 0$ ,  $i = 1, \dots, n$ . Tuttavia, se tale ipotesi non è verificata, ma la matrice  $\mathbf{A}$  è non singolare, è possibile riordinare preliminarmente le equazioni e le incognite del sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , in maniera che il metodo risulti definito. È da tenere presente, comunque, che la matrice di iterazione, e quindi le proprietà di convergenza, dipendono dall'ordinamento di righe e di incognite utilizzato.

<sup>12</sup>Il metodo, introdotto da Jacobi (1845), fu successivamente chiamato *metodo delle sostituzioni simultanee* (Geirenger, 1949) e *metodo iterativo di Richardson*, (Keller, 1958). È anche noto come *metodo a passo totale* (total-step), dal tedesco *Gesamtschrittverfahren*.

► **Esempio 2.20** Consideriamo l'applicazione del metodo di Jacobi al seguente sistema

$$\begin{cases} 20x_1 + 2x_2 - x_3 = 25 \\ 2x_1 + 13x_2 - 2x_3 = 30 \\ x_1 + x_2 + x_3 = 2 \end{cases} \quad (2.61)$$

che ha come soluzione esatta il vettore  $\mathbf{x} = [1, 2, -1]^T$ . Il metodo di Jacobi corrisponde alla seguente decomposizione della matrice dei coefficienti

$$\mathbf{A} = \begin{bmatrix} 20 & 2 & -1 \\ 2 & 13 & -2 \\ 1 & 1 & 1 \end{bmatrix} \Rightarrow \mathbf{M} = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 13 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} 0 & -2 & 1 \\ -2 & 0 & 2 \\ -1 & -1 & 0 \end{bmatrix}$$

Partendo dal vettore iniziale  $\mathbf{x}^0 = [0, 0, 0]^T$ , si ottengono i risultati contenuti nella Tabella 2.3, che mostrano la convergenza del metodo. Lasciamo come esercizio di verificare che il metodo risulta, invece, divergente quando è applicato al sistema lineare che si ottiene dal sistema (2.61) scambiando le prime due righe. È importante osservare che nel primo caso il raggio spettrale<sup>13</sup> della matrice di Jacobi è dato da 0.4490, mentre per il sistema con le due righe permutate il raggio spettrale vale 8.0710. In effetti, come vedremo nel seguito, condizione necessaria e sufficiente per la convergenza di un metodo iterativo è che il raggio spettrale della matrice di iterazione  $\mathbf{M}^{-1}\mathbf{N}$  sia strettamente minore di 1. ■

$k$	Jacobi	Gauss-Seidel
1	5.5577e + 00	4.7308e + 00
2	3.8038e + 00	8.9364e - 01
3	7.3203e - 01	2.0355e - 01
4	8.0164e - 01	5.0545e - 02
5	2.1970e - 01	1.2269e - 02
6	1.6726e - 01	2.9954e - 03
7	5.5982e - 02	7.3022e - 04
8	3.2669e - 02	1.7808e - 04
9	1.2908e - 02	4.3426e - 05
10	5.9462e - 03	1.0590e - 05
11	2.7599e - 03	2.5824e - 06
12	9.9184e - 04	6.2974e - 07
...		
20	1.7859e - 06	
21	9.5681e - 07	

Tabella 2.3: Successione degli errori  $\|\mathbf{x}^{(k)} - \mathbf{x}\|_1$  forniti dal metodo di Jacobi, e rispettivamente di Gauss-Seidel, applicati al sistema lineare (2.61).

<sup>13</sup>In MATLAB il raggio spettrale di una matrice  $\mathbf{A}$  può essere calcolato mediante il comando `abs(eig(A))`.

### Metodo di Gauss-Seidel

Nel metodo di Gauss-Seidel<sup>14</sup> si assume

$$\mathbf{M} = \mathbf{D} - \mathbf{E}; \quad \mathbf{N} = \mathbf{F} \Rightarrow \boxed{\mathbf{B}_G = (\mathbf{D} - \mathbf{E})^{-1}\mathbf{F} = (\mathbf{I} - \mathbf{D}^{-1}\mathbf{E})^{-1}\mathbf{D}^{-1}\mathbf{F}}$$

**Algoritmo 2.19** (Metodo di Gauss-Seidel) *Data una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , con  $a_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$  e un vettore  $\mathbf{b} \in \mathbb{R}^n$ , a partire da un vettore di tentativo  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , si costruisce la successione  $\{\mathbf{x}^{(k)}\}$  mediante le seguenti formule*

For  $i = 1, \dots, n$   
 $x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)})/a_{ii}$   
 end  $i$

Osserviamo che, a differenza del metodo di Jacobi, per l'implementazione del metodo di Gauss-Seidel è sufficiente un *unico* vettore, in quanto le componenti del vettore iterato sono utilizzate non appena esse sono state calcolate. Il metodo può essere implementato in linguaggio MATLAB nel seguente modo

```
M=tril(A); N=M-A; x=M\(N*x+b);
```

Come si vede dalla Tabella 2.3, il metodo di Gauss-Seidel può avere una convergenza superiore rispetto al metodo di Jacobi. Questa proprietà non è, tuttavia, generale, ma dipende dalla classe di matrici a cui si applicano i metodi. In effetti, come vedremo nel seguito, esistono matrici per le quali il metodo di Gauss-Seidel converge ma il metodo di Jacobi diverge, e viceversa matrici per le quali converge il metodo di Jacobi, ma non il metodo di Gauss-Seidel.

Quando la matrice  $\mathbf{A}$  è simmetrica definita positiva, il metodo di Gauss-Seidel può essere interpretato come un algoritmo per la ricerca del minimo della seguente funzione quadratica

$$F(\mathbf{x}) = (\mathbf{A}\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x}) = \sum_{i,j=1}^n a_{ij}x_ix_j - 2 \sum_{i=1}^n x_ib_i$$

Rinviando per maggiori dettagli al successivo Capitolo 5, ci limiteremo ora a fornire in Figura 2.10 una rappresentazione grafica del metodo in corrispondenza al caso bidimensionale e alle seguenti scelte della matrice  $\mathbf{A}$  e del vettore  $\mathbf{b}$

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

<sup>14</sup>Suggerito da Gauss in *Theoria Motus corporum coelestium in sectionibus conicis solem ambientium* (1809), nell'ambito della risoluzione del sistema delle equazioni normali ottenute applicando il metodo dei minimi quadrati. Sviluppato da P. L. Seidel (1821–1896) in *Über ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt, durch successive Annäherung aufzulösen* (1874). Il metodo è anche noto come *metodo delle sostituzioni successive* (Geiringer, 1949) e *metodo di Liebmann* (Frankel, 1950).

a cui corrisponde la soluzione  $\mathbf{x} = [1, 1]^T$  del sistema  $\mathbf{Ax} = \mathbf{b}$ . Il metodo effettua, in sostanza, successive *minimizazioni unidimensionali* lungo le direzioni degli assi. Nella figura sono rappresentate le curve di livello passanti per i successivi punti di minimo, ossia le curve definite da  $F(\mathbf{x}) = F(\bar{\mathbf{x}})$ , ove  $\bar{\mathbf{x}}$  è il generico punto di minimo. Nel seguito considereremo altri modi più convenienti di definire le direzioni di minimizzazione, in particolare le direzioni coniugate.

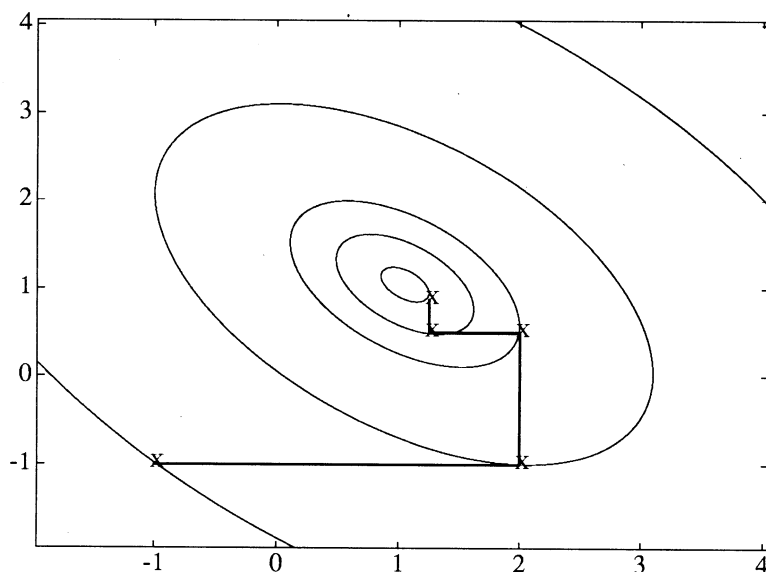


Figura 2.10: Interpretazione del metodo di Gauss-Seidel come metodo di minimizzazione di una funzione quadratica. I punti indicati sono i successivi punti di minimo lungo gli assi coordinati, ossia i punti  $(x_1^{(0)}, x_2^{(0)})$ ,  $(x_1^{(1)}, x_2^{(0)})$ ,  $(x_1^{(1)}, x_2^{(1)})$ ,  $(x_1^{(2)}, x_2^{(1)})$ , ...

### Metodo di rilassamento

Un modo semplice per accelerare la convergenza dei metodi di Jacobi e di Gauss-Seidel consiste nell'introduzione nella matrice di iterazione di un opportuno parametro, noto come *parametro di rilassamento*. Indicando con  $\omega$ , con  $\omega > 0$ , tale parametro e limitandoci al metodo di Gauss-Seidel, si assume

$$\mathbf{M} = \frac{\mathbf{D}}{\omega} - \mathbf{E}; \quad \mathbf{N} = \left(\frac{1}{\omega} - 1\right) \mathbf{D} + \mathbf{F} \Rightarrow \boxed{\mathbf{B}_\omega = (\mathbf{D} - \omega \mathbf{E})^{-1} [(1 - \omega) \mathbf{D} + \omega \mathbf{F}]}$$

Il corrispondente metodo iterativo è noto come *metodo di rilassamento*, o metodo SOR (successive over relaxation); quando  $\omega < 1$  il metodo è anche detto metodo

underrelaxation<sup>15</sup>.

**Algoritmo 2.20** (Metodo SOR) *Data una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , con  $a_{ii} \neq 0$ ,  $i=1, 2, \dots, n$  ed un vettore  $\mathbf{b} \in \mathbb{R}^n$ , a partire da un vettore di tentativo  $\mathbf{x}^{(0)} \in \mathbb{R}^n$  si costruisce, per  $\omega \in \mathbb{R}$ , la successione  $\{\mathbf{x}^{(k)}\}$  mediante le seguenti formule*

```

For  $i = 1, \dots, n$ 
   $x_i^{(k+1/2)} = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)})/a_{ii}$ 
   $x_i^{(k+1)} = \omega x_i^{(k+1/2)} + (1 - \omega)x_i^{(k)}$ 
end  $i$ 

```

La scelta  $\omega = 1$  fornisce ancora il metodo di Gauss-Seidel, e in ogni caso il numero di operazioni richieste per effettuare una iterazione è dello stesso ordine di grandezza di quelle richieste dal metodo di Gauss-Seidel. In linguaggio MATLAB si ha la seguente implementazione

```
M=diag(diag(A))/omega+tril(A,-1); N=M-A; x=M\(N*x+b);
```

Nell'applicazione del metodo SOR è naturalmente importante la scelta del parametro  $\omega$ . Tale scelta deve essere tale da rendere più elevata possibile la velocità di convergenza del metodo. Come vedremo nel seguito, tale obiettivo è raggiunto dal valore di  $\omega$  che minimizza il raggio spettrale della matrice di rilassamento  $\mathbf{B}_\omega$ . Il guadagno che si può ottenere è, tuttavia, dipendente dal numero di condizionamento della matrice di partenza  $\mathbf{A}$ . Questo aspetto è messo in rilievo dal seguente esempio.

► **Esempio 2.21** Consideriamo la seguente matrice

$$\mathbf{A} = \begin{bmatrix} -4 & 1 & 1 & 1 \\ 1 & -4 & 1 & 1 \\ 1 & 1 & -4 & 1 \\ 1 & 1 & 1 & -4 \end{bmatrix} \quad (2.62)$$

Nella Figura 2.11 è rappresentata la funzione

$$\omega \rightarrow \rho(\mathbf{B}_\omega) \quad (2.63)$$

ove  $\rho(\cdot)$  è il raggio spettrale, ossia il massimo dei moduli degli autovalori. Il minimo della funzione (2.63) è ottenuto per  $\omega = 1.22$ , in corrispondenza al quale assume il valore 0.2974. Il numero di condizionamento della matrice  $\mathbf{A}$  è dato da  $\mu_2(\mathbf{A}) = 5$ .

Consideriamo, quindi, la matrice

$$\mathbf{A} = \begin{bmatrix} 2 & 6 & 9 \\ 6 & 21 & 31 \\ 9 & 31 & 46 \end{bmatrix} \quad (2.64)$$

<sup>15</sup>Il metodo è anche noto come *metodo accelerato di Liebmann* (Frankel, 1950) e *metodo di Gauss-Seidel estrapolato* (Kahan, 1958).



per la quale si ha  $\mu_2(\mathbf{A}) = 1251.0$ . Il minimo della funzione (2.63) è ottenuto per  $\omega = 1.7$ , nel quale la funzione assume il valore 0.9571. Come si vede, in questo caso il guadagno è decisamente minore del caso precedente.

Con riferimento, in particolare, alla risoluzione dei sistemi lineari ottenuti nella discretizzazione dei problemi differenziali ai limiti (cfr. Esempio 2.17), si ha che l'efficienza del metodo SOR, come tecnica di accelerazione del metodo di Gauss-Seidel, diminuisce all'aumentare del numero di suddivisioni, ossia al diminuire del passo di discretizzazione, e quindi dell'errore di discretizzazione.

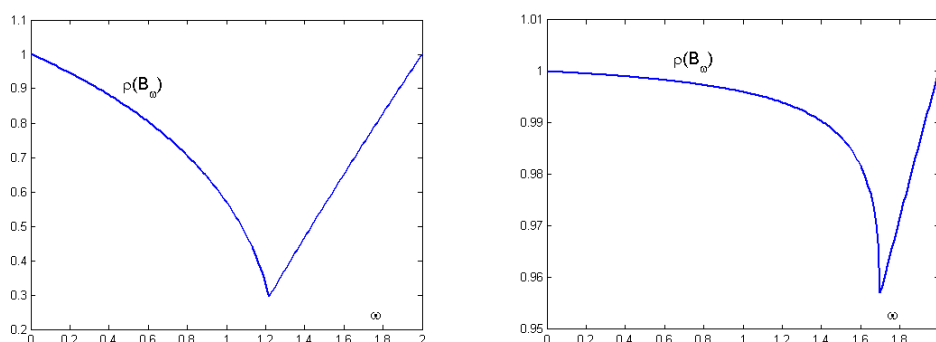


Figura 2.11: Rappresentazione della funzione  $\omega \rightarrow \rho(\mathbf{B}_\omega)$ , in corrispondenza (prima figura) alla matrice bencondizionata (2.62) e rispettivamente (seconda figura) alla matrice malcondizionata (2.64).

### 2.3.2 Metodi iterativi a blocchi

In diverse applicazioni, in particolare nella discretizzazione dei problemi alle derivate parziali e nello studio di sistemi di reti, il sistema  $\mathbf{Ax} = \mathbf{b}$  presenta una *struttura a blocchi* della seguente forma

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1p} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2p} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \cdots & \mathbf{A}_{pp} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_p \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_p \end{bmatrix} \quad (2.65)$$

ove  $\mathbf{A}_{11}, \mathbf{A}_{22}, \dots, \mathbf{A}_{pp}$  sono matrici *quadrate* e  $\mathbf{x}_i, \mathbf{b}_i$  sono vettori dello stesso ordine della matrice  $\mathbf{A}_{ii}$ , per ogni  $i = 1, \dots, p$ . A partire dalla seguente decomposizione

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F}$$

ove  $\mathbf{D} = [\mathbf{A}_{ii}]$ ,  $i = 1, \dots, p$ ;  $-\mathbf{E} = [\mathbf{A}_{ij}]$ ,  $i > j$ ;  $-\mathbf{F} = [\mathbf{A}_{ij}]$ ,  $i < j$ , i metodi considerati nel paragrafo precedente possono essere formulati a blocchi. Tali formulazioni

hanno interesse, in particolare, nelle implementazioni su calcolatori ad architettura parallela.

Come esemplificazione, le seguenti formule definiscono un particolare *metodo SOR a blocchi*

$$\mathbf{A}_{ii}\mathbf{x}_i^{(k+1)} = \omega \left( \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{A}_{ij}\mathbf{x}_j^{(k+1)} - \sum_{j=i+1}^p \mathbf{A}_{ij}\mathbf{x}_j^{(k)} \right) + (1 - \omega)\mathbf{A}_{ii}\mathbf{x}_i^{(k)} \quad (2.66)$$

per il quale è richiesta, per ogni  $i$ , la risoluzione di un sistema lineare, con matrice dei coefficienti data da  $\mathbf{A}_{ii}$ . Dal momento che tale matrice è indipendente dall'indice di iterazione  $k$ , si può effettuare una fattorizzazione preliminare  $\mathbf{A}_{ii} = \mathbf{L}_i \mathbf{U}_i$ , e quindi procedere, ad ogni iterazione, alla risoluzione di due sistemi triangolari.

Il successivo esempio analizza su due problemi particolari il problema della convergenza dei metodi a blocchi rispetto ai metodi per punti.

► **Esempio 2.22** Consideriamo la seguente matrice reale simmetrica

$$\mathbf{A} = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix}$$

e una sua partizione a blocchi in maniera che  $\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F}$ , con

$$\mathbf{D} = \left[ \begin{array}{cc|c} 5 & 2 & 0 \\ 2 & 5 & 0 \\ \hline 0 & 0 & 5 \end{array} \right], \quad \mathbf{E} = \left[ \begin{array}{cc|c} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \hline -2 & -3 & 0 \end{array} \right]$$

e  $\mathbf{F} = \mathbf{E}^T$ . Nella Tabella 2.4 sono riportati gli errori ottenuti con il metodo di Gauss-Seidel a blocchi (ossia con il metodo (2.66) con  $\omega = 1$ ) in relazione al sistema lineare  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , con  $\mathbf{b} = [9, 10, 10]^T$  e con valore iniziale  $\mathbf{x}^{(0)} = [0, 0, 0]^T$ . Il confronto con i risultati ottenuti con il metodo di Gauss-Seidel per punti mette in evidenza che asintoticamente il metodo per punti converge più rapidamente del metodo per blocchi. In effetti, si trova

$$\rho(\mathbf{B}_G^{\text{punti}}) = 0.3098, \quad \rho(\mathbf{B}_G^{\text{blocchi}}) = 0.3904$$

Consideriamo, ora, la seguente matrice reale simmetrica

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & -1/4 & -1/4 \\ 0 & 1 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 & 0 \\ -1/4 & -1/4 & 0 & 1 \end{bmatrix}$$

e la partizione a blocchi definita da  $\mathbf{A}_1 = \mathbf{D} - \mathbf{E} - \mathbf{F}$ , con

$$\mathbf{D} = \left[ \begin{array}{ccc|c} 1 & 0 & -1/4 & 0 \\ 0 & 1 & -1/4 & 0 \\ -1/4 & -1/4 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right], \quad \mathbf{E} = \left[ \begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 1/4 & 1/4 & 0 & 0 \end{array} \right]$$

$k$	G.-S. per punti	G.-S. per blocchi
0	$3.0000e + 00$	$3.0000e + 00$
1	$1.5680e + 00$	$1.1048e + 00$
2	$5.3171e - 01$	$4.3138e - 01$
3	$2.1641e - 01$	$1.6844e - 01$
4	$9.1316e - 02$	$6.5774e - 02$
5	$3.2553e - 02$	$2.5683e - 02$
6	$1.0244e - 02$	$1.0029e - 02$
7	$2.8577e - 03$	$3.9159e - 03$
8	$7.1179e - 04$	$1.5291e - 03$
9	$1.7121e - 04$	$5.9707e - 04$
10	$3.9791e - 05$	$2.3314e - 04$
11	$1.3868e - 05$	$9.1036e - 05$

Tabella 2.4: Errori  $\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_1$  corrispondenti al metodo di Gauss-Seidel per punti e per blocchi, ove  $\bar{\mathbf{x}}$  è la soluzione esatta del sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{b} = [9, 10, 10]^T$ .

e  $\mathbf{F} = \mathbf{E}^T$ . Nella Tabella 2.5 sono riportati gli errori ottenuti con il metodo di Gauss-Seidel a blocchi (ossia con il metodo (2.66) con  $\omega = 1$ ) in relazione al sistema lineare  $\mathbf{A}_1\mathbf{x} = \mathbf{b}_1$ , con  $\mathbf{b} = [0.5, 0.5, 0.5, 0.5]^T$  e valore iniziale  $\mathbf{x}^{(0)} = [0, 0, 0, 0]^T$ . Il confronto con i risultati ottenuti con il metodo di Gauss-Seidel per punti mette in evidenza che asintoticamente il metodo per blocchi converge più rapidamente del metodo per punti. In effetti, si trova

$$\rho(\mathbf{B}_G^{\text{punti}}) = 0.2500, \quad \rho(\mathbf{B}_G^{\text{blocchi}}) = 0.1428$$

### 2.3.3 Studio della convergenza

Un metodo iterativo della forma (2.60) è detto *convergente* se, qualunque sia il vettore iniziale  $\mathbf{x}^{(0)}$ , la successione  $\{\mathbf{x}^{(k)}\}$  è convergente. Vedremo, ora, che tale proprietà è legata agli autovalori della matrice di iterazione  $\mathbf{B}$ . Considerando, infatti, la decomposizione generale (2.59) e indicando con  $\bar{\mathbf{x}}$  la soluzione del sistema  $\mathbf{Ax} = \mathbf{b}$ , si hanno le seguenti relazioni

$$\begin{aligned} \bar{\mathbf{x}} &= \mathbf{B}\bar{\mathbf{x}} + \mathbf{M}^{-1}\mathbf{b} \\ \mathbf{x}^{(k+1)} &= \mathbf{B}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \end{aligned} \quad \Rightarrow \quad \boxed{\mathbf{e}^{(k+1)} = \mathbf{B}\mathbf{e}^{(k)}} \quad (2.67)$$

ove si è posto  $\mathbf{e}^{(k)} := \bar{\mathbf{x}} - \mathbf{x}^{(k)}$ . Applicando successivamente la relazione (2.67), si ottiene il seguente risultato

$$\boxed{\mathbf{e}^{(k)} = \mathbf{B}^k \mathbf{e}^{(0)}}$$

Pertanto, il metodo è convergente quando la successione di potenze  $\mathbf{B}^k$  converge alla matrice nulla e una condizione necessaria e sufficiente può essere formulata mediante il raggio spettrale della matrice  $\mathbf{B}$  (cfr. Appendice A).

$k$	G.-S. per punti	G.-S. per blocchi
0	$4.0000e + 00$	$4.0000e + 00$
1	$1.5000e + 00$	$8.5714e - 01$
2	$3.7500e - 01$	$1.2245e - 01$
3	$9.3750e - 02$	$1.7493e - 02$
4	$2.3438e - 02$	$2.4990e - 03$
5	$5.8594e - 03$	$3.5699e - 04$
6	$1.4648e - 03$	$5.0999e - 05$
7	$3.6621e - 04$	$7.2856e - 06$
8	$9.1553e - 05$	$1.0408e - 06$
9	$2.2888e - 05$	$1.4869e - 07$
10	$5.7220e - 06$	$2.1241e - 08$
11	$1.4305e - 06$	$3.0344e - 09$

Tabella 2.5: Errori  $\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_1$  corrispondenti al metodo di Gauss-Seidel per punti e per blocchi, ove  $\bar{\mathbf{x}}$  è la soluzione esatta del sistema lineare  $\mathbf{A}_1\mathbf{x} = \mathbf{b}_1$ , con  $\mathbf{b}_1 = [0.5, 0.5, 0.5, 0.5]^T$ .

**Teorema 2.5** (Convergenza) *Il metodo iterativo definito dalla matrice di iterazione  $\mathbf{B}$  converge per ogni  $\mathbf{x}^{(0)} \in \mathbb{R}^n$  se e solo se*

$$\boxed{\rho(\mathbf{B}) < 1} \quad (2.68)$$

► **Esempio 2.23** Consideriamo i metodi di Jacobi e di Gauss-Seidel per la matrice

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{bmatrix}, \quad \mathbf{B}_J = \begin{bmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{bmatrix}, \quad \mathbf{B}_G = \begin{bmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{bmatrix}$$

Il polinomio caratteristico della matrice  $\mathbf{B}_J$  è dato da  $-\lambda^3$  e della matrice  $\mathbf{B}_G$  da  $-\lambda(\lambda^2 + 4\lambda - 4)$ , per cui

$$\rho(\mathbf{B}_J) = 0; \quad \rho(\mathbf{B}_G) = 2(1 + \sqrt{2})$$

In questo caso, quindi, il metodo di Jacobi converge, mentre il metodo di Gauss-Seidel diverge.

Lasciamo come esercizio di verificare, procedendo allo stesso modo, che per la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{bmatrix}$$

il metodo di Jacobi diverge, mentre il metodo di Gauss-Seidel converge. ■

La condizione (2.68), necessaria e sufficiente per la convergenza di un metodo iterativo, non è in generale di facile verifica. Essa, tuttavia, permette di ottenere delle condizioni sufficienti di più facile verifica, ossia di individuare delle classi di matrici

per le quali un particolare metodo è convergente. Questo aspetto verrà considerato nel paragrafo successivo. In questo paragrafo, dopo aver ricordato una condizione necessaria per la scelta del parametro  $\omega$  nel metodo di rilassamento, discuteremo brevemente l'aspetto, importante nelle applicazioni, del *controllo della convergenza*.

Dal Teorema 2.5 si può ricavare facilmente la seguente *condizione necessaria* per il metodo di rilassamento.

**Proposizione 2.5** *Per ogni matrice  $\mathbf{A}$  di ordine  $n$  il raggio spettrale della matrice del metodo di rilassamento  $\mathbf{B}_\omega$  è maggiore o uguale a  $|\omega - 1|$ , e pertanto se il metodo è convergente, si ha necessariamente*

$$\boxed{0 < \omega < 2}$$

**Test di arresto per un metodo iterativo** Nelle applicazioni di un metodo iterativo, un test comune per arrestare il calcolo della successione  $\{\mathbf{x}^{(k)}\}$ , è il seguente

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \epsilon_1 + \epsilon_2 \|\mathbf{x}^{(k)}\|$$

ove  $\epsilon_1$  (rispettivamente  $\epsilon_2$ ), è una tolleranza prefissata sull'errore assoluto (rispettivamente sull'errore relativo). È importante osservare che il test precedente non garantisce che la soluzione sia stata approssimata con la precisione richiesta. Si può, infatti, mostrare che, per una norma di matrice e di vettore compatibili, si ha

$$\|\mathbf{e}^{(k)}\| \leq \frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{1 - \|\mathbf{B}\|}$$

per cui si può avere che  $\|\mathbf{e}^{(k-1)}\|$  è grande, anche se il test precedente è verificato.

Un test di arresto di tipo differente è, se  $\mathbf{b} \neq 0$ , dato da

$$\boxed{\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{b}\|} \leq \epsilon} \quad (2.69)$$

ove  $\mathbf{r}^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}$ . Dal momento che

$$\|\mathbf{e}^{(k)}\| = \|\mathbf{A}^{-1}\mathbf{r}^{(k)}\| \leq \|\mathbf{A}^{-1}\| \|\mathbf{r}^{(k)}\|$$

il test (2.69) implica che

$$\boxed{\frac{\|\mathbf{e}^{(k)}\|}{\|\bar{\mathbf{x}}\|} \leq \epsilon \mu(\mathbf{A})} \quad (2.70)$$

**Rapidità di convergenza** Dalla relazione  $\mathbf{e}^{(k)} = \mathbf{B}^{(k)} \mathbf{e}^{(0)}$ , utilizzando una norma di matrice e di vettore compatibili, si ha  $\|\mathbf{e}^{(k)}\| \leq \|\mathbf{B}^k\| \|\mathbf{e}^{(0)}\|$ , da cui

$$\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|} \leq \|\mathbf{B}^k\| = \max_{\mathbf{e}^{(0)} \neq 0} \frac{\|\mathbf{B}^k \mathbf{e}^{(0)}\|}{\|\mathbf{e}^{(0)}\|} = \max_{\mathbf{e}^{(0)} \neq 0} \frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|}$$

Si definisce *fattore medio di riduzione dell'errore per iterazione* il numero  $\sigma$  tale che

$$\sigma^k = \frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|}$$

In *media* l'errore è moltiplicato per  $\sigma$  ad ogni iterazione. Si ha  $\sigma \leq \|\mathbf{B}^k\|^{1/k}$ .

**Definizione 2.2** Si definisce *velocità media di convergenza per  $k$  iterazioni* il numero

$$R_k(\mathbf{B}) = -\ln \|\mathbf{B}^k\|^{1/k}$$

La velocità media di convergenza è inversamente proporzionale al numero di iterazioni. Per avere, infatti  $\|\mathbf{e}^{(k)}\|/\|\mathbf{e}^{(0)}\| \leq \epsilon$  si impone  $\|\mathbf{B}^k\| \leq \epsilon$ , da cui

$$-\ln \|\mathbf{B}^k\|^{1/k} \geq -\frac{1}{k} \ln \epsilon \Rightarrow k \geq \frac{-\ln \epsilon}{-\ln \|\mathbf{B}^k\|^{1/k}} = \frac{-\ln \epsilon}{R_k(\mathbf{B})}$$

Nel caso particolare in cui la matrice  $B$  è hermitiana, scegliendo come norma la norma spettrale, si ha che la velocità di convergenza è uguale a  $-\ln \rho(\mathbf{B})$ . Anche nel caso generale è tuttavia comodo utilizzare la seguente definizione.

**Definizione 2.3** Si chiama *velocità di convergenza asintotica del metodo iterativo definito dalla matrice di iterazione  $\mathbf{B}$*  il numero

$$\boxed{R(\mathbf{B}) = -\ln \rho(\mathbf{B})}$$

Si può allora dimostrare che il numero di iterazioni  $k$  necessarie per ridurre l'errore di un fattore  $\epsilon$  verifica la disuguaglianza  $k \geq -\ln \epsilon / R(\mathbf{B})$ .

## Matrici a predominanza diagonale

Rinviando all'Appendice A per la definizione e le principali proprietà delle matrici a predominanza diagonale, ricordiamo, in riferimento alla convergenza dei metodi iterativi, i seguenti risultati.

**Teorema 2.6** Se  $\mathbf{A}$  è una matrice di ordine  $n$  a predominanza diagonale stretta o irriducibilmente diagonalmente dominante, allora i metodi di Jacobi e di Gauss-Seidel sono convergenti.

► **Esempio 2.24** Si considerino i due sistemi equivalenti

$$\begin{cases} x-2y = -2 \\ 2x+y = 2 \end{cases} \iff \begin{cases} 2x+y = 2 \\ x-2y = -2 \end{cases}$$

La matrice dei coefficienti del secondo sistema è a predominanza diagonale stretta e pertanto il metodo di Gauss-Seidel converge. Per il primo sistema la matrice di iterazione di Gauss-Seidel è la seguente

$$\mathbf{B}_G = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & -4 \end{bmatrix}$$

ed ha come autovalori i numeri  $0, -4$ . Essendo  $\rho(\mathbf{B}_G) = 4 > 1$ , il metodo di Gauss-Seidel risulta non convergente. L'esempio evidenzia, in particolare, il fatto che la convergenza del metodo dipende dall'ordinamento.

È interessante osservare che il metodo SOR applicato al primo sistema converge per una scelta opportuna del parametro  $\omega$ . Più precisamente, la funzione  $\omega \rightarrow \rho(\mathbf{B}_\omega)$  raggiunge il valore minimo 0.39 per  $\omega = 0.61$ .

► **Esempio 2.25** Esaminiamo la convergenza del metodo di Jacobi applicato alle seguenti matrici

$$\mathbf{A}_1 = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 1 & -3/4 \\ -1/12 & 1 \end{bmatrix}$$

Indicando con  $\mathbf{B}_J^{(1)}, \mathbf{B}_J^{(2)}$ , le corrispondenti matrici di iterazione di Jacobi, si ha

$$\mathbf{B}_J^{(1)} = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}, \Rightarrow \rho(\mathbf{B}_J^{(1)}) = 0.5; \quad \mathbf{B}_J^{(2)} = \begin{bmatrix} 0 & 3/4 \\ 1/12 & 0 \end{bmatrix} \Rightarrow \rho(\mathbf{B}_J^{(2)}) = 0.25$$

Il metodo di Jacobi converge per entrambe le matrici, e convergenza è più rapida per la matrice  $\mathbf{A}_2$ . L'esempio mostra, quindi, che il metodo di Jacobi non necessariamente converge più rapidamente per le matrici "più" diagonalmente dominanti. ■

Per quanto riguarda il confronto tra il metodo di Jacobi e il metodo di Gauss-Seidel, si ha il seguente interessante risultato.

**Teorema 2.7** (Stein–Rosenberg) *Sia  $\mathbf{A} \in \mathbb{R}^{n \times n}$ . Se gli elementi principali di  $\mathbf{A}$  sono non nulli e gli elementi della matrice di iterazione di Jacobi  $\mathbf{B}_J$  sono non negativi, allora si verifica uno ed uno solo dei seguenti risultati*

1.  $0 < \rho(\mathbf{B}_G) < \rho(\mathbf{B}_J) < 1$
2.  $1 < \rho(\mathbf{B}_J) < \rho(\mathbf{B}_G)$
3.  $\rho(\mathbf{B}_G) = \rho(\mathbf{B}_J) = 0$
4.  $\rho(\mathbf{B}_G) = \rho(\mathbf{B}_J) = 1$

Per le *matrici tridiagonali* è possibile precisare la relazione tra la velocità di convergenza del metodo di Gauss-Seidel e del metodo di Jacobi. Si può, infatti, dimostrare che se gli elementi sulla diagonale principale sono differenti dallo zero, si ha

$$\rho(\mathbf{B}_G) = \rho(\mathbf{B}_J)^2 \quad (2.71)$$

ossia il metodo di Gauss-Seidel ha una velocità di convergenza asintotica doppia del metodo di Jacobi. Il risultato (2.71) può essere esteso al caso di decomposizioni a blocchi.

### Matrici definite positive

Per le matrici definite positive, che costituiscono una classe importante di matrici nelle applicazioni, si ha il seguente risultato relativo alla convergenza del metodo di rilassamento.

**Teorema 2.8** *Sia  $\mathbf{A}$  una matrice hermitiana decomposta, per punti o per blocchi, nella forma  $\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F}$  con  $\mathbf{D}$  matrice definita positiva. Allora, il metodo di rilassamento, per punti o per blocchi, è convergente, se e solo se  $0 < \omega < 2$  e  $A$  è definita positiva.*

Pertanto, per le matrici definite positive il valore del parametro ottimale  $\omega$  corrisponde al minimo nell'intervallo  $(0, 2)$  della seguente funzione

$$\omega \rightarrow \rho(\mathbf{B}_\omega) \quad (2.72)$$

Rinviando all'Esempio 2.21 per una opportuna esemplificazione della funzione (2.72), osserviamo che la determinazione del punto di minimo rappresenta, in generale, ossia per le matrici che non hanno una struttura particolare, un problema di non facile soluzione. Tale difficoltà è, in sostanza, uno dei motivi di interesse del metodo del gradiente coniugato che analizzeremo nel prossimo paragrafo.

### 2.3.4 Metodo del gradiente coniugato

Abbiamo già osservato in precedenza che quando la matrice  $\mathbf{A}$  è simmetrica definita positiva, il problema della soluzione del sistema lineare  $\mathbf{Ax} = \mathbf{b}$  è equivalente alla minimizzazione della funzione quadratica  $J(\mathbf{x}) = (\mathbf{Ax}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ . In particolare, abbiamo visto che il metodo di Gauss-Seidel corrisponde a minimizzare la funzione  $F(\mathbf{x})$  lungo le direzioni degli assi coordinati. Il metodo del gradiente coniugato corrisponde ad una scelta più conveniente di tali direzioni. In effetti, con tale metodo viene costruita una successione di vettori  $\{\mathbf{x}^{(k)}\}$ ,  $k = 0, 1, \dots$ , per i quali si ha  $\mathbf{x}^{(m)} = \bar{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b}$  per un valore  $m \leq n$ , ove  $n$  è l'ordine della matrice  $\mathbf{A}$ . In altre parole, dal punto di vista teorico il metodo del gradiente coniugato è un metodo diretto, in quanto fornisce la soluzione con un numero finito di operazioni. Tuttavia,



in aritmetica di calcolatore, e quindi in presenza di errori di arrotondamento, il metodo è usualmente implementato come metodo iterativo.

Lo studio del metodo del gradiente coniugato trova la sua collocazione più naturale nell'ambito dei metodi di ottimizzazione e sarà pertanto approfondito nel successivo Capitolo 5. In questo Capitolo ci limiteremo, pertanto, a fornire gli elementi essenziali per comprendere l'algoritmo.

A partire da una stima  $\mathbf{x}^{(0)}$  della soluzione del sistema  $\mathbf{Ax} = \mathbf{b}$ , si costruisce una successione di vettori  $\{\mathbf{x}^{(k)}\}$  nel seguente modo

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)} \quad (2.73)$$

ove  $\alpha_k \in \mathbb{R}$  è tale che

$$F(\mathbf{x}^{(k+1)}) = \min_{\alpha \in \mathbb{R}} F(\mathbf{x}^{(k)} + \alpha \mathbf{p}^{(k)}) \quad (2.74)$$

e  $\mathbf{p}^{(k)}$  è un *vettore direzione* scelto nel modo che ora vedremo. Indicato con  $\mathbf{r}^{(k)}$  il *vettore residuo*, ossia il vettore  $\mathbf{b} - \mathbf{Ax}^{(k)}$ , si pone

$$\mathbf{p}^{(k)} = \begin{cases} \mathbf{r}^{(0)} & \text{se } k = 0 \\ \mathbf{r}^{(k)} + \beta_k \mathbf{p}^{(k-1)} & \text{se } k \geq 1 \end{cases} \quad (2.75)$$

ove  $\beta_k$  è tale che

$$(\mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(k-1)} = 0 \quad (2.76)$$

Il vettore  $\mathbf{p}^{(k)}$  viene detto *A-coniugato* con il vettore  $\mathbf{p}^{(k-1)}$  (cfr. Capitolo 5 per l'interpretazione geometrica di tale proprietà). Sostituendo l'espressione di  $\mathbf{p}^{(k)}$  data dalla definizione (2.75) nell'equazione (2.76), si ricava

$$\beta_k = -\frac{(\mathbf{r}^{(k)})^T \mathbf{A} \mathbf{p}^{(k-1)}}{(\mathbf{p}^{(k-1)})^T \mathbf{A} \mathbf{p}^{(k-1)}} \quad k \geq 1 \quad (2.77)$$

Il valore di  $\alpha_k$  può essere ottenuto esplicitamente nel seguente modo. Derivando la funzione  $F(\mathbf{x}^{(k)} + \alpha \mathbf{p}^{(k)})$  rispetto ad  $\alpha$  si ottiene

$$\frac{dF}{d\alpha} = (\mathbf{x}^{(k)} + \alpha \mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(k)} - \mathbf{b}^T \mathbf{p}^{(k)}$$

da cui, imponendo  $dF/d\alpha = 0$ , si ricava

$$\alpha_k = \frac{(\mathbf{b} - \mathbf{Ax}^{(k)})^T \mathbf{p}^{(k)}}{(\mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(k)}} = \frac{(\mathbf{r}^{(k)})^T \mathbf{p}^{(k)}}{(\mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(k)}} \quad (2.78)$$

Dalla (2.73) si ha per  $k = 0, 1, \dots$

$$\mathbf{b} - \mathbf{Ax}^{(k+1)} = \mathbf{b} - \mathbf{Ax}^{(k)} - \alpha_k \mathbf{A} \mathbf{p}^{(k)}$$

e quindi

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A} \mathbf{p}^{(k)} \quad (2.79)$$

da cui per la (2.78)

$$(\mathbf{r}^{(k+1)})^T \mathbf{p}^{(k)} = (\mathbf{r}^{(k)} - \alpha_k \mathbf{A} \mathbf{p}^{(k)})^T \mathbf{p}^{(k)} = (\mathbf{r}^{(k)})^T \mathbf{p}^{(k)} - \alpha_k (\mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(k)} = 0 \quad (2.80)$$

ossia ad ogni passo il residuo  $\mathbf{r}^{(k+1)}$  è ortogonale al vettore direzione  $\mathbf{p}^{(k)}$ . Dalla (2.80) e dalla definizione (2.75) si ricava

$$(\mathbf{p}^{(k)})^T \mathbf{r}^{(k)} = (\mathbf{r}^{(k)})^T \mathbf{r}^{(k)} + \beta_k (\mathbf{p}^{(k-1)})^T \mathbf{r}^{(k)} = (\mathbf{r}^{(k)})^T \mathbf{r}^{(k)} \quad (2.81)$$

Si vede quindi che la definizione (2.78) è equivalente alla seguente

$$\alpha_k = \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(k)}} \quad (2.82)$$

Procedendo in modo analogo, si può ottenere la seguente definizione per  $\beta_k$  equivalente alla (2.77)

$$\beta_k = \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{r}^{(k-1)})^T \mathbf{r}^{(k-1)}} \quad (2.83)$$

Indicato con  $\mathcal{S}_k$  lo spazio generato dai  $k$  vettori  $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)}$ , si può dimostrare che le direzioni  $\mathbf{p}^{(i)}$ ,  $i = 0, \dots, k-1$  sono tali che

$$F(\mathbf{x}^{(k)}) = \min_{\mathbf{x} \in \mathcal{S}_k} F(\mathbf{x})$$

Da tale risultato, osservando che le direzioni  $\mathbf{p}^{(i)}$  sono linearmente indipendenti (cfr. Capitolo 5), si ricava che il metodo del gradiente coniugato determina la soluzione  $\mathbf{x}$  in al più  $n$  passi ( $n$  ordine della matrice  $\mathbf{A}$ ), cioè esiste un  $m \leq n$  tale che  $\mathbf{r}^{(m)} = 0$ . Tale proprietà può anche essere ricavata direttamente dal seguente risultato.

**Teorema 2.9** *Supponiamo che  $\mathbf{r}^{(0)} \neq 0$  e che per un  $h$  fissato, con  $h \geq 1$ , si abbia  $\mathbf{r}^{(k)} \neq 0$  per ogni  $k \leq h$ . Allora*

$$\begin{cases} (\mathbf{r}^{(k)})^T \mathbf{r}^{(j)} = 0 \\ (\mathbf{p}^{(k)})^T \mathbf{A} \mathbf{p}^{(j)} = 0 \end{cases} \quad \text{per } k \neq j \text{ e } k, j = 0, \dots, h \quad (2.84)$$

*In altre parole i primi  $h$  vettori residui costituiscono un insieme di vettori ortogonali e i vettori  $\mathbf{p}^{(k)}$  costituiscono un insieme di vettori  $A$ -coniugati.*

Dal momento che l'insieme dei primi  $h$  residui è formato da vettori ortogonali, non possono esistere più di  $n$  vettori  $\mathbf{r}^{(k)} \neq 0$  e quindi esiste un  $m \leq n$  tale che  $\mathbf{r}^{(m)} = 0$ . In maniera più precisa, si può dimostrare che se la matrice  $\mathbf{A}$  possiede al più  $p \leq n$  autovalori distinti, allora il metodo del gradiente coniugato converge in al più  $p$  iterazioni.

Riassumendo, il metodo del gradiente coniugato può essere descritto nella seguente forma particolare, nota come *metodo di Hestenes-Stiefel* (1952).

**Algoritmo 2.21** (Metodo del gradiente coniugato) *Data una matrice  $\mathbf{A}$  simmetrica definita positiva e di ordine  $n$ , l'algoritmo genera una successione di vettori  $\mathbf{x}^{(k)}$ , che converge alla soluzione del sistema  $\mathbf{Ax} = \mathbf{b}$ , in al più  $n$  iterazioni.*

```

 $\mathbf{x}^{(0)}$ , arbitrario
 $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - \mathbf{Ax}^{(0)}$ 
For  $k = 0, 1, \dots, n - 1$ 
  If  $\|\mathbf{r}^{(k)}\|_2 \leq \epsilon \|\mathbf{b}\|_2$ 
    then
      Set  $\mathbf{x} = \mathbf{x}^{(k)}$  Stop
    else
       $\mathbf{w}^{(k)} = \mathbf{Ap}^{(k)}$ 
       $\alpha_k = \frac{\|\mathbf{r}^{(k)}\|_2^2}{(\mathbf{p}^{(k)})^T \mathbf{w}^{(k)}}$ 
       $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ 
       $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{w}^{(k)}$ 
       $\beta_{k+1} = \frac{\|\mathbf{r}^{(k+1)}\|_2^2}{\|\mathbf{r}^{(k)}\|_2^2}$ 
       $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_{k+1} \mathbf{p}^{(k)}$ 
    end if
  end  $k$ 

```

Il passo dell'algoritmo che presenta un costo computazionale maggiore è quello relativo alla moltiplicazione  $\mathbf{Ap}$ , che richiede  $n^2$  moltiplicazioni. Se il numero di iterazioni è  $n$  si ha, allora, un costo complessivo dell'ordine di  $n^3$ , superiore a quello richiesto dal metodo di Cholesky ( $n^3/3$ ). Si intuisce, pertanto, che il metodo diventa interessante quando la matrice è sparsa e/o quando il numero di iterazioni necessarie è decisamente inferiore a  $n$ .

La seguente implementazione in MATLAB mette in evidenza l'aspetto vettoriale dell'algoritmo.

```

% x stima iniziale, kmax numero massimo di iterazioni
% eps precisione richiesta
r=b-a*x
rn=norm(r,2)
p=r;
for k=0:kmax
  if rn < eps
    x, rn % stampa della soluzione e della norma del residuo
    break
  else
    w=a*p;
    rni=1/rn^2;

```

```

al=rn^2/(w'*p);
x=x+al*p;
r=r-al*w;
rn=norm(r,2);
be=rn^2*rni;
p=r+be*p;
end
end

```

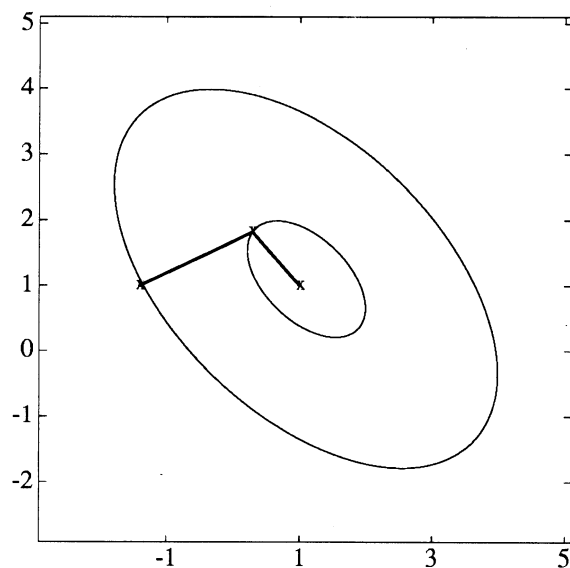


Figura 2.12: Illustrazione dei risultati ottenuti mediante l'algoritmo del gradiente coniugato nella risoluzione di un sistema lineare del secondo ordine. La soluzione viene ottenuta, a partire dal punto  $\mathbf{x}^{(0)} = [-1.5, 1]^T$  mediante due minimizzazioni lungo due direzioni coniugate.

► **Esempio 2.26** In Figura 2.12 sono illustrati i risultati che si ottengono applicando, a partire dal punto  $\mathbf{x}^{(0)} = [-1.5, 1]^T$ , il metodo del gradiente coniugato al sistema  $\mathbf{Ax} = \mathbf{b}$ , con

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

La direzione  $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$  è la direzione del gradiente nel punto  $\mathbf{x}^{(0)}$  e risulta ortogonale alla tangente in tale punto all'ellisse. La direzione successiva  $\mathbf{p}^{(1)}$  è A-coniugata a  $\mathbf{p}^{(0)}$  e passa per il centro dell'ellisse. La minimizzazione lungo tale direzione fornisce quindi il punto di minimo della  $F(x)$ , ossia la soluzione del sistema lineare dato. ■

**Limitazione dell'errore** Dal momento che il metodo del gradiente coniugato è utilizzato in pratica come un metodo iterativo, è interessante stabilire delle maggiorazioni dell'errore. Ricordiamo, in questo senso, la seguente maggiorazione che mette in evidenza l'influenza del numero di condizionamento della matrice  $\mathbf{A}$

$$\|\bar{\mathbf{x}} - \mathbf{x}^{(k)}\|_{\mathbf{A}} \leq 2 \left( \frac{\sqrt{\mu_2(\mathbf{A})} - 1}{\sqrt{\mu_2(\mathbf{A})} + 1} \right)^k \|\bar{\mathbf{x}} - \mathbf{x}^{(0)}\|_{\mathbf{A}} \quad (2.85)$$

ove  $\bar{\mathbf{x}}$  indica la soluzione esatta del sistema  $\mathbf{Ax} = \mathbf{b}$  e  $\|\mathbf{x}\|_{\mathbf{A}}$  indica la norma  $\sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}}$ .

Dalla maggiorazione (2.85) si vede che il metodo è tanto più efficiente quanto più  $\mu_2(\mathbf{A})$  è vicino a 1. Questo aspetto è illustrato dal seguente esempio.

► **Esempio 2.27** Consideriamo la risoluzione del sistema lineare  $\mathbf{Ax} = \mathbf{b}$  con

$$\mathbf{A} = \begin{bmatrix} d & -1 & & & \\ -1 & d & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & d \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad d \geq 2, \quad \mathbf{b} = \begin{bmatrix} d-1 \\ d-2 \\ \vdots \\ d-2 \\ d-1 \end{bmatrix}$$

La soluzione esatta è data da  $\mathbf{x} = [1, 1, \dots, 1]^T$ . In Figura 2.13 sono riportati i risultati ottenuti per  $n = 50$  e per i valori  $d = 2$ ,  $d = 3$ ,  $d = 4$ , in corrispondenza ai quali la matrice  $\mathbf{A}$  ha i seguenti numeri di condizionamento

$d$	$\mu_2(\mathbf{A})$	iterazioni
2	1053.5	26
3	4.9	18
4	2.9	14

ove sono riportati anche i numeri di iterazioni per ottenere un errore minore di  $\epsilon = 10^{-6}$ .

### 2.3.5 Precondizionamento

La maggiorazione (2.85) mette in evidenza l'importanza del *numero di condizionamento* per quanto concerne la *rapidità di convergenza* dei metodi di discesa. L'idea del *precondizionamento* consiste nel sostituire la risoluzione dell'equazione  $\mathbf{Ax} = \mathbf{b}$  con quella del sistema equivalente  $\mathbf{C}^{-1}\mathbf{Ax} = \mathbf{C}^{-1}\mathbf{b}$ , ove la matrice  $\mathbf{C}^{-1}$  è scelta con l'obiettivo di avere  $\mu_2(\mathbf{C}^{-1}\mathbf{A}) \ll \mu_2(\mathbf{A})$ . Naturalmente, la scelta migliore sarebbe  $\mathbf{C}^{-1} = \mathbf{A}^{-1}$ , perchè in questo caso si avrebbe  $\mu_2(\mathbf{C}^{-1}\mathbf{A}) = 1$ ; nella pratica, si tratta, al solito, di trovare un opportuno compromesso con il *costo* del calcolo di  $\mathbf{C}^{-1}$ .

Assumendo  $\mathbf{C}^{-1}$  simmetrica definita positiva, osserviamo che la matrice  $\mathbf{C}^{-1}\mathbf{A}$  non è, in generale, simmetrica e quindi non possiamo applicare l'algoritmo del gradiente coniugato direttamente a  $\mathbf{C}^{-1}\mathbf{A}$ . Tuttavia, si può definire una matrice  $\mathbf{C}^{-1/2}$  simmetrica e definita positiva tale che  $(\mathbf{C}^{-1/2})^2 = \mathbf{C}^{-1}$ ; di conseguenza, la matrice

$$\mathbf{C}^{1/2}(\mathbf{C}^{-1}\mathbf{A})\mathbf{C}^{-1/2} = \mathbf{C}^{-1/2}\mathbf{A}\mathbf{C}^{-1/2}$$

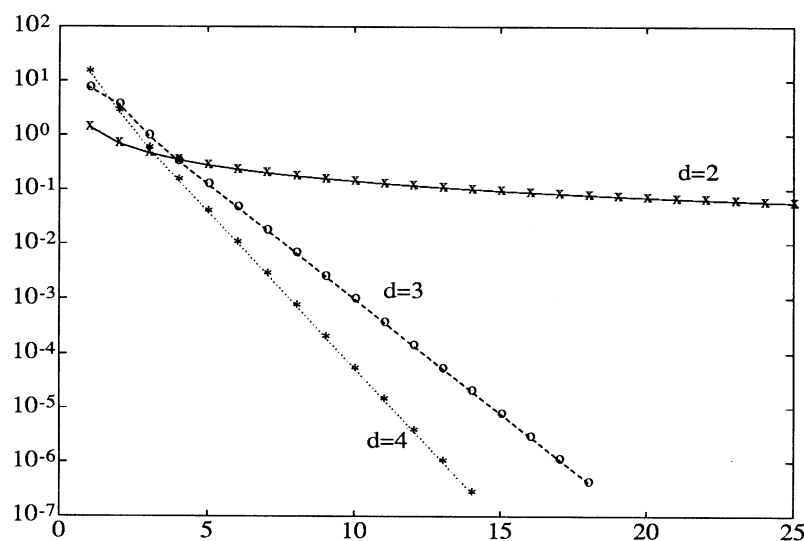


Figura 2.13: Illustrazione del comportamento del metodo del gradiente coniugato in dipendenza dal numero di condizionamento della matrice. Il metodo è applicato a un sistema con matrice tridiagonale  $\mathbf{A}$  di ordine  $n = 50$ , con  $a_{ii} = d$  e  $a_{i,i+1} = a_{i,i-1} = -1$ . Sono riportate le norme dei residui in corrispondenza ai valori  $d = 2$ ,  $d = 3$ ,  $d = 4$ .

è simmetrica definita positiva. Allora, invece di considerare il sistema  $\mathbf{C}^{-1}\mathbf{A}\mathbf{x} = \mathbf{C}^{-1}\mathbf{b}$  si può considerare il seguente sistema equivalente

$$\mathbf{C}^{1/2}(\mathbf{C}^{-1}\mathbf{A})\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\mathbf{x} = \mathbf{C}^{-1/2}\mathbf{b}$$

e, posto  $\mathbf{y} = \mathbf{C}^{1/2}\mathbf{x}$ , il problema diventa quello di trovare  $\mathbf{y}$  tale che

$$\mathbf{C}^{1/2}(\mathbf{C}^{-1}\mathbf{A})\mathbf{C}^{-1/2}\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{b}$$

Si può quindi applicare il metodo del gradiente coniugato alla nuova matrice

$$\tilde{\mathbf{A}} = \mathbf{C}^{-1/2}\mathbf{A}\mathbf{C}^{-1/2}$$

Si minimizza, cioè, il funzionale

$$\tilde{E}(\mathbf{y}) = (\tilde{\mathbf{A}}(\mathbf{y} - \bar{\mathbf{y}}), \mathbf{y} - \bar{\mathbf{y}})$$

ove  $\bar{\mathbf{y}}$  è la soluzione del sistema  $\tilde{\mathbf{A}}\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{b}$ .

Poiché il vettore a cui si è interessati è  $\bar{\mathbf{x}}$ , si può organizzare convenientemente l'algoritmo utilizzando le seguenti identità

$$\begin{aligned} \tilde{\mathbf{r}}^{(k)} &= \mathbf{C}^{-1/2}\mathbf{b} - \tilde{\mathbf{A}}\mathbf{y}^{(k)} = \mathbf{C}^{-1/2}\mathbf{b} - \mathbf{C}^{-1/2}\mathbf{A}\mathbf{C}^{-1/2}\mathbf{y}^{(k)} \\ &= \mathbf{C}^{-1/2}\mathbf{b} - \mathbf{C}^{-1/2}\mathbf{A}\mathbf{x}_k = \mathbf{C}^{-1/2}\mathbf{r}_k \\ \tilde{\mathbf{p}}^{(k)} &= \mathbf{C}^{1/2}\mathbf{p}^{(k)} \\ \mathbf{y}^{(k)} &= \mathbf{C}^{1/2}\mathbf{x}^{(k)} \end{aligned}$$

Si ha, allora, la seguente forma dell'algoritmo.

Data una matrice  $\mathbf{A}$  simmetrica definita positiva e di ordine  $n$ , e una matrice simmetrica definita positiva  $\mathbf{C}$ , l'algoritmo genera una successione di vettori  $\mathbf{x}^{(k)}$ , che converge alla soluzione del sistema  $\mathbf{Ax} = \mathbf{b}$ , in al più  $n$  iterazioni.

```

 $\mathbf{x}^{(0)}$ , arbitrario
 $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{Ax}^{(0)}$ 

 $\mathbf{Cp}^{(0)} = \mathbf{r}^{(0)}$ 
 $\mathbf{z}^{(0)} = \mathbf{p}^{(0)}$ 

for  $k = 0, 1, \dots, n - 1$ 
  if  $\|\mathbf{r}^{(k)}\|_2 < \epsilon \mathbf{b}$ 
    then
      set  $\mathbf{x} = \mathbf{x}^{(k)}$  stop
    else
       $\mathbf{w}^{(k)} = \mathbf{Ap}^{(k)}$ 
       $\alpha_k = \frac{(\mathbf{z}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{p}^{(k)})^T \mathbf{w}^{(k)}}$ 
       $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ 
       $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{w}^{(k)}$ 
       $\mathbf{Cz}^{(k+1)} = \mathbf{r}^{(k+1)}$ 
       $\beta_{k+1} = \frac{(\mathbf{z}^{(k+1)})^T \mathbf{r}^{(k+1)}}{(\mathbf{z}^{(k)})^T \mathbf{r}^{(k)}}$ 
       $\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} + \beta_{k+1} \mathbf{p}^{(k)}$ 
    end if
  end k

```

Sottolineiamo il fatto che ad ogni iterazione si deve risolvere un sistema del tipo

$$\mathbf{Cz} = \mathbf{r}$$

► **Esempio 2.28** Come illustrazione, consideriamo l'applicazione dell'idea del preconditionamento alla risoluzione del sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{A}$  matrice della seguente forma

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 0 & \dots & -1 \\ -1 & 3 & -1 & & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & -1 & 3 & -1 \\ -1 & \dots & 0 & -1 & 3 \end{bmatrix}$$

che risulta simmetrica e definita positiva. Una scelta naturale della matrice di *precondizionamento*  $\mathbf{C}$  è data dalla matrice tridiagonale  $\mathbf{T}$  (cfr. Figura 2.14) con elementi diagonali uguali a 3 e gli elementi sulle sottodiagonali uguali a  $-1$ .

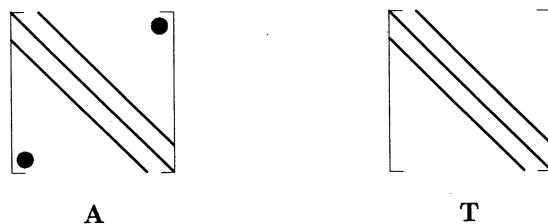


Figura 2.14: Matrice **A** e preconditionatore **T**.

Di seguito riportiamo l'implementazione in MATLAB del metodo per il sistema  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{b} = \mathbf{0}$ , e quindi con soluzione esatta  $\mathbf{x} = \mathbf{0}$ . Come punto di partenza, è scelto il vettore con componenti nulle, salvo la prima componente posta uguale a 1.

```
n=input('n')
a=zeros(n); % definizione della matrice A
for i=1:n
    x(i)=0.; % punto iniziale
    b(i)=0.; % termine noto
    for j=1:n
        if i==j
            a(i,j)=3;
        elseif abs(i-j)==1
            a(i,j)=-1;
        end
    end
end
end
t=a; % definizione della matrice T
a(1,n)=-1; a(n,1)=-1;
x(1)=1;
x=x'; b=b'
err=norm(x,1) % errore nella norma 1
r=b-a*x;
tp=inv(t); % calcolo dell'inversa di T
p=tp*r;
z=p;
for k=1:6
    w=a*p;
    pro=z'*r; % prodotto scalare
    al=pro/(w'*p);
    x=x+al*p;
    r=r-al*w;
    z=tp*r;
    bk=(r'*z)/pro;
```



```

p=z+bk*p;
err=norm(x,1);
end

```

Nella seguente tabella i risultati ottenuti, per  $n = 20$ , sono messi a confronto con quelli ottenuti senza tecnica di condizionamento, cioè per  $\mathbf{C} = \mathbf{I}$ .

k	$\mathbf{C} = \mathbf{T}$ $\ \mathbf{x}^{(k)}\ _1$	$\mathbf{C} = \mathbf{I}$ $\ \mathbf{x}^{(k)}\ _1$
0	1.	1.
1	0.688415	0.755555
2	0.157965	0.438095
3	1.2891 E-16	0.199044
4		0.080035
5		0.030952
6		0.011854

◆ **Esercizio 2.20** Confrontare i metodi di Jacobi e di Gauss-Seidel relativamente alla matrice

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 2/5 \\ 0 & 5 & 2/5 \\ 5/2 & 2 & 1 \end{bmatrix}$$

◆ **Esercizio 2.21** Studiare e confrontare la convergenza dei metodi di Jacobi, Gauss-Seidel e rilassamento per la matrice

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 0 & 0 & 0 & -1 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & 0 \\ 0 & -1 & 0 & -1 & 3 & -1 \\ -1 & 0 & 0 & 0 & -1 & 3 \end{bmatrix}$$

◆ **Esercizio 2.22** Data una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  singolare e  $\mathbf{A} = \mathbf{M} - \mathbf{N}$  una decomposizione tale che  $\mathbf{M}$  sia non singolare. Dimostrare che il metodo iterativo

$$\mathbf{x}^{(k+1)} = \mathbf{M}^{-1}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b}$$

per la risoluzione del sistema  $\mathbf{Ax} = \mathbf{b}$  non è convergente.

◆ **Esercizio 2.23** Dato il sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , con

$$\begin{bmatrix} 1 & \alpha & -\alpha \\ 1 & 1 & 1 \\ \alpha & \alpha & 1 \end{bmatrix}$$

determinare i valori di  $\alpha \in \mathbb{R}$  per i quali i metodi di Jacobi e di Gauss-Seidel sono convergenti. Inoltre, per i valori di  $\alpha$  per i quali entrambi i metodi sono convergenti, individuare quale dei due metodi presenta la velocità asintotica di convergenza maggiore.

Much of quantum mechanics boils down  
to looking for functions that are eigenfunctions of a given operator.  
**P. W. Atkins**

## Capitolo 3

# Autovalori e autovettori

Data una matrice  $\mathbf{A}$  di ordine  $n$ , viene detto *autovalore* di  $\mathbf{A}$  un numero  $\lambda \in \mathbb{C}$  per il quale il seguente sistema omogeneo

$$\mathbf{A} \mathbf{x} = \lambda \mathbf{x} \quad (3.1)$$

ammette soluzioni  $\mathbf{x} \neq 0$ . Tali soluzioni sono dette gli *autovettori* di  $\mathbf{A}$ , corrispondenti all'autovalore  $\lambda$ . In sostanza, un vettore non nullo  $\mathbf{x}$  è un autovettore quando nella trasformazione lineare  $\mathbf{x} \rightarrow \mathbf{A}\mathbf{x}$  esso viene trasformato in un multiplo di  $\mathbf{x}$ .

Le nozioni di autovalore e di autovettore sono importanti in diverse applicazioni. Segnaliamo, in particolare, il loro interesse nello studio della *stabilità delle strutture*, in questioni di *statistica*, nello studio di *sistemi dinamici* e della *propagazione dei segnali*, e in *chimica-fisica* nel calcolo degli orbitali molecolari. Rinviando all'Appendice A per gli elementi introduttivi e teorici, in questo capitolo analizzeremo, principalmente, gli *aspetti numerici*.

Il problema del calcolo degli autovalori è, in sostanza, un problema di tipo *non-lineare*, in quanto esso equivale alla ricerca delle radici della seguente equazione algebrica in  $\lambda$ , detta *equazione caratteristica*

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \quad (3.2)$$

Per la sua risoluzione, si potrebbero, quindi, utilizzare le tecniche che saranno analizzate nel successivo Capitolo 5. Tuttavia, la procedura matematica – costruzione del polinomio caratteristico, calcolo delle sue radici, e la risoluzione delle equazioni omogenee (3.1) – non è, in generale, una procedura numerica conveniente, sia per costo che per stabilità. Lo scopo principale di questo capitolo è, in effetti, quello di fornire metodi numerici alternativi. Il risultato alla base di tali metodi è *l'invarianza*

dello spettro di  $\mathbf{A}$ , ossia dell'insieme degli autovalori di  $\mathbf{A}$ , rispetto alle trasformazioni simili. Mediante una scelta opportuna di tali trasformazioni, la matrice di partenza viene ridotta ad una matrice di forma più semplice per quanto riguarda il calcolo degli autovalori, ad esempio una matrice diagonale o triangolare. I vari metodi differiscono tra loro per la diversa scelta e implementazione delle successive trasformazioni.

Come bibliografia utile per approfondire gli argomenti trattati in questo capitolo, segnaliamo ad esempio Golub e Van Loan [69]. Segnaliamo inoltre, come libreria di riferimento per il calcolo degli autovalori e autovettori, EISPACK [62], che raccoglie in forma integrata l'implementazione di numerosi algoritmi, generali e specializzati.

### 3.1 Condizionamento del problema degli autovalori

Data una matrice  $\mathbf{A}$ , per condizionamento degli autovalori si intende l'analisi di come si *propagano* sugli autovalori le perturbazioni (corrispondenti, ad esempio, agli errori di arrotondamento, o agli errori sperimentali) presenti negli elementi della matrice. Si tratta di un'analisi che prescinde dagli algoritmi utilizzati, ma che è indispensabile per la scelta dell'algoritmo più idoneo, nel senso che se il problema dato risulta essere malcondizionato, per la sua risoluzione saranno opportuni metodi *stabili*.

Constatiamo attraverso un esempio la possibilità che una matrice possa essere malcondizionata per quanto riguarda il calcolo dei suoi autovalori.

► **Esempio 3.1** Gli autovalori della seguente matrice

$$\mathbf{A} = \begin{bmatrix} 101 & 110 \\ -90 & -98 \end{bmatrix} \quad (3.3)$$

sono dati da  $\lambda_1 = 2$  e  $\lambda_2 = 1$ , e diventano  $\tilde{\lambda}_1 \approx 1. + 9.9499 i$ ,  $\tilde{\lambda}_2 \approx 1. - 9.9499 i$  per la matrice

$$\tilde{\mathbf{A}} = \begin{bmatrix} 100 & 110 \\ -90 & -98 \end{bmatrix}$$

ottenuta sostituendo all'elemento  $a_{11} = 101$  il valore  $\tilde{a}_{11} = 100$ . ■

Per studiare il problema del condizionamento, in particolare per introdurre, analogamente a quanto abbiamo visto per la risoluzione dei sistemi lineari, una *misura* del condizionamento, consideriamo per semplicità il caso di matrici *diagonalizzabili*, per le quali cioè esista una matrice  $\mathbf{P}$  non singolare, tale che

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{diag}[\lambda_1, \lambda_2, \dots, \lambda_n] \equiv: \mathbf{D}$$

Le colonne di  $\mathbf{P}$  sono gli autovettori di  $\mathbf{A}$ , corrispondenti agli autovalori  $\lambda_1, \dots, \lambda_n$ . Ad esempio, per la matrice (3.3) si ha

$$\begin{bmatrix} 101 & 110 \\ -90 & -98 \end{bmatrix} = \begin{bmatrix} -11 & 10 \\ 10 & -9 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 9 & 10 \\ 10 & 11 \end{bmatrix}$$

Ricordiamo, allora, il seguente risultato.

**Teorema 3.1** (Bauer-Fike) *Sia  $\mathbf{A}$  una matrice diagonalizzabile e  $\mathbf{A} + \mathbf{E}$  una perturbazione di  $\mathbf{A}$ , con  $\mathbf{E} \in \mathbb{R}^{n \times n}$ . Se  $\xi$  è un autovalore di  $\mathbf{A} + \mathbf{E}$ , allora esiste almeno un autovalore  $\lambda$  di  $\mathbf{A}$  tale che*

$$|\lambda - \xi| \leq K(\mathbf{A}) \|\mathbf{E}\| \quad (3.4)$$

ove  $K(\mathbf{A}) = \|\mathbf{P}\| \|\mathbf{P}^{-1}\|$  e  $\|\cdot\|$  indica una qualsiasi delle norme di matrice  $\|\cdot\|_p$ , con  $p = 1, 2, \infty$ .

Il numero  $K(\mathbf{A})$  può essere, quindi, assunto come *numero di condizionamento* (assoluto) del problema degli autovalori della matrice  $\mathbf{A}$ . Per la matrice (3.3) si ha, in effetti,  $K(\mathbf{A}) = 441$ , calcolato nella norma 1.

Ricordando che le matrici normali (ad esempio le matrici simmetriche, skew-simmetriche, ortogonali o unitarie) sono diagonalizzabili mediante matrici  $\mathbf{P}$  ortogonali (cfr. Appendice A), dal teorema precedente si ricava il seguente risultato.

**Corollario 3.1** *Se  $\mathbf{A}$  è una matrice normale e se  $\mathbf{A} + \mathbf{E}$  è una qualunque perturbazione di  $\mathbf{A}$ , allora per ogni autovalore  $\xi$  di  $\mathbf{A} + \mathbf{E}$  esiste almeno un autovalore  $\lambda$  di  $\mathbf{A}$  tale che*

$$|\lambda - \xi| \leq \|\mathbf{E}\|_2 \quad (3.5)$$

dal momento che  $\|\mathbf{P}\|_2 = 1$  se  $\mathbf{P}$  è ortogonale.

Il risultato (3.5) mostra che piccole perturbazioni (in senso assoluto) sulla matrice  $\mathbf{A}$  portano pure a piccole perturbazioni sugli autovalori. Pertanto per le matrici normali, in particolare quindi per le matrici *simmetriche* o *unitarie*, il calcolo degli autovalori è un problema *bencondizionato*, mentre, come evidenziato dall'esempio precedente, per una matrice generale il problema può essere *malcondizionato*. In effetti, il calcolo numerico degli autovalori di una matrice non simmetrica, o unitaria, è, in generale, un problema delicato.

## 3.2 Metodo delle potenze

Il *metodo delle potenze* è un metodo di tipo *iterativo* utile, in particolare, quando è richiesto il calcolo dell'autovalore di *modulo massimo* o, più in generale, di un autovalore *vicino* ad un valore prefissato<sup>1</sup>.

Per introdurre l'idea del metodo, consideriamo, per semplicità, il caso di una matrice  $\mathbf{A}$  di ordine  $n$  che abbia  $n$  autovettori  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  *linearmente indipendenti* e gli autovalori  $\lambda_1, \lambda_2, \dots, \lambda_n$  tali che

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$$

<sup>1</sup>Il metodo delle potenze, che utilizza le proprietà asintotiche delle potenze di matrici, è stato suggerito nel 1913 da Müntz.

Supponiamo, cioè, che l'autovalore di modulo massimo abbia molteplicità algebrica 1 e che non esistano altri autovalori con lo stesso modulo. Per le opportune estensioni del metodo rinviamo alla bibliografia.

Fissato un vettore  $\mathbf{z}_0 \in \mathbb{C}^n$ , con  $\mathbf{z}_0 \neq 0$ , si genera la successione di vettori  $\{\mathbf{y}_k\}$ ,  $k = 1, 2, \dots$  mediante la seguente *procedura ricorrente*

$$\begin{aligned} \mathbf{y}_0 &= \mathbf{z}_0 \\ \mathbf{y}_k &= \mathbf{A}\mathbf{y}_{k-1}, \quad k = 1, 2, \dots \end{aligned}$$

Esaminiamo il comportamento della successione  $\{\mathbf{y}_k\}$  per  $k \rightarrow \infty$ . Nelle ipotesi fatte su  $\mathbf{A}$ , il vettore  $\mathbf{z}_0$  ha la seguente rappresentazione

$$\mathbf{z}_0 = \sum_{i=1}^n \alpha_i \mathbf{x}_i$$

Supponendo che  $\mathbf{z}_0$  sia tale che  $\alpha_1 \neq 0$ , si ha

$$\mathbf{y}_k = \mathbf{A}^k \mathbf{z}_0 = \sum_{i=1}^n \alpha_i \lambda_i^k \mathbf{x}_i = \lambda_1^k \left[ \alpha_1 \mathbf{x}_1 + \sum_{i=2}^n \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i \right] \quad (3.6)$$

Dal momento che  $|\lambda_i/\lambda_1| < 1$ ,  $i \geq 2$ , la direzione del vettore  $\mathbf{y}_k$  tende a quella di  $\mathbf{x}_1$ . Inoltre, il *quoziente di Rayleigh* di  $\mathbf{y}_k$ , definito da

$$\sigma_k = \frac{(\mathbf{A}\mathbf{y}_k, \mathbf{y}_k)}{(\mathbf{y}_k, \mathbf{y}_k)}$$

tende all'autovalore  $\lambda_1$ .

Nella forma precedente l'algoritmo può presentare problemi di *underflow* o *overflow* dovuti al fatto che  $\lambda_1^k$  può tendere a zero o all'infinito; è necessaria, quindi, una operazione di *scaling*. Tale operazione può essere effettuata utilizzando a priori diversi tipi di norma di vettore. Se si considera ad esempio la *norma euclidea*, si ottiene il seguente *algoritmo*. Si costruisce una successione di vettori  $\mathbf{t}_k$ , con  $\|\mathbf{t}_k\|_2 = 1$ , mediante le relazioni

$$\begin{aligned} \mathbf{t}_k &= \frac{\mathbf{y}_k}{\|\mathbf{y}_k\|_2}, \quad \mathbf{y}_{k+1} = \mathbf{A}\mathbf{t}_k, \quad k = 0, 1, 2, \dots \\ \sigma_k &= \frac{(\mathbf{A}\mathbf{t}_k, \mathbf{t}_k)}{(\mathbf{t}_k, \mathbf{t}_k)} = (\mathbf{t}_k, \mathbf{y}_{k+1}) \end{aligned}$$

Da (3.6) si vede che gli *errori di troncamento* relativi a  $\mathbf{t}_k$  e  $\sigma_k$  tendono a zero come  $|\lambda_2/\lambda_1|^k$  (convergenza lineare). Lasciamo come *esercizio* mostrare che se la matrice  $\mathbf{A}$  è simmetrica, allora la convergenza è come  $|\lambda_2/\lambda_1|^{2k}$  (convergenza quadratica).

Nel seguito è riportato un esempio di implementazione dell'algoritmo.

```

PARAMETER (LA=50, ITMAX=50, EPS=1.E-5)
DIMENSION A(LA,LA),T(LA),Y(LA)
PRINT*, 'n= '
READ*,N
PRINT*, 'introduci matrice per righe'
DO 10 I=1,N
  PRINT*,I
  READ*,(A(I,J),J=1,N)
10 CONTINUE
PRINT*, 'stima iniziale '
READ*,(Y(I),I=1,N)
C   inizializzazione

CALL POW(LA,N,A,Y,T,SOLD)
C   iterazione
DO 50 IT=1,ITMAX
CALL POW(LA,N,A,Y,T,S)
PRINT*,IT,S
C   test d'arresto
IF(ABS(S-SOLD) .LE. (ABS(S)+1)*EPS)STOP
SOLD=S
50 CONTINUE
PRINT*, 'raggiunto numero massimo di iterazioni'
END

SUBROUTINE POW(LA,N,A,Y,T,S)
C   effettua una iterazione
C   S nuova approssimazione dell'autovalore di modulo massimo
C   T approssimazione del corrispondente autovettore
DIMENSION A(LA,*),Y(*),T(*)
C   calcolo della norma
P=(PS(N,Y,Y))**0.5
DO 20 I=1,N
20 T(I)=Y(I)/P
C   prodotto y=At
CALL MATVEC(LA,N,A,T,Y)
C   prodotto scalare (t,y)
S=PS(N,T,Y)
RETURN
END

REAL FUNCTION PS(N,X,Y)
C   prodotto scalare
REAL X(*),Y(*)
PS=0.
DO 10 I = 1,N
10 PS=PS+X(I)*Y(I)
RETURN
END

```

```

SUBROUTINE MATVEC(LA,N,A,X,Y)
C   prodotto y=Ax
REAL A(LA,*),X(*),Y(*)
DO 20 I=1,N
  S=0.
  DO 10 J=1, N
10   S=S+A(I,J)*X(J)
  Y(I)=S
20  CONTINUE
RETURN
END

```

Quando applicato alla seguente matrice

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 2 \\ -2 & 0 & 5 \\ 6 & -3 & 6 \end{bmatrix}$$

a partire dal valore iniziale  $(1/\sqrt{3}) [1, 1, 1]^T$ , il metodo fornisce i risultati riportati in Tabella 3.1. Gli autovalori esatti sono 5., 3., -1. Quando applicato alla seguente

k	$\sigma_k$	$\mathbf{t}_k$		
1	4.97372	0.203061	0.653211	0.729439
2	4.98854	0.202811	0.651639	0.730914
3	4.99223	0.203065	0.651281	0.731162
4	4.99551	0.203137	0.650947	0.731440
5	4.99727	0.203196	0.650769	0.731581
6	4.99837	0.203229	0.650658	0.731671
7	4.99902	0.203249	0.650592	0.731724
8	4.99941	0.203261	0.650552	0.731756
9	4.99965	0.203268	0.650528	0.731775
10	4.99979	0.203272	0.650514	0.731787
11	4.99987	0.203275	0.650505	0.731794
12	4.99992	0.203277	0.650500	0.731798

Tabella 3.1: Risultati ottenuti mediante il metodo delle potenze.

matrice *simmetrica*

$$\mathbf{A} = \begin{bmatrix} 6 & 4 & 4 & 1 \\ 4 & 6 & 1 & 4 \\ 4 & 1 & 6 & 4 \\ 1 & 4 & 4 & 6 \end{bmatrix}$$

a partire dal valore iniziale  $(1/\sqrt{4}) [1, 1, 1, 1]^T$ , il metodo fornisce i risultati riportati in Tabella 3.2. Gli autovalori esatti sono  $\lambda_1 = 15$ ,  $\lambda_2 = \lambda_3 = 5$  e  $\lambda_4 = -1$ .

▼ **Osservazione 3.1** *Il metodo delle potenze è convergente anche nel caso in cui l'autovalore di modulo massimo abbia molteplicità algebrica maggiore di 1, cioè  $\lambda_1 = \lambda_2 = \dots = \lambda_r$ ,*

k	$\sigma_k$	$\mathbf{t}_k$			
1	14.8276	0.413481	0.486577	0.486577	0.596254
2	14.9811	0.468218	0.500143	0.500143	0.529610
3	14.9979	0.489748	0.499907	0.499907	0.510229
4	14.9998	0.496578	0.499997	0.499997	0.503405
5	15.0000	0.498862	0.499999	0.499999	0.501137
6	15.0000	0.499621	0.500000	0.500000	0.500379

Tabella 3.2: Risultati ottenuti mediante il metodo delle potenze per una matrice simmetrica.

con

$$|\lambda_1| = |\lambda_2| = \dots = |\lambda_r| > |\lambda_{r+1}| \geq \dots \geq |\lambda_n|$$

Quando, invece, l'autovalore di modulo massimo non è unico, i risultati di convergenza precedenti non sono più applicabili. In certi casi, tuttavia, è ancora possibile determinare delle relazioni che permettono il calcolo degli autovalori dominanti. Si consideri, come esercizio, il caso di una matrice  $\mathbf{A}$  con autovalori reali  $\lambda_1 = -\lambda_2 > 0$  e  $\lambda_1 > |\lambda_3| \geq |\lambda_4| \geq \dots \geq |\lambda_n|$  e con  $n$  autovettori linearmente indipendenti. In questo caso si può vedere che  $\lambda_1^2$  è approssimato dalla successione  $(\mathbf{A}^2 \mathbf{y}_k)_j / (\mathbf{y}_k)_j$ , ove  $(\mathbf{y}_k)_j$  è una componente di  $\mathbf{y}_k$  diversa dallo zero. ■

### Metodo di Bernoulli

Ad ogni polinomio può essere associata una matrice (nella forma di Frobenius, o companion matrix, cfr. Appendice A), che ha come autovalori gli zeri del polinomio. L'applicazione del metodo delle potenze alla matrice *companion* equivale al metodo di Bernoulli per il calcolo della radice di modulo massimo del polinomio. Ricordiamo che il *metodo di Bernoulli*<sup>2</sup> per un polinomio  $P(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_n$ , con  $a_0 \neq 0$ , consiste nel generare una successione  $\{z_k\}$  nel seguente modo.

Si assegnano *arbitrariamente*  $n$  valori  $z_0, z_1, \dots, z_{n-1}$ ; quindi si calcola

$$z_{n+k} = -\frac{1}{a_0} (a_1 z_{n+k-1} + a_2 z_{n+k-2} + \dots + a_n z_k)$$

Si può dimostrare che, quando la radice di modulo massimo  $\lambda_1$  è unica, allora

$$\lim_{k \rightarrow \infty} \frac{z_{n+k}}{z_{n+k-1}} = \lambda_1$$

#### 3.2.1 Iterazione inversa

Quando il metodo delle potenze è applicato alla matrice  $\mathbf{A}^{-1}$ , esso fornisce, nel caso in cui si abbia  $0 < |\lambda_n| < |\lambda_{n-1}|$ , una approssimazione dell'autovalore di modulo

<sup>2</sup>descritto da Daniel Bernoulli nel 1728, noto anche come *metodo dei momenti*.



minimo. Più in generale, la successione

$$\mathbf{t}_k = \frac{\mathbf{y}_k}{\|\mathbf{y}_k\|_2}, \quad (\mathbf{A} - \lambda^* \mathbf{I})\mathbf{y}_{k+1} = \mathbf{t}_k, \quad k = 0, 1, 2, \dots$$

corrisponde<sup>3</sup> all'applicazione del metodo delle potenze alla matrice  $(\mathbf{A} - \lambda^* \mathbf{I})^{-1}$ , che ha come autovalori  $1/(\lambda_i - \lambda^*)$ . In questo caso la successione converge all'autovalore più vicino a  $\lambda^*$ . Osserviamo che il calcolo della successione  $\{\mathbf{y}_k\}$  richiede la *risoluzione, ad ogni iterazione, di un sistema lineare*. Poiché tuttavia la matrice è sempre la stessa, si può effettuare la decomposizione **LU** della matrice  $\mathbf{A} - \lambda^* \mathbf{I}$  e quindi risolvere, ad ogni iterazione, due sistemi triangolari.

### 3.2.2 Deflazione

Con il termine *deflazione* si intende l'operazione di eliminazione dell'autovalore  $\lambda_1$ , una volta che sia stata ottenuta una sua approssimazione. In pratica, essa consiste nella costruzione di una matrice  $\mathbf{A}'$  che abbia gli stessi autovalori di  $\mathbf{A}$  salvo  $\lambda_1$ . Si tratta di una operazione *numericamente delicata*, in quanto è necessario limitare la propagazione degli errori che derivano dal fatto che dell'autovalore  $\lambda_1$  si conosce solo, in generale, un valore approssimato. Una procedura *stabile* può essere ottenuta nel modo seguente. Mediante le *trasformazioni di Householder* si costruisce una matrice non singolare  $\mathbf{P}$  tale che  $\mathbf{P}\mathbf{x}_1 = \mathbf{e}_1$ , ove con  $\mathbf{e}_1$  si è indicato il primo vettore unitario e con  $\mathbf{x}_1$  un autovettore corrispondente a  $\lambda_1$ . Allora, da  $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$  si ha

$$\mathbf{P}\mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{x}_1 = \lambda_1\mathbf{P}\mathbf{x}_1 \Rightarrow (\mathbf{P}\mathbf{A}\mathbf{P}^{-1})\mathbf{e}_1 = \lambda_1\mathbf{e}_1$$

La matrice  $\mathbf{P}\mathbf{A}\mathbf{P}^{-1}$  ha gli stessi autovalori di  $\mathbf{A}$  e quindi deve avere la seguente forma

$$\mathbf{P}\mathbf{A}\mathbf{P}^{-1} = \begin{bmatrix} \lambda_1 & \mathbf{b}^T \\ 0 & \mathbf{A}' \end{bmatrix}$$

La matrice  $\mathbf{A}'$  è la matrice cercata.

### 3.2.3 Metodo di Lanczos

Il metodo di Lanczos<sup>4</sup> è una tecnica particolarmente conveniente per matrici *simmetriche* (o hermitiane) *sparse* e di *grandi dimensioni*. Il metodo genera una successione di matrici tridiagonali  $\mathbf{T}_j$ , di dimensioni  $j \times j$ , con la proprietà che gli autovalori *estremali* (ossia gli autovalori “più” positivi, e rispettivamente “più” negativi) di  $\mathbf{T}_j$  convergono monotonamente ai corrispondenti autovalori estremali di  $\mathbf{A}$ . In questo

<sup>3</sup>Tale estensione del metodo delle potenze è stata proposta da Wielandt nel 1944 ed è nota come metodo di iterazione inversa.

<sup>4</sup>C. Lanczos, *An iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators*, J. Res. Nat. Bur. Stand., 45, 1950.

modo è possibile avere informazioni sugli autovalori, calcolando gli autovalori di matrici tridiagonali di ordine relativamente basso. Si presuppone, naturalmente, che il calcolo degli autovalori di una matrice tridiagonale sia un problema più semplice (per tale problema cfr. paragrafi successivi). Nel seguito daremo una idea del metodo, rinviando per una trattazione più dettagliata alla bibliografia (cfr. in particolare Golub e Van Loan [69]).

Data una *matrice simmetrica*  $\mathbf{A}$ , si costruisce una matrice ortogonale  $\mathbf{Q}$  con la proprietà che

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{T} \quad (3.7)$$

con  $\mathbf{T}$  matrice simmetrica *tridiagonale*. Tale obiettivo è comune allo schema di Householder che analizzeremo nei successivi paragrafi. La differenza tra i due metodi consiste nella procedura utilizzata per raggiungere l'obiettivo. Mentre, come vedremo, nel metodo di Householder la matrice  $\mathbf{T}$  è ottenuta mediante successive premoltiplicazioni e postmoltiplicazioni di matrici elementari di tipo riflessione, ossia matrici di Householder, (o alternativamente, nel metodo di Givens con metodi di rotazione), nello schema di Lanczos le colonne di  $\mathbf{Q}$  e gli elementi di  $\mathbf{T}$  sono ottenuti mediante una procedura ricorrente. Moltiplicando (3.7) per  $\mathbf{Q}$ , si ottiene  $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{T}$ , e quindi ogni colonna della matrice  $\mathbf{A}\mathbf{Q}$  è uguale alla corrispondente colonna della matrice  $\mathbf{Q}\mathbf{T}$ . Se  $\mathbf{q}_j$  denota la colonna  $j$ -ma di  $\mathbf{Q}$ , la colonna  $j$ -ma di  $\mathbf{A}\mathbf{Q}$  è data da  $\mathbf{A}\mathbf{q}_j$ . Scritta la matrice  $\mathbf{T}$  nella seguente forma

$$\mathbf{T} = \begin{bmatrix} d_1 & u_1 & & & & \\ u_1 & d_2 & u_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & u_{n-2} & d_{n-1} & u_{n-1} & \\ & & & u_{n-1} & d_n & \end{bmatrix}$$

la colonna  $j$ -ma di  $\mathbf{Q}\mathbf{T}$  può essere espressa nella seguente forma

$$u_{j-1}\mathbf{q}_{j-1} + d_j\mathbf{q}_j + u_j\mathbf{q}_{j+1}$$

Uguagliando, quindi, la colonna  $j$ -ma di  $\mathbf{A}\mathbf{Q}$  alla colonna  $j$ -ma di  $\mathbf{Q}\mathbf{T}$ , si ottiene

$$\mathbf{A}\mathbf{q}_j = u_{j-1}\mathbf{q}_{j-1} + d_j\mathbf{q}_j + u_j\mathbf{q}_{j+1} \quad (3.8)$$

Assumendo, per convenzione, come vettore  $\mathbf{q}_0$  il vettore nullo, la relazione precedente è valida anche per  $j = 1$ . Dal momento che  $\mathbf{Q}$  è una matrice ortogonale, si ha

$$\mathbf{q}_j^T \mathbf{q}_{j-1} = \mathbf{q}_j^T \mathbf{q}_{j+1} = 0, \quad \mathbf{q}_j^T \mathbf{q}_j = 1$$

Premoltiplicando (3.8) per  $\mathbf{q}_j^T$ , si ottiene

$$d_j = \mathbf{q}_j^T \mathbf{A}\mathbf{q}_j$$

Definendo i vettori  $\mathbf{r}_j$  nel seguente modo

$$\mathbf{r}_j = \mathbf{A}\mathbf{q}_j - d_j\mathbf{q}_j - u_{j-1}\mathbf{q}_{j-1}$$

la relazione (3.8) implica che  $u_j\mathbf{q}_{j+1} = \mathbf{r}_j$ . Prendendo la norma euclidea, si ha

$$|u_j| \|\mathbf{q}_{j+1}\|_2 = \|\mathbf{r}_j\|_2$$

da cui si vede che  $u_j$  è  $\pm$  la lunghezza euclidea di  $\mathbf{r}_j$ . Risolvendo l'identità  $u_j\mathbf{q}_{j+1} = \mathbf{r}_j$ , si ottiene  $\mathbf{q}_{j+1} = \mathbf{r}_j/u_j$ . Le formule precedenti per  $d_j$ ,  $u_j$  e  $\mathbf{q}_{j+1}$  costituiscono il metodo di Lanczos per ridurre una matrice simmetrica  $\mathbf{A}$  di ordine  $n$  a una matrice tridiagonale. In conclusione, partendo da un vettore arbitrario, non nullo,  $\mathbf{q}_1$ , il metodo può essere riassunto nel seguente modo

```

 $\mathbf{q}_0 = 0, \mathbf{r} = \mathbf{q}_1, u_0 = \|\mathbf{r}\|_2$ 
for  $j = 1, \dots, n$ 
   $\mathbf{q}_j = \mathbf{r}/u_{j-1}, d_j = \mathbf{q}_j^T \mathbf{A} \mathbf{q}_j$ 
   $\mathbf{r} = (\mathbf{A} - d_j \mathbf{I}) \mathbf{q}_j - u_{j-1} \mathbf{q}_{j-1}$ 
   $u_j = \|\mathbf{r}\|_2$ 
end  $j$ 

```

Quando  $u_{j-1} = 0$ , e quindi l'algoritmo precedente non è in grado di procedere, si definisce  $\mathbf{q}_j$  arbitrariamente come un vettore di lunghezza unitaria e ortogonale ai vettori  $\mathbf{q}_1, \dots, \mathbf{q}_{j-1}$ .

Si può vedere facilmente che i vettori  $\mathbf{q}_j$  giacciono nello spazio generato dai vettori  $\mathbf{q}_1, \mathbf{A}\mathbf{q}_1, \dots, \mathbf{A}^{j-1}\mathbf{q}_1$ . Tale spazio, indicato usualmente con la notazione  $\mathcal{K}^j(\mathbf{q}_1)$ , è chiamato lo *spazio  $j$ -mo di Krylov* associato con  $\mathbf{q}_1$ . In sostanza, il metodo di Lanczos genera una base ortonormale per gli spazi di Krylov.

Chiariamo, ora, come il metodo di Lanczos possa essere utilizzato per calcolare alcuni degli autovalori estremi, ad esempio l'autovalore dominante, di una matrice simmetrica  $\mathbf{A}$ . Per ogni  $j$  fissato,  $j = 1, 2, \dots, n$ , consideriamo la matrice  $\mathbf{T}_j$  che ha come diagonale gli elementi  $d_1, d_2, \dots, d_j$  e con elementi sopra la diagonale  $u_1, u_2, \dots, u_j$  calcolati con l'algoritmo precedente. Si può mostrare che gli autovalori estremi (in particolare, quindi, l'autovalore dominante  $\lambda_{max}^{(j)}$ ) di  $\mathbf{T}_j$  convergono (in maniera più rapida che nel metodo delle potenze) agli autovalori estremi della matrice  $\mathbf{A}$ . Sfruttando tale risultato, è possibile ottenere delle approssimazioni degli autovalori estremi di  $\mathbf{A}$ , risolvendo successivamente il problema del calcolo degli autovalori estremi di matrici tridiagonali di dimensione crescente.

► **Esempio 3.2** Come illustrazione, consideriamo l'applicazione del metodo di Lanczos alla seguente matrice simmetrica

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 2 & 1 & 2 & -1 & 3 \\ 1 & 2 & 0 & 3 & 1 \\ 2 & -1 & 3 & 1 & 0 \\ 0 & 3 & 1 & 0 & 4 \end{bmatrix} \quad (3.9)$$

con  $\mathbf{q}_1 = [1, 1, 1, 1, 1]^T$ . Si ottengono successivamente i seguenti risultati, ove con  $\Lambda_j$  sono indicati gli autovalori della matrice  $\mathbf{T}_j$ .

$$\begin{aligned}
 j = 1 \quad \mathbf{T}_1 &= [6.600000]; & \Lambda_1 &= [6.600000] \\
 j = 2 \quad \mathbf{T}_2 &= \begin{bmatrix} 6.600000 & 1.019803 \\ 1.019803 & 3.053846 \end{bmatrix}; & \Lambda_2 &= \begin{bmatrix} 6.872357 \\ 2.781488 \end{bmatrix} \\
 j = 3 \quad \mathbf{T}_3 &= \begin{bmatrix} 6.600000 & 1.019803 & 0 \\ 1.019803 & 3.053846 & 3.179771 \\ 0 & 3.179771 & -2.229559 \end{bmatrix}; & \Lambda_3 &= \begin{bmatrix} 6.969281 \\ 4.195147 \\ -3.740141 \end{bmatrix} \\
 j = 4 \quad \mathbf{T}_4 &= \begin{bmatrix} 6.600000 & 1.019803 & 0. & 0 \\ 1.019803 & 3.053846 & 3.179771 & 0 \\ 0 & 3.179774 & -2.229559 & 1.101277 \\ 0 & 0 & 1.101277 & -1.210630 \end{bmatrix}; & \Lambda_4 &= \begin{bmatrix} 6.971348 \\ 4.234288 \\ -0.905203 \\ -4.086778 \end{bmatrix}
 \end{aligned}$$

Per  $j = 5$  si ottiene  $r \approx 1 \cdot 10^{-12}$ . Gli autovalori della matrice  $\mathbf{A}$  sono dati da

$$\Lambda(\mathbf{A}) = \begin{bmatrix} -0.987573459 \\ 0.872074485 \\ -4.090499422 \\ 4.234642665 \\ 6.971355730 \end{bmatrix}$$

Come confronto, applicando alla matrice (3.9) il metodo delle potenze a partire dal vettore iniziale  $\mathbf{q}_1$  si ottengono i risultati contenuti nella Tabella 3.3.

k	$\sigma_k$	$\mathbf{t}_k$				
1	6.82511	0.40179	0.46875	0.46875	0.33482	0.53572
2	6.91577	0.36137	0.50787	0.41997	0.30277	0.58601
3	6.95064	0.34681	0.50894	0.41448	0.25658	0.61891
4	6.96369	0.32959	0.52216	0.39588	0.24223	0.63507
5	6.96853	0.32364	0.52202	0.39275	0.22495	0.64643
6	6.97031	0.31716	0.52649	0.38584	0.21932	0.65210
7	6.97097	0.31484	0.52642	0.38450	0.21299	0.65615
8	6.97121	0.31246	0.52798	0.38198	0.21083	0.65820
9	6.97130	0.31157	0.52797	0.38144	0.20853	0.65967
10	6.97134	0.31079	0.52852	0.38053	0.20771	0.66043

Tabella 3.3: Risultati ottenuti mediante il metodo delle potenze per la matrice (3.9).

### 3.3 Metodi di trasformazione per similitudine

Le trasformazioni per similitudine mediante matrici ortogonali, o unitarie, rappresentano la base per la costruzione di diversi metodi efficienti per il calcolo degli autovalori e degli autovettori. In termini schematici, tali metodi cercano di trasformare una matrice  $\mathbf{A}$  in un'altra matrice con gli stessi autovalori, ma per la quale il

problema del calcolo degli autovalori sia più semplice. Più precisamente, a partire dalla matrice  $\mathbf{A}_1 \equiv \mathbf{A}$ , si costruisce una successione  $\{\mathbf{A}_k\}$ ,  $k = 2, 3, \dots$  mediante il seguente procedimento

$$\mathbf{A}_k = \mathbf{Q}_k^T \mathbf{A}_{k-1} \mathbf{Q}_k, \quad \mathbf{Q}_k^T \mathbf{Q}_k = \mathbf{I}, \quad k = 2, 3, \dots \quad (3.10)$$

Si vede facilmente che le matrici  $\mathbf{A}_k$  sono simili a  $\mathbf{A}$ , e quindi dal punto di vista teorico hanno lo stesso insieme di autovalori, e che i corrispondenti autovettori sono legati dalla relazione

$$\mathbf{x} = \mathbf{Q}_2 \mathbf{Q}_3 \cdots \mathbf{Q}_k \mathbf{x}_k$$

ove  $\mathbf{x}$  e  $\mathbf{x}_k$  sono rispettivamente un autovettore della matrice  $\mathbf{A}$  e della matrice  $\mathbf{A}_k$  corrispondenti al medesimo autovalore. Una proprietà importante dal punto di vista numerico è il fatto che la trasformazione (3.10) *preserva la simmetria*.

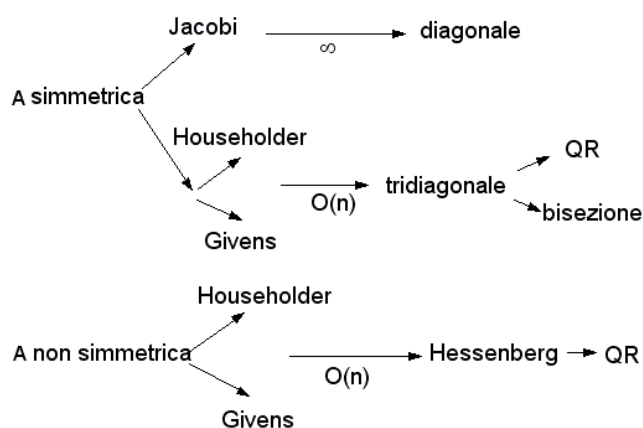


Figura 3.1: Schema dei metodi basati su trasformazioni per similitudine. Il simbolo  $O(n)$  indica che il metodo prevede un numero di iterazioni dell'ordine di  $n$ , mentre  $\infty$  significa che il risultato è ottenuto, in generale, come limite.

Le trasformazioni ortogonali utilizzate in pratica sono le *trasformazioni di Givens* (rotazioni) e le *trasformazioni di Householder* (riflessioni) (cfr. Appendice A per la definizione e le principali proprietà di tali trasformazioni). La Figura 3.1 riassume l'esposizione dei paragrafi successivi, rappresentando, in maniera schematica, le varie procedure numeriche corrispondenti al tipo di trasformazione utilizzato e al tipo di matrice.

Il metodo di Jacobi, basato sulle trasformazioni di tipo rotazione, trasforma una matrice simmetrica (o hermitiana) in una matrice diagonale. Il risultato è ottenuto come limite di una successione. Alternativamente, mediante le trasformazioni di

Householder o di Givens si può ridurre la matrice di partenza ad una matrice tridiagonale (quasi-triangolare, o di Hessenberg, nel caso di matrici non simmetriche). Tale risultato può essere ottenuto mediante un numero finito di iterazioni, proporzionale all'ordine  $n$  della matrice. Alla matrice nella forma ridotta, tridiagonale o di Hessenberg, si può applicare il metodo QR, mediante il quale si ottiene, come limite di una successione, una matrice diagonale nel caso simmetrico, e rispettivamente triangolare nel caso non simmetrico. Rispetto al metodo di Jacobi, il metodo QR presenta una superiore rapidità di convergenza, ma richiede ad ogni iterazione un costo superiore in termini di operazioni, dal momento che prevede ad ogni iterazione la decomposizione QR delle successive matrici. Tale costo si riduce, ovviamente, quando la matrice è sparsa. Questo è, in effetti, il motivo per il quale, prima di applicare il metodo QR, si preferisce ridurre la matrice alla forma tridiagonale, o rispettivamente di Hessenberg.

### 3.3.1 Metodo di Jacobi

Il *metodo di Jacobi*, noto anche come *metodo di Jacobi-Von Neumann*<sup>5</sup>, è uno dei metodi più noti ed utilizzati per il calcolo di tutti gli autovalori di *matrici simmetriche* (o più in generale, hermitiane) di piccole dimensioni. È un metodo del tipo (3.10) e utilizza trasformazioni di tipo *rotazione*. L'idea di base è quella di annullare, ad ogni iterazione, un elemento fuori dalla diagonale principale. Come matrice limite dell'iterazione, si ottiene allora una matrice diagonale, che fornisce direttamente gli autovalori, mentre il limite del prodotto  $\mathbf{Q}_2, \mathbf{Q}_3, \dots, \mathbf{Q}_k$  fornisce la corrispondente matrice degli autovettori.

Esporremo il metodo nel caso di una matrice  $\mathbf{A}$  simmetrica. Posto  $\mathbf{A}_1 = \mathbf{A}$ , si costruisce per  $k = 2, 3, \dots$  la successione di matrici

$$\mathbf{A}_k = \mathbf{G}_k^T \mathbf{A}_{k-1} \mathbf{G}_k$$

ove  $\mathbf{G}_k(p, q)$  è una matrice di rotazione di un opportuno angolo  $\theta$  nel piano di due vettori di base  $p, q$  scelti convenientemente. Tra gli elementi  $a_{ij}^{(k-1)}$  della matrice  $\mathbf{A}_{k-1}$  e gli elementi  $a_{ij}^{(k)}$  della matrice trasformata  $\mathbf{A}_k$  si hanno (cfr. Appendice A)

<sup>5</sup>C. G. J. J. Jacobi (1804-1851), J. Von Neumann (1903-1957).

le seguenti relazioni, ove si è posto  $c = \cos \theta$ ,  $s = \sin \theta$

$$\text{per } i \neq p, q, j \neq p, q \quad a_{ij}^{(k)} = a_{ij}^{(k-1)} \quad (3.11)$$

$$\text{per } j \neq p, q \quad a_{pj}^{(k)} = ca_{pj}^{(k-1)} - sa_{qj}^{(k-1)} \quad (3.12)$$

$$\text{per } j \neq p, q \quad a_{qj}^{(k)} = sa_{pj}^{(k-1)} + ca_{qj}^{(k-1)} \quad (3.13)$$

$$a_{pq}^{(k)} = cs(a_{pq}^{(k-1)} - a_{qq}^{(k-1)}) + (c^2 - s^2)a_{pq}^{(k-1)} \quad (3.14)$$

$$a_{pp}^{(k)} = c^2a_{pp}^{(k-1)} + s^2a_{qq}^{(k-1)} - 2csa_{pq}^{(k-1)} \quad (3.15)$$

$$a_{qq}^{(k)} = s^2a_{pp}^{(k-1)} + c^2a_{qq}^{(k-1)} + 2csa_{pq}^{(k-1)} \quad (3.16)$$

Ricordiamo che  $cs = \cos \theta \sin \theta = \frac{1}{2} \sin 2\theta$  e  $c^2 - s^2 = \cos^2 \theta - \sin^2 \theta = \cos 2\theta$ .

Alla generica iterazione  $k$ -ma l'angolo  $\theta$  viene determinato in maniera da ottenere

$$\boxed{a_{pq}^{(k)} = a_{qp}^{(k)} = 0} \quad (3.17)$$

ossia tale che

$$\boxed{\sin 2\theta (a_{pp}^{(k-1)} - a_{qq}^{(k-1)}) = -2 \cos 2\theta a_{pq}^{(k-1)}} \quad (3.18)$$

L'angolo  $\theta$  è allora determinato dalle seguenti relazioni

$$\left\{ \begin{array}{ll} \theta = \frac{\pi}{4} & \text{se } a_{pp}^{(k-1)} = a_{qq}^{(k-1)} \\ \tan 2\theta = \frac{2a_{pq}^{(k-1)}}{a_{qq}^{(k-1)} - a_{pp}^{(k-1)}}, \quad |\theta| < \frac{\pi}{4} & \text{se } a_{pp}^{(k-1)} \neq a_{qq}^{(k-1)} \end{array} \right. \quad (3.19)$$

La condizione  $|\theta| < \pi/4$  determina completamente il valore di  $\theta$  a partire da quello di  $\tan 2\theta$ . In realtà, non è nemmeno necessario calcolare esplicitamente l'angolo  $\theta$ , in quanto nelle formule precedenti intervengono solo le funzioni trigonometriche, o più precisamente relazioni in funzione di  $t = \tan \theta$ . In effetti, poiché  $|\theta| < \pi/4$ , si ha  $\cos \theta > 0$  e quindi

$$c = \cos \theta = \frac{1}{\sqrt{1 + \tan^2 \theta}} = \frac{1}{\sqrt{1 + t^2}}$$

$$s = \sin \theta = \cos \theta \tan \theta = \frac{t}{\sqrt{1 + t^2}}$$

Osserviamo, inoltre, che

$$\begin{aligned} a_{pp}^{(k)} - a_{pp}^{(k-1)} &= a_{pp}^{(k-1)}(\cos^2 \theta - 1) - a_{pq}^{(k-1)} \sin 2\theta + a_{qq}^{(k-1)} \sin^2 \theta \\ &= (a_{qq}^{(k-1)} - a_{pp}^{(k-1)}) \sin^2 \theta - a_{pq}^{(k-1)} \sin 2\theta \end{aligned}$$

da cui, ricordando (3.17)

$$a_{pp}^{(k)} = a_{pp}^{(k-1)} - a_{pq}^{(k-1)} \tan \theta \quad (3.20)$$

Allo stesso modo, si dimostra che

$$a_{qq}^{(k)} = a_{qq}^{(k-1)} + a_{pq}^{(k-1)} \tan \theta \quad (3.21)$$

Per determinare  $t = \tan \theta$  a partire da  $\tan 2\theta$ , si ha la seguente relazione trigonometrica

$$\tan 2\theta = \frac{2t}{1-t^2}$$

da cui

$$t^2 + 2t \cot 2\theta - 1 = 0, \text{ ove per la (3.19) } \cot 2\theta = \nu = \frac{a_{qq}^{(k-1)} - a_{pp}^{(k-1)}}{2a_{pq}^{(k-1)}} \quad (3.22)$$

Poiché  $t = \tan \theta$  verifica  $|t| < 1$  per  $|\theta| < \pi/4$  e il prodotto delle radici di (3.22) è uguale a  $-1$ , è necessario prendere la radice di modulo minimo. Si ha, quindi

$$\begin{cases} t = \frac{\text{sign}(\nu)}{|\nu| + \sqrt{1 + \nu^2}} & \text{se } \nu \neq 0 \\ t = 1 & \text{se } \nu = 0 \end{cases} \quad (3.23)$$

Successivamente, si calcola

$$c = \frac{1}{\sqrt{1+t^2}}; \quad s = ct$$

In definitiva, le formule che definiscono l'algoritmo diventano le seguenti

$$\begin{aligned} i \neq p, q, j \neq p, q & \quad a_{ij}^{(k)} = a_{ij}^{(k-1)} \\ j \neq p, q & \quad a_{pj}^{(k)} = ca_{pj}^{(k-1)} - sa_{qj}^{(k-1)} \\ j \neq p, q & \quad a_{qj}^{(k)} = sa_{pj}^{(k-1)} + ca_{qj}^{(k-1)} \\ & \quad a_{pp}^{(k)} = a_{pp}^{(k-1)} - ta_{pq}^{(k-1)} \\ & \quad a_{qq}^{(k)} = a_{qq}^{(k-1)} + ta_{pq}^{(k-1)} \end{aligned}$$

Si vede facilmente che il *numero di moltiplicazioni* per rotazione è approssimativamente dato da  $4n$ . Un'osservazione interessante è che le formule precedenti possono essere eseguite in *parallelo*.

Naturalmente, un elemento nullo di  $\mathbf{A}_{k-1}$  può diventare non nullo all'iterazione  $k$ ; per esempio, per  $j \neq p, q$ , se  $a_{pj}^{(k-1)} = 0$  e  $a_{qj}^{(k-1)} \neq 0$ , allora  $a_{pj}^{(k)} = -a_{qj}^{(k-1)} \sin \theta$ . Si tratta, quindi, in generale di un metodo non a terminazione finita.

Per la *scelta* della coppia di indici  $(p, q)$  ad ogni iterazione, vi sono diverse idee, tra le quali esamineremo le due seguenti.



### Metodo di Jacobi classico

Il metodo di Jacobi classico consiste nello scegliere  $(p, q)$ , con  $p < q$  in maniera che

$$|a_{pq}^{(k-1)}| = \max_{i \neq j} |a_{ij}^{(k-1)}| \quad (3.24)$$

Corrispondentemente a questa scelta, si ha il seguente risultato di convergenza.

**Teorema 3.2** *Se  $\mathbf{A}$  è una matrice simmetrica di ordine  $n$ , allora il metodo di Jacobi converge. Più precisamente, gli elementi non diagonali di  $\mathbf{A}$  convergono a zero, mentre ogni elemento diagonale  $a_{ii}^{(k)}$  converge verso un autovalore di  $\mathbf{A}$ .*

**DIMOSTRAZIONE.** La dimostrazione della convergenza è basata sulla seguente disuguaglianza, valida per la coppia di indici  $(p, q)$  definita in (3.24)

$$(a_{pq}^{(k-1)})^2 \geq \frac{S(\mathbf{A}_{k-1})}{n(n-1)}, \quad \text{ove } S(\mathbf{A}_{k-1}) = \sum_{\substack{r,j \\ r \neq j}} |a_{rj}^{(k-1)}|^2$$

Si ha allora

$$S(\mathbf{A}_k) = S(\mathbf{A}_{k-1}) - 2(a_{pq}^{(k-1)})^2 \leq S(\mathbf{A}_{k-1}) - 2 \frac{S(\mathbf{A}_{k-1})}{n(n-1)} = \gamma S(\mathbf{A}_{k-1})$$

ove  $\gamma = 1 - 2/(n(n-1)) < 1$ , per  $n \geq 2$ . Applicando in maniera ricorrente la maggiorazione precedente, si ha

$$S(\mathbf{A}_k) \leq \gamma^k S(\mathbf{A}_1)$$

da cui  $S(\mathbf{A}_k) \rightarrow 0$ . Per dimostrare poi che ogni elemento diagonale  $a_{ii}^{(k)}$  converge verso un autovalore di  $\mathbf{A}$  si utilizza il Teorema di Gerschgorin. ■

Quando la matrice hermitiana  $\mathbf{A}$  ha autovalori distinti  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , si può mostrare che da un certo passo  $k$  in poi si ha

$$S(\mathbf{A}_{k+N}) \leq \frac{2S(\mathbf{A}_{k-1})^2}{\delta^2}$$

ove  $N = n(n-1)/2$  e  $\delta = \min_{i \neq j} |\lambda_i - \lambda_j|$ . Tale risultato mostra che nelle ipotesi fatte il metodo di Jacobi ha convergenza quadratica.

### Metodo con soglia

Si fissa un valore  $s > 0$ . Si scelgono allora le coppie  $(p, q)$  nell'ordine naturale  $(1, 2)$ ,  $(1, 3)$ ,  $\dots$ ,  $(1, n)$ ,  $(2, 3)$ ,  $\dots$  ma si effettua la rotazione soltanto se  $|a_{pq}| \geq s$ . Il valore della soglia può essere ridefinito ad ogni rotazione. Nella implementazione riportata nel seguito si utilizza la seguente strategia. Nelle prime tre iterazioni si esegue la rotazione se  $|a_{pq}|$  è maggiore della soglia  $s = 0.2S_0/n^2$ , ove  $S_0$  è la somma dei moduli degli elementi fuori della diagonale. Dopo quattro iterazioni si evita la rotazione quando  $|a_{pq}|$  è piccolo rispetto agli elementi sulla diagonale  $|a_{pp}|$  e  $|a_{qq}|$ .

```

SUBROUTINE JAC(A,N,LA,D,V,NIT,B,Z,NITMAX,EPS)
C.....
C   calcola gli autovalori e gli autovettori di una matrice
C   A reale simmetrica di ordine N.
C   In output, gli elementi di A al di sopra della diagonale sono distrutti.
C   Gli autovalori sono ritornati nel vettore D.
C   Gli autovettori, normalizzati sono ritornati nella matrice V.
C   LA dimensione dell'array A
C   NIT numero di iterazioni impiegate.
C   B,Z vettori di lavoro di dimensione LA
C   NITMAX numero massimo di iterazioni
C   EPS test d'arresto
C.....
      DIMENSION A(LA,LA),D(LA),V(LA,LA),B(LA),Z(LA)
C   costruzione matrice identita'
      DO 20 I=1,N
        DO 10 J=1,N
          V(I,J)=0.
10      CONTINUE
          V(I,I)=1.
20      CONTINUE
C   inizializza B e D alla diagonale di A
      DO 30 I=1,N
        B(I)=A(I,I)
        D(I)=B(I)
        Z(I)=0.
30      CONTINUE
      NIT=0
      DO 200 IT=1,NITMAX
C   somma degli elementi fuori della diagonale
        SM=0.
        DO 15 I=1,N-1
          DO 14 J=I+1,N
            SM=SM+ABS(A(I,J))
14      CONTINUE
15      CONTINUE
        IF(SM.LE.EPS)RETURN
        IF(IT.LT.4)THEN
C   soglia nelle prime iterazioni
          TRESH=0.2*SM/N**2
        ELSE
          TRESH=0.
        ENDIF
        DO 220 I=1,N-1
          DO 210 J=I+1,N
            G=100.*ABS(A(I,J))
C   dopo quattro iterazioni si salta la rotazione se
C   l'elemento fuori dalla diagonale e' piccolo
            IF((IT.GT.4).AND.(ABS(D(I))+G.EQ.ABS(D(I)))
*           .AND.(ABS(D(J))+G.EQ.ABS(D(J))))THEN
              A(I,J)=0.

```

```

ELSE IF(ABS(A(I,J)) .GT. TRESH) THEN
  H=D(J)-D(I)
  IF(ABS(H)+G .EQ. ABS(H)) THEN
    T=A(I,J)/H
  ELSE
    THETA=0.5*H/A(I,J)
    T=1./(ABS(THETA)+SQRT(1.+THETA**2))
    IF(THETA.LT.0.) T=-T
  ENDIF
  C=1./SQRT(1+T**2)
  S=T*C
  TAU=S/(1.+C)
  H=T*A(I,J)
  Z(I)=Z(I)-H
  Z(J)=Z(J)+H
  D(I)=D(I)-H
  D(J)=D(J)+H
  A(I,J)=0.
  DO 160 JJ=1,I-1
    G=A(JJ,I)
    H=A(JJ,J)
    A(JJ,I)=G-S*(H+G*TAU)
    A(JJ,J)=H+S*(G-H*TAU)
160  CONTINUE
  DO 170 JJ=I+1,J-1
    G=A(I,JJ)
    H=A(JJ,J)
    A(I,JJ)=G-S*(H+G*TAU)
    A(JJ,J)=H+S*(G-H*TAU)
170  CONTINUE
  DO 180 JJ=J+1,N
    G=A(I,JJ)
    H=A(J,JJ)
    A(I,JJ)=G-S*(H+G*TAU)
    A(J,JJ)=H+S*(G-H*TAU)
180  CONTINUE
  DO 190 JJ=1,N
    G=V(JJ,I)
    H=V(JJ,J)
    V(JJ,I)=G-S*(H+G*TAU)
    V(JJ,J)=H+S*(G-H*TAU)
190  CONTINUE
    NIT=NIT+1
  ENDIF
210  CONTINUE
220  CONTINUE
  DO 230 I=1,N
    B(I)=B(I)+Z(I)
    D(I)=B(I)
    Z(I)=0.
230  CONTINUE

```

```

200 CONTINUE
    PRINT*, 'superato numero massimo di iterazioni'
    RETURN
    END

```

Di seguito riportiamo un esempio di driver.

```

PARAMETER (LA=10, EPS=1.E-5, NITMAX=50)
DIMENSION A(LA,LA), V(LA,LA), D(LA), B(LA), Z(LA)
READ*, N
DO 10 I=1, N
    READ*, (A(I, J), J=1, N)
10 CONTINUE

CALL JAC(A, N, LA, D, V, NIT, B, Z, NITMAX, EPS)
PRINT*, NIT
PRINT*, (D(I), I=1, N)
END

```

Applicato alla seguente matrice

$$\mathbf{A} = \begin{bmatrix} 6 & 4 & 4 & 1 \\ 4 & 6 & 1 & 4 \\ 4 & 1 & 6 & 4 \\ 1 & 4 & 4 & 6 \end{bmatrix}$$

fornisce i risultati:

```

      17
-1.00000   15.0000   5.00000   5.00000

```

### 3.3.2 Metodo di Householder

Data una matrice  $\mathbf{A}$  simmetrica di ordine  $n$  simmetrica, si assumono come matrici di trasformazione  $\mathbf{Q}_k$  in (3.10) le matrici elementari di Householder. Lo scopo è quello di ridurre la matrice  $\mathbf{A}$  alla forma *tridiagonale*.

Al primo passo, posto

$$\mathbf{A}_1 = \mathbf{A} = \begin{bmatrix} a_{11}^{(1)} & \mathbf{a}_1^T \\ \mathbf{a}_1 & \mathbf{B}_1 \end{bmatrix}$$

con  $\mathbf{B}_1$  matrice di ordine  $n - 1$  e  $\mathbf{a}_1 \in \mathbb{R}^{n-1}$ , si considera la matrice elementare di Householder  $\mathbf{H}_1$  di dimensione  $n - 1$  tale che

$$\mathbf{H}_1 \mathbf{a}_1 = \alpha_1 \mathbf{e}_1$$

ove  $\mathbf{e}_1$  è il primo vettore della base canonica di  $\mathbb{R}^{n-1}$ . La matrice

$$\mathbf{Q}_1 = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{H}_1 \end{bmatrix}$$

è tale che nella matrice

$$\mathbf{A}_2 = \mathbf{Q}_1^{-1} \mathbf{A}_1 \mathbf{Q}_1 = \mathbf{Q}_1 \mathbf{A}_1 \mathbf{Q}_1$$

siano nulli tutti gli elementi della prima colonna con indice di riga maggiore di due e i simmetrici di questi nella prima riga.

Al generico passo  $k$ , si parte dalla seguente situazione

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{T}_k & \mathbf{b}_k & 0 \\ \mathbf{b}_k^T & a_{kk}^{(k)} & \mathbf{a}_k^T \\ 0 & \mathbf{a}_k & \mathbf{B}_k \end{bmatrix}$$

ove  $\mathbf{T}_k$  è una matrice di ordine  $k - 1$  tridiagonale simmetrica,  $\mathbf{b}_k$  è un vettore di ordine  $k - 1$ , che ha nulle le prime  $k - 2$  componenti,  $\mathbf{a}_k \in \mathbb{R}^{n-k}$  e  $\mathbf{B}_k$  è una matrice simmetrica di ordine  $n - k$ . Sia allora  $\mathbf{H}_k$  la matrice di Householder di ordine  $n - k$  tale che

$$\mathbf{H}_k \mathbf{a}_k = \alpha_k \mathbf{e}_1$$

con  $\mathbf{e}_1$ , questa volta, primo vettore della base canonica di  $\mathbb{R}^{n-k}$ . Posto

$$\mathbf{Q}_k = \begin{bmatrix} \mathbf{I}_k & \mathbf{0}^T \\ \mathbf{0} & \mathbf{H}_k \end{bmatrix}$$

si ha

$$\mathbf{A}_{k+1} = \mathbf{Q}_k^T \mathbf{A}_k \mathbf{Q}_k = \begin{bmatrix} \mathbf{T}_k & \mathbf{b}_k & 0 \\ \mathbf{b}_k^T & a_{kk}^{(k)} & \mathbf{a}_k^T \mathbf{H}_k \\ 0 & \mathbf{H}_k \mathbf{a}_k & \mathbf{H}_k \mathbf{B}_k \mathbf{H}_k \end{bmatrix}$$

Dal momento che il vettore  $\mathbf{H}_k \mathbf{a}_k \in \mathbb{R}^{n-k}$  ha nulle le componenti di indice maggiore o uguale a due, la sottomatrice principale di ordine  $k + 2$  della matrice  $\mathbf{A}_{k+1}$  è tridiagonale simmetrica. Applicando il procedimento  $n - 2$  volte si ottiene come matrice trasformata per similitudine (e quindi con gli stessi autovalori della matrice  $\mathbf{A}$ ) una matrice *tridiagonale simmetrica*.

Tenendo presente che  $\mathbf{H}_k$  ha la forma  $\mathbf{H}_k = \mathbf{I} - \beta_k \mathbf{u}_k \mathbf{u}_k^T$ , per calcolare la matrice  $\mathbf{H}_k \mathbf{B}_k \mathbf{H}_k$  non si utilizza esplicitamente  $\mathbf{H}_k$ , ma solo il vettore  $\mathbf{u}_k$ . Si può, infatti, mostrare facilmente che

$$\mathbf{H}_k \mathbf{B}_k \mathbf{H}_k = \mathbf{B}_k - \mathbf{q}_k \mathbf{u}_k^T - \mathbf{u}_k \mathbf{q}_k^T$$

ove

$$\mathbf{q}_k = \mathbf{r}_k - \frac{1}{2} \beta_k (\mathbf{r}_k^T \mathbf{u}_k) \mathbf{u}_k, \quad \mathbf{r}_k = \beta_k \mathbf{B}_k \mathbf{u}_k$$

Il numero delle moltiplicazioni richiesto per ogni trasformazione  $k \rightarrow k + 1$  è, quindi,  $2(n-k)^2$ . In totale, pertanto, il metodo di Householder richiede per tridiagonalizzare una matrice simmetrica un numero di operazioni dato da  $\sum_{k=1}^{n-2} 2(n-k)^2 \approx 2n^3/3$ .

### 3.3.3 Metodo di Givens

Nel *metodo di Givens* si utilizzano, come nel metodo di Jacobi, le matrici di *rotazione*. Tuttavia, a differenza del metodo di Jacobi, il metodo di Givens ha come obiettivo quello di ridurre una matrice simmetrica ad una matrice *tridiagonale* simmetrica mediante un numero *finito* di trasformazioni.

Sia  $\mathbf{A}$  una matrice simmetrica di ordine  $n$  e poniamo  $\mathbf{A}_1 = \mathbf{A}$ ; si definisce quindi

$$\mathbf{A}_k = \mathbf{G}_{pq}^T \mathbf{A}_{k-1} \mathbf{G}_{pq}$$

Si hanno, allora le relazioni (3.12) – (3.16). La scelta degli indici  $(p, q)$  avviene in questo modo.

*Prima tappa.* Si vogliono annullare gli elementi della matrice che sono nella prima colonna salvo i primi due, cioè gli elementi di indici  $(3,1), (4,1), \dots, (n,1)$ .

Per ogni coppia  $(p, 1)$ , con  $p = 3, 4, \dots, n$ , si considera la rotazione nel piano  $(p, 2)$ , ove l'angolo  $\theta$  è determinato in modo che sia nullo l'elemento  $a_{p1}^{(k)}$  e quindi

$$a_{p1}^{(k-1)} \cos \theta - a_{21}^{(k-1)} \sin \theta = 0$$

Posto

$$r = \frac{1}{\sqrt{(a_{p1}^{(k-1)})^2 + (a_{21}^{(k-1)})^2}}$$

si ha

$$\sin \theta = r a_{p1}^{(k-1)}; \quad \cos \theta = r a_{21}^{(k-1)}$$

Ricordando le relazioni (3.12) – (3.16), si ha che ogni trasformazione nel piano  $(p, q)$  modifica soltanto le colonne  $p, q$ . Pertanto, *uno zero creato nella prima colonna (o prima riga) non è modificato nelle trasformazioni successive*.

Al termine della prima tappa si è ottenuta una matrice simile alla matrice di partenza con gli  $n - 2$  ultimi elementi della prima colonna nulli.

*Seconda tappa.* Si considera la seconda colonna. Per  $p = 4, 5, \dots, n$  si annulla l'elemento di posizione  $(p, 2)$  facendo una rotazione nel piano  $(p, 3)$ , che ha l'effetto di modificare gli elementi della terza e  $p$ -ma riga e colonna. L'elemento  $(p, 1)$ , che era stato annullato precedentemente è una combinazione lineare dell'elemento in posizione  $(p, 1)$  e dell'elemento in posizione  $(3, 1)$ . Poiché sono ambedue nulli, non si modificano gli zeri corrispondenti della prima colonna.

Le altre tappe consistono nel procedere colonna per colonna nell'ordine naturale. Dopo  $(n - 2) + (n - 3) + \dots + 1 = (n - 2)(n - 1)/2$  trasformazioni si ottiene una *matrice tridiagonale simmetrica*. Tenendo conto della simmetria, il numero totale di moltiplicazioni è dato da

$$\sum_{j=1}^{n-2} 4(n+1-j)(n-1-j) \approx \frac{4}{3}n^3$$

cioè circa il doppio delle operazioni richieste dal metodo di Householder. Tuttavia, come si è già osservato, le rotazioni permettono un azzeramento mirato e pertanto il metodo può risultare più conveniente per il calcolo degli autovalori e degli autovettori di una matrice sparsa.

### 3.3.4 Matrici non simmetriche

Quando la matrice  $\mathbf{A}$  è *non simmetrica*, l'applicazione dei metodi di Householder e di Givens dà origine a matrici equivalenti alla matrice  $\mathbf{A}$ , con una struttura che solitamente viene indicata come *matrice di Hessenberg superiore* (cfr. Figura 3.2). Detta  $\mathbf{H}$  tale matrice, si ha quindi  $h_{ij} = 0$ ,  $i > j + 1$ ,  $j = 1, 2, \dots, n - 2$ .

$$\mathbf{H} = \begin{bmatrix} * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * \\ & * & * & * & * & * & * & * \\ & & * & * & * & * & * & * \\ & & & * & * & * & * & * \\ & & & & * & * & * & * \\ & & & & & * & * & * \\ & & & & & & * & * \\ & & & & & & & * & * \end{bmatrix}$$

Figura 3.2: Matrice di Hessenberg.

Il costo per il metodo di Householder è di circa  $\frac{5}{3}n^3$ , mentre è di circa il doppio per quello di Givens.

La trasformazione nella forma di Hessenberg è preliminare alla applicazione del metodo QR, che come vedremo più avanti *conserva* tale struttura.

### 3.3.5 Matrici tridiagonali simmetriche

Come abbiamo visto, l'applicazione dei metodi di Lanczos, di Householder e di Givens prevede come fase terminale la ricerca degli autovalori e degli autovettori di una *matrice tridiagonale simmetrica*. In questo paragrafo incominceremo ad analizzare una prima tecnica per risolvere tale problema. Essa si basa sulla risoluzione dell'*equazione caratteristica* e presenta interesse quando si devono calcolare *determinati* autovalori, quali, ad esempio, gli autovalori di modulo massimo o di modulo minimo, o quelli contenuti in un particolare intervallo. Il metodo QR, che esamineremo nel paragrafo successivo è, invece, in generale più conveniente quando sono richiesti tutti gli autovalori della matrice.

Sia  $\mathbf{A}$  una *matrice tridiagonale reale simmetrica*, ossia della forma

$$\mathbf{A} = \begin{bmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & b_{n-2} & a_{n-1} & b_{n-1} & \\ & & & b_{n-1} & a_n & \end{bmatrix}$$

Tale matrice risulta (cfr. Appendice A) irriducibile quando  $b_i \neq 0$  per  $i = 1, 2, \dots, n-1$ . Viceversa, quando almeno uno degli elementi  $b_i$  è nullo la matrice è riducibile ed ha la seguente decomposizione a blocchi

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & & & 0 \\ & \mathbf{A}_{22} & & \\ & & \ddots & \\ 0 & & & \mathbf{A}_{rr} \end{bmatrix}$$

ove  $\mathbf{A}_{ii}$ ,  $i = 1, 2, \dots, r$ , sono matrici tridiagonali di ordine inferiore a  $n$ . In questo caso, quindi, il problema del calcolo degli autovalori di  $\mathbf{A}$  si riduce a quello delle matrici  $\mathbf{A}_{ii}$ . Possiamo, pertanto, supporre, senza perdita di generalità, che la matrice  $\mathbf{A}$  sia irriducibile.

Il *polinomio caratteristico* di una matrice simmetrica tridiagonale e irriducibile può essere calcolato mediante una opportuna iterazione, grazie al seguente risultato.

**Lemma 3.1** *Data una matrice  $\mathbf{A}$  tridiagonale simmetrica e irriducibile, indichiamo con  $\mathbf{A}_k$  la sua sottomatrice principale di ordine  $k$  e con  $p_k(\lambda)$  il corrispondente polinomio caratteristico. Posto, allora*

$$p_0(\lambda) = 1; \quad p_1(\lambda) = a_1 - \lambda$$

si ha per  $k \geq 2$

$$p_k(\lambda) = (a_k - \lambda)p_{k-1}(\lambda) - b_{k-1}^2 p_{k-2}(\lambda) \quad (3.25)$$

Per la dimostrazione basta sviluppare il determinante di ogni matrice  $\mathbf{A}_k - \lambda \mathbf{I}_k$  rispetto all'ultima riga.

Le radici del polinomio caratteristico  $p_n(\lambda)$ , che sono tutte reali, in quanto la matrice è simmetrica, possono essere approssimate mediante un procedimento di *bisezione* (cfr. successivo Capitolo 5). Se  $p_n(\alpha)p_n(\beta) < 0$  e  $\alpha < \beta$ , la seguente iterazione

$$\begin{aligned} &\text{while } |\alpha - \beta| > \epsilon(|\alpha| + |\beta|) \\ &\quad \lambda = (\alpha + \beta)/2 \\ &\quad \text{if } p_n(\lambda)p_n(\alpha) < 0 \end{aligned}$$



```

         $\beta = \lambda$ 
    else
         $\alpha = \lambda$ 
    end
end

```

termina quando si è ottenuta una approssimazione di un autovalore della matrice  $\mathbf{A}$  con un errore relativo inferiore a una precisione prefissata  $\epsilon$ .

Si possono, inoltre, dimostrare le seguenti proprietà.

**Lemma 3.2** *Le radici del polinomio  $p_k(\lambda)$  sono semplici e reali e i polinomi successivi  $p_k(\lambda)$  e  $p_{k-1}(\lambda)$  non hanno radici comuni. Inoltre, le radici di  $p_{k-1}(\lambda)$  separano strettamente le radici di  $p_k(\lambda)$ .*

► **Esempio 3.3** Come illustrazione, consideriamo la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix}, \quad \Lambda(\mathbf{A}) = \begin{bmatrix} 3.24697960371747 \\ 2.44504186791263 \\ 3.80193773580484 \\ 1.55495813208737 \\ 0.75302039628253 \\ 0.19806226419516 \end{bmatrix} \quad (3.26)$$

ove  $\Lambda(\mathbf{A})$  indica gli autovalori esatti della matrice  $\mathbf{A}$ . Dal Lemma 3.1 si ricava la seguente successione

$$\begin{aligned} p_0(\lambda) &= 1, & p_1(\lambda) &= 2 - \lambda \\ p_2(\lambda) &= \lambda^2 - 4\lambda + 3 \\ p_3(\lambda) &= -\lambda^3 + 6\lambda^2 - 10\lambda + 4 \\ p_4(\lambda) &= \lambda^4 - 8\lambda^3 + 21\lambda^2 - 20\lambda + 5 \\ p_5(\lambda) &= -\lambda^5 + 10\lambda^4 - 36\lambda^3 + 56\lambda^2 - 35\lambda + 6 \\ p_6(\lambda) &= \lambda^6 - 12\lambda^5 + 55\lambda^4 - 120\lambda^3 + 126\lambda^2 - 56\lambda + 7 \end{aligned}$$

In particolare, il polinomio  $p_6(\lambda)$  fornisce il polinomio caratteristico della matrice  $\mathbf{A}$ . I polinomi  $p_k$  sono rappresentati in Figura 3.3, nella quale si possono verificare i risultati precedenti. ■

Una successione di polinomi che verificano le proprietà del Lemma 3.2 è detta una *successione di Sturm*. Si ha come conseguenza il seguente risultato.

**Teorema 3.3** *Sia  $\mathbf{A}$  una matrice tridiagonale simmetrica e irriducibile. Il numero  $C(\mu)$  delle concordanze di segno nella successione  $p_0(\mu), p_1(\mu), \dots, p_n(\mu)$  è uguale al numero degli autovalori di  $\mathbf{A}$  strettamente maggiori di  $\mu$ . Quando  $p_k(\mu) = 0$ , si definisce come segno quello dell'opposto di  $p_{k-1}(\mu)$  che è non nullo. Pertanto, il numero degli autovalori di  $\mathbf{A}$  compresi in  $]a, b]$ , ove  $a < b$  è uguale a  $C(a) - C(b)$ .*

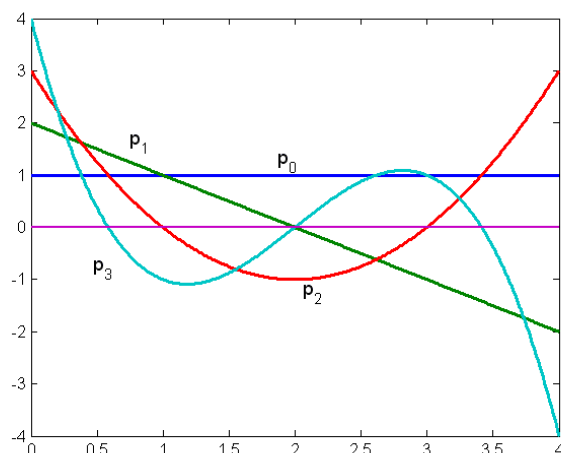


Figura 3.3: Rappresentazione della successione dei polinomi  $p_k(\lambda)$  corrispondenti alla matrice (3.26).

Il risultato, incorporato opportunamente nell'algoritmo della bisezione esaminato in precedenza, permette di ottenere una conveniente approssimazione degli autovalori di modulo massimo o minimo, o più in generale degli autovalori contenuti in un fissato intervallo.

### 3.3.6 Metodo QR

Il *metodo QR*<sup>6</sup> è uno dei metodi numerici più interessanti per il calcolo degli autovalori. Tuttavia, applicato ad una matrice generale, ha un costo elevato dal momento che richiede ad ogni iterazione la decomposizione di una matrice nel prodotto di una matrice ortogonale  $\mathbf{Q}$  e di una matrice triangolare  $\mathbf{R}$ . Per tale motivo, come abbiamo già osservato in precedenza, il metodo è applicato a matrici già ridotte a forma tridiagonale (caso simmetrico) e di Hessenberg (caso generale). Pur essendo l'idea del metodo semplice, la sua implementazione richiede particolari accorgimenti. Qui ci limiteremo a fornire le idee di base, rinviando, per una implementazione opportuna, ad esempio al codice EISPACK (cfr. Garbow, et al. [62])

Data una matrice  $\mathbf{A}$  di ordine  $n$ , mediante l'applicazione dell'algoritmo di Householder o di Givens (analizzati in Capitolo 2), si ottiene la seguente decomposizione

$$\mathbf{A}_1 = \mathbf{A} = \mathbf{Q}_1 \mathbf{R}_1$$

ove  $\mathbf{Q}_1$  è una matrice unitaria e  $\mathbf{R}_1$  è una matrice triangolare superiore.

<sup>6</sup>sviluppato da J. G. F. Francis, *The QR Transformation: A Unitary Analogue to the LR Transformation*, Comp. J., 4, 1961, rappresenta uno sviluppo del metodo LR di Rutishauser (1958), nel quale si utilizza una decomposizione della matrice in matrici triangolari.

Si calcola poi

$$\mathbf{A}_2 = \mathbf{R}_1 \mathbf{Q}_1$$

Tenendo presente che  $\mathbf{A}_2 = \mathbf{R}_1 \mathbf{Q}_1 = (\mathbf{Q}_1^T \mathbf{Q}_1)(\mathbf{R}_1 \mathbf{Q}_1) = \mathbf{Q}_1^T \mathbf{A}_1 \mathbf{Q}_1$ , si ha che  $\mathbf{A}_2$  è una matrice simile alla matrice  $\mathbf{A}_1$ .

Il procedimento viene quindi iterato, per  $k \geq 2$

$$\mathbf{A}_k = \mathbf{Q}_k \mathbf{R}_k \Rightarrow \mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{Q}_k = \mathbf{Q}_k^T \mathbf{A}_k \mathbf{Q}_k$$

for  $k = 1, 2, \dots$

$\mathbf{A} = \mathbf{Q}\mathbf{R}$

$\mathbf{A} = \mathbf{R}\mathbf{Q}$

end  $k$

► **Esempio 3.4** Data la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

si ha

$$\mathbf{A} = \mathbf{Q}_1 \mathbf{R}_1 = \begin{bmatrix} -0.4472 & -0.8944 \\ -0.8944 & 0.4472 \end{bmatrix} \begin{bmatrix} -2.2361 & -1.7889 \\ 0 & -1.3416 \end{bmatrix}$$

da cui

$$\begin{aligned} \mathbf{A}_2 = \mathbf{R}_1 \mathbf{Q}_1 &= \begin{bmatrix} 2.6000 & 1.2000 \\ 1.2000 & -0.6000 \end{bmatrix} \\ &= \mathbf{Q}_2 \mathbf{R}_2 = \begin{bmatrix} -0.9080 & -0.4191 \\ -0.4191 & 0.9080 \end{bmatrix} \begin{bmatrix} -2.8636 & -0.8381 \\ 0 & -1.0476 \end{bmatrix} \end{aligned}$$

La matrice  $\mathbf{A}_k$  per  $k = 11$  è data da

$$\mathbf{A}_{11} = \begin{bmatrix} 3.0000 & 0.0001 \\ 0.0001 & -1.0000 \end{bmatrix}$$

■

Come ulteriore esemplificazione, consideriamo l'applicazione del metodo alla matrice (3.26). Dopo 20 iterazioni si ottiene la seguente matrice

$$\mathbf{A}_{20} = \begin{bmatrix} 3.798685 & 0.042360 & 0. & 0. & 0. & 0. \\ 0.042360 & 3.250143 & 0.008426 & 0. & 0. & 0. \\ 0. & 0.008426 & 2.445130 & 0.000364 & 0. & 0. \\ 0. & 0. & 0.000364 & 1.554958 & 0.000001 & 0. \\ 0. & 0. & 0. & 0.000001 & 0.753020 & 0. \\ 0. & 0. & 0. & 0. & 0.000000 & 0.198062 \end{bmatrix}$$

i cui autovalori sono dati da

$$\mathbf{\Lambda}(\mathbf{A}_{20}) = [ 3.80193 \quad 3.24697 \quad 2.44504 \quad 1.55495 \quad 0.75302 \quad 0.19806 ]$$

Nei due esempi considerati, nei quali la matrice è simmetrica, si osserva la convergenza del metodo a una matrice diagonale. Tale convergenza può essere, in effetti, dimostrata in condizioni abbastanza generali per le quali, comunque, rinviamo alla bibliografia, limitandoci ad osservare che, utilizzando una opportuna tecnica di traslazione degli autovalori<sup>7</sup>, la convergenza è di ordine 3. Osserviamo, infine, che quando il metodo è applicato a una matrice non simmetrica, il limite è, in generale, una matrice triangolare.

### 3.4 Problema degli autovalori generalizzato

In numerose applicazioni ha interesse la seguente estensione del problema degli autovalori. Date due matrici  $\mathbf{A}$  e  $\mathbf{B}$  di ordine  $n$ , si cercano i valori  $\lambda \in \mathbb{C}$  per i quali il seguente sistema omogeneo

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \quad (3.27)$$

ammette soluzioni  $\mathbf{x}$  non identicamente nulle<sup>8</sup>. L'insieme di tutte le matrici della forma  $\mathbf{A} - \lambda\mathbf{B}$ , con  $\lambda \in \mathbb{C}$ , è detto un *pencil*. Gli autovalori del pencil sono, allora, elementi dell'insieme  $\lambda(\mathbf{A}, \mathbf{B})$  definito da

$$\lambda(\mathbf{A}, \mathbf{B}) := \{z \in \mathbb{C} \mid \det(\mathbf{A} - z\mathbf{B}) = 0\}$$

Se  $\lambda \in \lambda(\mathbf{A}, \mathbf{B})$ , un vettore  $\mathbf{x}$  soluzione del sistema (3.27) è chiamato un *autovettore* di  $\mathbf{A} - \lambda\mathbf{B}$ .

Un aspetto importante da sottolineare è che il problema generalizzato ha  $n$  autovalori se e solo se il rango della matrice  $\mathbf{B}$  è  $n$ . In caso contrario, l'insieme  $\lambda(\mathbf{A}, \mathbf{B})$  può essere sia finito che vuoto o infinito, come mostrato nei seguenti esempi

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow \lambda(\mathbf{A}, \mathbf{B}) = \{1\} \\ \mathbf{A} &= \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \Rightarrow \lambda(\mathbf{A}, \mathbf{B}) = \{\emptyset\} \\ \mathbf{A} &= \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow \lambda(\mathbf{A}, \mathbf{B}) = \{\mathbb{C}\} \end{aligned}$$

Osserviamo, inoltre, che se  $0 \neq \lambda \in \lambda(\mathbf{A}, \mathbf{B})$ , allora  $(1/\lambda) \in \lambda(\mathbf{B}, \mathbf{A})$ . Inoltre, se  $\mathbf{B}$  è non singolare, allora,  $\lambda(\mathbf{A}, \mathbf{B}) = \lambda(\mathbf{B}^{-1}\mathbf{A}, \mathbf{I})$ .

L'ultima osservazione suggerisce un metodo per risolvere il problema generalizzato quando  $\mathbf{B}$  è non singolare.

<sup>7</sup>In maniera schematica, si esegue ad ogni passo la decomposizione  $\mathbf{A}_k - s_k\mathbf{I} = \mathbf{Q}_k\mathbf{R}_k$  e si pone  $\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + s_k\mathbf{I}$ , ove  $s_k$  sono quantità scelte opportunamente ad ogni passo, e vicine al valore di un autovalore della matrice  $\mathbf{A}$ ; tenendo presente che la matrice  $\mathbf{A}_k$  converge ad una matrice triangolare, una possibile scelta consiste nell'assumere come  $s_k$  l'elemento  $a_{nn}^{(k)}$  della matrice  $\mathbf{A}_k$ .

<sup>8</sup>Nello studio della stabilità di strutture, la matrice  $\mathbf{A}$  dipende dalle proprietà del materiale ed è nota come *stiffness matrix*, mentre la matrice  $\mathbf{B}$  dipende dal carico ed è detta *mass matrix*.

- Si risolve il sistema  $\mathbf{BC} = \mathbf{A}$ , ossia si calcola  $\mathbf{B}^{-1}\mathbf{A}$ ;
- si utilizza il metodo QR per calcolare gli autovalori di  $\mathbf{C}$ .

Attraverso opportune esemplificazioni, si può mostrare che tale procedura è numericamente instabile, e quindi non opportuna, quando la matrice  $\mathbf{B}$  è malcondizionata.

Algoritmi stabili per il problema generalizzato degli autovalori si basano sul seguente risultato che è una estensione del classico teorema di decomposizione di Schur (cfr. Appendice A).

**Proposizione 3.1** *Se  $\mathbf{A}$  e  $\mathbf{B}$  sono matrici in  $\mathbb{C}^{n \times n}$ , allora esistono due matrici unitarie  $\mathbf{Q}$  e  $\mathbf{Z}$  tali che  $\mathbf{Q}^*\mathbf{AZ} = \mathbf{T}$  e  $\mathbf{Q}^*\mathbf{BZ} = \mathbf{S}$  sono triangolari superiori. Se per un valore di  $k$ ,  $t_{kk}$  e  $s_{kk}$  sono ambedue nulli, allora  $\lambda(\mathbf{A}, \mathbf{B}) = \mathbb{C}$ . Altrimenti,*

$$\lambda(\mathbf{A}, \mathbf{B}) = \{t_{ii}/s_{ii} : s_{ii} \neq 0\}$$

Le matrici  $\mathbf{Q}$  e  $\mathbf{Z}$  possono essere costruite mediante una opportuna generalizzazione del metodo QR, nota come *metodo QZ*.

Un caso di pencil particolarmente importante corrisponde a  $\mathbf{A}$  matrice simmetrica e  $\mathbf{B}$  simmetrica definita positiva (*symmetric definite pencil*). Questa proprietà si mantiene quando si operano trasformazioni di congruenza, ossia

$$\begin{array}{ccc} \mathbf{A} - \lambda\mathbf{B} & \Rightarrow & (\mathbf{X}^T\mathbf{A}\mathbf{X}) - \lambda(\mathbf{X}^T\mathbf{B}\mathbf{X}) \\ \text{simmetrica definita} & & \text{simmetrica definita} \end{array}$$

Osserviamo che, al contrario, la matrice  $\mathbf{B}^{-1}\mathbf{A}$  non è in generale simmetrica.

Il risultato precedente può essere utilizzato nel seguente modo. Si calcola la decomposizione di Cholesky della matrice  $\mathbf{B}$

$$\mathbf{B} = \mathbf{L}\mathbf{L}^T$$

con  $\mathbf{L}$  matrice triangolare non singolare. La matrice  $\mathbf{B}^{-1}\mathbf{A}$  è simile alla matrice  $\mathbf{G} = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^T$ . Infatti

$$\mathbf{L}^T(\mathbf{B}^{-1}\mathbf{A})(\mathbf{L}^T)^{-1} = \mathbf{L}^T(\mathbf{L}^T)^{-1}\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^T = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^T = \mathbf{G}$$

Il problema del calcolo degli autovalori della matrice  $\mathbf{B}^{-1}\mathbf{A}$  è in questo modo ricondotto al calcolo degli autovalori della matrice *simmetrica*  $\mathbf{G}$ . Il calcolo di  $\mathbf{G}$  può essere effettuato nel seguente modo. Si calcola

$$\mathbf{F} = \mathbf{A}(\mathbf{L}^{-1})^T$$

risolvendo il sistema triangolare  $\mathbf{F}\mathbf{L}^T = \mathbf{A}$  e successivamente si ottiene  $\mathbf{G}$  resolvendo il sistema  $\mathbf{L}\mathbf{G} = \mathbf{F}$ . Dal momento che  $\mathbf{G}$  è simmetrica, è sufficiente calcolare gli elementi sotto la diagonale di  $\mathbf{G}$  e allora basta calcolare in  $\mathbf{F}$  gli elementi  $f_{ik}$ ,  $k \leq i$ . Il calcolo di  $\mathbf{G}$  da  $\mathbf{A}$  e  $\mathbf{B}$  richiede circa  $\frac{2}{3}n^3$  moltiplicazioni.

► **Esempio 3.5** Se

$$\mathbf{A} = \begin{bmatrix} 229 & 163 \\ 163 & 116 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 81 & 59 \\ 59 & 43 \end{bmatrix}$$

allora  $\mathbf{A} - \lambda\mathbf{B}$  è un pencil simmetrico definito, in quanto gli autovalori di  $\mathbf{B}$  sono dati da  $[123.9839, 0.0161]$ .

La decomposizione  $\mathbf{LL}^T$  di Cholesky della matrice  $\mathbf{B}$  fornisce

$$\mathbf{L} = \begin{bmatrix} 9.0000 & 0 \\ 6.5556 & 0.1571 \end{bmatrix} \quad \text{con} \quad \mathbf{L}^{-1} = \begin{bmatrix} 0.1111 & 0 \\ -4.6355 & 6.3640 \end{bmatrix}$$

Si ha pertanto

$$\mathbf{G} = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^T = \begin{bmatrix} 2.827160 & -2.688751 \\ -2.688751 & 1.672839 \end{bmatrix}$$

che ha come autovalori  $[4.999999, -0.499999]$ . Osserviamo, anche, che posto

$$\mathbf{X} = \begin{bmatrix} 3 & -5 \\ -4 & 7 \end{bmatrix}$$

si ha  $\mathbf{X}^T\mathbf{A}\mathbf{X} = \mathbf{diag}(5, -1)$  e  $\mathbf{X}^T\mathbf{B}\mathbf{X} = \mathbf{diag}(1, 2)$ , da cui la conferma del risultato precedente. ■

### 3.5 Decomposizione SVD

Sia  $\mathbf{A}$  una matrice  $\in \mathbb{R}^{m \times n}$ . Senza perdita di generalità possiamo assumere  $m \geq n$  (altrimenti, si sostituisce  $\mathbf{A}$  con  $\mathbf{A}^T$ ). La decomposizione SVD può, allora, essere scritta nella forma

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma} \\ 0 \end{bmatrix} \mathbf{V}^T, \quad \mathbf{\Sigma} = \mathbf{diag}(\sigma_1, \dots, \sigma_n), \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0 \quad (3.28)$$

ove  $\mathbf{U}$  è una matrice ortogonale di ordine  $m$  e  $\mathbf{V}$  una matrice ortogonale di ordine  $n$ . Sappiamo che i valori singolari sono tali che  $\sigma_i^2$  sono gli autovalori della matrice  $\mathbf{A}^T\mathbf{A}$ . In teoria, quindi, essi potrebbero essere calcolati risolvendo il problema degli autovalori per la matrice simmetrica  $\mathbf{A}^T\mathbf{A}$ ; ma questa procedura può introdurre una perdita di accuratezza durante il calcolo della matrice  $\mathbf{A}^T\mathbf{A}$ .

Una procedura più opportuna per calcolare la SVD è stata introdotta da Golub e Kahan (1965). In tale procedura, il primo passo consiste nel ridurre, mediante un algoritmo basato sulle trasformazioni di Householder, la matrice  $\mathbf{A}$  ad una forma *bidiagonale*  $\mathbf{B}$ . Successivamente, si applica alla matrice tridiagonale  $\mathbf{B}^T\mathbf{B}$  un metodo QR con shift.

Più in dettaglio, mediante il metodo di Householder si calcola una matrice  $m \times m$  di Householder  $\mathbf{P}_1$  in modo da annullare gli elementi della sottodiagonale della prima

colonna di  $\mathbf{A}$ . La matrice  $\mathbf{A}' = \mathbf{P}_1 \mathbf{A}$  è illustrata nel seguente schema, nel quale  $m = 5$ ,  $n = 4$  e gli elementi che cambiano sono indicati con il simbolo \*

$$\mathbf{A} = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix} \rightarrow \mathbf{P}_1 \mathbf{A} = \mathbf{A}' = \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix}$$

Si determina, quindi, una matrice di Householder  $\mathbf{Q}_1$  di ordine  $n$  della seguente forma

$$\mathbf{Q}_1 = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{\mathbf{Q}} \end{bmatrix}$$

in maniera che

$$\mathbf{A}' = \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{bmatrix} \rightarrow \mathbf{A}' \mathbf{Q}_1 = \mathbf{A}'' = \begin{bmatrix} x & * & 0 & 0 \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix}$$

La matrice  $\mathbf{A}''$  ha quindi nulli gli elementi della prima riga che hanno indice di colonna maggiore o uguale a 3 e gli elementi della prima colonna che hanno indice di riga maggiore o uguale a 2. Ripetendo il procedimento per  $n - 2$  volte, cioè annullando alternativamente gli elementi delle colonne e delle righe, si arriva a una matrice bidiagonale.

◆ **Esercizio 3.1** *Esaminare il condizionamento del problema del calcolo degli autovalori della seguente matrice*

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

calcolando gli autovalori  $\lambda_i(\epsilon)$  e gli autovettori  $\mathbf{x}_i(\epsilon)$  delle seguenti matrici perturbate

$$\begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 0 & 1 + \epsilon \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 + \epsilon \\ 0 & 1 \end{bmatrix}$$

con  $\epsilon > 0$  e piccolo.

◆ **Esercizio 3.2** *Si applichi il metodo delle potenze alla seguente matrice*

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 & 4 \\ 1 & -3 & 1 & 5 \\ 3 & 1 & 6 & -2 \\ 4 & 5 & -2 & -1 \end{bmatrix}$$

con vettore iniziale  $\mathbf{z}_0 = [1, 1, 1, 1]^T$ . Spiegare il comportamento nella convergenza del metodo.

◆ **Esercizio 3.3** Trasformare la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}$$

in forma tridiagonale con i metodi di Lanczos, di Householder e di Givens.

◆ **Esercizio 3.4** Sia

$$\mathbf{A} = \mathbf{QR}, \quad \mathbf{Q}^* \mathbf{Q} = \mathbf{I}, \quad \mathbf{R} \text{ triangolare superiore, } \mathbf{A}' = \mathbf{RQ}$$

Mostrare che

1. se  $\mathbf{A}$  è una matrice di Hessenberg, allora anche  $\mathbf{A}'$  è una matrice di Hessenberg;
2. se  $\mathbf{A}$  è tridiagonale hermitiana, pure  $\mathbf{A}'$  è tridiagonale hermitiana.

◆ **Esercizio 3.5** Date due matrici di ordine  $n$   $\mathbf{A}$  e  $\mathbf{B}$ , considerare il problema degli autovalori  $(\mathbf{AB})\mathbf{x} = \lambda\mathbf{x}$ . Se  $\mathbf{B}$  è simmetrica e definita positiva e  $\mathbf{LL}^T$  è la corrispondente fattorizzazione di Cholesky, mostrare che le matrici  $\mathbf{AB}$  e  $\mathbf{L}^T \mathbf{AL}$  hanno gli stessi autovalori. Esaminare la relazione tra gli autovettori di  $\mathbf{AB}$  e gli autovettori di  $\mathbf{L}^T \mathbf{AL}$ .

◆ **Esercizio 3.6** Le matrici come la seguente

$$\mathbf{X}_3 = \begin{bmatrix} x_0 & x_1 & x_2 \\ x_2 & x_0 & x_1 \\ x_1 & x_2 & x_0 \end{bmatrix}$$

sono dette matrici circolanti e hanno interesse in diverse applicazioni (trasformata di Fourier). Trovare tutti gli autovalori e i corrispondenti autovettori di  $\mathbf{X}_3$ . Usare tali risultati per costruire una matrice  $\mathbf{U}$  tale che  $\mathbf{U}^{-1} \mathbf{X}_3 \mathbf{U}$  sia diagonale. Generalizzare i risultati ottenuti al caso di matrici circolanti generali.



Approach your problem from the right end and begin with the answers.

Then one day, perhaps you will find the final question.

R. van Gulik, *The Chinese Maze Murders*

## Capitolo 4

# Approssimazione di funzioni

Il problema dell'*approssimazione di funzioni*, di importanza fondamentale nella *matematica applicata*, consiste nella sostituzione di una funzione *complicata* con una *più semplice*, scelta nell'ambito di una classe fissata di funzioni a *dimensione finita*. Per chiarire il senso di tale affermazione, esaminiamo due situazioni in cui si può presentare la necessità di un'operazione di questo tipo.

- Della funzione da approssimare  $f(x)$  *non è nota* una espressione analitica, ma di essa si conoscono alcuni valori  $y_i$  in corrispondenza ad un insieme di punti  $x_i$ ,  $i = 0, \dots, n$ , e si vuole avere indicazioni sul comportamento della funzione in *altri* punti. È il caso di una funzione assegnata sotto forma di *tabella* finita, corrispondente, ad esempio, ad un insieme di *dati sperimentali*. In questo caso *approssimare* la funzione  $f(x)$  significa fornire un *modello* rappresentativo del fenomeno che ha dato origine ai dati sperimentali. In tale senso, quello che vedremo in questo capitolo costituisce una introduzione alla *modellizzazione* matematica.
- La funzione è *nota* in *forma analitica*, ma la sua sostituzione con funzioni più semplici è richiesta per il suo *calcolo numerico* su calcolatore, oppure per rendere più agevoli certe operazioni funzionali, quali ad esempio l'*integrazione* o la *derivazione*.

Naturalmente, il problema può essere affrontato con *differenti* tecniche, che per essere adeguate devono tenere conto della situazione specifica. Importante in questo senso è la scelta della *distanza* con la quale si vuole *approssimare* e che misura l'*errore*. Ad esempio, per approssimare dati sperimentali affetti da variabilità statistica, può essere opportuna una distanza di tipo *minimi quadrati*, definita dalla somma dei quadrati degli errori nei singoli nodi  $x_i$ ; se invece il numero dei dati è piccolo e i dati non sono affetti da errori casuali, può essere adeguata una operazione

di *interpolazione*, basata sulla richiesta che la funzione approssimante passi esattamente per i punti  $(x_i, y_i)$  assegnati. Infine, quando si approssima la funzione  $f(x)$  allo scopo di avere una procedura di calcolo da utilizzare su calcolatore, si richiede che l'approssimazione presenti un errore distribuito in maniera uniforme su un determinato intervallo; in questo caso una distanza appropriata è rappresentata dalla cosiddetta *norma del massimo*, definita come il massimo del modulo della differenza tra la funzione e la sua approssimante per tutti i punti dell'intervallo.

In questo capitolo esamineremo dapprima la tecnica dell'*interpolazione* (o *collocazione*), che costituisce, in particolare, la base di partenza per la costruzione di diversi metodi numerici per il calcolo di *integrali definiti* e per l'approssimazione della soluzione di *equazioni differenziali*. Successivamente, daremo una introduzione alle idee di base nell'approssimazione polinomiale nel senso dei minimi quadrati e nella norma del massimo (nota anche come *miglior approssimazione*). Gli aspetti numerici della approssimazione secondo i minimi quadrati sono già stati considerati nel precedente Capitolo 2 (cfr. anche Appendice A); gli aspetti statistici sono analizzati in particolare nel Capitolo 8; infine, diverse esemplificazioni del metodo come strumento per la identificazione di modelli matematici sono considerate nei Capitoli 11, 12 e 13.

## 4.1 Interpolazione

In maniera schematica, la procedura di interpolazione<sup>1</sup> è definita nel modo seguente. Dati i valori  $f_i = f(x_i)$  di una funzione  $f(x)$  nei punti  $x_1, x_2, \dots, x_{n+1}$  dell'asse reale, si cerca una funzione  $P(x)$  tale che

$$P(x_i) = f_i, \quad i = 1, 2, \dots, n + 1$$

Si dice allora che la funzione  $P(x)$  *interpola*  $f(x)$  nei punti (o *nodi*)  $x_i, i = 1, \dots, n + 1$ . Quando la funzione di interpolazione  $P(x)$  è utilizzata per stimare il valore della funzione  $f(x)$  al di fuori dell'intervallo formato dai punti  $x_i$ , si parla solitamente di una operazione di *estrapolazione*.

Come funzioni interpolanti  $P(x)$  possono essere utilizzati differenti tipi di funzioni; ad esempio, i polinomi algebrici, i polinomi trigonometrici, le funzioni razionali, ecc. Nel seguito, ci limiteremo, per brevità, alla considerazione dei polinomi algebrici e alle funzioni che sono costruite a partire dai polinomi, in particolare le *funzioni polinomiali a tratti*, o funzioni spline.

In effetti, l'insieme dei polinomi algebrici costituisce, per ragioni sia storiche che pratiche, la più importante classe di funzioni interpolanti. Essi hanno in particolare

---

<sup>1</sup>Indicata in forma espressiva come la scienza *of reading between the lines of a mathematical table* da G. Robinson, che insieme a E. Whittaker pubblicò nel 1924 quello che può essere considerato il primo testo dell'analisi numerica moderna.

il vantaggio di essere facilmente calcolabili; inoltre, per essi possono essere eseguite agevolmente le operazioni di somma, moltiplicazione, integrazione e derivazione.

Naturalmente, le ragioni precedenti non sono sufficienti per far ritenere i polinomi una “buona classe” di funzioni approssimanti. A tale scopo occorre anche che i polinomi siano in grado di “descrivere” adeguatamente funzioni anche non “troppo” regolari. Una risposta positiva a questo requisito è fornita dal seguente importante risultato, che in sostanza assicura la possibilità di approssimare mediante un polinomio di grado conveniente una *qualsiasi funzione continua*  $f(x)$  su un intervallo limitato e chiuso.

**Teorema 4.1 (Weierstrass)** *Se  $f(x)$  è una generica funzione continua sull'intervallo limitato e chiuso  $[a, b]$ , per ogni  $\epsilon > 0$  esiste un polinomio  $P_n(x)$  di grado  $n(\epsilon)$ , cioè di grado dipendente da  $\epsilon$ , tale che*

$$\max_{x \in [a, b]} |f(x) - P_n(x)| < \epsilon$$

Osserviamo, comunque, che il teorema di Weierstrass<sup>2</sup> assicura l'esistenza del polinomio approssimante, ma non implica che tale polinomio sia un particolare polinomio di interpolazione. In sostanza, mentre è confortante sapere che è possibile approssimare una funzione continua  $f(x)$  con una fissata accuratezza mediante un polinomio, non è garantito che tale polinomio possa essere ottenuto con un algoritmo conveniente.

#### 4.1.1 Interpolazione mediante polinomi

Consideriamo il seguente esempio introduttivo.

► **Esempio 4.1** Determinare il polinomio di secondo grado che interpola i seguenti valori

$x_i$	0.2	0.8	1.2
$f_i$	0.3456	0.5867	0.2151

Se scriviamo il polinomio nella seguente forma

$$P(x) = c_0 + c_1(x - 0.2) + c_2(x - 0.2)(x - 0.8)$$

---

<sup>2</sup>Karl Weierstrass (1815-1897) diede una dimostrazione non costruttiva del teorema nel lavoro *Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen reeller Argumente* (Sitzg. ber. Kgl. Preuss. Akad. d. Wiss. Berlin, 1885). Weierstrass è considerato uno dei fondatori della moderna teoria delle funzioni; in particolare, diede notevole impulso allo studio delle serie di potenze e alle applicazioni della matematica alla risoluzione di problemi in fisica e astronomia. Altre interessanti dimostrazioni del teorema di approssimazione furono date da E. Landau (1908) e H. Lebesgue (1908); ricordiamo anche una sua generalizzazione agli spazi topologici dovuta a M. H. Stone (1948). Una approssimazione *costruttiva* venne data, nell'ambito della teoria della probabilità, da S. H. Bernstein (*Démonstration du théorème de Weierstrass fondé sur le calcul des probabilités*, 1912) utilizzando la particolare successione di polinomi (detti poi *polinomi di Bernstein*) che analizzeremo nel seguito del capitolo (cfr. paragrafo 4.1.6).

le condizioni  $P(x_i) = f_i$  forniscono il sistema lineare

$$\begin{aligned} c_0 &= 0.3456 \\ c_0 + c_1(0.8 - 0.2) &= 0.5867 \\ c_0 + c_1(1.2 - 0.2) + c_2(1.2 - 0.2)(1.2 - 0.8) &= 0.2151 \end{aligned}$$

Trattandosi di un sistema di tipo triangolare, la soluzione è calcolata con una semplice sostituzione in avanti e si ottengono i valori

$$c_0 = 0.3456; \quad c_1 = 0.4018; \quad c_2 = -1.3308$$

da cui il polinomio (cfr. Figura 4.1)

$$P(x) = 0.3456 + 0.4018(x - 0.2) - 1.3308(x - 0.2)(x - 0.8)$$

Osserviamo che i primi due termini del polinomio  $P(x)$  sono l'equazione della retta che interpola  $f$  nei punti  $x = 0.2$  e  $x = 0.8$ . Come si vede, il polinomio interpolatore di secondo grado è ottenuto semplicemente aggiungendo al polinomio interpolatore di primo grado un opportuno polinomio di secondo grado. ■

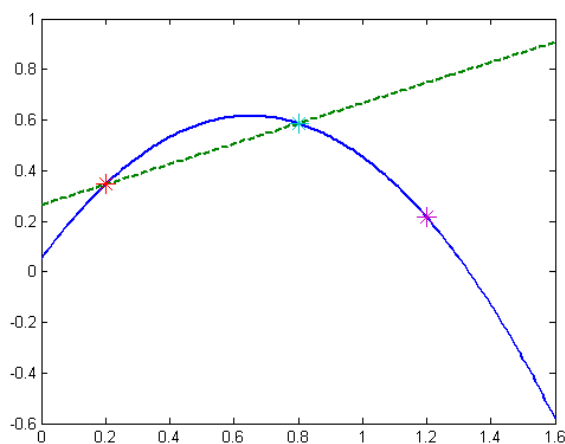


Figura 4.1: Con linea continua è rappresentato il polinomio interpolatore di secondo grado relativo ai punti  $(0.2, 0.3456)$ ,  $(0.8, 0.5867)$ ,  $(1.2, 0.2151)$  e con linea tratteggiata il polinomio interpolatore di primo grado relativo ai punti  $(0.2, 0.3456)$ ,  $(0.8, 0.5867)$ .

Il risultato di base per l'interpolazione polinomiale è il seguente teorema.

**Teorema 4.2 (Esistenza e unicità)** *Siano  $x_1, x_2, \dots, x_{n+1}$   $n + 1$  punti distinti arbitrari. Per ogni insieme di valori  $f_1, f_2, \dots, f_{n+1}$  esiste uno ed un solo polinomio  $P_n(x)$  di grado  $\leq n$  tale che*

$$P_n(x_i) = f_i, \quad i = 1, 2, \dots, n + 1$$

**DIMOSTRAZIONE.** Incominciamo a dimostrare l'*unicità*. Ragionando per assurdo, supponiamo che esistano due differenti polinomi  $P_n(x)$  e  $Q_n(x)$  di grado  $\leq n$  che verificano le condizioni

$$\begin{aligned} P_n(x_i) &= f_i \\ Q_n(x_i) &= f_i \end{aligned} \quad i = 1, 2, \dots, n+1$$

Questo significa che il polinomio  $P_n(x) - Q_n(x)$ , di grado minore o uguale a  $n$ , ha  $n+1$  zeri distinti, corrispondenti ai nodi  $x_1, x_2, \dots, x_{n+1}$ . In base al teorema fondamentale dell'algebra, tale polinomio deve essere identicamente nullo, cioè  $P_n(x) \equiv Q_n(x)$ .

L'*esistenza* può essere dimostrata in differenti maniere. Ci limitiamo a segnalare le due seguenti, di interesse costruttivo, cioè algoritmico (cfr. successivo paragrafo 4.1.4). Riprendendo l'idea suggerita dall'Esempio 4.1, si può procedere per induzione su  $n$  nel modo seguente. Per  $n=1$  si verifica direttamente che il polinomio di interpolazione è dato da

$$P_1(x) = f_1 + \frac{f_2 - f_1}{x_2 - x_1} (x - x_1)$$

Assumendo, ora, che  $P_{k-1}(x)$  sia un polinomio di grado  $\leq k-1$  tale che

$$P_{k-1}(x_i) = f_i, \quad i = 1, 2, \dots, k$$

mostriamo che esiste un polinomio  $P_k$ , di grado  $\leq k$  interpolante  $f$  nei nodi  $x_1, x_2, \dots, x_{k+1}$ . Poniamo

$$P_k(x) = P_{k-1}(x) + c(x - x_1)(x - x_2) \cdots (x - x_k)$$

Il polinomio  $P_k(x)$  ha grado  $\leq k$ . Inoltre

$$P_k(x_i) = f_i, \quad i = 1, 2, \dots, k$$

Poiché i punti  $x_1, x_2, \dots, x_{k+1}$  sono distinti, si può determinare  $c$  in maniera che

$$P_k(x_{k+1}) = f_{k+1} \Rightarrow c = \frac{f_{k+1} - P_{k-1}(x_{k+1})}{(x_{k+1} - x_1)(x_{k+1} - x_2) \cdots (x_{k+1} - x_k)}$$

Allora,  $P_k$  è il polinomio di grado  $\leq k$  che interpola  $f$  in  $x_1, x_2, \dots, x_{k+1}$ .

Un secondo modo di dimostrare l'esistenza consiste nel fornire una rappresentazione esplicita del polinomio di interpolazione. A tale scopo consideriamo i seguenti polinomi di grado  $n$

$$L_i(x) = \frac{(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_{n+1})}{(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_{n+1})} \quad (4.1)$$

per  $i = 1, 2, \dots, n+1$ . Tali polinomi sono i particolari polinomi di interpolazione che assumono il valore zero nei punti  $x_j$ , con  $j \neq i$ , e il valore 1 nel punto  $x_i$  (per un esempio si veda Figura 4.2). Si verifica, allora, facilmente la seguente rappresentazione del polinomio

di interpolazione  $P_n(x)$ , detta *rappresentazione di Lagrange*<sup>3</sup>

$$P_n(x) = \sum_{i=1}^{n+1} f_i L_i(x) \quad (4.2)$$

■

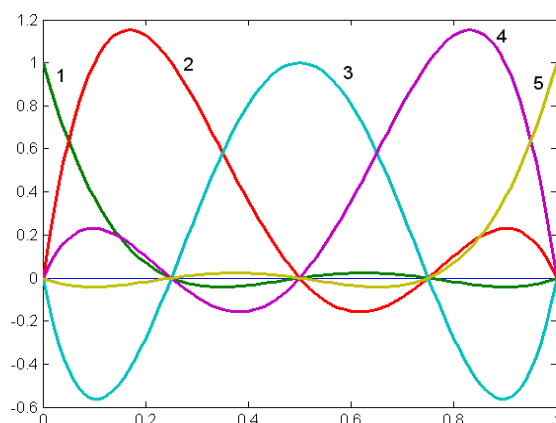


Figura 4.2: Polinomi di Lagrange  $L_i(x)$  relativi a 5 punti equidistanti sull'intervallo  $[0, 1]$ .

#### 4.1.2 Errore di troncamento nella interpolazione

Nei nodi  $x_i$  il polinomio di interpolazione assume, per definizione, i valori  $f_i$  e pertanto in tali punti l'errore è nullo, almeno se si prescinde dagli errori di arrotondamento. Lo studio dell'errore ha quindi senso per valori di  $x \neq x_i$  e quando  $f_i = f(x_i)$ , con  $f(x)$  funzione assegnata su un determinato intervallo. Si chiama allora *errore di troncamento* di interpolazione la quantità  $E(x) := f(x) - P_n(x)$ , e una sua stima è importante per un corretto utilizzo pratico del polinomio di interpolazione. Si intuisce, comunque, che una tale stima non è ottenibile con la sola conoscenza della funzione  $f(x)$  nei nodi  $x_i$ ; ad esempio, per una funzione discontinua l'errore di troncamento può essere arbitrariamente grande. In effetti, la regolarità della funzione

<sup>3</sup>La formula di rappresentazione (4.2) appare, in effetti, in un lavoro di Lagrange del 1795, ma era stata precedentemente utilizzata da Eulero (1755) e Waring (1779). Lo studio delle formule di interpolazione prende sostanzialmente avvio a partire dal 17° secolo, a seguito della necessità di costruire tabelle di valori delle funzioni trigonometriche e dei logaritmi. Lo sviluppo della teoria è legata, in particolare, ai nomi di T. Harriot (1611), H. Briggs (1617), I. Newton (1642-1727), J. Gregory (1638-1675), B. Taylor (1685-1731), C. Maclaurin (1698-1746), J. Stirling (1692-1770) e K. F. Gauss (1777-1855).

$f(x)$  può essere opportunamente utilizzata per fornire una rappresentazione dell'errore, dalla quale, almeno in teoria, è possibile ricavare una maggiorazione dell'errore e una indicazione sul suo comportamento al variare del numero e della collocazione dei nodi  $x_i$ . Importante in questo senso è il seguente risultato che ricordiamo senza dimostrazione, che è comunque una semplice conseguenza del Teorema di Rolle.

**Teorema 4.3** (Rappresentazione dell'errore di interpolazione) *Sia  $f(x)$  una funzione con  $n + 1$  derivate continue in un intervallo  $\mathcal{I} = [a, b]$  contenente i punti  $x_i, i = 1, 2, \dots, n + 1$ . Se  $P_n(x)$  è il polinomio di interpolazione della funzione  $f(x)$  nei punti  $\{x_i\}$ , allora per ogni  $x \in [a, b]$  esiste un punto  $\xi = \xi(x) \in (a, b)$ , tale che*

$$E(x) := f(x) - P(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_1)(x - x_2) \cdots (x - x_{n+1}) \quad (4.3)$$

La formula (4.3) mette in evidenza il fatto importante che l'errore di interpolazione dipende sia dal comportamento della derivata di ordine  $n + 1$  della funzione interpolata, sia dalla scelta, attraverso il polinomio  $\pi(x) := (x - x_1)(x - x_2) \cdots (x - x_{n+1})$ , dei nodi  $x_i$  e del punto di valutazione  $x$  del polinomio interpolatore. Illustreremo il risultato mediante un esempio.

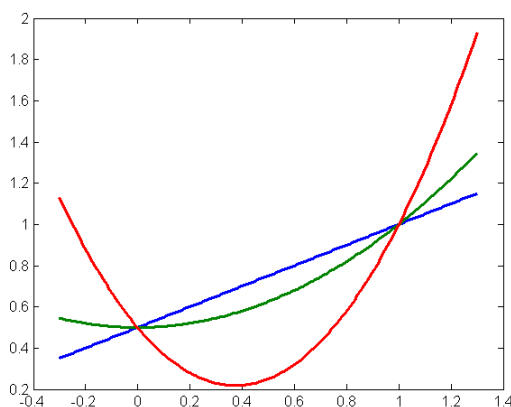


Figura 4.3: Illustrazione della dipendenza dalla derivata seconda dell'errore di troncamento nella interpolazione lineare.

► **Esempio 4.2** Se  $f(x)$  è due volte continuamente derivabile ed è approssimata per  $x_1 \leq x \leq x_2 = x_1 + h$  mediante una interpolazione lineare nei punti  $(x_1, f_1)$  e  $(x_2, f_2)$ , dalla rappresentazione (4.3) si ha

$$|f(x) - P_1(x)| = \frac{f''(\xi(x))}{2!} (x - x_1)(x - x_2) \Rightarrow |f(x) - P_1(x)| = \frac{f''(\xi(x_1 + sh))}{2!} h^2 s(s-1) \quad (4.4)$$

ove si è posto  $x = x_1 + sh$ . Tenendo conto che  $\max_{0 \leq s \leq 1} |s(s-1)| = \frac{1}{4}$ , si ha quindi

$$|f(x) - P_1(x)| \leq \frac{h^2}{8} \max_{x_1 \leq x \leq x_1+h} |f''(x)| \quad (4.5)$$

La Figura 4.3 evidenzia in maniera grafica l'importanza per l'errore di troncamento del comportamento della derivata seconda.

Come esemplificazione, consideriamo la funzione  $f(x) = \sqrt{x}$ , per  $x > 0$ . Utilizziamo il polinomio di interpolazione di primo grado corrispondente ai punti  $(0.25, 0.5)$ ,  $(0.49, 0.7)$  per calcolare  $\sqrt{0.35}$ . Dalla rappresentazione di Lagrange si ha

$$P_1(x) = 0.5 \frac{x - x_2}{x_1 - x_2} + 0.7 \frac{x - x_1}{x_2 - x_1} = -\frac{0.5}{0.24}(x - x_2) + \frac{0.7}{0.24}(x - x_1)$$

da cui  $P_1(\sqrt{0.35}) = 0.58\bar{3}$ , con un errore pari a 0.008274644. Tenendo conto che il modulo della derivata seconda  $f''(x) = -1/(4\sqrt{x^3})$  assume il valore massimo nel primo estremo  $x_1$ , la stima (4.5) fornisce la limitazione  $h^2/(16\sqrt{0.25^3}) = 0.0288$ . ■

### 4.1.3 Convergenza del polinomio di interpolazione

La rappresentazione (4.3) pone le basi per lo studio di un'altra importante questione relativa all'errore di troncamento. Dal punto di vista pratico, è essenziale poter ridurre tale errore inferiore a una quantità prefissata, scegliendo convenientemente il numero e la posizione dei nodi  $x_i$ . Dal punto di vista matematico, questo significa studiare la *convergenza* del metodo di approssimazione. In maniera più precisa, il problema si pone nella seguente forma. Data una funzione  $f(x)$  definita su un intervallo  $[a, b]$ , e sufficientemente regolare, consideriamo, per ogni  $n$  intero  $\geq 1$ , una suddivisione dell'intervallo  $[a, b]$  mediante i punti  $a \leq x_1 < x_2 < \dots < x_{n+1} \leq b$  e sia  $P_n(x)$  il polinomio di interpolazione nei nodi  $(x_i, f(x_i))$ ,  $i = 1, 2, \dots, n+1$ . Lo studio della *convergenza* riguarda allora l'analisi dell'errore  $f(x) - P_n(x)$ , per  $x \in [a, b]$  e al tendere di  $n$  all'infinito (più precisamente, al tendere a zero dell'ampiezza degli intervallini della suddivisione). Contrariamente all'intuizione (supportata anche dal Teorema di Weierstrass), la risposta a tale questione non è sempre positiva; ossia, anche se la funzione  $f(x)$  ammette derivate di ogni ordine su  $[a, b]$ , la successione  $\{P_n(x)\}$  dei valori assunti dal polinomio di interpolazione di grado  $n$  in un punto  $x \in [a, b]$  può non convergere a  $f(x)$ , per  $n \rightarrow \infty$ . In Figura 4.4, è illustrato un classico controesempio, noto come *esempio di Runge*<sup>4</sup>. Si considera la funzione

$$f(x) = \frac{1}{1 + 25x^2} \quad (4.6)$$

<sup>4</sup>C. D. T. Runge (1856-1927) tenne per primo la cattedra di matematica applicata nell'Università di Göttingen; i suoi interessi principali riguardarono lo studio dei problemi matematici in fisica, geodesia e astronomia e in generale l'applicazione dei metodi numerici alle scienze applicate ("The value of a mathematical discipline depends on its applicability to the natural sciences", 1880). L'esempio esaminato nel testo è contenuto in un lavoro del 1901; S. N. Bernstein (1912) dimostrò che l'interpolazione della funzione  $f(x) = |x|$  su nodi equidistanti dell'intervallo  $[-1, 1]$  diverge per ogni  $x \neq -1, 0, 1$ .



sull'intervallo  $[-1, 1]$ , interpolata mediante i polinomi  $P_n(x)$  nei punti equidistanti

$$x_{n,i} = -1 + (i - 1) \frac{2}{n}, \quad i = 1, \dots, n + 1$$

Come si vede in figura, l'errore è piccolo nei punti vicini all'origine, ma grande vicino agli estremi  $-1$  e  $+1$ . In effetti, si può dimostrare che

$$\lim_{n \rightarrow \infty} \left( \max_{-1 \leq x \leq 1} |f(x) - P_n(x)| \right) = \infty$$

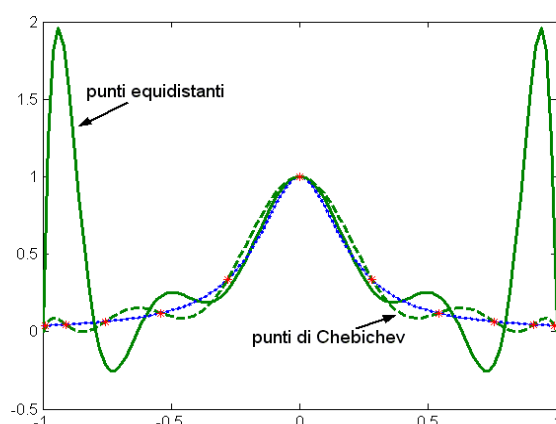


Figura 4.4: Esempio di Runge; polinomio di interpolazione di grado 10 rispettivamente in punti equidistanti (---) e negli zeri del polinomio di Chebichev (-.).

Più in generale, si può mostrare (Bernstein e Faber, 1914) che per una scelta arbitraria dei punti di interpolazione in  $[a, b]$ , è possibile costruire una funzione continua su  $[a, b]$  per cui il procedimento di interpolazione sui nodi fissati è tale che  $\max_{a \leq x \leq b} |f(x) - P_n(x)| \rightarrow \infty$ .

Come risultato significativo di convergenza, ricordiamo che per ogni funzione  $f(x)$  sufficientemente regolare, ad esempio dotata di derivata seconda continua, è possibile trovare un insieme di punti di interpolazione per i quali il polinomio di interpolazione converge uniformemente. Sempre in Figura 4.4 è indicato il comportamento del polinomio di interpolazione, per  $n = 10$ , relativo ai punti

$$x_{n,i}^* = \cos\left(\frac{2i - 1}{n + 1} \frac{\pi}{2}\right), \quad i = 1, \dots, n + 1$$

che corrispondono agli zeri del *polinomio di Chebichev* di grado  $n + 1$  relativo all'intervallo  $[-1, 1]$ . Tali polinomi verranno introdotti e analizzati successivamente (cfr. paragrafo 4.2.2).

#### 4.1.4 Costruzione del polinomio di interpolazione

Nel Teorema 4.2 sono stati introdotti due differenti modi per costruire il polinomio di interpolazione relativo ai nodi  $x_1, x_2, \dots, x_{n+1}$ . Analizzeremo, ora, tali procedure dal punto di vista computazionale.

Incominciamo a ricordare che la rappresentazione di Lagrange (4.2) è la base per la costruzione di formule di quadratura e di schemi alle differenze per l'approssimazione della soluzione di equazioni differenziali (cfr. rispettivamente il Capitolo 6 e il Capitolo 7), ed ha quindi un indubbio interesse teorico. Dal punto di vista pratico, ossia per il calcolo effettivo del polinomio di interpolazione, essa può risultare vantaggiosa quando si considera la costruzione di polinomi di interpolazione relativi allo stesso insieme di nodi  $x_i$  e a differenti dati  $f_i$ . In tale caso, infatti, i polinomi  $L_i(x)$ , che dipendono solo dai punti  $x_i$ , possono essere calcolati a priori una sola volta. Al contrario, quando si costruiscono polinomi di interpolazione di grado crescente, cioè si aggiungono ulteriori punti per migliorare la precisione, i polinomi  $L_i(x)$  devono essere successivamente ricostruiti. In tali situazioni risultano allora più convenienti altre procedure. In particolare, esamineremo un algoritmo basato sulla *rappresentazione di Newton*, e utile per fornire i coefficienti del polinomio, e l'*algoritmo di Neville*, che fornisce il valore del polinomio in un punto particolare.

#### Polinomio di interpolazione di Newton

Un polinomio di grado  $n \geq 1$  può essere scritto in differenti forme tra loro equivalenti. In particolare, dati i punti  $x_i, i = 1, 2, \dots, n + 1$ , si chiama *rappresentazione di Newton* la seguente forma

$$P(x) = c_0 + c_1(x - x_1) + c_2(x - x_1)(x - x_2) + \dots + c_n(x - x_1)(x - x_2) \dots (x - x_n) \quad (4.7)$$

A partire da tale rappresentazione, il valore di  $P(x)$  in un particolare punto  $x = z$  può essere ottenuto mediante l'algoritmo di Horner-Ruffini, adattato nel seguente modo. Mettendo in evidenza i fattori comuni, si ha

$$P(x) = c_0 + (x - x_1)(c_1 + (x - x_2)(c_2 + \dots + c_n(x - x_n) \dots)) \iff \begin{array}{l} \text{P=C(N)} \\ \text{DO 2 I=N,1,-1} \\ \text{2 P=P*(Z-X(I))+C(I-1)} \end{array}$$

Quando  $P(x)$  è il polinomio di interpolazione corrispondente ai punti  $(x_i, f_i), i = 1, 2, \dots, n + 1$ , i coefficienti  $c_i$  possono essere calcolati dalle condizioni  $P(x_i) = f_i$  con una procedura iterativa che introdurremo nel caso  $n = 2$ , ossia

$$\begin{aligned} c_0 &= f_1 \\ c_0 + c_1(x_2 - x_1) &= f_2 \\ c_0 + c_1(x_3 - x_1) + c_2(x_3 - x_1)(x_3 - x_2) &= f_3 \end{aligned}$$

da cui, mediante sostituzione

$$c_0 = f_1, \quad c_1 = \frac{f_2 - f_1}{x_2 - x_1}$$

$$c_2 = \frac{f_3 - f_1 - \frac{x_3 - x_1}{x_2 - x_1}(f_2 - f_1)}{(x_3 - x_1)(x_3 - x_2)} = \frac{\frac{f_3 - f_1}{x_3 - x_1} - \frac{f_2 - f_1}{x_2 - x_1}}{x_3 - x_2}$$

Dalle relazioni precedenti si vede che i coefficienti  $c_1$ ,  $c_2$  sono delle approssimazioni rispettivamente della derivata prima e seconda, ossia hanno un significato di “rapporti incrementali”; inoltre, il coefficiente  $c_2$  può essere facilmente calcolato quando si conoscono i rapporti incrementali di ordine inferiore. La procedura può essere generalizzata nel modo seguente. Indicando, per  $k = 1, 2, \dots, n$ , con  $P_k(x)$  il polinomio di grado non superiore a  $k$  che interpola i punti  $(x_i, f_i)$ ,  $i = 1, 2, \dots, k + 1$ , si ha

$$P_0(x) = c_0$$

$$P_k(x) = P_{k-1}(x) + c_k(x - x_1)(x - x_2) \cdots (x - x_k), \quad k = 1, \dots, n$$

da cui si vede che il valore di ciascun coefficiente  $c_k$  dipende solo dai valori di  $f$  nei punti  $x_1, x_2, \dots, x_{k+1}$ . È naturale, pertanto, introdurre la seguente notazione

$$c_k = f[x_1, x_2, \dots, x_{k+1}] \quad (4.8)$$

Per il *calcolo ricorsivo* dei coefficienti  $c_k$ , osserviamo che alternativamente  $P_k(x)$  può essere espresso nella seguente forma

$$P_k(x) = P_{k-1}^{(1)}(x) + \frac{x - x_1}{x_{k+1} - x_1} (P_{k-1}^{(2)}(x) - P_{k-1}^{(1)}(x)) \quad (4.9)$$

ove  $P_{k-1}^{(1)}$  (rispettivamente  $P_{k-1}^{(2)}$ ) è il polinomio di grado  $\leq k - 1$  che interpola  $f$  nei punti  $x_1, x_2, \dots, x_k$  (rispettivamente nei punti  $x_2, x_3, \dots, x_{k+1}$ ). In effetti, a secondo membro della relazione (4.9) si ha un polinomio di grado  $\leq k$  che assume, come si verifica facilmente, gli stessi valori del polinomio a primo membro nei punti  $x_1, \dots, x_{k+1}$ .

Uguagliando i coefficienti del termine  $x^k$  nei due polinomi, rispettivamente a primo e secondo membro della uguaglianza (4.9), si ottiene la seguente relazione ricorrente

$$f[x_1, x_2, \dots, x_{k+1}] = \frac{f[x_2, x_3, \dots, x_{k+1}] - f[x_1, x_2, \dots, x_k]}{x_{k+1} - x_1}$$

Per tale motivo il coefficiente  $f[x_1, x_2, \dots, x_k]$  è anche chiamato *differenza divisa* di ordine  $k$ . Osserviamo che il valore di  $f[x_1, x_2, \dots, x_k]$  non cambia se gli argomenti

$x_1, x_2, \dots, x_k$  sono assegnati in ordine differente, cioè la dipendenza dagli argomenti è *simmetrica*. Quando la funzione  $f$  è dotata di derivate continue fino all'ordine  $k$ , si può dimostrare che esiste un numero  $\xi$  nell'intervallo formato dai punti  $x_1, x_2, \dots, x_{k+1}$  tale che

$$f[x_1, x_2, \dots, x_{k+1}] = \frac{f^{(k)}(\xi)}{k!}$$

Il calcolo delle differenze divise è visualizzato, ad esempio per  $n = 3$ , nella seguente tabella

$x_1$	$f_1$			
	$f[x_1, x_2]$			
$x_2$	$f_2$	$f[x_1, x_2, x_3]$		
	$f[x_2, x_3]$	$f[x_1, x_2, x_3, x_4]$		
$x_3$	$f_3$	$f[x_2, x_3, x_4]$		
	$f[x_3, x_4]$			
$x_4$	$f_4$			

A partire dalla rappresentazione (4.7), il polinomio di interpolazione  $P_n(x)$  relativo ai punti  $(x_i, f_i)$ ,  $i = 1, \dots, n + 1$  è allora fornito dalla seguente espressione

$$P(x) = f_1 + f[x_1, x_2](x - x_1) + f[x_1, x_2, x_3](x - x_1)(x - x_2) + \dots + f[x_1, x_2, \dots, x_{n+1}](x - x_1)(x - x_2) \cdots (x - x_n)$$

► **Esempio 4.3** Determinare il polinomio di terzo grado che interpola i valori

$x$	0	2	-2	7
$f$	1	5	3	2

Si ottiene la seguente tabella

0	1			
	$\frac{5-1}{2-0} = 2$			
2	5	$\frac{1/2-2}{-2-0} = \frac{3}{4}$		
	$\frac{3-5}{-2-2} = \frac{1}{2}$	$\frac{-1/9-1/2}{7-2} = -\frac{11}{90}$	$\frac{-11/90-3/4}{7-0} = -\frac{157}{1260}$	
-2	3			
	$\frac{2-3}{7+2} = -\frac{1}{9}$			
7	2			

da cui il polinomio di interpolazione

$$P_3(x) = 1 + 2x + \frac{3}{4}x(x-2) - \frac{157}{1260}x(x-2)(x+2)$$

■

### Implementazione dell'algoritmo di Newton

Una prima forma di implementazione consiste nel calcolo di tutte le differenze divise  $c_{ij} = f[x_{i+1}, x_{i+2}, \dots, x_{j+1}]$  a partire dalla tabella  $(x_i, f_i), i = 1, 2, \dots, n + 1$ .

```

DO 2 I=0,N
2   C(I,I)=F(X(I+1))
DO 3 J=1,N
DO 3 I=0,N-J
3   C(I,I+J)=(C(I+1,I+J)-C(I,I+J-1))/(X(I+J+1)-X(I+1))

```

I coefficienti  $c_i$  della rappresentazione di Newton sono memorizzati nella prima riga dell'array C, cioè  $c_0 = C(0, 0), c_1 = C(0, 1), \dots$ . Alternativamente, per risparmiare posizioni di memoria, ma non quantità di operazioni, si può utilizzare la seguente implementazione, nella quale si ottengono direttamente i coefficienti nell'array C.

```

DO 2 I=0,N
C(I)=F(X(I+1))
2   CONTINUE
DO 3 J=1,N
DO 3 I=N,J,-1
C(I)=(C(I)-C(I-1))/(X(I+1)-X(I+1-J))
3   CONTINUE

```

### Algoritmo di Neville

Dalla relazione (4.9) si vede che il polinomio  $P_k$ , di grado  $\leq k$ , che interpola  $f$  nei punti  $x_1, x_2, \dots, x_{k+1}$  può essere calcolato usando due polinomi di interpolazione  $P_{k-1}$  di grado  $\leq k-1$ . Più precisamente, si ha che  $P_k(x)$  può essere pensato come il risultato della interpolazione lineare tra i due "punti"  $(x_1, P_{k-1}^{(1)}(x))$  e  $(x_2, P_{k-1}^{(2)}(x))$ . Il metodo di Neville, a partire da tale osservazione, calcola il valore del polinomio di interpolazione mediante successive *interpolazioni lineari*.

Per descrivere il metodo introduciamo alcune opportune notazioni. Se  $P_{12}(x)$  indica il polinomio di grado  $\leq 1$ , che interpola  $f$  in  $x_1$  e  $x_2$ , si ha

$$P_{12}(x) = f_1 + \frac{x - x_1}{x_2 - x_1} (f_2 - f_1) = \frac{1}{x_2 - x_1} \begin{vmatrix} x - x_1 & f_1 \\ x - x_2 & f_2 \end{vmatrix}$$

Analogamente se  $P_{123}(x)$  è il polinomio, di grado  $\leq 2$ , che interpola  $f$  nei punti  $x_1, x_2$  e  $x_3$ , si ha

$$P_{123}(x) = \frac{1}{x_3 - x_1} \begin{vmatrix} x - x_1 & P_{12}(x) \\ x - x_3 & P_{23}(x) \end{vmatrix}$$

Più in generale, se  $P_{i_1 i_2 \dots i_{k+1}}(x)$  è il polinomio, di grado  $\leq k$ , che soddisfa

$$P_{i_1 i_2 \dots i_{k+1}}(x_{i_j}) = f_{i_j}, \quad j = 1, 2, \dots, k + 1$$

esso può essere calcolato nella seguente maniera ricorsiva

$$P_i(x) = f_i$$

$$P_{i_1 i_2 \dots i_{k+1}}(x) = \frac{1}{x_{i_{k+1}} - x_{i_1}} \begin{vmatrix} x - x_{i_1} & P_{i_1 i_2 \dots i_k} \\ x - x_{i_{k+1}} & P_{i_2 i_3 \dots i_{k+1}} \end{vmatrix}$$

L'algoritmo di Neville segue, allora, ad esempio per  $n = 3$ , il seguente schema.

$$\begin{array}{l|ll} x - x_1 & x_1 & f_1 = P_1(x) \\ & x_2 & f_2 = P_2(x) \\ & x_3 & f_3 = P_3(x) \\ & x_4 & f_4 = P_4(x) \end{array} \begin{array}{l} \\ P_{12}(x) \\ P_{23}(x) \\ P_{34}(x) \end{array} \begin{array}{l} \\ P_{123}(x) \\ P_{234}(x) \end{array} P_{1234}(x)$$

L'implementazione dell'algoritmo, che lasciamo come esercizio, può essere effettuata in maniera del tutto analoga a quanto visto nel paragrafo precedente nel caso del polinomio di Newton.

► **Esempio 4.4** Calcolare il valore per  $x = 3$  del polinomio di interpolazione relativo ai punti  $(2, 0), (0, -8), (4, 8), (5, 27)$ . Si ha il seguente schema.

$3-x$	$x$	$f(x)$			
1	2	0			
			$\frac{1}{-2} \begin{vmatrix} 1 & 0 \\ 3 & -8 \end{vmatrix} = 4$		
3	0	-8		$\frac{1}{2} \begin{vmatrix} 1 & 4 \\ -1 & 4 \end{vmatrix} = 4$	
			$\frac{1}{4} \begin{vmatrix} 3 & -8 \\ -1 & 8 \end{vmatrix} = 4$		$\frac{1}{3} \begin{vmatrix} 1 & 4 \\ -2 & -5 \end{vmatrix} = 1$
-1	4	8		$\frac{1}{5} \begin{vmatrix} 3 & 4 \\ -2 & -11 \end{vmatrix} = -5$	
			$\frac{1}{1} \begin{vmatrix} -1 & 8 \\ -2 & 27 \end{vmatrix} = -11$		
-2	5	27			

I valori assegnati  $f_i$  corrispondono ai valori assunti dalla funzione  $(x - 2)^3$ , e quindi il valore dell'interpolazione nel punto  $x = 3$  è uguale a  $(3 - 2)^3 = 1$ . ■

### Punti di interpolazione equidistanti

Quando i punti di interpolazione sono equidistanti, le formule di interpolazione possono essere opportunamente semplificate. A tale scopo introduciamo l'operatore alle differenze  $\Delta$ , definito come l'operatore che trasforma la successione  $y = \{y_n\}$  nella successione  $\Delta y = \{y_{n+1} - y_n\}$ . Più semplicemente, si scrive  $\Delta y = y_{n+1} - y_n$ . Gli operatori alle differenze di ordine superiore  $\Delta^k$  sono definiti in maniera ricorsiva

$$\Delta^k y = \Delta^{k-1}(\Delta y)$$

La quantità  $\Delta^k y_n$  è una combinazione lineare di  $y_n, y_{n+1}, \dots, y_{n+k}$ . Supponiamo, ora, che i punti di interpolazione  $x_i$  siano equidistanti

$$x_i = x_1 + (i - 1)h, \quad i = 1, 2, \dots, n + 1$$

ove  $h$  è il passo di tabulazione. Si può dimostrare, ad esempio per induzione, la seguente uguaglianza

$$f[x_1, x_2, \dots, x_{k+1}] = \frac{\Delta^k f_1}{h^k k!}$$

Introducendo la nuova variabile  $s$  mediante la posizione  $x = x_1 + sh$ , e usando la rappresentazione di Newton del polinomio di interpolazione, si ha

$$P_n(x_1 + sh) = f_1 + s\Delta f_1 + \binom{s}{2}\Delta^2 f_1 + \dots + \binom{s}{n}\Delta^n f_1$$

ove le differenze  $\Delta^i f_1, i = 1, 2, \dots, s$  possono essere calcolate mediante solo operazioni di somma.

► **Esempio 4.5** A partire dai punti di interpolazione

$x$	0.5	0.6	0.7
$f(x)$	0.4794	0.5646	0.6442

si ottiene il seguente schema alle differenze

$x$	$f(x)$	$\Delta$	$\Delta^2$
0.5	0.4794		
		852	
0.6	0.5646		-56
		796	
0.7	0.6442		

ove le differenze sono date in unità di  $10^{-4}$ . In questo caso  $h = 0.1, x_1 = 0.5$  e il polinomio di interpolazione di Newton è il seguente

$$P_2(x_1 + sh) = 0.4794 + 0.0852s + \frac{s(s-1)}{2}(-0.0056)$$

■

### 4.1.5 Interpolazione mediante spline

Nel paragrafo precedente abbiamo visto (cfr. in particolare l'esempio di Runge (4.6)) che all'aumento del numero dei punti utilizzati nell'interpolazione non sempre corrisponde una migliore approssimazione della funzione interpolata. Il motivo è essenzialmente dovuto al comportamento altamente oscillatorio dei polinomi di grado elevato. Un modo per ovviare a questo inconveniente consiste nel suddividere

l'intervallo, in cui si vuole approssimare la funzione, in un certo numero di sottointervalli, interpolando quindi la funzione su ciascun sottointervallo con polinomi di grado basso. Si ottiene in questo modo una *approssimazione polinomiale a tratti*, che rispetto alla approssimazione polinomiale su tutto l'intervallo presenta una minore regolarità, ma che permette, in generale, una migliore descrizione della funzione, con vantaggi in particolare nella grafica e nell'approssimazione delle soluzioni di problemi ai limiti per equazioni differenziali (per questo secondo aspetto si veda Capitolo 7).

Assumendo di aver suddiviso l'intervallo di definizione  $[a, b]$  della funzione da approssimare  $f(x)$  mediante i punti  $a \leq x_1 < x_2 < \dots < x_n \leq b$ , e utilizzando, ad esempio, una interpolazione lineare su ogni sottointervallo  $[x_i, x_{i+1}]$ , si ottiene come funzione approssimante una *curva poligonale*, la cui derivata può essere discontinua nei nodi interni  $x_2, x_3, \dots, x_{n-1}$  (cfr. per una esemplificazione la Figura 4.5).

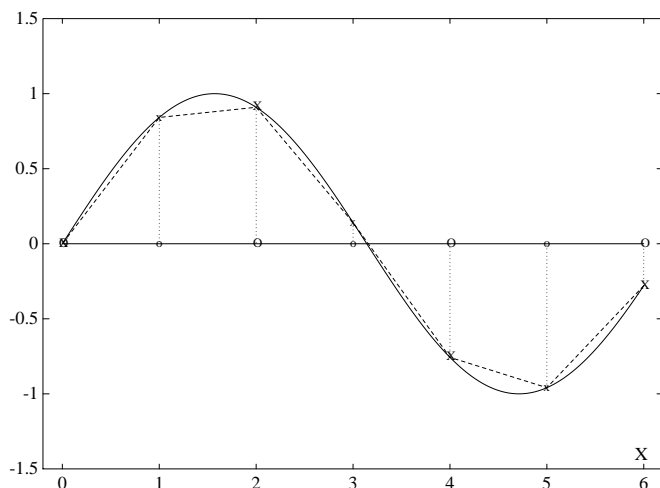


Figura 4.5: Interpolazione lineare a tratti della funzione  $\sin x$  sull'intervallo  $[0, 6]$  relativa a una suddivisione uniforme di ampiezza  $h = 1$ .

Se il grado del polinomio utilizzato è maggiore di uno, la regolarità della funzione approssimante può essere migliorata imponendo nei nodi interni la continuità anche di un numero opportuno di derivate. Ad esempio, come vedremo nel seguito, utilizzando un polinomio di terzo grado, è possibile ottenere una funzione approssimante continua, insieme alla derivata prima e seconda. Tale funzione è chiamata *spline*<sup>5</sup>

<sup>5</sup>*Spline* era il nome utilizzato dai disegnatori per indicare un nastro elastico, vincolato a passare, senza discontinuità e punti angolosi, attraverso una serie di punti prefissati. Il nastro assume la configurazione corrispondente al minimo dell'energia potenziale causata dalla deformazione del materiale, e quindi la configurazione di curvatura minima, compatibilmente con i vincoli. Tale configurazione può essere modellizzata matematicamente mediante una spline cubica interpolante (cfr. Teorema 4.4).



*cubica*. Più in generale, una funzione  $s(x)$  è chiamata *spline* di ordine  $p$ , se  $s(x)$  è definita come un polinomio di grado  $p$  su ciascun sottointervallo ed è continua su tutto l'intervallo insieme alle prime  $p - 1$  derivate. Quando  $s(x_i) = f(x_i)$ , per  $i = 1, 2, \dots, n$ , la funzione  $s$  è detta una *spline interpolante*. Per il seguito considereremo esclusivamente spline interpolanti, e quindi, per brevità, si tralascerà il termine interpolante.

In Figura 4.6 è rappresentata una esemplificazione di una situazione nella quale l'uso di una spline cubica pare più opportuno di un polinomio di interpolazione; più precisamente, è rappresentato il seguente polinomio di sesto grado

$$P(x) = -0.0064 + 5.0295x + 0.0169x^2 - 20.1005x^3 - 0.0486x^4 + 16.0767x^5 + 0.0436x^6$$

e la spline cubica che interpolano i seguenti punti

$x_i$	-1	-0.96	-0.86	-0.79	0.22	0.5	0.930
$f_i$	-1	-0.151	0.894	0.986	0.895	0.5	-0.306

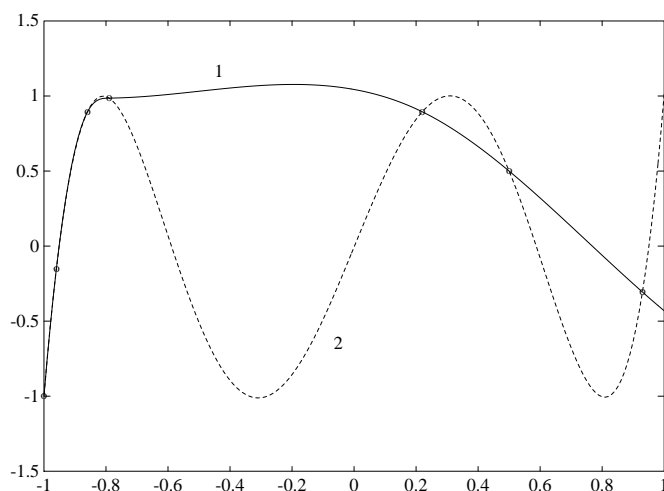


Figura 4.6: (1) Spline cubica interpolante i punti (o); (2) polinomio di interpolazione di Lagrange.

Nel seguito esamineremo più in dettaglio il caso delle spline di primo ordine, ossia le funzioni lineari a tratti, e le spline cubiche.

### Spline lineari

Dati i punti  $x_1 < x_2 < \dots < x_n$ , chiamati *nodi*, una *spline lineare* è una funzione  $s(x)$  con le seguenti proprietà

- $s(x)$  è continua su  $[x_1, x_n]$ ;

- $s(x)$  è un segmento di retta su ogni sottointervallo  $[x_i, x_{i+1}]$ .

Pertanto, su ogni sottointervallo si ha la rappresentazione

$$s(x) = a_i(x - x_i) + b_i, \quad x_i \leq x \leq x_{i+1}$$

e i coefficienti da determinare (i cosiddetti *gradi di libertà*) sono in numero di  $2(n-1)$ . La continuità fornisce  $n-2$  condizioni e i rimanenti  $n$  gradi di libertà possono essere determinati dalle condizioni di interpolazione  $s(x_i) = f_i$ ,  $i = 1, 2, \dots, n$ . Si ottiene, quindi

$$a_i = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}, \quad b_i = f_i$$

Un modo alternativo (di interesse in diverse applicazioni) di rappresentare la funzione  $s(x)$  è il seguente

$$s(x) = \sum_{i=1}^n f_i \psi_i(x)$$

ove le funzioni  $\psi_i(x)$  sono particolari funzioni spline lineari con la proprietà (cfr. Figura 4.7)

$$\psi_i(x_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

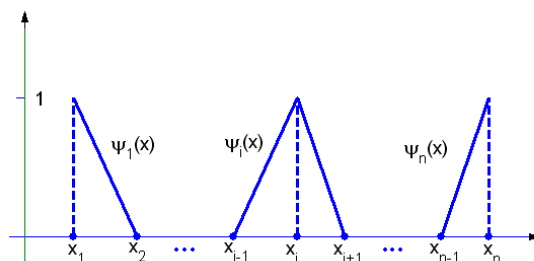


Figura 4.7: Funzioni base per le spline lineari.

Le funzioni  $\psi_i(x)$  costituiscono una *base* per lo spazio di tutte le spline lineari corrispondenti ad un insieme assegnato di nodi. Esse sono anche chiamate *B-spline lineari*.

Se i valori interpolati  $f_i$  sono i valori di una funzione  $f(x)$  nei punti  $x_i$  e tale funzione ha la derivata seconda continua, l'errore nell'approssimazione di  $f$  con le spline lineari ha la seguente rappresentazione

$$f(x) - s(x) = \frac{f''(\xi_i)}{2}(x - x_i)(x - x_{i+1}), \quad x_i \leq x \leq x_{i+1}$$

da cui, se  $M$  è una limitazione superiore di  $|f''(x)|$  su  $[x_1, x_n]$ , si ha

$$|f(x) - s(x)| \leq \frac{h^2}{8} M$$

ove  $h = \max_i |x_{i+1} - x_i|$ .

### Spline cubiche

Con le notazioni precedenti una *spline cubica* è una funzione con le proprietà

- $s(x)$ ,  $s'(x)$ ,  $s''(x)$  sono continue su  $[x_1, x_n]$ ;
- $s(x)$  è un polinomio di terzo grado su ogni  $[x_i, x_{i+1}]$ .

La funzione  $s(x)$  è composta da  $n - 1$  polinomi di terzo grado, e quindi ha  $4(n - 1)$  gradi di libertà. La continuità della funzione e delle derivate prima e seconda fornisce  $3(n - 2)$  condizioni; si hanno inoltre le  $n$  condizioni di interpolazione  $s(x_i) = f_i$ ,  $i = 1, 2, \dots, n$ . Pertanto, una *spline cubica interpolante* ha  $4(n - 1) - 3(n - 2) - n = 2$  gradi di libertà. Per determinare, quindi, in modo univoco la spline sono necessarie due ulteriori condizioni. Tali condizioni possono assumere le seguenti forme

$$\begin{array}{ll} s''(x_1) = 0, s''(x_n) = 0 & \text{spline naturale} \\ s'(x_1) = f'(x_1), s'(x_n) = f'(x_n) & \text{spline vincolata} \end{array}$$

Sotto una delle due condizioni precedenti si può dimostrare che esiste una ed una sola spline cubica interpolante. Di questo risultato forniremo una *dimostrazione costruttiva*.

### Algoritmo per la costruzione di una spline cubica

Su ogni sottointervallo  $[x_i, x_{i+1}]$  possiamo rappresentare una spline cubica interpolante nella seguente forma

$$s_i(x) = f_i + s'_i(x_i)(x - x_i) + \frac{s''_i(x_i)}{2}(x - x_i)^2 + \frac{s'''_i(x_i)}{3!}(x - x_i)^3$$

Si ha quindi

$$\begin{aligned} s'_i(x) &= s'_i(x_i) + s''_i(x_i)(x - x_i) + \frac{s'''_i(x_i)}{2}(x - x_i)^2 \\ s''_i(x) &= s''_i(x_i) + s'''_i(x_i)(x - x_i) \end{aligned}$$

Poniamo, ora, per brevità, per  $i = 1, 2, \dots, n - 1$

$$\begin{aligned} h_i &= x_{i+1} - x_i \\ z_i &= s''_i(x_i), z_n = s''_{n-1}(x_n) \end{aligned}$$

La continuità di  $s''$  nei punti interni implica che

$$s_i'''(x_i) = \frac{z_{i+1} - z_i}{h_i}, \quad i = 1, 2, \dots, n-1$$

Pertanto, si può scrivere  $s_i$  nella seguente forma

$$s_i(x) = f_i + s_i'(x_i)(x - x_i) + \frac{z_i}{2}(x - x_i)^2 + \frac{z_{i+1} - z_i}{6h_i}(x - x_i)^3$$

per  $i = 1, 2, \dots, n-1$  e la derivata  $s'(x_i)$  nella forma

$$s_i'(x) = s_i'(x_i) + z_i(x - x_i) + \frac{z_{i+1} - z_i}{2h_i}(x - x_i)^2$$

Essendo il vincolo  $s_i(x_i) = f_i$  soddisfatto direttamente dalla definizione di  $s_i(x)$ , da  $s_i(x_{i+1}) = f_{i+1}$  si ha

$$f_i + s_i'(x_i)h_i + \frac{z_i}{2}h_i^2 + \frac{z_{i+1} - z_i}{6h_i}h_i^3 = f_{i+1}$$

da cui

$$s_i'(x_i) = \frac{f_{i+1} - f_i}{h_i} - z_{i+1} \frac{h_i}{6} - z_i \frac{h_i}{3}$$

per  $i = 1, 2, \dots, n-1$ . Per determinare  $z_i$ , si può allora utilizzare la continuità di  $s'(x)$ , cioè  $s_i'(x_{i+1}) = s_{i+1}'(x_{i+1})$ . Con calcoli immediati, si ottengono in questo modo le seguenti relazioni

$$h_i z_i + 2(h_i + h_{i+1})z_{i+1} + h_{i+1}z_{i+2} = 6 \left( \frac{f_{i+2} - f_{i+1}}{h_{i+1}} - \frac{f_{i+1} - f_i}{h_i} \right)$$

per  $i = 1, 2, \dots, n-2$ . Per una spline naturale si ha per definizione  $z_1 = z_n = 0$ , mentre per una spline con condizioni ai limiti si hanno le seguenti due equazioni

$$\begin{aligned} -\frac{h_1}{3}z_1 - \frac{h_1}{6}z_2 &= \frac{f_1 - f_2}{h_1} + f'(x_1) \\ -\frac{h_{n-1}}{6}z_{n-1} - \frac{h_{n-1}}{3}z_n &= \frac{f_{n-1} - f_n}{h_{n-1}} + f'(x_n) \end{aligned}$$

In ambedue i casi il calcolo delle quantità  $z_i$  equivale alla risoluzione di un sistema lineare con matrice tridiagonale e a predominanza diagonale, per il quale il metodo di eliminazione di Gauss può essere applicato in una forma particolarmente semplice, che ora ricorderemo brevemente (cfr. Capitolo 2). Consideriamo, più in generale, il sistema

$$\mathbf{Ax} = \mathbf{f} \tag{4.10}$$

con  $\mathbf{f} = [f_1, f_2, \dots, f_n]^T$  e

$$\mathbf{A} = \begin{bmatrix} a_1 & c_1 & & & 0 \\ b_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & & & b_n & a_n \end{bmatrix}$$

Si ha la decomposizione  $\mathbf{A} = \mathbf{L}\mathbf{U}$  con

$$\mathbf{L} = \begin{bmatrix} 1 & & & & 0 \\ \beta_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & \beta_{n-1} & 1 & \\ 0 & & & \beta_n & 1 \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} \alpha_1 & c_1 & & & 0 \\ & \alpha_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & \alpha_{n-1} & c_{n-1} \\ 0 & & & & \alpha_n \end{bmatrix}$$

ove gli elementi  $\alpha_i, \beta_i$  sono determinati in forma iterativa nella forma seguente

$$\alpha_1 = a_1$$

$$\beta_k = \frac{b_k}{\alpha_{k-1}}, \quad \alpha_k = a_k - \beta_k c_{k-1}, \quad k = 2, 3, \dots, n$$

Il sistema (4.10) diventa, allora

$$\mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{f} \iff \begin{cases} \mathbf{L}\mathbf{y} = \mathbf{f} \\ \mathbf{U}\mathbf{x} = \mathbf{y} \end{cases}$$

da cui

$$y_1 = f_1, \quad y_i = f_i - \beta_i y_{i-1}, \quad i = 2, 3, \dots, n \quad (\text{forward})$$

$$x_n = \frac{y_n}{\alpha_n}, \quad x_i = \frac{y_i - c_i x_{i+1}}{\alpha_i}, \quad i = n-1, \dots, 2, 1 \quad (\text{backward})$$

Il costo dell'algoritmo è dato da  $3(n-1)$  moltiplicazioni e addizioni e  $2n-1$  divisioni.

Analogamente a quanto si è visto per le spline lineari, è possibile, e talvolta è conveniente, esprimere una generica spline cubica mediante una base locale, i cui elementi, cioè, siano diversi dallo zero in un intorno piccolo di ciascun nodo  $x_i$ . Come esemplificazione, riportiamo la definizione di un elemento di tale base (detta B-spline) nel caso in cui i nodi  $x_i$  siano equidistanti. Più precisamente, considerati i nodi  $-2, -1, 0, 1, 2$ , l'elemento  $\phi$  della base corrispondente al nodo 0 è definito da

$$\phi(x) = \frac{1}{4} \begin{cases} (x+2)^3 & x \in [-2, -1] \\ 1 + 3(x+1) + 3(x+1)^2 - 3(x+1)^3 & x \in [-1, 0] \\ 1 + 3(1-x) + 3(1-x)^2 - 3(1-x)^3 & x \in [0, 1] \\ (2-x)^3 & x \in [1, 2] \\ 0 & |x| > 2 \end{cases}$$

il cui grafico è riportato in Figura 4.8.

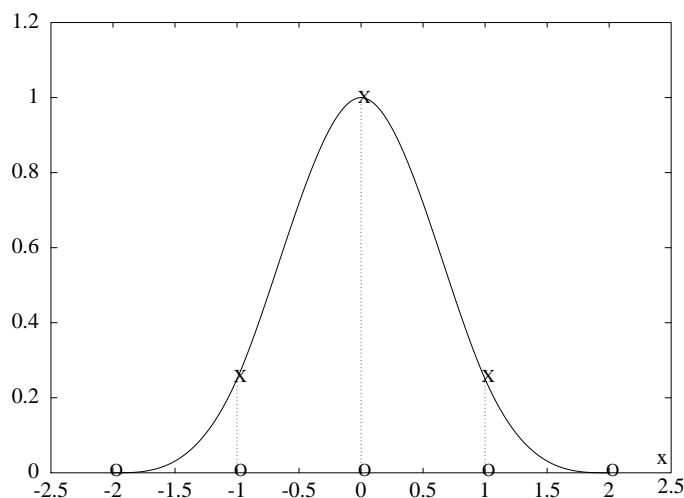


Figura 4.8: Rappresentazione grafica della B-spline cubica.

### Proprietà di approssimazione di una spline cubica

Nel caso in cui  $f_i$  siano i valori di una funzione  $f(x)$  sufficientemente regolare, è possibile dimostrare che al tendere a zero dell'ampiezza degli intervalli  $[x_i, x_{i+1}]$  la spline interpolante cubica converge alla funzione  $f(x)$ . Ad esempio, nel caso di una spline vincolata per una funzione  $f(x)$  derivabile fino al quarto ordine nell'intervallo  $[a, b]$ , si ha

$$\max_{a \leq x \leq b} \left| \frac{d^r s(x)}{dx^r} - \frac{d^r f(x)}{dx^r} \right| \leq K_r(\beta) M h^{4-r}, \quad r = 0, 1, 2, 3$$

ove  $\beta = \max_{i,j} \left( \frac{h_i}{h_j} \right)$ ,  $h = \max_i h_i$ ,  $|f^{(4)}(x)| \leq M$ ,  $a \leq x \leq b$  e

$$K_0(\beta) = \frac{5}{384}, \quad K_1(\beta) = \frac{1}{216}(9 + \sqrt{3})$$

$$K_2(\beta) = \frac{5}{384}(1 + 3\beta), \quad K_3(\beta) = \frac{1}{2}(1 + \beta^2)$$

In particolare, se i nodi rimangono equidistanti si ha  $\beta = 1$  e per nodi non equidistanti,  $\beta$  aumenta quando il rapporto tra il più grande e il più piccolo sottointervallo aumenta. Il risultato precedente mostra che la spline di interpolazione cubica converge uniformemente a  $f$  insieme alle derivate prima, seconda e terza.

Terminiamo ricordando la proprietà delle spline cubiche che ha dato il nome a questo tipo di approssimazione. Mediante una integrazione per parti si dimostra facilmente che se  $s(x)$  è una spline cubica naturale che interpola una funzione  $f(x)$

nei nodi  $a = x_1 < x_2 < \dots < x_n = b$ , si ha

$$\int_a^b (g''(x))^2 dx = \int_a^b (s''(x))^2 dx + \int_a^b (g''(x) - s''(x))^2 dx$$

per ogni funzione  $g(x)$  dotata di derivata prima e seconda continue su  $[a, b]$ , e che interpola  $f(x)$  nei nodi. Da tale uguaglianza si ricava il seguente risultato.

**Teorema 4.4 (Proprietà di minimo)** *Tra tutte le funzioni  $g(x)$  che sono due volte continuamente derivabili su  $[a, b]$ , e che interpolano  $f(x)$  nei punti  $a = x_1 < x_2 < \dots < x_n = b$ , la spline cubica naturale interpolante minimizza l'integrale*

$$\int_a^b (g''(x))^2 dx \quad (4.11)$$

Questo significa che le spline cubiche interpolanti sono tra le funzioni con derivata seconda continua che interpolano la funzione  $f(x)$  nei nodi  $x_i$  quelle che hanno minima curvatura, ossia quelle che oscillano meno. Osserviamo, inoltre, che l'integrale (4.11) è una approssimazione dell'integrale dell'energia potenziale elastica corrispondente alla configurazione  $g(x)$  di un nastro perfettamente elastico (spline) vincolato a passare attraverso i nodi  $(x_i, f_i)$ .

#### 4.1.6 Approssimazione di Bézier

Con il nome di *approssimazione di Bézier* si denota un insieme di tecniche particolarmente utili nel campo del disegno mediante calcolatore (CAD: computer aided design; CAM: computer aided manufacturing)<sup>6</sup>. Il motivo di tale interesse è legato alle *possibilità interattive* offerte da tale approssimazione. In questo paragrafo presenteremo una introduzione al metodo, rinviando ad esempio a Farin [55] per un opportuno approfondimento. Richiamiamo anche lo stretto legame esistente tra le curve di Bézier e le B-spline, ossia le basi delle spline a supporto locale (cfr. nel paragrafo precedente gli esempi delle spline di primo e terzo ordine); in effetti, le curve di Bézier costituiscono un caso speciale delle curve B-spline.

Le curve di Bézier possono essere generate da un algoritmo ricorsivo; è comunque utile, in particolare nell'ambito degli sviluppi teorici, avere di esse una rappresentazione *esplicita*, ossia un'espressione in termini di una formula non ricorsiva, piuttosto che in termini di un algoritmo. Tale rappresentazione può essere ottenuta mediante i *polinomi di Bernstein*, di cui ora ricorderemo la definizione e discuteremo alcune interessanti proprietà.

<sup>6</sup>Il metodo è stato sviluppato indipendentemente da P. Bézier alla Renault (*Definition numérique des courbes et surfaces* I, 1966, II, 1967) e da P. de Casteljau alla Citroën (rapporti tecnici, 1959, 1963, non pubblicati). Tali lavori, basati sull'analisi dell'uso di curve e superfici parametriche, possono essere visti come l'origine della nuova disciplina *Computer Aided Geometric Design* (CAGD).

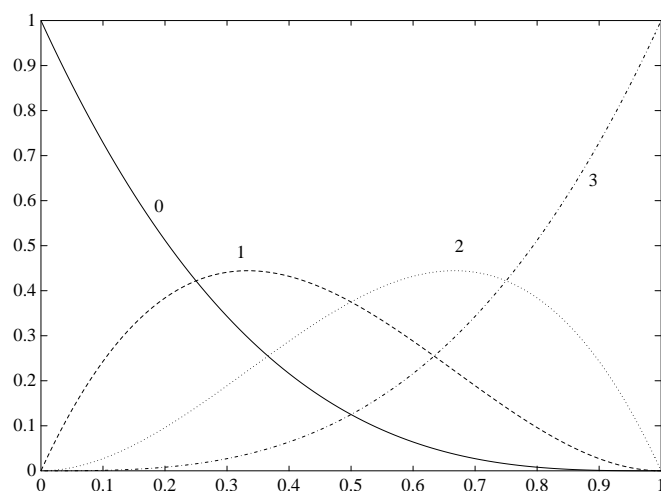


Figura 4.9: Rappresentazione degli elementi della base di Bernstein per  $n = 3$ .

**Polinomi di Bernstein** Chiamato  $[a, b]$  l'intervallo che contiene i punti  $x_i$ ,  $i = 0, 1, \dots, n$ , con  $x_0 = a$  e  $x_n = b$ , consideriamo i seguenti polinomi di grado  $n$

$$B_i^n(x) := \binom{n}{i} \frac{(b-x)^{n-i} (x-a)^i}{(b-a)^n}, \quad i = 0, \dots, n$$

ove  $\binom{n}{i}$  indica il coefficiente binomiale

$$\binom{n}{i} \equiv \frac{n!}{i!(n-i)!}$$

Tali polinomi sono linearmente indipendenti, ossia da  $\sum_{i=0}^n \alpha_i B_i^n(x) = 0$  su  $(a, b)$  segue  $\alpha_i = 0, i = 0, 1, \dots, n$ . Essi costituiscono le *funzioni base di Bernstein*. Ad esempio, per  $n = 3$  si ha (cfr. Figura 4.9)

$$B_0^3(x) = \frac{(b-x)^3}{(b-a)^3}, \quad B_1^3(x) = 3 \frac{(b-x)^2(x-a)}{(b-a)^3}, \quad B_2^3(x) = 3 \frac{(b-x)(x-a)^2}{(b-a)^3}, \quad B_3^3(x) = \frac{(x-a)^3}{(b-a)^3}$$

Per il calcolo numerico del valore di un elemento della base di Bernstein in un punto assegnato  $x$ , anziché la definizione, è più opportuno il seguente algoritmo numericamente più stabile

$$\begin{aligned} B_0^n(x) &= 1, & B_{n-1}^n(x) &= 0, \\ B_i^n(x) &= \frac{(b-x) B_i^{n-1}(x) + (x-a) B_{i-1}^{n-1}(x)}{(b-a)}, & i &= 0, 1, \dots, n \\ B_{n+1}^n(x) &= 0 \end{aligned}$$



L'algoritmo è implementato, per l'intervallo di riferimento  $(0, 1)$ , nella seguente procedura, che fornisce nel vettore  $B$  i valori degli elementi della base di Bernstein di grado  $n$  nel punto  $x$ .

```

SUBROUTINE BE(N,B,X)
REAL B(0:N),X
IF (N.EQ.0) THEN
  B(0)=1.0
ELSE IF (N.GT.0) THEN
  DO 2 J=1,N
    IF (J.EQ.1) THEN
      B(1)=X
    ELSE
      B(J)=X*B(J-1)
    END IF
    DO 1 M=J-1,1,-1
      B(M)=X*B(M-1)+(1.0-X)*B(M)
1    CONTINUE
    IF (J.EQ.1) THEN
      B(0)=1.0-X
    ELSE
      B(0)=(1.0-X)*B(0)
    END IF
2    CONTINUE
  END IF
  RETURN
END

```

Dal fatto che i polinomi  $B_i^n(x)$ , per  $i = 0, 1, \dots, n$  costituiscono una base per i polinomi di grado  $n$ , ossia un insieme di  $n + 1$  polinomi di grado  $n$  linearmente indipendenti, si ha che un generico polinomio  $P_n(x)$  di grado  $n$  può essere espresso in maniera univoca come combinazione lineare dei polinomi  $B_i^n(x)$ , ossia

$$P_n(x) = \sum_{i=0}^n p_i B_i^n(x) \quad (4.12)$$

ove  $p_i$  rappresentano le coordinate del polinomio  $P_n(x)$  rispetto alla base di Bernstein. Tali coordinate possono allora essere determinate imponendo, ad esempio, che  $P_n(x)$  verifichi le condizioni di interpolazione  $P_n(x_i) = y_i$ ,  $i = 0, \dots, n$ , ove  $y_i$  sono dei valori assegnati. Tali condizioni equivalgono alla risoluzione di un sistema lineare nelle  $n + 1$  incognite  $p_i$ .

Esaminiamo ora la relazione tra le quantità  $p_i$  e  $y_i$ . Mentre nella rappresentazione di Lagrange (4.2) si ha  $p_i = y_i$  per ogni  $i$ , nella rappresentazione (4.12) si ha  $p_0 = y_0$  e  $p_n = y_n$ , ma per  $i \neq 0, n$  i valori sono differenti. Tuttavia, l'aspetto interessante della rappresentazione (4.12) è il fatto che i coefficienti  $p_i$  sono utilizzabili in *maniera semplice* per variare la forma della curva. In altre parole, a partire dai valori  $p_i$  è possibile prevedere in maniera intuitiva il comportamento del polinomio  $P_n(x)$ , senza

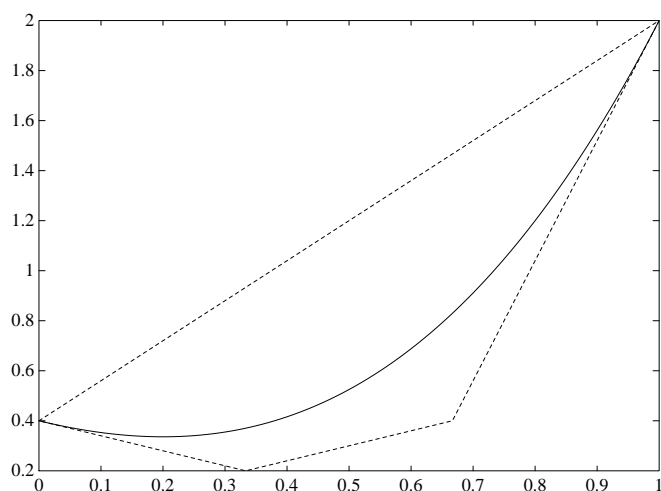


Figura 4.10: Poligono e curva di Bézier:  $n = 3$  e  $p_0 = 0.4, p_1 = 0.2, p_2 = 0.4, p_3 = 2$ .

necessariamente darne una rappresentazione grafica. Questo fatto è importante nel CAD, quando l'obiettivo finale, anziché una interpolazione di punti, è la realizzazione di una curva di forma prefissata.

L'idea di base consiste nello specificare  $p_i$  in maniera grafica. A tale scopo si associa ogni  $p_i$  ad un punto nel piano  $(t_i, p_i)$ , chiamato *punto di controllo*. I punti  $t_i$  sono assunti *equidistanti*

$$t_i = x_0 + i \frac{x_n - x_0}{n} = a + i \frac{b - a}{n}$$

e pertanto salvo per  $t_0$  e  $t_n$  si ha, in generale,  $t_i \neq x_i$ . Sulla base dei nodi  $(t_i, p_i)$  si costruisce il *poligono* di Bézier, partendo da  $(t_0, p_0)$  e congiungendo i successivi punti di controllo con segmenti di retta; infine, si congiunge  $(t_n, p_n)$  con  $(t_0, p_0)$ . Tale poligono fornisce un'idea della forma di  $P_n(x)$ . In particolare, si ha che se i valori  $p_i$  sono monotoni, pure il polinomio  $P_n(x)$  risulta una funzione monotona. Inoltre  $P_n(x)$  risulta completamente contenuto nel più piccolo insieme convesso (il cosiddetto *inviluppo convesso*) che contiene tutti i punti di controllo. In questo modo, muovendo opportunamente i punti di controllo si ha un'idea *diretta* e *intuitiva* della funzione  $P_n(x)$ .

Come esemplificazione, in Figura 4.10 sono rappresentati il poligono e la curva di Bézier per  $n = 3$  relativi ai valori  $p_0 = 0.4, p_1 = 0.2, p_2 = 0.4, p_3 = 2$ . Nella successiva Figura 4.11 è mostrato l'effetto dello spostamento del nodo  $(t_2, p_2)$ .

Nella forma illustrata in precedenza i punti di controllo  $(t_i, p_i)$  non possono essere spostati orizzontalmente. Si può, in effetti, superare tale inconveniente considerando la seguente generalizzazione. Siano  $\mathbf{p}_i$  dei punti assegnati nel piano e definiamo la

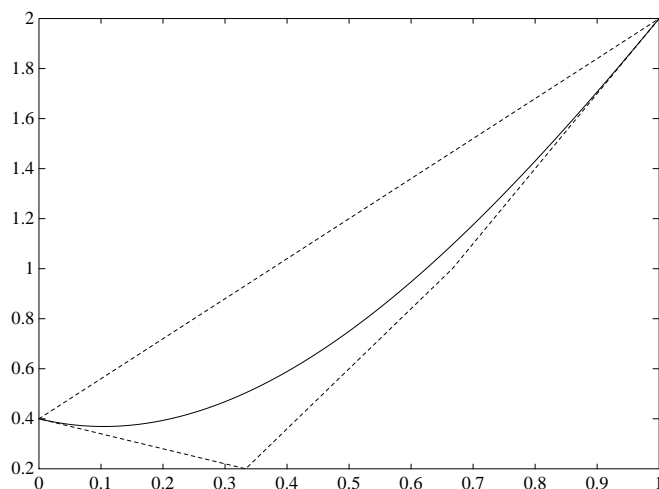


Figura 4.11: Poligono e curva di Bézier:  $n = 3$  e  $p_0 = 0.4, p_1 = 0.2, p_2 = 1., p_3 = 2.$

curva di Bézier vettoriale ponendo

$$\mathbf{P}_n(x) = \sum_{i=0}^n \mathbf{p}_i B_i^n(x)$$

Al variare del parametro  $x$  tra  $a$  e  $b$ ,  $\mathbf{P}_n(x)$  descrive una traiettoria nel piano con punto iniziale  $\mathbf{p}_0$  e punto finale  $\mathbf{p}_n$ .

In particolare, per  $n = 3$  si ottiene la curva di Bézier cubica, che ha la seguente rappresentazione parametrica

$$\begin{aligned} x(u) &= (1-u)^3 x_0 + 3(1-u)^2 u x_1 + 3(1-u) u^2 x_2 + u^3 x_3 \\ y(u) &= (1-u)^3 y_0 + 3(1-u)^2 u y_1 + 3(1-u) u^2 y_2 + u^3 y_3 \end{aligned}$$

ove, per evitare equivoci, si è indicato con  $u$  il parametro che varia nell'intervallo  $[0, 1]$  e con  $(x_i, y_i)$  le coordinate dei punti  $\mathbf{p}_i$ . Come esemplificazione si veda Figura 4.12.

Osserviamo che si può facilmente prolungare la curva cubica di Bézier mantenendo la continuità della derivata prima. È sufficiente considerare i tre punti  $\mathbf{p}_4, \mathbf{p}_5, \mathbf{p}_6$  sulla stessa linea (cfr. Figura 4.13).

◆ **Esercizio 4.1** A partire dai seguenti dati

$x_i$	0	1	2	3	6
$y_i$	1	9	22	91	245

costruire la corrispondente tabella delle differenze finite e mediante il polinomio di interpolazione nella forma di Newton approssimare il valore corrispondente a  $x = 3.8$ .

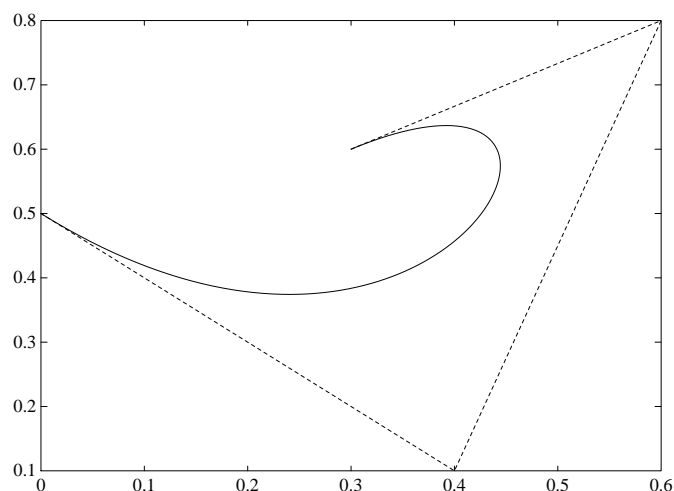


Figura 4.12: Poligono e curva vettore di Bézier:  $n = 3$  e  $\mathbf{p}_0 = (0., 0.5)$ ,  $\mathbf{p}_1 = (0.4, 0.1)$ ,  $\mathbf{p}_2 = (0.6, 0.8)$  e  $\mathbf{p}_3 = (0.3, 0.6)$ .

◆ **Esercizio 4.2** Verificare che i seguenti due polinomi

$$(a) P(x) = 5x^3 - 27x^2 + 45x - 21; \quad (b) Q(x) = x^4 - 5x^3 + 8x^2 - 5x + 3$$

interpolano i seguenti dati

$x_i$	1	2	3	4
$y_i$	2	1	6	47

Interpretare tale esempio alla luce del teorema di unicità del polinomio di interpolazione.

◆ **Esercizio 4.3** Mostrare che il valore della differenza divisa  $f[x_1, x_2, \dots, x_{k+1}]$  è indipendente dall'ordine dei punti  $x_1, x_2, \dots, x_{k+1}$ .

◆ **Esercizio 4.4** Il calore specifico dell'acqua come funzione della temperatura è fornito dalla seguente tabella

temperatura °C	calore specifico
20	0.99907
25	0.99852
30	0.99826
35	0.99818
40	0.99828
45	0.99849
50	0.99878

Approssimare mediante polinomi di interpolazione di differente grado il calore specifico a 37°C.

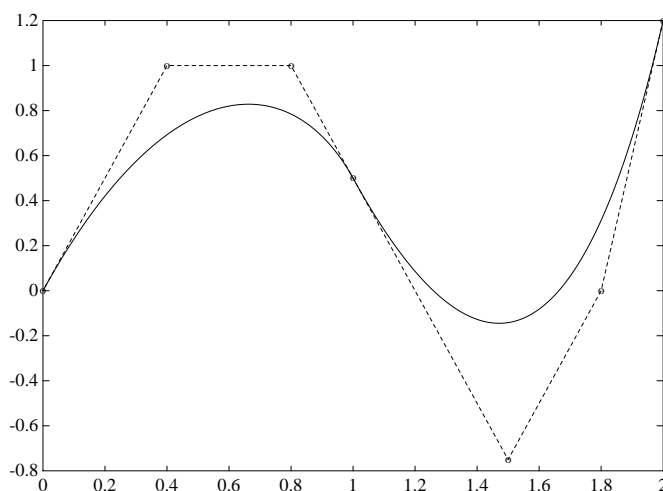


Figura 4.13: Prolungamento di una curva di Bézier cubica, mantenendo la continuità della derivata prima.

◆ **Esercizio 4.5** Mostrare che la spline lineare  $s$  corrispondente ai nodi  $x_1, x_2, \dots, x_n$  è la funzione che minimizza

$$\int_{x_1}^{x_n} (g'(x))^2 dx$$

tra tutte le funzioni continue  $g$  tali che  $g(x_i) = f_i$ , con  $f_i$  assegnati per  $i = 1, 2, \dots, n$ , e per le quali esiste l'integrale tra  $x_1$  e  $x_n$  del quadrato della derivata.

◆ **Esercizio 4.6** Trovare la spline cubica naturale che interpola i seguenti dati

$x_i$	1	2	3	4	5
$y_i$	0	1	0	1	0

◆ **Esercizio 4.7** Dati  $(m+1)(n+1)$  valori  $f_{i,j}$ , per  $i = 1, 2, \dots, m+1$  e  $j = 1, 2, \dots, n+1$ , posto

$$X_{m,i}(x) := \prod_{k=1, k \neq i}^{m+1} \frac{x - x_k}{x_i - x_k}, \quad i = 1, 2, \dots, m+1$$

$$Y_{n,i}(y) := \prod_{k=1, k \neq i}^{n+1} \frac{y - y_k}{y_i - y_k}, \quad i = 1, 2, \dots, n+1$$

mostrare che

$$p_{m,n}(x, y) := \sum_{i=1}^{m+1} \sum_{j=1}^{n+1} X_{m,i}(x) Y_{n,j}(y) f_{i,j}$$

è un polinomio di grado  $m$  in  $x$  e di grado  $n$  in  $y$  della forma  $\sum_{i=1}^{m+1} \sum_{j=1}^{n+1} a_{i,j} x^i y^j$ , che soddisfa alle seguenti condizioni di interpolazione

$$p_{m,n}(x_i, y_j) = f_{i,j}, \quad i = 1, 2, \dots, m+1, \quad j = 1, 2, \dots, n+1$$

## 4.2 Problema generale di approssimazione

Nell'approssimazione mediante interpolazione si cerca un polinomio, più in generale un elemento di una famiglia appropriata di funzioni (ad esempio, funzioni razionali fratte, funzioni trigonometriche, esponenziali ecc.), che *coincide* con la funzione da approssimare in punti prefissati.

Nelle applicazioni, questo tipo di approssimazione non è sempre possibile o conveniente. Supponiamo, ad esempio, che la funzione  $f(x)$  esprima una relazione tra quantità fisiche, o chimiche. I valori  $f_i$  sono allora determinati mediante misurazioni, e in generale si ha  $f_i = f(x_i) + \epsilon_i$ , ove gli errori  $\epsilon_i$  sono incogniti; pure i punti di osservazione  $x_i$  possono essere affetti da errore. A meno che gli errori siano piccoli e le misurazioni siano in numero basso, non è allora ragionevole descrivere  $f(x)$  mediante un polinomio che passi esattamente attraverso tali punti. Può essere, invece, più opportuna una approssimazione nella quale l'influenza degli errori di misurazione sia minimizzata. In effetti, questo risultato può essere ottenuto utilizzando il *metodo dei minimi quadrati*, che esamineremo nel seguito.

Una differente situazione, ma per la quale è ancora opportuna una approssimazione di tipo diverso dalla interpolazione, si presenta quando una funzione assegnata in forma analitica, ad esempio  $e^x$ ,  $\log x$ ,  $\sin x$ , ecc., è approssimata da un polinomio (o da una funzione razionale fratta) allo scopo di costruire una procedura di calcolo della funzione da utilizzare in un calcolatore. In questo caso il tipo di approssimazione più conveniente corrisponde a cercare il *polinomio di minimo grado* (e quindi con il minimo costo di valutazione) che approssima la funzione assegnata con un errore minore di una tolleranza assegnata su tutto un intervallo prefissato.

Le due situazioni precedenti sono casi particolari della seguente situazione generale. Supponiamo che la funzione  $f(x)$ , i cui valori possono essere assegnati su tutto un intervallo (caso *continuo*) o solo in corrispondenza ad un numero finito di punti  $x_0, x_1, \dots, x_m$  (caso *discreto*), appartenga ad uno *spazio lineare*  $V$ .

**Spazi lineari normati** Ricordiamo che uno spazio  $V$  è detto *lineare*, se per ogni coppia di elementi  $f$  e  $g$  in  $V$ , e un numero arbitrario reale  $\alpha$ ,  $\alpha f$  e  $f+g$  appartengono a  $V$ . Inoltre, un sottoinsieme non vuoto  $U$  di uno spazio lineare  $V$  è un *sottospazio* di  $V$  se per ogni coppia arbitraria  $f_1, f_2$  di elementi in  $U$ , e un numero arbitrario reale  $\alpha$ ,  $\alpha f_1$  e  $f_1 + f_2$  appartengono a  $U$ .

Vi sono diversi tipi di spazi lineari che hanno interesse nelle applicazioni; per il seguito, tuttavia, ci limiteremo a considerare lo spazio lineare  $V = C^0([a, b])$  delle funzioni a valori reali e continue sull'intervallo  $[a, b]$  per il caso continuo, e lo spazio dei vettori  $\mathbb{R}^n$  nel caso discreto.

Per precisare il senso dell'approssimazione occorre introdurre nello spazio  $V$  una *distanza*. Essa può essere definita a partire dalla nozione di *norma*.

**Definizione 4.1 (norma)** Dato uno spazio lineare  $V$ , si dice *norma* una trasformazione  $V \rightarrow \mathbb{R}$ , indicata usualmente con  $\|\cdot\|$ , con le seguenti proprietà, ove  $f$  e  $g$  sono elementi arbitrari in  $V$

$$\begin{aligned}\|f\| &\geq 0; \quad \|f\| = 0 \iff f = 0 \\ \|\alpha f\| &= |\alpha| \cdot \|f\| \quad \forall \alpha \in \mathbb{R} \\ \|f + g\| &\leq \|f\| + \|g\| \quad (\text{disuguaglianza triangolare})\end{aligned}$$

Si definisce allora *distanza* tra due elementi  $f$  e  $g$  in  $V$  la quantità  $\|f - g\|$ .

Nello spazio  $C^0([a, b])$  si possono, ad esempio, definire le seguenti norme

$$\begin{aligned}\|f\|_2 &= \left( \int_a^b |f(x)|^2 dx \right)^{1/2} && \text{norma 2 (euclidea)} \\ \|f\|_\infty &= \max_{a \leq x \leq b} |f(x)| && \text{norma del massimo (di Chebichev)}\end{aligned}$$

a cui corrispondono altrettante distanze e tipi diversi di approssimazione.

Nel caso discreto, cioè quando sono assegnati i valori  $f(x_i)$  in un insieme di punti  $S = \{x_i\}_{i=0}^m$ , si hanno definizioni analoghe (corrispondenti alla definizione di norma introdotta per  $\mathbb{R}^n$  nell'Appendice A) Ad esempio, si ha

$$\begin{aligned}\|f\|_{2,S} &= \left( \sum_{i=0}^m |f(x_i)|^2 \right)^{1/2} \\ \|f\|_{\infty,S} &= \max_{x_i \in S} |f(x_i)|\end{aligned}$$

Osserviamo che nel caso in cui  $f(x)$  sia definita su tutto un intervallo  $[a, b]$  contenente l'insieme di punti  $S$ , le quantità ora definite possono essere nulle senza che  $f(x)$  sia identicamente nulla su tutto  $[a, b]$ ; per tale motivo esse vengono dette *seminorme*, in quanto non verificano tutte le condizioni della Definizione 4.1.

Nelle applicazioni è talvolta opportuno generalizzare le definizioni precedenti introducendo una *funzione peso*  $w(x) > 0$ , con la quale assegnare una differente importanza ai valori  $f(x)$  nel calcolo della distanza. Si ha allora, ad esempio

$$\begin{aligned}\|f\|_{2,w} &= \left( \int_a^b w(x) |f(x)|^2 dx \right)^{1/2} \\ \|f\|_{2,S,w} &= \left( \sum_{i=0}^m w(x_i) |f(x_i)|^2 dx \right)^{1/2}\end{aligned}$$

con analoga definizione nel caso della norma del massimo. Naturalmente, la funzione  $w(x)$  deve essere tale da assicurare nel caso continuo l'esistenza dell'integrale per ogni  $f(x)$  nello spazio  $V$ .

Terminiamo questi brevi richiami sugli spazi lineari, ricordando la nozione importante di *prodotto scalare* e la conseguente nozione di *sistema ortogonale*, generalizzando le nozioni introdotte nel caso di  $V = \mathbb{R}^n$  in Appendice A. Per una generica coppia di funzioni  $f(x), g(x)$  nello spazio  $V = C^0([a, b])$  e per una opportuna funzione peso  $w(x)$  si può definire un *prodotto scalare* ponendo

$$(f, g) := \begin{cases} \int_a^b w(x)f(x)g(x) dx & \text{(caso continuo)} \\ \sum_{i=0}^m w(x_i)f(x_i)g(x_i) & \text{(caso discreto)} \end{cases}$$

da cui la seguente definizione di ortogonalità.

**Definizione 4.2 (sistema ortogonale)** *Le funzioni  $f(x), g(x) \in V$  si dicono ortogonali se  $(f, g) = 0$ . Un insieme di funzioni  $\phi_0, \phi_1, \dots, \phi_n$  appartenenti a  $V$  è chiamato un sistema ortogonale se  $(\phi_i, \phi_j) = 0$  per  $i \neq j$ , e  $(\phi_i, \phi_i) \neq 0$  per ogni  $i = 0, \dots, n$ . Quando  $(\phi_i, \phi_i) = 1$ , per ogni  $i$ , il sistema è detto ortonormale.*

Nel seguito (cfr. Paragrafo 4.2.2) considereremo differenti esempi di sistemi ortogonali nell'ambito dei polinomi. Osserviamo che per la norma, o seminorma, euclidea si ha  $\|f\|_2 = \sqrt{(f, f)}$ .

Si può allora dimostrare facilmente la seguente generalizzazione della uguaglianza di Pitagora. Se  $\{\phi_0, \phi_1, \dots, \phi_n\}$  è un sistema ortogonale, si ha

$$\left\| \sum_{j=0}^n c_j \phi_j \right\|_2^2 = \sum_{j=0}^n c_j^2 \|\phi_j\|_2^2 \quad (4.13)$$

Da tale uguaglianza si ricava, in particolare, che se  $\{\phi_0, \phi_1, \dots, \phi_n\}$  è un sistema ortogonale, si ha

$$\left\| \sum_{j=0}^n c_j \phi_j \right\| = 0 \quad \text{se e solo se } c_j = 0, j = 0, 1, \dots, n$$

ossia le funzioni  $\phi_0, \phi_1, \dots, \phi_n$  sono *linearmente indipendenti*.

**Approssimazione lineare** Ritornando alla definizione del problema di approssimazione, siano  $\phi_0, \phi_1, \dots, \phi_n$   $n + 1$  funzioni assegnate nello spazio  $V = C^0([a, b])$  e  $U$  il sottospazio di  $V$  costituito dalle combinazioni lineari

$$g_n(x) = c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_n\phi_n(x) \quad (4.14)$$

al variare delle costanti  $c_0, c_1, \dots, c_n$  in  $\mathbb{R}$ . Se, ad esempio, si sceglie  $\phi_i(x) = x^i$ ,  $i = 0, 1, \dots, n$ , allora  $g_n(x)$  è un polinomio di grado al più  $n$  e l'insieme  $U$  è lo spazio



dei polinomi di grado minore o uguale a  $n$ . Pur rimanendo nell'ambito dei polinomi, vi possono essere, tuttavia, per  $\phi(x)$  scelte più convenienti per le applicazioni (in particolare, come vedremo nel seguito i sistemi di polinomi ortogonali).

Fissata allora in  $V$  una particolare norma  $\|\cdot\|$ , il problema dell'approssimazione di una funzione  $f(x) \in V$  mediante funzioni del sottospazio  $U$  può essere formulato come un *problema di minimo*, corrispondente alla ricerca della combinazione  $g_n^*(x) \in U$ , tale che

$$\|f - g_n^*\| \leq \|f - g_n\| \quad \forall g_n \in U \quad (4.15)$$

Più precisamente, tale problema è chiamato un *problema di approssimazione lineare*, in quanto le funzioni approssimanti  $g_n(x)$  dipendono linearmente dai parametri incogniti  $c_0, c_1, \dots, c_n$ . In questo capitolo tratteremo in particolare i problemi di approssimazione di tipo lineare. I problemi di approssimazione di tipo non lineare, importanti per la modellistica matematica, saranno considerati sotto vari aspetti nei successivi Capitoli 5, 12 e 13.

Nel seguito del capitolo esamineremo più in dettaglio l'approssimazione lineare corrispondente rispettivamente alla norma *euclidea* e alla norma del *massimo*.

#### 4.2.1 Norma euclidea. Minimi quadrati

Si cerca  $g_n^*$  della forma (4.14) che verifica per tutte le funzioni  $g_n(x)$  dello stesso tipo le seguenti disuguaglianze

$$\text{(caso continuo)} \quad \|f - g_n^*\|_{2,w}^2 \leq \|f - g_n\|_{2,w}^2, \quad \text{ove} \quad \|f\|_{2,w}^2 = \int_a^b w(x) |f(x)|^2 dx$$

$$\text{(caso discreto)} \quad \|f - g_n^*\|_{2,S,w}^2 \leq \|f - g_n\|_{2,S,w}^2, \quad \text{ove} \quad \|f\|_{2,S,w}^2 = \sum_{i=0}^m w(x_i) |f(x_i)|^2$$

La funzione  $g_n^*$  viene detta *elemento di migliore approssimazione nel senso dei minimi quadrati*; dal punto di vista geometrico, la funzione  $g_n^*$  rappresenta la proiezione ortogonale di  $f$  sul sottospazio  $U$  (cfr. Figura 4.14).

Dal punto di vista teorico, ossia esistenza ed unicità dell'elemento di migliore approssimazione, si ha il seguente risultato, valido sia nel caso discreto che continuo.

**Teorema 4.5 (minimi quadrati)** *Supponiamo che le funzioni  $\phi_0, \phi_1, \dots, \phi_n$  siano linearmente indipendenti. Allora, esiste una ed una sola funzione*

$$g_n^*(x) = \sum_{j=0}^n c_j^* \phi_j(x) \quad (4.16)$$

tale che

$$\|f(x) - g_n^*(x)\|_2 \leq \|f(x) - g_n(x)\|_2 \quad \text{per ogni } g_n(x) = \sum_{j=0}^n c_j \phi_j(x)$$

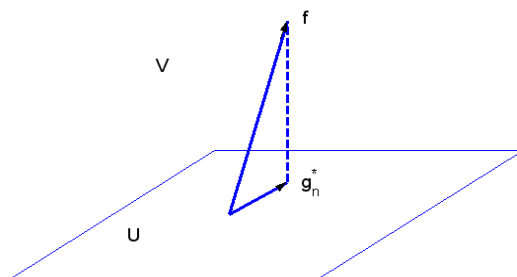


Figura 4.14: Interpretazione geometrica del problema dei minimi quadrati.

ove  $\|\cdot\|_2$  indica una delle due norme  $\|\cdot\|_{2,w}$ ,  $\|\cdot\|_{2,S,w}$ . Inoltre, la funzione  $g_n^*(x)$  è la soluzione del seguente sistema lineare (detto sistema delle equazioni normali)

$$(f(x) - g_n^*(x), \phi_k(x)) = 0, \quad k = 0, 1, 2, \dots, n \quad (4.17)$$

ove  $(\cdot, \cdot)$  indica il prodotto scalare in  $V$  corrispondente alla norma considerata.

**DIMOSTRAZIONE.** Consideriamo come illustrazione grafica del teorema la Figura 4.14. Il sottospazio  $U$  di  $V$  è generato dalle combinazioni lineari di  $\phi_0, \phi_1, \dots, \phi_n$ . Il teorema afferma che  $f - g_n^*$  è ortogonale alle funzioni  $\phi_i$ , e quindi a tutti gli elementi del sottospazio  $U$ . Si ha pertanto che  $g_n^*$  è la proiezione ortogonale di  $f$  su  $U$  e, in sostanza, il teorema afferma che la proiezione ortogonale  $g_n^*$  di  $f$  su  $U$  è l'elemento in  $U$  che ha la minima distanza euclidea da  $f$ . Osserviamo che le equazioni normali (4.17) possono essere scritte nel seguente modo

$$(g_n^*, \phi_k) = (f, \phi_k), \quad k = 0, 1, \dots, n$$

da cui, tenendo conto della rappresentazione (4.16)

$$\sum_{j=0}^n c_j^* (\phi_j, \phi_k) = (f, \phi_k), \quad k = 0, 1, \dots, n \quad (4.18)$$

Si ha, pertanto, che i coefficienti  $c_j^*$  sono soluzioni di un sistema lineare con matrice dei coefficienti  $[(\phi_j, \phi_k)]$ ,  $j, k = 0, 1, \dots, n$  e termine noto  $[(f, \phi_k)]$ ,  $k = 0, 1, \dots, n$ . Dimostriamo che la matrice dei coefficienti è non singolare e che quindi il sistema lineare (4.18) ammette una ed una sola soluzione. Ragionando per assurdo, se la matrice fosse singolare, il sistema omogeneo avrebbe una soluzione  $c_0, c_1, \dots, c_n$  non identicamente nulla, cioè tale che

$$\sum_{j=0}^n c_j (\phi_j, \phi_k) = 0, \quad k = 0, 1, \dots, n$$

Ma, allora, si avrebbe

$$\left\| \sum_{j=0}^n c_j \phi_j \right\|_2^2 = \sum_{k=0}^n c_k \left( \sum_{j=0}^n c_j (\phi_j, \phi_k) \right) = 0$$

e, contrariamente all'ipotesi, le funzioni  $\phi_j$  sarebbero linearmente dipendenti.

Dimostriamo ora che ogni funzione  $g_n = \sum_{j=0}^n c_j \phi_j$ , con  $c_j \neq c_j^*$  per almeno un indice  $j$ , ha una distanza da  $f$  maggiore che  $g_n^*$ . In effetti, dalla seguente identità

$$f - \sum_{j=0}^n c_j \phi_j = (f - g_n^*) + \sum_{j=0}^n (c_j^* - c_j) \phi_j$$

tenendo conto (cfr. (4.18)) che  $(f - g_n^*, \phi_j) = 0$  per  $j = 0, 1, \dots, n$ , si ha

$$(f - g_n^*, \sum_{j=0}^n (c_j^* - c_j) \phi_j) = 0$$

ossia gli elementi  $f - g_n^*$  e  $\sum_{j=0}^n (c_j^* - c_j) \phi_j$  sono ortogonali. Applicando allora l'uguaglianza di Pitagora (4.13), si ha

$$\left\| f - \sum_{j=0}^n c_j \phi_j \right\|_2^2 = \|f - g_n^*\|_2^2 + \left\| \sum_{j=0}^n (c_j^* - c_j) \phi_j \right\|_2^2$$

da cui

$$\left\| f - \sum_{j=0}^n c_j \phi_j \right\|_2^2 \geq \|f - g_n^*\|_2^2$$

L'uguaglianza si ottiene soltanto per  $c_j^* = c_j$ , dal momento che gli elementi  $\phi_j$  sono linearmente indipendenti. ■

► **Esempio 4.6** (*Caso continuo*) Determinare il polinomio  $g_2^*(x) = c_0 + c_1 x + c_2 x^2$  di migliore approssimazione nel senso dei minimi quadrati, con  $w(x) \equiv 1$ , della funzione  $f(x) = e^x$  su tutto l'intervallo  $[0, 1]$ .

In questo caso abbiamo  $\phi_0 = 1$ ,  $\phi_1 = x$ ,  $\phi_2 = x^2$ , e il sistema lineare delle equazioni normali (4.18) diventa il seguente

$$c_0 + \frac{1}{2}c_1 + \frac{1}{3}c_2 = e - 1$$

$$\frac{1}{2}c_0 + \frac{1}{3}c_1 + \frac{1}{4}c_2 = 1$$

$$\frac{1}{3}c_0 + \frac{1}{4}c_1 + \frac{1}{5}c_2 = e - 2$$

Risolvendo il sistema in doppia precisione ( $\approx 16$  cifre decimali), si ottiene come soluzione il vettore

$$c^* = [1.01299130990276, 0.85112505284626, 0.83918397639947]$$

In Figura 4.15 è rappresentato l'errore  $f(x) - g_2^*(x)$ . Osserviamo che se arrotondiamo il termine noto del sistema delle equazioni normali a sei cifre, cioè poniamo  $[e - 1, 1, e - 2] \approx$

[1.71828, 1, 0.71828] e risolviamo il corrispondente sistema ancora in doppia precisione, si ottiene come soluzione il vettore

$$\tilde{c} = [1.01292000000000, 0.85152000000000, 0.83880000000000]$$

Si vede, quindi, che gli errori relativi sui dati si amplificano sui risultati, indicando la presenza di malcondizionamento nella matrice del sistema. In effetti, tale matrice è la matrice di Hilbert che abbiamo già considerato dal punto di vista del condizionamento nel Capitolo 2. Una osservazione importante, comunque, è che il malcondizionamento del sistema delle equazioni normali è legato alla scelta particolare della base  $\phi_j$ . Come vedremo nel seguito, rispetto ad altre basi il sistema può essere bencondizionato. ■

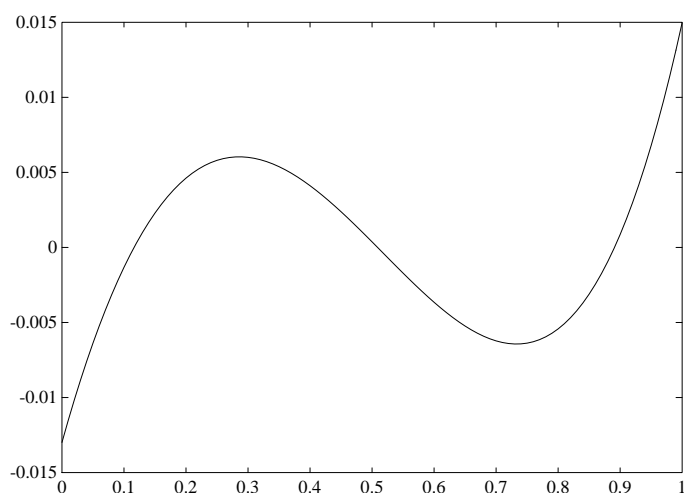


Figura 4.15: Errore  $f(x) - g_n^*(x)$  su  $[0, 1]$  per  $f(x) = e^x$  e  $g_n^*$  polinomio di secondo grado di migliore approssimazione nel senso dei minimi quadrati.

► **Esempio 4.7** (*Caso discreto*) Consideriamo un modo alternativo di ottenere le equazioni normali (4.18). I coefficienti  $c_0^*, c_1^*, \dots, c_n^*$  possono essere determinati minimizzando la funzione errore

$$d(c_0, c_1, \dots, c_n) = \sum_{i=0}^m w(x_i) (f(x_i) - g_n^*(x_i))^2$$

Ricordiamo allora che in un punto di minimo si ha

$$\frac{\partial d}{\partial c_k} = 0, \quad k = 0, 1, \dots, n$$

da cui

$$c_0 \sum_{i=0}^n w_i \phi_0(x_i) \phi_k(x_i) + \dots + c_n \sum_{i=0}^n w_i \phi_n(x_i) \phi_k(x_i) = \sum_{i=0}^n w_i \phi_k(x_i) f(x_i)$$

che coincide con il sistema (4.17). In termini matriciali il sistema precedente può essere scritto nel seguente modo. Introducendo per brevità la notazione

$$\mathbf{tab} f = [f(x_0), f(x_1), \dots, f(x_m)]^T$$

e indicando con  $\mathbf{A}$  la matrice di colonne  $\mathbf{tab} \phi_i$ ,  $i = 0, 1, \dots, n$ , si ha

$$\mathbf{tab} g_n^* = [\mathbf{tab} \phi_0, \mathbf{tab} \phi_1, \dots, \mathbf{tab} \phi_n] \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \mathbf{A} \mathbf{c}$$

Consideriamo, quindi, il seguente sistema di  $m + 1$  equazioni e  $n + 1$  incognite

$$\mathbf{tab} g_n^* = \mathbf{tab} f, \quad \iff \quad \mathbf{A} \mathbf{c} = \mathbf{b}$$

ove si è posto  $\mathbf{b} := \mathbf{tab} f$ . Il sistema delle equazioni normali corrisponde allora al seguente sistema di  $n$  equazioni e  $n$  incognite

$$\mathbf{A}^T \mathbf{A} \mathbf{c} = \mathbf{A}^T \mathbf{b} \quad (4.19)$$

In base al Teorema 4.5 la matrice  $\mathbf{A}^T \mathbf{A}$  è non singolare quando le funzioni  $\phi_i$ ,  $i = 0, 1, \dots, n$  sono linearmente indipendenti sui punti  $x_0, x_1, \dots, x_m$ , cioè quando i vettori  $\mathbf{tab} \phi_i$ ,  $i = 0, 1, \dots, n$  sono linearmente indipendenti. Ad esempio, nel caso in cui  $\phi_i = x^i$  una condizione sufficiente per la indipendenza lineare è che sia  $m \geq n$  e i punti  $x_i$  siano distinti. Infatti, se i vettori  $\mathbf{tab} \phi_i$  fossero linearmente dipendenti, si avrebbe  $\sum_{j=0}^n c_j \mathbf{tab} \phi_j = \mathbf{0}$  per  $c_0, c_1, \dots, c_n$  non tutti nulli; ma questo significa  $\sum_{j=0}^n c_j x_i^j = 0$  per  $i = 0, 1, \dots, m$ . Si avrebbe, quindi, un polinomio di grado  $\leq n$ , non identicamente nullo con un numero di zeri  $\geq n + 1$ .

Osserviamo anche che la matrice  $\mathbf{A}^T \mathbf{A}$  è una matrice *simmetrica definita positiva*. Per la risoluzione numerica del sistema (4.19) è, quindi, possibile utilizzare il *metodo di Choleski*. In caso di malcondizionamento, sono, tuttavia, opportune procedure più stabili, quali il metodo **QR** o la *decomposizione in valori singolari* (cfr. Capitolo 2 e Appendice A).

Ricordiamo, infine, che il calcolo del polinomio di migliore approssimazione nel senso dei minimi quadrati è noto in statistica come *problema di regressione polinomiale*. In tale contesto, e quando la dispersione dei punti è descrivibile tramite una distribuzione normale, si introduce la seguente misura della dispersione dei dati intorno al polinomio di regressione

$$s_{y/x} = \sqrt{\frac{S_r}{m + 1 - (n + 1)}}$$

ove  $S_r = \sum_{i=0}^m (f(x_i) - c_0 - c_1 x_i - c_2 x_i^2 - \dots - c_n x_i^n)^2$ . Il denominatore  $m + 1 - (n + 1)$  indica il numero dei gradi di libertà, tenendo conto dei coefficienti  $c_i$  del polinomio. Oltre alla quantità  $s_{y/x}$ , nota come *errore standard della stima*, si può calcolare il *coefficiente di correlazione*  $r$  ponendo

$$r^2 = \frac{S_t - S_r}{S_t}$$

ove  $S_t$  indica la somma dei quadrati delle differenze tra la variabile dipendente (la  $y$ ) e il suo valore medio. La differenza  $S_t - S_r$  quantifica la riduzione dell'errore ottenuta assumendo come modello un polinomio di grado  $n$ . Nel caso di un'approssimazione ideale si

ha  $S_r = 0$  e  $r^2 = 1$ , ossia il polinomio di regressione rappresenta perfettamente i dati. Al contrario, se  $r^2 = 0$  l'approssimazione data dal polinomio non porta alcun vantaggio. Per un approfondimento delle nozioni ora introdotte si veda il Capitolo 8. ■

### 4.2.2 Polinomi ortogonali

Nel paragrafo precedente abbiamo visto che per particolari scelte delle funzioni  $\phi_j$ , ad esempio per  $\phi_j(x) = x^j$ ,  $j = 0, 1, \dots, n$ , il sistema delle equazioni normali può essere *malcondizionato*, con conseguenti difficoltà numeriche per la sua risoluzione. Tali difficoltà possono essere evitate assumendo  $\phi_j$  come elementi di una base di polinomi ortogonali, cioè tali che per una fissata funzione peso  $w$  si abbia  $(\phi_j, \phi_k) = 0$  per  $j \neq k$ . In questo caso, infatti, le equazioni normali diventano

$$c_k(\phi_k, \phi_k) = (f, \phi_k), \quad k = 0, 1, \dots, n$$

e i coefficienti  $c_k$ , chiamati anche *coefficienti di Fourier*, possono essere ottenuti direttamente

$$c_k = \frac{(f, \phi_k)}{(\phi_k, \phi_k)}, \quad k = 0, 1, \dots, n$$

Rileviamo un altro vantaggio nell'uso dei polinomi ortogonali. Se  $p_n^*$  è il polinomio di migliore approssimazione di grado  $\leq n$ , per ottenere il polinomio di migliore approssimazione di grado  $\leq (n+1)$  è sufficiente calcolare  $c_{n+1}^*$  e porre  $p_{n+1}^* = p_n^* + c_{n+1}^* \phi_{n+1}$ .

► **Esempio 4.8** Come esempio introduttivo, calcoliamo i polinomi  $P_i$  ortogonali rispetto al seguente prodotto scalare

$$(f, g) = \int_{-1}^1 f(x)g(x) dx$$

Assunto  $P_0 = 1$ , cerchiamo  $P_1(x) = x + a_{11}$  che sia ortogonale a  $P_0$ . Si ha

$$\int_{-1}^1 (x + a_{11}) dx = 0$$

da cui  $a_{11} = 0$ . In modo analogo, posto  $P_2 = x^2 + a_{21}x + a_{22}$ , da  $(P_2, P_0) = (P_2, P_1) = 0$  si ricava  $a_{22} = -1/3$  e  $a_{21} = 0$ , e quindi  $P_2(x) = x^2 - 1/3$ . ■

Più in generale, si può dimostrare il seguente risultato che si dimostra con la stessa procedura seguita nell'esempio precedente.

**Teorema 4.6 (polinomi ortogonali)** *Dato un generico prodotto scalare, esiste una successione di polinomi ortogonali  $\{\phi_i\}$ ,  $i = 0, 1, \dots$ , con  $\phi_i$  polinomio di grado  $i$ . I coefficienti di grado massimo possono essere scelti arbitrariamente; quando sono fissati, il sistema di polinomi ortogonali è determinato in maniera univoca. I polinomi*

possono essere calcolati attraverso la seguente formula ricorrente

$$\begin{aligned}\phi_{n+1}(x) &= (\alpha_n x - c_{n,n}) \phi_n(x) - c_{n,n-1} \phi_{n-1}(x), \quad n \geq 0 \\ \phi_{-1}(x) &\equiv 0 \\ \phi_0(x) &\equiv A_0,\end{aligned}$$

ove

$$c_{n,n} = \frac{\alpha_n(x\phi_n, \phi_n)}{(\phi_n, \phi_n)}, \quad c_{n,n-1} = \frac{\alpha_n(\phi_n, \phi_n)}{\alpha_{n-1}(\phi_{n-1}, \phi_{n-1})}$$

Nel caso discreto, con nodi  $x_0, x_1, \dots, x_m$ , l'ultimo polinomio nella successione è  $\phi_m$ .

Una proprietà importante dei polinomi ortogonali è espressa dal seguente risultato.

**Proposizione 4.1** *Il polinomio di grado  $n$  in una famiglia di polinomi ortogonali, associata ad una funzione peso  $w$  su un intervallo  $[a, b]$ , ha  $n$  zeri reali semplici, tutti contenuti nell'intervallo aperto  $]a, b[$ .*

► **Esempio 4.9** Costruiamo i polinomi  $\phi_0, \phi_1, \phi_2$ , con coefficienti di grado massimo uguali a 1, che siano ortogonali rispetto al prodotto scalare

$$(f, g) = \sum_{i=0}^2 f(x_i)g(x_i)$$

ove  $x_0 = -\sqrt{3}/2$ ,  $x_1 = 0$  e  $x_2 = \sqrt{3}/2$ . Usando le formule del Teorema 4.6 e ponendo  $A_0 = 1$  e  $\alpha_n = 1$ , si ha  $\phi_0 = 1$  e

$$(x\phi_0, \phi_0) = \sum_{i=0}^2 x_i = 0$$

da cui  $\phi_1 = x\phi_0 = x$ . In modo analogo, osservando che

$$(x\phi_1, \phi_1) = \sum_{i=0}^2 x_i^3 = 0, \quad (\phi_0, \phi_0) = 3, \quad (\phi_1, \phi_1) = \sum_{i=0}^2 x_i^2 = \frac{3}{2}$$

si trova  $\phi_2 = x^2 - 1/2$ . ■

### Polinomi di Legendre

Osserviamo che un intervallo arbitrario  $[a, b]$  può essere trasformato nell'intervallo  $[-1, 1]$  mediante il seguente cambiamento di variabili

$$t = \frac{2}{b-a} \left( x - \frac{b+a}{2} \right), \quad x \in [a, b], \quad t \in [-1, 1]$$

Pertanto, nel seguito ci limiteremo a considerare l'intervallo di riferimento  $[-1, 1]$ .

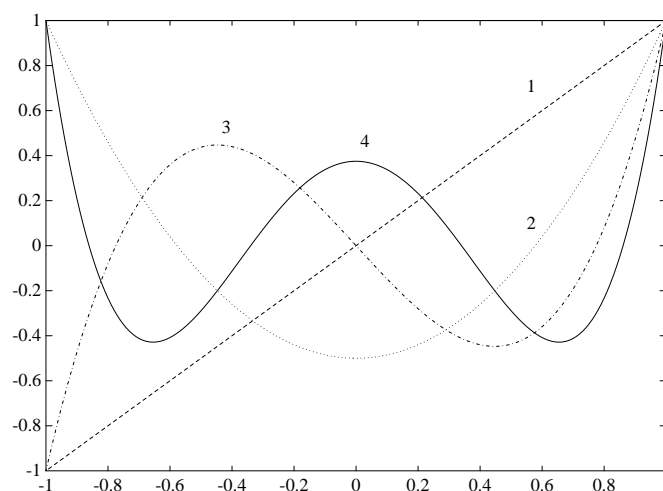


Figura 4.16: Polinomi di Legendre di grado 1, 2, 3, 4.

I *polinomi di Legendre*  $P_n(x)$  sono polinomi ortogonali su  $[-1, 1]$  rispetto alla funzione peso  $w(x) = 1$ , cioè

$$\int_{-1}^1 P_i(x)P_k(x) dx = 0, \quad i \neq k$$

Usualmente sono scalati in maniera da avere  $P_n(1) = 1$  per ogni  $n$ . In questo caso essi soddisfano alla seguente formula ricorrente

$$P_0(x) = 1, \quad P_1(x) = x$$

$$P_{n+1}(x) = \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x), \quad n \geq 1$$

Si ha, ad esempio

$$P_2(x) = \frac{1}{2} (3x^2 - 1), \quad P_3(x) = \frac{1}{2} (5x^3 - 3x), \quad P_4(x) = \frac{1}{8} (35x^4 - 30x^2 + 3)$$

Tali polinomi sono illustrati in Figura 4.16. Un modo alternativo di definire i polinomi di Legendre è dato dalla seguente formula

$$P_0(x) = 1; \quad P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n \geq 1$$

### Polinomi di Chebichev

I *polinomi di Chebichev*<sup>7</sup> sono definiti per  $n \geq 0$  e  $x \in [-1, 1]$  da

$$T_n(x) = \cos(n \arccos x)$$

<sup>7</sup>la notazione ormai comune  $T_n(x)$  per indicare i polinomi di Chebichev deriva da *Tschebyscheff*, uno dei tanti modi differenti di trascrivere il nome del matematico russo.



Per ricavare una formula ricorrente a tre termini per i polinomi di Chebichev, si utilizza l'identità  $e^{i\theta} = \cos \theta + i \sin \theta$ . Posto  $x = \cos \theta$ , per  $\theta \in [0, \pi]$ , poiché  $\cos \theta = \Re(e^{i\theta})$ , si ha, per  $n \geq 1$

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \cos(n+1)\theta + \cos(n-1)\theta \\ &= \Re(e^{i(n+1)\theta} + e^{i(n-1)\theta}) = \Re(e^{in\theta}(e^{i\theta} + e^{-i\theta})) \\ &= 2 \cos \theta \cos n\theta = 2xT_n(x) \end{aligned}$$

Si ha, pertanto

$$\begin{aligned} T_0(x) &= 1, \quad T_1(x) = x \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \quad n \geq 1 \end{aligned}$$

Si ha, ad esempio (cfr. Figura 4.17)

$$\begin{aligned} T_2(x) &= 2x^2 - 1, & T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1, & T_5(x) &= 16x^5 - 20x^3 + 5x \end{aligned}$$

Usando la formula ricorrente, si può vedere che il coefficiente di  $x^n$  in  $T_n(x)$  è  $2^{n-1}$ ,

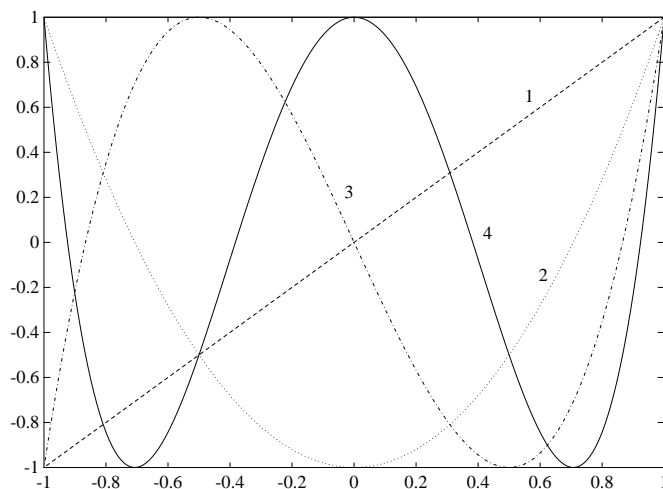


Figura 4.17: Polinomi di Chebichev di grado 1, 2, 3, 4.

per  $n \geq 1$ . Inoltre, si ha

$$T_n(-x) = (-1)^n T_n(x)$$

Gli zeri  $x_k$  dei polinomi di Chebichev sono ottenuti da

$$T_n(x_k) = \cos(n \arccos x_k) = 0 \Rightarrow x_k = \cos \frac{2k+1}{n} \cdot \frac{\pi}{2}, \quad k = 0, 1, \dots, n-1$$

Tra gli zeri,  $T_n$  assume i valori estremali  $+1$  e  $-1$  nei punti

$$x'_k = \cos \frac{k\pi}{n}, \quad k = 0, 1, \dots, n$$

I polinomi di Chebichev sono ortogonali rispetto alla funzione peso  $1/\sqrt{1-x^2}$  nell'intervallo  $[-1, 1]$ ; si ha infatti

$$\begin{aligned} (T_j, T_k) &= \int_{-1}^1 \frac{T_j(x)T_k(x)}{\sqrt{1-x^2}} dx = \int_0^\pi T_j(\cos \theta) T_k(\cos \theta) d\theta = \int_0^\pi \cos j\theta \cos k\theta d\theta \\ &= \frac{1}{2} \int_0^\pi (\cos(j+k)\theta + \cos(j-k)\theta) d\theta = \begin{cases} 0 & \text{per } j \neq k \\ \frac{\pi}{2} & \text{per } j = k \neq 0 \\ \pi & \text{per } j = k = 0 \end{cases} \end{aligned}$$

Si può dimostrare che  $T_0, T_1, \dots, T_m$  sono ortogonali rispetto al prodotto scalare

$$(f, g) = \sum_{i=0}^m f(x_i)g(x_i)$$

ove  $x_i$  sono gli zeri di  $T_{m+1}(x)$ . Osserviamo che nel caso discreto, la funzione peso è 1. Un caso particolare è stato considerato nell'Esempio 4.9.

L'interesse dei polinomi di Chebichev nel calcolo numerico è indicato dal seguente risultato.

**Teorema 4.7** (proprietà di minimax dei polinomi di Chebichev) *Tra tutti i polinomi con coefficiente di grado massimo uguale a 1, il polinomio  $2^{-(n-1)}T_n$  ha la minima norma del massimo sull'intervallo  $[-1, 1]$ .*

Volendo, ad esempio, calcolare il polinomio  $q_3(x)$ , di grado  $\leq 3$ , che minimizza

$$\max_{-1 \leq x \leq 1} |x^4 - q_3(x)|$$

si può utilizzare il fatto che il polinomio di Chebichev  $\frac{1}{8}T_4(x) = x^4 - x^2 + \frac{1}{8}$  è il polinomio di quarto grado, con coefficiente di grado massimo uguale a 1, che ha la più piccola norma del massimo su  $[-1, 1]$ . Si può, allora, scegliere  $q_3(x)$  in modo che

$$x^4 - q_3(x) = \frac{1}{8}T_4(x) \Rightarrow q_3(x) = x^2 - \frac{1}{8}$$

In Figura 4.18 è rappresentata la funzione errore  $x^4 - q_3(x)$ ; come si vede, tale funzione oscilla tra i suoi estremi  $\pm \frac{1}{8}$  in cinque punti. Come vedremo nel paragrafo successivo, si tratta di una proprietà caratteristica dell'approssimazione nella norma del massimo.

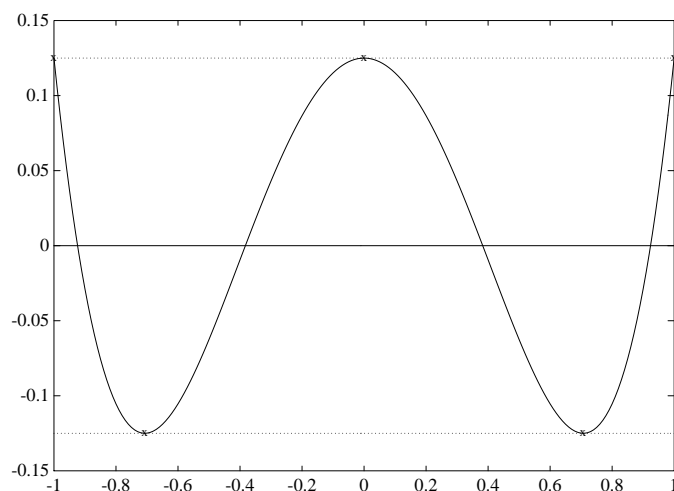


Figura 4.18: Funzione errore  $x^4 - q(x)$ , con  $q(x)$  polinomio di migliore approssimazione di  $x^4$  nella norma del massimo su  $[-1, 1]$ .

### 4.2.3 Norma del massimo. Approssimazione di Chebichev

In questo paragrafo consideriamo il problema della migliore approssimazione quando la norma è definita da

$$\|f\|_{\infty} = \max_{a \leq x \leq b} |f(x)|$$

ove  $f(x) \in C^0([a, b])$ , con  $[a, b]$  intervallo limitato e chiuso. Si può dimostrare che per ogni intero  $n$  fissato e ogni  $f \in C^0([a, b])$  esiste uno ed un solo polinomio  $p_n^*(x)$ , di grado  $\leq n$ , tale che

$$E_n(f) := \|f - p_n^*\|_{\infty} \leq \|f - p_n\|$$

per tutti i polinomi  $p_n$  di grado  $\leq n$ . Inoltre, applicando il teorema di Weierstrass, si ha  $E_n(f) \rightarrow 0$  per  $n \rightarrow \infty$ . Quando  $f$  è sufficientemente regolare, ad esempio dotata di derivata  $k$ -ma continua, si può dimostrare che  $E_n(f) = O(1/n^k)$ , per  $n \rightarrow \infty$ .

Il calcolo numerico del polinomio  $p_n^*$  è, in generale, decisamente più complicato di quello relativo al polinomio di migliore approssimazione nella norma euclidea. Ci limiteremo, pertanto, a considerare alcune idee sull'intervallo di riferimento  $[-1, 1]$ .

Incominciamo ad osservare che  $p_n^*$  può essere approssimato mediante un opportuno polinomio di interpolazione. Abbiamo visto (cfr. Esempio di Runge (4.6)) che, quando nella interpolazione sono utilizzati punti equidistanti si possono avere grossi errori vicino agli estremi dell'intervallo. Questo fatto suggerisce di utilizzare un numero maggiore di nodi vicino agli estremi, in maniera da costringere il polinomio a seguire meglio la funzione. Una distribuzione di questo tipo è ottenuta mediante gli zeri dei polinomi di Chebichev. Il fondamento teorico a questa scelta è il seguente.

Ricordiamo che l'errore di interpolazione, quando la funzione  $f(x)$  è sufficientemente regolare, è dato da

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x-x_1)(x-x_2)\cdots(x-x_{n+1})$$

Si vede, quindi, che l'errore dipende sia dal comportamento della derivata di ordine  $(n+1)$ -ma della funzione che dal polinomio  $(x-x_1)(x-x_2)\cdots(x-x_{n+1})$ . Questo secondo termine può essere appunto minimizzato scegliendo come nodi  $x_i$  gli zeri del polinomio di Chebychev di grado  $n+1$ . Infatti, in base al Teorema 4.7, si ha che, assumendo

$$(x-x_1)(x-x_2)\cdots(x-x_{n+1}) = 2^{-n} T_{n+1}(x)$$

viene minimizzato il termine

$$\max_{-1 \leq x \leq 1} |(x-x_1)(x-x_2)\cdots(x-x_{n+1})|$$

Lasciamo come esercizio verificare che, nel caso in cui  $f = x^{n+1}$ , il polinomio di interpolazione negli zeri del polinomio di Chebychev di grado  $n+1$  coincide con il polinomio di migliore approssimazione nella norma del massimo.

Un'altra proprietà utile nella costruzione del polinomio di migliore approssimazione secondo la norma del massimo è espressa dal seguente risultato (cfr. per una illustrazione la Figura 4.18).

**Teorema 4.8 (proprietà di alternanza)** *Sia  $f(x) \in C^0([a, b])$ , con  $[a, b]$  intervallo limitato e chiuso. Un polinomio  $p_n^*$  è il polinomio di migliore approssimazione nella norma del massimo della funzione  $f$  nell'ambito dei polinomi di grado  $\leq n$  se e solo se esistono  $n+2$  punti  $x_0, x_1, \dots, x_{n+1}$  tali che*

$$|f(x_k) - p_n^*(x_k)| = \|f - p_n^*\|_\infty, \quad k = 0, 1, \dots, n+1$$

e

$$f(x_k) - p_n^*(x_k) = -(f(x_{k+1}) - p_n^*(x_{k+1})), \quad k = 0, 1, \dots, n$$

In altre parole, la funzione errore  $f - p_n^*$  assume alternativamente i valori  $\pm \|f - p_n^*\|$  in almeno  $(n+2)$  punti.

► **Esempio 4.10** Come illustrazione del Teorema 4.8 calcoliamo il polinomio di migliore approssimazione di grado 1 della funzione  $e^x$  sull'intervallo  $[0, 1]$ . Si tratta, quindi, di trovare  $c_0, c_1$  tali che sia minimizzata la quantità

$$\max_{0 \leq x \leq 1} |e^x - (c_0 + c_1 x)|$$

Indicando con  $d$  il valore estrema assunto dalla funzione errore  $e(x) = e^x - (c_0 + c_1 x)$ , si vede facilmente che tale valore è assunto negli estremi dell'intervallo e in un punto interno  $\xi$ . Si ha, quindi

$$e(0) = d, \quad e(\xi) = -d, \quad e(1) = d$$

Nel punto  $\xi$  si deve, inoltre, avere  $e'(\xi) = 0$ . Si hanno pertanto le seguenti equazioni

$$\begin{aligned} 1 - c_0 &= d \\ e^\xi - c_0 - c_1\xi &= -d \\ e - c_0 - c_1 &= d \\ e^\xi - c_1 &= 0 \end{aligned}$$

da cui con semplici calcoli, si ottiene

$$\begin{aligned} c_0 &= \frac{1}{2}(e - (e - 1)\ln(e - 1)) \approx 0.894066; & c_1 &= e - 1 \approx 1.718281 \\ \xi &= \ln(e - 1) \approx 0.541324, & d &= 1 - c_0 \approx 0.105933 \end{aligned}$$

■

Per  $n > 1$  il Teorema 4.8 è utilizzato in un algoritmo iterativo, noto come *algoritmo di Remes*. L'idea dell'algoritmo è, in forma schematica, la seguente.

1. Si parte da un insieme di punti  $x_i^{(0)}$ , che possono ad esempio essere gli zeri del polinomio di Chebichev di grado  $n + 1$ .
2. Imponendo la condizione di alternanza, si calcolano, come nell'esempio precedente, i coefficienti del polinomio e la distanza  $d$ .
3. Si calcolano i punti interni  $x_j^{(1)}$  nei quali la funzione errore raggiunge il massimo o il minimo.
4. Se la funzione errore verifica la condizione di alternanza, si è ottenuto il polinomio richiesto. In caso contrario si riparte dai punti  $x_j^{(1)}$ .

◆ **Esercizio 4.8** Approssimare la funzione  $f(x) = \sqrt[3]{x}$  mediante un polinomio di primo grado sull'intervallo  $[0, 1]$

- (a) nel senso dei minimi quadrati con funzione peso 1;
- (b) nella norma del massimo.

◆ **Esercizio 4.9** Determinare i parametri  $a, b, c$  nella funzione

$$z = ax + by + c$$

mediante il metodo dei minimi quadrati a partire dai seguenti dati

$x$	0	1.2	2.1	3.4	4.0	4.2	5.6	5.8	6.9
$y$	0	0.5	6.0	0.5	5.1	3.2	1.3	7.4	10.2
$z$	1.2	3.4	-4.5	9.9	2.4	7.2	14.3	3.5	1.2

◆ **Esercizio 4.10** Si consideri la seguente legge sperimentale  $V = a + bT + cT^2$ , ove  $V$  è la viscosità di un liquido e  $T$  è la temperatura. Trovare mediante il metodo dei minimi quadrati i parametri  $a, b$  e  $c$  a partire dai seguenti dati sperimentali

$T$	1	2	3	4	5	6	7
$V$	2.31	2.01	1.80	1.66	1.55	1.47	1.41

◆ **Esercizio 4.11** L'azoto e l'ossigeno hanno i pesi atomici  $N \approx 14$  e  $O \approx 16$ . Usando il peso molecolare dei seguenti sei ossidi di azoto,

$$\begin{array}{lll} NO = 30.006 & N_2O = 44.013 & NO_2 = 46.006 \\ N_2O_3 = 76.012 & N_2O_5 = 108.010 & N_2O_4 = 92.011 \end{array}$$

stimare mediante il metodo dei minimi quadrati i pesi atomici dell'azoto e dell'ossigeno a quattro cifre decimali.

◆ **Esercizio 4.12** Costruire il polinomio di primo grado di migliore approssimazione nel senso dei minimi quadrati con i polinomi di Legendre e rispettivamente di Chebichev per le seguenti funzioni

- (a)  $f(x) = \ln x$  su  $[1, 2]$ ,  
 (b)  $f(x) = \sin \pi x$  su  $[0, 1]$

### 4.3 Calcolo numerico delle derivate

La necessità del *calcolo approssimato* della derivata di una funzione si pone, ad esempio, quando la funzione è data in forma di tabella, cioè quando essa è nota solo in un numero finito di punti. In questo caso, e quando è necessario calcolare le derivate in un numero elevato di punti, può essere conveniente calcolare una *spline interpolante* e approssimare le derivate della funzione con le derivate della corrispondente spline. L'uso dei polinomi di interpolazione, invece, non è in generale opportuno, in quanto, come abbiamo visto nei paragrafi precedenti, le derivate del polinomio di interpolazione non sono sempre buone approssimazioni delle derivate della funzione.

Un procedimento alternativo si basa sull'uso di *formule alle differenze*, cioè di opportuni rapporti incrementali. Nel seguito esamineremo, in particolare dal punto di vista dell'errore di troncamento, alcune di tali formule, che sono anche alla base, come vedremo nel successivo Capitolo 7, di metodi per la risoluzione numerica di equazioni differenziali.

Assumendo che la funzione  $f(x)$  sia nota nei punti  $x-h$ ,  $x$  e  $x+h$ , possiamo stimare  $f'(x)$ , cioè il coefficiente angolare della tangente alla curva  $y = f(x)$  nel punto  $x$ , mediante il coefficiente angolare di differenti rette (cfr. Figura 4.19). Usando la corda 1, relativa ai punti  $(x, f(x))$  e  $(x+h, f(x+h))$ , si ottiene

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} =: \nabla f$$

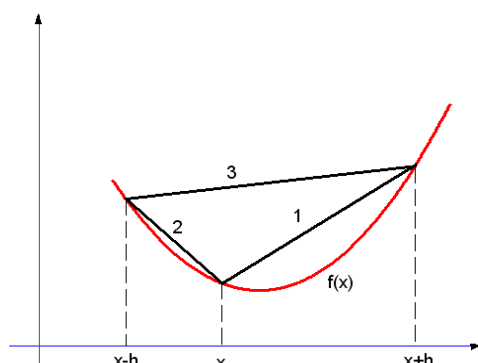


Figura 4.19: Approssimazione della derivata  $f'(x)$  mediante il coefficiente angolare di tre differenti corde.

L'approssimazione ottenuta è chiamata *differenza in avanti* (forward difference).

Utilizzando la retta 2 attraverso i punti  $(x, f(x))$  e  $(x - h, f(x - h))$ , si ottiene la seguente approssimazione, chiamata *differenza all'indietro* (backward difference):

$$f'(x) \approx \frac{f(x) - f(x - h)}{h} =: \bar{\nabla} f$$

Una approssimazione simmetrica è ottenuta mediante la retta 3 passante per i punti  $(x - h, f(x - h))$  e  $(x + h, f(x + h))$ ; si ha, allora

$$f'(x) \approx \frac{f(x + h) - f(x - h)}{2h} =: \delta f$$

che è chiamata *differenza centrale* (central difference).

Procedendo in maniera analoga si possono approssimare derivate di ordine superiore. Ad esempio, approssimando la  $f$  mediante un polinomio di secondo grado attraverso i punti  $(x - h, f(x - h))$ ,  $(x, f(x))$  e  $(x + h, f(x + h))$ , si ottiene la seguente approssimazione alle differenze centrali

$$f''(x) \approx \frac{f(x + h) - 2f(x) + f(x - h)}{h^2} =: \nabla \bar{\nabla} f$$

### 4.3.1 Studio dell'errore di troncamento

In Tabella 4.1 sono riportati, per successivi valori di  $h$ , gli errori che si commettono approssimando la derivata prima della funzione  $e^x$  nel punto  $x = 1$  mediante la differenza in avanti e rispettivamente la differenza centrale. Nel primo caso si vede

$h$	$\nabla e^x - e^x$	$r$	$\delta e^x - e^x$	$r$
0.4	0.624013	2.14	0.0730696	4.02
0.2	0.290893	2.06	0.0181581	4.00
0.1	0.140560	2.03	0.0045327	4.00
0.05	0.069103	2.01	0.0011327	4.00
0.025	0.034263		0.0002831	

Tabella 4.1: Errori di troncamento relativi all'approssimazione della derivata prima della funzione  $e^x$  nel punto  $x = 1$  mediante l'operatore di differenza in avanti e rispettivamente l'operatore della differenza centrale. Il valore  $r$  rappresenta il rapporto fra una approssimazione e la successiva.

che l'errore è, approssimativamente, dimezzato quando il passo è ridotto di un fattore due, mentre nel secondo caso è, sempre approssimativamente, ridotto di un fattore quattro. I risultati suggeriscono, quindi, una convergenza a zero di tipo lineare nel primo caso e quadratica nel secondo.

Più in generale, l'errore di troncamento può essere studiato, quando la funzione  $f(x)$  è sufficientemente regolare, mediante uno sviluppo in serie. Ad esempio, per una funzione  $f(x)$  due volte continuamente derivabile si ha

$$\begin{aligned} E_T &:= \frac{f(x+h) - f(x)}{h} - f'(x) \\ &= \frac{1}{h} (f(x) + hf'(x) + \frac{h^2}{2} f''(\xi) - f(x)) - f'(x) = \frac{h}{2} f''(\xi) \end{aligned}$$

ove  $\xi$  è un punto opportuno nell'intervallo  $(x, x+h)$ . Si vede, quindi, che l'errore è un infinitesimo con  $h$  del primo ordine per  $h \rightarrow 0$  (cioè  $E_T = O(h)$ ).

Procedendo in maniera analoga, si trova nel caso della approssimazione mediante la differenza centrale

$$\frac{f(x+h) - f(x-h)}{2h} - f'(x) = \frac{h^2}{6} f'''(\xi) = O(h^2)$$

e

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - f''(x) = \frac{h^2}{12} f^{(4)}(\xi)$$

### 4.3.2 Influenza degli errori di arrotondamento

L'influenza degli errori di arrotondamento nel calcolo approssimato della derivata è illustrata in Tabella 4.2, nella quale sono riportati i risultati ottenuti mediante la differenza centrale per la funzione  $e^x$  in  $x = 1$ . I calcoli sono stati eseguiti con precisione macchina  $\text{eps} \approx 2.22 \times 10^{-16}$ . Come si vede, l'errore diminuisce fino ad un certo valore di  $h$ , per poi aumentare. Questo comportamento è dovuto al fatto che i valori  $e^{1 \pm h}$  sono calcolati in modo approssimato e per  $h$  piccolo si verifica una cancellazione.



$h$	$\delta e^x - e^x$
$1 \cdot 10^{-1}$	0.004532
$1 \cdot 10^{-2}$	0.0000453
$1 \cdot 10^{-3}$	0.0000000453
$1 \cdot 10^{-4}$	0.00000000453
$1 \cdot 10^{-8}$	-0.0000000165
$1 \cdot 10^{-9}$	0.0000000740
$1 \cdot 10^{-10}$	0.000000224
$1 \cdot 10^{-14}$	-0.002172

Tabella 4.2: Influenza degli errori di arrotondamento nel calcolo approssimato della derivata prima della funzione  $e^x$  per  $x = 1$ .

Più in generale, indichiamo con  $\tilde{f}(x \pm h)$  i valori approssimati di  $f(x \pm h)$  e supponiamo che

$$|\tilde{f}(x \pm h) - f(x \pm h)| \leq \epsilon$$

ove  $\epsilon$  è una quantità dipendente dalla precisione macchina utilizzata. Allora, indicata con  $\tilde{\delta}$  la differenza centrale, calcolata numericamente, si ha la seguente maggiorazione dell'errore di troncamento

$$|\tilde{\delta}f(x) - f'(x)| \leq \frac{\epsilon}{h} + \frac{h^2}{6} |f^{(3)}(\xi)|$$

Per  $h \rightarrow 0$ , si ha pertanto una componente dell'errore che tende all'infinito; in altre parole, al diminuire del passo  $h$  gli errori di arrotondamento possono prevalere sull'errore di troncamento e fornire quindi risultati inattendibili. La situazione è illustrata in Figura 4.20, ove è indicata la scelta ottimale del passo  $h$  corrispondente al minimo della funzione  $\epsilon/h + h^2|f^{(3)}|/6$ .

◆ **Esercizio 4.13** Determinare l'errore di troncamento delle seguenti approssimazioni

(a)  $f'(x) \approx \frac{f(x+3h) - f(x-h)}{4h}$

(b)  $f'(x) \approx \frac{4f(x+h) - 3f(x) - f(x+2h)}{2h}$

(c)  $f'(x) \approx \frac{f(x-2h) - 8f(x-h) + 8f(x+h) - f(x+2h)}{12h}$

◆ **Esercizio 4.14** Analizzare la seguente approssimazione

$$\left(\frac{2+h}{2h^2}\right) f(x+h) - \frac{2}{h^2} f(x) + \left(\frac{2-h}{2h^2}\right) f(x-h)$$

dell'espressione  $f'(x) + f''(x)$ .

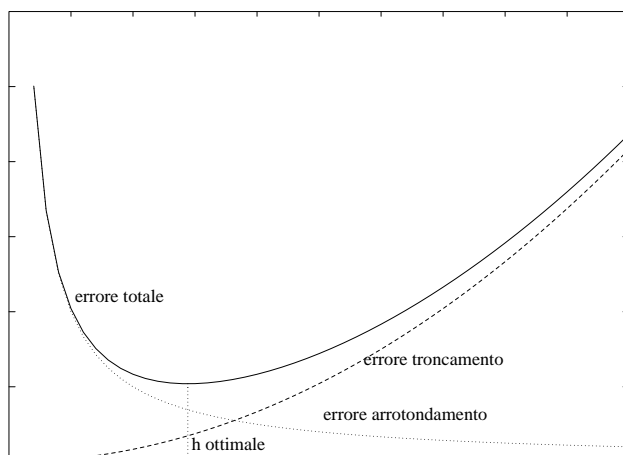


Figura 4.20: Errore di arrotondamento e errore di troncamento nell'approssimazione della derivata mediante le differenze centrali.

◆ **Esercizio 4.15** Analizzare le seguenti approssimazioni

$$(a) \quad f''(x) \approx \frac{f(x+2h) - 2f(x) + f(x-2h)}{4h^2}$$

$$(b) \quad f'''(x) \approx \frac{f(x+2h) - 2f(x+h) + 2f(x-h) - f(x-2h)}{2h^3}$$

◆ **Esercizio 4.16** Dati i punti  $x_1 < x_2 < x_3$ , con  $x_2 - x_1 = h$  e  $x_3 - x_2 = sh$ , analizzare la seguente formula

$$f''(x) \approx \frac{2}{h^2} \left[ \frac{f(x_1)}{1+s} - \frac{f(x_2)}{s} + \frac{f(x_3)}{s(s+1)} \right]$$

I have no satisfaction in formulas  
 unless I feel their numerical magnitude.  
 (Lord Kelvin)

## Capitolo 5

# Equazioni non lineari e ottimizzazione

Il problema che analizzeremo in questo capitolo è il seguente. Data una funzione  $\mathbf{f}$  definita in un insieme  $D \subset \mathbb{R}^n$  e a valori in  $\mathbb{R}^m$

$$\mathbf{f} : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad \mathbf{f} \equiv [f_1, f_2, \dots, f_m]^T$$

e un vettore  $\mathbf{b} \in \mathbb{R}^m$ , si cerca  $\mathbf{x} \in D$  tale che

$$\mathbf{f}(x) = \mathbf{b} \iff \begin{cases} f_1(x_1, \dots, x_n) = b_1 \\ f_2(x_1, \dots, x_n) = b_2 \\ \vdots \\ f_m(x_1, \dots, x_n) = b_m \end{cases}$$

Nel Capitolo 2 è stato studiato il caso particolare in cui la funzione  $\mathbf{f}$  è lineare, ossia quando  $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ , con  $\mathbf{A} \in \mathbb{R}^{m \times n}$ . Nel caso generale, si preferisce di solito scrivere il problema precedente nella seguente forma

$$\mathbf{F}(\mathbf{x}) = 0 \tag{5.1}$$

nella quale il termine noto  $\mathbf{b}$  è assorbito in  $\mathbf{f}$ . Il problema (5.1), detto *sistema di equazioni non lineari*, consiste nella ricerca degli *zeri* (o radici) della funzione  $\mathbf{f}$  nel dominio  $D$ . Come caso particolare, si ha per  $m = n = 1$  il problema del calcolo dello *zero di una funzione di una variabile*. Quando la funzione è un polinomio  $P$ , il problema più comune consiste nella ricerca dei valori  $z \in \mathbb{C}$  tali che  $P(z) = 0$ .

Il problema ora introdotto è *collegato* con il *problema di ottimizzazione*, consistente nella ricerca di un *minimo* (o *massimo*) di una funzione  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , ossia nel calcolo di un valore  $\mathbf{x}^* \in \mathbb{R}^n$  tale che

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^n \tag{5.2}$$

In effetti, un punto  $\mathbf{x}^*$  che verifica il sistema (5.1) è anche un minimo della seguente funzione

$$\sum_{i=1}^m (f_i(x))^2 \quad (5.3)$$

che rappresenta un caso particolare del problema più generale dei *minimi quadrati non lineari*. D'altra parte, quando la funzione  $f(\mathbf{x})$  è sufficientemente regolare, i punti di minimo (o di massimo) annullano il gradiente della funzione, ossia sono gli zeri di un sistema di equazioni.

Ne consegue che le tecniche di risoluzione delle equazioni non lineari e della ricerca dei punti di ottimalità sono essenzialmente basate sulle medesime idee. In pratica, come vedremo, l'idea di base è quella della *iterazione*, che permette di ricondurre la risoluzione del problema non lineare a quella di una successione di problemi lineari.

È superfluo sottolineare l'importanza dei modelli non lineari nelle applicazioni. A scopo introduttivo, negli esempi successivi saranno illustrati due semplici modelli per i quali è richiesta rispettivamente la risoluzione di un problema non lineare e di un problema di ottimizzazione. Per un approfondimento delle nozioni introdotte si può vedere ad esempio Ortega e Rheinboldt [124], Dennis e Schnabel [46].

► **Esempio 5.1** (*Equazione di stato di un gas*) Per modellizzare la relazione tra la pressione, il volume e la temperatura di un gas sono stati introdotti differenti tipi di *equazioni di stato*. Una delle più note di tali equazioni è l'*equazione di Beattie–Bridgeman*, espressa nella seguente forma

$$P = \frac{RT}{V} + \frac{\beta}{V^2} + \frac{\gamma}{V^3} + \frac{\delta}{V^4} \quad (5.4)$$

ove  $P$  è la pressione (atm),  $V$  è il volume molare (litro),  $T$  è la temperatura ( $^{\circ}\text{K}$ ),  $\beta$ ,  $\gamma$  e  $\delta$  sono parametri caratteristici di un gas e dipendenti dalla temperatura, e  $R$  è la costante universale dei gas in unità compatibili (atm litro /  $^{\circ}\text{K}$  g–mole). L'equazione di Beattie–Bridgeman è un'opportuna modifica della legge ideale dei gas

$$P = \frac{RT}{V}$$

I parametri  $\beta$ ,  $\gamma$  e  $\delta$  sono definiti da

$$\beta = RTB_0 - A_0 - \frac{Rc}{T^2}, \quad \gamma = -RTB_0b + aA_0 - \frac{RcB_0}{T^2}, \quad \delta = \frac{RB_0bc}{T^2}$$

ove  $A_0$ ,  $B_0$ ,  $a$ ,  $b$  e  $c$  sono delle costanti, caratteristiche di un particolare gas e determinate empiricamente a partire da dati sperimentali.

L'equazione (5.4) è *esplicita* nella pressione  $P$ , ma *implicita* nella temperatura  $T$  e nel volume  $V$ . In particolare, per trovare il volume corrispondente a valori assegnati della pressione e della temperatura sono necessari opportuni metodi numerici. Nel seguito esamineremo in particolare l'applicazione del *metodo di Newton*.

Per analizzare l'esistenza di una soluzione, riscriviamo l'equazione (5.4) nella seguente forma

$$f(V) := \left( RT + \frac{\beta}{V} + \frac{\gamma}{V^2} + \frac{\delta}{V^3} \right) \frac{1}{P} - V = 0 \quad (5.5)$$

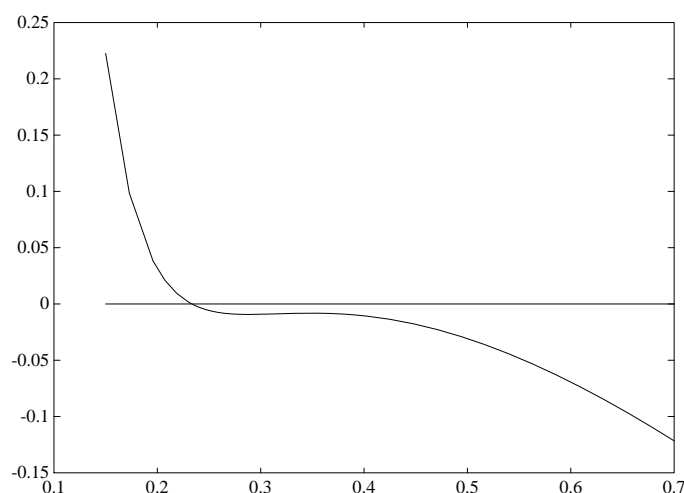


Figura 5.1: Grafico della funzione  $F(V)$  in corrispondenza al gas isobutano.

Nella Figura 5.1 è rappresentata la funzione  $F(V)$  corrispondente al gas *isobutano*, per il quale si hanno i seguenti valori

$$A_0 = 16.6037, \quad B_0 = 0.2354, \quad a = 0.11171, \quad b = 0.07697, \quad c = 3 \cdot 10^6$$

e  $T = 408^\circ\text{K}$ ,  $P = 36 \text{ atm}$  e  $R = 0.08206$ . Dalla figura si vede l'esistenza di una soluzione nell'intorno del valore  $V = 0.2$ . ■

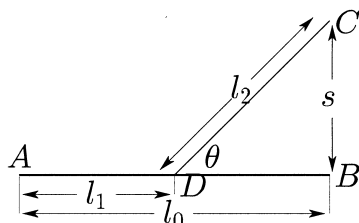


Figura 5.2: Ramificazione dei vasi sanguigni con ricerca dell'angolo ottimale  $\theta$ .

► **Esempio 5.2** (*Ramificazione ottimale dei vasi sanguigni*) Il sangue viene trasportato ai vari organi attraverso un sistema di arterie, capillari e vene. Tale sistema oppone una resistenza che dipende da diversi fattori, tra i quali il diametro dei vasi e la viscosità del sangue. Per il seguito considereremo in particolare l'influenza del diametro dei vasi, che supporremo, per semplicità, *rigidi*. Più precisamente, analizzeremo la situazione illustrata in Figura 5.2. Il segmento  $AB$  rappresenta un vaso principale di raggio  $r_1$  e un punto  $C$  è raggiunto da una diramazione  $DC$  caratterizzata da un raggio  $r_2$ . Indicato con  $\theta$  l'angolo di

diramazione  $B\hat{D}C$ , si tratta di determinare tale angolo in maniera che *sia ridotta al minimo la resistenza totale del sangue lungo la traiettoria ADC*. Tale resistenza è la somma della resistenza  $R_1$  lungo  $AD$  e della resistenza  $R_2$  lungo  $DC$ . Per una stima di tali resistenze si utilizza la *legge di Poiseuille*, che è una legge sperimentale valida per un flusso laminare, cioè un flusso nel quale le particelle del liquido si muovono parallele al tubo e la velocità aumenta regolarmente partendo da zero alla parete verso il centro. La legge di Poiseuille stabilisce che la resistenza  $R$  è proporzionale alla lunghezza  $l$  del vaso e inversamente proporzionale alla quarta potenza del raggio  $r$ , cioè

$$R = k \frac{l}{r^4}$$

ove  $k$  è un fattore costante dipendente dalla viscosità del sangue.

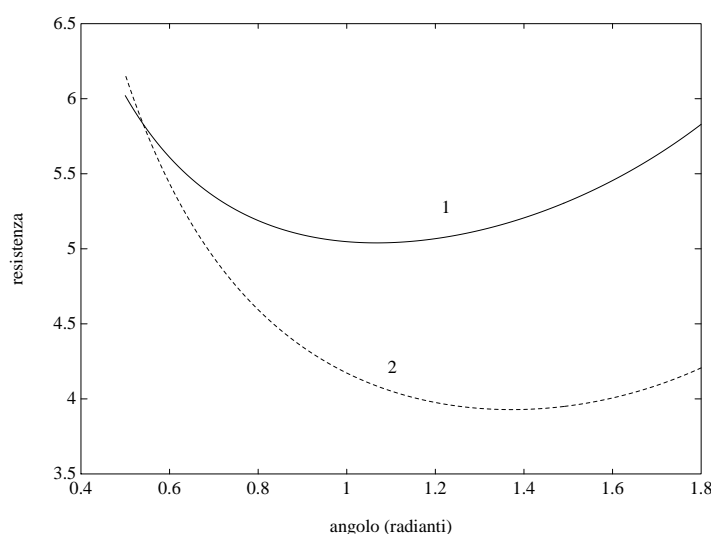


Figura 5.3: Resistenza  $R$  in funzione dell'angolo  $\theta$ . (1)  $r_2 = 1, r_1 = 1.5$ ; (2)  $r_2 = 1, r_1 = 2$ .

Posto  $l_0 = \overline{AB}$ ,  $l_1 = \overline{AD}$ ,  $l_2 = \overline{DC}$ , e  $s = \overline{CB}$ , dal triangolo rettangolo  $BDC$  si ha

$$l_2 = \frac{s}{\sin \theta}, \quad l_0 - l_1 = s \cot \theta$$

e quindi la resistenza  $R$  lungo la traiettoria  $ADC$  è data come funzione dell'angolo  $\theta$  dalla seguente relazione

$$R(\theta) = R_1 + R_2 = k \frac{l_1}{r_1^4} + k \frac{l_2}{r_2^4} = k \left( \frac{l_0 - s \cot \theta}{r_1^4} + \frac{s}{r_2^4 \sin \theta} \right)$$

In Figura 5.3 è rappresentata la funzione  $R(\theta)$  in corrispondenza a due coppie diverse di  $r_1, r_2$  e a  $k = 1, s = 3, l_0 = 5$ . La figura mette in evidenza l'esistenza di un valore ottimale di  $\theta$  (corrispondente al minimo di  $R(\theta)$ ), che aumenta con il diminuire del rapporto  $r_2/r_1$ . In questo caso tale valore può essere ottenuto analiticamente considerando il valore di  $\theta$  che annulla la derivata prima  $R'(\theta)$

$$R'(\theta) = k \left( \frac{s}{r_1^4 \sin^2 \theta} - \frac{s \cos \theta}{r_2^4 \sin^2 \theta} \right)$$

Si ottiene allora facilmente

$$R'(\theta) = 0 \Rightarrow \frac{1}{r_1^4} - \frac{\cos \theta}{r_2^4} = 0$$

da cui, indicando con  $\bar{\theta}$  il valore ottimale di  $\theta$ , si ha

$$\cos \bar{\theta} = \frac{r_2^4}{r_1^4}$$

L'analisi precedente è basata su alcune semplificazioni di natura fisiologica, in particolare l'ipotesi che i vasi siano rigidi e che il flusso del sangue sia laminare. Senza tali semplificazioni il modello matematico diventa più complesso, e per la sua risoluzione diventano essenziali le tecniche numeriche che svilupperemo nel seguito del capitolo. ■

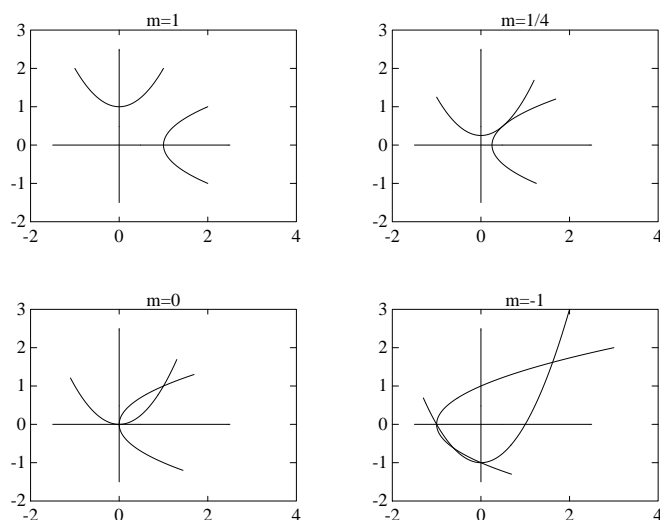


Figura 5.4: Esempi di sistemi non lineari.

► **Esempio 5.3** (*Esempi di sistemi non lineari*) Gli esempi hanno lo scopo di mettere in rilievo la maggiore varietà di situazioni che si possono presentare per un sistema non lineare, rispetto al caso lineare. Siano  $f_1, f_2$  definite da

$$f_1(x_1, x_2) \equiv x_1^2 - x_2 + m, \quad f_2(x_1, x_2) \equiv -x_1 + x_2^2 + m$$

ove  $m$  è un parametro. In corrispondenza a quattro valori diversi di  $m$  si ottengono le situazioni illustrate in Figura 5.4 e corrispondenti a

- (a)  $m = 1$  nessuna soluzione
- (b)  $m = 1/4$  una soluzione:  $x_1 = x_2 = 1/2$
- (c)  $m = 0$  due soluzioni:  $x_1 = x_2 = 0$ ;  $x_1 = x_2 = 1$
- (d)  $m = -1$  quattro soluzioni:  $x_1 = -1, x_2 = 0$ ;  $x_1 = 0, x_2 = -1$ ;  $x_1 = x_2 = (1 \pm \sqrt{5})/2$

D'altra parte il seguente sistema

$$f_1(x_1, x_2) \equiv \frac{1}{2}x_1 \sin\left(\frac{1}{2}\pi x_1\right) - x_2 = 0; \quad f_2(x_1, x_2) \equiv x_2^2 - x_1 + 1 = 0$$

ha un'infinità numerabile di soluzioni (cfr. Figura 5.5), mentre il seguente

$$f_1(x_1, x_2) \equiv x_1^2 - |x_2| = 0; \quad f_2(x_1, x_2) \equiv x_1^2 - x_2 = 0$$

ha un *continuo* di soluzioni. ■

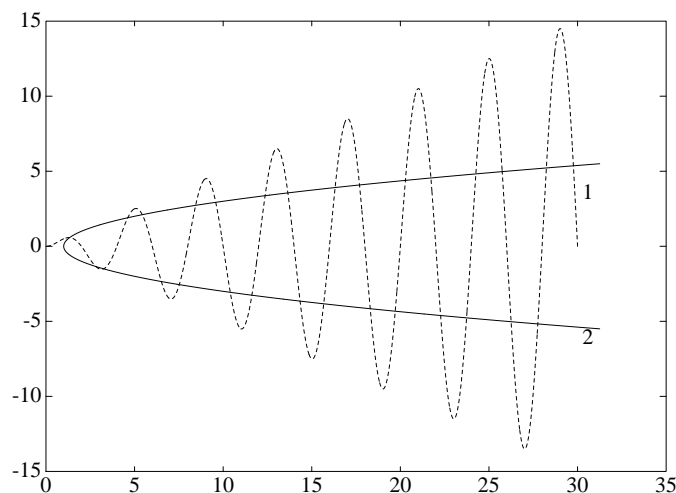


Figura 5.5: Rappresentazione delle soluzioni del sistema non lineare: (1)  $0.5 \sin(0.5\pi x_1) - x_2 = 0$ ; (2)  $x_2^2 - x_1 + 1 = 0$ .

## 5.1 Caso unidimensionale

In questo paragrafo introdurremo ed esamineremo alcuni metodi classici per il calcolo numerico di uno zero (o radice) di una funzione di variabile reale

$$f : I \subset \mathbb{R} \rightarrow \mathbb{R} \quad \boxed{f(x) = 0} \quad (5.6)$$

Alcuni di questi metodi, come il metodo di Newton, possono essere generalizzati, almeno formalmente, senza difficoltà al caso pluridimensionale; altri, come il primo metodo che considereremo (metodo di bisezione), sono essenzialmente ad hoc per il caso unidimensionale. Inoltre, la maggior parte di essi si inquadrano nella teoria generale dei metodi di punto unito, che tratteremo nel seguito.

Il *condizionamento* del problema consistente nella ricerca di una radice della funzione  $f(x)$  è illustrato intuitivamente nella Figura 5.6. In particolare, se la funzione è derivabile, si ha ben condizionamento quando il valore assoluto della derivata calcolata nella radice è grande.



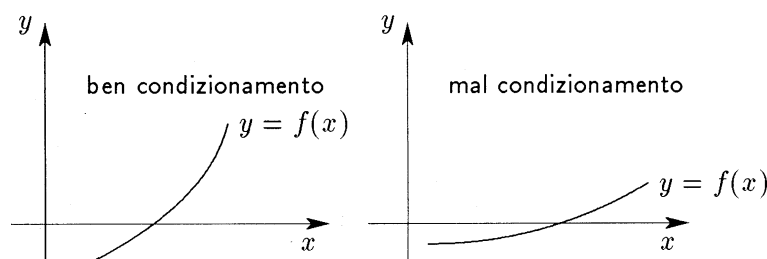


Figura 5.6: Illustrazione del condizionamento corrispondente al problema del calcolo di una radice dell'equazione  $f(x) = 0$ .

### 5.1.1 Metodo di bisezione

Un primo passo nell'approssimazione di una radice della funzione  $f(x)$  è la sua *localizzazione*, ossia la determinazione di un intervallo contenente la radice. A tale scopo, un metodo opportuno consiste in una *tabulazione* (in particolare, sotto forma di grafico) della funzione  $f(x)$ . Il passo di tabulazione deve essere, comunque, scelto opportunamente, in quanto un passo troppo grande può *nascondere* eventuali radici, mentre al contrario un passo troppo piccolo, in rapporto alla precisione utilizzata, può, per effetto degli errori di arrotondamento, individuare *false* radici. La situazione è illustrata nella Figura 5.7, in cui è rappresentato il grafico della funzione  $y = \cos x - \cos(3.1x)$  e la radice da localizzare è la più piccola radice strettamente positiva.

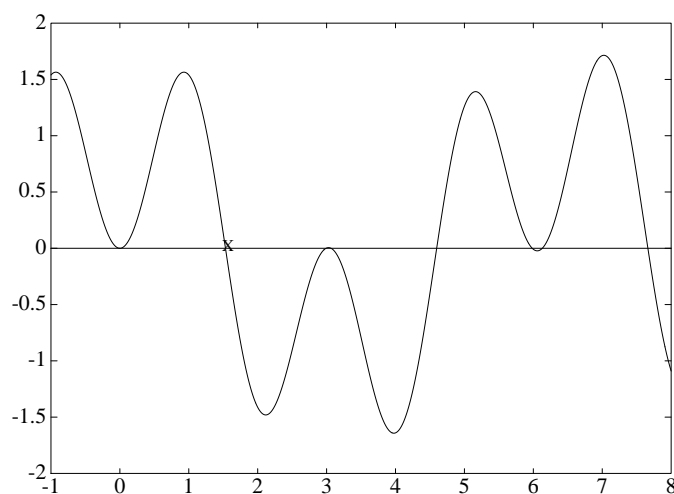


Figura 5.7: Grafico della funzione  $y = \cos x - \cos(3.1x)$  sull'intervallo  $[-1, 8]$ .

Una volta localizzato un intervallo  $[a, b]$  contenente una radice, la stima ottenuta può essere migliorata procedendo a suddivisioni successive dell'intervallo. Un criterio per la scelta ad ogni suddivisione dell'intervallo opportuno, ossia dell'intervallo che contiene ancora la radice, è fornito dal seguente risultato sulle radici di una funzione continua.

**Proposizione 5.1** *Se  $f(x)$  è una funzione continua sull'intervallo limitato e chiuso  $[a, b]$  e  $f(a) \cdot f(b) < 0$ , allora esiste almeno una radice di  $f(x)$  nell'intervallo  $[a, b]$ .*

Ne deriva il seguente *algoritmo*, illustrato in Figura 5.8.

**Algoritmo 5.1** (Algoritmo di bisezione) *Sia  $f(x)$  una funzione continua sull'intervallo limitato e chiuso  $[a, b]$  con  $f(a) \cdot f(b) < 0$ . L'algoritmo genera una successione di intervalli  $(a_k, b_k)$  con  $f(a_k) \cdot f(b_k) < 0$  e con  $[a_k, b_k] \subset [a_{k-1}, b_{k-1}]$  e  $|b_k - a_k| = \frac{1}{2}|b_{k-1} - a_{k-1}|$ . Date due tolleranze  $\epsilon_1, \epsilon_2$ , l'algoritmo si arresta o quando  $|b_k - a_k| \leq \epsilon_1$  o quando  $|f(\frac{1}{2}(a_k + b_k))| \leq \epsilon_2$  o infine quando  $k > nmax$ , ove  $nmax$  è un numero massimo di iterazioni fissato.*

```

a0 = a; b0 = b
for k = 0, 1, ..., nmax
  if bk - ak ≤ ε1
    then stop 1      (test sulla radice)
  else
    set mk+1 =  $\frac{1}{2}(a_k + b_k)$ 
    calcola f(mk+1)
    if |f(mk+1)| ≤ ε2      (test sulla funzione)
      then stop 2
    else
      if f(ak) · f(mk+1) < 0
        then set bk+1 = mk+1; ak+1 = ak
      else set ak+1 = mk+1; bk+1 = bk
      end if
    end if
  end if
end for

```

Nel caso in cui  $0 \notin [a, b]$ , è possibile adottare un criterio di arresto sull'errore relativo mediante la seguente condizione

$$\frac{b_k - a_k}{\min(|a_k|, |b_k|)} < \epsilon_1$$

Se si assume come stima della soluzione il punto di mezzo  $m_k$ , si ottiene ad ogni iterazione *una cifra binaria* della soluzione. Ad esempio per un numero floating point a 32-bit con una mantissa a 24-bit, se l'intervallo iniziale è  $[1, 2]$ , si ottiene l'accuratezza completa in 24 iterazioni. Poiché ogni iterazione richiede una valutazione della funzione, sono richieste al più 24 valutazioni. Si osservi che se la funzione

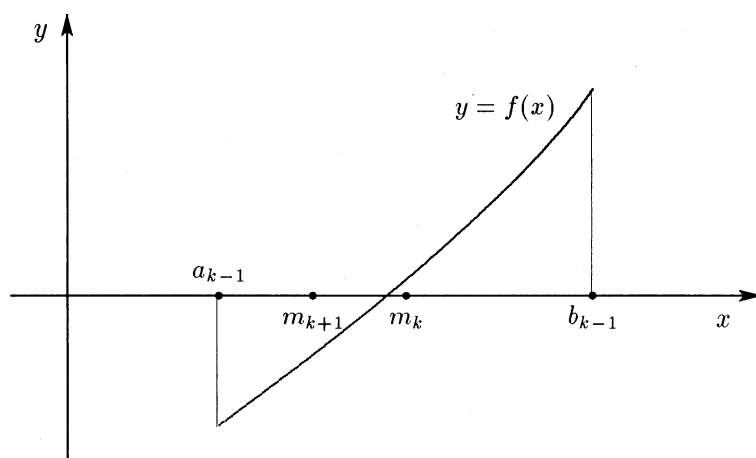


Figura 5.8: Metodo di bisezione.

fosse calcolata in corrispondenza a tutti i numeri macchina contenuti nell'intervallo, allora sarebbero necessarie  $2^{24} = 16777216$  valutazioni.

Indicando con  $e_k$  l'errore all'iterazione  $k$ , si ha

$$\frac{|e_{k+1}|}{|e_k|} \approx \frac{1}{2}$$

Si dice allora che la convergenza del metodo di bisezione è di tipo *lineare*; il numero  $C = \frac{1}{2}$  è detto *costante d'errore*. Di seguito è riportato un esempio di implementazione dell'algoritmo della bisezione.

```

FUNCTION BISEC(FUN,X1,X2,EPS)
C usando il metodo di bisezione calcola la radice di una funzione FUN,
C compresa nell'intervallo [X1, X2], con una precisione EPS.
C IMAX numero massimo di iterazioni.
  PARAMETER (IMAX=50)
  EXTERNAL FUN
  FMID=FUN(X2)
  F=FUN(X1)
  IF(F*FMID.GE.0.) THEN
    WRITE(*,*)'Segno concorde nei due estremi'
    RETURN
  ENDIF
  IF(F.LT.0.)THEN
    BISEC=X1
    DX=X2-X1
  ELSE
    BISEC=X2
    DX=X1-X2
  ENDIF

```

```

DO 10 J=1,IMAX
  DX=DX*.5
  XMID=BISEC+DX
  FMID=FUN(XMID)
  IF(FMID.LE.0.)BISEC=XMID
  IF(ABS(DX).LT.EPS .OR. FMID.EQ.0.) RETURN
10 CONTINUE
WRITE(*,*) 'superato il numero massimo di iterazioni'
END

```

Osserviamo che nell'implementazione precedente il punto di bisezione è calcolato come  $c = a + (b - a) * 0.5$ , anziché nella maniera convenzionale  $c = 0.5 * (a + b)$ . In effetti, la prima formula è preferibile quando si opera in aritmetica in floating point, come illustrato dal seguente esempio. Sia  $a = 0.982$ ,  $b = 0.984$  e si supponga di operare in un'aritmetica a tre cifre con arrotondamento. Allora,  $a + b$  è calcolato come 1.97 e la formula convenzionale fornisce il valore 0.985, che è un punto esterno all'intervallo  $[a, b]$ .

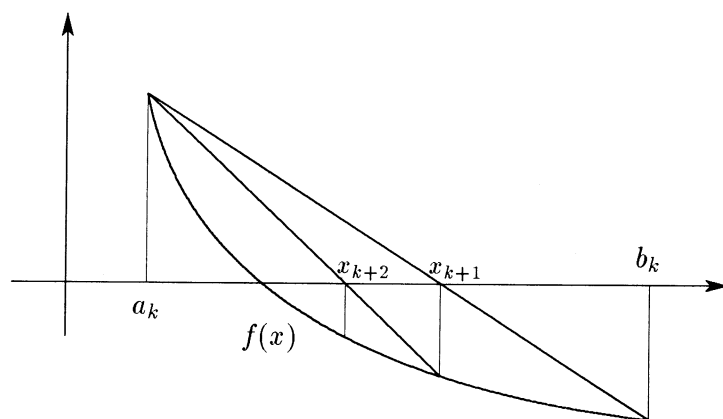


Figura 5.9: Illustrazione del metodo *regula falsi*.

### 5.1.2 Metodo regula falsi

Nel metodo di bisezione si utilizza, in sostanza, solo il segno della funzione. Un primo modo di utilizzare anche i *valori* della funzione è rappresentato dal cosiddetto *metodo regula falsi*, nel quale si considera come approssimazione della funzione nell'intervallo  $[a_k, b_k]$  la *retta che interpola i punti*  $(a_k, f(a_k))$ ,  $(b_k, f(b_k))$ , anziché, come nel metodo della bisezione, i punti  $(a_k, \text{sign } f(a_k))$ ,  $(b_k, \text{sign } f(b_k))$ . Si ottiene allora, come nuova approssimazione, lo zero di tale retta, cioè il valore

$$x_{k+1} = \frac{a_k f(b_k) - b_k f(a_k)}{f(b_k) - f(a_k)}$$

Procedendo poi come nell'algoritmo della bisezione, si genera ad ogni iterazione un intervallo che contiene, nelle stesse ipotesi sulla funzione, una radice della  $f$ . Tuttavia, diversamente che nel metodo di bisezione, non si ha in generale  $b_k - a_k \rightarrow 0$  per  $k \rightarrow \infty$  (si veda, per una esemplificazione la Figura 5.9).

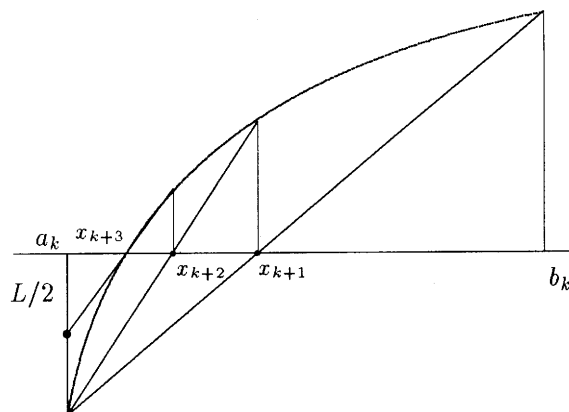


Figura 5.10: Metodo regula falsi modificato:  $L$  è dimezzato prima di calcolare  $x_{k+3}$ .

Un modo per eliminare l'eventualità che uno degli estremi dell'intervallo contenente la radice rimanga fisso durante le successive iterazioni, è implementato nel seguente algoritmo e illustrato in Figura 5.10. Tale metodo è anche noto come *metodo Illinois*. Un'altra variante del metodo regula falsi, nota come *metodo delle secanti*, verrà considerata nel seguito.

**Algoritmo 5.2** (Algoritmo regula falsi modificato) *Sia  $f(x)$  una funzione continua sull'intervallo limitato e chiuso  $[a, b]$  con  $f(a) \cdot f(b) < 0$ . L'algoritmo genera una successione di intervalli  $(a_k, b_k)$  con  $f(a_k) \cdot f(b_k) < 0$  e  $[a_k, b_k] \subset [a_{k-1}, b_{k-1}]$ . Date due tolleranze  $\epsilon_1, \epsilon_2$ , l'algoritmo si arresta o quando  $|b_k - a_k| \leq \epsilon_1$  o quando il modulo della funzione è  $\leq \epsilon_2$  o infine quando  $k > nmax$ , ove  $nmax$  è un numero massimo fissato di iterazioni.*

```

 $a_0 = a; b_0 = b$ 
 $L = f(a_0); R = f(b_0); x_0 = a_0$ 
for  $k = 0, 1, \dots, nmax$ 
  if  $b_k - a_k \leq \epsilon_1$ 
    then stop 1
  else
    set  $x_{k+1} = \frac{a_k R - b_k L}{R - L}$ 
    calcola  $f(x_{k+1})$ 
    if  $|f(x_{k+1})| \leq \epsilon_2$ 
      then stop 2
    else
      if  $f(a_k) \cdot f(x_{k+1}) < 0$ 

```

```

then set  $b_{k+1} = x_{k+1}$ ;  $a_{k+1} = a_k$ ;  $R = f(x_{k+1})$ 
  if  $f(x_k) \cdot f(x_{k+1}) > 0$  then  $L = L/2$ 
else set  $a_{k+1} = x_{k+1}$ ;  $b_{k+1} = b_k$ ;  $L = f(x_{k+1})$ 
  if  $f(x_k) \cdot f(x_{k+1}) > 0$  then  $R = R/2$ 
end if
end if
end if
end for

```

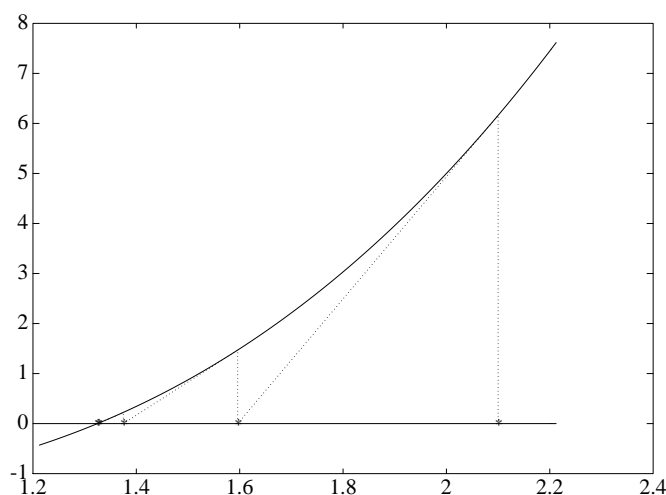


Figura 5.11: Illustrazione del metodo di Newton; calcolo della soluzione dell'equazione  $x^3 - x - 1 = 0$  nell'intervallo  $[1.2, 2]$ .

### 5.1.3 Metodo di Newton

Quando la funzione  $f(x)$  è derivabile, la derivata  $f'(x)$  può essere opportunamente utilizzata per *adattare* la ricerca della radice alla particolare funzione considerata e ottenere quindi una convergenza più rapida. Questo viene fatto, in particolare, nel *metodo di Newton*<sup>1</sup> che utilizza la seguente idea *geometrica*, illustrata in Figura 5.11 nel caso di  $f(x) = x^3 - x - 1$ .

<sup>1</sup>Il metodo è anche noto come metodo di Newton-Raphson. Newton introduce l'idea in *De analysi per aequationes numero terminorum infinitas* (1666-1669) per risolvere l'equazione:  $f(y) = y^3 - 2y - 5 = 0$ : *Sit  $y^3 - 2y - 5 = 0$ , resolvenda: Et sit 2, numerus quam decima sui parte differt a Radice quaesita. Tum pono  $2 + p = y$  & substituo hunc ipsi valorem in Aequationem, & inde nova prodit  $p^3 + 6p^2 + 10p - 1 = 0$ , cujus Radix  $p$  exquirenda est, ut Quotienti addatur: Nempe (neglectis  $p^3 + 6p^2$  ob parvitatem)  $10p - 1 = 0$ , sive  $p = 0.1$  prope veritatem est; itaque scribo 0,1 in Quotiente, & suppono  $0,1 + q = p$ , & hunc ejus valorem, ut prius substituo... L'idea è resa "più algoritmica" da Raphson in *Analysis Aequationum Universalis* (1690). Un'estensione del metodo è considerata da Fourier in *Question d'analyse Algebrique* (1818).*

A partire da una *stima iniziale*  $x_0$  della radice, si genera una successione di valori approssimati  $\{x_k\}$ , ove, supposto noto  $x_k$  il successivo valore  $x_{k+1}$  è ottenuto come l'intersezione con l'asse  $x$  della *tangente* nel punto  $(x_k, f(x_k))$ . In altre parole, all'equazione  $f(x) = 0$  si sostituisce per  $k = 0, 1, \dots$  la seguente equazione *lineare*

$$P_1(x) := f(x_k) + (x - x_k)f'(x_k) = 0 \quad (5.7)$$

dalla quale, se  $f'(x_k) \neq 0$  (ossia la tangente non è parallela all'asse  $x$ ) si ricava

$$x_{k+1} = x_k + h_k, \quad h_k = -\frac{f(x_k)}{f'(x_k)} \quad (5.8)$$

Il metodo di Newton utilizza quindi l'idea della *interpolazione* (della funzione  $f(x)$  mediante la tangente) combinata con l'idea della *iterazione*. Un modo equivalente di interpretare il metodo, più opportuno per la sua estensione alla risoluzione di sistemi di equazioni, consiste nel considerare il polinomio  $P_1(x)$  definito in (5.7) come il termine *lineare* nello sviluppo in serie della funzione  $f(x)$  nel punto  $x_k$ . Più in generale, ricordiamo che se la funzione  $f(x)$  è sufficientemente regolare, si ha

$$f(x_k + h_k) = f(x_k) + h_k f'(x_k) + \frac{h_k^2}{2} f''(x_k) + \dots + \frac{h_k^r}{r!} f^{(r)}(x_k + \theta h_k) \quad (5.9)$$

ove  $\theta$  è un opportuno valore nell'intervallo  $(0, 1)$ . A partire da tale risultato, si possono considerare per  $r > 1$  delle opportune *generalizzazioni* del metodo di Newton, ottenute assumendo sviluppi di grado più elevato. Dal punto di vista geometrico, significa approssimare la curva  $f(x)$  con un polinomio che nel punto  $(x_k, f(x_k))$  assume gli stessi valori della funzione insieme a quelli delle prime  $r$  derivate. Il metodo così ottenuto presenta, in generale, una migliore approssimazione locale, ma ha l'inconveniente di richiedere il calcolo delle derivate successive di  $f(x)$  e la risoluzione ad ogni passo di una equazione algebrica di grado  $r$  in  $h_k$ . Nel paragrafo precedente

k	$x_0 = 1.4$		$x_0 = 1.2$	
	$x_k$	$f(x_k)$	$x_k$	$f(x_k)$
1	-0.14136186D+01	-0.955D+00	-0.93758164D+00	-0.753D+00
2	0.14501293D+01	0.967D+00	0.47771595D+00	0.446D+00
3	-0.15506259D+01	-0.998D+00	-0.69651670D-01	-0.695D-01
4	0.18470540D+01	0.107D+01	0.22505187D-03	0.225D-03
5	-0.28935623D+01	-0.124D+01	-0.75990038D-11	-0.760D-11
6	0.87103258D+01	0.146D+01	0.00000000D+00	0.000D+00

Tabella 5.1: Risultati ottenuti con il metodo di Newton applicato all'equazione  $\arctan(x) = 0$ , a partire da due diversi valori iniziali  $x_0$ .

abbiamo visto che il metodo della bisezione, quando la funzione  $f(x)$  è continua ed è trascurato l'effetto degli errori di arrotondamento, è un metodo convergente, anche se il numero di valutazioni della funzione richiesto per ottenere una stima

$i$	$V^i$	$F(V^i)$	$\bar{V}^i$	$F(\bar{V}^i)$
0	0.930013	-0.2748	0.150000	0.2225
1	0.557331	-0.0509	0.179165	0.0775
2	0.429817	-0.0143	0.204716	0.0244
3	0.340674	-0.0082	0.222965	0.0061
4	1.076273	-0.3873	0.231636	0.0008
5	0.590329	-0.0649	0.233358	2.89 E-5
6	0.445398	-0.0171	0.233418	3.38 E-8
7	0.357114	-0.0082	0.233418	5. E-14
8	-0.262689	2.5926		
9	-0.743173	2.0979		
...				
22	0.232611	0.0003		
23	0.233405	6.05 E-6		
24	0.233418	1.47 E-9		

Tabella 5.2: Risultati ottenuti mediante il metodo di Newton in corrispondenza al gas isobutano.

prefissata può essere ritenuto in alcune applicazioni eccessivamente alto. Il metodo di Newton, introdotto per migliorare la velocità di convergenza, può al contrario presentare problemi per quanto riguarda la convergenza. Come semplice esemplificazione, consideriamo l'equazione  $f(x) := \arctan(x) = 0$ , che ammette ovviamente la soluzione  $x = 0$ . In Tabella 5.1 sono riportati i risultati ottenuti con il metodo di Newton a partire da due diversi valori  $x_0$ . Si vede allora che per  $x_0 = 1.4$  i valori  $x_k$  si allontanano, in maniera alternata, dalla soluzione esatta, mentre per  $x_0 = 1.2$  la successione presenta una rapida convergenza (come **esercizio**, si confronti tali risultati con quelli che possono essere ottenuti con il metodo della bisezione). Come ulteriore esempio, consideriamo la risoluzione dell'equazione (5.5) introdotta nell'Esempio 5.1. La questione importante della *convergenza* del metodo verrà esaminata in maniera più approfondita nei paragrafi successivi.

► **Esempio 5.1** (*Continuazione*). Indicando con  $\{V^i\}$ ,  $i = 0, 1, \dots$  la successione generata dal metodo di Newton, applicato all'equazione (5.5), si ha

$$V^{i+1} = V^i - \frac{(RT + \beta/V^i + \gamma/(V^i)^2 + \delta/(V^i)^3)/P - V^i}{-(\beta/(V^i)^2 + 2\gamma/(V^i)^3 + 3\gamma/(V^i)^4)/P - 1}$$

Nella Tabella 5.2 sono riportati i risultati ottenuti in corrispondenza a due differenti valori iniziali  $V^0$ . In particolare il valore  $V^0 \approx 0.930013$  corrisponde al valore fornito dall'equazione ideale dei gas. Dai risultati ottenuti appare ancora chiara l'importanza, per la convergenza del metodo di Newton, di una *buona* scelta del valore iniziale.

La seguente quantità, nota come *fattore di comprimibilità*

$$z = \frac{PV}{RT}$$

fornisce un indice della discrepanza del comportamento del gas reale da quello previsto dalla legge ideale dei gas (per la quale si ha  $z = 1$ ). Utilizzando il risultato precedentemente



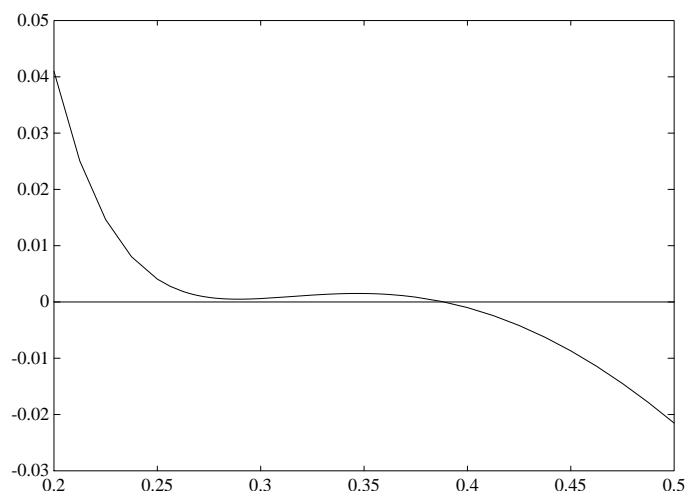


Figura 5.12: Grafico della funzione  $F(V)$  in corrispondenza al gas butano e  $R = 0.0287$ .

ottenuto per il volume  $V$ , si ha nel caso dell'isobutano

$$z = \frac{PV}{RT} = \frac{36 \times 0.233418}{0.08206 \times 408} \approx 0.25$$

Terminiamo analizzando un aspetto importante dei modelli matematici, che riguarda la *sensitività* delle quantità cercate (l'output del sistema studiato) rispetto alle quantità assegnate (l'input). Consideriamo, in particolare e come esemplificazione, la sensitività del volume  $V$  rispetto alla costante dei gas  $R$ . In Figura 5.12 è riportato il grafico della funzione  $F(V)$  corrispondente a  $R = 0.0827$  con un errore relativo, rispetto al valore  $R = 0.08206$ , di 0.77%. Applicando il metodo di Newton, con  $V^0 = RT/P$ , si ottengono i seguenti risultati.

$i$	$V^i$	$F(V^i)$
0	0.937266	-0.27171961
1	0.571142	-0.04755548
2	0.458686	-0.01055762
3	0.411643	-0.00231422
4	0.393159	-0.00036438
5	0.388918	-0.00001878
6	0.388674	-0.00000006

Sul valore ottenuto di  $V$  si ha, pertanto, rispetto al valore precedente, un errore relativo del 66%. Per il fattore di comprimibilità si ha  $z \approx 0.41$ . La Figura 5.13 evidenzia la sensitività della soluzione  $V$  rispetto alle variazioni della variabile  $P$ . ■

#### 5.1.4 Metodo di Newton in più dimensioni

Sia  $\mathbf{f}(\mathbf{x}) = [f_1(x), f_2(x), \dots, f_n(x)]^T$ ,  $\mathbf{x} \in \mathbb{R}^n$ , una funzione sufficientemente regolare. Il metodo di Newton introdotto nel paragrafo precedente nel caso scalare può

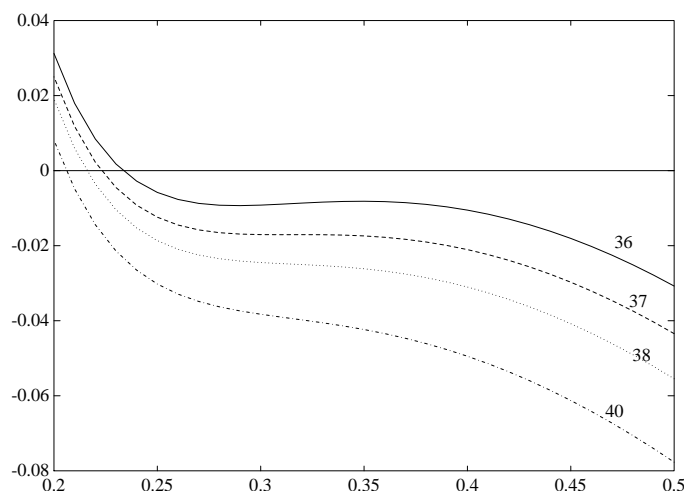


Figura 5.13: Grafico della funzione  $F(V)$  in corrispondenza ai successivi valori di  $P = 36, 37, 38, 40$ .

essere esteso alla risoluzione del sistema  $\mathbf{f}(\mathbf{x}) = 0$ , utilizzando il seguente sviluppo in serie

$$\mathbf{f}(\mathbf{x}^k + \mathbf{h}^k) = \mathbf{f}(\mathbf{x}^k) + \mathbf{J}(\mathbf{x}^k) \mathbf{h}^k + O(\|\mathbf{h}^k\|^2) \quad (5.10)$$

ove  $\mathbf{h}^k \in \mathbb{R}^n$  e  $\mathbf{J}$  è la *matrice jacobiana*, ossia la matrice

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Nell'ipotesi che la matrice  $\mathbf{J}$  sia non singolare, trascurando nello sviluppo (5.10) i termini di secondo grado, ossia *linearizzando* la funzione  $\mathbf{f}(\mathbf{x})$ , si ottiene a partire da una stima iniziale  $\mathbf{x}^0$  la seguente procedura, nota come *metodo di Newton-Raphson* per la risoluzione di un sistema di equazioni non lineari

$$\mathbf{J}(\mathbf{x}^k) \mathbf{h}^k = -\mathbf{f}(\mathbf{x}^k) \quad (5.11)$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{h}^k, \quad k = 0, 1, \dots \quad (5.12)$$

La costruzione del vettore  $\mathbf{x}^{k+1}$  richiede quindi la risoluzione del sistema lineare (5.11). Il seguente esempio illustra la procedura nel caso di un sistema di due equazioni.

► **Esempio 5.4** (*Metodo di Newton applicato a un sistema di due equazioni*) Nel caso di un sistema di due equazioni

$$\begin{cases} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{cases}$$

il metodo di Newton-Raphson comporta la risoluzione ad ogni iterazione del seguente sistema lineare

$$\begin{cases} \frac{\partial f_1(\mathbf{x}^k)}{\partial x_1} h_1^k + \frac{\partial f_1(\mathbf{x}^k)}{\partial x_2} h_2^k = -f_1(\mathbf{x}^k) \\ \frac{\partial f_2(\mathbf{x}^k)}{\partial x_1} h_1^k + \frac{\partial f_2(\mathbf{x}^k)}{\partial x_2} h_2^k = -f_2(\mathbf{x}^k) \end{cases}$$

Come applicazione, consideriamo la risoluzione del seguente sistema

$$\begin{cases} 2x_1 - \cos x_2 = 0 \\ 2x_2 - \sin x_1 = 0 \end{cases}$$

la cui soluzione è l'intersezione delle due curve  $2x_1 - \cos x_2 = 0$  e  $2x_2 - \sin x_1 = 0$  rappresentate in Figura 5.14.

k	$x_1^k$	$x_2^k$	$f_1$	$f_2$
0	0.7000	0.7000	0.6352	0.7558
1	0.5256	0.2554	0.0837	0.0091
2	0.4865	0.2340	0.00022	0.00037
3	0.4864	0.2337	0.2475 E-7	0.0171 E-7

Tabella 5.3: Metodo di Newton applicato al sistema  $2x_1 - \cos x_2 = 0$ ;  $2x_2 - \sin x_1 = 0$ .

In Tabella 5.3 sono riportati i risultati ottenuti con il seguente programma implementato in linguaggio MATLAB.

```
x=[0.7 0.7]'; % stima iniziale
for i=1:kmax
    J=[ 2          sin(x(2))
        -cos(x(1))  2      ]; % matrice jacobiana
    f=[2*x(1)-cos(x(2))
        2*x(2)-sin(x(1))]; % valore della funzione
    x=x-inv(J)*f % aggiornamento della soluzione
end
```

Le soluzioni sono rappresentate per  $k = 0, 1, 2$  in Figura 5.14. ■

### 5.1.5 Studio della convergenza del metodo di Newton

Sia  $f(x)$ ,  $x \in \mathbb{R}$ , una funzione *continua* con le derivate del primo e del secondo ordine in un intervallo contenente una *radice semplice*  $\alpha$ . Allora esiste un intorno  $I$  di  $\alpha$  in cui  $f'(x) \neq 0$ . Posto

$$e_k := x_k - \alpha$$

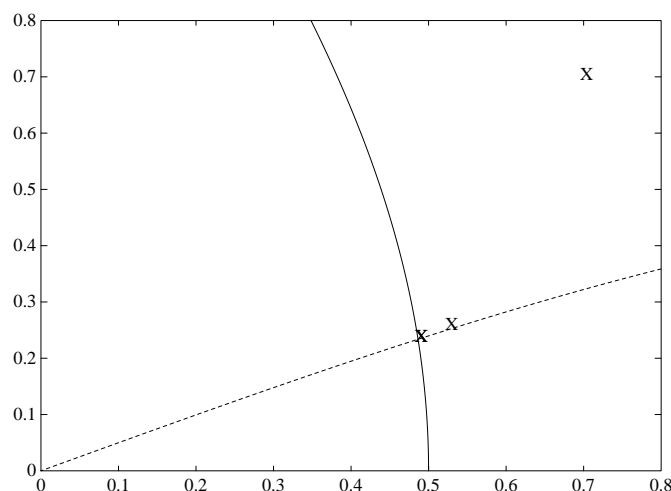


Figura 5.14: Metodo di Newton applicato al sistema  $2x_1 - \cos x_2 = 0$ ;  $2x_2 - \sin x_1 = 0$  La soluzione approssimata è indicata con  $X$ .

si ha, sviluppando la funzione  $f(x)$  in serie di Taylor intorno al punto  $x_k$

$$0 = f(\alpha) = f(x_k) + (\alpha - x_k)f'(x_k) + \frac{1}{2}(\alpha - x_k)^2 f''(\xi), \quad \xi \in (x_k, \alpha)$$

da cui, dividendo per  $f'(x_k)$

$$\frac{f(x_k)}{f'(x_k)} + \alpha - x_k = \alpha - x_{k+1} = \frac{-\frac{1}{2}(\alpha - x_k)^2 f''(\xi)}{f'(x_k)}$$

Si ha pertanto la relazione

$$e_{k+1} = \frac{1}{2} e_k^2 \frac{f''(\xi)}{f'(x_k)} \Rightarrow \boxed{\frac{e_{k+1}}{e_k^2} \rightarrow \frac{1}{2} \frac{f''(\alpha)}{f'(\alpha)}} \text{ per } x_k \rightarrow \alpha \quad (5.13)$$

Dal momento che l'errore al passo  $(k+1)$ -mo è proporzionale al quadrato dell'errore al passo  $k$ -mo, si dice che *il metodo di Newton ha una convergenza del secondo ordine*.

Più in generale si introduce la nozione di *ordine di convergenza* nel seguente modo.

**Definizione 5.1** Sia  $\{x_k\}$ ,  $k = 0, 1, \dots$  una successione convergente ad  $\alpha$  e sia  $e_k = x_k - \alpha$  l'errore commesso al passo  $k$ . Se esiste un numero  $p > 0$  e una costante  $C \neq 0$  tale che

$$\boxed{\lim_{n \rightarrow \infty} \frac{|e_{k+1}|}{|e_k|^p} = C}$$

allora,  $p$  è chiamato *ordine di convergenza della successione* e  $C$  è la costante d'errore.

Ad esempio, per  $p = 1, 2, 3$  si parla, rispettivamente, di convergenza lineare, quadratica e cubica. Per  $p > 1$ , si dice anche che la convergenza è *superlineare*. Ad ogni iterazione, le cifre esatte vengono approssimativamente raddoppiate per un metodo del secondo ordine e triplicate per uno del terzo ordine. È, comunque, opportuno osservare che l'ordine è una indicazione asintotica, e quindi nella valutazione pratica della rapidità di convergenza di un metodo è necessario tenere conto anche della costante di errore  $C$ . In altre parole, può accadere che un metodo lineare con costante di errore sufficientemente piccola dia nelle prime iterazioni risultati migliori di un metodo quadratico caratterizzato da una costante di errore più grande.

Ritornando allo studio della convergenza del metodo di Newton, indicata con  $M$  una costante tale che

$$M \geq \frac{1}{2} \left| \frac{f''(y)}{f'(x)} \right|, \quad \forall x, y \in I$$

dalla (5.13) si ha

$$|e_{k+1}| \leq M e_k^2 \Rightarrow |M e_{k+1}| \leq (M e_k)^2$$

da cui, per ricorrenza

$$|e_k| \leq \frac{1}{M} |M e_0|^{2^k} \quad (5.14)$$

Si vede pertanto che il metodo di Newton è convergente se il valore iniziale  $x_0$  è scelto sufficientemente vicino alla radice  $\alpha$ , più precisamente tale da avere  $|M e_0| = |M(x_0 - \alpha)| < 1$ .

Il risultato di convergenza ora stabilito viene detto di tipo *locale*, in quanto assicura la convergenza del metodo quando il valore iniziale è scelto convenientemente. In un risultato di tipo *globale*, invece, la convergenza è assicurata per tutti i valori  $x_0$  in un intorno noto a priori della soluzione. Un esempio di risultato globale è presentato nel seguente teorema.

**Teorema 5.1** Sia  $f(x) \in C^2([a, b])$ , con  $[a, b]$  intervallo chiuso e limitato. Se

1.  $f(a) \cdot f(b) < 0$
2.  $f'(x) \neq 0, \forall x \in [a, b]$
3.  $f''(x) \geq 0$ , oppure  $f''(x) \leq 0, \forall x \in [a, b]$
4.  $\left| \frac{f(a)}{f'(a)} \right| < b - a; \quad \left| \frac{f(b)}{f'(b)} \right| < b - a$

allora la funzione  $f(x)$  ha un'unica radice  $\alpha$  in  $[a, b]$  e il metodo di Newton converge ad  $\alpha$  per ogni scelta di  $x_0 \in [a, b]$ .

**DIMOSTRAZIONE.** Rileviamo che la condizione 4 significa che la tangente negli estremi dell'intervallo interseca l'asse  $x$  all'interno dell'intervallo  $[a, b]$ . Diamo un'idea della dimostrazione in *una* delle possibili quattro situazioni. Supponiamo cioè che

$$f' > 0, f'' \leq 0, f(a) < 0, f(b) > 0 \iff \text{img}$$

Sia  $x_0$  tale che  $a \leq x_0 < \alpha$  e quindi  $f(x_0) \leq 0$ . In caso contrario, si vede facilmente che  $x_1$  si trova in tale situazione. Si ha allora

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \geq x_0$$

Dimostriamo per induzione che  $x_k \leq \alpha$  e  $x_{k+1} \geq x_k$ . Poiché la proprietà è vera per  $k = 0$ , dimostriamo che, supposta vera per  $k$ , essa è verificata per  $k + 1$ . Si ha, infatti

$$-f(x_k) = f(\alpha) - f(x_k) = (\alpha - x_k)f'(x_k^*), \quad x_k \leq x_k^* \leq \alpha$$

Dal momento che  $f''(x) \leq 0$ , si ha che  $f'$  è decrescente e quindi  $f'(x_k^*) \leq f'(x_k)$ , da cui

$$\begin{aligned} -f(x_k) &\leq (\alpha - x_k)f'(x_k) \\ &\Downarrow \\ x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)} \leq x_k + (\alpha - x_k) = \alpha \end{aligned}$$

Di conseguenza

$$f(x_{k+1}) \leq 0 \Rightarrow x_{k+2} = x_{k+1} - \frac{f(x_{k+1})}{f'(x_{k+1})} \geq x_{k+1}$$

Per terminare la dimostrazione, basta osservare che la successione  $\{x_k\}$ , essendo monotona e limitata, è convergente; passando al limite nella formula iterativa si trova che tale limite è la radice  $\alpha$ . ■

È opportuno osservare che le condizioni di convergenza analizzate in questo paragrafo sono condizioni soltanto sufficienti; in altre parole, per particolari problemi il metodo può convergere anche in condizioni di regolarità meno restrittive. Come illustrazione, nell'esempio successivo sono esaminate alcune funzioni per le quali la funzione non è derivabile nella radice.

► **Esempio 5.5** La seguente funzione

$$f(x) = \begin{cases} \sqrt{x} & \text{per } x \geq 0 \\ -\sqrt{-x} & \text{per } x < 0 \end{cases}$$

non ammette derivata finita nella radice dell'equazione  $f(x) = 0$ . Il metodo di Newton corrisponde all'iterazione

$$x_{k+1} = -x_k$$

che fornisce la soluzione soltanto se  $x_0 = 0$ .

D'altra parte, sia

$$f(x) = \begin{cases} \sqrt[3]{x^2} & \text{per } x \geq 0 \\ -\sqrt[3]{x^2} & \text{per } x < 0 \end{cases}$$

Si ottiene, in questo caso

$$x_{k+1} = -\frac{1}{2}x_k$$

e quindi convergenza.

Consideriamo infine la funzione

$$f(x) = \begin{cases} \sqrt[3]{x} & \text{per } x \geq 0 \\ -\sqrt[3]{-x} & \text{per } x < 0 \end{cases}$$

Si ha

$$x_{k+1} = -2x_k$$

e quindi divergenza. ■

### 5.1.6 Metodo di Newton e radici multiple

Se  $\alpha$  è una radice multipla, e quindi in particolare si ha  $f'(\alpha) = 0$ , il metodo di Newton può essere ancora convergente, ma la sua rapidità di convergenza diminuisce. Illustriamo tale fatto, intuitivo pensando al significato geometrico del metodo, mediante un semplice esempio. Nel caso della funzione  $f(x) = x^2$ , il metodo si riduce alla seguente iterazione

$$x_{k+1} = x_k - \frac{x_k^2}{2x_k} \Rightarrow x_{k+1} = \frac{x_k}{2} \Rightarrow e_{k+1} = \frac{1}{2}e_k$$

da cui si vede che il metodo ha convergenza lineare. È comunque interessante osservare che il metodo può essere opportunamente modificato in maniera da migliorare la rapidità di convergenza; si ha infatti

$$x_{k+1} = x_k - \boxed{2} \frac{f(x_k)}{f'(x_k)} \Rightarrow x_{k+1} = x_k - 2 \frac{x_k^2}{2x_k} = 0$$

Il fattore moltiplicativo introdotto corrisponde alla molteplicità della radice  $\alpha = 0$ . L'idea può essere estesa al caso generale nel seguente modo. Sia  $f(x)$  una funzione sufficientemente regolare con

$$f(\alpha) = f'(\alpha) = \dots = f^{(p-1)}(\alpha) = 0, \quad f^{(p)}(\alpha) \neq 0$$

Si può allora dimostrare che la seguente modifica del metodo

$$g(x) = x - \boxed{p} \frac{f(x)}{f'(x)}$$

fornisce ancora un metodo del secondo ordine. Naturalmente, tale modifica può essere fatta solo se è nota a priori la *molteplicità* della radice. Nel caso generale, quando si accerta la presenza di una radice con molteplicità maggiore di uno, si può considerare la funzione

$$\Psi(x) = \frac{f(x)}{f'(x)} \Rightarrow \Psi(x) = \frac{(x - \alpha)h(x)}{ph(x) + (x - \alpha)h'(x)}, \quad \Psi'(\alpha) = \frac{1}{p} \neq 0$$

ed applicare il metodo di Newton alla nuova funzione. Il metodo comporta comunque un costo superiore, dal momento che utilizza i valori delle funzioni  $f, f', f''$ .

### 5.1.7 Alcune applicazioni del metodo di Newton

In questo paragrafo forniremo alcuni esempi di applicazione del metodo di Newton con la corrispondente analisi della convergenza.

► **Esempio 5.6** Sia  $c$  un numero reale positivo dato e

$$f(x) = x^2 - c; \quad x > 0$$

La soluzione dell'equazione  $f(x) = 0$  fornisce, ovviamente,  $x = \sqrt{c}$ . Il metodo di Newton assume la forma

$$x_{k+1} = \frac{1}{2} \left( x_k + \frac{c}{x_k} \right)$$

che corrisponde ad un noto metodo per il calcolo della radice quadrata che utilizza solo le quattro operazioni elementari<sup>2</sup>. Poiché  $f' > 0, f'' > 0$  per  $x > 0$ , possiamo applicare il Teorema 5.1. Su ogni intervallo  $[a, b]$  con  $0 < a < \sqrt{c} < b$  il più piccolo valore della derivata è assunto in  $x = a$ ; si vede facilmente che la condizione 4 del teorema è soddisfatta per ogni  $b > (a + c/a)/2$  e quindi la successione di Newton converge per ogni scelta di  $x_0 > 0$ .

Più in generale, sia  $f(x) = x^r - c$ , con  $c > 0$  e  $r$  un intero qualunque positivo. Il metodo di Newton fornisce

$$x_{k+1} = \left( 1 - \frac{1}{r} \right) x_k + \frac{1}{r} c x_k^{1-k}$$

Anche in questo caso si ha convergenza per ogni  $x_0 > 0$ . ■

► **Esempio 5.7** Per ogni numero reale assegnato  $c > 0$  si vuole calcolare  $1/c$ . Un problema equivalente consiste nella ricerca della soluzione dell'equazione non lineare

$$f(x) = \frac{1}{x} - c = 0$$

<sup>2</sup>Tale metodo è noto come *formula di Erone*, matematico alessandrino del II secolo, anche se pare fosse già noto ai matematici babilonesi. Segnaliamo anche il seguente algoritmo per il calcolo della radice senza l'utilizzo dell'operazione di divisione:  $x_{k+1} = x_k + c - x_k^2$ ; esso corrisponde ad aggiungere alla stima corrente l'errore. Lasciamo come esercizio la dimostrazione della convergenza di tale metodo per  $0 < c < 1$  e  $0 \leq x_0 \leq 1$ .



Il metodo di Newton assume la forma

$$x_{k+1} = x_k(2 - cx_k)$$

che permette il calcolo del reciproco senza *divisioni*<sup>3</sup>. In questo caso si ha  $f'(x) < 0$ ,  $f''(x) > 0$  per  $x > 0$  e le condizioni del teorema di convergenza sono verificate se è possibile trovare un intervallo  $[a, b]$  con  $a < c^{-1} < b$  e

$$\frac{f(b)}{f'(b)} = b(bc - 1) \leq b - a$$

Quest'ultima disuguaglianza è soddisfatta se

$$b = \frac{1 + \sqrt{1 - ac}}{c}$$

Poiché  $a > 0$  può essere preso arbitrariamente piccolo, significa che il metodo converge a  $c^{-1}$  per ogni scelta di  $x_0$  con  $0 < x_0 < 2c^{-1}$ . ■

### 5.1.8 Modifiche del metodo di Newton

Il metodo di Newton richiede il calcolo della derivata prima della funzione  $f(x)$ , e nel caso di sistemi di  $n$  equazioni quello della matrice jacobiana, ossia il calcolo di  $n^2$  derivate parziali. Nelle applicazioni tale calcolo può essere eccessivamente costoso; ad esempio, come vedremo nei successivi Capitoli 12 e 13 nell'ambito del problema della identificazione dei parametri, la funzione  $f$  può essere il risultato della risoluzione di equazioni differenziali. In questi casi diventano allora interessanti le varianti del metodo di Newton che non implicano il calcolo esplicito delle derivate o ne richiedono un numero minore di valutazioni. Nel seguito esamineremo alcune di tali varianti, limitandoci, per semplicità, al caso unidimensionale.

#### Metodo delle corde

Il metodo corrisponde a porre nella formula del metodo di Newton la derivata prima uguale ad una costante, cioè

$$x_{k+1} = x_k - \frac{f(x_k)}{m}$$

Come vedremo nel paragrafo successivo, il metodo risulta convergente per opportune scelte di  $m$  e in corrispondenza è un metodo del *primo ordine*. Un *compromesso* conveniente, utilizzato in particolare nella risoluzione dei *sistemi* di equazioni non lineari, consiste nell'applicare il metodo di Newton tenendo costante, per un numero opportuno di iterazioni, la matrice Jacobiana.

<sup>3</sup>Su alcuni calcolatori questa procedura è utilizzata per una implementazione hardware della divisione.

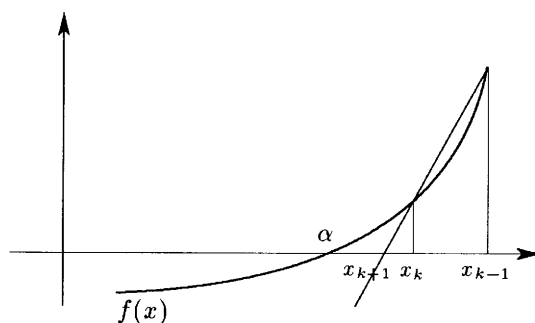


Figura 5.15: Illustrazione del metodo delle secanti.

### Metodo delle secanti

Il *metodo delle secanti* può essere visto come un'approssimazione del metodo di Newton o, alternativamente, come una differente realizzazione dell'idea dell'interpolazione. Date due *stime iniziali*  $x_0, x_1$ , si calcola per  $k = 2, 3, \dots$  la successione

$$x_{k+1} = x_k + h_k, \quad h_k = -f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \text{ se } f(x_k) \neq f(x_{k-1})$$

Geometricamente il punto  $x_{k+1}$  è lo zero della retta passante per i punti  $(x_k, f(x_k)), (x_{k-1}, f(x_{k-1}))$  (cfr. Figura 5.15).

Il costo dell'algoritmo è dato, per ogni iterazione, dal calcolo di un valore della funzione  $f(x)$ . Ricordiamo che nel metodo di Newton si ha il calcolo della funzione e della derivata. Procedendo in maniera analoga a quanto fatto per il metodo di Newton, si possono dimostrare *risultati di convergenza* validi nelle stesse ipotesi. L'*ordine di convergenza* è dato da  $r = (1 + \sqrt{5})/2 \approx 1.618$ . Assumendo come criteri di *valutazione* sia la *rapidità di convergenza*, sia il *costo per iterazione*, non è, quindi, in generale possibile stabilire a priori quale dei due metodi, Newton e secanti, è superiore; ciò è possibile solo esaminando il *caso particolare*.

### Metodo di Steffensen

Il *metodo di Steffensen* è un *metodo del secondo ordine*, che utilizza due valutazioni della funzione, ma non le derivate. Esso è definito nella forma seguente

$$x_{k+1} = x_k - \frac{f(x_k)}{g(x_k)}; \quad g(x_k) = \frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)}$$

Se si pone  $\beta_k = f(x_k)$  e si sviluppa  $g(x_k)$  in serie di Taylor intorno a  $x_k$ , si ha

$$g(x_k) = \frac{f(x_k + \beta_k) - f(x_k)}{\beta_k} = f'(x_k) \left( 1 - \frac{1}{2} h_k f''(x_k) + O(\beta_k^2) \right)$$

ove  $h_k = -f(x_k)/f'(x_k)$  è la correzione relativa al metodo di Newton. Pertanto

$$x_{k+1} = x_k + h_k \left(1 + \frac{1}{2} h_k f''(x_k) + O(\beta_k^2)\right)$$

da cui, usando la stima ottenuta in precedenza per il metodo di Newton, che può essere scritta nel seguente modo

$$h_k = -e_k + \frac{1}{2} e_k^2 \frac{f''(\xi)}{f'(x_k)}, \quad e_k = x_k - \alpha$$

si ha

$$\frac{e_{k+1}}{e_k^2} \rightarrow \frac{1}{2} \frac{f''(\alpha)}{f'(\alpha)} (1 + f'(\alpha))$$

e questo dimostra che il metodo è del secondo ordine. In Tabella 5.4 sono riportati i risultati ottenuti mediante il metodo di Steffensen (con  $x_0 = 2$ ) e il metodo delle secanti (con  $x_0 = 2, x_1 = 2.1$ ) relativamente alla seguente equazione

$$x \cosh \frac{d}{x} - x - l = 0, \quad d = 1, l = 0.1$$

k	Secanti		Steffensen	
	$x_k$	$f(x_k)$	$x_k$	$f(x_k)$
1	0.21000000D + 01	0.143D + 00	0.32640249D + 01	0.544D - 01
2	0.32298744D + 01	0.560D - 01	0.44154826D + 01	0.137D - 01
3	0.39612425D + 01	0.269D - 01	0.49454423D + 01	0.145D - 02
4	0.46360147D + 01	0.827D - 02	0.50155805D + 01	0.200D - 04
5	0.49356393D + 01	0.165D - 02	0.50165784D + 01	0.394D - 08
6	0.50103757D + 01	0.125D - 03	0.50165786D + 01	-0.351D - 14

Tabella 5.4: Metodo delle secanti e metodo di Steffensen applicati alla funzione  $x \cosh(d/x) - x - l$ , con rispettivamente  $x_0 = 2, x_1 = 2.1$  e  $x_0 = 2$ .

### Metodo di Muller

Il *metodo di Muller*, un'estensione del metodo delle secanti, utilizza un *polinomio di interpolazione di grado 2*. Esso richiede, quindi, tre stime iniziali. Inoltre fornisce, in generale, *due stime* della radice, tra le quali occorre effettuare una *scelta*; ad esempio nel metodo tradizionale viene scelta la più vicina alle stime delle iterazioni precedenti. Il metodo è descritto dalle formule

$$\begin{aligned}
 &x_0, x_1, x_2 \quad \text{stime iniziali} \\
 &s = f[x_k, x_{k-1}] + (x_k - x_{k-1})f[x_k, x_{k-1}, x_{k-2}] \quad k = 2, 3, \dots \\
 &x_{k+1} = x_k - \frac{2f(x_k)}{s + \text{sign}(s)\sqrt{s^2 - 4f(x_k)f[x_k, x_{k-1}, x_{k-2}]}}
 \end{aligned}$$

Un aspetto interessante del metodo consiste nel fatto che esso è in grado di approssimare le *radici complesse*. Ricordiamo che nel metodo di Newton questo è possibile solo se si assumono come stime iniziali dei numeri complessi.

### Metodi dell'interpolazione inversa

L'idea contenuta nel metodo di Muller potrebbe essere proseguita considerando polinomi di grado superiore. Sono chiare, comunque, le difficoltà che nascono dal calcolo delle radici del polinomio di interpolazione. Un modo differente di affrontare il problema, quando la funzione  $f(x)$  è *strettamente monotona*, è il seguente.

Supponiamo di conoscere  $k$  coppie di valori:  $(x_0, y_0), \dots, (x_{k-1}, y_{k-1})$  con  $y_i = f(x_i)$  e con i punti  $x_i$  distinti; allora, per l'ipotesi fatta sulla funzione  $f$ , anche i valori  $y_i$  sono *distinti*. Esiste quindi, ed è unico, il polinomio  $x = P_k(y)$  che interpola le coppie di valori  $(x_i, y_i)$  per  $i = 0, 1, \dots, k-1$ . Costruito tale polinomio, si pone

$$x_k = P_k(0)$$

In questo modo si possono costruire facilmente metodi ad un numero arbitrario di passi.

▼ **Osservazione 5.1** *Alcuni dei metodi precedenti, quando si approssimano problemi concreti, possono essere convenientemente usati in combinazione. Ad esempio se la funzione è concava o convessa (ipotesi del teorema 5.1), applicando il metodo di Newton e il metodo regula falsi, si ottengono due successioni monotone di senso opposto. La circostanza è interessante, in quanto fornisce un test pratico e affidabile (a meno degli errori di arrotondamento) per l'arresto delle iterazioni, in corrispondenza ad una tolleranza prefissata. Per una combinazione opportuna del metodo di bisezione e del metodo delle secanti si veda l'algoritmo di Dekker-Brent (cfr. Brent [21]).* ■

### 5.1.9 Radici di polinomi

In questo paragrafo considereremo il problema dell'approssimazione delle radici di una *equazione algebrica*, scritta nella seguente forma

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_1 z + a_0 = 0, \quad a_i \in \mathbb{R}, \quad z \in \mathbb{C} \quad (5.15)$$

per  $n$  intero  $\geq 1$ . Ricordiamo che nelle ipotesi fatte il polinomio  $P(z)$  ammette nel piano complesso esattamente  $n$  radici, e insieme ad ogni radice complessa vi è la corrispondente complessa coniugata. In altre parole, il polinomio  $P(z)$  può essere decomposto nel prodotto di polinomi lineari e quadratici a coefficienti reali<sup>4</sup>.

<sup>4</sup>È opportuno anche ricordare che, in base ad un risultato di Abel (1826), con sole operazioni aritmetiche ed estrazioni di radici non è possibile determinare le radici di una (*generica*) equazione algebrica di grado superiore a 4 e che quindi per tali equazioni è necessario ricorrere a metodi numerici.

Trattandosi di un caso particolare di equazioni non lineari, a priori possono essere utilizzati tutti i metodi che abbiamo analizzato nei paragrafi precedenti. In particolare, il *metodo di Newton* risulta particolarmente vantaggioso, in quanto il calcolo del polinomio  $P(z)$  e della derivata  $P'(z)$  per un valore particolare di  $z$  può essere ottenuto in maniera efficiente mediante il metodo di Horner-Ruffini che analizzeremo nel seguito (cfr. (5.16)). È interessante, tuttavia, vedere se è possibile costruire metodi più specifici, o anche se i metodi precedenti possono essere opportunamente adattati.

Un'osservazione importante riguarda l'equivalenza del problema della *risoluzione delle equazioni algebriche* con il problema del *calcolo degli autovalori di una matrice*. Rinviamo all'Appendice A per maggiori dettagli, ricordiamo che ad ogni polinomio è possibile associare una matrice, che ha tale polinomio come *polinomio caratteristico*. Pertanto, i metodi analizzati nel Capitolo 3 per l'approssimazione degli autovalori di una matrice sono, in effetti, anche metodi per la ricerca delle radici di una equazione algebrica. In particolare, ricordiamo il *metodo di Bernoulli*, associato al metodo delle potenze, utile, in particolare per l'approssimazione della *radice di modulo massimo*.

In questo paragrafo analizzeremo altri tipi di metodi, incominciando dal classico metodo di Bairstow<sup>5</sup>, che in sostanza, come vedremo, è un'applicazione del metodo di Newton per la risoluzione di un sistema di equazioni non lineari.

Ricordiamo prima i seguenti algoritmi per la divisione di un polinomio  $P(z)$  per un polinomio di primo e di secondo grado.

Per ogni valore  $\bar{z}$  fissato, la decomposizione

$$P(z) = (z - \bar{z})(b_{n-1}z^{n-1} + b_{n-2}z^{n-2} + \dots + b_0) + \bar{R}$$

viene ottenuta con il seguente algoritmo, noto come *algoritmo di Horner-Ruffini*, o anche *algoritmo della divisione sintetica* (*syntetic-division*)

$$\begin{aligned} b_{n-1} &= a_n \\ b_k &= a_{k+1} + \bar{z}b_{k+1}, \quad k = n-2, \dots, -1 \end{aligned} \tag{5.16}$$

Il resto  $\bar{R}$  della divisione è dato da  $\bar{R} = P(\bar{z}) = b_{-1}$ .

La divisione può essere continuata nel seguente modo

$$P(z) = (z - \bar{z})^2(c_{n-2}z^{n-2} + \dots + c_0) + (z - \bar{z})\bar{R}' + \bar{R}$$

ove

$$\begin{aligned} c_{n-2} &= b_{n-1} \\ c_k &= b_{k+1} + \bar{z}c_{k+1}, \quad k = n-3, \dots, -1 \\ \bar{R}' &= c_{-1} = P'(\bar{z}) \end{aligned}$$

<sup>5</sup>Il metodo è stato descritto da Bairstow nel 1914, in relazione allo studio della stabilità del volo degli aerei.

Procedendo in questa maniera, possono essere calcolate tutte le successive derivate in  $\bar{z}$  del polinomio e quindi il suo sviluppo di Taylor. Nel caso in cui  $\bar{z}$  sia una radice di  $P(z)$ , allora  $\bar{R} = 0$ . Il quoziente, ossia il polinomio di coefficienti  $b_i$ , contiene le rimanenti radici del polinomio  $P(z)$ . Si dice anche tale polinomio è stato ottenuto dal polinomio di partenza con una operazione di *deflazione*.

Generalizziamo ora l'algoritmo precedente al caso della divisione per un polinomio quadratico  $z^2 + pz + q$ . Partendo dall'identità

$$P(z) = (z^2 + pz + q)(b_{n-2}z^{n-2} + \dots + b_0) + b_{-1}(z + p) + b_{-2}$$

e introducendo, per comodità di scrittura, i valori  $b_n = b_{n-1} = 0$ , si ha

$$b_k = a_{k+2} - pb_{k+1} - qb_{k+2} \quad k = n-2, \dots, -1, -2$$

Il resto della divisione può essere scritto nella forma  $Rz + S$ , ove

$$R = a_1 - pb_0 - qb_1 \equiv b_{-1} \quad (5.17)$$

$$S = a_0 - qb_0 \equiv b_{-2} + pb_{-1} \quad (5.18)$$

Se  $z^2 + pz + q$  è un fattore di  $P(z)$ , allora  $R = S = 0$  e viceversa. Su tale osservazione è basato il seguente metodo numerico.

### Metodo di Bairstow

Il metodo ricerca una coppia di valori reali  $(p, q)$  tali da avere

$$\begin{cases} R(p, q) = 0 \\ S(p, q) = 0 \end{cases} \quad (5.19)$$

Si tratta, quindi, di risolvere un sistema non lineare di due equazioni nelle incognite  $(p, q)$ . Applicando il metodo di Newton, si hanno le seguenti formule, che descrivono il generico passo del metodo (cfr. Esempio 5.4)

$$p_{i+1} = p_i - \frac{1}{D} \left[ R \frac{\partial S}{\partial q} - S \frac{\partial R}{\partial q} \right]_{\substack{p=p_i \\ q=q_i}}$$

$$q_{i+1} = q_i - \frac{1}{D} \left[ S \frac{\partial R}{\partial p} - R \frac{\partial S}{\partial p} \right]_{\substack{p=p_i \\ q=q_i}}$$

ove

$$D = \begin{vmatrix} \frac{\partial R}{\partial p} & \frac{\partial R}{\partial q} \\ \frac{\partial S}{\partial p} & \frac{\partial S}{\partial q} \end{vmatrix}$$

rappresenta il determinante della matrice jacobiana della funzione  $[R, S]^T$ .

Utilizzando le formule (5.17) (5.18), si ha

$$\begin{aligned}\frac{\partial R}{\partial p} &= -p \frac{\partial b_0}{\partial p} - q \frac{\partial b_1}{\partial p} - b_0, & \frac{\partial R}{\partial q} &= -p \frac{\partial b_0}{\partial q} - q \frac{\partial b_1}{\partial q} - b_1 \\ \frac{\partial S}{\partial p} &= -q \frac{\partial b_0}{\partial p}, & \frac{\partial S}{\partial q} &= \frac{\partial b_{-2}}{\partial q} + p \frac{\partial b_{-1}}{\partial q}\end{aligned}$$

Osserviamo inoltre che

$$\begin{aligned}\frac{\partial b_{n-2}}{\partial p} &= \frac{\partial b_{n-1}}{\partial p} = 0 \\ \frac{\partial b_k}{\partial p} &= -b_{k+1} - p \frac{\partial b_{k+1}}{\partial p} - q \frac{\partial b_{k+2}}{\partial p}, & k &= n-3, \dots, 0, -1 \\ \frac{\partial b_{n-3}}{\partial q} &= \frac{\partial b_{n-2}}{\partial q} = 0 \\ \frac{\partial b_k}{\partial q} &= -b_{k+2} - p \frac{\partial b_{k+1}}{\partial q} - q \frac{\partial b_{k+2}}{\partial q}, & k &= n-4, \dots, -2\end{aligned}$$

Se definiamo la successione  $\{d_k\}$  mediante la formula ricorrente

$$d_{n-2} = d_{n-1} = 0, \quad d_k = -b_{k+1} - p d_{k+1} - q d_{k+2}, \quad k = n-3, \dots, 0, -1$$

si ottiene

$$\begin{aligned}\frac{\partial b_k}{\partial p} &= d_k; & \frac{\partial b_{k-1}}{\partial q} &= d_k, & k &= n-3, \dots, 0, -1 \\ \frac{\partial R}{\partial p} &= d_{-1}; & \frac{\partial R}{\partial q} &= d_0; & \frac{\partial S}{\partial p} &= -q d_0; & \frac{\partial S}{\partial q} &= d_{-1} + p d_0\end{aligned}$$

In definitiva, il metodo di Newton può essere scritto nella seguente forma

$$\begin{aligned}p_{i+1} &= p_i - \frac{1}{D} [b_{-1}(d_{-1} + p_i d_0) - (b_{-2} + p_i b_{-1} d_0)] \\ q_{i+1} &= q_i - \frac{1}{D} [(b_{-2} + p_i b_{-1}) d_{-1} + d_0 b_{-1} q_i] \\ D &= d_{-1}^2 + p_i d_0 d_{-1} + q_i d_0^2\end{aligned}$$

L'algoritmo è organizzato nel modo seguente. Dato il polinomio (5.15), si parte da una stima  $(p_0, q_0)$  e si applica il procedimento di Newton. Dalla decomponibilità di un polinomio a coefficienti reali in fattori quadratici reali segue l'esistenza di una soluzione del sistema (5.19). Per avere la convergenza del metodo si tratta quindi di operare una opportuna scelta del punto iniziale. Dal punto di vista pratico, se dopo un prefissato numero di iterazioni non si ha convergenza o il determinante diventa nullo, si riparte con una nuova coppia di valori iniziali. Ottenuta una soluzione accettabile, si passa ad applicare il metodo al polinomio quoziente, che è un polinomio

di grado  $n - 2$ . Ripetendo il procedimento, si arriva o ad un polinomio di grado 2 o ad un polinomio di grado 1.

Sottolineiamo che nella procedura precedente si ottiene una decomposizione *approssimata* del polinomio originario. Pertanto, a causa degli errori dovuti alle successive deflazioni, le radici dei successivi fattori della decomposizione possono differire dalle radici del polinomio di partenza.

### Metodo di Laguerre

Il metodo di Laguerre è basato sulla seguente iterazione

$$z_{k+1} = z_k - \frac{nP(z_k)}{P'(z_k) \pm \sqrt{H(z_k)}}$$

ove

$$H(z) = (n - 1)[(n - 1)(P'(z))^2 - nP(z)P''(z)]$$

Per una giustificazione del metodo si veda ad esempio [32]. Il segno nel denominatore è scelto in maniera che la quantità  $|z_{k+1} - z_k|$  sia minima. Il metodo richiede ad ogni passo il calcolo del polinomio e della derivata prima e seconda, ma risulta di ordine 3 per le radici semplici (reali o complesse). Per le equazioni con solo radici reali, il metodo è convergente per *ogni* scelta della stima iniziale.

Un possibile interesse del metodo consiste nella sua capacità di fornire, a differenza del metodo di Newton o simili, delle *radici complesse*, anche partendo da stime reali. Il motivo sta nel fatto che il termine  $H(z)$  può essere negativo.

#### 5.1.10 Sensitività delle radici di un polinomio

Quando gli algoritmi numerici visti in precedenza vengono implementati sul calcolatore, oltre agli errori di troncamento, dovuti al fatto che un algoritmo iterativo viene troncato dopo un numero finito di iterazioni, sono presenti gli errori di arrotondamento, sia quelli introdotti nella rappresentazione dei coefficienti  $a_i$  del polinomio, che quelli dovuti alle operazioni floating-point e propagati nei passi successivi dell'algoritmo. Nelle applicazioni, inoltre, i coefficienti  $a_i$ , che rappresentano in generale delle quantità fisiche, possono essere affetti da errori sperimentali (cfr. per una illustrazione l'Esempio 5.1). Per tali motivi, è importante, per una corretta interpretazione dei risultati ottenuti, conoscere il *condizionamento* del problema relativo al calcolo di una radice  $\alpha$  di un polinomio assegnato  $P(z)$ , ossia stimare l'influenza su  $\alpha$  di "piccole" perturbazioni relative ai coefficienti  $a_i$  del polinomio  $P(z)$ .

Il problema può essere studiato in maniera generale nel seguente modo. Indicato con  $g(z)$  un polinomio arbitrario e con  $\epsilon$  un numero positivo "piccolo", possiamo rappresentare il polinomio perturbato nella seguente forma

$$P_\epsilon(z) = P(z) + \epsilon g(z)$$



Se  $\alpha$  è una radice semplice di  $P$ , si può dimostrare che per  $\epsilon$  sufficientemente piccolo esiste una funzione analitica  $\alpha(\epsilon)$ , con  $\alpha(0) = \alpha$ , tale che  $\alpha(\epsilon)$  è una radice semplice del polinomio perturbato  $P_\epsilon$ , cioè

$$P(\alpha(\epsilon)) + \epsilon g(\alpha(\epsilon)) = 0$$

Differenziando rispetto a  $\epsilon$ , si ha

$$P'(\alpha(\epsilon)) \frac{d\alpha(\epsilon)}{d\epsilon} + g(\alpha(\epsilon)) + \epsilon g'(\alpha(\epsilon)) \frac{d\alpha(\epsilon)}{d\epsilon} = 0$$

da cui, facendo tendere  $\epsilon$  a zero

$$\left. \frac{d\alpha(\epsilon)}{d\epsilon} \right|_{\epsilon=0} = - \frac{g(\alpha)}{P'(\alpha)}$$

Sviluppando in serie la funzione  $\alpha(\epsilon)$  e trascurando i termini in potenze di  $\epsilon$  superiori a 1, si ha

$$\alpha(\epsilon) \approx \alpha + \epsilon \left( - \frac{g(\alpha)}{P'(\alpha)} \right)$$

Nel caso di una radice multipla  $\alpha$  di ordine  $m$ , si può mostrare che  $P(z) + \epsilon g(z)$  ha una radice della forma

$$\alpha(\epsilon) = \alpha + s(\epsilon^{1/m})$$

ove  $s(t)$  è, per  $|t|$  sufficientemente piccolo, una funzione analitica con  $s(0) = 0$ . Differenziando  $m$  volte rispetto a  $t$  e tenendo conto che  $P(\alpha) = P'(\alpha) = \dots = P^{(m-1)}(\alpha) = 0$ ,  $P^{(m)}(\alpha) \neq 0$ , da

$$0 \equiv P_\epsilon(\alpha(\epsilon)) = P(\alpha + s(t)) + t^m g(\alpha + s(t)), \quad t^m = \epsilon$$

si ottengono le seguenti relazioni

$$P^{(m)}(\alpha)k^m + m!g(\alpha) = 0, \quad k := \left. \frac{ds(t)}{dt} \right|_{t=0} = \left[ - \frac{m!g(\alpha)}{P^{(m)}(\alpha)} \right]^{1/m}$$

Utilizzando ancora una approssimazione del primo ordine, si ha

$$\alpha(\epsilon) \approx \alpha + \epsilon^{1/m} \left[ - \frac{m!g(\alpha)}{P^{(m)}(\alpha)} \right]^{1/m}$$

Supponiamo, ora, che il polinomio  $P(z)$  sia dato nella forma (5.15), cioè mediante i suoi coefficienti  $a_i$ . Se assumiamo

$$g_i(z) := a_i z^i$$

il polinomio  $P_\epsilon$  è il polinomio che si ottiene sostituendo il coefficiente  $a_i$  con  $a_i(1+\epsilon)$ . La formula precedente fornisce la seguente stima dell'effetto sulla radice  $\alpha$  di un errore relativo  $\epsilon$  di  $a_i$

$$\alpha(\epsilon) - \alpha \approx \epsilon^{1/m} \left[ -\frac{m! a_i \alpha^i}{P^{(m)}(\alpha)} \right]^{1/m}$$

Se la radice è diversa dallo zero, dalla formula precedente si può ottenere una stima per l'errore relativo. Il fattore di proporzionalità  $\epsilon^{1/m}$  cresce all'aumentare di  $m$ , in accordo con il fatto che le radici multiple sono peggio *condizionate* delle radici semplici. Naturalmente, nel condizionamento conta anche il fattore

$$k(i, \alpha) := \left| \frac{a_i \alpha^i}{P^{(m)}(\alpha)} \right|$$

che può rendere malcondizionata anche una radice semplice.

► **Esempio 5.8** [Wilkinson, 1959] Si consideri il polinomio

$$P(z) := (z-1)(z-2)\cdots(z-20) = \sum_{i=0}^{20} a_i z^i \quad (5.20)$$

Le radici sono ben separate, ma, perturbando, ad esempio, il coefficiente  $a_{15} \approx -10^{10}$ , si ha

$$\alpha_{16}(\epsilon) - \alpha_{16} \approx \epsilon 3.7 \cdot 10^{14}$$

Questo significa che le radici del polinomio sono talmente malcondizionate che un'aritmetica a 14 cifre decimali non garantisce nessuna cifra corretta di  $\alpha_{16}$ .

Al contrario, le radici del polinomio

$$P(z) := \prod_{j=1}^{20} (z - 2^{-j}), \quad \alpha_j = 2^{-j}$$

sono "vicine" ma ben condizionate. Ad esempio, perturbando  $a_0$  si ha

$$\left| \frac{\alpha_{20}(\epsilon) - \alpha_{20}}{\alpha_{20}} \right| \approx \left| \epsilon \frac{1}{(2^{-1}-1)(2^{-2}-1)\cdots(2^{-19}-1)} \right| \leq 4|\epsilon|$$

Osserviamo, comunque, che le radici sono bencondizionate solo rispetto a piccole perturbazioni *relative* dei coefficienti, ma non per piccole perturbazioni *assolute*. Se, ad esempio, sostituiamo  $a_0 = 2^{-210}$  con  $\bar{a}_0 = a_0 + \Delta a_0$ ,  $\Delta a_0 = 2^{-48}$  ( $\approx 10^{-14}$ ), allora il polinomio modificato ha radici con  $\bar{\alpha}_1 \cdots \bar{\alpha}_{20} = \bar{a}_0 = 2^{-210} + 2^{-48} = (2^{162} + 1)(\alpha_1 \cdots \alpha_{20})$ . Esiste quindi almeno un indice  $r$  con  $|\bar{\alpha}_r/\alpha_r| \geq (2^{162} + 1)^{1/20} > 2^8 = 256$  ■

▼ **Osservazione 5.2** *L'analisi precedente si riferisce alla sensitività delle radici rispetto ai coefficienti del polinomio rappresentato nella forma (5.15). Se il polinomio viene rappresentato in altra forma, i risultati possono, ovviamente, cambiare. Si pensi, ad esempio, alla possibilità di rappresentare un polinomio come polinomio caratteristico di una matrice. In questo caso, il problema si riconduce alla sensitività degli autovalori rispetto agli elementi della matrice e sappiamo che quando la matrice è simmetrica tale problema è sempre bencondizionato.* ■

◆ **Esercizio 5.1** Analizzare il metodo di Newton per la ricerca della radice della seguente equazione (cfr. Figura 5.16)

$$f(x) := \frac{1}{(x - 0.2)^2 + 0.01} + \frac{1}{(x - 0.9)^2 + 0.04} - 6 = 0$$

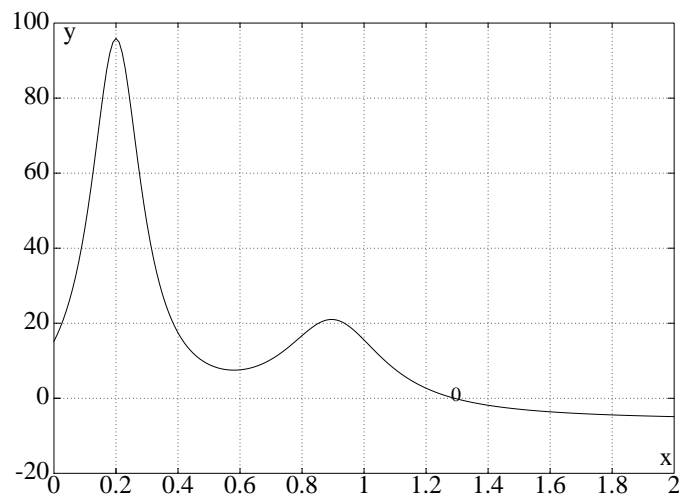


Figura 5.16: Grafico della funzione  $y = 1/((x - 0.2)^2 + 0.01) + 1/((x - 0.9)^2 + 0.04) - 6$ .

◆ **Esercizio 5.2** Localizzare mediante un metodo grafico tutte le radici dell'equazione

$$\ln(x + 1) + \tan(2x) = 0$$

◆ **Esercizio 5.3** Applicare il metodo di Newton per la ricerca della radice della seguente equazione

$$2x(1 - x^2 + x) \ln x = x^2 - 1$$

◆ **Esercizio 5.4** Analizzare il comportamento della successione generata dall'applicazione del metodo di Newton per la ricerca di uno zero della seguente funzione

$$2x^3 - 9x^2 + 12x + 15 = 0$$

in corrispondenza a valori iniziali  $x_0 = 3$ ,  $x_0 < 3$ , oppure  $x_0 > 3$ .

◆ **Esercizio 5.5** Esaminare sul caso particolare dell'equazione  $f(x) := x^2 - c = 0$ ,  $c > 0$  il seguente metodo iterativo

$$x_{k+1} = y - \frac{f(y)}{f'(x_k)}, \quad y = x_k - \frac{f(x_k)}{f'(x_k)}$$

◆ **Esercizio 5.6** L'equazione  $\sin(xy) = y - x$  definisce  $y$  come funzione implicita di  $x$ . La funzione  $y(x)$  ha un massimo per  $(x, y) \approx (1, 2)$ . Mostrare che le coordinate di tale massimo possono essere calcolate risolvendo il seguente sistema non lineare

$$\begin{cases} \sin(xy) - y + x = 0 \\ y \cos(xy) + 1 = 0 \end{cases}$$

Applicare per la risoluzione di tale sistema il metodo di Newton.

◆ **Esercizio 5.7** Determinare tutte le radici delle seguenti equazioni algebriche

$$2z^3 + 21z^2 - 26z - 240 = 0$$

$$2z^3 + 21z^2 - 10z - 210 = 0$$

◆ **Esercizio 5.8** Programmare il metodo di Bairstow, assumendo come verifica la risoluzione della seguente equazione

$$z^3 - 4z^2 + 6z - 4 = 0, \quad \text{con } p_0 = q_0 = 0$$

◆ **Esercizio 5.9** Localizzare le radici del seguente polinomio

$$P(z) = z^4 + 8z^3 - 8z^2 - 200z - 425$$

utilizzando i risultati noti per gli autovalori di una matrice.

◆ **Esercizio 5.10** Sia  $f(x) = \cos(\beta x) - x$ , con  $\beta = 2.7332 \pm 0.5 \cdot 10^{-4}$ . Indicare con quale accuratezza può essere risolta l'equazione. Determinare quindi la radice con un errore assoluto che non superi l'accuratezza ottenibile di più del 10%.

◆ **Esercizio 5.11** Interpretare il polinomio (5.20) come polinomio caratteristico di una matrice simmetrica.

◆ **Esercizio 5.12** Applicare il metodo di Laguerre all'equazione

$$P(z) = z^3 - 5z^2 - 17z + 21$$

partendo da  $z_1 = 10^6$ .

◆ **Esercizio 5.13** Approssimare le prime dieci radici positive della seguente equazione

$$\cosh \phi \cos \phi = -1$$

che ha origine nello studio delle vibrazioni elastiche di una linguetta in strumenti musicali.

◆ **Esercizio 5.14** In relazione al flusso con turbolenza di un fluido in un tubo, supponiamo che tra il fattore di frizione  $c_f$  e il numero di Reynolds  $Re$  valga la seguente relazione

$$\frac{1}{c_f} = -0.4 + 1.74 \ln(\text{Re} \sqrt{c_f})$$

Calcolare  $c_f$  per  $Re = 10^4, 10^5$  e  $10^6$ .

◆ **Esercizio 5.15** Un mezzo rappresentato da una semiretta  $x \geq 0$  è a una temperatura iniziale uniforme  $T_0$ . Per  $t > 0$  viene applicato alla superficie  $x = 0$  un flusso costante  $q$ . Se la conduttività termica e la diffusività termica sono rispettivamente  $k$  e  $\alpha$ , si può mostrare che la temperatura nel punto a distanza  $x$  e al tempo  $t$  è data da

$$T = T_0 + \frac{q}{k} \left[ 2 \sqrt{\frac{\alpha t}{\pi}} e^{-x^2/4\alpha t} - x \operatorname{erfc} \frac{x}{2\sqrt{\alpha t}} \right]$$

ove  $\operatorname{erfc}(z) = (2/\sqrt{\pi}) \int_z^\infty e^{-t^2} dt$  è la funzione errore complementare. Posto  $T_0 = 70$ ,  $q = 300$ ,  $k = 1.0$ ,  $\alpha = 0.04$ ,  $x^* = 1.0$ , esaminare uno schema per il calcolo del tempo  $t^*$  per il quale nel punto  $x^*$  si abbia la temperatura  $T^* = 120$ .

◆ **Esercizio 5.16** La velocità  $q d\lambda$  alla quale l'energia irradiante lascia l'unità di area della superficie di un corpo nero nell'intervallo di lunghezza d'onda  $(\lambda, \lambda + d\lambda)$  è data dalla legge di Planck

$$q d\lambda = \frac{2\pi h c^2 d\lambda}{\lambda^5 (e^{hc/k\lambda T} - 1)}$$

ove  $c = 2.997925 \cdot 10^{10}$  è la velocità della luce,  $h = 6.6256 \cdot 10^{-27}$  è la costante di Planck,  $k = 1.38054 \cdot 10^{-16}$  la costante di Boltzmann,  $T$  la temperatura assoluta e  $\lambda$  la lunghezza d'onda.

In corrispondenza ad una assegnata temperatura di superficie  $T$ , studiare uno schema per il calcolo della lunghezza d'onda  $\lambda_{\max}$  per la quale l'energia radiante ha la massima intensità (cercare  $\lambda_{\max}$  come soluzione dell'equazione  $dq/d\lambda = 0$ ). Calcolare i valori di  $\lambda_{\max}$  in corrispondenza a diversi valori di  $T$ , ad esempio  $T = 1000, 2000, 3000, 4000$ . Verificare, quindi, la validità della legge dello spostamento di Wien (Wien's displacement law)  $\lambda_{\max} T = \text{costante}$

◆ **Esercizio 5.17** In relazione allo studio dell'energia solare raccolta attraverso un campo di specchi piani su un collettore centrale, è stata proposta la seguente definizione di fattore di concentrazione geometrica  $C$

$$C = \frac{\pi (h/\cos A)^2 F}{0.5\pi D^2(1 + \sin A - 0.5 \cos A)}$$

ove  $A$  indica l'angolo al bordo del campo,  $F$  la frazione di copertura del campo mediante gli specchi,  $D$  il diametro del collettore, e  $h$  l'altezza del collettore. Trovare  $A$  se  $h = 300$ ,  $C = 1200$ ,  $F = 0.8$ , e  $D = 14$ .

◆ **Esercizio 5.18** Un modo classico per risolvere le equazioni di terzo grado è fornito dalle formule di Tartaglia-Cardano. L'equazione cubica

$$x^3 + ax^2 + bx + c = 0$$

è trasformata nella forma ridotta

$$y^3 + py + q = 0$$

mediante la sostituzione  $x = y - a/3$ . Si ottiene

$$p = b - \frac{a^2}{3}, \quad q = c - \frac{ab}{3} + 2 \left(\frac{a}{3}\right)^3$$

Una radice reale  $x_1$  della forma ridotta può essere trovata mediante le seguenti formule

$$s = \left[ \left( \frac{p}{3} \right)^3 + \left( \frac{q}{2} \right)^2 \right]^{1/2}, \quad y_1 = \left[ -\frac{q}{2} + s \right]^{1/3} + \left[ -\frac{q}{2} - s \right]^{1/3} \rightarrow \boxed{x_1 = y_1 - \frac{a}{3}}$$

Le altre due radici possono essere trovate mediante formule analoghe, o dividendo per  $x_1$  e risolvendo l'equazione quadratica quoziente.

Applicare il metodo di Tartaglia-Cardano per la ricerca della radice reale della seguente equazione

$$x^3 + 3x^2 + \alpha^2 x + 3\alpha^2 = 0$$

per vari valori di  $\alpha$ . In particolare, esaminare la eventuale perdita di accuratezza per valori grandi di  $\alpha$ .

Per la stessa equazione esaminare il comportamento del metodo di Newton, al variare di  $\alpha$  e del valore iniziale.

## 5.2 Metodi di punto fisso

Data una trasformazione  $x \rightarrow g(x)$ , ad esempio da  $\mathbb{R}$  in  $\mathbb{R}$ , il problema di trovare un numero  $\alpha$  tale che  $\alpha = g(\alpha)$  è noto come *problema di punto fisso* ed  $\alpha$  è chiamato un *punto fisso* (o punto unito) di  $g(x)$ . Come illustrato in Figura 5.17, un punto fisso corrisponde ad un'intersezione della curva  $y = g(x)$  con la bisettrice  $y = x$ .

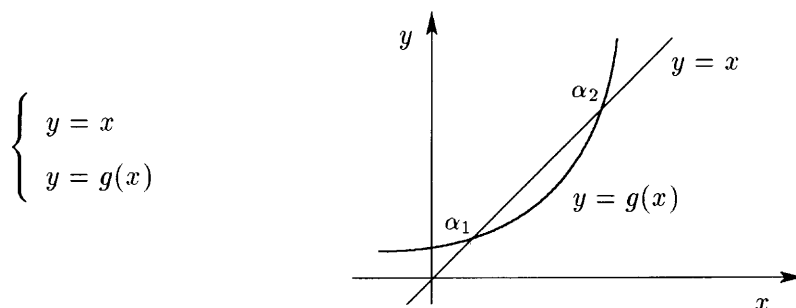


Figura 5.17: Punti uniti di una trasformazione  $x \rightarrow g(x)$ .

Il problema della ricerca del punto fisso di una trasformazione è un problema di grande importanza in numerose applicazioni. Nel seguito, tuttavia, ci limiteremo a ricordare alcuni tra i risultati che sono alla base dell'applicazione del metodo alla ricerca degli zeri di una funzione, rinviando per una trattazione più adeguata, ad esempio, a Ortega e Rheinboldt [124].

Il legame tra il problema del punto unito e quello della ricerca di uno zero di una funzione è il seguente. Data una funzione  $f(x)$ , si costruisce una funzione ausiliaria  $g(x)$ , in maniera che si abbia  $\alpha = g(\alpha)$  ogniqualvolta  $f(\alpha) = 0$ . La costruzione di

$g(x)$  non è, ovviamente, unica. Si ha, ad esempio

$$f(x) = x^2 - x - 2 \iff \begin{array}{l} 1. \quad g(x) = x^2 - 2 \\ 2. \quad g(x) = \sqrt{2+x} \\ 3. \quad g(x) = 1 + 2/x \\ 4. \quad g(x) = x - (x^2 - x - 2)/(2x - 1) \end{array}$$

In particolare, il caso 4 corrisponde al metodo di Newton, per il quale, come abbiamo visto, si ha  $g(x) = x - f(x)/f'(x)$ . Esamineremo nel seguito i *criteri* per una *scelta conveniente* della  $g(x)$ , incominciando a ricordare alcuni risultati di *esistenza* e di *unicità* del punto fisso.

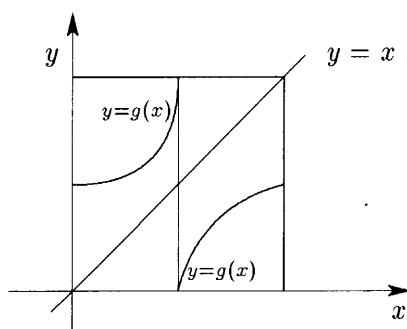


Figura 5.18: Discontinuità e punto fisso.

Data una generica funzione  $g(x)$  definita su un intervallo limitato e chiuso  $I \equiv [a, b]$  e a valori reali, essa può avere più punti fissi (come mostrato in Figura 5.17), oppure, come si può esemplificare facilmente, non avere nessun punto fisso. In altre parole, per assicurare che la  $g(x)$  abbia un punto fisso in  $I$  sono necessarie opportune restrizioni su  $g(x)$ . La prima ipotesi ragionevole è che si abbia  $g(I) \subseteq I$ , ossia che per ogni  $x \in I$  si abbia  $g(x) \in I$ . Tuttavia, la Figura 5.18 mostra che tale ipotesi non è ancora sufficiente, in quanto se la funzione  $g(x)$  non è continua, una parte del suo grafico può essere al di sopra della retta  $y = x$  e l'altra al di sotto. L'esistenza è, in effetti, assicurata dalla continuità.

**Teorema 5.2** *Se  $g(I) \subseteq I$ , con  $I$  intervallo limitato e chiuso, e  $g(x)$  è continua, allora  $g(x)$  ha almeno un punto fisso in  $I$ .*

**DIMOSTRAZIONE.** L'ipotesi  $g(I) \subseteq I$  significa  $a \leq g(a) \leq b$  e  $a \leq g(b) \leq b$ . Se  $a = g(a)$ , oppure  $b = g(b)$ , allora uno degli estremi dell'intervallo è un punto unito. Supponiamo, al contrario, che si abbia  $g(a) - a > 0$  e  $g(b) - b < 0$ . La funzione  $F(x) = g(x) - x$  è continua, con  $F(a) > 0$  e  $F(b) < 0$ . Allora, per il noto teorema sulle funzioni continue, esiste un punto  $\alpha$  con  $F(\alpha) = g(\alpha) - \alpha = 0$ . ■

Per assicurare l'*unicità* del punto fisso, occorre in sostanza impedire che la  $g(x)$  vari troppo rapidamente. Si ha, infatti, la seguente condizione sufficiente.

**Teorema 5.3** (Esistenza e unicità del punto fisso) *Se  $g(I) \subseteq I$  e  $g(x)$  è derivabile su  $I$ , con*

$$|g'(x)| \leq L < 1 \quad \text{per ogni } x \in I \quad (5.21)$$

*allora esiste uno ed un solo valore  $\alpha \in I$  tale che  $g(\alpha) = \alpha$ .*

**DIMOSTRAZIONE.** La condizione (5.21) assicura la continuità della funzione  $g(x)$  e quindi l'esistenza del punto fisso. Per dimostrare l'unicità, supponiamo che  $\alpha_1, \alpha_2$  siano due punti fissi distinti. Applicando il teorema del valore medio, si ottiene allora

$$|\alpha_1 - \alpha_2| = |g(\alpha_1) - g(\alpha_2)| = |g'(\xi)(\alpha_1 - \alpha_2)| \leq L|\alpha_1 - \alpha_2| < |\alpha_1 - \alpha_2|$$

che è una contraddizione. ■

La (5.21) è una condizione sufficiente affinché la funzione  $g(x)$  sia una *contrazione* su  $I$ . Ricordiamo, infatti, che  $g(x)$  è una contrazione su  $I$ , quando esiste una costante  $L$ , detta *costante di contrazione*, tale che

$$|g(x_1) - g(x_2)| \leq L|x_1 - x_2|, \quad 0 \leq L < 1$$

Una funzione contrattiva su  $I$  non è, comunque, necessariamente derivabile.

Una volta stabilita una condizione per la quale  $g(x)$  ha un unico punto fisso su  $I$ , rimane il problema del *calcolo numerico* di tale valore. La tecnica che utilizzeremo è nota come *iterazione del punto fisso* (o metodo di Picard, o anche di iterazioni successive) ed è definita nel seguente modo

$$\begin{aligned} x_0 & \text{ arbitrario in } I \\ x_{k+1} & = g(x_k) \quad \text{per } k = 0, 1, \dots \end{aligned} \quad (5.22)$$

Graficamente, il metodo può essere illustrato convenientemente congiungendo successivamente i punti  $(x_k, g(x_k)), (x_{k+1}, x_{k+1})$ . Per introdurre il problema della convergenza del procedimento (5.22), consideriamo il seguente esempio numerico.

► **Esempio 5.9** Partiamo dal problema del calcolo degli zeri della seguente equazione

$$x^3 - 3x^2 + 1 = 0$$

Si può verificare che gli zeri sono reali, con i seguenti valori

$$\alpha_1 \approx -0.5321; \quad \alpha_2 \approx 0.6527; \quad \alpha_3 \approx 2.8794$$

Trasformiamo, ora, il problema originario in un problema equivalente di calcolo di punti fissi. Più precisamente, riscriviamo l'equazione nella seguente forma

$$x = g(x) := \frac{x^2}{3} + \frac{1}{3x}$$



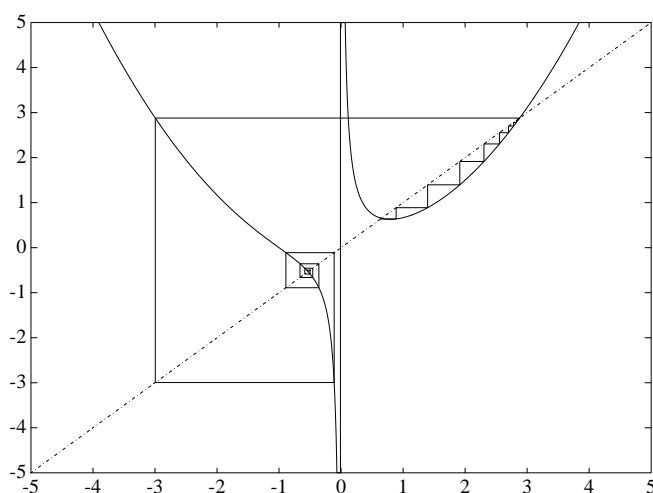


Figura 5.19: Procedimento iterativo  $x_{k+1} = x_k^2/3 + 1/(3x_k)$ , con  $x_0 = -0.5542$ .

$k$	$x_k$	$k$	$x_k$	$k$	$x_k$	$k$	$x_k$
1	-0.5542	9	-2.9952	17	2.8515	25	0.8883
2	-0.4991	10	2.8790	18	2.8272	26	0.6383
3	-0.5849	11	2.8794	19	2.7822	27	0.6580
4	-0.4559	12	2.8782	20	2.7001	28	0.6509
5	-0.6618	13	2.8771	21	2.5536	29	0.6533
6	-0.3577	14	2.8752	22	2.3042	30	0.6525
7	-0.8894	15	2.8714	23	1.9145	31	0.6528
8	-0.1111	16	2.8645	24	1.3959	32	0.6527

Tabella 5.5: Valori generati dal procedimento iterativo  $x_{k+1} = x_k^2/3 + 1/(3x_k)$ , con  $x_0 = -0.5542$ .

ove la funzione di iterazione  $g(x)$  è definita su ogni intervallo  $I$  che non contenga lo zero (cfr. per una rappresentazione grafica la Figura 5.19).

Il procedimento (5.22) fornisce, a partire dal valore  $x_0 = -0.5422$ , i risultati riportati in Tabella 5.5 e illustrati graficamente in Figura 5.19.

Rileviamo che, nonostante il punto iniziale sia stato scelto vicino al punto fisso  $\alpha_1$ , la successione dei valori  $x_k$  si allontana da tale punto, con un comportamento di tipo alternante, fino ad arrivare ad un valore molto vicino al punto fisso  $\alpha_3$ ; si allontana, comunque, anche da tale punto, con un comportamento, questa volta, di tipo monotono; infine, si vede che la successione converge, con un comportamento alternante, al punto fisso corrispondente alla radice  $\alpha_2$ .

Si conclude, quindi, che il metodo iterativo considerato è utile soltanto per il calcolo della radice  $\alpha_2$ . È interessante esaminare il valore della derivata  $g'(x)$  nei differenti punti fissi. Si hanno i seguenti valori

$$g'(\alpha_1) \approx -1.5320; \quad g'(\alpha_2) \approx -0.3473; \quad g'(\alpha_3) \approx 1.8794$$

dai quali si rileva che la convergenza si ha relativamente al punto in cui la derivata è in

modulo minore di 1, ossia quando la  $g(x)$  è una contrazione in un opportuno intorno del punto fisso. Si osserva, inoltre, che al segno negativo della derivata corrisponde un comportamento alternante della successione. ■

Le conclusioni dell'esempio precedente sono formalizzate più in generale nel seguente risultato.

**Teorema 5.4 (Convergenza)** *Sia  $g(I) \subseteq I \equiv [a, b]$  e  $|g'(x)| \leq L < 1$  per ogni  $x \in I$ . Per  $x_0 \in I$ , la successione  $x_{k+1} = g(x_k)$ ,  $k = 0, 1, \dots$  converge ad un punto fisso  $\alpha$ , e l'errore  $e_k := x_k - \alpha$  è maggiorato nel seguente modo*

$$|e_k| \leq \frac{L^k}{1-L} |x_1 - x_0| \quad (5.23)$$

**DIMOSTRAZIONE.** Dal Teorema 5.3 si ha l'esistenza di un unico punto fisso  $\alpha$  in  $I$ . Dato un qualsiasi  $k$ , applicando il teorema del valore medio si ottiene

$$|x_k - \alpha| = |g(x_{k-1}) - g(\alpha)| = |g'(\xi_k)| |x_{k-1} - \alpha| \leq L |x_{k-1} - \alpha|$$

Successive applicazioni di tale disuguaglianza forniscono il risultato

$$|x_k - \alpha| \leq L^k |x_0 - \alpha| \Rightarrow \lim_{k \rightarrow \infty} x_k = \alpha$$

dal momento che, essendo  $0 \leq L < 1$ , si ha  $\lim_{k \rightarrow \infty} L^k = 0$ . Per stabilire la maggiorazione dell'errore (5.23) osserviamo che

$$|x_0 - \alpha| \leq |x_0 - x_1| + |x_1 - \alpha| \leq |x_0 - x_1| + L|x_0 - \alpha|$$

da cui  $(1-L)|x_0 - \alpha| \leq |x_1 - x_0|$ . Il risultato (5.23) segue allora dal fatto che  $|x_k - \alpha| \leq L^k |x_0 - \alpha|$ . ■

La maggiorazione (5.23) fornisce ad ogni passo una stima dell'errore, che può essere utilizzata come *test* per terminare la iterazione.

Il Teorema 5.4 è chiamato un teorema di convergenza *globale*, in quanto esso specifica un intervallo fissato, *noto a priori*,  $I = [a, b]$ , e dimostra che il procedimento iterativo converge per ogni  $x_0 \in I$ . Si dicono, invece, teoremi di convergenza *locali*, i risultati che assicurano la convergenza quando  $x_0$  è scelto *sufficientemente vicino* al punto fisso  $\alpha$ . Le condizioni di tali risultati sono più facilmente verificabili, ma, in generale, non è possibile specificare quanto  $x_0$  debba essere scelto vicino a  $\alpha$  per avere la convergenza. Nel seguito forniamo alcuni esempi interessanti di risultati locali, la cui dimostrazione è lasciata come esercizio.

**Teorema 5.5** *Supponiamo che  $g'(x)$  sia continua in un intervallo aperto contenente  $\alpha$ , ove  $\alpha$  è un punto fisso di  $g(x)$ . Se  $|g'(\alpha)| < 1$ , allora esiste un intorno di  $\alpha$  tale che il procedimento iterativo converge per qualunque  $x_0$  scelto in tale intorno.*

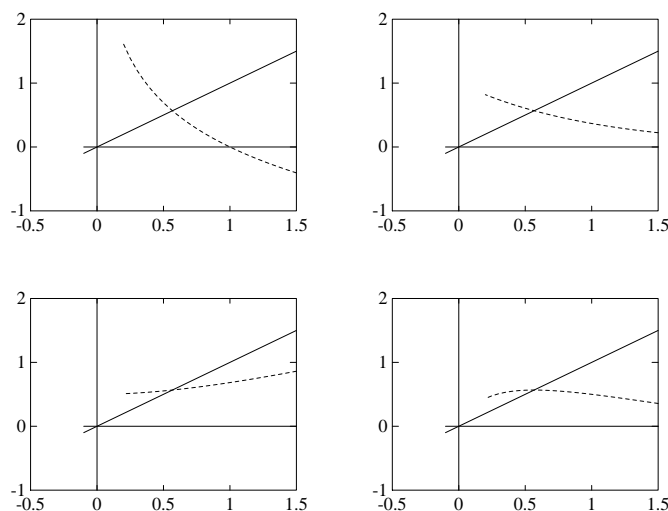


Figura 5.20: Esempi di procedimenti iterativi.

Osserviamo che nel caso di un sistema a  $n$  dimensioni, la condizione da verificare assume la forma

$$\|\mathbf{J}(\alpha)\| < 1, \quad \|\cdot\| \text{ norma in } \mathbb{R}^n$$

ove  $\mathbf{J} = \left[ \frac{\partial g_i}{\partial x_j} \right]$  rappresenta la matrice jacobiana di  $\mathbf{g} = [g_1, g_2, \dots, g_n]^T$ .

► **Esempio 5.10** L'equazione  $x + \log x = 0$  ha una radice  $\alpha$  in un intorno del punto 0.5. Consideriamo i seguenti procedimenti iterativi (cfr. Figura 5.20).

1.  $x_{k+1} = g_1(x_k) := -\log x_k$
2.  $x_{k+1} = g_2(x_k) := e^{-x_k}$
3.  $x_{k+1} = g_3(x_k) := (x_k + e^{-x_k})/2$
4.  $x_{k+1} = g_4(x_k) := x_k(1 - \log x_k)/(x_k + 1)$

Si hanno i seguenti risultati

$$g'_1(\alpha) \approx -2; \quad g'_2(\alpha) \approx -0.6; \quad g'_3(\alpha) \approx 0.2; \quad g'_4(\alpha) = 0.$$

Pertanto, mentre i metodi (2), (3), (4) convergono, il metodo (1) è divergente. La rapidità di convergenza è indicata dal modulo della derivata prima. I metodi (2), (3) sono lineari; il metodo (4) è del secondo ordine. ■

► **Esempio 5.11** Se la derivata prima di  $g$  nel punto fisso ha modulo uguale a 1, vi può essere sia convergenza che divergenza, a seconda del caso particolare. Consideriamo ad esempio la funzione  $g_1(x)$  definita nel modo seguente e rappresentata in Figura 5.21

$$g_1(x) = \begin{cases} \frac{1}{2}x^2 + \frac{1}{2} & \text{per } x \leq 1 \\ 2\sqrt{x} - 1 & \text{per } x \geq 1 \end{cases} \quad (5.24)$$

La funzione  $g_1$  è continua su  $\mathbb{R}$  ed ha  $\alpha = 1$  come unico punto fisso. La derivata prima è data da

$$g_1'(x) = \begin{cases} x & \text{per } x < 1 \\ \frac{1}{\sqrt{x}} & \text{per } x > 1 \end{cases} \quad g_1'(1) = \begin{cases} \lim_{x \rightarrow 1^-} g_1'(x) \\ \lim_{x \rightarrow 1^+} g_1'(x) \end{cases} = 1$$

Studiando direttamente il comportamento delle iterate si trova che il metodo converge per ogni scelta di  $x_0 \in \mathbb{R}$  (come appare evidente dalla Figura 5.21).

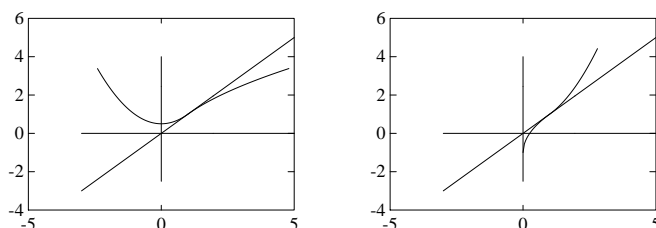


Figura 5.21: Rappresentazione della funzione  $g_1(x)$  definita in (5.24) e della funzione  $g_2(x)$  definita in (5.25).

Consideriamo, al contrario, la funzione  $g_2(x)$  definita nel modo seguente e rappresentata in Figura 5.21

$$g_2(x) = \begin{cases} 2\sqrt{x} - 1 & \text{per } x \leq 1 \\ \frac{1}{2}x^2 + \frac{1}{2} & \text{per } x > 1 \end{cases} \quad (5.25)$$

La funzione  $g_2$  è continua su  $\mathbb{R}$  ed ha  $\alpha = 1$  come unico punto fisso. La derivata prima è data da

$$g_2'(x) = \begin{cases} \frac{1}{\sqrt{x}} & \text{per } 0 < x < 1 \\ x & \text{per } x > 1 \end{cases} \quad g_2'(1) = 1$$

Studiando il comportamento delle iterate si trova che il metodo è divergente per ogni scelta di  $x_0 \in \mathbb{R}$ , distinto da 1. Tale comportamento è anche deducibile direttamente dalla rappresentazione grafica. ■

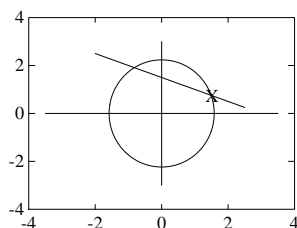


Figura 5.22: Rappresentazione grafica del sistema non lineare (5.26).

► **Esempio 5.12** Consideriamo il sistema non lineare

$$\begin{cases} x_1 + 2x_2 = 3 \\ 2x_1^2 + x_2^2 = 5 \end{cases} \quad (5.26)$$

Dalla rappresentazione grafica uno studio grafico (cfr. Figura 5.22) si vede che il sistema presenta una radice nel rettangolo

$$R := \begin{array}{|c|} \hline 1 \leq x_1 \leq 2 \\ \hline 0.5 \leq x_2 \leq 1.5 \\ \hline \end{array}$$

Consideriamo il procedimento iterativo definito da  $\mathbf{x} = \mathbf{g}(\mathbf{x})$ , con  $\mathbf{x} = [x_1, x_2]^T$  e  $\mathbf{g}$  definita nel modo seguente

$$\begin{cases} x_1 = \left( \frac{5 - x_2^2}{2} \right)^{1/2} \\ x_2 = \frac{1}{2}(3 - x_1) \end{cases}$$

La matrice jacobiana  $\mathbf{J}$  della funzione  $\mathbf{g}(\mathbf{x})$  è data da

$$\mathbf{J} = \begin{bmatrix} 0 & -x_2/(10 - 2x_2^2)^{1/2} \\ -1/2 & 0 \end{bmatrix}$$

Si ha  $\sup_{\mathbf{x} \in R} \|\mathbf{J}\|_1 = 3/\sqrt{22}$  e quindi il metodo iterativo risulta convergente alla radice  $[1.488, 0.7560]^T$ . ■

**Corollario 5.1** *Se  $g'(\alpha) \neq 0$ , allora esiste un intorno  $I$  di  $\alpha$  tale che per ogni  $x_0$  scelto in  $I$  le approssimazioni successive  $x_k$  convergono ad  $\alpha$ . Inoltre, se esiste  $g''(x)$  ed è limitata e  $g'(x)$  è continua in  $I$ , allora*

$$|x_{k+1} - \alpha| \leq C|x_k - \alpha|^2$$

ossia il metodo è del secondo ordine.

► **Esempio 5.13** Il metodo di Newton che abbiamo analizzato nel paragrafo precedente per la ricerca di una radice  $\alpha$  di  $f(x) = 0$  può essere considerato un metodo per la ricerca del punto fisso ponendo

$$g(x) := x - \frac{f(x)}{f'(x)}$$

Se  $f'(\alpha) \neq 0$  e la funzione è derivabile due volte si verifica che  $g'(\alpha) = 0$  e quindi la convergenza quadratica del metodo può essere ricavata dal Corollario 5.1. Nel caso in cui  $f^{(\nu)}(\alpha) = 0$ ,  $\nu = 0, 1, \dots, r-1$ , si vede facilmente che

$$g'(\alpha) = 1 - \frac{1}{r}$$

e quindi, per  $r > 1$ , il metodo di Newton è solo linearmente convergente. ■

**Teorema 5.6** *Sia  $g(x)$  una trasformazione dell'intervallo  $I = [x_0 - r, x_0 + r]$ , con  $r > 0$ , in se stesso. Supponiamo, inoltre, che in  $I$  sia  $|g'(x)| \leq L < 1$  e che si abbia  $|x_0 - g(x_0)| \leq (1 - L)r$ . Allora, la successione  $x_k$  generata dal metodo delle iterazioni successive è ben definita, ossia gli elementi della successione appartengono all'intervallo  $I$ , e converge ad un punto  $\alpha$  in  $I$ . Tale punto risulta l'unico punto fisso di  $g$  in  $I$ , ed inoltre si ha*

$$|x_k - \alpha| \leq L^k r$$

Del Teorema 5.6 verrà data nel successivo Capitolo 7 un'applicazione importante nell'ambito degli schemi impliciti per la risoluzione di problemi differenziali a valori iniziali.

### 5.2.1 Aspetti computazionali

Consideriamo un procedimento convergente del primo ordine, cioè, se  $\alpha$  è il valore limite, sia

$$|x_k - \alpha| \leq L^k |x_0 - \alpha|$$

Il numero di iterazioni per ridurre l'errore iniziale di un fattore  $10^{-q}$  è ottenuto ponendo  $L^k \leq 10^{-q}$ . Prendendo i logaritmi si trova

$$k \geq q / \log(1/L)$$

In questo caso, quindi, il numero di iterazioni è approssimativamente proporzionale a  $q$ . Per confronto, la convergenza *quadratica* (assumendo  $L|x_0 - \alpha| < 1$ ) fornisce

$$|x_k - \alpha| \leq L^{2^k - 1} |x_0 - \alpha|^{2^k}$$

Per ottenere una riduzione di un fattore di  $10^{-q}$ , poniamo

$$L^{2^k - 1} |x_0 - \alpha|^{2^k} \leq 10^{-q} |x_0 - \alpha|$$

e prendendo i logaritmi

$$k \geq \log_2 q - \log_2 \log_{10}((1/L)|x_0 - \alpha|)$$

Allora,  $k$  è proporzionale a  $\log_2 q$ . Per esempio, se  $q = 8$  sono sufficienti  $k \geq 3$  iterazioni. Per  $L = 1/2$ , un metodo del primo ordine richiederebbe approssimativamente 27 iterazioni.

### Errori di arrotondamento

Indicando con  $\{\tilde{x}_k\}$  la successione *calcolata*, si ha

$$\tilde{x}_{k+1} = g(\tilde{x}_k) + \delta_k$$

ove i termini  $\delta_k$  raccolgono gli effetti degli errori di arrotondamento che si commettono al passo  $k$  e che si propagano dai passi precedenti. A causa della presenza degli errori di arrotondamento, non è assicurata, in generale, la convergenza della successione al punto fisso  $\alpha$ . Comunque, se la trasformazione  $g$  è contrattiva e  $|\delta_k| < \delta$ , allora la successione  $\{\tilde{x}_k\}$  è contenuta in una sfera di centro  $\alpha$  e raggio  $\delta$ . Più precisamente, si può dimostrare il seguente risultato.

**Teorema 5.7** Sia  $g : \mathbb{R} \rightarrow \mathbb{R}$  una funzione derivabile con  $|g'(x)| \leq L$ , con  $0 \leq L < 1$ . Se  $\{\tilde{x}_k\}$  è la successione generata numericamente dal procedimento di iterazioni successive, con  $x_0 \in \mathbb{R}$  e se  $|\delta_k| < \delta$  allora

$$|\tilde{x}_k - \alpha| \leq \frac{\delta}{1-L} + \frac{L^k}{1-L} ((L+1)\delta + \|\tilde{x}_1 - \tilde{x}_0\|)$$

Se ad esempio  $\delta = 10^{-t}$ , ove  $t$  è la precisione utilizzata nel calcolo, allora per  $k$  sufficientemente elevato, l'errore sarà  $\leq 10^{-t}/(1-L) + \epsilon_k$ , con  $\epsilon_k \rightarrow 0$ . Se  $L = 1 - \gamma$ , allora  $1/(1-L) = 1/\gamma$  diventa arbitrariamente grande per  $\gamma$  che tende a zero. L'errore in questo caso è  $10^{-t}/\gamma$  e se assumiamo, ad esempio,  $\gamma = 10^{-t}$ , allora l'errore ad ogni iterata è di grandezza 1.

Si consideri come esempio  $g(x) := 30x^4(\sin^2(1/x) - 1/(1+x^2))$ . Tale funzione presenta un punto fisso nell'intervallo  $[10, 20]$ . In corrispondenza a  $x$  "grandi" il calcolo della funzione presenta un problema di *cancellazione*, per cui il contributo dei termini  $\delta_k$  può essere significativo.

Le osservazioni precedenti sono, chiaramente, importanti per una conveniente scelta delle *condizioni di arresto* dell'algoritmo iterativo.

### 5.2.2 Accelerazione della convergenza

Supponiamo che la funzione  $g(x)$  sia sufficientemente regolare e che la successione generata a partire da un punto  $x_0$  mediante il procedimento delle iterazioni successive converga a un punto fisso  $\alpha$ . Si ha

$$e_{k+1} = \alpha - x_{k+1} = g'(\xi_k)e_k = g'(\xi_k)(\alpha - x_k) \quad (5.27)$$

ove  $\xi_k$  è un opportuno punto tra  $\alpha$  e  $x_{k+1}$ . Poiché  $\lim_{k \rightarrow \infty} x_k = \alpha$ , si ha

$$\lim_{k \rightarrow \infty} g'(\xi_k) = g'(\alpha)$$

Di conseguenza, se  $g'(\alpha) \neq 0$ , si ha

$$e_{k+1} \approx g'(\alpha)e_k$$

e pertanto il metodo è del primo ordine. L'interesse ora è rivolto alla possibilità di costruire, partendo dalla successione  $\{x_k\}$ , una successione  $\{\hat{x}_k\}$  che converga *più rapidamente*. Risolvendo rispetto ad  $\alpha$  la (5.27), si ottiene

$$\alpha = x_{k+1} + \frac{g'(\xi_k)(x_{k+1} - x_k)}{1 - g'(\xi_k)} = x_{k+1} + \frac{x_{k+1} - x_k}{g'(\xi_k)^{-1} - 1}$$

Il valore  $g'(\xi_k)$  non è in generale noto; si può tuttavia dare una *stima* di tale valore osservando che, applicando il teorema del valor medio, si ha, per un opportuno  $\theta_k \in (x_k, x_{k+1})$

$$r_k := \frac{x_k - x_{k-1}}{x_{k+1} - x_k} = \frac{x_k - x_{k-1}}{g(x_k) - g(x_{k-1})} = g'(\theta_k)^{-1}$$

Asintoticamente, cioè per valori grandi di  $k$ , si ha

$$r_k = \frac{1}{g'(\theta_k)} \approx \frac{1}{g'(\alpha)} \approx \frac{1}{g'(\xi_k)}$$

e allora il valore

$$\hat{x}_k = x_{k+1} + \frac{x_{k+1} - x_k}{r_k - 1}; \quad r_k = \frac{x_k - x_{k-1}}{x_{k+1} - x_k} \quad (5.28)$$

dovrebbe fornire una *migliore approssimazione* di  $\alpha$  rispetto a  $x_k$  o  $x_{k+1}$ . Si può vedere facilmente che il valore  $\hat{x}_k$  corrisponde al punto fisso della retta

$$s(x) = g(x_k) + g[x_{k-1}, x_k](x - x_k)$$

ove  $[x_{k-1}, x_k]$  indica la differenza divisa del primo ordine (cfr. Capitolo 3).

► **Esempio 5.14** L'equazione

$$1.5x - \tan x = 0.1$$

ha una radice  $\alpha = 0.205921695\dots$ . Consideriamo il procedimento iterativo definito dalla funzione

$$g(x) = \frac{0.1 + \tan x}{1.5}$$

partendo da  $x_0 = 0$ . Si ha  $g'(\alpha) \approx 0.45636\dots$ , per cui il metodo è del primo ordine. Si ha, ad esempio  $x_2 = 0.1111\dots$ ;  $\hat{x}_2 = 0.2024\dots$ ;  $x_5 = 0.1751\dots$ ;  $\hat{x}_5 = 0.2053\dots$  ■

Il procedimento, consistente nel ricavare da una successione *linearmente convergente*  $\{x_k\}$  una successione  $\{\hat{x}_k\}$  *convergente più rapidamente*, è chiamato *procedimento  $\Delta^2$  di Aitken*. Usando le notazioni

$$\Delta x_k = x_{k+1} - x_k, \quad \Delta^2 x_k = \Delta(\Delta x_k) = \Delta x_{k+1} - \Delta x_k$$

il risultato (5.28) può essere espresso nella forma

$$\hat{x}_k = x_{k+1} - \frac{(\Delta x_k)^2}{\Delta^2 x_{k-1}}$$

Il procedimento si applica ad ogni successione linearmente convergente. Per quanto riguarda la convergenza si ha il seguente risultato.

**Teorema 5.8** Sia  $\{x_k\}$  una successione reale convergente a  $\alpha$ . Supponiamo che  $e_k = x_k - \alpha$  verifichi la condizione

$$e_{k+1} = (A + \delta_k)e_k \quad (5.29)$$

ove  $A$  è una costante, con  $|A| < 1$  e  $\delta_k \rightarrow 0$  per  $k \rightarrow \infty$ . Inoltre supponiamo  $e_k \neq 0, \forall k$ . Allora la successione  $\{\hat{x}_k\}$  generata dal procedimento di Aitken è ben definita per  $k$  sufficientemente grande e

$$\lim_{k \rightarrow \infty} \frac{\hat{x}_k - \alpha}{x_k - \alpha} = 0$$



DIMOSTRAZIONE. Dalla condizione (5.29) si ha:  $e_{k+2} = (A + \delta_{k+1})(A + \delta_k)e_k$ . Quindi

$$x_{k+2} - 2x_{k+1} + x_k = e_{k+2} - 2e_{k+1} + e_k = [(A - 1)^2 + \alpha_k]e_k$$

ove  $\alpha_k = A(\delta_k + \delta_{k+1}) - 2\delta_k + \delta_k\delta_{k+1}$ . Allora,  $\alpha_k \rightarrow 0$  e per  $k$  sufficientemente grande  $(A - 1)^2 + \alpha_k \neq 0$ . Questo implica che il metodo è ben definito. Un calcolo immediato mostra che

$$\hat{x}_k - \alpha = e_k - \frac{(A - 1 + \delta_k)^2 e_k}{(A - 1)^2 + \alpha_k}$$

e quindi il risultato richiesto. ■

Una variante del metodo di Aitken, chiamata *metodo di Steffensen*, consiste nell'utilizzare il valore ottenuto con il metodo di Aitken come nuovo punto di partenza per l'iterazione. Si ha cioè la formula

$$x_{k+1} = x_k - \frac{(g(x_k) - x_k)^2}{g(g(x_k)) - 2g(x_k) + x_k}$$

Se la funzione  $g(x)$  è sufficientemente regolare e  $g'(\alpha) \neq 0$  allora l'algoritmo produce una successione quadraticamente convergente. Relativamente alla funzione dell'Esempio 5.14 si ha  $x_3 = 0.2059217$ .

◆ **Esercizio 5.19** Studiare i seguenti procedimenti iterativi

1.  $x^4 = \sin x \iff x_{k+1} = \sqrt[4]{\sin x_k}$

2.  $x/2 = \sin x \iff x_{k+1} = 2 \sin x_k$

◆ **Esercizio 5.20** Considerare l'iterazione  $x_{k+1} = g(x_k)$  per le seguenti scelte di  $g$

1.  $g(x) := (4 + 4x - x^2)^{1/3}$  punto fisso  $x = 2$ .

2.  $g(x) := (6 - x^3)/5$  punto fisso  $x = 1$ .

3.  $g(x) := 2x^{-1} + x/2$  punto fisso  $x = 2$ .

4.  $g(x) := \frac{1}{3} + \frac{2}{3} \left( x + \frac{x-1}{3x^2 - 6x + 2} \right)$  punto fisso  $x = 1$ .

◆ **Esercizio 5.21** Considerare la funzione  $g(x)$  definita da

$$g(x) := x - \frac{f(x)}{f'(x)} - \frac{f''(x)}{2f'(x)} \left( \frac{f(x)}{f'(x)} \right)^2$$

Se  $\alpha$  è uno zero di  $f$  e  $f'(\alpha) \neq 0$ , mostrare che l'ordine di convergenza del metodo iterativo definito da  $g$  è almeno cubico, per valori iniziali sufficientemente vicini a  $\alpha$ .

◆ **Esercizio 5.22** Studiare, al variare di  $x_0 \in \mathbb{R}$ , la convergenza del procedimento iterativo

$$x_{k+1} = e^{x_k} - \frac{3}{2}$$

◆ **Esercizio 5.23** Studiare la convergenza dei metodi iterativi  $x_{k+1} = g(x_k)$ , con

$$\begin{aligned} g_1(x) &= x^3 - 5 \\ g_2(x) &= \sqrt[3]{x+5} \\ g_3(x) &= \frac{5}{x^2 - 1} \end{aligned}$$

per determinare le radici dell'equazione  $x^3 - x - 5 = 0$ .

◆ **Esercizio 5.24** Studiare il seguente metodo iterativo, trovando i punti limite e l'ordine di convergenza

$$x_{k+1} = \frac{x_k(x_k^2 + 3A)}{3x_k^2 + A}, \quad k \geq 0, A > 0$$

◆ **Esercizio 5.25** Discutere, al variare del parametro  $p > 0$ , la convergenza del metodo iterativo:

$$x_{k+1} = -\frac{1}{p}(e^{2x_k} + 2)$$

◆ **Esercizio 5.26** Trovare il numero di radici dell'equazione

$$xe^{-x} = e^{-p}$$

al variare del parametro  $p$ . Discutere il condizionamento delle radici al variare di  $p$ . Per  $p = 3$  confrontare il metodo di Newton con il metodo iterativo

$$x_{k+1} = e^{x_k - 3}$$

◆ **Esercizio 5.27** Studiare la risoluzione del sistema

$$\begin{cases} x_1 = \frac{1}{2} \cos x_2 \\ x_2 = \frac{1}{2} \sin x_1 \end{cases}$$

◆ **Esercizio 5.28** Il sistema non lineare

$$\begin{cases} x_1^2 + x_2^2 + x_3 - 4.12 = 0 \\ x_1^2 + x_2^2 + x_3 - 6.43 = 0 \\ x_1^2 + x_2^2 + x_3 - 6.52 = 0 \end{cases}$$

ha una soluzione nella sfera di centro  $[2, 1, 1]$  e raggio 0.2 nella norma  $\|\cdot\|_\infty$ . Applicare il metodo di Newton e studiare eventuali metodi iterativi alternativi.

◆ **Esercizio 5.29** Risolvere il sistema non lineare

$$x_1^2 + x_2^2 - 2x_1 - 2x_2 + 1 = 0; \quad x_1 + x_2 - 2x_1x_2 = 0$$

◆ **Esercizio 5.30** Riformulare le seguenti equazioni in modo da ottenere procedimenti iterativi convergenti

1.  $x^3 - x + 1 = 0$
2.  $\log(1 + x) - x^2 = 0$
3.  $e^x - 3x^3 = 0$

◆ **Esercizio 5.31** Applicare il metodo di Aitken e di Steffensen al seguente metodo iterativo

$$x_{k+1} = \sqrt{2 + x_k}$$

## 5.3 Sistemi dinamici discreti

Per *sistema dinamico* si intende un sistema, di varia natura (fisica, chimica, elettromeccanica, biologica, economica, eccetera), che evolve nel tempo. La descrizione di tale evoluzione può essere ottenuta con vari tipi di modelli matematici.<sup>6</sup> In particolare, quando la variabile tempo è *continua*, ossia una variabile reale, i modelli più comuni sono basati sull'utilizzo delle equazioni differenziali ordinarie o a derivate parziali, o più in generale su equazioni funzionali del tipo equazioni con ritardo o equazioni integro-differenziali. Tali modelli saranno considerati successivamente nel Capitolo 7. Quando la variabile tempo è *discreta*, indicando con  $k$  gli istanti successivi e con  $x_k$  il valore della variabile che descrive lo *stato* del sistema al tempo  $k$ , possiamo rappresentare matematicamente il sistema nel seguente modo

$$x_{k+1} = g_k(x_k, x_{k-1}, \dots, x_{k-r+1}, u_k), \quad k = r - 1, r, \dots \quad (5.30)$$

ove  $g_k$  è una funzione assegnata per ogni valore di  $k$ ,  $r$  è un intero fissato, e  $u_k$  è una funzione nota. L'equazione (5.30), detta usualmente, in analogia alle equazioni differenziali, *equazione alle differenze*, descrive allora un *sistema dinamico discreto*. A partire da un insieme di  $r$  valori  $\{x_0, x_1, \dots, x_{r-1}\}$ , essa fornisce i successivi stati del sistema. Attraverso la funzione  $u_k$  si può influire sul comportamento del sistema. Per tale motivo, tale funzione viene detta *variabile di controllo*, e i *problemi di controllo* consistono nella ricerca della funzione  $u_k$  in maniera da *ottimizzare* un determinato criterio.

Per uno studio approfondito dei sistemi dinamici discreti e per il loro utilizzo come strumento di costruzione di modelli matematici si veda, ad esempio, Lakshminantham e Trigiane [103], Eisen [53].

In questo paragrafo ci limiteremo ad illustrare alcuni risultati relativi al caso particolare dell'equazione (5.30), in cui  $r = 1$  e  $g$  è indipendente da  $k$ , ossia all'equazione

$$x_{k+1} = g(x_k), \quad k = 0, 1, \dots \quad (5.31)$$

<sup>6</sup>Il processo di costruzione di un modello di un sistema occupa un posto fondamentale nella scienza e nell'ingegneria. I modelli matematici cercano di "mimare" la realtà usando il linguaggio della matematica, ossia mediante un insieme di equazioni. Le variabili in tali equazioni sono associate con le quantità fisiche del sistema e una relazione tra le variabili nell'espressione matematica è analoga alla relazione tra le corrispondenti entità fisiche.

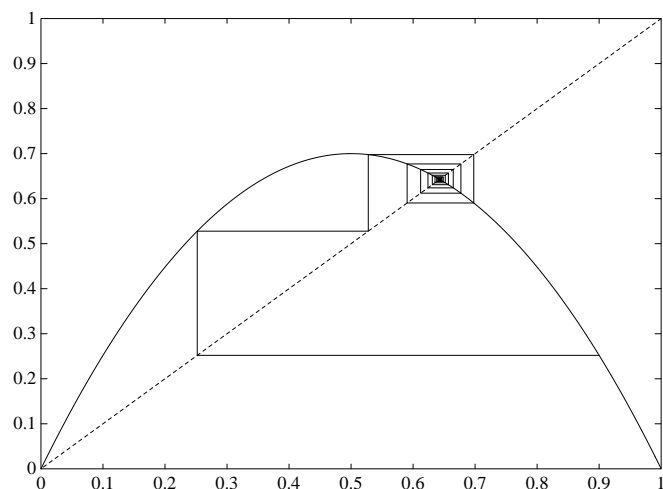


Figura 5.23: Rappresentazione della successione  $x_k$  generata dal procedimento iterativo  $x_{k+1} = rx_k(1 - x_k)$ , per  $r = 2.8$  e  $x_0 = 0.9$ .

che abbiamo considerato nel paragrafo precedente in relazione al calcolo degli zeri di una funzione.

► **Esempio 5.15** *Equazione della logistica discreta.* Consideriamo il seguente sistema discreto

$$x_{k+1} = rx_k(1 - x_k), \quad r > 0, \quad k = 0, 1, \dots \quad (5.32)$$

che può essere assunto come un modello di accrescimento di una popolazione, per la quale sono presenti effetti di rallentamento nella crescita, rappresentati dal termine  $(1 - x_k)$ , ove 1 è il risultato di una normalizzazione. Dato  $x_0$ , con  $0 < x_0 < 1$ , si è interessati alle soluzioni  $x_k \geq 0$ , e in particolare al comportamento della popolazione per  $k \rightarrow \infty$ .

Posto  $g(x) = rx(1 - x)$ , la trasformazione  $x \rightarrow g(x)$  trasforma l'intervallo  $[0, 1]$  in se stesso per ogni  $r$ , con  $0 \leq r \leq 4$ . I punti fissi della trasformazione sono le soluzioni della seguente equazione

$$x^* = rx^*(1 - x^*)$$

ossia il punto  $x^* = 0$  per  $r < 1$  e i punti  $x^* = 0$  e  $x^* = \frac{r-1}{r}$  per  $r > 1$ . In tali punti, detti anche *punti di equilibrio*, in quanto per  $x_0$  uguale ad uno di tali valori si ottiene una successione costante, la derivata  $g'(x)$  assume i seguenti valori

$$g'(0) = r, \quad g'\left(\frac{r-1}{r}\right) = 2 - r$$

Dai risultati del paragrafo precedente si ha che per  $0 < r < 1$  la successione  $x_k$  converge per ogni scelta di  $x_0$  nell'intervallo  $[0, 1]$  all'unico punto fisso  $x^* = 0$ ; esso è, quindi un *punto di equilibrio stabile*. In termini di interpretazione del modello di popolazioni significa, ad esempio, che ogni immigrazione di popolazione  $x_0 > 0$  è destinata all'estinzione.

Per  $r = 1$  il punto  $x^* = 0$  diventa *instabile*, in quanto si ha  $g'(0) = r = 1$  e di conseguenza la successione  $x_k$ , per  $x_0 > 0$  non converge a zero. Al contrario, per  $0 < r < 3$  diventa stabile il punto di equilibrio  $x^* = (r - 1)/r > 0$ , per il quale si ha  $-1 < g'(x^*) < 1$ .

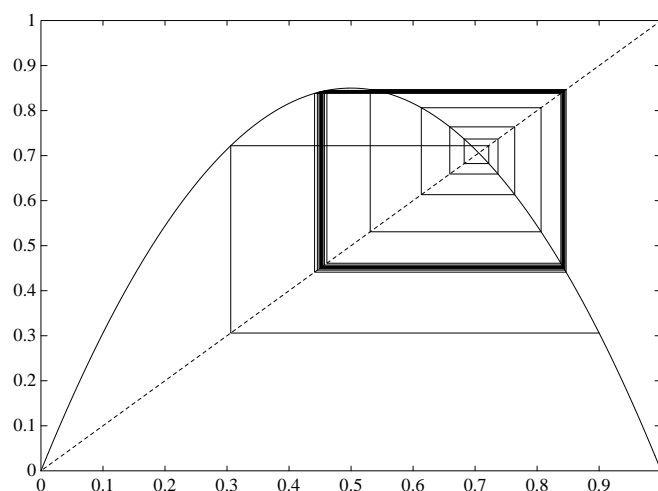


Figura 5.24: Rappresentazione della successione  $x_k$  generata dal procedimento iterativo  $x_{k+1} = rx_k(1 - x_k)$ , per  $r = 3.4$  e  $x_0 = 0.9$ .

Si dice, anche, che in  $r = 1$  si verifica una biforcazione. Come esemplificazione, in Figura 5.23 è rappresentata graficamente la successione  $x_k$ , in corrispondenza a  $r = 2.8$  e  $x_0 = 0.9$ . Il punto di equilibrio stabile è dato da  $x^* = 0.6429$ .

La seconda biforcazione si verifica per  $r = 3$ . Per un valore  $r > 3$  anche il punto di equilibrio  $x^* = (r - 1)/r$  diventa instabile, in quanto si ha  $g'(x^*) < -1$ . In Figura 5.24 è rappresentata la situazione per  $r = 3.4$  e  $x_0 = 0.9$ . La figura suggerisce l'esistenza di una soluzione periodica, con periodo 2. Più precisamente, si può dimostrare che esiste un valore  $x^{**}$  tale che

$$x^{**} = r[rx^{**}(1 - x^{**})] [1 - rx^{**}(1 - x^{**})]$$

Per  $r > 3$ , si ottiene

$$x^{**} = \frac{(r + 1) \pm [(r + 1)(r - 3)]^{1/2}}{2r} > 0$$

La precedente soluzione periodica di periodo 2 rimane *stabile* per  $3 < r < r_4$ , ove  $r_4 \approx 3.45$  (cfr. per una illustrazione la Figura 5.25). Successivamente, per  $r_4 < r < r_8$  appare una soluzione stabile periodica, con periodo quattro. La situazione si ripete, ossia per  $r$  che aumenta ogni soluzione di periodo  $p$  pari si biforca in una soluzione di periodo  $2p$ . Per ogni  $n$  vi è una una soluzione di periodo  $2^n$ , e associato con ciascuna di essa vi è un intervallo del parametro  $r$  nel quale essa è stabile. La distanza tra due successive biforcazioni sull'asse  $r$  diventa sempre più piccola. Vi è un valore limite  $r_c \approx 3.828$  al quale si ha instabilità per tutte le soluzioni periodiche di periodo  $2^n$  e per  $r > r_c$  appaiono cicli, localmente attrattivi, con periodi  $k, 2k, 4k, \dots$ , ma ora  $k$  è dispari. In Figura 5.26 è rappresentata la soluzione corrispondente a  $r = 4$  e  $x_0 = 0.9$ .

È stato dimostrato (Sarkovskii (1964), Li e Yorke (1975)) che se per un valore  $r_c$  esiste una soluzione di periodo 3, allora per ogni  $n \geq 1$  esistono soluzioni di periodo  $n$ , ed inoltre esistono soluzioni aperiodiche, cioè soluzioni che non presentano delle configurazioni ripetute

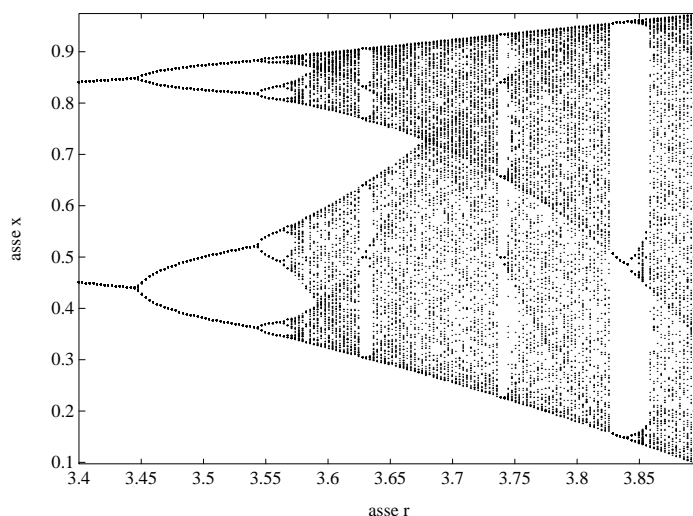


Figura 5.25: Rappresentazione schematica delle soluzioni stabili per il modello della logistica discreto al variare di  $r$ . Ad ogni biforcazione lo stato precedente diventa instabile.

e che sono indistinguibili dai valori generati da una funzione aleatoria. Una situazione di tale tipo è stata chiamata *caotica*.

Osserviamo che il genere di comportamento messo in evidenza per l'equazione della logistica è tipico delle equazioni alle differenze del tipo (5.31), per le quali la funzione  $g(x)$  ha una forma analoga a quella della funzione  $x(1-x)$ . Tra le varie e interessanti applicazioni dello studio del comportamento caotico, segnaliamo in particolare il suo utilizzo per un approccio deterministico alla *turbolenza* nell'ambito della meccanica dei fluidi (cfr. [16]).

## 5.4 Programmazione lineare

I modelli matematici sono spesso utilizzati per prendere decisioni; con il termine di *ottimizzazione* viene appunto indicato l'utilizzo di un modello matematico per la individuazione della migliore alternativa tra le varie possibilità assegnate.

In termini matematici, per ottimizzazione si indica allora la ricerca del massimo o del minimo di una funzione  $f(\mathbf{x})$ , per  $\mathbf{x} \in \Omega \subseteq \mathbb{R}^n$ . L'insieme  $\Omega$ , che definisce le varie possibilità tra le quali è possibile scegliere, è detto *insieme di ammissibilità* (feasible set). La funzione  $f(\mathbf{x})$ , che assegna a ogni membro dell'insieme di ammissibilità un numero che misura la "desiderabilità" della scelta corrispondente, è detta *funzione obiettivo*, o funzione costo.

Un *problema di programmazione lineare* (brevemente, problema LP) è un particolare problema di ottimizzazione vincolata nel quale sia la funzione obiettivo che

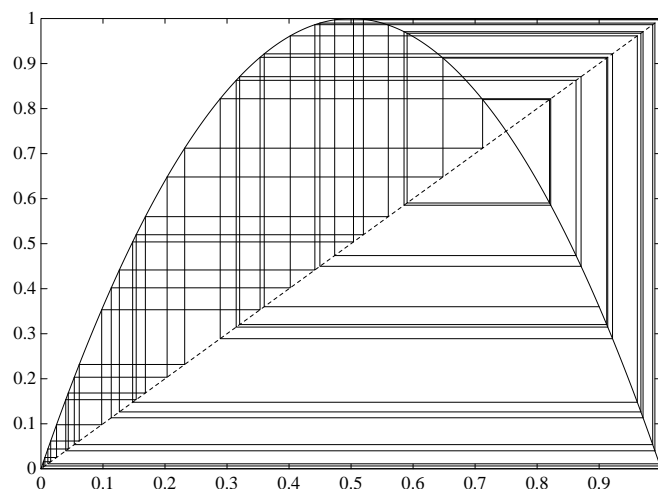


Figura 5.26: Rappresentazione della successione  $x_k$  generata dal procedimento iterativo  $x_{k+1} = rx_k(1 - x_k)$ , per  $r = 4$  e  $x_0 = 0.9$ .

le funzioni che traducono i vincoli sono *funzioni lineari*, ossia della forma

$$\ell(\mathbf{x}) := v_1x_1 + \cdots + v_nx_n = \sum_{j=1}^n v_jx_j = \mathbf{v}^T \mathbf{x} \quad (5.33)$$

ove  $\mathbf{v}$  è un vettore costante in  $\mathbb{R}^n$  e  $\mathbf{x} \in \mathbb{R}^n$  rappresenta un vettore variabile.

La programmazione lineare rappresenta, come mostreremo attraverso alcuni esempi, un interessante strumento di *modellizzazione matematica* in diversi campi applicativi. In questo paragrafo ci limiteremo ad *analizzare le idee di base*. In particolare, introdurremo il metodo numerico del *simplexso*, che rappresenta l'estensione del metodo di eliminazione di Gauss al caso di disequazioni lineari. Per una trattazione più adeguata dell'argomento rinviamo ad esempio a Dantzig [42], Luenberger [109] e Zoutendijk [157].

Introdurremo le principali idee del metodo attraverso un esempio semplice, per il quale è possibile una interpretazione grafica; tale interpretazione non è invece praticabile nei problemi reali, usualmente caratterizzati da un numero elevato di variabili e di vincoli.

► **Esempio 5.16** Di una determinata *risorsa* si utilizza una quantità, complessivamente non superiore a una limitazione fissata  $b_3$ , in due modi differenti. Si suppone, inoltre, che le quantità di risorsa da dedicare a ciascun modo di impiego, denotate rispettivamente con  $x_1, x_2$ , siano inferiori a  $b_1$  e a  $b_2$ , con  $b_1 + b_2 \leq b_3$ .

Indicando con  $c_1, c_2$  i *profitti* per unità di risorsa impiegata, si cercano le quantità  $x_1, x_2$  che forniscono il *massimo profitto*<sup>7</sup>. Dal punto di vista matematico si tratta di trovare il

<sup>7</sup>Come esempi di applicazione, si pensi all'utilizzo ottimale delle risorse di un calcolatore (CPU,

massimo della funzione  $f(x_1, x_2) := c_1x_1 + c_2x_2$ , con i seguenti vincoli

$$\begin{aligned} x_1, x_2, b_1, b_2, b_3 &\geq 0 \\ 0 \leq x_1 \leq b_1, \quad 0 \leq x_2 \leq b_2, \quad 0 \leq x_1 + x_2 \leq b_3 \end{aligned}$$

Posto

$$\mathbf{x} = [x_1, x_2]^T; \quad \mathbf{b} = [b_1, b_2, b_3]^T; \quad \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}; \quad \mathbf{c} = [c_1, c_2]^T$$

il problema può essere scritto nella seguente forma, nota come *forma primale* (primal form)

$$\boxed{\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^2} \quad & \mathbf{c}^T \mathbf{x} \\ & \mathbf{A} \mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned}} \quad (5.34)$$

ove  $\mathbf{c}^T \mathbf{x}$  è la *funzione costo*, o funzione obiettivo, e la matrice  $\mathbf{A}$ , di dimensione  $m \times n$  (in questo caso  $m = 3, n = 2$ ) è chiamata la *matrice dei vincoli* (constraint matrix). Osserviamo che mediante l'introduzione della matrice ampliata  $[\mathbf{A}, -\mathbf{I}]^T$  è possibile, naturalmente, incorporare in un'unica relazione matriciale anche i vincoli  $\mathbf{x} \geq 0$ ; tuttavia, per tradizione si preferisce scrivere in maniera separata tali vincoli, tenendo conto anche della loro maggiore semplicità. Il modello è illustrato *graficamente* in Figura 5.27.

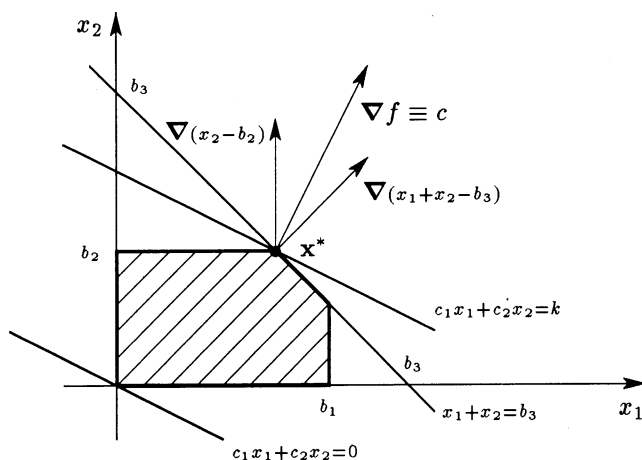


Figura 5.27: Rappresentazione grafica del problema di programmazione lineare introdotto nell'Esempio 5.16.

La zona tratteggiata indica l'*insieme di ammissibilità* (feasible region), ossia l'insieme  $\Omega$  dei punti  $\mathbf{x} \in \mathbb{R}^2$  che verificano i vincoli assegnati

$$\Omega := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A} \mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq 0\} \quad (5.35)$$

I/O, ecc.); oppure, in radioterapia, alla distribuzione ottimale della radiazione a cellule *sane* e *neoplastiche*, e in chimica alla formazione di miscele ottimali (cfr. il successivo Esempio 5.17).



Tale insieme risulta *vuoto* quando le disuguaglianze  $\mathbf{Ax} \leq \mathbf{b}$  e  $\mathbf{x} \geq 0$  sono *inconsistenti*, ossia quando non esiste nessun vettore  $\mathbf{x}$  che verifica l'insieme delle disuguaglianze. Nel caso particolare che stiamo considerando, tale eventualità si verifica se, ad esempio, si assume  $b_3 < 0$ . Al contrario, quando, ad esempio,  $b_3 = b_2 = \infty$ , allora l'insieme  $\Omega$  è *non limitato*.

Quando la regione di ammissibilità  $\Omega$  definita in (5.35) è non vuota, essa ha la proprietà importante della *convessità*, ossia per ogni coppia di punti  $\mathbf{x}$  e  $\mathbf{y}$  in  $\Omega$  e ogni scalare  $\mu$  tale che  $0 \leq \mu \leq 1$  il punto  $\mathbf{z}$  definito da

$$\mathbf{z} = (1 - \mu)\mathbf{x} + \mu\mathbf{y} \quad (5.36)$$

appartiene ancora a  $\Omega$ ; in altre parole, il segmento che ha come estremi i punti  $\mathbf{x}$  e  $\mathbf{y}$  appartiene all'insieme  $\Omega$ . La proprietà segue immediatamente dal seguente risultato

$$\mathbf{Az} = (1 - \mu)\mathbf{Ax} + \mu\mathbf{Ay} \leq (1 - \mu)\mathbf{b} + \mu\mathbf{b} = \mathbf{b} \Rightarrow \mathbf{Az} \leq \mathbf{b}$$

Più in generale, dati  $K$  punti  $\{\mathbf{x}_k\}$ ,  $k = 1, 2, \dots, K$ , in  $\Omega$ , appartiene a  $\Omega$  ogni loro *combinazione convessa*, definita come il seguente insieme di punti

$$\mathbf{z} = \sum_{k=1}^K \gamma_k \mathbf{x}_k, \quad \gamma_k \geq 0, \quad \sum_{k=1}^K \gamma_k = 1 \quad (5.37)$$

Per terminare l'analisi dell'insieme di ammissibilità  $\Omega$ , ricordiamo che un particolare vincolo  $a_i^T \mathbf{x} \leq \beta_i$  è detto *attivo* in un punto fissato  $\bar{\mathbf{x}}$ , quando si ha  $a_i^T \bar{\mathbf{x}} = \beta_i$ , ossia quando il vincolo è verificato con il segno di uguaglianza. Come si vede dalla Figura 5.27, i punti di ammissibilità nei quali è attivo almeno un vincolo costituiscono la frontiera dell'insieme  $\Omega$ . L'*insieme attivo* in un punto  $\bar{\mathbf{x}}$  è l'insieme dei vincoli che sono attivi nel punto  $\bar{\mathbf{x}}$ . Quando in un punto  $\mathbf{v}$  si hanno attivi esattamente  $n$  vincoli *linearmente indipendenti*, il punto  $\mathbf{v}$  è detto un *vertice non degenerato*<sup>8</sup>.

I vertici giocano un ruolo fondamentale nei problemi LP, in quanto la soluzione, se esiste, è in almeno uno dei vertici. In termini di matrice, se l'insieme dei vincoli è dato da  $\bar{\mathbf{A}}\mathbf{x} \leq \mathbf{b}$ , con  $\bar{\mathbf{A}}$  matrice di ordine  $\bar{m} \times n$ , l'esistenza di un vertice equivale al fatto che la sottomatrice di  $\bar{\mathbf{A}}$  costituita dalle righe corrispondenti ai vincoli attivi nel vertice  $\mathbf{v}$  ha rango  $n$ . In particolare, quindi, per l'esistenza di un vertice deve essere  $\bar{m} \geq n$  e il numero dei possibili vertici è limitato dalla quantità  $\binom{\bar{m}}{n}$ . Si può, inoltre, dimostrare facilmente il seguente risultato.

**Proposizione 5.2** *Consideriamo i vincoli  $\bar{\mathbf{A}}\mathbf{x} \leq \mathbf{b}$ , con  $\bar{\mathbf{A}}$  matrice di ordine  $\bar{m} \times n$ , con l'ipotesi che si abbia almeno un punto ammissibile. Se il rango della matrice  $\bar{\mathbf{A}}$  è  $n$ , ossia se la matrice contiene almeno un sottoinsieme di  $n$  righe linearmente indipendenti, allora esiste un vertice.*

Osserviamo che il risultato precedente è valido anche quando la regione di ammissibilità non è limitata. Come esemplificazione, si consideri l'insieme definito dai vincoli  $x_1 \geq 0$  e  $x_2 \geq 0$ , la cui matrice dei coefficienti ha rango 2; per essa l'origine è un vertice.

<sup>8</sup>Formalmente, un vertice è un *punto estremo* della regione di ammissibilità  $\Omega$ , ossia un punto che non è combinazione convessa stretta di altri due punti ammissibili. Quando nel vertice sono attivi più di  $n$  vincoli, il vertice è detto *degenerato*. Come esemplificazione, si considerino i vertici relativi all'insieme definito in  $\mathbb{R}^2$  dalle disuguaglianze  $x_1 \geq 1/2$ ,  $x_2 \geq 1/2$ ,  $x_1 + x_2 \geq 1$ ,  $x_1 + 2x_2 \leq 4$  e  $4x_1 + x_2 \leq 8$ .

L'adiacenza di due vertici può essere definita nel modo seguente. Se  $\bar{\mathbf{A}}_1$  (rispettivamente  $\bar{\mathbf{A}}_2$ ) è la matrice dei vincoli attivi nel vertice  $\mathbf{v}_1$  (rispettivamente in  $\mathbf{v}_2$ ) i due vertici sono *adiacenti* quando almeno  $n - 1$  righe in  $\bar{\mathbf{A}}_1$  sono anche in  $\bar{\mathbf{A}}_2$  e almeno una riga in  $\bar{\mathbf{A}}_1$  non è in  $\bar{\mathbf{A}}_2$ . Nel caso in cui  $\mathbf{v}_1$  e  $\mathbf{v}_2$  siano vertici non degenerati, allora essi sono adiacenti se e solo se le due matrici corrispondenti differiscono esattamente per una riga.

Un punto  $\mathbf{x}^* \in \Omega$  è detto *punto di ottimalità*, quando, nel caso della ricerca di un massimo, si ha  $\mathbf{c}^T \mathbf{x}^* \geq \mathbf{c}^T \mathbf{x}$  per ogni  $\mathbf{x} \in \Omega$ . Per la individuazione di un punto di ottimalità è importante la considerazione delle *curve di livello* della funzione costo, che nell'esempio che stiamo considerando sono date dalle seguenti rette

$$\mathbf{c}^T \mathbf{x} = c_1 x_1 + c_2 x_2 = k \quad (5.38)$$

al variare di  $k \in \mathbb{R}$ . Nel caso generale, i punti che verificano la condizione  $\mathbf{c}^T \mathbf{x} = k$ , per un valore fissato di  $k$ , definiscono un *piano* se  $n = 3$ , e un *iperpiano* se  $n \geq 4$ . Per semplicità, useremo il termine iperpiano per indicare il caso generale. Ogni punto  $\bar{\mathbf{x}}$  tale che  $\mathbf{c}^T \bar{\mathbf{x}} = k$  è detto giacere sull'associato iperpiano, e gli iperpiani ottenuti al variare di  $k$  sono *paralleli* (come esemplificazione, si vedano in Figura 5.27 le rette  $\mathbf{c}^T \mathbf{x} = 0$  e  $\mathbf{c}^T \mathbf{x} = k$ ). La quantità  $k$  è chiamata il *valore* dell'iperpiano.

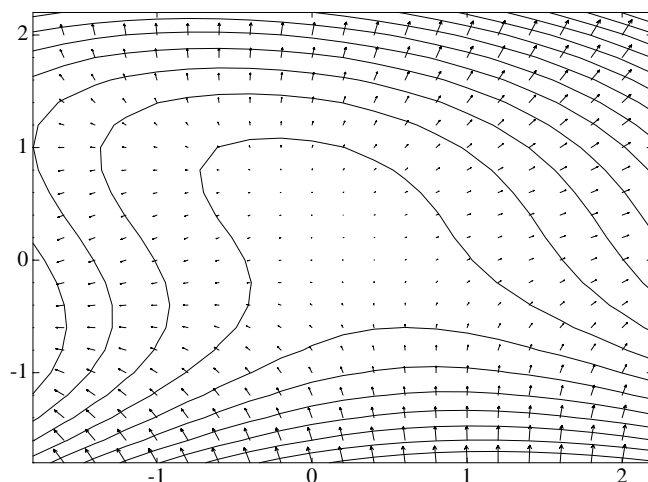


Figura 5.28: Curve di livello e vettore gradiente corrispondenti alla funzione  $f(\mathbf{x}) = x_1^2 + x_1 x_2 + x_2^3$ .

Ricordiamo che, data una generica funzione differenziabile  $f(\mathbf{x})$ , con  $\mathbf{x} \in \mathbb{R}^n$ , si chiama *vettore gradiente* di  $f(\mathbf{x})$  la seguente funzione a valori vettoriali

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (5.39)$$

Ad esempio, se  $f(\mathbf{x}) = x_1^2 + x_1x_2 + x_2^3$ , il suo gradiente è

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 3x_2^2 \end{bmatrix}$$

Tale vettore è rappresentato in Figura 5.28, insieme alle curve di livello. Come si vede dalla figura, il gradiente di una generica funzione non lineare varia con l'argomento  $\mathbf{x}$ ; inoltre, in ogni punto  $\mathbf{x}$  la direzione del vettore  $\nabla f(\mathbf{x})$  è la direzione di massima ascesa. Nel caso di una generica funzione lineare  $\ell(\mathbf{x}) = \mathbf{v}^T \mathbf{x}$ , si ha che la derivata parziale di  $\ell$  rispetto alla variabile  $x_j$  è semplicemente la componente  $j$ -ma di  $\mathbf{v}$ , cioè lo scalare  $v_j$ . Il vettore gradiente di  $\ell(\mathbf{x})$  è quindi il vettore  $\mathbf{v}$  per ogni scelta dell'argomento  $\mathbf{x}$ . Tale vettore è anche chiamato *vettore normale* di  $\ell$ , in quanto esso risulta *ortogonale* ad ogni retta che congiunge due qualunque punti sull'iperpiano  $\mathbf{v}^T \mathbf{x} = k$  (cioè è la direzione della normale all'iperpiano). Tale osservazione risulta dal fatto che, se  $\mathbf{y}$  e  $\mathbf{z}$  sono due punti nell'iperpiano, allora  $\mathbf{v}^T \mathbf{y} = \mathbf{v}^T \mathbf{z} = k$ , da cui  $\mathbf{v}^T (\mathbf{y} - \mathbf{z}) = 0$ .

Ritornando al problema illustrato in Figura 5.27, si vede che ogni direzione  $\mathbf{p}$ , con  $\mathbf{p}^T \mathbf{c} > 0$  è una direzione di ascesa per la funzione costo e la direzione del gradiente  $\nabla(\mathbf{c}^T \mathbf{x}) = \mathbf{c}$  è la direzione di *massima ascesa*. La soluzione ottimale  $\mathbf{x}^*$  è ottenuta, pertanto, nel vertice indicato. Se, ad esempio, poniamo:  $c_1 = 2$ ,  $c_2 = 4$ , la soluzione è ottenuta nel punto:  $\mathbf{x}^* = [b_3 - b_2, b_2]^T$ , che corrisponde ad investire il massimo possibile nel secondo modo, in corrispondenza al quale si ha il doppio di rendimento, e il resto nell'altro modo.

L'analisi grafica dell'esempio, al variare delle possibili situazioni corrispondenti a differenti scelte dei numeri  $b_1, b_2, b_3$ , suggerisce alcuni importanti risultati, che hanno in effetti una validità più generale.

### Risultati intuitivi

1. Se  $\Omega = \emptyset$ , non esistono soluzioni ammissibili e quindi soluzioni ottimali.
2. Se  $\Omega$  non è limitato lungo la direzione  $\mathbf{c}$ , non esistono soluzioni finite, cioè

$$\sup_{\mathbf{x} \in \Omega} \mathbf{c}^T \mathbf{x} = \infty$$

3. Se vi è un solo massimo, esso è sempre in un vertice.
4. Se vi è più di un massimo, allora ve ne sono infiniti e tra questi vi è almeno un vertice (nell'esempio il caso  $c_1 = c_2$ ).
5. Un punto  $\mathbf{x}^* \in \Omega$  è ottimale *se e solo se* la retta  $\mathbf{c}^T \mathbf{x} = \mathbf{c}^T \mathbf{x}^*$  non interseca l'interno di  $\Omega$ . Questo risultato può essere espresso nella seguente forma equivalente. Consideriamo, come indicato in Figura 5.27, i gradienti dei vincoli attivi in  $\mathbf{x}^*$ , ossia i vettori normali che individuano il vertice  $\mathbf{x}^*$ . Allora,  $\mathbf{x}^*$  è un punto di ottimalità se e solo se il gradiente  $\mathbf{c}$  della funzione costo appartiene al *cono*  $K(\mathbf{x}^*)$  individuato dai gradienti dei vincoli attivi, ossia se esso è una loro *combinazione positiva*. Questo risultato rappresenta l'estensione al caso di vincoli con disuguaglianze del risultato di Lagrange relativo alle condizioni necessarie per l'ottimalità di un problema con vincoli di uguaglianze (cfr. Capitolo 14).

Sottolineiamo, infine, che nell'esempio considerato, ma più in generale per ogni problema di programmazione lineare, un *ottimo locale* è anche un *ottimo globale*. ■

◆ **Esercizio 5.32** Si consideri l'insieme di ammissibilità definito dai vincoli  $x_1 + 2x_2 \geq 6$ ,  $2x_1 + x_2 \geq 6$ ,  $x_1 \geq 0$  e  $x_2 \geq 0$ . Esaminare, quindi, i problemi LP relativi alla minimizzazione delle seguenti funzioni costo a)  $x_1 + x_2$ ; b)  $3x_1 + x_2$ ; c)  $x_1 - x_2$ .

◆ **Esercizio 5.33** Massimizzare  $x_1 + x_2$  con i vincoli  $x_1 \geq 0$ ,  $x_2 \geq 0$ ,  $-3x_1 + 2x_2 \leq -1$ ,  $x_1 - x_2 \leq 2$ .

◆ **Esercizio 5.34** Si supponga di avere a disposizione in un laboratorio due differenti contatori di batteri, il primo dei quali sia in grado di valutare 6 campioni ogni ora e il secondo 10 campioni. Supponendo che il costo orario dei due contatori sia, rispettivamente, 20000 e 50000, si calcoli il tempo di utilizzo di ciascun contatore per stimare 1000 campioni al minimo costo, con il vincolo che nessuno dei due superi le 80 ore.

### 5.4.1 Trasformazione di problemi LP nella prima forma primale

Un problema di programmazione lineare formulato in una forma differente dalla forma primale (5.34) può essere riformulato in tale forma utilizzando le seguenti tecniche.

1. Si passa da un problema di minimo di  $\mathbf{c}^T \mathbf{x}$  a un problema di massimo, considerando il massimo di  $(-\mathbf{c})^T \mathbf{x}$ .
2. Un vincolo del tipo  $\mathbf{a}^T \mathbf{x} \geq \beta$  può essere sostituito dal vincolo equivalente  $(-\mathbf{a})^T \mathbf{x} \leq -\beta$ .
3. Il punto di ottimalità  $\mathbf{x}^*$  è indipendente dal valore della costante  $\lambda$  nella funzione costo  $\mathbf{c}^T \mathbf{x} + \lambda$ .
4. Un vincolo di uguaglianza  $\mathbf{a}^T \mathbf{x} = \beta$  è equivalente ai due vincoli  $\mathbf{a}^T \mathbf{x} \leq \beta$  e  $\mathbf{a}^T \mathbf{x} \geq \beta$ .
5. Se non è presente un vincolo di positività per una variabile  $x_i$ , si può sostituire  $x_i$  mediante la differenza di due variabili non negative,  $x_i = u_i - v_i$ , con  $u_i \geq 0$  e  $v_i \geq 0$ .

Come illustrazione delle tecniche precedenti, il seguente problema LP

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^3} \quad & 2x_1 + 3x_2 - x_3 + 4 \\ & x_1 - x_2 + 4x_3 \geq 2 \\ & x_1 + x_2 + x_3 = 15 \\ & x_2 \geq 0, x_3 \leq 0 \end{aligned}$$

può essere trasformato nella seguente forma equivalente

$$\begin{aligned} \min_{u,v,z,w} \quad & -2u + 2v - 3z - w \\ & -u + v + z + 4w \leq -2 \\ & u - v + z - w \leq 15 \\ & -u + v - z + w \leq -15 \\ & u \geq 0, v \geq 0, z \geq 0, w \geq 0 \end{aligned}$$

▼ **Osservazione 5.3** *Problemi che contengono valori assoluti delle variabili o valori assoluti di espressioni lineari possono essere spesso trasformati in problemi di programmazione lineare. Come illustrazione, consideriamo il problema di minimizzare  $|x_1 - x_2|$  sotto condizioni lineari su  $x_1$  e  $x_2$ . Si può introdurre una nuova variabile  $x_3 \geq 0$  e allora imporre i vincoli  $x_1 - x_2 \leq x_3$ ,  $-x_1 + x_2 \leq x_3$ . Si cerca, quindi, di minimizzare la forma lineare  $0x_1 + 0x_2 + 1x_3$ . ■*

### 5.4.2 Problema duale

Ad ogni problema LP formulato nella forma primale si può associare un altro problema chiamato il suo *duale*. Più precisamente, dato il problema

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^n} \quad & \mathbf{c}^T \mathbf{x} \\ & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned} \tag{5.40}$$

il corrispondente problema duale è definito nel seguente modo

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^m} \quad & \mathbf{b}^T \mathbf{u} \\ & \mathbf{A}^T \mathbf{u} \geq \mathbf{c} \\ & \mathbf{u} \geq 0 \end{aligned} \tag{5.41}$$

Ad esempio, il duale del problema

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^2} \quad & 2x_1 + 3x_2 \\ & 4x_1 + 5x_2 \leq 6 \\ & 7x_1 + 8x_2 \leq 9 \\ & 10x_1 + 11x_2 \leq 12 \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

è dato da

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^3} \quad & 6u_1 + 9u_2 + 12u_3 \\ & 4u_1 + 7u_2 + 10u_3 \geq 2 \\ & 5u_1 + 8u_2 + 11u_3 \geq 3 \\ & u_1 \geq 0, u_2 \geq 0, u_3 \geq 0 \end{aligned}$$

Formalmente, si passa dal problema primale al duale scambiando  $\mathbf{c}$  con  $\mathbf{b}$ , cambiando il segno della disequazione nei vincoli, trasponendo la matrice dei vincoli  $\mathbf{A}$  e passando da un problema di massimo a un problema di minimo. In questo modo, il numero delle disequazioni nel problema primale diventa il numero delle variabili nel problema duale, e quindi il problema duale ha dimensioni diverse da quello originale. Si vede anche che dualizzando il problema duale si ritorna al problema primale.

La considerazione del problema duale è interessante sotto vari aspetti. L'aspetto applicativo sarà illustrato nel paragrafo successivo. Dal punto di vista teorico si può

vedere che le variabili del problema duale possono essere interpretate come *moltiplicatori di Lagrange* del problema primale. Della importante teoria della dualità ci limiteremo a richiamare i seguenti risultati.

**Proposizione 5.3** *Se  $\mathbf{x}$  verifica i vincoli del problema primario e  $\mathbf{u}$  soddisfa i vincoli del duale, allora  $\mathbf{c}^T \mathbf{x} \leq \mathbf{b}^T \mathbf{u}$ . Di conseguenza, se  $\mathbf{c}^T \mathbf{x}^* = \mathbf{b}^T \mathbf{u}^*$ , allora  $\mathbf{x}^*$  e  $\mathbf{u}^*$  sono soluzioni ottimali, rispettivamente, del problema primale e duale.*

La dimostrazione segue immediatamente dal fatto che per ipotesi si ha  $\mathbf{x} \geq 0$ ,  $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ ,  $\mathbf{u} \geq 0$  e  $\mathbf{A}^T \mathbf{u} \geq \mathbf{c}$ , e quindi

$$\mathbf{c}^T \mathbf{x} \leq (\mathbf{A}^T \mathbf{u})^T \mathbf{x} = \mathbf{u}^T \mathbf{A}\mathbf{x} \leq \mathbf{u}^T \mathbf{b} = \mathbf{b}^T \mathbf{u}$$

Come esempio illustrativo, si consideri il seguente problema particolare

$$\begin{array}{ll} \text{(primale)} & \sum_{\mathbf{x} \in \mathbb{R}^2} x_1 + 4x_2 \\ & 2x_1 + x_2 \geq 6 \\ & 5x_1 + 3x_2 \geq 7 \\ & x_1 \geq 0, x_2 \geq 0 \end{array} \quad \begin{array}{ll} \text{(duale)} & \sum_{\mathbf{y} \in \mathbb{R}^2} 6y_1 + 7y_2 \\ & 2y_1 + 5y_2 \leq 1 \\ & y_1 + 3y_2 \leq 4 \\ & y_1 \geq 0, y_2 \geq 0 \end{array}$$

La scelta  $x_1 = 3$  e  $x_2 = 0$  è ammissibile, con costo  $x_1 + 4x_2 = 3$ . Nel problema duale  $y_1 = 1/2$  e  $y_2 = 0$  fornisce lo stesso valore  $6y_1 + 7y_2 = 3$ . Quindi i due vettori devono essere ottimali.

**Proposizione 5.4 (dualità)** *Se il problema originale ha una soluzione ottimale  $\mathbf{x}^*$ , allora il problema duale ha una soluzione ottimale  $\mathbf{u}^*$ , con  $\mathbf{c}^T \mathbf{x}^* = \mathbf{b}^T \mathbf{u}^*$ .*

Lasciamo come esercizio l'esame del problema duale, quando il problema primale ha la proprietà che  $\mathbf{c}^T \mathbf{x}$  è non limitata sull'insieme ammissibile.

### 5.4.3 Seconda forma primale

Mediante l'introduzione di nuove variabili, dette *variabili slack* è possibile trasformare alcune disequazioni nella formulazione primale (5.40) in equazioni. Come illustrazione, consideriamo il seguente problema

$$\begin{array}{ll} \max_{\mathbf{x} \in \mathbb{R}^2} & 2x_1 + 3x_2 \\ & 5x_1 + 3x_2 \leq 15 \\ & 3x_1 + 6x_2 \leq 18 \\ & x_1 \geq 0, x_2 \geq 0 \end{array}$$

Introducendo le variabili slack  $x_3, x_4$ , il problema può essere posto nella seguente forma *equivalente*

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^4} \quad & 2x_1 + 3x_2 + 0x_3 + 0x_4 \\ & 5x_1 + 3x_2 + x_3 = 15 \\ & 3x_1 + 6x_2 + x_4 = 18 \\ & x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0 \end{aligned}$$

In sostanza, mediante l'introduzione di opportune variabili il problema originale può essere posto nella seguente forma

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^{n+m}} \quad & \bar{\mathbf{c}}^T \mathbf{x} \\ & \bar{\mathbf{A}} \mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned} \tag{5.42}$$

ove  $\bar{\mathbf{c}}$  è il vettore in  $\mathbb{R}^{n+m}$  che ha le prime  $n$  componenti uguali alle componenti di  $\mathbf{c}$  e le rimanenti componenti nulle e la matrice  $\bar{\mathbf{A}}$  è data da  $[\mathbf{A}, \mathbf{I}]$ , con  $\mathbf{I}$  matrice identità di ordine  $m$ .

#### 5.4.4 Alcuni esempi applicativi

Per illustrare l'interesse applicativo della programmazione lineare esamineremo brevemente alcune situazioni *classiche*, che hanno in sostanza dato origine a tale teoria. Naturalmente, il contesto in cui tali situazioni verranno esaminate può essere diverso da quello indicato e lasciamo come *esercizio* trovare altre interessanti interpretazioni.

► **Esempio 5.17** Il problema è chiamato anche problema *menu-planning* o *optimal blending* e nasce quando, partendo da un insieme di componenti assegnate, si vuole ottenere una *miscela ottimale* rispettando opportuni vincoli.

Supponiamo ad esempio di avere  $n$  differenti cibi, con *prezzi unitari*  $c_1, c_2, \dots, c_n$ . Il problema consiste nel preparare una miscela a costo minimo, con il vincolo che siano rispettate determinate esigenze nutrizionali, cioè siano assicurate determinate quantità di calorie, vitamine, proteine, minerali, ecc.

Indichiamo con  $a_{ij}$  la quantità di nutrimento  $i$  nell'unità di cibo  $j$ , ove  $i = 1, \dots, m$ ;  $j = 1, \dots, n$ . Con  $b_i$  indichiamo la quantità minimale del nutrimento  $i$  e con  $x_j$  la quantità di cibo  $j$  da mettere nel menu. Il modello matematico assume, allora la seguente forma

$$\begin{aligned} \min \quad & \sum_{j=1}^n c_j x_j \\ & \sum_{j=1}^n a_{ij} x_j \geq b_i, \quad i = 1, \dots, m \\ & x_j \geq 0, \quad j = 1, \dots, n \end{aligned} \tag{5.43}$$

**Interpretazione duale** Il punto di vista adottato per stabilire il problema (5.43) è quello dell'*utilizzatore*. Guardiamo ora lo stesso problema dal punto di vista di un "venditore" di "pillole" di vitamine, calorie, ecc. Il suo problema è quello di stabilire il prezzo del

nutrimento  $i$ . Per essere competitivo, il prezzo  $u_i$  per unità dovrebbe essere tale che per ogni cibo  $j$  il prezzo unitario, calcolato a partire dal prezzo  $u_i$  delle sue componenti, non superi il prezzo di mercato  $c_j$ . Questo equivale ad imporre la condizione:  $\sum_{i=1}^m a_{ij}u_i \leq c_j$ .

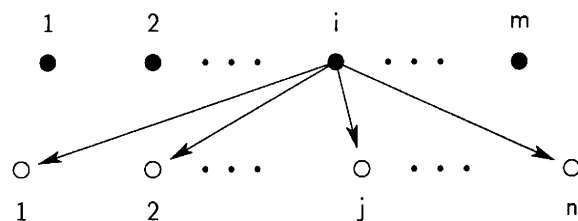
In conclusione, quindi, il problema per il venditore è il seguente

$$\begin{array}{l} \max \sum_{i=1}^m b_i u_i \\ \sum_{i=1}^m a_{ij} u_i \leq c_j, \quad j = 1, \dots, n \\ u_i \geq 0, \quad i = 1, \dots, m \end{array} \quad (5.44)$$

che rappresenta il *problema duale* del problema (5.43). Per il teorema di dualità, se  $x^*$  e  $u^*$  sono ottimali, allora  $\sum_{i=1}^m b_i u_i^* = \sum_{j=1}^n c_j x_j^*$  e quindi dal punto di vista economico (!) non ha importanza se si acquista un normale cibo o pillole. Dalla teoria della dualità si potrebbe anche vedere che se per un certo  $j$  si ha  $\sum a_{ij} u_i^* < c_j$ , allora si ha  $x_j^* = 0$ . Tale risultato si inquadra dal punto matematico nell'ambito della teoria dei moltiplicatori di Lagrange e dal punto di vista delle applicazioni si interpreta nel senso che il cibo  $j$  non ha sufficiente valore nutritivo per il suo prezzo e quindi non viene messo nel menu. Analogamente, se per un certo  $i$  si ha  $\sum a_{ij} x_j^* > b_i$ , allora  $u_i^* = 0$ , cioè la richiesta di nutrimento  $i$  è automaticamente soddisfatta, e quindi ridondante, e pertanto la corrispondente pillola non ha valore economico. ■

► **Esempio 5.18** Consideriamo il seguente classico problema di ottimizzazione del trasporto di un prodotto.

Un determinato prodotto deve essere spedito da  $m$  punti di stoccaggio a  $n$  differenti destinazioni. Supponiamo che al punto di stoccaggio  $i$  vi sia una disponibilità  $a_i$  del prodotto e che la richiesta alla destinazione  $j$  sia  $b_j$ ; supponiamo, inoltre, che il numero totale delle risorse sia almeno pari alle richieste.



Sia  $c_{ij}$  il costo di trasporto dell'unità di prodotto trasportato da  $i$  a  $j$  e  $x_{ij}$  la quantità di prodotto trasportato lungo questo cammino. Il problema è la determinazione di  $x_{ij}$  in modo che il *costo di trasporto* sia minimo. Il modello matematico sarà, allora, il seguente

$$\begin{array}{l} \min \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \\ \sum_{i=1}^m x_{ij} \geq b_j \quad j = 1, \dots, n \\ \sum_{j=1}^n x_{ij} \leq a_i \quad i = 1, \dots, m \\ x_{ij} \geq 0, \quad i = 1, \dots, m, \quad j = 1, \dots, n \end{array} \quad (5.45)$$



Per dualizzare il problema precedente possiamo pensare a una compagnia che compera il prodotto alla sorgente  $i$  per un prezzo  $u_i \geq 0$  e lo vende alla destinazione  $j$  ad un prezzo  $v_j \geq 0$ . Per essere competitiva, rispetto al trasporto in proprio, la compagnia deve fissare i prezzi in modo da avere  $v_j - u_i \leq c_{ij}$ . Per massimizzare il profitto sotto questi vincoli si deve risolvere il seguente problema

$$\boxed{\begin{array}{l} \max \sum_{j=1}^n b_j v_j - \sum_{i=1}^m a_i u_i \\ v_j - u_i \leq c_{ij}, u_i \geq 0, v_j \geq 0 \quad j = 1, \dots, n, i = 1, \dots, m \end{array}} \quad (5.46)$$

che corrisponde al *problema duale* di (5.45).

Molti altri problemi possono essere formulati come problemi di trasporto. Si pensi, ad esempio, nell'ambito dell'informatica, alla *gestione ottimale di reti di calcolatori*. Un altro esempio interessante è il *problema di assegnazione del personale*, nel quale si vuole assegnare  $n$  impiegati a  $n$  tipi diversi di lavoro. In base a test attitudinali o a prove di lavoro si conosce il beneficio  $c_{ij}$  alla compagnia, quando l'impiegato  $i$  è assegnato al lavoro  $j$ . Il problema è, allora, quello di distribuire gli impieghi in maniera da massimizzare il beneficio totale. In questo caso la variabile  $x_{ij}$  vale zero se l'impiegato  $i$  non è assegnato al lavoro  $j$  e uno nel caso contrario. L'insieme di ammissibilità è dato dalle relazioni  $\{\sum_j x_{ij} = 1, \sum_i x_{ij} = 1\}$ .

#### 5.4.5 Metodo del simplesso

Il *metodo del simplesso* è stato essenzialmente l'unico metodo disponibile per risolvere i problemi di programmazione lineare a partire dalla sua introduzione nel 1947 da parte di George B. Dantzig fino al 1984, quando sono state introdotte nuove idee basate sulla programmazione non lineare (*metodi interni*). Sebbene attualmente vi siano tecniche competitive, il metodo del simplesso rimane, tuttavia, ancora il metodo principale della moderna programmazione lineare.

In sostanza, il metodo del simplesso è una procedura sistematica per passare da un vertice assegnato ad un vertice adiacente in maniera che in tale passaggio si abbia un aumento nella funzione costo. Nel caso in cui l'insieme di ammissibilità non abbia vertici degenerati, e in assenza di errori di arrotondamento, il metodo termina dopo un numero finito di passaggi, limitato superiormente dal numero dei vertici ammissibili. Come abbiamo visto, tale numero può assumere il valore  $\binom{m}{n}$ , che è una funzione esponenziale delle dimensioni del problema. Tuttavia, nei problemi applicati si è osservato un numero di iterazioni dell'ordine di multipli (da 3 a 6) di  $m$  (numero dei vincoli). Osserviamo, comunque, che il caso peggiore è *reale*. In questo senso, segnaliamo in particolare gli esempi di Klee-Minty (1970), nei quali la regione ammissibile in  $\mathbb{R}^n$  contiene  $2^n$  vertici, e il metodo del simplesso esamina effettivamente tutti i vertici<sup>9</sup>. L'esistenza di tali esempi esclude che il metodo del simplesso

<sup>9</sup>Il modo più chiaro di visualizzare i vincoli di un esempio di Klee-Minty è probabilmente come deformazione di un ipercubo unitario. Sia  $\epsilon$ , con  $0 < \epsilon < 1/2$ . In  $n$  dimensioni, consideriamo la regione definita dai vincoli:  $\epsilon \leq x_1 \leq 1$ ,  $\epsilon x_{j-1} \leq x_j \leq 1 - \epsilon x_{j-1}$ , per  $j = 2, \dots, n$ , in modo che ogni successiva variabile sia limitata superiormente e inferiormente in termini della variabile precedente. Si può dimostrare che vi sono  $2^n$  vertici non degenerati; inoltre, nel caso in cui la funzione da

possa considerarsi efficiente (nel senso di richiedere un piccolo numero di iterazioni) su *tutti* i problemi di programmazione lineare. Una spiegazione matematica dell'efficienza *pratica* del metodo è ancora un problema aperto. Tale efficienza, tuttavia, rimane una osservazione sperimentale rilevata per molti anni nella risoluzione di importanti problemi applicativi.

Rinviando alla bibliografia per una trattazione più approfondita, daremo in questo paragrafo le idee di base del metodo, sufficienti per comprendere il funzionamento dei programmi di calcolo disponibili.

Consideriamo un problema di programmazione lineare nella seconda forma primale

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^n} \quad & \mathbf{c}^T \mathbf{x} \\ & \mathbf{A} \mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned} \tag{5.47}$$

ove si suppone che la matrice  $\mathbf{A}$ , di ordine  $m \times n$ , contenga una sottomatrice identità di ordine  $m$  nelle sue ultime  $m$  colonne, ossia della forma

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2k} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \\ a_{m1} & a_{m2} & \cdots & a_{mk} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

ove  $k = n - m$ . Si suppone inoltre che il vettore  $\mathbf{b} \in \mathbb{R}^m$  sia tale che  $\mathbf{b} \geq 0$ . L'insieme di ammissibilità è dato da

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A} \mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$$

Dalle ipotesi fatte segue che  $\Omega$  non è vuoto, in quanto è un punto ammissibile ad esempio il vettore  $x_1 = x_2 = \cdots = x_k = 0$  e  $x_{k+1} = b_1$ ,  $x_{k+2} = b_2$ , eccetera.

Indicando con  $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \dots, \mathbf{a}^{(n)}$  le colonne della matrice  $\mathbf{A}$ , il seguente risultato mette in relazione tali colonne con i vertici di  $\Omega$ .

**Proposizione 5.5** *Sia  $\mathbf{x} \in \Omega$  e  $\mathcal{I}(\mathbf{x}) := \{i \mid x_i > 0\}$ . Allora, le seguenti affermazioni sono equivalenti*

1.  $\mathbf{x}$  è un vertice di  $\Omega$ .
2. I vettori  $\{\mathbf{a}^{(i)} \mid i \in \mathcal{I}(\mathbf{x})\}$  sono linearmente indipendenti.

**DIMOSTRAZIONE.** Se l'affermazione 1 è falsa, allora si può scrivere  $\mathbf{x} = \frac{1}{2}(\mathbf{u} + \mathbf{v})$ , con  $\mathbf{u} \in \Omega, \mathbf{v} \in \Omega$  e  $\mathbf{u} \neq \mathbf{v}$ . Per ogni indice  $i$  non appartenente all'insieme  $\mathcal{I}(\mathbf{x})$ , si ha  $x_i = 0, u_i \geq 0, v_i \geq 0$  e  $x_i = \frac{1}{2}(u_i + v_i)$ . Ne segue che  $u_i$  e  $v_i$  devono essere nulli. Pertanto,

massimizzare sia data da  $x_n$ , il vertice ottimale è dato dal punto  $[\epsilon, \epsilon^2, \dots, \epsilon^{n-1}, 1 - \epsilon^n]^T$ .

tutte le componenti non nulle di  $\mathbf{u}$  e  $\mathbf{v}$  corrispondono a indici  $i$  in  $\mathcal{I}(\mathbf{x})$ . Poiché  $\mathbf{u}$  e  $\mathbf{v}$  appartengono a  $\Omega$ , si ha

$$\mathbf{b} = \mathbf{A}\mathbf{u} = \sum_{i=1}^n u_i \mathbf{a}^{(i)} = \sum_{i \in \mathcal{I}(\mathbf{x})} u_i \mathbf{a}^{(i)}, \quad \mathbf{b} = \mathbf{A}\mathbf{v} = \sum_{i=1}^n v_i \mathbf{a}^{(i)} = \sum_{i \in \mathcal{I}(\mathbf{x})} v_i \mathbf{a}^{(i)}$$

da cui

$$\sum_{i \in \mathcal{I}(\mathbf{x})} (u_i - v_i) \mathbf{a}^{(i)} = \mathbf{0}$$

e, quindi, i vettori  $\{\mathbf{a}^{(i)} \mid i \in \mathcal{I}(\mathbf{x})\}$  sono linearmente dipendenti e l'affermazione 2 è falsa. Di conseguenza, l'affermazione 2 implica l'affermazione 1.

In senso opposto, assumiamo che l'affermazione 2 sia falsa. Dalla dipendenza lineare delle colonne  $\mathbf{a}^{(i)}$  per  $i \in \mathcal{I}(\mathbf{x})$  si ha

$$\sum_{i \in \mathcal{I}(\mathbf{x})} y_i \mathbf{a}^{(i)} = \mathbf{0} \quad \text{con} \quad \sum_{i \in \mathcal{I}(\mathbf{x})} |y_i| \neq 0$$

con opportuni coefficienti  $y_i$ . Per ogni  $i \notin \mathcal{I}(\mathbf{x})$  poniamo  $y_i = 0$  e indichiamo con  $\mathbf{y}$  il vettore di componenti  $y_i$ . Allora, per ogni  $\lambda$  si vede che, dal momento che  $\mathbf{x} \in \Omega$

$$\mathbf{A}(\mathbf{x} \pm \lambda \mathbf{y}) = \sum_{i=1}^n (x_i \pm \lambda y_i) \mathbf{a}^{(i)} = \sum_{i=1}^n x_i \mathbf{a}^{(i)} \pm \lambda \sum_{i \in \mathcal{I}(\mathbf{x})} y_i \mathbf{a}^{(i)} = \mathbf{A}\mathbf{x} = \mathbf{b}$$

Scegliamo, ora, il numero reale  $\lambda$  positivo e sufficientemente piccolo in modo che  $\mathbf{x} \pm \lambda \mathbf{y} \geq \mathbf{0}$ . I vettori  $\mathbf{u} = \mathbf{x} + \lambda \mathbf{y}$  e  $\mathbf{v} = \mathbf{x} - \lambda \mathbf{y}$  appartengono a  $\Omega$ . Essi sono distinti e  $\mathbf{x} = \frac{1}{2}(\mathbf{u} + \mathbf{v})$ . Pertanto  $\mathbf{x}$  non è un vertice di  $\Omega$ , e l'affermazione 1 è falsa. Ne segue che l'affermazione 1 implica l'affermazione 2. ■

Come abbiamo già osservato in precedenza, l'algoritmo del semplice permette di valutare la funzione costo in successivi vertici adiacenti. Il risultato contenuto nella Proposizione (5.5) è allora importante, in quanto permette di operare su un'opportuna sottomatrice  $\mathbf{B}$ , di ordine  $m$ , della matrice dei vincoli  $\mathbf{A}$ . Il passaggio da un vertice ad un vertice adiacente è ottenuto mediante una sostituzione di una sola colonna della matrice  $\mathbf{B}$ .

A solo scopo di esemplificazione, riportiamo in maniera essenziale le operazioni richieste dal metodo ad ogni passaggio da un vertice al vertice adiacente, rinviando per una implementazione conveniente alla bibliografia.

**Metodo del semplice** Al generico passo, corrispondente ad un particolare vertice si ha un insieme di  $m$  indici  $\{k_1, k_2, \dots, k_m\}$ .

- 1 Si pongono le colonne  $\mathbf{a}^{(k_1)}, \mathbf{a}^{(k_2)}, \dots, \mathbf{a}^{(k_m)}$  in  $\mathbf{B}$  e si risolve  $\mathbf{B}\mathbf{x} = \mathbf{b}$ .
- 2 Se  $x_i > 0$  per  $1 \leq i \leq m$  si continua. Altrimenti l'algoritmo si arresta, in quanto il problema non ha soluzione, oppure vi sono vertici degenerati. Nel secondo caso l'algoritmo entra in ciclo senza un successivo miglioramento della funzione costo.

- 3 Si risolve  $\mathbf{B}^T \mathbf{y} = \mathbf{p}$ , con  $\mathbf{p} = [c_{k_1}, c_{k_2}, \dots, c_{k_m}]^T$ .
- 4 Si sceglie l'indice  $s$  in  $\{1, 2, \dots, n\}$ , ma non in  $\{k_1, k_2, \dots, k_m\}$  per il quale  $c_s - \mathbf{y}^T \mathbf{a}^{(s)}$  assume il valore più grande.
- 5 Se  $c_s - \mathbf{y}^T \mathbf{a}^{(s)} < 0$ , allora il vettore  $\mathbf{v} = [v_i]$ , con  $v_{k_i} = x_i$  per  $1 \leq i \leq m$  e  $v_i = 0$  per  $i \notin \{k_1, k_2, \dots, k_m\}$  è una soluzione ottimale. **Stop**.
- 6 Si risolve  $\mathbf{Bz} = \mathbf{a}^{(s)}$ .
- 7 Se  $z_i \leq 0$  per  $1 \leq i \leq m$ , allora **stop**, poiché la funzione costo non è limitata su  $\Omega$ .
- 8 Tra i rapporti  $x_i/z_i$ , con  $z_i > 0$  per  $1 \leq i \leq m$ , sia  $x_r/z_r$  il più piccolo.
- 9 Si sostituisce  $k_r$  con  $s$  e si ritorna al punto 1.

I passi 1, 3 e 6 richiedono la risoluzione di sistemi lineari con la *stessa* matrice dei coefficienti. Allora la fattorizzazione  $\mathbf{B} = \mathbf{LU}$  (o  $\mathbf{PB} = \mathbf{LU}$ , con  $\mathbf{P}$  opportuna matrice di permutazione corrispondente ad uno scambio di righe per motivi di stabilità numerica) porta direttamente alle soluzioni. Nei successivi passaggi non è necessario ricalcolare le matrici  $\mathbf{L}$  e  $\mathbf{U}$ , ma è possibile un loro conveniente *aggiornamento*.

#### 5.4.6 Risoluzione di sistemi lineari inconsistenti

Ricordiamo che un sistema lineare di  $m$  equazioni in  $n$  incognite

$$\sum_{j=1}^n a_{ij} x_j = b_i \quad 1 \leq i \leq m$$

è detto *inconsistente* se non esiste nessun vettore  $\mathbf{x} \in \mathbb{R}^n$  che verifica simultaneamente le  $m$  equazioni del sistema, ossia i *residui*

$$r_i := \sum_{j=1}^n a_{ij} x_j - b_i \quad 1 \leq i \leq m$$

non si annullano simultaneamente. I sistemi inconsistenti hanno origine in differenti applicazioni, in particolare nella valutazione di dati sperimentali mediante *modelli lineari*. Per un loro corretto utilizzo è essenziale una nuova definizione di *soluzione*, che rappresenti una opportuna estensione della definizione di soluzione classica. Tale definizione deve tenere conto del tipo di modello da cui nasce il sistema lineare e degli scopi dell'utilizzo del modello. Ad esempio, quando i dati sono affetti da errori casuali, con distribuzione gaussiana, è opportuna la definizione di soluzione secondo i *minimi quadrati* che abbiamo già analizzato in precedenza e che corrisponde al vettore che minimizza la somma dei quadrati dei residui  $\sum_{i=1}^m r_i^2$ .

Al contrario, quando i dati sono noti con accuratezza può essere più opportuna la ricerca del vettore  $\mathbf{x}$  che minimizza la quantità  $\max_{1 \leq i \leq m} |r_i|$  (problema  $\ell_\infty$ ), mentre quando si suppone che alcuni dati possano contenere errori anomali si può minimizzare la quantità  $\sum_{i=1}^n |r_i|$  (problema  $\ell_1$ ). Per risolvere tali problemi sono stati introdotti opportuni algoritmi. In questo paragrafo ci limiteremo, tuttavia, ad analizzare alcune semplici idee che permettono di ricondurre il problema della ricerca della soluzione dei problemi  $\ell_\infty$  e  $\ell_1$  alla risoluzione di opportuni problemi di programmazione lineare.

**Problema  $\ell_1$**  Il problema può essere riformulato nel seguente modo, mediante l'introduzione di  $m$  nuove variabili  $\epsilon_1, \epsilon_2, \dots, \epsilon_m$

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\epsilon} \in \mathbb{R}^m} \quad & - \sum_{i=1}^m \epsilon_i \\ & \sum_{j=1}^n a_{ij} x_j - b_i \leq \epsilon_i \quad 1 \leq i \leq m \\ & - \sum_{j=1}^n a_{ij} x_j + b_i \leq \epsilon_i \quad 1 \leq i \leq m \end{aligned} \quad (5.48)$$

**Problema  $\ell_\infty$**  Posto  $\epsilon = \max_{1 \leq i \leq m} |r_i|$  si ha il seguente problema di programmazione lineare equivalente

$$\begin{aligned} \max_{x \in \mathbb{R}^n, \epsilon \in \mathbb{R}} \quad & -\epsilon \\ & \sum_{j=1}^n a_{ij} x_j - \epsilon \leq b_i \quad 1 \leq i \leq m \\ & - \sum_{j=1}^n a_{ij} x_j - \epsilon \leq -b_i \quad 1 \leq i \leq m \end{aligned} \quad (5.49)$$

◆ **Esercizio 5.35** Dato il seguente sistema lineare inconsistente

$$\begin{cases} 5x_1 + 2x_2 & = 6 \\ x_1 + x_2 + x_3 & = 2 \\ & 7x_2 - 5x_3 = 11 \\ 6x_1 & + 9x_3 = 9 \end{cases}$$

scrivere i corrispondenti problemi di programmazione lineare per la risoluzione del sistema nel senso  $\ell_1$  e  $\ell_\infty$ , quando si impone l'ulteriore vincolo che le variabili siano non negative.

◆ **Esercizio 5.36** Dati i valori di una funzione  $f(x_i)$  in  $m$  punti  $x_i$ , formulare sotto forma di problema di programmazione lineare il problema della ricerca di un polinomio  $p_n(x)$  di grado al più  $n$ , con  $n \leq m - 1$ , che minimizza l'espressione  $\max_{1 \leq i \leq m} |f(x_i) - p_n(x_i)|$ .

## 5.5 Metodi di ottimizzazione

Come abbiamo visto in precedenza, un problema di ottimizzazione è caratterizzato da un insieme di ammissibilità  $\Omega$  e da una funzione obiettivo  $f(\mathbf{x})$ . Nel paragrafo precedente abbiamo analizzato la situazione particolare in cui la funzione  $f(\mathbf{x})$  è lineare e  $\Omega$  è descritto da disequazioni lineari. Vi sono, tuttavia, numerose applicazioni nelle quali è necessario, o opportuno, tenere conto della non linearità sia della funzione obiettivo  $f(\mathbf{x})$  che dei vincoli.

Come semplice esempio illustrativo, si consideri il problema del calcolo del raggio e dell'altezza di un recipiente di forma cilindrica, di volume  $V$  assegnato, in maniera che risulti minima l'area della superficie totale, e quindi sia minima la quantità di materiale usato per la sua costruzione. Chiamati  $x_1$  e  $x_2$  rispettivamente il raggio  $r$  e l'altezza  $h$  del cilindro si ha il seguente problema di ottimizzazione

$$\begin{cases} \min 2\pi x_1^2 + 2\pi x_1 x_2 \\ \mathbf{x} \in \Omega = \{\mathbf{x} \in \mathbb{R}^2 \mid \pi x_1^2 x_2 = V, x_1 \geq 0, x_2 \geq 0\} \end{cases} \quad (5.50)$$

che rappresenta un caso particolare di problema di ottimizzazione *vincolato* (constrained). Più in generale, l'insieme di ammissibilità può essere descritto da disequazioni non lineari, oppure da vincoli di natura diversa, ad esempio, dalla condizione che alcune variabili siano numeri interi (*problema di programmazione intera*) o variabili booleane  $\{0, 1\}$ . Nel caso particolare in cui  $\Omega \equiv \mathbb{R}^n$ , il problema viene detto *non vincolato* (unconstrained).

Ricordiamo che un punto di *minimo globale* di una funzione  $f(\mathbf{x})$  su  $\Omega$  è un punto  $\mathbf{x}^* \in \Omega$  tale che  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  per ogni  $\mathbf{x} \in \Omega$ , mentre  $\mathbf{x}^*$  è un punto di *minimo locale*, quando esiste un intorno  $I(\mathbf{x}^*)$  del punto  $\mathbf{x}^*$  tale che  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  per tutti gli  $\mathbf{x} \in I(\mathbf{x}^*)$ . Analoga definizione si ha per i punti di massimo<sup>10</sup>. I metodi che analizzeremo nel seguito sono metodi idonei a calcolare una approssimazione dei punti di ottimalità locali. Essi possono, comunque, essere utilizzati per il calcolo dei punti di ottimalità globali nel senso seguente. Quando da ulteriori informazioni sul problema, quali, ad esempio, quelle che provengono dai dati sperimentali, è possibile avere una stima del punto di ottimalità globale, le tecniche di ottimizzazione locali possono servire per ottenere un raffinamento di tale stima<sup>11</sup>.

<sup>10</sup>Osserviamo che una tecnica per *minimizzare* può essere direttamente utilizzabile come tecnica per *massimizzare*, dal momento che minimizzare  $f(\mathbf{x})$  è equivalente a massimizzare  $-f(\mathbf{x})$ .

<sup>11</sup>Osserviamo che, dal momento che l'insieme dei numeri macchina è finito, è possibile, almeno in teoria, calcolare numericamente un punto di ottimalità globale, quando l'insieme  $\Omega$  è limitato, con un numero finito di valutazioni della funzione  $f(\mathbf{x})$ . Per illustrare, comunque, la *non praticità* di tale idea, consideriamo il caso di una funzione  $f(x)$ , con  $x \in \mathbb{R}$  vincolata all'intervallo  $1 \leq x \leq 2$ . Se la precisione macchina utilizzata è  $\text{eps} = 10^{-16}$ , si dovrebbe valutare  $f(x)$  all'incirca per  $10^{16}$  valori  $x$  diversi. Supponendo che ogni calcolo di  $f(x)$  richieda  $10^{-7}$  secondi, sarebbero necessari  $10^9$  secondi, ossia un certo numero di anni!

Nel seguito considereremo, in particolare, problemi di ottimizzazione non vincolata. Il motivo principale di tale scelta è nel fatto che la trattazione di tali problemi è più semplice. In effetti, una trattazione adeguata dei problemi con vincoli richiederebbe un insieme di risultati il cui studio esula dagli scopi del presente volume.

Tuttavia, un secondo motivo per considerare più in dettaglio i problemi non vincolati sta nel fatto che mediante opportune trasformazioni alcuni dei problemi con vincoli possono essere ridotti a problemi non vincolati. Ad esempio, nel caso del problema (5.50), dal vincolo di uguaglianza si può ricavare la variabile  $x_2$  e ridurre il problema a un problema di minimo in una sola variabile.

Più in generale, segnaliamo le seguenti trasformazioni

$$\begin{aligned} x_i \geq 0 & \Rightarrow x_i = y_i^2, \quad x_i = e^{y_i}, \quad x_i = |y_i| \\ 0 \leq x_i \leq 1 & \Rightarrow x_i = \sin^2 y_i, \quad x_i = e^{y_i} / (e^{y_i} + e^{-y_i}) \\ b_i \leq x_i \leq u_i & \Rightarrow x_i = b_i + (u_i - b_i) \sin^2 y_i \\ 0 \leq x_i \leq x_j \leq x_k & \Rightarrow x_i = y_i^2, \quad x_j = y_i^2 + y_j^2, \quad x_k = y_i^2 + y_j^2 + y_k^2 \\ -1 \leq x_i \leq 1 & \Rightarrow x_i = \sin y_i \end{aligned}$$

Le trasformazioni precedenti operano sulle variabili; altri tipi di trasformazioni operano, invece, sulla funzione da minimizzare; in questa direzione segnaliamo, in particolare, i metodi basati sulle funzioni di *penalizzazione* e sulle funzioni di *barriera*, e i metodi che utilizzano i *moltiplicatori di Lagrange*.

### 5.5.1 Ottimizzazione unidimensionale

In questo paragrafo considereremo, sostanzialmente, due tipi di tecniche per minimizzare una funzione  $f(x)$ , con  $x \in \mathbb{R}$ . Nel primo tipo si usano esplicitamente i valori delle derivate, mentre nel secondo vengono usati solo i valori della funzione. In analogia a quanto abbiamo visto per i metodi relativi alle equazioni non lineari, i metodi che utilizzano le derivate sono più efficienti per quanto riguarda la rapidità di convergenza, ma la loro convergenza può dipendere in maniera essenziale dalla stima iniziale del punto di minimo.

Lo studio dei problemi unidimensionali è importante per diversi motivi. In effetti, oltre che naturalmente essere più semplici, essi possono avere un interesse specifico, in quanto modelli particolari, ed inoltre molte idee usate in più dimensioni possono essere illustrate più adeguatamente in una dimensione; ma, soprattutto, è da tenere presente che la maggior parte dei metodi in  $n$  dimensioni prevedono la risoluzione successiva di problemi unidimensionali.

A scopo illustrativo e introduttivo, in Figura 5.29 sono riportate due situazioni, in un certo senso opposte, nelle quali il problema di minimizzazione di una funzione in una variabile può presentare difficoltà. Il primo esempio, corrispondente alla

funzione  $f(x) = x^2 + \sin 53x$ , illustra, in maniera “esasperata”, la possibilità di esistenza di un alto numero di minimi locali. Il secondo esempio, corrispondente alla funzione  $f(x) = (x - 2)^4 - 9$ , rappresenta l’esistenza di un minimo “eccessivamente piatto”.

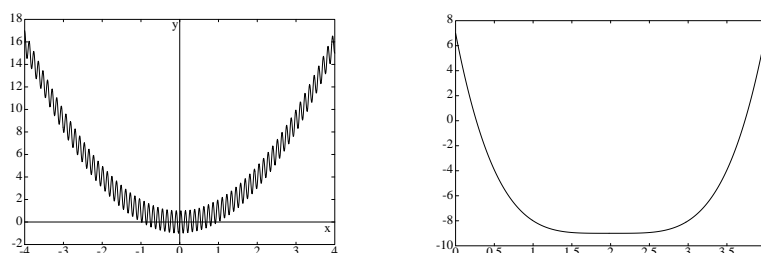


Figura 5.29: Situazioni che possono presentare difficoltà nel calcolo del minimo. Nella prima figura è rappresentata la funzione  $f(x) = x^2 + \sin 53x$ ; nella seconda la funzione  $f(x) = (x - 2)^4 - 9$ .

### Metodo di Newton

Sia  $f(x)$  una funzione definita e continua, insieme alle prime due derivate, su  $\mathbb{R}$ . Il metodo di Newton è un metodo iterativo, nel quale ad ogni passo dell’iterazione si sostituisce al problema del minimo della funzione  $f(x)$  quello del minimo di una opportuna funzione quadratica, della quale, come è noto, è possibile calcolare il minimo in maniera analitica. Più precisamente, sia  $x_k$  la stima corrente della soluzione  $x^*$ , e consideriamo il seguente sviluppo in serie di Taylor intorno al punto  $x_k$

$$f(x_k + p) = f(x_k) + pf'(x_k) + \frac{1}{2}p^2 f''(x_k) + \dots$$

Si ha, allora

$$\begin{aligned} f(x^*) &= \min_x f(x) = \min_p f(x_k + p) = \min_p \left[ f(x_k) + pf'(x_k) + \frac{1}{2}p^2 f''(x_k) + \dots \right] \\ &\approx \min_p \left[ f(x_k) + pf'(x_k) + \frac{1}{2}p^2 f''(x_k) \right] \end{aligned}$$

Il minimo  $p^*$  della forma quadratica, ottenuto derivando rispetto a  $p$  e ponendo la derivata uguale a zero, è dato dal seguente valore

$$p^* = -\frac{f'(x_k)}{f''(x_k)}$$

Il valore  $x_{k+1} = x_k + p^*$  fornisce una nuova stima del punto  $x^*$ . Si ha quindi il seguente procedimento iterativo, usualmente chiamato *metodo di Newton*

$$\boxed{x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}} \quad (5.51)$$



Osserviamo che la formula (5.51) è esattamente la stessa formula che abbiamo ottenuto nei paragrafi precedenti, applicando il metodo delle tangenti all'equazione non lineare

$$f'(x) = 0$$

Come illustrazione, il metodo di Newton è applicato alla funzione  $f(x) = \sin x - \cos x$ , con  $x_0 = -0.3$  (cfr. Figura 5.30). I valori ottenuti dalle successive iterazioni del

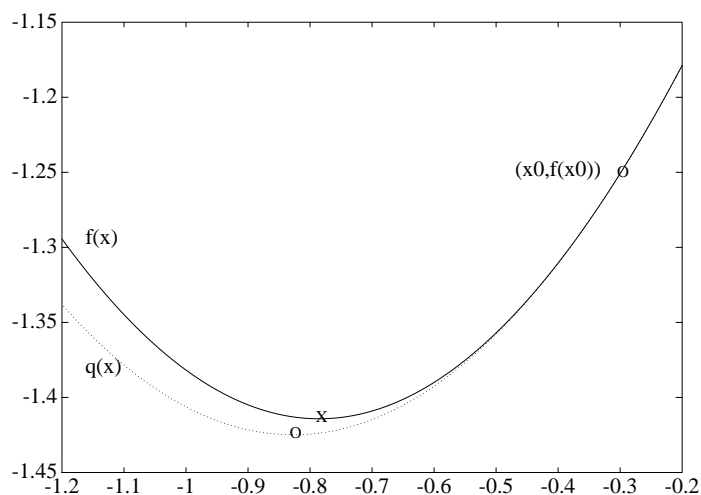


Figura 5.30: Illustrazione del metodo di Newton. La funzione  $q(x)$  corrisponde al polinomio di secondo grado  $q(x) = f(x_0) + (x - x_0)f'(x_0) + (x - x_0)^2 f''(x_0)/2$ , con  $x_0 = -0.3$ . Il punto  $x_1 = x_0 - f'(x_0)/f''(x_0)$  corrisponde al minimo del polinomio  $q(x)$ .

metodo sono forniti nella seguente tabella.

$k$	$x_k$
0	-0.3
1	-0.8274915
2	-0.7853732
3	-0.7853981
4	-0.7853981

Come si vede, la convergenza è rapida. In effetti, si può dimostrare che se  $x_k$  è sufficientemente vicino a  $x^*$  e se  $f''(x^*) > 0$ , allora

$$|x_{k+1} - x^*| \leq c |x_k - x^*|^2$$

ove  $c$  è una costante non negativa che dipende dalla funzione che si minimizza. In maniera schematica, questo significa che il numero delle cifre accurate in  $x_k$  si raddoppia ad ogni iterazione.

L'importanza dell'ipotesi  $f''(x^*) > 0$  è evidenziata dal seguente esempio. Consideriamo la funzione  $f(x) = (x - 2)^4 - 9$ , che ha un minimo nel punto  $x^* = 2$  (cfr. Figura 5.29). Si verifica facilmente che per tale funzione il metodo di Newton si riduce alla seguente formula

$$x_{k+1} = x_k - \frac{1}{3}(x_k - 2) \Rightarrow x_{k+1} - 2 = \frac{2}{3}(x_k - 2)$$

da cui si vede che la convergenza è di tipo *lineare*.

Naturalmente, senza ulteriori ipotesi sulla funzione  $f(x)$ , il metodo di Newton può essere non convergente. In effetti, non è detto che per ogni  $k$  si abbia  $f''(x_k) > 0$ , e quindi che la forma quadratica approssimante abbia un minimo. Inoltre, l'approssimazione della funzione  $f(x)$  mediante un polinomio di secondo grado può essere non sufficientemente adeguata, e di conseguenza non è detto che il punto  $x_{k+1}$  sia più vicino a  $x^*$  di  $x_k$ . Un ulteriore aspetto negativo, per quanto riguarda le applicazioni reali, del metodo di Newton è l'utilizzo esplicito delle derivate prime e seconde.

### Interpolazione parabolica

Nel metodo di Newton si costruisce la forma quadratica approssimante usando i valori della funzione e delle derivate in un punto. Alternativamente, la forma quadratica può essere ottenuta nel seguente modo. Si parte con tre valori  $x_1, x_2, x_3$  arbitrari e al passo generico si costruisce il polinomio di secondo grado che interpola i punti  $(x_i, f(x_i))$ ,  $i = k - 2, k - 1, k$ . Come punto  $x_{k+1}$  si assume il punto di minimo del polinomio ottenuto. Si continua, quindi, l'iterazione partendo da  $x_{k-1}, x_k$  e  $x_{k+1}$ . L'algoritmo ottenuto è detto algoritmo di *successiva interpolazione parabolica*. Si può dimostrare che l'iterazione converge con velocità  $\approx 1.324\dots$ , purché sia applicata sufficientemente vicino a  $x^*$  e  $f''(x^*) > 0$ .

### Algoritmo della sezione aurea

I metodi che considereremo in questo paragrafo non richiedono la conoscenza delle derivate di  $f(x)$ , e non corrispondono direttamente a metodi per la risoluzione di equazioni non lineari. Supporremo, inoltre, che la funzione  $f(x)$  abbia un unico minimo sull'intervallo di definizione  $[a, b]$ ; più precisamente, supporremo che la  $f(x)$  verifichi la seguente definizione.

**Definizione 5.2** (Funzioni unimodali) *Una funzione  $f(x) : \mathbb{R} \rightarrow \mathbb{R}$  è strettamente unimodale su  $[a, b]$  se esiste un  $x^* \in [a, b]$  tale che  $f(x^*) = \min f(x)$ , per  $x \in [a, b]$ , e se per ogni coppia di valori  $x_1, x_2$ , con  $a \leq x_1 < x_2 \leq b$*

$$\begin{cases} x_2 \leq x^* & \text{implica che } f(x_1) > f(x_2) \\ x^* \leq x_1 & \text{implica che } f(x_2) > f(x_1) \end{cases} \quad (5.52)$$

In altre parole, una funzione strettamente unimodale è strettamente decrescente per  $x \leq x^*$  e strettamente crescente per  $x^* \leq x$ . Con riferimento alla Figura 5.29, la prima curva non è unimodale, mentre è unimodale la seconda. Osserviamo, anche, che una funzione unimodale non è necessariamente continua e che è unimodale ogni funzione strettamente convessa (e quindi, in particolare, quando  $f''(x) > 0$ , per ogni  $x \in [a, b]$ ).

Per una funzione unimodale si ha la seguente proprietà importante. Per ogni coppia di punti  $x_1, x_2$  nell'intervallo con  $a \leq x_1 < x_2 \leq b$ , si ha

$$\begin{cases} f(x_1) > f(x_2) & \text{implica che } x^* \in [x_1, b] \\ f(x_1) = f(x_2) & \text{implica che } x^* \in [x_1, x_2] \\ f(x_1) < f(x_2) & \text{implica che } x^* \in [a, x_2] \end{cases} \quad (5.53)$$

La proprietà precedente può essere opportunamente utilizzata per ridurre l'intervallo nel quale si trova il punto  $x^*$  (il cosiddetto *intervallo di incertezza*).

Un primo modo semplice di procedere consiste nel considerare, ad esempio, i punti  $x_1 = (a + b)/2 - \epsilon$  e  $x_2 = (a + b)/2 + \epsilon$ , ove  $\epsilon$  è un numero positivo fissato, opportunamente piccolo. Dal confronto dei valori  $f(x_1)$  e  $f(x_2)$  si ha la possibilità di ridurre l'intervallo di incertezza iniziale  $[a, b]$  all'intervallo  $[a, (a + b)/2 + \epsilon]$ , o all'intervallo  $[(a + b)/2 - \epsilon, b]$ . La procedura che ne risulta è detta *metodo di bisezione*, in quanto per  $\epsilon$  piccolo l'intervallo di incertezza è approssimativamente dimezzato; in pratica, è l'estensione naturale del metodo che abbiamo visto in precedenza per il calcolo di una zero di una funzione continua con segno discorde agli estremi dell'intervallo.

Un aspetto negativo del metodo della bisezione è il fatto che ad ogni passaggio esso richiede il calcolo di *due* nuovi valori della funzione  $f(x)$ . In altre parole, se, ad esempio, l'intervallo ottenuto è  $[a, x_2]$ , il valore  $f(x_1)$  non viene ulteriormente utilizzato.

In effetti, è possibile scegliere i successivi punti  $x_i$ , in maniera da richiedere ad ogni passaggio *un solo* valore della funzione. Ricordiamo in questo senso il metodo basato sull'utilizzo dei numeri della successione di Fibonacci. Tale metodo è ottimale nel senso che per un assegnato numero di valutazioni della funzione esso riduce alla minima lunghezza l'intervallo di incertezza.

Sia

$$\tau_{k+1} = \tau_k + \tau_{k-1}, \quad \tau_0 = \tau_1, \quad k = 1, 2, \dots \quad (5.54)$$

la *successione di Fibonacci* e  $M$  il numero previsto di riduzioni dell'intervallo. Indicando con  $a^k, b^k$ , con  $a^0 = a$  e  $b^0 = b$ , i successivi intervalli di incertezza, i punti in cui valutare successivamente la funzione sono dati dalle seguenti equazioni

$$\begin{aligned} x_1^{k+1} &= (\tau_{M-1-k}/\tau_{M+1-k})(b^k - a^k) + a^k \\ x_2^{k+1} &= (\tau_{M-k}/\tau_{M+1-k})(b^k - a^k) + a^k \end{aligned} \quad k = 0, 1, \dots, M-2 \quad (5.55)$$

Si avrà, quindi, successivamente

$$\begin{aligned} a^{k+1} &= a^k, \quad b^{k+1} = x_2^{k+1}, \quad \text{se } f(x_1^{k+1}) < f(x_2^{k+1}) \\ a^{k+1} &= x_1^{k+1}, \quad b^{k+1} = b^k, \quad \text{se } f(x_1^{k+1}) \geq f(x_2^{k+1}) \end{aligned} \quad (5.56)$$

per  $k = 0, 1, \dots, M-2$ . Per  $k = M-1$ , ossia quando i due punti definiti in (5.55) coincidono, si può assumere

$$x_1^M = \frac{1}{2}(a^{M-1} + b^{M-1}) - \epsilon, \quad x_2^M = \frac{1}{2}(a^{M-1} + b^{M-1}) + \epsilon$$

per  $\epsilon$  opportunamente piccolo. È facile dimostrare, per  $k = 1, 2, \dots, M-2$ , il seguente risultato

$$\begin{aligned} x_1^{k+1} &= a^k + (x_2^k - x_1^k), \quad x_2^{k+1} = x_1^k, \quad f(x_1^k) < f(x_2^k) \\ x_1^{k+1} &= x_2^k, \quad x_2^{k+1} = b^k - (x_2^k - x_1^k), \quad f(x_1^k) \geq f(x_2^k) \end{aligned}$$

dal quale si vede che, in effetti, il procedimento utilizza una sola valutazione per ogni passo, salvo per  $k = 0$ . Infine, si può dimostrare che

$$b^{k+1} - a^{k+1} = \begin{cases} (\tau_{M-k}/\tau_{M+1-k})(b^k - a^k) & \text{per } k \leq M-2 \\ \frac{1}{2}(b^{M-1} - a^{M-1}) + \epsilon & \text{per } k = M-1 \end{cases}$$

che implica la relazione

$$b^M - a^M = \frac{b - a}{2\tau_{M+1}} + \epsilon$$

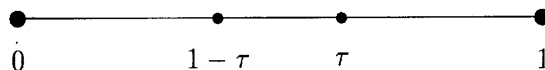
dalla quale è possibile determinare il valore di  $M$  necessario per ottenere una lunghezza desiderata dell'intervallo di incertezza.

Dal procedimento precedente si può ottenere una variante più conveniente dal punto di vista computazionale e che mantiene, asintoticamente, ossia per  $M$  grande, la proprietà di ottimalità nei riguardi del numero delle valutazioni. Dalla definizione (5.54) si ottiene

$$1 = \frac{\tau_k}{\tau_{k+1}} + \frac{\tau_{k-1}}{\tau_k} \frac{\tau_k}{\tau_{k+1}}$$

dalla quale, posto  $\tau = \lim_{k \rightarrow \infty} \tau_k/\tau_{k+1}$ , si ha

$$1 - \tau = \tau^2$$



La quantità  $\tau = (\sqrt{5} - 1)/2$  è detta *sezione aurea*<sup>12</sup> dell'intervallo  $[0, 1]$ .

Modificando, allora, le formule (5.55) nel seguente modo

$$\begin{aligned} x_1^{k+1} &= \tau^2 (b^k - a^k) + a^k \\ x_2^{k+1} &= \tau (b^k - a^k) + a^k \end{aligned} \quad k = 0, 1, \dots, M - 2 \quad (5.57)$$

si ottiene il cosiddetto *metodo della sezione aurea*, nel quale l'intervallo di incertezza ha un fattore di riduzione costante, corrispondente a  $\tau$ .

Nel seguito è riportata una interessante implementazione (dovuta a Brent, 1973) del metodo della sezione aurea, combinato con il metodo della interpolazione parabolica visto in precedenza.

```

DOUBLE PRECISION FUNCTION FMIN(AX,BX,F,TOL)
DOUBLE PRECISION AX,BX,F,TOL
C   calcola una approssimazione del punto in cui f(x) raggiunge
C   il minimo sull'intervallo [AX,BX].
C input..
C AX   primo estremo dell'intervallo iniziale
C BX   secondo estremo dell'intervallo iniziale
C F    sottoprogramma che calcola f(x) per ogni x dell'intervallo (AX,BX)
C TOL  lunghezza desiderata dell'intervallo di incertezza del risultato finale
C output..
C FMIN ascissa che approssima il punto ove f raggiunge il minimo
C
C   il metodo utilizza una combinazione del metodo della sezione aurea
C   e dell'interpolazione parabolica.
C   La convergenza non e' piu' lenta di quella relativa
C   al metodo di Fibonacci. Se f e' continua insieme alle
C   derivate prima e seconda, la convergenza e' superlineare.
C
C   La funzione f non e' mai calcolata in due punti distanti
C   meno di eps*abs(fmin)+(tol/3), ove eps e' approssimativamente
C   la radice quadrata della precisione macchina.
C   Se f e' unimodale e anche i valori calcolati sono unimodali
C   allora FMIN approssima l'ascissa del minimo globale di f
C   sull'intervallo (AX,BX) con un errore minore di 3*eps*abs(fmin)+tol
C.....

```

<sup>12</sup>La quantità  $\tau$  è la media proporzionale tra la lunghezza dell'intervallo e la parte rimanente  $1 - \tau : \tau = \tau : 1$ . Il reciproco della sezione aurea  $s = \frac{1}{2}(\sqrt{5} + 1)$  verifica diverse curiose proprietà; ricordiamo, ad esempio, le seguenti.

$$s = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \dots}}}; \quad s = \sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \dots}}}}$$

$$s = s^{-1} + s^{-2} + s^{-3} + \dots; \quad s^n = s^{n-1} + s^{n-2}$$

```

DOUBLE PRECISION A,B,C,D,E,EPS,XM,P,Q,R,TOL1,T2,U,V,W,FU,FV,FW,
2 FX,X,TOL3
DOUBLE PRECISION DABS,DSQRT
EXTERNAL F
C C = 1-r = r**2
C
C=0.5D0*(3.0D0-DSQRT(5.0D0))
C EPS e' approssimativamente la radice quadrata della precisione macchina
10 EPS=1.D-18
TOL1=EPS+1.0D0
EPS=DSQRT(EPS)
C.....
A=AX
B=BX
V=A+C*(B-A)
W=V
X=V
E=0.0D0
FX=F(X)
FV=FX
FW=FX
TOL3=TOL/3.0D0
C loop principale
20 XM=0.5D0*(A+B)
TOL1=EPS*DABS(X)+TOL3
T2=2.0D0*TOL1
C test d'arresto
IF (DABS(X-XM).LE.(T2-0.5D0*(B-A))) GO TO 190
P=0.0D0
Q=0.0D0
R=0.0D0
IF (DABS(E).LE.TOL1) GO TO 50
C interpolazione parabolica
R=(X-W)*(FX-FV)
Q=(X-V)*(FX-FW)
P=(X-V)*Q-(X-W)*R
Q=2.0D0*(Q-R)
IF (Q.LE.0.0D0) GO TO 30
P=-P
GO TO 40
30 Q=-Q
40 R=E
E=D
50 IF ((DABS(P).GE.DABS(0.5D0*Q*R)).OR.(P.LE.Q*(A-X))
2 .OR.(P.GE.Q*(B-X))) GO TO 60
C passo relativo all'interpolazione parabolica
D=P/Q
U=X+D
C controllo se F e' troppo vicino a AX o BX
IF (((U-A).GE.T2).AND.((B-U).GE.T2)) GO TO 90
D=TOL1

```

```
        IF (X.GE.XM) D=-D
        GO TO 90
C passo relativo alla sezione aurea
60 IF (X.GE.XM) GO TO 70
    E=B-X
    GO TO 80
70 E=A-X
80 D=C*E
C.....
90 IF (DABS(D).LT.TOL1) GO TO 100
    U=X+D
    GO TO 120
100 IF (D.LE.O.OO0) GO TO 110
    U=X+TOL1
    GO TO 120
110 U=X-TOL1
120 FU=F(U)
C aggiorna A, B, V, W, X
    IF (FX.GT.FU) GO TO 140
    IF (U.GE.X) GO TO 130
    A=U
    GO TO 140
130 B=U
140 IF (FU.GT.FX) GO TO 170
    IF (U.GE.X) GO TO 150
    B=X
    GO TO 160
150 A=X
160 V=W
    FV=FW
    W=X
    FW=FX
    X=U
    FX=FU
    GO TO 20
170 IF ((FU.GT.FW).AND.(W.NE.X)) GO TO 180
    V=W
    FV=FW
    W=U
    FW=FU
    GO TO 20
180 IF ((FU.GT.FV).AND.(V.NE.X).AND.(V.NE.W)) GO TO 20
    V=U
    FV=FU
    GO TO 20
C fine del loop principale
190 FMIN=X
    RETURN
    END
```

A scopo illustrativo, consideriamo l'applicazione della routine precedente per la ricerca del minimo della funzione  $f(x) = x^3 - 2x - 5$  nell'intervallo  $[0, 1]$ .

```

double precision a,b,xs,tol,fmin
external f
a=0.
b=1
tol=1.0d-10
xs=fmin(a,b,f,tol)
print*, 'xs=', xs
stop
end
double precision function f(x)
double precision x
f=x*(x*x-2.d0)-5.d0
return
end

```

Si ottiene come stima il valore  $\hat{x}^* = 0.816496581987164$  da confrontare con il valore esatto  $x^* = \sqrt{2/3} = 0.81649658092773$ . Quando applicato alla funzione  $f(x) = (x - 2)^4 - 9$  (cfr. Figura 5.29), con  $a = 1, b = 3$  e gli altri parametri uguali ai precedenti, si trova  $\hat{x}^* = 2.00001160879744$

### 5.5.2 Ottimizzazione in più dimensioni

Data una funzione  $f(\mathbf{x})$  definita su  $\mathbb{R}^n$ , con  $n > 1$ , consideriamo il seguente problema di minimo non vincolato

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (5.58)$$

Procedendo in maniera schematica, i metodi numerici per risolvere il problema (5.58) possono essere suddivisi in due gruppi. Il primo gruppo comprende i metodi che non richiedono la conoscenza delle derivate della funzione, ma che utilizzano soltanto i valori della funzione. Tali metodi, detti anche metodi a ricerca diretta (*direct search methods*), sono basati sul confronto diretto tra i valori della funzione, in punti successivi scelti in maniera opportuna. Ogni metodo corrisponde ad una particolare strategia per generare i valori della funzione da confrontare. Il metodo della sezione aurea e il metodo di Fibonacci sono esempi di metodi diretti nel caso unidimensionale. Per  $n > 1$ , segnaliamo, in particolare, il *metodo della griglia* (grid search), nel quale si confrontano i valori della funzione nei nodi di una reticolazione di  $\mathbb{R}^n$  ottenuta mediante rette parallele agli assi, e il *metodo del semplice*, nel quale ad ogni iterazione si confrontano i valori della funzione in  $n + 1$  punti, che possono essere considerati come i vertici di un semplice<sup>13</sup> in  $\mathbb{R}^n$ . Il metodo del semplice

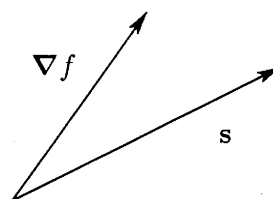
<sup>13</sup>Un *simpleso* in  $\mathbb{R}^n$  consiste di  $n + 1$  punti  $\mathbf{x}_i, i = 1, 2, \dots, n + 1$ , linearmente indipendenti, ossia non sullo stesso iperpiano, e dei punti che si ottengono considerando le combinazioni convesse dei punti  $\mathbf{x}_i$ , che vengono chiamati i *vertici* del semplice. Il semplice è *regolare* quando i suoi



rappresenta, a differenza del metodo della griglia, una ricerca del punto di minimo di tipo *adattivo*, nel senso che dal confronto dei valori della funzione nei vertici del semplice si traggono indicazioni per la costruzione di un nuovo semplice muovendosi in una direzione lungo la quale la funzione decresce (*direzione di discesa*).

Nel secondo gruppo di metodi le direzioni di discesa sono costruite utilizzando i valori delle derivate della funzione  $f(\mathbf{x})$ . Essi si basano, sostanzialmente, sul fatto che lungo una direzione  $\mathbf{s} = [s_1, s_2, \dots, s_n]^T$  la funzione  $f(\mathbf{x})$  decresce o aumenta a seconda che la derivata direzionale

$$\nabla f(\mathbf{x})^T \mathbf{s} = \frac{\partial f}{\partial x_1} s_1 + \frac{\partial f}{\partial x_2} s_2 + \dots +$$



è negativa o positiva, ossia a seconda che l'angolo tra il vettore  $\nabla f(\mathbf{x})$  del gradiente e il vettore  $\mathbf{s}$  è maggiore o minore di  $\pi/2$ . Per questo motivo tali metodi sono anche indicati come *metodi del gradiente* e differiscono tra loro per le scelte diverse della direzione  $\mathbf{s}$ . Nel seguito considereremo alcune di tali scelte, divenute ormai classiche, rinviando alla bibliografia per un approfondimento.

Nei paragrafi successivi la funzione  $f(\mathbf{x})$  sarà supposta sufficientemente regolare, ossia dotata delle derivate che verranno di volta in volta utilizzate.

### Metodo steepest descent

Come tutti i metodi del gradiente, il metodo della discesa più ripida (*steepest descent*, proposto da Cauchy nel 1845) è un metodo *iterativo*, nel senso che genera, a partire da una stima iniziale  $\mathbf{x}^1$ , una successione di approssimazioni  $\mathbf{x}^2, \mathbf{x}^3, \dots$ , ognuna delle quali è ottenuta mediante la risoluzione di un problema di minimo unidimensionale.

Più precisamente, nel metodo steepest descent la successione  $\{\mathbf{x}^k\}$  è definita, per  $k = 1, 2, \dots$ , nel modo seguente

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \lambda_k^* \mathbf{g}_k \quad (5.59)$$

ove con  $\mathbf{g}_k^*$  si è indicato il valore del gradiente  $\nabla f(\mathbf{x})$  calcolato nel punto  $\mathbf{x}^k$  e con  $\lambda_k^*$  il valore della variabile reale  $\lambda$  corrispondente al minimo della funzione  $\lambda \rightarrow f(\mathbf{x}^k - \lambda \mathbf{g}_k)$ , ossia al minimo della funzione  $f(\mathbf{x})$  lungo la direzione  $-\mathbf{g}_k$  a partire da  $\mathbf{x}^k$ . Nel metodo steepest descent, quindi, si assume come direzione di ricerca  $\mathbf{s}$  quella opposta al gradiente, la quale rappresenta, *localmente*, ossia in un intorno

---

vertici hanno distanza uguale. Esempi di semplici regolari sono un triangolo equilatero in  $\mathbb{R}^2$  e un tetraedro regolare in  $\mathbb{R}^3$ .

opportuno del punto  $\mathbf{x}^k$ , la direzione lungo la quale si ha la massima variazione della funzione. Sotto opportune ipotesi sulla funzione  $f(\mathbf{x})$ , è possibile dimostrare che la successione  $\{\mathbf{x}^k\}$  generata dal procedimento (5.59) converge a un punto  $\mathbf{x}^*$  in cui si annulla il gradiente, e quindi, eventualmente a un punto di minimo locale. Tuttavia, come mostreremo nell'esempio successivo, e come si può mostrare più in generale, la convergenza del metodo è di tipo lineare e in alcuni casi può essere, in effetti, estremamente lenta. Il motivo di tale comportamento è in sostanza da ricercare nel fatto che il metodo steepest descent non tiene in alcun conto delle derivate seconde della funzione  $f(\mathbf{x})$ , cioè della curvatura della funzione, che determina il comportamento della funzione vicino al minimo. Questo aspetto negativo viene superato nel metodo di Newton che considereremo nel successivo paragrafo.

► **Esempio 5.19** (*Metodo steepest descent*) Sia  $f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{9}{2}x_2^2$ . Si tratta evidentemente di una forma quadratica definita positiva con punto di minimo  $\mathbf{x}^* = [0, 0]^T$ . Si vede facilmente che il metodo di Cauchy, per  $\mathbf{x}^1 = [9, 1]^T$ , si riduce alla seguente iterazione

$$\mathbf{x}^k = \begin{bmatrix} 9 \\ (-1)^{k-1} \end{bmatrix} (0.8)^{k-1}, \quad k = 1, 2, \dots$$

La successione  $\{\mathbf{x}^k\}$  è rappresentata in Figura 5.31. Si ha, quindi,  $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2 / \|\mathbf{x}^k - \mathbf{x}^*\|_2 = \|\mathbf{x}^{k+1}\|_2 / \|\mathbf{x}^k\|_2 = c$ , con  $c$  costante. ■

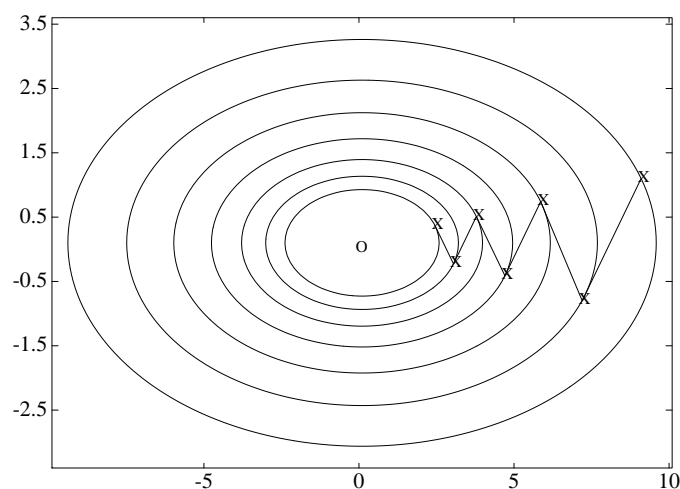


Figura 5.31: Rappresentazione dei risultati ottenuti mediante il metodo steepest descent applicato alla funzione  $f(x_1, x_2) = (x_1^2 + 9x_2^2)/2$ , con  $\mathbf{x}^1 = [9, 1]^T$ .

### Metodo di Newton

Come abbiamo già visto nel caso unidimensionale, l'idea di base del metodo di Newton consiste nell'approssimare, ad ogni passo della iterazione, la funzione data

$f(\mathbf{x})$  mediante la funzione quadratica che si ottiene dallo sviluppo in serie della funzione arrestato ai termini di secondo grado.

Nell'ipotesi che la funzione  $f(\mathbf{x})$  abbia le derivate continue fino al secondo ordine in un intorno opportuno del punto  $\mathbf{x}^k$ , si considera quindi il minimo della seguente funzione

$$q_k(\mathbf{x}) := f(\mathbf{x}^k) + (\mathbf{x} - \mathbf{x}^k)^T \mathbf{g}_k + \frac{1}{2}(\mathbf{x} - \mathbf{x}^k)^T \mathbf{G}_k (\mathbf{x} - \mathbf{x}^k)$$

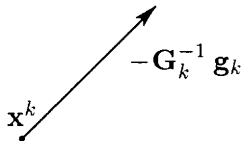
ove con  $\mathbf{G}_k$  si è indicata la matrice hessiana  $[\partial^2 f / \partial x_i \partial x_j]$ ,  $i, j = 1, \dots, n$ , e con  $\mathbf{g}_k$  il vettore gradiente, entrambi calcolati nel punto  $\mathbf{x} = \mathbf{x}^k$ . Supposto che  $q_k(\mathbf{x})$  abbia un punto di minimo in  $\bar{\mathbf{x}}$ , si ha  $\nabla q_k(\bar{\mathbf{x}}) = 0$ , cioè

$$\mathbf{G}_k(\bar{\mathbf{x}} - \mathbf{x}^k) + \mathbf{g}_k = 0 \Rightarrow \bar{\mathbf{x}} = \mathbf{x}^k - \mathbf{G}_k^{-1} \mathbf{g}_k$$

Il valore  $\bar{\mathbf{x}}$  è assunto come nuova stima del punto di minimo di  $f(\mathbf{x})$ , da cui la seguente formula iterativa

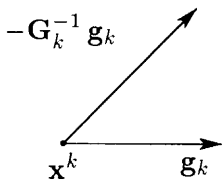
$$\boxed{\mathbf{x}^{k+1} = \mathbf{x}^k - \mathbf{G}_k^{-1} \mathbf{g}_k} \quad (5.60)$$

per  $k = 1, 2, \dots$ . Una variante del metodo (5.60) è la seguente

$$\boxed{\mathbf{x}^{k+1} = \mathbf{x}^k - \lambda_k^* \mathbf{G}_k^{-1} \mathbf{g}_k} \quad (5.61)$$


ove  $\lambda_k^*$  è determinata mediante la minimizzazione unidimensionale a partire da  $\mathbf{x}^k$  nella direzione  $-\mathbf{G}_k^{-1} \mathbf{g}_k$ .

Il calcolo del vettore  $\mathbf{x}^{k+1}$  richiede la risoluzione di un sistema di  $n$  equazioni in  $n$  incognite, il cui condizionamento dipende dalla matrice hessiana  $\mathbf{G}_k$ . Tale condizionamento può essere quindi migliorato mediante una opportuna operazione di scalatura sulle variabili. Il metodo di Newton coincide con il metodo steepest descent quando  $\mathbf{G}_k^{-1} = \mathbf{I}$ . Più in generale, la direzione  $-\mathbf{G}_k^{-1} \mathbf{g}_k$  è una direzione di discesa quando

$$\mathbf{g}_k^T \mathbf{G}_k^{-1} \mathbf{g}_k > 0 \quad (5.62)$$


ossia quando la matrice  $\mathbf{G}_k$  è *definita positiva*. Il successivo esempio illustra il comportamento del metodo quando la matrice hessiana non è definita positiva.

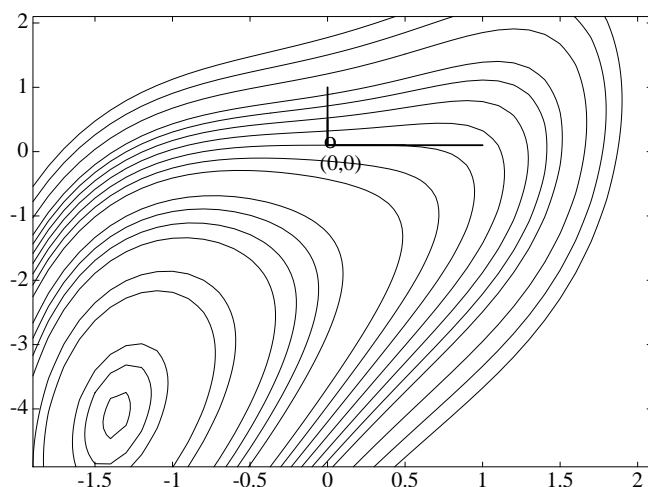


Figura 5.32: Esempio di non convergenza del metodo di Newton. Nel punto  $[0, 0]^T$  per la funzione  $f(\mathbf{x}) := x_1^4 - 3x_1x_2 + (x_2 + 2)^2$  la direzione fornita dal metodo di Newton risulta ortogonale alla direzione del gradiente.

► **Esempio 5.20** Consideriamo il problema del calcolo del minimo della seguente funzione

$$f(\mathbf{x}) := x_1^4 - 3x_1x_2 + (x_2 + 2)^2$$

a partire dal punto  $\mathbf{x}^1 = [0, 0]^T$ . Per tale funzione si ha

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} 4x_1^3 - 3x_2 \\ -3x_1 + 2(x_2 + 2) \end{bmatrix}, \quad \mathbf{G}(\mathbf{x}) = \begin{bmatrix} 12x_1^2 & -3 \\ -3 & 2 \end{bmatrix}$$

e quindi

$$\mathbf{g}_1 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \quad \mathbf{G}_1 = \begin{bmatrix} 0 & -3 \\ -3 & 2 \end{bmatrix}, \quad \mathbf{G}_1^{-1} = -\frac{1}{9} \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}, \quad -\mathbf{G}_1^{-1} \mathbf{g}_1 = \begin{bmatrix} 4/3 \\ 0 \end{bmatrix}$$

Come illustrato in Figura 5.32, nel punto  $\mathbf{x}^1$  la direzione di ricerca indicata dal metodo di Newton è ortogonale alla direzione locale del gradiente. Come conseguenza, lungo la direzione  $[4/3, 0]^T$  a partire dal punto  $\mathbf{x}^1$  si ha

$$f\left(\frac{4}{3}\lambda, 0\right) = \frac{256}{81}\lambda^4 + 4$$

e il minimo si verifica per  $\lambda = 0$ , e quindi  $\mathbf{x}^2 \equiv \mathbf{x}^1$ . ■

In opportune condizioni di regolarità sulla funzione  $f(\mathbf{x})$ , il metodo di Newton ha le medesime proprietà di convergenza viste in una dimensione, ossia “vicino” alla soluzione si ha

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2 \leq c \|\mathbf{x}^k - \mathbf{x}^*\|_2^2$$

con  $c$  costante non negativa dipendente dalla funzione  $f(\mathbf{x})$ . Come illustrazione, si veda il successivo esempio.

► **Esempio 5.21** Consideriamo il calcolo del minimo della seguente funzione

$$f(\mathbf{x}) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \quad (5.63)$$

che presenta un minimo nel punto  $\mathbf{x}^* = [1, 1]^T$ , con  $f(\mathbf{x}^*) = 0$ . Il grafico della funzione (5.63), introdotta da Rosenbrock (1960), rappresenta una valle con pareti molto ripide, il cui fondo segue approssimativamente la curva parabolica  $x_2 = x_1^2$ . Come punto di partenza viene usualmente scelto  $\mathbf{x}^1 = [-1.2, 1]^T$ .

Per la funzione (5.63) il metodo steepest descent si rivela altamente inefficiente, in quanto le traiettorie di ricerca, ossia le direzioni del gradiente, continuano a riflettersi lungo le pareti, con un lento avanzamento verso il punto di minimo.

Esaminiamo, quindi, l'applicazione del metodo di Newton nella forma (5.60). Il vettore gradiente e la matrice hessiana sono dati dalle seguenti espressioni

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} -400x_1(x_2 - x_1^2) + 2(x_1 - 1) \\ 200(x_2 - x_1^2) \end{bmatrix}, \quad \mathbf{G}(x) = \begin{bmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}$$

I risultati riportati nella Tabella 5.6 (cfr. anche Figura 5.33) mostrano la rapidità della convergenza del metodo nell'intorno del punto di minimo. Osserviamo che la successione  $f(x^k)$  non è monotona decrescente. ■

$k$	$\mathbf{x}^k$	$f(\mathbf{x}^k)$
1	$[-1.2, 1.]$	24.2
2	$[-1.1753, 1.3807]$	4.7319
3	$[0.7631, -3.1750]$	$1.4118 \cdot 10^3$
4	$[0.7634, 0.5828]$	0.0560
5	$[1.0000, 0.9440]$	0.3132
6	$[1.0000, 1.0000]$	$1.8527 \cdot 10^{-11}$

Tabella 5.6: Risultati ottenuti mediante il metodo di Newton applicato alla funzione di Rosenbrock.

Gli aspetti negativi del metodo di Newton derivano, principalmente, da una parte dalla difficoltà, usuale nelle applicazioni, di calcolare la matrice hessiana e dall'altra dalla necessità di risolvere ad ogni passo un sistema lineare. Allo scopo di superare tali difficoltà, sono stati introdotti metodi nei quali la matrice  $\mathbf{G}^{-1}$  dell'equazione (5.60) viene sostituita da una opportuna matrice definita positiva  $\mathbf{H}_k$ , che è aggiornata ad ogni iterazione solamente sulla base dei risultati ottenuti, senza la necessità della risoluzione di un sistema lineare.

I metodi basati su tale idea sono chiamati *metodi quasi-Newton*, o anche *metodi a matrice variabile* ed hanno la seguente forma

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \lambda_k^* \mathbf{H}_k^{-1} \mathbf{g}_k \quad (5.64)$$

per  $k = 1, 2, \dots$ , e con  $\mathbf{x}^1$ ,  $\mathbf{H}_1$  assegnati. Il valore  $\lambda^*$  corrisponde al minimo della  $f(\mathbf{x})$  lungo la direzione  $-\mathbf{H}_k \mathbf{g}_k$ . Vi sono, naturalmente, diverse possibilità di

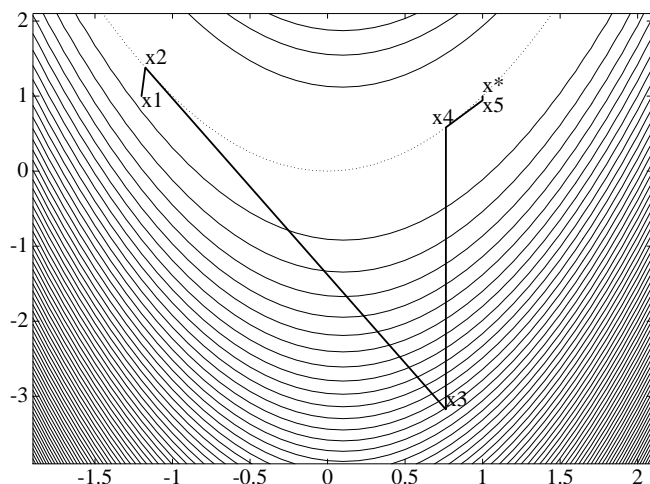


Figura 5.33: Illustrazione grafica dei risultati ottenuti mediante il metodo di Newton applicato alla funzione di Rosenbrock. La curva tratteggiata corrisponde alla funzione  $x_2 = x_1^2$ .

sceita per le matrici  $\mathbf{H}_k$ . Rinviamo alla letteratura specializzata per un adeguato approfondimento, nel successivo paragrafo descriveremo in dettaglio il *metodo di Davidon-Fletcher-Powell*, che è uno dei più noti metodi quasi-Newton<sup>14</sup>.

### Metodo di Davidon-Fletcher-Powell

Introduciamo il metodo, supponendo che la funzione  $f(\mathbf{x})$  sia una funzione *quadratica*. Più precisamente, consideriamo il problema del calcolo del minimo della seguente funzione

$$f(\mathbf{x}) := \frac{1}{2} \mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \quad (5.65)$$

ove  $\mathbf{G}$  è una matrice simmetrica definita positiva.

L'iterazione  $k$ -ma del metodo consiste dei seguenti passi.

**1** Si pone

$$\mathbf{d}_k = -\mathbf{H}_k \mathbf{g}_k$$

con  $\mathbf{H}_1 = \mathbf{I}$ . Il vettore  $\mathbf{d}_k$  fornisce la direzione di ricerca a partire dal punto corrente  $\mathbf{x}^k$ . Per  $k = 1$  il metodo coincide con il metodo di discesa rapida.

**2** Si minimizza la funzione ad una variabile  $\lambda \rightarrow f(\mathbf{x}^k + \lambda \mathbf{d}_k)$ . Sia  $\lambda_k^*$  ( $> 0$ ) il valore corrispondente al punto di minimo.

<sup>14</sup>Tale metodo, spesso indicato semplicemente con la sigla DFP, è stato introdotto in origine da Davidon nel 1959 e successivamente elaborato da Fletcher e Powell nel 1963.

3 Si pone

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \lambda_k^* \mathbf{d}_k$$

4 Si calcolano  $f(\mathbf{x}^{k+1})$  e  $\mathbf{g}_{k+1}$  (gradiente nel punto  $\mathbf{x}^{k+1}$ ). Nel punto  $\mathbf{x}^{k+1}$  il vettore  $\mathbf{d}_k$  è tangente alla superficie di livello  $f(\mathbf{x}) = f(\mathbf{x}^{k+1})$ , e pertanto è ortogonale a  $\mathbf{g}_{k+1}$ , ossia

$$\mathbf{d}_k^T \mathbf{g}_{k+1} = 0$$

5 Posto

$$\boldsymbol{\gamma}_k = \mathbf{g}_{k+1} - \mathbf{g}_k, \quad \boldsymbol{\sigma}_k = \lambda_k^* \mathbf{d}_k$$

si definisce

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \mathbf{A}_k + \mathbf{B}_k \quad (5.66)$$

ove

$$\mathbf{A}_k = \frac{\boldsymbol{\sigma}_k \boldsymbol{\sigma}_k^T}{\boldsymbol{\sigma}_k^T \boldsymbol{\gamma}_k} \quad (5.67)$$

$$\mathbf{B}_k = \frac{\mathbf{H}_k \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T \mathbf{H}_k}{\boldsymbol{\gamma}_k^T \mathbf{H}_k \boldsymbol{\gamma}_k} \quad (5.68)$$

Il metodo si arresta quando sono verificati opportuni test sulle variazioni  $\mathbf{x}^{k+1} - \mathbf{x}^k$ ,  $f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)$  e sulla lunghezza del vettore  $\mathbf{g}_k$ .

Rinviando alla bibliografia per la corrispondente dimostrazione, ricordiamo i seguenti risultati.

**Proposizione 5.6** *Nel metodo DFP le matrici  $\mathbf{H}_k$  sono simmetriche definite positive per ogni  $k$ .*

Da tale proprietà segue facilmente che il metodo DFP è *stabile*, nel senso che il valore di  $f(\mathbf{x})$  è ridotto ad ogni iterazione.

**Proposizione 5.7** *Se il metodo DFP è utilizzato per minimizzare la forma quadratica (5.65), con  $\mathbf{G}$  matrice simmetrica definita positiva di ordine  $n$ , allora  $\mathbf{H}_{n+1} = \mathbf{G}^{-1}$ .*

In altre parole, per una funzione quadratica il metodo DFP raggiunge il punto ottimale in al più  $n$  iterazioni, ossia, come anche si dice, è un metodo a *terminazione quadratica*. Sottolineiamo, comunque, che tale proprietà è in effetti vera solo se le operazioni sono eseguite esattamente, ossia in assenza di errori di arrotondamento, e se le successive ottimizzazioni unidimensionali (passo 2) forniscono il punto di minimo *esatto*.

La dimostrazione della Proposizione 5.7 utilizza in particolare il seguente risultato relativo alle direzioni di ricerca  $\boldsymbol{\sigma}_k$

$$\boldsymbol{\sigma}_k^T \mathbf{G} \boldsymbol{\sigma}_l = 0 \quad 1 \leq k < l < m \quad (5.69)$$

Come vedremo nel paragrafo successivo, un insieme di vettori che verificano tali condizioni sono detti *mutuamente G-coniugati* e sono, in particolare, *linearmente indipendenti*.

Naturalmente, quando la funzione  $f(\mathbf{x})$  non è una forma quadratica, il metodo è, in generale, di tipo iterativo; la convergenza della successione  $\{\mathbf{x}^k\}$  è comunque assicurata per funzioni che verificano opportune condizioni di regolarità (ad esempio, per le funzioni che hanno un minimo e sono strettamente convesse).

Relativamente alla relazione di aggiornamento (5.66), osserviamo che le matrici  $\mathbf{A}_k$  e  $\mathbf{B}_k$  sono due matrici simmetriche di ordine  $n$  e di rango 1 e che la matrice  $\mathbf{A}_k + \mathbf{B}_k$  è di rango 2. Ne segue che  $\mathbf{H}_k$  è aggiornata mediante l'addizione di una matrice simmetrica di rango 2. Si può dimostrare, in particolare, che la scelta delle matrici  $\mathbf{A}_k$  è tale che

$$\mathbf{G}^{-1} = \sum_{k=1}^n \mathbf{A}_k$$

I metodi quasi-Newton sono spesso classificati in termini della matrice utilizzata per aggiornare  $\mathbf{H}_k$ ; allora, il metodo DFP è un metodo di *rango 2*. In generale, i metodi di rango superiore hanno proprietà di convergenza migliori di quelle di rango inferiore, ma tale guadagno è spesso ottenuto mediante una complessità computazionale superiore. Il metodo DFP è illustrato nel seguente esempio.

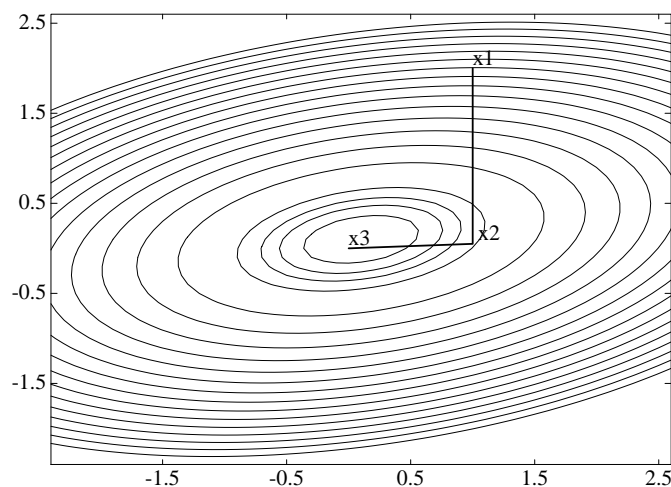


Figura 5.34: Punti generati dal metodo DFP applicato alla funzione quadratica  $f(\mathbf{x}) = x_1^2 - x_1x_2 + 3x_2^2$ , a partire dal punto iniziale  $\mathbf{x}^1 = [1, 2]^T$ .

► **Esempio 5.22** Analizziamo l'applicazione del metodo DFP per la minimizzazione della seguente funzione quadratica

$$f(\mathbf{x}) := x_1^2 - x_1x_2 + 3x_2^2$$



a partire dal punto iniziale  $\mathbf{x}^1 = [1, 2]^T$ . Tenendo conto che il vettore gradiente è dato da

$$\mathbf{g}(\mathbf{x}) \equiv \nabla f(\mathbf{x}) = [2x_1 - x_2, -x_1 + 6x_2]^T$$

si ha

$$\begin{aligned} \mathbf{g}_1 &= [0, 11]^T, \quad \mathbf{d}_1 = -\mathbf{H}_1 \mathbf{g}_1 = -\mathbf{g}_1 = [0, -11]^T, \quad \lambda_1^* = 1/6, \quad \boldsymbol{\sigma}_1 = [0, -11/6]^T \\ \mathbf{x}^2 &= [1, 1/6], \quad \mathbf{g}_2 = [11/6, 0], \quad \boldsymbol{\gamma}_1 = \mathbf{g}_2 - \mathbf{g}_1 = [11/6, -11]^T \\ \mathbf{A}_1 &= \begin{bmatrix} 0 & 0 \\ 0 & 1/6 \end{bmatrix}, \quad \mathbf{B}_1 = \begin{bmatrix} -1/37 & 6/37 \\ 6/37 & -36/37 \end{bmatrix}, \quad \mathbf{H}_2 = \begin{bmatrix} 36/37 & 6/37 \\ 6/37 & 43/222 \end{bmatrix} \\ \mathbf{d}_2 &= [-66/37, -11/37]^T, \quad \lambda_2^* = 37/66, \quad \boldsymbol{\sigma}_2 = [-1, -1/6]^T \end{aligned}$$

da cui  $\mathbf{x}^3 = [0, 0]^T$ ,  $\mathbf{g}_3 = [0, 0]^T$ . Pertanto  $\mathbf{x}^3$  è il punto di minimo (cfr. per una illustrazione Figura 5.34). Lasciamo come esercizio la verifica della proprietà  $\mathbf{H}_3 = \mathbf{G}^{-1}$ . ■

### Direzioni coniugate

Consideriamo il problema del calcolo del minimo della seguente funzione quadratica

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \quad (5.70)$$

con  $\mathbf{G}$  matrice simmetrica e definita positiva. Come introduzione al concetto di direzioni coniugate, consideriamo il caso bidimensionale, nel quale le curve di livello  $f(\mathbf{x}) = k$ , al variare di  $k$ , sono delle ellissi concentriche (cfr. Figura 5.35); il centro comune  $C$  rappresenta il punto di minimo della funzione  $f(\mathbf{x})$ .

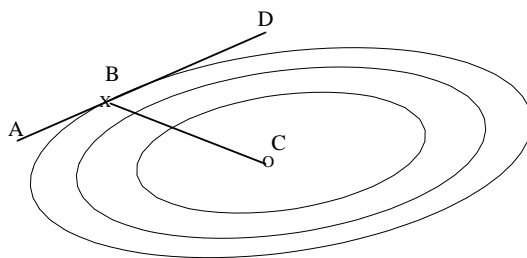


Figura 5.35: Direzioni coniugate.

Ricordiamo che due diametri di una ellisse sono detti *coniugati* (in senso geometrico) se ognuno è parallelo alla tangente all'ellisse condotta per l'estremità dell'altro. Con riferimento alla Figura 5.35, si ha allora che la direzione  $BC$  è coniugata alla direzione  $AD$ , in quanto il diametro attraverso  $B$  è coniugato al diametro parallelo ad  $AD$ . L'interesse delle direzioni coniugate nell'ambito del calcolo del minimo di  $f(\mathbf{x})$  risiede

nel fatto che il punto B rappresenta il punto di minimo della funzione  $f(\mathbf{x})$  lungo la direzione AD, in quanto tale direzione è tangente in B all'ellisse. In altre parole, se è nota la direzione coniugata alla direzione AD, il punto di minimo di  $f(\mathbf{x})$  può essere calcolato mediante al più due minimizzazioni unidimensionali. L'idea delle direzioni coniugate può essere facilmente estesa al caso di  $n$  dimensioni mediante la seguente definizione.

Sia  $\mathbf{G}$  una matrice simmetrica definita positiva. Allora, due direzioni  $\mathbf{p} \neq 0$  e  $\mathbf{q} \neq 0$  si dicono *coniugate* rispetto a  $\mathbf{G}$ , o semplicemente coniugate, se  $\mathbf{p}^T \mathbf{G} \mathbf{q} = 0$ . Lasciamo come esercizio di verificare, per  $n = 2$ , che tale definizione corrisponde a quella data in precedenza in forma geometrica; ancora come esercizio lasciamo la dimostrazione che gli autovettori di una matrice simmetrica  $\mathbf{A}$ , con autovalori distinti, sono vettori  $\mathbf{A}$ -coniugati.

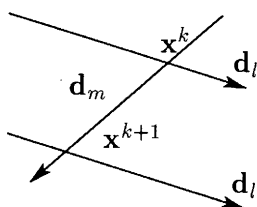
Esistono diversi modi di generare insiemi di direzioni mutuamente coniugate, a cui corrispondono differenti algoritmi di minimizzazione. Nel paragrafo precedente abbiamo già osservato che le direzioni generate dal metodo DFP sono mutuamente coniugate. Nel successivo paragrafo considereremo un altro classico metodo, noto come metodo di Fletcher-Reeves. In questo paragrafo analizzeremo alcune proprietà delle direzioni coniugate e il loro utilizzo negli algoritmi di minimizzazione.

**Teorema 5.9** *Se i vettori  $\mathbf{d}_j$  sono mutuamente coniugati, allora essi sono linearmente indipendenti.*

La dimostrazione di tale risultato può essere ottenuta facilmente per assurdo.

**Teorema 5.10** *Siano  $\mathbf{x}^k$  e  $\mathbf{x}^{k+1}$  due punti consecutivi ottenuti mediante un algoritmo di minimizzazione della funzione quadratica (5.70). Se*

- (i)  $\mathbf{x}^k$  minimizza  $f(\mathbf{x})$  nella direzione  $\mathbf{d}_l$ ,
- (ii)  $\mathbf{x}^{k+1}$  minimizza  $f(\mathbf{x})$  nella direzione  $\mathbf{d}_m$ ,
- (iii)  $\mathbf{d}_l$  e  $\mathbf{d}_m$  sono direzioni coniugate



*allora anche  $\mathbf{x}^{k+1}$  minimizza  $f(\mathbf{x})$  nella direzione  $\mathbf{d}_l$ .*

DIMOSTRAZIONE. La condizione (i) implica  $\mathbf{d}_l^T \mathbf{g}_k = 0$ , e inoltre da (iii) si ha  $\mathbf{d}_l^T \mathbf{G} \mathbf{d}_m = 0$ . D'altra parte, per la forma quadratica (5.70) si ha  $\mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{G}(\mathbf{x}^{k+1} - \mathbf{x}^k)$  e la condizione (ii) implica  $\mathbf{x}^{k+1} = \mathbf{x}^k + \lambda_m^* \mathbf{d}_m$ , ove  $\lambda_m^*$  è il valore ottimale. In definitiva, si ha

$$\mathbf{d}_l^T \mathbf{g}_{k+1} = \mathbf{d}_l^T (\mathbf{g}_k + \lambda_m^* \mathbf{G} \mathbf{d}_m) = 0$$

che dimostra il teorema. ■

Il teorema precedente ha i seguenti importanti corollari.

**Corollario 5.2** *Se*

- (i)  $\mathbf{x}^k$  minimizza  $f(\mathbf{x})$  nelle direzioni  $\mathbf{d}_1, \dots, \mathbf{d}_r$ ,
- (ii)  $\mathbf{x}^{k+1}$  minimizza  $f(\mathbf{x})$  nella direzione  $\mathbf{d}_k$ ,
- (iii)  $\mathbf{d}_k$  è coniugata alle direzioni  $\mathbf{d}_1, \dots, \mathbf{d}_r$ ,

allora  $\mathbf{x}^{k+1}$  minimizza  $f(\mathbf{x})$  nelle direzioni  $\mathbf{d}_1, \dots, \mathbf{d}_r$  e  $\mathbf{d}_k$ .

**Corollario 5.3** *Se, per  $k \geq 2$ ,*

- (i)  $\mathbf{x}^{r+1}$  minimizza  $f(\mathbf{x})$  nella direzione  $\mathbf{d}_r$ , per  $r = 1, \dots, k$ ,
- (ii) ognuna delle coppie di direzioni  $(\mathbf{d}_1, \mathbf{d}_2)$ ,  $(\mathbf{d}_2, \mathbf{d}_3)$ ,  $\dots$ ,  $(\mathbf{d}_{k-1}, \mathbf{d}_k)$  è una coppia di direzioni coniugate,

allora  $\mathbf{x}^{k+1}$  minimizza  $f(\mathbf{x})$  nelle direzioni  $\mathbf{d}_1, \dots, \mathbf{d}_r$  per  $r = 1, \dots, k$ .

Il corollario segue applicando il teorema successivamente, con  $(k, l, m)$  che assume i valori  $(2, 1, 2)$ ;  $(3, 2, 3)$ ,  $(3, 2, 3)$ ;  $(4, 1, 2)$ ,  $(4, 2, 3)$ ,  $(4, 3, 4)$ ;  $\dots$ ;  $(k, 1, 2)$ ,  $(k, 2, 3)$ ,  $\dots$ ,  $(k, k-1, k)$ .

**Teorema 5.11** *Siano  $\mathbf{d}_i$ ,  $i=1, 2, \dots, m (\leq n)$  direzioni mutuamente coniugate. Allora il minimo della funzione  $f(\mathbf{x})$ , definita in (5.70), nel sottospazio  $\mathbb{R}^m$  contenente il punto iniziale  $\mathbf{x}^1$  e le direzioni  $\mathbf{d}_i$  può essere calcolato mediante la successiva minimizzazione unidimensionale lungo le direzioni  $\mathbf{d}_i$ .*

DIMOSTRAZIONE. Il minimo richiesto si ottiene nel punto

$$\mathbf{x}^1 + \sum_i \lambda_i \mathbf{d}_i$$

ove i parametri  $\lambda_i$  sono scelti in maniera da minimizzare la funzione

$$f(\mathbf{x}^1 + \sum_i \lambda_i \mathbf{d}_i) = \frac{1}{2} \sum_i \lambda_i^2 \mathbf{d}_i^T \mathbf{G} \mathbf{d}_i + \sum_i \lambda_i \mathbf{d}_i^T (\mathbf{G} \mathbf{x}^1 + \mathbf{b}) + f(\mathbf{x}^1)$$

Il teorema segue, allora, osservando che nel secondo membro non esistono termini in  $\lambda_r \lambda_s$  ( $r \neq s$ ), grazie al fatto che le direzioni  $\mathbf{d}_i$  sono coniugate. ■

Si ha, allora, il seguente importante risultato che giustifica l'utilizzo delle direzioni coniugate.

**Corollario 5.4** *Un algoritmo che utilizza direzioni di ricerca mutuamente coniugate possiede la proprietà della terminazione quadratica, ossia per una funzione quadratica della forma (5.70) fornisce il minimo in al più  $n$  iterazioni.*

### Metodo di Fletcher-Reeves

Nel metodo di Fletcher-Reeves (1964) le direzioni coniugate vengono generate mediante una semplice formula ricorrente. Esso rappresenta, in sostanza, l'estensione alle funzioni non quadratiche del metodo del gradiente coniugato, introdotto da Hestenes e Stiefel (1952) per la risoluzione dei sistemi lineari (cfr. Capitolo 2).

Consideriamo il problema della ricerca del minimo della funzione definita in (5.70). Il passo iniziale coincide con quello del metodo di steepest descent

$$\mathbf{d}_1 = -\mathbf{g}_1$$

Successivamente, le direzioni sono scelte della seguente forma

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k \quad k = 1, 2, \dots \quad (5.71)$$

ove

$$\beta_k = \frac{\mathbf{g}_{k+1}^T \mathbf{g}_{k+1}}{\mathbf{g}_k^T \mathbf{g}_k}$$

Si può dimostrare, procedendo per induzione, che le direzioni generate dalla relazione (5.71) sono mutuamente coniugate e quindi il metodo è a terminazione quadratica. Quando applicato a funzioni non quadratiche, il metodo è iterativo. Per evitare che le direzioni  $\mathbf{d}_k$  divengano linearmente dipendenti, si riparte periodicamente, ad esempio, per  $k = 1, n+1, 2n+1, \dots$  con la direzione del metodo di steepest descent.

A differenza del metodo DFP che produce le direzioni coniugate mediante formule basate su relazioni tra *matrici* di ordine  $n$ , il metodo di Fletcher-Reeves utilizza solo formule *vettoriali*. Di conseguenza, la quantità di memoria utilizzata dal metodo di Fletcher-Reeves è dell'ordine di  $n$ , mentre è dell'ordine di  $n^2$  per il metodo DFP.

► **Esempio 5.23** Come illustrazione del metodo di Fletcher-Reeves, consideriamo il calcolo del minimo della funzione

$$f(\mathbf{x}) = x_1^2 - x_1 x_2 + 3x_2^2$$

con  $\mathbf{x}^1 = [1, 2]^T$ . Dal momento che  $g(\mathbf{x}) = [2x_1 - x_2, -x_1 + 6x_2]^T$ , si ha

$$\mathbf{d}_1 = -\mathbf{g}_1 = [0, -11]^T$$

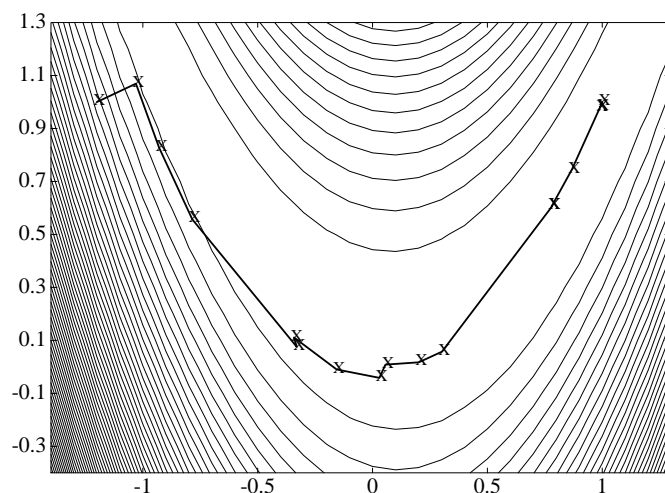


Figura 5.36: Risultati ottenuti mediante il metodo di Fletcher-Reeves nel caso della funzione di Rosenbrock  $f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$ , con  $\mathbf{x}^1 = [-1.2, 1]^T$ .

che fornisce un minimo nel punto  $\mathbf{x}^2 = [1, 1/6]^T$ . La direzione di ricerca a partire da  $\mathbf{x}^2$  è calcolata nel seguente modo

$$\mathbf{g}_2 = \left[ \frac{11}{6}, 0 \right]^T, \quad \beta_1 = -\frac{\mathbf{g}_2^2}{\mathbf{g}_1^2} = \frac{1}{36} \Rightarrow \mathbf{d}_2 = -\mathbf{g}_2 + \beta_1 \mathbf{d}_1 = \left[ -\frac{11}{6}, -\frac{11}{36} \right]$$

Il minimo lungo la direzione  $\mathbf{d}_2$  è ottenuto nel punto  $\mathbf{x}^3 = [0, 0]^T$  che rappresenta il punto di minimo globale della funzione  $f(\mathbf{x})$ , in quanto  $\mathbf{g}_3 = [0, 0]^T$ . ■

Nella Tabella 5.7 sono riportati i risultati ottenuti mediante il metodo di Fletcher-Reeves relativamente alla funzione di Rosenbrock. Tali risultati sono rappresentati in Figura 5.36.

### 5.5.3 Metodo SOR

Sia  $\mathbf{x} \rightarrow f(\mathbf{x})$  una funzione *continua e convessa* da  $\mathbb{R}^n \rightarrow \mathbb{R}$  e consideriamo il problema

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

Nelle seguenti ipotesi

$$f(\mathbf{x}) \rightarrow +\infty \quad \text{se} \quad \|\mathbf{x}\| \rightarrow \infty \quad (5.72)$$

$$f \text{ è strettamente convessa} \quad (5.73)$$

si può dimostrare che *esiste un unico*  $\mathbf{x}^*$  tale che

$$f(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

$k$	$x^k$	$f(x^k)$
1	[-1.2, 1.]	24.2
2	[-1.03009, 1.06935]	4.1281
3	[-0.92947, 0.82812]	3.85102
4	[-0.78477, 0.55703]	3.53164
5	[-0.32980, 0.07541]	1.87961
6	[-0.34189, 0.11097]	1.80419
7	[-0.15646, -0.00972]	1.45441
8	[0.02956, -0.04140]	1.12050
9	[0.05632, 0.00850]	0.893413
10	[0.20116, 0.01717]	0.692392
11	[0.30221, 0.05744]	0.601708
12	[0.78255, 0.60722]	0.499576E - 01
13	[0.77988, 0.60651]	0.487407E - 01
14	[0.86782, 0.74492]	0.241763E - 01
15	[0.99178, 0.98242]	0.216667E - 03
16	[0.98965, 0.97918]	0.112046E - 03
17	[1.00002, 0.999977]	0.323624E - 06

Tabella 5.7: Risultati ottenuti mediante il metodo di Fletcher-Reeves applicato alla funzione di Rosenbrock.

Partendo da  $\mathbf{x}^1 = \{x_1^1, \dots, x_n^1\}$  e assumendo  $\mathbf{x}^k$  noto, si calcola  $x_i^{k+1}$ , successivamente per  $i = 1, 2, \dots, n$ , come la soluzione del seguente problema unidimensionale

$$f(x_1^{k+1}, \dots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k) \leq f(x_1^{k+1}, \dots, x_{i-1}^{k+1}, z, x_{i+1}^k, \dots, x_n^k), \quad \forall z \in \mathbb{R}$$

per  $i = 1, 2, \dots, n$ . Esaminiamo la *convergenza* di tale algoritmo, incominciando da un *controesempio* che mostra che l'algoritmo, nelle sole ipotesi fatte, può essere *divergente*.

► **Esempio 5.24** Consideriamo, per  $n = 2$ , la funzione

$$f(x_1, x_2) = x_1^2 + x_2^2 - 2(x_1 + x_2) + 2|x_1 - x_2|$$

che verifica le ipotesi richieste, ma che *non è differenziabile*. Si ha  $\mathbf{x}^* = [1, 1]^T$ . Applicando l'algoritmo per  $\mathbf{x}^0 = [0, 0]^T$ , si ha da calcolare il minimo rispetto a  $x_1$  della seguente funzione

$$f(x_1, 0) = x_1^2 - 2x_1 + 2|x_1|$$

che fornisce  $x_1^1 = 0$ . Successivamente si definisce  $x_2^1$  mediante la minimizzazione rispetto a  $x_2$  di  $f(x_1^1, x_2) = f(0, x_2)$ , che fornisce  $x_2^1 = 0$ . L'origine è, quindi, un punto in cui l'algoritmo si *blocca*. ■

Osserviamo comunque che la differenziabilità non è una condizione necessaria per la convergenza; il metodo può convergere anche per particolari funzioni non differenziabili. Comunque nel caso di differenziabilità si può dimostrare il seguente risultato.

**Teorema 5.12** *Se la funzione  $f(\mathbf{x})$  verifica le ipotesi (5.72) e (5.73) ed è di classe  $C^1$ , allora il metodo di rilassamento converge alla soluzione  $\mathbf{x}^*$  del problema di minimo.*

L'algoritmo precedente può essere facilmente generalizzato al caso di un problema di minimo *vincolato*, quando l'insieme di ammissibilità  $K$  è definito nel modo seguente

$$K = \prod_{i=1}^n K_i \quad (5.74)$$

$$K_i = [a_i, b_i] \subset \mathbb{R}$$

ove  $a_i, b_i$  non sono necessariamente numeri finiti. Il convesso  $K$  è allora detto di tipo *locale*. L'estensione dell'algoritmo precedente è ovvia. Si definisce  $x_i^{k+1}$  come la soluzione in  $K_i$  del seguente problema di minimo unidimensionale

$$f(x_1^{k+1}, \dots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k) \leq f(x_1^{k+1}, \dots, x_{i-1}^{k+1}, z, x_{i+1}^k, \dots, x_n^k), \quad \forall z \in K_i$$

per  $i = 1, \dots, n$ . Per tale algoritmo si estende il risultato di *convergenza* del Teorema 5.12.

**▼ Osservazione 5.4** *Se il convesso  $K$  non è della forma (5.74), allora il metodo può non convergere per determinati punti di partenza; si consideri come controesempio il calcolo del minimo della funzione  $f(\mathbf{x}) = x_1^2 + x_2^2$  sull'insieme  $K$  definito da*

$$K = \{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^2, x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \geq 1\}$$

*che ammette come minimo il punto  $\mathbf{x}^* = [1/2, 1/2]^T$ . Partendo da  $\mathbf{x}^0 = [0, 1]$ , l'algoritmo si blocca in tale punto. ■*

Terminiamo con un'estensione al caso vincolato del metodo SOR che abbiamo visto nel Capitolo 2 per la risoluzione dei sistemi lineari. Supponiamo che  $f$  sia della forma

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{A}\mathbf{x}, \mathbf{x}) - (\mathbf{b}, \mathbf{x})$$

ove  $\mathbf{A}$  è una matrice simmetrica definita positiva, e  $K$  di tipo *locale*

$$K = \prod_{i=1}^n K_i$$

L'algoritmo risulta definito dalle formule seguenti, ove  $i = 1, 2, \dots, n$

$$\begin{cases} x_i^{k+1/2} = -\frac{1}{a_{ii}} \left( \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} + \sum_{j=i+1}^n a_{ij} x_j^k - b_i \right) \\ x_i^{k+1} = P_{K_i}((1 - \omega)x_i^k + \omega x_i^{k+1/2}) \end{cases}$$

Nelle ipotesi precedenti, si può dimostrare come per il caso non vincolato, che il metodo converge per  $0 < \omega < 2$ .

### 5.5.4 Minimi quadrati non lineari

Il problema può avere *origine* in diversi contesti applicativi: *identificazione di parametri*, *curve-fitting*, ecc. Per esempi nell'ambito della identificazione dei parametri si vedano i Capitoli 12 e 13. Il problema può essere formulato, in forma generale, nel seguente modo. È dato un *modello matematico*

$$\mathbf{y} = \mathbf{F}(\mathbf{t}, \mathbf{x}) \quad (5.75)$$

cioè una dipendenza funzionale di  $\mathbf{y}$  da  $\mathbf{t}$  e  $\mathbf{x}$ . Tale funzione può essere data esplicitamente, oppure più in generale può essere la soluzione, ad esempio, di un problema differenziale.

Il vettore  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$  è il *vettore dei parametri*. Il vettore  $\mathbf{t} = [t_1, t_2, \dots, t_k]^T$  rappresenta le *variabili indipendenti*, mentre  $\mathbf{y} = [y_1, y_2, \dots, y_r]^T$  è il vettore delle *variabili dipendenti*, o *osservate* (per il seguito supporremo  $r = 1$ ). Supponendo di conoscere, in corrispondenza ai valori  $\mathbf{t}^1, \mathbf{t}^2, \dots, \mathbf{t}^m$  della variabile indipendente, i valori *sperimentali*  $y^*$  di  $y$ , si definiscono per  $i = 1, 2, \dots, m$  i *residui*

$$f_i(\mathbf{x}) = F(\mathbf{t}^i, \mathbf{x}) - y^*(\mathbf{t}^i) \quad (5.76)$$

e si pone  $\mathbf{f}(\mathbf{x}) = [f_1, f_2, \dots, f_m]^T$ . Il problema è, allora, quello di determinare il vettore  $\mathbf{x}$  in modo che i residui siano *minimi*. Per precisare il problema, è necessario introdurre un particolare *stimatore*, cioè una particolare *distanza* in  $\mathbb{R}^m$ . Quando si sceglie come distanza la norma euclidea<sup>15</sup>  $\|\cdot\|_2$ , il problema diventa il seguente, che viene detto *problema dei minimi quadrati*

$$\min_{\mathbf{x} \in \mathbb{R}^n} S(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{f}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) \quad (5.77)$$

Naturalmente, nelle applicazioni possono esistere ulteriori vincoli per i parametri  $\mathbf{x}$ . In particolare, se i parametri  $\mathbf{x}$  del modello hanno un significato fisico, si ha usualmente  $\mathbf{x} \geq 0$ . Per il seguito, tuttavia considereremo, per semplicità, solo il caso di parametri *non vincolati*.

Supporremo, inoltre, che il modello matematico  $F$  sia regolare, in particolare che sia possibile calcolare le *derivate*

$$J_{ij} = \frac{\partial f_i}{\partial x_j}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n$$

La matrice  $\mathbf{J}(\mathbf{x}) = [J_{ij}]$  di ordine  $m \times n$  è la *matrice Jacobiana*.

<sup>15</sup>Più in generale, si potrebbe prendere una norma euclidea *pesata*, cioè del tipo  $\|\cdot\|_{\mathbf{A}}$ , con  $\mathbf{A}$  matrice definita positiva, corrispondente, ad esempio, alla *matrice di covarianza* dei dati sperimentali. Tale scelta ha proprietà statistiche interessanti; in effetti (cfr. Capitolo 8), se i dati  $y^*$  sono affetti da errori distribuiti con legge gaussiana, essa fornisce la minima varianza per  $\mathbf{x}$ .



▼ **Osservazione 5.5** Quando  $m = n$ , per ogni soluzione del sistema

$$\mathbf{f}(\mathbf{x}) = 0$$

si ha  $S(\mathbf{x}) = 0$ . Nel quadro precedente, quindi, rientra come caso particolare il problema della risoluzione di un sistema non lineare. ■

▼ **Osservazione 5.6** Ovviamente, tutti i metodi di ottimizzazione visti nelle sezioni precedenti (gradiente, gradiente-coniugato, quasi-Newton, ...) si applicano anche al problema dei minimi quadrati. C'è, comunque, l'interesse a vedere se, data la forma particolare di tale problema, si possano costruire metodi ad hoc più efficienti. ■

### Gradiente e hessiana di $S(\mathbf{x})$

Con un semplice calcolo si ottiene

$$\begin{aligned}\nabla S(\mathbf{x}) &= 2\mathbf{J}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) \\ \nabla^2 S(\mathbf{x}) &= 2\mathbf{J}(\mathbf{x})^T \mathbf{J}(\mathbf{x}) + 2 \sum_{i=1}^m \nabla^2 f_i(\mathbf{x}) f_i(\mathbf{x})\end{aligned}$$

ove  $\nabla^2 f_i(\mathbf{x})$  sono le matrici hessiane delle funzioni  $f_i(\mathbf{x})$ .

► **Esempio 5.25** Come illustrazione, supponiamo di voler fittare i dati  $(t^i, y(t^i))$ ,  $i = 1, \dots, 4$  mediante il modello  $F(t, \mathbf{x}) = \exp(tx_1) + \exp(tx_2)$ , con  $t \in \mathbb{R}$ . In questo caso si ha  $\mathbf{J}(\mathbf{x}) \in \mathbb{R}^{4 \times 2}$

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} t^1 e^{t^1 x_1} & t^1 e^{t^1 x_2} \\ t^2 e^{t^2 x_1} & t^2 e^{t^2 x_2} \\ t^3 e^{t^3 x_1} & t^3 e^{t^3 x_2} \\ t^4 e^{t^4 x_1} & t^4 e^{t^4 x_2} \end{bmatrix}$$

Inoltre,  $\nabla S(\mathbf{x}) \in \mathbb{R}^2$

$$\nabla S(\mathbf{x}) = 2\mathbf{J}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) = 2 \left[ \sum_{i=1}^4 f_i(x) t^i e^{t^i x_1}, \sum_{i=1}^4 f_i(x) t^i e^{t^i x_2} \right]^T$$

e  $\nabla^2 S(\mathbf{x}) \in \mathbb{R}^{2 \times 2}$

$$\nabla^2 S(\mathbf{x}) = 2 \begin{bmatrix} \sum_{i=1}^4 (t^i)^2 e^{t^i x_1} (f_i(x) + e^{t^i x_1}) & \sum_{i=1}^4 (t^i)^2 e^{t^i (x_1 + x_2)} \\ \sum_{i=1}^4 (t^i)^2 e^{t^i (x_1 + x_2)} & \sum_{i=1}^4 (t^i)^2 e^{t^i x_2} (f_i(x) + e^{t^i x_2}) \end{bmatrix}$$

■

Una condizione *necessaria* in un punto di minimo per (5.77) è la seguente

$$\nabla S(\mathbf{x}) = 0 \Rightarrow \mathbf{J}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) = 0 \quad (5.78)$$

Se allora per risolvere (5.78) utilizziamo il *metodo di Newton*, si ottiene

$$\begin{cases} \mathbf{x}^1 & \text{arbitrario} \\ \mathbf{x}^{k+1} = \mathbf{x}^k + \lambda_k \mathbf{s}^k & \boxed{\nabla^2 S(\mathbf{x}^k) \mathbf{s}^k = -\nabla S(\mathbf{x}^k)} \end{cases}$$

In questa forma, tuttavia, il metodo risulta costoso, in quanto richiede il calcolo del termine  $\sum_{i=1}^m \nabla^2 f_i(\mathbf{x}) f_i(\mathbf{x})$ . Osserviamo, d'altra parte, che il contributo di tale termine è, “piccolo” vicino alla soluzione, in due situazioni: quando il modello è “buono”, ossia quando i residui  $f_i(\mathbf{x})$  sono piccoli, e quando le funzioni  $f_i$  sono “quasi lineari”, ossia i termini  $\nabla^2 f_i(\mathbf{x})$  sono piccoli. In tali casi risulta, pertanto, ragionevole la seguente approssimazione

$$\nabla^2 S(\mathbf{x}) \approx 2 \mathbf{J}(\mathbf{x})^T \mathbf{J}(\mathbf{x})$$

Assumendo tale approssimazione si ottiene il cosiddetto *metodo di Gauss-Newton*. Quindi, nel metodo di Gauss-Newton la direzione  $\mathbf{s}^k$  è calcolata come soluzione del seguente sistema lineare delle *equazioni normali*

$$\mathbf{J}(\mathbf{x}^k)^T \mathbf{J}(\mathbf{x}^k) \mathbf{s}^k = -\mathbf{J}(\mathbf{x}^k)^T \mathbf{f}(\mathbf{x}^k)$$

Naturalmente, per risolvere il sistema delle equazioni normali sono opportune le “precauzioni numeriche” indicate nel caso lineare. Ricordiamo, in particolare le tecniche basate sulle decomposizioni **QR**, **SVD**.

Nel caso generale, ossia quando il termine  $\sum_{i=1}^m \nabla^2 f_i(\mathbf{x}) f_i(\mathbf{x})$  non è trascurabile, esistono metodi che introducono opportune simulazioni di tale termine. Ad esempio, nel *metodo ibrido di Powell* si ha l'iterazione

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \lambda_k \mathbf{p}^k$$

ove

$$\mathbf{p}^k = \beta_k \mathbf{s}^k - \gamma_k \nabla S_k$$

Si utilizza, cioè, una direzione che è una combinazione opportuna della direzione di Gauss-Newton e quella del gradiente (cfr. Figura 5.37).

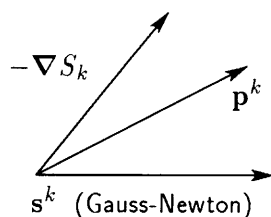


Figura 5.37: Metodo ibrido di Powell.

Nel *metodo di Levenberg-Marquardt*<sup>16</sup>, la direzione di ricerca  $\mathbf{s}^k$  è ottenuta risolvendo il seguente sistema lineare

$$(\mathbf{J}(\mathbf{x}^k)^T \mathbf{J}(\mathbf{x}^k) + \delta_k \mathbf{D}_k) \mathbf{s}^k = -\mathbf{J}(\mathbf{x}^k)^T \mathbf{f}(\mathbf{x}^k)$$

<sup>16</sup>Tale metodo è stato proposto da Levenberg (1944) e da Marquardt (1963).

ove  $\mathbf{D}_k$  è una opportuna matrice definita positiva e  $\delta_k$  è un parametro opportuno. Osserviamo che se  $\delta_k$  è sufficientemente grande la matrice  $\mathbf{J}(\mathbf{x}^k)^T \mathbf{J}(\mathbf{x}^k) + \delta_k \mathbf{D}_k$  è definita positiva e  $\mathbf{s}^k$  è una direzione di discesa, mentre per  $\delta_k = 0$  si riottiene il metodo di Gauss-Newton. Si ha ancora, quindi, come direzione  $\mathbf{s}^k$  una *combinazione tra Gauss-Newton e gradiente*. Nei due metodi la parte, numericamente più “delicata”, riguarda, ovviamente, la *scelta dei vari parametri*. Nelle implementazioni più utilizzate la scelta dei parametri è di tipo *adattivo*; essi vengono, cioè, aggiornati sulla base dei risultati di minimizzazione successivamente ottenuti.

◆ **Esercizio 5.37** *Mostrare che se  $\bar{\mathbf{x}}$  è un punto di minimo di una funzione  $f(\mathbf{x})$ ,  $\mathbf{x} \in \mathbb{R}^n$ , lungo la retta  $\mathbf{x} = x_1 + \lambda \mathbf{d}$ , con  $\mathbf{d} \in \mathbb{R}^n$ , allora tale retta è tangente in  $\bar{\mathbf{x}}$  alla superficie di livello  $f(\mathbf{x}) = f(\bar{\mathbf{x}})$ .*

◆ **Esercizio 5.38** *Fornire un esempio di una funzione strettamente convessa che non ha punti di minimo.*

◆ **Esercizio 5.39** *Applicare il metodo di Newton per minimizzare le seguenti funzioni*

- (a)  $(n-1)x + bx^{1-n}$ , (b)  $(n-2)x^2 + 2bx^{2-n}$  ( $n > 2$ )  
 (c)  $x^{n+1} - (n+1)bx$ , (d)  $n \lg n + bx^{-n}$ ,  
 (e)  $(n-3)x^{(n+3)/2} + (n+3)x^{-(n-3)/2}$  ( $n \neq 3$ )

ove  $x > 0$ ,  $b > 0$ , e  $n > 1$ .

◆ **Esercizio 5.40** *Applicare il metodo di Newton per la minimizzazione della seguente funzione*

$$f(x_1, x_2) = x_1^4 + 6x_1x_2 + 1.5x_2^2 + 36x_2 + 405$$

con  $\mathbf{x}^1 = [0, 0]$ .

◆ **Esercizio 5.41** *Determinare il punto di minimo di  $f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 4x_2^2)$  lungo la retta  $3x_1 + 2x_2 = 4$ . Mostrare che i vettori  $\mathbf{p} = [1, 1]^T$  e  $\mathbf{q} = [4, -3]^T$  sono coniugati.*

◆ **Esercizio 5.42** *Sia  $\mathbf{P}$  una matrice le cui colonne  $\mathbf{p}_1, \dots, \mathbf{p}_n$  sono mutuamente coniugate, in maniera che  $\mathbf{D} = \mathbf{P}^T \mathbf{A} \mathbf{P}$  è una matrice diagonale con  $d_1, \dots, d_n$  elementi diagonali. Mostrare che  $\mathbf{A}^{-1} = \mathbf{P} \mathbf{D}^{-1} \mathbf{P}^T$ . Usare tale risultato per mostrare che*

$$\mathbf{A}^{-1} = \sum_{k=1}^n \frac{\mathbf{p}_k \mathbf{p}_k^T}{d_k}$$

◆ **Esercizio 5.43** *Sia  $S : \mathbb{R}^2 \rightarrow \mathbb{R}$  definita da*

$$S(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \mathbf{f}(\mathbf{x})$$

con  $f_1(x) = x_2 - x_1^2$  e  $f_2(x) = 1 - x_1$ . Analizzare il metodo di Newton e il metodo di Gauss-Newton per la ricerca del punto di minimo  $[1, 1]^T$  di  $S$ .

The function of an expert is not to be more right than other people, but to be wrong for more sophisticated reasons.

D. Butler

## Capitolo 6

# Integrazione numerica

In questo capitolo analizzeremo i metodi numerici per l'approssimazione di integrali definiti<sup>1</sup>, della forma

$$I(f) = \int_a^b f(x) dx \quad (6.1)$$

con  $f(x)$  definita su un intervallo limitato e chiuso  $[a, b]$ . L'integrale (6.1) sarà inteso nel senso di *integrale di Riemann*, di cui ricorderemo ora la definizione e le proprietà essenziali.

**Integrale di Riemann** Si suddivide l'intervallo  $[a, b]$  in  $n$  sottointervalli mediante i seguenti punti

$$a = x_0 < x_1 < \dots < x_n = b$$

Sia  $\xi_i$  un generico punto del sottointervallo  $i$ -mo:  $x_i \leq \xi_i \leq x_{i+1}$ , e formiamo la somma

$$S(f; P) := \sum_{i=0}^{n-1} f(\xi_i) (x_{i+1} - x_i) \quad (6.2)$$

---

<sup>1</sup>Il calcolo numerico di integrali definiti, collegato con il calcolo dell'area di regioni limitate da curve, è uno dei problemi più antichi nella matematica. Sicuramente l'esempio più noto di tale problema fu il calcolo dell'area contenuta in un cerchio, ossia dell'approssimazione del numero  $\pi$ . Utilizzando un metodo numerico basato sull'approssimazione di un cerchio mediante dei poligoni inscritti, e rispettivamente circoscritti, Archimede (287-212 A.C.) ottenne la sorprendente limitazione  $3 \frac{10}{71} < \pi < 3 \frac{1}{7}$ .

ove, per brevità, con  $P$  si è indicata la partizione dell'intervallo. Tali somme sono chiamate le *somme di Riemann*. Indicando con  $\Delta$  la massima lunghezza dei sottointervalli, ossia posto  $\Delta := \max_i(x_{i+1} - x_i)$ , consideriamo le successioni  $\{S(f; P_n)\}$ ,  $n = 1, 2, \dots$  di somme del tipo (6.2), tali che  $\lim_{n \rightarrow \infty} \Delta_n = 0$ . Si dice che la funzione  $f(x)$  è integrabile secondo Riemann quando tutte le successioni  $\{S(f; P_n)\}$ , corrispondenti ad una qualsiasi scelta dei punti  $\xi_i$ , hanno, per  $n \rightarrow \infty$ , un limite<sup>2</sup> comune  $I$ . Tale limite viene chiamato l'integrale di Riemann della funzione  $f(x)$  (detta funzione integranda) sull'intervallo  $[a, b]$  (detto intervallo di integrazione) e si scrive

$$I = \int_a^b f(x) dx$$

ove il simbolo  $\int$  è originato da una deformazione del simbolo  $S$ . La *rapidità* di convergenza di ogni successione  $S(f; P_n)$  dipende naturalmente dal modo particolare con cui sono costruite le partizioni  $P_n$  e dalla scelta dei punti  $\xi_i$ . I metodi numerici che introdurremo nel seguito possono, in sostanza, essere interpretati come scelte *opportune* della partizione e dei punti di valutazione della funzione.

Una maniera equivalente di presentare la definizione di integrale di Riemann è la seguente. In corrispondenza ad ogni sottointervallo  $[x_i, x_{i+1}]$ , si indica con  $m_i$  (rispettivamente con  $M_i$ ) il limite inferiore (rispettivamente il limite superiore) della funzione  $f(x)$  su  $[x_i, x_{i+1}]$ . Si costruiscono, quindi le seguenti due somme

$$L(f; P) = \sum_{i=0}^{n-1} m_i(x_{i+1} - x_i), \quad U(f; P) = \sum_{i=0}^{n-1} M_i(x_{i+1} - x_i)$$

La funzione  $f(x)$  è, allora, integrabile secondo Riemann, quando il limite superiore di  $L(f; P)$  al variare di  $P$  nell'ambito di tutte le possibili partizioni di  $[a, b]$  coincide con il limite inferiore di  $U(f; P)$ . Il loro valore comune rappresenta l'integrale di Riemann di  $f(x)$  su  $[a, b]$ .

Osserviamo che per definizione, in corrispondenza ad una fissata partizione  $P$  si ha

$$L(f; P) \leq \int_a^b f(x) dx \leq U(f; P)$$

Quando la funzione  $f(x)$  è positiva, le due quantità  $L$  e  $U$  rappresentano una stima, rispettivamente inferiore e superiore, dell'area della superficie definita da  $0 \leq x \leq f(x)$ , con  $a \leq x \leq b$  e il cui valore è fornito dall'integrale di Riemann della funzione.

Si può dimostrare che, in particolare, è integrabile secondo Riemann una *funzione continua sull'intervallo limitato e chiuso*  $[a, b]$ ; più in generale, che è integrabile una

---

<sup>2</sup>In maniera più precisa, significa che per ogni  $\epsilon > 0$  si può trovare un  $\delta > 0$  tale che  $|S(f; P) - I| < \epsilon$  per ogni suddivisione  $P$  con  $\Delta < \delta$  e per ogni scelta dei punti  $\xi_i$ .

funzione  $f(x)$  limitata su  $[a, b]$  e continua, con eccezione al più di un numero finito, o numerabile, di punti di discontinuità<sup>3</sup>.

Ricordiamo le seguenti proprietà fondamentali, ove  $f$  e  $g$  sono funzioni limitate e integrabili secondo Riemann su  $[a, b]$

$$\begin{aligned}\int_a^a f(x) dx &= 0, & \int_a^b f(x) dx &= - \int_b^a f(x) dx \\ \int_a^b f(x) dx &= \int_a^c f(x) dx + \int_c^b f(x) dx \\ \int_a^b (\alpha f(x) + \beta g(x)) dx &= \alpha \int_a^b f(x) dx + \beta \int_a^b g(x) dx\end{aligned}$$

con  $a \leq c \leq b$  e  $\alpha$  e  $\beta$  costanti reali qualunque. In particolare, la terza proprietà indica che l'operazione di integrazione definisce un *funzionale lineare*.

Se  $f(x)$  e  $g(x)$  sono funzioni continue e  $f(x) \leq g(x)$  su  $[a, b]$ , allora

$$\int_a^b f(x) dx \leq \int_a^b g(x) dx$$

In particolare, se  $f(x) \geq 0$  su  $[a, b]$ , allora  $\int_a^b f(x) dx \geq 0$ . Inoltre, se  $f(x)$  è limitata e integrabile secondo Riemann, allora è pure integrabile  $|f(x)|$  e

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx$$

**Proposizione 6.1** (Teorema generalizzato della media) *Siano  $f(x)$  e  $g(x)$  funzioni continue su  $[a, b]$ , con  $g(x) \geq 0$ . Allora, esiste un valore  $\xi$ , con  $a < \xi < b$  tale che*

$$\int_a^b f(x)g(x) dx = f(\xi) \int_a^b g(x) dx \quad (6.3)$$

In particolare, si ha

$$\int_a^b f(x) dx = (b - a) f(\xi)$$

Se

$$m \leq f(x) \leq M \quad \text{per } a \leq x \leq b$$

---

<sup>3</sup>Come esempio classico di funzione non integrabile secondo Riemann, si consideri la *funzione di Dirichlet*

$$d(x) = \begin{cases} 0 & \text{se } x \text{ è razionale} \\ 1 & \text{se } x \text{ è irrazionale} \end{cases}$$

Per tale funzione, per ogni intervallo  $[a, b]$  e per ogni partizione  $P$  di  $[a, b]$  si ha  $L(d; P) = 0$  e  $U(d; P) = b - a$ .

allora

$$m(b-a) \leq \int_a^b f(x) dx \leq M(b-a)$$

Ricordiamo che se  $f(x)$  è, ad esempio, una funzione continua su  $[a, b]$ , una funzione  $F(x)$  è una *primitiva* (o antiderivata) di  $f(x)$ , quando  $F'(x) = f(x)$ . Naturalmente, se  $F(x)$  è una primitiva, le funzioni  $F(x) + c$ , con  $c$  costante arbitraria, sono pure primitive. In effetti, si può mostrare che l'insieme delle primitive di  $f(x)$  può essere descritto nel seguente modo

$$F(x) = \int_a^x f(t) dt + c \quad (6.4)$$

La funzione  $F(x)$  è chiamato *l'integrale indefinito* di  $f(x)$  e indicato con il simbolo  $\int f(x) dx$ .

In altri termini, si può dimostrare che per una funzione continua  $f(x)$  (ma il risultato può essere opportunamente generalizzato) si ha

$$\frac{d}{dx} \int_a^x f(t) dt = f(x), \quad a \leq x \leq b$$

Da (6.4) si ricava il seguente importante risultato.

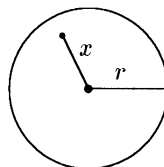
**Teorema 6.1** (Teorema fondamentale del calcolo integrale) *Se  $f(x)$  è una funzione continua su  $[a, b]$  e se  $F(x)$  è l'integrale indefinito di  $f(x)$ , allora*

$$\boxed{\int_a^b f(x) dx = F(b) - F(a)} \quad (6.5)$$

Il teorema precedente permette, in particolare, il calcolo dell'integrale definito, quando l'integrale indefinito è facilmente ottenibile ed è sufficientemente semplice da calcolare. Osserviamo, comunque, che l'operazione di integrazione, inversa a quella di derivazione, può portare a nuove funzioni (trascendentali). Ad esempio,  $\int dx/x$  porta alla funzione logaritmo, che non è una funzione algebrica, mentre l'integrazione  $\int e^{-x^2} dx$  porta a una funzione che non può essere espressa con un numero finito di funzioni algebriche, logaritmiche o esponenziali. Una situazione applicativa è illustrata dal seguente esempio.

► **Esempio 6.1** Nella teoria dei campi elettrici si mostra che l'intensità  $H$  del campo magnetico indotto dalla corrente su un anello cilindrico può essere espressa nel seguente modo

$$H(x) = \frac{4 I r}{r^2 - x^2} \int_0^{\pi/2} \sqrt{1 - \left(\frac{x}{r}\right)^2 \sin^2 \theta} d\theta$$



ove  $I$  è l'intensità della corrente,  $r$  il raggio del cilindro, e  $x$  la distanza dal centro del punto ove si calcola l'intensità magnetica ( $0 \leq x \leq r$ ). Noti  $I$ ,  $r$  e  $x$ , la valutazione di  $H(x)$  richiede il calcolo di un integrale che non può essere espresso in termini di funzioni elementari. Esso rientra nella categoria dei cosiddetti *integrali ellittici*. In effetti, ad un integrale dello stesso tipo si perviene quando si vuole calcolare la lunghezza di una ellisse (cfr. il successivo Esempio 6.2). Lo studio di tali integrali ha dato origine nei secoli scorsi allo studio di opportune funzioni trascendenti, dette appunto *funzioni ellittiche* e di cui si conoscono vari tipi di approssimazioni disponibili in opportune raccolte di tavole (si veda, ad esempio, *CRC Standard Mathematical Tables*). Attraverso tali tavole si ricava, nel caso particolare  $I = 15.3$ ,  $r = 120$  e  $x = 84$ , il valore  $H = 1.355661135$ . Lasciamo come esercizio il confronto tra tale risultato e i valori che si possono ottenere con i vari metodi che saranno esaminati nel seguito del presente capitolo. ■

Osserviamo, inoltre, che anche quando l'integrale indefinito è esprimibile mediante funzioni elementari<sup>4</sup>, non è detto che il suo utilizzo sia numericamente conveniente. Come esempio illustrativo si consideri

$$\int_0^x \frac{dt}{1+t^4} = \frac{\sqrt{2}}{8} \ln \frac{x^2 + \sqrt{2x+1}}{x^2 - \sqrt{2x+1}} + \frac{\sqrt{2}}{4} \arctan \frac{\sqrt{2x}}{1-x^2}$$

per il quale si deve calcolare logaritmi e arcotangenti, che possono essere eseguiti, in aritmetica di macchina, solo con approssimazione.

Terminiamo, osservando che in certe applicazioni la funzione  $f(x)$  può essere nota non in forma analitica, ma solo in forma di tabella, e di conseguenza il risultato (6.5) non è applicabile. In tali situazioni, i metodi che svilupperemo nel seguito rappresentano la sola possibilità di approssimare l'integrale.

Dal risultato (6.5) si ricava facilmente la seguente formula di integrazione.

**Proposizione 6.2** (Integrazione per parti) *Se  $f(x)$  e  $g(x)$  sono due funzioni derivabili con derivata continua su  $[a, b]$  si ha*

$$\int_a^b f(x)g'(x) dx = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x) dx \quad (6.6)$$

Un'altra formula importante è contenuta nel seguente risultato.

**Proposizione 6.3** (Integrazione per sostituzione) *Sia  $g'(u)$  una funzione continua nell'intervallo  $c \leq u \leq d$ , e poniamo  $g(c) = a$  e  $g(d) = b$ . Sia inoltre  $f(x)$  una funzione continua in tutti i punti  $x = g(u)$ , con  $c \leq u \leq d$ . Allora,*

$$\int_a^b f(x) dx = \int_c^d f(g(u))g'(u) du$$

<sup>4</sup>Segnaliamo in questo senso l'utilizzo dei linguaggi simbolici; in particolare MATHEMATICA, DERIVE, MACSYMA.



► **Esempio 6.2** La lunghezza dell'arco di mezza ellisse  $(x/a)^2 + (y/b)^2 = 1$ ,  $y \geq 0$ ,  $b > 0$ , può essere calcolata mediante il seguente integrale

$$L = \int_{-a}^a (dx^2 + dy^2)^{\frac{1}{2}} = \int_{-a}^a (1 + (y')^2)^{\frac{1}{2}} dx$$

Passando alle coordinate polari  $x = a \cos \phi$ ,  $y = b \sin \phi$ ,  $\phi \in [0, \pi]$ , si ottiene

$$L = \int_{-\pi}^0 (a^2 \sin^2 \phi + b^2 \cos^2 \phi)^{\frac{1}{2}} d\phi$$

dal quale, posto  $k^2 = 1 - a^2/b^2$  e  $\cos^2 \phi = 1 - \sin^2 \phi$ , si ha  $L = b \int_{-\pi}^0 (1 - k^2 \sin^2 \phi)^{1/2} d\phi$ . Tale integrale è chiamato un *integrale ellittico di primo genere* e il suo valore può essere approssimato mediante opportuni sviluppi in serie o mediante metodi numerici. ■

Mediante un opportuno utilizzo delle due regole precedenti è, talvolta, possibile trasformare un integrale dato in integrali risolvibili analiticamente, o in una forma più conveniente dal punto di vista numerico. Esse rappresentano, quindi, un prezioso strumento di studio preliminare del problema dato (*preprocessing*).

**Integrali multipli** Limitandoci al caso di due dimensioni, sia  $f(x, y)$  una funzione definita sul rettangolo  $R := a \leq x \leq b$ ,  $c \leq y \leq d$ . Consideriamo, quindi, la partizione di  $R$  ottenuta a partire dalle suddivisioni dei lati  $a = x_0 < x_1 < \dots < x_n = b$  e  $c = y_0 < y_1 < \dots < y_m = d$ . Sia  $R_{ij}$  il rettangolino  $x_i \leq x \leq x_{i+1}$ ,  $y_j \leq y \leq y_{j+1}$  e  $p_{ij}$  un generico punto in  $R_{ij}$  (cfr. Figura 6.1).

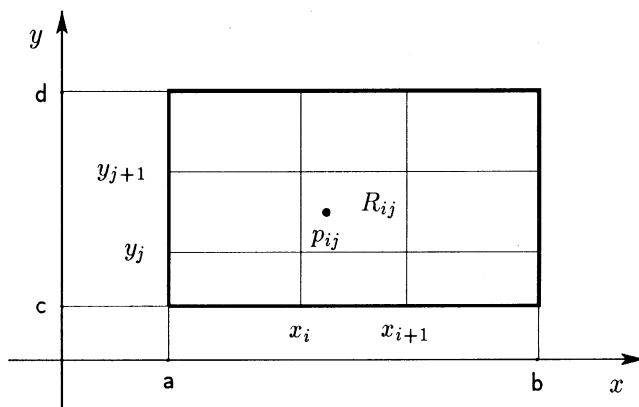


Figura 6.1: Partizione di un rettangolo.

Se  $G$  è una particolare partizione, indichiamo con  $d(G)$  la massima diagonale dei rettangoli  $R_{ij}$ . Allora, l'integrale doppio  $\iint_R f(x, y) dx dy$  esiste ed ha il valore  $I$  se

e solo se, dato un  $\epsilon > 0$ , si può trovare un valore  $\delta > 0$ , tale che

$$\left| I - \sum_{i,j} f(p_{ij})(x_{i+1} - x_i)(y_{j+1} - y_j) \right| \leq \epsilon \quad (6.7)$$

per ogni partizione  $G$  con  $d(G) \leq \delta$  e per ogni scelta dei punti  $p_{ij}$ . Si può dimostrare che una funzione continua sul rettangolo limitato  $R$  è integrabile su  $R$  (più in generale, si può mostrare che risulta integrabile una funzione che è limitata su  $R$  ed è inoltre continua, con eccezione al più di un insieme di punti con area nulla).

Ricordiamo la seguente importante proprietà (*teorema di Fubini*) valida per una funzione  $f(x, y)$ , ad esempio, continua su  $R$

$$\iint_R f(x, y) dx dy = \int_a^b \left( \int_c^d f(x, y) dy \right) dx = \int_c^d \left( \int_a^b f(x, y) dx \right) dy \quad (6.8)$$

Attraverso tale proprietà il calcolo dell'integrale doppio è ricondotto a due integrazioni successive in una sola variabile.

Per definire gli integrali doppi su una regione limitata e sufficientemente regolare  $\Omega$  di forma non rettangolare, si può utilizzare, in maniera schematica, la seguente idea. Si introduce un rettangolo  $R$  che contiene  $\Omega$  e si prolunga in  $R$  la definizione di  $f(x, y)$ , ponendo  $f(x, y) = 0$  se  $(x, y) \notin \Omega$ . Si definisce, quindi  $\iint_{\Omega} f(x, y) dx dy = \iint_R f(x, y) dx dy$ .

La proprietà (6.8) si estende nel seguente modo. Siano  $\phi(x)$  e  $\psi(x)$  due funzioni continue su  $[a, b]$ , con  $\phi(x) \leq \psi(x)$ . Indicata con  $\Omega$  la regione definita da  $a \leq x \leq b$ ,  $\phi(x) \leq y \leq \psi(x)$ , si ha

$$\iint_{\Omega} f(x, y) dx dy = \int_a^b \left( \int_{\phi(x)}^{\psi(x)} f(x, y) dy \right) dx \quad (6.9)$$

Una regione  $\Omega$  come quella ora definita è chiamata *semplice rispetto alle parallele all'asse  $y$* . In modo analogo si definisce una regione semplice rispetto alle parallele all'asse  $x$  e per essa vale un risultato analogo a (6.9).

Per terminare, ricordiamo la seguente formula di *integrazione per sostituzione*. Sia

$$x = \phi(u, v), \quad y = \psi(u, v)$$

una trasformazione biunivoca tra l'insieme  $\Omega'$  e l'insieme  $\Omega$ , con  $\phi(u, v)$  e  $\psi(u, v)$  derivabili con continuità su  $\Omega'$ . Sia, inoltre,  $J(u, v)$  il determinante della matrice jacobiana della trasformazione, ossia

$$J(u, v) = \begin{vmatrix} \partial\phi/\partial u & \partial\phi/\partial v \\ \partial\psi/\partial u & \partial\psi/\partial v \end{vmatrix}$$

Nell'ipotesi, allora, che  $J \neq 0$  in  $\Omega'$ , si ha la seguente formula

$$\iint_{\Omega} f(x, y) dx dy = \iint_{\Omega'} f(\phi(u, v), \psi(u, v)) |J(u, v)| du dv \quad (6.10)$$

Segnaliamo, in particolare, le seguenti trasformazioni.

$$\begin{array}{l} \text{coordinate polari nel piano} \\ \text{coordinate cilindriche in } \mathbb{R}^3 \\ \text{coordinate sferiche in } \mathbb{R}^3 \end{array} \left\{ \begin{array}{l} x = r \cos \phi \\ y = r \sin \phi \\ \\ x = r \cos \phi \\ y = r \sin \phi \\ z = z \\ \\ x = r \sin \phi \sin \theta \\ y = r \sin \phi \cos \theta \\ z = r \cos \phi \end{array} \right. \quad \begin{array}{l} J(r, \phi) = r \\ J(r, \phi, z) = r \\ J(r, \phi, \theta) = -r^2 \sin \phi \end{array}$$

**Integrali impropri** Gli integrali per i quali o l'intervallo, o la funzione integranda non sono limitati, vengono chiamati integrali impropri e sono definiti come limiti di opportuni integrali propri.

**Definizione 6.1** Se  $f(x)$  è una funzione definita su  $[a, b]$  e non limitata nell'intorno del punto  $x = a$ , si pone

$$\int_a^b f(x) dx = \lim_{r \rightarrow a^+} \int_r^b f(x) dx$$

quando il limite esiste finito.

Una definizione analoga si ha quando la funzione  $f(x)$  è non limitata nell'intorno di un punto  $x = c$ , con  $a < c \leq b$ . Condizioni sufficienti per l'esistenza dell'integrale improprio possono essere date in termini di ordine di infinito della funzione integranda.

Si definisce, inoltre, *valore principale di Cauchy* dell'integrale, il seguente limite, quando esiste finito

$$P \left( \int_a^b f(x) dx \right) = \lim_{r \rightarrow 0^+} \left[ \int_a^{c-r} f(x) dx + \int_{c+r}^b f(x) dx \right]$$

Osserviamo che può esistere il valore principale di Cauchy, senza che esista l'integrale improprio.

**Definizione 6.2** Se  $f(x)$  è una funzione definita su  $[0, +\infty]$ , si pone quando il limite esiste finito

$$\int_0^{\infty} f(x) dx = \lim_{r \rightarrow \infty} \int_0^r f(x) dx$$

In maniera analoga si definiscono gli integrali  $\int_a^{\infty} f(x) dx$ , con  $a$  qualunque, e il valore principale di Cauchy su  $(-\infty, \infty)$ .

**Formule di quadratura** La maggior parte dei metodi numerici per approssimare l'integrale (6.1) possono essere inquadrati nel seguente contesto generale. Per la funzione integranda  $f(x)$  si trova una famiglia di approssimanti  $\{f_n(x) \mid n \geq 1\}$  e si definisce

$$I_n(f) := \int_a^b f_n(x) dx = I(f_n); \quad E_n(f) = I(f) - I_n(f) = \int_a^b [f(x) - f_n(x)] dx$$

ove  $E_n(f)$  rappresenta l'errore di troncamento. Lo studio di  $E_n$  può essere, in questo modo, ricondotto a quello dell'approssimazione  $f(x) - f_n(x)$ . Naturalmente, la sostituzione di  $I(f)$  con  $I(f_n)$  è utile soltanto se  $I(f_n)$  è calcolabile in modo semplice. Per questo motivo, le funzioni  $f_n$  vengono scelte, in generale, nell'ambito dei polinomi o dei polinomi a tratti, e la tecnica utilizzata per la costruzione dell'approssimanti  $f_n$  è quella della interpolazione, o delle spline. I vari metodi differiscono tra loro, oltre che per la tecnica utilizzata, per il numero e la locazione dei punti di interpolazione. Nel seguito, considereremo in particolare le *formule di Newton-Cotes*, che corrispondono alla tecnica di interpolazione, con punti equidistanti. Una scelta più opportuna dei punti di interpolazione porta alle formule di Gauss, che risultano esatte per polinomi di grado superiore rispetto a quelle di Newton-Cotes. L'idea della interpolazione polinomiale a tratti (spline) porta alle cosiddette *formule composte*, che rispetto alle formule precedenti hanno il vantaggio di utilizzare, su ogni intervallo, polinomi di grado relativamente basso. Successivamente, analizzeremo le formule di quadratura che risultano dall'applicazione dell'idea della *estrapolazione*.

Come bibliografia sulle formule di quadratura segnaliamo, in particolare, Davis e Rabinowitz [44], Engels [54], Stroud e Secrest [149].

## 6.1 Formule di Newton–Cotes

L'intervallo  $[a, b]$  è diviso in  $n$  sottointervalli di ampiezza  $h = (b - a)/n$  usando i punti  $x_i = a + ih$ ,  $i = 0, 1, \dots, n$ . Con riferimento all'integrale (6.1), la funzione integranda  $f$  è approssimata dal polinomio di interpolazione  $p_n(x)$  di grado al più  $n$  relativo ai *nodi*  $(x_i, f(x_i))$ ,  $i = 0, 1, \dots, n$ . Utilizzando la rappresentazione di Lagrange del polinomio di interpolazione, data da

$$p_n(x) = \sum_{i=0}^n L_i(x) f(x_i)$$

si ottiene

$$\int_a^b p_n(x) dx = \sum_{i=0}^n \left( \int_a^b L_i(x) dx \right) f(x_i) = \sum_{i=0}^n A_i f(x_i)$$

ove i coefficienti  $A_i$ , detti i *pesi* della formula, possono essere ottenuti integrando i polinomi  $L_i(x)$ . Integrando l'uguaglianza  $f(x) = p_n(x) + R(x)$ , ove  $R(x)$  indica l'errore di interpolazione relativo alla funzione  $f(x)$ , si ricava la seguente relazione

$$\int_a^b f(x) dx = \sum_{i=0}^n A_i f(x_i) + R_T(f)$$

nella quale il termine  $\sum_{i=0}^n A_i f(x_i)$  rappresenta una particolare *formula di quadratura di Newton–Cotes*<sup>5</sup> e  $R_T(f) = \int_a^b R(x) dx$  è il corrispondente errore di troncamento. Per costruzione, si ha  $R_T(f) = 0$  per  $f(x) = x^k$ , con  $k = 0, 1, \dots, n$ . In realtà, si può facilmente vedere che le formule corrispondenti ad un valore di  $n$  pari risultano esatte anche per  $x^{n+1}$ .

**Formula del trapezio** Per  $n = 1$  si ottiene la seguente formula, nota come *formula del trapezio*

$$\int_{x_0}^{x_1} f(x) dx = \frac{h}{2}(f_0 + f_1) + R_T(f)$$

ove per brevità si è posto  $f_i := f(x_i)$ . Assumendo che la funzione integranda  $f(x)$  sia dotata di derivata seconda continua su  $[a, b]$ , si può mostrare che l'errore di troncamento  $R_T$  ha la seguente rappresentazione

$$R_T(f) = -\frac{h^3}{12} f''(\eta), \quad x_0 < \eta < x_1 \quad (6.11)$$

**Formula di Simpson** Per  $n = 2$ , ossia interpolando la funzione  $f(x)$  mediante un polinomio di secondo grado nei punti equidistanti:  $x_0, x_1, x_2$ , si ottiene la seguente formula

$$\int_{x_0}^{x_2} f(x) dx = \frac{h}{3}(f_0 + 4f_1 + f_2) + R_S(f) \quad (6.12)$$

nota come *formula di Simpson* (o anche, di Cavalieri-Simpson<sup>6</sup>).

<sup>5</sup>I. Newton (1642-1727), R. Cotes (1682-1716). I lavori di Cotes furono pubblicati postumi sotto il titolo di *Harmonia mensurarum*. Pare che alla morte prematura di Cotes, Newton abbia esclamato “*Had Cotes lived, we might have known something*”.

<sup>6</sup>In effetti, tale formula è stata introdotta da Cavalieri nel 1639, e successivamente da James Gregory (1668) e Thomas Simpson (1743). Il nome di T. Simpson (1710-1761) è noto in matematica in particolare per la formula di quadratura, ma il suo lavoro diede contributi importanti in geometria, teoria della probabilità e astronomia. I nomi di seno, coseno, tangente e cotangente per le funzioni trigonometriche furono introdotti da Simpson. La formula (6.12) è anche nota come *Kepler's Barrel Rule*. In effetti, J. Kepler (1571-1630) utilizzò tale regola per il calcolo della capacità di un barile di vino (*Stereometria doliorum*, 1612).

Supponendo la funzione integranda  $f$  dotata di derivata quarta continua su  $[a, b]$ , si ha

$$R_S(f) = -\frac{h^5}{90} f^{(4)}(\eta) \quad x_0 < \eta < x_2$$

Le formule del trapezio e di Simpson sono illustrate in Figura 6.2, insieme alla formula corrispondente a  $n = 3$ .

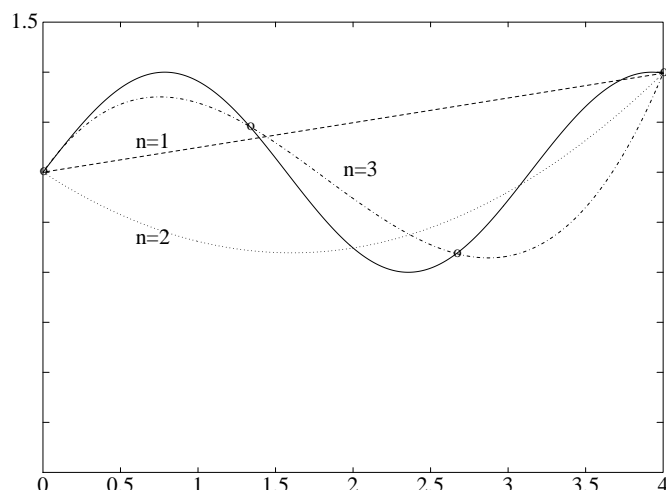


Figura 6.2: Illustrazione delle formule di Newton-Cotes, per  $n = 1$ ,  $n = 2$  e  $n = 3$ .

▼ **Osservazione 6.1** Le formule di Newton-Cotes considerate in precedenza sono anche chiamate formule di tipo chiuso, in relazione al fatto che utilizzano come nodi anche gli estremi dell'intervallo  $[a, b]$ . Per talune applicazioni, in particolare nell'ambito della risoluzione delle equazioni differenziali, possono presentare interesse anche formule di tipo aperto nelle quali i nodi sono solo punti interni all'intervallo. Le formule di Gauss che considereremo nel seguito rappresentano un esempio importante di tale tipo di formule. Come vedremo, tuttavia, per tali formule i nodi non sono equidistanti. Tra le formule di Newton-Cotes di tipo aperto, con nodi quindi equidistanti e interni all'intervallo, ricordiamo le seguenti

$$n = 2 \quad \int_{x_0}^{x_2} f(x) dx = 2hf(x_1) + \frac{h^3}{3} f''(\eta) \quad (6.13)$$

$$n = 3 \quad \int_{x_0}^{x_3} f(x) dx = \frac{2h}{2}[f(x_1) + f(x_2)] + \frac{3h^3}{4} f''(\eta) \quad (6.14)$$

Di particolare interesse è la formula (6.13), nota come formula delle tangenti, o formula midpoint. Lasciamo come esercizio la sua interpretazione geometrica, come pure la dimostrazione che la formula risulta esatta per i polinomi di grado 1. ■

► **Esempio 6.3** Consideriamo il calcolo della funzione  $\text{erf}(x)$  nel punto  $x = 1$ , ossia l'integrale

$$I = \text{erf}(1) = \frac{2}{\sqrt{\pi}} \int_0^1 e^{-t^2} dt = 0.842700792949714 \dots$$

Applicando le formule di quadratura esaminate in precedenza, si ottengono i seguenti risultati

$$\begin{aligned} \text{formula midpoint } I &\approx \frac{2}{\sqrt{\pi}} e^{-1/4} = 0.87872578\dots \\ \text{formula dei trapezi } I &\approx \frac{2}{\sqrt{\pi}} \left(\frac{1}{2}\right) [1 + e^{-1}] = 0.771743332\dots \\ \text{formula di Simpson } I &\approx \frac{2}{\sqrt{\pi}} \left(\frac{1}{6}\right) [1 + 4e^{-1/4} + e^{-1}] = 0.84310283\dots \end{aligned}$$

■

### 6.1.1 Convergenza delle formule di quadratura

Indicando come *grado di precisione* di una formula il grado massimo del polinomio per il quale la formula è esatta, si ha che il grado di precisione della formula aumenta con il numero  $n + 1$  dei nodi utilizzati per l'interpolazione. Si potrebbe pensare, allora, che anche l'errore  $E_n = I(f) - I(f_n)$ , per  $f$  funzione non polinomiale, ma sufficientemente regolare, tenda a zero, per  $n \rightarrow \infty$ . Questo risultato può essere, tuttavia, falso, come mostra il seguente esempio, nel quale la funzione è indefinitamente derivabile.

► **Esempio 6.4** Calcoliamo le approssimazioni  $I_n$  del seguente integrale

$$I = \int_{-1}^1 \frac{1}{1 + 25x^2} dx$$

usando le formule di quadratura di Newton–Cotes per  $n = 1, 2, \dots, 8$ . Si ottengono i seguenti risultati.

$n$	$I_n$	$E_n$
1	0.038462	0.510898
2	0.679487	-0.130126
3	0.208145	0.341215
4	0.237400	0.311960
6	0.387045	0.162315
7	0.289899	0.250461
8	0.150049	0.399311

Il valore esatto è  $I = \frac{1}{5}(\arctan 5 - \arctan(-5)) \approx 0.549360$ . I risultati mostrano che l'errore non diminuisce per  $n$  che aumenta. Si può, d'altra parte, dimostrare (cfr. Capitolo 4) che la successione dei polinomi di interpolazione della funzione  $1/(1 + 25x^2)$  corrispondenti ad una successione di suddivisioni dell'intervallo  $[-1, 1]$  in parti uguali non è uniformemente convergente. ■

### 6.1.2 Formule composte

L'esempio analizzato nel paragrafo precedente mostra che vi sono funzioni continue  $f$  per le quali l'approssimazione  $I_n$  calcolata mediante le formule di Newton–Cotes non converge all'integrale  $I$  per  $n \rightarrow \infty$ . In questo senso, una procedura più opportuna consiste nell'approssimare la funzione  $f$  su *sottointervalli* di  $[a, b]$  con polinomi di grado piccolo, ad esempio  $n = 1$ ,  $n = 2$ . Procedendo in questo modo si ottengono le cosiddette *formule di quadratura composte*. Posto, ancora,  $h = (b-a)/n$  e  $x_i = a+ih$  e utilizzando su ogni sottointervallo  $[x_i, x_{i+1}]$  la formula del trapezio, si ottiene, nel caso in cui la funzione integranda abbia la derivata seconda continua

$$\int_a^b f(x) dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x) dx = T(h) - \frac{b-a}{12} h^2 f''(\eta), \quad \eta \in ]a, b[ \quad (6.15)$$

ove

$$T(h) := h\left(\frac{1}{2}f_0 + f_1 + \cdots + f_{n-1} + \frac{1}{2}f_n\right) \quad (6.16)$$

è chiamata la *formula del trapezio composta*. L'espressione dell'errore in (6.15) è ottenuta nel seguente modo. Dalla rappresentazione (6.11) dell'errore su ogni intervallo  $[x_i, x_{i+1}]$  si ha che l'errore corrispondente alla formula composta è dato da

$$\sum_{i=0}^{n-1} \left( -\frac{h^3}{12} f''(\eta_i) \right) = -\frac{b-a}{12} h^2 \frac{\sum_{i=0}^{n-1} f''(\eta_i)}{n} = -\frac{b-a}{12} h^2 f''(\eta)$$

ove  $\eta_i$  sono opportuni punti negli intervalli  $[x_i, x_{i+1}]$  e  $\eta$  è un punto opportuno nell'intervallo  $(a, b)$ . L'ultimo passaggio nella formula precedente utilizza la continuità della derivata seconda  $f''(x)$ .

Se  $n = 2m$ , cioè se l'intervallo  $[a, b]$  è suddiviso in un numero *pari* di sottointervalli, allora l'integrale su ogni sottointervallo  $[x_{2i}, x_{2i+2}]$  può essere approssimato mediante la regola di Simpson; se  $f^{(4)}$  è continua, si ottiene

$$\int_{x_{2i}}^{x_{2i+2}} f(x) dx = \frac{h}{3} (f_{2i} + 4f_{2i+1} + f_{2i+2}) - \frac{h^5}{90} f^{(4)}(\eta_i)$$

con  $\eta_i \in ]x_{2i}, x_{2i+2}[$ , da cui

$$\int_a^b f(x) dx = \sum_{i=0}^{m-1} \int_{x_{2i}}^{x_{2i+2}} f(x) dx = S(h) - \frac{b-a}{180} h^4 f^{(4)}(\eta), \quad \eta \in ]a, b[ \quad (6.17)$$

ove

$$S(h) := \frac{h}{3} (f_0 + 4f_1 + 2f_2 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m}) \quad (6.18)$$

è la *formula di Simpson composta*. Le formule del trapezio e di Simpson composte sono illustrate in Figura 6.3.

Il calcolo della formula  $S(h)$  può essere effettuato come indicato nel seguente algoritmo.



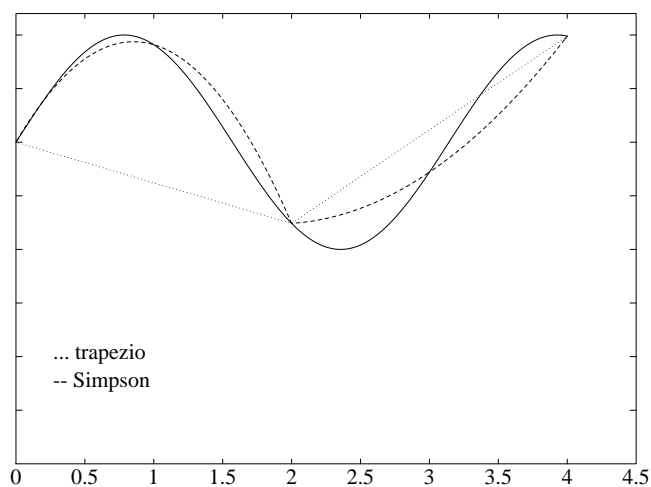


Figura 6.3: Illustrazione delle formule composte del trapezio e di Simpson.

**Algoritmo 6.1** (Formula di Simpson composta) *Calcolo dell'integrale  $I = \int_a^b f(x) dx$ . Input:  $a, b, m$ ; output:  $I_p$ , approssimazione di  $I$ .*

```

Set  $h = (b - a)/(2m)$ 
 $I_0 = f(a) + f(b)$ 
 $I_1 = 0$  somma di  $f(x_{2i-1})$ 
 $I_2 = 0$  somma di  $f(x_{2i})$ 
do  $i = 1, \dots, 2m - 1$ 
   $x = a + ih$ 
  if  $i$  pari then  $I_2 = I_2 + f(x)$ 
  else  $I_1 = I_1 + f(x)$ 
end if
end do
 $I_p = h(I_0 + 2I_2 + 4I_1)/3$ 

```

Come mostrano le formule (6.15) e (6.16), l'errore corrispondente alla formula del trapezio composta (rispettivamente alla formula di Simpson) tende a zero come  $h^2$  (rispettivamente come  $h^4$ ), purché naturalmente la funzione integranda sia opportunamente regolare. Tale comportamento dell'errore è evidenziato nel seguente esempio.

► **Esempio 6.5** Per approssimare il seguente integrale

$$I = \int_0^1 \frac{1}{1+x} dx$$

utilizziamo successivamente la *regola del trapezio*, per  $h = 1, 0.5, 0.25, 0.125$ , e la *regola di Simpson*, per  $h = 0.5, 0.25, 0.125$ . Tenendo presente che il risultato esatto è dato da  $I = \ln 2 \approx 0.693147$ , si ottengono i seguenti risultati

$h$	$T(h)$	$S(h)$	$T(h) - I$	$S(h) - I$
1.	0.750000		0.056853	
0.5	0.708333	0.694444	0.015186	0.001297
0.25	0.697024	0.693254	0.003877	0.000107
0.125	0.694122	0.693155	0.000975	0.000008

I risultati mostrano, come la teoria prevede, che l'errore ottenuto con la regola dei trapezi è diviso per quattro quando il passo  $h$  è dimezzato, mentre nella regola di Simpson è diviso per 16. ■

## 6.2 Formule di Gauss

Nelle formule di Newton–Cotes si utilizzano i valori della funzione integranda in *punti equidistanti* dell'intervallo di integrazione  $[a, b]$  e i pesi  $A_i$  sono calcolati integrando il corrispondente polinomio di interpolazione. In questo modo, se i punti di suddivisione  $x_i$  sono in numero di  $n + 1$ , si ottengono formule che risultano *esatte* per i polinomi di grado almeno  $n$ .

Le *formule di Gauss*<sup>7</sup> sono ottenute, come le formule di Newton–Cotes, mediante interpolazione, ma i nodi  $x_i$  sono scelti in modo da massimizzare il *grado di precisione*, cioè il grado del polinomio per cui le formule sono esatte.

Introduciamo l'idea mediante un esempio.

► **Esempio 6.6** Per  $n = 1$  consideriamo la formula

$$\int_{-1}^1 f(x) dx \approx A_0 f(x_0) + A_1 f(x_1)$$

Possiamo cercare di determinare i quattro parametri  $A_0, A_1, x_0, x_1$ , imponendo che la formula sia esatta per i polinomi  $1, x, x^2, x^3$ . Si ottengono allora le seguenti equazioni

$$\begin{aligned} A_0 + A_1 &= 2 \\ A_0 x_0 + A_1 x_1 &= 0 \\ A_0 x_0^2 + A_1 x_1^2 &= \frac{2}{3} \\ A_0 x_0^3 + A_1 x_1^3 &= 0 \end{aligned}$$

che hanno come soluzione

$$A_0 = A_1 = 1, \quad x_1 = -x_0 = \frac{\sqrt{3}}{3}$$

Si è ottenuta allora la seguente formula

$$\int_{-1}^1 f(x) dx \approx f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right)$$

<sup>7</sup>introdotta da K. F. Gauss in *Methodus nova integralium valores per approximationem inveniendi* (1814).

con grado di precisione 3. Come confronto, ricordiamo che la formula di Simpson ha pure grado di precisione 3, ma tale precisione è ottenuta con tre nodi.

Per terminare l'esempio, osserviamo che la formula ora ottenuta sull'intervallo particolare  $[-1, 1]$ , può essere utilizzata su un intervallo generico  $[a, b]$  mediante la seguente trasformazione di variabile

$$\int_a^b f(t) dt = \frac{b-a}{2} \int_{-1}^1 f\left(\frac{a+b+x(b-a)}{2}\right) dx$$

da cui la seguente espressione della formula di Gauss

$$\int_a^b f(t) dt \approx \frac{b-a}{2} \left[ f\left(\frac{a+b}{2} - \frac{b-a}{2\sqrt{3}}\right) + f\left(\frac{a+b}{2} + \frac{b-a}{2\sqrt{3}}\right) \right]$$

che è esatta per i polinomi di grado tre. Con riferimento all'esempio 6.3, si ottiene per erf(1) il valore 0.84244189... , che è più accurato del risultato fornito dalla formula di Simpson a tre punti. ■

La procedura illustrata nell'esempio precedente può essere generalizzata utilizzando la nozione di polinomi ortogonali. Tali polinomi sono stati introdotti nel precedente Capitolo 4; ora ne ricorderemo gli elementi essenziali.

A partire dalla base  $1, x, x^2, x^3, \dots$ , è possibile costruire una successione di polinomi  $P_n(x)$ , di grado  $n$  per  $n \geq 0$ , con la seguente proprietà

$$\int_{-1}^1 P_n(x) P_m(x) dx = 0, \quad n \neq m$$

$$\int_{-1}^1 [P_n(x)]^2 dx = c(n) \neq 0$$

Tale proprietà significa che i polinomi  $P_n(x)$ ,  $n = 0, 1, 2, \dots$ , detti *polinomi di Legendre*, risultano *ortogonali* sull'intervallo  $[-1, 1]$ . La proprietà di ortogonalità generalizza quella di perpendicolarità di due vettori nello spazio euclideo  $\mathbb{R}^n$ . I primi elementi della successione sono dati da

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x)$$

I successivi possono essere costruiti attraverso la seguente formula ricorrente

$$P_n(x) = \frac{2n-1}{n} x P_{n-1}(x) - \frac{n-1}{n} P_{n-2}(x)$$

È evidente dalla definizione che per ogni  $n$  fissato il polinomio  $P_n(x)$  ha grado esattamente  $n$ . Si può inoltre mostrare che la successione  $\{P_n\}$ ,  $n = 0, 1, \dots$  costituisce una base dei polinomi, nel senso che un generico polinomio di grado  $n$  può essere

espresso in maniera unica come combinazione lineare dei polinomi  $P_0(x)$ ,  $P_1(x)$ ,  $\dots$ ,  $P_n(x)$ . Ad esempio, si può mostrare che

$$x^4 + 3x^3 - 2x^2 + 2x - 1 \equiv -\frac{22}{15}P_0(x) + \frac{19}{5}P_1(x) - \frac{16}{21}P_2(x) + \frac{6}{5}P_3(x) + \frac{8}{35}P_4(x)$$

Un'altra proprietà importante, che è conseguenza della ortogonalità, è il fatto che per ogni  $n$  fissato il polinomio  $P_n(x)$  ha  $n$  zeri reali e distinti nell'intervallo aperto  $(-1, 1)$ .

Si può allora mostrare che, se come nodi della formula di integrazione vengono utilizzati gli zeri del polinomio di Legendre  $P_n(x)$  e i pesi sono calcolati in modo che la formula sia di interpolazione, si ottiene una particolare *formula di Gauss*, che risulta esatta per polinomi di grado fino a  $2n - 1$ . Nella Tabella 6.1 sono riportati i nodi e i pesi relativi alle prime formule.

n	$x_i$	$A_i$
1	$\pm.57735\ 02692$	1.0
2	$\pm.77459\ 66692$ 0.0	0.55555 55556 0.88888 88889
3	$\pm.86113\ 63116$ $\pm.33998\ 10436$	0.34785 48451 0.65214 51549
4	$\pm.90617\ 98459$ $\pm.53846\ 93101$ 0.0	0.23692 68851 0.47862 86705 0.56888 88889
5	$\pm.93246\ 95142$ $\pm.66120\ 93865$ $\pm.23861\ 91861$	0.17132 44924 0.36076 15730 0.46791 39346

Tabella 6.1: Nodi e pesi per la formula di Gauss-Legendre.

Altre formule di tipo Gauss possono essere ottenute utilizzando differenti polinomi ortogonali. Di particolare interesse sono quelle che utilizzano i *polinomi di Chebichev*. Tali polinomi, indicati usualmente con  $T_n(x)$ , sono ortogonali rispetto alla funzione peso  $w(x) = 1/\sqrt{1-x^2}$ , ossia verificano le relazioni

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_n(x) T_m(x) dx = 0, \quad n \neq m$$

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}} [T_n(x)]^2 dx = c(n) \neq 0$$

Partendo da  $T_0(x) = 1$ ,  $T_1(x) = x$ , i polinomi successivi sono generati dalla relazione ricorrente

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$$

Un altro modo di esprimere i polinomi di Chebichev è il seguente

$$T_n(x) = \cos(n \arccos x)$$

dal quale si vede che gli zeri del polinomio di Chebichev  $T_{n+1}(x)$  di grado  $n+1$  sono dati esplicitamente dalla formula

$$x_i = \cos \frac{(2i+1)\pi}{2n+2}, \quad i = 0, 1, \dots, n$$

Assumendo tali punti come nodi, si ottiene la seguente formula

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}} f(x) dx \approx \sum_{i=0}^n A_i f(x_i), \quad A_i = \frac{\pi}{n+1}, \quad i = 0, 1, \dots, n$$

che risulta esatta per le funzioni  $f(x)$  che sono polinomi di grado inferiore o uguale a  $2n+1$ .

Terminiamo, ricordando che esistono formule di Gauss per l'approssimazione di integrali su intervalli illimitati, del tipo  $(0, +\infty)$  e  $(-\infty, +\infty)$ . Esse utilizzano come nodi gli zeri di opportuni polinomi ortogonali, noti come polinomi di Laguerre e di Hermite.

### 6.2.1 Formule di Lobatto

Come abbiamo visto, le formule di Gauss sono formule di tipo aperto. In alcune applicazioni vi può essere la necessità, o l'opportunità, di utilizzare come nodi anche gli estremi dell'intervallo. Le formule di Gauss possono allora essere modificate nel seguente modo. Riferendoci, ancora, all'intervallo  $[-1, 1]$ , i punti  $-1$  e  $1$  sono utilizzati come nodi, mentre gli altri nodi sono dati dagli zeri dell'equazione  $P'_{n-1}(x) = 0$ , ove  $P_{n-1}(x)$  è il polinomio di Legendre di grado  $n-1$ . Per interpolazione, si ottiene allora la seguente formula, detta *formula di Lobatto*<sup>8</sup>

$$\int_{-1}^1 f(x) dx \approx \alpha f(-1) + \alpha f(1) + \sum_{k=1}^{n-2} \alpha_k f(x_k) \quad (6.19)$$

ove  $\alpha = 2/n(n-1)$  e  $\alpha_k = \alpha/P_{n-1}(x_k)^2$ . Le ascisse e i pesi sono indicati nella Tabella 6.2. Calcolando la funzione integranda nei punti  $\pm 1$  si perdono due gradi di libertà, e una formula di Lobatto a  $n$  punti è esatta solo per polinomi di grado  $2n-3$ , in confronto a  $2n-1$  per una corrispondente formula di Gauss-Legendre.

### 6.2.2 Formule di quadratura di Gauss-Kronrod

Indichiamo con  $G_n$  la formula di Gauss a  $n$  punti, ossia

$$G_n = \sum_{i=1}^n w_i f(x_i) \quad (6.20)$$

<sup>8</sup>R. Lobatto, "Lessen over de integral-rekening" den Haag (1852).

n	$x_k$	$\alpha_k$
3	$\pm 1.00000\ 00000$	0.33333 33333
	0.00000 00000	1.33333 33333
4	$\pm 1.00000\ 00000$	0.16666 66667
	$\pm 0.44721\ 35955$	0.83333 33333
5	$\pm 1.00000\ 00000$	0.10000 00000
	$\pm 0.65465\ 36707$	0.54444 44444
	0.00000 00000	0.71111 11111
6	$\pm 1.00000\ 00000$	0.06666 66667
	$\pm 0.76505\ 53239$	0.37847 49563
	$\pm 0.28523\ 15165$	0.55485 83770

Tabella 6.2: Nodi e pesi per la formula di Lobatto.

Tale formula risulta esatta per i polinomi fino al grado  $2n - 1$ , con un costo di  $n$  valutazioni della funzione. Allo scopo di *valutare* l'errore commesso nell'applicazione della formula, Kronrod (1965) ha considerato formule del seguente tipo

$$K_{2n+1} = \sum_{i=1}^n a_i f(x_i) + \sum_{j=1}^{n+1} b_j f(y_j)$$

che condividono  $n$  nodi con le formule  $G_n$ , dimostrando che è possibile trovare i  $3n+2$  parametri  $a, b$  e  $y_j$  in modo tale che la formula  $K_{2n+1}$  abbia grado di precisione  $3n+1$  (cfr. la Tabella 6.3 per  $n = 7$ ). Le due formule  $(G_n, K_{2n+1})$  sono chiamate una *coppia Gauss-Kronrod*. Il costo per il calcolo della coppia è dato da  $2n+1$  valutazioni della funzione integranda. Osserviamo che tale costo è lo stesso di quello relativo al calcolo di  $G_{2n+1}$ , che ha grado di precisione  $4n+1$ . Il motivo, comunque, per preferire la coppia di Gauss-Kronrod è dovuto al fatto che la differenza  $|G_n - K_{2n+1}|$  è una stima dell'errore che si commette, quando l'integrale è approssimato dalla formula  $K_{2n+1}$ . Più precisamente, la sperimentazione ha suggerito la seguente stima

$$\left| \int_a^b f(x) dx - K_{2n+1} \right| \approx (200 |G_n - K_{2n+1}|)^{1.5}$$

Ad esempio, per il calcolo di  $\operatorname{erf}(1)$  si ottengono i seguenti valori

$$G_7 = 0.842700792948824892; \quad K_{15} = 0.842700792949714861;$$

con  $(200 |G_7 - K_{15}|)^{1.5} = 2.10^{-15}$ . L'errore attuale, quando si assume  $K_{15}$ , è circa  $2 \cdot 10^{-17}$ . Per un opportuno confronto, osserviamo che applicando una formula dei trapezi composta con 128 valutazioni della funzione si ha un errore dell'ordine di  $4 \cdot 10^{-6}$ . In effetti, le formule di Gauss-Kronrod rappresentano uno degli algoritmi più efficienti per il calcolo di integrali di tipo generale. La scelta standard è la coppia  $(G_7, K_{15})$ .

$G_7$		$K_{15}$	
nodi	pesi	nodi	pesi
$\pm 0.94910\ 79123\ 42758$	0.12948 49661 68870	$\pm 0.99145\ 53711\ 20813$	0.02293 53220 10529
$\pm 0.74153\ 11855\ 99394$	0.27970 53914 89277	$\pm 0.94910\ 79123\ 42758$	0.06309 20926 29979
$\pm 0.40584\ 51513\ 77397$	0.38183 00505 05119	$\pm 0.86486\ 44233\ 59769$	0.10479 00103 22250
0.00000 00000 00000	0.41795 91836 73469	$\pm 0.74153\ 11855\ 99394$	0.14065 32597 15525
		$\pm 0.58608\ 72354\ 67691$	0.16900 47266 39267
		$\pm 0.40584\ 51513\ 77397$	0.19035 05780 64785
		$\pm 0.20778\ 49550\ 07898$	0.20443 29400 75298
		0.00000 00000 00000	0.20948 21410 84728

Tabella 6.3: Nodi e pesi per la coppia  $(G_7, K_{15})$  di Gauss-Kronrod.

### 6.3 Formule adattive

La scelta della suddivisione in parti *uguali* in una formula di quadratura *composta* può non essere appropriata quando si integra una funzione con comportamento altamente variato nell'intervallo di integrazione, cioè con variazioni *larghe* su alcune parti e variazioni *piccole* su altre. Il significato di variazione è precisato dal comportamento delle *derivate della funzione*. Analizzando, come esemplificazione, la formula composta di Simpson, abbiamo visto che l'errore relativo a tale formula dipende dal comportamento nell'intervallo di integrazione della derivata quarta della funzione integranda. Se si vuole quindi che l'errore sia stimato in maniera *uniforme* su tutto l'intervallo, e al minimo costo, sarebbe conveniente utilizzare un passo di suddivisione *piccolo* ove la derivata quarta è *grande* e viceversa un passo più grande, con conseguente risparmio di valutazioni, ove la derivata quarta è *piccola*.

Ciò che rende nelle implementazioni concrete spesso inattuabile una procedura del tipo ora considerato è il fatto che in generale *non è possibile* conoscere a priori il comportamento della funzione integranda e delle sue derivate.

Con il termine *formula adattiva* si intende un procedimento mediante il quale, con successive applicazioni di una determinata formula, è possibile *avere un'idea* del comportamento della funzione e *scoprire* eventuali singolarità della funzione o delle successive derivate. In questo modo diviene possibile *adattare automaticamente* il passo di suddivisione.

Può essere interessante a questo punto riflettere su come si presenta *nel concreto* la risoluzione approssimata di un particolare integrale (e analogamente, più in generale, di un qualunque problema numerico).

È fissata una precisione  $\epsilon$  da raggiungere, determinata in base agli scopi del calcolo e all'accuratezza dei dati del problema. Si vuole, quindi, calcolare una quantità che differisca dal valore dell'integrale esatto per meno di  $\epsilon$ . Dal punto di vista pratico è ovviamente importante che la quantità approssimata sia ottenuta con il *minimo* numero di *valutazioni* della funzione. Una formula adattiva è in sostanza una risposta a questa esigenza pratica.

A titolo di esemplificazione, in questo paragrafo svilupperemo l'idea utilizzando in particolare la formula di Simpson.

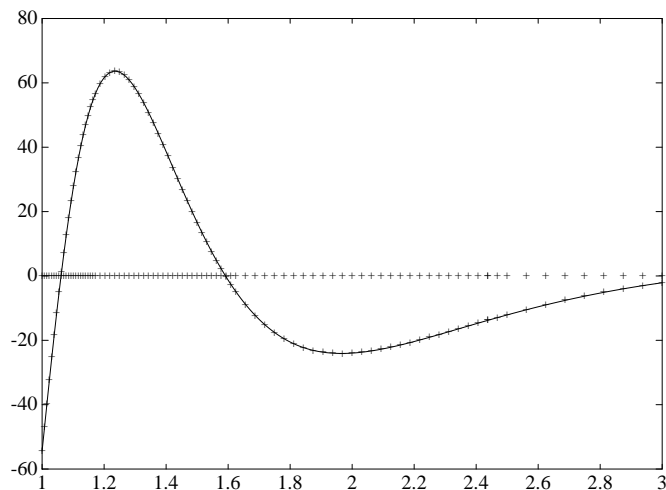


Figura 6.4: Rappresentazione dei punti di valutazione in una formula di Simpson adattiva.

### 6.3.1 Formula di Simpson adattiva

Dato  $\epsilon > 0$ , si vuole ottenere una approssimazione  $AI$  dell'integrale di  $f(x)$  su un intervallo limitato  $(a, b)$  tale che

$$\left| \int_a^b f(x) dx - AI \right| \leq \epsilon$$

A tale scopo si incomincia ad applicare la formula di Simpson con passo  $h = (b-a)/2$ . Si ottiene

$$\int_a^b f(x) dx = S(a, b) - \frac{b-a}{180} h^4 f^{(4)}(\xi), \quad \xi \in (a, b)$$

ove

$$S(a, b) = \frac{h}{3} [f(a) + 4f(a+h) + f(b)]$$

Il *passo successivo* consiste nel cercare di stimare l'errore  $I - S(a, b)$  senza determinare esplicitamente la funzione  $f^{(4)}(x)$ . Per fare questo, applichiamo la formula di Simpson con passo  $h/2 = (b-a)/4$  ottenendo

$$\begin{aligned} \int_a^b f(x) dx = & \frac{h}{6} \overbrace{[f(a) + 4f(a + \frac{h}{2}) + f(a+h)]}^{S(a, \frac{a+b}{2})} + \frac{h}{6} \overbrace{[f(a+h) + 4f(a + \frac{3}{2}h) + f(b)]}^{S(\frac{a+b}{2}, b)} - \\ & - \left(\frac{h}{2}\right)^4 \frac{b-a}{180} f^{(4)}(\bar{\xi}), \quad \bar{\xi} \in (a, b) \end{aligned}$$



Si ha quindi

$$\int_a^b f(x) dx = S(a, \frac{a+b}{2}) + S(\frac{a+b}{2}, b) - \frac{1}{16} \frac{h^5}{90} f^{(4)}(\bar{\xi})$$

Supponiamo ora

$$f^{(4)}(\xi) = f^{(4)}(\bar{\xi}) \quad (6.21)$$

Si tratta di un'ipotesi in generale non verificata, ma che permette di ottenere utili indicazioni sull'errore. Il *successo* della tecnica dipenderà da quanto *poco* l'ipotesi precedente si discosta dal vero.

Con facili calcoli, dalle equazioni precedenti si ricava

$$\frac{h^5}{90} f^{(4)}(\xi) \approx \frac{16}{15} [S(a, b) - S(a, \frac{a+b}{2}) - S(\frac{a+b}{2}, b)]$$

da cui la seguente stima

$$\left| \int_a^b f(x) dx - [S(a, \frac{a+b}{2}) + S(\frac{a+b}{2}, b)] \right| \approx \frac{1}{15} |S(a, b) - S(a, \frac{a+b}{2}) - S(\frac{a+b}{2}, b)|$$

Se pertanto

$$\left| S(a, b) - S(a, \frac{a+b}{2}) - S(\frac{a+b}{2}, b) \right| < 15\epsilon \quad (6.22)$$

allora si ha

$$\left| \int_a^b f(x) dx - [S(a, \frac{a+b}{2}) + S(\frac{a+b}{2}, b)] \right| \leq \epsilon$$

In pratica, per tenere conto dell'ipotesi (6.21), anziché  $15\epsilon$  si prenderà una stima più *conservativa*, ad esempio  $10\epsilon$ .

Se la disuguaglianza (6.22) non è verificata, si applica la procedura di stima dell'errore su ogni intervallo  $[a, \frac{a+b}{2}]$ ,  $[\frac{a+b}{2}, b]$ . Nel caso in cui la stima su ciascuno degli intervalli segnali un errore  $< \epsilon/2$ , allora il procedimento termina. Altrimenti, se su uno degli intervalli la stima dell'errore non passa il test, tale intervallo viene ulteriormente suddiviso e ognuno dei sottointervalli viene esaminato per vedere se l'errore è  $< \epsilon/4$ , e così di seguito.

► **Esempio 6.7** A titolo di illustrazione si consideri l'integrale

$$\int_1^3 \frac{100}{x^2} \sin \frac{10}{x} dx$$

il cui valore esatto è dato da  $-1.42602475\dots$ . Utilizzando la procedura adattiva precedente con  $\epsilon = 1.E-4$  si trova il valore  $-1.426021$  con un errore  $\approx 3.E-6$ . Tale risultato viene ottenuto con 84 valutazioni della funzione integranda. Nella Figura 6.4 è riportata la distribuzione dei punti di valutazione. Con la formula di Simpson a passo uniforme con 256 valutazioni si ottiene un errore  $2.4 E-6$ . ■

## 6.4 Formule di estrapolazione

L'idea del *metodo di estrapolazione* o metodo di Richardson, una delle più interessanti del calcolo numerico, si applica in una situazione generale del seguente tipo. Sia  $\tau_0$  una quantità incognita che si calcola come *limite* di una successione di quantità calcolabili  $T(y)$  dipendenti da un parametro  $y$ . Per fissare le idee supporremo  $y \rightarrow 0$ .

Nella situazione ora descritta rientrano in particolare le *formule composte* studiate nel paragrafo precedente ( $y$  è allora il passo della discretizzazione), ma più in generale le *formule alle differenze* per la derivazione numerica studiate nel Capitolo 4 e i *sistemi dinamici discreti* per la ricerca del punto fisso di una trasformazione considerati nel Capitolo 5.

L'idea di Richardson parte dal presupposto che il risultato finale del calcolo numerico richieda la valutazione di  $T(y)$  per successivi valori del parametro  $y$ . Questa è, in effetti, la situazione *reale* quando, come accade in generale, *non è noto a priori* il valore del parametro  $y$  che fornisce la stima di  $\tau_0$  corrispondente ad una precisione prefissata. Osserviamo a questo proposito che le maggiorazioni dell'errore di troncamento, come quelle che abbiamo ricavato nel paragrafo precedente a proposito delle formule di quadratura, non possono, in generale, essere utilizzate direttamente, in quanto esse richiedono la conoscenza di quantità difficili da valutare o da stimare.

Nel metodo di Richardson le *maggiorazioni* dell'errore di troncamento sono utilizzate *indirettamente* per ricavare informazioni sul comportamento asintotico di  $T(y)$ . Tale comportamento permette di estrapolare, dai *valori già calcolati* di  $T(y)$ , un valore più *esatto*. L'*aspetto importante* da sottolineare è che tale valore è ottenuto senza fare intervenire il calcolo di  $T(y)$ , che rappresenta, in generale, la parte più costosa dal punto di vista computazionale.

Per dettagliare maggiormente l'idea, supponiamo che  $T(y)$  abbia uno sviluppo, rispetto ad  $y$ , del seguente tipo

$$T(y) = \tau_0 + \tau_1 y + \tau_2 y^2 + \cdots + \tau_k y^k + R_{k+1}(y) \quad (6.23)$$

con

$$|R_{k+1}(y)| \leq c_{k+1} y^{k+1}$$

e tale che i coefficienti  $\tau_i$  e  $c_{k+1}$  siano indipendenti da  $y$ .

► **Esempio 6.8** Per funzioni  $f \in C^{m+1}([x-a, x+a])$  e  $|h| \leq |a|$ , si ha dallo sviluppo di Taylor

$$T(h) := \frac{f(x+h) - f(x)}{h} = \tau_0 + \tau_1 h + \cdots + \tau_m h^m + h^{m+1}(\tau_{m+1} + o(1))$$

con  $\tau_k = f^{(k+1)}(x)/(k+1)!$ ,  $k = 0, 1, 2, \dots, m+1$ .

Per funzioni  $f \in C^{2m+3}([x-a, x+a])$  si ha

$$T(h) := \frac{f(x+h) - f(x-h)}{2h} = \tau_0 + \tau_1 h^2 + \cdots + \tau_m h^{2m} + h^{2m+2}(\tau_{m+1} + o(1))$$

con  $\tau_k = f^{(2k+1)}(x)/(2k+1)!$ ,  $k = 0, 1, 2, \dots, m+1$ .

Nel secondo caso, che rientra nel quadro generale assumendo  $y = h^2$ , lo sviluppo asintotico contiene soltanto le potenze pari di  $h$ ; questo fatto, come vedremo, rappresenta un vantaggio per il metodo. ■

► **Esempio 6.9** Se indichiamo con  $T(h)$ , con  $h = (b-a)/n$ , la formula dei trapezi

$$\int_a^b f(x) dx \approx T(h) := h \sum_{i=1}^{n-1} f(x_i) + \frac{h}{2}(f(x_0) + f(x_n)), \quad x_i = a + ih$$

nel caso di una funzione  $f \in C^{2m+2}([a, b])$  si ha, grazie ad una formula nota come *formula di Eulero-Maclaurin*

$$T(h) = \tau_0 + \tau_1 h^2 + \dots + \tau_m h^{2m} + \alpha_{m+1} h^{2m+2}$$

ove

$$\begin{aligned} \tau_0 &= \int_a^b f(x) dx \\ \tau_k &= \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(b) - f^{(2k-1)}(a)), \quad k = 1, 2, \dots, m \\ \alpha_{m+1} &= \frac{B_{2m+2}}{(2m+2)!} (b-a) f^{(2m+2)}(\xi(h)), \quad a < \xi(h) < b \end{aligned}$$

Dalla ipotesi di regolarità sulla  $f$  si ricava la seguente limitazione

$$|\alpha_{m+1}(h)| \leq M_{n+1}$$

con  $M_{n+1}$  costante indipendente da  $h$ . ■

Tenendo presente che il resto  $R_{k+1}(y)$  nello sviluppo asintotico tende a zero per  $y \rightarrow 0$ , l'idea è di sostituire la quantità  $T(y)$  con il polinomio

$$\tilde{T}(y) = \tau_0 + \tau_1 y + \tau_2 y^2 + \dots + \tau_k y^k$$

prendendo quindi il valore  $\tilde{T}(0)$  come nuova stima della quantità  $\tau_0$ .

Il polinomio  $\tilde{T}(y)$  può essere costruito con un *procedimento di interpolazione*, utilizzando i valori di  $T(y)$  calcolati in corrispondenza a particolari valori di  $y$ . In maniera generale si può procedere nel seguente modo.

Si scelgono due numeri reali  $r$  e  $y_0$  tali che:  $0 < r < 1, y_0 > 0$  e si costruiscono per  $m = 0, 1, \dots$  le successioni definite da

$$\begin{aligned} T_{m,0} &= T(r^m y_0) \\ T_{m,n+1} &= \frac{T_{m,n} - r^{n+1} T_{m-1,n}}{1 - r^{n+1}}, \quad n \geq 0 \end{aligned} \tag{6.24}$$

che corrispondono al calcolo della quantità  $T(y)$  nei punti  $y = r^m y_0$  e per  $n \geq 0$  alle operazioni di interpolazione e estrapolazione descritte in precedenza (si ricordi l'algoritmo di Neville). In forma tabellare si ha

$$\begin{array}{cccc}
 T_{0,0} & & & \\
 \downarrow & & & \\
 T_{1,0} & \rightarrow & T_{1,1} & \\
 \downarrow & & \downarrow & \\
 T_{2,0} & \rightarrow & T_{2,1} & \rightarrow & T_{2,2} \\
 \downarrow & & \downarrow & & \downarrow \\
 T_{3,0} & \rightarrow & T_{3,1} & \rightarrow & T_{3,2} & \rightarrow & T_{3,3} \\
 \vdots & & \vdots & & \vdots & & \vdots
 \end{array}$$

Si può dimostrare il seguente risultato.

**Proposizione 6.4** Per ogni  $n \geq 0$  si ha

$$T_{m,n} = \tau_0 + O((r^m y_0)^{n+1})$$

Si ha, quindi, la *convergenza* verso  $\tau_0$  di ciascuna delle colonne della tavola; per la prima colonna la convergenza è  $O(r^m y_0)$ , per la seconda  $O(r^{2m} y_0^2)$ , per la  $n$ -ma  $O((r^m y_0)^n)$ . La convergenza della colonna  $n$ -ma è asintoticamente  $n$  volte più rapida della prima colonna (che è la più costosa da calcolare!).

### Metodo di Romberg

Il metodo di Romberg<sup>9</sup> consiste nell'applicare il metodo di Richardson alla formula dei trapezi descritta nell'Esempio 6.9, assumendo  $y = h^2$ ,  $y_0 = (b-a)^2$ ,  $r = 1/4$ ; si ha allora la tabella triangolare  $T_{m,n}$ , definita da

$$\begin{aligned}
 T_{m,0} &= T\left(\frac{b-a}{2^m}\right), \quad m \geq 0 \\
 T_{m,n+1} &= \frac{4^{n+1} T_{m,n} - T_{m-1,n}}{4^{n+1} - 1}, \quad n = 0, 1, \dots, m-1
 \end{aligned}$$

Per il risultato precedente, si ha

$$T_{m,n} = \int_a^b f(x) dx + O(h^{2n+2}), \quad h = \frac{b-a}{2^m}$$

<sup>9</sup>Con riferimento al metodo di Archimede per il calcolo dell'area di un cerchio (cfr. nota (1)) si può vedere facilmente che, se  $F_n$  è l'area del poligono regolare a  $n$  lati circoscritto al cerchio, la successione  $F_6, F_{12}, F_{24}, \dots$  converge monotonamente a  $\pi$ . C. Huygens (1629-1695) osservò che la velocità di convergenza di tale successione è di ordine  $(1/n)^2$ , cioè quadratica. Mediante opportune combinazioni lineari di coppie di successivi termini della successione, egli fu in grado di costruire una successione convergente con ordine  $(1/n)^4$ . Il metodo di Romberg è, in sostanza, uno sviluppo sistematico di tale idea.

Se la funzione  $f$  è integrabile secondo Riemann su  $(a, b)$ , si ha

$$\lim_{m \rightarrow \infty} T_{m,n} = \int_a^b f(x) dx, \quad \forall n \geq 0$$

Si può verificare facilmente che  $T_{m,1}$  è ancora il metodo di Simpson; per  $n > 2$ , tuttavia, la quantità  $T_{m,n}$  non corrisponde più a un metodo di Newton-Cotes.

Si noti che i valori della funzione  $f$  che intervengono nel calcolo di  $T_{m,0}$  intervengono pure nel calcolo di  $T_{m+1,0}$ , per cui si può organizzare l'algoritmo nella forma seguente

$$T_{0,0} = \frac{1}{2}(f(a) + f(b))$$

$$T'_{m,0} = h \sum_{i=1}^{2^m} f\left(a + \left(i - \frac{1}{2}\right)h\right), \quad h = (b-a)/2^m$$

$$T_{m+1,0} = \frac{1}{2}(T_{m,0} + T'_{m,0})$$

Nella Tabella 6.4 sono riportati i risultati corrispondenti alla funzione  $e^x \cos x$  su  $(0, \pi)$ , e rispettivamente nella Tabella 6.5 quelli corrispondenti alla funzione  $\sqrt{x}$  su  $(0, 1)$ , ottenuti utilizzando il seguente programma.

```

SUBROUTINE ROMBERG(F,A,B,M,R,IR)
DIMENSION R(IR,M)
*****
* Metodo di Romberg
* F funzione integranda
* H=(B-A)/2**(M-1)
*****
H = B - A
R(1,1) = 0.5*H*(F(A) + F(B))
L = 1
DO 4 I = 2,M
H = 0.5*H
L = L + L
SUM = 0.0
DO 2 K = 1,L-1,2
SUM = SUM + F(A + H*REAL(K))
2 CONTINUE
N = 1
R(I,1) = 0.5*R(I-1,1) + H*SUM
DO 3 J = 2,I
N = 4*N
R(I,J) = R(I,J-1) + (R(I,J-1) - R(I-1,J-1))
&/REAL(N - 1)
3 CONTINUE
4 CONTINUE
RETURN
END

```

m/n	0	1	2	3	4
0	22.708				
1	5.318	-0.477E 00			
2	1.265	-8.540E-02	-5.926E-02		
3	0.311	-6.134E-03	-8.497E-04	7.724E-05	
4	7.766E-02	-3.919E-04	-9.536E-06	3.814E-06	3.814E-06

Tabella 6.4: Errori ottenuti con il metodo di Romberg per l'integrale  $\int_0^\pi e^x \cos x dx$ .

m/n	0	1	2	3	4
0	0.166E 00				
1	6.311E-02	2.859E-02			
2	2.338E-02	1.014E-02	8.910E-03		
3	8.536E-03	3.587E-03	3.150E-03	3.059E-03	
4	3.085E-03	1.268E-03	1.113E-03	1.081E-03	1.073E-03

Tabella 6.5: Errori ottenuti con il metodo di Romberg per l'integrale di  $\sqrt{x}$  su  $(0, 1)$ .

Nella Tabella 6.6 sono riportati i risultati ottenuti applicando l'estrapolazione al calcolo della derivata di  $\sin x$  nel punto  $\pi/3$  con il metodo delle differenze centrali.

h/n	0	1	2	3
1.000	0.42073548			
0.500	0.47942552	0.49898887		
0.250	0.49480790	0.49993536	0.49999845	
0.125	0.49869895	0.49999598	0.500000000	0.50000000

Tabella 6.6: Errori ottenuti con il metodo di estrapolazione per il calcolo della derivata di  $\sin x$  in  $\pi/3$ .

## 6.5 Difficoltà nell'integrazione numerica

Le formule di quadratura studiate nei paragrafi precedenti possono dare risultati insoddisfacenti quando la funzione integranda, o alcune delle sue derivate, presentano delle singolarità. Ricordiamo, infatti, che la maggiorazione dell'errore di una particolare formula, e quindi il corrispondente ordine di convergenza, sono ottenuti sotto opportune condizioni di regolarità della funzione.

Per gli integrali singolari sono necessarie, quindi, opportune strategie. Nel seguito, illustreremo alcune idee su un esempio particolare.

Consideriamo il seguente integrale

$$I = \int_0^1 \frac{\arctan x}{x^{3/2}} dx$$

che presenta una singolarità nel punto  $x = 0$ , in quanto per sviluppo in serie si ha

$$\frac{\arctan x}{x^{3/2}} = \frac{x - x^3/3 + x^5/5 - \dots}{x^{3/2}} = x^{-1/2} + O(x^{3/2})$$

Una *integrazione per parti* fornisce il seguente risultato

$$I = \int_0^1 \frac{\arctan x}{x^{3/2}} dx = [-2x^{-1/2} \arctan x]_0^1 + \int_0^1 \frac{2}{x^{1/2}(1+x^2)} dx$$

Con il cambiamento di variabili  $z = x^{1/2}$  si ottiene

$$I = -\frac{\pi}{2} + \int_0^1 \frac{4}{1+t^4} dt$$

L'ultimo integrale non presenta singolarità, e può essere calcolato numericamente con una formula di quadratura.

Un modo alternativo di procedere consiste in una *sottrazione della singolarità* mediante una opportuna funzione. Nell'esempio si può procedere nel seguente modo

$$I = \int_0^1 \frac{\arctan x - x}{x^{3/2}} dx + \int_0^1 \frac{1}{\sqrt{x}} dx = \int_0^1 \frac{\arctan x - x}{x^{3/2}} dx + 2$$

La nuova funzione integranda è, vicino all'origine, della forma  $-\frac{1}{3}x^{3/2} + O(x^{7/2})$ . In questo modo si è eliminata la singolarità nella funzione, ma è rimasta singolare la derivata seconda. Per avere maggiore regolarità è quindi necessario sottrarre un polinomio di grado maggiore; ricordando lo sviluppo in serie di  $\arctan x$ , si può, ad esempio, sottrarre  $x - x^3/3$ .

Segnaliamo, infine, la procedura basata su uno sviluppo in serie della funzione integranda. Nel caso dell'esempio

$$I = \int_0^1 \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} x^{-3/2} dx = \sum_{n=0}^{\infty} (-1)^n \frac{2}{(2n+1)(4n+1)}$$

Un altro tipo di singolarità si presenta quando l'intervallo di integrazione è infinito. In questo caso può essere opportuna una trasformazione di variabili per ridurre l'intervallo di integrazione a un intervallo finito. Ad esempio,  $(0, \infty)$  è trasformato in  $(1, 0)$  mediante la trasformazione  $t = 1/(1+x)$  o  $t = e^{-x}$ .

Un altro modo di procedere è illustrato dal seguente esempio. L'integrale improprio  $\int_0^{\infty} e^{-x^2} dx$  può essere approssimato dall'integrale proprio  $\int_0^b e^{-x^2} dx$  con un errore dato da

$$E = \int_b^{\infty} e^{-x^2} dx < \frac{1}{b} \int_b^{\infty} x e^{-x^2} dx = \frac{1}{2b} e^{-b^2}$$

Tale errore tende a zero per  $b$  che tende all'infinito, e quindi il calcolo dell'integrale su  $(0, \infty)$  è ridotto alla ricerca di un opportuno valore di  $b$  e al calcolo, mediante una opportuna formula di quadratura, dell'integrale su  $(0, b)$ .

Ricordiamo, infine, la possibilità di utilizzare *formule di Gauss* basate sui polinomi ortogonali di Laguerre e di Hermite.

## 6.6 Integrali multipli

Le idee che abbiamo esaminato nei paragrafi precedenti per il calcolo numerico di integrali in una variabile possono essere opportunamente generalizzate al calcolo di *integrali multipli*, ma la quantità di operazioni richieste aumenta rapidamente con il numero delle dimensioni. È importante, quindi, ridurre, quando possibile, il numero delle dimensioni mediante tecniche analitiche.

Si ha, ad esempio

$$\begin{aligned} & \int_0^\infty \int_0^\infty \int_0^\infty e^{-(x_1+x_2+x_3)} \sin(x_1x_3) \sin(x_2x_3) dx_1 dx_2 dx_3 \\ &= \int_0^\infty e^{-x_1} dx_1 \int_0^\infty e^{-x_2} \sin(x_2x_1) dx_2 \int_0^\infty e^{-x_3} \sin(x_3x_1) dx_3 \\ &= \int_0^\infty \left(\frac{x_1}{1+x_1^2}\right)^2 e^{-x_1} dx_1 \end{aligned}$$

e l'integrale semplice che rimane può essere calcolato mediante le tecniche viste in precedenza.

Talvolta la riduzione di dimensione può essere ottenuta mediante un opportuno cambiamento di variabili. Ad esempio, se  $D$  è il cerchio unitario, il seguente integrale

$$\iint_D \frac{x_2 \sin(kx_2)}{x_1^2 + x_2^2} dx_1 dx_2$$

può essere ridotto ad un integrale semplice mediante il passaggio a coordinate polari.

I procedimenti ora indicati non sono, tuttavia, di utilità generale, o almeno di applicabilità semplice. Sono necessarie, quindi, tecniche più generali, tra le quali segnaliamo le seguenti

1. Quando il dominio di integrazione è semplice rispetto agli assi coordinati, ad esempio è un parallelepipedo, l'integrale può essere ridotto ad integrazioni successive in una dimensione, alle quali si applicano le usuali formule.
2. Il dominio è suddiviso in figure elementari, tipo parallelepipedo, prismi, ecc., e l'integrale è calcolato come somma di integrali sulle singole figure elementari. Si tratta dell'estensione delle formule composte.
3. Quando il numero delle dimensioni è elevato, può essere più conveniente una tecnica di simulazione statistica (cfr. nel successivo Capitolo 10 il *metodo Monte Carlo*).

► **Esempio 6.10** Considerando in particolare l'approccio (2), segnaliamo, come esemplificazione, la seguente formula di integrazione su un triangolo generico  $T$

$$\iint_T f(x, y) dx dy \approx \frac{1}{3} \text{mis}(T) \left[ f\left(\frac{P_1 + P_2}{2}\right) + f\left(\frac{P_2 + P_3}{2}\right) + f\left(\frac{P_3 + P_1}{2}\right) \right]$$



ove  $P_i$  sono i vertici del triangolo  $T$  e  $\text{mis}(T)$  indica l'area del triangolo. Tale formula risulta esatta per i polinomi di grado 2 ed è nota come *formula dei punti medi*. Il risultato può essere ottenuto nel seguente modo. Ogni punto  $P \equiv (x, y)$  del piano può essere espresso in maniera univoca mediante la relazione

$$P = \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3, \quad \lambda_1 + \lambda_2 + \lambda_3 = 1$$

Le variabili  $\lambda_i$  sono le *coordinate baricentriche* di  $P$  e sono determinate come soluzioni del seguente sistema lineare, non singolare quando il triangolo  $T$  è non degenere

$$\begin{cases} \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 = x \\ \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3 = y \\ \lambda_1 + \lambda_2 + \lambda_3 = 1 \end{cases}$$

ove  $P_i = (x_i, y_i)$ ,  $i = 1, 2, 3$  e  $P = (x, y)$ . I punti interni del triangolo  $T$  sono caratterizzati dalle condizioni

$$\lambda_i > 0, \quad i = 1, 2, 3$$

Il baricentro corrisponde a  $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$ . Il lato  $P_2 P_3$  ha come equazione  $\lambda_1 = 0$  e analoghe equazioni si hanno per gli altri lati.

Se la funzione  $f(x, y)$  è una funzione affine  $f(P) = a_1 x + a_2 y + b$ , allora si ha

$$f(P) = \lambda_1 f(P_1) + \lambda_2 f(P_2) + \lambda_3 f(P_3)$$

che rappresenta la formula di interpolazione lineare su griglie triangolari. Per la interpolazione *quadratica* si utilizzano oltre i nodi  $P_i$  i punti di mezzo dei lati  $\frac{1}{2}(P_i + P_j)$ ,  $i \neq j$ . Posto, per brevità

$$\Delta''_{ij} = f(P_i) + u(P_j) - 2f\left(\frac{1}{2}(P_i + P_j)\right), \quad i \neq j$$

si vede facilmente che la formula di interpolazione

$$f(P) \approx \lambda_1 f(P_1) + \lambda_2 f(P_2) + \lambda_3 f(P_3) - 2(\lambda_2 \lambda_3 \Delta''_{23} + \lambda_3 \lambda_1 \Delta''_{31} + \lambda_1 \lambda_2 \Delta''_{12})$$

è esatta per i polinomi di secondo grado.

Per dimostrare allora che la formula di quadratura assegnata è esatta per i polinomi di secondo grado si può procedere in questo modo.

Osservato che per ragioni di simmetria  $\iint_T \lambda_i dx dy$  assume lo stesso valore per  $i = 1, 2, 3$  e lo stesso si ha per le varie combinazioni di  $\iint_T \lambda_i \lambda_j dx dy$ , per un polinomio di secondo grado si ha

$$\begin{aligned} \iint_T f(x, y) dx dy &= a(f(P_1) + f(P_2) + f(P_3)) - 2b(\Delta''_{23} + \Delta''_{31} + \Delta''_{12}) \\ &= (a - 4b)(f(P_1) + f(P_2) + f(P_3)) + \\ &\quad + 4b\left[f\left(\frac{P_2 + P_3}{2}\right) + f\left(\frac{P_3 + P_1}{2}\right) + f\left(\frac{P_1 + P_2}{2}\right)\right] \end{aligned}$$

ove

$$a = \iint_T \lambda_1 dx dy, \quad b = \iint_T \lambda_1 \lambda_2 dx dy$$

Per il calcolo di questi ultimi integrali operiamo un cambiamento di variabili, assumendo come variabili di integrazione  $\lambda_1, \lambda_2$ . Tenendo conto della relazione  $\lambda_3 = 1 - \lambda_1 - \lambda_2$ , si ha

$$\begin{aligned}x &= \lambda_1(x_1 - x_3) + \lambda_2(x_2 - x_3) + x_3 \\y &= \lambda_1(y_1 - y_3) + \lambda_2(y_2 - y_3) + y_3\end{aligned}$$

Il determinante della jacobiana della trasformazione è dato da  $2 \operatorname{mis}(T)$ , e quindi

$$\begin{aligned}a &= \int_0^1 \int_0^{1-\lambda_1} \lambda_1 \cdot 2 \operatorname{mis}(T) d\lambda_1 d\lambda_2 = 2 \operatorname{mis}(T) \int_0^1 \lambda_1 (1 - \lambda_1) d\lambda_1 = \frac{\operatorname{mis}(T)}{3} \\b &= \int_0^1 \int_0^{1-\lambda_1} \lambda_1 \lambda_2 \cdot 2 \operatorname{mis}(T) d\lambda_1 d\lambda_2 = 2 \operatorname{mis}(T) \int_0^1 \lambda_1 \frac{(1 - \lambda_1)^2}{2} d\lambda_1 = \frac{\operatorname{mis}(T)}{12}\end{aligned}$$

da cui il risultato richiesto. ■

◆ **Esercizio 6.1** Confrontare vari metodi numerici per l'approssimazione della lunghezza dell'ellisse  $x^2 + y^2/4 = 1$ .

◆ **Esercizio 6.2** Un corpo nero (ossia un oggetto capace di emettere e di assorbire tutte le frequenze di radiazione uniformemente) emette energia a una velocità proporzionale alla quarta potenza della sua temperatura assoluta, in accordo alla seguente equazione di Stefan-Boltzmann

$$E = 36.9 \cdot 10^{-12} T^4$$

ove  $E$  è la potenza emissiva ( $\text{watt}/\text{cm}^2$ ) e  $T$  la temperatura  $^\circ\text{K}$ . Supponiamo di essere interessati a calcolare la frazione di questa energia totale contenuta nello spettro visibile, assunto corrispondente all'intervallo  $[4 \cdot 10^{-5}, 7 \cdot 10^{-5}]$  cm. Tale energia si ottiene integrando l'equazione di Plank, e quindi

$$E_{\text{visibile}} = \int_{4 \cdot 10^{-5}}^{7 \cdot 10^{-5}} \frac{2.39 \cdot 10^{-11}}{x^5 (e^{1.432/T x} - 1)} dx$$

ove  $x$  indica la lunghezza d'onda in cm. L'efficienza luminosa è definita come il rapporto dell'energia nello spettro visibile rispetto all'energia totale. Se si moltiplica per 100 per dare l'efficienza luminosa in percentuale e si associano le costanti, si ha

$$\text{EFF} = \frac{64.77 \int_{4 \cdot 10^{-5}}^{7 \cdot 10^{-5}} \frac{1}{x^5 (e^{1.432/T x} - 1)} dx}{T^4}$$

Fornire una procedura numerica per il calcolo di EFF in corrispondenza a dei valori fissati di  $T$ .

◆ **Esercizio 6.3** La funzione di Debye presenta interesse nella termodinamica statistica per il calcolo del calore specifico a volume costante di particolari sostanze. La funzione ha la seguente espressione

$$D(x) := 3x^{-3} \int_0^x \frac{y^3}{e^y - 1} dy$$

Analizzare e implementare per il calcolo di  $D(x)$  le formule di Simpson e di Gauss per  $x = 0.5, 10.050.0$  e  $100.0$ . Tenere presente, come test, che il valore per  $x = 0.5$  è  $\approx 0.4899$ .

◆ **Esercizio 6.4** *Approssimare il valore del seguente integrale*

$$I = \int_0^1 \ln x \cos x \, dx$$

◆ **Esercizio 6.5** *Mostrare che se sono disponibili le derivate  $f'(x_0)$  e  $f'(x_1)$ , allora la seguente formula, detta formula del trapezio migliorata, o anche formula di Chevallier*

$$\int_{x_0}^{x_1} f(x) \, dx \approx \frac{h}{2}[f(x_0) + f(x_1)] + \frac{h^2}{12}[f'(x_0) - f'(x_1)]$$

con  $h = x_1 - x_0$ , ha grado di precisione tre, cioè è esatta quando  $f(x)$  è un polinomio di grado minore o uguale a tre.

◆ **Esercizio 6.6** *Mostrare che se nella formula del trapezio i nodi  $x_1$  e  $x_2$  non coincidono necessariamente con gli estremi dell'intervallo  $[a, b]$ , si ottiene la seguente formula di quadratura*

$$\int_a^b f(x) \, dx \approx \frac{b-a}{x_2-x_1} \left[ x_2 f(x_1) - x_1 f(x_2) + \frac{b+a}{2} [f(x_2) - f(x_1)] \right]$$

Studiare il corrispondente errore.

◆ **Esercizio 6.7** *Ottenere la formula di quadratura di interpolazione mediante un polinomio di grado 2 con nodi  $x_1, x_2$  e  $x_3$  non necessariamente equidistanti o coincidenti con gli estremi dell'intervallo di integrazione  $a$  e  $b$ . Studiare l'errore di troncamento associato.*

◆ **Esercizio 6.8** *Mostrare che una formula di quadratura di Gauss*

$$\int_{-1}^1 f(x) \, dx \approx \sum_{i=0}^n w_i f(x_i)$$

per la quale  $x_0 = -1$  e i rimanenti nodi sono le  $n$  radici della funzione

$$\phi_n(x) = \frac{P_n(x) + P_{n+1}(x)}{1+x}$$

è esatta quando  $f(x)$  è un polinomio di grado inferiore o uguale a  $2n$ . Trovare i pesi  $w_i, i = 0, 1, \dots, n$ , per  $n = 1, 2, 3$ .

◆ **Esercizio 6.9** *Calcolare l'integrale*

$$\int_0^1 \frac{\sin x}{x} \, dx$$

mediante una formula di Gauss a tre termini.

◆ **Esercizio 6.10** *Supponendo di conoscere i valori  $f(x_1), f(x_2), \dots, f(x_n)$  nei punti  $x_1, x_2, \dots, x_n$  (ordinati in modo crescente, ma non necessariamente equidistanti), approssimare l'integrale  $\int_a^b f(x) \, dx$  mediante successive applicazioni della formula del trapezio. Implementare il metodo sotto forma di subroutine.*

◆ **Esercizio 6.11** Applicare il metodo di quadratura adattivo a  $\int_{0.1}^2 \sin(1/x) dx$ .

◆ **Esercizio 6.12** Usare il metodo di Romberg per approssimare l'integrale  $\int_1^{10} \log x dx$ .

... prior to undertaking any work it is advisable to study the integrals from the following points of view:

1. Confirm the existence of the integral.
2. Ascertain the important ranges of the parameters involved.
3. Reduce the integral to its simplest form.
4. Determine the essential parameters which are involved.
5. Determine the accuracy to which numerical values are to be given.

**M. Abramowitz**

Data æquatione quotcunque fluentes quantitates involvente fluxiones invenire et vice versa.

6a cc d æ 13e ff 7i 3l 9n 4o 4qrr 4s 9t 12vx

Newton a Leibniz, 24 ottobre, 1676.

## Capitolo 7

# Equazioni differenziali

Le *equazioni differenziali* rappresentano uno degli strumenti più efficaci nella costruzione di modelli matematici per la simulazione di *sistemi dinamici*, cioè di sistemi che evolvono nel tempo secondo determinate *leggi*. Esse sono utilizzate nei più svariati campi applicativi, dall'ingegneria alla chimica, alla fisica, alla biologia, alle scienze sociali, all'informatica. È superfluo, quindi, sottolineare l'importanza di un loro studio sia sotto l'aspetto teorico che numerico. Data, comunque, la vastità dell'argomento, ciò che segue non può che rappresentare una *introduzione elementare*. Inoltre, la maggior parte dei risultati saranno forniti in *maniera intuitiva*, rinviando per le necessarie giustificazioni e gli opportuni approfondimenti alla bibliografia. Incominceremo con alcune semplici definizioni e notazioni. Nel seguito esamineremo alcuni modelli basati su tipi particolari di equazioni differenziali.

### 7.1 Aspetti introduttivi

Le equazioni differenziali sono equazioni che esprimono un legame tra alcune funzioni incognite e le loro derivate. Quando le derivate si riferiscono ad una sola variabile si parla di *equazioni differenziali ordinarie* (ODE); nel caso di più variabili si hanno le *equazioni alle derivate parziali* (PDE). Ad esempio

$$y'(t) = k y(t) \tag{7.1}$$

ove  $t$  è la *variabile indipendente* e  $y$  la *funzione incognita*, è una equazione differenziale ordinaria, mentre

$$\frac{\partial u}{\partial t} - a \frac{\partial^2 u}{\partial x^2} = 0 \tag{7.2}$$

ove  $t, x$  sono le *variabili indipendenti* e  $u$  la *funzione incognita*, è un caso particolare di equazione alle derivate parziali. Vedremo successivamente l'interesse per le applicazioni delle equazioni precedenti.

Per il seguito saranno utilizzate le seguenti notazioni equivalenti

$$y' \text{ (oppure } y'(t)) \text{ per } \frac{dy(t)}{dt}; \quad y'' \text{ (oppure } y''(t)) \text{ per } \frac{d^2y(t)}{dt^2}$$

$$u_t \text{ (oppure } u_t(x, t)) \text{ per } \frac{\partial u(x, t)}{\partial t}; \quad u_{xt} \text{ (oppure } u_{xt}(x, t)) \text{ per } \frac{\partial}{\partial t} \left[ \frac{\partial u(x, t)}{\partial x} \right]$$

Naturalmente, le notazioni utilizzate per indicare la funzione incognita e le variabili indipendenti possono variare nelle diverse applicazioni per sottolineare opportunamente il loro significato. In questo senso, usualmente  $t$  indica la variabile *tempo*, mentre  $x$  indica una variabile *spaziale*.

Oltre che in ordinarie e a derivate parziali, le equazioni differenziali sono *classificate* in differenti altri modi, a seconda della opportunità. L'*ordine* di una equazione è l'ordine massimo della derivata della funzione incognita che appare nell'equazione. Ad esempio, l'equazione (7.1) è del primo ordine, come pure la seguente

$$y'(t) = k y(t)(1 - by(t))$$

anche se la funzione incognita  $y$  compare con potenza 2. Al contrario, la seguente equazione, ove  $m, k$  sono costanti assegnate, con  $m \neq 0$

$$m y''(t) + k e^{-at} y(t) = 0 \tag{7.3}$$

è un'equazione differenziale del secondo ordine. Più in generale, un'equazione differenziale di ordine  $n$  è un'equazione della seguente forma

$$F(t, y(t), y'(t), y''(t), \dots, y^{(n)}(t)) = 0$$

ove  $F$  è una funzione assegnata di  $n + 2$  variabili. Quando tale equazione può essere esplicitata rispetto alla variabile  $y^{(n)}(t)$ , ossia si può scrivere

$$y^{(n)}(t) = f(t, y(t), y'(t), \dots, y^{(n-1)}(t))$$

si dice che l'equazione è in *forma normale*. Ad esempio, l'equazione (7.3) è equivalente alla seguente equazione in forma normale

$$y'' = -\frac{k}{m} e^{-at} y$$

mentre la seguente è un esempio di equazione non equivalente ad una singola equazione in forma normale

$$y(t) = t^2 y(t) - [y'(t)]^2$$

Osserviamo che il fatto che un'equazione differenziale possa essere messa in forma normale è importante in quanto in tale forma essa può essere più facilmente studiata sia teoricamente che numericamente. Ricordiamo infine che mediante opportune sostituzioni è possibile trasformare una data equazione differenziale di ordine  $n$  in un sistema di  $n$  equazioni differenziali del primo ordine. Ad esempio, per l'equazione (7.3), posto  $y_1 = y$ ,  $y_2 = y'$ , si ha

$$my''(t) + ke^{-at}y(t) = 0 \Leftrightarrow \begin{cases} y_1' = y_2 \\ y_2' = -\frac{k}{m}e^{-at}y_1 \end{cases}$$

Il risultato ora enunciato giustifica il fatto che per il seguito analizzeremo più in dettaglio il caso particolare dell'equazione differenziale del primo ordine

$$y' = f(t, y) \quad (7.4)$$

In effetti, i risultati ottenuti per tale equazione si estendono al caso di equazioni di ordine superiore e ai sistemi considerando  $y$  e  $f$  come funzioni vettoriali.

### 7.1.1 Definizione di soluzione

Una funzione  $y(t)$  definita su un intervallo  $(a, b)$  è una *soluzione* dell'equazione differenziale (7.4) se  $y(t)$  è continua insieme alla derivata  $y'(t)$  e l'equazione (7.4) è verificata, quando si sostituisce la funzione incognita con la funzione  $y$ , per ogni  $t \in (a, b)$ . Ad esempio, la funzione  $y = e^{3t}$  è una soluzione dell'equazione differenziale

$$y'(t) = 3y(t) \quad (7.5)$$

per ogni intervallo  $(a, b)$ , con  $-\infty \leq a < b \leq \infty$ . In effetti, ogni funzione della forma  $y = ce^{3t}$ , con  $c$  costante arbitraria, è una soluzione dell'equazione data. Tale risultato può essere verificato direttamente, oppure osservando che l'equazione differenziale (7.5) è equivalente alla seguente equazione

$$\frac{d}{dt}[e^{-3t}y(t)] = 0, \quad a < t < b$$

da cui  $e^{-3t}y(t) = c$ , con  $c$  costante generica.

Dalla definizione appare chiaro che come soluzione di una equazione differenziale si deve intendere, più propriamente, la coppia *funzione e intervallo*. Osservando che una funzione che è soluzione su un intervallo  $(a, b)$  è pure soluzione su ogni sottointervallo di  $(a, b)$ , si dice che una soluzione  $y(t)$ ,  $(a, b)$  è *massimale*, quando non esiste una soluzione  $z(t)$ ,  $(\alpha, \beta)$  tale che l'intervallo  $(\alpha, \beta)$  contiene l'intervallo  $(a, b)$ , con  $y(t) \equiv z(t)$  su  $(a, b)$ . Ad esempio,  $e^{3t}$ ,  $(-\infty, \infty)$  è una soluzione massimale per l'equazione differenziale  $y'(t) = 3y(t)$ .

Come esempio più significativo, consideriamo la seguente equazione differenziale

$$y'(t) = y^2(t) \quad (7.6)$$

Per tale equazione la coppia  $y(t) \equiv 0, (-\infty, \infty)$  è una soluzione massimale, ma ha interesse vedere se esistono anche soluzioni non identicamente nulle. Se  $y(t)$  è per ipotesi una di tali soluzioni, essa deve essere diversa dallo zero in almeno un punto  $t_0$ , ed essendo continua esisterà un intervallo  $I$  contenente  $t_0$  nel quale la  $y(t)$  non si annulla. Su  $I$  possiamo allora dividere ambo i membri di (7.6) per  $y^2(t)$  ed ottenere

$$\frac{y'(t)}{y^2(t)} = 1 \quad \forall t \in I$$

da cui, per la nota regola di derivazione

$$-\frac{d}{dt}\left(\frac{1}{y(t)}\right) = 1 \Leftrightarrow \frac{1}{y(t)} = -t + c \quad \forall t \in I$$

con  $c$  costante arbitraria. In definitiva, si ha

$$y = \frac{1}{c - t} \quad \forall t \in I$$

In questo caso si ha, quindi, che le coppie  $y = 1/(c - t), (-\infty, c)$  e  $y = 1/(c - t), (c, +\infty)$ , sono *due soluzioni massimali* distinte, separate dal punto  $t = c$ , nel quale le soluzioni sono illimitate.

◆ **Esercizio 7.1** Trovare l'ordine delle seguenti equazioni differenziali, cercando, quando possibile, di porle in forma normale

$$(a) (y')^2 - t^3 y = 2; \quad (b) (y' \cos t)' + y \tan t = 1$$

◆ **Esercizio 7.2** Trovare le soluzioni massimali delle seguenti equazioni differenziali

$$(a) y' = 10y^2, \quad (b) y' = 3|t|y$$

### 7.1.2 Curve soluzioni e campi di direzioni

Come *curva soluzione* dell'equazione differenziale (7.4) si intende il grafico di una soluzione. La rappresentazione di una soluzione sotto forma di grafico è utile per avere direttamente informazioni sul comportamento della soluzione. Tale rappresentazione può essere, naturalmente, ottenuta in modo facile quando si ha una rappresentazione analitica semplice della soluzione. In caso contrario, sono indispensabili i metodi numerici, che analizzeremo nel seguito. Segnaliamo, comunque, dapprima un'altra possibilità, basata sulla seguente interpretazione della funzione  $f(t, y)$ . Supponendo



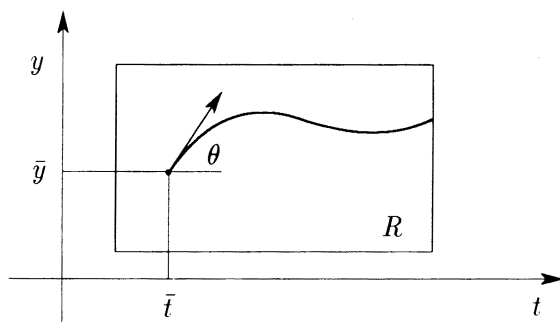


Figura 7.1: Soluzione curva dell'equazione differenziale  $y' = f(t, y)$ . Si ha  $f(\bar{t}, \bar{y}) = \tan \theta$ .

che tale funzione sia definita in una regione  $R$  del piano  $t$ - $y$ , ad esempio in un rettangolo a lati paralleli agli assi, la funzione  $f(t, y)$  fornisce in ogni punto  $(t, y) \in R$  la pendenza della tangente alla curva soluzione che passa per tale punto (cfr. Figura 7.1). Se, allora, il grafico  $G$  di una funzione derivabile con continuità  $y(t)$ , è interno ad  $R$ , si ha che  $G$  è una curva soluzione di (7.4) se in ogni punto  $(t, y)$  su  $G$  la pendenza della tangente ha il valore  $f(t, y)$ .

Tale modo di vedere l'equazione differenziale (7.4) può essere utilizzato per *costruire* soluzioni particolari nel seguente modo. In corrispondenza ai punti di una reticolazione prefissata di  $R$ , si disegnano dei *vettori* con pendenza  $f(t, y)$ . Il diagramma ottenuto, chiamato *campo di direzioni* (cfr. Figura 7.2 per un esempio) può dare indicazioni utili sul comportamento delle soluzioni che passano per punti fissati dell'insieme  $R$ .

In particolare, pensando alle curve soluzione come a linee di flusso di un fluido con velocità  $f(t, y)$  nel punto  $(t, y)$ , si può vedere che quando  $\partial f / \partial y > 0$  in un punto  $t = \bar{t}$  e per ogni  $y$ , le direzioni di flusso sono divergenti, come illustrato in Figura 7.3. Un comportamento opposto si verifica quando  $\partial f / \partial y < 0$ . Nel primo caso si parla di *soluzioni instabili* e nel secondo di *soluzioni stabili*. Tale denominazione è motivata dal fatto che la distanza, su rette parallele all'asse delle  $y$ , tra due soluzioni fissate aumenta, o rispettivamente diminuisce, all'aumentare di  $t$ . Di conseguenza, nel primo caso gli errori iniziali sono amplificati, mentre nel secondo sono ridotti.

### 7.1.3 Problemi ai valori iniziali

Data un'equazione differenziale del tipo (7.4), nei modelli reali è spesso interessante la ricerca di una *particolare* soluzione che assuma per un determinato  $t = t_0$  un

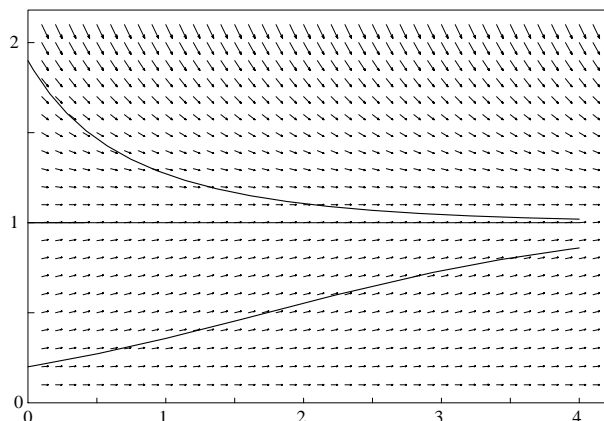


Figura 7.2: Campo di direzioni relativo all'equazione differenziale  $y' = 0.8y(1 - y)$ . Con linea continua sono rappresentate le due soluzioni particolari corrispondenti al valore iniziale  $y(0) = 0.2$  e rispettivamente  $y(0) = 1.8$ .

valore assegnato  $y_0$ , ossia la risoluzione del seguente problema

$$\begin{cases} y'(t) = f(t, y) \\ y(t_0) = y_0 \end{cases} \quad (7.7)$$

Tale problema è detto *problema ai valori iniziali* o *problema di Cauchy*; dal punto di vista geometrico, esso equivale alla ricerca della curva soluzione che passa per il punto *iniziale*  $(t_0, y_0)$ . I *dati* del problema, cioè l'*input* del modello descritto dal problema matematico (7.7), sono la funzione  $f(t, y)$  e la coppia di valori  $(t_0, y_0)$ , e il *risultato*, corrispondente all'*output* del modello, è la funzione  $y(t)$  definita su un intervallo  $I$ , con  $t_0 \in I$ , e che verifica la condizione iniziale  $y(t_0) = y_0$  e l'equazione differenziale  $y'(t) = f(t, y)$  per  $t$  in  $I$ .

Un modo per risolvere il problema (7.7) consiste nel cercare dapprima una espressione analitica di tutte le soluzioni massimali dell'equazione differenziale  $y' = f(t, y)$ , cioè l'espressione dell'*integrale generale*. Una volta costruito l'integrale generale, si cerca il valore della costante (o delle costanti, nel caso di un sistema di equazioni differenziali) che fornisce la soluzione passante per il punto  $(t_0, y_0)$ .

Ad esempio, per l'equazione differenziale  $y' = ky$ , l'integrale generale è dato dalla famiglia di funzioni  $y = c \exp(kt)$ , con  $c$  costante arbitraria. Da esso si ricava immediatamente  $y_0 = c \exp(kt_0)$ , da cui  $c = y_0 \exp(-kt_0)$ , e quindi  $y(t) = y_0 \exp(k(t-t_0))$  è la soluzione particolare cercata.

Le principali tecniche utili per costruire l'integrale generale di un'equazione differenziale sono introdotte e analizzate in Appendice B. Esse si applicano, tuttavia,

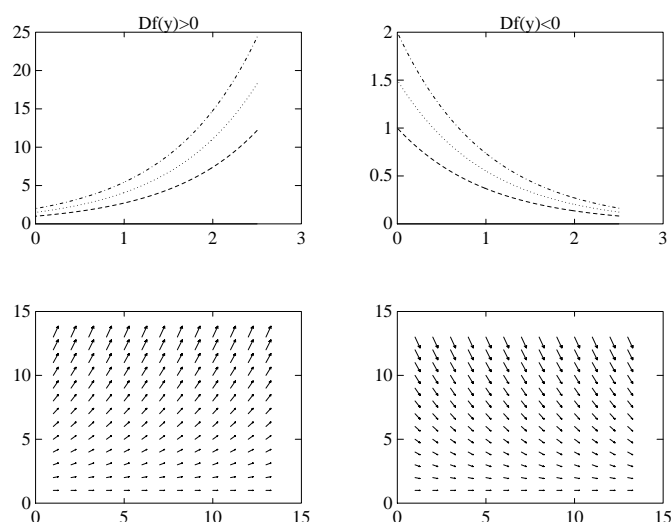


Figura 7.3: Soluzioni instabili e rispettivamente stabili di un problema a valori iniziali.

solo a tipi particolari di equazioni differenziali. Nel caso generale, è necessario quindi ricorrere a procedimenti numerici, ossia all' approssimazione del problema (7.7) mediante opportuni problemi in dimensione finita. Allo studio di tali metodi, che costituisce l'oggetto principale di questo capitolo, è comunque importante premettere la discussione delle seguenti questioni di base:

**esistenza** ossia stabilire quali condizioni sui dati garantiscono l'esistenza di almeno una soluzione del problema (7.7) nell'ambito di una particolare classe di funzioni;

**unicità** ossia stabilire quando il problema ammette al più una soluzione;

**sensitività** ossia, nel caso di esistenza ed unicità della soluzione, studiare come tale soluzione varia in corrispondenza a variazioni sui dati; in particolare, verificare se tale dipendenza è continua.

Una risposta a tali questioni può essere ottenuta imponendo particolari condizioni alla funzione  $f(t, y)$ . Precisamente, supponiamo che

1.  $f(t, y)$  sia una funzione *continua* nell'insieme  $S := \{(t, y) \mid a \leq t \leq b, y \in \mathbb{R}\}$ .
2.  $f(t, y)$  sia una funzione *lipschitziana* rispetto ad  $y$ . Tale proprietà significa l'esistenza di una costante  $L > 0$ , detta *costante di Lipschitz*, tale che

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2| \quad (7.8)$$

per ogni  $t \in [a, b]$  e per ogni  $y_1, y_2 \in \mathbb{R}$ .

Ricordiamo che una *condizione sufficiente* affinché la funzione  $f(t, y)$  sia lipschitziana rispetto a  $y$  è che la derivata parziale  $\partial f/\partial y$  esista e sia limitata  $|\partial f(t, y)/\partial y| \leq L$ . Si ha allora il seguente risultato.

**Teorema 7.1 (Esistenza e unicità)** *Se la  $f(t, y)$  verifica le condizioni 1 e 2, allora per ogni  $t_0 \in [a, b]$  e ogni  $y_0 \in \mathbb{R}$ , esiste una ed una sola soluzione del problema di Cauchy (7.7). Tale soluzione è continua insieme alla derivata prima nell'intervallo  $[a, b]$ .*

Osserviamo che la condizione 2 è essenziale per l'unicità della soluzione. Come controesempio si può considerare il seguente problema di Cauchy

$$\begin{cases} y'(t) = 3/2 y^{1/3} \\ y(0) = 0 \end{cases} \quad (7.9)$$

Esso ammette come soluzioni la funzione identicamente nulla  $y(t) \equiv 0$  e la funzione  $y(t) = t^{3/2}$ . Più in generale, detto  $\alpha$  un numero positivo, le funzioni

$$y_\alpha(t) = \begin{cases} (t - \alpha)^{3/2} & \text{per } \alpha < t \\ 0 & \text{per } 0 \end{cases}$$

sono tutte soluzioni del problema (7.9). Rileviamo che la funzione  $f(t, y) = \frac{3}{2}y^{1/3}$  non è lipschitziana rispetto ad  $y$  (si prenda nella definizione, ad esempio,  $y_2 = 0$ ).

Per quanto riguarda la dipendenza della soluzione dai dati, consideriamo il seguente *problema perturbato*

$$\begin{cases} \tilde{y}'(t) = f(t, \tilde{y}) + \delta(t) \\ \tilde{y}(t_0) = y_0 + \epsilon_0 \end{cases} \quad (7.10)$$

ove  $\delta(t)$  e  $\epsilon_0$  sono piccole perturbazioni. Nell'ipotesi che la funzione  $f(t, y)$  sia lipschitziana, si può dimostrare che esistono due costanti positive  $k$  e  $\bar{\epsilon}$  tali che per ogni  $0 \leq \epsilon \leq \bar{\epsilon}$  le soluzioni  $y(t)$  e  $\tilde{y}(t)$  verificano la condizione

$$|\tilde{y}(t) - y(t)| \leq k \epsilon \quad \forall t \in [a, b] \quad (7.11)$$

quando  $|\epsilon_0| < \epsilon$ ,  $|\delta(t)| < \epsilon$ . La condizione (7.11) esprime una *dipendenza continua* della soluzione dai dati. Tale dipendenza è sicuramente importante nelle applicazioni, dal momento che i dati  $f$  e  $y_0$  possono essere affetti da errori. Tuttavia, non è sufficiente ad assicurare che gli errori trasmessi dai dati sulla soluzione siano *piccoli*. In effetti, la costante  $k$  in (7.11) dipende dalla costante di Lipschitz  $L$  della funzione  $f(t, y)$ . In particolare, quando  $L$  è grande e  $\partial f/\partial y > 0$  gli errori sul valore iniziale  $y_0$  aumentano esponenzialmente per  $t$  che cresce.

Come esempio illustrativo, si consideri il seguente problema a valori iniziali

$$\begin{cases} y'' - 4y' - 5y = 0 \\ y(0) = 1, \quad y'(0) = -1 \end{cases}$$

La soluzione generale dell'equazione differenziale è la funzione  $y = c_1 e^{-t} + c_2 e^{5t}$ , da cui la soluzione particolare, corrispondente alle condizioni iniziali assegnate,  $y = e^{-t}$ . Ora, se perturbiamo la condizione iniziale  $y(0) = 1$  ponendo  $y(0) = 1 + \epsilon$ , si trova che la soluzione del nuovo problema a valori iniziali è la funzione

$$y = \left(1 + \frac{5\epsilon}{6}\right) e^{-t} + \frac{\epsilon}{6} e^{5t}$$

che è rappresentata in Figura 7.4 in corrispondenza ad alcuni valori di  $\epsilon$ . Il problema a valori iniziali assegnato è quindi *instabile* o malcondizionato, dal momento che a *piccole* perturbazioni relative nelle condizioni iniziali corrispondono *grandi* variazioni relative nella soluzione. Per un problema di questo tipo *la risoluzione numerica è praticamente impossibile*. La presenza di tale instabilità indica, in sostanza, una *difficoltà* nella formulazione del modello.

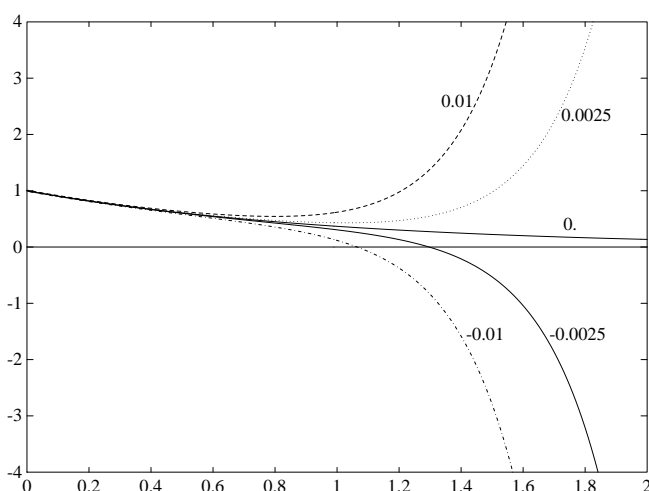


Figura 7.4: Soluzioni del problema a valori iniziali  $y'' - 4y' - 5y = 0$ ,  $y(0) = 1 + \epsilon$ ,  $y'(0) = -1$  in corrispondenza a diversi valori di  $\epsilon$ .

▼ **Osservazione 7.1** Sappiamo che il problema a valori iniziali  $y' = y^2$ ,  $y(0) = y_0$ , con  $y_0 \neq 0$  ha l'unica soluzione  $y(t) = y_0(1 - ty_0)^{-1}$ , definita sull'intervallo  $(-\infty, y_0^{-1})$ , se  $y_0 > 0$ , e sull'intervallo  $(y_0^{-1}, +\infty)$  se  $y_0 < 0$ . D'altra parte la funzione  $f(t, y) = y^2$  non è una funzione lipschitziana in  $y$  per  $y \in (-\infty, +\infty)$ . In realtà, il Teorema 7.1 può essere generalizzato nel seguente modo. Se la funzione  $f$  è continua e lipschitziana in una regione<sup>1</sup>

<sup>1</sup>Un insieme  $R$  del piano  $t$ - $y$  è una regione se  $R$  è un insieme aperto (cioè per ogni punto  $P$  in  $R$  esiste un cerchio con centro in  $P$  e con interno contenuto in  $R$ ), e se  $R$  è connesso (cioè due punti qualsiasi in  $R$  possono essere collegati con un cammino costituito da punti in  $R$ ).

$R$  e  $(t_0, y_0)$  è un punto di  $R$ , allora il problema a valori iniziali (7.4) ha una ed una sola soluzione su un intervallo  $I$  contenente  $t_0$  al suo interno. Inoltre, se  $D$  è una qualunque regione limitata contenuta in  $R$  e  $(t_0, y_0)$  è un punto interno a  $D$ , allora la soluzione di (7.4) può essere estesa in avanti e all'indietro fino a che la sua curva soluzione esce dalla frontiera di  $D$ . ■

## 7.2 Alcuni modelli

In questo paragrafo presenteremo alcuni esempi significativi di modelli basati sull'utilizzo delle equazioni differenziali.

► **Esempio 7.1** (*Oscillatore armonico*) Si consideri il moto di un punto materiale di massa  $m$  lungo una linea retta. Indichiamo con  $y(t)$  la posizione del punto in funzione del tempo. Le derivate  $y'(t)$ ,  $y''(t)$  sono rispettivamente la *velocità* e l'*accelerazione* del punto. Se sul punto agisce una forza  $F$ , il moto è regolato dalla *legge di Newton*, che stabilisce che la forza è uguale al prodotto della massa per l'accelerazione

$$my''(t) = F \quad (7.12)$$

Supponiamo ora che il punto materiale sia il punto medio di un *filo elastico* fissato agli estremi (cfr. Figura 7.5).

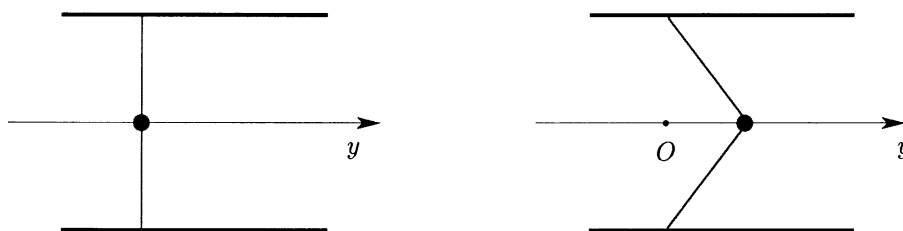


Figura 7.5: Oscillatore armonico.

L'origine  $O$  dell'asse  $y$  è fissata nella posizione di riposo del sistema corda-punto materiale. Spostando il punto dalla posizione di equilibrio, la corda esercita una forza di richiamo, e pertanto la forza  $F$  ha il segno opposto a quello dello spostamento  $y(t)$ . Se si *suppone* che l'intensità della forza sia proporzionale allo spostamento e si indica con  $k$  ( $> 0$ ) la costante di proporzionalità, si ha

$$F = -k y(t)$$

e quindi il modello matematico del sistema è il seguente

$$\boxed{y''(t) + \frac{k}{m}y(t) = 0} \quad (7.13)$$

cioè un'equazione differenziale lineare omogenea del secondo ordine. Le soluzioni sono (cfr. Appendice B) le funzioni della seguente famiglia (*integrale generale*)

$$y(t) = c_1 \cos\left(\sqrt{\frac{k}{m}} t\right) + c_2 \sin\left(\sqrt{\frac{k}{m}} t\right) \quad (7.14)$$

ove le costanti  $c_1, c_2$  possono essere fissate in maniera univoca imponendo all'istante iniziale  $t = 0$  la posizione  $y(0)$  e la velocità  $y'(0)$  iniziali. Ad esempio, per  $y(0) = y_0$  e  $y'(0) = 0$  (cioè il punto è inizialmente fermo) si ha la seguente *soluzione particolare*

$$y(t) = y_0 \cos\left(\sqrt{\frac{k}{m}} t\right)$$

Allo scopo di rendere più *realistico* il modello precedente, supponiamo che la forza che agisce sul punto materiale dipenda anche da un termine di *attrito* proporzionale alla velocità, e quindi

$$F = -k_1 y'(t) - k_2 y(t)$$

con  $k_1, k_2$  costanti positive. In corrispondenza, si ottiene il seguente modello

$$\boxed{y''(t) + \frac{k_1}{m} y'(t) + \frac{k_2}{m} y(t) = 0} \quad (7.15)$$

In questo caso il comportamento della soluzione dipende dal segno della seguente quantità

$$\Delta = \frac{k_1^2}{m^2} - \frac{4k_2}{m}$$

In particolare, se  $\Delta > 0$ , le soluzioni sono date da

$$y(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

ove  $\lambda_1, \lambda_2$  sono le soluzioni dell'equazione quadratica  $\lambda^2 + k_1/m\lambda + k_2/m = 0$ , cioè

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = -\frac{k_1}{2m} \pm \frac{1}{2} \sqrt{\frac{k_1^2}{m^2} - \frac{4k_2}{m}}$$

Essendo  $\Delta < k_1^2/m^2$ , si ha  $\sqrt{\Delta} < k_1/m$ , e quindi le radici  $\lambda_1, \lambda_2$  sono entrambe *negative*. In questo caso pertanto la soluzione  $y(t)$  tende a zero (cioè alla posizione di riposo) per  $t \rightarrow +\infty$ . Un analogo risultato si ottiene per  $\Delta = 0$ .

Per  $\Delta < 0$  si ha un *moto armonico smorzato* descritto dalle seguenti funzioni

$$y(t) = e^{-(k_1/2m)t} \left[ c_1 \cos\left(\frac{\sqrt{-\Delta}}{2} t\right) + c_2 \sin\left(\frac{\sqrt{-\Delta}}{2} t\right) \right]$$

Terminiamo l'esempio con l'analisi di un modello di origine differente dal precedente, ma analogo dal punto di vista matematico. Più precisamente, consideriamo il circuito elettrico illustrato in Figura 7.6. Il simbolo  $E$  rappresenta una sorgente di forza elettromotrice. Tale sorgente, una batteria o un generatore, produce una differenza di potenziale che genera un flusso di corrente  $I$  attraverso il circuito, quando l'*interruttore*  $S$  è chiuso. Il simbolo  $R$  rappresenta una *resistenza* al flusso di corrente. Quando la corrente fluisce attraverso un avvolgimento  $L$ , la differenza di potenziale prodotta dall'avvolgimento è proporzionale alla variazione dell'intensità  $I$  di corrente; la costante di proporzionalità è chiamata *induttanza*  $L$  dell'avvolgimento. Il simbolo  $C$  indica un *condensatore*, la cui carica  $Q(t)$  viene assunta come incognita del problema.

Per determinare  $Q(t)$  si utilizza un modello matematico basato sulla *seconda legge di Kirchhoff*. Tale legge stabilisce che in un circuito chiuso la differenza di potenziale fornita attraverso la sorgente  $E$  uguaglia la somma delle differenze di potenziale nel resto del circuito. Tenendo presente che

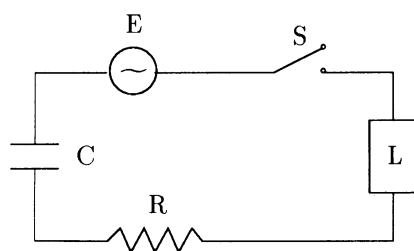


Figura 7.6: Circuito elettrico.

1. la differenza di potenziale lungo una resistenza di  $R$  ohm è data da  $RI$  (legge di Ohm),
2. la differenza di potenziale lungo una induttanza di  $L$  henry è data da  $L(dI/dt)$ ,
3. la differenza di potenziale lungo un condensatore di  $C$  farad è data da  $Q/C$ ,

si ha l'equazione differenziale

$$E(t) = L \frac{dI}{dt} + RI + \frac{Q}{C}$$

da cui, essendo  $I(t) = dQ(t)/dt$

$$\boxed{L \frac{d^2 Q}{dt^2} + R \frac{dQ}{dt} + \frac{Q}{C} = E(t)} \quad (7.16)$$

Sottolineiamo la stretta analogia dell'equazione (7.16) con l'equazione del moto armonico (7.15). È in sostanza questo tipo di analogie, che si riscontrano nello studio di differenti fenomeni, a giustificare uno studio *astratto* delle corrispondenti equazioni. Tale studio ha, in effetti, l'obiettivo di mettere in rilievo proprietà di validità generale, che prescindono dal particolare fenomeno studiato.

A titolo di esemplificazione, ricordiamo che nel caso  $R = 0$  e  $E = E_0 \sin \omega t$ , si ottiene la seguente soluzione, che verifica le condizioni iniziali  $Q(0) = Q'(0) = 0$

$$Q(t) = \frac{-E_0}{L(p^2 - \omega^2)} \frac{\omega}{p} \sin pt + \frac{E_0}{L(p^2 - \omega^2)} \sin \omega t$$

ove  $p = 1/\sqrt{LC}$ . La soluzione risulta essere la somma di due funzioni periodiche di periodo differente. Se, ad esempio  $L = 1$ ,  $E_0 = 1$ ,  $C = 0.25$ ,  $\omega = \sqrt{3.5}$ , si ottiene  $p = 2$  e

$$Q(t) = -1.870828 \sin 2t + 2 \sin 1.870828t$$

La soluzione è rappresentata in Figura 7.7.

Una situazione interessante corrisponde al caso in cui si ha  $p = \omega$ , cioè quando la frequenza  $\omega$  della sorgente esterna *uguaglia* la frequenza naturale del sistema. Questa situazione è nota come *fenomeno di risonanza*. In questo caso si può vedere che la soluzione del sistema ha una componente di tipo oscillatorio, la cui ampiezza di oscillazione cresce indefinitamente. In Figura 7.8 è rappresentata la funzione

$$Q(t) = \frac{1}{4} \sin 2t - 2t \cos 2t$$



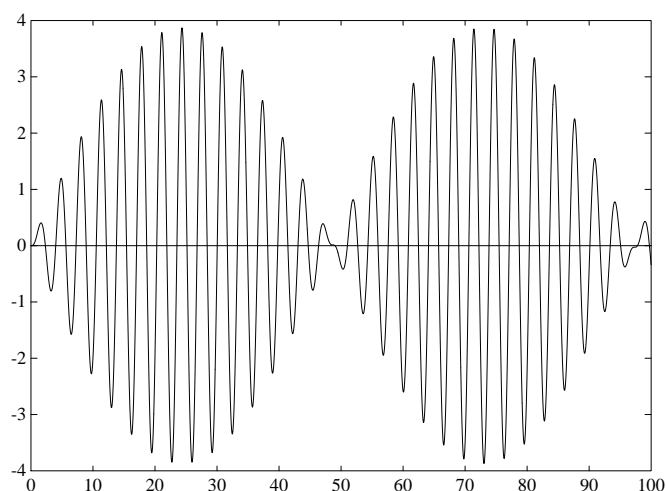


Figura 7.7: Grafico della funzione  $Q(t) = -1.870828 \sin 2t + 2 \sin 1.870828$ .

che corrisponde alla soluzione dell'equazione (7.16) con  $R = 0$  e  $\omega = p = 2$  e le altre quantità scelte come nell'esempio precedente.

È interessante osservare che, mentre nello studio di strutture meccaniche il fenomeno della risonanza è, in generale, da evitare, in quanto possibile causa di collasso delle strutture stesse<sup>2</sup>, nel caso dei circuiti elettrici esso può essere di utilità per amplificare opportunamente un segnale. Ricordiamo, in particolare, il suo utilizzo in chimica in differenti tecniche (*nuclear magnetic resonance* (NMR)<sup>3</sup>, *proton magnetic resonance* (H-NMR), *electron spin resonance* (ESR)<sup>4</sup> per identificare le strutture molecolari. ■

► **Esempio 7.2** (*Decadimento radioattivo*) Alcuni elementi, o loro isotopi, sono *instabili*, nel senso che *decadono* in isotopi di altri elementi mediante l'emissione di particelle *alpha* (nuclei di elio), particelle *beta* (elettroni), o *fotoni*. Tali elementi sono detti *radioattivi*. Ad esempio, un atomo di radio può decadere in un atomo di radon, con emissione di una particella alpha  $^{226}\text{Ra} \xrightarrow{\alpha} ^{222}\text{Rn}$ . Il decadimento di un singolo nucleo radioattivo è un evento random (cfr. successivo Capitolo 8) e quindi il tempo di decadimento non può essere previsto con certezza. Si può, tuttavia, descrivere il processo di decadimento di un numero elevato di nuclei radioattivi basandosi sulla seguente legge sperimentale

<sup>2</sup>Un esempio famoso di risonanza *distruttiva* fu il crollo, a seguito di oscillazioni di ampiezza sempre più grandi causate da correnti di vento, del ponte Tacoma Narrows a Puget Sound nello stato di Washington nel 7 Novembre del 1940, cinque mesi dopo la sua costruzione. In fisiologia, in particolari patologie neuromuscolari, la risonanza può essere la causa di ampie oscillazioni nel movimento degli arti.

<sup>3</sup>NMR: *the study of the properties of molecules containing magnetic nuclei by applying a magnetic field and observing the frequency at which they come into resonance with an electromagnetic field* (P. W. Atkins).

<sup>4</sup>ESR: *the study of molecules containing unpaired electrons by observing the magnetic fields at which they come into resonance with monochromatic radiation* (P. W. Atkins).

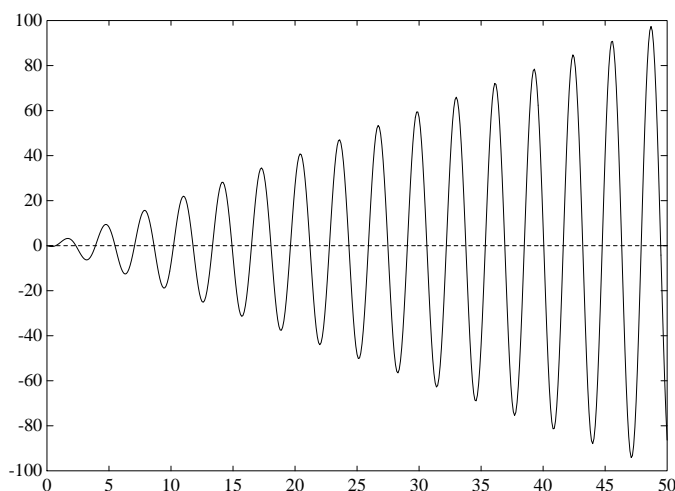


Figura 7.8: Grafico della funzione  $Q(t) = \sin(2t)/4 - 2t \cos(2t)$ .

*In un campione con un numero elevato di nuclei radioattivi, la diminuzione nel numero di nuclei radioattivi durante un dato intervallo di tempo è direttamente proporzionale alla lunghezza dell'intervallo di tempo e al numero di nuclei presenti all'inizio dell'intervallo.*

Indicando con  $N(t)$  il numero di nuclei radioattivi nel campione al tempo  $t$  e l'intervallo di tempo con  $\Delta t$ , la legge si esprime mediante la relazione

$$N(t + \Delta t) - N(t) = -k N(t) \Delta t \quad (7.17)$$

ove  $k$  è una costante positiva di proporzionalità. Si ha cioè (cfr. successivo Esempio 7.4) una *reazione del primo ordine* con *costante di velocità*  $k$ . Osserviamo che  $N(t)$  è un numero intero, ma che  $k \Delta t$  non è necessariamente intero. Per rendere, quindi, consistente il modello matematico (7.17), è necessario *idealizzare* il fenomeno, interpretando  $N(t)$  come una quantità *continua*, anziché discreta. Dal punto di vista fisico,  $N(t)$  può indicare ad esempio una misura della massa.

Dividendo ambo i membri della relazione (7.17) per  $\Delta t$  e passando al limite per  $\Delta t \rightarrow 0$ , si ottiene la seguente equazione differenziale

$$N'(t) = \lim_{\Delta t \rightarrow 0} \frac{N(t + \Delta t) - N(t)}{\Delta t} = -k N(t) \quad (7.18)$$

L'insieme delle soluzioni di tale equazione è dato da

$$N(t) = C e^{-kt}, \quad -\infty < t < \infty \quad (7.19)$$

ove  $C$  è una costante arbitraria. Se si conosce il valore  $N_0$  di  $N(t)$  ad un istante  $t_0$ , si ricava  $N_0 = C e^{-kt_0}$ , da cui  $C = N_0 e^{kt_0}$  e pertanto

$$N(t) = N_0 e^{-k(t-t_0)} \quad (7.20)$$

Si definisce *tempo di dimezzamento* (half-time)  $t_{1/2}$  il tempo richiesto affinché una assegnata quantità di nuclei radioattivi sia dimezzata. Da (7.19) si ottiene

$$\frac{N(t + t_{1/2})}{N(t)} = \frac{1}{2} = \frac{N_0 e^{k(t_0 - t - t_{1/2})}}{N_0 e^{k(t_0 - t)}} = e^{-kt_{1/2}} \Rightarrow k = \frac{1}{t_{1/2}} \ln 2$$

Secondo il modello assunto si ha che il tempo di dimezzamento è indipendente da  $N_0$ ,  $t_0$  e  $t$ . Il modello (7.20), con  $k$  calcolato nel modo precedente, può essere utilizzato per avere previsioni di valori di  $N(t)$  per valori di  $t \neq t_0$ . Come applicazione, consideriamo il problema della determinazione dell'età di reperti archeologici di organismi viventi mediante l'analisi del contenuto di carbone radioattivo. Ricordiamo che le cellule viventi assorbono direttamente o indirettamente carbone dall'anidride carbonica ( $CO_2$ ) contenuta nell'aria. Gli atomi di carbone contenuti nell'anidride carbonica sono divisi, in una proporzione che supporremo costante sia nelle cellule che nell'aria, in una forma radioattiva  $^{14}C$ , insieme alla forma comune  $^{12}C$ . La forma radioattiva è prodotta dalla collisione dei raggi cosmici (neutroni) con l'azoto nell'atmosfera. I nuclei di  $^{14}C$  decadono ad atomi di azoto emettendo particelle beta. Si ha quindi che gli esseri viventi, o che sono vissuti, contengono una certa quantità di nuclei radioattivi di carbone  $^{14}C$ . Quando l'organismo muore, cessa l'assorbimento di  $CO_2$  e continua solo il decadimento radioattivo. Il tempo di dimezzamento  $t_{1/2}$  del  $^{14}C$  è dato da  $5568 \pm 30$  anni.

Sia quindi  $q(t)$  la quantità di  $^{14}C$  per grammo di carbone al tempo  $t$  (misurato in anni) in un determinato reperto biologico. Sia  $t = 0$  il tempo corrente e  $T < 0$  l'istante della morte. Allora

$$q(t) \equiv q(T) \equiv q_T, \quad \forall t \leq T; \quad q'(t) = -k q(t), \quad T < t \leq 0; \quad q(0) = q_0$$

ove  $q_0$  indica la quantità di nuclei radioattivi al tempo corrente. L'espressione analitica della soluzione del problema differenziale precedente è data da

$$q(t) = q_T e^{-k(t-T)}, \quad T \leq t \leq 0$$

da cui

$$q_0 = q(0) = q_T e^{kT}$$

Tenendo conto che la costante  $k$  può essere stimata a partire dal tempo di dimezzamento  $t_{1/2}$  di  $^{14}C$  attraverso la relazione  $kt_{1/2} = \ln 2$ , si può valutare  $T$  quando sono noti  $q_0$  e  $q_T$ ; in realtà, è sufficiente conoscere il rapporto  $q_0/q_T$ . Si procede allora nel seguente modo. Mediante contatore Geiger si valuta il rapporto  $q'(0)/q'(T)$ , utilizzando il fatto che il numero di disintegrazioni di  $^{14}C$  per unità di tempo è proporzionale alla velocità di decadimento  $q'(t)$ . D'altra parte dall'equazione differenziale si ha

$$q'(0) = -k q(0) = -k q_0; \quad q'(T+) = -k q(T) = -k q_T$$

per cui

$$T = \frac{1}{k} \ln \frac{q_0}{q_T} = \frac{t_{1/2}}{\ln 2} \ln \frac{q'(0)}{q'(T+)}$$

Il procedimento precedente è basato sull'ipotesi che il rapporto tra  $^{14}C$  e  $^{12}C$  nell'atmosfera sia costante nel tempo. In realtà, tuttavia, tale rapporto è soggetto a variazioni, di cui il procedimento precedente deve tenere opportunamente conto nella stima di  $q_T$ . ■

► **Esempio 7.3** (*Modello a compartimenti per lo studio della concentrazione del piombo nel corpo umano*) Il piombo è assorbito dal corpo attraverso la respirazione, i cibi e le bevande. Dal polmone e dall'intestino il piombo entra nel sangue e quindi distribuito al fegato e ai reni. Esso è assorbito lentamente dagli altri tessuti e molto lentamente dalle ossa. Il piombo è eliminato dal corpo principalmente attraverso il sistema urinario e i capelli, le unghie, e il sudore. Il flusso del piombo attraverso il corpo può essere studiato dal punto di vista matematico mediante la *tecnica dei compartimenti*, che verrà analizzata più in dettaglio nel successivo Capitolo 12. La Figura 7.9 fornisce una rappresentazione schematica del fenomeno fisiologico; ogni organo interessato è rappresentato da un box (compartimento), mentre gli scambi fra i vari organi sono indicati da opportuni vettori. Il modello matematico è basato sul seguente principio di conservazione

*La velocità complessiva di ricambio di una sostanza in un compartimento è uguale alla velocità di ingresso meno la velocità di uscita.*

Indichiamo con  $x_i(t)$  la quantità di piombo nel compartimento  $i$  al tempo  $t$ . Assumeremo che la velocità di trasferimento del piombo dal compartimento  $j$  al compartimento  $i$ , con  $i \neq j$ , sia proporzionale alla quantità  $x_j$ , mentre la velocità di flusso contrario da  $i$  a  $j$  sia proporzionale a  $x_i$ . Indicheremo con  $a_{ij}$  il fattore di proporzionalità da  $j$  in  $i$  e con  $a_{ji}$  quello da  $i$  in  $j$ .

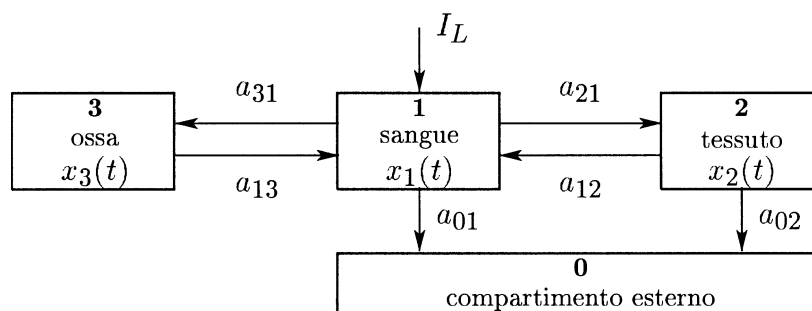


Figura 7.9: Assorbimento, distribuzione e escrezione del piombo nei tessuti.

Nelle precedenti ipotesi si ha il seguente *modello matematico*

$$\begin{aligned}
 (\text{sangue}) \quad x_1' &= -(a_{01} + a_{21} + a_{31})x_1 + a_{12}x_2 + a_{13}x_3 + I_L \\
 (\text{tessuto}) \quad x_2' &= a_{21}x_1 - (a_{02} + a_{12})x_2 \\
 (\text{ossa}) \quad x_3' &= a_{31}x_1 - a_{13}x_3
 \end{aligned} \tag{7.21}$$

ove  $I_L(t)$  indica la velocità di assorbimento del piombo nel sangue. Il compartimento 0 rappresenta l'ambiente esterno. Il fatto che non abbiamo preso in considerazione la corrispondente equazione di bilancio significa che nel modello esaminato non è ipotizzato un ritorno di piombo da questo compartimento nel sistema.

I coefficienti  $a_{ij}$  sono per il loro significato fisico delle quantità *positive*, da determinare sulla base di un procedimento di identificazione a partire da dati sperimentali. Una volta che tali coefficienti sono determinati, le soluzioni del sistema dinamico (7.21), corrispondenti a particolari condizioni iniziali, ad esempio  $x_1(0) = x_2(0) = x_3(0) = 0$ , forniscono per ogni input  $I_L$  la distribuzione nel tempo del piombo nei tre compartimenti considerati.

Si può dimostrare analiticamente che quando i coefficienti  $a_{ij}$  sono costanti (positive) e la velocità di input  $I_L$  è pure costante, le soluzioni del sistema (7.21) raggiungono una soluzione *stazionaria*  $[\bar{x}_1, \bar{x}_2, \bar{x}_3]$ , che si ottiene risolvendo il seguente sistema lineare nelle incognite  $\bar{x}_i$ ,  $i = 1, 2, 3$

$$\begin{aligned} 0 &= -(a_{01} + a_{21} + a_{31})\bar{x}_1 + a_{12}\bar{x}_2 + a_{13}\bar{x}_3 + I_L \\ 0 &= a_{21}\bar{x}_1 - (a_{02} + a_{12})\bar{x}_2 \\ 0 &= a_{31}\bar{x}_1 - a_{13}\bar{x}_3 \end{aligned}$$

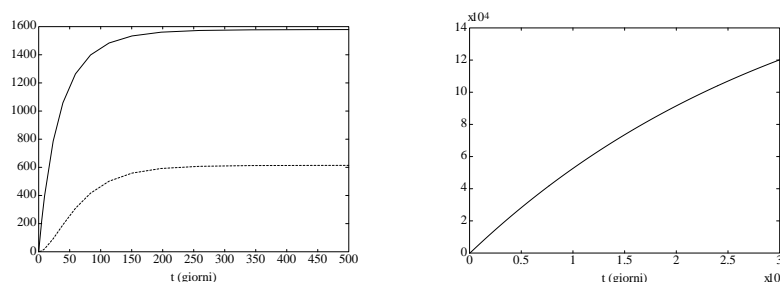


Figura 7.10: Nella prima figura è rappresentato l'assorbimento del piombo nel sangue (—) e nel tessuto (---) e nella seconda figura l'assorbimento nel tessuto osseo.

Come illustrazione, consideriamo il seguente caso particolare

$$\begin{aligned} x'_1 &= -\frac{65}{1800}x_1 + \frac{1088}{87500}x_2 + \frac{7}{200000}x_3 + 49.3 \\ x'_2 &= \frac{20}{1800}x_1 - \frac{20}{700}x_2 \\ x'_3 &= \frac{7}{1800}x_1 - \frac{7}{200000}x_3 \end{aligned}$$

ove  $x'_i$ ,  $i = 1, 2, 3$  e  $I_L$  sono espresse in unità di microgrammi ( $\mu g$ ) di piombo per giorno. Si ottiene lo *stato stazionario* per

$$\bar{x}_1 = 1800, \quad \bar{x}_2 = 701, \quad \bar{x}_3 = 200010 \quad (7.22)$$

In Figura 7.10 sono riportate le soluzioni  $x_1(t)$  e  $x_2(t)$ , ottenute numericamente mediante un metodo di Runge–Kutta (cfr. paragrafo successivo). Come si vede, sia nel sangue che nel tessuto la condizione di equilibrio è raggiunta rapidamente, contrariamente a quanto avviene nel tessuto osseo.

Nella Figura 7.11 sono riportate le soluzioni numeriche ottenute in corrispondenza a  $I_L = 0$  e al valore iniziale posto uguale allo stato stazionario indicato in (7.22). Come prevedibile, mentre il sangue e i tessuti si liberano rapidamente del contenuto di piombo, il rilascio relativo al tessuto osseo è decisamente più lento. ■

► **Esempio 7.4** (*Cinetica chimica*) La cinetica chimica è la branca della chimica che tratta la velocità e il meccanismo di una reazione chimica allo scopo di scoprire e spiegare i

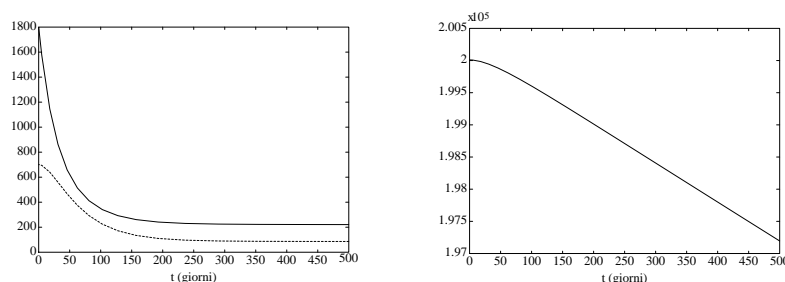


Figura 7.11: Nella prima figura è rappresentato l'eliminazione del piombo nel sangue (—) e nel tessuto (---) e nella seconda figura l'eliminazione nel tessuto osseo.

fattori che influenzano la velocità e il modo nel quale procede la reazione<sup>5</sup>. Una reazione che avviene in una *singola fase* può essere caratterizzata in ogni punto dalle seguenti variabili: la *velocità*, la *concentrazione* delle specie chimiche, e una *variabile termodinamica*: l'energia interna o la temperatura.

Un sistema chimico è chiamato *uniforme* se non vi sono variazioni spaziali entro il sistema. Nella breve introduzione che presenteremo nel seguito assumeremo che il sistema chimico considerato sia omogeneo, uniforme e con volume e temperatura costanti. In particolare, studieremo l'evoluzione del sistema nel tempo, e le variabili usate per lo studio di tale evoluzione saranno le *concentrazioni* delle specie interessate alle reazioni. Tali variabili sono chiamate *variabili di stato*. In definitiva, la *cinetica chimica studia il modo nel quale una sistema reagente passa da uno stato ad un altro e il tempo richiesto per effettuare tale transizione*.

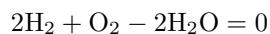
Nel seguito utilizzeremo notazioni e idee che sono tipiche nella chimica cinetica e nella stechiometria<sup>6</sup>; supporremo il sistema composto da  $N$  specie chimiche  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N$ . Per esempio, la reazione chimica



consiste di tre specie  $\mathcal{M}_1 = \text{H}_2\text{O}$ ,  $\mathcal{M}_2 = \text{H}_2$ , e  $\mathcal{M}_3 = \text{O}_2$ . Una reazione chimica è usualmente scritta nella forma

$$\sum_{i=1}^N \nu_i \mathcal{M}_i = 0 \quad (7.24)$$

ove  $\nu_i$  sono i *coefficienti stechiometrici* delle specie  $\mathcal{M}_i$  nell'equazione di bilancio della reazione. La reazione (7.23) può essere scritta nella seguente forma



<sup>5</sup>One reason for studying the rates of reactions is the practical importance of being able to predict how quickly a reaction mixture approaches equilibrium. The rate might depend on variables under our control, such as the pressure, the temperature, and the presence of a catalyst, and we may be able to optimize it by the appropriate choice of conditions. Another reason is that the study of reaction rates leads to an understanding of the **mechanisms** of reactions, their analysis into a sequence of elementary reactions (P. W. Atkins).

<sup>6</sup>La *stechiometria* è quella parte della chimica che tratta i rapporti ponderali fra gli atomi nei composti e fra molecole ed atomi nelle reazioni chimiche.

Per convenzione, diremo che una specie  $\mathcal{M}_i$  è un *reagente* se  $\nu_i < 0$  e un *prodotto* se  $\nu_i > 0$ , per  $i = 1, 2, \dots, N$ . La notazione (7.24) può essere generalizzata per descrivere un *sistema* di  $R$  reazioni, ossia

$$\sum_{i=1}^N \nu_{ij} \mathcal{M}_i = 0, \quad j = 1, 2, \dots, R \quad (7.25)$$

ove  $\nu_{ij}$  sono i coefficienti stechiometrici delle specie  $\mathcal{M}_i$  nella reazione  $j$ -ma, per  $i = 1, 2, \dots, N$  e  $j = 1, 2, \dots, R$ . Osserviamo che dall'equazione stechiometrica (7.24) si ha:

$$\frac{\delta n_i}{\delta n_k} = \frac{\nu_i}{\nu_k}, \quad 1 \leq i \leq N, 1 \leq k \leq N$$

dove  $\delta n_i$  rappresenta un cambiamento nel numero di moli delle specie  $\mathcal{M}_i$  nel sistema chimico. Nell'esempio (7.23), per ogni mole<sup>7</sup> di acqua che è decomposta per elettrolisi, si ha la generazione di una mezza-mole di ossigeno e di una mole di idrogeno.

**Definizione 7.1 (velocità di reazione)** La *velocità di reazione*  $r_i(t)$ ,  $i = 1, 2, \dots, N$ , è la *velocità con cui varia la concentrazione di una specie fissata  $\mathcal{M}_i$  interessata alla reazione*.

Se  $c_i(t)$  è la funzione che rappresenta la concentrazione della specie  $\mathcal{M}_i$  al tempo  $t$ , allora  $r_i(t) = dc_i(t)/dt$ . Nella reazione (7.23) si ha

$$r_1(t) = \frac{dc_1}{dt}(t) = -\frac{dc_2}{dt}(t) = -2\frac{dc_3}{dt}(t)$$

cioè  $r_1(t) = -r_2(t) = -2r_3(t)$ . La velocità di reazione, a temperatura fissata  $T$ , è funzione soltanto delle concentrazioni delle varie specie, e quindi può essere espressa nella forma

$$r(t) = K_T f(c_1(t), c_2(t), \dots, c_N(t)) \quad (7.26)$$

L'indice  $T$  della costante  $K$  indica che tale costante può dipendere dalla temperatura  $T$  alla quale avviene la reazione, ma che è indipendente dalla concentrazione. Una forma usualmente utilizzata, nota come legge di Arrhenius (1889) è la seguente

$$K_T = Ae^{-E/RT}$$

ove  $A$  e  $E$  sono costanti indipendenti dalla temperatura.

Uno dei principali scopi della chimica sperimentale consiste nell'ottenere la forma dell'espressione (7.26). La procedura usuale consiste nel postulare una particolare forma e nel pianificare opportune sperimentazioni per validare tale scelta. Si vede, quindi, che dal punto di vista matematico il principale problema della cinetica chimica è un *problema inverso*, cioè un *problema di identificazione* del modello<sup>8</sup>.

Una forma particolare delle funzioni (7.26) è la seguente

$$r(t) = K_T [c_1(t)]^{\alpha_1} [c_2(t)]^{\alpha_2} \dots [c_N(t)]^{\alpha_N} \quad (7.27)$$

<sup>7</sup>Si definisce *mole* la quantità di materia di un sistema che contiene tante unità elementari quanti atomi sono contenuti in 0.012Kg di <sup>12</sup>C. Tale numero è il *numero di Avogadro*:  $6.02204 \cdot 10^{23}$ .

<sup>8</sup>*Rate laws have two main applications. A practical application is that one we know the rate law and the rate constant we can predict the rate of reaction from the composition of the mixture. The theoretical application of a rate law is that it is a guide to the mechanism of the reaction, and any proposed mechanism must be consistent with the observed rate law* (P. W. Atkins).

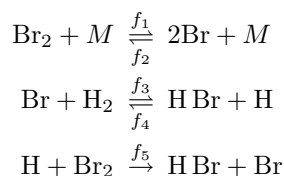
ove  $K_T, \alpha_1, \alpha_2, \dots, \alpha_N$  sono costanti da identificare. Il numero  $\alpha = \alpha_1 + \alpha_2 + \dots + \alpha_N$  viene anche detto *ordine della reazione*, mentre  $\alpha_i$  è l'ordine della reazione rispetto alla componente  $\mathcal{M}_i$ .

Nel caso del sistema di reazioni (7.25) si può associare ad ogni reazione una funzione  $f_j$  delle  $N$  concentrazioni  $c_i(t)$  tale che

$$\frac{dc_i}{dt}(t) = \sum_{j=1}^R \nu_{ij} f_j(c_1(t), c_2(t), \dots, c_N(t)) \quad (7.28)$$

per  $i = 1, 2, \dots, N$ . Le funzioni  $f_1, f_2, \dots, f_R$  definiscono la cinetica del sistema. Noti i valori delle concentrazioni  $c_i(t)$  ad un valore iniziale  $t_0$ , l'evoluzione del sistema è allora descritta dalle soluzioni di un *problema a valori iniziali*. Il modello ottenuto è noto come *modello deterministico di massa e azione* della reazione<sup>9</sup>.

Come illustrazione, consideriamo la classica reazione  $\text{H}_2 + \text{Br}_2 \xrightarrow{f} 2\text{HBr}$  tra idrogeno e bromo per formare bromuro di idrogeno HBr. Tale reazione è basata sul seguente meccanismo, costituito da un insieme di *reazioni elementari*, ognuna delle quali coinvolge solo una o due molecole<sup>10</sup>



ove  $M$  è  $\text{H}_2$  o  $\text{Br}_2$ . *Sperimentalmente*, si è determinato che le funzioni di velocità  $f_j$  delle reazioni elementari sono della forma

$$\begin{aligned} f_1(c_1(t), \dots, c_5(t)) &= K_1 c_2(t) [c_1(t) + c_2(t)] \\ f_2(c_1(t), \dots, c_5(t)) &= K_2 [c_5(t)]^2 [c_1(t) + c_2(t)] \\ f_3(c_1(t), \dots, c_5(t)) &= K_3 c_5(t) c_1(t) \\ f_4(c_1(t), \dots, c_5(t)) &= K_4 c_3(t) c_4(t) \\ f_5(c_1(t), \dots, c_5(t)) &= K_5 c_4(t) c_2(t) \end{aligned}$$

ove  $\mathcal{M}_1 = \text{H}_2, \mathcal{M}_2 = \text{Br}_2, \mathcal{M}_3 = \text{HBr}, \mathcal{M}_4 = \text{H}$  e  $\mathcal{M}_5 = \text{Br}$ . In corrispondenza, si ha:

$$\begin{aligned} r_3(t) &= \frac{dc_3}{dt} = \sum_{j=1}^5 \nu_{3j} f_j(c_1(t), c_2(t), \dots, c_5(t)) \\ &= 0 \cdot f_1(c_1(t), \dots, c_5(t)) + 0 \cdot f_2(c_1(t), \dots, c_5(t)) \\ &\quad + f_3(c_1(t), \dots, c_5(t)) - f_4(c_1(t), \dots, c_5(t)) + f_5(c_1(t), \dots, c_5(t)) \\ &= K_3 c_5(t) c_1(t) - K_4 c_3(t) c_4(t) + K_5 c_4(t) c_2(t) \end{aligned}$$

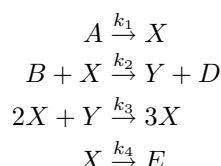
<sup>9</sup>La prima equazione cinetica di tipo massa e azione fu introdotta da Wilhelmy (1850) per lo studio della velocità di mutarotazione di zuccheri; la legge di massa e azione fu suggerita da Gulderg & Waage (1867). Sottolineiamo che la legge di massa e azione è un *postulato* nella teoria fenomenologica della cinetica delle reazioni chimiche.

<sup>10</sup>The chemical equation for an elementary reaction has a different significance from that for an overall reaction. It specifies the individual event, not merely the overall, net stoichiometry. Which meaning should be ascribed to the arrow  $\rightarrow$  will always be clear from the context (P. W. Atkins).



In modo analogo si costruiscono le equazioni differenziali relative alle concentrazioni delle altre componenti della reazione.

Come ulteriore esempio, esaminiamo il seguente modello di cinetica chimica, noto anche come *Brusselator*. Si hanno sei sostanze  $A, B, D, E, X, Y$  che reagiscono secondo il seguente meccanismo



Se indichiamo con  $A(t), B(t), \dots$  le concentrazioni di  $A, B, \dots$ , come funzioni del tempo, alle reazioni precedenti corrisponde il seguente sistema differenziale

$$\begin{aligned} A' &= -k_1 A \\ B' &= -k_2 B X \\ D' &= k_2 B X \\ E' &= k_4 X \\ X' &= k_1 A - k_2 B X + k_3 X^2 Y - k_4 X \\ Y' &= k_2 B X - k_3 X^2 Y \end{aligned}$$

Il sistema può essere semplificato nel seguente modo. Le equazioni per  $D$  e  $E$  sono trascurate, in quanto esse non influenzano le altre; inoltre,  $A$  e  $B$  sono supposte costanti e infine le velocità di reazione  $k_i$  sono poste uguali a 1. Posto  $y_1(t) := X(t)$  e  $y_2(t) := Y(t)$ , si ottiene

$$\begin{aligned} y_1' &= A + y_1^2 y_2 - (B + 1)y_1 \\ y_2' &= B y_1 - y_1^2 y_2 \end{aligned} \quad (7.29)$$

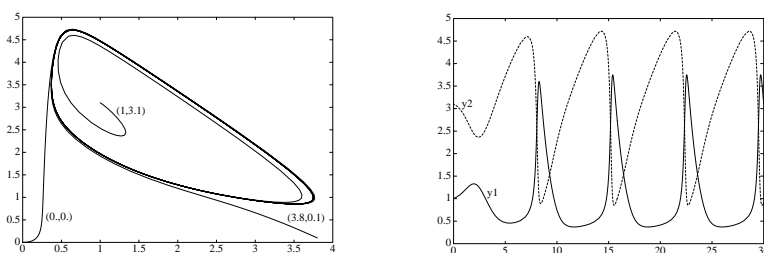


Figura 7.12: Soluzioni del sistema differenziale Brusselator.

Ricordiamo (cfr. Appendice B) che sono detti *punti critici* o punti stazionari i valori delle variabili  $(y_1, y_2)$  per le quali si annullano le derivate  $y_1', y_2'$ . Assumendo come valori iniziali tali punti, il sistema rimane invariato. Per il sistema che stiamo considerando si verifica facilmente che si ha un punto stazionario in  $y_1 = A, y_2 = B/A$ . Nelle applicazioni è importante analizzare la *stabilità* di tale punto. In maniera schematica, si tratta di vedere se *piccoli spostamenti* dal punto portano a soluzioni che evolvono *allontanandosi* o *avvicinandosi* dal

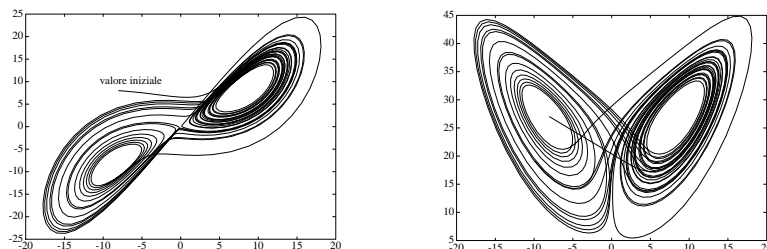


Figura 7.13: Modello di Lorenz corrispondente ai valori  $b = 8/3$ ,  $\sigma = 10$  e  $r = 28$  e ai valori iniziali  $y_1 = -8$ ,  $y_2 = 8$  e  $y_3 = r - 1$ . Nella prima figura sono rappresentate le traiettorie  $(y_1, y_2)$  e nella seconda le traiettorie  $(y_1, y_3)$ .

punto stazionario. Nel secondo caso si ha che il punto stazionario è un *punto attrattore*, o punto stabile. Nel caso del sistema (7.29) si può dimostrare che il punto stazionario  $(A, B/A)$  è un *punto instabile* quando  $B > A^2 + 1$ . Nella prima parte della Figura 7.12 sono indicate nel piano delle fasi le traiettorie  $(y_1, y_2)$ , calcolate numericamente e corrispondenti a tre diversi valori iniziali e per  $A + 1, B = 3$ . Nella seconda parte della Figura 7.12 sono rappresentate le traiettorie  $(y_1(t), y_2(t))$  corrispondenti a valori iniziali  $(1, 3.1)$ , che rappresenta una piccola perturbazione del punto stazionario. Le Figure mostrano il fatto importante che le traiettorie tendono a stabilizzarsi su un *ciclo periodico*. Si può vedere che quando  $B$  tende ad  $A^2 + 1$  il ciclo limite diviene sempre più piccolo e si riduce al punto critico. Tale fenomeno è noto come *biforcazione di Hopf*.

Un comportamento più generale è indicato nella Figura 7.13. Essa corrisponde al seguente sistema differenziale

$$\begin{aligned} y_1' &= -\sigma y_1 + \sigma y_2 \\ y_2' &= -y_1 y_3 + r y_1 - y_2 \\ y_3' &= y_1 y_2 - b y_3 \end{aligned}$$

ove  $\sigma, r$  e  $b$  sono costanti positive. Il sistema, noto in letteratura come *modello di Lorenz*, rappresenta un modello per lo studio della turbolenza atmosferica in una regione di aria all'interno di un accumulo di nubi (*J. Atmos. Sci.* (1963)). Il sistema presenta un ovvio punto critico in  $y_1 = y_2 = y_3 = 0$ , che è instabile quando  $r > 1$ . In questo caso vi sono due punti critici addizionali dati da

$$y_1 = y_2 = \pm \sqrt{b(r-1)}, \quad y_3 = r - 1$$

che risultano instabili quando  $\sigma > b + 1$  e

$$r \geq r_c = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1}$$

La Figura 7.13 corrisponde ai valori  $b = 8/3$ ,  $\sigma = 10$  e  $r = 28$ . Come valori iniziali si sono assunti i valori  $y_1 = -8$ ,  $y_2 = 8$  e  $y_3 = r - 1$ . I risultati, ottenuti numericamente sull'intervallo  $(0, 20)$ , mostrano l'assenza di soluzioni periodiche; tuttavia, le traiettorie rimangono limitate. In questo caso si dice che il comportamento delle soluzioni del sistema è di tipo

*caotico*. In effetti, a piccoli cambiamenti nelle condizioni iniziali corrispondono grandi e imprevedibili variazioni nelle orbite. In altre parole, comunque siano esatte le misurazioni delle condizioni iniziali, non è possibile stimare il valore della soluzione in tempi successivi. È opportuno, tuttavia, tenere presente che il modello di Lorenz rappresenta il risultato di una *semplificazione* della realtà. ■

## 7.3 Metodi numerici

In questo paragrafo introdurremo, attraverso opportuni esempi, le idee che sono alla base dell'*approssimazione numerica* del problema a valori iniziali

$$\begin{cases} y'(t) = f(t, y) \\ y(t_0) = y_0 \end{cases} \quad (7.30)$$

ove la funzione  $f(t, y)$  è supposta continua e lipschitziana in  $y$  (cfr. (7.8)). I metodi che analizzeremo hanno in comune la seguente idea di partenza. Si discretizza l'intervallo di integrazione  $(t_0, T)$  mediante i *nodi*  $t_{i+1} = t_i + h_i$ ,  $i = 0, 1, \dots, n-1$ , ove  $h_i$  sono detti *passi* della discretizzazione. Nel seguito, per motivi di semplicità, supporremo che tali passi siano *costanti*, cioè  $h_i = h$ ,  $i = 0, 1, \dots, n-1$ . In realtà, nelle applicazioni è usualmente necessario, per ottenere la soluzione approssimata al *minimo costo*, variare opportunamente il passo in funzione del comportamento della soluzione.

In corrispondenza all'insieme dei nodi  $\{t_i\}$ , un particolare metodo genera una *sequenza* di valori  $\{\eta_i\}$ . Tali valori definiscono la *soluzione discreta* e rappresentano le *approssimazioni* dei valori  $y(t_i)$  della soluzione del problema continuo. I vari metodi differiscono tra loro per il modo particolare con il quale vengono calcolati i valori  $\eta_i$ . Incominceremo dall'idea che abbiamo già utilizzato in precedenza per introdurre il *piano delle velocità* (cfr. Figura 7.1), cioè dall'approssimazione locale della curva soluzione mediante la *tangente*. Il metodo che ne risulta è noto come *metodo di Eulero*<sup>11</sup>. Data la sua semplicità, ci servirà anche per introdurre in maniera più elementare i vari concetti.

### 7.3.1 Metodo di Eulero

Il metodo è illustrato in forma grafica in Figura 7.14. L'equazione della tangente alla curva soluzione  $y(t)$  nel punto  $(t_0, y_0)$  è data da

$$y = y_0 + f(t_0, y_0)(t - t_0)$$

e fornisce per  $t = t_0 + h$  il valore

$$\eta_1 = y_0 + hf(t_0, y_0)$$

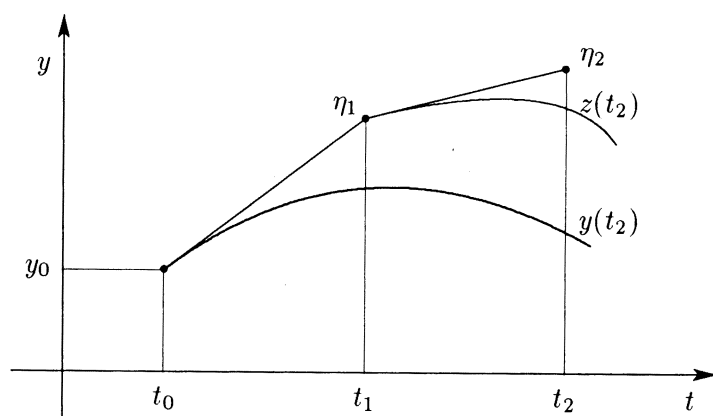


Figura 7.14: Metodo di Eulero. Errore di troncamento locale  $l(t_2) = z(t_2) - \eta_2$ ; errore globale  $e(t_2) = y(t_2) - \eta_2$ .

Si considera, allora, la tangente nel punto  $(t_1, \eta_1)$  alla curva soluzione dell'equazione differenziale che passa per il punto  $(t_1, \eta_1)$ . Osserviamo che tale curva, che indicheremo, per maggiore chiarezza, con  $z(t)$ , differisce in generale dalla  $y(t)$ , in quanto quest'ultima è individuata dal passaggio per il punto  $(t_0, y_0)$ . Per  $t = t_2$  l'equazione della tangente fornisce il valore

$$\eta_2 = \eta_1 + hf(t_1, \eta_1)$$

Procedendo in maniera analoga, si ottiene il seguente algoritmo, che fornisce la soluzione approssimata del problema a valori iniziali (7.30) in punti equidistanti sull'intervallo  $(t_0, T)$ .

**Algoritmo 7.1** (Metodo di Eulero)

Input:  $t_0, y_0, T$  e il numero  $n$  delle suddivisioni dell'intervallo  $(t_0, T)$ .

Output: soluzione approssimata  $\eta_i$ .

```

set  $h = (T - t_0)/n$ 
 $\eta_0 = y_0$ 
do  $i = 0, \dots, n - 1$ 
 $\eta_{i+1} = \eta_i + hf(t_i, \eta_i)$ 
 $t_{i+1} = t_i + h$ 
end do

```

In particolare, il valore  $\eta_n$  rappresenta una approssimazione della soluzione  $y(t)$  per  $t = t_n \equiv T$ .

<sup>11</sup>Il metodo è descritto da Eulero (1768) in *Institutiones Calculi Integralis* (Sectio Secunda, Caput VII).

► **Esempio 7.5** Come esempio illustrativo, si consideri il problema

$$\begin{cases} y'(t) = y(t) \sin 4t \\ y(0) = 1 \end{cases}$$

che ha come soluzione esatta la funzione  $y = e^{(1-\cos(4t))/4}$ . In Figura 7.15 sono riportati i risultati ottenuti mediante il metodo di Eulero sull'intervallo  $(0, 4)$  in corrispondenza a due valori del passo  $h$ . Più precisamente, nella figura sono rappresentate le funzioni lineari a tratti che si ottengono interpolando i valori  $\eta_i$  forniti dal metodo di Eulero. ■

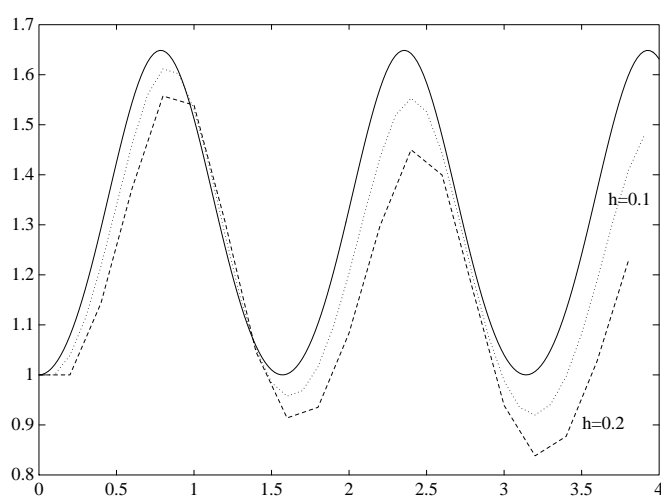


Figura 7.15: Risultati ottenuti con il metodo di Eulero per il problema a valori iniziali  $y' = y \sin(4t), y(0) = 1$ , in corrispondenza a  $h = 0.2$  e  $h = 0.1$ . A tratto continuo è rappresentata la soluzione analitica.

Come si vede dall'esempio precedente, l'approssimazione migliora al diminuire del passo  $h$ . In effetti, dal punto di vista delle applicazioni è importante che l'approssimazione possa essere migliorata fin che si vuole scegliendo opportunamente il passo  $h$ . In forma schematica, se questo avviene, si dice che lo schema è *convergente*. Analizzeremo, ora, più in dettaglio questo aspetto fondamentale.

### Studio della convergenza

Per ogni  $t_i, i = 0, 1, \dots, n$ , definiamo *errore globale* in  $t_i$  la differenza tra la soluzione esatta e la soluzione calcolata mediante il metodo numerico, ossia la quantità

$$e(t_i, h) := y(t_i) - \eta_i$$

La *convergenza* riguarda il comportamento della funzione  $h \rightarrow e(t_i, h)$  al tendere a zero di  $h$ . Considerato, ad esempio, il punto  $t = T$ , per  $h \rightarrow 0$  si ha  $n \rightarrow \infty$ , e lo

schema verrà detto *convergente* nel punto  $T$  quando

$$\lim_{h \rightarrow 0} e(t_n, h) = 0 \iff \lim_{n \rightarrow \infty} \eta_n = y(T)$$

Un altro aspetto importante nello studio della convergenza riguarda l'analisi della *rapidità* di convergenza rispetto ad  $h$ . Tale analisi, oltre che fornire indicazioni sulla scelta di un particolare passo per ottenere una accuratezza desiderata, è anche alla base per l'utilizzo di formule di estrapolazione (analoghe a quelle esaminate nel capitolo precedente nell'ambito delle formule di quadratura). Analizzeremo, ora, la questione della convergenza nel caso specifico del metodo di Eulero, ma le idee, come vedremo, sono di validità più generale.

Osserviamo in Figura 7.14 il comportamento della soluzione approssimata rispetto alla soluzione continua nel punto  $t_2$ , che rappresenta, in realtà, la situazione in un generico punto  $t_i$  per  $i > 1$ .

Nell'errore globale  $y(t_2) - \eta_2$  si individuano due contributi. Il primo è l'errore introdotto dal metodo, dovuto al fatto che la soluzione passante dal punto  $(t_1, \eta_1)$  è stata sostituita dalla tangente. Questo errore, definito ad ogni avanzamento del passo, dalla quantità

$$l_i(h, f) := z(t_{i+1}) - [\eta_i + hf(t_i, \eta_i)]$$

è chiamato *errore di troncamento locale*. La quantità

$$\tau_i(h, f) := \frac{l_i(h, f)}{h} = \frac{z(t_i + h) - z(t_i)}{h} - f(t_i, z(t_i))$$

chiamata *errore di discretizzazione locale*, fornisce una misura di come l'equazione alle differenze, corrispondente al metodo numerico, approssima l'equazione differenziale data.

Quando l'errore di discretizzazione locale tende a zero per  $h \rightarrow 0$ , si dice che il metodo è *consistente*. Dal momento che nelle ipotesi fatte su  $f$  la soluzione  $y$  dell'equazione differenziale risulta derivabile, ne segue che *il metodo di Eulero è un metodo consistente*. In effetti, tale risultato equivale a dire che il rapporto incrementale tende alla derivata. Si può ulteriormente precisare tale convergenza, quando la funzione  $f$  è più regolare, ad esempio derivabile in  $t$  e  $y$ . Si ha, allora, che la soluzione  $z$  è dotata di derivata seconda e quindi vale il seguente sviluppo in serie

$$z(t_i + h) = z(t_i) + hz'(t_i) + \frac{h^2}{2}z''(t_i + \theta_i h) \quad (7.31)$$

ove  $\theta_i$  è un opportuno valore in  $(0, 1)$ . Ricordando che  $z'(t_i) = f(t_i, z(t_i))$ , si ha pertanto

$$l_i(h, f) := z(t_{i+1}) - [\eta_i + hf(t_i, \eta_i)] = \frac{h^2}{2}z''(t_i + \theta_i h) \quad (7.32)$$

da cui

$$|\tau_i(h, f)| \leq M h \quad (7.33)$$

con  $M \geq |z''(t)| = |f_t(t, y) + f(t, y)f_y(t, y)|$ ,  $t \in [t_0, T]$ ,  $y \in \mathbb{R}$ . Più in generale, quando l'errore di discretizzazione di un metodo tende a zero come  $h^r$ , cioè  $\tau(f, h) = O(h^r)$ , si dice che il metodo è di *ordine*  $r$ . In particolare, quindi, il metodo di Eulero è un metodo del primo ordine. Sottolineiamo che la precisazione sull'ordine del metodo ha richiesto un supplemento di regolarità della soluzione; quindi, se la regolarità non è quella richiesta, la precisione dell'approssimazione può essere inferiore.

Il secondo contributo all'errore globale  $y(t_{i+1}) - \eta_{i+1}$  è dovuto al fatto che la tangente è considerata nel punto  $(t_i, \eta_i)$ , anziché nel punto  $(t_i, y(t_i))$  della soluzione del problema assegnato (7.30). Tale contributo non è locale, in quanto esso rappresenta l'accumulo di tutti gli errori di troncamento locale commessi in precedenza. La Figura 7.14 suggerisce, anche, che l'entità di tale contributo dipende dal condizionamento del problema, cioè da come l'errore  $y(t_1) - \eta_1$  si propaga nell'errore  $y(t_2) - z(t_2)$ .

Quando  $h$  tende a zero, il numero degli errori di troncamento locale tende all'infinito. Non è, quindi, evidente a priori che la consistenza di un metodo, cioè la convergenza a zero dei singoli errori locali, comporti necessariamente la convergenza a zero dell'errore globale. In realtà, affinché si abbia la convergenza, occorre che *l'accumulo degli errori di troncamento locale si mantenga limitato quando  $h$  tende a zero*. Quando questo avviene, si dice che il metodo è *stabile*. La denominazione sottolinea il fatto che per un metodo stabile le perturbazioni sui risultati numerici, corrispondenti a delle perturbazioni sui dati, si mantengono limitate, non "esplodono", per  $h$  che tende a zero.

Dimostreremo, ora, che il metodo di Eulero è un *metodo stabile*; più precisamente, dimostreremo che il metodo è *convergente*. Verificheremo, quindi, sull'esempio particolare del metodo di Eulero, il seguente risultato di validità più generale

$$\text{convergenza} = \text{stabilità} + \text{consistenza}$$

Nel seguito signaleremo esempi di metodi consistenti, ma non stabili, e non convergenti.

Per dimostrare la stabilità utilizzeremo il seguente risultato sulle equazioni alle differenze.

**Lemma 7.1** *Se  $\alpha$  e  $\beta$  sono numeri reali positivi e  $\{w_i\}_{i=0}^k$  è una successione con  $w_0 \geq -\beta/\alpha$  e*

$$w_{i+1} \leq (1 + \alpha)w_i + \beta \quad \text{per ogni } i = 0, 1, 2, \dots, k-1 \quad (7.34)$$

*allora*

$$w_{i+1} \leq e^{(i+1)\alpha} \left( \frac{\beta}{\alpha} + w_0 \right) - \frac{\beta}{\alpha} \quad (7.35)$$

DIMOSTRAZIONE. Per ogni  $i$  fissato, dalla disuguaglianza (7.34) si ha

$$\begin{aligned} w_{i+1} &\leq (1 + \alpha)w_i + \beta \\ &\leq (1 + \alpha)[(1 + \alpha)w_{i-1} + \beta] + \beta \\ &\vdots \\ &\leq (1 + \alpha)^{i+1}w_0 + [1 + (1 + \alpha) + (1 + \alpha)^2 + \cdots + (1 + \alpha)^i]\beta \end{aligned}$$

Utilizzando la formula per il calcolo della somma della serie geometrica  $\sum_{j=0}^i (1 + \alpha)^j$  di ragione  $(1 + \alpha)$ , si ha

$$w_{i+1} \leq (1 + \alpha)^{i+1}w_0 + \frac{(1 + \alpha)^{i+1} - 1}{\alpha} \beta = (1 + \alpha)^{i+1} \left( \frac{\beta}{\alpha} + w_0 \right) - \frac{\beta}{\alpha}$$

Per concludere, è sufficiente tenere conto che per ogni  $x \geq -1$  si ha

$$0 \leq 1 + x \leq e^x$$

come si dimostra facilmente per sviluppo in serie della funzione  $e^x$ . ■

**Teorema 7.2 (Convergenza del metodo di Eulero)** *Sia  $y(t)$  la soluzione del problema (7.30), con  $f(t, y)$  continua e lipschitziana in  $y$  con costante  $L$ . Se  $\eta_i$ , per  $i = 0, 1, \dots, n$ , con  $h = (T - t_0)/n$ , indica la soluzione ottenuta con il metodo di Eulero, si ha la seguente maggiorazione dell'errore*

$$|y(t_i) - \eta_i| \leq \frac{e^{L|t_i - t_0|} - 1}{L} |\tau(h, f)| \quad (7.36)$$

ove  $\tau(h, f) = \max_{0 \leq i \leq n} \tau_i(h, f)$ . In particolare, quando la soluzione  $y(t)$  ammette la derivata seconda e  $|y''(t)| \leq M$ , allora

$$|y(t_i) - \eta_i| \leq \frac{e^{L|t_i - t_0|} - 1}{L} \frac{hM}{2} \quad (7.37)$$

DIMOSTRAZIONE. Per la soluzione  $y(t)$  del problema (7.30) si ha (cfr. (7.31))

$$y(t_{i+1}) = y(t_i) + h f(t_i, y(t_i)) + h \tau_i \quad (7.38)$$

Per sottrazione dalla relazione  $\eta_{i+1} = \eta_i + h f(t_i, \eta_i)$ , che definisce il metodo di Eulero, si ha

$$y_{i+1} - \eta_{i+1} = y_i - \eta_i + h [f(t_i, y_i) - f(t_i, \eta_i)] + h \tau_i$$

da cui, dalla proprietà di lipschitzianità della  $f$

$$|y_{i+1} - \eta_{i+1}| \leq |y_i - \eta_i| (1 + |h|L) + |h| |\tau|$$

ove  $|\tau| \geq |\tau_i|$ . Con riferimento al Lemma 7.1, con  $w_j = |y_j - \eta_j|$  per  $j = 0, 1, \dots, n$  e  $\alpha = |h|L$  e  $\beta = |h| |\tau|$ , si ha, quando  $\eta_0 = y_0$

$$|y_i - \eta_i| \leq \frac{e^{i|h|L} - 1}{L} |\tau|$$

da cui il risultato richiesto, tenendo conto che  $ih = t_i - t_0$ . ■



La maggiorazione (7.36) mostra che per il metodo di Eulero l'errore globale è maggiorato dall'errore locale, moltiplicato per la quantità  $(e^{L|t_i-t_0|} - 1)/L$ , che risulta indipendente dal passo  $h$ . Questo è, in sostanza, il senso della stabilità del metodo. La maggiorazione (7.37) indica, inoltre, che, quando la soluzione  $y(t)$  ha la derivata seconda, l'errore globale tende a zero linearmente con  $h$ . In questo caso la maggiorazione potrebbe essere utilizzata per indicare il valore del passo  $h$  necessario per ottenere una approssimazione  $\eta_i$  tale che  $|y(x_i) - \eta_i| \leq \epsilon$ , con  $\epsilon$  prefissato. Tuttavia, tale possibilità è solo teorica, in quanto presuppone la conoscenza di una buona stima di  $M$ , cioè di una limitazione stretta delle funzioni  $f$ ,  $f_x$  e  $f_y$  nella regione di integrazione.

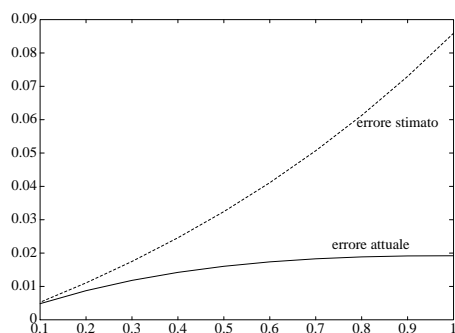
Un'altra ragione che rende, in generale, non utilizzabile da un punto di vista pratico la maggiorazione (7.37) deriva dal fatto che in alcuni casi essa non è *stretta*. Come illustrazione, si consideri il seguente problema a valori iniziali

$$y'(t) = -y(t) + t + 1, \quad 0 \leq t \leq 1, \quad y(0) = 1$$

che ha la soluzione esatta  $y(t) = t + e^{-t}$ , e quindi  $y''(t) = e^{-t}$  e  $|y''(t)| \leq e^{-0} = 1$ , per  $t \in [0, 1]$ . Essendo, inoltre,  $f_y = -1$ , la funzione  $f(t, y)$  è una funzione lipschitziana con costante  $L = 1$ . La maggiorazione (7.37) diventa pertanto

$$|y_i - \eta_i| \leq 0.5 h (e^{t_i} - 1)$$

In Figura 7.16 sono indicati sia l'errore vero che l'errore stimato mediante la maggiorazione precedente per  $h = 0.1$  e sull'intervallo  $[0, 1]$ . Mentre l'errore stimato cresce in maniera esponenziale, l'errore attuale rimane praticamente limitato. La ragione di tale comportamento è da attribuire al fatto che la derivata  $f_y$  è negativa, e come abbiamo già osservato, per un problema di questo tipo gli errori, per  $t$  che aumenta, tendono a smorzarsi.



$t_i$	attuale	stimato
0.1	0.004837	0.005258
0.2	0.008730	0.011070
0.3	0.011818	0.017492
0.4	0.014220	0.024591
0.5	0.016040	0.032436
0.6	0.017370	0.041105
0.7	0.018288	0.050687
0.8	0.018861	0.061277
0.9	0.019149	0.072980
1.	0.019201	0.085914

Figura 7.16: Errore attuale e errore stimato relativi al metodo di Eulero per il problema a valori iniziali  $y' = -y + t + 1$ ;  $y(0) = 1$ .

### Una semplice strategia di cambiamento del passo

Un aspetto importante nelle applicazioni è la possibilità di stimare la precisione ottenuta durante il calcolo. Nella pratica, infatti, è assegnata una tolleranza  $\epsilon$ , e si vuole che la soluzione numerica approssimi la soluzione del problema continuo a meno di  $\epsilon$ ; più precisamente, si vuole che  $|y_i - \eta_i| \approx \epsilon$ , nel caso che si desideri controllare l'errore assoluto e  $\approx \epsilon|y_i|$  per l'errore relativo. In realtà, dal momento che in generale non è facile stimare l'errore globale, ci si limita a controllare l'errore di troncamento locale.

Per il metodo di Eulero l'errore di troncamento locale è, come abbiamo visto (cfr. (7.32)), proporzionale a  $h^2 y''$ . Si vede, quindi, che per controllare l'errore locale mediante una opportuna scelta del passo  $h$  è necessario avere una stima della derivata seconda. Per ottenere tale stima si può, ad esempio, utilizzare il seguente algoritmo.

- (1) Supponiamo di avere calcolato  $\eta_{i-1}$ ,  $\eta_i$ , e quindi  $\eta'_{i-1} = f(t_{i-1}, \eta_{i-1})$ ,  $\eta'_i = f(t_i, \eta_i)$  e di volere calcolare la soluzione nel punto  $t_{i+1}$  con un errore locale minore di una quantità  $\epsilon$  assegnata. Una stima di  $y''$  è fornita dalla quantità

$$\eta''_{i-1,i} := \frac{\eta'_i - \eta'_{i-1}}{t_i - t_{i-1}}$$

e, quindi, si sceglierà  $h = t_{i-1} - t_i$  in maniera che

$$|\eta''_{i-1,i}| \frac{h^2}{2} \leq \epsilon \Rightarrow h \leq \sqrt{\frac{2\epsilon}{|\eta''_{i-1,i}|}}$$

Nell'applicazione pratica  $h$  è scelto, "prudentemente", minore di tale valore, ad esempio 0.9 di esso. Inoltre, si introduce un intervallo  $[h_{\min}, h_{\max}]$  nel quale è permessa la variazione del passo. In definitiva, si ha la seguente formula

$$h = \max \left[ \min \left( 0.9 \sqrt{\frac{2\epsilon}{|\eta''_{i-1,i}|}}, h_{\max} \right), h_{\min} \right]$$

Per controllare l'errore relativo si sostituisce  $\epsilon$  con  $\epsilon|\eta_i|$ .

- (2) Si calcola il valore  $\eta_{i+1}$  nel punto  $t_{i+1}$  mediante il metodo di Eulero.
- (3) Si utilizzano le informazioni, ora aggiornate, nel punto  $t_{i+1}$  per verificare se il passo utilizzato è corretto; ricordiamo, infatti, che la stima di  $y''$  ha utilizzato punti precedenti a  $t_{i+1}$ . Si calcola, quindi

$$\eta''_{i,i+1} := \frac{\eta'_{i+1} - \eta'_i}{t_{i+1} - t_i}$$

e si controlla se la quantità  $|y''_{i,i+1}| h^2/2$  è sufficientemente piccola. In caso contrario, si deve tornare indietro al punto  $t_i$  e usare un passo più piccolo.

Nella forma precedente il metodo di Eulero diventa un *metodo adattivo*, cioè un metodo in grado di scegliere il passo in base al comportamento della soluzione del problema particolare trattato. L'idea, ora vista per il metodo di Eulero, può essere opportunamente adattata, come vedremo nel seguito, ai metodi di ordine superiore, sfruttando, quando possibile, anche la possibilità di cambiare, oltre che il passo, anche l'ordine del metodo. Osserviamo, per concludere, che le strategie più comuni per il cambiamento del passo sono solo parzialmente basate su un fondamento teorico. Molto spesso esse sono una opportuna combinazione di *teoria, empirismo e intuizione*.

### 7.3.2 Influenza degli errori di arrotondamento

Quando il metodo numerico è implementato su calcolatore, anziché la successione  $\eta_i$  si ottiene una successione di numeri macchina  $\tilde{\eta}_i$ , corrispondenti alla seguente iterazione

$$\tilde{\eta}_{i+1} = \tilde{\eta}_i + hf(t_i, \tilde{\eta}_i) + \rho_i \quad i = 0, 1, \dots, n-1 \quad (7.39)$$

ove  $\rho_i$  sono gli *errori locali di arrotondamento*. Se poniamo

$$\rho(h) = \max_{0 \leq i \leq n-1} |\rho_i|$$

si ha che  $\rho(h)$  non decresce, in generale, per  $h \rightarrow 0$ . Per vedere l'effetto degli errori di arrotondamento, si sottrae la (7.39) dalla equazione (7.38) e si ottiene

$$\tilde{e}_{i+1} = \tilde{e}_i + h[f(t_i, y_i) - f(t_i, \tilde{\eta}_i)] + h\tau_i - \rho_i$$

ove  $\tilde{e}_i = y(t_i) - \tilde{\eta}_i$ . Procedendo come in Teorema 7.2, si ottiene la seguente limitazione dell'errore

$$|\tilde{e}_i| \leq e^{L|t_i-t_0|} |y_0 - \tilde{\eta}_0| + \left[ \frac{e^{L|t_i-t_0|} - 1}{L} \right] \left[ \tau(h) + \frac{\rho(h)}{h} \right] \quad (7.40)$$

Tale limitazione indica che al di sotto di un valore  $h^*$  l'errore totale può aumentare. Il valore  $h^*$  dipende, in particolare, da  $\rho(h)$ , e quindi dalla precisione macchina utilizzata. Tenendo presente che il metodo di Eulero è un metodo del primo ordine, per avere un errore locale, ad esempio, dell'ordine di  $10^{-4}$ , è necessario un passo  $h \approx 10^{-4}$ . D'altra parte, con una precisione macchina  $\text{eps} \approx 1.10^{-7}$  (semplice precisione) si ha  $\text{eps}/h \approx 1.10^{-3}$ . Si vede quindi che in semplice precisione per un passo  $h$  dell'ordine indicato gli errori di arrotondamento possono prevalere sull'errore di troncamento locale. Questa è in sostanza una delle motivazioni per introdurre i metodi di ordine superiore, con i quali è possibile ottenere la medesima precisione, utilizzando passi più grandi. È opportuno, comunque, osservare che la limitazione (7.40) si riferisce alla situazione peggiore. In pratica, gli errori di arrotondamento, variando in segno e grandezza, possono compensarsi.

### 7.3.3 Metodi di sviluppo in serie

Il metodo di Eulero corrisponde ad approssimare in ogni punto  $t_i$  la soluzione mediante la tangente, cioè mediante lo sviluppo in serie della funzione arrestato al termine di primo grado. Quando la funzione  $f(t, y)$  è sufficientemente regolare, si possono ottenere metodi di ordine superiore al primo utilizzando uno sviluppo in serie di grado superiore. Il punto di partenza di tali metodi è, pertanto, il seguente *sviluppo in serie*

$$y(t_{i+1}) = y(t_i) + h \Delta(t_i, y_i; h, f) \quad (7.41)$$

ove si è posto

$$\Delta(t, y; h, f) = y'(t) + \frac{h}{2} y''(t) + \frac{h^2}{3!} y'''(t) + \dots$$

Le derivate in tale sviluppo non sono note esplicitamente, in quanto non è nota la soluzione. Comunque, esse possono essere ottenute derivando successivamente l'equazione  $y'(t) = f(t, y(t))$ . Si ha, ad esempio

$$\begin{aligned} y'' &= f' = f_t + f_y y' = f_t + f_y f \\ y''' &= f'' = f_{tt} + f_{ty} f + f_{yt} f + f_{yy} f^2 + f_y f_t + f_y^2 f \\ &= f_{tt} + 2f_{ty} f + f_{yy} f^2 + f_t f_y + f_y^2 f \end{aligned}$$

In maniera analoga, si può esprimere ogni derivata di  $y$  in termini di  $f(t, y)$  e delle sue derivate parziali. È, comunque, evidente che, salvo per funzioni  $f$  particolari, ad esempio per le funzioni *lineari in  $y$* , le derivate di ordine superiore possono avere espressioni complicate.

Da (7.41) si può ottenere, per ogni intero  $p$  fissato, un particolare *metodo numerico, troncando* la serie ai primi  $p$  termini. Si ha quindi

$$\boxed{\eta_{i+1} = \eta_i + h \Phi(t_i, \eta_i; h, f)} \quad i = 0, 1, \dots$$

ove

$$\Phi(t, \eta; h, f) = f(t, \eta) + \frac{h}{2} f'(t, \eta) + \dots + \frac{h^{p-1}}{p!} f^{(p-1)}(t, \eta)$$

Per  $p = 1$  si riottiene il metodo di Eulero, mentre per  $p = 2$  si ha il seguente metodo del *secondo ordine*

$$\eta_{i+1} = \eta_i + h \left[ f(t_i, \eta_i) + \frac{h}{2} (f_t(t_i, \eta_i) + f_y(t_i, \eta_i) f(t_i, \eta_i)) \right]$$

► **Esempio 7.6** Consideriamo la risoluzione numerica del seguente problema

$$y' = \frac{1}{t^2} - \frac{y}{t} - y^2$$

con la condizione iniziale  $y(1) = -1$ . Si verifica facilmente che la soluzione esatta è data da  $y = -1/t$ .

$h$	$e^{(1)}$	$e^{(1)}(h_{i+1})/e^{(1)}(h_i)$	$e^{(2)}$	$e^{(2)}(h_{i+1})/e^{(2)}(h_i)$
1/16	-0.035862		0.001877	
1/32	-0.018075	0.5040	0.000463	0.246
1/64	-0.009075	0.5020	0.000115	0.248
1/128	-0.004547	0.5010	0.000028	0.249

Tabella 7.1: Metodi di sviluppo in serie, con  $p = 1$  e  $p = 2$ , per il calcolo della soluzione in  $t = 2$  del problema a valori iniziali  $y' = 1/t^2 - y/t - y^2$ , con la condizione  $y(1) = -1$ .

Nella Tabella 7.1 sono riportati i risultati ottenuti in  $t = 2$  mediante il metodo di sviluppo in serie per  $p = 1$  e  $p = 2$  e per successivi valori di  $h$ . Più precisamente, con  $e^{(1)}$ , rispettivamente  $e^{(2)}$ , è indicato l'errore globale ottenuto con il metodo corrispondente a  $p = 1$ , rispettivamente  $p = 2$ . In questo caso si ha

$$f(t, y) = \frac{1}{t^2} - \frac{y}{t} - y^2; \quad f'(t, y) = -\frac{2}{t^3} - \frac{y'}{t} + \frac{y}{t^2} - 2yy'$$

Si vede che in corrispondenza al dimezzarsi del passo  $h$  l'errore per  $p = 1$  è approssimativamente diviso per 2, mentre è diviso per 4 per  $p = 2$ . Tale comportamento dell'errore è rappresentato in Figura 7.17, che riporta in scala logaritmica i valori assoluti degli errori.

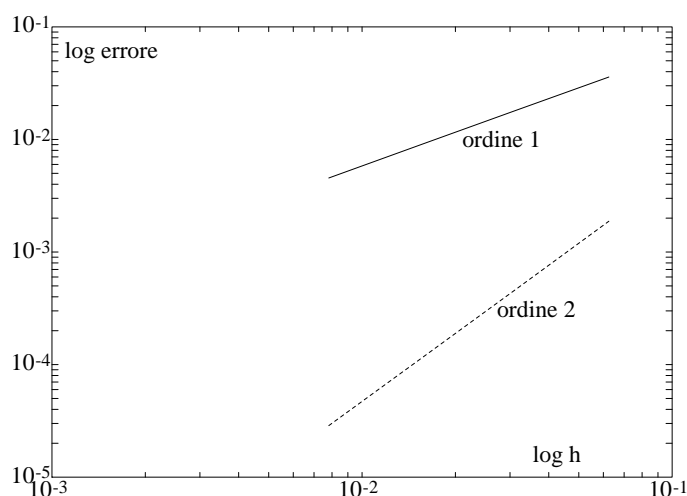


Figura 7.17: Convergenza dei metodi di sviluppo in serie per  $p = 1$  e  $p = 2$ .

◆ **Esercizio 7.3** *Mostrare che il metodo di sviluppo in serie di ordine  $p$  applicato al problema  $y' = \lambda y$ ,  $y(0) = y_0$  genera la successione  $\eta_i = \rho(\bar{h})^i y_0$ , ove  $\bar{h} = h\lambda$  e*

$$\rho(\bar{h}) = 1 + \bar{h} + \frac{\bar{h}^2}{2!} + \dots + \frac{\bar{h}^p}{p!}$$

Per  $p = 2$ , mostrare che  $|\rho(\bar{h})| < 1$  se e solo se  $-2 < \bar{h} < 0$ .

### 7.3.4 Metodi di Runge-Kutta

Come è già stato osservato, l'utilità pratica dei metodi di sviluppo in serie è limitata dalla necessità di calcolare le derivate della funzione  $f(t, y)$ . I *metodi di Runge-Kutta*<sup>12</sup> evitano tale inconveniente approssimando le derivate mediante opportuni valori della funzione  $f(t, y)$  nell'intervallo  $(t_i, t_{i+1})$ . Altri metodi, che esamineremo nel seguito, utilizzano, per approssimare le derivate di  $f$ , i valori della soluzione approssimata  $\eta_j$  in opportuni punti precedenti  $t_j$ ,  $j = i, i-1, \dots, i-r$  della reticolazione. Per tale motivi quest'ultimi vengono chiamati *metodi a più passi*, mentre i metodi ricavati dallo sviluppo in serie che abbiamo visto nel paragrafo precedente e i metodi di Runge-Kutta che esamineremo in questo paragrafo sono chiamati *metodi a un passo*. Rinviando al seguito per una discussione più appropriata, si può dire, in maniera schematica, che a parità di ordine un metodo a più passi, sfruttando meglio la *memoria*, è meno costoso di un metodo a un passo. D'altra parte, un metodo a un passo è, in generale, di più facile implementazione e richiede, diversamente dai metodi a più passi, la conoscenza solo del valore iniziale  $\eta_0$ .

Introduciamo l'idea dei metodi Runge-Kutta mediante l'analisi di un caso particolare. Riprendendo le notazioni del paragrafo precedente, poniamo

$$\Phi(t, y; h) = c_1 f(t, y) + c_2 f(t + ha_2, y + hb_{21} f(t, y))$$

ove  $c_1, c_2, a_2$  e  $b_{21}$  sono costanti da determinare in maniera che tra i seguenti sviluppi in serie

$$\Phi(t, y; h) = (c_1 + c_2)f(t, y) + hc_2[a_2 f_t(t, y) + b_{21} f_y(t, y)f(t, y)] + O(h^2)$$

$$\Delta(t, y; h) = f(t, y) + \frac{1}{2}h[f_t(t, y) + f_y(t, y)f(t, y)] + O(h^2)$$

si abbia la coincidenza del massimo numero di coefficienti di potenze di  $h$ . Si ottiene facilmente il sistema di equazioni

$$c_1 + c_2 = 1, \quad c_2 a_2 = \frac{1}{2}, \quad c_2 b_{21} = \frac{1}{2}$$

per il quale, indicato con  $\alpha$  un parametro  $\neq 0$ , si ha il seguente *insieme* di soluzioni

$$c_1 = 1 - \alpha, \quad c_2 = \alpha, \quad a_2 = b_{21} = \frac{1}{2\alpha}$$

e la corrispondente *famiglia di metodi*

$$\Phi(t, y; h) = (1 - \alpha)f(t, y) + \alpha f\left(t + \frac{h}{2\alpha}, y + \frac{h}{2\alpha} f(t, y)\right)$$

<sup>12</sup>introdotti da Runge nel 1895 e successivamente sviluppati da Heun (1900) e da Kutta (1901).

In particolare, per  $\alpha = 1/2$  si ottiene il seguente metodo, noto anche come *metodo di Heun*

$$\eta_{i+1} = \eta_i + \frac{h}{2} [f(t_i, \eta_i) + f(t_i + h, \eta_i + hf(t_i, \eta_i))]$$

e per  $\alpha = 1$  il *metodo di Eulero modificato*

$$\eta_{i+1} = \eta_i + h [f(t_i + \frac{1}{2}h, \eta_i + \frac{1}{2}hf(t_i, \eta_i))]$$

Ambedue i metodi utilizzano due valutazioni della funzione  $f$ . Come esemplificazione, il metodo di Heun è illustrato in Figura 7.18.

```

Y=Y0
T=T0
DO 10 I=1,N
  YP1=F(T,Y)
  T=T+H
  YS=Y+H*YP1
  YP2=F(T,YS)
  Y=Y+0.5*H*(YP1+YP2)
10 CONTINUE

```

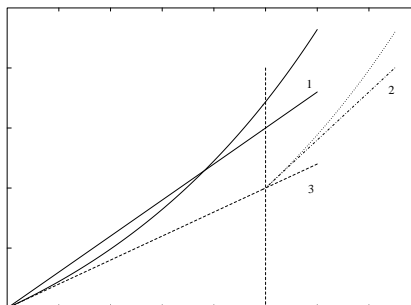


Figura 7.18: Illustrazione del metodo di Heun. La retta (1) ha coefficiente angolare dato dalla media tra il coefficiente angolare della tangente nel punto  $(t_i, \eta_i)$ , rappresentata dalla retta (3) e quello della retta (2), che rappresenta la tangente alla curva soluzione del problema a valori iniziali assegnato passante per il punto  $(t_{i+1}, \eta_i + hf(t_i, \eta_i))$ .

Il procedimento sviluppato in precedenza in un caso particolare può essere esteso in modo da utilizzare un qualsiasi numero  $m$  di valutazioni della funzione  $f$ . In questo modo si ottiene il seguente metodo generale di Runge–Kutta *esplicito* a  $m$ -stadi

$$\begin{aligned} \eta_{i+1} &= \eta_i + h \Phi(t_i, \eta_i; h) \\ \Phi(t, \eta; h) &:= \sum_{r=1}^m c_r k_r, \quad k_1 = f(t, y) \\ k_r &= f(t + ha_r, y + h \sum_{s=1}^{r-1} b_{rs} k_s), \quad r = 2, \dots, m \end{aligned} \quad (7.42)$$

Un modo usuale di rappresentare il metodo (7.42) è sotto forma di *tableau*, noto anche come *Butcher array* e illustrato in Tabella 7.2.

Esempi corrispondenti ad alcuni valori di  $m$  sono indicati in Tabella 7.3; in particolare, tra i metodi corrispondenti a  $m = 4$  si trova uno dei metodi più noti per

0					
$a_2$	$b_{21}$				
$a_3$	$b_{31}$	$b_{32}$			
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
$a_m$	$b_{m1}$	$b_{m2}$	$\cdots$	$b_{m\ m-1}$	
	$c_1$	$c_2$	$\cdots$	$c_{m-1}$	$c_m$

Tabella 7.2: Tableau relativo a un generico metodo Runge–Kutta esplicito.

approssimare la soluzione dei problemi a valori iniziali, definito dalle seguenti formule

$$\eta_{i+1} = \eta_i + \frac{h}{6} (k_1 + 2k_2 + 2k_3 + k_4)$$

$$k_1 = f(t_i, \eta_i), \quad k_2 = f(t_i + \frac{1}{2}h, \eta_i + \frac{1}{2}hk_1)$$

$$k_3 = f(t_i + \frac{1}{2}h, \eta_i + \frac{1}{2}hk_2), \quad k_4 = f(t_i + h, \eta_i + hk_3)$$

Tale metodo è di ordine 4, ossia, se la funzione  $f$  è sufficientemente regolare, per il corrispondente errore locale si ha  $\tau(x, y) = O(h^4)$ . Il risultato è a priori prevedibile, dal momento che nel caso particolare in cui la funzione  $f$  è indipendente dalla variabile  $y$ , la formula si riduce alla formula di integrazione di Simpson.

RK ordine 2	RK ordine 4	Heun ordine 3
0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
$\frac{1}{2}$	0	$\frac{1}{3}$
1	0	0
0	$\frac{1}{6}$	$\frac{1}{4}$
1	$\frac{1}{3}$	0
	$\frac{1}{3}$	$\frac{2}{3}$
	$\frac{1}{6}$	$\frac{3}{4}$

Tabella 7.3: Esempi di metodi di Runge–Kutta espliciti.

► **Esempio 7.7** Nella Tabella 7.4 sono messi a confronto, a parità di costo, i risultati ottenuti mediante alcuni metodi di Runge–Kutta per il seguente problema

$$y' = -y + 1; \quad x \in [0, 1]; \quad y(0) = 0$$

che ha come soluzione  $y = 1 - e^{-x}$ . I risultati chiariscono in maniera evidente il significato dell'ordine di un metodo. ■



	metodo Eulero	metodo Eulero modificato	metodo Runge–Kutta 4	valore esatto
t	h=0.025	h=.05	h=.1	
.1	<u>.09631</u>	<u>.09512</u>	<u>.0951625</u>	.095162582
.2	<u>.18334</u>	<u>.18119</u>	<u>.1812691</u>	.181269247
.3	<u>.26200</u>	<u>.25908</u>	<u>.2591815</u>	.259181779
.4	<u>.33307</u>	<u>.32956</u>	<u>.3296797</u>	.329679954
.5	<u>.39731</u>	<u>.39333</u>	<u>.3934690</u>	.393469340

Tabella 7.4: Confronto tra metodi di Runge–Kutta di differente ordine e a parità di costo.

### Metodi di Runge–Kutta impliciti

I metodi considerati nel paragrafo precedente sono di tipo *esplicito*, in quanto il calcolo di  $\eta_{i+1}$  richiede la valutazione della funzione  $\Phi$ , che dipende solo dai valori già calcolati della soluzione approssimata. Per risolvere problemi particolari (cfr. il successivo Esempio 7.9), hanno interesse metodi, analoghi nella sostanza ai precedenti, ma nei quali la funzione  $\Phi$  dipende anche dal valore  $\eta_{i+1}$ . Essi sono detti metodi *impliciti*, in quanto richiedono per il calcolo di  $\eta_{i+1}$  la risoluzione di equazioni, che sono non lineari quando la funzione  $f(t, y)$  è non lineare in  $y$ .

Un *metodo Runge-Kutta a m-stadi implicito* è definito, in maniera generale, nel seguente modo, a cui corrisponde il tableau illustrato in Tabella 7.5

$$\begin{aligned} \eta_{i+1} &= \eta_i + h\Phi(t_i, \eta_i; h) \\ \Phi(t, \eta; h) &:= \sum_{r=1}^m c_r k_r \\ k_r &= f\left(t + ha_r, y + h \sum_{s=1}^m b_{rs} k_s\right), \quad r = 1, \dots, m \end{aligned} \tag{7.43}$$

$a_1$	$b_{11}$	$b_{12}$	$\cdots$	$b_{1\,m-1}$	$b_{1m}$
$a_2$	$b_{21}$	$b_{22}$	$\cdots$	$b_{2\,m-1}$	$b_{2m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_m$	$b_{m1}$	$b_{m2}$	$\cdots$	$b_{m\,m-1}$	$b_{mm}$
	$c_1$	$c_2$	$\cdots$	$c_{m-1}$	$c_m$

Tabella 7.5: Tableau relativo a un generico metodo Runge–Kutta implicito.

Quando  $b_{is} = 0$  per  $s = i + 1, \dots, m$ , il metodo corrispondente è chiamato *semi-*

*esplicito*, in quanto il calcolo delle quantità  $k_i$  è ridotto al calcolo di *single* equazioni non lineari, anziché di un sistema completo.

► **Esempio 7.8** Come esemplificazione di un metodo semi-esplicito, consideriamo il seguente metodo

$$\begin{aligned}\eta_{i+1} &= \eta_i + \frac{h}{6}(k_1 + 4k_2 + k_3) \\ k_1 &= f(t_i, \eta_i) \\ k_2 &= f\left(t_i + \frac{h}{2}, \eta_i + \frac{h}{4}k_1 + \frac{h}{4}k_2\right) \\ k_3 &= f(t_i + h, \eta_i + hk_2)\end{aligned}\tag{7.44}$$

Si tratta di un metodo del *quarto ordine a 3-stadi*. Per il calcolo del valore  $\eta_{i+1}$  è necessario risolvere l'equazione (7.44) nell'incognita  $k_2$ . L'esistenza di una soluzione di tale equazione può essere dimostrata mediante il teorema delle contrazioni analizzato nel Capitolo 5. Infatti, la funzione a secondo membro di (7.44) è una funzione lipschitziana di costante  $Lh/4$ , ove  $L$  è la costante di Lipschitz di  $f$  rispetto a  $y$ . Pertanto, tale funzione è una contrazione quando

$$h < \frac{4}{L}\tag{7.45}$$

cioè per  $h$  sufficientemente piccolo. Osserviamo, tuttavia, che, essendo il teorema delle contrazioni una condizione sufficiente, la soluzione può esistere anche per valori del passo  $h$  che non verificano la condizione precedente. Per il calcolo numerico di  $k_2$  si potrebbe utilizzare il metodo delle iterazioni successive, la cui convergenza è pure assicurata dalla condizione (7.45). Tale condizione, tuttavia, può essere eccessivamente restrittiva sul passo  $h$  quando la costante  $L$  è grande, come appunto avviene per i problemi stiff. Come vedremo più estesamente nel successivo Esempio 7.9, diventano allora più convenienti altri metodi, quali ad esempio il metodo di Newton o sue opportune varianti. ■

### 7.3.5 Metodo di Eulero implicito e formula dei trapezi

Il metodo di Eulero implicito e la formula, o metodo, dei trapezi sono due *metodi impliciti* particolarmente interessanti per la risoluzione dei problemi stiff. Essi possono essere introdotti nel seguente modo. Incominciamo ad osservare che integrando tra  $t_i$  e  $t_{i+1}$  l'equazione differenziale  $y' = f(t, y)$  si ottiene la seguente identità

$$y(t_{i+1}) = y_i + \int_{t_i}^{t_{i+1}} f(t, y(t)) dt\tag{7.46}$$

Il metodo di Eulero considerato nei paragrafi precedenti, e che nel seguito chiameremo metodo di Eulero esplicito, è ottenuto approssimando l'integrale nell'identità (7.46) con il valore  $(t_{i+1} - t_i) f(t_i, y(t_i))$ , che corrisponde all'applicazione della formula di quadratura dei rettangoli con nodo nel punto  $t_i$ . Metodi differenti possono allora essere ottenuti scegliendo in maniera diversa il nodo da utilizzare nella formula di quadratura (cfr. Figura 7.19). In particolare, se come nodo si sceglie il punto

$t_{i+1}$  si ottiene il seguente metodo numerico

$$\begin{cases} \eta_0 = y_0 \\ \eta_{i+1} = \eta_i + hf(t_{i+1}, \eta_{i+1}) \end{cases}$$

detto *metodo di Eulero implicito*, o metodo di Eulero all'indietro. Lasciamo come esercizio la sua interpretazione geometrica in termini di approssimazione locale della soluzione mediante la tangente. Quando l'integrale in (7.46) è approssimato

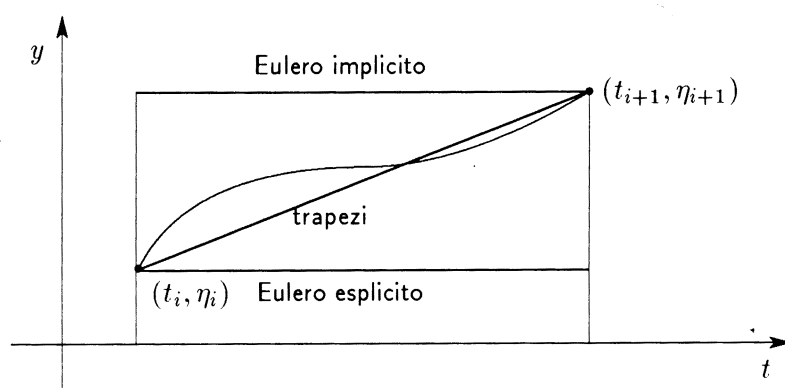


Figura 7.19: Metodi di Eulero esplicito e implicito e formula dei trapezi come particolari formule di quadratura.

mediante la formula di quadratura dei trapezi, si ottiene il seguente metodo

$$\begin{cases} \eta_0 = y_0 \\ \eta_{i+1} = \eta_i + \frac{h}{2} [f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1})] \end{cases}$$

detto *metodo dei trapezi*, e che corrisponde ad assumere come approssimazione della derivata la media delle derivate nei punti  $(t_i, \eta_i)$  e  $(t_{i+1}, \eta_{i+1})$ . Più in generale, si potrebbe assumere come approssimazione una media pesata, cioè

$$\eta_{i+1} = \eta_i + h [(1 - \theta)f(t_i, \eta_i) + \theta f(t_{i+1}, \eta_{i+1})]$$

con  $\theta \in [0, 1]$ . Il metodo corrispondente è anche noto come  $\theta$ -metodo. Per  $\theta = 0$ ,  $\theta = 1$ ,  $\theta = 1/2$  si riottengono rispettivamente il metodo di Eulero esplicito, implicito e il metodo del trapezio (noto anche, in particolare nell'ambito delle equazioni alle derivate parziali, come *metodo di Crank-Nicolson*).

Per costruzione, e come si può verificare facilmente per sviluppo in serie, il metodo di Eulero implicito è un metodo del *primo ordine* (come pure ogni elemento della famiglia  $\theta$ , con  $\theta \neq 1/2$ ), mentre la formula dei trapezi è un metodo del *secondo*

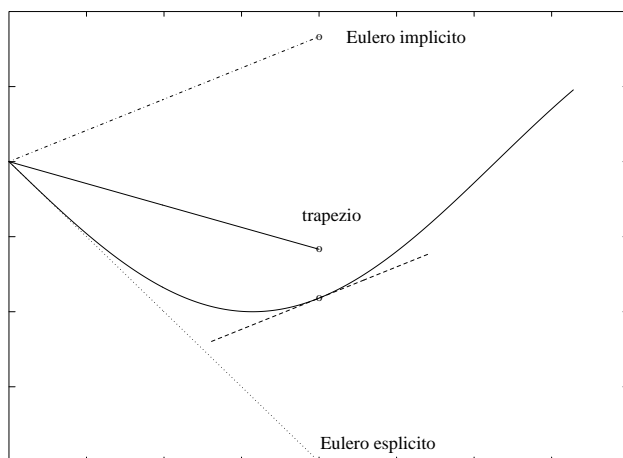


Figura 7.20: Illustrazione grafica dei metodi di Eulero, di Eulero implicito e del trapezio.

*ordine*. Essendo impliciti, richiedono la risoluzione ad ogni passo di una equazione non lineare. Vedremo, ora, su un esempio la loro utilità pratica.

► **Esempio 7.9** (*Un esempio di equazione stiff*) Consideriamo il seguente problema a valori iniziali

$$y'(t) = \lambda(y - \sin t) + \cos t, \quad y(0) = 1 \quad (7.47)$$

che per ogni  $\lambda$ , reale o complesso, ha la soluzione  $y(t) = e^{\lambda t} + \sin t$ . In particolare, per  $\lambda \in \mathbb{R}$ , con  $\lambda \ll -1$  si ha che per  $t$  vicino allo zero la  $y$  decresce rapidamente, per assumere nel seguito il comportamento della funzione  $\sin t$ . Ad una *fase transiente* molto breve segue, quindi, una fase in cui la soluzione *varia lentamente* (cfr. Figura 7.21). Supponiamo, come è richiesto in diverse applicazioni, di essere interessati sia alla fase transiente, che alla fase stazionaria.

Usando un metodo numerico a passo variabile, il passo di discretizzazione necessario per ottenere una accuratezza assegnata dovrebbe poter aumentare passando dalla fase transiente alla fase stazionaria. Nell'esempio che stiamo considerando, quando il termine  $e^{\lambda t}$  è diventato trascurabile rispetto a  $\sin t$ , il passo richiesto dal metodo dovrebbe essere quello per risolvere l'equazione differenziale  $y'(t) = \cos t$ .

Vediamo, allora, come si comportano su questo esempio i metodi di Eulero esplicito, Eulero implicito e la formula dei trapezi. In pratica, è sufficiente esaminare tale comportamento relativamente alla seguente equazione differenziale

$$y' = \lambda y$$

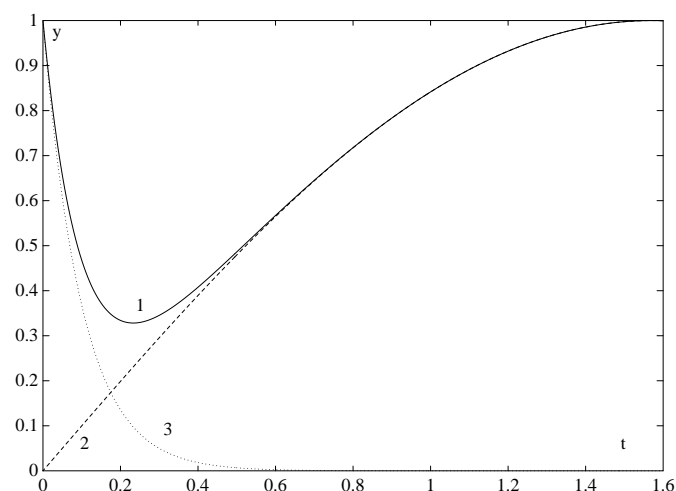


Figura 7.21: (1) soluzione  $y = \exp(-10t) + \sin t$  del problema a valori iniziali  $y' = -10(y - \sin t) + \cos t$ ,  $y(0) = 1$ ; (2)  $\sin t$ ; (3) componente transiente  $\exp(-10t)$ .

che rappresenta il termine transiente. Si ha

$$\begin{array}{ll} \text{Eulero esplicito} & \eta_{i+1} = \eta_i + h(\lambda\eta_i) \Rightarrow \eta_i = (1 + h\lambda)^i \eta_0 \\ \text{Eulero implicito} & \eta_{i+1} = \eta_i + h(\lambda\eta_{i+1}) \Rightarrow \eta_i = \frac{1}{(1 - h\lambda)^i} \eta_0 \\ \text{trapezio} & \eta_{i+1} = \eta_i + \frac{h}{2}(\lambda\eta_i + \lambda\eta_{i+1}) \Rightarrow \eta_i = \left(\frac{1 + (h/2)\lambda}{1 - (h/2)\lambda}\right)^i \eta_0 \end{array}$$

Per  $\lambda < 0$  si vede che la soluzione discreta fornita dal metodo di Eulero esplicito tende a zero, per  $i$  che tende all'infinito, soltanto se è verificata la seguente condizione

$$|1 + h\lambda| < 1 \iff -2 < h\lambda < 0 \Rightarrow h < \frac{2}{|\lambda|} \quad (7.48)$$

L'intervallo  $(-2, 0)$  dell'asse reale viene chiamato intervallo di *assoluta stabilità* del metodo.

Al contrario, le soluzioni fornite dal metodo di Eulero implicito e dalla formula del trapezio tendono a zero, qualunque sia il valore del prodotto  $h\lambda$ , e, quindi, il corrispondente intervallo reale di assoluta stabilità è tutto il semiasse  $(-\infty, 0)$ . In questo caso si dice che i metodi sono *A-stabili*.

Per  $\lambda \in \mathbb{C}$  la zona di assoluta stabilità dei metodi di Eulero implicito e della formula del trapezio contiene tutto il semipiano con  $\Re(\lambda h) \leq 0$ , mentre quella del metodo di Eulero esplicito è data dal cerchio  $|z + 1| \leq 1$  (cfr. Figura 7.22, nella quale è rappresentata anche la zona di stabilità del metodo di Runge-Kutta del quarto ordine).

La *conseguenza numerica* delle proprietà ora evidenziate è la possibilità, per i metodi di Eulero implicito e la formula del trapezio, di scegliere l'ampiezza del passo  $h$  solo in base alla accuratezza richiesta, in particolare più grande nella zona stazionaria. Viceversa, per il metodo di Eulero esplicito la condizione (7.48) deve essere rispettata su tutto l'intervallo

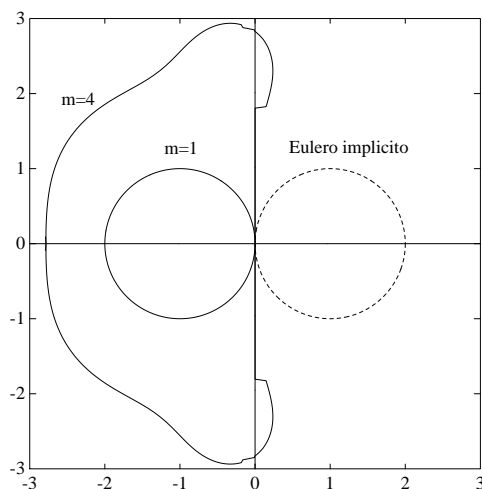


Figura 7.22: Zone di stabilità nel piano complesso relative ai metodi di Eulero esplicito ed implicito e al metodo di Runge–Kutta del quarto ordine.

di integrazione; in caso contrario, come abbiamo visto, la componente transiente numerica prevale facendo tendere all'infinito la soluzione. In questo caso, nella zona stazionaria non è solo l'accuratezza ad indicare l'ampiezza dell'intervallo, ma anche la stabilità.

Rimane, comunque, da considerare, per i metodi di Eulero implicito e della formula dei trapezi, l'aspetto importante del calcolo numerico, ad ogni passo, della soluzione  $\eta_{i+1}$ . Tale è, in sostanza, il prezzo da pagare per la maggiore stabilità. Come vedremo, il problema va affrontato adeguatamente, se non si vuole perdere il vantaggio ottenuto nella stabilità.

Analizzeremo, ad esempio, la formula del trapezio; considerazioni analoghe si hanno per il metodo di Eulero implicito. Per  $h$  e  $i$  fissati, si cerca  $\eta_{i+1}$  che verifica la seguente equazione

$$\eta_{i+1} = \eta_i + \frac{h}{2} [f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1})] \quad (7.49)$$

Applicando il teorema delle contrazioni, si vede che una condizione sufficiente per l'esistenza (e unicità) della soluzione dell'equazione (7.49) è la seguente

$$\frac{hL}{2} < 1 \Rightarrow h < \frac{2}{L} \quad (7.50)$$

ove  $L$  è la costante di Lipschitz della funzione  $f$ , e quindi per il problema (7.47)  $L = |\lambda|$ . La condizione (7.50) assicura, anche, la convergenza del *metodo delle iterazioni successive*

$$\eta_{i+1}^{(r+1)} = \eta_i + \frac{h}{2} [f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1}^{(r)})]$$

ove  $\eta_{i+1}^{(0)}$  è un valore scelto opportunamente. Ad esempio, come  $\eta_{i+1}^{(0)}$  si può assumere il valore  $\eta_{i+1}$  fornito dal metodo di Eulero esplicito. Si ha, allora, che  $\eta_{i+1}^{(1)}$  è lo stesso valore fornito dal metodo di Heun.

La procedura, ora descritta su un esempio, e consistente nella applicazione del metodo delle iterazioni successive, a partire da un valore iniziale calcolato mediante un opportuno

metodo esplicito, è nota come tecnica del *predictor-corrector* (ove predictor è il metodo esplicito e corrector il metodo implicito) ed è, come vedremo successivamente, una tecnica usuale nell'applicazione dei metodi a più passi.

Tuttavia, a causa della condizione (7.50), la tecnica del predictor-corrector non è conveniente per le equazioni del tipo (7.47), con  $\lambda \ll -1$ . Si avrebbe, infatti, una restrizione sul passo analoga a quella che abbiamo visto per i metodi espliciti, che diventerebbero, quindi, competitivi. È, allora, necessario utilizzare, per la risoluzione dell'equazione (7.49), altri procedimenti. Analizziamo, ad esempio, l'applicazione del *metodo di Newton*. Indicando con  $x$  l'incognita  $\eta_{i+1}$  e con  $c$  le quantità note, si ha da risolvere la seguente equazione

$$F(x) := x - \frac{h}{2} f(x) + c = 0$$

Supponendo  $f$  derivabile rispetto a  $y$ , si ha  $F'(x) = 1 - (h/2)f_y(x)$ . Nel caso del problema (7.47 si ha  $f_y = \lambda$ , e, quindi, per  $\lambda < 0$  la derivata  $F'(x)$  è diversa dallo zero. Il metodo di Newton è, allora, *convergente*.

L'equazione (7.47) è un modello semplificato di equazioni nelle quali la soluzione durante l'intervallo di integrazione raggiunge *rapidamente* stati stazionari, e più in generale di sistemi di equazioni, nei quali sono presenti soluzioni con scale di tempi molto diverse tra loro. Equazioni con queste caratteristiche sono indicate come *equazioni stiff* e saranno considerate più in dettaglio nel seguito. ■

### 7.3.6 Metodi di Runge-Kutta-Fehlberg

Come abbiamo già osservato a proposito del metodo di Eulero, nell'applicazione pratica dei metodi numerici è importante dare una *stima* dell'errore locale per permettere una scelta adattiva del passo. Tale stima, quando si applicano i metodi di Runge-Kutta, può essere ottenuta utilizzando differenti procedimenti, che hanno in comune l'idea di calcolare la soluzione in  $t_{i+1}$  con una differente precisione. Un procedimento classico consiste nel calcolare la soluzione in  $t_{i+1}$ , a partire da  $t_i$ , una volta con passo  $h = t_{i+1} - t_i$  e successivamente con passo  $h/2$ . Dal confronto dei due valori e sfruttando opportunamente la conoscenza dell'ordine del metodo, è possibile ricavare una stima dell'errore locale. Il procedimento è, in sostanza, il *procedimento di estrapolazione* che abbiamo esaminato per le formule di quadratura.

Un modo alternativo per ottenere due valori della soluzione discreta con precisione differente consiste nell'applicazione di due metodi di *ordine differente*, ad esempio un metodo di terzo e quarto ordine, oppure di quarto e quinto ordine. A tale scopo i metodi di Runge-Kutta, nella forma tradizionale che abbiamo visto in precedenza, non sono convenienti, in quanto richiedono un numero eccessivo di valutazioni della funzione  $f$ . Ad esempio, un metodo del quarto ordine accoppiato ad un metodo del quinto ordine richiederebbe dieci valutazioni della  $f$ . Risultano, invece, di particolare utilità i cosiddetti *metodi di Runge-Kutta-Fehlberg*<sup>13</sup>, che ora illustreremo brevemente.

<sup>13</sup>E. Fehlberg, *Classical fifth-, sixth-, seventh-, and eighth order Runge-Kutta formulas with step size control*; NASA Technical Report 287 (1968).

r	$a_r$	$b_{rs}$				
		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
2	1/4	1/4				
3	3/8	3/32	9/32			
4	12/13	1932/2197	-7200/2197	7296/2197		
5	1	439/216	-8	3680/513	-845/4104	
6	1/2	-8/27	2	-3544/2565	1859/4104	-11/40

Tabella 7.6: Coefficienti  $a_r$  e  $b_{rs}$  per il metodo RKF45.

Per semplicità di presentazione considereremo una particolare coppia di formule, ad esempio le formule di Runge-Kutta-Fehlberg (RKF) di ordine 4 e 5. Con le usuali notazioni poniamo

$$k_1 = f(t_i, \eta_i), \quad k_r = f(t_i + ha_r, y + h \sum_{s=1}^{r-1} b_{rs} k_s), \quad r = 2, 3, \dots, 6$$

Le formule RKF del quarto e quinto ordine sono, rispettivamente

$$\eta_{i+1} = \eta_i + h \sum_{r=1}^5 c_r k_r$$

$$\hat{\eta}_{i+1} = \eta_i + h \sum_{r=1}^6 \hat{c}_r k_r$$

L'applicazione delle due formule richiede ad ogni passo 6 valutazioni della funzione  $f$ . I valori dei parametri delle formule sono riportati nelle Tabelle 7.6, 7.7.

r	1	2	3	4	5	6
$c_r$	25/216	0	1408/2565	2197/4104	-1/5	
$\hat{c}_r$	16/135	0	6656/12825	28561/56430	-9/50	2/55

Tabella 7.7: Coefficienti  $c_r$  e  $\hat{c}_r$  per il metodo RKF45.

Diamo ora un'idea di come le formule precedenti possono essere utilizzate per stimare l'errore locale. Supponiamo, più in generale, che siano disponibili due metodi discreti per l'approssimazione della soluzione  $y(t)$  di un problema a valori iniziali, il primo dei quali

$$\eta_0 = y_0; \quad \eta_{i+1} = \eta_i + h_i \Phi(t_i, \eta_i; h)$$

abbia errore locale  $\tau_i$  di ordine  $O(h^p)$ , mentre il secondo

$$\hat{\eta}_0 = y_0; \quad \hat{\eta}_{i+1} = \hat{\eta}_i + h_i \hat{\Phi}(t_i, \hat{\eta}_i; h)$$



abbia errore locale  $\hat{\tau}_i$  di ordine  $O(h^{p+1})$ . Supponendo  $\eta_i = \hat{\eta}_i = y(t_i)$  (tale ipotesi corrisponde a considerare l'errore locale), si ha

$$\begin{aligned} y(t_{i+1}) - \eta_{i+1} &= y(t_{i+1}) - \eta_i - h\Phi(t_i, \eta_i, h) \\ &= y(t_{i+1}) - y(t_i) - h\Phi(t_i, y_i, h) = h\tau_i \end{aligned}$$

Quindi

$$\tau_i = \frac{1}{h}[y(t_{i+1}) - \eta_{i+1}] = \frac{1}{h}[y(t_{i+1}) - \hat{\eta}_{i+1}] + \frac{1}{h}[\hat{\eta}_{i+1} - \eta_{i+1}] = \hat{\tau}_i + \frac{1}{h}[\hat{\eta}_{i+1} - \eta_{i+1}]$$

Ma  $\tau_i$  è di ordine  $O(h^p)$ , mentre  $\hat{\tau}_i$  è di ordine  $O(h^{p+1})$  e quindi la “parte più significativa” di  $\tau_i$  deve essere attribuita al termine  $(\hat{\eta}_{i+1} - \eta_{i+1})/h$ . Di conseguenza si può assumere come *stima* dell'errore  $\tau_i$  la quantità

$$\boxed{\tau_i \approx \frac{1}{h}(\hat{\eta}_{i+1} - \eta_{i+1})} \quad (7.51)$$

La stima ora ottenuta può essere utilizzata “in pratica” nel seguente modo. Poiché  $\tau_i$  è di ordine  $p$ , esiste una costante  $k$  tale che  $\tau_i \approx kh^p$ . Da (7.51) si ha

$$kh^p \approx \frac{1}{h}(\hat{\eta}_{i+1} - \eta_{i+1})$$

Cerchiamo, ora, un valore  $q > 0$  tale che, per una *tolleranza* prefissata  $\epsilon$ , si abbia

$$|\tau_i(qh)| \leq \epsilon$$

Poiché

$$\tau_i(qh) \approx k(qh)^p = q^p(kh^p) \approx \frac{q^p}{h}(\hat{\eta}_{i+1} - \eta_{i+1})$$

si ha la limitazione

$$q \leq \left( \frac{\epsilon h}{|\hat{\eta}_{i+1} - \eta_{i+1}|} \right)^{1/p} \quad (7.52)$$

La stima di  $q$  così determinata al passo  $i$ -mo è utilizzata per i seguenti due scopi

1. per scartare, se necessario, la scelta iniziale di  $h$  al passo  $i$ -mo e ripetere il calcolo usando  $qh$ ;
2. per predire un valore appropriato per la scelta iniziale di  $h$  al passo  $(i+1)$ -mo.

Un'esempio di implementazione del metodo è data nell'algoritmo seguente, ove si è modificata leggermente la formula (7.52) per rendere meno “conservativa” l'indicazione fornita; inoltre, per evitare una eccessiva frequenza nel cambiamento del passo si sono introdotti dei limiti su  $q$ .

**Algoritmo 7.2** (Metodo adattivo RKF45) *Si approssima la soluzione del problema a valori iniziali  $y'(t) = f(t, y)$ ;  $y(t_0) = y_0$ . Parametri di input: gli estremi dell'intervallo di integrazione  $[t_0, T]$ ; condizione iniziale  $y_0$ ; tolleranza (per l'errore locale assoluto)  $TOL$ ; massimo  $h_{max}$  e minimo  $h_{min}$ . Come output si ha  $t, \eta, h$ , cioè il valore approssimato in successivi valori di  $t$ , insieme al valore del passo  $h$  utilizzato oppure il messaggio che il programma richiede un passo più piccolo del valore  $h_{min}$ .*

```

Set  $t = t_0$ ;  $\eta = y_0$ ;  $h = h_{max}$ 
output( $t, \eta$ )
while ( $t \leq T$ ) do
  Set  $k_1 = hf(t, y)$ 
     $k_2 = hf(t + \frac{1}{4}h, \eta + \frac{1}{4}k_1)$ 
     $k_3 = hf(t + 3/8h, \eta + 3/32k_1 + 9/32k_2)$ 
     $k_4 = hf(t + 12/13h, \eta + 1932/2197k_1 - 7200/2197k_2 + 7296/2197)k_3$ 
     $k_5 = hf(t + h, \eta + 439/216k_1 - 8k_2 + 3680/513k_3 - 845/4104k_4)$ 
     $k_6 = hf(t + 1/2h, \eta - 8/27k_1 + 2k_2 - 3544/2565k_3 + 1859/4104k_4$ 
       $- 11/40k_5)$ 
  Set  $R = |1/360k_1 - 128/4275k_3 - 2197/75240k_4 + 1/50k_5 + 2/55k_6|/h$ 
    ( $R = |\hat{\eta}_{i+1} - \eta_{i+1}|/h$ )
  Set  $\delta = 0.84(TOL/R)^{1/4}$ 
  If  $R \leq TOL$  then
    Set  $t = t + h$  (approssimazione accettata)
     $\eta = \eta + 25/216k_1 + 1408/2565k_3 + 2197/4104k_4 - 1/5k_5$ 
    output( $t, \eta, h$ )
  end if
  If  $\delta \leq 0.1$  then set  $h = 0.1h$ 
    else if  $\delta \geq 4$  then set  $h = 4h$ 
      else set  $h = \delta h$ 
    (calcolo del nuovo h)
  end if
  If  $h > h_{max}$  then set  $h = h_{max}$ 
  If  $h < h_{min}$  then
    output( $h < h_{min}$ )
    stop
  end if
end do
stop

```

► **Esempio 7.10** Anche come termine di confronto per altri algoritmi, riportiamo i risultati ottenuti con l'algoritmo precedente per il seguente problema a valori iniziali

$$y' = \frac{y}{4} \left(1 - \frac{y}{20}\right); \quad y(0) = 1$$

che ha come soluzione esatta  $y = 20/(1 + 19e^{-(t/4)})$ . In Tabella 7.8 è riportato il numero  $nf$  di valutazioni della funzione  $f$  in corrispondenza a diversi valori della tolleranza  $TOL$  e a differenti valori dell'estremo di integrazione  $T$ .

Si lascia come esercizio il confronto, ad esempio, con il metodo RK di ordine 4 a passo costante. ■

T	TOL=1. E-6		TOL=1. E-9	
	nf	Errore	nf	Errore
5	42	-5.12 E-7	144	-6.972 E-10
10	72	-1.01 E-6	240	-1.210 E-9
20	114	3.60 E-7	384	-8.55 E-10

Tabella 7.8: Numero di valutazioni della funzione f nel metodo RKF45.

▼ **Osservazione 7.2** *Le tecniche di cambiamento del passo e/o dell'ordine dei metodi nell'approssimazione numerica, oltre che fornire una strategia ottimale dal punto di vista del costo, hanno anche interesse come strumento di indagine per scoprire eventuali punti di singolarità della soluzione o delle derivate. L'esistenza di tali punti, infatti, è segnalata dalla richiesta di passi "eccessivamente" piccoli.* ■

◆ **Esercizio 7.4** *Applicare il metodo di Heun e di Eulero modificato al problema a valori iniziali  $y' = t/y$ ,  $y(0) = 1$  su  $(0, 2)$ , usando  $h = 10^{-k}$ ,  $k = 1, 2, 3$ . Confrontare con la soluzione esatta.*

◆ **Esercizio 7.5** *Scrivere il seguente problema a valori iniziali*

$$3y''' + 4ty'' + \sin y = f(t), \quad y(t_0) = y_0, \quad y'(t_0) = y'_0, \quad y''(t_0) = y''_0$$

come problema a valori iniziali per un sistema di equazioni differenziali del primo ordine. Applicare il metodo di Eulero esplicito per l'approssimazione della soluzione.

◆ **Esercizio 7.6** *Applicare il metodo di Runge-Kutta del quarto ordine al seguente sistema di tipo predatore-preda*

$$\begin{cases} x' = x(1 - y), & x(0) = 5 \\ y' = y(0.75x - 1.5), & y(0) = 2 \end{cases}$$

per  $t \in (0, 1)$ .

◆ **Esercizio 7.7** *Analizzare il seguente metodo a un passo*

$$\begin{aligned} \eta_{i+1/2} - \frac{1}{2}\eta_{i+1} &= \frac{1}{2}\eta_i + \frac{h}{8}(f_i - f_{i+1}) \\ \eta_{i+1} &= \eta_i + \frac{h}{6}(f_i + 4f_{i+1/2} + f_{i+1}) \end{aligned}$$

◆ **Esercizio 7.8** *Analizzare l'ordine del seguente metodo di Runge-Kutta*

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

### 7.3.7 Metodi a più passi

In questo paragrafo considereremo, in particolare, i metodi a più passi *lineari* (linear multistep methods, brevemente LMM), definiti per  $r$  intero  $\geq 1$ , dalla formula

$$\sum_{j=0}^r \alpha_j \eta_{i+j} = h \sum_{j=0}^r \beta_j f(t_{i+j}, \eta_{i+j}) \quad (7.53)$$

per  $h = (T - t_0)/n$ , e  $t_i = t_0 + ih$ ,  $i = 0, 1, \dots, n - r$ . I coefficienti  $\alpha_j$  e  $\beta_j$  sono costanti fissate, con  $\alpha_r \neq 0$ . Più precisamente, il metodo definito dalla formula (7.53) è detto un *metodo lineare a  $r$  passi*. L'attributo *lineare* si riferisce al fatto che il valore  $\eta_{i+r}$  è una *combinazione lineare* dei valori  $\eta_j$  e dei valori  $f(t_j, \eta_j)$ ,  $j = i, i + 1, \dots, i + r$ ;  $r$  *passi*, in quanto per il calcolo di  $\eta_{i+r}$  sono richiesti  $r$  valori  $\eta_i, \eta_{i+1}, \dots, \eta_{i+r-1}$  (cfr. Figura 7.23). Quando  $\beta_r = 0$  il metodo è di tipo *esplicito*,

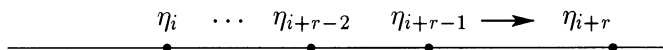


Figura 7.23: Metodo a  $r$  passi.

altrimenti è *implicito*, in quanto l'incognita  $\eta_{i+r}$  compare anche nel secondo membro nel termine  $f(t_{i+r}, \eta_{i+r})$ .

La formula ricorrente (7.53) richiede per  $i = 0$  la conoscenza di  $r$  valori, ossia  $\eta_0, \eta_1, \dots, \eta_{r-1}$ , mentre il problema a valori iniziali fornisce solo il valore  $y_0$ . Gli altri valori iniziali di  $\eta$  devono perciò essere calcolati utilizzando opportunamente l'equazione differenziale. Per tale calcolo si possono, ad esempio, utilizzare i metodi a un passo che abbiamo analizzato nel paragrafo precedente; oppure, più convenientemente si può adottare una strategia di cambiamento dell'ordine e del passo. In maniera schematica, tale procedura consiste nell'utilizzare, ad esempio, il metodo di Eulero per calcolare  $\eta_1$ ; successivamente, avendo a disposizione  $\eta_0, \eta_1$ , si applica un metodo a due passi e di ordine 2, e così di seguito; all'aumento dell'ordine corrisponde la possibilità di aumentare il passo.

La costruzione di metodi particolari della forma (7.53) si basa sostanzialmente sulle seguenti due idee. La prima, illustrata nel successivo Esempio 7.11, consiste nell'approssimazione della derivata  $y'$  mediante opportuni rapporti incrementali; la seconda idea parte dalla formulazione integrale del problema a valori iniziali e approssima l'integrale mediante una formula di quadratura. Su tale idea si basano i metodi di Adams che analizzeremo nel successivo paragrafo.

► **Esempio 7.11** Dati i tre nodi  $t_i, t_{i+1}, t_{i+2}$ , se si approssima la derivata  $y'$  nel punto  $t_{i+1}$  mediante le differenze centrali, si ottiene la seguente formula

$$\eta_{i+2} = \eta_i + 2h f(t_{i+1}, \eta_{i+1}) \quad (7.54)$$

che corrisponde ad un metodo lineare a 2 passi. Con uno sviluppo in serie si può dimostrare facilmente che l'errore di discretizzazione locale, definito da

$$\tau(t, y, h) := \frac{1}{h}(y(t+h) - y(t-h) - 2hf(t, y))$$

è infinitesimo con  $h^2$  (quando la soluzione  $y$  ammette la derivata terza). Si tratta, quindi, di un metodo del *secondo ordine*. Vedremo nel seguito che, pur essendo un metodo *convergente*, ha proprietà di stabilità non del tutto soddisfacenti (cfr. successivo Esempio 7.13). In letteratura è noto come *metodo delle differenze centrali* (mid-point) o *metodo leap-frog*.

Un metodo, ancora del secondo ordine, ma con caratteristiche di stabilità migliori, è ottenuto approssimando la derivata prima con una *differenza all'indietro* di ordine due

$$\nabla\eta_{i+2} + \frac{1}{2}\nabla^2\eta_{i+2} = hf(t_{i+2}, \eta_{i+2})$$

che fornisce il seguente metodo lineare a 2 passi, di tipo *implicito*

$$\frac{3}{2}\eta_{i+2} - 2\eta_{i+1} + \frac{1}{2}\eta_i = hf(t_{i+2}, \eta_{i+2}) \quad (7.55)$$

Tale metodo è un caso particolare della seguente famiglia di metodi, noti come *metodi alle differenze all'indietro* (backward differentiation formulas, brevemente BDF)

$$\nabla\eta_{i+2} + \frac{1}{2}\nabla^2\eta_{i+2} + \frac{1}{3}\nabla^3\eta_{i+2} + \dots + \frac{1}{r}\nabla^r\eta_{i+2} = hf(t_{i+2}, \eta_{i+2})$$

Per ogni  $r \geq 1$  la formula fornisce un metodo implicito a  $r$  passi e di ordine  $r$ . Si vede, comunque, che per motivi di stabilità, solo i metodi per  $r \leq 6$  sono convergenti; su tali metodi è basata un noto programma (dovuto a Gear [63]) per la risoluzione numerica di problemi differenziali di tipo stiff. ■

### Metodi di Adams

I *metodi di Adams*, una sottofamiglia dei metodi lineari a più passi della forma seguente

$$\eta_{i+r} - \eta_{i+r-1} = h \sum_{j=0}^r \beta_j f_{i+j} \quad (7.56)$$

ove si è posto, per brevità  $f_{i+j} = f(t_{i+j}, \eta_{i+j})$ , rappresentano, per le loro proprietà di stabilità e facilità di implementazione, i metodi a più passi più comunemente utilizzati per la risoluzione di problemi a valori iniziali di tipo non stiff. I coefficienti  $\beta_j$  possono essere ottenuti mediante un procedimento di interpolazione, a partire dalla relazione

$$y(t_{i+r}) - y(t_{i+r-1}) = \int_{t_{i+r-1}}^{t_{i+r}} f(t, y(t)) dt \quad (7.57)$$

che si ricava integrando l'equazione differenziale  $y' = f(t, y)$  tra i nodi  $t_{i+r-1}$  e  $t_{i+r}$ . Per ogni  $r$  fissato si ottiene allora un particolare metodo sostituendo alla funzione

integranda il polinomio di interpolazione relativo alle coppie di punti  $(t_j, f_j)$ ,  $j = i, i+1, \dots, i+r$ , ossia utilizzando una formula di quadratura di tipo interpolazione (cfr. Capitolo 6). In particolare, quando non si utilizza come nodo della formula il punto  $t_{i+r}$ , si ottengono dei metodi espliciti, che sono noti in letteratura come *metodi Adams-Bashforth*, mentre i metodi impliciti sono chiamati *metodi Adams-Moulton*<sup>14</sup>. Di seguito sono riportati, come esemplificazione, i primi metodi delle due famiglie, con il corrispondente errore di troncamento locale, che può essere ricavato per sviluppo in serie o, direttamente, dall'errore di troncamento relativo alla formula di quadratura utilizzata. Nella presentazione degli algoritmi impliciti, rispetto alla scrittura standard (7.53), si è aumentato di 1 l'indice  $i$ . Il motivo sarà chiarito nel seguito, in relazione alla risoluzione dello schema implicito mediante la tecnica del *predictor-corrector*, nella quale si abbina al metodo implicito considerato un opportuno metodo esplicito.

#### Adams-Bashforth

$$\eta_{i+2} - \eta_{i+1} = \frac{h}{2} [3f_{i+1} - f_i], \quad \tau = \frac{5h^2}{12} y''' \quad (7.58)$$

$$\eta_{i+3} - \eta_{i+2} = \frac{h}{12} [23f_{i+2} - 16f_{i+1} + 5f_i], \quad \tau = \frac{3h^3}{8} y^{(4)} \quad (7.59)$$

$$\eta_{i+4} - \eta_{i+3} = \frac{h}{24} [55f_{i+3} - 59f_{i+2} + 37f_{i+1} - 9f_i], \quad \tau = \frac{251h^4}{720} y^{(5)} \quad (7.60)$$

#### Adams-Moulton

$$\eta_{i+2} - \eta_{i+1} = \frac{h}{2} [f_{i+2} + f_{i+1}], \quad \tau = -\frac{h^2}{12} y''' \quad (7.61)$$

$$\eta_{i+3} - \eta_{i+2} = \frac{h}{12} [5f_{i+3} + 8f_{i+2} - f_{i+1}], \quad \tau = -\frac{h^3}{24} y^{(4)} \quad (7.62)$$

$$\eta_{i+4} - \eta_{i+3} = \frac{h}{24} [9f_{i+4} + 19f_{i+3} - 5f_{i+2} + f_{i+1}], \quad \tau = -\frac{19h^4}{720} y^{(5)} \quad (7.63)$$

#### Metodi predictor-corrector

Dalle formule (7.58)–(7.63) si vede che a parità di costo, in termini di valutazioni della funzione  $f$ , i metodi di Adams-Moulton sono più precisi dei corrispondenti metodi di Adams-Bashforth; vedremo, inoltre, nel seguito (cfr. successiva Figura 7.25), che

<sup>14</sup>J. C. Adams (1819–1892), matematico applicato e astronomo inglese, predisse nel 1845 l'esistenza dell'orbita del pianeta Nettuno, basandosi sullo studio matematico delle perturbazioni dell'orbita di Urano. Utilizzò, insieme a Bashforth, l'algoritmo, che porta ora i loro nomi, per lo studio delle forze di capillarità. F. R. Moulton (1872–1952) adattò tale algoritmo per approssimare le soluzioni di problemi balistici.

essi presentano anche migliori proprietà di stabilità a passo fissato. Per tali motivi gli schemi impliciti sono preferibili a quelli espliciti, anche se il loro utilizzo richiede la risoluzione, ad ogni passo, di un'equazione (un sistema nel caso di un sistema di equazioni differenziali), in generale non lineare. Una tecnica interessante per la risoluzione di quest'ultimo problema consiste nell'associare allo schema implicito utilizzato un opportuno schema esplicito. Per illustrare tale tecnica, nota come *metodo predictor-corrector*<sup>15</sup>, consideriamo la formula di Adams-Moulton (7.62) riscritta nel seguente modo

$$\eta_{i+3} = \frac{5h}{12} f(t_{i+3}, \eta_{i+3}) + \eta_{i+2} + \frac{h}{12} [8f_{i+2} - f_{i+1}] \quad (7.64)$$

Supponendo noti i valori  $\eta_{i+1}$  e  $\eta_{i+2}$ , la (7.64) è un'equazione non lineare nell'incognita  $\eta_{i+3}$ . Per la sua risoluzione si può utilizzare il procedimento di punto fisso considerato nel Capitolo 5 e che consiste nella costruzione, a partire da una stima iniziale  $\eta_{i+3}^{(0)}$ , della successione  $\{\eta_{i+3}^{(r)}\}$  attraverso il procedimento iterativo

$$\eta_{i+3}^{(r+1)} = \frac{5h}{12} f(t_{i+3}, \eta_{i+3}^{(r)}) + \eta_{i+2} + \frac{h}{12} [8f_{i+2} - f_{i+1}], \quad r = 0, 1, \dots \quad (7.65)$$

che converge quando  $h$  verifica la seguente condizione

$$\frac{5|h|}{12} L < 1$$

ove  $L$  è la costante di Lipschitz di  $f$  rispetto alla variabile  $y$ . Un valore iniziale  $\eta_{i+3}^{(0)}$ , già opportunamente vicino alla soluzione, può essere ottenuto utilizzando la soluzione fornita nel punto  $t_{i+3}$  da un metodo esplicito di ordine confrontabile al metodo implicito utilizzato. Nel caso particolare considerato possiamo, ad esempio, utilizzare la formula (7.59), che risulta dello stesso ordine della (7.64). La formula (7.59) viene allora detta *predictor*, in quanto fornisce un valore  $\eta_{i+3}^*$  utilizzato nella formula (7.64), detta *corrector*, nel seguente modo

$$\begin{array}{l} \eta_{i+3}^* = \eta_{i+2} + \frac{h}{12} [23f_{i+2} - 16f_{i+1} + 5f_i] \\ \eta_{i+3} = \frac{5h}{12} f(t_{i+3}, \eta_{i+3}^*) + \eta_{i+2} + \frac{h}{12} [8f_{i+2} - f_{i+1}] \end{array} \quad (7.66)$$

L'applicazione del metodo richiede due valutazioni della funzione  $f$ . Naturalmente, se la costante di contrazione  $5|h|L/12$  non è sufficientemente piccola, possono essere necessarie ulteriori iterazioni del procedimento (7.65). D'altra parte, come abbiamo osservato nell'Esempio 7.9, nei problemi stiff può essere necessario ricorrere ai metodi di tipo Newton.

<sup>15</sup>Il metodo predictor-corrector è stato introdotto da F. R. Moulton (1926) e da W. E. Milne (1926); per la risoluzione delle equazioni implicite, J. C. Adams utilizzava il metodo di Newton.

Un altro aspetto importante della tecnica predictor-corrector è la possibilità che essa offre di stimare l'errore locale, e quindi di adattare il passo di discretizzazione  $h$  al comportamento della soluzione. Assumendo, come illustrazione, ancora le formule particolari (7.66), dalle espressioni degli errori locali della formula (7.59) e della formula (7.62), è possibile mostrare che l'errore locale  $\tau(h, f)$  relativo al metodo predictor-corrector è stimato dalla seguente quantità

$$\tau(h, f) \approx \frac{1}{10}[\eta_{i+3}^* - \eta_{i+3}] \quad (7.67)$$

### 7.3.8 Convergenza dei metodi lineari a più passi

In maniera schematica, la convergenza significa che è possibile ottenere, in assenza di errori di arrotondamento, una approssimazione prefissata utilizzando un passo di discretizzazione  $h$  opportuno. In maniera più precisa, si ha la seguente definizione.

**Definizione 7.2 (convergenza)** Dato il problema a valori iniziali  $y' = f(t, y)$ ,  $y(t_0) = y_0$ , il metodo lineare a  $r$  passi

$$\sum_{j=0}^r \alpha_j \eta_{i+j} = h \sum_{j=0}^r \beta_j f_{i+j} \quad (7.68)$$

con  $\eta_0(h), \eta_1(h), \dots, \eta_{r-1}(h)$  tendenti a  $y(t_0) = y_0$  per  $h \rightarrow 0$ , è detto convergente in un punto fissato  $\bar{t}$ , se, posto  $\bar{t} = t_0 + (i+r)h$ , si ha

$$\lim_{h \rightarrow 0} \eta_{i+r} = \lim_{i \rightarrow \infty} \eta_{i+r} = y(\bar{t})$$

Come per i metodi a un passo, la convergenza può essere stabilita sulla base della consistenza e della stabilità. Ricordiamo che il metodo è consistente quando l'errore di discretizzazione locale tende a zero per  $h \rightarrow 0$ . Per lo schema (7.68) l'errore di discretizzazione locale  $\tau(h)$  è definito dalla relazione

$$\sum_{j=0}^r \alpha_j y(t+jh) - h \sum_{j=0}^r \beta_j y'(t+jh) = h\tau$$

ove  $y(t)$  è una soluzione dell'equazione  $y' = f(t, y)$ . In sostanza, l'errore  $h\tau$  misura di quanto la soluzione discreta differisce dalla soluzione continua, quando i valori precedenti sono supposti esatti, cioè  $\eta_j = y(t_j)$ ,  $j = i, i+1, \dots, i+r-1$ . Quando  $\tau = O(h^p)$ , il metodo è detto di ordine  $p$ . Per il calcolo dell'ordine di un metodo è utile il seguente risultato, che può essere ottenuto facilmente mediante sviluppo in serie. Se  $y(t)$  è sufficientemente regolare, si può esprimere  $h\tau$  nella seguente forma

$$h\tau = C_0 y(t) + C_1 h y'(t) + C_2 h^2 y''(t) + \dots + C_q h^q y^{(q)}(t) + \dots$$



ove le costanti  $C_q$  sono date da

$$\begin{aligned} C_0 &= \alpha_0 + \alpha_1 + \cdots + \alpha_r \\ C_1 &= \alpha_1 + 2\alpha_2 + \cdots + r\alpha_r - (\beta_0 + \beta_1 + \cdots + \beta_r) \end{aligned}$$

e per  $q = 2, 3, \dots$

$$C_q = \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + \cdots + r^q\alpha_r) - \frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2 + \cdots + r^{q-1}\beta_r)$$

Il metodo è, allora, di ordine  $p$  quando  $C_0 = C_1 = \cdots = C_p = 0$  e  $C_{p+1} \neq 0$ ; la costante  $C_{p+1}$  è chiamata la *costante d'errore*. Per un metodo di ordine  $p$  si ha, pertanto,  $h\tau = C_{p+1}h^{p+1}y^{(p+1)}(t) + O(h^{p+2})$ .

Osserviamo che un metodo lineare a  $r$  passi è individuato dai coefficienti  $\alpha_j$  e  $\beta_j$ . È opportuno associare a tali coefficienti i seguenti polinomi

$$\begin{aligned} \rho(\theta) &= \alpha_r\theta^r + \cdots + \alpha_1\theta + \alpha_0 \\ \sigma(\theta) &= \beta_r\theta^r + \cdots + \beta_1\theta + \beta_0 \end{aligned} \tag{7.69}$$

Le condizioni  $C_0 = C_1 = 0$ , che assicurano che il metodo è consistente (e del primo ordine), possono essere espresse nella forma equivalente

$$\begin{array}{|l} \rho(1) = 0 \\ \rho'(1) = \sigma(1) \end{array} \tag{7.70}$$

Tornando al problema della convergenza, si può vedere che la condizione (7.70) è una *condizione necessaria* per la convergenza. In altre parole, un *metodo lineare a più passi convergente è consistente*.

Nel caso dei metodi a un passo abbiamo visto che la consistenza è anche una condizione sufficiente per la convergenza, quando la funzione  $\Phi$  che definisce il metodo è una funzione lipschitziana. Al contrario, per i metodi a più passi è richiesta, per la convergenza, una condizione aggiuntiva sulle radici del polinomio  $\rho(\theta) = 0$ . La condizione è, in sostanza, dovuta al fatto che un metodo a più passi è una *equazione alle differenze* di ordine superiore a 1, mentre l'equazione differenziale è del primo ordine. All'analisi di tale condizione premettiamo alcune nozioni elementari sulle equazioni alle differenze.

### Equazioni alle differenze lineari

Un'equazione della forma

$$a_r z_{i+r} + a_{r-1} z_{i+r-1} + \cdots + a_0 z_i = b_i, \quad i = 0, 1, \dots \tag{7.71}$$

è chiamata una *equazione alle differenze lineare di ordine  $r$* . Le costanti  $a_0, a_1, \dots, a_r$  sono quantità assegnate, con  $a_0 \neq 0$  e  $a_r \neq 0$ . Pure la successione  $b_i$  è assegnata; in particolare, quando  $b_i = 0, i = 0, 1, \dots$ , l'equazione alle differenze è chiamata *omogenea*. Una *soluzione* dell'equazione alle differenze è una successione  $\{z_i\}$  che soddisfa la relazione (7.71) per tutti i valori di  $i$ .

Osserviamo che l'esistenza di soluzioni dell'equazione (7.71) non pone problemi, in quanto per ogni insieme di valori iniziali  $z_0, z_1, \dots, z_{r-1}$  una soluzione è generata dalla seguente iterazione

$$z_i = \frac{1}{a_r} [b_i - (a_{r-1}z_{i-1} + \dots + a_0z_{i-r})], \quad i = r, r+1, \dots$$

È comunque interessante fornire una espressione generale della soluzione, in analogia all'integrale generale nell'ambito delle equazioni differenziali. Le idee che sono alla base della costruzione dell'espressione generale della soluzione di una equazione alle differenze sono introdotte nel successivo esempio.

► **Esempio 7.12** (*Soluzione generale di un'equazione alle differenze*) Data la seguente equazione alle differenze omogenea del secondo ordine

$$z_{i+2} + a_1z_{i+1} + a_0z_i = 0, \quad i = 0, 1, 2, \dots \quad (7.72)$$

cerchiamo soluzioni particolari della forma  $z_i = \theta^i$ , con  $\theta \neq 0$ . Inserendo tale espressione nell'equazione (7.72), si ha

$$\theta^{i+2} + a_1\theta^{i+1} + a_0\theta^i = 0$$

da cui, dividendo per  $\theta^i$

$$\theta^2 + a_1\theta + a_0 = 0$$

Quest'ultima equazione è chiamata *equazione caratteristica* corrispondente all'equazione alle differenze (7.72). Supponiamo dapprima che l'equazione caratteristica abbia due radici  $\theta_1, \theta_2$  distinte. Allora, si vede immediatamente che una qualunque combinazione lineare  $z_i = c_1\theta_1^i + c_2\theta_2^i$  è una soluzione dell'equazione alle differenze. Si ha, infatti

$$\begin{aligned} z_{i+2} + a_1z_{i+1} + a_0z_i &= (c_1\theta_1^{i+2} + c_2\theta_2^{i+2}) + a_1(c_1\theta_1^{i+1} + c_2\theta_2^{i+1}) + a_0(c_1\theta_1^i + c_2\theta_2^i) \\ &= c_1(\theta_1^{i+2} + a_1\theta_1^{i+1} + a_0\theta_1^i) + c_2(\theta_2^{i+2} + a_1\theta_2^{i+1} + a_0\theta_2^i) = 0 \end{aligned}$$

Le costanti  $c_1$  e  $c_2$  possono essere determinate dai valori iniziali

$$c_1 + c_2 = z_0; \quad c_1\theta_1 + c_2\theta_2 = z_1$$

con  $z_0$  e  $z_1$  valori assegnati.

Ad esempio, la seguente equazione

$$z_0 = z_1 = 1, \quad z_{i+2} = z_{i+1} + z_i, \quad i = 0, 1, \dots \quad (7.73)$$

che genera i cosiddetti *numeri di Fibonacci*, ha la seguente soluzione

$$z_i = \frac{1}{\sqrt{5}} \left( \frac{1 + \sqrt{5}}{2} \right)^{i+1} - \frac{1}{\sqrt{5}} \left( \frac{1 - \sqrt{5}}{2} \right)^{i+1}$$

Quando l'equazione caratteristica dell'equazione alle differenze ha una *radice doppia*  $\theta_1$ , allora si verifica facilmente che  $z_i = i\theta_1^i$  è una soluzione dell'equazione alle differenze, e che quindi in questo caso la soluzione generale dell'equazione alle differenze è data dalla successione  $z_i = c_1 i\theta_1^i + c_2 \theta_1^i$ .

L'espressione generale della soluzione di un'equazione alle differenze permette, in particolare, di esaminare la *propagazione degli errori*. Questo è, naturalmente, importante nelle applicazioni agli schemi alle differenze per l'approssimazione dei problemi a valori iniziali. Consideriamo, ad esempio, la seguente equazione alle differenze

$$z_{i+2} - 2.5z_{i+1} + z_i = 0$$

che ha come soluzione generale la successione

$$z_i = c_1 2^{-i} + c_2 2^i$$

In corrispondenza ai valori iniziali  $z_0 = 2, z_1 = 1$  la soluzione è data dalla successione decrescente  $z_i = 2^{-i+1}$ . Se, tuttavia, consideriamo la seguente perturbazione dei valori iniziali

$$z_0 = 2, \quad z_1 = 1 + \epsilon$$

si ottiene la seguente soluzione

$$z_i = \left(2 - \frac{2\epsilon}{3}\right) 2^{-i} + \frac{2\epsilon}{3} 2^i$$

che, per  $i$  sufficientemente grande, è una successione crescente.

Terminiamo l'esempio, mostrando come in alcuni casi dall'espressione della soluzione dell'equazione omogenea sia possibile costruire una soluzione particolare della equazione non omogenea. Consideriamo la seguente equazione

$$z_{i+2} - 3z_{i+1} + 2z_i = i$$

la cui equazione omogenea ha la soluzione generale  $z_i = c_1 1^i + c_2 2^i$ . Cerchiamo come soluzione dell'equazione non omogenea un polinomio di grado 2 nella  $i$ , cioè

$$\bar{z}_i = d_1 i^2 + d_2 i$$

Inserendo tale successione nell'equazione e identificando i coefficienti, si ottiene  $d_1 = d_2 = -1/2$ . Pertanto la soluzione generale dell'equazione non omogenea assegnata è data dalla successione

$$z_i = c_1 + c_2 2^i - \frac{1}{2}(i^2 + i)$$

■

### Condizione delle radici

Utilizzando i risultati precedenti sulle equazioni alle differenze, mostriamo che per un metodo convergente le radici del polinomio  $\rho(\theta)$  devono soddisfare ad una opportuna condizione. Consideriamo il problema particolare  $y' = 0, y(0) = 0$ , che ha come

soluzione ovvia la funzione  $y(t) \equiv 0$ . Quando applichiamo un metodo a più passi a tale problema, si ha l'equazione alle differenze

$$\sum_{j=0}^r \alpha_j \eta_{i+j} = 0$$

la cui equazione caratteristica è data da  $\rho(\theta) = 0$ . Se  $\theta_j$  sono gli zeri del polinomio  $\rho(\theta)$ , le successioni

$$z_i = h \theta_j^i, \quad i = 0, 1, \dots$$

sono soluzioni dell'equazione alle differenze, e, inoltre, sono tali che i valori iniziali, per  $i = 0, 1, \dots, r-1$  tendono a  $y(0) = 0$  per  $h \rightarrow 0$ . Dalla definizione di convergenza si ha che per ogni  $\bar{t}$  fissato, posto  $\bar{t} = (i+r)h$ ,  $\eta_{i+r} = h \theta_j^{i+r} \rightarrow y(\bar{t}) = 0$  per  $i \rightarrow \infty$ , o equivalentemente per  $h \rightarrow 0$ . Si vede pertanto che la convergenza implica la condizione  $|\theta_j| \leq 1$ . Ragionando in maniera analoga, si trova che nel caso di radici multiple si deve avere  $|\theta_j| < 1$ .

In conclusione, se un metodo a più passi è convergente, le radici  $\theta_j$  del polinomio  $\rho(\theta)$  verificano la seguente condizione, nota come *condizione delle radici*

1.  $|\theta_j| \leq 1$  se  $\theta_j$  è una radice semplice;
2.  $|\theta_j| < 1$  se  $\theta_j$  è una radice multipla.

La condizione delle radici insieme alla consistenza è anche sufficiente per la convergenza. Si ha, infatti, il seguente importante risultato.

**Teorema 7.3 (Convergenza)** *Un metodo a più passi lineare è convergente se e solo se il metodo è consistente e soddisfa la condizione delle radici.*

Applicando il teorema precedente, si vede immediatamente che i *metodi di Adams sono metodi convergenti*. Anche il metodo alle differenze centrali è un metodo convergente, in quanto è consistente per costruzione e le soluzioni dell'equazione  $\rho(\theta) = \theta^2 - 1$  sono sul cerchio unitario, ma distinte. Si può, invece, verificare che i metodi alle differenze all'indietro (BDF) verificano la condizione delle radici soltanto per i metodi di ordine minore o uguale a 6. Più in generale, la condizione delle radici impedisce di utilizzare tutti coefficienti  $a_j$  e  $b_j$ , presenti nella formula generale di un metodo a più passi lineari, per avere un metodo di ordine massimo possibile. Si può, infatti, mostrare che per l'ordine  $p$  di un metodo lineare a  $r$  passi, che verifica la condizione delle radici, si ha  $p \leq r + 2$ , quando  $r$  è pari e  $p \leq r + 1$ , per  $r$  dispari.

### 7.3.9 Stabilità per passo fissato

La condizione delle radici assicura la stabilità per  $h \rightarrow 0$ . Per tale motivo i metodi che verificano tale condizione sono anche chiamati *zero-stabili*, o asintoticamente stabili. Tuttavia, nella implementazione effettiva dei metodi viene, in effetti, utilizzato

un passo  $h$  diverso dallo zero e in alcune applicazioni si vuole che la soluzione discreta descriva opportunamente la soluzione continua anche per  $h$  non eccessivamente piccolo (cfr. Esempio 7.9). Pertanto, per essere utile, un metodo deve mantenere la stabilità anche per opportuni valori di  $h$  diversi dallo zero. Si tratta, chiaramente, di una questione molto importante per le applicazioni, ma la cui trattazione esula dai limiti del presente testo. Ci limiteremo, quindi, ad evidenziare il problema mediante un esempio, rinviando alla bibliografia per un opportuno approfondimento. Aggiungiamo, soltanto, che è proprio la stabilità a passo fissato ad impedire, in generale, l'uso di metodi a più passi, ad esempio di Adams, di ordine troppo elevato; in effetti, anche per problemi regolari e non stiff, negli algoritmi con cambiamento automatico del passo e dell'ordine l'ordine non è, in generale, superiore al valore 10.

► **Esempio 7.13** Consideriamo l'applicazione del metodo delle differenze centrali al problema a valori iniziali  $y' = -2y$ ,  $y(0) = 1$ , che ha come soluzione  $y(t) = e^{-2t}$ . L'applicazione del metodo fornisce la seguente equazione alle differenze

$$\eta_{i+2} = \eta_i + 2h(-2\eta_{i+1}), \quad i = 0, 1, 2, \dots$$

Come valori iniziali assumiamo  $\eta_0 = 1$  e per  $\eta_1$  il valore fornito dal metodo di Eulero  $\eta_1 = \eta_0 + h(-2\eta_0)$ . Con  $h = 0.1$  si ottengono i risultati mostrati in Figura 7.24.

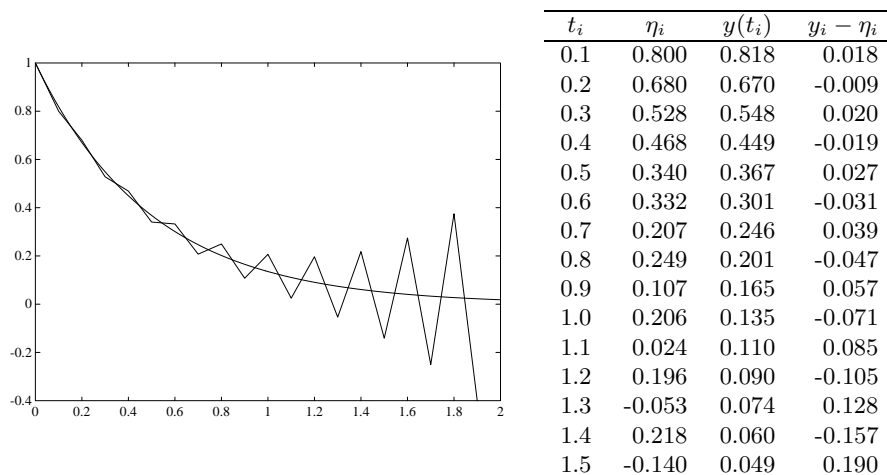


Figura 7.24: Instabilità numerica del metodo delle differenze centrali.

Come si vede, la soluzione numerica presenta un comportamento oscillatorio e il modulo dell'errore aumenta indefinitamente, all'aumentare di  $t$ . Se ne ricava che il metodo, anche se convergente e del secondo ordine, non è in pratica utilizzabile, almeno se si è interessati alla soluzione per  $t$  sufficientemente grande.

Vediamo brevemente il motivo di questo comportamento. Consideriamo, più in generale, il problema  $y' = \lambda y$ ,  $y(0) = 1$ , per il quale il metodo fornisce l'equazione alle differenze del secondo ordine

$$\eta_{i+2} = \eta_i + 2h\lambda\eta_{i+1}$$

Le radici della corrispondente equazione caratteristica  $\theta^2 - 2h\lambda\theta - 1 = 0$  sono

$$\theta_{1,2} = h\lambda \pm \sqrt{h^2\lambda^2 + 1}$$

Mediante uno sviluppo in serie, per  $h^2\lambda^2 < 1$ , si ha

$$\begin{aligned}\theta_1 &= 1 + h\lambda + \frac{h^2\lambda^2}{2} - \frac{h^4\lambda^4}{8} + \dots = e^{h\lambda} + O(h^3) \\ \theta_2 &= -1 + h\lambda - \frac{h^2\lambda^2}{2} + \frac{h^4\lambda^4}{8} + \dots = -e^{h\lambda} + O(h^3)\end{aligned}$$

La soluzione generale dell'equazione alle differenze è

$$\eta_i = c_1\theta_1^i + c_2\theta_2^i = c_1e^{ih\lambda} + c_2(-1)^ie^{-ih\lambda} + O(h^2)$$

Le condizioni iniziali  $\eta_0 = 1, \eta_1 = e^{h\lambda}$  danno

$$c_1 = 1 - c_2; \quad c_2 = \frac{e^{h\lambda} - \theta_1}{\theta_2 - \theta_1}$$

Osserviamo anche che, essendo  $t_i = ih$ , si ha

$$\eta_i = c_1e^{t_i\lambda} + c_2(-1)^ie^{-t_i\lambda} + O(h^2)$$

Si vede, pertanto, che la formula alle differenze centrali produce una soluzione con due componenti: la prima corrisponde alla soluzione esatta dell'equazione differenziale, mentre l'altra è una soluzione spuria con carattere oscillatorio. Per  $\lambda < 0$ , la soluzione esatta è decrescente, mentre la soluzione spuria aumenta con  $t$  e può dominare la soluzione numerica. Osserviamo che anche nel caso in cui i valori iniziali sono tali che  $c_2 = 0$ , la soluzione spuria è ugualmente presente a causa degli errori di arrotondamento. L'analisi ha messo, quindi, in evidenza che il metodo delle differenze centrali è un metodo *instabile* per ogni valore di  $h$ . Il metodo non è, quindi, appropriato per  $\lambda < 0$ . Osserviamo che, al contrario, per  $\lambda > 0$  la soluzione spuria tende a zero. Per tale motivo si dice, anche, che il metodo è *relativamente stabile* sull'intervallo  $(0, +\infty)$ .

In Figura 7.25 sono rappresentati gli intervalli di stabilità a passo fissato di alcuni metodi di Adams. Più precisamente, sono rappresentati i valori di  $h\lambda$  per i quali le soluzioni, ottenute con i corrispondenti metodi per il problema test  $y' = \lambda y$ , convergono a zero per  $i \rightarrow \infty$ . Come si vede la zona di stabilità è maggiore, a parità di ordine, per i metodi impliciti, e, in ogni caso, diminuisce all'aumentare dell'ordine. ■

### 7.3.10 Sistemi di equazioni del primo ordine

I metodi numerici analizzati in precedenza nel caso di una equazione differenziale si adattano facilmente alla risoluzione di sistemi di equazioni differenziali. Come esemplificazione, vedremo l'implementazione del metodo di Runge-Kutta e del metodo predictor-corrector di Adams nel caso di un sistema di due equazioni differenziali

$$\begin{aligned}\frac{dy}{dt} &= f(t, y, z), & y(t_0) &= y_0 \\ \frac{dz}{dt} &= g(t, y, z), & z(t_0) &= z_0\end{aligned}$$

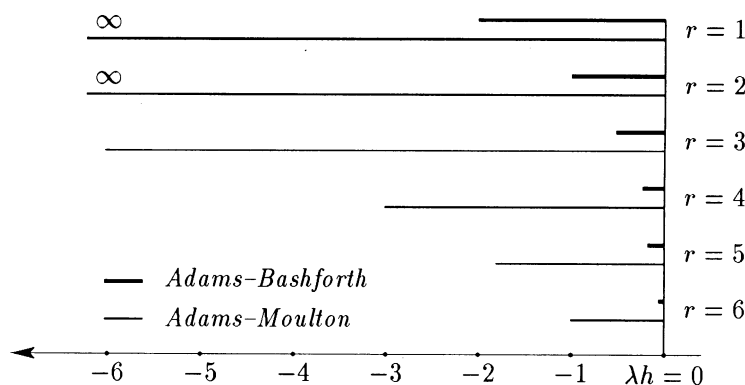


Figura 7.25: Intervalli di stabilità a passo fissato dei metodi di Adams espliciti e impliciti.

Le soluzioni numeriche sono indicate con il vettore  $(\eta_i, \xi_i)$ ,  $i = 0, 1, \dots$

**Algoritmo 7.3** (Metodo di Runge-Kutta)

Input:  $f(t, y, z)$ ,  $g(t, y, z)$ ,  $t_0, y_0, z_0, T$  e  $n$ , numero delle suddivisioni dell'intervallo  $(t_0, T)$ .

Output: soluzione approssimata  $\eta_i, \xi_i$ .

```

set  $h = (T - t_0)/n$ 
 $\eta_0 = y_0$ 
do  $i = 0, \dots, n - 1$ 
 $k_1 = hf(t_i, \eta_i, \xi_i)$  ;  $l_1 = hg(t_i, \eta_i, \xi_i)$ 
 $k_2 = hf(t_i + h/2, \eta_i + k_1/2, \xi_i + l_1/2)$ ;  $l_2 = hg(t_i + h/2, \eta_i + k_1/2, \xi_i + l_1/2)$ 
 $k_3 = hf(t_i + h/2, \eta_i + k_2/2, \xi_i + l_2/2)$ ;  $l_3 = hg(t_i + h/2, \eta_i + k_2/2, \xi_i + l_2/2)$ 
 $k_4 = hf(t_i + h, \eta_i + k_3, \xi_i + l_3)$ ;  $l_4 = hg(t_i + h, \eta_i + k_3, \xi_i + l_3)$ 
 $t_{i+1} = t_i + h$ 
 $\eta_{i+1} = \eta_i + (k_1 + 2k_2 + 2k_3 + k_4)/6$ 
 $\xi_{i+1} = \xi_i + (l_1 + 2l_2 + 2l_3 + l_4)/6$ 
end do

```

La successiva routine RK4 implementa l'algoritmo di Runge-Kutta per un sistema di equazioni differenziali di ordine  $n$ . La routine avanza l'integrazione dal generico punto  $t$  al punto  $t + h$ ; può essere, quindi, utilizzata per ottenere la soluzione approssimata su punti particolari dell'intervallo di integrazione.

```

SUBROUTINE RK4(Y,DY,N,T,H,TOUT,YOUT,F)
C.....
C    dati i valori delle N variabili Y e le loro derivate DY,
C    mediante il metodo di Runge-Kutta si avanza la soluzione da T a T+H.
C    TOUT = T+H
C    YOUT soluzione in TOUT

```

```

C      F subroutine che definisce la funzione f(t,y)
C      SUBROUTINE F(T,Y,DY)
C
C.....
      PARAMETER (NMAX=20)
      DIMENSION Y(N),DY(N),YOUT(N),YT(NMAX),DYT(NMAX),DYM(NMAX)
      HH=H*0.5
      H6=H/6.
      TH=T+HH
      DO 10 I=1,N
        YT(I)=Y(I)+HH*DY(I)
10     CONTINUE
      CALL F(TH,YT,DYT)
      DO 20 I=1,N
        YT(I)=Y(I)+HH*DYT(I)
20     CONTINUE
      CALL F(TH,YT,DYM)
      DO 30 I=1,N
        YT(I)=Y(I)+H*DYM(I)
        DYM(I)=DYT(I)+DYM(I)
30     CONTINUE
      CALL F(T+H,YT,DYT)
      DO 40 I=1,N
        YOUT(I)=Y(I)+H6*(DY(I)+DYT(I)+2.*DYM(I))
40     CONTINUE
      TOUT=T+H
      RETURN
      END

```

In maniera analoga si estendono i metodi a più passi. Come esemplificazione, consideriamo il seguente metodo predictor-corrector del quarto ordine.

#### Algoritmo 7.4 (Metodo di Adams-Bashforth-Moulton)

Input:  $f(t, y, z)$ ,  $g(t, y, z)$ ,  $(t_i, \eta_i, \xi_i)$ ,  $i = 0, 1, 2, 3$  e  $h, n$ .

Output: soluzione approssimata  $\eta_i, \xi_i$ .

```

do i = 0, 1, 2, 3 (inizializzazione)
  fi = f(ti, ηi, ξi) ; gi = g(ti, ηi, ξi)
end do
do i = 3, ..., n - 1
  ti+1 = ti + h
  ηi+1 = ηi + h(55fi - 59fi-1 + 37fi-2 - 9fi-3)/24
  ξi+1 = ξi + h(55gi - 59gi-1 + 37gi-2 - 9gi-3)/24
  fi+1 = f(ti+1, ηi+1, ξi+1)
  gi+1 = g(ti+1, ηi+1, ξi+1)
  ηi+1 = ηi + h(9fi+1 + 19fi - 5fi-1 + fi-2)/24
  ξi+1 = ξi + h(9gi+1 + 19gi - 5gi-1 + gi-2)/24
  fi+1 = f(ti+1, ηi+1, ξi+1)
  gi+1 = g(ti+1, ηi+1, ξi+1)
end do

```



### 7.3.11 Metodo di Cowell-Numerov

Le equazioni della forma

$$y''(t) = f(t, y) \quad (7.74)$$

non contenenti cioè la derivata prima, possono essere trattate con tecniche speciali. Come esemplificazione, analizzeremo un noto metodo, che, analogamente ai metodi a più passi considerati in precedenza, può essere considerato un caso particolare di una famiglia di metodi di ordine differente<sup>16</sup>.

Mediante uno sviluppo in serie si può dimostrare che, se la soluzione  $y(t)$  è sufficientemente regolare, si ha la seguente relazione

$$\frac{y(t_{i+1}) - 2y(t_i) + y(t_{i-1}))}{h^2} = \frac{1}{12} (f(t_{i+1}, y(t_{i+1})) + 10f(t_i, y(t_i)) + f(t_{i-1}, y(t_{i-1}))) - \frac{h^2}{240} y^{(4)}(t_i + \theta h)$$

con  $\theta \in (-1, 1)$ . Se ne ricava il seguente metodo numerico, noto come *metodo di Cowell-Numerov*

$$\eta_{i+1} = 2\eta_i - \eta_{i-1} + \frac{h^2}{12} (f_{i+1} + 10f_i + f_{i-1}) \quad (7.75)$$

Tale metodo, a due passi e di tipo implicito, risulta del quarto ordine. Il primo passo richiede l'applicazione di un altro metodo, ad esempio uno sviluppo in serie.

La *convergenza* può essere dimostrata esaminando la *consistenza* e la *stabilità*. La prima proprietà è verificata per costruzione. Per quanto riguarda la stabilità, applicando il metodo al problema particolare  $y'' = 0$ ,  $y(0) = y'(0) = 0$ , si vede che essa è equivalente alla proprietà che gli zeri del polinomio caratteristico  $\rho(\theta) = \theta^2 - 2\theta + 1$  siano tutti appartenenti al cerchio unitario e le radici sulla circonferenza unitaria abbiano molteplicità non superiore a 2.

◆ **Esercizio 7.9** Applicare il seguente metodo predictor-corrector

$$\begin{aligned} \eta_{i+1}^* &= \eta_i + h f(t_i, \eta_i) \\ \eta_{i+1} &= \eta_i + h f(t_{i+1}, \eta_{i+1}^*) \end{aligned}$$

al problema a valori iniziali  $y' = 3t^2 y^2$ ,  $y(0) = -1$  per  $t \in (0, 5)$ , e al problema a valori iniziali  $y' = \lambda y$ ,  $y(0) = 1$ , ove  $\lambda = 20$  e  $\lambda = -20$ .

<sup>16</sup>Equazioni differenziali del tipo (7.74) sono utilizzate, ad esempio, nello studio di problemi di *meccanica celeste*. Per tale motivo, alcuni metodi sono legati ai nomi di astronomi: Störmer (1907) *Sur les trajectoires des corpuscules électrisés* (studio dell'aurora boreale), Cowell e Crommelin (1910) *Investigation of the motion of Halley's comet from 1759 to 1910*, Numerov (1927) *Note on the numerical integration of  $d^2x/dt^2 = f(x, t)$* .

◆ **Esercizio 7.10** Trovare la soluzione generale delle seguenti equazioni alle differenze

$$\text{a) } z_{i+2} + z_{i+1} - 6z_i = 0 \quad \text{b) } z_{i+3} - 6z_{i+2} + 11z_{i+1} - 6z_i = 0$$

◆ **Esercizio 7.11** Esaminare il seguente metodo a due passi

$$\eta_{i+2} + \alpha_1 \eta_{i+1} + a \eta_i = h(\beta_2 f_{i+2} + \beta_1 f_{i+1} + \beta_0 f_i)$$

ove  $a$  è un parametro. Determinare  $\alpha_1, \beta_2, \beta_1, \beta_0$  in maniera che il metodo sia del terzo ordine. Trovare per quali valori di  $a$  il metodo è del quarto ordine. Trovare infine i valori di  $a$  in corrispondenza ai quali è soddisfatta la condizione delle radici.

◆ **Esercizio 7.12** Esaminare la convergenza del seguente metodo

$$\eta_{i+4} - \frac{8}{19}(\eta_{i+3} - \eta_{i+1}) - \eta_i = \frac{6h}{19}(f_{i+4} + 4f_{i+3} + 4f_{i+1} + f_i)$$

◆ **Esercizio 7.13** Applicare il seguente metodo

$$\eta_{i+2} - (1 + \alpha)\eta_{i+1} + \alpha\eta_i = \frac{h}{2}[(3 - \alpha)f(t_{i+1}, \eta_{i+1}) - (1 + \alpha)f(t_i, \eta_i)]$$

rispettivamente per  $\alpha = 0$  e  $\alpha = -5$  per il calcolo della soluzione numerica del problema a valori iniziali  $y' = 4ty^{1/2}$ ,  $y(0) = 1$  per  $0 \leq t \leq 2$ , utilizzando successivamente i passi  $h = 0.1, 0.05, 0.025$ .

◆ **Esercizio 7.14** Esaminare la convergenza del seguente metodo

$$\eta_{i+2} - \eta_{i+1} = \frac{h}{3}[3f(t_{i+1}, \eta_{i+1}) - 2f(t_i, \eta_i)]$$

## 7.4 Equazioni stiff

Abbiamo già analizzato una situazione stiff nell'Esempio 7.9; in questo paragrafo ne approfondiremo il significato attraverso ulteriori esempi, allo scopo di fornire indicazioni sui metodi numerici da utilizzare. Come bibliografia, segnaliamo in particolare Hairer, Wanner [73], Lambert [104].

La terminologia *stiff* è stata introdotta da Curtis e Hirschfelder<sup>17</sup> nell'ambito dello studio di un sistema meccanico costituito da due *molle*, una delle quali molto più rigida (stiff) dell'altra. Il termine stiff è passato poi ad indicare, più in generale, un sistema di equazioni differenziali che descrive un sistema fisico caratterizzato da *costanti di tempo* molto differenti tra di loro. Ricordiamo che la costante di tempo è il termine comunemente usato per indicare la velocità di decadimento. Con riferimento all'equazione scalare  $y' = \lambda y$ , con soluzione  $ce^{\lambda t}$ , se  $\lambda$  è un numero reale e negativo, allora la  $y$  decade di un fattore  $e^{-1}$  nel tempo  $-1/\lambda$ . In questo caso la

<sup>17</sup> *stiff equations are equations where certain implicit methods, in particular BDF, perform better, usually tremendously better, than explicit ones* (Curtis, Hirschfelder, 1952).

costante di tempo è data dal valore  $-1/\lambda$ ; più  $\lambda$  è grande e più è piccola la costante di tempo. Più in generale, per un sistema di equazioni differenziali lineari  $\mathbf{y}' = \mathbf{A}\mathbf{y}$  le costanti di tempo sono date dal reciproco delle parti reali degli autovalori di  $\mathbf{A}$ .

Sistemi differenziali di tipo stiff hanno origine in numerosi e importanti settori applicativi<sup>18</sup>. Segnaliamo in particolare le applicazioni relative all'analisi di circuiti elettronici, alla cinetica chimica e biochimica, al calcolo dei reattori nucleari e in generale, ai problemi di controllo.

Le difficoltà relative alla soluzione di sistemi stiff sono di natura prettamente numerica. In maniera schematica, si ha, infatti che tali sistemi sono dal punto di vista analitico *stabili*, nel senso che la loro soluzione tende asintoticamente (per  $t \rightarrow \infty$ ) a una situazione stazionaria. Tuttavia i metodi numerici *classici*, ossia i metodi di Adams e Runge-Kutta che abbiamo analizzato nei paragrafi precedenti, non sono sufficientemente *stabili*, nel senso del passo fissato, per integrare *convenientemente* tali sistemi. In altre parole, utilizzando tali metodi per un sistema stiff, si può essere costretti ad utilizzare un passo di integrazione eccessivamente piccolo (con conseguente alto costo di risoluzione) in relazione alla regolarità della soluzione continua. In definitiva, il problema per quanto riguarda la risoluzione numerica dei sistemi stiff è la ricerca di metodi numerici per i quali l'ampiezza del passo di integrazione dipenda soltanto dalla precisione richiesta. Cercheremo di chiarire questi aspetti nell'esempio successivo.

► **Esempio 7.14** (*Esempio di sistema stiff*) Consideriamo il seguente sistema lineare

$$\begin{cases} y_1' = \frac{\lambda_1 + \lambda_2}{2} y_1 + \frac{\lambda_1 - \lambda_2}{2} y_2 \\ y_2' = \frac{\lambda_1 - \lambda_2}{2} y_1 + \frac{\lambda_1 + \lambda_2}{2} y_2 \end{cases} \quad (7.76)$$

ove supponiamo  $\lambda_i \in \mathbb{C}$ ,  $\Re(\lambda_i) < 0$ ,  $i = 1, 2$ . La soluzione generale è fornita dalle funzioni (cfr. Appendice B)

$$y_1(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}, \quad y_2(t) = c_1 e^{\lambda_1 t} - c_2 e^{\lambda_2 t}$$

In Figura 7.26 sono rappresentate le soluzioni corrispondenti ai valori iniziali  $y_1(0) = 1$ ,  $y_2(0) = 0$  e per  $\lambda_1 = -1$ ;  $\lambda_2 = -100$ . Per  $t \rightarrow \infty$  la soluzione tende a zero, cioè il sistema è asintoticamente stabile. Se le due costanti di tempo sono, come nella figura, molto diverse tra loro, si ha una zona *transiente*, ove la soluzione ha una *grande variazione*, ossia le derivate sono grandi. L'ampiezza della zona transiente è proporzionale a  $1/|\lambda_2|$ . Successivamente, la soluzione si stabilizza sulla soluzione particolare con costante di tempo più alta (nel caso della figura data da  $1/|\lambda_1|$ ).

Supponiamo ora di voler risolvere numericamente il sistema (7.76), in maniera che l'errore locale di discretizzazione si mantenga su tutto l'intervallo di integrazione dell'ordine di

<sup>18</sup> . . . around 1960, things became completely different and everyone became aware that the world was full of stiff problems (G. Dahlquist, 1985).

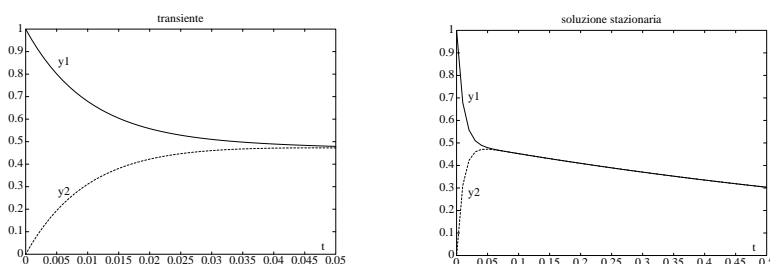


Figura 7.26: Soluzioni del sistema stiff (7.76), in corrispondenza a  $\lambda_1 = -1$ ;  $\lambda_2 = -100$ .

una precisione  $\epsilon$  prefissata. Come prima illustrazione, consideriamo il comportamento del metodo di Eulero esplicito, per il quale abbiamo visto in precedenza che l'errore locale di discretizzazione  $\tau$  è della forma

$$\tau_i(y_j, h) \approx \frac{h}{2} y_j''(\xi), \quad \xi \in (t_i, t_{i+1}) \quad (7.77)$$

per  $j = 1, 2$ . In particolare, per le soluzioni del sistema (7.76) si ha  $y_j'' = c_1 \lambda_1^2 e^{\lambda_1 t} \pm c_2 \lambda_2^2 e^{\lambda_2 t}$ , da cui si vede che le derivate seconde delle componenti della soluzione sono *grandi* nella zona transiente, mentre successivamente esse si comportano come la derivata della componente più lenta.

Appare allora evidente da (7.77) che per quanto riguarda l'errore locale dovrebbe essere possibile, per  $t$  che aumenta, utilizzare un passo di discretizzazione  $h$  sempre più grande, con notevole risparmio di operazioni, soprattutto nel caso in cui l'intervallo di integrazione  $T$  sia molto maggiore dell'ampiezza della zona transiente<sup>19</sup>. Tuttavia, come abbiamo già osservato su un'equazione scalare nell'Esempio 7.9, la scelta del passo  $h$  è per il metodo di Eulero esplicito vincolata da motivi di stabilità. Per il sistema (7.76) si verifica facilmente che le componenti  $\eta_i^{(1)}$ ,  $\eta_i^{(2)}$  della soluzione ottenuta con il metodo di Eulero esplicito, hanno la seguente forma

$$\begin{aligned} \eta_i^{(1)} &= c_1(1 + h\lambda_1)^i + c_2(1 + h\lambda_2)^i \\ \eta_i^{(2)} &= c_1(1 + h\lambda_1)^i - c_2(1 + h\lambda_2)^i \end{aligned}$$

da cui si vede che la soluzione discreta si comporta, per  $i \rightarrow \infty$ , come la soluzione continua soltanto se il passo  $h$  verifica i seguenti vincoli

$$\boxed{\begin{aligned} |1 + h\lambda_1| < 1 \\ |1 + h\lambda_2| < 1 \end{aligned} \Rightarrow 0 \leq h < \frac{2}{|\lambda_i|}, \quad i = 1, 2} \quad (7.78)$$

In conclusione, per il metodo di Eulero esplicito si ha che per *motivi di stabilità*, e non per *motivi di accuratezza*, il passo è indicato dalla componente che decade più rapidamente;

<sup>19</sup>In effetti, questo è un aspetto che caratterizza i sistemi stiff nelle applicazioni. Ad esempio, nello studio di problemi di chimica dei polimeri, la zona transiente è di qualche millesimo di secondo, mentre la reazione è di interesse su un arco di tempo dell'ordine delle ore (cfr. ad esempio Comincioli, Faucitano [34])

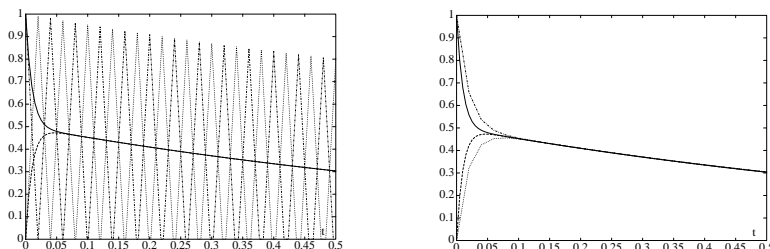


Figura 7.27: In riferimento al sistema stiff (7.76), nella prima figura è rappresentata, insieme alla soluzione continua, la soluzione discreta ottenuta mediante il metodo di Eulero esplicito con  $h = 0.02$ , mentre nella seconda figura è rappresentata la soluzione ottenuta con lo stesso passo mediante il metodo di Eulero implicito.

nel caso dell'esempio numerico precedente è necessario assumere lungo tutto l'intervallo di integrazione:  $h < 2/100 = 0.02$ , mentre per la componente che decade più lentamente sarebbe sufficiente assumere  $h < 2$ . Come illustrazione, in Figura 7.27 è mostrata la soluzione ottenuta con il metodo di Eulero esplicito per  $h = 0.02$ . Come si vede, le componenti della soluzione numerica presentano un comportamento oscillatorio; d'altra parte, si può vedere che per  $h > 0.02$  i moduli di tali componenti tendono all'infinito, per  $i \rightarrow \infty$ . Un risultato decisamente migliore si ottiene con il metodo di Eulero implicito, che ha come soluzione la seguente successione

$$\eta_i^{(1)} = \frac{c_1}{(1 - h\lambda_1)^i} + \frac{c_2}{(1 - h\lambda_2)^i}$$

$$\eta_i^{(2)} = \frac{c_1}{(1 - h\lambda_1)^i} - \frac{c_2}{(1 - h\lambda_2)^i}$$

che tende a zero, qualunque sia la scelta di  $h > 0$ . Con tale metodo è quindi possibile scegliere il passo in base solo all'accuratezza richiesta; in particolare, come si rileva dalla figura, è opportuno scegliere un passo inferiore a 0.02 nella fase transiente, ma successivamente l'ampiezza del passo può essere convenientemente elevata. ■

L'esempio precedente suggerisce una definizione generale di sistema stiff e indica la strada da seguire per ricercare metodi numerici opportuni. Rinviando al seguito la seconda questione, vediamo dapprima come può essere formalizzata la definizione di stiffness nel caso di un problema a valori iniziali per un sistema di equazioni non lineari

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad t \in (t_0, T) \quad (7.79)$$

$$\mathbf{y}(t_0) = \mathbf{y}_0 \quad (7.80)$$

ove  $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ ,  $\mathbf{f} = [f_1, f_2, \dots, f_n]^T$ ,  $n \geq 1$ . Per determinare se tale problema è di tipo stiff, è necessario conoscere la natura delle soluzioni  $\bar{\mathbf{y}}$  dell'equazione (7.79) nell'intorno della soluzione particolare  $\mathbf{y}(t)$ . In tale intorno, l'equazione (7.79)

può essere approssimata dalla seguente equazione lineare, ottenuta per sviluppo in serie e detta *equazione variazionale*

$$\bar{\mathbf{y}}' - \mathbf{J}(\bar{\mathbf{y}} - \mathbf{y}(t)) - \mathbf{f}(t, \mathbf{y}(t)) = 0$$

ove  $\mathbf{J}$  indica la matrice Jacobiana  $\mathbf{f}_{\mathbf{y}} = [\partial f_i / \partial y_j]$ ,  $i, j = 1, 2, \dots, n$ , calcolata in  $(t, \mathbf{y}(t))$ . Indichiamo con  $\lambda_i(t)$ ,  $i = 1, 2, \dots, n$  gli autovalori locali (che supporremo distinti) della matrice Jacobiana  $\mathbf{J}$ . Allora, le soluzioni  $\bar{\mathbf{y}}$  in un intorno della soluzione esatta  $\mathbf{y}(t)$  sono della forma

$$\bar{\mathbf{y}} \approx \mathbf{y}(t) + \sum_{i=1}^n c_i e^{\lambda_i t} \boldsymbol{\xi}_i$$

ove  $c_i$  sono costanti arbitrarie e  $\boldsymbol{\xi}$  sono gli autovalori di  $\mathbf{J}$ . Le autofunzioni  $e^{\lambda_i t}$  caratterizzano quindi la risposta locale del sistema a piccole variazioni o perturbazioni intorno a  $\mathbf{y}(t)$ . Supporremo che il sistema sia localmente stabile, ossia tale che  $\Re(\lambda_i) < 0$ ,  $i = 1, 2, \dots, n$ . I valori  $1/\Re(-\lambda_i)$  sono chiamate le *costanti di tempo locali* e sono alla base della seguente definizione di stiffness.

**Definizione 7.3** (stiffness) *Il problema a valori iniziali (7.79), (7.80) è detto stiff in un intervallo  $I \subset [t_0, T]$  se, per  $t \in I$*

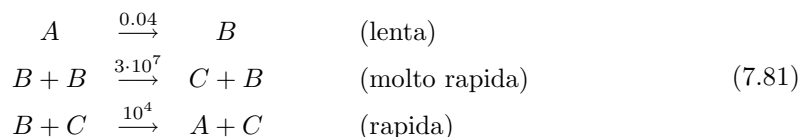
1.  $\Re(\lambda_i) < 0$ ;
2.  $S(t) := \max_{i=1, \dots, n} \Re(-\lambda_i) / \min_{i=1, \dots, n} \Re(-\lambda_i) \gg 1$

ove  $\lambda_i$  sono gli autovalori della matrice Jacobiana  $\mathbf{J}$  corrispondente alla soluzione  $\mathbf{y}$  calcolata in  $t$ .

Il rapporto  $S(t)$  può essere assunto come una *misura* di quanto il sistema dato è stiff. In realtà, la presenza o no di stiffness in un sistema differenziale è una questione abbastanza delicata, non sempre riconducibile allo schema contenuto nella definizione precedente. In effetti, si possono costruire problemi per i quali  $S(t)$  è grande, ma che possono essere risolti con metodi classici senza altre restrizioni sul passo, oltre quelle imposte dalla accuratezza richiesta. Rinviamo ad esempio a Lambert [104] per una discussione più approfondita<sup>20</sup>, ci limiteremo a fornire alcuni esempi di problemi, comunemente ritenuti stiff.

<sup>20</sup>If a numerical method with a finite region of absolute stability, applied to a system with any initial conditions, is forced to use in a certain interval of integration a step-length which is excessively small in relation to the smoothness of the exact solution in that interval, then the system is said to be stiff in that interval (Lambert, 1990).

► **Esempio 7.15** (*Sistemi di reazioni chimiche*) Consideriamo il seguente modello di reazioni chimiche, caratterizzato da velocità di reazione molto diverse fra loro



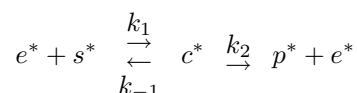
A tale modello, introdotto da Robertson<sup>21</sup> e divenuto successivamente un noto problema test su cui sperimentare gli algoritmi per le equazioni stiff, corrisponde il seguente sistema differenziale, per un particolare insieme di condizioni iniziali

$$\begin{array}{ll}
 \text{A: } y_1' = -0.04y_1 + 10^4 y_2 y_3 & y_1(0) = 1 \\
 \text{B: } y_2' = 0.04y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2 & y_2(0) = 0 \\
 \text{C: } y_3' = & 3 \cdot 10^7 y_2^2 \quad y_3(0) = 0
 \end{array} \quad (7.82)$$

ove  $y_i(t)$ ,  $i = 1, 2, 3$  indicano le concentrazioni delle sostanze  $A, B, C$  al tempo  $t$ . La matrice jacobiana  $\mathbf{J}$  ha un autovalore  $\lambda_3$  uguale a zero (dal momento che  $\sum_{i=1}^3 y_i' = 0$ , e quindi  $\sum_{i=1}^3 y_i(t) = 1$ ) e due autovalori  $\lambda_1(t)$ ,  $\lambda_2(t)$  reali negativi che dipendono dalla soluzione e quindi variano nel tempo. Si può mostrare che sull'intervallo  $(0, 100)$  il rapporto di stiffness  $S(t)$ , calcolato a partire dai due autovalori  $\lambda_1$ ,  $\lambda_2$ , varia da  $O(10^4)$  a  $O(10^5)$ . Il comportamento della soluzione è rappresentato in Figura 7.28. Si osserva che la soluzione  $y_2(t)$  raggiunge rapidamente una posizione quasi-stazionaria in vicinanza a  $y_2' = 0$ , ossia, ponendo  $y_1 = 1$ ,  $y_3 = 0$ ,  $0.04 \approx 3 \cdot 10^7 y_2^2$ , da cui  $y_2 \approx 3.65 \cdot 10^{-5}$ . Successivamente, molto lentamente la soluzione  $y_2$  ritorna al valore zero; in effetti, si vede facilmente che per  $t \rightarrow \infty$  si ha  $y_1(t) \rightarrow 0$ ,  $y_2(t) \rightarrow 0$  e  $y_3(t) \rightarrow 1$ .

Lasciamo come esercizio il confronto tra i risultati ottenuti mediante il metodo di Eulero esplicito e rispettivamente il metodo di Eulero implicito, con utilizzo del metodo di Newton per la risoluzione ad ogni passo del sistema non lineare. ■

► **Esempio 7.16** (*Un esempio dalla biochimica*) Consideriamo il seguente modello di cinetica degli enzimi (Henri 1902)



che rappresenta una reazione tra un *enzima* ( $e^*$ ) e un *substrato* ( $s^*$ ) con la formazione di un complesso enzima-substrato ( $c^*$ ), che può reagire per formare il prodotto ( $p^*$ ) o dissociarsi.

<sup>21</sup> When the equations represent the behaviour of a system containing a number of fast and slow reactions, a forward integration of these equations becomes difficult (H. H. Robertson, 1966).

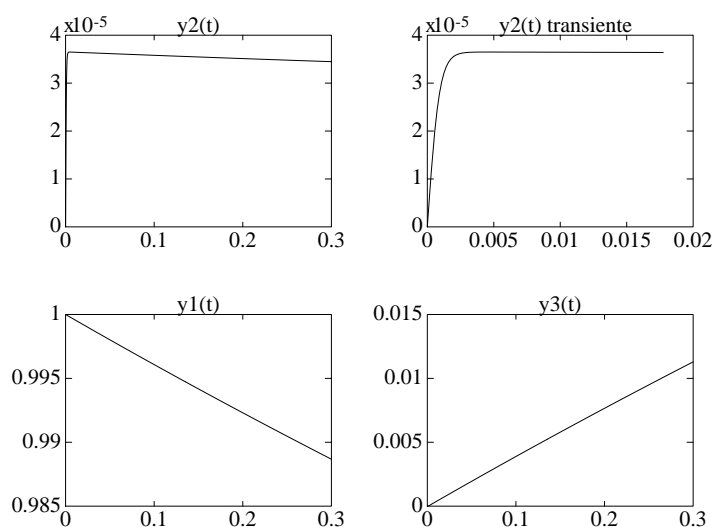


Figura 7.28: Rappresentazione della soluzione del sistema differenziale (7.82).

Dalla legge massa-azione, si ottiene il seguente modello differenziale

$$\begin{aligned}\frac{ds^*}{dt} &= -k_1 s^* e^* + k_{-1} c^* \\ \frac{de^*}{dt} &= -k_1 s^* e^* + (k_{-1} + k_2) c^* \\ \frac{dc^*}{dt} &= k_1 s^* e^* - (k_{-1} + k_2) c^* \\ \frac{dp^*}{dt} &= k_2 c^*\end{aligned}$$

Le condizioni sperimentali all'inizio della reazione ( $t = 0$ ) sono  $s^* = s^*(0)$ ,  $e^* = e^*(0)$ ,  $c^* = p^* = 0$ . Dal momento che dal sistema differenziale si ha  $de^*/dt + dc^*/dt = 0$ , e quindi  $e^* + c^* = e^*(0)$ , si può eliminare la variabile  $e^*$ , ottenendo il seguente sistema ridotto

$$\begin{aligned}\frac{ds^*}{dt^*} &= -k_1 e^*(0) s^* + (k_1 s^* + k_{-1}) c^* \\ \frac{dc^*}{dt^*} &= k_1 e^*(0) s^* - (k_1 s^* + k_{-1} + k_2) c^*\end{aligned}$$

Nella Tabella 7.9 sono riportati rapporti di stiffness  $S(0)$  corrispondenti ad alcuni valori, sperimentalmente importanti, delle costanti di velocità e dei valori iniziali. ■

### 7.4.1 Metodi numerici

Lo scopo di questo paragrafo è quello di fornire una breve panoramica dei metodi disponibili per la risoluzione di sistemi stiff della forma

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0 \quad (7.83)$$



caso	$k_1$	$k_{-1}$	$k_2$	$e^*(0)/s^*(0)$	$S(0)$
1	1.0E+05	1.0E+00	1.0E+00	1.0E-03	1.E+08
2	1.0E+05	1.0E+00	1.0E+00	1.0E+00	4.E+05
3	1.0E+05	1.0E+00	1.0E-03	1.0E+00	4.E+08
4	1.0E+05	1.0E+00	1.0E-02	1.0E-03	1.E+10

Tabella 7.9: Rapporti di stiffness, per  $t = 0$ , nella cinetica degli enzimi.

con  $\mathbf{y} \in \mathbb{R}^n$ ,  $t_0, \mathbf{y}_0$  assegnati e  $\mathbf{f}$  funzione assegnata a valori in  $\mathbb{R}^n$ .

Le considerazioni sviluppate nei paragrafi precedenti portano alla conclusione che i metodi convenienti per l'approssimazione dei sistemi stiff vanno ricercati tra quelli con un'ampia zona di stabilità a passo fisso. Metodi con tale proprietà possono essere costruiti utilizzando differenti idee. Per brevità, ci limiteremo a segnalare i metodi che rientrano nelle due classi di metodi considerati in precedenza. Per una più ampia descrizione dello stato dell'arte, si veda in particolare Hairer, Wanner [73], Lambert [104].

**Formule BDF** Le formule alle differenze all'indietro (BDF), introdotte nell'Esempio 7.11, sono particolari metodi a più passi della seguente forma

$$\boldsymbol{\eta}_i = \sum_{j=1}^r \alpha_j \boldsymbol{\eta}_{i-j} + h \beta_0 \dot{\boldsymbol{\eta}}_i$$

ove per brevità si è posto  $\dot{\boldsymbol{\eta}}_i = \mathbf{f}(t, \boldsymbol{\eta}_i)$  e ove  $r$ , numero di passi utilizzati, corrisponde anche all'ordine del metodo. Quando il passo utilizzato varia ad ogni passo, i coefficienti  $\alpha_j$  e  $\beta_0$  dipendono dai rapporti  $h_i/h_{i-1}, \dots, h_i/h_{i-r+1}$ . Il caso  $r = 1$  corrisponde al metodo di Eulero implicito per il quale, come abbiamo visto, non vi sono vincoli di stabilità. La stessa proprietà è soddisfatta dal metodo che si ottiene per  $r = 2$ , e anche dai metodi corrispondenti a  $r = 3, 4, 5, 6$  se applicati a sistemi con autovalori reali. Nel caso, invece, in cui il sistema (la matrice jacobiana) ha autovalori complessi (e quindi le soluzioni hanno componenti ad andamento oscillatorio) la zona di stabilità nel piano complesso si riduce all'aumentare di  $r$  da 3 a 6. Ricordiamo che per  $r > 6$  i metodi BDF non sono utilizzabili, in quanto non convergenti.

A partire dalle prime implementazioni delle formule BDF da parte di Gear (DIFSUB, 1968; STIFF, 1969), diverse altre versioni sono state successivamente sviluppate allo scopo di sfruttare meglio le nuove architetture dei calcolatori e la struttura (in particolare la presenza di sparsità) dei problemi. Ci limiteremo a segnalare ODEPACK (cfr. Hindmarsh [84]; per una descrizione si veda anche Comincioli [31]), importante raccolta di routine per la risoluzione di problemi stiff e non stiff, con possibilità di passare automaticamente tra metodi non stiff (Adams) e metodi stiff (BDF).

**Metodi Runge-Kutta** Come abbiamo già osservato in precedenza, i metodi di Runge-Kutta espliciti non sono appropriati per la risoluzione dei problemi stiff. Essi sono stati, tuttavia, generalizzati in forme opportune. Abbiamo già discusso i Runge-Kutta *impliciti*. Segnaliamo un'altra interessante generalizzazione, costituita dai cosiddetti metodi di Rosenbrock, in breve ROW. In maniera schematica, tali metodi utilizzano, oltre i valori della  $\mathbf{f}$ , anche i valori della jacobiana  $\partial \mathbf{f} / \partial \mathbf{y}$ , ossia le derivate seconde della funzione  $\mathbf{y}$ .

### 7.4.2 Sistemi altamente oscillatori

Consideriamo, come introduzione al tipo di problemi che esamineremo in questo paragrafo, il seguente problema modello

$$y''(t) + \lambda^2 y(t) = \lambda^2 \sin t, \quad t \in (0, T) \quad (7.84)$$

e la seguente famiglia di soluzioni

$$y(t) = c \sin \lambda t + \frac{\sin t}{1 - 1/\lambda^2} \quad (7.85)$$

rappresentata in Figura 7.29 per valori particolari di  $c$  e di  $\lambda$ . Si vede che per  $\lambda$  sufficientemente grande la soluzione consiste di un'onda portante ad alta frequenza  $c \sin \lambda t$ , modulata da un'onda lenta  $\sin t / (1 - 1/\lambda^2)$ .

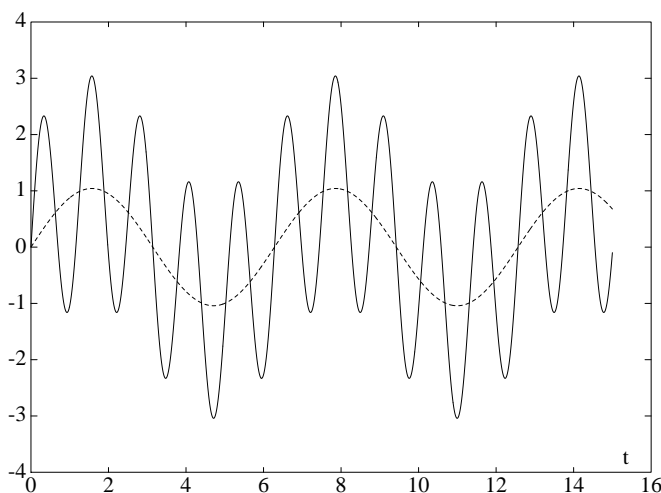


Figura 7.29: Rappresentazione di una soluzione dell'equazione differenziale (7.84), insieme alla soluzione particolare  $\sin t / (1 - 1/\lambda^2)$ .

Per questo tipo di problemi il calcolo della soluzione in un punto particolare è un problema *malcondizionato* (quanto più è grande  $\lambda$ ). Mentre nei problemi stiff che

abbiamo analizzato nel paragrafo precedente il cambiamento rapido nelle soluzioni è un fenomeno *transitorio*, nei problemi altamente oscillatori i cambiamenti rapidi sono un aspetto *permanente*. Le prestazioni dei metodi stiff possono quindi degradare notevolmente, a seguito della necessità di assumere frequentemente passi piccoli. D'altra parte, i metodi classici a un passo o a più passi sono basati sui valori della funzione e delle derivate, ossia su funzionali *instabili* della soluzione. L'idea per ottenere algoritmi convenienti può essere quindi quella di abbandonare l'obiettivo di ottenere informazioni sulla soluzione di tipo puntuale e introdurre in alternativa opportuni funzionali più stabili, quale ad esempio una media o uno smoothing della soluzione. Per una introduzione a questo tipo di algoritmi rinviamo ad esempio a Miranker [113].

## 7.5 Problemi ai limiti

Il tipo di problemi che analizzeremo in questo paragrafo è illustrato dal seguente problema particolare. Dato un intervallo  $[a, b]$ , una funzione  $f(x, y, y')$  e due valori  $\alpha, \beta$ , si cerca una funzione  $y(x)$  tale che

$$y''(x) = f(x, y(x), y'(x)), \quad a < x < b \quad (7.86a)$$

$$y(a) = \alpha, \quad y(b) = \beta \quad (7.86b)$$

Tale problema viene chiamato *problema ai limiti*; le condizioni (7.86b) sono le *condizioni ai limiti*. Esse, come vedremo successivamente in alcuni modelli, possono assumere forme differenti, ad esempio del tipo  $k_1 y(a) + k_2 y'(a) = \alpha$  e  $l_1 y(b) + l_2 y'(b) = \beta$ , o anche espressioni più complicate di tipo non lineare. In sostanza, ciò che caratterizza i problemi ai limiti, rispetto ai problemi a valori iniziali, è il fatto che la soluzione dell'equazione differenziale (7.86a) viene individuata imponendo che essa verifichi delle condizioni assegnate in *punti distinti*.

Nel caso dei problemi a valori iniziali abbiamo visto che una condizione di regolarità sulla  $f$  (continuità nel complesso delle variabili e lipschitzianità nella variabili  $y, y'$ ) assicura l'esistenza e l'unicità della soluzione e l'esistenza di opportuni metodi numerici. Mostriamo, ora, con un semplice esempio, che la situazione è differente per i problemi ai limiti; la sola regolarità non è, in generale, sufficiente perché il problema sia ben posto.

► **Esempio 7.17** L'equazione differenziale

$$y''(x) + y(x) = 0 \quad (7.87)$$

ha come soluzione generale

$$y = c_1 \sin x + c_2 \cos x, \quad c_1, c_2 \text{ costanti arbitrarie}$$

La soluzione particolare  $y(x) = \sin x$  è l'unica soluzione dell'equazione che verifica le seguenti condizioni ai limiti

$$y(0) = 0; \quad y\left(\frac{\pi}{2}\right) = 1$$

Osserviamo che in questo caso la soluzione  $y \equiv 0$  è l'unica soluzione del corrispondente *problema omogeneo*, definito dall'equazione (7.87) e dalle condizioni ai limiti  $y(0) = 0$  e  $y(\pi/2) = 0$ . Se consideriamo, invece, le condizioni ai limiti

$$y(0) = 0; \quad y(\pi) = 1$$

*non esiste* soluzione al problema; il corrispondente problema omogeneo ha infinite soluzioni  $c_1 \sin x$ . ■

### 7.5.1 Alcuni modelli

In questo paragrafo analizzeremo esempi di processi che possono essere studiati matematicamente mediante opportuni problemi differenziali ai limiti.

► **Esempio 7.18** (*Equazione della diffusione*) La diffusione è un processo nel quale la materia è trasportata da una parte di un sistema ad un'altra come risultato di un moto molecolare random. Il processo è analogo a quello relativo al trasferimento di calore. In effetti, un primo studio della diffusione da un punto di vista quantitativo è stato introdotto da Fick (1855), adattando le equazioni matematiche della conduzione del calore derivate alcuni anni prima da Fourier (1822).

La teoria matematica della diffusione nelle sostanze isotrope<sup>22</sup> è quindi basata sull'ipotesi che la velocità di trasferimento della sostanza che si diffonde attraverso l'area unitaria di una sezione sia proporzionale al gradiente della concentrazione misurato normale alla sezione, cioè

$$F = -D \frac{\partial C}{\partial x} \quad (7.88)$$

ove  $F$  è la velocità di trasferimento per unità di area di sezione,  $C$  la concentrazione della sostanza che si diffonde,  $x$  la coordinata spaziale misurata normalmente alla sezione;  $D$  è chiamato il *coefficiente di diffusione*. In alcuni casi, come ad esempio per la diffusione in soluzioni diluite,  $D$  può essere assunto costante, mentre in altri casi, ad esempio nella diffusione in polimeri, esso dipende dalla concentrazione. Se  $F$  e  $C$  sono espresse nella stessa unità di quantità, ad esempio grammi o grammi molecole, allora da (7.88) si ha che  $D$  ha la dimensione di  $l^2 t^{-1}$ , ad esempio  $\text{cm}^2 \text{s}^{-1}$ . Osserviamo, infine che il segno in (7.88) è negativo in quanto la diffusione avviene in direzione opposta a quella dell'aumento di concentrazione.

L'*equazione differenziale* fondamentale della diffusione in un mezzo isotropo può essere derivata dalla equazione (7.88), detta anche *legge di Fick*, nel seguente modo. Consideriamo (cfr. Figura 7.30) un elemento di volume nella forma di un parallelepipedo rettangolo con spigoli paralleli agli assi delle coordinate e di lunghezza  $2dx$ ,  $2dy$ ,  $2dz$ . Indichiamo con  $\mathbf{P}(x, y, z)$  il centro dell'elemento. In  $\mathbf{P}$  la concentrazione della sostanza che si diffonde è  $C(x, y, z)$ . Siano  $ABCD$  e  $A'B'C'D'$  le facce perpendicolari all'asse  $x$ .

<sup>22</sup>Un mezzo è isotropo quando la struttura e le proprietà di diffusione nell'intorno di ciascun punto sono le medesime in tutte le direzioni. Viceversa, in un mezzo anisotropo (non isotropo) le proprietà della diffusione dipendono dalla direzione nella quale esse sono misurate. Esempi di materiali anisotropi sono i cristalli e i polimeri, nei quali le molecole hanno una direzione preferenziale di orientamento.

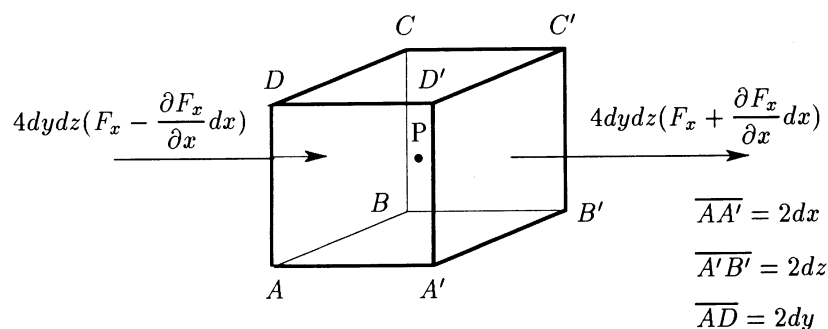


Figura 7.30: Illustrazione dell'equazione della diffusione.

La velocità alla quale la sostanza che diffonde entra nell'elemento attraverso la faccia  $ABCD$  nel piano  $x-dx$  è data da

$$4dydz \left( F_x - \frac{\partial F_x}{\partial x} dx \right)$$

ove  $F_x$  è la velocità di trasferimento attraverso l'unità di area del corrispondente piano attraverso  $P$ . In modo analogo la velocità di perdita della sostanza attraverso la faccia  $A'B'C'D'$  è data da  $4dydz \left( F_x + \left( \frac{\partial F_x}{\partial x} \right) dx \right)$ . Pertanto, il contributo alla velocità di aumento della sostanza nell'elemento dalle due facce è uguale a  $-8dxdydz \frac{\partial F_x}{\partial x}$ . In maniera analoga si mostra che il contributo relativo alle altre facce è dato rispettivamente da  $-8dxdydz \frac{\partial F_y}{\partial y}$  e  $-8dxdydz \frac{\partial F_z}{\partial z}$ . Tenendo allora conto che la velocità di aumento della sostanza nell'elemento è anche uguale a

$$8dxdydz \frac{\partial C}{\partial t}$$

si ha

$$\frac{\partial C}{\partial t} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} = 0 \quad (7.89)$$

Se il coefficiente di diffusione è costante, tenendo conto della legge di Fick (7.88), si ricava la seguente equazione, detta *equazione della diffusione*

$$\frac{\partial C}{\partial t} = D \left( \frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) \quad (7.90)$$

Quando durante il processo di diffusione, la sostanza che diffonde è modificata dall'esterno, esiste cioè una sorgente positiva, o negativa, l'equazione della diffusione viene modificata nel seguente modo. Indicata con  $f(x, y, z, t)$  la velocità di creazione, o di rimozione, per unità di volume, si ha

$$\frac{\partial C}{\partial t} = D \left( \frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) + f(x, y, z, t) \quad (7.91)$$

Nel caso in cui  $D$  non sia costante, ma dipenda dalla concentrazione della sostanza che diffonde  $C$ , o nel caso in cui il mezzo non sia omogeneo, l'equazione (7.90) diventa

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial C}{\partial x} \right) + \frac{\partial}{\partial y} \left( D \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial z} \left( D \frac{\partial C}{\partial z} \right) \quad (7.92)$$

con  $D = D(x, y, z, C)$ . Se  $D$  dipende durante la diffusione solo dal tempo e non dalle altre variabili, cioè  $D = g(t)$ , introducendo una nuova scala di tempi  $\tau$ , in modo che  $d\tau = g(t)dt$ , si ha

$$\frac{\partial C}{\partial \tau} = D \left( \frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) \quad (7.93)$$

Per semplicità, nel seguito considereremo in particolare il problema della diffusione in *una dimensione spaziale*. Fisicamente, il problema può corrispondere allo studio della diffusione in un foglio piano di spessore  $l$  (cfr. Figura 7.31). Sulle due superfici si ha una concentrazione  $C_0$  e  $C_1$ .

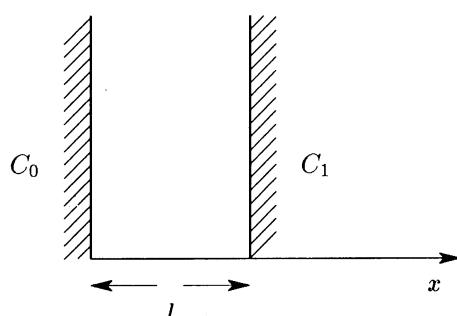


Figura 7.31: Diffusione attraverso un foglio infinito di spessore  $l$ .

Se supponiamo, inoltre, che la sostanza che diffonde possa essere immobilizzata da una reazione irreversibile del primo ordine, in modo che la velocità di rimozione della sostanza sia  $kC$ , dove  $k$  è una costante, allora l'equazione della diffusione in una dimensione diventa

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} - kC + f(x, t) \quad (7.94)$$

ove  $D$  è il coefficiente di diffusione, supposto costante e  $f(x, t)$  è una eventuale sorgente. L'equazione precedente è analoga all'equazione che rappresenta la conduzione del calore lungo un conduttore che perde calore attraverso la superficie ad una velocità proporzionale alla sua temperatura. In condizioni *stazionarie*, ossia quando la concentrazione risulta indipendente dal tempo, si ottiene un problema ai limiti del tipo (7.86), con  $C(0) = C_0$  e  $C(l) = C_1$ . Quando  $C_0 = C_1$  e  $D$  è costante, si può, osservando che vi è una simmetria attorno all'asse  $x = l/2$ , risolvere l'equazione (7.94) sull'intervallo  $[0, l/2]$ , con le condizioni ai limiti  $C(0) = C_0$  e  $\partial C / \partial x = 0$  in  $x = l/2$ . ■

► **Esempio 7.19** (*Equilibrio di una sbarra elastica*) Consideriamo una sbarra elastica fissata verticalmente ad un estremo (cfr. Figura 7.32). Indichiamo con  $u(x)$  lo *spostamento* (displacement); un punto inizialmente nella posizione  $x$ , che misura la distanza del punto dal vertice, si trova per effetto delle forze esterne, ad esempio di gravità, nella posizione  $x + u(x)$ . L'allungamento in ogni punto è misurato dalla derivata  $\epsilon = du/dx$ . Questo corrisponde all'allungamento di una molla ed è chiamato lo *strain* nella sbarra. Ove  $u$  è costante, la sbarra non è allungata e lo strain è zero. Altrimenti l'allungamento produce una forza interna  $\sigma(x)$ , detta *stress* e dipendente dallo strain. La relazione tra lo stress e lo strain

rappresenta l'*equazione costitutiva* del materiale. L'equazione più semplice corrisponde alla *legge di Hooke*, nella quale lo stress è supposto proporzionale allo strain  $\sigma(x) = E(x)du/dx$ . Il fattore di proporzionalità  $E(x)$  è chiamato *modulo di Young*<sup>23</sup>.

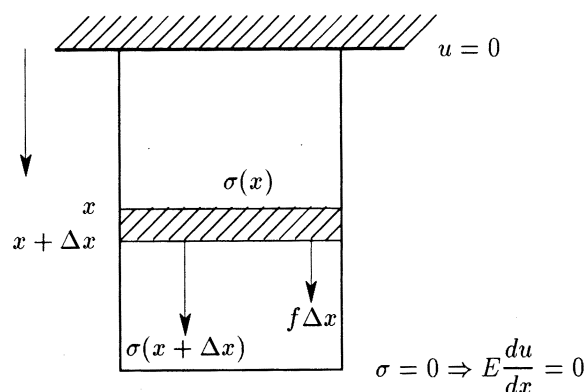


Figura 7.32: Equilibrio di una sbarra elastica.

Considerato un elemento infinitesimo  $\Delta x$ , imponiamo l'equilibrio tra la differenza delle *forze interne* ai due estremi e le *forze esterne*, di intensità  $f(x)$

$$\left( E(x) \frac{du}{dx} \right)_{x+\Delta x} - \left( E(x) \frac{du}{dx} \right)_x + f \Delta x = 0$$

Nel caso della forza di gravità si ha  $f = \rho a g$ , ove  $\rho$  è la densità del materiale,  $a$  la sezione trasversale e  $g$  la costante di gravità. Dividendo per  $\Delta x$ , si ottiene al limite per  $\Delta \rightarrow 0$  la seguente *equazione di equilibrio*

$$-\frac{d}{dx} \left( E(x) \frac{du}{dx} \right) = f(x) \quad (7.95)$$

Nell'estremo fissato si ha, poi,  $u = 0$ , mentre nell'estremo libero si ha  $\sigma = 0$ , cioè  $E du/dx = 0$ . Il *modello matematico* che approssima il problema fisico della determinazione dello spostamento, e quindi degli sforzi interni, è, allora rappresentato da un *problema differenziale ai limiti* del tipo (7.86). ■

### 7.5.2 Metodo shooting

Il *metodo shooting* riconduce la risoluzione di un problema ai limiti a quella di una successione di opportuni problemi a valori iniziali. Introduciamo l'*idea* del

<sup>23</sup>L'idea che lo stress in un corpo sia dipendente dallo strain fu espressa per la prima volta da R. Hooke (1635–1703) nel 1676 in forma di anagramma: *ceiinossttuv*. Il significato fu spiegato nel 1678 nella forma *ut tensio sic vis*.

metodo, considerando il problema ai limiti (7.86). Ad ogni valore di un *parametro*  $s$ , associamo la soluzione  $y(x; s)$  del seguente *problema a valori iniziali*

$$y''(x; s) = f(x, y(x; s), y'(x; s)) \quad (7.96a)$$

$$y(a; s) = \alpha, \quad y'(a; s) = s \quad (7.96b)$$

Non è detto, naturalmente, che tale soluzione esista su tutto l'intervallo  $[a, b]$ , per almeno un insieme non vuoto di valori del parametro  $s$ . Ciò dipenderà dalle proprietà della funzione data  $f$ . Per il momento ammetteremo tale ipotesi. Del resto, vedremo nel seguito che il metodo può essere applicato in diverse altre forme.

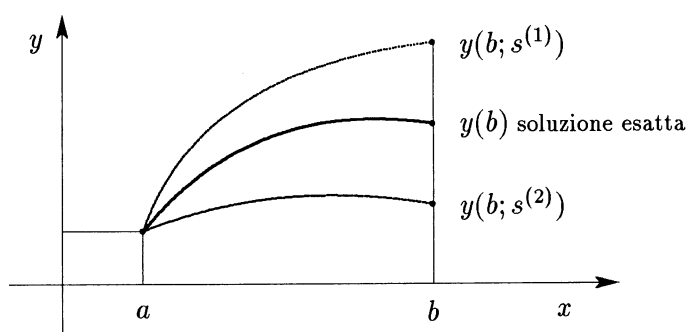


Figura 7.33: Metodo shooting.

La soluzione del problema ai limiti assegnato corrisponde, allora, al valore del parametro  $s$  che soddisfa alla seguente equazione

$$\boxed{y(b; s) = \beta} \quad \iff \quad \boxed{F(s) := y(b; s) - \beta = 0} \quad (7.97)$$

Per risolvere l'equazione (7.97) possiamo utilizzare ad esempio il *metodo della bisezione*. A questo proposito osserviamo che se la funzione  $f$  è lipschitziana, allora la funzione  $s \rightarrow y(x; s)$  è continua. Per applicare il metodo è sufficiente quindi trovare due valori di  $s$  in cui la funzione  $F(s)$  ha segno contrario (cfr. Figura 7.33).

In Figura 7.34 sono rappresentati i risultati ottenuti relativamente al seguente problema

$$\begin{aligned} y'' &= \frac{1}{8}(32 + 2x^3 - yy'), & x \in (1, 3) \\ y(1) &= 17; \quad y(3) = \frac{43}{3} \end{aligned} \quad (7.98)$$

che ha come soluzione esatta la funzione  $y(x) = x^2 + 16/x$ . Il valore del parametro  $s$  corrispondente alla soluzione esatta è dato da  $y'(1) = -14$ . In figura sono riportate le curve corrispondenti ai valori  $s = -1$  e  $s = -20$  e al punto di mezzo  $s = -10.5$ . Per ogni valore del parametro la soluzione del corrispondente problema a valori iniziali,



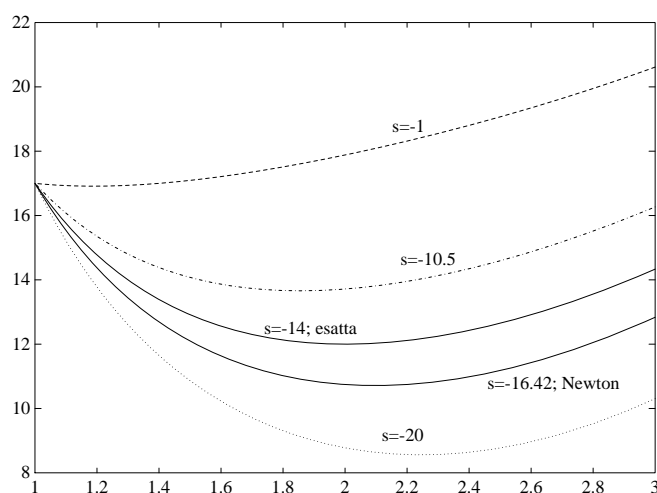


Figura 7.34: Applicazione del metodo shooting.

trasformato opportunamente in un sistema del primo ordine in due equazioni, è calcolata numericamente mediante la subroutine RKF45.

Nella Figura 7.34 è pure riportata la curva corrispondente al valore del parametro  $s = -16.42$  che è stato trovato a partire dal valore  $s = -1$  applicando il metodo di Newton all'equazione (7.97). La *derivata* della funzione  $F(s)$  è stata calcolata nel seguente modo. Abbiamo

$$F'(s) = \frac{\partial}{\partial s} y(b; s)$$

Supponendo  $f$  sufficientemente regolare e ponendo, per brevità  $v(x; s) = \partial y(x; s) / \partial s$ , si ha, derivando la (7.96a), il seguente sistema variazionale

$$\begin{cases} v'' = f_y(x, y, y') v + f_{y'}(x, y, y') v' \\ v(a) = 0; \quad v'(a) = 1 \end{cases} \quad (7.99)$$

Nel caso dell'esempio precedente si ottiene

$$v'' = -\frac{1}{8} (yv' + y'v), \quad v(1) = 0; \quad v'(1) = 1$$

In questo modo, la derivata  $F'(s)$  viene calcolata mediante la risoluzione di un problema a valori iniziali *ausiliario*. Osserviamo che tale problema è di tipo *lineare* e che la soluzione  $v(x; s)$  fornisce per ogni valore di  $x$  la *sensitività* della soluzione  $y$  rispetto al valore del parametro  $s$ . Naturalmente, in alternativa al metodo di Newton, si possono utilizzare i vari metodi che abbiamo analizzato nel Capitolo 5. Un particolare interesse può presentare il metodo delle secanti, che non richiede il calcolo esplicito della derivata.

Terminiamo sottolineando il fatto che per una buona riuscita del calcolo numerico della soluzione di un problema ai limiti mediante il metodo shooting è necessario *controllare* opportunamente i vari tipi di errore presenti nel calcolo, ossia: l'errore dovuto al metodo di risoluzione dell'equazione non lineare (7.97), l'errore nel calcolo numerico della soluzione dei problemi ai valori iniziali e infine gli errori di arrotondamento. Il controllo di questi errori è essenziale perché la soluzione trovata *abbia significato* e il *costo* del calcolo sia *ottimale*.

Nell'applicazione del metodo shooting, nella forma precedente, si possono trovare difficoltà quando la soluzione cresce troppo rapidamente. In questo caso la ricerca dello zero della funzione  $F(s)$  può risultare numericamente impossibile. Può, allora, risultare utile applicare il metodo su intervalli *più piccoli*. Il metodo che ne risulta è detto metodo *shooting multiplo*.

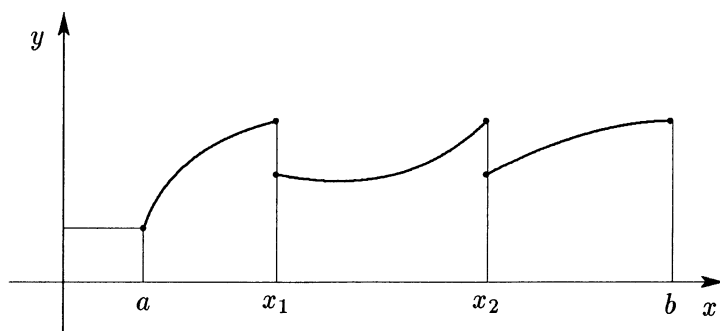


Figura 7.35: Illustrazione del metodo shooting multiplo.

### Shooting multiplo

Consideriamo una suddivisione dell'intervallo  $[a, b]$  mediante i punti (cfr. Figura 7.35)  $a = x_0 < x_1 < x_2 < \dots < x_m = b$  ed indichiamo con  $y(x; x_k, s_{1k}, s_{2k})$  le soluzioni dei seguenti problemi a valori iniziali

$$\begin{aligned} y'' &= f(x, y, y'), & x \in (x_k, x_{k+1}) \\ y(x_k) &= s_{1k}; & y'(x_k) = s_{2k} \end{aligned} \quad (7.100)$$

per  $k = 1, 2, \dots, m-1$ . Per la determinazione dei parametri  $s_{ik}$ ,  $i = 1, 2$ ,  $k = 1, 2, \dots, m$  imponiamo le seguenti condizioni

$$\left. \begin{aligned} s_{11} &= \alpha; & s_{1m} &= \beta & \text{condizioni ai limiti} \\ y(x_{k+1}, x_k, s_{1k}) &= s_{1k+1} \\ y'(x_{k+1}, x_k, s_{2k}) &= s_{2k+1} \end{aligned} \right\} \text{condizioni di regolarità}$$

che costituiscono un *sistema* di equazioni, in generale, non lineari, nelle incognite  $s_{ik}$ ,  $i = 1, 2$ ;  $k = 1, 2, \dots, m - 1$ .

### 7.5.3 Metodo alle differenze

Il *metodo alle differenze* è basato sulla suddivisione dell'intervallo  $[a, b]$  in sottointervalli  $[x_i, x_{i+1}]$  (cfr. Figura 7.36) e nell'approssimazione delle derivate mediante opportuni operatori alle differenze. Nel seguito, per semplicità, supporremo gli intervalli  $[x_i, x_{i+1}]$  di ampiezza uguale  $h = (b - a)/n$  e quindi i punti di suddivisione sono dati dalla relazione

$$x_i = a + ih, \quad i = 0, 1, \dots, n$$

Con  $\bar{y}_i$  indicheremo i valori che approssimano i valori  $y_i = y(x_i)$  della soluzione del problema continuo.

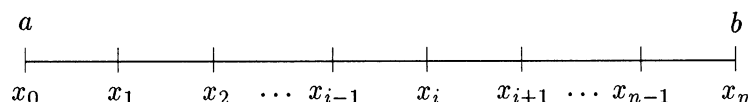


Figura 7.36: Suddivisione dell'intervallo  $[a, b]$  in  $n$  sottointervalli.

Analizzeremo il metodo nel caso del seguente problema modello lineare

$$-y''(x) + q(x)y(x) = f(x), \quad a < x < b \quad (7.101a)$$

$$y(a) = \alpha, \quad y(b) = \beta \quad (7.101b)$$

ove  $q(x)$  e  $f(x)$  sono funzioni assegnate, continue sull'intervallo  $[a, b]$ . Si può dimostrare che nell'ipotesi

$$q(x) \geq 0 \quad (7.102)$$

si ha per il problema (7.101) *esistenza e unicità* della soluzione. La soluzione, inoltre, è continua insieme alla derivata prima e seconda e l'eventuale maggiore regolarità dei dati  $q(x)$  e  $f(x)$  comporta una maggiore regolarità per la soluzione; ad esempio, se  $f(x)$  e  $q(x)$  hanno la derivata seconda, la soluzione ha la derivata quarta.

In corrispondenza ad ogni nodo  $x_i$ , si *discretizza* la derivata seconda mediante il seguente operatore alle differenze (cfr. Capitolo 4)

$$y''(x_i) \approx \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1}))}{h^2} \quad (7.103)$$

Ponendo tale approssimazione nell'equazione (7.101a) per  $i = 1, 2, \dots, n - 1$  e sostituendo  $y(x_i)$  con  $\bar{y}_i$ , si ottengono le seguenti equazioni lineari

$$-\frac{\bar{y}_{i+1} - 2\bar{y}_i + \bar{y}_{i-1}}{h^2} + q(x_i)\bar{y}_i = f(x_i), \quad i = 1, 2, \dots, n - 1 \quad (7.104)$$

o equivalentemente,

$$-\bar{y}_{i-1} + (2 + h^2 q_i) \bar{y}_i - \bar{y}_{i+1} = h^2 f_i, \quad i = 1, 2, \dots, n-1 \quad (7.105)$$

ove, per brevità, si è posto  $q_i = q(x_i)$ ,  $f_i = f(x_i)$ . In particolare, per  $i = 1$  e  $i = n-1$ , tenendo conto delle condizioni ai limiti, si ottengono le seguenti equazioni

$$\begin{aligned} (2 + h^2 q_1) \bar{y}_1 - \bar{y}_2 &= h^2 f_1 + \alpha \\ -\bar{y}_{n-2} + (2 + h^2 q_{n-1}) \bar{y}_{n-1} &= h^2 f_{n-1} + \beta \end{aligned}$$

Il *problema discreto* è, quindi, un *sistema lineare*, che può essere scritto in forma vettoriale nel seguente modo. Indicato con  $\bar{\mathbf{y}}$  il vettore delle incognite

$$\bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{n-2}, \bar{y}_{n-1}]^T$$

con  $\mathbf{A}$  la matrice dei coefficienti

$$\mathbf{A} = \begin{bmatrix} (2 + h^2 q_1) & -1 & & & & 0 \\ -1 & (2 + h^2 q_2) & -1 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & (2 + h^2 q_{n-2}) & -1 \\ 0 & & & & -1 & (2 + h^2 q_{n-1}) \end{bmatrix}$$

e con  $\mathbf{b}$  il vettore dei termini noti

$$\mathbf{b} = [h^2 f_1 + \alpha, h^2 f_2, \dots, h^2 f_{n-1}, h^2 f_{n-1} + \beta]^T$$

si ha

$$\mathbf{A} \bar{\mathbf{y}} = \mathbf{b} \quad (7.106)$$

Osserviamo che la matrice  $\mathbf{A}$  è, nell'ipotesi  $q(x) \geq 0$ , a predominanza diagonale, stretta per  $i = 1$  e  $i = n-1$ . Tale proprietà assicura la non singolarità della matrice e quindi l'*esistenza* e l'*unicità* della soluzione del sistema lineare (7.106). Il modello discreto è, allora, risolubile nelle stesse ipotesi del problema continuo. La proprietà di predominanza diagonale permette, inoltre, l'applicazione del metodo di eliminazione senza pivoting. Tenendo conto che la matrice è tridiagonale, mediante la decomposizione  $\mathbf{A} = \mathbf{L}\mathbf{U}$  si può risolvere il sistema lineare con un numero di operazioni di ordine  $O(n)$ .

La soluzione discreta  $\bar{y}_i$ ,  $i = 0, 1, \dots, n$  dipende dal passo di discretizzazione  $h$ . Una questione importante è, allora, l'analisi della *convergenza* per  $h \rightarrow 0$ . In pratica, si tratta di vedere se è possibile approssimare la soluzione continua, ad una precisione prefissata, per  $h$  sufficientemente piccolo. L'approssimazione dovrebbe essere di tipo puntuale, dal momento che la soluzione del problema ai limiti è una funzione continua su  $[a, b]$ . Lo studio della convergenza è introdotto dal seguente semplice esempio.

► **Esempio 7.20** Consideriamo il problema

$$-y''(x) + y(x) = 0, \quad y(0) = 0, \quad y(1) = \sinh(1)$$

che ha come soluzione analitica la funzione  $y(x) = \sinh(x)$ . Per  $n = 4$ , e quindi  $h = 0.25$ , il problema discreto consiste nella risoluzione di un sistema  $\mathbf{A}\bar{\mathbf{y}} = \mathbf{b}$  nelle incognite  $\bar{y}_1, \bar{y}_2, \bar{y}_3$ , con

$$\mathbf{A} = \begin{bmatrix} (2+h^2) & -1 & 0 \\ -1 & (2+h^2) & -1 \\ 0 & -1 & (2+h^2) \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \sinh(1) \end{bmatrix}$$

La soluzione è riportata nella Tabella 7.10, insieme alla soluzione ottenuta per  $n = 8$ . I risultati mostrano che, quando il passo è dimezzato, l'errore è ridotto approssimativamente di un fattore quattro. ■

$x_i$	$h = 0.25$		$h = 0.125$	
	$\bar{y}_i$	$\bar{y}_i - y_i$	$\bar{y}_i$	$\bar{y}_i - y_i$
0.125			0.1253508	$2.50 \cdot 10^{-5}$
0.250	0.2528031	$1.90 \cdot 10^{-4}$	0.2526602	$4.79 \cdot 10^{-5}$
0.375			0.3839175	$6.64 \cdot 10^{-5}$
0.500	0.5214064	$3.11 \cdot 10^{-4}$	0.5211735	$7.82 \cdot 10^{-5}$
0.625			0.6665728	$8.05 \cdot 10^{-5}$
0.750	0.8225976	$2.80 \cdot 10^{-4}$	0.8223873	$7.06 \cdot 10^{-5}$
0.875			0.9910516	$4.50 \cdot 10^{-5}$

Tabella 7.10: Soluzioni ottenute mediante il metodo delle differenze finite per il problema ai limiti  $-y'' + y = 0$ ,  $y(0) = 0$ ,  $y(1) = \sinh(1)$ .

Il seguente teorema mostra in effetti che, se la soluzione  $y(x)$  è sufficientemente regolare, l'errore di troncamento è dell'ordine  $O(h^2)$ .

**Teorema 7.4 (Convergenza)** Sia  $y(x)$  la soluzione del problema ai limiti

$$-y''(x) + q(x)y(x) = f(x), \quad y(a) = \alpha, \quad y(b) = \beta$$

con  $q(x)$  e  $f(x)$  funzioni continue su  $[a, b]$  e  $q(x) \geq 0$ . Supponiamo, inoltre, che la funzione  $y(x)$  abbia la derivata quarta, con  $|y^{(4)}| \leq M$  per ogni  $x \in [a, b]$ . Se  $\bar{y}_i$  indica l'approssimazione di  $y(x_i)$  ottenuta con il metodo alle differenze, allora

$$|y(x_i) - \bar{y}_i| \leq \frac{Mh^2}{24} (x_i - a)(b - x_i) \quad i = 1, 2, \dots, n-1 \quad (7.107)$$

La dimostrazione del teorema segue lo schema usuale che abbiamo utilizzato nel caso dei problemi ai valori iniziali. Si studia l'errore di discretizzazione locale, cioè l'errore commesso quando in ogni nodo  $x_i$  si sostituisce l'equazione differenziale con

l'equazione alle differenze. Quando, per  $h \rightarrow 0$ , tale errore tende a zero, lo schema è detto *consistente*. Si mostra, quindi, che gli errori locali si mantengono limitati per  $h \rightarrow 0$ , cioè che lo schema numerico è *stabile*. La dimostrazione si basa su una procedura, nota come *principio del massimo*. La consistenza e la stabilità forniscono, allora, la convergenza del metodo.

Terminiamo con due esempi che mostrano ulteriori applicazioni del metodo.

► **Esempio 7.21** (*Diffusione e convezione*) Analizziamo il metodo alle differenze finite applicato al seguente problema ai limiti

$$\begin{cases} y''(x) - p(x)y'(x) - q(x)y = f(x), & a < x < b \\ y(a) = \alpha, \quad y(b) = \beta \end{cases}$$

ove  $p(x)$ ,  $q(x)$ ,  $f(x)$  sono funzioni continue su  $[a, b]$  con  $q(x) \geq 0$ . Si può dimostrare che in tali ipotesi il problema ammette una ed una sola soluzione, continua insieme alle derivate dei primi due ordini. Per  $h = (b-a)/n$ , per  $n$  intero assegnato, consideriamo il seguente schema alle differenze ottenuto utilizzando l'operatore alle differenze centrali

$$\frac{\bar{y}_{j+1} - 2\bar{y}_j + \bar{y}_{j-1}}{h^2} - p(x_j)\frac{\bar{y}_{j+1} - \bar{y}_{j-1}}{2h} - q(x_j)\bar{y}_j = f(x_j)$$

per  $j = 1, 2, \dots, n-1$ . Le condizioni ai limiti sono date da  $\bar{y}_0 = \alpha$ ,  $\bar{y}_n = \beta$ . Il problema discreto è equivalente ad un sistema lineare con matrice dei coefficienti tridiagonale. Indicando con  $a_j$  gli elementi sulla diagonale principale e con  $-c_j$ ,  $-b_j$  gli elementi rispettivamente sulla diagonale superiore e inferiore, si ha

$$a_j := 1 + \frac{h^2}{2}q(x_j), \quad b_j := \frac{1}{2}\left[1 + \frac{h}{2}p(x_j)\right], \quad c_j := \frac{1}{2}\left[1 - \frac{h}{2}p(x_j)\right]$$

Se facciamo la seguente ipotesi

$$\frac{h}{2}|p(x_j)| \leq 1 \quad j = 1, 2, \dots, n-1 \quad (7.108)$$

allora si ha  $|b_j| + |c_j| = b_j + c_j = 1$  e la matrice dei coefficienti risulta a *prevalenza diagonale*. La matrice è pertanto *non singolare* e il *problema approssimato ammette una ed una sola soluzione*.

La limitazione (7.108) può essere eliminata, *discretizzando in maniera differente* il termine  $p(x)y'$ . Più precisamente, poniamo

$$\begin{aligned} \frac{\bar{y}_{j+1} - 2\bar{y}_j + \bar{y}_{j-1}}{h^2} - p(x_j)\frac{\bar{y}_{j+1} - \bar{y}_j}{h} - q(x_j)\bar{y}_j &= f(x_j), & \text{se } p(x_j) \leq 0 \\ \frac{\bar{y}_{j+1} - 2\bar{y}_j + \bar{y}_{j-1}}{h^2} - p(x_j)\frac{\bar{y}_j - \bar{y}_{j-1}}{h} - q(x_j)\bar{y}_j &= f(x_j), & \text{se } p(x_j) > 0 \end{aligned}$$

In maniera compatta, lo schema precedente, noto come *metodo upwind difference*, può essere scritto nella seguente forma

$$\frac{\bar{y}_{j+1} - 2\bar{y}_j + \bar{y}_{j-1}}{h^2} + \frac{(|p_j| - p_j)\bar{y}_{j+1} - 2|p_j|\bar{y}_j + (|p_j| + p_j)\bar{y}_{j-1}}{2h} - q_j\bar{y}_j = f_j$$

Lo schema upwind è meno accurato del precedente, ma è risolvibile senza restrizioni sul passo  $h$ . Infatti, la matrice del corrispondente sistema lineare è, per ogni  $h$ , a *predominanza diagonale*. Questo significa che l'operatore discreto verifica per ogni valore del parametro di discretizzazione  $h$  un principio del massimo, analogo a quello verificato dall'operatore continuo. Più precisamente, se  $f(x) = 0$ ,  $\alpha = \beta = 0$  e  $q(x) \geq 0$ , una soluzione numerica  $\bar{y}_i$  ottenuta con il metodo upwind difference non può assumere un massimo positivo o un minimo negativo nei nodi interni  $x_j$ ,  $j = 1, 2, \dots, n-1$ , e quindi è identicamente nulla. Per il metodo alle differenze centrali, visto in precedenza, questo risultato è vero nell'ipotesi (7.108). ■

► **Esempio 7.22** (*Problema ai limiti non lineare*) Approssimiamo mediante il metodo delle differenze finite la soluzione del seguente problema ai limiti non lineare

$$\begin{cases} y''(x) = \frac{1}{2}(1+x+y)^3, & x \in (0, 1) \\ y(0) = y(1) = 0 \end{cases}$$

Posto  $f(x, y) := \frac{1}{2}(1+x+y)^3$ , si ha

$$\frac{\partial f}{\partial y} = \frac{3}{2}(1+x+y)^2 \geq 0$$

Si può, allora, mostrare che il problema ai limiti assegnato ammette una ed una sola soluzione. Per applicare il metodo alle differenze finite discretizziamo l'intervallo  $(0, 1)$  mediante i punti  $x_i = ih$ , con  $h = 1/(n+1)$  e  $i = 0, 1, \dots, n+1$ . Indichiamo con  $\bar{y}_i$  i valori della soluzione approssimata nei punti  $x_i$ . Discretizzando la derivata seconda  $y''$  mediante le differenze centrali, si ottiene il seguente sistema non lineare

$$\begin{cases} \bar{y}_0 = 0, \quad \bar{y}_{n+1} = 0 \\ -\frac{\bar{y}_{i+1} - 2\bar{y}_i + \bar{y}_{i-1}}{h^2} + \frac{1}{2}(1+x_i + \bar{y}_i)^3 = 0 \end{cases}$$

Eliminando le variabili  $\bar{y}_0, \bar{y}_{n+1}$ , il sistema può essere scritto nella seguente forma matriciale

$$\mathbf{A}\bar{\mathbf{y}} + h^2\mathbf{B}(\bar{\mathbf{y}}) = 0 \quad (7.109)$$

con  $\bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n]^T$  e

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}, \quad \mathbf{B}(\mathbf{y}) = \text{diag}(f(x_i, \bar{y}_i)), \quad i = 1, 2, \dots, n$$

Il sistema *non lineare* (7.109) può essere risolto mediante, ad esempio, il metodo di Newton. Infatti, la matrice jacobiana del sistema, data da

$$\mathbf{J} = \mathbf{A} + h^2\mathbf{B}_y$$

è, grazie alla positività della funzione  $f_y$ , una matrice a predominanza diagonale, ed anche definita positiva. Il metodo di Newton è pertanto convergente per scelte di valori iniziali sufficientemente vicini alla soluzione. L'implementazione del metodo è la seguente. Si sceglie una stima iniziale  $\bar{\mathbf{y}}^{(0)}$  del vettore soluzione; supposta, quindi, nota la iterata  $\bar{\mathbf{y}}^{(r)}$ , si calcola la successiva iterata  $\bar{\mathbf{y}}^{(r+1)}$  risolvendo il seguente sistema lineare nel vettore incognito  $\delta$

$$(\mathbf{A} + h^2 \mathbf{B}_y(\bar{\mathbf{y}}^{(r)})) \delta = -\mathbf{A}\bar{\mathbf{y}}^{(r)} - h^2 \mathbf{B}(\bar{\mathbf{y}}^{(r)})$$

e ponendo  $\bar{\mathbf{y}}^{(r+1)} = \bar{\mathbf{y}}^{(r)} + \delta$ . Al variare di  $n$  si ottengono i risultati riportati nella seguente tabella. Come si verifica facilmente, la soluzione esatta del problema continuo è data da  $y(x) = 2/(2-x) - x - 1$ . L'errore riportato  $E_\infty$  è pertanto la distanza nella norma del massimo della soluzione discreta dalla soluzione continua. Come vettore iniziale  $\bar{\mathbf{y}}^{(0)}$  è stato assunto il vettore nullo.

$n$	$E_\infty$
10	0.60698591 E-03
20	0.16806671 E-03
30	0.07715980 E-03
40	0.04418134 E-03
50	0.02855734 E-03
60	0.01996486 E-03
70	0.01473998 E-03
80	0.01132411 E-03

I risultati mostrano una convergenza per  $h \rightarrow 0$ . In effetti, la convergenza può essere dimostrata, grazie alla monotonia del termine non lineare  $f(x, y)$ , con tecniche analoghe a quelle utilizzate per i problemi ai limiti lineari. ■

#### 7.5.4 Metodo degli elementi finiti

Il *metodo degli elementi finiti* è stato introdotto, in particolare, per la risoluzione di problemi in ingegneria meccanica. In maniera schematica, l'idea di base consiste nella suddivisione della struttura da studiare, ad esempio una trave o una piastra elastica, in piccole parti (gli *elementi finiti*): triangoli, quadrilateri, in due dimensioni, tetraedri, parallelepipedi, in tre dimensioni (cfr. per un esempio Figura 7.37 [125]). Si descrive quindi, mediante opportune equazioni, come un carico applicato alla struttura influenza ciascuna parte. In pratica, la funzione incognita, lo sforzo o la deformazione, viene approssimata su ogni parte mediante un polinomio di interpolazione, individuato da un numero finito di gradi di libertà corrispondenti a punti particolari (*nodi*), scelti opportunamente nel singolo elemento finito; ad esempio, per un triangolo e per un polinomio di primo grado, i nodi possono essere i vertici del triangolo. L'equazione su ogni elemento si riduce, pertanto, in generale, ad una equazione algebrica in cui le incognite sono i valori del polinomio nei nodi. Tenendo, poi, conto che ogni elemento dipende dagli elementi vicini, e quindi *assemblando* i



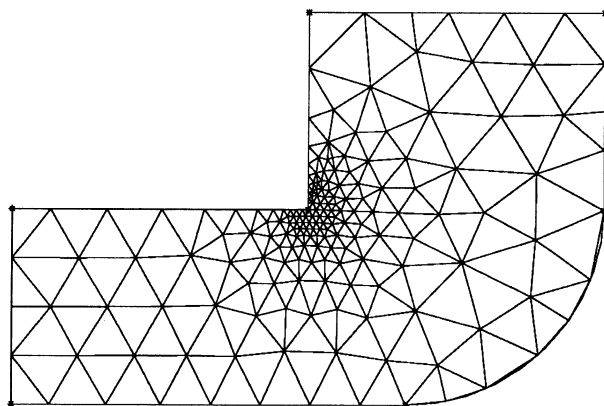


Figura 7.37: Esempio di reticolazione per l'applicazione del metodo degli elementi finiti.

vari elementi finiti che costituiscono la struttura, si ottiene un sistema di equazioni che approssima l'effetto totale del carico sulla struttura.

L'idea può, naturalmente, essere applicata all'analisi di altri fenomeni, in particolare, per lo studio della diffusione. In effetti, il metodo degli elementi finiti è diventato uno dei metodi più importanti per l'approssimazione della soluzione delle equazioni alle derivate parziali. Dal punto di vista matematico, esso può essere considerato una particolare forma di applicazione di un metodo generale, noto come *metodo di Galerkin* o anche *metodo di Rayleigh-Ritz*, che utilizza una *formulazione integrale* del problema, corrispondente al *principio dei lavori virtuali* o al *principio del minimo dell'energia*<sup>24</sup>.

Lo scopo di quanto segue non può essere, naturalmente, una trattazione adeguata del metodo. Semplicemente, sarà una introduzione agli aspetti matematici e numerici di base. Come problema di riferimento, considereremo un problema ai limiti in una dimensione, sottolineando, comunque, che l'interesse del metodo è, in particolare, per i problemi in più dimensioni.

Dato il seguente problema ai limiti *modello*

$$-y''(x) + q(x)y(x) = f(x), \quad a < x < b \quad (7.110a)$$

$$y(a) = 0, \quad y(b) = 0 \quad (7.110b)$$

con  $q(x)$  e  $f(x)$  funzioni continue sull'intervallo  $[a, b]$  e  $q(x) \geq 0$  per  $x \in [a, b]$ , consideriamo una sua formulazione in forma integrale. A tale scopo, indichiamo, per

<sup>24</sup>Galerkin, matematico russo, utilizzò il metodo per lo studio della deformazione di sbarre elastiche mentre era in carcere per le sue idee politiche (1906). Ritz sviluppò il metodo nel 1909 per la soluzione di problemi di equilibrio, e Lord Rayleigh nel 1870 per la soluzione di problemi di vibrazioni.

brevità, con  $L$  l'operatore che trasforma una funzione  $y$  due volte derivabile nella funzione continua  $-y''(x) + q(x)y$  e introduciamo il seguente spazio di funzioni, detto *spazio delle funzioni test*

$$V = \{v \mid v' \text{ continua a tratti e limitata in } [a, b] \text{ e } v(a) = v(b) = 0\}$$

Inoltre, per due qualunque funzioni  $h(x), g(x)$  integrabili in  $[a, b]$  definiamo il seguente prodotto scalare

$$(f, g) := \int_a^b h(x)g(x) dx$$

Se  $y$  è soluzione dell'equazione (7.110a), allora, si ha

$$(Ly - f, v) = 0 \quad \forall v \in V$$

e viceversa, se  $y$  è tale che

$$(Ly, v) = (f, v) \quad \forall v \in V \tag{7.111}$$

allora  $Ly = f$ . La formulazione (7.111) esprime, in sostanza, il fatto che la *funzione residuo*  $r = Ly - f$  è ortogonale a tutte le funzioni in  $V$ . Possiamo, ora, trasformare opportunamente l'equazione (7.111) mediante integrazione per parti

$$\begin{aligned} (Ly, v) &= - \int_a^b y''(x)v(x)dx + \int_a^b q(x)y(x)v(x) dx \\ &= [-y'(x)v(x)]_a^b + \int_a^b y'(x)v'(x) dx + \int_a^b q(x)y(x)v(x) dx \\ &= \int_a^b y'(x)v'(x) dx + \int_a^b q(x)y(x)v(x) dx \end{aligned}$$

ove si è tenuto conto che per una funzione  $v \in V$  si ha  $v(a) = v(b) = 0$ . Si ha, allora, la seguente formulazione del problema, detta *formulazione debole* del problema ai limiti assegnato.

▼ **Problema 7.1** (Formulazione debole) *Si cerca  $y \in V$  tale che*

$$(y', v') + (qy, v) = (f, v) \tag{7.112}$$

per ogni  $v \in V$ .

▼ **Osservazione 7.3** *La formulazione (7.112) è detta debole, in quanto in essa, a differenza della formulazione (7.110a), non è richiesto che la funzione  $y$  abbia la derivata seconda, ma soltanto che il prodotto  $y'v'$  sia integrabile. Ugualmente, non è necessario supporre che le funzioni  $q(x)$  e  $f(x)$  siano funzioni continue. Per valutare l'importanza nelle applicazioni di questo fatto, si pensi, ad esempio, al problema relativo alla deformazione di una trave quando il carico  $f(x)$ , come avviene spesso, è una funzione discontinua, ad esempio a gradini, e quindi non derivabile. ■*

La formulazione (7.112) corrisponde al *principio dei lavori virtuali*. Si può mostrare che nel caso del problema particolare considerato esso è equivalente al seguente *principio del minimo dell'energia*. Posto, per brevità  $a(y, v) := (y', v') + (qy, v)$  e definito il seguente funzionale, che rappresenta l'energia totale del sistema

$$J(v) = \frac{1}{2}a(v, v) - (f, v), \quad \forall v \in V$$

si ha che la soluzione  $y$  dell'equazione (7.112) corrisponde al minimo di  $J(v)$ , cioè

$$J(y) \leq J(v), \quad \forall v \in V$$

e viceversa. Motivo di tale equivalenza è il fatto che per ogni  $v, w \in V$  si ha  $a(v, w) = a(w, v)$ , cioè, come anche si dice, la forma bilineare  $a(v, w)$  è *simmetrica*. L'*unicità* della soluzione del problema (7.112) si dimostra facilmente, per il fatto che, essendo  $q(x) \geq 0$ , da  $a(v, v) = 0$  si ricava immediatamente  $v \equiv 0$ . Si può, infatti, dimostrare che esiste una costante  $\alpha > 0$  tale che

$$a(v, v) \geq \alpha \|v\|_1^2 \quad \forall v \in V \quad (7.113)$$

ove  $\|v\|_1^2 = (v, v) + (v', v')$ . L'*esistenza* della soluzione può essere dimostrata in uno spazio di funzioni più generale dello spazio  $V$ , il cui studio esula dagli scopi del presente volume. Ci limiteremo ad osservare che, nel caso in cui le funzioni  $q(x)$  e  $f(x)$  siano funzioni continue, la soluzione del problema debole ha la derivata seconda continua, e quindi coincide con la soluzione del problema (7.110).

Ora ci interesseremo dell'*approssimazione numerica* del problema (7.112). In forma generale, sia  $V_h$  uno sottospazio di dimensione *finita* dello spazio  $V$  e  $\{\phi_i\}_{i=1}^{n-1}$  una corrispondente *base*. In particolare, si avrà  $\phi_i(a) = \phi_i(b) = 0$ . Consideriamo, allora, il seguente problema.

▼ **Problema 7.2** (Problema discreto) *Si cerca  $y_h \in V_h$  che verifica l'equazione*

$$a(y_h, v_h) = (f, v_h) \quad \forall v_h \in V_h \quad (7.114)$$

Mostreremo, ora, che la condizione (7.114) è equivalente alla risoluzione di un *sistema lineare*. A tale scopo, basta osservare che per definizione di base si può scrivere

$$y_h = \sum_{j=1}^{n-1} c_j \phi_j$$

e che la relazione (7.114) è verificata per ogni  $v_h \in V_h$  se lo è per tutti gli elementi  $\phi_i$  della base. Si ottengono, quindi, le seguenti equazioni lineari

$$\sum_{j=1}^{n-1} c_j (\phi_j', \phi_i') + \sum_{j=1}^{n-1} c_j (q\phi_j, \phi_i) = (f, \phi_i), \quad i = 1, 2, \dots, n-1$$

ossia un sistema lineare nei coefficienti  $c_j$

$$\mathbf{A}\mathbf{c} = \mathbf{b} \quad (7.115)$$

La matrice  $\mathbf{A}$  è la somma di due matrici  $\mathbf{A}_1 + \mathbf{A}_2$ , ove  $\mathbf{A}_1$ , detta anche *matrice di rigidità* (stiffness) è data da

$$\mathbf{A}_1 = \begin{bmatrix} (\phi'_1, \phi'_1) & (\phi'_1, \phi'_2) & \cdots & (\phi'_1, \phi'_{n-1}) \\ (\phi'_2, \phi'_1) & (\phi'_2, \phi'_2) & \cdots & (\phi'_2, \phi'_{n-1}) \\ \vdots & \vdots & \ddots & \vdots \\ (\phi'_{n-1}, \phi'_1) & (\phi'_{n-1}, \phi'_2) & \cdots & (\phi'_{n-1}, \phi'_{n-1}) \end{bmatrix}$$

e  $\mathbf{A}_2$ , detta *matrice di massa* (mass) è

$$\mathbf{A}_2 = \begin{bmatrix} (q\phi_1, \phi_1) & (q\phi_1, \phi_2) & \cdots & (q\phi_1, \phi_{n-1}) \\ (q\phi_2, \phi_1) & (q\phi_2, \phi_2) & \cdots & (q\phi_2, \phi_{n-1}) \\ \vdots & \vdots & \ddots & \vdots \\ (q\phi_{n-1}, \phi_1) & (q\phi_{n-1}, \phi_2) & \cdots & (q\phi_{n-1}, \phi_{n-1}) \end{bmatrix}$$

e il vettore dei termini noti  $\mathbf{b}$  è dato da

$$\mathbf{b} = [b_i], \quad b_i = (f, \phi_i), \quad i = 1, 2, \dots, n-1$$

La matrice  $\mathbf{A}$  è *simmetrica*, grazie alla simmetria della forma  $a(y, v)$ . Si può anche mostrare che è *definita positiva*, grazie alla (7.113), per cui il sistema lineare (7.115) ammette una ed una sola soluzione.

► **Esempio 7.23** Come esemplificazione, possiamo assumere come spazio  $V_h$  lo spazio delle funzioni spline di primo ordine corrispondenti ad una suddivisione in parti uguali dell'intervallo  $[a, b]$ . Posto  $h = (b - a)/n$ , si hanno i nodi  $x_j = a + jh$ ,  $j = 0, 1, \dots, n$ , e le funzioni di  $V_h$  sono, in particolare, nulle nei nodi  $x_0$  e  $x_n$ , lineari su ogni intervallo  $[x_i, x_{i+1}]$ ,  $i = 0, 1, \dots, n-1$  e continue su  $[a, b]$ . Le derivate di tali funzioni sono funzioni costanti su ogni intervallo  $[x_j, x_{j+1}]$ . Come base possiamo assumere la base canonica, nella quale ogni elemento  $\phi_j$ , per  $j = 1, 2, \dots, n-1$ , corrisponde alla particolare spline lineare che vale 1 nel punto  $x_j$  e 0 nei nodi  $x_i$ , con  $i \neq j$  (cfr. Figura 7.38).

Nella costruzione della matrice  $\mathbf{A}$  osserviamo che

$$(\phi'_j, \phi'_k) = 0, \quad (\phi_j, \phi_k) = 0, \quad \text{per } |j - k| \geq 2$$

La matrice è, quindi, *tridiagonale*. Gli elementi sulla diagonale principale sono dati da

$$(\phi'_j, \phi'_j) = \int_{x_{j-1}}^{x_j} \left(\frac{1}{h}\right)^2 dx + \int_{x_j}^{x_{j+1}} \left(\frac{1}{h}\right)^2 dx = \frac{2}{h}$$

mentre quelli della sottodiagonale sono

$$(\phi'_{j-1}, \phi'_j) = \int_{x_{j-1}}^{x_j} \left(\frac{1}{h}\right) \left(-\frac{1}{h}\right) dx = -\frac{1}{h}$$

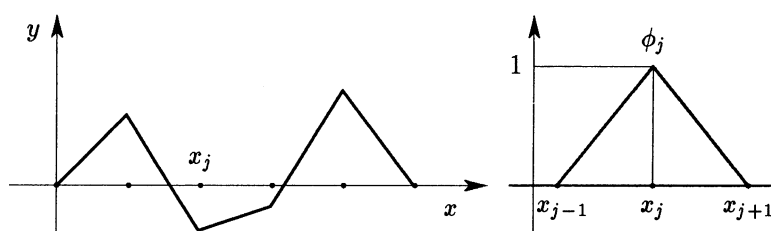


Figura 7.38: Esempio di funzione  $v_h \in V_h$  e di funzione base  $\phi_j$ .

Allora, la matrice di rigidità  $\mathbf{A}_1$  è

$$\mathbf{A}_1 = \frac{1}{h} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & \\ & & & & -1 & 2 \end{bmatrix}$$

In modo analogo, supponendo per semplicità  $q(x)$  costante, per la matrice di massa si ottiene

$$\begin{aligned} (\phi_j, \phi_j) &= q \int_{x_{j-1}}^{x_j} \left( \frac{x - x_{j-1}}{h} \right)^2 dx + q \int_{x_j}^{x_{j+1}} \left( \frac{x_{j+1} - x}{h} \right)^2 dx = q \frac{2h}{3} \\ (\phi_{j-1}, \phi_j) &= q \int_{x_{j-1}}^{x_j} \frac{x_j - x}{h} \frac{x - x_{j-1}}{h} dx = q \frac{h}{6} \end{aligned}$$

e, quindi

$$\mathbf{A}_2 = \frac{qh}{6} \begin{bmatrix} 4 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 4 \end{bmatrix}$$

Come applicazione, consideriamo il seguente problema ai limiti

$$-y'' + y = 1, \quad y(0) = y(1) = 0$$

che ammette come soluzione esatta la funzione  $y(x) = c_1 e^x + c_2 e^{-x} + 1$ , con  $c_1 = (1 - e)/(e^2 - 1)$  e  $c_2 = (e - e^2)/(e^2 - 1)$ . In Figura 7.39 sono riportati i risultati ottenuti con il metodo degli elementi finiti lineari per  $h = 0.25$  e  $h = 0.125$ . I risultati evidenziano, in particolare, la *convergenza* del metodo. In effetti, per la soluzione  $y_h$  ottenuta con il metodo degli elementi finiti lineari, si possono mostrare le seguenti *maggiorazioni dell'errore*

$$\|y - y_h\| \leq C h^2 \|f\|, \quad \|y' - y'_h\| \leq C h \|f\|$$

ove  $\|y\| = \sqrt{(y, y)}$ . ■

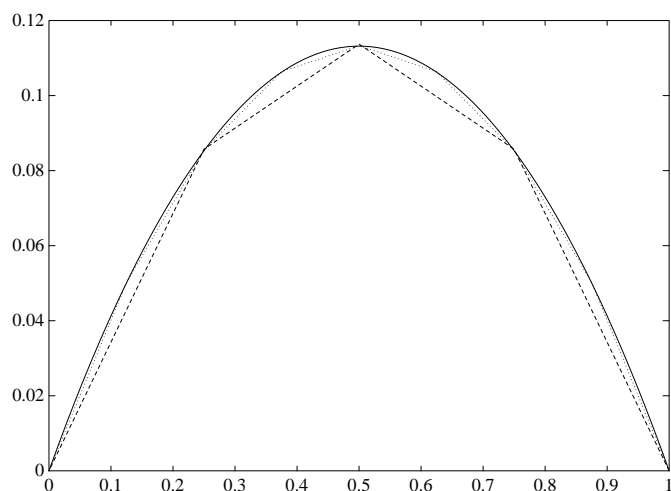


Figura 7.39: Approssimazione mediante il metodo degli elementi finiti della soluzione del problema ai limiti  $-y'' + y = 1$ ;  $y(0) = y(1) = 0$ , per  $h = 0.25$  e  $h = 0.125$ . La soluzione esatta è indicata a tratto continuo.

### 7.5.5 Problema degli autovalori

Consideriamo il seguente problema ai limiti

$$y''(x) + a^2 y(x) = 0, \quad 0 < x < 1; \quad y(0) = y(1) = 0$$

L'equazione differenziale ha la soluzione  $y = A \cos ax + B \sin ax$ . Dalla condizione  $y(0) = 0$  si ha  $A = 0$ , mentre  $y(1) = 0$  fornisce  $B \sin a = 0$ . Se  $\sin a \neq 0$ , si ha  $B = 0$ , e quindi la sola soluzione possibile è la funzione identicamente nulla  $y(x) \equiv 0$ . Se, invece,  $\sin a = 0$ , cioè  $a = n\pi$ , con  $n$  intero,  $B$  può essere scelto arbitrariamente. I valori particolari  $a^2 = n^2\pi^2$  sono chiamati *autovalori* e le soluzioni corrispondenti sono le *autofunzioni*<sup>25</sup>.

Il problema degli autovalori ha una notevole importanza nelle applicazioni, ad esempio nello studio della stabilità di strutture, nello studio delle equazioni delle onde nella fisica moderna e in alcune questioni di statistica. Usualmente, l'equazione differenziale è della seguente forma

$$\frac{d}{dx} \left( p \frac{dy}{dx} \right) - qy + \lambda \rho y = 0 \quad (7.116)$$

ove  $p, q$  e  $\rho$  sono funzioni assegnate, con  $p(x)$  funzione derivabile e  $q(x), \rho(x)$  funzioni continue. Il problema della risoluzione di tale equazione con condizioni ai limiti, ad

<sup>25</sup>Sottolineiamo l'analogia con il problema degli autovalori e autovettori di una matrice  $\mathbf{Ax} = \lambda \mathbf{x}$ . L'esistenza di autovettori equivale alla singolarità della matrice dei coefficienti  $\mathbf{A} - \lambda \mathbf{I}$ .

esempio del tipo

$$\begin{cases} a_0 y(a) + b_0 y'(a) = 0, \\ a_1 y(b) + b_1 y'(b) = 0 \end{cases}$$

è chiamato *problema di Sturm-Liouville*<sup>26</sup>. Più precisamente, il problema consiste nella ricerca dei valori di  $\lambda$  per cui esistono soluzioni  $y(x)$  non identicamente nulle.

► **Esempio 7.24** Le soluzioni del problema di autovalori  $(1-x^2)y'' - 2xy' = \lambda y$  per  $\lambda = -i(i+1)$ ,  $i = 0, 1, 2, \dots$  sono i *polinomi di Legendre*<sup>27</sup>  $P_i(x)$ . Analogamente, le soluzioni del problema  $(1-x^2)y'' - xy' = \lambda y$ , per  $\lambda = -i^2$ ,  $i = 0, 1, 2, \dots$  sono i *polinomi di Chebichev*  $T_i(x)$  (cfr. Capitolo 4). ■

Per la discretizzazione del problema di Sturm-Liouville possono essere utilizzati i vari metodi che abbiamo analizzato in precedenza. Come esemplificazione, consideriamo la discretizzazione ottenuta mediante il metodo delle differenze finite. Introdotta una suddivisione dell'intervallo  $[a, b]$  mediante i punti  $x_i = a + ih$ , e sostituendo all'operatore differenziale un opportuno rapporto incrementale, si ottengono le seguenti equazioni lineari nel vettore  $\boldsymbol{\eta}$  di componenti  $\eta_i$ , che approssimano i valori della soluzione  $y_i$

$$\frac{p_i}{h^2} (\eta_{i-1} - 2\eta_i + \eta_{i+1}) + \frac{p'_i}{2h} (\eta_{i+1} - \eta_{i-1}) - q_i \eta_i + \lambda \rho_i \eta_i = 0$$

che possono essere scritte, tenendo anche conto delle condizioni ai limiti nella forma

$$(\mathbf{A} - \lambda \mathbf{I})\boldsymbol{\eta} = 0$$

con  $\mathbf{A}$  matrice opportuna. In questa maniera il problema è ricondotto alla risoluzione di un problema di autovalori in dimensione finita.

► **Esempio 7.25** Consideriamo il seguente problema di autovalori

$$y'' + \lambda \frac{y}{(x+1)^2} = 0, \quad y(0) = y(1) = 0$$

per il quale si può mostrare che gli autovalori esatti sono dati da  $\pi^2 i^2 / (\ln 2)^2 + 1/4$ , per  $i = 1, 2, \dots$ . Applichiamo il metodo alle differenze finite con  $h = 1/n$ . Ad esempio, per  $n = 4$  si ottiene il sistema

$$\begin{cases} -2\eta_1 + \eta_2 + \lambda\eta_1/25 = 0 \\ \eta_1 - 2\eta_2 + \eta_3 + \lambda\eta_2/36 = 0 \\ \eta_2 - 2\eta_3 + \lambda\eta_3/49 = 0 \end{cases}$$

In Tabella 7.11 sono riportati i risultati ottenuti in corrispondenza ad alcuni valori di  $n$ . Tali risultati mostrano buone approssimazioni per gli autovalori più piccoli. ■

<sup>26</sup>C. Sturm (1803-1855) e J. Liouville (1809-1882) introdussero tale sistema per risolvere problemi concernenti la soluzione di equazioni differenziali alle derivate parziali delle onde e della diffusione.

<sup>27</sup>A. M. Legendre (1752-1833) introdusse tale problema nello studio del potenziale gravitazionale di una sfera solida.

n	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\lambda_5$	$\lambda_6$	$\lambda_7$
2	18						
4	19.7301	67.1751	133.0948				
8	20.5074	77.8330	162.3222	260.0084	364.786	495.58382	690.9590
esatto	20.7922	82.4191	185.1305	328.9266	513.8072	739.7723	1006.8221

Tabella 7.11: Approssimazione degli autovalori del problema di Sturm-Liouville  $y'' + \lambda y/(x+1)^2 = 0, y(0) = y(1) = 0$  mediante il metodo delle differenze finite.

◆ **Esercizio 7.15** *Approssimare mediante il metodo di shooting la soluzione del seguente problema ai limiti*

$$y'' = 1 + yy', \quad y(0) = 1, \quad y(0.6) = 2$$

Utilizzare il metodo di Runge-Kutta per risolvere i problemi a valori iniziali e il metodo delle secanti per la risoluzione dell'equazione che determina il parametro.

◆ **Esercizio 7.16** *Sviluppare il metodo degli elementi finiti, utilizzando come spazio  $V_h$  l'insieme delle spline cubiche.*

## 7.6 Equazioni integrali

In forma schematica, un'equazione integrale è un'equazione nella quale la funzione incognita compare come funzione integranda (di un integrale); è di *primo tipo* quando l'incognita compare *solo* nell'integrale, altrimenti è detta di *secondo tipo*. Un'altra distinzione si riferisce ai limiti nell'integrale; se sono fissati, l'equazione è detta di *Fredholm*, in caso contrario di *Volterra*<sup>28</sup>. Indicando con  $y(x)$  la funzione incognita, si hanno, allora, i seguenti tipi di equazioni

$$\begin{aligned} f(x) &= \int_a^b K(x, t) y(t) dt && \text{Fredholm, primo tipo, non omogenea} \\ y(x) &= f(x) + \lambda \int_a^b K(x, t) y(t) dt && \text{Fredholm, secondo tipo, non omogenea} \\ y(x) &= \lambda \int_a^b K(x, t) y(t) dt && \text{Fredholm, secondo tipo, omogenea} \\ f(x) &= \int_a^x K(x, t) y(t) dt && \text{Volterra, primo tipo, non omogenea} \\ y(x) &= f(x) + \lambda \int_a^x K(x, t) y(t) dt && \text{Volterra, secondo tipo, non omogenea} \end{aligned}$$

<sup>28</sup>Il nome di *equazione integrale* è stato suggerito da du Bois-Raymond (1888). Il contributo di Volterra in questo settore è stato fatto negli anni 1884–1896 e quello di Fredholm negli anni 1900–1903.



ove  $f(x)$  e  $K(x, t)$  sono funzioni assegnate;  $K(x, t)$  è detto il *nucleo* dell'equazione. Il numero  $\lambda$  ha il significato di *autovalore*; in alcune applicazioni può avere interesse cercare i valori di  $\lambda$  per i quali l'equazione ha una soluzione.

► **Esempio 7.26** (*Accrescimento di una popolazione*) Quando si studia l'accrescimento di una popolazione  $n(t)$  da un tempo 0 ad un tempo  $t$ , vi sono due termini da considerare. Il primo è  $n_0 s(t)$ , cioè il numero di individui nati al tempo  $t = 0$  e sopravvissuti al tempo  $t$ ; più in particolare,  $n_0$  è il numero degli individui nati al tempo  $t = 0$  e  $s(t)$ , la *funzione di sopravvivenza*, è la frazione di individui nati al tempo  $t = 0$  e che sopravvivono all'età  $t$  (cfr. Figura 7.40). Il secondo termine è dovuto agli individui nati nell'intervallo di tempo  $0 < \tau < t$  e che sopravvivono al tempo  $t$ .

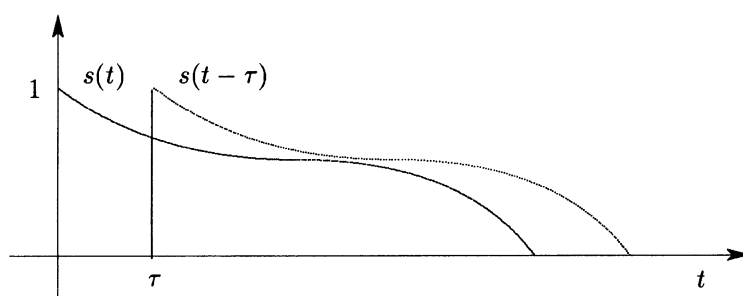


Figura 7.40: Funzione di sopravvivenza  $s(t)$ .

Se assumiamo che la variazione di popolazione al tempo  $\tau$  sia proporzionale alla popolazione presente, cioè  $\Delta n(\tau) = kn(\tau)\Delta\tau$ , avremo che tale termine è della forma

$$\int_0^t k s(t - \tau)n(\tau) d\tau$$

in quanto  $s(t - \tau)n(\tau)$  è il numero di individui presenti al tempo  $\tau$  e che sopravvivono al tempo  $t \geq \tau$ . La funzione  $n(t)$  è soluzione, quindi, della seguente equazione di Volterra

$$n(t) = n_0 s(t) + k \int_0^t s(t - \tau)n(\tau) d\tau$$

Più in generale, consideriamo due popolazioni  $n_1(t)$  e  $n_2(t)$ , di tipo *preda* e *predatore*. Quando le popolazioni sono separate, la prima aumenta, mentre la seconda diminuisce. Quando vengono a contatto, la seconda popolazione si nutre della prima e aumenta. La corrispondente velocità di accrescimento può dipendere non solo dalla popolazione  $n_1(t)$  presente, ma anche dai valori precedenti della prima popolazione. In questo caso la dinamica delle due popolazioni è descritta dal seguente sistema nel quale  $n_1$  e  $n_2$  compaiono sia sotto il segno di integrazione che di derivazione; per tale motivo si parla di *sistema di equazioni integro-differenziali*

$$\begin{aligned} \frac{dn_1}{dt} &= n_1(t) \left[ k_1 - \gamma_1 n_2(t) - \int_{t-T_0}^t s_1(t - \tau)n_2(\tau) d\tau \right], & k_1 > 0 \\ \frac{dn_2}{dt} &= n_2(t) \left[ -k_2 + \gamma_2 n_1(t) + \int_{t-T_0}^t s_2(t - \tau)n_1(\tau) d\tau \right], & k_2 > 0 \end{aligned}$$

ove  $k_1$  e  $-k_2$  sono i coefficienti di crescita e di diminuzione delle due popolazioni, nel caso in cui esse sono separate. Il valore  $T_0$  rappresenta la durata dell'effetto di *ereditarietà*.

Una situazione analoga alla precedente si riscontra nei *problemi di manutenzione* di un magazzino. Più precisamente, si tratta di calcolare la velocità  $dr/dt$  di sostituzione dei pezzi deteriorati, in modo da mantenere il numero dei pezzi efficienti ad una quantità  $f(t)$  assegnata. Indichiamo con  $s(t-\tau)$  la funzione di sopravvivenza al tempo  $t$  dei pezzi acquistati al tempo  $\tau < t$ . Supponendo nota la funzione  $s(t)$ , il problema può essere formulato mediante la seguente equazione integrale in  $dr/dt$

$$f(t) = f(0) s(t) + \int_0^t \frac{dr}{d\tau}(\tau) s(t-\tau) d\tau$$

Il termine  $(dr/d\tau)(\tau) d\tau$  rappresenta il numero dei nuovi pezzi aggiunti nell'intervallo di tempo  $(\tau, \tau + d\tau)$ . Tali pezzi saranno di età  $t - \tau$  al tempo  $t$ , e quindi la loro funzione di sopravvivenza è  $s(t - \tau)$  e il numero al tempo  $t$  è dato da  $(dr/d\tau)(\tau) s(t - \tau) d\tau$ . ■

► **Esempio 7.27** (*Problema di Bernoulli*) Si tratta di trovare la forma di una curva  $y = u(x)$  per la quale l'area  $A$  della superficie sottesa dalla curva è una frazione prescritta  $k$  dell'area del rettangolo circoscritto (cfr. Figura 7.41). Assumendo  $u(0) = 0$ , si ha la seguente equazione integrale

$$k x u(x) = \int_0^x u(t) dt$$

In particolare, per  $k = 1/3$  si trova la parabola  $y = x^2$ . ■

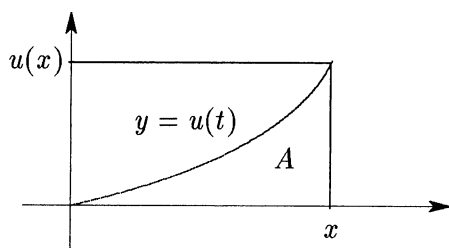


Figura 7.41: Problema di Bernoulli.

► **Esempio 7.28** (*Problema inverso relativo al potenziale gravitazionale*) In  $\mathbb{R}^3$  il potenziale  $V$  in un punto  $\mathbf{x} = [x_1, x_2, x_3]$  dovuto ad una distribuzione di massa  $\rho$  è dato da

$$V(\mathbf{x}) = -G \int_{\mathbb{R}^3} \frac{\rho(\xi, \eta, \zeta)}{r} d\xi d\eta d\zeta$$

ove  $G$  è una costante e  $r^2 = (x_1 - \xi)^2 + (x_2 - \eta)^2 + (x_3 - \zeta)^2$ . La conoscenza di  $\rho$  fornisce il potenziale mediante una integrazione diretta. Il *problema inverso* della determinazione di  $\rho$  a partire da un potenziale assegnato  $V$  è una equazione integrale. Si può anche mostrare che  $\rho$  e  $V$  sono legati dalla seguente relazione

$$\Delta V = 4\pi G\rho, \quad \mathbf{x} \in \mathbb{R}^3 \quad (7.117)$$

ove  $\Delta V = \sum_{i=1}^3 \partial^2 V / \partial^2 x_i$  è l'operatore di Laplace. L'equazione (7.117) è chiamata *equazione di Poisson*. ■

► **Esempio 7.29** (*Inversione di dati termodinamici*) Uno dei maggiori problemi della *meccanica statistica* consiste nella determinazione dell'equazione di stato di un gas in termini del potenziale  $V(r)$  tra le molecole del gas. Se la densità  $\tilde{\rho}$ , in moli per unità di volume del gas, è piccola, allora l'equazione di stato è data dalla seguente relazione, detta anche *equazione viriale*

$$\frac{p}{\tilde{\rho}RT} = 1 - 2\pi B(kT)\tilde{\rho} + O(\tilde{\rho}^2) \quad (7.118)$$

ove  $p$  è la pressione,  $T$  la temperatura,  $R$  la costante dei gas,  $k$  la costante di Boltzman e il secondo coefficiente viriale è dato da

$$B(kT) = \int_0^\infty (1 - e^{-V(r)/kT}) r^2 dr \quad (7.119)$$

Il termine  $O(\tilde{\rho}^2)$  denota termini di ordine superiore a 1 in  $\tilde{\rho}$  che possono essere trascurati quando  $\tilde{\rho}$  è piccola. Dato un potenziale  $V(r)$ , l'equazione (7.118) fornisce l'equazione di stato con  $B$  dato da (7.119). Questo è il *problema diretto*. Il *problema inverso*, consistente nella determinazione del potenziale inframolecolare  $V(r)$  dall'equazione di stato, richiede la risoluzione dell'equazione integrale non lineare (7.119) in  $V$ , per  $B$  assegnato. ■

Sulle equazioni integrali esiste una vasta letteratura sia per gli aspetti teorici che numerici. Nel seguito, dovremo limitarci ad alcuni risultati introduttivi, rinviando alla letteratura specializzata per gli opportuni approfondimenti (cfr. ad esempio Jerri [94]). Incominceremo con un esempio che mostra come le equazioni di Fredholm di primo tipo presentino maggiori difficoltà di quelle di secondo tipo.

► **Esempio 7.30** Consideriamo il caso di un nucleo costante  $K(x, t) = k$  e  $a = 0, b = 1$ . L'equazione di Fredholm di primo tipo è, allora, della forma

$$k \int_0^1 u(t) dt = f(x)$$

Tale equazione *non ha soluzione*, a meno che  $f$  sia una costante, cioè  $f(x) = c$ , nel qual caso ogni funzione per la quale  $\int_0^1 u(t) dt = c/k$  è una soluzione. Per quanto riguarda, invece, l'equazione di secondo tipo

$$u(x) - k \int_0^1 u(t) dt = f(x)$$

posto  $c = \int_0^1 u(t) dt$ , si ha  $u(x) = f(x) + kc$  e, integrando l'equazione per  $x \in [0, 1]$ , si ottiene

$$c = \int_0^1 f(t) dt + kc$$

e, quindi

$$u(x) = f(x) + \frac{k}{1-k} \int_0^1 f(t) dt$$

Pertanto, la soluzione è unica, salvo nel caso  $k = 1$ . Per  $k = 1$  non vi è soluzione, a meno che  $\int_0^1 f(t) dt = 0$ , nel qual caso la soluzione generale è  $u(x) = f(x) + p$ , con  $p$  costante arbitraria. ■

L'esempio mostra che la risoluzione delle equazioni integrali di Fredholm di primo tipo è un problema, in generale, mal posto, mentre per quelle di secondo tipo il problema può essere mal posto per particolari dati.

Una tecnica numerica, comunemente utilizzata per approssimare le soluzioni delle equazioni integrali, consiste nell'approssimare l'integrale mediante una formula di quadratura. Dettaglieremo tale procedura nel caso delle equazioni di Fredholm di secondo tipo, ricordando, tra gli altri metodi, la possibilità di adattare il metodo di Galerkin, nel quale, come abbiamo visto, la soluzione incognita è espressa come combinazione di funzioni *semplici* assegnate.

► **Esempio 7.31** Consideriamo l'equazione di Fredholm di secondo tipo lineare

$$\lambda u(s) - \int_a^b k(s,t)u(t) dt = f(s), \quad a \leq s \leq b$$

con  $k$  e  $f$  funzioni assegnate,  $\lambda$  numero assegnato in  $\mathbb{C}$  e  $u$  funzione incognita. In termini di operatori, l'equazione può essere scritta nella seguente forma

$$(\lambda - K)u = f(s) \quad (7.120)$$

Sono noti risultati di *esistenza* e *unicità* della soluzione nel caso in cui  $K$  è un operatore *compatto*. In particolare, ricordiamo che se  $f(s)$  e  $k(s,t)$  sono funzioni continue, vi è un'unica soluzione  $u(s)$  continua, eccetto quando  $\lambda$  è un autovalore di  $K$  oppure zero. Per  $\lambda = 0$  l'equazione è di *primo tipo*. Rinviando alla letteratura specializzata per un opportuno approfondimento di tali aspetti teorici, aggiungiamo soltanto che la soluzione dell'equazione integrale risulta un *problema malcondizionato* quando  $\lambda$  è *vicino* ad un autovalore e quando  $\lambda = 0$ . Il significato del malcondizionamento è, come al solito, il seguente. Supponendo, ad esempio, di passare da  $f$  a  $f + \delta f$ , il problema è malcondizionato quando a *piccole* variazioni  $\delta f$  corrispondono *grandi* variazioni  $\delta u$  nella soluzione. Il *fattore di amplificazione* è dato dall'inverso dell'operatore  $(\lambda - K)$ , in quanto si vede facilmente che

$$\delta u = (\lambda - K)^{-1} \delta f$$

Per il seguito ci interesseremo in particolare dell'aspetto *numerico*, supponendo che le funzioni  $k(s,t)$ ,  $u(s)$ ,  $f(s)$  siano almeno *continue*.

L'idea più naturale per ottenere uno schema numerico consiste nell'approssimare l'integrale mediante una *formula di quadratura*. Indicati con  $w_j$  e  $t_j$  rispettivamente i *pesi* e i *nod*i della formula di quadratura e con  $E(s)$  l'*errore di troncamento*, si ottiene per ogni  $s \in [a, b]$ , la seguente uguaglianza

$$\lambda u(s) - \sum_{j=1}^n w_j k(s, t_j) u(t_j) + E(s) = f(s)$$

Se trascuriamo  $E(s)$  e calcoliamo l'equazione nei punti  $s = t_j$  (*metodo di collocazione*), si ottiene il seguente *sistema di equazioni lineari*

$$\lambda \xi_i - \sum_{j=1}^n w_j k(t_i, t_j) \xi_j = f(t_i), \quad i = 1, \dots, n$$

ove  $\xi_i$  sono valori che *approssimano*  $u(t_i)$ . In notazione matriciale, posto  $\boldsymbol{\xi}_n = [\xi_1, \dots, \xi_n]^T$  e  $\mathbf{f}_n = [f(t_1), \dots, f(t_n)]^T$ , possiamo scrivere il sistema precedente nella forma

$$(\lambda \mathbf{I}_n - \mathbf{K}_n) \boldsymbol{\xi}_n = \mathbf{f}_n$$

ove  $\mathbf{I}_n$  è la matrice identità di ordine  $n$  e  $\mathbf{K}_n$  è una matrice di ordine  $n$ .

Analogamente a quanto avviene nel caso continuo, la matrice  $(\lambda \mathbf{I}_n - \mathbf{K}_n)$  può essere *malcondizionata* per certi valori di  $\lambda$ , in particolare per  $\lambda = 0$ . In caso di malcondizionamento la risoluzione del sistema lineare richiede l'utilizzo di metodi *stabili*: in particolare, *decomposizione in valori singolari* della matrice, fattorizzazione **QR**.

Esaminiamo, ora, il problema della *convergenza* della soluzione approssimata alla soluzione continua. Indichiamo con  $\mathbf{r}_n$  l'*operatore di restrizione*, definito nel modo seguente

$$\mathbf{r}_n u(t) = [u(t_1), u(t_2), \dots, u(t_n)]^T$$

Applicando tale operatore all'equazione continua (7.120), si ottiene

$$\lambda \mathbf{r}_n u - \mathbf{r}_n \mathbf{K} u = \mathbf{r}_n f =: \mathbf{F}_n$$

Sottraendo tale equazione dall'equazione discreta, si ottiene

$$(\lambda \mathbf{I}_n - \mathbf{K}_n)(\mathbf{r}_n u - \bar{\boldsymbol{\xi}}_n) = (\mathbf{r}_n \mathbf{K} - \mathbf{K}_n \mathbf{r}_n) u$$

Nell'ipotesi che la matrice  $\lambda \mathbf{I}_n - \mathbf{K}_n$  sia non singolare, si ottiene

$$\mathbf{r}_n u - \bar{\boldsymbol{\xi}}_n = (\lambda \mathbf{I}_n - \mathbf{K}_n)^{-1} [(\mathbf{r}_n \mathbf{K} - \mathbf{K}_n \mathbf{r}_n) u]$$

In questo modo si vede che l'*errore di discretizzazione*  $\mathbf{r}_n u - \bar{\boldsymbol{\xi}}_n$  è il prodotto di due parti. Il secondo fattore è l'errore dovuto alla formula di quadratura  $E = [e_i]$

$$e_i = [(\mathbf{r}_n \mathbf{K} - \mathbf{K}_n \mathbf{r}_n) u]_i = \int_a^b k(t_i, t) u(t) dt - \sum_{j=1}^n w_j k(t_i, t_j) u(t_j)$$

Ad esempio, per la formula di Cavalieri–Simpson composta corrispondente ad una suddivisione dell'intervallo  $(a, b)$  con passo uniforme  $h$  si ha

$$e_i = -\frac{b-a}{180} h^4 \frac{\partial^4}{\partial t^4} k(t_i, \tau_i) u(\tau_i), \quad \tau_i \in (a, b)$$

Sottolineiamo che il risultato precedente di rappresentazione dell'errore è valido solo se la *funzione*  $k(s, t)u(t)$  è *sufficientemente regolare*. Tale regolarità dipende sia dalla regolarità del nucleo  $k(s, t)$ , sia dalla regolarità della funzione incognita  $u(t)$ . L'altro termine nella maggiorazione dell'errore è la matrice inversa  $(\lambda \mathbf{I}_n - \mathbf{K}_n)^{-1}$ . Si dice che la successione di operatori  $\{(\lambda \mathbf{I}_n - \mathbf{K}_n)\}$  è *stabile* quando esistono  $N, C$  tali che

$$\sup_{n \geq N} \|(\lambda \mathbf{I}_n - \mathbf{K}_n)^{-1}\| \leq C < \infty$$

ove  $\|\cdot\|$  rappresenta una norma di matrice. Ad esempio, la formula di quadratura di Cavalieri–Simpson fornisce una approssimazione stabile.

Dalla *stabilità* e dalla *consistenza* (errore di quadratura infinitesimo con  $h$ ) si ricava la *convergenza* dello schema numerico. Si ha inoltre che l'errore nella soluzione ha lo stesso ordine di convergenza dell'errore della formula di quadratura utilizzata. Come esemplificazione, consideriamo la seguente equazione integrale

$$u(s) - \int_0^1 (st + 1) u(t) dt = s^4 - \frac{s}{6} - \frac{1}{5}$$

che ha come soluzione continua la funzione  $u(s) = s^4$ . Utilizzando come formula di quadratura la formula composta di Cavalieri–Simpson, si ottengono, in corrispondenza a successivi valori del passo  $h$ , i seguenti risultati.

$s$	$h = 0.5$	$h = 0.25$	$h = 0.125$	$u$ esatta
0.	-0.063888	-0.003993	-0.002495	0.
0.125			-0.000013	0.000244
0.250		-0.000347	0.003640	0.003906
0.375			0.019501	0.019775
0.500	-0.009722	0.057986	0.062217	0.062500
0.625			0.152297	0.152587
0.750		0.311631	0.316107	0.316406
0.875			0.585875	0.586181
1.	0.919444	0.994965	0.999685	1.

■

► **Esempio 7.32** (*Equazione di Volterra*) Incominciamo dall'equazione di Volterra del secondo tipo

$$y(x) = f(x) + \lambda \int_a^x K(x, t) y(t) dt$$

Essa può essere considerata come un caso particolare dell'equazione di Fredholm, definendo il nucleo  $K(x, t) = 0$  per  $a < t < b$ . Vi sono, comunque vantaggi a trattare l'equazione nella forma precedente, in maniera analoga a quanto avviene per le matrici triangolari rispetto al caso generale.

Dividendo l'intervallo  $[a, b]$  in parti uguali, mediante i punti  $x_r = a + rh$ ,  $r = 0, 1, \dots, n$ , e approssimando l'integrale mediante la formula del trapezio, si ottiene successivamente

$$\begin{aligned} \eta_0 &= f_0 \\ \eta_1 &= f_1 + \lambda h \left[ \frac{1}{2} K_{10} \eta_0 + \frac{1}{2} K_{11} \eta_1 \right] \\ \eta_2 &= f_2 + \lambda h \left[ \frac{1}{2} K_{20} \eta_0 + K_{21} \eta_1 + \frac{1}{2} K_{22} \eta_2 \right] \\ &\vdots \end{aligned}$$

ove  $\eta_r$  indicano i valori approssimati della soluzione nei punti  $x_r$ . Come illustrazione, consideriamo il problema

$$y(x) = x + \int_0^x (t - x) y(t) dt$$

Derivando si ottiene

$$y'(x) = 1 - \int_0^x y(t) dt; \quad y'(0) = 1$$

$$y''(x) = -y(x) \Rightarrow y = A \cos x + B \sin x$$

ed essendo  $y(0) = 0$ ,  $y'(0) = 1$ , si ha  $y(x) = \sin x$ . Applicando il metodo precedente, si ottengono i seguenti risultati. L'errore indicato è ottenuto utilizzando lo sviluppo in serie di  $\sin x$ .

$$\eta_0 = 0$$

$$\eta_1 = h \qquad \eta_1 - \sin h = \frac{h^3}{6} + \dots$$

$$\eta_2 = 2h - h^3 \qquad \eta_2 - \sin 2h = \frac{2h^3}{6} + \dots$$

$$\eta_3 = 3h - 4h^3 + h^5 \qquad \eta_3 - \sin 3h = \frac{3h^3}{6} + \dots$$

Un'altra classica procedura di tipo costruttivo per approssimare la soluzione di un'equazione integrale è fornita dal metodo delle *approssimazioni successive*. A partire da una stima iniziale  $y^{(0)}$  si costruisce la successione  $\{y^{(r)}\}$  mediante l'iterazione

$$y^{(r+1)}(x) = f(x) + \lambda \int_a^x K(x, t) y^{(r)}(t) dt$$

In generale, è necessaria, ad ogni iterazione, l'applicazione di una formula di quadratura. Nel caso dell'esempio, si ha, tuttavia

$$y^{(0)} = x,$$

$$y^{(1)} = x + \int_0^x (t-x)t dt = x - \frac{x^3}{6},$$

$$y^{(2)} = x + \int_0^x (t-x) \left( t - \frac{t^3}{6} \right) dt = x - \frac{x^3}{6} + \frac{x^5}{120},$$

$$\vdots$$

cioè lo sviluppo in serie di  $\sin x$ .

Terminiamo, osservando che un'equazione di Volterra di primo tipo

$$f(x) = \int_a^x K(x, t)y(t) dt$$

può essere, talvolta, utilmente trasformata in un'equazione di secondo tipo, mediante una derivazione

$$f'(x) = K(x, x)y(x) + \int_a^x \frac{\partial K(x, t)}{\partial x} y(t) dt$$

nell'ipotesi che  $K(x, x) \neq 0$  nell'intervallo considerato. ■

◆ **Esercizio 7.17** *Mostrare che l'equazione integrale  $y(x) = 1 - x + \int_0^1 K(x, t)y(t) dt$ , ove*

$$K(x, t) = \begin{cases} t(1-x) & \text{per } t \leq x \\ x(1-t) & \text{per } t > x \end{cases}$$

*è equivalente al problema ai limiti  $y'' + y = 0$ ,  $y(0) = 1$ ,  $y(1) = 0$ . (Suggerimento:  $y(x) = 1 - x + \int_0^x t(1-x)y(t) dt + \int_x^1 x(1-t)y(t) dt$ ,  $y' = -1 + \int_x^1 y(t) dt - \int_0^1 ty(t) dt$ ).*

◆ **Esercizio 7.18** *Risolvere l'equazione di Volterra*

$$y(x) = 1 + \int_0^x (x-t)y(t) dt$$

## 7.7 Equazioni con ritardo

Le equazioni differenziali che abbiamo analizzato in precedenza sono particolari equazioni funzionali, nelle quali la funzione incognita e le sue derivate sono calcolate nel *medesimo* istante  $t$ . Introdurremo ora, mediante un esempio, una classe più generale di equazioni funzionali, nelle quali la funzione incognita compare con differenti tipi di argomenti.

► **Esempio 7.33** (*Miscela di liquidi*) Consideriamo un recipiente contenente  $a$  litri di acqua salata (cfr. Figura 7.42). Acqua pura fluisce nel recipiente ad una velocità di  $q$  litri al minuto. L'acqua salata è mescolata in continuazione nel recipiente e la miscela fluisce all'esterno alla stessa velocità  $q$ .

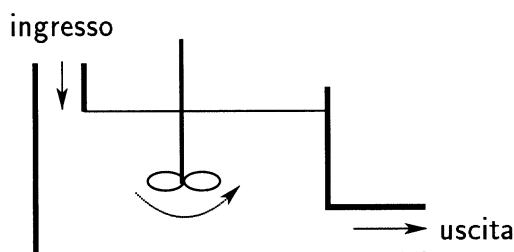


Figura 7.42: Miscela di liquidi.

Sia  $y(t)$  la quantità (in peso) del sale nella miscela al tempo  $t$ . Se supponiamo che la miscela sia perfettamente omogenea, allora l'acqua salata che esce dal recipiente contiene  $y(t)/a$  di sale per litro e quindi

$$y'(t) = -q \frac{y(t)}{a}$$

Più realisticamente, tuttavia, si deve tenere conto che il mescolamento non può avvenire istantaneamente in tutto il recipiente. Allora, la concentrazione dell'acqua salata che esce dal serbatoio al tempo  $t$  sarà uguale alla concentrazione media ad un certo istante precedente,



cioè a  $t - r$ . Possiamo supporre che  $r$  sia una costante positiva, anche se potrebbe essere una funzione opportuna del tempo. Si ottiene, allora, la seguente equazione

$$y'(t) = -cy(t - r), \quad c = \frac{q}{a} \quad (7.121)$$

che possiamo chiamare *equazione differenziale con ritardo*. Il termine  $r$  è chiamato *ritardo* (delay, time lag).

La prima importante questione che si pone riguarda le *condizioni iniziali* da usare affinché il corrispondente problema matematico a valori iniziali abbia una soluzione unica (come il problema fisico che il modello matematico rappresenta). Una risposta naturale nel caso specifico consiste nel fornire una *funzione iniziale*  $\theta(t)$  su tutto un intervallo  $[t_0 - r, t_0]$  e porre

$$y(t) = \theta(t) \quad \text{per } t_0 - r \leq t \leq t_0 \quad (7.122)$$

Si considera, quindi, il problema di trovare una estensione continua di  $\theta(t)$  a una funzione  $y(t)$  che soddisfa l'equazione (7.121) per  $t \geq t_0$  (cfr. Figura 7.43). In  $t_0$  la funzione  $y'(t_0)$  è interpretata come derivata a destra. In pratica, si suppone nota la storia passata della miscela, rappresentata dalla funzione  $\theta$ , la quale non necessariamente soddisfa l'equazione di ritardo.

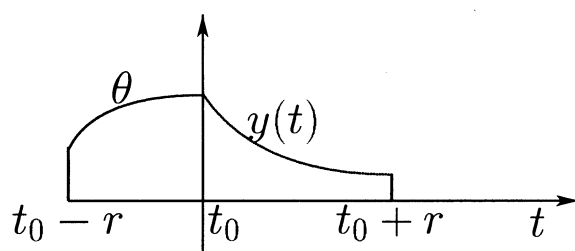


Figura 7.43: Problema a valori iniziali.

Assumiamo, ad esempio,  $\theta(t) \equiv \theta_0$ , con  $\theta_0$  costante positiva; ossia, supponiamo che nel recipiente, prima dell'istante  $t_0$  nel quale vengono aperte le valvole di ingresso e di uscita, vi sia una quantità  $\theta_0$  di sale perfettamente mescolato. In questo caso il problema (7.121), (7.122) può essere facilmente risolto per integrazioni successive. Infatti, per  $t \in [t_0, t_0 + r]$  si ha  $y'(t) = -c\theta_0$  con la condizione iniziale  $y(t_0) = \theta_0$ , e quindi la soluzione è data da  $y(t) = \theta_0 - c\theta_0(t - t_0)$ , per  $t_0 \leq t \leq t_0 + r$ . Tale procedura, nota come *metodo dei passi*, può essere ripetuta sull'intervallo  $[t_0 + r, t_0 + 2r]$ , utilizzando la funzione ora calcolata sull'intervallo precedente (cfr. per  $\theta_0 = 1$  e  $c = 1$  la Figura 7.44).

Osserviamo che, affinché la soluzione del modello matematico abbia significato fisico, è necessario che  $y(t) \geq 0$  per  $t \geq t_0$ . Tale condizione impone dei vincoli su  $r$  e  $c$ . Ora, mentre si vede facilmente che la condizione  $cr < 1$  assicura la positività di  $y$  in  $[t_0, t_0 + r]$  e la condizione  $cr < 2 - \sqrt{2}$  la positività su  $[t_0, t_0 + 2r]$ , non è agevole stabilire, mediante il metodo dei passi, la condizione affinché sia  $y(t) \geq 0$  per  $t \geq t_0$ . Si può, tuttavia, vedere in altro modo che tale condizione è data da  $cr \leq 1/e$ . ■

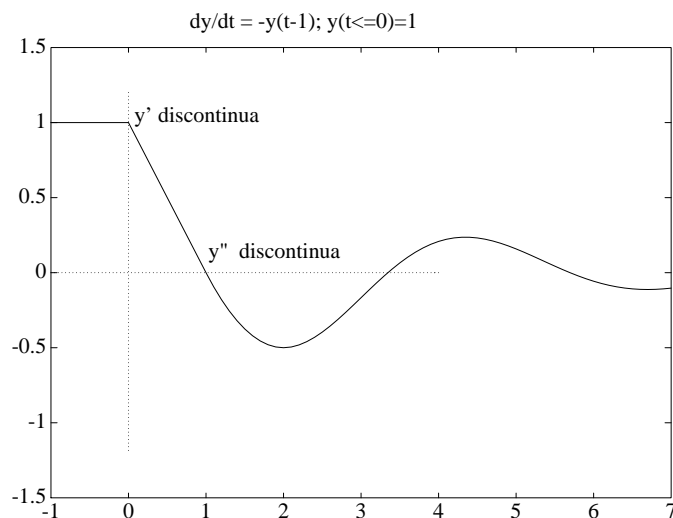


Figura 7.44: Esempio di equazione con ritardo.

► **Esempio 7.34** (*Problema dei due corpi in elettrodinamica*) L'interazione tra due elettroni, o altre particelle con una carica elettrica, è un'azione a distanza. Per il fatto, quindi, che le interazioni tra le due particelle viaggiano non istantaneamente, ma con una velocità finita  $c$  (la velocità della luce), l'influenza della particella 2 sulla particella 1 è stata generata da 2 ad un istante precedente  $t - r$  (cfr. Figura 7.45).

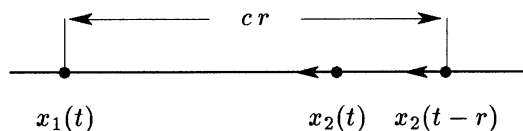


Figura 7.45: Problema dei due corpi.

Consideriamo, per semplicità, due particelle che si muovono lungo l'asse  $x$  e siano  $x_1(t)$  e  $x_2(t)$  le posizioni delle due particelle al tempo  $t$ , rispetto ad un riferimento inerziale. Allora, la forza che raggiunge  $x_1(t)$  al tempo  $t$  proveniente dalla particella 2 è stata generata al tempo  $t - r$ . Il ritardo  $r$  è determinato come *soluzione dell'equazione funzionale*

$$cr = |x_1(t) - x_2(t - r)|$$

In questo caso il ritardo  $r$  non solo non è costante, ma dipende dal tempo  $t$  attraverso le traiettorie incognite  $x_1$  e  $x_2$ . Un'equazione analoga si ha per il ritardo con cui la particella 1 influisce sulla particella 2. Indicando, allora, con  $r_{21}(t)$  e  $r_{12}(t)$  i due ritardi, si hanno le seguenti equazioni funzionali

$$cr_{21}(t) = |x_1(t) - x_2(t - r_{21}(t))|$$

$$cr_{12}(t) = |x_2(t) - x_1(t - r_{12}(t))|$$

e in corrispondenza le *equazioni di moto*, che hanno la seguente forma

$$\begin{aligned}x_1''(t) &= f_1(x_1(t) - x_2(t - r_{21}(t)), x_1'(t), x_2'(t - r_{21}(t))) \\x_2''(t) &= f_2(x_2(t) - x_1(t - r_{12}(t)), x_2'(t), x_1'(t - r_{12}(t)))\end{aligned}$$

con  $f_1$  e  $f_2$  funzioni opportune. In definitiva la dinamica delle particelle è descritta da un sistema di equazioni differenziali con ritardo e nelle quali i termini di ritardo dipendono dalle funzioni incognite stesse. Per uno studio più approfondito di tali equazioni si veda, ad esempio, Driver [51]. ■

### 7.7.1 Introduzione ai metodi numerici

La presenza di discontinuità nelle derivate limita l'utilità dei metodi numerici che corrispondono a formule ad elevata accuratezza. L'esempio analizzato in precedenza mostra, comunque, che la soluzione può diventare progressivamente più regolare. È possibile quindi prevedere la possibilità di applicare metodi più precisi nel proseguo della integrazione.

La maggior parte dei metodi numerici per i problemi a valori iniziali possono essere adattati in modo da fornire corrispondenti tecniche per le equazioni con ritardo. Si hanno quindi, in particolare, metodi di tipo Runge-Kutta, e metodi a più passi lineari. In ciascuno di questi metodi la formula standard deve essere “completata” con una formula di interpolazione per definire il valore della soluzione nel ritardo. Illustriamo l'idea facendo riferimento al problema introdotto nell'Esempio 7.33, con  $\theta_0 = 1$  e  $r = 1$ .

Come abbiamo visto il metodo analitico procede per passi; per  $1 \leq t \leq 2$  l'equazione e la condizione iniziale danno il problema  $y'(t) = 1$  che può essere risolto analiticamente (numericamente nel caso generale con un programma, ad esempio, a cambio di passo e ordine automatico). Nel caso di risoluzione numerica si ottiene la soluzione approssimata  $\eta(t_r)$  in punti  $t_r \in [1, 2]$ . In modo analogo, passando all'intervallo  $[2, 3]$ , si ha da risolvere il problema  $y'(t) = -y(t - 1)$ . L'algoritmo utilizzato avrà bisogno di valori di  $y(t - 1)$  in punti che, in generale, non saranno tra quelli  $\eta(t_r)$  calcolati nel passo precedente e quindi essi dovranno essere ottenuti per *interpolazione* o *estrapolazione*.

Una difficoltà connessa con l'operazione di interpolazione deriva dalla possibile esistenza, come abbiamo visto, di punti di discontinuità nelle derivate, che possono influire negativamente sul risultato dell'interpolazione. In alcuni casi, come nel caso di ritardi costanti, è possibile conoscere a priori la posizione dei punti di discontinuità, ma, nel caso generale la loro individuazione deve essere affidata o a un studio preliminare dell'equazione o all'*adattività* dell'algoritmo.

Dalle considerazioni ora svolte deriva, da una parte la difficoltà di costruire algoritmi di applicabilità generale e dall'altra l'*opportunità* di considerare, nel caso delle equazioni con ritardo, ma più in generale per le equazioni di tipo integro-differenziali e alle differenze, la soluzione numerica come una soluzione *globale*, anziché come una

soluzione in un insieme di punti discreto. In altre parole, è opportuno pensare la *soluzione discreta* come, ad esempio, un *polinomio*, di cui il particolare metodo utilizzato aggiorna successivamente i coefficienti. Per una subroutine basata su questa idea si veda, ad esempio Hairer et al. [74].

Terminiamo questa breve introduzione allo studio delle equazioni con ritardo con esempi significativi, risolti numericamente con un adattamento opportuno della formula RKF45.

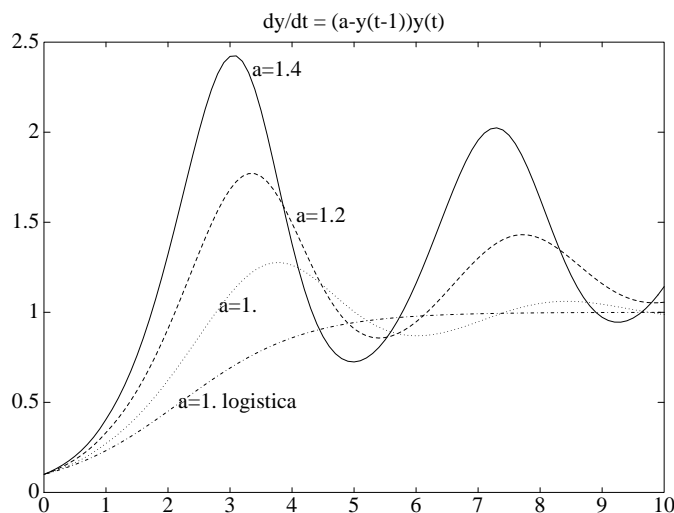


Figura 7.46: Equazione della logistica con ritardo.

► **Esempio 7.35** (*Equazione della logistica con ritardo*) Supponiamo che  $y(t)$  rappresenti la misura di una popolazione (ad esempio la quantità di cellule in un tessuto) al variare del tempo  $t$ . Il modello più semplice per studiare, ad esempio, l'accrescimento della popolazione è quello esponenziale descritto dall'equazione differenziale  $y' = \lambda y$ , ove  $\lambda$  è il tasso di accrescimento (growth rate). Nel caso in cui  $\lambda$  è supposto costante, si ha il ben noto *modello di Malthus* che ha come soluzione una funzione esponenziale. Se si tiene conto che l'accrescimento della popolazione può essere influenzato, ad esempio dalla mancanza di nutrimento, di spazio (overcrowding effect) o alla presenza di patologie (in altre parole si introduce nel modello l'effetto dell'ambiente circostante), il modello può essere rappresentato nella forma

$$y'(t) = k(a - y(t))y(t)$$

che rappresenta il modello di Verhulst(1845), Pearl(1922). La soluzione dell'equazione è la nota *curva logistica*. Partendo da un punto  $y_0 < a$ , la soluzione tende asintoticamente ad  $a$  per  $t \rightarrow \infty$ .

Questo, comunque, non è sempre il comportamento riscontrato nell'osservazione *sperimentale*. In alcune circostanze si *osserva* una *oscillazione* intorno al valore asintotico. Un modo per *interpretare* questo *dato sperimentale* consiste nell'introdurre nel modello una dipendenza del fattore di accrescimento dalla generazione precedente. La giustificazione *biologica* di questa "correzione" può, ad esempio, essere basata sul fatto che una cellula impiega

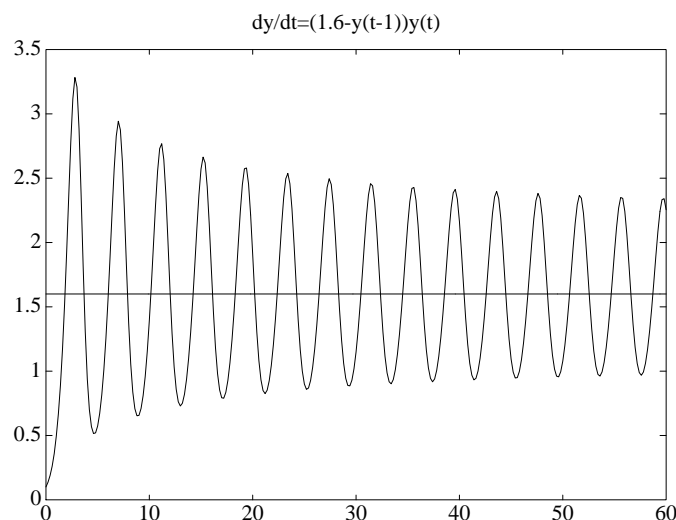


Figura 7.47: Equazione della logistica con ritardo: ciclo limite periodico.

per duplicarsi un certo tempo (il ciclo cellulare). Dal punto di vista *matematico*, il nuovo modello può essere della forma

$$y'(t) = k(a - y(t - \tau))y(t)$$

Introducendo la nuova funzione  $z(t) = k\tau y(\tau t)$  nell'equazione precedente e chiamando ancora  $y$  la nuova variabile si ha l'equazione

$$y'(t) = (a - y(t - 1))y(t)$$

Studiando l'equazione *linearizzata* si vede che  $y = a$  è un punto di equilibrio, stabile se  $0 < a \leq \pi/2$ ; inoltre per  $a < 1/e = 0.368\dots$  si ha una soluzione monotona, altrimenti la soluzione è oscillatoria. Per  $a > \pi/2$  la soluzione di equilibrio diventa instabile e si ha una soluzione periodica limite. Questa analisi è confermata dalla risoluzione numerica. In Figura 7.46 sono rappresentate le curve corrispondenti ad alcuni valori di  $a < \pi/2$ , mentre in Figura 7.47 è rappresentata la curva corrispondente al valore  $a = 1.6 > \pi/2$  e dalla quale si vede l'esistenza del limite periodico. ■

◆ **Esercizio 7.19** L'equazione differenziale  $y'(t) = [y(t)]^{2/3}$  con la condizione iniziale  $t(0) = 0$  ha infinite soluzioni. Studiare l'esistenza e l'unicità della soluzione dell'equazione

$$y'(t) = [y(t - r)]^{2/3}, \quad t \geq 0, \quad r > 0$$

◆ **Esercizio 7.20** Studiare il seguente modello di diffusione di infezione, propagata da un "vettore" (ad esempio insetti). Supponiamo che la popolazione host sia in numero costante, posto per normalizzazione = 1. Facciamo inoltre le seguenti assunzioni.

1. La malattia non è letale ed assicura un'immunità trascurabile.

2. La popolazione “vettore” è larga e il numero dei vettori infettivi è direttamente proporzionale al numero degli individui ospiti. Questo significa una mescolanza omogenea tra popolazioni ospite e vettore.
3. Vi è un ritardo  $T > 0$  tra il tempo nel quale un ospite è esposto ed il tempo nel quale l'individuo diventa infetto.
4. Gli individui infetti sono curati e possono ritornare nel pool degli individui suscettibili di infezione ad una velocità che è proporzionale al loro numero.

La popolazione ospite è partizionata, al tempo  $t$ , nelle tre classi

1.  $y(t)$  = frazione di popolazione suscettibile ;
2.  $z(t)$  = frazione di popolazione che è in incubazione;
3.  $x(t)$  = frazione di popolazione infetta ;

Si ha  $x(t) + y(t) + z(t) = 1$ . Per l'ipotesi (2) il numero dei vettori infettivi è  $b_1x(t)$ ,  $b_1 > 0$ . Da queste ipotesi si hanno le seguenti equazioni

$$\begin{aligned}x'(t) &= bx(t-T)y(t-T) - cx(t) \\y'(t) &= cx(t) - bx(t) - bx(t)y(t) \\z'(t) &= b[x(t)y(t) - x(t-T)y(t-T) - cx(t)]\end{aligned}$$

ove  $b = b_1b_2$  e  $b_2 > 0$  è la costante di “contatto” tra i vettori e i suscettibili. Le soluzioni sono “ammissibili” se  $0 \leq x, y, z \leq 1$  e  $x + y + z = 1$ .

◆ **Esercizio 7.21** Progettare un algoritmo numerico per l'approssimazione della soluzione di un'equazione del tipo

$$y'(t) = f\left(t, y(t), \int_{t-\tau}^t K(t, \xi, y(\xi)) d\xi\right)$$

Considerare come caso particolare, ad esempio, la seguente equazione (Volterra)

$$y'(t) = \left(\epsilon - \alpha y(t) - \int_0^t y(\xi)k(t-\xi) d\xi\right) y(t)$$

per lo studio della dinamica delle popolazioni, ove il termine integrale rappresenta ad esempio una diminuzione nel tasso di accrescimento dovuta all'inquinamento.

## 7.8 Equazioni alle derivate parziali

Un'equazione alle derivate parziali ha come incognita una funzione di più variabili indipendenti e contiene alcune sue derivate parziali. L'ordine dell'equazione è l'ordine massimo delle derivate che compaiono nell'equazione. Dato che la realtà è a più dimensioni, è evidente l'opportunità, e talvolta la necessità, di considerare modelli matematici basati su equazioni alle derivate parziali. In effetti, esse rappresentano uno degli strumenti più interessanti e efficaci nell'indagine di differenti fenomeni.

Nel paragrafo precedente (cfr. Esempio 7.18) abbiamo già introdotto l'*equazione della diffusione*, che è alla base dello studio del flusso del calore in un corpo, ma anche della diffusione di una sostanza, del flusso di un fluido (in condizioni di flusso laminare), del flusso di elettricità in cavi, o in neuroni (equazione del cavo, *cable equation*), e di molti altri fenomeni di tipo analogo.

Nel presente paragrafo introdurremo altri tipi di equazioni importanti nelle applicazioni, in particolare l'*equazione del potenziale* e l'*equazione delle onde*. Per le equazioni alle derivate parziali, a differenza delle equazioni ordinarie, non vi è una teoria generale per quanto riguarda le questioni dell'esistenza e l'unicità della soluzione. I seguenti semplici esempi, oltre che ad introdurre l'argomento, servono a mostrare alcune situazioni che possono presentarsi nell'ambito delle equazioni alle derivate parziali.

► **Esempio 7.36** L'equazione alle derivate parziali del secondo ordine

$$u_{yx} = 0$$

nella funzione incognita  $u(x, y) \in C^2(\mathbb{R}^2)$ , ossia dotata di derivate prime e seconde in tutto il piano, può essere facilmente risolta riscrivendo l'equazione nella forma  $(u_y)_x = 0$ , da cui si vede che  $u_y$  è una funzione indipendente da  $x$ , cioè  $u_y = k(y)$ . Allora, l'equazione  $u_y = k(y)$  è una equazione differenziale ordinaria in  $y$  per ogni valore fissato di  $x$ . Integrando rispetto a  $y$  si ottiene come soluzione  $u(x, y) = f(x) + g(y)$ , con  $f$  e  $g$  funzioni arbitrarie di classe  $C^2$ . Viceversa, una funzione della forma precedente è una soluzione dell'equazione differenziale  $u_{yx} = 0$ . Come si vede, a differenza delle equazioni differenziali ordinarie, la *soluzione generale* di un'equazione alle derivate parziali può coinvolgere, anziché costanti, delle *funzioni arbitrarie*. ■

► **Esempio 7.37** Consideriamo la seguente equazione del primo ordine

$$xu_x + yu_y + u = 0, \quad (x, y) \in \mathbb{R}^2 \quad (7.123)$$

Mostreremo che la *sola* soluzione dell'equazione che appartiene alla classe  $C^1(\mathbb{R}^2)$  è la funzione  $u(x, y) \equiv 0$ . Sia, infatti,  $u(x, y)$  una soluzione continua insieme alle derivate del primo ordine e supponiamo che  $m$  sia il minimo di  $u$  sul quadrato  $Q = [-1, 1] \times [-1, 1]$ . Se  $u = m$  in un punto *interno* di  $Q$ , allora in tale punto si ha  $u_x = u_y = 0$  e quindi dall'equazione (7.123)  $u = -xu_x - yu_y = 0$  e, pertanto,  $m = 0$ . Supponiamo, ora, che  $u = m$  in un punto sulla frontiera di  $Q$ . Per fissare le idee, sia  $(\bar{x}, -1)$  tale punto. Avremo, allora,  $u_y(\bar{x}, -1) \geq 0$  e

$$u_x(\bar{x}, -1) = \begin{cases} 0 & \text{se } |\bar{x}| < 1 \\ \geq 0 & \text{se } \bar{x} = -1 \\ \leq 0 & \text{se } \bar{x} = 1 \end{cases}$$

Dall'equazione (7.123) si ha

$$u(\bar{x}, -1) = m = -\bar{x}u_x(\bar{x}, -1) + u_y(\bar{x}, -1) \geq 0$$

In maniera analoga si procede nel caso in cui  $m$  è raggiunto in un punto qualunque della frontiera di  $Q$ . Si ha, quindi  $u \geq m \geq 0$ . Lo stesso ragionamento si può applicare a  $-u$ , che

è pure una soluzione dell'equazione (7.123). Allora,  $M = \min(-u(x, y)) \geq 0$ , su  $Q$ , da cui  $u(x, y) \leq -M \leq 0$ , per ogni  $(x, y) \in Q$  e  $m = M = 0 \Rightarrow u \equiv 0$  su  $Q$ . Il ragionamento può essere evidentemente esteso ad un quadrato qualunque, per cui, in definitiva,  $u \equiv 0$  su  $\mathbb{R}^2$ . ■

### 7.8.1 Propagazione delle onde

La *propagazione delle onde* e la *diffusione* rappresentano i due processi fondamentali in natura. In questo paragrafo analizzeremo alcuni fenomeni relativi alle onde, introducendo i relativi modelli matematici e indicando le idee di base per la loro risoluzione numerica.

Per *onda* (wave) si intende un segnale identificabile o una perturbazione in un mezzo che si propaga nel tempo, trasportando con sé energia. Esempi familiari sono le onde elettromagnetiche, le onde su una superficie di un liquido, le onde sonore, le onde degli sforzi nei solidi, provocate ad esempio da scosse telluriche. Sottolineiamo che con l'onda non si ha necessariamente trasporto di materiale; è la perturbazione, che trasporta energia che viene propagata<sup>29</sup>.

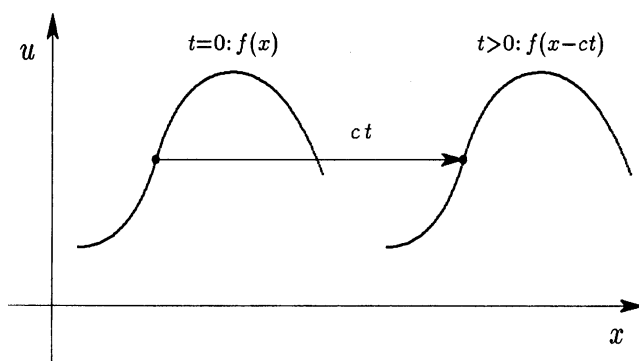


Figura 7.48: Propagazione di un'onda a velocità costante.

Un semplice modello matematico di un'onda è la funzione

$$u(x, t) = f(x - ct) \quad (7.124)$$

che rappresenta un'onda che viaggia non distorta alla velocità costante  $c > 0$ . (cfr. Figura 7.48). La coordinata  $x$  rappresenta la posizione,  $t$  il tempo, e  $u$  l'intensità della perturbazione. Per  $t = 0$  il profilo dell'onda è  $u = f(x)$  e per  $t > 0$  la perturbazione si è spostata sulla destra di una lunghezza pari a  $ct$ . Per trovare un'equazione alle derivate parziali che modella (7.124), calcoliamo  $u_t$  e  $u_x$

$$u_t = -cf'(x - ct), \quad u_x = f'(x - ct)$$

<sup>29</sup>“ a similitudine delle onde fatte il maggio nelle biade dal corso dei venti, che si vede correr l'onde per le campagne, e le biade non si mutano di lor sito” (Leonardo).



da cui

$$\boxed{u_t + cu_x = 0} \quad (7.125)$$

ossia, un'equazione differenziale alle derivate parziali del primo ordine lineare, che ha come soluzione generale la funzione (7.124), con  $f$  funzione arbitraria. Essa è anche chiamata *equazione del trasporto* (advection equation), in quanto essa può essere utilizzata, in particolare, per studiare la concentrazione di una sostanza disciolta in un fluido che si muove con velocità  $c$ . In maniera analoga, un'onda della forma  $u = f(x + ct)$  è un'onda che viaggia a sinistra ed è una soluzione dell'equazione alle derivate parziali  $u_t - cu_x = 0$ .

► **Esempio 7.38** *Onde sinusoidali*. In molte applicazioni le onde viaggianti sono periodiche, cioè della forma

$$u = A \cos(kx - \omega t) \quad (7.126)$$

Il numero positivo  $A$  è l'ampiezza e  $\lambda = 2\pi/k$  è la lunghezza d'onda. Se si scrive (7.126) nella forma

$$u = A \cos k\left(x - \frac{\omega}{k}t\right)$$

si vede che (7.126) è un'onda che viaggia a destra con velocità  $c = \omega/k$ . Questo numero è chiamato *velocità di fase* e rappresenta la velocità con cui è necessario muoversi per rimanere allo stesso punto sull'onda viaggiante. ■

Non tutte le onde si propagano in modo da mantenere il profilo della stessa forma, ossia *non distorto*; si pensi, ad esempio, alle onde di superficie del mare, o alle onde di pressione che si propagano nei solidi o nei gas. La distorsione del profilo di un'onda viene prodotta quando nei materiali i segnali sono trasmessi ad una velocità che varia con il variare della pressione. Si tratta di fenomeni di tipo *non lineare*. Nel seguito (cfr. Esempio 7.40) considereremo una importante famiglia di problemi non lineari, per i quali la distorsione del segnale può portare a delle soluzioni discontinue, dette *fronti d'onda*. Ora, tuttavia, introdurremo sul problema lineare il concetto di *curve caratteristiche*, che è di importanza basilare per lo studio sia teorico che numerico del problema.

### Curve caratteristiche

Consideriamo l'equazione del trasporto

$$u_t + cu_x = 0, \quad x \in \mathbb{R}, \quad t > 0 \quad (7.127)$$

con la seguente condizione iniziale

$$u(x, 0) = \phi(x), \quad x \in \mathbb{R}$$

La soluzione di tale problema è data dalla funzione

$$u(x, t) = \phi(x - ct)$$

Le rette  $x - ct = r$ , con  $r$  costante generica, rappresentano le linee lungo le quali i valori iniziali si propagano con valore costante. In sostanza, esse possono essere interpretate come le linee nello spazio-tempo lungo le quali vengono trasportati i segnali. Inoltre, lungo tali linee l'equazione alle derivate parziali (7.127) si riduce all'equazione differenziale ordinaria  $du/dt = 0$ . In altre parole, se  $\mathcal{C}$  è la curva  $x = ct + r$ , per un valore particolare di  $r$ , allora la derivata direzionale lungo tale curva è

$$\frac{du}{dt}(x(t), t) = u_x(x(t), t) \frac{dx}{dt} + u_t(x(t), t) = u_x(x(t), t)c + u_t(x(t), t)$$

cioè il primo membro della (7.127) calcolato lungo  $\mathcal{C}$ . La famiglia di rette  $x - ct = r$ , con  $r$  costante, è chiamata la *famiglia delle curve caratteristiche*.

Il concetto di curve caratteristiche si estende al caso di equazioni di trasporto nelle quali la velocità  $c$  è una funzione  $c(x, t)$  nel seguente modo. Successivamente, vedremo l'estensione al caso in cui  $c$  può dipendere dall'incognita  $u$  (Esempio 7.40). Consideriamo il problema a valori iniziali

$$\begin{aligned} u_t + c(x, t) u_x &= 0, & x \in \mathbb{R}, & t > 0 \\ u(x, 0) &= \phi(x), & x \in \mathbb{R} \end{aligned}$$

ove  $c(x, t)$  è una funzione assegnata. Sia  $\mathcal{C}$  la famiglia di curve definite dall'equazione differenziale

$$\frac{dx}{dt} = c(x, t)$$

Allora, lungo un elemento di  $\mathcal{C}$  si ha

$$\frac{du}{dt} = u_x \frac{dx}{dt} + u_t = u_x c(x, t) + u_t = 0$$

e quindi  $u$  è costante su ogni elemento di  $\mathcal{C}$ .

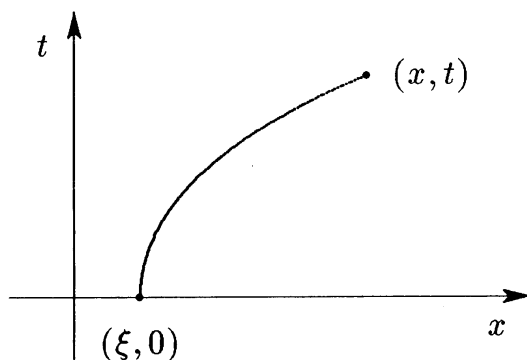


Figura 7.49: Curve caratteristiche  $x = t^2 + r$ .

► **Esempio 7.39** Consideriamo il problema a valori iniziali

$$\begin{aligned} u_t + 2tu_x &= 0, & x \in \mathbb{R}, & t > 0 \\ u(x, 0) &= e^{-x^2}, & x \in \mathbb{R} \end{aligned}$$

Le curve caratteristiche sono definite dall'equazione differenziale  $dx/dt = 2t$  che ha come soluzioni la famiglia di parabole (cfr. Figura 7.49)

$$x = t^2 + r, \quad r \text{ costante}$$

Sapendo che  $u$  è costante lungo tali curve, è possibile trovare la soluzione del problema a valori iniziali. Sia infatti  $(x, t)$  un punto arbitrario con  $t > 0$ . La curva caratteristica nel punto  $(x, t)$  passa attraverso il punto iniziale  $(\xi, 0)$ , con  $\xi$  definito da  $x = t^2 + \xi$ . Si avrà, quindi

$$u(x, t) = e^{-\xi^2} = e^{-(x-t^2)^2}$$

La velocità del segnale in  $(x, t)$  è data da  $2t$ ; essa aumenta con il tempo, ma l'onda mantiene la configurazione iniziale (cfr. Figura 7.50). ■

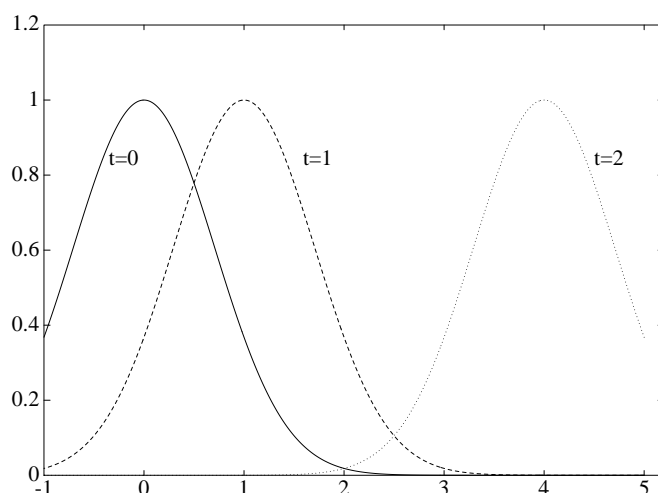


Figura 7.50: Soluzioni corrispondenti ad istanti successivi del problema a valori iniziali  $u_t + 2tu_x = 0$ ,  $u(x, 0) = e^{-x^2}$ .

## 7.8.2 Approssimazione numerica

Esamineremo, ora, le implicazioni *numeriche* della presenza delle caratteristiche. Introdotta un *passo di discretizzazione*  $h$  nella variabile  $x$  e un passo  $k$  nella variabile  $t$ , consideriamo l'*insieme discreto* definito dai seguenti punti (cfr. Figura 7.51)

$$x_j = jh, \quad t_n = nk; \quad |j|, n = 0, 1, \dots$$

Discretizziamo quindi l'equazione (7.127), utilizzando una differenza in avanti sia in  $x$  che in  $t$ . In corrispondenza, si ottiene il seguente schema numerico

$$\frac{\bar{u}_{j,n+1} - \bar{u}_{j,n}}{k} + c \frac{\bar{u}_{j+1,n} - \bar{u}_{j,n}}{h} = 0 \quad (7.128)$$

ove con  $\bar{u}$  si è indicata la *soluzione approssimata* e  $\bar{u}_{j,n} \equiv \bar{u}(x_j, t_n)$ . Posto per brevità  $\lambda = ck/h$ , da (7.128) si ottiene

$$\bar{u}_{j,n+1} = (1 + \lambda)\bar{u}_{j,n} - \lambda\bar{u}_{j+1,n}$$

A partire dai valori  $\bar{u}_{j,0}$ , definiti come approssimazioni della funzione  $\phi(x)$  nei nodi  $x_j$ , dalla formula precedente si ottengono in maniera *esplicita* tutti i valori di  $\bar{u}$ . Il *dominio di dipendenza discreta* dello schema precedente è indicato in Figura 7.51. Se *supponiamo*  $c > 0$ , si vede che il dominio di dipendenza discreto *non contiene* il dominio di dipendenza continuo.

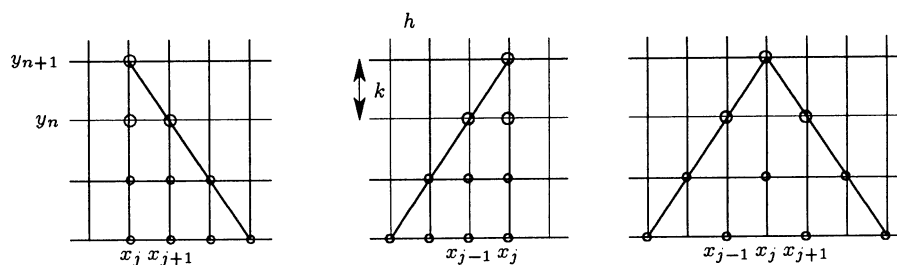


Figura 7.51: Rappresentazione di alcuni schemi alle differenze.

La conseguenza è che la soluzione discreta fornita dallo schema (7.128) *non* può, in generale, convergere per  $h, k \rightarrow 0$  alla soluzione del problema (7.127), quando  $c > 0$ . Infatti, per  $x, t$  fissato, la soluzione continua dipende dai valori di  $\phi$  alla sinistra di  $x$ , mentre la discreta da quelli a destra.

Sostituendo in (7.128) la differenza in avanti in  $x$  con la differenza all'indietro, si ottiene il seguente schema

$$\frac{\bar{u}_{j,n+1} - \bar{u}_{j,n}}{k} + c \frac{\bar{u}_{j,n} - \bar{u}_{j-1,n}}{h} = 0 \quad (7.129)$$

da cui

$$\bar{u}_{j,n+1} = (1 - \lambda)\bar{u}_{j,n} + \lambda\bar{u}_{j-1,n} \quad (7.130)$$

In questo caso si ha che il dominio di dipendenza continuo è contenuto nel dominio di dipendenza discreto se  $k/h \leq 1/c$ , cioè se

$$\boxed{\lambda \leq 1} \quad (7.131)$$

Si tratta, ora, di vedere se la condizione (7.131) è anche *sufficiente* per la *convergenza*. A tale scopo, supponendo la soluzione  $u(x, t)$  (cioè, in pratica il dato  $\phi$ ) sufficientemente regolare, incominciamo ad osservare che lo schema (7.129) è *consistente*; si ha, infatti, per l'*errore di troncamento locale* la seguente valutazione

$$\tau(x, t) = \frac{u(x, t+k) - u(x, t)}{k} + c \frac{u(x, t) - u(x-h, t)}{h} = O(k+h)$$

Definendo allora l'*errore globale*  $e(x, t) := \bar{u}(x, t) - u(x, t)$ , si ottiene

$$e_{j,n+1} = (1-\lambda)e_{j,n} + \lambda e_{j-1,n} - k\tau_{j,n}; \quad |j|, n = 1, 2, \dots$$

Ponendo  $E_n := \sup_j |e_{j,n}|$  e *utilizzando* la condizione (7.131) si ha

$$\begin{aligned} |e_{j,n+1}| &\leq (1-\lambda)|e_{j,n}| + \lambda|e_{j-1,n}| + k|\tau_{j,n}| \leq \\ &\leq (1-\lambda)E_n + \lambda E_n + k O(k+h) \leq E_n + k O(k+h) \end{aligned}$$

da cui

$$E_{n+1} \leq E_n + k O(k+h) \Rightarrow E_{n+1} \leq E_0 + t_{n+1} O(k+h)$$

Per ogni  $(x, t)$  fissato, si ha pertanto la maggiorazione

$$|\bar{u}(x, t) - u(x, t)| \leq \|\bar{u}(x, 0) - \phi(x)\|_\infty + t O(k+h)$$

Nella dimostrazione precedente si riconosce la solita implicazione: *stabilità + consistenza = convergenza*.

In conclusione lo schema (7.130) converge per particolari scelte del rapporto  $k/h$  (quando  $h, k \rightarrow 0$ ); in questo caso si dice che lo schema è *condizionatamente stabile* e quindi *condizionatamente convergente*.

Naturalmente, se  $c < 0$  i due schemi precedenti si scambiano: il primo diventa condizionatamente convergente e il secondo divergente. Si vede quindi che la stabilità dello schema dipende anche dal problema.

Uno schema *simmetrico* e quindi, in un certo senso "meno dipendente" dalla direzione della linea caratteristica, è il seguente

$$\nabla_t \bar{u}(x, t) + c \frac{1}{2} [\nabla_x \bar{u}(x, t) + \bar{\nabla}_x \bar{u}(x, t)] = 0 \quad (7.132)$$

ove  $\nabla$  indica una differenza in avanti e  $\bar{\nabla}$  una differenza all'indietro. Si può vedere facilmente che l'errore di troncamento locale corrispondente a tale schema è dato da

$$\tau = O(k+h^2)$$

Se  $|\lambda| \leq 1$ , la *condizione del dominio di dipendenza è verificata*. Tuttavia, tale schema *non è convergente*. Il senso di tale affermazione è il seguente. Se si fa tendere a zero  $h$  e  $k$  in maniera che il rapporto sia costante, esistono dati iniziali

$\phi$ , per cui la soluzione *numerica* non converge a quella corrispondente *continua*. In questo modo si è visto che la condizione del dominio di dipendenza e la consistenza non sono condizioni sufficienti per la convergenza.

D'altra parte osserviamo che lo schema (7.132) può essere modificato per ottenere uno schema convergente nel modo seguente (Figura 7.51)

$$\frac{1}{k} \left\{ \bar{u}(x, t+k) - \frac{1}{2} [\bar{u}(x+h, t) + \bar{u}(x-h, t)] \right\} + c \frac{1}{2h} [\bar{u}(x+h, t) - \bar{u}(x-h, t)] = 0$$

cioè, in forma equivalente

$$\bar{u}_{j,n+1} = \frac{1}{2}(1-\lambda)\bar{u}_{j+1,n} + \frac{1}{2}(1+\lambda)\bar{u}_{j-1,n}$$

L'errore di troncamento è ancora  $O(k + h^2)$ ; se

$$\boxed{|\lambda| \leq 1} \quad (7.133)$$

il dominio di dipendenza discreto include quello continuo. Inoltre, poiché i coefficienti  $1 \pm \lambda$  sono non negativi e con somma 1, si può, utilizzando una dimostrazione analoga a quella precedente, dimostrare che in questo caso la condizione (7.133) assicura la *convergenza* del metodo. In letteratura tale condizione è nota come *condizione di Courant–Friedrichs–Lewy*<sup>30</sup> e rappresenta una pietra miliare nello studio teorico dei metodi numerici per le equazioni alle derivate parziali.

Per terminare segnaliamo due altri schemi interessanti nelle applicazioni.

**Metodo di Lax-Wendroff** In maniera esplicita, il metodo è definito dalla formula

$$\bar{u}_{j,n+1} = (1-\lambda^2)\bar{u}_{j,n} - \frac{1}{2}\lambda(1-\lambda)\bar{u}_{j+1,n} + \frac{1}{2}\lambda(1+\lambda)\bar{u}_{j-1,n}$$

Lasciamo come esercizio verificare che il metodo è del secondo ordine, ossia che l'errore di troncamento locale si comporta come  $\tau = O(k^2 + h^2)$ .

**Schema leap-frog** Il metodo utilizza le differenze centrali sia in  $x$  che in  $t$  ed è definito dalla formula

$$\bar{u}_{j,n+1} = \bar{u}_{j,n-1} - \lambda(\bar{u}_{j+1,n} - \bar{u}_{j-1,n})$$

Anche quest'ultimo metodo è del secondo ordine, ma utilizza tre livelli nel tempo.

<sup>30</sup>... werden wir bei dem Anfangswertproblem hyperbolischer Gleichungen erkennen, daß die Konvergenz allgemein nur dann vorhanden ist, wenn die Verhältnisse der Gittermaschen in verschiedenen Richtungen gewissen Ungleichungen genügen, Courant, Friedrichs, Lewy (1928).

### 7.8.3 Equazioni non lineari

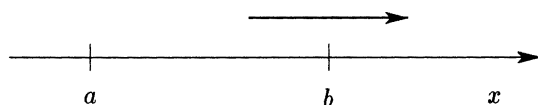
In questo paragrafo considereremo alcune estensioni dei risultati precedenti al caso di equazioni del primo ordine non lineari. In particolare, analizzeremo il caso, molto importante nelle applicazioni, delle equazioni di conservazione.

► **Esempio 7.40** (*Equazioni di conservazione*) Viene indicata come equazione di conservazione un'equazione alle derivate parziali del tipo

$$u_t + F_x(u) = 0 \quad (7.134)$$

che esprime il fatto che l'*aumento* di una quantità fisica  $u_t$  è uguale alla *variazione* nel flusso  $-F_x(u)$  della quantità attraverso la frontiera (il flusso è misurato da sinistra a destra). Nella *fluidodinamica*  $u(x, t)$  può rappresentare la densità di un fluido nel punto  $x$ , mentre  $F(x, t)$  potrebbe essere il flusso, cioè la quantità di fluido che passa attraverso il punto  $x$  al tempo  $t$ . Come semplice modello rappresentativo, considereremo un caso particolare dell'equazione (7.134), utile per prevedere il flusso del traffico stradale (flusso di macchine, anziché flusso di molecole).

Consideriamo un tratto di autostrada sul quale le macchine muovono da sinistra a destra, senza rampe di accesso né di uscita.



Poniamo

$u(x, t)$  = densità di automobili in  $x$  (automobili per unità di lunghezza in  $x$ )

$F(x, t)$  = flusso di automobili in  $x$  (automobili per minuto passanti per  $x$ )

Allora, per un segmento  $[a, b]$  di strada la *variazione* nel numero di automobili (rispetto al tempo) può essere espressa, in maniera equivalente, nei seguenti due modi

variazione nel numero delle automobili in  $[a, b]$  :  $\frac{d}{dt} \int_a^b u(x, t) dx$

variazione nel numero delle automobili in  $[a, b]$  :  $F(a, t) - F(b, t) = - \int_a^b \frac{\partial F}{\partial x}(x, t) dx$

Uguagliando i due termini, si ottiene la seguente equazione di bilancio

$$\int_a^b \frac{\partial u}{\partial t}(x, t) dx = - \int_a^b \frac{\partial F}{\partial x}(x, t) dx$$

In definitiva, poiché l'intervallo  $[a, b]$  è arbitrario, si ottiene la seguente *equazione di conservazione*

$$\boxed{u_t + F_x = 0} \quad (7.135)$$

che contiene le due incognite  $u$  e  $F$ . È, quindi, necessaria un'ulteriore relazione. Nel controllo del traffico, la quantità di automobili che passano per un punto assegnato, cioè il flusso, è trovato *sperimentalmente* come funzione della densità  $u$ . Un tipico modello può essere, ad esempio, dato da

$$F(u) = ku(1 - u)$$

ove  $k$  è una costante e 1 rappresenta la densità massima. Altri possibili modelli sono

$$(1) F(u) = ku \quad (\text{flusso lineare}); \quad (2) F(u) = ku^2 \quad (\text{flusso quadratico})$$

Nella fluidodinamica l'equazione che esprime il flusso come funzione della densità  $u$  è pure ottenuta mediante leggi sperimentali, dette *leggi costitutive* del materiale.

Quando  $F$  è una funzione di  $u$ , l'equazione (7.135) può essere riscritta nel seguente modo

$$u_t + \frac{dF}{du}u_x = u_t + g(u)u_x = 0$$

Ad esempio, nel caso di flusso quadratico  $F(u) = u^2$ , l'equazione di conservazione diventa

$$u_t + 2uu_x = 0$$

Pertanto, se la densità iniziale delle automobili è  $u(x, 0) = \phi(x)$ , la densità  $u(x, t)$  è soluzione del seguente problema a valori iniziali

$$\begin{aligned} u_t + 2uu_x &= 0 & -\infty < x < \infty, & 0 < t < \infty \\ u(x, 0) &= \phi(x) & -\infty < x < \infty \end{aligned}$$

Vediamo, ora, come estendere la nozione di curva caratteristica per tale problema, e più in generale, per un'equazione del tipo

$$u_t + g(u)u_x = 0 \quad -\infty < x < \infty \quad 0 < t < \infty \quad (7.136)$$

con  $g(u)$  funzione continua. In analogia a quanto visto relativamente all'equazione  $u_t + cu_x = 0$ , possiamo pensare ad una particella che partendo al punto  $x_0$  si muove con velocità  $g(u)$ . Al tempo  $t$  la posizione della particella sarà

$$x = x_0 + g(u)t \quad \text{equazione caratteristica.}$$

Osservando che la densità  $u(x, t)$  non cambia lungo una curva caratteristica, si avrà

$$x = x_0 + g(u(x_0, 0))t \quad \text{curva caratteristica da } (x_0, 0).$$

Ritornando all'esempio del traffico stradale, e quindi  $g(u) = 2u$ , consideriamo la seguente condizione iniziale

$$u(x, 0) = \begin{cases} 1 & x \leq 0 \\ 1 - x & 0 < x < 1 \\ 0 & 1 \leq x \end{cases}$$

Per  $x_0 < 0$ , le caratteristiche sono

$$x = x_0 + g(u(x_0, 0))t = x_0 + g(1)t = x_0 + 2t$$



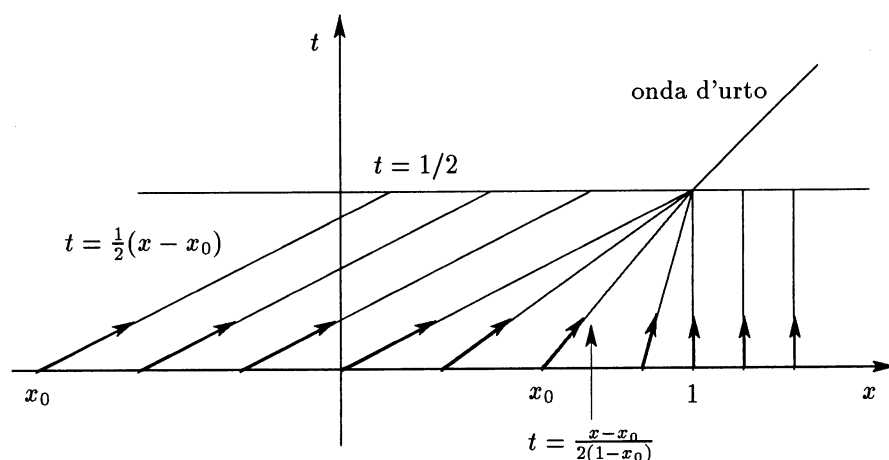


Figura 7.52: Linee caratteristiche dell'equazione  $u_t + 2uu_x = 0$ .

da cui  $t = (x - x_0)/2$ . Per  $0 < x_0 < 1$  si ha

$$x = x_0 + g((1 - x_0))t = x_0 + 2(1 - x_0)t$$

da cui  $t = (x - x_0)/(2(1 - x_0))$ . Infine, per  $1 \leq x_0 < \infty$ , si ha

$$x = x_0 + g(0)t = x_0$$

cioè delle *rette verticali*. Le caratteristiche del problema sono rappresentate in Figura 7.52, mentre in Figura 7.53 è rappresentata la funzione  $u(x, t)$  ad alcuni istanti successivi. Notiamo

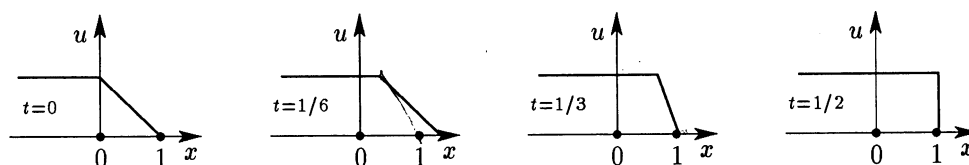


Figura 7.53: Densità del traffico a successivi istanti.

che le curve caratteristiche relative all'intervallo  $x_0 \in [0, 1]$  si intersecano tutte nel medesimo punto per  $t = 1/2$ . Poiché lungo tali caratteristiche la funzione  $u$  assume valori diversi, si verifica per la funzione una *discontinuità* in  $(1, 1/2)$ . In maniera intuitiva, poiché  $g'(u) > 0$ , i valori più grandi di  $u$  sono propagati con velocità maggiore che non per i valori più piccoli e si ha una distorsione nel profilo dell'onda. Quando le caratteristiche si incrociano, si ha il fenomeno dell'*onda d'urto* (shock wave), con discontinuità della soluzione. A partire dall'istante in cui si verifica la discontinuità, l'equazione differenziale non è più verificata, mentre mantiene significato la formulazione integrale

$$\frac{d}{dt} \int_a^b u(x, t) dt = F(u(a, t)) - F(u(b, t))$$

Utilizzando tale formulazione, si può dimostrare che la *velocità* con cui si sposta il fronte d'onda è data da

$$S = \frac{F(u_R) - F(u_L)}{u_R - u_L}$$

ove  $u_R$  e  $u_L$  sono i valori della soluzione rispettivamente a destra e a sinistra del fronte d'onda. Nell'esempio particolare considerato si ha quindi  $S = 1$ . Questo significa che per  $t > 1/2$  il fronte d'onda si muove da sinistra a destra con velocità 1. ■

▼ **Osservazione 7.4** *Supponiamo che un liquido non viscoso fluisca attraverso un tubo e che il liquido possa disperdersi attraverso la parete secondo una legge descritta dalla funzione  $G(u)$ . L'equazione, che non è più conservativa, diventa*

$$u_t + F_x = -G(u)$$

Lasciamo come esercizio lo studio e l'interpretazione della soluzione di tale equazione, quando ad esempio  $F(u) = u$ . ■

▼ **Osservazione 7.5** *Si può verificare mediante sostituzione diretta che una soluzione implicita del problema non lineare (7.136) è data da*

$$u = \phi(x - g(u)t)$$

Ad esempio, nel caso particolare in cui  $g(u) = u$ , si ottiene

$$u = x - ut \Rightarrow u(x, t) = \frac{x}{1+t}$$

Nel caso generale, tuttavia, non è possibile ricavare esplicitamente  $u$  in termini di  $x$  e  $t$ . ■

## 7.8.4 Equazione delle onde

L'equazione delle onde

$$u_{tt} - c^2 u_{xx} = 0 \tag{7.137}$$

considerata, per semplicità, nel caso unidimensionale, ha origine in maniera naturale in diverse situazioni fisiche; in particolare, in problemi di acustica, di elettromagnetismo, nello studio di vibrazioni in mezzi elastici. Nel seguito, ricavata l'equazione nel caso più semplice e intuitivo dello studio delle vibrazioni trasversali di una corda elastica, analizzeremo le proprietà fondamentali delle soluzioni del modello ottenuto, dalle quali possono essere ottenuti anche suggerimenti per una conveniente approssimazione numerica.

► **Esempio 7.41** (*Modello lineare della corda vibrante*) Data una corda elastica tesa tra i punti  $x = 0$  e  $x = L$  della retta  $x$ , costruiremo un modello del moto della corda sotto l'azione della tensione, quando non sono applicate forze esterne. Per semplificare il problema, supporremo che la corda sia *flessibile*; cioè, che la forza di tensione nella corda agisca tangenzialmente alla corda e quindi la corda non opponga resistenza alla piegatura (bending). Poiché la posizione di equilibrio della corda sotto la sola azione della tensione è l'intervallo  $0 \leq x \leq L$ , si può descrivere il moto della corda mediante tre funzioni  $u_i(x, t)$ ,  $i = 1, 2, 3$ ,

corrispondenti alle coordinate spaziali al tempo  $t$  del punto  $x$  sulla corda in posizione di equilibrio. Dal punto di vista fisico il moto della corda sarà individuato dalla posizione e dalla velocità iniziali e dalle condizioni in cui la corda è tenuta nei punti estremi  $x = 0$  e  $x = L$ . L'equazione di moto assume una forma semplice quando si introducono le seguenti ipotesi.

1. Il moto della corda avviene in un piano fissato, e i punti sulla corda sono vincolati a muoversi soltanto in una direzione trasversale alla corda. Allora, il moto può essere descritto da una singola funzione  $u(x, t)$  definita nella regione  $R = \{(x, t) : 0 < x < L, t > 0\}$ .
2. La funzione  $u$  è supposta regolare (continua con le derivate del primo e del secondo ordine), e si considerano soltanto *piccoli* spostamenti, nel senso che si trascurano termini del secondo ordine in  $u_x$ , in presenza di termini di ordine inferiore.
3. La densità  $\rho$  della corda è supposta costante. Inoltre, si suppone che non vi siano forze trasversali che agiscono sulla corda. In particolare viene trascurato il peso della corda.

Per ricavare l'equazione di moto della corda, consideriamo il bilancio delle forze che agiscono sul segmento  $[x, x']$ .

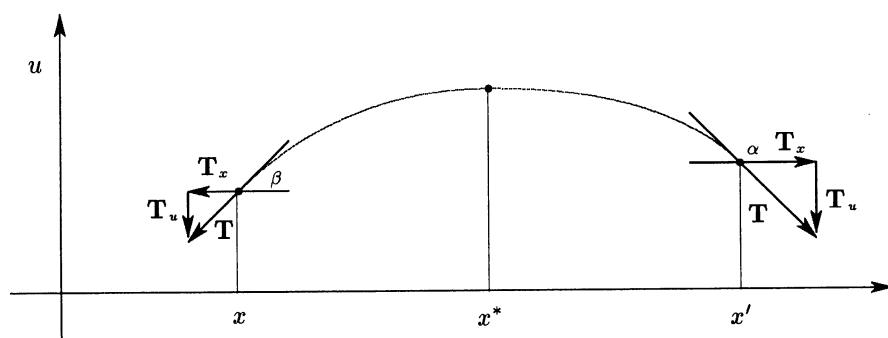


Figura 7.54: Profilo della corda nell'intervallo  $[x, x']$  al tempo  $t$ .

Per ricavare l'equazione di moto della corda, consideriamo le forze che agiscono sul segmento  $[x, x']$ . Con riferimento alla Figura 7.54, si ha

$$\begin{aligned} \|\mathbf{T}_u(x', t)\| &= \|\mathbf{T}(x', t)\| |\sin \alpha|; & \|\mathbf{T}_x(x', t)\| &= \|\mathbf{T}(x', t)\| |\cos \alpha| \\ \|\mathbf{T}_u(x, t)\| &= \|\mathbf{T}(x, t)\| |\sin \beta|; & \|\mathbf{T}_x(x, t)\| &= \|\mathbf{T}(x, t)\| |\cos \beta| \end{aligned}$$

Tenendo conto che

$$|\sin \alpha| = \frac{|u_x(x', t)|}{\sqrt{1 + |u_x(x', t)|^2}}, \quad |\cos \alpha| = \frac{1}{\sqrt{1 + |u_x(x', t)|^2}}$$

con analoghe relazioni per  $\sin \beta, \cos \beta$ , per l'ipotesi (2) si ha

$$|\sin \alpha| = |u_x(x', t)|, \quad |\sin \beta| = |u_x(x, t)|, \quad |\cos \beta| = 1, \quad |\cos \alpha| = 1$$

Introducendo, allora, un orientamento appropriato, si ottiene come *componente trasversale* delle forze di tensione che agiscono sul segmento  $[x, x']$  il seguente vettore

$$\|\mathbf{T}(x', t)\| u_x(x', t) - \|\mathbf{T}(x, t)\| u_x(x, t)$$

e come *componente parallela* il vettore

$$\|\mathbf{T}(x', t)\| - \|\mathbf{T}(x, t)\|$$

L'ipotesi di movimenti solo trasversali porta a concludere che  $\|\mathbf{T}(x', t)\| = \|\mathbf{T}(x, t)\|$ , e quindi, essendo  $x$  e  $x'$  arbitrari, che  $T \equiv \|\mathbf{T}\|$  è indipendente da  $x$  in  $[0, L]$ . Applicando, poi, la *seconda legge di Newton* al segmento  $[x, x']$ , si ha l'equazione

$$\rho(x' - x)u_{tt}(x^*, t) = T[u_x(x', t) - u_x(x, t)] \quad (7.138)$$

ove  $x^*$  è la coordinata del baricentro del segmento di corda tra  $x$  e  $x'$ . Dividendo (7.138) per  $\rho(x' - x)$ , indicando con  $c^2$  la quantità  $T/\rho$ , e passando al limite per  $x' \rightarrow x$ , si ottiene l'*equazione delle onde in una dimensione spaziale* (7.137). ■

L'equazione (7.137), una particolare equazione alle derivate parziali omogenea del secondo ordine, ha infinite soluzioni. Per esempio, se  $F$  e  $G$  sono due funzioni continue con le derivate del primo e del secondo ordine, allora  $u = F(x - ct) + G(x + ct)$  è una soluzione, come può essere dimostrato facilmente mediante la regola di derivazione delle funzioni composte.

Un'osservazione importante per la costruzione di soluzioni dell'equazione delle onde è la seguente. Se  $\Omega$  è una regione del piano  $(x, t)$  e  $u$  è una soluzione dell'equazione (7.137), continua con le derivate prime e seconde in  $\Omega$ , allora

- $u_t - cu_x$  è costante su ogni segmento in  $\Omega$  della forma  $x - ct = \text{costante}$ .
- $u_t + cu_x$  è costante su ogni segmento in  $\Omega$  della forma  $x + ct = \text{costante}$ .

La dimostrazione segue facilmente, osservando che (7.137) può essere scritta come  $v_t + cv_x = 0$ , ove  $v(x, t) = u_t(x, t) - cu_x(x, t)$ , e che  $v_t + cv_x$  rappresenta la derivata di  $v$  lungo una direzione parallela alla retta  $x - ct = \text{costante}$ . In maniera analoga si procede nel secondo caso. Le famiglie di rette  $x \pm ct = \text{costante}$  hanno quindi il significato di *caratteristiche* dell'equazione delle onde e in corrispondenza si può definire il *triangolo caratteristico* nel modo seguente (cfr. Figura 7.55).

**Definizione 7.4** *Siano  $a, b$  due punti in  $\mathbb{R}$ , con  $a < b$ . Si chiama triangolo caratteristico per l'equazione delle onde l'interno del triangolo con base  $(a, b)$  e con lati costituiti da segmenti delle linee caratteristiche  $x - ct = a$  e  $x + ct = b$ .*

La conoscenza delle caratteristiche permette la costruzione, a meno di una integrazione, della soluzione di un problema a valori iniziali relativo all'equazione delle onde, per  $-\infty < x < +\infty$ . Più precisamente, data una funzione  $f \in C^2(\mathbb{R})$  e una funzione  $g \in C^1(\mathbb{R})$ , si chiama *problema a valori iniziali* o *problema di Cauchy*, la ricerca della soluzione delle seguenti equazioni

$$\begin{aligned} u_{tt} &= c^2 u_{xx}, & x \in \mathbb{R}, & t > 0 \\ u(x, 0) &= f(x), & x \in \mathbb{R} \\ u_t(x, 0) &= g(x), & x \in \mathbb{R} \end{aligned} \quad (7.139)$$

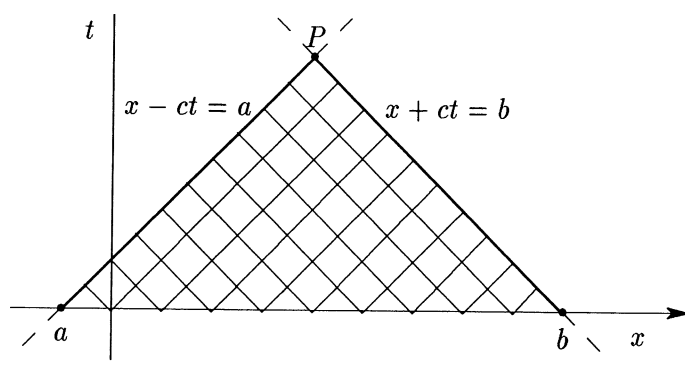


Figura 7.55: Triangolo caratteristico relativo all'equazione delle onde.

Si può mostrare che il problema ammette una ed una sola soluzione, che può essere espressa nella seguente forma, detta anche *soluzione di D'Alembert*

$$u(x, t) = \frac{f(x+ct) + f(x-ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} g(s) ds, \quad x \in \mathbb{R}, t > 0 \quad (7.140)$$

Osserviamo che la funzione  $u(x, t)$  in (7.140) può essere scritta come

$$u(x, t) = P(x+ct) + Q(x-ct)$$

ove

$$P(s) = \frac{1}{2}f(s) + \frac{1}{2c} \int_0^s g(s') ds', \quad Q(s) = \frac{1}{2}f(s) - \frac{1}{2c} \int_0^s g(s') ds'$$

L'osservazione implica che la soluzione di un problema a valori iniziali è la somma di due *onde* che viaggiano, una a destra e l'altra a sinistra, con velocità  $c$ . Considerando, ad esempio, le condizioni iniziali

$$u(x, 0) = \sin x; \quad u_t(x, 0) = 0$$

si ottiene la soluzione

$$u(x, t) = \frac{1}{2}[\sin(x-ct) + \sin(x+ct)]$$

In Figura 7.56 è rappresentata la soluzione per gli istanti  $t = 0$  e  $t = 1$ , in corrispondenza a  $c = 1$ .

### Approssimazione numerica

Vedremo ora su un esempio l'importanza delle curve caratteristiche nella scelta dello schema numerico. In sostanza, si tratta di un'estensione di quanto abbiamo visto

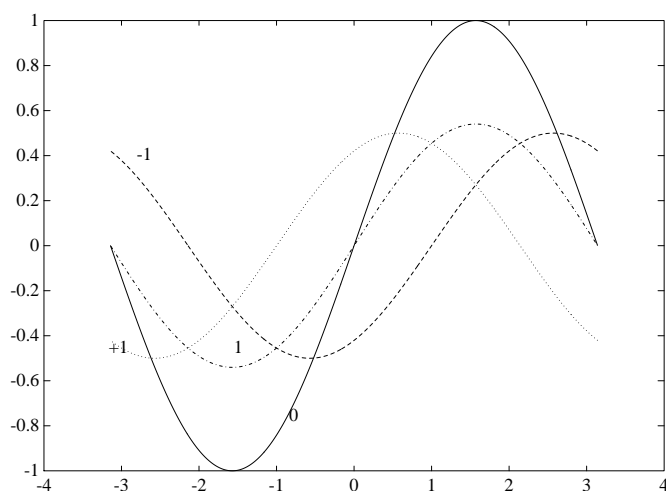


Figura 7.56: Soluzione del problema  $u_{tt} = u_{xx}$ , con le condizioni iniziali  $u(x, 0) = \sin x$ ,  $u_t(x, 0) = 0$ . La curva (0) rappresenta la soluzione in  $t = 0$  e la (1) la soluzione per  $t = 1$ . La curva (-1) rappresenta la curva  $\sin(x - t)/2$  e la (+1) la curva  $\sin(x + t)/2$ , per  $t = 1$ .

in precedenza nell'approssimazione di un'equazione alle derivate parziali del primo ordine. A questo proposito, ricordiamo che in effetti si può mostrare che l'equazione delle onde è equivalente ad un sistema di due equazioni alle derivate parziali del primo ordine.

Dati due passi  $h > 0, k > 0$ , consideriamo la reticolazione costituita dai nodi  $(x_j, t_n)$ , con  $x_j = jh$  e  $t_n = nk$ , per  $j = 0, \pm 1, \pm 2, \dots; n = 0, 1, \dots$ . Considerato il problema (7.139), discretizziamo l'equazione mediante le differenze centrali per  $u_{tt}$  e  $u_{xx}$  (cfr. Figura 7.57). Indicata con  $\bar{u}_{j,n}$  la soluzione approssimata, si ottiene l'equazione

$$\frac{\bar{u}_{j,n+1} - 2\bar{u}_{j,n} + \bar{u}_{j,n-1}}{k^2} - c^2 \frac{\bar{u}_{j+1,n} - 2\bar{u}_{j,n} + \bar{u}_{j-1,n}}{h^2} = 0 \quad (7.141)$$

Mediante uno sviluppo in serie si può vedere che l'errore di discretizzazione locale, che misura come l'equazione alle differenze approssima l'equazione differenziale, è, per una soluzione sufficientemente regolare,  $O(h^2 + k^2)$ .

L'equazione (7.141) porta al seguente schema *esplicito*

$$\bar{u}_{j,n+1} = 2(1 - r^2)\bar{u}_{j,n} + r^2(\bar{u}_{j-1,n} + \bar{u}_{j+1,n}) - \bar{u}_{j,n-1} \quad (7.142)$$

ove

$$r = c \frac{k}{h}$$

I valori di  $\bar{u}_{j,n}$  per  $n = 0$  e  $n = 1$  sono determinati dalle condizioni iniziali del

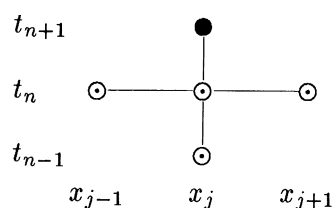


Figura 7.57: Schema alle differenze esplicito per l'equazione delle onde.

problema (7.139). Più precisamente, si ha

$$\bar{u}_{j,0} = f(x_j), \quad j = 0, \pm 1, \dots \quad (7.143)$$

$$\frac{\bar{u}_{j,1} - \bar{u}_{j,0}}{k} = g(x_j) \Rightarrow \bar{u}_{j,1} = f(x_j) + k g(x_j), \quad j = 0, \pm 1, \dots \quad (7.144)$$

Osserviamo che l'approssimazione (7.144) è del primo ordine  $O(k)$ , mentre come abbiamo osservato l'approssimazione dell'equazione delle onde è del secondo ordine. Per ottenere una approssimazione della derivata prima di ordine adeguato si può procedere nel seguente modo. Per sviluppo in serie si ha

$$u(x, k) = u(x, 0) + ku_t(x, 0) + \frac{k^2}{2}u_{tt}(x, 0) + O(k^3)$$

Ma  $u_{tt} = c^2 u_{xx} = c^2 f''(x)$ , e quindi

$$u(x, k) = u(x, 0) + ku_t(x, 0) + \frac{k^2 c^2}{2} f''(x) + O(k^3)$$

da cui la seguente approssimazione del secondo ordine

$$\begin{aligned} \bar{u}_{j,1} &= f(x_j) + kg(x_j) + \frac{k^2 c^2}{2} f''(x) \\ &= f(x_j) + kg(x_j) + \frac{r^2}{2} (f(x_{j-1}) - 2f(x_j) + f(x_{j+1})) \end{aligned}$$

Esaminiamo ora il problema della *convergenza* del metodo numerico dianzi introdotto. Seguendo lo schema più volte utilizzato in precedenza, si tratta di esaminare, dal momento che la *consistenza* è verificata per costruzione, se lo schema è *stabile*, ossia se la soluzione numerica rimane limitata al tendere a zero dei passi  $h$  e  $k$ . L'analisi può essere effettuata utilizzando differenti tecniche, quali ad esempio le maggiorazioni dell'energia e l'analisi di stabilità di von Neumann. Qui, tuttavia, ci limiteremo ad evidenziare una condizione *necessaria* affinché lo schema risulti stabile, e quindi *convergente*. In effetti, con le tecniche ora citate si può vedere che la condizione è anche *sufficiente*. In Figura 7.58 sono rappresentati il dominio di dipendenza continuo e discreto relativi ad un generico nodo  $(x_j^*, t_n^*)$ . Più precisamente, il dominio di

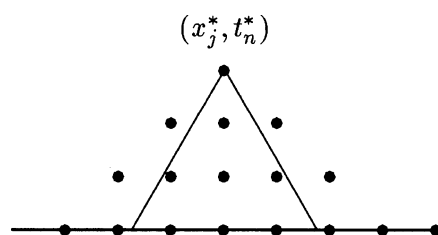


Figura 7.58: Dominio di dipendenza continuo, individuato dalle caratteristiche uscenti dal punto  $(x_j^*, t_n^*)$  con velocità  $\pm c$  e dominio di dipendenza discreto indicato dai nodi della reticolazione.

dipendenza discreto è costituito dai nodi dai quali dipende la costruzione del valore della soluzione discreta nel punto  $(x_j^*, t_n^*)$ . Osserviamo che il dominio di dipendenza discreto contiene il dominio di dipendenza continuo se la velocità con cui lo schema numerico propaga il segnale è superiore alla velocità caratteristica  $c$ , cioè se

$$\boxed{r = c \frac{k}{h} \leq 1} \quad (7.145)$$

D'altra parte, il fatto che il dominio discreto debba contenere il dominio di dipendenza continuo è ragionevole, dal momento che in caso contrario vi sarebbero dati iniziali nell'intervallo di dipendenza che lo schema numerico non utilizza. In effetti, la condizione (7.145), che rappresenta la *condizione di Courant-Friedrichs-Lewy* per l'equazione delle onde, assicura la stabilità, e quindi la convergenza, dello schema numerico.

La condizione (7.145) comporta una restrizione sui passi tanto più severa, quanto più è elevata la velocità  $c$ . In questi casi possono presentare interesse gli schemi numerici *incondizionatamente stabili*, cioè gli schemi che sono stabili senza alcuna condizione sui passi della reticolazione. Per tali metodi la scelta dei passi avviene soltanto sulla base della accuratezza richiesta. Tuttavia, per ottenere metodi con tali caratteristiche occorre sacrificare uno degli aspetti interessanti del metodo (7.141), ossia il fatto che la soluzione numerica possa essere calcolata in maniera esplicita. I metodi incondizionatamente stabili sono, infatti, metodi che hanno come dominio di dipendenza tutta la striscia  $0 \leq t \leq t_{n+1}$  e, quindi, in essi il calcolo di  $\bar{u}_{j,n+1}$  richiede la conoscenza di alcuni valori della soluzione allo stesso livello  $t_{n+1}$ . Per questo motivo essi non sono implementabili per i problemi a valori iniziali su tutta la retta  $\mathbb{R}$ , cioè nell'ipotesi di una corda infinita, ma hanno senso solo per i *problemi a valori iniziali e ai limiti*, quali si ottengono, ad esempio, nello studio delle vibrazioni di una corda di lunghezza finita  $L$ . Ricordiamo che nel caso in cui la corda sia *fissata*



agli estremi  $x = 0, x = L$ , il problema matematico assume la seguente forma

$$\begin{aligned} u_{tt} - c^2 u_{xx} &= 0, & t > 0, & 0 < x < L \\ u(x, 0) &= f(x), & u_t(x, 0) &= g(x), & 0 \leq x \leq L \\ u(0, t) &= 0, & u(L, t) &= 0 \end{aligned}$$

Definiamo, allora, la griglia  $x_j = jh$ , con  $j = 0, 1, \dots, J+1$  e  $h = L/(J+1)$ , e  $t_n = nk$ , con  $n = 0, 1, 2, \dots$  e  $k$  assegnato. Si approssima, quindi, l'equazione differenziale sostituendo  $u_{tt}$  mediante una differenza centrale nel punto  $(x_j, t_n)$  e  $u_{xx}$  mediante la media delle differenze centrali calcolate nei due punti  $(x_j, t_{n+1})$  e  $(x_j, t_{n-1})$ . Si ottiene lo schema indicato in Figura 7.59. Più esplicitamente, si ottengono le seguenti

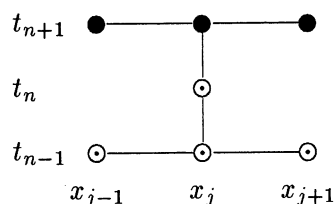


Figura 7.59: Schema alle differenze implicite per l'equazione delle onde.

equazioni alle differenze

$$\frac{\bar{u}_{j,n+1} - 2\bar{u}_{j,n} + \bar{u}_{j,n-1}}{k^2} = \frac{c^2}{2} \left[ \frac{\bar{u}_{j+1,n+1} - 2\bar{u}_{j,n+1} + \bar{u}_{j-1,n+1}}{h^2} + \frac{\bar{u}_{j+1,n-1} - 2\bar{u}_{j,n-1} + \bar{u}_{j-1,n-1}}{h^2} \right]$$

da cui, per  $j = 1, 2, \dots, J$

$$\begin{aligned} -r^2 \bar{u}_{j-1,n+1} + 2(1+r^2)\bar{u}_{j,n+1} - r^2 \bar{u}_{j+1,n+1} \\ = r^2(\bar{u}_{j-1,n-1} + \bar{u}_{j+1,n-1}) + 4\bar{u}_{j,n} - 2(1+r^2)\bar{u}_{j,n-1} \end{aligned}$$

Per ogni  $n \geq 1$ , le equazioni precedenti rappresentano un sistema di  $J$  equazioni nelle  $J$  incognite  $\bar{u}_{1,n+1}, \bar{u}_{2,n+1}, \dots, \bar{u}_{J,n+1}$ , mentre dalle condizioni ai limiti si ottiene  $\bar{u}_{0,n+1} = \bar{u}_{J+1,n+1} = 0$ . Per le condizioni iniziali si procede come nel caso precedente. Si verifica facilmente che il sistema è risolvibile, in quanto la matrice è irriducibile e a predominanza diagonale. Per la risoluzione si può utilizzare la decomposizione **LU**, oppure un metodo iterativo del tipo SOR o gradiente coniugato. Si può, infine, mostrare che lo schema risulta *convergente* per ogni successione  $(h, k)$  tendente a zero.

### 7.8.5 Equazione della diffusione

Rinviando all'Esempio 7.18 per la motivazione fisica, in questo paragrafo analizzeremo in particolare il seguente problema a valori iniziali

$$u_t - u_{xx} = 0, \quad t > 0 \quad (7.146a)$$

$$u(x, 0) = f(x), \quad -\infty < x < \infty \quad (7.146b)$$

Si può dimostrare che, se la funzione  $f(x)$  è limitata e continua, allora la soluzione può essere espressa nella seguente forma

$$u(x, t) = \int_{-\infty}^{\infty} \frac{e^{-(\xi-x)^2/4t}}{\sqrt{4\pi t}} f(\xi) d\xi \quad (7.147)$$

Da tale espressione si vede che se  $f(x) > 0$  in un intervallo aperto  $(a, b)$  e  $f(x) \equiv 0$  al di fuori di  $(a, b)$ , allora  $u(x, t) > 0$  per tutti gli  $x$  quando  $t > 0$ . In sostanza, questo significa che per l'equazione della diffusione *i segnali si propagano con velocità infinita*, ossia il *dominio di dipendenza* relativo ad un punto  $(x, t)$ , con  $t > 0$  è l'intero asse  $x$ .

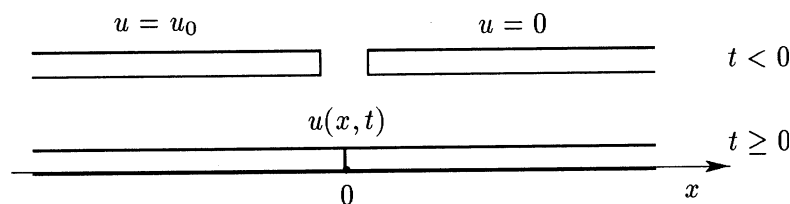


Figura 7.60: Studio della diffusione della temperatura in una sbarra, con valori iniziali  $u = u_0 > 0$  per  $x < 0$  e  $u = 0$  per  $t > 0$ .

► **Esempio 7.42** Consideriamo due semisbarre di lunghezza infinita, una delle quali a temperatura nulla e l'altra a temperatura costante  $u_0 > 0$ . All'istante  $t = 0$  vengono riunite e si vuole calcolare come varia nel tempo la temperatura della sbarra riunita (cfr. Figura 7.60).

Dall'espressione (7.147) si ottiene

$$u(x, t) = \frac{u_0}{\sqrt{\pi}} \int_{x/2\sqrt{t}}^{\infty} e^{-\xi^2} d\xi \Rightarrow u(x, t) = \frac{u_0}{2} \operatorname{erfc}\left(\frac{x}{2\sqrt{t}}\right)$$

ove la funzione  $\operatorname{erfc}$  è la *funzione errore complementare* definita da

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-\xi^2} d\xi$$

In Figura 7.61 è rappresentata la funzione  $u(x, t)$  per  $u_0 = 1$ . ■

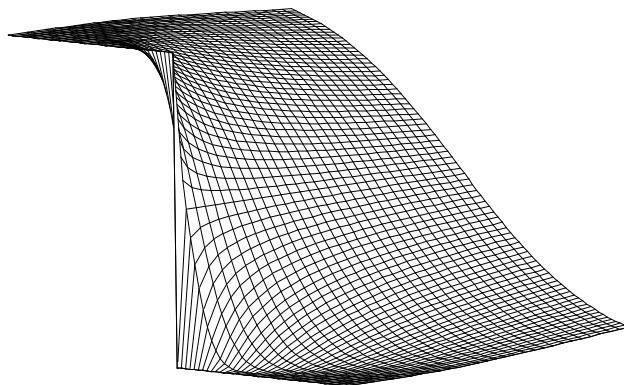


Figura 7.61: Soluzione del problema  $u_t = u_{xx}$ , con le condizioni iniziali  $u(x, 0) = 0$  per  $x > 0$  e  $u(x, 0) = u_0$  per  $x \leq 0$ .

Nelle applicazioni l'equazione della diffusione viene risolta per  $x$  che varia in un intervallo limitato. In corrispondenza, si ha da risolvere un *problema a valori iniziali e ai limiti* del seguente tipo (cfr. Figura 7.62)

$$u_t - D u_{xx} = 0, \quad 0 < t < T, \quad 0 < x < L \quad (7.148a)$$

$$u(x, 0) = f(x), \quad 0 < x < L \quad (7.148b)$$

$$u(0, t) = g_1(t), \quad u(L, t) = g_2(t), \quad 0 < t < T \quad (7.148c)$$

ove  $D$  rappresenta il *coefficiente di diffusione* e  $f(x), g_1(t), g_2(t)$  sono funzioni assegnate.

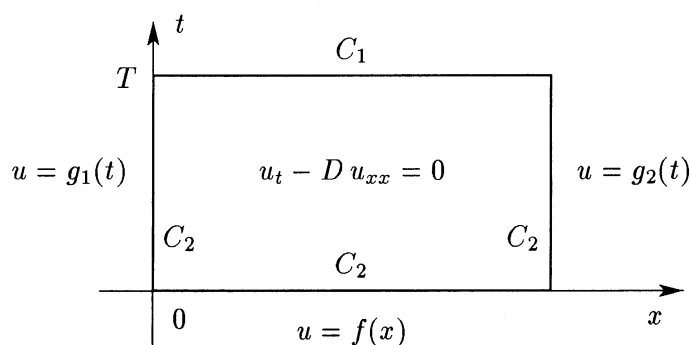


Figura 7.62: Problema a valori iniziali e ai limiti relativo all'equazione della diffusione.

La natura diversa della diffusione, rispetto alla propagazione di onde che abbiamo analizzato nel paragrafo precedente, è evidenziata dal seguente risultato, noto come principio del massimo.

**Proposizione 7.1** (Principio del massimo) *Sia  $u(x, t)$  una funzione continua nell'insieme  $\bar{R} = R \cup C_1 \cup C_2$ , ove  $R \equiv \{x, t \mid 0 < x < L, 0 < t < T\}$  e  $C_1, C_2$  sono porzioni di frontiera di  $R$  indicate in Figura 7.62. Se in  $R \cup C_2$  la funzione  $u(x, t)$  ha una derivata seconda continua rispetto a  $x$  e una derivata prima continua rispetto a  $t$  e verifica l'equazione*

$$u_t - D u_{xx} = 0$$

con  $D > 0$ , allora il  $\max_{x \in \bar{R}} u(x, t)$  è ottenuto in un punto di  $C_2$ . In maniera analoga, considerando la funzione  $-u(x, t)$ , si può dimostrare che il  $\min_{x \in \bar{R}} u(x, t)$  è raggiunto in  $C_2$ .

Il precedente risultato suggerisce dei criteri per la scelta dei metodi numerici. In effetti, come ora vedremo, per avere dei metodi convergenti è necessario che la soluzione numerica si propaghi con velocità sufficientemente elevata. Tale risultato è automaticamente verificato per i metodi di tipo implicito, mentre per i metodi espliciti è ottenuto imponendo opportuni vincoli sui passi di discretizzazione.

### Metodo alle differenze esplicito

Con riferimento al problema (7.148), consideriamo la reticolazione corrispondente ai nodi  $(x_j, t_n)$ , con  $x_j = j h$ ,  $t_n = n k$  e  $h = L/J$ ,  $k = T/N$ . Al solito indichiamo con  $\bar{u}_{j,n}$  la soluzione numerica nel nodo  $(x_j, t_n)$ . La condizione iniziale (7.148b) porta direttamente ai valori

$$\bar{u}_{j,0} = f(x_j), \quad j = 0, 1, \dots, J$$

e le condizioni ai limiti (7.148c) ai valori

$$\bar{u}_{0,n} = g_1(t_n), \quad \bar{u}_{J,n} = g_2(t_n), \quad n = 1, 2, \dots, N$$

Per  $0 < j < J$  e  $n > 0$  discretizziamo l'equazione (7.148a) nel punto  $(x_j, t_n)$  mediante una differenza in avanti per la derivata  $u_t$  e una differenza centrale per la derivata seconda  $u_{xx}$ , cioè

$$u_t \approx \frac{\bar{u}_{j,n+1} - \bar{u}_{j,n}}{k}, \quad u_{xx} \approx \frac{\bar{u}_{j+1,n} - 2\bar{u}_{j,n} + \bar{u}_{j-1,n}}{h^2}$$

Sostituendo in (7.148) si ottiene lo schema illustrato in Figura 7.63 e dal quale si ricava la seguente relazione, per  $j = 1, \dots, J - 1$  e  $n = 0, 1, \dots, N - 1$

$$\bar{u}_{j,n+1} = (1 - 2r)\bar{u}_{j,n} + r(\bar{u}_{j+1,n} + \bar{u}_{j-1,n}) \quad (7.149)$$

ove si è posto

$$r = D \frac{k}{h^2} \quad (7.150)$$

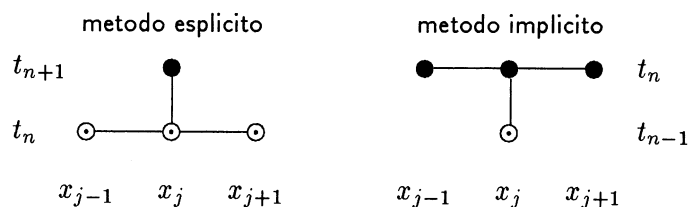


Figura 7.63: Schemi alle differenze esplicito ed implicito per l'equazione della diffusione.

► **Esempio 7.43** Un tubo di lunghezza  $L$  è inizialmente riempito con aria contenente una concentrazione  $u(x, 0) = f$  di vapori di alcool etilico. Si suppone, inoltre, che nel primo estremo, corrispondente a  $x = 0$ , i vapori possano diffondersi liberamente nell'aria, in maniera che la concentrazione nel primo estremo sia sostanzialmente nulla. Nel secondo estremo, per  $x = L$ , la concentrazione è mantenuta costante, cioè  $u(L, t) = g_2$ . Considerando soltanto gli effetti della diffusione molecolare, si tratta di determinare la concentrazione  $u(x, t)$  di alcool come funzione della variabile  $x$  e  $t$ . Dal punto di vista matematico la  $u(x, t)$  è soluzione di un problema a valori iniziali e ai limiti del tipo (7.148). Supponendo di mantenere il sistema alla temperatura costante di  $30^\circ$ , si ha per l'alcool etilico un coefficiente di diffusione  $D = 0.119 \text{ cm}^2/\text{sec}$ . Si può mostrare che la soluzione, nel caso dei dati  $f = 2$ ,  $g_2 = 10$ ,  $L = 20$ , assume la seguente rappresentazione analitica sotto forma di serie (cfr. Appendice B)

$$u(x, t) = \frac{x}{2} + \frac{20}{\pi} \sum_{n=1}^{\infty} e^{-0.01175n^2t} \sin \frac{n\pi x}{10} - \frac{12}{\pi} \sum_{n=1}^{\infty} e^{-0.00294(2n-1)^2t} \sin \frac{(2n-1)\pi x}{20}$$

dalla quale si vede che per  $t \rightarrow \infty$  la concentrazione  $u(x, t)$  tende alla *soluzione stazionaria*  $u = x/2$ .

La soluzione numerica ottenuta mediante il metodo esplicito (7.149) è rappresentata in Figura 7.64. Più precisamente, sono rappresentati i risultati corrispondenti a due diversi valori della quantità  $r$  definita in (7.150). Mentre per  $r = 1/2$  si osserva un comportamento regolare della soluzione discreta, con convergenza, per  $t \rightarrow \infty$ , alla soluzione stazionaria, per  $r > 1/2$  la soluzione mostra una *instabilità*, che si accentua per  $t$  che aumenta (la figura corrisponde a  $T = 139$ ). ■

L'esempio numerico precedente mostra l'importanza della quantità (7.150). In effetti, si può dimostrare che la condizione

$$\boxed{r = D \frac{k}{h^2} \leq \frac{1}{2}} \quad (7.151)$$

è una *condizione necessaria* per la *stabilità* del metodo numerico esplicito (7.149). In sostanza, essa assicura che il segnale *discreto* si propaghi con velocità sufficientemente alta.

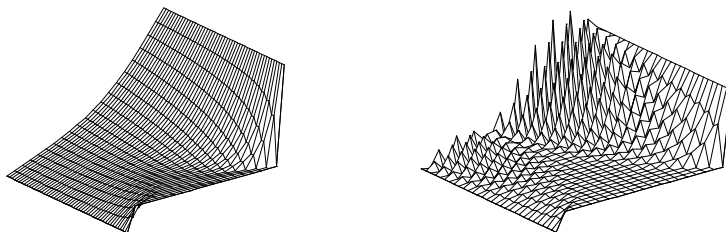


Figura 7.64: Metodo alle differenze esplicito. La prima figura corrisponde ad una reticolazione con  $h = 1$  e  $k$  tale che  $Dk/h^2 = 0.5$ ; la seconda corrisponde allo stesso valore di  $h$ , mentre  $k$  è determinato da  $Dk/h^2 = 0.55$ .

Per quanto riguarda la *convergenza*, è sufficiente osservare che lo schema (7.150) è *consistente*. Più precisamente, si può mostrare che per una funzione sufficientemente regolare, l'*errore di discretizzazione locale* si comporta come  $O(k + h^2)$ . Si conclude, quindi, che lo schema (7.150) è *convergente*, in quanto stabile e consistente, *sotto la condizione* (7.151). Si dice anche che lo schema è *condizionatamente convergente*.

### Metodo alle differenze implicito

Metodi *incondizionatamente convergenti* possono essere ottenuti utilizzando schemi di tipo implicito, per i quali è necessario risolvere ad ogni livello temporale un sistema lineare. Uno schema di questo tipo può essere ottenuto utilizzando, per approssimare la derivata  $u_t$ , anziché una differenza in avanti, una *differenza all'indietro* (cfr. Figura 7.63). Si ottiene in questo modo lo schema

$$\frac{\bar{u}_{j,n} - \bar{u}_{j,n-1}}{k} - D \frac{\bar{u}_{j+1,n} - 2\bar{u}_{j,n} + \bar{u}_{j-1,n}}{h^2} = 0 \quad (7.152)$$

da cui

$$-r\bar{u}_{j-1,n} + (1 + 2r)\bar{u}_{j,n} - r\bar{u}_{j+1,n} = \bar{u}_{j,n-1} \quad (7.153)$$

per  $n = 1, 2, \dots, N$  e  $j = 1, 2, \dots, J - 1$ . Essendo i valori  $\bar{u}_{0,n}$ ,  $\bar{u}_{J,n}$  determinati dalle condizioni ai limiti, le equazioni (7.153) rappresentano un sistema di equazioni nelle  $J - 1$  incognite  $\bar{u}_{1,n}, \bar{u}_{2,n}, \dots, \bar{u}_{J-1,n}$ . Come si verifica facilmente, la matrice del sistema è tridiagonale, simmetrica e a predominanza diagonale; il sistema può, quindi, essere risolto utilizzando la decomposizione  $\mathbf{A} = \mathbf{L}\mathbf{U}$ , che può essere effettuata, quando i coefficienti dell'equazione sono indipendenti dalla variabile  $t$ , una sola volta.

Consideriamo, ora, più in dettaglio il problema della *convergenza*. L'*errore locale di discretizzazione*  $\tau_{j,n}$  è definito dalla seguente relazione

$$\tau_{j,n} := u_t(x_j, t - n) - Du_{xx}(x_j, t_n) - \left( \frac{u_{j,n} - u_{j,n-1}}{k} - D \frac{u_{j+1,n} - 2u_{j,n} + u_{j-1,n}}{h^2} \right) \quad (7.154)$$

da cui, per sviluppo in serie

$$\tau_{j,n} = \frac{1}{2}u_{tt}(x_j, \bar{t}_n)k - \frac{D}{12}u_{xxxx}(\bar{x}_j, t_n)h^2$$

ove  $\bar{t}_n$  è un opportuno valore nell'intervallo  $(t_{n-1}, t_n)$  e, analogamente,  $\bar{x}_j$  nell'intervallo  $(x_{j-1}, x_{j+1})$ . Si ha, pertanto, per una soluzione  $u$  sufficientemente regolare

$$\tau_{j,n} = O(k) + O(h^2) \quad (7.155)$$

Definendo l'errore globale  $e_{j,n} := u_{j,n} - \bar{u}_{j,n}$ , dalle equazioni (7.152), (7.154) si ottiene

$$(1 + 2r)e_{j,n} = e_{j,n-1} + r(e_{j+1,n} + e_{j-1,n}) - k\tau_{j,n}$$

da cui, posto  $E_n = \max_{0 \leq j \leq J} |e_{j,n}|$  e  $\hat{\tau} = \max_{j,n} |\tau_{j,n}|$ , si ricava la maggiorazione

$$(1 + 2r)|e_{j,n}| \leq E_{n-1} + 2rE_n + k\hat{\tau}$$

Prendendo il massimo per  $0 \leq j \leq J$ , si ha

$$E_n \leq E_{n-1} + k\hat{\tau}$$

e, quindi, per ricorrenza

$$E_n \leq E_0 + nk\hat{\tau} \leq T\hat{\tau} \quad (7.156)$$

La disuguaglianza (7.156) mostra che

$$\boxed{|u_{j,n} - \bar{u}_{j,n}| \leq T\hat{\tau}}$$

cioè che lo schema implicito è *stabile*. La convergenza segue, allora, dalla consistenza.

Il metodo implicito permette, a differenza del metodo esplicito, la scelta del passo  $k$  solo in base all'accuratezza richiesta. In Figura 7.65 sono rappresentati i risultati ottenuti relativamente all'Esempio 7.43 e con la reticolazione  $h = 1$  e  $k$  tale che  $Dk/h^2 = 1$ .

### Metodo di Crank-Nicolson

Una estensione dei metodi esplicito e implicito, visti in precedenza, consiste nel mediare opportunamente le approssimazioni di  $u_{xx}$  lungo i due livelli temporali  $t_{n+1}$  e  $t_n$  (cfr. Figura 7.66). In questo modo si possono ottenere schemi con caratteristiche di precisione e di stabilità diverse. Se  $\theta$  è un parametro con  $0 \leq \theta \leq 1$ , si ha la seguente famiglia di metodi

$$\begin{aligned} \frac{\bar{u}_{j,n+1} - \bar{u}_{j,n}}{k} - D \left[ \theta \frac{\bar{u}_{j+1,n+1} - 2\bar{u}_{j,n+1} + \bar{u}_{j-1,n+1}}{h^2} \right. \\ \left. + (1 - \theta) \frac{\bar{u}_{j+1,n} - 2\bar{u}_{j,n} + \bar{u}_{j-1,n}}{h^2} \right] = 0 \end{aligned}$$

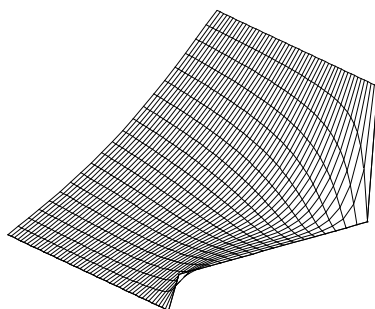


Figura 7.65: Metodo alle differenze implicito con  $h = 1$  e  $k$  tale che  $Dk/h^2 = 1$ .

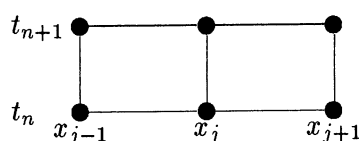


Figura 7.66: Schemi alle differenze di Crank-Nicolson.

per  $j = 1, \dots, J - 1$  e  $n = 0, \dots, N - 1$ .

Per  $\theta = 0$  si riottiene il metodo esplicito. Per  $\theta > 0$  gli schemi sono impliciti e per  $\theta = 1$  si ha il metodo implicito considerato in precedenza. Il metodo corrispondente al valore  $\theta = 1/2$  è noto come *metodo di Crank-Nicolson*. Si può mostrare che per tale metodo l'errore di troncamento locale è  $O(h^2 + k^2)$ , e quindi del secondo ordine sia nella variabile spaziale che temporale. Per  $\theta \neq 1/2$  i metodi sono del primo ordine nella variabile  $t$ , ma la costante d'errore (cioè il fattore moltiplicativo del termine  $h^2 + k$ ) diminuisce per  $\theta \rightarrow 1/2$ . Per quanto riguarda la stabilità, e quindi anche la convergenza, si ha che per  $\theta \geq 1/2$  i metodi sono *incondizionatamente stabili*, mentre per  $\theta < 1/2$  i metodi sono stabili sotto la condizione

$$D \frac{k}{h^2} \leq \frac{1}{2(1 - 2\theta)}$$

Ne segue l'interesse del metodo di Crank-Nicolson, che risulta incondizionatamente stabile e del secondo ordine. In pratica, tuttavia, poiché  $\theta = 1/2$  è esattamente il valore di separazione tra i metodi condizionatamente stabili e quelli incondizionatamente stabili, può essere opportuno, per evitare instabilità dovute agli errori di arrotondamento, utilizzare un metodo corrispondente ad un valore di  $\theta$  leggermente superiore a  $\theta = 1/2$ .

**▼ Osservazione 7.6** *I metodi alle differenze considerati in questo paragrafo nel caso di condizioni ai limiti di tipo Dirichlet possono essere estesi facilmente a problemi corrispondenti a condizioni ai limiti di tipo differente. Con riferimento alla formulazione (7.148),*



supponiamo, ad esempio, che per  $x = L$  si abbia la seguente condizione

$$\frac{\partial u}{\partial x} = g_2(t) \quad (7.157)$$

In termini fisici la condizione corrisponde ad assegnare il flusso (di calore o di sostanza) attraverso il secondo estremo. Con il metodo delle differenze finite tale condizione può essere discretizzata nel seguente modo. Si estende la reticolazione  $(x_j, t_n)$ , aggiungendo i punti di frontiera  $(x_{J+1}, t_n)$ , per  $n = 0, 1, \dots, N$  (cfr. Figura 7.67). Si discretizza, quindi, la condizione (7.157) mediante una differenza centrale

$$\bar{u}_{J+1,n} - \bar{u}_{J-1,n} = 2h g_2(t_n)$$

per  $n = 1, 2, \dots, N$ . L'equazione (7.152) viene, di conseguenza, scritta per  $j = 1, 2, \dots, J$  e si ottiene un sistema lineare nelle  $J$  incognite  $\bar{u}_{1,n}, \bar{u}_{2,n}, \dots, \bar{u}_{J,n}$ . ■

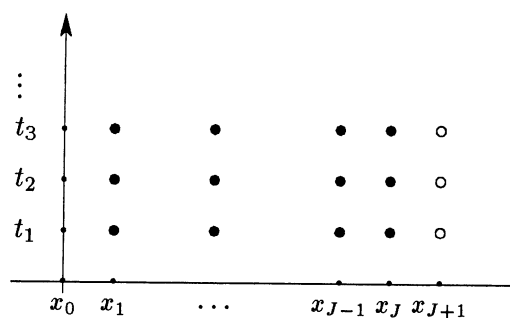


Figura 7.67: Insieme discreto per una condizione ai limiti di tipo  $u_x(L, t) = g_2(t)$ .

▼ **Osservazione 7.7** In questo paragrafo si sono introdotte le idee che sono alla base dell'approssimazione numerica dell'equazione della diffusione, limitandoci, per brevità, al metodo delle differenze finite. Analoghe problematiche si hanno relativamente ad altri metodi, quali ad esempio il metodo degli elementi finiti. Per una trattazione più adeguata di tali metodi rinviamo, tuttavia, alla letteratura specializzata. ■

### 7.8.6 Equazione di Laplace

L'equazione di Laplace

$$\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} = 0, \quad \mathbf{x} = [x_1, x_2, x_3]^T \in \Omega \quad (7.158)$$

ove  $\Omega$  è una regione assegnata nello spazio  $\mathbb{R}^3$ , è il prototipo delle equazioni di tipo *ellittico* e rappresenta una delle equazioni fondamentali della fisica matematica. Essa è alla base, infatti, di numerosi modelli per lo studio, ad esempio, dei fenomeni in elettrostatica, degli stati stazionari nella diffusione e nella fluidodinamica. Ricordiamo che in uno *stato stazionario* la funzione  $u$  non cambia con il tempo, ma può

variare da punto a punto dell'insieme  $\Omega$ . Spesso la funzione  $u$  ha il significato di un potenziale; per tale motivo l'equazione (7.158) è anche chiamata *equazione del potenziale*. L'equazione di Laplace non omogenea  $\Delta u = f$  è nota come *equazione di Poisson*. La funzione  $f$  tiene conto della presenza di eventuali sorgenti in  $\Omega$ . Ricordiamo, infine, l'equazione  $\Delta u + \lambda u = 0$ , nota come *equazione di Helmholtz* (o equazione delle onde ridotta), che descrive, in particolare, le configurazioni di una membrana elastica soggetta ad un campo di sforzi.

Il carattere dell'equazione (7.158), in rapporto all'equazione delle onde e della diffusione che abbiamo analizzato in precedenza, è evidenziato dal seguente esempio.

► **Esempio 7.44** (*Esempio di Hadamard*). Consideriamo l'equazione alle derivate parziali

$$u_{yy} + u_{xx} = 0, \quad y > 0, \quad x \in \mathbb{R}$$

soggetta alle *condizioni iniziali*

$$u(x, 0) = 0, \quad u_y(x, 0) = 0, \quad x \in \mathbb{R}$$

La soluzione di tale problema è evidentemente la funzione identicamente nulla  $u(x, t) \equiv 0$  per  $t \geq 0$ ,  $x \in \mathbb{R}$ . Modificando le condizioni iniziali nel seguente modo

$$u(x, 0) = 0, \quad u_y(x, 0) = 10^{-4} \sin 10^4 x$$

la soluzione diventa la seguente

$$u(x, y) = 10^{-8} \sin(10^4 x) \sinh(10^4 y)$$

Ora, la funzione  $\sinh(10^4 y)$  si comporta, per  $y$  sufficientemente grande, come la funzione  $\exp(10^4 y)$ , e, quindi, la soluzione cresce esponenzialmente con  $y$ . Si vede, pertanto, che una perturbazione arbitrariamente piccola nei dati porta ad una variazione arbitrariamente grande nella soluzione, e quindi che il problema a valori iniziali è instabile (cfr. Figura 7.68). ■

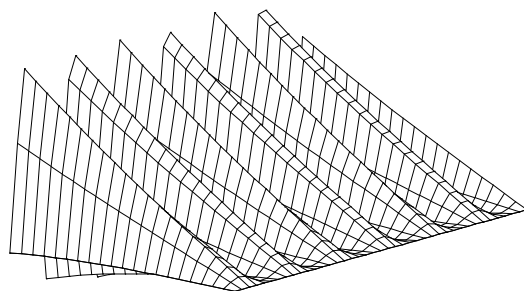


Figura 7.68: Esempio di Hadamard  $\Delta u = 0$ ,  $u(x, 0) = 0$ ;  $u_y(x, 0) = 10^{-4} \sin 10^4 x$  per  $0 \leq t \leq 10^{-4}$ .

L'esempio precedente suggerisce che per l'equazione di Laplace i problemi a valori iniziali possono essere malposti. In sostanza, l'equazione  $\Delta u = 0$  esprime il fatto

che la soluzione  $u$  in un punto fissato è uguale alla media della soluzione nei punti vicini; essa descrive, cioè, un *fenomeno di equilibrio*<sup>31</sup>. Pertanto, per l'equazione di Laplace sono più naturali le condizioni di tipo al contorno. Tali condizioni possono essere di vario tipo. Ad esempio, possono essere assegnati i valori della funzione  $u(\mathbf{x})$  sulla frontiera  $\partial\Omega$  di  $\Omega$ , cioè

$$u(\mathbf{x}) = g_1(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega \quad (7.159)$$

ove  $g_1(\mathbf{x})$  è una funzione assegnata su  $\partial\Omega$ . Il problema consistente nella risoluzione dell'equazione (7.158) con la condizione (7.159) è chiamato *problema di Dirichlet*.

Per quanto riguarda la risolubilità del problema, ricordiamo il seguente risultato, per la cui dimostrazione si veda ad esempio Courant-Hilbert [40].

**Proposizione 7.2** *Se l'insieme  $\Omega$  è sufficientemente regolare e  $g_1$  è una funzione continua su  $\partial\Omega$ , il problema di Dirichlet ammette una ed una sola soluzione, che dipende con continuità dal dato  $g_1$ .*

In maniera schematica, la dipendenza continua della soluzione  $u$  da  $g_1$  significa che se si ha una successione di dati  $g_1^{(n)}$  tali che, per  $n \rightarrow \infty$ ,  $g_1^{(n)} \rightarrow 0$  nel senso della convergenza uniforme, allora anche le corrispondenti soluzioni  $u^{(n)}$  tendono a zero nel senso della convergenza uniforme su  $\bar{\Omega}$ .

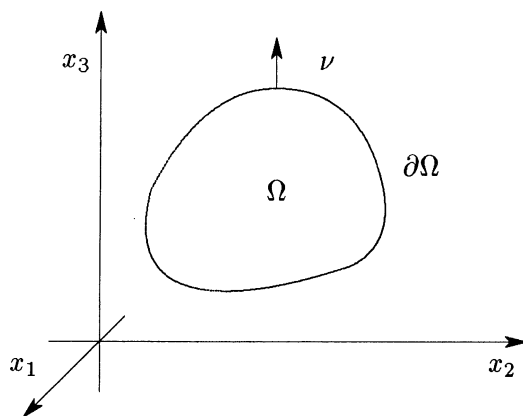


Figura 7.69: Regione  $\Omega$  in  $\mathbb{R}^3$ .

Un altro tipo di condizioni corrisponde ad assegnare su  $\partial\Omega$  il valore della derivata di  $u$  lungo la direzione della normale  $\nu$ , orientata, ad esempio, verso l'esterno (cfr.

<sup>31</sup>Più precisamente, si può dimostrare che se una funzione  $u$  ha le derivate prime e seconde continue in  $\Omega$ , è continua su  $\bar{\Omega}$  e verifica  $\Delta u = 0$ , con  $\Omega$  insieme aperto di  $\mathbb{R}^3$ , allora  $u$  non presenta né massimo locale stretto, né minimo locale stretto in  $\Omega$ , e pertanto i valori di  $u$  in  $\Omega$  sono compresi tra il minimo e il massimo di  $u$  su  $\partial\Omega$  (*principio del massimo*).

Figura 7.69), cioè

$$\frac{\partial u}{\partial \nu} = g_2(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega$$

con  $g_2(\mathbf{x})$  funzione assegnata. Il problema corrispondente è, allora, noto come *problema di Neumann*. Poiché  $\partial u/\partial \nu$  è proporzionale al *flusso* attraverso  $\partial\Omega$ , la condizione di *frontiera isolata* è

$$\frac{\partial u}{\partial \nu} = 0, \quad \mathbf{x} \in \partial\Omega$$

In questo caso la soluzione è determinata a meno di una costante additiva. Nel caso non omogeneo per l'esistenza della soluzione è necessario che la funzione assegnata  $g_2$  verifichi la seguente *condizione di compatibilità*

$$\int_{\partial\Omega} g_2 d\sigma = 0$$

ove  $d\sigma$  è l'elemento superficiale di  $\partial\Omega$ .

### Cambiamento di coordinate

Per la trattazione di problemi in particolari geometrie è utile esprimere l'operatore  $\Delta u$  in differenti sistemi di coordinate, ad esempio in *coordinate polari*, *sferiche* e *cilindriche*.

Consideriamo, come esemplificazione, la trasformazione in coordinate polari, definite da (cfr. Figura 7.70)

$$\begin{array}{ll} r^2 = x_1^2 + x_2^2 & x_1 = r \cos \theta \\ \theta = \tan^{-1}(x_2/x_1) & x_2 = r \sin \theta \end{array}$$

Utilizzando la regola di derivazione delle funzioni composte, si ottiene

$$u_{x_1} = u_r r_{x_1} + u_\theta \theta_{x_1} = u_r(\cos \theta) - u_\theta(\sin \theta/r)$$

Allo stesso modo, si ottiene

$$u_{x_2} = u_r r_{x_2} + u_\theta \theta_{x_2} = u_r(\sin \theta) + u_\theta(\cos \theta/r)$$

Procedendo in maniera analoga per le derivate seconde, si ha in definitiva la seguente formula

$$\Delta u = u_{x_1 x_1} + u_{x_2 x_2} = u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta}$$

che rappresenta l'operatore di Laplace in coordinate polari. Esso ha lo stesso significato intuitivo del laplaciano in coordinate cartesiane, ma una differente forma.

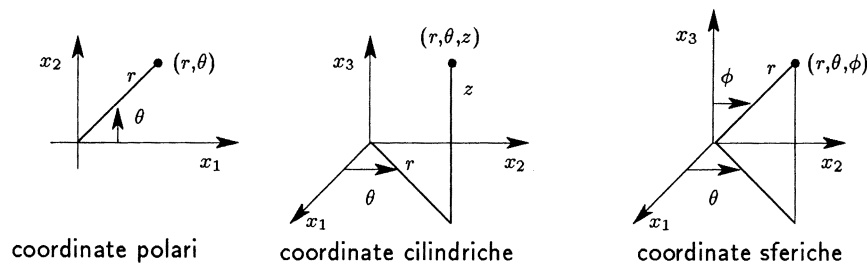


Figura 7.70: Trasformazione di coordinate.

Una analisi simile a quella precedente mostra che per le *coordinate cilindriche* definite da

$$\begin{array}{ll} r^2 = x_1^2 + x_2^2 & x_1 = r \cos \theta \\ \theta = \tan^{-1}(x_2/x_1) & x_2 = r \sin \theta \\ z = x_3 & x_3 = z \end{array}$$

si ha

$$\Delta u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} + u_{zz}$$

Infine, per le *coordinate sferiche* definite da

$$\begin{array}{ll} r^2 = x_1^2 + x_2^2 + x_3^2 & x_1 = r \sin \phi \cos \theta \\ \cos \phi = x_3/r & y = r \sin \phi \sin \theta \\ \tan \theta = x_2/x_1 & x_3 = r \cos \phi \end{array}$$

si ha

$$\Delta u = u_{rr} + \frac{2}{r}u_r + \frac{1}{r^2}u_{\phi\phi} + \frac{\cot \theta}{r^2}u_{\phi} + \frac{1}{r^2 \sin^2 \phi}u_{\theta\theta}$$

### Approssimazione mediante il metodo delle differenze finite

Per l'analisi delle idee di fondo dell'approssimazione numerica ci limiteremo al caso di un problema in  $\mathbb{R}^2$ , con  $\Omega$  corrispondente ad un rettangolo a lati paralleli agli assi. Più in particolare, indicando con  $(x, y)$  le variabili spaziali e posto

$$\Omega = \{(x, y) \mid a < x < b, c < y < d\}$$

considereremo la risoluzione del seguente problema

$$\Delta u := \frac{\partial^2 u(x, y)}{\partial x^2} + \frac{\partial^2 u(x, y)}{\partial y^2} = f(x, y), \quad (x, y) \in R \quad (7.160a)$$

$$u(x, y) = g(x, y), \quad (x, y) \in \partial R \quad (7.160b)$$

ove  $f(x, y), g(x, y)$  sono due funzioni continue assegnate. Introduciamo una *discretizzazione* del rettangolo  $\Omega$  mediante i nodi

$$\begin{aligned} x_i &= a + ih & \text{per } i = 0, 1, \dots, n \\ y_j &= a + jh & \text{per } j = 0, 1, \dots, m \end{aligned}$$

con  $m, n$  interi assegnati e  $h = (b - a)/n$  e  $k = (d - c)/m$ . Si approssima poi l'operatore  $\Delta$  mediante l'operatore alle differenze centrali.

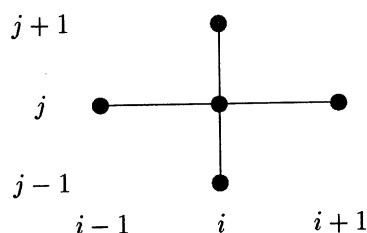


Figura 7.71: Schemi alle differenze centrali.

Procedendo come nei paragrafi precedenti, si ottengono le seguenti equazioni nella soluzione approssimata  $\bar{u}_{i,j}$  relativa al nodo  $(x_i, y_j)$

$$\frac{\bar{u}_{i+1,j} - 2\bar{u}_{i,j} + \bar{u}_{i-1,j}}{h^2} + \frac{\bar{u}_{i,j+1} - 2\bar{u}_{i,j} + \bar{u}_{i,j-1}}{k^2} = f(x_i, y_j) \quad (7.161)$$

per  $i = 1, 2, \dots, n - 1$  e  $j = 1, 2, \dots, m - 1$ . Mediante uno sviluppo in serie si verifica che l'*errore di troncamento locale* dipende dalle derivate quarte della funzione  $u(x, y)$ . In ogni equazione sono interessati cinque nodi, come mostrato in Figura 7.71. Dalla condizione (7.160b) si ricavano le seguenti condizioni ai limiti per la funzione approssimata

$$\bar{u}_{i,j} = g(x_i, y_j) \quad (7.162)$$

per  $i = 0, i = n$  e  $j = 0, 1, \dots, m$  e per  $j = 0, j = m$  e  $i = 1, 2, \dots, n - 1$ . L'insieme delle equazioni (7.161), (7.162) costituisce un *sistema lineare* nel vettore incognito  $\bar{u}_{i,j}$ , per  $i = 0, 1, \dots, n, j = 0, 1, \dots, m$ . Si vede facilmente che la matrice di tale sistema è simmetrica. Si può anche mostrare che è definita positiva e, quindi, non singolare. Per una conveniente risoluzione numerica del sistema è opportuno tenere conto che la matrice è *sparsa* e usualmente di grandi dimensioni, in quanto l'ordine è  $(n - 1) \times (m - 1)$ . Sono, quindi, interessanti i metodi iterativi, in particolare i metodi SOR e i metodi di tipo gradiente coniugato.

► **Esempio 7.45** Consideriamo il seguente problema di Dirichlet (cfr. Figura 7.72)

$$-u_{xx} - u_{yy} + \mu u - xy(\mu y^2 - 6) = 0, \quad (x, y) \in (0, 1) \times (0, 1)$$

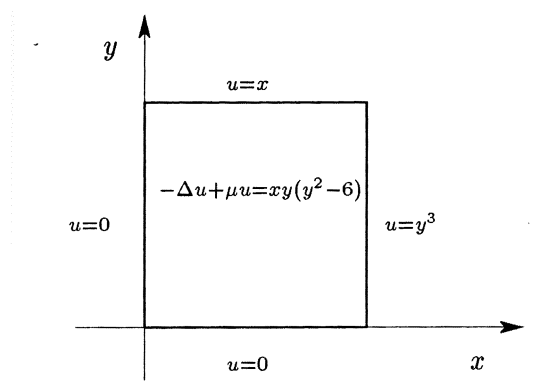


Figura 7.72: Esempio di problema di Dirichlet.

con  $\mu$  funzione costante, con le condizioni ai limiti

$$u(x, 0) = 0, \quad u(x, 1) = x, \quad u(0, y) = 0, \quad u(1, y) = y^3, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1$$

Si verifica facilmente che la soluzione esatta è data dalla funzione  $u = xy^3$ .

Posto  $h = 1/n$ ,  $k = 1/m$  e  $x_i = ih$ ,  $y_j = jk$ , si hanno le seguenti equazioni alle differenze

$$-\frac{\bar{u}_{i+1,j} - 2\bar{u}_{i,j} + \bar{u}_{i-1,j}}{h^2} - \frac{\bar{u}_{i,j+1} - 2\bar{u}_{i,j} + \bar{u}_{i,j-1}}{k^2} + \mu\bar{u}_{i,j} - x_i y_j (\mu y_j^2 - 6) = 0$$

per  $i = 1, \dots, n-1$ ,  $j = 1, \dots, m-1$ . Le condizioni ai limiti diventano

$$\bar{u}_{i,0} = 0, \quad \bar{u}_{i,m} = x_i, \quad \bar{u}_{0,j} = 0, \quad \bar{u}_{n,j} = y_j^3$$

Poiché l'errore di troncamento locale dipende dalle derivate di ordine quattro della soluzione  $u$ , in questo caso tale errore risulta identicamente nullo, e quindi la soluzione discreta coincide in ogni nodo con la soluzione continua. Il problema può, pertanto, essere di utilità come *problema test* per l'analisi dei metodi di risoluzione del sistema lineare. Si può dimostrare che la matrice dei coefficienti del sistema discreto ha i seguenti autovalori

$$\lambda_{rs} = \frac{4}{h^2} \sin^2 \left( \frac{1}{2} r \pi h \right) + \frac{4}{k^2} \sin^2 \left( \frac{1}{2} s \pi k \right) + \mu$$

per  $r = 1, 2, \dots, n-1$  e  $s = 1, 2, \dots, m-1$ . Per  $h$  e  $k$  sufficientemente piccoli si ha che il più piccolo autovalore  $\lambda_{11}$  è approssimativamente uguale a  $2\pi^2 + \mu$ . Pertanto la matrice ha autovalori positivi, e di conseguenza è *definita positiva*, per  $\mu > -2\pi^2 \approx -19.74$ . In tale caso il metodo SOR converge per ogni valore del parametro di rilassamento  $\omega$ , con  $0 < \omega < 2$ . Si può, anche, dimostrare che per il problema particolare che stiamo esaminando il valore del *parametro ottimale* è fornito da

$$\omega_{opt} \approx 2 - 2 \left( \pi^2 + \frac{1}{2} \mu \right)^{1/2} h$$

Di seguito, come esemplificazione, riportiamo un segmento di programma FORTRAN che implementa il metodo SOR per il sistema precedente. Osserviamo che la soluzione  $\bar{u}$  è

memorizzata, per convenienza, in una array a due indici. La matrice dei coefficienti  $\mathbf{A}$  non è esplicitamente costruita e memorizzata.

```

C   ITMAX: numero massimo di iterazioni
C   ERRMAX: errore tra due successive iterazioni
C   EPS:   test di arresto
C   OM:   parametro di rilassamento
C   AMU:  coefficiente dell'equazione

      DO 20 IT=1,ITMAX
        ERRMAX=0.
        DO 15 I=1,N-1
          DO 15 J=1,M-1
            GS=( (U(I+1,J)+U(I-1,J))/H**2+
&              (U(I,J+1)+U(I,J-1))/K**2+
&              +X(I)*Y(J)*(AMU*Y(J)**2-6.) )/(2/H**2+2/K**2+AMU)
            OLD=U(I,J)
            U(I,J)=OM*GS+(1-OM)*U(I,J)
            ERR=ABS(U(I,J)-OLD)
            ERRMAX=MAX(ERMAX,ERR)
15      CONTINUE
        IF(ERMAX.LT.EPS) STOP
20     CONTINUE

```

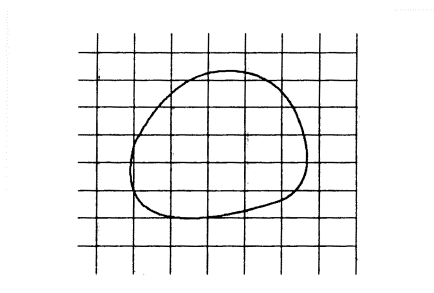


Figura 7.73: Reticolazione di un dominio generale.

L'estensione del metodo a domini  $\Omega$  di forma più generale presenta una difficoltà nella definizione dell'insieme discreto, in particolare dei punti di frontiera. Come esemplificazione, si veda la Figura 7.73. La difficoltà può essere superata in diverse maniere. Ad esempio, si possono assumere come punti di frontiera le intersezioni delle rette che definiscono la griglia con la frontiera  $\partial\Omega$ . Di conseguenza, è necessario modificare la definizione dell'operatore alle differenze, assumendo che i quattro punti vicini possano essere a distanze differenti. In questo modo possono essere superate le difficoltà relative al problema di Dirichlet, ma rimangono notevoli difficoltà nel discretizzare, ad esempio, il problema di Neumann. Un modo più generale di procedere consiste nell'utilizzare una decomposizione del dominio  $\Omega$  in figure geometriche



più opportune. Ad esempio, l'uso di triangoli permette di descrivere la frontiera in maniera più adeguata. Si tratta, in sostanza, di applicare il *metodo degli elementi finiti*, che abbiamo analizzato in precedenza nel caso unidimensionale. L'estensione di tale metodo ai problemi in più dimensioni pone, tuttavia, una serie di difficoltà tecniche, per la cui trattazione rinviamo alla letteratura specializzata.

... and it is probable that there is some secret here  
which remains to be discovered.

**C. S. Peirce**

He uses statistics as a drunken man uses lamp-posts  
for support rather than for illumination.

Andrew Lang

## Capitolo 8

# Probabilità e statistica

In questo capitolo vengono introdotti e analizzati gli elementi essenziali della *statistica* e del *calcolo delle probabilità*, ai quali è fatto riferimento nel resto del testo e che sono alla base delle tecniche per la costruzione e la validazione dei modelli matematici e la valutazione dei risultati sperimentali. Successivamente, verranno analizzate alcune importanti applicazioni, in particolare, i *processi di Markov* e la *teoria dell'informazione*.

In maniera schematica, la *statistica* può essere definita come l'arte e la scienza di generare o raccogliere, di descrivere, di analizzare, di sintetizzare, e di interpretare i dati provenienti da fenomeni naturali o da esperimenti allo scopo di ottenere nuova conoscenza.

Per quanto riguarda in particolare le misure sperimentali, queste possono essere sia *inesatte* che in *numero limitato*. In effetti, i risultati di una procedura sperimentale complicata possono essere influenzati da diversi fattori di disturbo, dovuti, in particolare alla diversità e alla variabilità del materiale da osservare e delle condizioni di osservazione. Di conseguenza, i risultati ottenuti in successivi esperimenti presentano una *dispersione*, che porta ad una *incertezza* nelle possibili conclusioni da trarre dall'esperimento. D'altra parte, la certificazione (test), ad esempio di parti meccaniche, può comportare la distruzione delle parti stesse e di conseguenza il test è possibile solo per una frazione piccola (ossia, un *campione*) delle parti in esame. Nelle due situazioni considerate, i metodi statistici possono essere utili, nel primo caso per stimare l'affidabilità dei risultati sperimentali ottenuti e nel secondo per estendere le informazioni ricavate dagli esperimenti disponibili a casi più generali.

In una tipica applicazione della statistica, ad una fase essenzialmente *descrittiva*, che comporta la riduzione delle informazioni, cioè la rappresentazione dei dati mediante un numero piccolo di *parametri* o caratteristiche (media, mediana, moda, varianza, ecc.), segue la ricerca di un *modello*, formulato in termini di *leggi statisti-*

che. Il successivo passo è quello dell'*interpretazione* (inference), che può essere la base per opportune *previsioni*.

## 8.1 Elementi di calcolo della probabilità

Uno degli obiettivi della statistica è quello di permettere decisioni in caso di risultati dispersi, cioè in presenza di incertezza. Tali decisioni sono espresse quantitativamente mediante la *teoria della probabilità*, che rappresenta pertanto lo strumento di base nelle applicazioni della statistica.

### 8.1.1 Probabilità matematica e probabilità statistica

Consideriamo, come esempio introduttivo, l'affermazione: *la probabilità di ottenere 1 da un lancio di un dado è un 1/6*. Tale affermazione trae fondamento dal fatto che, se il dado è fisicamente perfetto, in una successione di lanci il rapporto tra il numero di volte che esce 1 e il numero totale dei lanci tende a 1/6 all'aumentare dei lanci. Tale rapporto è chiamato *frequenza relativa*. L'ipotesi che il dado sia fisicamente perfetto significa che nel generico lancio ogni faccia ha la *stessa possibilità* di uscire, o equivalentemente che in una successione di lanci ogni faccia tende a uscire lo stesso numero di volte. Gli elementi dell'insieme  $\{1, 2, 3, 4, 5, 6\}$  sono i cosiddetti *casi possibili*, mentre l'elemento 1 è il *caso favorevole*. Osserviamo che i casi possibili si escludono a vicenda poiché non possono apparire simultaneamente due o più facce.

La definizione di *probabilità*, ora introdotta su un esempio, corrisponde alla seguente definizione storica, dovuta a Bernoulli (1654–1705) e a Laplace (1749–1827)<sup>1</sup>

$$P = \frac{\text{numero di casi favorevoli}}{\text{numero di casi possibili}} \quad (8.1)$$

Secondo tale definizione ogni probabilità  $P$  è un numero compreso tra zero e uno. Un caso impossibile ha probabilità zero, mentre un caso sicuro ha probabilità uno. Talvolta, la probabilità  $P$  è moltiplicata per 100 ed espressa come percentuale ( $0\% \leq P \leq 100\%$ ).

La definizione (8.1) di probabilità, chiamata anche *probabilità a priori* o *probabilità matematica*, assume che tutti i possibili risultati di un esperimento siano, come nel caso del dado fisicamente simmetrico, ugualmente probabili, cioè *statisticamente*

<sup>1</sup>*La théorie des hasards consiste à réduire tous les événements du même genre à un certain nombre de cas également possibles, c'est-à-dire tels que nous soyons également indécis sur leur existence, et à déterminer le nombre de cas favorables à l'événement dont on cherche la probabilité. Le rapport de ce nombre à celui de tous les cas possibles est la mesure de cette probabilité, qui n'est ainsi qu'une fraction dont le numérateur est le nombre des cas favorables, et dont le dénominateur est le nombre de tous les cas possibles.* (Théorie analytique des probabilités par M. Le Marquis de Laplace, 1820.) Una delle motivazioni storiche dell'introduzione del calcolo delle probabilità fu lo studio dei giochi d'azzardo; ci limitiamo a segnalare, in questo senso, il manuale *De ludo aleæ* dovuto a G. Cardano (1501–1576) (inveterato giocatore!) e pubblicato postumo nel 1663.

*simmetrici*. In caso contrario, il solo modo di calcolare la probabilità di un particolare evento consiste nel tenere conto dell'informazione ottenuta mediante un elevato numero di tentativi. La probabilità ottenuta in questo modo viene indicata come *probabilità a posteriori* o *probabilità statistica*. Ad esempio, l'affermazione: *la probabilità di una nascita di gemelli è 1/100* è da intendersi nel senso che la frequenza relativa osservata nell'arco di alcuni anni è stata 1:100; da tale constatazione si può assumere che una nascita futura sarà una nascita di gemelli con probabilità uguale a tale frequenza. I concetti ora introdotti verranno approfonditi nel seguito.

### 8.1.2 Elementi di calcolo combinatorio

Numerosi problemi in probabilità richiedono il calcolo del numero degli elementi in un insieme finito  $E$ , ossia, come anche si dice, della *cardinalità*  $\text{card } E$  ( $|E|$ , o anche  $\#(E)$ ) di  $E$ . In un esperimento a  $k$  stadi, nel quale al primo stadio si hanno  $n_1$  possibili risultati,  $n_2$  nel secondo per ogni possibile risultato al primo stadio, e così di seguito, il numero totale dei possibili risultati dell'esperimento è dato da

$$N = n_1 n_2 \cdots n_k = \prod_{i=1}^k n_i \quad (8.2)$$

In particolare, quando  $n_1 = n_2 = \cdots = n_k = n$ , si ha  $N = n^k$ , che rappresenta il numero delle *disposizioni con ripetizione* di  $n$  oggetti a gruppi di  $k$ .

**Disposizioni senza ripetizione** Siano  $E$  ed  $F$  due insiemi finiti, con  $p = \text{card } E$  e  $n = \text{card } F$ . Supponiamo anche, senza perdita di generalità, che  $E = \{1, 2, \dots, p\}$ . Calcoliamo allora il numero delle funzioni  $f: E \rightarrow F$ , che sono *iniezioni*, ossia tali che non esiste nessuna coppia  $i, j \in E$  con  $i \neq j$  e  $f(i) = f(j)$ . Naturalmente, per l'esistenza di tali funzioni deve essere  $p \leq n$ , nel qual caso, indicando con  $P(n, p)$  il numero delle iniezioni, si ha

$$P(n, p) = n(n-1) \cdots (n-p+1) = \frac{n!}{(n-p)!}$$

Infatti, per costruire una iniezione  $f: E \rightarrow F$ , si incomincia a scegliere  $f(1)$  in  $F$ , e questo può essere fatto in  $n$  modi diversi. Successivamente, una volta che  $f(1)$  è stato scelto, vi sono  $n-1$  scelte possibili, dal momento che  $f(2)$  deve differire da  $f(1)$ . Ragionando allo stesso modo, si vede che il generico  $f(i)$  è scelto in  $F - \{f(1), \dots, f(i-1)\}$  con  $(n-i+1)$  possibilità, e quindi il risultato enunciato.

**Permutazioni** Nel caso particolare in cui  $n = p$  si ottiene  $P(n, n) = n!$ . Se  $\text{card } E = \text{card } F$ , una iniezione di  $E$  in  $F$  è necessariamente una trasformazione biunivoca e, quando  $E \equiv F$ , viene detta una permutazione di  $E$  in se stesso. Pertanto, il numero  $P_n = P(n, n)$  delle permutazioni di un insieme con  $n$  elementi è dato da

$$P_n = n!$$

**Numero di sottoinsiemi di un insieme finito** Dato un insieme finito  $F$  con  $n = \text{card } F$  elementi, il numero dei differenti sottoinsiemi di  $p$  elementi, con  $p \leq n$  e senza ripetizione (*combinazioni* di  $n$  oggetti in gruppi di  $p$ ), può essere calcolato nel seguente modo. Si incomincia ad osservare che per quanto precede il numero dei sottoinsiemi *ordinati* di  $F$  con  $p$  elementi è dato da  $P(n, p)$ , in quanto ogni sottoinsieme ordinato  $\{x_1, \dots, x_p\}$ , con  $x_i \in F$  è identificabile con una iniezione  $f: \{1, \dots, p\} \rightarrow F$ , con  $f(i) = x_i$ . Osservando, quindi, che tutte le permutazioni di un sottoinsieme ordinato rappresentano lo stesso sottoinsieme non ordinato, il numero richiesto delle combinazioni di  $n$  oggetti in gruppi di  $p$  è dato da

$$C(n, p) := \binom{n}{p} = \frac{n!}{(n-p)!p!}$$

Da tale risultato si ricava che il numero di tutti i sottoinsiemi di un insieme finito  $F$ , con  $\text{card } F = n$ , è dato da  $\sum_{p=0}^n \binom{n}{p}$ , con la convenzione che  $\binom{n}{0} = 1$  (o equivalentemente  $0! = 1$ ). Osservando che il numero dei sottoinsiemi di  $F$  è uguale al numero delle sequenze di lunghezza  $n$  formate dai numeri 0 e 1, dalla formula (8.2) si ha il risultato

$$2^n = \sum_{p=0}^n \binom{n}{p} \quad (8.3)$$

che è un caso particolare della seguente formula, nota come *formula binomiale*

$$(x + y)^n = \sum_{p=0}^n \binom{n}{p} x^p y^{n-p} \quad (8.4)$$

valida per ogni  $x, y \in \mathbb{R}$ . Il caso precedente si ottiene per  $x = y = 1$ . Osservando che la formula (8.4) è simmetrica rispetto a  $x$  e  $y$ , si ottiene la seguente relazione

$$\binom{n}{p} = \binom{n}{n-p}$$

Terminiamo, ricordando la seguente formula, nota come *formula di Tartaglia* (1499–1557), o *formula di Pascal* (1623–1662), valida per ogni  $0 < p < n$

$$\binom{n}{p} = \binom{n-1}{p-1} + \binom{n-1}{p} \quad (8.5)$$

Per la dimostrazione consideriamo un elemento  $x_0$  dell'insieme  $F$ . Nella scelta di un sottoinsieme di  $p$  elementi, si hanno due possibilità, a seconda che l'elemento  $x_0$  appartiene o no al sottoinsieme. Se si vuole  $x_0$  nel sottoinsieme, per completare il sottoinsieme si devono scegliere  $p-1$  elementi dai rimanenti  $n-1$  elementi. Questo può essere fatto in  $\binom{n-1}{p-1}$  modi. Se non si vuole  $x_0$  nel sottoinsieme, si devono scegliere  $p$  elementi dai rimanenti  $n-1$ , e questo può essere fatto in  $\binom{n-1}{p}$  modi.

Poiché l'elemento  $x_0$  è nel sottoinsieme scelto oppure non vi appartiene, la somma dei due numeri precedenti fornisce il numero dei sottoinsiemi di  $p$  elementi che possono essere scelti da un insieme di  $n$  elementi, cioè  $\binom{n}{p}$ .

La relazione (8.5) insieme con la conoscenza che  $\binom{n}{0} = \binom{n}{n} = 1$  può essere utilizzata per la costruzione ricorrente dei numeri  $\binom{n}{p}$  (*triangolo di Pascal*).

► **Esempio 8.1** Si cerca la probabilità che in quattro successivi lanci di un dado simmetrico i risultati appaiano in ordine crescente. L'insieme dei casi possibili corrisponde all'insieme delle disposizioni con ripetizione di sei oggetti a quattro a quattro, il cui numero è dato da  $6^4$ . I casi favorevoli si verificano quando i risultati dei quattro lanci sono distinti e in ordine crescente. Il numero di tali eventi è dato dal numero delle combinazioni di sei oggetti a quattro a quattro, dal momento che come elemento rappresentativo di ogni combinazione si può scegliere la disposizione dei quattro numeri che segue l'ordine crescente. Pertanto, la probabilità cercata è data da

$$P = \frac{\binom{6}{4}}{6^4} \approx 0.0115$$

► **Esempio 8.2** Da un'urna contenente  $N_1$  palline nere e  $N_2$  palline rosse vengono estratte a caso  $n$  palline ( $n \leq N_1 + N_2$ ). Cerchiamo la probabilità di ottenere  $k$  palline nere ( $0 \leq k \leq \inf(N_1, n)$ ). L'insieme dei casi possibili corrisponde all'insieme delle combinazioni di  $N_1 + N_2$  in gruppi di  $n$ , il cui numero è dato da  $\binom{N_1 + N_2}{n}$ . Si ha un caso favorevole quando in un gruppo vi sono  $k$  palline nere e  $n - k$  palline rosse. Per calcolare il numero di tali gruppi, incominciamo ad osservare che con  $N_1$  palline nere si possono formare  $\binom{N_1}{k}$  gruppi di  $k$  elementi. Ad ogni sottoinsieme di  $k$  palline nere, si deve associare un sottoinsieme di  $n - k$  palline rosse; il numero di tali sottoinsiemi è dato da  $\binom{N_2}{n - k}$ . In definitiva, il numero dei casi favorevoli è dato da  $\binom{N_1}{k} \binom{N_2}{n - k}$  e la probabilità cercata è data da

$$P = \frac{\binom{N_1}{k} \binom{N_2}{n - k}}{\binom{N_1 + N_2}{n}}$$

► **Esempio 8.3** Da un'urna contenente  $N$  palline numerate da 1 a  $N$ , si estraggono simultaneamente  $n$  palline, con  $1 \leq n \leq N$ . Si cerca la probabilità che il numero più basso estratto sia  $k$  con  $k \leq N - n$ . Il numero dei sottoinsiemi di  $n$  palline tra le  $N$  palline è dato da  $\binom{N}{n}$ . Se la pallina numerata  $k$  è nel sottoinsieme e se essa è la pallina con il numero più basso, le rimanenti  $n - 1$  palline devono essere scelte tra  $N - k$  palline, cioè  $k + 1, \dots, N$ . Il numero di tale scelte è dato da  $\binom{N - k}{n - 1}$ . La probabilità cercata è allora data da  $\binom{N - k}{n - 1} / \binom{N}{n}$ .

### 8.1.3 Teoria assiomatica della probabilità

L'insieme di tutti i possibili risultati di un esperimento formano lo *spazio degli eventi elementari*, o spazio campione (*sample space*)  $S$ . I singoli risultati corrispondono agli elementi, o ai punti, di  $S$ . Ad esempio, lo spazio campione  $S$  corrispondente ad un singolo lancio di un dado consiste di 6 elementi, corrispondenti ai numeri da 1 a 6. Si tratta in questo caso di uno spazio *finito*. Se, tuttavia, consideriamo come evento il numero di volte che un dado deve essere lanciato prima di ottenere un 6, si

ha uno spazio di eventi elementari *infinito*, in quanto ogni numero intero positivo è un possibile risultato. Infine, se l'evento rappresenta, ad esempio, la misura di una grandezza, lo spazio  $S$  può corrispondere a tutti i punti di un intervallo della retta reale.

In maniera più precisa, indichiamo come *evento* un qualunque sottoinsieme  $E$  dello spazio degli eventi elementari  $S$ . Si dice che l'evento  $E$  si è *verificato* quando il risultato dell'esperimento appartiene al sottoinsieme  $E$  di  $S$ . In particolare, lo spazio intero  $S$  rappresenta l'*evento certo*. Nell'esempio del lancio di un dado l'evento  $E = S$  rappresenta l'evento che esca uno qualunque dei numeri  $\{1, 2, 3, 4, 5, 6\}$ .

Se  $E_1$  e  $E_2$  sono due eventi, il sottoinsieme degli eventi elementari  $E_1 \cup E_2$  ( $E_1$  unione  $E_2$ ), che consiste di tutti i punti che sono in  $E_1$  o in  $E_2$  (o in ambedue), corrisponde all'evento che si verifica quando esce almeno uno degli eventi  $E_1$  o  $E_2$ . Ad esempio, se  $E_1 = \{1, 3\}$ ,  $E_2 = \{3, 5\}$ , si ha  $E = E_1 \cup E_2 = \{1, 3, 5\}$ . L'insieme  $E$  caratterizza l'evento “ $E_1$  o  $E_2$  o ambedue”. Analogamente, l'insieme  $E_1 \cap E_2$  ( $E_1$  intersezione  $E_2$ , indicato talvolta anche con  $E_1 E_2$ ) caratterizza l'evento che si verifica quando esce un evento che appartiene sia ad  $E_1$  che ad  $E_2$ . Nell'esempio precedente si ha  $E_1 \cap E_2 = \{1, 3\} \cap \{3, 5\} = \{3\}$ . Quando  $E_1$  e  $E_2$  non hanno punti in comune (sono cioè insiemi disgiunti), si dice che gli eventi  $E_1$  e  $E_2$  sono *mutuamente esclusivi*. In questo caso l'operazione  $E_1 \cap E_2$  fornisce l'*insieme vuoto*  $\emptyset$ , che corrisponde all'*evento impossibile*.

Per ogni evento  $E$  si definisce, infine, l'insieme *complementare*  $\overline{E}$  di  $E$  (complemento logico *non*  $E$ , indicato anche con  $E^c$ ), costituito da tutti i punti di  $S$  che non sono in  $E$ . Nell'esempio del lancio del dado, se  $E = \{1, 2\}$ , allora  $\overline{E} = \{3, 4, 5, 6\}$ . Per definizione, si ha

$$\begin{aligned} E \cup \overline{E} &= S && \text{evento certo} \\ E \cap \overline{E} &= \emptyset && \text{evento impossibile} \end{aligned}$$

Sia ora  $S$  uno spazio campione e  $\mathcal{C}$  una collezione di eventi  $\{E_i\}$  (gli eventi di interesse) con le seguenti proprietà

- (i)  $S \in \mathcal{C}$
- (ii) Se  $E \in \mathcal{C}$ , allora  $\overline{E} \in \mathcal{C}$
- (iii)  $E_1, E_2 \in \mathcal{C}$ , allora  $E_1 \cup E_2 \in \mathcal{C}$

Si dice allora che  $\mathcal{C}$  è un'*algebra* di eventi. Come semplice illustrazione, si consideri nell'esperimento corrispondente al lancio di un dado la collezione formata dai quattro eventi  $\{\emptyset\}$ ,  $\{\text{dispari}\}$ ,  $\{\text{pari}\}$ ,  $\{S\}$ ; tale scelta è opportuna quando si è interessati solo a controllare se il risultato di un lancio è *dispari* o *pari*. In pratica, per costruire un'algebra di eventi si può partire dalla collezione di eventi che interessano in una data applicazione, allargandola successivamente, se necessario, per includere l'evento certo, tutti i complementari di eventi già compresi e tutte le unioni e le intersezioni

finite di eventi già compresi. Naturalmente, la collezione di tutti i sottoinsiemi di  $S$  è un'algebra. Tuttavia, quando lo spazio campione  $S$  non è finito, la scelta  $\mathcal{C} = S$  può essere impossibile per motivi matematici legati alla definizione di probabilità<sup>2</sup>.

Si *assume* che ad ogni evento  $E \in \mathcal{C}$  sia associato un numero reale  $P(E)$ , che diremo *probabilità* dell'evento  $E$ , quando sono verificati i seguenti *assiomi*<sup>3</sup>

$$0 \leq P(E) \leq 1 \quad (8.6)$$

$$P(S) = 1 \quad (8.7)$$

se  $\{E_i\}$  è una successione di eventi di  $\mathcal{C}$  con  $E_i \cap E_j = \emptyset$ , per  $i \neq j$ ,  $i, j = 1, 2, \dots$  e  $E_1 \cup E_2 \cup \dots = \cup_{i=1}^{\infty} E_i \in \mathcal{C}$ , allora

$$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) \quad (8.8)$$

L'individuazione della collezione  $\mathcal{C}$  e l'assegnazione di una particolare funzione  $P$  costituisce un *modello matematico*, la cui validità deve essere verificata sperimentalmente, in maniera analoga a quanto viene fatto, per esempio, nel caso di leggi fisiche. In definitiva, il problema di assegnare una specifica probabilità in una situazione pratica è un problema di natura statistica<sup>4</sup>. Osserviamo, inoltre, che gli assiomi precedenti sono verificati (in pratica, suggeriti) dalla nozione di frequenza relativa.

► **Esempio 8.4** Supponiamo che lo spazio campionario  $S$  sia costituito da un numero finito  $n$  di elementi  $S = \{s_1, \dots, s_n\}$ , ove  $s_j$  sono gli *eventi elementari*, e sia  $\mathcal{C}$  l'insieme di tutti i sottoinsiemi di  $S$ . Definiamo, allora,  $P$  nel seguente modo

$$P(s_j) = p_j, \quad j = 1, 2, \dots, n, \quad p_j \geq 0, \quad \sum_{j=1}^n p_j = 1$$

Per ogni  $E \in \mathcal{C}$ , e quindi  $E = \cup_j \{s_{i_j}\}$ , unione finita di eventi elementari, si definisce

$$P(E) = \sum_j P(s_{i_j})$$

Si verifica immediatamente che  $P$  è una probabilità. In particolare, quando  $p_1 = p_2 = \dots = p_n = 1/n$  si ha

$$P(E) = \frac{\#(E)}{n}$$

<sup>2</sup>In sostanza, quando  $S$  è infinito si richiede che se  $E_1, E_2, \dots, E_n, \dots \in \mathcal{C}$ , allora  $\cup_{n=1}^{\infty} E_n \in \mathcal{C}$ . Allora,  $\mathcal{C}$  viene detta una  $\sigma$ -algebra (o tribu). In effetti, si può mostrare che esistono insiemi di punti sulla retta reale che non sono unioni e intersezioni numerabili di intervalli.

<sup>3</sup>La teoria assiomatica della probabilità ha avuto inizio abbastanza recentemente, in pratica con il lavoro: A. Kolmogoroff *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Ergeb. Math und ihrer grensg., **2**, 1933.

<sup>4</sup>Il punto di vista ora illustrato, ossia che la probabilità di un evento sia una "misura" della verosimiglianza che l'evento avvenga, e quindi che, sebbene  $P(E)$  non sia osservabile, sia comunque empiricamente verificabile, è detto punto di vista *frequentistico*, o obiettivo. Nel punto di vista cosiddetto *soggettivo*, la probabilità  $P(E)$  è, invece, un numero compreso tra 0 e 1 che rappresenta la valutazione di una persona della verosimiglianza dell'evento.



ove  $\#(E)$  è il numero degli elementi in  $E$ . Questa assegnazione, detta *uniforme*, è la base dello studio dei vari giochi di azzardo, come il lancio di una moneta, di un dado e dei vari giochi di carte. In presenza di assegnazione uniforme, il problema del calcolo della probabilità di un evento si riduce a quello del calcolo del numero degli elementi in  $E$ .

Come illustrazione, consideriamo l'esperimento del lancio di una moneta, le cui facce sono individuate da  $T$  (testa) e  $C$  (croce), e quindi  $S = \{T, C\}$ . Definiamo  $P$  ponendo

$$P(T) = p = 1 - P(C), \quad 0 \leq p \leq 1$$

Allora  $P$  definisce una probabilità su  $(S, \mathcal{C})$ , ove  $\mathcal{C} = \{\{\emptyset\}, \{T\}, \{C\}, \{T, C\}\}$ . Se la moneta è simmetrica, possiamo assumere  $p = 1/2$ , dal momento che  $\#(S) = 2$ .

Se la moneta è lanciata tre volte, allora si ha

$$S = \{CCC, CCT, CTC, TCC, CTT, TCT, TTC, TTT\}$$

e possiamo assumere  $p_1 = p_2 = \dots = p_8 = \frac{1}{8}$ . Allora, ad esempio

$$P(\text{esattamente due croci}) = P(CCT, CTC, TCC) = \frac{3}{8}$$

$$P(\text{nessuna croce}) = P(TTT) = \frac{1}{8}$$

► **Esempio 8.5** Consideriamo l'esperimento del lancio successivo di una moneta, con  $P(C) = p$ ,  $0 \leq p \leq 1$ . L'esperimento termina quando appare per la prima volta  $C$ . In questo caso  $S = \{C, TC, TTC, TTTC, \dots\}$  è un insieme infinito numerabile. Consideriamo, per motivi che vedremo nel seguito, la seguente assegnazione

$$p_k = P(\underbrace{TTT \dots T}_{k-1}C) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$

Si ha allora  $0 \leq p_k \leq 1$ , per  $k = 1, 2, \dots$  e

$$P(S) = \sum_{k=1}^{\infty} p_k = p \sum_{k=1}^{\infty} (1-p)^{k-1} = \frac{p}{1 - (1-p)} = 1$$

Ne segue che  $P$  definisce una probabilità su  $\mathcal{S}$ , ove  $\mathcal{S}$  è la classe di tutti i sottoinsiemi di  $S$ . Ad esempio, se  $E = \{\text{sono richiesti almeno } n \text{ lanci}\}$ ,  $n = 1, 2, \dots$  si ha

$$P(E) = \sum_{k=n}^{\infty} p_k = p(1-p)^{n-1} \sum_{k=0}^{\infty} (1-p)^k = (1-p)^{n-1}$$

► **Esempio 8.6** Consideriamo il numero degli arrivi a una stazione di servizio in un intervallo di lunghezza  $t$ , misurato in ore. Allora,  $S = \{0, 1, 2, \dots\}$ ; come  $\mathcal{C}$  consideriamo la classe di tutti i sottoinsiemi di  $S$ . Come vedremo nel seguito, una assegnazione "ragionevole" di probabilità è la seguente

$$p_k = P(k) = \exp(-\lambda t) \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \dots$$

ove  $\lambda > 0$  è una costante da scegliere opportunamente nelle singole applicazioni. Si ha

$$P(S) = \sum_{k=0}^{\infty} p_k = e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\lambda t} e^{\lambda t} = 1$$

Posto, per  $E \in \mathcal{C}$ ,  $P(E) = \sum_{k \in E} p_k$ , si verifica che  $P$  definisce una probabilità su  $\mathcal{C}$ . Se, ad esempio,  $E = \{\text{non più di un arrivo in 15 minuti}\}$ , allora  $t = 1/4$  e

$$P(E) = P(\{0, 1\}) = e^{-\lambda/4} + e^{-\lambda/4} \frac{(\lambda/4)}{1!} = e^{-\lambda/4} \left(1 + \frac{\lambda}{4}\right)$$

► **Esempio 8.7** Consideriamo l'esperimento consistente nella scelta di un punto da un intervallo fissato  $[a, b]$  della retta reale; si ha quindi  $S = [a, b]$ . Quando si è interessati al calcolo della probabilità che un punto scelto appartenga ad un sottointervallo  $[c, d] \subset [a, b]$ , gli eventi di interesse sono i sottointervalli di  $S$ ; la collezione  $\mathcal{C}$  è pertanto costituita dall'insieme dei sottointervalli, insieme alle loro unioni e intersezioni (in effetti,  $\mathcal{C}$  è assunto come la più piccola  $\sigma$ -algebra contenente sottointervalli di  $S$ ; tale insieme è detto  $\sigma$ -algebra di Borel).

Per assegnare una probabilità agli eventi in  $\mathcal{C}$ , si introduce una funzione  $f(x)$  integrabile su  $S$  e tale che  $\int_a^b f(x) dx = 1$ . Per ogni sottointervallo  $I$  di  $S$ , si definisce

$$P(I) = \int_I f(x) dx$$

e se  $E \in \mathcal{C}$  è l'unione di sottointervalli disgiunti  $I_k$  di  $S$ , si definisce

$$P(E) = \int_E f(x) dx = \sum_k \int_{I_k} f(x) dx$$

Si può allora mostrare che  $P$  definisce una probabilità su  $S$ . In particolare, l'assegnazione di probabilità uniforme corrisponde a definire, per ogni  $[c, d] \in \mathcal{C}$ ,  $P([c, d]) = (d - c)/(b - a)$ . In questo modo agli intervalli di uguale ampiezza è assegnata la medesima probabilità. Altre esemplificazioni verranno considerate nel seguito nell'ambito dello studio delle funzioni di densità di probabilità. ■

Come conseguenza degli assiomi (8.6), (8.8), (8.9), si ha  $1 = P(S) = P(E \cup \overline{E}) = P(E) + P(\overline{E})$ , cioè

$$P(E) = 1 - P(\overline{E}) \quad (8.9)$$

La probabilità che si verifichi almeno uno degli eventi  $E_1$  e  $E_2$ , quando  $E_1$  e  $E_2$  non sono mutuamente esclusivi è data da

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \quad (8.10)$$

Dalla Figura 8.1 si vede, infatti, che aggiungendo semplicemente  $P(E_1)$  e  $P(E_2)$ , la probabilità  $P(E_1 \cap E_2)$  verrebbe contata due volte. La (8.10) è la *regola additiva* per eventi arbitrari che non sono necessariamente mutuamente esclusivi.

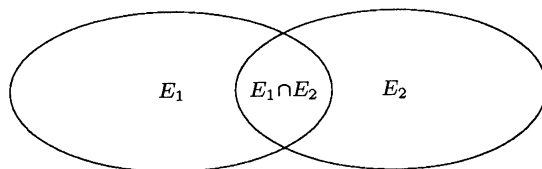


Figura 8.1: Diagramma di Venn.

Consideriamo, come illustrazione, il lancio di un dado simmetrico. I due eventi  $E_1 = \{1, 2\}$  e  $E_2 = \{2, 3\}$  non sono mutuamente esclusivi, in quanto hanno  $E_1 \cap E_2 = 2$  come risultato comune. L'evento  $E_1 \cup E_2$  è dato da  $\{1, 2, 3\}$ . Trattandosi di un dado simmetrico, si ha  $P(E_1) = 1/3$ ,  $P(E_2) = 1/3$  e  $P(E_1 \cup E_2) = 1/3 + 1/3 - 1/6 = 1/2$ .

La legge (8.10) si generalizza al caso di più eventi. Ad esempio, nel caso di tre eventi  $A, B, C$  arbitrari si ha

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

e nel caso di  $A_1, \dots, A_n$ , con  $n \geq 2$ , eventi (non necessariamente disgiunti) si ha

$$P(\cup A_i) = \sum_i P(A_i) - \sum_{j < i} \sum_{k < j} P(A_i \cap A_j) + \sum_{k < j < i} \sum_{l < k} \sum_{m < l} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} P(\cap_i A_i)$$

detta formula di Poincaré, o anche *inclusion-exclusion formula*.

► **Esempio 8.8** Supponiamo che  $n$  individui si presentino a ritirare la valigia da un deposito, che restituisce le valigie in modo casuale. Si cerca la probabilità che almeno una persona riceva la propria valigia. Sia  $A_i$  l'evento che una persona abbia la propria valigia. Si vuole, allora, calcolare  $P(\cup A_i)$ . Dal momento che le valigie sono restituite a caso, si ha  $P(A_i) = 1/n$  e pertanto

$$\sum_i P(A_i) = \frac{n}{n} = 1$$

Inoltre, la persona  $i$ -ma ha  $n$  scelte, e dopo che essa ha scelto, la persona  $j$ -ma ha  $n-1$  scelte. Pertanto, la probabilità che le persone  $i$ -ma e  $j$ -ma ricevano ambedue la valigia corretta è  $1/n(n-1)$ , cioè  $P(A_i \cap A_j) = 1/n(n-1)$ , da cui

$$\sum_{j < i} \sum_{k < j} P(A_i \cap A_j) = \frac{\binom{n}{2}}{n(n-1)} = \frac{1}{2!}$$

In maniera analoga, si ottiene

$$\sum_{k < j < i} \sum_{l < k} \sum_{m < l} P(A_i \cap A_j \cap A_k) = \frac{\binom{n}{3}}{n(n-1)(n-2)} = \frac{1}{3!}$$

Si vede, allora, che la probabilità  $p_n$  che almeno uno degli  $n$  individui abbia la propria valigia è data da

$$p_n = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots + \frac{(-1)^{n+1}}{n!} = \sum_{i=1}^n \frac{(-1)^{i+1}}{i!}$$

Tenendo presente che  $e^{-1} = \sum_{i=0}^{\infty} (-1)^i / i! = 1 - \sum_{i=1}^{\infty} (-1)^{i+1} / i!$ , si ha il seguente risultato

$$\lim_{n \rightarrow \infty} p_n = 1 - e^{-1} \approx 0.6321$$

Si può, pertanto, concludere che, mentre la probabilità che una *determinata* persona abbia la propria valigia tende a zero per  $n$  che tende all'infinito, la probabilità che *almeno* una persona abbia la propria valigia tende a 0.6321. ■

### 8.1.4 Probabilità condizionata e indipendenza statistica

Consideriamo il seguente esempio introduttivo. Un'urna contiene 15 palline rosse e 5 palline nere. Indichiamo con  $E_1$  l'evento corrispondente all'estrazione di una pallina rossa e con  $E_2$  quello relativo alla pallina nera. Si cerca la probabilità di ottenere in due consecutive estrazioni prima una pallina rossa e successivamente una pallina nera, nell'ipotesi che la pallina estratta non venga reintrodotta nell'urna. La probabilità di estrarre una pallina rossa è data da  $P(E_1) = 15/20 = 3/4$ . La probabilità di estrarre una pallina nera, una volta che sia stata rimossa una pallina rossa, è data da  $5/19$ . In pratica, la conoscenza del verificarsi dell'evento  $E_1$  ha ridotto lo spazio degli eventi. La probabilità ottenuta in questo modo è detta *probabilità condizionata* (o subordinata) e indicata con il simbolo  $P(E_2 | E_1)$ , ove il simbolo  $|$ , come è usuale nella teoria degli insiemi, indica "a condizione che". La probabilità  $P(E_1 \cap E_2)$  di ottenere in due estrazioni senza sostituzione una pallina rossa e successivamente una pallina nera è data da  $P(E_1)P(E_2 | E_1) = \frac{3}{4} \frac{5}{19} = \frac{15}{76}$ .

Più in generale, si ha la seguente definizione.

**Definizione 8.1** *La probabilità di un evento  $E_2$ , nell'ipotesi che sia già verificato l'evento  $E_1$ , è chiamata probabilità condizionata ed è definita nel modo seguente*

$$P(E_2 | E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} \quad (8.11)$$

quando  $P(E_1) \neq 0$ .

Analogamente, per  $P(E_2) \neq 0$  si ha

$$P(E_1 | E_2) = \frac{P(E_2 \cap E_1)}{P(E_2)}$$

Tale risultato porta alla seguente *regola della moltiplicazione* relativa alla probabilità del verificarsi simultaneo degli eventi  $E_1$  e  $E_2$

$$P(E_1 \cap E_2) = P(E_1)P(E_2 | E_1) = P(E_2)P(E_1 | E_2) = P(E_2 \cap E_1) \quad (8.12)$$

Il risultato è illustrato in Figura 8.2.

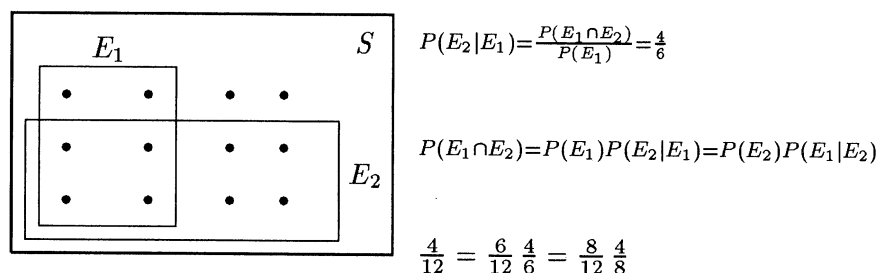


Figura 8.2: Probabilità condizionata.

**Definizione 8.2** Due eventi sono chiamati stocasticamente<sup>5</sup> indipendenti se

$$P(E_2 | E_1) = P(E_2) \quad (8.13)$$

In tale caso si ha pure

$$P(E_1 | E_2) = P(E_1)$$

Osserviamo che se  $E_1$  e  $E_2$  sono stocasticamente indipendenti, allora sono stocasticamente indipendenti gli eventi

- $\bar{E}_1, E_2 \iff P(E_2 | E_1) = P(E_2 | \bar{E}_1) = P(E_2)$
- $E_1, \bar{E}_2 \iff P(E_1 | E_2) = P(E_1 | \bar{E}_2) = P(E_1)$

Dalla formula (8.11), se due eventi  $E_1, E_2$  sono stocasticamente indipendenti, si ha

$$\boxed{P(E_1 \cap E_2) = P(E_1) P(E_2)} \quad (8.14)$$

Il risultato si generalizza al caso di un evento che risulta da  $n$  esperimenti stocasticamente indipendenti con risultati  $E_i, i = 1, 2, \dots, n$ , nel seguente modo

$$P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1) P(E_2) \dots P(E_n) = \\ P(E_1) P(E_2 | E_1) P(E_3 | E_1 \cap E_2) \dots P(E_n | E_1 \cap E_2 \cap \dots \cap E_{n-1})$$

▼ **Osservazione 8.1** Osserviamo che eventi disgiunti non sono indipendenti. Infatti, in caso contrario per ogni coppia di eventi disgiunti  $E_1$  e  $E_2$  almeno uno di essi dovrebbe essere di probabilità zero, in quanto  $0 = P(\emptyset) = P(E_1 \cap E_2) = P(E_1) P(E_2)$ . In realtà, due eventi disgiunti sono fortemente dipendenti, in quanto “disgiunti” significa “incompatibili”: se uno di essi è realizzato, allora si sa che l'altro non lo è. ■

► **Esempio 8.9** Data la seguente tabella

<sup>5</sup>stocastico ( $\sigma\tau\acute{o}\chi\omicron\varsigma =$  congettura) significa associato con esperimenti aleatori.

	$P(E_1)$	$P(E_2)$	$P(E_1 \cup E_2)$
caso 1	0.1	0.9	0.91
caso 2	0.4	0.6	0.76
caso 3	0.5	0.3	0.73

esaminare per quali casi gli eventi  $E_1$  e  $E_2$  sono stocasticamente indipendenti. Tenendo conto che  $P(E_1 \cap E_2) = P(E_1) + P(E_2) - P(E_1 \cup E_2)$ , dal risultato (8.14), si ottiene

	$P(E_1 \cap E_2)$	$P(E_1)P(E_2)$	indipendenza
caso 1	0.09	0.09	si
caso 2	0.24	0.24	si
caso 3	0.07	0.15	no

► **Esempio 8.10** Consideriamo l'esperimento: un dado simmetrico è lanciato due volte, oppure, equivalentemente, due dadi simmetrici sono lanciati simultaneamente. Lo spazio  $S$  è costituito dalle coppie  $\{(1,1), (1,2), \dots, (1,6), (2,1), \dots, (5,6), (6,6)\}$ . Gli eventi  $E_1, E_2$  sono definiti nel modo seguente

- $E_1 =$  primo lancio pari  $= \{2, 4, 6\}$
- $E_2 =$  secondo lancio minore o uguale a 2  $= \{1, 2\}$

Essendo i lanci indipendenti, si ha  $P(E_1 \cap E_2) = P(E_1)P(E_2) = \frac{3}{6} \frac{2}{6} = \frac{1}{6}$ . Osserviamo che i casi favorevoli all'evento  $E_1 \cap E_2$  corrispondono alle sei coppie  $\{(2,1), (2,2), (4,1), (4,2), (6,1), (6,2)\}$ , mentre i casi possibili sono 36.

► **Esempio 8.11** Si cerca la probabilità  $P$  di ottenere tre sei in tre lanci successivi di un dado simmetrico. Trattandosi di eventi indipendenti, si ha  $P = \frac{1}{6} \frac{1}{6} \frac{1}{6} = \frac{1}{216}$ .

► **Esempio 8.12** Un dado simmetrico è lanciato quattro volte. Si cerca la probabilità di ottenere un sei almeno una volta. Sostituendo l'evento "un sei almeno una volta" con l'evento complementare "nessun sei", si ha che la probabilità di non ottenere un sei in un singolo lancio è  $\frac{5}{6}$ , e quindi la probabilità di non ottenere un sei in quattro lanci è uguale a  $(\frac{5}{6})^4$ . Pertanto, la probabilità di ottenere almeno un sei con quattro lanci è  $1 - (\frac{5}{6})^4 \approx 0.518$ .

► **Esempio 8.13** Una persona, indicata con A, lancia una moneta simmetrica  $n$  volte e ottiene un numero  $C_A$  di croci. Una seconda persona, B, lancia la moneta  $n+1$  volte e ottiene  $C_B$  croci. Si cerca la probabilità che  $C_A \geq C_B$ . Se indichiamo con  $T_A$ , e rispettivamente  $T_B$ , il numero di teste ottenute da A e da B, per simmetria si avrà

$$P(C_A \geq C_B) = P(T_A \geq T_B) \quad (8.15)$$

Ma  $T_A = n - C_A$  e  $T_B = n + 1 - C_B$ , per cui  $T_A \geq T_B \iff C_A < C_B$ . Quindi,  $P(T_A \geq T_B) = P(C_A < C_B) = 1 - P(C_A \geq C_B)$  e da (8.15) si ottiene  $P(C_A \geq C_B) = \frac{1}{2}$ .

### Teorema di Bayes

Siano  $A_1, A_2, \dots, A_n$  eventi mutuamente esclusivi e tali che l'unione di tutti gli eventi  $A_i$  sia l'evento certo; in altre parole, gli  $n$  eventi  $A_i$  formino una partizione di un spazio  $S$  di eventi elementari. Supponiamo, inoltre, che un evento casuale  $E$ , con  $P(E) > 0$  e che può accadere solo in combinazione con un evento  $A_i$ , sia già avvenuto. Allora, la probabilità che accada un evento  $A_k$  è data da

$$P(A_k | E) = \frac{P(A_k \cap E)}{P(E)} = \frac{P(A_k)P(E | A_k)}{P(A_1)P(E | A_1) + \dots + P(A_n)P(E | A_n)} \quad (8.16)$$

Il risultato (8.16), noto come *Teorema di Bayes*<sup>6</sup>, è illustrato dai seguenti esempi.

► **Esempio 8.14** Si hanno a disposizione due urne, indicate rispettivamente con I e II. Supponiamo che la probabilità di scegliere I sia  $\frac{1}{10}$  e la probabilità di scegliere II sia allora  $\frac{9}{10}$ . Supponiamo, inoltre, che le urne contengano delle palline bianche e nere. Più precisamente, nell'urna I il 70% delle palline sono nere, mentre nell'urna II le nere sono il 40%. Si cerca la probabilità che una pallina nera estratta a caso provenga dall'urna I. In questo caso  $E$  è l'evento che la pallina sia nera,  $A_1$  l'evento che essa sia stata estratta dall'urna I, e  $A_2$  l'evento che essa provenga dall'urna II. Applicando il risultato (8.16), si ha (cfr. Figura 8.3 per una illustrazione sotto forma di grafo)

$$P(\text{dall'urna I} | \text{nera}) = \frac{0.1 \cdot 0.7}{0.1 \cdot 0.7 + 0.9 \cdot 0.4} = 0.163$$

Il risultato può essere interpretato nel seguente modo. Se si effettuano numerosi tentativi, è giustificato concludere che nel 16.3% di tutti i casi nei quali è estratta una pallina nera, essa provenga dall'urna I.

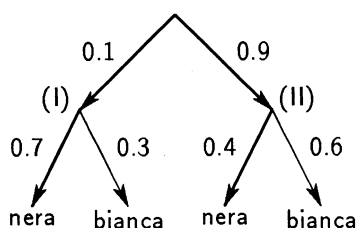


Figura 8.3: Grafo associato all'Esempio 8.14.

► **Esempio 8.15** Due macchine, indicate rispettivamente con I e II, sono impiegate per la produzione di un determinato articolo, che è sottoposto successivamente a collaudo, e quindi eventualmente scartato. Si conoscono le percentuali di scarto della produzione delle

<sup>6</sup>Il risultato è delineato in un lavoro pubblicato postumo (1763) di Thomas Bayes (1702-1761). In maniera esplicita fu enunciato da Laplace (1812).

due macchine ed il rapporto fra le loro produzioni. Si vuole conoscere la probabilità che un pezzo sia difettoso e la probabilità che un pezzo difettoso sia stato prodotto, ad esempio, dalla macchina I. Indichiamo con  $E$  l'evento "il pezzo è difettoso", con  $A_1$  l'evento "il pezzo è prodotto da I", e con  $A_2$  l'evento "il pezzo è prodotto da II". Se  $p_1$  è la percentuale di scarto per la macchina I, si ha  $P(E | A_1) = p_1$ , e analogamente, se  $p_2$  è la percentuale di scarto per la macchina II, abbiamo  $P(E | A_2) = p_2$ . Se indichiamo, poi, con  $n_1/n_2$  il rapporto tra le produzioni di I e II, possiamo assumere  $P(A_1) = n_1/(n_1 + n_2)$  e  $P(A_2) = n_2/(n_1 + n_2)$ . Se, ad esempio, assumiamo  $p_1 = 0.08$ ,  $p_2 = 0.06$  e  $n_1/n_2 = 3/2$ , si ha

$$P(E) = \frac{3}{5} 0.08 + \frac{2}{5} 0.06 = 0.072$$

mentre la probabilità che un pezzo difettoso sia stato prodotto dalla macchina I corrisponde al valore

$$P(A_1 | E) = \frac{3}{5} \frac{0.08}{0.072} = \frac{2}{3}$$

► **Esempio 8.16** *Legge di Hardy-Weinberg.* Supponiamo che un gene sia o dominante (A) o recessivo (a). Ogni individuo nella popolazione ha due geni, e quindi le possibili combinazioni genetiche sono AA, Aa, e aa. Supponiamo che nella prima generazione le proporzioni di ogni tipo siano  $p=P(AA)$ ,  $q=P(Aa)$ , e  $r=P(aa)$  (con  $p+q+r=1$ ) sia per i maschi che per le femmine. Supponiamo che tutti gli incontri siano casuali, in maniera che il contributo della madre sia indipendente dal contributo del padre. Allora, la probabilità che il padre trasmetta un A al figlio è  $p + q/2$ , dal momento che è sicuro di trasmettere un A se egli è AA, e trasmette A con probabilità  $\frac{1}{2}$  se esso è Aa. In maniera analoga, la probabilità che la madre trasmetta un A è  $p + q/2$ . Poiché i contributi del padre e della madre sono indipendenti, la probabilità che il discendente sia AA è data da

$$p^* = \left(p + \frac{q}{2}\right)^2$$

In modo analogo, si vede che le probabilità  $q^*$  e  $r^*$  che il discendente sia rispettivamente Aa e aa sono date da

$$q^* = 2 \left(p + \frac{q}{2}\right) \left(r + \frac{q}{2}\right), \quad r^* = \left(r + \frac{q}{2}\right)^2$$

Indicando con  $p^{**}$ ,  $q^{**}$  e  $r^{**}$  le proporzioni di AA, Aa e aa nella terza generazione, dalle formule precedenti si trova

$$\begin{aligned} p^{**} &= \left(p^* + \frac{q^*}{2}\right)^2 = \left[\left(p + \frac{q}{2}\right)^2 + \left(p + \frac{q}{2}\right) \left(r + \frac{q}{2}\right)\right]^2 \\ &= \left[\left(p + \frac{q}{2}\right) (p + q + r)\right]^2 = \left(p + \frac{q}{2}\right)^2 = p^* \end{aligned}$$

In maniera analoga si ha  $q^{**} = q^*$  e  $r^{**} = r^*$ . Pertanto, le proporzioni nelle successive generazioni sono ancora  $p^*$ ,  $q^*$  e  $r^*$ . In maniera, a priori sorprendente, le proporzioni si sono stabilizzate in una sola generazione. Sottolineiamo che tale risultato, noto come legge di Hardy-Weinberg, è valido sotto l'ipotesi di incontri casuali, e quindi di indipendenza dei contributi genetici. In realtà, se ad esempio il gene influisce sull'altezza di un individuo, gli incontri possono non essere casuali, dal momento che possono essere favoriti gli incontri tra individui di altezze non troppo diverse.



► **Esempio 8.17** In un sistema di comunicazione digitale, si trasmettono i bit 0 e 1 attraverso un canale disturbato (noisy) in maniera che (cfr. Figura 8.4) il segnale trasmesso differisca dal segnale ricevuto con probabilità  $p$  (*canale binario simmetrico*). Supponendo che uno 0 sia emesso con probabilità  $\pi_0$  e un 1 con probabilità  $\pi_1 = 1 - \pi_0$ , si cerca la probabilità di ottenere 1 come segnale ricevuto. Indichiamo con  $X$  e  $Y$  le variabili casuali input e output. Allora

$$P(Y = 1) = P(Y = 1 | X = 0)P(X = 0) + P(Y = 1 | X = 1)P(X = 1)$$

da cui

$$P(Y = 1) = p\pi_0 + (1 - p)\pi_1$$

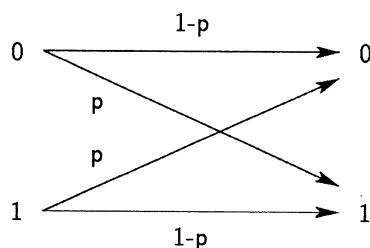


Figura 8.4: Rappresentazione di un canale di comunicazione di tipo binario simmetrico.

◆ **Esercizio 8.1** Una ruota della roulette ha 37 posizioni numerate  $0, 1, \dots, 36$ . Supponendo che la pallina abbia la stessa probabilità di fermarsi nelle varie posizioni, calcolare la probabilità di ottenere

- un numero pari,
- un numero maggiore di 30,
- un numero minore o uguale a 10.

◆ **Esercizio 8.2** Supponendo che in una popolazione i gruppi sanguigni  $A$  (antigene  $A$  presente),  $B$  (antigene  $B$  presente),  $AB$  (entrambi gli antigeni presenti),  $0$  (nessun antigene presente) siano presenti con le seguenti probabilità

$$P(A) = 0.35, \quad P(B) = 0.42, \quad P(AB) = 0.18, \quad P(0) = 0.05$$

trovare la probabilità che un individuo scelto a caso abbia a) l'antigene  $A$ , b) l'antigene  $B$ , c) nessun antigene.

◆ **Esercizio 8.3** In una popolazione umana 35 persone hanno il sangue del gruppo  $A$ , 47 del gruppo  $B$ , 21 del gruppo  $AB$  e 4 del gruppo  $0$ . Trovare la probabilità che un individuo scelto a caso abbia il sangue del gruppo  $AB$ .

◆ **Esercizio 8.4** Intorno ad un tavolo rotondo si dispongono a caso  $n$  uomini e  $n$  donne. Calcolare la probabilità che ogni donna si trovi seduta tra due uomini.

◆ **Esercizio 8.5** Sia  $E_n$ ,  $n \geq 1$  una successione arbitraria di eventi. Dimostrare la seguente proprietà, chiamata sub- $\sigma$ -additività

$$P(\cup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} P(E_n)$$

◆ **Esercizio 8.6** Siano  $E_1$  e  $E_2$  due eventi incompatibili, con  $P(E_1) = 0.5$  e  $P(E_1 \cap E_2) = 0.6$ . Calcolare  $P(E_2)$ .

◆ **Esercizio 8.7** Sapendo che un motore dotato di sei candele ne ha due difettose, calcolare la probabilità che, levando a caso due candele, entrambe siano difettose.

◆ **Esercizio 8.8** Sapendo che  $1/3$  di un dato insieme di prodotti è difettoso, calcolare la probabilità che, prelevando tre oggetti a caso

- esattamente uno di essi sia difettoso;
- almeno uno di essi sia difettoso.

◆ **Esercizio 8.9** Calcolare la probabilità che su un gruppo di 25 individui si abbiano 25 compleanni diversi. (Si assuma un anno di 365 giorni e che tutti i giorni siano ugualmente probabili.)

◆ **Esercizio 8.10** Si consideri il circuito elettrico illustrato in Figura 8.5, nel quale  $I_1$ ,  $I_2$ ,  $I_3$  sono interruttori che possono operare indipendentemente,  $B$  è una sorgente (batteria) e  $H$  un apparecchio per riscaldamento. Supposto che le probabilità che gli interruttori  $I_r$ ,  $r = 1, 2, 3$  siano chiusi siano rispettivamente 0.9, 0.8, 0.85, calcolare la probabilità che  $H$  sia attivo.

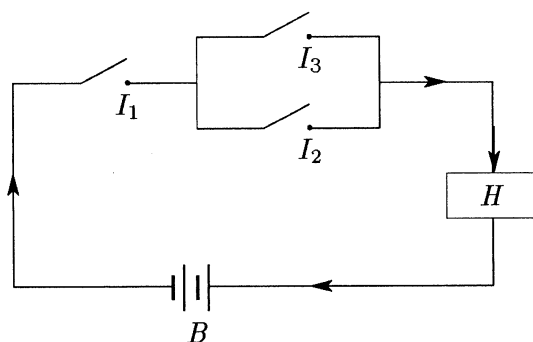


Figura 8.5: Schema di un circuito elettrico.

◆ **Esercizio 8.11** Un impianto nucleare ha un sistema protettivo costituito da 8 differenti congegni che operano in maniera indipendente. Un incidente può succedere solo se almeno 7 di tali congegni sono fuori uso. Supposto 0.1 la probabilità che ognuno dei singoli congegni sia fuori uso, calcolare la probabilità che si abbia un incidente e la probabilità che almeno 5 dei congegni sia fuori d'uso.

◆ **Esercizio 8.12** Si consideri l'esperimento del lancio di due dadi simmetrici. Si indichi con  $E_1$  l'evento corrispondente ad un totale dispari,  $E_2$  l'evento corrispondente all'uscita di 1 sul primo dado, e infine  $E_3$  l'evento corrispondente ad un totale uguale a 7. Esaminare se risultano indipendenti le seguenti coppie di eventi

- a)  $E_1$  e  $E_2$ ;      b)  $E_1$  e  $E_3$ ;      c)  $E_2$  e  $E_3$

◆ **Esercizio 8.13** Supponiamo che la probabilità di ottenere testa nel lancio di una moneta non simmetrica sia  $p$ . Si consideri, quindi, il seguente gioco. Tre giocatori, denotati rispettivamente con (I), (II), (III), lanciano la moneta successivamente, a cominciare da (I). Vince chi ottiene testa per primo. Calcolare la probabilità di vincere per ognuno dei giocatori.

◆ **Esercizio 8.14** L'urna (I) contiene due palline bianche e due nere; l'urna (II) contiene tre palline bianche e due nere. Si trasferisce una pallina da (I) a (II), e quindi si estrae da (I) una pallina che risulta essere bianca. Calcolare la probabilità che sia bianca la pallina trasferita.

◆ **Esercizio 8.15** Dato  $P(E_1) = 0.5$  e  $P(E_1 \cap E_2) = 0.6$ , trovare  $P(E_2)$  quando

- a)  $E_1$  e  $E_2$  sono incompatibili;  
b)  $E_1$  e  $E_2$  sono indipendenti;  
c)  $P(E_1 | E_2) = 0.4$ .

◆ **Esercizio 8.16** In un'urna vi sono quattro foglietti di carta del medesimo formato; ciascuno di essi è segnato con uno dei numeri 110, 101, 011, 000 e non vi sono due foglietti con lo stesso numero. Sia  $A_1$  l'evento che sul foglietto estratto a caso la cifra 1 appaia al primo posto,  $A_2$  che 1 appaia al secondo posto,  $A_3$  che 1 appaia al terzo posto. Esaminare se gli eventi  $A_1, A_2, A_3$  sono indipendenti. Esaminare, inoltre, se tali eventi sono a coppie indipendenti.

◆ **Esercizio 8.17** Supposto che nel lancio di tre dadi simmetrici non risultino due dadi con la stessa faccia, calcolare

- a) la probabilità che la somma dei punti sia 7;  
b) la probabilità che si ottenga un 1.

◆ **Esercizio 8.18** Due tetraedri simmetrici con le facce numerate da 1 a 4 vengono successivamente lanciati finché non si ottiene un totale uguale a 5 sulle facce rivolte verso il basso. Calcolare la probabilità che siano necessari più di due lanci.

◆ **Esercizio 8.19** Si consideri il segmento  $ACB$  di lunghezza  $a+b$ , ove  $AC = a$  e  $CB = b$ , con  $a < b$ . Si sceglie a caso un punto  $X$  dal segmento  $AC$  e un punto  $Y$  dal segmento  $BC$ . Mostrare che la probabilità che i segmenti  $AX$ ,  $XY$ , e  $YB$  formino un triangolo è data da  $a/(2b)$ . (Si ricordi che una condizione necessaria e sufficiente affinché tre segmenti siano i lati di un triangolo è che la lunghezza di ciascuno dei segmenti sia minore della somma delle lunghezze degli altri due).

## 8.2 Variabili aleatorie e funzioni di distribuzione

Una *variabile aleatoria*, o stocastica (random), è una funzione reale definita sullo spazio campione:  $S \xrightarrow{X} \mathbb{R}$ ; essa associa ad ogni possibile risultato di un esperimento un numero reale. In alcuni casi gli eventi elementari sono già numeri reali (ad esempio i numeri da 1 a 6 nel lancio di un dado, o, più in generale, quando gli eventi corrispondono alla misura di una quantità fisica) e allora sono essi stessi variabili aleatorie. In altri casi, è necessaria una opportuna codifica. Ad esempio, nel lancio di una moneta, indicando con  $C$  e  $T$  le due differenti facce,  $X(C) = a$ ,  $X(T) = b$ , con  $a, b$  numeri reali distinti, è una particolare variabile aleatoria.

Il valore  $x$  che una variabile aleatoria  $X$  assume in un dato esperimento è detto una *realizzazione* di  $X$ . Il *rango* di  $X$  è l'insieme di tutte le possibili realizzazioni di  $X$ . La probabilità dell'evento:  $X$  assume un qualunque valore nell'intervallo  $(a, b)$  è indicata con  $P(a < X < b)$ . Di conseguenza  $P(-\infty < X < +\infty) = 1$  e la probabilità  $P(X > c)$  che  $X$  assuma un valore più grande di un numero reale assegnato  $c$  verifica la relazione

$$P(X > c) = 1 - P(X \leq c)$$

Ad esempio, se  $X$  è il numero che si ottiene nel lancio di un dado simmetrico, allora

$$\begin{aligned} P(5 < X < 6) &= 0, & P(5 \leq X < 6) &= \frac{1}{6} \\ P(1 \leq X \leq 6) &= 1, & P(5 < X \leq 6) &= \frac{1}{6} \\ P(X > 1) &= 1 - P(X \leq 1) = 1 - \frac{1}{6} = \frac{5}{6} \end{aligned}$$

Una variabile aleatoria che assume solo un numero finito o un insieme numerabile di valori, come nell'esperimento del lancio di dadi, è chiamata una variabile aleatoria *discreta*.

La distribuzione di probabilità di una variabile aleatoria specifica la probabilità con la quale i valori della variabile sono realizzati. Più precisamente, si definisce *funzione di ripartizione*, o funzione di distribuzione (*cumulative distribution function*, *CDF*), la funzione  $F(x)$  definita nel modo seguente

$$\boxed{F(x) = P(X \leq x)} \tag{8.17}$$

Essa specifica la probabilità che la variabile aleatoria  $X$  assuma un valore minore o uguale a  $x$ . La funzione  $F(x)$  è definita su  $\mathbb{R}$  ed è monotona crescente da 0 a 1. In Figura 8.6 è illustrata la funzione di distribuzione corrispondente all'esperimento di un dado simmetrico. In questo caso la funzione  $F(x)$  è una *funzione a gradini*: è costante su ogni intervallo che non contiene alcun valore che può assumere la variabile aleatoria ed ha un salto nei valori che la variabile assume; l'ampiezza del salto corrisponde alla probabilità con la quale questo valore è realizzato. Nell'esempio tale valore è  $1/6$ .

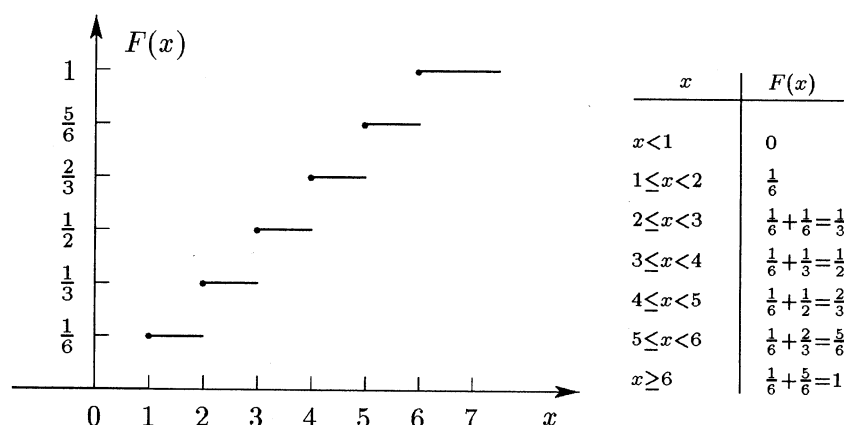


Figura 8.6: La funzione di distribuzione nell'esperimento del lancio di un dado simmetrico.

Nell'esempio del lancio di un dado si vede che la funzione di distribuzione  $F(x)$  è uguale alla somma delle probabilità  $f(x_i) = P(X = x_i) = \frac{1}{6}$  per tutti gli  $x_i \leq x$ . Il risultato è vero per ogni variabile aleatoria discreta. La funzione  $f(x)$  è detta *funzione di probabilità* o *funzione di frequenza*. Per una variabile aleatoria discreta si ha, quindi, la seguente relazione tra la funzione di distribuzione  $F(x)$  e la funzione di probabilità  $f(x)$

$$X : F(x) = \sum_{x_i \leq x} f(x_i) \quad (8.18)$$

Per le variabili aleatorie *continue*, come quelle che si ottengono dalle misure di grandezze fisiche, la funzione di distribuzione  $F(x)$  è ottenuta mediante l'integrazione di una opportuna funzione  $f(x) \geq 0$ , detta funzione di densità di probabilità (*probability density function, PDF*)

$$X : F(x) = \int_{-\infty}^x f(t) dt \quad (8.19)$$

Naturalmente, la definizione ha senso in convenienti condizioni di regolarità della

funzione  $f(x)$ . Ricordiamo, inoltre, che  $F(-\infty) = 0$  e  $F(+\infty) = 1$ , e quindi

$$\int_{-\infty}^{\infty} f(t) dt = 1$$

La probabilità dell'evento  $a \leq X \leq b$ , con  $a$  e  $b$  numeri reali distinti, è uguale all'area dell'insieme sotteso dalla curva  $y = f(x)$  per  $x \in [a, b]$

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(t) dt$$

La quantità  $f(x) dx$  è detta *elemento di probabilità di  $x$* . In maniera approssimata, si può dire che la probabilità che  $X$  cada nell'intervallo  $(x, x + dx)$  è data dal differenziale  $f(x) dx$ .

► **Esempio 8.18** In  $\mathbb{R}$  definiamo una funzione di densità di probabilità  $f(x)$  nel seguente modo

$$f(x) = \begin{cases} 0 & \text{per } x < 0 \\ \frac{1}{2}x & \text{per } 0 \leq x \leq 2 \\ 0 & \text{per } x > 2 \end{cases}$$

La funzione di distribuzione corrispondente ha la forma

$$F(x) = \begin{cases} 0 & \text{per } x < 0 \\ \frac{1}{4}x^2 & \text{per } 0 \leq x \leq 2 \\ 1 & \text{per } x > 2 \end{cases}$$

Le due funzioni sono illustrate in Figura 8.7.

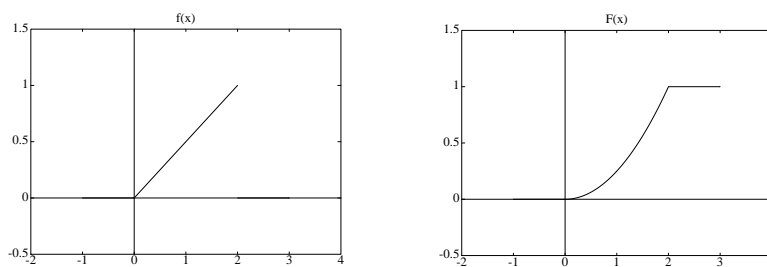


Figura 8.7: Densità di probabilità e funzione di distribuzione relative all'Esempio 8.18.

### 8.2.1 Parametri di una distribuzione

Data una variabile aleatoria  $X$ , alla funzione  $F(x)$  di distribuzione di probabilità sono associati alcuni numeri, detti *parametri* o *valori caratteristici della distribuzione*, che hanno un ruolo importante nella statistica matematica. Essi sono i momenti, le loro funzioni ed i parametri di posizione.

### Valore medio di una variabile aleatoria

Se  $X$  è una variabile aleatoria continua con densità di probabilità  $f(x)$ , si chiama *valore medio* o *speranza matematica* (*expectation*) di  $X$  il seguente numero

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x) dx \quad (8.20)$$

La definizione richiede, naturalmente, che la funzione  $xf(x)$  sia integrabile su  $\mathbb{R}$ . Interpretando intuitivamente la distribuzione di probabilità di  $X$  come distribuzione della massa unitaria sull'asse reale, il valore medio di una variabile aleatoria  $X$  può essere pensato come il *baricentro* della distribuzione di probabilità della  $X$ . Nel caso di una variabile aleatoria discreta la definizione di valore medio diventa la seguente

$$\mu = E(X) = \sum_i x_i P(X = x_i) \quad (8.21)$$

Si può dimostrare, facilmente, che se  $X_1$  e  $X_2$  sono variabili aleatorie con valori medi finiti, allora

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$$

con  $a, b$  costanti generiche.

### Varianza e deviazione standard

Se  $X$  è una variabile aleatoria continua con densità di probabilità  $f(x)$ , si chiama *varianza* di  $X$ , quando è definito, il valore

$$\sigma^2 = E[(X - \mu)^2] = \text{var}(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \quad (8.22)$$

La radice quadrata non negativa  $\sigma$  è detta *deviazione standard*, o *scarto quadratico medio*. Analoghe definizioni si hanno nel caso di variabili aleatorie discrete. La varianza, o la deviazione standard, fornisce una misura della concentrazione della distribuzione di probabilità intorno al valore medio  $\mu$ . Utilizzando la linearità dell'integrale si può esprimere la varianza nel seguente modo equivalente

$$\sigma^2 = E(X^2) - \mu^2$$

Si ha, inoltre

$$\text{var}(aX + b) = a^2 \text{var}(X)$$

ove  $a, b$  sono due costanti generiche. Ricordiamo, anche, che se  $X_1, X_2$  sono variabili aleatorie indipendenti<sup>7</sup>, allora

$$\text{var}(X_1 \pm X_2) = \text{var}(X_1) + \text{var}(X_2)$$

<sup>7</sup>Intuitivamente, le variabili aleatorie  $X_1$  e  $X_2$  sono indipendenti quando un'informazione riguardante una delle variabili non dice niente rispetto all'altra; ad esempio, può essere ragionevole

### Momenti di una variabile aleatoria

Data una variabile aleatoria  $X$  con densità di probabilità  $f(x)$ , si dice *momento di ordine  $k$* , della variabile aleatoria  $X$ , quando esiste, il valore

$$\mu'_k = E(X^k) = \int_{-\infty}^{+\infty} x^k f(x) dx \quad (8.23)$$

In maniera analoga, si definisce *momento centrale di ordine  $k$*  il seguente numero

$$\mu_k = E[(X - \mu)^k] \quad (8.24)$$

Si vede che in particolare il valore medio è  $\mu'_1$  e la varianza è  $\mu_2$ . Il momento centrale  $\mu_3$ , di ordine 3, è utilizzato, in generale, per dare una misura della obliquità (*skewness*) intorno al valore medio. Per una distribuzione simmetrica si ha  $\mu_3 = 0$  (ma non è necessariamente vero il viceversa). Una distribuzione  $f$  con  $\mu_3 > 0$  è detta essere obliqua (*skewed*) a destra, mentre una  $f_2$  con  $\mu_3 < 0$  è detta essere obliqua a sinistra. Il momento centrale  $\mu_4$  di ordine 4, chiamato usualmente *kurtosis*, fornisce indicazioni sulla forma a “picco” della distribuzione. Lo studio dei momenti di una distribuzione è importante, oltre che per l’aspetto descrittivo, per il fatto che *sotto condizioni sufficientemente generali i momenti  $\mu'_k$  ( $k = 1, 2, \dots$ ) determinano un’unica funzione di distribuzione.*

Un ulteriore risultato, collegato con il precedente, è il seguente. Sia  $Z_N$  ( $N = 1, 2, \dots$ ) una successione di variabili aleatorie con funzione di distribuzione  $F_N$ ; sia  $\mu'_{kN}$  il momento centrale  $k$ -mo di  $Z_N$ , e sia inoltre  $F$  una funzione di distribuzione con momenti centrali  $\mu'_k$ . Allora, in ipotesi sufficientemente generali, si ha  $F_N \rightarrow F$  per  $N \rightarrow \infty$ , quando  $\mu'_{kN} \rightarrow \mu'_k$ , per  $N \rightarrow \infty$  e per  $k = 1, 2, \dots$ . Questo risultato è di grande importanza nella costruzione di approssimazioni di funzioni di distribuzione.

### Disuguaglianza di Chebichev

Quando di una variabile aleatoria  $X$  sono noti il valore medio  $\mu$  e la varianza  $\sigma^2$ , per stimare il comportamento di una distribuzione di probabilità agli estremi è utile il seguente risultato, noto come *disuguaglianza di Chebichev* (o di I. J. Bienayme)

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \quad (8.25)$$

valido per ogni  $\epsilon > 0$ . In particolare, assumendo  $\epsilon = k\sigma$ , si ha  $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ . Usualmente, la limitazione (8.25) è molto più larga della vera probabilità, ma in alcuni problemi può essere ugualmente utile. Segnaliamo, come significativa applicazione, la relazione nota come *legge (debole) dei grandi numeri*.

---

supporre che l’altezza di una persona scelta a caso e il colore dei capelli siano indipendenti. Più precisamente,  $X_1$  e  $X_2$  sono variabili aleatorie indipendenti se gli eventi  $X_1 = x_1$  e  $X_2 = x_2$  sono eventi indipendenti.



### Legge dei grandi numeri

Supponiamo che una variabile aleatoria  $X$  possa assumere i valori  $x_1, x_2, \dots, x_k$  con probabilità  $p_1, p_2, \dots, p_k$ . Il valore medio è allora dato da

$$\mu = x_1 p_1 + x_2 p_2 + \dots + x_k p_k$$

Supponiamo ora di effettuare  $N$  osservazioni indipendenti della variabile aleatoria  $X$ . Se  $N_1, N_2, \dots, N_k$  è il numero di volte che si ottengono rispettivamente i valori  $x_1, x_2, \dots, x_k$ , la *media aritmetica*  $\bar{X}_N$  è data da

$$\bar{X}_N = \frac{N_1 x_1 + N_2 x_2 + \dots + N_k x_k}{N}, \quad N = N_1 + N_2 + \dots + N_k$$

Ricordando che il significato dell'affermazione "una variabile aleatoria assume il valore  $x$  con probabilità  $p_i$ " è

$$\frac{N_i}{N} \approx p_i \quad \text{per } N \rightarrow \infty$$

si ha

$$\bar{X}_N = x_1 \left( \frac{N_1}{N} \right) + x_2 \left( \frac{N_2}{N} \right) + \dots + x_k \left( \frac{N_k}{N} \right) \approx x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \mu$$

e, quindi, per  $N \rightarrow \infty$ ,  $\bar{X}_N \rightarrow \mu$ . Naturalmente, la precedente non è una dimostrazione, ma solo un suggerimento di una possibile dimostrazione.

Il Teorema di Chebichev fornisce la seguente stima della velocità di come  $\bar{X}_N$  approssima  $\mu$

$$P(|\bar{X}_N - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 N}$$

ove  $\sigma^2 = \text{Var}(X)$ . Tale relazione è detta legge dei grandi numeri (*law of large numbers*). Nel seguito, quando discuteremo il teorema limite centrale vedremo come ottenere stime migliori. Più in generale, si ha il seguente risultato.

**Proposizione 8.1** *Siano  $X_1, X_2, \dots, X_N$  un insieme di variabili aleatorie indipendenti con valori medi  $\mu_1, \mu_2, \dots, \mu_N$  e varianze  $\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2$ . Posto*

$$\bar{X}_N = \frac{X_1 + X_2 + \dots + X_N}{N}; \quad V_N = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_N^2; \quad \mu_N = \frac{\mu_1 + \mu_2 + \dots + \mu_N}{N}$$

si ha, per ogni  $\epsilon > 0$

$$P(|\bar{X}_N - \mu_N| \geq \epsilon) \leq \frac{\text{var}(\bar{X}_N)}{\epsilon^2} = \frac{V_N}{\epsilon^2 N^2}$$

Il risultato segue dal teorema di Chebichev, osservando che

$$\text{Var}(\bar{X}_N) = \frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_N^2}{N^2} = \frac{V_N}{N^2}$$

▼ **Osservazione 8.2** *Il risultato precedente stabilisce un legame tra la probabilità e la frequenza empirica e suggerisce, quindi, un modo sperimentale per stimare il valore medio di una variabile aleatoria. A questo proposito, sottolineiamo l'importanza della scelta del campionamento delle esperienze, che devono essere indipendenti.* ■

### Parametri di posizione

**Definizione 8.3** *Si dice che una variabile aleatoria  $X$  ha una distribuzione simmetrica, se esiste un punto  $a$  tale che, qualunque sia  $x$ , la funzione di distribuzione  $F(x)$  della  $X$  soddisfa la relazione seguente*

$$F(a - x) = 1 - F(a + x) + P(X = a + x)$$

Il punto  $a$  è detto centro di simmetria. In particolare, per  $a = 0$  si ha

$$F(-x) = 1 - F(x) + P(X = x)$$

Se la variabile aleatoria simmetrica è di tipo continuo, allora la corrispondente funzione di densità  $f(x)$  soddisfa alla condizione

$$f(a - x) = f(a + x)$$

tranne che negli eventuali punti di discontinuità della  $f(x)$ .

Si vede, facilmente, che se una variabile aleatoria  $X$  ha una distribuzione simmetrica ed esiste il valore medio di  $X$ , allora il valore medio coincide col centro di simmetria.

**Definizione 8.4** *Ogni valore  $x$  che soddisfa le disuguglianze*

$$P(X \leq x) \geq \frac{1}{2}, \quad P(X \geq x) \geq \frac{1}{2}$$

è detto mediana, ed è usualmente indicato con  $x_{1/2}$ .

Se la variabile  $X$  è di tipo continuo, allora mediana risulta essere ogni numero  $x$  che è soluzione della seguente equazione

$$F(x) = \frac{1}{2} \tag{8.26}$$

► **Esempio 8.19** Supponiamo che la variabile aleatoria  $X$  assuma solo i valori 0 e 1 con probabilità

$$P(X = 0) = \frac{1}{5}, \quad P(X = 1) = \frac{4}{5}$$

Allora, il punto  $x = 1$  è mediana, in quanto si ha

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 1 > \frac{1}{2}$$

$$P(X \geq 1) = P(X = 1) = \frac{4}{5} > \frac{1}{2}$$

► **Esempio 8.20** Consideriamo la variabile aleatoria  $X$  corrispondente alla seguente funzione di densità

$$f(x) = \begin{cases} 0 & \text{per } x < 0 \\ \cos x & \text{per } 0 \leq x \leq \pi/2 \\ 0 & \text{per } x > \pi/2 \end{cases}$$

La mediana si ottiene risolvendo l'equazione (8.26), e quindi

$$\int_{-\infty}^{x_{1/2}} f(x) dx = \int_0^{x_{1/2}} \cos x dx = \sin x_{1/2} = \frac{1}{2}$$

da cui  $x_{1/2} = \pi/6$ . ■

La mediana è un caso particolare di una classe di parametri detti *frattili* o *quantili*, o anche *percentili*.

**Definizione 8.5** Ogni valore  $x$  che verifica le disuguaglianze

$$P(X \leq x) \geq p, \quad P(X \geq x) \geq 1 - p \iff p \leq F(x) \leq p + P(X = x)$$

per  $0 < p < 1$ , è detto *quantile di ordine  $p$*  ed è indicato usualmente con  $x_p$ .

Se  $P(X = x) = 0$ , e quindi in particolare se la variabile aleatoria  $X$  è di tipo continuo, quantile di ordine  $p$  è ogni numero  $x$  che soddisfa l'equazione

$$F(x) = p$$

► **Esempio 8.21** Supponiamo che i tempi di sopravvivenza di pazienti affetti da cancro alla vescica in stato avanzato possano essere modellizzati dalla seguente funzione di densità

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & x \geq 0 \\ 0 & \text{altrove} \end{cases}$$

Allora  $E(X) = \lambda$ , e quindi il tempo medio di sopravvivenza è dato da  $\lambda$ . Osservando che

$$F(x) = \int_{-\infty}^x f(t) dt = \int_0^x \frac{1}{\lambda} e^{-t/\lambda} dt = 1 - e^{-x/\lambda}$$

per  $x \geq 0$ , dato  $p$ , con  $0 < p < 1$ , il percentile  $x_p$  è definito come la soluzione della seguente equazione

$$F(x_p) = 1 - e^{-x_p/\lambda} = p$$

da cui

$$x_p = \lambda \ln \left( \frac{1}{1-p} \right)$$

Si hanno in particolare i valori  $x_{0.5} = \lambda \ln 2$ ,  $x_{0.25} = \lambda \ln(4/3)$ ,  $x_{0.75} = \lambda \ln 4$ . Questo risultato si interpreta dicendo che circa il 50% dei pazienti moriranno entro il tempo  $\lambda \ln 2$  e circa il 25% dopo  $\lambda \ln(4/3)$ .

### 8.2.2 Variabili aleatorie multivariate

Ad esempio nel tiro ad un bersaglio piano il punto colpito, fissato un sistema cartesiano di riferimento, è individuato dalle sue due coordinate; è naturale, quindi, la considerazione di una coppia, ordinata, di variabili aleatorie  $X, Y$  (ascissa e ordinata del punto). In questi casi si parla di *variabile aleatoria doppia*, o bidimensionale, o anche di vettore aleatorio a due componenti (unidimensionali). Naturalmente, per ogni intero  $n > 1$ , si può definire una *variabile aleatoria  $n$ -dimensionale*, o vettore aleatorio a  $n$  componenti  $[X_1, X_2, \dots, X_n]$ . Nel seguito, tuttavia, tratteremo in particolare il caso  $n = 2$ .

**Definizione 8.6** Si dice funzione di ripartizione, o di distribuzione, congiunta della variabile aleatoria doppia  $(X, Y)$ , e si indica con  $F(x, y)$ , la funzione definita dalla relazione

$$F(x, y) = P(X \leq x \cap Y \leq y) \quad (8.27)$$

La funzione  $F(x, y)$  è definita su  $\mathbb{R}^2$  e come funzione delle due variabili, separatamente, è monotona non decrescente. Inoltre,

$$\lim_{\substack{x \rightarrow +\infty \\ y \rightarrow +\infty}} F(x, y) = 1; \quad \lim_{x \rightarrow -\infty} F(x, y) = \lim_{y \rightarrow -\infty} F(x, y) = 0$$

Come nel caso unidimensionale, la distribuzione può essere *discreta*, oppure *continua*, ossia dotata di densità. Come esempio di distribuzione discreta, si consideri la coppia di variabili aleatorie:  $X$  punto realizzato col lancio di un dado, e  $Y$  punto realizzato col lancio di un secondo dado, o col secondo lancio del medesimo dado. Le determinazioni possibili sono le trentasei coppie ordinate  $(i, j)$ , con  $i, j = 1, 2, \dots, 6$ . Supponendo il dado simmetrico e i due lanci indipendenti, la probabilità  $p_{ij}$  di ottenere la coppia  $(i, j)$  è data da  $p_{ij} = 1/36$ .

Per una distribuzione discreta si ha

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{ij}$$

ove  $x_i$  sono le possibili determinazioni della variabile aleatoria  $X$ ,  $y_j$  quelle della variabile aleatoria  $Y$  e  $p_{ij} = P(X = x_i \cap Y = y_j)$ . Naturalmente, si deve avere  $\sum_i \sum_j p_{ij} = 1$ . La  $F(x, y)$  è una funzione costanti a pezzi, con discontinuità lungo intervalli o semirette uscenti dai punti  $(x_i, y_j)$  ove sono concentrate le masse (cfr. nel caso dell'esempio precedente la Figura 8.8).

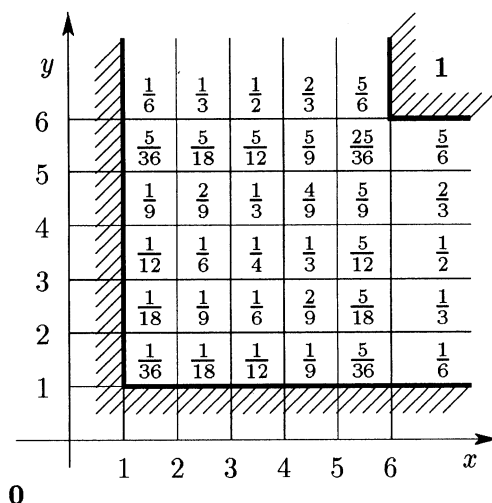


Figura 8.8: La funzione di distribuzione  $F(x, y)$  nell'esperimento del lancio di due dadi simmetrici.

Una distribuzione si dice dotata di densità se esiste una funzione  $f(x, y)$  non negativa, normalizzata, cioè tale che  $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$ , in modo che per la funzione di distribuzione  $F(x, y)$  risulti

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(\xi, \eta) d\xi d\eta$$

La  $F(x, y)$  è allora una funzione continua delle due variabili. Nei punti di continuità della densità  $f(x, y)$  si ha

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y)$$

Trascurando infinitesimi di ordine superiore al secondo rispetto alla distanza euclidea  $\sqrt{\delta x^2 + \delta y^2}$ , si ha

$$P(x < X \leq x + \delta x \cap y < Y \leq y + \delta y) = f(x, y) \delta x \delta y$$

che illustra il significato della funzione di densità.

### Distribuzioni marginali

Sia  $F(x, y)$  la funzione di distribuzione della variabile aleatoria doppia  $(X, Y)$ . Si dice *distribuzione marginale* della variabile aleatoria  $X$  la distribuzione unidimensionale la cui funzione di distribuzione  $F_1(x)$  è data dalla relazione

$$F_1(x) := \lim_{y \rightarrow +\infty} F(x, y) = P(X \leq x \cap Y < +\infty)$$

In modo analogo si definisce la funzione di distribuzione  $F_2(y)$  della distribuzione marginale della  $Y$ . Nel caso delle distribuzioni dotate di densità si ha, ad esempio

$$F_1(x) = \int_{-\infty}^x d\xi \int_{-\infty}^{+\infty} f(\xi, \eta) d\eta$$

la cui densità

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, \eta) d\eta$$

è chiamata *densità della distribuzione marginale* della  $X$ . In maniera analoga si definisce la densità  $f_2(y)$  della distribuzione marginale della  $Y$ .

Il seguente risultato fornisce un utile criterio per verificare l'indipendenza statistica di due variabili aleatorie  $X, Y$ .

**Proposizione 8.2** *Se  $(X, Y)$  è una variabile aleatoria doppia che ha  $f(x, y)$  come funzione di densità di probabilità, condizione necessaria e sufficiente affinché  $X$  e  $Y$  siano indipendenti è che*

$$f(x, y) = f_1(x) f_2(y)$$

Il risultato segue dal fatto che condizione necessaria e sufficiente affinché  $X, Y$  siano indipendenti è che si abbia

$$F(x, y) = F_1(x) F_2(y)$$

Nel caso discreto il risultato equivale alla seguente relazione

$$p_{ij} = p_{i.} p_{.j}, \quad i, j = 1, 2, \dots$$

ove  $p_{i.} = P(X = x_i \cap Y < +\infty) = P(X = x_i) = p_{1i} + p_{2i} + \dots + p_{im}$  fornisce, al variare di  $i$ , la distribuzione marginale della  $X$ ; e analogamente  $p_{.j} = P(Y = y_j)$  descrive, al variare di  $j$ , la distribuzione marginale di probabilità della  $Y$ .

► **Esempio 8.22** Consideriamo due successivi lanci di una moneta. La variabile aleatoria  $X$  assume i valori 0 oppure 1, a seconda che si ottenga rispettivamente testa o croce. Stessa definizione per la variabile  $Y$ , rispetto al secondo lancio. Come risultato di due lanci la variabile aleatoria doppia  $(X, Y)$  può assumere uno qualsiasi dei valori  $(1, 1), (1, 0), (0, 1), (0, 0)$ . Supponendo che la moneta sia simmetrica e che i lanci siano effettuati nelle medesime condizioni, la probabilità di ciascuno dei quattro eventi considerati è la stessa e vale  $1/4$ , mentre

la probabilità che  $X$  e  $Y$  assumano i valori 0 oppure 1 vale  $1/2$ . Si ha, pertanto, ad esempio la relazione

$$P(X = 1, Y = 1) = \frac{1}{4} = \frac{1}{2} \frac{1}{2} = P_1(X = 1) P_2(Y = 1)$$

Analoghi risultati si ottengono per i rimanenti eventi, per cui si conclude che  $X$  e  $Y$  sono indipendenti. Sottolineiamo, comunque, che l'indipendenza non è una nozione astratta, ma è inerente alle circostanze in cui si effettuano gli esperimenti. In realtà, l'elemento di base e fondamentale della indipendenza è l'indipendenza degli esperimenti. ■

### Distribuzioni subordinate o condizionate

Con riferimento ad una variabile aleatoria doppia  $(X, Y)$  di tipo discreto e ricordando la definizione di probabilità condizionata, si ha

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{.j}}$$

Dal momento che  $\sum_i P(X = x_i | Y = y_j) = 1$ , i valori  $p_{ij}/p_{.j}$  forniscono al variare di  $i$  una distribuzione di probabilità, detta *distribuzione della variabile  $X$  subordinata all'ipotesi  $\{Y = y_j\}$* . La corrispondente funzione di distribuzione è indicata con

$$F(x | y_j) = P(X \leq x | Y = y_j)$$

In maniera analoga, nel caso di distribuzioni dotate di densità si dimostra il seguente risultato

$$F(x | y) = \frac{\int_{-\infty}^x f(\xi, y) d\xi}{f_2(y)}$$

Nell'ipotesi che  $f_2(y)$  sia diversa dallo zero nel punto  $y$ , la  $F(x | y)$  è una funzione di distribuzione continua corrispondente alla funzione di densità

$$f(x | y) = \frac{f(x, y)}{f_2(y)}$$

Analoghe definizioni si hanno per le funzioni  $F(y | x)$  e  $f(y | x)$ .

► **Esempio 8.23** Si ha un insieme di individui, ognuno dei quali è caratterizzato da due parametri, ad esempio la statura  $X$  e il peso  $Y$ . Siano  $x_1, x_2, \dots, x_n$  le possibili determinazioni della statura e  $y_1, y_2, \dots, y_n$  quelle relative al peso. Si procede ad un sorteggio, in maniera che ognuno degli individui abbia la medesima probabilità di essere sorteggiato. Con  $p_{ij}$  si indica la probabilità che la statura  $X$  dell'individuo scelto a caso misuri  $x_i$  e che il peso  $Y$  sia  $y_j$ . La  $p_{.i}$  è la probabilità che un individuo scelto a caso abbia statura pari a  $x_i$ , mentre  $p_{ij}/p_{.j}$  è la probabilità che sia  $x_i$  la statura di un individuo il cui peso è  $y_j$ .

► **Esempio 8.24** Consideriamo il problema del tiro al bersaglio. Le variabili aleatorie  $X, Y$  corrispondono alle coordinate del punto colpito in un sistema di riferimento, la cui

origine è costituita dal bersaglio. Supponiamo che la variabile aleatoria doppia  $(X, Y)$  abbia densità di probabilità

$$f(x, y) = \begin{cases} 1/(\pi r^2) & \text{se } x^2 + y^2 \leq r^2 \\ 0 & \text{altrove} \end{cases}$$

Calcoliamo  $f_1$ , la densità marginale di  $X$ . Si ha

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_{x^2+y^2 \leq r^2} \left( \frac{1}{\pi r^2} \right) dy = \frac{1}{\pi r^2} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} dy = \frac{2\sqrt{r^2-x^2}}{\pi r^2}$$

per  $|x| \leq r$ , e zero altrove. Per simmetria si ha

$$f_2(y) = \frac{2\sqrt{r^2-y^2}}{\pi r^2}$$

per  $|y| \leq r$ , e zero altrove. Di conseguenza, per  $x \in (-r, r)$  fissato

$$f(y | x) = \frac{f(x, y)}{f_1(x)} = \begin{cases} 1/(2\sqrt{r^2-x^2}) & \text{per } |y| \leq \sqrt{r^2-x^2} \\ 0 & \text{altrove} \end{cases}$$

Se ne ricava

$$P\left(Y \geq \frac{r}{2} \mid X = \frac{r}{2}\right) = \int_{r/2}^r f\left(y \mid \frac{r}{2}\right) dy = \frac{1}{2\sqrt{3}}$$

### Valori caratteristici relativi ad una distribuzione doppia

Sia  $(X, Y)$  una variabile aleatoria doppia e sia  $F(x, y)$  la corrispondente funzione di distribuzione di probabilità. Data, poi, una funzione  $g(x, y)$ , a valori reali e sufficientemente regolare, consideriamo la variabile aleatoria  $Z = g(X, Y)$ . Ci proponiamo, ora, di calcolare i valori caratteristici, in particolare il valore medio e la varianza, di  $Z$  a partire dalla distribuzione della variabile  $(X, Y)$ . Si hanno i seguenti risultati. Nel caso in cui la distribuzione sia discreta e la serie

$$\sum_{i,j} |g(x_i, y_j)| p_{ij}$$

sia convergente, allora

$$E(Z) = E[g(X, Y)] = \sum_{i,j} g(x_i, y_j) p_{ij} \quad (8.28)$$

Analogamente, se la  $F(x, y)$  è dotata di densità ed esiste finito l'integrale

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |g(x, y)| f(x, y) dx dy$$

il valore medio di  $Z$  è dato da

$$E(Z) = E[g(X, Y)] = \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy \quad (8.29)$$



Il risultato precedente può essere verificato facilmente, considerando il caso particolare della variabile aleatoria  $Z = X + Y$ . In questo caso, si può mostrare che se la distribuzione congiunta di  $X, Y$  è dotata di densità  $f(x, y)$ , la funzione di distribuzione di  $Z$ , definita dalla seguente relazione

$$F(z) = P(X + Y \leq z)$$

ha la seguente funzione di densità di probabilità

$$f(z) = \int_{-\infty}^{+\infty} f(x, z - x) dx$$

Il valore medio  $E(Z)$  può essere, allora, anche calcolato secondo la definizione data in precedenza per una distribuzione unidimensionale, cioè

$$E(Z) = \int_{-\infty}^{+\infty} z f(z) dz = \int_{-\infty}^{+\infty} z \left[ \int_{-\infty}^{+\infty} f(x, z - x) dx \right] dz$$

Assumendo, ora  $Z = g(X, Y) := X^r Y^s$ , con  $r, s$  interi, si definisce *momento di ordine  $r + s$*  della distribuzione doppia  $(X, Y)$ , e si indica con  $\mu'_{rs}$ , la quantità  $E(X^r Y^s)$ .

In particolare, i momenti primi  $\mu'_{10}, \mu'_{01}$  forniscono i valori medi di  $X$ , e rispettivamente di  $Y$ , e sono indicati usualmente con la notazione  $\mu_x$  e  $\mu_y$ . I *momenti centrali secondi* (o varianze) sono definiti nel modo seguente

$$\mu_{20} := \text{var}(X) = E[(X - \mu_x)^2], \quad \mu_{02} = \text{var}(Y) = E[(Y - \mu_y)^2]$$

**Definizione 8.7** Viene detta *covarianza delle variabili aleatorie  $X, Y$* , e indicata usualmente con  $\text{cov}(X, Y)$ , il *momento secondo centrale misto*

$$\mu_{11} = E[(X - \mu_x)(Y - \mu_y)]$$

*nell'ipotesi che il valore medio indicato esista.*

Dalla definizione si ricavano facilmente le seguenti relazioni

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y) \quad (8.30)$$

$$\text{cov}(X, Y) = E(XY) - \mu_x \mu_y \quad (8.31)$$

**Definizione 8.8** Due variabili aleatorie  $X, Y$  si dicono *non correlate, o anche ortogonali*, quando  $\text{cov}(X, Y) = 0$ .

Dalla (8.31) si ricava che se  $X$  e  $Y$  sono indipendenti, allora,  $\text{cov}(X, Y) = 0$ . Mostriamo con un esempio che il contrario non è necessariamente vero.

► **Esempio 8.25** Sia  $W$  una variabile aleatoria con densità  $f(x) = 1$  su  $(0, 1)$  (distribuzione uniforme), e definiamo  $X = \sin 2\pi W$ ,  $Y = \cos 2\pi W$ . Le variabili  $X$  e  $Y$  non sono indipendenti, dal momento che se è noto un valore di  $X$ , allora la  $W$  ha due soli possibili valori e quindi anche  $Y$  ha soltanto due possibili valori. Si ha  $E(Y) = \int_0^1 \cos 2\pi t dt = 0$  e  $E(X) = \int_0^1 \sin 2\pi t dt = 0$ . Pertanto,  $\text{cov}(X, Y) = \int_0^1 \sin 2\pi t \cos 2\pi t dt = \frac{1}{2} \int_0^1 \sin 4\pi t dt = 0$ . In questo caso, quindi, la covarianza è nulla, ma le due variabili non sono indipendenti. ■

### Correlazione e regressione

La covarianza  $\text{cov}(X, Y)$  è, nel senso che preciseremo nel seguito, una misura di una *relazione lineare* tra  $X$  e  $Y$ . A tale scopo, tuttavia, è opportuno applicare una operazione di scalatura, definendo la seguente quantità adimensionata, detta *indice di correlazione lineare*

$$\rho_{X,Y} = \text{cov} \left( \frac{X - \mu_x}{\sigma(X)}, \frac{Y - \mu_y}{\sigma(Y)} \right) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)} \quad (8.32)$$

► **Esempio 8.26** Sia  $X$  il tempo totale che un cliente consuma presso una banca, e  $Y$  il tempo di attesa nella coda prima di raggiungere il cassiere. Supponendo che il cliente si metta immediatamente in coda dopo essere entrato nella banca, il *tempo di servizio* è dato da  $X - Y$  e il *tempo medio* di servizio è  $E(X - Y)$ . Supponiamo che la variabile aleatoria doppia  $(X, Y)$  abbia densità di probabilità

$$f(x, y) = \begin{cases} \lambda^2 e^{-\lambda x} & \text{per } 0 \leq y \leq x < \infty \\ 0 & \text{altrove} \end{cases}$$

Si ha, allora

$$E(X - Y) = \int_0^\infty \left[ \int_0^x (x - y) \lambda^2 e^{-\lambda x} dy \right] dx = \frac{1}{\lambda}$$

Tale risultato può anche essere ottenuto calcolando la densità marginale di  $X$

$$f_1(x) = \lambda^2 e^{-\lambda x} \int_0^x dy = \begin{cases} \lambda^2 x e^{-\lambda x}, & \text{per } x \geq 0 \\ 0 & \text{per } x < 0 \end{cases}$$

e quella di  $Y$

$$f_2(y) = \lambda^2 \int_y^\infty e^{-\lambda x} dx = \begin{cases} \lambda e^{-\lambda y}, & \text{per } y \geq 0 \\ 0, & \text{per } y < 0 \end{cases}$$

e quindi

$$E(X - Y) = E(X) - E(Y) = \frac{2}{\lambda} - \frac{1}{\lambda} = \frac{1}{\lambda}$$

Si ottiene inoltre

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{3}{\lambda^2} - \left(\frac{2}{\lambda}\right) \left(\frac{1}{\lambda}\right) = \frac{1}{\lambda^2} > 0$$

Il risultato indica il fatto ovvio che più grande è il tempo di attesa e più grande è il tempo totale. Il coefficiente di correlazione è dato da

$$\rho = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)} = \frac{1/\lambda^2}{\sqrt{2}/\lambda^2} = \frac{1}{\sqrt{2}} \approx 0.71$$

■

Il coefficiente di correlazione verifica il seguente risultato.

**Proposizione 8.3** *Se  $X$  e  $Y$  hanno momenti secondi finiti, allora*

$$|\rho_{X,Y}| \leq 1$$

*L'uguaglianza si ha soltanto se  $P(Y = cX) = 1$  per un qualche valore della costante  $c$ .*

In altre parole, quando  $\rho = \pm 1$ , la dipendenza stocastica tra  $X$  e  $Y$  è una dipendenza funzionale, di tipo lineare  $Y = cX$ . Un valore di  $|\rho|$  vicino a 1 indica una relazione lineare tra  $X$  e  $Y$ .

▼ **Osservazione 8.3** *È opportuno sottolineare che  $\rho$  non è una misura di come due variabili  $X$  e  $Y$  dipendono l'una dall'altra. Esso è solo una misura di come esse variano insieme, o sono associate. Esso non dovrebbe, pertanto, essere utilizzato per indicare una relazione di causa e effetto tra le due variabili. Si consideri il seguente esempio. Sia  $X$  una variabile aleatoria con densità uniforme su  $(0,1)$*

$$f(x) = 1, \quad 0 < x < 1, \text{ e zero altrove}$$

*Consideriamo, quindi,  $Y = X^k$  per  $k > 0$ . Le due variabili  $X$  e  $Y$  sono allora dipendenti. Si può mostrare facilmente che il coefficiente di correlazione è dato da*

$$\rho = \frac{\sqrt{6k+3}}{k+2}$$

*Per  $k = 2$  si ha  $\rho \approx 0.968$ , mentre per  $k \rightarrow \infty$  si ha  $\rho \rightarrow 0$ . Come si vede,  $\rho$  può essere piccolo anche se  $X$  e  $Y$  sono fortemente dipendenti, e d'altra parte può essere vicino a 1 anche se la relazione tra  $X$  e  $Y$  è non lineare. ■*

Il significato dell'indice di correlazione  $\rho$  è ulteriormente evidenziato dalle considerazioni che seguono e che riguardano la nozione di *regressione lineare* tra due variabili aleatorie. Con riferimento alla coppia aleatoria  $(X, Y)$ , si fissi una determinazione  $y$  della variabile  $Y$ , e si determini, in corrispondenza, il valore medio  $E(X | y)$  della variabile aleatoria  $X$  subordinatamente all'ipotesi  $Y = y$ . Al variare di  $y$ , il valor medio  $E(X | y)$  descrive un insieme di punti, chiamata *curva di regressione*<sup>8</sup> di  $X$  su  $Y$  e rappresentata nel piano  $x, y$  dalla curva di equazioni parametriche

$$\begin{cases} x = E(X | y) = m_1(y) \\ y = y \end{cases}$$

Nel caso discreto la curva si riduce ad un insieme di punti isolati. Se  $X$  e  $Y$  sono indipendenti, la curva di regressione di  $X$  su  $Y$  è, naturalmente la retta parallela all'asse  $y$ , di equazione  $x = \mu_x$ . In modo del tutto analogo si definisce la *curva di*

<sup>8</sup>Il termine *regressione* ha la sua origine nella seguente osservazione attribuita al genetista Sir Francis Galton (1822–1911): *Population extremes regress toward their mean.*

*regressione di Y su X*. Le curve di regressione godono della seguente proprietà, che enunciamo con riferimento alla curva di regressione di X su Y. Si ha

$$E([X - m_1(Y)]^2) = \min_{g(Y)} E([X - g(Y)]^2) \quad (8.33)$$

ossia la deviazione quadratica media di X da una funzione  $g(Y)$  della variabile aleatoria Y è minima quando  $g(Y)$  è uguale a  $m_1(Y)$ .

In generale le due curve di regressione sono distinte, e non sono necessariamente delle funzioni lineari. Quest'ultimo risultato è, tuttavia, vero, come vedremo nel seguito, per la famiglia importante delle distribuzioni normali. In ogni caso, nelle applicazioni si preferisce spesso considerare in luogo delle curve di regressione le *rette di regressione*. Più precisamente, si definisce *retta di regressione lineare di X su Y*, o *retta dei minimi quadrati*, la retta di equazione  $x = \alpha y + \beta$  per la quale risulta minima la deviazione quadratica media

$$E([X - (\alpha Y + \beta)]^2) = E(X^2) - 2\alpha E(XY) - 2\beta E(X) + \alpha^2 E(Y^2) + 2\alpha\beta E(Y) + \beta^2$$

La soluzione si ottiene annullando le derivate parziali rispetto ad  $\alpha, \beta$  e si trova

$$\alpha = \rho \frac{\sigma_x}{\sigma_y}, \quad \beta = \mu_x - \rho \frac{\sigma_x}{\sigma_y} \mu_y$$

e per la corrispondente funzione  $E([X - (\alpha Y + \beta)]^2)$  si trova il valore  $\sigma_1^2(1 - \rho^2)$ , che viene detta *varianza residua coi minimi quadrati* di X su Y. Pertanto, la retta di regressione lineare di X su Y è la retta di equazione

$$x - \mu_x = \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y)$$

In maniera analoga si trova che

$$y - \mu_y = \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

è la retta di regressione lineare di Y su X. Quando  $\rho = 0$ , (variabili non correlate) le rette di regressione sono parallele agli assi coordinati. Se  $\rho = \pm 1$ , le due rette coincidono; in effetti, come abbiamo visto, in questo caso la dipendenza stocastica tra X e Y è un legame di dipendenza lineare.

► **Esempio 8.27** Consideriamo una variabile aleatoria bidimensionale  $(X, Y)$  discreta, le cui componenti possono assumere i valori 1, 2, 3, 4, 5 con le probabilità indicate nella Tabella 8.1. Nella prima riga della tabella sono indicate le distribuzioni marginali per la X e nell'ultima colonna quelle della Y. Calcoliamo, ora, le probabilità condizionate di Y supposto che X assuma dei valori fissati. Si ha, ad esempio

$$P(Y = 1 | X = 1) = \frac{1/12}{1/3} = \frac{1}{4}$$

$y$	1/3	1/6	1/6	1/6	1/6	
5	1/24	1/24	1/24	1/24	1/30	1/5
4	1/12	0	1/24	1/24	1/30	1/5
3	1/12	1/24	1/24	0	1/30	1/5
2	1/24	1/24	1/24	1/24	1/30	1/5
1	1/12	1/24	0	1/24	1/30	1/5
	1	2	3	4	5	$x$

Tabella 8.1: Valori della probabilità relativi alla variabile aleatoria  $(X, Y)$  dell'Esempio 8.27.

$y$	1	1	1	1	1	
5	1/8	1/4	1/4	1/4	1/5	
4	1/4	0	1/4	1/4	1/5	
3	1/4	1/4	1/4	0	1/5	
2	1/8	1/4	1/4	1/4	1/5	
1	1/4	1/4	0	1/4	1/5	
	1	2	3	4	5	$x$

Tabella 8.2: Distribuzione condizionata di  $Y | X$  relativa all'Esempio 8.27.

Gli altri valori sono indicati nella Tabella 8.2. In maniera analoga si calcola la tabella relativa alla distribuzione condizionata di  $X | Y$ . Utilizzando la Tabella 8.2, si trova

$$E(Y | X = 1) = 1 \frac{1}{4} + 2 \frac{1}{8} + 3 \frac{1}{4} + 4 \frac{1}{4} + 5 \frac{1}{8} = \frac{23}{8}$$

$$E(Y | X = 2) = \frac{11}{4}, \quad E(Y | X = 3) = \frac{7}{2}$$

$$E(Y | X = 4) = 3, \quad E(Y | X = 5) = 3$$

In maniera analoga si ottengono i valori

$$E(X | Y = 1) = \frac{5}{2}, \quad E(X | Y = 2) = \frac{35}{12}$$

$$E(X | Y = 3) = \frac{55}{24}, \quad E(X | Y = 4) = \frac{65}{24}$$

$$E(X | Y = 5) = \frac{35}{12}$$

Si hanno, quindi, le seguenti curve di regressione

$$x = k, \quad y = E(Y | X = k), \quad k = 1, 2, \dots, 5 \quad (8.34)$$

$$x = E(X | Y = k), \quad y = k, \quad k = 1, 2, \dots, 5 \quad (8.35)$$

Per calcolare le rette di regressione, teniamo conto dei seguenti risultati

$$E(X) = \frac{8}{3}, \quad E(Y) = 3$$

$$\sigma_x \approx 1.49, \quad \sigma_y \approx 1.41, \quad \rho \approx 0.06$$

Si hanno allora le seguenti due rette

$$y - 3 = 0.056 \left( x - \frac{8}{3} \right)$$

$$y - 3 = 15.7 \left( x - \frac{8}{3} \right)$$

In Figura 8.9 sono rappresentate le curve di regressione e le corrispondenti rette di regressione.

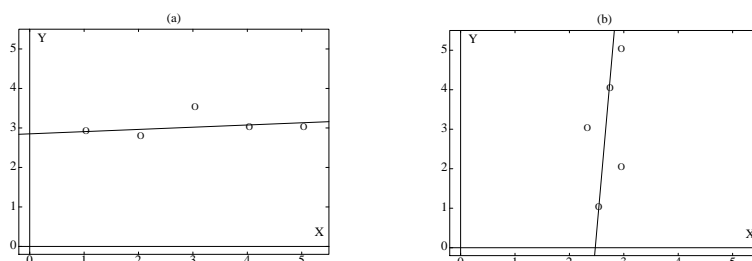


Figura 8.9: (a) Curva e retta di regressione  $E(Y | X)$ . (b) Curva e retta di regressione  $E(X | Y)$ .

### 8.2.3 Distribuzioni $n$ -dimensionali ( $n > 2$ )

Vedremo, ora, brevemente, come le nozioni introdotte in precedenza nel caso di distribuzioni bivariate possono essere estese al caso generale di  $n$  variabili aleatorie  $X_1, X_2, \dots, X_n$ , con  $n \geq 2$ . La *funzione di ripartizione congiunta* della  $n$ -pla di variabili è definita nel seguente modo

$$F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap \dots \cap X_n \leq x_n)$$

e la *distribuzione marginale delle  $k < n$  variabili*  $X_1, X_2, \dots, X_k$  è la distribuzione,  $k$ -dimensionale, la cui funzione di ripartizione è

$$F(x_1, x_2, \dots, x_k) = \lim_{x_{k+1} \rightarrow +\infty, \dots, x_n \rightarrow +\infty} F(x_1, x_2, \dots, x_k, x_{k+1}, \dots, x_n)$$

Le  $n$  variabili sono *indipendenti* quando la distribuzione (marginale) congiunta relativa ad un sottoinsieme qualunque di variabili tra le  $n$  assegnate è uguale al prodotto delle distribuzioni marginali unidimensionali delle variabili considerate. La nozione di *covarianza* si generalizza considerando l'insieme degli  $\binom{n}{2}$  valori  $\text{cov}(X_i, X_j) := E[(X_i - \mu_{x_i})(X_j - \mu_{x_j})]$ . In corrispondenza si ottiene la seguente matrice, detta

matrice di correlazione (lineare)

$$\mathbf{R} = \begin{vmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{vmatrix}$$

Se per ogni  $i \neq j$  si ha  $\rho_{ij} = 0$ , le variabili  $X_1, X_2, \dots, X_n$  sono dette *non correlate*. In questo caso il determinante  $\det(\mathbf{R})$  vale 1, mentre nel caso generale si ha  $0 \leq \det(\mathbf{R}) \leq 1$ .

Ricordiamo, infine, che la nozione di retta di regressione lineare si generalizza nella nozione di *iperpiano di regressione lineare*, o *iperpiano dei minimi quadrati*, di una variabile  $X_i$  sulle altre variabili. L'equazione di tale iperpiano è la seguente

$$x_i = \sum_{\substack{k=1 \\ k \neq i}}^n \alpha_{ik} x_k + c_i$$

I coefficienti  $\alpha_{ik}$  e  $c_i$  si ottengono risolvendo un problema di minimo, o equivalentemente, il sistema che si ottiene annullando le derivate parziali della funzione da minimizzare. Ricordiamo che per la soluzione numerica di tale sistema sono disponibili diversi metodi stabili, che si basano sulla decomposizione ortogonale della matrice, o nella sua decomposizione in valori singolari (cfr. Appendice A).

**▼ Osservazione 8.4** *Le considerazioni contenute in questo paragrafo costituiscono le basi teoriche del metodo dei minimi quadrati per la ricerca di una dipendenza funzionale tra due variabili  $X$  e  $Y$ . Osserviamo che nella trattazione rientra come caso particolare la ricerca di una dipendenza di tipo polinomiale*

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_r x^r, \quad r \in \mathbb{N}, r \geq 1$$

■

### 8.2.4 Analisi di alcune distribuzioni

In questo paragrafo passeremo in rassegna alcune distribuzioni di probabilità che hanno un interesse particolare nella statistica matematica e nelle applicazioni.

#### Distribuzione binomiale

Consideriamo il seguente esempio introduttivo. Una macchina produce dei pezzi, alcuni dei quali sono difettosi. Più precisamente, supponiamo che la probabilità di scegliere a caso dal lotto dei pezzi prodotti un pezzo buono sia  $p$ , con  $0 < p < 1$ . Inoltre, supponiamo che i risultati di successive estrazioni siano eventi indipendenti,

come è plausibile se prima di ogni estrazione si rimette nel lotto il pezzo precedentemente estratto. Cerchiamo ora la probabilità di trovare  $k$  pezzi buoni in  $n$  successive estrazioni. Per l'ipotesi di indipendenza delle singole estrazioni, la probabilità di ottenere una particolare  $n$ -pla costituita da  $k$  pezzi buoni e  $(n - k)$  pezzi difettosi è data da

$$p^k (1 - p)^{n-k}$$

D'altra parte il numero complessivo di  $n$ -ple formate da  $k$  pezzi buoni e  $(n - k)$  difettosi, comunque ordinati nella  $n$ -pla, è dato dalle combinazioni di  $n$  elementi a  $k$  a  $k$ , cioè

$$\binom{n}{k} = \frac{n!}{(n - k)! k!}$$

Quindi la probabilità di trovare  $k$  pezzi buoni su  $n$  pezzi estratti è

$$\binom{n}{k} p^k (1 - p)^{n-k}$$

L'esempio precedente si generalizza in questo modo. Si suppone di effettuare  $n$  volte una stessa prova; in ciascuna prova si può ottenere un evento  $A$ , che per convenzione indichiamo con *successo*, con probabilità  $p$ , con  $0 < p < 1$ , oppure l'evento complementare  $\bar{A}$ , indicato come *insuccesso*. Si suppone, inoltre, che i risultati degli  $n$  esperimenti siano indipendenti. Il numero  $k$  di volte in cui l'evento  $A$  si realizza nelle  $n$  prove è una variabile aleatoria  $X$ , che può assumere i valori  $k = 0, 1, \dots, n$ , con probabilità data dalla formula

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (8.36)$$

**Definizione 8.9** Una variabile aleatoria discreta  $X$  che ammette (8.36) come funzione di probabilità è detta variabile aleatoria binomiale, o di Bernoulli, di parametri  $p$  e  $n$  ed è indicata usualmente con  $X \sim \mathcal{B}(n, p)$ .

Per una variabile binomiale si ottengono facilmente i seguenti valori dei momenti

$$\mu = np, \quad \sigma^2 = np(1 - p)$$

Dalla (8.36) si ricava la seguente *funzione di distribuzione*

$$F(x) = P(X \leq x) = \sum_{k \leq x} \binom{n}{k} p^k (1 - p)^{n-k}$$

ove la sommatoria è estesa a tutti gli interi non negativi minori o uguali a  $x$ . Si può dimostrare il seguente risultato.



**Proposizione 8.4** (Proprietà additiva) *La somma di due variabili aleatorie indipendenti con distribuzioni binomiali di parametri  $(p, n_1)$  e  $(p, n_2)$  è ancora una variabile aleatoria con distribuzione binomiale di parametri  $(p, n_1 + n_2)$ .*

Per  $n = 1$  la distribuzione binomiale si riduce alla *distribuzione zero-uno*. Ricordiamo che una variabile aleatoria  $X$  ha una *distribuzione a due punti*, se esistono due punti  $x_1, x_2 \in \mathbb{R}$  tali che

$$P(X = x_1) = p, \quad P(X = x_2) = 1 - p \quad (8.37)$$

con  $0 < p < 1$ . Il caso in cui si ha  $x_1 = 1$  e  $x_2 = 0$  corrisponde alla distribuzione zero-uno. Se la variabile aleatoria  $X$  assume il valore 1 quando si verifica un evento  $E$ , e il valore 0, quando si verifica l'evento  $\bar{E}$ , allora si dice che  $X$  è l'*indicatore* di  $E$ .

Per  $n \geq 2$  la distribuzione binomiale può essere ottenuta a partire dalla distribuzione zero-uno nel seguente modo. Consideriamo le variabili aleatorie  $X_1, \dots, X_n$  indipendenti e equidistribuite con distribuzione zero-uno. La variabile aleatoria

$$S_n = X_1 + X_2 + \dots + X_n \quad (8.38)$$

può allora assumere i valori  $k = 0, 1, \dots, n$ . Più precisamente, l'evento  $\{S_n = k\}$  si verifica se e solo se  $k$  delle  $n$  variabili  $X_j$  ( $j = 1, \dots, n$ ) assumono il valore 1 e  $(n - k)$  di esse assumono il valore 0. Questo risultato può verificarsi in  $\binom{n}{k}$  modi diversi, e quindi, essendo le variabili indipendenti, la probabilità dell'evento  $\{S_n = k\}$  è data dal valore (8.36).

La variabile aleatoria  $S_n$  definita in (8.38) conta il numero  $k$  di volte che si è verificato l'evento  $A$  in  $n$  prove e prende quindi il nome di *frequenza assoluta di successo*, mentre la variabile  $Y_n = S_n/n$  è detta *frequenza relativa di successo*. La variabile aleatoria  $Y_n$  assume i valori

$$0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1$$

con probabilità  $P(Y_n = \frac{k}{n}) = P(S_n = k)$ .

### Distribuzione ipergeometrica

La distribuzione binomiale, considerata nel paragrafo precedente, rappresenta un modello descrittivo dei fenomeni casuali che possono essere ricondotti allo schema di estrazioni di palline di vario tipo da un'urna, quando la composizione dell'urna è assegnata e la pallina viene rimessa nell'urna dopo ogni estrazione. Il caso in cui, al contrario, la pallina estratta non viene rimessa nell'urna può essere descritto nel seguente modo.

Consideriamo un'urna con  $N$  palline, di cui  $w$  bianche e  $b$  nere, con  $N = w + b$ . Eseguiamo, quindi,  $n$  estrazioni successive secondo lo *schema ipergeometrico*, ossia senza rimettere nell'urna le palline successivamente estratte (senza reimpulamento). Indichiamo con  $X$  la variabile aleatoria che assume il valore  $k$ , con  $k = 0, 1, \dots, n$ , se come risultato delle  $n$  estrazioni effettuate troviamo  $k$  palline bianche.

Incominciamo, allora, ad osservare che la probabilità di estrarre di seguito, prima  $k$  palline bianche e poi  $(n - k)$  palline nere è data dal seguente valore

$$\frac{w}{N} \frac{w-1}{N-1} \cdots \frac{w-(k-1)}{N-(k-1)} \frac{b}{N-k} \frac{b-1}{N-(k+1)} \cdots \frac{b-(n-k-1)}{N-(n-1)} \quad (8.39)$$

In effetti, (8.39) rappresenta anche la probabilità di estrarre  $k$  palline bianche e  $(n - k)$  palline nere in un ordine qualunque, in quanto nel calcolo intervengono a numeratore e a denominatore sempre gli stessi fattori convenientemente permutati. Tenendo conto che  $k$  palline bianche e  $(n - k)$  palline nere possono essere estratte in  $\binom{n}{k}$  modi diversi, la probabilità dell'evento  $\{X = k\}$  è data da

$$P(X = k) = \binom{n}{k} \frac{w(w-1) \cdots (w-k+1)b(b-1) \cdots (b-n+k+1)}{N(N-1) \cdots (N-n+1)} \quad (8.40)$$

Essa ha senso se sono verificate le seguenti limitazioni

$$w - k + 1 \geq 1, \quad b - n + k + 1 \geq 1$$

cioè se  $k$  soddisfa alle disuguaglianze

$$\max(0, n - Nq) \leq k \leq \min(n, Np)$$

ove si è posto  $Np = w$ ,  $Nq = b$ . Il numero  $p$  e rispettivamente  $q = 1 - p$  indicano le probabilità di ottenere nella prima estrazione una pallina bianca o rispettivamente nera. Utilizzando tali notazioni, la (8.40) può essere scritta nel seguente modo equivalente

$$P(X = k) = \frac{\binom{Np}{k} \binom{Nq}{n-k}}{\binom{N}{n}} \quad (8.41)$$

**Definizione 8.10** Una variabile aleatoria  $X$ , la cui funzione di probabilità è data dalla (8.41), segue la distribuzione ipergeometrica.

Si può mostrare che per i momenti della distribuzione ipergeometrica si ottengono i seguenti valori

$$\mu = np, \quad \sigma^2 = \frac{N-n}{N-1} npq$$

Si vede che per  $N$  "grande" tali valori approssimano i corrispondenti valori della distribuzione binomiale.

▼ **Osservazione 8.5** Una interessante applicazione della distribuzione ipergeometrica riguarda il controllo statistico di qualità nella produzione industriale. Si supponga, più precisamente, di dover controllare un lotto di  $N$  pezzi di cui  $w$  buoni e  $b = N - w$  difettosi. Dal lotto si estraggono  $n$  pezzi a caso e si sottopongono ad un controllo di qualità, senza rimmetterli nel lotto (campionamento distruttivo). Si può, allora, calcolare mediante la (8.41) la probabilità che tra  $n$  pezzi estratti  $k$  siano buoni. Osserviamo, tuttavia, che in pratica i numeri  $w$  e  $b$ , e quindi la probabilità  $p$ , non sono noti a priori, ma rappresentano le incognite interessanti del problema. In effetti, il problema pratico è proprio quello di ottenere informazioni su  $w$ , dopo avere osservato  $X$  in alcuni campionamenti. A tale scopo, come vedremo più in dettaglio nel seguito, si stima il valore medio  $\mu$  mediante prove ripetute eseguite su opportuni campionamenti e si ricava in tal modo una stima della probabilità  $p$ .

Quello ora delineato su un esempio è un tipico problema statistico. Possiamo generalizzare e dire che un problema statistico consiste nell'ottenere informazioni riguardanti un universo sulla base di un campione. In un certo senso, la probabilità e la statistica procedono in direzioni opposte. La teoria della probabilità procede da un universo conosciuto per derivare distribuzioni relative a campioni estratti dall'universo, mentre la statistica procede dal campionamento osservato per desumere informazioni circa le caratteristiche sconosciute dell'universo; questo modo di ottenere conoscenza è indicato come inferenza statistica e sarà approfondito nel seguito. ■

### Distribuzione uniforme

Dati i numeri reali  $a, b$  si definisce la densità di probabilità nel seguente modo (cfr. Figura 8.10)

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se } x \in [a, b] \\ 0 & \text{altrove} \end{cases} \quad (8.42)$$

Una variabile aleatoria  $X$  con densità di probabilità data dall'equazione (8.42) è

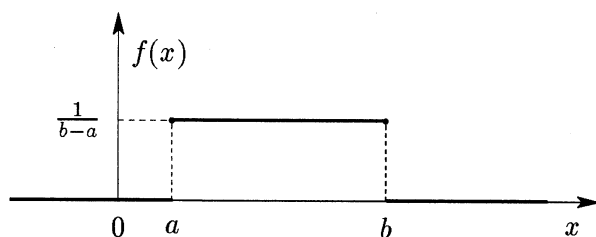


Figura 8.10: Densità di probabilità uniforme su  $[a, b]$ .

detta essere *uniformemente distribuita su  $[a, b]$*  e viene usualmente indicata con il simbolo  $X \sim \mathcal{U}([a, b])$ . Per il valore medio e la varianza si ottengono i seguenti valori

$$\mu = \frac{a+b}{2}, \quad \sigma^2 = \frac{(a-b)^2}{12}$$

A partire dalla distribuzione uniforme possono essere costruite altre distribuzioni utili nelle applicazioni e ottenute considerando la somma di due o più variabili aleatorie distribuite uniformemente. Se, ad esempio,  $X_1$  e  $X_2$  sono due variabili aleatorie distribuite uniformemente sull'intervallo  $[0, 1]$ , la variabile aleatoria  $X = X_1 + X_2$  segue una distribuzione detta *triangolare*, con determinazioni sull'intervallo  $[0, 2]$ . Se indichiamo con  $f(x)$  la densità (comune) delle variabili  $X_1$  e  $X_2$ , la densità  $\bar{f}(x)$  della somma  $X_1 + X_2$  è data da

$$\bar{f}(x) = \int_{-\infty}^{\infty} f(x-z)f(z) dz = \int_0^1 f(x-z) dz = \int_{x-1}^x f(y) dy$$

Si ricava facilmente

$$\bar{f}(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2-x & 1 \leq x \leq 2 \\ 0 & \text{altrove} \end{cases}$$

Per una applicazione importante della distribuzione uniforme si veda il Capitolo 10 relativo al metodo Monte Carlo.

### Distribuzione di Poisson

Una variabile aleatoria che assume i valori in  $\mathbb{N}$  e ammette la distribuzione

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0 \quad (8.43)$$

ove  $\lambda$  è un numero reale non negativo, è chiamata una *variabile aleatoria di Poisson* con parametro  $\lambda$  ed indicata con  $X \sim \mathcal{P}(\lambda)$  (cfr. Figura 8.11). In particolare, ricordando la convenzione  $0! = 1$ , si ha  $P(X = 0) = e^{-\lambda}$ . Si dimostrano i seguenti risultati

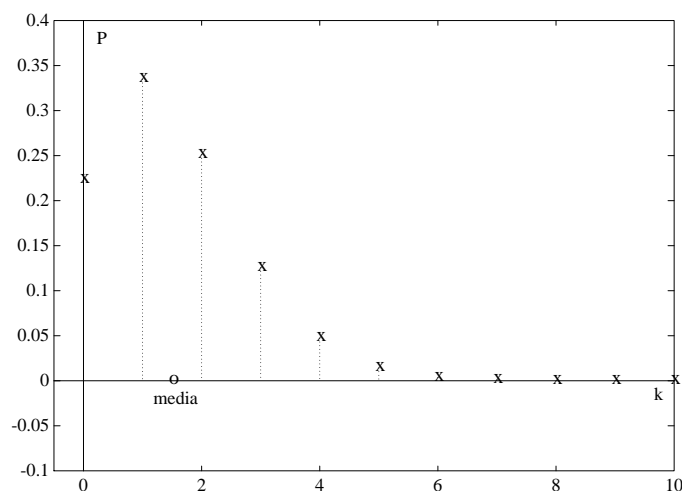
$$\mu = \lambda, \quad \sigma^2 = \lambda$$

La distribuzione di Poisson è ampiamente utilizzata nelle applicazioni. Citiamo, come esempio, la seguente applicazione. Nelle soluzioni molto diluite (ad esempio, cellule, batteri, ecc.) osservate al microscopio, la frequenza degli elementi di superficie (supposti uguali) che contengono 0, 1, 2, 3, ... organismi si distribuisce secondo la legge di Poisson se la preparazione è omogenea: in effetti la verifica di questa proprietà costituisce un *test di omogeneità*.

La funzione di probabilità (8.43) può essere ottenuta come caso limite di una successione di funzioni di probabilità della distribuzione binomiale. Più precisamente, si ha il seguente risultato.

**Teorema 8.1 (Poisson)** *Data la variabile aleatoria  $X$  con funzione di probabilità (8.36), se vale per ogni  $n$  la relazione*

$$np = \lambda$$

Figura 8.11: Distribuzione di Poisson per  $\lambda = 1.5$ .

con  $\lambda > 0$  costante, allora

$$\lim_{n \rightarrow +\infty} P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Una proprietà importante della distribuzione di Poisson è l'*additività*. Se  $X_1, X_2, \dots, X_N$  sono variabili aleatorie indipendenti con distribuzioni di Poisson, allora  $X_1 + X_2 + \dots + X_N$  ha ancora una distribuzione di Poisson con media  $m_1 + m_2 + \dots + m_N$ , se  $m_j$  è la media della variabile  $X_j$ .

Come illustrazione della distribuzione di Poisson, esaminiamo la seguente generalizzazione dello schema a prove ripetute, chiamato anche modello di Poisson a tempo continuo.

► **Esempio 8.28** *Modello di Poisson*. Consideriamo un pescatore intento a pescare sulla riva di un fiume e supponiamo che in un determinato intervallo di tempo la probabilità che arrivi un pesce dipenda solo dall'ampiezza  $\delta t$  dell'intervallo e sia proporzionale a  $\delta t$ . Più precisamente, supponiamo che tale probabilità sia espressa da

$$\lambda \delta t + o(\delta t)$$

ove  $\lambda$  è la costante di proporzionalità e  $o(\delta t)/\delta t \rightarrow 0$  per  $\delta t \rightarrow 0$ . Supponiamo, inoltre, che la probabilità di un arrivo multiplo (due o più pesci) nell'intervallo  $\delta t$  sia  $o(\delta t)$  e che l'arrivo di un pesce in un dato intervallo sia *indipendente* dagli arrivi precedenti. Il modello precedente, pur nelle semplificazioni adottate, rappresenta diverse situazioni realistiche (ad esempio, l'andamento dell'utilizzo di un particolare unità, I/O, CPU, ecc., in un calcolatore, o di un nodo particolare in una rete di calcolatori). Lasciamo al lettore la ricerca di ulteriori applicazioni. In effetti, si tratta di uno *schema delle prove ripetute a tempo continuo*.

Si vuole determinare la probabilità  $P_n(t)$  che al tempo  $t$  siano stati pescati  $n$  pesci. A tale scopo, osserviamo che si hanno  $n$  pesci al tempo  $t + \delta t$  quando

- vi sono già  $n$  pesci al tempo  $t$ , e nell'intervallo  $(t, t + \delta t)$  non ne sono arrivati;
- vi erano  $n - 1$  pesci al tempo  $t$ , e nel frattempo ne è arrivato uno;
- vi erano meno di  $n - 1$  pesci al tempo  $t$ , e si verifica un arrivo multiplo.

Per  $n > 0$  si avrà pertanto

$$P_n(t + \delta t) = P_n(t) [1 - \lambda \delta t + o(\delta t)] + P_{n-1}(t) [\lambda \delta t + o(\delta t)] + o(\delta t)$$

mentre per  $n = 0$  si ha

$$P_0(t + \delta t) = P_0(t) [1 - \lambda \delta t + o(\delta t)]$$

Dividendo per  $\delta t$ , si ottengono le relazioni

$$\begin{aligned} \frac{P_n(t + \delta t) - P_n(t)}{\delta t} &= -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(\delta t)}{\delta t} \\ \frac{P_0(t + \delta t) - P_0(t)}{\delta t} &= -\lambda P_0(t) + \frac{o(\delta t)}{\delta t} \end{aligned}$$

Facendo tendere a zero  $\delta t$ , si ottengono le seguenti equazioni differenziali ricorrenti

$$\begin{aligned} P'_n(t) &= -\lambda P_n(t) + \lambda P_{n-1}(t) \\ P'_0(t) &= -\lambda P_0(t) \end{aligned}$$

con le seguenti condizioni iniziali

$$P_0(0) = 1, \quad P_n(0) = 0, \quad \text{per } n > 0$$

La soluzione può essere ottenuta con un ragionamento di ricorrenza. Posto

$$w_n(t) = e^{\lambda t} P_n(t)$$

si ottiene

$$w'_0(t) = 0, \quad w'_n(t) = \lambda w_{n-1}(t)$$

Tenendo conto delle condizioni iniziali  $w_0(0) = 1$ ,  $w_n(0) = 0$ , si ricavano le relazioni

$$\begin{aligned} w_0 &= 1 \\ w_k(t) &= \lambda \int_0^t w_{k-1}(s) ds \end{aligned}$$

che sono verificate dalle funzioni

$$w_k(t) = \frac{(\lambda t)^k}{k!}$$

Pertanto, la probabilità  $P_n(t)$  di avere  $n$  pesci al tempo  $t$  è data da

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

Si può mostrare che la quantità  $1/\lambda$  rappresenta il valore medio del tempo che si deve attendere per la cattura del primo pesce (*tempo medio di attesa*), mentre il *numero medio* di pesci al tempo  $t$  è dato dal valore  $\lambda t$ . ■

**Densità di probabilità esponenziale**

Consideriamo come variabile aleatoria  $X$ , ad esempio, il tempo tra due successivi arrivi di automobili ad una stazione di pedaggio su una autostrada (o tra due successive chiamate ad un apparecchio telefonico pubblico, ecc.). Supponiamo che il tempo di arrivo di un'automobile alla stazione non sia influenzata dal tempo di arrivo dell'automobile precedente, ossia

$$P(X > t + s | X > t) = P(X > s)$$

Posto

$$A_1 \equiv \text{insieme } (X > t); \quad A_2 \equiv \text{insieme } (X > t + s)$$

si ha  $A_1 \cap A_2 = A_2$  e quindi

$$P(A_2 | A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{P(A_2)}{P(A_1)}$$

da cui

$$\frac{P(X > t + s)}{P(X > t)} = P(X > s)$$

Indicata con  $F$  la funzione di ripartizione di probabilità della variabile  $X$  e definita la funzione  $G = 1 - F$ , dalla relazione precedente si ha

$$G(t + s) = G(t) G(s)$$

Si può, allora, dimostrare che l'unica soluzione continua, monotona, non crescente e non identicamente nulla di tale equazione funzionale è data da

$$G(t) = e^{-\lambda t}, \quad \text{per } \lambda \geq 0$$

e, quindi

$$F(t) = 1 - e^{-\lambda t}$$

a cui corrisponde la funzione di densità

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{se } t \geq 0 \\ 0 & \text{altrove} \end{cases} \quad (8.44)$$

Le funzioni di ripartizione e di densità sono illustrate nella Figura 8.12. Una variabile aleatoria  $X$  con densità di probabilità data dall'equazione (8.44) è detta *variabile aleatoria esponenziale* ed è denotata  $X \sim \mathcal{E}(\lambda)$ . Si può mostrare che

$$\mu = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}$$

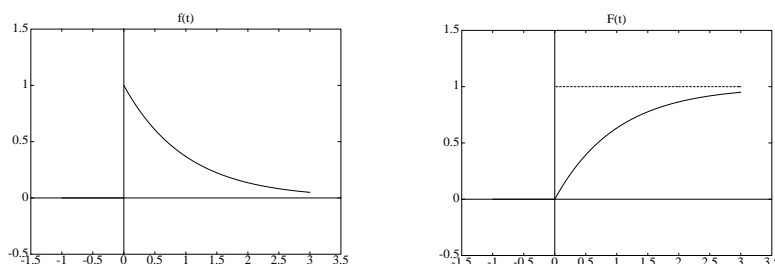


Figura 8.12: Densità di probabilità e funzione di distribuzione esponenziale per  $\lambda = 1$ .

► **Esempio 8.29** Usualmente le parti componenti di un prodotto industriale sono caratterizzate dal tempo durante il quale tali componenti sono efficienti, detto tempo di sopravvivenza (*lifetime*) della componente. Supponiamo che la lunghezza di tale tempo sia una variabile aleatoria con distribuzione esponenziale. Spesso, il fallimento di una qualsiasi componente comporta il fallimento di tutto il prodotto. In tale caso, è importante conoscere qual è il tempo di sopravvivenza minimo nell'insieme delle variabili aleatorie corrispondenti alle diverse componenti.

Siano  $X_1, X_2, \dots, X_N$  le  $N$  variabili aleatorie con distribuzioni esponenziali con parametri  $\lambda_1, \lambda_2, \dots, \lambda_N$ . Le distribuzioni di probabilità sono, allora

$$F_{X_i}(t) = P(X_i \leq t) = 1 - e^{-\lambda_i t}$$

Il problema consiste nella ricerca della distribuzione della variabile  $Z_N = \min(X_1, X_2, \dots, X_N)$ . Si ha

$$F_{Z_N}(t) = P(Z_N \leq t) = 1 - P(Z_N > t)$$

Essendo, nell'ipotesi che le variabili  $X_i$  siano indipendenti

$$P(Z_N > t) = P(X_1 > t)P(X_2 > t) \cdots P(X_N > t)$$

si ottiene

$$F_{Z_N}(t) = 1 - e^{-(\lambda_1 + \lambda_2 + \cdots + \lambda_N)t}$$

La variabile  $Z_N$  ha, pertanto, una distribuzione esponenziale. L'esempio ha messo in rilievo una proprietà importante della distribuzione esponenziale: *il minimo di un insieme di variabili aleatorie indipendenti che seguono distribuzioni esponenziali con parametri  $\lambda_1, \lambda_2, \dots, \lambda_N$  è una variabile aleatoria con distribuzione ancora esponenziale, con parametro  $\lambda_1 + \lambda_2 + \cdots + \lambda_N$* . In particolare, quando le componenti sono identiche il parametro è dato da  $N\lambda$ . ■

◆ **Esercizio 8.20** Una molecola in un gas ha una velocità  $V$  che è una variabile aleatoria con densità di probabilità  $f(v)$  data da

$$f(v) = a v e^{-v^2}, \quad v \geq 0$$

a) Trovare  $a$  in modo che  $f(v)$  sia una funzione di densità di probabilità.



b) L'energia cinetica  $E$  della particella è data dalla relazione  $E = mV^2/2$ . Trovare la funzione di distribuzione e la densità di probabilità di  $E$ .

◆ **Esercizio 8.21** L'efficienza  $X$  di un enzima per la digestione può essere descritto dalla seguente funzione di densità di probabilità

$$f(x) = \frac{1}{4}(5 - 3x^2), \quad 0 \leq x \leq 1$$

Trovare la probabilità che l'enzima abbia una efficienza maggiore del 50%. Trovare la media e la varianza dell'efficienza.

◆ **Esercizio 8.22** Mostrare che per  $\alpha > 0$ ,  $\beta > 0$ , la funzione

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{per } x \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

ove  $\Gamma$  è la funzione gamma definita<sup>9</sup> da

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

è una densità di probabilità. La corrispondente variabile casuale, indicata con  $\gamma(\alpha, \beta)$ , è chiamata variabile aleatoria gamma. Mostrare che per tale variabile si ha

$$\mu = \frac{\alpha}{\beta}, \quad \sigma^2 = \frac{\alpha}{\beta^2}$$

◆ **Esercizio 8.23** Un dado simmetrico è lanciato tre volte. Sia  $X$  la variabile aleatoria corrispondente al numero di volte che la faccia superiore assume i valori 1 o 2, e  $Y$  il numero di volte che la faccia assume i valori 4, 5 o 6.

- Trovare la distribuzione congiunta di  $(X, Y)$ .
- Trovare le due distribuzioni marginali.
- Trovare la distribuzione condizionata di  $X$  dato  $Y = y$  e quella di  $Y$  dato  $X = x$ .
- Trovare  $P(1 \leq X \leq 3 \mid Y = 1)$ .

◆ **Esercizio 8.24** La tensione superficiale  $X$  e l'acidità  $Y$  di un determinato composto chimico sono distribuiti congiuntamente con la seguente densità di probabilità

$$f(x, y) = \begin{cases} (3 - x - y)/3 & \text{per } 0 \leq x \leq 1, 0 \leq y \leq 2, \\ 0 & \text{altrove} \end{cases}$$

- Trovare le due densità marginali e le due densità condizionate.
- Trovare  $P(X \leq .92 \mid Y = 1)$ ,  $P(1 \leq X + Y \leq 2 \mid X = 0.1)$ .

<sup>9</sup>Ricordiamo che quando  $\alpha$  è un intero, mediante integrazione per parti, si ottiene  $\Gamma(\alpha) = (\alpha - 1)!$  e inoltre per  $\alpha$  qualunque si ha  $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ .

◆ **Esercizio 8.25** Un raggio di raggi  $X$  diretto ad un cristallo proteico è riflesso dagli atomi nelle molecole della proteina. L'intensità della riflessione  $R$  è registrata su una scala continua da 0 a 3 e la distribuzione di probabilità associata ad un particolare cristallo è data da

$$f(r) = \frac{2}{3} - \frac{2}{9}r, \quad 0 \leq r \leq 3$$

Trovare la media e la varianza dell'intensità della riflessione.

◆ **Esercizio 8.26** Il numero medio di impulsi ricevuti da una cellula nervosa durante l'unità di tempo è 2.5. Assumendo una distribuzione di Poisson, calcolare la probabilità che la cellula riceva meno di 3 impulsi in cinque unità di tempo.

◆ **Esercizio 8.27** Sia  $(X, Y)$  una variabile aleatoria doppia con densità di probabilità data da  $f(x, y)$ , con

$$f(x, y) = e^{-(x+y)} \quad x \geq 0, y \geq 0$$

Siano  $U = X - Y$  e  $V = X + Y$ .

- Trovare la funzione di distribuzione di  $U$  e  $V$ .
- Trovare la densità di probabilità di  $U$  e  $V$ .
- Esaminare l'indipendenza delle variabili  $X$  e  $Y$ .
- Trovare la densità di probabilità di  $U$  e  $V$ .
- Esaminare l'indipendenza delle variabili  $U$  e  $V$ .

◆ **Esercizio 8.28** Sia  $X$  una distribuzione uniforme nell'intervallo  $-1 \leq x \leq 1$ . Trovare la densità di probabilità delle seguenti variabili

$$\text{a) } X^3; \quad \text{b) } |X|; \quad \text{c) } \cos \pi X$$

◆ **Esercizio 8.29** Siano  $X$  e  $Y$  i tempi di esaurimento di due batterie e supponiamo che  $X$  e  $Y$  siano indipendenti con rispettive densità

$$\begin{aligned} f_1(x) &= e^{-x}, \quad x \geq 0, \quad \text{e zero altrove} \\ f_2(x) &= ye^{-y}, \quad y \geq 0, \quad \text{e zero altrove} \end{aligned}$$

Trovare  $P(X - Y > 0)$  e  $P(|X - Y| > 2)$ .

◆ **Esercizio 8.30** Si consideri il lancio di due tetraedri con facce numerate da 1 a 4. Supponendo ognuno dei due tetraedri simmetrico, si indichi con  $X$  il più piccolo dei due numeri usciti (corrispondenti alla faccia rivolta verso il basso) e con  $Y$  il più grande.

- Trovare la densità congiunta di  $X$  e  $Y$ .
- Trovare  $P(X \geq 2, Y \geq 2)$ .
- Trovare la media e la varianza di  $X$  e  $Y$ .
- Trovare la distribuzione condizionata di  $Y$  dato  $X$  per ognuno dei possibili valori di  $X$ .
- Trovare il coefficiente di correlazione di  $X$  e  $Y$ .

### 8.3 Densità di Gauss o normale

La somma di variabili aleatorie indipendenti è una variabile aleatoria che può approssimare una variabile aleatoria con distribuzione normale. Questo risultato, che esamineremo nel seguito, giustifica l'importanza della distribuzione normale nella statistica e nelle applicazioni. Le sue proprietà saranno, quindi, analizzate più in dettaglio. Consideriamo la funzione (illustrata in Figura 8.13)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0 \quad (8.45)$$

Una variabile aleatoria  $X$  con densità di probabilità (8.45) è detta *variabile aleatoria normale* o di Gauss<sup>10</sup> ed è denotata  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

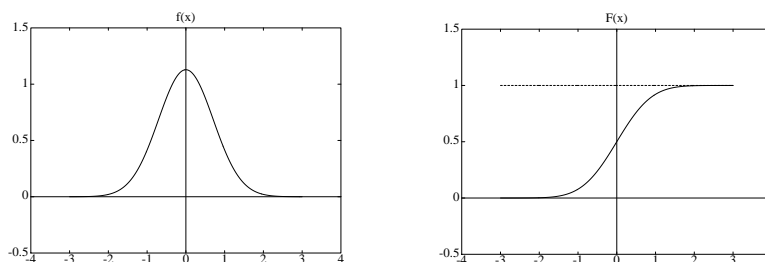


Figura 8.13: Densità di probabilità e funzione di distribuzione di Gauss.

I valori  $\mu$  e  $\sigma^2$  rappresentano esattamente il valore medio e la varianza della distribuzione. Il valore medio  $\mu$  rappresenta il centro di simmetria di  $f(x)$ . Il numero  $\sigma$  è uguale alla distanza dei *punti di flesso* dal valore medio  $\mu$ . Il *massimo* della curva è dato da  $y_{max} = 1/(\sigma\sqrt{2\pi})$  e per  $\sigma = 1$  assume il valore  $\approx 0.398942$ . L'ordinata dei punti di flesso è approssimativamente uguale a  $0.6 y_{max}$ . Dalla definizione si dimostra facilmente che per una distribuzione normale  $\mathcal{N}(0, \sigma^2)$  si hanno per i *momenti* i seguenti valori

$$\mu'_{2k+1} = 0, \quad \mu_{2k} = 1 \cdot 3 \cdot 5 \cdots (2k-1)\sigma^{2k}, \quad k = 0, 1, 2, \dots$$

La funzione di distribuzione  $F(x)$  è data da

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \quad (8.46)$$

<sup>10</sup>La distribuzione è anche nota come *legge degli errori*, in quanto essa descrive, in particolare, la distribuzione degli errori casuali relativi a successive misure di una quantità fisica. L'introduzione della legge è legata ai nomi di De Moivre (1667-1754), Laplace (1749-1827), e Gauss (1777-1855).

Si hanno, ad esempio, i seguenti valori

$$P(|X - \mu| > \sigma) \approx 0.3173$$

$$P(|X - \mu| > 2\sigma) \approx 0.0455$$

$$P(|X - \mu| > 3\sigma) \approx 0.0027$$

In altre parole, l'area dell'insieme sotteso dalla funzione di densità  $f(x)$  per  $x \in [\mu - \sigma, \mu + \sigma]$  rappresenta circa il 68.26% dell'area totale e per  $x \in [\mu - 3\sigma, \mu + 3\sigma]$  il 99.74%.

Per la distribuzione normale vale la proprietà *additiva*. Più precisamente, si ha il seguente risultato.

**Teorema 8.2** *Se  $X_1, X_2, \dots, X_N$  sono variabili aleatorie indipendenti con distribuzione  $\mathcal{N}(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, N$ , allora la variabile aleatoria*

$$Z = a_0 + a_1X_1 + a_2X_2 + \dots + a_NX_N$$

*è una variabile aleatoria normale  $\mathcal{N}(\mu_Z, \sigma_Z^2)$ , con*

$$\mu_Z = a_0 + \sum_{i=1}^N a_i\mu_i, \quad \sigma_Z^2 = \sum_{i=1}^N a_i^2\sigma_i^2$$

Quando si considera una variabile aleatoria  $X$  distribuita normalmente con media  $\mu$  e varianza  $\sigma^2$ , è usualmente più conveniente utilizzare la variabile *standardizzata*  $X^*$ , definita da

$$X^* = \frac{X - \mu}{\sigma}$$

Dal risultato precedente si ha che  $X^*$  è una variabile aleatoria con distribuzione normale  $\mathcal{N}(0, 1)$ . Il calcolo della probabilità per  $X$  può essere fatto in termini di  $X^*$ . Si ha infatti

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq X^* \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

ove, per brevità, con  $\Phi(x)$  si è indicata la funzione di distribuzione corrispondente a  $X^*$ . In questo modo è sufficiente implementare il calcolo numerico della funzione  $\Phi(x)$ . Ricordiamo, anche, che per la simmetria si ha

$$\Phi(x) = 1 - \Phi(-x)$$

Un'altra proprietà della distribuzione normale, che risulta dalle osservazioni precedenti, è che la *probabilità delle deviazioni dal valore medio dipende solo dalla loro*

grandezza come multiplo di  $\sigma$ . Più precisamente, se  $X$  segue una distribuzione normale  $\mathcal{N}(\mu, \sigma^2)$  e  $a$  è un numero generico, si ha

$$P(|X - \mu| > a) = P\left(\frac{|X - \mu|}{\sigma} > \frac{a}{\sigma}\right) = 2\Phi\left(-\frac{a}{\sigma}\right)$$

La probabilità dipende, quindi, solo da  $a/\sigma$ . Ne segue anche che i percentili di  $X$  possono essere scritti in termini dei percentili di  $X^*$ . Più precisamente, se indichiamo con  $x_c$  e  $x_c^*$  i percentili rispettivamente di  $X$  e di  $X^*$ , si ha la relazione

$$x_c = \mu + x_c^* \sigma$$

Il seguente esempio illustra l'utilizzo delle proprietà precedenti.

► **Esempio 8.30** Supponiamo che un determinato prodotto sia costituito da tre componenti. La lunghezza totale del prodotto  $Z$  sia uguale alla somma delle tre lunghezze  $X_1$ ,  $X_2$  e  $X_3$  delle sue componenti. Supponiamo, inoltre, che tali lunghezze siano, a causa della variabilità della produzione, delle variabili aleatorie indipendenti, ognuna distribuita normalmente con

$$\mu_1 = 1, \sigma_1^2 = 0.002; \quad \mu_2 = 2, \sigma_2^2 = 0.010; \quad \mu_3 = 3, \sigma_3^2 = 0.010$$

Si cerca la probabilità che  $Z$  soddisfi alla richiesta  $6.00 \pm 0.20$ , ossia, più precisamente, il valore di  $P(6.00 - 0.20 \leq Z \leq 6.00 + 0.20)$ . Per calcolare tale valore è necessario conoscere la distribuzione di probabilità della variabile aleatoria  $Z$ . Dai risultati precedenti si ha che  $Z$  ha una distribuzione normale, con

$$\mu_z = \mu_1 + \mu_2 + \mu_3; \quad \sigma_z^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$$

Sostituendo i valori assegnati, si trova  $\sigma_z^2 = 0.002 + 0.01 + 0.01$ , da cui  $\sigma_z \approx 0.148$ . Si ha, pertanto

$$\begin{aligned} P(5.80 \leq Z \leq 6.20) &= P\left(\frac{5.80 - 6.00}{0.148} \leq \frac{Z - 6.00}{0.148} \leq \frac{6.20 - 6.00}{0.148}\right) \\ &= P(-1.3 \leq Z^* \leq 1.3) \end{aligned}$$

ove  $Z^*$  è la distribuzione normale standardizzata. La probabilità cercata è allora data da  $\Phi(1.3) - \Phi(-1.3) \approx 0.8064$ . In termini pratici, questo significa che su 100 prodotti scelti a caso ci si può aspettare che 81 soddisfino alle specifiche richieste. ■

### 8.3.1 Distribuzione normale multivariata

La variabile aleatoria doppia  $(X, Y)$  si dice *normale doppia* o *bivariata* quando la distribuzione di probabilità congiunta è dotata della seguente densità

$$f(x, y) := \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]} \quad (8.47)$$

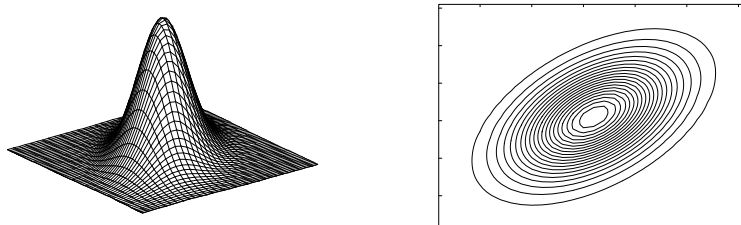


Figura 8.14: Densità di distribuzione normale bivariata per  $\sigma_1 = 0.5$ ,  $\sigma_2 = 1$ ,  $\rho = 0.5$  e corrispondenti curve di livello.

ove  $x, y \in \mathbb{R}$  e  $-1 < \rho < 1$ ,  $0 < \sigma_1$ ,  $0 < \sigma_2$ ,  $\mu_1, \mu_2 \in \mathbb{R}$ . La funzione  $f(x, y)$  è rappresentata in Figura 8.14 per valori particolari dei parametri.

Per definizione di densità, la probabilità che un punto  $(X, Y)$  appartenga ad una regione  $\Omega \subset \mathbb{R}^2$  è data dal seguente integrale

$$P[(X, Y) \in \Omega] = \int_{\Omega} f(x, y) dx dy$$

Si dimostra, facilmente, utilizzando la seguente sostituzione di variabili

$$u = \frac{x - \mu_1}{\sigma_1}, \quad v = \frac{y - \mu_2}{\sigma_2}$$

che  $P[(X, Y) \in \mathbb{R}^2] = 1$ . Si hanno, inoltre, i seguenti risultati.

**Proposizione 8.5** *Se  $(X, Y)$  ha una distribuzione normale bivariata, allora*

$$\begin{aligned} E(X) &= \mu_1, & E(Y) &= \mu_2 \\ \text{var}(X) &= \sigma_1^2, & \text{var}(Y) &= \sigma_2^2, & \text{cov}(X, Y) &= \rho \sigma_1 \sigma_2 \end{aligned}$$

*Il coefficiente di correlazione tra  $X$  e  $Y$  è, quindi, dato dal parametro  $\rho$ .*

Si può, inoltre, mostrare che le *distribuzioni marginali* di  $X$  e  $Y$ , sono distribuzioni normali con densità rispettivamente date dalle seguenti funzioni

$$f_1(x) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad f_2(y) = \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

Ricordiamo che  $X$  e  $Y$  sono variabili non correlate se e solo se  $\text{cov}(X, Y) = 0$ , oppure equivalentemente se e solo se  $\rho(X, Y) = 0$ . Si può mostrare che se  $\rho = 0$ , allora la densità congiunta  $f(x, y)$  diventa il prodotto delle due distribuzioni marginali  $f_1(x)$  e  $f_2(y)$ , il che implica la indipendenza di  $X$  e  $Y$ . Si ha, pertanto, il seguente risultato.

**Proposizione 8.6** *Se la variabile aleatoria doppia  $(X, Y)$  ha distribuzione normale bivariata,  $X$  e  $Y$  sono indipendenti se e solo se  $X$  e  $Y$  sono non correlate.*

La distribuzione di  $Y$  subordinata all'ipotesi  $\{X = x\}$  ammette come densità la funzione

$$f(y | x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{\sigma_2 \sqrt{2\pi(1 - \rho^2)}} e^{-\frac{1}{2(1-\rho^2)\sigma_2^2} \left[ y - \mu_2 - \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1) \right]^2}$$

che coincide con la densità della distribuzione normale di una variabile aleatoria unidimensionale con valore medio uguale a  $\mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)$ . Ne risulta che la *curva* di regressione di  $Y$  su  $X$  è la *retta* di regressione di  $Y$  su  $X$ . Analogamente, si mostra che la retta di regressione di  $X$  su  $Y$  ha equazione  $\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2)$  (cfr. Figura 8.15, ove viene anche mostrata la densità condizionata di  $X$  dato  $Y = y$ , per due valori particolari  $y_0$  e  $y_1$  di  $Y$ ). Inoltre, le varianze delle variabili aleatorie condizionate  $Y | X$  e  $X | Y$  coincidono con le varianze residue relative ai minimi quadrati.

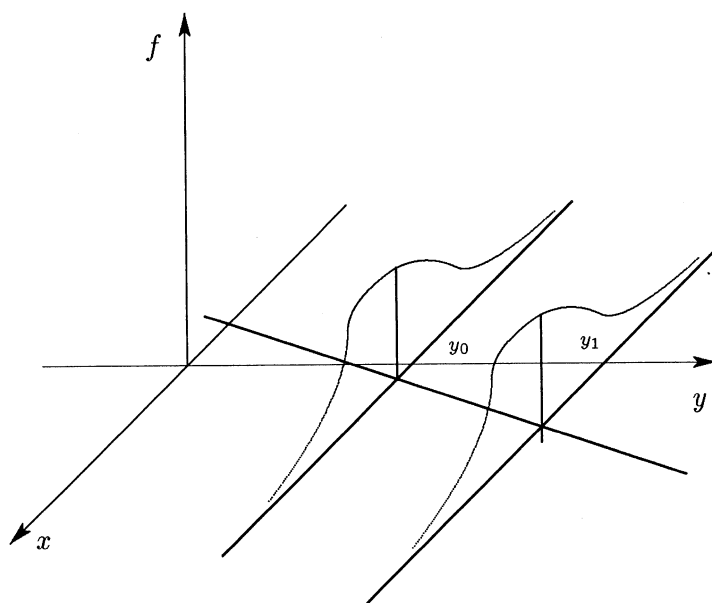


Figura 8.15: Retta di regressione  $x = \mu_1 + \rho (\sigma_1/\sigma_2) (y - \mu_2)$  e densità condizionate di  $X$  rispetto a  $Y$  per due valori particolari  $y_0$  e  $y_1$  di  $Y$ .

◆ **Esercizio 8.31** Supponiamo che la pressione sistolica per una popolazione di maschi normali sia una variabile aleatoria con media 120 mm Hg e deviazione standard  $\sigma = 20$  mm Hg. Se un individuo con pressione sistolica  $2.5\sigma$  al di sopra della media è considerato

affetto da ipertensione, calcolare la proporzione di maschi che si trovano in tale categoria. Calcolare inoltre la proporzione degli individui che hanno la pressione al di sotto di 100 mm Hg e al di sopra di 160 mm Hg.

◆ **Esercizio 8.32** Si chiama log-normale la distribuzione di una variabile aleatoria  $X$ , quando  $\ln X$  segue la distribuzione normale, e quindi la densità di  $X$  è data da

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-[\ln(x-\mu)]^2/2\sigma^2}$$

ove  $\sigma, \mu \in \mathbb{R} > 0$  con  $\sigma > 0$ . Si determinino  $E(X)$  e  $\text{var}(X)$ .

◆ **Esercizio 8.33** Supposto che la variabile doppia  $(X, Y)$  segua la distribuzione normale bivariata, con densità (8.47), si dimostri che anche la variabile doppia  $(W, Z)$ , le cui componenti sono funzioni lineari omogenee di  $X$  e  $Y$ , segue una distribuzione normale bivariata.

◆ **Esercizio 8.34** Si consideri il modello del tiro ad un bersaglio rappresentato dall'origine di un sistema di coordinate  $x y$ . Si supponga che le coordinate  $(x, y)$  del punto colpito siano realizzazioni di una coppia di variabili aleatorie normali standardizzate indipendenti. Per due proiettili sparati indipendentemente l'uno dall'altro, siano  $(X_1, Y_1)$  e  $(X_2, Y_2)$  i punti colpiti, e sia  $Z$  la distanza che li separa. Trovare la distribuzione di  $Z^2$ .

◆ **Esercizio 8.35** Se  $X$  ha una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$ , trovare la distribuzione, la media, e la varianza di  $Y = e^X$ .

◆ **Esercizio 8.36** Siano  $X_1$  e  $X_2$  due variabili aleatorie indipendenti, ognuna avente distribuzione normale con parametri  $\mu = 0$  e  $\sigma^2 = 1$ . Trovare la distribuzione congiunta di  $Y_1 = X_1^2 + X_2^2$  e di  $Y_2 = X_1/X_2$ . Trovare la distribuzione marginale di  $Y_1$  e  $Y_2$ . Esaminare se  $Y_1$  e  $Y_2$  sono distribuzioni indipendenti.

◆ **Esercizio 8.37** Siano  $X_1$  e  $X_2$  due variabili aleatorie normali standardizzate indipendenti. Sia  $U$  una variabile aleatoria indipendente da  $X_1$  e  $X_2$ , e con distribuzione uniforme su  $(0, 1)$ . Si consideri, quindi, la variabile  $Z = UX_1 + (1 - U)X_2$ . Trovare la distribuzione condizionata di  $Z$ , nell'ipotesi che  $U = u$ . Trovare, inoltre,  $E(Z)$  e  $\text{var}(Z)$ .

## 8.4 Campionamenti e distribuzioni dei campionamenti

In questo paragrafo esamineremo il concetto di *campionamento* ed introdurremo alcuni risultati relativi alle distribuzioni generate dal campionamento. Tali nozioni si basano sui risultati teorici sviluppati nei paragrafi precedenti e costituiscono il fondamento teorico delle applicazioni della statistica, in particolare, nella *teoria della stima* e nella *verifica delle ipotesi*. In maniera schematica, il problema di base della *statistica inferenziale* consiste nella determinazione delle caratteristiche (forma della distribuzione e parametri della distribuzione) di una popolazione in base ad un campione estratto da essa. Da qui l'importanza della significatività del campionamento, che può essere analizzata attraverso lo studio della distribuzione dei campioni, considerati come particolari variabili aleatorie.



### 8.4.1 Introduzione al problema

Consideriamo, ad esempio, una variabile aleatoria  $X$  con distribuzione normale di media  $\mu$  e varianza  $\sigma^2$ , e supponiamo di osservare la variabile  $X$  un numero  $n$  di volte, ottenendo i risultati  $x_1, x_2, \dots, x_n$ . Tali risultati costituiscono un particolare campione e possono essere rappresentati mediante un punto  $(x_1, x_2, \dots, x_n)$  nello spazio  $\mathbb{R}^n$ . Nell'ipotesi che le  $n$  osservazioni siano indipendenti, la *densità di probabilità congiunta*  $f_n(x_1, x_2, \dots, x_n)$  è data da

$$f_n(x_1, x_2, \dots, x_n) = \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

Il risultato può essere esteso ad una generica popolazione con densità  $f(x)$ . Un punto  $(x_1, x_2, \dots, x_n)$  viene detto *campione aleatorio* di ampiezza  $n$ , quando la distribuzione congiunta  $f_n(x_1, x_2, \dots, x_n)$ , detta anche *distribuzione campionaria*, è data da  $f(x_1) \cdots f(x_n)$ . La definizione implica che per un campione aleatorio le variabili  $X_1, X_2, \dots, X_n$ , che hanno come realizzazioni i valori  $x_1, x_2, \dots, x_n$ , siano stocasticamente indipendenti.

Supponiamo, ora, di essere interessati a stimare il valore medio  $\mu$  della densità  $f(x)$  attraverso l'osservazione del campione. Un modo ovvio consiste nell'assumere come stima di  $\mu$  la media aritmetica  $\bar{x}_n = (x_1 + x_2 + \dots + x_n)/n$ . Notiamo che  $\bar{x}_n$  (più correttamente  $\bar{X}_n$ ) è una variabile aleatoria, in quanto funzione del campione aleatorio  $(x_1, x_2, \dots, x_n)$ ; si dice che  $\bar{x}_n$  è una *variabile campionaria*, o, anche, una particolare statistica<sup>11</sup>. Si pone, quindi, il problema di introdurre una misura della bontà della stima  $\bar{x}_n$ . Una risposta può essere data in termini di probabilità, ma questo richiede lo studio della distribuzione della variabile aleatoria  $\bar{x}_n$ .

Collegato con il problema della stima di  $\mu$  vi è quello della stima della varianza  $\sigma^2$  della densità  $f(x)$ . In accordo con la definizione di  $\sigma^2$  sembrerebbe naturale usare la seguente funzione campionaria

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Ma, come ora vedremo, è più opportuna un'altra definizione. È naturale, infatti, richiedere che il valore medio di una stima sia uguale al parametro da stimare. Mentre, allora, è immediato verificare che  $E(\bar{x}_n)$  è uguale per tutti i valori di  $n$  al valore medio  $\mu$  della popolazione<sup>12</sup>, si ha invece

$$E(\hat{s}^2) = \frac{n-1}{n} \sigma^2$$

<sup>11</sup>Una *statistica* è una funzione di variabili aleatorie osservabili, e quindi a sua volta variabile aleatoria osservabile, che non contiene alcun parametro incognito.

<sup>12</sup>Si può, anche, mostrare che  $\text{var}(\bar{x}_n) = \sigma^2/n$ . In altre parole, la distribuzione di  $\bar{x}_n$  è centrata intorno a  $\mu$  e la dispersione dei valori di  $\bar{x}_n$  intorno a  $\mu$  è piccola se l'ampiezza del campione è grande. La quantità  $\sigma/\sqrt{n}$ , cioè la deviazione standard di  $\bar{x}_n$  è, usualmente, chiamata *errore standard* della media, e indicato con  $\text{SE}(\bar{x}_n)$ .

Questo significa che, anche su un campionamento ripetuto, la quantità  $\hat{s}^2$  è una sottostima di  $\sigma^2$ , apprezzabile per valori piccoli di  $n$ . Per evitare questa *distorsione* (*bias*) si utilizza come stima la seguente quantità

$$\frac{n}{n-1} \hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \equiv: s^2 \quad (8.48)$$

per la quale  $E(s^2) = \sigma^2$ . Si dice, allora, che  $s^2$ , chiamata anche *varianza campionaria*, è una stima *non distorta* (unbiased) della varianza  $\sigma^2$  della popolazione. Ancora, per studiare l'adeguatezza di  $s^2$  come stima di  $\sigma^2$  è necessario esaminare la sua distribuzione.

Oltre la media e la varianza, si possono studiare altre funzioni campionarie che possono essere utili nella inferenza statistica. Nel seguito esamineremo alcune di tali funzioni in relazione, in particolare, alla distribuzione normale, che presenta notevole interesse sia teorico che pratico.

#### 8.4.2 Campionamento di distribuzioni normali

Ricordiamo che la distribuzione normale è completamente individuata dai due parametri  $\mu$  e  $\sigma^2$ . Come abbiamo visto, la *media campionaria*  $\bar{x}_n$  fornisce una stima non distorta del valor medio  $\mu$ . Per quanto riguarda la *distribuzione* della media campionaria, si ha il seguente risultato, che abbiamo in parte già evidenziato in precedenza.

**Proposizione 8.7** *La media campionaria  $\bar{x}_n$  di un campione aleatorio di ampiezza  $n$  relativo ad una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$  ha una distribuzione normale con media  $\mu$  e varianza  $\sigma^2/n$ .*

Nella derivazione della *distribuzione* della varianza campionaria  $s^2$  gioca un ruolo centrale la cosiddetta distribuzione  $\chi^2$  (chi-quadrato), che ora introdurremo.

#### Distribuzione chi-quadrato

Date  $m$  variabili  $X_1, X_2, \dots, X_m$  indipendenti e con distribuzione normale standard (ossia  $\mu = 0, \sigma^2 = 1$ ), la variabile aleatoria

$$U = X_1^2 + X_2^2 + \dots + X_m^2 \quad (8.49)$$

viene chiamata *variabile chi-quadrato*, e denotata con  $\chi^2$ , con  $m$  gradi di libertà<sup>13</sup>. Si può dimostrare, attraverso la considerazione dei momenti della distribuzione, che

<sup>13</sup>La distribuzione  $\chi^2$  è stata introdotta e analizzata in particolare da I. J. Bienayme (1858), F. R. Helmert (1876) e K. Pearson (1900).

la variabile  $\chi^2$  ha come *densità di probabilità* la funzione  $f_m(u)$ , definita nel seguente modo (cfr. Figura 8.16)

$$f_m(u) := \frac{(u/2)^{m/2-1}}{2\Gamma(m/2)} e^{-u/2} \quad 0 \leq u \leq +\infty \quad (8.50)$$

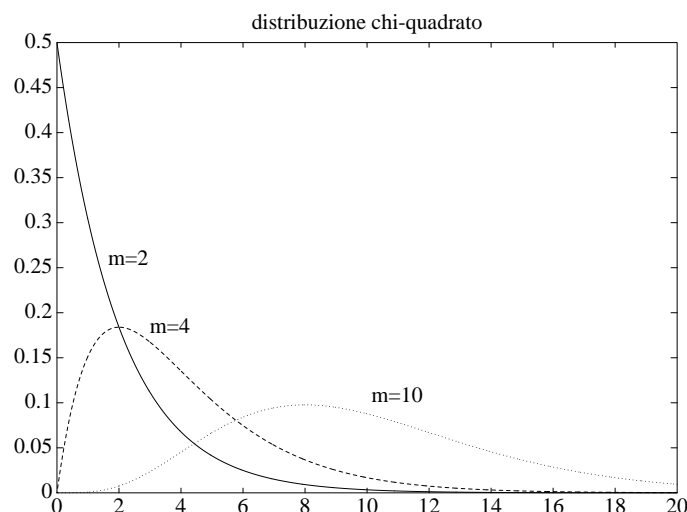


Figura 8.16: Distribuzioni  $\chi^2$  con 2, 4 e 10 gradi di libertà.

Si può, inoltre, mostrare che

$$E(\chi^2) = m, \quad \text{var}(\chi^2) = 2m$$

Una proprietà importante della distribuzione  $\chi^2$  è l'*additività*, espressa nel seguente risultato, che segue direttamente dalla definizione.

**Proposizione 8.8** *Se  $\chi_1^2, \chi_2^2, \dots, \chi_k^2$  sono variabili distribuite indipendentemente con  $m_1, m_2, \dots, m_k$  rispettivi gradi di libertà, allora la loro somma*

$$\chi_1^2 + \chi_2^2 + \dots + \chi_k^2$$

*ha la distribuzione  $\chi^2$  con  $m_1 + m_2 + \dots + m_k$  gradi di libertà.*

Esaminiamo, ora, l'utilizzo della distribuzione  $\chi^2$  nello studio della funzione campionaria  $s^2$ . Supponiamo di avere  $n$  osservazioni indipendenti relative ad una variabile aleatoria con distribuzione normale di media  $\mu$  e varianza  $\sigma^2$ . Consideriamo, allora,  $(n-1)s^2 = \sum (x_i - \bar{x})^2$ , che, come somma di quadrati, suggerisce una possibile relazione con  $\chi^2$ . In effetti, la quantità

$$\sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2$$

ha una distribuzione  $\chi^2$  con  $n$  gradi di libertà e verifica l'identità

$$\sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2 = \frac{(n-1)s^2}{\sigma^2} + \left( \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right)^2$$

Il secondo addendo nell'identità è il quadrato di una variabile normale standard e quindi ha la distribuzione  $\chi^2$  con 1 grado di libertà. Se  $\bar{x}$  e  $s^2$  fossero indipendenti, dalla proprietà additiva seguirebbe che la variabile  $(n-1)s^2/\sigma^2$  avrebbe una distribuzione  $\chi^2$  con  $n-1$  gradi di libertà. In effetti, si può dimostrare che *per campioni estratti da una distribuzione normale  $\bar{x}$  e  $s^2$  sono distribuite indipendentemente*. Si può, pertanto, concludere con il seguente risultato<sup>14</sup>.

**Proposizione 8.9** *Se  $s^2$  è la varianza campionaria di un campione aleatorio da una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$ , allora la variabile campionaria  $(n-1)s^2/\sigma^2$  ha la distribuzione  $\chi^2$  con  $n-1$  gradi di libertà.*

Il risultato espresso nella proposizione (8.9) può essere utilizzato per dare una limitazione per  $s^2/\sigma^2$ , cioè, più precisamente, mediante tale risultato è possibile calcolare la probabilità che il valore incognito  $\sigma^2$  sia contenuto in un intervallo fissato. In effetti, supposto di aver effettuato  $n$  osservazioni su una variabile normale, e indicata con  $\chi_{n-1}^2$  la distribuzione chi-quadrato con  $n-1$  gradi di libertà, per ogni probabilità fissata  $1-\alpha$  si possono trovare due numeri  $a$  e  $b$ , dipendenti da  $n$ , tali che

$$P\left(a \leq (n-1)\frac{s^2}{\sigma^2} \leq b\right) = 1 - \alpha$$

Il numero  $1-\alpha$  rappresenta, allora, la probabilità che l'intervallo

$$\left[ (n-1)\frac{s^2}{b}, (n-1)\frac{s^2}{a} \right]$$

che, sottolineiamo, è una variabile aleatoria, contenga il valore  $\sigma^2$ . Tale intervallo è anche denotato *intervallo di confidenza*<sup>15</sup>, o intervallo fiduciario, di  $\sigma^2$ . È possibile calcolare numericamente  $a$  e  $b$  in maniera che la lunghezza dell'intervallo di confidenza sia minima. Tuttavia, per semplicità, si pone usualmente

$$a = \chi_{n-1, 1-\alpha/2}^2, \quad b = \chi_{n-1, \alpha/2}^2$$

ove, in maniera generale, il valore  $\chi_{n-1, \beta}$  è definito nel seguente modo

$$P(\chi_{n-1}^2 > \chi_{n-1, \beta}^2) = \beta$$

<sup>14</sup>Osserviamo che tale risultato si applica solo a popolazioni distribuite normalmente, in quanto si può mostrare che per nessuna altra distribuzione la media campionaria e la varianza campionaria sono distribuite indipendentemente e la media campionaria ha una distribuzione esatta normale.

<sup>15</sup>La nozione di intervallo di confidenza è legata in particolare ai nomi di J. Neyman e E. S. Pearson (1895-1980).

e può essere calcolato mediante la tabella riportata in Appendice D. Si ottiene in questo modo l'intervallo di confidenza per  $\sigma^2$  di livello  $1 - \alpha$  a code (*tail*) uguali

$$\left[ (n-1) \frac{s^2}{\chi_{n-1, \alpha/2}^2}, (n-1) \frac{s^2}{\chi_{n-1, 1-\alpha/2}^2} \right] \quad (8.51)$$

► **Esempio 8.31** Supponiamo che un campione di 15 fusibili di 25 amp. mostri una deviazione standard di 0.85 amp. Se si suppone che l'amperaggio sia una variabile distribuita normalmente, allora un intervallo di confidenza per  $\sigma^2$  al 95 per cento è dato da (8.51), con

$$n = 15, s^2 = (.85)^2, \alpha = 0.05,$$

e, quindi  $\chi_{14, 0.025}^2 \approx 26.119$  e  $\chi_{14, 0.975}^2 \approx 5.629$ . L'intervallo di confidenza è pertanto dato da  $[0.39, 1.80]$ . Ne consegue, ad esempio, che una certificazione da parte del costruttore di una varianza 0.8 è, in pratica, supportata dal campionamento effettuato. ■

### Distribuzione t di Student

Sia  $X_1, X_2, \dots, X_n$  un campione aleatorio relativo ad una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$ . Indicata con  $\bar{x}_n$  la media campionaria, ci proponiamo di calcolare la probabilità che il valore  $\bar{X}_n - \mu$  sia contenuto in un intervallo fissato. Incominciamo dal caso in cui è supposto noto il valore  $\sigma^2$  della distribuzione. Fissata una probabilità  $1 - \alpha$ , possiamo cercare una costante  $a > 0$ , dipendente da  $n$  e  $\sigma^2$  tale che

$$P(-a \leq \bar{x}_n - \mu \leq a) = 1 - \alpha$$

Per ottenere  $a$  si può convertire  $\bar{x}_n$  alla sua forma standard, dividendo l'uguaglianza precedente per  $\sigma/\sqrt{n}$ , la deviazione standard della variabile  $\bar{x}_n$ . Si ottiene

$$P\left(\frac{-a\sqrt{n}}{\sigma} \leq \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{a\sqrt{n}}{\sigma}\right) = 1 - \alpha$$

Per esempio, per  $1 - \alpha = 0.99$  si ottiene

$$P\left(-2.6 \leq \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \leq 2.6\right) = 0.99$$

da cui  $a\sqrt{n}/\sigma = 2.6$ . Per un valore assegnato di  $n$ , si può calcolare il valore di  $a$ ; viceversa, dato  $a$ , si può scegliere il valore conveniente della dimensione  $n$  del campionamento. La procedura precedente può essere sistematizzata, osservando che la variabile

$$Z = \frac{\sqrt{n}(\bar{x}_n - \mu)}{\sigma}$$

ha una distribuzione normale standard (indipendente da  $\mu$  e da  $\sigma^2$ ). Allora, le limitazioni su  $Z$  possono essere espresse, per ogni valore di  $\alpha$  nella forma

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

ove  $z_{\alpha/2}$  è il percentile  $\alpha/2$  della variabile normale standard. Di conseguenza, si ottiene che vale  $1 - \alpha$  la probabilità che il seguente intervallo

$$\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right] \quad (8.52)$$

contenga il valore  $\mu$ . L'intervallo (8.52) è detto un *intervallo di confidenza* per  $\mu$ , con *coefficiente di confidenza*  $1 - \alpha$ .

▼ **Osservazione 8.6** *Sottolineiamo che  $\mu$  è una costante incognita, mentre gli estremi dell'intervallo (8.52) dipendono dai differenti campioni, sono cioè delle variabili aleatorie. In termini di frequenza, l'interpretazione del risultato precedente è la seguente: in una successione di campionamenti la proporzione di volte che gli intervalli (8.52) includono  $\mu$  è data da  $1 - \alpha$ .* ■

I risultati precedenti utilizzano essenzialmente il fatto che la variabile  $Z$  ha una distribuzione che non dipende da  $\mu$  e da  $\sigma^2$ . In effetti, come abbiamo visto, è una distribuzione normale standard. L'utilizzo di  $Z$  richiede, tuttavia, la conoscenza della quantità  $\sigma^2$ . Quando  $\sigma^2$  non è nota, è, allora, naturale la considerazione della seguente variabile

$$T := \frac{\bar{x}_n - \mu}{s/\sqrt{n}}$$

In effetti, si può dimostrare che  $T$  ha come densità di probabilità la seguente funzione

$$f_m(t) := \frac{\Gamma((m+1)/2)}{\Gamma(m/2)\sqrt{m\pi}} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2} \quad t \in \mathbb{R} \quad (8.53)$$

per  $m = n - 1$ .

Una distribuzione con densità data dalla funzione (8.53) è detta una *distribuzione t di Student* con  $m$  gradi di libertà<sup>16</sup>. La funzione  $T$  è, quindi, una distribuzione  $t$  di Student con  $n-1$  gradi di libertà. La funzione di densità  $f_m$  è illustrata in Figura 8.17 per alcuni valori di  $m$ . Come si vede, si tratta di una distribuzione simmetrica simile alla distribuzione normale standard. Si può, in effetti, dimostrare che  $f_n(t) \rightarrow (1/\sqrt{2\pi})e^{-t^2/2}$  per  $n \rightarrow \infty$ . Per valori piccoli di  $m$ , comunque, la distribuzione  $t$  di Student assegna più probabilità alle code, rispetto alla distribuzione normale standard.

Se  $F$  è una distribuzione con densità  $f_m$  data da (8.53), indichiamo con  $t_{m,\alpha}$  i seguenti valori

$$P(F > t_{m,\alpha}) = \alpha$$

<sup>16</sup>Tale distribuzione è stata introdotta da W. S. Gosset (1876–1937), impiegato presso una fabbrica di birra di Dublino, per studiare le fluttuazioni nei materiali e nella temperatura su piccoli campioni inerenti al processo di fabbricazione della birra. “Student” è lo pseudonimo sotto il quale pubblicava i suoi lavori.

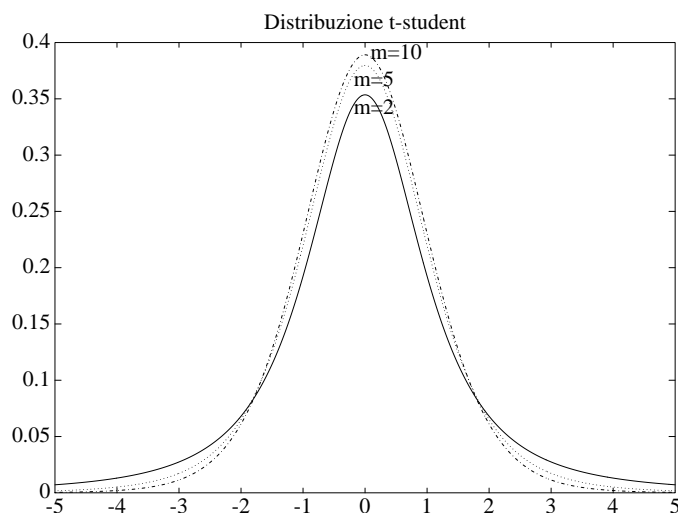


Figura 8.17: Distribuzione t di Student.

per valori fissati di  $\alpha$ ; allora  $t_{m,\alpha}$  è il quantile di ordine  $1 - \alpha$  (per il suo calcolo si veda Appendice D). Per la simmetria si ha

$$t_{m,1-\alpha} = -t_{m,\alpha}$$

Possiamo, allora, concludere che nel caso in cui  $\sigma^2$  è stimato da  $s^2$ , la probabilità che il seguente intervallo

$$\left[ \bar{x} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{x} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \right] \quad (8.54)$$

contenga il valore  $\mu$  è data da  $1 - \alpha$ . L'intervallo (8.54) è detto, quindi, *intervallo di confidenza* per  $\mu$  di livello  $1 - \alpha$ .

► **Esempio 8.32** Supponiamo che la lunghezza di un determinato prodotto sia una variabile aleatoria con distribuzione normale. Su un campione di 9 elementi si trova una lunghezza media di 3.7 e una deviazione standard di 1. Trovare l'intervallo di confidenza per la media vera della popolazione di livello 90 per cento. In questo caso, si ha  $\bar{x} = 3.7$ ,  $s = 1$  e  $n = 9$ . Poiché  $\sigma$  non è noto, utilizziamo un intervallo di confidenza basato sulla distribuzione t con  $n - 1 = 8$  gradi di libertà. Si ha, inoltre  $\alpha = 0.10$  e  $t_{8,0.05} = 1.860$ , per cui

$$\begin{aligned} \left[ \bar{x} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{x} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \right] &= \left[ 3.7 - \frac{1}{\sqrt{9}}(1.86), 3.7 + \frac{1}{\sqrt{9}}(1.86) \right] \\ &= [3.08, 4.32] \end{aligned}$$

In termini di frequenza, il risultato significa che su 100 campioni di 9 elementi, 90 dei corrispondenti intervalli di confidenza  $[\bar{x} - 1.86 s/3, \bar{x} + 1.86 s/3]$  conterranno il valore  $\mu$ . ■

► **Esempio 8.33** Consideriamo il seguente problema. Si vuole conoscere l'ampiezza  $n$  del campione tale che l'errore commesso stimando la media  $\mu$  di una variabile aleatoria  $X$  con distribuzione normale mediante la media campionaria  $\bar{x}_n$  l'errore massimo sia  $\epsilon$  con probabilità  $1 - \alpha$ . Si cerca, cioè,  $n$  in modo che

$$P(|\bar{x}_n - \mu| \leq \epsilon) = 1 - \alpha$$

Supponendo di conoscere  $\sigma$ , si ha

$$1 - \alpha = P\left(\frac{|\bar{x}_n - \mu|}{\sigma/\sqrt{n}} \leq \frac{\epsilon}{\sigma} \sqrt{n}\right) = P\left(|Z| \leq \frac{\epsilon \sqrt{n}}{\sigma}\right)$$

ove  $Z$  è la distribuzione normale standard. Ne segue

$$\epsilon \frac{\sqrt{n}}{\sigma} = z_{\alpha/2} \Rightarrow n \approx \frac{\sigma^2 z_{\alpha/2}^2}{\epsilon^2}$$

Consideriamo, ad esempio, come variabile aleatoria  $X$  il valore della pressione sistolica nella popolazione dei maschi. Supponiamo nota  $\sigma = 10$  mm Hg. Si vuole conoscere di quale ampiezza debba essere il campionamento in maniera che con probabilità 0.95 si abbia  $|\bar{x}_n - \mu| < 3$ . In questo caso si ha  $\epsilon = 3$ ,  $\alpha = 0.05$  e  $z_{\alpha/2} = z_{0.025} \approx 1.96$  e pertanto

$$n \approx \frac{10^2 1.96^2}{3^2} \approx 42.68$$

È necessario, quindi, un campionamento di ampiezza 43. ■

## Distribuzione F

I risultati del paragrafo precedente possono essere estesi al caso in cui si hanno due campioni indipendenti. Più precisamente, siano  $X_1, X_2, \dots, X_{n_1}$  e  $Y_1, Y_2, \dots, Y_{n_2}$  due campioni indipendenti tratti da due popolazioni distribuite normalmente, con parametri, rispettivamente  $\mu_1, \sigma_1$  e  $\mu_2, \sigma_2$ . Un esempio di applicazione si ha quando si vuole esaminare la variabilità di un nuovo prodotto rispetto ad un precedente. È naturale, in questo caso, confrontare le differenze o i rapporti delle stime campionarie  $s_1^2$  e  $s_2^2$ . In particolare, lo studio del rapporto è conveniente quando si è interessati principalmente al cambiamento in termini relativi, piuttosto che alla differenza assoluta nella variazione. Sappiamo che  $(n_1 - 1)s_1^2/\sigma_1^2$  e  $(n_2 - 1)s_2^2/\sigma_2^2$  hanno delle distribuzioni  $\chi^2$ , con  $n_1 - 1$  e  $n_2 - 1$  gradi di libertà. Consideriamo, allora, il seguente rapporto di variabili  $\chi^2$

$$\frac{(n_1 - 1)s_1^2/\sigma_1^2}{(n_2 - 1)s_2^2/\sigma_2^2}$$

Ricordiamo, ora, il seguente risultato.



**Proposizione 8.10** Se  $U$  è una variabile  $\chi^2$  con  $m_1$  gradi di libertà e  $W$  è una variabile  $\chi^2$  con  $m_2$  gradi di libertà, con  $U$  e  $W$  indipendenti, allora la variabile

$$\frac{U/m_1}{W/m_2} = F_{m_1, m_2}$$

ha la seguente densità di probabilità

$$f_{m_1, m_2}(t) = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} \left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}} t^{\frac{m_1}{2}-1} \left(1 + \frac{m_1}{m_2}t\right)^{-\frac{m_1+m_2}{2}} \quad (8.55)$$

per  $t > 0$  e  $f_{m_1, m_2}(t) = 0$  per  $t \leq 0$ .

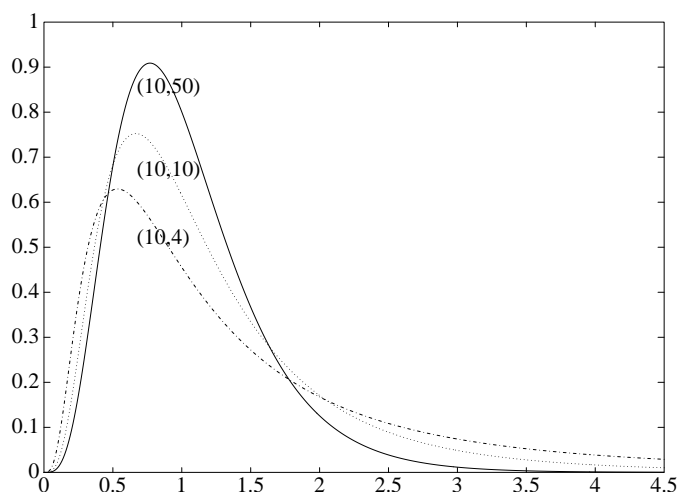


Figura 8.18: Densità di probabilità  $f_{m_1, m_2}(t)$ .

Una variabile aleatoria con densità di probabilità (8.55) è detta una *variabile aleatoria distribuita secondo una F* con  $m_1$  e  $m_2$  gradi di libertà. In Figura 8.18 è rappresentata la funzione  $f_{m_1, m_2}$  in corrispondenza ad alcune coppie di valori  $(m_1, m_2)$ . Si può dimostrare che il valore medio e la varianza sono dati da

$$E(F) = \frac{m_2}{m_2 - 2} \quad \text{per } m_2 > 2; \quad \text{var}(F) = \frac{2m_2^2(m_2 + m_1 - 2)}{m_1(m_2 - 2)^2(m_2 - 4)} \quad \text{per } m_2 > 4$$

Indicati con  $F_{m_1, m_2, \alpha}$  i percentili (cfr. Appendice D)

$$P(F_{m_1, m_2} > F_{m_1, m_2, \alpha}) = \alpha$$

si può dimostrare la seguente proprietà

$$F_{m_1, m_2, 1-\alpha} = \frac{1}{F_{m_2, m_1, \alpha}}$$

Dalla Proposizione 8.10 segue che la variabile

$$\frac{((n_1 - 1) s_1^2 / \sigma_1^2) / (n_1 - 1)}{((n_2 - 1) s_2^2 / \sigma_2^2) / (n_2 - 1)} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$

ha una distribuzione  $F$  con  $(n_1 - 1, n_2 - 1)$  gradi di libertà. Questo risultato può essere utilizzato per costruire un intervallo di confidenza per il rapporto  $\sigma_2^2 / \sigma_1^2$

$$P\left(a \leq \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} \leq b\right) = 1 - \alpha$$

Scegliendo  $a = F_{n_1-1, n_2-1, 1-\alpha/2}$  e  $b = F_{n_1-1, n_2-1, \alpha/2}$ , si ha che l'intervallo

$$\left[ F_{n_1-1, n_2-1, 1-\alpha/2} \frac{s_2^2}{s_1^2}, F_{n_1-1, n_2-1, \alpha/2} \frac{s_2^2}{s_1^2} \right]$$

è un *intervallo di confidenza di livello  $1 - \alpha$  per il rapporto  $\sigma_2^2 / \sigma_1^2$* .

Concludiamo ricordando la costruzione dell'intervallo di confidenza per  $\mu_1 - \mu_2$ . Limitiamoci al caso in cui  $\sigma_1$  e  $\sigma_2$  siano incognite, ma si sappia che  $\sigma_1 = \sigma_2$ . Nelle ipotesi di indipendenza tra le variabili  $X$  e  $Y$  si può, allora, dimostrare che posto

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

il seguente intervallo

$$\left[ \bar{x} - \bar{y} - s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2, \alpha/2}, \bar{x} - \bar{y} + s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2, \alpha/2} \right]$$

è un *intervallo di confidenza per  $\mu_1 - \mu_2$  di livello  $1 - \alpha$* .

### 8.4.3 Teorema limite centrale e applicazioni

Nei paragrafi precedenti abbiamo visto che se  $X_1, X_2, \dots, X_N$  sono variabili aleatorie normali, allora la media aritmetica  $\bar{x}_N$  è pure distribuita normalmente. Il *teorema limite centrale*, che è uno dei più importanti in statistica e nel calcolo delle probabilità, implica, in sostanza, che *sotto condizioni molto generali  $\bar{x}_N$  è distribuita normalmente, per  $N$  grande, anche se  $X_1, X_2, \dots, X_N$  non hanno una distribuzione normale*<sup>17</sup>.

**Teorema 8.3** *Sia  $\{X_N\}$  una successione di variabili aleatorie indipendenti e identicamente distribuite con comune media  $\mu$  e comune varianza  $\sigma^2$ ,  $0 < \sigma^2 < \infty$ . Se  $\bar{x}_N$  denota la media aritmetica e  $\bar{x}_N^*$  la variabile aleatoria standardizzata*

$$\bar{x}_N^* = \frac{\bar{x}_N - \mu}{\sigma / \sqrt{N}}$$

<sup>17</sup>Il risultato è legato ai nomi di De Moivre (1667-1754), Laplace (1749-1827).

allora per ogni  $z \in \mathbb{R}$

$$\lim_{N \rightarrow \infty} P(\bar{x}_N^* \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$$

L'importanza del teorema nelle applicazioni pratiche è data dal fatto che la media  $\bar{x}_N$  di un campione estratto da una qualsiasi distribuzione con varianza finita  $\sigma^2$  e media  $\mu$  è approssimativamente distribuita, per  $N$  sufficientemente grande, come una variabile aleatoria normale con media  $\mu$  e varianza  $\sigma^2/N$ . Per  $N$  grande e per  $b > a$  qualunque si ha quindi

$$P(a \leq \bar{x}_N^* \leq b) \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt = \Phi(b) - \Phi(a)$$

L'adeguatezza dell'approssimazione, come funzione di  $N$ , dipende dalla natura particolare della distribuzione. Nel seguito considereremo, in particolare, le approssimazioni della distribuzione uniforme, della distribuzione binomiale e di Poisson.

### Approssimazione normale della distribuzione uniforme

Sia  $X_1, X_2, \dots, X_N$  un campione aleatorio estratto da una distribuzione uniforme nell'intervallo  $(0, 1]$ . Si può mostrare che la densità esatta di  $\bar{x}_N$  è data da

$$f_{\bar{x}_N}(x) = \sum_{k=0}^{N-1} \frac{N}{(N-1)!} [(Nx)^{N-1} - \binom{N}{1} (Nx-1)^{N-1} + \binom{N}{2} (Nx-2)^{N-1} - \dots + (-1)^k \binom{N}{k} (Nx-k)^{N-1}] I_{[k/N, (k+1)/N]}(x)$$

ove con  $I_{[a,b]}$  si è indicata la funzione caratteristica dell'insieme  $[a, b]$ , ossia la funzione che vale 1 su  $[a, b]$  e 0 altrove. In Figura 8.19 sono rappresentate le funzioni  $f_{\bar{x}_N}$  per alcuni valori di  $N$ .

► **Esempio 8.34** Inchiamo con  $X_1, X_2, \dots, X_N$  gli errori di arrotondamento di  $N$  numeri all'intero più vicino. Possiamo ritenere gli errori  $X_i$  delle variabili aleatorie indipendenti con distribuzione uniforme su  $[-1/2, 1/2]$ . L'errore totale è dato da  $S_N = \sum_{k=1}^N X_k$ . Si vuole sapere qual è la probabilità che l'errore totale sia, ad esempio, maggiore di  $N/4$ , cioè il valore di  $P(S_N > N/4)$ . Se  $f_N$  è la densità di  $S_N$ , allora

$$P\left(S_N > \frac{N}{4}\right) = \int_{N/4}^{\infty} f_N(x) dx$$

Tale risultato può essere approssimato utilizzando il teorema limite centrale. Dal momento che ogni variabile  $X_i$  è distribuita uniformemente su  $[-1/2, 1/2]$ , si ha  $E(X_i) = 0$  e  $\sigma^2 = \text{var}(X_i) = 1/12$ . La variabile  $S_N$  è, allora, approssimata da una distribuzione normale con media 0 e varianza  $N/12$ . Ad esempio, per  $N = 1000$  si ottiene  $P(S_N > 25) = 0.0031$ . ■

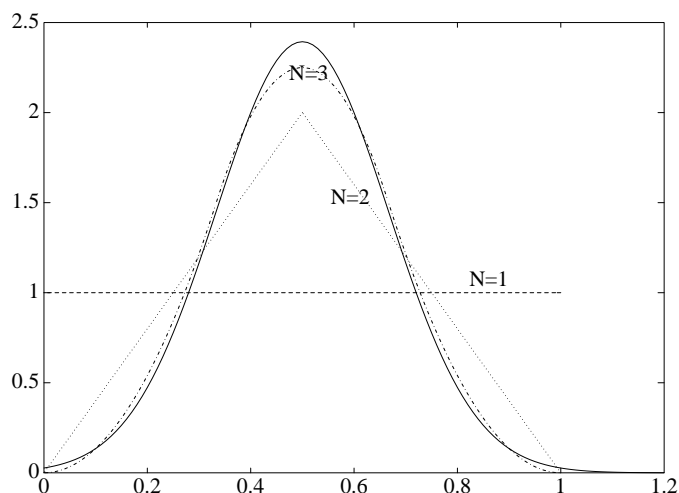


Figura 8.19: Distribuzioni della media campionaria relativa alla distribuzione uniforme. La curva continua rappresenta la distribuzione normale con  $\mu = 0.5$  e con varianza  $\sigma^2/N$ , per  $N = 3$  e ove  $\sigma = 1/\sqrt{12}$  è la varianza della distribuzione uniforme su  $(0, 1]$ .

### Approssimazione normale della distribuzione binomiale

Il numero di successi  $S_N$  in  $N$  tentativi binomiali indipendenti può essere espresso come una somma

$$S_N = X_1 + X_2 + \cdots + X_N$$

ove  $X_i$ , ( $i = 1, \dots, N$ ), sono distribuite indipendentemente e assumono i due valori 1 e 0, con probabilità  $p$  e  $q$  ( $p + q = 1$ ). Si ha  $E(S_N) = Np$  e  $\text{var}(S_N) = Npq$ . La variabile aleatoria standardizzata  $S_N^*$  è data nella forma seguente

$$S_N^* = \frac{S_N - Np}{\sqrt{Npq}}$$

Allora, dal teorema limite centrale si ricava che  $S_N^*$  è, per  $N$  grande, approssimata dalla distribuzione normale standard. In effetti, l'approssimazione è adeguata quando il valore di  $p$  non è vicino a 0 o a 1. In questi ultimi casi si ha che  $S_N$  è approssimativamente una distribuzione di Poisson.

► **Esempio 8.35** Consideriamo il seguente modello, interessante in diverse applicazioni. Si hanno  $N$  richieste di un determinato servizio, che può essere fornito attraverso un numero  $k$  di punti di servizio. Il problema riguarda la scelta ottimale del numero  $k$  dei punti di servizio. Se  $k = N$ , si ha la certezza che ogni richiesta è esaudita. In pratica, tuttavia, può essere improbabile l'accesso simultaneo a tutti i punti di servizio, e quindi alcuni punti di servizio possono rimanere inattivi. D'altra parte, se  $k$  è piccolo rispetto a  $N$ , si possono formare code. Si tratta quindi, nella programmazione del numero dei punti di servizio, di bilanciare opportunamente i due aspetti precedenti. Il problema può essere modellizzato in termini

probabilistici nel seguente modo. Si suppone che le richieste siano indipendenti e che ciascuna abbia una probabilità  $p$  di accedere al servizio. Empiricamente,  $p$  può essere interpretata come la frazione di tempo durante il quale la richiesta utilizza il servizio. Supporremo, per semplicità, che  $p$  sia la stessa per ogni richiesta. Cerchiamo, allora, la probabilità che la domanda totale  $S_N$  superi il numero dei punti di servizio. Dalla distribuzione binomiale si ha

$$P(S_N > k) = \sum_{j=k+1}^N \binom{N}{j} p^j q^{N-j}$$

Si cerca, quindi,  $k$  in modo che la precedente probabilità sia ad un livello specificato  $\alpha$ . Il valore  $1 - \alpha$  è anche chiamato *livello di servizio* (service level) e il valore necessario di  $k$  è indicato con  $k(\alpha, N)$ . Esso è definito come la soluzione della seguente equazione

$$\sum_{j=k(\alpha, N)+1}^N \binom{N}{j} p^j q^{N-j} = \alpha$$

Il calcolo di  $k(\alpha, N)$  può essere semplificato utilizzando il teorema limite centrale. Per  $N$  grande, si ha infatti

$$P(S_N > k) = P\left(\frac{S_N - Np}{\sqrt{Npq}} > \frac{k - Np}{\sqrt{Npq}}\right) = P\left(S_N^* > \frac{k - Np}{\sqrt{Npq}}\right) \approx 1 - \Phi\left(\frac{k - Np}{\sqrt{Npq}}\right)$$

Il problema è, quindi, ricondotto alla soluzione della seguente equazione

$$1 - \Phi\left(\frac{k(\alpha, N) - Np}{\sqrt{Npq}}\right) = \alpha$$

Se indichiamo con  $z_{1-\alpha}$  il percentile  $(1 - \alpha)$  della variabile normale standard, si ha

$$\frac{k(\alpha, N) - Np}{\sqrt{Npq}} \approx z_{1-\alpha} \Rightarrow k(\alpha, N) \approx Np + z_{1-\alpha} \sqrt{Npq}$$

In particolare, quando  $N$  è grande e  $p$  piccolo si ha  $k(\alpha, N) \approx Np + z_{1-\alpha} \sqrt{Np}$ , ossia  $k(\alpha, N) - Np$  cresce come la radice quadrata della domanda media. ■

### Approssimazione normale della distribuzione di Poisson

Consideriamo  $N$  variabili aleatorie di Poisson  $Y_1, Y_2, \dots, Y_N$ , ognuna con parametro  $\lambda$ . Allora, la somma  $Y_1 + Y_2 + \dots + Y_n$  è ancora una variabile di Poisson con parametro  $N\lambda$ . Pertanto una variabile di Poisson con parametro  $m$  approssima una distribuzione normale per  $m$  che aumenta. Più precisamente, la variabile standardizzata  $Y^* = (Y - m)/\sqrt{m}$  tende, per  $m \rightarrow \infty$ , alla distribuzione normale standard.

◆ **Esercizio 8.38** Consideriamo il seguente movimento aleatorio di una particella. Partendo dall'origine, ad ogni tentativo si muove di un'unità a destra o rimane nella stessa posizione con probabilità rispettivamente  $p$  e  $1 - p$ . La decisione è supposta ad ogni stato essere indipendente dalle precedenti decisioni e dalla posizione attuale. Sia  $N_r$  il numero dei passi per raggiungere una posizione  $r$  per la prima volta.

- a) Calcolare  $E(N_r)$  e  $\text{var}(N_r)$ .
- b) Utilizzando il teorema limite centrale, trovare una approssimazione della probabilità  $P(N_r > n)$ , per  $n$  fissato.

◆ **Esercizio 8.39** Siano  $X$  e  $Y$  due variabili aleatorie indipendenti con distribuzione normale con varianze rispettivamente  $\sigma_x^2$  e  $\sigma_y^2$ . Si esegue un campionamento di ampiezza  $n$  di  $X$  e di  $Y$ . Si vuole sapere come deve essere grande  $n$  in maniera che il rapporto  $(s_x^2/s_y^2)/(\sigma_x^2/\sigma_y^2)$  sia nell'intervallo  $(0.93, 1.6)$  con probabilità 0.90.

◆ **Esercizio 8.40** Consideriamo una serie di variabili aleatorie indipendenti  $Y_1, Y_2, \dots$ , ognuna con media zero e varianza  $\sigma^2$ . Si costruiscono, quindi, le variabili  $X_1, X_2, \dots$  ponendo

$$X_0 = 0, \quad X_j = \alpha X_{j-1} + Y_j, \quad j = 1, 2, \dots$$

ove  $-1 < \alpha < +1$ .

- a) Trovare  $E(X_n)$  e  $\text{var}(X_n)$ .
- b) Usando il teorema limite centrale, ottenere una espressione approssimata per il calcolo della probabilità  $P(X_n \leq t)$ , per  $n$  grande e  $t$  fissato.

◆ **Esercizio 8.41** L'arrivo dei clienti ad un negozio segue la distribuzione di Poisson con velocità di uno per ogni cinque minuti. Si cerca la probabilità che il numero degli arrivi durante un periodo di 12 ore sia tra 114 e 176.

◆ **Esercizio 8.42** Un test a risposte multiple consiste di 75 questioni, con cinque risposte per ogni questione, una delle quali è corretta. Si supponga che uno studente risponda a tutte le questioni mediante tentativi a caso. Si vuole sapere il numero atteso (valore medio) delle risposte corrette, la probabilità che le prime quattro risposte siano corrette, la probabilità che almeno 20 risposte siano corrette. Supponiamo che si voglia premiare con un punto una risposta corretta e con zero punti una risposta errata. Allo scopo di eliminare l'effetto dei tentativi a caso, si vuole sapere quanti punti dovrebbero essere tolti per una risposta non corretta, in maniera che una lunga serie di tentativi per questione abbia come media zero punti.

◆ **Esercizio 8.43** Una moneta simmetrica è lanciata fino ad ottenere 100 volte testa. Trovare la probabilità che siano necessari almeno 226 lanci e la probabilità che siano necessari esattamente 226 lanci.

◆ **Esercizio 8.44** Se per una popolazione si ha  $\sigma = 2$  e  $\bar{x}$  è la media dei campioni di ampiezza 100, trovare mediante il teorema limite centrale i limiti entro i quali sarà compreso  $\bar{x} - \mu$  con probabilità 90%.

## 8.5 Elementi di statistica inferenziale

Lo scopo di questo breve paragrafo non può essere certamente quello di analizzare in maniera esauriente uno degli aspetti più interessanti, e anche più difficili, della statistica. Più semplicemente, è parso opportuno, a completamento delle nozioni di base analizzate nei paragrafi precedenti, raccogliere alcune idee relative ai procedimenti inferenziali, rinviando alla bibliografia per un opportuno approfondimento.

### 8.5.1 Modello deterministico e modello stocastico

Per meglio inquadrare e chiarire ulteriormente il significato dell'inferenza statistica, può essere conveniente premettere alcune considerazioni generali sulla natura di un *modello stocastico*. In termini generali, un *modello* può essere definito come una idealizzazione matematica utilizzata per approssimare un fenomeno osservabile. La costruzione di un modello è basata su alcune ipotesi semplificatrici; di conseguenza, alcuni aspetti del fenomeno reale possono essere ignorati nel modello. Il successo nell'utilizzo del modello dipenderà in maniera essenziale dalla validità delle ipotesi fatte e dall'importanza degli aspetti ignorati.

Lo strumento di controllo della validità di un modello sono le osservazioni. Più precisamente, si chiama *esperimento* il processo di raccolta di osservazioni allo scopo di acquisire nuova conoscenza o di verificare ciò che si è ipotizzato.

Un *modello deterministico* è basato sull'assunzione che le condizioni in cui un esperimento è attuato determinino completamente il risultato dell'esperimento. Viceversa, un *modello stocastico* (non deterministico) è utilizzato per descrivere i fenomeni aleatori (*random*), per i quali non è possibile (o non conveniente) prevedere il risultato esatto dell'esperimento. La distinzione può essere illustrata in vari modi. Si pensi, ad esempio, all'esperimento del lancio di una moneta. Un eventuale modello deterministico richiederebbe la conoscenza dello stato iniziale della moneta, l'altezza dal suolo, la forza impressa inizialmente, eccetera. Nella pratica, può essere più conveniente ritenere random il risultato dell'esperimento e utilizzare per la sua descrizione un modello stocastico basato sul calcolo della probabilità.

Un altro contesto nel quale può essere opportuno l'utilizzo di un modello stocastico riguarda la *misurazione* di una grandezza fisica. Consideriamo come esempio la nota legge di Ohm  $V = IR$ . Mediante tale legge si può calcolare, ad esempio, il voltaggio  $V$  di un circuito quando sono note la resistenza  $R$  e l'intensità  $I$  della corrente. La legge corrisponde ad un modello deterministico per il calcolo di  $V$ , purché si ammetta che i valori di  $R$  e di  $I$  possano essere calcolati esattamente. In realtà, per vari motivi, dovuti allo strumento di misurazione e alle condizioni di misurazione, ripetute misurazioni possono portare a valori diversi. Di conseguenza, la legge fornisce successivamente valori differenti per la quantità  $V$ . In un modello stocastico si considerano le misure delle grandezze  $I, R$ , come variabili aleatorie; facendo, quindi, ipotesi sulla loro distribuzione statistica è possibile ricavare dalla relazione  $V = IR$  il comportamento aleatorio della variabile  $V$ . In questo modo l'accuratezza della misura  $V$  è data in termini di probabilità.

Come ultima esemplificazione, consideriamo lo studio del decadimento radioattivo di un elemento. Un classico modello deterministico è basato sull'ipotesi che la velocità di decadimento di una determinata massa  $m$  di una sostanza radioattiva sia proporzionale alla massa, cioè in termini matematici

$$\frac{dm}{dt} = -\lambda m \Rightarrow m = m_0 e^{-\lambda t}, \quad t \geq 0 \quad (8.56)$$

ove  $\lambda > 0$  è il tasso di velocità di decadimento e  $m_0$  è la massa al tempo  $t = 0$ . La legge formulata in (8.56) permette di calcolare esattamente, quando siano noti  $m_0$  e  $\lambda$ , la massa  $m(t)$  come funzione del tempo. La validità del modello può essere verificata confrontando i risultati forniti dalla legge (8.56) con i dati sperimentali. Tuttavia, dal punto di vista sperimentale non è possibile prevedere esattamente il numero di fissioni in un dato intervallo di tempo, a causa della natura random dell'istante in cui un elemento si disintegra. In un modello stocastico si introducono delle ipotesi concernenti la probabilità che un dato elemento decada nell'intervallo di tempo  $(0, t)$ . Da tale ipotesi si può ricavare la probabilità di avere esattamente  $k$  decadimenti nell'intervallo  $(0, t)$ . L'applicazione di un modello deterministico comporta la risoluzione di un problema matematico, nel quale i parametri hanno valori assegnati. In un modello stocastico, invece, i risultati seguono dal calcolo delle probabilità.

Dagli esempi precedenti si vede che la costruzione di un modello stocastico richiede, in sostanza, l'individuazione di tutti gli eventi di interesse e l'assegnazione a tali eventi di una opportuna distribuzione di probabilità. L'assegnazione di probabilità rappresenta, in generale, l'aspetto più interessante e più difficile nella costruzione del modello. Essa viene usualmente formulata come *ipotesi* sulla base di considerazioni teoriche e/o di esperienze eseguite su opportuni campioni estratti dalla popolazione totale degli eventi di interesse.

Abbiamo visto nei paragrafi precedenti che le leggi di probabilità possono essere caratterizzate da alcuni parametri significativi, quali ad esempio la media e la varianza. Pertanto, quando si ipotizza per il modello una determinata legge di probabilità, può essere conveniente cercare di stimare, mediante i campioni, tali parametri. Le basi del calcolo delle probabilità necessarie per risolvere questo primo problema dell'inferenza statistica, detto *problema di stima*, sono state fornite, in sostanza, nel paragrafo precedente, in particolare per quanto concerne gli intervalli di confidenza. Tuttavia, anche a scopo riassuntivo, raccoglieremo nel prossimo paragrafo le nozioni e le proprietà di base relative agli *stimatori*.

### 8.5.2 Stimatori e loro proprietà

Sia  $X_1, X_2, \dots, X_n$  un campione aleatorio estratto da una popolazione descritta da una variabile aleatoria  $X$ , con densità di probabilità  $f(\cdot; \theta)$ ; la funzione  $f$  sia nota a meno del vettore dei *parametri*  $\theta = [\theta_1, \theta_2, \dots, \theta_k]$ , con  $k \geq 1$  e  $\theta \in \Omega$ , ove  $\Omega$  è lo *spazio dei parametri*, ossia l'insieme dei possibili valori che il parametro  $\theta$  può assumere<sup>18</sup>. Supposto, inoltre, che si possano osservare i valori  $x_1, x_2, \dots, x_n$  del campione aleatorio  $X_1, X_2, \dots, X_n$ , si vuole stimare attraverso tali valori il valore del parametro  $\theta$ , oppure, più in generale, il valore di una funzione  $\tau(\theta)$  del para-

<sup>18</sup>Ad esempio, se  $X$  è il numero di chiamate telefoniche in  $T$  secondi, allora  $X$  può avere una distribuzione di Poisson con media  $\lambda T$ , ove  $\lambda$  è il numero medio di chiamate per secondo. In questo caso se  $\theta = \lambda T$ , si ha  $\theta \geq 0$ .



metro incognito. A tale scopo si può procedere in due maniere differenti. Nella prima, detta *stima puntuale*, si stima l'incognita  $\tau(\theta)$  mediante il valore di una *statistica* (funzione nota di variabili aleatorie osservabili, essa stessa variabile aleatoria), che indichiamo con  $T = t(X_1, X_2, \dots, X_n)$  e che viene detta uno *stimatore puntuale*. La seconda maniera di procedere, nota come *stima per intervalli*, consiste nel definire due statistiche, indicate con  $t_1(X_1, \dots, X_n)$  e  $t_2(X_1, \dots, X_n)$ , in modo che  $[t_1(X_1, \dots, X_n), t_2(X_1, \dots, X_n)]$  costituisca un intervallo per il quale si può determinare la probabilità che contenga l'incognita  $\tau(\theta)$ .

Ad esempio, se  $f(\cdot, \theta)$  è la densità normale  $\phi(x, \mu, \sigma)$ , ove il parametro  $\theta$  è il vettore  $[\mu, \sigma]$ , e si vuole stimare la media, cioè  $\tau(\theta) = \mu$ , allora la statistica  $\bar{X} = (1/n) \sum_{i=1}^n X_i$  è un possibile stimatore puntuale di  $\tau(\theta)$ , mentre  $[\bar{X} - 2\sqrt{s^2/n}, \bar{X} + 2\sqrt{s^2/n}]$  è un possibile stimatore per intervalli.

In questo paragrafo considereremo in particolare la stima puntuale. Per essa la problematica riguarda, in sostanza, la scelta dei criteri e delle tecniche per definire uno stimatore *ottimale*. Nel seguito passeremo in rassegna alcune di tali tecniche, in particolare il *metodo della massima verosimiglianza* (*maximum-likelihood*). Per quanto riguarda le notazioni,  $\hat{\theta}$  sarà utilizzato per indicare una stima di  $\theta$ ; inoltre, se  $\hat{\theta}$  è una stima di  $\theta$ , allora  $\hat{\Theta}$  indicherà il corrispondente stimatore di  $\theta$ , ossia  $\hat{\Theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ .

### Metodo dei momenti

Se  $f(\cdot; \theta)$  è la densità di una variabile aleatoria  $X$ , il momento assoluto  $r$ -mo  $\mu'_r = E(X^r)$  risulta una funzione nota del vettore  $\theta = [\theta_1, \dots, \theta_k]$ , cioè  $\mu'_r = \mu'_r(\theta_1, \dots, \theta_k)$ . Se  $X_1, \dots, X_n$  è un campione aleatorio con densità  $f(\cdot; \theta)$ , si indica con  $M'_j$  il momento campionario  $j$ -mo, ossia

$$M'_j = \frac{1}{n} \sum_{i=1}^n X_i^j$$

Si considera, quindi, il seguente sistema nelle  $k$  incognite  $\theta_1, \dots, \theta_k$

$$M'_j = \mu'_j(\theta_1, \dots, \theta_k), \quad j = 1, \dots, k \quad (8.57)$$

Nell'ipotesi che il sistema sia risolvibile in maniera univoca, la soluzione  $(\hat{\Theta}_1, \dots, \hat{\Theta}_k)$  rappresenta uno stimatore di  $(\theta_1, \dots, \theta_k)$ .

► **Esempio 8.36** Sia  $X_1, \dots, X_n$  un campione aleatorio corrispondente ad una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$ . Poniamo  $[\theta_1, \theta_2] = [\mu, \sigma]$ . Ricordando che  $\sigma^2 = \mu'_2 - (\mu'_1)^2$  e che  $\mu = \mu'_1$ , il sistema (8.57) diventa

$$\begin{aligned} M'_1 &= \mu'_1(\mu, \sigma) = \mu \\ M'_2 &= \mu'_2(\mu, \sigma) = \sigma^2 + \mu^2 \end{aligned}$$

da cui si ricava  $M'_1 = \bar{X}$  come stimatore di  $\mu$  e

$$\sqrt{M'_2 - \bar{X}^2} = \sqrt{\frac{1}{n} \sum X_i^2 - \bar{X}^2} = \sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2}$$

come stimatore di  $\sigma$ . Rileviamo che lo stimatore di  $\sigma$  ottenuto con il metodo dei momenti non coincide con lo stimatore  $s$  introdotto nel paragrafo precedente. ■

Il metodo dei momenti non individua univocamente gli stimatori, sia perché si potrebbero usare i momenti centrali, anziché come in precedenza i momenti assoluti, e sia anche perché potrebbero essere utilizzati momenti diversi dai primi  $k$ .

### Massima verosimiglianza

Si dice *funzione di verosimiglianza* di  $n$  variabili aleatorie  $X_1, \dots, X_n$  la densità congiunta delle  $n$  variabili aleatorie  $F_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta)$ , considerata come funzione di  $\theta$ . Nel caso particolare in cui  $X_1, \dots, X_n$  è un campione estratto corrispondente alla densità  $f(x; \theta)$ , la funzione di verosimiglianza è data dal prodotto  $f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$ . La funzione di verosimiglianza, denotata usualmente con  $L(\theta) = L(\theta; x_1, \dots, x_n)$  (da *likelihood*=verosimiglianza) fornisce la “verosimiglianza” che le variabili casuali assumano un particolare valore  $x_1, x_2, \dots, x_n$ . La verosimiglianza è il valore di una funzione di densità; per variabili aleatorie discrete essa corrisponde a una probabilità.

Supponiamo, ora, che per ogni realizzazione campionaria  $x_1, \dots, x_n$  esista il massimo della funzione campionaria  $L(\theta)$ , per  $\theta \in \Omega$ . Il valore  $\hat{\theta}$  del massimo risulta essere una funzione di  $x_1, \dots, x_n$ , cioè  $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ . In questo caso, la variabile aleatoria  $\hat{\Theta} = \hat{\theta}(X_1, \dots, X_n)$  viene chiamata *stimatore di massima verosimiglianza* del parametro  $\theta$ .

► **Esempio 8.37** Un campione aleatorio di lunghezza  $n$  corrispondente alla distribuzione normale di media  $\mu$  e varianza  $\sigma^2$  ha la densità congiunta (funzione di verosimiglianza)

$$L(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = \left( \frac{1}{2\pi\sigma} \right)^{n/2} e^{-\sum (x_i - \mu)^2 / 2\sigma^2}$$

ove  $\theta = [\mu, \sigma]$ , con  $\sigma > 0$  e  $\mu \in \mathbb{R}$ . Il calcolo del massimo di  $L(\theta)$  può essere semplificato passando al logaritmo, cioè alla considerazione della funzione

$$L^* = \ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

Ponendo uguali a zero le derivate parziali  $\partial L^* / \partial \mu$  e  $\partial L^* / \partial \sigma^2$  si ottengono le seguenti stime

$$\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x}; \quad \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Come si vede, in questo caso si ottengono gli stessi valori ottenuti in precedenza mediante il metodo dei momenti. ■

► **Esempio 8.38** Sia  $X_1, \dots, X_n$  un campione corrispondente ad una distribuzione uniforme sull'intervallo  $[\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma]$ , con

$$f(x; \theta) = f(x; \mu, \sigma) = \frac{1}{2\sqrt{3}\sigma} I_{[\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma]}(x)$$

Mediante il metodo dei momenti si ottiene  $\bar{X}$  come stima di  $\mu$  e

$$\sqrt{M'_2 - \bar{X}^2} = \sqrt{\frac{1}{n} \sum X_i^2 - \bar{X}^2} = \sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2}$$

come stima di  $\sigma$ . Stime del tutto diverse si ottengono con il metodo della massima verosimiglianza. Si può, infatti, vedere facilmente, che, se  $y_1$  è il valore osservato più piccolo e  $y_n$  il più grande, si hanno le seguenti stime di massima verosimiglianza

$$\hat{\mu} = \frac{y_1 + y_n}{2}; \quad \hat{\sigma} = \frac{y_n - y_1}{2\sqrt{3}}$$

■

Metteremo, ora, in evidenza alcune utili proprietà degli stimatori di massima verosimiglianza.

**Teorema 8.4** (Proprietà di invarianza) Sia  $\hat{\Theta} = (\hat{\Theta}_1, \dots, \hat{\Theta}_k)$ , con  $\hat{\Theta}_j = \hat{\theta}_j(X_1, \dots, X_n)$ , uno stimatore di massima verosimiglianza di  $\theta = (\theta_1, \dots, \theta_k)$  nella densità  $f(\cdot; \theta_1, \dots, \theta_k)$ . Se  $\tau(\theta) = (\tau_1(\theta), \dots, \tau_r(\theta))$ , per  $1 \leq r \leq k$ , è una trasformazione dello spazio dei parametri  $\Omega$ , allora uno stimatore di massima verosimiglianza di  $\tau(\theta) = (\tau_1(\theta), \dots, \tau_r(\theta))$ , è  $\tau(\hat{\Theta}) = (\tau_1(\hat{\Theta}), \dots, \tau_r(\hat{\Theta}))$ .

La proprietà di invarianza permette di considerare la stima  $(\theta_1, \dots, \theta_k)$ , invece di quella più generale  $\tau_1(\theta_1, \dots, \theta_k), \dots, \tau_r(\theta_1, \dots, \theta_k)$ .

Consideriamo, ad esempio, la densità normale con  $\theta = [\mu, \sigma^2]$  e  $\tau(\theta) = \mu + z_q\sigma$ , ove  $z_q$  è dato da  $\phi(z_q) = q$ . In altre parole  $\tau(\theta)$  è il quantile  $q$ -mo. Applicando il risultato precedente, si ha che lo stimatore di massima verosimiglianza di  $\tau(\theta)$  è dato da  $\bar{X} + z_q \sqrt{(1/n) \sum (X_i - \bar{X})^2}$ .

Il successivo risultato indica una proprietà di ottimalità dello stimatore di massima verosimiglianza. Il risultato è enunciato nel caso in cui  $\theta$  sia un numero reale.

**Teorema 8.5** (Consistenza) In ipotesi opportune di regolarità sulla funzione di densità  $f(x; \theta)$ , se  $\hat{\Theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$  è lo stimatore di massima verosimiglianza di  $\theta$  per un campione aleatorio di lunghezza  $n$  corrispondente alla densità  $f(x; \theta)$ , allora la distribuzione della variabile aleatoria  $\hat{\Theta}_n$  approssima, per  $n \rightarrow \infty$ , la distribuzione normale con media  $\theta$  e varianza

$$\sigma_*^2 = \frac{1}{n E_\theta \left[ \left( \frac{\partial}{\partial \theta} \ln f(X; \theta) \right)^2 \right]}$$

Il risultato precedente può essere ulteriormente precisato, dimostrando che per ogni  $\epsilon > 0$  si ha

$$\lim_{n \rightarrow \infty} P_\theta[|\hat{\Theta}_n - \theta| > \epsilon] = 0 \quad \forall \theta \in \Omega$$

Si dice, allora, che la successione  $\{\hat{\Theta}_n\}$  è *semplicemente consistente*. Inoltre, si può provare che se  $\{T_n\}$  è una qualsiasi altra successione di stimatori per cui la distribuzione di  $[T_n - \theta]$  tende alla distribuzione normale con media 0 e varianza  $\sigma^2(\theta)$ , allora  $\sigma^2(\theta)$  non è inferiore a  $\sigma_*^2$  per tutti i valori di  $\theta$  in un intervallo aperto.

Le proprietà precedenti si riassumono dicendo che la successione di stimatori  $\{\hat{\Theta}_n\}$  è *migliore e asintoticamente normale* (in sigla: BAN, *Best Asymptotically Normal*). L'attributo "migliore" si riferisce, naturalmente, alla varianza minima.

► **Esempio 8.39** Sia  $X_1, \dots, X_n$  un campionamento corrispondente alla distribuzione esponenziale  $f(x; \theta) = \theta e^{-\theta x} I_{[0, \infty)}(x)$ . Si può dimostrare che lo stimatore di massima verosimiglianza di  $\theta$  è dato da  $n / \sum_1^i X_i = 1/\bar{X}_n$ . Il Teorema 8.5 dice che tale stimatore ha una distribuzione asintotica normale con media  $\theta$  e varianza uguale a

$$\frac{1}{n E_\theta \left[ \left( \frac{\partial}{\partial \theta} \ln f(X; \theta) \right)^2 \right]} = \frac{\theta^2}{n}$$

■

## Minimi quadrati

Il metodo dei minimi quadrati è senza dubbio la procedura di stima più popolare<sup>19</sup>. Alcuni aspetti del metodo sono stati già considerati nei paragrafi precedenti (si veda in particolare l'analisi della retta di regressione). Ora cercheremo di inquadrare il metodo nell'ambito della teoria della stima dei parametri, limitandoci a considerare il seguente caso particolare di *modello statistico lineare*.

Supponiamo che ad ogni  $x$  di un sottoinsieme  $I$  della retta reale corrisponda una variabile aleatoria  $Y_x$  associata ad una funzione di ripartizione di probabilità  $F_{Y_x}(\cdot)$ . Come esemplificazione, si pensi ad una serie di osservazioni sperimentali in tempi successivi. Ipotizzando per semplicità che la varianza di  $F_{Y_x}(\cdot)$  sia una costante  $\sigma^2$  e che la media  $\mu(x)$  sia su una retta definita da  $\beta_0 + \beta_1 x$ , si vogliono stimare i parametri  $\beta_0, \beta_1$  sulla base di campioni aleatori  $Y_i (\equiv Y_{x_i})$  di lunghezza 1 estratti dalla funzione di ripartizione  $F_{Y_{x_i}}(\cdot)$ , in corrispondenza ad una serie di valori  $x_i$ , per

<sup>19</sup>Legendre (1805) fu il primo nella letteratura ad usare tale metodo per stimare i coefficienti nell'adattamento (fitting) di curve a dati assegnati. Gauss (1809) introdusse una base statistica nella stima dei parametri, mostrando che le stime ottenute mediante il metodo dei minimi quadrati massimizzano la densità di probabilità per una distribuzione normale degli errori. In questo senso, Gauss anticipava il metodo della massima verosimiglianza. Successivamente, Gauss stesso, Cauchy, Bienaymé, Chebychev, Gram, Schmidt, e altri diedero contributi relativamente all'aspetto computazionale del metodo dei minimi quadrati (lineare), mediante l'introduzione, in particolare, dei polinomi ortogonali.

$i = 1, 2, \dots, n$ . Pertanto, le coppie di valori  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  sono un insieme di osservazioni per le quali si ha

$$E(Y_i) = \beta_0 + \beta_1 x_i \quad (8.58)$$

$$\text{var}(Y_i) = \sigma^2 \quad (8.59)$$

per  $i = 1, 2, \dots, n$ .

▼ **Osservazione 8.7** L'attributo "lineare" dato al modello precedente si riferisce al fatto che la funzione  $\mu(\cdot)$  è lineare nei parametri incogniti. Nella definizione precedente la funzione è stata supposta lineare anche in  $x$ , ma questo non è essenziale nella definizione di modello lineare; ad esempio,  $E(Y) = \mu(x)$ , con  $\mu(x) = \beta_0 + \beta_1 \ln x$  è ancora un modello statistico lineare. Il caso più generale di modello lineare è quindi della forma  $\mu(x) = \beta_0 g_0(x) + \beta_1 g_1(x) + \dots + \beta_r g_r(x)$ , con  $r \geq 0$  e  $g_i(x)$  funzioni assegnate della variabile  $x$ . ■

Nel seguito supporremo che  $Y_1, Y_2, \dots, Y_n$  siano variabili aleatorie a due a due non correlate, ossia  $\text{cov}(Y_i, Y_j) = 0$ , per ogni  $i \neq j$ . Vengono allora detti *stimatori dei minimi quadrati* di  $\beta_0$  e  $\beta_1$  i valori di  $\beta_0$  e  $\beta_1$  che minimizzano la seguente somma dei quadrati degli scarti

$$S(\beta_0, \beta_1) := \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 \quad (8.60)$$

Se indichiamo con  $\hat{B}_0$  e  $\hat{B}_1$  tali stimatori, risolvendo il sistema lineare che si ottiene annullando le derivate parziali, si hanno i seguenti valori

$$\hat{B}_1 = \frac{\sum (Y_i - \bar{Y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \quad \hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{x}$$

Gli stimatori puntuali di  $\beta_0$  e  $\beta_1$  ottenuti con il metodo dei minimi quadrati hanno interessanti proprietà di ottimalità, come evidenziato da un teorema noto in letteratura come *Teorema di Gauss-Markov* e che afferma che gli stimatori  $\hat{B}_0$  e  $\hat{B}_1$  sono i migliori stimatori lineari non distorti di  $\beta_0$  e  $\beta_1$ . Il senso del risultato è il seguente. Consideriamo ad esempio il parametro  $\beta_0$ ; analogo risultato si ha per il parametro  $\beta_1$ . Il fatto che gli stimatori siano lineari significa che sono della forma  $\sum a_j Y_j$ , con  $a_j$  costanti; sono cioè funzioni lineari nelle variabili aleatorie  $Y_j$ . Il teorema afferma, allora, che  $\hat{B}_0$  è, tra tutti gli stimatori lineari non distorti, cioè tali da avere media uguale a  $\beta_0$ , quello a varianza minima.

▼ **Osservazione 8.8** Il Teorema di Gauss-Markov può essere esteso al caso in cui le variabili  $Y_i$  siano correlate nel seguente modo. Sia  $\mathbf{V}$  la matrice di covarianza relativa al vettore  $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^T$  e si indichi con  $\mathbf{Y} - \beta_0 - \beta_1 \mathbf{X}$  il vettore di componenti  $Y_i - \beta_0 - \beta_1 x_i$ , per  $i = 1, 2, \dots, n$ . Gli stimatori di  $\beta_0$  e  $\beta_1$  sono ottenuti minimizzando la seguente funzione quadratica

$$S(\theta_0, \theta_1) := (\mathbf{Y} - \beta_0 - \beta_1 \mathbf{X})^T \mathbf{V}^{-1} (\mathbf{Y} - \beta_0 - \beta_1 \mathbf{X})^T$$

e ancora sono tali da rendere minima la varianza tra tutti gli stimatori lineari non distorti. Aggiungiamo che nel caso in cui la distribuzione di  $\mathbf{Y}$  è normale, gli stimatori così ottenuti sono efficienti, ossia la corrispondente varianza è la minima ottenibile (con stimatori anche distorti). ■

### 8.5.3 Verifica di ipotesi

In questo paragrafo analizzeremo un altro aspetto importante della statistica inferenziale, noto come *verifica* (testing) di un'ipotesi statistica. In maniera schematica, il problema consiste nel controllare, mediante osservazioni e esperimenti, la verità di un'affermazione relativa al comportamento di un fenomeno. L'affermazione può riguardare qualche parametro di una distribuzione assegnata (*ipotesi parametriche*) o la forma stessa della distribuzione di una variabile aleatoria o di una media campionaria associata ad una popolazione statistica (*ipotesi non parametriche*).

L'affermazione che si vuole controllare è detta *ipotesi nulla*, ed è indicata usualmente con il simbolo  $H_0$ . L'affermazione che si contrappone all'ipotesi  $H_0$  è detta *ipotesi alternativa* ed è indicata con  $H_1$ . Come esemplificazione, si formuli un'ipotesi relativa alla media di una popolazione. Più precisamente, si voglia controllare la veridicità dell'affermazione: *il peso medio degli adulti in una nazione è  $\mu = 68.5$  Kg<sup>20</sup>*. In questo esempio si ha

$$\begin{aligned} H_0 : \mu &= 68.5 \\ H_1 : \mu &\neq 68.5 \end{aligned} \quad (8.61)$$

L'ipotesi è verificata mediante l'analisi di campioni estratti dalla popolazione.

L'ipotesi nulla  $H_0$  può essere a priori vera o falsa e l'utilizzazione dei test dell'inferenza statistica può portare solo a una di queste conclusioni:

1. l'ipotesi nulla  $H_0$  è rifiutata, oppure
2. non è possibile rifiutare l'ipotesi nulla  $H_0$ .

Le possibili differenti situazioni sono illustrate nella seguente tabella.

L'induzione conduce a	$H_0$ è vera	$H_0$ è falsa
accettare $H_0$	conclusione esatta	conclusione errata
respingere $H_0$	conclusione errata	conclusione esatta

Gli eventi corrispondenti a tali situazioni sono i seguenti.

1. L'induzione conduce ad accettare  $H_0$ , quando questa è vera;

<sup>20</sup>In modo analogo, si potrebbe considerare l'ipotesi: *il peso medio degli adulti in una nazione è almeno 68.5 Kg*, cioè  $\mu \geq 68.5$  Kg.

2. l'induzione conduce a rifiutare  $H_0$ , quando questa è vera;
3. l'induzione conduce ad accettare  $H_0$ , quando questa è falsa;
4. l'induzione conduce a rifiutare  $H_0$ , quando questa è falsa.

Nel caso degli eventi 2 e 3 l'induzione conduce a delle conclusioni errate. L'evento 2 viene denominato *errore di primo tipo*, mentre l'evento 3 è chiamato *errore di secondo tipo*.

Nel caso dell'esempio (8.61), si può o rifiutare l'affermazione che  $\mu = 68.5$  (e concludere che  $\mu \neq 68.5$ ) oppure ammettere che non possiamo rifiutare l'affermazione sulla base dell'informazione disponibile. Sottolineiamo che quando si conclude di non essere in grado di rifiutare l'ipotesi nulla, non si afferma che tale ipotesi è vera, ma semplicemente si ammette di non essere in grado di rifiutarla. Continuando sull'esempio, supponiamo di avere ottenuto su un campione aleatorio di adulti il seguente valore della *media campionaria*

$$\bar{x} = 68.3$$

Tale valore differisce dal valore  $\mu$  ipotizzato, ma la differenza 0.2 potrebbe non essere, nelle intenzioni di chi ha fatto l'ipotesi, "significativa". Si tratta, in altre parole, di introdurre nell'operazione di rifiuto dell'ipotesi nulla un *livello di significatività*. Per fare questo è necessario considerare la legge di probabilità secondo la quale è distribuita la variabile aleatoria  $\bar{x}$  al variare del campione a cui si riferisce.

Dal teorema centrale limite sappiamo che la media campionaria  $\bar{x}$  può essere approssimata per la dimensione  $n$  del campione sufficientemente grande da una distribuzione normale con media  $\mu_{\bar{x}} = \mu$  e deviazione standard  $\sigma_{\bar{x}}$ , ove  $\mu$  e  $\sigma$  sono i parametri della popolazione da cui sono estratti i campioni. Ricordiamo, anche, che nel caso in cui la popolazione sia distribuita normalmente (come può essere ammesso per l'esempio che stiamo considerando), allora anche la variabile  $\bar{x}$  è distribuita normalmente. Osserviamo, infine, che la deviazione standard della popolazione  $\sigma$  può essere stimata dallo stimatore  $s$ .

Ora possiamo precisare le condizioni nelle quali siamo disposti a rifiutare l'ipotesi nulla. Ad esempio, potremmo accettare un rischio non superiore al 5% di rifiutare l'ipotesi nulla quando non dovremmo, cioè di commettere un errore rifiutandola. In questo caso si dice che *si sta verificando l'ipotesi nulla al livello di significatività*  $\alpha = 0.05$ . Il livello di significatività, usualmente indicato con  $\alpha$ , è la probabilità (rischio) di rifiutare *erroneamente* l'ipotesi nulla, cioè la probabilità di commettere un errore di primo tipo.

Possiamo, allora, procedere, nel seguente modo. Si costruisce un intervallo di confidenza di livello 95% per  $\mu$  e si controlla se il valore *ipotizzato*  $\mu$  è o no in tale intervallo. Nell'esempio supponiamo di avere calcolato su un campione di lunghezza  $n = 100$  una media campionaria  $\bar{x} = 69.1$ , con una deviazione standard campionaria

$s = 7.0$ . Allora, utilizzando la distribuzione normale<sup>21</sup> con deviazione standard  $s/\sqrt{n}$  si ottiene che l'intervallo di confidenza di livello 95% di  $\mu$  è dato da

$$\bar{x} - 1.96 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{s}{\sqrt{n}} \Rightarrow 67.7 \leq \mu \leq 70.5$$

e possiamo essere “confidenti al 95%” che il valore di  $\mu$  è nell'intervallo (67.7, 70.5). Più precisamente, si ha che la probabilità che l'intervallo  $(\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n})$ , considerato come variabile casuale al variare del campione, contenga il numero  $\mu$ , è uguale a 0.95. Poiché il valore ipotizzato  $\mu = 68.5$  giace in tale intervallo, non vi è evidenza per rifiutare l'ipotesi nulla  $H_0$ . Pertanto *al livello di significatività  $\alpha = 0.05$ , l'ipotesi nulla  $H_0 : \mu = 68.5$  non è rifiutata.*

Sottolineiamo, ancora, che non abbiamo dimostrato l'ipotesi  $\mu = 68.5$ ; semplicemente abbiamo fallito nel rifiutarla. In altre parole, si può dire che la differenza tra  $\bar{x} = 69.1$  e il valore ipotizzato  $\mu = 68.5$  *non è significativa al livello 0.05.*

Osserviamo che la costruzione dell'intervallo di confidenza può essere semplificata passando dalla variabile campionaria  $\bar{x}$  alla distribuzione normale standard, mediante la trasformazione

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Per  $\alpha = 0.05$ , è allora sufficiente vedere se  $z > 1.96$  o  $z < -1.96$ . Come applicazione, supponiamo che nell'esempio precedente i valori  $\bar{x} = 69.1$  e  $s = 7$  siano stati ottenuti su un campione di  $n = 625$  elementi. In questo caso si ha  $z = (69.1 - 68.5)/(7.0/\sqrt{625}) \approx 2.14$ . Poiché  $z = 2.14 > 1.96$ , l'ipotesi nulla  $H_0 : \mu = 68.5$  viene rifiutata al livello 0.05.

Confrontando con i risultati precedenti, si vede l'influenza sul test di induzione da parte della dimensione del campione. Con gli stessi valori di  $\bar{x}$  e di  $s$ , ma per  $n = 100$ , l'ipotesi non era stata rifiutata allo stesso livello di significatività. Osserviamo che non vi è contraddizione nei due risultati ottenuti, in quanto avevamo accettato un rischio del 5% di esserci sbagliati!

◆ **Esercizio 8.45** *Si ipotizza che il tempo richiesto da una determinata reazione chimica sia, in media, 24.0 secondi. Per verificare l'ipotesi  $\mu = 24.0$  al livello di significatività 0.05, si esegue la reazione 36 volte, ottenendo una media campionaria  $\bar{x} = 23.5$  secondi con deviazione standard  $s = 1.4$  secondi. Analizzare la conclusione.*

◆ **Esercizio 8.46** *Un ricercatore desidera verificare l'ipotesi che un dato farmaco ha effetti apprezzabili entro al più 4 minuti dopo la sua somministrazione. Il ricercatore considera un campionamento di 100 pazienti e ottiene una media campionaria di 4 minuti e 6 secondi. Analizzare la conclusione a livello di significatività  $\alpha = 0.05$ , se la deviazione standard campionaria  $s$  è rispettivamente (a) 0.64 (b) 0.62 (c) 0.60 minuti.*

<sup>21</sup>Come si è visto nei paragrafi precedenti, quando non è nota la deviazione  $\sigma$ , ma essa è stimata dalla media campionaria  $s$ , sarebbe più opportuno considerare la distribuzione  $t$  di Student.



◆ **Esercizio 8.47** Si ipotizza che l'età media dei frequentatori di una biblioteca sia  $\mu = 39.0$  anni. Per verificare tale ipotesi si ottiene sulla base di un campionamento  $\bar{x} = 38.0$  anni con  $s = 10.2$  anni. Si cerca qual è la dimensione campionaria più piccola affinché la precedente informazione campionaria porti a rifiutare l'ipotesi nulla al livello di significatività  $\alpha = 0.05$ .

La tecnica illustrata in precedenza su una applicazione particolare può essere utilizzata, opportunamente generalizzata, per la verifica di altre ipotesi importanti nelle applicazioni. Ricordiamo, in particolare, lo studio della significatività della differenza  $\mu_1 - \mu_2$  delle medie di due distinte popolazioni, o più in generale per un numero qualunque di popolazioni, la verifica dell'ipotesi  $\mu_1 = \mu_2 = \mu_3 = \dots$  mediante l'analisi della varianza (ANOVA). Ci limiteremo ad illustrare alcune di queste applicazioni mediante opportuni esempi.

▼ **Osservazione 8.9** Ricordiamo che l'impostazione utilizzata in questo paragrafo è anche nota come inferenza classica o frequentista. In tale impostazione, si osserva una variabile casuale  $X$  con funzione di densità  $f(x; \theta)$ , ove  $\theta$  è un parametro non osservato, e si cerca di trarre conclusioni per  $\theta$  a partire dai dati osservati  $X$ .

Un approccio differente, indicato usualmente come approccio bayesiano, consiste nel ritenere  $\theta$  ancora non osservabile, ma, anziché costante come nella impostazione classica, come una variabile casuale  $\Theta$  con distribuzione nota  $\pi(\theta)$ .

Un modello statistico bayesiano consiste allora di

1. una variabile casuale osservata  $X$ ,
2. di una variabile casuale non osservata  $\Theta$ ,
3. della densità condizionata  $f(x | \theta)$  di  $X$  data  $\theta$ ,
4. della densità marginale  $\pi(\theta)$  di  $\Theta$ .

La distribuzione  $\pi(\Theta)$  è anche chiamata distribuzione a priori di  $\Theta$ . Poiché il parametro  $\Theta$  è costante nella statistica classica e un vettore casuale nella statistica bayesiana, i valori aspettati nella statistica classica sono sostituiti da valori aspettati condizionati data  $\Theta$ . In particolare, se  $T(X)$  è uno stimatore di  $\tau(\Theta)$ , allora  $T(X)$  è non distorto se  $E(T(X | \Theta)) = \tau(\Theta)$ .

L'approccio classico e l'approccio bayesiano alla statistica sono quindi punti di vista completamente differenti. I vari concetti di stimatore, test, e parametro hanno un significato differente nei due sistemi. Uno dei vantaggi dell'analisi bayesiana è che i concetti sono in generale di più immediata comprensione ed inoltre in essa è possibile includere automaticamente la conoscenza acquisita in precedenza. Un possibile svantaggio è la soggettività relativa alla distribuzione di probabilità del parametro, ossia l'inclusione automatica di pregiudizi<sup>22</sup>. Per una discussione più approfondita si veda ad esempio Barnett [12], Arnold [6].

<sup>22</sup>In practice, it seems that statistics is more of an art than a science. There are few experiments in which all the assumptions of an analysis (classical or bayesian) are satisfied. Therefore, it is important to remain flexible, using whatever approach seems appropriate. If a person or organization is doing an experiment solely in order to learn something for him- or herself or for the organization, then a bayesian analysis incorporating all the the person's or company prior beliefs

► **Esempio 8.40** *Differenze tra medie.* Supponiamo di avere due popolazioni, con medie, rispettivamente  $\mu_1$  e  $\mu_2$  e di voler analizzare la *differenza tra le medie*  $\mu_1 - \mu_2$ . Come esemplificazione, supponiamo che le due popolazioni corrispondano alle durate in ore di due differenti marche di pile, denotate rispettivamente con A e B. Allora, in media, le pile A durano  $\mu_1$  ore e le pile B  $\mu_2$  ore. Quando  $\mu_1 = \mu_2$ , le due marche di pile hanno in media la stessa durata e quindi possono essere considerate della stessa qualità. Per il *controllo di qualità* è quindi importante stabilire se  $\mu_1 \neq \mu_2$  ad un fissato livello di confidenza. Si possono, quindi, considerare le seguenti ipotesi

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2$$

L'ipotesi nulla è verificata per campionamento delle due popolazioni. Siano  $\bar{x}_1, \bar{x}_2$ , le corrispondenti medie campionarie. Consideriamo, quindi, la distribuzione della variabile aleatoria  $\mu_1 - \mu_2$ . Come conseguenza del teorema limite centrale, si può dimostrare che *data una popolazione con media  $\mu_1$  e deviazione standard  $\sigma_1$ , e una seconda popolazione con media  $\mu_2$  e deviazione standard  $\sigma_2$ , se un esperimento consiste nel prendere un campione aleatorio di lunghezza  $n_1$ , sufficientemente grande, dalla prima popolazione e un campione aleatorio di ampiezza  $n_2$  dalla seconda popolazione e vengono indicate con  $\bar{x}_1, \bar{x}_2$  le corrispondenti media campionarie, le differenze  $\bar{x}_1 - \bar{x}_2$ , ottenute in una successione di esperimenti, tendono ad essere distribuite normalmente con media e deviazione standard*

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2, \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} := \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Il risultato precedente permette la verifica dell'ipotesi  $H_0(\mu_1 = \mu_2)$  a livelli di significatività fissati (ad esempio, 0.05 e 0.01). Posto

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

si rifiuta l'ipotesi nulla  $H_0(\mu_1 = \mu_2)$  solo se  $z$  è maggiore di 1.96 o minore di -1.96 (al livello di significatività  $\alpha = 0.05$ ) e al di fuori dell'intervallo  $(-2.58, 2.58)$  (al livello di significatività  $\alpha = 0.01$ ). ■

► **Esempio 8.41** *Ipotesi sulle proporzioni.* Si è interessati a conoscere la proporzione  $\pi$  di una popolazione con determinate caratteristiche, ad esempio, la percentuale di automobilisti che utilizza una determinata strada per raggiungere una fissata località. In questo caso le ipotesi nulle possono assumere, ad esempio, le seguenti forme

$$H_0 : \pi = 0.60, \quad \text{oppure} \quad H_0 : \pi \geq 0.60$$

La verifica di tali ipotesi si basa sul risultato che per  $n$  sufficientemente grande le proporzioni campionarie  $p$  hanno una distribuzione che è approssimata dalla distribuzione normale con media  $\mu_p = \pi$  (la vera proporzione della popolazione) e deviazione standard

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

*may be appropriate. If, on the other hand, the person or company wants to convince other people or organizations of the truth of his or her analysis, then a classical analysis would appear to be more appropriate* (Arnold [6]).

ove  $n$  è la lunghezza del campione. Si utilizza, quindi, la trasformazione

$$z = \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}$$

Come esemplificazioni, consideriamo le seguenti applicazioni. Supponiamo che una casa farmaceutica sostenga che un determinato farmaco sperimentale produca effetti apprezzabili nel 60% dei casi. Per verificare tale ipotesi al livello 0.05, si sperimenta il farmaco su 200 pazienti e si osserva un effetto su 134 di essi, con una proporzione campionaria  $p = 134/200 = 0.67$ . Vediamo la conclusione che si può trarre sulla ipotesi nulla  $H_0 : (\pi = 0.60)$ . Si ha la trasformazione

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.67 - 0.60}{\sqrt{\frac{0.60(1-0.60)}{200}}} \approx 2.02 > 1.96$$

sicché l'ipotesi nulla è rifiutata. In altre parole, possiamo "credere al 95%" che l'affermazione della casa farmaceutica non sia corretta.

Come ulteriore applicazione, supponiamo che per verificare l'ipotesi ( $H_0 : \pi \geq 0.70$ ) si sia fatto un campionamento di lunghezza 100, ottenendo  $p = 0.67$ . Vediamo la conclusione al livello di confidenza  $\alpha = 0.05$ . Applicando la trasformazione precedente, si ottiene

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.67 - 0.70}{\sqrt{\frac{0.70(1-0.70)}{100}}} \approx -0.65$$

Siccome  $-0.65$  non è inferiore a  $-1.64$  sulla sinistra (cfr. Figura 8.20), il risultato  $p = 0.67$  non giustifica il rifiuto dell'ipotesi nulla. ■

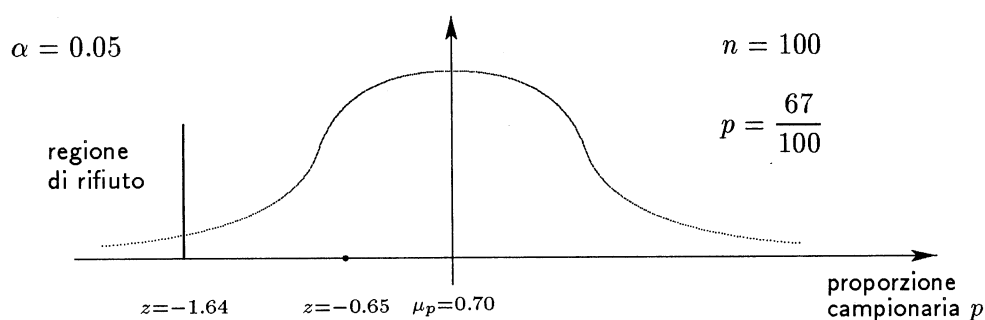


Figura 8.20: Rappresentazione della zona di rifiuto.

► **Esempio 8.42** *Ipotesi sulle varianze.* Date due differenti popolazioni, si vuole verificare l'ipotesi nulla

$$H_0 : \sigma_1^2 = \sigma_2^2$$

In questo caso si tiene conto che il rapporto delle varianze campionarie  $s_1^2/s_2^2$ , oppure  $s_2^2/s_1^2$  (si sceglie tra i due rapporti quello per il quale il numeratore è maggiore del denominatore) segue la distribuzione  $F$ . Il grado di libertà del numeratore è dato dalla lunghezza del campione utilizzato per formare il numeratore diminuita di uno, e in modo simile si calcola il grado di libertà del denominatore. Si calcola, quindi, un valore critico di  $F_{\alpha, \nu_1, \nu_2}$  corrispondente al livello fissato  $\alpha$  di confidenza.

Come applicazione, supponiamo di voler verificare se due tipi differenti di pile, indicate con A e B, hanno una durata che è dispersa in maniera significativamente differente. L'ipotesi nulla è, pertanto,  $H_0 : \sigma_A^2 = \sigma_B^2$ , ove  $\sigma_A$  e  $\sigma_B$  sono le deviazioni standard delle due popolazioni. Consideriamo come livello di confidenza  $\alpha = 0.05$ . Supponiamo che da un campione aleatorio di  $n_1 = 25$  pile del tipo A si ottenga  $s_A^2 = 31.4$ , e analogamente da un campione di  $n_2 = 24$  pile di tipo B si ottenga  $s_B^2 = 38.3$ . Si calcola

$$F = \frac{s_B^2}{s_A^2} = \frac{38.3}{31.4} \approx 1.22$$

Il grado di libertà del numeratore è dato da  $n_2 - 1 = 23$  e del denominatore  $n_1 - 1 = 24$ . Si trova, allora,  $F_{0.05} = 2.01$ . Poiché tale valore è maggiore del valore calcolato 1.22 non si è in grado di rifiutare l'ipotesi nulla al livello 0.05. In altre parole, le varianze campionarie non differiscono a sufficienza per sostenere con 95% di confidenza che le vere varianze  $\sigma_A^2$  e  $\sigma_B^2$  siano differenti. ■

► **Esempio 8.43** *Analisi della varianza*. Supponiamo di avere un certo numero di popolazioni con medie  $\mu_1, \mu_2, \dots$  e con medie campionarie  $\bar{x}_1, \bar{x}_2, \dots$ . Per il seguito supporremo che i campioni considerati per le differenti popolazioni abbiano la medesima lunghezza. Si vuole verificare la seguente ipotesi nulla

$$H_0 : \mu_1 = \mu_2 = \mu_3 \dots$$

contro l'ipotesi alternativa

$$H_1 : \text{almeno due } \mu_i \text{ sono differenti}$$

La tecnica utilizzata per la verifica di tale ipotesi è nota in statistica come *analisi della varianza*, e in forma abbreviata ANOVA. Per illustrare l'idea di base della tecnica, consideriamo il caso di tre differenti popolazioni, indicate con  $P_1$ ,  $P_2$  e  $P_3$ . L'ipotesi nulla è allora

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Supponiamo di aver ottenuto i seguenti risultati campionari (su campioni di lunghezza 7).

$P_1$ :	47	53	43	50	41	56	53
$P_2$ :	44	50	46	40	45	42	48
$P_3$ :	54	44	55	50	46	53	55

I risultati campionari possono essere considerati come gli elementi di una matrice  $\mathbf{X} = [x_{i,j}]$  di  $m = 3$  righe e  $n = 7$  colonne. Poiché per le differenti popolazioni si tiene conto di una sola caratteristica, detta *fattore* (factor), si dice anche che l'analisi effettuata è a *una entrata* (one-way analysis of variance). Le  $m$  popolazioni vengono anche chiamate *livelli* (levels)

del fattore<sup>23</sup>. La riga  $i$ -ma rappresenta i risultati campionari relativi alla popolazione  $P_i$ . Poniamo  $N = n \times m$ , e costruiamo le seguenti quantità, dette *totali per righe*.

$$T_1 = \sum_{j=1}^n x_{1,j} = 343; \quad T_2 = \sum_{j=1}^n x_{2,j} = 315; \quad T_3 = \sum_{j=1}^n x_{3,j} = 357$$

In corrispondenza si ottengono i seguenti valori per le medie campionarie

$$\bar{x}_1 = \frac{1}{n}T_1 = 49.0; \quad \bar{x}_2 = \frac{1}{n}T_2 = 45.0; \quad \bar{x}_3 = \frac{1}{n}T_3 = 51.0$$

Si tratta ora di vedere se le differenze tra tali medie campionarie sono *significantive* ad un livello  $\alpha$ .

A tale scopo, calcoliamo le seguenti *somme di quadrati* (SS)

$$\begin{aligned} SS_{\text{total}} &= \sum_{\substack{i=1,m \\ j=1,n}} x_{i,j}^2 - \frac{1}{N} \left( \sum_{\substack{i=1,m \\ j=1,n}} x_{i,j} \right)^2 \\ SS_{\text{factor}} &= \frac{\sum_{i=1,m} T_i^2}{n} - \frac{1}{N} \left( \sum_{\substack{i=1,m \\ j=1,n}} x_{i,j} \right)^2 \\ SS_{\text{error}} &= SS_{\text{total}} - SS_{\text{factor}} \end{aligned}$$

Nell'esempio numerico si ottengono i seguenti valori

$$SS_{\text{total}} = 506.67; \quad SS_{\text{factor}} = 130.67; \quad SS_{\text{error}} = 376.00$$

I *gradi di libertà* (df) sono calcolati nel seguente modo

$$\begin{aligned} df_{\text{factor}} &= m - 1 = 2 \\ df_{\text{total}} &= N - 1 = 20 \\ df_{\text{error}} &= df_{\text{total}} - df_{\text{factor}} = 18 \end{aligned}$$

I risultati sono raccolti nella Tabella 8.3, detta *tabella ANOVA*. Si calcolano, infine, le seguenti quantità, chiamate *quadrati medi* (mean squares) MS

$$\begin{aligned} MS_{\text{factor}} &= \frac{SS_{\text{factor}}}{df_{\text{factor}}} = 65.335 \\ MS_{\text{error}} &= \frac{SS_{\text{error}}}{df_{\text{error}}} = 20.889 \end{aligned}$$

Si calcola, allora, il rapporto F (*F-ratio*)

$$F = \frac{MS_{\text{factor}}}{MS_{\text{error}}} \approx 3.13$$

<sup>23</sup>Situazioni di questo tipo si hanno, ad esempio, in farmacologia, quando si studiano gli effetti di farmaci diversi. Le differenti popolazioni corrispondono ai differenti trattamenti. L'ipotesi nulla equivale a dire che non esiste una differenza significativa tra i vari trattamenti.

source	SS	df	MS	F
Factor	130.67	2	65.335	3.13
Error	376.00	18	20.889	
Total	506.67	20		

Tabella 8.3: Tabella ANOVA.

Tale rapporto può essere interpretato come

$$\frac{\text{variabilità tra campioni}}{\text{variabilità entro i campioni}}$$

e viene utilizzato per verificare l'ipotesi nulla nel solito modo. Fissato un livello  $\alpha$ , si utilizza la distribuzione  $F_{\nu_1, \nu_2}$  corrispondente ai gradi di libertà  $\nu_1$  e  $\nu_2$ , dati rispettivamente dal numeratore ( $df_{\text{factor}}$ ) e dal denominatore ( $df_{\text{error}}$ ). Nel caso dell'esempio si trova per  $\alpha = 0.05$  il valore  $F_{2, 18, 0.05} = 3.55$ . Poiché

$$F_{\text{calcolato}} = 3.13 < 3.55$$

l'ipotesi nulla  $\mu_1 = \mu_2 = \mu_3$  non viene rifiutata al livello di significatività 0.05, cioè l'evidenza campionaria non fornisce un argomento adeguato (al livello 0.05) contro l'ipotesi che le tre popolazioni abbiano la stessa media. Se al contrario, avessimo ottenuto un valore calcolato di  $F$  maggiore del valore critico di  $F$ , avremmo potuto concludere con 95% di confidenza che le medie *non* sono tutte le stesse.

Ricordiamo che l'applicazione del test  $F$  richiede le seguenti ipotesi sulle popolazioni

1. la distribuzione di ogni popolazione è con buona approssimazione una distribuzione normale;
2. le varianze  $\sigma^2$  di tutte le popolazioni sono uguali (tale proprietà è nota in letteratura come *homoscedasticity*).

Tuttavia, l'analisi della varianza è una tecnica statistica "robusta", nel senso che essa può fornire risultati soddisfacenti anche nel caso in cui le ipotesi precedenti siano soddisfatte solo in maniera approssimata. ■

► **Esempio 8.44** *Coefficiente di correlazione.* Dato un insieme campionario di coppie di valori  $(x_1, y_1), \dots, (x_n, y_n)$ , la retta di regressione dei minimi quadrati di  $y$  in  $x$ , o rispettivamente di  $x$  in  $y$ , può essere scritta nella seguente forma

$$y = \bar{y} + \frac{s_{xy}}{s_x^2} (x - \bar{x}), \quad x = \bar{x} + \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

ove

$$s_x^2 = \frac{\sum (x - \bar{x})^2}{n}, \quad s_y^2 = \frac{\sum (y - \bar{y})^2}{n}, \quad s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

e  $\bar{x}$ ,  $\bar{y}$  sono le medie campionarie di  $x_i$  e rispettivamente di  $y_i$ , per  $i = 1, 2, \dots, n$ . Il *coefficiente di correlazione campionario* è definito da

$$r = \frac{s_{xy}}{s_x s_y}$$

Quando  $r = \pm 1$  si ha una correlazione lineare perfetta. Si può mostrare che nell'ipotesi  $\rho = 0$  la statistica

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

ha una distribuzione  $t$  di Student con  $n-2$  gradi di libertà; tale proprietà può essere quindi opportunamente utilizzata per verificare l'ipotesi  $\rho = 0$ .

Come illustrazione, supponiamo che il coefficiente di correlazione calcolato su un determinato campione di lunghezza 18 sia dato da 0.32; si vuole allora verificare la seguente ipotesi nulla

$$H_0 : \rho = 0; \quad H_1 : \rho > 0$$

ove  $\rho$  è il coefficiente di correlazione da stimare. Osserviamo che il test sull'ipotesi  $\rho = 0$  corrisponde a verificare se la correlazione osservata attraverso il campione è significativa, oppure se essa può essere spiegata attraverso le fluttuazioni del campionamento. Se la correlazione è trovata significativa, si pone successivamente il problema dell'interpretazione della relazione tra le variabili  $X$  e  $Y$  (*ma questo non è più un problema di natura statistica*).

Nel caso considerato si ha

$$\bar{t} = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.32 \sqrt{16}}{\sqrt{1-(0.32)^2}} \approx 1.3510$$

Ad un livello di significatività 0.05, si deve confrontare  $\bar{t}$  con il valore  $t_{0.95}$  per 16 gradi di libertà. Poiché tale valore è dato approssimativamente da 1.746, non è possibile respingere l'ipotesi  $h_0$  al livello 0.05. Lasciamo come esercizio verificare che il minimo valore della lunghezza  $n$  del campione per cui il coefficiente di correlazione campionario 0.32 è significativamente maggiore di zero al livello 0.05 è dato da  $n = 28$ .

Terminiamo l'esempio ricordando che la statistica

$$Z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$$

ha una distribuzione approssimativamente normale con media e deviazione standard

$$\mu_Z = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right), \quad \sigma_Z = \frac{1}{\sqrt{n-3}}$$

La precedente trasformazione, nota come *trasformazione Z di Fisher*, può essere utilizzata per la verifica dell'ipotesi  $\rho = \rho_0$ , con  $\rho_0$  valore fissato  $\neq 0$ , e per trovare gli intervalli di confidenza per il coefficiente di correlazione. Come illustrazione, supponiamo che per un campione di lunghezza 12 si sia ottenuto  $r = 0.856$ . In corrispondenza, si ottiene  $Z \approx 1.278$  e  $\sigma_Z \approx 0.333$ . L'intervallo di confidenza al 95% di  $\mu_Z$  è dato da

$$1.278 - [1.96 \times 0.333] < \mu_Z < 1.278 + [1.96 \times 0.333] < \mu_Z \Rightarrow 0.625 < \mu_Z < 1.930$$

Applicando la trasformata inversa di Fisher, si ottiene il seguente intervallo di confidenza al 95 % per  $\rho$

$$0.554 < \rho < 0.958$$

► **Esempio 8.45** *Wilcoxon signed rank test*. Il test di Wilcoxon, introdotto nel 1945, è una delle prime procedure *non parametriche* sviluppate, e rimane uno dei test non parametrici più utilizzati. Essa verrà utilizzata per introdurre le idee che sono alla base della statistica non parametrica (o *distribution free*).

Sia  $Y_1, Y_2, \dots, Y_n$  un campione di lunghezza  $n$  estratto da una densità di probabilità  $f_Y(y)$ , che è supposta continua e simmetrica. Sia  $\tilde{\mu}$  la mediana di  $f_Y(y)$ . Si desidera testare l'ipotesi

$$\begin{array}{l} H_0 : \tilde{\mu} = \tilde{\mu}_0 \\ \text{contro} \\ H_1 : \tilde{\mu} \neq \tilde{\mu}_0 \end{array}$$

ove  $\tilde{\mu}_0$  è un valore prefissato di  $\tilde{\mu}$ .

La statistica di Wilcoxon è basata sulla grandezza, e le direzioni, delle deviazioni di  $Y_i$  da  $\tilde{\mu}_0$ . Le deviazioni  $|Y_1 - \tilde{\mu}_0|, |Y_2 - \tilde{\mu}_0|, \dots, |Y_n - \tilde{\mu}_0|$  possono essere ordinate in ordine crescente; definiamo  $R_i$  il grado (*rank*) di  $|Y_i - \tilde{\mu}_0|$  nell'insieme  $\{|Y_j - \tilde{\mu}_0|, j = 1, \dots, n\}$ , ossia al più piccolo  $|Y_j - \tilde{\mu}_0|$  è assegnato un rank 1, al secondo più piccolo un rank 2, eccetera.

Associato con ogni  $R_i$  vi è un indicatore di segno  $Z_i$ , ove

$$Z_i = \begin{cases} 0 & \text{se } Y_i - \tilde{\mu}_0 < 0 \\ 1 & \text{se } Y_i - \tilde{\mu}_0 > 0 \end{cases}$$

Si definisce come statistica di Wilcoxon  $W$  la combinazione lineare

$$W = \sum_{i=1}^n Z_i R_i$$

Come illustrazione, consideriamo il caso in cui  $n = 3$  e  $y_1 = 6.$ ,  $y_2 = 4.9$  e  $y_3 = 11.2$ . Inoltre, nell'ipotesi da testare sia  $\tilde{\mu}_0 = 10$ . Si ha

$$|y_1 - \tilde{\mu}_0| = 4., \quad |y_2 - \tilde{\mu}_0| = 5.1, \quad |y_3 - \tilde{\mu}_0| = 1.2 \Rightarrow r_1 = 2, \quad r_2 = 3, \quad r_3 = 1$$

e  $z_1 = 0, z_2 = 0, z_3 = 1$ . Pertanto

$$w = \sum_{i=1}^n z_i r_i = (0)(2) + (0)(3) + (1)(1) = 1$$

Sottolineiamo il fatto che  $W$  dipende solo dai rank delle deviazioni da  $\tilde{\mu}_0$  e non dalle deviazioni stesse; per tenere conto di quest'ultime sarebbero necessarie ulteriori conoscenze per la distribuzione  $f_Y(y)$ , contrariamente allo spirito delle procedure non parametriche.

È chiaro che  $W$  assume i suoi valori da 0 (tutte le deviazioni negative) a  $\sum_{i=1}^n 1 = [n(n+1)]/2$  (tutte le deviazioni positive). Intuitivamente, se  $H_0$  fosse vera,  $W$  dovrebbe avere valori vicini a  $[n(n+1)]/4$ . Il successivo risultato permette di determinare quanto  $W$  dovrebbe essere vicino a 0 o a  $[n(n+1)]/2$  affinché  $H_0$  possa essere respinta.

**Teorema 8.6** *Se  $\{Y_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \{Z_i\}_{i=1}^n$  sono definiti come in precedenza, la distribuzione di probabilità di  $W$ , quando  $H_0 : \tilde{\mu} = \tilde{\mu}_0$  è vera, è data da*

$$P(W = w) = f_W(w) = \left(\frac{1}{2^n}\right) \cdot c(w)$$



ove  $c(w)$  è il coefficiente di  $e^{wt}$  nello sviluppo di

$$\prod_{i=1}^n (1 + e^{it})$$

Si può dimostrare che la distribuzione della statistica normalizzata  $W'$ , definita da

$$W' = \frac{W - E(W)}{\sqrt{\text{var}(W)}}$$

converge, per  $n \rightarrow \infty$ , alla distribuzione normale  $\mathcal{N}(0, 1)$ . In pratica, per  $n > 12$  la distribuzione

$$W' = \frac{W - [n(n+1)]/4}{\sqrt{[n(n+1)(2n+1)]/24}}$$

può essere adeguatamente approssimata dalla distribuzione normale standard. Pertanto, per testare

$$\begin{aligned} H_0 : \tilde{\mu} &= \tilde{\mu}_0 \\ \text{contro} \\ H_1 : \tilde{\mu} &\neq \tilde{\mu}_0 \end{aligned}$$

ad esempio al livello  $\alpha = 0.05$  di significatività, si dovrebbe rifiutare  $H_0$  se  $w'$  è minore o uguale a  $-1.96$  oppure maggiore o uguale a  $1.96$ .

◆ **Esercizio 8.48** Dato il campione  $(-4.4, 4.0, 2.0, -4.8)$  estratto da una popolazione normale con varianza unitaria, verificare se la media della popolazione è minore di 0 al livello 0.05. Verificare, cioè, al livello 0.05 l'ipotesi nulla  $H_0 : \mu \leq 0$ , contro l'ipotesi alternativa  $H_1 : \mu > 0$ .

◆ **Esercizio 8.49** Dati i campioni  $(1.8, 2.9, 1.4, 1.1)$ ,  $(5.0, 8.6, 9.2)$ , estratti da popolazioni normali, verificare l'ipotesi di uguaglianza delle varianze al livello 0.05.

◆ **Esercizio 8.50** Un dado è lanciato 300 volte, ottenendo i seguenti risultati

evento	1	2	3	4	5	6
frequenza	43	49	56	45	66	41

Verificare al livello 0.05 l'ipotesi che il dado non sia truccato.

◆ **Esercizio 8.51** La quantità di ossigeno disciolta nell'acqua è utilizzata come una misura dell'inquinamento dell'acqua. Supponiamo che siano stati presi dei campioni in quattro posizioni diverse in un lago, ottenendo, per quanto riguarda la quantità di ossigeno disciolta, i seguenti risultati

posizione	quantità di ossigeno disciolta (%)					
A	7.8	6.4	8.2	6.9	6.7	8.3
B	6.7	6.8	7.1	6.9	7.3	7.4
C	7.2	7.4	6.9	6.4	6.5	6.7
D	6.0	7.4	6.5	6.9	7.2	6.8

Analizzare se i dati precedenti indicano una differenza significativa nella quantità media dell'ossigeno disciolto nelle quattro differenti posizioni.

## 8.6 Software numerico

Tra le numerose fonti di software statistico, segnaliamo, in particolare, la libreria **NAG**, e la libreria **IMSL** (International Mathematical and Statistical Libraries)<sup>24</sup>. Ambedue le librerie sono *general purpose*, cioè coprono praticamente tutti i settori dell'analisi numerica.

Tra i packages più specializzati nel calcolo statistico, con opportune interfacce per l'utente, segnaliamo i seguenti

- (a) **SPSS** (Statistical Package for the Social Sciences) (si veda [121]).
- (b) **BMDP** (Biomedical Computer Package) (si veda [50]). È un package in generale più sofisticato del precedente e più orientato alle applicazioni mediche.
- (c) **GENSTAT** (si veda [3]). È un package più matematico e più flessibile dei packages SPSS e BMDP. Varie strutture di dati e operazioni sono definite in un particolare linguaggio formale. GENSTAT è particolarmente utile per l'analisi della varianza, in quanto utilizza un metodo generale per fittare modelli ANOVA e può trattare effetti casuali e dati mancanti.
- (d) **GLIM** (General Linear Interactive Modelling) (si veda [8]). È un programma particolarmente interessante per fittare i dati con modelli lineari generali.
- (e) **SAS** (Statistical Analysis System) (si veda [133]). È di uso particolarmente semplice e implementato anche su personal computer.
- (f) **SIR** (Scientific Information Retrieval) (si veda [136]). È più precisamente un package per il trattamento di banche dati, in particolare di tipo gerarchico. Mediante il SIR si possono organizzare file di dati da utilizzare con SPSS o BMDP.
- (g) **MINITAB** (si veda [140]). È uno sviluppo di un semplice programma da usare in maniera interattiva per l'insegnamento della statistica.

Tra i periodici che riportano implementazioni di procedure statistiche segnaliamo, in particolare, *Applied Statistics* (per una raccolta di algoritmi pubblicati su tale rivista si veda Griffiths e Hill [71]). Infine, particolarmente interessanti risultano i prodotti **MATLAB** e **MATHEMATICA**<sup>25</sup>.

È opportuno sottolineare il contributo importante dato, a partire dagli anni '80, dai calcolatori allo sviluppo di nuovi metodi statistici mediante l'introduzione e l'implementazione di algoritmi di calcolo opportuni<sup>26</sup>.

<sup>24</sup>distribuite rispettivamente da Numerical Algorithms Group, Banbury Road, Oxford, England e da IMSL, Inc., Sixth Floor, NBC Building, 7500 Bellaire Boulevard, Houston, Texas, 77036.

<sup>25</sup>distribuiti rispettivamente da *The MathWorks, Inc. 21 Eliot St. South Natick, MA 01760* e da *Wolfram Research, Inc., Champaign, IL 61821, USA*.

<sup>26</sup>Per una discussione introduttiva di questo aspetto si veda ad esempio B. Efron, R. Tibshirani *Statistical Analysis in the Computer Age*, Science **253**, 1991.

## 8.7 Catene di Markov

Consideriamo un *sistema fisico stocastico* (o probabilistico), cioè caratterizzato da una legge di probabilità, quale ad esempio il moto browniano di una particella in un liquido o in un gas, o le emissioni da una sorgente radioattiva, e supponiamo di osservare in istanti successivi lo *stato* del sistema. Per il moto browniano lo *stato* del sistema ad un istante determinato (osservazione) è la posizione della particella a quell'istante; per l'emissione di radiazioni lo stato del sistema ad un tempo  $t$  è il numero totale di emissioni che si sono verificate nell'intervallo di tempo da 0 a  $t$ .

Indicati con  $E_1, E_2, \dots$  gli stati possibili del sistema, useremo la variabile  $X_i$  per rappresentare il risultato dell'osservazione  $i$ -ma del sistema. Pertanto,  $X_i$  può assumere come valore uno degli stati  $E_1, E_2, \dots$ . Dopo  $n$  osservazioni del sistema si ha un campione  $(X_1, \dots, X_n)$ . Ad esempio, per le emissioni radioattive gli stati possibili sono

$E_1$	nessuna emissione
$E_2$	una emissione
$E_3$	due emissioni
$\dots$	
$E_r$	$r-1$ emissioni
$\dots$	

Una questione importante nello studio di un sistema dinamico probabilistico riguarda la conoscenza di come la storia passata del sistema influisce la determinazione della probabilità degli eventi futuri. Ad esempio, nel caso dell'emissione di radiazioni è evidente che la conoscenza dello stato particolare osservato alla  $r-1$ -ma osservazione influisce sulla determinazione della probabilità di osservare uno stato  $E_i$  alla successiva osservazione  $r$ -ma. Infatti, se ad esempio alla quinta osservazione si è ottenuto lo stato  $E_3$  (totale di due emissioni), allora la probabilità di ottenere lo stato  $E_2$  alla successiva sesta osservazione è determinata ( $=0$ ).

Il problema precedente può essere posto nella forma seguente. Si tratta di calcolare la probabilità di osservare uno stato  $E_i$  alla osservazione  $k$ -ma, conoscendo gli stati particolari osservati in corrispondenza a ciascuna delle precedenti  $k-1$  osservazioni. Si tratta, quindi, di calcolare la seguente probabilità condizionata

$$P(X_k = E_i \mid X_1 = E_{j_1}, X_2 = E_{j_2}, \dots, X_{k-1} = E_{j_{k-1}})$$

ove  $E_{j_i}$ , per  $i = 1, 2, \dots, k-1$ , indica lo stato osservato alla osservazione  $i$ -ma.

Ricordiamo che nel caso particolare in cui i risultati delle osservazioni in un sistema sono indipendenti il valore della precedente probabilità condizionata è uguale alla probabilità non condizionata  $P(X_k = E_i)$ . Come esempio di un tale sistema, si consideri l'esperimento del lancio ripetuto di una moneta simmetrica. In questo caso gli stati possibili sono testa ( $E_1$ ) e croce ( $E_2$ ). La probabilità di osservare testa

al terzo tentativo, sapendo che i primi due tentativi hanno dato croce, è ancora semplicemente la probabilità di osservare testa al terzo esperimento, cioè

$$P(X_3 = E_1 \mid X_1 = E_2, X_2 = E_2) = P(X_3 = E_1) = \frac{1}{2}$$

In generale, tuttavia, i sistemi fisici mostrano una dipendenza, e la probabilità di trovarsi in uno stato particolare alla osservazione  $k$  dipende da alcuni o da tutti i  $k - 1$  stati osservati in precedenza. Quando tale dipendenza è *solo* dallo stato alla precedente osservazione  $k - 1$ , si dice che il modello probabilistico che descrive il sistema è una *catena di Markov*<sup>27</sup>. Allora, per una catena di Markov si ha

$$P(X_k = E_i \mid X_1 = E_{j_1}, \dots, X_{k-1} = E_{j_{k-1}}) = P(X_k = E_i \mid X_{k-1} = E_{j_{k-1}})$$

Quando questo tipo di sistema si trova in un determinato stato, i cambiamenti futuri nel sistema dipendono solo da tale stato e non dal modo con il quale il sistema è pervenuto a questo particolare stato.

► **Esempio 8.46** *Modello meteorologico.* Il tempo (meteorologico) di una determinata località può essere caratterizzato, ad esempio, dai tre stati: soleggiato (S), nuvoloso (N), e piovoso (P). Con  $X_k$  si indicano le osservazioni giornaliere. Introduciamo le seguenti ipotesi, che definiscono un particolare modello probabilistico. Se ad un determinato giorno il tempo è nello stato S, allora gli stati S e N sono ugualmente probabili per il giorno successivo. Se è nello stato N, allora vi è un 50% di probabilità che il giorno successivo sia nello stato S, un 25% di continuare nello stato N, e un 25% di trovarsi nello stato P. Infine, se è nello stato P, non sarà possibile al giorno successivo lo stato S, ma sono ugualmente probabili gli stati P e N. Il modello probabilistico precedente può essere riassunto nella seguente matrice, i cui elementi rappresentano la probabilità di passare da uno stato ad un altro.

	S	N	P
S	$\frac{1}{2}$	$\frac{1}{2}$	0
N	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
P	0	$\frac{1}{2}$	$\frac{1}{2}$

<sup>27</sup>Andrei Andreevich Markov, matematico russo (1856-1922), allievo di Chebichev. Iniziò lo studio dei processi stocastici, indicati nel seguito col suo nome, nel lavoro: *L'estensione della legge dei grandi numeri a variabili mutuamente indipendenti* (1906); per motivi, quindi, di carattere teorico. Successivamente, studiò l'applicazione della sua teoria alla distribuzioni di vocali e consonanti nell'*Eugenio Onegin* di Puskin. I processi di Markov sono stati utilizzati, nel seguito, per modellizzare diversi fenomeni in *fisica* (processi a cascata, trasformazioni radioattive, fissione nucleare, . . .), in *astronomia* (studio di fluttuazioni nella luminosità della Via Lattea, distribuzione spaziale delle galassie, . . .), in *chimica* (cinetica delle reazioni, teoria statistica delle catene polimeriche, . . .), in *biologia* (crescita di popolazioni, struttura delle popolazioni biologiche, embriogenesi, genetica molecolare, farmacologia, crescita tumorale e delle epidemie, . . .), e nelle *scienze sociali* (comportamento nelle votazioni, mobilità geografica entro una regione, competizione industriale, epidemiologia delle malattie mentali, . . .).

Il modello può essere alternativamente descritto in termini di un diagramma, come indicato in Figura 8.21. In tale diagramma, i nodi corrispondono agli stati e gli archi orientati tra i nodi indicano possibili transizioni, con la probabilità di una assegnata transizione indicata sul corrispondente arco.

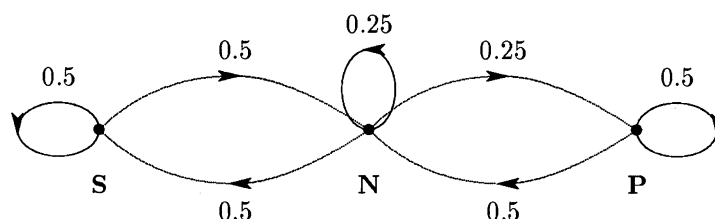


Figura 8.21: Diagramma relativo alla catena di Markov nell'Esempio 8.46.

Il processo partendo da uno stato iniziale evolve, ogni giorno, a una nuova condizione. A differenza, tuttavia, di quello che avviene per i sistemi dinamici di tipo *deterministico* (descritti, ad esempio, da equazioni alle differenze o da equazioni differenziali), nel modello probabilistico non è possibile predire esattamente quale sarà lo stato successivo. Si hanno solamente affermazioni di tipo probabilistico, basate, ad esempio, su inferenze di tipo statistico.

► **Esempio 8.47** Consideriamo due urne, denotate rispettivamente con A e B. L'urna A contiene 10 palline di colore rosso e 10 di colore nero. L'urna B contiene 3 palline rosse e 9 nere. Il sistema incomincia con l'urna rossa da cui viene estratta una pallina, viene annotato il colore, e successivamente la pallina viene rimessa nell'urna. Se la pallina estratta è rossa, la seconda pallina è estratta dall'urna A, altrimenti, se la pallina è nera, la seconda estrazione avviene dall'urna B. Il processo è ripetuto con scelta dell'urna per l'estrazione determinata dal colore della pallina ottenuta nell'estrazione precedente. I due possibili stati in questo caso sono “pallina rossa” ( $E_1$ ), “pallina nera” ( $E_2$ ). Nell'ipotesi di estrazioni aleatorie (cioè che le palline abbiano la stessa probabilità di essere estratte), la probabilità che, ad esempio, nella quinta estrazione si ottenga una pallina rossa, data l'informazione che i risultati delle precedenti estrazioni siano  $(E_2, E_2, E_1, E_2)$ , è semplicemente la probabilità di ottenere una pallina rossa alla quinta estrazione, noto che la quarta estrazione ha prodotto una pallina nera; cioè

$$P(X_5 = E_1 \mid X_1 = E_2, X_2 = E_2, X_3 = E_1, X_4 = E_2) = P(X_5 = E_1 \mid X_4 = E_2) = \frac{3}{12} = \frac{1}{4}$$

Osserviamo che

$$P(X_5 = E_1 \mid X_1 = E_1, X_2 = E_1, X_3 = E_1, X_4 = E_2) = P(X_5 = E_1 \mid X_4 = E_2) = \frac{1}{4}$$

mentre

$$P(X_5 = E_1 \mid X_1 = E_i, X_2 = E_j, X_3 = E_k, X_4 = E_1) = P(X_5 = E_1 \mid X_4 = E_1) = \frac{10}{20} = \frac{1}{2}$$

per ogni scelta di stati  $E_i, E_j, E_k$ .

### 8.7.1 Concetti di base

**Definizione 8.11** Lo spazio degli stati  $\mathcal{S}$  di una catena di Markov è l'insieme di tutti gli stati del sistema.

Nell'Esempio 8.46 lo spazio degli stati è  $\mathcal{S} = \{S, N, P\} = \{E_1, E_2, E_3\}$ , e nell'Esempio 8.47  $\mathcal{S} = \{\text{pallina rossa, pallina nera}\} = \{E_1, E_2\}$ . Gli stati di una catena di Markov sono supposti vicendevolmente esclusivi, nel senso che non si possono avere due stati differenti al medesimo tempo.

**Definizione 8.12** Una catena di Markov viene chiamata una catena di Markov finita se lo spazio degli stati della catena è finito.

Più in generale, si possono considerare catene di Markov applicate a sistemi per i quali gli stati costituiscono un insieme infinito numerabile:  $\{E_n\}$ ,  $n = 1, 2, \dots$ . Per il seguito, al fine di alleggerire le notazioni, anziché dire che l'osservazione  $i$ -ma ha come risultato lo stato  $E_k$ , si dirà semplicemente che il sistema al tempo  $i$  è nello stato  $E_k$ , o più semplicemente che si trova nello stato  $k$ . Inoltre, per uniformarsi alle notazioni più usuali, indicheremo con  $X_0$  l'osservazione iniziale, e quindi un campione generico sarà della forma  $(X_0, X_1, \dots, X_n)$ .

**Definizione 8.13** La funzione di probabilità di transizione a un passo (*one-step transition probability function*) per una catena di Markov è una funzione che assegna la probabilità di passare da uno stato  $j$  allo stato  $k$  in un passo (un intervallo di tempo) per ogni  $j$  e  $k$ .

La funzione di probabilità di transizione a un passo sarà indicata con<sup>28</sup>

$$p_{jk} = P(E_k | E_j) = P(k | j) \quad \text{per ogni } j \text{ e } k \quad (8.62)$$

Le probabilità di transizione possono essere considerati come gli elementi di una matrice definita nel modo seguente

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

<sup>28</sup>La forma più generale della funzione di transizione a un passo è la seguente

$$p_{jk}(n-1, n) = P(X_n = E_k | X_{n-1} = E_j)$$

che assegna la probabilità di essere nello stato  $k$  al tempo  $n$ , se il sistema era nello stato  $j$  al tempo  $n-1$ . Questa funzione è dipendente dal tempo, mentre la funzione data dall'equazione (8.62) è indipendente dal tempo. Per questo motivo, quest'ultima viene detta funzione di transizione omogenea. Nel seguito si considereranno solo funzioni di transizione omogenee.

Nell'Esempio 8.47 si ha

$$\begin{aligned} p_{11} &= P(E_1 | E_1) = \frac{1}{2}, & p_{12} &= P(E_2 | E_1) = \frac{1}{2} \\ p_{21} &= P(E_1 | E_2) = \frac{1}{4}, & p_{22} &= P(E_2 | E_2) = \frac{3}{4} \end{aligned}$$

e quindi

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

Dal momento che gli elementi  $p_{ij}$  rappresentano i valori di probabilità, si ha che  $p_{ij} \geq 0$ . Osserviamo, inoltre, che nella generica riga  $i$ -ma si trova la condizione che al tempo precedente (osservazione precedente) si è osservato lo stato  $i$ . Si tratta, quindi, della probabilità condizionata  $P(\cdot | E_i)$  sullo spazio degli stati  $\mathcal{S}$ . Poiché il sistema deve trovarsi in almeno uno degli stati, si avrà, pertanto

$$p_{i1} + p_{i2} + \dots = \sum_j p_{ij} = \sum_j P(E_j | E_i) = P\left(\sum_j E_j | E_i\right) = 1$$

Una matrice con le proprietà ora evidenziate viene chiamata *matrice stocastica*<sup>29</sup>.

**Definizione 8.14** *La funzione di probabilità iniziale è una funzione che assegna la probabilità che il sistema sia inizialmente (al tempo zero) nello stato  $i$ , per ogni  $i$ .*

La funzione di probabilità iniziale sarà indicata con  $p_i^{(0)} = P(X_0 = E_i)$ , per ogni  $i$ . Le  $p_i^{(0)}$  rappresentano le componenti del vettore iniziale  $\mathbf{p}^{(0)} = [p_1^{(0)}, p_2^{(0)}, \dots]$ . Come esemplificazione, consideriamo l'Esempio 8.47. Avendo scelto inizialmente l'urna A, si ha  $\mathbf{p}^{(0)} = [\frac{1}{2}, \frac{1}{2}]$ . Osserviamo che nel seguito di questo paragrafo, per evitare notazioni inutilmente pesanti e per seguire una consuetudine, i vettori considerati saranno intesi come vettori *riga*.

### Esempi di cammini aleatori

Consideriamo una particella che si muove su una retta con passi unitari. La probabilità di un passo a destra è data da  $p$  e di un passo a sinistra da  $q$ , con  $p + q = 1$ . Tale tipo di sistema è chiamato *cammino aleatorio* unidimensionale<sup>30</sup>.

<sup>29</sup>Osserviamo (cfr. Teorema di Perron-Frobenius in Appendice A) che se  $\mathbf{P}$  è una matrice stocastica, allora il valore  $\lambda_0 = 1$  è un autovalore e nessun altro autovalore della matrice ha modulo maggiore di 1. Inoltre, se  $\mathbf{x}$  è un *vettore riga di probabilità* (tale cioè da avere tutte le componenti non negative e con somma 1), allora il vettore riga  $\mathbf{xP}$  è ancora un vettore di probabilità. Pertanto, le matrici stocastiche possono essere pensate come le trasformazioni naturali nell'insieme dei vettori di probabilità.

<sup>30</sup>Più in generale, le probabilità  $p$  e  $q$  possono dipendere dallo stato  $i$  in cui si trova la particella, ed inoltre la probabilità  $r_i$  che la particella rimanga in  $i$  può essere diversa dallo zero. Si hanno,

Consideriamo come esempio un cammino a sei stati, illustrato in Figura 8.22. Si possono avere varianti del modello in corrispondenza a differenti comportamenti dei punti estremi (barriere), che nell'esempio sono gli stati 1 e 6.

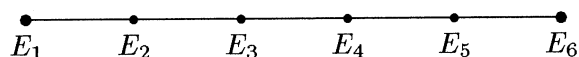


Figura 8.22: Spazio degli stati di cammini aleatori.

► **Esempio 8.48** *Cammini aleatori con barriere parzialmente riflettenti.* Se la particella si trova nello stato 1 essa si muove a destra con probabilità  $p$  e rimane nello stato 1 con probabilità  $q$ . Analogamente, se la particella è nello stato 6 si muove a sinistra con probabilità  $q$  e rimane nello stato 6 con probabilità  $p$ . La matrice di transizione  $\mathbf{P}$  è allora data da

$$\mathbf{P} = \begin{bmatrix} q & p & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & p & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & q & p \end{bmatrix}$$

► **Esempio 8.49** *Cammino aleatorio con barriere assorbenti.* Gli stati limiti sono assorbenti, nel senso che una volta che una particella raggiunge uno stato limite non può lasciarlo; quindi  $p_{11} = 1$  e  $p_{66} = 1$ . In questo caso si ha

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & p & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

► **Esempio 8.50** *Modello di diffusione di Ehrenfest* (meccanica statistica). Consideriamo un sistema fisico nel quale  $k$  molecole sono distribuite tra due contenitori  $A$  e  $B$ . I due contenitori sono separati da una membrana permeabile che permette un passaggio libero delle molecole tra i due contenitori. Il modello probabilistico è basato sulla seguente assunzione. Ad ogni istante del tempo  $t$  una delle  $k$  molecole è scelta a caso (numerando, ad esempio, da 1 a  $k$  le molecole e scegliendo a caso un numero tra 1 e  $k$ ); se essa si trova in  $A$  viene trasferita in  $B$ , se si trova in  $B$  viene trasferita in  $A$ . Lo stato del sistema è determinato dal numero delle molecole in  $A$  e in  $B$ ; più precisamente, lo stato  $E_i$  indica lo stato del sistema quando vi sono  $i$  molecole in  $A$ . I possibili stati del sistema sono allora  $E_0, \dots, E_k$ . Nel modello di Ehrenfest, se  $A$  ha  $j$  molecole, la probabilità di ottenere nel successivo tentativo lo stato

quindi, per ogni  $i$  tre numeri  $p_i, q_i, r_i$ , con  $p_i + r_i + q_i = 1$ . La corrispondente catena di Markov è nota, per le sue applicazioni allo studio di popolazioni viventi, come *catena di morte e vita* (birth and death chain).



$E_{j-1}$  è data da  $j/k$  e  $(k-j)/k$  per avere lo stato  $E_{j+1}$ ; cioè  $p_{j,j-1} = j/k$  e  $p_{j,j+1} = (k-j)/k$ . Il modello è, pertanto, un caso particolare di cammino aleatorio (barriere completamente riflettenti) e la corrispondente matrice di transizione è data da

$$\begin{array}{c}
 E_0 \\
 E_1 \\
 E_2 \\
 E_3 \\
 \vdots \\
 E_{k-1} \\
 E_k
 \end{array}
 \begin{bmatrix}
 E_0 & E_1 & E_2 & E_3 & E_4 & \cdots & E_{k-1} & E_k \\
 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
 1/k & 0 & (k-1)/k & 0 & 0 & \cdots & 0 & 0 \\
 0 & 2/k & 0 & (k-2)/k & 0 & \cdots & 0 & 0 \\
 0 & 0 & 3/k & 0 & (k-3)/k & \cdots & 0 & 0 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1/k \\
 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0
 \end{bmatrix}$$

La matrice di transizione mostra la tendenza del sistema a spostarsi verso lo stato di equilibrio di  $1/2$  delle molecole in ogni contenitore.

► **Esempio 8.51** *Rovina del giocatore.* Supponiamo che un giocatore inizi con un certo capitale iniziale e faccia una serie di puntate unitarie. Assumiamo che egli abbia una probabilità  $p$  di vincere e  $q = 1 - p$  di perdere, e che se il suo capitale raggiunge il valore zero, egli sia “rovinato” e il suo capitale rimanga nullo nel seguito. Con  $X_n$ ,  $n \geq 0$ , si indica il capitale al tempo  $n$ . Il sistema è una catena di Markov nella quale 0 è uno stato assorbente. Si può modificare il modello in differenti maniere; ad esempio, se il giocatore lascia il gioco, quando il capitale raggiunge un valore  $d$ , si ha una catena di Markov finita nella quale 0 e  $d$  sono stati assorbenti.

► **Esempio 8.52** *Catena ramificata.* Consideriamo un gene composto da  $d$  sotto-unità, con  $d$  intero positivo. Ogni sotto-unità può essere o normale o mutante. Consideriamo una cellula con un gene composto da  $m$  sotto-unità mutanti e  $d - m$  sotto-unità normali. Prima che la cellula si dupli in due cellule figlie, il gene si duplica. Il corrispondente gene di una delle cellule figlie è composto da  $d$  unità scelte a caso dalle  $2m$  sotto-unità mutanti e le  $2(d - m)$  sotto-unità normali. Supponiamo di seguire una linea fissata di discesa da un gene assegnato. Sia  $E_0$  il numero delle sotto-unità mutanti inizialmente presenti e  $E_n$ ,  $n \geq 1$ , il numero presente nel gene discendente  $n$ -mo. Si ottiene una catena di Markov con  $\mathcal{S} = \{0, 1, 2, \dots, d\}$  e

$$p_{ij} = \frac{\binom{2i}{j} \binom{2d-2i}{d-j}}{\binom{2d}{d}}$$

Gli stati 0 e  $d$  sono stati assorbenti per tale catena. Si può, inoltre, dimostrare la seguente relazione

$$\sum_{j=0}^d j p_{ij} = i, \quad i = 0, \dots, d \quad (8.63)$$

A partire da tale proprietà<sup>31</sup>, si può far vedere che il valore medio (aspettato) di  $X_{n+1}$ , dati i valori  $X_0, \dots, X_n$ , è uguale al valore presente di  $X_n$ .

<sup>31</sup>Ricordiamo che una sequenza di variabili aleatorie (non necessariamente una catena di Mar-

### 8.7.2 Descrizione di un sistema mediante una catena di Markov

Come abbiamo visto, una catena di Markov è completamente determinata quando sono fissati lo *spazio degli stati*, la *probabilità iniziale* e la *matrice di transizione*. Pertanto, la prima questione che si pone quando si vuole rappresentare un sistema dinamico mediante una catena di Markov, riguarda la determinazione o la stima di tali caratteristiche. Esamineremo, ora, altre questioni che presentano interesse nel caso di sistemi descritti da una catena di Markov. In particolare, può avere interesse analizzare

1. la probabilità di passare da uno stato  $j$  a uno stato  $k$  in  $n$  passi;
2. la probabilità incondizionata che al tempo  $n$  (cioè, dopo  $n$  passi dalla prima osservazione) il sistema sia nello stato  $j$ ;
3. il tempo medio per raggiungere uno stato assorbente  $k$  (ove termina, quindi, la catena), quando la catena ha avuto inizio da uno stato particolare  $j$ ;
4. l'esistenza di un *stato stazionario*; in altre parole, l'esistenza di una funzione di probabilità  $\pi_j$  tale che  $\lim_{n \rightarrow \infty} P(X_n = E_j) = \pi_j$ .

Le quattro questioni precedenti devono trovare risposta nelle quantità  $\mathcal{S}$ ,  $\mathbf{p}^{(0)}$ , e  $\mathbf{P}$  che definiscono la catena di Markov.

#### Matrice di transizione a $n$ -passi

La *funzione di probabilità di transizione a  $n$ -passi* è la probabilità di passare da uno stato  $j$  a uno stato  $k$  in esattamente  $n$  passi, cioè

$$p_{jk}^{(n)} = P(X_{t+n} = E_k, X_t = E_j)$$

Naturalmente, un sistema può passare dallo stato  $E_j$  allo stato  $E_k$  in differenti modi. Ad esempio, se un sistema ha  $r$  stati possibili, allora in *due* passi si può andare da  $E_j$  a  $E_k$  nei seguenti modi

$$E_j \rightarrow E_1 \rightarrow E_k, \quad E_j \rightarrow E_2 \rightarrow E_k, \quad \dots \quad E_j \rightarrow E_r \rightarrow E_k$$

Per definizione di catena di Markov si ha

$$P(E_j \rightarrow E_i \text{ e } E_i \rightarrow E_k) = P(E_j \rightarrow E_i) P(E_i \rightarrow E_k) = p_{ji} p_{ik}$$

---

kov) con la proprietà (8.63) è detta una *martingala*. Tale nozione, introdotta inizialmente in corrispondenza a modelli di gioco, ricopre un ruolo importante nella teoria moderna della probabilità.

Pertanto

$$p_{jk}^{(2)} = p_{j1}p_{1k} + p_{j2}p_{2k} + \cdots + p_{jr}p_{rk}$$

In maniera *ricorsiva*, la probabilità  $p_{jk}^{(3)}$  può essere calcolata nel seguente modo

$$p_{jk}^{(3)} = p_{j1}p_{1k}^{(2)} + p_{j2}p_{2k}^{(2)} + \cdots + p_{jr}p_{rk}^{(2)}$$

Per induzione su  $n$ , si avrà  $p_{jk}^{(n+1)} = \sum_{\forall i \in \mathcal{S}} p_{ji}p_{ik}^{(n)}$ . Più in generale, ancora per induzione si possono ottenere le seguenti relazioni

$$p_{jk}^{(m+n)} = \sum_{\forall i \in \mathcal{S}} p_{ji}^{(m)} p_{ik}^{(n)}$$

dette *equazioni di Chapman-Kolmogorov*. Indicata con  $\mathbf{P}^{(n)}$  la matrice di componenti  $p_{jk}^{(n)}$ , si ha  $\mathbf{P}^{(n)} = \mathbf{P}^n$ , ove  $\mathbf{P}^n$  è la matrice di probabilità di transizione a un passo moltiplicata per se stessa  $n$  volte, e quindi le formule precedenti equivalgono ai seguenti risultati

$$\mathbf{P}^{n+1} = \mathbf{P}\mathbf{P}^n, \quad \mathbf{P}^{m+n} = \mathbf{P}^m\mathbf{P}^n$$

Nell'Esempio 8.47, si ha

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \Rightarrow \mathbf{P}^2 = \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix}$$

Allora, la probabilità di passare da una pallina rossa a una pallina rossa in *due* passi è  $p_{11}^{(2)} = \frac{3}{8}$ .

### Funzioni di probabilità incondizionata

La *probabilità incondizionata*  $p_k^{(n)} := P(X_n = E_k)$  che al tempo  $n$  il sistema si trovi nello stato  $k$  è data dalla seguente equazione

$$p_k^{(n)} = \sum_{\forall j} p_j^{(0)} p_{jk}^{(n)}$$

e può essere scritta in forma vettoriale:  $\mathbf{p}^{(n)} = [p_1^{(n)}, p_2^{(n)}, \dots]$ . Si ha allora

$$\mathbf{p}^{(n)} = \mathbf{p}^{(0)} \mathbf{P}^n$$

Riferendoci, come esemplificazione, all'Esempio 8.47, il seguente valore

$$p_1^{(2)} = \sum_{j=1}^2 p_j^{(0)} p_{j1}^{(2)} = \frac{3}{4} \frac{3}{8} + \frac{1}{4} \frac{5}{16} = \frac{23}{64}$$

fornisce la probabilità di essere nello stato 1 alla terza osservazione (al tempo 2).

### Classificazione degli stati

**Definizione 8.15** Uno stato  $k$  è detto accessibile da uno stato  $j$  se esiste un intero positivo  $n$  tale che  $p_{jk}^{(n)} > 0$ .

Nell'Esempio 8.48 lo stato 3 è accessibile dallo stato 1, poiché  $p_{13}^{(2)} = p^2 > 0$ , mentre non lo è nell'Esempio 8.49, in quanto lo stato 1 è assorbente.

La proprietà di accessibilità non è simmetrica. La corrispondente nozione simmetrica è chiamata comunicazione.

**Definizione 8.16** Si dice che due stati  $j$  e  $k$  comunicano se  $j$  è accessibile da  $k$  e  $k$  è accessibile da  $j$ .

Si può verificare che nell'Esempio 8.48 tutte le coppie comunicano. La proprietà di comunicazione divide gli stati di una catena di Markov in classi distinte di equivalenza.

**Definizione 8.17** Un insieme non vuoto  $\mathcal{C}$  di stati è detto chiuso se nessun stato fuori dell'insieme è accessibile da un qualunque stato entro l'insieme.

Nel caso di un insieme chiuso ridotto ad un singolo stato  $k$  si parla di *stato assorbente*, e una catena di Markov che ha una o più stati assorbenti è chiamata *catena di Markov assorbente*. Una volta che una catena di Markov entra in un insieme chiuso essa rimane nell'insieme.

**Definizione 8.18** Una catena di Markov è irriducibile se tutte le coppie di stati comunicano, ossia vi è una sola classe comunicante. Altrimenti, essa è riducibile.

Ad esempio, la catena di Markov discussa nell'Esempio 8.48 è irriducibile. L'interesse della considerazione degli insiemi chiusi è il seguente. Se una catena di Markov consiste di uno o più insiemi chiusi, allora questi insiemi sono sotto-catene di Markov che possono essere studiate indipendentemente, in particolare per quanto riguarda le loro corrispondenti proprietà limite.

**Definizione 8.19** Uno stato  $j$  è detto transiente, o non ricorrente, se la probabilità condizionata di ritornare a  $j$ , supposto che il sistema parta da  $j$ , è minore di uno, altrimenti è detto ricorrente.

Ad esempio, uno stato assorbente è necessariamente ricorrente. Si può dimostrare che uno stato  $j$  è transiente se e solo se

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty \quad \text{e} \quad \sum_{n=1}^{\infty} p_{kj}^{(n)} < \infty$$

per tutti gli stati nello spazio  $\mathcal{S}$ . Una catena di Markov è chiamata una *catena transiente* se tutti i suoi stati sono transienti e una *catena ricorrente* se tutti i suoi

stati sono ricorrenti. Si può dimostrare che una catena di Markov finita deve avere almeno uno stato ricorrente e quindi non può essere una catena transiente. Inoltre, se una catena di Markov finita ha uno stato transiente, essa è non irriducibile.

**Definizione 8.20** Il periodo di uno stato  $k$  in una catena di Markov finita è il massimo comune divisore dell'insieme degli interi positivi  $n$  per i quali  $p_{kk}^{(n)} > 0$ . Uno stato  $k$  è detto aperiodico se esso ha periodo 1.

Consideriamo, ad esempio, la seguente matrice di transizione di una catena di Markov a due stati (che si alternano nei successivi istanti)

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \Rightarrow \mathbf{P}^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{P}^4 = \mathbf{P}^{2m} \quad (8.64)$$

ove  $m$  è un intero positivo qualunque. Allora,  $p_{11}^{(n)} > 0$  per  $n = 2m$ ; il massimo comune divisore di tale insieme è 2. Lo stato 1 ha quindi periodo 2. Allo stesso modo si vede che anche lo stato 2 ha periodo 2. La precedente è quindi un esempio di catena di Markov periodica irriducibile<sup>32</sup>.

Una catena di Markov finita è chiamata una *catena aperiodica* se esiste uno stato  $k$  con periodo 1. Si può mostrare che una catena di Markov è aperiodica mostrando che esiste uno stato  $k$  per il quale  $p_{kk} = p_{kk}^{(1)} > 0$ . Si ha, pertanto, che se almeno uno degli elementi diagonali in  $\mathbf{P}$  è diverso di zero, allora la catena è aperiodica. Il viceversa non è, comunque, vero; cioè una catena può essere aperiodica, anche se tutti gli elementi diagonali di  $\mathbf{P}$  sono nulli. Si può anche mostrare che una catena di Markov è aperiodica mostrando che esiste un intero  $n$  tale che  $p_{jk}^{(n)} > 0$  per tutti i valori di  $j$  e  $k$ .

► **Esempio 8.53** Se

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{bmatrix}$$

la corrispondente catena di Markov è aperiodica, dal momento che  $p_{22} = 1/2 > 0$ .

Se

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \Rightarrow \mathbf{P}^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

la corrispondente catena di Markov è aperiodica, poiché  $p_{jk}^{(2)} > 0$  per tutti gli indici  $j$  e  $k$ .

<sup>32</sup>Ricordiamo che una catena di Markov è chiamata *regolare* quando esiste un  $m > 0$  tale che per la corrispondente matrice di transizione si ha  $\mathbf{P}^m > \mathbf{0}$ . Una catena di Markov regolare è necessariamente irriducibile, ma una catena di Markov irriducibile non è necessariamente regolare. La matrice (8.64) è un controesempio.

### Tempi di assorbimento per catene di Markov finite

Sia  $\mathbf{P}$  la matrice di transizione corrispondente a una catena di Markov finita e assorbente e che contiene, con l'insieme di stati assorbenti, un insieme di stati transienti  $\mathcal{T}$ . Allora, la matrice  $\mathbf{P}$  può essere scritta nella seguente forma

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{bmatrix}$$

ove  $\mathbf{I}$  è la matrice quadrata di transizione corrispondente agli stati assorbenti,  $\mathbf{Q} = [p_{ik}]$  è una matrice quadrata, ove  $j$  e  $k$  sono stati nell'insieme  $\mathcal{T}$ , e  $\mathbf{R}$  è la matrice rettangolare delle probabilità di transizione dagli stati transienti agli stati assorbenti. Dal momento che la catena di Markov è finita,  $\mathcal{T}$  non è un insieme comunicante chiuso (irriducibile) e, quindi, non determina una sotto-catena di Markov. Si ha, pertanto, che  $\mathbf{Q}$  non è una matrice stocastica.

► **Esempio 8.54** Consideriamo un cammino aleatorio con barriere assorbenti e a 5 stati. Si ha

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ 0 & q & 0 & p & 0 \\ 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Usando il cambiamento di indici  $1 \leftrightarrow 1'$ ,  $5 \leftrightarrow 2'$ ,  $2 \leftrightarrow 3'$ ,  $3 \leftrightarrow 4'$ , e  $4 \leftrightarrow 5'$ , si può riscrivere  $\mathbf{P}$  nella seguente forma

$$\mathbf{P}' = \begin{bmatrix} \boxed{\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}} & \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} \\ \begin{matrix} q & 0 \\ 0 & 0 \\ 0 & p \end{matrix} & \boxed{\begin{matrix} 0 & p & 0 \\ q & 0 & p \\ 0 & q & 0 \end{matrix}} \end{bmatrix} \Rightarrow \mathbf{R} = \begin{bmatrix} q & 0 \\ 0 & 0 \\ 0 & p \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0 & p & 0 \\ q & 0 & p \\ 0 & q & 0 \end{bmatrix}$$

Se  $\mathbf{p}^{(0)}$  è il vettore di probabilità iniziali, possiamo scrivere  $\mathbf{p}^{(0)} = [\mathbf{p}_A^{(0)}, \mathbf{p}_T^{(0)}]$ , ove  $\mathbf{p}_T^{(0)}$  è il vettore delle probabilità di essere inizialmente negli stati transienti. In generale, il vettore  $\mathbf{p}_A^{(0)}$  sarà il vettore nullo.

Supponendo che la catena di Markov venga osservata agli istanti successivi  $1, 2, 3, \dots$  indichiamo con  $T$  il *tempo di assorbimento*, ossia il numero di passi dall'insieme degli stati transienti all'insieme degli stati assorbenti. La variabile  $T$  può assumere i valori  $1, 2, 3, \dots$ . Si ha il seguente risultato, per la cui dimostrazione rinviamo ad esempio a Kemeny e Snell [96].

**Proposizione 8.11** *La probabilità di passare dall'insieme degli stati transienti a uno stato assorbente nel tempo  $n$  (esattamente  $n$  passi) è data da*

$$P(T = n) = \mathbf{p}_T^{(0)} \mathbf{Q}^{n-1} \mathbf{R} \mathbf{1}, \quad n = 1, 2, \dots \quad (8.65)$$

ove  $\mathbf{1}$  è un vettore colonna di componenti uguali a 1.

Il prodotto di matrici in (8.65) somma semplicemente, per ogni stato transiente  $j$ , la probabilità di partire in  $j$ , di stare nell'insieme degli stati transienti per  $n - 1$  intervalli di tempo e quindi di andare nell'insieme degli stati assorbenti.

► **Esempio 8.55** Continuando l'Esempio 8.54, con  $p = 1/4$  e  $q = 3/4$ , assumiamo come vettore iniziale  $\mathbf{p}^{(0)} = [0, 0, 1/4, 1/2, 1/4]$ , da cui  $\mathbf{p}_T^{(0)} = [1/4, 1/2, 1/4]$ . Applicando la formula (8.65), si ottiene

$$P(T = 1) = \mathbf{p}_T^{(0)} \mathbf{Q}^0 \mathbf{R} \mathbf{1} = \begin{bmatrix} 1/4 & 1/2 & 1/4 \end{bmatrix} \begin{bmatrix} 3/4 \\ 0 \\ 1/4 \end{bmatrix} = \frac{3}{16} + \frac{1}{16} = \frac{1}{4}$$

$$P(T = 2) = \mathbf{p}_T^{(0)} \mathbf{Q}^1 \mathbf{R} \mathbf{1} = \begin{bmatrix} 1/4 & 1/2 & 1/4 \end{bmatrix} \begin{bmatrix} 0 & 1/4 & 0 \\ 3/4 & 0 & 1/4 \\ 0 & 3/4 & 0 \end{bmatrix} \begin{bmatrix} 3/4 \\ 0 \\ 1/4 \end{bmatrix} = \frac{5}{16}$$

$$P(T = 3) = \mathbf{p}_T^{(0)} \mathbf{Q}^2 \mathbf{R} \mathbf{1} = \begin{bmatrix} 1/4 & 1/2 & 1/4 \end{bmatrix} \begin{bmatrix} 0 & 1/4 & 0 \\ 3/4 & 0 & 1/4 \\ 0 & 3/4 & 0 \end{bmatrix}^2 \begin{bmatrix} 3/4 \\ 0 \\ 1/4 \end{bmatrix} = \frac{5}{32}$$

Allo stesso modo si ottiene  $P(T = 4) = 15/128$ . Notiamo che  $\lim_{n \rightarrow \infty} P(T = n) = 0$ . ■

Dal momento che gli eventi  $T = i$  e  $T = j$ , per  $i \neq j$  non possono verificarsi simultaneamente e vi deve essere un passaggio dagli stati transienti agli stati assorbenti per almeno un valore di  $n$ , si ha  $\sum_{n=1}^{\infty} P(T = n) = 1$ .

Si definisce *tempo medio* (tempo atteso) di *assorbimento* da un insieme di stati transienti la seguente quantità

$$m_T = \sum_{n=1}^{\infty} n P(T = n)$$

Il tempo medio  $m_T$  può essere calcolato (cfr. ad esempio Kemeny e Snell [96]) nel seguente modo

$$m_T = \mathbf{p}_T^{(0)} (\mathbf{I} - \mathbf{Q})^{-1} \mathbf{1}$$

La matrice  $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$  è chiamata *matrice fondamentale di una catena di Markov assorbente*.

Analogamente, per il momento secondo  $m_T^{(2)} = \sum_{n=1}^{\infty} n^2 P(T = n)$  si ha

$$m_T^{(2)} = \mathbf{p}_T^{(0)} (2\mathbf{N} - \mathbf{I}) \mathbf{N} \mathbf{1}$$

e la *varianza del tempo di assorbimento*  $v_T = \sum_{n=1}^{\infty} (n - m_T)^2 P(T = n)$  può essere ottenuta ponendo  $v_T = m_T^{(2)} - (m_T)^2$ .

► **Esempio 8.56** Continuando l'Esempio 8.54, con  $p = 1/4$  e  $q = 3/4$  e vettore iniziale  $\mathbf{p}^{(0)} = [0, 0, 1/4, 1/2, 1/4]$ , da cui  $\mathbf{p}_T^{(0)} = [1/4, 1/2, 1/4]$ , si ha

$$\mathbf{I} - \mathbf{Q} = \begin{bmatrix} 1 & -\frac{1}{4} & 0 \\ -\frac{3}{4} & 1 & -\frac{1}{4} \\ 0 & -\frac{3}{4} & 1 \end{bmatrix} \Rightarrow \mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} \frac{13}{10} & \frac{2}{5} & \frac{1}{10} \\ \frac{6}{5} & \frac{8}{5} & \frac{2}{5} \\ \frac{9}{10} & \frac{6}{5} & \frac{13}{10} \end{bmatrix}$$

Pertanto

$$m_T = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{18}{10} \\ \frac{16}{5} \\ \frac{34}{10} \end{bmatrix} = 2.9$$

In modo analogo, si trovano i valori  $m_T^{(2)} = 12.66$  e  $v_T = 4.25$ . ■

I valori  $m_T$  e  $v_T$  danno la media e la varianza del tempo di assorbimento dall'insieme degli stati transienti all'insieme degli stati assorbenti. Si vede facilmente che le componenti  $m_j$  della matrice  $\mathbf{M} = \mathbf{N}\mathbf{1}$  sono i tempi (numero di passi) medi di assorbimento a partire da uno stato transiente  $j$  a uno stato assorbente. In modo analogo si calcolano i momenti secondi e le varianze.

### Distribuzioni limite per catene di Markov finite

Per il seguito di questo paragrafo supporremo che la catena di Markov finita di cui discuteremo abbia  $r$  stati. Se  $k$  è uno stato transiente, allora per ogni stato  $j$  si ha  $\lim_{n \rightarrow \infty} p_{jk}^{(n)} = 0$ , in quanto per uno stato transiente  $k$  si ha  $\sum_{n=1}^{\infty} p_{jk}^{(n)} < \infty$  per ogni stato  $j$ . Esamineremo ora il caso in cui  $k$  non sia uno stato transiente.

**Definizione 8.21** Una catena di Markov finita è ergodica se esistono delle probabilità  $\pi_j$  tali che

$$\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j \quad \text{per ogni stato } i \text{ e } j \quad (8.66)$$

I valori  $\pi_j$  rappresentano le probabilità di essere in uno stato dopo che è stato raggiunto l'equilibrio. Osserviamo che esse sono indipendenti dallo stato iniziale  $i$ . In effetti si può dimostrare il seguente risultato.

**Proposizione 8.12** Se  $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j$ , allora  $\lim_{n \rightarrow \infty} p_j^{(n)} = \pi_j$ .

Le probabilità limite  $\pi_j$  possono essere calcolate risolvendo il seguente sistema di equazioni<sup>33</sup>

$$\pi_j = \sum_{k=1}^r \pi_k p_{kj} \quad \text{per } j = 1, 2, \dots, r \quad (8.67)$$

<sup>33</sup>Osserviamo che la matrice dei coefficienti del sistema omogeneo (8.67) ha determinante nullo, in quanto la somma degli elementi di ogni colonna è nulla.



soggetta alle condizioni

$$\pi_j \geq 0, \forall j; \quad \sum_{j=1}^r \pi_j = 1 \quad (8.68)$$

La distribuzione di probabilità  $\{\pi_k\}$  definita dalle equazioni (8.67) e (8.68) è chiamata una *distribuzione stazionaria*. Se una catena di Markov è ergodica, allora si può dimostrare (cfr. ad esempio Parzen [129]) che essa possiede un'unica distribuzione stazionaria. Osserviamo, comunque, che vi sono catene di Markov che possiedono delle distribuzioni che verificano le equazioni (8.67) e (8.68), ma che non sono ergodiche. Come esempio, si consideri la matrice di transizione

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \Rightarrow p_{11}^{(n)} = \begin{cases} 1 & \text{se } n \text{ è pari} \\ 0 & \text{se } n \text{ è dispari} \end{cases}$$

e, quindi, la catena è non ergodica. Tuttavia, la risoluzione delle equazioni (8.67) e (8.68) fornisce le probabilità stazionarie  $\pi_1 = \pi_2 = \frac{1}{2}$ . Sottolineiamo che la precedente matrice  $\mathbf{P}$  è la matrice di transizione per una catena di Markov irriducibile periodica.

I seguenti teoremi, per la cui dimostrazione si veda ad esempio Parzen [129], forniscono delle *condizioni sufficienti* affinché una catena di Markov finita sia ergodica.

**Teorema 8.7** *Una catena di Markov finita irriducibile aperiodica è ergodica. Quindi una catena di Markov regolare è ergodica.*

Si può inoltre mostrare che ogni riga della matrice  $\mathbf{P}^n$  tende alla distribuzione stazionaria.

► **Esempio 8.57** Esaminiamo il modello meteorologico considerato nell'Esempio 8.46, definito dalla matrice di transizione

$$\mathbf{P} = \begin{bmatrix} 0.500 & 0.500 & 0. \\ 0.500 & 0.250 & 0.250 \\ 0. & 0.500 & 0.500 \end{bmatrix} \Rightarrow \mathbf{P}^2 = \begin{bmatrix} 0.500 & 0.375 & 0.125 \\ 0.375 & 0.438 & 0.187 \\ 0.250 & 0.375 & 0.375 \end{bmatrix}$$

La corrispondente catena di Markov è irriducibile e aperiodica e quindi ergodica. Per trovare le probabilità limite, risolviamo le equazioni (8.67)

$$\begin{aligned} \pi_1 &= \frac{1}{2}\pi_1 + \frac{1}{2}\pi_2 + 0\pi_3 \\ \pi_2 &= \frac{1}{2}\pi_1 + \frac{1}{4}\pi_2 + \frac{1}{2}\pi_3 \\ \pi_3 &= 0\pi_1 + \frac{1}{4}\pi_2 + \frac{1}{2}\pi_3 \end{aligned}$$

con le condizioni (8.68). Si trovano, facilmente, le seguenti soluzioni

$$\pi_1 = \pi_2 = 0.400, \quad \pi_3 = 0.200$$

Allora, la probabilità asintotica di essere negli stati 1 e 2 è 0.4, mentre quella di essere nello stato 3 è 0.2. Ricordando il significato del modello e degli stati, si ha che asintoticamente ci si può aspettare 40% di giorni soleggiati, 40% di giorni nuvolosi, e 20% di giorni piovosi.

► **Esempio 8.58** *Rinnovo di apparecchiature.* Una apparecchiatura, la cui vita fisica non può superare, ad esempio, le 4 settimane, viene controllata ogni settimana; se trovata guasta, essa viene sostituita con una apparecchiatura nuova dello stesso tipo. Il sistema in evoluzione è l'età dell'apparecchiatura (misurata in settimane); gli stati possibili sono 0 (apparecchiatura nuova) 1, 2, 3, 4. Indicata con  $p_i$  la probabilità che l'apparecchiatura di età  $i$  superi in vita una settimana, la matrice di transizione è la seguente

$$\mathbf{P} = \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \\ \left[ \begin{array}{ccccc} q_0 & p_0 & 0 & 0 & 0 \\ q_1 & 0 & p_1 & 0 & 0 \\ q_2 & 0 & 0 & p_2 & 0 \\ q_3 & 0 & 0 & 0 & p_3 \\ 1 & 0 & 0 & 0 & 0 \end{array} \right] \end{array}$$

Si verifica facilmente che la corrispondente catena di Markov è ergodica. Le equazioni (8.67) sono

$$\begin{aligned} \pi_0 &= q_0\pi_0 + q_1\pi_1 + q_2\pi_2 + q_3\pi_3 + \pi_4 \\ \pi_1 &= p_0\pi_0 \\ \pi_2 &= p_1\pi_1 \\ \pi_3 &= p_2\pi_2 \\ \pi_4 &= p_3\pi_3 \end{aligned}$$

Dalle ultime quattro equazioni e dalle (8.68) si ricava

$$\pi_0 = \frac{1}{1 + p_0 + p_0p_1 + p_0p_1p_2 + p_0p_1p_2p_3}, \quad \pi_{i+1} = p_0p_1 \cdots p_i\pi_0 \quad i = 0, 1, 2, 3$$

La probabilità  $\pi_0$  è la probabilità che, in condizioni di regime, l'apparecchiatura installata sia nuova. Se  $N$  sono le apparecchiature installate,  $N\pi_0$  è, a regime, il numero medio di rinnovi che si devono effettuare ogni settimana.

► **Esempio 8.59** Per il modello di diffusione di Ehrenfest considerato nell'Esempio 8.50, si può dimostrare (cfr. Feller [58]) che se il numero di molecole  $k$  distribuite tra i due contenitori A e B è elevato, allora la probabilità stazionaria  $\pi_{k/2}$  per lo stato  $E_{k/2}$ , è approssimativamente uno. Più precisamente, si può dimostrare che  $\lim_{k \rightarrow \infty} \pi_{k/2} = 1$ . In effetti, se  $k = 10^6$ , allora la probabilità di trovare più di 505 000 molecole in A è approssimativamente  $10^{-23}$  (cfr. Feller [58]). ■

**Definizione 8.22** Una matrice di transizione è detta doppiamente stocastica (o anche bistocastica) se la somma degli elementi di ogni colonna è uguale a 1, cioè se  $\sum_{i=1}^r p_{ij} = 1$  per ogni  $j$ .

Per le catene di Markov corrispondenti a tali matrici si ha il seguente risultato.

**Teorema 8.8** Se la matrice di transizione  $\mathbf{P}$  corrispondente a una catena di Markov finita irriducibile aperiodica con  $r$  stati è doppiamente stocastica, allora le probabilità stazionarie sono date da  $\pi_k = 1/r$ , per  $k = 1, \dots, r$ .

Come esemplificazione, si consideri la seguente matrice

$$\mathbf{P} = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ \frac{3}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

che corrisponde a una catena di Markov irriducibile, aperiodica e doppiamente stocastica.

### 8.7.3 Catene di Markov infinite

In alcune applicazioni è naturale la considerazione di modelli basati su catene di Markov con un numero infinito (numerabile) di stati  $E_1, E_2, \dots$ . Una catena di Markov infinita ha spesso una maggiore simmetria, e porta a formule più semplici che non le corrispondenti catene finite ottenute, in generale, imponendo condizioni ai limiti artificiali.

► **Esempio 8.60** *Cammino aleatorio infinito con barriera riflettente.* Un oggetto si muove su una linea con passi unitari discreti; la posizione  $i$  dell'oggetto corrisponde agli interi  $0, 1, 2, \dots$ . Se l'oggetto è nella posizione  $i > 0$ , vi è una probabilità  $p$  che nel successivo passaggio si trovi in  $i + 1$ , una probabilità  $q$  che si trovi in  $i - 1$  e una probabilità  $r$  che esso rimanga in  $i$ . Se è nella posizione 0, allora si muoverà alla posizione 1 con probabilità  $p$  e rimarrà in 0 con probabilità  $1 - p$ . In corrispondenza, si ha la seguente matrice di transizione (infinita)

$$\mathbf{P} = \begin{bmatrix} 1-p & p & 0 & 0 & \dots \\ q & r & p & 0 & \dots \\ 0 & q & r & p & \dots \\ 0 & 0 & q & r & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

Buona parte della teoria sviluppata in precedenza nel caso di una catena finita può essere estesa al caso infinito. In particolare, si estendono immediatamente i concetti di accessibilità, di classi comunicanti, di irriducibilità e di periodicità. Di particolare importanza è la seguente estensione del teorema limite.

**Teorema 8.9** *Per una catena di Markov irriducibile, aperiodica i limiti*

$$v_j = \lim_{m \rightarrow \infty} p_{ij}^{(m)}$$

*esistono e sono indipendenti dallo stato iniziale  $i$ .*

Osserviamo che, a differenza del caso finito, i limiti  $v_j$  possono essere nulli. Nell'Esempio 8.60, se  $p > 0$ ,  $q > 0$ , è chiaro che ogni stato comunica con ogni altro, in

quanto vi è un cammino di probabilità di transizione diverse dallo zero da un qualunque stato ad un qualunque altro stato. Allora, la catena è irriducibile. La catena è aperiodica, poiché si può ritornare a uno stato restando a 0 indefinitamente. Si applica pertanto il Teorema 8.9. Comunque, se  $p > q$ , i limiti  $v_j$  sono tutti nulli, in quanto il processo si sposta continuamente a destra e ogni stato ha carattere di stato transiente.

**Definizione 8.23** Una matrice di Markov irriducibile e aperiodica è detta positiva ricorrente se

- (a)  $v_j = \lim_{m \rightarrow \infty} p_{ij}^{(m)} > 0$  per tutti i valori di  $j$ ;  
 (b)  $\sum_j v_j = 1$

Si ha allora il seguente risultato.

**Teorema 8.10** Data una catena di Markov irriducibile e aperiodica,

- (a) essa è positiva ricorrente se e solo se esiste un'unica distribuzione di probabilità  $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots]$ , con  $\pi_i > 0$ , per ogni  $i$  e  $\sum_i \pi_i = 1$ , che è soluzione delle seguenti equazioni

$$\pi_j = \sum_i p_{ij} \pi_i \quad (8.69)$$

In questo caso,

$$\pi_j = v_j = \lim_{m \rightarrow \infty} p_{ij}^{(m)} > 0$$

per tutti i valori di  $j$ .

- (b) Se la catena non è positiva ricorrente, allora

$$v_j = \lim_{m \rightarrow \infty} p_{ij}^{(m)} = 0$$

per tutti i valori di  $j$ .

► **Esempio 8.61** Continuando l'Esempio 8.60, supponiamo  $p > 0, r > 0$ , e  $q > 0$ . Cerchiamo, allora, una soluzione delle equazioni (8.69). Si ha

$$\begin{aligned} pv_0 - qv_1 &= 0 \\ (1-r)v_j - pv_{j-1} - qv_{j+1} &= 0, \quad j = 1, 2, \dots \end{aligned}$$

La successione  $\{v_j\}$  è, quindi, una soluzione di un'equazione alle differenze lineari a coefficienti costanti. La corrispondente equazione caratteristica è data da  $-q\lambda^2 + (1-r)\lambda - p = 0$ , che ha le radici  $\lambda = 1, p/q$ . Se  $p < q$ , una soluzione è data da

$$v_j = \left(1 - \frac{p}{q}\right) \left(\frac{p}{q}\right)^j$$

che verifica la condizione  $\sum_{j=1}^{\infty} v_j = 1$ . Pertanto, per  $p < q$ , la catena è positiva ricorrente. Se  $p > q$ , non esiste soluzione e la catena non è positiva ricorrente. ■

◆ **Esercizio 8.52** Considerare la catena di Markov a due stati corrispondente alla seguente matrice di transizione

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

ove  $p + q > 0$ . Trovare  $\mathbf{P}^n$ .

◆ **Esercizio 8.53** Considerare una catena di Markov a tre stati  $\{0, 1, 2\}$  corrispondente alla seguente matrice di transizione

$$\mathbf{P} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} & \begin{bmatrix} .4 & .4 & .2 \\ .3 & .4 & .3 \\ .2 & .4 & .4 \end{bmatrix} \end{matrix}$$

Mostrare che tale catena ha un'unica distribuzione stazionaria  $\pi$  e calcolare  $\pi$ .

◆ **Esercizio 8.54** Un topolino è messo nel labirinto illustrato in Figura 8.23. Il topolino si muove, in maniera aleatoria, tra i nove compartimenti attraverso le connessioni indicate. Supponiamo, più precisamente che se vi sono  $k$  modi per lasciare un determinato compartimento, esso sceglie ognuno di esse con uguale probabilità. Rappresentare i cammini del topolino mediante una catena di Markov e costruire la corrispondente matrice di transizione. Esaminare se la catena è ergodica, e trovare eventualmente la probabilità stazionaria.

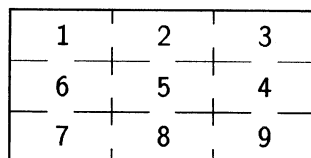


Figura 8.23: Esempio di labirinto.

◆ **Esercizio 8.55** Sia  $\pi$  una distribuzione stazionaria di una catena di Markov. Supponiamo che  $j$  e  $k$  siano due stati tali che  $p_{ij} = cp_{ik}$ , ove  $i$  è un qualunque stato in  $\mathcal{S}$  e  $c$  una costante. Mostrare che  $\pi_j = c\pi_k$ .

◆ **Esercizio 8.56** Mostrare che se uno stato in una classe comunicante è periodico, allora tutti gli stati nella classe sono periodici.

◆ **Esercizio 8.57** Esaminare se le matrici  $\mathbf{A}^2$  e  $\mathbf{AB}$  sono bistocastiche, quando  $\mathbf{A}$ ,  $\mathbf{B}$  sono supposte bistocastiche.

◆ **Esercizio 8.58** Esaminare i cammini aleatori infiniti corrispondenti ai seguenti due casi particolari

$$(a) r = 0, p > q > 0; \quad (b) r > 0, p = q > 0$$

## 8.8 Introduzione alla teoria dell'Informazione

In questo paragrafo daremo alcuni elementi essenziali della *teoria dell'Informazione* (o teoria della codifica). Tale teoria ha avuto inizio, sostanzialmente, con i lavori di C. Shannon (1948) e di Wiener (1949)<sup>34</sup>. Per gli sviluppi della teoria e le applicazioni si vedano, ad esempio, Ash [7], Hamming [78]. La teoria dell'Informazione concerne l'analisi di una "entità", chiamata *sistema di comunicazione*, rappresentata in maniera schematica in Figura 8.24. La *sorgente dei messaggi* è la persona o la macchina che produce l'informazione. Il *codificatore* (encoder) associa ad ogni messaggio un "oggetto" conveniente per la trasmissione sul canale (ad esempio, una successione di cifre binarie (bits), o un'onda continua). Il *canale* è il mezzo sul quale è trasmesso il messaggio codificato. Il *decodificatore* opera sull'output del canale e cerca di estrarre il messaggio originale da mandare alla *destinazione*. In generale, questo non può essere fatto con completa affidabilità, a causa dell'effetto di *rumori* (noise), un termine che indica tutto ciò che tende a produrre errori nella trasmissione. La teoria dell'Informazione ha come scopo la costruzione di modelli matematici per ognuno dei blocchi in Figura 8.24.

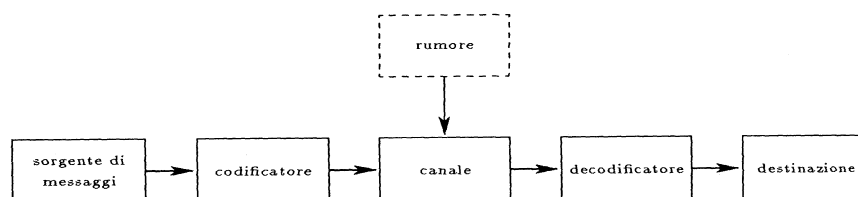


Figura 8.24: Sistema di comunicazione.

Uno dei risultati più significativi è il cosiddetto *teorema fondamentale della teoria dell'informazione*, che, in maniera schematica, afferma che è *possibile trasmettere informazione attraverso un canale disturbato (noisy channel) ad una qualsiasi velocità minore della capacità del canale con una probabilità di errore arbitrariamente piccola*. In sostanza, si può raggiungere una *affidabilità* arbitrariamente elevata, a spese della diminuzione della effettiva velocità di trasmissione (in pratica la riduzione della velocità a un numero, detto *capacità del canale*). Questo può essere ottenuto mediante un'opportuna *codifica*. Il codificatore assegna a ciascuno dei messaggi di un gruppo specificato una sequenza di simboli, chiamata *codice di parola*, appropriata per la trasmissione lungo il canale. In generale, per ottenere affidabilità, senza

<sup>34</sup>C. Shannon *A mathematical theory of communication*, Bell System Tech. J., **23**, 379–423, 623–656 (1948); N. Wiener *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley, New York, 1949.

sacrificare la velocità di trasmissione, i codici di parola sono assegnati a lunghi blocchi di digits, anziché ai singoli digits; in altre parole, il codificatore aspetta che la sorgente di segnali produca un blocco di digits di una lunghezza specificata, e allora assegna un codice di parola all'intero blocco.

I concetti e i risultati precedenti possono essere formalizzati, dal punto di vista matematico, introducendo una *misura matematica dell'informazione contenuta in un messaggio*, o, equivalentemente, una misura dell'*incertezza*<sup>35</sup>. Nel seguito esamineremo, in particolare, la misura proposta da Shannon, che è importante per il suo significato operativo nella costruzione di codici. Per una discussione e una dimostrazione del teorema fondamentale della teoria dell'informazione si veda, ad esempio, Ash [7].

### 8.8.1 Quantità di Informazione

Siano  $X_1, \dots, X_n$  variabili aleatorie discrete che assumono i loro valori nell'insieme finito  $E$ , chiamato un *alfabeto*, con il generico elemento, o *lettera*, *simbolo*, denotato con  $\alpha_i$

$$E = \{\alpha_1, \dots, \alpha_k\}$$

Posto  $X = (X_1, \dots, X_n)$ ,  $X$  è un vettore aleatorio che assume i suoi valori nell'insieme finito  $E^n$ , l'insieme delle *parole di lunghezza  $n$* , scritte nell'alfabeto  $E$ . Per ogni vettore aleatorio  $Y$ , con valori in un insieme finito  $F$ , con distribuzione di probabilità  $P(Y = y) = p(y)$ , si definisce *quantità di informazione media* (o *entropia*<sup>36</sup>, o anche *incertezza* di  $Y$ ),  $H(Y)$  contenuta in  $Y$  il valore

$$H(Y) = -E[\log p(Y)] := - \sum_{y \in F} p(y) \log p(y) \quad (8.70)$$

ove la base del logaritmo è specificata solo quando necessario; la notazione  $H_b$  è utilizzata per indicare che la definizione utilizza il logaritmo a base  $b$ . Per convenzione, si pone  $0 \log 0 = 0$ .

La funzione  $I(y) := -\log P(Y = y)$  può essere interpretata come la quantità di informazione fornita dall'evento  $Y = y$ . In accordo a tale interpretazione, meno probabile è un evento e maggiore è l'informazione che si ottiene quando esso si verifica.

<sup>35</sup>Come semplice esempio introduttivo, si consideri una variabile aleatoria  $X$  che assume con uguale probabilità i valori 1, 2, 3, 4, 5, e si cerchi la quantità di informazione, intorno al valore assunto da  $X$ , contenuta nell'affermazione  $1 \leq X \leq 2$ . Si ha, naturalmente, che la probabilità di avere il valore esatto di  $X$  è aumentata rispetto alla situazione originale in cui la probabilità è  $1/5$ . Allora, l'informazione  $1 \leq X \leq 2$  ha ridotto l'*incertezza* intorno al valore attuale di  $X$ .

<sup>36</sup>Il termine *entropia* come concetto scientifico fu utilizzato per la prima volta in termodinamica (Clausius, 1850). La sua interpretazione probabilistica nel contesto della meccanica statistica è attribuita a Boltzmann (1877). Comunque, la relazione esplicita tra entropia e probabilità fu stabilita successivamente (Planck, 1906). Shannon (1948) utilizzò il concetto di entropia per fornire una descrizione sintetica delle proprietà di una sequenza lunga di simboli e applicò i risultati a numerosi problemi relativi alla teoria della codifica e alla trasmissione dei dati.

Un evento certo (corrispondente a probabilità 1) fornisce nessuna informazione, al contrario di un evento improbabile.

► **Esempio 8.62** Quando  $P(X_1 = \alpha_i) = p_i$ , si ha

$$H(X_1) = - \sum_{i=1}^k p_i \log p_i =: H(p_1, p_2, \dots, p_k)$$

Ad esempio, se  $E = \{0, 1\}$  e  $P(X_1 = 0) = p$  e  $P(X_1 = 1) = 1 - p$ , allora  $H(X_1) = H(p) = -p \log p - (1 - p) \log(1 - p)$ . In Figura 8.25 è riportato il grafico della funzione  $p \rightarrow H(X_1)$ . Per  $p = 1/2$  si ha  $I(0) = I(1) = H(X_1) = \log_2 2 = 1$ . Cioè l'osservazione del "bit"  $X_1$  fornisce un "bit" di informazione.

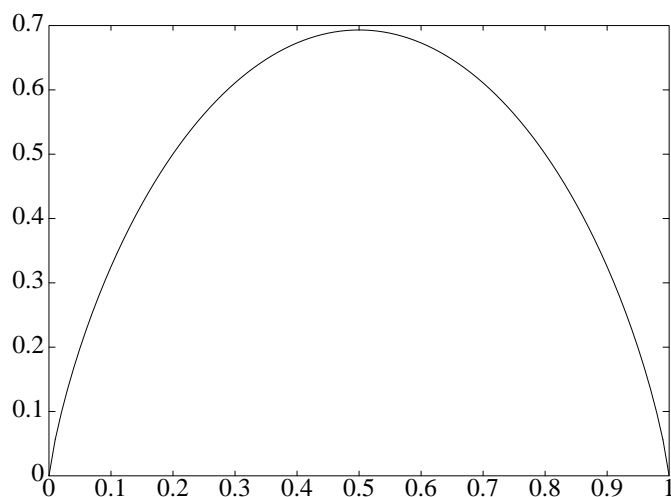


Figura 8.25: Funzione di entropia  $H_2(p)$ .

Più in generale, siano  $X$  e  $Y$  due variabili aleatorie discrete associate al medesimo esperimento, con probabilità congiunta

$$p(x_i, y_j) = P(X = x_i \text{ e } Y = y_j) = p_{ij} \quad i = 1, \dots, k, \quad j = 1, \dots, r$$

Si ha, quindi, un esperimento con  $kr$  possibili risultati; il risultato  $[X = x_i, Y = y_j]$  ha probabilità  $p(x_i, y_j)$ . Si definisce *entropia congiunta* di  $X$  e  $Y$  la seguente quantità

$$H(X, Y) := - \sum_{i=1}^k \sum_{j=1}^r p(x_i, y_j) \log p(x_i, y_j)$$

In modo analogo, si definisce l'entropia congiunta di  $n$  variabili aleatorie  $X_1, X_2, \dots, X_n$  la quantità

$$H(X_1, \dots, X_n) = - \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_1, x_2, \dots, x_n)$$



ove  $p(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$  è la probabilità congiunta di  $X = [X_1, X_2, \dots, X_n]$ . In particolare, quando  $X_1, \dots, X_n$  sono variabili aleatorie indipendenti e identicamente distribuite, si ha

$$\begin{aligned} H(X) &= H(X_1, \dots, X_n) = -E[\log p(X_1, \dots, X_n)] = -E[\log p(X_1) \cdots p(X_n)] \\ &= -E[\log p(X_1) + \cdots + \log p(X_n)] = -\sum_{j=1}^n E[\log p(X_j)] \\ &= \sum_{j=1}^n H(X_j) = nH(X_1) \end{aligned}$$

Nel caso generale, si può mostrare la seguente relazione tra l'entropia individuale e l'entropia congiunta

$$H(X_1, \dots, X_n) \leq H(X_1) + \cdots + H(X_n)$$

ove l'uguaglianza si ha se e solo se  $X_1, \dots, X_n$  sono indipendenti.

Tornando al caso di due variabili aleatorie  $X$  e  $Y$ , si definisce *entropia condizionata di  $Y$  dato  $X = x_i$*  la quantità

$$H(Y | X = x_i) = -\sum_{j=1}^r p(y_j | x_i) \log p(y_j | x_i)$$

ove  $p(y_j | x_i)$ ,  $j = 1, 2, \dots, r$  è la distribuzione di probabilità di  $Y$  dato  $X = x_i$ . Si definisce, allora, *entropia condizionata di  $Y$  dato  $X$*  la media pesata di  $H(Y | X = x_i)$ , ossia

$$H(Y | X) = -\sum_{i=1}^k p(x_i) \sum_{j=1}^r p(y_j | x_i) \log p(y_j | x_i)$$

Tenendo conto che  $p(x_i, y_j) = p(x_i)p(y_j | x_i)$ , si ha

$$H(Y | X) = -\sum_{i=1}^k \sum_{j=1}^r p(x_i, y_j) \log p(y_j | x_i)$$

Il risultato può essere esteso, facilmente, al caso di un numero generico di variabili.

A proposito dell'entropia condizionata, ricordiamo i seguenti risultati. Supponendo che  $X$  e  $Y$  siano due variabili aleatorie osservate, ma che solo il valore di  $X$  sia rivelato, si può dimostrare che

$$H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y)$$

e

$$H(Y | X) \leq H(Y)$$

ove l'uguaglianza si ha se e solo se  $X$  e  $Y$  sono indipendenti. Si definisce, allora, *informazione trasmessa da  $Y$  su  $X$*  la quantità

$$I(X | Y) := H(X) - H(X | Y) = E \left[ -\log \frac{p(X)}{p(X | Y)} \right]$$

Dai risultati precedenti si ha  $I(X | Y) \geq 0$  e  $I(X | Y) = I(Y | X)$ , ossia l'informazione trasmessa su  $X$  da  $Y$  è la stessa dell'informazione trasmessa su  $Y$  da  $X$ .

Il concetto di misura di informazione trova applicazione, in particolare, nello studio della trasmissione di segnali attraverso canali di comunicazione con rumori. Qui ci limiteremo ad una semplice illustrazione. Consideriamo l'esperimento del lancio di due monete, di cui una A simmetrica (testa e croce) e l'altra B, che presenta due teste. Scelta a caso una moneta, si effettuano due lanci e si registra il numero di teste ottenute. Si vuole calcolare la quantità di informazione trasmessa alla conoscenza dell'identità della moneta scelta (A o B) dalla conoscenza del numero di teste ottenute. È chiaro, infatti, che il numero di teste ottenuto fornisce informazione: se si sono ottenute meno di due teste, allora è uscita la moneta simmetrica A; d'altra parte, se ambedue i lanci hanno fornito testa, l'evidenza favorisce la moneta B.

Formalizziamo l'esperimento nel seguente modo. Sia  $X$  la variabile aleatoria che assume il valore 0 o 1, a seconda che sia stata estratta la moneta A o la moneta B. Sia, inoltre,  $Y$  il numero di teste ottenuto nei due lanci della moneta scelta. L'incertezza iniziale intorno l'identità della moneta è  $H(X)$ . Dopo che il numero di teste è rivelato, l'incertezza diventa  $H(X | Y)$ . Si ottengono, allora, i seguenti risultati numerici (il logaritmo si intende nella base 2).

$$\begin{aligned} H(X) &= \log 2 = 1 \\ H(X | Y) &= P(Y=0)H(X | Y=0) + P(Y=1)H(X | Y=2) + P(Y=2)H(X | Y=2) \\ &= \frac{1}{8} \cdot 0 + \frac{1}{4} \cdot 0 - \frac{5}{8} \left( \frac{1}{5} \log \frac{1}{5} + \frac{4}{5} \log \frac{4}{5} \right) = 0.45 \\ I(X | Y) &= 0.55 \end{aligned}$$

▼ **Osservazione 8.10** Si può mostrare che la funzione

$$H(p_1, \dots, p_k) := -C \sum_{i=1}^k p_i \log p_i$$

ove  $C$  è un numero arbitrario positivo, è la sola funzione che verifica i seguenti quattro assiomi

1. La funzione  $f(k) := H(1/k, 1/k, \dots, 1/k)$  è una funzione monotona crescente di  $k$  ( $k = 1, 2, \dots$ ).
2.  $f(kr) = f(k) + f(r)$ ,  $k, r = 1, 2, \dots$
3. Per  $r = 1, 2, \dots, k-1$  si ha (assioma del raggruppamento)

$$\begin{aligned} H(p_1, \dots, p_k) &= H(p_1 + \dots + p_r, p_{r+1} + \dots + p_k) \\ &\quad + (p_1 + \dots + p_r) H\left(\frac{p_1}{\sum_{i=1}^r p_i}, \dots, \frac{p_r}{\sum_{i=1}^r p_i}\right) \\ &\quad + (p_{r+1} + \dots + p_k) H\left(\frac{p_{r+1}}{\sum_{i=r+1}^k p_i}, \dots, \frac{p_k}{\sum_{i=r+1}^k p_i}\right) \end{aligned}$$

4.  $H(p, 1-p)$  è una funzione continua di  $p$ .

### 8.8.2 Disuguaglianza di Gibbs

Cerchiamo una limitazione superiore di  $H(X_1)$  in termini della dimensione  $k$  di  $E$ . Utilizzando la nota disuguaglianza  $\log z \leq z - 1$  ( $z > 0$ ), ove l'uguaglianza è verificata se e solo se  $z = 1$ , si può dimostrare la seguente disuguaglianza (nota anche come *disuguaglianza di Gibbs*)

$$-\sum_{i=1}^k p_i \log p_i \leq -\sum_{i=1}^k p_i \log q_i$$

ove  $(p_i, 1 \leq i \leq k)$  e  $(q_i, 1 \leq i \leq k)$  sono due qualunque distribuzioni di probabilità discrete. In effetti, supponendo, senza di perdita di generalità, che  $p_i > 0$ , per  $1 \leq i \leq k$ , si ha

$$\sum_{i=1}^k p_i \log \frac{q_i}{p_i} \leq \sum_{i=1}^k p_i \left( \frac{q_i}{p_i} - 1 \right) = \sum_{i=1}^k q_i - \sum_{i=1}^k p_i = 0$$

L'uguaglianza si ha se e solo se  $p_i = q_i, 1 \leq i \leq k$ . Ad esempio, poniamo

$$p_1 = \frac{1}{2}, \quad p_2 = p_3 = \frac{1}{4}, \quad q_1 = \frac{1}{3}, \quad q_2 = \frac{4}{9}, \quad q_3 = \frac{2}{9}$$

Allora

$$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4} = 1.5$$

e

$$-\frac{1}{2} \log \frac{1}{3} - \frac{1}{4} \log \frac{4}{9} - \frac{1}{4} \log \frac{2}{9} = 1.63$$

Prendendo  $q_i = 1/k, 1 \leq i \leq k$  nella disuguaglianza di Gibbs si ottiene

$$H(X_1) = H(p_1, \dots, p_k) \leq \log k$$

e l'uguaglianza si ottiene se e solo se  $P(X_1 = \alpha_i) = p_i \equiv \frac{1}{k}, 1 \leq i \leq k$ . Osserviamo, infine, che ovviamente  $H(X_1) \geq 0$ . L'uguaglianza a 0 si ottiene se e solo se  $p_i \log p_i = 0$ , per  $1 \leq i \leq n$ , cioè se  $p_i = 0$  oppure  $p_i = 1$ . Tenendo conto che  $\sum_{i=1}^k p_i = 1$ , si ha che esiste uno ed un solo valore  $j$  per cui  $p_j = 1$ . Pertanto,  $H(X_1) = 0$  se e solo se per un valore di  $j \in \{1, 2, \dots, k\}$  si ha  $P(X_1 = \alpha_j) = 1$ .

### 8.8.3 Codifica e disuguaglianza di Kraft

Un *codice binario* (binary code) per  $E$  è una trasformazione  $c$  da  $E$  nell'insieme  $\{0, 1\}^*$ , definito come l'insieme delle sequenze finite (incluso la sequenza vuota), di 0 e 1. La sequenza  $c(\alpha_i)$  rappresenta la parola codice per  $\alpha_i$ ; con  $l_i(c)$  si indica la lunghezza di tale parola codice. Il codice  $c$  è detto *codice istantaneo* o *codice prefisso* (prefix code) se non esiste nessuna coppia di indici  $(i, j)$ , con  $i \neq j$  tali che  $c(\alpha_i)$  sia l'inizio di  $c(\alpha_j)$ .

► **Esempio 8.63** Se  $E = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ , il codice definito nel seguente modo

$$\begin{aligned}\alpha_1 &\rightarrow c(\alpha_1) = 0 \\ \alpha_2 &\rightarrow c(\alpha_2) = 1 \\ \alpha_3 &\rightarrow c(\alpha_3) = 10 \\ \alpha_4 &\rightarrow c(\alpha_4) = 11\end{aligned}\tag{8.71}$$

non è un codice prefisso, poiché  $c(\alpha_2)$  è l'inizio sia di  $c(\alpha_3)$  che di  $c(\alpha_4)$ . È, invece, un codice prefisso il seguente

$$\begin{aligned}\alpha_1 &\rightarrow c(\alpha_1) = 00 \\ \alpha_2 &\rightarrow c(\alpha_2) = 01 \\ \alpha_3 &\rightarrow c(\alpha_3) = 10 \\ \alpha_4 &\rightarrow c(\alpha_4) = 11\end{aligned}\tag{8.72}$$

■

I codici prefissi sono *univocamente decodificabili* (uniquely decodable), ossia, se una successione finita di 0 e 1 è ottenuta mediante codifica di una stringa finita di lettere da  $E$ , la stringa originale di lettere può essere ricostruita senza ambiguità. Per esempio, la codifica di  $\alpha_2\alpha_1$  mediante il codice (8.71) fornisce 10, e tale risultato può anche essere ottenuto dalla codifica di  $\alpha_3$  col medesimo codice. Pertanto, il codice (8.71) non è univocamente decodificabile<sup>37</sup>.

Un codice binario può essere rappresentato mediante un *albero binario* (binary tree). Come esemplificazione, nella Figura 8.26 in (a) è rappresentato il codice (8.71), e in (b) il codice (8.72).

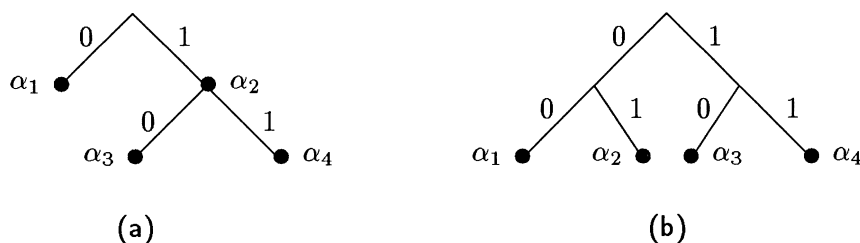


Figura 8.26: Due differenti codici su un albero.

**Proposizione 8.13** Se  $c$  è un codice prefisso, allora

$$\sum_{i=1}^k 2^{-l_i(c)} \leq 1\tag{8.73}$$

<sup>37</sup>Osserviamo che, viceversa, un codice univocamente decodificabile non è necessariamente istantaneo. Si consideri, come esempio, il codice  $\alpha_1 \leftrightarrow 0$ ;  $\alpha_2 \leftrightarrow 01$ , che non è istantaneo, dal momento che 0 è un prefisso di 01. Tuttavia, il codice è univocamente decodificabile, poiché ogni sequenza può essere decodificata osservando la posizione dei bits 1 nella sequenza. Ad esempio, la sequenza 0010000101001 è decodificata come  $\alpha_1\alpha_2\alpha_1\alpha_1\alpha_1\alpha_1\alpha_2\alpha_2\alpha_1\alpha_2$ . L'attributo "istantaneo" si riferisce al fatto che una sequenza di caratteri codificati può essere decodificata passo per passo.

Viceversa, se un insieme di interi  $l_i$ ,  $1 \leq i \leq k$ , verifica la seguente disuguaglianza, nota anche come disuguaglianza di Kraft

$$\sum_{i=1}^k 2^{-l_i} \leq 1 \quad (8.74)$$

allora esiste almeno un codice binario prefisso  $c$  su  $E$  per il quale  $l_i(c) = l_i$ ,  $1 \leq i \leq k$ .

Per la dimostrazione della prima parte si consideri il sottoalbero con  $\alpha_i$  come radice e i  $2^m$  nodi dell'albero binario al livello  $m$ , ove  $m = \sup_{1 \leq i \leq k} l_i(c)$ . Dal momento che  $c$  è un codice prefisso, non vi sono altri  $\alpha_j$  sul sottoalbero originato da  $\alpha_i$ . Pertanto,  $2^m \geq \sum_{i=1}^k 2^{m-l_i(c)}$ . Lasciamo la dimostrazione della seconda parte come esercizio.

Indichiamo con  $\mathcal{P}$  l'insieme dei codici binari prefissi per  $E$ . Definiamo per ogni codice binario  $c$  per  $E$  la sua lunghezza media  $L(c)$  ponendo

$$L(c) = \sum_{i=1}^k p_i l_i(c)$$

Si può, allora, dimostrare il seguente risultato.

**Proposizione 8.14** *Assumendo 2 come base dei logaritmi e utilizzando  $H_2(X_1)$  come notazione per indicare  $H(X_1)$ , si ha*

$$H_2(X_1) \leq \inf_{c \in \mathcal{P}} L(c) \leq H_2(X_1) + 1 \quad (8.75)$$

**DIMOSTRAZIONE.** Incominciamo a considerare il seguente problema di minimo con vincoli

$$\left\{ \begin{array}{l} \min_{l_i} \sum_{i=1}^k p_i l_i \\ \sum_{i=1}^k 2^{-l_i} = 1 \quad l_i \in \mathbb{R}, 1 \leq i \leq k \end{array} \right.$$

Utilizzando il metodo dei moltiplicatori di Lagrange, si considera la funzione

$$f(\lambda, l_1, \dots, l_k) = \sum_{i=1}^k p_i l_i + \lambda \left( \sum_{i=1}^k e^{-l_i \ln 2} - 1 \right)$$

La minimizzazione di tale funzione porta a  $l_i = -\log_2 p_i$ ,  $1 \leq i \leq k$  e  $\lambda = 1/\ln 2$ . Assumendo  $l_i$  come il più piccolo intero più grande di  $-\log_2 p_i$ , si ottiene

$$-\sum_{i=1}^k p_i \log_2 p_i \leq \sum_{i=1}^k p_i l_i \leq \sum_{i=1}^k p_i (-\log_2 p_i + 1) = -\sum_{i=1}^k p_i \log_2 p_i + 1$$

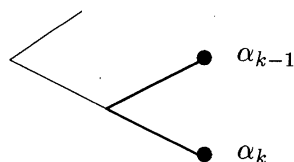
■

### 8.8.4 Codici ottimali; algoritmo di Huffman

Supponiamo che le probabilità  $p_i$  siano state ordinate in maniera decrescente, ossia  $p_1 \geq \dots \geq p_k$ . Sia  $c$  un codice binario prefisso per  $E$  con  $l_i(c) = l_i$ . Consideriamo le seguenti due proprietà

$$\begin{aligned} (\mathcal{P}_1) \quad & l_1 \leq \dots \leq l_k \\ (\mathcal{P}_2) \quad & l_k = l_{k-1} \end{aligned}$$

Si può mostrare che se tali condizioni non sono ambedue verificate, allora esiste un codice binario prefisso con lunghezza media minore o uguale (non peggiore) di  $c$  che le soddisfa. Inoltre, se le condizioni  $(\mathcal{P}_1)$ ,  $(\mathcal{P}_2)$  sono verificate, allora esiste un codice non peggiore di  $c$  che presenta una configurazione del tipo mostrato nella seguente figura.



Per dimostrare le proprietà precedenti è sufficiente scambiare i nodi nell'albero. Esse sono alla base di un algoritmo ricorsivo (*algoritmo di Huffman*) per la costruzione di un codice binario prefisso con lunghezza media minima. In sostanza, si considera l'insieme

$$E' = E - \{\alpha_{k-1}, \alpha_k\} + \{\alpha'_{k-1}\} = \{\alpha_1, \dots, \alpha_{k-2}, \alpha'_{k-1}\}$$

con distribuzione di probabilità

$$p_1, p_2, \dots, p_{k-2}, p'_{k-1} = p_{k-1} + p_k$$

cioè si combinano gli ultimi due simboli  $\alpha_{k-1}, \alpha_k$  in un simbolo equivalente  $\alpha_{k,k-1}$  con probabilità  $p_k + p_{k-1}$ . Supponendo, ora, di avere già costruito un codice ottimale  $C_2$  per il nuovo insieme di simboli, si costruisce un codice  $C_1$  per l'insieme originale dei simboli  $\alpha_1, \dots, \alpha_k$  nel seguente modo. I codici parole associate con  $\alpha_1, \dots, \alpha_{k-2}$  sono esattamente gli stessi dei corrispondenti codici parole di  $C_2$ . I codici parole associate con  $\alpha_{k-1}$  e  $\alpha_k$  sono formati aggiungendo uno zero e un uno, rispettivamente, al codice parola associato con il simbolo  $\alpha_{k,k-1}$  in  $C_2$ . Si può dimostrare che  $C_1$  è un codice ottimale per l'insieme di probabilità  $p_1, \dots, p_k$ . Per una descrizione più adeguata dell'algoritmo, rinviamo alla bibliografia citata.

### 8.8.5 Codifica di blocchi

Consideriamo la codifica  $E^n$ , l'insieme di blocchi (parole) di lunghezza  $n$ . Supponiamo che le lettere delle parole di  $E^n$  siano scelte indipendentemente l'una dall'altra

e estratte dall'alfabeto  $E$  con probabilità  $P(X_j = \alpha_i) = p_i$ ,  $1 \leq j \leq n$ ,  $1 \leq i \leq k$ . Si ha, allora, il seguente risultato.

**Proposizione 8.15** *Se  $c^{(n)}$  è un codice binario prefisso ottimale per  $E^{(n)}$ , allora*

$$\lim_{n \rightarrow \infty} \frac{L(c^{(n)})}{n} = H_2(X_1)$$

Per la dimostrazione, basta osservare che

$$H_2(X_1, \dots, X_n) \leq L(c^{(n)}) \leq H_2(X_1, \dots, X_n) + 1$$

Ma  $H_2(X_1, \dots, X_n) = nH_2(X_1)$ , per cui

$$H_2(X_1) \leq \frac{L(c^{(n)})}{n} \leq H_2(X_1) + \frac{1}{n}$$

da cui il risultato, che, in altre parole, esprime il fatto che  $H_2(X_1)$  è asintoticamente il minimo numero medio di cifre binarie necessarie per simbolo di  $E$ .

### 8.8.6 Applicazione alla costruzione di questionari

Supponiamo che uno tra i  $k$  "oggetti"  $\alpha_1, \dots, \alpha_k$  venga estratto casualmente secondo la distribuzione di probabilità  $p_1, \dots, p_k$ . L'oggetto non è mostrato, ma deve essere identificato attraverso domande alle quali può essere risposto con un "sì" o con un "no". Si tratta di ottimizzare la procedura di individuazione e di valutare il minimo numero medio di domande necessarie per identificare l'oggetto. Naturalmente, dal

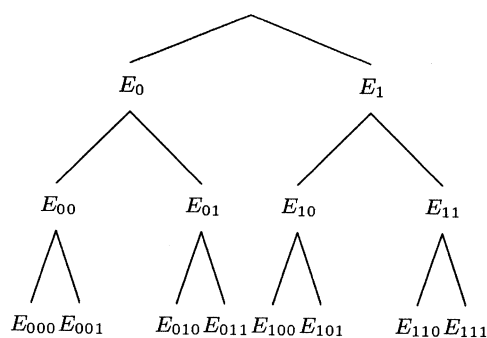


Figura 8.27: Albero corrispondente a un questionario.

momento che una domanda "sì" o "no" è associata con una partizione in due classi, la prima domanda sarà associata con una partizione di  $E = \{\alpha_1, \dots, \alpha_k\}$  in due classi  $E_1$  e  $E_0$ , con  $E_0 \cap E_1 = \emptyset$  e  $E_0 + E_1 = E$ . Supponiamo che la risposta sia

“sì”, cioè che l’oggetto sia in  $E_1$ . La successiva domanda, se si vuole procedere in maniera ottimale, riguarderà solo la classe  $E_1$ . Tale domanda sarà associata con una partizione  $E_1 = E_{10} + E_{11}$ ; analoga situazione se la risposta alla prima domanda fosse stata “no”: si avrebbe in questo caso la ripartizione  $E_0 = E_{00} + E_{01}$ . Un questionario ottimale può essere, pertanto, rappresentato mediante un albero (cfr. Figura 8.27). L’albero termina quando un insieme in un nodo contiene soltanto un oggetto. In definitiva, il problema della costruzione di un questionario è ricondotto allo studio di un codice prefisso per  $E = \{\alpha_1, \dots, \alpha_k\}$  e un questionario ottimale può essere ottenuto con la procedura di Huffman.

Come esemplificazione, si supponga di avere  $k$  oggetti, che sono perfettamente identici, salvo che uno ed uno solo di essi ha un peso leggermente differente dagli altri. Per la sua identificazione si ha a disposizione una bilancia. Non sapendo a priori se l’oggetto diverso è più pesante o più leggero degli altri la risposta della bilancia è di tipo “sì”, “no”, in quanto le pendenze a sinistra o a destra hanno, per la ricerca, lo stesso effetto. Lasciamo come esercizio la ricerca di una strategia ottimale di pesatura, supponendo che gli oggetti abbiano la stessa probabilità di essere estratti per la pesatura.

Associated with each discrete memoryless channel, there is a nonnegative number  $C$  (called channel capacity) with the following property. For any  $\epsilon > 0$  and  $R < C$ , for large enough  $n$ , there exists a code of length  $n$  and rate  $\geq R$  (i. e. with at least  $2^{Rn}$  distinct codewords), and an appropriate decoding algorithm, such that, when the code is used on the given channel, the probability of decoder error is  $< \epsilon$ .

**Shannon’s channel coding theorem** (1948)



Fallaces sunt rerum species.

Seneca

## Capitolo 9

# Algoritmi nella cluster analysis

La *cluster analysis* (analisi dei raggruppamenti), un'importante tecnica nel campo dell'*analisi dei dati*, è lo studio formale degli algoritmi e dei metodi per raggruppare, o classificare, oggetti in base a misure di "similarità". L'organizzazione dei dati in gruppi è, notoriamente, uno degli strumenti fondamentali della conoscenza e dell'apprendimento; da qui l'interesse della cluster analysis nelle scienze applicate, in particolare nell'ambito del riconoscimento di forme (*pattern recognition*) e dell'*intelligenza artificiale*<sup>1</sup>. L'obiettivo di questo capitolo è una breve introduzione alle idee essenziali della tecnica della cluster analysis, con particolare riguardo all'aspetto *numerico*. Come bibliografia di base segnaliamo, ad esempio, Anderberg [4], Hartigan [79], Jain e Dubes [93].

### 9.1 Rappresentazione dei dati

Alla base della cluster analysis vi sono le nozioni relative alla *rappresentazione dei dati*, quali le misure di "prossimità" o di similitudine dei dati, i tipi e le scale dei dati. Il presente paragrafo è dedicato ad una introduzione a tali nozioni.

---

<sup>1</sup>Lo sviluppo della cluster analysis è il risultato di numerosi contributi interdisciplinari, in particolare da parte di biologi, psicologi, statistici, cultori di scienze sociali e, naturalmente, di matematici e ingegneri. Per indicare la cluster analysis è stato anche suggerito (I. J. Good, 1977) il nome *botriologia* (botriology) dal greco *βότρυς* (grappolo). La cluster analysis è connessa con altre importanti tecniche statistiche per l'analisi multivariata, in particolare con l'*analisi fattoriale*, l'*analisi discriminante* e l'*analisi delle componenti principali*.

### 9.1.1 Matrice campione

Dato un insieme  $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$  di  $n$  oggetti, ognuno dei quali individuato da  $d$  misure (o attributi) *osservabili*, l'insieme può essere rappresentato da una matrice bidimensionale di  $n$  righe e  $d$  colonne, detta matrice campione degli individui (*pattern matrix*). La generica riga  $i$ -ma di tale matrice definisce un oggetto, o un individuo,  $I_i$  e la generica colonna  $C_j$  indica una misura, o una caratteristica (*feature*). Ad esempio, se  $\mathcal{I}$  rappresenta un gruppo di pazienti, le colonne possono corrispondere ai risultati di determinati test e le righe ai vari pazienti. Le  $d$  caratteristiche possono essere considerate come le coordinate rispetto ad un insieme di assi ortogonali, e allora gli  $n$  individui corrispondono a  $n$  punti in uno spazio euclideo  $\mathbb{R}^d$ , a  $d$  dimensioni, detto spazio degli individui (*pattern space*) (cfr. Figura 9.1 per una semplice illustrazione).

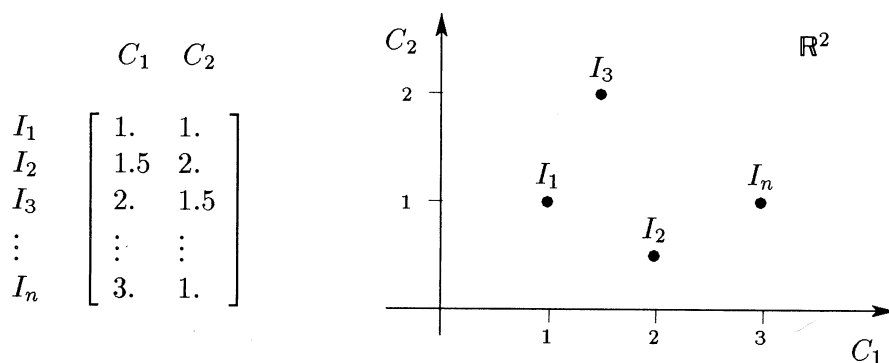


Figura 9.1: Esempio di matrice pattern e corrispondente spazio pattern.

### 9.1.2 Matrice di prossimità

Un modo differente di presentare i dati consiste nel fornire gli *indici di affinità*, o di *prossimità*, tra coppie di individui. Questi indici, essenziali per gli algoritmi di clustering, possono essere calcolati, come vedremo nel seguito, in vari modi a partire dalla matrice campione, ma possono essere anche assegnati, a seconda delle applicazioni, direttamente a partire dai dati. Gli indici possono essere pensati come gli elementi  $d_{ij}$  di una matrice quadrata di ordine  $n$ , detta *matrice di prossimità* (*proximity matrix*). Più precisamente,  $d_{ij}$  corrisponde all'indice di prossimità tra l'individuo  $i$  e l'individuo  $j$ . Usualmente, gli indici di prossimità sono indipendenti dall'ordine degli elementi nella coppia  $(i, j)$  e la matrice è quindi simmetrica. Un indice di prossimità può indicare una *similarità* o una *dissimilarità*. Per esempio, la distanza euclidea tra due individui nello spazio pattern è un indice di dissimilarità,

nel senso che più è piccola la distanza e più gli individui sono simili tra loro. Al contrario, il coefficiente di correlazione è un indice di similarità.

### 9.1.3 Il problema del clustering

Sia  $m$  un intero minore di  $n$ . A partire dai dati contenuti nella matrice pattern, il problema del clustering consiste nel determinare  $m$  cluster (sottoinsiemi) di individui in  $\mathcal{I}$ , indicati ad esempio con  $\pi_1, \pi_2, \dots, \pi_m$ , in modo che il generico individuo  $I_i$  appartenga ad uno e ad un solo sottoinsieme e che gli individui assegnati allo stesso cluster siano, in base alle informazioni contenute nella matrice di prossimità, *simili*, mentre gli individui appartenenti a cluster differenti siano *dissimili*.

La definizione precedente può essere ulteriormente precisata, introducendo un *criterio di ottimalità* da raggiungere. Tale criterio può, naturalmente, essere dato in vari modi, in termini di una relazione funzionale che rifletta i livelli di desiderabilità delle varie partizioni o raggruppamenti.

► **Esempio 9.1** Supponiamo di avere un insieme  $\mathcal{I}$  con  $n = 8$  elementi e  $d = 1$  caratteristiche, con matrice pattern data dal vettore colonna  $[3, 4, 7, 4, 3, 3, 4, 4]^T$ . Indichiamo con  $x_i$  il valore della caratteristica corrispondente all'individuo  $I_i$ . Un criterio di ottimalità può essere ottenuto a partire dalle varianze entro ciascun gruppo. Più precisamente, se  $\pi_j$  è un generico elemento della suddivisione, con  $n_j$  individui, si definisce

$$W_j = \sum_{i=1}^{n_j} (x_i - \bar{x}_j)^2 = \sum_{i=1}^{n_j} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n_j} x_i \right)^2$$

Si definisce *funzione obiettivo* la funzione (dipendente dai vari modi di raggruppare gli elementi)  $W = W_1 + \dots + W_m$ , ove  $m$  è il numero dei gruppi corrispondenti ad un determinato raggruppamento. Come criterio di ottimalità, si può allora assumere la *minimizzazione* della funzione  $W$  nell'ambito dell'insieme di tutte le possibili suddivisioni dell'insieme  $\mathcal{I}$ . In questo modo, si è naturalmente assunto come indice di prossimità la distanza euclidea. Nel caso particolare in considerazione, si ottiene  $W = 12$ , in corrispondenza alla suddivisione costituita da un unico gruppo formato dagli 8 elementi, mentre si ha  $W = 0$  quando

$$\pi_1 = \{3, 3, 3\}; \quad \pi_2 = \{4, 4, 4, 4\}; \quad \pi_3 = \{7\} \quad (9.1)$$

Pertanto, se si desidera una partizione in tre gruppi, il valore ottimale è 0, corrispondente alla suddivisione (9.1). Osserviamo, tuttavia, che nelle applicazioni è usualmente necessario considerare sia il valore della funzione obiettivo che il numero dei gruppi desiderato. Il numero  $m$  dei cluster può essere in alcuni casi scelto convenientemente a priori, ma in generale tale numero è determinato dallo stesso procedimento di clustering.

Osserviamo, infine, che un modo diretto per risolvere il problema di cluster, in corrispondenza ad un numero  $m$  fissato, può consistere nella considerazione di tutte le possibili suddivisioni in  $m$  gruppi, scegliendo quella (o quelle) che realizzano il minimo della funzione obiettivo. Tale idea (nota anche come procedura di clustering per *completa enumerazione*) ha, comunque, interesse pratico solo quando i numeri  $m$  (numero dei cluster) e  $n$  (numero

degli individui) sono *piccoli*. In effetti, basta osservare che, ad esempio, vi sono 1701 modi di partizionare  $n = 8$  oggetti in  $m = 4$  sottoinsiemi<sup>2</sup>. ■

### 9.1.4 Tipi di dati e scale

Uno degli aspetti più importanti nell'applicazione della tecnica di cluster analysis e nell'interpretazione dei risultati riguarda la *quantizzazione* dei dati (cioè degli elementi sia della matrice pattern che della matrice proximity). Spesso tali dati si presentano in forma *binaria* (ad esempio, risposte si-no in un questionario), o in forma *qualitativa* (variabili categoriali, ad esempio il partito di appartenenza). Esistono varie tecniche per rappresentare tali variabili in forma *continua* (cioè, come variabili continue su un determinato intervallo), che è, in generale, la forma più idonea nella risoluzione numerica. Rinviando alla bibliografia per una trattazione più adeguata, ci limiteremo a fornire un semplice esempio che evidenzia l'importanza della scelta di una scala conveniente. I quattro punti rappresentati in Figura 9.2 si raggruppano in maniera diversa operando una trasformazione di coordinate.

### 9.1.5 Indici di prossimità

Un indice di prossimità  $d_{ik}$  tra l'individuo  $i$  e l'individuo  $k$ , verifica le seguenti proprietà

1.  $d_{ii} = 0$ , se l'indice rappresenta una *dissimilarità*, oppure  $d_{ii} \geq \max_k d_{ik}$  se è un indice di *similarità*.
2.  $d_{ik} = d_{ki}$ , per ogni  $(i, k)$ .
3.  $d_{ik} \geq 0$ , per ogni  $(i, k)$ .

---

<sup>2</sup>Il procedimento di partizionare un insieme di  $n$  oggetti in  $m$  sottoinsiemi non vuoti può essere visto come il problema di porre  $n$  distinte palline in  $m$  scatole identiche, in modo che nessuna di esse rimanga vuota. Si può mostrare che, quando l'ordine delle palline entro ciascuna scatola è irrilevante, come pure l'ordine delle  $m$  scatole, il numero dei vari modi possibili è dato da

$$S(n, m) := \frac{1}{m!} \sum_{j=0}^m (-1)^j \binom{m}{j} (m-j)^n$$

I numeri  $S(n, m)$  sono anche detti *numeri di Stirling di seconda specie*. Si può vedere che

$$S(n, 0) = 0, \quad x^n = \sum_{i=1}^n S(n, i) x(x-1) \cdots (x-i+1), \quad S(n, n+k) = 0 \text{ per } k > 0$$

e vale inoltre la seguente formula ricorrente

$$S(n+1, i) = iS(n, i) + S(n, i-1)$$

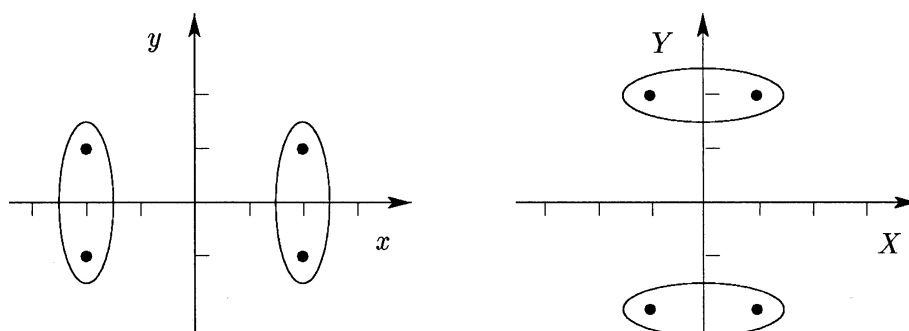


Figura 9.2: Cambiamento di scala:  $X = x/2$ ;  $Y = 2y$ .

Un indice di prossimità può essere determinato in vari modi. Indichiamo con  $x_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, d$  gli elementi della matrice campione. Il generico oggetto  $i$ -mo è quindi individuato dal vettore  $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$ ,  $i = 1, 2, \dots, m$ . Supponiamo che  $x_{ij}$  siano variabili continue opportunamente scalate. Il più comune indice di prossimità è allora dato dalla *metrica di Minkowski*, definita nel modo seguente

$$d(i, k) := \left( \sum_{j=1}^d |x_{ij} - x_{kj}|^r \right)^{1/r}$$

ove  $r$  è un numero reale con  $r \geq 1$ . Osserviamo che per tale metrica si ha

4.  $d(i, k) = 0$  se e solo se  $\mathbf{x}_i = \mathbf{x}_k$
5.  $d(i, k) \leq d(i, m) + d(m, k)$ , per ogni  $i, k, m$ .

Come casi particolari, si hanno le seguenti metriche (cfr. Figura 9.3)

- a)  $r = 2$  (distanza euclidea)  $d(i, k) := \left[ \sum_{j=1}^d (x_{ij} - x_{kj})^2 \right]^{1/2}$
- b)  $r = 1$  (distanza Manhattan)  $d(i, k) := \sum_{j=1}^d |x_{ij} - x_{kj}|$
- c)  $r \rightarrow \infty$  (distanza del massimo, o di Chebichev)  $d(i, k) := \max_{1 \leq j \leq d} |x_{ij} - x_{kj}|$

La metrica più comune nelle applicazioni è la metrica euclidea, la quale è invariante rispetto alle traslazioni e alle rotazioni nello spazio campione. Quando le caratteristiche sono quantizzate da *variabili binarie*, la metrica di Manhattan è chiamata

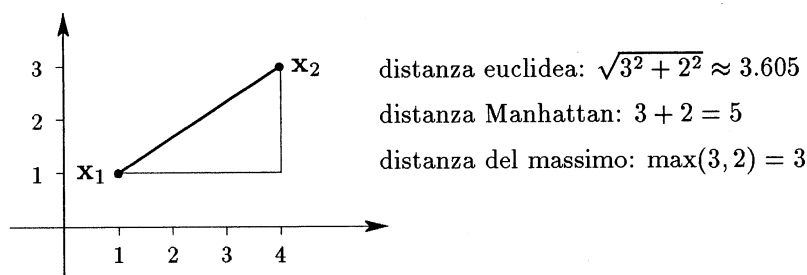


Figura 9.3: Esempi particolari di metrica di Minkowski.

*distanza di Hamming*; essa corrisponde al numero di caratteristiche per le quali i due individui differiscono.

Per terminare, ricordiamo la seguente metrica, nota come *distanza di Mahalanobis*

$$d(i, k) := (\mathbf{x}_i - \mathbf{x}_k)^T \mathcal{S}^{-1} (\mathbf{x}_i - \mathbf{x}_k)$$

ove  $\mathcal{S}$  è la matrice di covarianza relativa ai dati  $[x_{ik}]$ , definita dalla relazione

$$\mathcal{S} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \quad \text{ove } \bar{\mathbf{x}} = \sum_{i=1}^n \frac{\mathbf{x}_i}{n}$$

La distanza di Mahalanobis è invariante rispetto a trasformazioni lineari non singolari e tiene conto della correlazione tra le caratteristiche; inoltre essa standardizza ogni caratteristica a variabili con media zero e varianza unitaria.

### 9.1.6 Variabili nominali

Quando le variabili sono di tipo nominale, gli indici di prossimità possono essere ottenuti in vari modi a partire dal numero delle variabili che coincidono. Supponiamo, ad esempio, che le variabili siano di tipo binario (0, 1). Indichiamo, allora, per una generica coppia di individui  $\mathbf{x}_i$  e  $\mathbf{x}_k$  con  $a_{11}$  il numero delle caratteristiche che valgono 1 per ambedue gli individui, e con  $a_{10}$  il numero delle caratteristiche che sono 1 per l'individuo  $\mathbf{x}_i$  e 0 per  $\mathbf{x}_k$ . In modo analogo sono definiti  $a_{01}$  e  $a_{00}$ . Si ottiene, pertanto, la seguente matrice

$$\begin{array}{cc} & \mathbf{x}_k \\ & \begin{array}{cc} 1 & 0 \end{array} \\ \mathbf{x}_i & \begin{array}{|cc|} \hline 1 & \begin{array}{cc} a_{11} & a_{10} \end{array} \\ 0 & \begin{array}{cc} a_{01} & a_{00} \end{array} \\ \hline \end{array} \end{array}$$

A partire da tale matrice si possono definire vari tipi di indici di prossimità. Rinviamo alla bibliografia per un'analisi più approfondita, ricordiamo i seguenti casi particolari

1. *Coefficiente semplice di accoppiamento* (simple matching coefficient)

$$d_{ik} = \frac{a_{00} + a_{11}}{a_{00} + a_{11} + a_{01} + a_{10}} = \frac{a_{00} + a_{11}}{d}$$

2. *Coefficiente di Jaccard*

$$d_{ik} = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} = \frac{a_{11}}{d - a_{00}}$$

Il coefficiente di Jaccard, a differenza del coefficiente semplice, ignora le coincidenze degli zeri; questo, naturalmente, può essere ragionevole nelle applicazioni per le quali è importante la coincidenza dei valori 1.

► **Esempio 9.2** Supponiamo che le risposte a 20 domande di un test possano essere “sì” (1) oppure “no” (0) e che due particolari insiemi di risposte abbiano dato i risultati contenuti nella seguente tabella

$\mathbf{x}_1$	01100100100111001010
$\mathbf{x}_2$	01110000111111011010

Si ottiene allora la seguente matrice

		$\mathbf{x}_2$	
		1	0
$\mathbf{x}_1$	1	8	1
	0	4	7

da cui

Coefficiente semplice:  $15/20=0.75$   
Coefficiente di Jaccard:  $8/13=0.615$

■

### 9.1.7 Proiezioni lineari

Gli algoritmi di proiezione trasformano un insieme di  $n$  vettori dello spazio  $\mathbb{R}^d$  in vettori di un sottospazio  $\mathbb{R}^m$  con  $m < d$ . L'interesse di tali algoritmi nell'ambito della cluster analysis consiste nel fatto che essi permettono, in particolare per  $m = 2$ , una *visualizzazione* dei dati, permettendo un'analisi qualitativa dei risultati ottenuti dagli algoritmi di cluster. Indicando con  $\mathbf{y} \in \mathbb{R}^m$  il vettore trasformato, si ha

$$\mathbf{y}_i = \mathcal{A} \mathbf{x}_i \quad \text{per } i = 1, \dots, n$$

ove  $\mathcal{A}$  è una matrice  $m \times d$ . Una scelta interessante di tale matrice consiste nell'assumere come assi coordinati in  $\mathbb{R}^m$  gli *autovettori* corrispondenti agli autovalori più grandi della matrice di covarianza dei dati. Con una opportuna scelta di  $m$  si può

in tal modo mantenere la dispersione dei dati originali in uno spazio a dimensione più piccola. Tale tecnica è anche nota come *metodo di Karhunen-Loeve* o *metodo delle componenti principali* ed è collegata con la tecnica statistica dell'*analisi discriminante*.

Terminiamo, osservando che quando i dati presentano una “struttura complessa”, ossia i punti corrispondenti agli individui giacciono su superfici curve, può essere conveniente utilizzare opportune trasformazioni *non lineari*. Si tratta, in sostanza, di trovare delle trasformazioni che riducano la dimensione, mantenendo la struttura. La maggior parte degli algoritmi di proiezione non lineare si basano sulla massimizzazione o minimizzazione di una funzione a più variabili, sono cioè problemi di ottimizzazione.

## 9.2 Metodi e algoritmi di clustering

La maggior parte degli algoritmi di clustering sono basati sulle seguenti due tecniche: la ricerca della partizione che minimizza, per un numero fissato di partizioni, la dispersione entro i cluster, o massimizza la dispersione tra i cluster (algoritmi di partizione basati sulla varianza, *square-error partitioning algorithms*), oppure una agglomerazione di tipo gerarchico (*agglomerative hierarchical algorithms*), nella quale i dati sono organizzati in una sequenza nidificata di gruppi. In questo paragrafo esamineremo brevemente alcuni modi differenti di realizzare le due idee precedenti, corrispondenti ad algoritmi diversi. Vale la pena osservare che nelle applicazioni non esiste l'algoritmo di cluster “migliore”; in effetti, una strategia conveniente può consistere nel confronto comparativo dei risultati ottenuti mediante l'applicazione di più algoritmi.

### 9.2.1 Algoritmi di tipo gerarchico

Indichiamo con  $\mathcal{H}$  l'insieme degli  $n$  oggetti da clusterizzare, cioè

$$\mathcal{H} := \{x_1, x_2, \dots, x_n\}$$

ove  $x_i$  rappresenta l'oggetto  $i$ -mo. Una partizione  $\mathcal{C}$  di  $\mathcal{H}$ , è una suddivisione di  $\mathcal{H}$  in sottoinsiemi  $\{C_1, C_2, \dots, C_m\}$  che verificano le seguenti proprietà

$$\begin{aligned} C_i \cap C_j &= \emptyset & i, j &= 1, \dots, m, i \neq j \\ C_1 \cup C_2 \cup \dots \cup C_m &= \mathcal{H} \end{aligned}$$

Una partizione  $\mathcal{B}$  è *nidificata* (nested) nella partizione  $\mathcal{C}$  se ogni componente di  $\mathcal{B}$  è un sottoinsieme proprio di una componente di  $\mathcal{C}$ . In altre parole,  $\mathcal{C}$  è ottenuta per *fusione* (merging) di componenti di  $\mathcal{B}$ . Come illustrazione, si considerino le seguenti



due partizioni

$$\mathcal{C} = \{(x_1, x_3, x_5, x_7), (x_2, x_4, x_6, x_8), (x_9, x_{10})\}$$

$$\mathcal{B} = \{(x_1, x_3), (x_5, x_7), (x_2), (x_4, x_6, x_8), (x_9, x_{10})\}$$

Una procedura di clustering di tipo *gerarchico* consiste in una successione di partizioni nella quale ogni partizione è nidificata nella successiva partizione della successione. La procedura può essere applicata procedendo in due modi differenti. Negli algoritmi di tipo *agglomerativo* (agglomerative) si parte dalla partizione nella quale ognuno degli  $n$  oggetti sono posti in un singolo cluster; tale partizione è detta *cluster disgiunto*. L'algoritmo procede, quindi, ad immergere successivamente due o più cluster sulla base delle informazioni ottenute dalla matrice di prossimità. Il procedimento è ripetuto per formare una successione di cluster nidificati nella quale il numero di cluster diminuisce progressivamente fino ad ottenere un unico cluster contenente tutti gli  $n$  oggetti (*cluster congiunto*). Negli algoritmi di tipo *suddivisione* (divise) si procede alla rovescia partendo dal cluster congiunto e dividendo successivamente le partizioni, fino ad arrivare al cluster disgiunto.

Le procedure di tipo gerarchico sono usualmente illustrate mediante un *dendrogramma*, un tipo particolare di albero consistente in livelli di nodi, ognuno dei quali rappresenta un cluster (cfr. Figura 9.4 per una illustrazione). Si ottiene un clustering particolare tagliando orizzontalmente il dendrogramma. Nel seguito daremo un'idea di alcuni algoritmi di tipo gerarchico, rinviando alla bibliografia per ulteriori approfondimenti. Poiché i metodi che considereremo utilizzano alcuni risultati della *teoria dei grafi*, richiameremo nel prossimo paragrafo alcune nozioni di base di tale teoria.

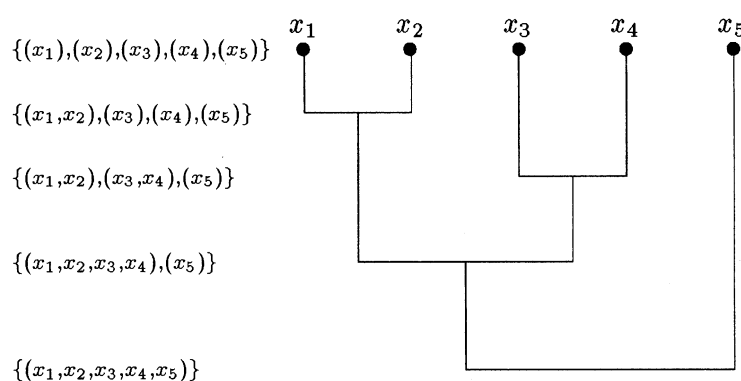


Figura 9.4: Esempio di dendrogramma.

### 9.2.2 Elementi di teoria dei grafi

Un *grafo lineare* è un diagramma costituito da punti, detti *vertici* e da segmenti di retta, chiamati *lati*, o archi (*edge*), che congiungono tali vertici. Nelle applicazioni un grafo costituisce un modello astratto, o matematico, per rappresentare la struttura di un sistema; i vertici indicano le componenti del sistema e gli archi le relazioni tra tali componenti<sup>3</sup>.

Indicando con  $V = \{v_i\}$  l'insieme dei nodi, con  $E = \{e_j\}$  l'insieme degli archi, e con  $f$  una funzione che associa ad ogni arco una coppia non ordinata di vertici  $f : E \rightarrow V \times V$ , un *grafo non orientato* (undirected, o nonoriented graph)  $G$  è definito dalla terna  $G = \langle V, E, f \rangle$ . In altre parole, un grafo non orientato rappresenta una relazione binaria, simmetrica, non riflessiva sull'insieme dei vertici. Un grafo è detto *orientato* (directed) quando le coppie di vertici sono ordinate; in questo caso nella rappresentazione grafica del grafo gli archi sono opportunamente orientati mediante una freccia (cfr. Figura 9.5 per alcuni esempi di rappresentazione di grafi).

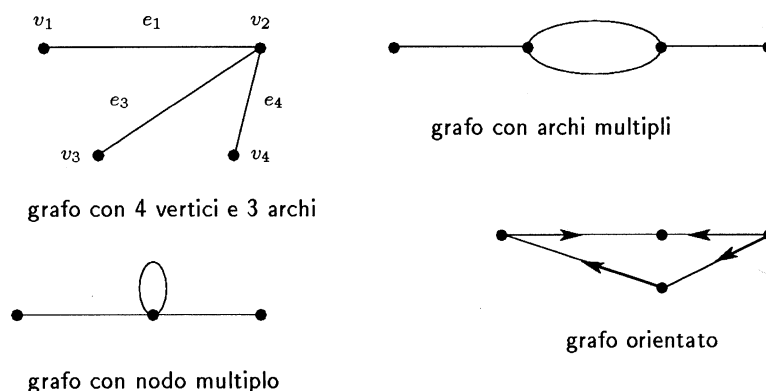


Figura 9.5: Esempi di rappresentazioni di grafi.

Quando i due vertici ai quali  $f$  assegna un arco sono distinti, si dice che il grafo non presenta vertici multipli *self-loop*; inoltre, non si hanno archi *multipli* quando  $f$  assegna ogni arco a una distinta coppia di vertici. Per il seguito i grafi considerati saranno supposti senza vertici *self-loop* e senza archi multipli. Se l'arco  $e$  è associato alla coppia di vertici  $(v_1, v_2)$ ,  $f(e) = (v_1, v_2)$ , l'arco  $e$  è detto *incidente* ai vertici  $v_1$

<sup>3</sup>Uno dei primi lavori nella teoria dei grafi è dovuto a Eulero (1736), in relazione alla soluzione del cosiddetto problema dei ponti di Königsberg. Nel 1847, Firkhoff impiegò la teoria dei grafi nell'analisi dei circuiti elettrici. Un impulso allo sviluppo della teoria è stato dato dalla congettura (1852) relativa al problema dei quattro colori (risolto da Appel e Haken nel 1976 mediante una dimostrazione che prevede anche una verifica su calcolatore). A seconda del tipo di applicazioni, un grafo lineare è anche indicato come *grafo di flusso* (flow graph), *flusso dei dati* (flow chart), *diagramma di stato* (state diagram), eccetera.

e  $v_2$ . Nel caso generale possono esistere vertici senza alcun arco incidente. Per un grafo  $G$  costituito da  $n$  vertici, si definisce *matrice di adiacenza*  $\mathbf{A} = [a_{ij}]$  associata con  $G$  la matrice di ordine  $n$  definita nel seguente modo

$$a_{ij} = \begin{cases} 1 & (v_i, v_j) \in E \\ 0 & (v_i, v_j) \notin E \end{cases}$$

Nell'Esempio illustrato in Figura 9.5 (4 vertici e 3 archi) si ha

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Un grafo  $H = \langle V', E', f' \rangle$  è detto *sottografo* del grafo  $G = \langle V, E, f \rangle$  se  $V'$  è un sottoinsieme di  $V$ ,  $E'$  è un sottoinsieme di  $E$  e  $f'$  è la restrizione di  $f$  al sottoinsieme  $E'$ .

Un *cammino* (path) in un grafo  $G$  tra i vertici  $v_1$  e  $v_n$  è una sequenza di vertici e archi

$$v_1 e_1 v_2 e_2 \cdots v_{n-1} e_{n-1} v_n$$

che non contiene archi e vertici ripetuti e per il quale l'arco  $e_i$  è incidente ai vertici  $v_i$  e  $v_{i+1}$  (cfr. Figura 9.6).

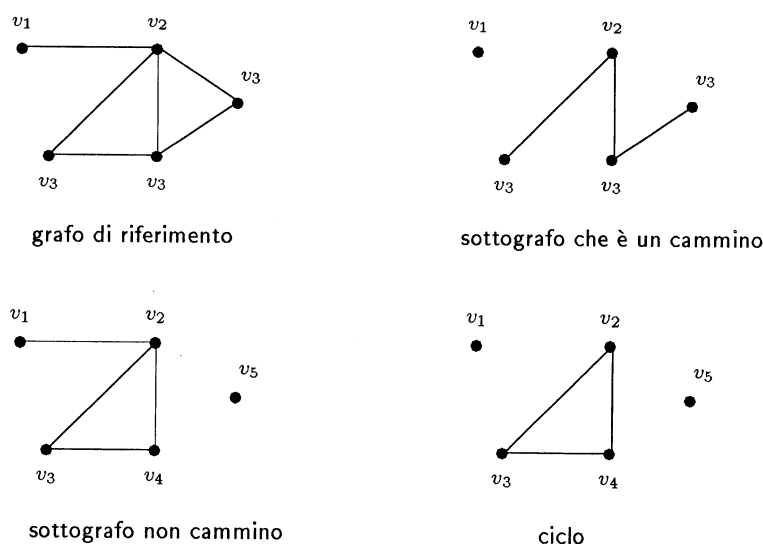


Figura 9.6: Proprietà dei grafi.

Un grafo è *connesso* se esiste un cammino tra ogni coppia di vertici nel grafo. Una *componente* è un sottografo *massimale* di un grafo connesso; in altre parole una componente non è un sottografo proprio di un altro grafo connesso. Un grafo  $G$  è *completo* se ad ogni coppia di vertici è assegnato un arco. Pertanto, un grafo completo costituito da  $n$  vertici contiene esattamente  $n(n-1)/2$  archi (si veda, per un esempio curioso, Figura 9.7).

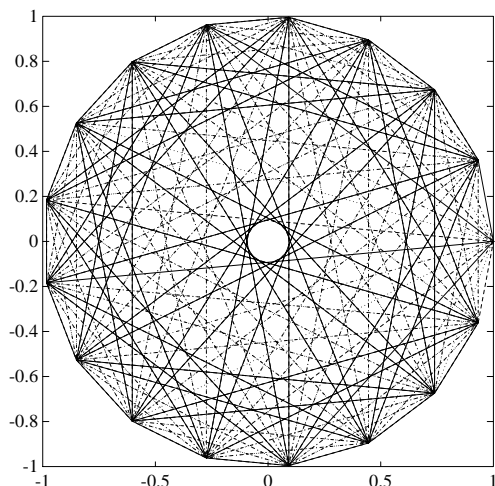


Figura 9.7: Esempio di grafo completo. I nodi corrispondono alle 17 radici nel piano complesso dell'equazione  $z^{17} - 1 = 0$ . Ogni nodo è connesso con ciascun altro.

Un sottografo completo  $H$  di un grafo  $G$  che non sia un sottografo proprio di ogni altro sottografo completo di  $G$  è detto *sottografo completo massimale* di  $G$ . In altre parole,  $H$  è un sottografo completo massimale di  $G$  se un arco è incidente a ogni coppia di vertici in  $H$ , ma non è possibile aggiungere a  $H$  ulteriori vertici di  $G$  senza distruggere la completezza (cfr. Figura 9.8). Un *ciclo* (cycle) è un cammino nel quale i vertici  $v_1$  e  $v_n$  coincidono. Un *albero* (tree) è un grafo connesso senza cicli. Se un sottografo ha  $m$  vertici, si può facilmente dimostrare che un albero contenente tali vertici ha esattamente  $m-1$ . Un albero generatore (*spanning tree*) è un albero che contiene tutti i vertici del grafo (cfr. per una illustrazione Figura 9.9). Quando gli archi in un grafo sono pesati (ad esempio, nelle applicazioni al problema di clustering, dai coefficienti di dissimilarità) il *peso* di un albero è la somma dei pesi relativi agli archi nell'albero. Un albero generatore minimale (*minimal spanning tree* (MST)) di un grafo  $G$  è un albero che ha il peso minimo tra tutti gli spanning tree di  $G$  (cfr. per una illustrazione Figura 9.10). Naturalmente, come si vede facilmente su esempi, un grafo può avere più di un MST. Uno dei più noti algoritmi per il calcolo di un minimal spanning tree è il seguente algoritmo proposto da Kruskal (1956).

**Algoritmo 9.1** (Algoritmo di Kruskal) *In input si ha l'insieme  $E$  degli archi e il peso  $f$ ,*

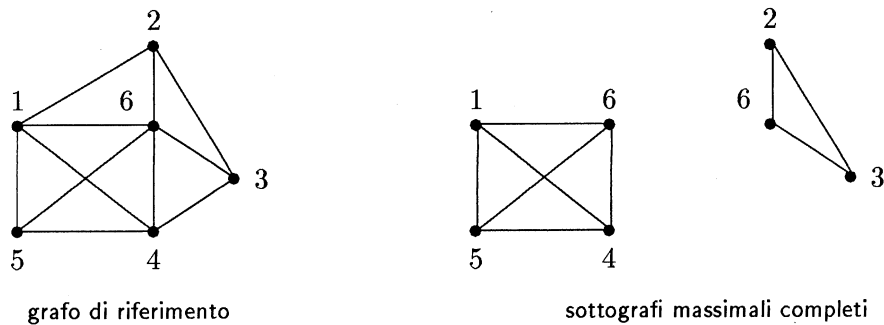


Figura 9.8: Esempi di grafi.

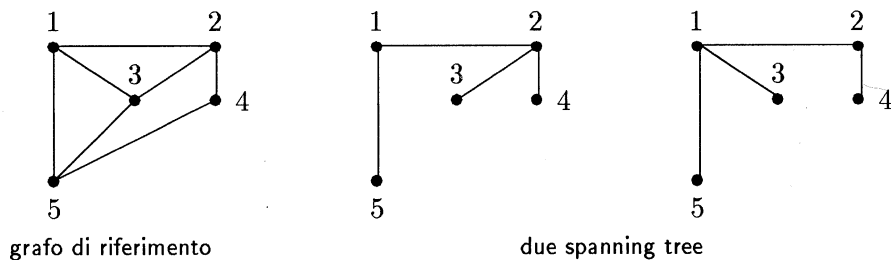


Figura 9.9: Esempi di spanning tree.

relativi a un grafo pesato  $G$ . I dati sono memorizzati in un file sequenziale della forma  $\{(v, w, f(\{v, w\}))\}$ , con  $v \neq w$ . In output si ha un insieme di archi  $E'$  che rappresenta un minimal spanning tree.

```

sorting input file  $\{(v, w, f)\}$  in ordine crescente rispetto a  $f$ ;
 $E' = \emptyset$ 
 $C = \{\{v_1\}, \{v_2\}, \dots, \{v_n\}\}$  % insieme dei vertici;
while( $|C| > 1$ ) do %  $|C| =$  numero di elementi in  $C$ ;
  read un arco  $\{v, w\}$  dal file input;
  if  $v \in K, w \in K', K, K' \in C, K \neq K'$ 
  then
     $K'' = K \cup K'$ ;
    si cancella  $K$  e  $K'$  da  $C$  e si aggiunge  $K''$  a  $C$ ;
     $E' = E' \cup \{\{v, w\}\}$ ;
  endif
repeat

```

L'algoritmo parte dal grafo triviale  $T = (V, E') = (V, \emptyset)$  e il numero degli archi in  $E'$  aumenta man mano procede l'iterazione principale. Si vede facilmente che

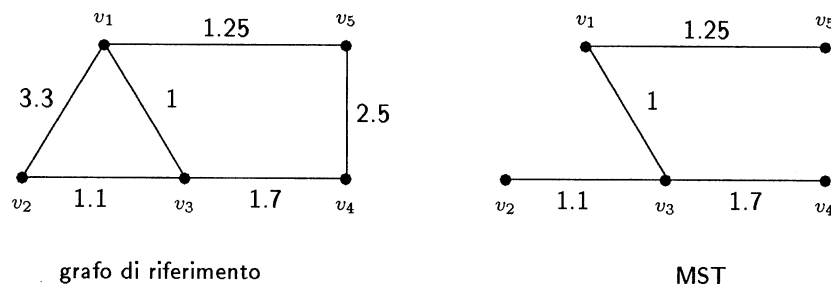


Figura 9.10: Esempio di grafo pesato.

ad ogni passo dell'iterazione principale ogni componente connessa di  $T = (V, E')$  è un albero, e che l'insieme dei vertici di ogni componente connessa è dato da un insieme  $K \in C$ . L'algoritmo termina quando  $T$  diventa connesso, nel qual caso  $T$  è un minimal spanning tree (grazie all'ordinamento in ordine crescente effettuato sulla successione degli archi). Per una descrizione più dettagliata dell'algoritmo precedente e di altri algoritmi disponibili per la costruzione di MST si può vedere ad esempio Sedgewick [143].

Gli spanning tree di grafi completi risultano, come vedremo anche successivamente, particolarmente importanti nella cluster analysis. Dato, infatti, un MST nel quale i vertici rappresentino gli oggetti e il peso  $f$  una dissimilarità, si può ottenere un clustering eliminando gli archi per i quali il peso supera una *soglia*  $\alpha$  prefissata. L'eliminazione di tali archi produce un insieme di componenti connesse dell'albero, corrispondenti agli elementi del clustering. Nell'esempio illustrato in Figura 9.10, per  $\alpha = 1.2$  si ottiene un clustering formato dai tre gruppi  $\{1, 2, 3\}$ ,  $\{5\}$ ,  $\{4\}$ , mentre per  $\alpha = 1.5$  si hanno i due gruppi  $\{1, 2, 3, 5\}$ ,  $\{4\}$ . Naturalmente, nelle applicazioni risulta cruciale, per un utilizzo appropriato dei risultati, la scelta del valore di soglia  $\alpha$ , che definisce il livello di dissimilarità che si ritiene accettabile.

### 9.2.3 Algoritmi Single-Link e Complete-Link

Data una matrice di prossimità  $\mathcal{D} = [d_{ij}]$  simmetrica di ordine  $n$ , supponiamo che gli  $n(n-1)/2$  elementi della matrice triangolare superiore contengano una permutazione degli interi da 1 a  $n(n-1)/2$ , cioè che i coefficienti di prossimità siano dati in forma di scala ordinale; supponiamo, inoltre, per fissare le idee, che i coefficienti di prossimità indichino una dissimilarità, e quindi che, ad esempio,  $d_{12} > d_{13}$  significhi che gli oggetti 1 e 3 sono più "simili" che gli oggetti 1 e 2.

► **Esempio 9.3** Come esempio illustrativo di matrice di prossimità ordinale si consideri

per  $n = 5$  la seguente matrice

$$\mathcal{D}_1 = \begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} & \begin{bmatrix} 0 & 6 & 8 & 2 & 7 \\ 6 & 0 & 1 & 5 & 3 \\ 8 & 1 & 0 & 10 & 9 \\ 2 & 5 & 10 & 0 & 4 \\ 7 & 3 & 9 & 4 & 0 \end{bmatrix} \end{matrix}$$

■

Si definisce, per ogni livello di dissimilarità  $v$ , *grafo soglia* (threshold graph) il grafo  $G(v)$  non orientato e non pesato di  $n$  nodi, senza nodi o archi multipli, nel quale ogni nodo rappresenta un individuo dell'insieme  $\mathcal{H}$  da clusterizzare ed è tale che per ogni coppia di nodi  $i$  e  $j$  si ha un arco  $(i, j)$  soltanto se gli individui  $i$  e  $j$  sono meno dissimili del livello  $v$ . Ossia

$$(i, j) \in G(v) \quad \text{se e solo se} \quad d_{ij} \leq v$$

Trattandosi di una misura di dissimilarità, si avrà  $d_{ii} = 0$ . Per ogni numero reale  $v$ , si ha che  $G(v)$  definisce nell'insieme  $\mathcal{H} \times \mathcal{H}$ , una relazione binaria che è simmetrica e riflessiva.

► **Esempio 9.3** (*continuazione*) Indicando con il simbolo  $*$  nella posizione  $(i, j)$  l'appartenenza della coppia  $(x_i, x_j)$  alla relazione binaria, si ha, per  $v = 5$ , la rappresentazione della matrice  $\mathcal{D}_1$  e del corrispondente grafo soglia riportate in Figura 9.11. ■

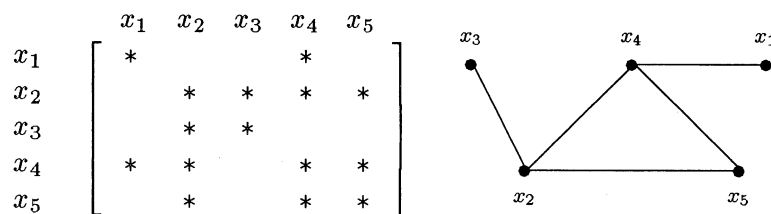


Figura 9.11: Relazione binaria e grafico soglia relativi all'Esempio 9.3.

Vediamo ora come i grafi soglia possono essere utilizzati per la costruzione di algoritmi di clustering di tipo agglomerativo. In effetti, ci limiteremo a dare l'idea di base considerando due semplici algoritmi; per algoritmi più idonei all'implementazione su calcolatore rinviamo alla bibliografia.

**Algoritmo 9.2** (Algoritmo agglomerativo single-link) (*nearest neighbour*)

- Step 1. Si incomincia con il clustering disgiunto corrispondente al grafo soglia  $G(0)$ , che contiene nessun arco e che pone ogni oggetto in un unico cluster. Si pone  $k \leftarrow 1$ .
- Step 2. Si crea il grafo soglia  $k$ . Se il numero delle componenti in  $G(k)$  è minore del numero dei cluster nel raggruppamento attuale, si ridefinisce il raggruppamento attuale assumendo come cluster ogni componente di  $G(k)$ .
- Step 3. Se  $G(k)$  consiste di un singolo grafo connesso stop. Altrimenti, si pone  $k \leftarrow k + 1$  e si ritorna allo Step 2.

**Algoritmo 9.3** (Algoritmo agglomerativo complete-link) (*furthest neighbour*)

- Step 1. Si incomincia con il clustering disgiunto corrispondente al grafo soglia  $G(0)$ , che contiene nessun arco e che pone ogni oggetto in un unico cluster. Si pone  $k \leftarrow 1$ .
- Step 2. Si crea il grafo soglia  $k$ . Se due dei cluster attuali formano un sottografo massimalmente completo in  $G(k)$ , si ridefinisce il raggruppamento attuale assumendo riunendo tali due cluster in un singolo cluster.
- Step 3. Se  $k = n(n-1)/2$ , e quindi  $G(k)$  è un grafo completo sugli  $n$  nodi stop. Altrimenti, si pone  $k \leftarrow k + 1$  e si ritorna allo Step 2.

Si può formare un *dendrogramma soglia* (threshold dendrogram) ricordando i clustering nell'ordine nei quali essi sono formati. Aggiungendo i livelli di dissimilarità ai quali viene formato ogni clustering, si ottiene un *dendrogramma di prossimità* (proximity dendrogram).

Come illustrazione, in Figura 9.12 sono riportati i grafi soglia e il dendrogramma ottenuti mediante l'algoritmo single-link sull'Esempio 9.3. Lasciamo come utile esercizio l'analoga costruzione relativa all'algoritmo complete-link. I cluster single-link sono caratterizzati come sottografi massimalmente connessi, mentre i cluster complete-link sono sottografi massimalmente completi. Osserviamo che la completezza è una proprietà più forte della connessione. In pratica, i cluster single-link possono essere più dispersi e meno omogenei dei cluster complete-link, i quali, a loro volta, possono essere meno separati.

#### 9.2.4 Algoritmi di clustering di tipo partizione

Abbiamo visto che gli algoritmi di clustering di tipo gerarchico organizzano i dati in una successione nidificata di gruppi. Una caratteristica interessante dei metodi gerarchici è la possibilità di visualizzare i risultati sotto forma di dendrogramma, che può rappresentare nelle applicazioni un valido aiuto nella scelta del clustering opportuno. In questo paragrafo esamineremo brevemente altri tipi di tecniche, di tipo non gerarchico e indicate usualmente come algoritmi di partizione (*partitional clustering*). Le tecniche di tipo gerarchico sono utilizzate, in particolare, in biologia, scienze sociali e del comportamento, scienze naturali, a motivo della necessità di



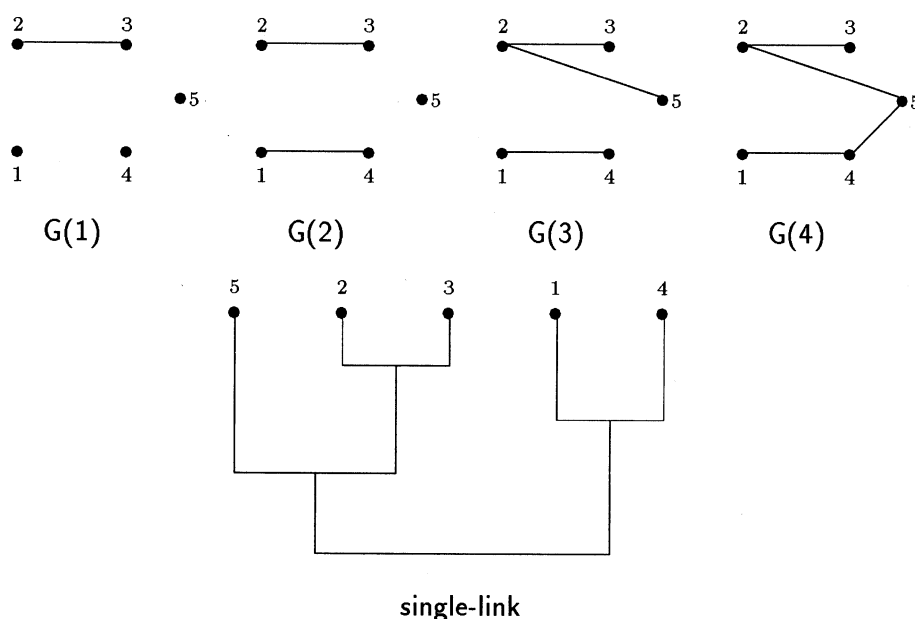


Figura 9.12: Grafi soglia e dendrogramma ottenuti con l'algoritmo single-link per i dati dell'Esempio 9.3.

avere in tali applicazioni delle classificazioni (*taxonomy*). Le tecniche di tipo partizione sono usate frequentemente nelle applicazioni nelle quali risultano importanti anche le singole partizioni; inoltre, esse risultano particolarmente appropriate nella rappresentazione e compressione di basi di dati di grandi dimensioni, per le quali i dendrogrammi risultano impraticabili.

Il problema della partitional clustering può essere formulato nel seguente modo. Dati  $n$  individui (pattern) in uno spazio metrico  $\mathbb{R}^d$ , si vuole determinare una partizione degli individui in  $K$  gruppi, o cluster, in maniera che gli individui in un cluster siano più simili tra di loro che non agli individui in cluster differenti. Come abbiamo già osservato in precedenza, questo risultato può essere ottenuto mediante l'ottimizzazione di una funzione particolare (*criterio*). Il criterio più comunemente utilizzato è basato sull'errore quadratico (*square-error criterion*), corrispondente alla varianza entro i cluster. L'obiettivo della clustering è, allora, quello di ottenere una partizione che, per un numero fissato di cluster, minimizzi l'errore quadratico, o equivalentemente massimizzi la varianza tra i cluster.

Più in particolare, supponiamo che l'insieme degli  $n$  individui, rappresentati nello spazio euclideo  $\mathbb{R}^d$ , siano partizionati in qualche modo in  $K$  cluster  $\{C_1, C_2, \dots, C_K\}$ , in modo che il cluster  $C_k$  abbia  $n_k$  individui e che ogni individuo sia in un solo cluster, cioè che  $\sum_{k=1}^K n_k = n$ . Il vettore medio, o *centro*, del cluster  $C_k$  è definito come il

baricentro del cluster, ossia

$$\mathbf{m}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_i^{(k)}$$

ove  $\mathbf{x}_i^{(k)}$  è il vettore che rappresenta il generico individuo  $i$ -mo nel cluster  $C_k$ . L'errore quadratico relativo al cluster  $C_k$  è la somma dei quadrati delle distanze euclidee tra ogni individuo in  $C_k$  e il centro del cluster  $\mathbf{m}^{(k)}$ , ossia la seguente quantità

$$e_k^2 = \sum_{i=1}^{n_k} (\mathbf{x}_i^{(k)} - \mathbf{m}^{(k)})^T (\mathbf{x}_i^{(k)} - \mathbf{m}^{(k)})$$

L'errore quadratico  $E_K^2$  relativo all'intero clustering contenente  $K$  cluster è la somma degli errori quadratici relativi ad ogni cluster, ossia

$$E_K^2 = \sum_{k=1}^K e_k^2$$

Un metodo di clustering di tipo errore quadratico consiste nella ricerca di una partizione in  $K$  cluster, con  $K$  fissato, che minimizzi  $E_K^2$ . La partizione che si ottiene in tal modo è detta partizione con varianza minima ed è caratterizzata dal fatto che i cluster corrispondenti hanno una forma di iperellissoidi. Naturalmente, la partizione con varianza minima può cambiare se le variabili sono scalate in maniera differente, in quanto l'errore quadratico non è invariante rispetto alle trasformazioni lineari non singolari. Si possono, tuttavia, costruire criteri invarianti rispetto a tali trasformazioni. In questo senso, segnaliamo, ad esempio, il criterio fornito dalla traccia della matrice  $\mathcal{S}_W^{-1} \mathcal{S}_B$ , ove, rispettivamente,  $\mathcal{S}_W$  è la matrice di dispersione entro i cluster e  $\mathcal{S}_B$  la matrice di dispersione tra i cluster, ossia

$$\mathcal{S}_W = \sum_{k=1}^K \mathcal{S}^{(k)}, \quad \text{ove } \mathcal{S}^{(k)} = \sum_{j=1}^{n_k} (\mathbf{x}_j^{(k)} - \mathbf{m}^{(k)}) (\mathbf{x}_j^{(k)} - \mathbf{m}^{(k)})^T$$

$$\mathcal{S}_B = \mathcal{S} - \mathcal{S}_W, \quad \text{ove } \mathcal{S} = \sum_{k=1}^K \sum_{j=1}^{n_k} (\mathbf{m}_j^{(k)} - \mathbf{m}) (\mathbf{m}_j^{(k)} - \mathbf{m})^T$$

ove  $\mathbf{m} = (1/n) \sum_{k=1}^K n_k \mathbf{m}^{(k)}$ . Differenti criteri possono, ovviamente, portare a differenti clustering e a priori non è facile, in generale, conoscere quale criterio sia maggiormente appropriato per una determinata applicazione.

Occupiamoci ora di come ottenere la partizione che minimizza, per ogni  $K$  fissato, un determinato criterio. Abbiamo già osservato in precedenza che una enumerazione esaustiva di tutte le possibili partizioni è computazionalmente possibile solo per un numero molto piccolo di individui. In generale, quindi, sarà necessario utilizzare un procedimento di tipo iterativo, nel quale, partendo da una partizione iniziale, si

muovono gli individui da un cluster ad un altro in maniera da migliorare il valore della funzione obiettivo. Ogni successiva partizione è allora una partizione della precedente e, quindi, viene esaminato solo un numero piccolo di partizioni. Algoritmi basati su tale tecnica sono computazionalmente efficienti ma possono convergere a *minimi locali* della funzione criterio. Un metodo di clustering partizionale iterativo può essere implementato in maniere differenti, alle quali possono corrispondere differenti partizioni. Alla base di tali algoritmi vi sono, tuttavia, i seguenti passi generali.

**Algoritmo 9.4** (Algoritmo iterativo di clustering partizionale)

- Step 1. *Si sceglie una partizione iniziale con  $K$  cluster.  
Si ripetono i passi da 2 a 5 fino a che la suddivisione in cluster si stabilizza.*
- Step 2. *Si genera una nuova partizione assegnando ogni individuo al centro più vicino.*
- Step 3. *Si assumono come nuovi centri i baricentri dei cluster.*
- Step 4. *Si ripetono i passi 2 e 3 fino a trovare un valore ottimale della funzione criterio.*
- Step 5. *Si aggiusta il numero dei cluster per unione o separazione dei cluster esistenti per soddisfare determinate condizioni (eliminazione di outlier, o separazione di cluster troppo grandi, ecc.)*

I vari algoritmi differiscono per il modo con il quale i vari passi sono dettagliati. Ad esempio, nello Step 2, quando si assegnano gli individui al centro più vicino, si può calcolare il centro del nuovo cluster dopo ogni assegnazione (come nell'*algoritmo K-means di McQueen*, 1967), oppure ricalcolare i centri dei cluster dopo che tutti gli individui sono stati esaminati (come nell'*algoritmo di Forgy*, 1965). La complessità computazionale del metodo K-means è dell'ordine  $O(ndKT)$ , ove  $n$  è il numero degli individui,  $d$  il numero delle caratteristiche per ogni individuo,  $K$  il numero dei cluster richiesti e  $T$  è il numero delle iterazioni, che dipende dalla partizione iniziale e dalla dimensione del problema; in pratica, tuttavia, viene specificato una limitazione superiore a  $T$ . Per la risoluzione di problemi di clustering di grandi dimensioni, risultano particolarmente interessanti le architetture di calcolatori di tipo *parallelo*.

Come illustrazione dell'algoritmo K-means, riportiamo nel seguito una semplice implementazione in FORTRAN, rielaborata da una implementazione contenuta in Anderberg [4]. La subroutine `build` chiama i sottoprogrammi che costruiscono la partizione ottimale di un determinato numero di elementi e aumenta la dimensione della partizione spezzando uno dei cluster. La subroutine `kmeans` assegna ogni individuo in maniera ottimale. La subroutine `single` fornisce una statistica relativa ai cluster. La subroutine `output` stampa le informazioni relative ai cluster. I dati sono contenuti nella matrice `a`; più precisamente, nella prima riga, per l'indice di

colonna  $j = 2, 3, \dots, n$  sono memorizzate le label relative alle variabili; nella prima colonna, per l'indice di riga  $i = 2, 3 \dots m$ , sono memorizzate le label relative a ciascun individuo. La matrice pattern si ottiene quindi per  $i = 2, \dots, m$  e  $j = 2, \dots, n$ . Pertanto, il numero di individui è dato da  $m-1$  e il numero delle variabili da  $n-1$ . Le variabili `mm` e `nm` servono per dimensionare opportunamente l'array `a`. Se i dati sono scalati in maniera da avere media nulla e varianza 1, le medie e le varianze dei dati iniziali sono contenute nei vettori `aver` e `sdev`, che nella subroutine `output` vengono utilizzati per fornire le informazioni sulle variabili originarie. La variabile `k` indica il numero di cluster richiesto; nel programma si ha  $k \leq km$ . L'array di lavoro `sum` viene utilizzata per memorizzare le informazioni relative ai singoli cluster. L'algoritmo incomincia considerando il cluster formato da tutti gli individui e aggiunge ad ogni passo un cluster spezzando il cluster di massima varianza. La variabile `xmiss` indica i dati omessi. Il vettore `nclus` specifica per ogni individuo il numero del cluster a cui l'individuo appartiene. Il vettore `dclus` contiene la distanza di ogni individuo dal cluster più vicino. La variabile `iter` indica il numero massimo di iterazioni (aggiustamenti successivi dei cluster) per un numero fissato di cluster.

```

subroutine build(a,m,n,k,sum,xmiss,nclus,dclus,x,iter,
& aver,sdev,mm,nm,km,avre)
dimension sum(8,nm,km),a(mm,nm),x(nm),nclus(mm),dclus(mm)
dimension aver(1),sdev(1)
dimension avre(nm,km)
common/str/passw
passw=0.
do 20 i=1,8
  do 20 j=2,n
    do 20 kk=1,k
20      sum(i,j,kk)=0.
  kl=k
do 10 kk=1,kl
  do 14 nc=1,iter
    err=0.
    do 13 kkk=1,kk
      do 13 j=2,n
        if(nc.eq.1.or.sum(1,j,kkk).ne.sum(3,j,kkk))err=1
13      continue
    if(err.eq.0.)go to 15
    do 16 kkk=1,kk
    do 16 j=2,n
      sum(2,j,kkk)=0.
16      sum(1,j,kkk)=sum(3,j,kkk)
c deposito di ogni caso nel vettore x e chiamata di kmeans
  do 11 i=2,m
    do 12 j=2,n
12      x(j)=a(i,j)
      nclus(i)=nc
      call kmeans(n,kk,sum,x,nclus(i),dclus(i),xmiss,mm,nm,km)
11      continue

```

```

14     continue
15     continue
      call output(m,n,kk,sum,a,nclus,dclus,aver,sdev,mm,nm,km,avre)
c  ricerca del cluster km di massima varianza
      do 300 jj=1,kk
          if(sum(2,2,jj).eq.0)stop
300    continue
      sm=0.
      do 30 j=2,n
          do 30 kkk=1,kk
              if(sum(4,j,kkk).lt.sm)go to 30
              sm=sum(4,j,kkk)
              jm=j
              km=kkk
30    continue
      kn=kk+1
      do 31 jj=2,n
          sum(2,jj,kn)=0.
          sum(2,jj,km)=0.
          sum(3,jj,km)=0
31    sum(3,jj,kn)=0.
      do 32 i=2,m
          if(nclus(i).ne.km)go to 32
c  calcolo del baricentro del nuovo cluster e
c  aggiornamento del baricentro di km
      do 33 jj=2,n
          if(a(i,jj).eq.xmiss)go to 33
          if(a(i,jj).lt.sum(1,jj,km))go to 34
          sum(2,jj,kn)=sum(2,jj,km)+1
          sum(3,jj,kn)=sum(3,jj,km)+a(i,jj)
          go to 33
34    sum(2,jj,km)=sum(2,jj,km)+1
          sum(3,jj,km)=sum(3,jj,km)+a(i,jj)
33    continue
32    continue
      do 35 jj=2,n
          if(sum(2,jj,kn).ne.0.)sum(3,jj,kn)=sum(3,jj,kn)/sum(2,jj,kn)
          if(sum(2,jj,km).ne.0.)sum(3,jj,km)=sum(3,jj,km)/sum(2,jj,km)
35    continue
10    continue
      return
      end
      subroutine kmeans(n,k,sum,x,jmin,dmin,xmiss,mm,nm,km)
c  calcolo delle distanze di x da ogni baricentro dei cluster
c  e assegnazione di x al cluster piu' vicino. Aggiornamento
c  per tale cluster della media, della deviazione standard e del
c  minimo e del massimo rispetto ad ogni variabile.
      dimension sum(8,nm,km),x(nm)
      jmin=1
      dmin=1.e4
      do 20 j=1,k

```

```

        xp=1.e-10
        dd=0.
        do 21 i=2,n
            if(x(i).eq.xmiss)go to 21
            dd=dd+(x(i)-sum(1,i,j))**2
            xp=xp+1
21        continue
            dd=(dd/xp)**0.5
            if(dd.gt.dmin)go to 20
            dmin=dd
            jmin=j
20        continue
        do 31 i=2,n
            if(x(i).eq.xmiss)go to 31
30        call single(x(i),sum(2,i,jmin),sum(3,i,jmin),sum(4,i,jmin),
&            sum(5,i,jmin),sum(6,i,jmin),sum(7,i,jmin))
31        continue
        return
        end
        subroutine single(x,count,ave,sd,xmin,xmax,ssq)
c  aggiorna per il cluster jmin la media ave, la deviazione standard sd,
c  il valore minimo xmin, massimo xmax e il numero di osservazioni count
        if(count.ne.0.)go to 10
        ave=0
        sd=0
        xmin=1.e20
        xmax=-xmin
        ssq=0.
10        count=count+1.
        ave=ave+(x-ave)/count
        if(count.ne.1.)ssq=ssq+count*(x-ave)**2/(count-1.)
        sd=(ssq/count)**0.5
        if(xmin.gt.x)xmin=x
        if(xmax.lt.x)xmax=x
        return
        end
        subroutine output(m,n,kk,sum,a,nclus,dclus,aver,sdev,mm,nm,km,avre)
        dimension sum(8,nm,kk),nclus(mm),dclus(mm),a(mm,nm)
        dimension aa(10),dd(10)
        dimension avre(nm,km)
        dimension aver(1),sdev(1)
        dimension r(50)
        common/str/passw
        data lc/0/
        write(*,9)kk
9        format(' overall mean square ,with ',i3,' clusters' )
        assw=0.
        do 40 j=2,n
            sd=0.
            sc=0.
            ssb=0.

```

```

    ssw=0.
    do 41 k=1,kk
        sd=sd+sum(3,j,k)*sum(2,j,k)
        avre(j,k)=sum(3,j,k)
        ssb=ssb+sum(3,j,k)**2*sum(2,j,k)
        ssw=ssw+sum(7,j,k)
41    sc=sc+sum(2,j,k)
        dfb=kk-1
        dfw=sc-dfb-1
        th=1.e-10
        if(sc.eq.0.)sc=th
        if(dfw.eq.0.)dfw=th
        if(dfb.eq.0.)dfb=th
c    assw=errore totale nella partizione
        assw=assw+ssw
        ssb=ssb-sd**2/sc
        sddsc=(sd**2)/sc
        ssb=ssb/dfb
        ssw=ssw/dfw
        if(ssw.eq.0.)ssw=th
        ratio=0.
c    ratio=rapporto di varianza per la variabile j
        if(lc.ne.0)ratio=(r(j)/ssw-1)*(1+dfw)+1
        r(j)=ssw
        scsa=a(1,j)
        idfw=dfw
        idfb=dfb
        write(*,8)scsa,ssw,idfw,ssb,idfb,ratio
8        format(1h ,a4,1pe13.2,'wmsq',i4,'wdf',1pe13.2,'bmsq',i4,' bdf'
& ,e12.3,'fr.')
40    continue
        write(*,10)assw
10    format('overall within sum of squares',1pe13.2)
        lc=lc+1
        omsr=(passw/assw-1)*dfw
        write(*,66)omsr
66    format(' ',///' overall mean square ratio',e15.4,///)
c    il valore di omsr misura la riduzione della varianza entro i cluster
c    passando dalla partizione k alla partizione k+1; puo' essere utilizzato
c    come indice di significativita' della suddivisione; valori grandi
c    giustificano il passaggio da k a k+1 cluster. In maniera approssimata
c    segue la distribuzione F con gradi di liberta' dati da n e (m-k-1)n.
        passw=assw
2        format(' cluster members with their dist. to the cluster centre '/3x)
        do 20 k=1,kk
            write(*,11)
11            format(1x,71(1h-))
            write(*,1)k,kk
1            format(i5,' th cluster of',i5)
            write(*,2)
            l=0

```

```

do 21 i=2,m
  if(nclus(i).ne.k)go to 22
  l=l+1
  ddd=0.
  do 300 j=2,n
300    ddd=ddd+((a(i,j)-sum(1,j,k))*sdev(j))**2
    dd(1)=ddd**0.5
22    if(1.lt.5.and.i.lt.m)go to 21
    if(1.eq.0)go to 21
    write(*,3)(a(11,1),11=1,1)
3    format(1h0,2x,5(7x,a4))
    write(*,12)(dd(11),11=1,1)
12    format(1h ,2x,1p5e11.2)
    l=0
21    continue
    write(*,4)
4    format( ' summary statistics for the cluster')
    write(*,5)
5    format(' label , centre , count , ave., sd , xmi ,xmax ,ssq')
    do 30 j=2,n
      aj=sdev(j)*sum(1,j,k)+aver(j)
      aj5=aver(j)+sdev(j)*sum(5,j,k)
      aj6=aver(j)+sdev(j)*sum(6,j,k)
      aj4=sum(4,j,k)*sdev(j)
      aj3=aver(j)+sdev(j)*sum(3,j,k)
      scsa=a(1,j)
      irt=sum(2,j,k)
30    write(*,6)scsa,aj,irt,aj3,aj4,aj5,aj6,sum(7,j,k)
6    format(1x,a4,1pe11.3,i4,1p5e11.3)
20    continue
    return
end

```

Come esempio di applicazione, riportiamo nel seguito i risultati ottenuti nell'analisi di alcuni dati clinici in neuropsicobiologia (cfr. [39]).

```

numero dei casi      = 59
numero della variabili = 3
numero massimo dei cluster = 3

```

casi	umht	snas	hato	casi	umht	snas	hato
1	1498	.180	17	31	1208	.110	28
2	1619	.110	22	32	653	.100	16
3	1285	.140	12	33	773	.210	18
4	1415	.140	19	34	1185	.100	22
5	1142	.100	18	35	889	.220	17
6	1502	.140	16	36	1521	.100	16
7	1114	.120	20	37	1060	.220	17
8	1478	.100	19	38	1570	.180	20
9	1355	.100	21	39	2678	.100	20
10	844	.100	15	40	1513	.100	17
11	1198	.100	21	41	1628	.100	21
12	1115	.210	22	42	1212	.310	16



13	1590	.250	16	43	1094	.100	11
14	1310	.230	14	44	1331	.120	26
15	1406	.360	21	45	1628	.100	18
16	1428	.100	22	46	1869	.220	21
17	746	.100	16	47	1257	.320	15
18	613	.100	22	48	1592	.330	28
19	407	.370	19	49	1806	.100	22
20	634	.230	22	50	1217	.350	23
21	967	.170	17	51	1541	.230	24
22	1554	.100	21	52	1374	.240	32
23	1090	.100	15	53	3388	.220	22
24	1958	.100	18	54	1386	.170	24
25	985	.170	17	55	1786	.110	17
26	717	.100	23	56	2511	.260	19
27	1395	.100	16	57	1570	.100	19
28	2149	.140	24	58	1092	.180	21
29	868	.110	23	59	1509	.110	19
30	1170	.110	21				

studio statistico

	media	deviazione standard	minimo	massimo
umht	1362.593	503.632	407.00	3388.00
snas	.162	.078	.10	.37
hato	19.627	3.868	11.00	32.00

matrice di correlazione			autovalori		autovettori			
	umht	snas	hato					
				1	1.177	0.532	-0.648	-0.544
umht	1.00	-.02	.14	2	1.018	0.438	0.761	-0.477
snas	-.02	1.00	.12	3	0.804	0.723	0.015	0.689
hato	.14	.12	1.00					

overall mean square with 3 cluster

umht	7.55E-01	wmsq	56	wdf	7.86E+00	bmsq	2	bdf	1.393E+01	fr.
snas	3.04E-01	wmsq	56	wdf	2.05E+01	bmsq	2	bdf	4.805E+01	fr.
hato	7.81E-01	wmsq	56	wdf	7.12E+00	bmsq	2	bdf	2.395E+00	fr.

overall within sum of squares 1.03E+02

overall mean square ratio .1467E+02

-----  
1th cluster of 3

cluster members with their distance to the cluster centre

1	2	3	4	5	6	7
1.52E+02	2.73E+02	6.14E+01	6.90E+01	2.04E+02	1.56E+02	2.32E+02
8	9	10	11	16	17	18
1.32E+02	9.18E+00	5.02E+02	1.48E+02	8.21E+01	6.00E+02	7.33E+02
22	23	24	26	27	29	30
2.08E+02	2.56E+02	6.12E+02	6.29E+02	4.91E+01	4.78E+02	1.76E+02
31	32	34	36	38	39	40
1.38E+02	6.93E+02	1.61E+02	1.75E+02	2.24E+02	1.33E+03	1.67E+02
41	43	44	45	49	54	55
2.82E+02	2.52E+02	1.64E+01	2.82E+02	4.60E+02	4.03E+01	4.40E+02
57	58	59				
2.24E+02	2.54E+02	1.63E+02				

summary statistics for the cluster

label	centre	count	ave.	sd	xmi	xmax	ssq
umht	1.346E+03	38	1.346E+03	3.887E+02	6.130E+02	2.678E+03	2.263E+01

```
snas 1.139E-01 38 1.139E-01 2.455E-02 1.000E-01 1.800E-01 3.696E+00
hato 1.932E+01 38 1.932E+01 3.480E+00 1.100E+01 2.800E+01 3.076E+01
```

```
-----
2th cluster of 3
cluster members with their distance to the cluster centre
```

```
      12      13      14      19      20      21      25
9.85E+01 5.73E+02 2.93E+02 6.10E+02 3.83E+02 4.96E+01 3.16E+01
      33      35      37      42      47
2.44E+02 1.28E+02 4.34E+01 1.95E+02 2.40E+02
```

```
summary statistics for the cluster
label centre count ave. sd xmi xmax ssq
umht 1.017E+03 12 1.017E+03 3.054E+02 4.070E+02 1.590E+03 4.414E+00
snas 2.425E-01 12 2.425E-01 5.833E-02 1.700E-01 3.700E-01 6.586E+00
hato 1.750E+01 12 1.750E+01 2.363E+00 1.400E+01 2.200E+01 4.478E+00
```

```
-----
3th cluster of 3
cluster members with their distance to the cluster centre
```

```
      15      28      46      48      50      51      52
4.88E+02 2.55E+02 2.53E+01 3.02E+02 6.77E+02 3.53E+02 5.20E+02
      53      56
1.49E+03 6.17E+02
```

```
summary statistics for the cluster
label centre count ave. sd xmi xmax ssq
umht 1.894E+03 9 1.894E+03 6.550E+02 1.217E+03 3.388E+03 1.522E+01
snas 2.611E-01 9 2.611E-01 6.822E-02 1.400E-01 3.600E-01 6.758E+00
hato 2.378E+01 9 2.378E+01 3.765E+00 1.900E+01 3.200E+01 8.525E+00
```

Nella Figura 9.13 è rappresentata la suddivisione in 3 cluster nel piano individuato dai due autovettori corrispondenti ai primi due autovalori della matrice di covarianza; sono inoltre rappresentati i baricentri corrispondenti ai tre gruppi.

### 9.2.5 Fuzzy clustering

Gli algoritmi descritti in precedenza assegnano ogni individuo ad un unico cluster, ossia gli individui sono partizionati in insiemi disgiunti. Questo è ragionevole quando i cluster ottenuti presentano una forma compatta e sono “ben separati”.

In pratica, tuttavia, i cluster possono “toccarsi” o avere sovrapposizioni, nel qual caso l’assegnazione degli individui risulta maggiormente difficoltosa. Con riferimento alla Figura 9.14, è chiaro che gli individui  $\mathbf{x}_i$  e  $\mathbf{x}_k$  sono da pensare in cluster differenti, mentre l’individuo  $\mathbf{x}_j$  potrebbe appartenere sia al cluster che contiene  $\mathbf{x}_i$  che a quello contenente  $\mathbf{x}_i$ . Si dice che tali tipi di cluster contengono frontiere *fuzzy* (confuse, indistinte).

Nell’ambito della teoria degli *insiemi fuzzy*<sup>4</sup> un oggetto può appartenere a un

<sup>4</sup>Le ricerche sugli insiemi fuzzy hanno avuto origine, in sostanza, dal lavoro *Fuzzy Sets* di L. A. Zadeh (1965). Alcune delle maggiori applicazioni degli insiemi fuzzy riguardano la teoria dei controlli, l’intelligenza artificiale, e l’utilizzo delle banche dati (fuzzy information retrieval). Si veda, ad esempio Miyamoto [116].

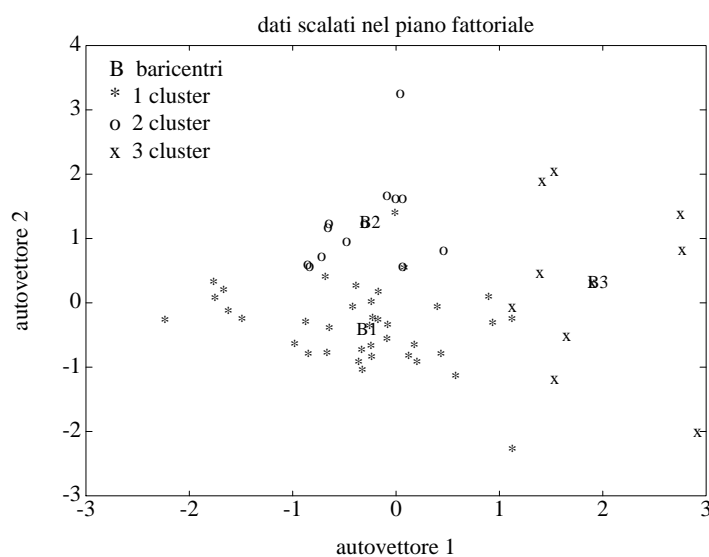


Figura 9.13: Rappresentazione dei dati scalati nel piano degli autovettori corrispondenti ai primi due autovalori della matrice di covarianza.

cluster con un *grado di appartenenza*. Il grado di appartenenza assume un valore nell'intervallo  $[0, 1]$ . Per una maggiore comprensione dell'idea, ricordiamo i concetti principali relativi alla teoria fuzzy.

**Insiemi crisp e insiemi fuzzy** Se  $A$  indica un sottoinsieme di un insieme  $S$ , nella teoria classica degli insiemi ogni elemento  $x \in S$  è tale che o  $x$  appartiene ad  $A$ , oppure  $x$  non appartiene ad  $A$ . Il sottoinsieme  $A$  può essere, quindi, rappresentato da una funzione  $f_A : S \rightarrow \{0, 1\}$ , definita nel seguente modo e detta *funzione*

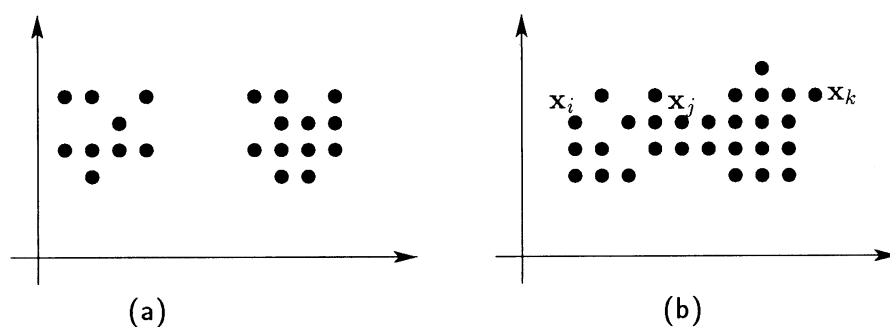


Figura 9.14: Esempi di strutture di cluster: (a) cluster ben separati; (b) cluster di tipo fuzzy.

caratteristica del sottoinsieme  $A$

$$f_A(x) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases}$$

Gli insiemi fuzzy sono introdotti generalizzando opportunamente la definizione della funzione caratteristica  $f_A$ . Sia  $\mu_B$  una funzione definita su  $S$  con valori nell'intervallo  $[0, 1]$ , ossia  $\mu_B : S \rightarrow [0, 1]$ , ove  $B$  è un'etichetta (label) della funzione. Si dice che la label  $B$  è un fuzzy set, quando si ha una determinata interpretazione di tale label. Chiariamo la definizione con un semplice esempio.

► **Esempio 9.4** Consideriamo come  $S$  l'insieme degli interi non negativi che rappresentano le età di una popolazione e cerchiamo di definire un sottoinsieme  $B$  che implica *essere giovani*. Mentre è chiaro che gli individui di cinque, dodici anni sono giovani, e quelli di trenta o quaranta anni sono meno giovani, non vi è un criterio definito che separa i giovani dai vecchi. Definiamo, ora, una funzione  $\mu_B$  che corrisponde al concetto di *giovane*. In altre parole, per  $x \in S$ ,  $\mu_B(x)$  mostra il grado di rilevanza dell'età  $x$  al concetto *giovane*. Si può, ad esempio, porre (cfr. Figura 9.15)

$$\mu_B(x) = \begin{cases} 1 & \text{se } 0 \leq x < 20 \\ (50 - x)/30 & \text{se } 20 \leq x < 50 \\ 0 & \text{se } 50 \leq x \end{cases}$$

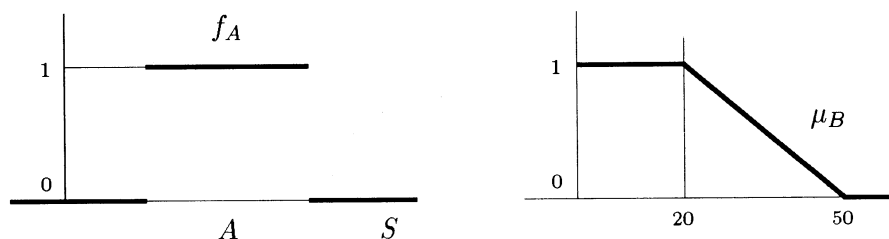


Figura 9.15: Rappresentazione di un insieme crisp e di un insieme fuzzy.

Un fuzzy set  $B$  è, pertanto, un grado di pertinenza degli elementi a un concetto rappresentato da una label  $B$ . La funzione  $\mu_B(x)$  è una generalizzazione della funzione caratteristica  $f_A(x)$  di un sottoinsieme ordinario (che possiamo chiamare *crisp*, distinto). Se  $\mu_B(x) = 1$ , allora  $x$  appartiene certamente all'insieme  $B$ ; se  $\mu_B(x) = 0$ , allora  $x$  non appartiene a  $B$ , mentre se  $0 < \mu_B(x) < 1$  l'appartenenza di  $x$  a  $B$  è ambigua. Si ha, comunque, che se  $\mu_B(x_1) > \mu_B(x_2)$ , allora la pertinenza di  $x_1$  al concetto rappresentato da  $B$  è maggiore di quella relativa a  $x_2$ . Un altro

modo di interpretare la label  $B$  è il seguente. Se  $\alpha$  è un parametro appartenente all'intervallo  $[0, 1]$ , si definisce taglio- $\alpha$  ( $\alpha$ -cut), o livello  $\alpha$ , il seguente insieme crisp

$$B_\alpha = \{x \mid \mu_B(x) \geq \alpha, x \in S\}$$

Allora, un insieme fuzzy  $B$  è interpretato come una famiglia di  $\alpha$ -cuts:  $\{B_\alpha\}_{\alpha \in [0,1]}$ .

Concludiamo la definizione di fuzzy set, osservando che ogni insieme crisp è un fuzzy set; in particolare, l'insieme intero  $S$  è un insieme fuzzy, la cui funzione di appartenenza è  $\mu_S(x) = 1$  per ogni  $x \in S$ . Allo stesso modo, l'insieme vuoto  $\emptyset$  è un insieme fuzzy, con  $\mu_\emptyset = 0$  per ogni  $x \in S$ . Lasciamo come utile esercizio l'estensione agli insiemi fuzzy delle definizioni di inclusione, di unione, di intersezione e di complemento.

**Algoritmi di clustering fuzzy** Nel caso di clustering fuzzy, il generico individuo  $\mathbf{x}_i$  ha un grado di appartenenza  $f_q(\mathbf{x}_i) \geq 0$ , al cluster  $q$ -mo. Più è grande il valore di  $f_q(\mathbf{x}_i)$  e più grande è la presunzione che  $\mathbf{x}_i$  appartenga al cluster  $q$ . Naturalmente,  $\sum_q f_q(\mathbf{x}_i) = 1$  e quando  $f_q(\mathbf{x}_i) = 1$  l'individuo  $\mathbf{x}_i$  appartiene al cluster  $q$  con assoluta certezza. Sottolineiamo, da una parte la natura *soggettivistica* del grado di appartenenza (che, basata su definizioni piuttosto che su misure, rappresenta il punto cruciale della teoria) e dall'altra la differenza tra l'approccio mediante la teoria degli insiemi fuzzy e l'approccio probabilistico. In effetti, in un contesto probabilistico, l'individuo  $\mathbf{x}_i$  appartiene ad un unico cluster, in dipendenza dal risultato di un esperimento casuale, mentre in clustering fuzzy, ogni individuo può appartenere a due cluster simultaneamente.

I metodi clustering visti in precedenza possono essere generalizzati introducendo una appropriata definizione del criterio, che tenga conto anche della funzione di appartenenza. L'output di un algoritmo fuzzy includerà insieme alla partizione una ulteriore informazione riguardante, per ogni individuo, i valori di appartenenza. Per una panoramica sugli algoritmi clustering fuzzy si veda ad esempio Miyamoto [116].

### 9.2.6 Clustering software

Sono disponibili numerosi package contenenti algoritmi di clustering; per una panoramica si può vedere ad esempio [9]. Per comodità, il software disponibile può essere raggruppato nelle seguenti tre categorie.

1. *Raccolta di subroutine e algoritmi.* Sono programmi usualmente sviluppati da gruppi di ricerca. Segnaliamo in particolare le raccolte contenute nei libri Anderberg [4] e Hartigan [79]. Alcune subroutine sono disponibili attraverso le librerie NAG e IMSL.
2. *Package statistici generali che contengono metodi clustering.* Segnaliamo, in particolare, i package BMDP, SAS e OSIRIS, di cui esistono versioni su personal computer.

3. *Package di cluster analysis specializzati.* Segnaliamo in particolare la collezione di programmi in Fortran chiamata CLUSTAN, che contiene un'ampia raccolta di tecniche di clustering. Altri package di interesse sono NT-SYS, CLUS, TAXON, BC-TRY, ICICLE.

### 9.3 Validazione del clustering

Come abbiamo già osservato in precedenza, nella cluster analysis è disponibile un campione di oggetti di classificazione incognita e lo scopo è quello di raggruppare tali oggetti in classi naturali, o cluster. Il fatto che non vi sia una classificazione *a priori* del campione suggerisce che la cluster analysis è, in sostanza, uno strumento per l'analisi dei dati; in altre parole, si desidera studiare i dati per vedere se effettivamente esiste un raggruppamento *naturale e utile*. Un aspetto importante da sottolineare è il fatto che, in generale, su un campione possono essere imposti (ipotizzati) vari tipi di classificazione. Per esemplificare, un campione di individui umani può essere classificato per il sesso, oppure per classi sociali, per il colore della pelle, per l'età. La lista potrebbe continuare, ed è ovvio che il tipo di raggruppamento che emerge da un'analisi dipenderà in maniera essenziale dalle variabili utilizzate per rappresentare l'oggetto. Questo è uno dei punti cruciali, in quanto una scelta non opportuna delle variabili può portare a un clustering senza utilità per un determinato scopo.

Un altro aspetto importante nell'applicazione della tecnica di cluster, e che abbiamo discusso in precedenza, riguarda il modo con il quale sono rappresentate le variabili, in particolare il modo con il quale esse vengono scalate. Un problema collegato al precedente riguarda la definizione di "vicinanza" tra due oggetti, con l'introduzione della matrice di prossimità. Come abbiamo visto, tale matrice è la base di partenza per tutte le tecniche numeriche di clustering.

Un successivo problema riguarda la scelta di un particolare algoritmo. Come avviene spesso nel calcolo numerico, ma qui in particolare, non esiste in assoluto l'algoritmo *migliore*. L'opportunità di certe tecniche rispetto ad altre dipende, in effetti, dal tipo particolare di dati a disposizione e dagli scopi della ricerca. Una strategia conveniente può consistere, quindi, nell'applicazione di differenti algoritmi, con un'analisi comparativa dei risultati ottenuti. Quest'ultimo punto introduce un altro aspetto fondamentale nell'utilizzo della cluster analysis, ossia la *validazione* dei risultati ottenuti dall'applicazione di un determinato algoritmo. In sostanza, si tratta di analizzare la validità di una struttura clustering, ossia di una determinata gerarchia, o partizione. Ricordiamo che, tagliando a diversi livelli un dendrogramma, oppure fissando un numero diverso di cluster in un algoritmo di partizione, è possibile ottenere una grande varietà di classificazioni. La scelta di una classificazione particolare è, in effetti, uno dei problemi più difficili e delicati della cluster analysis<sup>5</sup>. Per una discussione più adeguata di tale questione rinviamo, ad esem-

---

<sup>5</sup>“without a strong effort in this direction, cluster analysis will remain a black art accessible only

---

pio, a Jain e Dubes [93], ove, in particolare, il problema è affrontato dal punto di vista statistico, con tecniche simili a quelle che abbiamo analizzato nei paragrafi precedenti nell'ambito della validazione di una ipotesi statistica.

---

*to those true believers who have experience and great courage” (A. K. Jain, R. C. Dubes).*

Say not, "I have found the truth",  
but rather, "I have found a truth".  
(Kahlil Gibran)

## Capitolo 10

# Metodo Monte Carlo

In maniera schematica, il *metodo Monte Carlo*, o metodo delle prove statistiche (statistical trials), consiste nell'approssimazione della soluzione di un problema matematico mediante simulazioni basate sull'utilizzo di numeri casuali (*simulazioni stocastiche*). La soluzione dei problemi numerici mediante tale metodo è quindi, nello spirito, più vicina al metodo sperimentale che ai metodi numerici classici. In effetti, l'errore ottenuto con il metodo Monte Carlo non può essere, in generale, stimato a priori, ma è valutato mediante l'analisi della deviazione standard delle quantità che vengono simulate; inoltre, la soluzione potrebbe non essere riproducibile nei suoi dettagli. Introduciamo l'idea del metodo attraverso una prima applicazione, nota come metodo Monte Carlo *hit or miss*<sup>1</sup>. L'analisi di tale metodo verrà approfondita nel seguito, insieme all'analisi di procedure alternative.

► **Esempio 10.1** Per approssimare l'area di un cerchio di raggio 1, ossia il numero  $\pi$ , si può seguire la seguente procedura. Consideriamo un numero  $N$  di coppie di numeri  $(r_1, r_2)$ , ognuno dei quali estratti a caso, con una distribuzione uniforme sull'intervallo  $[0, 1]$ . Ogni coppia è rappresentata in coordinate cartesiane da un punto del quadrato di lato 1 (cfr. Figura 10.1). Si assume, quindi, come *stima* di  $\pi/4$  il rapporto  $p_N$  tra il numero dei punti tali che  $r_1^2 + r_2^2 \leq 1$  e il numero totale  $N$  dei punti. Come semplice illustrazione, in Figura 10.1 sono riportate le coppie di punti che si ottengono dalla Tabella 10.1 di *numeri casuali*. Precisamente, le coordinate di ciascun punto sono calcolate nel seguente modo. Prendendo, ad esempio, il primo numero 26099, si considerano le prime quattro cifre 2609 e si assume  $r_1 = 0.26$  e  $r_2 = 0.09$ . Procedendo in maniera analoga per ciascuno dei 45 numeri della Tabella 10.1, si ottengono i punti riportati nella Figura 10.1: 31 di essi appartengono al

---

<sup>1</sup>Historically, *hit or miss* methods were once the ones most usually propounded in explanation of Monte Carlo techniques; they were of course, the easiest methods to understand (particularly if explained in the kind of graphical language involving a curve in a rectangle) Hammersley, Handscomb (1964).



quarto di cerchio di raggio 1. In corrispondenza, si ottiene la stima  $p_{45} = 31/45 \approx 0.689$ , mentre  $\pi/4 \approx 0.785$ ; si ha quindi un errore del 12.3%. Come vedremo nel seguito, tale stima può essere migliorata aumentando il numero  $N$  dei tentativi, in particolare vedremo che, essendo ogni esperimento un prova di Bernoulli con probabilità  $p = \pi/4$  di successo, la varianza  $\sigma^2$ , che misura la dispersione intorno al valore medio  $p$  delle successive stime, è data da  $\sigma^2 = p(1 - p)/N$ . ■

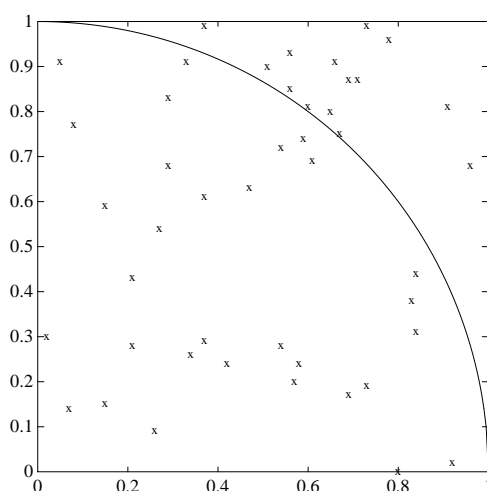


Figura 10.1: Illustrazione del metodo Monte Carlo *hit or miss* applicato al calcolo dell'area di un quarto di cerchio di raggio 1.

All'inizio del secolo il metodo Monte Carlo fu utilizzato per analizzare l'equazione di Boltzmann; nel 1908 esso venne utilizzato dallo statistico W. S. Gosset, noto con lo pseudonimo di *Student*, per stimare il coefficiente di correlazione nella *distribuzione t* (cfr. Capitolo 8). Il termine *Monte Carlo* venne introdotto da Von Neumann e Ulam durante la 2<sup>a</sup> guerra mondiale come parola in codice per indicare il lavoro segreto condotto a Los Alamos<sup>2</sup>; il metodo venne quindi utilizzato, in particolare, per la simulazione della diffusione di neutroni nella fissione nucleare. In seguito, il metodo è stato ampiamente utilizzato per approssimare il valore di *integrali multipli*, e la soluzione di *equazioni differenziali* e *integrali* originate dalla fisica e non risolubili in maniera analitica. Attualmente, il metodo Monte Carlo rappresenta una delle tecniche più interessanti per l'analisi di sistemi *complessi*, e trova quindi applicazione in svariati campi. Lo scopo di questo capitolo è quello di introdurre le idee di base del metodo; come bibliografia utile ad approfondire le nozioni di seguito introdotte si veda in particolare Rubinstein [139].

<sup>2</sup>Il nome Monte Carlo appare per la prima volta nel lavoro N. Metropolis, S. Ulam, *The Monte Carlo method*, J. Amer. statistical assoc., **44**, 247, 335–341 (1949).

---

26099	65801	69870	84446	58248	21282	56938	54729	67757
71874	61692	80001	214306	02305	59741	34262	15157	27545
08774	29689	42245	51903	69179	96682	91819	60812	47631
37294	92028	56850	83380	05912	29830	37612	15593	73198
33912	37996	78967	57201	66916	73998	54289	07147	84313

---

Tabella 10.1: Numeri casuali.

## 10.1 Numeri casuali e pseudo-casuali

Come evidenziato dal precedente Esempio 10.1, un aspetto di base nell'applicazione del metodo di Monte Carlo è costituito dalla disponibilità di successioni di *numeri casuali* (*random variate*). In questo paragrafo analizzeremo, pertanto, alcune tecniche per generare tali successioni, rinviando per una trattazione più approfondita ad esempio a Knuth [100], Rubinstein [139], Devroye [48].

### 10.1.1 Distribuzioni uniformi

Un primo modo per generare numeri casuali con *distribuzione uniforme*, utilizzato in particolare nel passato prima dell'affermarsi dei calcolatori, consiste nel successivo lancio di una moneta, o di un dado, o nell'utilizzo di altri processi fisici. Le successioni generate in tal modo presentano per le applicazioni l'inconveniente della *non riproducibilità*, oltre che della lentezza con cui sono ottenute. Tali inconvenienti possono essere in parte superati mediante la costruzione di opportune tabelle<sup>3</sup>, memorizzate successivamente nella memoria di un calcolatore; un esempio di tali tabelle è riportato in Tabella 10.1. Il vantaggio di questo metodo è la riproducibilità; il suo svantaggio, in particolare per quanto si riferisce al suo utilizzo nell'ambito del metodo di Monte Carlo, nel quale è richiesta in generale una grande quantità di numeri casuali, è ancora la sua lentezza e il rischio che la tabella venga esaurita.

L'idea per superare le difficoltà precedenti consiste nell'utilizzare per la generazione dei numeri casuali le operazioni aritmetiche di un calcolatore. Un primo metodo in questa direzione, noto come metodo *mid-square*, venne suggerito da von Neumann nel 1951. In maniera schematica, il metodo consiste nel prendere il quadrato del numero casuale precedente e nell'estrarre le cifre centrali; ad esempio, se si generano numeri casuali con quattro cifre e si è arrivati al numero 5232, si calcola il suo quadrato 27 373 824 e si assume come successivo numero casuale il numero 3738. Essendo tale procedura completamente deterministica, i numeri da essa generati non sono in realtà casuali<sup>4</sup>, e per tale motivo essi vengono usualmente chiamati *numeri*

<sup>3</sup>Segnaliamo in particolare la ben nota tavola pubblicata dalla RAND Corporation nel 1955 e contenente un milione di cifre.

<sup>4</sup>*Anyone who considers arithmetical methods of producing random digits is, of course, in a state*

*pseudo-casuali*. Quanto la distribuzione di tali numeri, e analogamente di quelli generati dalle procedure che esamineremo nel seguito, si discosta dalla distribuzione uniforme viene valutato da opportuni test statistici, in particolare dal test basato sulla distribuzione chi-quadrato. Ricordiamo, tuttavia, che il successo di un test statistico indica solo che *non si hanno motivi per rifiutare l'ipotesi che i numeri pseudo-casuali siano distribuiti uniformemente*; ma questo *non è una garanzia della validità di tale ipotesi* (per un approfondimento di questi aspetti si veda ad esempio Knuth [100]).

Attualmente, le procedure più comunemente utilizzate per la generazione di numeri pseudo-casuali con distribuzione uniforme sono basate su una relazione di congruenza della forma seguente

$$x_i = (ax_{i-1} + c) \pmod{m}, \quad i = 1, 2, \dots \quad (10.1)$$

ove il *moltiplicatore*  $a$ , l'incremento  $c$ , e il *modulo*  $m$  sono interi non negativi. Ricordiamo che l'operazione modulo  $(\text{mod } m)$  significa che

$$x_i = ax_{i-1} + c - mk_i \quad (10.2)$$

ove  $k_i = [(ax_{i-1} + c)/m]$  indica la parte intera di  $(ax_{i-1} + c)/m$ ; ossia,  $x_i$  rappresenta il resto della divisione di  $ax_{i-1} + c$  per  $m$ .

Naturalmente, la successione generata dalla relazione (10.1) a partire da un intero iniziale  $x_0$ , usualmente chiamato *seed*, si ripete in al più  $m$  passi ed è quindi *periodica*, con periodo di lunghezza  $p \leq m$ . Per esempio, se  $a = c = x_0 = 3$  e  $m = 5$ , allora la successione generata dalla relazione (10.1) è  $x_i = 3, 2, 4, 0, 3$  con periodo  $p = 4$ ; se  $m = 16$ ,  $a = 3$ ,  $c = 1$  e  $x_0 = 2$ , si ha la successione

$$2, 7, 6, 3, 10, 15, 14, 11, 2, 7, \dots$$

con periodo  $p = 8$ . Per le applicazioni si ha interesse a scegliere i parametri  $a, c$  e  $m$  in maniera che il periodo  $p$  sia il più lungo possibile. Quando  $p = m$ , si dice che il generatore (10.1) ha un periodo completo (*full period*). Per delle condizioni necessarie e sufficienti affinché il generatore definito in (10.1) abbia un periodo completo  $m$ , si veda ad esempio Knuth [100]. In particolare, ricordiamo che gli interi  $c$  e  $m$  non devono avere divisori in comune. La costruzione di efficienti generatori di numeri pseudo-casuali è pertanto dipendente dal calcolatore, nel senso che utilizza la particolare aritmetica implementata su un determinato calcolatore. Come esemplificazione, ricordiamo il seguente generatore, suggerito da Lehmer nel 1951 e

---

*of sin*, J. Von Neumann (1951). *A random sequence is a vague notion embodying the ideas of a sequence in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests, traditional with statisticians and depending somewhat on the uses to which the sequence is to be put*, D. H. Lehmer (1951).

utilizzato in particolare in ambiente FORTRAN sulla serie IBM/360, per la quale la lunghezza della parola è uguale a 32 bits, con 1 bit riservato al segno

$$x_i = (7^5 x_{i-1}) \bmod (2^{31} - 1) \quad (10.3)$$

Il numero intero  $x_0$  è scelto in maniera arbitraria tra 1 e il numero primo (di Mersenne)  $2^{31} - 1$ . I numeri generati sono interi nell'intervallo aperto  $(1, 2^{31} - 1)$ . Per avere una successione nell'intervallo aperto  $(0, 1)$  sarà, allora, sufficiente dividere  $x_i$  per  $2^{31} - 1$ . L'algoritmo è implementato nel seguente sottoprogramma<sup>5</sup>

```

FUNCTION RAND(L)
L=MOD(16807*L,2147483647)
RAND=REAL(L)*4.6566128752459E-10
RETURN
END

```

Per ottenere una successione di numeri casuali sul generico intervallo  $(a, b)$  si utilizzerà l'istruzione  $x=(b-a)*RAND(L)+a$ . Osserviamo che il periodo corrispondente alla successione generata da RAND è dell'ordine di  $2^{30}$  e che, d'altra parte, i test statistici hanno fornito risultati soddisfacenti<sup>6</sup>. Una conferma sperimentale è fornita dalla Figura 10.2, nella quale è rappresentata la frequenza della successione  $x_i$  ottenuta dal generatore (10.3) per  $i = 1, 2, \dots, 10\,000$ . Altri valori comunemente utilizzati sono  $a = 69069$ ,  $c = 0$  e  $m = 2^{32}$ .

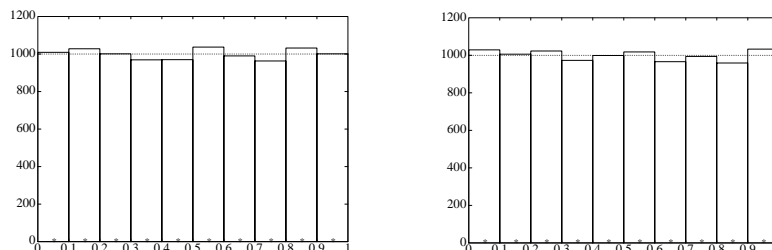


Figura 10.2: Nella prima figura è rappresentata la frequenza corrispondente alla successione  $x_i$  ottenuta con il generatore (10.3) per  $i = 1, 2, \dots, 10\,000$ ; analogamente nella seconda figura è rappresentata la frequenza corrispondente al generatore (10.4).

Concludiamo, segnalando un'altra idea interessante per generare su calcolatore successioni di numeri pseudo-casuali. In essa il generico elemento della successione è calcolato come combinazione lineare di due elementi precedenti; per tale motivo la

<sup>5</sup>Ricordiamo che in FORTRAN la funzione MOD(I,J) calcola il resto della divisione di I per J, cioè  $I - [I/J]*J$ , dove  $[I/J]$  è la parte intera del del numero reale  $I/J$ .

<sup>6</sup>si veda ad esempio, S. K. Park, K. W. Miller *Random Number Generations: Good ones are hard to find*. Communications of the ACM, vol. 31, 10(1988).

procedura è anche nota come *generatore di tipo Fibonacci*. Come esemplificazione, consideriamo il seguente algoritmo

$$x_{i+1} = x_{i-17} - x_{i-5} \quad (10.4)$$

nel quale i valori della successione dipendono da 17 valori iniziali  $x_i$ , assunti nell'intervallo  $(0, 1)$  e dalla strategia adottata quando  $x_i < 0$ ; una strategia standard consiste nel porre  $x_i + 1.0 \rightarrow x_i$ ; in questo modo tutti gli elementi della successione sono nell'intervallo unitario. Dal momento che  $x_{i+1}$  dipende dai 17 valori precedenti, il periodo teorico  $p$  è dato da  $p = (2^{17} - 1)2^n$ , ove  $n$  è il numero di bits nella parte frazionaria di  $x_i$ , cioè la mantissa. Ad esempio, in aritmetica floating-point (IEEE) a 32-bit si ha  $n = 24$ , e quindi  $p \approx 2 \cdot 10^{12}$ . In Figura 10.2 è rappresentata la frequenza per  $i = 1, 2, \dots, 10\,000$ .

### 10.1.2 Numeri casuali secondo una distribuzione assegnata

Nelle applicazioni è spesso necessario generare numeri casuali  $\xi$  distribuiti secondo una densità di probabilità assegnata  $f(x)$ , non necessariamente uniforme sull'intervallo  $(a, b)$ , ove  $a$  e/o  $b$  possono essere infiniti. Ricordiamo (cfr. Capitolo 8) che questo vuol dire che la probabilità che  $\xi$  sia nell'intervallo  $(x, x + dx)$  è data da  $f(x) dx$ . Per la distribuzione uniforme su  $(0, 1)$  si ha  $f(x) = 1$  su  $(0, 1)$  e  $f(x) = 0$  altrove.

I numeri  $\xi$  possono essere costruiti a partire dai numeri casuali su  $(0, 1)$  in diversi modi. Incominciamo da quello più diretto, chiamato anche *metodo della trasformata inversa*.

**Metodo della trasformata inversa** Se  $F(x)$  è la funzione di distribuzione, o di ripartizione, corrispondente alla densità di probabilità  $f(x)$ , ossia la funzione

$$F(x) = \int_a^x f(t) dt$$

si ha che  $F(x)$  è una funzione non decrescente in  $(a, b)$ , con  $0 \leq F(x) \leq 1$ . Ad ogni numero casuale  $r$  a distribuzione uniforme su  $(0, 1)$ , si associa (cfr. Figura 10.3)  $\xi$  definito nel modo seguente

$$\xi = F^{-1}(r) := \inf\{x \in (a, b) \mid F(x) \geq r\}, \quad 0 \leq r \leq 1 \quad (10.5)$$

In effetti, dalla figura si vede che la probabilità che  $\xi$  sia nell'intervallo  $(x, x + dx)$  è uguale alla probabilità che  $r$  sia nell'intervallo  $(F(x), F(x + dx))$ ; ma quest'ultima, essendo  $r$  distribuito uniformemente, è data da  $F(x + dx) - F(x) = F'(x) dx$ , ossia da  $f(x) dx$ , e di conseguenza  $\xi$  segue la distribuzione  $f(x)$ . In altre parole, se  $U$  è una variabile aleatoria con distribuzione uniforme su  $[0, 1]$ , allora la variabile  $F^{-1}(U)$

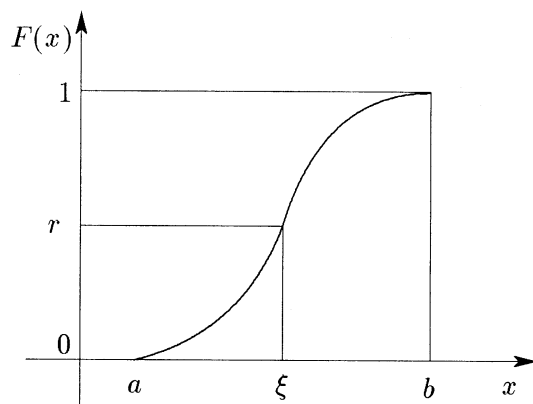


Figura 10.3: Illustrazione del metodo della trasformazione inversa.

ha  $F$  come funzione di distribuzione; viceversa, se  $X$  è una variabile aleatoria con funzione di ripartizione  $F$ , allora  $F(X)$  è uniformemente distribuita su  $[0, 1]$ .

Il risultato (10.5) può essere utilizzato per generare numeri random nel caso in cui  $F^{-1}$  sia nota esplicitamente. Per ottenere la distribuzione  $f(x)$ , è necessario, in pratica, generare un numero elevato  $N$  di numeri casuali  $r$  e dedurne i corrispondenti numeri casuali  $\xi$ . Si tiene conto quindi del numero  $\Delta n(x)$  di numeri  $\xi$  nell'intervallo  $x - \Delta x/2, x + \Delta x/2$ , da cui  $\Delta n(x)/N = f(x) \Delta x$ . In questo modo si può costruire l'istogramma della distribuzione; per esemplificazioni, si vedano le Figure 10.4 e 10.5.

Osserviamo, infine, che quando la distribuzione  $f(x)$  è data a meno di un fattore di normalizzazione, l'equazione (10.5) è sostituita dalla seguente

$$r = \left( \int_a^\xi f(t) dt \right) / \left( \int_a^b f(t) dt \right) \quad (10.6)$$

► **Esempio 10.2** Come illustrazione, consideriamo la generazione di numeri distribuiti secondo la seguente densità di probabilità

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{altrove} \end{cases} \Rightarrow F(x) = \begin{cases} 0 & x < 0 \\ \int_0^x 2t dt = x^2 & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases} \quad (10.7)$$

Applicando la (10.5), si ha  $\xi = r^{1/2}$ , per  $0 \leq r \leq 1$ . In Figura 10.4 è rappresentata la distribuzione ottenuta nel modo ora ottenuto, a partire da una successione di 10 000 numeri pseudo-casuali ottenuti mediante l'algoritmo (10.3).

Lasciamo come esercizio verificare che procedendo in modo analogo per la densità

$$f(x) = \begin{cases} \frac{2}{a} \left(1 - \frac{x}{a}\right) & 0 \leq x \leq a \\ 0 & \text{altrove} \end{cases} \Rightarrow F(x) = \begin{cases} 0 & x < 0 \\ \frac{2}{a} \left(x - \frac{x^2}{2a}\right) & 0 \leq x \leq a \\ 1 & x > a \end{cases} \quad (10.8)$$

si ottiene  $\xi = a(1 - \sqrt{r})$ . Si tenga presente che, se  $r$  è un numero casuale su  $[0, 1]$ , anche  $1 - r$  ha la stessa proprietà.

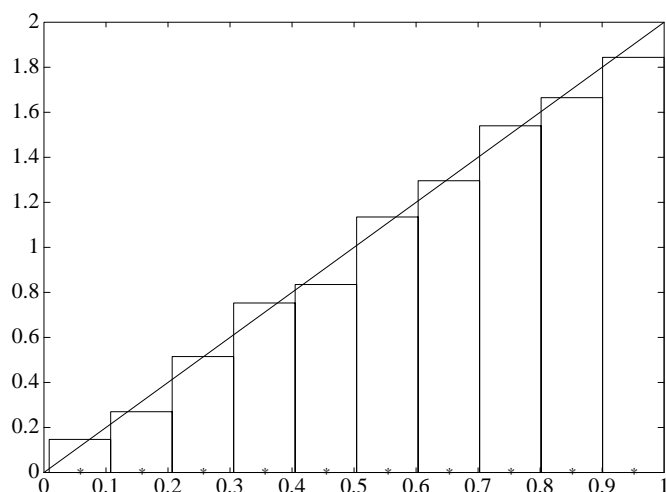


Figura 10.4: Frequenza relativa corrispondente alla densità di probabilità definita in (10.7) e ottenuta a partire da 10000 numeri pseudo-casuali distribuiti uniformemente su  $(0,1)$ .

► **Esempio 10.3** (*distribuzione esponenziale*) La distribuzione esponenziale  $\mathcal{E}(\lambda)$ , utilizzata in particolare nello studio di processi di code (modelli di *interarrival* e *service times*, cfr. Capitolo 8), è definita nel seguente modo

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & 0 \leq x < \infty, \lambda > 0 \\ 0 & \text{altrove} \end{cases} \quad (10.9)$$

a cui corrisponde la funzione di ripartizione

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \quad (10.10)$$

In questo caso il metodo della trasformata inversa fornisce direttamente il risultato

$$\xi = -\frac{\ln(1-r)}{\lambda}$$

o anche, dal momento che  $1 - r$  è pure un numero casuale a distribuzione uniforme su  $(0, 1)$

$$\xi = -\frac{1}{\lambda} \ln r \quad (10.11)$$

In Figura 10.5 è rappresentata la frequenza relativa ottenuta, per  $\lambda = 1$ , utilizzando 10 000 numeri pseudo-casuali a distribuzione uniforme. Osserviamo che, in pratica, è possibile ricavare  $\xi$  senza il calcolo del logaritmo, e quindi in maniera più efficiente (cfr. ad esempio Rubinstein [135]).

Come esercizio lasciamo verificare che per la seguente distribuzione, detta anche *distribuzione di Rayleigh*

$$f(x) = \frac{x}{\sigma} e^{-x^2/2\sigma^2}, \quad \sigma \neq 0, \quad x \geq 0 \quad (10.12)$$

si ha  $\xi = \sigma\sqrt{-\ln r}$ . ■

Come abbiamo già osservato, il metodo della trasformata inversa è utilizzabile nella forma illustrata in precedenza quando  $F(x)$  è tale che la corrispondente trasformata inversa può essere calcolata analiticamente, o in maniera equivalente, quando esiste la soluzione analitica dell'equazione  $F(\xi) = r$ . Vi sono, comunque, distribuzioni, tra le quali l'importante distribuzione gaussiana, per le quali questo non avviene. D'altra parte, anche quando si conosce la forma analitica di  $F^{-1}$ , il metodo della trasformata inversa può non essere necessariamente il metodo più efficiente. Esistono quindi opportuni metodi alternativi, tra i quali analizzeremo brevemente un metodo introdotto da von Neumann, rinviando, per una panoramica più approfondita, alla bibliografia già citata.

**Metodo acceptance-rejection di Von Neumann** Supponiamo che la densità di probabilità  $f(x)$ , definita sull'intervallo  $(a, b)$ , sia limitata

$$f(x) \leq M$$

Si ha quindi  $f(x)/M \leq 1$  e  $f(a) = 0$ . Si generano allora delle coppie di numeri casuali  $(r_1, r_2)$  a distribuzione uniforme su  $(0, 1)$ . Si confronta quindi  $r_1$  con la quantità  $f(a + (b - a)r_2)/M$  e se

$$\begin{cases} \frac{1}{M}f(a + (b - a)r_2) \geq r_1 & \text{si pone } \xi = a + (b - a)r_2 \\ \frac{1}{M}f(a + (b - a)r_2) < r_1 & \text{si scarta la coppia } (r_1, r_2) \end{cases}$$

Si può dimostrare che i numeri casuali  $\xi$  selezionati in questo modo seguono la distribuzione  $f(x)$ .

► **Esempio 10.4** Per la seguente densità di probabilità

$$f(x) = 3x^2, \quad 0 \leq x \leq 1$$

si ha  $M = 3, a = 0$  e  $b = 1$ . Si ha allora il seguente algoritmo

1. Si generano due numeri casuali  $r_1, r_2$  secondo la distribuzione  $\mathcal{U}(0, 1)$ .
2. Se  $r_1 \leq r_2^2$ , allora  $r_2$  è un numero casuale secondo la distribuzione  $f(x)$ .



3. In caso contrario, si rifiuta la coppia  $r_1 r_2$  e si ritorna al passo 1.

Il metodo può essere generalizzato nel seguente modo. Sia  $X$  una variabile aleatoria con densità di probabilità  $f_X(x)$ ,  $x \in I$  che viene rappresentata nella forma

$$f_X(x) = C h(x) g(x) \quad (10.13)$$

ove  $C \geq 1$ ,  $h(x)$  è una densità di probabilità, e  $0 < g(x) \leq 1$ . Si generano quindi due variabile aleatorie  $U$  con distribuzione  $\mathcal{U}(0,1)$  e  $Y$  con distribuzione  $h(y)$  e si esegue il test  $U \leq g(Y)$ . Quando tale test è verificato, si assume  $Y$  come numero casuale secondo la distribuzione  $f_X(x)$ ; in caso contrario, si rigetta la coppia  $U, Y$  e si ripete il procedimento. Naturalmente, la procedura ha interesse pratico quando è facile generare numeri casuali secondo la distribuzione  $h(x)$  e l'efficienza, ossia il reciproco del numero di tentativi prima di trovare una coppia che passa il test, è sufficientemente elevata. Si può vedere che la probabilità di successo in ogni tentativo è data da  $p = 1/C$ . Il caso che abbiamo considerato in precedenza corrisponde alle seguenti scelte

$$C = M(b-a), \quad h(x) = \frac{1}{b-a}, \quad g(x) = \frac{f(x)}{M} \quad a \leq x \leq b$$

Concludiamo l'argomento della generazione dei numeri casuali, considerando il caso importante della distribuzione gaussiana.

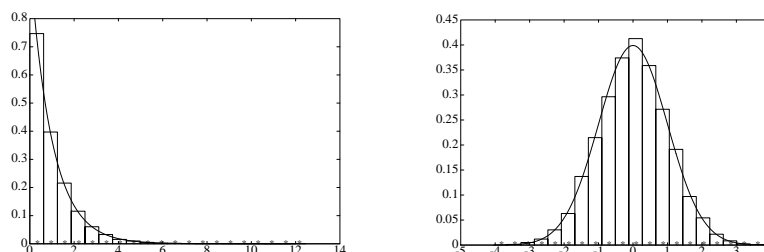


Figura 10.5: Simulazione, mediante la generazione di numeri pseudo-casuali con distribuzione uniforme, della distribuzione esponenziale (metodo della trasformata inversa) e della distribuzione gaussiana (metodo di Box e Muller).

**Distribuzione gaussiana** Ricordiamo che una variabile casuale  $X$  ha una distribuzione gaussiana, o normale,  $\mathcal{N}(\mu, \sigma^2)$ , se la densità di probabilità è data da

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty \quad (10.14)$$

Dal momento che  $X = \mu + \sigma Z$ , ove  $Z$  è la variabile normale standard  $\mathcal{N}(0,1)$ , possiamo limitarci a considerare la generazione di numeri casuali distribuiti secondo

la distribuzione  $\mathcal{N}(0, 1)$ . Come abbiamo già osservato in precedenza, in questo caso l'inversa  $F^{-1}$  della funzione di ripartizione non è esprimibile in forma analitica e non è possibile quindi utilizzare il metodo diretto della trasformata inversa. Rinviamo alla bibliografia citata per l'analisi di efficienti metodi alternativi, ci limiteremo a ricordare, come esemplificazione, un metodo introdotto da Box e Muller (1958) e basato sul seguente risultato. Se  $r_1$  e  $r_2$  sono due numeri casuali indipendenti distribuiti secondo la distribuzione  $\mathcal{U}(0, 1)$ , allora i numeri

$$z_1 = (-2 \ln U_1)^{1/2} \cos 2\pi U_2, \quad z_2 = (-2 \ln U_1)^{1/2} \sin 2\pi U_2 \quad (10.15)$$

seguono la distribuzione  $\mathcal{N}(0, 1)$ . La dimostrazione del risultato è basata sul passaggio a due dimensioni e alla trasformazione in coordinate polari. In Figura 10.5 è rappresentata la frequenza relativa ottenuta utilizzando 10 000 coppie di numeri pseudo-casuali a distribuzione uniforme.

## 10.2 Calcolo di integrali

Nel precedente Capitolo 6 abbiamo analizzato differenti metodi deterministici (formule di quadratura) per approssimare il valore di un integrale definito. Ricordiamo che l'errore commesso con tali formule dipende dalla regolarità della funzione integranda; ad esempio, se  $h$  indica il passo di suddivisione dell'intervallo di integrazione e quindi una quantità inversamente proporzionale al numero  $N$  delle valutazioni della funzione, per la formula composta di Cavalieri-Simpson l'errore ha ordine  $O(h^4)$  quando la funzione integranda ha le derivate fino all'ordine 4. In caso contrario, la formula può fornire un risultato con precisione inferiore. Nel caso di integrali in una dimensione abbiamo analizzato alcune tecniche per eliminare le eventuali discontinuità. In due o più dimensioni, tuttavia, le singolarità della funzione o delle derivate possono verificarsi anche lungo curve o superficie di forma complicata, e quindi esse possono essere difficilmente rimosse con particolari suddivisioni del dominio o con sostituzioni di variabili. In questi casi, il metodo di Monte Carlo diventa competitivo con le formule di quadratura, e in particolare quando la dimensione del dominio di integrazione è elevata, può essere in pratica l'unico metodo utilizzabile. Come vedremo, tutto quello che il metodo richiede è che la funzione integranda sia definita nel dominio di integrazione e che l'integrale *esista*. Un altro aspetto interessante del metodo è il fatto che l'ordine di convergenza, rispetto al numero di valutazioni della funzione integranda, è in sostanza indipendente dalle dimensioni del dominio di integrazione. Per le formule di quadratura, invece, il numero di valutazioni necessarie per ottenere una accuratezza assegnata cresce esponenzialmente con il numero delle dimensioni.

Nel seguito introdurremo le idee di base più in dettaglio nel caso di integrali semplici, essendo, come vedremo, immediata l'estensione al caso di integrali multipli immediata. In particolare, esamineremo due idee; la prima, nota come *metodo Monte*

*Carlo hit or miss* è basato sull'interpretazione geometrica di un integrale come area; la seconda, più efficiente in generale della precedente e chiamata *metodo Monte Carlo sample-mean*, utilizza la rappresentazione di un integrale come valore medio.

### 10.2.1 Metodo Monte Carlo hit or miss

Sia

$$I = \int_a^b g(x) dx \quad (10.16)$$

l'integrale da approssimare, ove per semplicità assumeremo che la funzione integranda  $g(x)$  sia limitata  $0 \leq g(x) \leq c$ ,  $a \leq x \leq b$ . Indichiamo con  $R$  il rettangolo (cfr. Figura 10.6)

$$R = \{(x, y) : a \leq x \leq b, 0 \leq y \leq c\}$$

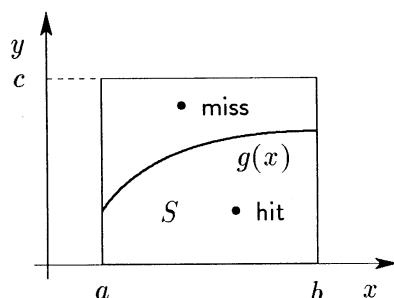


Figura 10.6: Rappresentazione grafica del metodo Monte Carlo hit or miss.

Sia  $(X, Y)$  un vettore casuale distribuito uniformemente sul rettangolo  $R$  con densità di probabilità

$$f_{XY}(x, y) = \begin{cases} 1/[c(b-a)], & \text{se } (x, y) \in R \\ 0, & \text{altrove} \end{cases} \quad (10.17)$$

La probabilità  $p$  che il vettore casuale  $(X, Y)$  cada nell'insieme  $S := \{(x, y) \mid y \leq g(x)\}$  è data da

$$p = \frac{\text{area } S}{\text{area } R} = \frac{\int_a^b g(x) dx}{c(b-a)} = \frac{I}{c(b-a)} \quad (10.18)$$

Se  $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$  sono  $N$  vettori casuali indipendenti, il parametro  $p$  può essere stimato dal seguente rapporto

$$\hat{p} = \frac{N_H}{N} \quad (10.19)$$

ove  $N_H$  è il numero dei successi (*hits*), ossia il numero dei vettori per i quali  $g(X_i) \geq Y_i$ ,  $i = 1, 2, \dots, N$  (analogamente  $N - N_H$  è il numero degli insuccessi (*misses*)). Ne

segue da (10.18) e (10.19) che l'integrale  $I$  può essere stimato come segue

$$I \approx \theta_1 = c(b-a) \frac{N_H}{N} \quad (10.20)$$

Dal momento che ognuna delle  $N$  prove costituisce una prova di Bernoulli con probabilità  $p$  di successo, allora

$$E(\theta_1) = c(b-a)E\left(\frac{N_H}{N}\right) = c(b-a)\frac{E(N_H)}{N} + pc(b-a) = I$$

da cui si ricava che  $\theta_1$  è uno stimatore non distorto di  $I$ . La varianza di  $\hat{p}$  è data da

$$\text{var}(\hat{p}) = \text{var}\left(\frac{N_H}{N}\right) = \frac{1}{N^2} \text{var}(N_H) = \frac{1}{N} p(1-p)$$

da cui, tenendo conto della (10.18)

$$\text{var}(\theta_1) = [c(b-a)]^2 \text{var}(\hat{p}) = \frac{I}{N} [c(b-a) - I] \quad (10.21)$$

e la deviazione standard

$$\sigma_{\theta_1} = N^{-1/2} \{I [c(b-a) - I]\}^{1/2}$$

Utilizzando la disuguaglianza di Chebichev (cfr. Capitolo 8)

$$P[|\theta_1 - I| < \epsilon] \geq 1 - \frac{\text{var}(\theta_1)}{\epsilon^2}$$

si vede facilmente che per ottenere

$$P[|\theta_1 - I| < \epsilon] \geq \alpha$$

con  $\alpha$  livello di probabilità fissato, è richiesto un numero di valutazioni  $N$ , con

$$N \geq \frac{(1-p)p[c(b-a)]^2}{(1-\alpha)\epsilon^2}$$

Quando  $N$  è sufficientemente elevato, grazie al teorema limite centrale la variabile aleatoria

$$\hat{\theta}_1 = \frac{\theta_1 - I}{\sigma_{\theta_1}}$$

è distribuita approssimativamente come la distribuzione normale standard  $\Phi(x)$ , e quindi l'*intervallo di confidenza* con livello  $1 - 2\alpha$  è dato da

$$\theta_1 \pm z_\alpha \frac{[\hat{p}(1-\hat{p})]^{1/2} (b-a)c}{N^{1/2}}, \quad z_\alpha = \Phi^{-1}(\alpha) \quad (10.22)$$

### 10.2.2 Metodo Monte Carlo sample-mean

Il metodo sample-mean consiste nel rappresentare l'integrale (10.16) come un valore atteso di una particolare variabile aleatoria. In maniera generale, possiamo riscrivere l'integrale nel seguente modo

$$I = \int_a^b g(x) dx \Rightarrow I = \int_a^b \frac{g(x)}{f_X(x)} f_X(x) dx \quad (10.23)$$

ove  $f_X(x)$  è una qualsiasi densità di probabilità, tale che  $f_X(x) > 0$  quando  $g(x) \neq 0$ . Allora

$$I = E\left(\frac{g(X)}{f_X(X)}\right) \quad (10.24)$$

ove la variabile aleatoria  $X$  è distribuita secondo  $f_X(x)$ . Assumendo, in particolare

$$f_X(x) = \begin{cases} 1/(b-a) & \text{se } a < x < b \\ 0 & \text{altrove} \end{cases} \quad (10.25)$$

si ha

$$E(g(X)) = \frac{I}{b-a} \Rightarrow I = (b-a) E(g(X)) \quad (10.26)$$

e uno stimatore non distorto di  $I$  è la seguente media campionaria (*sample mean*)

$$\theta_2 = (b-a) \frac{1}{N} \sum_{i=1}^N g(X_i) \quad (10.27)$$

La varianza di  $\theta_2$  è uguale a  $E(\theta_2^2) - [E(\theta_2)]^2$  e quindi

$$\text{var}(\theta_2) = \frac{1}{N} \left[ (b-a) \int_a^b g^2(x) dx - I^2 \right] \quad (10.28)$$

In conclusione, l'algoritmo Monte Carlo sample-mean è definito dai seguenti passi

1. Si genera una successione  $\{U_i\}_{i=1}^N$  di  $N$  numeri casuali.
2. Si calcola  $X_i = a + U_i(b-a)$ ,  $i = 1, \dots, N$ .
3. Si calcola  $g(X_i)$ ,  $i = 1, \dots, N$ .
4. Si calcola la media  $\theta_2$  come indicato in (10.27); il valore  $\theta_2$  fornisce una stima dell'integrale  $I$  con varianza data da (10.28).

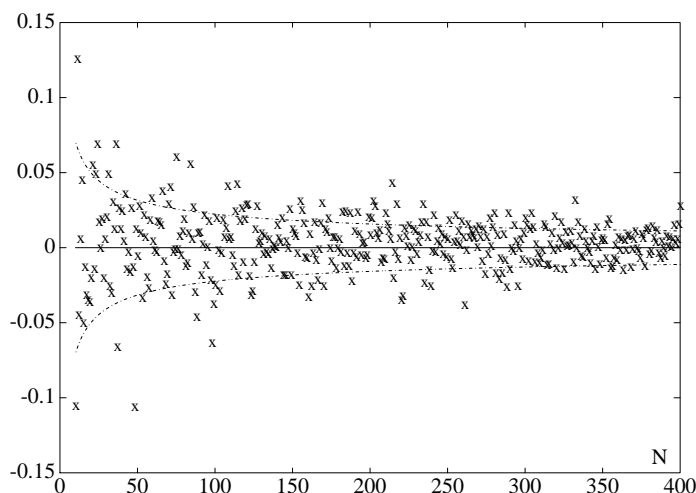


Figura 10.7: Rappresentazione degli errori  $\theta_2 - I$ , con  $I = \int_0^1 \sqrt{1-x^2} dx$ , per valori crescenti di  $N$ . Con linea tratteggiata sono indicate le curve corrispondenti alle deviazioni standard.

**Efficienza di un metodo Monte Carlo** Supponiamo che due differenti metodi Monte Carlo forniscano due stime  $\theta_1$  e  $\theta_2$  dell'integrale  $I$ , con

$$E(\theta_1) = E(\theta_2) = I$$

e indichiamo con  $T_1$  e  $T_2$  rispettivamente i tempi richiesti dai due metodi per valutare  $\theta_1$  e  $\theta_2$ . Allora, si dice che il primo metodo è meno efficiente del secondo, quando

$$\frac{T_1 \text{var}(\theta_1)}{T_2 \text{var}(\theta_2)} > 1$$

Considerando in particolare il metodo hit or miss e il metodo sample-mean, dal momento che, come si verifica facilmente, si ha  $\text{var}(\theta_1) \geq \text{var}(\theta_2)$ , supponendo approssimativamente uguali i tempi  $T_1$  e  $T_2$ , si conclude che il metodo sample-mean è più efficiente del metodo hit or miss.

► **Esempio 10.5** Come esemplificazione, utilizzando 10 numeri estratti dalla Tabella 10.1 si ottiene per l'integrale

$$I = \int_0^1 \sqrt{1-x^2} dx$$

che fornisce l'area del quadrante di cerchio, cioè il numero  $\pi/4$ , il valore  $\theta_2 = 0.8048$ , con un errore del 2.4%. Sottolineiamo che la funzione integranda presenta una singolarità per la derivata prima nel punto  $x = 1$ . In Figura 10.7 è riportato l'andamento degli errori  $\theta_2(N) - I$  per valori crescenti di  $N$  e utilizzando una successione di numeri pseudo-casuali generati con l'algoritmo (10.3).

N	$x = -\ln t$	$x = -2 \ln t$
100	0.6994	0.7089
1000	0.7066	0.7113
10000	0.7062	0.6962

Tabella 10.2: Risultati ottenuti con il metodo Monte Carlo sample-mean per l'integrale (10.29), in corrispondenza a due diverse sostituzioni.

Il rapporto di efficienza tra il metodo hit or miss e il metodo sample-mean è su questo esempio dato da  $(p(1-p)/(2/3-p^2)) \approx 3.38$ , ove  $p = \pi/4$ .

Come ulteriore esemplificazione, consideriamo il calcolo del seguente integrale

$$I = \int_0^\infty e^{-x} \cos^2(x^2) dx = 0.70260 \dots \quad (10.29)$$

Mediante la trasformazione di variabili  $x = -\ln t$ , l'integrale precedente può essere ridotto al seguente integrale su un intervallo finito

$$I = \int_0^1 \cos^2(\ln^2 t) dt$$

per il quale la funzione integranda oscilla indefinitamente tra i valori 0 e 1 su ogni intervallo contenente l'origine. Analogamente, con la trasformazione  $x = -2 \ln t$ , si ottiene un integrale su  $(0, 1)$  con funzione integranda  $2t \cos^2(4 \ln^2 t)$ , che tende a zero per  $t \rightarrow 0$ . Il metodo Monte Carlo sample-mean fornisce i risultati contenuti nella Tabella 10.2. ■

### 10.2.3 Calcolo di integrali multipli

I metodi precedenti si estendono facilmente al calcolo di un integrale multiplo della forma

$$I = \int_D g(\mathbf{x}) d\mathbf{x} \quad (10.30)$$

ove  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^n$  e  $D = \prod_{k=1}^n [a_k, b_k]$ . Ad esempio, l'applicazione del metodo sample-mean consiste nell'estrarre a caso, per la prova  $i$ -ma,  $n$  variabili aleatorie a distribuzione rettangolare su  $(0, 1)$

$$\mathbf{U}_i = [U_{i1}, U_{i2}, \dots, U_{in}], \quad i = 1, 2, \dots, N$$

a cui si associa il vettore  $\mathbf{X}_i = [X_{i1}, X_{i2}, \dots, X_{in}]$ , con  $X_{ik} = a_k + (b_k - a_k)U_{ik}$ . In corrispondenza, si calcola il seguente stimatore di  $I$

$$\bar{I} = |D| \frac{1}{N} \sum_{i=1}^N g(\mathbf{X}_i), \quad |D| = \prod_{k=1}^n (b_k - a_k) \quad (10.31)$$

e il valore della varianza può essere ottenuto procedendo in maniera del tutto analoga a quanto fatto nel paragrafo precedente. Come si vede, all'aumentare della dimensione  $n$  corrisponde un aumento della quantità dei numeri casuali necessari, ma non delle valutazioni di  $g$  (che sono in generale la parte più costosa dell'algoritmo). In altre parole, il risultato di base sulla riduzione dell'errore in maniera proporzionale a  $N^{-1/2}$  si applica indipendentemente dalle dimensioni del dominio di integrazione.

► **Esempio 10.6** Come illustrazione, consideriamo l'applicazione del metodo sample-mean al seguente integrale multiplo

$$I = \int_0^1 \int_0^1 \int_0^1 \int_0^1 e^{x_1 x_2} \cos\left(\frac{\pi}{2} x_3 x_4\right) dx_1 dx_2 dx_3 dx_4 \quad (10.32)$$

Riducendo l'integrale a due dimensioni e utilizzando il metodo di Romberg si ha il valore  $\bar{I} = 1.150073$ . In Tabella 10.3 sono riportati i risultati ottenuti mediante il seguente programma implementato in linguaggio MATLAB.

```
for ii=1:14
    n=2^ii;
    x=rand(n,4); % generazione di quadruple di numeri casuali
    f=exp(x(:,1)*x(:,2))*cos(x(:,3)*x(:,4)*pi/2);
    I(ii)=sum(f)/n % calcolo del valore medio
end
```

punti	valore approssimato	punti	valore approssimato
2	0.977036	$2^8$	1.178055
$2^2$	1.010176	$2^9$	1.162432
$2^3$	0.984609	$2^{10}$	1.150753
$2^4$	1.158108	$2^{11}$	1.155908
$2^5$	1.126952	$2^{12}$	1.155255
$2^6$	1.090402	$2^{13}$	1.153096
$2^7$	1.098141	$2^{14}$	1.150869

Tabella 10.3: Risultati ottenuti con il metodo Monte Carlo applicato all'integrale (10.32).

Dalla tabella si vede che il risultato si stabilizza sulle tre cifre 1.15 a partire da  $2^{10} = 1024$  valutazioni della funzione; ricordiamo che nell'applicazione, ad esempio, della formula dei trapezi con una suddivisione di passo  $1/10$  sono necessarie  $10^4 = 10000$  valutazioni della funzione integranda. ■

#### 10.2.4 Tecniche di riduzione della varianza

Abbiamo visto che la velocità di convergenza del metodo Monte Carlo nella formulazione di base è dell'ordine  $1/\sqrt{N}$ . Questo significa in pratica che, per aumentare



l'accuratezza di 10 volte, è necessario aumentare  $N$  (e quindi la quantità di calcolo) 100 volte. Tuttavia, quando sulla funzione integranda si hanno ulteriori informazioni, è possibile applicare opportune tecniche che, *riducendo la varianza*, permettono di migliorare l'accuratezza della stima ottenuta, senza aumentare il numero delle prove. Nel seguito daremo una breve panoramica di tali tecniche, rinviando per un approfondimento alla bibliografia citata. Ricordiamo, anche, che tecniche più recenti, basate sull'utilizzo di numeri che non sono casuali, ma con particolari proprietà teoriche (*numeri quasi-random*, o equidistribuiti), permettono di fornire stime degli integrali multipli dell'ordine di  $N^{-1}$ .

**Campionamento secondo l'importanza** L'idea di base consiste nel concentrare la distribuzione dei punti campione nelle parti del dominio di integrazione che sono ritenute "più importanti". Con riferimento alla formulazione generale (10.23), (10.24), si ha il seguente stimatore dell'integrale  $I$

$$\theta_3 = \frac{1}{N} \sum_{i=1}^N \frac{g(X_i)}{f_X(X_i)} \quad (10.33)$$

ove  $X_i$  sono distribuiti secondo la densità di probabilità  $f_X(x)$ , con varianza

$$\text{var}(\theta_3) = \frac{1}{N} \left[ \int_a^b \frac{g^2(x)}{f_X(x)} dx - I^2 \right] \quad (10.34)$$

Il problema diventa, allora, quello di *scegliere* la distribuzione  $f_X(x)$  che minimizza  $\text{var}(\theta_3)$ . Naturalmente, la scelta ottimale sarebbe  $f_X(x) = |g(x)| / \int_a^b |g(x)| dx$ , ma tale scelta richiede il calcolo dell'integrale di  $|g(x)|$ . In pratica, si sceglie quindi  $f_X(x)$  tale da avere un comportamento "simile" a quello della funzione  $|g(x)|$ .

**Campionamento stratificato** L'idea di questa tecnica, simile a quella utilizzata per introdurre le formule di quadratura composte (cfr. Capitolo 6), consiste nel suddividere l'intervallo di integrazione in  $(a, b)$  in un numero  $K$  di intervalli mediante i punti  $a = a_0 < a_1 < \dots < a_K = b$ . Si ha quindi

$$I = \int_a^{a_1} f(x) dx + \int_{a_1}^{a_2} f(x) dx + \dots + \int_{a_{K-1}}^b f(x) dx = \sum_{k=1}^K I_k$$

con  $I_k = \int_{a_{k-1}}^{a_k} f(x) dx$ . Ogni passo del metodo consiste allora nella generazione di  $K$  numeri casuali distribuiti uniformemente su  $(0, 1)$ , dai quali si ricavano  $K$  numeri distribuiti uniformemente negli intervalli  $(a_{k-1}, a_k)$ . Lo stimatore si costruisce quindi in maniera ovvia. Il punto importante nell'applicazione dell'idea è la scelta della suddivisione. In pratica, una scelta conveniente consiste nello scegliere i punti  $a_k$  in maniera che la funzione integranda  $g$  abbia una variazione confrontabile in ciascun

intervallo. In questo senso, la tecnica ora illustrata è simile a quella del campionamento secondo l'importanza; la riduzione della varianza è ottenuta concentrando più campioni nei sottoinsiemi "più importanti", piuttosto che scegliendo la distribuzione  $f_X(x)$  ottimale.

Un altro aspetto importante del campionamento stratificato è il fatto che in ogni passo del metodo i valori della funzione integranda nei singoli intervalli possono essere calcolati in *parallelo*.

**Trasformazioni antitetiche** La tecnica è basata sul seguente risultato statistico. Se  $Y'$  e  $Y''$  sono due stimatori non distorti del parametro  $I$  (corrispondente al valore incognito dell'integrale), si ha che  $\frac{1}{2}(Y' + Y'')$  è pure uno stimatore non distorto di  $I$  con varianza

$$\text{var} \left[ \frac{1}{2}(Y' + Y'') \right] = \frac{1}{4} \text{var}(Y') + \frac{1}{4} \text{var}(Y'') + \frac{1}{2} \text{cov}(Y', Y'')$$

Ne segue che se la covarianza  $\text{cov}(Y', Y'')$  è negativa, si ha una riduzione della varianza quando si considera la media delle variabili  $Y', Y''$ . Per tale motivo, due variabili aleatorie con correlazione negativa vengono dette *variabili antitetiche*.

Come illustrazione, consideriamo l'integrale  $I = \int_0^1 g(x) dx$ , che è uguale a

$$I = \frac{1}{2} \int_0^1 [g(x) + g(1-x)] dx$$

La seguente variabile aleatoria

$$Y = \frac{1}{2}(Y' + Y'') = \frac{1}{2}[g(U) + g(1-U)]$$

con  $U$  variabile aleatoria con distribuzione  $\mathcal{U}(0, 1)$ , è uno stimatore non distorto di  $I$ . Per stimare  $I$  si prende quindi un campione di  $N$  numeri con distribuzione uniforme e si calcola

$$\theta_4 = \frac{1}{2N} \sum_{i=1}^N [g(U_i) + g(1-U_i)]$$

La situazione peggiore si verifica quando la funzione  $g(x)$  è simmetrica rispetto alla linea  $x = 1/2$ , e quindi  $g(x) = g(1-x)$ , mentre il caso più vantaggioso si ha quando la funzione è antisimmetrica, ossia  $g(x) - g(1/2) = g(1/2) - g(1-x)$ . Consideriamo, come illustrazione, l'integrale  $I = \int_0^1 \sqrt{1-x^2} dx$ . Per tale integrale la stima  $\theta_2$  ottenuta mediante il metodo sample-mean, con un campionamento di 100 numeri casuali, corrisponde all'errore  $\theta_2 - I \approx -0.0136$ , contro l'errore  $\theta_4 - I \approx -0.00218$  corrispondente alla stima  $\theta_4$  con  $N = 50$ .

La tecnica delle trasformazioni antitetiche può naturalmente essere combinata con la tecnica del campionamento stratificato. Ad esempio, dividendo l'intervallo

$(a, b)$  nei due sottointervalli  $(a, c)$  e  $(c, b)$ , ad ogni numero casuale  $U_i$  con distribuzione uniforme su  $(0, 1)$  si associa il punto in  $(a, c)$  :  $X'_i = a + (c - a)U_i$  e il punto in  $(c, b)$  :  $X''_i = c + (b - c)(1 - U_i)$ . Si calcolano quindi i valori  $I'_i = (c - a)g(X'_i)$  e  $I''_i = (b - c)g(X''_i)$  e lo stimatore

$$\theta_5 = \frac{1}{N} \sum_{i=1}^N (I'_i + I''_i)$$

Nel caso dell'integrale  $I = \int_0^1 \sqrt{1 - x^2} dx$  per  $N = 50$  si ha l'errore  $\theta_5 - I \approx 0.00111$  per  $c = 1/2$  e  $\theta_5 - I \approx -0.000566$  per  $c = 1/\sqrt{2}$ .

## 10.3 Simulazione

In questo paragrafo illustreremo mediante esempi l'idea della *simulazione*. In maniera schematica, si considera una situazione fisica nella quale è presente un elemento di casualità<sup>7</sup> e si cerca di imitare la situazione sul calcolatore. Opportune conclusioni statistiche possono essere ricavate se l'esperimento è ripetuto un elevato numero di volte.

### 10.3.1 Problema dei due dadi

Consideriamo l'esperimento del lancio di due dadi statisticamente simmetrici, e quindi con la stessa probabilità di ottenere in ciascun lancio uno dei numeri 1, 2, 3, 4, 5 e 6. Si cerca la probabilità di ottenere un 12 (cioè un 6 su ciascuno dei due dadi) in successivi 24 lanci dei due dadi.

Il problema può essere, come abbiamo visto nel precedente Capitolo 8, risolto in maniera analitica, e questo fatto rende l'esempio interessante per un test sulla simulazione. Brevemente, una soltanto delle 36 combinazioni è quella esatta; pertanto, con 24 lanci la probabilità di un risultato non corretto è  $(\frac{35}{36})^{24}$  e  $1 - (\frac{35}{36})^{24} = 0.49140$  è quindi la probabilità di ottenere un 12.

La simulazione del processo consiste nella ripetizione di un numero elevato di esperimenti, ognuno dei quali consiste nel lancio dei due dadi 24 volte. Per il risultato del lancio di un singolo dado, si utilizzano degli interi distribuiti uniformemente nell'insieme  $\{1, 2, 3, 4, 5, 6\}$ . Se  $U$  è una variabile aleatoria in  $(0, 1)$ , allora  $6U + 1$  è una variabile aleatoria in  $(1, 7)$  e la parte intera è un intero casuale in

<sup>7</sup>Come abbiamo già discusso nel Capitolo 8, il termine *casuale* è usato nel linguaggio comune per indicare un evento non prevedibile, quale ad esempio il risultato di un lancio di una moneta, l'ingresso di un nuovo cliente in un negozio, o il tempo tra due successivi clienti. In realtà, nessuno di tali eventi è realmente imprevedibile; ad esempio, il lancio della moneta dipende dall'orientamento iniziale della moneta stessa, dall'altezza del suolo, dalla resistenza dell'aria, ... Tuttavia, la complessiva complicazione di tali cause rende più utile pensare il risultato come casuale, in particolare quando si è interessati a conoscere il *comportamento medio*.

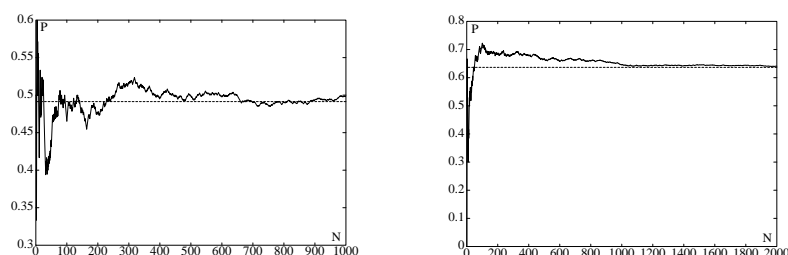


Figura 10.8: Le due figure rappresentano le frequenze corrispondenti rispettivamente all'esperimento del lancio dei due dadi e al problema di Buffon. La linea tratteggiata indica il valore teorico della probabilità.

$\{1, 2, 3, 4, 5, 6\}$ . L'esperimento è illustrato dal seguente programma implementato in linguaggio MATLAB.

```

m=0;
for N=1:1000
    for k=1:24
        i1=fix(6.*rand+1.); % lancio del primo dado
        i2=fix(6.*rand+1.); % lancio del secondo dado
        if ((i1+i2)==12), % successo, se escono due 6
            m=m+1 ; break, end
        end
    end
    p(N)=m/N; % calcolo della frequenza
end

```

In particolare, si ottiene  $p(1000) \approx 0.4980$ , in buon accordo con il valore teorico. In Figura 10.8 è mostrato l'andamento della frequenza  $p(N)$  al variare del numero  $N$  degli esperimenti.

### 10.3.2 Problema di Buffon

Il problema del lancio di un ago, noto come *problema di Buffon*<sup>8</sup>, rappresenta uno dei primi esempi di simulazione mediante il metodo Monte Carlo; la sua soluzione fornisce una stima della quantità  $1/\pi$ , ossia del rapporto del diametro di un cerchio e la corrispondente circonferenza. Nella forma più semplice, il problema è il seguente.

Su una superficie piana è tracciato un sistema di rette parallele, a distanza  $d$  uguale tra loro. Si suppone quindi che un ago di lunghezza  $l$  venga lanciato sul piano in maniera casuale; più precisamente, con riferimento alla Figura 10.9, si assume che la distanza  $u$  del centro dell'ago dalla retta più vicina e l'angolo  $v$  che fornisce l'orientamento dell'ago siano due variabili aleatorie con distribuzione uniforme. Si cerca allora la probabilità che l'ago intersechi una delle rette.

<sup>8</sup>Compte de Buffon (1707–1788), *Essai d'arithmétique morale* (1777).

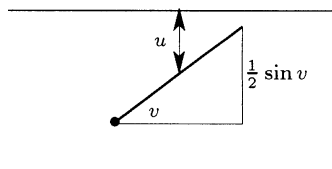


Figura 10.9: Problema di Buffon.

Se poniamo per semplicità  $d = l = 1$ , si vede facilmente che l'ago interseca una delle due linee se e solo se

$$u \leq \frac{1}{2} \sin v \quad (10.35)$$

con  $u$  variabile aleatoria distribuita uniformemente sull'intervallo  $(0, 1/2)$  e  $v$  variabile aleatoria distribuita uniformemente su  $(0, \pi/2)$ . La probabilità richiesta può essere quindi calcolata mediante la valutazione dell'area dell'insieme definito dalla disuguaglianza (10.35). Nel caso generale si trova  $p = 2l/\pi d$ , e quindi nel caso particolare considerato  $p = 2/\pi$ . Tale valore può essere stimato sperimentalmente con il metodo Monte Carlo, generando una successione di coppie di numeri  $(u, v)$  casuali e contando il numero di volte che è verificata la disuguaglianza (10.35). Nella seguente implementazione in linguaggio MATLAB si è posto  $w = 2u$ , con  $w$  variabile aleatoria uniforme su  $(0, 1)$ .

```

m=0;
for i=1:N
    w=rand; v=pi*rand/2;
    if (w <= sin(v)), m=m+1; end
p(N)=m/N;
end

```

Il comportamento della frequenza  $p(N)$  al variare di  $N$  è mostrato in Figura 10.8; in particolare si ha  $p(2000) \approx 0.6405$ , rispetto al risultato analitico  $p \approx 0.6366$ .

### 10.3.3 Simulazione di traiettorie con collisioni

I problemi di *diffusione* sono legati a traiettorie di particelle che subiscono delle collisioni tra loro o con altre particelle. Una particella, ad esempio un neutrone di un reattore nucleare, un ione in una soluzione, un elettrone in un metallo o in un semiconduttore, una molecola di un gas o di un fluido, eccetera, segue durante un intervallo di tempo una traiettoria deterministica, regolata da un campo di forze assegnato. Successivamente, in corrispondenza ad un istante aleatorio, subisce una *collisione*, in seguito alla quale la sua velocità  $\mathbf{v}$  viene modificata ad un valore aleatorio  $\mathbf{v}'$  (con eventuale creazione di altre particelle). In questo senso la traiettoria

di una particella ha un carattere fortemente aleatorio. Di conseguenza, i metodi Monte Carlo si presentano interessanti, a volte gli unici metodi in grado di trattare tali problemi.

L'idea più comunemente utilizzata consiste nel decomporre la traiettoria di una particella in un numero elevato di sequenze, ognuna delle quali comprende un *volo libero* e una *collisione*. Più precisamente, si definisce lo stato iniziale (posizione e velocità), in generale in maniera arbitraria, e successivamente

1. si estrae un primo numero aleatorio  $r_1$ , che determina la durata  $t_1$  del volo libero (o, equivalentemente, l'istante  $t_1$  al quale avviene la collisione).
2. Tra gli istanti 0 e  $t_1$  la particella segue una traiettoria deterministica corrispondente al campo di forze assegnato. Si può calcolare quindi (eventualmente, mediante opportuni metodi numerici) la posizione e la velocità  $\bar{\mathbf{v}}_1$  al tempo  $t_1$ .
3. Nel caso in cui siano possibili più collisioni, si seleziona una particolare collisione mediante un nuovo numero aleatorio  $r_2$ .
4. Scelta una determinata collisione, un insieme di numeri aleatori determina la posizione e la velocità  $\bar{\mathbf{v}}'_1$  dopo la collisione, supposta istantanea. A questo punto si conosce la posizione e la velocità all'inizio della seconda sequenza e si può ripartire dal passo 1 con istante iniziale  $t_1$ .
5. Nel caso in cui la collisione crei altre particelle, si utilizzano altri numeri casuali per simulare la loro velocità.

La procedura è illustrata in Figura 10.10, nella quale è rappresentata la proiezione sul piano  $(x, y)$  dello spazio delle componenti della velocità, determinata, in volo libero, da un campo di forze costanti dirette nel senso di  $x \geq 0$ . La particella parte dal punto  $O$  e la velocità aumenta fino al punto  $P_1$ . In tale punto subisce una collisione a seguito della quale l'estremo del vettore velocità viene portato nel punto  $P'_1$ , dal quale si ha un volo libero fino al punto  $P_2$ , ecc.

Dalla simulazione di un insieme di particelle si può ricavare, a partire da una descrizione microscopica del sistema, una stima dell'evoluzione nel tempo della velocità e dell'energia media.

◆ **Esercizio 10.1** *Descrivere un algoritmo per generare numeri casuali distribuiti secondo la seguente densità di probabilità (distribuzione logistica)*

$$f_X(x) = \frac{\exp[-(x - \alpha)/\beta]}{\beta [1 + \exp[-(x - \alpha)/\beta]]^2}, \quad -\infty < x < \infty, \beta > 0, \alpha > 0$$

◆ **Esercizio 10.2** *Esaminare il problema della generazione di numeri casuali interni a un cerchio, a partire da numeri casuali distribuiti uniformemente su  $(0, 1)$ .*

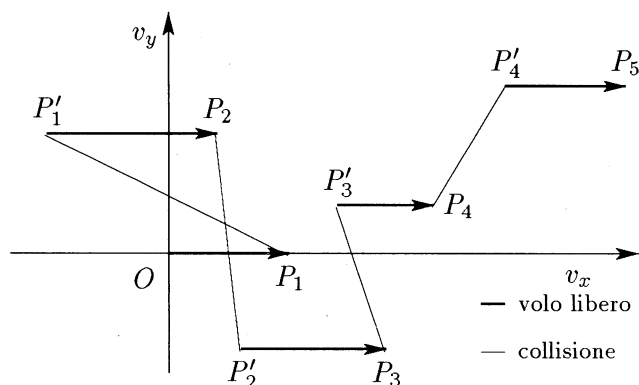


Figura 10.10: Simulazione Monte Carlo di una traiettoria nello spazio delle velocità.

◆ **Esercizio 10.3** Usare il metodo Monte Carlo per approssimare il seguente integrale

$$\int_{-1}^1 \int_{-1}^1 \int_{-1}^1 (x^2 + y^2 + z^2) dx dy dz$$

◆ **Esercizio 10.4** Usando il metodo Monte Carlo, trovare l'area della regione definita dalle disequazioni

$$1 \leq x \leq 3, \quad -1 \leq y \leq 4, \quad x^3 + y^3 \leq 29, \quad y \geq e^x - 2$$

◆ **Esercizio 10.5** Per l'approssimazione dell'integrale  $\int_a^b f(x) dx$  si analizzi la seguente procedura Monte Carlo. Si generano  $n$  numeri casuali su  $(a, b)$ , con  $n$  dispari, e si riordinano in maniera che  $a < x_1 < x_2 < \dots < x_n < b$ . Si calcola quindi

$$I \approx f(x_1)(x_2 - a) + f(x_3)(x_4 - x_2) + \dots + f(x_n)(b - x_{n-1})$$

Verificare la procedura sull'integrale  $\int_0^1 \sqrt{1-x^2} dx$ .

◆ **Esercizio 10.6** Stimare mediante la simulazione con il metodo Monte Carlo il numero medio di lanci di una moneta, statisticamente simmetrica, necessari per ottenere i tre differenti eventi (a) HHTT, (b) HTHT, (c) HHHH. Confrontare con la soluzione analitica.

◆ **Esercizio 10.7** Due punti sono scelti a caso sulla circonferenza di un cerchio. Stimare mediante una simulazione Monte Carlo la distanza media dal centro del cerchio al baricentro dei due punti.

◆ **Esercizio 10.8** Scrivere un programma per stimare la probabilità che tre punti casuali sui lati di un quadrato siano i vertici di un triangolo ottuso.

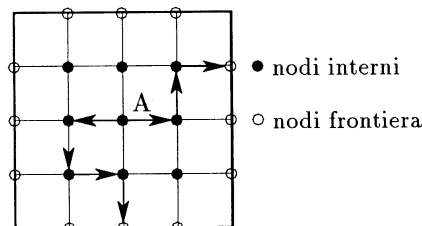
◆ **Esercizio 10.9** (Ottimizzazione). Stimare mediante il metodo Monte Carlo il minimo della funzione  $100(y - x^2)^2 + (1 - x^2)$ .

◆ **Esercizio 10.10** Dato il triangolo di vertici  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ , mostrare che se  $U$  e  $V$  sono variabili casuali con distribuzione  $\mathcal{U}(0, 1)$ , con  $U < V$ , il vettore  $U\mathbf{v}_1 + (V-U)\mathbf{v}_2 + (1-V)\mathbf{v}_3$  è distribuito uniformemente nel triangolo.

◆ **Esercizio 10.11** (Equazione di Laplace). Si consideri il seguente problema di Dirichlet per l'equazione di Laplace

$$\Delta u := u_{xx} + u_{yy} = 0, \quad \text{in } \Omega$$

$$u(x, y) = g(x, y), \quad \text{su } \partial\Omega$$



ove  $g(x, y)$  è una funzione definita sulla frontiera del dominio quadrato  $\Omega = (0, 1) \times (0, 1)$ . Utilizzando uno schema alle differenze finite (cfr. Capitolo 7), con lo stesso passo di discretizzazione nelle due variabili  $x$  e  $y$ , la soluzione discreta  $\bar{u}_{i,j} \approx u(ih, jh)$  risolve il seguente sistema lineare

$$u_{i,j} = \frac{1}{4}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}), \quad (ih, jh) \text{ nodo interno}$$

$$u_{i,j} = g(ih, jh), \quad (ih, jh) \text{ nodo frontiera}$$

Per la risoluzione di tale sistema analizzare la seguente procedura basata sul metodo Monte Carlo.

1. Si genera una successione di cammini casuali che hanno come punto di partenza un punto interno specificato  $A$  e come punto finale un punto di frontiera  $p_i$ . Per ogni punto di frontiera  $p_i$  fissato si calcola il numero di cammini casuali che hanno  $p_i$  come punto finale e si indica con  $P_A(p_i)$  la frazione tra tale numero e il numero totale dei cammini casuali.
2. Si calcola la soluzione approssimata  $\bar{u}(A)$  mediante la formula

$$\bar{u}(A) = g_1 P_A(p_1) + g_2 P_A(p_2) + \cdots + g_m P_A(p_m)$$

ove  $g_i$  è il valore della funzione  $g(x, y)$  nel punto frontiera  $p_i$  e  $m$  è il numero totale dei punti frontiera.

Si consideri l'estensione del metodo precedente al caso di  $n > 2$  dimensioni, esaminando in particolare la dipendenza del costo del metodo all'aumentare di  $n$  e al variare della funzione al bordo  $g$ .



A great discovery solves a great problem  
but there is a grain of discovery in the solution of any problem.

G. Polya

## Capitolo 11

# Introduzione al trattamento dei segnali

Uno degli obiettivi fondamentali nel trattamento dei segnali (signal processing), è costituito dal *filtraggio* (filtering), che rappresenta, in maniera schematica, il tentativo di minimizzare l'influenza dei segnali estranei e dei disturbi (noise). Tale obiettivo può ottenuto attraverso, sostanzialmente, due tipi di procedure. Per via hardware, mediante filtri di tipo *analogico*, realizzati usualmente mediante circuiti dotati di opportuni trasformatori e amplificatori, oppure per mezzo di opportuni *algoritmi numerici*, implementati su calcolatori digitali. Quest'ultimo tipo di filtri è indicato solitamente con il nome di *filtri digitali*, e con *trattamento dei segnali digitali* (digital signal processing) viene indicata la teoria che studia le proprietà e le caratteristiche di tali filtri.

L'uso dei filtri digitali, sviluppatosi a partire dagli anni sessanta, in concomitanza con l'affermarsi dell'hardware informatico, offre, rispetto ai filtri di tipo analogico, svariati vantaggi. In particolare, essi permettono una perfetta riproducibilità e una superiore facilità nella progettazione di nuovi tipi di filtri (maggiore adattività), nonché la possibilità di utilizzare contemporaneamente il medesimo sistema hardware per differenti funzioni di filtraggio.

Lo scopo di questo capitolo è, essenzialmente, quello di introdurre, oltre che il concetto di filtro digitale, gli *strumenti matematici e algoritmici* che sono alla base dell'analisi e della progettazione di un filtro digitale. In particolare, verranno analizzati differenti tipi di trasformate matematiche, ossia la *trasformata  $z$* , la *trasformata di Fourier discreta* e la trasformata basata sull'utilizzo delle funzioni *wavelet*. Le wavelet, una sintesi di idee sviluppatesi negli ultimi 20 anni in diversi settori, che vanno dall'ingegneria alla matematica pura, rappresentano uno strumento matematico di grande interesse per l'analisi numerica e per l'analisi dei segnali,

con applicazioni alla *compressione* di immagini e al *riconoscimento* di forme. Algoritmo per il calcolo veloce della trasformata di Fourier discreta, comunemente indicata come *Fast Fourier Transform* (FFT), deve essere attribuito, in sostanza, buona parte dell'*interesse pratico* dei filtri digitali. Come illustrazione applicativa, viene presentato un insieme di routine in linguaggio MATLAB per la costruzione di filtri numerici, e più in generale per il trattamento dei segnali. Successivamente, viene introdotto e discusso il problema della costruzione di *filtri ottimali* (*filtro di Kalman*).

Per un ampliamento e un approfondimento delle nozioni introdotte in questo capitolo si può vedere, per quanto riguarda le wavelet, Daubechies [43], e per l'analisi dei filtri numerici Jackson [90], Rabiner e Gold [132], Parks e Burrus [128], Scharf [142].

Motivi di spazio non permettono di trattare in maniera adeguata un argomento strettamente collegato con quelli trattati in questo capitolo e indicato usualmente con il termine di *elaborazione numerica delle immagini* (image processing). Ricordiamo, in particolare, il problema della ricostruzione numerica di una immagine tridimensionale a partire dalle sue proiezioni bidimensionali. Tale problema<sup>1</sup> ha assunto una grande importanza negli ultimi anni, a seguito delle numerose e importanti applicazioni, quali ad esempio: *tomografia computerizzata* (TAC), *ecografia ad ultrasuoni*, *risonanza magnetica nucleare* (MRI, magnetic resonance imaging). Per una introduzione rinviamo ad esempio a Herman [80] e Cappellini [26].

## 11.1 Teoria dei sistemi lineari a tempo discreto

La teoria dei sistemi lineari a tempo discreto (*discrete-time linear systems*) riguarda la rappresentazione e il trattamento di sequenze, sia in tempo che in frequenza, allo scopo di ottenere algoritmi di simulazione su calcolatori digitali (*digital signal processing*).

Generalmente il tempo è quantizzato in modo uniforme, ossia  $t = nT$ , ove  $T$  è l'intervallo tra due campionamenti. I segnali a tempo discreto sono rappresentati matematicamente come sequenze di numeri la cui ampiezza può assumere i valori di una variabile continua<sup>2</sup>. Una sequenza  $h(n)$  (o  $h(nT)$ ), per  $N_1 \leq n \leq N_2$ , può essere ottenuta in vari modi: per *definizione esplicita* (ad esempio  $h(n) = n$ ,  $0 \leq n \leq N$ ), per *ricorrenza* (ad esempio  $h(n) = h(n-1)/2$ ), oppure per *campionamento*

<sup>1</sup>La prima soluzione matematica al problema della ricostruzione di una funzione dalle sue proiezioni venne data da J. Radon, *Über die Bestimmung von Funktionen durch ihre Integralwerte langs gewisser Mannigfaltigkeiten*, Ber. Saechsische Akad. Wiss. **29** (1917). In effetti, la *trasformata di Radon* è uno degli strumenti più importanti nell'ambito della ricostruzione delle immagini.

<sup>2</sup>Nello sviluppo della teoria è conveniente supporre che tali ampiezze, come pure le altre grandezze che definiscono i sistemi, abbiano precisione infinita. Nelle applicazioni, tuttavia, sarà necessario analizzare l'influenza degli errori introdotti dalla rappresentazione finita su calcolatore.

di una forma d'onda continua  $h(nT) = h(t)|_{t=nT}$  (come avviene, ad esempio, nei convertitori A/D da segnali analogici a segnali digitali.)

Analizziamo alcune sequenze che hanno un ruolo importante nello sviluppo della teoria del trattamento digitale dei segnali. La sequenza  $u_0(n)$  definita nel seguente modo (cfr. Figura 11.1)

$$u_0(n) = \begin{cases} 1 & \text{per } n = 0 \\ 0 & \text{per } n \neq 0 \end{cases} \quad (11.1)$$

è detta *impulso digitale* (o campionamento unitario) ed ha nei sistemi discreti lo stesso ruolo che ricopre l'impulso analogico (o funzione delta di Dirac  $\delta(t)$ , cfr. Appendice B) nei sistemi continui. La differenza sta nel fatto che l'impulso digitale è, in effetti, un particolare segnale, mentre il corrispondente impulso analogico esiste solo nel senso delle distribuzioni. Dalla definizione di  $u_0(n)$  si ottiene l'*impulso*

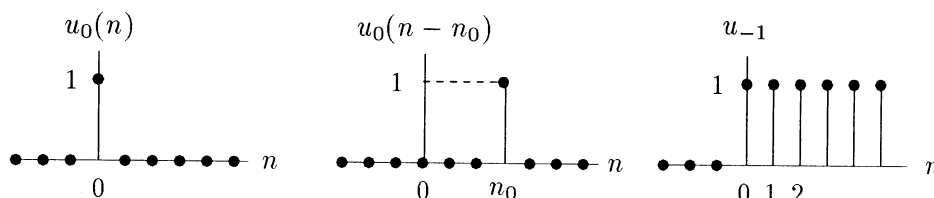


Figura 11.1: Alcune sequenze importanti nel trattamento digitale dei dati.

ritardato  $u_0(n - n_0)$ , per  $n_0$  (il cosiddetto ritardo, *delay*) fissato, definito come

$$u_0(n - n_0) = \begin{cases} 1 & \text{per } n = n_0 \\ 0 & \text{per } n \neq n_0 \end{cases} \quad (11.2)$$

La sequenza a salto unitario (*unit step*)  $u_{-1}(n)$  è definita come segue

$$u_{-1}(n) = \begin{cases} 1 & \text{per } n \geq 0 \\ 0 & \text{per } n < 0 \end{cases} \quad (11.3)$$

Si vede facilmente che tra la sequenza unit step  $u_{-1}(n)$  e l'impulso digitale  $u_0(n)$  sussiste la seguente relazione

$$u_{-1}(n) = \sum_{r=-\infty}^n u_0(r) \quad (11.4)$$

Altre importanti sequenze sono l'esponenziale decrescente (*decaying exponential*)  $g(n)$ , definita per ogni  $a \in \mathbb{R}$ , con  $a > 0$  da

$$g(n) = \begin{cases} a^n & \text{per } n \geq 0 \\ 0 & \text{per } n < 0 \end{cases} \quad (11.5)$$

e la funzione *sinusoidale*  $h(n)$  (cfr. Figura 11.2)

$$h(n) = \cos\left(\frac{2\pi n}{n_0}\right) \quad n \in \mathbb{Z} \quad (11.6)$$

Più in generale, si ha la funzione *esponenziale complessa*

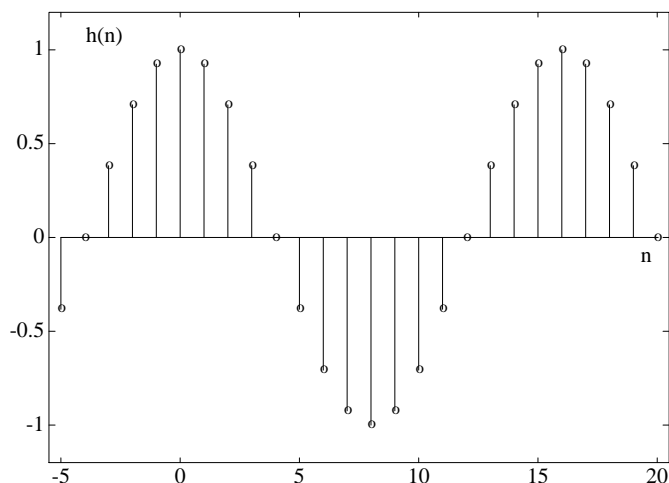


Figura 11.2: Sequenza sinusoidale corrispondente a  $\omega_0 := 2\pi/n_0 = 2\pi/16$ .

$$e^{i\omega n} = \cos(\omega n) + i \sin(\omega n), \quad i = \sqrt{-1}$$

Terminiamo, osservando che una sequenza arbitraria può essere rappresentata in termini di impulsi ritardati, opportunamente scalati. Consideriamo, infatti, una generica successione  $\{a(n)\}$ ,  $n = 0, 1, \dots$  nella quale  $a(n)$  rappresenta l'ampiezza del termine  $n$ -mo della successione. Si ha allora

$$\{a(n)\} = \sum_{m=-\infty}^{\infty} a(m) u_0(n-m) \quad (11.7)$$

### 11.1.1 Sistemi lineari invarianti nel tempo

Un sistema a tempo discreto (*discrete-time system*) è, in sostanza, un algoritmo per convertire una sequenza, detta *input* e indicata con  $x(n)$ , in un'altra sequenza, detta *output* e indicata con  $y(n)$  (cfr. Figura 11.3). In termini funzionali si ha

$$y(n) = \phi[x(n)] \quad (11.8)$$

ove  $\phi(\cdot)$  è determinata da un sistema specifico. Un sistema *lineare* è definito nel seguente modo. Se  $x_1(n)$  e  $x_2(n)$  sono input particolari al sistema lineare e  $y_1(n)$ ,

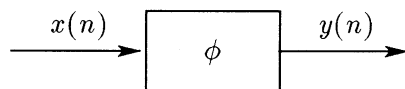


Figura 11.3: Rappresentazione di un sistema a tempo discreto.

$y_2(n)$  sono i rispettivi output, allora, se si applica come input la sequenza  $ax_1(n) + bx_2(n)$ , si ottiene come output la sequenza  $ay_1(n) + by_2(n)$ , ove  $a$  e  $b$  sono costanti arbitrarie.

In un sistema invariante nel tempo (*time-invariant*), se la sequenza in input  $x(n)$  produce una sequenza output  $y(n)$ , allora la sequenza input  $x(n-n_0)$  produce la sequenza output  $y(n-n_0)$  per ogni  $n_0$ .

Mostreremo ora che i sistemi lineari e invarianti nel tempo, indicati nel seguito per brevità, con LTI (linear, time-invariant), possono essere completamente caratterizzati dalla sequenza  $h(n)$ , definita come la risposta del sistema a un impulso digitale (ossia,  $h(n)$  è l'output del sistema, quando in input si ha la sequenza  $u_0(n)$  definita in (11.1)). La sequenza  $h(n)$  è chiamata risposta all'impulso (*impulse response* o *unit sample response*).

Se  $x(n)$  è l'input ad un sistema LTI, dalla rappresentazione (11.7) si ha

$$x(n) = \sum_{m=-\infty}^{\infty} x(m) u_0(n-m) \quad (11.9)$$

Dal momento che  $h(n)$  è la risposta alla sequenza  $u_0(n)$ , per la proprietà di invarianza temporale si ha che  $h(n-m)$  è la risposta alla sequenza  $u_0(n-m)$ . Ugualmente, per la proprietà di linearità la risposta alla sequenza  $x(m)u_0(n-m)$  è data da  $x(m)h(n-m)$ . Pertanto la risposta a  $x(n)$  è data da

$$y(n) = \sum_{m=-\infty}^{\infty} x(m)h(n-m) \quad (11.10)$$

L'equazione (11.10), che rappresenta una relazione di convoluzione<sup>3</sup>, può essere riscritta, mediante un cambiamento di variabile nella seguente forma equivalente

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)x(n-m)$$

Si vede, quindi, che la successione  $h(n)$  caratterizza completamente un sistema LTI (cfr. Figura 11.4).

<sup>3</sup>La *convoluzione* di due successioni  $\{a_n\}$  e  $\{b_n\}$  è la successione  $\{c_n\}$ , ove  $c_n = \sum_k a_k b_{n-k}$ .

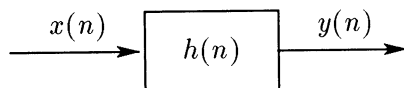


Figura 11.4: Rappresentazione di un sistema LTI.

Come esemplificazione, in Figura 11.5 è rappresentato l'output  $y(n)$  di un LTI corrispondente ad un particolare input  $x(n)$  e ad una particolare risposta all'impulso  $h(n)$ .

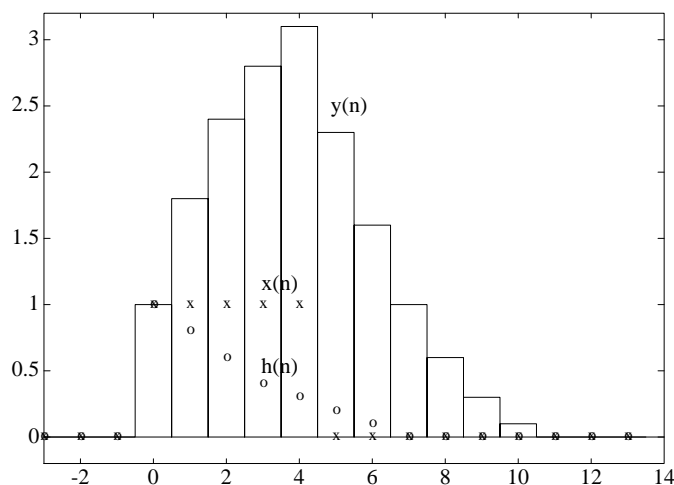


Figura 11.5: Output  $y(n)$  in un sistema LTI corrispondente ad un particolare input  $x(n)$  e a una particolare funzione  $h(n)$ .

**Causalità e stabilità** Un sistema LTI viene detto causale o *realizzabile* (realizable) se l'output per  $n = n_0$  dipende unicamente dai valori dell'input per  $n \leq n_0$ ; la proprietà è equivalente a dire che la risposta all'impulso  $h(n)$  è nulla per  $n < 0$ . La realizzabilità è una proprietà importante dal punto di vista applicativo, ma esistono sistemi non realizzabili che godono di interessanti proprietà teoriche. In pratica, nella teoria dei filtri si cerca di ottenere delle buone approssimazioni di sistemi ottimali non realizzabili attraverso sistemi realizzabili.

Un sistema LTI viene detto *stabile* quando ogni input limitato, ossia  $|x(n)| \leq C$ , con  $C$  costante positiva, produce un output limitato. Si può dimostrare facilmente che *condizione necessaria e sufficiente* affinché un sistema LTI sia stabile è che per

la funzione  $h(n)$  si abbia

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty \quad (11.11)$$

La dimostrazione, che lasciamo come esercizio, è basata sulla rappresentazione (11.10).

### 11.1.2 Equazioni alle differenze

Nell'ambito dei sistemi LTI una particolare importanza hanno i sistemi nei quali la relazione funzionale tra la sequenza di input  $x(n)$  e la sequenza di output  $y(n)$  è rappresentata da una *equazione alle differenze lineare a coefficienti costanti*. L'espressione più generale di una equazione alle differenze a coefficienti costanti di ordine  $M$  relativa a un sistema causale ha la seguente forma

$$y(n) = \sum_{i=0}^M b_i x(n-i) - \sum_{i=1}^M a_i y(n-i) \quad n \geq 0 \quad (11.12)$$

con  $a_M \neq 0$ . Dato un insieme di condizioni iniziali  $(x(i), y(i), i = -1, -2, \dots, -M)$  e la sequenza  $x(n)$ , l'equazione (11.12) fornisce in maniera esplicita (ricorsiva) la sequenza output  $y(n)$ . Alternativamente, si può ottenere la funzione  $y(n)$  in forma chiusa (cioè come funzione di  $n$ ) considerando la soluzione generale dell'equazione alle differenze e imponendo le condizioni iniziali. L'argomento è già stato trattato nel precedente Capitolo 7. Per comodità richiamiamo ora le idee essenziali attraverso un esempio.

► **Esempio 11.1** Consideriamo la seguente equazione alle differenze

$$y(n) = x(n) - 3y(n-1)$$

con la condizione iniziale  $y(-1) = 0$  e con input  $x(n) = n^2 + n$ . L'equazione omogenea  $y(n) + 3y(n-1) = 0$  ha come soluzione generale la successione  $c(-3)^n$ , con  $c$  costante arbitraria. Una soluzione particolare dell'equazione completa è cercata della forma  $An^2 + Bn + C$ . Sostituendo nell'equazione si trova  $A = 1/4$ ,  $B = 5/8$  e  $C = 9/32$ . La soluzione generale è allora data dalla seguente successione

$$y(n) = \frac{n^2}{4} + \frac{5n}{8} + \frac{9}{32} + c(-3)^n$$

Imponendo la condizione iniziale  $y(-1) = 0$ , si trova  $c = -9/32$  e quindi la soluzione finale è data da

$$y(n) = \frac{n^2}{4} + \frac{5n}{8} + \frac{9}{32}[1 - (-3)^n]$$

■

▼ **Osservazione 11.1** *Nel trattamento dei segnali le equazioni alle differenze presentano un interesse pratico, in quanto a partire da esse è facile la realizzazione di un sistema.*

Come esemplificazione, in Figura 11.6 è rappresentata una realizzazione dell'equazione alle differenze generale del primo ordine

$$y(n) = -a_1y(n-1) + b_0x(n) + b_1x(n-1) \quad (11.13)$$

In modo analogo si può realizzare la generica equazione alle differenze del secondo ordine

$$y(n) = -a_1y(n-1) - a_2y(n-2) + b_0x(n) + b_1x(n-1) + b_2x(n-2) \quad (11.14)$$

I sistemi di ordine superiore possono essere decomposti in combinazioni, in cascata o in parallelo, di sistemi del primo e del secondo ordine. Per questo aspetto importante nelle applicazioni rinviamo alla bibliografia, in particolare a Rabiner e Gold [132]. ■

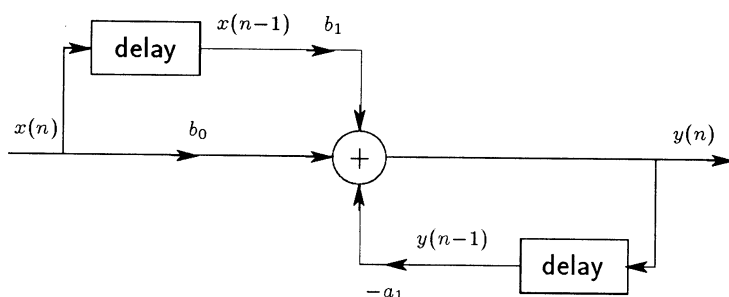


Figura 11.6: Realizzazione di una equazione alle differenze del primo ordine.

### 11.1.3 Rappresentazione in dominio di frequenza

Consideriamo la classe di sequenze input della seguente forma

$$x(n) = e^{i\omega n} \quad -\infty < n < \infty \quad (11.15)$$

Se tale input è applicato a un sistema LTI con risposta all'impulso  $h(n)$ , dalla (11.10) si ottiene

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)e^{i\omega(n-m)} = e^{i\omega n} \sum_{m=-\infty}^{\infty} h(m)e^{-i\omega m} = x(n)H(e^{i\omega}) \quad (11.16)$$

Ossia, per gli input della forma (11.15) l'output è uguale all'input moltiplicato per il seguente fattore complesso, detto risposta in frequenza (*frequency response*).

$$H(e^{i\omega}) := \sum_{n=-\infty}^{\infty} h(n)e^{-i\omega n} \quad (11.17)$$



Osserviamo che la funzione  $H(e^{i\omega})$  è periodica con periodo  $2\pi$ . In effetti, una sequenza input di frequenza  $\omega + 2m\pi$ , con  $m = \pm 1, \pm 2, \dots$  coincide con la sequenza input di frequenza  $\omega$

$$\hat{x}(n) = e^{i(\omega+2m\pi)n} = e^{i\omega n} = x(n)$$

Inoltre, quando  $h(n)$  è una funzione reale (come nell'esempio precedente), il modulo di  $H(e^{i\omega})$  è simmetrico e la fase è antisimmetrica sull'intervallo  $0 \leq \omega < 2\pi$ .

► **Esempio 11.2** Come esemplificazione, consideriamo il sistema LTI rappresentato dalla seguente equazione alle differenze del primo ordine

$$y(n) = x(n) + ay(n-1), \quad |a| < 1 \quad (11.18)$$

con la condizione iniziale  $y(-1) = 0$ . Si vede facilmente che la risposta all'impulso  $h(n)$  è data da  $a^n u_{-1}(n)$ . Per esso si ottiene

$$H(e^{i\omega}) := \sum_{n=0}^{\infty} a^n e^{-i\omega n} = \sum_{n=0}^{\infty} (ae^{-i\omega})^n = \frac{1}{1 - ae^{-i\omega}}$$

In Figura 11.7 sono rappresentati i grafici del modulo e della fase<sup>4</sup> di  $H(e^{i\omega})$  per  $\omega$  nell'intervallo  $0 \leq \omega \leq 2\pi$  e  $a = 0.5$ . Dalla Figura si vede che il sistema (11.18) attenua le frequenze alte, ossia, come si dice, ha caratteristiche di filtro passa-basso (*low-pass filter*).

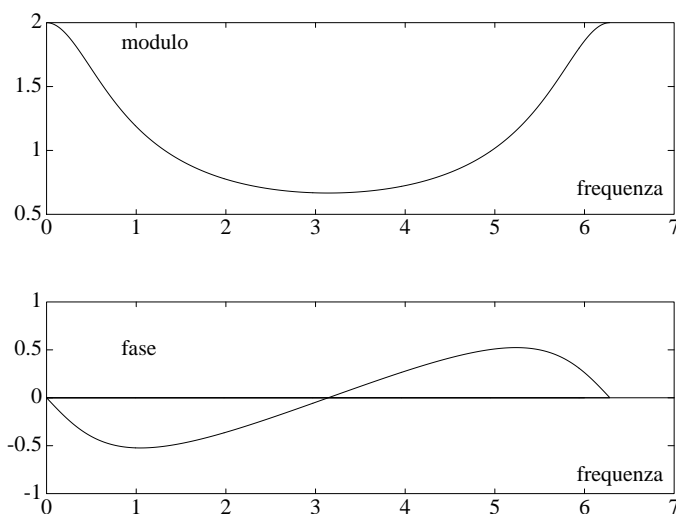


Figura 11.7: Esempio di risposta in frequenza in un sistema LTI del primo ordine.

In modo analogo si studia il seguente sistema del secondo ordine

$$y(n) = x(n) + a_1y(n-1) + a_2y(n-2) \quad (11.19)$$

<sup>4</sup>Ricordiamo che per un numero complesso  $h = x + iy = me^{ip}$ , si chiama modulo (*magnitude*) il numero reale non negativo  $m$  e fase (*phase*) il numero reale  $p$ . Usualmente il modulo è indicato con  $|h|$  e la fase con il simbolo  $\angle h$ .

con le condizioni iniziali  $y(-1) = y(-2) = 0$ . Indicate con  $p_1$  e  $p_2$  le radici (supposte, per semplicità, distinte) del polinomio  $p^2 - a_1p - a_2$ , la risposta all'impulso  $h(n)$  è della forma

$$h(n) = \alpha_1 p_1^n + \alpha_2 p_2^n \quad (I)$$

se  $p_1, p_2$  sono reali e della forma

$$h(n) = \alpha_1 r^n \sin(bn + \phi) \quad (II)$$

quando le radici sono complesse di modulo  $r$  e fase  $b$ . Le quantità  $\alpha_1, \alpha_2$  e  $\phi$  sono costanti da determinare in base alle condizioni iniziali.

Il caso (I) rappresenta, in sostanza, due sistemi lineari del primo ordine e la risposta all'impulso è determinata da  $p_1^n$  e  $p_2^n$ . Nel caso (II), nel quale si deve avere  $a_2 < 0$  e  $a_2 < -a_1^2/4$ , la risposta all'impulso è una sinusoidale per la quale si può mostrare che la risposta in frequenza può essere scritta nella seguente forma

$$H(e^{i\omega}) = \frac{1}{1 - 2r \cos b e^{-i\omega} + r^2 e^{-2i\omega}}$$

ove  $r = \sqrt{-a_2}$ ,  $\cos b = a_1/(2r)$ ,  $\phi = b$  e  $\alpha_1 = 1/\sin b$ . In Figura 11.8 sono rappresentati i grafici del modulo e della fase corrispondenti a  $b = \pi/4$  e  $r = 0.7$ . La figura suggerisce che il sistema considerato si comporta come un risonatore (*resonator*), ossia esalta un segnale corrispondente ad una particolare frequenza. ■

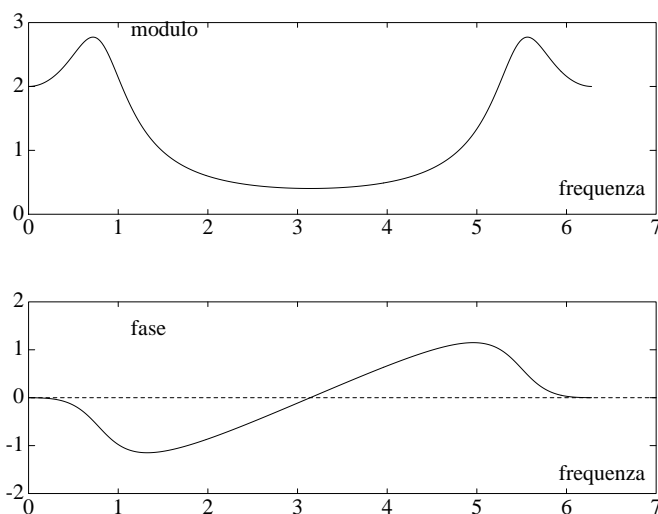


Figura 11.8: Esempio di risposta in frequenza in un sistema LTI del secondo ordine.

Dal momento che la risposta in frequenza è una funzione periodica di  $\omega$ , l'equazione (11.17) può essere interpretata come la rappresentazione in serie di Fourier di  $H(e^{i\omega})$ , ove i coefficienti  $h(n)$  sono ottenuti da  $H(e^{i\omega})$  nel modo seguente

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(e^{i\omega}) e^{i\omega n} d\omega \quad (11.20)$$

Più in generale, per ogni input  $x(n)$  per il quale esiste finita la seguente serie

$$X(e^{i\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-i\omega n} \quad (11.21)$$

si ha la rappresentazione

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{i\omega})e^{i\omega n} d\omega \quad (11.22)$$

e le (11.21), (11.22) rappresentano una coppia di relazioni di *serie di Fourier discreta*. Utilizzando la linearità, dalla relazione (11.22) si ottiene per la risposta  $y(n)$  la seguente rappresentazione

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} Y(e^{i\omega})e^{i\omega n} d\omega \quad (11.23)$$

ove

$$Y(e^{i\omega}) = X(e^{i\omega})H(e^{i\omega}) \quad (11.24)$$

Come si vede, la convoluzione nel dominio del tempo è convertito a una moltiplicazione nel dominio della frequenza. Sottolineiamo, inoltre, l'importanza della risposta in frequenza, dal momento che una arbitraria sequenza input  $x(n)$  può essere rappresentata (cfr. (11.22)) come una sovrapposizione delle sequenze input  $e^{i\omega n}$ , per  $0 \leq \omega \leq 2\pi$ .

▼ **Osservazione 11.2** Se  $T$  è l'ampiezza dell'intervallo di campionamento, le relazioni (11.17), (11.20) possono essere riscritte nella seguente forma

$$H(e^{i\omega T}) = \sum_{n=-\infty}^{\infty} h(nT)e^{-i\omega nT} \quad (11.25)$$

$$h(nT) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} H(e^{i\omega T})e^{i\omega nT} d\omega \quad (11.26)$$

La funzione  $H(e^{i\omega T})$  risulta, allora, periodica in  $\omega$  (misurato in radianti per secondo) con un periodo di  $2\pi/T$ . Alternativamente, si può sostituire  $\omega$  con  $2\pi f$ , con  $f$  misurata in hertz (Hz). ■

#### 11.1.4 Relazione tra sistemi continui e discreti

Quando la sequenza  $x(nT)$  è ottenuta per campionamento, con passo  $T$ , da una forma d'onda continua (analogica)  $x(t)$ , si possono mettere in relazione tra loro la risposta in frequenza della sequenza  $X(e^{i\omega T})$  e la trasformata di Fourier  $X_A(i\Omega)$  della forma d'onda continua nel seguente modo.

Le relazioni che definiscono la trasformata di Fourier della forma d'onda continua sono date da

$$X_A(i\Omega) = \int_{-\infty}^{\infty} x(t)e^{-i\Omega t} dt \quad (11.27)$$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_A(i\Omega)e^{i\Omega t} d\Omega \quad (11.28)$$

mentre per la forma d'onda discreta le relazioni sono date da

$$X(e^{i\omega T}) = \sum_{n=-\infty}^{\infty} x(nT)e^{-i\omega nT} \quad (11.29)$$

$$x(nT) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} X(e^{i\omega T})e^{i\omega nT} d\omega \quad (11.30)$$

Se  $x(nT) = x(t)|_{t=nT}$ , è possibile porre in relazione  $X_A(i\Omega)$  e  $X(e^{i\omega T})$  valutando l'equazione (11.28) in  $t = nT$  e scrivendo

$$x(nT) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} \int_{(2m-1)\pi/T}^{(2m+1)\pi/T} X_A(i\Omega)e^{i\Omega nT} d\Omega \quad (11.31)$$

Per sostituzione e chiamando  $\omega$  la variabile di integrazione, si ha

$$x(nT) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \left[ \frac{1}{T} \sum_{m=-\infty}^{\infty} X_A(\omega + \frac{2\pi}{T}m) \right] e^{i\omega nT} d\omega \quad (11.32)$$

Per uguaglianza tra i termini entro la parentesi quadra della (11.32) e la (11.30) si ottiene, infine, la relazione cercata

$$X(e^{i\omega T}) = \frac{1}{T} \sum_{m=-\infty}^{\infty} X_A(\omega + \frac{2\pi}{T}m) \quad (11.33)$$

Tale relazione mette in evidenza il fatto che le risposte in frequenza delle sequenze consistono della somma di un numero infinito di componenti della risposta in frequenza della forma d'onda analogica. Nel caso particolare in cui il passo di campionamento  $T$  è tale che  $X_A(i\Omega) = 0$  se  $|\Omega| > \pi/T$ , allora per  $|\omega| \leq \pi/T$  si ha (*sampling theorem*, cfr. Figura 11.9 per una esemplificazione)

$$X(e^{i\omega T}) = \frac{1}{T} X_A(\omega) \quad (11.34)$$

Il valore minimo della frequenza di campionamento  $1/T$  per la quale si ha  $X_A(i\Omega) = 0$ , per  $|\Omega| > \pi/T$ , è detta *frequenza di Nyquist*. Per valori superiori di  $T$  si può verificare uno spostamento (shifting) di informazioni relative ad alte frequenze in  $X_A(i\Omega)$  nelle basse frequenze in  $X(e^{i\omega T})$  (cfr. Figura 11.10). Tale fenomeno, detto *aliasing* può essere, naturalmente, evitato campionando il segnale analogico ad una velocità sufficientemente alta, ossia assumendo  $T$  sufficientemente piccolo.

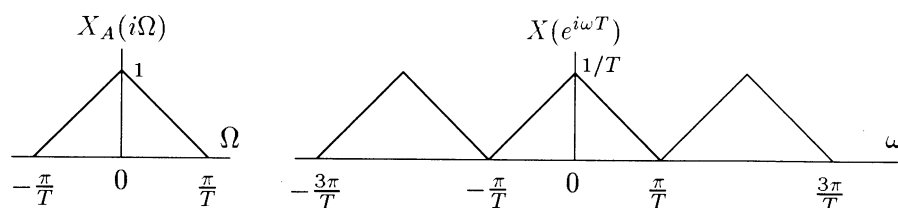


Figura 11.9: Una esemplificazione di relazione tra sistema analogico e digitale.

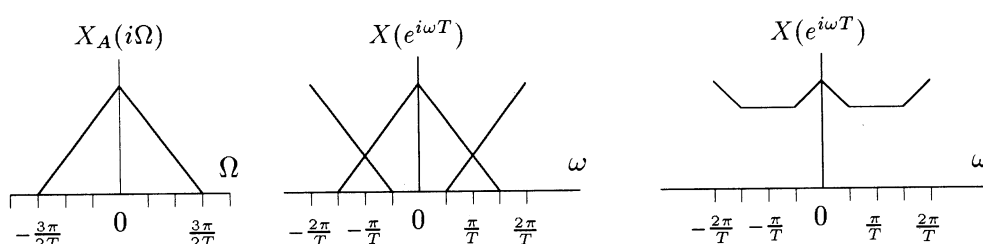


Figura 11.10: Effetto del sottocampionamento sulla risposta in frequenza digitale.

### 11.1.5 La trasformata $z$

La *trasformata  $z$*  rappresenta una tecnica particolarmente utile per rappresentare ed analizzare le sequenze, e quindi i sistemi discreti<sup>5</sup>. Data una sequenza  $x(n)$ , definita per tutti i valori di  $n$ , la sua trasformata  $z$  è definita in forma di *serie di potenze* nel modo seguente

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (11.35)$$

ove  $z$  è una variabile complessa. La funzione complessa di variabile complessa  $X(z)$  è definita solo per i valori di  $z$  per i quali la serie di potenze risulta convergente. Tale problema può essere quindi affrontato adeguatamente mediante i risultati della *teoria delle funzioni di variabile complessa*. Per il seguito ci limiteremo a considerare alcuni casi particolari di interesse nell'analisi dei filtri digitali.

**Sequenze di durata finita** Nel caso particolare in cui la sequenza  $x(n)$  è differente dallo zero solo nell'intervallo  $[n_1, n_2]$ , con  $n_1$  e  $n_2$  interi finiti e  $n_1 < n_2$ , ossia

<sup>5</sup>La trasformata  $z$  nei sistemi discreti gioca, in sostanza, un ruolo simile a quello svolto dalla trasformata di Laplace (cfr. Appendice B) nei sistemi continui: trasforma le equazioni alle differenze in equazioni algebriche.

$x(n)$  è di *durata finita*, si ha la somma di un numero finito di termini e quindi  $X(z)$  è definita in tutto il piano complesso  $\mathbb{C}$ , salvo eventualmente  $z = 0$ , oppure  $z = \infty$ .

Un sistema LTI, la cui risposta ad un impulso  $h(n)$  è di durata finita è detto sistema, o filtro, *finite impulse response* (FIR), o non ricorsivo. Al contrario, se  $n_1 = -\infty$  o  $n_2 = +\infty$ , il filtro è detto *infinite impulse response* (IIR), o ricorsivo. In particolare nell'ambito dei processi stocastici di filtraggio, un filtro FIR è anche indicato come filtro *moving average* (MA) e un filtro IIR come filtro *autoregressive* (AR).

Se ogni elemento della sequenza  $\{h(n)\}$  è un numero finito, si ha che il criterio di stabilità (11.11) è verificato in quanto si ha una somma di un numero finito di elementi. Pertanto, un filtro FIR è *stabile*; esso può essere inoltre reso realizzabile ritardando opportunamente la risposta all'impulso (di  $-n_1$ , se  $n_1 < 0$ ).

Nell'esempio successivo è indicato il calcolo della trasformata  $z$  di alcune sequenze interessanti.

► **Esempio 11.3** Se  $x(n)$  è la sequenza impulso  $u_0(n)$ , si ha

$$X(z) = 1$$

mentre per la sequenza  $u_{-1}(n)$ , nulla per  $n < 0$  e uguale a 1 per  $n \geq 0$  si trova

$$X(z) = \sum_{n=0}^{\infty} z^{-n} = \frac{1}{1 - z^{-1}}$$

che è definita per  $|z| > 1$ . Più in generale, per la sequenza  $x(n) = 0$ ,  $n < 0$  e  $x(n) = a^n$ ,  $n \geq 0$ , con  $a > 0$ , si ha

$$X(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \sum_{n=0}^{\infty} (az^{-1})^n = \frac{1}{1 - az^{-1}}$$

che è definita per  $|z| > a$ . Analogamente, per la sequenza esponenziale complessa  $x(n) = 0$ ,  $n < 0$  e  $x(n) = e^{in\omega}$ ,  $n \geq 0$  si trova

$$X(z) = \frac{1}{1 - z^{-1}e^{i\omega}}$$

■

### Proprietà della trasformata $z$

**Linearità** Se  $X_1(z)$  è la trasformata  $z$  di  $x_1(n)$  e  $X_2(z)$  è la trasformata  $z$  di  $x_2(n)$ , allora la trasformata  $z$  di  $ax_1(n) + bx_2(n)$  è data da  $aX_1(z) + bX_2(z)$  per ogni numero reale  $a$  e  $b$ .

**Delay** Se una sequenza  $x_1(n)$  ha  $X_1(z)$  come trasformata  $z$ , allora la sequenza  $x_1(n - n_0)$  ha come trasformata  $z$  la funzione  $z^{-n_0}X_1(z)$  per ogni  $n_0$ . Tale proprietà

è utile, in particolare, per fornire la trasformata  $z$  di una equazione alle differenze che rappresenta un sistema LTI. Ad esempio, la seguente equazione alle differenze

$$y(n) = x(n) - b_1 y(n-1) - b_2 y(n-2) \quad (11.36)$$

ha come trasformata  $z$  la funzione

$$Y(z) = X(z) - b_1 z^{-1} Y(z) - b_2 z^{-2} Y(z) \Rightarrow Y(z) = \frac{X(z)}{1 + b_1 z^{-1} + b_2 z^{-2}}$$

ove

$$Y(z) = \sum_{n=-\infty}^{\infty} y(n) z^{-n}, \quad X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}$$

**Convoluzione di sequenze** Se  $x(n)$  è l'input di un sistema discreto LTI con risposta all'impulso  $h(n)$  e  $y(n)$  è l'output, allora

$$Y(z) = X(z)H(z) \quad (11.37)$$

ove  $H(z)$  è la trasformata  $z$  di  $h(n)$  ed è nota come funzione di trasferimento (*transfer function*). Ad esempio, il sistema corrispondente all'equazione (11.36) è caratterizzato dalla seguente funzione di trasferimento

$$H(z) = \frac{1}{1 + b_1 z^{-1} + b_2 z^{-2}}$$

Più in generale, per un generico sistema LTI si ha

$$H(z) = \frac{\sum_{k=0}^M a_k z^{-k}}{1 + \sum_{k=1}^M b_k z^{-k}}$$

Il principale compito nel progetto di un filtro consiste nella determinazione di una funzione di trasferimento  $H(z)$  (di complessità minima) che soddisfi a determinate specifiche, rispetto all'ampiezza e alla fase della risposta, o alla banda di frequenze che si desidera tagliare<sup>6</sup>.

### 11.1.6 La trasformata di Fourier discreta

Una successione periodica  $x_p(n)$  di periodo  $N$  può essere rappresentata nella seguente forma

$$x_p(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_p(k) e^{i(2\pi/N)kn} \quad (11.38)$$

<sup>6</sup>I filtri possono essere utilizzati anche per rappresentare la dinamica di un sistema. In questo caso il problema della costruzione di un filtro è un *problema inverso*: noti i dati di input e di output, si tratta di trovare il sistema che riproduce la relazione input-output rilevata sperimentalmente (*identificazione di un sistema*).

ove  $X_p(k)$  rappresenta l'ampiezza della sinusoida corrispondente alla frequenza  $\omega_k = 2\pi k/N$ . Si può vedere facilmente (moltiplicando la relazione (11.38) per  $e^{-i(2\pi/N)kn}$  e sommando in  $n$ ) che

$$X_p(k) = \sum_{n=0}^{N-1} x_p(n) e^{-i(2\pi/N)nk} \quad \boxed{X_p(k) = \text{DFT}(x(n))} \quad (11.39)$$

L'equazione (11.39) è nota come *trasformata di Fourier discreta* (DFT), e la (11.38) è chiamata la *trasformata di Fourier discreta inversa* (IDFT). Dalle equazioni (11.38), (11.39) si vede che le sequenze  $x_p(n)$  e  $X_p(k)$  sono ambedue periodiche di periodo  $N$  e che  $X_p(k)$  può essere determinata esattamente da un periodo di  $x_p(n)$ . La sequenza  $X_p(k)$  fornisce le componenti della frequenza di  $x_p(n)$  e, quindi, è chiamata lo *spettro* del segnale  $x_p(n)$ , o la descrizione nel dominio della frequenza di  $x_p(n)$ . Un aspetto interessante delle relazioni DFT, IDFT è la possibilità che esse offrono di rappresentare sequenze di lunghezza finita. Data, infatti, una sequenza  $x(n)$  di lunghezza finita  $N$ , è sufficiente applicare le considerazioni precedenti al suo prolungamento periodico  $x_p$  di periodo  $N$ .

Ad esempio, si verifica facilmente che la trasformata di Fourier discreta di un impulso  $x_0(n)$  è data da  $X(k) = 1$  per ogni  $k$ . Ossia, un impulso ha componenti uguali in tutte le frequenze.

Come ulteriore illustrazione, in Figura 11.11 sono riportati i grafici del modulo e dell'angolo della trasformata di Fourier discreta della sequenza  $x(n) = a^n$ , per  $a = 0.9$  e  $N = 30$ .

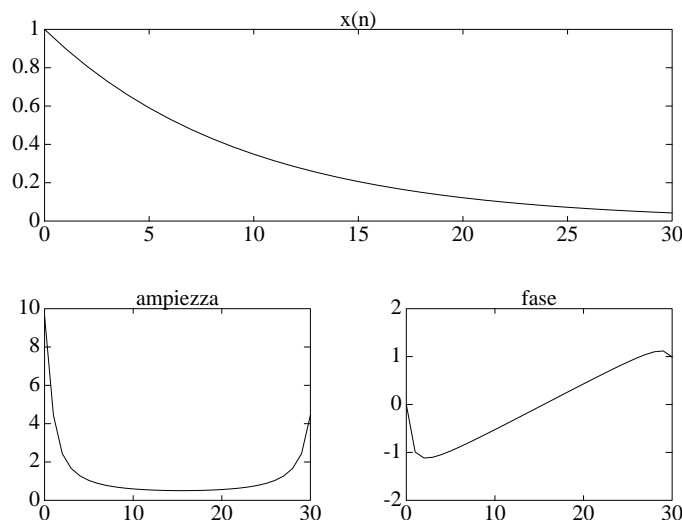


Figura 11.11: Modulo e fase della trasformata di Fourier discreta della sequenza  $x(n) = a^n$ , per  $a = 0.9$  e  $N = 30$ .



### Proprietà della DFT

**Linearità** La trasformata DFT di una combinazione lineare di sequenze periodiche (dello stesso periodo) coincide con la combinazione lineare delle corrispondenti DFT.

**Proprietà shifting** Se  $x_p(n)$  è una sequenza periodica, di periodo  $N$ , con DFT  $X_p(k)$ , allora la sequenza  $x_p(n-n_0)$  ha come DFT la sequenza  $X_p(k)e^{-i(2\pi/N)n_0k}$ . In altre parole uno spostamento nella sequenza del segnale risulta in uno spostamento della fase nello spettro.

**Proprietà simmetrica** Se  $x_p(n)$  è una sequenza *reale*, periodica di periodo  $N$ , allora la trasformata DFT  $X_p(k)$  verifica le seguenti proprietà di simmetria

$$\begin{aligned}\Re[X_p(k)] &= \Re[X_p(N-k)]; & \Im[X_p(k)] &= -\Im[X_p(N-k)] \\ |X_p(k)| &= |X_p(N-k)|; & \angle X_p(k) &= -\angle X_p(-k)\end{aligned}$$

Se, inoltre,  $x_p(n)$  è simmetrica, ossia  $x_p(n) = x_p(N-n)$ , allora  $X_p(k)$  è *reale*.

**Convoluzione di sequenze** Se  $x_p(n)$  e  $h_p(n)$  sono due sequenze periodiche di periodo  $N$  e  $X_p(k)$ ,  $H_p(k)$  sono le corrispondenti DFT, la DFT della sequenza  $y_p(n)$ , la *convoluzione periodica* di  $x_p(n)$  e  $h_p(n)$ , definita da

$$y_p(n) := \sum_{l=0}^{N-1} x_p(l)h_p(n-l) \quad (11.40)$$

è data da

$$\boxed{Y_p(k) = H_p(k) \cdot X_p(k)} \quad (11.41)$$

Questa proprietà è utilizzata per ridurre la quantità di operazioni necessarie per il calcolo della convoluzione, utilizzando la seguente equazione

$$y(n) = \text{IDFT}[\text{DFT}(x(n)) \text{DFT}(h(n))]$$

purché, naturalmente, si utilizzino algoritmi efficienti per il calcolo di DFT e di IDFT.

### Calcolo della DFT

Per il calcolo della DFT di una sequenza di  $N$  punti, dalla definizione sono necessarie  $N$  moltiplicazioni complesse e  $N-1$  addizioni complesse per ciascun coefficiente, e quindi approssimativamente un totale di  $N^2$  moltiplicazioni complesse e  $N(N-1)$  addizioni complesse per il calcolo di tutta la sequenza  $X_p(k)$ . In alcune applicazioni tale quantità di calcoli può rendere impraticabile l'utilizzo della trasformata. Vi

sono, tuttavia, algoritmi, solitamente indicati con il generico nome di *Fast Fourier Transform* (FFT) e originati in sostanza da un algoritmo accreditato a Cooley e Tukey (1965), con i quali il numero totale di operazioni si riduce a  $(N/2) \log_2 N$  moltiplicazioni complesse e  $N \log_2 N$  addizioni complesse. Ad esempio, per  $N = 1024$  (che può corrispondere, nell'analisi di una figura bidimensionale, a una griglia di soli  $32 \times 32$  punti) il risparmio è di 100 a 1.

In questo paragrafo daremo l'idea di base degli algoritmi FFT in un caso particolare. Per un approfondimento si veda ad esempio Rabiner e Gold [132].

**FFT nel caso in cui  $N = 2^k$**  Semplifichiamo le notazioni, riscrivendo la (11.39) nella seguente forma

$$c_j = \sum_{\beta=0}^{N-1} a_{\beta} w^{j\beta}, \quad w = e^{-2\pi i/N} \quad (11.42)$$

Posto  $\beta = 2\beta_1$ , quando  $\beta$  è pari, e  $\beta = 2\beta_1 + 1$  quando  $\beta$  è dispari, con  $0 \leq \beta_1 \leq N/2 - 1$ , si ottiene

$$c_j = \sum_{\beta_1=0}^{N/2-1} a_{2\beta_1} (w^2)^{j\beta_1} + \sum_{\beta_1=0}^{N/2-1} a_{2\beta_1+1} (w^2)^{j\beta_1} w^j$$

Se  $\alpha$  e  $j_1$  sono tali che  $j = \alpha N/2 + j_1$ , essendo  $w^N = 1$ , si ha

$$(w^2)^{j\beta_1} = (w^2)^{\alpha(N/2)\beta_1} (w^2)^{j_1\beta_1} = (w^N)^{\alpha\beta_1} (w^2)^{j_1\beta_1} = (w^2)^{j_1\beta_1}$$

Pertanto, se poniamo per  $j_1 = 0, 1, \dots, N/2 - 1$ ,

$$\begin{aligned} \phi(j_1) &= \sum_{\beta_1=0}^{N/2-1} a_{2\beta_1} (w^2)^{j_1\beta_1} \\ \psi(j_1) &= \sum_{\beta_1=0}^{N/2-1} a_{2\beta_1+1} (w^2)^{j_1\beta_1} \end{aligned}$$

ove  $(w^2)^{N/2} = 1$ , si ottiene

$$\boxed{c_j = \phi(j_1) + w^j \psi(j_1)} \quad j = 0, 1, \dots, N - 1$$

Osserviamo ora che il calcolo di  $\phi(j_1)$  e di  $\psi(j_1)$  equivale al calcolo di due trasformate di Fourier con  $\frac{N}{2} = 2^{k-1}$ , anziché di una trasformata con  $N = 2^k$  termini. Si procede, quindi, in maniera *ricorsiva*, applicando la stessa idea alle nuove trasformate. Se si indica con  $p_k$  il tempo totale di operazioni necessarie per il calcolo dei coefficienti  $c_j$ , quando  $N = 2^k$ , si ha

$$p_k \leq 2p_{k-1} + 2 \cdot 2^k, \quad k = 1, 2, \dots$$

Dal momento che  $p_0 = 0$ , si ha per induzione che  $p_k \leq 2k \cdot 2^k = 2N \cdot \log_2 N$ . La procedura può essere estesa al caso in cui  $N = r_1 r_2 \cdots r_p$ . Di seguito è riportata, come esemplificazione una implementazione in FORTRAN della FFT, quando  $N$  è una potenza di 2.

```

      SUBROUTINE FFT(A,M)
C     A vettore di numeri complessi di dimensione N
C     in INPUT: vettore X(N)
C     In OUTPUT: trasformata di fourier discreta di X(N)
      COMPLEX A(*),U,W,T
      N=2**M
      NV2=N/2
      NM1=N-1
      J=1
      DO 7 I=1,NM1
        IF(I.GE.J) GO TO 5
        T=A(J)
        A(J)=A(I)
        A(I)=T
5       K=NV2
6       IF(K.GE.J) GO TO 7
        J=J-K
        K=K/2
        GO TO 6
7      J=J+K
      DO 20 L=1,M
        LE=2**L
        LE1=LE/2
        U=(1.,0.)
        ANG=3.14159265358979/LE1
        W=CMPLX(COS(ANG),SIN(ANG))
        DO 20 J=1,LE1
          DO 10 I=J,N,LE
            IP=I+LE1
            T=A(IP)*U
            A(IP)=A(I)-T
10         A(I)=A(I)+T
20      U=U*W
      RETURN
      END

```

◆ **Esercizio 11.1** Determinare le proprietà dei sistemi descritti dalle seguenti relazioni input/output

$$\begin{aligned}
 \text{(a) } y(n) &= e^{(n)}, & \text{(b) } y(n) &= x(n+1)x(n-1) \\
 \text{(c) } y(n) &= \sum_{k=-\infty}^{n+2} x(k), & \text{(d) } y(n) &= x(2n)
 \end{aligned}$$

◆ **Esercizio 11.2** Stabilire se i seguenti sistemi LTI sono stabili

$$(a) h(n) = (n+1)2^{-n}u_{-1}(n), \quad (b) h(n) = (-1)^n u_{-1}(n), \quad (c) h(n) = \frac{u_{-1}(n)}{n!}$$

ove  $u_{-1}(n)$  indica la funzione unit-step.

◆ **Esercizio 11.3** Analizzare la risposta all'impulso (impulse response)  $h(n)$  corrispondente ai seguenti sistemi LTI, indicando se si tratta di sistemi FIR o IIR

$$(a) y(n) = x(n) - 2x(n-1) + 2x(n-2) - x(n-3)$$

$$(b) y(n) - \frac{1}{2}y(n-2) = 2x(n) - x(n-2), \quad (c) y(n) - \sqrt{2}y(n-1) + y(n-2) = x(n)$$

## 11.2 Analisi dei segnali mediante MATLAB

Nel sistema integrato MATLAB esistono numerosi strumenti numerici utili per l'analisi dei segnali. Segnaliamo, in particolare, l'aritmetica complessa, il calcolo della convoluzione, gli algoritmi di algebra lineare e il calcolo della FFT. Alcune altre funzioni più specializzate sono raccolte, sotto forma di M-files, ossia di programmi in linguaggio MATLAB (e quindi in forma opportuna per eventuali adattamenti), nel package SIGNAL PROCESSING TOOLBOX<sup>7</sup>. Lo scopo di questo paragrafo è quello di fornire una breve introduzione all'uso di tale package, con particolare riguardo alle nozioni introdotte nel paragrafo precedente. Per ulteriori informazioni, specie per quanto riguarda l'utilizzo del package per la progettazione di filtri rinviamo alla relativa *User's guide*.

### 11.2.1 Segnali come vettori

Come abbiamo visto, un segnale digitale è rappresentabile mediante una sequenza  $x(n)$ . Un primo modo per introdurre una sequenza in MATLAB consiste nell'elenco esplicito degli elementi. Ad esempio l'istruzione  $\mathbf{x}=[1\ 4\ -5\ 7\ 0]$  crea una sequenza di cinque elementi nel vettore riga  $\mathbf{x}$ . Si passa ad un vettore colonna mediante l'operazione di trasposizione  $\mathbf{x}'$ .

Sottolineiamo il fatto che in MATLAB (nella implementazione corrente) l'indice  $n$  in un vettore deve assumere valori *strettamente positivi*, cioè da 1 a  $N$ , anziché da 0 a  $N-1$ . A causa di tale convenzione, è necessario riscrivere opportunamente le relazioni introdotte nei paragrafi precedenti.

<sup>7</sup>Di interesse collegato sono i packages SYSTEM IDENTIFICATION TOOLBOX e CONTROL SYSTEM TOOLBOX rivolti al trattamento del problema della costruzione e al controllo di modelli matematici di *sistemi dinamici*, costruiti a partire da dati sperimentali. Si tratta di software interattivo, con capacità grafiche e facilità di importazione e esportazione dei dati, e quindi particolarmente interessante per il problema dell'*identificazione* di modelli, che, come noto, è essenzialmente un processo di natura iterativa, con successive valutazioni e aggiustamenti.

Se i dati sono contenuti in un file, chiamato ad esempio `file1.dat`, in formato ASCII, con l'istruzione `load file1.dat` i dati vengono introdotti nello spazio di lavoro di MATLAB nel vettore `file1`.

Un *segnale continuo* può essere campionato nel seguente modo. Dato, ad esempio, il segnale  $\sin(2\pi 20t) + 3\sin(2\pi 30t)$ , per  $t \in [0, 1]$ , le seguenti istruzioni

```
t=0:0.01:1;
y=sin(2*pi*20*t)+3*sin(2*pi*30*t);
```

creano nel vettore `y` il campionamento del segnale con passo 0.01. In Figura 11.12, insieme al vettore `y`, è riportato il vettore `yn` che si ottiene sommando a `y` un *rumore bianco*, cioè un errore casuale con distribuzione normale. Tale vettore è ottenuto mediante le seguenti istruzioni

```
rand('normal')
yn=y+0.6*rand(t);
```

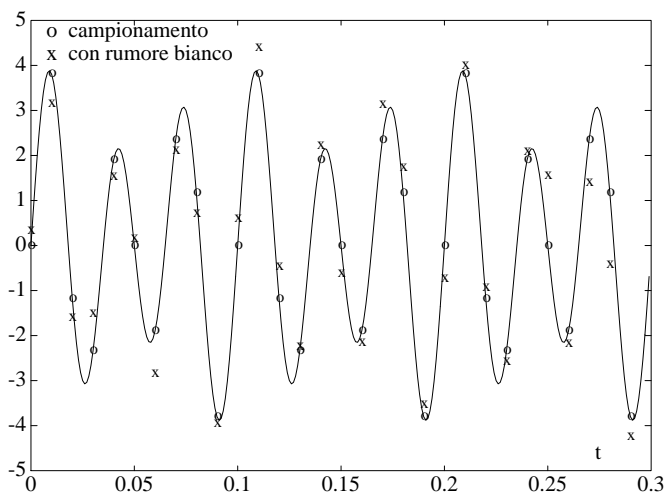


Figura 11.12: Campionamento di un segnale continuo.

Tipi particolari di segnali possono essere creati direttamente. Ad esempio, l'impulso digitale  $u_0(n)$  di dimensione 50 e la sequenza unitaria  $u_{-1}(n)$  sono dati dalle seguenti istruzioni

```
y=[1 zeros(1,49)]; % impulso digitale
y=ones(1,50);      % impulso unitario
```

### 11.2.2 Sistemi lineari discreti

Come si è visto nei paragrafi precedenti, nel dominio del tempo un *filtro digitale* è caratterizzato da un'equazione alle differenze lineare. A causa della convenzione

sugli indici di un vettore, tale equazione è riscritta nella seguente forma

$$y(n) = b_1x(n) + b_2x_{n-1} + \dots + b_{n_b+1}x(n - n_b) \\ - a_2y_{n-1} - \dots - a_{n_a+1}y(n - n_a)$$

L'ordine del filtro è il massimo tra gli interi  $n_a$  e  $n_b$ . La trasformata  $z$  dell'equazione alle differenze fornisce la seguente descrizione nel dominio delle frequenze

$$Y(z) = H(z)X(z), \quad H(z) = \frac{b_1 + b_2z^{-1} + \dots + b_{n_b+1}z^{-n_b}}{1 + a_2z^{-1} + \dots + a_{n_a+1}z^{-n_a}}$$

Ricordiamo che i filtri non ricorsivi FIR (finite impulse response) corrispondono a  $n_a = 0$ .

In MATLAB un filtro può essere rappresentato mediante due vettori riga **a** e **b**. Ad esempio, mediante le seguenti istruzioni

```
b=[3 5 7]; a=[1 2 2 1];
```

si rappresenta il filtro IIR corrispondente alla seguente funzione di trasferimento

$$H(z) = \frac{3 + 5z^{-1} + 7z^{-2}}{1 + 2z^{-1} + 2z^{-2} + z^{-3}}$$

La funzione  $H(z)$  è una funzione razionale in  $z^{-1}$ . Se indichiamo con  $N(z)$  e  $D(z)$ , rispettivamente, il numeratore e il denominatore, le radici di  $N(z)$  e di  $D(z)$  sono, rispettivamente, gli *zeri* e i *poli* di  $H(z)$ . La conoscenza di tali valori è importante per l'analisi del filtro corrispondente a  $H(z)$ .

Ricordiamo che in MATLAB la ricerca degli zeri di un polinomio può essere effettuata mediante l'istruzione **roots**. Nell'esempio precedente le radici del polinomio  $N(z) = 3 + 5z^{-1} + 7z^{-2}$  sono ottenute dalla seguente istruzione

```
z=roots(b)
z=
-0.8333+1.2802 i
-0.8333-1.2802 i
```

### 11.2.3 Analisi dei filtri

La funzione **y=filter(b,a,x)** filtra i dati nel vettore **x** con il filtro descritto dai vettori **b** e **a** per creare i dati filtrati **y**. La funzione **filter** permette di imporre le *condizioni iniziali* **zi** sui termini di ritardo e di ottenere le *condizioni finali* **zf** ottenute dopo un'esecuzione

```
[y,zf]=filter(b,a,x,zi)
```

Per caratterizzare il comportamento nel *dominio della frequenza* di un filtro, la seguente istruzione (che utilizza un algoritmo basato sulla FFT)

```
[h,w]=freqz(b,a,n)
```

fornisce la risposta in frequenza  $H(e^{i\omega})$  del filtro digitale definito nel modo seguente

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_1 + b_2 z^{-1} + \dots + b_{n_b+1} z^{-n_b}}{a_1 + a_2 z^{-1} + \dots + a_{n_a+1} z^{-n_a}}$$

La risposta in frequenza è fornita nel vettore complesso  $\mathbf{h}$ , mentre in  $\mathbf{w}$  sono contenuti gli  $n$  punti di frequenza. Più precisamente,  $\mathbf{w}$  contiene  $n$  punti ugualmente distanziati tra  $0$  e  $\pi$ . Per ottenere il grafico del modulo  $|H(e^{i\omega})|$  e della fase  $\angle H(e^{i\omega})$  in gradi si utilizzano le seguenti istruzioni

```
m=abs(h);
p=angle(h);
semilogy(w,m)
plot(w,p*180/pi)
```

Come esemplificazione, in Figura 11.13 sono riportati i risultati ottenuti in corrispondenza al filtro definito dai vettori  $\mathbf{b}=[2 \ 3 \ 4]$  e  $\mathbf{a}=[1 \ 3 \ 3 \ 1]$ .

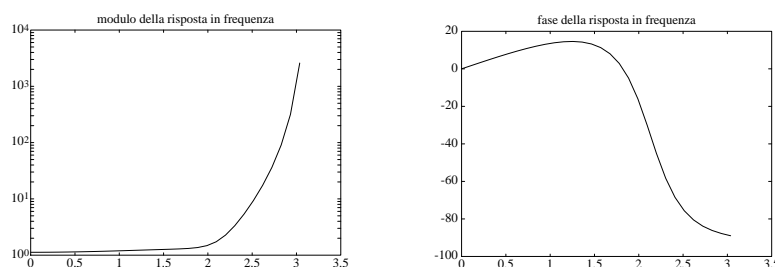


Figura 11.13: Modulo e fase della risposta in frequenza ad un particolare filtro.

### Interpolazione e decimazione

L'operazione di *interpolazione* e la sua inversa di *decimazione* sono utilizzate per convertire un segnale da un campionamento ad un altro. Tali operazioni sono basate sull'utilizzo di particolari filtri.

Come esemplificazione, supponiamo che un segnale  $y$  campionato a 150 Hz debba essere convertito in un segnale  $s$  campionato a 100 Hz. Sia, ad esempio

```
t=0:(1/150):2;
y=square(2*pi*t)+3*sin(10*pi*t)+0.5*rand(t);
```

Si può allora interpolare  $t$  e  $y$  a 300 Hz

```
t1=interp(t,2); y1=interp(y,2);
```

e quindi ridurre il nuovo segnale di un fattore 3

```
t2=decimate(t1,3); s=decimate(y1,3);
```

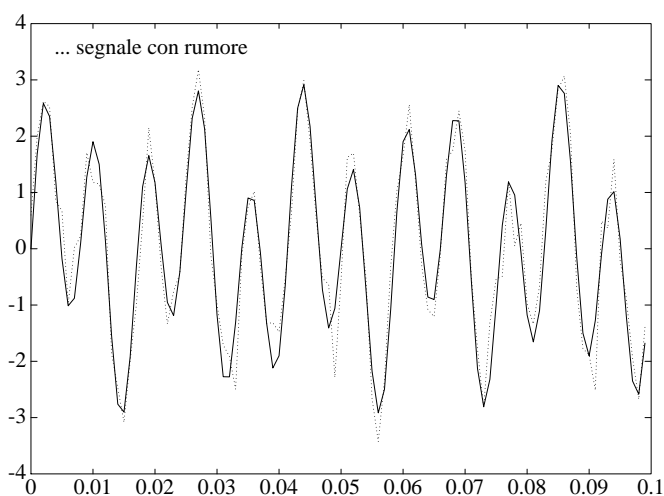


Figura 11.14: Grafico del segnale principale e del segnale con rumore.

#### 11.2.4 FFT e analisi spettrale

La funzione `fft(x)` calcola la trasformata discreta di Fourier del vettore  $x$  mediante un algoritmo Fast Fourier Transform, mentre `ifft(x)` calcola la trasformata di Fourier inversa.

Più precisamente, le due funzioni implementano il calcolo della seguente coppia di trasformazioni

$$X(k+1) = \sum_{n=0}^{N-1} x(n+1)w_N^{kn} \quad x(n+1) = \frac{1}{N} \sum_{k=0}^{N-1} X(k+1)w_N^{-kn}$$

ove  $w_N = e^{-i(2\pi/N)}$  e  $N$  è la lunghezza del vettore  $x$ .

Le funzioni `fft(x,n)` e `ifft(x,n)` calcolano la coppia di trasformate per un vettore di lunghezza  $n$ . Se il vettore dato  $x$  ha una lunghezza inferiore a  $n$ , il vettore



viene completato mediante l'aggiunta di zeri; in caso contrario, la sequenza  $x$  viene troncata.

Come esemplificazione, consideriamo l'utilizzo delle funzioni precedenti per la stima dello spettro di frequenza (*power spectrum*) di un segnale. Consideriamo il segnale  $y_n$  generato nel seguente modo

```
t=0:0.001:1;
y=sin(2*pi*50*t)+2*sin(2*pi*120*t);
rand('normal')
yn=y+0.5*rand(t);
```

ossia il segnale  $y$  è disturbato da un rumore bianco originato da una distribuzione normale. I due segnali  $y$  e  $y_1$  sono rappresentati in Figura 11.14. Calcoliamo, quindi, la trasformata di Fourier di  $y_n$  di lunghezza 1024 ( $= 2^{10}$ )

```
Yn=fft(yn,1024);
```

I quadrati dei moduli del vettore  $Y_n$ , normalizzati alla lunghezza  $n$  del vettore, possono essere calcolati mediante la seguente istruzione

```
Pyy=Yn.*conj(Yn)/n;
```

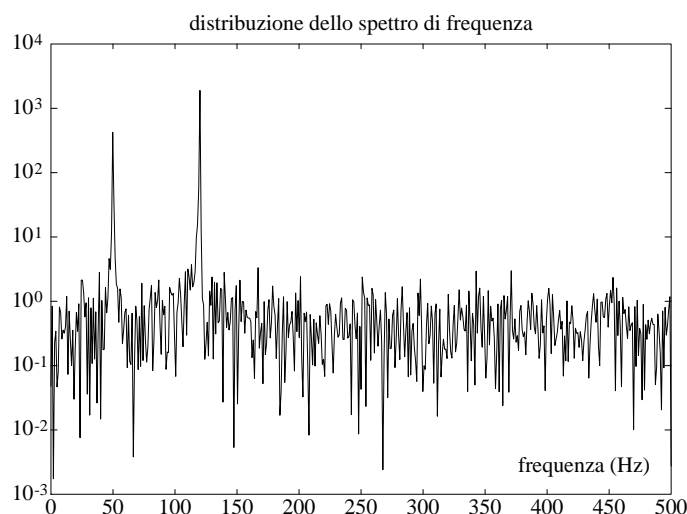


Figura 11.15: Esempio di spettro di frequenza.

Cerchiamo, ora, di rappresentare graficamente il vettore  $P_{yy}$  rispetto alle frequenze. Poiché il segnale è stato campionato ogni  $T = 0.001$  sec., interessano le frequenze nell'intervallo da 0 a 1000 Hz. In realtà, per la proprietà di simmetria della trasformata di Fourier applicata ad un vettore reale, possiamo limitarci a considerare le frequenze da 0 a 500 Hz. Il vettore delle frequenze  $f$  può essere allora costruito nel seguente modo

```
f=500*(0:512)/512
```

Mediante le seguenti istruzioni

```
Pyy(514:1024)=[]; Pyy(2:512)=2*Pyy(2:512);
```

si costruisce un vettore della stessa lunghezza di  $f$  e con le componenti moltiplicate per 2, come compenso per le frequenze trascurate. Infine, l'istruzione

```
semilogy(f,Pyy)
```

produce il grafico riprodotto in Figura 11.15, dal quale si vedono i due picchi corrispondenti alle frequenze 50 Hz e 120 Hz corrispondenti al segnale  $y$ .

La procedura seguita in precedenza per la stima dello spettro di frequenza è stata indicata a solo scopo illustrativo. In realtà, stime migliori possono essere ottenute mediante altre funzioni (`spectrum`, `cceps`) per le quali rinviamo alla guida di utilizzo del package.

### 11.3 Introduzione al filtro di Kalman

In questo paragrafo vengono introdotte le idee essenziali relative alla *teoria del filtro di Kalman*, che rappresenta uno degli strumenti più interessanti nel *filtraggio* e nella regolarizzazione (*smoothing*) dei segnali, nonché nel campo della *previsione* in condizioni *dinamiche*<sup>8</sup>. Dal punto di vista matematico lo studio del filtro di Kalman può essere inquadrato nell'ambito del *metodo dei minimi quadrati*. Più precisamente, come vedremo, da una parte si estende il metodo al caso in cui la stima ottimale di una quantità fissata viene modificata da successive valutazioni della quantità (*metodo dei minimi quadrati ricorsivo*) e dall'altra al caso in cui la quantità stessa da stimare è la soluzione di un sistema dinamico e quindi varia successivamente (*identificazione ricorsiva*). Pertanto, per una migliore comprensione possono essere opportuni alcuni richiami del metodo dei minimi quadrati (cfr. Appendice A e Capitolo 8).

---

<sup>8</sup>La teoria è stata introdotta da Kalman (1960) e Kalman, Bucy (1961) per superare, in sostanza, le difficoltà inerenti alla *teoria di Wiener-Kolmogorov* ([155], 1949), per la quale è cruciale l'ipotesi che i processi che descrivono i segnali e i rumori siano di tipo *stazionario*. Tra le numerose applicazioni *real world* del filtro di Kalman segnaliamo, in particolare, il suo utilizzo nell'*analisi delle immagini*, nei *processi chimici* (stima, predizione e controllo di inquinamento), in *fluidodinamica* (predizione del flusso), nei *sistemi di comunicazione* (demodulazione e equalizzazione dei segnali), nell'*industria aerospaziale* (stima della posizione e della velocità di un oggetto mobile).

### 11.3.1 Metodo dei minimi quadrati con peso

In forma matriciale il metodo dei minimi quadrati può essere visto come un metodo per definire e calcolare la soluzione di un sistema sovradeterminato  $\mathbf{Ax} = \mathbf{b}$ , con  $\mathbf{A}$  matrice di  $m$  righe e  $n$  colonne e  $m \geq n$ . Per il seguito supporremo che la matrice  $\mathbf{A}$  abbia rango  $n$ , ossia che le sue colonne siano dei vettori indipendenti. Il vettore  $\mathbf{b} \in \mathbb{R}^m$  corrisponde ai valori delle *misurazioni* e  $\mathbf{x} \in \mathbb{R}^n$  è il vettore dei *parametri* da stimare. La soluzione del sistema secondo i minimi quadrati è il vettore  $\hat{\mathbf{x}}$  che minimizza la lunghezza euclidea  $\|\mathbf{e}\|^2$  del vettore errore  $\mathbf{e} = \mathbf{b} - \mathbf{Ax}$ . Geometricamente,  $\mathbf{A}\hat{\mathbf{x}}$  è la proiezione di  $\mathbf{b}$  sullo spazio generato dalle colonne di  $\mathbf{A}$ ; inoltre, il vettore  $\hat{\mathbf{x}}$  è la soluzione del seguente sistema lineare, detto *sistema delle equazioni normali*

$$\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b} \Rightarrow \hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \quad (11.43)$$

Supponiamo, ora, che le  $m$  misurazioni abbiano una differente attendibilità. Si può tenere conto di tale informazione introducendo una matrice  $\mathbf{W}$  con la quale pesare opportunamente le varie componenti del vettore residuo. Si considera, cioè,  $\|\mathbf{W}\mathbf{e}\|^2$  come quantità da minimizzare e il sistema delle equazioni normali si modifica nel seguente modo

$$(\mathbf{WA})^T \mathbf{WAx} = (\mathbf{WA})^T \mathbf{Wb} \Rightarrow \mathbf{A}^T \mathbf{W}^T \mathbf{WAx} = \mathbf{A}^T \mathbf{W}^T \mathbf{Wb}$$

Posto  $\mathbf{C} = \mathbf{W}^T \mathbf{W}$ , e supponendo che la matrice  $\mathbf{A}^T \mathbf{CA}$  sia invertibile, la soluzione è formalmente<sup>9</sup> data da  $\hat{\mathbf{x}} = \mathbf{Lb}$ , con  $\mathbf{L} = (\mathbf{A}^T \mathbf{CA})^{-1} \mathbf{A}^T \mathbf{C}$ .

#### Scelta della matrice $\mathbf{C} = \mathbf{W}^T \mathbf{W}$

Il problema della scelta *ottimale* della matrice  $\mathbf{W}$  è risolto sotto opportune condizioni dal *teorema di Gauss-Markov* analizzato nel Capitolo 8. Brevemente, supponendo che gli errori  $\mathbf{e} = \mathbf{b} - \mathbf{Ax}$  provengano da una popolazione di errori, con una distribuzione di probabilità di tipo *normale*, e indicata con  $\mathbf{V}$  la *matrice di covarianza* (che supporremo definita positiva, e quindi invertibile<sup>10</sup>)

$$\mathbf{V} = \mathbf{E}(\mathbf{e}\mathbf{e}^T)$$

la scelta

$$\boxed{\mathbf{C} = \mathbf{W}^T \mathbf{W} = \mathbf{V}^{-1}} \quad (11.44)$$

fornisce uno stimatore

$$\hat{\mathbf{x}} = \mathbf{Lb} = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}^{-1} \mathbf{b} \quad (11.45)$$

<sup>9</sup>Ricordiamo che per quanto riguarda la *stabilità numerica* possono essere più convenienti opportune decomposizioni della matrice  $\mathbf{WA}$  (cfr. Capitolo 2).

<sup>10</sup>In effetti, la matrice di covarianza è non invertibile quando almeno una delle varianze è nulla, ossia quando alcune delle misure  $b_i$  sono esatte; questo significa che le equazioni corrispondenti potrebbero essere risolte esattamente, anziché con il metodo dei minimi quadrati.

*lineare, non distorto e ottimale.* Ricordiamo che non distorto significa che il valore atteso dell'errore  $\mathbf{x} - \hat{\mathbf{x}}$ , tra il valore incognito del parametro  $\mathbf{x}$  e il valore  $\hat{\mathbf{x}}$  stimato a partire dalle osservazioni  $\mathbf{b}$ , è nullo. In effetti, si ha

$$E(\mathbf{x} - \hat{\mathbf{x}}) = E(\mathbf{x} - \mathbf{L}\mathbf{b}) = E(\mathbf{x} - \mathbf{L}\mathbf{A}\mathbf{x} - \mathbf{L}\mathbf{e}) = E((\mathbf{I} - \mathbf{L}\mathbf{A})\mathbf{x}) = 0$$

Infine, ottimale significa che la scelta  $\mathbf{C} = \mathbf{V}^{-1}$  minimizza la varianza dell'errore

$$\mathbf{P} = E[(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^T] = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1}$$

La matrice  $\mathbf{P}$  fornisce gli errori aspettati in  $\hat{\mathbf{x}}$ , mentre la matrice  $\mathbf{V}$  fornisce gli errori aspettati nelle osservazioni  $\mathbf{b}$ . L'inversa della matrice  $\mathbf{P}$ , ossia la matrice  $\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A} = \mathbf{A}^T \mathbf{C} \mathbf{A}$  è un indicatore importante nella teoria del filtraggio ed è chiamata *matrice di informazione* (o matrice di Fisher), in quanto misura il contenuto di informazione dell'esperimento. In maniera schematica, essa aumenta all'aumentare della attendibilità di  $\mathbf{b}$ , ossia al diminuire di  $\mathbf{V}$ .

Il caso in cui gli errori sono indipendenti e distribuiti normalmente con varianza uno corrisponde al caso di *rumori bianchi*; si ha allora  $\mathbf{C} = \mathbf{I}$  e la matrice di informazione è  $\mathbf{A}^T \mathbf{A}$ .

### 11.3.2 Metodo dei minimi quadrati ricorsivo

Supponiamo di aver stimato il vettore  $\mathbf{x}$  mediante il metodo dei minimi quadrati sulla base di un primo insieme di misurazioni. Indicheremo con  $\mathbf{b}_0$  tale insieme di misurazioni, con  $\mathbf{V}_0$  la corrispondente matrice di covarianza e con  $\mathbf{x}_0$  la stima ottenuta. Utilizzando i risultati ottenuti nel paragrafo precedente possiamo esprimere  $\mathbf{x}_0$  nella seguente forma

$$\mathbf{x}_0 = (\mathbf{A}_0^T \mathbf{V}_0^{-1} \mathbf{A}_0)^{-1} \mathbf{A}_0^T \mathbf{V}_0^{-1} \mathbf{b}_0 \quad (11.46)$$

L'errore  $\mathbf{x} - \mathbf{x}_0$  ha media nulla e la sua matrice di covarianza è data da

$$\mathbf{P}_0 = E((\mathbf{x} - \mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^T) = (\mathbf{A}_0^T \mathbf{V}_0^{-1} \mathbf{A}_0)^{-1} \quad (11.47)$$

Nel caso in cui siano disponibili altri dati  $\mathbf{b}_1$ , poniamoci il problema di calcolare una stima ottimale della soluzione  $\mathbf{x}_1$  del sistema combinato  $\mathbf{A}_0 \mathbf{x} = \mathbf{b}_0$ ,  $\mathbf{A}_1 \mathbf{x} = \mathbf{b}_1$  a partire da  $\mathbf{x}_0$  e  $\mathbf{b}_1$  *senza ripartire con il calcolo da  $\mathbf{b}_0$* .

Nell'ipotesi che gli errori  $\mathbf{e}_1$  siano indipendenti dagli errori  $\mathbf{e}_0$ , la matrice di covarianza  $\mathbf{V}$  degli errori totali  $[\mathbf{e}_0, \mathbf{e}_1]^T$  è di tipo diagonale a blocchi

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{V}_1 \end{bmatrix}$$

Pertanto la matrice  $\mathbf{A}^T \mathbf{C} \mathbf{A}$  nell'equazione relativa al vettore  $\mathbf{x}_1$  è

$$\mathbf{P}_1^{-1} = \begin{bmatrix} \mathbf{A}_0 \\ \mathbf{A}_1 \end{bmatrix}^T \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{V}_1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}_0 \\ \mathbf{A}_1 \end{bmatrix} = \mathbf{A}_0^T \mathbf{V}_0^{-1} \mathbf{A}_0 + \mathbf{A}_1^T \mathbf{V}_1^{-1} \mathbf{A}_1 \quad (11.48)$$

dalla quale si ha la seguente importante relazione

$$\mathbf{P}_1^{-1} = \mathbf{P}_0^{-1} + \mathbf{A}_1^T \mathbf{V}_1^{-1} \mathbf{A}_1 \quad (11.49)$$

che fornisce l'incremento nell'informazione fornito dal secondo insieme di misure. Sottolineiamo che la relazione (11.49) non dipende dai valori attuali di  $\mathbf{b}_0$  o  $\mathbf{b}_1$ , ma soltanto dalle loro proprietà statistiche.

Osserviamo, ora, che dall'equazione normale  $\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{V}^{-1} \mathbf{b}$ , si ha

$$\mathbf{x}_1 = \mathbf{P}_1 \begin{bmatrix} \mathbf{A}_0 \\ \mathbf{A}_1 \end{bmatrix}^T \mathbf{V}^{-1} \begin{bmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \end{bmatrix} = \mathbf{P}_1 (\mathbf{A}_0^T \mathbf{V}_0^{-1} \mathbf{b}_0 + \mathbf{A}_1^T \mathbf{V}_1^{-1} \mathbf{b}_1) \quad (11.50)$$

da cui

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{P}_1 (\mathbf{P}_0^{-1} \mathbf{x}_0 + \mathbf{A}_1^T \mathbf{V}_1^{-1} \mathbf{b}_1) \\ &= \mathbf{P}_1 (\mathbf{P}_1^{-1} \mathbf{x}_0 - \mathbf{A}_1^T \mathbf{V}_1^{-1} \mathbf{A}_1 \mathbf{x}_0 + \mathbf{A}_1^T \mathbf{V}_1^{-1} \mathbf{b}_1) \\ &= \mathbf{x}_0 + \mathbf{K}_1 (\mathbf{b}_1 - \mathbf{A}_1 \mathbf{x}_0) \end{aligned} \quad (11.51)$$

La matrice

$$\boxed{\mathbf{K}_1 = \mathbf{P}_1 \mathbf{A}_1^T \mathbf{V}_1^{-1}}$$

viene chiamata *matrice di guadagno* (gain matrix). In questa maniera si è ottenuta una *formula ricorsiva*; essa permette di calcolare  $\mathbf{x}_1$  a partire da  $\mathbf{x}_0$ , anziché da  $\mathbf{b}_0$ .

Naturalmente, se  $\mathbf{b}_1 = \mathbf{A}_1 \mathbf{x}_0$  la stima migliore è ancora  $\mathbf{x}_1 = \mathbf{x}_0$ ; ossia le nuove misure sono esattamente consistenti con il valore originale  $\mathbf{x}_0$ . In caso contrario, l'*errore di predizione*  $\mathbf{b}_1 - \mathbf{A}_1 \mathbf{x}_0$  è amplificato mediante la matrice di guadagno  $\mathbf{K}_1$  per fornire la correzione a  $\mathbf{x}_0$ . Il procedimento può essere iterato a partire da  $\mathbf{P}_1$  e  $\mathbf{x}_1$  per ottenere  $\mathbf{P}_2$  e  $\mathbf{x}_2$ . In maniera generale, a partire dalla coppia  $(\mathbf{P}_{i-1}, \mathbf{x}_{i-1})$  e dai valori osservati  $\mathbf{b}_i$  si ottiene  $(\mathbf{P}_i, \mathbf{x}_i)$  mediante le seguenti formule

$$\begin{aligned} \mathbf{P}_i^{-1} &= \mathbf{P}_{i-1}^{-1} + \mathbf{A}_i^T \mathbf{V}_i^{-1} \mathbf{A}_i \\ \mathbf{x}_i &= \mathbf{x}_{i-1} + \mathbf{K}_i (\mathbf{b}_i - \mathbf{A}_i \mathbf{x}_{i-1}) \quad \text{ove } \mathbf{K}_i = \mathbf{P}_i \mathbf{A}_i^T \mathbf{V}_i^{-1} \end{aligned}$$

► **Esempio 11.4** Come semplice esemplificazione, consideriamo una serie  $b_1, \dots, b_m$  di valutazioni sperimentali di una determinata grandezza incognita  $x$ . Supponiamo inoltre che le successive valutazioni siano indipendenti e distribuite normalmente con la stessa varianza  $\sigma^2$ . Si verifica facilmente che in questo caso si ha

$$\mathbf{A}^T = [ 1 \quad 1 \quad \dots \quad 1 ], \quad \mathbf{P} = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} = \frac{\sigma^2}{m}$$

e la soluzione  $\hat{x}$  è la media dei valori  $b_i$ . Supponendo di aggiungere successivamente una nuova stima, le formule iterative precedenti diventano

$$P_i^{-1} = P_{i-1}^{-1} + \frac{1}{\sigma^2}, \quad x_i = x_{i-1} + K_i (b_i - A_i x_{i-1})$$

ove  $A_i = [1]$ . La prima relazione dice che  $P_i^{-1}$  è uguale a  $i/\sigma^2$ , mentre dalla seconda relazione, poiché  $K_i = P_i A_i^T V_i^{-1} = [1/i]$ , si ha

$$x_i = x_{i-1} + \frac{1}{i}(b_i - x_{i-1}) = \frac{1}{i}b_i + \frac{i-1}{i} \left( \frac{b_1 + b_2 + \dots + b_{i-1}}{i-1} \right) = \frac{b_1 + b_2 + \dots + b_i}{i}$$

Naturalmente, l'idea ha la sua efficacia nel caso più generale in cui  $\mathbf{b}$  e  $\mathbf{x}$  sono vettori e  $\mathbf{V}$  e  $\mathbf{P}$  sono matrici. ■

### 11.3.3 Filtro di Kalman

Nello schema considerato nel paragrafo precedente le successive misurazioni  $\mathbf{b}_i$  si riferiscono alla medesima grandezza  $\mathbf{x}$ . Una situazione più generale si ha quando la grandezza che si vuole stimare cambia successivamente, in maniera che le stime  $\mathbf{b}_i$  si riferiscono, per ogni  $i$  fissato, a grandezze differenti. In altre parole, come nel caso dei segnali digitali studiati nei paragrafi precedenti, si suppone di avere una grandezza  $\mathbf{x}$  che può assumere valori diversi in successivi passi discreti. Per mantenere omogeneità con le notazioni introdotte in questo paragrafo, indichiamo con  $\mathbf{x}_i$  (anziché, come nel paragrafo precedente,  $\mathbf{x}(i)$ ) il valore che la variabile assume al generico passo  $i$ . Supponiamo, inoltre, di *sapere* che il valore  $\mathbf{x}_{i+1}$  è legato al valore precedente  $\mathbf{x}_i$  da una relazione lineare del tipo

$$\mathbf{x}_{i+1} = \mathbf{F}_i \mathbf{x}_i \quad (\text{con errore } \epsilon_i) \quad (11.52)$$

ove  $\mathbf{F}_i$  è una matrice nota e  $\epsilon_i$  è un eventuale termine di *errore*. Ad ogni passo si acquisisce nuova informazione su  $\mathbf{x}_i$  mediante una misurazione  $\mathbf{b}_i$  di una funzione lineare di  $\mathbf{x}_i$ , cioè di  $\mathbf{A}_i \mathbf{x}_i$ , ove  $\mathbf{A}_i$  è una matrice nota. Indicando con  $\mathbf{e}_i$  gli errori di misurazione, si avrà la seguente relazione

$$\mathbf{b}_i = \mathbf{A}_i \mathbf{x}_i \quad (\text{con errore } \mathbf{e}_i) \quad (11.53)$$

Il problema consiste nel dare una *stima ottimale* della successione  $\{\mathbf{x}_i\}$ , cioè di separare il segnale dal rumore. Il modello costituito dalle equazioni (11.52) e (11.53) è rappresentato in maniera schematica in Figura 11.16.

Utilizzando il formalismo introdotto nel paragrafo precedente, si tratta di trovare mediante il metodo dei minimi quadrati (opportunosamente pesati) la soluzione del sistema lineare costituito dalle seguenti equazioni

$$\begin{aligned} \mathbf{A}_i \mathbf{x}_i &= \mathbf{b}_i && \text{misurazione} \\ -\mathbf{F}_i \mathbf{x}_i + \mathbf{x}_{i+1} &= 0 && \text{sistema dinamico} \end{aligned}$$

Il filtro di Kalman corrisponde, in sostanza, alla risoluzione di tale problema in maniera *ricorsiva*.

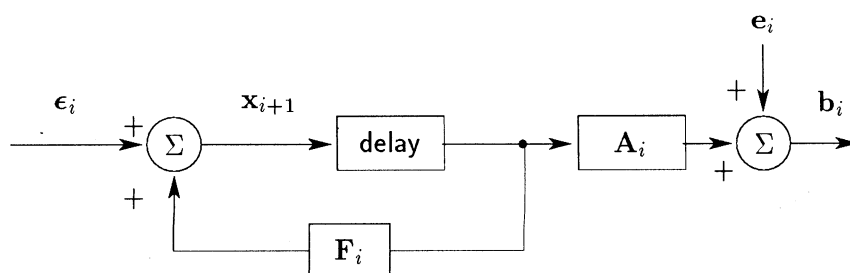


Figura 11.16: Schema del modello.

A scopo esemplificativo, consideriamo il problema per  $i = 0, 1$ . Per la stima dei valori  $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$  si hanno, allora, le seguenti equazioni

$$\begin{aligned} \mathbf{A}_0 \mathbf{x}_0 &= \mathbf{b}_0 \\ -\mathbf{F}_0 \mathbf{x}_0 + \mathbf{x}_1 &= 0 \\ \mathbf{A}_1 \mathbf{x}_1 &= \mathbf{b}_1 \\ -\mathbf{F}_1 \mathbf{x}_1 + \mathbf{x}_2 &= 0 \\ \mathbf{A}_2 \mathbf{x}_2 &= \mathbf{b}_2 \end{aligned}$$

e, in forma di matrici, il sistema

$$\begin{bmatrix} \mathbf{A}_0 & & & & \\ -\mathbf{F}_0 & \mathbf{I} & & & \\ & \mathbf{A}_1 & & & \\ & -\mathbf{F}_1 & \mathbf{I} & & \\ & & & \mathbf{A}_2 & \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_0 \\ 0 \\ \mathbf{b}_1 \\ 0 \\ \mathbf{b}_2 \end{bmatrix} \quad (11.54)$$

Abbiamo, quindi, un sistema sovradeterminato di cinque equazioni nelle tre incognite  $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ . Sottolineiamo il fatto che la misurazione  $\mathbf{b}_2$  influisce non solo sul valore  $\mathbf{x}_2$ , ma su tutta la soluzione. Questa influenza sui valori precedenti è chiamata regolarizzazione (*smoothing*), e la stima di  $\mathbf{x}_2$  filtraggio (*filtering*). Per sottolineare la dipendenza della soluzione dal valore  $\mathbf{b}_2$ , indicheremo con  $\mathbf{x}_{0/2}, \mathbf{x}_{1/2}, \mathbf{x}_{2/2}$  la soluzione di (11.54) ottenuta mediante i minimi quadrati.

▼ **Osservazione 11.3** *Il problema che stiamo considerando è una estensione del problema considerato nel paragrafo precedente. In effetti, quest'ultimo corrisponde al caso in cui il sistema dinamico si riduce all'equazione  $\mathbf{x}_{i+1} = \mathbf{x}_i = \mathbf{x}$  e non sono presenti in tale equazione gli errori  $\epsilon_i$ . In questo caso il sistema (11.54) si riduce ai termini contenenti i coefficienti  $\mathbf{A}_i$ . Osserviamo, tuttavia, che quando si suppone la presenza degli errori  $\epsilon_i$  il sistema rimane il precedente (naturalmente, con  $\mathbf{F}_i = \mathbf{I}$ ) e i risultati sono pertanto, in generale, differenti.* ■

Nell'ipotesi che gli errori sperimentali  $\mathbf{e}_i$  siano indipendenti e distribuiti normalmente con varianza  $\sigma^2$ , e che pure gli errori  $\boldsymbol{\epsilon}_i$ , presenti nell'equazione dinamica, siano indipendenti con varianza differente dalla precedente, e che scriveremo per comodità  $(\sigma/c)^2$ , possiamo ricavare dal teorema di Gauss-Markov l'indicazione per una scelta ottimale dei *pesi* da utilizzare nell'applicazione del metodo dei minimi quadrati. In effetti, basta dividere per  $\sigma$  le equazioni che interessano  $\mathbf{A}$  e  $\mathbf{b}$  e per  $\sigma/c$  le rimanenti. Eliminando il fattore  $\sigma$ , si ha che il sistema (11.54) si trasforma nel seguente

$$\begin{bmatrix} \mathbf{A}_0 & & & \\ -c\mathbf{F}_0 & c\mathbf{I} & & \\ & \mathbf{A}_1 & & \\ & -c\mathbf{F}_1 & c\mathbf{I} & \\ & & \mathbf{A}_2 & \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_0 \\ 0 \\ \mathbf{b}_1 \\ 0 \\ \mathbf{b}_2 \end{bmatrix} \quad (11.55)$$

Questo è, in definitiva, il sistema  $\mathbf{A}\mathbf{x} = \mathbf{b}$  da risolvere mediante il metodo dei minimi quadrati. Il metodo diretto consisterebbe nella risoluzione delle equazioni normali  $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ , ma l'interesse è quello di risolvere il sistema in forma ricorsiva, in maniera cioè da utilizzare il più possibile quanto si è calcolato nei passi precedenti.

Come illustrazione, consideriamo il passaggio da  $i = 0$  a  $i = 1$ . Per  $i = 0$  la matrice  $\mathbf{A}$  si riduce a

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_0 & \\ -c\mathbf{F}_0 & c\mathbf{I} \\ & \mathbf{A}_1 \end{bmatrix}$$

Nel passaggio a  $i = 1$  il nuovo blocco  $-c\mathbf{F}_1$  influisce sull'elemento di indici (2,2) della matrice  $\mathbf{A}^T\mathbf{A}$ , mentre sono completamente nuove la terza riga e la terza colonna. Questa è, in sostanza, la regola generale: nel passaggio da  $i$  a  $i + 1$  la matrice tridiagonale a blocchi  $\mathbf{A}^T\mathbf{A}$  cambia il blocco inferiore destro e aggiunge una nuova colonna e una nuova riga. L'aspetto fondamentale nella costruzione dell'algoritmo diventa allora quello di calcolare in maniera efficiente e *numericamente stabile* gli effetti di tali cambiamenti sulla stima  $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$  e sulla matrice di covarianza<sup>11</sup>  $\mathbf{P} = (\mathbf{A}^T\mathbf{A})^{-1}$ .

Rinviamo alla bibliografia per maggiori dettagli (si veda in particolare Ljung e Söderström [108]), ci limiteremo ad analizzare alcuni aspetti numerici. Come abbiamo visto nella risoluzione numerica del problema standard dei minimi quadrati (cfr. Capitolo 2), dal punto di vista numerico (cioè della stabilità dell'algoritmo) non è conveniente, in generale, operare sulla matrice  $\mathbf{A}^T\mathbf{A}$ , per la quale il numero di condizionamento è il quadrato di quello della matrice  $\mathbf{A}$ . In effetti, esistono algoritmi nei quali non si costruisce esplicitamente la matrice  $\mathbf{A}^T\mathbf{A}$ , ma si utilizzano opportune decomposizioni della matrice  $\mathbf{A}$ . In questo senso, particolarmente interessanti

<sup>11</sup>Osserviamo che tale matrice non dipende dai dati  $\mathbf{b}$  e potrebbe essere calcolata a priori, ossia prima di effettuare gli esperimenti.



risultano la *decomposizione in valori singolari (SVD)* e la *decomposizione ortogonale*  $\mathbf{A} = \mathbf{QR}$ , ottenuta ortogonalizzando le colonne di  $\mathbf{A}$ . Esaminando ad esempio quest'ultima decomposizione, essa trasforma il sistema  $\mathbf{Ax} = \mathbf{b}$  nella seguente forma triangolare superiore

$$\begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \quad \text{ove } \mathbf{R} = \begin{bmatrix} \mathbf{R}_0 & \mathbf{R}_{0,1} & 0 \\ 0 & \mathbf{R}_1 & \mathbf{R}_{1,2} \\ 0 & 0 & \mathbf{R}_2 \end{bmatrix} \quad (11.56)$$

ossia riduce il problema alla risoluzione del sistema triangolare  $\mathbf{Rx} = \mathbf{c}$ . La matrice di covarianza  $\mathbf{P} = (\mathbf{A}^T \mathbf{A})^{-1}$  diventa  $(\mathbf{R}^T \mathbf{R})^{-1}$ . Tutta l'informazione è pertanto contenuta nella matrice  $\mathbf{R}$ . Tale matrice può essere calcolata in maniera ricorsiva, dal momento che il solo cambiamento rispetto al passo precedente è in  $\mathbf{R}_1$ , a parte naturalmente i blocchi diversi dallo zero nell'ultima colonna.

## 11.4 Introduzione alle funzioni wavelet

Nei paragrafi precedenti abbiamo visto che un segnale unidimensionale  $f(t)$ , con  $t \in \mathbb{R}$ , può essere rappresentato nei seguenti due modi

- (i) rappresentazione *fisica*, ossia tramite i valori numerici di  $f(t)$ ;
- (ii) rappresentazione *spettrale*, ossia tramite i valori della trasformata di Fourier  $\hat{f}(\omega)$  di  $f(t)$

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt$$

Sotto opportune ipotesi di regolarità è inoltre possibile *ricostruire*  $f(t)$  dalla sua rappresentazione spettrale

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) e^{i\omega t} dt$$

Ognuna delle due precedenti rappresentazioni è adatta per un certo tipo di segnale; ad esempio, la rappresentazione fisica per i segnali impulsivi, mentre quella spettrale per i segnali periodici. In ogni caso, esse possono non essere convenienti per la rappresentazione di quei segnali nei quali ampiezza e frequenza dipendono dalla variabile  $t$  (ad esempio durante un transiente). In altre parole, la trasformata di Fourier contiene una informazione globale sulle frequenze del segnale su tutti gli istanti, anziché mostrare come le frequenze variano con il tempo.

Per l'analisi di frequenze dipendenti dal tempo sono stati proposti differenti metodi. Una prima idea seguita da Gabor<sup>12</sup> consiste nel *localizzare* il segnale  $f(t)$  utilizzando una funzione *finestra*  $g(t)$ . Si ottiene in questo modo la seguente trasformata,

<sup>12</sup>D. Gabor, *Theory of communication*, J. Inst. Elec. Eng. (Londra) **93** III (1946).

chiamata *windowed Fourier transform*, o anche *short time Fourier transform*

$$\Phi_{\text{WFT}}f(p, q) = \int_{-\infty}^{+\infty} f(t) e^{-ipt} g(t - q) dt \quad (11.57)$$

per  $p, q \in \mathbb{R}$ . Nel caso in cui la funzione  $g(t)$  sia *localizzata* nell'intorno, ad esempio, del punto  $t = 0$ , allora  $\Phi_{\text{WFT}}f(\cdot, q)$  descrive il contenuto in frequenza della funzione  $f(t)$  nell'intorno di  $t = q$ . Se la trasformata di Fourier  $\widehat{g}(\omega)$  è poi *localizzata* nell'intorno, ad esempio, di  $\omega = 0$ , allora  $\Phi_{\text{WFT}}f(p, \cdot)$  descrive il contenuto in  $t$  di  $\widehat{f}(\omega)$  nell'intorno di  $\omega = p$ . Pertanto, con una scelta conveniente della funzione  $g(t)$  i valori  $\Phi_{\text{WFT}}f(p, q)$  descrivono la funzione  $f(t)$  nell'intorno del punto  $(p, q)$  dello spazio delle fasi tempo-frequenza (ricordiamo che le funzioni  $g$  considerate da Gabor sono delle gaussiane, ossia della forma  $g(t) = e^{-rt^2}$ ). I parametri  $p, q$  possono variare con continuità su tutta la retta reale, oppure in maniera discreta  $p = mp_0$ ,  $q = nq_0$ , con  $n, m \in \mathbb{Z}$  e  $p_0, q_0 > 0$  fissati. Nel caso discreto si ha una successione di coefficienti  $\{c_{mn}(f)\}$ , per  $m, n \in \mathbb{Z}$

$$c_{mn} = \int_{-\infty}^{+\infty} g_{mn}(t) f(t) dt, \quad g_{mn}(t) = e^{-imq_0 t} g(t - np_0)$$

ove i parametri  $p_0$  e  $q_0$  sono da scegliere in maniera che la ricostruzione di  $f$  sia possibile a partire dai coefficienti campionati  $c_{mn}(f)$ . Come esemplificazione, in Figura 11.17 è mostrata la funzione finestra  $g(t) = \pi^{-1/4} e^{-t^2/2}$  e l'elemento  $g_{23}$  corrispondente a  $p_0 = 1$  e  $q_0 = 2\pi$ . Come per la rappresentazione spettrale, il segnale

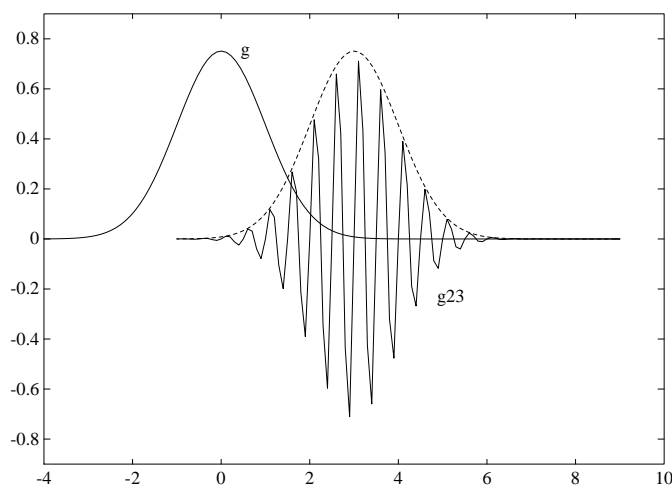


Figura 11.17: Funzione “finestra” di Gabor  $g(t) = \pi^{-1/4} e^{-t^2/2}$  e elemento  $g_{23}$  corrispondente a  $p_0 = 1$  e  $q_0 = 2\pi$ .

$f(t)$  è completamente caratterizzato dai valori di  $\Phi_{\text{WFT}}(p, q)$  e può essere ricostruito, in opportune ipotesi di regolarità per la funzione  $g$  e per  $\Phi_{\text{WFT}}$ , attraverso la

seguinte formula

$$f(t) = \frac{1}{2\pi\|g\|^2} \int_{-\infty}^{+\infty} \Phi_{\text{WFT}}(p, q) e^{ipt} g(t - q) dpdq$$

dove  $\|g\|$  indica (qui e nel seguito) la norma nello spazio  $L^2(\mathbb{R})$  (cfr. per la definizione di tale spazio Capitolo 4) della funzione  $g$ .

Lo studio del segnale per scale piccole richiede un procedimento in grado di adattare la grandezza della finestra alla grandezza di ciò che si vuole analizzare. Sotto questo aspetto, è da sottolineare che la trasformata di Gabor non è in grado di analizzare dettagli più piccoli della *finestra* utilizzata. Un'idea per superare la difficoltà ora rilevata è stata proposta agli inizi degli anni 80 da parte di Grossmann e Morlet<sup>13</sup> con l'introduzione dell'analisi tramite wavelet (analisi inizialmente applicata a segnali provenienti dalla ricerca geologica e petrolifera). La trasformata wavelet (*wavelet transform*) di una funzione  $f(t)$ ,  $t \in \mathbb{R}$  viene definita nel modo seguente

$$\Phi_{\text{WAT}}f(a, b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{+\infty} f(t)h\left(\frac{t-b}{a}\right) dt \quad (11.58)$$

ove  $b$  un parametro reale di *traslazione*,  $a$  un parametro reale di *scala* e  $h(t)$  è la funzione wavelet analizzante (*analyzing wavelet*), che si suppone verificare la seguente *condizione di ammissibilità*

$$C_h = \int_{-\infty}^{+\infty} |\widehat{h}(\omega)|^2 |\omega|^{-1} d\omega < +\infty$$

La trasformata wavelet corrisponde quindi al prodotto scalare di  $f$  con una versione *contratta* e *traslata* di  $h(t)$ . Anche in questo caso è possibile ricostruire il segnale  $f(t)$  a partire dalla sua trasformata wavelet. In opportune ipotesi di regolarità si ha infatti la seguente formula

$$f(t) = \frac{1}{2\pi C_h} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Phi_{\text{WAT}}f(a, b) |a|^{-1/2} h\left(\frac{t-b}{a}\right) \frac{da db}{a^2} \quad (11.59)$$

Nel caso discreto, data la funzione  $f(t)$  si generano i coefficienti

$$d_{mn}(f) = \int_{-\infty}^{+\infty} h_{mn}(t)f(t) dt, \quad f_{mn}(t) = a_0^{-m/2} h(a_0^{-m} t - nb_0)$$

ove il parametro di contrazione  $a_0$  e il parametro di traslazione  $b_0$  sono da scegliere convenientemente. Come esemplificazione, in Figura 11.18 è rappresentata la funzione  $h(t)$  originariamente proposta da Grossmann e Morlet

$$h(t) = \frac{2}{\sqrt{3}\pi^{1/4}} (1 - t^2) e^{-t^2/2}$$

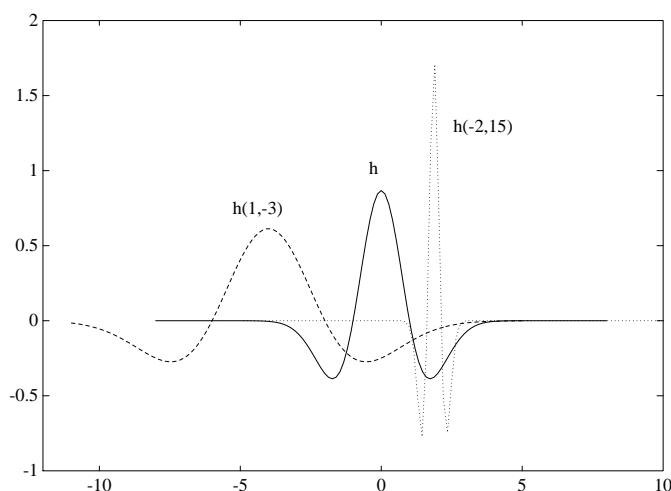


Figura 11.18: Wavelet analizzatrice di Morlet e Grossmann per valori differenti dei parametri.

e le funzioni contratte e traslate  $h_{1,-3}(t)$  e  $h_{-2,15}(t)$ , per  $a_0 = 2$  e  $b_0 = 0.5$ . La figura mette in rilievo un aspetto importante della trasformata wavelet nei confronti della trasformata di Gabor. Mentre il supporto della funzione finestra modulata rimane immutato, quello della trasformata wavelet può essere opportunamente ridotto per le frequenze alte. In sostanza, mediante la trasformata wavelet è possibile entrare nei dettagli su fenomeni di breve durata e alta frequenza, quali i segnali caratterizzati da un rapido transiente. Tale possibilità ha avuto applicazioni nella teoria degli operatori singolari (cfr. ad esempio C. Fefferman, R. Llave, *Relativistic stability of matter*, Rev. Mat. Iberoamericana **2** (1986)). Ricordiamo, inoltre, che sia la trasformata wavelet che la trasformata di Fourier con *finestra* sono esempi della *decomposizione in stati coerenti* usata in fisica quantistica (cfr. J. R. Klauder, B. S. Skagerstam, *Coherent States*, World Scientific (Singapore) (1985)). Per tali motivi, nuovi sviluppi per lo studio delle wavelet si sono avuti in applicazioni alla teoria quantistica (cfr. G. Battle, P. Federbush, *Ondelettes and phase cell cluster expansions: a vindication*, Comm. Math. Phys. **109** (1987)).

Nel caso discreto un problema importante per le applicazioni riguarda la possibilità di ricostruire la funzione  $f$  dai valori dei coefficienti  $d_{mn}$ . È chiaro che condizione necessaria è che la trasformazione  $f \rightarrow \{d_{mn}(f)\}$  sia biunivoca. Dal punto di vista pratico, è tuttavia anche necessario che la procedura di ricostruzione sia *stabile*, nel senso che la “vicinanza” dei valori  $d_{mn}(f)$  ai valori  $d_{mn}(\tilde{f})$  deve implicare la

<sup>13</sup>A. Grossman, J. Morlet, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal. **15** (1984).

“vicinanza” di  $f$  a  $\tilde{f}$ . Più concretamente, si richiede che

$$A \|f\|^2 \leq \sum_{m,n} |d_{m,n}(f)|^2 \leq B \|f\|^2 \quad (11.60)$$

con  $0 < A < B < +\infty$ , e  $A, B$  indipendenti da  $f$ . Una famiglia di funzioni  $h_{mn}(t)$ ,  $m, n \in \mathbb{Z}$ , appartenenti allo spazio  $L^2(\mathbb{R})$  e che verifica la proprietà (11.60) per ogni  $f \in L^2(\mathbb{R})$  viene detta un *frame* e le costanti  $A$  e  $B$  sono chiamate *frame bounds* e giocano un ruolo importante nelle applicazioni pratiche. Osserviamo che un frame non è necessariamente una base per  $L^2(\mathbb{R})$ ; in effetti, esso può essere ridondante, e quindi gli elementi del frame possono essere non linearmente indipendenti.

► **Esempio 11.5** Si consideri il caso in dimensione finita con  $v_1 = (1, 0)$ ,  $v_2 = (1/2, 3)$ ,  $v_3 = (1/2, -3)$ . L'insieme  $\{v_i, 1 \leq i \leq 3\}$ , è un frame di  $\mathbb{R}^2$ , infatti

$$\frac{3}{2} \|v\|^2 \leq \sum_{j=1}^3 |(v_j, v)|^2 \leq 18 \|v\|^2$$

dove  $(\cdot, \cdot)$ ,  $\|\cdot\|$  sono il prodotto scalare di  $\mathbb{R}^2$  e la corrispondente norma indotta. L'insieme considerato non è una base, dal momento che i vettori non sono linearmente indipendenti. ■

Quando  $A = B$ , nel qual caso il frame è chiamato *tight*, vale la seguente formula di ricostruzione

$$f = A^{-1} \sum_{m,n} d_{mn}(f) h_{mn}(t) \quad (11.61)$$

Nel caso più generale si ha la formula

$$f = 2(A+B)^{-1} \sum_{m,n} d_{mn}(f) h_{mn}(t) + R(f)$$

ove  $R(f)$  rappresenta il termine di errore, per il quale si può dimostrare la seguente stima

$$\int_{-\infty}^{+\infty} |R(f)(t)|^2 dt \leq \frac{BA^{-1} - 1}{BA^{-1} + 1} \int_{-\infty}^{+\infty} |f(t)|^2 dt$$

Dal momento che la costante nel termine di maggiorazione è minore di 1, è possibile iterare il procedimento fino ad ottenere nella ricostruzione la precisione desiderata. Il concetto di frame può essere utilizzato anche per il caso della trasformata windowed di Fourier con *funzione finestra*  $g$ . Per tale funzione esistono valori dei parametri  $p_0, q_0$ , con  $p_0 q_0 \leq 2\pi$ , per cui l'insieme  $\{g_{mn}, m, n \in \mathbb{Z}\}$  forma un frame di  $L^2(\mathbb{R})$ . Tuttavia, per il caso particolare  $p_0 q_0 = 2\pi$ , che corrisponde al campionamento della trasformata con passo di Nyquist, è stato dimostrato (*Teorema di Balian-Low*<sup>14</sup>)

<sup>14</sup>R. Balian, *Un principe d'incertitude fort en théorie du signal ou en mécanique quantique*, C.R. Sc. Paris **292**, série II (1981).

che non esiste una base ortonormale per la trasformata windowed Fourier con buone proprietà di localizzazione in tempo-frequenza. Al contrario, per la trasformata wavelet è possibile trovare una funzione  $h(t)$  localizzata nello spazio delle fasi tempo-frequenza, in maniera che l'insieme  $\{h_{mn}, m, n \in \mathbb{Z}\}$  costituisca un frame senza valori critici per i parametri  $a_0, b_0$ . In particolare, è possibile dimostrare l'esistenza di basi ortonormali di wavelet con  $h$  regolare e localizzata in tempo e frequenza. Il primo esempio di tale base è stato fornito nel caso unidimensionale  $L^2(\mathbb{R})$  da Y. Meyer agli inizi degli anni '80 ed ha la seguente forma

$$\{\psi_{jk} = 2^{j/2} \psi(2^j t - k), j, k \in \mathbb{Z}\} \quad (11.62)$$

dove  $\psi(t)$  è una opportuna funzione tale che  $\psi(t)$  e la sua trasformata di Fourier hanno un decadimento esponenziale all'infinito. In seguito Meyer, in collaborazione con S. Mallat, ha sviluppato il concetto di *Analisi Multirisoluzione*, grazie al quale è possibile derivare opportuni *algoritmi* per la generazione di basi ortonormali di wavelet.

**Definizione 11.1** *Un'analisi multirisoluzione di  $L^2(\mathbb{R})$  è una collezione  $V_j, j \in \mathbb{Z}$  di sottospazi di  $L^2(\mathbb{R})$*

$$\dots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \dots \quad (11.63)$$

tale che

$$\bigcap_{j \in \mathbb{Z}} V_j = \{0\}, \quad \overline{\bigcup_{j \in \mathbb{Z}} V_j} = L^2(\mathbb{R}) \quad , \quad (11.64)$$

$$\begin{aligned} \text{a)} \quad & f(t) \in V_j \iff f(2t) \in V_{j+1} \\ \text{b)} \quad & f(t) \in V_0 \iff f(t-k) \in V_0, \forall k \in \mathbb{Z} \end{aligned} \quad (11.65)$$

Inoltre, esiste  $\theta \in V_0$  tale che l'insieme

$$\{\theta(t-k), k \in \mathbb{Z}\} \quad (11.66)$$

è una base incondizionale di  $V_0$ .

Ricordiamo che in uno spazio lineare normato  $B$  una sequenza  $\{e_i\}_{i \in \mathbb{N}}$  è detta una *base incondizionale* per  $B$  se

i) ogni elemento  $x \in B$  ha un'unica rappresentazione nella forma

$$x = \sum_{n=1}^{+\infty} a_n e_n$$

dove  $a_n, n \in \mathbb{N}$  sono delle costanti;

ii) la convergenza della serie

$$\sum_{n=1}^{+\infty} a_n e_n$$

implica la convergenza della serie

$$\sum_{n=1}^{+\infty} a_{\pi(n)} e_{\pi(n)}$$

per ogni permutazione  $\pi$  dell'insieme  $\mathbb{N}$ .

Osserviamo, inoltre, che per una funzione  $f \in L^2(\mathbb{R})$  la sua proiezione sui  $V_j$  consecutivi rappresenta un'approssimazione della  $f$  sempre più *fine*:  $V_j$  rappresenta l'approssimazione di  $f$  alla scala  $2^{-j}$ ,  $j \in \mathbb{Z}$ .

► **Esempio 11.6** Definiamo

$$V_0 = \{f \in L^2(\mathbb{R}) : f \text{ costante su } [k, k+1[, k \in \mathbb{Z}\}$$

Gli spazi  $V_j$  sono formati dalle funzioni costanti su ogni sottointervallo

$$[k2^j, (k+1)2^j[$$

La funzione *generatrice*  $\theta(t)$  è allora la funzione caratteristica

$$\theta(t) = \begin{cases} 1 & t \in [0, 1[ \\ 0 & \text{altrimenti.} \end{cases}$$

■

La base  $\{\theta(t-k)\}_{k \in \mathbb{Z}}$  non è necessariamente una base ortonormale per lo spazio  $L^2(\mathbb{R})$ ; tuttavia, da essa è possibile ricavare una base ortonormale mediante il procedimento di ortogonalizzazione di Gram-Schmidt, effettuato utilizzando la trasformata di Fourier della base incondizionale che abbiamo a disposizione. In particolare, è possibile dimostrare<sup>15</sup> che, posto

$$\Omega(\rho) = \sum_{k \in \mathbb{Z}} |\widehat{\theta}(\rho = 2\pi k)|^2, \quad \widehat{\phi}(\rho) = \frac{\widehat{\theta}(\rho)}{\sqrt{\Omega(\rho)}}$$

si ottiene la successione  $\{\phi(t-k)\}_{k \in \mathbb{Z}}$  che forma una base ortonormale per  $L^2(\mathbb{R})$ . Inoltre, l'insieme  $\{\phi_{jk}(t)\}_{j,k \in \mathbb{Z}}$  definito da

$$\phi_{jk}(t) = 2^{j/2} \phi(2^j t - k) \tag{11.67}$$

<sup>15</sup>Si veda ad esempio S. Mallat, *Multiresolution approximation and wavelet orthonormal bases of  $L^2(\mathbb{R})$* , Trans. of the AMS, vol. 315, (1989).

permette di definire la seguente base per lo spazio  $V_j$

$$\{\phi_{jk}, k \in \mathbb{Z}\}$$

La funzione  $\phi$  è chiamata *scaling function* dell'analisi multirisoluzione; se  $f \in L^2(\mathbb{R})$  allora

$$\sum_{k \in \mathbb{Z}} \int_{-\infty}^{+\infty} f(t) \phi_{jk}(t) dt \phi_{jk}$$

rappresenta un'approssimazione della  $f(t)$  alla scala  $2^{-j}$ .

► **Esempio 11.7** (Shannon, 1949) Si definisce

$$V_0 = \{f \in L^2(\mathbb{R}) : \text{supp}(\hat{f}) \subset [-\pi, \pi]\}$$

da cui

$$V_j = \{f \in L^2(\mathbb{R}) : \text{supp}(\hat{f}) \subset [-\pi 2^j, \pi 2^j]\}$$

La base ortonormale di  $V_0$  si ricava, tramite trasformata inversa di Fourier, da

$$\phi(t) = \frac{\sin(\pi t)}{\pi t}$$

■

► **Esempio 11.8** (*Funzioni spline*) Per la costruzione di opportune basi di wavelet è possibile utilizzare le funzioni spline. Come illustrazione, consideriamo il caso delle spline lineari, ossia

$$V_0 = \{f \in C^0(\mathbb{R}), f \text{ lineare in } [k, k+1]\}$$

per il quale la funzione  $\theta(t)$  è data da

$$\theta(t) = \begin{cases} t & 0 \leq t \leq 1 \\ 2-t & 1 \leq t \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

È possibile allora costruire la corrispondente base ortonormale calcolando

$$\hat{\theta}(\rho) = e^{-i\rho} \left( \frac{\sin(\rho/2)}{\rho/2} \right)^2$$

$$\Omega(\rho) = \sum_{k \in \mathbb{Z}} |\hat{\theta}(\rho - 2\pi k)|^2 = \sin^4(\rho/2) \sum_k \frac{1}{(\rho/2 + k\pi)^4}$$

■

Per un approfondimento delle nozioni precedentemente introdotte, e, più in generale, per una introduzione più adeguata alla teoria e alle applicazioni delle trasformate wavelet rinviamo in particolare a Daubechies [43]. Segnaliamo, inoltre, la serie di volumi dedicata all'argomento: *Wavelet Analysis and its Applications*, C. K. Chui, Series Editor, Academic Press, Inc.



Models are to be used,  
but non to be believed.  
**Henri Theil**

## Capitolo 12

# Introduzione alla teoria dei compartimenti

L'obiettivo di questo capitolo è quello di introdurre e analizzare alcuni aspetti matematici e numerici di base nella *teoria dei sistemi compartimentali*. La teoria compartimentale è, in sostanza, una tecnica di modellizzazione matematica che si basa sull'analisi di un sistema strutturalmente complicato mediante la sua separazione in un numero finito di parti componenti, chiamate *compartimenti*<sup>1</sup> o stati, che interagiscono tra loro mediante scambio di materiale. Tale tipo di modellizzazione ha interesse in diverse applicazioni, in particolare nello studio e nel controllo dei *sistemi metabolici* negli organismi viventi, nell'analisi della cinetica delle *reazioni chimiche*, ad esempio della *cinetica dei farmaci*, e nell'analisi dei sistemi ecologici, o di altra natura.

Le problematiche relative all'applicazione dell'analisi compartimentale sono quelle comuni, più in generale, ad una *rappresentazione matematica* di un particolare fenomeno. Come semplificazione del fenomeno, essa richiede, innanzitutto, una *conoscenza adeguata* nel campo dal quale il fenomeno proviene, se si vuole che il modello matematico sia significativo e utile. Un primo passo nella costruzione del modello riguarda, allora, la sua *specificazione*; nella teoria dei compartimenti questo significa la determinazione del numero dei compartimenti e delle interconnessioni tra i compartimenti che permettono lo scambio del materiale. Un successivo problema riguarda la *identificabilità strutturale* del modello, ossia la verifica se i parametri contenuti nel modello siano definiti univocamente, a partire da dati sperimentali, quando tali dati sono supposti privi di errori. Dal punto di vista matematico, il

---

<sup>1</sup>“...quantities of a substance having uniform and distinguishable kinetics of transformation or transport” (Atkins, 1969), “...each of which is homogeneous and well mixed, and interacts by exchanging material.” (Jacquez, 1972).

problema della identificabilità è equivalente alla verifica dell'esistenza e dell'unicità delle soluzioni di equazioni non lineari. Un terzo problema, infine, di natura essenzialmente numerica, riguarda il calcolo dei valori ottimali dei parametri, ossia dei valori che minimizzano un determinato stimatore (*stima dei parametri*).

Per un adeguato approfondimento della teoria dei compartimenti, le cui prime applicazioni, come strumento di ricerca, sono abbastanza recenti (Hevesy, 1948), rinviamo ad esempio ad Anderson [5] e Jacquez [92].

## 12.1 Elementi introduttivi

In questo paragrafo verrà precisata la terminologia di base e verranno analizzati alcuni esempi elementari di modelli a compartimenti.

Per *compartimento* si intende una quantità di materiale che si comporta dal punto di vista cinetico (ossia per quanto riguarda il *turnover*, o ricambio) in maniera *omogenea*, come un composto perfettamente mescolato (*well mixed*).

Un concetto più ampio è quello di *spazio* (*space*), o volume. Il contenuto di uno spazio non è necessariamente omogeneo e ben mescolato. Un esempio di volume in biologia è fornito dallo spazio extracellulare in un organo; la quantità di un *determinato* composto (ad esempio il plasma sanguigno) in tale volume può costituire un particolare compartimento. Si vede, quindi, che un compartimento non è necessariamente identificabile con un reale volume fisico. La misura della quantità di sostanza in un compartimento (come volume o massa) è detta *size* (ma anche, talvolta, volume).

Un *sistema compartimentale* consiste di due o più compartimenti *interconnessi*. La interconnessione significa che allo stato stazionario vi è uno flusso di materiale tra determinati compartimenti. Tale scambio può verificarsi in maniera differente, a seconda del tipo di compartimenti: per passaggio attraverso alcune barriere fisiche o per trasformazioni fisiche o chimiche.

Nella usuale rappresentazione di un sistema a compartimenti, una scatola (*box*) denota un compartimento e una freccia indica il trasferimento del materiale all'interno o all'esterno del compartimento (cfr. Figura 12.1). Tale rappresentazione è, per quanto osservato in precedenza, puramente schematica. Un sistema di compartimenti è detto *chiuso*, quando non vi è scambio di materiale con l'ambiente esterno; in caso contrario, è detto un sistema *aperto*<sup>2</sup>. I sistemi biologici sono, per la maggior parte, sistemi aperti.

► **Esempio 12.1** (*Cinetica di un farmaco*) Il processo di ingestione per via orale di un farmaco e il suo successivo metabolismo può essere studiato mediante il sistema a due compartimenti rappresentato in maniera schematica in Figura 12.2. Il farmaco, introdotto con

<sup>2</sup>Più in particolare, un determinato compartimento è detto *leaky* se il materiale può lasciare il compartimento verso l'ambiente esterno; in caso contrario è detto *leakproof*. Un sistema di compartimenti è quindi chiuso se ogni compartimento del sistema è leakproof.

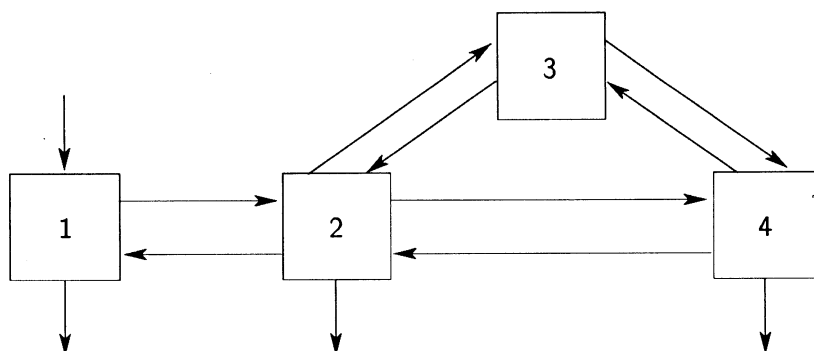


Figura 12.1: Rappresentazione schematica di un sistema di compartimenti.

una intensità (quantità per unità di tempo)  $I(t)$ , entra nel tratto gastrointestinale, rappresentato dal compartimento 1, viene assorbito nella circolazione e distribuito attraverso il corpo per essere metabolizzato e infine eliminato. Indichiamo con  $y_1(t)$  la massa (o la concentrazione) del farmaco, al generico tempo  $t$ , nel compartimento 1 e con  $y_2(t)$  la massa del farmaco nel compartimento 2. Tali quantità rappresentano le *variabili di stato* del sistema.

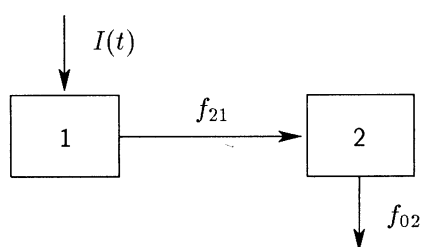


Figura 12.2: Sistema a due compartimenti corrispondente alla cinetica di un farmaco. Il compartimento 1 rappresenta il tratto intestinale (GI) e il compartimento 2 la circolazione sanguigna.

Un modello matematico relativo allo schema a compartimenti ora introdotto può essere ottenuto mediante il seguente principio di conservazione (*equazione di bilancio di massa*)

la velocità totale di cambiamento della sostanza in un compartimento è *uguale* alla velocità con cui la sostanza entra nel compartimento *meno* la velocità con cui la sostanza esce dal compartimento.

In altre parole, il principio implica che nessun compartimento è una sorgente (*source*) o un pozzo (*sink*). In termini di equazioni differenziali<sup>3</sup>, il principio, applicato al compartimento

<sup>3</sup>Ricordiamo che per taluni processi può essere più conveniente l'utilizzo di modelli basati sulle equazioni alle differenze.

1, si traduce nel modo seguente

$$\frac{dy_1}{dt} = I(t) - \text{velocità di distribuzione da 1 a 2} \quad (12.1)$$

e analogamente per il compartimento 2

$$\frac{dy_2}{dt} = \text{velocità di ingresso} - \text{velocità di uscita} \quad (12.2)$$

A questo punto rimane l'aspetto importante di come modellizzare matematicamente le velocità di ingresso e di uscita. Assumendo, *ad esempio*, una *cinetica del primo ordine*, si ha che tali velocità sono proporzionali alla massa del farmaco nel corrispondente compartimento. In questo caso, le equazioni (12.1) e (12.2) assumono la seguente forma

$$\frac{dy_1}{dt} = I(t) - f_{21}y_1 \quad (12.3)$$

$$\frac{dy_2}{dt} = f_{21}y_1 - f_{02}y_2 \quad (12.4)$$

ove  $f_{21} > 0$  e  $f_{02} > 0$  sono i fattori di proporzionalità, che possono dipendere, per alcuni tipi di farmaci dalla variabile  $t$ .

Le equazioni (12.3) e (12.4), insieme ad appropriate *condizioni iniziali*  $y_1(0), y_2(0)$ , costituiscono un *modello matematico* del metabolismo di un farmaco. Nelle applicazioni interessa, in particolare, conoscere  $y_2(t)$ , che fornisce la variazione nel tempo della massa del farmaco nella circolazione sanguigna. La sua conoscenza, insieme alle indicazioni sull'effetto del farmaco, permette di impostare in maniera razionale il problema del *dosaggio ottimale* (mediante la risoluzione di un problema di controllo, cfr. il successivo Capitolo 14) Perché questo sia possibile, rimane tuttavia da risolvere il problema del *legame tra il modello e i dati*, ossia della individuazione dei parametri  $f_{21}, f_{02}$ .

In Figura 12.3 sono rappresentate le soluzioni  $y_1(t)$  e  $y_2(t)$ , con  $y_1(0) = y_2(0) = 0$  e corrispondenti ad una particolare forma della funzione  $I(t)$  e a due coppie differenti di valori dei coefficienti  $f_{21}, f_{02}$ .

Tali soluzioni, ottenute mediante un procedimento numerico, danno un'idea di come le variabili di stato dipendono dai parametri. Ulteriori indicazioni possono essere ottenute studiando il comportamento asintotico (ossia per  $t \rightarrow \infty$ ) delle soluzioni. Dal punto di vista *sperimentale* si possono avere indicazioni, mediante opportuni campionamenti del sangue, sulla variabile  $y_2(t)$ . Il confronto tra tali valori e quelli relativi alla variabile  $y_2(t)$  fornita dal modello matematico è alla base del procedimento di identificazione dei parametri. ■

► **Esempio 12.2** (*Cinetica di un antibiotico*) Sperimentalmente si è trovato che gli antibiotici del tipo sulfamidina (SDM) sono biotrasformati nell'organismo degli animali; in particolare, la SDM viene trasformato nel metabolita inattivo N-acetil-SDM. Si è inoltre verificato che tale metabolita, quando conservato in concime liquido, dopo un periodo di induzione di circa 60 giorni viene reattivato, grazie alla mediazione di microorganismi, che sviluppano la capacità di operare tale trasformazione nel corso di un processo evolutivo.

Indicando con  $x(t)$  la concentrazione nel tempo del metabolita, con  $y(t)$  la concentrazione di SDM, con  $\mu$  la velocità di eliminazione di SDM nel concime liquido e con  $k(t)$  la velocità di trasformazione del metabolita in SDM, si può studiare il fenomeno attraverso il seguente

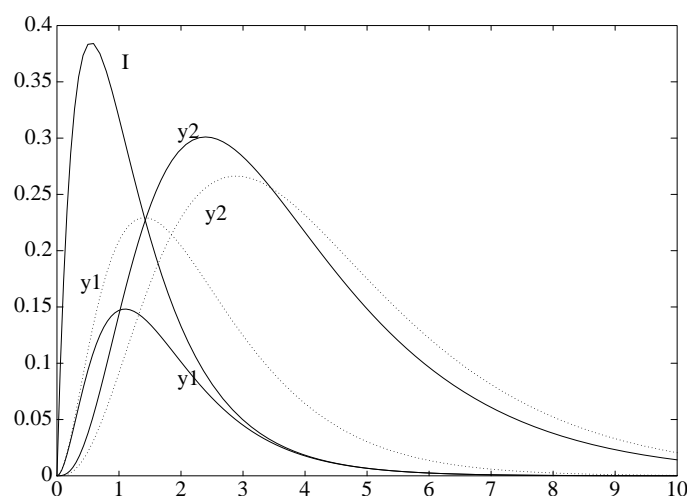


Figura 12.3: Modello a compartimenti della cinetica di un farmaco. Le curve a tratto continuo corrispondono ai valori dei parametri  $f_{21} = 1$  e  $f_{02} = 0.5$ , mentre le curve a punti corrispondono ai valori  $f_{21} = 2$  e  $f_{02} = 0.5$ . L'intensità di introduzione del farmaco è simulata dalla funzione  $I(t) = e^{-t} - e^{-3t}$ .

modello matematico

$$\frac{dx}{dt} = -k(t)x \quad (12.5)$$

$$\frac{dy}{dt} = k(t)x - \mu y \quad (12.6)$$

con opportune condizioni iniziali. La funzione  $k(t)$  è modellizzata nel seguente modo

$$k(t) = k_{\max}(1 - e^{-(t/t_c)^\beta}) \quad (12.7)$$

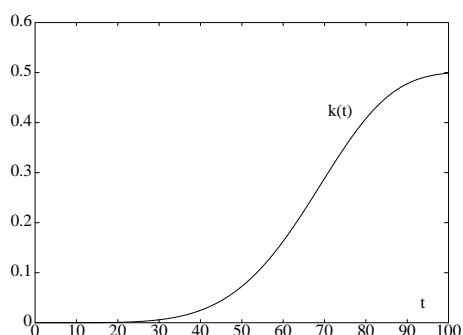
ove  $k_{\max}$  ha il significato di velocità massima di reattivazione,  $t_c$  è relativo al tempo di sviluppo della capacità di reattivazione, e  $\beta$  è un parametro utilizzato per dare la forma opportuna alla curva  $k(t)$ . Il modello dipende, quindi, dai 4 parametri  $k_{\max}$ ,  $t_c$ ,  $\beta$  e  $\mu$ .

Indichiamo con  $x^*$  e  $y^*$  i valori di SDM, rispettivamente inattivo e attivo, ottenuti sperimentalmente (cfr. Figura 12.4). I parametri del modello possono allora essere identificati mediante la minimizzazione del seguente stimatore, del tipo minimi quadrati

$$F(k_{\max}, t_c, \beta, \mu) = \sum [x(t_i) - x^*(t_i)]^2 + [y(t_i) - y_i^*(t_i)]^2 \quad (12.8)$$

ove la sommatoria è estesa ai punti  $t_i$  relativi alle osservazioni sperimentali. Sottolineiamo il fatto che per costruire lo stimatore  $F$  è necessaria la risoluzione del sistema differenziale (12.5), (12.6).

A titolo di esemplificazione, riportiamo i risultati di identificazione ottenuti mediante la seguente procedura in linguaggio MATLAB. Per la risoluzione del sistema differenziale viene utilizzato il metodo di Runge-Kutta, mentre la minimizzazione è ottenuta mediante il metodo del semplice geometrico.



SDM: dati sperimentali		
$t_i$	inattivo	attivo
0	69.0	26.1
3	62.5	23.3
6	70.1	24.3
11	62.3	25.9
26	62.6	26.6
38	54.9	30.5
48	42.0	36.8
59	8.9	75.3
68	2.8	70.9
76	1.7	85.1

Figura 12.4: Rappresentazione della velocità di reazione  $k(t)$  corrispondente ai parametri identificati sulla base dei dati sperimentali.

```

global kmm tcc bee muu
global td data
td =      data =
    0      26.1000  69.0000
    3      23.3000  62.5000
    6      24.3000  70.1000
   11      25.9000  62.3000
   26      26.6000  62.6000
   38      30.5000  54.9000
   48      36.8000  42.0000
   59      75.3000   8.9000
   68      70.9000   2.8000
   76      85.1000   1.7000

p0=[0.4 80 0.003 6];
p=fmins('etas',p0);

function ydot=leta(t,y) %% sistema differenziale
ydot(1)=-k(t)*y(1);
ydot(2)= k(t)*y(1)-muu*y(2);

function kt=k(t)
kt=kmm*(1-exp(-(t/tcc).^bee));

function zs=etas(p) %% costruzione dello stimatore
kmm=p(1);tcc=p(2);muu=p(3);bee=p(4);
yz(1)=data(1,2);yz(2)=data(1,1);
yc(1)=0;zc(1)=0;
nz=1;
for i=1:length(td)-1
t0=td(i);
y0=yz(nz,:);
%%% risoluzione sistema differenziale

```

```
[t,yz]=ode45('leta',t0,td(i+1),y0,1.e-7);
nz=length(t);
yc(i+1)=yz(nz,1)-data(i+1,2);zc(i+1)=yz(nz,2)-data(i+1,1);
end
zs=sum(yc.^2)+sum(zc.^2); %%% stimatore
```

Si ottengono per i parametri le seguenti stime

$$k_{\max} = 0.511; \quad t_c = 72.511; \quad \mu = 0.00518; \quad \beta = 5.014$$

Le corrispondenti soluzioni del modello matematico sono rappresentate in Figura 12.5. ■

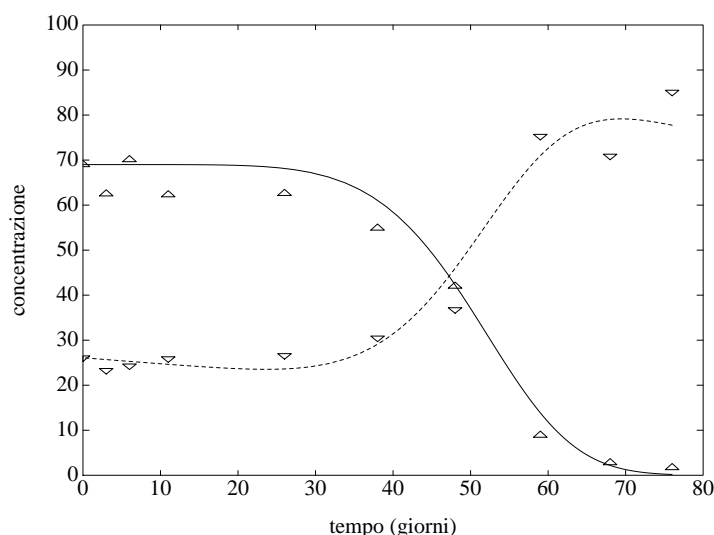


Figura 12.5: Cinetica dell'antibiotico sulfadimidina ( $\nabla$ ) e del suo metabolita ( $\Delta$ ) in concime liquido. Le curve continue sono soluzioni del modello differenziale corrispondenti ai valori dei parametri ottenuti con una procedura di stima basata sui minimi quadrati.

► **Esempio 12.3** (*Diffusione*) Consideriamo la diffusione di un materiale tra due compartimenti separati da una membrana attraverso la quale il materiale può diffondere, non necessariamente con la stessa permeabilità nelle due direzioni (cfr. Figura 12.6). Siano  $c_1(t)$ ,  $c_2(t)$  le concentrazioni del materiale nei due compartimenti, che vengono supposti perfettamente miscelati e di volume, rispettivamente  $V_1$ ,  $V_2$ , fissato. Dal momento che il volume dei due compartimenti è supposto costante, si ha  $y_i(t) = c_i(t)V_i$ ,  $i = 1, 2$ , ove  $y_i(t)$  indica la massa del materiale nel compartimento  $i$  al generico istante  $t$ . La *legge di Fick* stabilisce che la velocità del trasferimento per diffusione del materiale attraverso il piano della membrana è proporzionale al prodotto dell'area  $A$  della superficie della membrana per il gradiente della concentrazione tra le due regioni. Supponendo che la permeabilità possa

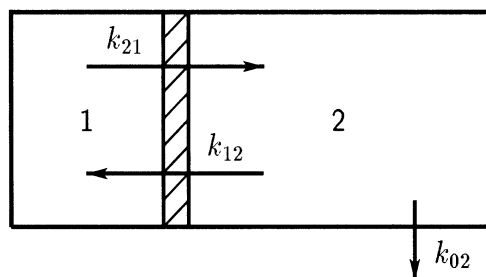


Figura 12.6: Sistema a due compartimenti corrispondente alla diffusione di un materiale.

essere diversa nelle due direzioni, si ottengono le seguenti equazioni

$$\frac{dy_1}{dt} = -k_{21}Ac_1 + k_{12}Ac_2 \quad (12.9)$$

$$\frac{dy_2}{dt} = k_{21}Ac_1 - k_{12}Ac_2 - k_{02}c_2 \quad (12.10)$$

ove si è supposto che l'escrezione del materiale dal compartimento 2 avvenga seguendo una cinetica del primo ordine. Introducendo le seguenti quantità, chiamate *coefficienti frazionali di trasferimento* (fractional transfer coefficients)

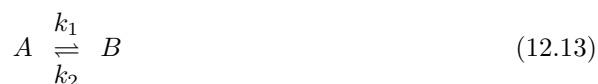
$$f_{ij} = \frac{k_{ij}A}{V_j}, \quad i, j = 1, 2, \quad f_{02} = \frac{k_{02}}{V_2} \quad (12.11)$$

il modello precedente può essere scritto nella seguente forma, che è analoga in struttura a quella del modello studiato nell'esempio 12.1.

$$\begin{bmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \end{bmatrix} = \begin{bmatrix} -f_{21} & f_{12} \\ f_{21} & -f_{02} - f_{12} \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} \quad (12.12)$$

■

► **Esempio 12.4** (*Reazioni chimiche del primo ordine*) Come primo esempio, consideriamo la seguente reazione monomolecolare reversibile (cfr. Figura 12.7)



Se  $y_1$  e  $y_2$  indicano le concentrazioni, rispettivamente di  $A$  e  $B$ , dalla legge di massa azione si ottiene

$$\frac{dy_1}{dt} = -k_1y_1 + k_2y_2 \quad (12.14)$$

$$\frac{dy_2}{dt} = k_1y_1 - k_2y_2 \quad (12.15)$$



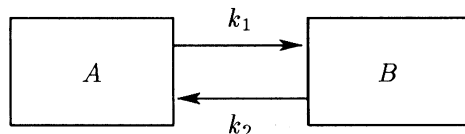


Figura 12.7: Sistema a due compartimenti corrispondente ad una reazione chimica del primo ordine.

Si tratta di un sistema *chiuso*; come conseguenza,  $\dot{y}_1 + \dot{y}_2 = 0$  e per ogni  $t > 0$  si ha  $y_1(t) + y_2(t) = y_1(0) + y_2(0)$ .

Consideriamo, come ulteriore esempio illustrativo, la decomposizione del *calcare* nei suoi prodotti principali, l'*ossido di calcio* e di *magnesio*



ove  $k_i$  sono le corrispondenti costanti di velocità. Supponiamo che il calcare consista di una frazione  $\beta$  di  $\text{Ca CO}_3$  e una frazione  $1 - \beta$  di  $\text{Mg CO}_3$  ( $0 < \beta < 1$ ). Ogni mole del reagente che si decompone fornisce una mole di prodotto più una mole di anidride carbonica; si suppone che l'anidride carbonica  $\text{CO}_2$  non influenzi la velocità con cui avviene la reazione.

Le reazioni (12.16) avvengono in un reattore tenuto ad una temperatura costante. A partire da  $t = 0$ , indichiamo con  $y_1(t)$ ,  $y_2(t)$ ,  $y_3(t)$ ,  $y_4(t)$ , rispettivamente le masse di  $\text{Ca CO}_3$ ,  $\text{Mg CO}_3$ ,  $\text{Ca O}$ ,  $\text{Mg O}$  presenti nel reattore al tempo  $t$ . Indichiamo, inoltre, con  $u(t)$  la velocità alla quale il calcare è aggiunto al reattore. Nell'*ipotesi* che la cinetica delle reazioni sia del primo ordine (ossia che la velocità di reazione sia proporzionale alla massa del reagente), le reazioni (12.16) sono descritte dalle seguenti equazioni differenziali

$$\begin{aligned} \dot{y}_1(t) &= \beta u(t) - k_1 y_1(t) \\ \dot{y}_2(t) &= (1 - \beta)u(t) - k_2 y_2(t) \end{aligned} \quad (12.17)$$

In maniera analoga, si ottengono le equazioni che descrivono la formazione dei prodotti della reazione

$$\begin{aligned} \dot{y}_3(t) &= k_1 y_1(t) \\ \dot{y}_4(t) &= k_2 y_2(t) \end{aligned} \quad (12.18)$$

Posto  $f_{31} = k_1$  e  $f_{42} = k_2$ , le equazioni precedenti costituiscono il seguente sistema di equazioni differenziali, che descrive il *modello della riduzione del calcare*

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{bmatrix} = \begin{bmatrix} -f_{31} & 0 & 0 & 0 \\ 0 & -f_{42} & 0 & 0 \\ f_{31} & 0 & 0 & 0 \\ 0 & f_{42} & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} + \begin{bmatrix} \beta \\ 1 - \beta \\ 0 \\ 0 \end{bmatrix} u \quad (12.19)$$

Sottolineiamo una proprietà della matrice dei coefficienti (verificata anche per gli esempi precedenti): gli elementi fuori della diagonale sono non negativi e la somma degli elementi di

ogni colonna è non positiva (e, quindi, gli elementi sulla diagonale sono non positivi). Da tale proprietà, come vedremo nel seguito, si possono ottenere informazioni sulla *stabilità* (ossia sul comportamento per  $t \rightarrow \infty$ ) delle soluzioni del modello matematico. Nella Figura 12.8 è illustrato il comportamento del modello in due differenti situazioni. ■

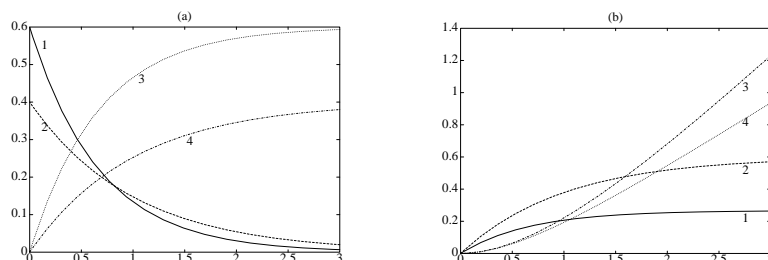


Figura 12.8: Modello di decomposizione del calcare. In figura (a) si ha  $u(t) \equiv 0$  e per  $t = 0$  si ha una quantità  $M = 1$  di calcare. In (b) si ha  $u(t) = 1$  e  $y_i(0) = 0$ , per  $i = 1, \dots, 4$ .

► **Esempio 12.5** (*Reattore chimico a flusso continuo*, continuous flow chemical reactor)

In un recipiente, tenuto a volume e temperatura costante, vi è un flusso continuo di reagenti. Si suppone, inoltre, che tali reagenti vengano istantaneamente e perfettamente miscelati tra loro e che una parte del prodotto della reazione venga prelevata dal recipiente. Indichiamo con  $R$  la velocità del flusso entrante e uscente dal reattore.

Come esempio illustrativo, consideriamo la reazione irreversibile tra una mole di  $\text{H}_2\text{O}$  con una mole di  $\text{SO}_3$  per la formazione di una mole di acido solforico  $\text{H}_2\text{SO}_4$



Indichiamo con  $u_1$  (rispettivamente  $u_2$ ) la concentrazione molare di  $\text{H}_2\text{O}$  (rispettivamente di  $\text{SO}_3$ ) in *input* e  $c_1, c_2, c_3$  le concentrazioni molari, rispettivamente di  $\text{H}_2\text{O}$ ,  $\text{SO}_3$ ,  $\text{H}_2\text{SO}_4$ .

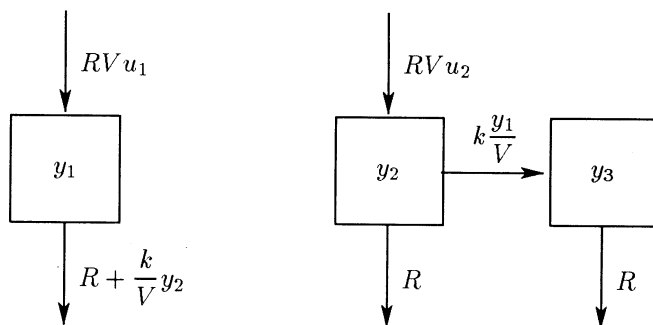


Figura 12.9: Sistema a compartimenti non lineare.

Alla base della costruzione del modello matematico vi è ancora la *legge di massa azione*. In termini schematici, abbiamo una molecola  $\alpha$  di un composto chimico e una molecola  $\beta$  di un secondo composto chimico che si combinano per dare origine a una molecola  $\gamma$  di un terzo composto chimico. La reazione avviene a seguito di una collisione tra  $\alpha$  e  $\beta$ . Più precisamente, si *suppone* che la produzione di molecole  $\gamma$  sia proporzionale<sup>4</sup> al numero di incontri tra  $\alpha$  e  $\beta$ . Allora, in termini di concentrazioni  $[\cdot]$  delle sostanze chimiche, si ha la seguente equazione (*reazione del secondo ordine*)

$$\frac{d[\gamma]}{dt} = k [\alpha] [\beta]$$

ove  $k > 0$  è la velocità di reazione. Applicando tale risultato all'esempio precedente, abbiamo per la concentrazione di  $\text{H}_2\text{SO}_4$

$$\frac{dc_3}{dt} = \text{guadagno} - \text{perdita} = k c_1 c_2 - R c_3 \quad (12.21)$$

In maniera analoga, si ottengono le equazioni relative ai reagenti

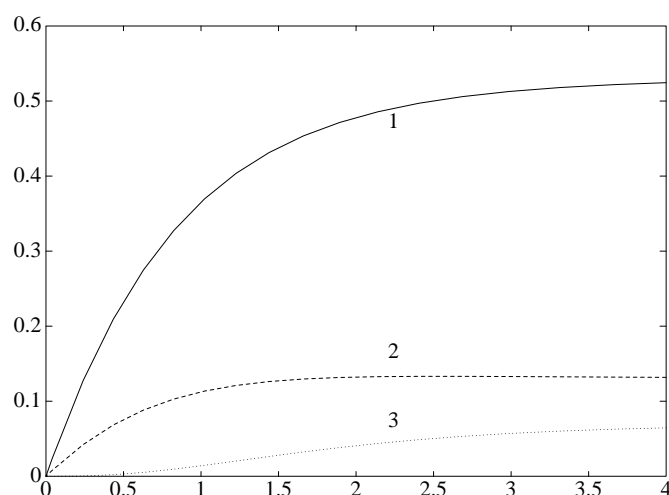


Figura 12.10: Reattore chimico a flusso continuo relativo ai reagenti  $\text{H}_2\text{O}$  (1),  $\text{SO}_3$  (2) e al prodotto  $\text{H}_2\text{SO}_4$  (3).

$$\frac{dc_1}{dt} = \text{guadagno} - \text{perdita} = R u_1 - (R c_1 + k c_1 c_2) \quad (12.22)$$

$$\frac{dc_2}{dt} = \text{guadagno} - \text{perdita} = R u_2 - (R c_2 + k c_1 c_2) \quad (12.23)$$

<sup>4</sup>Tale ipotesi è alla base di numerosi altri modelli matematici, in particolare, nello studio di popolazioni. Sottolineiamo la natura essenzialmente *empirica* dell'ipotesi, e quindi la necessità di una sua *verifica sperimentale*. Nell'ambito delle reazioni chimiche (cfr. Capitolo 7), tale verifica risale ai lavori di Guldberg e Waage (1867).

In termini delle quantità  $y_i(t) = Vc_i(t)$ ,  $i = 1, 2, 3$ , e posto

$$f_{11} = -\left(R + \frac{k}{V}y_2\right), \quad f_{22} = -\left(R + \frac{k}{V}y_1\right), \quad f_{32} = \frac{k}{V}y_1, \quad f_{33} = -R$$

le equazioni precedenti si scrivono nel modo seguente

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} f_{11} & 0 & 0 \\ 0 & f_{22} & 0 \\ 0 & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} RVu_1 \\ RVu_2 \\ 0 \end{bmatrix} \quad (12.24)$$

che corrisponde al diagramma di compartimenti rappresentato in Figura 12.9. Il sistema (12.24) è *non lineare*, in quanto i coefficienti  $f_{ij}$  sono funzioni delle variabili  $y_i$ . Tuttavia, essi verificano ancora le proprietà rilevate negli esempi precedenti: gli elementi fuori della diagonale sono non negativi e ogni colonna ha somma non positiva. I risultati numerici rappresentati in Figura 12.10 indicano il comportamento asintotico delle soluzioni. ■

► **Esempio 12.6** (*Teoria matematica delle epidemie*) Supponiamo che in una determinata popolazione esistano due gruppi di individui: infettivi e suscettibili di infezione. Indichiamo con  $y_1(t)$  il numero dei suscettibili e con  $y_2(t)$  quello degli infettivi, ad ogni tempo  $t \geq 0$ . Se la malattia si diffonde per contatto diretto tra gli individui delle due classi, si può ipotizzare che la velocità alla quale i suscettibili ricevono l'infezione sia  $\beta y_1 y_2$ , con  $\beta > 0$  opportuno parametro dipendente dal tipo di malattia e dalle caratteristiche della popolazione. Gli infettivi possono morire o diventare immuni alla malattia alla velocità  $-\gamma y_2$ , con  $\gamma > 0$  costante opportuna. Inoltre, il numero dei suscettibili può essere ridotto (mediante l'introduzione di vaccinazioni) ad una velocità  $-u(t)$ . Con ragionamento analogo a quello seguito negli esempi precedenti, si arriva, allora, al seguente modello deterministico, di tipo compartimentale (non lineare)

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} f_{11} & 0 \\ f_{21} & f_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} -u \\ 0 \end{bmatrix} \quad (12.25)$$

ove  $f_{11} = -\beta y_2$ ,  $f_{21} = \beta y_2$ ,  $f_{22} = -\gamma$ .

I risultati riportati in Figura 12.11 mostrano l'influenza sul comportamento delle due popolazioni variare della funzione  $u(t)$ . In effetti, quando nel modello sono stati identificati i parametri  $\beta$  e  $\gamma$ , la funzione  $u(t)$  rappresenta una *funzione di controllo* del sistema. ■

## 12.2 Modello a compartimenti generale

I modelli esaminati nel paragrafo precedente sono esempi particolari della seguente situazione generale. Supponendo di avere un sistema a  $n$  compartimenti, numerati da 1 a  $n$ , la descrizione generale della dinamica dello scambio di materiale in corrispondenza al generico compartimento  $i$ -mo è la seguente *equazione di bilancio* di massa:

$$\dot{y}_i = \text{velocità di entrata} - \text{velocità di uscita} \quad i = 1, 2, \dots, n \quad (12.26)$$

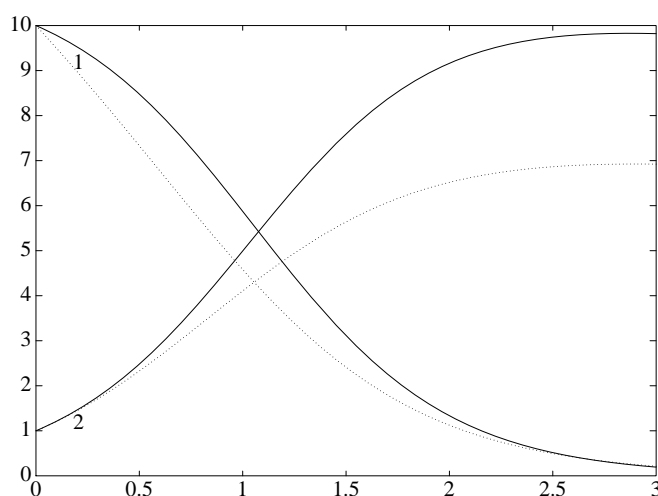


Figura 12.11: Modello di diffusione di una epidemia. Le curve continue si riferiscono al caso in cui  $u(t) \equiv 0$ , mentre le curve a punti corrispondono a  $u(t) \equiv 0.3$ . Per gli altri parametri si sono assunti i valori  $\gamma = 0.05$ ,  $\beta = 0.2$ ,  $y_1(0) = 10$ ,  $y_2(0) = 1$ .

ove  $y_i(t) \geq 0$  è la quantità del materiale nel compartimento  $i$  al tempo  $t$ . La velocità di trasferimento di materiale dal compartimento  $j$  al compartimento  $i$  ( $i \neq j$ ) è modellizzata da  $f_{ij}y_j$ , dove  $f_{ij}$  è una quantità non negativa, chiamata *coefficiente di trasferimento frazionale* (fractional transfer coefficient), che può essere una funzione delle componenti del vettore delle quantità  $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ , del tempo  $t$  e di un vettore di parametri  $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_\nu]^T$ . Dalle relazioni (12.26) si ottiene il seguente *modello a compartimenti di tipo generale* (cfr. Figura 12.12)

$$\dot{y}_i = \sum_{\substack{j=1 \\ j \neq i}}^n f_{ij}(\mathbf{y}(t), t, \boldsymbol{\alpha}) y_j(t) + I_i(t) - \sum_{\substack{j=0 \\ j \neq i}}^n f_{ji}(\mathbf{y}(t), t, \boldsymbol{\alpha}) y_i(t) \quad (12.27)$$

per  $i = 1, 2, \dots, n$  e per  $t \in [0, \bar{t}]$ , con  $\bar{t}$  assegnato. Le funzioni  $I_i(t)$  rappresentano la velocità di *input* del materiale nel compartimento  $i$ -mo dall'esterno, e  $f_{0i}$  è il coefficiente frazionale di uscita, in modo che  $f_{0i}y_i$  rappresenta la velocità di uscita all'esterno del sistema dal compartimento  $i$ -mo.

Posto

$$f_{ii} = - \sum_{\substack{j=0 \\ j \neq i}}^n f_{ji} \quad i = 1, 2, \dots, n$$

il flusso totale dal compartimento  $i$  agli altri compartimenti e verso l'esterno del sistema è dato da  $f_{ii}y_i$ . In forma matriciale, il modello (12.27) equivale al seguente

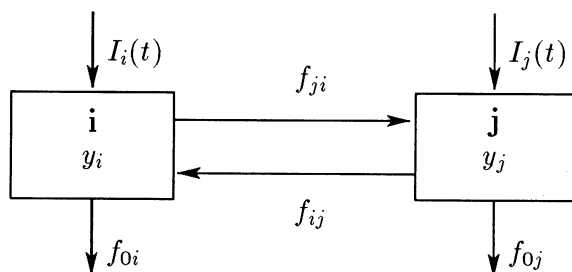


Figura 12.12: Sistema a compartimenti generale.

sistema di equazioni differenziali

$$\dot{\mathbf{y}} = \mathbf{F} \mathbf{y} + \mathbf{I} \quad (12.28)$$

ove  $\mathbf{F} = [f_{ij}]$  è la matrice di ordine  $n$  dei coefficienti frazionali di trasferimento e  $\mathbf{I} = [I_1, I_2, \dots, I_n]^T$ . Quando le componenti del vettore  $\mathbf{y}$  sono assegnate per un particolare valore di  $t$ , il modello matematico corrispondente al modello a compartimenti diventa un *problema differenziale a valori iniziali*, la cui soluzione può essere ottenuta mediante i metodi numerici esaminati nel Capitolo 7. Talvolta, l'interesse può essere limitato all'analisi della soluzioni allo *stato stazionario* (steady-state), ossia del comportamento delle soluzioni per  $t \rightarrow \infty$ . Quando  $I_i$  sono costanti, lo stato stazionario è associato alle quantità  $y_i$  tali che  $dy_i/dt = 0$ . La ricerca di tali soluzioni diventa, allora, un problema di risoluzione di sistemi (eventualmente non lineari) del tipo  $\mathbf{F} \mathbf{y} + \mathbf{I} = 0$ . Nel seguito esamineremo, in particolare, il caso in cui  $f_{ij}$  sono *costanti*, ossia dipendenti solo da alcuni parametri del modello, ma non dal vettore  $\mathbf{y}$  e dal tempo  $t$ .

### 12.3 Cinetica dei traccianti

Un organismo vivente non è isolato dall'ambiente che lo circonda, ma è in continua interazione con esso. In particolare, assorbe energia dall'ambiente, restituendo prodotti di scarto. Un organismo vivente è, quindi, un *sistema aperto*, e la sua interazione con il suo ambiente può essere caratterizzata approssimativamente come uno stato stazionario, nel quale vi sono continui scambi di energia e di materia con l'ambiente.

Gli organismi di livello superiore tendono a mantenere i loro processi e le loro caratteristiche, come ad esempio la temperatura, il livello di glucosio nel sangue, costantemente entro certi limiti, nonostante piccoli disturbi. Tale proprietà di relativa indipendenza dall'ambiente è detta *omeostasi*.

Supponiamo ora che le proprietà di un sistema fisiologico siano studiate mediante un modello a compartimenti. In tale tipo di modellizzazione, il sistema è caratterizzato, come abbiamo visto nei paragrafi precedenti, dai coefficienti di trasferimento  $f_{ij}$ .

I *dati sperimentali* per la stima di tali parametri possono essere ottenuti mediante la cosiddetta *tecnica dei traccianti*, che, in sostanza, consiste in una perturbazione dello stato stazionario in modo da ottenere uno stato *transiente* osservabile. Più precisamente, viene introdotta, diciamo al tempo  $t = 0$ , in uno o più compartimenti del sistema una quantità fissata di *tracciante* (un colorante o un isotopo radioattivo), di volume trascurabile rispetto a quello dei compartimenti del sistema. L'introduzione del materiale etichettato (*labeled* o *tagged*) al tempo zero perturba il sistema, in quanto improvvisamente vi è un piccolo eccesso di materiale; si suppone, comunque, che le caratteristiche dello stato stazionario rimangano immutate. Si suppone, cioè, che il comportamento dinamico del materiale etichettato in un compartimento sia rappresentativo del comportamento di tutto il materiale.

Per  $t > 0$ , in ogni compartimento del sistema vi sono due tipi di particelle: quelle del materiale etichettato e quelle del materiale non etichettato (*unlabeled*, indicato anche come materiale *tracce*). I due tipi di materiale sono, in generale, indistinguibili, ma si suppone che la quantità del materiale etichettato possa essere osservata e rilevata sperimentalmente in istanti successivi. Tali osservazioni sperimentali sono alla base della costruzione di un conveniente stimatore per l'identificazione dei parametri del modello, e risolvere, quindi, il *problema inverso*: dati i risultati forniti dal modello, stimare i parametri.

A scopo illustrativo, consideriamo un sistema a due compartimenti e supponiamo che le funzioni di input  $I_1, I_2$  siano costanti. Con le notazioni del paragrafo precedente il modello può essere rappresentato nella seguente forma

$$\dot{y}_i = F_i(y_1, y_2) + I_i \quad i = 1, 2 \quad (12.29)$$

Se  $\bar{y}_i$  indica la quantità costante di materiale non etichettato nel compartimento  $i$ , allo stato stazionario si ha

$$0 = \dot{\bar{y}}_i = F_i(\bar{y}_1, \bar{y}_2) + I_i \quad i = 1, 2 \quad (12.30)$$

Supponiamo, ora, che al tempo  $t = 0$  il sistema sia perturbato dall'iniezione di una piccola quantità di tracciante e indichiamo con  $b_i(t)$  la velocità di input del tracciante nel compartimento  $i$ . Sia, inoltre,  $x_i(t)$  la quantità<sup>5</sup> di tracciante nel compartimento  $i$  al tempo  $t \geq 0$ . Per il fatto che il materiale tracciante è tracciato

---

<sup>5</sup>Quando il materiale è etichettato mediante un marcatore radioattivo, la quantità  $x(t)$  rappresenta la quantità totale di materiale radioattivo che è misurato in unità di radioattività o semplicemente *attività*, come i microcuries ( $\mu\text{Ci}$ ), disintegrazioni per secondo (dis/sec), o colpi per minuto (cpm). Il *curie* è definito come il numero di disintegrazioni per secondo che avvengono in 1 g di radio, oppure  $1\text{Ci} = 3.7 \times 10^{10}$  dis/sec.

sono indistinguibili, possiamo supporre che la quantità totale  $Q_i = x_i(t) + \bar{q}_i$  del tracciante e del tracciato nel compartimento  $i$ -mo soddisfi alle seguenti equazioni

$$\begin{aligned}\dot{Q}_1 &= F_1(Q_1, Q_2) + I_1 + b_1 \\ \dot{Q}_2 &= F_2(Q_1, Q_2) + I_2 + b_2\end{aligned}\quad (12.31)$$

Possiamo ora ricavare un modello per la dinamica del tracciante, utilizzando l'ipotesi che le quantità  $x_i(t)$  siano piccole. Tale ipotesi può essere precisata, scrivendo che  $x_i(t) = \epsilon z_i(t)$ , con  $0 < |\epsilon| \ll 1$ . Dalle equazioni (12.31) si ricava allora per  $i = 1, 2$

$$\epsilon \dot{z}_i = F_i(\bar{y}_1 + \epsilon z_1, \bar{y}_2 + \epsilon z_2) + I_i + b_i$$

da cui, mediante uno sviluppo in serie, arrestato ai termini di primo ordine

$$\begin{aligned}\epsilon \dot{z}_i &= F_i(\bar{y}_1, \bar{y}_2) + \epsilon z_1 \frac{\partial F_i}{\partial y_1}(\bar{y}_1, \bar{y}_2) \\ &+ \epsilon z_2 \frac{\partial F_i}{\partial y_2}(\bar{y}_1, \bar{y}_2) + O(\epsilon^2) + I_i + b_i\end{aligned}$$

Trascurando, infine, i termini del secondo ordine  $O(\epsilon^2)$ , si ottiene

$$\dot{x}_i = \frac{\partial F_i}{\partial y_1}(\bar{y}_1, \bar{y}_2)x_1 + \frac{\partial F_i}{\partial y_2}(\bar{y}_1, \bar{y}_2)x_2 + b_i \quad (12.32)$$

per  $i = 1, 2$ . In questo modo si vede che le quantità  $x_i$  sono soluzioni di un sistema di equazioni differenziali *lineari* con coefficienti

$$a_{ij} = \frac{\partial F_i}{\partial y_j}(\bar{y}_1, \bar{y}_2) \approx f_{ij}, \quad (i, j = 1, 2) \quad (12.33)$$

Si può, pertanto, concludere con il seguente risultato importante: *nei limiti delle ipotesi fatte, i coefficienti di trasferimento  $a_{ij}$  relativi al modello lineare del tracciante forniscono una stima dei coefficienti di trasferimento  $f_{ij}$  del modello compartimentale.*

### 12.3.1 Modelli a compartimenti lineari

Le considerazioni del paragrafo precedente possono, naturalmente, essere estese al caso di  $n$  compartimenti e mettono in evidenza l'interesse dello studio dei *sistemi a compartimenti lineari*, che sono della forma

$$\dot{x}_i(t) = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j - \sum_{\substack{j=0 \\ j \neq i}}^n a_{ji}x_i + b_i(t) \quad (12.34)$$



con  $i = 1, 2, \dots, n$ . Posto

$$a_{ii} = - \sum_{\substack{j=0 \\ j \neq i}}^n a_{ji}$$

si ha che la matrice dei coefficienti  $\mathbf{A} = [a_{ij}]$ , detta *matrice compartimentale*, verifica le seguenti proprietà<sup>6</sup>

$$a_{ij} \geq 0 \quad \text{per } i, j = 1, 2, \dots, n \text{ con } i \neq j \quad (12.35)$$

$$a_{ii} \leq 0 \quad \text{per } i = 1, 2, \dots, n \quad (12.36)$$

$$\sum_{i=1}^n a_{ij} = -a_{0j} \leq 0 \quad \text{per } j = 1, 2, \dots, n \quad (12.37)$$

Di conseguenza, la matrice  $\mathbf{A}$  risulta *diagonalmente dominante* per colonne, ossia verifica la seguente proprietà

$$|a_{ii}| \geq \sum_{\substack{r=1 \\ r \neq i}}^n |a_{ri}| \quad i = 1, 2, \dots, n$$

In conclusione, *la cinetica dei traccianti in un sistema generale a  $n$  compartimenti in uno stato stazionario è descritta dal seguente modello deterministico*

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t), & t \geq 0 \\ \mathbf{x}(0) = \mathbf{x}_0 \end{cases} \quad (12.38)$$

ove il vettore  $\mathbf{x}_0$  fornisce le condizioni *iniziali*.

► **Esempio 12.7** (*Cinetica del potassio nei globuli rossi*) Attraverso un esempio concreto introdurremo il problema della identificabilità dei coefficienti.

Nella circolazione del sangue nell'uomo vi è un continuo passaggio di ioni di potassio dal plasma ai globuli rossi e viceversa. La determinazione della velocità di ingresso e di uscita è di interesse in medicina per comprendere la funzione dei globuli rossi e la individuazione di alcune malattie del sangue.

Dalla fisiologia della circolazione del sangue si ha il suggerimento di un modello a due compartimenti (cfr. Figura 12.13); inoltre, si conosce, per via sperimentale, che la concentrazione del potassio nei globuli rossi è dovuta a un meccanismo di trasporto non lineare e che il livello di potassio nel plasma e nei globuli rossi è praticamente costante nel tempo. Possiamo quindi assumere che il sistema sia in stato stazionario.

<sup>6</sup>Ricordiamo che una matrice  $\mathbf{A}$  con la proprietà  $a_{ij} \geq 0$ , per tutti gli indici  $i, j$ , con  $i \neq j$ , è detta *essenzialmente non negativa*, o anche *Z-matrice* o *matrice di Metzler*. Tali matrici hanno interesse in varie applicazioni, come ad esempio nei modelli economici e nella risoluzione numerica delle equazioni alle derivate parziali. Si può mostrare che se  $\mathbf{b}(t) \geq 0$ , per  $t \geq 0$ , allora per la soluzione del sistema differenziale  $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$ , si ha  $\mathbf{x}(t) \geq 0$  se e solo se  $\mathbf{A}$  è una matrice di Metzler.

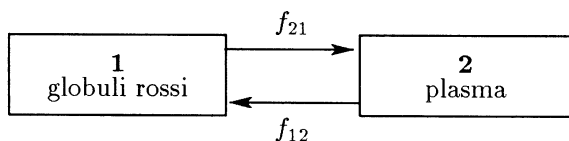


Figura 12.13: Sistema a due compartimenti rappresentante la cinetica degli ioni di potassio.

Indicando con  $y_1$  la quantità di ioni di potassio nel primo compartimento (globuli rossi) e, rispettivamente con  $y_2$  la quantità contenuta nel secondo compartimento (plasma), si ha il seguente modello matematico

$$\dot{y}_1 = -f_{21}(\mathbf{y}, t)y_1 + f_{12}(\mathbf{y}, t)y_2 \quad (12.39)$$

$$\dot{y}_2 = f_{21}(\mathbf{y}, t)y_1 - f_{12}(\mathbf{y}, t)y_2 \quad (12.40)$$

ove  $\mathbf{y} = [y_1, y_2]^T$ . In condizioni stazionarie si ha che  $y_1$ ,  $y_2$ ,  $f_{21}$  e  $f_{12}$  sono *costanti*. Esaminiamo, ora, l'applicazione del metodo del tracciante per stimare i coefficienti di permeabilità  $f_{21}$ ,  $f_{12}$ . Al tempo  $t = 0$ , si introduce nel plasma sanguigno una piccola quantità  $D$  di ioni radioattivi  $K^{42+}$  (il tracciante). Indicando con  $x_1(t)$ ,  $x_2(t)$  la quantità di potassio radioattivo contenuto, rispettivamente, nel compartimento 1 e 2 al tempo  $t \geq 0$ , si ha per la cinetica del tracciante il seguente sistema lineare

$$\dot{x}_i = a_{i1}x_1 + a_{i2}x_2, \quad i = 1, 2 \quad (12.41)$$

$$x_1(0) = 0, \quad x_2(0) = D$$

ove  $a_{11} = -f_{21}$ ,  $a_{12} = f_{12}$ ,  $a_{21} = f_{21}$ ,  $a_{22} = -f_{12}$ .

Esamineremo, ora, la questione della identificabilità dei coefficienti  $a_{ij}$  attraverso i valori  $x_2^*(t_i)$  corrispondenti alla quantità di ioni rilevata sperimentalmente nel plasma sanguigno in successivi istanti  $t_i$ ,  $i = 1, 2, \dots, r$ .

Dal momento che il sistema è stato supposto *chiuso*, si ha per ogni istante  $t \geq 0$ :  $x_1(t) + x_2(t) = D$ , da cui  $x_1 = D - x_2$  e quindi dalla seconda equazione in (12.41)

$$\dot{x}_2 = f_{21}(D - x_2) - f_{12}x_2 \quad (12.42)$$

da cui (cfr. Appendice B)

$$\frac{x_2(t)}{P} - 1 = e^{mt+b} \Rightarrow mt + b = \ln \left( \frac{x_2(t)}{P} - 1 \right) \quad (12.43)$$

ove

$$P = f_{21} \frac{D}{f_{21} + f_{12}}, \quad m = -(f_{21} + f_{12}), \quad b = \ln \frac{f_{12}}{f_{21}}$$

Osserviamo che il parametro  $P$  corrisponde alla quantità asintotica (per  $t \rightarrow \infty$ ) del tracciante nel secondo compartimento; esso, pertanto, può essere stimato sperimentalmente. Si vede, allora, che se i dati sperimentali sono in numero sufficiente, dalla equazione (12.43) è possibile, mediante il procedimento dei minimi quadrati lineari, identificare in maniera univoca

i parametri  $m$  e  $b$ . Da essi si passa, poi, facilmente al calcolo dei coefficienti  $f_{ij}$ . In Figura 12.14 sono rappresentati i risultati ottenuti in corrispondenza ai dati sperimentali riportati in Tabella 12.1. Per i parametri si ottengono i valori stimati  $m = -0.0017$ ,  $b = 1.6070$ .

$t$	0	250	700	1200	1700
$x_2/P - 1$	5	3.2	1.6	0.6	0.2

Tabella 12.1: Dati sperimentali relativi alla cinetica del potassio nel sangue.

Concludiamo l'esempio, sottolineando il fatto che la riduzione del modello ad una sola variabile è stata possibile grazie all'ipotesi che il sistema è chiuso. Lasciamo come esercizio lo studio del modello, quando è ammesso un flusso di ioni potassio da uno dei due compartimenti all'ambiente esterno. ■

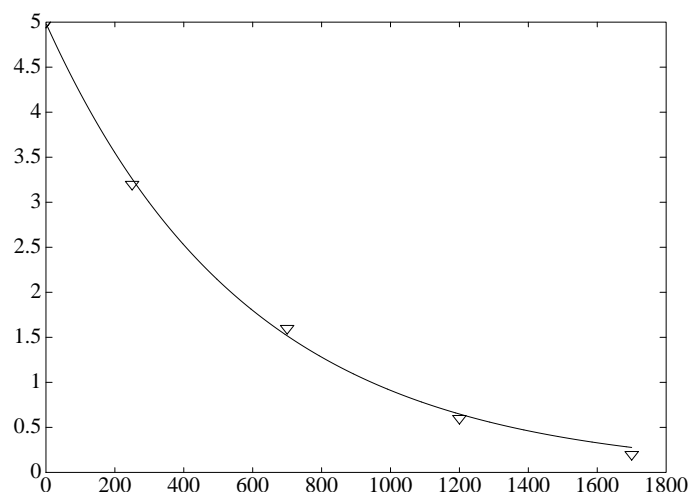


Figura 12.14: Dati sperimentali ( $\nabla$ ) e risultati forniti dal modello (-).

### 12.3.2 Equazioni della concentrazione del tracciante

Grazie all'ipotesi di stato stazionario, possiamo assumere che  $V_i$ , il volume (*size*) del compartimento  $i$ , e  $\bar{y}_i$ , la quantità di materiale non etichettato nel compartimento  $i$ , siano *costanti* nel tempo. Quando nelle *esperienze* viene rilevata la *concentrazione* del tracciante, le equazioni del modello devono essere riscritte in termini di concentrazioni, anziché di quantità. Poiché  $V_i > 0$  è costante, la concentrazione del tracciante nel compartimento  $i$  è data da

$$c_i(t) = \frac{x_i(t)}{V_i} \quad i = 1, 2, \dots, n$$

Indicando con  $\mathbf{V}$  la matrice diagonale  $\mathbf{V} = \mathbf{diag}(V_1, V_2, \dots, V_n)$  e posto  $\mathbf{c}(t) = [c_1(t), c_2(t), \dots, c_n(t)]^T$ , mediante la trasformazione lineare  $\mathbf{x} = \mathbf{V}\mathbf{c}$  si ottiene

$$\mathbf{V}\dot{\mathbf{c}} = \mathbf{A}\mathbf{V}\mathbf{c} + \mathbf{b} \Rightarrow \boxed{\dot{\mathbf{c}} = (\mathbf{V}^{-1}\mathbf{A}\mathbf{V})\mathbf{c} + \mathbf{V}^{-1}\mathbf{b}} \quad (12.44)$$

La matrice  $\mathbf{M} = \mathbf{V}^{-1}\mathbf{A}\mathbf{V}$  è *simile* alla matrice compartimentale  $\mathbf{A}$ ; le due matrici hanno, inoltre, la medesima diagonale. In modo analogo si costruiscono le equazioni compartimentali relative all'*attività specifica* del tracciante, definite nel modo seguente

$$a_i(t) := \frac{x_i(t)}{\bar{y}_i} \quad i = 1, 2, \dots, n$$

Posto  $\mathbf{Q} = \mathbf{diag}(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n)$ , e indicato con  $\mathbf{a}(t)$  il vettore di componenti  $a_i(t)$ , si ha

$$\dot{\mathbf{a}} = (\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q})\mathbf{a} + \mathbf{Q}^{-1}\mathbf{b} \quad (12.45)$$

È importante osservare che le matrici relative alle varie forme delle equazioni compartimentali sono matrici *simili* e hanno, quindi, il medesimo insieme di *autovalori*. Come vedremo nel seguito, tali autovalori caratterizzano le soluzioni del sistema compartimentale quando  $t \rightarrow \infty$  (studio della *stabilità* del sistema).

### 12.3.3 Struttura di un sistema e connettività

Alcune proprietà del modello a compartimenti

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t), \quad t \geq 0 \\ \mathbf{x}(0) &= \mathbf{x}_0 \end{aligned} \quad (12.46)$$

dipendono solo dalla disposizione degli elementi diversi dallo zero nella matrice  $\mathbf{A}$ . In particolare, se  $a_{ij} = 0$ , per  $i \neq j$ , non vi è flusso dal compartimento  $j$  al compartimento  $i$ .

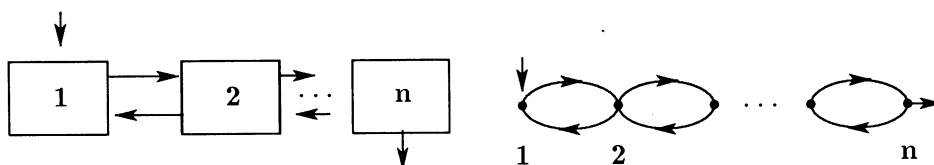


Figura 12.15: Sistema a compartimenti a catena e corrispondente diagramma di connettività.

La struttura di un compartimento può essere utilmente rappresentata mediante un grafo orientato, detto *diagramma di connettività*, nel quale i *nod*i rappresentano i compartimenti e gli archi orientati i flussi tra i compartimenti. Come esemplificazione, in Figura 12.15 è rappresentato il diagramma di connettività di un sistema

a  $n$  compartimenti di tipo particolare, detto sistema *a catena*. In tale struttura comunicano tra loro solo i compartimenti adiacenti, mentre l'input e l'output sono possibili solo nei nodi posti agli estremi della catena. La corrispondente matrice compartimentale  $\mathbf{A}$  è di tipo *tridiagonale*. Un esempio di sistema a catena è fornito dal sistema biologico corrispondente ai compartimenti identificati nel plasma sanguigno (1), nel fluido interstiziale (2) e dalle cellule (3).

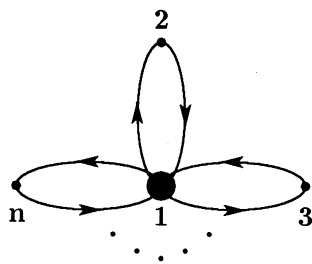


Figura 12.16: Sistema a compartimenti di tipo mammillare.

In Figura 12.16 è rappresentato il diagramma di connettività corrispondente al sistema a compartimenti noto come sistema di tipo *mammillare*. In tale struttura vi è un compartimento che agisce come compartimento *centrale*, e ad esso sono connessi tutti gli altri compartimenti. Indicando con 1 il compartimento centrale, la matrice compartimentale  $\mathbf{A}$  ha elementi diversi dallo zero solo sulla prima riga, la prima colonna e la diagonale principale, ossia ha la seguente forma

$$\begin{bmatrix} * & * & * & \cdots & * \\ * & * & 0 & \cdots & 0 \\ * & 0 & * & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ * & 0 & 0 & \cdots & * \end{bmatrix}$$

Ritornando al caso generale, ricordiamo una proprietà importante nello studio dei sistemi a compartimenti. Quando in un sistema due compartimenti, comunque scelti, sono connessi da un flusso, si dice che il corrispondente diagramma di connettività è *fortemente connesso*. Si può vedere che tale proprietà è equivalente a dire che la matrice  $\mathbf{A}$  è *irriducibile*.

La riducibilità della matrice  $\mathbf{A}$  comporta, previo opportuno ordinamento dell'insieme  $S$  dei compartimenti, l'esistenza di un sottoinsieme  $T$  di  $S$  tale che non vi sia flusso di materiale dai compartimenti in  $T$  agli altri compartimenti in  $S - T$ . L'esistenza di tali sottoinsiemi  $T$ , chiamati anche insiemi trappola (*trap*), riduce, in sostanza, la dimensione del problema, e quindi può semplificare lo studio del comportamento e della stabilità del sistema.

### 12.3.4 Stabilità

Con riferimento al modello (12.46), supponiamo che gli autovalori della matrice  $\mathbf{A}$  siano ordinati in ordine di grandezza

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n| \quad (12.47)$$

Dalle proprietà della matrice  $\mathbf{A}$  si ricava il seguente risultato.

**Proposizione 12.1** *La parte reale degli autovalori di  $\mathbf{A}$  è non positiva; inoltre, la matrice non ha autovalori puramente immaginari.*

**DIMOSTRAZIONE.** Per una matrice compartimentale  $\mathbf{A}$  si ha  $a_{ik} \geq 0$ , per ogni  $i \neq k$ , e inoltre la somma degli elementi della generica colonna  $k$ -ma è  $-a_{0k}$ , per cui

$$\sum_{\substack{i=1 \\ i \neq k}}^n a_{ik} = -a_{0k} - a_{kk} \leq -a_{kk} = |a_{kk}|$$

per ogni  $k = 1, 2, \dots, n$ . D'altra parte, dal teorema di localizzazione degli autovalori di Gershgorin (cfr. Appendice A) si ha che gli autovalori sono contenuti nell'unione dei dischi  $D_k$  definiti nel seguente modo

$$D_k := \left\{ z \in \mathbb{C} \mid |z - a_{kk}| \leq \sum_{\substack{i=1 \\ i \neq k}}^n |a_{ik}| \right\} \quad k = 1, 2, \dots, n$$

Il risultato richiesto è allora conseguenza del fatto che per il risultato precedente il disco  $D_k$  è contenuto nel cerchio  $|z - a_{kk}| \leq |a_{kk}|$ . ■

La non esistenza di autovalori puramente immaginari assicura che le eventuali soluzioni oscillatorie sono necessariamente *smorzate*.

Dal risultato precedente si può anche facilmente vedere che se ogni compartimento nel sistema ha una uscita verso l'ambiente esterno, allora lo zero non è un autovalore di  $\mathbf{A}$ , e quindi  $\mathbf{A}$  è non singolare. D'altra parte, se nessuno compartimento nel sistema ha una uscita, allora vi è almeno un autovalore nullo e di conseguenza  $\mathbf{A}$  è una matrice singolare.

Utilizzando i risultati sulle matrici *non negative*, in particolare il teorema di Perron-Frobenius (cfr. Appendice A), si possono dimostrare i seguenti risultati.

**Teorema 12.1** *Per ogni matrice compartimentale  $\mathbf{A}$  di ordine  $n$ , con autovalori ordinati come in (12.47), l'autovalore  $\lambda_n$  è reale e non positivo. A tale autovalore corrisponde un autovettore  $\mathbf{v} \geq 0$ . Per ogni altro autovalore  $\lambda_i$  si ha  $\Re(\lambda_i) \leq \lambda_n$ .*

**Teorema 12.2** *Se un sistema a  $n$  compartimenti è fortemente connesso, allora l'autovalore reale  $\lambda_n$  della corrispondente matrice  $\mathbf{A}$  è tale che*

(a) *a  $\lambda_n$  corrisponde un autovettore  $\mathbf{v} > 0$ ;*

- (b)  $\lambda_n$  è un autovalore semplice di  $\mathbf{A}$ ;
- (c) per ogni altro autovalore si ha  $\Re(\lambda_i) < \lambda_n$ ;
- (d)  $\lambda_n$  cresce all'aumentare di un qualsiasi elemento di  $\mathbf{A}$ .

Dal secondo teorema si ha, in particolare, che la soluzione del sistema che varia più lentamente corrisponde alla radice semplice reale  $\lambda_n$ .

Nell'analisi di un sistema a compartimenti può avere interesse conoscere quando anche gli altri autovalori della matrice  $\mathbf{A}$  sono *reali*, in quanto in tale caso il sistema non ha comportamenti oscillatori. La questione può essere affrontata utilizzando il seguente risultato relativo alle matrici.

**Proposizione 12.2** *Se per una matrice  $\mathbf{A}$  di ordine  $n$ , esiste una matrice simmetrica definita positiva  $\mathbf{Q}$  tale che  $\mathbf{QA}$  sia simmetrica, allora, gli autovalori di  $\mathbf{A}$  sono reali. Inoltre, gli autovettori corrispondenti a distinti autovalori di  $\mathbf{A}$  sono tra loro ortogonali rispetto al prodotto scalare  $(\mathbf{x}, \mathbf{y})_{\mathbf{Q}} := (\mathbf{x}, \mathbf{Q}\mathbf{y})$ .*

Una matrice che verifica le condizioni del risultato precedente viene detta *matrice simmetrizzabile*. Si possono allora dimostrare facilmente i seguenti risultati.

**Teorema 12.3** *Un sistema a  $n$  compartimenti a catena con compartimenti ordinati in maniera tale che la matrice compartmentale  $\mathbf{A}$  verifichi la condizione  $a_{i,i-1} \neq 0$ , per  $i = 2, 3, \dots, n$ , è simmetrizzabile.*

Nel caso in cui si abbia  $a_{i,i+1} \neq 0$ ,  $a_{i+1,i} \neq 0$ , per  $i = 1, \dots, n$ , si può anche dimostrare che gli autovalori sono distinti.

**Teorema 12.4** *Un sistema a compartimenti di tipo mammillare è simmetrizzabile.*

Più in generale, si può dimostrare il seguente risultato.

**Teorema 12.5** *Se il ciclo<sup>7</sup> di lunghezza massima nel diagramma di connettività corrispondente alla matrice  $\mathbf{A}$  è di lunghezza due, allora ogni autovalore di  $\mathbf{A}$  è reale.*

Come esempio di applicazione, si consideri la matrice

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

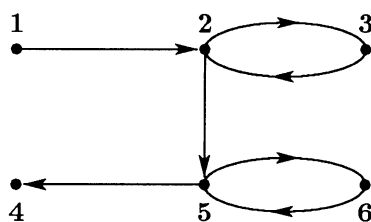


Figura 12.17: Un diagramma di connettività nel quale non esistono cicli di lunghezza superiore a due.

a cui è associato il diagramma illustrato in Figura 12.17. Per tale matrice si hanno gli autovalori semplici  $0, -1$  gli autovalori doppi  $-2.618, -0.382$ .

Consideriamo, ora, il caso in cui nel modello  $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}$  il vettore input  $\mathbf{b}$  è costante e  $\mathbf{A}$  è una matrice invertibile (e, quindi, necessariamente il sistema è aperto). Dalla teoria sui sistemi differenziali lineari si può ricavare l'espressione della soluzione generale in termini degli autovettori e autovalori della matrice  $\mathbf{A}$ . Più precisamente, se lo spazio generato dagli autovalori  $\{\mathbf{v}_i\}$  ha dimensione  $n$ , si ha

$$\mathbf{x}(t) = \sum_{k=1}^n c_k e^{\lambda_k t} \mathbf{v}_k - \mathbf{A}^{-1} \mathbf{b} \quad (12.48)$$

ove  $c_k$  rappresentano delle costanti arbitrarie e, come abbiamo visto in precedenza,  $\Re(\lambda_k) < 0$  per ogni  $k$ . Dalla (12.48) si vede, quindi, che

$$\mathbf{x}(t) \rightarrow \mathbf{x}_e := -\mathbf{A}^{-1} \mathbf{b}, \quad \text{per } t \rightarrow \infty$$

per ogni valore iniziale  $\mathbf{x}(0)$  del tracciante. La convergenza di  $\mathbf{x}(t)$  a  $\mathbf{x}_e$  per  $t \rightarrow \infty$  dice che il modello è *asintoticamente stabile*. La rapidità di convergenza di  $\mathbf{x}(t)$  a  $\mathbf{x}_e$  è, per  $t$  sufficientemente grande, dipendente dall'autovalore di modulo minimo  $\lambda_n$ . Il vettore  $\mathbf{x}_e$  è un punto di *equilibrio* del modello, dal momento che  $\dot{\mathbf{x}}_e = 0$ . Quando  $\mathbf{A}$  è non singolare, vi è un *unico* punto di equilibrio  $\mathbf{x}_e = -\mathbf{A}^{-1} \mathbf{b}$  per il sistema, mentre nel caso generale l'esistenza dei punti di equilibrio dipende dalla risolubilità del sistema  $\mathbf{A}\mathbf{x} = -\mathbf{b}$ .

Per quanto riguarda la invertibilità della matrice  $\mathbf{A}$ , ricordiamo, senza dimostrazione, il seguente risultato.

**Teorema 12.6** *Sia  $S$  un sistema aperto a  $n$  compartimenti. La corrispondente matrice compartimentale  $\mathbf{A}$  è invertibile se e solo se  $S$  non contiene sottoinsiemi di tipo trappola. Nel caso in cui sia invertibile, si ha  $\mathbf{A}^{-1} \leq \mathbf{0}$ , ossia  $-\mathbf{A}$  è una  $M$ -matrice.*

<sup>7</sup>Ricordiamo che un ciclo di lunghezza  $k$  in un grafo  $S$  è un insieme ordinato di  $k$  ( $\leq n$ ) indici distinti  $\{j_1, j_2, \dots, j_k\}$  tali che gli archi orientati  $(j_1 \vec{j}_2, \dots, j_k \vec{j}_1)$  sono contenuti nel grafo  $S$ .



▼ **Osservazione 12.1** Per la risoluzione numerica del sistema lineare algebrico  $\mathbf{A}\mathbf{x}_e = -\mathbf{b}$  è opportuno tenere conto che, quando il grado di connettività del sistema è piccolo, la matrice  $\mathbf{A}$  è sparsa. Possono, quindi, essere di interesse, in particolare, i metodi iterativi. Osserviamo, inoltre, che la matrice  $\mathbf{A}$  può essere estremamente malcondizionata. In questo caso piccoli cambiamenti negli elementi di  $\mathbf{A}$  (dovuti usualmente a errori sperimentali) possono causare grossi errori negli elementi dell'inversa della matrice e, quindi, nelle componenti del vettore  $\mathbf{x}_e$ .

Come illustrazione, consideriamo la matrice  $\mathbf{A}$  corrispondente al sistema a compartimenti ad una sola uscita illustrata in Figura 12.18.

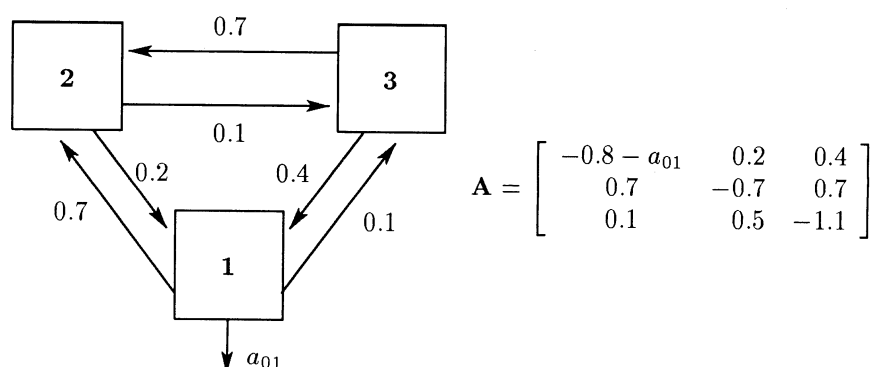


Figura 12.18: Esempio di sistema a compartimenti mal condizionato.

Con  $a_{01} = 5.0 \times 10^{-3}$ , la matrice inversa  $\mathbf{A}^{-1}$  è data da

$$\mathbf{A}^{-1} = \begin{bmatrix} -200.0000 & -200.0000 & -200.0000 \\ -400.0000 & -402.6190 & -401.6667 \\ -200.0000 & -201.1905 & -201.6667 \end{bmatrix}$$

Supponendo, ora, che  $a_{01}$  abbia un errore di misurazione di  $1.0 \times 10^{-3}$ , sicché  $a_{01}$  sia in realtà  $4.0 \times 10^{-3}$ . L'elemento  $a_{11}$  della matrice  $\mathbf{A}$  passa allora dal valore  $-0.8050$  al valore  $-0.8040$ , mentre la matrice inversa diventa

$$\mathbf{A}^{-1} = \begin{bmatrix} -250.0000 & -250.0000 & -250.0000 \\ -500.0000 & -502.6190 & -501.6667 \\ -250.0000 & -251.1905 & -251.6667 \end{bmatrix}$$

Si passa, quindi, da un errore del 0.12% nei dati a un errore del 25% nei risultati. Il numero di condizionamento della matrice  $\mathbf{A}$ , nella norma 2, è in effetti dato dal seguente valore

$$\mu_2(\mathbf{A}) = 1.6880 \times 10^3$$

mentre il determinante è dato da  $\det(\mathbf{A}) = -0.0017$ . D'altra parte, osserviamo che quando  $a_{01} \rightarrow 0$ , la matrice  $\mathbf{A}$  tende a una matrice singolare.

Lasciamo come esercizio la dimostrazione che per una matrice  $\mathbf{A}$  relativa ad un sistema a compartimenti con un'unica uscita nell'ambiente esterno dal compartimento  $m$ -mo, si ha  $\mu_\infty \geq n \|\mathbf{A}\|_\infty / a_{0m}$ . ■

Gli elementi  $\tau_{ij}$  della matrice  $\mathbf{T} = -\mathbf{A}^{-1}$  possono essere interpretati dal punto di vista fisico come *tempi medi di residenza*. Come illustrazione, consideriamo il caso scalare, illustrato in Figura 12.19 e corrispondente all'equazione  $\dot{x}(t) = a_{11}x(t)$ , con  $a_{11} = -a_{01} < 0$  e la condizione iniziale  $x(0) = x_0$ . Nella stessa figura è rappresentata la soluzione  $x(t) = x_0 \exp(a_{11}t)$ . La quantità  $\tau = \tau_{11} = -a_{11}^{-1}$  corrisponde al numero

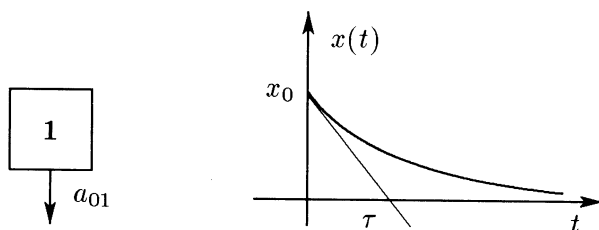


Figura 12.19: Un sistema a un compartimento aperto.

medio (valore aspettato) di unità di tempo impiegato da una particella scelta a caso per lasciare il compartimento. Per esempio, se scegliamo a caso nel compartimento 1000 particelle, e se per questo campione sono richiesti in media 4 minuti affinché una particella esca dal compartimento, allora  $-1/\tau = -1/4 = -0.25$  è un valore ragionevole per  $a_{11}$ . Osserviamo anche che

$$\tau = -a_{11}^{-1} = -\frac{x(0)}{\dot{x}(0)}$$

è l'intersezione con l'asse  $t$  della retta  $y = x(t) + x(0)$  (cfr. Figura 12.19). Si vede, quindi, che  $\tau$  è il tempo richiesto affinché la quantità iniziale  $x_0$  sia ridotta a zero, cioè sia completamente uscita dal compartimento (supponendo, naturalmente, che  $\dot{x}(t)$  rimanga costante al valore iniziale  $\dot{x}(0)$ ).

Più in generale, il significato degli elementi della matrice  $T$  è illustrato dalle seguenti considerazioni, che mostrano il legame tra la teoria dei compartimenti e la teoria delle *catene di Markov* (cfr. Capitolo 8). Supponiamo che  $\mathbf{A}$  sia una matrice compartimentale *aperta*, cioè una matrice per la quale vi è almeno un indice  $j$ , con  $j = 1, 2, \dots, n$ , tale che la costante di flusso

$$a_{0j} = -\sum_{i=1}^n a_{ij}$$

dal compartimento  $j$  all'ambiente esterno sia positiva. A partire da  $\mathbf{A}$  si può costruire una matrice  $\bar{\mathbf{A}}$ , detta la *chiusura* di  $\mathbf{A}$  nel seguente modo. L'ambiente esterno diventa un compartimento e il numero dei compartimenti presenti nel sistema diventa  $n + 1$ . La corrispondente matrice compartimentale  $\bar{\mathbf{A}}$ , di ordine  $n + 1$ , viene allora ottenuta aggiungendo a sinistra una colonna  $(n + 1) \times 1$  di zeri e in alto un

vettore  $1 \times n$ , il cui elemento  $j$ -mo è  $a_{0j}$ . Si può dimostrare facilmente che per  $h > 0$  sufficientemente piccolo la matrice  $\bar{\mathbf{Q}}(h) := (\mathbf{I} + h\bar{\mathbf{A}})$  è la matrice di transizione dello stato  $(n+1)$ -mo della catena di Markov  $\mathbf{x}(t+h)^T = \mathbf{x}(t)^T \bar{\mathbf{Q}}(h)$  ed ha la seguente forma canonica

$$\bar{\mathbf{Q}}(h) := \left[ \begin{array}{c|c} 1 & 0 \\ \hline \mathbf{R}(h) & \mathbf{Q}(h) \end{array} \right]$$

ove  $\mathbf{Q}(h) = (\mathbf{I} + h\mathbf{A})^T$ . Tale catena di Markov è assorbente, in quanto l'ambiente esterno è uno stato assorbente, ed è possibile passare da ognuno dei compartimenti iniziali all'esterno (infatti, essendo la matrice  $\mathbf{A}$  invertibile, non vi sono trappole). La matrice

$$\mathbf{M}(h) := [\mathbf{I} - \bar{\mathbf{Q}}(h)]^{-1} = [\mathbf{I} - h\mathbf{A}^T]^{-1} = \frac{1}{h} (-\mathbf{A}^{-1})^T$$

è la *matrice fondamentale* associata con  $\bar{\mathbf{Q}}(h)$ , e il suo generico elemento non negativo  $m_{ij}(h)$  è interpretato come il numero aspettato di volte per cui il processo di Markov è nel compartimento (stato)  $j$  se ha avuto inizio nel compartimento  $i$  ed è avanzato fino ad essere assorbito (cioè è uscito nell'ambiente esterno). Posto  $\mathbf{T} := -\mathbf{A}^{-1} = [\tau_{ij}]$ , si ha  $m_{ij} = \tau_{ij}/h$ , per  $i, j = 1, 2, \dots, n$ . Scelto, allora,  $h = 1$  si può concludere con il seguente risultato.

**Teorema 12.7** *Se  $S$  è un sistema a compartimenti aperto con nessuna trappola e  $\mathbf{A}$  è la matrice associata compartimentale, allora il generico elemento  $\tau_{ij}$  della matrice  $\mathbf{T} = -\mathbf{A}^{-1}$  è il tempo medio di residenza che una particella scelta a caso spende nel compartimento  $i$ , avendo cominciato nel compartimento  $j$  al tempo zero, prima di uscire dal sistema  $S$ .*

## 12.4 Identificazione del modello

Il problema centrale nello studio dei compartimenti, e più in generale dei modelli matematici, riguarda la possibilità di determinare in maniera univoca (cioè di *identificare*) i parametri  $a_{ij}$ , a partire dalle osservazioni sperimentali di alcune componenti del vettore soluzione  $\mathbf{x}$ .

Il problema è, in sostanza, l'*inverso* di quello esaminato nei paragrafi precedenti. In effetti, in tali paragrafi abbiamo analizzato il comportamento di un sistema, supposta nota la struttura, ossia il numero dei compartimenti, la loro connessione e i valori dei coefficienti frazionali di trasferimento. Nelle applicazioni, tuttavia, il problema si pone in maniera differente. In generale, infatti, la ricerca parte da un insieme di dati sperimentali che si riferiscono ad un problema reale (biologico, chimico, economico, ...). Tali dati possono suggerire un modello matematico a compartimenti. Lo scopo del ricercatore è, allora, quello di progettare opportuni esperimenti al fine di esaminare la natura del sistema, ossia di stimare il numero dei compartimenti, la loro connessione e i valori dei coefficienti frazionali di trasferimento. Si

tratta, ovviamente, di un problema la cui soluzione presenta, in generale, notevole difficoltà, anche a causa del numero limitato di osservazioni e della presenza di errori nelle misurazioni. Le principali tappe nel processo di costruzione di un modello a compartimenti sono schematizzate in Figura 12.20 e saranno ora discusse più in dettaglio.

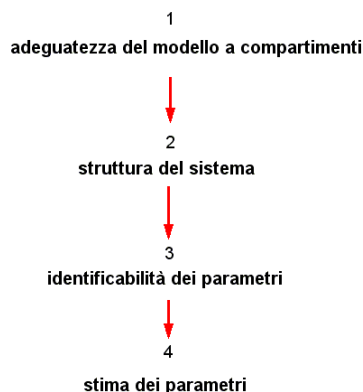


Figura 12.20: Successive questioni nella costruzione di un modello a compartimenti.

1. Per adeguatezza del modello si intende la possibilità di rappresentare il sistema reale mediante un sistema a compartimenti. La risoluzione della questione si basa su una approfondita conoscenza del sistema reale e degli scopi della ricerca.
2. La struttura del sistema a compartimenti riguarda il numero dei compartimenti e delle connessioni, ossia il diagramma di connettività. Possono, naturalmente, esistere differenti modelli, e una fase importante del problema è la individuazione del modello “migliore”.
3. Una volta fissata la struttura del modello, rimane la determinazione dei coefficienti frazionali di trasferimento. Tali coefficienti possono essere identificati quando tutti i compartimenti sono osservabili sperimentalmente (se si prescinde dagli *errori sperimentali*). Si può, infatti, vedere che in tale caso si ha un numero di condizioni pari al numero delle incognite.

Tuttavia, per i sistemi reali non è in generale possibile accedere a tutti i compartimenti. Si pone, allora, il problema di vedere se i coefficienti di trasferimento possono essere identificati a partire dalle misurazioni relative ad un opportuno *sottoinsieme* di compartimenti del sistema. Tale problema, di natura squisitamente teorica, è noto come problema di *identificabilità strutturale*.

In maniera schematica, un parametro è *non identificabile*, quando un numero infinito di valori del parametro è compatibile con i risultati di un esperimento. In corrispondenza, un modello è detto non identificabile, quando esistono uno o più parametri non identificabili. In questo caso, sono necessarie altre osservazioni sperimentali o ulteriori vincoli sui parametri.

4. Supponendo di avere risolto la questione della identificabilità dei parametri e, quindi, di conoscere quali compartimenti è necessario osservare sperimentalmente per individuare i parametri richiesti, si pone il problema *numerico-statistico* di fornire una stima dei parametri.

Più in generale, possiamo considerare il problema della identificazione in riferimento al seguente modello

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & t \geq 0 \\ \mathbf{x}(0) = \mathbf{0} \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \end{cases} \quad (12.49)$$

ove le componenti del vettore  $\mathbf{u} = [u_1(t), u_2(t), \dots, u_q(t)]^T$  costituiscono le *funzioni input* del sistema<sup>8</sup> e la matrice  $\mathbf{B}$  di dimensioni  $n \times q$  è la *matrice input di distribuzione*, in quanto essa indica come gli input  $u_i(t)$  sono distribuiti attraverso il sistema: l'elemento  $b_{ik}$  è positivo se l'input  $u_k(t)$  entra nel compartimento  $i$ , e zero altrimenti. La matrice  $\mathbf{C}$ , di dimensione  $p \times n$  è chiamata la *matrice output di connessione*, o di osservazione, e indica quali componenti, o combinazioni di componenti, sono osservate sperimentalmente. In generale, le quantità  $\mathbf{y}(t)$  sono misurate in un insieme di punti discreti  $\{t_1, t_2, \dots, t_m\}$ . Utilizzando, allora, come stimatore quello fornito dal metodo dei minimi quadrati, il problema della stima degli elementi della matrice  $\mathbf{A}$  si trasforma nella ricerca del minimo della seguente funzione

$$S := \sum_{i=1}^m \|y(t_i) - y_i^*\|^2$$

ove  $y_i^*$  sono i valori rilevati sperimentalmente in corrispondenza ai tempi discreti  $t_i$ . Sottolineiamo il fatto che per la costruzione del valore della funzione  $S$  è necessaria la risoluzione (effettuata, in generale, mediante metodi numerici) del sistema differenziale (12.49). La procedura è stata illustrata su un caso particolare nell'Esempio 12.2.

Per un ampliamento e un approfondimento delle precedenti questioni, in particolare per quanto riguarda la identificabilità del sistema, rinviamo in particolare a Anderson [5] e Jacquez [92].

<sup>8</sup>Nelle applicazioni, quando il modello è stato identificato, tali funzioni possono assumere il ruolo di funzioni *controllo*, cfr. Capitolo 14.

Tali problematiche sono connesse, più in generale, alla *controllabilità* del sistema (ossia alla possibilità di ricostruire ogni stato del sistema a partire dagli output), e alla *raggiungibilità* (reachability, ossia alla possibilità di trasferire il sistema da un particolare stato corrente, a un prefissato stato entro un certo intervallo di tempo e utilizzando un determinato insieme di input).

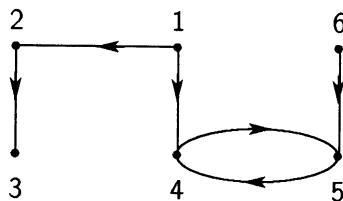


Figura 12.21: Diagramma di connettività.

◆ **Esercizio 12.1** In riferimento al diagramma di connettività che è rappresentato nella Figura 12.21, si supponga che il tracciante sia iniettato nel compartimento 1 e che venga osservato il compartimento 5. Analizzare la identificabilità del corrispondente sistema a compartimenti.

◆ **Esercizio 12.2** Analizzare la identificabilità di un sistema a tre compartimenti di tipo mammillare con un'unica uscita situata nel compartimento centrale.

... who can tell  
Which of her forms has shown her substance right?  
William Butler Yeats

## Capitolo 13

# Identificazione dei parametri nelle equazioni differenziali

La *identificazione dei parametri nei modelli dinamici* rappresenta uno dei problemi più importanti nella costruzione e nell'utilizzo dei modelli matematici. In effetti, il problema è già stato introdotto e discusso in altre parti del testo; si vedano in particolare i Capitoli 7 e 12 per opportune esemplificazioni e il Capitolo 8 per l'inquadramento del problema nell'ambito della statistica. Lo scopo di questo capitolo è quello di esporre *in maniera più organica* i risultati e le idee di base. Come riferimento, per un opportuno approfondimento, segnaliamo in particolare Bard [10].

### 13.1 Formulazione del problema

Dal punto di vista matematico, il problema si presenta nella seguente forma. È dato un problema a valori iniziali per un sistema di equazioni differenziali che modella la dinamica di un processo deterministico (*sistema dinamico*)

$$\begin{cases} \frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t), \mathbf{a}, t) & \text{modello matematico} \\ \mathbf{x}(t_0) = \mathbf{x}^{(0)} & \text{di un sistema dinamico} \end{cases} \quad (13.1)$$

ove  $\mathbf{a} \in \mathbb{R}^p$  rappresenta il *vettore dei parametri*. Con  $\mathbf{a}^*$  si indica il valore di  $\mathbf{a}$  da calcolare; per ipotesi,  $\mathbf{a}^*$  è un vettore a priori incognito, da stimare mediante opportune osservazioni sulle *variabili di stato*  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ . Nelle applicazioni, tuttavia, i valori delle variabili  $x_i$  non sono di solito direttamente osservabili; si suppone, invece, che siano disponibili le seguenti quantità, dette le *variabili osservate*

$$(y_j)_r = h_j[\mathbf{x}(t_r), \boldsymbol{\zeta}_r], \quad j = 1, 2, \dots, m; \quad r = 1, 2, \dots, R \quad (13.2)$$

ove  $\zeta_r$  rappresenta un vettore di variabili casuali che possono influenzare i risultati delle misurazioni sperimentali effettuate nei successivi istanti  $t_r$ . Si suppone che le funzioni  $\mathbf{f} = [f_1, f_2, \dots, f_n]^T$  e  $\mathbf{h} = [h_1, h_2, \dots, h_m]^T$  siano di forma nota, insieme con le proprietà statistiche di  $\zeta_r$ . Alcune delle componenti del *vettore iniziale*  $\mathbf{x}^{(0)} = [x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}]^T$  possono essere incognite a priori e corrispondere, quindi, ad ulteriori parametri da stimare; nel seguito, tuttavia, considereremo, per semplicità, il caso in cui  $\mathbf{x}^{(0)}$  sia un vettore assegnato. I risultati ottenuti si estendono, comunque, facilmente al caso generale.

In conclusione, il problema della identificazione consiste nel ricavare una stima  $\bar{\mathbf{a}}$  del vettore dei parametri  $\mathbf{a}^*$  dalle osservazioni  $y_j$  effettuate agli istanti  $t_r$ . Un esperimento (*experiment*), relativo al sistema dinamico (13.1), consiste nella valutazione delle variabili  $\mathbf{y}$  in corrispondenza a un valore fissato di  $t$ . Un gruppo di esperimenti eseguiti con identiche condizioni iniziali, ma corrispondenti a valori diversi di  $t$ , viene detta una esecuzione (*run*).

► **Esempio 13.1** Come illustrazione del concetto di sistema dinamico, consideriamo il seguente modello di reazione chimica corrispondente a tre sostanze le cui concentrazioni, indicate con  $c_1, c_2, c_3$ , soddisfano le seguenti equazioni differenziali

$$\begin{cases} dc_1/dt = -k_1c_1^2 + k_2c_2c_3 \\ dc_2/dt = k_1c_1^2 - k_2c_2c_3 - k_3c_2 \\ dc_3/dt = k_1c_1^2 - k_2c_2c_3 + k_3c_2 \end{cases} \quad (13.3)$$

Le concentrazioni iniziali  $c_2$  e  $c_3$  non sono note esattamente, ma, supponendo che la loro somma sia uguale a uno, si ha

$$c_1(0) = \alpha, \quad c_2(0) = \beta, \quad c_3(0) = 1 - \alpha - \beta \quad (13.4)$$

ove  $\alpha$  è una quantità nota, mentre  $\beta$  è una quantità da stimare.

Al generico istante  $t_r$  si esaminano due campioni della miscela. Nel primo si determina  $c_1$  direttamente per titolazione, mentre il secondo è analizzato mediante uno strumento ottico, che misura il coefficiente di assorbimento della luce da parte della miscela. Supponendo che tale coefficiente sia una funzione lineare delle concentrazioni  $c_1$  e  $c_2$ , le variabili osservate  $\mathbf{y}$  corrispondono alle seguenti definizioni

$$(y_1)_r = c_1(t_r) + (\zeta_1)_r, \quad (y_2)_r = d + pc_1(t_r) + qc_2(t_r) + (\zeta_2)_r$$

In questo modello,  $c_1, c_2, c_3$  sono le variabili di stato;  $y_1, y_2$  sono le variabili osservate;  $\alpha$  e  $t$  sono le variabili indipendenti;  $\beta, k_1, k_2, k_3, p, q$  e  $r$  sono i parametri incogniti. In particolare, i parametri di interesse possono essere le costanti di velocità  $k_1, k_2$  e  $k_3$ . Una buona stima per  $\beta$  può essere ricavata dal modo con il quale si è preparata la soluzione e i parametri  $d, p$  e  $q$ , che caratterizzano lo strumento ottico, possono essere preventivamente identificati.

Altri esempi significativi di modelli dinamici sono stati esaminati nel precedente Capitolo 12, al quale rinviamo anche per una discussione più approfondita sulla collocazione del problema di identificazione nell'ambito più generale della costruzione e validazione di un modello matematico. ■



### 13.1.1 Difficoltà nella identificazione

Il calcolo di una stima  $\bar{\mathbf{a}}$  del vettore dei parametri  $\mathbf{x}^*$  dalle osservazioni delle variabili  $y_j$  nei tempi  $t_r$  può essere effettuato nel seguente modo. In corrispondenza ad un vettore fissato  $\mathbf{a}$  dei parametri, la soluzione  $\mathbf{x}(t, \mathbf{a})$  del problema a valori iniziali (13.1) può essere approssimata mediante i metodi numerici che sono stati analizzati nel Capitolo 7. In questo modo, ad ogni tempo  $t_r$  è possibile confrontare i valori calcolati  $g_j[\mathbf{x}(\mathbf{a}, t_r)] := h_j[\mathbf{x}(\mathbf{a}, t_r), 0]$  con i valori misurati  $(y_j)_r$ . Si introduce quindi una opportuna distanza  $F(\mathbf{a})$  tra i valori calcolati e i valori misurati. Nell'ambito del *metodo dei minimi quadrati* (equivalente al criterio della massima verosimiglianza nel caso in cui gli errori sono distribuiti normalmente) la funzione  $F(\mathbf{a})$  ha la seguente forma

$$F(\mathbf{a}) := \sum_{r=1}^R (\mathbf{y}(t_r) - \mathbf{g}[\mathbf{x}(\mathbf{a}, t_r)])^T \mathbf{W}_r (\mathbf{y}(t_r) - \mathbf{g}[\mathbf{x}(\mathbf{a}, t_r)]) \quad (13.5)$$

ove  $\mathbf{W}_r$  sono opportune matrici definite positive di dimensioni  $m \times m$ , la cui scelta ottimale sarà discussa nel seguito. In questo modo il problema della stima dei parametri è ricondotto al problema numerico del calcolo del minimo di una funzione non lineare nelle variabili  $\mathbf{a}$ , che può essere risolto mediante i metodi analizzati nel Capitolo 5 (cfr. in particolare il Paragrafo 5.5.4 riguardante i minimi quadrati non lineari). Il valore  $\bar{\mathbf{a}}$  corrispondente al minimo di  $F(\mathbf{a})$  è una stima del vero valore (incognito)  $\mathbf{a}^*$ .

La precedente procedura pone, tuttavia, diverse questioni di estrema importanza per un corretto utilizzo delle stime ottenute<sup>1</sup>. A parte, infatti, le difficoltà analitiche inerenti all'esistenza di un unico minimo della funzione di  $F(\mathbf{a})$  e alle difficoltà numeriche per il suo calcolo, si pone la questione delicata della *affidabilità (reliability)*, sia del modello (13.1) utilizzato (*adeguatezza* del modello) che della stima  $\bar{\mathbf{a}}$  ottenuta. In questo capitolo analizzeremo in particolare quest'ultima questione, ossia cercheremo di associare ad ogni stima  $\bar{\mathbf{a}}$  dei parametri una opportuna *regione di confidenza*, che rappresenti una misura, in termini statistici, della affidabilità della stima ottenuta. Un aspetto importante, che sarà messo in evidenza da tale analisi, è che il grado di affidabilità di una stima è determinato da più fattori. In primo luogo, naturalmente, vi sono le proprietà statistiche degli errori relativi alle misurazioni sperimentali e la scelta dello stimatore (13.5). Ugualmente importante, tuttavia, è il *condizionamento* del modello (13.1), che caratterizza la dipendenza della soluzione  $\mathbf{x}(t, \mathbf{a})$  dal vettore dei parametri  $\mathbf{a}$ . Ricordiamo che per un problema malcondizionato a piccoli errori relativi nei dati (nel nostro caso le variabili  $x_i$ , o più precisamente le osservate  $y_i$ ) possono corrispondere grandi variazioni relative nei risultati (ossia nella stima cercata  $\bar{\mathbf{a}}$ ). A questo proposito, richiamiamo l'attenzione sull'Esempio 1.7 analizzato nel Capitolo 1 che mette in rilievo come sia possibile, in determinati modelli e in

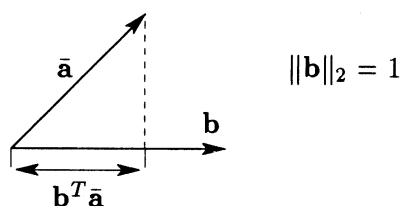
<sup>1</sup>It is not enough to compute a vector  $\bar{\mathbf{a}}$  and to state that this is the estimate value of the unknown parameters  $\mathbf{a}^*$ . We must also investigate the reliability and precision of our estimates, Y. Bard.

relazione a un determinato insieme di osservazioni sperimentali, ottenere dei fitting ugualmente buoni con valori decisamente differenti dei parametri.

In maniera schematica, possiamo riassumere le considerazioni precedenti nel seguente modo. Un buon fitting, ossia una buona coincidenza tra i valori osservati sperimentalmente e quelli calcolati, non è, da solo, una condizione sufficiente per ritenere affidabile un modello e i parametri ottenuti mediante tale modello, ma la loro affidabilità è legata a diversi fattori; in particolare, al condizionamento del modello, alla scelta dello stimatore, alle caratteristiche statistiche dei dati. Inoltre, dal momento che il condizionamento del modello dipende anche dalla scelta dei dati sperimentali, un altro fattore importante nella determinazione dell'affidabilità dei risultati ottenuti è la *pianificazione (design)* degli esperimenti, per la quale utili suggerimenti possono venire dall'analisi della regione di confidenza.

### 13.2 Intervalli di confidenza per i parametri

Indicata con  $\bar{\mathbf{a}}$  una stima del vero valore  $\mathbf{a}^*$ , ci proponiamo di associare un intervallo di confidenza  $\gamma(\mathbf{b})$  con la proiezione di  $\bar{\mathbf{a}}$  su un qualsiasi vettore  $\mathbf{b}$  normalizzato a uno (ossia tale che  $\|\mathbf{b}\|_2 = 1$ ). La nozione di intervallo di confidenza è stata introdotta



e discussa nel Capitolo 8. Ricordiamone brevemente il significato. Si considera la stima  $\mathbf{a}$  come una variabile casuale corrispondente ad una serie di esperimenti differenti. Dire che l'intervallo

$$[\mathbf{b}^T \bar{\mathbf{a}} - \gamma(\mathbf{b}), \mathbf{b}^T \bar{\mathbf{a}} + \gamma(\mathbf{b})] \quad (13.6)$$

è un intervallo di confidenza, ad esempio 95%, significa che la probabilità che il vero valore  $\mathbf{a}^*$  si trovi in tale intervallo è 0.95. In forma più precisa, si supponga di calcolare le stime  $\bar{\mathbf{a}}$  corrispondenti a un numero elevato, diciamo cento, di esperimenti. Si hanno pertanto cento intervalli del tipo  $[\mathbf{b}^T \bar{\mathbf{a}} - \gamma(\mathbf{b}), \mathbf{b}^T \bar{\mathbf{a}} + \gamma(\mathbf{b})]$ . Allora il vero valore  $\mathbf{a}^*$  dovrebbe essere contenuto in circa 95 di tali intervalli.

La scelta di un intervallo per un determinato livello di confidenza è in un certo senso arbitraria. La definizione (13.6) permette di avere informazioni su opportune combinazioni lineari degli elementi di  $\bar{\mathbf{a}}$  su  $\mathbf{b}$ . Osserviamo che, scegliendo in particolare  $\mathbf{b} = \mathbf{e}_i$ , ove  $\mathbf{e}_i$  è il vettore di componenti tutte nulle salvo la componente

$i$ -ma, l'intervallo di confidenza  $\gamma(\mathbf{b})$  diventa semplicemente l'intervallo di confidenza dell'elemento  $\bar{a}_i$  di  $\bar{\mathbf{a}}$ .

Valutazioni di  $\gamma(\mathbf{b})$ , che siano utili nelle applicazioni, possono essere ottenute nell'ipotesi che  $\bar{\mathbf{a}}$  sia una buona approssimazione di  $\mathbf{a}^*$  ossia che si abbia  $\bar{\mathbf{a}} = \mathbf{a}^* + \boldsymbol{\alpha}$ , con  $\boldsymbol{\alpha}$  vettore sufficientemente piccolo. In tale ipotesi, infatti, è possibile *linearizzare* il problema e utilizzare quindi i risultati che abbiamo visto nel Capitolo 8 in relazione ai modelli lineari.

### 13.2.1 Equazioni di sensitività

La difficoltà maggiore nell'applicazione dell'idea della linearizzazione è relativa al calcolo delle derivate delle variabili di stato  $\mathbf{x}(\mathbf{a}, t)$  (soluzioni del sistema (13.1)), rispetto ai parametri  $\mathbf{a}$ , ossia degli elementi della seguente matrice di ordine  $n \times p$

$$\mathbf{D}(t) = \begin{bmatrix} \frac{\partial x_1(\mathbf{a}, t)}{\partial a_1} & \frac{\partial x_1(\mathbf{a}, t)}{\partial a_2} & \dots & \frac{\partial x_1(\mathbf{a}, t)}{\partial a_p} \\ \frac{\partial x_2(\mathbf{a}, t)}{\partial a_1} & \frac{\partial x_2(\mathbf{a}, t)}{\partial a_2} & \dots & \frac{\partial x_2(\mathbf{a}, t)}{\partial a_p} \\ \dots & \dots & \dots & \dots \\ \frac{\partial x_n(\mathbf{a}, t)}{\partial a_1} & \frac{\partial x_n(\mathbf{a}, t)}{\partial a_2} & \dots & \frac{\partial x_n(\mathbf{a}, t)}{\partial a_p} \end{bmatrix} \quad (13.7)$$

La matrice  $\mathbf{D}(t)$ , chiamata anche *matrice di sensitività*, in quanto i suoi elementi forniscono in ogni istante  $t$  il grado di dipendenza di ogni variabile di stato  $x_i(t)$  dal parametro  $a_j$ , può essere calcolata risolvendo un opportuno sistema differenziale lineare. Più precisamente, supponendo che la funzione  $\mathbf{f}(\mathbf{x}, \mathbf{a}, t)$  in (13.1) sia sufficientemente regolare si può dimostrare che gli elementi della matrice  $\mathbf{D}(t)$  sono le soluzioni del seguente sistema *lineare* di  $n \times p$  equazioni differenziali, chiamato *sistema di sensitività*

$$\frac{d}{dt} \left( \frac{\partial x_i(\mathbf{a}, t)}{\partial a_j} \right) = \sum_{s=1}^n \frac{\partial}{\partial x_s} f_i(\mathbf{x}, \bar{\mathbf{a}}, t) \frac{\partial x_s(\mathbf{a}, t)}{\partial a_j} + \frac{\partial}{\partial a_j} f_i(\mathbf{x}, \bar{\mathbf{a}}, t) \quad (13.8)$$

per  $i = 1, 2, \dots, n$  e  $j = 1, 2, \dots, p$ , con condizioni iniziali

$$\frac{\partial x_i(\mathbf{a}, t_0)}{\partial a_j} = 0 \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, p \quad (13.9)$$

Osserviamo che le condizioni iniziali sono nulle per il fatto che si è supposto lo stato iniziale  $\mathbf{x}^{(0)}$  noto a priori. Introdotte le seguenti matrici

$$\mathbf{A} = [A_{ij}] = \left[ \frac{\partial}{\partial x_j} f_i(\mathbf{x}, \bar{\mathbf{a}}, t) \right] \quad i, j = 1, 2, \dots, n$$

$$\mathbf{B} = [B_{ij}] = \left[ \frac{\partial}{\partial a_j} f_i(\mathbf{x}, \bar{\mathbf{a}}, t) \right] \quad i = 1, 2, \dots, n, j = 1, 2, \dots, p$$

il sistema (13.8) può essere scritto nella seguente forma matriciale

$$\begin{cases} \frac{d\mathbf{D}(t)}{dt} = \mathbf{A}(t)\mathbf{D}(t) + \mathbf{B}(t) \\ \mathbf{D}(t_0) = 0 \end{cases} \quad (13.10)$$

Osserviamo che la risoluzione del sistema (13.10) richiede la conoscenza delle variabili di stato  $\mathbf{x}$  e quindi la risoluzione del sistema di stato (13.1). Pertanto, il calcolo degli elementi di  $\mathbf{D}$  richiede complessivamente la risoluzione di un sistema di  $n + np = n(p+1)$  equazioni differenziali.

► **Esempio 13.2** Come semplice illustrazione, consideriamo il seguente problema a valori iniziali

$$\frac{dx}{dt} = -ax, \quad x(0) = 1 \quad (13.11)$$

che ha come soluzione la funzione  $x = e^{-at}$ , da cui direttamente  $\partial x / \partial a = -te^{-at}$ . In questo caso il sistema di sensitività si riduce all'equazione

$$\frac{dv}{dt} \left( \frac{\partial x}{\partial a} \right) = -av - x \quad (13.12)$$

ove, per brevità, si è posto  $v = \partial x / \partial a$ . Tenendo conto dell'espressione della soluzione  $x$ , si ha

$$\frac{dv}{dt} \left( \frac{\partial x}{\partial a} \right) = -av - e^{-at}, \quad v(0) = 0$$

che ha come soluzione  $v = -ae^{-at}$ , in accordo con il risultato ottenuto direttamente. ■

### 13.2.2 Problema linearizzato

L'analisi che svilupperemo nel seguente paragrafo è basata sull'ipotesi che  $\bar{\mathbf{a}}$  sia una buona approssimazione di  $\mathbf{a}^*$  ed inoltre che gli errori  $\zeta_r$  relativi alle valutazioni sperimentali siano sufficientemente piccoli, in maniera che per l'equazione (13.2) sia accettabile la seguente approssimazione del primo ordine

$$(y_j)_r = h_j[\mathbf{x}(t_r), 0] + \sum_k (\zeta_k)_r \frac{\partial h_j[\mathbf{x}(t_r), 0]}{\partial \zeta_k} = g_j[\mathbf{x}(t_r)] + (\eta_j)_r, \quad j = 1, 2, \dots, m$$

Gli errori  $\boldsymbol{\eta}_r$  corrispondenti a differenti insiemi di misure sperimentali sono supposti statisticamente indipendenti, ossia tali che

$$E(\boldsymbol{\eta}_r \boldsymbol{\eta}_s^T) = 0, \quad r \neq s \quad (13.13)$$

mentre non sono supposti necessariamente indipendenti gli elementi di ogni vettore  $\boldsymbol{\eta}_r$ ; la corrispondente matrice di covarianza, di dimensioni  $m \times m$ , sarà indicata con  $\mathbf{M}_r$

$$\mathbf{M}_r := E(\boldsymbol{\eta}_r \boldsymbol{\eta}_r^T) \quad (13.14)$$

Nell'ipotesi di una distribuzione Gaussiana con media zero, la densità di probabilità è data da (cfr. Capitolo 8)

$$P(\boldsymbol{\eta}_r) = \frac{1}{(2\pi)^{m/2} \sqrt{|\det(\mathbf{M})|}} \exp\left\{-\frac{1}{2} \boldsymbol{\eta}_r^T \mathbf{M}_r^{-1} \boldsymbol{\eta}_r\right\}$$

Considerando ora la funzione  $F(\mathbf{a}^* + \boldsymbol{\alpha})$  definita dall'equazione (13.5), con  $\boldsymbol{\alpha}$  tale che  $\mathbf{a}^* + \boldsymbol{\zeta} = \bar{\mathbf{a}}$ , per linearizzazione si ha, per  $i = 1, 2, \dots, p$

$$\frac{\partial}{\partial \alpha_i} \sum_{r=1}^R \left\{ (\mathbf{y}(t_r) - \mathbf{g}[\mathbf{x}(\mathbf{a}^* + \boldsymbol{\alpha}), t_r])^T \mathbf{W}_r (\mathbf{y}(t_r) - \mathbf{g}[\mathbf{x}(\mathbf{a}^* + \boldsymbol{\alpha}), t_r]) \right\} = 0$$

da cui

$$\frac{\partial}{\partial \alpha_i} \sum_{r=1}^R \left\{ (\mathbf{g}_r + \boldsymbol{\eta}_r - \mathbf{g}_r - \mathbf{G}_r \mathbf{D}_r \boldsymbol{\alpha})^T \mathbf{W}_r (\mathbf{g}_r + \boldsymbol{\eta}_r - \mathbf{g}_r - \mathbf{G}_r \mathbf{D}_r \boldsymbol{\alpha}) \right\} = 0 \quad (13.15)$$

ove si è posto

$$\mathbf{g}_r = \mathbf{g}[\mathbf{x}(\mathbf{a}^*, t_r)], \quad \mathbf{G}_r = \left[ \frac{\partial}{\partial x_j} g_i[\mathbf{x}(\mathbf{a}^*, t_r)] \right], \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n$$

e  $\mathbf{D}_r$  è la matrice di sensitività definita nel paragrafo precedente. Ricordiamo che la matrice  $\mathbf{D}_r$  ha dimensioni  $n \times p$ , mentre  $\mathbf{G}_r$  è una matrice  $m \times n$ . Dall'equazione (13.15) si ricava

$$\mathbf{H} \boldsymbol{\alpha} = \sum_{r=1}^R \mathbf{D}_r^T \mathbf{G}_r^T \mathbf{W}_r \boldsymbol{\eta}_r \quad \text{ove} \quad \mathbf{H} = \sum_{r=1}^R \mathbf{D}_r^T \mathbf{G}_r^T \mathbf{W}_r \mathbf{G}_r \mathbf{D}_r \quad (13.16)$$

Se la matrice  $\mathbf{H}$  è non singolare, l'equazione (13.16) determina l'errore sui parametri  $\boldsymbol{\alpha}$  in funzione degli errori sui dati  $\boldsymbol{\eta}_r$ . Si vede quindi che il *valore medio* di  $\boldsymbol{\alpha}$  è zero, mentre la *matrice di covarianza* degli elementi di  $\boldsymbol{\alpha}$ , ossia la matrice

$$\mathbf{V} = E(\boldsymbol{\alpha} \boldsymbol{\alpha}^T) \quad (13.17)$$

si ricava dall'equazione (13.16)

$$E(\mathbf{H}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{H}) = E\left(\sum_{r=1}^R \sum_{s=1}^R \mathbf{D}_r^T \mathbf{G}_r^T \mathbf{W}_r \boldsymbol{\eta}_r \boldsymbol{\eta}_s^T \mathbf{W}_s^T \mathbf{G}_s \mathbf{D}_s\right) \quad (13.18)$$

Tenendo conto, infine, delle ipotesi (13.13), (13.14), si ottiene la seguente equazione

$$\mathbf{H}\mathbf{V}\mathbf{H} = \sum_{r=1}^R \mathbf{D}_r^T \mathbf{G}_r^T \mathbf{W}_r \mathbf{M}_r \mathbf{W}_r \mathbf{G}_r \mathbf{D}_r \quad (13.19)$$

che determina  $\mathbf{V}$  se  $\mathbf{H}$  è non singolare.

► **Esempio 13.2** (*continuazione*) Per illustrare e interpretare il risultato (13.19), consideriamo in relazione al sistema dinamico (13.11) il caso semplice in cui  $R = 1$ ,  $y = g[x(t)] \equiv x(t)$  e  $M_r = \sigma^2$ ,  $W_r = M_r^{-1}$ . Come si verifica facilmente, si ottiene il seguente risultato

$$\mathbf{V} = \sigma^2 \left(\frac{\partial a}{\partial x}\right)^2$$

da cui si vede che la varianza dei parametri è determinata sia dalla *varianza dei dati* che dal *condizionamento del modello dinamico* misurato dalla quantità  $\partial a/\partial x$ . Per diminuire l'incertezza sui parametri si deve quindi migliorare l'accuratezza dei dati sperimentali, e scegliere modelli opportunamente ben condizionati. ■

Dall'equazione (13.16) si vede che  $\boldsymbol{\alpha}$  è una funzione lineare degli errori  $\boldsymbol{\eta}_r$ , e di conseguenza la sua densità di probabilità è Gaussiana, data da

$$\frac{1}{(2\pi)^{p/2} \sqrt{\det(\mathbf{V})}} \exp\left(-\frac{1}{2} \boldsymbol{\alpha}^T \mathbf{V}^{-1} \boldsymbol{\alpha}\right) \quad (13.20)$$

Sia ora  $\mathbf{b}$  un vettore di lunghezza euclidea unitaria; la funzione lineare  $\mathbf{b}^T \boldsymbol{\alpha}$  di  $\boldsymbol{\alpha}$  ha una distribuzione Gaussiana con varianza

$$\sigma_b^2 = \mathbf{b}^T \mathbf{V} \mathbf{b} \quad (13.21)$$

L'intervallo di confidenza  $\gamma(\mathbf{b})$  associato con  $\mathbf{b}^T \bar{\mathbf{a}}$  è allora

$$\gamma(\mathbf{b}) = k \sigma_b \quad (13.22)$$

ove la costante positiva  $k$  dipende dal livello di confidenza scelto. Per esempio, al livello 95% si ha  $k = 1.96$ . Pertanto, se si asserisce che

$$\mathbf{b}^T \bar{\mathbf{a}} - 1.96 \sigma_b \leq \mathbf{b}^T \mathbf{a}^* \leq \mathbf{b}^T \bar{\mathbf{a}} + 1.96 \sigma_b \quad (13.23)$$

allora, in un numero convenientemente alto di esperimenti simili, si è nel vero in 95% degli esperimenti. L'intervallo di confidenza per un elemento  $\bar{a}_i$  di  $\bar{\mathbf{a}}$  è calcolato assumendo  $b_i = 1$  e  $b_j = 0$ , per  $j \neq i$ . Da cui

$$\sigma_{a_i}^2 = (\mathbf{V})_{ii} \quad (13.24)$$

cioè le varianze relative a ogni singolo parametro sono date dagli elementi sulla diagonale principale della matrice di covarianza  $\mathbf{V}$ .

Se la matrice  $\mathbf{V}$  è definita positiva, i suoi autovalori sono tutti positivi e i suoi autovettori sono reali e ortogonali. La seguente equazione

$$\mathbf{z}^T \mathbf{V}^{-1} \mathbf{z} = 1, \quad \mathbf{z} \in \mathbb{R}^p \quad (13.25)$$

definisce un iperellissoide  $\mathcal{E}$  che ha gli autovettori di  $\mathbf{V}$  come le direzioni dei suoi assi principali, mentre le lunghezze dei semidiametri sono date dai corrispondenti autovalori di  $\mathbf{V}$  (cfr. Figura 13.1 corrispondente al problema considerato nell'Esempio 13.3).

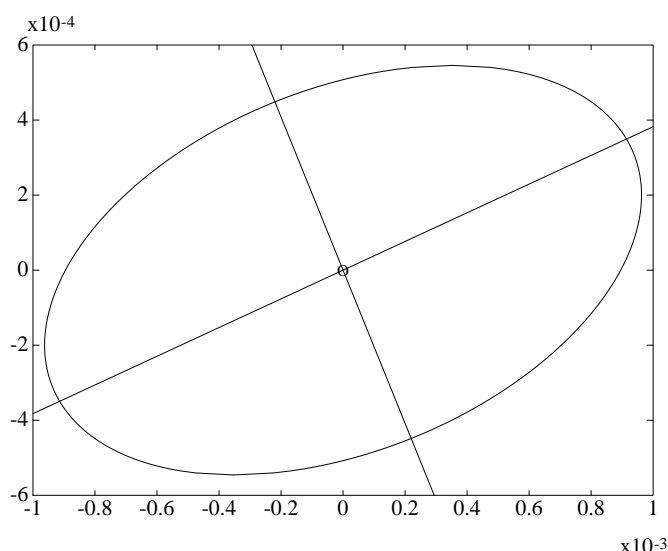


Figura 13.1: Rappresentazione dell'ellisse  $\mathbf{z}^T \mathbf{V}^{-1} \mathbf{z} = 1$ .

Si può allora dimostrare la seguente importante limitazione

$$\sqrt{\lambda_{\min}(\mathbf{V})} \leq \sigma_b \leq \sqrt{\lambda_{\max}(\mathbf{V})} \quad (13.26)$$

Il rapporto

$$\frac{\lambda_{\max}(\mathbf{V})}{\lambda_{\min}(\mathbf{V})} \quad (13.27)$$

fra l'autovalore più grande e l'autovalore più piccolo della matrice  $\mathbf{V}$  indica il *condizionamento* del problema di identificazione. Più tale rapporto è grande, e più l'iperellissoide  $\mathcal{E}$  è allungato; in questo caso,  $\mathbf{a}^*$  può essere ben condizionato rispetto ad alcune direzioni, ma vi sono direzioni rispetto alle quali è mal condizionato. Per questa ragione gli intervalli di confidenza corrispondenti ai singoli elementi  $\bar{a}_i$  di  $\bar{\mathbf{a}}$  sono meno informativi di quelli corrispondenti agli autovettori di  $\mathbf{V}$ .

### 13.2.3 Scelta della matrice $\mathbf{W}$

L'analisi precedente è indipendente dalla scelta delle matrici  $\mathbf{W}_r$  che definiscono la distanza  $F(\mathbf{a})$  da minimizzare. Una scelta conveniente di tali matrici è comunque importante nelle applicazioni per diversi motivi.

1. La scelta di  $\mathbf{W}_r$  influisce sulla configurazione dell'iperellissoide  $\mathcal{E}$  definito in (13.25), e di conseguenza sulla grandezza degli intervalli di confidenza  $\gamma(\mathbf{b})$ . Le matrici  $\mathbf{W}_r$  possono quindi essere definite in modo da minimizzare tali intervalli, cercando le matrici per le quali è minimo il diametro massimo dell'iperellissoide  $\mathcal{E}$ . Si può dimostrare che la scelta

$$\mathbf{W}_r := \mathbf{M}_r^{-1} \quad (13.28)$$

minimizza la quantità  $\sigma_b$  per *ogni* scelta di  $\mathbf{b}$ . Tale risultato non dipende dalle ipotesi relative alla distribuzione statistica degli errori  $\boldsymbol{\eta}_r$ ; la distribuzione degli errori interviene, infatti, solo quando si ricavano gli intervalli di confidenza dalla varianza  $\sigma_b^2$  di  $\mathbf{b}^T \bar{\mathbf{a}}$ . La scelta (13.28) corrisponde a pesare ogni risultato sperimentale secondo la loro affidabilità. Ricordiamo (cfr. Capitolo 8) che a tale scelta si perviene anche applicando il principio di massima verosimiglianza.

2. La scelta delle matrici  $\mathbf{W}_r$  può influenzare l'efficienza dei metodi numerici utilizzati per la minimizzazione della funzione  $F(\mathbf{a})$ . A questo proposito, osserviamo che, come si può vedere facilmente, le superfici di livello di  $F(\mathbf{a})$  in vicinanza al punto  $\mathbf{a}$  hanno la seguente forma

$$(\mathbf{a} - \bar{\mathbf{a}})^T \mathbf{H} (\mathbf{a} - \bar{\mathbf{a}}) = c^2$$

ove  $\mathbf{H}$  è la matrice definita nell'equazione (13.16). La scelta (13.28) rende le superfici di livello della forma dell'iperellissoide  $\mathcal{E}$ .

### 13.2.4 Formule riassuntive

Con scopo riassuntivo e per comodità, raccogliamo in questo paragrafo i vari passi della procedura per la identificazione e la valutazione statistica dei parametri in un sistema dinamico. Dato il sistema dinamico

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{a}, t); \quad \mathbf{x}(t_0) = \mathbf{x}^{(0)} \quad (13.29)$$

e il problema di minimo

$$\min_{\mathbf{a}} F(\mathbf{a}), \quad F(\mathbf{a}) := \sum_{r=1}^R (\mathbf{y}(t_r) - \mathbf{g}[\mathbf{x}(\mathbf{a}, t_r)])^T \mathbf{M}_r^{-1} (\mathbf{y}(t_r) - \mathbf{g}[\mathbf{x}(\mathbf{a}, t_r)]) \quad (13.30)$$



si suppone di avere calcolata una stima  $\bar{\mathbf{a}}$  del parametro  $\mathbf{a}^*$ . Per il calcolo del corrispondente intervallo di confidenza si procede nel modo seguente.

$$\mathbf{A} = \left[ \frac{\partial}{\partial x_j} f_i(\mathbf{x}, \bar{\mathbf{x}}, t) \right] \quad (13.31)$$

$$\mathbf{B} = \left[ \frac{\partial}{\partial a_j} f_i(\mathbf{x}, \bar{\mathbf{x}}, t) \right] \quad (13.32)$$

$$\dot{\mathbf{D}}(t) = \mathbf{A}(t)\mathbf{D}(t) + \mathbf{B}(t); \quad \mathbf{D}(t_0) = 0; \quad \mathbf{D}_r = \mathbf{D}(t_r) \quad (13.33)$$

$$\mathbf{G}_r = \left[ \frac{\partial}{\partial x_j} g_i[\mathbf{x}(\bar{\mathbf{a}}, t_r)] \right] \quad (13.34)$$

$$\mathbf{H} = \sum_{r=1}^R \mathbf{D}_r^T \mathbf{G}_r^T \mathbf{M}_r^{-1} \mathbf{G}_r \mathbf{D}_r \quad (13.35)$$

$$\mathbf{V} = \mathbf{H}^{-1} \quad (13.36)$$

$$\sigma_b^2 = \mathbf{b}^T \mathbf{V} \mathbf{b} \quad (13.37)$$

Dal momento che la matrice di sensitività  $\mathbf{D}$  ha come elementi le derivate delle variabili di stato rispetto ai parametri, essa può essere utilizzata per il calcolo delle derivate della funzione  $F(\mathbf{a})$  rispetto ai parametri  $\mathbf{a}$ . Si ha, infatti,

$$\frac{\partial F}{\partial a_j} = \sum_{i=1}^n \frac{\partial F}{\partial x_i} \frac{\partial}{\partial a_j} x_i(\mathbf{a}, t)$$

Questo fatto può essere sfruttato, in particolare, in quei metodi di minimizzazione che fanno uso esplicito delle derivate. Ricordiamo, tuttavia, che per il calcolo delle derivate della funzione  $F(\mathbf{a})$  rispetto ai parametri  $\mathbf{a}$  esistono metodi, in generale più convenienti, basati sulla costruzione del sistema aggiunto nello spirito della teoria dei controlli (cfr. successivo Capitolo 14).

### 13.2.5 Pianificazione degli esperimenti

Pianificare un esperimento significa, in sostanza, scegliere in modo razionale i valori di  $\mathbf{x}$  nei quali valutare l'osservata  $\mathbf{y}$ . Dal momento che i parametri sono stimati sulla base dei dati ottenuti negli esperimenti, è naturale porre la questione se sia possibile pianificare gli esperimenti in modo tale da rendere la procedura di identificazione la più efficiente possibile, ossia tale da rendere minima l'incertezza con la quale i parametri sono stimati.

Il problema della scelta ottimale degli esperimenti può essere teoricamente inquadrato nell'ambito della *teoria dell'informazione* analizzata nel Capitolo 8, mentre per una soluzione pratica può essere di aiuto la procedura di stima dei parametri sviluppata nei paragrafi precedenti di questo capitolo. Rinviamo alla bibliografia

per un opportuno approfondimento (cfr. in particolare Bard [10]), ci limiteremo in questo paragrafo ad alcune considerazioni di carattere generale.

Nell'ipotesi che i parametri  $\mathbf{a}$  siano distribuiti normalmente con media  $\mathbf{a}^*$  e matrice di covarianza  $\mathbf{V}$ , la corrispondente *entropia*, che misura l'incertezza associata con la densità di probabilità  $P(\mathbf{a}) = \mathcal{N}(\mathbf{a}^*, \mathbf{V})$  (cfr. Capitolo 8), è data da<sup>2</sup>

$$H(P) = \log \det(\mathbf{V}) \quad (13.38)$$

Tenendo presente che in corrispondenza ad una distribuzione normale è possibile mostrare che  $(\det(\mathbf{V}))^{1/2}$  è proporzionale al volume di una regione di confidenza nello spazio dei parametri  $\mathbf{a}$ , l'equazione (13.38) dice che l'incertezza cresce con il logaritmo del volume della regione di confidenza. Pertanto, un esperimento che diminuisce l'incertezza riduce anche il volume della regione di confidenza. In definitiva, per massimizzare la quantità di informazione guadagnata con una serie di esperimenti, si tratta di scegliere  $(\mathbf{y})_r$  in modo tale che l'incertezza sia minimizzata, ossia sia minima la quantità  $H(P)$ , o equivalentemente il determinante della matrice  $\mathbf{V}$ . In questo senso l'analisi della matrice di covarianza  $\mathbf{V}$  può suggerire scelte opportune dei dati sperimentali.

$t$	$y = x_1$		
	sperimentale	deviazione standard	calcolata
0	0.0000	1.000	0.0000
10	0.1000	0.043	0.1050
20	0.1920	0.040	0.1921
40	0.3555	0.035	0.3215
80	0.4330	0.030	0.4515
160	0.4455	0.028	0.4497
320	0.2470	0.040	0.2319

Tabella 13.1: Dati sperimentali e risultati calcolati mediante il modello identificato per il sistema dinamico (13.39).

► **Esempio 13.3** Come illustrazione della procedura di identificazione dei parametri, consideriamo il problema della identificazione dei parametri  $k_1$ ,  $k_2$  nel seguente problema a valori iniziali corrispondente alla reazione chimica monomolecolare tra due sostanze<sup>3</sup>.

$$\left\{ \begin{array}{l} \frac{dx_1}{dt} = k_1 x_2 - k_2 x_1 \\ \frac{dx_2}{dt} = -k_1 x_2 \end{array} \right. \quad \begin{array}{c} \uparrow k_2 \\ \boxed{x_1} \leftarrow \boxed{x_2} \quad k_1 \end{array} \quad (13.39)$$

<sup>2</sup>Il determinante di una matrice di covarianza è anche detto *varianza generalizzata*.

<sup>3</sup>Una terza sostanza prodotta dalla reazione non ha influenza sui risultati ed è ignorata.

ove  $x_1(t)$  e  $x_2(t)$  rappresentano le quantità delle due sostanze presenti al tempo  $t$ . I valori iniziali sono

$$x_1(0) = 0, \quad x_2(0) = 1 \quad (13.40)$$

e vengono supposti noti esattamente. Si suppone inoltre di avere a disposizione dei risultati sperimentali in corrispondenza alla sostanza  $x_1$  in tempi tra 10 min e 320 min dopo che i due reagenti sono stati mescolati. Nell'ipotesi che gli errori sperimentali siano distribuiti normalmente, in Tabella 13.1 sono rappresentati i valori medi  $(y_1)_r$  e le corrispondenti deviazioni standard  $\sigma_r$  (cfr. anche Figura 13.2).

I parametri da identificare corrispondono alle costanti di reazione  $k_1, k_2$ . Per la minimizzazione della funzione

$$F(k_1, k_2) := \sum_{r=1}^6 \left[ \frac{x_1(t_r) - y_1(t_r)}{\sigma_r} \right]^2$$

ove  $y_1(t_r)$  rappresenta il dato sperimentale, si utilizza il *metodo di Levenberg-Marquardt* studiato nel precedente Capitolo 5. Le soluzioni  $x_1(t), x_2(t)$  del problema a valori iniziali (13.39) sono approssimate mediante un metodo a passo e ordine variabile basato sulle formule di Adams (cfr. Capitolo 7). Si ottengono le seguenti stime dei valori ottimali  $k_1^*, k_2^*$

$$\bar{k}_1 = 0.011487, \quad \bar{k}_2 = 0.006445 \quad \Rightarrow \quad F(\bar{k}_1, \bar{k}_2) = 1.2257 \quad (13.41)$$

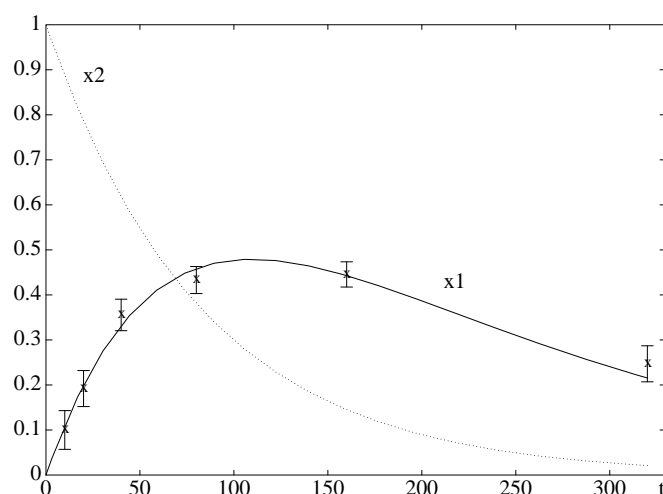


Figura 13.2: Rappresentazione dei risultati sperimentali e dei risultati ottenuti mediante il modello identificato nel caso del sistema dinamico (13.39).

La matrice di covarianza  $\mathbf{V}$  corrispondente ai parametri  $\bar{k}_1$  e  $\bar{k}_2$  è la seguente

$$\mathbf{V} = \begin{bmatrix} 0.926 & 0.1928 \\ 0.1928 & 0.2979 \end{bmatrix} \times 10^{-6} \quad (13.42)$$

dalla quale si ricavano le opportune informazioni sull'accuratezza dei parametri ottenuti. In particolare i valori  $\sqrt{V_{11}}, \sqrt{V_{22}}$  forniscono le deviazioni standard relative ai parametri  $k_1$  e

$k_2$ ; si ha quindi

$$\begin{aligned}\bar{k}_1 &= 1.14871 \cdot 10^{-2}; & \sigma_{k_1} &= 9.62491 \cdot 10^{-4} \\ \bar{k}_2 &= 6.44530 \cdot 10^{-3}; & \sigma_{k_2} &= 5.45812 \cdot 10^{-4}\end{aligned}$$

Gli autovalori di  $\mathbf{V}$  sono dati da

$$\lambda_{\min} = 0.24342 \cdot 10^{-6}; \quad \lambda_{\max} = 0.98084 \cdot 10^{-6}$$

da cui si ricava la seguente limitazione

$$\sqrt{\lambda_{\min}} \leq \sigma_b \leq \sqrt{\lambda_{\max}}; \quad 4.93408 \cdot 10^{-4} \leq \sigma_b \leq 9.90377 \cdot 10^{-4}$$

Il valore  $\lambda_{\max}/\lambda_{\min}$  fornisce l'indicazione sul *condizionamento* del problema. Per l'esempio che stiamo considerando tale rapporto vale  $\approx 4.0289$ , per cui il problema non è mal condizionato. Gli autovettori della matrice  $\mathbf{V}$  sono rappresentati dalle colonne della seguente matrice

$$\begin{bmatrix} -0.9623 & 0.2717 \\ -0.2717 & -0.9623 \end{bmatrix}$$

In Figura 13.1 è rappresentata l'ellisse  $\mathbf{z}^T \mathbf{V}^{-1} \mathbf{z} = 1$  che ha gli autovettori di  $\mathbf{V}$  come le direzioni dei suoi assi principali e la radice quadrata dei corrispondenti autovalori di  $\mathbf{V}$  come semidiametri.

Osserviamo, infine, che dalla matrice  $\mathbf{V}$  è possibile ricavare il *coefficiente di correlazione* tra i parametri. Nell'esempio, il coefficiente di correlazione tra  $k_1$  e  $k_2$  è dato da 0.3671. ■

Tutti i fiumi vanno al mare,  
eppure il mare non è mai pieno.  
Qoèlet 1,7

## Capitolo 14

# Introduzione alla teoria del controllo ottimo

Uno degli scopi principali dell'introduzione e dell'analisi dei modelli matematici è certamente quello di approfondire e accrescere la nostra conoscenza dei fenomeni naturali. Altrettanto importante, tuttavia, è l'obiettivo di utilizzare tali modelli per influenzare il comportamento del sistema reale rappresentato dal modello matematico. La *teoria dei controlli* (*optimal control theory*) ha come campo di ricerca questo secondo obiettivo. In maniera più precisa, nella teoria dei controlli si cerca di determinare i valori di alcune variabili, dette *variabili di controllo*, in maniera che il sistema *minimizzi* (o massimizzi) un determinato criterio. Il concetto verrà illustrato nel seguito attraverso opportune esemplificazioni. È importante, comunque, realizzare il fatto che premessa indispensabile per operare un controllo significativo di un sistema è la disponibilità di un *buon modello matematico*. In maniera sintetica, ricordiamo che un modello matematico è un *buon modello*, quando esso è formulato in termini di un problema matematico ben posto, ossia per il quale si ha esistenza, unicità, dipendenza continua della soluzione dai dati, nell'ambito di una classe di funzioni con proprietà interpretabili in termini del modello reale, e per il quale è possibile fornire un'espressione analitica della soluzione, o alternativamente, sono disponibili opportuni metodi numerici. Questi aspetti sono stati l'oggetto prevalente della trattazione e dell'analisi condotta nei precedenti capitoli del presente volume. Lo scopo di quest'ultimo capitolo è quello di dare, prevalentemente in forma descrittiva e esemplificativa, le idee matematiche e numeriche di base della teoria dei controlli, che rappresenta da tempo uno degli strumenti matematici più interessanti in vari settori dell'ingegneria, ma, più recentemente anche in altri campi di ricerca, quali la chimica, la biologia, la medicina. Per un opportuno approfondimento segnaliamo, ad esempio Bellman [15], Bryson e Ho [23], Eisen [53], Lee e Markus [107], Macki e Strauss [110], Swan [150].

## 14.1 Modelli introduttivi

I modelli che analizzeremo in questo paragrafo hanno la funzione di introdurre in maniera elementare il significato e la nomenclatura di base di un problema di controllo.

► **Esempio 14.1** (*Impiego ottimale di una risorsa*) Supponiamo che una risorsa possa essere ripartita in tempi successivi secondo due modi differenti di utilizzo, indicati simbolicamente con (I) e rispettivamente (II). Si supponga inoltre che l'impiego della risorsa porti, da una parte ad una variazione (ad esempio una riduzione) della quantità di risorsa disponibile, e dall'altra ad un reddito. In maniera schematica, il problema è allora quello di *trovare ad ogni successivo passo la strategia di ripartizione in modo da massimizzare il reddito relativo a tutto il tempo di impiego della risorsa*.

Per formulare il problema in termini matematici, indichiamo con  $t = 0, 1, \dots, T$  i tempi successivi e con  $x(t) \geq 0$  la quantità di risorsa disponibile al generico tempo  $t$  e che può essere ripartita per l'utilizzo nei due modi durante il periodo  $(t, t+1)$ . La variabile  $x(t)$  rappresenta la variabile di *stato*. Con  $u(t)$  indichiamo la quantità di risorsa destinata all'utilizzo (I); si deve avere  $0 \leq u(t) \leq x(t)$ . La funzione  $u(t)$  rappresenta la *decisione* (il *controllo*) da prendere al tempo  $t$ . Nell'ipotesi di utilizzare *tutta* la risorsa disponibile, per l'utilizzo (II) si ha allora a disposizione la quantità  $x(t) - u(t)$ .

Supponiamo, ora, che l'utilizzo della risorsa durante il generico periodo  $(t, t+1)$ ,  $t = 0, 1, \dots, T-1$  porti ad una variazione della quantità di risorsa  $x(t)$ . Ad esempio, per fissare le idee, supponiamo che l'utilizzo nel modo (I) comporti una riduzione della risorsa disponibile di un fattore 0.8 e nel modo (II) di un fattore 0.5. La quantità di risorsa  $x(t+1)$  disponibile al tempo  $t+1$  è allora fornita dalla seguente relazione

$$x(t+1) = 0.8u(t) + 0.5(x(t) - u(t)) = 0.5x(t) + 0.3u(t) \quad (14.1)$$

L'equazione (14.1) descrive l'evoluzione del sistema rappresentato dalla variabile  $x(t)$  e viene chiamata *equazione di stato*. Più in generale, l'equazione di stato per un sistema *discreto* (nel quale, come nell'esempio che stiamo esaminando, le variabili sono definite su un insieme finito o più in generale numerabile) è rappresentata da una equazione alle differenze della forma

$$x(t+1) = f(t, x(t), u(t)), \quad t = 0, 1, \dots \quad (14.2)$$

ove la funzione  $f(t, x(t), u(t))$  definisce il *modello matematico* attraverso il quale viene simulato il sistema reale. Essa può essere una funzione *lineare*, come nell'esempio considerato, o più in generale *non lineare*; inoltre le variabili  $x$  e  $u$  possono essere vettori e allora la  $f$  è una funzione di più variabili e a valori vettoriali. La Figura 14.1 fornisce una rappresentazione in forma di flusso (*flow chart*) di un generico processo a più stadi. Ad ogni scelta fissata della successione dei controlli  $u(t)$ ,  $t = 0, 1, \dots$  corrisponde una determinata successione di valori  $x(t)$ ,  $t = 0, 1, \dots$ , che viene chiamata la *traiettoria* del sistema corrispondente alla successione di controlli fissata. Sottolineiamo che gli elementi della successione  $x$  dipendono oltre che dal tempo  $t$  anche dalla successione dei controlli  $u$ . In Tabella 14.1 sono riportate, come esemplificazione, le traiettorie corrispondenti ad alcune particolari scelte dei controlli  $u(t)$ .

Supponiamo ora che l'utilizzo della risorsa nell'intervallo di tempo  $(t, t+1)$  dia origine a un *reddito*, indicato con  $J(t)$ . Tale reddito dipende in generale dalla decisione  $u(t)$  e dallo

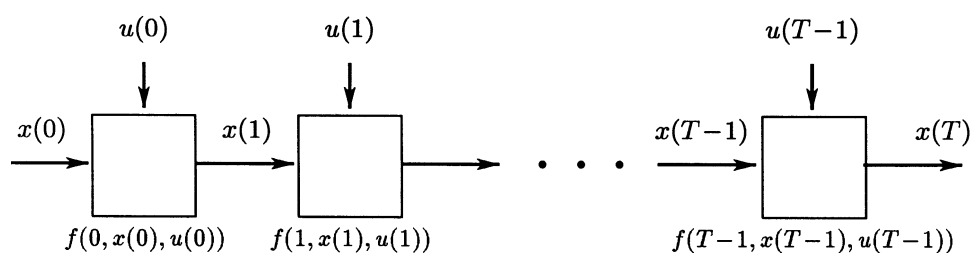


Figura 14.1: Flow chart di un processo a più stadi (multistage).

stato  $x(t)$ . Nell'esempio che stiamo considerando possiamo supporre, per fissare le idee, che l'impiego nel modo (I) dia un reddito 2, per unità di risorsa impiegata, e 3 nel modo (II). Si ha pertanto

$$J(t, x(t), u(t)) := 2u(t) + 3(x(t) - u(t)) = 3x(t) - u(t) \quad (14.3)$$

Il reddito totale  $J(x, u)$  relativo all'intervallo di tempo  $0, T - 1$  è ottenuto *sommando* i redditi su ogni intervallo di tempo  $(t, t + 1)$ ; si ha cioè

$$J(x, u) := \sum_{t=0}^{T-1} J(t, x(t), u(t)) \quad (14.4)$$

Sottolineiamo che, fissato  $T$ ,  $J(x, u)$  è una funzione della sola variabile  $u$ , in quanto la  $x$  è definita dalla  $u$  attraverso le equazioni di stato. Possiamo allora formulare il seguente problema di ottimizzazione.

▼ **Problema 14.1** Si cerca una successione di decisioni  $u^*(t), t = 0, \dots, T - 1$  tale che, indicata con  $x^*(t), t = 0, \dots, T - 1$  la corrispondente traiettoria, si abbia

$$J(x^*, u^*) = \max_{u \in \mathcal{U}} J(x, u) \iff J(x^*, u^*) \geq J(x, u) \quad \forall u \in \mathcal{U} \quad (14.5)$$

ove  $\mathcal{U}$  indica l'insieme dei controlli ammissibili.

Nell'esempio considerato l'insieme  $\mathcal{U}$  è definito dai vincoli  $0 \leq u(t) \leq x(t)$ , per ogni  $t = 0, 1, \dots, T - 1$ . La funzione da massimizzare (o minimizzare)  $J(x, u)$  è detta *funzione obiettivo*, o *funzione costo*, e le funzioni  $u^*, x^*$  che soddisfano la condizione (14.5) sono dette rispettivamente *controllo ottimale* e *traiettoria ottimale*. Una funzione obiettivo che può essere definita come somma (nel caso di sistemi discreti) e come integrale (nel caso di sistemi continui) di obiettivi locali, ossia definiti ad ogni passo  $t$ , è detta di tipo *additivo*. In questo capitolo considereremo solo funzioni obiettivo di questo tipo.

Per l'esempio che stiamo esaminando, si vede dalla Tabella 14.1 che in un processo a 3 stadi non sono ottimali le scelte consistenti nel ripartire ad ogni passo tutta la risorsa costantemente ad uno solo dei modi di impiego. Il motivo è da ricercare nel fatto che da una parte il modo (I) è quello che rende meno, mentre d'altra parte il modo (II) rende di più, ma consuma una parte maggiore di risorsa. Si intuisce, quindi, che se il processo è sufficientemente lungo è necessario un compromesso. Ancora dalla tabella si vede che il controllo

	t	0	1	2	3	J
(1)	u	x(0)	0	0		5.6
	x	1	0.8	0.4	0.2	
	J(t)	2	2.4	1.2		
(2)	u	x(0)	x(1)	x(2)		4.88
	x	1	0.8	0.64	0.512	
	J(t)	2	1.6	1.28		
(3)	u	0	0	0		5.25
	x	1	0.5	0.25	0.125	
	J(t)	3	1.5	0.75		

Tabella 14.1: Traiettorie e relativi costi corrispondenti a differenti controlli in un modello multistage.

consistente nell'impiego al primo passo nel modo (I), che consuma meno, e successivamente nell'impiego (II), che rende di più, fornisce un reddito maggiore rispetto alle strategie precedenti. Nel seguito mostreremo che effettivamente tale scelta è ottimale. Il risultato verrà ottenuto applicando due idee fondamentali per il trattamento dei problemi di controllo, la *programmazione dinamica* e il *principio del massimo*. In ambedue le procedure il problema di ottimizzazione globale, ossia relativo a tutto l'intervallo  $(0, T - 1)$  del processo, viene ricondotto, in maniera differente, alla risoluzione di problemi di ottimizzazione locali, ossia relativi a ciascun intervallo  $(t, t + 1)$ . ■

► **Esempio 14.2** (*Sistema lineare e criterio quadratico*) Supponiamo che le variabili di stato di un processo siano rappresentate dalle funzioni a valori vettoriali  $t \rightarrow \mathbf{x}(t) \in \mathbb{R}^n$ ,  $n \geq 1$ , e che le equazioni di stato siano date dal seguente problema a valori iniziali per un sistema di equazioni differenziali

$$\begin{cases} \frac{d\mathbf{x}(t)}{dt} = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{u}(t) & t \in (t_0, T) \\ \mathbf{x}(t_0) = \mathbf{x}_0 \end{cases} \quad (14.6)$$

ove  $\mathbf{u}(t) \in \mathbb{R}^m$ ,  $m \geq 1$  rappresenta il vettore dei controlli e  $t_0, T$ , rispettivamente il tempo iniziale e finale del processo, sono quantità fissate. Le matrici  $\mathbf{F} \in \mathbb{R}^{n \times n}$  e  $\mathbf{G} \in \mathbb{R}^{m \times m}$  sono supposte indipendenti da  $\mathbf{x}$  e da  $\mathbf{u}$  e funzioni continue in  $t$ .

Consideriamo quindi una funzione obiettivo della seguente forma

$$J(\mathbf{x}, \mathbf{u}) := \int_{t_0}^T [(\mathbf{x}(s))^T \mathbf{Q}(s) \mathbf{x}(s) + (\mathbf{u}(s))^T \mathbf{R}(s) \mathbf{u}(s)] ds + \mathbf{x}^T(T) \mathbf{A} \mathbf{x}(T) \quad (14.7)$$

ove le matrici  $\mathbf{Q}(t) \in \mathbb{R}^{n \times n}$ ,  $\mathbf{R}(t) \in \mathbb{R}^{m \times m}$  sono supposte continue in  $t$  e simmetriche definite positive. Più in generale, la matrice  $\mathbf{Q}$  può essere una matrice semidefinita positiva, ossia tale che  $\mathbf{x}(t)^T \mathbf{Q}(t) \mathbf{x}(t) \geq 0$  per ogni  $\mathbf{x}(t)$ . La matrice  $\mathbf{A}$  è simmetrica definita positiva.



Il problema della ricerca della funzione  $\mathbf{u}^*$  che minimizza il costo  $J(\mathbf{x}, \mathbf{u})$  definito in (14.7), quando la dipendenza della variabile di stato  $\mathbf{x}$  dal controllo  $\mathbf{u}$  è data dal problema a valori iniziali (14.6), è noto come problema di controllo a *sistema lineare e criterio quadratico*. Tale tipo di problema è ben noto nella letteratura, sia per le numerose applicazioni pratiche, sia anche, come vedremo nel seguito, per i significativi risultati teorici e numerici che per esso possono essere ottenuti. Osservando che la definizione del funzionale (14.7) rappresenta un compromesso tra il costo della funzione di stato  $\mathbf{x}$  e il costo della variabile di controllo  $\mathbf{u}$ , si intuisce il suo interesse in tutti quei problemi in cui si vuole mantenere le variabili di stato ad un livello assegnato, senza un impiego eccessivo della variabile di controllo. Per importanti applicazioni nell'ambito biomedico si veda ad esempio Swan [150] e per problemi di ingegneria Bryson e Ho [23]. ■

► **Esempio 14.3** (*Controllo in chemioterapia*) In maniera schematica, il problema consiste nella ricerca di un protocollo di assegnazione di un farmaco per ridurre dopo un tempo  $T$  una massa di tessuto neoplastico al di sotto di una soglia prestabilita, tenendo presente che la terapia può avere un effetto dannoso sui tessuti sani. Il problema può essere formulato come problema di controllo attraverso i seguenti passi.

1. Si costruisce un *modello matematico* che descrive l'effetto della terapia sulla dinamica cellulare del tessuto. Questo comporta la *scelta* di un particolare modello di *accrescimento cellulare* e la traduzione in termini matematici dell'*efficacia* del farmaco sul tessuto.
2. Sulla base delle *osservazioni sperimentali* si stabiliscono i limiti di validità del modello (per una discussione di questo punto cfr. i Capitoli 12 e 13).
3. Il *modello matematico* viene assunto come modello *rappresentativo* del modello reale, e diventa quindi il *sistema di stato* nel problema di controllo.
4. Si definisce l'obiettivo da raggiungere, costruendo un appropriato criterio  $J$ .

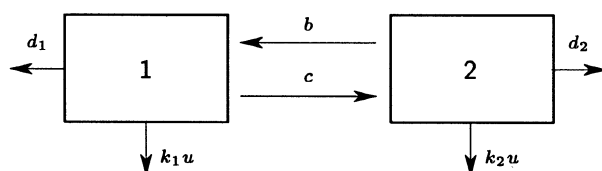


Figura 14.2: Accrescimento cellulare a due compartimenti: (1) cellule che proliferano; (2) cellule che non proliferano.

Come esemplificazione dei passi precedenti, che segnano in sostanza la traccia di una generica ricerca di modellistica matematica, consideriamo il seguente modello.

Si parte dall'*ipotesi biologica* che l'efficacia del farmaco utilizzato possa dipendere dalle varie fasi del ciclo cellulare. Per tenere conto di tale ipotesi sono quindi necessari modelli di accrescimento del tessuto, nei quali venga considerata esplicitamente la posizione delle cellule nel ciclo cellulare. Rinviamo per un'ampia panoramica di modelli di questo tipo ad esempio ad Eisen [53], ci limiteremo a discutere il seguente modello, nel quale le cellule del

tessuto neoplastico sono pensate suddivise in due sottopopolazioni: le cellule che *proliferano* e le cellule che si trovano nel *ciclo cellulare*. Più precisamente, si suppone che il tessuto possa essere rappresentato dal modello a due compartimenti indicato nella Figura 14.2. Indicato con  $x_1(t)$ ,  $x_2(t)$  il numero totale, al tempo  $t$ , delle cellule rispettivamente nel compartimento 1 e 2, in base alla tecnica analizzata nel Capitolo 12, la dinamica delle due popolazioni può essere descritta dal seguente problema a valori iniziali

$$\begin{cases} \frac{dx_1}{dt} = \overbrace{a x_1}^{\text{crescita}} - \overbrace{k_1 u x_1}^{\text{terapia}} + \overbrace{b x_2}^{\text{flusso}} - \overbrace{c x_1}^{\text{flusso}} - \overbrace{d_1 x_1}^{\text{morte}}, & x_1(0) = x_{10} \\ \frac{dx_2}{dt} = \overbrace{c x_1}^{\text{flusso}} - \overbrace{b x_2}^{\text{flusso}} - \overbrace{k_2 u x_2}^{\text{terapia}} - \overbrace{d_2 x_2}^{\text{morte}}, & x_2(0) = x_{20} \end{cases} \quad (14.8)$$

ove la funzione  $u(t)$  indica l'intensità della terapia e  $k_1$ ,  $k_2$  sono i coefficienti di efficacia del farmaco sulle due popolazioni. L'assunzione di due coefficienti, possibilmente differenti, corrisponde all'ipotesi biologica che il farmaco possa agire in maniera differenziata sulle due popolazioni. Le quantità  $x_{10}$ ,  $x_{20}$  rappresentano i valori iniziali (al tempo  $t = 0$ ) delle due popolazioni. I vari coefficienti  $a$ ,  $b$ ,  $c$ , ... sono parametri da identificare in base ad opportuni dati sperimentali e utilizzando le tecniche numeriche analizzate nel Capitolo 13.

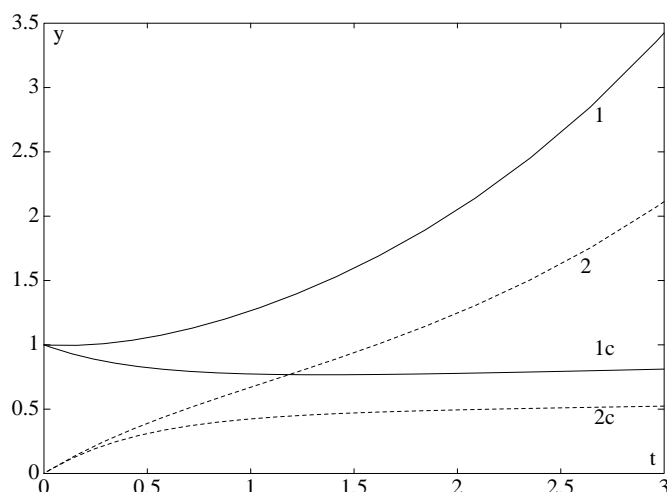


Figura 14.3: Le curve (1) e (2) rappresentano rispettivamente le popolazioni  $x_1(t)$  e  $x_2(t)$  quando l'effetto del farmaco è nullo. Le curve (1c) e (2c) rappresentano le popolazioni sotto l'effetto di un farmaco, supposto costante nel tempo. Nel modello si è assunto  $k_1 > k_2$  e  $a > d_1 + cd_2/(d_2 + b)$ .

Il sistema (14.8) è un esempio di sistema *bilineare*, ossia lineare separatamente nelle due variabili  $\mathbf{x} = [x_1, x_2]^T$  e  $u$ , ma quadratico nel complesso delle due variabili. I sistemi bilineari, insieme ai sistemi lineari considerati nell'esempio precedente, costituiscono una classe importante di sistemi nella teoria dei controlli. Nella Figura 14.3 è rappresentata, a solo scopo esemplificativo, la dinamica delle popolazioni  $x_1(t)$ ,  $x_2(t)$ , sia in assenza di terapia (ossia, quando  $u(t) \equiv 0$ ), che in presenza di una terapia. I coefficienti del sistema sono stati

assunti in maniera che la popolazione totale in assenza di terapia cresca illimitatamente. Nel secondo caso l'intensità della terapia è supposta costante e i coefficienti di efficacia sono stati supposti diversi, con  $k_1 > k_2$ .

Assumendo il sistema (14.8) come sistema di stato, è possibile formulare il problema della terapia ottimale definendo opportunamente la funzione obiettivo. Nella definizione dell'obiettivo occorre tenere presente dell'eventuale danno che il farmaco può arrecare ai tessuti sani. Indicando allora con  $x_{11}$ ,  $x_{21}$  delle soglie di sicurezza per le popolazioni di cellule neoplastiche da raggiungere ad un tempo prefissato  $T$ , si può definire la funzione obiettivo  $J(u)$  nel seguente modo

$$J(u) = |x_1(T) - x_{11}|^2 + |x_2(T) - x_{21}|^2 + \int_0^T (u(s))^T R(s) u(s) ds \quad (14.9)$$

ove  $R$  è una funzione positiva che *pesa* ad ogni tempo  $t$  il danno arrecato dal farmaco alle cellule sane; la funzione obiettivo è quindi di tipo quadratico. In definitiva, il problema della terapia ottimale può essere formulato in termini matematici come il problema della ricerca della funzione di controllo  $u(t)$  ( $\geq 0$ ) che minimizza il funzionale (14.9), ove le variabili di stato  $x_1(t)$ ,  $x_2(t)$  dipendono da  $u(t)$  attraverso il sistema di stato (14.8).

Naturalmente, la (14.9) non è la sola possibilità di definire l'obiettivo. Come ulteriore esempio interessante per il seguito (cfr. il successivo Esempio 14.20) segnaliamo la seguente alternativa. Dato il numero iniziale  $x_1(0)$ ,  $x_2(0)$  delle cellule, l'obiettivo della terapia è quello di ridurre nel tempo assegnato  $T$  il numero delle cellule a

$$x_1(T) = x_{11}; \quad x_2(T) = x_{21} \quad (14.10)$$

utilizzando la minima quantità di farmaco. Allora, la funzione obiettivo diventa

$$J(u) := \int_0^T u(s) ds \quad (14.11)$$

Rispetto alla formulazione precedente, ora si sono fissati i valori delle variabili di stato non solo all'istante iniziale  $t = 0$ , ma anche al tempo finale  $t = T$ . Il sistema di stato diventa quindi un problema ai limiti, anziché un problema a valori iniziali. Osserviamo che per  $u(t)$  fissata in maniera generica tale problema ai limiti non ha in generale soluzione, in quanto per un sistema di due equazioni differenziali del primo ordine la soluzione è individuata da due condizioni. Dobbiamo comunque tenere conto che nel problema di controllo la funzione è una ulteriore incognita. Se il problema di controllo ha una soluzione ottimale  $u^*$ , significa che la corrispondente traiettoria  $\mathbf{x}^*(t)$  porta il sistema dal valore iniziale  $[x_{10}, x_{20}]$  al valore finale  $[x_{11}, x_{21}]$ . Nella terminologia della teoria dei controlli il punto finale  $[x_{11}, x_{21}]$  è chiamato l'insieme bersaglio (*target set*). Nella formulazione precedente (14.9), si è solo fissato il valore finale del tempo  $T$ , mentre i valori delle variabili  $x_1(T)$ ,  $x_2(T)$  sono a priori arbitrari. In questo caso, quindi, il target set è costituito dall'insieme dei valori  $\{T, x_1, x_2\}$  al variare di  $x_1, x_2$ .

Problemi di controllo analoghi a quello ora considerato si hanno in numerosi altri settori della biologia e della medicina; segnaliamo, in particolare, il problema di controllo ottimale del *diabete* (funzione di controllo = intensità di inoculazione di insulina); modelli di controllo ottimale di disturbi del sistema *endocrino* (in particolare della tiroide); problemi di controllo del *sistema circolatorio* (controllo della ipertensione); modelli di controllo nel *sistema immunitario* (per un approfondimento si veda ad esempio Swan [53]).

Terminiamo l'esempio con un'analisi riguardante la possibilità di mantenere nel tempo la concentrazione  $C(t)$  di un farmaco nel sangue ad un livello compreso tra due limitazioni, corrispondenti rispettivamente ad un livello inferiore  $C_L$  di efficacia del farmaco e ad un livello superiore  $C_H$  di sicurezza. Supponendo che il farmaco venga amministrato in dosi  $C_0$  ad intervalli di tempo di lunghezza fissata  $t_0$ , il problema precedente può essere considerato come un particolare *problema di controllo*, nel quale le variabili di controllo sono le quantità  $C_0$  e  $t_0$  e l'obiettivo è il raggiungimento di una concentrazione del farmaco  $C(t)$  che sia nel contempo sicura (cioè  $\leq C_H$ ) ed efficace (cioè  $\geq C_L$ ). Più precisamente la funzione controllo  $u(t)$  è rappresentata da impulsi (cfr. Appendice B per la definizione di funzione impulsiva) nei successivi istanti  $0, t_0, 2t_0, \dots$

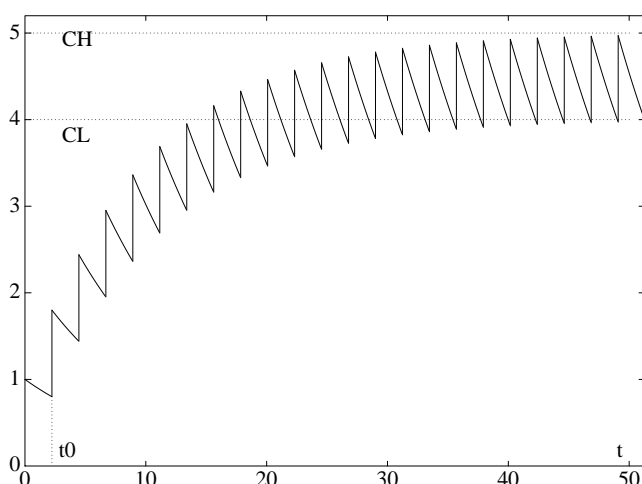


Figura 14.4: Concentrazione del farmaco negli intervalli successivi  $[i t_0, (i + 1)t_0]$ ,  $i = 0, 1, \dots$ . Asintoticamente la concentrazione è compresa tra il livello di sicurezza  $C_H$  e il livello di efficacia  $C_L$ .

Faremo ora vedere brevemente che sotto le seguenti ipotesi il problema di controllo ora formulato può essere risolto in maniera esplicita.

1. La diminuzione della concentrazione del farmaco nel sangue è proporzionale alla concentrazione.
2. La dose inoculata si diffonde in maniera sufficientemente rapida da poter assumere che la sua concentrazione nel sangue dopo la sua amministrazione sia immediatamente uguale  $C_0$ .

Dalla prima ipotesi si ha il seguente semplice modello

$$\frac{dC(t)}{dt} = -k C(t) \quad (14.12)$$

ove la costante  $k > 0$  è chiamata la *costante di eliminazione*. Se con  $C(0)$  indichiamo il valore della concentrazione all'istante iniziale  $t = 0$ , la soluzione dell'equazione (14.12) è, come noto, la funzione

$$C(t) = C(0)e^{-kt} \quad 0 \leq t < \infty \quad (14.13)$$

In base alla seconda ipotesi si ha  $C(0) = C_0$  e al tempo  $t = t_0$  la concentrazione assume il valore residuo  $R_1$  dato da  $C_0 e^{-kt_0}$ . Al tempo  $t_0$  si somministra la seconda dose che innalza la concentrazione al valore  $C_1 = R_1 + C_0$ . Applicando la stessa procedura negli istanti successivi, si ottengono le seguenti relazioni ricorrenti

$$C_{i-1}e^{-kt_0} = R_i \quad i = 1, 2, \dots \quad (14.14)$$

$$R_i + C_0 = C_i \quad i = 1, 2, \dots \quad (14.15)$$

dalle quali si ha

$$C_i = C_0 + C_{i-1} e^{-kt_0} \quad i = 1, 2, \dots \quad (14.16)$$

Per induzione si può mostrare facilmente che per un generico intero  $n$  si ha

$$C_n = C_0(1 + e^{-kt_0} + e^{-2kt_0} + \dots + e^{-nkt_0}) = C_0 \left( \frac{1 - e^{-(n+1)kt_0}}{1 - e^{-kt_0}} \right) \quad (14.17)$$

Dalla (14.15) si ricava la relazione

$$R_n = C_0 \left( \frac{1 - e^{-nkt_0}}{1 - e^{-kt_0}} \right) e^{-kt_0} \quad (14.18)$$

Dalle equazioni (14.17), (14.18) si ricava che le successioni  $C_n$  e  $R_n$  sono crescenti; inoltre, prendendo il limite per  $i \rightarrow \infty$  nell'uguaglianza (14.15), si ottiene

$$R + C_0 = C \quad \text{ove} \quad C = \frac{C_0}{1 - e^{-kt_0}}, \quad R = \frac{C_0 e^{-kt_0}}{1 - e^{-kt_0}} \quad (14.19)$$

Come anche mostra la Figura 14.4, la concentrazione  $C(t)$  tende a oscillare tra  $R$  e  $C_0 + R$ . Dal risultato ora ottenuto è immediato ricavare i valori di  $C_0$  e di  $t_0$  per i quali la concentrazione tende a oscillare tra due valori assegnati  $C_L$  e  $C_H$ . Basterà porre

$$C_0 + R = C_H, \quad R = C_L \Rightarrow C_0 = C_H - C_L$$

Dall'espressione di  $R$  in (14.19) si ha

$$C_L = \frac{(C_H - C_L) e^{-kt_0}}{1 - e^{-kt_0}} \Rightarrow e^{kt_0} = \frac{C_H}{C_L} \Rightarrow t_0 = \frac{1}{k} \ln \frac{C_H}{C_L}$$

► **Esempio 14.4** (*Sistema lineare e criterio di tempo minimo*) Quest'ultimo esempio introduttivo permette, tra l'altro, di completare la terminologia relativa ai problemi di controllo e di introdurre i problemi di controllo con *criterio di tempo minimo*. Ulteriori problemi di controllo verranno studiati nel seguito, dopo aver introdotto alcune tecniche risolutive. Con riferimento alla Figura 14.5, un'automobile si muove dal punto  $O$  lungo la direzione  $d$  sotto l'effetto di una accelerazione  $\alpha(t)$  e di una decelerazione  $\beta(t)$ .

Per semplificare il modello, l'automobile è approssimata da un punto, il cui movimento è descritto dal seguente modello

$$\ddot{d}(t) = \alpha(t) + \beta(t)$$

ove  $d(t)$  indica la distanza percorsa al tempo  $t$ . Scegliendo la *posizione* e la *velocità* come *variabili di stato*, cioè

$$x_1(t) := d(t); \quad x_2(t) := \dot{d}(t)$$

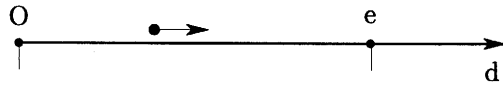


Figura 14.5: Problema di controllo con criterio di tempo minimo.

e ponendo

$$u_1(t) := \alpha(t); \quad u_2(t) := \beta(t)$$

si ha il seguente sistema di stato

$$\dot{\mathbf{x}}(t) = \mathbf{A} \mathbf{x}(t) + \mathbf{G} \mathbf{u}(t) \quad (\text{sistema di stato}) \quad (14.20)$$

ove abbiamo posto

$$\mathbf{x} := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{u} := \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \mathbf{A} := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{G} := \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

**Vincoli fisici** Dopo aver scelto un modello matematico, il passo successivo consiste nel definire i vincoli fisici sulla funzione di stato e di controllo. Supponendo che l'automobile parta da  $O$  al tempo  $t_0$  con velocità nulla e che si fermi (quindi, ancora con velocità nulla) in  $e$  al tempo  $t_f$ , si hanno le seguenti *condizioni ai limiti*

$$\begin{cases} x_1(t_0) = 0 \\ x_2(t_0) = 0 \end{cases} \quad \text{condizioni iniziali} \quad \begin{cases} x_1(t_f) = e \\ x_2(t_f) = 0 \end{cases} \quad \text{condizioni finali} \quad (14.21)$$

Sottolineiamo una situazione caratteristica dei problemi di controllo e che abbiamo già rilevato nell'esempio precedente, ossia la possibilità di avere per il sistema differenziale di stato un numero di condizioni superiore a quello usualmente previsto dalla teoria relativa ai problemi ai limiti (in questo caso, quattro condizioni per un sistema del secondo ordine). In realtà, come abbiamo già osservato, nel problema di controllo si ha come ulteriore incognita la funzione  $\mathbf{u}(t)$  e, per l'esempio che stiamo considerando, anche il tempo finale (chiamato anche *time horizon*)  $t_f$ . Esiste, naturalmente, il problema dell'esistenza di funzioni  $\mathbf{u}(t)$  e di valori  $t_f$  tali che il sistema di stato (14.20) sia risolubile. Questo tipo di problema, la cui soluzione è preliminare alla formulazione del problema di controllo, è noto in letteratura come problema di *controllabilità* del sistema.

In aggiunta alle condizioni ai limiti (14.21), si può aggiungere il seguente vincolo

$$0 \leq x_2(t) \quad (14.22)$$

che equivale ad imporre che l'automobile non possa tornare indietro. Per quanto riguarda la variabile di controllo possiamo imporre i seguenti vincoli

$$\begin{aligned} 0 &\leq u_1(t) \leq M_1 \\ -M_2 &\leq u_2(t) \leq 0 \end{aligned} \quad (14.23)$$

che traduce l'esistenza di una limitazione superiore  $M_1$  alla capacità di accelerare, e analogamente di una limitazione inferiore  $-M_2$  per la decelerazione. Infine, possiamo tenere conto

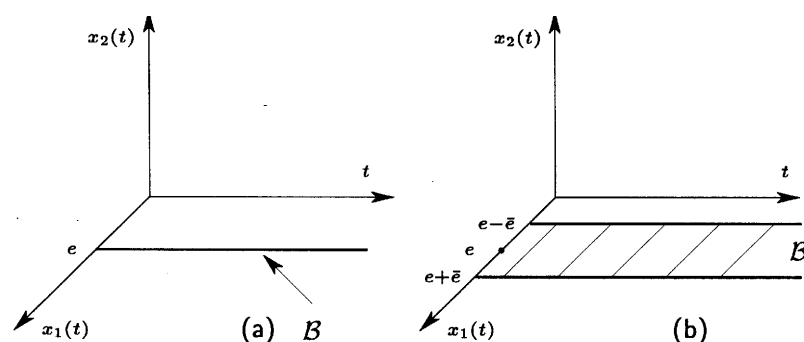


Figura 14.6: (a) target set per  $x_1 = e$ ; (b) target set per  $|x_1 - e| \leq \bar{e}$ ,  $x_2 = 0$ .

di *vincoli* che coinvolgono sia la variabile di stato che la variabile di controllo; si può, ad esempio, supporre che la macchina abbia un pieno  $B$  di benzina e che non vi siano stazioni di servizio. Poiché il consumo di benzina è proporzionale alla velocità e alla accelerazione, possiamo modellizzare il consumo di benzina nel seguente modo

$$\int_{t_0}^{t_f} [k_1 u_1(t) + k_2 x_2(t)] dt \leq B \quad (14.24)$$

**Definizione 14.1** Una funzione di controllo  $u(t)$ , che verifica durante l'intervallo  $[t_0, t_f]$  i vincoli, è chiamata controllo ammissibile (*feasible*).

**Definizione 14.2** Una funzione di stato  $x(t)$  che soddisfa i vincoli imposti alle traiettorie su  $[t_0, t_f]$  è chiamata una traiettoria ammissibile (*feasible*).

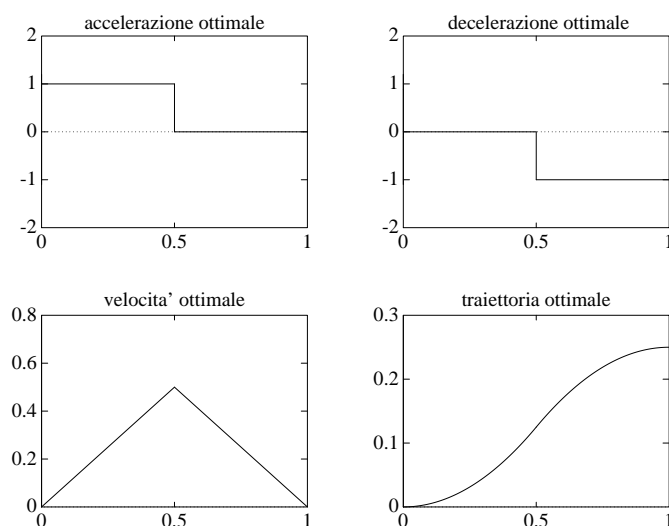


Figura 14.7: Problema di controllo di distanza fissata in tempo minimo.

**Definizione 14.3** Si chiama insieme bersaglio (*target set*) la regione  $\mathcal{B}$  dello spazio-tempo (a  $n + 1$  dimensioni) che rappresenta lo stato finale del sistema.

Nell'esempio che stiamo esaminando, se fissiamo il punto  $e$  e lasciamo libero il punto  $t_f$ , l'insieme  $\mathcal{B}$ , rappresentato nella Figura 14.6(a), è la linea  $x_1 = e$ ;  $x_2 = 0$ ;  $t_0 \leq t_f \leq +\infty$ ; in Figura 14.6(b) è rappresentato l'insieme bersaglio relativo al problema nel quale si richiede che l'automobile arrivi con velocità nulla nell'intervallo  $[e - \bar{e}, e + \bar{e}]$ .

**Funzione obiettivo** Per terminare la formulazione del problema di controllo rimane da definire il funzionale da ottimizzare. Per l'esempio che stiamo esaminando, supponendo che l'obiettivo sia quello di far raggiungere all'automobile il punto  $e$  il più rapidamente possibile, si ha

$$J(\mathbf{x}, \mathbf{u}) = \int_{t_0}^{t_f} ds = t_f - t_0 \quad (14.25)$$

ove la scrittura in termini di integrale evidenzia l'*additività* del funzionale  $J$ .

**Problema di controllo** In definitiva, il problema di controllo per l'esempio considerato consiste nella ricerca della funzione  $\mathbf{u}^*(t)$  nell'insieme  $\mathcal{U}$  dei controlli ammissibili, che verificano cioè i vincoli (14.23) e che generano traiettorie  $\mathbf{x}(t)$  ammissibili, ossia che verificano i vincoli (14.21), (14.22), (14.24), in modo da avere

$$J(\mathbf{x}^*, \mathbf{u}^*) \leq J(\mathbf{x}, \mathbf{u}) \quad \forall \mathbf{u} \in \mathcal{U}$$

ove  $\mathbf{x}^*$  è la traiettoria corrispondente al controllo ottimo  $\mathbf{u}^*$ .

È interessante analizzare la soluzione del problema di controllo ora formulato nell'ipotesi che si abbia  $M_1 = M_2 = M$  e non si tenga conto del vincolo (14.24). La soluzione, che può essere ottenuta facilmente procedendo in maniera analitica utilizzando un ragionamento di simmetria<sup>1</sup>, è illustrata in Figura 14.7. Sottolineiamo, in particolare, la forma del controllo ottimale; esso assume *solo* i valori degli estremi che definiscono l'insieme di ammissibilità, ossia non ci sono valori intermedi. Tale forma, chiamata nelle applicazioni di ingegneria *controllo bang-bang* (on-off, relay), è caratteristica, come vedremo più avanti, dei problemi di controllo nei quali il sistema di stato è lineare nella funzione  $\mathbf{u}$ , ossia della forma

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t)$$

con  $\mathbf{a}$  e  $\mathbf{B} \in \mathbb{R}^{n \times m}$  funzioni assegnate, e per i quali la funzione controllo è sottoposta a vincoli di tipo bilaterale, ossia

$$M_{i-} \leq u_i(t) \leq M_{i+}, \quad i = 1, 2, \dots, m, \quad t \in [t_0, t_f]$$

ove  $M_{i-}$ ,  $M_{i+}$  sono le limitazioni assegnate. ■

<sup>1</sup>intorno al punto  $\bar{t} = (t_0 + t_f)/2$ ; il valore di  $t_f$  è allora ottenuto risolvendo le equazioni di stato e imponendo le condizioni ai limiti.



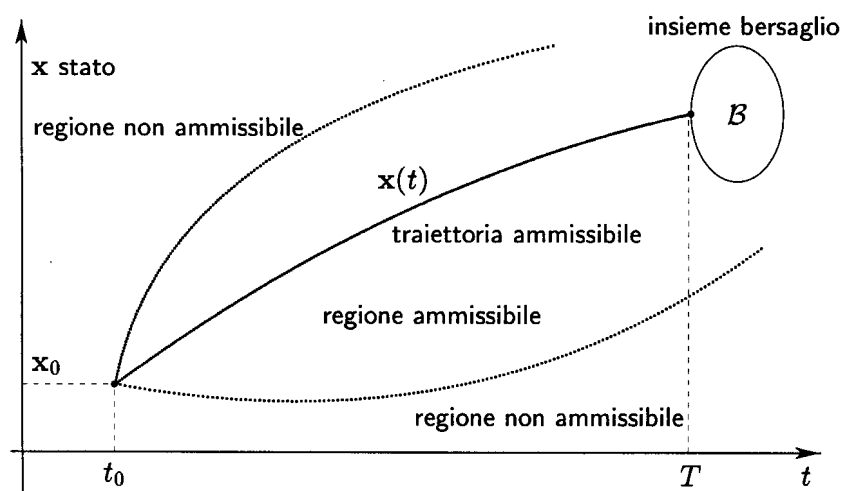


Figura 14.8: Rappresentazione schematica di un problema di controllo.

## 14.2 Formulazione di un problema di controllo

I problemi di controllo che analizzeremo nel seguito corrispondono alla seguente formulazione generale. È dato un *processo di decisioni sequenziali*, ossia un sistema dinamico definito da una relazione ricorrente se l'insieme  $\mathcal{I}$  degli istanti delle decisioni è discreto (ossia  $\mathcal{I} \subset \mathbb{Z}$ ) o da un sistema di equazioni differenziali ordinarie (o più in generale a derivate parziali) se  $\mathcal{I} \subset (t_0, +\infty)$

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{f}(\mathbf{x}(t); \mathbf{u}(t), t) \\ \frac{d\mathbf{x}(t)}{dt} &= \mathbf{f}(\mathbf{x}(t); \mathbf{u}(t), t) \end{aligned} \quad (14.26)$$

ove

1.  $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$  è la variabile di *stato*, la *traiettoria*;  $\mathcal{X}$  è l'insieme degli stati ammissibili (*feasible*).
2.  $\mathbf{u} \in \mathcal{U}_t \subset \mathbb{R}^m$  è il *comando*, la *variabile di decisione*, il *controllo*;  $\mathcal{U}_t$  è, per ogni  $t$ , l'insieme dei controlli ammissibili. L'insieme  $\mathcal{U}_t$ , denotato nel seguito, per brevità, semplicemente con  $\mathcal{U}$ , è supposto un insieme convesso e chiuso.

Le condizioni *iniziali* e *finali* sono precisate nella seguente maniera

- L'istante iniziale  $t_0$  e lo stato iniziale  $\boldsymbol{\xi}$  sono fissati

$$\mathbf{x}(t_0) = \boldsymbol{\xi} \quad (14.27)$$

- Lo stato e l'istante finale sono definiti come il primo punto (se esiste) (cfr. Figura 14.8) della traiettoria del sistema tale che

$$(\mathbf{x}(T), T) \in \mathcal{B} \subset \mathcal{X} \times \mathcal{I} \quad (14.28)$$

ove  $\mathcal{B}$  è chiamato *insieme bersaglio* (*target set*).

Infine, è dato un *criterio additivo*, ossia un funzionale  $J(\mathbf{u})$  della seguente forma

$$J(\mathbf{u}) = \sum_{s=t_0}^{T-1} L(\mathbf{x}(s), \mathbf{u}(s), s) + \lambda(\mathbf{x}(T), T) \quad \text{caso discreto} \quad (14.29)$$

$$J(\mathbf{u}) = \int_{t_0}^T L(\mathbf{x}(s), \mathbf{u}(s), s) ds + \lambda(\mathbf{x}(T), T) \quad \text{caso continuo} \quad (14.30)$$

Il termine  $\lambda(\mathbf{x}(T), T)$  è anche detto, in particolare nelle applicazioni economiche, *terminal payoff*.

Un *problema di controllo* (o di *ottimizzazione dinamica*) consiste allora nella ricerca di una successione di comandi  $\{\mathbf{u}^*(t_0), \mathbf{u}^*(t_0+1), \dots, \mathbf{u}^*(T-1)\}$ , rispettivamente una funzione  $\{\mathbf{u}^*(s), t_0 \leq s \leq T\}$ , che *minimizza* il criterio  $J(\mathbf{u})$  nell'insieme dei comandi ammissibili che trasferiscono il sistema da  $(\boldsymbol{\xi}, t_0)$  a  $(\mathbf{x}(T), T) \in \mathcal{B}$  con traiettorie *ammissibili*. Una funzione  $\mathbf{u}^*$  che realizza tale minimo è chiamata un *controllo ottimale* e la corrispondente traiettoria  $\mathbf{x}^*$  una *traiettoria ottimale*.

Il problema di controllo ora formulato pone diverse questioni, la prima delle quali riguarda la *controllabilità*, ossia l'esistenza di almeno un controllo ammissibile che trasferisca il sistema dallo stato iniziale allo stato finale con una traiettoria ammissibile. La controllabilità di un sistema è, in generale, un problema matematico di non facile soluzione, in quanto comporta lo studio dell'esistenza della soluzione di problemi ai limiti (cfr. Esercizio 14.7 per un esempio). Rinviando per un opportuno approfondimento alla bibliografia, nel seguito assumeremo che i problemi considerati siano controllabili. Inoltre, supporremo che le funzioni  $\mathbf{f}$ ,  $L$ ,  $\lambda$  e la varietà  $\mathcal{B}$  siano sufficientemente regolari, ad esempio derivabili, con derivate prime continue.

La seconda questione importante riguarda l'*esistenza* di un controllo ottimo. Ricordiamo che *l'esistenza di un controllo ottimo non è automaticamente assicurata*, (cfr. Esercizio 14.9) nemmeno nel caso in cui il funzionale  $J(\mathbf{u})$  sia limitato. L'esistenza di un controllo ottimo dipende, in effetti, sia dal funzionale  $J(\mathbf{u})$  che dall'insieme di ammissibilità su cui il funzionale è minimizzato.

### 14.2.1 Forme diverse di un controllo ottimo

Il controllo ottimo  $\mathbf{u}^*(t)$  può essere cercato nelle due forme seguenti, illustrate in maniera schematica in Figura 14.9.

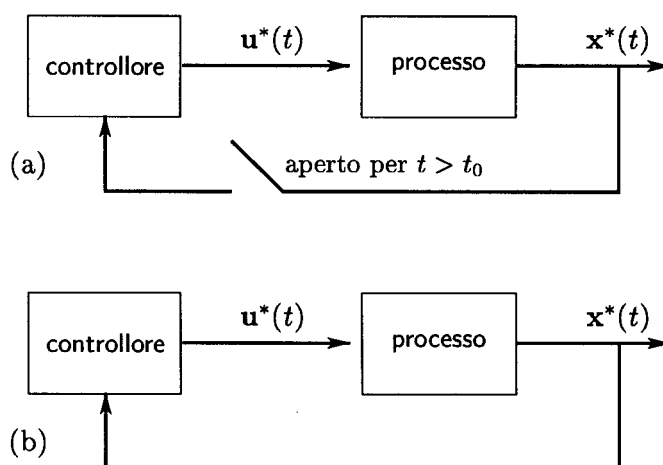


Figura 14.9: (a) Controllo aperto; (b) Controllo chiuso (feedback).

1. Il controllo ottimo è una funzione del tempo per uno *stato iniziale specificato*, cioè

$$\mathbf{u}^*(t) = \mathbf{u}^*(\mathbf{x}(t_0), t) \quad (14.31)$$

Il controllo è detto allora in forma aperta (*open-loop*).

2. Il controllo ottimo dipende ad ogni tempo  $t$  dalle informazioni sullo stato  $\mathbf{x}(t)$  in cui il sistema si trova al tempo  $t$ , ossia

$$\mathbf{u}^*(t) = \mathbf{u}^*(\mathbf{x}(t), t) \quad (14.32)$$

In questo caso il controllo è detto in forma chiusa o *feedback* (retroazione).

In altre parole, il controllo in forma aperta è ottimale solo per un *particolare* stato iniziale ed è determinato completamente da un processo esterno, mentre un controllo in forma chiusa permette di generare un controllo ottimale a partire da un qualsiasi stato iniziale. Come semplice esemplificazione<sup>2</sup> si consideri il riscaldamento di un ambiente mediante una caldaia a temperatura regolabile. Un controllo in forma chiusa corrisponde a regolare la temperatura della caldaia in base alle rilevazioni

<sup>2</sup>Probabilmente, uno degli esempi più antichi di controllo feedback è il ben noto meccanismo di controllo del livello del liquido in un recipiente mediante una valvola galleggiante, attribuito a *Κτεσιβιος* (terzo secolo A. C.). In termini matematici, se  $x(t)$  rappresenta il livello e  $u(t)$  la velocità di flusso all'interno del recipiente, il fenomeno è descritto dall'equazione differenziale  $\dot{x}(t) = u(t)$ . Supponendo di voler mantenere il livello ad un valore  $\bar{x}$  assegnato, basterà assumere  $u(t) = \alpha(\bar{x} - x(t))$ , ove  $\alpha$  è una costante opportuna dipendente dalla pressione del liquido in ingresso e dalle dimensioni del condotto. La soluzione dell'equazione differenziale  $\dot{x}(t) = \alpha(\bar{x} - x(t))$  aumenta se  $\bar{x} > x(t)$  ed ha come punto stazionario il valore  $\bar{x}$ .

di un termostato posto nell'ambiente. Un controllo in forma aperta corrisponde a fissare a priori, solo sulla base dello stato iniziale della temperatura nell'ambiente, la legge di variazione della temperatura della caldaia. Il controllo in forma chiusa è in grado di reagire, diversamente dal controllo in forma aperta, ad eventi imprevisti, quali ad esempio un improvviso abbassamento della temperatura nell'ambiente per l'apertura di una finestra.

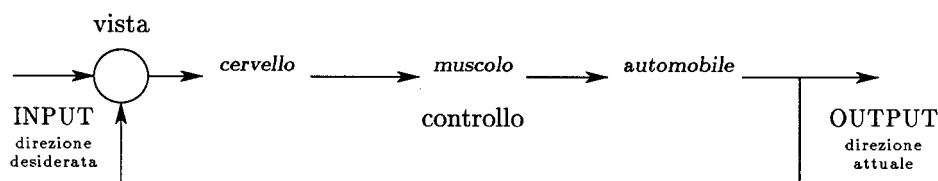


Figura 14.10: Problema feedback di guida.

Un esempio familiare di controllo feedback è rappresentato in Figura 14.10, ove è schematizzato il problema della guida di un'automobile. L'obiettivo è quello di mantenere una direzione desiderata: la vista controlla gli errori tra questa direzione e l'attuale; attraverso il cervello, il controllo passa ai muscoli e il segnale viene amplificato dalla macchina.

In effetti, il controllo feedback è uno dei più importanti fenomeni che avvengono negli organismi viventi, essendo alla base di numerosi processi di autoregolazione (*omeostasi*<sup>3</sup>).

A solo scopo di esemplificazione e rinviando ad esempio a Eisen [53] e Swan [150] per altre interessanti applicazioni, in Figura 14.11 è schematizzato il sistema di regolazione della pressione arteriosa (p. a.) mediante rilascio dell'ormone renina da parte del rene. Se  $v_1$  e  $v_2$  sono due generiche variabili che possono influenzarsi, il simbolo  $v_1 \rightarrow v_2$  indica che  $v_1$  e  $v_2$  cambiano nella stessa direzione, ossia se, ad esempio  $v_1$  aumenta, anche  $v_2$  aumenta. Al contrario, il simbolo  $v_1 - - \rightarrow v_2$  indica che  $v_1$  e  $v_2$  cambiano in direzione opposta. Si vede allora che il ciclo in figura rappresenta un *feedback negativo*: un aumento della pressione arteriosa comporta una diminuzione nella produzione della renina, da cui una diminuzione del fluido extracellulare e una diminuzione della pressione.

<sup>3</sup>The living being is an agency of such sort that each disturbing influence induces by itself the calling forth of compensatory activity to neutralize or repair the disturbance. The higher in the scale of living beings, the more numerous, the more perfect and the more complicated do these regulatory agencies become. They tend to free the organism completely from the unfavorable influences and changes occurring in the environment, Fredericq, 1885. Il termine *omeostasi* venne introdotto da Cannon (1929) per indicare *coordinated physiological reactions which maintain most of the steady states of the body*.

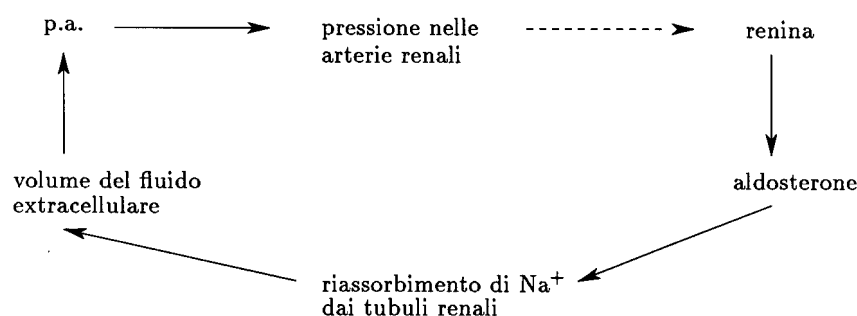


Figura 14.11: Regolazione della pressione arteriosa (p.a.) mediante un controllo feedback.

## 14.3 Metodo della programmazione dinamica

Il metodo della *programmazione dinamica*, sviluppato in particolare da Bellman<sup>4</sup>, è un metodo numerico per il calcolo del controllo ottimale in forma chiusa. Nel seguito il metodo verrà introdotto in relazione ad un problema di controllo discreto, mentre l'estensione al problema continuo sarà ottenuta mediante un opportuno passaggio al limite.

### 14.3.1 Principio di ottimalità

L'idea di partenza del metodo consiste nel pensare il problema di controllo relativo al valore iniziale  $\mathbf{x} = \boldsymbol{\xi}$ , per  $t = t_0$ , come un caso particolare di una *famiglia* di problemi di controllo ottenuti assumendo come punto iniziale un generico punto  $(\mathbf{x}, t)$ . Indicando, per evitare confusione di notazioni, con  $\mathbf{z}(s)$  la variabile di stato, abbiamo quindi di variare di  $(\mathbf{x}, t) \in \mathcal{X} \times \mathcal{I}$ , i seguenti problemi

$$\left\{ \begin{array}{l} \mathbf{z}(s+1) = \mathbf{f}(\mathbf{z}(s), \mathbf{u}(s), s), \quad \mathbf{u}(s) \in \mathcal{U}(s) \\ s = t, t+1, \dots, T-1 \\ \mathbf{z}(t) = \mathbf{x} \\ (\mathbf{z}(T), T) \in \mathcal{B} \\ J_{(\mathbf{x}, t)}(\mathbf{u}) := \sum_{s=t}^{T-1} L(\mathbf{z}(s), \mathbf{u}(s), s) + \lambda(\mathbf{z}(T), T) \end{array} \right. \quad (14.33)$$

<sup>4</sup>R. E. Bellman, *On the theory of dynamic programming*. Proc. Natl. Acad. Sci, USA 38, 1952.

In corrispondenza a ciascuno dei problemi (14.33), definiamo la seguente funzione delle variabili  $\mathbf{x}, t$

$$(\mathbf{x}, t) \rightarrow V(\mathbf{x}, t) := \min_{\mathbf{u} \in \mathcal{U}} J_{(\mathbf{x}, t)}(\mathbf{u})$$

nota come *optimal-return function* (o anche, in particolare nelle applicazioni economiche, *payoff function*). Sottolineiamo che tale funzione rappresenta il valore del funzionale costo  $J$  per il problema di controllo relativo al punto iniziale  $(\mathbf{x}, t)$ . L'esistenza della funzione  $V(\mathbf{x}, t)$  richiede pertanto un'ipotesi *supplementare*, ossia che tutti i problemi, al variare dei valori iniziali  $(\mathbf{x}, t)$ , abbiano una soluzione, mentre la soluzione del problema di partenza richiederebbe a priori soltanto l'esistenza di  $V(\xi, t_0)$ . Ne segue che le *condizioni* che otterremo nel seguito saranno, in generale, *solo sufficienti*. Tuttavia, nel caso del controllo relativo a un sistema di stato lineare e funzionale costo quadratico saremo in grado di dimostrare che la condizione ottenuta mediante la programmazione dinamica è anche necessaria.

Alla base del metodo della programmazione dinamica vi è il *principio di ottimalità*, valido per i funzionali  $J$  *additivi*, e ben noto nella *meccanica*<sup>5</sup>. Introduciamo il principio con una semplice considerazione. Con riferimento alla Figura 14.12, sia, per ipotesi,  $a^* \rightarrow b^* \rightarrow e^*$  la traiettoria che rende minimo un funzionale additivo  $J_{a^* e^*}$  relativo ad un processo che ha come stato iniziale  $a^*$  e stato finale  $e^*$ . Consideriamo, quindi, un processo che abbia come stato iniziale  $b^*$ , che appartiene alla precedente traiettoria ottimale, e come stato finale ancora  $e^*$ . Allora la traiettoria che rende minimo il funzionale  $J_{b^* e^*}$  è ancora  $b^* \rightarrow e^*$ . Se infatti esistesse una traiettoria, diciamo,  $b^* \rightarrow c^* \rightarrow e^*$  con costo  $J_{b^* c^* e^*} < J_{b^* e^*}$ , si avrebbe, per l'additività

$$J_{a^* e^*} = J_{a^* b^*} + J_{b^* e^*} > J_{a^* b^*} + J_{b^* c^* e^*}$$

e alla traiettoria  $a^* \rightarrow b^* \rightarrow c^* \rightarrow e^*$  corrisponderebbe un costo inferiore, contrariamente all'ipotesi che  $a^* \rightarrow b^* \rightarrow e^*$  fosse la traiettoria ottimale.

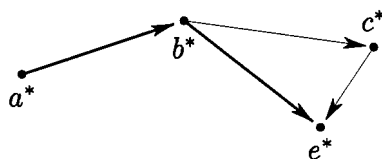


Figura 14.12: Processo additivo; due possibili cammini ottimali da  $b^*$  a  $e^*$ .

Con riferimento ai problemi di controllo, al principio di ottimalità è stata data da Bellman la seguente formulazione che illustreremo successivamente su opportuni esempi.

<sup>5</sup>ove assume la seguente forma: *Toute courbe qui doit donner un maximum conserve aussi dans toutes ses parties les lois de ce même maximum* (Bernoulli, 1706).

**Proposizione 14.1** (Bellman, 1957) *An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

Ricordando che la funzione  $V(\mathbf{x}, t)$  rappresenta il valore del funzionale costo  $J_{(\mathbf{x}, t)}$  per il problema di controllo che ha come stato iniziale al tempo  $t$  lo stato  $\mathbf{x}$ , si ha per il principio di ottimalità la seguente *equazione ricorrente del primo ordine*

$$\begin{aligned} V(\mathbf{x}, t) &= \min_{\mathbf{u}(t), \mathbf{u}(t+1), \dots} \left\{ \sum_{s=t}^{T-1} L(\mathbf{z}, \mathbf{u}, s) + \lambda(\mathbf{z}(T), T) \right\} \\ &= \min_{\mathbf{u}(t)} \left\{ L(\mathbf{x}, \mathbf{u}(t), t) + \min_{\mathbf{u}(t+1), \mathbf{u}(t+2), \dots} \left[ \sum_{s=t+1}^{T-1} L(\mathbf{z}(s), \mathbf{u}(s), s) + \lambda(\mathbf{z}(T), T) \right] \right\} \end{aligned}$$

ossia l'equazione

$$V(\mathbf{x}, t) = \min_{\mathbf{u} \in \mathcal{U}} \{ L(\mathbf{x}, \mathbf{u}, t) + V(\mathbf{f}(\mathbf{x}, \mathbf{u}, t), t+1) \} \quad (14.34)$$

che è nota come *equazione di Hamilton-Jacobi-Bellman*. Supponendo per semplicità  $T$  fissato, per  $t = T$  si ottiene l'uguaglianza

$$V(\mathbf{x}, T) = \lambda(\mathbf{x}, T) \quad (14.35)$$

che fornisce il termine di partenza per la procedura ricorrente (14.34).

Riassumendo, l'applicazione del metodo della programmazione dinamica comporta i seguenti passi:

- si risolve l'equazione funzionale (14.34) a partire dalla condizione iniziale (14.35);
- il controllo ottimale  $\mathbf{u}^*(\mathbf{x}, t)$  è dato ad ogni tempo  $t$  dall'argomento delle minimizzazioni (14.35) e (14.34).

Il controllo ottimale è fornito in forma chiusa.

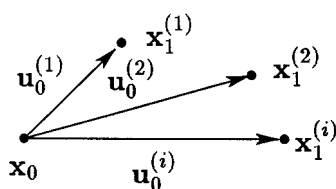
Per chiarire ulteriormente la natura del metodo, consideriamo il seguente semplice esempio.

► **Esempio 14.5** Per un processo a due stati si ha

$$J(\mathbf{u}_0, \mathbf{u}_1) = J_0 + J_1 = L(\mathbf{x}_0, \mathbf{u}_0, t_0) + L(\mathbf{x}_1, \mathbf{u}_1, t_1) \quad \text{ove} \quad \mathbf{x}_1 = \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, t_0)$$

Il problema di controllo consiste nel calcolare i vettori  $\mathbf{u}_0, \mathbf{u}_1$  che minimizzano la funzione  $J(\mathbf{u}_0, \mathbf{u}_1)$  con il vincolo dell'equazione di stato  $\mathbf{x}_1 = \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0)$ .

Supponiamo ora di risolvere, al tempo  $t_1$  e per un generico  $\mathbf{x}_1$  il problema del minimo di  $L(\mathbf{x}_1, \mathbf{u}_1, t_1)$ , rispetto a  $\mathbf{u}_1$ . Memorizziamo, quindi, il valore del minimo  $J_1^*(\mathbf{x}_1)$ , insieme all'argomento del minimo  $\mathbf{u}_1^*(\mathbf{x}_1)$ .



$$\mathbf{x}_1^{(i)} = \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0^{(i)}, t_0)$$

$$J^* = \min_{\mathbf{u}_0^{(i)}} L(\mathbf{x}_0, \mathbf{u}_0^{(i)}, t_0) + J_1^*(\mathbf{x}_1^{(i)})$$

Figura 14.13: Rappresentazione schematica di un passaggio del metodo di programmazione dinamica.

L'applicazione del metodo prevede a questo punto la minimizzazione del funzionale

$$L(\mathbf{x}_0, \mathbf{u}_0, t_0) + J_1^*(\mathbf{x}_1)$$

rispetto a  $\mathbf{u}_0$ . In Figura 14.13 sono rappresentate le diverse traiettorie  $\mathbf{x}_1^{(i)}$  corrispondenti ai controlli  $\mathbf{u}_0^{(i)}$ ,  $i = 1, 2, \dots$ . Le quantità  $J_1^*(\mathbf{x}_1^{(i)})$  forniscono il costo ottimale relativo ad ogni stato  $\mathbf{x}_1^{(i)}$ . Come si vede, il *guadagno* del metodo consiste nel ricondurre un problema di ottimo *globale*, sull'intervallo  $[t_0, t_1]$  e rispetto ai due vettori  $\mathbf{u}_0, \mathbf{u}_1$ , alla successione dei due problemi di ottimo in  $t_1$  e in  $t_0$ , rispettivamente nella variabile  $\mathbf{u}_1$  e  $\mathbf{u}_2$ . Naturalmente, questo è possibile grazie ad un maggiore utilizzo della memoria: ad ogni stadio del processo sono memorizzati i valori ottimali della funzione obiettivo in corrispondenza ad ogni stato ammissibile. Questo fatto rappresenta, quando il numero degli stati ammissibili è elevato, una possibile limitazione all'applicabilità del metodo (*curse of dimensionality*). ■

Esamineremo ora alcune applicazioni significative del metodo.

► **Esempio 14.6** (*Impiego ottimale di una risorsa*) Appliciamo il metodo della programmazione dinamica al problema introdotto nell'Esempio 14.1 riguardante l'impiego ottimale di una risorsa in due differenti modi. In forma riassuntiva, il modello è il seguente

$$x(t+1) = 0.5x(t) + 0.3u(t), \quad 0 \leq u(t) \leq x(t) \quad (\text{Equazione di stato})$$

$$J(t, u) = 2u(t) + 3(x(t) - u(t)) = 3x(t) - u(t)$$

$$\max_u \sum_{t=0}^{T-1} J(t, u)$$

In questo caso possiamo porre  $V(x, T) = 0$ ; il primo termine non triviale è  $V(x, T-1)$ . Si ha pertanto

$$V(x, T-1) = \max_{0 \leq u \leq x} \{3x - u\} = 3x \quad \rightarrow \quad u^*(x, T-1) = 0$$

$$\begin{aligned} V(x, T-2) &= \max_{0 \leq u \leq x} \{3x - u + 3(0.5x + 0.3u)\} = \\ &= \max_{0 \leq u \leq x} \{4.5x - 0.1u\} = 4.5x \quad \rightarrow \quad u^*(x, T-2) = 0 \end{aligned}$$

$$\begin{aligned} V(x, T-3) &= \max_{0 \leq u \leq x} \{3x - u + 4.5(0.5x + 0.3u)\} = \\ &= \max_{0 \leq u \leq x} \{5.25x + 0.35u\} = 5.60x \quad \rightarrow \quad u^*(x, T-3) = x \end{aligned}$$



Se consideriamo ad esempio un problema a 3 stati con stato iniziale  $x(0) = \xi$ , il comando ottimale è dato da

$$\begin{aligned} u^*(0) &= x(0) \\ u^*(1) &= 0 \\ u^*(2) &= 0 \end{aligned}$$

a cui corrisponde il valore del costo ottimale  $J^* = 5.6\xi$ . ■

► **Esempio 14.7** (*Problema di cammino ottimale*) Con riferimento alla Figura 14.14 che rappresenta lo schema di una rete, a partire dal *nodo iniziale* **a** è possibile arrivare al *nodo finale* **h** seguendo differenti percorsi. Il costo di ogni percorso è ottenuto sommando i costi relativi ai singoli segmenti che compongono il percorso. Ad esempio, il percorso  $\mathbf{a} \rightarrow \mathbf{b} \rightarrow \mathbf{c} \rightarrow \mathbf{f} \rightarrow \mathbf{g} \rightarrow \mathbf{h}$  ha costo 22, mentre il percorso  $\mathbf{a} \rightarrow \mathbf{d} \rightarrow \mathbf{e} \rightarrow \mathbf{h}$  ha costo 19. Si tratta di trovare il cammino (o eventualmente i cammini) a costo minimo. Dal momento che

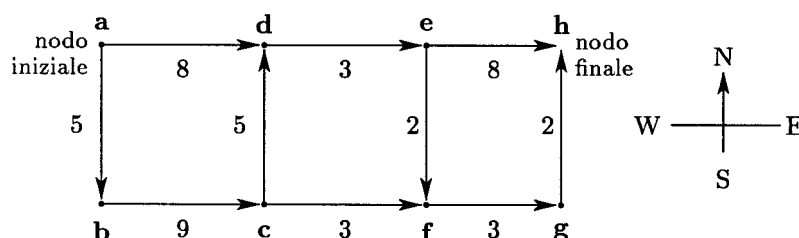


Figura 14.14: Esempio di una rete.

i cammini possibili sono in numero finito, i cammini a costo ottimale possono, naturalmente, essere individuati mediante una ricerca di minimo tra i costi di tutti i possibili cammini. Un'alternativa più efficiente è, tuttavia, fornita dal metodo della programmazione dinamica. Mediante tale metodo è possibile, come vedremo, associare ad ogni nodo della rete una scelta ottimale della direzione, ossia, nella terminologia dei problemi di controllo, fornire il controllo in forma chiusa. Al contrario, la soluzione basata sulla completa enumerazione dei cammini possibili fornisce il controllo in forma aperta, in quanto individua solo i particolari cammini ottimali che hanno **a** come nodo iniziale e **h** come nodo finale.

Il problema precedente può essere inquadrato nella formulazione generale dei problemi di controllo nel seguente modo. Si assume come variabile di *stato* la posizione nei vari nodi **a**, **b**, **c**, ... La *decisione* (il controllo) corrisponde alla scelta, in ogni nodo, di una delle quattro direzioni *N*, *E*, *S*, *W*. Come si vede dalla figura, non tutte le direzioni possono essere ammissibili. Ad esempio, dal nodo **c** è solo possibile andare in **d** o in **f**. Tali direzioni obbligate definiscono in ogni nodo l'insieme  $\mathcal{U}$  dei controlli ammissibili.

Indichiamo con  $J_{cd}$  il costo per andare da **c** in **d** e analogamente con  $J_{cf}$  il costo da **c** in **f**. Se si suppone di conoscere *costi ottimali*  $J_{dh}^*$  e  $J_{fh}^*$  per raggiungere il nodo finale **h** a partire rispettivamente dal nodo **d** e dal nodo **f** (nell'esempio,  $J_{dh}^* = 10$ ,  $J_{fh}^* = 5$ ), il minimo costo  $J_{ch}^*$  per raggiungere **h** da **c** è allora dato da

$$J_{ch}^* = \min \begin{cases} C_{cdh}^* = J_{cd} + J_{dh}^* = \text{minimo costo per raggiungere } \mathbf{h} \text{ da } \mathbf{c} \text{ via } \mathbf{d} \\ C_{cff}^* = J_{cf} + J_{fh}^* = \text{minimo costo per raggiungere } \mathbf{h} \text{ da } \mathbf{c} \text{ via } \mathbf{f} \end{cases}$$

$\alpha$	$u_i$	$\mathbf{x}_i$	$C_{\alpha\mathbf{x}_i\mathbf{h}}^*$	$J_{\alpha\mathbf{h}}^*$	$u^*(\alpha)$
<b>g</b>	N	<b>h</b>	2	2	N
<b>f</b>	E	<b>g</b>	5	5	E
<b>e</b>	E	<b>h</b>	8		
	S	<b>f</b>	7	7	S
<b>d</b>	E	<b>e</b>	10	10	E
<b>c</b>	N	<b>d</b>	15		
	E	<b>f</b>	8	8	E
<b>b</b>	E	<b>c</b>	17	17	E
<b>a</b>	E	<b>d</b>	18	18	E
	S	<b>b</b>	22		

Tabella 14.2: Decisioni ottimali ottenute mediante il metodo di programmazione dinamica nel problema di percorso a minimo costo.

Nel caso dell'esempio si ha  $J_{\mathbf{ch}}^* = \min\{15, 8\} = 8$  e la decisione ottimale da assumere nel nodo **c** corrisponde ad andare in **f**. Nel caso generale, i valori  $J_{\mathbf{dh}}^*$  e  $J_{\mathbf{fh}}^*$  possono essere calcolati in maniera ricorrente mediante l'applicazione del principio di ottimalità. A tale scopo, introduciamo le seguenti notazioni.

- $\alpha$  è lo stato corrente.
- $u_i$  sono i controlli ammissibili nello stato  $\alpha$ , ossia un sottoinsieme delle quattro direzioni possibili  $\{N, E, S, W\}$ .
- $\mathbf{x}_i$  è lo stato (il nodo) raggiungibile da  $\alpha$  mediante l'applicazione di  $u_i$  in  $\alpha$ .
- $J_{\alpha\mathbf{x}_i}$  è il costo per andare da  $\alpha$  a  $\mathbf{x}_i$ .
- $J_{\mathbf{x}_i\mathbf{h}}^*$  è il *costo minimo* per raggiungere lo stato finale **h** da  $\mathbf{x}_i$ .
- $C_{\alpha\mathbf{x}_i\mathbf{h}}^*$  è il costo minimo per andare da  $\alpha$  a **h** via  $\mathbf{x}_i$ .
- $J_{\alpha\mathbf{h}}^*$  è il costo minimo per andare da  $\alpha$  a **h** (rispetto ad ogni percorso ammissibile).
- $u^*(\alpha)$  è la *decisione ottimale* in  $\alpha$ .

Con le notazioni precedenti, il principio di ottimalità implica che

$$C_{\alpha\mathbf{x}_i\mathbf{h}}^* = J_{\alpha\mathbf{x}_i} + J_{\mathbf{x}_i\mathbf{h}}^*$$

e la decisione ottimale in  $\alpha$ , ossia  $u^*(\alpha)$ , è la decisione che corrisponde al seguente minimo

$$J_{\alpha\mathbf{h}}^* = \min \{C_{\alpha\mathbf{x}_1\mathbf{h}}^*, C_{\alpha\mathbf{x}_2\mathbf{h}}^*, \dots, C_{\alpha\mathbf{x}_i\mathbf{h}}^*, \dots\}$$

ove il minimo è effettuato su tutti i controlli ammissibili  $u_i$  nello stato  $\alpha$ . Il metodo permette allora di costruire la Tabella 14.2, che fornisce per il problema assegnato la soluzione in *forma chiusa*. In particolare, le ultime due colonne forniscono per ogni nodo  $\alpha$  rispettivamente il costo e la decisione ottimali per andare da  $\alpha$  al nodo finale **h**.

Per analizzare il guadagno in termini di operazioni del metodo della programmazione dinamica nei confronti del metodo consistente nella completa enumerazione dei percorsi, consideriamo il semplice problema di percorso rappresentato nella Figura 14.15. In ogni nodo vi sono soltanto due possibili scelte, indicate in figura da vettori. Ogni percorso da A in B è quindi composto dallo stesso numero  $N$  di stadi. Il numero totale dei possibili

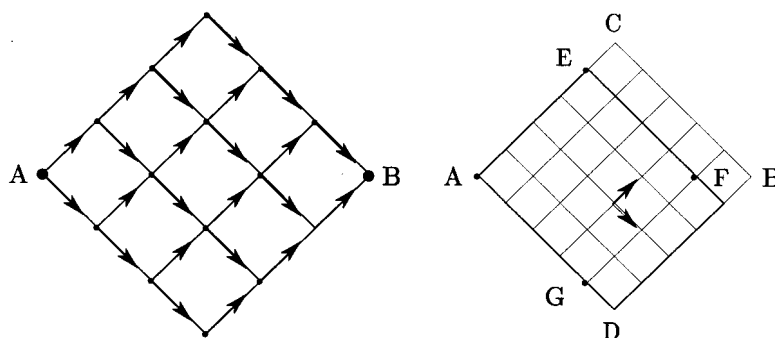


Figura 14.15: Percorsi a  $N=6$  e rispettivamente  $N=10$  stadi.

percorsi da A a B è quindi dato da  $\binom{N}{N/2}$ . In effetti basta osservare che, indicando con  $D$  la direzione secondo la diagonale verso il basso e con  $U$  la direzione verso l'alto, ogni percorso può essere rappresentato da una sequenza di  $N$  simboli, metà dei quali sono  $D$  e l'altra metà sono  $U$ . Pertanto, il numero dei percorsi corrisponde al numero delle combinazioni di  $N$  in gruppi di  $N/2$ . Per gli esempi dati in figura si ha

$$N = 6 \quad \text{numero totale dei percorsi} = \binom{N}{N/2} = 20$$

$$N = 10 \quad \text{numero totale dei percorsi} = \binom{N}{N/2} = 252$$

Il calcolo del costo di ogni percorso richiede  $N - 1$  addizioni e il calcolo del costo ottimale richiede  $\binom{N}{N/2} - 1$  confronti. In totale, quindi, il metodo basato sulla completa enumerazione richiede

$$(N - 1) \binom{N}{N/2} \text{ addizioni,} \quad \binom{N}{N/2} - 1 \text{ confronti}$$

Per il calcolo delle operazioni richieste dal metodo della programmazione dinamica, osserviamo che vi sono  $N$  nodi (nella figura, i nodi sulle linee CB e DB, escluso il nodo B) nei quali il metodo richiede solo una addizione e nessun confronto. Vi sono  $(N/2)^2$  nodi rimanenti (quelli nel rombo AEF G) nei quali sono richiesti due addizioni e un confronto. Pertanto, nel metodo della programmazione dinamica sono richieste

$$\frac{N^2}{2} + N \text{ addizioni,} \quad \frac{N^2}{4} \text{ confronti}$$

Ad esempio, per  $N = 10$  il metodo della programmazione dinamica richiede 60 addizioni e 25 confronti, contro le 2268 addizioni e i 251 confronti richiesti dal metodo basato sull'altro

metodo. Il vantaggio cresce comunque esponenzialmente con  $N$ . Ad esempio, per  $N = 20$  il numero delle addizioni nel metodo della programmazione dinamica è 220 ed è 3 510 864 nell'altro metodo. ■

► **Esempio 14.8** (*Un problema di allocazione di risorse*) Supponiamo che una risorsa fissata  $A$  debba essere distribuita tra un numero  $N$  di attività differenti e sia  $g_k(u(k))$  il guadagno ricavato dall'allocazione della quantità  $u(k)$  all'attività  $k$ -ma. Il problema di ottimizzazione è allora quello di massimizzare la funzione

$$g_0(u(0)) + g_1(u(1)) + \cdots + g_{N-1}(u(N-1))$$

sotto il vincolo

$$u(0) + u(1) + \cdots + u(N-1) = A$$

Per un'applicazione in biologia, si pensi ad una distribuzione di una quantità fissata  $A$  di radiazioni ad  $N$  tumori situati in differenti posizioni. In questo caso  $g(u)$  è il numero di cellule cancerose eliminate da una quantità di radiazione  $u$ ; una ipotesi comunemente utilizzata è che sia  $g(u) = u^{1/2}$ . Il problema di ottimizzazione consiste nella ricerca della distribuzione della radiazione corrispondente all'eliminazione del massimo numero di cellule cancerose. Si tratta quindi di massimizzare la funzione

$$\sum_{k=0}^{N-1} u(k)^{1/2} \quad (14.36)$$

con il vincolo

$$\sum_{k=0}^{N-1} u(k) = A \quad (14.37)$$

Tale problema è equivalente al problema di controllo corrispondente alla seguente equazione di stato

$$x(k+1) = x(k) - u(k)$$

con le condizioni, rispettivamente iniziale e finale

$$x(0) = A, \quad x(N) = 0$$

e la funzione obiettivo

$$J = \sum_{k=0}^{N-1} u(k)^{1/2}$$

Tale formulazione si basa sull'ipotesi che l'allocazione sia fatta in maniera sequenziale. Lo stato  $x(k)$  rappresenta la quantità di risorsa disponibile per l'allocazione alle attività da  $k$  a  $N$ .

La funzione return  $V(x, k)$  è il valore ottimale che può essere ottenuto mediante l'allocazione di una quantità  $x$  di risorse tra le ultime  $N - k$  attività. Con riferimento alla funzione obiettivo (14.36), si ha  $V(x, N) = 0$  e

$$V(x, N-1) = x^{1/2}$$

Successivamente, si ha

$$V(x, N-2) = \max_u [u^{1/2} + V(x-u, N-1)] = \max_u [u^{1/2} + (x-u)^{1/2}]$$

Il valore massimo si ha per  $u^* = x/2$  per il quale si ha

$$V(x, N-2) = \sqrt{2x}$$

In modo analogo, per  $N-3$  si ha

$$V(x, N-3) = \max_u [u^{1/2} + \sqrt{2}(x-u)^{1/2}]$$

da cui

$$V(x, N-3) = \sqrt{3x}$$

Più in generale, si ha quindi

$$V(x, N-k) = \sqrt{kx}, \quad u^*(N-k) = \frac{x(N-k)}{k}$$

Ad ogni stadio del processo, si determina la quantità delle risorse che rimangono e si divide per il numero delle risorse che rimangono. Il risultato determina l'allocazione allo stadio corrente. La procedura fornisce quindi la soluzione in forma feedback. Per il problema assegnato con  $x(0) = A$ , si ha  $u(k) = A/N$  per ogni  $k$ . ■

### 14.3.2 Programmazione dinamica nel caso continuo

Il metodo della programmazione dinamica può essere esteso ai problemi di controllo relativi a sistemi continui nel seguente modo. Seguendo la logica utilizzata nel caso continuo, consideriamo la seguente famiglia di problemi di controllo, al variare del punto iniziale  $(\mathbf{x}, t)$

$$\left\{ \begin{array}{l} \frac{d\mathbf{z}(s)}{ds} = \mathbf{f}(\mathbf{z}(s), \mathbf{u}(s), s), \quad \mathbf{u}(s) \in \mathcal{U}(s), \quad t < s < T \\ \mathbf{z}(t) = \mathbf{x} \\ (\mathbf{x}(T), T) \in \mathcal{B} \\ J_{(\mathbf{x}, t)}(\mathbf{u}) := \int_t^T L(\mathbf{z}(s), \mathbf{u}(s), s) ds + \lambda(\mathbf{z}(T), T) \end{array} \right. \quad (14.38)$$

In corrispondenza si ha la funzione *return*

$$(\mathbf{x}, t) \rightarrow V(\mathbf{x}, t) := \min_{\mathbf{u} \in \mathcal{U}} J_{(\mathbf{x}, t)}(\mathbf{u}) \quad (14.39)$$

che rappresenta il valore ottimale della funzionale obiettivo a partire dallo stato  $\mathbf{x}$  all'istante  $t$ . Per ottenere un'equazione utile per il calcolo della funzione  $V(\mathbf{x}, t)$  e di conseguenza il controllo ottimale  $\mathbf{u}^*(t)$  in forma feedback, utilizzeremo i risultati

sviluppati nel paragrafo precedente in relazione al caso discreto. Più precisamente, si discretizza il problema (14.38) supponendo costanti le funzioni di stato e di controllo su ogni intervallo  $[t, t + \delta t]$ , con  $\delta t > 0$ ; le derivate sono allora approssimate da rapporti incrementali e il sistema di stato diventa una relazione ricorrente, mentre l'integrale che definisce il funzionale obiettivo  $J$  si riduce a una somma finita. Applicando il principio di ottimalità, si ha quindi la seguente relazione ricorrente

$$\begin{aligned} V(\mathbf{x}, t) &= \min_{\mathbf{u}(t) \in \mathcal{U}} \left\{ L(\mathbf{x}, \mathbf{u}(t), t) \delta t + \min_{\mathbf{u}(t+\delta t), \dots} \left[ \int_{t+\delta t}^T + \lambda \right] \right\} \\ &= \min_{\mathbf{u}(t) \in \mathcal{U}} \{ L(\mathbf{x}, \mathbf{u}(t), t) \delta t + V(\mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \delta t, t + \delta t) \} \end{aligned}$$

Nell'ipotesi che  $V(\mathbf{x}, t)$  sia una funzione regolare (in altre parole, nell'ipotesi che le funzioni  $\mathbf{f}(\mathbf{x}, t)$ ,  $L(\mathbf{x}, \mathbf{u}, t)$ ,  $\lambda(\mathbf{x}, T)$  siano sufficientemente regolari), si ha

$$V(\mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \delta t, t + \delta t) = V(\mathbf{x}, t) + V_t \delta t + V_x \mathbf{f} \delta t + O(\delta^2)$$

ove con  $V_x$  si è indicato il vettore riga  $[\partial V / \partial x_1, \partial V / \partial x_2, \dots, \partial V / \partial x_n]$ . Passando al limite per  $\delta t \rightarrow 0$  e osservando che  $V(\mathbf{x}, t)$  non dipende da  $\mathbf{u}$ , si ottiene la seguente equazione, nota come *equazione di Hamilton–Jacobi–Bellman* (HJB)

$$\min_{\mathbf{u}(t) \in \mathcal{U}(t)} \{ L(\mathbf{x}, \mathbf{u}, t) + V_x(\mathbf{x}, t) \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \} + V_t(\mathbf{x}, t) = 0 \quad (14.40)$$

Si ha, allora, il seguente risultato.

**Teorema 14.1** (Programmazione dinamica) *Se l'equazione funzionale (HJB)*

$$V_t(\mathbf{x}, t) + \min_{\mathbf{u} \in \mathcal{U}} \{ L(\mathbf{x}, \mathbf{u}, t) + V_x(\mathbf{x}, t) \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \} = 0 \quad (14.41)$$

*ha una soluzione che verifica le condizioni ai limiti*

$$V(\mathbf{x}, T) = \lambda(\mathbf{x}, T) \quad \text{su } \mathcal{B} \quad (14.42)$$

*allora il problema di ottimizzazione dinamica (14.38) ha una soluzione  $\mathbf{u}^*(\mathbf{x}, t)$  data da*

$$\mathbf{u}^*(\mathbf{x}, t) = \text{argomento del } \min_{\mathbf{u} \in \mathcal{U}} \{ L + V_x \mathbf{f} \} \quad (14.43)$$

L'equazione (HJB) è un'equazione alle derivate parziali del primo ordine che richiede ad ogni punto  $(\mathbf{x}, t)$  la risoluzione di un problema di minimo. La sua trattazione risulta, pertanto, semplificata quando tale problema di minimo può essere risolto in forma analitica; è il caso dei sistemi lineari e costo quadratico che analizzeremo nell'esempio successivo. Un modo più sintetico di scrivere l'equazione (HJB) utilizza la seguente funzione, chiamata *funzione hamiltoniana* associata al problema di controllo (14.38)

$$\mathcal{H}(\mathbf{x}, \mathbf{u}, \mathbf{y}, t) := L(\mathbf{x}, \mathbf{u}, t) + \mathbf{y}^T \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \quad (14.44)$$

ove  $\mathbf{y} \in \mathbb{R}^n$  è chiamato *vettore aggiunto*. Posto allora

$$\mathcal{H}^0(\mathbf{x}, \mathbf{y}, t) := \min_{\mathbf{u} \in \mathcal{U}} \mathcal{H}(\mathbf{x}, \mathbf{u}, \mathbf{y}, t)$$

l'equazione (HJB) può essere scritta nella seguente forma

$$V_t + \mathcal{H}^0(\mathbf{x}, V_{\mathbf{x}}^T, t) = 0$$

che è anche nota come *equazione di Hamilton-Jacobi*.

► **Esempio 14.9** (*Sistema lineare e costo quadratico*) Introduciamo il risultato incominciando dal problema particolare in una dimensione corrispondente alla seguente equazione differenziale

$$\dot{x}(t) = x(t) + u(t) \quad (14.45)$$

e al seguente funzionale obiettivo da minimizzare

$$J(u) = \frac{1}{4}x^2(T) + \int_0^T \frac{1}{4}u^2(t) dt \quad (14.46)$$

Il tempo finale  $T$  è fissato e nessun vincolo è posto sulla funzione di stato  $x(t)$  e di controllo  $u(t)$ . La funzione hamiltoniana assume in questo caso la seguente forma

$$\mathcal{H}(x(t), u(t), V_x, t) = \frac{1}{4}u^2(t) + V_x(x(t), t) [x(t) + u(t)] \quad (14.47)$$

e dal momento che il controllo non è vincolato, esso deve verificare la seguente condizione necessaria

$$\frac{\partial \mathcal{H}}{\partial u} = \frac{1}{2}u^*(t) + V_x(x, t) = 0 \quad (14.48)$$

Osserviamo che

$$\frac{\partial^2 \mathcal{H}}{\partial u^2} = \frac{1}{2} > 0$$

e, quindi, il controllo che soddisfa la condizione (14.48) *minimizza*  $\mathcal{H}$ . Dalla (14.48) si ricava

$$u^*(t) = -2V_x(x, t) \quad (14.49)$$

che, sostituito nell'equazione di Hamilton-Jacobi-Bellman, fornisce la seguente equazione

$$0 = V_t + \frac{1}{4}[-2V_x]^2 + V_x x(t) - 2[V_x]^2 = V_t - [V_x]^2 + V_x x(t) \quad (14.50)$$

La condizione ai limiti (14.42) diventa

$$V(x(T), T) = \frac{1}{4}x^2(T) \quad (14.51)$$

L'equazione (14.50) può essere risolta mediante il *metodo di separazione delle variabili*, che consiste nel cercare la soluzione nella seguente forma

$$V(x, t) = \frac{1}{2}P(t)x^2 \quad (14.52)$$

ove  $P(t)$  è una funzione incognita da determinare. Sostituendo in (14.50) le seguenti derivate

$$V_x(x, t) = P(t)x, \quad V_t(x, t) = \frac{1}{2}\dot{P}(t)x^2 \quad (14.53)$$

si ottiene

$$0 = \frac{1}{2}\dot{P}(t)x^2 - P^2x^2 + P(t)x^2 \quad (14.54)$$

Dovendo tale equazione essere verificata per ogni  $x$ , si ottiene la seguente equazione differenziale del primo ordine nella funzione incognita  $P(t)$

$$\frac{1}{2}\dot{P}(t) - P^2 + P(t) = 0 \quad (14.55)$$

che è un caso particolare di *equazione di Riccati* (cfr. Appendice B). Dalla condizione ai limiti (14.51) si ricava inoltre la condizione

$$P(T) = \frac{1}{2} \quad (14.56)$$

Abbiamo in questo modo trovato che la funzione  $P(T)$  è la soluzione del problema a valori iniziali (14.55), (14.56). In questo caso particolare è facile mostrare che la soluzione di tale problema è data dalla seguente funzione

$$P(t) = \frac{e^{(T-t)}}{e^{(T-t)} + e^{-(T-t)}} \quad (14.57)$$

e la legge di controllo ottimale è data da

$$u^*(t) = -2V_x(x, t) = -2P(t)x(t) \quad (14.58)$$

In Figura 14.16 è mostrata la funzione  $P(t)$ , il controllo ottimale  $u^*(t)$  e la traiettoria  $x^*(t)$  soluzione del sistema (14.45) in corrispondenza a  $u(t) = u^*(t)$  e al valore iniziale  $x(0) = 0.5$ . I risultati rappresentati si riferiscono a  $T = 3$ . Osserviamo che dalla definizione (14.57) per  $T \rightarrow \infty$  si ha  $P(t) \rightarrow 1$ ; in corrispondenza la traiettoria ottimale è soluzione dell'equazione differenziale  $\dot{x}^*(t) = x^*(t) - 2x^*(t) = -x^*(t)$ , da cui  $x^*(t) \rightarrow 0$  e  $u^*(t) \rightarrow 0$  per  $t \rightarrow \infty$ .

Consideriamo ora l'estensione dei risultati precedenti al seguente problema in più dimensioni

$$\left\{ \begin{array}{l} \frac{d\mathbf{x}(t)}{dt} = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}\mathbf{u}(t), \quad \mathbf{x} \in \mathbb{R}^n; \quad \mathbf{u} \in \mathbb{R}^m \\ T: \text{fissato}; \text{ stato finale } \mathbf{x}(T): \text{libero} \\ J(\mathbf{u}) = \frac{1}{2} \int_{t_0}^T [\mathbf{x}^T(s)\mathbf{Q}(s)\mathbf{x}(s) + \mathbf{u}^T(s)\mathbf{R}(s)\mathbf{u}(s)] ds + \frac{1}{2}\mathbf{x}^T(T)\mathbf{A}\mathbf{x}(T) \end{array} \right. \quad (14.59)$$

ove  $\mathbf{A}$ ,  $\mathbf{Q}$ ,  $\mathbf{R}$  sono matrici simmetriche definite positive,  $\mathbf{F}$  è una matrice di ordine  $n$  e  $\mathbf{G}$  di ordine  $m$ ;  $\mathbf{Q}(t)$ ,  $\mathbf{R}(t)$ ,  $\mathbf{F}(t)$ ,  $\mathbf{G}(t)$  sono funzioni derivabili sull'intervallo  $(t_0, T)$ . Le variabili di stato  $\mathbf{x}(t)$  e di controllo  $\mathbf{u}(t)$  sono supposte non vincolate.

L'hamiltoniana del problema è data da

$$\mathcal{H}(\mathbf{x}, \mathbf{u}, V_{\mathbf{x}}^T, t) = \frac{1}{2}\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \frac{1}{2}\mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t) + V_{\mathbf{x}}[\mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{u}(t)] \quad (14.60)$$



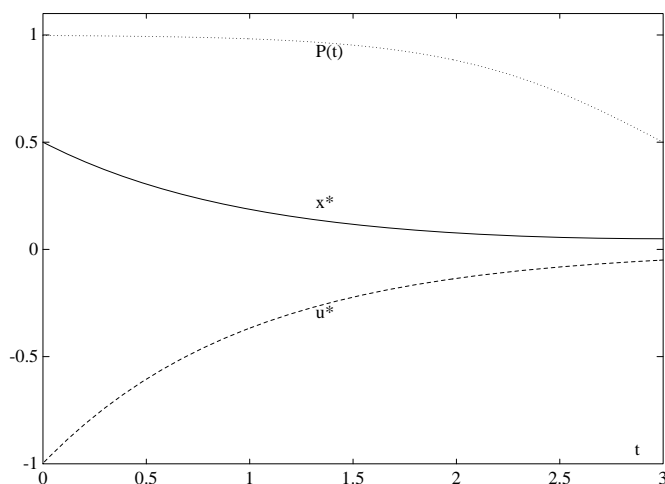


Figura 14.16: Rappresentazione del controllo ottimale  $u^*(t)$ , della traiettoria  $x^*(t)$  e della funzione  $P(t)$  corrispondente al problema di controllo (14.45), (14.46) con  $x(0) = 0.5$  e  $T = 3$ .

Si ha quindi la seguente *condizione necessaria*

$$\frac{\partial \mathcal{H}}{\partial \mathbf{u}} = \mathbf{R}(t)\mathbf{u}(t) + \mathbf{G}^T(t) V_{\mathbf{x}}^T = 0 \quad (14.61)$$

Poiché la matrice

$$\frac{\partial^2 \mathcal{H}}{\partial \mathbf{u}^2} = \mathbf{R}(t)$$

è definita positiva e  $\mathcal{H}$  è una forma quadratica in  $\mathbf{u}$ , il controllo che soddisfa l'equazione (14.61) è un minimo (globale) del funzionale  $\mathcal{H}(\mathbf{u})$ . Risolvendo l'equazione (14.61), si ha

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t) \mathbf{G}^T(t) V_{\mathbf{x}}^T \quad (14.62)$$

che, sostituita nella definizione (14.60), fornisce la seguente uguaglianza

$$\mathcal{H}(\mathbf{x}, \mathbf{u}^*, V_{\mathbf{x}}^T, t) = \frac{1}{2} \mathbf{x}^T(t) \mathbf{Q}(t) \mathbf{x}(t) + V_{\mathbf{x}} \mathbf{F} \mathbf{x} - \frac{1}{2} V_{\mathbf{x}} \mathbf{G} \mathbf{R}^{-1} \mathbf{G}^T V_{\mathbf{x}}^T \quad (14.63)$$

L'equazione di Hamilton-Jacobi-Bellman diventa pertanto

$$V_t + \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + V_{\mathbf{x}} \mathbf{F} \mathbf{x} - \frac{1}{2} V_{\mathbf{x}} \mathbf{G} \mathbf{R}^{-1} \mathbf{G}^T V_{\mathbf{x}}^T = 0 \quad (14.64)$$

con la seguente condizione finale

$$V(\mathbf{x}, T) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (14.65)$$

La condizione (14.65) suggerisce la ricerca della soluzione  $V(\mathbf{x}, t)$  della seguente forma

$$V(\mathbf{x}, t) = \frac{1}{2} \mathbf{x}^T \mathbf{P}(t) \mathbf{x}$$

con  $\mathbf{P}(t)$  matrice simmetrica di ordine  $n$  da determinare. Tenendo conto che

$$V_t = \frac{1}{2} \mathbf{x}^T \frac{d\mathbf{P}}{dt} \mathbf{x}; \quad V_{\mathbf{x}} = \mathbf{x}^T \mathbf{P}(t)$$

per sostituzione in (14.64), si ottiene

$$\frac{1}{2} \mathbf{x}^T \frac{d\mathbf{P}}{dt} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} - \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{G} \mathbf{R}^{-1} \mathbf{G}^T \mathbf{P} \mathbf{x} + \mathbf{x}^T \mathbf{P} \mathbf{F} \mathbf{x} = 0 \quad (14.66)$$

La matrice prodotto  $\mathbf{P}\mathbf{F}$  che appare nell'ultimo termine può essere scritta come somma della parte simmetrica e della parte non simmetrica, ossia nella seguente forma

$$\mathbf{P}\mathbf{F} = \frac{1}{2} [\mathbf{P}\mathbf{F} + (\mathbf{P}\mathbf{F})^T] + \frac{1}{2} [\mathbf{P}\mathbf{F} - (\mathbf{P}\mathbf{F})^T]$$

Si può allora mostrare facilmente che solo la parte simmetrica di  $\mathbf{P}\mathbf{F}$  contribuisce nella somma (14.66), che può essere scritta quindi nella seguente forma

$$\frac{1}{2} \mathbf{x}^T \frac{d\mathbf{P}}{dt} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} - \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{G} \mathbf{R}^{-1} \mathbf{G}^T \mathbf{P} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{F} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{F}^T \mathbf{P} \mathbf{x} = 0 \quad (14.67)$$

L'equazione precedente deve essere verificata per ogni  $\mathbf{x}$ , per cui in definitiva  $V(\mathbf{x}, t)$  è soluzione dell'equazione di ottimalità se la funzione  $\mathbf{P}(t)$  è la soluzione del seguente problema a valori iniziali per un sistema di equazioni differenziali

$$\boxed{\begin{aligned} \frac{d\mathbf{P}}{dt} + \mathbf{F}^T(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}(t) - \mathbf{P}(t)\mathbf{G}(t)\mathbf{R}^{-1}\mathbf{G}^T(t)\mathbf{P}(t) + \mathbf{Q}(t) &= 0 \\ \mathbf{P}(T) &= \mathbf{A} \end{aligned}} \quad (14.68)$$

che rappresenta l'estensione al caso di più dimensioni dell'*equazione di Riccati*. Osserviamo che, dal momento che la matrice  $\mathbf{P}$  è simmetrica, è sufficiente l'integrazione, da  $t = T$  a  $t = t_0$ , di  $n(n+1)/2$  equazioni differenziali. Una volta risolto numericamente il problema a valori iniziali (14.68), la legge di controllo ottimale  $\mathbf{u}^*(\mathbf{x}, t)$  è ottenuta in forma chiusa risolvendo il seguente sistema lineare simmetrico

$$\boxed{\mathbf{R} \mathbf{u}^*(\mathbf{x}, t) = -\mathbf{G}^T \mathbf{P} \mathbf{x}^*(t)} \quad (14.69)$$

ove la variabile di stato  $\mathbf{x}^*(t)$  è ottenuta risolvendo il sistema di stato (14.59) con  $\mathbf{u}(t) = \mathbf{u}^*(t)$ .

Come illustrazione, consideriamo il problema di controllo corrispondente al seguente sistema di stato

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2 = 2x_1(t) - x_2(t) + u(t) \end{cases} \quad (14.70)$$

e al seguente funzionale da minimizzare

$$J(u) = \frac{1}{2} \int_0^T [2x_1^2(t) + x_2^2(t) + \frac{1}{2}u^2(t)] dt \quad (14.71)$$

Per tale problema si ha

$$\mathbf{F} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{R} = \frac{1}{2}$$

e la matrice  $\mathbf{A}$  è la matrice nulla. L'equazione di Riccati (14.68) corrisponde in questo caso al seguente sistema differenziale per le componenti  $p_{ij}$  della matrice  $\mathbf{P}$

$$\begin{cases} \dot{p}_{11}(t) = 2[p_{12}^2(t) - 2p_{12}(t) - 1] \\ \dot{p}_{12}(t) = 2p_{12}(t)p_{22}(t) - p_{11}(t) + p_{12}(t) - 2p_{22}(t) \\ \dot{p}_{22}(t) = 2p_{22}^2(t) - 2p_{12}(t) + 2p_{22}(t) - 1 \end{cases} \quad (14.72)$$

con le condizioni ai limiti  $p_{11}(T) = p_{12}(T) = p_{22}(T) = 0$ . La legge di controllo ottimale è data da

$$u^*(t) = -2[p_{12}(t), p_{22}(t)] \mathbf{x}(t) \quad (14.73)$$

ove  $\mathbf{x} = [x_1, x_2]^T$ . In Figura 14.17 è rappresentata la soluzione dell'equazione di Riccati, il controllo ottimale e le corrispondenti traiettorie per  $\mathbf{x}(0) = [-4, 4]^T$  e  $T = 15$ . Le equazioni differenziali sono risolte mediante un metodo di Runge-Kutta-Fehlberg a passo variabile (cfr. Capitolo 7). ■

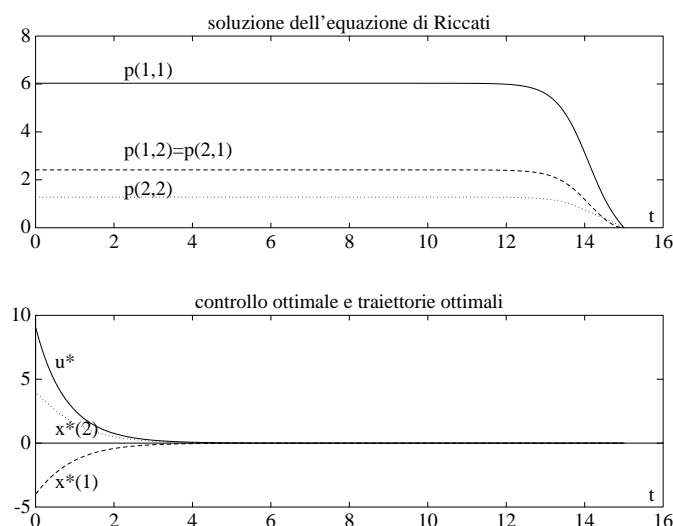


Figura 14.17: Rappresentazione della soluzione del problema di controllo relativo all'equazione di stato (14.70) e al funzionale costo (14.71).

## 14.4 Principio del minimo di Pontryagin

Il principio del minimo di Pontryagin rappresenta l'estensione al problema del minimo (o del massimo) di un funzionale della nota condizione necessaria per l'esistenza di un punto estremo (minimo o massimo) locale di una funzione di variabile reale (cfr. Capitolo 5). Ricordiamo brevemente che una funzione  $f(x)$  ha un *minimo* (rispettivamente un *massimo*) *locale* in  $x^*$ , con  $x^*$  punto interno al dominio di definizione di  $f(x)$ , quando in un intorno  $I_\epsilon \equiv: |x - x^*| \leq \epsilon$ ,  $\epsilon > 0$  opportuno, si ha

$$f(x^*) \leq f(x) \quad (\text{rispettivamente } f(x^*) \geq f(x)) \quad \forall x \in I_\epsilon$$

Se la funzione  $f(x)$  è derivabile, si dimostra che in un punto  $x^*$  di estremo locale deve essere verificata la condizione  $f'(x^*) = 0$ , che rappresenta quindi una *condizione necessaria* per l'esistenza di punti di minimo, o massimo, locali. In effetti, per  $\delta \neq 0$  si considera lo sviluppo in serie di Taylor della funzione  $f(x)$  intorno al punto  $x = x^*$

$$f(x^* + \delta) = f(x^*) + \delta f'(x^*) + O(\delta^2)$$

dal quale si vede che per  $|\delta|$  sufficientemente piccolo l'incremento della funzione  $f(x^* + \delta) - f(x^*)$  ha lo stesso segno di  $\delta f'(x^*)$ . Essendo il segno di  $\delta$  arbitrario, si vede che in un punto di estremo locale questo è possibile solo se  $f'(x^*) = 0$ . Brevemente, si dice anche che in un punto di estremo locale deve essere nulla la variazione infinitesima  $df = f'(x^*) dx$  corrispondente ad una generica variazione infinitesima della variabile indipendente  $x$ .

Le radici  $x^*$  dell'equazione  $f'(x) = 0$  sono dette *punti stazionari*, e non corrispondono necessariamente a punti di estremo locale (in effetti possono essere punti di flesso con tangente orizzontale). La natura esatta di un punto stazionario può essere ricavata dallo studio delle derivate successive; se ad esempio, in particolare,  $f''(x^*) > 0$ , allora  $x^*$  è un punto di minimo.

Con l'obiettivo di introdurre in maniera graduale il principio di Pontryagin, ossia dapprima nel caso discreto e successivamente nel caso continuo, incominceremo a considerare l'estensione dei risultati precedenti al caso di funzioni in più variabili e in presenza di vincoli. Le notazioni utilizzate sono motivate dalle successive applicazioni ai problemi di controllo.

**Problemi senza vincoli** Se la funzione  $L(\mathbf{u})$  è definita per ogni  $\mathbf{u} \in \mathbb{R}^m$ ,  $m \geq 1$ , ed è derivabile su tutto  $\mathbb{R}^m$ , la *condizione necessaria per un minimo* è la seguente

$$\boxed{\frac{\partial L}{\partial \mathbf{u}} = 0} \iff \frac{\partial L}{\partial u_i} = 0, \quad i = 1, 2, \dots, m \quad (14.74)$$

ossia l'annullarsi del gradiente della funzione  $L(\mathbf{u})$ . In analogia al caso unidimensionale, la condizione (14.74) è ottenuta osservando che la variazione infinitesima  $dL = L_{\mathbf{u}} d\mathbf{u}$ , corrispondente ad una variazione arbitraria  $d\mathbf{u}$ , deve essere nulla in un punto di minimo. I punti  $\mathbf{u}^*$  che verificano la condizione (14.74) sono detti *punti stazionari*. Indicata con  $\partial^2 L / \partial \mathbf{u}^2$  la matrice *hessiana*, ossia la matrice di elementi  $\partial^2 L / \partial u_i \partial u_j$ , per  $i, j = 1, 2, \dots, m$ , un punto stazionario  $\mathbf{u}^*$  per il quale la matrice hessiana è *definita positiva* è un punto di minimo; analogamente, se l'hessiana è definita negativa, è un punto di massimo. Quando la matrice hessiana è singolare il punto  $\mathbf{u}^*$  è detto un punto *singolare* e sono necessarie ulteriori informazioni per stabilire se il punto è un minimo.

► **Esempio 14.10** Come illustrazione, consideriamo i seguenti casi di funzioni  $L(u_1, u_2)$ .

(a) *minimo*

$$L(u_1, u_2) := \mathbf{u}^T \mathbf{A} \mathbf{u} = [u_1, u_2] \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Gli autovalori della matrice hessiana, che in questo caso coincide con la matrice  $\mathbf{A}$ , sono dati da  $\lambda_1 \approx 0.697$ ,  $\lambda_2 \approx 4.3$  e quindi la matrice è definita positiva. Il punto  $\mathbf{u}^* = [0, 0]^T$  è quindi un punto di minimo.

(b) *punto sella*

$$L(u_1, u_2) := \mathbf{u}^T \mathbf{A} \mathbf{u} = [u_1, u_2] \begin{bmatrix} -1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Gli autovalori di  $\mathbf{A}$  sono  $\lambda_1 \approx -1.23$  e  $\lambda_2 \approx 3.23$  e il punto  $\mathbf{u}^* = [0, 0]^T$  è un punto sella; la Figura 14.18 fornisce una illustrazione attraverso le curve di livello  $L(\mathbf{u}) = k$ .

(c) *punto singolare*  $L(\mathbf{u}) = (u_1 - u_2^2)(u_1 - 3u_2^2)$ . Gli autovalori sono non negativi, ma uno di essi è zero (cfr. Figura 14.18).

■

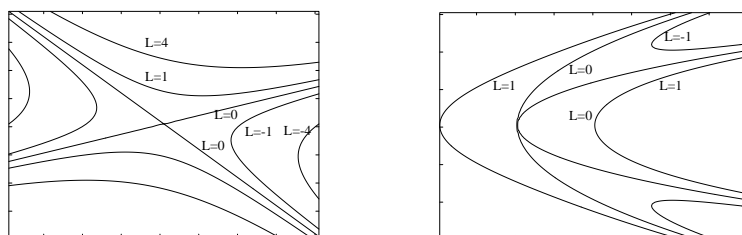


Figura 14.18: Illustrazione mediante le curve di livello  $L(\mathbf{u}) = k$  di un punto sella e rispettivamente di un punto singolare.

**Problemi con vincoli di uguaglianza** Sia  $L(\mathbf{x}, \mathbf{u})$ , con  $\mathbf{x} \in \mathbb{R}^n$  e  $\mathbf{u} \in \mathbb{R}^m$ , una funzione da minimizzare rispetto ai parametri di controllo  $\mathbf{u}$ , mentre i parametri di stato  $\mathbf{x}$  sono determinati a partire dai parametri  $\mathbf{u}$  attraverso i seguenti insiemi di *vincoli*

$$\mathbf{f}(\mathbf{x}; \mathbf{u}) = 0 \iff \begin{cases} f_1(x_1, \dots, x_n; u_1, \dots, u_m) = 0 \\ \vdots \\ f_i(x_1, \dots, x_n; u_1, \dots, u_m) = 0 \\ \vdots \\ f_n(x_1, \dots, x_n; u_1, \dots, u_m) = 0 \end{cases} \quad (14.75)$$

ove  $f_i(\mathbf{x}; \mathbf{u})$  sono funzioni assegnate e definite su  $\mathbb{R}^{n+m}$ . Il problema che si considera è quindi il seguente

$$\min_{\mathbf{u} \in \mathbb{R}^m} L(\mathbf{x}; \mathbf{u}) \quad (14.76)$$

$$\mathbf{f}(\mathbf{x}; \mathbf{u}) = 0 \quad (14.77)$$

Per il seguito supporremo le funzioni  $L(\mathbf{x}; \mathbf{u})$  e  $\mathbf{f}(\mathbf{x}; \mathbf{u})$  derivabili; inoltre la matrice jacobiana  $\mathbf{f}_x$ , ossia la matrice  $n \times n$  di elementi  $\partial f_i / \partial x_j$ ,  $i, j = 1, \dots, n$ , è supposta *non singolare*. Quest'ultima ipotesi assicura che la condizione (14.77) definisce le variabili  $\mathbf{x}$  come funzioni implicite delle variabili  $\mathbf{u}$ .

Un punto *stazionario* per il problema (14.76)-(14.77) è un punto nel quale è nulla la variazione infinitesima  $dL = 0$  corrispondente ad una variazione infinitesima arbitraria  $d\mathbf{u}$ , con la condizione che  $d\mathbf{x}$  sia tale che  $d\mathbf{f} = 0$ . Tenendo conto che

$$dL = L_x d\mathbf{x} + L_u d\mathbf{u} \quad (14.78)$$

$$d\mathbf{f} = \mathbf{f}_x d\mathbf{x} + \mathbf{f}_u d\mathbf{u} \quad (14.79)$$

ricavando  $d\mathbf{x}$  dalla (14.79) e sostituendo nella (14.78), si ottiene

$$dL = (L_u - L_x \mathbf{f}_x^{-1} \mathbf{f}_u) d\mathbf{u} \quad (14.80)$$

e quindi la seguente *condizione necessaria*

$$\boxed{L_u - L_x \mathbf{f}_x^{-1} \mathbf{f}_u = 0} \quad (14.81)$$

Tali  $m$  equazioni, insieme con le  $n$  equazioni (14.77), determinano le  $m$  variabili  $\mathbf{u}$  e le  $n$  variabili  $\mathbf{x}$  nei punti stazionari. Per maggiore chiarezza, osserviamo che la (14.81) rappresenta la derivata parziale di  $L$  rispetto a  $\mathbf{u}$ , *quando si mantiene  $\mathbf{f}$  costante*, mentre  $L_u$  rappresenta la derivata parziale di  $L$  rispetto a  $\mathbf{u}$ , *quando  $\mathbf{x}$  è considerato costante*.

Un altro modo (equivalente al precedente) per ottenere la condizione (14.81) consiste nell'osservare che in un punto stazionario le due equazioni  $dL = 0$  e  $d\mathbf{f} = 0$  in  $d\mathbf{x}$  e  $d\mathbf{u}$ , con  $dL$  e  $d\mathbf{f}$  definite rispettivamente in (14.78) e (14.79), devono essere consistenti. Trattandosi di un sistema omogeneo, la consistenza equivale (cfr. Appendice A) alla dipendenza lineare dei due vettori  $[L_x, L_u]$  e  $[\mathbf{f}_x, \mathbf{f}_u]$ , ossia all'esistenza di un vettore  $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$  tale che

$$L_x + \mathbf{y}^T \mathbf{f}_x = 0 \quad n \text{ equazioni} \quad (14.82)$$

$$L_u + \mathbf{y}^T \mathbf{f}_u = 0 \quad m \text{ equazioni} \quad (14.83)$$

Dall'equazione (14.82) si ottiene  $\mathbf{y}^T = -L_x(\mathbf{f}_x)^{-1}$  e sostituendo nella (14.83) si riottiene la condizione (14.81). Segnaliamo la seguente interessante interpretazione del vettore  $\mathbf{y}$ . Se nelle uguaglianze (14.78) e (14.79) poniamo  $d\mathbf{u} = 0$  e eliminiamo  $d\mathbf{x}$ , si ottiene

$$-\mathbf{y}^T = L_x(\mathbf{f}_x)^{-1} = \frac{\partial L}{\partial \mathbf{f}}$$

Ossia, le componenti del vettore  $\mathbf{y}$  sono le derivate parziali di  $L$  rispetto a  $\mathbf{f}$ , quando  $\mathbf{u}$  è costante; esse rappresentano quindi dei *fattori di sensitività* della funzione obiettivo  $L$  rispetto ai vincoli  $\mathbf{f}$ .

Analizziamo, infine, un terzo modo per ottenere la condizione (14.81), noto come *metodo dei moltiplicatori di Lagrange* e sulla base del quale introdurremo successivamente la condizione di ottimalità per un problema di controllo. Consideriamo la seguente funzione obiettivo, ottenuta “aggiungendo” i vincoli (14.75) alla funzione obiettivo  $L(\mathbf{x}, \mathbf{u})$  mediante un insieme di moltiplicatori indeterminati  $y_1, \dots, y_n$  (detti i moltiplicatori di Lagrange)

$$\mathcal{H}(\mathbf{x}, \mathbf{u}, \mathbf{y}) := L(\mathbf{x}, \mathbf{u}) + \sum_{i=1}^n y_i f_i(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \mathbf{y}^T \mathbf{f}(\mathbf{x}, \mathbf{u}) \quad (14.84)$$

La variazione di  $\mathcal{H}$  rispetto alle variazioni infinitesimali  $d\mathbf{x}$  e  $d\mathbf{u}$  è data da

$$d\mathcal{H} = \frac{\partial \mathcal{H}}{\partial \mathbf{x}} d\mathbf{x} + \frac{\partial \mathcal{H}}{\partial \mathbf{u}} d\mathbf{u} \quad \text{con} \quad \begin{cases} \frac{\partial \mathcal{H}}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{x}} + \mathbf{y}^T \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \\ \frac{\partial \mathcal{H}}{\partial \mathbf{u}} = \frac{\partial L}{\partial \mathbf{u}} + \mathbf{y}^T \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \end{cases} \quad (14.85)$$

Supponiamo ora di avere scelto un vettore di riferimento  $\mathbf{u}$  e di aver determinato il vettore  $\mathbf{x}$  che verifica i vincoli (14.75); in corrispondenza si ha  $\mathcal{H} = L$ . Dal momento che si è interessati ad esaminare le variazioni di  $\mathcal{H}$  (e quindi di  $L$ ) rispetto alle variazioni del vettore  $\mathbf{u}$ , è conveniente *scegliere* il vettore  $\mathbf{y}$ , in maniera che

$$\frac{\partial \mathcal{H}}{\partial \mathbf{x}} = 0 \quad \Rightarrow \quad \mathbf{y}^T = -\frac{\partial L}{\partial \mathbf{x}} \left( \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right)^{-1} \quad (14.86)$$

che coincide con la scelta vista in precedenza. Ne segue il seguente risultato

$$dL \equiv d\mathcal{H} = \frac{\partial \mathcal{H}}{\partial \mathbf{u}} d\mathbf{u} \quad (14.87)$$

dal quale si ha che  $\partial \mathcal{H} / \partial \mathbf{u}$  è il gradiente di  $L$  rispetto a  $\mathbf{u}$ , *quando è verificato il vincolo*  $\mathbf{f}(\mathbf{x}, \mathbf{u}) = 0$ . In un punto stazionario si avrà allora la condizione

$$\frac{\partial \mathcal{H}}{\partial \mathbf{u}} d\mathbf{u} \equiv \frac{\partial L}{\partial \mathbf{u}} + \mathbf{y}^T \frac{\partial \mathbf{f}}{\partial \mathbf{u}} = 0$$

Riassumendo, per il problema di minimo (14.76), (14.77) si hanno le seguenti *condizioni necessarie di stazionarietà*

$$\boxed{\begin{aligned} \frac{\partial \mathcal{H}}{\partial \mathbf{y}} &= \mathbf{f}(\mathbf{x}, \mathbf{u}) = 0 \\ \frac{\partial \mathcal{H}}{\partial \mathbf{x}} &= \frac{\partial L}{\partial \mathbf{x}} + \mathbf{y}^T \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = 0 \\ \frac{\partial \mathcal{H}}{\partial \mathbf{u}} &= \frac{\partial L}{\partial \mathbf{u}} + \mathbf{y}^T \frac{\partial \mathbf{f}}{\partial \mathbf{u}} = 0 \end{aligned}} \quad (14.88)$$

ove  $\mathcal{H} = L(\mathbf{x}, \mathbf{u}) + \mathbf{y}^T \mathbf{f}(\mathbf{x}, \mathbf{u})$ . Le (14.88) sono  $2n + m$  equazioni nelle  $2n + m$  incognite  $\mathbf{x}$ ,  $\mathbf{y}$  e  $\mathbf{u}$ . Condizioni sufficienti per l'esistenza di un minimo possono essere ottenute analizzando gli autovalori della matrice hessiana  $\partial^2 L / \partial \mathbf{u}^2$ .

► **Esempio 14.11** Come illustrazione, consideriamo il problema del calcolo del minimo della seguente funzione nella variabile  $u \in \mathbb{R}$

$$L(x, u) = \frac{1}{2} \left( \frac{x^2}{a^2} + \frac{u^2}{b^2} \right)$$

ove  $x \in \mathbb{R}$  verifica il seguente vincolo

$$f(x, u) = c - xu = 0$$

e  $a, b$  e  $c$  sono costanti positive assegnate. Le curve di livello  $L(x, u) = k$  sono delle ellissi, mentre  $c - xu = 0$  è una iperbole equilatera. Il valore minimo di  $L$  si ottiene quindi quando l'ellisse è tangente alla iperbole. Utilizzando il metodo dei moltiplicatori di Lagrange, si ha

$$\mathcal{H}(x, u, y) := \frac{1}{2} \left( \frac{x^2}{a^2} + \frac{u^2}{b^2} \right) + y(c - xu)$$

da cui le seguenti condizioni necessarie

$$\frac{\partial \mathcal{H}}{\partial y} = c - xu = 0, \quad \frac{\partial \mathcal{H}}{\partial x} = \frac{x}{a^2} - yu = 0, \quad \frac{\partial \mathcal{H}}{\partial u} = \frac{u}{b^2} - yx = 0$$

Si trova facilmente

$$x^* = \pm \sqrt{\frac{ac}{b}}, \quad u^* = \pm \sqrt{\frac{bc}{a}}, \quad y^* = \frac{1}{ab}, \quad L^* = \frac{c}{ab}$$

Dalla matrice hessiana  $\partial^2 L / \partial u^2$  si può verificare che i valori  $u^*$  corrispondono a due punti di minimo nei quali la funzione obiettivo assume lo stesso valore. Osserviamo anche che

$$y^* = \frac{\partial L^*}{\partial c}$$

e quindi  $y^*$  è il coefficiente di sensitività del valore ottimale  $L^*$  rispetto alla variazione del vincolo.

Come ulteriore esempio si consideri il caso generale di una *funzione obiettivo quadratica* e *vincoli lineari*, ossia il seguente problema (cfr. per analogia l'Esempio 14.9)

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^m} L(\mathbf{x}, \mathbf{u}) &= \frac{1}{2} (\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}) \\ \mathbf{f}(\mathbf{x}, \mathbf{u}) &= \mathbf{x} + \mathbf{G} \mathbf{u} + \mathbf{c} = 0, \quad \mathbf{x} \in \mathbb{R}^n \end{aligned} \tag{14.89}$$

ove  $\mathbf{Q}$ ,  $\mathbf{R}$  sono matrici simmetriche definite positive e  $\mathbf{G}$ ,  $\mathbf{c}$  sono rispettivamente una matrice di ordine  $m$  e un vettore di ordine  $n$  assegnati. Lasciamo come esercizio la verifica dei seguenti risultati

$$\mathbf{u}^* = -(\mathbf{R} + \mathbf{G}^T \mathbf{Q} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{Q} \mathbf{c}; \quad \mathbf{x}^* = -(\mathbf{I} - \mathbf{G}(\mathbf{R} + \mathbf{G}^T \mathbf{Q} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{Q}) \mathbf{c}$$

In questo caso quindi il calcolo dei vettori  $\mathbf{u}^*$ ,  $\mathbf{x}^*$  è ricondotto alla risoluzione di due sistemi lineari. Si verifichi inoltre che  $\mathbf{y}^T = \partial L(\mathbf{x}^*, \mathbf{u}^*) / \partial \mathbf{c}$ . ■



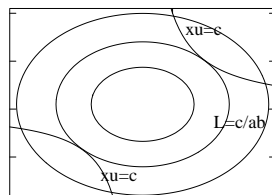


Figura 14.19: Esempio di minimizzazione con vincoli di uguaglianza.

**Metodi numerici** In generale, le equazioni (14.88), corrispondenti alle condizioni necessarie di ottimalità, non possono essere risolte in maniera analitica, ma richiedono l'utilizzo di opportuni metodi numerici. Tali metodi sono stati analizzati nel Capitolo 5 nel caso di problemi di ottimalità senza vincoli. Discuteremo ora brevemente la loro estensione al caso di problemi di ottimalità con vincoli di uguaglianza, ossia della forma (14.76), (14.77). Per illustrare l'idea, consideriamo in particolare il *metodo del gradiente*; allo stesso modo si estendono i metodi basati sulle direzioni coniugate e i metodi Quasi-Newton.

Ricordiamo che il metodo del gradiente è un metodo iterativo che a partire da un vettore di tentativo  $\mathbf{u}^0$  genera una successione di vettori  $\{\mathbf{u}^k\}$ , ciascuno dei quali è ottenuto mediante una minimizzazione unidimensionale lungo la direzione del gradiente. Quando il metodo converge, il limite della successione è un punto di stazionarietà, ossia uno zero dell'equazione  $\partial\mathcal{H}/\partial\mathbf{u} = 0$ . I passi necessari per calcolare il vettore  $\mathbf{u}^{k+1}$  a partire dal vettore  $\mathbf{u}^k$  possono essere riassunti nel seguente modo, che esponiamo in una forma adatta alla generalizzazione ai problemi di controllo.

- (a) Si calcola il vettore  $\mathbf{x}^k$  soluzione del sistema di equazioni

$$\mathbf{f}(\mathbf{x}^k, \mathbf{u}^k) = 0 \quad (14.90)$$

- (b) Si calcola il vettore  $\mathbf{y}^k$  risolvendo il seguente sistema lineare (cfr. (14.86))

$$\left(\frac{\partial\mathbf{f}}{\partial\mathbf{x}}\right)^T \mathbf{y}^k = \left(-\frac{\partial L}{\partial\mathbf{x}}\right)^T \quad (14.91)$$

- (c) Si calcola, per  $\mathbf{u} = \mathbf{u}^k$ ,  $\mathbf{x} = \mathbf{x}^k$  e  $\mathbf{y} = \mathbf{y}^k$ , il vettore gradiente

$$\left(\frac{\partial\mathcal{H}}{\partial\mathbf{u}}\right)_k = \left(\frac{\partial L}{\partial\mathbf{u}}\right)_k + (\mathbf{y}^k)^T \left(\frac{\partial\mathbf{f}}{\partial\mathbf{u}}\right)_k \quad (14.92)$$

(d) Si calcola il nuovo vettore  $\mathbf{u}^{k+1}$  mediante la seguente formula

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \lambda_k \left( \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \right)_k \quad (14.93)$$

ove  $\lambda_k$  è un parametro da scegliere convenientemente. In particolare, quando esso è ottenuto in maniera che il punto  $\mathbf{u}^{k+1}$  sia il minimo di  $L$  lungo la direzione del gradiente, si ha il metodo della massima discesa (*steepest descent*).

La procedura precedente prosegue fino a che la norma del gradiente è inferiore ad una precisione  $\epsilon$  prefissata. La risoluzione del sistema (14.90) (in generale non lineare) e del sistema lineare (14.91) rappresentano la parte computazionalmente più onerosa dell'algoritmo. Dal punto di vista pratico è quindi importante ridurre al minimo il numero delle iterazioni. Tale risultato può essere ottenuto, come abbiamo visto nel Capitolo 5, mediante la costruzione di direzioni di discesa tra loro coniugate o l'utilizzo delle direzioni fornite dal metodo di Newton.

I risultati che abbiamo analizzato in precedenza in relazione ai problemi di ottimizzazione con vincoli di uguaglianza possono essere estesi in maniera conveniente ai problemi con vincoli di disuguaglianza. Per una esemplificazione, rinviamo al modello della programmazione lineare trattato nel Capitolo 5. Ulteriori esempi saranno considerati successivamente nell'ambito dei problemi di controllo. Per una trattazione generale, che richiede una introduzione adeguata alla programmazione convessa, si veda ad esempio [31] e la relativa bibliografia.

#### 14.4.1 Problemi di controllo discreti

**Sistemi ad un solo stato** Come introduzione al caso generale, consideriamo il caso semplice di un sistema ad un solo stato. Un sistema, descritto dalla variabile di stato  $\mathbf{x}(t) \in \mathbb{R}^n$ , si trova inizialmente nello stato  $\mathbf{x}(0)$  e mediante la decisione descritta dal vettore di controllo  $\mathbf{u}(t) \in \mathbb{R}^m$  viene portato allo stato  $\mathbf{x}(1)$  attraverso la seguente relazione

$$\mathbf{x}(1) = \mathbf{f}(0, \mathbf{x}(0), \mathbf{u}(0)) \quad (14.94)$$

ove  $\mathbf{f}$  è una funzione assegnata. Il problema di controllo consiste nello scegliere  $\mathbf{u}(0)$  in modo da minimizzare la seguente funzione obiettivo

$$J(\mathbf{x}, \mathbf{u}) := L(0, \mathbf{x}(0), \mathbf{u}(0)) + \lambda(\mathbf{x}(1)) \quad (14.95)$$

ove  $L$  e  $\lambda$  sono funzioni assegnate, con il vincolo (14.94). Il problema da risolvere è allora del tutto analogo a quelli di ottimizzazione con vincoli di uguaglianza che abbiamo considerato nel paragrafo precedente. Possiamo pertanto ottenere le condizioni necessarie di ottimalità utilizzando la tecnica dei moltiplicatori di Lagrange. Per convenienza, considereremo la seguente funzione "aumentata"

$$\bar{J} := L(0, \mathbf{x}(0), \mathbf{u}(0)) + \mathbf{y}^T(1) [\mathbf{f}(0, \mathbf{x}(0), \mathbf{u}(0)) - \mathbf{x}(1)] + \lambda(\mathbf{x}(1)) \quad (14.96)$$

che, posto

$$\mathcal{H}(0, \mathbf{x}(0), \mathbf{u}(0), \mathbf{y}(1)) = L(0, \mathbf{x}(0), \mathbf{u}(0)) + \mathbf{y}^T(1) \mathbf{f}(0, \mathbf{x}(0), \mathbf{u}(0)) \quad (14.97)$$

può essere riscritta nel seguente modo

$$\bar{J} = \mathcal{H}(0, \mathbf{x}(0), \mathbf{u}(0), \mathbf{y}(1)) + \lambda(\mathbf{x}(1)) - \mathbf{y}^T(1) \mathbf{x}(1) \quad (14.98)$$

Consideriamo ora una variazione infinitesima in  $\bar{J}$  dovuta a variazioni infinitesime in  $\mathbf{u}(0)$ ,  $\mathbf{x}(0)$  e  $\mathbf{x}(1)$

$$d\bar{J} = \frac{\mathcal{H}(0)}{\partial \mathbf{u}(0)} d\mathbf{u}(0) + \frac{\mathcal{H}(0)}{\partial \mathbf{x}(0)} d\mathbf{x}(0) + \left[ \frac{\partial \lambda}{\partial \mathbf{x}(1)} - \mathbf{y}^T(1) \right] d\mathbf{x}(1) \quad (14.99)$$

Per evitare di calcolare, dall'equazione (14.94),  $d\mathbf{x}(1)$  in termini di  $\mathbf{u}(0)$ , si *sceglie*

$$\mathbf{y}^T(1) = \frac{\partial \lambda}{\partial \mathbf{x}(1)} \quad (14.100)$$

Con tale scelta, si ha la relazione

$$d\bar{J} = \frac{\mathcal{H}(0)}{\partial \mathbf{u}(0)} d\mathbf{u}(0) + \frac{\mathcal{H}(0)}{\partial \mathbf{x}(0)} d\mathbf{x}(0) \quad (14.101)$$

dalla quale si vede che  $\partial \mathcal{H}(0)/\partial \mathbf{u}(0)$  è il gradiente di  $J$  rispetto a  $\mathbf{u}(0)$ , quando  $\mathbf{x}(0)$  è mantenuto costante e soddisfacente (14.94), e  $\partial \mathcal{H}(0)/\partial \mathbf{x}(0)$  è il gradiente di  $J$  rispetto a  $\mathbf{x}(0)$ , quando  $\mathbf{u}(0)$  è mantenuto costante e soddisfacente (14.94). In particolare, quando il vettore iniziale  $\mathbf{x}(0)$  è assegnato, si ha  $d\mathbf{x}(0) = 0$ .

In definitiva, se  $\mathbf{x}(0)$  è assegnato, le condizioni di ottimalità sono date dall'equazione di stato (14.94), dall'equazione (14.100) e dall'annullarsi del gradiente

$$\frac{\partial \mathcal{H}(0)}{\partial \mathbf{u}(0)} = 0 \quad (14.102)$$

Osserviamo che tali condizioni corrispondono a  $n + n + m$  equazioni per le  $n + n + m$  incognite  $\mathbf{x}(1)$ ,  $\mathbf{y}(1)$  e  $\mathbf{u}(0)$ .

**Sistemi a più stati** Le considerazioni sviluppate nel paragrafo precedente possono essere facilmente generalizzate al caso di un problema di controllo discreto a più stati, definito dalle seguenti equazioni alle differenze

$$\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(0) \text{ assegnato}, \quad t = 0, \dots, T-1 \quad (14.103)$$

con  $\mathbf{x} \in \mathbb{R}^n$  e  $\mathbf{u} \in \mathbb{R}^m$ , e dalla seguente funzione obiettivo

$$J(\mathbf{x}, \mathbf{u}) = \sum_{t=0}^{T-1} L(t, \mathbf{x}(t), \mathbf{u}(t)) + \lambda(\mathbf{x}(T)) \quad (14.104)$$

Procedendo in maniera del tutto analoga, posto

$$\mathcal{H}(t, \mathbf{x}(t), \mathbf{u}(t)) := L(t, \mathbf{x}(t), \mathbf{u}(t)) + \mathbf{y}^T(t+1)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad t = 0, \dots, T-1 \quad (14.105)$$

si trovano le seguenti condizioni necessarie di ottimalità

$$\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \quad t = 0, \dots, T-1 \quad (14.106)$$

$$\mathbf{y}(t) = \left[ \frac{\partial \mathbf{f}(t)}{\partial \mathbf{x}(t)} \right]^T \mathbf{y}(t+1) + \left[ \frac{\partial L(t)}{\partial \mathbf{x}(t)} \right]^T, \quad t = 1, \dots, T-1 \quad (14.107)$$

$$\frac{\partial \mathcal{H}(t)}{\partial \mathbf{u}(t)} = \frac{\partial L(t)}{\partial \mathbf{u}(t)} + \mathbf{y}^T(t+1) \frac{\partial \mathbf{f}(t)}{\partial \mathbf{u}(t)} = 0, \quad t = 0, \dots, T-1 \quad (14.108)$$

$$\mathbf{x}(0) \text{ assegnato}, \quad \mathbf{y}(T) = \left[ \frac{\partial \lambda}{\partial \mathbf{x}(T)} \right]^T \quad (14.109)$$

Il vettore  $\mathbf{y}$  è chiamato il *vettore di stato aggiunto* e le equazioni alle differenze (14.107) costituiscono il *sistema di stato aggiunto*. La funzione  $\mathcal{H}$  è chiamata la *hamiltoniana* del problema. Osserviamo che le equazioni (14.106) e (14.107) sono *accoppiate*, dal momento che  $\mathbf{u}(t)$  dipende da  $\mathbf{y}(t)$  attraverso (14.108), e i coefficienti di (14.107) dipendono, in generale, da  $\mathbf{x}(t)$  e  $\mathbf{u}(t)$ . Tenendo presenti le condizioni (14.109), si ha che le equazioni precedenti costituiscono un *problema ai limiti (two-point boundary-value problem)*, la cui soluzione può essere ottenuta, in generale, mediante metodi numerici di tipo shooting; in maniera schematica (cfr. per maggiori dettagli il Capitolo 7), si considera la famiglia di problemi a valori iniziali ottenuta ponendo  $\mathbf{y}(0) = \mathbf{s}$ , con  $\mathbf{s}$  vettore da determinare in maniera che la soluzione  $\mathbf{y}(t, \mathbf{s})$  delle equazioni (14.106), (14.107) e (14.108) verifichi per  $t = T$  la condizione (14.109).

► **Esempio 14.12** Come illustrazione, riprendiamo l'Esempio 14.8 relativo ad una terapia ottimale di un tumore mediante radiazioni.  $N$  tumori situati in differenti posizioni vengono irradiati mediante una quantità fissata  $A$  di radiazioni. Si tratta di trovare la quantità  $u(k)$  con cui irradiare la locazione  $k$  in modo da distruggere il massimo numero di cellule cancerose. Il problema può essere formulato come problema di minimo della funzione

$$J = \sum_{k=0}^{N-1} u(k)^{1/2} \quad (14.110)$$

con il vincolo

$$\sum_{k=0}^{N-1} u(k) = A \quad (14.111)$$

Tale problema è equivalente al problema di controllo corrispondente alla seguente equazione di stato

$$x(k+1) = x(k) - u(k)$$

con le seguenti condizioni, rispettivamente iniziale e finale

$$x(0) = A, \quad x(N) = 0 \quad (14.112)$$

e la funzione obiettivo  $J$  definita in (14.110).

L'hamiltoniana del problema è definita come segue

$$\mathcal{H} = u(k)^{1/2} + y(k+1)[x(k) - u(k)] \quad (14.113)$$

e l'equazione dello stato aggiunto è

$$y(k) = y(k+1), \quad k = 0, 1, \dots, N-1 \quad (14.114)$$

Si può vedere facilmente che, dal momento che lo stato  $x$  è assegnato nel punto finale  $N$  (cfr. (14.112)), non si hanno condizioni su  $y(N)$ . Dall'equazione (14.114) si ha quindi che  $y(k) = \bar{y}$ , con  $\bar{y}$  costante. La condizione di ottimalità fornisce la seguente equazione

$$\mathcal{H}_u = \frac{1}{2} u(k)^{-1/2} - y = 0 \quad (14.115)$$

Pertanto  $u(k) = 1/(4\bar{y}^2)$  e quindi anche  $u(k)$  è costante. La costante è determinata dalla condizione  $\sum_{k=0}^{N-1} u(k) = A$ , da cui  $u^*(k) = A/N$ . ■

#### 14.4.2 Problemi di controllo continui

In questo paragrafo esamineremo in maniera intuitiva l'estensione delle considerazioni sviluppate nei paragrafi precedenti al caso di un problema di controllo continuo nella seguente forma generale (cfr. per le notazioni Sezione 14.2)

$$\left\{ \begin{array}{l} \frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \quad \mathbf{x}(t) \in \mathbb{R}^n, \quad \mathbf{u}(t) \in \mathcal{U}(t) \subset \mathbb{R}^m, \quad t \in (t_0, T) \\ \mathbf{x}(t_0) = \boldsymbol{\xi}, \quad t_0, \boldsymbol{\xi} \text{ assegnati} \\ (\mathbf{x}(T), T) \in \mathcal{B} \\ J(\mathbf{u}) = \int_{t_0}^T L(\mathbf{x}(s), \mathbf{u}(s), s) ds + \lambda(\mathbf{x}(T), T) \end{array} \right. \quad (14.116)$$

Per stabilire le condizioni di ottimalità, studiamo la *variazione*  $\delta J$  del criterio  $J(\mathbf{u})$  corrispondente ad una *variazione ammissibile e infinitesimale*  $\delta\mathbf{u}(t)$  del comando  $\mathbf{u}(t)$ . Per *variazione ammissibile*  $\delta\mathbf{u}(t)$  si intende una variazione tale che il comando  $\mathbf{u}(t) + \delta\mathbf{u}(t)$  appartiene a  $\mathcal{U}(t)$  e trasferisce il sistema dallo stato iniziale  $(\boldsymbol{\xi}, t_0)$ , lungo una traiettoria ammissibile  $\mathbf{x}(t) + \delta\mathbf{x}(t)$ , in uno stato e istante  $(\mathbf{x}(T) + d\mathbf{x}_T, T + dT) \in \mathcal{B}$ . Per variazione infinitesima intendiamo una variazione  $\delta\mathbf{u}(t)$  dipendente, ad esempio, da un parametro  $\epsilon$  tale che  $\|\delta\mathbf{u}(t)\| \rightarrow 0$  quando  $\epsilon \rightarrow 0$ .

Si definisce *hamiltoniana* del problema (14.116) la seguente funzione

$$\mathcal{H}(\mathbf{x}, \mathbf{u}, \mathbf{y}, t) = L(\mathbf{x}, \mathbf{u}, t) + \mathbf{y}^T(t) \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \quad (14.117)$$

definita per  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  e  $\mathbf{u} \in \mathcal{U}$ . Tale funzione “aggiunge”, mediante la funzione vettoriale  $\mathbf{y}$ , alla funzione costo il vincolo costituito dall'equazione di stato. La funzione  $\mathbf{y}(t)$  svolge quindi il ruolo di “moltiplicatore di Lagrange”. Essa permette, in

sostanza, di calcolare la variazione  $\delta \mathbf{x}$  che corrisponde alla variazione  $\delta \mathbf{u}$ ; sottolineiamo che  $\mathbf{x}$  dipende da  $\mathbf{u}$  attraverso l'equazione di stato. Si ha allora il seguente risultato.

**Lemma 14.1** *Sia  $\mathbf{u}(t)$  un controllo ammissibile di riferimento (nominale) e  $\mathbf{x}(t)$  la corrispondente traiettoria che trasferisce il sistema dal punto iniziale  $(\boldsymbol{\xi}, t_0)$  al punto finale  $(\mathbf{x}(T), T) \in \mathcal{B}$ . La variazione  $\delta J$  del criterio, dovuta a una variazione ammissibile  $\delta \mathbf{u}(t)$  del comando, può essere espressa nel seguente modo*

$$\delta J = \int_{t_0}^T \delta \mathcal{H}(t) dt \quad (14.118)$$

ove  $\delta \mathcal{H}$  indica la variazione dell'hamiltoniana  $\mathcal{H}$  dovuta a  $\delta \mathbf{u}$ , ossia

$$\delta \mathcal{H} = (L_{\mathbf{u}} + \mathbf{y}^T \mathbf{f}_{\mathbf{u}}) \delta \mathbf{u} \quad (14.119)$$

con  $\mathbf{f}_{\mathbf{u}} = [\partial \mathbf{f} / \partial u_i, i = 1, \dots, m]$  (matrice  $n \times m$ ) e  $L_{\mathbf{u}} = [\partial L / \partial u_i, i = 1, \dots, m]$ . La funzione vettoriale  $\mathbf{y}(t)$  (vettore di stato aggiunto) è definita come soluzione della seguente equazione differenziale lineare (equazione di stato aggiunto)

$$\frac{d\mathbf{y}(t)}{dt} = -\mathbf{f}_{\mathbf{x}}^T \mathbf{y} - L_{\mathbf{x}}^T \quad (14.120)$$

con  $\mathbf{f}_{\mathbf{x}} = [\partial \mathbf{f} / \partial x_i, i = 1, \dots, n]$  (matrice  $n \times n$ ) e  $L_{\mathbf{x}} = [\partial L / \partial x_i, i = 1, \dots, n]$ , e dalle condizioni finali, dette condizioni di trasversalità

$$\mathbf{y}^T(T) d\mathbf{x}_T - \mathcal{H}(T) dT = \lambda_{\mathbf{x}} d\mathbf{x}_T + \lambda_t dT \quad (14.121)$$

per ogni vettore  $(d\mathbf{x}_T, dT)$  tangente alla varietà  $\mathcal{B}$  nel punto finale.

**DIMOSTRAZIONE.** Procedendo in maniera formale,  $\delta \mathbf{x}(t)$  è soluzione (al primo ordine) del seguente problema a valori iniziali

$$\begin{cases} \frac{d}{dt}(\delta \mathbf{x}(t)) = \mathbf{f}_{\mathbf{x}} \delta \mathbf{x} + \mathbf{f}_{\mathbf{u}} \delta \mathbf{u}(t) \\ \delta \mathbf{x}(t_0) = 0 \end{cases}$$

Dalla regola di derivazione di un prodotto si ha

$$\frac{d}{dt}(\mathbf{y}^T \delta \mathbf{x}) = \frac{d}{dt}(\mathbf{y}^T) \delta \mathbf{x} + \mathbf{y}^T \frac{d}{dt} \delta \mathbf{x}$$

e quindi dal sistema precedente e da (14.120) si ottiene

$$\frac{d}{dt}(\mathbf{y}^T \delta \mathbf{x}) = -L_{\mathbf{x}} \delta \mathbf{x} + \mathbf{y}^T \mathbf{f}_{\mathbf{u}} \delta \mathbf{u}$$

da cui

$$\frac{d}{dt}(\mathbf{y}^T \delta \mathbf{x}) + L_{\mathbf{x}} \delta \mathbf{x} + L_{\mathbf{u}} \delta \mathbf{u} = \mathbf{y}^T \mathbf{f}_{\mathbf{u}} \delta \mathbf{u} + L_{\mathbf{u}} \delta \mathbf{u} = \delta \mathcal{H}$$

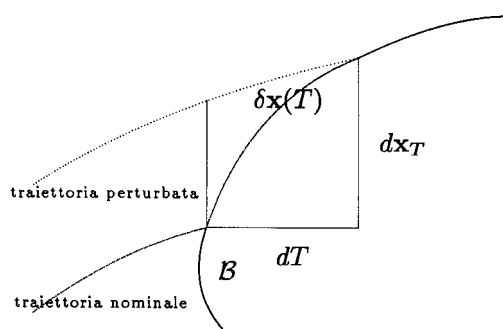


Figura 14.20: Illustrazione schematica delle condizioni di trasversalità.

Integrando tale uguaglianza tra  $t_0$  e  $T$  e tenendo conto della condizione iniziale  $\delta \mathbf{x}(t_0) = 0$ , si ottiene

$$\mathbf{y}^T(T)\delta \mathbf{x}(T) + \int_{t_0}^T (L_{\mathbf{x}}\delta \mathbf{x} + L_{\mathbf{u}}\delta \mathbf{u}) dt = \int_{t_0}^T \delta \mathcal{H}(t) dt \quad (14.122)$$

Consideriamo, ora, lo spostamento  $(d\mathbf{x}_T, dT)$  del punto finale su  $\mathcal{B}$ ; si ha (al primo ordine)

$$\delta \mathbf{x}(T) = d\mathbf{x}_T - \left( \frac{d\mathbf{x}}{dt} \right)_T dT = d\mathbf{x}_T - \mathbf{f} dT$$

con  $(d\mathbf{x}_T, dT)$  tangente a  $\mathcal{B}$ . Tenendo conto delle condizioni finali (14.121) si ha

$$\mathbf{y}^T \delta \mathbf{x}(T) = \mathbf{y}^T(T)(d\mathbf{x}_T - \mathbf{f} dT) = \lambda_{\mathbf{x}} d\mathbf{x}_T + \lambda_t dT + L dT \quad (14.123)$$

dal momento che  $\mathcal{H}(T) = \mathbf{y}^T(T)\mathbf{f} + L$ . Il risultato richiesto si ottiene allora da (14.122) e (14.123), osservando che dalle usuali regole di derivazione si ha

$$\delta J = \int_{t_0}^T (L_{\mathbf{x}}\delta \mathbf{x} + L_{\mathbf{u}}\delta \mathbf{u}) dt + L dT + \lambda_{\mathbf{x}} d\mathbf{x}_T + \lambda_t dT$$

■

Le condizioni di trasversalità (14.121) esprimono il fatto che il vettore  $(\mathbf{y}^T - \lambda_{\mathbf{x}}, \mathcal{H} + \lambda_t)_T$  è ortogonale alla varietà  $\mathcal{B}$ . Per approfondirne il significato, consideriamo alcuni casi particolari.

**Problemi con tempo finale  $T$  fissato** Se il tempo finale  $T$  è specificato,  $\mathbf{x}(T)$  può essere specificato, libero, o vincolato a giacere su una ipersuperficie assegnata nello spazio degli stati.

- (i) *Stato finale specificato.* Poiché  $d\mathbf{x}_T = 0$  e  $dT = 0$ , la condizione (14.121) è automaticamente verificata. Teniamo, comunque, presente, che in questo caso si hanno le  $n$  condizioni ai limiti

$$\mathbf{x}(T) = \mathbf{x}_T \quad (14.124)$$

con  $\mathbf{x}_T$  vettore assegnato.

- (ii) *Stato finale libero.* Poiché  $dT = 0$ , mentre  $d\mathbf{x}_T$  è arbitrario, si hanno le seguenti  $n$  equazioni

$$\mathbf{y}(T) = \lambda_{\mathbf{x}}^T(\mathbf{x}(T)) \quad (14.125)$$

- (iii) *Stato finale appartenente ad una ipersuperficie definita da  $\mathbf{m}(\mathbf{x}(t)) = 0$ .* Come illustrazione, consideriamo il caso particolare di un sistema di stato del secondo ordine, con il vincolo che lo stato finale  $\mathbf{x}(T) = [x_1(T), x_2(T)]^T$  appartenga alla seguente circonferenza

$$m(\mathbf{x}(t)) = (x_1(t) - 3)^2 + (x_2(t) - 4)^2 - 4 = 0$$

Le variazioni ammissibili in  $\mathbf{x}(T)$  sono (al primo ordine) tangenti alla circonferenza nel punto  $(\mathbf{x}(T), T)$ . Tenendo presente che la tangente è ortogonale al gradiente, che è fornito dal seguente vettore

$$\frac{\partial m}{\partial \mathbf{x}} = \begin{bmatrix} 2(x_1(T) - 3) \\ 2(x_2(T) - 4) \end{bmatrix}$$

si ha

$$\left[ \frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}(T)) \right]^T d\mathbf{x}_T = 2(x_1(T) - 3) dx_1(T) + 2(x_2(T) - 4) dx_2(T) = 0$$

da cui

$$dx_2(T) = -\frac{x_1(T) - 3}{x_2(T) - 4} dx_1(T)$$

Sostituendo nell'equazione (14.121), e tenendo conto che  $dx_1(T)$  è arbitrario, si ha la seguente condizione

$$\left[ \lambda_{\mathbf{x}(T)}^T - \mathbf{y}(T) \right]^T \begin{bmatrix} 1 \\ -\frac{x_1(T) - 3}{x_2(T) - 4} \end{bmatrix} = 0 \quad (14.126)$$

La seconda condizione è data dall'appartenza del punto finale all'insieme bersaglio, cioè la condizione  $m(\mathbf{x}(T)) = 0$ .

**Problemi con tempo finale  $T$  libero** Per brevità, ci limiteremo a considerare alcune situazioni; nel seguito esse saranno illustrate su opportuni modelli.

- (i) *Stato finale fissato.* Poiché  $d\mathbf{x}_T = 0$ , mentre  $dT$  è arbitrario, la condizione (14.121) fornisce la seguente equazione

$$\mathcal{H}(\mathbf{x}(T), \mathbf{u}(T), \mathbf{y}, T) + \frac{\partial \lambda}{\partial t}(\mathbf{x}(T), T) = 0 \quad (14.127)$$



(ii) *Stato finale libero*. Poiché  $d\mathbf{x}_T$ , e  $dT$  sono arbitrari e indipendenti, si hanno le seguenti condizioni

$$\mathbf{y}(T) = \left[ \frac{\lambda}{\mathbf{x}}(\mathbf{x}(T), T) \right]^T \quad (n \text{ equazioni}) \quad (14.128)$$

$$\mathcal{H}(\mathbf{x}(T), \mathbf{u}(T), \mathbf{y}, T) + \frac{\partial \lambda}{\partial t}(\mathbf{x}(T), T) = 0 \quad (1 \text{ equazione}) \quad (14.129)$$

In particolare, quando  $\lambda \equiv 0$ , si ha

$$\mathbf{y}(T) = 0 \quad (14.130)$$

$$\mathcal{H}(\mathbf{x}(T), \mathbf{u}(T), \mathbf{y}, T) = 0 \quad (14.131)$$

L'interesse della formula (14.119) sta nel fatto che essa permette di valutare l'effetto su  $J$  della variazione locale di  $\mathbf{u}$ . In senso formale, si può interpretare  $\mathcal{H}_{\mathbf{u}} = L_{\mathbf{u}} + \mathbf{y}^T \mathbf{f}_{\mathbf{u}}$  come la derivata di  $J$  rispetto a  $\mathbf{u}$ . La giustificazione di tale affermazione richiede naturalmente una opportuna definizione di derivata negli spazi funzionali. La funzione  $\mathcal{H}_{\mathbf{u}}$  è anche chiamata funzione di risposta impulsiva (*impulse response function*), in quanto ogni componente di  $\mathcal{H}_{\mathbf{u}}$  rappresenta la variazione in  $J$  dovuta all'impulso unitario (funzione di Dirac) nella corrispondente componente di  $\delta\mathbf{u}$  al tempo  $t$ . Le funzioni  $\mathbf{y}(t)$  sono anche chiamate funzioni di influenza (*influence function*) (*marginal value*, o anche *shadow price* nelle applicazioni economiche), in quanto, come vedremo nel seguito, le componenti del vettore  $\mathbf{y}(t)$  forniscono le variazioni di  $J$ , al tempo  $t$ , rispetto alle componenti della funzione di stato  $\mathbf{x}(t)$ .

La precedente interpretazione della funzione  $\mathcal{H}_{\mathbf{u}}$  è alla base del principio del minimo di Pontryagin che ora richiameremo nella sostanza, rimandando per una sua più corretta interpretazione e per una dimostrazione in un ambiente funzionale opportuno, alla bibliografia.

Dalla definizione di minimo locale si ha che  $\mathbf{u}^*(t)$  è un *controllo ottimale* quando

$$J(\mathbf{u}^*(t) + \delta\mathbf{u}(t)) - J(\mathbf{u}^*(t)) \geq 0 \quad \forall \delta\mathbf{u}(t) \text{ infinitesima ammissibile} \quad (14.132)$$

Dal risultato (14.118) si ha pertanto la seguente *condizione necessaria* di minimo

$$\int_{t_0}^T \delta\mathcal{H}(t) dt \geq 0 \quad \forall \delta\mathbf{u}(t) \text{ infinitesima ammissibile} \quad (14.133)$$

Ricordiamo che la positività di un integrale *non* comporta, in generale, la positività della funzione integranda. In questo caso, tuttavia, le variazioni  $\delta\mathbf{u}(t)$  sono arbitrarie. Ragionando in maniera intuitiva, se assumiamo  $\delta\mathbf{u}(t)$  nella seguente forma

$$\delta\mathbf{u}(t) = \begin{cases} \overline{\delta\mathbf{u}} & \text{per } t_1 - \epsilon < t < t_1 + \epsilon \\ 0 & \text{altrimenti} \end{cases}$$

con  $t_1$  fissato,  $\overline{\delta \mathbf{u}}$  arbitrario, e  $\epsilon$  sufficientemente piccolo, si ricava

$$\delta \mathcal{H}(t_1) \geq 0 \quad \forall \overline{\delta \mathbf{u}}$$

Tale risultato esprime il fatto che in ciascun istante  $t_1 \in (t_0, T)$  la funzione  $\mathbf{u} \rightarrow \mathcal{H}(\mathbf{x}(t_1; \mathbf{u}), \mathbf{u}, \mathbf{y}(t_1; \mathbf{u}), t_1)$  deve avere un minimo locale per  $\mathbf{u} = \mathbf{u}^*(t)$ . In maniera schematica, l'interesse del risultato è dato dal fatto che il problema iniziale di minimizzazione del *funzionale*  $J$  è stato trasformato in una infinità (per  $t \in (t_0, T)$ ) di problemi di minimizzazione della funzione *scalare*  $\mathcal{H}$ . In conclusione, abbiamo il seguente risultato.

**Teorema 14.2** (Principio del minimo di Pontryagin) *Condizione necessaria affinché il controllo  $\mathbf{u}^*(t)$  e la traiettoria corrispondente  $\mathbf{x}^*(t)$  siano ottimali, è che esista un vettore aggiunto  $\mathbf{y}^*(t)$  tale che*

1.  $\mathbf{y}^*(t)$  è soluzione dell'equazione aggiunta

$$\frac{d\mathbf{y}^*}{dt} = -\mathbf{f}_x^T \mathbf{y}^* - L_x^T \quad t \in (t_0, T) \quad (14.134)$$

con la condizione finale

$$(\mathbf{y}^*)^T(T) d\mathbf{x}_T - \mathcal{H}(T) dT = \lambda_x d\mathbf{x}_T + \lambda_t dT \quad (14.135)$$

per ogni vettore  $(d\mathbf{x}_T, dT)$  tangente a  $\mathcal{B}$  nel punto  $(\mathbf{x}(T), T)$ .

2. per ogni  $t \in (t_0, T)$ , la funzione  $\mathcal{H}(\mathbf{x}(t), \mathbf{u}, \mathbf{y}(t), t)$  raggiunge il suo minimo rispetto a  $\mathbf{u} \in \mathcal{U}(t)$  per  $\mathbf{u} = \mathbf{u}^*(t)$ , ossia

$$\mathcal{H}(\mathbf{x}^*(t), \mathbf{u}^*, \mathbf{y}^*(t), t) \leq \mathcal{H}(\mathbf{x}(t), \mathbf{u}, \mathbf{y}(t), t) \quad \forall t \in (t_0, T), \quad \forall \mathbf{u} \in \mathcal{U} \quad (14.136)$$

Se  $\mathbf{u}^*$  è un punto interno di  $\mathcal{U}$ , in particolare quando  $\mathcal{U} \equiv \mathbb{R}^m$  (caso di controlli non vincolati), la condizione precedente (14.136) equivale alla seguente equazione

$$\frac{\partial \mathcal{H}}{\partial \mathbf{u}} = L_{\mathbf{u}} + (\mathbf{y}^*)^T \mathbf{f}_{\mathbf{u}} = 0 \quad (14.137)$$

Possiamo riassumere le condizioni del teorema precedente, nel caso in cui  $\mathbf{u}^*$  sia interno a  $\mathcal{U}$  nella seguente forma

$$\boxed{\begin{aligned} \dot{\mathbf{x}}^*(t) &= \left( \frac{\partial \mathcal{H}}{\partial \mathbf{y}} \right)^T \\ \dot{\mathbf{y}}^*(t) &= - \left( \frac{\partial \mathcal{H}}{\partial \mathbf{x}} \right)^T \\ 0 &= \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \end{aligned}} \quad (14.138)$$

Esse rappresentano  $2n+m$  equazioni nelle funzioni incognite  $\mathbf{u}^*(t) \in \mathbb{R}^m$  e  $\mathbf{x}^*(t)$ ,  $\mathbf{y}^* \in \mathbb{R}^n$ . Ad esse vanno aggiunte le condizioni iniziali per la funzione di stato  $\mathbf{x}^*$  e le condizioni di trasversalità (14.135).

Una prima osservazione importante riguarda il fatto che, come abbiamo già rilevato in precedenza nel caso di funzioni di un numero finito di variabili, le condizioni di ottimalità (14.138) sono condizioni, in generale, solo *necessarie*. In altre parole, possono essere soluzioni delle equazioni (14.138) anche funzioni che non corrispondono a un minimo (o a un massimo) del funzionale  $J$ . Una indagine più approfondita sulla natura delle soluzioni di (14.138) richiede, in generale, lo studio delle variazioni seconde. Esistono, tuttavia, situazioni per le quali è possibile stabilire facilmente l'esistenza di un minimo. Segnaliamo, in particolare, il caso dei problemi di controllo relativi a un funzionale costo quadratico e a un sistema di stato lineare, che abbiamo già analizzato in precedenza nell'ambito del metodo della programmazione dinamica.

Una seconda osservazione importante nelle applicazioni, si riferisce al caso particolare in cui la funzione  $\mathcal{H}$  non dipenda esplicitamente dal tempo  $t$ , ossia si abbia  $\partial\mathcal{H}/\partial t = 0$ . In questo caso, dalle condizioni (14.138) si ricava

$$\frac{d\mathcal{H}}{dt} = \frac{\partial\mathcal{H}}{\partial\mathbf{y}} \frac{d\mathbf{y}^*}{dt} + \frac{\partial\mathcal{H}}{\partial\mathbf{x}} \frac{d\mathbf{x}^*}{dt} + \frac{\partial\mathcal{H}}{\partial\mathbf{u}} \frac{d\mathbf{u}^*}{dt} = 0 \quad (14.139)$$

Si ha pertanto che l'hamiltoniana, lungo una traiettoria ottimale, è *costante*. In particolare, se  $\lambda \equiv 0$  e  $T$  è *libero*, dalle condizioni di trasversalità si ottiene  $\mathcal{H}(T) = 0$  e, quindi, in definitiva

$$\mathcal{H}(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) = 0 \quad (14.140)$$

### 14.4.3 Metodi numerici

Nel seguito analizzeremo alcuni modelli interessanti nei quali le condizioni (14.138) permettono di ricavare in maniera analitica interessanti indicazioni sul controllo e le corrispondenti traiettorie ottimali. In generale, comunque, per la loro risoluzione sono necessari opportuni metodi numerici. In tale contesto, l'analisi precedente assume una particolare importanza, in quanto la condizione (14.137) è alla base dell'implementazione dei vari metodi di minimizzazione che utilizzano il gradiente della funzione obiettivo (cfr. Capitolo 5). A solo scopo illustrativo, esamineremo, in particolare l'implementazione del metodo del gradiente nel caso di controlli non vincolati; l'estensione agli altri metodi, quali ad esempio il metodo del gradiente coniugato, o ai metodi quasi-Newtoniani, è immediata.

**Metodo del gradiente** A partire da una funzione di tentativo  $\mathbf{u}^{(0)}(t)$ , definita per  $t \in (t_0, T)$ , si genera una successione di funzioni  $\{\mathbf{u}^{(k)}(t)\}$  mediante la seguente procedura iterativa. Supponendo di conoscere  $\mathbf{u}^{(k)}$ , si calcola  $\mathbf{u}^{(k+1)}$  attraverso i seguenti passi.

1. Si risolvono i sistemi di stato e di stato aggiunto per  $\mathbf{u} = \mathbf{u}^{(k)}$ . Indichiamo con  $\mathbf{x}^{(k)}(t)$ ,  $\mathbf{y}^{(k)}(t)$  le soluzioni ottenute.
2. Si determina  $\mathbf{u}^{(k+1)}$  mediante la seguente formula

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \rho(L_{\mathbf{u}} + (\mathbf{y}^{(k)})^T \mathbf{f}_{\mathbf{u}})_k^T \quad (14.141)$$

Se il parametro  $\rho$  è individuato come il valore che minimizza  $J(\rho)$  lungo la direzione  $\mathbf{u}^{(k)} - \rho(L_{\mathbf{u}} + (\mathbf{y}^{(k)})^T \mathbf{f}_{\mathbf{u}})_k^T$ , si ha il metodo di *massima discesa*.

Il procedimento converge, sotto opportune condizioni, ad una terna di funzioni  $\mathbf{u}^*$ ,  $\mathbf{x}^*$ ,  $\mathbf{y}^*$  che sono soluzioni delle condizioni di ottimalità (14.138), e quindi possono corrispondere a delle soluzioni ottimali. In pratica l'algoritmo viene iterato fino ad ottenere  $\|L_{\mathbf{u}} + \mathbf{y}^T \mathbf{f}_{\mathbf{u}}\| \leq \epsilon$ , con  $\epsilon$  prefissato. Naturalmente, il *costo maggiore* nell'applicazione dell'algoritmo è nella risoluzione ad ogni passo dei sistemi di stato e di stato aggiunto. A questo proposito, sottolineiamo che, salvo casi particolari, *non è possibile* in generale *disaccoppiare* i due sistemi in  $\mathbf{x}$  e  $\mathbf{y}$  come problemi a valori iniziali; si pensi, ad esempio, al caso in cui sono fissati sia  $T$  che  $\mathbf{x}(T)$ . Si tratta, quindi, in generale di risolvere un *problema ai limiti* (two-point boundary). Per esso si possono utilizzare le tecniche che abbiamo sviluppato nel Capitolo 7 sulle equazioni differenziali ordinarie, in particolare i *metodi shooting*.

► **Esempio 14.13** Come semplice illustrazione dei passi necessari nell'applicazione del metodo del gradiente, consideriamo il seguente problema di controllo. Data l'equazione di stato

$$\dot{x}(t) = -x(t) + u(t)$$

con la condizione iniziale  $x(0) = 4$ , si cerca la funzione  $u(t)$ ,  $t \in [0, 1]$  che minimizza il seguente funzionale

$$J = x^2(1) + \int_0^1 \frac{1}{2} u^2 dt$$

L'equazione di stato aggiunto è la seguente

$$\dot{y}(t) = y(t)$$

con la condizione ai limiti  $y(1) = 2x(1)$ , e dalla condizione necessaria di ottimalità si ha

$$\frac{\partial \mathcal{H}}{\partial u} = u(t) + y(t) = 0$$

Scegliendo come stima iniziale del controllo la funzione  $u^{(0)} = 1$  sull'intervallo  $[0, 1]$  e integrando il sistema di stato in corrispondenza a tale funzione, si ottiene

$$x^{(0)}(t) = 3e^{-t} + 1$$

da cui  $y^{(0)}(1) = 2x^{(0)}(1) = 2(3e^{-1} + 1)$ . Utilizzando tale valore nel sistema di stato aggiunto, si ottiene

$$y^{(0)}(t) = 2e^{-1}(3e^{-1} + 1)e^t \Rightarrow \frac{\partial \mathcal{H}^{(0)}}{\partial u}(t) = 1 + 2e^{-1}(3e^{-1} + 1)e^t$$

La successiva stima del controllo è allora fornita dalla relazione

$$u^{(1)}(t) = u^{(0)}(t) - \rho \frac{\partial \mathcal{H}^{(0)}}{\partial u}(t) \Rightarrow u^{(1)}(t) = 1 - \rho[1 + 2e^{-1}(3e^{-1} + 1)e^t]$$

In Figura 14.21 sono rappresentate le funzioni di stato  $x^{(0)}(t)$ ,  $x^{(1)}(t)$  e le funzioni di controllo  $u^{(0)}(t)$  e  $u^{(1)}(t)$  per  $\rho = 0.5$ . In corrispondenza, si ha  $J(u^{(0)}) = 4.9253$  e  $J(u^{(1)}) = 1.1878$ . ■

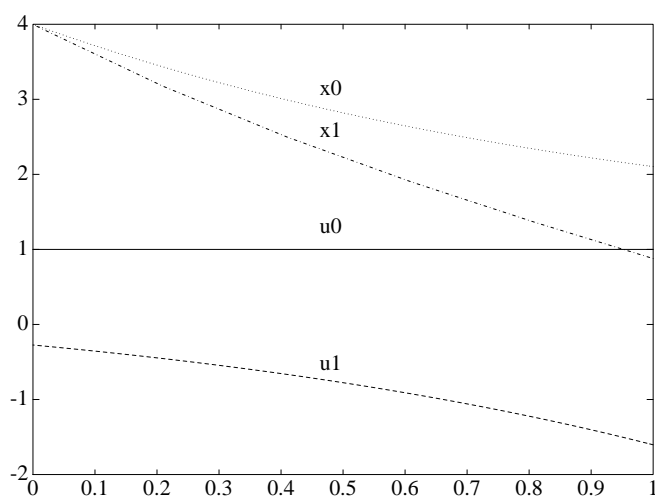


Figura 14.21: Rappresentazione della variabili di stato e di controllo corrispondenti alle prime due iterazioni del metodo del gradiente applicato all'Esempio 14.13.

#### 14.4.4 Programmazione dinamica e principio del minimo

Riprendiamo l'equazione della programmazione dinamica

$$V_t + \min_{\mathbf{u}} \{L + V_{\mathbf{x}} \mathbf{f}(\mathbf{x}, \mathbf{u}, t)\} = 0 \quad (14.142)$$

Se  $V$  è sufficientemente regolare e il minimo  $\mathbf{u}^*$  è un punto interno a  $\mathcal{U}(t)$ , la condizione (14.142) è equivalente (nel punto di minimo) alle seguenti due condizioni

$$V_t + L + V_{\mathbf{x}} \mathbf{f} = 0 \quad (14.143)$$

$$L_{\mathbf{u}} + V_{\mathbf{x}} \mathbf{f}_{\mathbf{u}} = 0 \quad (14.144)$$

Tenendo conto dell'equazione di stato, la derivata di  $V_{\mathbf{x}}$  rispetto a  $t$  lungo la traiettoria ottimale è data da

$$\frac{d}{dt} V_{\mathbf{x}} = V_{\mathbf{x}t} + V_{\mathbf{x}\mathbf{x}} \mathbf{f}$$

e, derivando (14.143) rispetto a  $\mathbf{x}$ , si ha

$$V_{\mathbf{x}t} + V_{\mathbf{x}\mathbf{x}}\mathbf{f} + L_{\mathbf{x}} + L_{\mathbf{u}}\mathbf{u}_{\mathbf{x}} + V_{\mathbf{x}}\mathbf{f}_{\mathbf{x}} + V_{\mathbf{x}}\mathbf{f}_{\mathbf{u}}\mathbf{u}_{\mathbf{x}} = 0$$

da cui, tenendo conto di (14.144), si ottiene

$$\frac{d}{dt}(V_{\mathbf{x}})^T = -\mathbf{f}_{\mathbf{x}}^T(V_{\mathbf{x}})^T - (L_{\mathbf{x}})^T$$

che coincide con l'equazione aggiunta (14.134). Pertanto, lungo una traiettoria ottimale si ha la seguente identificazione

$$\mathbf{y}(t) = V_{\mathbf{x}}^T(\mathbf{x}, t) \quad (14.145)$$

che giustifica il significato di coefficiente di sensitività dato al vettore aggiunto  $\mathbf{y}(t)$ .

#### 14.4.5 Legame con il calcolo delle variazioni

Con il termine *calcolo delle variazioni* si intende l'insieme delle tecniche matematiche relative allo studio dell'ottimizzazione di un funzionale nell'ambito di un opportuno spazio di funzioni. Tali tecniche, storicamente originate dal lavoro di Newton sulla determinazione della configurazione di un corpo che si muove nell'aria con la minima resistenza e dal lavoro di J. Bernoulli sulla brachistocrona (cfr. successivo Esempio 14.23), rappresentano uno degli strumenti più importanti nella modellistica matematica. Considerando l'equazione di stato come un vincolo, un problema di controllo è ovviamente un particolare problema di calcolo delle variazioni. In questo paragrafo mostreremo brevemente come, viceversa, alcuni problemi di calcolo delle variazioni possono essere formulati in termini di problemi di controllo, in modo da utilizzare le condizioni di ottimalità ricavate nei paragrafi precedenti.

Consideriamo il problema della minimizzazione del seguente funzionale

$$J(\mathbf{x}) = \int_{t_0}^T L(\mathbf{x}, \dot{\mathbf{x}}, t) dt$$

ove  $L$  è una funzione assegnata, nell'ambito di un insieme di funzioni  $\mathbf{x}(t)$ , opportunamente regolari e che verificano delle condizioni ai limiti assegnate, ad esempio  $\mathbf{x}(t_0) = \boldsymbol{\xi}_0$ ,  $\mathbf{x}(T) = \boldsymbol{\xi}_T$ . Se si definisce  $\mathbf{u} := d\mathbf{x}/dt$ , si ottiene un problema di controllo per il quale si ha

$$\mathcal{H} = \mathbf{y}^T \mathbf{u} + L(\mathbf{x}, \mathbf{u}, t)$$

con

$$\frac{d\mathbf{y}}{dt} = -L_{\mathbf{x}}^T$$

Il principio del minimo applicato a  $\mathcal{H}$  fornisce la condizione

$$\mathbf{y} + L_{\mathbf{u}}^T = 0$$

da cui la seguente condizione necessaria, nota nel calcolo delle variazioni come *equazione di Eulero* (o Eulero–Lagrange)

$$\boxed{\frac{d}{dt}L_{\dot{x}}^T - L_x^T = 0}$$

► **Esempio 14.14** Come illustrazione, consideriamo il problema della ricerca della curva  $x(t)$  di minima lunghezza che congiunge i punti  $A = (0,0)$  e  $B = (1,1)$ . Se  $x(t)$  è sufficientemente regolare, la lunghezza della curva è fornita dal seguente integrale

$$J = \int_0^1 \sqrt{1 + \left(\frac{dx}{dt}\right)^2} dt \quad (14.146)$$

Pertanto, il problema posto consiste nella ricerca del minimo del funzionale (14.146) nello spazio delle funzioni che sono opportunamente regolari (ossia con derivata prima di quadrato integrabile) e che verificano le seguenti condizioni ai limiti

$$x(0) = 0, \quad x(1) = 1 \quad (14.147)$$

In questo caso la funzione  $L(x) = \sqrt{1 + \dot{x}^2}$  non dipende esplicitamente da  $x$  e l'equazione di Eulero si riduce alla seguente equazione

$$\frac{d}{dt}(L_{\dot{x}}) = \frac{d}{dt}\dot{x}(1 + \dot{x}^2)^{-1/2} = 0$$

dalla quale si ha  $\dot{x}^* = k$ , con  $k$  costante. Si ha pertanto  $x^*(t) = kt + d$ , ove le costanti  $k$  e  $d$  sono determinate dalle condizioni ai limiti (14.147), e quindi  $d = 0$  e  $k = 1$ . La curva ottimale è, come era intuitivo, la retta  $x^*(t) = t$ .

Se come condizioni ai limiti, anziché le (14.146), imponiamo  $x(0) = 0$ , mentre  $x(t)$  è libera in  $t = 1$ , la condizione di trasversalità fornisce la condizione  $y(1) = 0$ , ossia

$$\dot{x} [1 + \dot{x}^2]^{-1/2} = 0 \quad \text{per } t = 1$$

da cui  $\dot{x}(1) = 0$ . Dall'equazione di Eulero e dalla condizione  $x(0) = 0$  si ottiene ancora  $x(t) = kx$  e quindi in definitiva si ha la soluzione ottimale  $x^*(t) = 0$  sull'intervallo  $[0, 1]$ .

Come ulteriore esempio interessante, consideriamo la configurazione  $x(t)$  assunta da un filo perfettamente elastico soggetto ad un campo di forze trasversale di intensità  $f(t)$  e fissato agli estremi  $t = 0, t = T$ . Tale configurazione corrisponde al minimo del seguente integrale, che rappresenta l'energia elastica del sistema

$$J(x) = \frac{1}{2} \int_0^T \dot{x}^2(t) dt - \int_0^T f(t)x(t) dt$$

nell'ambito delle funzioni sufficientemente regolari e tali che  $x(0) = x(T) = 0$ . L'equazione di Eulero fornisce in questo caso il seguente noto problema ai limiti (cfr. Capitolo 7)

$$\begin{matrix} L_{\dot{x}} = \dot{x} \\ L_x = -f(t) \end{matrix} \quad \Rightarrow \quad \boxed{\begin{cases} -\dot{x} = f(t) \\ x(0) = x(T) = 0 \end{cases}}$$

■

### 14.4.6 Applicazioni; modelli matematici

In questo paragrafo illustreremo il principio del minimo mediante la sua applicazione a differenti problemi, alcuni dei quali rappresentativi di situazioni importanti nella matematica applicata.

► **Esempio 14.15** Come esempio introduttivo, consideriamo il problema di controllo definito dal seguente sistema di stato

$$\begin{cases} \frac{dx_1}{dt} = x_2 \\ \frac{dx_2}{dt} = u \end{cases} \quad (14.148)$$

con le condizioni  $x_1(0) = x_2(0) = 0$ ,  $x_1(1) = 1$ , mentre  $x_2(1)$  è non fissato, e dal seguente funzionale da minimizzare

$$J(u) = \int_0^1 (x_2(t) + u^2(t)) dt \quad (14.149)$$

La corrispondente funzione hamiltoniana è definita come segue

$$\mathcal{H} = x_2(t) + u^2(t) + y_1(t)x_2(t) + y_2(t)u \quad (14.150)$$

Osservando che

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} x_2 \\ u \end{bmatrix}, \quad L = x_2 + u, \quad \Rightarrow \quad \mathbf{f}_{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad L_{\mathbf{x}} = [0, 1]$$

il sistema aggiunto è dato da

$$\begin{cases} \frac{dy_1}{dt} = 0 \\ \frac{dy_2}{dt} = -(y_1 + 1) \end{cases} \quad (14.151)$$

Dal momento che l'istante finale  $T = 1$  è fissato,  $x_2(1)$  è libero ed inoltre  $\lambda \equiv 0$ , la condizione di trasversalità in questo caso fornisce la seguente quarta condizione ai limiti

$$y_2(1) = 0 \quad (14.152)$$

Le soluzioni del sistema (14.151) sono  $y_1(t) = k_1$ ,  $y_2(t) = -(k_1 + 1)t + k_2$  con  $k_1, k_2$  costanti da determinare. Supponendo che il controllo  $u$  sia non vincolato, si ha la seguente condizione di ottimalità

$$\frac{\partial \mathcal{H}}{\partial u} = 2u + y_2 = 0 \quad \Rightarrow \quad u = -\frac{y_2}{2} \quad (14.153)$$

Sostituendo tale valore nel sistema di stato (14.148) e integrando, si ottiene

$$x_2(t) = \frac{1}{4}(k_1 + 1)t^2 - \frac{1}{2}k_2t + k_3 \quad (14.154)$$

Dalla condizione iniziale  $x_2(0) = 0$  si ottiene  $k_3 = 0$ . Dalla prima equazione del sistema (14.148) si ottiene, sostituendo la precedente espressione di  $x_2$  e tenendo conto della condizione iniziale  $x_1(0) = 0$

$$x_1(t) = \frac{1}{12}(k_1 + 1)t^3 - \frac{1}{4}k_2t^2$$



Utilizzando, infine, la condizione  $x_1(1) = 1$  e la condizione  $y_2(1) = 0$ , si ricavano le seguenti condizioni

$$\begin{cases} k_1 - 3k_2 = 11 \\ -k_1 + k_2 = 1 \end{cases} \Rightarrow k_1 = -7; k_2 = -6$$

In definitiva, il sistema di ottimalità ha la seguente soluzione

$$u^* = -3(t-1); \quad \mathbf{x}^* = \begin{bmatrix} -1/2t^3 + 3/2t^2 \\ -3/2t^2 + 3t \end{bmatrix}; \quad \mathbf{y}^* = \begin{bmatrix} -7 \\ 6(t-1) \end{bmatrix}$$

rappresentata in Figura 14.22. Tale soluzione corrisponde effettivamente ad un minimo, in quanto  $\partial^2 \mathcal{H} / \partial u^2 = 2 > 0$ . ■

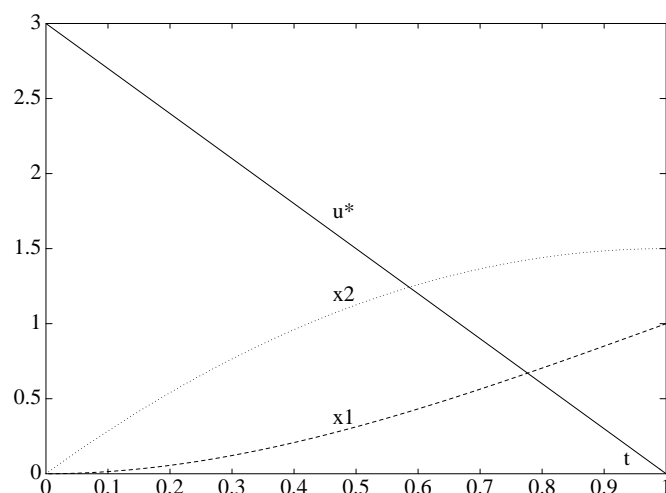


Figura 14.22: Rappresentazione della soluzione ottimale del problema di controllo esaminato nell'Esempio 14.15.

► **Esempio 14.16** *Sistema lineare e controllo quadratico* Problemi di controllo di questo tipo sono stati risolti nel precedente Esempio 14.9 mediante il metodo della programmazione dinamica. In particolare, si è mostrato che la sua risoluzione può essere ricondotta alla risoluzione di un'equazione differenziale di Riccati. Mostriamo ora che la medesima condizione può essere ottenuta applicando il principio del minimo. Ne segue che tale condizione risulta

una condizione necessaria e sufficiente. Riassumiamo i dati del problema

$\frac{d\mathbf{x}}{dt} = \mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{u}$ $\mathbf{x}(t_0) = \boldsymbol{\xi}$ $T : \text{fissato}, \quad \mathbf{x}(T) : \text{libero}$ $L = \frac{1}{2}(\mathbf{x}^T \mathbf{Q}\mathbf{x} + \mathbf{u}^T \mathbf{R}\mathbf{u}); \quad \lambda = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x}$	$\frac{d\mathbf{y}}{dt} = -\mathbf{F}^T \mathbf{y} - \mathbf{Q}\mathbf{x}$ $\mathbf{y}(T) = \mathbf{A}\mathbf{x}$
--	---

↓

$$\mathcal{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{y}(t), t) = \frac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} + \frac{1}{2}\mathbf{u}^T \mathbf{R}\mathbf{u} + \mathbf{y}^T (\mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{u})$$

Dalla condizione di ottimalità si ricava

$$0 = \frac{\partial \mathcal{H}}{\partial \mathbf{u}} = \mathbf{R}\mathbf{u}^* + \mathbf{G}^T \mathbf{y}^* \Rightarrow \mathbf{u}^* = -\mathbf{R}^{-1} \mathbf{G}^T \mathbf{y}^*$$

da cui, sostituendo nelle equazioni di stato e stato aggiunto, si ottengono le seguenti equazioni ottimali

$$\begin{aligned} \frac{d\mathbf{x}^*}{dt} &= \mathbf{F}\mathbf{x}^* - \mathbf{G}\mathbf{R}^{-1} \mathbf{G}^T \mathbf{y}^*; & \mathbf{x}(t_0) &= \boldsymbol{\xi} \\ \frac{d\mathbf{y}^*}{dt} &= -\mathbf{Q}\mathbf{x}^* - \mathbf{F}^T \mathbf{y}^*; & \mathbf{y}^*(T) &= \mathbf{A}\mathbf{x}^* \end{aligned}$$

Si può verificare facilmente che le soluzioni  $\mathbf{x}^*(t)$  e  $\mathbf{y}^*(t)$  sono legate dalla seguente relazione

$$\mathbf{y}^* = \mathbf{P}(t)\mathbf{x}^*$$

ove  $\mathbf{P}(t)$  è la soluzione del sistema di Riccati (14.68) introdotto nell'Esempio 14.9. ■

► **Esempio 14.17** (*Massima distanza e minimo sforzo*) Il problema che esamineremo in questo esempio è rappresentativo di diverse situazioni pratiche, nelle quali si hanno due (o più) obiettivi da raggiungere tra di loro antitetici. Assumendo come funzione di controllo l'accelerazione, si vuole far raggiungere, in un tempo fissato, ad una vettura la *massima* distanza, con il *minimo* sforzo. In questo caso si ha che ad un aumento dell'accelerazione corrisponde un aumento sia della distanza che dello sforzo impiegato. Il problema è, quindi, quello di ottenere una soluzione ottimale di compromesso. Un modo per risolvere il problema consiste nel considerare, come funzionale  $J$  da ottimizzare, una opportuna combinazione dei due obiettivi

$$J = c_1 J_1 + c_2 J_2$$

ove mediante i coefficienti  $c_1$  e  $c_2$  è possibile *pesare* opportunamente i due differenti obiettivi.

Indicando con  $d(t)$  la distanza percorsa dalla vettura, la legge di moto (ossia il sistema di stato) è, previa una opportuna normalizzazione, la seguente

$$\ddot{d}(t) = u(t), \quad d(0) = 0, \quad \dot{d}(0) = 0 \quad (14.155)$$

L'obiettivo  $J_1(u)$  da massimizzare nel tempo  $T$  è quindi dato da  $J_1(u) = d(T)$ . Lo sforzo relativo all'intervallo  $(0, T)$  può essere modellizzato come una quantità proporzionale all'integrale del quadrato dell'accelerazione  $u(t)$ ; pertanto, previa normalizzazione,  $J_2(u) = 1/2 \int_0^T u^2(t) dt$ . In definitiva, possiamo assumere come obiettivo da *massimizzare* il seguente funzionale

$$J(u) := c_1 d(T) - \frac{c_2}{2} \int_0^T u^2(t) dt \quad (14.156)$$

con  $c_1 > 0$ ,  $c_2 > 0$  costanti fissate. Posto  $x_1 = d$ ;  $x_2 = \dot{d}$ , il problema può essere scritto nella seguente forma

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad \mathbf{f}_{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{f}_u = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \mathbf{x}(0) &= 0 \\ J &= c_1 x_1(T) - \frac{c_2}{2} \int_0^T u^2 dt, \quad L = -\frac{c_2}{2} u^2, \quad \lambda = c_1 x_1 \end{aligned}$$

a cui corrisponde il seguente sistema aggiunto

$$\begin{cases} \dot{y}_1(t) = 0 \\ \dot{y}_2(t) = -y_1 \end{cases}$$

con le condizioni finali

$$y_1(T) = c_1; \quad y_2(T) = 0$$

Risolvendo le equazioni aggiunte, si ottiene

$$y_1(t) = c_1; \quad y_2(t) = c_1(T - t)$$

La funzione hamiltoniana è data da

$$\mathcal{H}(x, u, y, t) = -\frac{c_2}{2} u^2 + y_1 x_2 + y_2 u$$

Nell'ipotesi che il controllo  $u$  non sia vincolato, si ha la seguente condizione di ottimalità

$$\frac{\partial \mathcal{H}}{\partial u} = 0$$

da cui

$$\boxed{u^*(t) = \frac{y_2(t)}{c_2} \Rightarrow u^*(t) = \frac{c_1}{c_2} (T - t)}$$

Lasciamo come esercizio l'estensione delle considerazioni precedenti al caso in cui nell'equazione della dinamica si tiene conto dell'attrito, ossia  $\ddot{d} = u - k\dot{d}$ , con  $k > 0$ . ■

► **Esempio 14.18** (*Controllo vincolato*) In questo esempio elementare cominceremo a considerare il caso in cui la funzione controllo è soggetta a vincoli e quindi non è possibile, in generale, scrivere la condizione di ottimalità (14.136) nella forma di equazione (14.137).

Consideriamo il seguente problema di calcolo delle variazioni. Si cerca una curva di equazione  $x(t)$ ,  $t \in [0, T]$  opportunamente regolare e tale che, a partire da  $x(0) = 0$  raggiunga

la massima altezza in  $T$ , con il vincolo  $x'(t) \leq 1$ . Il problema può essere tradotto in forma di problema di controllo nel modo seguente.

$$\boxed{\begin{array}{l} \dot{x}(t) = u(t), \quad t \in (0, T) \\ x(0) = 0 \\ u(t) \leq 1 \\ J = x(T) \end{array}} \Rightarrow \begin{array}{l} f_x = 0 \\ L = 0; \quad \lambda = x(T) \end{array}$$

Il sistema aggiunto è dato allora da

$$\dot{y} = 0; \quad y(T) = 1 \Rightarrow y(t) = 1$$

L'hamiltoniana è data da

$$\mathcal{H} = yu = u$$

Il massimo della funzione  $\mathcal{H}$  per  $u \leq 1$  si ottiene, ovviamente, per  $u^* = 1$ , a cui corrisponde, come era intuitivo, la traiettoria ottimale  $x^* = t$ . ■

► **Esempio 14.19** (*Insetti come ottimizzatori*) Questo esempio, tratto da Macevicz e Oster, *Modelling social insect population II: Optimal reproductive strategies in annual eusocial insect colonies*, Behavioral Ecol. Sociobiol., I, 265-282, 1976, introduce il concetto di *controllo bang-bang*, corrispondente al caso in cui la funzione controllo è soggetta a vincoli bilaterali.

Diversi tipi di insetti, come le vespe, i calabroni, vivono in colonie consistenti di due classi: i lavoratori e i riproduttori (regine e maschi). Eccetto che per le giovani regine, la colonia muore alla fine dell'estate. Le regine possono iniziare una nuova colonia nella successiva primavera. Per massimizzare il numero delle colonie fondate nel successivo anno, ogni colonia dovrebbe massimizzare il numero dei riproduttori alla fine dell'anno corrente. Possiamo interpretare tale processo di massimizzazione mediante il seguente modello matematico.

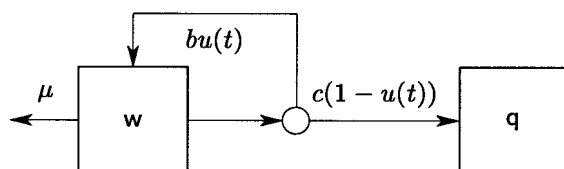


Figura 14.23: Sistema dinamico di una colonia di insetti.

Indicata con  $T$  la durata della stagione di vita, chiamiamo  $w(t)$  e  $q(t)$  rispettivamente il numero di lavoratori e di riproduttori viventi al tempo  $t \in [0, T]$ . Assumiamo che ad ogni istante  $t$  una frazione  $u(t)$  dei lavoratori dedichi i suoi sforzi per aumentare la forza di lavoro e la rimanente frazione  $1 - u(t)$  per generare riproduttori. Indichiamo con  $b$  e  $c$  due costanti positive, dipendenti dall'ambiente e che rappresentano il tasso di velocità al quale ogni lavoratore produce rispettivamente nuovi lavoratori e nuovi riproduttori. Le larve femmine possono diventare lavoratori o riproduttori, in dipendenza della quantità (e qualità) di cibo con cui vengono nutrite (ad esempio, la pappa reale per le api). Il tasso di mortalità dei lavoratori è supposto costante e indicato con  $\mu$ . Per semplicità, sarà supposto nullo il tasso di

mortalità dei riproduttori; supporremo, inoltre che sia  $b > \mu$ , ossia supporremo che la colonia sia riproduttiva durante la stagione. Con le precedenti ipotesi, il modello, rappresentato in maniera schematica in Figura 14.23, può essere tradotto in termini matematici nella seguente forma

$$\begin{cases} \dot{w}(t) = b u(t) w(t) - \mu w(t) \\ \dot{q}(t) = c(1 - u(t)) w(t) \end{cases} \quad (14.157)$$

con le seguenti condizioni iniziali

$$w(0) = 1, \quad q(0) = 0 \quad (14.158)$$

e il vincolo

$$0 \leq u \leq 1 \quad (14.159)$$

Le condizioni iniziali (14.158) sono motivate dal fatto che la regina che fonda la colonia è contata come un lavoratore, in quanto, a differenza dei successivi riproduttori, deve cercare il cibo per la prima covata.

Il problema è pertanto quello di trovare la funzione  $u^*(t)$  che *massimizza*, sotto il vincolo (14.159), il seguente funzionale obiettivo

$$J(u) = q(T)$$

L'hamiltoniana del problema è fornita dalla seguente funzione

$$\mathcal{H} = y_1(bu - \mu)w + y_2c(1 - u)w = w(y_1b - y_2c)u + (y_2c - y_1\mu)w \quad (14.160)$$

Dal momento che  $\mathcal{H}$  è una funzione lineare in  $u$  e  $w > 0$  (come si vede facilmente dall'equazione di stato), essa assume il massimo per  $u^* = 0$  oppure per  $u^* = 1$ , a seconda che la quantità  $y_1b - y_2c$  è rispettivamente negativa o positiva. Per il calcolo di  $u(t)$  è necessario quindi studiare la variazione delle variabili di stato aggiunto. Il sistema di stato aggiunto è il seguente

$$\dot{y}_1(t) = -(bu(t) - \mu)y_1(t) - c(1 - u(t))y_2(t) \quad (14.161)$$

$$\dot{y}_2(t) = 0 \quad (14.162)$$

$$y_1(T) = 0, \quad y_2(T) = 1 \quad (14.163)$$

Dalle condizioni (14.163) si ha  $y_1(T)b - y_2(T)c = -c < 0$  e di conseguenza  $u^*(T) = 0$ . L'equazione aggiunta  $\dot{y}_2 = 0$  e la condizione  $y_2(T) = 1$  implicano  $y_2(t) = 1$  per  $t \in [0, T]$ . Quindi, in un opportuno intorno di  $t = T$ , l'equazione aggiunta (14.161) diventa

$$\dot{y}_1(t) = \mu y_1(t) - c \Rightarrow y_1(t) = \frac{c}{\mu} (1 - e^{\mu(t-T)}) \quad (14.164)$$

Ne segue che, andando all'indietro, ossia per  $t$  che diminuisce, la funzione  $y_1(t)$  aumenta a partire dal valore 0 in  $T$ ; indichiamo con  $t_s$  il valore di  $t$  tale che  $y_1(t_s) = c/b$ . In corrispondenza a tale valore il controllo ottimale  $u^*$  passa da 0 a 1 e la prima equazione dello stato aggiunto diventa

$$\dot{y}_1(t) = -(b - \mu)y_1(t)$$

e poiché  $b > \mu$  si ha che, andando all'indietro,  $y_1(t)$  continua ad aumentare e  $u$  rimane uguale a 1. Si verifica facilmente che il punto  $t_s$  in cui avviene lo switch è dato da

$$t_s = T + \frac{1}{\mu} \ln \left( 1 - \frac{\mu}{b} \right)$$

Nella Figura 14.24 è rappresentata la dinamica della popolazione di insetti in corrispondenza ai valori dei parametri  $c = 0.05$ ,  $b = 0.05$ ,  $\mu = 0.008$  e  $T = 150$ , a cui corrisponde il valore di switch ottimale  $t_s = 128.20$ .

In conclusione, il modello matematico indica che la procedura ottimale seguita dalla colonia consiste nel produrre solo lavoratori fino ad un certo tempo critico  $t_s$  e a partire da tale istante solo riproduttori. Tale istante dipende dai parametri  $b$ ,  $\mu$  e  $c$  del modello, e quindi anche dalle condizioni ambientali. È superfluo osservare che per gli insetti la risoluzione del problema avviene in maniera inconscia, in quanto è scritto nel loro codice genetico come risultato del processo naturale di evoluzione e di selezione naturale. ■

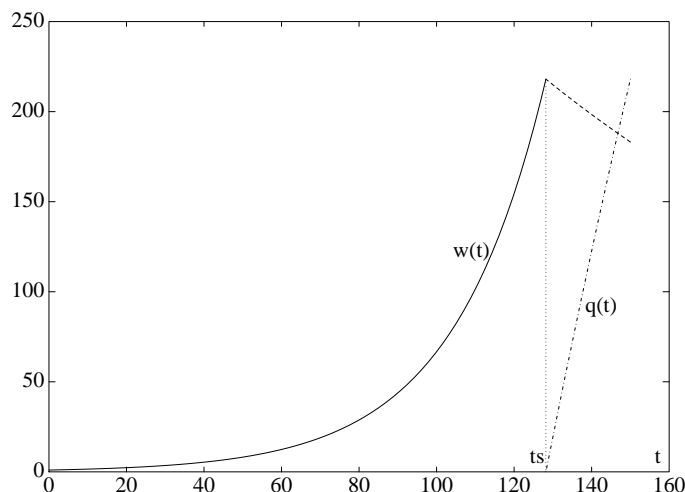


Figura 14.24: Dinamica della popolazione di insetti descritta dal modello (14.157) per  $c = 0.05$ ,  $b = 0.05$ ,  $\mu = 0.008$  e  $T = 150$ . L'istante di switch ottimale è dato da  $t_s = 128.20$ .

► **Esempio 14.20** (*chemioterapia ottimale*) Come ulteriore illustrazione dei problemi di controllo di tipo bang-bang, riprendiamo l'Esempio 14.3, precisato nel seguente modo

$$\begin{cases} \frac{dx_1}{dt} = (a - ku)x_1 + bx_2 \\ \frac{dx_2}{dt} = cx_1 - dx_2 \end{cases} \quad (14.165)$$

ove  $x_1(t)$ , rispettivamente  $x_2(t)$ , rappresenta il numero delle cellule che proliferano, rispettivamente che si trovano nel ciclo cellulare;  $b$ ,  $c$  e  $d$  sono costanti non negative, mentre  $a$  è

una costante qualunque. La costante  $k > 0$  dipende dal dosaggio massimo del farmaco somministrato. La quantità di farmaco è descritta dalla funzione  $u = u(t)$ , che per il significato dato alla costante  $k$ , verifica la condizione

$$0 \leq u \leq 1 \quad (14.166)$$

Dato il numero iniziale di cellule

$$x_1(0) = x_{10}; \quad x_2(T) = x_{20} \quad (14.167)$$

l'obiettivo della terapia è quello di ridurre in un intervallo di tempo  $T$  fissato il numero delle cellule a

$$x_1(T) = x_{11}; \quad x_2(T) = x_{21} \quad (14.168)$$

usando la minima quantità di farmaco. La funzione costo è quindi la seguente

$$J(u) = \int_0^T u \, dt \quad (14.169)$$

da minimizzare. In questo caso l'insieme bersaglio è ridotto a un punto e la funzione hamiltoniana è definita come segue

$$\mathcal{H} = u + y_1[(a - ku)x_1 + bx_2] + y_2[cx_1 - dx_2] \quad (14.170)$$

Come nell'esempio precedente, la hamiltoniana è lineare in  $u$  e quindi la determinazione del controllo ottimale richiede la conoscenza, ad ogni istante  $t \in [0, T]$ , del segno della seguente funzione (*funzione switching*)

$$S(t) = 1 - y_1(t)x_1(t)k \quad \Rightarrow \quad u^* = \begin{cases} 1 & \text{se } S(t) < 0 \\ 0 & \text{se } S(t) > 0 \end{cases} \quad (14.171)$$

La funzione controllo  $u(t)$  risulta non definita quando  $S(t) = 0$  su un intervallo di lunghezza finita in  $[0, T]$ ; in questo caso si ha un *problema di controllo singolare*. Per determinare il comportamento della funzione  $S(t)$ , è necessario studiare il sistema aggiunto, ossia il sistema differenziale

$$\dot{y}_1 = -(a - ku)y_1 - cy_2 \quad (14.172)$$

$$\dot{y}_2 = -by_1 + dy_2 \quad (14.173)$$

e la derivata di  $S(t)$ , o equivalentemente del prodotto  $y_1(t)x_1(t)$ ; si vede facilmente che

$$\frac{d}{dt}(y_1x_1) = by_1x_2 - cy_2x_1, \quad \frac{d}{dt}(y_2x_2) = -\frac{d}{dt}(y_1x_1) \quad (14.174)$$

Si possono, a questo punto, distinguere differenti casi a seconda del valore dei parametri  $c, b$ . Il caso più semplice corrisponde ai valori dei parametri  $b = c = 0$ , per il quale  $S(t)$  è una funzione costante su  $[0, T]$ . Tuttavia, come si vede su esempi, si può avere  $S(t) \equiv 0$ , nel qual caso la funzione  $u$  non è univocamente determinata. Per una discussione degli altri casi, rinviamo a Eisen [53]. ■

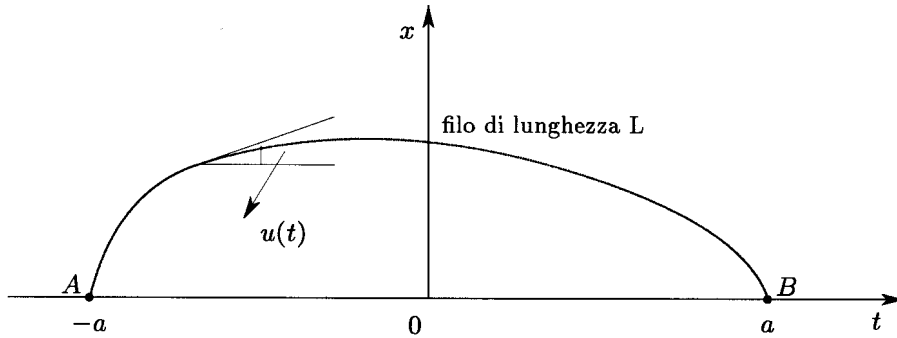


Figura 14.25: Problema isoparametrico.

► **Esempio 14.21** (*Problema isoparametrico: massima area con perimetro assegnato*) In questo esempio, che tratta un problema classico di calcolo delle variazioni, viene considerato un vincolo sulla funzione controllo di tipo più generale.

Dato un filo di lunghezza  $L$  fissato agli estremi di un segmento di retta di lunghezza  $2a < L$ , si tratta di trovare la forma del filo per la quale l'area della superficie compresa tra il filo e la retta è *massima*. Usando il sistema di coordinate indicato in Figura 14.25, il problema è quello di trovare  $u(t)$  che massimizza

$$J = \int_{-a}^a x(t) dt \quad (14.175)$$

con il *vincolo* che il perimetro (la lunghezza del filo) sia fissato, ossia

$$L = \int_{-a}^a \sqrt{1 + \left(\frac{dx}{dt}\right)^2} dt = \int_{-a}^a \sec u(t) dt \quad (14.176)$$

ove  $\sec u = 1/\cos u$  e  $u$  è l'angolo corrispondente alla tangente<sup>6</sup>

$$\frac{dx}{dt} = \tan u \quad (14.177)$$

In questo caso il target set è ridotto al punto fissato  $(0, a)$ , mentre il valore iniziale è dato da  $(0, -a)$ .

Il vincolo (14.176) sulla funzione controllo  $u$  può essere eliminato introducendo un'ulteriore variabile di stato, che rappresenta la lunghezza del filo per ogni  $t$

$$\begin{aligned} \dot{z} &= \sec u(t) \\ z(-a) &= 0, \quad z(a) = L \end{aligned} \quad (14.178)$$

L'hamiltoniana del sistema è quindi

$$\mathcal{H} = x + y_1 \tan u + y_2 \sec u \quad (14.179)$$

<sup>6</sup>La formulazione assume l'ipotesi  $-\pi/2 < u < \pi/2$ , che è assicurata se  $L < \pi a$ .



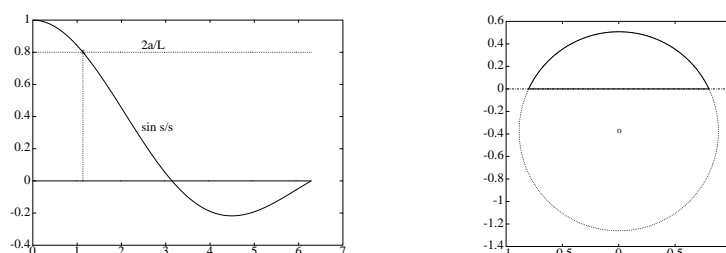


Figura 14.26: Problema isoperimetrico corrispondente ai dati  $a = 0.8$ ,  $L = 2$ . Nella prima figura è rappresentata la radice dell'equazione (14.187); nella seconda figura è rappresentata la circonferenza con centro in  $(0, -L \cos s / 2s)$  e raggio  $L/2s$ ; in solido è indicato l'arco di cerchio definito in (14.188).

Le equazioni aggiunte sono

$$\begin{aligned} \dot{y}_1 &= -1 \Rightarrow y_1 = -t + c, \quad \text{ove } c = \text{costante} \\ \dot{y}_2 &= 0 \Rightarrow y_2 = \text{costante} \end{aligned} \quad (14.180)$$

Essendo ora  $u$  non vincolata, la condizione di ottimizzazione è ottenuta ponendo uguale a zero la derivata di  $\mathcal{H}$  rispetto a  $u$

$$0 = \frac{\partial \mathcal{H}}{\partial u} = y_1 \sec^2 u + y_2 \tan u \sec u \Rightarrow \sin u = -\frac{y_1}{y_2} \quad (14.181)$$

Eliminando  $y_1$  dalle equazioni precedenti, si ottiene

$$t = y_2 \sin u + c \quad (14.182)$$

Dal momento che  $\mathcal{H}$  non è una funzione esplicita di  $t$ , si ha  $\mathcal{H} = \text{costante}$ , per cui tenendo conto delle relazioni (14.181), (14.179) si ha

$$x = -y_2 \cos u + \mathcal{H}, \quad \text{ove } \mathcal{H} = \text{costante} \quad (14.183)$$

Il perimetro è ottenuto sostituendo (14.182) nella (14.176)

$$L = \int_A^B \sec u \frac{dt}{du} du = y_2 \int_A^B du = y_2 (u_B - u_A) \quad (14.184)$$

Per valutare le cinque incognite  $c$ ,  $\mathcal{H}$ ,  $y_2$ ,  $u_A$  e  $u_B$ , si hanno a disposizione la condizione (14.184) e le quattro condizioni ai limiti

$$t(u_A) = -a, \quad t(u_B) = a, \quad x(u_A) = 0, \quad x(u_B) = 0 \quad (14.185)$$

Si ottiene facilmente la seguente soluzione

$$c = 0, \quad \mathcal{H} = -\frac{L \cos s}{2s}, \quad y_2 = -\frac{L}{2s}, \quad u_A = s, \quad u_B = -s \quad (14.186)$$

ove  $s$  è determinata dalla seguente equazione

$$\frac{\sin s}{s} = \frac{2a}{L} \quad (14.187)$$

In conclusione, si ha

$$t = -\frac{L}{2s} \sin u, \quad x = \frac{L}{2s} (\cos u - \cos s) \Rightarrow t^2 + \left(x + \frac{L \cos s}{2s}\right)^2 = \frac{L^2}{4s^2} \quad (14.188)$$

che è l'equazione di un *arco circolare* con centro in  $t = 0$  e  $x = -(L \cos s / 2s)$ , e raggio  $L/2s$ . Come esemplificazione, in Figura 14.26 sono rappresentati i risultati corrispondenti ai dati  $a = 0.8$ ,  $L = 2$ . ■

► **Esempio 14.22** (*Strategie ottimali in Immunologia*) Questo esempio introduce in un contesto biologico di grande interesse i problemi di controllo a *tempo ottimale*.

La funzione primaria del sistema immunitario è quella di eliminare l'*antigene* presente nell'organismo. Esso esercita questa funzione mediante la produzione di *anticorpi* da parte di un complesso sistema cellulare, rappresentato in maniera schematica in Figura 14.27. I linfociti  $L(t)$  possono, quando il sistema immunitario è attivato, *duplicarsi* oppure "trasformarsi" in plasmacellule  $P(t)$ , che sono caratterizzate da una maggiore capacità, rispetto ai linfociti  $L$ , di produrre anticorpo  $A(t)$ .

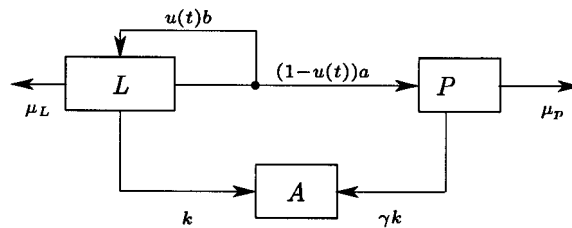


Figura 14.27: Rappresentazione schematica del sistema immunitario.

Si può pensare che durante l'evoluzione il sistema immunitario si sia perfezionato in maniera che la *risposta*, in condizioni normali, sia *ottimale*. L'ottimalità può significare, ed è questo il problema che ora considereremo, che la produzione di una quantità di anticorpo, diciamo  $A^*$ , sufficiente ad eliminare una quantità assegnata di antigene, *avvenga nel tempo minimo*. In termini matematici, il sistema immunitario determina ad ogni istante  $t$  la frazione ottimale  $u^*(t)$  di  $L(t)$  che si duplica ancora in  $L(t)$ , mentre la frazione  $1 - u^*(t)$  si trasforma in  $P(t)$ .

Un possibile modello matematico (equazione di stato) che descrive l'evoluzione nel tempo del sistema immunitario, a seguito di una attivazione, che supponiamo istantanea al tempo  $t = 0$ , è il seguente

$$\begin{aligned} \frac{dL}{dt} &= bu(t)L - a(1-u(t))L - \mu_L L, & L(0) &= L_0 \\ \frac{dP}{dt} &= a(1-u(t))L - \mu_P P, & P(0) &= 0 \\ \frac{dA}{dt} &= k(L + \gamma P), & A(0) &= 0 \end{aligned}$$

Il problema di controllo consiste nella ricerca di  $u(t)$ , con il vincolo  $0 \leq u(t) \leq 1$ , in maniera che

$$\begin{array}{l} \min_u \int_0^T dt \\ A(T) = A^* \end{array}$$

ove  $A^*$  indica la quantità di anticorpo necessaria per eliminare la quantità assegnata di antigene. Per questo problema l'insieme bersaglio è costituito dall'insieme  $\{L(T), P(T), A(T), T\}$ , ove  $T, L(T), P(T)$  sono liberi e  $A(T)$  vincolato.

Mediante l'applicazione del principio di minimo, si può vedere (Perelson, 1976) che la soluzione *ottimale* dipende dal rapporto reciproco tra le costanti di produzione e morte delle cellule e la quantità  $A^*$  di anticorpo previsto. In certe situazioni, quando, ad esempio, la quantità  $A^*$  è grande, si può verificare una soluzione di tipo *bang-bang*

$$\begin{array}{ll} u^* = 1, & 0 \leq t \leq t^* \\ u^*(t) = 0, & t^* \leq t \leq T \end{array}$$

ove  $t^*$  dipende da  $d, b, \mu_L, \mu_P, k, \gamma$  e da  $A^*$  e  $L_0$ . ■

► **Esempio 14.23** (*Problema di Zermelo, ossia come attraversare un fiume nel tempo minimo*) Con riferimento alla Figura 14.28, un'imbarcazione deve attraversare un fiume la cui corrente è caratterizzata in ciascun punto dal vettore  $\mathbf{c}(x_1, x_2)$ . Il modulo della velocità dell'imbarcazione rispetto all'acqua è supposto costante e indicato con  $V$ , mentre la funzione  $u(t)$  fornisce l'angolo formato dalla direzione seguita dall'imbarcazione rispetto all'asse  $x_1$ . Il problema consiste nel calcolare la funzione  $u(t)$  (la funzione controllo) in modo da minimizzare il tempo necessario per andare dal punto  $A$  al punto  $B$ . Le equazioni di

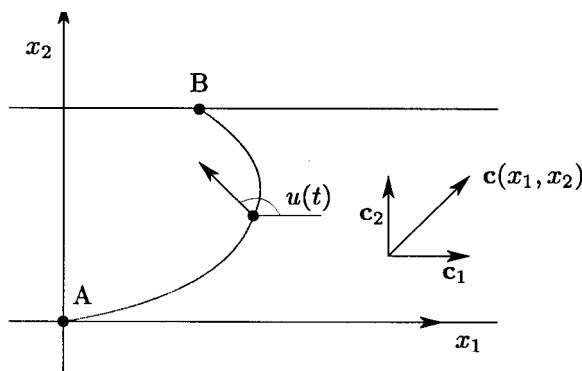


Figura 14.28: Attraversamento di un fiume; problema di tempo minimo.

moto sono

$$\begin{array}{l} \dot{x}_1 = V \cos u + c_1(x_1, x_2) \\ \dot{x}_2 = V \sin u + c_2(x_1, x_2) \end{array} \quad (14.189)$$

ove  $c_1$  e  $c_2$  rappresentano i moduli delle componenti rispetto agli assi del vettore  $\mathbf{c}$ . Le posizioni dei punti  $A$  e  $B$  forniscono le condizioni iniziali e finali, ossia quattro condizioni

(Per una discussione del sistema di stato, si veda anche Appendice B). Il funzionale costo da minimizzare, è dato da

$$J = \int_0^T dt, \quad \text{ossia} \quad L = 1, \quad \lambda = 0 \quad (14.190)$$

L'hamiltoniana del sistema è definita nel seguente modo

$$\mathcal{H} = y_1 (V \cos u + c_1) + y_2 (V \sin u + c_2) + 1 \quad (14.191)$$

Si hanno quindi le seguenti condizioni di ottimalità

$$\dot{y}_1 = -\frac{\partial \mathcal{H}}{\partial x_1} = -y_1 \frac{\partial c_1}{\partial x_1} - y_2 \frac{\partial c_2}{\partial x_1} \quad (14.192)$$

$$\dot{y}_2 = -\frac{\partial \mathcal{H}}{\partial x_2} = -y_1 \frac{\partial c_1}{\partial x_2} - y_2 \frac{\partial c_2}{\partial x_2} \quad (14.193)$$

$$0 = \frac{\partial \mathcal{H}}{\partial u} = V(-y_1 \sin u + y_2 \cos u) \Rightarrow \tan u(t) = \frac{y_2(t)}{y_1(t)} \quad (14.194)$$

Tenuto conto che l'hamiltoniana definita in (14.191) non dipende esplicitamente dal tempo  $t$  e che  $\lambda = 0$ , si ha che il valore di  $\mathcal{H}$  calcolata sulla traiettoria ottimale è uguale a zero. Risolvendo allora l'equazione  $\mathcal{H} = 0$  e l'equazione (14.194) rispetto a  $y_1$  e  $y_2$ , si ottengono le seguenti espressioni

$$y_1 = \frac{-\cos u}{V + c_1 \cos u + c_2 \sin u} \quad (14.195)$$

$$y_2 = \frac{-\sin u}{V + c_1 \cos u + c_2 \sin u} \quad (14.196)$$

e sostituendo in (14.192) (o equivalentemente in (14.193)), si ottiene la seguente equazione differenziale nella variabile  $u(t)$

$$\frac{du}{dt} = \sin^2 u \frac{\partial c_2}{\partial x_1} + \sin u \cos u \left( \frac{\partial c_1}{\partial x_1} - \frac{\partial c_2}{\partial x_2} \right) - \cos^2 u \frac{\partial c_1}{\partial x_2} \quad (14.197)$$

L'equazione (14.197), accoppiata alle equazioni di moto (14.189) e alle condizioni ai limiti determinate dalle posizioni dei punti A e B, fornisce il cammino ottimale cercato. Osserviamo che, se  $c_1$  e  $c_2$  sono costanti, si ha  $u = \text{costante}$ , e i cammini ottimali sono delle linee rette. Riassumendo, le incognite del problema sono le funzioni  $x_1(t)$ ,  $x_2(t)$ ,  $u(t)$  e il valore ottimale  $T$ ; in corrispondenza abbiamo tre equazioni differenziali e quattro condizioni ai limiti.

Nel caso particolare in cui  $c_1 = c_1(x_2)$ ,  $c_2 = c_2(x_2)$ , ossia quando la corrente varia rispetto alla sola variabile  $x_2$ , si ha  $\dot{y}_1 = 0$ , e quindi  $y_1$  è una funzione costante. Dalla (14.195) si ricava allora

$$\frac{\cos u}{V + c_1(x_2) \cos u + c_2(x_2) \sin u} = K \quad (14.198)$$

ove  $K$  è una costante che dipende dalle condizioni ai limiti. La relazione (14.198) definisce (implicitamente) la funzione controllo in termini delle velocità locali della corrente.

Come semplice illustrazione, consideriamo il caso in cui  $c_1(x_2) = c$ , con  $c$  costante,  $c_2(x_2) = 0$  e  $A = (0, 0)$ ,  $B = (-1, 1)$ . Come abbiamo già osservato, in questo caso  $u$  è una

funzione costante e la traiettoria ottimale è il segmento di retta che congiunge i punti A e B. Il valore della costante  $u$  può essere allora calcolato osservando che  $\dot{x}_2/\dot{x}_1 = -1$ , ossia risolvendo la seguente equazione

$$\frac{V \sin u}{V \cos u + c} = -1$$

Il valore ottimale di  $T$  è ottenuto dalla relazione  $T = 1/V \sin u$ , che deriva dalla condizione ai limiti  $x_2(T) = 1$ . Ad esempio, per  $V = 1$  e  $c = 0.2$  si ricava  $u = 2.4981$  ( $> \pi/2 + \pi/4 \approx 2.3562$ ) e  $T = 1.6667$ , mentre per  $V = 1$  e  $c = 0.8$  si ha  $u = 2.9575$  e  $T = 5.629$ .

Come ulteriore esemplificazione, consideriamo il caso in cui  $c_2(x_2) = 0$ , e  $c_1(x_2) = kx_2$ , con  $k$  costante assegnata, ossia il caso in cui la corrente aumenta linearmente con  $x_2$ ; assumiamo, inoltre,  $A=(0,0)$  e  $B=(0,1)$ . In questo caso le condizioni necessarie di ottimalità corrispondono alle seguenti equazioni differenziali

$$\begin{aligned} \dot{x}_1 &= V \cos u + kx_2 \\ \dot{x}_2 &= V \sin u \\ \dot{u} &= -\cos^2 u \frac{\partial c_1}{\partial x_2} \end{aligned} \quad (14.199)$$

con le condizioni ai limiti  $x_1(0) = x_2(0) = 0$  e  $x_1(T) = 0$ ,  $x_2(T) = 1$ . Per calcolare le incognite  $x_1(t)$ ,  $x_2(t)$ ,  $u(t)$  e  $T$  si può utilizzare una procedura di tipo shooting (cfr. Capitolo 7). Si assume cioè come valore iniziale  $u(0)$  un valore di tentativo  $s$ ; in corrispondenza si risolve il problema a valori iniziali relativo alle equazioni differenziali (14.199) e alle condizioni iniziali  $x_1(0) = x_2(0) = 0$  e  $u(0) = s$ . Si cercano quindi i parametri  $T$  e  $s$  risolvendo le seguenti equazioni

$$x_1(T) = 0, \quad x_2(T) = 1 \quad (14.200)$$

A scopo illustrativo in Figura 14.29 sono rappresentate le traiettorie corrispondenti ad

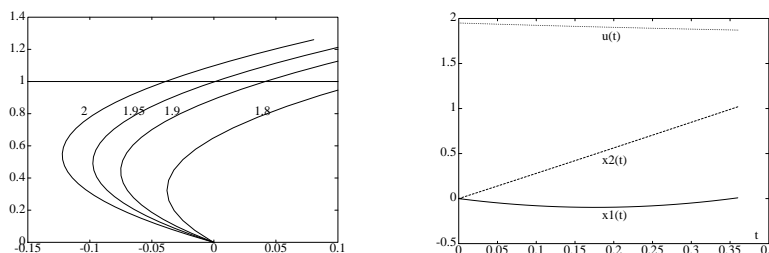


Figura 14.29: Nella prima figura sono rappresentate le traiettorie corrispondenti ai valori  $s = 2, s = 1.95, s = 1.9, s = 1.8$  per il problema di minimo tempo (14.199). Nella seconda figura sono rappresentate le funzioni  $x_1(t)$ ,  $x_2(t)$  e  $u(t)$  per  $s = 1.95$ .

alcuni valori del parametro  $s$  e in corrispondenza ai dati  $k = 2$ ,  $V = 3$ . Dalla figura si vede che la soluzione cercata si ottiene per  $s \approx 1.95$ , per il quale si ha  $T \approx 0.36$ . Concludiamo l'esempio considerando un problema analogo a quello di Zermelo. Una particella attraversa una regione con velocità data in ogni punto dal vettore  $\mathbf{V}$ , il cui modulo  $V = V(x_1, x_2)$  è una funzione assegnata della posizione  $(x_1, x_2)$ ; le equazioni di moto sono pertanto le seguenti

$$\dot{x}_1 = V(x_1, x_2) \cos u, \quad \dot{x}_2 = V(x_1, x_2) \sin u$$

ove  $u$  è l'angolo che il vettore  $\mathbf{V}$  forma con asse  $x_1$ . Si tratta ancora di trovare la funzione  $u(t)$  per la quale è minimo il tempo impiegato per passare da un punto A a un punto B. Si può mostrare che lungo un cammino ottimale la funzione  $u(t)$  deve verificare la seguente equazione differenziale

$$\dot{u} = \frac{\partial V}{\partial x_1} \sin u - \frac{\partial V}{\partial x_2} \cos u$$

Come per il problema di Zermelo, il valore  $u(0)$  e il tempo finale  $T$  sono determinati dalle condizioni finali  $(x_1(T), x_2(T)) \equiv B$ .

Considerando, in particolare, il caso in cui  $V = V(x_2)$ , ossia il caso in cui la velocità è funzione di una sola coordinata, osserviamo che allora un integrale del cammino ottimale è fornito dalla seguente equazione

$$\frac{\cos u}{V(x_2)} = k \quad (14.201)$$

con  $k$  costante opportuna. Si ha, infatti

$$\begin{cases} \dot{x}_2 = V(x_2) \sin u \\ \dot{u} = -V'(x_2) \cos u \end{cases} \Rightarrow \frac{d}{dt} \left( \frac{\cos u}{V(x_2)} \right) = \frac{-\dot{u} \sin u V(x_2) - \cos u V'(x_2) \dot{x}_2}{V(x_2)^2} = 0$$

Il risultato ora ottenuto è noto in ottica come *legge di rifrazione di Snell*.

Un altro caso particolare corrisponde al cosiddetto problema della *brachistocrona*<sup>7</sup>. Con riferimento alla Figura 14.30, una pallina scivola senza attrito su un filo che congiunge due punti A e B in un campo di forza gravitazionale costante. La pallina ha una velocità iniziale  $V_0$  nel punto A. Si tratta di trovare la configurazione del filo per la quale la pallina raggiunge il punto B nel tempo minimo. In assenza di attrito, il filo esercita sulla pallina una forza

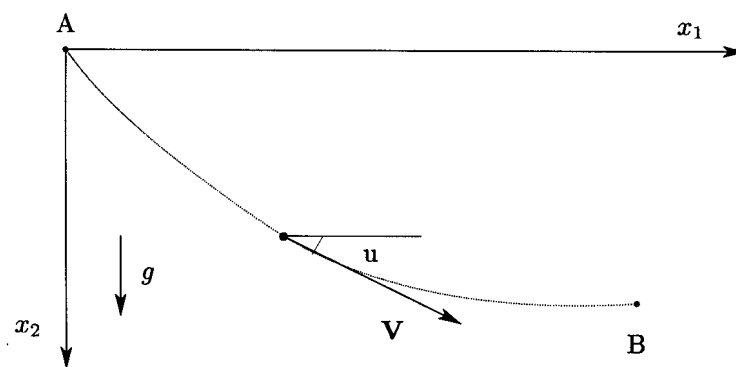


Figura 14.30: Il problema della brachistocrona.

ortogonale alla sua velocità e quindi il sistema è conservativo, ossia l'energia totale è costante

$$\frac{V^2}{2} - g x_2 = \frac{V_0^2}{2} \Rightarrow V = (V_0^2 + 2g x_2)^{1/2} =: V(x_2)$$

<sup>7</sup>chiamato in tal modo da Bernoulli, 1696 ( $\beta\rho\alpha\chi\acute{\upsilon}\varsigma$  = breve,  $\chi\rho\acute{o}\nu\omicron\varsigma$  = tempo, ossia tempo più breve).

Le equazioni di moto sono allora le seguenti

$$\dot{x}_1 = V(x_2) \cos u, \quad \dot{x}_2 = V(x_2) \sin u \quad (14.202)$$

ove  $u(t)$  è l'angolo che il vettore  $\mathbf{V}$  forma con l'asse  $x_1$ . Il problema è quindi quello della ricerca della funzione  $u(t)$  che minimizza il tempo per andare da A in B. Utilizzando le equazioni (14.202) e l'equazione  $\dot{u}(t) = -V'(x_2) \cos u$ , si vede facilmente che lungo una traiettoria ottimale deve essere  $\dot{u} = \text{costante}$  e quindi che le traiettorie sono delle *cicloidi*, ossia tratti di una curva generata da un punto su una circonferenza che ruota senza strisciare su un direzione orizzontale. ■

◆ **Esercizio 14.1** Determinare mediante il metodo della programmazione dinamica il controllo ottimale del problema definito dal seguente sistema discreto

$$x(t+1) = -0.5x(t) + u(t)$$

e dal funzionale da minimizzare

$$J(u) = \sum_{t=0}^2 |x(t)|$$

La funzione di stato e la funzione di controllo sono vincolati nel seguente modo

$$\begin{aligned} -0.2 \leq x(t) \leq 0.2, & \quad t = 0, 1, 2 \\ -0.1 \leq u(t) \leq 0.1, & \quad t = 0, 1 \end{aligned}$$

Trovare, in particolare, il controllo ottimale corrispondente al valore iniziale  $x(0) = 0.2$ .

◆ **Esercizio 14.2** Il seguente sistema

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -x_1(t) + (1 - x_1^2(t))x_2(t) + u(t) \end{cases}$$

è controllato in modo da minimizzare il seguente funzionale costo

$$J(u) = \frac{1}{2} \int_0^1 [2x_1^2(t) + x_2^2(t) + u^2(t)] dt$$

I valori iniziali e finali dello stato sono specificati. Determinare il sistema di stato aggiunto. Determinare quindi il controllo ottimale nei seguenti casi

- (a)  $u(t)$  non è vincolato.
- (b)  $|u(t)| \leq 1$ .

◆ **Esercizio 14.3** Con riferimento alla Figura 14.31, il livello  $x(t)$  dell'acqua in un bacino, da cui può filtrare acqua attraverso le pareti e nel quale vi è una immissione di acqua dall'esterno, è descritto dal seguente equazione differenziale

$$\dot{x}(t) = -0.5x(t) + u(t)$$

ove  $u(t)$  è la velocità di immissione di acqua dall'esterno al tempo  $t$ . Assumendo che  $0 \leq u(t) \leq M$ , trovare la legge di controllo ottimale per massimizzare il seguente funzionale

$$J_1(u) = \int_0^{50} x(t) dt$$

Ripetere l'analisi precedente con l'aggiunta dell'ipotesi che

$$\int_0^{50} u(t) dt = Q$$

con  $Q$  costante assegnata. Infine, sempre con il vincolo precedente, determinare il controllo ottimale relativo al funzionale costo  $J_2 = x(50)$ .

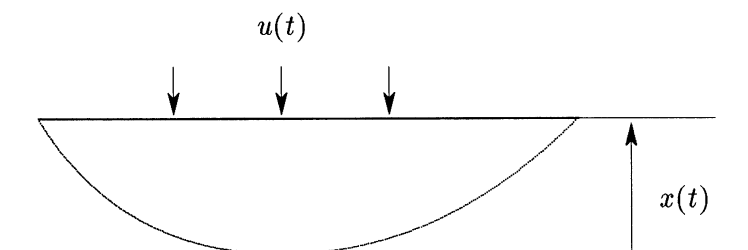


Figura 14.31: Problema di controllo del livello di acqua in un bacino.

◆ **Esercizio 14.4** Trovare le soluzioni ottimali del funzionale

$$J = \int_0^1 (\dot{x}_1 \dot{x}_2 + x_1 x_2) dt$$

con le condizioni  $x_1(0) = 4$ ,  $x_2(0) = 2$ .

◆ **Esercizio 14.5** Minimizzare il funzionale

$$J = \int_0^{\pi/2} (\dot{x}_1^2 + \dot{x}_2^2 + 2x_1 x_2) dt$$

con i vincoli  $x_1(0) = x_2(9) = 1$  e  $x_1(\pi/2) = x_2(\pi/2) = 0$ .

◆ **Esercizio 14.6** Si supponga che il numero delle cellule tumorali  $x(t)$  sia determinato da una legge esponenziale  $\dot{x}(t) = kx(t)$ , con  $k$  costante positiva, e che l'effetto di un farmaco di concentrazione  $u(t)$  sia descritto dal sistema

$$\dot{x}(t) = kx(t) - u(t), \quad x(0) = x_0$$

determinare il controllo  $u(t)$  che minimizza il seguente funzionale

$$J(u) = x(T) + \int_0^T u^2 dt$$

ove  $T$  è fissato.



◆ **Esercizio 14.7** Con riferimento alla Figura 14.32, due pendoli, accoppiati da una molla, sono controllati da due forze uguali e opposte  $u$ , applicate come mostrato in figura. Mostrare che le equazioni di moto sono date da

$$\begin{aligned} m L^2 \ddot{\theta}_1 &= -k a^2 (\theta_1 - \theta_2) - m g L \theta_1 - u \\ m L^2 \ddot{\theta}_2 &= -k a^2 (\theta_2 - \theta_1) - m g L \theta_2 - u \end{aligned}$$

Mostrare inoltre che, posto  $y_1 = \theta_1 + \theta_2$  e  $y_2 = \theta_1 - \theta_2$ , si ha

$$\ddot{y}_1 = -\frac{g}{L} y_1, \quad m L^2 \ddot{y}_2 = -(2k a^2 + m g L) y_2 - 2u$$

da cui si deduce che non è possibile controllare  $\theta_1 + \theta_2$ . Esaminare in quali situazioni il sistema risulta completamente controllabile.

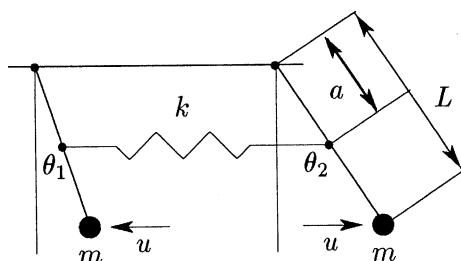


Figura 14.32: Esempio di sistema non completamente controllabile.

◆ **Esercizio 14.8** Analizzare il problema di controllo relativo al minimo del funzionale

$$J(u) = \int_0^T dt$$

per il seguente sistema di stato

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = u(t)$$

con  $x_1(0) = a$ ,  $x_2(0) = b$ , con  $a$  e  $b$  assegnati e  $x_1(T) = x_2(T) = 0$  e  $|u(t)| \leq 1$ .

◆ **Esercizio 14.9** Mostrare che il seguente problema non ammette un controllo ottimale

$$\begin{aligned} \frac{dx}{dt} &= u(t), \quad t \geq 0, \quad 0 \leq u(t) \leq 1 \\ x(0) &= 0, \quad x(T) = 1, \quad T \text{ libero} \\ \min J(u) &= \int_0^T \phi(t) u(t) dt \end{aligned}$$

ove  $\phi(t)$  è una funzione continua assegnata, positiva e decrescente per  $t \geq 0$ . (Suggerimento: osservare che  $\inf J(u) = 0$ , mentre  $J(u) > 0$  per ogni controllo ammissibile). D'altra parte, un controllo ottimale esiste se  $T$  è fissato.

◆ **Esercizio 14.10** Trovare la legge di controllo ottimale per trasferire il sistema

$$\begin{cases} \dot{x}_1(t) = -x_1(t) - u(t) \\ \dot{x}_2(t) = -2x_2(t) - 2u(t) \end{cases}$$

da un punto iniziale arbitrario all'origine nel tempo minimo. L'insieme dei controlli ammissibili  $\mathcal{U}$  è definito dai controlli opportunamente regolari con  $|u(t)| \leq 1$ .

◆ **Esercizio 14.11** Dato il seguente sistema di stato

$$\dot{x} = x + u, \quad x(0) = 5, \quad 0 \leq u \leq 2$$

trovare il controllo che minimizza il funzionale

$$J = \int_0^2 (-2x + 3u + \alpha u^2) dt$$

rispettivamente per  $\alpha = 0$  e  $\alpha = 1$ .

## 14.5 Identificazione di parametri

Nel precedente Capitolo 13 abbiamo analizzato il problema della identificazione dei parametri quando il modello matematico è rappresentato da un sistema differenziale della forma

$$\begin{cases} \frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}, t, \mathbf{a}), & \mathbf{x} \in \mathbb{R}^n, \mathbf{f} = [f_1, f_2, \dots, f_n]^T \\ \mathbf{x}(t_0) = \mathbf{x}_0 \end{cases} \quad (14.203)$$

ove  $\mathbf{x}$  rappresenta il *vettore di stato* e  $\mathbf{a} \in \mathbb{R}^m$  è il *vettore dei parametri*. Per semplicità, il valore iniziale  $\mathbf{x}_0$  è supposto indipendente dal vettore  $\mathbf{a}$ . In questo paragrafo riprenderemo il problema per mostrare come il problema della identificazione del vettore  $\mathbf{a}$  possa essere inquadrato nell'ambito della teoria dei controlli; ricaveremo, in particolare, un metodo numerico per il calcolo del gradiente dello stimatore rispetto ad  $\mathbf{a}$ , alternativo al metodo numerico basato sulle equazioni di sensitività che abbiamo considerato nel Capitolo 13.

In effetti, se poniamo

$$\frac{d\mathbf{a}}{dt} = 0, \quad \mathbf{a}(t_0) = \mathbf{a}$$

si ha un usuale problema di controllo, con funzione di controllo  $\mathbf{a}(t) \equiv \mathbf{a}$ . Nell'ambito del metodo dei minimi quadrati, assumiamo lo stimatore della seguente forma

$$\mathbf{a} \rightarrow J(\mathbf{a}) = \int_{t_0}^T (\mathbf{x}(t, \mathbf{a}) - \mathbf{z})^T \mathbf{R}(t) (\mathbf{x}(t, \mathbf{a}) - \mathbf{z}) dt \quad (14.204)$$

ove  $\mathbf{x}(t, \mathbf{a})$  indica la soluzione di (14.203) corrispondente ad un fissato vettore dei parametri  $\mathbf{a}$ ;  $\mathbf{R}(t)$  è una matrice simmetrica definita positiva (la funzione *peso*) e

$\mathbf{z}(t) \in \mathbb{R}^n$  è il vettore dei dati sperimentali (la funzione *osservata*). Lo stimatore  $J$  definito in (14.204) corrisponde al funzionale obiettivo del problema di controllo, che consiste quindi nel seguente problema di minimo

$$\min_{\mathbf{a} \in \mathcal{U}} J(\mathbf{a}) \quad (14.205)$$

ove  $\mathcal{U}$  è l'insieme dei *parametri ammissibili*. Per il seguito supporremo, per semplicità, che tale insieme sia tutto  $\mathbb{R}^m$ . Esamineremo ora il problema del calcolo del *gradiente* di  $J(\mathbf{a})$  rispetto ad  $\mathbf{a}$ . Ricordiamo che le componenti del vettore gradiente sono utili, sia per avere indicazioni su come lo stimatore dipende dai singoli parametri  $a_i$ , sia anche per applicare gli algoritmi di minimizzazione che utilizzano esplicitamente il gradiente (cfr. Capitolo 5).

Consideriamo dapprima brevemente il metodo delle *equazioni di sensitività*, rinviando per maggiori dettagli al precedente Capitolo 13.

### 14.5.1 Equazioni di sensitività

Dalla definizione di  $J(\mathbf{a})$  si ha

$$\frac{\partial J}{\partial a_i} = 2 \int_{t_0}^T [\mathbf{R}(\mathbf{x}(t, \mathbf{a}) - \mathbf{z}(t))]^T \frac{\partial \mathbf{x}}{\partial a_i} dt, \quad i = 1, \dots, m \quad (14.206)$$

e quindi il problema del calcolo del gradiente di  $J$  è ricondotto a quello del calcolo delle seguenti funzioni

$$\frac{\partial x_j(t, \mathbf{a})}{\partial a_i}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n \quad (14.207)$$

dette le *funzioni di sensitività* del modello (14.203); esse forniscono ad ogni istante  $t$  la dipendenza di ogni componente del vettore di stato  $\mathbf{x}$  dalle singole componenti del vettore dei parametri  $\mathbf{a}$ , e rappresentano quindi un utile strumento di valutazione del modello utilizzato e di pianificazione dei dati sperimentali.

In opportune ipotesi di regolarità sulla funzione  $\mathbf{f}(\mathbf{x}, t, \mathbf{a})$ , si deriva ambo i membri dell'equazione differenziale (14.203) rispetto a  $a_i$  e, scambiando l'ordine di derivazione, si ottiene per ogni  $i = 1, \dots, m$  il seguente sistema di equazioni differenziali nel vettore incognito  $\partial \mathbf{x} / \partial a_i = [\partial x_1 / \partial a_i, \partial x_2 / \partial a_i, \dots, \partial x_n / \partial a_i]^T$

$$\frac{d}{dt} \left( \frac{\partial \mathbf{x}(t, \mathbf{a})}{\partial a_i} \right) = \sum_{j=1}^n \frac{\partial \mathbf{f}}{\partial x_j} \frac{\partial x_j}{\partial a_i} + \frac{\partial \mathbf{f}}{\partial a_i}$$

$$\frac{\partial \mathbf{x}(t_0)}{\partial a_i} = 0$$

Procedendo in questo modo, il calcolo delle funzioni (14.207) richiede la risoluzione di  $m + 1$  problemi a valori iniziali per sistemi di equazioni differenziali di ordine  $n$  (di cui  $m$  sono lineari).

### 14.5.2 Metodo basato sulla teoria dei controlli

Seguendo la procedura con la quale abbiamo ricavato le condizioni di ottimalità per un problema di controllo di tipo generale, consideriamo la variazione del funzionale  $J$  corrispondente ad una variazione ammissibile  $\delta \mathbf{a}$  dei parametri  $\mathbf{a}$ . Si ottiene

$$\delta J = \int_{t_0}^T \delta \mathcal{H}(t) dt, \quad \text{con} \quad \delta \mathcal{H}(t) = (L_{\mathbf{a}} + \mathbf{y}^T \mathbf{f}_{\mathbf{a}}) \delta \mathbf{a} \quad (14.208)$$

ove  $L(t, \mathbf{x}, \mathbf{a})$  è la funzione integranda del funzionale  $J$  e  $\mathbf{f}_{\mathbf{a}}$  è la seguente matrice

$$\mathbf{f}_{\mathbf{a}} = \begin{bmatrix} \partial f_1 / \partial a_1 & \partial f_1 / \partial a_2 & \cdots & \partial f_1 / \partial a_m \\ \partial f_2 / \partial a_1 & \partial f_2 / \partial a_2 & \cdots & \partial f_2 / \partial a_m \\ \cdots & \cdots & \cdots & \cdots \\ \partial f_n / \partial a_1 & \partial f_n / \partial a_2 & \cdots & \partial f_n / \partial a_m \end{bmatrix}$$

Se supponiamo in particolare che la matrice  $\mathbf{R}$  sia indipendente da  $\mathbf{a}$ , si ha  $L_{\mathbf{a}} = 0$ . Il vettore di stato aggiunto  $\mathbf{y}$  è la soluzione del seguente sistema differenziale lineare

$$\begin{cases} \frac{d\mathbf{y}}{dt} = -\mathbf{f}_{\mathbf{x}}^T \mathbf{y} - L_{\mathbf{x}}^T \\ \mathbf{y}(T) = 0 \end{cases} \quad (14.209)$$

ove  $L_{\mathbf{x}}^T = 2\mathbf{R}(\mathbf{x}(t, \mathbf{a}) - \mathbf{z})$ . Dal risultato (14.208) si ha

$$J_{\mathbf{a}} = \int_{t_0}^T \mathbf{f}_{\mathbf{a}}^T \mathbf{y} dt \iff \frac{\partial J}{\partial a_i} = \int_{t_0}^T \left[ \frac{\partial f_1}{\partial a_i} y_1 + \frac{\partial f_2}{\partial a_i} y_2 + \cdots + \frac{\partial f_n}{\partial a_i} y_n \right] dt \quad (14.210)$$

In conclusione, il metodo ora illustrato permette di calcolare il gradiente di  $J$  rispetto al vettore  $\mathbf{a}$  mediante l'integrazione del sistema differenziale (14.208) (*sistema di stato*) e del sistema (14.209) (*sistema di stato aggiunto*), indipendentemente dal numero  $m$  dei parametri da stimare. Rispetto al metodo basato sulle equazioni di sensitività, si ha quindi, in generale, una notevole riduzione nel numero di equazioni differenziali da risolvere. Sottolineiamo, comunque, che la coppia di problemi (14.208), (14.209) costituiscono un *problema ai limiti*, in quanto le funzioni  $\mathbf{x}(t)$  e  $\mathbf{y}(t)$  sono fissate nei due punti distinti  $t_0$  e  $T$ .

► **Esempio 14.24** Come illustrazione, consideriamo il seguente sistema di stato

$$\begin{cases} \frac{dx_1}{dt} = a_1 x_1 + a_2 x_1^2 + a_3 x_2 + g_1(t), & x_1(0) = 1 \\ \frac{dx_2}{dt} = a_4 x_2^2 + a_5 x_1 + g_2(t), & x_2(0) = 0 \end{cases} \quad (14.211)$$

ove  $g_1, g_2$  sono funzioni assegnate. In questo caso si ha  $m = 5$  e  $n = 2$ . Supponendo che  $x_2(t)$  sia la variabile osservata, ossia, più precisamente, che si conoscano i valori sperimentali

di  $z_j = x_2(t_j)$  negli istanti successivi  $t_1, t_2, \dots, t_p$  dell'intervallo di tempo  $[0, T]$ , il funzionale costo  $J$  assume (supponendo, ad esempio  $\mathbf{R} = \mathbf{I}$ ) la seguente forma

$$J(\mathbf{a}) = \sum_{j=1}^p (x_2(t_j) - z_j)^2 \quad (14.212)$$

che può essere considerata come un caso particolare della definizione (14.204), quando la funzione integranda è di tipo impulsivo (cfr. Appendice B). Si ha allora

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} a_1 + 2a_2x_1 & a_3 \\ & a_5 \\ & 2a_4x_2 \end{bmatrix}, \quad \frac{\partial \mathbf{f}}{\partial \mathbf{a}} = \begin{bmatrix} x_1 & x_1^2 & x_2 & 0 & 0 \\ 0 & 0 & 0 & x_2^2 & x_1 \end{bmatrix}$$

e il sistema aggiunto assume la seguente forma

$$\begin{cases} \frac{dy_1}{dt} = -(a_1 + 2a_2x_1)y_1 + a_5y_2 & y_1(T) = 0 \\ \frac{dy_2}{dt} = -a_3y_1 + 2a_4x_2y_2 + 2 \sum_{j=1}^p \delta_{t-t_j} (x_2(t_j) - z_j) & y_2(T) = 0 \end{cases} \quad (14.213)$$

ove  $\delta_{t-t_j}$  è la funzione di Dirac relativa al punto  $t = t_j$ . Infine, le componenti del gradiente  $J_{\mathbf{a}}$  sono date dai seguenti integrali

$$\begin{aligned} \frac{\partial J}{\partial a_1} &= \int_0^T x_1 y_1 dt, & \frac{\partial J}{\partial a_2} &= \int_0^T x_1^2 y_1 dt, & \frac{\partial J}{\partial a_3} &= \int_0^T x_2 y_1 dt, \\ \frac{\partial J}{\partial a_4} &= \int_0^T x_2^2 y_2 dt, & \frac{\partial J}{\partial a_5} &= \int_0^T x_1 y_2 dt \end{aligned}$$

Ricordiamo che l'utilizzo delle equazioni di sensitività richiederebbe la risoluzione di un problema a valori iniziali per 12 equazioni differenziali. ■

L'interesse della procedura basata sulla teoria dei controlli è ulteriormente sottolineato dal successivo esempio relativo a sistemi di stato basati su equazioni alle derivate parziali.

► **Esempio 14.25** (*Esempi di problemi inversi: identificazione della trasmissività termica di un corpo e problema dell'elettrocardiografia*) Con riferimento alla Figura 14.33, si studia la diffusione del calore in un corpo rappresentato dall'insieme  $\Omega \subset \mathbb{R}^3$ , con  $\Omega$  un dominio limitato di frontiera  $\Gamma$  sufficientemente regolare. Con  $u(\mathbf{x}, t)$ ,  $\mathbf{x} \in \Omega$ ,  $t \in [0, T]$ , con  $T > 0$  fissato, si indica la *temperatura* del corpo nel punto  $\mathbf{x}$  all'istante  $t$ . Supponendo il corpo *isotropo*, ma non necessariamente omogeneo, la variazione della temperatura  $u(\mathbf{x}, t)$  è descritta dalla seguente equazione alle derivate parziali (cfr. Capitolo 7)

$$\frac{\partial u(\mathbf{x}, t)}{\partial t} - \sum_{i=1}^3 \frac{\partial}{\partial x_i} \left( a(\mathbf{x}) \frac{\partial u(\mathbf{x}, t)}{\partial x_i} \right) = f(\mathbf{x}, t), \quad \forall \mathbf{x} \in \Omega, \quad t \in (0, T) \quad (14.214)$$

$$\frac{\partial u(\mathbf{x}, t)}{\partial \nu} = 0, \quad \mathbf{x} \in \Gamma, \quad t \in (0, T) \quad (14.215)$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \quad (14.216)$$

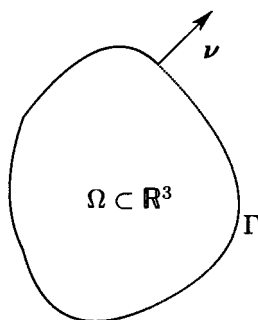


Figura 14.33: Problema della diffusione del calore in un corpo rappresentato dall'insieme  $\Omega$  nello spazio a tre dimensioni.

ove  $a(\mathbf{x})$  rappresenta la trasmissività termica del corpo nel punto  $\mathbf{x}$  ed è supposta indipendente da  $t$ . La condizione (14.215), nella quale  $\nu$  rappresenta il versore della normale esterna a  $\Gamma$ , corrisponde all'ipotesi che il flusso di calore attraverso  $\Gamma$  sia nullo (corpo termicamente isolato); la funzione  $f(\mathbf{x}, t)$  rappresenta l'intensità di calore immesso o sottratto dal corpo, e la funzione  $u_0(\mathbf{x})$  indica la ripartizione iniziale della temperatura.

Le funzioni  $f(\mathbf{x}, t)$  e  $u_0(x)$  sono supposte note, mentre  $a(\mathbf{x})$  è una funzione incognita. Il problema è precisamente quello di *identificare*  $a(\mathbf{x})$  attraverso i valori osservati della temperatura  $u(\mathbf{x}, t)$  sulla frontiera  $\Gamma$  per  $t \in (0, T)$ . Il problema può essere formulato in forma di *problema di controllo*, introducendo il seguente *funzionale costo*

$$J(a) = \int_0^T \int_{\Gamma} [u(\mathbf{x}, t, a) - z(\mathbf{x}, t)]^2 d\Gamma dt \quad (14.217)$$

ove  $z(\mathbf{x}, t)$  rappresenta la *funzione osservata* e  $u(\mathbf{x}, t, a)$  è la soluzione del problema (14.214)-(14.216) corrispondente alla funzione  $a(\mathbf{x})$ . Fissato quindi un insieme di controlli ammissibili  $\mathcal{U}$ , ossia un opportuno spazio funzionale, si cerca  $a(\mathbf{x}) \in \mathcal{U}$  che minimizza il funzionale  $J(a)$  definito in (14.217).

Situazioni analoghe a quella ora descritta si riscontrano assai frequentemente nelle applicazioni. Ci limiteremo a segnalare la identificazione delle proprietà reologiche dei materiali a partire dall'osservazione della propagazione delle onde sismiche o dalle onde opportunamente prodotte allo scopo di individuare la presenza di pozzi petroliferi. In campo medico segnaliamo come esempio importante quello dell'*elettrocardiografia*. In maniera schematica (cfr. per una trattazione più dettagliata Colli Franzone [30] e la bibliografia ivi contenuta), lo scopo principale dell'elettrocardiografia è quello di ottenere informazioni sullo stato elettrico e di conseguenza fisiologico del cuore a partire dalle misure del potenziale elettrico eseguite sulla superficie del corpo umano durante i vari battiti cardiaci. Indichiamo con  $\Omega_1$  il dominio di  $\mathbb{R}^3$  che rappresenta il corpo umano e con  $\Gamma_1$  la sua frontiera; sia inoltre  $\Omega_0$  un dominio di  $\mathbb{R}^3$  che contiene il cuore ed avente frontiera  $\Gamma_0$  (cfr. per una rappresentazione schematica di una sezione bidimensionale la Figura 14.34). La  $\Gamma_0$  è assunta come una approssimazione fissa della superficie epicardica, che invece varia nel tempo. Posto  $\Omega := \Omega_1 - \Omega_0$ , si studia quindi la propagazione del segnale elettrico in  $\Omega$ . Il corpo umano, ad eccezione del cuore, può essere, con buona approssimazione, considerato un mezzo conduttore isotropo, resistivo e lineare; inoltre, benché le sorgenti bioelettriche nel miocardio siano variabili nel tempo, ad ogni istante del battito cardiaco si può assumere che il cuore generi nel corpo umano un

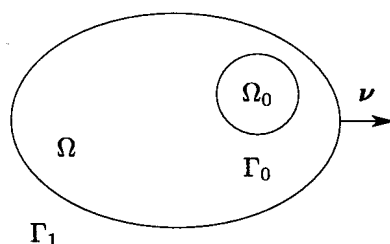


Figura 14.34: Rappresentazione schematica di una sezione bidimensionale del torace;  $\Gamma_0$  rappresenta la superficie epicardica.

sistema di correnti *stazionarie*. Il campo elettrico  $\mathbf{E}$  e la densità di corrente  $\mathbf{j}$  soddisfano in  $\Omega$  le seguenti equazioni

$$\mathbf{j} = \sigma \mathbf{E}, \quad \text{rot } \mathbf{E} = 0, \quad \text{div } \mathbf{j} = 0 \quad (14.218)$$

ove con  $\sigma$  si è indicata la conducibilità elettrica. Essendo  $\mathbf{E}$  irrotazionale, si ha che esso deriva da un potenziale  $V(\mathbf{x}, t)$ , ossia si ha  $\mathbf{E} = -\text{grad } V(\mathbf{x})$ . Inoltre, essendo  $\Omega_1$  immerso nell'aria, che si comporta come un mezzo isolante, su  $\Gamma_1$  la componente di  $\mathbf{E}$  è nulla, e quindi  $\partial V / \partial \nu = 0$ , ove  $\nu$  indica la normale esterna a  $\Gamma_1$ .

Nell'ipotesi che il potenziale  $V(\mathbf{x})$  sia noto sulla superficie epicardica  $\Gamma_0$ , allora la funzione  $V(\mathbf{x})$  è la soluzione del seguente *problema ai limiti* di tipo ellittico

$$\begin{cases} \text{div } \sigma(\mathbf{x}) \text{ grad } V(\mathbf{x}) = 0 & \mathbf{x} \in \Omega \\ V(\mathbf{x}) = v(\mathbf{x}) & \mathbf{x} \in \Gamma_0, \quad \frac{\partial V(\mathbf{x})}{\partial \nu} = 0 & \mathbf{x} \in \Gamma_1 \end{cases} \quad (14.219)$$

ove  $v(\mathbf{x})$  indica il valore del potenziale su  $\Gamma_0$ . Il problema (14.219) rappresenta il cosiddetto *problema diretto* dell'elettrocardiografia; esso consiste nel simulare le elettromappe toraciche a partire dalla conoscenza degli eventi elettrici intracardiaci, della geometria e conducibilità del torace umano.

Il *problema inverso* consiste, invece, nell'identificare la distribuzione di potenziale sulla superficie epicardica dalla conoscenza delle elettromappe di superficie. Dal punto di vista matematico, se non è noto il potenziale su  $\Gamma_0$ , ma è possibile misurare il potenziale su una porzione  $\Sigma \subset \Gamma_1$  della superficie toracica, allora  $V(\mathbf{x})$  soddisfa al seguente *problema a valori iniziali* (problema di Cauchy)

$$\begin{cases} \text{div } \sigma(\mathbf{x}) \text{ grad } V(\mathbf{x}) = 0 & \mathbf{x} \in \Omega \\ V(\mathbf{x}) = z(\mathbf{x}) & \mathbf{x} \in \Sigma, \quad \frac{\partial V(\mathbf{x})}{\partial \nu} = 0 & \mathbf{x} \in \Gamma_1 \end{cases} \quad (14.220)$$

ove  $z(\mathbf{x})$  indica la distribuzione del potenziale sulla superficie toracica  $\Sigma$ , ossia in pratica il potenziale rilevato dalla apparecchiatura<sup>8</sup>. Il problema inverso consiste quindi nello stimare  $V(\mathbf{x})$  su  $\Gamma_0$ .

<sup>8</sup>Si suppone che l'informazione elettrocardiografica sia raccolta in *numerosi* punti della superficie toracica; in effetti, esistono apparecchiature che permettono ad esempio il rilevamento dalla superficie toracica di 240 segnali elettrocardiografici in 2 millisecondi.

Il problema di Cauchy (14.220) definisce  $V(\mathbf{x})$  in maniera univoca, ma è notoriamente un *problema mal posto*, nel senso che la soluzione  $V(\mathbf{x})$  non dipende con continuità dal valore iniziale  $z(\mathbf{x})$ . In altre parole, piccole perturbazioni su  $z$  possono produrre perturbazioni amplificate in modo incontrollato sulla soluzione (cfr. Capitolo 7). Tenendo presente che il dato osservato  $z$  è inevitabilmente affetto da errori, si capisce l'impossibilità pratica di utilizzare il modello matematico nella forma (14.220) per la risoluzione del problema inverso. Il problema può essere, invece, convenientemente riformulato come *problema di controllo*.

Indichiamo con  $\mathcal{U}$  uno spazio di funzioni definite su  $\Gamma_0$  che contenga le distribuzioni di potenziale sull'epicardio ammissibili; per  $u \in \mathcal{U}$  si indica con  $w(\mathbf{x}, u)$ , o semplicemente  $w(u)$ , la soluzione del seguente problema diretto

$$\begin{cases} \operatorname{div} \sigma(\mathbf{x}) \operatorname{grad} w(u) = 0 & \mathbf{x} \in \Omega \\ w(u) = u & \mathbf{x} \in \Gamma_0, \quad \frac{\partial w(u)}{\partial \nu} = 0 & \mathbf{x} \in \Gamma_1 \end{cases} \quad (14.221)$$

Introduciamo quindi come *operatore di osservazione* l'operatore che, data un'elettromappa cardiaca, associa la corrispondente elettromappa toracica ad essa compatibile, ossia l'operatore  $\mathcal{A}u = w(u)|_{\Sigma}$  dei valori di  $w$  su  $\Sigma$ . Il problema inverso (14.221) può essere allora riformulato nel seguente modo

$$\text{Trovare } u \in \mathcal{U} \text{ tale che } J(u) = \min_{v \in \mathcal{U}} J(v) \quad (14.222)$$

ove il funzionale

$$J(v) := \int_{\Sigma} |\mathcal{A}v - z|^2 d\sigma \quad (14.223)$$

misura la distanza tra l'osservata  $z$  e la predizione  $w(v)|_{\Sigma} = \mathcal{A}v$ . Procedendo in questo modo, si è ottenuto un problema di controllo definito dal *sistema di stato* (14.221), dalla *funzione controllo*  $u \in \mathcal{U}$  e dal *funzionale costo*<sup>9</sup>  $J(v)$ .

Ritornando ora al problema precedente della identificazione della trasmissività termica di un corpo, diamo un'idea di come calcolare il *gradiente*, rispetto alla funzione  $a(\mathbf{x})$ , del funzionale  $J(a)$  definito in (14.217). La procedura che indichiamo è del tutto formale, ma può essere giustificata introducendo opportuni spazi funzionali per la soluzione  $u(\mathbf{x}, t)$  e il controllo  $a(\mathbf{x})$ .

Si definisce come *funzione stato aggiunto*  $y(\mathbf{x}, t)$  la soluzione del seguente problema

$$\begin{cases} -\frac{\partial y}{\partial t} - \sum_{i=1}^3 \frac{\partial}{\partial x_i} \left( a(x) \frac{\partial y}{\partial x_i} \right) = 0, & \forall \mathbf{x} \in \Omega, \quad t \in (0, T) \\ a(\mathbf{x}) \frac{\partial p}{\partial \nu}(\mathbf{x}, t) = -2[u(\mathbf{x}, t; a) - z(\mathbf{x}, t)], & \mathbf{x} \in \Gamma, \quad t \in (0, T) \\ y(\mathbf{x}, T) = 0, & \mathbf{x} \in \Omega \end{cases} \quad (14.224)$$

che è un problema del tipo del calore con segno del tempo invertito; esso risulta ben posto quando risolto, come è appunto richiesto in questo caso, procedendo all'indietro a partire da

<sup>9</sup>In realtà, per motivi di stabilità, è opportuno *regolarizzare* il funzionale  $J(v)$  definito in (14.223) mediante l'aggiunta di un termine del tipo  $\epsilon \int_{\Gamma_0} |\mathcal{B}|^2 d\sigma$ , ove  $\mathcal{B}$  è un opportuno operatore e  $\epsilon$  è un parametro di regolarizzazione da scegliere in maniera conveniente. Per questi aspetti, importanti per la risoluzione numerica del problema, rinviamo a Colli Franzone [30].



$T$ . Ripetendo la procedura che abbiamo descritto in precedenza nel caso di problemi relativi ad equazioni differenziali ordinarie, si può mostrare, con opportune integrazioni per parti, che la variazione  $\delta J = J(a + \delta a) - J(a)$  corrispondente alla variazione  $\delta a$  della funzione controllo  $a(\mathbf{x})$  può essere espressa nella seguente forma

$$\delta J = \int_{\Omega} \left[ \int_0^T \sum_{i=1}^3 \frac{\partial u}{\partial x_i} \frac{\partial y}{\partial x_i} dt \right] \delta a(\mathbf{x}) d\mathbf{x} \quad (14.225)$$

Da tale espressione si può ricavare il gradiente funzionale di  $J$  rispetto alla funzione  $a(\mathbf{x})$ . Formalmente, si ha

$$\frac{\partial J}{\partial a}(\mathbf{x}, a) = \int_0^T \sum_{i=1}^3 \frac{\partial u(\mathbf{x}, t, a)}{\partial x_i} \frac{\partial y(\mathbf{x}, t, a)}{\partial x_i} dt \quad (14.226)$$

Tale risultato può essere utilizzato in un metodo iterativo di minimizzazione del funzionale  $J$ . Come esemplificazione, il metodo del gradiente porta a una iterazione del seguente tipo

$$a(\mathbf{x})^{(k+1)} = a(\mathbf{x})^{(k)} - \lambda_k \frac{\partial J}{\partial a}(\mathbf{x}, a^{(k)}) \quad (14.227)$$

ove  $a(\mathbf{x})^{(0)}$  è una funzione di tentativo e  $\lambda_k$  è l'usuale parametro che si utilizza nei metodi del gradiente (cfr. Capitolo 5). ■

Linear algebra allows and even encourages a very satisfying combination of both elements of mathematics – abstraction and application.  
**G. Strang**

## Appendice A

# Elementi di algebra lineare

Questa appendice contiene in forma essenziale e concreta i risultati e le idee dell'algebra lineare che sono di particolare interesse nella costruzione e nell'analisi degli algoritmi numerici. Per una introduzione più approfondita si veda ad esempio Lancaster e Tismenetsky [105], Golub e Van Loan [69], Strang [147].

### A.1 Matrici. Definizioni fondamentali

Una *matrice*<sup>1</sup>  $\mathbf{A}$  è un insieme di  $m \times n$  numeri reali (o complessi) ordinati secondo lo schema

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Si può abbreviare, scrivendo  $\mathbf{A} = (a_{ij})$  (oppure  $\mathbf{A} = [a_{ij}]$ ), per  $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, n$ . La notazione  $\mathbf{A} \in \mathbb{R}^{m \times n}$  (rispettivamente  $\mathbf{A} \in \mathbb{C}^{m \times n}$ ) significa che la matrice ha  $m$  righe e  $n$  colonne di elementi reali (rispettivamente complessi). Se  $m = n$ , allora si dice che  $\mathbf{A}$  è una matrice quadrata di ordine  $n$ . Nel seguito faremo riferimento usualmente a matrici con elementi reali, indicando, quando si ritiene di interesse, l'estensione dei risultati al caso delle matrici ad elementi complessi.

---

<sup>1</sup>Il termine *matrice* (matrix) pare sia stato utilizzato per la prima volta da Sylvester (1850). Lo sviluppo della teoria delle matrici è legato ai nomi di Hamilton (1853), Cayley (1854) e successivamente Laguerre, Hermite e Frobenius.

Un *vettore colonna*  $\mathbf{x} = [x_i]$ ,  $i = 1, 2, \dots, n$  è una matrice che consiste di una sola colonna. Si ha quindi

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

L'insieme di tutti i vettori colonna a  $n$  componenti reali (rispettivamente complessi) costituisce lo *spazio euclideo*  $\mathbb{R}^n$  (rispettivamente lo spazio  $\mathbb{C}^n$ ).

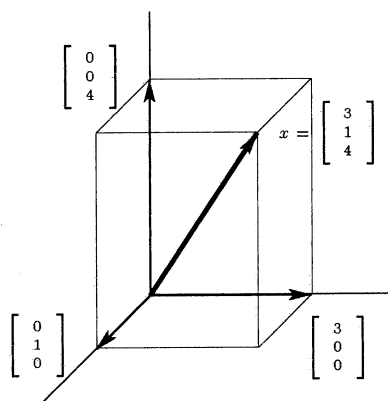


Figura A.1: Un vettore in  $\mathbb{R}^3$ .

Un vettore  $\mathbf{x} \in \mathbb{R}^n$  può essere identificato con un punto dello spazio a  $n$ -dimensioni, assumendo le componenti  $x_i$  come coordinate del punto. Come illustrazione, in Figura A.1 è rappresentato geometricamente il seguente vettore in  $\mathbb{R}^3$

$$\mathbf{x} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}$$

Data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , è talvolta utile vedere le colonne della matrice come vettori colonna. In tale contesto si può scrivere  $\mathbf{A}$  nella seguente forma

$$\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n]$$

ove  $\mathbf{a}_j \in \mathbb{R}^m$  rappresenta la colonna  $j$ -ma.

### A.1.1 Matrici particolari

Una matrice *diagonale* è una matrice del tipo

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} = \mathbf{diag}(d_1, d_2, \dots, d_n)$$

Se  $d_1 = d_2 = \cdots = d_n = a$ , la matrice è detta matrice *scalare*. In particolare, la matrice *unità*, o matrice *identità*,  $\mathbf{I}_n$ , o semplicemente  $\mathbf{I}$ , è definita da  $\mathbf{I}_n = \mathbf{diag}(1, 1, \dots, 1)$ ; cioè  $\mathbf{I} = (\delta_{ij})$ , dove  $\delta_{ij}$  è il simbolo di Kronecker:  $\delta_{ij} = 0$ , per  $i \neq j$ ,  $\delta_{ij} = 1$  per  $i = j$ .

Una matrice di *permutazione*  $\mathbf{P}$  è una matrice che si ottiene dalla matrice identità  $\mathbf{I}$  con scambi di righe (o di colonne); pertanto, in ogni riga e ogni colonna  $\mathbf{P}$  ha un solo elemento diverso dallo zero e uguale a 1.

Una matrice *triangolare* è una matrice della forma

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \quad \text{oppure} \quad \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

$\mathbf{L}$  è detta *triangolare inferiore* e  $\mathbf{U}$  *triangolare superiore*.

Una matrice di *Hessenberg* (superiore) è una matrice quasi triangolare, più precisamente della forma

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & h_{nn-1} & h_{nn} \end{bmatrix}$$

In maniera analoga, si definisce una matrice di Hessenberg inferiore.

### A.1.2 Operazioni su matrici

Due matrici  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$  sono dette *uguali*, e si scrive  $\mathbf{A} = \mathbf{B}$ , se  $a_{ij} = b_{ij}$  per  $i = 1, 2, \dots, m, j = 1, 2, \dots, n$ .

Il prodotto  $\alpha \mathbf{A}$  di una matrice  $\mathbf{A}$  per uno scalare  $\alpha$ , ossia un elemento di  $\mathbb{R}$  o di  $\mathbb{C}$ , è la matrice di elementi  $(\alpha a_{ij})$ .

La *somma* di due matrici  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$  è la matrice  $\mathbf{C} \in \mathbb{R}^{m \times n}$  di elementi  $c_{ij} = a_{ij} + b_{ij}$ , per  $i = 1, 2, \dots, m$  e  $j = 1, 2, \dots, n$ . L'elemento neutro rispetto alla somma

in  $\mathbb{R}^{m \times n}$  è dato dalla matrice i cui elementi sono tutti uguali a zero. Per il seguito, in assenza di possibilità di equivoci, tale matrice verrà indicata semplicemente con lo zero.

Il *prodotto* (righe per colonne) di due matrici  $\mathbf{A} \in \mathbb{R}^{m \times p}$  e  $\mathbf{B} \in \mathbb{R}^{p \times n}$  è la matrice  $\mathbf{C} \in \mathbb{R}^{m \times n}$  con elementi

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj} \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n$$

Sottolineiamo che il prodotto righe per colonne è definito solo se il numero delle colonne di  $\mathbf{A}$  è uguale al numero delle righe di  $\mathbf{B}$ . Si verifica facilmente che per ogni matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  si ha

$$\mathbf{I}_m \mathbf{A} = \mathbf{A}, \quad \mathbf{A} \mathbf{I}_n = \mathbf{A}$$

La moltiplicazione fra matrici gode della proprietà associativa e di quella distributiva rispetto all'addizione

$$\mathbf{A}(\mathbf{B}\mathbf{C}) = (\mathbf{A}\mathbf{B})\mathbf{C}, \quad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{A}\mathbf{B} + \mathbf{A}\mathbf{C}, \quad (\mathbf{B} + \mathbf{C})\mathbf{D} = \mathbf{B}\mathbf{D} + \mathbf{C}\mathbf{D}$$

mentre, in generale, si ha  $\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$ , e quindi la moltiplicazione non è *commutativa*. In effetti, la possibilità di definire la matrice  $\mathbf{A}\mathbf{B}$  non implica necessariamente quella della matrice  $\mathbf{B}\mathbf{A}$ , e viceversa. Inoltre, anche quando i prodotti sono ambedue definiti, essi possono differire tra loro, come mostra il seguente semplice esempio.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \Rightarrow \mathbf{A}\mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}\mathbf{A} = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix}$$

È interessante osservare che le somme e i prodotti di matrici triangolari, ad esempio triangolari superiori, sono ancora matrici triangolari dello stesso tipo, ossia le matrici triangolari costituiscono una famiglia chiusa rispetto all'operazione di somma e di prodotto.

▼ **Osservazione A.1** Se  $\mathbf{A} \in \mathbb{R}^{m \times n}$  e  $\mathbf{x} \in \mathbb{R}^n$ , il prodotto

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

definisce una trasformazione da  $\mathbb{R}^n$  in  $\mathbb{R}^m$ . Dal momento che per ogni coppia di vettori  $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$  e ogni coppia di scalari  $\alpha, \beta$ , si ha  $\mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{w}) = \alpha\mathbf{A}\mathbf{x} + \beta\mathbf{A}\mathbf{w}$ , la trasformazione  $\mathbf{x} \rightarrow \mathbf{A}\mathbf{x}$  è lineare. Viceversa, si può mostrare che ogni trasformazione lineare da  $\mathbb{R}^n$  in  $\mathbb{R}^m$  può essere rappresentata da una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ . Il prodotto di due matrici corrisponde allora alla matrice associata alla trasformazione che risulta dalla composizione di due trasformazioni lineari.

Osserviamo, ancora, che dal punto di vista concettuale e computazionale, ha interesse la seguente interpretazione del prodotto di una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  per un vettore  $\mathbf{x} \in \mathbb{R}^n$

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \sum_{i=1}^n x_i \mathbf{a}_i$$

Il vettore  $\mathbf{y} \in \mathbb{R}^m$  risulta una combinazione lineare delle colonne  $\mathbf{a}_i$  della matrice  $\mathbf{A}$ . Ad esempio, se

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & -2 \\ 1 & -1 & 6 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

si ha

$$\mathbf{A}\mathbf{x} = 1 \begin{bmatrix} 3 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} + 3 \begin{bmatrix} -2 \\ 6 \end{bmatrix} = \begin{bmatrix} -1 \\ 17 \end{bmatrix}$$

■

Si definisce matrice *trasposta*  $\mathbf{A}^T$  di una matrice  $\mathbf{A}$  la matrice che si ottiene dalla matrice  $\mathbf{A}$  per scambio delle righe con le colonne. Gli elementi  $a_{ij}^T$  della matrice  $\mathbf{A}^T$  sono, pertanto, dati dagli elementi  $a_{ji}$ , per  $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, n$ . Se la matrice  $\mathbf{A}$  è ad elementi complessi, si definisce la matrice *trasposta coniugata*  $\mathbf{A}^*$  (denotata anche  $\mathbf{A}^H$ , ponendo  $a_{ij}^* = \bar{a}_{ji}$ , ove  $\bar{a}_{ji}$  è il coniugato del numero complesso  $a_{ji}$ ). Ad esempio, si ha

$$\mathbf{A} = \begin{bmatrix} 2-i & 3 & 2+2i \\ i & 4i & 1 \end{bmatrix}, \quad \mathbf{A}^* = \begin{bmatrix} 2+i & -i \\ 3 & -4i \\ 2-2i & 1 \end{bmatrix}$$

Una matrice  $\mathbf{A} \in \mathbb{C}^{n \times n}$  viene detta *hermitiana* quando  $\mathbf{A}^* = \mathbf{A}$ ; nel caso di una matrice ad elementi reali la proprietà si riduce a  $\mathbf{A}^T = \mathbf{A}$  e la matrice viene detta *simmetrica*. Ricordiamo, inoltre, che una matrice  $\mathbf{A} \in \mathbb{C}^{n \times n}$  viene detta *normale* quando  $\mathbf{A}^* \mathbf{A} = \mathbf{A} \mathbf{A}^*$ . Una matrice hermitiana è ovviamente un caso particolare di matrice normale. Nel seguito, vedremo altri esempi di matrici normali, insieme a proprietà interessanti delle matrici hermitiane, o in particolare simmetriche, e delle matrici normali.

Per la trasposta di un prodotto, si ha la seguente proprietà

$$(\mathbf{A}\mathbf{B})^T = \mathbf{B}^T \mathbf{A}^T$$

con un analogo risultato per la trasposta coniugata.

Se  $\mathbf{x}$  è un vettore *colonna*,  $\mathbf{x}^T$  è allora un vettore *riga*. Se  $\mathbf{x}$  e  $\mathbf{y}$  sono due vettori *colonna* con  $n$  elementi reali, il prodotto di matrice

$$\mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i$$

è detto *prodotto scalare* (o *dot*, *interno*) tra i due vettori, e generalizza la moltiplicazione tra scalari. Per indicare il prodotto scalare si usa anche la notazione  $(\mathbf{x}, \mathbf{y})$ . Nel caso di vettori  $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ , si ha

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^* \mathbf{y} = \sum_{i=1}^n \bar{x}_i y_i$$

Se  $\mathbf{x} \in \mathbb{R}^m$  e  $\mathbf{y} \in \mathbb{R}^n$ , si chiama *prodotto esterno* tra i due vettori la matrice  $\mathbf{xy}^T \in \mathbb{R}^{m \times n}$  definita nel modo seguente

$$\mathbf{xy}^T = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

Osserviamo che ogni colonna di  $\mathbf{xy}^T$  è un multiplo di  $\mathbf{x}$ , e ogni riga è un multiplo di  $\mathbf{y}^T$ . In maniera analoga si definisce per  $\mathbf{x} \in \mathbb{C}^m$  e  $\mathbf{y} \in \mathbb{C}^n$  il prodotto esterno  $\mathbf{xy}^*$ . Come illustrazione, dati i vettori

$$\mathbf{x} = [1, i, -i]^T, \quad \mathbf{y} = [i, 1, i]^T, \quad i = \sqrt{-1}$$

si ha

$$\mathbf{x}^*\mathbf{y} = -1, \quad \mathbf{xy}^* = \begin{bmatrix} -i & 1 & -i \\ 1 & i & 1 \\ -1 & -i & -1 \end{bmatrix}$$

Il prodotto interno fra vettori gode delle seguenti proprietà

1.  $\mathbf{x}^*\mathbf{x} \in \mathbb{R}$ , e  $\geq 0$ ,  $= 0$  se e solo se  $\mathbf{x} = 0$ ;
2.  $\overline{\mathbf{x}^*\mathbf{y}} = \mathbf{y}^*\mathbf{x}$ ;
3.  $\mathbf{x}^*(\alpha\mathbf{y}) = \alpha\mathbf{x}^*\mathbf{y}$  per  $\alpha \in \mathbb{C}$ ;
4.  $\mathbf{x}^*(\mathbf{y} + \mathbf{z}) = \mathbf{x}^*\mathbf{y} + \mathbf{x}^*\mathbf{z}$  per  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{C}^n$ .

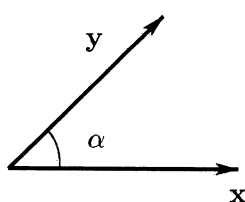


Figura A.2: Angolo fra due vettori.

Se  $\mathbf{x} \in \mathbb{R}^n$ , la quantità  $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$  è detta *lunghezza euclidea* del vettore  $\mathbf{x}$ ; un vettore la cui lunghezza euclidea è uguale a 1 viene detto *normalizzato*. Il seguente risultato fornisce una interessante interpretazione del prodotto scalare. Se  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , si ha

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos \alpha$$

ove  $\alpha$  denota l'angolo tra i due vettori (cfr. Figura A.2 per una rappresentazione in  $\mathbb{R}^2$ ).

Se  $\mathbf{A}$  è una matrice ad  $m$  righe  $\mathbf{a}_i^T$ , ( $1 \leq i \leq m$ ) di lunghezza  $p$  e  $\mathbf{B}$  è una matrice contenente  $n$  colonne  $\mathbf{b}_j$  della stessa lunghezza  $p$ , allora il prodotto  $\mathbf{AB}$  può essere anche definito come la matrice  $\mathbf{C}$  ad elementi  $c_{ij} = \mathbf{a}_i^T \mathbf{b}_j$ , per ( $1 \leq i \leq m$ ,  $1 \leq j \leq n$ ).

**Matrici definite positive** Una matrice simmetrica (hermitiana)  $\mathbf{A}$  si dice *definita positiva* quando si ha  $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$  per tutti i vettori reali  $\mathbf{x} \neq 0$  ( $\mathbf{x}^* \mathbf{A} \mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{C}^n \neq 0$ ).

Ad esempio, la matrice hermitiana

$$\mathbf{A} = \begin{bmatrix} 3 & i \\ -i & 3 \end{bmatrix}$$

è definita positiva. Infatti, per ogni  $\mathbf{x} = [x_1, x_2]^T \neq 0$  risulta

$$\begin{aligned} \mathbf{x}^* \mathbf{A} \mathbf{x} &= [\bar{x}_1, \bar{x}_2] \begin{bmatrix} 3 & i \\ -i & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3\bar{x}_1 x_1 - i x_1 \bar{x}_2 + i \bar{x}_1 x_2 + 3\bar{x}_2 x_2 \\ &= (x_1 - i x_2)(\bar{x}_1 + i \bar{x}_2) + 2(x_2 - i x_1)(\bar{x}_2 + i \bar{x}_1) = |x_1 - i x_2|^2 + 2|x_2 - i x_1|^2 > 0 \end{aligned}$$

Chiamando *sottomatrice principale* di ordine  $i$  di una matrice  $\mathbf{A}$  la matrice che si ottiene eliminando  $n - i$  righe e le corrispondenti  $n - i$  colonne, si ha il seguente risultato importante.

**Proposizione A.1** *Se una matrice  $\mathbf{A} \in \mathbb{C}^{n \times n}$  è definita positiva, anche tutte le sue sottomatrici principali sono definite positive.*

Da tale risultato si ha, in particolare, che gli elementi  $a_{ii}$  di una matrice  $\mathbf{A}$  definita positiva sono reali e positivi. Nel seguito vedremo altre proprietà delle matrici hermitiane definite positive in relazione alle nozioni di determinante e di autovalore.

**Ortogonalità** Quando per due vettori  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  si ha  $\mathbf{x}^T \mathbf{y} = 0$ , si dice che i due vettori sono *ortogonali*. Se, in aggiunta, si ha  $\mathbf{x}^T \mathbf{x} = \mathbf{y}^T \mathbf{y} = 1$ , allora i due vettori sono detti *ortonormali*. I vettori non nulli  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}$  sono ortogonali se  $(\mathbf{x}^{(i)})^T \mathbf{x}^{(j)} = 0$  per  $i \neq j$ . A partire da un insieme di vettori ortogonali si può costruire un insieme di vettori ortonormali ponendo

$$\mathbf{y}^{(i)} = \frac{\mathbf{x}^{(i)}}{\sqrt{(\mathbf{x}^{(i)})^T \mathbf{x}^{(i)}}}$$

Una matrice quadrata  $\mathbf{Q} \in \mathbb{R}^{n \times n}$  si dice *ortogonale* se  $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ , ossia se le colonne sono vettori normalizzati e ortogonali (brevemente, *ortonormali*). Vedremo nel seguito che per una matrice ortogonale si ha pure  $\mathbf{Q} \mathbf{Q}^T = \mathbf{I}$ , ossia che pure le righe sono vettori ortonormali.



Analogamente, una matrice  $\mathbf{U} \in \mathbb{C}^{n \times n}$  è detta *unitaria* se  $\mathbf{U}^* \mathbf{U} = \mathbf{I}$ . Rileviamo che una matrice unitaria è un caso particolare di matrice normale.

Le matrici ortogonali (e le matrici unitarie) hanno un ruolo importante nel calcolo numerico. Uno dei motivi di tale importanza è il seguente risultato.

**Proposizione A.2** *La moltiplicazione di un vettore mediante una matrice ortogonale  $\mathbf{Q}$  preserva la lunghezza del vettore, ossia per ogni vettore  $\mathbf{x} \in \mathbb{R}^n$  si ha*

$$\boxed{\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2} \quad (\text{A.1})$$

*Inoltre, mantiene il prodotto interno*

$$(\mathbf{Q}\mathbf{x})^T (\mathbf{Q}\mathbf{y}) = \mathbf{x}^T \mathbf{y}$$

per ogni  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ .

La dimostrazione è immediata, in quanto

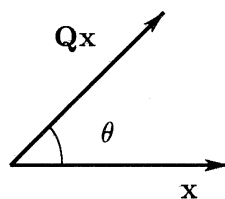
$$(\mathbf{Q}\mathbf{x})^T (\mathbf{Q}\mathbf{y}) = \mathbf{x}^T \mathbf{Q}^T \mathbf{Q} \mathbf{y} = \mathbf{x}^T \mathbf{I} \mathbf{y} = \mathbf{x}^T \mathbf{y}$$

e la (A.1) si ottiene prendendo  $\mathbf{x} = \mathbf{y}$ .

Nel seguito vedremo due importanti classi di matrici ortogonali, ossia le matrici di rotazione, o di Givens, e le matrici di riflessione, o di Householder. Un esempio semplice di matrice di rotazione è la seguente.

$$\mathbf{Q} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \quad \mathbf{Q}^T = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

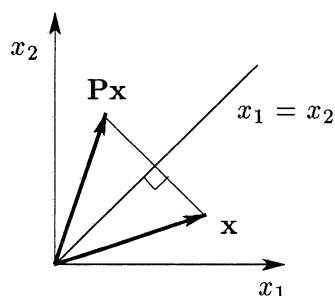
Il prodotto di tale matrice per un vettore  $\mathbf{x}$  ha come risultato la rotazione del vettore di un angolo prefissato  $\theta$ , mentre il prodotto per la matrice  $\mathbf{Q}^T$  ruota il vettore di un angolo  $-\theta$ . Un altro esempio di matrici ortogonali è dato dalle matrici



di *permutazione*. Si ha ad esempio

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{P}^T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (\text{A.2})$$

In questo caso particolare  $\mathbf{P}$  coincide con  $\mathbf{P}^T$ , ma in generale le due matrici sono differenti. Il caso (A.2) rappresenta, anche, un esempio semplice di matrice di riflessione. Il vettore  $\mathbf{P}\mathbf{x}$  corrisponde all'immagine speculare di  $\mathbf{x}$  attraverso la retta  $x_2 = x_1$ .



### A.1.3 Matrici partizionate

In particolare nella risoluzione di sistemi lineari, è utile pensare a una matrice come composta di matrici di ordine inferiore

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1n} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_{m1} & \mathbf{A}_{m2} & \cdots & \mathbf{A}_{mn} \end{bmatrix}$$

ove  $\mathbf{A}_{ij}$  è una matrice  $\in \mathbb{R}^{p_i \times q_j}$ . Usualmente le matrici diagonali  $\mathbf{A}_{ii}$  sono quadrate. In questo caso  $m = n$ ,  $p_i = q_i$ ,  $i = 1, 2, \dots, n$ . Per matrici partizionate nello stesso modo, le operazioni possono essere eseguite formalmente trattando i blocchi come scalari. Ad esempio

$$\mathbf{C} = \mathbf{A}\mathbf{B}, \quad \mathbf{C}_{ij} = \sum_{k=1}^n \mathbf{A}_{ik}\mathbf{B}_{kj}$$

Se  $\mathbf{A}_{ij} = \mathbf{O}$  per  $i \neq j$ , allora  $\mathbf{A}$  è chiamata *matrice diagonale a blocchi*. In modo analogo, si definiscono le *matrici tridiagonali a blocchi*.

### A.1.4 Indipendenza lineare, base e dimensione

**Indipendenza lineare** I vettori  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^m$ , con  $n \leq m$  si dicono *linearmente indipendenti* se dalla condizione

$$\sum_{i=1}^n \alpha_i \mathbf{x}^{(i)} = \mathbf{0}, \quad \alpha_i \in \mathbb{R}$$

segue che  $\alpha_i = 0$ , per  $i = 1, \dots, n$ . In caso contrario, i vettori assegnati sono *linearmente dipendenti*. In questo caso il vettore zero può essere scritto come una combinazione lineare non triviale dei vettori dati (cfr. per una illustrazione in  $\mathbb{R}^2$  la Figura A.3). Quando i vettori  $\mathbf{x}^{(i)}$  sono le colonne  $\mathbf{a}_i$  di una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , si ricava che la dipendenza lineare delle colonne di  $\mathbf{A}$  è equivalente alla condizione

$$\mathbf{A}\mathbf{z} = \mathbf{0} \quad \text{per un vettore } \mathbf{z} \neq \mathbf{0} \quad (\text{A.3})$$

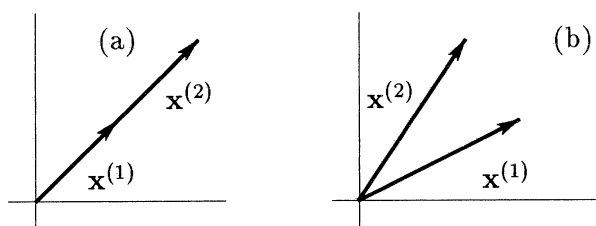


Figura A.3: (a) Vettori linearmente dipendenti. (b) Vettori linearmente indipendenti.

e che la indipendenza lineare delle colonne di  $\mathbf{A}$  è equivalente alla condizione

$$\mathbf{A}\mathbf{z} = \mathbf{0} \Rightarrow \mathbf{z} = \mathbf{0} \quad (\text{A.4})$$

Sia  $\mathcal{S}$  un insieme di vettori dello spazio  $\mathbb{R}^m$ . Si dice che  $\mathcal{S}$  è un *sottospazio* di  $\mathbb{R}^m$  se, per ogni scalare  $\alpha$  e  $\beta$

$$\mathbf{x}, \mathbf{y} \in \mathcal{S} \Rightarrow \alpha\mathbf{x} + \beta\mathbf{y} \in \mathcal{S} \quad (\text{A.5})$$

Tale proprietà implica che ogni sottospazio deve contenere il vettore zero.

L'insieme di tutti i vettori che sono combinazioni lineari delle colonne di una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  è chiamato lo *spazio colonna*, o spazio immagine di  $\mathbf{A}$ , e sarà indicato per il seguito con la notazione  $\mathcal{R}(\mathbf{A})$ . Lo spazio immagine, che sottolineiamo è un sottospazio di  $\mathbb{R}^m$ , può essere espresso nella seguente forma

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\}$$

**Base** Dato un sottospazio  $\mathcal{S}$  di  $\mathbb{R}^m$ ,  $k$  vettori  $\{\mathbf{x}^{(i)}\}, i = 1, \dots, k$  costituiscono una *base* di  $\mathcal{S}$  se ogni vettore  $\mathbf{z} \in \mathcal{S}$  può essere espresso, in maniera unica, come combinazione lineare dei vettori della base

$$\mathbf{z} = \sum_{i=1}^k \alpha_i \mathbf{x}^{(i)}$$

Una base è, quindi, un insieme di vettori linearmente indipendenti che *genera* (span) lo spazio  $\mathcal{S}$ . I numeri  $\alpha_i$  sono le coordinate del punto  $\mathbf{x} \in \mathbb{R}^m$  rispetto alla base fissata.

Una base di  $\mathbb{R}^m$  particolarmente importante è la *base canonica*, formata dai vettori

$$\mathbf{e}_i = [0, \dots, \underset{i}{0, 1, 0, \dots}, 0]^T, \quad i = 1, 2, \dots, m$$

che sono le colonne della matrice identità di ordine  $m$ .

**Dimensione** Si può dimostrare che tutte le basi di un sottospazio hanno lo stesso numero di elementi; tale numero, indicato con  $\dim(\mathcal{S})$ , è detto *dimensione* del sottospazio e esprime il numero dei *gradi di libertà* del sottospazio. Ad esempio, lo spazio  $\mathbb{R}^m$  ha dimensione  $m$ , e ogni insieme di  $m$  vettori linearmente indipendenti di  $\mathbb{R}^m$  costituisce una base di  $\mathbb{R}^m$ .

Ricordiamo il fatto importante che *in un sottospazio di dimensione  $k$ , nessun insieme di più di  $k$  vettori può essere linearmente indipendente e nessun insieme di meno di  $k$  vettori può generare lo spazio.*

Siano  $\mathcal{S}$  e  $\mathcal{T}$  due sottospazi di  $\mathbb{R}^m$ . La somma

$$X = \mathcal{S} + \mathcal{T} := \{\mathbf{x} = \mathbf{s} + \mathbf{t} \mid \mathbf{s} \in \mathcal{S}, \mathbf{t} \in \mathcal{T}\}$$

e l'intersezione  $\mathcal{S} \cap \mathcal{T}$  sono ancora sottospazi. Per le loro dimensioni vale la seguente relazione

$$\dim(\mathcal{S} + \mathcal{T}) = \dim \mathcal{S} + \dim \mathcal{T} - \dim(\mathcal{S} \cap \mathcal{T})$$

da cui segue che

$$\begin{aligned} \max\{\dim \mathcal{S}, \dim \mathcal{T}\} &\leq \dim(\mathcal{S} + \mathcal{T}) \leq \min\{\dim \mathcal{S} + \dim \mathcal{T}, m\} \\ \max\{0, \dim \mathcal{S} + \dim \mathcal{T} - m\} &\leq \dim(\mathcal{S} \cap \mathcal{T}) \leq \min\{\dim \mathcal{S}, \dim \mathcal{T}\} \end{aligned}$$

Se  $\mathcal{S} \cap \mathcal{T} = \{0\}$ , il sottospazio  $X = \mathcal{S} + \mathcal{T}$  è detto *somma diretta* di  $\mathcal{S}$  e  $\mathcal{T}$  e viene usualmente indicato con  $\mathcal{S} \oplus \mathcal{T}$ . In questo caso si ha

$$\dim X = \dim \mathcal{S} + \dim \mathcal{T}$$

e gli elementi  $\mathbf{x}$  di  $X$  possono essere espressi univocamente con la somma

$$\mathbf{x} = \mathbf{s} + \mathbf{t}, \quad \mathbf{s} \in \mathcal{S}, \mathbf{t} \in \mathcal{T}$$

**Proiezione ortogonale** Sia  $\mathcal{S}$  un sottospazio di  $\mathbb{R}^m$ . Il sottospazio

$$\mathcal{S}^\perp := \{\mathbf{u} \in \mathbb{R}^m \mid \mathbf{u}^T \mathbf{v} = 0 \text{ per ogni } \mathbf{v} \in \mathcal{S}\}$$

è detto *sottospazio ortogonale* a  $\mathcal{S}$ . Si hanno le seguenti relazioni

$$\begin{aligned} \mathcal{S} \cap \mathcal{S}^\perp &= \{0\} \\ \mathcal{S} \oplus \mathcal{S}^\perp &= \mathbb{R}^m \\ \dim \mathcal{S}^\perp &= m - \dim \mathcal{S} \end{aligned}$$

Ne segue che ogni vettore  $\mathbf{x} \in \mathbb{R}^m$  può essere espresso in maniera univoca come somma

$$\mathbf{x} = \mathbf{s} + \mathbf{t}, \quad \mathbf{s} \in \mathcal{S}, \mathbf{t} \in \mathcal{S}^\perp \tag{A.6}$$

Il vettore  $\mathbf{s}$  è detto *proiezione ortogonale* di  $\mathbf{x}$  su  $\mathcal{S}$ .

Come illustrazione, consideriamo in  $\mathbb{R}^3$  il sottospazio  $\mathcal{S}$  generato dal vettore  $\mathbf{x}^{(1)} = [0, 0, 1]^T$ , ossia l'insieme di tutti i vettori con le prime due componenti nulle. La sua dimensione è 1. Lo spazio  $\mathcal{S}^\perp$  è costituito dai vettori con la terza componente nulla ed è generato dai vettori  $\mathbf{x}^{(2)} = [1, 0, 0]^T$  e  $\mathbf{x}^{(3)} = [0, 1, 0]^T$  ed ha dimensione 2. La relazione (A.6) è allora illustrata in Figura A.4.

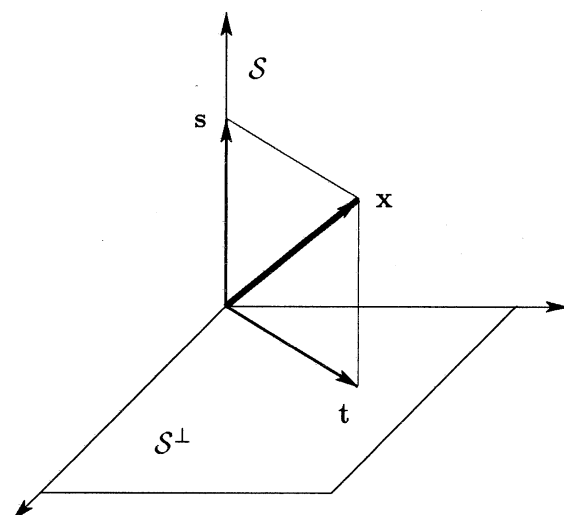


Figura A.4: Proiezione ortogonale di  $\mathbf{x}$  su  $\mathcal{S}$ .

Se, data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , indichiamo con  $\mathcal{N}(\mathbf{A})$ , il nucleo della matrice  $\mathbf{A}$ , o *spazio nullo*, cioè l'insieme dei vettori che verificano  $\mathbf{A}\mathbf{x} = \mathbf{0}$

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$$

si ha che  $\mathcal{N}(\mathbf{A})$  e lo spazio  $\mathcal{R}(\mathbf{A}^T)$ , cioè lo spazio generato dalle righe della matrice  $\mathbf{A}$ , sono sottospazi ortogonali di  $\mathbb{R}^n$ . Supponiamo, infatti, che  $\mathbf{w}$  sia un vettore in  $\mathcal{N}(\mathbf{A})$  e  $\mathbf{v}$  in  $\mathcal{R}(\mathbf{A}^T)$ . Allora,  $\mathbf{A}\mathbf{w} = \mathbf{0}$ , ed inoltre esiste un vettore  $\mathbf{x}$  tale che  $\mathbf{v} = \mathbf{A}^T\mathbf{x}$ . Quindi

$$\mathbf{w}^T\mathbf{v} = \mathbf{w}^T(\mathbf{A}^T\mathbf{x}) = (\mathbf{w}^T\mathbf{A}^T)\mathbf{x} = (\mathbf{A}\mathbf{w})^T\mathbf{x} = \mathbf{0}^T\mathbf{x} = 0$$

Analogamente, si dimostra che lo spazio  $\mathcal{N}(\mathbf{A}^T)$  e lo spazio  $\mathcal{R}(\mathbf{A})$  sono sottospazi ortogonali di  $\mathbb{R}^m$ . Come conseguenza si hanno le seguenti relazioni

$$\dim(\mathcal{R}(\mathbf{A}^T)) + \dim(\mathcal{N}(\mathbf{A})) = n \quad (\text{A.7})$$

$$\dim(\mathcal{R}(\mathbf{A})) + \dim(\mathcal{N}(\mathbf{A}^T)) = m \quad (\text{A.8})$$

Il numero  $\dim(\mathcal{R}(\mathbf{A}^T))$  viene detto *rango* di  $\mathbf{A}$ ; tale nozione verrà riconsiderata nel seguito, ove mostreremo che  $\dim(\mathcal{R}(\mathbf{A}^T)) = \dim(\mathcal{R}(\mathbf{A}))$ .

**Basi ortonormali** Fra le basi di  $\mathbb{R}^m$  sono particolarmente importanti le *basi ortonormali*, nelle quali, cioè, gli elementi della base sono vettori ortonormali. È importante osservare che a partire da una qualunque base è possibile costruire una base ortonormale, come è mostrato dal seguente procedimento, noto come *procedimento di Gram-Schmidt*.

**Proposizione A.3 (Gram-Schmidt)** *Se i vettori  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)} \in \mathbb{R}^m$ , con  $k \leq m$ , sono  $k$  vettori linearmente indipendenti, i vettori  $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(k)}$ , costruiti nel seguente modo*

$$\mathbf{z}^{(1)} = \mathbf{x}^{(1)}, \quad \mathbf{y}^{(1)} = \frac{\mathbf{z}^{(1)}}{\sqrt{(\mathbf{z}^{(1)})^T \mathbf{z}^{(1)}}}$$

$$\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \sum_{j=1}^{i-1} ((\mathbf{y}^{(j)})^T \mathbf{x}^{(i)}) \mathbf{y}^{(j)}, \quad \mathbf{y}^{(i)} = \frac{\mathbf{z}^{(i)}}{\sqrt{(\mathbf{z}^{(i)})^T \mathbf{z}^{(i)}}}$$

per  $i = 2, \dots, k$ , sono ortonormali.

Il risultato, che può essere dimostrato facilmente per induzione, è alla base della possibilità di decomporre una matrice data  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , con le colonne linearmente indipendenti, nel prodotto di una matrice ortogonale  $\mathbf{Q} \in \mathbb{R}^{m \times m}$  e di una matrice triangolare  $\mathbf{R} \in \mathbb{R}^{m \times n}$ . A partire da tale decomposizione è possibile risolvere in maniera *numericamente stabile* i sistemi malcondizionati (cfr. Capitolo 2).

### A.1.5 Determinante, inversa e rango

Sia  $\mathbf{A}$  una matrice quadrata di ordine  $n$ . Il *determinante*<sup>2</sup> di  $\mathbf{A}$ , denotato usualmente con la notazione  $\det(\mathbf{A})$ , è definito da

$$\det(\mathbf{A}) := \sum_{\mathbf{j}} (-1)^{t(\mathbf{j})} a_{1j_1} a_{2j_2} \cdots a_{nj_n} \quad (\text{A.9})$$

ove  $t(\mathbf{j})$  è il numero di inversioni nella permutazione  $\mathbf{j} = (j_1, j_2, \dots, j_n)$  e  $\mathbf{j}$  varia su tutte le  $n!$  permutazioni degli interi  $1, 2, \dots, n$ . La formula (A.9) è l'estensione al caso di  $n$  generico delle formule note per matrici di ordine  $n = 2$  e  $n = 3$

$$n = 2: \quad \det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$$

$$n = 3: \quad \det(\mathbf{A}) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}$$

<sup>2</sup>Il termine *determinante* è stato introdotto per la prima volta nel 1812 da Cauchy, grazie al quale i determinanti diventarono di uso comune nella ricerca matematica. In precedenza, il concetto di determinante, come grandezza che caratterizza una matrice, fu introdotto da Cramer (1750) e sviluppato da Vandermonde (1771) e da Laplace (1772) (che utilizzava il termine di *risultante*).

Osserviamo che il calcolo del determinante mediante la definizione (A.9) richiede la formazione di  $n!$  prodotti, ognuno dei quali richiede  $n-1$  moltiplicazioni, per un totale di  $n!$  addizioni e  $(n-1)n!$  moltiplicazioni. Tuttavia, come è mostrato nel Capitolo 2, il determinante di una generica matrice di ordine  $n$  può essere ottenuto con un numero di addizioni e di moltiplicazioni dell'ordine di  $n^3/3$ . Questo risultato è possibile grazie all'utilizzo delle seguenti proprietà del determinante.

- Il valore del determinante rimane immutato se si aggiunge a una riga (colonna) un'altra riga (colonna) moltiplicata per uno scalare.
- Il determinante di una matrice triangolare è uguale al prodotto degli elementi sulla diagonale principale.
- Se si scambiano due righe (colonne), il determinante cambia di segno.
- Si ha:  $\det(\mathbf{A}) = \det(\mathbf{A}^T)$ ,  $\det(\mathbf{A}^*) = \overline{\det(\mathbf{A})}$ .
- $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$  (regola di Binet).

Il determinante di una matrice può essere espresso in maniera ricorsiva mediante la *regola di Laplace*. Indicata con  $\mathbf{A}_{ij}$  la sottomatrice quadrata di ordine  $n-1$  ottenuta dalla matrice  $\mathbf{A}$  eliminando la  $i$ -ma riga e la  $j$ -ma colonna, per un qualunque indice di riga  $i$  si ha

$$\det(\mathbf{A}) = \begin{cases} a_{11} & \text{se } n = 1 \\ \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij}) & \text{se } n > 1 \end{cases}$$

Il termine  $(-1)^{i+j} \det(\mathbf{A}_{ij})$  è detto *cofattore* dell'elemento  $a_{ij}$ . Si chiamano, inoltre, *minori* i determinanti delle sottomatrici quadrate che si ottengono fissando in maniera qualunque uno stesso numero di righe e di colonne.

Ricordiamo la seguente interessante caratterizzazione, nota come *criterio di Sylvester*<sup>3</sup>, delle matrici definite positive mediante i determinanti delle sottomatrici principali.

**Teorema A.1** (Criterio di Sylvester) *Una matrice simmetrica  $\mathbf{A}$  di ordine  $n$  è definita positiva se e solo se*

$$\det(\mathbf{A}_k) > 0, \quad k = 1, 2, \dots, n$$

ove  $\mathbf{A}_k$  è la matrice principale di ordine  $k$ , cioè la matrice formata dalle prime  $k$  righe e  $k$  colonne della matrice  $\mathbf{A}$ .

Dal risultato precedente si ricava in particolare la seguente maggiorazione.

<sup>3</sup>James Joseph Sylvester (1814-1897).

**Teorema A.2** Per una matrice  $\mathbf{A}$  simmetrica definita positiva di ordine  $n$ , si ha

$$|a_{ij}|^2 < a_{ii} a_{jj}, \quad i, j = 1, 2, \dots, n$$

e quindi il massimo elemento di  $\mathbf{A}$  si trova sulla diagonale.

Ricordiamo, infine, che il determinante di una matrice  $\mathbf{A}$  è uguale al volume del parallelepipedo in  $n$  dimensioni i cui spigoli corrispondono alle righe della matrice  $\mathbf{A}$  (cfr. per  $n = 2$  la Figura A.5).

Il risultato è ovvio quando le righe della matrice sono ortogonali. In tale caso, infatti, il prodotto  $\mathbf{A}\mathbf{A}^T$  è una matrice diagonale con gli elementi sulla diagonale forniti dai quadrati  $l_i^2$  delle lunghezze delle righe. Dalle proprietà del determinante si ha allora

$$l_1^2 l_2^2 \cdots l_n^2 = \det(\mathbf{A}\mathbf{A}^T) = (\det \mathbf{A}) (\det \mathbf{A}^T) = (\det \mathbf{A})^2$$

Il segno di  $\det(\mathbf{A})$  indica l'orientamento dell'insieme delle coordinate. Il caso generale si riconduce al caso ortogonale, applicando alle righe della matrice il procedimento di ortogonalizzazione di Gram-Schmidt.

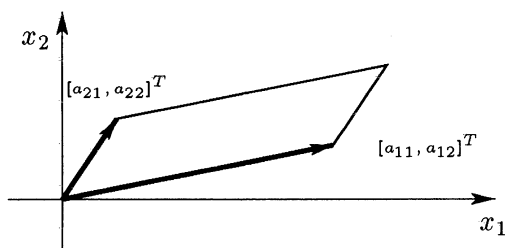


Figura A.5: Il  $\det(\mathbf{A})$  fornisce l'area del parallelogramma.

**Inversa** Una matrice quadrata  $\mathbf{A}$  di ordine  $n$  è detta *non singolare* quando per essa si ha  $\det(\mathbf{A}) \neq 0$ .

Una matrice  $\mathbf{B}$  quadrata di ordine  $n$ , è una *inversa* di  $\mathbf{A}$  se

$$\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} = \mathbf{I}$$

in tal caso si scrive  $\mathbf{B} = \mathbf{A}^{-1}$ , e la matrice  $\mathbf{A}$  si dice *invertibile*. Il determinante fornisce un test per l'invertibilità della matrice; si ha, infatti, il seguente importante risultato.

**Teorema A.3** Una matrice quadrata è invertibile se e solo se essa è non singolare.



Ricordiamo che viene chiamata *matrice aggiunta* di  $\mathbf{A}$  la matrice  $\text{adj}(\mathbf{A})$  di ordine  $n$ , il cui elemento  $(i, j)$ -mo è dato da

$$(-1)^{i+j} \det(\mathbf{A}_{ji})$$

ove  $\mathbf{A}_{ji}$  è la sottomatrice ottenuta da  $\mathbf{A}$  cancellando la riga  $j$ -ma e la colonna  $i$ -ma. Mediante la matrice aggiunta si può fornire la seguente formula esplicita (di interesse, usualmente, solo teorico) per l'inversa di una matrice non singolare

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \text{adj}(\mathbf{A}) \quad (\text{A.10})$$

**Rango** Il *rango* di una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  può essere definito come l'*ordine massimo dei minori non nulli della matrice*. Esso viene denotato usualmente con la notazione:  $r = \text{rank}(\mathbf{A})$ .

Una matrice  $\mathbf{A}$  è di *rango completo* quando:  $\text{rank}(\mathbf{A}) = \min(m, n)$ ; in caso contrario viene detta a *rango deficiente* (rank-deficient).

L'importanza della nozione di rango risiede nel fatto che il rango di una matrice  $\mathbf{A}$  rappresenta *il numero massimo delle colonne, e quindi anche delle righe, della matrice  $\mathbf{A}$  che sono linearmente indipendenti*. Il rango di  $\mathbf{A}$  è quindi la dimensione di  $\mathcal{R}(\mathbf{A})$  (e ugualmente di  $\mathcal{R}(\mathbf{A}^T)$ ), lo spazio lineare generato dalle colonne (rispettivamente, le righe) di  $\mathbf{A}$ .

Si ha il seguente importante risultato (cfr. (A.7)), valido per una matrice rettangolare qualunque

$$\boxed{\text{rank}(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = n}$$

In particolare, quando  $m = n$  e  $\mathbf{A}$  è una matrice non singolare, allora  $\text{rank}(\mathbf{A}) = n$ ,  $\mathcal{N}(\mathbf{A}) = \{0\}$ .

◆ **Esercizio A.1** Verificare che per una matrice tridiagonale, detta anche di Jacobi di ordine  $n$

$$\mathbf{J}_n = \begin{bmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ c_1 & a_2 & b_2 & \ddots & \vdots \\ 0 & c_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & 0 & c_{n-1} & a_n \end{bmatrix}$$

si ha  $\det(\mathbf{J}_n) = a_n \det(\mathbf{J}_{n-1}) - b_{n-1} c_{n-1} \det(\mathbf{J}_{n-2})$ , per  $n \geq 3$ .

◆ **Esercizio A.2** Verificare che il determinante della matrice di Vandermonde  $\mathbf{V}_n(x_1, x_2, \dots, x_n)$  definita da

$$\mathbf{V}_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{bmatrix}$$

è dato da  $\prod_{1 \leq j < i \leq n} (x_i - x_j)$ .

◆ **Esercizio A.3** Dimostrare che se tutti i minori di ordine  $k$  ( $1 \leq k < \min(m, n)$ ) di una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  sono uguali a zero, allora ogni minore di  $\mathbf{A}$  di ordine più grande di  $k$  è pure uguale a zero.

◆ **Esercizio A.4** Una matrice di Hadamard di ordine  $n$  è una matrice con elementi che sono tutti  $\pm 1$  e soddisfa  $\mathbf{A}^T \mathbf{A} = n\mathbf{I}$ . Mostrare che  $|\det(\mathbf{A})| = n^{n/2}$ .

◆ **Esercizio A.5** Dimostrare che se  $\mathbf{A}$ ,  $\mathbf{B}$  sono non singolari, allora

$$\det(\mathbf{A}^{-1}) = (\det(\mathbf{A}))^{-1} \\ (\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}$$

◆ **Esercizio A.6** Mostrare che

- una matrice ortogonale  $\mathbf{A}$  è invertibile e che  $\det(\mathbf{A}) = \pm 1$  e  $\mathbf{A}^{-1} = \mathbf{A}^T$
- Se  $n = 2$ , tutte le matrici ortogonali hanno una delle due forme

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \quad \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix}$$

◆ **Esercizio A.7** Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  è idempotente se  $\mathbf{A}^2 = \mathbf{A}$ . Mostrare che una matrice idempotente è singolare, a meno che essa sia la matrice identità.

◆ **Esercizio A.8** Sia  $\mathbf{A} = \mathbf{I} + \mathbf{B}$ , con  $\mathbf{B} \in \mathbb{C}^{n \times n}$  è una matrice triangolare in senso stretto, ossia  $b_{ii} = 0$ ,  $i = 1, 2, \dots, n$ . Si dimostri che la matrice  $\mathbf{A}$  è invertibile e che

$$\mathbf{A}^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \cdots + (-1)^k \mathbf{B}^k, \quad \text{ove } k \leq n - 1$$

◆ **Esercizio A.9** Mostrare che se  $\mathbf{L}$  è una matrice quadrata non singolare, allora  $\mathbf{L}^T \mathbf{L}$  è una matrice simmetrica definita positiva. Se  $\mathbf{A}$  è una matrice simmetrica definita positiva, esiste un'unica matrice  $\mathbf{J}$  simmetrica definita positiva tale che  $\mathbf{J}^2 = \mathbf{A}$ .

◆ **Esercizio A.10** Se  $\mathbf{A}$  è una matrice definita positiva e  $\mathbf{C}$  è una matrice non singolare, dimostrare che  $\mathbf{B} = \mathbf{C}^T \mathbf{A} \mathbf{C}$  è ancora una matrice definita positiva.

◆ **Esercizio A.11** Verificare che se  $\mathbf{x} \in \mathbb{R}^n$  e  $\mathbf{y} \in \mathbb{R}^m$  il prodotto esterno  $\mathbf{xy}^T$  è una matrice di rango uno.

◆ **Esercizio A.12** *Provare che per ogni matrice rettangolare  $\mathbf{A}$*

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T) = \text{rank}(\mathbf{A}^*)$$

◆ **Esercizio A.13** *Esaminare per quali valori dei parametri  $\alpha$  e  $\beta$  la seguente matrice*

$$\mathbf{A} = \begin{bmatrix} 1/2 & (\sqrt{3}/2)i \\ \alpha & \beta \end{bmatrix}$$

è a) normale, b) unitaria, c) hermitiana, d) definita positiva.

◆ **Esercizio A.14** *Verificare che il rango di una matrice simmetrica è uguale all'ordine massimo dei minori principali non nulli.*

◆ **Esercizio A.15** *Se  $\mathbf{A}_1$  e  $\mathbf{A}_4$  sono matrici quadrate, mostrare che*

$$\det \begin{bmatrix} \mathbf{A}_1 & \mathbf{O} \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} = \det(\mathbf{A}_1) \det(\mathbf{A}_4)$$

Più in generale mostrare che se  $\mathbf{A}$  è una matrice triangolare a blocchi con blocchi diagonali  $\mathbf{A}_i$  quadrati, si ha

$$\det(\mathbf{A}) = \prod_{i=1}^k \det(\mathbf{A}_i).$$

◆ **Esercizio A.16** *Mostrare che se  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , allora*

$$\det(\mathbf{I}_n + \mathbf{x}\mathbf{y}^T) = 1 + \mathbf{x}^T\mathbf{y}$$

### A.1.6 Matrici elementari

La maggior parte degli algoritmi relativi al trattamento delle matrici e alla risoluzione dei sistemi lineari sono basati su successive moltiplicazioni di matrici *elementari*  $\mathbf{H}(\alpha, \mathbf{u}, \mathbf{v})$ , della forma seguente.

$$\boxed{\mathbf{H}(\alpha, \mathbf{u}, \mathbf{v}) = \mathbf{I} - \alpha \mathbf{u} \mathbf{v}^T}$$

dove  $\alpha$  è uno scalare e  $\mathbf{u}, \mathbf{v}$  sono vettori  $\in \mathbb{R}^n$ . La matrice  $\mathbf{u}\mathbf{v}^T$  è una matrice  $\in \mathbb{R}^{n \times n}$  di rango 1.

Una proprietà interessante di tali matrici è la seguente

$$\mathbf{H}(\alpha, \mathbf{u}, \mathbf{v}) \mathbf{H}(\beta, \mathbf{u}, \mathbf{v}) = \mathbf{H}(\gamma, \mathbf{u}, \mathbf{v})$$

ove

$$\gamma = \alpha + \beta - \alpha \beta (\mathbf{v}^T \mathbf{u})$$

Se  $\gamma = 0$ , allora  $\mathbf{H}(\beta, \mathbf{u}, \mathbf{v})$  è l'inversa di  $\mathbf{H}(\alpha, \mathbf{u}, \mathbf{v})$ ; cioè

$$\beta = \frac{\alpha}{\alpha \mathbf{v}^T \mathbf{u} - 1}$$

Quando  $\mathbf{v}^T \mathbf{u} = 1/\alpha$ , il denominatore è zero e  $\mathbf{H}(\alpha, \mathbf{u}, \mathbf{v})$  è singolare.

L'espressione dell'inversa di  $\mathbf{H}(\alpha, \mathbf{u}, \mathbf{v})$  ha come conseguenza la possibilità di calcolare direttamente l'effetto sull'inversa di una modifica della matrice mediante una matrice di rango 1.

Scrivendo, infatti

$$\mathbf{B} = \mathbf{A} - \mathbf{u}\mathbf{v}^T = \mathbf{A}(\mathbf{I} - \mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T)$$

si ha che la matrice tra parentesi è una matrice elementare della forma  $\mathbf{H}(1, \mathbf{A}^{-1}\mathbf{u}, \mathbf{v})$ . Quindi l'inversa di  $\mathbf{B}$  è data da

$$\mathbf{B}^{-1} = \left[ \mathbf{I} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T}{\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u} - 1} \right] \mathbf{A}^{-1} \quad (\text{A.11})$$

$$= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u} - 1} \quad (\text{A.12})$$

Il risultato ora ottenuto è noto come *formula di Sherman–Morrison*. Considereremo nel seguito alcuni casi importanti di matrici elementari.

### Matrici di Gauss

Le matrici utilizzate nei metodi di eliminazione sono definite da

$$\mathbf{H}(1, \mathbf{m}_i, \mathbf{e}_i) = \mathbf{I} - \mathbf{m}_i \mathbf{e}_i^T \equiv: \mathbf{M}_i$$

ove

$$\mathbf{m}_i^T = (0, \dots, 0, m_{i+1i}, \dots, m_{ni})$$

e  $\mathbf{e}_i$  è la componente  $i$ -ma della base canonica di  $\mathbb{R}^n$ .

La pre-moltiplicazione di  $\mathbf{A}$  con  $\mathbf{M}_i$  ha l'effetto di sottrarre da ogni riga  $r$  ( $r = i + 1, \dots, n$ ) la riga  $i$  moltiplicata per  $m_{ri}$ ; ogni colonna di  $\mathbf{A}$  è interessata indipendentemente dalle altre.

Se si assume, per ogni  $j$  fissato, con  $i \leq j \leq n$

$$m_{ri} = \frac{a_{rj}}{a_{ij}} \quad r = i + 1, i + 2, \dots, n$$

allora vengono azzerati gli elementi  $(i + 1, j), (i + 2, j), \dots, (n, j)$ .

Le matrici  $\mathbf{M}_i$  vengono utilizzate nei metodi di eliminazione congiuntamente con le matrici elementari  $\mathbf{I}_{ii'}$  definite da

$$\mathbf{I}_{ii'} = \mathbf{I} - (\mathbf{e}_i - \mathbf{e}_{i'}) (\mathbf{e}_i - \mathbf{e}_{i'})^T$$

La pre-moltiplicazione con  $\mathbf{I}_{ii'}$  scambia le righe  $i$  e  $i'$ .

L'inversa della matrice  $\mathbf{M}_i$  è data da

$$\mathbf{M}_i^{-1} = \mathbf{I} + \mathbf{m}_i \mathbf{e}_i^T$$

che è ancora una matrice di tipo  $\mathbf{M}_i$ . Si ha anche

$$\mathbf{M}_i \mathbf{M}_j = (\mathbf{I} - \mathbf{m}_i \mathbf{e}_i^T)(\mathbf{I} - \mathbf{m}_j \mathbf{e}_j^T) = \mathbf{I} - \mathbf{m}_i \mathbf{e}_i^T - \mathbf{m}_j \mathbf{e}_j^T \quad (j \geq i)$$

Ne segue che la matrice

$$\mathbf{M}_1^{-1} \mathbf{M}_2^{-1} \cdots \mathbf{M}_{n-1}^{-1} = \mathbf{I} + \mathbf{m}_1 \mathbf{e}_1^T + \cdots + \mathbf{m}_{n-1} \mathbf{e}_{n-1}^T$$

è una matrice triangolare inferiore con elementi  $\mathbf{m}_{ij}$  e 1 sulla diagonale principale.

### Matrici di Householder

Le *matrici di Householder* (o trasformate di Householder) sono matrici elementari del tipo

$$\mathbf{H}(2, \mathbf{u}, \mathbf{u}) = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$$

ove  $\mathbf{u} \in \mathbb{R}^n$  è un vettore assegnato di lunghezza unitaria  $\|\mathbf{u}\|_2 = 1$ .

Si tratta di matrici *ortogonali e simmetriche*. Vengono anche indicate come matrici di *riflessione*, in quanto rappresentano la simmetria rispetto all'iperpiano passante per l'origine e ortogonale al vettore  $\mathbf{u}$  (cfr. Figura A.6).

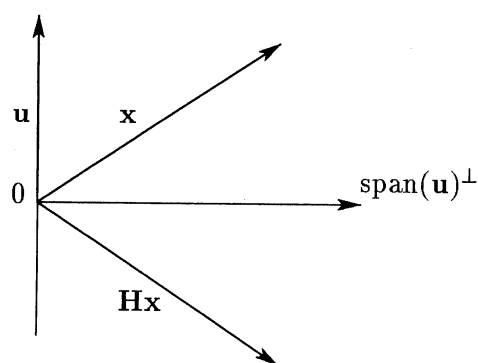


Figura A.6: Trasformata di Householder.

▼ **Osservazione A.2** Nel caso di vettori in  $\mathbb{C}^n$ , si definisce la matrice  $\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^*$ , che è hermitiana e unitaria se e solo se  $\|\mathbf{u}\|_2^2 = \mathbf{u}^* \mathbf{u} = 1$ . ■

Dal punto di vista numerico le matrici di Householder sono importanti in quanto possono essere utilizzate per annullare in maniera numericamente stabile elementi specificati in una matrice o in un vettore. Come illustrazione, si consideri  $\mathbf{x} = [3, 1, 5, 1]^T$ ,  $\mathbf{u} = [9, 1, 5, 1]^T / \sqrt{108}$ ; allora

$$\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T = \frac{1}{54} \begin{bmatrix} -27 & -9 & -45 & -9 \\ -9 & 53 & -5 & -1 \\ -45 & -5 & 29 & -5 \\ -9 & -1 & -5 & 53 \end{bmatrix}$$

è tale che  $\mathbf{H}\mathbf{x} = [-6, 0, 0, 0]^T$ . Si osservi che  $\|\mathbf{H}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ . Più in generale si ha la seguente proprietà.

**Proposizione A.4** *Siano  $\mathbf{x}$  e  $\mathbf{y}$  due vettori qualunque di  $\mathbb{R}^n$ , linearmente indipendenti con  $\|\mathbf{y}\|_2 = 1$ . Si può allora determinare un vettore  $\mathbf{u} \in \mathbb{R}^n$ , con  $\|\mathbf{u}\|_2 = 1$ , e uno scalare  $\alpha$  tali che se  $\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$ , allora  $\mathbf{H}\mathbf{x} = \alpha\mathbf{y}$ .*

**DIMOSTRAZIONE.** Poiché  $\mathbf{H}$  è una matrice ortogonale, si ha  $\|\mathbf{x}\| = \|\mathbf{H}\mathbf{x}\| = \|\alpha\mathbf{y}\| = |\alpha|$ , da cui si hanno due possibili scelte per  $\alpha$ :  $\|\mathbf{x}\|$  o  $-\|\mathbf{x}\|$ . L'uguaglianza  $\mathbf{H}\mathbf{x} = \alpha\mathbf{y}$  si scrive

$$\mathbf{x} - 2(\mathbf{u}^T\mathbf{x})\mathbf{u} = \alpha\mathbf{y}, \quad \text{oppure} \quad \mathbf{x} - \alpha\mathbf{y} = 2\lambda\mathbf{u}, \quad \text{con} \quad \lambda = \mathbf{u}^T\mathbf{x}$$

Per calcolare  $\lambda$  si esegue il prodotto scalare dell'uguaglianza precedente per  $\mathbf{x}$

$$\mathbf{x}^T\mathbf{x} - \alpha\mathbf{x}^T\mathbf{y} = 2\lambda\mathbf{x}^T\mathbf{u} = 2\lambda^2.$$

Poiché i vettori  $\mathbf{x}, \mathbf{y}$  non sono linearmente dipendenti, si ha:  $|\mathbf{x}^T\mathbf{y}| < \|\mathbf{x}\| \|\mathbf{y}\| = |\alpha|$ , e l'uguaglianza precedente permette di determinare  $\lambda \neq 0$ , poiché  $\mathbf{x}^T\mathbf{x} - \alpha\mathbf{x}^T\mathbf{y} = \alpha^2 - \alpha\mathbf{x}^T\mathbf{y} > 0$ . Il vettore  $\mathbf{u}$  è allora definito da  $\mathbf{u} = (\mathbf{x} - \alpha\mathbf{y})/(2\lambda)$ . La matrice  $\mathbf{H}$  non dipende dal segno di  $\lambda$ . ■

▼ **Osservazione A.3** *Nelle applicazioni si ha interesse a porre  $\mathbf{H}$  sotto la forma  $\mathbf{I} - \frac{1}{\beta}\mathbf{v}\mathbf{v}^T$ , con  $\mathbf{v} = 2\lambda\mathbf{u} = \mathbf{x} - \alpha\mathbf{y}$  e  $\beta = 2\lambda^2 = \alpha^2 - \alpha(\mathbf{x}^T\mathbf{y})$ , che evita di dover estrarre una radice quadrata per il calcolo di  $\lambda$ .*

*In generale non si ha bisogno di calcolare  $\mathbf{H}$  esplicitamente, ma solamente di poter calcolare il trasformato di un generico vettore  $\mathbf{z}$ , cioè*

$$\mathbf{H}\mathbf{z} = \mathbf{z} - \frac{\mathbf{v}\mathbf{v}^T\mathbf{z}}{\beta}$$

*Tale risultato è ottenuto calcolando dapprima  $\gamma = (\mathbf{v}^T\mathbf{z})/\beta$  e poi  $\mathbf{H}\mathbf{z} = \mathbf{z} - \gamma\mathbf{v}$ .* ■

▼ **Osservazione A.4** *Si è visto che nel calcolo di  $\beta$  si hanno due possibili scelte di  $\alpha$ ; poiché  $\beta > 0$  interviene al denominatore, tra le due scelte possibili, si mantiene quella che fornisce il valore più grande di  $\beta$ , cioè si prende il segno di  $\alpha$  opposto a quello di  $\mathbf{x}^T\mathbf{y}$ . Questo per motivi di stabilità rispetto agli errori di arrotondamento. Si osservi che le due possibili soluzioni,  $\alpha > 0$  e  $\alpha < 0$ , danno origine a due vettori  $\mathbf{v}_1$  e  $\mathbf{v}_2$ , ortogonali, poiché  $\mathbf{v}_1^T\mathbf{v}_2 = (\mathbf{x} - \alpha\mathbf{y})^T(\mathbf{x} + \alpha\mathbf{y}) = \|\mathbf{x}\|^2 - \alpha^2 = 0$ .* ■

▼ **Osservazione A.5** *Nel caso che il vettore  $\mathbf{y}$  sia ad esempio il vettore  $\mathbf{e}_1$ , le operazioni da eseguire sono le seguenti*

$$\begin{aligned} \alpha^2 &= \sum_{i=1}^n x_i^2, & e \quad \text{sign}(\alpha) &= -\text{sign}(x_1) \\ \beta &= \alpha^2 - \alpha x_1 = \alpha^2 + |\alpha x_1| \\ v_1 &= x_1 - \alpha \\ v_i &= x_i, & \text{per } 2 \leq i \leq n \end{aligned}$$

*Il costo è dato da  $n + 1$  addizioni e moltiplicazioni e una radice quadrata.* ■

Il seguente algoritmo utilizza le matrici di Householder per ridurre a zero un qualsiasi blocco di componenti contigue di un vettore.

**Algoritmo A.1** (trasformazione di Householder) *Dato  $\mathbf{x} \in \mathbb{R}^n$  e due indici  $k, j$ , con  $1 \leq k \leq j \leq n$ , l'algoritmo calcola  $\mathbf{v}^T = [0, \dots, v_k, \dots, v_j, 0, \dots, 0]$  e  $\delta = 2/\sqrt{\mathbf{v}^T \mathbf{v}}$  tale che le componenti da  $k+1$  fino a  $j$  di  $(\mathbf{I} - \delta \mathbf{v} \mathbf{v}^T) \mathbf{x}$  siano nulle. Si suppone che  $(x_k, \dots, x_j) \neq 0$ .*

```

m := max(|x_k|, ..., |x_j|)
α := 0
For i = k to j
    v_i := x_i/m
    α := α + v_i^2
end i
α := √α
δ := 1/(α(α + |v_k|))
v_k := v_k + sign(v_k)α

```

L'algoritmo richiede approssimativamente  $2(j - k)$  flops.

Se si pre-moltiplica una matrice  $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$  per una matrice di Householder  $\mathbf{H}$ , si ottiene una matrice  $\mathbf{A}'$  le cui colonne sono trasformate nel seguente modo

$$\mathbf{a}'_k = (\mathbf{I} - 2\mathbf{u}\mathbf{u}^T)\mathbf{a}_k = \mathbf{a}_k - 2(\mathbf{u}^T \mathbf{a}_k)\mathbf{u}$$

Allo stesso modo nella post-moltiplicazione sono le righe ad essere trasformate indipendentemente.

**Algoritmo A.2** (prodotto per una matrice di Householder) *Dato una matrice  $\mathbf{A} \in \mathbb{R}^{n \times q}$ , un vettore  $\mathbf{v}^T = [0, \dots, v_k, \dots, v_j, 0, \dots, 0]$  e  $\delta = 2/\sqrt{\mathbf{v}^T \mathbf{v}}$ , l'algoritmo sostituisce alla matrice  $\mathbf{A}$  la matrice  $(\mathbf{I} - \delta \mathbf{v} \mathbf{v}^T)\mathbf{A}$ .*

```

For p = 1, ..., q
    s := v_k a_kp + ... + v_j a_jp
end p
s := δ s
For i = k, ..., j
    a_ip := a_ip - s v_i
end i

```

L'algoritmo richiede  $2q(j - k + 1)$  flops. Un analogo algoritmo si ha per  $\mathbf{A} := \mathbf{A}(\mathbf{I} - \delta \mathbf{v} \mathbf{v}^T)$ .

### Trasformazioni di Givens

Le *trasformazioni di Givens* corrispondono a correzioni di rango due della matrice identità e permettono, rispetto alle trasformazioni di Householder, di operare in maniera *più selettiva* sugli elementi di una matrice. Risultano, pertanto, di particolare interesse nel trattamento delle matrici *sparse*.

Le trasformazioni di Givens sono, in sostanza, matrici di *rotazione* della forma

$$\mathbf{J}(i, k, \theta) = \begin{bmatrix} 1 & \vdots & & \vdots & & \\ \cdots & c & \cdots & s & \cdots & \\ & \vdots & & \vdots & & \\ \cdots & -s & \cdots & c & \cdots & \\ & \vdots & & \vdots & & 1 \\ & & & i & & k \end{bmatrix}$$

ove:  $c = \cos \theta$ ,  $s = \sin \theta$ , con  $\theta$  angolo fissato. Si tratta evidentemente di matrici *ortogonali*.

La *premultiplicazione* mediante  $\mathbf{J}(i, k, \theta)$  corrisponde a una rotazione di  $\theta$  gradi nel piano delle coordinate  $(i, k)$ . Infatti, se  $\mathbf{x} \in \mathbb{R}^n$  e  $\mathbf{y} = \mathbf{J}(i, k, \theta)\mathbf{x}$ , allora

$$\begin{aligned} y_i &= cx_i + sx_k \\ y_k &= -sx_i + cx_k \\ y_j &= x_j, \quad j \neq i, k \end{aligned}$$

Da queste formule si ricava la possibilità di trasformare a zero l'elemento di indice  $k$ , prendendo

$$c = \frac{x_i}{\sqrt{x_i^2 + x_k^2}}; \quad s = \frac{x_k}{\sqrt{x_i^2 + x_k^2}}$$

Ad esempio se  $\mathbf{x} = [1, 2, 3, 4]$ , si ha  $\mathbf{J}(2, 4, \theta)\mathbf{x} = [1, \sqrt{20}, 3, 0]^T$ , quando  $\cos(\theta) = 1/\sqrt{5}$  e  $\sin(\theta) = 2/\sqrt{5}$ . Osserviamo che il calcolo di  $c$  e  $s$  può essere effettuato in forma *numericamente stabile*, mediante il seguente algoritmo.

**Algoritmo A.3** (trasformata di Givens) *Dato  $\mathbf{x} \in \mathbb{R}^n$  e gli indici  $1 \leq i < k \leq n$ , l'algoritmo calcola  $c, s$  tali che la componente  $k$ -ma del vettore  $\mathbf{y} = \mathbf{J}(i, k, \theta)\mathbf{x}$  sia nulla.*

```

If  $x_k = 0$ 
then
   $c := 1$  and  $s := 0$ 
else
  if  $|x_k| \geq |x_i|$ 
  then
     $t := x_i/x_k$ ,  $s := 1/\sqrt{1+t^2}$ ,  $c := st$ 
  else
     $t := x_k/x_i$ ,  $c := 1/\sqrt{1+t^2}$ ,  $s := ct$ 

```

L'algoritmo richiede 4 flops e una sola radice quadrata. L'angolo  $\theta$  non viene esplicitamente calcolato.

Quando si premoltiplica per la matrice  $\mathbf{J}(i, k, \theta)$  una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , solo le righe  $i$  e  $k$  risultano modificate attraverso il seguente algoritmo.



**Algoritmo A.4** (premultiplicazione) Data  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $c = \cos(\theta)$ ,  $s = \sin(\theta)$ , e gli indici  $1 \leq i < k \leq n$ , la matrice  $\mathbf{A}' = \mathbf{J}(i, k, \theta)\mathbf{A}$  è data da

$$\begin{aligned} \text{For } j = 1, \dots, n \\ a'_{ij} &:= c a_{ij} + s a_{kj} \\ a'_{kj} &:= -s a_{ij} + c a_{kj} \end{aligned}$$

L'algoritmo richiede  $4n$  flops.

## A.2 Sistemi lineari

La risoluzione di un sistema lineare consiste nella ricerca di un vettore  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ , le cui componenti risolvono le seguenti equazioni lineari

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{array} \right. \quad (\text{A.13})$$

che in forma matriciale si scrivono

$$\mathbf{Ax} = \mathbf{b} \quad (\text{A.14})$$

con  $\mathbf{A}$  chiamata *matrice dei coefficienti* e  $\mathbf{b}$  *vettore dei termini noti*. Più precisamente, (A.13) è detto un *sistema lineare di m equazioni in n incognite*. Se  $\mathbf{b} = 0$ , il sistema è detto *omogeneo*, e ammette sempre almeno la soluzione nulla. Se  $\text{rank}(\mathbf{A}) = r < n$ , allora il sistema omogeneo  $\mathbf{Ax} = 0$  ha  $(n-r)$  soluzioni *linearmente indipendenti*.

Indicando con  $\mathbf{a}_j$  la colonna  $j$ -ma della matrice  $\mathbf{A}$ , il sistema (A.13) può essere scritto nella seguente forma

$$\mathbf{b} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \dots + x_n\mathbf{a}_n \quad (\text{A.15})$$

che esprime il vettore termine noto come *combinazione lineare* dei vettori colonna di  $\mathbf{A}$ .

Dalla rappresentazione (A.15) si vede che la condizione di *risolubilità* del sistema è la seguente

$$\mathbf{b} \in \mathcal{R}(\mathbf{A})$$

ove, come abbiamo visto in precedenza,  $\mathcal{R}(\mathbf{A})$  è lo spazio generato dalle colonne di  $\mathbf{A}$ .

In altra forma, si ha che il sistema è risolubile quando  $\text{rank}([\mathbf{A}, \mathbf{b}]) = \text{rank}(\mathbf{A})$ , ove con  $[\mathbf{A}, \mathbf{b}]$  si indica la cosiddetta matrice *augmentata*, ottenuta aggiungendo alle colonne di  $\mathbf{A}$  il vettore  $\mathbf{b}$ .

L'*unicità* della soluzione equivale a richiedere che  $\mathcal{N}(\mathbf{A}) = \{0\}$ , ossia che si abbia  $\dim(\mathcal{N}(\mathbf{A})) = 0$ .

Se  $m = n = r$ , allora  $\mathcal{R}(\mathbf{A}) = \mathbb{R}^n$  e pertanto il criterio di risolubilità è soddisfatto per qualunque vettore  $\mathbf{b}$ . In questo caso la soluzione è univocamente determinata e può essere espressa come  $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ . Utilizzando la formula (A.10) si ottiene la seguente rappresentazione esplicita della soluzione

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \frac{\text{adj}(\mathbf{A})\mathbf{b}}{\det(\mathbf{A})}$$

Tale risultato è usualmente scritto sotto una forma equivalente, nota come *regola di Cramer*<sup>4</sup>. In tale regola la componente  $j$ -ma di  $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$  è data da

$$x_j = \frac{\det(\mathbf{B}_j)}{\det(\mathbf{A})}, \quad \text{ove } \mathbf{B}_j = \begin{bmatrix} a_{11} & a_{12} & b_1 & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & b_n & a_{nn} \end{bmatrix}$$

La matrice  $\mathbf{B}_j$  è ottenuta sostituendo alla colonna  $j$ -ma di  $\mathbf{A}$  il vettore  $\mathbf{b}$ .

Se il criterio di risolubilità è soddisfatto, ma  $\text{rank}(\mathbf{A}) = r < n$ , allora le soluzioni del sistema formano uno spazio vettoriale di dimensione  $n-r$ .

▼ **Osservazione A.6** *Il teorema fondamentale dell'algebra lineare è spesso stabilito nella forma di Fredholm: per ogni  $\mathbf{A}$  e  $\mathbf{b}$ , uno ed uno solo dei seguenti sistemi ha soluzione*

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad \text{oppure } \mathbf{A}^T\mathbf{y} = 0, \quad \mathbf{y}^T\mathbf{b} \neq 0$$

In altre parole, o  $\mathbf{b}$  è nello spazio colonna  $\mathcal{R}(\mathbf{A})$  oppure vi è un  $\mathbf{y}$  in  $\mathcal{N}(\mathbf{A}^T)$  tale che  $\mathbf{y}^T\mathbf{b} \neq 0$ . ■

### A.3 Autovalori e trasformazioni per similitudine

Sia  $\mathbf{A}$  una matrice quadrata di ordine  $n$  ad elementi reali o complessi.

**Definizione A.1** *Si chiama autovalore<sup>5</sup> di  $\mathbf{A}$  un numero  $\lambda \in \mathbb{C}$  per il quale esiste un vettore  $\mathbf{u} \in \mathbb{C}^n$ , non nullo, chiamato autovettore destro di  $\mathbf{A}$  tale che*

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$$

<sup>4</sup>Gabriel Cramer (1704-1752).

<sup>5</sup>La nozione di autovalore appare per la prima volta nel secolo XVIII in relazione alla teoria delle equazioni differenziali lineari omogenee a coefficienti costanti (Lagrange, 1762; studio del moto di un sistema a  $n$  parametri in prossimità di una posizione di equilibrio). Nell'ambito della teoria delle matrici la nozione di autovalore si sviluppa ad opera di Cauchy (1829), Sylvester (1851) e Hamilton (1853). Per indicare l'autovalore e l'autovettore si è usato per lungo tempo i termini *latent root* e *latent point*, che derivano dal fatto che se  $\mathbf{u}$  è un autovettore di  $\mathbf{A}$ , allora  $\mathbf{A}\mathbf{u}$  è sovrapposto a  $\mathbf{u}$ . I termini *eigenvalue* e *eigenvector* sono mutuati da termini in tedesco *Eigenwert* e *Eigenvektor*, che letteralmente significano valore proprio e vettore proprio.

Lo *spettro* di  $\mathbf{A}$ , indicato usualmente con  $\text{Sp}(\mathbf{A})$ , è l'insieme degli autovalori della matrice  $\mathbf{A}$ . Il *raggio spettrale* di  $\mathbf{A}$  è dato dal seguente numero

$$\rho(\mathbf{A}) = \max_{1 \leq i \leq n} |\lambda_i|$$

Osserviamo che un autovettore è definito a meno di un fattore moltiplicativo. In effetti, se  $\alpha \neq 0$ , allora  $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$  è equivalente a  $\mathbf{A}(\alpha\mathbf{u}) = \lambda(\alpha\mathbf{u})$ .

Quando  $\lambda \in \text{Sp}(\mathbf{A})$ , la matrice  $\mathbf{A} - \lambda\mathbf{I}$  è di rango strettamente inferiore a  $n$ . La matrice è invertibile se e solo se  $0 \notin \text{Sp}(\mathbf{A})$ .

**Definizione A.2** Si dice *polinomio caratteristico della matrice  $\mathbf{A}$*  il seguente polinomio di grado  $n$

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) \quad (\text{A.16})$$

Le radici del polinomio caratteristico sono gli autovalori di  $\mathbf{A}$ . La matrice ha, pertanto,  $n$  autovalori in  $\mathbb{C}$ ; se  $\lambda$  è uno zero di  $p_{\mathbf{A}}$  di molteplicità  $k$ , si dice che l'autovalore  $\lambda$  è di *molteplicità algebrica*  $k$ .

Dalla definizione segue immediatamente che gli autovalori di una matrice  $\mathbf{A}$  diagonale o triangolare sono uguali agli elementi sulla diagonale principale. In effetti, la matrice  $\mathbf{A} - \lambda\mathbf{I}$  è ancora diagonale o triangolare e quindi il suo determinante è dato dal prodotto degli elementi sulla diagonale principale.

Sviluppando  $\det(\mathbf{A} - \lambda\mathbf{I})$ , si ottiene

$$\det(\mathbf{A} - \lambda\mathbf{I}) = a_0\lambda^n + a_1\lambda^{n-1} + \dots + a_n$$

in cui

$$a_0 = (-1)^n, \quad a_i = (-1)^{n-i}\sigma_i, \quad i = 1, \dots, n$$

e  $\sigma_i$  è la somma dei determinanti delle  $\binom{n}{i}$  sottomatrici principali di  $\mathbf{A}$  di ordine  $i$ .

Lasciamo come esercizio dimostrare che il polinomio caratteristico della seguente matrice

$$\begin{bmatrix} p_1 & p_2 & \cdots & p_{n-1} & p_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad (\text{A.17})$$

è dato da

$$(-1)^n[\lambda^n - p_1\lambda^{n-1} - \dots - p_n]$$

La forma (A.17) è nota come *matrice di Frobenius*<sup>6</sup> (o *companion matrix*). È importante osservare che una generica matrice può essere ricondotta alla forma di Frobenius mediante trasformazioni che lasciano inalterati gli autovalori della matrice.

<sup>6</sup>Georg Ferdinand Frobenius (1849-1917).

Tali trasformazioni sono dette trasformazioni per similitudine e saranno analizzate nel paragrafo successivo.

Per terminare, ricordiamo il seguente classico risultato.

**Teorema A.4** (Cayley-Hamilton) *Sia  $\mathbf{A} \in \mathbb{C}^{n \times n}$  e sia  $p_{\mathbf{A}}(\lambda)$  il suo polinomio caratteristico. Allora*

$$p_{\mathbf{A}}(\mathbf{A}) = 0$$

ove lo zero a secondo membro indica la matrice di ordine  $n$  con tutti gli elementi nulli.

Come esemplificazione, si consideri la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

il cui polinomio caratteristico è dato da  $(\lambda - 1)^2(\lambda - 4)$ .

Ricordiamo che viene detto *polinomio minimo* di  $\mathbf{A}$  il polinomio di grado minimo, e con coefficiente di grado massimo uguale a 1, che è annullato da  $\mathbf{A}$ . Nell'esempio precedente si verifica che il corrispondente polinomio di grado minimo è dato da  $\lambda^2 - 5\lambda + 4$ .

### A.3.1 Trasformazioni per similitudine

Sia  $\mathbf{S}$  una matrice *non singolare*. Allora la matrice  $\mathbf{A}' = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$  è detta *simile* alla matrice  $\mathbf{A}$  e la trasformazione corrispondente è detta una *trasformazione per similitudine*. Le trasformazioni per similitudine corrispondono ad un *cambiamento di base* in  $\mathbb{C}^n$ . Abbiamo il seguente importante risultato.

**Proposizione A.5** *Le matrici simili hanno lo stesso polinomio caratteristico.*

**DIMOSTRAZIONE.** Se  $\mathbf{A}'$  è una matrice simile a  $\mathbf{A}$ , si ha

$$\begin{aligned} p_{\mathbf{A}'}(\lambda) &= \det(\mathbf{A}' - \lambda\mathbf{I}) = \det(\mathbf{S}^{-1}\mathbf{A}\mathbf{S} - \lambda\mathbf{S}^{-1}\mathbf{S}) \\ &= \det(\mathbf{S}^{-1}) \det(\mathbf{A} - \lambda\mathbf{I}) \det(\mathbf{S}) \end{aligned}$$

da cui  $p_{\mathbf{A}'}(\lambda) = p_{\mathbf{A}}(\lambda)$ . ■

Gli autovalori sono, pertanto, *indipendenti* dalla base nella quale la matrice è rappresentata; da qui anche il loro interesse<sup>7</sup>.

---

<sup>7</sup>Se  $\mathbf{C}$  è una generica matrice non singolare, gli autovalori di  $\mathbf{A}$  e  $\mathbf{C}^T\mathbf{A}\mathbf{C}$  sono, in generale, differenti. Tuttavia, la trasformazione  $\mathbf{A} \rightarrow \mathbf{C}^T\mathbf{A}\mathbf{C}$ , detta *trasformazione per congruenza* e corrispondente al cambiamento di variabili  $\mathbf{x} = \mathbf{C}\mathbf{y}$ , mantiene la *simmetria*; in effetti, si può mostrare che le due matrici  $\mathbf{A}$  e  $\mathbf{C}^T\mathbf{A}\mathbf{C}$  hanno lo stesso numero di autovalori positivi, lo stesso numero di autovalori negativi e lo stesso numero di autovalori nulli. Tale risultato è noto come *legge di conservazione dei numeri inerziali* (Sylvester's law of inertia). Mentre le trasformazioni simili conservano gli autovalori, le trasformazioni per congruenza conservano solo il segno degli autovalori.

Un autovettore  $\mathbf{v}$  di  $\mathbf{A}'$  si ricava da un autovettore  $\mathbf{u}$  di  $\mathbf{A}$  mediante la trasformazione  $\mathbf{v} = \mathbf{S}^{-1}\mathbf{u}$ .

Lasciamo come esercizio dimostrare che

$$\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i, \quad \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i$$

ove la quantità  $\text{tr}(\mathbf{A}) := \sum_{i=1}^n a_{ii}$  è detta *traccia* di  $\mathbf{A}$ . Come il determinante, anche la traccia è indipendente dalla base.

Nella successiva proposizione sono raccolti alcuni importanti risultati relativi agli autovalori di operazioni su matrici.

**Proposizione A.6** *Se  $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ , allora si hanno i seguenti risultati*

$$(\mathbf{A} - \mu\mathbf{I})\mathbf{u} = (\lambda - \mu)\mathbf{u} \quad (\text{A.18})$$

$$\mathbf{A}^k\mathbf{u} = \lambda^k\mathbf{u} \quad (\text{A.19})$$

$$\mathbf{A}^{-1}\mathbf{u} = \frac{1}{\lambda}\mathbf{u}, \quad \text{se } \mathbf{A} \text{ è invertibile} \quad (\text{A.20})$$

$$P(\mathbf{A})\mathbf{u} = P(\lambda)\mathbf{u}, \quad \text{per ogni polinomio } P \quad (\text{A.21})$$

La dimostrazione è lasciata come esercizio.

Per quanto riguarda gli autovalori del prodotto di due matrici ricordiamo il seguente risultato.

**Teorema A.5** *Siano  $\mathbf{A} \in \mathbb{R}^{n \times m}$  e  $\mathbf{B} \in \mathbb{R}^{m \times n}$ , con  $m > n$ ; allora*

$$\det(\lambda\mathbf{I} - \mathbf{BA}) = \lambda^{m-n} \det(\lambda\mathbf{I} - \mathbf{AB})$$

*Gli autovalori non nulli delle due matrici  $\mathbf{AB}$  e  $\mathbf{BA}$  sono gli stessi.*

Per la dimostrazione si considerano le due uguaglianze

$$\begin{bmatrix} \mathbf{I}_n & 0 \\ -\mathbf{B} & \mu\mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mu\mathbf{I}_n & \mathbf{A} \\ \mathbf{B} & \mu\mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \mu\mathbf{I}_n & \mathbf{A} \\ 0 & \mu^2\mathbf{I}_m - \mathbf{BA} \end{bmatrix}$$

$$\begin{bmatrix} \mu\mathbf{I}_n & -\mathbf{A} \\ 0 & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mu\mathbf{I}_n & \mathbf{A} \\ \mathbf{B} & \mu\mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \mu^2\mathbf{I}_n - \mathbf{AB} & 0 \\ \mathbf{B} & \mu\mathbf{I}_m \end{bmatrix}$$

e si calcola il determinante di ciascuna uguaglianza, ponendo poi  $\mu^2 = \lambda$ .

### A.3.2 Autovettori a sinistra

Dall'uguaglianza

$$\overline{\det(\mathbf{A} - \lambda\mathbf{I})} = \det(\mathbf{A} - \lambda\mathbf{I})^* = \det(\mathbf{A}^* - \bar{\lambda}\mathbf{I})$$

si ottiene che gli autovalori di  $\mathbf{A}^*$  sono i coniugati degli autovalori di  $\mathbf{A}$ .

Gli autovettori di  $\mathbf{A}^*$  verificano, per definizione

$$\mathbf{A}^* \mathbf{v} = \bar{\lambda} \mathbf{v}, \quad \mathbf{v} \neq 0, \quad \text{per } \lambda \in \text{Sp}(\mathbf{A})$$

Prendendo l'aggiunto dei due membri si ha

$$\mathbf{v}^* \mathbf{A} = \lambda \mathbf{v}^*$$

Il vettore  $\mathbf{v}$  è detto *autovettore a sinistra di  $\mathbf{A}$* .

Per una matrice hermitiana si ha che gli autovalori sono reali e gli autovettori rispettivamente a sinistra e a destra coincidono.

**Teorema A.6** *Sia  $\mathbf{u}_i$  un autovettore della matrice  $\mathbf{A}$  corrispondente all'autovalore  $\lambda_i$ . Sia  $\mathbf{v}_j$  un autovettore a sinistra corrispondente all'autovalore  $\lambda_j$ .*

*Allora, se  $\lambda_i \neq \lambda_j$ , si ha*

$$(\mathbf{u}_i, \mathbf{v}_j) = 0 \tag{A.22}$$

La dimostrazione è lasciata come esercizio.

Dal teorema precedente, si ha, in particolare che se  $\mathbf{A}$  è una matrice hermitiana, allora gli autovettori corrispondenti a due autovalori *distinti* sono *ortogonali*.

### A.3.3 Riduzione delle matrici

Una matrice  $\mathbf{A}$  si dice *diagonalizzabile*, se esiste una matrice non singolare  $\mathbf{S}$  tale che la trasformata per similitudine di  $\mathbf{A}$ :  $\mathbf{S}^{-1} \mathbf{A} \mathbf{S}$  è una matrice diagonale  $\mathbf{D}$ . Poiché

$$\mathbf{S}^{-1} \mathbf{A} \mathbf{S} = \mathbf{D} \quad \Leftrightarrow \quad \mathbf{A} \mathbf{S} = \mathbf{S} \mathbf{D}$$

si ha che le colonne della matrice  $\mathbf{S}$  sono autovettori di  $\mathbf{A}$ . Lasciamo come esercizio la ricerca di un esempio di matrice non diagonalizzabile.

Le matrici diagonalizzabili possono essere caratterizzate mediante i loro autovettori nel seguente modo.

**Teorema A.7** *Una matrice  $\mathbf{A}$  di ordine  $n$  è diagonalizzabile se e solo se  $\mathbf{A}$  possiede  $n$  autovettori linearmente indipendenti.*

Ricordando che

**Proposizione A.7** *Gli autovettori associati ad autovalori distinti sono linearmente indipendenti.*

si ricava la seguente condizione *sufficiente*.

**Corollario A.1** *Una matrice è diagonalizzabile, se i suoi autovalori sono distinti.*

Per le matrici non diagonalizzabili ricordiamo il seguente risultato.

**Proposizione A.8** (Forma di Jordan) *Una matrice  $\mathbf{A}$  quadrata di ordine  $n$  è simile a una matrice della forma*

$$\mathbf{J} = \text{diag}(\mathbf{J}_1, \mathbf{J}_2, \dots, \mathbf{J}_p) \quad (\text{A.23})$$

ove ogni  $\mathbf{J}_i$  è una matrice quadrata di ordine  $r_i$  della forma

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ \cdots & \cdots & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & \lambda_i \end{bmatrix}$$

e  $\sum_{i=1}^p r_i = n$ .

La matrice (A.23) è chiamata la *forma canonica di Jordan*. Come caso particolare di (A.23) si ha per  $p = n$  e  $r_i = 1$ ,  $i = 1, \dots, n$  la forma diagonale. All'altro estremo si ha il caso corrispondente a  $p = 1$  e  $r_1 = n$ . Un blocco  $\mathbf{J}_i$  ha  $\lambda_i$  come autovettore di molteplicità algebrica  $r_i$ ; la *molteplicità geometrica* di  $\lambda_i$ , cioè la dimensione dello spazio lineare generato dagli autovettori corrispondenti a  $\lambda_i$  è data dal numero dei blocchi che compaiono nella matrice  $\mathbf{J}$  e relativi allo stesso autovalore  $\lambda_i$ .

Come esercizio lasciamo il calcolo della molteplicità geometrica degli autovalori relativi alle seguenti matrici

$$\mathbf{A} = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}$$

Attraverso gli autovettori di una matrice è possibile caratterizzare la commutatività del prodotto di due matrici nel seguente modo.

**Teorema A.8** *Una condizione necessaria e sufficiente affinché due matrici  $\mathbf{A}$ ,  $\mathbf{B}$  dello stesso ordine e diagonalizzabili, commutino, è che esista una base formata da autovettori per entrambe.*

La dimostrazione è lasciata come esercizio.

### A.3.4 Fattorizzazione unitaria di una matrice

Fra le trasformazioni per similitudine hanno una particolare importanza quelle ottenute mediante una matrice unitaria, o ortogonale. In effetti, il teorema successivo mostra che è possibile mediante una trasformazione unitaria ricondurre una qualsiasi matrice a una forma triangolare superiore.

**Teorema A.9** (Forma normale di Schur) *Data una matrice  $\mathbf{A} \in \mathbb{C}^{n \times n}$ , esiste una matrice unitaria  $\mathbf{U}$  e una matrice triangolare superiore  $\mathbf{T}$  tali che*

$$\mathbf{A} = \mathbf{UTU}^* \quad (\text{A.24})$$

*Gli elementi sulla diagonale principale di  $\mathbf{T}$  sono gli autovalori della matrice  $\mathbf{A}$ .*

La dimostrazione può essere ottenuta per induzione (cfr. ad esempio Lancaster e Tismenetsky [105]).

Attraverso le trasformazioni unitarie è possibile caratterizzare le matrici normali nel seguente modo. Ricordiamo che una matrice  $\mathbf{A}$  è detta normale se  $\mathbf{A}^* \mathbf{A} = \mathbf{A} \mathbf{A}^*$ . Le matrici hermitiane, unitarie, e antihermitiane ( $\mathbf{A}^* = -\mathbf{A}$ ) sono esempi di matrici normali.

**Teorema A.10** *Una matrice  $\mathbf{A} \in \mathbb{C}^{n \times n}$  è normale se e solo se esiste una matrice unitaria  $\mathbf{U}$  tale che*

$$\mathbf{A} = \mathbf{UDU}^* \quad (\text{A.25})$$

*ove  $\mathbf{D}$  è la matrice diagonale formata dagli autovalori della matrice. In particolare, una matrice normale è diagonalizzabile ed i suoi autovettori corrispondenti ad autovalori distinti sono ortogonali.*

Nel caso in cui la matrice  $\mathbf{A}$  è normale e ad elementi reali, il teorema precedente si estende nel senso che la matrice  $\mathbf{U}$  è ortogonale e  $\mathbf{D}$  è una matrice diagonale a blocchi di ordine 1, o 2, in corrispondenza, rispettivamente ad autovalori reali e ad autovalori complessi coniugati. Nel caso particolare di matrici simmetriche la  $\mathbf{D}$  è una matrice diagonale.

Come corollario si hanno i seguenti risultati

- Una matrice hermitiana è diagonalizzabile. I suoi autovalori sono reali e gli autovettori corrispondenti ad autovalori distinti sono ortogonali.
- Una matrice antihermitiana è diagonalizzabile. I suoi autovalori sono immaginari puri e gli autovettori corrispondenti ad autovalori distinti sono ortogonali.
- Una matrice unitaria è diagonalizzabile. I suoi autovalori hanno modulo 1. Gli autovettori corrispondenti ad autovalori distinti sono ortogonali.
- Una matrice hermitiana è definita positiva (semidefinita positiva) se e solo se tutti i suoi autovalori sono strettamente positivi (nonnegativi).

Dai risultati precedenti si ricava, in particolare, una utile interpretazione geometrica delle matrici definite positive. Sia  $\mathbf{A}$  una matrice simmetrica definita positiva e  $\mathbf{Q}$  la matrice le cui colonne sono date dagli autovettori di  $\mathbf{A}$  normalizzati a 1. Si ha,



quindi,  $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ , con  $\mathbf{\Lambda}$  matrice diagonale degli autovalori. Allora, la rotazione  $\mathbf{y} = \mathbf{Q}^T\mathbf{x}$  produce la somma di quadrati

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{x} = \mathbf{y}^T\mathbf{\Lambda}\mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2$$

Si vede, quindi che l'equazione

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = 1$$

descrive un ellissoide con centro nell'origine e i cui assi sono nelle direzioni degli autovettori. I semiassi hanno lunghezza  $1/\sqrt{\lambda_j}$ . Come illustrazione, in Figura A.7 è rappresentata l'ellisse corrispondente alla seguente matrice definita positiva

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

per la quale si ha  $\mathbf{x}^T\mathbf{A}\mathbf{x} = 2x_1^2 - 2x_1x_2 + 2x_2^2 = 1$ . Gli autovalori sono dati da  $\lambda_1 = 1$  e  $\lambda_2 = 3$ , mentre gli autovettori hanno la direzione delle bisettrici.

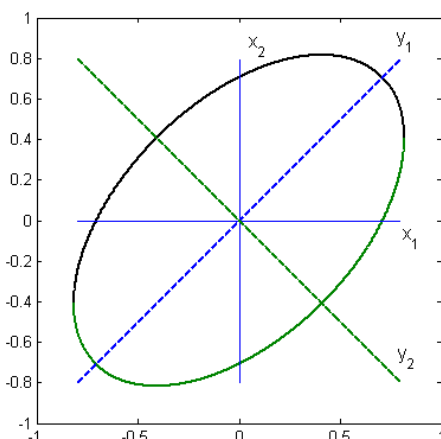


Figura A.7: Rappresentazione dell'ellisse  $\mathbf{x}^T\mathbf{A}\mathbf{x} = 2x_1^2 - 2x_1x_2 + 2x_2^2 = 1$ .

Se gli autovalori  $\lambda_i$  di una matrice hermitiana sono disposti in ordine decrescente

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

si può dimostrare facilmente il seguente risultato, noto anche come teorema di Courant–Fischer

$$\lambda_1 = \max_{0 \neq \mathbf{x} \in \mathbb{C}^n} \frac{\mathbf{x}^* \mathbf{A} \mathbf{x}}{\mathbf{x}^* \mathbf{x}}, \quad \lambda_n = \min_{0 \neq \mathbf{x} \in \mathbb{C}^n} \frac{\mathbf{x}^* \mathbf{A} \mathbf{x}}{\mathbf{x}^* \mathbf{x}} \quad (\text{A.26})$$

## A.4 Localizzazione degli autovalori

Il problema della localizzazione degli autovalori riguarda la determinazione di regioni del piano complesso contenenti gli autovalori di una matrice data. Risultati di questo tipo sono utili sia per il calcolo numerico degli autovalori che per evidenziare le proprietà di alcune matrici.

Ricordiamo a questo proposito due risultati, noti come teoremi di Gershgorin e Hadamard.

**Teorema A.11** (Primo teorema di Gershgorin–Hadamard) *Sia  $\mathbf{A}$  una matrice quadrata di ordine  $n$ . Per ogni  $k = 1, 2, \dots, n$ , consideriamo il disco  $D_k$  del piano complesso definito da*

$$|z - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}| \quad (\text{A.27})$$

*Gli autovalori di  $\mathbf{A}$  appartengono all'unione degli  $n$  dischi  $D_k$ , cioè*

$$\lambda \in \bigcup_{k=1}^n D_k$$

**DIMOSTRAZIONE.** Sia  $\lambda$  un autovalore e  $\mathbf{u}$  un autovalore associato tale che

$$\max_{1 \leq i \leq n} |u_i| = |u_k| = 1$$

La componente  $k$ -ma dell'uguaglianza  $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$  si scrive

$$(\lambda - a_{kk})u_k = \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}|$$

Poiché si ha  $|u_j| \leq |u_k| = 1$ , prendendo il modulo nell'uguaglianza precedente si ottiene

$$|\lambda - a_{kk}| = \left| \sum_{\substack{j=1 \\ j \neq k}}^n a_{kj} u_j \right| \leq \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}| |u_j| \leq \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}|$$

Quindi  $\lambda \in D_k$  e da questo il risultato del teorema. ■

**Corollario A.2** *Se l'unione  $M_1$  di  $k$  cerchi di Gershgorin è disgiunta dall'unione  $M_2$  dei rimanenti  $n - k$ , allora  $k$  autovalori appartengono a  $M_1$  e  $n - k$  autovalori appartengono a  $M_2$ .*

Per la dimostrazione si considera la famiglia di matrici

$$\mathbf{A}(t) = \mathbf{D} + t(\mathbf{A} - \mathbf{D}), \quad 0 \leq t \leq 1$$

ove  $\mathbf{D} = \mathbf{diag}(a_{ii})$ . Si applica quindi il teorema precedente su  $\mathbf{A}(0)$ , utilizzando poi un ragionamento di continuità.

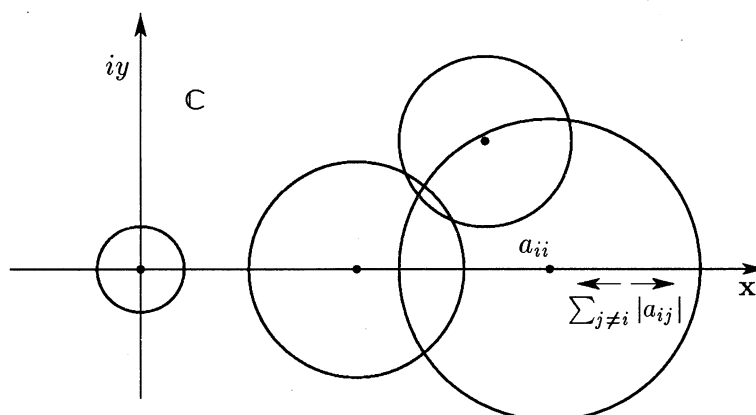


Figura A.8: Esempi di cerchi di Gershgorin.

▼ **Osservazione A.7** Poiché gli autovalori di  $\mathbf{A}^T$  sono i medesimi della matrice  $\mathbf{A}$ , applicando il teorema precedente alla matrice  $\mathbf{A}^T$  si ottiene una nuova regione. Gli autovalori di  $\mathbf{A}$  appartengono alla intersezione delle due regioni. ■

Il secondo risultato, la cui dimostrazione è leggermente più impegnativa è il seguente.

**Teorema A.12** (Secondo teorema di Gershgorin–Hadamard) Sia  $\mathbf{A}$  una matrice di ordine  $n$ , irriducibile. Se un autovalore  $\lambda$  è situato sulla frontiera della riunione dei dischi, allora tutti i cerchi  $D_k$  passano per  $\lambda$ .

#### A.4.1 Norma di vettore e di matrice

Una applicazione  $\mathbb{R}^n \rightarrow \mathbb{R}_+$ , è chiamata *norma*, indicata usualmente con  $\|\mathbf{x}\|$ , quando verifica le seguenti condizioni

- (i)  $\|\mathbf{x}\| \geq 0$ , per ogni  $\mathbf{x} \in \mathbb{R}^n$ .
- (ii)  $\|\mathbf{x}\| = 0$  se e solo se  $\mathbf{x} = 0$ .
- (iii)  $\|\alpha\mathbf{x}\| = |\alpha| \|\mathbf{x}\|$ , per ogni  $\mathbf{x} \in \mathbb{R}^n$ ,  $\alpha \in \mathbb{R}$ .
- (iv) Per ogni  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , vale la *disuguaglianza triangolare*

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Il numero  $\|\mathbf{x} - \mathbf{y}\|$  definisce allora una *distanza* tra i punti  $\mathbf{x}$  e  $\mathbf{y}$ .

Un esempio di norma in  $\mathbb{R}^n$  è fornito dalla cosiddetta *norma p*, definita, per  $1 \leq p < \infty$ , nel modo seguente

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} \quad (\text{A.28})$$

Per  $p = 2$  si ha la usuale *norma euclidea*; per  $p = 1$  la norma corrispondente è anche nota come *norma di Manhattan*.

Nel caso  $p = \infty$  si ha la *norma del massimo*, detta anche *norma di Chebichev*, definita da

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| \quad (\text{A.29})$$

Ad esempio, se  $\mathbf{x} = [1, -2]^T$  si ha

$$\|\mathbf{x}\|_1 = 3, \quad \|\mathbf{x}\|_2 = \sqrt{5}, \quad \|\mathbf{x}\|_\infty = 2$$

Per  $p = 1$  e  $p = \infty$ , non ci sono problemi a mostrare che le proprietà richieste dalla definizione di di norma sono verificate. Per  $1 < p < \infty$  la proprietà triangolare diventa

$$\left( \sum_{i=1}^n |x_i + y_i|^p \right)^{1/p} \leq \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} + \left( \sum_{i=1}^n |y_i|^p \right)^{1/p}$$

nota anche come *disuguaglianza di Minkowski*.

Se  $\mathbf{A}$  è una matrice simmetrica di ordine  $n$  e *definita positiva*, si può definire una norma di vettore ponendo

$$\|\mathbf{x}\|_{\mathbf{A}} := \left( \sum_{i,j} a_{ij} x_i x_j \right)^{1/2}$$

In Figura A.9 sono rappresentate le sfere unitarie in  $\mathbb{R}^2$  corrispondenti a differenti tipi di norme.

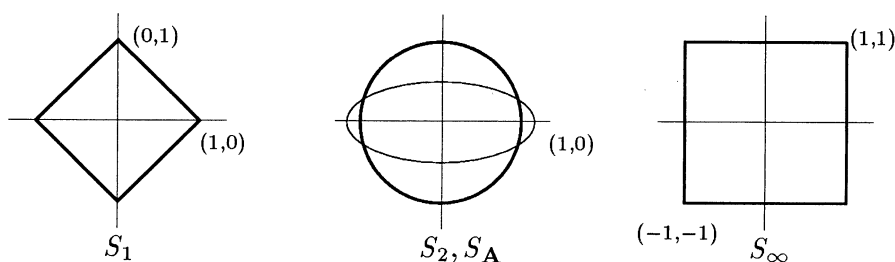


Figura A.9: Le sfere unitarie  $S_p$  e  $S_{\mathbf{A}}$  relative alla norma  $\|\cdot\|_p$  e alla norma  $\|\cdot\|_{\mathbf{A}}$ .

Osserviamo che una norma di vettore è una funzione continua in  $\mathbb{R}^n$  e che per ogni coppia di norme di vettore,  $\|\mathbf{x}\|$ ,  $\|\mathbf{x}\|'$ , esistono due costanti positive  $m$  e  $M$  tali che per ogni  $\mathbf{x} \in \mathbb{R}^n$

$$m\|\mathbf{x}\|' \leq \|\mathbf{x}\| \leq M\|\mathbf{x}\|'$$

In altre parole, in  $\mathbb{R}^n$ , per  $n$  fissato, le norme sono tra di loro equivalenti.

### Norma di matrice

Una *matrice* quadrata di ordine  $n$  può essere considerata un vettore in uno spazio di dimensione  $n^2$  (avendo fissata una convenzione relativamente all'ordine degli elementi). Allora, per definire una *norma di matrice*, potremmo utilizzare la definizione data per un vettore. Tuttavia, per le applicazioni conviene restringere ulteriormente la definizione. In particolare, date due matrici quadrate di ordine  $n$   $\mathbf{A}$  e  $\mathbf{B}$ , è utile porre nella definizione la seguente condizione

$$(v) \quad \|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$$

È da sottolineare che non tutte le norme di vettore verificano la condizione (v); si consideri, ad esempio

$$\mathbf{A} = \mathbf{B} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{C} = \mathbf{AB} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

Si ha

$$\max_{i,j} |c_{i,j}| = 2 \quad \max_{i,j} |a_{i,j}| = \max_{i,j} |b_{i,j}| = 1$$

Per riassumere, possiamo chiamare *norma di matrice* un'applicazione:  $\mathbf{A} \rightarrow \|\mathbf{A}\|$ , che verifica condizioni analoghe alle condizioni (i), (ii), (iii), (iv) date nella definizione di norma di vettore, con l'aggiunta della precedente condizione (v). Un modo naturale, geometrico, di definire una norma di matrice, che verifica le condizioni precedenti, è il seguente.

Sia  $\|\cdot\|$  una norma fissata di vettore. Definiamo *norma naturale* (o norma indotta dalla norma di vettore) della matrice  $\mathbf{A}$  la quantità

$$\|\mathbf{A}\| \equiv \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \quad (\text{A.30})$$

Poiché per ogni  $\mathbf{x} \neq 0$  si può definire  $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$ , sicché  $\|\mathbf{u}\| = 1$ , la definizione (A.30) è equivalente alla seguente

$$\|\mathbf{A}\| = \max_{\|\mathbf{u}\|=1} \|\mathbf{Au}\| = \|\mathbf{Ay}\|, \quad \|\mathbf{y}\| = 1$$

Per definizione se  $\mathbf{I}$  è la matrice identità e  $\|\cdot\|$  è una norma naturale, allora  $\|\mathbf{I}\| = 1$ .

Osserviamo che per una norma definita come in (A.30) si ha il seguente risultato

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \|\mathbf{x}\| \quad (\text{A.31})$$

Più in generale, quando una norma di matrice verifica la condizione (A.31) si dice che essa è *consistente* (o compatibile) con la corrispondente norma di vettore. La norma naturale è, in sostanza, la *più piccola* norma consistente con una assegnata norma di vettore.

Lasciamo come esercizio mostrare che la definizione (A.30) verifica le proprietà (i)–(v).

Vediamo, ora, quali sono le norme naturali di matrice che corrispondono alle norme  $p$  di vettore per  $p = 1, 2, \infty$ . Indichiamo con  $a_{ij}$ ,  $i, j = 1, 2, \dots, n$  gli elementi della matrice  $\mathbf{A}$ .

**Proposizione A.9** *La norma di matrice indotta dalla norma del massimo ( $p = \infty$ ) è la seguente*

$$\|\mathbf{A}\|_{\infty} = \max_i \sum_{j=1}^n |a_{ij}| \quad (\text{A.32})$$

*cioè la massima tra le somme dei moduli delle righe.*

In maniera analoga, si dimostra che

$$\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

cioè  $\|\mathbf{A}\|_1 = \|\mathbf{A}^T\|_{\infty}$ .

Consideriamo ora la norma 2, corrispondente alla norma euclidea di vettore. Si ha il seguente risultato.

**Proposizione A.10** *La norma 2 di matrice, corrispondente alla norma euclidea di vettore, può essere calcolata nel modo seguente*

$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^T \mathbf{A})} \quad (\text{A.33})$$

*Per tale motivo la norma  $\|\mathbf{A}\|_2$  è nota anche come norma spettrale. Nel caso particolare di una matrice simmetrica la norma spettrale coincide con il raggio spettrale della matrice.*

Fra le norme che sono state introdotte si hanno le seguenti relazioni

$$\begin{aligned} \frac{1}{\sqrt{n}} \|\mathbf{A}\|_{\infty} &\leq \|\mathbf{A}\|_2 \leq \sqrt{n} \|\mathbf{A}\|_{\infty} \\ \frac{1}{\sqrt{n}} \|\mathbf{A}\|_1 &\leq \|\mathbf{A}\|_2 \leq \sqrt{n} \|\mathbf{A}\|_1 \\ \max_{i,j} |a_{ij}| &\leq \|\mathbf{A}\|_2 \leq n \max_{i,j} |a_{ij}| \\ \|\mathbf{A}\|_2 &\leq \sqrt{\|\mathbf{A}\|_1 \|\mathbf{A}\|_{\infty}} \end{aligned}$$

Una norma di matrice che non è subordinata ad una norma di vettore è la norma di *Frobenius* (o di Schur<sup>8</sup>), definita per una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$  nel modo seguente

$$\|\mathbf{A}\|_F = \left( \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 \right)^{1/2} = [\text{tr}(\mathbf{A}^T \mathbf{A})]^{1/2}$$

La norma di Frobenius è essenzialmente la norma euclidea della matrice considerata come un vettore di  $mn$  componenti. È interessante osservare che la norma di Frobenius è compatibile con la norma di vettore euclidea; si ha, infatti

$$\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_F \|\mathbf{x}\|_2$$

### Matrici convergenti

Per studiare la convergenza di procedure iterative (cfr. in particolare i Capitoli 2 e 5) è utile stabilire quando per una matrice  $\mathbf{A}$  si ha la convergenza a zero delle successive potenze, cioè

$$\lim_{m \rightarrow \infty} \mathbf{A}^m = 0 \quad (\text{A.34})$$

In questo caso si dice che la matrice è *convergente*. Per stabilire la convergenza di una matrice si hanno le seguenti condizioni.

**Teorema A.13** *I seguenti risultati sono equivalenti.*

- (a)  $\mathbf{A}$  è convergente
- (b)  $\lim_{m \rightarrow \infty} \|\mathbf{A}^m\| = 0$ , per una norma di matrice
- (c)  $\rho(\mathbf{A}) < 1$

Una *condizione sufficiente*, in generale non necessaria, ma importante nelle applicazioni, è la seguente.

**Corollario A.1**  $\mathbf{A}$  è convergente se per una particolare norma di matrice si ha

$$\|\mathbf{A}\| < 1$$

## A.5 I valori singolari e la pseudoinversa

Per una matrice rettangolare la nozione di autovalore perde di significato. Una nozione più generale e che presenta interesse numerico in relazione al *rango* di una matrice e al suo *condizionamento* è la nozione di *valori singolari*. Dal punto di vista

<sup>8</sup>Friedrich Heinrich Schur (1856-1932).

teorico è pure interessante una estensione del concetto di inversa di una matrice, la cosiddetta *pseudoinversa*, che può anche essere definita a partire dai valori singolari.

In questo paragrafo richiameremo le idee essenziali relativamente a questi due concetti, che hanno assunto dal punto di vista numerico un'importanza notevole, in particolare per quanto riguarda la risoluzione dei problemi *malcondizionati*. Per un approfondimento si rinvia ad esempio Golub e Van Loan [69].

### A.5.1 Decomposizione in valori singolari SVD

Data una matrice arbitraria  $\mathbf{A} \in \mathbb{C}^{m \times n}$ , la matrice  $\mathbf{A}^* \mathbf{A}$  è una matrice hermitiana *semidefinita positiva* poiché  $\mathbf{x}^*(\mathbf{A}^* \mathbf{A})\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0$ , per ogni  $\mathbf{x} \in \mathbb{C}^n$ . I suoi autovalori  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$  possono pertanto essere espressi nella forma  $\lambda_k = \sigma_k^2$ , con  $\sigma_k \geq 0$ . I numeri  $\sigma_1 \geq \dots \geq \sigma_n \geq 0$  sono chiamati i *valori singolari di A*.

Sostituendo in (A.26) alla matrice  $\mathbf{A}$  la matrice  $\mathbf{A}^* \mathbf{A}$ , si ottiene

$$\sigma_1 = \max_{\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} =: \|\mathbf{A}\|_2, \quad \sigma_n = \min_{\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \quad (\text{A.35})$$

In particolare, se  $m = n$  e  $\mathbf{A}$  è non singolare, si ha

$$\frac{1}{\sigma_n} = \|\mathbf{A}^{-1}\|_2, \quad \boxed{\mu_2(\mathbf{A}) = \frac{\sigma_1}{\sigma_n}} \quad (\text{A.36})$$

Ricordiamo anche la seguente relazione

$$\|\mathbf{A}\|_E^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_p^2, \quad (\text{Norma di Frobenius})$$

ove  $p = \min(m, n)$ . Il seguente teorema fornisce una interessante interpretazione del più piccolo valore singolare  $\sigma_n$  di una matrice quadrata  $\mathbf{A}$ ; tale valore è, in sostanza, la distanza di  $\mathbf{A}$  dalla matrice singolare *più vicina*.

**Teorema A.14** *Siano  $\mathbf{A}$  e  $\mathbf{E}$  due matrici  $\in \mathbb{R}^{n \times n}$  e siano  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$  i valori singolari della matrice  $\mathbf{A}$ . Allora*

- (i)  $\|\mathbf{E}\|_2 \geq \sigma_n$  se  $\mathbf{A} + \mathbf{E}$  è singolare;
- (ii) esiste una matrice  $\mathbf{E}$  con  $\|\mathbf{E}\|_2 = \sigma_n$ , tale che  $\mathbf{A} + \mathbf{E}$  è singolare.

**DIMOSTRAZIONE.** Se  $\mathbf{A} + \mathbf{E}$  è singolare, si ha un vettore  $\mathbf{x} \neq 0$  tale che  $(\mathbf{A} + \mathbf{E})\mathbf{x} = 0$ . Allora da (A.35) si ha

$$\sigma_n \|\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}\|_2 = \|-\mathbf{E}\mathbf{x}\|_2 \leq \|\mathbf{E}\|_2 \|\mathbf{x}\|_2$$



da cui  $\sigma_n \leq \|\mathbf{E}\|_2$ . La proprietà (ii) è evidente se  $\sigma_n = 0$ , dal momento che in questo caso la matrice è già singolare. Supponiamo quindi  $\sigma_n > 0$ . Da (A.35) si ha che esistono due vettori  $\mathbf{u}, \mathbf{v}$  tali che

$$\begin{aligned} \|\mathbf{A}\mathbf{u}\|_2 &= \sigma_n, & \|\mathbf{u}\|_2 &= 1 \\ \mathbf{v} &:= \frac{1}{\sigma_n} \mathbf{A}\mathbf{u}, & \|\mathbf{v}\|_2 &= 1 \end{aligned}$$

Definendo la matrice  $\mathbf{E} = -\sigma_n \mathbf{v}\mathbf{u}^*$ , si ha  $(\mathbf{A} + \mathbf{E})\mathbf{u} = 0$ , cioè la matrice  $\mathbf{A} + \mathbf{E}$  è singolare; inoltre

$$\|\mathbf{E}\|_2 = \sigma_n \max_{\mathbf{x} \neq 0} \|\mathbf{v}\|_2 \frac{|\mathbf{u}^* \mathbf{x}|}{\|\mathbf{x}\|_2} = \sigma_n$$

■

▼ **Osservazione A.8** Si consideri ad esempio la seguente matrice

$$\begin{bmatrix} 1/\epsilon & 0 \\ 0 & \epsilon \end{bmatrix}$$

Il determinante è uguale a 1, mentre  $\sigma_1 = 1/\epsilon, \sigma_2 = \epsilon$ . Per  $\epsilon \rightarrow 0$  il rango della matrice tende a 1; come si vede  $\sigma_2$  è una misura conveniente della deficienza del rango. Un caso più generale è dato dal seguente esempio di Wilkinson.

Si consideri la matrice triangolare superiore  $\mathbf{B}$  di ordine  $n$  definita da

$$b_{ij} = \begin{cases} 1 & \text{se } i = j \\ -1 & \text{se } i < j \\ 0 & \text{se } i > j \end{cases}$$

La matrice ha rango  $n$ . Se, tuttavia l'elemento di indici  $(n, 1)$  viene perturbato della quantità  $\epsilon = -2^{2-n}$ , la matrice diventa di rango  $n - 1$ . Il valore singolare  $\sigma_n$  di  $\mathbf{B}$  assume il valore  $0.917 \dots 10^{-4}$  per  $n = 15$  e  $0.296 \cdot 10^{-5}$  per  $n = 20$ . Al crescere di  $n$  la matrice  $\mathbf{B}$  è sempre più vicina ad una matrice di rango  $n - 1$ . Tale comportamento non si vede dagli elementi sulla diagonale principale, e quindi anche dal determinante, che vale sempre 1. Osserviamo anche che il numero di condizionamento  $\mu_2(\mathbf{A}) = \sigma_1/\sigma_n$ , calcolato, ad esempio, per  $n = 20$ , ha il valore  $\approx 3.99 \cdot 10^6$ . ■

Una generica matrice  $\mathbf{A} \in \mathbb{C}^{m \times n}$  può essere trasformata con matrici unitarie in una forma *diagonale* formata dai valori singolari.

**Teorema A.15 (SVD)** Sia  $\mathbf{A}$  una matrice arbitraria  $\in \mathbb{C}^{m \times n}$ . Allora esiste una matrice unitaria (ortogonale se  $\mathbf{A}$  è reale)  $\mathbf{U} \in \mathbb{C}^{m \times m}$  (rispettivamente,  $\in \mathbb{R}^{m \times m}$ ) e una matrice unitaria (ortogonale se  $\mathbf{A}$  è reale)  $\mathbf{V} \in \mathbb{C}^{n \times n}$  (rispettivamente,  $\in \mathbb{R}^{n \times n}$ ) tali che

$$\boxed{\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*} \quad (\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \text{ se } \mathbf{A} \text{ è reale}) \quad (\text{A.37})$$

con

$$\mathbf{\Sigma} = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{D} = \mathbf{diag}(\sigma_1, \sigma_2, \dots, \sigma_r), \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0.$$

Il numero  $r$  dei valori singolari diversi dallo zero è il rango della matrice  $\mathbf{A}$ .

La decomposizione (A.37) viene chiamata la SVD (singular value decomposition) della matrice  $\mathbf{A}^9$ . Il teorema può essere dimostrato per induzione su  $m$  e  $n$ ; si veda ad esempio Golub e Van Loan [69].

Le matrici  $\mathbf{U}$ ,  $\mathbf{V}$ , hanno il seguente significato. Le colonne di  $\mathbf{U}$  (rispettivamente di  $\mathbf{V}$ ) rappresentano gli autovettori della matrice  $\mathbf{A}\mathbf{A}^*$  (rispettivamente di  $\mathbf{A}^*\mathbf{A}$ ).

L'interpretazione geometrica dei valori singolari è la seguente. Supponendo ad esempio che  $\mathbf{A}$  sia reale, si ha che essa trasforma la sfera unitaria in  $\mathbb{R}^n$  (definita da  $\|\mathbf{x}\|_2 = 1$ ), nell'insieme dei vettori  $\mathbf{y} = \mathbf{A}\mathbf{x}$ , che definiscono un ellissoide di dimensione  $r$  nello spazio  $\mathbb{R}^m$ . I valori singolari sono le lunghezze degli assi dell'ellissoide. Il *numero di condizionamento* è legato all'*eccentricità* dell'ellissoide. In termini di trasformazioni lineari, la SVD si interpreta nel modo seguente. Per ogni trasformazione lineare  $\mathbf{A} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , esistono due basi ortonormali  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  in  $\mathbb{R}^n$  e  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$  in  $\mathbb{R}^m$  tali che  $\mathbf{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i$  per  $i = 1, 2, \dots, r$  e  $\mathbf{A}\mathbf{v}_i = 0$  per  $i = r + 1, \dots, n$ .

◆ **Esercizio A.17** Si dimostri che i valori singolari di  $\mathbf{A}$  e di  $\mathbf{A}^*$  sono uguali.

◆ **Esercizio A.18** Si dimostri che se  $\mathbf{A}$  è una matrice  $\in \mathbb{C}^{n \times n}$  allora

$$|\det(\mathbf{A})| = \prod_{i=1}^n \sigma_i$$

Un risultato interessante che generalizza il Teorema A.14, è il seguente.

**Teorema A.16** Sia  $\mathbf{A}$  una matrice  $\in \mathbb{C}^{m \times n}$  e sia

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$$

la sua decomposizione in valori singolari, ove

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_p = 0, \quad p = \min(m, n)$$

e sia  $k$  un intero  $\leq r = \text{rank}(\mathbf{A})$ . Posto

$$\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^*$$

si ha

$$\min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2 = \|\mathbf{A} - \mathbf{A}_k\|_2 = \sigma_{k+1}$$

<sup>9</sup>La decomposizione ai valori singolari è stata introdotta indipendentemente da Beltrami (1873) e da Jordan (1874) nel caso di matrici quadrate. La tecnica fu estesa alle matrici rettangolari negli anni '30 da Eckart e Young. Come tecnica numerica il suo utilizzo risale agli anni '60. Il termine *valore singolare* pare sia stato introdotto da Weyl (1949) nello studio di alcune relazioni fra gli autovalori e i valori singolari di una matrice.

Il teorema, per la cui dimostrazione si veda ad esempio Golub e Van Loan [69], dice in sostanza che il più piccolo valore singolare positivo di  $\mathbf{A}$  è la *distanza*, nella norma 2, di  $\mathbf{A}$  dall'insieme di tutte le matrici a rango deficiente. Esso permette, quindi, di stimare l'errore che si commette quando la matrice  $\mathbf{A}$ , a seguito di approssimazioni, viene sostituita con una matrice di rango  $k$ .

► **Esempio A.1** Supponiamo che gli elementi della seguente matrice

$$\mathbf{A} = \begin{bmatrix} 1.02 & 2.03 & 4.20 \\ 0.25 & 0.51 & 1.06 \\ 1.74 & 3.46 & 7.17 \end{bmatrix}$$

corrispondino ai valori di misurazioni sperimentali soggette ad errori di grandezza, ad esempio, inferiori o uguali a 0.015. La matrice  $\mathbf{A}$  ha la seguente decomposizione in valori singolari

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{bmatrix} 0.5015 & -0.1871 & -0.8447 \\ 0.1263 & 0.9817 & -0.1424 \\ 0.8559 & -0.0352 & 0.5160 \end{bmatrix} \begin{bmatrix} 9.5213 & 0 & 0 \\ 0 & 0.0071 & 0 \\ 0 & 0 & 0.0023 \end{bmatrix} \begin{bmatrix} 0.2135 & -0.9422 & 0.2582 \\ 0.4247 & -0.1485 & -0.8931 \\ 0.8798 & 0.3003 & 0.3685 \end{bmatrix}$$

Dal punto di vista teorico, la matrice  $\mathbf{A}$  ha rango 3; comunque, per il Teorema A.16, vi è una matrice di rango 2 che dista, nella norma 2, solo di 0.0023 e una matrice di rango 1 che dista solo di 0.0071, e gli elementi di tali matrici sono entro gli errori sperimentali relativi alla matrice  $\mathbf{A}$ . In conclusione, possiamo soltanto dire che il rango di  $\mathbf{A}$  è almeno 1, ed inoltre è ragionevole *sostituire*  $\mathbf{A}$  con la matrice di rango 1 più vicina.

$$\mathbf{A}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T = \begin{bmatrix} 1.0193 & 2.0280 & 4.2011 \\ 0.2567 & 0.5107 & 1.0580 \\ 1.7395 & 3.4610 & 7.1696 \end{bmatrix}, \quad \mathbf{A} - \mathbf{A}_1 = \begin{bmatrix} 0.0007 & 0.0020 & -0.0011 \\ -0.0067 & -0.0007 & 0.0020 \\ 0.0005 & -0.0010 & 0.0004 \end{bmatrix}$$

■

La capacità della decomposizione in valori singolari di fornire informazioni su come ottenere approssimazioni, di rango inferiore, di una matrice assegnata, è utile in molteplici applicazioni. Segnaliamo, ad esempio, il suo utilizzo negli algoritmi di *compressione* dei dati, in particolare nella codifica di immagini (*trasformata di Karhunen-Loeve*, cfr. ad esempio Hall [75]). Come esemplificazione e in termini schematici, supponiamo che  $\mathbf{B}$  sia la matrice che definisce i livelli di grigio (*blackness matrix*)<sup>10</sup>. Mediante la decomposizione in valori singolari si vede se una approssimazione di rango inferiore

$$\mathbf{B}_r = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$$

<sup>10</sup>Ogni foto può essere *discretizzata* decomponendo l'immagine in quadretti e assegnando un livello di grigio ad ogni quadrettino. Imponendo, ad esempio, una griglia di  $1000 \times 1000$  su una foto e assegnando un livello di grigio da 0 a 10, si ha una matrice blackness di 1000000 interi per ogni foto.

può rappresentare adeguatamente l'immagine. In caso affermativo, la matrice  $\mathbf{B}_r$  può essere codificata mediante  $2r$  vettori  $\mathbf{u}_i$  e  $\mathbf{v}_i$  e  $r$  numeri  $\sigma_i$ . Se ad esempio è adeguato  $r = 5$ , per una matrice  $\mathbf{B}$  di dimensioni  $1000 \times 1000$  sarà sufficiente memorizzare  $2 \times 5 \times 1000 + 5 = 10\,005$  valori anziché  $1\,000\,000$ , con un risparmio di quasi il 99%.

**▼ Osservazione A.9** Come si è visto i quadrati dei valori singolari sono, in sostanza, gli autovalori delle matrici  $\mathbf{A}^*\mathbf{A}$ ,  $\mathbf{A}\mathbf{A}^*$ . Nel caso particolare in cui la matrice  $\mathbf{A}$  è hermitiana e definita positiva, i valori singolari sono anche gli autovalori e gli autovettori sono dati dalle colonne di  $\mathbf{V}$ . In generale, comunque, non vi è una relazione semplice tra gli autovalori e i valori singolari.

In teoria, si potrebbe pensare di calcolare i valori singolari risolvendo il problema degli autovalori per la matrice hermitiana  $\mathbf{A}^*\mathbf{A}$ , ma questa procedura può portare a una perdita di accuratezza, come è illustrato dal seguente esempio.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}, \quad |\epsilon| < \sqrt{\text{eps}}, \quad \text{eps} = \text{precisione macchina}$$

Come prodotto  $\mathbf{A}^*\mathbf{A}$  si ha

$$\mathbf{A}^*\mathbf{A} = \begin{bmatrix} 1 + \epsilon^2 & 1 \\ 1 & 1 + \epsilon^2 \end{bmatrix}$$

che ha come valori singolari:  $\sigma_1 = \sqrt{2 + \epsilon^2}$ ,  $\sigma_2 = |\epsilon|$ . Nell'aritmetica floating-point, con precisione eps, invece di  $\mathbf{A}^*\mathbf{A}$  si ha

$$\text{fl}(\mathbf{A}^*\mathbf{A}) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

i cui autovalori sono  $\lambda_1 = 2$ ,  $\lambda_2 = 0$ , e  $\sigma_2$  non concorda, nella precisione macchina usata, con  $\sqrt{\lambda_2}$ .

Una procedura numerica più stabile consiste nel ridurre, mediante trasformazioni ortogonali di tipo Householder, la matrice  $\mathbf{A}$  ad una matrice bidiagonale e nell'applicare a questa il metodo QR (algoritmo di Golub e Reinsch) (cfr. Capitolo 3). ■

## A.5.2 Risultati di perturbazione per i valori singolari

Il problema del calcolo degli autovalori è ben condizionato quando la matrice è una matrice normale (cfr. Capitolo 3). Se la matrice non è tale, "piccole perturbazioni" nei dati della matrice possono indurre "grandi" variazioni negli autovalori, con conseguente difficoltà di calcolo. Al contrario il problema dei valori singolari è sempre ben condizionato. Si ha infatti il seguente risultato, per il quale rinviamo ad esempio a Golub e Van Loan [69].

**Teorema A.17** Siano  $\mathbf{A}$  e  $\delta\mathbf{A}$ , matrici  $\in \mathbb{C}^{m \times n}$ . Se  $\sigma_i, \tau_i, \psi_i, i = 1, 2, \dots, n$ , sono i valori singolari di  $\mathbf{A}$ , di  $\delta\mathbf{A}$  e di  $\mathbf{A} + \delta\mathbf{A}$ , risulta

$$|\psi_i - \sigma_i| \leq \tau_i = \|\delta\mathbf{A}\|_2, \quad i = 1, 2, \dots, n$$

### A.5.3 Applicazioni della SVD

Nel seguito, per semplificare le notazioni, supporremo le dimensioni della matrice tali che  $m \geq n$ ; questo del resto corrisponde alla maggior parte delle applicazioni della SVD all'analisi dei dati sperimentali, nelle quali  $a_{ij}$  rappresenta la  $i$ -ma osservazione della variabile  $j$ -ma. Supporremo inoltre  $\mathbf{A}$  reale.

#### Sistemi lineari generali

La decomposizione in valori singolari permette di esaminare in maniera numericamente più conveniente le nozioni di consistenza di un sistema lineare, di unicità delle soluzioni, e più in generale di dimensione dello spazio delle soluzioni del sistema.

Sia  $\mathbf{A}$  una matrice  $\in \mathbb{R}^{m \times n}$ , con  $m \geq n$  e rango  $r$ , e sia  $\mathbf{b}$  un vettore dello spazio  $\mathbb{R}^m$ . Il problema che consideriamo è la soluzione del sistema lineare

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (\text{A.38})$$

Usando la SVD di  $\mathbf{A}$ , il sistema (A.38) diventa

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \mathbf{x} = \mathbf{b}$$

da cui

$$\mathbf{\Sigma}\mathbf{z} = \mathbf{d}$$

ove  $\mathbf{z} = \mathbf{V}^T \mathbf{x}$  e  $\mathbf{d} = \mathbf{U}^T \mathbf{b}$ . Il sistema di equazioni  $\mathbf{\Sigma}\mathbf{z} = \mathbf{d}$  è diagonale e quindi può essere studiato facilmente. Si ha infatti

$$\begin{aligned} \sigma_j z_j &= d_j, & \text{se } j \leq n \text{ e } \sigma_j \neq 0 \\ 0z_j &= d_j, & \text{se } j \leq n \text{ e } \sigma_j = 0 \\ 0 &= d_j, & \text{se } j > n \end{aligned}$$

Il secondo insieme di valori è vuoto se  $r = n$  e il terzo è vuoto se  $n = m$ .

Le equazioni risultano *consistenti*, cioè esiste una soluzione se e solo se  $d_j = 0$  quando  $\sigma_j = 0$  o  $j > n$ . Se  $r < n$ , allora le  $z_j$  associate con un  $\sigma_j$  nullo possono assumere un valore arbitrario. Ritornando alle variabili  $\mathbf{x} = \mathbf{V}\mathbf{z}$ , questi valori arbitrari servono a parametrizzare lo spazio di tutte le possibili soluzioni  $\mathbf{x}$ .

Siano  $\mathbf{u}_j$  e  $\mathbf{v}_j$  le colonne di  $\mathbf{U}$  e  $\mathbf{V}$ . Allora la decomposizione SVD può essere scritta

$$\mathbf{A}\mathbf{v}_j = \sigma_j \mathbf{u}_j, \quad j = 1, 2, \dots, n.$$

Se  $\sigma_j = 0$ , allora  $\mathbf{A}\mathbf{v}_j = 0$  e  $\mathbf{v}_j$  è nello *spazio nullo* di  $\mathbf{A}$ , cioè in  $\mathcal{N}(\mathbf{A})$ , mentre se  $\sigma_j \neq 0$ , allora  $\mathbf{u}_j \in \mathcal{R}(\mathbf{A})$ .

Si può pertanto dare una descrizione completa dello *spazio nullo* e dello *spazio immagine*, nel modo seguente. Sia  $V_0$  l'insieme delle colonne  $\mathbf{v}_j$  per le quali  $\sigma_j = 0$ , e  $V_1$  le rimanenti colonne  $\mathbf{v}_j$ . Analogamente, sia  $U_1$  l'insieme delle colonne  $\mathbf{u}_j$  per le quali  $\sigma_j \neq 0$ , e  $U_0$  le rimanenti colonne  $\mathbf{u}_j$ , incluse quelle con  $j > n$ . Vi sono  $r$  colonne in  $V_0$ ,  $n - r$  in  $V_1$  e in  $U_1$  e  $m - n + r$  in  $U_0$ . Inoltre

1.  $V_0$  è una base ortonormale per  $\mathcal{N}(\mathbf{A})$ .
2.  $V_1$  è una base ortonormale per  $\mathcal{N}(\mathbf{A})^\perp$ .
3.  $U_1$  è una base ortonormale per  $\mathcal{R}(\mathbf{A})$ .
4.  $U_0$  è una base ortonormale per  $\mathcal{R}(\mathbf{A})^\perp$ .

Lasciamo come esercizio la esemplificazione dei risultati precedenti relativamente alla seguente matrice

$$\mathbf{A} = \begin{bmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{bmatrix}$$

e ad esempio,  $\mathbf{b} = [5 \ 5 \ 5 \ 5 \ 5]^T$  e  $\mathbf{b} = [4 \ 5 \ 5 \ 5 \ 5]^T$ .

La SVD di  $\mathbf{A}$  è la seguente

$$\mathbf{A} = \begin{bmatrix} 0.355 & -0.689 & 0.541 & 0.193 & 0.265 \\ 0.399 & -0.376 & -0.802 & -0.113 & 0.210 \\ 0.443 & -0.062 & 0.160 & -0.587 & -0.656 \\ 0.487 & 0.251 & -0.079 & 0.742 & -0.378 \\ 0.531 & 0.564 & 0.180 & -0.235 & 0.559 \end{bmatrix} \cdot \begin{bmatrix} 35.127 & 0 & 0 \\ 0 & 2.465 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.202 & 0.890 & 0.408 \\ 0.517 & 0.257 & -0.816 \\ 0.832 & -0.376 & 0.408 \end{bmatrix}$$

### Problema dei minimi quadrati lineari

Nel *problema dei minimi quadrati lineari*, un'estensione del problema precedente, si cerca un vettore  $\mathbf{x}$  che minimizza la lunghezza euclidea del vettore *residuo*

$$\mathbf{r} = \mathbf{Ax} - \mathbf{b} \quad (\text{A.39})$$

Brevemente, l'origine del problema, che è anche noto come problema della *regressione lineare*, è la seguente (cfr. i Capitoli 4 e 8 per maggiori dettagli). Sono date  $m$  coppie di valori

$$(t_i, y_i^*), \quad i = 1, \dots, m \quad (\text{A.40})$$

e si suppone che esista una relazione funzionale  $t \rightarrow y(t)$  tale che

$$y_i^* = y(t_i)$$

Si vuole, quindi, approssimare la funzione  $y(t)$  mediante la seguente *combinazione lineare* di  $n$  funzioni  $\phi_j$ , con  $n \leq m$

$$y(t) \approx \Phi(t) := x_1\phi_1(t) + x_2\phi_2(t) + \dots + x_n\phi_n(t) \quad (\text{A.41})$$

Osserviamo che nelle applicazioni i valori  $y_i^*$  rappresentano usualmente delle rilevazioni sperimentali corrispondenti a diversi valori della variabile  $t$ . La funzione  $y(t)$  non è, quindi, nota a priori e la funzione  $\Phi(t)$  definita in (A.41) rappresenta una ipotesi, o più precisamente un *modello matematico*, mediante il quale si vuole conoscere, in particolare, l'andamento del fenomeno, che si sta studiando, per valori della variabile  $t$  diversi da  $t_i$ . Quando, invece, la funzione  $y(t)$  è nota a priori, la  $\Phi(t)$  rappresenta una approssimazione di  $y(t)$  mediante le funzioni  $\phi_j(t)$ , che vengono scelte nell'ambito di classi di funzioni "semplici", quali ad esempio i polinomi algebrici o trigonometrici.

I valori  $x_j$ ,  $j = 1, \dots, n$ , i cosiddetti *parametri* del modello matematico, vengono scelti in maniera da minimizzare<sup>11</sup> una distanza, opportunamente definita, tra i valori calcolati  $\Phi(t_i)$  e i dati  $y_i^*$ , per  $i = 1, \dots, m$ . In particolare, il *metodo dei minimi quadrati* corrisponde a minimizzare la distanza *euclidea*.

Se introduciamo la matrice  $\mathbf{A}$  (nota anche come matrice *design*) di componenti

$$a_{ij} = \phi_j(t_i) \quad (\text{A.42})$$

e indichiamo con  $\mathbf{b}$  e  $\mathbf{x}$  i vettori di componenti  $y_i^*$  e  $x_j$ , il problema precedente può essere interpretato come la ricerca di un vettore  $\mathbf{x} \in \mathbb{R}^n$  che risolve il seguente problema di minimo

$$\min_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^m r_i^2 \quad (\text{A.43})$$

ove  $r_i$  rappresentano le componenti del vettore residuo corrispondente al seguente sistema sovradeterminato

$$\mathbf{Ax} = \mathbf{b} \quad (\text{A.44})$$

Il problema può avere più soluzioni quando le funzioni base  $\phi_i(t)$  sono linearmente dipendenti sui dati  $t_i$ ; questa situazione può anche essere la conseguenza di errori sperimentali e di arrotondamento ed è la causa del possibile *malcondizionamento* del problema dei minimi quadrati.

Un modo per risolvere (A.43) consiste nello scrivere la *condizione necessaria* per l'esistenza del minimo, cioè nel porre uguali a zero le derivate di  $r^2$  rispetto a  $x_j$ ; si ottiene in questo modo il seguente sistema lineare

$$\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b} \quad (\text{A.45})$$

usualmente chiamato *sistema delle equazioni normali*. Si tratta di un sistema di  $n$  equazioni in  $n$  incognite. La matrice dei coefficienti  $\mathbf{A}^T \mathbf{A}$  è simmetrica, e nel caso in cui la caratteristica di  $\mathbf{A}$  sia  $n$  (che corrisponde al caso in cui le funzioni base  $\phi_i(t)$  siano linearmente indipendenti), è pure *definita positiva*. Per la sua risoluzione si hanno, quindi, a disposizione vari metodi, in particolare, il metodo di Cholesky (cfr. Capitolo 2).

<sup>11</sup>L'operazione di ricerca ottimale dei parametri è nota come procedura di *fitting*.

Tuttavia, osservando che il *condizionamento* della matrice  $\mathbf{A}$  viene *peggiorato*<sup>12</sup> nel calcolo di  $\mathbf{A}^T \mathbf{A}$ , possono talvolta essere più opportuni, per motivi di stabilità numerica, quei metodi che permettono il calcolo della soluzione  $\mathbf{x}$ , senza la costruzione esplicita della matrice  $\mathbf{A}^T \mathbf{A}$ .

Una possibilità, in questo senso, è offerta dalla la SVD di  $\mathbf{A}$ . Infatti, dal momento che le matrici ortogonali conservano la distanza, si ha

$$\|\mathbf{r}\|_2 = \|\mathbf{U}^T(\mathbf{A}\mathbf{V}\mathbf{V}^T \mathbf{x} - \mathbf{b})\|_2 = \|\boldsymbol{\Sigma}\mathbf{z} - \mathbf{d}\|_2$$

ove  $\mathbf{z} = \mathbf{V}^T \mathbf{x}$  e  $\mathbf{d} = \mathbf{U}^T \mathbf{b}$ . La SVD quindi, riduce il problema generale dei minimi quadrati a un problema relativo a una matrice *diagonale*. Il vettore  $\mathbf{z}$  che fornisce il minimo di  $\|\mathbf{r}\|_2$  è dato da

$$\begin{aligned} z_j &= \frac{d_j}{\sigma_j}, & \text{se } \sigma_j \neq 0 \\ z_j &= \text{arbitrari}, & \text{se } \sigma_j = 0 \end{aligned}$$

La trasformazione  $\mathbf{x} = \mathbf{V}\mathbf{z}$  risolve il problema originario. Se la matrice  $\mathbf{A}$  ha rango minore di  $n$ , la soluzione non è *unica*. In questo caso è possibile aggiungere un'ulteriore condizione. Si può ad esempio *definire come soluzione del problema dei minimi quadrati*, il vettore  $\mathbf{x}$  di *minima lunghezza*. Tale vettore può essere ottenuto, ponendo

$$z_j = 0, \quad \text{se } \sigma_j = 0$$

#### A.5.4 Pseudoinversa

Data una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , viene detta *matrice pseudoinversa*<sup>13</sup> di  $\mathbf{A}$  l'unica matrice  $\mathbf{A}^+ \in \mathbb{R}^{n \times m}$  che verifica le seguenti condizioni

1.  $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$
2.  $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$
3.  $\mathbf{A}\mathbf{A}^+$  è simmetrica
4.  $\mathbf{A}^+\mathbf{A}$  è simmetrica

Si verifica, facilmente, che nel caso di una matrice  $\mathbf{A}$  quadrata e non singolare si ha  $\mathbf{A}^+ = \mathbf{A}^{-1}$ . Inoltre, se  $\mathbf{A}$  è rettangolare e di rango  $n$  ( $n \leq m$ ), allora  $\mathbf{A}^+ = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ , mentre se  $m \leq n$  e il rango di  $\mathbf{A}$  è  $m$ , allora  $\mathbf{A}^+ = \mathbf{A}^T (\mathbf{A}\mathbf{A}^T)^{-1}$ . Negli

<sup>12</sup>Si ha, infatti (cfr. Capitolo 2)  $\mu_2(\mathbf{A}^T \mathbf{A}) = (\mu_2(\mathbf{A}))^2$ .

<sup>13</sup>Uno studio sistematico del concetto di inversa generalizzata è stato iniziato da Moore nel 1920. Un nuovo impulso a tale argomento è dovuto a Penrose nel 1955, da cui anche l'attuale nome di *pseudoinversa secondo Moore-Penrose*.



altri casi non c'è una procedura semplice per il calcolo della matrice pseudoinversa. Un metodo generale consiste nell'utilizzo della SVD di  $\mathbf{A}$ . Si ha, infatti

$$\boxed{\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T} \quad (\text{A.46})$$

ove la matrice  $\mathbf{\Sigma}^+$  è la pseudoinversa della matrice  $\mathbf{\Sigma}$ . Si verifica facilmente che la  $\mathbf{\Sigma}^+$  è la matrice diagonale  $\in \mathbb{R}^{n \times m}$  ad elementi  $\sigma_j^+$ , con

$$\sigma_j^+ = \begin{cases} \frac{1}{\sigma_j} & \text{se } \sigma_j \neq 0 \\ 0, & \text{se } \sigma_j = 0 \end{cases}$$

Dimostreremo, ora, che la soluzione del problema dei minimi quadrati può essere espressa nella forma  $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ .

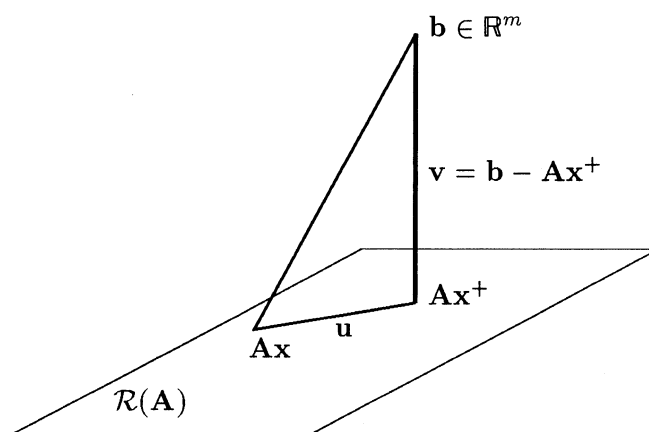


Figura A.10: Soluzione secondo i minimi quadrati.

Siano  $\mathcal{N}(\mathbf{A})$ ,  $\mathcal{R}(\mathbf{A})$ , rispettivamente lo spazio nullo e lo spazio immagine della matrice  $\mathbf{A}$ ; con  $\mathcal{N}(\mathbf{A})^\perp$ ,  $\mathcal{R}(\mathbf{A})^\perp$ , indichiamo i corrispondenti spazi lineari ortogonali.

Indichiamo con  $\mathbf{P}$  la matrice  $\in \mathbb{R}^{n \times n}$ , che proietta  $\mathbb{R}^n$  su  $\mathcal{N}(\mathbf{A})^\perp$  e con  $\overline{\mathbf{P}}$  la matrice  $\in \mathbb{R}^{m \times m}$  che proietta  $\mathbb{R}^m$  su  $\mathcal{R}(\mathbf{A})$ , cioè

$$\begin{aligned} \mathbf{P} = \mathbf{P}^T = \mathbf{P}^2, \quad \mathbf{P}\mathbf{x} = 0 & \iff \mathbf{x} \in \mathcal{N}(\mathbf{A}) \\ \overline{\mathbf{P}} = \overline{\mathbf{P}}^T = \overline{\mathbf{P}}^2, \quad \overline{\mathbf{P}}\mathbf{y} = \mathbf{y} & \iff \mathbf{y} \in \mathcal{R}(\mathbf{A}) \end{aligned}$$

Per ogni  $\mathbf{b} \in \mathcal{R}(\mathbf{A})$  vi è un unico  $\mathbf{x}_1 \in \mathcal{N}(\mathbf{A})^\perp$ , che soddisfa  $\mathbf{A}\mathbf{x}_1 = \mathbf{b}$ , cioè vi è una trasformazione univoca  $f: \mathcal{R}(\mathbf{A}) \rightarrow \mathbb{R}^n$ , con

$$\mathbf{A}f(\mathbf{b}) = \mathbf{b}, \quad f(\mathbf{b}) \in \mathcal{N}(\mathbf{A})^\perp, \quad \forall \mathbf{b} \in \mathcal{R}(\mathbf{A})$$

Infatti, dato  $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ , vi è un vettore  $\mathbf{x}$ , che soddisfa  $\mathbf{b} = \mathbf{A}\mathbf{x}$ ; quindi  $\mathbf{b} = \mathbf{A}(\mathbf{P}\mathbf{x} + (\mathbf{I} - \mathbf{P})\mathbf{x}) = \mathbf{A}\mathbf{P}\mathbf{x} = \mathbf{A}\mathbf{x}_1$ , dove  $\mathbf{x}_1 = \mathbf{P}\mathbf{x} \in \mathcal{N}(\mathbf{A})^\perp$ , poiché  $(\mathbf{I} - \mathbf{P})\mathbf{x} \in \mathcal{N}(\mathbf{A})$ . Inoltre se  $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{N}(\mathbf{A})^\perp$ , con  $\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2$ , ne segue che  $\mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A})^\perp = \{0\}$ , che implica  $\mathbf{x}_1 = \mathbf{x}_2$ . La trasformazione  $f$  è ovviamente lineare.

La trasformazione composta  $f \circ \overline{\mathbf{P}} : \mathbf{y} \in \mathbb{R}^m \rightarrow f(\overline{\mathbf{P}}\mathbf{y}) \in \mathbb{R}^n$  è definita, dal momento che  $\overline{\mathbf{P}}\mathbf{y} \in \mathcal{R}(\mathbf{A})$ ; essendo, inoltre, lineare, essa è rappresentata da una matrice  $\in \mathbb{R}^{n \times m}$ . Lasciamo come esercizio la dimostrazione che tale matrice verifica le condizioni di Moore-Penrose, e quindi che essa coincide con la matrice pseudoinversa  $\mathbf{A}^+$  di  $\mathbf{A}$ . Osserviamo, anche, che  $\mathbf{A}^+\mathbf{A}$  è il *proiettore ortogonale*  $\mathbf{P}$  e analogamente che  $\mathbf{A}\mathbf{A}^+$  coincide con  $\overline{\mathbf{P}}$ . Si può allora dimostrare il seguente risultato.

**Teorema A.18** *Il vettore  $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$  soddisfa a*

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \geq \|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\|_2 \quad (\text{A.47})$$

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\|_2, \quad \text{e } \mathbf{x} \neq \mathbf{x}^+ \Rightarrow \|\mathbf{x}\|_2 > \|\mathbf{x}^+\|_2 \quad (\text{A.48})$$

ossia il vettore  $\mathbf{x}^+$  è, tra le soluzioni del problema dei minimi quadrati, quella di minima lunghezza euclidea.

**DIMOSTRAZIONE.** Per il fatto che  $\mathbf{A}\mathbf{A}^+$  è il proiettore ortogonale su  $\mathcal{R}(\mathbf{A})$  si ha  $\forall \mathbf{x} \in \mathbb{R}^n$  (cfr. Figura A.10)

$$\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{u} - \mathbf{v}$$

$$\mathbf{u} := \mathbf{A}(\mathbf{x} - \mathbf{A}^+\mathbf{b}) \in \mathcal{R}(\mathbf{A}), \quad \mathbf{v} := (\mathbf{I} - \mathbf{A}\mathbf{A}^+)\mathbf{b} = \mathbf{b} - \mathbf{A}\mathbf{x}^+ \in \mathcal{R}(\mathbf{A})^\perp$$

Di conseguenza  $\forall \mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 \geq \|\mathbf{v}\|_2^2 = \|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\|_2^2$$

e  $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\|_2$  ha luogo precisamente se  $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{A}^+\mathbf{b}$ . Per un tale  $\mathbf{x}$ , poniamo

$$\mathbf{x} = \mathbf{x}^+ + (\mathbf{x} - \mathbf{x}^+), \quad \mathbf{x}^+ \in \mathcal{N}(\mathbf{A})^\perp$$

Si ha  $\mathbf{A}(\mathbf{x} - \mathbf{x}^+) = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{A}^+\mathbf{b} = 0$ , quindi  $\mathbf{x} - \mathbf{x}^+ \in \mathcal{N}(\mathbf{A})$ ; pertanto

$$\|\mathbf{x}\|_2^2 = \|\mathbf{x}^+\|_2^2 + \|\mathbf{x} - \mathbf{x}^+\|_2^2 > \|\mathbf{x}^+\|_2^2$$

■

## Regolarizzazione di un sistema lineare

Dato il sistema lineare  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , con  $\mathbf{A} \in \mathbb{R}^{m \times n}$  e  $\mathbf{b} \in \mathbb{R}^m$ , e con soluzione nel senso dei minimi quadrati  $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$ , supponiamo che il termine noto  $\mathbf{b}$  sia perturbato mediante il vettore  $\delta\mathbf{b} \in \mathbb{R}^m$ . In corrispondenza, si ha da risolvere il sistema  $\mathbf{A}(\mathbf{x}^+ + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$ , da cui  $\delta\mathbf{x} = \mathbf{A}^+\delta\mathbf{b}$ . Si ha, allora, la seguente maggiorazione

$$\|\delta\mathbf{x}\|_2 \leq \|\mathbf{A}^+\|_2 \|\delta\mathbf{b}\|_2 = \sigma_r^{-1} \|\delta\mathbf{b}\|_2$$

ove  $\sigma_r$  rappresenta il minimo valore singolare di  $\mathbf{A}$ . D'altra parte, dalla relazione  $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$  si ricava  $\|\mathbf{A}\mathbf{A}^+\mathbf{b}\|_2 = \|\mathbf{A}\mathbf{x}^+\|_2 \leq$  (cfr. (A.35))  $\sigma_1\|\mathbf{x}^+\|_2$ . In definitiva, si ha la seguente maggiorazione dell'*errore relativo*

$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}^+\|_2} \leq \frac{\sigma_1}{\sigma_r} \frac{\|\delta\mathbf{b}\|_2}{\|\mathbf{A}\mathbf{A}^+\mathbf{b}\|_2} \quad (\text{A.49})$$

La maggiorazione (A.49) suggerisce l'assunzione del quoziente  $\sigma_1/\sigma_r$  come *numero di condizionamento*  $\mu_2(\mathbf{A})$  della matrice  $\mathbf{A}$  nei riguardi della soluzione del sistema lineare  $\mathbf{A}\mathbf{x} = \mathbf{b}$  mediante i minimi quadrati. Si verifica facilmente che, quando  $m = n$  e la matrice è non singolare, la definizione ora data coincide con quella fornita nel Capitolo 2.

La definizione precedente di numero di condizionamento suggerisce come costruire problemi che approssimino il problema di minimo dato  $\min\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ , e che siano meglio condizionati. Più precisamente, si può procedere nel seguente modo. Data la decomposizione  $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ , si costruisce la matrice diagonale  $\mathbf{\Sigma}_r = \text{diag}(\eta_1, \eta_2, \dots, \eta_\mu, \dots)$  ponendo

$$\eta_\mu := \begin{cases} \sigma_\mu^{-1} & \text{se } \sigma_\mu \geq \tau \\ 0 & \text{altrimenti} \end{cases}$$

ove  $\tau > 0$  è un parametro da scegliere opportunamente. In altre parole il passaggio da  $\mathbf{\Sigma}$  a  $\mathbf{\Sigma}_r$  comporta l'eliminazione dei valori singolari "piccoli". Di conseguenza, anziché considerare la soluzione  $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$ , si considera l'approssimazione  $\mathbf{x}_r^+ = \mathbf{A}_r^+\mathbf{b}$ , ove  $\mathbf{A}_r^+ := \mathbf{V}\mathbf{\Sigma}_r^+\mathbf{U}^T$ . La matrice  $\mathbf{A}_r^+$  è chiamata la *effettiva pseudoinversa* di  $\mathbf{A}$  e, per quanto abbiamo detto in precedenza, il corrispondente problema è meglio condizionato. Si può mostrare che la matrice  $\mathbf{A}_r^+$  verifica le seguenti condizioni

$$\begin{aligned} \mathbf{A}_r^+\mathbf{A} &= (\mathbf{A}_r^+\mathbf{A})^T, & \mathbf{A}\mathbf{A}_r^+ &= (\mathbf{A}\mathbf{A}_r^+)^T, & \mathbf{A}_r^+\mathbf{A}\mathbf{A}_r^+ &= \mathbf{A}_r^+ \\ \|\mathbf{A}\mathbf{A}_r^+\mathbf{A} - \mathbf{A}\|_2 &\leq \tau \end{aligned}$$

L'eliminazione dei valori singolari piccoli è chiamata una *regolarizzazione* del problema. Naturalmente, il processo migliora il condizionamento, ma introduce un errore nel metodo. Vi sono altre possibilità per regolarizzare un problema malcondizionato. Segnaliamo in particolare un metodo dovuto a Tichonov (1963) che corrisponde a smorzare l'influenza dei valori singolari "piccoli".

◆ **Esercizio A.19** Dimostrare che le matrici  $\mathbf{A}$  e  $\mathbf{A}^*$  hanno gli stessi valori singolari.

◆ **Esercizio A.20** Dimostrare che se  $\mathbf{A} \in \mathbb{C}^{n \times n}$ , allora

$$|\det(\mathbf{A})| = \prod_{i=1}^n \sigma_i$$

◆ **Esercizio A.21** Dimostrare che  $\mathbf{A}^* \mathbf{A}$  è definita positiva se e solo se le colonne di  $\mathbf{A}$  sono linearmente indipendenti, ossia se  $\mathbf{A}$  ha rango massimo.

◆ **Esercizio A.22** Calcolare la SVD e la pseudoinversa delle seguenti matrici

1. un vettore  $\mathbf{v} \in \mathbb{C}^n$ ,
2. una matrice nulla  $\in \mathbb{C}^{m \times n}$ ,
3. una matrice a rango 1,  $\mathbf{A} = \mathbf{xy}^*$ ,  $\mathbf{x} \in \mathbb{C}^m$ ,  $\mathbf{y} \in \mathbb{C}^n$ .

◆ **Esercizio A.23** Se  $\mathbf{A} \in \mathbb{C}^{m \times n}$ ;  $\mathbf{B} \in \mathbb{C}^{n \times r}$ , sono due matrici di rango massimo, dimostrare che

$$(\mathbf{AB})^+ = \mathbf{B}^+ \mathbf{A}^+$$

Tale relazione può non essere valida quando una delle due matrici non è di rango massimo. Ad esempio, se

$$\mathbf{A} = \frac{1}{3} \begin{bmatrix} 2 & 2 \\ -2 & -2 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

allora

$$(\mathbf{AB})^+ = \frac{1}{3} [-2, 2, -1], \quad \mathbf{B}^+ \mathbf{A}^+ = \frac{1}{30} [-2, 2, -1]$$

◆ **Esercizio A.24** Se si definisce  $\Phi(x) = \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2$ ,  $\mathbf{x} \in \mathbb{R}^n$ , mostrare che il gradiente  $\nabla \Phi(x)$  è dato da  $\nabla \Phi(x) = \mathbf{A}^T (\mathbf{Ax} - \mathbf{b})$ .

◆ **Esercizio A.25** Dimostrare che una matrice  $\mathbf{A} \in \mathbb{C}^{n \times n}$  può essere scritta in maniera univoca nella seguente forma

$$\mathbf{A} = \mathbf{HQ}$$

ove  $\mathbf{H}$  è semidefinita positiva e  $\mathbf{Q}$  è unitaria. Tale decomposizione viene detta decomposizione polare. (Sugg. Se  $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$  è la SVD di  $\mathbf{A}$ , si ponga  $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^*$ ,  $\mathbf{Q} = \mathbf{UV}^*$ ).

## A.6 Matrici non negative

Un vettore  $\mathbf{x} \in \mathbb{R}^n$ , o una matrice  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , è *non negativa* se tutte le componenti sono non negative, ossia se  $x_i \geq 0$ ,  $i = 1, \dots, n$ , o  $a_{ij} \geq 0$ ,  $i = 1, \dots, m$ ;  $j = 1, \dots, n$ . Nel caso in cui le componenti sono tutte strettamente positive, il vettore  $\mathbf{x}$  o la matrice  $\mathbf{A}$  è detta *positiva*. Per indicare le matrici non negative (positive), si userà la notazione  $\mathbf{A} \geq \mathbf{0}$  (rispettivamente  $\mathbf{A} > \mathbf{0}$ ). Di conseguenza la notazione  $\mathbf{x} \geq \mathbf{y}$ , o  $\mathbf{A} \geq \mathbf{B}$ , significherà  $\mathbf{x} - \mathbf{y} \geq \mathbf{0}$ , rispettivamente  $\mathbf{A} - \mathbf{B} \geq \mathbf{0}$ . Infine, con  $|\mathbf{A}|$  si indicherà la matrice che ha come elementi i valori assoluti degli elementi della matrice  $\mathbf{A}$ .

Un esempio importante di matrici non negative è dato dalle matrici *stocastiche* o *matrici di Markov*. Ricordiamo (cfr. Capitolo 8) che gli elementi di tali matrici rappresentano delle probabilità, e quindi sono non negativi; inoltre, si ha che le somme di tutte le colonne sono uguali a 1, ossia  $\sum_{i=1}^n a_{ij} = 1$ ,  $j = 1, \dots, n$ .

Per le matrici  $\mathbf{A}$  di Markov, con  $\mathbf{A} > \mathbf{0}$  si ha il seguente importante risultato che introduce l'interesse numerico delle matrici positive.

**Proposizione A.11** Se  $\mathbf{A}$  è una matrice stocastica positiva, allora esiste un unico vettore positivo  $\hat{\mathbf{x}}$ , con  $\sum_{i=1}^n \hat{x}_i = 1$ , tale che per ogni vettore non negativo  $\mathbf{x}^{(0)}$  con  $\sum_{i=1}^n x_i^{(0)} = 1$ , la successione definita nel modo seguente

$$\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)}, \quad k = 0, 1, \dots$$

verifica

$$\mathbf{x}^{(k)} \rightarrow \hat{\mathbf{x}}, \quad \text{per } k \rightarrow \infty$$

La dimostrazione di tale risultato è basata sul seguente risultato relativo al raggio spettrale  $\rho(\mathbf{A})$  delle matrici positive.

**Proposizione A.12 (Perron)** Sia  $\mathbf{A}$  una matrice positiva di ordine  $n$ . Allora,  $\rho(\mathbf{A})$  è un autovalore semplice di  $\mathbf{A}$ , e tutti gli altri autovalori sono minori in modulo di  $\rho(\mathbf{A})$ . Inoltre, è possibile prendere un autovettore corrispondente a  $\rho(\mathbf{A})$  positivo.

Da tale risultato si ricava che per le matrici positive l'autovalore di modulo massimo è sempre reale.

Il risultato di Perron può essere esteso alle matrici non negative che hanno la proprietà della *irriducibilità*, e che analizzeremo nel successivo paragrafo.

I risultati di base di Perron e Frobenius sulle matrici non negative risalgono ai primi anni del '900, ma l'argomento ha conosciuto negli ultimi anni un risveglio di interesse, a causa delle molte applicazioni delle matrici non negative nell'analisi numerica, nella probabilità e nella statistica, nei modelli economici, e in molti altri tipi di modelli matematici.

### A.6.1 Matrici irriducibili

Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  è detta *riducibile* se vi è una matrice permutazione  $\mathbf{P}$  tale che

$$\mathbf{PAP}^T = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ 0 & \mathbf{A}_{22} \end{bmatrix} \quad (\text{A.50})$$

ove  $\mathbf{A}_{11}$  e  $\mathbf{A}_{22}$  sono sottomatrici quadrate. Una matrice è *irriducibile* se essa è *non riducibile*.

Naturalmente, una matrice ad elementi tutti diversi da zero è irriducibile; d'altra parte una matrice con tutta una colonna o riga nulla è riducibile. In effetti, il concetto di riducibilità non è connesso con la grandezza o il segno degli elementi di una matrice, ma dipende solo dalla disposizione degli elementi nulli. Questa idea è alla base dell'utilizzo del *grafo orientato* associato alla matrice.

Il grafo orientato di una matrice  $\in \mathbb{R}^{n \times n}$  è ottenuto congiungendo  $n$  punti (vertici)  $P_1, \dots, P_n$  mediante una linea orientata da  $P_i$  a  $P_j$ , se  $a_{ij} \neq 0$ , per  $i \neq j$ . Ricordiamo che un grafo orientato è *fortemente connesso* se per ogni coppia di punti distinti  $P_i$  e  $P_j$  vi è un cammino orientato da  $P_i$  a  $P_j$ . Si ha allora il seguente risultato.

**Proposizione A.13** Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  è irriducibile se e solo se il grafo orientato relativo a  $\mathbf{A}$  è fortemente connesso.

Ad esempio (cfr. Figura A.11), la matrice

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

è irriducibile, mentre sono riducibili le matrici

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 2 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}$$

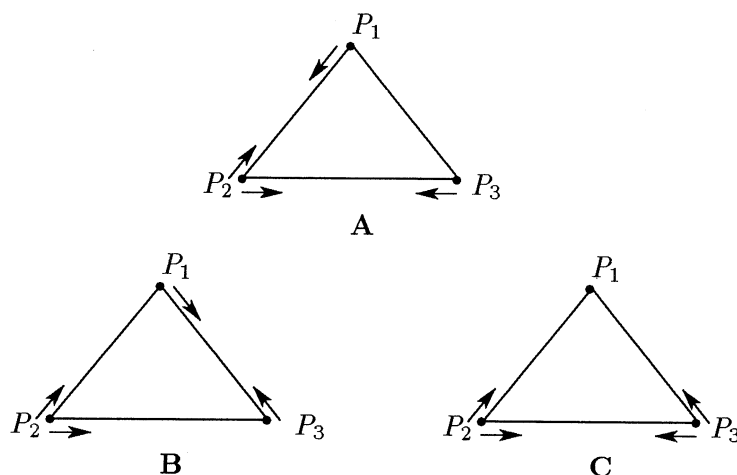


Figura A.11: Grafi relativi alle matrici A,B,C.

Per le matrici non negative e irriducibili si ha il seguente risultato che contiene il precedente risultato di Perron come caso particolare.

**Teorema A.19** (Perron-Frobenius) Sia  $\mathbf{A}$  una matrice  $\in \mathbb{R}^{n \times n}$  non negativa e irriducibile. Allora il raggio spettrale  $\rho(\mathbf{A})$  è un autovalore semplice di  $\mathbf{A}$  e un autovettore associato può essere preso positivo. Inoltre, se  $\mathbf{A}$  ha almeno una riga con tutti gli elementi diversi dallo zero allora ogni altro autovalore  $\lambda$  di  $\mathbf{A}$  è tale che  $|\lambda| < \rho(\mathbf{A})$ .

Per una dimostrazione si veda ad esempio Ortega [123]; essa sfrutta il seguente interessante risultato.

**Lemma A.1** Se  $\mathbf{B}$  è una matrice  $\in \mathbb{R}^{n \times n}$  irriducibile non negativa con elementi diagonali positivi, allora  $\mathbf{B}^{n-1} > \mathbf{0}$ .

Come illustrazione del risultato di Perron-Frobenius, si consideri la seguente matrice

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

che, come si verifica facilmente, è irriducibile. I suoi autovalori corrispondono alle radici dell'equazione caratteristica  $\lambda^3 - 1 = 0$ , cioè alle radici cubiche dell'unità 1 e  $(-1 \pm i\sqrt{3})/2$ . I tre autovalori hanno quindi modulo uguale a 1, ma l'autovalore 1 è semplice e ad esso corrisponde l'autovettore  $[1, 1, 1]^T$ .

### A.6.2 Matrici con inverse non negative; M-matrici

Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  è ad *inversa non negativa* se essa è invertibile e  $\mathbf{A}^{-1} \geq \mathbf{0}$ . Un esempio di condizioni che assicurano che una matrice  $\mathbf{A}$  è a inversa non negativa, è dato dal seguente risultato.

**Teorema A.20** Sia  $\mathbf{A}$  una matrice  $\in \mathbb{R}^{n \times n}$  con elementi diagonali positivi e poniamo  $\mathbf{D} = \text{diag}(a_{11}, \dots, a_{nn})$ . Supponiamo che  $\mathbf{B} = \mathbf{D} - \mathbf{A} \geq \mathbf{0}$  e  $\rho(\mathbf{D}^{-1}\mathbf{B}) < 1$ . Allora esiste  $\mathbf{A}^{-1}$ , con  $\mathbf{A}^{-1} \geq \mathbf{0}$ . Inoltre, se  $\mathbf{A}$  è irriducibile, allora  $\mathbf{A}^{-1} > \mathbf{0}$ .

Per le matrici non negative o ad inversa non negativa vi sono diversi risultati di tipo *confronto*.

**Proposizione A.14** Se  $\mathbf{A} \geq \mathbf{B}$  sono matrici  $\in \mathbb{R}^{n \times n}$  con inverse non negative, allora  $\mathbf{A}^{-1} \leq \mathbf{B}^{-1}$ .

**DIMOSTRAZIONE.** Da  $\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B}) \geq \mathbf{0}$  si conclude che  $\mathbf{A}^{-1}\mathbf{B} \leq \mathbf{I}$ . Quindi,  $(\mathbf{I} - \mathbf{A}^{-1}\mathbf{B})\mathbf{B}^{-1} \geq \mathbf{0}$ , che è equivalente a  $\mathbf{A}^{-1} \leq \mathbf{B}^{-1}$ . ■

**Proposizione A.15** Se  $\mathbf{0} \leq \mathbf{A} \leq \mathbf{B}$ , allora  $\rho(\mathbf{A}) \leq \rho(\mathbf{B})$

Vediamo ora una classe importante di matrici a inversa positiva.

**Definizione A.3** Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  è una *M-matrice* se  $a_{ij} \leq 0$ ,  $i, j = 1, \dots, n$ ,  $i \neq j$ , e  $\mathbf{A}^{-1} \geq \mathbf{0}$ . Una M-matrice simmetrica è detta *matrice di Stieltjes*<sup>14</sup>.

Posto  $\mathbf{A} = \mathbf{D} - \mathbf{B}$ , ove  $\mathbf{D}$  è la matrice diagonale costituita dagli elementi sulla diagonale principale di  $\mathbf{A}$  e  $\mathbf{B}$  è la matrice degli elementi fuori dalla diagonale, si ha

$$\mathbf{I} = (\mathbf{D} - \mathbf{B})\mathbf{A}^{-1} = \mathbf{D}\mathbf{A}^{-1} - \mathbf{B}\mathbf{A}^{-1}$$

Tale risultato mostra che gli elementi di  $\mathbf{D}$  sono positivi, in quanto  $\mathbf{B}\mathbf{A}^{-1} \geq \mathbf{0}$ . In effetti, si può dimostrare il seguente risultato.

<sup>14</sup>Thomas Jan Stieltjes (1856-1894).

**Proposizione A.16** Una matrice reale  $\mathbf{A}$  di ordine  $n$  con elementi fuori dalla diagonale non positivi è una  $M$ -matrice se e solo se  $\mathbf{A}$  ha elementi positivi sulla diagonale e  $\rho(\mathbf{D}^{-1}\mathbf{B}) < 1$ .

► **Esempio A.2** Sia  $\mathbf{A}$  la seguente matrice tridiagonale

$$\mathbf{A} = \begin{bmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & \ddots & & \\ & \ddots & \ddots & b_{n-1} & \\ & & c_{n-1} & a_n & \end{bmatrix} \quad (\text{A.51})$$

ove

$$a_i > 0, i = 1, \dots, n, \quad b_i < 0, c_i < 0, \quad i = 1, \dots, n-1 \quad (\text{A.52})$$

e

$$a_1 + b_1 > 0, \quad a_n + c_{n-1} > 0, \quad a_i + b_i + c_{i-1} \geq 0, \quad i = 2, \dots, n-1 \quad (\text{A.53})$$

Matrici di tale forma sono alla base di diversi modelli applicativi (si veda, ad esempio, nel Capitolo 7 la risoluzione di problemi differenziali mediante i metodi alle differenze finite).

Si può dimostrare che nelle ipotesi (A.52), (A.53) la matrice (A.51) è una  $M$ -matrice, con  $\mathbf{A}^{-1} > \mathbf{0}$ . ■

Come applicazione delle  $M$ -matrici esaminiamo la seguente situazione. Supponiamo che in un sistema lineare  $\mathbf{Ax} = \mathbf{b}$ , il vettore  $\mathbf{b}$  sia dato con approssimazione, ad esempio  $\mathbf{b}_1 \leq \mathbf{b} \leq \mathbf{b}_2$ . Se  $\mathbf{A}$  è una  $M$ -matrice, si ha allora la seguente stima della soluzione:  $\mathbf{A}^{-1}\mathbf{b}_1 \leq \mathbf{A}^{-1}\mathbf{b} = \mathbf{x} \leq \mathbf{A}^{-1}\mathbf{b}_2$ .

Una condizione sufficiente affinché una matrice sia una  $M$ -matrice è fornita dalla proprietà di predominanza diagonale, definita nel seguente modo.

**Definizione A.4** Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$  è diagonalmente dominante se

$$\sum_{j=1, j \neq i}^n |a_{ij}| \leq |a_{ii}|, \quad i = 1, \dots, n \quad (\text{A.54})$$

e strettamente diagonalmente dominante se in (A.54) si ha la disuguaglianza stretta per tutti gli indici  $i$ . Infine, la matrice è irriducibilmente diagonalmente dominante se essa è irriducibile, diagonalmente dominante, e la (A.54) ha luogo con disuguaglianza stretta per almeno un indice  $i$ .

Relativamente a tale classe di matrici si ha il seguente importante risultato.

**Teorema A.21** Una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , strettamente o irriducibilmente diagonalmente dominante, è invertibile.



**DIMOSTRAZIONE.** Considerando ad esempio il caso in cui  $\mathbf{A}$  è strettamente diagonalmente dominante, supponiamo che esista un  $\mathbf{x} \neq \mathbf{0}$ , tale che  $\mathbf{Ax} = \mathbf{0}$ . Sia  $|x_m| = \max_{1 \leq j \leq n} |x_j|$ . Allora  $|x_m| > 0$  e la disuguaglianza

$$|a_{mm}| |x_m| = \left| \sum_{j \neq m} a_{mj} x_j \right| \leq |x_m| \sum_{j \neq m} |a_{mj}|$$

contraddice la stretta dominanza diagonale. In maniera analoga si procede quando  $\mathbf{A}$  è irriducibilmente diagonalmente dominante. ■

Abbiamo inoltre la seguente proprietà.

**Teorema A.22** *Se una matrice  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , strettamente o irriducibilmente diagonalmente dominante, è tale che  $a_{ij} \leq 0$ ,  $i \neq j$  e  $a_{ii} > 0$ ,  $i = 1, \dots, n$ , allora  $\mathbf{A}$  è una M-matrice.*

Per una dimostrazione si veda ad esempio Ortega [123].

◆ **Esercizio A.26** *Verificare che una matrice tridiagonale è irriducibile se e solo se gli elementi della sopra- e sotto-diagonale sono diversi dallo zero.*

◆ **Esercizio A.27** *Sia  $\mathbf{A}$  una matrice  $\in \mathbb{R}^{n \times n}$  e  $\mathbf{P}$  una matrice permutazione. Mostrare che il grafo orientato di  $\mathbf{A}$  è fortemente connesso se e solo se il grafo di  $\mathbf{PAP}^T$  è fortemente connesso.*

◆ **Esercizio A.28** *Mostrare che se  $\mathbf{A}$  è una matrice non negativa e irriducibile, allora  $\mathbf{A}^{n-1}$  è positiva.*

◆ **Esercizio A.29** *Mostrare che se  $\mathbf{A}$  è una matrice non negativa e tale che gli elementi sulla diagonale sono tutti nulli, allora  $\mathbf{A}$  è riducibile.*

◆ **Esercizio A.30** *Dimostrare che una matrice  $\mathbf{A}$  con elementi fuori dalla diagonale non positivi è una matrice di Stieltjes se e solo se  $\mathbf{A}$  è definita positiva.*

Unum et unum duo, duo et duo quatuor,  
"odiosa cantio mihi erat"  
S. Agostino *Confessioni*, I, 13

## Appendice B

# Equazioni differenziali Tecniche analitiche

In questa appendice sono raccolti i risultati e le tecniche principali relative all'*integrazione analitica*, ossia al calcolo della soluzione in termini di funzioni polinomiali, esponenziali, trigonometriche, e simili, di particolari tipi di equazioni differenziali. Sottolineiamo che il calcolo numerico non è da considerare uno strumento sostitutivo del metodo analitico, ma come illustrato attraverso alcune applicazioni in questo capitolo, il metodo analitico può essere utile, oltre che a fornire talvolta direttamente la soluzione, ad evidenziare le proprietà e il comportamento qualitativo della soluzione, dando importanti indicazioni per una scelta conveniente dei metodi numerici.

### B.1 Separazione delle variabili

L'equazione differenziale

$$\frac{dy}{dt} = f(t, y) \quad (\text{B.1})$$

è detta *separabile* se  $f(t, y) = L(t)M(y)$ , ossia se  $f(t, y)$  può essere espressa come il prodotto di due funzioni continue, una della sola variabile  $t$  e l'altra della sola variabile  $y$ . Se  $M(y_0) = 0$ , allora  $y(t) = y_0$  è una soluzione. Se  $M(y) \neq 0$ , allora le variabili possono essere *separate* riscrivendo l'equazione (B.1) nel seguente modo

$$\frac{1}{M(y)} \frac{dy}{dt} = L(t)$$

da cui

$$\int \frac{dy}{M(y)} = \int L(t) dt + c \quad (\text{B.2})$$

con  $c$  costante arbitraria<sup>1</sup>. Se gli integrali possono essere calcolati in maniera analitica, la (B.2) definisce esplicitamente o implicitamente le soluzioni  $y = y(t)$  dell'equazione (B.1).

► **Esempio B.1** Come illustrazione consideriamo la seguente equazione differenziale

$$y' = e^{t+y}$$

Dividendo per  $e^y$ , si ottiene

$$e^{-y} y' = e^t \Rightarrow e^{-y} dy = e^t dt$$

da cui

$$-e^{-y} = e^t + c$$

■

► **Esempio B.2** Per l'equazione differenziale

$$y' = (t + 1) \cos y$$

dal momento che  $\cos y = 0$  per  $y$  multiplo dispari di  $\pi/2$ , alcune soluzioni sono date da  $y_k(t) = (2k + 1)(\pi/2)$ , con  $k$  intero. Le altre soluzioni sono ottenute per separazione delle variabili

$$\int \frac{dy}{\cos y} = \int (t + 1) dt$$

da cui l'equazione

$$\frac{1}{2} \ln \frac{1 + \sin y}{1 - \sin y} = \frac{t^2}{2} + t + c \Rightarrow \frac{1 + \sin y}{1 - \sin y} = e^{2((t^2/2)+t+c)}$$

che definisce le soluzioni *implicitamente*. La risoluzione, sia algebrica che grafica, di tale equazione non è un problema semplice. Per uno studio qualitativo delle soluzioni è più opportuno, invece, uno studio del campo di velocità che si ricava direttamente dall'equazione differenziale (cfr. Figura B.1). ■

## B.2 Equazione lineare del primo ordine

L'equazione differenziale (B.1) è detta *lineare* se  $f(t, y)$  è una funzione lineare in  $y$ . L'equazione è quindi della forma

$$\boxed{\frac{dy}{dt} + a(t)y = b(t)} \quad (\text{B.3})$$

<sup>1</sup>In questa appendice, seguendo una consuetudine, indicheremo con  $\int f(t) dt$  una *particolare* primitiva della funzione  $f(t)$ , anziché l'integrale generale, ossia l'insieme delle primitive di  $f(t)$ .

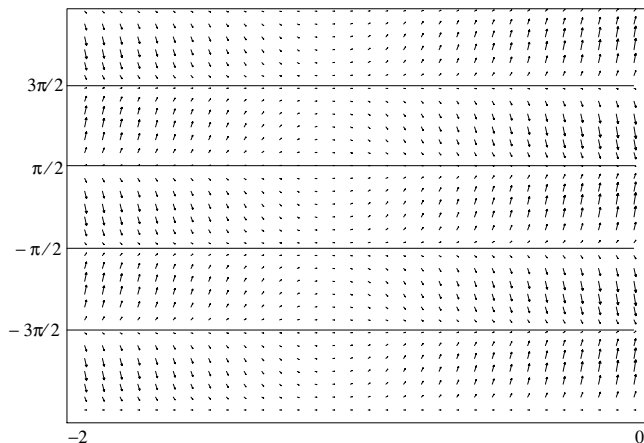


Figura B.1: Campo di velocità relativo all'equazione differenziale  $y' = (t + 1) \cos y$ .

ove  $a(t)$  e  $b(t)$  sono funzioni continue assegnate della variabile  $t$ . Se  $b(t) \equiv 0$ , allora l'equazione è chiamata *omogenea*, ed è un caso particolare di equazioni a variabili separabili.

Un metodo per risolvere l'equazione (B.3) consiste nell'ottenere una funzione  $\mu(t)$ , detto *fattore integrante*, tale che

$$\mu(t) \frac{dy}{dt} + \mu(t)a(t)y = \frac{d}{dt}[\mu(t)y] = \mu(t) \frac{dy}{dt} + \frac{d\mu}{dt} y$$

da cui

$$\frac{d\mu}{dt} = \mu(t)a(t)$$

Quest'ultima equazione è separabile ed ha la seguente soluzione particolare

$$\ln \mu = \int a(t) dt \Rightarrow \mu(t) = e^{\int a(t) dt}$$

Moltiplicando (B.3) per  $\mu(t)$ , si ottiene

$$\frac{d}{dt} \left[ y e^{\int a(t) dt} \right] = b(t) e^{\int a(t) dt}$$

e, quindi, integrando rispetto a  $t$  e dividendo per  $e^{\int a(t) dt}$  si ottiene la seguente soluzione generale

$$\boxed{y = e^{-\int a(t) dt} \int [b(t) e^{\int a(\tau) d\tau}] dt + c e^{-\int a(t) dt}} \quad (\text{B.4})$$

ove  $c$  è una costante arbitraria.

► **Esempio B.3** In corrispondenza all'equazione differenziale lineare  $y' + 2y = 3e^t$  si ottiene la famiglia di soluzioni  $y = ce^{-2t} + e^t$ , illustrata in Figura B.2 per alcuni valori della costante  $c$ .

Per l'equazione differenziale lineare  $y' + 2ty = (1 + t^2)^{-1}$  si ottiene

$$y = ce^{-t^2} + e^{-t^2} \int e^{t^2} (1 + t^2)^{-1} dt$$

In questo caso l'integrale non è esprimibile in termini di funzioni semplici (polinomi, esponenziali, logaritmi, funzioni trigonometriche, ...). Per uno studio quantitativo delle soluzioni è, quindi, necessario il ricorso a procedimenti numerici. ■

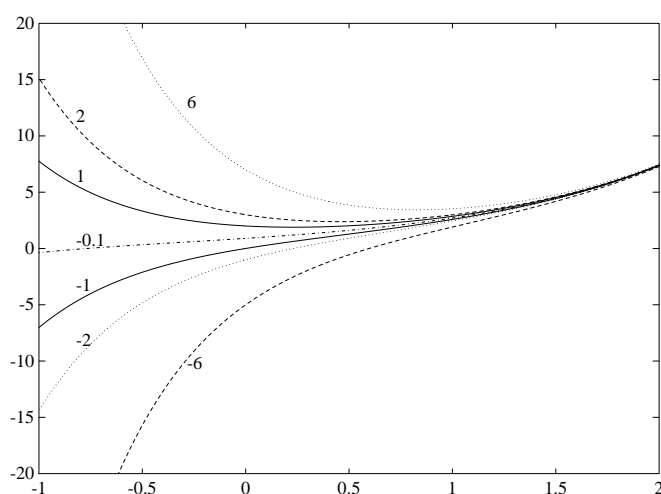


Figura B.2: Soluzioni dell'equazione differenziale  $y' + 2y = 3e^t$  corrispondenti ad alcuni valori della costante  $c$ .

Se si scambia il ruolo delle variabili  $t$  e  $y$ , l'equazione lineare assume la forma

$$\frac{dt}{dy} + A(y)t = B(y)$$

con  $A(y)$ ,  $B(y)$  funzioni continue assegnate. Tale equazione può essere riscritta nella forma

$$[B(y) - tA(y)] \frac{dy}{dt} = 1 \quad (\text{B.5})$$

Se ne deduce che un'equazione della forma (B.5) può essere risolta trattandola come un'equazione lineare in  $t$  e assumendo  $y$  come variabile indipendente.

► **Esempio B.4** La seguente equazione

$$(\cos^2 y - t \sec y)y' = \sin y$$

è lineare della forma

$$\sin y \frac{dt}{dy} + t \sec y = \cos^2 y$$

per la quale si ha la seguente soluzione generale

$$t \tan y = \sin y + c$$

■

Un modo differente per risolvere l'equazione (B.3) consiste nel considerare dapprima la corrispondente equazione lineare *omogenea*

$$\frac{dy}{dt} + a(t)y = 0$$

che ha come soluzione generale la funzione  $y = ce^{-\int a(t) dt}$ . Si cerca quindi la funzione  $u(t)$  tale che  $y = u(t)e^{-\int a(t) dt}$  sia soluzione dell'equazione (B.3). Si trova  $u' = b(t)e^{-\int a(t) dt}$ . È immediato verificare che risolvendo rispetto a  $u$  si arriva alla soluzione generale precedentemente ottenuta. La procedura ora enunciata è interessante, in quanto, come vedremo successivamente, essa è generalizzabile alle equazioni lineari di ordine superiore. Osserviamo che la soluzione (B.4) si presenta come la somma di una soluzione particolare dell'equazione non omogenea, corrispondente al termine  $e^{-\int a(t) dt} \int [b(t)e^{\int a(t) dt}] dt$  e della soluzione generale dell'equazione omogenea, data dal termine  $ce^{-\int a(t) dt}$ .

### B.3 Equazione di Bernoulli

Un'equazione del tipo

$$\boxed{\frac{dy}{dt} + p(t)y = q(t)y^a} \quad (\text{B.6})$$

ove  $p(t), q(t)$  sono due funzioni continue assegnate e  $a$  un numero reale fissato è chiamata *equazione di Bernoulli*. In particolare per  $a = 0$  si ha un'equazione lineare e per  $a = 1$  un'equazione separabile. Il caso  $a = 2$  contiene come caso particolare l'*equazione logistica*, data da  $y' = \alpha y - \beta y^2$ , con  $\alpha$  e  $\beta$  costanti positive.

Per  $a \neq 0, 1$ , dividendo l'equazione per  $y^a$ , si ottiene

$$y^{-a}y' + py^{1-a} = q$$

Si introduce, quindi, la nuova variabile  $z = y^{1-a}$ , per la quale si ha  $z' = (1-a)y^{-a}y'$ . Si ottiene, allora, la seguente equazione nella variabile  $z$

$$z' + p(1-a)z = (1-a)q$$

che è lineare e può essere integrata nel modo visto nel paragrafo precedente.

► **Esempio B.5** Data l'equazione

$$y' = ty^2 + 2ty$$

e posto  $z = y^{-1}$ , si ha

$$z' + 2tz = -t \Rightarrow ze^{t^2} = -\frac{1}{2}e^{t^2} + c \Rightarrow 1/y = -1/2 + ce^{-t^2}$$

■

## B.4 Equazione di Riccati

L'equazione differenziale

$$\boxed{\frac{dy}{dt} = P(t)y^2 + Q(t)y + R(t)} \quad (\text{B.7})$$

con  $P(t)$ ,  $Q(t)$  e  $R(t)$  funzioni continue assegnate, è detta un'*equazione di Riccati* o un'*equazione generalizzata di Riccati*<sup>2</sup>. Le equazioni della forma (B.7) sono collegate con la teoria delle funzioni di Bessel; ricordiamo, inoltre, il loro interesse nell'ambito della geometria proiettiva differenziale, del calcolo delle variazioni e del controllo ottimale (cfr. Capitolo 14).

Quando  $P = 0$  si ha come caso particolare un'equazione lineare e per  $R = 0$  un'equazione di Bernoulli. In caso contrario, si può costruire l'integrale generale quando si conosce un integrale particolare. In effetti, sia per ipotesi  $y = u$  un integrale particolare di (B.7). La sostituzione

$$y = u + \frac{1}{z}$$

dà origine, dopo opportuna semplificazione, alla seguente equazione lineare

$$z' + (2Pu + Q)z + P = 0$$

che può essere integrata nel modo visto in precedenza.

► **Esempio B.6** L'equazione

$$t^2y' = t^2y^2 + ty - 3$$

ha come soluzione particolare la funzione  $y = 1/t$ . La sostituzione  $y = 1/t + 1/z$  porta all'equazione lineare

$$t^2z' + 3tz = -t^2$$

---

<sup>2</sup>Jacopo Francesco Riccati, matematico veneziano (1676–1754), considerò (*Acta eruditorum*, 1724) l'equazione particolare  $y'(t) + t^{-n}y^2(t) - nt^{m+n-1} = 0$ , con  $m$  e  $n$  costanti assegnate. Il nome di equazione di Riccati all'equazione (B.7) venne dato da d'Alembert nel 1763.

che ha come integrale generale  $t^3 z = c - t^4/4$ . Pertanto, la soluzione dell'equazione di partenza è data da (cfr. Figura B.3)

$$y = \frac{1}{t} + \frac{t^3}{c - \frac{1}{4}t^4}$$

■

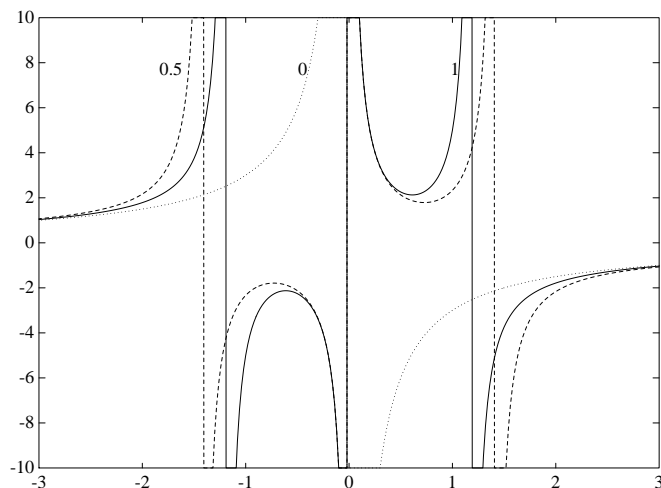


Figura B.3: Soluzioni dell'equazione differenziale di Riccati  $t^2 y' = t^2 y^2 + ty - 3$  corrispondenti ad alcuni valori della costante  $c$ .

Terminiamo ricordando la seguente interessante proprietà delle soluzioni di un'equazione di Riccati. Se  $m_1, m_2, m_3, m_4$  sono quattro numeri, introduciamo le notazioni

$$\{m_1, m_2, m_3\} = \frac{m_3 - m_1}{m_3 - m_2}$$

$$\{m_1, m_2, m_3, m_4\} = \{m_1, m_2, m_3\} \{m_2, m_1, m_4\}$$

Il rapporto  $\{m_1, m_2, m_3, m_4\}$  è detto *birapporto*, o rapporto anarmonico dei numeri  $m_1, m_2, m_3, m_4$ . Si può, allora, dimostrare il seguente risultato.

**Proposizione B.1** *Se  $y_i(t)$  ( $i = 1, \dots, 4$ ) sono quattro soluzioni dell'equazione (B.7) con  $y_3(t) \neq y_2(t)$  e  $y_4(t) \neq y_1(t)$  su un sottointervallo  $I_0$ , allora il birapporto  $\{y_1(t), y_2(t), y_3(t), y_4(t)\}$  è costante su  $I_0$ .*

## B.5 Equazione omogenea

Ricordiamo che una funzione  $\phi(\xi, \eta)$  è detta *omogenea* di grado  $n$  nelle variabili  $\xi, \eta$ , se per ogni  $(\xi, \eta) \neq (0, 0)$  e ogni  $k \neq 0$  si ha

$$\phi(k\xi, k\eta) \equiv k^n \phi(\xi, \eta)$$



Ad esempio, la funzione  $\xi^2 - \xi\eta + 2\eta^2$  è omogenea di grado 2, la funzione  $(\xi - 3\eta)^{1/2}$  di grado 1/2, e la funzione  $\ln \xi - \ln \eta$  di grado zero.

L'equazione

$$\frac{dy}{dt} = f(t, y) \quad (\text{B.8})$$

è detta *omogenea* se  $f(t, y)$  è omogenea di grado zero. Operando la sostituzione  $y = tz$ , da cui  $y' = z + tz'$ , si ha

$$z + tz' = f(t, tz) = f(1, z) \quad (\text{B.9})$$

in quanto, dall'identità  $f(k\xi, k\eta) = f(\xi, \eta)$ , posto  $\xi = 1, \eta = z$  e  $k = t$ , si ha  $f(t, tz) = f(1, z)$ . In altre parole, per un'equazione omogenea la variabile interessante è il rapporto  $y/t$ . L'equazione differenziale (B.9) è separabile con soluzione

$$\int \frac{dz}{f(1, z) - z} = \ln t + c$$

Per completare la risoluzione si elimina  $z$  tra questa equazione e  $y = tz$ . Osserviamo che per i valori  $k$  che sono radici dell'equazione  $k = f(1, k)$  la funzione  $y = kt$  può essere una soluzione singolare dell'equazione differenziale assegnata.

► **Esempio B.7** Come illustrazione, consideriamo l'equazione differenziale

$$y' = \frac{t^2 + y^2}{2ty}, \quad t \neq 0, \quad y \neq 0 \quad (\text{B.10})$$

Introducendo la nuova variabile  $z$  tale che  $y = tz$ , si ottiene l'equazione

$$tz' + z = \frac{t^2 + t^2z^2}{2zt^2} = \frac{1 + z^2}{2z}$$

Separando le variabili, si ottiene

$$\frac{2z}{1 - z^2} z' = \frac{1}{t}, \quad \Rightarrow \quad -\ln |1 - z^2| = \ln |t| + c, \quad z \neq \pm 1$$

ove  $c$  è la costante di integrazione. Ritornando alle variabili originarie  $y$  e  $t$  ponendo  $z = y/t$ , si ha

$$\ln |t| + \ln \left| 1 - \frac{y^2}{t^2} \right| = \ln \left| t \left( 1 - \frac{y^2}{t^2} \right) \right| = -c, \quad y \neq \pm t$$

da cui

$$\left| t \left( 1 - \frac{y^2}{t^2} \right) \right| = e^{-c}, \quad y \neq \pm t \quad (\text{B.11})$$

Sostituendo direttamente in (B.10), si vede che ciascuna delle rette  $y = \pm t$  è una soluzione di (B.10) su ogni intervallo non contenente  $t = 0$ . Tenendo conto di questo risultato, si può eliminare il valore assoluto da (B.11), sostituire  $e^{-c}$  mediante una costante reale  $k$ , e ottenere la seguente equazione che definisce le soluzioni di (B.10)

$$t^2 - y^2 = kt, \quad t \neq 0, \quad y \neq 0$$

che rappresentano per ogni valore di  $k$  una iperbole. ■

► **Esempio B.8** *Attraversamento di un fiume.* Con riferimento alla Figura B.4, si cerca la traiettoria di un battello che parte dal punto  $A = (a, 0)$  e attraversa un fiume nel quale vi è una corrente  $s$ , di intensità costante  $s$ . Si suppone che anche l'intensità  $v$  della velocità  $\mathbf{v}$  del battello sia costante.

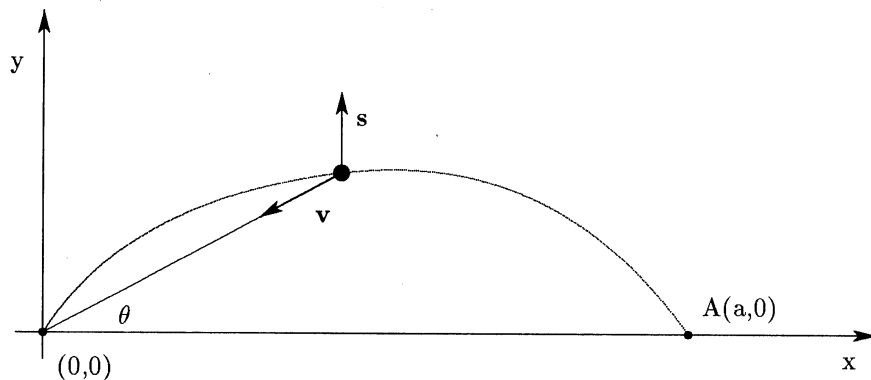


Figura B.4: Traiettoria nell'attraversamento di un fiume soggetto a una corrente di intensità costante  $s$ .

Indicando con  $(x(t), y(t))$  la posizione del battello al tempo  $t$  nel sistema di riferimento indicato in figura, si hanno le seguenti equazioni di moto

$$\begin{aligned}\frac{dx(t)}{dt} &= -v \cos \theta = \frac{-vx}{\sqrt{x^2 + y^2}} \\ \frac{dy(t)}{dt} &= -v \sin \theta + s = \frac{-vy}{\sqrt{x^2 + y^2}} + s\end{aligned}$$

Dividendo le equazioni, si ottiene la seguente equazione differenziale del primo ordine in  $x$  e  $y$  e nella quale il tempo  $t$  non compare esplicitamente

$$\frac{dy}{dx} = \frac{vy - s\sqrt{x^2 + y^2}}{vx} \equiv: f(x, y)$$

Si verifica facilmente che  $f(kx, ky) = f(x, y)$ , per cui l'equazione precedente è omogenea. Se poniamo  $y = zx$ , e per brevità  $r = s/v$ , si ottiene

$$\frac{dy}{dx} = \frac{dz}{dx} x + z = \frac{zx - r\sqrt{x^2 + z^2x^2}}{x} = z - r\sqrt{1 + z^2}$$

da cui la seguente equazione a variabili separabili

$$\frac{1}{\sqrt{1 + z^2}} \frac{dz}{dx} = \frac{-r}{x}$$

le cui soluzioni sono definite dall'equazione  $\ln[z + \sqrt{1 + z^2}] = -r \ln x + c$ , ove  $c$  è una costante di integrazione. Imponendo che per  $x = a$  si abbia  $z = y = 0$  si ottiene  $c = r \ln a$ , e pertanto

$$\ln[z + \sqrt{1 + z^2}] = \ln \left(\frac{x}{a}\right)^{-r} \Rightarrow z + \sqrt{1 + z^2} = \left(\frac{x}{a}\right)^{-r} \Rightarrow z = \frac{1}{2} \left[ \left(\frac{x}{a}\right)^{-r} - \left(\frac{x}{a}\right)^r \right]$$

Poiché  $z = y/x$ , l'equazione della traiettoria del battello è la seguente

$$y = \frac{a}{2} \left[ \left( \frac{x}{a} \right)^{1-r} - \left( \frac{x}{a} \right)^{1+r} \right]$$

In Figura B.5 sono rappresentate le traiettorie corrispondenti ad alcuni valori di  $r = s/v$ , dalle quali si vede che vi è possibilità di raggiungere il punto  $(0,0)$  *solo* se  $r < 1$ . ■

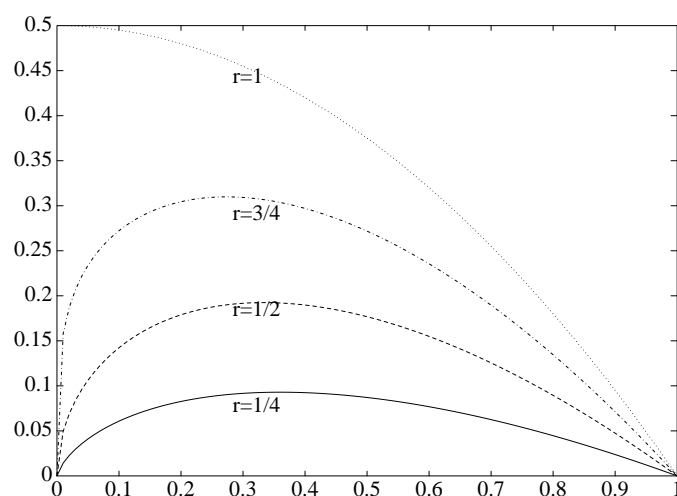


Figura B.5: Traiettorie corrispondenti a diversi valori di  $r = s/v$ .

## B.6 Equazione esatta

Un'equazione differenziale della forma

$$\boxed{M(t, y) + N(t, y)y' = 0} \quad (\text{B.12})$$

ove  $M$  e  $N$  sono due funzioni differenziabili con continuità in una regione  $R$  del piano  $ty$ , è detta *esatta* nella regione  $R$  se vi è una funzione differenziabile con continuità  $F(t, y)$  tale che

$$\frac{\partial F}{\partial t} = M(t, y) \quad \text{e} \quad \frac{\partial F}{\partial y} = N(t, y) \quad \forall (t, y) \in R$$

Per esempio, l'equazione

$$2ty + (1 + t^2)y' = 0 \quad (\text{B.13})$$

è esatta, con  $F(t, y) = y + t^2y$ .

Le soluzioni di un'equazione esatta possono essere caratterizzate implicitamente nel seguente modo. Una funzione  $y(t)$  per  $a < t < b$  è una soluzione di  $M + Ny' = 0$  se e solo se  $(t, y(t))$  appartiene a  $R$  per  $a < t < b$  ed inoltre, per un particolare valore della costante  $c$ , si ha

$$F(t, y(t)) = c, \quad t \in (a, b)$$

Il risultato segue dalla regola di derivazione delle funzioni composte

$$0 = \frac{d}{dt}F(t, y(t)) = F_t + F_y y' = M(t, y(t)) + N(t, y(t))y'(t)$$

Ad esempio, per l'equazione (B.13) le soluzioni  $y(t)$  soddisfano la relazione

$$y(t) + t^2 y'(t) = c \Rightarrow y = \frac{c}{1 + t^2}$$

Ricordiamo il seguente *test* di esattezza dell'equazione differenziale (B.12).

**Proposizione B.2** *Siano  $M(t, y)$  e  $N(t, y)$  due funzioni differenziabili con continuità nel rettangolo  $R = (a, b) \times (c, d)$ . Allora,  $M + Ny' = 0$  è esatta se e solo se*

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial t} \quad \text{in } R$$

## B.7 Equazione di Clairaut

Per ogni funzione  $f$  definita e continua sull'asse reale, l'equazione differenziale

$$\boxed{y = ty' + f(y')} \quad (\text{B.14})$$

è chiamata *equazione di Clairaut*<sup>3</sup>. Si verifica immediatamente che l'integrale generale di tale equazione è fornito dalla seguente famiglia di rette

$$y = ct + f(c) \quad (\text{B.15})$$

al variare della costante  $c$ . Inoltre, nel caso in cui la funzione  $f$  sia derivabile, esiste una soluzione particolare che non può essere ottenuta dall'integrale generale (B.15) per un valore particolare della costante  $c$ . Per questo motivo, tale soluzione è detta *singolare*. Usando la notazione  $y' = p$ , si ha

$$y = tp + f(p)$$

Derivando rispetto a  $t$ , si ottiene

$$p = p + [t + f'(p)] \frac{dp}{dt}$$

<sup>3</sup>A. C. Clairaut (1713–1765); introdusse il procedimento di integrazione dell'equazione (B.14) nell'ambito dello studio della cometa di Halley.

da cui o  $dp/dt = 0$ , oppure  $t = -f'(p)$ . La soluzione  $dp/dt = 0$  corrisponde a  $p = c$  e porta alla soluzione generale (B.15). La seconda soluzione fornisce

$$t = -f'(p), \quad y = f(p) - pf'(p)$$

che rappresenta l'equazione parametrica della soluzione singolare. Si può dimostrare che l'integrale generale rappresenta la famiglia delle tangenti alla curva corrispondente alla soluzione singolare.

► **Esempio B.9** Come illustrazione, consideriamo l'equazione

$$y = y't + (y')^3$$

La soluzione generale è data da  $y = ct + c^3$ , mentre la soluzione singolare è rappresentata dalle equazioni

$$t = -3p^2; \quad y = pt + p^3$$

Eliminando il parametro  $p$ , si ottiene l'equazione  $4t^3 + 27y^2 = 0$ . L'integrale singolare e alcuni elementi della soluzione generale sono rappresentati in Figura B.6. ■

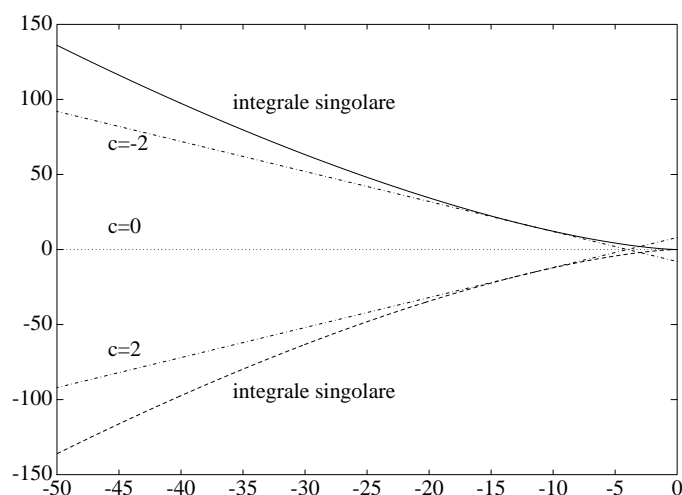


Figura B.6: Soluzione singolare e integrale generale dell'equazione di Clairaut  $y = ty' + (y')^3$ .

◆ **Esercizio B.1** Risolvere le seguenti equazioni

1.  $y' = 2te^y(1+t^2)^{-1}$ .
2.  $y' + y = e^t y^{1/2}$ .
3.  $y' + t^{-1}y = y^{-4}$ .

4.  $t \tan(y/t) + yy' = 0$ .
5.  $(3t + y - 2)y' = y - t + 2$ .
6.  $t^2 y' = (y - 1)(t + y - 1)$ .
7.  $(e^t \sin y - 2y \sin t) + (y^2 + e^t \cos y + 2 \cos t)y' = 0$ .
8.  $y = ty' + \ln y'$ .
9.  $y = ty' + y'/(1 + y')$ .

◆ **Esercizio B.2** Trovare la soluzione  $y(t)$  del seguente problema a valori iniziali

$$y \frac{dy}{dt} + (1 + y^2) \sin t = 0, \quad y(0) = 1$$

## B.8 Equazioni lineari del secondo ordine

In questo paragrafo considereremo i risultati di base relativi alla seguente equazione differenziale

$$\boxed{y''(t) + a(t)y' + b(t)y = f(t)} \quad (\text{B.16})$$

ove i coefficienti  $a(t)$  e  $b(t)$  e il termine noto  $f(t)$  sono funzioni continue su un intervallo assegnato  $I$ . L'equazione (B.16) è detta *equazione differenziale del secondo ordine lineare*. Quando  $f(t)$  non è identicamente nulla, l'equazione è detta *non omogenea*, e *omogenea* nel caso contrario. Infine, quando i coefficienti  $a$  e  $b$  sono costanti l'equazione è detta *equazione differenziale del secondo ordine con coefficienti costanti*.

**Definizione B.1** Se  $a(t), b(t), f(t)$  sono funzioni continue su un intervallo  $I$ , una funzione  $y(t)$  è una soluzione per l'equazione (B.16) su  $I$ , se  $y(t)$  è continua insieme alle derivate fino al secondo ordine su  $I$  e soddisfa l'equazione (B.16) per ogni  $t$  sull'intervallo  $I$ .

Lo studio dell'equazione (B.16) è importante per diversi motivi. Innanzitutto, essa è alla base di modelli matematici interessanti (cfr. Capitolo 7); inoltre, la conoscenza del comportamento della soluzione dell'equazione lineare può essere utile nello studio delle soluzioni di equazioni più generali. Infine, lo studio dell'equazione (B.16) serve come modello per l'analisi di equazioni lineari di ordine più elevato. Incominceremo dal seguente importante risultato.

**Proposizione B.3** (Problema a valori iniziali) Se  $a(t), b(t), f(t)$  sono funzioni continue su un intervallo  $I$ , per ogni valore  $t_0 \in I$  e ogni coppia di costanti  $y_0, v_0$  il problema a valori iniziali

$$\begin{cases} y'' + a(t)y' + b(t)y = f \\ y(t_0) = y_0, \quad y'(t_0) = v_0 \end{cases} \quad (\text{B.17})$$

ha una ed una sola soluzione  $y(t)$  in  $C^2(I)$ .

Un modo analitico per risolvere il problema a valori iniziali (B.17) consiste nel costruire l'integrale generale dell'equazione (B.16) e quindi nel determinare le costanti in maniera che siano verificate le condizioni iniziali imposte. Ad esempio, come vedremo nel seguito, l'equazione differenziale  $y'' + 4y = 4t$  ha come integrale generale la famiglia di funzioni  $y = c_1 \sin 2t + c_2 \cos 2t + t$ , con  $c_1, c_2$  costanti arbitrarie. La soluzione che verifica le condizioni iniziali  $y(0) = 0$ ,  $y'(0) = 0$  corrisponde allora ai valori delle costanti  $c_2 = 1$ ,  $c_1 = -1/2$ .

Dal punto di vista analitico è quindi importante la costruzione dell'integrale generale dell'equazione (B.16). Nel seguito considereremo tale problema, con riferimento in particolare alle equazioni a coefficienti costanti. A tale scopo, incominceremo a studiare l'equazione omogenea

$$L(y) := y'' + ay' + by = 0 \quad (\text{B.18})$$

con  $a, b$  costanti assegnate. Per tale equazione possiamo considerare le soluzioni definite su tutto l'asse reale  $\mathbb{R}$ .

Le soluzioni dell'equazione (B.18), e più in generale di una equazione lineare omogenea, non necessariamente a coefficienti costanti, verificano la seguente importante proprietà.

**Proposizione B.4** *Se  $y_1(t), y_2(t)$  sono due qualunque soluzioni dell'equazione omogenea (B.18), allora anche la combinazione lineare  $c_1 y_1(t) + c_2 y_2(t)$ , con  $c_1, c_2$  costanti arbitrarie, è una soluzione dell'equazione.*

Si tratta, ora, di vedere quando l'espressione  $c_1 y_1(t) + c_2 y_2(t)$  fornisce tutte le soluzioni dell'equazione (B.18). La risposta è fornita dal seguente risultato.

**Proposizione B.5** *Siano  $y_1(t), y_2(t)$  due soluzioni dell'equazione differenziale lineare omogenea (B.18), soddisfacenti per  $t = 0$  la condizione*

$$\begin{vmatrix} y_1(0) & y_2(0) \\ y_1'(0) & y_2'(0) \end{vmatrix} = y_1(0)y_2'(0) - y_2(0)y_1'(0) \neq 0 \quad (\text{B.19})$$

*Allora, tutte le soluzioni dell'equazione (B.18) sono del tipo  $c_1 y_1(t) + c_2 y_2(t)$ , al variare dei parametri  $c_1, c_2$ .*

La dimostrazione, i cui dettagli sono lasciati come esercizio, utilizza l'unicità della soluzione del problema di Cauchy (cfr. Proposizione B.3). Il determinante in (B.19) è anche detto il *Wronskiano*<sup>4</sup> relativo alle funzioni  $y_1(t), y_2(t)$  e indicato, usualmente, con  $W(y_1(t), y_2(t))$ .

Ricordiamo che due funzioni  $f_1(t), f_2(t)$  continue su un intervallo  $I$  sono dette *linearmente dipendenti* se e solo se esistono due costanti  $c_1, c_2$ , non ambedue nulle, tali che

$$c_1 f_1(t) + c_2 f_2(t) = 0 \quad \text{per ogni } t \in I$$

<sup>4</sup>H. Wronski (1778–1853) fu successivamente militare, matematico, filosofo.

In altre parole,  $f_1, f_2$  sono linearmente dipendenti se e solo se una delle due funzioni è un multiplo dell'altra su  $I$ . La coppia di funzioni  $f_1, f_2$  sono dette *linearmente indipendenti* se e solo se esse non sono linearmente dipendenti.

Il risultato espresso nella Proposizione (B.5) ha allora la seguente interpretazione. Le soluzioni  $y_1(t), y_2(t)$  dell'equazione lineare omogenea (B.18) sono linearmente indipendenti se e solo se il Wronskiano  $W(y_1(t), y_2(t))$  è diverso dallo zero per un valore particolare di  $t$  (nella proposizione si è assunto  $t = 0$ ).

In definitiva, la costruzione dell'integrale generale dell'equazione (B.18) è ricondotta a quella della ricerca di due soluzioni particolari  $y_1(t), y_2(t)$  che verificano la condizione (B.19). A tale scopo, cerchiamo soluzioni della seguente forma

$$y(t) = e^{\lambda t}$$

con  $\lambda$  parametro da determinare. Essendo  $y' = \lambda e^{\lambda t}$  e  $y'' = \lambda^2 e^{\lambda t}$ , si ha

$$L(e^{\lambda t}) = \lambda^2 e^{\lambda t} + a \lambda e^{\lambda t} + b e^{\lambda t} = e^{\lambda t}(\lambda^2 + a \lambda + b)$$

Pertanto, l'equazione differenziale  $L(e^{\lambda t}) = 0$  è soddisfatta se e solo se  $\lambda$  è una soluzione della seguente equazione algebrica di secondo grado

$$\lambda^2 + a \lambda + b = 0 \quad (\text{B.20})$$

che è nota come *equazione caratteristica* associata all'equazione differenziale (B.18). Indichiamo con  $\lambda_1, \lambda_2$  le radici

$$\lambda_1 = \frac{-a - \sqrt{a^2 - 4b}}{2}, \quad \lambda_2 = \frac{-a + \sqrt{a^2 - 4b}}{2}$$

Se il discriminante  $\Delta = a^2 - 4b$  non è negativo, le radici sono *reali*. In particolare, quando  $\Delta = 0$  si ha  $\lambda_1 = \lambda_2$ . Quando  $\Delta < 0$ , le radici sono *complesse coniugate*  $\alpha \pm i\beta$ , con

$$\alpha = -\frac{a}{2}, \quad \beta = \frac{\sqrt{-\Delta}}{2} = \frac{\sqrt{4b - a^2}}{2}$$

Verifichiamo la condizione (B.19) nel caso in cui  $\Delta > 0$ , ossia nel caso in cui le radici sono reali e distinte. In effetti, si ha

$$W(e^{\lambda_1 t}, e^{\lambda_2 t}) = \begin{vmatrix} e^{\lambda_1 t} & e^{\lambda_2 t} \\ \lambda_1 e^{\lambda_1 t} & \lambda_2 e^{\lambda_2 t} \end{vmatrix} = (\lambda_2 - \lambda_1)e^{(\lambda_1 + \lambda_2)t}$$

Pertanto, quando  $\Delta > 0$  l'integrale generale dell'equazione (B.18) è fornito dalla seguente famiglia di funzioni

$$y(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$



Il risultato può essere esteso agli altri due casi nel seguente modo. Più precisamente, se  $\Delta = 0$  e  $\lambda_1$  è la radice doppia, l'integrale generale si scrive nella maniera seguente

$$y(t) = c_1 e^{\lambda_1 t} + c_2 t e^{\lambda_1 t}$$

Infine, quando le radici sono complesse coniugate si hanno come soluzioni linearmente indipendenti le funzioni  $e^{\alpha t} \cos \beta t$  e  $e^{\alpha t} \sin \beta t$ , e pertanto l'integrale generale è il seguente

$$y(t) = c_1 e^{\alpha t} \cos \beta t + c_2 e^{\alpha t} \sin \beta t$$

### B.8.1 Equazioni lineari non omogenee particolari

L'integrale generale dell'equazione (B.16) può essere costruito sommando all'integrale generale dell'equazione omogenea un integrale particolare dell'equazione non omogenea. In questo paragrafo indicheremo come costruire tale integrale nel caso in cui la  $f(t)$  assuma forme particolari.

Ad esempio, quando la funzione  $f(t)$  è un *polinomio* si cerca come integrale particolare un polinomio, con coefficienti da determinare. A scopo illustrativo, consideriamo l'equazione

$$y'' + y = t^3 - 2$$

Indicando con  $\bar{y}(t)$  l'integrale particolare che cerchiamo, poniamo

$$\bar{y}(t) = a_3 t^3 + a_2 t^2 + a_1 t + a_0$$

per cui

$$\bar{y}(t)'' + \bar{y}(t) = a_3 t^3 + a_2 t^2 + (a_1 + 6a_3)t + (a_0 + 2a_2)$$

Imponendo l'identità tra il secondo membro dell'equazione precedente e il termine noto  $f(t) = t^3 - 2$ , si ha

$$a_3 = 1; \quad a_2 = 0; \quad a_1 + 6a_3 = 0; \quad a_0 + 2a_2 = -2$$

da cui  $\bar{y}(t) = t^3 - 6t - 2$ . In definitiva, utilizzando i risultati del paragrafo precedente, si ha che l'integrale generale dell'equazione proposta è dato da

$$y(t) = c_1 \cos t + c_2 \sin t + t^3 - 6t - 2$$

Quando  $f(t)$  è una combinazione lineare delle funzioni  $\sin t$  e  $\cos t$ , si cerca  $\bar{y}(t)$  della forma  $\bar{y}(t) = A \sin t + B \cos t$ , con coefficienti da determinare. Ad esempio, per l'equazione

$$y'' + y' + 2y = 2 \cos t$$

si ricava

$$\bar{y}'' + \bar{y}' + 2\bar{y} = (A - B) \sin t + (A + B) \cos t$$

da cui  $A - B = 0$ ,  $A + B = 2$ , e quindi  $A = B = 1$  e  $\bar{y}(t) = \sin t + \cos t$ . Si può vedere che in questo caso l'integrale generale è fornito dalla formula

$$y(t) = e^{-t/2} \left( c_1 \cos \frac{\sqrt{7}}{2} t + c_2 \sin \frac{\sqrt{7}}{2} t \right) + \sin t + \cos t$$

Se il termine noto è del tipo  $Ae^{Bt}$ , si cerca ancora una soluzione particolare dello stesso tipo; è necessario, tuttavia, che  $e^{Bt}$  non sia soluzione dell'equazione omogenea associata. Consideriamo, ad esempio l'equazione

$$y'' - 5y' = 3e^{2t}$$

Posto  $\bar{y}(t) = Ae^{2t}$ , si ottiene

$$\bar{y}'' - 5\bar{y}' = 4Ae^{2t} - 10Ae^{2t} = -6Ae^{2t}$$

da cui  $A = -1/2$  e  $\bar{y}(t) = -e^{2t}/2$ . L'integrale generale è allora

$$y(t) = c_1 + c_2 e^{5t} - \frac{1}{2} e^{2t}$$

Per terminare, segnaliamo un metodo generale per costruire un integrale particolare dell'equazione non omogenea a partire dall'integrale generale della omogenea. Tale metodo, noto come *metodo della variazione delle costanti*, consiste nella ricerca di  $\bar{y}(t)$  della forma

$$\bar{y}(t) = c_1(t)y_1(t) + c_2(t)y_2(t)$$

e le funzioni  $c_1(t)$  e  $c_2(t)$  sono determinate in modo che siano verificate le seguenti condizioni

$$\begin{aligned} c_1'(t)y_1(t) + c_2'(t)y_2(t) &= 0 \\ c_1'(t)y_1'(t) + c_2'(t)y_2'(t) &= f(t) \end{aligned}$$

che rappresentano un sistema lineare nelle incognite  $c_1'(t)$  e  $c_2'(t)$ . Ad esempio, per l'equazione differenziale

$$y'' + y = \frac{1}{\sin t}$$

si cerca l'integrale particolare della forma  $\bar{y}(t) = c_1(t) \cos t + c_2(t) \sin t$ . Procedendo nella maniera indicata e risolvendo il sistema lineare nelle funzioni  $c_1'(t)$  e  $c_2'(t)$ , si trova  $c_1' = -1$ ,  $c_2' = \cos t / \sin t$ . Calcolando una primitiva di tali funzioni, si trova come integrale generale dell'equazione data la seguente famiglia di funzioni

$$y(t) = c_1 \cos t + c_2 \sin t - t \cos t + \sin t \ln |\sin t|$$

### B.8.2 Sistemi differenziali lineari del primo ordine

Consideriamo, come esemplificazione, il seguente sistema lineare del primo ordine, in due equazioni e due incognite a coefficienti costanti.

$$\begin{cases} y_1'(t) = ay_1(t) + by_2(t) + f_1(t) \\ y_2'(t) = cy_1(t) + dy_2(t) + f_2(t) \end{cases} \quad (\text{B.21})$$

ove  $a, b, c, d$  sono i *coefficienti* del sistema e le funzioni  $f_1(t), f_2(t)$  sono i *termini noti*. Supporremo tali funzioni derivabili con continuità su un intervallo  $I$ . Nel caso in cui  $f_1(t) = f_2(t) \equiv 0$ , il sistema è detto *omogeneo*. Una *soluzione* del sistema (B.21) è una coppia di funzioni derivabili  $[y_1(t), y_2(t)]^T$  che verificano le due equazioni per ogni  $t \in I$ .

Osserviamo che se  $[y_1(t), y_2(t)]^T$  è una soluzione, allora le due funzioni  $y_1(t), y_2(t)$  sono derivabili due volte. Derivando allora la prima equazione, si ottiene

$$y_1'' = ay_1' + by_2' + f_1'$$

da cui, sostituendo il valore  $y_2'$  ricavato dalla seconda equazione del sistema (B.21)

$$y_1'' = ay_1' + b(cy_1 + dy_2 + f_2) + f_1' = ay_1' + bcy_1 + bdy_2 + bf_2 + f_1'$$

Dalla prima equazione si può ricavare  $by_2$  (nell'ipotesi che  $b \neq 0$ ; lasciamo come esercizio la considerazione del caso in cui  $b = 0$  e/o  $c = 0$ ). Si ottiene

$$y_1'' = ay_1' + bcy_1 + d(y_1' - ay_1 - f_1) + bf_2 + f_1'$$

da cui la seguente equazione lineare del secondo ordine a coefficienti costanti

$$y_1'' - (a + d)y_1' + (ad - bc)y_1 = bf_2 - df_1 + f_1' \quad (\text{B.22})$$

che può essere risolta con i metodi visti nei paragrafi precedenti.

► **Esempio B.10** La soluzione generale  $\mathbf{y}(t) = [y_1, y_2]^T$  del sistema

$$\begin{cases} y_1'(t) = 5y_1(t) + 3y_2(t) \\ y_2'(t) = -y_1(t) + y_2(t) \end{cases}$$

è fornita dalla relazione

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{2t} + c_2 \begin{bmatrix} 3 \\ -1 \end{bmatrix} e^{4t}$$

A partire dall'integrale generale si può vedere che l'integrale particolare che soddisfa, ad esempio, le condizioni iniziali  $y_1(0) = 1, y_2(0) = 2$  si ottiene per  $c_1 = -7/2$  e  $c_2 = 3/2$ . ■

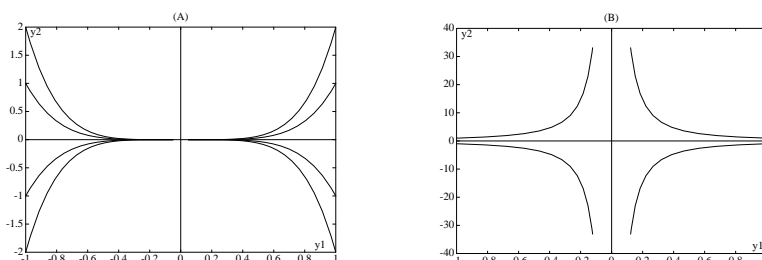


Figura B.7: (A) Traiettorie relative al sistema  $y_1' = -y_1$ ,  $y_2' = -5y_2$ . Il punto di equilibrio  $(0, 0)$  è un *nodo stabile*. (B) Traiettorie relative al sistema  $y_1' = -3y_1$ ,  $y_2' = 5y_2$ . Il punto di equilibrio  $(0, 0)$  è un punto *sella instabile*.

Osserviamo che l'equazione caratteristica dell'equazione (B.22), data da

$$\lambda^2 - (a + d)\lambda + (ad - bc) = 0$$

corrisponde al *polinomio caratteristico* della matrice dei coefficienti del sistema<sup>5</sup>

$$\mathbf{A} - \lambda \mathbf{I} := \begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} = 0$$

Il comportamento, quindi, delle traiettorie  $y_1(t)$ ,  $y_2(t)$  è determinato dagli autovalori della matrice  $\mathbf{A}$ . In particolare, tali autovalori sono indicatori del tipo di *stabilità* che presenta il sistema dinamico descritto dal sistema omogeneo associato al sistema differenziale (B.21) nei *punti di equilibrio*. Ricordiamo che i punti di equilibrio, o punti critici, corrispondono alle soluzioni del sistema lineare  $ay_1 + by_2 = 0$ ,  $cy_1 + dy_2 = 0$ .

Ad esempio, se  $\lambda_2 < 0 < \lambda_1$ , il punto di equilibrio è un *punto sella instabile*, cioè per  $t \rightarrow \infty$  vi è almeno una traiettoria che cresce illimitatamente. Viceversa, se  $\lambda_2 < \lambda_1 < 0$  si ha un *nodo asintoticamente stabile*. Quando  $\lambda_1 = \alpha + i\beta = \bar{\lambda}_2$  con  $\alpha < 0$  e  $\beta \neq 0$  si ha un *fuoco*, mentre se  $\lambda_1 = i\beta = \bar{\lambda}_2$ , con  $\beta \neq 0$ , si ha un *centro*. Per un'illustrazione si vedano le Figure B.7 e B.8.

### Stabilità per sistemi non lineari

Lo studio del sistema lineare (B.21) può fornire indicazioni sul comportamento qualitativo delle soluzioni di un sistema non lineare  $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ , con  $\mathbf{y} \in \mathbb{R}^n$  e  $\mathbf{f}(\mathbf{y}) = [f_1, f_2, \dots, f_n]^T$ . Esaminiamo come esempio illustrativo il seguente sistema non lineare

$$\begin{cases} y_1'(t) = ay_1(t) - by_1(t)y_2(t) \\ y_2'(t) = cy_1(t)y_2(t) - dy_2(t) \end{cases} \quad (\text{B.23})$$

<sup>5</sup>Del resto basta osservare che il vettore  $\mathbf{y} = e^{\lambda t} \mathbf{v}$ , ove  $\mathbf{v}$  è un vettore costante, è soluzione del sistema  $\mathbf{y}' = \mathbf{A}\mathbf{y}$  se e solo se  $d(e^{\lambda t} \mathbf{v})/dt = \lambda e^{\lambda t} \mathbf{v} = \mathbf{A}(e^{\lambda t} \mathbf{v}) = e^{\lambda t} \mathbf{A}\mathbf{v}$ , e quindi se  $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$ , ossia se  $\lambda$  è un autovalore della matrice  $\mathbf{A}$ .

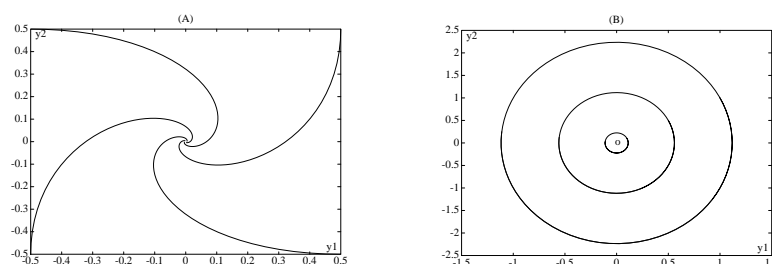


Figura B.8: (A) Traiettorie relative al sistema  $y_1' = -y_1 + y_2$ ,  $y_2' = -y_1 - y_2$ . Il punto di equilibrio  $(0,0)$  è un punto *fuoco* stabile. (B) Traiettorie relative al sistema  $y_1' = y_2$ ,  $y_2' = -4y_1$ . Il punto di equilibrio  $(0,0)$  è un *centro*.

con  $a, b, c, d$  costanti positive, che rappresenta un modello importante nella dinamica delle popolazioni. In tale contesto  $y_1(t)$  rappresenta una popolazione che si comporta da *preda* e  $y_2(t)$  una popolazione di *predatore*. Il sistema è non lineare a causa del termine  $y_1 y_2$ , che è svantaggioso per la preda e vantaggioso per il predatore. Vogliamo studiare il comportamento qualitativo della soluzione  $\mathbf{y}(t) = [y_1(t), y_2(t)]^T$  corrispondente a condizioni iniziali  $\mathbf{y}(0)$  positive. Lasciata da sola, la popolazione preda dovrebbe aumentare  $y_1' = ay_1$ , mentre la popolazione dei predatori dovrebbe diminuire  $y_2' = -dy_2$ . Il numero degli incontri tra loro è proporzionale a  $y_1 y_2$  e ogni incontro dà una opportunità al predatore di recuperare<sup>6</sup>.

Lo studio della stabilità del sistema (B.23) mediante il procedimento della *linearizzazione* si sviluppa attraverso i seguenti punti

1. Si cercano i *punti critici*  $\mathbf{y}^*$  risolvendo  $\mathbf{f}(\mathbf{y}^*) = 0$
2. Si calcola la matrice

$$\mathbf{A} = \begin{bmatrix} \partial f_1 / \partial y_1 & \partial f_1 / \partial y_2 \\ \partial f_2 / \partial y_1 & \partial f_2 / \partial y_2 \end{bmatrix}$$

per  $\mathbf{y} = \mathbf{y}^*$ .

3. Si studia la stabilità locale del sistema esaminando gli autovalori della matrice  $\mathbf{A}$ .

Nel caso del sistema (B.23) si hanno due punti critici:  $\mathbf{y}_{(1)}^* = [0, 0]^T$  e  $\mathbf{y}_{(2)}^* = [d/c, a/b]^T$  e la matrice  $\mathbf{A}$  è data da

$$\begin{bmatrix} a - by_2 & -by_1 \\ cy_2 & cy_1 - d \end{bmatrix}$$

<sup>6</sup>Il modello (B.23) venne proposto da Vito Volterra nel 1920 per spiegare l'aumento di pescecani (i predatori) nel sud del Mediterraneo a seguito della diminuzione della pesca di sardine (le prede) durante la prima guerra mondiale.

Nel punto  $\mathbf{y}_{(1)}^*$  si hanno come autovalori della matrice  $\mathbf{A}$  i valori reali  $a$  e  $-d$  di segno opposto. Pertanto, l'origine  $[0, 0]^T$  è un punto *sella instabile*. Partendo anche da una quantità piccola, la popolazione preda aumenta, in quanto si ha  $a > 0$ .

Nell'altro punto critico  $\mathbf{y}_{(2)}^*$  la matrice  $\mathbf{A}$  diventa

$$\mathbf{A} = \begin{bmatrix} 0 & -bd/c \\ ca/b & 0 \end{bmatrix}$$

ed ha come autovalori i numeri immaginari  $+i\sqrt{ad}$  e  $-i\sqrt{ad}$  e quindi il punto critico è un centro. Se il sistema fosse genuinamente lineare, cioè non linearizzato, la sua soluzione oscillerebbe intorno al punto  $\mathbf{y}_{(2)}^*$  come  $\cos\sqrt{ad}t$  e  $\sin\sqrt{ad}t$ . Per il sistema non lineare dato, vi è un *ciclo periodico* (cfr. per una illustrazione Figura B.9).

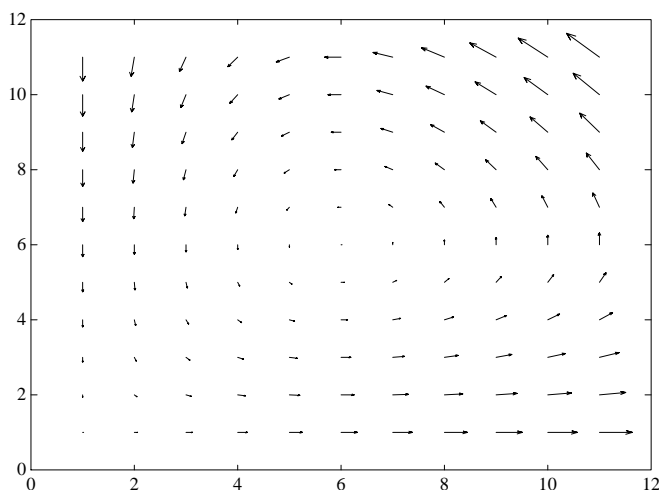


Figura B.9: Campo di velocità per il sistema predatore-preda.

### B.8.3 Equazioni lineari di ordine $n$

La procedura utilizzata nel paragrafo precedente nel caso di un'equazione lineare a coefficienti costanti del secondo ordine può essere generalizzata al caso di un'equazione lineare a coefficienti costanti di ordine  $n$  qualunque

$$L(y) := y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y' + a_0y = f(t) \quad (\text{B.24})$$

ove i coefficienti  $a_i$  sono numeri *reali* assegnati e  $f(t)$  è una funzione continua su un intervallo  $I$ .

Il *polinomio caratteristico* corrispondente all'equazione differenziale (B.24) e definito nel modo seguente

$$P(\lambda) := \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$$

con  $\lambda$  numero reale o complesso, ha la seguente proprietà

$$L(e^{\lambda t}) = P(\lambda)e^{\lambda t}$$

Si ha, quindi, che la funzione  $e^{\lambda t}$  è una soluzione particolare dell'equazione omogenea corrispondente all'equazione (B.24) se e solo se  $\lambda$  è una radice del polinomio  $P(\lambda)$ .

Dal momento che i coefficienti  $a_i$  sono numeri reali, le radici complesse del polinomio  $P(\lambda)$  sono coniugate, con la stessa molteplicità. Supponiamo, ora, che il polinomio abbia le  $\lambda_j, \bar{\lambda}_j$  radici complesse, ognuna con molteplicità  $n_j, j = 1, 2, \dots, k$  e le radici reali  $\mu_j$  con molteplicità  $m_j, j = 1, 2, \dots, l$ . Si può allora dimostrare che l'integrale generale dell'equazione lineare omogenea a coefficienti costanti  $L(y) = 0$  è dato dalla seguente famiglia di funzioni

$$y(t) = h_1(t)e^{\mu_1 t} + \dots + h_l(t)e^{\mu_l t} + p_1(t)e^{\alpha_1 t} \cos \beta_1 t + q_1(t)e^{\alpha_1 t} \sin \beta_1 t + \dots \\ + p_k(t)e^{\alpha_k t} \cos \beta_k t + q_k(t)e^{\alpha_k t} \sin \beta_k t$$

ove  $\lambda_j = \alpha_j + i\beta_j, j = 1, 2, \dots, k$  e  $h_j$  sono polinomi generici di grado minore o uguale a  $m_j - 1$ , per  $j = 1, 2, \dots, l$  e  $p_j, q_i$  polinomi generici di grado minore o uguale a  $n_j - 1$ , per  $j = 1, 2, \dots, k$ .

► **Esempio B.11** Per l'equazione differenziale

$$y^{(5)} + 8y^{(3)} + 16y' = 0$$

si ha  $P(\lambda) = \lambda^5 + 8\lambda^3 + 16\lambda$  e, utilizzando le notazioni precedenti, l'equazione  $P(\lambda) = 0$  ha una radice reale  $\mu_1 = 0$  con molteplicità  $m_1 = 1$  e una coppia di radici complesse coniugate  $\lambda_1 = \pm 2i$  con molteplicità  $n_1 = 2$ . Pertanto, l'integrale generale è della seguente forma

$$y(t) = c_1 + c_2 \cos 2t + c_3 \sin 2t + c_4 t \cos 2t + c_5 t \sin 2t$$

con  $c_j, j = 1, 2, \dots, 5$  costanti arbitrari. Dall'integrale generale si può ottenere, ad esempio, la soluzione del seguente *problema a valori iniziali*

$$y^{(5)} + 8y^{(3)} + 16y' = 0 \tag{B.25}$$

$$y(0) = 1; y'(0) = 1, y^{(2)}(0) = y^{(3)}(0) = y^{(4)}(0) = 0 \tag{B.26}$$

Dalle condizioni iniziali si ottengono, infatti, le condizioni

$$c_1 + c_2 = 1, \quad 2c_3 + c_4 = 1, \quad -4c_2 + 4c_5 = 0, \quad -8c_3 - 12c_4 = 0, \quad -16c_2 - 32c_5 = 0$$

da cui  $c_1 = 1, c_2 = c_5 = 0, c_3 = 3/4, c_4 = -1/2$ . Pertanto la soluzione del problema a valori iniziali assegnato è data dalla funzione

$$y(t) = 1 + \frac{3}{4} \sin 2t - \frac{1}{2} t \cos 2t$$

◆ **Esercizio B.3** Trovare le soluzioni dei seguenti problemi.

- a)  $y'' + 2y' + 2y = e^t$ ,  $y(0) = 0$ ,  $y'(0) = 0$ .  
 b)  $y'' - 3y' + 2y = 0$ ,  $y(0) - y'(0) = 1$ ,  $y'(1) = -2$ .  
 c)  $y'' = 2$ ,  $y(0) = 0$ ,  $y(1) + y'(1) = 0$ .

◆ **Esercizio B.4** *Mostrare che se i numeri reali  $a, b$  sono tali che  $a > 0, b > 0$ , allora ogni soluzione dell'equazione  $y'' + ay' + by = 0$  tende a zero per  $t \rightarrow \infty$ .*

◆ **Esercizio B.5** *Una soluzione  $y(t)$  di  $y'' + ay' + by = 0$ , con  $a$  e  $b$  numeri reali assegnati, è detta positivamente limitata se vi è una costante positiva  $M$  tale che  $|y(t)| \leq M$  per tutti i valori di  $t \geq 0$ . Mostrare che tutte le soluzioni sono positivamente limitate se  $a \geq 0, b \geq 0$  e  $a^2 + b^2 \neq 0$ .*

◆ **Esercizio B.6** *Trovare la soluzione generale dei seguenti problemi a valori iniziali*

- a)  $y' = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} y$ ,  $y(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$   
 b)  $y' = \begin{bmatrix} 1 & -3 \\ -2 & 2 \end{bmatrix} y$ ,  $y(0) = \begin{bmatrix} 0 \\ 5 \end{bmatrix}$

## B.9 Trasformata di Laplace

La *trasformata di Laplace*<sup>7</sup> è una trasformazione lineare mediante la quale una funzione della variabile temporale  $t$  è trasformata in una funzione di variabile complessa  $s$ . La trasformata di Laplace riduce una *equazione differenziale lineare* in una *equazione algebrica* e sostituisce alla convoluzione di due funzioni il prodotto delle loro corrispondenti trasformate.

Sia  $f(t)$  una funzione a valori reali definita per  $0 \leq t < \infty$  con le seguenti proprietà.

1. La funzione  $f(t)$  è *continua a tratti*, ossia su ogni intervallo di lunghezza finita la funzione  $f(t)$  ha al più un numero finito di discontinuità e in ogni punto di discontinuità esiste sia il limite sinistro che il limite destro.
2. La funzione  $f(t)$  è di *ordine esponenziale*, ossia esistono due costanti  $\alpha$  e  $M$ , dipendenti da  $f$ , tali che

$$|f(t)| < M e^{\alpha t} \quad 0 \leq t < \infty \quad (\text{B.27})$$

**Definizione B.2** *La trasformata di Laplace della funzione  $f(t)$  è la funzione*

$$\mathcal{L}[f] = F(s) = \int_0^{\infty} e^{-st} f(t) dt \quad (\text{B.28})$$

<sup>7</sup>Pierre Simon Laplace (1749–1827).



ove  $s$  è una variabile complessa<sup>8</sup>.

Il seguente risultato stabilisce una condizione sufficiente per l'esistenza della trasformata di Laplace.

**Teorema B.1** *Se  $f(t)$  è una funzione continua a tratti di ordine esponenziale, allora la trasformata di Laplace esiste per  $\Re(s) > \alpha$ . Inoltre, si ha*

$$|F(s)| < \frac{M}{\Re(s) - \alpha} \quad (\text{B.29})$$

$$\lim_{s \rightarrow \infty} F(s) = 0 \quad (\text{B.30})$$

Dalla condizione (B.30) si ricava in particolare che il quoziente di due polinomi può essere una trasformata di Laplace solo se il grado del denominatore è maggiore del grado del numeratore.

Le ipotesi del Teorema B.1 sono condizioni sufficienti, ma non necessarie, come mostra il seguente esempio

$$\mathcal{L}[t^{-1/2}] = \left(\frac{\pi}{s}\right)^{1/2}$$

nel quale la funzione  $f(t) = t^{-1/2}$  non è continua nel punto zero.

► **Esempio B.12** Se  $f(t)$  è la funzione costante  $c$  per  $t \geq 0$ , allora per  $\Re(s) > 0$  si ha

$$\mathcal{L}[f](s) = \int_0^{\infty} e^{-st} c dt = \left[-\frac{c}{s} e^{-st}\right]_0^{\infty} = \frac{c}{s}$$

dal momento che  $\lim_{t \rightarrow \infty} e^{-st} = 0$  per  $\Re(s) > 0$ . In particolare, per  $c = 1$  la funzione  $f(t)$  è la *funzione di Heaviside* (unit step function)

$$H(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (\text{B.31})$$

Si veda la Figura B.10 per una rappresentazione nel caso di  $s \in \mathbb{R}$ . ■

► **Esempio B.13** Sia  $f(t) = t$ , per  $t \geq 0$ . Per  $\Re(s) \geq 0$  si ha

$$\mathcal{L}[f](s) = \int_0^{\infty} e^{-st} t dt = \left[-\frac{1}{s^2} e^{-st} (st + 1)\right]_0^{\infty} = \frac{1}{s^2}$$

---

<sup>8</sup>Ricordiamo che l'integrale che appare nella definizione (B.28) è da intendersi nel seguente modo

$$\int_0^{\infty} e^{-st} f(t) dt = \lim_{T \rightarrow \infty} \int_0^T e^{-st} f(t) dt$$

La funzione  $f(t) = e^{t^2}$  fornisce un esempio di una funzione per la quale tale limite non esiste.

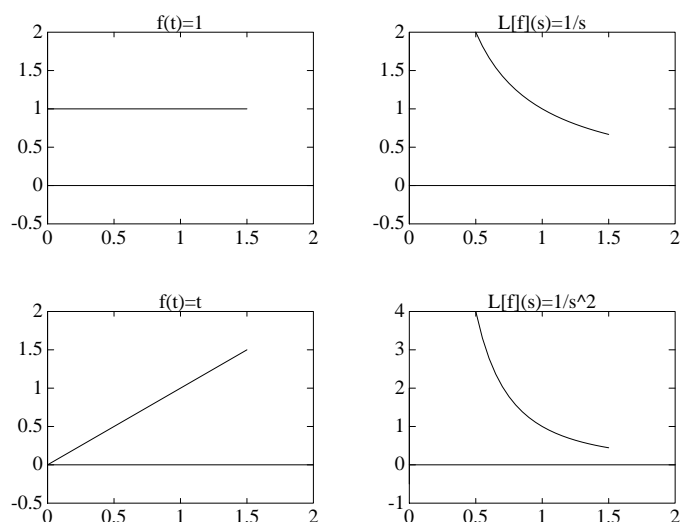


Figura B.10: Trasformate di Laplace delle funzioni  $f(t) = 1$  e  $f(t) = t$ .

► **Esempio B.14** Sia  $f(t) = 5e^{at}$ , per  $t \geq 0$  e  $a \neq 0$ . Allora, per  $\Re(s) > a$  si ha

$$\mathcal{L}[f](s) = \int_0^{\infty} e^{-st} 5e^{at} dt = \left[ \frac{5}{a-s} e^{(a-s)t} \right]_0^{\infty} = \frac{5}{s-a}$$

Si veda la Figura B.11 per una esemplificazione grafica per  $s \in \mathbb{R}$ . ■

► **Esempio B.15** Sia  $f(t)$  la funzione continua a tratti definita nel modo seguente

$$f(t) = \begin{cases} 1, & 0 \leq t \leq 1 \\ 0, & t > 1 \end{cases}$$

Allora, per  $\Re(s) \geq 0$

$$\mathcal{L}[f](s) = \int_0^{\infty} e^{-st} f(t) dt = \int_0^1 e^{-st} dt = \left[ -\frac{1}{s} e^{-st} \right]_0^1 = \frac{1}{s} [1 - e^{-s}]$$

Osserviamo che  $\mathcal{L}[f](s)$  è una funzione indefinitamente derivabile nella variabile  $s$ , sebbene la funzione  $f(t)$  non sia continua.

In modo analogo si dimostra che se  $f(t)$  è la seguente funzione

$$f(t) = \begin{cases} 1, & a \leq t \leq b \\ 0, & 0 \leq t < a, \quad b < t < \infty \end{cases} \quad (\text{B.32})$$

ossia la *funzione impulso* sull'intervallo  $[a, b]$ , con  $0 < a < b$ , si ha

$$\mathcal{L}[f](s) = \frac{e^{-as}(1 - e^{(a-b)s})}{s}$$

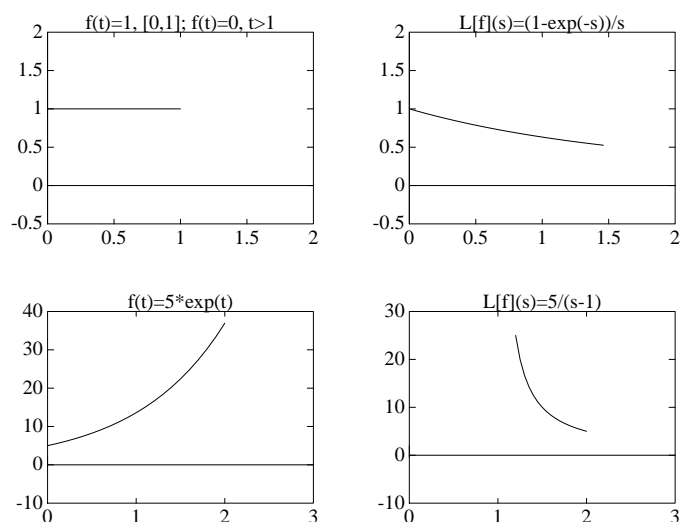


Figura B.11: Trasformate di Laplace di una funzione costante a tratti e della funzione esponenziale  $f(t) = 5e^t$ .

e per la funzione a gradini (*step function*)

$$f(t) = \begin{cases} 1, & 0 \leq t \leq a \\ 2, & a \leq t \leq 2a \\ \vdots & \end{cases} \Rightarrow \mathcal{L}[f](s) = \frac{1}{s(1 - e^{-as})} \quad (\text{B.33})$$

► **Esempio B.16** *Funzione di Dirac*. Introduciamo il significato della cosiddetta funzione di Dirac in maniera intuitiva considerando le equazioni di moto di una palla di baseball che arriva al battitore con una velocità  $v(t_0)$  e riparte con una velocità  $v(t_1)$ , dipendente dalla forza esercitata dal battitore. Utilizzando la seconda legge di Newton si ha

$$\frac{d[mv(t)]}{dt} = f(t)$$

ove  $m$  è la massa della palla,  $v$  la sua velocità, e  $f(t)$  la forza totale che agisce sulla palla. Integrando l'equazione precedente tra  $t_0$  e  $t_1$ , si ottiene

$$mv(t_1) - mv(t_0) = \int_{t_0}^{t_1} f(t) dt$$

La quantità  $mv$  è chiamato il *momento* di moto della palla. L'equazione precedente stabilisce che la variazione del momento della palla da  $t_0$  a  $t_1$  è uguale all'integrale della  $f(t)$  su  $(t_0, t_1)$ . Pertanto, la quantità fisicamente importante è l'integrale della forza, il cosiddetto *impulso*, piuttosto che la forza stessa.

Se poniamo  $t_1 = t_0 + \epsilon$ , con  $\epsilon > 0$ , e definiamo la seguente funzione

$$\delta_\epsilon(t - t_0) = \begin{cases} 0 & t < t_0 \\ 1/\epsilon & t_0 < t < t_0 + \epsilon \\ 0 & t > t_0 + \epsilon \end{cases}$$

è facile verificare che

$$\int_{-\infty}^{\infty} \delta_{\epsilon}(t - t_0) dt = 1$$

e  $f(t) = I \delta_{\epsilon}(t - t_0)$ , per la quale si ha  $\int_{t_0}^{t_0+\epsilon} f(t) dt = I$ , rappresenta una approssimazione matematica di una forza “grande” che agisce su un intervallo di tempo “piccolo”, in maniera che il suo impulso sia una costante  $I$ ;  $f(t)$  è tanto più grande sull’intervallo  $(t_0, t_0+\epsilon)$ , quanto più è piccolo  $\epsilon$ . Inoltre per ogni funzione  $g(t)$  continua su  $\mathbb{R}$  si ha (applicando il teorema della media)

$$\int_{-\infty}^{\infty} g(t) \delta_{\epsilon}(t - t_0) dt = \frac{1}{\epsilon} \int_{t_0}^{t_0+\epsilon} g(t) dt = g(\tau)$$

con  $\tau$  valore opportuno nell’intervallo  $(t_0, t_0 + \epsilon)$ .

Passando *formalmente* al limite, per  $\epsilon \rightarrow 0$ , e indicando con  $\delta$  la “funzione” limite, si ottiene

$$\int_{-\infty}^{\infty} \delta(t - t_0) dt = 1, \quad \int_{-\infty}^{\infty} \delta(t - t_0) g(t) dt = g(t_0) \quad (\text{B.34})$$

Il limite  $\delta$ , noto come *funzione delta di Dirac*, fu introdotto dal fisico P. A. M. Dirac nel 1930 per descrivere funzioni impulsive nell’ambito della meccanica quantistica. Naturalmente, si tratta di una funzione non ordinaria, in quanto non è definita puntualmente (infatti,  $\delta(t - t_0) = 0$  per  $t \neq t_0$  e  $\infty$  per  $t = t_0$ ). Una sua interpretazione, matematicamente corretta, può essere data utilizzando le proprietà (B.34) che sono di tipo integrale. Questo è stato fatto, in particolare, da L. Schwartz (1940) nell’ambito della teoria delle *distribuzioni*, o funzioni generalizzate, nella quale il concetto di funzione viene ampliato considerando i *funzionali lineari e continui* su opportune classi di funzioni.

Rinviando alla letteratura specializzata per un approfondimento della nozione delle funzioni impulso  $\delta$ , dalla (B.34) ponendo  $g(t) = e^{-st}$  per  $t \geq 0$  e  $g(t) = 0$  per  $t < 0$ , si ha

$$\mathcal{L}[\delta(t - t_0)] = e^{-st_0}$$

■

## La trasformazione inversa

Data la funzione  $f$ , la sua trasformata di Laplace è univocamente determinata da (B.28). Reciprocamente,  $F(s)$  determina univocamente  $f(t)$  come mostrato nel seguente teorema.

**Teorema B.2** *Se  $f_1, f_2$  sono due funzioni continue a tratti di ordine esponenziale e aventi la stessa trasformata di Laplace  $F$ , allora  $f_1(t) = f_2(t)$  nei punti  $t$  nei quali  $f_1$  e  $f_2$  sono ambedue continue.*

In altre parole, due funzioni continue a tratti di ordine esponenziale con la stessa trasformata di Laplace possono differire solo nei loro punti di discontinuità.

La *trasformata inversa*  $f$  di  $F$  è scritta simbolicamente nella forma

$$f = \mathcal{L}^{-1}[F] \quad (\text{B.35})$$

L'operatore  $\mathcal{L}^{-1}$  è chiamato *operatore di Laplace inverso*, e  $f$  è chiamata *l'inversa di  $F$* . La costruzione di tale inversa è basata su alcune proprietà della trasformata, mediante le quali è possibile ricondurre il calcolo dell'inversa a quello delle trasformate inverse di funzioni elementari contenute in opportune tavole (cfr. Tabella B.2 e più in generale Abramowitz e Stegun [1], Erdélyi e Bateman [55], Widder [154]).

### B.9.1 Proprietà della trasformata di Laplace

In questo paragrafo considereremo le principali proprietà della trasformata di Laplace, mediante le quali si può esprimere la trasformata di una combinazione di funzioni in termini di trasformate note. La dimostrazione di tali proprietà, basata su proprietà elementari dell'operazione di integrazione, è lasciata come esercizio.

**A. Traslazione** Se  $h$  è un numero reale, allora

$$\mathcal{L}[f(t-h)] = e^{-sh}F(s)$$

ove  $f(t)$  è definita uguale a 0 per  $t < 0$ . Ad esempio, la trasformata di Laplace della funzione  $g(t) = 1$  per  $t \geq 5$  e 0 altrove, può essere calcolata osservando che  $g(t) = f(t-5)$ , ove  $f(t)=1$  per  $0 \leq t < \infty$ . Dall'Esempio B.12 si ha allora

$$\mathcal{L}[g] = \mathcal{L}[f(t-5)] = \frac{e^{-5s}}{s}$$

Più in generale, si ha

$$\mathcal{L}[H(t-h)] = e^{-hs} \frac{1}{s} \tag{B.36}$$

mentre per  $h > 0$ , si ha direttamente

$$\mathcal{L}[H(h-t)] = \int_0^h e^{-st} dt = \frac{1 - e^{-hs}}{s}, \quad \Re(s) > 0 \tag{B.37}$$

**B. Cambiamento di scala** Se  $a$  è un numero reale positivo, allora

$$\mathcal{L}\left[f\left(\frac{t}{a}\right)\right] = aF(as)$$

Ad esempio, tenendo conto che la trasformata di Laplace di  $\sin t$  è  $F(s) = (s^2+1)^{-1}$ , la trasformata di  $\sin at$  è

$$\frac{F(s/a)}{a} = \frac{1/[(s/a)^2 + 1]}{a} = \frac{a}{s^2 + a^2}$$

**C. Traslazione complessa** Sia  $a$  un numero complesso. Allora

$$\mathcal{L}[e^{-at} f(t)] = F(s + a)$$

Tenendo ad esempio conto che la trasformata della funzione  $f(t) = 1$  è  $F(s) = 1/s$ , si ha

$$\mathcal{L}[e^{-at}] = \frac{1}{s + a}$$

In modo analogo, si trova

$$\mathcal{L}[e^{-2t} \cos 3t] = \frac{s + 2}{(s + 2)^2 + 9}$$

**D. Linearità** L'operatore di Laplace  $\mathcal{L}$  è lineare, ossia verifica la seguente proprietà

$$\mathcal{L}[c_1 f_1 + c_2 f_2] = c_1 \mathcal{L}[f_1] + c_2 \mathcal{L}[f_2]$$

Come illustrazione, si ha

$$\mathcal{L}[5t^3 + 7 \sin 2t] = 5\mathcal{L}[t^3] + 7\mathcal{L}[\sin 2t] = \frac{30}{s^4} + \frac{14}{s^2 + 4}$$

Dal fatto che  $\mathcal{L}$  è un operatore lineare si ricava che pure  $\mathcal{L}^{-1}$  è un operatore lineare.

**E. Convoluzione** Il prodotto di convoluzione  $f * g$  delle funzioni  $f$  e  $g$ , supposte continue a tratti e di ordine esponenziale, è definito nel modo seguente

$$(f * g)(t) = \int_0^t f(x)g(t-x) dx = \int_0^t g(x)f(t-x) dx \quad (\text{B.38})$$

**Teorema B.3** (teorema di convoluzione) *Se le funzioni  $f$  e  $g$  sono continue a tratti su  $(0, \infty)$  e di ordine esponenziale, allora*

$$\mathcal{L}[f * g] = \mathcal{L}[f] \cdot \mathcal{L}[g] \quad (\text{B.39})$$

o, equivalentemente

$$\mathcal{L}^{-1}[FG] = f * g = g * f \quad (\text{B.40})$$

ove  $F$  e  $G$  rappresentano rispettivamente la trasformata di Laplace di  $f$  e  $g$ .

Tenendo, ad esempio, presente che  $\mathcal{L}[1] = 1/s$  e che  $\mathcal{L}[\sin t] = (s^2 + 1)^{-1}$ , si ha

$$\mathcal{L}^{-1}\left[\frac{1}{s(s^2 + 1)}\right] = \mathcal{L}^{-1}\left[\frac{1}{s}\right] * \mathcal{L}^{-1}\left[\frac{1}{s^2 + 1}\right] = \int_0^t 1 \cdot \sin x dx = 1 - \cos t$$

In maniera analoga si ha

$$\mathcal{L}\left[\int_0^t x \sin(t-x) dx\right] = \mathcal{L}[t] \mathcal{L}[\sin t] = s^{-2}(s^2 + 1)^{-1}$$

e

$$\mathcal{L}^{-1}[(s + 5)^{-1}(s^2 + 1)^{-1}] = \mathcal{L}^{-1}[1/(s + 5)] * \mathcal{L}^{-1}[1/(s^2 + 1)] = e^{-5t} * \sin t$$

proprietà	funzione	trasformata di Laplace
1. traslazione reale	$f(t-h)$	$e^{-sh} F(s)$
2. cambiamento di scala	$f(t/a)$	$aF(as)$
3. traslazione complessa	$e^{-at} f(t)$	$F(s+a)$
4. linearità	$c_1 f_1(t) + c_2 f_2(t)$	$c_1 F_1 + c_2 F_2$
5. convoluzione	$f * g$	$F G$
6. derivata	$f^{(n)}(t)$	$s^n F(s) - s^{n-1} f(0+) - s^{n-2} f'(0+) - \dots - f^{(n-1)}(0+)$
7. integrazione	$f^{(-p)}(t)$	$s^{-p} F(s) + s^{-p} f^{(-p)}(0+) + \dots + s^{-1} f^{(-p)}(0+)$
8. valore iniziale	$\lim_{t \rightarrow 0+} f(t)$	$\lim_{s \rightarrow \infty} sF(s)$
9. valore finale	$\lim_{t \rightarrow \infty} f(t)$	$\lim_{s \rightarrow 0+} sF(s)$
10. derivata di $\mathcal{L}[f]$	$-tf(t)$	$F'(s)$
11. periodicità: $f(t) = f(t+T)$	$f(t)$	$\int_0^T e^{-st} f(t) dt / (1 - e^{-sT})$

Tabella B.1: Proprietà della trasformata di Laplace.

**F. Trasformate di derivate** Mediante la formula di integrazione per parti, si verifica facilmente che se la derivata  $f^{(n)}$  di ordine  $n$  è una funzione continua a tratti e di ordine esponenziale, si ha

$$\mathcal{L}[f^{(n)}(t)](s) = -f^{(n-1)}(0+) + s\mathcal{L}[f^{(n-1)}(t)]$$

ove con  $f^{(n-1)}(0+)$  si è indicato il limite a destra nel punto 0. Per induzione, si ha allora la seguente formula

$$\mathcal{L}[f^{(n)}] = s^n \mathcal{L} - s^{n-1} f(0+) - s^{n-2} f'(0+) - \dots - f^{(n-1)}(0+) \quad (\text{B.41})$$

► **Esempio B.17** Consideriamo il seguente problema a valori iniziali

$$\begin{cases} y'' + y' - 2y = 0 \\ y(0) = 1, y'(0) = 0 \end{cases} \quad (\text{B.42})$$

Prendendo la trasformata di Laplace di ambo i membri dell'equazione differenziale, tenendo conto della linearità e del fatto che

$$\mathcal{L}[y''] = s^2 Y(s) - sy(0) - y'(0); \quad \mathcal{L}[y'] = sY(s) - y(0)$$

ove  $Y(s)$  indica la trasformata della soluzione  $y(t)$ , si ha la seguente algebrica in  $Y(s)$

$$s^2 Y(s) - s + sY(s) - 1 - 2Y(s) = 0$$

dalla quale si ricava

$$Y(s) = \frac{1+s}{s^2+s-2} = \frac{2}{3(s-1)} + \frac{1}{3(s+2)}$$

Usando la linearità dell'operatore inverso di Laplace e il fatto che  $\mathcal{L}^{-1}[(s+a)^{-1}] = e^{-at}$ , ne segue che

$$\mathcal{L}^{-1}[Y(s)] = \frac{2}{3}\mathcal{L}^{-1}\left[\frac{1}{s-1}\right] + \frac{1}{3}\mathcal{L}^{-1}\left[\frac{1}{s+2}\right]$$

da cui la seguente espressione della soluzione del problema a valori iniziali (B.42)

$$y(t) = \frac{2}{3}e^t + \frac{1}{3}e^{-2t}$$

L'applicazione della trasformata di Laplace per la risoluzione di equazioni differenziali sarà ulteriormente illustrata nel paragrafo successivo. ■

**G. Integrazione** Introducendo la seguente notazione

$$\begin{aligned} f^{(-1)}(t) &= \int_{a_1}^t f(x_1) dx_1 \\ f^{(-2)}(t) &= \int_{a_2}^t \int_{a_1}^{x_2} f(x_1) dx_1 dx_2 \\ &\vdots \\ f^{(-p)}(t) &= \int_{a_p}^t \int_{a_{p-1}}^{x_p} \cdots \int_{a_1}^{x_2} f(x_1) dx_1 dx_2 \cdots dx_p \end{aligned}$$

si ha

$$\mathcal{L}[f^{(-p)}(t)] = \frac{1}{s^p}F(s) + \frac{1}{s^p}f^{(-1)}(0+) + \cdots + \frac{1}{s}f^{(-p)}(0+)$$

Osservando, ad esempio, che la funzione  $t^2$ , per  $t \geq 0$  può essere scritta nella forma

$$t^2 = 2 \int_0^t \int_0^{x_2} dx_1 dx_2$$

si ha

$$\mathcal{L}[t^2] = \frac{2}{s^2}\mathcal{L}[1] = \frac{2}{s^3}$$

Più in generale, si può vedere che  $\mathcal{L}[t^n] = n!/s^{n+1}$ .

**H. Valore iniziale** Se  $f$  ha una derivata continua a tratti, allora

$$\lim_{t \rightarrow 0+} f(t) = \lim_{s \rightarrow \infty} sF(s)$$

Tale risultato è utile quando la trasformata di Laplace è difficile da invertire, ma d'altra parte è interessante il valore iniziale. Ad esempio,

$$\mathcal{L}[f] = \frac{2s^2 + 5}{3s^3 + 2s^2 + 5} \Rightarrow \lim_{t \rightarrow 0+} f(t) = \frac{2}{3}$$



Ossia il valore iniziale di  $f$  è calcolato senza invertire  $F(s)$ .

Un risultato analogo al precedente riguarda il *valore finale*. Nelle stesse ipotesi si ha

$$\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0^+} sF(s)$$

### B.9.2 Applicazioni della trasformata di Laplace

I seguenti esempi illustrano l'utilizzo del metodo della trasformata di Laplace per risolvere le equazioni differenziali.

► **Esempio B.18** Consideriamo il seguente problema a valori iniziali

$$\begin{cases} y'' - 2y' - 3y = 6e^t \\ y(0) = 1, y'(0) = 3 \end{cases} \quad (\text{B.43})$$

Si ha

$$\mathcal{L}[y'' - 2y' - 3y] = \mathcal{L}[6e^t] = \frac{6}{s-1} \Rightarrow s^2Y - s - 3 - 2sY + 2 - 3Y = \frac{6}{s-1}$$

da cui

$$Y = \frac{s^2 + 5}{(s-1)(s+1)(s-3)} \Rightarrow y = \mathcal{L}^{-1}[Y] = \frac{7e^{3t}}{4} - \frac{3e^t}{2} + \frac{3e^{-t}}{4}$$

■

► **Esempio B.19** Consideriamo il seguente problema *integrodifferenziale*

$$\begin{cases} y' - 3y + 2 \int_0^t y dt = 1 + t \\ y(0) = 1 \end{cases} \quad (\text{B.44})$$

Tenendo conto che

$$\mathcal{L}[y'] = sY - 1; \quad \mathcal{L}\left[\int_0^t y dt\right] = \frac{Y}{s}$$

si ha

$$sY - 1 - 3Y + \frac{2}{s}Y = \frac{1}{s} + \frac{1}{s^2} \Rightarrow Y = \frac{s^2 + s + 1}{s(s-1)(s-2)}$$

da cui

$$y = \mathcal{L}^{-1}[Y] = \mathcal{L}^{-1}\left[\frac{1}{2s} + \frac{7}{2(s-2)} - \frac{3}{s-1}\right] = \frac{1}{2} + \frac{7}{2}e^{2t} - 3e^t$$

■

► **Esempio B.20** Consideriamo la soluzione del seguente sistema differenziale

$$\begin{cases} -2x(t) + y'(t) = -1 \\ 2x'(t) - y(t) = t \end{cases}$$

con le condizioni iniziali  $x(0) = 1/2$  e  $y(0) = 0$ . Applicando la trasformata di Laplace, si ottiene il sistema lineare

$$\begin{cases} -2X + sY = -\frac{1}{s} \\ 2sX - 1 - Y = \frac{1}{s^2} \end{cases} \Rightarrow \begin{cases} X = \frac{1}{4(s-1)} + \frac{1}{4(s+1)} \\ Y = \frac{1}{2(s-1)} - \frac{1}{2(s+1)} \end{cases}$$

da cui

$$x = \frac{e^t}{4} + \frac{e^{-t}}{4}; \quad y = \frac{e^t}{2} - \frac{e^{-t}}{2}$$

■

► **Esempio B.21** Consideriamo il movimento di una molla oscillante sottoposta ad un tempo fissato  $T > 0$  ad una forza di tipo impulsivo e corrispondente al seguente modello matematico

$$\begin{cases} x''(t) + \omega^2 x(t) = A \delta(t - T) \\ x(0) = \alpha; \quad x'(0) = \beta \end{cases}$$

ove  $A$  è una costante positiva e  $\omega$  è un parametro legato alle proprietà di rigidità della molla e alla sua massa. La forza  $A\delta(t - T)$  è chiamata una *forza impulsiva*, mentre il suo integrale è il corrispondente *impulso* (cfr. Esempio B.16). Applicando la trasformata di Laplace, si ottiene

$$(s^2 + \omega^2)\mathcal{L}[x] - s\alpha - \beta = Ae^{-Ts} \Rightarrow \mathcal{L}[x] = \frac{s\alpha + \beta}{s^2 + \omega^2} + \frac{Ae^{-Ts}}{s^2 + \omega^2}$$

Applicando la trasformazione inversa, si ottiene

$$x(t) = \alpha \cos \omega t + \frac{\beta}{\omega} \sin \omega t + \frac{A}{\omega} H(t - T) \sin \omega(t - T)$$

ove  $H$  è la funzione di Heaviside. Osserviamo che  $x(t)$  è una funzione continua, mentre  $x'(t)$  ha un salto per  $t = T$ . ■

► **Esempio B.22** Sia, al tempo  $t \geq 0$ ,  $y(t)$  la quantità di una popolazione che aumenta secondo una legge esponenziale e che è soggetta durante un periodo iniziale, diciamo i primi 30 giorni, ad un prelievo di  $h$  individui per giorno. Indicando con  $A$  la quantità al tempo  $t = 0$  e con  $k$  la costante di aumento, misurata in  $(\text{giorni})^{-1}$ , si ha il seguente modello

$$\begin{cases} y'(t) = ky(t) - h H(30 - t) \\ y(0) = A \end{cases} \quad (\text{B.45})$$

ove  $H$  indica la funzione di Heaviside. Per calcolare la funzione  $y(t)$  per  $t > 0$ , utilizziamo la trasformata di Laplace. Si ha

$$s\mathcal{L}[y] - A = k\mathcal{L}[y] - \frac{h}{s}(1 - e^{-30s})$$

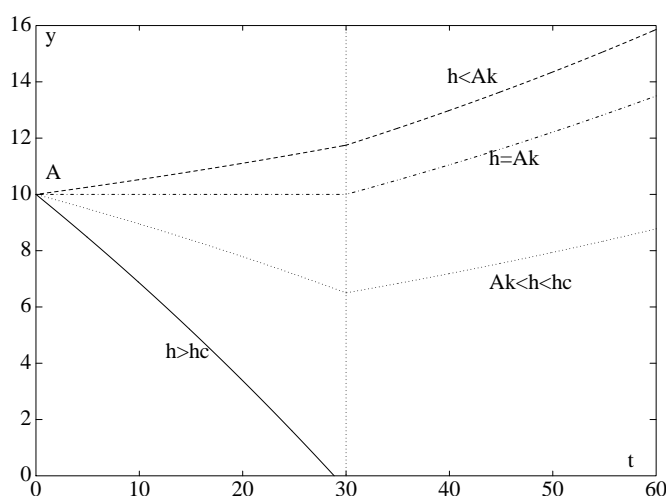


Figura B.12: Accrescimento di una popolazione soggetta a prelievo durante i primi 30 giorni.

da cui

$$\mathcal{L}[y] = \frac{A}{s-k} - \frac{h}{s(s-k)}(1 - e^{-30s}) = \frac{A-h/k}{s-k} + \frac{h}{k} \frac{1}{s}(1 - e^{-30s}) + \frac{h}{k} e^{-30s} \frac{1}{s-k}, \quad s > k$$

Passando alla trasformata inversa, si ha la seguente formula per la soluzione  $y(t)$

$$y(t) = \left(A - \frac{h}{k}\right) e^{kt} + \frac{h}{k} H(30-t) + \frac{h}{k} H(t-30) e^{k(t-30)} \quad (\text{B.46})$$

ove l'ultimo termine è utilizzato per  $t > 30$ . In Figura B.12 sono rappresentati i comportamenti della popolazione in corrispondenza a valori diversi della velocità di raccolta  $h$ . Si può vedere facilmente che se

$$h \geq h_c := \frac{kA}{1 - e^{-30k}}$$

allora la popolazione diventa negativa (ossia si estingue) prima dei 30 giorni.

Lasciamo come esercizio la ricerca della relazione tra i parametri  $A$ ,  $h$  e  $k$  per i quali dopo, ad esempio, 330 giorni dalla fine del raccolto, la popolazione raggiunga nuovamente il livello  $A$  del tempo  $t = 0$ . ■

◆ **Esercizio B.7** Sapendo che  $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$ , trovare  $\mathcal{L}[t^{-1/2}]$ .

◆ **Esercizio B.8** Trovare la trasformata di Laplace di  $f(t) = \epsilon^{-1}$ , per  $a < t < a + \epsilon$  e 0 altrove. Trovare, quindi, la trasformata di Laplace della funzione delta  $\delta(t-a)$  facendo tendere, nel risultato precedente,  $\epsilon$  a zero.

◆ **Esercizio B.9** Trovare la trasformata di Laplace delle seguenti funzioni

$$(1) \sinh at \quad (2) t^2 e^{at} \quad (3) te^{2t} f'(t)$$

$$(4) (t+1)H(t-1) \quad (5) te^{at}H(t-1) \quad (6) (t-2)[H(t-1) - H(t-3)]$$

◆ **Esercizio B.10** Utilizzare la trasformata di Laplace per risolvere i seguenti problemi a valori iniziali

$$(1) y'' + y = \sin t, \quad y(0) = 0, \quad y'(0) = 1$$

$$(2) y'' - 2y' + 2y = 0, \quad y(0) = 0, \quad y'(0) = 1$$

$$(2) y'' - 2y' + y = H(t-1), \quad y(0) = 1, \quad y'(0) = 0$$

$$(3) y'' + 2y' - 3y = H(1-t), \quad y(0) = 1, \quad y'(0) = 0$$

$$(4) y'' + y = e^t + \delta(t-1), \quad y(0) = 0, \quad y'(0) = 0$$

◆ **Esercizio B.11** Supponiamo che un campione di un elemento radioattivo abbia una velocità di decadimento proporzionale alla quantità presente, e che al tempo  $t = a$  si verifichi una introduzione dell'elemento dall'esterno con flusso costante, che si interrompe all'istante  $t = b$ , con  $b > a$ . Mostrare che il sistema può essere modellizzato nel seguente modo

$$y' = -k_1 y + k_2 [H(t-a) - H(t-b)], \quad 0 < a < b$$

ove  $k_1$  è la costante di decadimento e  $k_2$  la quantità di elemento introdotta per unità di tempo nell'intervallo  $a \leq t < b$ . Usare la trasformata di Laplace per rappresentare la funzione  $y(t)$ .

◆ **Esercizio B.12** In relazione allo studio di un circuito elettrico, si risolva mediante la trasformata di Laplace i seguenti problemi a valori iniziali

(a) si applica una forza elettromotrice costante  $E_0$  all'istante  $t = a$

$$L \frac{d^2 q}{dt^2} + \frac{1}{C} q = E_0 H(t-a), \quad q(0) = 0, \quad q'(0) = 0$$

(b) la forza elettromotrice è applicata all'istante  $t = a$  e esclusa all'istante  $t = b$

$$L \frac{d^2 q}{dt^2} + \frac{1}{C} q = E_0 [H(t-a) - H(t-b)], \quad q(0) = 0, \quad q'(0) = 0$$

funzione $f(t)$ , $t \geq 0$	trasformata di Laplace, $F(s)$
$\delta(t - t_0)$ , impulso unitario	$e^{-st_0}$
$H(t - a)$	$\frac{e^{-as}}{s}$
$t^{n-1}$ , $n = 1, 2, \dots$	$\frac{(n-1)!}{s^n}$
$t^{p-1}$ , $p > 0$	$\frac{\Gamma(p)}{s^p}$
$e^{-at}$	$\frac{1}{s+a}$
$\frac{e^{-at} - e^{-bt}}{b-a}$	$\frac{1}{(s+a)(s+b)}$
$\frac{1}{(n-1)!} t^{n-1} e^{-at}$	$\frac{1}{(s+a)^n}$
$\frac{1}{b-a} [(\alpha-a)e^{-at} - (\alpha-b)e^{-bt}]$	$\frac{s+\alpha}{(s+a)(s+b)}$
$\sin \omega t$	$\frac{\omega}{s^2 + \omega^2}$
$\cos \omega t$	$\frac{s}{s^2 + \omega^2}$
$\omega t e^{-\omega t}$	$\frac{\omega}{(s+\omega)^2}$
$\frac{t^{n-1} e^{-t/T}}{T^n (n-1)!}$	$\frac{1}{(1+Ts)^n}$
$1 - e^{-t/T}$	$\frac{1}{s(1+Ts)}$
$1 - \frac{t+T}{T} e^{-t/T}$	$\frac{1}{s(1+Ts)^2}$
$e^{-at} \sin \omega t$	$\frac{\omega}{(s+a)^2 + \omega^2}$
$e^{-at} \cos \omega t$	$\frac{s+a}{(s+a)^2 + \omega^2}$
$\sin at \sin bt$	$\frac{2abs}{[s^2 + (a+b)^2][s^2 + (a-b)^2]}$
$\frac{a}{2\sqrt{\pi t^3}} e^{-a^2/4t}$ , $a > 0$	$e^{-a/\sqrt{s}}$
$\frac{e^{-bt} - e^{-at}}{t}$	$\ln \left( \frac{s+a}{s+b} \right)$
$\frac{e^{-at} - e^{-bt}}{2\sqrt{\pi t^3}}$	$\sqrt{s+b} - \sqrt{s+a}$

Tabella B.2: Trasformate di Laplace.

## B.10 Serie di Fourier

Data una funzione  $f(x)$  definita sull'intervallo  $-l \leq x \leq l$ , se i seguenti integrali esistono finiti

$$a_n = \frac{1}{l} \int_{-l}^l f(x) \cos \frac{n\pi x}{l} dx, \quad n = 0, 1, 2, \dots \quad (\text{B.47})$$

$$b_n = \frac{1}{l} \int_{-l}^l f(x) \sin \frac{n\pi x}{l} dx, \quad n = 1, 2, \dots \quad (\text{B.48})$$

la seguente serie

$$\frac{a_0}{2} + a_1 \cos \frac{\pi x}{l} + b_1 \sin \frac{\pi x}{l} + \dots = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[ a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right] \quad (\text{B.49})$$

è chiamata *serie di Fourier*<sup>9</sup> della funzione  $f$  sull'intervallo  $-l \leq x \leq l$ . Il seguente teorema fornisce delle condizioni sufficienti affinché la serie (B.49) rappresenti la funzione  $f(x)$ .

**Teorema B.4** *Supponiamo che la funzione  $f$  e la sua derivata  $f'$  siano delle funzioni continue a tratti<sup>10</sup> sull'intervallo  $-l \leq x \leq l$ . Allora la corrispondente serie di Fourier (B.49) converge a  $f(x)$  in ogni punto in cui  $f(x)$  è continua e a  $\frac{1}{2}[f(x+0) + f(x-0)]$  in un punto di discontinuità. Per  $x = \pm l$ , la serie di Fourier (B.49) converge a  $\frac{1}{2}[f(l) + f(-l)]$ .*

Osserviamo che la serie di Fourier (B.49) converge in ogni punto dell'intervallo  $-l < x < l$  alla funzione che coincide con la  $f(x)$  in ogni punto di continuità, e che in un punto di discontinuità assume il valore  $\frac{1}{2}[f(x+0) + f(x-0)]$ .

Utilizzando le seguenti relazioni

$$\begin{cases} e^{n\pi x/l} = \cos n\frac{\pi x}{l} + i \sin n\frac{\pi x}{l} \\ e^{-n\pi x/l} = \cos n\frac{\pi x}{l} - i \sin n\frac{\pi x}{l} \end{cases} \iff \begin{cases} \cos n\frac{\pi x}{l} = \frac{1}{2}(e^{n\pi x/l} + e^{-n\pi x/l}) \\ \sin n\frac{\pi x}{l} = \frac{1}{2i}(e^{n\pi x/l} - e^{-n\pi x/l}) \end{cases}$$

ove  $i$  rappresenta l'unità immaginaria  $i = \sqrt{-1}$ , è possibile riscrivere la serie (B.49) nel seguente modo equivalente

$$\sum_{n=-\infty}^{\infty} c_n e^{in\frac{\pi x}{l}}, \quad c_n = \frac{1}{2}(a_n - i b_n)$$

<sup>9</sup>J. Fourier (1768–1830) annunciò il 21 dicembre 1807 davanti all'Accademia delle Scienze di Parigi il risultato sorprendente che una funzione arbitraria può essere considerata come il limite delle somme parziali della serie (B.49); in effetti, tale risultato è valido sotto opportune ipotesi sulla funzione  $f(x)$  (cfr. Teorema B.4 per un tipo particolare di condizioni sufficienti).

<sup>10</sup>ossia le funzioni  $f$  e  $f'$  abbiano al più un numero finito di discontinuità sull'intervallo dato, e in ogni eventuale punto di discontinuità esistano i limiti a sinistra e a destra; considerando ad esempio la funzione  $f(x)$ , tali limiti verranno indicati rispettivamente con  $f(x-0)$  e  $f(x+0)$ .

Osserviamo, infine, che per una funzione  $f(x)$  tale che su tutto l'intervallo  $-l \leq x \leq l$  si abbia

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[ a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right] \quad (\text{B.50})$$

i coefficienti  $a_n, b_n$  sono necessariamente dati rispettivamente dalle formule (B.47), (B.48), ossia è *unica* la serie di Fourier di  $f(x)$  sull'intervallo fissato  $-l \leq x \leq l$ .

Dalla uguaglianza (B.50), elevando al quadrato lo sviluppo in serie e integrando termine a termine<sup>11</sup> si ottiene la seguente identità, nota come *identità di Parseval*

$$\frac{1}{l} \int_{-l}^l f^2(x) dx = \frac{a_0^2}{2} + \sum_{n=1}^{\infty} (a_n^2 + b_n^2)$$

► **Esempio B.23** Dall'unicità dello sviluppo in serie di Fourier e dalla seguente identità trigonometrica

$$\cos^2 x = \frac{1}{2} + \frac{\cos 2x}{2}$$

si ricava che la serie di Fourier della funzione  $\cos^2 x$  sull'intervallo  $-\pi \leq x \leq \pi$  è data da  $\frac{1}{2} + \frac{1}{2} \cos 2x$ . ■

► **Esempio B.24** Data la seguente funzione

$$f(x) = \begin{cases} 0 & \text{per } -1 \leq x < 0 \\ 1 & \text{per } 0 \leq x \leq 1 \end{cases} \quad (\text{B.51})$$

si ha

$$\begin{aligned} a_0 &= \int_{-1}^1 f(x) dx = \int_0^1 dx = 1 \\ a_n &= \int_{-1}^1 f(x) \cos n\pi x dx = \int_0^1 \cos n\pi x dx = 0, \quad n \geq 1 \\ b_n &= \int_{-1}^1 f(x) \sin n\pi x dx = \int_0^1 \sin n\pi x dx = \frac{1 - \cos n\pi}{n\pi} = \frac{1 - (-1)^n}{n\pi}, \quad n \geq 1 \end{aligned}$$

Si ha quindi  $b_n = 0$  per  $n$  pari e  $b_n = 2/n\pi$  per  $n$  dispari. Pertanto, la serie di Fourier della funzione  $f$  sull'intervallo  $-1 \leq x \leq 1$  è

$$\frac{1}{2} + \frac{2 \sin \pi x}{\pi} + \frac{2 \sin 3\pi x}{3\pi} + \frac{2 \sin 5\pi x}{5\pi} + \dots$$

Tale serie, come indicato dal Teorema B.4, converge a 0 se  $-1 < x < 0$ , e a 1 se  $0 < x < 1$ . Nei punti  $x = -1, 0, e +1$ , la serie si riduce al valore  $\frac{1}{2}$ . In Figura B.13 sono rappresentate le somme della serie per alcuni successivi valori di  $n$ . Osserviamo che vicino al punto di discontinuità della funzione vi è una sovrastima, che non scompare per  $n$  che aumenta (si muove verso il salto); tale comportamento è noto in letteratura come *Gibbs phenomenon* ed è tipico della serie di Fourier in un punto di discontinuità. ■

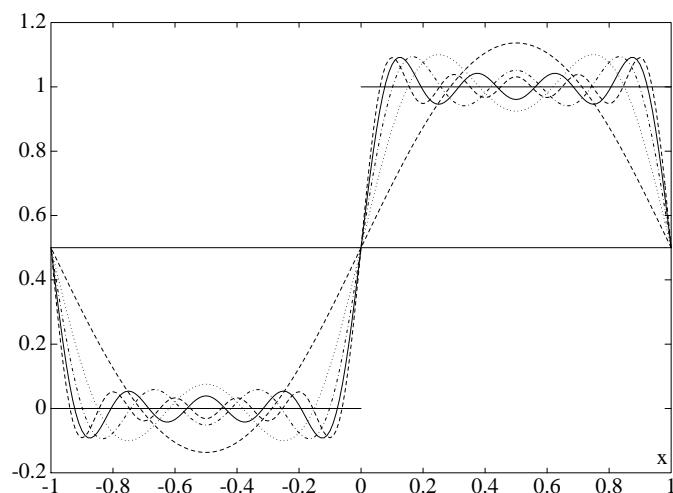


Figura B.13: Approssimanti di Fourier, per  $n = 0, 1, 3, 5, 7, 9$ , della funzione (B.51).

► **Esempio B.25** Per la seguente funzione

$$f(x) = \begin{cases} 1 & \text{per } -2 \leq x < 0 \\ x & \text{per } 0 \leq x \leq 2 \end{cases} \quad (\text{B.52})$$

si ottiene facilmente

$$a_0 = 2, \quad a_n = \frac{1}{2} \int_{-2}^0 \cos \frac{n\pi x}{2} dx + \frac{1}{2} \int_0^2 x \cos \frac{n\pi x}{2} dx = \frac{2}{(n\pi)^2} (\cos n\pi - 1), \quad n \geq 1$$

e

$$b_n = \frac{1}{2} \int_{-2}^0 \sin \frac{n\pi x}{2} dx + \frac{1}{2} \int_0^2 x \sin \frac{n\pi x}{2} dx = -\frac{1}{n\pi} (1 + \cos n\pi), \quad n \geq 1$$

da cui  $a_n = 0$  se  $n$  è pari,  $a_n = -4/n^2\pi^2$  se  $n$  è dispari,  $b_n = 0$  se  $n$  è dispari e  $b_n = -2/n\pi$  se  $n$  è pari. Pertanto la serie di Fourier di  $f$  sull'intervallo  $-2 \leq x \leq 2$  è

$$1 - \frac{4}{\pi^2} \cos \frac{\pi x}{2} - \frac{1}{\pi} \sin \pi x - \frac{4}{9\pi^2} \cos \frac{3\pi x}{2} - \frac{1}{2\pi} \sin 2\pi x + \dots$$

$$1 - \frac{4}{\pi^2} \sum_{n=0}^{\infty} \frac{\cos(2n+1)\pi x/2}{(2n+1)^2} - \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{\sin n\pi x}{n}$$

In Figura B.14 sono rappresentate le somme parziali per  $n = 0, 1, 2, 3, 4, 5$ . In particolare, per  $x = 0$  la serie converge a  $\frac{1}{2}$ , da cui la seguente interessante identità

$$1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \dots = \frac{\pi^2}{8}$$

■

<sup>11</sup>e utilizzando le note relazioni trigonometriche  $\sin \alpha \cos \beta = [\sin(\alpha + \beta) + \sin(\alpha - \beta)]/2$ ,  $\sin \alpha \sin \beta = [\cos(\alpha - \beta) - \cos(\alpha + \beta)]/2$ ,  $\cos \alpha \cos \beta = [\cos(\alpha + \beta) + \cos(\alpha - \beta)]/2$ .



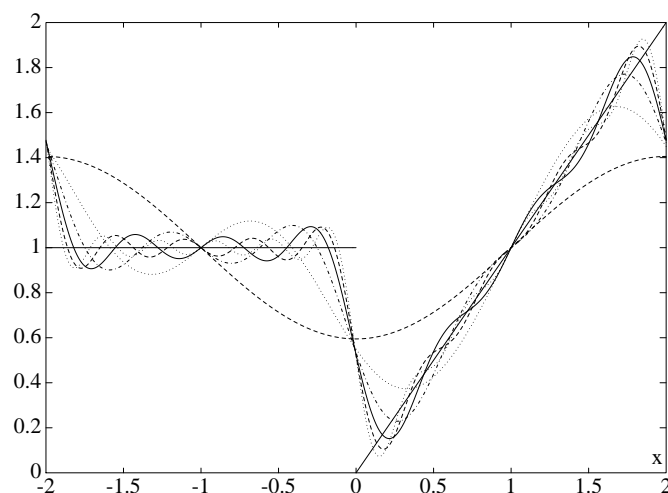


Figura B.14: Approssimanti di Fourier, per  $n = 0, 1, 2, 3, 4, 5$ , della funzione (B.52).

**Prolungamento periodico** Le funzioni  $\cos n\pi x/l$  e  $\sin n\pi x/l$ ,  $n = 1, 2, \dots$ , sono *periodiche* di periodo  $2l$ , come segue immediatamente dalla seguente identità

$$\cos \frac{n\pi}{l}(x + 2l) = \cos \left( \frac{n\pi x}{l} + 2n\pi \right) = \cos \frac{n\pi x}{l}$$

e dall'analogia per la funzione seno. Se, allora, a partire da una funzione  $f(x)$  che verifica le condizioni del Teorema B.4 sull'intervallo  $-l \leq x \leq l$ , definiamo la funzione  $\tilde{f}(x)$  nel seguente modo

$$\tilde{f}(x) = \begin{cases} f(x) & \text{per } -l < x < l \\ \frac{1}{2}[f(l) + f(-l)] & \text{per } x = \pm l \end{cases}, \quad \tilde{f}(x + 2l) = \tilde{f}(x)$$

la serie di Fourier (B.49) converge (nel senso del Teorema B.4) per ogni  $x$  alla funzione periodica  $\tilde{f}(x)$ , detta *prolungamento periodico* della funzione  $f(x)$  (cfr. Figura B.15 per esemplificazioni).

### Funzioni pari e dispari

**Definizione B.3** Una funzione  $f$  è detta *pari* se  $f(-x) = f(x)$ .

Sono funzioni pari, ad esempio, le funzioni  $f(x) = x^2$  e  $f(x) = \cos n\pi x/l$ .

**Definizione B.4** Una funzione  $f$  è detta *dispari* se  $f(-x) = -f(x)$ .

Sono funzioni dispari, ad esempio, le funzioni  $f(x) = x$  e  $f(x) = \sin n\pi x/l$ .

Si vede facilmente che se  $f(x)$  è una funzione dispari si ha  $\int_{-l}^l f(x) dx = 0$ , mentre per una funzione pari si ha  $\int_{-l}^l f(x) dx = 2 \int_0^l f(x) dx$ . Da tali proprietà si ricava immediatamente il seguente risultato.

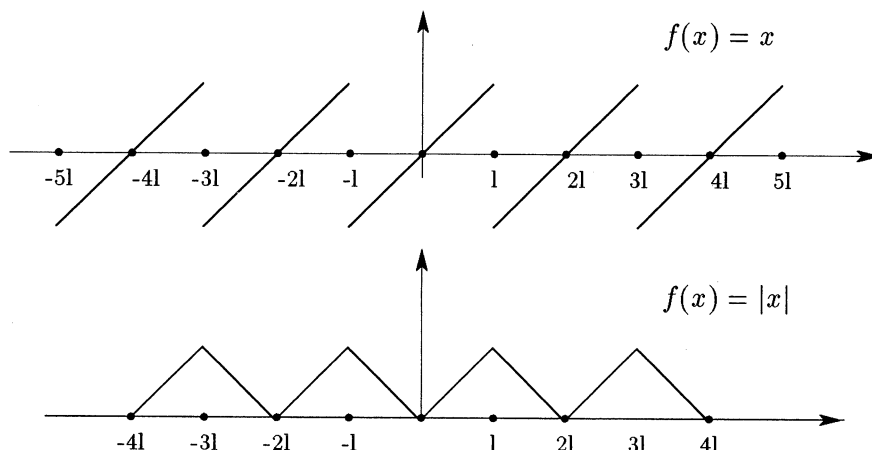


Figura B.15: Esempi di prolungamenti periodici.

**Lemma B.1** *La serie di Fourier di una funzione pari è una serie di soli coseni, ossia non contiene termini della forma  $\sin n\pi x/l$ .*

*Analogamente, la serie di Fourier di una funzione dispari è una serie di soli seni.*

Dal lemma precedente si ricava il seguente risultato importante per le applicazioni alle equazioni differenziali.

**Teorema B.5** *Sia  $f(x)$  una funzione continua a tratti insieme alla derivata  $f'(x)$  nell'intervallo  $0 \leq x \leq l$ . Allora, su tale intervallo  $f(x)$  può essere sviluppata sia in una serie di soli coseni*

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{l}, \quad a_n = \frac{2}{l} \int_0^l f(x) \cos \frac{n\pi x}{l} dx, \quad n = 0, 1, 2, \dots$$

*che di soli seni*

$$\sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{l}, \quad b_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx, \quad n = 1, 2, \dots$$

► **Esempio B.26** Per la funzione  $f(x) = 1$  sull'intervallo  $0 < x < \pi$  si ha lo sviluppo  $f(x) = \sum_{n=1}^{\infty} b_n \sin nx$ , ove

$$b_n = \frac{2}{\pi} \int_0^{\pi} \sin nx dx = \frac{2}{n\pi} (1 - \cos n\pi) = \begin{cases} 0 & n \text{ pari} \\ 4/n\pi & n \text{ dispari} \end{cases}$$

da cui

$$1 = \frac{4}{\pi} \left[ \frac{\sin x}{1} + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \dots \right], \quad 0 < x < \pi$$

■

◆ **Esercizio B.13** *Mostrare che per la seguente funzione*

$$f(x) = \begin{cases} -2x/l, & x \in [-l/2, 0) \\ 2x/l, & x \in [0, l/2] \end{cases}$$

si ha il seguente sviluppo in serie di Fourier

$$f(x) = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{2 \cos(\pi k) - 2}{(\pi k)^2} \cos\left(\frac{2\pi k x}{l}\right)$$

◆ **Esercizio B.14** *Mostrare che per la funzione  $f(x) = x^2$  sull'intervallo  $[-l/2, l/2]$  si ha il seguente sviluppo in serie di Fourier*

$$f(x) = \frac{l^2}{12} + \sum_{k=1}^{\infty} \frac{\cos(\pi k)}{(\pi k/l)^2} \cos\left(\frac{2\pi k x}{l}\right)$$

### B.10.1 Equazione della diffusione

Consideriamo il seguente problema a valori iniziali e ai limiti

$$\frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2}; \quad u(x, 0) = f(x), \quad 0 < x < l; \quad u(0, t) = u(l, t) = 0 \quad (\text{B.53})$$

che (cfr. Capitolo 7) descrive la propagazione del calore in una sbarra di lunghezza  $l$ , quando è nota la distribuzione iniziale  $f(x)$  della temperatura e gli estremi della sbarra sono tenuti a temperatura nulla.

Una tecnica analitica per risolvere il problema (B.53) consiste nel cercare la soluzione nella forma  $u(x, t) = X(x)T(t)$ , ove  $X(x)$  è una funzione della sola variabile  $x$  e analogamente  $T(t)$  è una funzione della sola  $t$  (da cui il nome di *separazione delle variabili* dato alla tecnica).

Si vede facilmente che  $u(x, t) = X(x)T(t)$  è una soluzione dell'equazione  $u_t = \alpha^2 u_{xx}$  se

$$X T' = \alpha^2 X'' T \Rightarrow \frac{X''}{X} = \frac{T'}{\alpha^2 T} \quad (\text{B.54})$$

ove  $T' = dT/dt$ ,  $X'' = d^2X/dx^2$ . Dalla (B.54) si ricava

$$\frac{X''}{X} = -\lambda, \quad \text{e} \quad \frac{T'}{\alpha^2 T} = -\lambda \quad (\text{B.55})$$

con  $\lambda$  costante. Inoltre, le condizioni ai limiti implicano che  $X(0) = 0$  e  $X(l) = 0$ . Pertanto,  $u(x, t) = X(x)T(t)$  è una soluzione del problema

$$\frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2}; \quad u(0, t) = u(l, t) = 0 \quad (\text{B.56})$$

se

$$X'' + \lambda X = 0, \quad X(0) = 0, \quad X(l) = 0 \quad (\text{B.57})$$

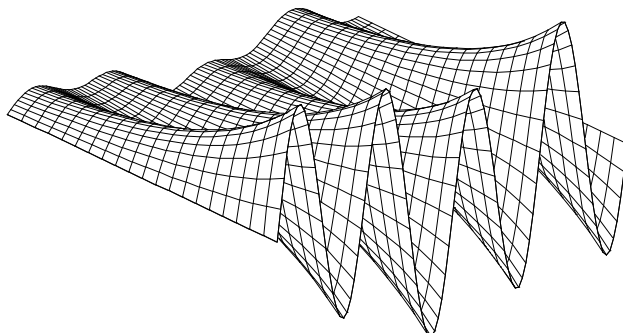


Figura B.16: Rappresentazione della soluzione del problema della diffusione (B.60).

e

$$T' + \lambda \alpha^2 T = 0 \quad (\text{B.58})$$

Si può mostrare (cfr. Capitolo 7) che il problema ai limiti (B.57) ha una soluzione non identicamente nulla  $X(x)$  solo se  $\lambda = \lambda_n = n^2 \pi^2 / l^2$ ,  $n = 1, 2, \dots$ ; in tal caso si ha

$$X(x) = X_n(x) = \sin \frac{n\pi x}{l}$$

Dall'equazione (B.58) si ottiene

$$T(t) = T_n(t) = e^{-\alpha^2 n^2 \pi^2 t / l^2}$$

In definitiva, la funzione

$$u_n(x, t) = \sin \frac{n\pi x}{l} e^{-\alpha^2 n^2 \pi^2 t / l^2}$$

è una soluzione non identicamente nulla del problema (B.56) per ogni intero positivo  $n$ , ed essendo il problema lineare si ha che la funzione

$$u(x, t) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{l} e^{-\alpha^2 n^2 \pi^2 t / l^2}$$

è (formalmente) una soluzione per ogni scelta delle costanti  $c_1, c_2, \dots$ . Tali costanti possono essere scelte in maniera da verificare la condizione iniziale  $u(x, 0) = f(x)$ ,  $0 < x < l$ . Se  $f(x)$  verifica le condizioni del Teorema B.5, posto

$$c_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx$$

si ha che la serie di Fourier  $\sum_{n=1}^{\infty} c_n \sin n\pi x/l$  converge a  $f(x)$  in ogni punto di continuità della funzione, e quindi in conclusione la funzione

$$u(x, t) = \frac{2}{l} \sum_{n=1}^{\infty} \left[ \int_0^l f(x) \sin \frac{n\pi x}{l} dx \right] \sin \frac{n\pi x}{l} e^{-\alpha^2 n^2 \pi^2 t/l^2} \quad (\text{B.59})$$

può essere ritenuta (dopo le opportune giustificazioni sui vari passaggi ai limiti richiesti) la soluzione cercata del problema (B.53).

Come semplice illustrazione, consideriamo il problema

$$\begin{aligned} \frac{\partial u}{\partial t} &= 0.2 \frac{\partial^2 u}{\partial x^2} \\ u(x, 0) &= \sin 3\pi x + 5 \sin 8\pi x, \quad 0 < x < 1 \\ u(0, t) &= u(1, t) = 0 \end{aligned} \quad (\text{B.60})$$

In questo caso si ottiene direttamente la soluzione

$$u(x, t) = \sin 3\pi x e^{-9(0.2)\pi^2 t} + 5 \sin 8\pi x e^{-64(0.2)\pi^2 t}$$

rappresentata in Figura B.16 per  $0 \leq t \leq 0.2$ .

### B.10.2 Equazione delle onde

Consideriamo il seguente problema a valori iniziali e ai limiti

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}; \quad \begin{cases} u(x, 0) = f(x), & u_t(x, 0) = g(x), \quad 0 \leq x \leq l \\ u(0, t) = u(l, t) = 0 \end{cases} \quad (\text{B.61})$$

che (cfr. Capitolo 7) descrive la propagazione delle onde in vari mezzi e le vibrazioni meccaniche di una corda elastica. Vedremo ora brevemente come applicare a tale problema il metodo della separazione delle variabili.

Posto  $u(x, t) = X(x)T(t)$ , dall'equazione differenziale (B.61) si ottiene l'equazione

$$\frac{T''}{c^2 T} = \frac{X''}{X} \quad (\text{B.62})$$

dalla quale

$$\frac{T''}{c^2 T} = -\lambda = \frac{X''}{X}$$

con  $\lambda$  costante. Dalle condizioni ai limiti si ha inoltre

$$0 = u(0, t) = X(0)T(t), \quad 0 = u(l, t) = X(l)T(t)$$

da cui  $X(0) = X(l) = 0$ . Pertanto  $u(x, t) = X(x)T(t)$  è una soluzione del seguente problema

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}; \quad u(0, t) = u(l, t) = 0 \quad (\text{B.63})$$

se

$$X'' + \lambda X = 0, \quad X(0) = X(l) = 0 \quad (\text{B.64})$$

e

$$T'' + \lambda c^2 T = 0 \quad (\text{B.65})$$

Il problema ai limiti (B.64) ha una soluzione non identicamente nulla  $X(x)$  solo se  $\lambda = \lambda_n = n^2\pi^2/l^2$ , e allora

$$X(x) = X_n(x) = \sin \frac{n\pi x}{l}$$

D'altra parte dall'equazione (B.65) si ha

$$T(t) = T_n(t) = a_n \cos \frac{n\pi ct}{l} + b_n \sin \frac{n\pi ct}{l}$$

Pertanto, la funzione

$$u_n(x, t) = \sin \frac{n\pi x}{l} \left[ a_n \cos \frac{n\pi ct}{l} + b_n \sin \frac{n\pi ct}{l} \right]$$

è una soluzione non identicamente nulla di (B.63) per ogni intero positivo  $n$ , e ogni coppia di costanti  $a_n, b_n$ . La combinazione lineare

$$u(x, t) = \sum_{n=1}^{\infty} \sin \frac{n\pi x}{l} \left[ a_n \cos \frac{n\pi ct}{l} + b_n \sin \frac{n\pi ct}{l} \right]$$

soddisfa formalmente il problema ai limiti (B.63) e le condizioni iniziali

$$u(x, 0) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l}, \quad u_t(x, 0) = \sum_{n=1}^{\infty} \frac{n\pi c}{l} b_n \sin \frac{n\pi x}{l}$$

Allora per soddisfare le condizioni iniziali  $u(x, 0) = f(x)$  e  $u_t(x, 0) = g(x)$ , si scelgono le costanti  $a_n$  e  $b_n$  nel seguente modo

$$a_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx, \quad b_n = \frac{2}{n\pi c} \int_0^l g(x) \sin \frac{n\pi x}{l} dx$$

In particolare, per  $g(x) \equiv 0$ , ossia nel caso di una corda vibrante la velocità di rilascio iniziale è nulla, lo spostamento della corda per ogni istante  $t > 0$  è data dalla seguente formula

$$u(x, t) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l} \cos \frac{n\pi ct}{l}, \quad a_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx \quad (\text{B.66})$$

Il primo termine ( $n = 1$ ) rappresenta il primo modo di vibrazione nel quale la corda oscilla intorno alla posizione di equilibrio con frequenza  $\omega_1 = c/2l$  cicli per secondo. Tale frequenza è chiamata la *frequenza fondamentale*, o prima armonica della corda. In modo analogo, la frequenza  $n$ -ma è data da  $\omega_n = n\pi c/2\pi l = n\omega_1$ . Tali frequenze dipendono dalla tensione della corda attraverso la costante  $c$ .

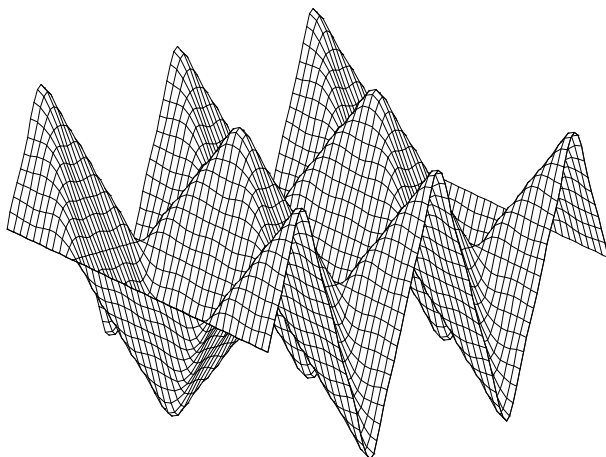


Figura B.17: Rappresentazione della funzione (B.67), per  $n = 6$  e  $0 \leq t \leq 20$ .

Dall'identità trigonometrica

$$\sin \frac{n\pi x}{l} \cos \frac{n\pi ct}{l} = \frac{1}{2} \left[ \sin \frac{n\pi}{l}(x - ct) + \sin \frac{n\pi}{l}(x + ct) \right]$$

e indicando con  $\tilde{f}$  l'estensione periodica dispari di  $f(x)$ , ossia

$$\tilde{f}(x) = \begin{cases} f(x) & 0 < x < l \\ -f(-x) & -l < x < 0 \end{cases} \quad \tilde{f}(x + 2l) = \tilde{f}(x)$$

si ottiene facilmente il seguente risultato

$$u(x, t) = \frac{1}{2} \left[ \tilde{f}(x - ct) + \tilde{f}(x + ct) \right]$$

ove (cfr. Capitolo 7)  $\tilde{f}(x - ct)$  (rispettivamente  $\tilde{f}(x + ct)$ ) descrive un'onda che viaggia con velocità  $c$  nella direzione positiva (rispettivamente negativa) delle  $x$ .

Come illustrazione, la deformazione di una corda elastica di lunghezza  $l = 10$ , innalzata nel punto di mezzo ad una distanza di 1 e quindi rilasciata, è descritta dalla seguente funzione

$$u(x, t) = \frac{8}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \sin \frac{n\pi}{2} \sin \frac{n\pi x}{2} \cos \frac{n\pi t}{10} \quad (\text{B.67})$$

rappresentata, per  $n = 6$  e per  $0 \leq t \leq 20$ , in Figura B.17.

The computer is no better than its program.

E. E. Morison, 1966

## Appendice C

# Algoritmi di ricerca e di ordinamento

Gli algoritmi di ricerca (*searching*) e di ordinamento (*sorting*) costituiscono uno strumento fondamentale per il trattamento dei dati (*data processing*), e il loro studio è uno dei principali obiettivi della Computer Science. La letteratura su tali argomenti è estremamente vasta; segnaliamo, ad esempio, Knuth [99], Sedgewick [143], Gonnet e Baeza-Yates [70]. L'intento di questa appendice è quello di fornire una breve introduzione alle idee che sono alla base dei principali algoritmi, nella convinzione che la conoscenza di tali idee possa, da una parte essere di aiuto alla costruzione di procedimenti numerici *efficienti*, e dall'altra servire alla formazione di una corretta *mentalità algoritmica*.

## C.1 Algoritmi di ricerca

### C.1.1 Ricerca sequenziale

**Basic sequential search** L'algoritmo di base, noto anche come *linear search* o *brute force search*, consiste nella ricerca di un determinato elemento in una *array*<sup>1</sup> esaminando *sequenzialmente* gli elementi dell'array fino a che viene trovato l'elemento cercato, oppure si raggiunge la fine dell'array. L'implementazione dell'algoritmo

---

<sup>1</sup>L'*array* è una delle strutture dati fondamentali, definita come una primitiva in diversi linguaggi di programmazione. Più precisamente, una *array* è definita da un numero fissato di elementi assegnati, memorizzati in maniera contigua e accessibili mediante un indice. Il generico elemento di un'array  $a$  è indicato come  $a[i]$ . La caratteristica principale di una *array* è che *se l'indice è noto*, l'accesso ad ogni elemento avviene in un tempo costante.

Nel seguito del capitolo considereremo, per semplicità di esposizione, in particolare *array numeriche*, per le quali gli elementi  $a[i]$  sono supposti dei numeri reali. Tuttavia, gli algoritmi considerati si estendono a strutture dati più generali (in particolare, le cosiddette strutture *linked list* e *queues*), introducendo per ogni record della struttura una etichetta *key*, che costituisce una piccola parte del record e rispetto alla quale viene eseguito il *searching* o il *sorting*.



può presentare differenti forme, a seconda che si desidera conoscere tutte le volte che un elemento è presente nella struttura, oppure solo la prima volta che si presenta. Se  $n$  è il numero degli elementi nell'array, indichiamo con  $A_n$  la variabile casuale che rappresenta il numero dei confronti fatti in una ricerca con successo e con  $A'_n$  la variabile casuale corrispondente al numero di confronti nel caso di insuccesso, ossia nel caso in cui l'elemento cercato non è presente nell'array. Si ha allora

$$\begin{aligned} P(A_n = i) &= \frac{1}{n} \quad i = 1, 2, \dots, n \\ E(A_n) &= \frac{n+1}{2}, \quad \sigma^2(A_n) = \frac{n^2-1}{12} \\ A'_n &= n \end{aligned}$$

L'algoritmo sequenziale è l'algoritmo di ricerca più semplice, e il suo utilizzo è giustificato in condizioni particolari, ad esempio quando il numero  $n$  degli elementi dell'array è piccolo, oppure quando, nel caso di una ricerca completa, è previsto un numero elevato ( $O(n)$ ) di presenze dell'elemento cercato. Un altro vantaggio dell'algoritmo sequenziale è il fatto che esso non richiede restrizioni sull'ordine nel quale gli elementi sono memorizzati nella struttura dati assegnata.

**Self-organizing sequential search** Nel caso in cui si abbiano successivi accessi alla stessa struttura dati, l'efficienza del metodo sequenziale può essere migliorata modificando opportunamente l'ordine degli elementi nell'array. Ad esempio, nel cosiddetto metodo *move-to-front* in caso di successo l'elemento trovato viene spostato all'inizio dell'array, con conseguente slittamento degli elementi che precedevano l'elemento trovato. La procedura trova la sua motivazione nel fatto che se l'accesso ad alcuni elementi è più frequente di altri, muovendo tali elementi all'inizio dell'array verrà ridotto il tempo nelle ricerche successive. Un'idea analoga è utilizzata nel cosiddetto *transpose method*, nel quale in caso di successo l'elemento trovato viene trasposto immediatamente con quello che precede nell'array.

**Optimal sequential search** Indichiamo con  $p_i$  la probabilità che l'elemento in posizione  $i$ -ma venga assunto come elemento di ricerca (probabilità di accesso). Nel caso in cui le probabilità  $p_i$  siano note a priori ed inoltre sia noto che tali probabilità rimangono invariate nelle ricerche successive, si può ottimizzare l'efficienza del metodo sequenziale *riordinando* preliminarmente gli elementi in ordine decrescente rispetto alle probabilità di accesso (in maniera che nella prima posizione vi sia l'elemento con probabilità maggiore di essere cercato, ecc.). In questo modo si ottiene

$$E(A_n) = \mu'_1 = \sum_{i=1}^n i p_i, \quad A'_n = n, \quad \sigma^2(A_n) = \sum_{i=1}^n i^2 p_i - (\mu'_1)^2$$

I metodi move-to-front e transpose visti in precedenza rappresentano una ricerca, di tipo adattivo, di ordinamento ottimale, quando le probabilità di accesso non rimangono successivamente costanti.

### C.1.2 Ricerca in strutture dati ordinate

Gli algoritmi considerati in questo paragrafo sono indicati per la ricerca di un elemento di un'array, nell'ipotesi che gli elementi siano disposti secondo un ordine prestabilito. Senza perdita di generalità, supporremo un ordinamento crescente.

**Binary search** L'algoritmo utilizza la procedura di bisezione (*tail recursion*) introdotta nel Capitolo 5 per il calcolo di una radice di una funzione di variabile reale. Ad ogni passo della ricerca, viene eseguito un confronto con l'elemento di mezzo dell'array; l'algoritmo decide allora quale delle due metà dell'array contiene l'elemento richiesto e scarta l'altra metà. Il procedimento è ripetuto, dividendo per metà il numero degli elementi da esaminare, fino ad arrivare ad un solo elemento. Se l'array contiene  $n$  elementi e<sup>2</sup>  $k = \lfloor \log_2 n \rfloor$ , si ha

$$\begin{aligned} \lfloor \log_2(n+1) \rfloor &\leq A_n = A'_n \leq k+1 \\ E(A_n) = E(A'_n) &= k+2 - \frac{2^{k+1}}{n+1} \approx \log_2 n, \quad \sigma^2(A'_n) \leq \frac{1}{12} \end{aligned}$$

L'algoritmo binary search è un algoritmo di ricerca *ottimale* quando si utilizzano solo operazioni di confronto; una efficienza superiore può essere ottenuta soltanto utilizzando operazioni speciali (cfr. gli algoritmi successivi *interpolation search* e *hashing*). Inoltre, è un algoritmo *stabile*, nel senso che i tempi di ricerca sono "raccolti" intorno al tempo di ricerca medio (la varianza è  $O(1)$ ).

Come illustrazione dell'algoritmo, consideriamo l'implementazione in FORTRAN del seguente problema di ricerca tabellare (importante, ad esempio, nell'applicazione delle funzioni polinomiali a tratti (spline), cfr. Capitolo 4). Data una array di ascisse  $XX(J)$ ,  $J=1, 2, \dots, N$ , con elementi crescenti monotonamente oppure decrescenti monotonamente, e dato un numero  $X$ , si tratta di trovare un intero  $J$  tale che  $X$  sia fra  $XX(J)$  e  $XX(J+1)$ .

Possiamo definire due valori fittizi  $XX(0)$  e  $XX(N+1)$  ponendo, se la successione  $XX(j)$  è crescente,  $XX(0)$  uguale a  $-\infty$  e  $XX(N+1)$  uguale a  $+\infty$ , scambiando, naturalmente i due valori nel caso in cui la successione sia decrescente. In questo modo l'indice cercato  $J$  è sempre compreso fra 0 e  $N$ , estremi inclusi. Una risposta  $J=0$ , oppure  $J=N$ , indica che il valore  $X$  è fuori scala.

---

<sup>2</sup>Ricordiamo che, dato un numero reale  $x$ , si definisce  $\lfloor x \rfloor$  (*floor* di  $x$ ) il più grande intero minore o uguale a  $x$ , mentre  $\lceil x \rceil$  (*ceil* di  $x$ ) è il più piccolo intero maggiore o uguale a  $x$ ; ad esempio,  $\lfloor \sqrt{2} \rfloor = 1$ ,  $\lceil \sqrt{2} \rceil = 2$ .

**Algoritmo C.1** (ricerca tabellare) (FORTRAN)

```

SUBROUTINE FIND(XX,N,X,J)
DIMENSION XX(N)
C
C    dato un vettore XX di lunghezza N e un valore X ritorna un valore
C    J tale che X e' fra XX(J) e XX(J+1).
C    Il vettore XX deve essere monotono, crescente o decrescente.
C    J=0 oppure J=N indica che X e' all'esterno del vettore XX.
C
C    inizializzazione
JL=0
JU=N+1
C
20  IF(JU-JL.GT.1)THEN
C    bisezione
    JM=(JU+JL)/2
    IF((XX(N).GT.XX(1)).EQV.(X.GT.XX(JM)))THEN
        JL=JM
    ELSE
        JU=JM
    ENDIF
C    si ripete la bisezione fino alla verifica della condizione 20
GO TO 20
C
ENDIF
J=JL
RETURN
END

```

Ricordiamo che la relazione logica  $L1.EQV.L2$  è vera quando le espressioni logiche  $L1$  e  $L2$  sono o ambedue vere o ambedue false. Il suo utilizzo permette di trattare successioni  $XX$  sia crescenti che decrescenti.

**Interpolation search** Il metodo, noto anche come *estimated entry search*, corrisponde ai metodi di interpolazione esaminati nel Capitolo 5, nei quali per la ricerca dell'intervallo contenente uno zero della funzione si tiene conto non solo del segno, ma anche dei valori assunti dalla funzione. In pratica, è quello che avviene quando si consulta ad esempio un elenco telefonico; se il nome cercato incomincia con  $B$ , si cerca vicino all'inizio, ma se comincia con  $Z$ , è più conveniente una ricerca verso la fine.

Se con  $\bar{a}$  viene indicato l'elemento da cercare ed  $[l, r]$  ( $l < r$ ) è l'intervallo di indici iniziale, utilizzando una interpolazione lineare si assume, come elemento di confronto,  $a[j]$ , ove l'indice  $j$  è dato da  $j = l + (\bar{a} - a[l])(r - l)/(a[r] - a[l])$ . Naturalmente, l'algoritmo può risultare non definito quando vi sono elementi ripetuti.

Come illustrazione dell'algoritmo, consideriamo il seguente vettore

```

a=[ 1  1  1  3  5  5  5  7  8  9  12  13  14  16  18  19  24 ]
   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17

```

e  $\bar{a} = 13$ , corrispondente a  $a[12]$ . L'applicazione dell'interpolazione lineare a partire dall'intervallo  $[1, 17]$  fornisce successivamente i valori arrotondati  $j = 9, 11, 12$  e la ricerca è completata in tre passi. Più in generale, si può dimostrare che per l'algoritmo interpolation search si ha  $E(A_n) = \log_2 \log_2 n + O(1) \approx \log_2 \log_2(n + 3)$ . In pratica, tuttavia, l'efficienza del metodo dipende dalla distribuzione degli elementi  $a[i]$  (così come la convergenza degli algoritmi di interpolazione per la ricerca dello zero di una funzione dipende dalla regolarità della funzione). Una variante del metodo, nota come *interpolation-sequential search*, abbina l'idea dell'interpolazione a quella della ricerca sequenziale, nel senso che ad una prima applicazione dell'interpolazione, segue una ricerca sequenziale nel verso appropriato. Tale variante può presentare interesse in certe applicazioni, ad esempio in operazioni di I/O su un calcolatore.

**Altri algoritmi** Per concludere l'introduzione agli algoritmi di ricerca segnaliamo due famiglie interessanti di metodi, basati rispettivamente sull'utilizzo della **hashing function** e della ricorsione (**recursive structures search**). Per una trattazione adeguata di tali metodi rinviamo alla bibliografia già segnalata.

## C.2 Algoritmi di ordinamento

Data una array numerica  $ar[i]$ ,  $i = 1, 2, \dots, n$ , gli algoritmi considerati in questo paragrafo modificano l'array data permutando i suoi elementi in maniera da ottenere un ordinamento, per fissare le idee, crescente. In pratica, anziché permutare gli elementi dell'array, si può costruire una tabella di indici (*index table*)  $I_j$ ,  $j = 1, 2, \dots, n$ , in maniera che l'array  $ar[I_j]$ ,  $j = 1, 2, \dots, n$ , sia in ordine crescente quando indicizzata da  $j$ .

Gli algoritmi più efficienti, in particolare l'algoritmo *heapsort* e l'algoritmo *quicksort*, raggiungono lo scopo con un numero medio di operazioni  $O(n \log_2 n)$ . Alla introduzione di tali algoritmi premettiamo, tuttavia, una breve discussione di alcuni metodi più elementari di ordine  $O(n^2)$ , e che possono essere di utilità per  $n$  non eccessivamente elevato.

**Bubble sort** L'algoritmo bubble sort ordina una array mediante successivi scambi degli elementi adiacenti che non sono in ordine corretto. Ogni passaggio completo porta almeno un elemento nella sua locazione finale. In questo modo, ogni passaggio ha almeno un elemento in meno del precedente. Indicando con  $C_n$  il numero dei confronti necessari per ordinare una array di dimensione  $n$  si ha

$$n - 1 \leq C_n \leq \frac{n(n - 1)}{2}$$

Un esempio di implementazione del metodo è fornito di seguito in un linguaggio descrittivo, ove  $[lo, up]$  rappresenta l'intervallo di indici per i quali si vuole effettuare

l'ordinamento, ad esempio  $lo = 1$  e  $up = n$ . Inoltre, nella implementazione indicata il metodo esegue un passaggio dall'inizio alla fine, e quindi esegue un ritorno dalla fine all'inizio. In questa forma l'algoritmo è anche noto come *double-direction bubble sort* (o cocktail shaker sort).

**Algoritmo C.2** (double direction bubble sort)

```

while  $up > lo$  do
   $j := lo$ 
  for  $i := lo$  to  $up - 1$  do
    if  $ar[i] > ar[i + 1]$  then
       $t := ar[i + 1]$ 
       $ar[i] := ar[i + 1]$ 
       $ar[i + 1] := t$ 
       $j := i$ 
    end if
  end for
   $up := j$ 
  for  $i := up$  downto  $lo + 1$  do
    if  $ar[i] < ar[i - 1]$  then
       $t := ar[i]$ 
       $ar[i] := ar[i - 1]$ 
       $ar[i - 1] := t$ 
       $j := i$ 
    end if
  end for
   $lo := j$ 
end while

```

L'algoritmo bubble sort è un algoritmo semplice, ma non efficiente per  $n$  grande, in quanto il tempo richiesto è  $O(n^2)$ . Tuttavia, per una array con pochi elementi fuori posto la versione double-direction può essere efficiente; in effetti, se solo  $k$  degli  $n$  elementi sono fuori ordine, il tempo richiesto dall'algoritmo double-direction è  $O(kn)$ .

**Selection sort** Uno degli algoritmi di ordinamento più semplici, utilizza la seguente idea: si trova l'elemento più piccolo nella array e si scambia con l'elemento nella prima posizione, ripetendo successivamente l'operazione sui rimanenti elementi fino a che l'intera array è ordinata, come illustrato nella seguente implementazione in un linguaggio simbolico.

**Algoritmo C.3** (Selection sort)

```

for  $i := 1$  to  $n - 1$  do
   $min := i$ 
  for  $j := i + 1$  to  $n$  do
    if  $ar[j] < ar[min]$  then  $min := j$ 
  end for
  swap  $ar[i]$  and  $ar[min]$ 
end for

```

```

    end for
    t := ar[min]
    ar[min] := ar[i]
    ar[i] := t
  end for

```

Si può vedere che l'algoritmo selection sort utilizza approssimativamente  $n^2/2$  confronti e  $n$  scambi. Un aspetto interessante dell'algoritmo è il fatto che ogni elemento è spostato al più una volta; questo può essere vantaggioso ad esempio nell'ordinamento di file con record molto grandi e piccole chiavi (key) di individuazione.

**Linear insertion sort** Il metodo linear insertion corrisponde alla procedura usualmente impiegata per ordinare un mazzo di carte: si considera una carta alla volta, inserendola tra quelle già esaminate (e che vengono mantenute ordinate). L'elemento considerato è inserito semplicemente muovendo gli elementi più grandi di una posizione a destra e allora inserendo l'elemento nella posizione vuota. L'algoritmo è illustrato dalla seguente implementazione in FORTRAN.

**Algoritmo C.4** (linear insertion sort) (FORTRAN)

```

SUBROUTINE LINEASR(N,AR)
DIMENSION AR(N)
C   ordina una array AR di lunghezza N in ordine numerico crescente
C   mediante il metodo linear insertion. L'array AR e' sostituita
C   dalla array ordinata.
DO 20 J=2,N
  T=AR(J)
  DO 15 I=J-1,1,-1
    IF(AR(I).LE.T) GO TO 10
    AR(I+1)=AR(I)
15  CONTINUE
  I=0
10  AR(I+1)=T
20  CONTINUE
RETURN
END

```

L'algoritmo linear insertion utilizza in media circa  $n^2/4$  confronti e  $n^2/8$  scambi. Come per il metodo bubble sort, l'algoritmo linear sort è stabile, nel senso che gli elementi uguali rimangono dopo l'ordinamento nello stesso ordine relativo.

Una variante del metodo linear sort, comunemente utilizzata, consiste nella ricerca della posizione finale di ogni elemento mediante l'algoritmo binary search che abbiamo analizzato in precedenza. Tale variante, nota come *binary insertion sort*, usa un numero di confronti "quasi" ottimale, ma non riduce il numero degli scambi necessari per far posto all'elemento inserito, per cui il tempo totale impiegato rimane  $O(n^2)$ .

**Quicksort** L'algoritmo quicksort, introdotto da C. A. R. Hoare (1961), utilizza una tecnica di *divide et impera* (divide-and-conquer). All'inizio di ogni iterazione viene selezionato un elemento, detto elemento *perno* (pivot). L'array è allora suddivisa in due parti, una delle quali contiene gli elementi che sono più piccoli del pivot, e l'altra gli elementi maggiori. In questo modo, l'elemento scelto viene sistemato nella posizione appropriata tra le due parti risultanti. La procedura è quindi applicata in maniera ricorsiva sulle due sottomatrici.

Se  $C_n$  indica il numero dei confronti necessari per ordinare una array casuale, e  $k = \lfloor \log_2 n \rfloor$ , si può dimostrare che

$$(n+1)k - 2^{k+1} + 2 \leq C_n \leq \frac{n(n-1)}{2}$$

$$E(C_n) = n - 1 + \frac{2}{n} \sum_{k=1}^{n-1} E(C_k) \approx 2n \log_2 n$$

da cui si vede che l'algoritmo quicksort esegue un numero medio di confronti che cresce come  $n \log_2 n$ ; in effetti, l'algoritmo stabilisce una relazione di ordinamento tra un grande numero di coppie di elementi senza confrontarli direttamente, ma inferendo tale relazione da confronti già fatti con altri elementi e sfruttando la proprietà transitiva. Nel caso peggiore (corrispondente al caso in cui l'array è già ordinata!) l'algoritmo ha un costo  $O(n^2)$ .

Nel caso peggiore l'algoritmo quicksort può usare  $O(n)$  livelli di ricorsione, e questo implica  $O(n)$  memorizzazioni aggiuntive. L'aumento dei livelli di ricorsione può essere diminuito scegliendo come sotto-array da ordinare ricorsivamente quella di dimensioni più piccole. In questo modo si ottengono  $O(\log_2 n)$  livelli. Nel seguito è riportata, come esemplificazione, una implementazione in linguaggio PASCAL, che permette l'utilizzo della procedura in maniera ricorsiva.

Successivamente, è riportata una implementazione in linguaggio FORTRAN, nella quale è utilizzata una memoria ausiliaria come catasta (*stack*) per memorizzare le sotto-array ancora da esaminare; inoltre, in tale implementazione, quando una sotto-array ha una dimensione inferiore a un intero prefissato  $M$ , viene applicato (seguendo Knuth [99]) il procedimento di linear insertion che abbiamo esaminato in precedenza. Infine, per l'individuazione del pivot viene utilizzata una procedura di generazione di numeri casuali.

**Algoritmo C.5** (quicksort) (Pascal)

```

procedure sort (var ar : ArrayToSort; lo,up : integer);
var i,j : integer;
    tem : real;
begin
while up > lo do begin
    i := lo;
    j := up;

```

```

    tem := ar[lo];
    {**** divisione del file in due ****}
    while i < j do begin
        while ar[j] > tem do
            j := j - 1;
        ar[i] := ar[j];
        while (i < j) and (ar[i] <= tem) do
            i := i + 1;
        ar[j] := ar[i];
        end;
    ar[i] := tem
    {**** sort recursivo ****}
    if i - lo < up - i then begin
        sort(ar, lo, i - 1);
        lo := i + 1
    end
    else begin
        sort(ar, i + 1, up);
        up := i - 1
    end
    end
end;

```

### Algoritmo C.6 (quicksort) (FORTRAN)

```

SUBROUTINE QUSRT(N,AR)
C   ordina un array AR di lunghezza N in ordine crescente mediante
C   l'algoritmo quicksort; AR e' sostituita dall'array ordinata.
PARAMETER (M=7,NST=50,FM=7875.,FA=211.,FC=1663.,FMI=1.2698413E-4)
C   M: la dimensione della sottomatrice ordinata mediante
C   l'algoritmo linear insertion.
C   NST: dimensione della memoria ausiliaria.
C   FM, FA, FC, FMI: parametri utilizzati per la generazione di
C   numeri casuali.
DIMENSION AR(N),ISTA(NST)
JST=0
L=1
IR=N
FX=0.
C   ordinamento mediante linear insertion
10 IF(IR-L.LT.M)THEN
    DO 13 J=L+1,IR
        A=AR(J)
        DO 11 I=J-1,1,-1
            IF(AR(I).LE.A)GO TO 12
            AR(I+1)=AR(I)
11 CONTINUE
        I=0
12 AR(I+1)=A

```



```
13     CONTINUE
      IF(JST.EQ.0)RETURN
      IR=ISTA(JST)
      L=ISTA(JST-1)
      JST=JST-2
      ELSE
        I=L
        J=IR
C     generazione numeri casuali
      FX=MOD(FX*FA+FC,FM)
      IQ=L+(IR-L+1)*(FX*FMI)
      A=AR(IQ)
      AR(IQ)=AR(L)
20     CONTINUE
21     IF(J.GT.0)THEN
          IF(A.LT.AR(J))THEN
            J=J-1
            GO TO 21
          ENDIF
        ENDIF
      IF(J.LE.I)THEN
        AR(I)=A
        GO TO 30
      ENDIF
      AR(I)=AR(J)
      I=I+1
22     IF(I.LE.N)THEN
          IF(A.GT.AR(I))THEN
            I=I+1
            GO TO 22
          ENDIF
        ENDIF
      IF(J.LE.I)THEN
        AR(J)=A
        I=J
        GO TO 30
      ENDIF
      AR(J)=AR(I)
      J=J-1
      GO TO 20
30     JST=JST+2
      IF(JST.GT.NST)PAUSE 'NST deve essere superiore.'
      IF(IR-I.GE.I-L)THEN
        ISTA(JST)=IR
        ISTA(JST-1)=I+1
        IR=I-1
      ELSE
        ISTA(JST)=I-1
        ISTA(JST-1)=L
        L=I+1
      ENDIF
```

```

ENDIF
GO TO 10
END

```

**Shellsort** L'algoritmo shellsort (D. L. Shell 1959), o *diminishing increment sort*, rappresenta una interessante variante dell'algoritmo linear insertion. L'idea di base è la seguente. Supponiamo, ad esempio, di dover ordinare i numeri  $n_1, n_2, \dots, n_{16}$ . Mediante il metodo linear insertion si ordina dapprima ognuno degli 8 gruppi di due elementi  $(n_1, n_9), (n_2, n_{10}), \dots, (n_8, n_{16})$ . Successivamente, si ordina ognuno dei 4 gruppi di 4 elementi  $(n_1, n_5, n_9, n_{13}), \dots, (n_4, n_8, n_{12}, n_{16})$ . In modo analogo, si ordinano i due gruppi di 8 elementi; infine, si ordina tutta l'array di 16 elementi.

Più in generale, l'array è vista come una collezione di  $d$  sotto-array interconnesse e tali che la prima è costituita dagli elementi di indici  $1, d+1, 2d+1, \dots$ , la seconda dagli elementi di indici  $2, d+2, 2d+2, \dots$ , e così di seguito. L'algoritmo linear insertion è applicato ad ognuno di tali sotto-array per diversi valori di  $d$ . Differenti sequenze di incrementi  $d$  possono dare differenti prestazioni dell'algoritmo. Se ad esempio si assume  $d = \{2^k, 2^{k-1}, \dots, 8, 4, 2, 1\}$  e  $n = 2^k$ , si può mostrare che per il numero  $C_n$  dei confronti e il numero degli scambi  $I_n$  si ha

$$E(I_n) = O(n^{3/2}), \quad E(C_n) = E(I_n) + n \log_2 n - \frac{3(n-1)}{2}$$

Rinviando per maggiori dettagli ad esempio a Knuth [99], riportiamo come esemplificazione una implementazione in FORTRAN.

#### Algoritmo C.7 (shellsort) (FORTRAN)

```

SUBROUTINE SHELL(N,AR)
C   ordina una array AR di dimensione N in ordine crescente
C   mediante l'algoritmo Shellsort (diminishing increment sort).
PARAMETER (ALN2I=1.4426950, TINY=1.E-5)
DIMENSION AR(N)
LOGNB2=INT(ALOG(FLOAT(N))*ALN2I+TINY)
M=N
DO 10 NN=1,LOGNB2
  M=M/2
  K=N-M
  DO 20 J=1,K
    I=J
30    CONTINUE
    L=I+M
    IF(AR(L).LT.AR(I)) THEN
      T=AR(I)
      AR(I)=AR(L)
      AR(L)=T
    I=I-M

```

```

                IF(I.GE.1)GO TO 30
            ENDIF
20         CONTINUE
10         CONTINUE
            RETURN
            END

```

**Heapsort** È un metodo elegante e semplice che non usa memoria ausiliaria e non richiede linguaggi che supportano la ricorsività; inoltre garantisce l'ordinamento di array di  $n$  elementi in un numero di confronti minore di  $2n \log_2 n$  passaggi, con  $O(n \log_2 n)$  anche nel caso peggiore.

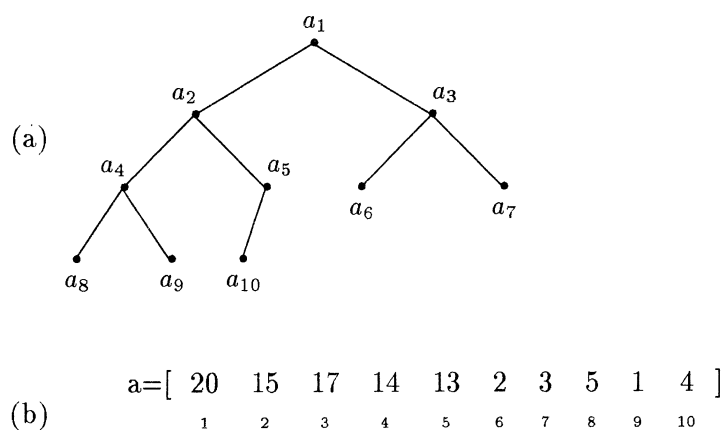


Figura C.1: (a) Rappresentazione di uno heap come albero binario; (b) sua allocazione sequenziale.

Rinviando ad esempio a Knuth [99] per maggiori dettagli, ricorderemo le idee di base. Un insieme di  $n$  numeri  $a_i$ ,  $i = 1, \dots, n$ , è detto formare un **heap** (letteralmente “mucchio”) se soddisfa la seguente relazione

$$a_{\lfloor j/2 \rfloor} \geq a_j \quad \text{per } 1 \leq \left\lfloor \frac{j}{2} \right\rfloor < j \leq n \quad (\text{C.1})$$

Per comprendere la definizione (C.1), si pensi ai numeri  $a_i$  disposti in un albero binario (*binary tree*) completo, nel quale la radice (*root*) è data dall'elemento  $a_1$ , i due nodi sottostanti (*children*) sono gli elementi  $a_2$  e  $a_3$ , eccetera; per una illustrazione si veda la Figura C.1. In questa forma, un heap ha in ogni nodo un elemento più grande o uguale degli elementi corrispondenti ai suoi due nodi sottostanti.

La costruzione di un heap avviene per successive inserzioni partendo da un heap vuoto. Una volta riarrangiata l'array come struttura heap, l'algoritmo di ordinamento procede con successive *estrazioni* e *ricostituzioni* di heap: si estrae il massimo

dallo heap, ritenendolo come massimo dell'insieme, si ricostituisce quindi lo heap e si ripetono le operazioni per trovare i successivi elementi nell'ordinamento.

La procedura è illustrata dalla seguente implementazione in FORTRAN.

**Algoritmo C.8** (heapsort) (FORTRAN)

```

SUBROUTINE HEAPS(N,AR)
C   ordina l'array AR di dimensione N in ordine crescente
C   usando l'algoritmo heapsort. AR e' sostituita dall'array ordinata.
DIMENSION AR(N)
L=N/2+1
IR=N
10  CONTINUE
    IF(L.GT.1)THEN
        L=L-1
        RRA=AR(L)
    ELSE
        RRA=AR(IR)
        AR(IR)=AR(1)
        IR=IR-1
        IF(IR.EQ.1)THEN
            AR(1)=RRA
            RETURN
        ENDIF
    ENDIF
    I=L
    J=L+L
20  IF(J.LE.IR)THEN
        IF(J.LT.IR)THEN
            IF(AR(J).LT.AR(J+1))J=J+1
        ENDIF
        IF(RRA.LT.AR(J))THEN
            AR(I)=AR(J)
            I=J
            J=J+J
        ELSE
            J=IR+1
        ENDIF
        GO TO 20
    ENDIF
    AR(I)=RRA
GO TO 10
END

```

It is difficult to understand why statisticians commonly limit their enquiries to averages and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances could be got rid of at once.

Sir Francis Galton

## Appendice D

# Tabelle statistiche

In questa appendice sono raccolte per comodità le tabelle corrispondenti ad alcune tra le più importanti distribuzioni statistiche.

Più in generale, oltre alle librerie NAG, IMSL e ai vari programmi statistici analizzati nel Capitolo 8, ricordiamo le seguenti fonti di tabelle statistiche

- W. H. Beyer (editore). *CRC Standard Mathematical Tables*. Chemical Rubber Co., Cleveland, Ohio 1973.
- R. A. Fisher, F. Yates. *Statistical Tables for Biological, Agricultural, and Medical Research*. Longman Group Ltd., London 1974.
- J. A. Greenwood, H. O. Hartley. *Guide to Tables in Mathematical Statistics*. University Press, Princeton, N.J. 1962.
- A. Hald. *Statistical Tables and Formulas*. John Wiley & Sons, New York 1952.
- D. B. Owen. *Handbook of Statistical Tables*. Addison-Wesley, Inc., Reading, MA 1962.
- E. S. Pearson, H. O. Hartley. *Biometrika Tables for Statisticians*. Cambridge University Press 1954.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Tabella D.1: Distribuzione normale standard  $Z$ ; la tabella fornisce la probabilità  $P(0 < Z < z)$  per valori di  $z = .00(.01)3.09$ . Per il calcolo di  $P(-\infty < Z < z)$  si tiene conto che  $P(-\infty < Z < z) = 0.5 + P(0 < Z < z)$ . Si ha inoltre  $P(|Z| < z) = 2P(0 < Z < z)$  e  $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z \leq z_1)$ .

$n$	$\alpha$	.45	.40	.35	.3	.25	.2	.15	.1	.05	.025	.01
1		.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821
2		.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965
3		.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541
4		.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747
5		.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365
6		.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143
7		.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998
8		.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896
9		.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821
10		.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764
11		.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718
12		.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681
13		.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650
14		.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624
15		.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602
16		.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583
17		.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567
18		.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552
19		.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539
20		.127	.258	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528
21		.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518
22		.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508
23		.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500
24		.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492
25		.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485
26		.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479
27		.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473
28		.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467
29		.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462
30		.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457
40		.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423
60		.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390
120		.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358
$\infty$		.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326

Tabella D.2: Distribuzione t-student:  $P(t_n > t_{n,\alpha}) = \alpha$ .

$n$	$\alpha$	.99	.95	.70	.30	.05	.025	.01	.001
1		.000157	.00393	.148	1.074	3.841	5.0238	6.635	10.827
2		.0201	.103	.713	2.408	5.991	7.3780	9.210	13.815
3		.115	.352	1.424	3.665	7.815	9.348	11.345	16.266
4		.297	.711	2.195	4.878	9.488	11.143	13.277	18.467
5		.554	1.145	3.000	6.064	11.070	12.832	15.086	20.515
6		.872	1.635	3.828	7.231	12.592	14.449	16.812	22.457
7		1.239	2.167	4.671	8.383	14.067	16.013	18.475	24.322
8		1.646	2.733	5.527	9.524	15.507	17.535	20.090	26.125
9		2.088	3.325	6.393	10.656	16.919	19.023	21.666	27.877
10		2.558	3.940	7.267	11.781	18.307	20.483	23.209	29.588
11		3.053	4.575	8.148	12.899	19.675	21.920	24.725	31.264
12		3.571	5.226	9.034	14.011	21.026	23.337	26.217	32.909
13		4.107	5.892	9.926	15.119	22.362	24.736	27.688	34.528
14		4.660	6.571	10.821	16.222	23.685	26.119	29.141	36.123
15		5.229	7.261	11.721	17.322	24.996	27.448	30.578	37.697
16		5.812	7.962	12.624	18.418	26.296	28.845	32.000	39.252
17		6.408	8.672	13.531	19.511	27.587	30.191	33.409	40.790
18		7.015	9.390	14.440	20.601	28.869	31.526	34.805	42.312
19		7.633	10.117	15.352	21.689	30.144	32.852	36.191	43.820
20		8.260	10.851	16.266	22.775	31.410	34.170	37.566	45.315
21		8.897	11.591	17.182	23.858	32.671	35.479	38.932	46.797
22		9.542	12.338	18.101	24.939	33.924	36.781	40.289	48.268
23		10.196	13.091	19.021	26.018	35.172	38.076	41.638	49.728
24		10.856	13.848	19.943	27.096	36.415	39.364	42.980	51.179
25		11.524	14.611	20.867	28.172	37.652	40.646	44.314	52.620
26		12.198	15.379	21.792	29.246	38.885	41.923	45.642	54.052
27		12.879	16.151	22.719	30.319	40.113	43.194	46.963	55.476
28		13.565	16.928	23.647	31.391	41.373	44.461	48.278	56.893
29		14.256	17.708	24.577	32.461	42.557	45.722	49.588	58.302
30		14.953	18.493	25.508	33.530	43.773	46.979	50.892	59.703

Tabella D.3: Distribuzione  $\chi^2$ :  $P(\chi_n^2 > \chi_{n,\alpha}) = \alpha$ . Per valori di  $n$  maggiori, si tiene conto che la variabile casuale  $\sqrt{2}\chi^2 - \sqrt{2n-1}$  è con buona approssimazione una distribuzione normale standard.



$n$	$m$	1	2	3	4	5	10	16	30	200	$\infty$
1		161	200	216	225	230	242	246	250	254	254
		<b>4052</b>	<b>4999</b>	<b>5403</b>	<b>5625</b>	<b>5764</b>	<b>6056</b>	<b>6169</b>	<b>6261</b>	<b>6352</b>	<b>6366</b>
2		18.51	19.00	19.16	19.25	19.30	19.39	19.43	19.46	19.49	19.50
		<b>98.49</b>	<b>99.00</b>	<b>99.17</b>	<b>99.25</b>	<b>99.30</b>	<b>99.40</b>	<b>99.44</b>	<b>99.47</b>	<b>99.49</b>	<b>99.50</b>
3		10.13	9.55	9.28	9.12	9.01	8.78	8.69	8.62	8.54	8.53
		<b>34.12</b>	<b>30.82</b>	<b>29.46</b>	<b>28.71</b>	<b>28.24</b>	<b>27.23</b>	<b>26.83</b>	<b>26.50</b>	<b>26.18</b>	<b>26.12</b>
4		7.71	6.94	6.59	6.39	6.26	5.96	5.84	5.74	5.65	5.63
		<b>21.20</b>	<b>18.00</b>	<b>16.69</b>	<b>15.98</b>	<b>15.52</b>	<b>14.54</b>	<b>14.15</b>	<b>13.93</b>	<b>13.52</b>	<b>13.46</b>
5		6.61	5.79	5.41	5.19	5.05	4.74	4.60	4.50	4.38	4.36
		<b>16.26</b>	<b>13.27</b>	<b>12.06</b>	<b>11.39</b>	<b>10.97</b>	<b>10.05</b>	<b>9.68</b>	<b>9.38</b>	<b>9.07</b>	<b>9.02</b>
6		5.99	5.14	4.76	4.53	4.39	4.06	3.92	3.81	3.69	3.67
		<b>13.74</b>	<b>10.92</b>	<b>9.78</b>	<b>9.15</b>	<b>8.75</b>	<b>7.87</b>	<b>7.52</b>	<b>7.23</b>	<b>6.94</b>	<b>6.88</b>
8		5.32	4.46	4.07	3.84	3.69	3.34	3.20	3.08	2.96	2.93
		<b>11.26</b>	<b>8.65</b>	<b>7.59</b>	<b>7.01</b>	<b>6.63</b>	<b>5.82</b>	<b>5.48</b>	<b>5.20</b>	<b>4.91</b>	<b>4.86</b>
10		4.96	4.10	3.71	3.48	3.33	2.97	2.82	2.70	2.56	2.54
		<b>10.04</b>	<b>7.56</b>	<b>6.55</b>	<b>5.99</b>	<b>5.64</b>	<b>4.85</b>	<b>4.52</b>	<b>4.25</b>	<b>3.96</b>	<b>3.91</b>
16		4.49	3.63	3.24	3.01	2.85	2.49	2.33	2.20	2.04	2.01
		<b>8.53</b>	<b>6.23</b>	<b>5.29</b>	<b>4.77</b>	<b>4.44</b>	<b>3.69</b>	<b>3.37</b>	<b>3.10</b>	<b>2.80</b>	<b>2.75</b>
30		4.17	3.32	2.92	2.69	2.53	2.16	1.99	1.84	1.66	1.62
		<b>7.56</b>	<b>5.39</b>	<b>4.51</b>	<b>4.02</b>	<b>3.70</b>	<b>2.98</b>	<b>2.66</b>	<b>2.38</b>	<b>2.07</b>	<b>2.01</b>
100		3.94	3.09	2.70	2.46	2.30	1.92	1.75	1.57	1.34	1.28
		<b>6.90</b>	<b>4.82</b>	<b>3.98</b>	<b>3.51</b>	<b>3.20</b>	<b>2.51</b>	<b>2.19</b>	<b>1.89</b>	<b>1.51</b>	<b>1.43</b>
$\infty$		3.84	2.99	2.60	2.37	2.21	1.83	1.64	1.46	1.17	1.00
		<b>6.64</b>	<b>4.60</b>	<b>3.78</b>	<b>3.32</b>	<b>3.02</b>	<b>2.32</b>	<b>1.99</b>	<b>1.69</b>	<b>1.25</b>	<b>1.00</b>

Tabella D.4: Distribuzione  $F(m, n)$ ; valori  $F_{m,n,\alpha}$  tali che  $P(F(m, n) > F_{m,n,\alpha}) = \alpha$  per  $\alpha = 0.05$  e  $\alpha = 0.01$  (in neretto).

n	k	p									
		.01	.05	.10	.15	.20	.25	1/3	.35	.45	.50
2	0	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4444	0.4225	0.3025	0.2500
	1	0.9999	0.9975	0.9900	0.9775	0.9600	0.9375	0.8889	0.8775	0.7975	0.7500
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.2963	0.2746	0.1664	0.1250
	1	0.9997	0.9928	0.9720	0.9392	0.8960	0.8438	0.7407	0.7182	0.5748	0.5000
	2	1.0000	0.9999	0.9990	0.9966	0.9920	0.9844	0.9630	0.9751	0.9089	0.8750
	3		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.1975	0.1785	0.0915	0.0625
	1	0.9994	0.9860	0.9477	0.8905	0.8192	0.7383	0.5926	0.5630	0.3910	0.3125
	2	1.0000	0.9995	0.9963	0.9880	0.9728	0.9492	0.8889	0.8735	0.7585	0.6875
	3		1.0000	0.9999	0.9995	0.9984	0.9961	0.9876	0.9850	0.9590	0.9375
	4			1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.0878	0.0754	0.0277	0.0156
	1	0.9986	0.9672	0.8857	0.7765	0.6553	0.5339	0.3512	0.3191	0.1636	0.1094
	2	1.0000	0.9978	0.9842	0.9527	0.9011	0.8306	0.6804	0.6471	0.4415	0.3438
	3		0.9999	0.9987	0.9941	0.9830	0.9624	0.8999	0.8826	0.7447	0.6562
	4		1.0000	0.9999	0.9996	0.9984	0.9954	0.9822	0.9777	0.9308	0.8906
	5			1.0000	1.0000	0.9999	0.9998	0.9986	0.9982	0.9917	0.9844
	6					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0173	0.0135	0.0025	0.0010
	1	0.9958	0.9139	0.7361	0.5443	0.3758	0.2440	0.1040	0.0860	0.0464	0.0107
	2	1.0000	0.9885	0.9298	0.8202	0.6778	0.5256	0.2991	0.2616	0.0996	0.0547
	3		0.9990	0.9872	0.9500	0.8791	0.7759	0.5593	0.5138	0.2660	0.1719
	4		0.9999	0.9984	0.9901	0.9672	0.9219	0.7869	0.7515	0.5044	0.3770
	5		1.0000	0.9999	0.9986	0.9936	0.9803	0.9234	0.9051	0.7384	0.6230
	6			1.0000	0.9999	0.9991	0.9965	0.8803	0.9740	0.8990	0.8281
	7				1.0000	0.9999	0.9996	0.9966	0.9952	0.9726	0.9453
	8					1.0000	1.0000	0.9996	0.9995	0.9955	0.9893
	9							1.0000	1.0000	0.9997	0.9990
10									1.0000	1.0000	

Tabella D.5: Probabilità cumulativa binomiale  $P(X \leq k) = \sum_{j=0}^k \binom{n}{j} p^j (1-p)^{n-j}$ .

k	$\lambda$											
	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0	1.1	1.2	1.3	1.4
0	0.905	0.819	0.741	0.670	0.607	0.549	0.449	0.368	0.333	0.301	0.273	0.247
1	0.995	0.982	0.963	0.938	0.910	0.878	0.809	0.736	0.699	0.663	0.627	0.592
2	1.000	0.999	0.996	0.992	0.986	0.977	0.953	0.920	0.900	0.879	0.857	0.833
3		1.000	1.000	0.999	0.998	0.997	0.991	0.981	0.974	0.966	0.957	0.946
4				1.000	1.000	1.000	0.999	0.996	0.995	0.992	0.989	0.986
5							1.000	0.999	0.999	0.998	0.998	0.997
6								1.000	1.000	1.000	1.000	0.999
k	1.5	1.6	1.7	1.8	1.9	2.0	2.4	2.8	3.0	3.2	3.4	3.6
0	0.223	0.202	0.183	0.165	0.150	0.135	0.091	0.061	0.050	0.041	0.033	0.027
1	0.553	0.525	0.493	0.463	0.434	0.406	0.308	0.231	0.199	0.171	0.147	0.126
2	0.809	0.783	0.757	0.731	0.704	0.677	0.570	0.469	0.423	0.380	0.340	0.303
3	0.934	0.921	0.907	0.891	0.875	0.857	0.779	0.692	0.647	0.603	0.558	0.515
4	0.981	0.976	0.970	0.964	0.956	0.947	0.904	0.848	0.815	0.781	0.744	0.706
5	0.996	0.994	0.992	0.990	0.987	0.983	0.964	0.935	0.916	0.895	0.871	0.844
6	0.999	0.999	0.998	0.997	0.997	0.995	0.988	0.976	0.966	0.955	0.942	0.927
7	1.000	1.000	1.000	0.999	0.999	0.999	0.997	0.992	0.988	0.983	0.977	0.969
8				1.000	1.000	1.000	0.999	0.998	0.996	0.994	0.992	0.988
9							1.000	0.999	0.999	0.998	0.997	0.996
k	3.8	4.0	4.2	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8	6.0
0	0.022	0.018	0.015	0.012	0.010	0.008	0.007	0.006	0.005	0.004	0.003	0.003
1	0.107	0.092	0.078	0.066	0.056	0.048	0.040	0.034	0.029	0.024	0.021	0.017
2	0.269	0.238	0.210	0.185	0.163	0.143	0.125	0.109	0.095	0.082	0.072	0.062
3	0.473	0.433	0.395	0.359	0.326	0.294	0.265	0.238	0.213	0.191	0.170	0.151
4	0.668	0.629	0.590	0.551	0.513	0.476	0.440	0.406	0.373	0.342	0.313	0.285
5	0.816	0.785	0.753	0.720	0.686	0.651	0.616	0.581	0.546	0.512	0.478	0.446
6	0.909	0.889	0.867	0.844	0.818	0.791	0.762	0.732	0.702	0.670	0.638	0.606
7	0.960	0.949	0.936	0.921	0.905	0.887	0.867	0.845	0.822	0.797	0.771	0.744
8	0.984	0.979	0.972	0.964	0.955	0.944	0.932	0.918	0.903	0.886	0.867	0.847
9	0.994	0.992	0.989	0.985	0.980	0.975	0.968	0.960	0.951	0.941	0.929	0.916
10	0.998	0.997	0.996	0.994	0.992	0.990	0.986	0.982	0.977	0.972	0.965	0.957
11	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.990	0.988	0.984	0.980
12	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.991
13				1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996
14							1.000	1.000	0.999	0.999	0.999	0.999

Tabella D.6: Probabilità cumulativa di Poisson  $P(X \leq k) = \sum_{x=0}^k e^{-\lambda} \frac{\lambda^x}{x!}$ .

$\nu$	$\alpha$	0.10	0.05	0.02	0.01
1		0.9877	0.9969	0.9995	0.9999
2		0.9000	0.9500	0.9800	0.9900
3		0.8054	0.8783	0.9343	0.9587
4		0.7293	0.8114	0.8822	0.9172
5		0.6694	0.7545	0.8329	0.8745
6		0.6215	0.7067	0.7887	0.8343
7		0.5822	0.6664	0.7498	0.7977
8		0.5494	0.6319	0.7155	0.7646
9		0.5214	0.6021	0.6851	0.7348
10		0.4973	0.5760	0.6581	0.7079
11		0.4762	0.5529	0.6339	0.6835
12		0.4575	0.5324	0.6120	0.6614
13		0.4409	0.5139	0.5923	0.6411
14		0.4259	0.4973	0.5742	0.6226
15		0.4124	0.4821	0.5577	0.6055
16		0.4000	0.4683	0.5425	0.5897
17		0.3887	0.4555	0.5285	0.5751
18		0.3783	0.4438	0.5155	0.5614
19		0.3687	0.4329	0.5034	0.5487
20		0.3598	0.4227	0.4921	0.5368
25		0.3233	0.3809	0.4451	0.4869
30		0.2960	0.3494	0.4093	0.4487
35		0.2746	0.3246	0.3810	0.4182
40		0.2573	0.3044	0.3578	0.3932
45		0.2428	0.2875	0.3384	0.3721
50		0.2306	0.2732	0.3218	0.3541
60		0.2108	0.2500	0.2948	0.3248
70		0.1954	0.2319	0.2737	0.3017
80		0.1829	0.2172	0.2565	0.2830
90		0.1726	0.2050	0.2422	0.2673
100		0.1638	0.1946	0.2301	0.2540

Tabella D.7: Coefficiente di correlazione. La tavola fornisce i valori di  $r_\alpha$  tali che  $P(-r_\alpha < r < r_\alpha) = 1 - \alpha$ . Esempio di utilizzo: se  $n$  è la lunghezza del campionamento  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , il grado di libertà  $\nu$  è dato da  $\nu = n - 2$ . Ad esempio, per  $n = 12$  si ha  $\nu = 10$  e al livello 0.01 si ottiene  $r_{0.01} = 0.7079$ . Per avere una indicazione di buona correlazione, il valore  $r$  calcolato deve essere nell'intervallo  $[0.7079, 1]$ , o simmetricamente in  $[-1, -0.7079]$ . Se, ad esempio, il valore calcolato  $r$  è 0.860, dal momento che  $0.860 > 0.7079$ , si conclude che la correlazione è “buona” al livello 0.01; ossia, più precisamente, si rifiuta al livello 0.01 l'ipotesi nulla  $H_0$  che non vi sia affatto correlazione.

**Intervallo di confidenza per  $\mu$**

Sia  $X_1, X_2, \dots, X_n$  un campione di lunghezza  $n$  da una popolazione distribuita normalmente con media  $\mu$  e varianza  $\sigma^2$ .

Se  $\sigma$  è noto

$$\left[ \bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right]$$

è un intervallo di confidenza minimo al livello  $1 - \alpha$  per  $\mu$ . Se  $\sigma$  è incognito, tale intervallo è dato da

$$\left[ \bar{X} - \frac{S}{\sqrt{n}} t_{n-1, \alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1, \alpha/2} \right]$$

ove  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$ .

**Intervallo di confidenza per  $\sigma^2$  a code uguali**

Se  $\mu$  è incognito

$$\left[ \frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2} \right]$$

è un intervallo di confidenza minimo a code uguali al livello  $1 - \alpha$  per  $\sigma^2$ . Se  $\mu$  è noto, tale intervallo è dato da

$$\left[ \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, \alpha/2}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, 1-\alpha/2}^2} \right]$$

**Intervallo di confidenza per  $\mu_1 - \mu_2$**

Siano  $X_1, X_2, \dots, X_m$  e  $Y_1, Y_2, \dots, Y_n$  campioni indipendenti estratti da due popolazioni con distribuzione normale con medie  $\mu_1, \mu_2$  e varianze  $\sigma_1^2, \sigma_2^2$ .

Se  $\sigma_1, \sigma_2$  sono note, allora

$$\left[ \bar{X} - \bar{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} z_{\alpha/2}, \bar{X} - \bar{Y} + \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} z_{\alpha/2} \right]$$

è un intervallo di confidenza minimo al livello  $1 - \alpha$  per  $\mu_1 - \mu_2$ . Se  $\sigma_1 = \sigma_2$  (incognito), tale intervallo è dato da

$$\left[ \bar{X} - \bar{Y} - S \sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2, \alpha/2}, \bar{X} - \bar{Y} + S \sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2, \alpha/2} \right]$$

**Intervallo di confidenza per  $\sigma_2^2 / \sigma_1^2$**

Siano  $X_1, X_2, \dots, X_m$  e  $Y_1, Y_2, \dots, Y_n$  campioni indipendenti estratti da due popolazioni con distribuzione normale. Allora

$$\left[ F_{m-1, n-1, 1-\alpha/2} \frac{S_2^2}{S_1^2}, F_{m-1, n-1, \alpha/2} \frac{S_2^2}{S_1^2} \right]$$

ove  $S_1^2$  e  $S_2^2$  sono le varianze campionarie corrispondenti ai due campioni, è un intervallo di confidenza minimo al livello  $1 - \alpha$  per  $\sigma_2^2 / \sigma_1^2$ .

**Test di ipotesi relativi alla media  $\mu$**

$H_0$	$H_1$	si rifiuta $H_0$ al livello $\alpha$ se
$\mu \leq \mu_0$	$\mu > \mu_0$	$t_0 > t_{n-1,\alpha}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$t_0 < -t_{n-1,\alpha}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$ t_0  > t_{n-1,\alpha/2}$

Si calcola  $t_0 = \sqrt{n}(\bar{x} - \mu_0)/s$ , ove  $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ ,  $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$ .  
Se  $\sigma$  è noto, si calcola  $z_0 = \sqrt{n}(\bar{x} - \mu_0)/\sigma$  e si usa  $z_\alpha$ , con  $P(Z > z_\alpha) = \alpha$ ,  
invece di  $t_{n-1,\alpha}$ .

**Test di ipotesi relativi a  $\mu_1 - \mu_2$  quando  $\sigma_1 = \sigma_2$  (incognito)**

$H_0$	$H_1$	si rifiuta $H_0$ al livello $\alpha$ se
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$t_0 > t_{m+n-2,\alpha}$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$t_0 < -t_{m+n-2,\alpha}$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$ t_0  > t_{m+n-2,\alpha/2}$

Si calcola

$$t_0 = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{m} + \frac{1}{n}}}, \quad s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$$

Se  $\sigma_1, \sigma_2$  sono note, si calcola  $z_0 = (\bar{x} - \bar{y}) / (\sigma_1^2/m + \sigma_2^2/n)^{1/2}$  e si utilizza  $z_\alpha$ ,  
anziché  $t_{m+n-2,\alpha}$ .

**Test di ipotesi relativi a  $\sigma$**

$H_0$	$H_1$	si rifiuta $H_0$ al livello $\alpha$ se
$\sigma \leq \sigma_0$	$\sigma > \sigma_0$	$v_0 > \chi_{n-1,\alpha}^2$
$\sigma \geq \sigma_0$	$\sigma < \sigma_0$	$v_0 < \chi_{n-1,\alpha}^2$
$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$v_0 > \chi_{n-1,\alpha/2}^2$
		o
		$v_0 < \chi_{n-1,1-\alpha/2}^2$

Si calcola  $v_0 = (n-1)s^2/\sigma_0^2$ . Se  $\mu$  è nota, si calcola  $v'_0 = \sum_{i=1}^n (x_i - \mu)^2 / \sigma_0^2$  e  
si usa  $\chi_{n,\alpha}^2$ , anziché  $\chi_{n-1,\alpha}^2$ .

**Test di ipotesi relativi all'uguaglianza di varianze  $\sigma_1 = \sigma_2$**

$H_0$	$H_1$	si rifiuta $H_0$ al livello $\alpha$ se
$\sigma_1 \leq \sigma_2$	$\sigma_1 > \sigma_2$	$f_0 > F_{m-1,n-1,\alpha}$
$\sigma_1 \geq \sigma_2$	$\sigma_1 < \sigma_2$	$f_0 < F_{m-1,n-1,\alpha}$
$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$f_0 > F_{m-1,n-1,\alpha/2}$
		o
		$f_0 < F_{m-1,n-1,1-\alpha/2}$

Si calcola  $f_0 = s_1^2/s_2^2$ .

# Notazioni

$\mathbb{N}$	insieme dei numeri naturali $1, 2, \dots$
$\mathbb{R}, \mathbb{C}$	insieme dei numeri reali, complessi
$\mathbb{R}^n$	insieme dei vettori $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$
$\mathcal{A} \Rightarrow \mathcal{B}, \mathcal{A} \Leftrightarrow \mathcal{B}$	$\mathcal{A}$ implica $\mathcal{B}$ , $\mathcal{A}$ se e solo se $\mathcal{B}$
$f(x) := g(x)$	$f(x) = g(x)$ per definizione
$\emptyset$	insieme vuoto
$x \in S$	$x$ è un elemento dell'insieme $S$
$x \notin S$	$x$ non è un elemento dell'insieme $S$
$\{x \mid \dots\}$	insieme di tutti gli $x$ con la proprietà $\dots$
$S \subseteq T$	l'insieme $S$ è contenuto nell'insieme $T$
$S \subset T$	l'insieme $S$ è contenuto propriamente nell'insieme $T$
$\cap, \cup, -$	intersezione, unione, differenza
$\text{card}(S),  S , \#(S)$	cardinalità di $S$
$X \times Y$	insieme prodotto, $X \times Y = \{(x, y) \mid x \in X, y \in Y\}$
$f: S \subseteq X \rightarrow Y$	trasformazione dall'insieme $S$ nell'insieme $Y$
$\Big _{x=x_0}$	simbolo di valutazione; ad es. $\frac{d}{dx}f(x) _{x=x_0} = f'(x_0)$
$F(x) _a^b$	$F(b) - F(a)$
$\mathbf{A}, \mathbf{X}$	matrici
$\mathbf{a}, \mathbf{x}$	vettori
$\mathbf{A}^{-1}$	inversa
$\mathbf{A}^+$	pseudoinversa
$\mathcal{R}(\mathbf{A})$	spazio immagine della matrice $\mathbf{A}$
$\mathcal{N}(\mathbf{A})$	spazio nullo della matrice $\mathbf{A}$
$\text{rank}(\mathbf{A})$	rango della matrice $\mathbf{A}$
$\dim(S)$	dimensione dello spazio lineare $S$
$S^\perp$	spazio ortogonale a $S$
$\rho(\mathbf{A})$	raggio spettrale della matrice $\mathbf{A}$
$\mu_\alpha(\mathbf{A})$	numero di condizionamento della matrice $\mathbf{A}$ nella norma $\ \cdot\ _\alpha$
$\det(\mathbf{A})$	determinante della matrice $\mathbf{A}$
$(\cdot, \cdot)$	prodotto scalare
$C^r([a, b]), r \text{ intero} \geq 0$	spazio delle funzioni a valori reali e continue in $[a, b]$ con le derivate fino all'ordine $r$ .
$\nabla, \bar{\nabla}, \delta$	differenze in avanti, all'indietro, centrali
$\text{sign}(a)$	segno del numero reale $a$
$\Re(z), \Im(z)$	parte reale, immaginaria, del numero complesso $z$
$\nabla f(x), \text{grad } f(x)$	vettore gradiente di $f(\mathbf{x})$
$\text{div } \mathbf{u}$	divergenza del vettore $\mathbf{u}$
$\mathbf{x}^T, \mathbf{x}^*$	trasposto, trasposto coniugato, del vettore $\mathbf{x}$
$f(x) = o(g(x)), O(g(x))$	simboli di Landau, pag. 2
eps	precisione macchina
flops	numero di operazioni floating point
$P(n, r), C(n, r)$	permutazioni, combinazioni di $n$ oggetti a $r$
$\mu, E(X)$	valore medio, aspettato, della variabile aleatoria $X$
$\sigma^2, \text{var}(X)$	varianza della variabile aleatoria $X$

# Bibliografia

- [1] P. Abramowitz, I. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1970.
- [2] R. C. Aiken, editor. *Stiff Computation*. Oxford University Press, New York, 1985.
- [3] N. G. Alvey et al. *GENSTAT, A General Statistical Program*. The Statistics Department, Rothamsted Experimental Station, 1977.
- [4] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, Inc., New York, 1973.
- [5] D. H. Anderson. *Compartmental Modeling and Tracer Kinetics*. Springer-Verlag, New York, Heidelberg, Berlin, 1983.
- [6] S. F. Arnold. *Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, N. J., 1990.
- [7] R. Ash. *Information Theory*. Interscience Publishers, New York, 1965.
- [8] R. J. Baker, J. A. Nelder. *The GLIM System*. Numerical Algorithms Group, Oxford, 1978.
- [9] R. K. Balshfield, M. S. Aldenderfer, L. C. Morey. Cluster analysis software. In P. R. Krishnaiah, L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 245–266. North-Holland Publishing Company, Amsterdam, 1982.
- [10] Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, Inc., New York, 1974.
- [11] N. C. Barford. *Experimental Measurement: Precision, Error and Truth*. John Wiley & Sons, New York, second edition, 1985.
- [12] V. Barnett. *Comparative Statistical Inference*. John Wiley & Sons, New York, 1982.



- [13] B. A. Barry. *Errors in Practical Measurement in Science, Engineering, and Technology*. John Wiley & Sons, New York, 1978.
- [14] R. Bartels, J. Beatty, B. Barsky. *An Introduction to Splines for use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann, New York, 1981.
- [15] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.
- [16] P. Bergé, Y. Pomeau, C. Vidal, editors. *L'ordre dans le chaos; vers une approche déterministe de la turbulence*. Hermann, Paris, 1984.
- [17] P. R. Bevington. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, Inc., Los Altos CA, 1987.
- [18] G. Birkhoff, Gian-Carlo Rota. *Ordinary Differential Equations*. John Wiley & Sons, New York, fourth edition, 1989.
- [19] Å. Björck. *Least Squares Methods: Handbook of Numerical Analysis*. Elsevier North-Holland, Amsterdam, 1988.
- [20] Å. Björck, R. J. Plemmons, H. Schneider. *Large Scale Matrix Problems*. North-Holland, New York, 1981.
- [21] R.P. Brent. *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, N. J., 1973.
- [22] J. C. Bruch, V. Comincioli, C. S. Gupta. An analysis of three-dimensional unsteady seepage through an earth dam. In J. R. Whiteman, editor, *Finite Elements and Applications VI*, pages 405–412. Academic Press, Inc., 1988.
- [23] A. Bryson, Y. Ho. *Applied Optimal Control*. Blaisdell, New York, 1969.
- [24] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*. John Wiley & Sons, Chichester, 1987.
- [25] G. D. Byrne, A. C. Hindmarsh. Stiff ode solvers: A review of current and coming attractions. *Journal of Computational Physics*, 70:1–62, 1987.
- [26] V. Cappellini. *Elaborazione numerica delle immagini*. Boringhieri, Torino, 1985.
- [27] B. Carnahan, H. A. Luther, J. O. Wilkes. *Applied Numerical Methods*. John Wiley & Sons, New York, 1969.
- [28] E. A. Coddington, N. Levinson. *Theory of Ordinary Differential Equations*. Mc Graw-Hill, Inc., New York, 1966.

- [29] P. Colli, V. Comincioli, G. Naldi, A. Torelli. A mathematical study of the plasticity effects in muscle contraction. *Appl. Math. Optim.*, 2:103–118, 1990.
- [30] P. Colli Franzone, L. Guerri, C. Viganotti. Oblique dipole layer potentials applied to electrocardiology. *J. Math. Biol.*, 17:93–124, 1983.
- [31] V. Comincioli. *Analisi Numerica. Metodi Modelli Applicazioni*. McGraw-Hill Libri Italia, Milano, 1990.
- [32] V. Comincioli. *Analisi Numerica. Complementi e problemi*. McGraw-Hill Libri Italia, Milano, 1991.
- [33] V. Comincioli. *FORTRAN 77: Introduzione e applicazioni numeriche*. McGraw-Hill Libri Italia, Milano, 1991.
- [34] V. Comincioli, A. Faucitano. Simulation of heterogeneous: Free radical polymerization reactions. In *Modelling of Chemical Reaction Systems*, pages 261–267, Berlin Heidelberg New York, 1981. Springer-Verlag.
- [35] V. Comincioli, G. Naldi. Mathematical models in muscle contraction: parallelism in the numerical approach. *Mathematical and Computer Modelling*, 7:661–683, 1990.
- [36] V. Comincioli, L. Nespoli, P. F. Periti, G. Serazzi. A mathematical model of the two signal theory for t–b cells cooperation. In *Systems Theory in Immunology*, number 32 in Lecture Notes in Biomathematics, pages 175–189. Springer-Verlag, 1978.
- [37] V. Comincioli, C. Poggesi, C. Reggiani, A. Torelli. A four-state cross bridge model for muscle contraction. Mathematical study and validation. *J. Math. Biol.*, 20:277–304, 1984.
- [38] V. Comincioli, A. Torelli. A mathematical model of contracting muscle with viscoelastic elements. *SIAM J. Math. Anal.*, 19:593–612, 1988.
- [39] G. L. Corona, V. Comincioli, et al. Amitriptyline and nortriptyline plasma levels. Clinical response in depressed women. *Neuropsychobiology*, 16:97–102, 1986.
- [40] R. Courant, D. Hilbert. *Methods of Mathematical Physics*. John Wiley & Sons, New York, 1953.
- [41] J. Cullum, R. A. Willoughby. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Vol. 1: Theory, Vol. 2: Programs*. Birkhäuser, Boston, Mass., 1985.

- [42] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton NJ, 1963.
- [43] I. Daubechies. *Wavelets*. SIAM, Philadelphia, PA, 1992.
- [44] P. J. Davis, P. Rabinowitz. *Methods of Numerical Integration*. Academic Press, Inc., London, 1975.
- [45] C. De Boor. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978.
- [46] J. E. Dennis, R. B. Schnabel. *Numerical Methods for Unconstrained Optimization*. Prentice-Hall, Englewood Cliffs, N. J., 1983.
- [47] DeLos F. DeTar. *Computer Programs for Chemistry*. Benjamin W. A., New York, 1968.
- [48] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.
- [49] T. R. Dickson. *The Computer and Chemistry*. W. H. Freeman and Company, San Francisco, 1968.
- [50] W. J Dixon et al. *BMDP Statistical Software*. University of California Press, Berkeley, 1981.
- [51] R. D. Driver. *Ordinary and Delay Differential Equations*. Springer-Verlag, New York Berlin Heidelberg, 1977.
- [52] I. S. Duff, A. M. Erisman, J. K. Reid. *Direct Methods for Sparse Matrices*. Oxford University Press, Oxford U. K., 1987.
- [53] M. Eisen. *Mathematical Methods & Models in the Biological Sciences*. Prentice-Hall, Englewood Cliffs, N. J., 1988.
- [54] H. Engels. *Numerical Quadrature and Cubature*. Academic Press, Inc., New York, 1980.
- [55] A. Erdélyi, H. Bateman, editors. *Tables of Integral Transforms*. McGraw-Hill, Inc., New York, 1954.
- [56] P. Érdi, J. Tóth. *Mathematical Models of Chemical Reactions. Theory and Applications of Deterministic and Stochastic Models*. Manchester University Press, Oxford Road, Manchester, UK, 1989.
- [57] G. Farin. *Curves and Surfaces Geometric Design. A Practical Guide*. Academic Press, Inc., New York, 1988.

- [58] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, New York, 1957.
- [59] G. E. Forsythe, W. R. Wasow. *Finite Difference Methods for Partial Differential Equations*. John Wiley & Sons, New York, 1960.
- [60] Y. C. Fung. *A First Course in Continuum Mechanics*. Prentice-Hall, Englewood Cliffs, N. J., 1977.
- [61] Y. C. Fung. *Biomechanics; Mechanical Properties of Living Tissues*. Springer-Verlag, New York Heidelberg Berlin, 1981.
- [62] B. S. Garbow, J. M. Boyle, J. J. Dongarra, C. B. Moler. *Matrix Eigensystem Routines. EISPACK Guide Extension*. Springer-Verlag, New York, 1977.
- [63] C. W. Gear. The automatic integration of ordinary differential equations. *Comm. ACM*, 14:176–179, 1971.
- [64] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Englewood Cliffs, N. J., 1971.
- [65] A. George, J. W. H. Liu. *Computer solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs, N. J., 1981.
- [66] C. F. Gerald, P. O. Wheatley. *Applied Numerical Analysis*. Addison-Wesley, Reading, MA, 1989.
- [67] P. E. Gill, W. Murray. *Numerical Methods for Constrained Optimization*. Academic Press, Inc., Cambridge MA, 1974.
- [68] P. E. Gill, W. Murray, M. H. Wright. *Practical Optimization*. Academic Press, Inc., Cambridge MA, 1981.
- [69] G. H. Golub, C. F. Van Loan. *Matrix Computations*. The John Hopkins Press, Baltimore, second edition, 1989.
- [70] G. H. Gonnet, R. Baeza-Yates. *Handbook of Algorithms and Data Structures. In Pascal and C*. Addison-Wesley, Reading, MA, second edition, 1991.
- [71] P. Griffiths, I. D. Hill, editors. *Applied Statistical Algorithms*. Ellis Horwood Limited, Chichester, 1985.
- [72] W. W. Hager. *Applied Numerical Linear Algebra*. Prentice-Hall International Editions, Englewood Cliffs, N. J., 1988.
- [73] E. Hairer, Wanner G. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, 1991.

- [74] E. Hairer, S. P. Norsett, G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer-Verlag, Berlin, 1987.
- [75] E. L. Hall. *Computer Image Processing and Recognition*. Academic Press, Inc., New York, 1979.
- [76] J. M. Hammersley, D. C. Handscomb. *Monte Carlo Methods*. John Wiley & Sons, New York, 1964.
- [77] R. W. Hamming. *Numerical Methods for Scientists and Engineers*. McGraw-Hill, Inc., New York, 1973.
- [78] R. W. Hamming, editor. *Coding and Information Theory*. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1980.
- [79] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.
- [80] G. T. Herman. *Image Reconstruction from Projections. The Fundamentals of Computerized Tomography*. Academic Press, Inc., New York, 1980.
- [81] M. R. Hestenes. *Calculus of Variations and Optimal Control Theory*. John Wiley & Sons, New York, 1966.
- [82] M. R. Hestenes. *Conjugate Direction Methods in Optimization*. Springer-Verlag, New York, 1980.
- [83] D. R. Hill. *Experiments in Computational Matrix Algebra*. Random House, New York, 1988.
- [84] A. C. Hindmarsh. Odepack a systematized collection of ode solvers. In R. S. Stepleman, editor, *Scientific Computing*, page 55. North-Holland Publishing Company, Amsterdam, 1983.
- [85] H. Hochstadt. *Integral Equations*. John Wiley & Sons, New York, 1973.
- [86] P. G. Hoel, S. C. Port, C. J. Stone. *Introduction to Probability Theory*. Houghton Mifflin Company, Boston, 1971.
- [87] P. G. Hoel, S. C. Port, C. J. Stone. *Introduction to Statistical Theory*. Houghton Mifflin Company, Boston, 1971.
- [88] T. C. Hu. *Combinatorial Algorithms*. Addison-Wesley, Reading, MA, 1982.
- [89] T. J. R. Hughes. *The Finite Element Method*. Prentice-Hall, N. J., 1987.
- [90] L. B. Jackson. *Signals, Systems, and Transforms*. Addison-Wesley, Reading, MA, 1991.

- [91] S. L. S. Jacoby, J. S. Kowalik, J. T. Pizzo. *Iterative Methods for Nonlinear Optimization Problems*. Prentice-Hall, N. J., 1972.
- [92] J. A. Jacquez. *Compartmental Analysis in Biology and Medicine*. The University of Michigan Press, Ann Arbor, 1985.
- [93] A. K. Jain, R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, N.J., 1988.
- [94] A. J. Jerri. *Introduction to Integral Equations with Applications*. Marcel Dekker, Inc., New York, 1985.
- [95] D. Kahaner, C. B. Moler, S. Nash. *Numerical Methods and Software*. Prentice-Hall, Englewood Cliffs, N. J., 1988.
- [96] J. Kemeny, J. L. Snell. *Finite Markov Chains*. Van Nostrand, Princeton, N. J., 1960.
- [97] W. J. Kennedy, J. E. Gentle. *Statistical Computing*. Marcel Dekker, New York, 1988.
- [98] D. E. Knuth. *The Art of Computer Programming Vol. 1: Fundamental Algorithms*. Addison-Wesley, London, 1973.
- [99] D. E. Knuth. *The Art of Computer Programming Vol. 3: Searching and Sorting*. Addison-Wesley, London, 1973.
- [100] D. E. Knuth. *The Art of Computer Programming Vol. 2: Seminumerical Algorithms*. Addison-Wesley, London, 1981.
- [101] H. Koçak. *Differential and Difference Equations through Computer Experiments*. Springer-Verlag, New York, 1989.
- [102] L. I. Kronsjö. *Algorithms. Their Complexity and Efficiency*. John Wiley & Sons, Chichester, 1979.
- [103] V. Lakshmikantham, D. Trigiante. *Theory of Difference Equations: Numerical Methods and Applications*. Academic Press, Inc., New York, 1988.
- [104] J. D. Lambert. *Numerical Methods for Ordinary Differential Systems. The Initial Value Problem*. John Wiley & Sons, New York, 1991.
- [105] P. Lancaster, M. Tismenetsky. *The Theory of Matrices. Second Edition with Applications*. Academic Press, Inc., New York, 1985.
- [106] C. L. Lawson, R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N. J., 1974.

- [107] E. B. Lee, L. Markus. *Foundations of Optimal Control Theory*. John Wiley & Sons, New York, 1968.
- [108] L. Ljung, T. Söderström. *Theory and Practice of Recursive Identification*. The MIT Press, Cambridge MA, 1987.
- [109] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 1973.
- [110] J. Macki, A. Strauss. *Introduction to Optimal Control Theory*. Springer-Verlag, New York Berlin Heidelberg, 1982.
- [111] O. L. Mangasarian. *Nonlinear Programming*. McGraw-Hill, Inc., New York, 1969.
- [112] G. I. Marchuk. *Mathematical Models in Immunology*. Springer-Verlag, New York Berlin Heidelberg, 1983.
- [113] W. L. Miranker. *Numerical Methods for Stiff Equations*. Reidel, Boston, 1981.
- [114] A. R. Mitchell. *Computational Methods in Partial Differential Equations*. John Wiley & Sons, New York, 1969.
- [115] A. R. Mitchell, D. F. Griffiths. *The Finite Difference Method in Partial Differential Equations*. John Wiley & Sons, Chichester, 1980.
- [116] S. Miyamoto. *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Kluwer Academic Publishers, London, 1990.
- [117] A. M. Mood, F. A. Graybill, D.C. Boes. *Introduzione alla statistica*. McGraw-Hill Libri Italia, Milano, 1991.
- [118] J. J. Moré, B. S. Garbow, K. E. Hillstom. User guide for minpack-1. Technical Report 80-74, Argonne National Laboratory, 1980.
- [119] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag, New York, 1984.
- [120] R. G. Mortimer. *Mathematics for Physical Chemistry*. Macmillan Publishing Co., Inc., New York, 1981.
- [121] N. H. Nie, editor. *SPSS-X User Guide*. McGraw-Hill, Inc., New York, 1983.
- [122] A. V. Oppenheim, R. W. Schaffer. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, N. J., 1975.
- [123] J. M. Ortega. *Matrix Theory*. Plenum Press, New York, 1988.

- [124] J. M. Ortega, W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, Inc., New York, 1970.
- [125] M. Paolini, C. Verdi. An automatic triangular mesh generator for planar domains. *Rivista di Informatica*, 1990.
- [126] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., New York, third edition, 1991.
- [127] T. S. Parker, L. O. Chua. *Practical Numerical Algorithms for Chaotic Systems*. Springer-Verlag, New York, 1989.
- [128] T. W. Parks, C. S. Burrus. *Digital Filter Design*. John Wiley & Sons, New York, 1987.
- [129] E. Parzen. *Stochastic Processes*. Holden-Day, San Francisco, 1962.
- [130] L. S. Pontryagin, V. G. Boltyanski, R. V. Gambrelidze, E. F. Mischenko. *The Mathematical Theory of Optimal Processes*. Interscience Pub., New York, 1962.
- [131] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling. *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1986.
- [132] L. R. Rabiner, B. Gold. *Theory and Application of Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, N. J., 1975.
- [133] A. A. Ray, editor. *SAS User's Guide*. SAS Inst. Inc., Cary N. Carolina, 1982.
- [134] D. Revuz. *Markov Chains*. North-Holland, New York, 1975.
- [135] J. R. Rice. *Numerical Methods, Software, and Analysis*. McGraw-Hill, Inc., Aukland, 1985.
- [136] B. N. Robinson et al. *SIR Scientific Information Retrivial Users Manual*. SIR Inc., Evanston, Illinois, 1980.
- [137] V. K. Rohatgi. *Statistical Inference*. John Wiley & Sons, New York, 1984.
- [138] S. M. Ross. *Introduction to Probability Models*. Academic Press, Inc., New York, fourth edition, 1980.
- [139] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. John Wiley & Sons, New York, 1981.
- [140] T. A. Ryan, B. L. Joiner, B. F. Ryan. *MINITAB Student Handbook*. Duxbury Press, North Scituate, MA, 1976.



- [141] L. Sachs. *Applied Statistics. A Handbook of Techniques*. Springer-Verlag, New York, 1982.
- [142] L. L. Scharf. *Statistical Signal Processing. Detection, Estimation, and Time Series Analysis*. Addison-Wesley, Reading, MA, 1991.
- [143] R. Sedgewick. *Algorithms*. Addison-Wesley, Reading, MA, second edition, 1988.
- [144] S. M. Selby, R. C. Weast, editors. *Handbook of Tables for Mathematics*. CRC Press, Cleveland, Ohio, 1975.
- [145] J. Smoller. *Shock Waves and Reaction Diffusion Equations*. Springer-Verlag, New York, 1983.
- [146] P. Sterbenz. *Floating-Point Computation*. John Wiley & Sons, New York, 1977.
- [147] G. Strang. *Linear Algebra and Its Applications*. Academic Press, Inc., New York, 1980.
- [148] G. Strang. *Introduction to Applied Mathematics*. Wellesley Cambridge Press, 1986.
- [149] A. D. Stroud, D. Secrest. *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, N. J., 1966.
- [150] G. W. Swan. *Applications of Optimal Control Theory in Biomedicine*. Marcel Dekker, Inc., New York, 1984.
- [151] J. H. van Lint, editor. *Introduction to Coding Theory*. Springer-Verlag, New York, 1982.
- [152] R. S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, N. J., 1962.
- [153] G. R. Walsh. *Methods of Optimization*. John Wiley & Sons, London, 1975.
- [154] D. V. Widder. *The Laplace Transform*. Princeton University Press, Princeton, N.J., 1941.
- [155] N. Wiener. *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. John Wiley & Sons, New York, 1949.
- [156] M. A. Wolfe. *Numerical Methods for Unconstrained Optimization. An Introduction*. Van Nostrand Reinhold Company, New York, 1978.
- [157] G. Zoutendijk. *Mathematical Programming Methods*. North-Holland, Amsterdam, 1976.

# Indice analitico

- accelerazione, di un metodo iterativo, 234
  - Aitken, metodo di, 236
- additivi, obiettivi, 720
- Adams, metodi di, 368
  - Bashforth, 369
  - Moulton, 369
- adeguatezza, di un modello, 3, 701
- aggiunto, vettore, 744, 760
- algoritmo, definizione, 1
  - instabilità di un, 22
- ammissibili, controlli, 720
- ammissibilità, insieme di, 242
- analisi degli errori
  - all'indietro, backward, 15
  - sistemi lineari, 77
- analisi multirisoluzione, wavelet, 671
- analisi spettrale, 657
- angolo, tra due vettori, 801
- ANOVA, 539, 542
- antibiotico, cinetica di un, 677
- approssimazione, problema generale di, 169
- aritmetica, in virgola mobile, 15
- Arrhenius, legge di, 339
- arrotondamento, definizione, 13
- autovalori, autovettori, 109, 820
  - condizionamento del calcolo di, 110
  - problema degli, generalizzato, 135
  - problema degli, equazioni differenziali, 409
- Bairstow, metodo di, 217
- Banachiewicz, metodo di, 52
- bang-bang, on-off, relay, controllo, 729
- base, di numerazione, 4
- base, di uno spazio lineare, 805
  - canonica, 806
  - ortonormale, 808
- basic sequential search, 899
- Bauer-Fike, teorema di, 111
- Bayes, teorema di, 474
- BDF, backward differentiation formulas, 368
- Beattie-Bridgeman, legge di, 191
- Bernoulli
  - equazione di, 856
  - metodo di, radici di un polinomio, 115
  - problema di, equazione integrale, 413
- Bernstein, polinomi di, 163
- Bézier, approssimazione di, 162
- binario, sistema, 5
- binary search, 901
- Binet, regola di, 809
- birapporto, rapporto armonico tra 4 numeri, 858
- bisezione, 196, 395
  - metodo della, calcolo degli autovalori, 132
- bit, 5
- blackness, matrix, 837
- Box-Muller, metodo, 619
- brachistocrona, 784
- Brusselator, modello di cinetica chimica, 341
- bubble sort, 903

- Buffon, problema di, 629  
 byte, 12  
  
 calcare, modello di riduzione del, 682  
 calcolo, combinatorio, 464  
 cammino, ottimale, 738  
 campionamento, 514  
 cancellazione, in aritmetica di macchina, 16  
 caos, nei sistemi dinamici, 241  
 caratteristica, di un numero reale, 6  
 caratteristiche, curve, 429, 440  
 Cauchy  
   problema a valori iniziali, o di, 325  
   valore principale di, 295  
 Cayley-Hamilton, teorema di, 822  
 centro, punto di equilibrio, 871  
 Chebichev, disuguaglianza di, 483  
 chemioterapia, controllo in, 722, 776  
 Chevillet, formula di quadratura di, 319  
 Cholesky, metodo di, 52  
   per matrici a banda, 57  
   per matrici tridiagonali, 58  
 ciclo, cellulare, 723  
 cicloide, 784  
 cifre  
   arabiche, decimali, 4  
   esadecimali, 5  
   significative, 3  
 cinetica chimica, 337, 681  
 circolatorio, sistema, 725  
 circuiti, elettrici, 26, 331  
 Clairaut, equazione differenziale di, 862  
 cluster analysis, 579  
   algoritmi di tipo gerarchico, 586  
   algoritmi single-link, complete-link, 592  
   algoritmi di tipo partizione, 595  
   K-means, 597  
 cofattore, 809  
 companion, matrix, Frobenius, 821  
 compartimenti, tecnica dei, 335, 674  
  
 condizionamento, 698  
   connettività di un sistema a, 693  
   mammillare, sistema a compartimenti, 693  
   stabilità di un sistema a, 694  
 compressione, dei dati, 837  
 comprimibilità, fattore di, 204  
 comunicazione, sistema di, 568  
 condizionamento  
   di una equazione non lineare, 196  
   di un problema di identificazione, 712  
   di un problema matematico, 21  
   di un sistema lineare, 67, 835  
   numero di, 69  
   delle radici di un polinomio, 219  
 confidenza, intervallo di, 519, 707, 920  
 coniugate, direzioni, 277  
 connettività, diagramma di, 693  
 conservazione, equazioni di, 434  
 contrazione, costante di, 227  
 controllabilità, di un sistema, 727  
 controllo ottimo, 685, 718, 730  
   open-loop, aperto, 732  
   feedback, closed-loop, chiuso, 732  
 convezione, 401  
 convoluzione, di sequenze, 647  
 coordinate, trasformazione di, 456  
   baricentriche, 317  
   cilindriche, polari, sferiche, 456  
 corda vibrante, modello, 438  
 corde, metodo delle, 212  
 correlazione, tra variabili aleatorie, 492, 544, 717, 919  
 costante d'errore, 198  
 Courant-Fischer, teorema di, 827  
 Courant-Friedrichs-Lewy, condizione di, 433, 444  
 covarianza, matrice di, 284, 497  
 Cowell-Numerov, metodo di, 380  
 Cramer, regola di, 820  
 Crank-Nicolson, metodo di, 359, 451

- criterio  
   additivo, 731  
   di tempo minimo, 726  
   quadratico, 721, 744  
 Crout, metodo di, 50  
 curie, 688  
  
 D'Alembert, equazione della corda, 441  
 Davidon-Fletcher-Powell, metodo di, 274  
 debole, formulazione di un problema  
   ai limiti, 406  
 Debye, funzione di, 318  
 decadimento radioattivo, 333  
 decimazione, filtraggio di segnali, 656  
 decomposizione, fattorizzazione, di una  
   matrice  
   **LDL**<sup>T</sup>, 51  
   **LDM**<sup>T</sup>, 49  
   **LU**, 48  
   **QR**, 79  
   **RR**<sup>T</sup>, 52  
 deflazione, calcolo degli autovalori, 116  
   zeri di polinomi, 217  
 derivate, calcolo numerico delle, 185  
 design, matrice, 841  
 determinante, 808  
 diabete, 724  
 differenze  
   divise, interpolazione, 150  
   metodo alle, problema ai limiti, 398  
   operatore alle, 186  
 diffusione,  
   coefficiente di, 391  
   equazione della, 391, 401, 445, 894  
   metodi numerici, 448  
   condizione di convergenza, 449  
   principio del massimo, 447  
   e convezione, 401  
   tra compartimenti, 680  
 dimensione, di uno spazio lineare, 806  
 Dirac, funzione di, 878  
 Dirichlet  
   funzione di, 290  
   problema di, 457  
 disposizioni, calcolo combinatorio, 463  
 distorsione, bias, 516  
 distribuzione, funzione di, 479  
   binomiale, 498, 917  
   chi-quadrato, 517, 915  
   esponenziale, 505, 616  
   F, 518  
   gamma, 507  
   di Gauss, normale, 509, 618, 913  
   multivariata, 512  
   di Poisson, 502, 918  
   standardizzata, 510  
   ipergeometrica, 500  
   marginale, 488  
   subordinata, 489  
   uniforme, 501  
   t di Student, 519, 914  
 distribuzioni, funzioni generalizzate, 878  
 Doolittle, metodo di, 50  
 duale, problema, 249  
  
 efficienza luminosa, 318  
 electron spin resonance (ESR), 333  
 elementi finiti, metodo degli, 404  
 elettrocardiografia, 791  
 elettrodinamica, problema dei due cor-  
   pi, 422  
 eliminazione, procedimento di, 31  
   all'indietro, 34  
   in avanti, 33  
 ellittici, integrali, 291, 293  
 endocrino, sistema, 724  
 entropia, 569, 715  
 enzimi, cinetica degli, 386  
 epidemie, teoria matematica, 685  
 equazione differenziale, 320  
   campo di direzioni, 324  
   carattere altamente oscillatorio, 389  
   curva soluzione, 323  
   esatta, 862

- forma normale di una, 321
- omogenea, 859
- ordine di una, 321
- problema ai limiti, 76, 390
- problema a valori iniziali, 324
- soluzione, definizione, 322
- equazioni alle derivate parziali, 426
- equazioni alle differenze, 346, 640
  - soluzione generale, 373
- equazioni normali, minimi quadrati, 23, 841
- equilibrio
  - di forze elastiche, 73, 394
  - di strutture, 28
- Erone, metodo di, 211
- errore, funzione, 446
- errore standard, 516
- errori
  - analisi degli, sistemi lineari, 68
  - assoluti, 3
  - di arrotondamento, 13, 188
  - legge degli, 509
  - propagazione degli, 18
  - random, 4
  - relativi, 3
  - sistematici, 3
  - sorgenti di, 3
  - di troncamento, 4
- estrapolazione, 142
  - formule di quadratura di, 310
- euclideo, spazio, 797
- Eulero, equazione di, 768
- Eulero, metodo di
  - esplicito, 343
    - cambiamento del passo, 349
    - consistenza, 346
    - convergenza, 346
    - influenza degli errori di arrotondamento, 351
    - stabilità, 347
  - implicito, 358
  - modificato, metodo di, 354
- Eulero-Maclaurin, formula di, 311
- farmaco
  - cinetica di un, 675
  - controllo della concentrazione di un, 725
- fattorizzazione, a blocchi, 64
- FFT, Fast Fourier Transform, 650
- Fibonacci, Leonardo Pisano detto, 5
  - successione di, 86, 374, 614
  - metodo di ottimizzazione di, 263
- Fick, legge di, 392, 680
- fill-in, nel metodo di eliminazione, 59
- filtraggio, di segnali, 634, 664
  - AR, FIR, IIR, MA, 640
  - uso di MATLAB, 653
- filtrazione, controllo, 785
- finestra, trasformata di Fourier, 666
- Fisher, trasformata Z di, 545
- fitting, 841
- Fletcher-Reeves, metodo di, 280
- floating-point, virgola mobile, 10
- flop, definizione, 34
- forza impulsiva, 885
- Fourier
  - serie di, 889
  - trasformata di, 644, 648
- Fubini, teorema di integrazione di, 294
- fuoco, punto di equilibrio, 871
- fuzzy clustering, 604
- Galerkin, metodo di, 404
- gas, equazione di stato di un, 191
- Gauss
  - metodo di eliminazione di, 34, 36
    - numero delle operazioni, 38
  - formule di quadratura di, 302
  - Jordan, metodo di, 67
  - Kronrod, formule di quadratura di, 306
  - Markov, 535
  - Newton, 286

- Seidel, metodo iterativo, sistemi lineari, 89
- Gershgorin-Hadamard, teorema di, 828
- Gibbs
  - disuguaglianza di, 573
  - fenomeno di, 890
- Givens
  - metodo di, autovalori, 129
  - metodo di, sistemi lineari, 84
- gradiente coniugato
  - metodo del, sistemi lineari, 99
- gradiente, vettore, 246
- grafi, teoria dei, 587, 847
- Gram-Schmidt, 808
  - metodo numerico, sistemi lineari, 81
- griglia, grid search, metodo di ottimizzazione, 268
- guida, problema feedback di, 733
- Hadamard
  - esempio di, 452
  - matrice di, 812
- Hamilton-Jacobi-Bellman
  - equazione di, caso continuo, 743
  - equazione di, caso discreto, 736
- hamiltoniana, 743
- Hamming, distanza di, 584
- Hardy-Weinberg, legge di, 475
- heapsort, 910
- Heaviside, funzione di, 876
- Helmholtz, equazione di, 453
- Hessenberg, matrice nella forma di, 130, 798
- Hestenes-Stiefel, metodo di, 101
- Heun, metodo di, 354
- Hilbert, matrice di, 74
- Hooke
  - legge di, 394
  - materiale di tipo, 74
- Hopf, biforcazione di, 341
- Horner-Ruffini, 7, 149, 216
- Householder,
  - metodo di, autovalori, 127
  - metodo di, sistemi lineari, 83
- Huffman, algoritmo di, 576
- identificazione, di un modello, 648, 700
  - dei parametri in un modello, 701, 704
- immunologia, strategie ottimali in, 779
- impulse response function, 763
- indipendenza stocastica, 472
- infezione, diffusione di una, modello, 425
- inferenza, statistica, 528
  - bayesiana, 539
- informazione, teoria dell', 568
- inquinamento, modello, 426
- insetti, come ottimizzatori, 774
- instabilità, amplificazione degli errori, 17, 22
- integrale
  - improprio, 294
  - indefinito, 291
  - metodo Monte Carlo
  - multiplo, 293, 624
    - formule di cubatura, 316
  - teorema fondamentale del calcolo integrale, 291
- integrali, equazioni, 411
  - Fredholm, 412
  - Volterra, 412
- integro-differenziale, problema, 413
- interpolation search, 902
- interpolazione, 141
  - convergenza del polinomio di, 147
  - errore di troncamento nella, 145
  - inversa, metodo della, 215
  - polinomio di, di Newton, 149
- inversa, di una matrice, 30, 810
- inversi, problemi, 791
- ipotesi, verifica statistica di, 536, 921

- isoparametrico, problema di controllo, 777
- iterazione inversa, calcolo degli autovalori, 115
- Jacobi  
 matrice di, 811  
 metodo iterativo di, sistemi lineari, 86  
 metodo iterativo di, autovalori, 121
- jacobiana, matrice, 205
- Jaccard, coefficiente di, 584
- Jennings profile, metodo, 62
- Jordan, forma canonica di, 825
- Kalman, filtro ottimale di, 663
- Karhunen-Loeve, trasformata di, 585
- Kirchhoff, legge di, 26, 331
- Kraft, disuguaglianza di, 573
- Kruskal, algoritmo di, 590
- Krylov, spazio di, 118
- Laguerre, metodo di, 219
- Lanczos, metodo di, calcolo degli autovalori, 116
- Landau, simboli, 2
- Laplace  
 equazione di, 453  
 metodo Monte Carlo, 633  
 regola di, 809  
 trasformata di, 875
- lavori virtuali, principio dei, 406
- Lax-Wendroff, metodo di, 434
- leap-frog, metodo, 368, 434
- Leontief, modello economico di, 29
- Levenberg-Marquardt, metodo di, 286, 716
- limite centrale, teorema, 525
- linear insertion sort, 905
- Lipschitz, condizione di, 326
- LMM, linear multistep methods, 367  
 condizione delle radici, 375  
 convergenza, 371, 375
- costante di errore, 372  
 ordine, 371  
 stabilità per passo fissato, 376
- livello, curve di, 245
- Lobatto, formule di quadratura di, 305
- localizzazione, di una radice, 196
- logistica, equazione della  
 con ritardo, 423  
 discreta, 240
- Lorenz, modello di, turbolenza atmosferica, 341
- lunghezza euclidea, di un vettore, 801
- Mahalanobis, distanza di, 584
- mantissa, di un numero reale, 6
- manutenzione, problemi di, 413
- marginal value, 763
- Markov, catene di, 549  
 cammini aleatori, 554  
 distribuzioni limite, 562  
 equazioni di Chapman-Kolmogorov, 557  
 ergodiche, 562  
 funzione di probabilità di transizione, 552  
 infinite, 559  
 martingala, 556  
 matrice stocastica, 553, 846  
 modello di diffusione di Ehrenfets, 554  
 modello di rovina del giocatore, 555  
 modello meteorologico, 550  
 sistemi a compartimenti, 699  
 spazio degli stati, 552  
 tempi di assorbimento di, 560
- massa e azione, equazione di, 339
- matrice, definizione, 796  
 a banda, 53  
 memorizzazione di una, 55  
 metodo di eliminazione per, 56  
 aggiunta, 811  
 convergente, 833

- definita positiva, 99, 802
- diagonalizzabile, 824
- diagonalmente dominante, 97, 850
- elementare, 813
  - di Gauss, 49, 814
  - di Givens, 817
  - di Householder, 815
- hermitiana, 800
- identità, 798
- inversa di una, calcolo, 30
- irriducibile, 27, 847
- M-, 27, 849
- non negative, 846
- normale, 826
- ortogonale, 802
- partizionata, 804
- permutazione, 798
- a predominanza diagonale, 27, 97
- scalare, 798
- simmetrica, 51, 800
- sparsa, 58
- stiffness, 29
- di Stieltjes, 849
- trasposta, 800
- triangolare, 798
- unitaria, 803
- Z-, di Metzler, 690
- matrice di guadagno, informazione, 662
- matrice, di prossimità, 580
- media, teorema della, 290
- mediana, di una variabile aleatoria, 485
- menu-planning, problema di programmazione lineare, 251
- metodi
  - compatti, sistemi lineari, 48
  - diretti, sistemi lineari, 30, 31
    - a blocchi, 92
  - iterativi, sistemi lineari, 30, 86
- minimi quadrati, metodo, 82, 173, 494, 534, 659, 840
  - con peso, 659
  - decomposizione **QR**, 81
    - non lineari, 283
    - ricorsivo, 661
- Minkowski, metrica di, 583
  - disuguaglianza di, 830
- minori, di una matrice, 809
- modello, adeguatezza di un, 3
- modello deterministico, modello stocastico, 529
- mole, 338
- molteplicità algebrica, di un autovalore, 821
- moltiplicatori, nel metodo di eliminazione, 31
- moltiplicatori, di Lagrange, 752
- momenti, di una variabile aleatoria, 482
  - metodo dei, 533
- Monte Carlo, metodo, 609
  - efficienza del, 623
  - hit or miss, 620
  - sample-mean, 622
  - tecniche di riduzione della varianza, 625
    - campionamento secondo l'importanza, 626
    - campionamento stratificato, 626
    - trasformazioni antitetiche, 627
- Muller, metodo di, 214
- Neville, algoritmo di, interpolazione, 152
- Newton, metodo, 201
  - convergenza, 207
  - metodo di ottimizzazione, 259, 270
  - radici multiple, 210, 233
  - per equazioni differenziali, 358, 362
- Newton-Cotes, formule di quadratura di, 296
- nodo, punto di equilibrio, 871
- norma
  - di matrice, 831
  - consistente, 831
  - naturale, 831



- di Frobenius, o di Schur, 832
  - spettrale, 832
- di vettore, 829
- in uno spazio lineare, 170
  - del massimo (di Chebichev), 170
  - euclidea (norma 2), 170
- notazione, posizionale, 4
- nuclear magnetic resonance (NMR), 333
- numeri macchina, floating point, 10
- numeri, legge dei grandi, 483
- numeri casuali, pseudocasuali, 611
- Nyquist, frequenza di, 645
  
- Ohm, legge di, 26
- omeostasi, 733
- omogenea, equazione differenziale, 8529
- onde, propagazione delle, 427
  - equazione delle, 438, 896
  - fronti di, 429
  - sinusoidali, 428
  - velocità di fase, 429
- operazioni, in virgola mobile, 15
- optimal blending, problema di programmazione lineare, 251
- optimal sequential search, 900
- ordine di convergenza, 208
- ordine, di una reazione chimica, 339
- oscillatore armonico, 329
- ottimizzazione, metodi di, 257
- overflow, 13
  
- Parseval, identità, 890
- parti, integrazione per, 292
- Pascal, triangolo di, 464
- pencil, autovalori, 136
- permutazioni, calcolo combinatorio, 463
- Perron, teorema di, 847
- pianificazione, degli esperimenti, 714
- Picard, metodo di, iterazioni successive, 227, 362
- piombo, concentrazione nel corpo umano, 335
  
- pivot, strategia del, 40
  - parziale, totale, 42
  - scalato, 44
    - algoritmo, 45
- Planck, legge di, 224
- Poiseuille, legge di, 193
- Poisson
  - equazione di, 453
  - modello probabilistico di, 503
- polinomi
  - ortogonali, 177
    - di Chebichev, 180, 304, 410
    - di Legendre, 179, 303, 410
- polinomio caratteristico, 821
- Pontryagin, principio del minimo di, 748
- popolazione, accrescimento di una, 411, 885
- potassio, cinetica nei globuli rossi, 690
- potenze, metodo delle, autovalori, 111
- potenziale, gravitazionale, 414
- Powell, metodo di, 286
- precisione
  - macchina eps, definizione, 14
  - semplice, doppia, 10
- precondizionamento, di una matrice, 104
- predatore-preda, modello, 413, 871
- predictor-corrector, metodo, 370
- preprocessing, operazione di, sistemi lineari, 71
- probabilità,
  - condizionata, 471
  - densità di (PDF), 480
  - funzione di ripartizione (CDF), 479
  - teoria assiomatica, 462
- processo, additivo, 735
- prodotto, in spazi vettoriali
  - esterno, 801
    - scalare, dot, interno, 171, 800
- proiezione, ortogonale, 806
- programmazione dinamica, 742

- programmazione lineare, 242  
proton magnetic resonance (H-NMR), 333  
pseudoinversa, 842  
punti, stazionari, 750  
punto, estremo, di un insieme convesso, 245  
punto fisso  
  definizione, 225  
  calcolo numerico, 227  
quadratura, formule di, 296  
  composte, 299  
  adattive, 307  
quantile, di una variabile aleatoria, 486  
questionari, costruzione di, 577  
**QR**, metodo, calcolo degli autovalori, 133  
quicksort, 906  
radici, zeri, di polinomi, 215  
  condizione delle, equazioni differenziali, 375  
raggio spettrale, 821  
rango, rank, di una matrice, 811  
rappresentazione normalizzata, 6  
Rayleigh, quoziente di, 112  
reazioni elementari, cinetica chimica, 338, 681  
reattore chimico, a flusso costante, 683  
regolarizzazione, di un sistema lineare, 845  
regressione, curva di, 494  
regula falsi, metodo, 200  
RKF45, metodo adattivo, 365  
Remes, algoritmo di 184  
return, function, 735  
Riccati, equazione di, 745, 747, 857  
Riemann, integrale di, 288  
rilassamento, metodo iterativo, sistemi lineari, 90  
  parametro di, 92  
risonanza, fenomeno della, 332  
risorse, impiego ottimale delle, 243, 719, 737  
  allocazione delle, 741  
ritardo, equazione differenziale con, 419  
Romberg, metodo di, 312  
Rosenbrock, metodi di, 389  
rumore bianco, 653  
Runge, esempio di, interpolazione, 147  
Runge-Kutta, 353  
  -Fehlberg, 362  
  impliciti, 356  
  per sistemi di equazioni differenziali, 378  
scalatura, procedimento di, 25  
Schur, forma normale di, 826  
secanti, metodo delle, 213  
selection sort, 904  
self-organizing sequential search, 900  
sella, punto critico, 871  
sensitività  
  equazioni di, 708, 789  
  funzioni di, 789  
sforzo, minimo, 772  
shadow price, 763  
Shellsort, 909  
Sherman-Morrison, formula di, 63, 814  
shock wave, onda d'urto, 436  
shooting, metodo, 395  
  multiplo, 397  
separazione delle variabili, metodo, 852, 894  
sezione aurea, algoritmo della, 262  
similitudine, trasformazioni per, calcolo degli autovalori, 119, 820  
simplexso  
  metodo di programmazione lineare, 253  
  ente geometrico, 268  
Simpson, formula di quadratura di, 297, 300, 417

- adattiva, 308
- simulazione, metodo Monte Carlo, 628  
di traiettorie con collisioni, 630
- sistema ortogonale, 171
- sistemi  
dinamici discreti, 238  
lineari, 25, 819  
inconsistenti, 256  
non lineari, 190  
triangolari, 33
- sistema lineare a tempo discreto, 635  
LTI, 638
- skewness, di una variabile aleatoria, 483
- smearing, perdita di cifre, 16
- Snell, legge di rifrazione di, 784
- somma diretta, spazi vettoriali, 806
- sostituzione, integrazione per, 292
- spanning-tree, 590
- spazi lineari normati, 170
- spazio colonna, immagine, 805, 839
- spazio euclideo, 797
- spazio nullo, nucleo di una matrice, 805, 839
- spettro, di una matrice, 821
- spline, 154  
costruzione di wavelet, 673  
cubiche, 158  
B-, 161  
lineari, 156, 408
- SOR, metodo di rilassamento, 90, 281  
applicazione al problema di Dirichlet, 459
- splitting, di una matrice, 87
- stabilità  
assoluta, 360  
per passo fissato, 376
- statistica inferenziale, 515, 528
- stazionarietà, condizioni necessarie, 749
- stimatori, 530  
BAN, 534  
consistenza, invarianza, 533
- stechiometria, 338
- steepest descent, metodo, 269, 755
- Stefan-Boltzmann, equazione di, 318
- Steffensen, metodo di, 213, 236
- Stein-Rosenberg, risultato di, 98
- stiff, sistema, 381  
biochimica, 386  
definizione, 385  
esempio di, 359, 382  
metodi numerici, 388  
reazioni chimiche, 386
- strain, 394
- stress, 394
- Strassen, algoritmo di, 39
- strutture, analisi di, 28
- Sturm, successione di, 132  
-Liouville, problema di, 410
- SVD, decomposizione in valori singolari, 834  
algoritmo, 137
- sviluppo in serie, metodo numerico, 351
- Sylvester, criterio di, 809
- switching, funzione, 777
- target set, insieme bersaglio, 728
- Tartaglia, formula di, 464
- Tartaglia-Cardano, formule di, 225
- tempi medi di residenza, compartimenti, 698
- terminal payoff, 731
- time horizon, 727
- theta- ( $\theta$ -)metodo, 558
- traccia, di una matrice, 817
- traccianti, cinetica dei, 687
- traffico, modello di, 435
- traiettoria, di un sistema, 719
- trapezio, formula di quadratura del, 297
- trapezi, formula dei, 358
- trasferimento frazionale, fractional transfer coefficient, 681, 685
- trasformazioni  
per congruenza, 822

- per similitudine, 820
- trasmissività termica, identificazione, 791
- trasporto, equazione del, 428
- trasporto ottimale, problema di, 252
- trasversalità, condizioni di, 760
- turnover, compartimenti, 675
  
- underflow, 12
- unimodali, funzioni, 262
- updating, tecnica, 62
  
- valore medio, di una variabile aleatoria, 481
- Vandermonde, matrice di, 76, 811
- variabili aleatorie, 478
  - multivariate, 486
- varianza, di una variabile aleatoria, 482
  - analisi della, 541
- variazioni, calcolo delle, 768
- vasi sanguigni, ramificazione ottimale, 192
- velocità media di convergenza, 98
- velocità di reazione, cinetica chimica, 338
- verosimiglianza, maximum-likelihood, 531
- vertice, definizione, 244
- vettore, definizione, 797
- viriale, equazione, 414
- Von Neumann, metodo acceptance-rejection, 617
  
- wavelet, funzioni, 666
- Weierstrass, teorema di, 142
- Wien, legge di, 224
- Wilcoxon, test statistico, 545
- wronskiano, 866
  
- Young, modulo di, 394
  
- z, trasformata, 647
- Zermelo, problema di, 781, 860
- zero, di una funzione, 190