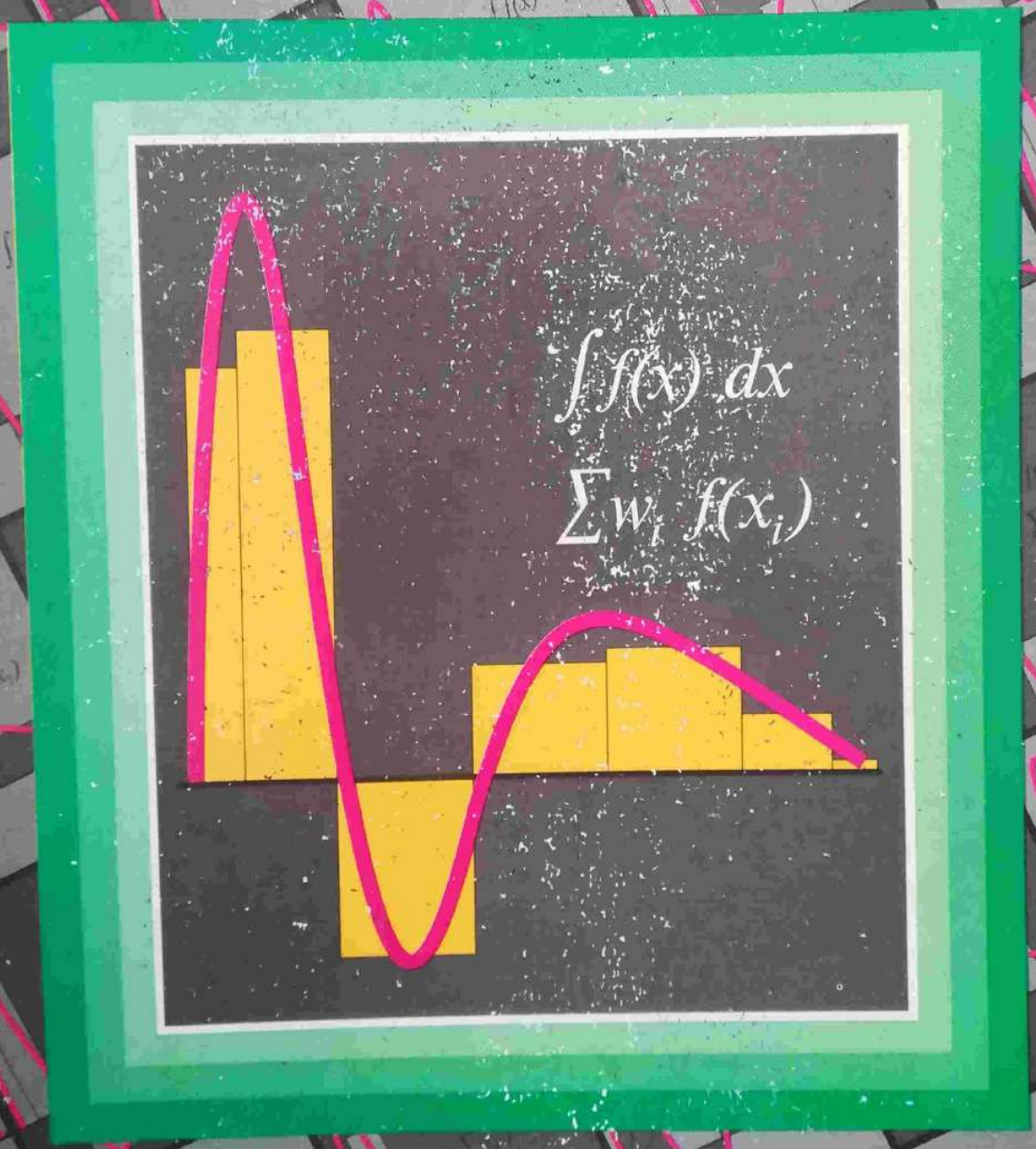
ROBERTO BEVILACQUA
DARIO BINI MILVIO CAPOVANI
ORNELLA MENCHI

METODI NUMERICI



ZANICHELLI

Prima edizione: gennaio 1992

Ristampe:
6 5 4 3 2 1 1992 1993 1994 1995 1996 1997

Il testo è stato composto dagli autori utilizzando un sistema MACINTOSH™ collegato a una stampante LASER WRITER™, entrambi di produzione della Apple Computer Inc.

Copertina: Duilio Leonardi

Il disegno di copertina rappresenta l'approssimazione dell'integrale $\int_a^b f(x)\,dx$ con una formula di quadratura della forma $\sum_{i=1}^{n} w_i f(x_i)$, in cui gli x_i sono i nodi e i w_i sono i pesi.
I rettangoli hanno basi w_i e altezze $f(x_i)$. La somma delle basi è uguale all'ampiezza dell'intervallo di integrazione.

Realizzare un libro è un'operazione complessa, che richiede numerosi controlli: sul testo, sulle immagini e sulle relazioni che si stabiliscono tra essi.
L'esperienza suggerisce che è praticamente impossibile pubblicare un libro privo di errori. Siamo quindi grati ai lettori che vorranno segnalarceli. Per segnalazioni e suggerimenti relativi a questo volume l'indirizzo a cui scrivere è:

Zanichelli Editore S.p.A.
Via Irnerio 34
40126 Bologna
tel. 051/293111 - fax 051/249782

Stampato a Bologna
dalla Tipostampa Bolognese, via Collemarini, 5/A,
per conto della Zanichelli Editore S. p. A.,
via Irnerio 34, 40126 Bologna.

INDICE

Introduzione vii

Capitolo 1 I PROBLEMI DEL CALCOLO

1. Il calcolatore: la generazione degli errori 1
2. Problemi mal condizionati 10
3. Complessità di calcolo 12

Esercizi proposti 21
Commento bibliografico 27
Bibliografia 28

Capitolo 2 ANALISI DELL'ERRORE

1. Rappresentazione in base di un numero 30
2. Conversione di base 34
3. Numeri di macchina 36
4. Errori di rappresentazione 40
5. Operazioni di macchina 44
6. Calcolo del valore di una funzione 50
7. Errore nelle operazioni di macchina 53
8. Uso dei grafi per l'analisi dell'errore 55
9. Errore nelle funzioni non razionali 67
10. Analisi dell'errore all'indietro 71
11. Analisi statistica dell'errore 75
12. Analisi automatica dell'errore 79

Esercizi proposti 83
Commento bibliografico 100
Bibliografia 102

Capitolo 3 EQUAZIONI E SISTEMI NON LINEARI

1. Metodo di bisezione 104
2. Metodi di iterazione funzionale 106
3. Criteri di arresto 112
4. Effetto degli errori di arrotondamento 115
5. Ordine di convergenza 118
6. Metodo delle corde 123
7. Metodo delle tangenti 125
8. Metodo delle secanti 135
9. Efficienza di un metodo iterativo 142
10. Metodo di Aitken 146
11. Metodi iterativi per i sistemi non lineari 150

12. Metodo di Newton-Raphson 155
13. Condizionamento e localizzazione degli zeri di un polinomio . . 166
14. Successione di Sturm 174
15. Metodo di Newton per il calcolo degli zeri di un polinomio . . . 179
16. Metodo di Bairstow 184
17. Metodo di Bernoulli e metodo qd 188
Esercizi proposti 197
Commento bibliografico 246
Bibliografia 250

Capitolo 4 CALCOLO DELLE DIFFERENZE

1. Somme e serie 253
2. Operatore differenza 256
3. Operatore somma 261
4. Funzione gamma 264
5. Polinomi di Bernoulli 267
6. Trasformazione di Eulero 271
7. Equazioni alle differenze lineari 274
8. Stabilità del calcolo delle formule ricorrenti 286
Esercizi proposti 298
Commento bibliografico 349
Bibliografia 350

Capitolo 5 INTERPOLAZIONE

1. Il problema dell'interpolazione 352
2. Polinomio di Lagrange 354
3. Resto nell'interpolazione polinomiale 358
4. Polinomi osculatori 366
5. Polinomio di Newton 370
6. Errori di arrotondamento del polinomio di interpolazione . . . 379
7. Proprietà delle differenze divise 384
8. Interpolazione inversa 390
9. Interpolazione razionale 392
10. Frazioni continue 395
11. Differenze inverse 404
12. Differenze reciproche 410
13. Interpolazione trigonometrica e trasformata discreta di Fourier 417
14. Funzioni spline 434
Esercizi proposti 447
Commento bibliografico 504
Bibliografia 507

Capitolo 6 APPROSSIMAZIONE

1. Introduzione 509
2. Il problema dell'approssimazione lineare 511
3. Polinomi ortogonali 522
4. Approssimazione ai minimi quadrati 540
5. Approssimazione minimax polinomiale 552
6. Algoritmo di Remez 563
7. Approssimazione quasi minimax 570
8. Approssimazione minimax rispetto all'errore relativo 579
9. Approssimazione minimax con vincoli 583
10. Approssimazione minimax razionale 585
11. Approssimazione razionale con frazioni continue infinite 593
12. Approssimazione di Padé 597
13. Approssimazione nel discreto 612
Esercizi proposti 622
Commento bibliografico 713
Bibliografia 716

Capitolo 7 INTEGRAZIONE E DERIVAZIONE APPROSSIMATE

1. Formule di quadratura interpolatorie 720
2. Formule di Newton-Cotes 728
3. Formule newtoniane composte 736
4. Formule gaussiane 743
5. Formule gaussiane pesate 752
6. Integrali impropri 759
7. Formule gaussiane con nodi prefissati 768
8. Quadratura automatica 773
9. Integrazione in più dimensioni 779
10. Metodo Monte Carlo 783
11. Approssimazione delle derivate 787
Esercizi proposti 795
Commento bibliografico 846
Bibliografia 848

Bibliografia generale 850

Indice analitico 852

INTRODUZIONE

La matematica nella società moderna ha assunto una crescente importanza: si può ormai affermare che il livello di cultura matematica è una misura del progresso scientifico e tecnologico di un paese.

Nell'analisi dei problemi del mondo reale la matematica svolge un ruolo determinante. In ogni disciplina scientifica e in ogni settore della tecnologia i modelli matematici che approssimano l'evolversi dell'evento oggetto di studio consentono di simulare, e quindi prevedere, lo sviluppo del fenomeno senza dover effettuare fisicamente esperimenti complessi, costosi e in alcuni casi anche pericolosi: nella progettazione di un velivolo la forma delle ali può essere successivamente adattata, senza dover costruire alcun prototipo, in base ai risultati delle simulazioni numeriche fatte con il calcolatore; lo studio dell'inquinamento ambientale, quale il propagarsi di una sostanza tossica nella acque di un fiume, può essere condotto mediante calcolatore con un adeguato modello matematico, senza dover realizzare anche in forma ridotta, un reale pericoloso esperimento.

Più in generale l'intervento della matematica nella risoluzione dei problemi del mondo reale avviene anche nella vita quotidiana, pur in modo non evidente per l'utente inconsapevole. Si pensi ad esempio, solo per citare qualche caso, ai riproduttori di compact disc, dove la elaborazione digitale richiede opportuni trattamenti numerici dei dati, alle diagnosi ottenute attraverso la tomografia assiale o la risonanza magnetica, alle previsioni del tempo, alla gestione delle linee della metropolitana in funzione dell'affluenza di persone, alla trasmissione e al filtraggio di segnali o immagini via satellite.

Il processo di risoluzione di un problema del mondo reale può essere così schematizzato: una prima fase (di modellizzazione) in cui al problema reale si associa un modello matematico che ne approssima l'evoluzione; una seconda fase in cui il modello matematico viene analizzato e da questo si ricavano proprietà qualitative della soluzione, come esistenza, unicità, regolarità; una terza fase in cui si individuano dei metodi di risoluzione e se ne analizza l'efficienza; infine una quarta fase in cui il metodo di risoluzione viene implementato su calcolatore mediante un adeguato linguaggio di programmazione.

Queste fasi non esauriscono il processo, né il processo si inquadra rigidamente sempre in questo schema. Infatti i risultati ottenuti nella elaborazione automatica possono suggerire modifiche al modello matematico, in modo da renderlo più aderente al problema reale; come pure modifiche possono essere suggerite da possibili semplificazioni dei metodi di risoluzione.

Un aspetto importante in questo processo è che nella quasi totalità dei casi la soluzione del problema non è esprimibile in forma esplicita, ad esempio mediante funzioni elementari, e comunque le sole proprietà quali-

tative non sono sufficienti per gli scopi richiesti. Si pensi ad esempio alla traiettoria che deve descrivere una sonda spaziale per raggiungere un certo obiettivo: la funzione che la descrive è la soluzione di un opportuno sistema di equazioni differenziali; in questo caso è necessario conoscere le coordinate della navicella in un insieme stabilito di istanti temporali.

Si presenta quindi la necessità di risolvere algoritmicamente il problema matematico, cioè di ottenere mediante un numero finito di operazioni aritmetiche e/o logiche una informazione anche se parziale (spesso approssimata), ma adeguata alle richieste, della soluzione. Si parla così di "risoluzione numerica" del problema intendendo che la soluzione ottenuta è calcolata a partire da un insieme finito di numeri (rappresentati con un numero finito di cifre) attraverso un numero finito di operazioni aritmetiche ed è espressa ancora medienate un insieme finito di numeri.

Questo processo di risoluzione di un problema del mondo reale è sempre stato adottato anche nel passato: l'ultima fase era necessariamente risolta con carta e penna, e ciò ha limitato lo sviluppo dei metodi numerici a quei metodi che potevano essere utilizzati solo con mezzi di calcolo manuali. La sintesi e l'analisi dei metodi numerici è antica quanto la matematica: i babilonesi usavano una sorta di metodo delle secanti per risolvere equazioni di primo grado; tecniche di calcolo numerico, sviluppate in passato da alcuni astronomi sono tuttora in uso.

Nella risoluzione di problemi di analisi matematica che generalmente hanno natura continua questo processo algoritmico, necessariamente discreto poiché finito, si inquadra nel settore noto come *analisi numerica*.

Nel tentativo di dare una definizione di analisi numerica si potrebbe dire che l'analisi numerica è la disciplina che sviluppa e studia tutti quegli strumenti matematici atti ad individuare e analizzare metodi di risoluzione numerica nel senso sopra precisato.

L'analisi numerica ha assunto le caratteristiche di una disciplina autonoma solo con l'introduzione e l'uso dei calcolatori, quando l'elaborazione di grandi quantità di dati ha messo in luce nuovi problemi non emersi nel calcolo manuale e quando il dover risolvere in modo finito problemi di natura continua ha creato delle nuove problematiche. Fra queste, particolare rilevanza ha lo studio del condizionamento numerico di un problema, della stabilità numerica di un algoritmo, della complessità computazionale di un problema e le questioni legate alla discretizzazione di un problema continuo. L'introduzione e l'uso di calcolatori ha dato e sta dando un impulso notevole al settore dell'analisi numerica. Sono stati e vengono tuttora realizzati nuovi ed efficienti metodi numerici, che permettono di trattare problemi sempre più complessi che era impensabile trattare pochi anni fa e gran parte dei metodi numerici utilizzati nel passato sono stati sostituiti da metodi più efficienti (il metodo di eliminazione di Gauss è tuttora in uso, anche se è

stato fortemente modificato da strategie atte ad aumentarne l'efficienza).

La possibilità, fornita da metodi numerici efficienti e da calcolatori sempre più potenti, di trattare problemi troppo complessi per gli strumenti tradizionali della matematica applicata e dell'ingegneria, ha aumentato le richieste della scienza e della tecnologia per la risoluzione di nuovi e sempre più complessi problemi, creando un *feed-back* che ha alimentato e continua ad alimentare l'analisi e lo sviluppo di nuovi algoritmi, favorendo una crescita esplosiva delle ricerche in questo campo.

La conoscenza dei metodi numerici è diventata un elemento indispensabile per ogni ricercatore che opera nelle scienze applicate. Le problematiche di tipo numerico formano ormai un bagaglio culturale di base per tutti quelli che svolgono il loro lavoro nel campo tecnico-scientifico.

L'analisi numerica fa parte di un'area più vasta della matematica in cui vengono elaborati ed analizzati i processi algorimici, e che può essere definita con il termine di *matematica computazionale*. Tale area è sempre stata presente nelle varie discipline classiche della matematica, ma solo con l'avvento dei calcolatori ha avuto notevole sviluppo. Oltre all'analisi numerica, altre componenti importanti e in forte sviluppo della matematica computazionale sono la *geometria computazionale* e la *computer algebra*.

In questo libro sono esposti e analizzati i principali metodi dell'analisi numerica. Nel primo capitolo vengono trattati i problemi globali del calcolo e vengono evidenziate, anche attraverso esempi, le problematiche tipiche del settore, in particolare i problemi legati al condizionamento di un problema ed alla stabilità numerica di un algoritmo, e questioni di complessità computazionale e modelli di calcolo. Lo studio della propagazione degli errori è trattato nel seondo capitolo. Nel terzo capitolo vengono studiati problemi relativi alla risoluzione di equazioni e di sistemi non lineari. Il quarto capitolo è dedicato all'esposizione e all'analisi di strumenti classici per lo studio dell'approssimazione, quali le differenze finite, la funzione gamma e i polinomi di Bernoulli. I capitoli successivi sono dedicati all'interpolazione (quinto capitolo) e all'approssimazione di funzioni (sesto capitolo), all'integrazione e alla derivazione numeriche (settimo capitolo).

Nel libro non vengono trattati i metodi numerici per l'algebra lineare, che sono oggetto del testo *Metodi numerici per l'algebra lineare* di D. Bini, M. Capovani e O. Menchi, edito da Zanichelli, i metodi di ottimizzazione e quelli per la risoluzione numerica di equazioni differenziali.

I metodi descritti sono accompagnati da esempi, i cui calcoli sono stati eseguiti su un calcolatore IBM serie /370 in precisione semplice, cioè con numeri rappresentati in virgola mobile con 6 cifre esadecimali, corrispondenti a circa 7 cifre decimali. Per questo la relazione di uguaglianza indicata con il simbolo "=", che interviene fra numeri decimali, è da intendersi relativamente alle sole 7 cifre riportate. Quando di un risultato non è essenziale

determinare molte cifre, ma importa conoscere il solo ordine di grandezza, verranno riportate 3 sole cifre e l'uguaglianza sarà indicata dal simbolo "$\approx$". Gli stessi esempi, implementati su calcolatori diversi, danno ovviamente risultati numerici diversi. Per evidenziare le particolari caratteristiche numeriche occorre allora modificare opportunamente i dati in relazione alle diverse precisioni di macchina.

Gli autori desiderano ringraziare tutti coloro che hanno contribuito alla realizzazione di questo libro: in particolare Mario Arioli e Francesco Romani per il fattivo contributo alla stesura dei capitoli 6 e 7, Fabio Di Benedetto per un'accurata revisione del testo, Bruno Codenotti, Claudia Fassino, Paola Favati, Giotto Fiorio, Luca Gemignani, Mauro Leoncini, Grazia Lotti e Giovanni Resta per gli utili suggerimenti. Un ringraziamento va anche ai molti studenti che nei passati anni accademici hanno utilizzato le varie stesure preliminari come dispense degli insegnamenti numerici dei corsi di laurea in Matematica e in Scienze dell'Informazione dell'Università di Pisa.

Capitolo 1

I PROBLEMI DEL CALCOLO

1. Il calcolatore: la generazione degli errori

Il calcolatore, essendo una macchina *finita*, deve operare su numeri rappresentati per mezzo di una sequenza *finita* di cifre; ad esempio, numeri reali come π o $\sqrt{2}$, la cui rappresentazione richiede infinite cifre, non possono essere trattati nelle elaborazioni numeriche se non commettendo un errore. Lo stesso problema si presenta anche con numeri razionali la cui rappresentazione ha uno sviluppo periodico: ad esempio il numero $1/3$ ha la rappresentazione $0.333\ldots$ in base 10, e può essere rappresentato solo troncandone lo sviluppo e quindi commettendo un errore. La presenza degli errori dovuti alla rappresentazione con un numero finito di cifre è un fatto acquisito ed evidente anche quando si opera con i calcolatori tascabili, così che non ci si meraviglia più di fronte a risultati quali

$$3 \times \frac{1}{3} = 0.9999999.$$

Utilizzando calcolatori che adoperano basi di rappresentazione interna diverse dalla base 10, la presenza degli errori è meno evidente. Ad esempio, battendo 0.1 alla tastiera di un *personal computer*, che utilizza la base 2 per la rappresentazione interna e la base 10 per la rappresentazione sullo schermo, si può leggere sullo schermo il numero 0.1, cioè $1/10$, però il numero effettivamente messo nella memoria del calcolatore non è $1/10$, perché in base 2 tale numero è rappresentato dalla sviluppo periodico $0.00110011\ldots$, il cui troncamento genera un errore di rappresentazione. Quindi in generale si commettono errori già nell'immissione dei dati, prima ancora di eseguire effettivamente i calcoli. Il controllo degli errori generati dall'uso di una rappresentazione di un dato con un numero finito di cifre è molto importante al fine di determinare la *attendibilità* del risultato ottenuto col calcolatore.

Nella maggior parte dei calcolatori la rappresentazione interna dei numeri viene fatta usando una base diversa dalla base 10: di solito si usa la base 2 o la base 16; mentre nei dispositivi di uscita (schermo, stampante ecc.), i numeri compaiono nella loro rappresentazione in base 10.

Poiché l'insieme dei numeri rappresentabili sul calcolatore è un insieme finito e quindi limitato, ovviamente non tutti i numeri sono rappresentabili in modo esatto, e in particolare non potranno essere rappresentati numeri arbitrariamente grandi o arbitrariamente piccoli. A causa di questa limitazione l'elaborazione può essere interrotta perché si è verificata una situazione di *overflow*, con generazione di risultati intermedi di modulo troppo

elevato, o di *underflow* con generazione di risultati intermedi di modulo troppo piccolo. La rappresentazione con un numero finito di cifre impone inoltre l'uso di una aritmetica approssimata (*aritmetica finita*) che introduce errori anche nell'esecuzione delle operazioni aritmetiche. Ad esempio, supponendo per semplicità di usare un calcolatore con una rappresentazione in base 10 con due cifre, il prodotto effettivamente calcolato dei numeri 0.57 e 0.41 è 0.23 ossia il numero di due cifre che meglio approssima il risultato esatto di 0.2337. Per poter valutare la attendibilità del risultato occorre quindi poter controllare la propagazione degli errori generati nell'esecuzione delle operazioni aritmetiche.

I seguenti esempi mostrano come la propagazione degli errori dovuta alla rappresentazione finita e all'uso di una aritmetica finita possa alterare i risultati al di là di ogni ragionevole previsione.

1.1 Esempio. La seguente identità è ovvia in aritmetica esatta

$$b = (1 + a) - 1 = a.$$

Se le operazioni indicate sono eseguite col calcolatore, si ottengono invece i valori:

$$\text{per } a = 10^{-6}, \quad \widetilde{b} = 0.9536743\ 10^{-6},$$
$$\text{per } a = 10^{-7}, \quad \widetilde{b} = 0. \qquad \blacksquare$$

1.2 Esempio. Siano a un numero reale positivo e $b = (1 + a)^2 - 1$. Se $a = 10^{-6}$, il valore di b è dato da

$$b = 2a + a^2 = 0.2000001\ 10^{-5},$$

mentre il risultato fornito dal calcolatore è $\widetilde{b} = 0.190734910^{-5}$. Se $a = 10^{-7}$ il risultato fornito dal calcolatore è $\widetilde{b} = 0$. $\blacksquare$

1.3 Esempio. Sia p un numero reale non nullo. L'equazione di secondo grado

$$x^2 - qx + 1 = 0, \quad \text{dove} \quad q = (p^2 + 1)/p,$$

ha le soluzioni $x_1 = p$ e $x_2 = 1/p$, che possono essere così calcolate

$$x_1 = \frac{q + \sqrt{q^2 - 4}}{2}, \quad x_2 = \frac{q - \sqrt{q^2 - 4}}{2}. \tag{1}$$

I valori $x_1 = p$ e $x_2 = 1/p$ e i risultati $\widetilde{x}_1$ e $\widetilde{x}_2$ effettivamente ottenuti col calcolatore mediante la (1) sono riportati nella tabella seguente.

q	$x_1 = p$	$x_2 = 1/p$	$\widetilde{x}_1$	$\widetilde{x}_2$
100.01	100.	0.01	99.99998	$0.1000977\ 10^{-1}$
1000.001	1000.	0.001	1000.	$0.9765625\ 10^{-3}$
10000.0001	10000.	0.0001	10000.	0.

Si osservi che, mentre la soluzione calcolata $\widetilde{x}_1$ è affetta da un errore trascurabile, la $\widetilde{x}_2$ contiene un errore elevato e per $p = 10000$ nessuna cifra del risultato $\widetilde{x}_2$ fornito dal calcolatore è corretta. Nella figura 1.1 è riportato il grafico, in scala logaritmica, della quantità

$$\left| \frac{x_2 - \widetilde{x}_2}{x_2} \right|,$$

che rappresenta l'errore relativo da cui è affetta la soluzione effettivamente calcolata $\widetilde{x}_2$. Si noti come l'errore tenda ad aumentare al crescere di q.

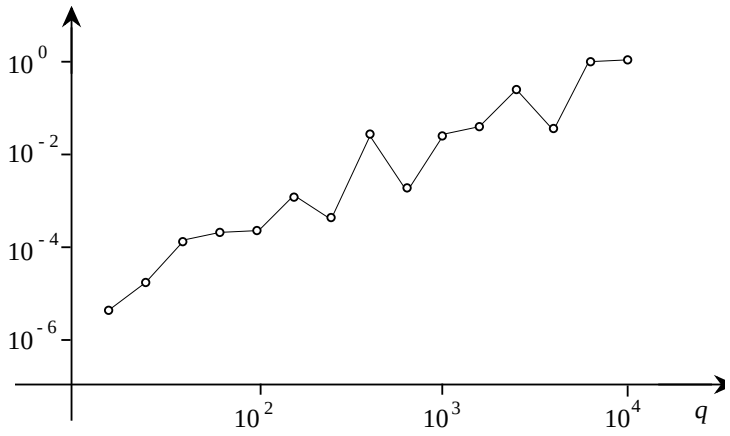


Fig. 1.1 - Errore relativo di $\widetilde{x}_2$ calcolata con (1).

Poiché il prodotto delle radici è uguale a 1, la soluzione x_2 può essere calcolata anche per mezzo della formula

$$x_2 = \frac{1}{x_1}\,. \tag{2}$$

Applicando la (1) per calcolare x_1 e la (2) per calcolare x_2, si ottengono risultati affetti da errori trascurabili. ∎

1.4 Esempio. Si supponga di voler utilizzare il calcolatore per avere delle indicazioni sul valore del limite

$$\lim_{x\to\infty} f(x), \quad f(x) = x(\sqrt{x^2+1} - x), \tag{3}$$

calcolando il valore di $f(x)$ per valori crescenti di x. Si osservi che la funzione $f(x)$ data in (3) può essere scritta anche nelle forme

$$f(x) = x\sqrt{x^2+1} - x^2, \tag{4}$$

$$f(x) = \frac{x}{\sqrt{x^2+1}+x}\,. \tag{5}$$

È quindi possibile calcolare $f(x)$ in questi tre modi diversi. Nella tabella che segue sono riportati alcuni valori r_1, r_2 e r_3 forniti dal calcolatore calcolando $f(x)$ rispettivamente mediante le (3), (4) e (5).

x	r_1	r_2	r_3
1000	0.4882813	0.4375000	0.4999995
2000	0.4882813	0.	0.4999999
3000	0.7324219	0.	0.5
4000	0.9765625	0.	0.5
4050	0.9887695	0.	0.5
4095	0.9997559	0.	0.5

Le informazioni ottenute sono ovviamente discordanti: secondo i valori di r_1, ottenuti con la (3), si potrebbe dedurre che

$$\lim_{x\to\infty} f(x) = 1,$$

secondo i valori di r_2, ottenuti applicando la (4), si potrebbe dedurre che

$$\lim_{x\to\infty} f(x) = 0,$$

secondo i valori di r_3, ottenuti applicando la (5), si deduce il valore corretto del limite

$$\lim_{x\to\infty} f(x) = \frac{1}{2}\,. \qquad \blacksquare$$

1.5 Esempio. Si approssimi il valore di e^{-9} utilizzando lo sviluppo in serie

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \tag{6}$$

Ponendo nella (6) $x = -9$, si ottiene

$$\mathrm{e}^{-9} = 1 - 9 + \frac{9^2}{2!} - \frac{9^3}{3!} + \frac{9^4}{4!} + \dots \tag{7}$$

e ponendo $x = 9$, poiché $\mathrm{e}^{-x} = \dfrac{1}{\mathrm{e}^x}$, si ottiene

$$\mathrm{e}^{-9} = \frac{1}{\mathrm{e}^9} = \frac{1}{1 + 9 + \dfrac{9^2}{2!} + \dfrac{9^3}{3!} + \dfrac{9^4}{4!} + \dots} \tag{8}$$

Le prime 7 cifre del risultato ottenuto sommando n termini della formula (7), si stabilizzano sul valore $-0.9639455\ 10^{-4}$ per $n \geq 38$, mentre le prime 7 cifre del risultato ottenuto sommando n termini della formula (8), si stabilizzano sul valore $0.1234107\ 10^{-3}$ per $n \geq 23$. Poiché e^{-9} è un numero positivo, il metodo che si basa sulla (7) fornisce un risultato completamente inattendibile. Invece il valore ottenuto con il metodo che si basa sulla (8) risulta avere 4 cifre decimali corrette. Nella figura 1.2 sono riportati, al crescere di n, gli errori relativi dei risultati ottenuti sommando n termini della (7) (grafico con i pallini), e sommando n termini della (8) (grafico con i quadratini neri).

■

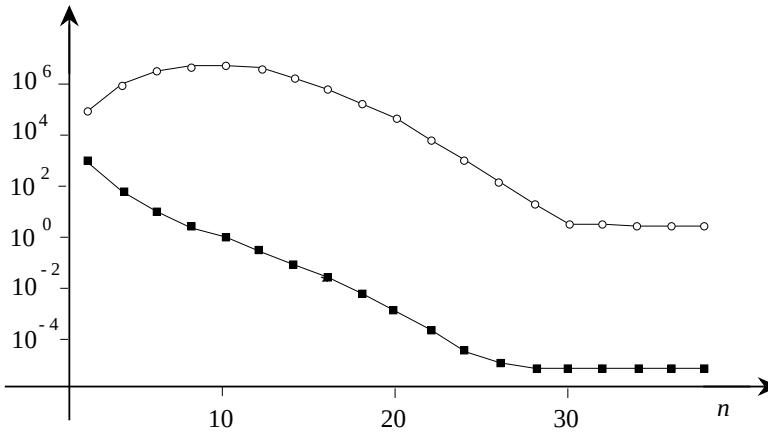


Fig. 1.2 - Grafico degli errori relativi delle approssimazioni di e^{-9} ottenute dalle formule (7) e (8).

1.6 Esempio. Se $f(x)$ è una funzione derivabile due volte con continuità, dalla formula di Taylor si ha:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi), \quad x < \xi < x+h,$$

da cui

$$f'(x) = \frac{f(x+h) - f(x)}{h} \; - \frac{h}{2} f''(\xi).$$

Il valore di $f'(x)$ può quindi essere approssimato, a meno di un errore che tende a zero con h, dal rapporto incrementale:

$$\Delta_f = \frac{f(x+h) - f(x)}{h}\,. \tag{9}$$

Se $f(x)$ è derivabile 3 volte con continuità, dalla formula di Taylor si ha

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(\xi), \quad x < \xi < x+h,$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(\eta), \quad x-h < \eta < x,$$

da cui, sottraendo membro a membro, si ha

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{12}\,(f'''(\xi) + f'''(\eta)).$$

Il valore di $f'(x)$ può quindi essere approssimato, a meno di un errore che tende a zero col quadrato di h, anche dal rapporto:

$$\delta_f = \frac{f(x+h) - f(x-h)}{2h}\,. \tag{10}$$

Per la funzione $f(x) = x^2$ è $f'(x) = 2x$ e quindi $f'(1) = 2$. Poiché è $f''(x) = 2$, la formula (9) consente di approssimare il valore $f'(1)$, a meno di un errore che tende a zero con h, mentre la formula (10), poiché $f'''(x) \equiv 0$, consente di calcolare esattamente $f'(1)$. Nella figura 1.3 sono riportati i valori effettivamente ottenuti con il calcolatore, per valori di h compresi fra 10^{-1} e 10^{-5} (con il pallino i valori ottenuti con la (9) e con il quadratino nero i valori ottenuti con la (10)). Si noti come i risultati ottenuti si discostino sensibilmente dal valore esatto per valori di h piccoli, minori di 10^{-4}. ∎

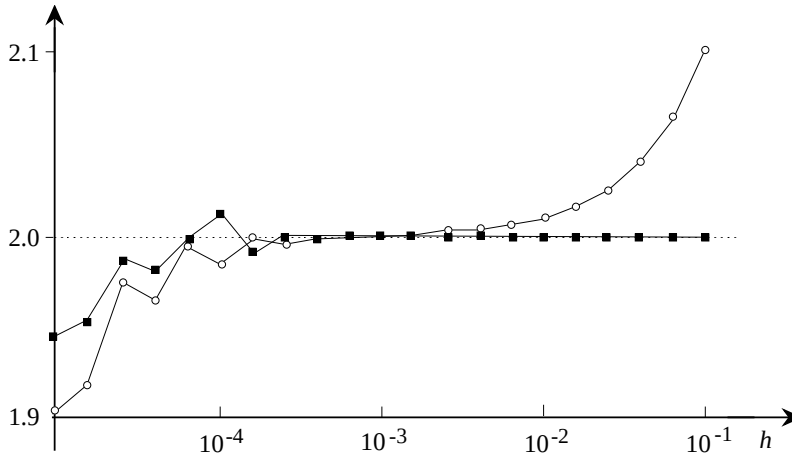


Fig. 1.3 - Approssimazione della derivata.

Da questi esempi risulta che gli errori generati dalla rappresentazione con un numero finito di cifre e dall'uso di una aritmetica finita, possono essere così elevati da togliere ogni validità ai risultati ottenuti ed è quindi molto importante riuscire a controllarne e valutarne la propagazione. Si è visto anche che, utilizzando un procedimento di calcolo opportuno, è possibile contenere tali errori. Esistono quindi, per uno stesso problema,

procedimenti di calcolo (*algoritmi*) che possono generare errori in misura diversa. In questi problemi elementari, dei quali si conosce la soluzione, è stato facile distinguere gli algoritmi *numericamente instabili*, ossia quelli per cui si presenta una elevata propagazione dell'errore, da quelli *numericamente stabili*. In generale nei problemi concreti tale distinzione non è facile ed è quindi molto importante disporre di tecniche per stabilire a priori se un dato algoritmo è numericamente stabile o instabile.

1.7 Esempio. Si vuole determinare la configurazione che assume nello spazio una pellicola di acqua saponata sottesa da una curva chiusa, determinata da una funzione definita sul bordo del quadrato $[0,1] \times [0,1]$. Tale problema è matematicamente approssimato da una equazione alle derivate parziali (equazione di Poisson), la cui soluzione può essere a sua volta approssimata mediante la risoluzione di un sistema di equazioni lineari. Esistono vari metodi per risolvere tale sistema con basso *costo computazionale*, cioè con un ridotto numero di operazioni. I grafici delle figure 1.4 e 1.5 mostrano le soluzioni, nel caso in cui la curva chiusa sia costituita dalla funzione $\frac{1}{2}\,\sin \pi x$ per $y = 0$ o $y = 1$ e $\frac{1}{2}\,\sin \pi y$ per $x = 0$ o $x = 1$, e il sistema, costituito da 256 equazioni con 256 incognite, venga risolto rispettivamente con il metodo dell'*analisi di Fourier* o con il metodo di *marching*.

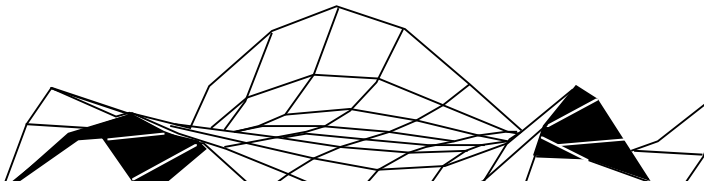


Fig. 1.4 - Configurazione di equilibrio di una pellicola: metodo dell'*analisi di Fourier*.

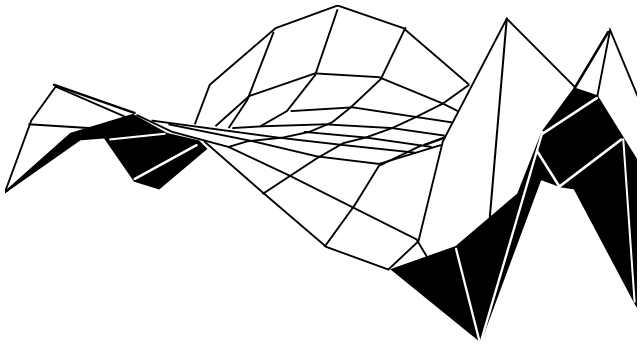


Fig. 1.5 - Configurazione di equilibrio di una pellicola: metodo di *marching*.

Si noti come nel secondo caso siano presenti notevoli irregolarità in prossimità di un lato del quadrato. Ciò è dovuto al fatto che il metodo di marching è numericamente instabile, infatti si può dimostrare che gli errori numerici generati crescono esponenzialmente col numero delle incognite del sistema, e si accumulano maggiormente sulle ultime incognite calcolate, che sono quelle che regolano lo spostamento dei punti vicini ad un lato del quadrato.

∎

Dagli esempi precedenti appare chiaro come non sempre sia possibile approssimare in modo adeguato la soluzione di un problema. In generale questa incertezza non è causata solamente dall'utilizzazione di una aritmetica finita, ma dipende anche dai procedimenti di risoluzione usati, in quanto vengono introdotti errori non solo nella formulazione del problema concreto e nella costruzione di un modello matematico associato, ma anche nel calcolo effettivo della soluzione stessa.

Per studiare un "fenomeno", sia esso naturale quale il moto di un pianeta o sia esso un prodotto umano quale il propagarsi dell'inquinamento atmosferico o la costruzione di un ponte, la matematica fornisce strumenti per realizzare ed analizzare dei modelli che descrivono, sia pure in modo approssimato, il fenomeno stesso. Ad esempio, nel modello matematico per la costruzione di un ponte, è richiesto il bilancio delle forze esercitate sui vari elementi del ponte stesso (peso degli elementi, peso degli automezzi in transito, azione del vento, ecc.). Non si tiene però conto delle forze gravitazionali esercitate sugli elementi del ponte dalla Luna o dal Sole; tali forze sono realmente esistenti e non nulle, e quindi dovrebbero essere considerate in un modello matematico esatto, ma possono di fatto essere trascurate in un modello che *approssima* il problema reale con errori contenuti.

Generalmente un modello matematico è un modello continuo (di solito un sistema di equazioni differenziali) che, pur consentendo di individuare proprietà qualitative della soluzione: esistenza, unicità, regolarità, ecc., non consente sempre di determinare analiticamente la soluzione. È perciò necessario effettuare una approssimazione sostituendo il modello continuo con un modello discreto (*processo di discretizzazione*). Si possono presentare anche casi in cui la risoluzione del problema richiede di utilizzare un metodo iterativo, cioè un metodo che fornisce una sequenza di approssimazioni della soluzione, che converge alla soluzione stessa. Ad esempio, per approssimare una soluzione dell'equazione $x - \cos x = 0$, può essere utilizzata la successione $\{x_i\}$ così definita

$$x_0 = 1, \qquad x_{i+1} = \cos x_i, \quad i = 0, 1, \ldots,$$

la cui interpretazione geometrica è riportata nella figura 1.6.

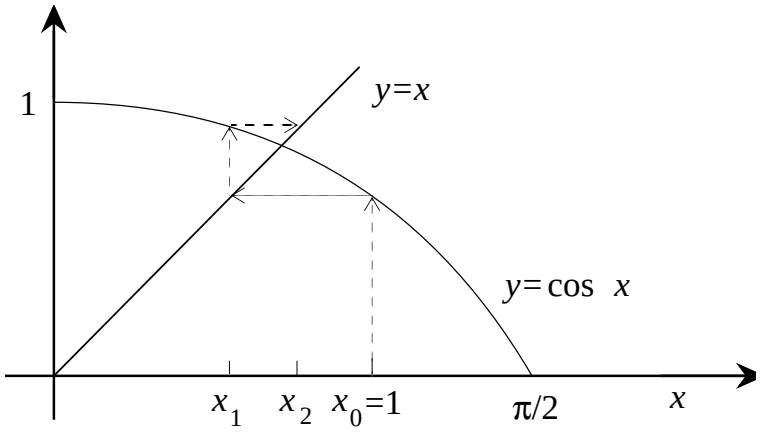


Fig. 1.6 - Interpretazione geometrica del metodo iterativo $x_{i+1} = \cos x_i$.

Nel processo di discretizzazione o quando si utilizzano procedimenti di tipo iterativo, si introduce un errore detto *errore analitico*.

Questa sequenza di approssimazioni è illustrata nello schema della figura 1.7. In ogni passaggio da un blocco all'altro può essere commesso un errore di approssimazione. È importante che gli errori di approssimazione commessi siano dello stesso ordine. Infatti ha scarsa utilità calcolare la soluzione del modello matematico con precisione elevata quando questo è una grossolana approssimazione del fenomeno naturale, o quando il modello discreto approssima con scarsa precisione il modello continuo.

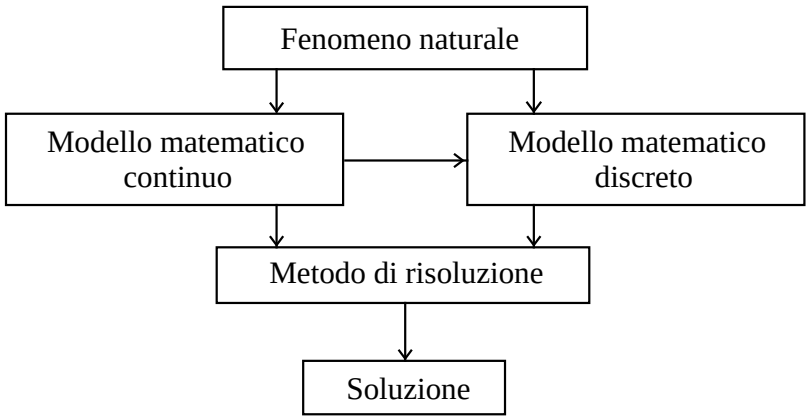


Fig. 1.7 - Successive approssimazioni nell'analisi di un fenomeno.

Infine i risultati effettivamente ottenuti utilizzando il calcolatore sono alterati, oltre che dagli errori generati dalle operazioni aritmetiche svolte con una aritmetica finita (*errore algoritmico*), anche dagli errori generati dalla rappresentazione dei dati con un numero finito di cifre (*errore inerente*).

2. Problemi mal condizionati

Dagli esempi dei paragrafi precedenti risulta che per uno stesso problema esistono procedimenti di risoluzione che producono diversi errori algoritmici. Esistono cioè algoritmi più stabili e algoritmi meno stabili.

Esistono anche problemi tali che, qualunque algoritmo venga utilizzato per risolverli, l'errore generato nel risultato risulta elevato e talvolta tale da rendere privo di significato il risultato stesso. Questo fenomeno è una particolarità intrinseca del problema e non dipende dagli algoritmi utilizzati. Per questi problemi, detti *mal condizionati*, piccole variazioni nei dati inducono grosse variazioni nei risultati. Quindi in questo caso già la inevitabile perturbazione dei dati, dovuta alla rappresentazione finita in base, genera errori elevati nei risultati, qualunque sia l'algoritmo utilizzato. Vi sono problemi per i quali esiste la soluzione ed esiste un algoritmo per calcolarla, ma che sono praticamente irrisolubili perché fortemente malcondizionati, per cui i soli errori di rappresentazione dei dati generano sul risultato errori superiori al 100%.

Si consideri ad esempio il seguente sistema di equazioni lineari

$$\begin{cases} x + y = 2 \\ \\ 1001\,x + 1000\,y = 2001, \end{cases}$$

che ha la soluzione $x = 1,\ y = 1$. Si alteri il coefficiente della x, nella prima equazione, dell'1%, e si consideri il nuovo sistema perturbato

$$\begin{cases} (1 + 1/100)\,x + y = 2 \\ \\ 1001\,x + 1000\,y = 2001, \end{cases}$$

che ha soluzione $\widetilde{x} = -1/9,\ \widetilde{y} = 1901/900$. La soluzione del sistema perturbato presenta, rispetto alla soluzione del sistema non perturbato, una variazione maggiore del 110% sia nella x che nella y.

La natura del mal condizionamento del problema precedente può essere descritta mediante un'interpretazione geometrica. Le due equazioni del sistema corrispondono alle due rette tracciate con linea continua nella figura 1.8 e la loro intersezione ha per componenti la soluzione del sistema. Poiché le due rette formano tra loro un angolo molto piccolo, perturbando la prima equazione, si altera di poco la posizione della prima retta (ottenendo la retta tratteggiata), ma cambia molto la posizione del punto intersezione e quindi le coordinate di questo che sono le soluzioni del sistema perturbato.

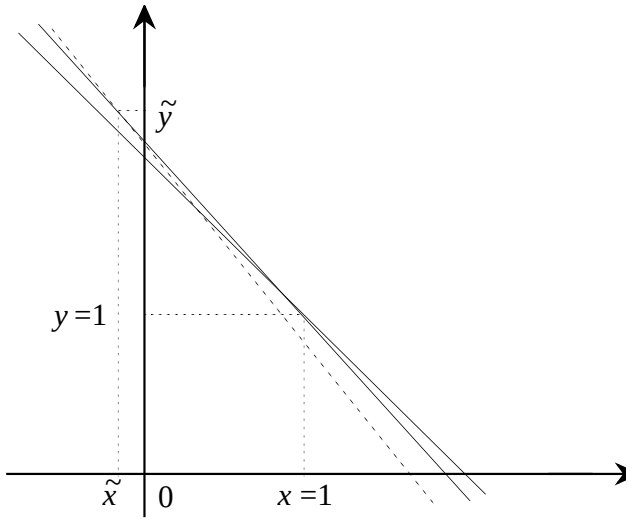


Fig. 1.8 - Interpretazione geometrica del mal condizionamento di un sistema lineare.

Un modo per misurare il condizionamento nel caso in cui il problema sia quello di calcolare il valore che una funzione f di una variabile reale assume in un punto $x \neq 0$, consiste nel valutare la variazione che si produce nel valore della funzione quando il dato x viene perturbato con una perturbazione relativa d. Indicato con $x' = x(1+d)$ il nuovo dato, la corrispondente variazione relativa del risultato è

$$r = \frac{f(x(1+d)) - f(x)}{f(x)} ,$$

e il condizionamento viene misurato per mezzo del quoziente r/d. Se la funzione $f(x)$ è derivabile, si ha

$$\lim_{d \to 0} \frac{r}{d} = \frac{x f'(x)}{f(x)} .$$

Tale quantità, che non dipende dalla perturbazione d, dà un'indicazione di quanto è amplificato l'errore introdotto nella variabile indipendente.

1.8 Esempio. Sia $f(x) = \sqrt{1-x}$. Se $x \neq 1$ è

$$\frac{x f'(x)}{f(x)} = - \frac{x}{2(1-x)} .$$

Quindi il calcolo di $f(x)$ può essere un problema mal condizionato se x è vicino a 1. Infatti per $x = 0.9999$ e $d = 10^{-5}$, l'errore relativo del risultato è in modulo superiore a $0.5\ 10^{-1}$. ∎

Nel caso di malcondizionamento, per trattare il problema conviene aumentare la precisione della rappresentazione dei dati, aumentando il numero di cifre, in modo da ridurre l'errore di rappresentazione. È per questo che su molti calcolatori è possibile utilizzare, oltre alla *precisione semplice*, in cui la rappresentazione in base è fatta con un numero standard di cifre, anche la *precisione doppia*, in cui la rappresentazione in base avviene con un numero all'incirca doppio di cifre, e, più in generale, la *precisione multipla*, in cui si usa un numero arbitrario di cifre. Nel caso dell'esempio 1.7, dove l'errore numerico cresce esponenzialmente con la dimensione del problema, l'aumento del numero di cifre non fornisce alcun concreto vantaggio nella riduzione dell'errore. Esistono comunque molti casi in cui questo modo di operare permette di contenere l'errore.

Si presentano però due inconvenienti: per memorizzare un numero occorre una maggior quantità di memoria, con conseguente riduzione delle dimensioni massime dei problemi trattabili; il tempo necessario per eseguire una operazione su due numeri aumenta all'aumentare del numero delle cifre, con conseguente notevole incremento del tempo di elaborazione. Sul calcolatore una moltiplicazione viene normalmente eseguita con un metodo analogo a quello usato nel calcolo "con carta e penna" di un prodotto di due numeri di n cifre decimali che richiede n^2 moltiplicazioni di numeri di una cifra. Quindi raddoppiando il numero di cifre risulta quadruplicato il tempo richiesto per ogni moltiplicazione: se ad esempio un personal computer esegue mille moltiplicazioni al secondo, raddoppiando il numero di cifre eseguirà solamente 250 moltiplicazioni al secondo. L'uso di un numero maggiore di cifre per ridurre la propagazione degli errori comporta quindi un impiego maggiore delle risorse *quantità di memoria* e *tempo di elaborazione*.

3. Complessità di calcolo

Una scelta accorta di un metodo per la risoluzione di un problema consente una migliore utilizzazione delle risorse tempo di elaborazione e quantità di memoria, permettendo così di trattare anche problemi di elevate dimensioni.

Si consideri ad esempio il problema del calcolo della soluzione di un sistema di n equazioni lineari in n incognite

$$A\mathbf{x} = \mathbf{b}, \tag{11}$$

dove A è una matrice di ordine n non singolare di elementi a_{ij}, $i,\ j = 1, \ldots, n$, $\mathbf{x}$ è il vettore delle incognite e $\mathbf{b}$ è il vettore dei termini noti, i cui elementi sono rispettivamente x_i e b_i, $i = 1, \ldots, n$. In componenti il sistema (11) si scrive

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, \ldots, n.$$

Un metodo di risoluzione consiste nell'applicare la regola di Cramer mediante la quale le componenti della soluzione vengono espresse come quozienti di determinanti di matrici di ordine n

$$x_j = \frac{\det A_j}{\det A}, \quad j = 1, 2, \ldots, n,$$

dove A_j è la matrice ottenuta dalla matrice A sostituendo la j-esima colonna col vettore $\mathbf{b}$. Tali determinanti possono essere calcolati utilizzando la regola di Laplace attraverso lo sviluppo per righe

$$\det A = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} \det A_{1j},$$

dove A_{1j} è la matrice di ordine $n-1$ ottenuta da A eliminando la prima riga e la j-esima colonna. In tal modo un determinante di una matrice di ordine n viene espresso come combinazione lineare di n determinanti di matrici di ordine $n-1$. Se C_n indica il numero di moltiplicazioni richieste per il calcolo del determinante di una matrice di ordine n con la regola di Laplace, vale la relazione

$$\begin{aligned} C_2 &= 2, \\ C_n &= nC_{n-1} + n \geq nC_{n-1}, \end{aligned}$$

da cui $C_n \geq n(n-1)\ldots 2 = n!$. Quindi risolvere un sistema di n equazioni ed n incognite col metodo di Cramer costa almeno $(n+1)!$ moltiplicazioni.

Se il sistema viene risolto col metodo di sostituzione (metodo di eliminazione di Gauss), si può verificare che il numero di moltiplicazioni richieste è $n^3/3 + n^2 - n/3$.

Se per semplicità il tempo di esecuzione viene stimato moltiplicando il tempo richiesto da una singola moltiplicazione per il numero di moltiplicazioni, nell'ipotesi che una moltiplicazione venga eseguita in 10^{-6} secondi, per la risoluzione di un sistema di n equazioni ed n incognite con i due metodi in esame si avrebbero i seguenti tempi

n	metodo di Cramer	metodo di sostituzione
12	104 minuti	$7.2\ 10^{-4}$ secondi
13	24 ore	$9.0\ 10^{-4}$ secondi
14	15 giorni	$1.1\ 10^{-3}$ secondi
20	$1.7\ 10^{6}$ anni	$3.1\ 10^{-3}$ secondi
30	$2.7\ 10^{20}$ anni	$9.9\ 10^{-3}$ secondi
40	$1.1\ 10^{36}$ anni	$2.3\ 10^{-2}$ secondi
50	$5.0\ 10^{52}$ anni	$4.4\ 10^{-2}$ secondi

Utilizzando il metodo di Cramer, il problema della risoluzione di un sistema di equazioni lineari non risulta trattabile già per valori piccoli di n. La possibilità di risolvere sistemi di grosse dimensioni, come quelli che derivano da problemi concreti, non dipende solo dalla disponibilità di calcolatori potenti (anche con tali calcolatori il metodo di Cramer richiederebbe tempi inaccettabili di elaborazione) ma dall'uso di metodi computazionalmente efficienti. È quindi con lo sviluppo di algoritmi efficienti e la realizzazione del relativo *software*, oltre che con lo sviluppo del *hardware*, che è possibile trattare problemi di sempre più grosse dimensioni.

Il metodo di Gauss permette di ridurre il costo computazionale del problema da $(n+1)!$ a circa $n^3/3$ moltiplicazioni. È naturale chiedersi se tale metodo sia ottimo, ossia se non esistano algoritmi che richiedono un numero inferiore di moltiplicazioni (l'ottimalità è stata dimostrata relativamente ai metodi che usano solo combinazioni di righe e di colonne). Tale tipo di problema fa parte di una classe più ampia, oggetto di studio del settore della *complessità computazionale* che ha avuto recentemente consistenti sviluppi. Nel caso della risoluzione di un sistema di n equazioni lineari in n incognite sono stati recentemente individuati metodi che richiedono non più di kn^θ operazioni aritmetiche, dove k è una costante positiva e $\theta < 3$ (il più piccolo valore noto di θ è $2.49\ldots$). Tali metodi si basano sul fatto che il costo computazionale della risoluzione di un sistema di n equazioni lineari in n incognite è asintoticamente uguale al costo computazionale del calcolo del prodotto di due matrici di ordine n, e per tale problema esistono algoritmi che richiedono un numero basso di operazioni aritmetiche. Nell'esempio seguente viene descritto l'algoritmo di Strassen per moltiplicare matrici di ordine n con n^θ moltiplicazioni, $\theta = log_2 7 = 2.807\ldots$

1.9 Esempio (Algoritmo di *Strassen*). Siano $A,\ B,\ C$, matrici di ordine $n = 2^k$, con k intero positivo, tali che $C = AB$, ossia

$$c_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj}, \quad i,\ j = 1, \ldots, n. \tag{12}$$

Se gli n^2 elementi c_{ij} vengono calcolati utilizzando direttamente la formula (12), occorrono n^3 moltiplicazioni ed $n^3 - n^2$ addizioni. È possibile però calcolare gli n^2 elementi c_{ij} con non più di $4.7n^\theta$ operazioni aritmetiche, $\theta = \log_2 7$, mediante l'algoritmo di Strassen qui descritto.

Se $k = 1$ gli elementi c_{ij} dati da

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

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

$$\begin{array}{ll} s_1 = (a_{11} + a_{22})(b_{11} + b_{22}) & s_2 = (a_{21} + a_{22})b_{11} \\ s_3 = a_{11}(b_{12} - b_{22}) & s_4 = a_{22}(b_{21} - b_{11}) \\ s_5 = (a_{11} + a_{12})b_{22} & s_6 = (a_{21} - a_{11})(b_{11} + b_{12}) \\ s_7 = (a_{12} - a_{22})(b_{21} + b_{22}) & \\ c_{11} = s_1 + s_4 - s_5 + s_7 & c_{12} = s_3 + s_5 \\ c_{21} = s_2 + s_4 & c_{22} = s_1 - s_2 + s_3 + s_6. \end{array} \tag{13}$$

Poiché nelle relazioni (13) non viene utilizzata la proprietà commutativa della moltiplicazione, è possibile applicare tali formule anche nel caso in cui gli elementi a_{ij}, b_{ij}, c_{ij} sono matrici A_{ij}, B_{ij}, C_{ij}.

Si supponga che $k > 1$ e si partizionino le matrici A, B e C in quattro sottomatrici quadrate di ordine $n/2$.

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ \\ B_{21} & B_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ \\ C_{21} & C_{22} \end{bmatrix}$$

Poiché vale

$$\begin{array}{ll} C_{11} = A_{11}B_{11} + A_{12}B_{21}, & C_{12} = A_{11}B_{12} + A_{12}B_{22}, \\ C_{21} = A_{21}B_{11} + A_{22}B_{21}, & C_{22} = A_{21}B_{12} + A_{22}B_{22}, \end{array}$$

è possibile calcolare la matrice prodotto C, mediante le relazioni (13), con 7 moltiplicazioni di matrici di ordine $n/2$.

Poiché le moltiplicazioni di matrici di ordine $n/2$ possono essere effettuate mediante lo stesso metodo, ossia partizionando ciascuna matrice di ordine $n/2$ in quattro sottomatrici di ordine $n/4$ e calcolando per ciascuna moltiplicazione 7 prodotti di matrici di ordine $n/4$, si può procedere in questo modo finché si ottengono matrici di ordine 1. Se con C_n si indica il numero di moltiplicazioni impiegate dal metodo descritto per moltiplicare matrici di ordine n, vale la relazione

$$C_n = 7C_{n/2},$$

da cui, essendo $n = 2^k$ e $C_1 = 1$, segue

$$C_n = 7C_{n/2} = 7^2C_{n/4} = \ldots = 7^kC_1 = n^\theta, \quad \theta = \log_2 7.$$

Il metodo può essere applicato anche nel caso in cui n non è potenza di 2, bordando le matrici con elementi nulli in modo da ottenere matrici di

dimensione 2^k dove k è il minimo intero maggiore o uguale a $\log_2 n$. Procedendo in modo analogo è possibile dimostrare che il numero delle operazioni aritmetiche richieste (addizioni incluse) è minore di $4.7n^\theta$. ∎

1.10 Esempio (calcolo del valore di un polinomio in un punto). Si consideri il polinomio di grado n

$$p(x) = \sum_{i=0}^{n} a_i x^i.$$

Il calcolo di $p(x)$, per un valore assegnato di x, può essere effettuato calcolando separatamente le potenze x^i, mediante il seguente algoritmo

$$\begin{aligned} &p_0 = a_0, \quad y_0 = 1 \\ &y_i = y_{i-1}x, \quad p_i = a_i y_i + p_{i-1}, \quad i = 1, 2, \ldots, n, \\ &p(x) = p_n, \end{aligned}$$

che richiede $2n - 1$ moltiplicazioni ed n addizioni. Poiché il polinomio $p(x)$ può anche essere rappresentato nel modo seguente

$$p(x) = (\ldots((a_n x + a_{n-1})x + a_{n-2})x + \ldots + a_1)x + a_0,$$

si può ottenere un altro algoritmo per il calcolo di $p(x)$, detto *metodo di Ruffini-Horner* o della *divisione sintetica*:

$$\begin{aligned} &p_0 = a_n, \\ &p_i = p_{i-1}x + a_{n-i}, \quad i = 1, 2, \ldots, n, \\ &p(x) = p_n, \end{aligned}$$

che impiega n moltiplicazioni ed n addizioni, cioè il numero di moltiplicazioni è dimezzato rispetto al metodo precedente. È stato dimostrato [12] che il metodo di Horner è ottimale nel caso in cui il polinomio venga individuato mediante i suoi coefficienti. ∎

È però possibile, mediante una preelaborazione dei coefficienti del polinomio (*precondizionamento*) calcolare il valore in un punto con un numero di operazioni pari a circa la metà di quello richiesto dal metodo di Horner. Per il precondizionamento è però richiesto il calcolo preliminare delle radici di un'equazione algebrica che è di grado superiore al secondo se il polinomio ha grado superiore a 8. Per questo motivo tali metodi hanno principalmente interesse teorico in quanto per certi valori della variabile possono non essere stabili, inoltre il loro uso è conveniente solo nel caso che il numero dei punti in cui si calcola il polinomio sia sufficiente a far sì che il costo

della preelaborazione possa essere assorbito dalla riduzione del costo nella seconda fase. Per il calcolo del valore di un polinomio di grado n in $n+1$ punti è poi possibile utilizzare metodi veloci basati sull'algoritmo FFT per il calcolo della trasformata discreta di Fourier, che richiedono un numero di operazioni aritmetiche dell'ordine di $n \log_2^2 n$ (si vedano gli esercizi del capitolo 5.)

1.11 Esempio (prodotto di numeri complessi). Si considerino i numeri complessi $x = x_1 + \boldsymbol{i} x_2$, $y = y_1 + \boldsymbol{i} y_2$, $z = z_1 + \boldsymbol{i} z_2$, dove $x_1,\ x_2,\ y_1,\ y_2,\ z_1,\ z_2$, sono numeri reali e $\boldsymbol{i}$ è l'unità immaginaria cioè $\boldsymbol{i}^2 = -1$. Se $z = xy$, allora

$$\begin{aligned} z_1 &= x_1 y_1 - x_2 y_2, \\ z_2 &= x_1 y_2 + x_2 y_1. \end{aligned}$$

I valori di z_1 e di z_2 possono essere calcolati con 4 moltiplicazioni e 2 addizioni di numeri reali. È però possibile ridurre a 3 il numero delle moltiplicazioni, con un conseguente aumento del numero delle addizioni, col seguente algoritmo

$$\begin{aligned} &s_1 = (x_1 + x_2)(y_1 - y_2), \\ &s_2 = x_1 y_2, \quad s_3 = x_2 y_1, \\ &z_1 = s_1 + s_2 - s_3, \quad z_2 = s_2 + s_3. \end{aligned}$$

∎

Si è visto che per uno stesso problema esistono metodi di risoluzione di costo computazionale diverso. Esistono però problemi per i quali non si conoscono (e probabilmente non esistono) algoritmi di risoluzione che hanno un costo inferiore al costo esponenziale con le dimensioni del problema. Tali problemi, *intrinsecamente complessi*, non sono in generale trattabili, anche per dimensioni non elevate.

Un problema appartenente a questa classe, molto importante per le sue applicazioni, è il problema del "commesso viaggiatore", dove, assegnato un insieme di città con le relative distanze e una percorrenza massima possibile, si chiede di determinare, se esiste, un percorso che tocchi tutte le città e che abbia lunghezza non superiore alla percorrenza massima. Gli algoritmi per risolvere tale problema, nel caso generale, prendono in considerazione tutti i percorsi possibili e hanno quindi un costo esponenziale nel numero delle città.

Una possibilità per trattare alcuni problemi computazionalmente difficili è quella di indebolire le condizioni che individuano il modello di calcolo. Le considerazioni svolte finora presuppongono un modello dove ad ogni passo è possibile effettuare una singola operazione aritmetica su dati iniziali

o quantità calcolate ai passi precedenti. Tale modello è particolarmente adatto per lo studio del costo di un algoritmo realizzato su un calcolatore "tradizionale", cioè dotato di una sola unità in grado di effettuare un'operazione aritmetica alla volta.

Oltre a questo modello di calcolo, che si definisce *sequenziale*, esistono modelli di *calcolo parallelo*. In un modello di calcolo parallelo si dispone di $p > 1$ *processori* aritmetici, in grado di effettuare ad ogni istante p operazioni aritmetiche su dati iniziali o quantità calcolate ai passi precedenti, con risultati che vengono memorizzati in una memoria comune a cui tutti i processori simultaneamente possono accedere. Tale situazione ideale è descritta nel modello più generale di calcolo parallelo noto come modello PRAM (Parallel RAM), in cui il costo di un algoritmo viene misurato dalla coppia (s, p) dove s è il numero di *passi* impiegati dall'algoritmo, e p è il numero di processori utilizzati. Questo modello, particolarmente semplice, non è molto aderente alle situazioni reali, dove il tempo di elaborazione non è legato solo al numero di passi richiesti ma anche all'esecuzione di altre operazioni come, ad esempio, il trasferimento dei dati. Esistono però modelli più sofisticati che meglio rappresentano i reali calcolatori paralleli o i calcolatori vettoriali prodotti con le nuove tecnologie.

Come esempio di quanto possa essere ridotta la complessità di un problema in ambiente di calcolo parallelo si consideri il caso della somma di n numeri

$$s = \sum_{i=1}^{n} x_i.$$

Il classico algoritmo utilizzato per il calcolo di s, cioè

$$\begin{aligned} &s_1 = x_1, \\ &s_i = s_{i-1} + x_i, \quad i = 2, \ldots, n, \\ &s = s_n, \end{aligned}$$

è intrinsecamente sequenziale e quindi non sfrutta la possibilità di effettuare simultaneamente più operazioni e richiede $n-1$ passi. L'algoritmo parallelo, qui descritto nel caso in cui $n = 2^k$

$$\begin{aligned} &v_j^{(1)} = x_{2j-1} + x_{2j}, \quad j = 1, \ldots, n/2; \\ &v_j^{(i)} = v_{2j-1}^{(i-1)} + v_{2j}^{(i-1)}, \quad j = 1, \ldots, n/2^i, \quad i = 2, \ldots, k, \\ &s = v_1^{(k)}; \end{aligned}$$

utilizzando $n/2$ processori, permette di calcolare la somma in $k = \log_2 n$ passi.

Il problema fondamentale del calcolo parallelo è quello di individuare metodi di risoluzione intrinsecamente paralleli che possano sfruttare appieno la disponibilità di più processori. Molti metodi veloci in un ambiente di calcolo sequenziale risultano estremamente lenti in ambiente di calcolo parallelo proprio perché concepiti in modo intrinsecamente sequenziale. Ad esempio, il metodo di eliminazione di Gauss per l'inversione di matrici $n \times n$ o per la risoluzione di un sistema di n equazioni in n incognite, è intrinsecamente sequenziale, e richiede un numero di passi proporzionale ad n utilizzando n^2 processori. Un algoritmo di basso costo computazionale per l'inversione in parallelo di matrici è stato proposto da Csanky [6] e richiede un numero di passi proporzionale a $\log_2^2 n$ utilizzando n^4 processori. Non è noto se utilizzando n^c processori, con c costante, tale problema possa essere risolto in meno di $\log_2^2 n$ passi, mentre si può dimostrare che $2 \log_2 n$ passi sono necessari [3].

Il modello di calcolo può essere generalizzato richiedendo che l'algoritmo non fornisca necessariamente un risultato corretto, ma ammettendo anche risultati affetti da errore purché questo sia controllabile a priori (*modello di calcolo approssimato*). Il seguente esempio [1] mostra come sia possibile approssimare con precisione arbitraria il prodotto di una matrice 2×2 per un vettore con sole 3 moltiplicazioni mentre per il calcolo esatto ne sono richieste 4.

1.12 Esempio. Siano

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Le componenti del vettore $\mathbf{y} = A\mathbf{x}$ sono

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ y_2 &= a_{21}x_1 + a_{22}x_2. \end{aligned}$$

Posto $\epsilon \neq 0$, le relazioni

$$\begin{aligned} y_1 + \epsilon a_{11}a_{12} &= (a_{11} + \epsilon^{-1}x_2)(\epsilon a_{12} + x_1) - \epsilon^{-1}x_1x_2 \\ y_2 + \epsilon a_{21}a_{22} &= (a_{21} + \epsilon^{-1}x_2)(\epsilon a_{22} + x_1) - \epsilon^{-1}x_1x_2 \end{aligned}$$

consentono di approssimare y_1 e y_2 con errore di modulo arbitrariamente piccolo, ma non nullo, eseguendo solo 3 moltiplicazioni, se non si contano le moltiplicazioni per ϵ o ϵ^{-1}. Scegliendo ϵ potenza della base della rappresentazione, tali moltiplicazioni si riducono a semplici traslazioni delle cifre della rappresentazione e quindi hanno tempi di esecuzione bassi. ∎

Altri modelli di calcolo in cui non si ha certezza della correttezza del risultato sono quelli *probabilistici*. Ad esempio esistono algoritmi, detti algoritmi *Monte Carlo*, che forniscono un risultato la cui correttezza è nota in senso probabilistico. Cioè la probabilità, che l'errore da cui è affetto il risultato fornito sia superiore a una soglia prefissata, è nota a priori, e generalmente è piccola. Un esempio di facile descrizione, anche se computazionalmente non efficiente, riguarda il calcolo delle prime t cifre di π.

1.13 Esempio. Per approssimare π si consideri un quadrato di lato 2 e il cerchio di raggio 1 iscritto nel quadrato, come nella figura 1.9.

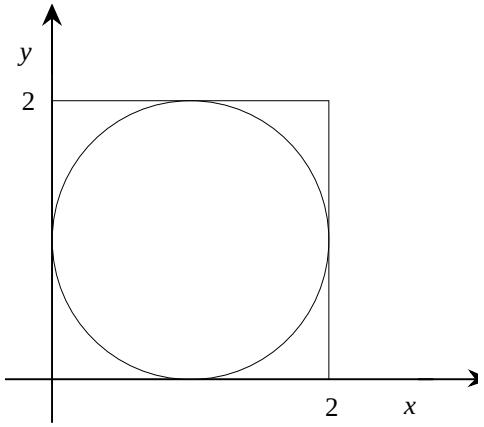


Fig. 1.9 - Costruzione geometrica per il calcolo di π con il metodo Monte Carlo.

Il rapporto fra l'area del cerchio e quella del quadrato è $\dfrac{\pi}{4}$. Si generano n coppie di numeri pseudocasuali indipendenti (x, y) con distribuzione uniforme in $[0, 2] \times [0, 2]$. Se δ_n è il numero di punti di coordinate (x, y) interni al cerchio, si approssima π col rapporto $p_n = 4\,\dfrac{\delta_n}{n}$. È possibile dimostrare che per n sufficientemente grande la probabilità che risulti

$$\left|\frac{p_n - \pi}{\pi}\right| < \frac{k}{\sqrt{n}}\,\sqrt{\frac{4}{\pi} - 1} \tag{14}$$

è data da $\mathrm{erf}\Big(\dfrac{k}{\sqrt{2}}\Big)$ (si vedano il paragrafo 11 del capitolo 2 e l'esercizio 2.40). In particolare per $k = 3$ il valore della funzione erf è 0.9973, e quindi la probabilità di errore è inferiore allo 0.27%, per $k = 4$ il valore della funzione erf è 0.99994, e quindi la probabilità di errore è inferiore allo 0.006%. Per avere quindi 2 cifre corrette con probabilità 0.99994 bastano $1.6\ 10^5$ estrazioni, per avere 4 cifre corrette bastano $1.6\ 10^9$ estrazioni. ∎

Il metodo Monte Carlo può essere applicato ad una vasta gamma di problemi, per i quali non esistono metodi analitici o, se esistono, sono eccessivamente complicati, ma in generale nella pratica fornisce risultati modesti a causa della sua bassa velocità di convergenza. Viene usato soprattutto in simulazione e nel calcolo approssimato degli integrali (e particolarmente in ambiente di calcolo parallelo).

Vi sono però casi in cui l'incertezza sul risultato permette generalmente una riduzione della complessità computazionale. Un esempio significativo a questo proposito è l'algoritmo di Solovay e Strassen [17] per determinare se un numero intero è primo.

Esistono poi algoritmi, detti *algoritmi Las Vegas*, che non sempre forniscono un risultato; quando però il risultato è fornito, questo è certamente corretto, ma la probabilità che l'algoritmo si arresti senza dare alcun risultato è non nulla. In questo caso, anche se la probabilità di insuccesso è elevata, bastano poche applicazioni dell'algoritmo per avere elevata probabilità di ottenere il risultato. Infatti la probabilità di insuccesso in tutte le applicazioni dell'algoritmo decresce esponenzialmente a zero con il numero delle applicazioni. Poiché le applicazioni sono indipendenti, gli algoritmi Las Vegas sono particolarmente adatti in un ambiente di calcolo parallelo. Un algoritmo Las Vegas può essere ottenuto da un algoritmo Monte Carlo applicando un test di correttezza dei risultati forniti.

Esercizi proposti

1.1 Si verifichi con un esempio che utilizzando la proprietà commutativa della moltiplicazione è possibile ridurre il numero di moltiplicazioni sufficienti a calcolare una funzione.

(Traccia: si consideri la funzione $f(x, y) = x^2 - y^2$.)

1.2 Siano $\mathbf{x}$ e $\mathbf{y}$ due vettori di ordine n (n pari). Si valuti il numero di operazioni sufficienti a calcolare il prodotto scalare

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i,$$

mediante il seguente algoritmo di Winograd.

$$\mathbf{x}^T\mathbf{y} = h(\mathbf{x}, \mathbf{y}) - w(\mathbf{x}) - w(\mathbf{y}),$$

dove

$$h(\mathbf{x},\mathbf{y}) = (x_1+y_2)(x_2+y_1)+(x_3+y_4)(x_4+y_3)+\ldots+(x_{n-1}+y_n)(x_n+y_{n-1}),$$

$$w(\mathbf{x}) = x_1x_2 + x_3x_4 + \ldots + x_{n-1}x_n, \quad w(\mathbf{y}) = y_1y_2 + y_3y_4 + \ldots + y_{n-1}y_n.$$

Si dica se viene fatto uso della proprietà commutativa della moltiplicazione.

(Traccia: per il calcolo di $w(\mathbf{x})$ sono sufficienti $n/2$ moltiplicazioni e $n/2-1$ addizioni, per il calcolo di $h(\mathbf{x},\mathbf{y})$ sono sufficienti $n/2$ moltiplicazioni e $3n/2-1$ addizioni. Complessivamente $3n/2$ moltiplicazioni e $5n/2-1$ addizioni.)

1.3 Siano A e B due matrici di ordine n. Gli elementi della matrice $C = AB$ sono dati da

$$c_{ij} = \mathbf{a}_i^T \mathbf{b}_j,$$

dove $\mathbf{a}_i^T$ è la i-esima riga di A e $\mathbf{b}_j$ è la j-esima colonna di B. Si dica quante operazioni sono sufficienti per calcolare C se il prodotto dei due vettori è fatto con l'algoritmo di Winograd dell'esercizio 1.2.

(Traccia: per le funzioni $w(\mathbf{a}_i)$, $i = 1,\ldots,n$, sono sufficienti $n^2/2$ moltiplicazioni e $n(n/2-1)$ addizioni, per le funzioni $h(\mathbf{a}_i,\mathbf{b}_j)$, $i,j = 1,\ldots,n$, sono sufficienti $n^3/2$ moltiplicazioni e $n^2(3n/2-1)$ addizioni. Complessivamente $n^3/2+n^2$ moltiplicazioni e $3n^3/2+2n^2-2n$ addizioni.)

1.4 Siano $k > 1$ un intero e $x \neq 1$. Per calcolare il valore del polinomio

$$p(x) = \sum_{i=0}^{n} x^i, \quad n = 2^k - 1,$$

si possono usare le seguenti relazioni

a) $p(x) = (\ldots((x+1)x+1)\ldots)x+1$, (regola di Ruffini-Horner)

b) $p(x) = \dfrac{x^{n+1}-1}{x-1}$,

c) $p(x) = (1+x)\,(1+x^2)\,(1+x^4)\ldots(1+x^{2^{k-1}})$.

In b) e c) le potenze x^{2^i}, $i = 1,\ldots,k$, vengono calcolate come

$$x^{2^i} = \left(x^{2^{i-1}}\right)^2.$$

Si valuti il costo computazionale dei tre algoritmi così ottenuti.

1.5 Siano $x > 0$ e $n > 1$ un intero. Per calcolare x^n si può procedere anche nel modo seguente: posto

$$k = \lfloor \log_2 n \rfloor,$$

sia

$$n = (d_k d_{k-1} \dots d_1 d_0)_2$$

la rappresentazione in base 2 di n; si calcolino le potenze $z_i = x^{2^i}$, $i = 0, \dots, k$, e si ottenga x^n come

$$x^n = \prod_{\substack{i=0 \\ d_i=1}}^{k} z_i \ .$$

Si valuti il costo computazionale di questo algoritmo.

1.6 Siano

$$a(x) = \sum_{i=0}^{n} a_i x^i, \quad b(x) = \sum_{i=0}^{n} b_i x^i, \quad c(x) = \sum_{i=0}^{2n} c_i x^i,$$

tre polinomi di grado rispettivamente non superiore a n, n, $2n$, tali che

$$c(x) = a(x)b(x).$$

I coefficienti dei polinomi sono legati dalla relazione

$$c_k = \sum_{i=\max(0,k-n)}^{\min(k,n)} a_i b_{k-i}, \quad k = 0, \dots, 2n. \tag{15}$$

a) Si dica quante operazioni aritmetiche sono richieste per calcolare i coefficienti c_k con le (15).

b) Per calcolare i coefficienti di $c(x)$ si può utilizzare anche un algoritmo basato sulla seguente osservazione: posto, per semplicità, n potenza di 2, si ha per $n \geq 2$

$$a(x) = a'(x) + z a''(x), \quad a'(x) = \sum_{i=0}^{n/2-1} a_i x^i, \quad a''(x) = \sum_{i=n/2}^{n} a_i x^{i-n/2},$$

$$b(x) = b'(x) + z b''(x), \quad b'(x) = \sum_{i=0}^{n/2-1} b_i x^i, \quad b''(x) = \sum_{i=n/2}^{n} b_i x^{i-n/2},$$

$$z = x^{n/2},$$

e vale

$$c(x) = a'(x)b'(x) + z[a'(x)b''(x) + a''(x)b'(x)] + z^2 a''(x)b''(x). \tag{16}$$

Si calcola

$$a'(x)b'(x), \quad a''(x)b''(x)$$

e

$$\begin{aligned} a'(x)b''(x) + a''(x)b'(x) =& [a'(x) + a''(x)][b'(x) + b''(x)] \\ & - a'(x)b'(x) - a''(x)b''(x) \end{aligned}$$

con tre moltiplicazioni di polinomi di grado minore o uguale a $\frac{n}{2}$ e da questi, per mezzo delle (16), si calcolano i coefficienti di $c(x)$. Si procede ricorsivamente per il calcolo del prodotto di polinomi di grado inferiore. Si valuti il costo computazionale di tale algoritmo.

(Traccia: b) si segua una tecnica analoga a quella dell'esempio 1.9; l'algoritmo richiede kn^α operazioni aritmetiche, dove k è una costante positiva e $\alpha = \log_2 3 \leq 1.585$. Esistono metodi per moltiplicare polinomi di grado n, il cui costo computazionale è dell'ordine di $n \log_2 n$ e si basano sull'algoritmo FFT per il calcolo della trasformata discreta di Fourier, si vedano gli esercizi del capitolo 5.)

1.7 Utilizzando il metodo descritto nell'esercizio 1.6 b), si determini un algoritmo per moltiplicare numeri interi la cui rappresentazione in base abbia non più di n cifre, e se ne dia il costo computazionale. A differenza del caso dell'esercizio precedente, si tenga conto della presenza del riporto.

(Traccia: l'algoritmo richiede kn^α operazioni fra singole cifre, dove k è una costante positiva e $\alpha = \log_2 3 \leq 1.585$. Il prodotto di interi con n cifre può essere calcolato con $k'n \log_2 n \log_2(\log_2 n)$ operazioni fra singole cifre, dove k' è una costante positiva, usando il metodo di Schönhage-Strassen [15].)

1.8 Una matrice A di Toeplitz di ordine n ha elementi individuati da $2n-1$ parametri $\alpha_{1-n}, \ldots, \alpha_{n-1}$ nel modo seguente

$$a_{ij} = \alpha_{i-j}, \quad i,\ j = 1, \ldots, n.$$

Si dimostri che il calcolo del prodotto

$$\mathbf{u} = A\mathbf{v}, \quad \mathbf{u} = [u_1, \ldots, u_n]^T, \quad \mathbf{v} = [v_1, \ldots, v_n]^T,$$

è riconducibile al calcolo del prodotto di due polinomi, e se ne valuti il costo computazionale.

(Traccia: si considerino i tre polinomi

$$a(x) = \sum_{i=0}^{2n-2} \alpha_{i+1-n}\, x^i, \quad b(x) = \sum_{i=0}^{n-1} v_{i+1}\, x^i, \quad c(x) = \sum_{i=0}^{3n-3} u_{i+2-n}\, x^i,$$

di gradi rispettivamente $2n-2$, $n-1$ e $3n-3$. È

$$c(x) = a(x)b(x),$$

infatti si ha

$$\begin{bmatrix} u_{2-n} \\ \vdots \\ u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_n \\ u_{n+1} \\ \vdots \\ u_{2n-1} \end{bmatrix} = \begin{bmatrix} \alpha_{1-n} & & & \\ \vdots & \alpha_{1-n} & & \\ \alpha_{-1} & \ddots & \ddots & \\ \alpha_0 & \alpha_{-1} & \ddots & \alpha_{1-n} \\ \alpha_1 & \alpha_0 & \ddots & \vdots \\ \vdots & \alpha_1 & \ddots & \alpha_{-1} \\ \alpha_{n-1} & \ddots & \ddots & \alpha_0 \\ & \alpha_{n-1} & \ddots & \alpha_1 \\ & & \ddots & \vdots \\ & & & \alpha_{n-1} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}.$$

1.9 Sia n potenza di 2; per risolvere il sistema $T_n\mathbf{x} = \mathbf{b}$ di ordine n, dove

$$T_n = \begin{bmatrix} \alpha_0 & & & \\ \alpha_1 & \alpha_0 & & \\ \vdots & \ddots & \ddots & \\ \alpha_{n-1} & \dots & \alpha_1 & \alpha_0 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix},$$

si può usare un metodo basato sulla seguente osservazione: la matrice T_n^{-1} è della forma

$$T_n^{-1} = \begin{bmatrix} t_0 & & & \\ t_1 & t_0 & & \\ \vdots & \ddots & \ddots & \\ t_{n-1} & \dots & t_1 & t_0 \end{bmatrix},$$

ed è quindi definita dalla sua sola prima colonna. Si dimostri che posto

$$T_n = \begin{bmatrix} T_{n/2} & O \\ W_{n/2} & T_{n/2} \end{bmatrix},$$

vale

$$T_n^{-1} = \begin{bmatrix} T_{n/2}^{-1} & O \\ -T_{n/2}^{-1} W_{n/2} T_{n/2}^{-1} & T_{n/2}^{-1} \end{bmatrix},$$

per cui se

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \quad \text{e} \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

si ha

$$\begin{cases} \mathbf{x}_1 = T_{n/2}^{-1}\, \mathbf{b}_1 \\ \mathbf{x}_2 = T_{n/2}^{-1} \left(\mathbf{b}_2 - W_{n/2} \mathbf{x}_1 \right). \end{cases}$$

La risoluzione di un sistema di ordine n è ricondotta allora a risolvere un sistema di ordine $n/2$ (calcolo di $T_{n/2}^{-1}$) e al calcolo di tre prodotti di una matrice del tipo di quella dell'esercizio 1.8 per un vettore. Si valuti il costo computazionale del metodo descritto.

(Traccia: indicato con $I(n)$ il costo computazionale di tale metodo, vale $I(n) = I(\frac{n}{2}) + kn^{\alpha},\ \alpha = \log_2 3$, se si usa il metodo dell'esercizio 1.6, oppure $I(n) = I(\frac{n}{2}) + kn \log_2 n$, se si usano i metodi FFT, si veda il capitolo 5.)

1.10 Si analizzi il costo computazionale dell'algoritmo per il calcolo del quoziente $q(x)$ e del resto $r(x)$ della divisione di due polinomi $a(x)$ e $b(x)$ di grado rispettivamente m ed n $(m \geq n)$, basato sulla seguente osservazione. Poiché

$$a(x) = b(x)q(x) + r(x),$$

per i coefficienti dei polinomi vale la relazione

$$\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \\ a_n \\ \vdots \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} b_0 & & & \\ b_1 & b_0 & & \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & b_0 \\ b_n & & & b_1 \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & b_n \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{m-n} \end{bmatrix} + \begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

in cui la matrice dei coefficienti di $b(x)$ è formata da $m - n + 1$ colonne. Il vettore $[q_0, \ldots, q_{m-n}]^T$ si ricava risolvendo il sistema triangolare (dove $b_i = 0$ per $i < 0$)

$$\begin{bmatrix} b_n & \ldots & b_{2n-m} \\ & \ddots & \vdots \\ & & b_n \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{m-n} \end{bmatrix} = \begin{bmatrix} a_n \\ \vdots \\ a_m \end{bmatrix}$$

e il vettore $[r_0, \ldots, r_{n-1}]^T$, se $n-1 \geq m-n$, si ricava mediante

$$\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ \vdots \\ r_{n-1} \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ a_{n-1} \end{bmatrix} - \begin{bmatrix} b_0 & & \\ \vdots & \ddots & \\ b_{2n-m-1} & \ddots & b_0 \\ \vdots & \ddots & \vdots \\ b_{n-1} & \ldots & b_{2n-m-1} \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{m-n} \end{bmatrix},$$

e, se $n-1 < m-n$, si ricava mediante

$$\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{bmatrix} - \begin{bmatrix} b_0 & & & \\ b_1 & b_0 & & \\ \vdots & \ddots & \ddots & \\ b_{n-1} & \ldots & b_1 & b_0 \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{n-1} \end{bmatrix}.$$

(Traccia: si sfruttino gli esercizi 1.8 e 1.9).

Commento bibliografico

La complessità computazionale degli algoritmi numerici è un argomento di studio che solo recentemente ha avuto un grande sviluppo. Il metodo di Horner per il calcolo dei polinomi è del 1819. Il metodo però era già stato descritto da Ruffini nel 1804, ma sembra che Horner non ne fosse a conoscenza. Un secolo prima, nel 1711, Newton aveva usato uno schema analogo per rappresentare polinomi. Gli studi sulla complessità del calcolo dei polinomi risalgono al 1954 (Ostrowski [11]) e al 1956 (Motzkin [10]).

Grande impulso alla ricerca nel settore della complessità è stato dato dal problema del prodotto di matrici introdotto da Strassen nel 1969 [19] e dal problema più generale del calcolo di insiemi di forme bilineari. La ricerca di algoritmi di risoluzione efficienti (limitazioni superiori alla complessità) e la ricerca di criteri per determinare limitazioni inferiori non banali alla complessità di un problema sono i due principali aspetti della ricerca nel settore della complessità.

L'importanza degli studi in questo settore non è solamente concreta (si pensi al ruolo degli algoritmi FFT nelle applicazioni tecniche e scientifiche), ma anche teorica: strumenti matematici di ogni tipo vengono correntemente utilizzati. Per una rassegna ampia e ben curata, in particolare sotto il profilo teorico si veda [2], [3], [20]. In [20] è contenuta una bibliografia dettagliata e aggiornata sull'argomento. Per gli aspetti più algoritmici si veda [9]. Per una rassegna sul prodotto di matrici si veda [13].

Per quanto riguarda i problemi legati alla propagazione dell'errore, si vedano i libri di Knuth [8] e di Sterbenz [18]. In particolare per gli esempi

concreti e per un'impostazione più orientata all'interazione con il calcolatore e all'uso delle librerie di software numerico si veda [14].

Per quanto riguarda il calcolo parallelo, si veda [5], [21] per i problemi algebrici e [22] per problemi di algebra lineare numerica. Per gli algoritmi probabilistici si veda [7] e [16], per quanto riguarda l'uso degli algoritmi Las Vegas in ambiente di calcolo parallelo si veda [4].

Bibliografia

[1] D. Bini, "On Commutativity and Approximation", *Theoretical Computer Science*, 28, 1984, pp.135-150.

[2] D. Bini, M. Capovani, G. Lotti, F. Romani, *Complessità Numerica*, Boringhieri, Torino, 1981.

[3] A. Borodin, I. Munro, *The Computational Complexity of Algebraic and Numeric Problems*, Elsevier Computer Science Library, New York, 1975.

[4] A. Borodin, J. von zur Gathen, J. Hopcroft, "Fast Parallel Matrix and GCD Computation", *Information and Control*, 52, 1982, pp. 241-256.

[5] B. Codenotti, M. Leoncini, "Fondamenti di calcolo parallelo", Addison-Wesley, Milano, 1990.

[6] L. Csanky, "Fast Parallel Matrix Inversion Algorithm", *SIAM J. Comput.*, 5, 1976, pp. 618-623.

[7] J. M. Hammersley, D. C. Handscomb, *Monte Carlo Methods*, Methuen & Co., Ltd, London, 1964.

[8] D. E. Knuth, *The Art of Computer Programming, vol. 2, Seminumerical Algorithms*, Addison-Wesley, Reading, Mass., 1969.

[9] L. Kronsjö, *Algorithms, their Complexity and Efficiency*, John Wiley & Sons, New York, 1979.

[10] T. S. Motzkin, "Evaluation of Polynomials and Evaluation of Rational Functions", *Bull. Amer. Math. Soc.*, 61, 1955, p. 163.

[11] A. M. Ostrowski, "On two Problems in Abstract Algebra Connected with Horner's Rule", *Studies Presented to R. von Mises*, Academic Press, New York, 1954, pp. 40-48.

[12] V. Y. Pan, "On Methods of Computing Values of Polynomials", *Russian Math. Surveys*, 21, 1966, pp. 105-136.

[13] V. Y. Pan, "How to Multiply Matrices Faster", Lecture Notes in *Computer Science*, 179, Springer-Verlag, Berlin, 1984.

[14] J. R. Rice, *Numerical Methods, Software, and Analysis*. Mc Graw-Hill, New York, 1983.

[15] A. Schönage, V. Strassen, "Schnelle Multiplikation grosser Zahlen", *Computing*, 7, 1971, pp. 281-292.

[16] Y. A. Shreider (cur.), *The Monte Carlo Method*, Pergamon Press, Oxford, 1966.

[17] R. Solovay, V. Strassen, "A Fast Monte-Carlo Test for Primality", *SIAM J. Comput.*, 6, 1977, pp. 84-85.

[18] P. H. Sterbenz, *Floating-Point Computation*. Prentice-Hall, Englewood Cliffs, N. J., 1974.

[19] V. Strassen, "Gaussian Elimination is not Optimal", *Numer. Math.*, 13, 1969, pp. 354-356.

[20] V. Strassen, "Algebraische Berechnungs Komplexität", in *Perspectives in Mathematics*, Birkhäuser Verlag, Basel, 1984.

[21] J. von zur Gathen, "Parallel Arithmetic Computation: A Survey", *Proc. Math. Foundation of Comp. Science, Lecture Notes in Computer Science*, 233, pp. 93-112, Springer-Verlag, Berlin, 1986.

[22] P. Zellini, *Algoritmi paralleli nell'algebra numerica lineare*, Monografia di software matematico n. 11, IAC del C.N.R., Roma, 1983.

Capitolo 2

ANALISI DELL'ERRORE

1. Rappresentazione in base di un numero

L'insieme dei numeri rappresentabili con un calcolatore è finito: con n quantità $x_1, x_2, \ldots, x_n$, che possono assumere i valori binari 0 e 1, il massimo numero di configurazioni diverse ottenibili è 2^n. Ad esempio, con $n = 3$ si hanno le otto configurazioni seguenti:

$$000,\ 001,\ 010,\ 011,\ 100,\ 101,\ 110,\ 111.$$

Un problema fondamentale è quello di associare ad ogni configurazione di cifre binarie un numero reale in modo che l'insieme dei numeri reali rappresentabili sia opportunamente distribuito nell'intervallo della retta reale a cui appartengono i numeri da rappresentare e in modo che l'errore che si commette risulti sufficientemente contenuto.

Per affrontare tale problema è opportuno fare riferimento al sistema di rappresentazione in base dei numeri reali. Sia $\beta \geq 2$ il numero *intero* che si assume come *base della rappresentazione*. Valgono i seguenti teoremi.

2.1 Teorema. *Sia x un numero reale positivo; allora esistono e sono unici il numero intero p e il numero reale y, con*

$$\beta^{-1} \leq y < 1, \tag{1}$$

tali che

$$x = \beta^p y. \tag{2}$$

Dim. Se x deve soddisfare la (2) e y la (1) dovrà essere

$$\beta^{p-1} \leq x < \beta^p. \tag{3}$$

Un intero p che verifica la (3) deve essere tale che

$$p - 1 \leq \log_\beta x < p,$$

e quindi è unico e risulta

$$p = \lfloor \log_\beta x \rfloor + 1,$$

dove con la notazione $\lfloor z \rfloor$ si intende la parte intera di z. Dalla (2) è

$$y = x\beta^{-p},$$

e dall'unicità di p segue l'unicità di y. ∎

2.2 Teorema. *Sia y un numero reale tale che $\beta^{-1} \le y < 1$; allora esiste ed è unica la successione $\{d_i\}_{i=1,2,\dots}$ di numeri interi, $0 \le d_i < \beta$, $d_1 \neq 0$, non definitivamente uguali a $\beta - 1$ (cioè non tutti uguali a $\beta - 1$ da un certo indice in poi), tali che*

$$y = \sum_{i=1}^{\infty} d_i \beta^{-i}. \tag{4}$$

Dim. Si considerino le successioni $\{d_i\}_{i=1,2,\dots}$ e $\{z_i\}_{i=1,2,\dots}$ così definite:

$$\left.\begin{aligned} &z_0 = y, \\ &d_i = \lfloor \beta z_{i-1} \rfloor, \\ &z_i = \beta z_{i-1} - d_i, \end{aligned}\right\} \quad i = 1, 2, \dots \tag{5}$$

Poiché per ipotesi è $\beta^{-1} \le z_0 < 1$, risulta $1 \le d_1 < \beta$. Inoltre per ogni i è $0 \le z_i < 1$ e quindi $0 \le d_i < \beta$. Dalla (5) si ottiene

$$y = z_0 = \beta^{-1}(d_1 + z_1) = \beta^{-1} d_1 + \beta^{-2}(d_2 + z_2) = \dots = \sum_{i=1}^{k} d_i \beta^{-i} + \beta^{-k} z_k. \tag{6}$$

Poiché $0 \le z_k < 1$ e $\beta \ge 2$, è

$$\lim_{k \to \infty} \beta^{-k} z_k = 0,$$

per cui, passando al limite, dalla (6) si ottiene la (4).

Gli interi d_i non possono essere tutti uguali a $\beta - 1$, altrimenti sarebbe

$$y = (\beta - 1) \sum_{i=1}^{\infty} \beta^{-i} = (\beta - 1)\beta^{-1} \sum_{i=0}^{\infty} \beta^{-i} = \frac{(\beta - 1)\beta^{-1}}{1 - \beta^{-1}} = 1.$$

Inoltre, se per assurdo esistesse un indice $k \ge 2$ per cui

$$d_{k-1} \neq \beta - 1 \quad \text{e} \quad d_i = \beta - 1, \quad \text{per ogni } i \ge k,$$

si avrebbe

$$\begin{aligned} y &= \sum_{i=1}^{k-1} d_i \beta^{-i} + \sum_{i=k}^{\infty} d_i \beta^{-i} = \sum_{i=1}^{k-1} d_i \beta^{-i} + (\beta - 1)\beta^{-k} \sum_{i=0}^{\infty} \beta^{-i} \\ &= \sum_{i=1}^{k-1} d_i \beta^{-i} + \beta^{-k+1} = \sum_{i=1}^{k-2} d_i \beta^{-i} + (d_{k-1} + 1)\beta^{-(k-1)}, \end{aligned}$$

e dalla (6) seguirebbe

$$\beta^{-(k-2)} z_{k-2} = y - \sum_{i=1}^{k-2} d_i \beta^{-i} = (d_{k-1} + 1)\beta^{-(k-1)},$$

da cui

$$\beta z_{k-2} = d_{k-1} + 1,$$

che è assurdo perché per la (5) risulterebbe

$$d_{k-1} = \lfloor \beta z_{k-2} \rfloor = d_{k-1} + 1.$$

Per quanto riguarda l'unicità, si supponga che esistano due successioni distinte $\{d_i\}_{i=1,2,\ldots}$, $\{c_i\}_{i=1,2,\ldots}$, i cui elementi non siano definitivamente uguali a $\beta - 1$, per cui

$$\sum_{i=1}^{\infty} d_i \beta^{-i} = \sum_{i=1}^{\infty} c_i \beta^{-i}.$$

Indicato con h il minimo intero per cui $d_h - c_h \neq 0$, si avrebbe

$$0 = \sum_{i=1}^{\infty} (d_i - c_i)\beta^{-i} = \sum_{i=h}^{\infty} (d_i - c_i)\beta^{-i} = (d_h - c_h)\beta^{-h} + \sum_{i=h+1}^{\infty} (d_i - c_i)\beta^{-i}. \quad (7)$$

Si può supporre, senza ledere la generalità, che $d_h > c_h$, cioè $d_h - c_h \geq 1$. Inoltre, poiché $c_i < \beta$, è

$$d_i - c_i \geq -c_i \geq -(\beta - 1),$$

ed avendo escluso il caso in cui $d_i - c_i = -(\beta - 1)$ definitivamente, esiste $k > h$ tale che

$$d_k - c_k \geq -(\beta - 1) + 1.$$

Sostituendo nella (7) risulterebbe

$$\begin{aligned} 0 &\geq \beta^{-h} + \sum_{i=h+1}^{\infty} (d_i - c_i)\beta^{-i} \geq \beta^{-h} - (\beta - 1) \sum_{i=h+1}^{\infty} \beta^{-i} + \beta^{-k} \\ &= \beta^{-h} - \beta^{-h} + \beta^{-k} = \beta^{-k}, \end{aligned}$$

che è assurdo. ∎

2.3 Teorema *(di rappresentazione in base). Sia x un numero reale non nullo; allora esistono e sono unici il numero intero p e la successione $\{d_i\}_{i=1,2,\ldots}$ di numeri interi, $0 \le d_i < \beta$, $d_1 \neq 0$, non definitivamente uguali a $\beta - 1$, tali che*

$$x = sgn(x)\beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}, \tag{8}$$

dove $sgn(x) = 1$ se $x > 0$, $sgn(x) = -1$ se $x < 0$.
Dim. Se $x > 0$, la tesi segue combinando i risultati dei teoremi 2.1 e 2.2. Se $x < 0$, è sufficiente applicare i due teoremi a $|x|$. ∎

2.4 Definizione. La (8) viene detta *rappresentazione in base* del numero x. Le quantità che compaiono nella (8) vengono dette

p	*esponente*,
d_i	*cifre della rappresentazione*,
$\sum_{i=1}^{\infty} d_i \beta^{-i}$	*mantissa*.

Se esiste un indice k tale che $d_k \neq 0$ e $d_i = 0$ per $i > k$, la rappresentazione in base β si dice *finita di lunghezza k*. ∎

La rappresentazione (8) in base β viene indicata con la notazione posizionale

$$x = \pm(.d_1 d_2 \ldots)_\beta \ \beta^p \quad \text{o} \quad x = \pm(.d_1 d_2 \ldots)\beta^p;$$

se l'esponente è nullo e la base β è chiara dal contesto, di solito il numero viene così indicato

$$x = \pm(.d_1 d_2 \ldots).$$

Poiché $d_1 \neq 0$, la rappresentazione (8) è *normalizzata*. La normalizzazione, oltre ad essere necessaria per l'unicità, si rivela vantaggiosa quando si rappresenta solo un numero limitato di cifre. Ad esempio il numero $x = 1/7000$ può essere così rappresentato in base 10:

$x = (.142857142857\ldots)\ 10^{-3}$	(*rappresentazione normalizzata*),
$x = (.000142857142\ldots)$	(*rappresentazione non normalizzata*).

Se si tronca la rappresentazione a 6 cifre, si ottiene nel primo caso il numero $x_1 = (.142857)\ 10^{-3}$ e, nel secondo caso, il numero $x_2 = (.000142)$. È evidente come x_1 fornisca una migliore approssimazione, rispetto a x_2, del valore di x. L'informazione fornita dagli zeri dopo il punto può essere trasferita, con minor spreco di memoria, nell'esponente.

2. Conversione di base

Basandosi sui teoremi precedenti, si può costruire un algoritmo per determinare gli elementi della rappresentazione in base di un numero reale che per semplicità si suppone positivo. Conviene distinguere il caso $0 < x < 1$, dal caso $x > 1$. Nel primo caso, riferendosi alla dimostrazione del teorema 2.2, si può definire il seguente algoritmo, che calcola k cifre della rappresentazione del numero x nella base β.

2.5 Algoritmo (di determinazione delle prime k cifre in base β di un numero x, $0 < x < 1$).

{ determinazione dell'esponente }
$p := 1$;
while $\lfloor x \rfloor = 0$ **do**
begin
    $x := \beta x$;
    $p := p - 1$
end;
{ determinazione delle cifre d_i }
$i := 0$;
repeat
    $i := i + 1$;
    $d_i := \lfloor x \rfloor$;
    $x := \beta(x - d_i)$
until $i = k$;

Se la rappresentazione di x è finita di lunghezza minore di k, le ultime cifre di x risultano nulle. ∎

Si osservi che un numero può avere una rappresentazione finita in una base e una rappresentazione non finita in un'altra base. Ad esempio il numero 1/10, che in base 10 è rappresentato da (.1), in base 2 è periodico di periodo 1100 (si veda l'esempio 2.7).

Negli esempi seguenti il numero alla sinistra del segno "=" è inteso rappresentato in base 10.

2.6 Esempio. Se $\beta = 10$ si ottengono le seguenti rappresentazioni in base.

$$\frac{1}{10} = (.1)_{10}\, 10^0, \quad \frac{1}{200} = (.5)_{10}\, 10^{-2}, \quad -\frac{1}{3} = -(.333\ldots)_{10}\, 10^0 \qquad ∎$$

2.7 Esempio. Se $\beta = 2$ si ottengono le seguenti rappresentazioni in base.

$$\frac{1}{10} = (.110011001100\ldots)_2\, 2^{-3},$$
$$\frac{1}{100} = (.1010001111010111000010100011110101110000\ldots)_2\, 2^{-6}. \qquad ∎$$

2.8 Esempio. Se $\beta = 8$ si ottengono le seguenti rappresentazioni in base.

$$\frac{1}{10} = (.63146314\ldots)_8\ 8^{-1}, \quad \frac{7}{64} = (.7)_8\ 8^{-1}. \qquad \blacksquare$$

2.9 Esempio. Se $\beta = 16$, usando i simboli alfabetici per le cifre successive al 9, A=10, B=11, C=12, D=13, E=14, F=15, si ottengono le seguenti rappresentazioni in base:

$$\frac{1}{10} = (.1999\ldots.)_{16}\ 16^0, \qquad \frac{1}{100} = (.28\mathrm{F}5\mathrm{C}28\mathrm{F}5\mathrm{C}\ldots)_{16}\ 16^{-1},$$
$$\frac{5}{64} = (.14)_{16}\ 16^0, \qquad -\frac{2}{3} = -(.\mathrm{AAAA}\ldots)_{16}\ 16^0. \qquad \blacksquare$$

Se $x > 1$ è intero, per il teorema 2.3 si può scrivere

$$x = \beta^p \sum_{i=1}^{p} d_i \beta^{-i} = \sum_{i=1}^{p} d_i \beta^{p-i} = \beta\Big(\sum_{i=1}^{p-1} d_i \beta^{p-i-1}\Big) + d_p$$
$$= \beta\Big(\beta\Big(\sum_{i=1}^{p-2} d_i \beta^{p-i-2}\Big) + d_{p-1}\Big) + d_p,$$

quindi d_p è il resto della divisione fra gli interi x e β, d_{p-1} è il resto della divisione fra gli interi

$$\sum_{i=1}^{p-1} d_i \beta^{p-i-1} \quad \text{e} \quad \beta,$$

e così via. Si può quindi definire il seguente algoritmo.

2.10 Algoritmo (di conversione di un numero intero positivo).

$i := 0$;
repeat
    $i := i + 1$;
    $c_i := x$ **mod** β;
    $x := x$ **div** β
until $x = 0$;
$p := i$;
for $i := 1$ **to** p **do**
    $d_i := c_{p+1-i}$;

dove con le notazioni x **mod** β e x **div** β si intendono rispettivamente il resto e il quoziente intero della divisione fra gli interi x e β. Si osservi che la rappresentazione di un numero intero è finita qualunque sia la base. $\blacksquare$

Se $x > 1$ non è intero, posto $x = x_1 + x_2$, dove $x_1 = \lfloor x \rfloor$, risulta $0 < x_2 < 1$. Allora il numero intero x_1 può essere convertito con l'algoritmo 2.10 e il numero x_2 può essere convertito con l'algoritmo 2.5, ottenendo le rappresentazioni:

$$x_1 = \beta^p \sum_{i=1}^{p} d_i \beta^{-i},$$

$$x_2 = \beta^q \sum_{i=1}^{\infty} c_i \beta^{-i}.$$

Si ha allora:

$$x = \beta^p \sum_{i=1}^{\infty} e_i \beta^{-i},$$

dove

$$e_i = \begin{cases} d_i & \text{per } i = 1, \ldots, p, \\ 0 & \text{per } i = p+1, \ldots, p-q, \quad \text{se } p+1 \leq p-q, \\ c_{i-p+q} & \text{per } i \geq p-q+1. \end{cases}$$

2.11 Esempio. Sia $x = \dfrac{3270}{7}$. Posto $x = x_1 + x_2$, dove $x_1 = 467$, $x_2 = \dfrac{1}{7}$, utilizzando gli algoritmi 2.5 e 2.10 si ottengono le seguenti rappresentazioni in base 2:

$$467 = (.111010011)_2 \, 2^9,$$

$$\frac{1}{7} = (.100100100\ldots)_2 \, 2^{-2},$$

da cui

$$x = (.111010011001001001\ldots)_2 \, 2^9. \qquad \blacksquare$$

3. Numeri di macchina

Per il teorema 2.3 un qualsiasi numero reale x è rappresentabile nella forma (8). Il calcolatore, essendo una macchina finita, consente la rappresentazione di un numero finito, anche se grande, di numeri reali.

2.12 Definizione. Siano $\beta,\ t,\ m,\ M$, numeri interi tali che $\beta \geq 2$, $t \geq 1$, $m,\ M > 0$. Si definisce insieme dei *numeri di macchina*, con rappresentazione normalizzata, in base β con t cifre significative, l'insieme

$$\mathcal{F}_{(\beta,t,m,M)} = \{0\} \cup \{x \in \mathbf{R} \ : \ x = \operatorname{sgn}(x)\beta^p \sum_{i=1}^{t} d_i \beta^{-i},$$
$$\text{con } 0 \leq d_i < \beta \text{ per } i = 1, 2, \ldots, t,\ d_1 \neq 0,\ -m \leq p \leq M\}. \qquad \blacksquare$$

Si osservi che l'insieme dei numeri di macchina contiene per definizione lo zero, che non è rappresentabile con una mantissa normalizzata, e quindi non è rappresentabile in modo univoco. Di solito lo zero è rappresentato con mantissa nulla ed esponente $-m$.

Un numero di macchina non nullo $x \in \mathcal{F}_{(\beta,t,m,M)}$ viene indicato nel seguito con la notazione

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

2.13 Esempio. Nella figura 2.1 sono riportati i numeri di $\mathcal{F}_{(2,3,1,1)}$.

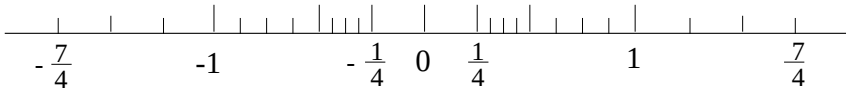


Fig. 2.1 - L'insieme dei numeri di macchina $\mathcal{F}_{(2,3,1,1)}$.

Si noti che i numeri positivi sono compresi fra $\frac{1}{4}$ e $\frac{7}{4}$, quelli negativi fra $-\frac{7}{4}$ e $-\frac{1}{4}$. ∎

Un numero di macchina può essere rappresentato in un calcolatore disponendo di t campi di memoria per la mantissa, ciascuno dei quali può assumere β configurazioni diverse e quindi può memorizzare una cifra d_i, di un campo di memoria che può assumere $m + M + 1$ configurazioni diverse e quindi può memorizzare gli $m + M + 1$ valori diversi dell'esponente, e infine di un campo di memoria che può assumere due configurazioni diverse per il segno.

L'esponente può essere rappresentato come un intero relativo, ma generalmente viene usata la rappresentazione *in traslazione*, in modo che la configurazione nulla corrisponda all'esponente $-m$. L'esponente così rappresentato prende il nome di *caratteristica*. La mantissa viene di solito rappresentata *in modulo e segno*.

La maggior parte dei calcolatori utilizza:

a) aritmetiche in base 2: ogni cifra della mantissa richiede 1 bit per la sua rappresentazione;

b) aritmetiche in base 16: ogni cifra della mantissa richiede 4 bit per la sua rappresentazione.

Nella figura 2.2 è schematizzata una possibile rappresentazione interna su 32 bit (cioè 4 *byte*) del numero $(.d_1 d_2 \dots d_t)\beta^p \in \mathcal{F}_{(\beta,t,m,M)}$ nel caso $\beta = 2$, $t = 24$, $m + M + 1 = 256$. Poiché la rappresentazione è normalizzata, risulta $d_1 = 1$ e quindi non è necessario memorizzare il primo bit della mantissa.

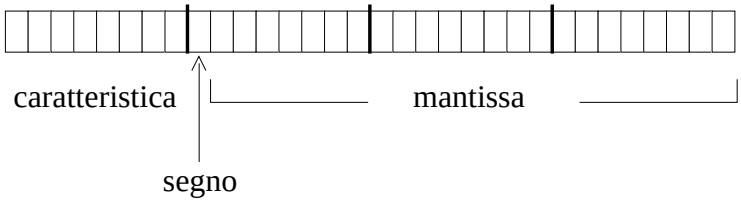


Fig. 2.2 - Rappresentazione interna di un numero di $\mathcal{F}_{(2,24,128,127)}$.

Il primo gruppo di otto bit determina il valore della caratteristica $m+p$ e quindi dell'esponente p, il bit successivo determina il segno (ad esempio positivo se il bit è 0, negativo se il bit è 1), i successivi 23 bit rappresentano le cifre $d_2, d_3, \ldots, d_{24}$ della mantissa. L'insieme dei numeri di macchina con questa configurazione è $\mathcal{F}_{(2,24,128,127)}$ ed è quello utilizzato per la precisione semplice dei calcolatori DEC, Sistema VAX. Per questi calcolatori l'insieme dei numeri in precisione doppia è dato da $\mathcal{F}_{(2,56,128,127)}$. Molti altri calcolatori usano questa stessa rappresentazione interna.

L'aritmetica in base 16 è principalmente adottata nei calcolatori IBM della serie 360/370, ICL serie 2900 e Siemens serie 7500, in cui l'insieme dei numeri in precisione semplice è $\mathcal{F}_{(16,6,64,63)}$ e in precisione doppia è $\mathcal{F}_{(16,14,64,63)}$. La rappresentazione è la seguente:

bit 0 : segno,

bit $1, \ldots, 7$: caratteristica in base 2,

bit $8, \ldots, 31$: mantissa di 6 cifre esadecimali (precisione semplice),

bit $8, \ldots, 63$: mantissa di 14 cifre esadecimali (precisione doppia).

Ogni cifra della mantissa richiede per la sua rappresentazione un semibyte. In precisione semplice occorrono dunque 32 bit (4 byte) per rappresentare un numero di macchina, mentre in precisione doppia occorrono 64 bit (8 byte).

Il minimo numero positivo di $\mathcal{F}_{(\beta,t,m,M)}$ è

$$\omega = (.1)\beta^{-m} = \beta^{-m-1},$$

il massimo è

$$\Omega = \beta^M(\beta - 1)\sum_{i=1}^{t}\beta^{-i} = \beta^M(1 - \beta^{-t}).$$

Poiché con t cifre in base β si possono formare β^t numeri diversi, e fra questi vanno esclusi quelli la cui prima cifra è nulla, l'insieme $\mathcal{F}_{(\beta,t,m,M)}$, oltre allo zero, contiene $(m + M + 1)(\beta^t - \beta^{t-1})$ numeri positivi compresi fra ω e Ω e altrettanti numeri negativi compresi fra $-\Omega$ e $-\omega$.

Se il numero reale e non nullo $x = \pm(d_1 d_2 \ldots)\beta^p$ è tale che $-m \le p \le M$, $d_1 \neq 0$ e $d_i = 0$, per $i > t$, allora $x \in \mathcal{F}_{(\beta,t,m,M)}$. Se x non appartiene a $\mathcal{F}_{(\beta,t,m,M)}$, si pone il problema di associare, in modo adeguato, ad x un numero di macchina $\widetilde{x}$. Nel seguito si suppone per semplicità che x sia un numero positivo (risultati analoghi si ottengono per $x < 0$), e che la base β sia un numero pari (questa ipotesi è in generale verificata).

Si possono presentare i seguenti casi:

a) L'esponente p non appartiene all'intervallo $[-m, M]$. Se $p < -m$, si associa zero ad x e dal sistema di calcolo viene di solito segnalata la situazione di *underflow*. Se $p > M$, x non viene rappresentato e dal sistema di calcolo viene di solito segnalata la situazione di *overflow*.

b) L'esponente p appartiene all'intervallo $[-m, M]$ ma x non appartiene ad $\mathcal{F}_{(\beta,t,m,M)}$ perché le sue cifre d_i, $i > t$, non sono tutte nulle. In questo caso è possibile associare al numero x un numero di macchina $\widetilde{x}$ seguendo due criteri diversi:

troncamento di x alla t-esima cifra

$$\widetilde{x} = \mathrm{trn}(x) = \beta^p \sum_{i=1}^{t} d_i \beta^{-i},$$

arrotondamento di x alla t-esima cifra

$$\widetilde{x} = \mathrm{arr}(x) = \beta^p \ \mathrm{trn}\Big(\sum_{i=1}^{t+1} d_i \beta^{-i} + \frac{1}{2}\beta^{-t}\Big).$$

Quindi

se $\quad d_{t+1} < \dfrac{\beta}{2}, \quad$ allora si ha $\quad \mathrm{arr}(x) = \mathrm{trn}(x),$

se invece $d_{t+1} \ge \dfrac{\beta}{2}, \quad$ allora si ha $\quad \mathrm{arr}(x) = \mathrm{trn}(x) + \beta^{p-t}.$

Si noti che nell'effettuare l'arrotondamento di un numero può verificarsi la situazione di overflow; ad esempio se

$$x = (.d_1 d_2 \ldots d_{t+1})\beta^M, \quad \text{dove} \quad d_i = \beta - 1,\ i = 1, 2, \ldots, t \text{ e } d_{t+1} \ge \frac{\beta}{2},$$

risulta $\mathrm{arr}(x) = (.10\ldots0)\beta^{M+1}$.

2.14 Esempio. In $\mathcal{F}_{(10,3,5,5)}$ se $x = 98273$ si ha

$$\mathrm{trn}(x) = (.982)\ 10^5 \quad \text{e} \quad \mathrm{arr}(x) = (.983)\ 10^5;$$

se $x = 99960$ si ha

$$\mathrm{trn}(x) = (.999)\ 10^5$$

e nell'arrotondamento di x si ha overflow. ∎

4. Errori di rappresentazione

Per valutare l'errore commesso nel rappresentare un numero reale $x \neq 0$, che per semplicità si suppone positivo, con un numero di macchina $\widetilde{x}$, si considerano le seguenti quantità

$$\widetilde{x} - x \quad \textit{errore assoluto},$$

$$\frac{\widetilde{x} - x}{x} \quad \textit{errore relativo}.$$

Si possono dare delle maggiorazioni dei moduli degli errori nella rappresentazione in $\mathcal{F}_{(\beta,t,m,M)}$ dei numeri quando non si verificano situazioni di underflow o di overflow. Per questo si considera l'insieme

$$\mathcal{R} = \{x \in \mathbf{R} \ : \ x = \beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}, \ d_1 \neq 0, \ \omega \leq x \leq \Omega\}.$$

Si esamina per semplicità solo il caso dei numeri positivi, in quanto le maggiorazioni che si ottengono valgono anche per i numeri negativi.

2.15 Teorema. *Per ogni $x \in \mathcal{R}$ risulta*

$$|\mathrm{trn}(x) - x| < \beta^{p-t}, \tag{9}$$

$$|\mathrm{arr}(x) - x| \leq \frac{1}{2}\beta^{p-t}, \tag{10}$$

dove il segno di uguaglianza vale se e solo se $d_{t+1} = \dfrac{\beta}{2}$ e $d_{t+i} = 0, \ i \geq 2$.

Dim. Se x coincide con Ω, allora $\mathrm{trn}(x) - x = \mathrm{arr}(x) - x = 0$. Altrimenti siano a e b due numeri di macchina consecutivi tali che $a \leq x < b$. Allora

$$a = \beta^p \sum_{i=1}^{t} d_i \beta^{-i}, \quad b = \beta^p \Big(\sum_{i=1}^{t} d_i \beta^{-i} + \beta^{-t} \Big),$$

e risulta

$$\mathrm{trn}(x) = a, \qquad x - \mathrm{trn}(x) < b - a = \beta^{p-t},$$

da cui segue la (9). Risulta poi

$$\mathrm{arr}(x) = \begin{cases} a & \text{se } x < \dfrac{a+b}{2}, \\ b & \text{se } x \geq \dfrac{a+b}{2}, \end{cases}$$

e quindi

$$|\mathrm{arr}(x) - x| \leq \frac{b-a}{2} = \frac{1}{2}\beta^{p-t},$$

e l'uguaglianza vale se e solo se

$$x = \frac{a+b}{2}\,, \quad \text{cioè } d_{t+1} = \frac{\beta}{2} \text{ e } d_{t+i} = 0,\ i \geq 2. \qquad \blacksquare$$

Si osservi che in situazione di underflow, ossia se $0 < x < \omega$, il teorema precedente non è applicabile; in tal caso per l'errore assoluto risulta

$$|x - 0| < \beta^{-m-1}.$$

In situazione di overflow, se $x > \Omega$, non è possibile dare una limitazione superiore per l'errore assoluto.

Dal teorema 2.15 è possibile ottenere anche una limitazione superiore del modulo dell'errore relativo che si genera nella rappresentazione di un numero reale x con un numero di macchina $\widetilde{x}$.

2.16 Teorema. *Per ogni $x \in \mathcal{R}$ valgono le limitazioni*

$$\left|\frac{\widetilde{x} - x}{x}\right| < u, \tag{11}$$

$$\left|\frac{\widetilde{x} - x}{\widetilde{x}}\right| < u, \tag{12}$$

dove

$$u = \begin{cases} \beta^{1-t} & \text{se } \widetilde{x} = \mathrm{trn}(x), \\ \dfrac{1}{2}\,\beta^{1-t} & \text{se } \widetilde{x} = \mathrm{arr}(x). \end{cases} \tag{13}$$

Dim. Poiché $d_1 \neq 0$, si ha

$$x \geq \beta^p d_1 \beta^{-1} \geq \beta^{p-1},$$

e quindi

$$\left|\frac{\widetilde{x} - x}{x}\right| \leq \frac{|\widetilde{x} - x|}{\beta^{p-1}}\,.$$

Dalla (9) segue che

$$\left|\frac{\mathrm{trn}(x) - x}{x}\right| < \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{-t+1}.$$

Dalla (10) segue che

$$\left|\frac{\mathrm{arr}(x)-x}{x}\right| \leq \frac{1}{2}\,\frac{\beta^{p-t}}{\beta^{p-1}} = \frac{1}{2}\,\beta^{-t+1}. \tag{14}$$

Il segno di uguaglianza nella (14) potrebbe valere se e solo se

$$d_{t+1} = \frac{\beta}{2} \quad \text{e} \quad d_{t+i} = 0, \text{ per } i \geq 2,$$

ma in tal caso risulterebbe

$$x \geq \beta^p(d_1\beta^{-1} + d_{t+1}\beta^{-t-1}) > \beta^p(d_1\beta^{-1}) \geq \beta^{p-1},$$

per cui nella (14) vale la disuguaglianza stretta, cioè

$$\left|\frac{\mathrm{arr}(x)-x}{x}\right| < \frac{1}{2}\,\beta^{-t+1}.$$

La (12) si ottiene in modo analogo, essendo $\widetilde{x} \geq \beta^{p-1}$ e $\widetilde{x} \neq \beta^{p-1}$ se $d_{t+1} = \dfrac{\beta}{2}$. ∎

2.17 Definizione. La quantità u, che limita superiormente il modulo dell'errore relativo e che è stata definita nel teorema precedente, è detta *precisione di macchina*; l'errore che si commette nel rappresentare il numero reale x col numero di macchina $\widetilde{x}$ è detto *errore di rappresentazione*. La rappresentazione di un numero reale $x \in \mathcal{R}$ con un numero $\widetilde{x} \in \mathcal{F}_{(\beta,t,m,M)}$ è detta *rappresentazione in virgola mobile* con troncamento se $\widetilde{x} = \mathrm{trn}(x)$, con arrotondamento se $\widetilde{x} = \mathrm{arr}(x)$. ∎

Dal teorema 2.16, ponendo

$$\epsilon = \frac{\widetilde{x}-x}{x}\,, \quad \eta = \frac{x-\widetilde{x}}{\widetilde{x}}\,,$$

si ha

$$\widetilde{x} = x(1+\epsilon), \quad \text{con } |\epsilon| < u, \tag{15}$$

$$\widetilde{x} = \frac{x}{1+\eta}\,, \quad \text{con } |\eta| < u, \tag{16}$$

dove u è la precisione di macchina data dalla (13) e corrispondente alla rappresentazione con troncamento o arrotondamento di x.

Dal teorema 2.16 risulta che la rappresentazione con arrotondamento è in generale più accurata. Per questo, anche se richiede un costo maggiore, essa viene di solito utilizzata nella conversione dei dati in ingresso-uscita, normalmente realizzata al livello del software.

Per l'estremo superiore degli errori relativi commessi vale anche il seguente

2.18 Teorema. *Per ogni $x \in \mathcal{R}$ risulta*

$$\sup_{x\in\mathcal{R}} \left| \frac{trn(x) - x}{x} \right| = \frac{u}{1+u}, \quad u = \beta^{-t+1},$$

$$\max_{x\in\mathcal{R}} \left| \frac{arr(x) - x}{x} \right| = \frac{u}{1+u}, \quad u = \frac{1}{2}\,\beta^{-t+1}.$$

Dim. Sia $x = \beta^p y,\ \beta^{-1} \le y < 1$. Per il caso del troncamento si considerino i due insiemi

$$\mathcal{F}_1 = \{v \in \mathbf{R} \ :\ v = \sum_{i=1}^{t} d_i \beta^{-i},\ d_1 \neq 0\},$$

$$\mathcal{F}_2 = \{z \in \mathbf{R} \ :\ z = \sum_{i=t+1}^{\infty} d_i \beta^{-i},\ d_i \neq \beta - 1 \text{ definitivamente}\}.$$

y può essere univocamente decomposto come

$$y = v + z, \quad \text{dove} \quad v \in \mathcal{F}_1,\ z \in \mathcal{F}_2,$$

(è v =trn(y), $z = y-$trn(y)) e per ogni coppia $(v, z) \in \mathcal{F}_1 \times \mathcal{F}_2$ il numero $y = v + z$ è tale che $\beta^{-1} \le y < 1$. Si ha

$$\min_{v\in\mathcal{F}_1} v = \beta^{-1}, \qquad \sup_{z\in\mathcal{F}_2} z = \beta^{-t},$$

$$\sup_{x\in\mathcal{R}} \left| \frac{\mathrm{trn}(x) - x}{x} \right| = \sup_{\beta^{-1}\le y<1} \left| \frac{\mathrm{trn}(y) - y}{y} \right| = \sup_{\substack{v\in\mathcal{F}_1 \\ z\in\mathcal{F}_2}} \frac{z}{v+z}$$

$$= \sup_{z\in\mathcal{F}_2} \Big(\sup_{v\in\mathcal{F}_1} \frac{z}{v+z} \Big) = \sup_{z\in\mathcal{F}_2} \frac{z}{\beta^{-1}+z} = \frac{\beta^{-t}}{\beta^{-1}+\beta^{-t}} = \frac{\beta^{-t+1}}{1+\beta^{-t+1}}.$$

Nel caso dell'arrotondamento, se $d_{t+1} < \dfrac{\beta}{2}$, la dimostrazione può essere condotta in modo analogo al caso precedente. Sia infatti $x = \beta^p y,\ \beta^{-1} \le y < 1$; in questo caso è

$$\mathcal{F}_2 = \{z \in \mathbf{R} \ :\ z = \sum_{i=t+1}^{\infty} d_i \beta^{-i},\ d_{t+1} < \frac{\beta}{2},\ d_i \neq \beta - 1 \text{ definitivamente}\},$$

e

$$\sup_{z\in\mathcal{F}_2} z = \frac{1}{2}\,\beta^{-t}.$$

Risulta

$$\sup_{\substack{x\in\mathcal{R}\\ 0\le d_{t+1}<\beta/2}} \left|\frac{\mathrm{arr}(x)-x}{x}\right| = \sup_{\substack{v\in\mathcal{F}_1\\ z\in\mathcal{F}_2}} \frac{z}{v+z} = \frac{\frac{1}{2}\beta^{-t+1}}{1+\frac{1}{2}\beta^{-t+1}}\,.$$

Se invece $d_{t+1} \ge \dfrac{\beta}{2}$, è

$$x \ge \beta^p\big(\beta^{-1} + \frac{1}{2}\,\beta^{-t}\big),$$

e quindi per la (10)

$$\left|\frac{\mathrm{arr}(x)-x}{x}\right| \le \frac{\frac{1}{2}\beta^{-t}}{\beta^{-1}+\frac{1}{2}\beta^{-t}} = \frac{\frac{1}{2}\beta^{-t+1}}{1+\frac{1}{2}\beta^{-t+1}}\,.$$

Poiché tale valore è effettivamente raggiunto se

$$x = \beta^p(\beta^{-1} + \frac{1}{2}\beta^{-t}),$$

risulta

$$\max_{x\in\mathcal{R}} \left|\frac{\mathrm{arr}(x)-x}{x}\right| = \frac{\frac{1}{2}\beta^{-t+1}}{1+\frac{1}{2}\beta^{-t+1}}\,. \qquad \blacksquare$$

Si osservi che, nella situazione di underflow, l'errore relativo non è limitato dalla precisione di macchina u ed è del 100%. Infatti in questo caso si ha $\left|\dfrac{0-x}{x}\right| = 1$.

In generale, se $\widetilde{x}$ è una approssimazione del numero x con un errore relativo minore di β^{1-t}, si dice che t cifre della rappresentazione nella base β di $\widetilde{x}$ sono *significative*. Se $\widetilde{x}$ è un numero di macchina che approssima il numero reale x, per cui le prime t cifre della rappresentazione in base di x e di $\widetilde{x}$ coincidono, allora l'errore relativo della approssimazione è limitato da β^{1-t}. Non vale però il viceversa, ossia se $\widetilde{x}$ è un numero di macchina che approssima x con errore relativo minore di β^{1-t}, allora non è detto che le prime t cifre coincidano. Ad esempio, in $\mathcal{F}_{(10,5,m,M)}$, il numero .999995 è rappresentato da $\widetilde{x} = \mathrm{arr}(x) = .10000\ 10^1$ e nessuna delle cifre della rappresentazione in base di x e di $\widetilde{x}$ coincide. Comunque anche in questo caso tutte le cifre della rappresentazione sono significative.

5. Operazioni di macchina

La somma di due numeri di macchina $x,\ y \in \mathcal{F}_{(\beta,t,m,M)}$ è un numero che può non appartenere all'insieme $\mathcal{F}_{(\beta,t,m,M)}$. Ad esempio se $\beta = 10$ e $t = 2$,

la somma dei due numeri $x = (.11)\ 10^0$, $y = (.11)\ 10^{-2}$, è $x+y = (.1111)\ 10^0$ che non appartiene ad $\mathcal{F}_{(10,2,m,M)}$. Questo può accadere anche con le altre operazioni aritmetiche.

Si presenta dunque il problema di approssimare, nel modo più conveniente possibile, il risultato di una operazione aritmetica fra due numeri di macchina con un numero di macchina. Occorre perciò individuare le "operazioni aritmetiche" su $\mathcal{F}_{(\beta,t,m,M)}$, ossia definire una *aritmetica di macchina*, detta anche *aritmetica approssimata* o *aritmetica finita*, che meglio approssimi l'*aritmetica esatta* dei numeri reali.

È possibile definire più aritmetiche di macchina. Esse però devono essere tali che, se $\textcircled{\text{op}}$ è l'operazione di macchina che approssima l'operazione esatta op, per tutti i numeri di macchina x e y per cui l'operazione non dia luogo a condizioni di underflow o di overflow, risulti:

$$x \ \textcircled{\text{op}}\ y = (x \text{ op } y)(1+\epsilon), \quad |\epsilon| < u, \tag{17}$$

dove u è la precisione di macchina.

Le operazioni di macchina che soddisfano la (17) sono dette *operazioni in virgola mobile* e l'aritmetica associata è detta *aritmetica in virgola mobile*, l'errore relativo ϵ commesso è detto *errore locale* dell'operazione.

Le operazioni:

$$x \ \textcircled{\text{op}}\ y = \text{trn}(x \text{ op } y) \tag{18}$$

$$x \ \textcircled{\text{op}}\ y = \text{arr}(x \text{ op } y) \tag{19}$$

sono operazioni di macchina che soddisfano la limitazione (17).

Un'aritmetica di macchina con l'arrotondamento (19) approssima l'aritmetica dei numeri reali meglio di un'aritmetica di macchina con il troncamento (18), ma la sua implementazione richiede l'uso di *registri* più lunghi e un maggior tempo di macchina per l'esecuzione delle operazioni. L'utilizzazione del troncamento (18) rappresenta un compromesso fra il tempo di esecuzione e il contenimento dell'errore. In realtà le operazioni di macchina effettivamente realizzate non soddisfano esattamente la (18), anche se soddisfano la (17) con la precisione di macchina relativa al troncamento.

Il seguente algoritmo realizza l'addizione di macchina definita come nella (19). Per snellire la descrizione, si escludono i casi in cui uno o entrambi gli addendi sono nulli, il caso in cui il risultato è nullo, o si genera overflow o underflow. Anche la sottrazione con l'arrotondamento (19) può essere calcolata con questo algoritmo, in quanto

$$\text{arr}(x-y) = \text{arr}(x+(-y)).$$

2.19 Algoritmo (che calcola $z = x \oplus y = \mathrm{arr}(x+y)$, utilizzando un registro di lunghezza $t+2$,

$$\begin{aligned} x &= \mathrm{sgn}(x)\beta^p\ f, \quad \beta^{-1} \le f < 1, \\ y &= \mathrm{sgn}(y)\beta^q\ g, \quad \beta^{-1} \le g < 1, \\ z &= \mathrm{sgn}(z)\beta^r\ h, \quad \beta^{-1} \le h < 1. \end{aligned}$$

Si supponga $|x| \ge |y|$, altrimenti si scambino x e y).

$r\ := p$;
$\mathrm{sgn}(z)\ := \mathrm{sgn}(x)$;
if $p - q \ge t + 2$
  then $h\ := f$
  else begin
    { si traslano a destra le cifre di g di $p-q$ posizioni e si eliminano le cifre successive alla $(t+2)$-esima, aumentando di 1 la $(t+2)$-esima cifra nel caso che x e y abbiano segno opposto e vi siano cifre non nulle dopo la $(t+2)$-esima }
    if $\mathrm{sgn}(x)\mathrm{sgn}(y) > 0$
      then $g'\ := \beta^{-t-2}\lfloor g\beta^{q-p+t+2} \rfloor$
      else $g'\ := \beta^{-t-2}\lceil g\beta^{q-p+t+2} \rceil$;
    { si sommano le mantisse }
    $h\ := |\mathrm{sgn}(x)f + \mathrm{sgn}(y)g'|$;
    { *post-normalizzazione* del risultato ottenuto }
    if $h \ge 1$ **then begin** $h\ := h\beta^{-1}$; $r\ := r+1$ **end** ;
    while $h < \beta^{-1}$ **do begin** $h\ := h\beta$; $r\ := r-1$ **end**;
    { si arrotonda il risultato ottenuto }
    if $\mathrm{sgn}(z) > 0$
      then $h\ := \beta^{-t}\lfloor h\beta^t + \frac{1}{2} \rfloor$
      else $h\ := \beta^{-t}\lceil h\beta^t - \frac{1}{2} \rceil$;
    if $h \ge 1$ **then begin** $h\ := h\beta^{-1}$; $r\ := r+1$ **end**
  end;

Si osservi che nel caso che x e y siano dello stesso segno, per ottenere $\mathrm{arr}(x+y)$ è sufficiente un registro di lunghezza $t+1$. ∎

I seguenti esempi in $\mathcal{F}_{(10,2,m,M)}$ illustrano alcuni casi critici.

2.20 Esempi.
a) Siano $x = (.90)\ 10^0,\ y = (.99)\ 10^{-1}$. Risulta

$$x + y = 0.999 \quad \text{e quindi } \mathrm{arr}(x + y) = (.10)\ 10^1.$$

Dopo l'allineamento e l'eliminazione delle cifre successive alla $(t+2)$-esima, risulta $g' = 0.099$, per cui si ha $h = 0.999$ e infine con la post-normalizzazione si ottiene il risultato

$$z = (.10)\ 10^1.$$

b) Siano $x = (.10)\ 10^0,\ y = -(.99)\ 10^{-4}$. Risulta

$$x + y = 0.099901 \quad \text{e quindi } \mathrm{arr}(x + y) = x.$$

È questo un caso in cui $p - q \geq t + 2$.
c) Siano $x = (.10)\ 10^0,\ y = -(.51)\ 10^{-3}$. Risulta

$$x + y = 0.09949 \quad \text{e quindi } \mathrm{arr}(x + y) = (.99)\ 10^{-1}.$$

Dopo l'allineamento e l'eliminazione delle cifre successive alla $(t+2)$-esima, risulta $g' = 0.0006$, per cui si ha $h = 0.0994$ e infine con la post-normalizzazione si ottiene il risultato

$$z = (.99)\ 10^{-1}.$$

Si osservi che l'avere aggiunto un'unità all'ultima cifra rappresentata di g è stato essenziale ai fini del raggiungimento del risultato richiesto. ∎

L'algoritmo 2.19 utilizza un registro di lunghezza $t + 2$. L'algoritmo seguente realizza invece l'addizione di macchina con un registro di lunghezza $t + 1$ (valgono le stesse limitazioni previste per l'algoritmo 2.19).

2.21 Algoritmo (che calcola $z = x \oplus y$, disponendo di un registro di lunghezza $t + 1$,

$$x = \mathrm{sgn}(x)\beta^p\ f, \quad \beta^{-1} \leq f < 1,$$
$$y = \mathrm{sgn}(y)\beta^q\ g, \quad \beta^{-1} \leq g < 1,$$
$$z = \mathrm{sgn}(z)\beta^r\ h, \quad \beta^{-1} \leq h < 1.$$

Si supponga $|x| \geq |y|$, altrimenti si scambino x e y).

$r := p$;
$\mathrm{sgn}(z) := \mathrm{sgn}(x)$;
if $p - q \geq t + 1$
  then $h := f$

else begin
{ si traslano a destra le cifre di g di $p-q$ posizioni e si eliminano le cifre successive alla $(t+1)$-esima }
$g' := \beta^{-t-1}\lfloor g\beta^{q-p+t+1}\rfloor$
{ si sommano le mantisse }
$h := |\mathrm{sgn}(x)f + \mathrm{sgn}(y)g'|$;
{ *post-normalizzazione* del risultato ottenuto }
if $h \geq 1$ **then begin** $h := h\beta^{-1}$; $r := r+1$ **end** ;
while $h < \beta^{-1}$ **do begin** $h := h\beta$; $r := r-1$ **end**;
{ si tronca il risultato ottenuto }
$h := \beta^{-t}\lfloor h\beta^{t}\rfloor$
end;

∎

Per l'addizione realizzata con l'algoritmo 2.21 è

$$x \oplus y = \mathrm{trn}(x+y),$$

esclusi i casi in cui $xy < 0$ e $p-q > 1$, come risulta anche nei seguenti esempi in $\mathcal{F}_{(10,2,m,M)}$. Si può comunque dimostrare (si veda l'esercizio 2.8) che il risultato ottenuto soddisfa sempre la condizione (17) con la precisione di macchina u relativa al troncamento.

2.22 Esempi. In $\mathcal{F}_{(10,2,m,M)}$
a) siano $x = (.10)\ 10^0,\ y = (.10)\ 10^{-3}$. Risulta

$$x + y = 0.1001 \quad \text{e quindi } \mathrm{trn}(x+y) = x.$$

In questo caso è anche $x \oplus y = x$;
b) siano $x = (.10)\ 10^0,\ y = -(.99)\ 10^{-3}$. Risulta

$$x + y = 0.09901 \quad \text{e quindi } \mathrm{trn}(x+y) = (.99)\ 10^{-1}.$$

In questo caso, in cui $p - q \geq t + 1$, risulta $x \oplus y = x = (.10)\ 10^0$, cioè $x \oplus y \neq \mathrm{trn}(x+y)$;
c) siano $x = (.10)\ 10^0,\ y = -(.51)\ 10^{-2}$. Risulta

$$x + y = 0.0949 \quad \text{e quindi } \mathrm{trn}(x+y) = (.94)\ 10^{-1}.$$

Risulta invece $x \oplus y = (.95)\ 10^{-1}$. ∎

Utilizzando registri di lunghezza $t+2$ o $t+1$ si possono realizzare (si vedano gli esercizi 2.10 e 2.11), con algoritmi analoghi ai 2.19 o 2.21, le operazioni di macchina

$$x \otimes y = \mathrm{arr}(xy) \quad \text{e} \quad x \otimes y = \mathrm{trn}(xy)$$
$$x \oslash y = \mathrm{arr}(x/y) \quad \text{e} \quad x \oslash y = \mathrm{trn}(x/y).$$

Le operazioni di macchina non soddisfano tutte le proprietà algebriche delle operazioni nel campo reale. Per tutti i numeri di macchina per cui l'operazione non dia luogo a condizioni di overflow o di underflow risulta:

$$
\begin{aligned}
&x \oplus y = y \oplus x,\\
&x \ominus y = x \oplus (-y),\\
&-(x \oplus y) = -x \oplus (-y),\\
&x \oplus y = 0 \quad \text{se e solo se} \quad x = -y,\\
&x \oplus 0 = x,\\
&x \otimes y = y \otimes x,\\
&(-x) \otimes y = x \otimes (-y) = -(x \otimes y),\\
&1 \otimes y = y,\\
&x \otimes y = 0 \quad \text{se e solo se} \quad x = 0 \text{ oppure } y = 0,\\
&(-x) \oslash y = x \oslash (-y) = -(x \oslash y),\\
&0 \oslash y = 0,\\
&x \oslash 1 = x,\\
&x \oslash x = 1.
\end{aligned}
$$

Per l'aritmetica di macchina non valgono però le seguenti proprietà:

a) Associatività dell'addizione: $(x + y) + z = x + (y + z)$;

b) Associatività della moltiplicazione: $(xy)z = x(yz)$;

c) Legge di cancellazione: se $xy = yz,\ y \neq 0$ allora $x = z$;

d) Distributività: $x(y + z) = xy + xz$;

e) Semplificazione: $x(y/x) = y$.

2.23 Esempio. In $\mathcal{F}_{(10,2,m,M)}$ con aritmetica con arrotondamento si ha:

a) posto $x = (.11)\ 10^0,\ y = (.13)\ 10^{-1},\ z = (.14)\ 10^{-1}$, risulta

$$
\begin{aligned}
&(x \oplus y) \oplus z = (.12)\ 10^0 \oplus z = (.13)\ 10^0,\\
&x \oplus (y \oplus z) = x \oplus (.27)\ 10^{-1} = (.14)\ 10^0;
\end{aligned}
$$

b) posto $x = (.11)\ 10^1,\ y = (.31)\ 10^1,\ z = (.25)\ 10^1$, risulta

$$
\begin{aligned}
&(x \otimes y) \otimes z = (.34)\ 10^1 \otimes z = (.85)\ 10^1,\\
&x \otimes (y \otimes z) = x \otimes (.78)\ 10^1 = (.86)\ 10^1;
\end{aligned}
$$

c) posto $x = (.51)\ 10^1,\ y = (.22)\ 10^1,\ z = (.52)\ 10^1$, risulta

$$
\begin{aligned}
&x \otimes y = (.11)\ 10^2,\\
&z \otimes y = (.11)\ 10^2,
\end{aligned}
$$

e quindi $x \otimes y = z \otimes y,\ y \neq 0,\ x \neq z$;

d) posto $x = (.11)\ 10^1,\ y = (.23)\ 10^1,\ z = (.24)\ 10^1$, risulta

$$(x \otimes y) \oplus (x \otimes z) = (.25)\ 10^1 \oplus (.26)\ 10^1 = (.51)\ 10^1,$$
$$x \otimes (y \oplus z) = x \otimes (.47)\ 10^1 = (.52)\ 10^1;$$

e) posto $x = (.70)\ 10^1,\ y = (.80)\ 10^1$, si ha

$$x \otimes (y \oslash x) = x \otimes (.11)\ 10^1 = (.77)\ 10^1 \neq y. \qquad \blacksquare$$

6. Calcolo del valore di una funzione

Le funzioni effettivamente calcolabili su un calcolatore sono solo le funzioni razionali, il cui valore è ottenuto mediante un numero finito di operazioni aritmetiche. Le funzioni non razionali, come ad esempio le funzioni trigonometriche, possono essere approssimate da opportune funzioni razionali. Nel calcolo di una generica funzione $f : \mathbf{R}^n \to \mathbf{R}$, il valore effettivamente calcolato in corrispondenza ad un vettore $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ può essere affetto da errori indotti da diversi fenomeni:

a) errore generato dalla rappresentazione dei dati $x_1, x_2, \ldots, x_n$ come numeri di macchina (*errore inerente*);

b) errore generato dal fatto che le operazioni sono effettuate in aritmetica finita (*errore algoritmico*);

c) errore generato, se la funzione $f(\mathbf{x})$ non è razionale, dalla sua approssimazione con una funzione razionale (*errore analitico*).

Se $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ sono le rappresentazioni come numeri di macchina dei dati $x_1, x_2, \ldots, x_n$, e $\psi(\mathbf{x})$ è la funzione effettivamente calcolata, il valore che si ottiene al posto di $f(\mathbf{x})$, è allora $\psi(\widetilde{\mathbf{x}})$, dove $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n)$.

Si esamina prima il caso in cui la funzione $f(\mathbf{x})$ è razionale.

2.24 Teorema. *Siano gli $x_i \neq 0$, $i = 1, 2, \ldots, n$, e sia $f(\mathbf{x})$ una funzione razionale tale che $f(\mathbf{x}) \neq 0$ e $f(\widetilde{\mathbf{x}}) \neq 0$. Indicati con*

$$\epsilon_{tot} = \frac{\psi(\widetilde{\mathbf{x}}) - f(\mathbf{x})}{f(\mathbf{x})}$$

l'errore totale relativo di $\psi(\widetilde{\mathbf{x}})$ rispetto a $f(\mathbf{x})$, con

$$\epsilon_{in} = \frac{f(\widetilde{\mathbf{x}}) - f(\mathbf{x})}{f(\mathbf{x})}$$

l'errore inerente e con

$$\epsilon_{alg} = \frac{\psi(\widetilde{\mathbf{x}}) - f(\widetilde{\mathbf{x}})}{f(\widetilde{\mathbf{x}})}$$

l'errore algoritmico, risulta

$$\epsilon_{tot} = \epsilon_{alg}(1 + \epsilon_{in}) + \epsilon_{in}, \tag{20}$$

e se $f(\mathbf{x})$ è differenziabile in un intorno di $\mathbf{x}$ che contiene tutti i punti del segmento

$$S = \{\mathbf{y} = \alpha\mathbf{x} + (1-\alpha)\widetilde{\mathbf{x}},\ 0 \le \alpha \le 1\},$$

esiste un punto $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n) \in S$, tale che

$$\epsilon_{in} = \sum_{i=1}^{n} \frac{x_i}{f(\mathbf{x})}\ \epsilon_i\ \frac{\partial f(\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\xi}, \tag{21}$$

dove $\epsilon_i = \dfrac{\widetilde{x}_i - x_i}{x_i}$, *con* $|\epsilon_i| < u$, *per* $i = 1, 2, \ldots, n$.

Dim. Si ha

$$\begin{aligned}\epsilon_{tot} = \frac{\psi(\widetilde{\mathbf{x}}) - f(\mathbf{x})}{f(\mathbf{x})} &= \frac{\psi(\widetilde{\mathbf{x}})}{f(\mathbf{x})} - 1 = \frac{\psi(\widetilde{\mathbf{x}})}{f(\widetilde{\mathbf{x}})}\ \frac{f(\widetilde{\mathbf{x}})}{f(\mathbf{x})} - 1 \\ &= (1 + \epsilon_{alg})(1 + \epsilon_{in}) - 1 = \epsilon_{alg}(1 + \epsilon_{in}) + \epsilon_{in}.\end{aligned}$$

Per la formula di Taylor esiste un punto $\boldsymbol{\xi} \in S$, tale che

$$f(\widetilde{\mathbf{x}}) - f(\mathbf{x}) = \sum_{i=1}^{n} (\widetilde{x}_i - x_i)\ \frac{\partial f(\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\xi},$$

da cui segue la (21). ∎

La formula (20) contiene termini lineari e un termine non lineare negli errori ϵ_{alg} e ϵ_{in}. Se tali errori sono dell'ordine della precisione di macchina u, che è un numero molto più piccolo di 1, è ragionevole condurre un'*analisi dell'errore al primo ordine*, cioè trascurare il contributo dei termini non lineari, il cui modulo è minore di ku^2, dove k è una costante indipendente da u. Tale tipo di semplificazione è accettabile se la funzione ψ è una buona approssimazione della funzione f e se l'errore inerente è sufficientemente piccolo.

Se $f(\mathbf{x})$ è differenziabile due volte, per il teorema di Taylor risulta

$$f(\widetilde{\mathbf{x}}) - f(\mathbf{x}) = \sum_{i=1}^{n} (\widetilde{x}_i - x_i)\ \frac{\partial f(\mathbf{x})}{\partial x_i} + O(h^2),$$

dove $h^2 = \sum_{i=1}^{n} (\widetilde{x}_i - x_i)^2 = \sum_{i=1}^{n} x_i^2 \epsilon_i^2$. Allora, indicando con il simbolo $\doteq$ la "uguaglianza" di due quantità che differiscono per termini di ordine superiore al primo negli errori, con un'analisi dell'errore al primo ordine, da (20) e (21) si ottiene

$$\epsilon_{tot} \doteq \epsilon_{alg} + \epsilon_{in} \quad \text{e} \quad \epsilon_{in} \doteq \sum_{i=1}^{n} c_i \epsilon_i, \tag{22}$$

dove

$$c_i = \frac{x_i}{f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial x_i}. \tag{23}$$

I coefficienti c_i sono detti *coefficienti di amplificazione* e danno una misura di quanto influisce l'errore relativo ϵ_i, da cui è affetto il dato x_i, sul risultato: se i coefficienti c_i sono di modulo elevato, anche piccoli errori ϵ_i inducono grossi errori su $f(\mathbf{x})$. In questo caso il problema del calcolo di $f(\mathbf{x})$ è detto problema *mal condizionato*.

Ad esempio, se $f(x)$ è una funzione di una sola variabile, si ha

$$\epsilon_{in} \doteq c_x \epsilon_x, \quad \text{dove} \quad c_x = \frac{x f'(x)}{f(x)}, \quad \epsilon_x = \frac{\widetilde{x} - x}{x}.$$

La funzione $f(x) = x^n$ ha coefficiente di amplificazione $c_x = n$ e quindi è tanto più malcondizionata quanto più grande è n. Ad esempio, posto $x = 1.00001$, l'errore di rappresentazione di x in base 16 con 6 cifre è dell'ordine di 10^{-6}, l'errore da cui è affetto x^n è dell'ordine di 10^{-5} per $n = 10$ e di 10^{-4} per $n = 100$.

L'errore algoritmico ϵ_{alg} è generato dal calcolo della funzione $\psi(\widetilde{\mathbf{x}})$, che è esprimibile come composizione di un numero finito di operazioni di macchina: l'analisi al primo ordine permette di esprimere ϵ_{alg} come combinazione lineare degli errori locali generati dalle singole operazioni. Poiché gli errori locali delle operazioni sono in modulo minori della precisione di macchina u, il modulo dell'errore algoritmico può essere così maggiorato

$$|\epsilon_{alg}| < \theta(x)u + O(u^2),$$

in cui $\theta(x)$ è una funzione indipendente da u e $O(u^2)$ è una funzione di u di ordine maggiore o uguale al secondo. Nel seguito per tale relazione si userà la notazione

$$|\epsilon_{alg}| \dot{<} \theta(x)u.$$

L'algoritmo risulta tanto più *stabile* in corrispondenza ad un insieme di dati, quanto più piccoli sono i valori assunti dalla funzione $\theta(x)$ in corrispondenza a quei dati. Per un problema mal condizionato la distinzione fra algoritmo

stabile e instabile non è molto significativa, in quanto l'errore totale risulta dominato dall'errore inerente.

Se la funzione $f(\mathbf{x})$ non è razionale, è necessario approssimarla con una funzione razionale $g(\mathbf{x})$: tale approssimazione introduce *l'errore analitico*

$$\epsilon_{an} = \frac{g(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} .$$

Detta ancora $\psi(\mathbf{x})$ la funzione effettivamente calcolata al posto della $g(\mathbf{x})$, se $g(\widetilde{\mathbf{x}}) \neq 0$, con procedimento analogo a quello utilizzato nella dimostrazione del teorema 2.24 si ha:

$$\epsilon_{tot} = \frac{\psi(\widetilde{\mathbf{x}}) - f(\mathbf{x})}{f(\mathbf{x})} \doteq \epsilon_{in} + \epsilon_{alg} + \frac{g(\widetilde{\mathbf{x}}) - f(\widetilde{\mathbf{x}})}{f(\widetilde{\mathbf{x}})} ,$$

dove in questo caso è

$$\epsilon_{alg} = \frac{\psi(\widetilde{\mathbf{x}}) - g(\widetilde{\mathbf{x}})}{g(\widetilde{\mathbf{x}})} .$$

Nell'ipotesi che la differenza

$$\left| \epsilon_{an} - \frac{g(\widetilde{\mathbf{x}}) - f(\widetilde{\mathbf{x}})}{f(\widetilde{\mathbf{x}})} \right|$$

sia maggiorata da una funzione almeno quadratica in u, si ha

$$\epsilon_{tot} \doteq \epsilon_{in} + \epsilon_{an} + \epsilon_{alg}. \tag{24}$$

7. Errore nelle operazioni di macchina

Si esamina ora il caso in cui $f(\mathbf{x})$ sia una delle operazioni aritmetiche. Da (22) e (23) si ha

a) se $f(x_1, x_2) = x_1 \pm x_2$,

$$\epsilon_{tot} \doteq \epsilon + c_1\epsilon_1 + c_2\epsilon_2, \quad c_1 = \frac{x_1}{x_1 \pm x_2} , \quad c_2 = \frac{\pm x_2}{x_1 \pm x_2} , \tag{25}$$

dove ϵ è l'errore locale generato dall'operazione di addizione o di sottrazione;

b) se $f(x_1, x_2) = x_1 x_2$,

$$\epsilon_{tot} \doteq \epsilon + c_1\epsilon_1 + c_2\epsilon_2, \quad c_1 = 1, \quad c_2 = 1,$$

dove ϵ è l'errore locale generato dall'operazione di moltiplicazione;

c) se $f(x_1, x_2) = x_1/x_2$,

$$\epsilon_{tot} = \epsilon + c_1\epsilon_1 + c_2\epsilon_2, \quad c_1 = 1, \quad c_2 = -1,$$

dove ϵ è l'errore locale generato dall'operazione di divisione.

L'errore locale ϵ di ogni operazione di macchina è tale che $|\epsilon| < u$.

Dalla (25) si ha che nel caso dell'addizione di due numeri x_1 e x_2 dello stesso segno (o nel caso della sottrazione di due numeri x_1 e x_2 di segno opposto), i coefficienti di amplificazione sono limitati in modulo da 1 e l'errore totale è limitato in modulo da $2u$; nel caso dell'addizione di due numeri x_1 e x_2 di segno opposto (o nel caso della sottrazione di due numeri x_1 e x_2 dello stesso segno) non è possibile dare una maggiorazione dell'errore indipendente da x_1 e x_2. Nel caso dell'addizione, se ad esempio x_1 è quasi uguale a $-x_2$, i coefficienti di amplificazione c_1 e c_2 assumono valori molto grandi in modulo e quindi la limitazione superiore del modulo dell'errore è molto elevata.

In questi casi l'elevato errore che si può ritrovare nel risultato non è generato dalla operazione aritmetica in virgola mobile, infatti l'errore locale dell'operazione ha modulo minore di u, ma è dovuto alla presenza degli errori relativi non nulli ϵ_1 ed ϵ_2 in $\widetilde{x}_1$ e $\widetilde{x}_2$ che sono molto amplificati dalle operazioni aritmetiche di addizione e di sottrazione.

2.25 Esempio. Rappresentando i numeri

$$x_1 = 0.123456, \quad x_2 = -0.123454$$

in $\mathcal{F}_{(10,5,m,M)}$ con arrotondamento, si ottiene rispettivamente

$$\widetilde{x}_1 = \mathrm{arr}(x_1) = (.12346)\ 10^0, \quad \widetilde{x}_2 = \mathrm{arr}(x_2) = -(.12345)\ 10^0.$$

Vale inoltre $\widetilde{x}_1 \oplus \widetilde{x}_2 = (.1)\ 10^{-4}$, e poiché $x_1 + x_2 = (.2)\ 10^{-5}$, nessuna cifra di $\widetilde{x}_1 \oplus \widetilde{x}_2$ è esatta e si ha $|\epsilon_{tot}| = 4$. ∎

L'amplificazione dell'errore generata dalla addizione di numeri di segno opposto o dalla sottrazione di numeri dello stesso segno è detta *fenomeno di cancellazione numerica*; effettuando una sottrazione di numeri dello stesso segno con le prime cifre uguali, queste si cancellano a due a due. Nelle operazioni aritmetiche di moltiplicazione e divisione questo fenomeno non si verifica e l'errore presente negli operandi non viene amplificato.

Il fenomeno della cancellazione numerica è alla base di gran parte dei casi di instabilità numerica illustrati dagli esempi del primo capitolo.

Nel calcolo delle radici α_1 e α_2 dell'equazione $ax^2+bx+c = 0$, mediante le formule

$$\alpha_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad \alpha_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \tag{26}$$

(nell'esempio 1.3 è $a = c = 1$, $b = -q$), si può presentare cancellazione numerica nel calcolo di α_1 se $\sqrt{b^2 - 4ac}$ è quasi uguale a b, o nel calcolo di α_2

se $\sqrt{b^2 - 4ac}$ è quasi uguale a $-b$. Per eliminare il fenomeno di cancellazione numerica, poiché

$$\alpha_1 \alpha_2 = \frac{c}{a},$$

nel primo caso, calcolato α_2 con la (26), si calcola α_1 mediante la formula

$$\alpha_1 = \frac{c}{a\alpha_2},$$

e nel secondo, calcolato α_1 con la (26), si calcola α_2 mediante la formula

$$\alpha_2 = \frac{c}{a\alpha_1}.$$

Nel calcolo della funzione (esempio 1.4)

$$f(x) = x(\sqrt{x^2 + 1} - x)$$

si presenta il fenomeno della cancellazione numerica quando x assume valori grandi, per i quali $\sqrt{x^2 + 1}$ è quasi uguale a x. Invece con la funzione

$$f(x) = \frac{x}{\sqrt{x^2 + 1} + x}$$

non si presenta il fenomeno di cancellazione numerica.

Anche nell'esempio 1.5, utilizzando la serie a segni alterni, si presenta il fenomeno di cancellazione numerica.

8. Uso dei grafi per l'analisi dell'errore

Si esamina ora il problema del calcolo di una funzione razionale $f(\mathbf{x})$ che non sia una singola operazione aritmetica. Il valore $\psi(\widetilde{\mathbf{x}})$ effettivamente calcolato al posto di $f(\mathbf{x})$ è ottenuto sostituendo ai dati x_i i corrispondenti numeri di macchina $\widetilde{x}_i$, $i = 1, 2, \ldots, n$, e alle operazioni aritmetiche le corrispondenti operazioni di macchina, specificando anche l'ordine in cui esse vengono eseguite. La funzione $\psi(\mathbf{x})$, che dipende dall'aritmetica di macchina, è ottenuta implementando un algoritmo costituito da p passi del tipo:

$$z^{(i)} = y_1^{(i)} \text{ op } y_2^{(i)}, \quad \text{per } i = 1, 2, \ldots, p, \tag{27}$$

dove gli operandi $y_1^{(i)}$ e $y_2^{(i)}$ possono essere dati iniziali o risultati delle operazioni ai passi precedenti, cioè

$$y_1^{(i)},\ y_2^{(i)} \in \{x_1, x_2, \ldots, x_n, z^{(1)}, z^{(2)}, \ldots, z^{(i-1)}\},$$

e

$$f(\mathbf{x}) = z^{(p)}.$$

2.26 Esempio. Sia $f(x_1, x_2) = x_1^2 - x_2^2$. Questa funzione può essere calcolata con il seguente algoritmo:

$$\begin{aligned} z^{(1)} &= x_1^2 \\ z^{(2)} &= x_2^2 \\ z^{(3)} &= z^{(1)} - z^{(2)}. \end{aligned}$$

La funzione può essere calcolata anche con il seguente algoritmo:

$$\begin{aligned} v^{(1)} &= x_1 + x_2 \\ v^{(2)} &= x_1 - x_2 \\ v^{(3)} &= v^{(1)} \times v^{(2)}. \end{aligned}$$

∎

Per l'errore $\epsilon_{tot}^{(i)}$ del risultato effettivamente calcolato all'i-esimo passo dalla (22) si ha:

$$\epsilon_{tot}^{(i)} \doteq \epsilon^{(i)} + c_1^{(i)}\epsilon_1^{(i)} + c_2^{(i)}\epsilon_2^{(i)}, \quad i = 1,\ 2, \ldots, p, \tag{28}$$

dove $\epsilon^{(i)}$ è l'errore locale generato dalla i-esima operazione e $\epsilon_1^{(i)}$ e $\epsilon_2^{(i)}$ sono gli errori presenti negli operandi della i-esima operazione e possono essere errori di dati iniziali oppure errori accumulati dai risultati intermedi del calcolo: per $\epsilon_1^{(i)}$ ad esempio,

$$\begin{aligned} \text{se} \quad & y_1^{(i)} = x_j \text{ per qualche } j,\ 1 \le j \le n, \text{ è } \epsilon_1^{(i)} = \epsilon_j, \\ \text{altrimenti se} \quad & y_1^{(i)} = z^{(j)} \text{ per qualche } j,\ 1 \le j \le i-1, \text{ è } \epsilon_1^{(i)} = \epsilon_{tot}^{(j)}. \end{aligned}$$

Inoltre, essendo $z^{(p)} = f(\mathbf{x})$, si ha

$$\epsilon_{tot} = \epsilon_{tot}^{(p)};$$

con applicazioni ripetute della formula (28), si può esprimere l'errore ϵ_{tot} nei termini degli errori dei dati ϵ_i e degli errori locali $\epsilon^{(i)}$ delle operazioni (27).

Per valutare l'errore inerente, che è indipendente dall'algoritmo usato, conviene utilizzare la (22), mentre per valutare l'errore algoritmico, che ovviamente dipende dall'algoritmo, conviene utilizzare un *grafo*, che descrive la sequenza delle operazioni dell'algoritmo ed è costruito nel modo

seguente: i nodi corrispondono ai dati x_i, $i = 1, 2, \ldots, n$, e ai risultati intermedi $z^{(i)}$, $i = 1, 2, \ldots, p$; in corrispondenza ai nodi x_i sono riportati gli errori ϵ_i di rappresentazione dei dati, e ai nodi $z^{(i)}$ sono riportati gli errori locali $\epsilon^{(i)}$ delle operazioni aritmetiche. Ai nodi $z^{(i)}$ arrivano gli archi orientati provenienti dai nodi corrispondenti agli operandi della i-esima operazione; in corrispondenza a ciascun arco è riportato il relativo coefficiente di amplificazione. Nella figura 2.3 è riportato il grafo corrispondente ad un algoritmo composto da tre operazioni per il calcolo di $f(x_1, x_2, x_3, x_4)$.

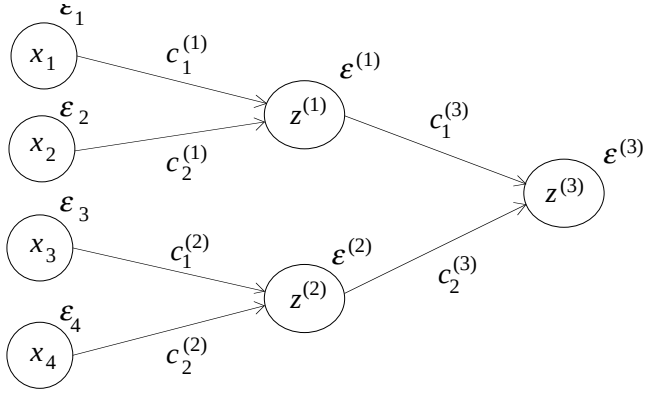


Fig. 2.3 - Grafo associato ad un algoritmo composto da tre operazioni.

L'errore relativo totale si ottiene percorrendo il grafo dall'ultimo nodo verso i nodi iniziali. Ad ogni nodo che si incontra, viene calcolato $\epsilon_{tot}^{(i)}$ dato dalla (28), cioè si sommano l'errore $\epsilon^{(i)}$ che corrisponde al nodo e gli errori precedentemente accumulati nei nodi ad esso collegati, moltiplicati per i coefficienti di amplificazione corrispondenti agli archi percorsi.

Per il grafo rappresentato nella figura 2.3 si ha:

al nodo $z^{(3)}$, $\epsilon_{tot}^{(3)}$ è dato da

$$\epsilon_{tot}^{(3)} \doteq \epsilon^{(3)} + c_1^{(3)}\epsilon_{tot}^{(1)} + c_2^{(3)}\epsilon_{tot}^{(2)},$$

dove $\epsilon_{tot}^{(1)}$ è l'errore accumulato al nodo $z^{(1)}$ e $\epsilon_{tot}^{(2)}$ è l'errore accumulato al nodo $z^{(2)}$;

al nodo $z^{(1)}$, $\epsilon_{tot}^{(1)}$ è dato da

$$\epsilon_{tot}^{(1)} \doteq \epsilon^{(1)} + c_1^{(1)}\epsilon_1 + c_2^{(1)}\epsilon_2,$$

dove ϵ_1 è l'errore di rappresentazione di x_1 e ϵ_2 è l'errore di rappresentazione di x_2;

al nodo $z^{(2)}$, $\epsilon_{tot}^{(2)}$ è dato da

$$\epsilon_{tot}^{(2)} \doteq \epsilon^{(2)} + c_1^{(2)}\epsilon_3 + c_2^{(2)}\epsilon_4,$$

dove ϵ_3 è l'errore di rappresentazione di x_3 e ϵ_4 è l'errore di rappresentazione di x_4.

2.27 Esempio. L'errore inerente della funzione $f(x_1, x_2) = x_1^2 - x_2^2$ è dato da

$$\epsilon_{in} \doteq \frac{2x_1^2}{x_1^2 - x_2^2}\,\epsilon_1 - \frac{2x_2^2}{x_1^2 - x_2^2}\,\epsilon_2.$$

Quindi il problema del calcolo di $f(x_1, x_2)$ è mal condizionato quando il modulo di $x_1^2 - x_2^2$ è piccolo. Si vuole ora determinare l'errore totale che si produce nel calcolo di $f(x_1, x_2)$, con i due algoritmi diversi descritti nell'esempio 2.26. Il grafo corrispondente al primo algoritmo è riportato nella figura 2.4.

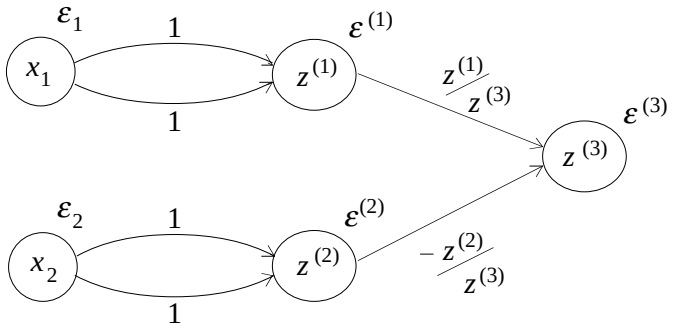


Fig. 2.4 - Grafo per il calcolo dell'errore della funzione $f(x_1, x_2) = x_1^2 - x_2^2$.

Dal grafo si ricava l'errore totale:

$$\begin{aligned}\epsilon_{tot1} &\doteq \epsilon^{(3)} + \frac{z^{(1)}}{z^{(3)}}\,(\epsilon^{(1)} + 2\epsilon_1) - \frac{z^{(2)}}{z^{(3)}}\,(\epsilon^{(2)} + 2\epsilon_2) \\ &\doteq \epsilon^{(3)} + \frac{x_1^2}{x_1^2 - x_2^2}\,\epsilon^{(1)} - \frac{x_2^2}{x_1^2 - x_2^2}\,\epsilon^{(2)} + \frac{2x_1^2}{x_1^2 - x_2^2}\,\epsilon_1 - \frac{2x_2^2}{x_1^2 - x_2^2}\,\epsilon_2,\end{aligned} \tag{29}$$

dove $\epsilon^{(1)}$, $\epsilon^{(2)}$, $\epsilon^{(3)}$ sono gli errori locali delle operazioni.

Il grafo corrispondente al secondo algoritmo è riportato nella figura 2.5. Dal grafo si ricava l'errore totale:

$$\begin{aligned}\epsilon_{tot2} &\doteq \eta^{(3)} + \left[\eta^{(1)} + \frac{x_1}{v^{(1)}}\,\epsilon_1 + \frac{x_2}{v^{(1)}}\,\epsilon_2\right] + \left[\eta^{(2)} + \frac{x_1}{v^{(2)}}\,\epsilon_1 - \frac{x_2}{v^{(2)}}\,\epsilon_2\right] \\ &\doteq \eta^{(1)} + \eta^{(2)} + \eta^{(3)} + \frac{2x_1^2}{x_1^2 - x_2^2}\,\epsilon_1 - \frac{2x_2^2}{x_1^2 - x_2^2}\,\epsilon_2,\end{aligned} \tag{30}$$

dove $\eta^{(1)}$, $\eta^{(2)}$, $\eta^{(3)}$ sono gli errori locali delle operazioni.

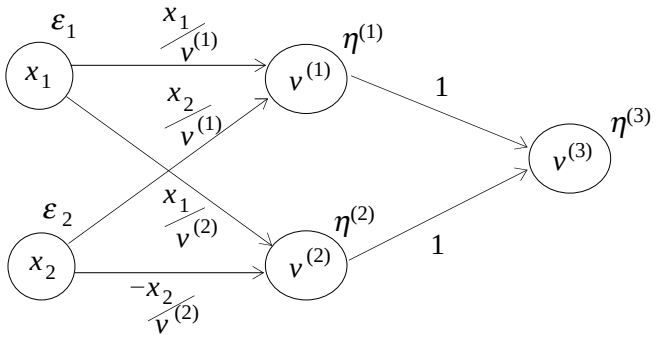


Fig. 2.5 - Grafo per il calcolo dell'errore della funzione $f(x_1, x_2) = (x_1 - x_2)(x_1 + x_2)$.

Dalle (29) e (30) si ricavano gli errori algoritmici dei due algoritmi:

$$\epsilon_{alg1} \doteq \epsilon^{(3)} + \frac{x_1^2}{x_1^2 - x_2^2}\,\epsilon^{(1)} - \frac{x_2^2}{x_1^2 - x_2^2}\,\epsilon^{(2)},$$
$$\epsilon_{alg2} \doteq \eta^{(1)} + \eta^{(2)} + \eta^{(3)},$$

da cui

$$|\epsilon_{alg1}| \dot{<} \Big(1 + \frac{x_1^2 + x_2^2}{|x_1^2 - x_2^2|}\Big)\, u, \quad |\epsilon_{alg2}| \dot{<} 3u.$$

Dal confronto risulta evidente che il secondo algoritmo è più stabile del primo se il modulo di $x_1^2 - x_2^2$ è piccolo. In questo caso però il problema è mal condizionato.

Per illustrare il diverso comportamento dei due algoritmi si è calcolata la funzione $f(x_1, x_2)$ per i numeri di macchina $\widetilde{x}_1$ e $\widetilde{x}_2$ che approssimano i valori $x_1 = 1 + 2i\,10^{-6}$ e $x_2 = 1 + i\,10^{-6}$, $i = 1, \ldots, 500$. L'errore relativo effettivamente generato nel calcolo di $(\widetilde{x}_1 + \widetilde{x}_2)(\widetilde{x}_1 - \widetilde{x}_2)$ risulta sempre inferiore a 10^{-5}, mentre quello generato nel calcolo di $\widetilde{x}_1^2 - \widetilde{x}_2^2$ arriva fino a $0.73\ 10^{-3}$, e quando $i = 400, \ldots, 500$ è sempre superiore a $0.5\ 10^{-3}$. ∎

L'analisi dell'errore condotta con il grafo permette di valutare sia l'errore algoritmico che l'errore inerente; poiché però per l'errore inerente conviene utilizzare la (22), si può semplificare l'analisi utilizzando il grafo per il calcolo del solo errore algoritmico, assumendo nulli gli errori di rappresentazione dei dati, come sarà fatto negli esempi che seguono.

2.28 Esempio. Particolarmente importante è lo studio della propagazione dell'errore nel calcolo della funzione

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i.$$

Sia $f(\mathbf{x}) \neq 0$. Dalla (21) risulta che

$$\epsilon_{in} = \frac{1}{f(\mathbf{x})} \sum_{i=1}^{n} x_i \epsilon_i.$$

In generale non si possono dare limitazioni dell'errore inerente che non dipendano dai dati x_i. Però se essi sono tutti dello stesso segno risulta

$$|\epsilon_{in}| \leq \sum_{i=1}^{n} \frac{|x_i|}{|f(\mathbf{x})|} \max_{i=1,\ldots,n} |\epsilon_i| \leq \max_{i=1,\ldots,n} |\epsilon_i|,$$

ed essendo $|\epsilon_i| < u$, per $i = 1, \ldots, n$, si ha

$$|\epsilon_{in}| < u.$$

Si consideri ora l'algoritmo che calcola le somme parziali così definite:

$$\begin{aligned} &z^{(1)} = x_1 + x_2, \\ &z^{(i)} = z^{(i-1)} + x_{i+1}, \quad i = 2, \ldots, n-1, \\ &f(\mathbf{x}) = z^{(n-1)}, \end{aligned}$$

e il corrispondente grafo che per il caso $n = 4$ è riportato nella figura 2.6.

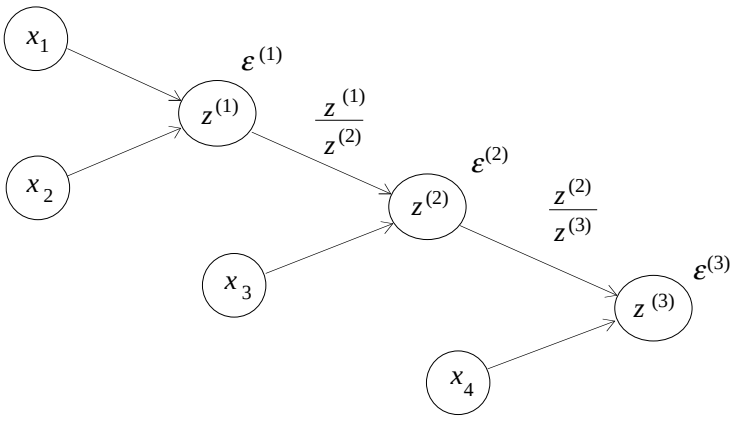


Fig. 2.6 - Grafo del calcolo della funzione $f(\mathbf{x}) = \sum_{i=1}^{4} x_i$.

Poiché l'errore inerente è già stato determinato, nel grafo non sono state riportate le indicazioni relative agli errori di rappresentazione degli x_i e ai corrispondenti coefficienti di amplificazione. Per l'errore algoritmico si ottiene dal grafo:

$$\epsilon_{alg1} \doteq \frac{1}{f(\mathbf{x})} \sum_{i=1}^{n-1} \Big(\sum_{j=1}^{i+1} x_j \Big) \epsilon^{(i)}.$$

In generale, anche per l'errore algoritmico non è possibile dare limitazioni che non dipendano dai dati. Se però essi sono tutti dello stesso segno, poiché risulta

$$\left| \sum_{j=1}^{i+1} x_j \right| \leq |f(\mathbf{x})|, \quad \text{per } i = 1, 2, \ldots, n-1,$$

vale la limitazione

$$|\epsilon_{alg1}| \dot{<} (n-1)u. \tag{31}$$

Se prima di eseguire la somma si riordinano gli addendi, che si suppongono ancora dello stesso segno, in ordine di modulo non decrescente, poiché la successione delle medie aritmetiche è ancora non decrescente in modulo, risulta

$$\frac{1}{i} \left| \sum_{j=1}^{i} x_j \right| \leq \frac{1}{i+1} \left| \sum_{j=1}^{i+1} x_j \right| \leq \frac{1}{n} |f(\mathbf{x})|$$

e quindi

$$|\epsilon_{alg1}| \doteq \frac{1}{|f(\mathbf{x})|} \sum_{i=1}^{n-1} \left| \sum_{j=1}^{i+1} x_j \right| |\epsilon^{(i)}| < \frac{1}{|f(\mathbf{x})|} \sum_{i=1}^{n-1} \frac{(i+1)|f(\mathbf{x})|}{n} \, u < \frac{n+1}{2} \, u. \tag{32}$$

Quest'ultima limitazione migliora di un fattore $\frac{1}{2}$ la limitazione (31).

Una limitazione dell'errore ancora migliore si può ottenere, sempre nel caso che i dati siano dello stesso segno, ricorrendo all'algoritmo seguente (*algoritmo di addizione in parallelo*), descritto per semplicità per $n = 2^p$, con p intero positivo: si calcolino i valori $v_j^{(i)}$, $j = 1, \ldots, n/2^i$, $i = 0, \ldots, p$, così definiti:

$$\begin{aligned}
&v_j^{(0)} = x_j, \quad j = 1, \ldots, n; \\
&v_j^{(i)} = v_{2j-1}^{(i-1)} + v_{2j}^{(i-1)}, \quad j = 1, \ldots, \frac{n}{2^i}, \quad i = 1, \ldots, p, \\
&f(\mathbf{x}) = v_1^{(p)}.
\end{aligned}$$

Nel caso $p = 3$, questo algoritmo è descritto dal grafo riportato nella figura 2.7.

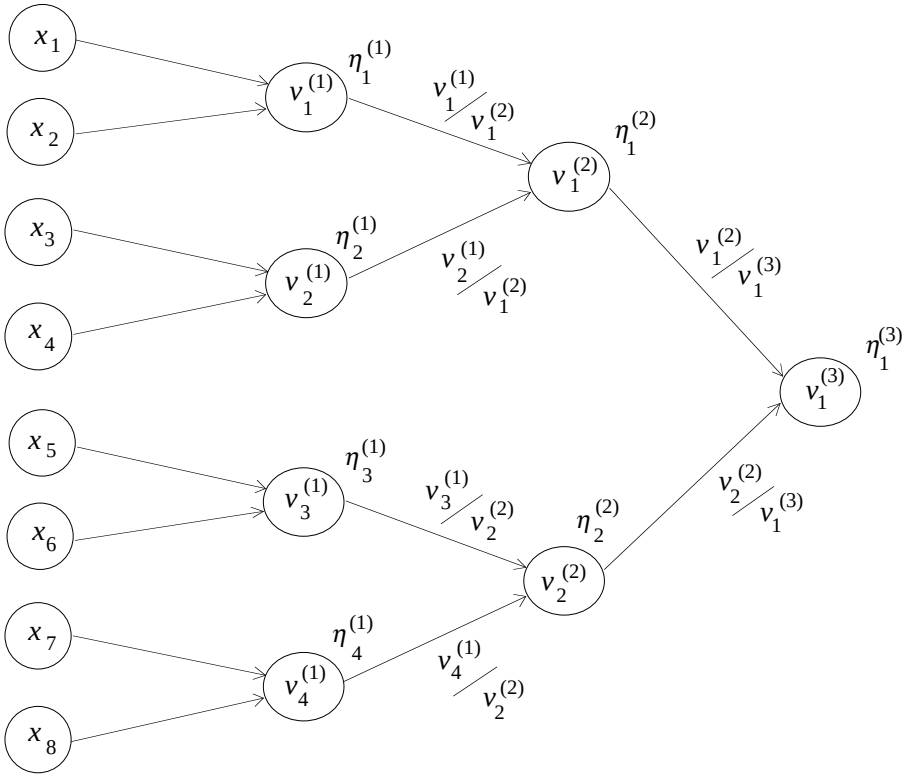


Fig. 2.7 - Grafo del calcolo della funzione $\sum_{i=1}^{8} x_i$.

L'errore algoritmico è

$$\epsilon_{alg2} \doteq \frac{1}{f(\mathbf{x})} \sum_{i=1}^{p} \sum_{j=1}^{n/2^i} v_j^{(i)} \eta_j^{(i)},$$

e poiché

$$\sum_{j=1}^{n/2^i} |v_j^{(i)}| = |f(\mathbf{x})|, \quad \text{per } i = 1, 2, \ldots, p,$$

si ottiene

$$|\epsilon_{alg2}| \dot{<} \max_{i,j} |\eta_j^{(i)}| \sum_{i=1}^{p} \sum_{j=1}^{n/2^i} \frac{|v_j^{(i)}|}{|f(\mathbf{x})|} \leq \max_{i,j} |\eta_j^{(i)}| p < u \log_2 n. \tag{33}$$

Dal confronto di (33) con (31) e (32) risulta che, nel caso di numeri dello stesso segno, l'algoritmo di addizione in parallelo è più stabile, come risulta anche nel caso seguente.

Gli algoritmi precedenti sono utilizzati per calcolare in la somma degli n numeri $x_i = \text{arr}(i^{-3/2})$ per $n = 8192 = 2^{13}$. Indicate con s_1, s_2 e s_3 le somme effettivamente calcolate, prima seguendo l'ordine naturale $i = 1, 2, \ldots, n$, poi seguendo l'ordine inverso $i = n, n-1, \ldots, 1$, cioè disponendo gli addendi in ordine crescente, e infine con l'algoritmo in parallelo, e con ϵ_j gli errori relativi di s_j rispetto ad s per $j = 1, 2, 3$, risulta:

$$\begin{aligned}\epsilon_1 &= 0.1553448\ 10^{-2}\\ \epsilon_2 &= 0.1079697\ 10^{-4}\\ \epsilon_3 &= 0.2328962\ 10^{-5}.\end{aligned}$$

∎

2.29 Esempio. Si studia la propagazione dell'errore nel calcolo del polinomio

$$p(x) = \sum_{i=0}^{n} a_i x^i$$

con i due algoritmi descritti nell'esempio 1.10, nell'ipotesi che i coefficienti a_i siano numeri di macchina. L'errore inerente è dato da

$$\epsilon_{in} \doteq \frac{xp'(x)}{p(x)}\ \epsilon_x,$$

e non risulta limitabile negli intorni delle radici non nulle del polinomio. Per lo studio dell'errore algoritmico, si considera prima il caso in cui il calcolo è fatto per mezzo delle potenze x^i, mediante l'algoritmo

$$\begin{aligned}&y_0 = 1, \quad p_0 = a_0,\\ &y_i = y_{i-1}x, \quad p_i = a_i y_i + p_{i-1}, \quad i = 1, 2, \ldots, n,\\ &p(x) = p_n.\end{aligned}$$

Gli errori algoritmici η_i delle potenze y_i, come risulta dal grafo riportato nella figura 2.8, sono

$$\eta_i = \eta^{(i)} + \eta_{i-1}, \quad i = 2, \ldots, n,$$

in cui $\eta^{(i)}$ sono gli errori locali delle moltiplicazioni, e poiché $|\eta_1| < u$, è

$$|\eta_i| \dot{<} i\,u, \quad i = 2, \ldots, n.$$

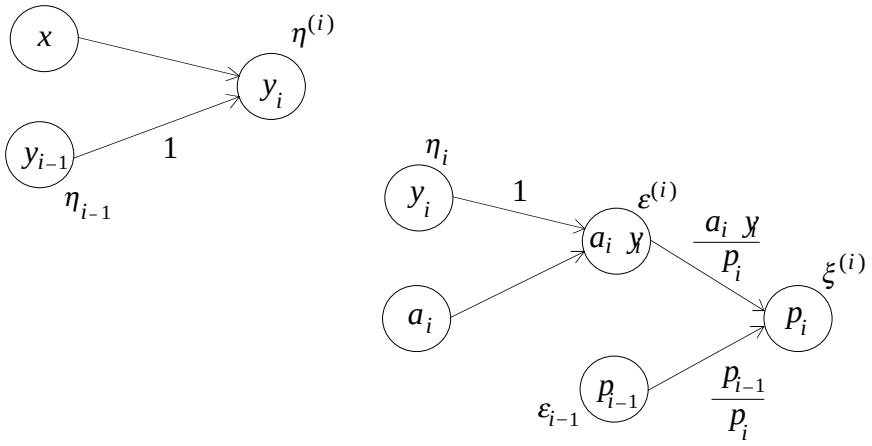


Fig. 2.8 - Calcolo del polinomio con le potenze di x.

Gli errori algoritmici ϵ_i delle somme parziali p_i, come risulta dal grafo riportato nella figura 2.8, sono

$$\epsilon_i \doteq \xi^{(i)} + \frac{p_{i-1}}{p_i}\,\epsilon_{i-1} + \frac{a_i y_i}{p_i}\,(\epsilon^{(i)} + \eta_i), \quad i = 1, \ldots, n, \quad |\epsilon_0| = 0,$$

in cui $\epsilon^{(i)}$ sono gli errori locali delle moltiplicazioni, e $\xi^{(i)}$ gli errori locali delle addizioni. Quindi è

$$|\epsilon_i| \stackrel{\cdot}{<} u + \left|\frac{p_{i-1}}{p_i}\right| |\epsilon_{i-1}| + \left|\frac{a_i y_i}{p_i}\right| (i+1)u, \quad i = 1, \ldots, n.$$

Anche per l'errore algoritmico non è possibile dare delle limitazioni superiori nel caso generale. Se i coefficienti a_i sono tutti positivi e $x > 0$, allora, poiché

$$\left|\frac{p_{i-1}}{p_i}\right| + \left|\frac{a_i y_i}{p_i}\right| = 1,$$

risulta

$$|\epsilon_i| \stackrel{\cdot}{<} (i+2)u, \quad i = 1, \ldots, n,$$

per cui

$$|\epsilon_{alg1}| \stackrel{\cdot}{<} (n+2)u.$$

Si considera poi il caso in cui il calcolo del polinomio viene fatto con il metodo di Ruffini-Horner:

$$\begin{aligned} &p_0 = a_n, \\ &p_i = p_{i-1}x + a_{n-i}, \quad i = 1, 2, \ldots, n, \\ &p(x) = p_n. \end{aligned}$$

Gli errori ϵ_i dei p_i, come risulta dal grafo in figura 2.9, sono

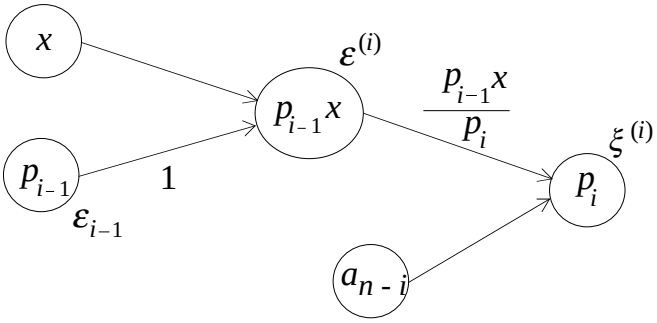


Fig. 2.9 - Calcolo del polinomio con il metodo di Ruffini-Horner.

$$\epsilon_i \doteq \xi^{(i)} + \frac{p_{i-1}x}{p_i}\,(\epsilon^{(i)} + \epsilon_{i-1}), \quad i = 1, \ldots, n, \quad |\epsilon_0| = 0,$$

in cui $\epsilon^{(i)}$ sono gli errori locali delle moltiplicazioni, e $\xi^{(i)}$ gli errori locali delle addizioni. Si ha quindi

$$|\epsilon_i| \;\dot{<}\; u + \alpha(u + |\epsilon_{i-1}|), \quad i = 1, \ldots, n,$$

dove

$$\alpha = \max_{i=1,\ldots,n} \left|\frac{p_{i-1}x}{p_i}\right|.$$

Se i coefficienti sono tutti positivi e $x > 0$, allora $\alpha \leq 1$ e α è tanto più piccolo quanto più i coefficienti crescono al decrescere del grado. È

$$|\epsilon_{alg2}| \;\dot{<}\; 2u(1 + \alpha + \alpha^2 + \ldots + \alpha^{n-1}).$$

Dal punto di vista dell'errore algoritmico il metodo di Ruffini-Horner è da preferire se il rapporto α è piccolo, in particolare quando il grado n è elevato. Infatti, se $\alpha < \frac{1}{2}$ è

$$|\epsilon_{alg2}| \;\dot{<}\; 4u,$$

e quindi la limitazione non dipende da n. Però per valori più grandi di α la limitazione può essere più elevata di quella dell'altro metodo, come risulta nel caso $\alpha = 1$ in cui

$$|\epsilon_{alg2}| \;\dot{<}\; 2nu.$$

Per illustrare il diverso comportamento dei due algoritmi, sono stati calcolati i valori del polinomio a coefficienti positivi decrescenti

$$p(x) = \sum_{i=0}^{n} \frac{x^i}{(n-i)!}\,, \quad n = 15,$$

al variare di x nell'intervallo [1,4]. Nella figura 2.10 sono riportati gli errori relativi effettivamente prodotti nel calcolo del polinomio fatto con le potenze di x (grafico con i pallini) e con il metodo di Ruffini-Horner (grafico con i quadratini neri). I due metodi in questo caso generano errori confrontabili.

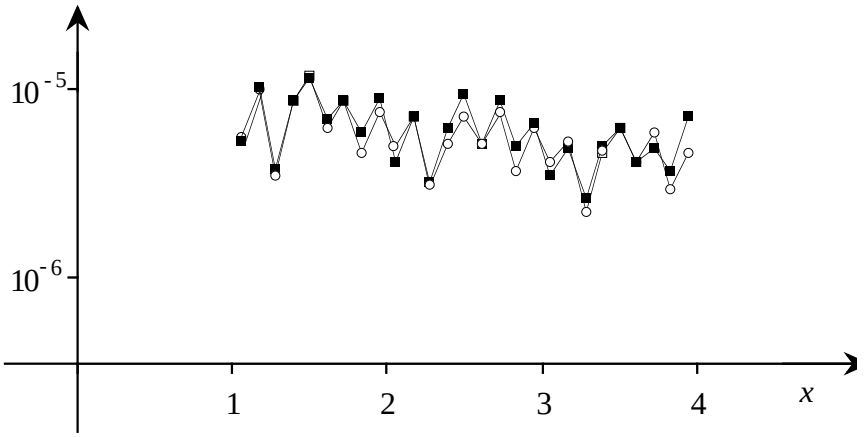


Fig. 2.10 - Errori generati nel calcolo di un polinomio con coefficienti positivi decrescenti.

Il confronto è stato fatto anche nel caso del polinomio a coefficienti positivi crescenti

$$p(x) = \sum_{i=0}^{n} \frac{x^i}{i!}, \quad n = 15.$$

Gli errori relativi effettivamente prodotti sono riportati nella figura 2.11. In questo caso, che però è importante in quanto si presenta spesso con serie convergenti a coefficienti positivi, troncate all'n-esimo termine, il metodo di Ruffini-Horner genera un errore algoritmico minore.

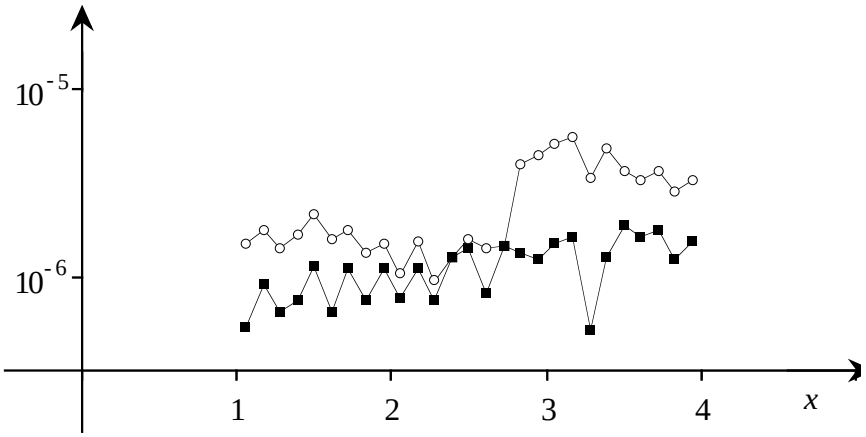


Fig. 2.11 - Errori che si producono nel calcolo di un polinomio con coefficienti positivi crescenti.

In generale il metodo di Ruffini-Horner, oltre a richiedere circa la metà delle moltiplicazioni rispetto al calcolo fatto con le potenze di x, ha un maggiore

intervallo di applicabilità, perché nel calcolo delle potenze x^i si possono produrre errori di overflow e di underflow per valori di x per cui il valore del polinomio è invece ancora rappresentabile. Nel caso del polinomio

$$p(x) = \sum_{i=0}^{n} \frac{x^{2i}}{(2i)!}, \quad n = 25,$$

il calcolo per mezzo delle potenze di x genera errori di underflow per $|x| \leq 0.5$ e di overflow per $|x| \geq 33$, quando il valore di $p(x)$ è solo dell'ordine di 10^{15}. Il metodo di Ruffini-Horner consente il calcolo del polinomio per tutti gli x tali che $10^{-6} < |x| \leq 640$, cioè fino a quando il valore del polinomio è rappresentabile. ∎

9. Errore nelle funzioni non razionali

Se la funzione non razionale $f(x)$ è continua e derivabile un numero sufficiente di volte, essa può essere approssimata con un polinomio (ottenuto ad esempio troncando la formula di Taylor di $f(x)$), o più in generale con una funzione razionale $g(x)$, ricorrendo a metodi che saranno oggetto di studio nei prossimi capitoli.

Nell'approssimazione di una funzione non razionale l'errore totale, in un'approssimazione al primo ordine, è espresso dalla (24), in cui l'errore analitico ϵ_{an} è diverso da zero e in generale tende a diminuire quanto più elevato è il grado n del polinomio (o della funzione razionale) usato, mentre l'errore algoritmico, in generale, tende ad aumentare con n. Indicando con $e_1(n)$ ed $e_2(n)$ le maggiorazioni di $|\epsilon_{an}|$ e $|\epsilon_{alg}|$ al variare di n, il diverso comportamento dei due errori può essere qualitativamente rappresentato come nella figura 2.12.

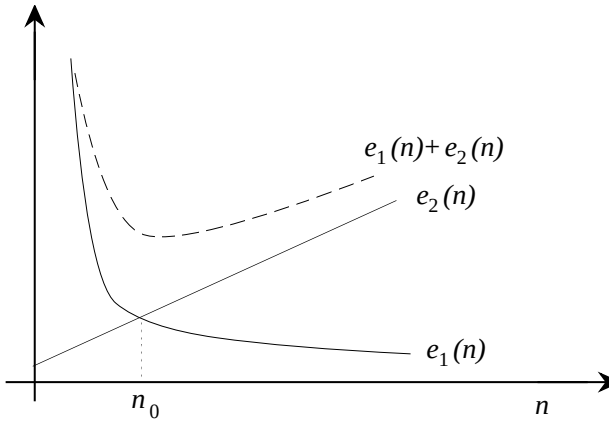


Fig. 2.12 - Comportamento degli errori analitico e algoritmico.

Quindi conviene scegliere un valore di n vicino a n_0, perché per n molto più grande di n_0, ad un maggior volume di calcolo non corrisponde in generale una diminuzione dell'errore effettivamente prodotto.

2.30 Esempio. La funzione logaritmo può essere approssimata con un polinomio troncando la serie di Taylor all'n-esimo termine nel modo seguente:

$$\log(1+x) = \sum_{i=1}^{n} (-1)^{i-1} \frac{x^i}{i} + r(x), \quad \text{per } |x| < 1,$$

dove

$$r(x) = \frac{(-1)^n x^{n+1}}{(n+1)\,(1+\xi)^{n+1}}\,, \quad |\xi| < |x|.$$

L'errore analitico è dato da

$$\epsilon_{an} = \frac{r(x)}{\log(1+x)}\,,$$

e quindi $|\epsilon_{an}|$ è una funzione decrescente al crescere di n e tende a zero per $n \to \infty$.

Per $0 < x < 1$ è $\log(1+x) > x - \dfrac{x^2}{2}$, e quindi

$$|\epsilon_{an}| < \frac{x^n}{(n+1)\big(1-\frac{x}{2}\big)} < \frac{2x^n}{n+1}\,,$$

per $-1 < x < 0$ è $|\log(1+x)| > |x|$, e quindi

$$|\epsilon_{an}| < \frac{|x|^n}{(n+1)(1+x)^{n+1}}\,.$$

Per il calcolo si usa l'algoritmo

$$\begin{aligned} &s_0 = 0, \quad t_0 = -1, \\ &t_i = -x t_{i-1}, \quad s_i = s_{i-1} + \frac{t_i}{i}\,, \quad \text{per } i = 1, \ldots, n. \end{aligned}$$

Utilizzando i risultati dell'esempio 2.28, si ha che l'errore algoritmico è dato da

$$\epsilon_{alg} \doteq \frac{1}{s_n} \sum_{i=1}^{n} \Big(\frac{t_i}{i}\,\epsilon_i + s_i \epsilon^{(i)} \Big),$$

in cui ϵ_i è l'errore da cui è affetto $\dfrac{t_i}{i}$ e $\epsilon^{(i)}$ è l'errore locale della i-esima addizione, e quindi

$$\epsilon_1 = 0, \quad |\epsilon_i| < iu, \quad |\epsilon^{(i)}| < u, \quad \text{per} \quad i = 1, \ldots, n.$$

Se $x < 0$ i termini s_i sono tutti dello stesso segno, quindi

$$|\epsilon_{alg}| \dot{<} \frac{u}{|s_n|} \sum_{i=1}^{n} (|t_i| + |s_i|) = \frac{u}{|s_n|} \left[\sum_{i=1}^{n} |x|^i + \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{|x|^i}{i} \right]$$
$$= \frac{u}{|s_n|} \left[\sum_{i=1}^{n} |x|^i + \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{|x|^i}{i} \right] = \frac{(n+1)u}{|s_n|} \sum_{i=1}^{n} \frac{|x|^i}{i} = (n+1)u.$$

L'errore algoritmico è quindi maggiorato in modulo da una funzione crescente con n. Nella figura 2.13 sono riportati i grafici dei moduli degli errori analitico (con i quadratini neri) e algoritmico (con i pallini) effettivamente prodotti per $x = -0.4$. La scelta più conveniente per n è $n_0 = 12$. ∎

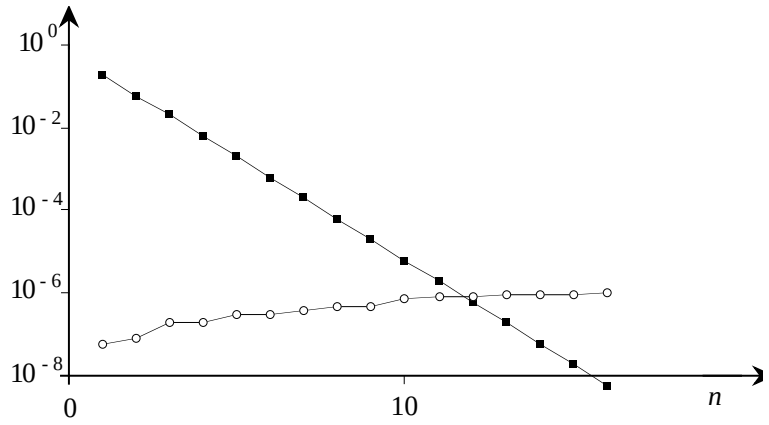


Fig. 2.13 - Errore analitico ed errore algoritmico nel calcolo di log(0.6).

Molte funzioni non razionali fra le più comuni (come le funzioni trigonometriche, esponenziale, logaritmo e radice quadrata) possono essere calcolate utilizzando programmi inclusi in librerie di software, che implementano algoritmi per i quali gli errori algoritmici e analitici corrispondenti sono limitati superiormente in modulo da quantità dell'ordine della precisione di macchina. La valutazione dell'errore totale da cui è affetta una funzione non razionale può ancora essere fatta usando i grafi: ad ogni utilizzazione di una funzione di libreria si fa corrispondere un nodo, a cui arrivano tanti archi quanti sono gli argomenti della funzione.

2.31 Esempio. Supponendo di avere a disposizione una funzione di libreria EXP(x), tale che per ogni numero di macchina x soddisfi alla relazione

$$\mathrm{EXP}(x) = \mathrm{e}^x(1+\epsilon), \quad |\epsilon| = |\epsilon_{alg} + \epsilon_{an}| < u,$$

si calcoli $f(a,b) = \mathrm{e}^{a+b}$. Si possono usare i due algoritmi seguenti:

$$\begin{array}{ll} z^{(1)} = \mathrm{e}^a & v^{(1)} = a+b \\ z^{(2)} = \mathrm{e}^b & \\ z^{(3)} = z^{(1)} \times z^{(2)}, & v^{(2)} = \mathrm{e}^{v^{(1)}}. \end{array}$$

Ad ogni nodo corrispondente ad un valore di e^x arriva un solo arco con il coefficiente di amplificazione

$$c_x = \frac{xf'(x)}{f(x)} = \frac{x\mathrm{e}^x}{\mathrm{e}^x} = x.$$

I grafi corrispondenti ai due algoritmi sono quelli riportati nella figura 2.14.

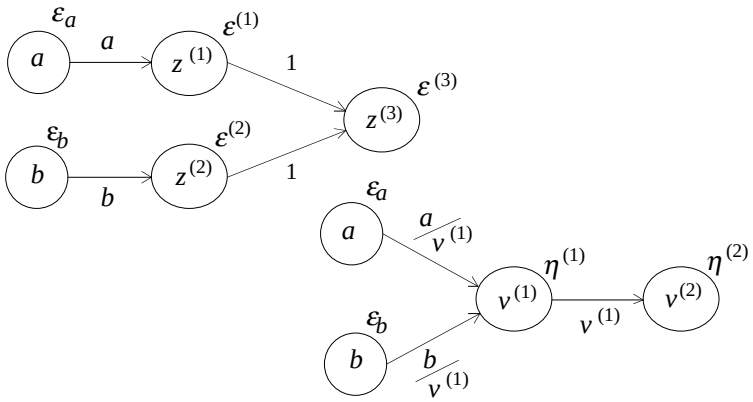


Fig. 2.14 - Grafi per il calcolo della funzione $f(a,b) = \mathrm{e}^{a+b}$.

Per il primo algoritmo si ha

$$\epsilon_{tot1} = \epsilon^{(1)} + \epsilon^{(2)} + \epsilon^{(3)} + a\epsilon_a + b\epsilon_b,$$

e quindi $|\epsilon_{alg1}| < 3u$.

Per il secondo algoritmo si ha

$$\epsilon_{tot2} = \eta^{(2)} + v^{(1)}(\eta^{(1)} + \frac{a}{v^{(1)}}\,\epsilon_a + \frac{b}{v^{(1)}}\,\epsilon_b) = (a+b)\eta^{(1)} + \eta^{(2)} + a\epsilon_a + b\epsilon_b,$$

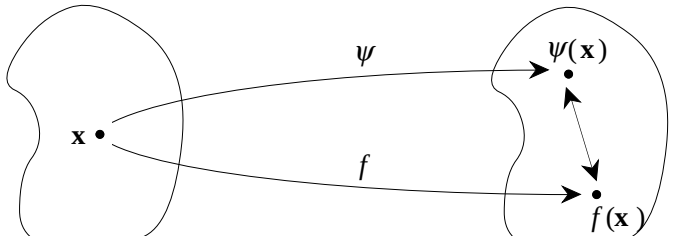
e quindi $|\epsilon_{alg2}| < (1+|a+b|)u$. Il secondo algoritmo risulta quindi più stabile quando $|a+b|$ è piccolo, meno stabile quando $|a+b|$ è elevato. ∎

10. Analisi dell'errore all'indietro

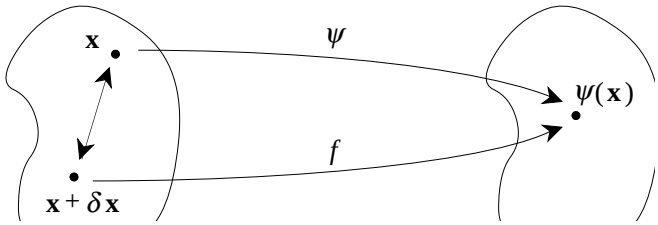
Sia $\mathbf{x} = (x_1, \ldots, x_n)$ un vettore di numeri di macchina e siano f una funzione di $\mathbf{R}^n$ in $\mathbf{R}$ e ψ la funzione di macchina effettivamente calcolata al posto della f. Con l'analisi dell'errore algoritmico svolta nei paragrafi precedenti si possono determinare delle maggiorazioni dell'errore presente nel risultato $\psi(\mathbf{x})$: questa tecnica di analisi è detta *analisi dell'errore in avanti*. È possibile anche condurre un tipo diverso di analisi dell'errore algoritmico, detta *analisi dell'errore all'indietro*, in cui si valuta la perturbazione $\delta\mathbf{x} = (\delta x_1, \ldots, \delta x_n)$ dei dati $\mathbf{x}$ tale che risulti $f(\mathbf{x} + \delta\mathbf{x}) = \psi(\mathbf{x})$, ottenendo maggiorazioni del tipo

$$\frac{|\delta x_i|}{|x_i|} \dot{<} \theta(\mathbf{x})u, \quad i = 1, \ldots, n.$$

Nella figura 2.15 è schematicamente illustrata la differenza fra i due punti di vista.



analisi in avanti



analisi all'indietro

Fig. 2.15 - Analisi dell'errore algoritmico.

Poiché

$$\epsilon_{alg} = \frac{\psi(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} = \frac{f(\mathbf{x} + \delta\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})},$$

con l'analisi all'indietro l'errore algoritmico viene valutato, interpretandolo come l'errore inerente indotto dalla variazione relativa $\delta\mathbf{x}$ nei dati, cioè

$$\epsilon_{alg} \doteq \sum_{i=1}^{n} c_i \, \frac{\delta x_i}{x_i}\,, \quad c_i = \frac{x_i}{f(\mathbf{x})} \, \frac{\partial f(\mathbf{x})}{\partial x_i}\,. \tag{34}$$

L'analisi all'indietro è particolarmente utile nello studio della propagazione dell'errore nella risoluzione dei problemi dell'algebra lineare.

2.32 Esempi.
a) Somma di due numeri di macchina x_1 e x_2. Posto $\mathbf{x} = (x_1, x_2)$ e $f(\mathbf{x}) = x_1 + x_2$, risulta

$$\psi(\mathbf{x}) = x_1 \oplus x_2 = (x_1 + x_2)\,(1+\epsilon) = x_1(1+\epsilon) + x_2(1+\epsilon), \quad |\epsilon| < u,$$

e quindi

$$\psi(x) = f(\mathbf{x} + \delta\mathbf{x}), \quad \text{dove } \delta\mathbf{x} = (\epsilon x_1, \epsilon x_2), \quad \left|\frac{\delta x_1}{x_1}\right| = \left|\frac{\delta x_2}{x_2}\right| = |\epsilon| < u.$$

b) Prodotto di due numeri di macchina x_1 e x_2. Posto $\mathbf{x} = (x_1, x_2)$ e $f(\mathbf{x}) = x_1 x_2$, risulta

$$\psi(\mathbf{x}) = x_1 \otimes x_2 = x_1\, x_2\, (1+\epsilon) = x_1(1+\epsilon)\, x_2, \quad |\epsilon| < u,$$

e quindi

$$\psi(x) = f(\mathbf{x} + \delta\mathbf{x}), \quad \text{dove } \delta\mathbf{x} = (\epsilon x_1, 0), \quad \left|\frac{\delta x_1}{x_1}\right| = |\epsilon| < u, \quad \delta x_2 = 0.$$

c) Somma di n numeri di macchina $x_1, \ldots, x_n$. Posto $\mathbf{x} = (x_1, \ldots, x_n)$ e $f(\mathbf{x}) = \sum_{i=1}^{n} x_i$, utilizzando il seguente algoritmo

$$\begin{aligned} &s_1 = x_1, \\ &s_i = s_{i-1} + x_i, \quad i = 2, \ldots, n, \\ &f(\mathbf{x}) = s_n, \end{aligned}$$

si ha

$$\widetilde{s}_i = \widetilde{s}_{i-1} \oplus x_i = (\widetilde{s}_{i-1} + x_i)\,(1+\epsilon_i), \quad |\epsilon_i| < u, \quad \text{per } i = 2, \ldots, n,$$

da cui

$$\psi(\mathbf{x}) = \widetilde{s}_n = \sum_{i=1}^{n} x_i \prod_{j=i}^{n}(1+\epsilon_j), \quad \epsilon_1 = 0.$$

Indicando con

$$x_i \prod_{j=i}^{n}(1+\epsilon_j) = x_i + \delta x_i,$$

risulta

$$\frac{\delta x_i}{x_i} \doteq \sum_{j=i}^{n} \epsilon_j \,,$$

e quindi

$$\psi(\mathbf{x}) = f(\mathbf{x}+\delta\mathbf{x}), \quad \text{dove } \left|\frac{\delta x_i}{x_i}\right| \dot{<} (n-i+1)u.$$

Valutando l'errore algoritmico con la (34), come se fosse un errore inerente indotto dalle variazioni relative $\dfrac{\delta x_i}{x_i}$ dei dati x_i, si ha

$$|\epsilon_{alg}| \doteq \left|\sum_{i=1}^{n} c_i \frac{\delta x_i}{x_i}\right| \dot{<} \sum_{i=1}^{n} \left|\frac{x_i}{f(\mathbf{x})}\right| (n-i+1)u. \tag{35}$$

Sommando i numeri in ordine di modulo non decrescente, nella (35) a valori $|x_i|$ maggiori corrispondono coefficienti $n-i+1$ minori e quindi l'algoritmo di somma in ordine di modulo non decrescente è più stabile di quello in ordine di modulo non crescente (si confronti con l'esempio 2.28).

d) Risoluzione di un sistema lineare con matrice triangolare utilizzando il metodo di sostituzione.

Siano A una matrice non singolare triangolare inferiore di ordine n e $\mathbf{b}$ un vettore di ordine n, aventi come elementi numeri di macchina. Si indica con $\mathbf{x} = \mathbf{f}(A,\mathbf{b})$ la soluzione del sistema $A\mathbf{x} = \mathbf{b}$ e con $\widetilde{\mathbf{x}} = \boldsymbol{\psi}(A,\mathbf{b})$ la soluzione effettivamente calcolata con l'algoritmo

$$\left.\begin{aligned} &x_1 = \frac{b_1}{a_{11}}\,, \\ &s_{i0} = 0, \\ &y_{ik} = a_{ik}x_k, \quad s_{ik} = s_{i,k-1} + y_{ik}, \text{ per } k = 1,\dots,i-1, \\ &x_i = \frac{b_i - s_{i,i-1}}{a_{ii}}\,, \end{aligned}\right\} \quad \begin{array}{c}\text{per} \\ i = 2,\dots,n.\end{array}$$

Per i valori effettivamente calcolati da (15) e (16) si ha

$$\widetilde{x}_1 = \frac{b_1}{a_{11}(1+\delta_1)}\,, \qquad |\delta_1| < u,$$

$$\left.\begin{aligned} \widetilde{y}_{ik} &= a_{ik}\widetilde{x}_k(1+\epsilon_{ik}),\\ \widetilde{s}_{ik} &= (\widetilde{s}_{i,k-1} + \widetilde{y}_{ik})(1+\zeta_{ik}),\\ \widetilde{x}_i &= \frac{b_i - \widetilde{s}_{i,i-1}}{a_{ii}(1+\eta_i)(1+\delta_i)}\,, \end{aligned}\right\} \quad \begin{aligned} &|\delta_i|, |\epsilon_{ik}|, |\zeta_{ik}|, |\eta_i| < u,\\ &\text{per} \quad i = 2, \ldots, n,\\ &\quad k = 1, \ldots, i-1, \end{aligned}$$

in cui $\zeta_{i1} = 0$. Vale quindi

$$\widetilde{s}_{21} = a_{21}\widetilde{x}_1(1+\epsilon_{21}),$$

e ponendo $\gamma_{21} = \epsilon_{21}$, si ha

$$\widetilde{s}_{21} = a_{21}(1+\gamma_{21})\widetilde{x}_1, \quad \text{con } |\gamma_{21}| < u.$$

Per $i = 3, \ldots, n$,

$$\widetilde{s}_{i,i-1} = \sum_{j=1}^{i-1}\left[a_{ij}\widetilde{x}_j(1+\epsilon_{ij})\prod_{r=j}^{i-1}(1+\zeta_{ir})\right] = \sum_{j=1}^{i-1} a_{ij}(1+\gamma_{ij})\widetilde{x}_j,$$

dove

$$\gamma_{ij} \doteq \epsilon_{ij} + \sum_{r=j}^{i-1}\zeta_{ir}, \quad \text{per} \quad j = 1, \ldots, i-1.$$

Quindi

$$|\gamma_{ij}| \dot{<} (i-j+1)u \le nu, \quad \text{per} \quad i = 2, \ldots, n,\ j = 1, \ldots, i-1.$$

Si ha poi

$$\widetilde{x}_1 = \frac{b_1}{a_{11}(1+\gamma_{11})}$$

e

$$\widetilde{x}_i = \frac{b_i - \widetilde{s}_{i,i-1}}{a_{ii}(1+\gamma_{ii})}\,, \quad \gamma_{ii} \doteq \eta_i + \delta_i, \quad \text{per} \quad i = 2, \ldots, n,$$

e quindi

$$|\gamma_{11}| < u \quad \text{e} \quad |\gamma_{ii}| \dot{<} 2u, \quad \text{per} \quad i = 2, \ldots, n.$$

In conclusione risulta

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

cioè il vettore $\widetilde{\mathbf{x}}$ effettivamente calcolato è quello che si otterrebbe utilizzando l'algoritmo per risolvere in aritmetica esatta il sistema

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

in cui δA è la matrice triangolare inferiore i cui elementi

$$\delta a_{ij} = a_{ij}\gamma_{ij},$$

verificano la relazione

$$\left|\frac{\delta a_{ij}}{a_{ij}}\right| = |\gamma_{ij}| \dot{<} nu, \quad j = 1, \ldots, i, \ i = 1, \ldots, n. \qquad \blacksquare$$

11. Analisi statistica dell'errore

Nelle forme di analisi dell'errore che si sono considerate finora si sono determinate delle maggiorazioni dei moduli degli errori che sono pessimistiche, in quanto potrebbero essere raggiunte solo se ogni singolo errore di rappresentazione o errore locale avesse il massimo modulo possibile, situazione questa, che è statisticamente molto improbabile. Un'analisi dell'errore più realistica può essere fatta considerando ogni singolo errore, locale o di rappresentazione, come una variabile casuale e l'errore inerente o l'errore algoritmico come funzione di tali variabili casuali.

In un'analisi dell'errore al primo ordine tale funzione è del tipo

$$\delta(\epsilon_1, \ldots, \epsilon_m) = d_1\epsilon_1 + \ldots + d_m\epsilon_m,$$

dove $\epsilon_1, \ldots, \epsilon_m$ sono variabili casuali e $d_1, \ldots, d_m$ sono coefficienti che possono dipendere dalla funzione $f(x_1, \ldots, x_n)$ che si sta calcolando (come nel caso dell'errore inerente, in cui i d_i sono i coefficienti di amplificazione) o anche dall'algoritmo utilizzato (come nel caso dell'errore algoritmico).

Obiettivo dell'analisi statistica dell'errore è valutare la probabilità che l'errore sia inferiore ad un valore prefissato: ciò può essere fatto valutando la funzione di distribuzione della funzione errore, considerata come una variabile casuale, o, se ciò non è possibile, calcolandone la media e la varianza. Strumenti fondamentali per l'analisi statistica sono i seguenti teoremi.

2.33 Teorema. *Siano $\epsilon_1, \ldots, \epsilon_m$ variabili casuali indipendenti, di media $\mu_1, \ldots, \mu_m$ e varianza $\sigma_1^2, \ldots, \sigma_m^2$. La variabile casuale*

$$\delta = \delta(\epsilon_1, \ldots, \epsilon_m) = d_1\epsilon_1 + \ldots + d_m\epsilon_m, \tag{36}$$

in cui $d_1, \ldots, d_m$ sono costanti, ha media

$$\mu(\delta) = d_1\mu_1 + \ldots + d_m\mu_m$$

e varianza

$$\sigma^2(\delta) = d_1^2\sigma_1^2 + \ldots + d_m^2\sigma_m^2.$$ ∎

La conoscenza di $\mu(\delta)$ e $\sigma^2(\delta)$ fornisce informazioni notevoli circa la localizzazione e dispersione dell'errore.

2.34 Teorema *(Disuguaglianza di Chebyshev). Sia δ una variabile casuale di media $\mu(\delta)$ e varianza $\sigma^2(\delta)$ e sia c un numero positivo. Allora la probabilità che*

$$|\delta - \mu(\delta)| > c$$

è minore di $\dfrac{\sigma^2(\delta)}{c^2}$. ∎

2.35 Esempio. Nel caso dell'errore inerente della somma

$$s = \sum_{i=1}^{n} x_i$$

di n numeri x_i tutti positivi si ha (si veda l'esempio 2.28)

$$\epsilon_{in} = \sum_{i=1}^{n} c_i\epsilon_i, \quad c_i = \frac{x_i}{s}.$$

Se le variabili casuali ϵ_i hanno tutte media μ e varianza σ^2, per il teorema 2.34 la probabilità che

$$|\epsilon_i - \mu| > c, \quad c > 0,$$

è minore di $\dfrac{\sigma^2}{c^2}$. Per il teorema 2.33 si ha

$$\mu(\epsilon_{in}) = \sum_{i=1}^{n} \frac{x_i}{s}\,\mu = \mu,$$

$$\sigma^2(\epsilon_{in}) = \sum_{i=1}^{n} \frac{x_i^2}{s^2}\,\sigma^2 < \sigma^2,$$

e quindi la probabilità che

$$|\epsilon_{in} - \mu| > c, \quad c > 0,$$

è ancora minore di $\dfrac{\sigma^2}{c^2}$.

Nel caso dell'errore algoritmico è

$$\epsilon_{alg} = \sum_{i=1}^{n-1} d_i \epsilon^{(i)}, \quad 0 \le d_i \le 1.$$

Se le variabili casuali $\epsilon^{(i)}$ hanno tutte media μ e varianza σ^2, si ha

$$\mu(\epsilon_{alg}) = \sum_{i=1}^{n-1} d_i \mu \le (n-1)\mu,$$
$$\sigma^2(\epsilon_{alg}) = \sum_{i=1}^{n-1} d_i^2 \sigma^2 < (n-1)\sigma^2. \qquad \blacksquare$$

La maggiorazione data nel teorema 2.34 è del tutto generale, perché non vengono fatte ipotesi sul tipo di distribuzione della variabile casuale δ. Per questo motivo la disuguaglianza può risultare pessimistica, come nel caso in cui la distribuzione della variabile casuale è *normale*, cioè quando la *densità di probabilità* è data da

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \qquad \sigma > 0.$$

Il grafico della funzione $p(x)$ nel caso che la media μ sia zero, è riportato nella figura 2.16, dove l'area tratteggiata indica la probabilità che il valore della variabile casuale sia compreso fra -2σ e 2σ.

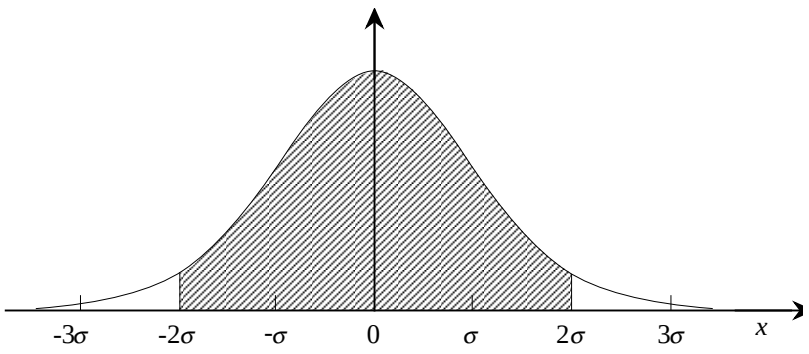


Fig. 2.16 - Densità di probabilità di una distribuzione normale

Definendo la *funzione errore*

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2}\, dt$$

e la funzione *funzione errore complementare*

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x),$$

nel caso che la distribuzione della variabile casuale sia normale, la probabilità che

$$|\delta - \mu(\delta)| > k\sigma(\delta)$$

è data da

$$\pi_k = 1 - \int_{\mu(\delta)-k\sigma(\delta)}^{\mu(\delta)+k\sigma(\delta)} p(x)\ dx = \operatorname{erfc}\Big(\frac{k}{\sqrt{2}}\Big)$$

e, al variare di k, è la seguente

k	π_k
1	0.31731
2	0.04500
3	0.00270
4	0.00006

(37)

D'altra parte la determinazione della specifica funzione di distribuzione della variabile casuale (36), è difficile, a parte i casi più semplici, anche se sono note le funzioni di distribuzione delle variabili casuali ϵ_i, $i = 1, \ldots, m$. Però è possibile dimostrare (teorema *centrale di convergenza*) che per valori grandi di m la funzione di distribuzione di δ, dopo un'opportuna normalizzazione, è approssimata da una funzione che è completamente indipendente dalle distribuzioni delle singole variabili casuali $\epsilon_1, \ldots, \epsilon_m$, ed è data dalla distribuzione normale (si veda [7]).

2.36 Esempio. Quando si opera con arrotondamento si può supporre che gli errori di rappresentazione dei dati e gli errori locali delle operazioni siano distribuiti uniformemente nell'intervallo $\big[-\frac{1}{2}u, \frac{1}{2}u\big]$, cioè che la funzione di densità di probabilità $p(\epsilon)$ sia quella riportata nella figura 2.17.

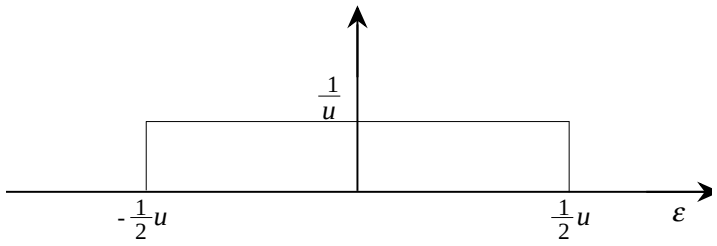


Fig. 2.17 - Densità di probabilità di una distribuzione uniforme

Allora la media degli errori è

$$\mu = \int_{-\infty}^{\infty} \epsilon p(\epsilon)\, d\epsilon = 0$$

e la loro varianza è

$$\sigma^2 = \int_{-\infty}^{\infty} (\epsilon - \mu)^2 p(\epsilon)\, d\epsilon = \frac{1}{12}\, u^2$$

(si veda però l'esercizio 2.43).

Per l'errore algoritmico nel caso della sommatoria si ha (dall'esempio 2.35)

$$\mu(\epsilon_{alg}) = 0, \quad \sigma(\epsilon_{alg}) < \sqrt{n-1}\, \sigma = \sqrt{\frac{n-1}{12}}\, u.$$

Se n è sufficientemente elevato per il teorema del limite centrale si può assumere che l'errore algoritmico abbia una distribuzione normale. In tal caso dalla tabella (37) segue che l'errore algoritmico ha modulo maggiore di $\sigma(\epsilon_{alg})$, $2\sigma(\epsilon_{alg})$, $3\sigma(\epsilon_{alg})$, rispettivamente nel 32% dei casi, 4.5% dei casi, 0.27% dei casi.

Si osservi come nella quasi totalità dei casi l'errore non sia superiore a $3\sigma(\epsilon_{alg}) \leq 0.9\sqrt{n-1}\, u$, mentre con l'analisi diretta dell'errore, svolta nell'esempio 2.28, si è ottenuta la maggiorazione

$$|\epsilon_{alg}| \dot{<} (n-1)u,$$

valida nella totalità dei casi.

Un risultato così preciso non si ottiene invece se si opera con il troncamento: infatti si può assumere che gli errori siano distribuiti uniformemente nell'intervallo $[0, u]$ e quindi $\mu = \frac{1}{2}u$, mentre la varianza è la stessa. Risulta

$$\mu(\epsilon_{alg}) \leq \frac{n-1}{2}\, u. \qquad \blacksquare$$

12. Analisi automatica dell'errore

I limiti all'utilità dell'analisi diretta dell'errore consistono nella pesantezza del procedimento con cui si ricava l'espressione del risultato da calcolare, e quindi anche l'espressione dell'errore, e nel fatto che le maggiorazioni dell'errore così ottenute si rivelano molto spesso pessimistiche. Si otterrebbero maggiorazioni più aderenti alla realtà se si seguissero tutti i passi del procedimento di calcolo, durante l'esecuzione, al fine di individuare,

per ogni risultato intermedio, la migliore limitazione dell'errore propagato, in corrispondenza dei dati effettivi. Un simile modo di analizzare l'errore sarebbe troppo oneroso, per essere affrontato manualmente. Hanno perciò interesse alcune tecniche di analisi automatica, per mezzo delle quali lo studio della propagazione dell'errore viene condotto dal calcolatore durante il calcolo stesso.

a) Una prima tecnica [15] è l'*analisi (automatica) di significatività*, che consiste nell'uso di un'aritmetica con numeri non normalizzati. Così facendo, quando si verifica cancellazione di cifre, il numero di zeri che precedono la cifra più significativa del risultato fornisce ovviamente una stima dell'errore introdotto. È inevitabile che, rinunciando alla normalizzazione, il risultato sia meno accurato, anche se corredato da informazioni sull'errore commesso.

b) Un'altra tecnica, implementata con il nome di *aritmetica "noisy mode"*, consiste nel confrontare il risultato calcolato per mezzo dell'aritmetica usuale con quello che si ottiene modificando le operazioni aritmetiche nel seguente modo: se in un'operazione è richiesta la post-normalizzazione, le cifre inserite da destra valgono $\beta - 1$ anziché 0. La differenza fra i due risultati calcolati fornisce una stima dell'errore. Anche con questa tecnica la stima così prodotta è approssimativa.

c) Una terza tecnica usata è l'*aritmetica degli intervalli*. Essa consiste nel rappresentare ogni numero, invece che con un solo numero di macchina, con una coppia di numeri di macchina, definenti un intervallo che contiene il numero da rappresentare. Le operazioni di macchina sono quindi operazioni su intervalli di numeri di macchina e sono così definite:

$$[x_1, x_2] \textcircled{\text{op}} [y_1, y_2] \;=\; [z_1, z_2],$$

dove per ogni x, y tali che $x_1 \le x \le x_2$ e $y_1 \le y \le y_2$, sia $z_1 \le (x \text{ op } y) \le z_2$.

2.37 Esempio. L'addizione di macchina nell'aritmetica degli intervalli viene realizzata nel modo seguente:

$$[x_1, x_2] \oplus [y_1, y_2] \;=\; [z_1, z_2],$$

dove

$$z_1 = \begin{cases} \text{trn } (x_1 + y_1) & \text{se } x_1 + y_1 \ge 0, \\ \text{trn } (x_1 + y_1) - \beta^{p-t} & \text{se } x_1 + y_1 < 0, \end{cases}$$

$$z_2 = \begin{cases} \text{trn } (x_2 + y_2) + \beta^{p-t} & \text{se } x_2 + y_2 > 0, \\ \text{trn } (x_2 + y_2) & \text{se } x_2 + y_2 \le 0. \end{cases}$$

In modo analogo si possono implementare le altre operazioni di macchina e il calcolo delle funzioni di libreria. ∎

L'aritmetica degli intervalli risente del difetto, del resto presente in ogni analisi rigorosa dell'errore, di supporre che ogni operazione produca un risultato affetto dal massimo errore possibile. Per questo il risultato finale può essere un intervallo molto ampio anche in casi in cui l'errore effettivo sia contenuto. Però a differenza dell'analisi di significatività e dell'aritmetica noisy mode, l'aritmetica degli intervalli garantisce una limitazione corretta dell'errore.

2.38 Esempio. Si deve calcolare la successione (x_i, y_i), definita dalle formule ricorrenti

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix},$$

con $x_0 = 1$ e $y_0 = 0$. Il vettore $[x_{i+1}, y_{i+1}]^T$ risulta ruotato di 45^o rispetto al vettore $[x_i, y_i]^T$. Siano $\widetilde{x}_0 = x_0 + \eta_0$ e $\widetilde{y}_0 = y_0 + \xi_0$, con $|\eta_0|, |\xi_0| < \epsilon$, $\epsilon > 0$ assegnato. Si considerino i vettori errore $[\eta_i, \xi_i]^T = [\widetilde{x}_i - x_i, \widetilde{y}_i - y_i]^T$, dove $\widetilde{x}_i$ e $\widetilde{y}_i$ sono i valori calcolati di x_i e y_i, supponendo che il calcolo delle formule ricorrenti non introduca altri errori. Anche $[\eta_{i+1}, \xi_{i+1}]^T$ risulta ruotato di 45^o rispetto al vettore $[\eta_i, \xi_i]^T$, ed è quindi $|\eta_i|, |\xi_i| < \epsilon$ per ogni i. Se però si usa l'analisi degli intervalli per stimare la propagazione dell'errore si ottiene

$$\eta_{i+1} = \frac{1}{\sqrt{2}} (\eta_i - \xi_i), \quad \xi_{i+1} = \frac{1}{\sqrt{2}} (\eta_i + \xi_i),$$

da cui

$$|\eta_{i+1}|,\ |\xi_{i+1}| \leq \sqrt{2} \max\{|\eta_i|,\ |\xi_i|\},$$

e quindi, poiché $|\eta_0|, |\xi_0| < \epsilon$, risulta

$$|\eta_{i+1}|,\ |\xi_{i+1}| < (\sqrt{2})^{i+1} \epsilon.$$

Geometricamente $(\widetilde{x}_0, \widetilde{y}_0)$ appartiene ad un quadrato di lato 2ϵ e con i lati paralleli agli assi x e y. Al primo passo tale quadrato è ruotato di 45^o e quindi è contenuto in un altro quadrato di lato $2\sqrt{2}\epsilon$, con i lati paralleli agli assi, come è illustrato dalla figura 2.18 anche per i passi successivi [6]. ∎

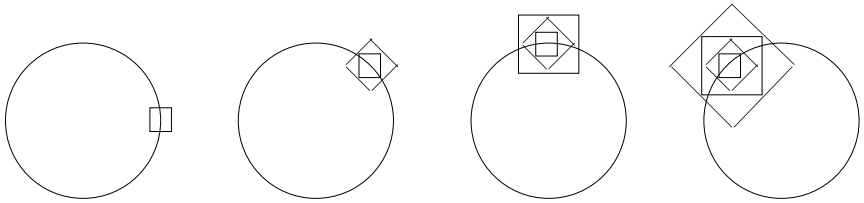


Fig. 2.18 - Analisi degli intervalli

d) Un'altra possibilità di controllo automatico dell'errore consiste nell'uso della *precisione variabile*, in cui i numeri vengono rappresentati con parole di lunghezza variabile. La lunghezza di questa rappresentazione viene determinata in modo da memorizzare tutte e sole le cifre esatte disponibili. Ad esempio, nel caso del risultato di un'operazione si tiene conto del numero di cifre esatte degli operandi, dell'errore propagato dall'operazione e dell'errore locale, ai fini dell'aumento o della diminuzione della lunghezza della rappresentazione. Dal punto di vista implementativo ciò può essere realizzato trattando i numeri come vettori di interi, la cui dimensione è indicata dal primo elemento del vettore.

Tutte le tecniche finora accennate richiedono la disponibilità di un'aritmetica diversa da quella normalmente presente sulla macchina, e che quindi deve essere implementata con un software opportuno e con elevati tempi di esecuzione. Le due tecniche seguenti consentono di ottenere delle indicazioni sul comportamento dell'errore, senza richiedere software non standard.

e) Una tecnica consiste nel *perturbare sperimentalmente* i dati del problema e nel confrontare i risultati così ottenuti con quelli corrispondenti ai dati originari. Perturbando separatamente insiemi di dati diversi si può misurare la sensibilità del problema alle variazioni sui dati.

f) Il modo più comunemente usato per avere delle indicazioni sull'ordine di grandezza dell'errore è quello di rieseguire il calcolo con un *numero maggiore di cifre*, cosa che in genere è possibile fare, perché nella maggior parte delle macchine è presente un'aritmetica con precisione doppia di quella standard. Anche se non si può garantire che i risultati ottenuti con la precisione doppia siano migliori, nella stragrande maggioranza dei casi il confronto fra i risultati ottenuti con le due aritmetiche permette di dare una stima dell'errore.

2.39 Esempio. Le seguenti due espressioni sono formalmente identiche

$$721 - 228\sqrt{10} \quad \text{e} \quad \frac{1}{721 + 228\sqrt{10}}\,.$$

Calcolandole in precisione semplice si ottengono

per la prima il valore $0.9765625\ 10^{-3}$,

per la seconda il valore $0.6934816\ 10^{-3}$.

Per stabilire quale dei due valori è più corretto, si riesegue il calcolo in precisione doppia ottenendo

per la prima il valore $0.69348160957361\ 10^{-3}$,

per la seconda il valore $0.69348160951230\ 10^{-3}$.

Se ne deduce che la seconda espressione è più stabile della prima. ∎

Esercizi proposti

2.1 Si scrivano le rappresentazioni nelle basi 2, 8 e 16 dei seguenti numeri

$$0.5, \quad 1, \quad 0.05, \quad 0.3, \quad 0.6, \quad \frac{1}{9}, \quad \frac{1}{8}, \quad \frac{1}{6}, \quad 41.41.$$

2.2 Si costruiscano tabelle dell'addizione e della moltiplicazione per varie basi $\beta \geq 2$. Ad esempio per la base $\beta = 2$ si ha

$+$	0	1
0	0	1
1	1	10

$*$	0	1
0	0	0
1	0	1

2.3 Esistono numeri in base 10 con 8 cifre che non si possono ottenere convertendo con arrotondamento alcun numero in base 16 con 6 cifre?

(Traccia: si tenga conto del fatto che le mantisse rappresentabili in base 16 con 6 cifre sono meno di quelle rappresentabili in base 10 con 8 cifre, in quanto $16^6 - 16^5 < 10^8 - 10^7$. Risulta ad esempio che tutti i numeri in base 10 con 8 cifre compresi fra $(0.62500061)\ 10^{-1}$ e $(0.62500118)\ 10^{-1}$ non si possono ottenere convertendo con arrotondamento alcun numero in base 16 con 6 cifre.)

2.4 Sia $\beta \geq 2$ e siano $s \neq 1$ e t due interi positivi.

a) si dica se
$$\mathcal{F}_{(\beta,st,sm,sM)} = \mathcal{F}_{(\beta^s,t,m,M)};$$

b) si scriva un algoritmo di conversione da $\mathcal{F}_{(\beta,st,sm,sM)}$ a $\mathcal{F}_{(\beta^s,t,m,M)}$ e viceversa;

c) si esamini in particolare il caso in cui $\beta = 2$, $s = 4$ e $t = 6$.

(Traccia: a) i due insiemi non coincidono, ad esempio il numero

$$a = (.1)_{\beta^s}\ (\beta^s)^{-m}$$

appartiene all'insieme $\mathcal{F}_{(\beta^s,t,m,M)}$, ma non all'insieme $\mathcal{F}_{(\beta,st,sm,sM)}$, infatti $a = (.1)_\beta\ \beta^{-sm-s+1}$ e l'esponente $-sm - s + 1$ è minore del minimo rappresentabile $-sm$; viceversa il numero $b = (.1\ldots1)_\beta\beta^1$, con st cifre della mantissa uguali a 1, appartiene a $\mathcal{F}_{(\beta,st,sm,sM)}$, ma non a $\mathcal{F}_{(\beta^s,t,m,M)}$; b) per convertire un numero dalla base β^s alla base β si sostituisca ad ogni cifra in base β^s la sua rappresentazione in s cifre in base β e si normalizzi il numero così ottenuto; viceversa per convertire un numero dalla base β alla

base β^s, se p è l'esponente rispetto alla base β e $|p|$ non è un multiplo di s, si determini il nuovo esponente rispetto alla base β^s, si trasli opportunamente la mantissa a destra e si sostituisca ad ogni gruppo di s cifre nella base β una cifra nella base β^s.

2.5 Sia $x > 0$ un numero che scritto nella base β abbia antiperiodo di k e periodo di n cifre, cioè

$$x = (.d_1 \dots d_k \overline{d_{k+1} \dots d_{k+n}})\ \beta^p;$$

si dimostri che

$$x = \frac{(.d_1 \dots d_{k+n})\ \beta^n - (.d_1 \dots d_k)}{\beta^n - 1}\ \beta^p.$$

(Traccia: è

$$\begin{aligned} x &= (.d_1 \dots d_k)\beta^p + \beta^{p-k} \sum_{i=0}^{\infty} (.d_{k+1} \dots d_{k+n})\ \beta^{-in} \\ &= (.d_1 \dots d_k)\beta^p + \beta^{p-k}(.d_{k+1} \dots d_{k+n})\ \frac{\beta^n}{\beta^n - 1}\ . \ ) \end{aligned}$$

2.6 Si dimostri che date due basi, β e γ, tali che γ ammetta come fattori primi tutti i fattori primi di β, ogni numero avente rappresentazione finita in base β ha anche rappresentazione finita in base γ.

(Traccia: si dimostri che per ogni termine della forma $d_i\beta^{-i}$, d_i intero, esistono j e c interi tali che $d_i\beta^{-i} = c\gamma^{-j}$.)

2.7 Si supponga che il numero x, $\beta^{-1} \le x < 1$, sia rappresentabile nella base β con t cifre. Si converta x in base γ. Si dimostri che esistono numeri x per cui il valore arrotondato alla T-esima cifra della mantissa nella base γ è 1 se e solo se vale la disuguaglianza

$$\beta^t \ge 2\gamma^T.$$

(Traccia: si verifichi prima che l'arrotondamento della mantissa in base γ è 1 se e solo se è $x \ge 1 - \frac{1}{2}\ \gamma^{-T}$. Inoltre si osservi che se esistono degli x per i quali l'arrotondamento della mantissa in base γ è 1, uno di tali x è certamente il più grande, cioè $x = 1 - \beta^{-t}$.)

2.8 Si verifichi che per l'algoritmo 2.21, che calcola $x \oplus y$ disponendo di un registro di lunghezza $t + 1$, vale

a) se $xy > 0$, $\quad x \oplus y = \text{trn}\ (x + y)$,
b) se $xy < 0$,
  1) se $p - q = 0$, allora $x \oplus y = x + y$,
  2) se $p - q = 1$, allora
    se $p - r = 0$, $\quad x \oplus y = \text{trn}\ (x + y)$,
    se $p - r > 0$, $\quad x \oplus y = x + y$,
  3) se $p - q \geq 2$, allora

$$\left| \frac{x \oplus y - (x + y)}{x + y} \right| < u = \beta^{-t+1}.$$

Si verifichi inoltre che nel caso in cui $p - r \geq 2$, in cui si può presentare il fenomeno della cancellazione, l'errore locale è nullo.

(Traccia: b) 1) in questo caso le cifre di g non vengono traslate, e quindi non si perde alcuna cifra; b) 2) poiché le cifre di g sono traslate di un solo posto, anche in questo caso non si perde alcuna cifra di g. Quindi il risultato è esatto o troncato in conseguenza della prima cifra di $f - g'$; b) 3) sia $x > 0 > y$; poiché $f \geq \beta^{-1}$ e in questo caso $g' < \beta^{-2}$, allora $f - g' > \beta^{-1} - \beta^{-2} > \beta^{-2}$; quindi $p - r < 2$. L'errore assoluto commesso nella sottrazione delle mantisse è

$$|(f - g') - (f - g\beta^{q-p})| = g\beta^{q-p} - g' < \beta^{-(t+1)}.$$

Si tenga conto del fatto che può essere richiesta, per la post-normalizzazione, la traslazione di al più un posto, perché $p - r < 2$.)

2.9 Si dimostri che se si calcola $x \ominus y$ disponendo di un registro di lunghezza $t + 1$, nel caso che sia

$$x \geq y \geq \frac{x}{2} \geq 0,$$

risulta

$$x \ominus y = x - y.$$

(Traccia: il risultato è ovvio se i due numeri hanno lo stesso esponente. Altrimenti la differenza degli esponenti deve essere 1, ma in tal caso la prima cifra della differenza delle mantisse è zero.)

2.10 Si scrivano due algoritmi per la moltiplicazione di macchina $\otimes$ analoghi agli algoritmi per l'addizione 2.19 e 2.21. Si verifichi che se il calcolo è eseguito con un registro di lunghezza $t + 1$, allora

$$x \otimes y = \text{trn}(xy),$$

e se il calcolo è eseguito con un registro di lunghezza $t+2$, allora

$$x \otimes y = \text{arr}(xy).$$

(Traccia: si tenga conto del fatto che il prodotto delle mantisse è compreso fra β^{-2} e 1, e quindi può essere richiesta una fase di post-normalizzazione.)

2.11 Si scrivano due algoritmi per la divisione di macchina $\oslash$ analoghi agli algoritmi per l'addizione 2.19 e 2.21. Si verifichi che se il calcolo è eseguito con un registro di lunghezza $t+1$, allora

$$x \oslash y = \text{trn}\,\frac{x}{y}\,,$$

e se il calcolo è eseguito con un registro di lunghezza $t+2$, allora

$$x \oslash y = \text{arr}\,\frac{x}{y}\,.$$

Si può verificare overflow per arrotondamento?

(Traccia: se la mantissa di y è minore della mantissa di x, si moltiplichi x per β^{-1}. Non si può verificare overflow per arrotondamento, perché la mantissa del quoziente non può avere t cifre uguali a $\beta - 1$.)

2.12 Operando in $\mathcal{F}_{(\beta,t,m,M)}$, β pari, con troncamento, si verifichi che

a) se $x = \beta^{-(t+2)}$, $y = 1$ e $z = -1$, allora

$$(x \oplus y) \oplus z = 0 \quad \text{e} \quad x \oplus (y \oplus z) = \beta^{-(t+2)},$$

b) se $x = 1$, $y = z = \dfrac{1}{2}\,\beta^{-(t-1)}$, allora

$$(x \oplus y) \oplus z = 1 \quad \text{e} \quad x \oplus (y \oplus z) = 1 + \beta^{-(t-1)}.$$

(Infatti non vale la proprietà associativa dell'addizione.)

2.13 Operando in $\mathcal{F}_{(\beta,t,m,M)}$, $\beta > 2$, $t \geq 2$, con troncamento, si verifichi che se $x = 1 + \beta^{-(t-1)}$, $y = z = 1 - \beta^{-t}$, allora

$$(x \otimes y) \otimes z = 1 - \beta^{-t} \quad \text{e} \quad x \otimes (y \otimes z) = 1.$$

(Infatti non vale la proprietà associativa della moltiplicazione.)

2.14 Operando in $\mathcal{F}_{(\beta,t,m,M)}$, $\beta > 2$, $t \geq 4$, con troncamento, si verifichi che se $x = 2$, $y = \beta^2 - 1$, $z = \beta^2 - 1 + \beta^{-(t-2)}$, allora

$$x \otimes y = x \otimes z = 2\beta^2 - 2 \quad \text{e} \quad y \neq z.$$

(Infatti non vale la legge di cancellazione.)

2.15 Operando in $\mathcal{F}_{(\beta,t,m,M)}$, $t \geq 4$, con troncamento, si verifichi che se $x = \beta^2 - 1$, $y = \beta^2 - 1 + \beta^{-(t-2)}$, $z = (\beta - 1)\beta^{-(t-2)}$, allora

$$(x \otimes y) \oplus (x \otimes z) = (\beta^2 - 1)^2 + (\beta - 2)\beta^{-(t-4)}$$

e

$$x \otimes (y \oplus z) = (\beta^2 - 1)^2 + (\beta - 1)\beta^{-(t-4)}.$$

(Infatti non vale la legge distributiva.)

2.16 Si dimostri che, operando in $\mathcal{F}_{(\beta,t,m,M)}$ con troncamento o con arrotondamento, valgono le seguenti disuguaglianze

a) se $x < y$, allora $x \oplus z \leq y \oplus z$, per ogni z;

b) se $x < y$ e $w < z$, allora $x \oplus w \leq y \oplus z$;

c) se $x < y$, allora $x \otimes z \leq y \otimes z$, per ogni $z > 0$.

Si verifichi che, se si opera con trocamento,

a) se $\beta > 2$, $x = \beta^{-t}$, $y = 2\beta^{-t}$ e $z = 1$, allora

$$x < y \quad \text{e} \quad x \oplus z = y \oplus z = 1,$$

b) se $x = 1 - \beta^{-t}$, $y = 1$, $w = \beta^{-t}$, $z = \beta^{-t} + \beta^{-(t+1)}$ e $t \geq 2$, allora

$$x < y, \quad w < z \quad \text{e} \quad x \oplus w = y \oplus z = 1.$$

Si individuino esempi analoghi per l'arrotondamento.

2.17 Si verifichi con un controesempio che non sempre in aritmetica di macchina il punto di mezzo dell'intervallo $[a, b]$ calcolato tramite l'espressione

$$(a \oplus b) \oslash 2,$$

appartiene all'intervallo stesso. Per lo stesso caso si verifichi che invece il punto di mezzo dell'intervallo calcolato con l'espressione

$$a \oplus ((b \ominus a) \oslash 2)$$

appartiene all'intervallo.

(Traccia: in $\mathcal{F}_{(10,1,m,M)}$ siano $a = 0.6\ 10^{-1}$ e $b = 0.9\ 10^{-1}$. Risulta $(a \oplus b) \oslash 2 \notin [a, b]$, sia che le operazioni di macchina siano definite con troncamento che con arrotondamento.)

2.18 Si dimostri che se $a, b \in \mathcal{F}_{(\beta,t,m,M)}$, $a \neq 0$, e si opera con troncamento, allora l'equazione

$$a \otimes x = b$$

a) non ha sempre soluzione in $\mathcal{F}_{(\beta,t,m,M)}$,
b) se ha soluzione, questa non è sempre unica,

anche se non si verificano errori di overflow e di underflow.

(Traccia: a) si verifichi, con un controesempio, che se $t \geq 4$ e la mantissa di a è diversa da β^{-1} e da $1 - \beta^{-t}$, allora esistono valori di b per cui non c'è soluzione; b) si verifichi che se il prodotto $a \otimes x$ non richiede post-normalizzazione, esso può restare inalterato anche se si aumenta la mantissa di x di β^{-t}.)

2.19 Siano $a, x \in \mathcal{F}_{(\beta,t,m,M)}$, $\beta > 2$, $t \geq 2$, $a = 1 - \beta^{-t}$.

a) Si dica qual è il prodotto $a \otimes x$ nel caso che si operi con troncamento o con arrotondamento;
b) si dimostri che l'equazione $a \otimes x = 1$ non ha soluzioni in $\mathcal{F}_{(\beta,t,m,M)}$, sia che si operi con troncamento che con arrotondamento.

(Traccia: a) sia $x = f\beta^p$;
se $f = \beta^{-1}$, allora $a \otimes x = ax$ (con tronc. o con arrot.),
se $\beta^{-1} < f \leq \frac{1}{2}$, allora $a \otimes x = x$ (con arrot.), $a \otimes x = x - \beta^{p-t}$ (con tronc.),
se $\frac{1}{2} < f < 1$, allora $a \otimes x = x - \beta^{p-t}$ (con tronc. o con arrot.).)

2.20 Siano x, y e z tre numeri di macchina. È possibile che nel calcolo di $(x \otimes y) \otimes z$ si verifichi errore di overflow o di underflow, ma non nel calcolo di $x \otimes (y \otimes z)$?

2.21 In $\mathcal{F}_{(\beta,t,m,M)}$ si determini il massimo intervallo $I = [a, b]$, $\omega < a < b < \Omega$, tale che per ogni numero di macchina $x \in I$, il calcolo di $x \otimes x$ non dia luogo a errori di overflow o di underflow se si opera con troncamento.

(Traccia: se M è pari, risulta

$$b = (1 - \beta^{-t})\beta^{M/2},$$

se M è dispari, risulta

$$b = z\beta^{(M+1)/2}, \quad \text{dove } z = \max\{y \in \mathcal{F}_{(\beta,t,m,M)} \ : \ y \otimes y < \beta^{-1}\}.$$

Per a si proceda in modo analogo.)

2.22 Nell'ipotesi che la condizione di underflow non sia segnalata come errore, si possono avere notevoli perdite di accuratezza nel risultato anche nel calcolo di espressioni semplici. Si calcolino ad esempio in $\mathcal{F}_{(\beta,t,m,M)}$ le due espressioni identiche

$$\frac{ay+b}{ay+c} \;=\; \frac{a+b/y}{a+c/y}$$

per i seguenti numeri di macchina

$$a=\beta^{-m-1},\quad b=\beta^{-m-1},\quad c=(\beta-1)\beta^{-m-1},\quad y=(\beta-1)\beta^{-1},$$

nell'ipotesi di porre a zero il risultato di un'operazione di macchina che dia luogo ad underflow.

2.23 a) Siano $\beta \geq 2$ e d un intero, tale che $0 < d < \beta - 1$. Si verifichi che il numero $\dfrac{d}{\beta-1}$ ha nella base β la rappresentazione periodica $(.\overline{d})_\beta\ \beta^0$; b) si calcolino in $\mathcal{F}_{(16,6,m,M)}$, operando con troncamento, le espressioni

$$\alpha = \big[\mathrm{arr}\big(\frac{9}{15}\big) \ominus \mathrm{arr}\big(\frac{7}{15}\big)\big] \ominus \big[\mathrm{arr}\big(\frac{2}{15}\big) \ominus 16^{-6}\big],$$
$$\beta = \big[\mathrm{trn}\big(\frac{9}{15}\big) \ominus \mathrm{trn}\big(\frac{7}{15}\big)\big] \ominus \big[\mathrm{trn}\big(\frac{2}{15}\big) \ominus 16^{-6}\big].$$

Si calcolino gli errori relativi da cui sono affetti α e β e si confrontino con le maggiorazioni degli errori ottenute con l'analisi diretta.

(Risposta: b) si ha $\alpha = (.2)_{16}\ 16^{-5}$ e $\beta = (.1)_{16}\ 16^{-5}$, con errori $\epsilon_\alpha = 1$ e $\epsilon_\beta = 0$. Con l'analisi diretta risulta invece $|\epsilon_\alpha| < 14$ e $|\epsilon_\beta| < 24$.)

2.24 Assegnato un intero n, per calcolare $x_i = ih$, $h = 1/n$, per $i = 1, \ldots, n$, si possono usare i seguenti algoritmi:

a) $x_i = i\ h, \quad i = 1, \ldots, n,$

b) $x_1 = h, \qquad x_{i+1} = x_i + h, \quad i = 1, \ldots, n-1.$

Si confrontino gli errori algoritmici dei due algoritmi, e si dica per quale dei due è meglio limitato l'errore nel calcolo di x_n. Si confronti con i valori effettivamente calcolati in $\mathcal{F}_{(16,6,m,M)}$ con arrotondamento, per $n = 8, 10, 12$.

2.25 [6] Per quali delle seguenti espressioni

$$(\sqrt{2}-1)^6 = \frac{1}{(\sqrt{2}+1)^6} = (3-2\sqrt{2})^3$$
$$= \frac{1}{(3+2\sqrt{2})^3} = 99-70\sqrt{2} = \frac{1}{99+70\sqrt{2}}$$

si ha una più contenuta propagazione dell'errore dovuta all'approssimazione di $\sqrt{2}$?

2.26 Si studi la propagazione dell'errore nel calcolo delle seguenti espressioni (le parentesi indicano l'ordine da seguire nelle operazioni)

a) 1) $(x+2)^2$, 2) $(x^2+4x)+4$, 3) $x\,(x+4)+4$;

b) 1) $(x+1)^2\,(x+1)$, 2) $((x+3)\,x+3)\,x+1$;

c) 1) $(((x^2)\,x)\,x)\sqrt{x}$, 2) $(x^2\,x^2)\sqrt{x}$, 3) $\sqrt{((x^2)^2)^2\,x}$, $x>0$;

d) 1) $x^2\sqrt{x}$, 2) $\sqrt{\exp(5\,\log x)}$, 3) $\exp(2.5\,\log x)$, $x>0$;

e) 1) $\log(x^2/y^2)$, 2) $\log x^2 - \log y^2$, 3) $2(\log x - \log y)$, $x,y>0$;

f) 1) $\left(\frac{1}{x}\right)^n$, 2) $\frac{1}{x^n}$, n intero;

g) 1) $\sqrt{x-\sqrt{y}}$, 2) $\sqrt{\frac{x+z}{2}}-\sqrt{\frac{x-z}{2}}$, in cui $z=\sqrt{x^2-y}$, $x^2>y>0$;

h) 1) $\sqrt{x+y}-\sqrt{x}$, 2) $\frac{y}{\sqrt{x+y}+\sqrt{x}}$, $x,y>0$;

i) 1) $(\sqrt{x^2+1}-x)x$, 2) $\frac{x}{\sqrt{x^2+1}+x}$.

In ogni caso si supponga di disporre di funzioni di libreria per il calcolo del logaritmo, dell'esponenziale e della radice quadrata, che producono risultati affetti da errore analitico e algoritmico limitati in modulo dalla precisione di macchina.

2.27 Siano x_i, $i=1,\ldots,n$, n numeri di macchina e si verifichi, senza usare l'approssimazione al primo ordine, che l'errore relativo di arrotondamento commesso nel calcolo di

$$p=\prod_{i=1}^{n} x_i,$$

è dato da

$$\sigma=\prod_{i=1}^{n-1}(1+\rho_i)-1,$$

dove ρ_i è l'errore locale introdotto con la i-esima moltiplicazione, e inoltre è

$$|\sigma| < \mathrm{e}^{(n-1)u} - 1 \quad \text{dove} \quad u = \max_{i=1,\ldots,n-1} |\rho_i|.$$

(Traccia: si verifichi che per ogni $x > 0$ e per ogni intero k è

$$(1+x)^k < \mathrm{e}^{kx}. \ )$$

2.28 Si dica se è conveniente, dal punto di vista della propagazione dell'errore, razionalizzare le espressioni

$$a) \quad \frac{\sqrt{x} - \sqrt{y}}{\sqrt{x} + \sqrt{y}}, \qquad b) \quad \frac{\sqrt{x} + \sqrt{y}}{\sqrt{x} - \sqrt{y}}, \qquad x, y > 0.$$

(Traccia: non è conveniente razionalizzare la a); è conveniente razionalizzare la b) se $\sqrt{x}\sqrt{y} > |x - y|$.)

2.29 Per calcolare $x^{2^k/n}$, n, k interi positivi, si possono usare i due algoritmi seguenti:

a) $z^{(1)} = x^{\frac{1}{n}}; \quad z^{(i)} = (z^{(i-1)})^2, \ i = 2, \ldots, k+1,$

b) $z^{(1)} = x; \quad z^{(i)} = (z^{(i-1)})^2, \ i = 2, \ldots, k+1; \quad z^{(k+2)} = (z^{(k+1)})^{\frac{1}{n}}.$

Supponendo di disporre di funzioni di libreria di elevamento a potenza reale i cui errori analitici e algoritmici sono limitati in modulo dalla precisione di macchina, si dica quale dei due algoritmi è preferibile dal punto di vista dell'errore.

(Traccia: risulta

$$|\epsilon_{alg1}| \dot{<} (2^{k+1} - 1 + \frac{2^k}{n} \, |\log x|)u, \quad |\epsilon_{alg2}| \dot{<} (\frac{2^k - 1}{n} + 1 + \frac{2^k}{n} \, |\log x|)u.$$

Pertanto il secondo algoritmo è da preferire.)

2.30 Sia A una matrice di ordine n i cui elementi $a_{ij} \geq 0, \ i, j = 1, \ldots, n,$ sono numeri di macchina non tutti nulli. La somma s degli elementi di A può essere calcolata con due diversi algoritmi:

a) algoritmo sequenziale

$z^{(0)} = 0,$

$z^{(k)} = z^{(k-1)} + a_{ij}$, dove $(k = (i-1)n + j, \ j = 1, \ldots, n,), \ i = 1, \ldots, n,$

$s = z^{(n^2)};$

b) algoritmo parallelo: si costruisce il vettore $\mathbf{b}$ la cui i-esima componente è la somma degli elementi della i-esima riga di A, poi si sommano gli elementi di $\mathbf{b}$

$$\left.\begin{array}{l} v^{(i,0)} = 0, \\ v^{(i,j)} = v^{(i,j-1)} + a_{ij},\ j = 1, \ldots, n, \\ b_i = v^{(i,n)}, \end{array}\right\} \quad i = 1, \ldots, n,$$

$$w^{(0)} = 0,$$
$$w^{(i)} = w^{(i-1)} + b_i,\ i = 1, \ldots, n,$$
$$s = w^{(n)}.$$

Si dica quale dei due algoritmi è preferibile dal punto di vista dell'errore.

(Traccia: risulta

$$|\epsilon_{alg1}| \dot{<} (n^2 - 1)u, \quad |\epsilon_{alg2}| \dot{<} 2(n-1)u.$$

Pertanto il secondo algoritmo è da preferire.)

2.31 Siano $x > 0$ un numero di macchina e $n > 1$ un intero. Per il calcolo di x^n si può procedere in uno dei modi seguenti:

a) $x^n = \prod_{i=1}^{n} x;$

b) sia $n = \sum_{i=0}^{k} d_i 2^i$, con $k = \lfloor \log_2 n \rfloor$, la rappresentazione in base 2 di n, si calcolino le potenze $z_i = x^{2^i}$, $i = 0, \ldots, k$, e si ottenga x^n come

$$x^n = \prod_{\substack{i=0 \\ d_i = 1}}^{k} z_i \ ;$$

c) $x^n = \mathrm{e}^{n \log\ x}$.

Supponendo che le funzioni di libreria che calcolano $\log\ x$ ed e^x abbiano errori analitici e algoritmici limitati in modulo dalla precisione di macchina, si dica quale dei tre algoritmi è preferibile dal punto di vista dell'errore (per il costo computazionale dell'algoritmo b) si veda l'esercizio 1.5).

(Traccia: risulta

$$|\epsilon_{alg1}| \dot{<} (n-1)u, \quad |\epsilon_{alg2}| \dot{<} (n-1)u, \quad |\epsilon_{alg3}| \dot{<} (1 + 2n|\log x|)u.$$

Pertanto il terzo algoritmo è da preferire solo se x è sufficientemente vicino a 1.)

2.32 Siano $k > 1$ un intero e $x > 0$, $x \neq 1$, un numero di macchina. Per calcolare il valore del polinomio

$$p(x) = \sum_{i=0}^{n} x^i, \quad n = 2^k - 1,$$

si possono usare i seguenti algoritmi:

a) $p(x) = (\dots((x+1)x+1)\dots)x+1$, (metodo di Ruffini-Horner)

b) $p(x) = \dfrac{x^{n+1}-1}{x-1}$,

c) $p(x) = (1+x)\,(1+x^2)\,(1+x^4)\dots(1+x^{2^{k-1}})$.

In b) e c) le potenze x^{2^i}, $i = 1, \dots, k$, vengono calcolate come

$$x^{2^i} = \left(x^{2^{i-1}}\right)^2.$$

Si confrontino i tre algoritmi dal punto di vista dell'errore prodotto (per il costo computazionale si veda l'esercizio 1.4).

(Traccia: risulta

$$|\epsilon_{alg1}| \dot{<} (2n-1)u, \quad |\epsilon_{alg2}| \dot{<} \left(3 + n\,\frac{x^{n+1}}{|x^{n+1}-1|}\,\right)u, \quad |\epsilon_{alg3}| \dot{<} (n+k-1)u.$$

Pertanto il terzo algoritmo è sempre preferibile rispetto al primo ed è da preferire al secondo se x è vicino a 1.)

2.33 Siano $\mathbf{x}$ un vettore di numeri di macchina di ordine n. Per calcolare $s = \|\mathbf{x}\|_2$ si può procedere in uno dei modi seguenti:

a) $s = \sqrt{\sum_{i=1}^{n} x_i^2}$,

b) $\alpha = \max_{i=1,\dots,n} |x_i|, \quad s = \alpha \sqrt{\sum_{i=1}^{n} \left(\frac{x_i}{\alpha}\right)^2}$.

Il secondo algoritmo presenta minori rischi di overflow. Si confrontino i due algoritmi dal punto di vista dell'errore prodotto.

(Traccia: risulta

$$|\epsilon_{alg1}| \dot{<} (1 + \frac{n}{2})u, \quad |\epsilon_{alg2}| \dot{<} (3 + \frac{n}{2})u.$$

Pertanto i due algoritmi sono equivalenti dal punto di vista dell'errore.)

2.34 Si studi l'errore algoritmico del calcolo delle due espressioni

$$\frac{\tan\alpha - \tan\beta}{\tan\alpha + \tan\beta} = \frac{\sin(\alpha-\beta)}{\sin(\alpha+\beta)}\,, \quad 0 < \beta < \alpha < \frac{\pi}{4}\,,$$

nell'ipotesi che le funzioni di libreria che calcolano $\tan x$ e $\sin x$ abbiano errori analitici e algoritmici limitati in modulo dalla precisione di macchina.

(Traccia: per la prima espressione è

$$|\epsilon_{alg1}| \dot{<} u\Big[3 + \frac{4\tan\alpha\tan\beta}{\tan^2\alpha - \tan^2\beta}\Big],$$

e quindi l'errore algoritmico non è limitabile in modulo se α e β sono vicini, mentre per la seconda espressione è

$$|\epsilon_{alg2}| \dot{<} u\Big[3 + \frac{\alpha-\beta}{\tan(\alpha-\beta)} + \frac{\alpha+\beta}{\tan(\alpha+\beta)}\Big] < 5u.\ )$$

2.35 Per calcolare le radici quadrate complesse $x + \boldsymbol{i}y$ del numero $a + \boldsymbol{i}b$, $b \neq 0$, si può usare l'algoritmo

$$y = \pm\sqrt{\frac{1}{2}(\sqrt{a^2+b^2} - a)}, \quad x = \frac{b}{2y}\,.$$

a) Si studi l'errore algoritmico, con particolare riferimento al caso in cui b sia vicino a 0;

b) si modifichi l'algoritmo proposto in modo da evitare possibili errori di cancellazione.

(Traccia: a) se $a > 0$ l'errore algoritmico non è limitato quando b è in un intorno di 0, b) per $a > 0$ si calcoli

$$|y| = \frac{|b|}{\sqrt{2(\sqrt{a^2+b^2} + a)}}\,.\ )$$

2.36 Sia α una radice reale con molteplicità s della funzione $f(x)$, derivabile con continuità fino all'ordine s. Si dica se il fattore di amplificazione c_x dell'errore inerente di $f(x)$ è limitato in un intorno di α, e si calcoli, se esiste, il

$$\lim_{x\to\alpha} c_x.$$

Si studino in particolare i casi $f(x) = \sin x$ e $f(x) = \sin^2 x$ in un intorno di $x = 0$ e di $x = \pi$, $f(x) = \arctan x$ e $f(x) = \mathrm{e}^{-x} - \dfrac{1}{1+x}$ in un intorno di $x = 0$.

(Traccia: per la formula di Taylor risulta

$$f(x) = \frac{(x-\alpha)^s}{s!}\, f^{(s)}(\xi_1), \quad |\xi_1 - \alpha| < |x - \alpha|,$$

$$f'(x) = \frac{(x-\alpha)^{s-1}}{(s-1)!}\, f^{(s)}(\xi_2), \quad |\xi_2 - \alpha| < |x - \alpha|,$$

per cui

$$c_x = \frac{x f'(x)}{f(x)} = \frac{sx}{x-\alpha}\, \frac{f^{(s)}(\xi_2)}{f^{(s)}(\xi_1)},$$

e quindi c_x è limitato in un intorno di α solo se $\alpha = 0$ ed è $\lim_{x\to\alpha} c_x = s$. Per $\alpha \neq 0$ è $\lim_{x\to\alpha} |c_x| = \infty$.)

2.37 Sia $f(x) = g(h(x))$, dove $h(x)$ e $g(y)$ sono due funzioni da $\mathbf{R} \to \mathbf{R}$.

a) Si dimostri che il coefficiente di amplificazione dell'errore inerente di f calcolato in x è uguale al prodotto del coefficiente di amplificazione di h calcolato in x per il coefficiente di amplificazione di g calcolato in $h(x)$;

b) se $f(x) = g_1(h_1(x)) = g_2(h_2(x))$, è possibile calcolare $f(x)$ con due algoritmi diversi. Se il modulo del coefficiente di amplificazione di g_1 è maggiore del modulo del coefficiente di amplificazione di g_2, si dica quale dei due algoritmi è più conveniente nell'ipotesi che gli errori algoritmici siano dell'ordine della precisione di macchina.

(Traccia: b) siano ϵ_{alg_1} e ϵ_{alg_2} gli errori algoritmici del calcolo di $h_1(x)$ e $h_2(x)$ e siano η_{alg_1} e η_{alg_2} gli errori algoritmici del calcolo di $g_1(y)$ e $g_2(y)$. Indicati con γ_1 e γ_2 i coefficienti di amplificazione di $g_1(y)$ e $g_2(y)$, gli errori algoritmici del calcolo di $g_1(h_1(x))$ e $g_2(h_2(x))$ sono rispettivamente $\gamma_1\epsilon_{alg_1} + \eta_{alg_1}$ e $\gamma_2\epsilon_{alg_2} + \eta_{alg_2}$. Se tutti gli errori algoritmici sono dello stesso ordine di grandezza della precisione di macchina, il secondo algoritmo è preferibile.)

2.38 Si giustifichi la diversa propagazione dell'errore nel calcolo di e^{-9} con i due algoritmi utilizzati nell'esempio 1.5. Si esegua il calcolo anche con l'algoritmo ottenuto sommando separatamente i termini positivi e quelli negativi, e sommando poi i due totali parziali. Si verifichi che in tal caso l'errore è ancora più elevato e se ne dia una spiegazione.

(Traccia: si tenga conto del punto b) dell'esercizio 2.37.)

2.39 Si valutino gli errori analitici commessi approssimando in un intorno dello zero le seguenti funzioni non razionali con polinomi ottenuti troncando

le serie di Maclaurin al secondo o terzo termine non nullo:

a) $\sinh x = \dfrac{\mathrm{e}^x - \mathrm{e}^{-x}}{2}$, b) $\tanh x = \dfrac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}}$, c) $\dfrac{1}{1+x} - \mathrm{e}^{-x}$,
d) $\mathrm{e}^x - \sqrt{1+x}$, e) $x\cos x - \sin x$.

(Traccia: troncando al secondo termine, per $|x| < \dfrac{1}{2}$ si ha:

a) $|\epsilon_{an}| < 0.01\ x^4$, b) $|\epsilon_{an}| < 0.15\ x^4$, c) $|\epsilon_{an}| < 11.5\ x^2$,
d) $|\epsilon_{an}| < 1.67\ x^2$, e) $|\epsilon_{an}| < 0.0042\ x^4$.

Si proceda in modo analogo quando si tronca al terzo termine.)

2.40 a) Si determinino la media e la varianza di una variabile casuale x che può assumere i valori 1 con probabilità α e 0 con probabilità $1-\alpha$;

b) siano $x_1, \ldots, x_n$ variabili casuali indipendenti con la stessa distribuzione di probabilità della variabile x definita in a); si determinino la media μ_n e la varianza σ_n della variabile casuale

$$\delta_n = \sum_{i=1}^{n} x_i;$$

c) si determini la probabilità che, per valori grandi di n sia

$$|\delta_n - \mu_n| < k\sigma_n;$$

d) si verifichi la (14, cap. 1).

(Traccia: a) risulta $\mu = \alpha$ e $\sigma^2 = \alpha(1-\alpha)$; b) si utilizzi il teorema 2.33, risulta $\mu_n = n\alpha$ e $\sigma_n^2 = n\alpha(1-\alpha)$; c) per il teorema centrale di convergenza si può assumere che la distribuzione della variabile δ_n sia normale. Quindi la probabiltà richiesta è $\mathrm{erf}\Big(\dfrac{k}{\sqrt{2}}\Big)$; d) nell'esempio 1.13 è

$$\alpha = \frac{\pi}{4}, \quad \delta_n = \frac{np_n}{4}, \quad \mu_n = \frac{n\pi}{4}, \quad \sigma_n^2 = \frac{n\pi}{4}\big(1 - \frac{\pi}{4}\big).\ )$$

2.41 Si dimostri che

a) $\mathrm{erf}(-x) = -\mathrm{erf}(x), \quad \mathrm{erf}(0) = 0, \quad \lim\limits_{x\to\infty} \mathrm{erf}(x) = 1,$

b) $\mathrm{erf}(x) = \dfrac{2}{\sqrt{\pi}}\left(x - \dfrac{x^3}{3\cdot 1!} + \dfrac{x^5}{5\cdot 2!} - \dfrac{x^7}{7\cdot 3!} + \ldots\right).$

(Traccia: la prima segue dalla simmetria di e^{-t^2}, la terza da

$$\int_0^\infty \mathrm{e}^{-t^2}\ dt = \frac{\sqrt{\pi}}{2}\,;$$

b) si scriva la formula di Maclaurin tenendo conto che

$$\mathrm{erf}'(x) = \frac{2}{\sqrt{\pi}}\ \mathrm{e}^{-x^2}.\ )$$

2.42 Sia x una variabile casuale distribuita uniformemente nell'intervallo $[1/2, 1)$ e siano $\mu(\epsilon)$ la media e $\sigma^2(\epsilon)$ la varianza della funzione

$$\epsilon(x) = \frac{x - \widetilde{x}}{x}\,,$$

dove $\widetilde{x}$ è il valore troncato di x in $\mathcal{F}_{(2,t,m,M)}$. Si verifichi che

$$\mu(\epsilon) = \frac{u}{2}\ \log 2 + O(u^2), \quad u = 2^{-t+1},$$
$$\sigma^2(\epsilon) = \frac{u^2}{4}\ \left(\frac{2}{3}\ - \log^2 2\right) + O(u^3).$$

(Traccia: la funzione densità di probabilità di x è

$$p(x) = \begin{cases} 2 & \text{per } 1/2 \le x < 1, \\ \\ 0 & \text{altrimenti.} \end{cases}$$

Si considerano tutti i numeri di macchina x_i nell'intervallo $[1/2, 1)$:

$$x_i = \frac{1}{2}\ + ih, \quad i = 0, \ldots, n, \quad h = 2^{-t} = \frac{1}{2}\ u, \quad n = \frac{1}{2h}\ - 1 = \frac{1}{u}\ - 1.$$

Si calcoli

$$\mu(\epsilon) = \int_{-\infty}^{\infty} \epsilon(x) p(x)\ dx = \int_{\frac{1}{2}}^{1} 2\ \frac{x - \widetilde{x}}{x}\ dx$$
$$= 1 - 2\sum_{i=0}^{n} \int_{x_i}^{x_i + h} \frac{x_i}{x}\ dx = 1 - 2\sum_{i=0}^{n} x_i \big[\log(x_i + h) - \log x_i\big].$$

Per il calcolo della sommatoria si applichi la formula di Eulero-Maclaurin (si veda l'esercizio 4.42) alla funzione $f(x) = x\big[\log(x+h) - \log x\big]$, sapendo che una primitiva di $f(x)$ è

$$\frac{1}{2}\ \big[(x^2 - h^2)\log(x + h) - x^2 \log x + hx\big].$$

Risulta

$$\sum_{i=0}^{n} x_i \big[\log(x_i + h) - \log x_i \big] = \frac{1}{2} - \frac{h}{2} \log 2 + \frac{h^2}{12} + O(h^3),$$

da cui segue che

$$\mu(\epsilon) = h \log 2 - \frac{h^2}{6} + O(h^3).$$

Per la varianza si ha analogamente

$$\begin{aligned}
\sigma^2(\epsilon) &= \int_{\frac{1}{2}}^{1} \big[\epsilon(x) - \mu(\epsilon)\big]^2 p(x)\, dx = -(\mu(\epsilon) - 1)^2 + 2 \int_{\frac{1}{2}}^{1} \frac{\widetilde{x}^{\,2}}{x^2}\, dx \\
&= -(\mu(\epsilon) - 1)^2 + 2 \sum_{i=0}^{n} \int_{x_i}^{x_i + h} \frac{x_i^2}{x^2}\, dx \\
&= -(\mu(\epsilon) - 1)^2 + 2h \sum_{i=0}^{n} \Big(1 - \frac{h}{x_i + h}\Big),
\end{aligned}$$

e si approssimi la sommatoria con la formula di Eulero-Maclaurin.)

2.43 Siano $u = 2^{-t}$, $t > 0$ e x una variabile casuale distribuita uniformemente nell'intervallo $[1/2, 1 - u/2]$ e siano $\mu(\epsilon)$ la media e $\sigma^2(\epsilon)$ la varianza della funzione

$$\epsilon(x) = \frac{x - \widetilde{x}}{x},$$

dove $\widetilde{x}$ è il valore arrotondato di x in $\mathcal{F}_{(2,t,m,M)}$. Si verifichi che

$$\begin{aligned}
\mu(\epsilon) &= \frac{u^2}{3} + O(u^3), \\
\sigma^2(\epsilon) &= \frac{u^2}{6} + O(u^3).
\end{aligned}$$

(Traccia: la funzione densità di probabilità di x è

$$p(x) = \begin{cases} \dfrac{2}{1-u} & \text{per } \dfrac{1}{2} \le x \le 1 - \dfrac{u}{2}, \\ 0 & \text{altrimenti.} \end{cases}$$

Posto $x_i = \dfrac{1}{2} + iu$, $i = 0, \ldots, n$, $n = \dfrac{1}{2u} - 1$, si proceda come nell'esercizio

2.42. È

$$\begin{aligned}\mu(\epsilon) &= \int_{-\infty}^{\infty} \epsilon(x)p(x)\,dx = \int_{\frac{1}{2}}^{1-\frac{u}{2}} \frac{x-\widetilde{x}}{x}\,\frac{2}{1-u}\,dx \\ &= 1 - \frac{2}{1-u}\left[\int_{\frac{1}{2}}^{\frac{1}{2}+\frac{u}{2}} \frac{x_0}{x}\,dx + \sum_{i=1}^{n}\int_{x_i-\frac{u}{2}}^{x_i+\frac{u}{2}} \frac{x_i}{x}\,dx\right] \\ &= 1 - \frac{2}{1-u}\left[\frac{1}{2}\log(1+u) + \sum_{i=1}^{n} x_i\big[\log(x_i+\frac{u}{2}) - \log(x_i-\frac{u}{2})\big]\right].\end{aligned}$$

Applicando la formula di Eulero-Maclaurin si verifichi che

$$\sum_{i=1}^{n} x_i\big[\log(x_i+\frac{u}{2}) - \log(x_i-\frac{u}{2})\big] = \frac{1}{2} - u + \frac{u^2}{12} + O(u^3).$$

Si proceda in modo analogo per la varianza.)

2.44 Una variabile casuale x si dice con *distribuzione logaritmica* nell'intervallo $[1/2, 1)$ se la sua densità di probabilità è

$$p(x) = \begin{cases} \dfrac{1}{x\log 2} & \text{se } x \in [\frac{1}{2}, 1), \\ 0 & \text{altrimenti,} \end{cases}$$

e cioè la funzione di distribuzione è

$$F(x) = \int_{-\infty}^{x} p(\xi)\,d\xi = 1 + \frac{\log x}{\log 2}, \quad \text{se } x \in \Big[\frac{1}{2}, 1\Big).$$

Si dimostri che se x e y sono variabili casuali indipendenti tali che le loro mantisse in base 2 abbiano distribuzione logaritmica nell'intervallo $[1/2, 1)$, allora

a) la probabilità che nella moltiplicazione di x e di y sia richiesta la post-normalizzazione è $\dfrac{1}{2}$;

b) la mantissa di $z = xy$ ha distribuzione logaritmica.

Una proprietà analoga alla b) non vale se le distribuzioni delle mantisse di x e di y sono uniformi. Anche per questo motivo è condivisa l'ipotesi che la distribuzione delle mantisse dei numeri in virgola mobile sia logaritmica. Un'altra motivazione di questo è dovuta al fatto che molti dei numeri che

compaiono nei calcoli rappresentano quantità la cui distribuzione deve essere indipendente dall'unità di misura utilizzata, ed è possibile dimostrare che questa indipendenza implica necessariamente la distribuzione logaritmica.

(Traccia: a) siano

$$x = f\ 2^p, \quad y = g\ 2^q, \quad f, g \in \left[\frac{1}{2}, 1\right).$$

La post-normalizzazione è richiesta se risulta $g < \dfrac{1}{2f}$, e la sua probabilità è quindi data da

$$\int_{\frac{1}{2}}^{1} \left(\int_{\frac{1}{2}}^{\frac{1}{2f}} p(f)p(g)\ dg \right)\ df = \int_{\frac{1}{2}}^{1} \int_{\frac{1}{2}}^{\frac{1}{2f}} \frac{dg}{fg \log^2 2}\ df = \frac{1}{2}\,;$$

b) è sufficiente verificare che la probabilità che la mantissa di z sia minore di $\alpha \in [1/2, 1)$ è

$$F(\alpha) = 1 + \frac{\log \alpha}{\log 2}\,.$$

Tale probabilità è data dalla somma delle probabilità che $fg < \alpha$ se non è richiesta la post-normalizzazione, cioè se $fg \in [1/2, 1)$, e che $2fg < \alpha$ se è richiesta la post-normalizzazione, cioè se $fg \in [1/4, 1/2)$. Poiché la probabilità che la post-normalizzazione sia richiesta oppure no è di $\dfrac{1}{2}$, risulta

$$F(\alpha) = \frac{1}{2} \int_{\frac{1}{2}}^{1} \int_{\frac{1}{2}}^{\frac{\alpha}{f}} p(f)p(g)\ dg\ df + \frac{1}{2} \int_{\frac{1}{2}}^{1} \int_{\frac{1}{2}}^{\frac{\alpha}{2f}} p(f)p(g)\ dg\ df = 1 + \frac{\log \alpha}{\log 2}\,.\ )$$

2.45 Si definiscano le operazioni di sottrazione, moltiplicazione e divisione per l'aritmetica degli intervalli.

Commento bibliografico

La notazione posizionale oggi usata nella rappresentazione dei numeri può essere fatta risalire ai babilonesi, che usavano la base 60, base utilizzata ancora oggi nelle misure del tempo in ore, minuti e secondi, e nelle misure degli angoli in gradi, primi e secondi.

L'introduzione in Europa della notazione posizionale in base 10 viene fatta risalire agli arabi, che nel periodo attorno al 1000 avrebbero perfezionato questa notazione già abbastanza diffusa in India. Un ruolo notevole

nella diffusione dell'aritmetica basata sulla notazione posizionale fu svolto dal Fibonacci (Leonardo Pisano) con il suo *Liber Abaci* del 1202. La notazione posizionale in base 10 fu però usata per alcuni secoli solo per i numeri interi, mentre la notazione frazionaria in base 60 fu usata fin verso il 1500. Le frazioni decimali non divennero di uso comune in Europa fino al 17^o secolo. Nella pratica varie unità di misura non decimali furono usate ancora per molto tempo e in certi paesi vengono utilizzate ancora oggi.

La notazione in base 2 venne usata per la prima volta nel 1605 da Harriot, anche se la data di origine della numerazione binaria è il 1703, anno in cui Leibnitz pubblicò un articolo dove veniva completamente definita l'aritmetica in base 2. Nel 1658 Pascal riconobbe che l'uso di una certa base in una rappresentazione è puramente convenzionale e suggerì l'adozione della base 12, che avrebbe consentito alcune semplificazioni nella divisione dei numeri. Nei secoli successivi si sviluppò un'ampia letteratura sull'uso di basi diverse, che però restò puramente teorica fino all'avvento dei calcolatori. Nei calcolatori l'aritmetica in base 2 prevalse su quella in base 10 su suggerimento di Von Neumann, che nel 1946 pubblicò con Burks e Goldstine un articolo in cui si definirono le proprietà dell'aritmetica di macchina. Recentemente sono stati proposti come basi anche numeri interi negativi o addirittura complessi, con i quali è possibile definire aritmetiche che godono di proprietà diverse rispetto a quella usuale (si veda il libro di Knuth [13]).

Sui primi calcolatori fu implementata l'aritmetica in virgola fissa, perché è più semplice da realizzare e richiede meno memoria di quella in virgola mobile. L'uso di un'aritmetica in virgola fissa richiede il costante controllo dell'utente, che è così in grado anche di verificare l'accuratezza dei calcoli svolti. Invece l'aritmetica in virgola mobile, implementata sui calcolatori odierni, è di uso più agevole, ma rende più difficile il coltrollo dell'accuratezza. Con l'aumento del volume di calcolo aumenta il rischio che si generino grossi errori, fenomeno ignorato da molti utilizzatori. Già nell'800 Gauss aveva studiato in modo sistematico gli errori di arrotondamento e la loro propagazione nella costruzione e nell'uso di tabelle. Gauss poté quindi osservare che formule, ottime da altri punti di vista, non potevano essere utilizzate perché instabili e individuò nel fenomeno della cancellazione la causa di questa instabilità.

Un'analisi degli errori generati in un'aritmetica in virgola mobile fu condotta già nel 1953 da Bauer e Samelson [3]. Fondamentale è stato il contributo di Wilkinson che nel 1960 [17] introdusse la tecnica dell'analisi dell'errore all'indietro, detta *backward error analysis*, per studiare la propagazione degli errori negli algoritmi dell'algebra lineare. Il libro di Wilkinson, *Rounding Errors in Algebraic Processes*, del 1963 [18] è ancora oggi di grande importanza nello studio della propagazione dell'errore. Per una presentazione dei concetti di condizionamento e stabilità, per applicazioni

dell'analisi inversa dell'errore agli algoritmi dell'algebra lineare e per commenti bibliografici, si veda il capitolo 4 di [4]. Il libro di Sterbenz [15] presenta una trattazione completa dei sistemi floating-point, una descrizione delle proprietà dell'aritmetica di macchina e delle difficoltà che l'utente può incontrare implementando metodi numerici, con particolare attenzione ai calcolatori IBM serie /370. Per l'uso dei grafi nella descrizione degli algoritmi e nello studio della propagazione degli errori si veda [2]. Un testo didattico sullo studio dell'errore condotto con l'uso dei grafi è quello di Dorn e Mc Cracken [8].

Le nozioni elementari di calcolo delle probabilità che stanno alla base dello studio probabilistico dell'errore di arrotondamento sono riportate in [7] e [10]. Si veda anche il paragrafo corrispondente nel libro di Henrici [9]. Il fatto che la distribuzione delle cifre dei numeri in virgola mobile sia logaritmica anziché uniforme fu rilevata per la prima volta da Newcomb nel 1881 e successivamente da Benford nel 1938. Per un'accurata interpretazione del fenomeno si veda [13].

Un'introduzione, con indicazioni bibliografiche, all'analisi automatica dell'errore è fatta in [15]. L'aritmetica degli intervalli è stata studiata estensivamente da Moore a partire dal 1965 [14]. Si veda anche [1] e [12]. Stime statistiche degli errori per mezzo dell'aritmetica degli intervalli si trovano in [16]. Sull'aritmetica degli intervalli è basato il sistema ACRITH, recentemente prodotto dalla IBM [11]. Tale sistema fornisce limitazioni automatiche dell'errore, che viene mantenuto dell'ordine della precisione di macchina. Su aritmetiche in precisione variabile si basano il sistema MP definito da Brent [5] e il recente sistema Mathematica [19].

Bibliografia

[1] G. Alefeld, J. Herzberger, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] F. L. Bauer, "Computational Graphs and Rounding Error", *SIAM J. Numer. Anal.*, 11, 1974, pp. 87-96.

[3] F. L. Bauer, K. Samelson "Optimale Rechengenauigkeit bei Rechenanlagen mit gleitendem Komma", *Zeitschrift für Angewandte Math. und Physik*, 4, 1953, pp. 312-316.

[4] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, Bologna, 1988.

[5] R. P. Brent, "A Fortran Multiple-Precision Arithmetic Package", *ACM Trans. Math. Soft.*, 4, 1978, pp. 57-70.

[6] G. Dahlquist, Å. Björk, N. Anderson, *Numerical Methods*, Prentice Hall, Englewood Cliffs, N. J., 1974.

[7] G. Dall'Aglio, *Calcolo delle probabilità*, Zanichelli, Bologna, 1987.

[8] W. S. Dorn, D. D. Mc Cracken, *Numerical Methods and FORTRAN Programming, with Applications in Engineering and Science*, J. Wiley & Sons, New York, 1964.

[9] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, J. Wiley & Sons, New York, 1962.

[10] P. Hoel, S. Port, C. Stone, *Introduction to Probability Theory*, Houghton Mifflin, Boston, 1971.

[11] IBM High-Accuracy Arithmetic Subroutine Library, General Information Manual, Order No. GC 33-6163-02, 1986.

[12] D. Jacobs, *The State of the Art in Numerical Analysis*, Academic Press, New York, 1977.

[13] D. E. Knuth, *The Art of Computer Programming, vol. 2, Seminumerical Algorithms*, Addison-Wesley, Reading, Mass., 1969.

[14] R. E. Moore, *Interval Analysis*, Prentice Hall, Englewood Cliffs, N. J., 1966.

[15] P. H. Sterbenz, *Floating-Point Computation*, Prentice Hall, Englewood Cliffs, N. J., 1974.

[16] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[17] J. H. Wilkinson, "Error Analysis of Floating-Point Computation", *Numer. Math.*, 2, 1960, pp. 219-340.

[18] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, N. J., 1963.

[19] S. Wolfram, *Mathematica, A System for Doing Mathematics by Computer*, Addison-Wesley, Reading, Mass., 1988.

Capitolo 3

EQUAZIONI E SISTEMI NON LINEARI

Il calcolo delle soluzioni di un'equazione

$$f(x) = 0 \tag{1}$$

è uno dei più importanti problemi della matematica applicata. In generale non sono disponibili formule esplicite per calcolare le soluzioni di (1), per cui si deve ricorrere a metodi iterativi che consentano di approssimare le soluzioni con una precisione prestabilita.

1. Metodo di bisezione

Sia $f(x) \in C[a,b]$ tale che $f(a)f(b) < 0$. Allora esiste almeno una soluzione α di (1), $a < \alpha < b$. Il metodo più immediato per approssimare α è il metodo di bisezione, che procede suddividendo, ad ogni passo, l'intervallo $[a,b]$ a metà e determinando in quale dei due sottointervalli si trova la soluzione: l'ampiezza dell'intervallo che contiene α risulta così dimezzata ogni volta.

Si pone $a_0 = a,\ b_0 = b$. Per $i = 1, 2, \ldots$ si calcolano

$$c_i = \frac{a_{i-1} + b_{i-1}}{2} \quad \text{e} \quad f(c_i);$$

se $f(a_{i-1})f(c_i) < 0$, si pone $a_i = a_{i-1}$, $b_i = c_i$;

se $f(a_{i-1})f(c_i) > 0$, si pone $a_i = c_i$, $b_i = b_{i-1}$;

se $f(c_i) = 0$, è $\alpha = c_i$.

Se la condizione $f(c_i) = 0$ non è mai verificata, il procedimento viene interrotto utilizzando opportuni *criteri di arresto*: uno di tali criteri è che l'ampiezza dell'intervallo sia sufficientemente piccola, cioè che risulti

$$b_i - a_i < \epsilon, \tag{2}$$

dove $\epsilon > 0$ è una costante prefissata.

La condizione (2), che è verificata certamente per $i > \log_2\left(\dfrac{b-a}{\epsilon}\right)$,

garantisce che α è approssimata da c_i con un errore assoluto minore in modulo di ϵ.

Nell'ipotesi che sia $0 \notin [a, b]$, al posto della (2), si può usare la condizione

$$\frac{b_i - a_i}{\min(|a_i|, |b_i|)} < \epsilon, \tag{3}$$

che garantisce che la soluzione α è approssimata da c_i con un errore relativo minore in modulo di ϵ. Un altro possibile criterio di arresto è dato dalla relazione

$$|f(c_i)| < \epsilon. \tag{4}$$

Si osservi che la costante ϵ non può essere scelta arbitrariamente piccola, perché, a causa degli errori di arrotondamento, le condizioni (2), (3) e (4) possono non essere mai soddisfatte.

3.1 Esempio. Nella figura 3.1 è riportato il grafico della funzione

$$f(x) = x^3 + 4x \cos x - 2$$

nell'intervallo [0,1].

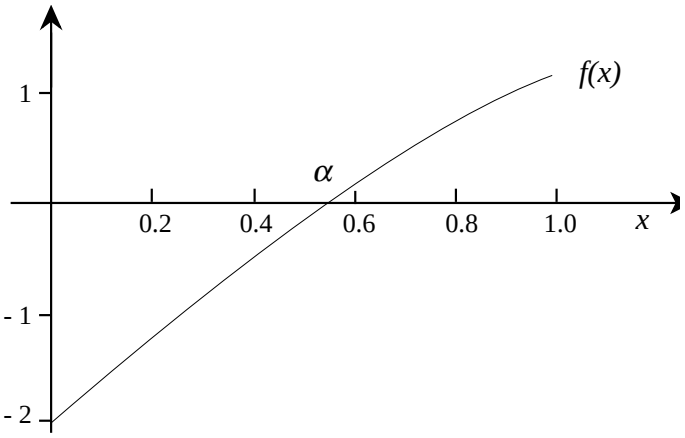


Fig. 3.1 - Grafico della funzione $f(x) = x^3 + 4x \cos x - 2$.

Poiché $f'(x) > 0$ per ogni $x \in [0, 1]$ e $f(0)f(1) < 0$, l'equazione $f(x) = 0$ ha una sola soluzione reale $\alpha \in (0, 1)$. Applicando il metodo di bisezione si ha

i	c_i	$f(c_i)$
1	0.5000000	-0.1198349
2	0.7500000	0.6169415
3	0.6250000	0.2715483
4	0.5625000	$0.8130836\ 10^{-1}$
5	0.5312500	$-0.1794720\ 10^{-1}$
6	0.5468750	$0.3201580\ 10^{-1}$
7	0.5390625	$0.7117271\ 10^{-2}$
8	0.5351563	$-0.5393982\ 10^{-2}$
9	0.5371094	$0.8668900\ 10^{-3}$
10	0.5361328	$-0.2263069\ 10^{-2}$

L'ampiezza dell'intervallo, che inizialmente è 1, dopo 10 passi è ridotta a $2^{-10} < 0.00098$ e quindi $c_{10} = 0.5361328$ approssima α con un errore assoluto minore di 10^{-3}. ∎

Per l'applicabilità di questo metodo è richiesta solo la continuità della funzione. Tuttavia il numero di passi richiesti per raggiungere una prefissata precisione risulta in generale molto elevato: nell'esempio 3.1, dopo 10 passi, e quindi dopo 10 valutazioni della funzione f, l'errore assoluto è ancora dell'ordine di 10^{-3}.

2. Metodi di iterazione funzionale

Vi sono molti altri metodi per approssimare la soluzione dell'equazione (1) con un numero di passi minore, a parità di precisione, di quello richiesto dal metodo di bisezione. In generale si tratta di metodi iterativi, detti *metodi di iterazione funzionale*, della forma

$$x_{i+1} = g(x_i), \quad i = 0, 1, \ldots \tag{5}$$

con cui, a partire da un valore iniziale x_0, è possibile approssimare le soluzioni dell'equazione

$$x = g(x). \tag{6}$$

Le soluzioni di (6) sono anche dette *punti fissi* della funzione $g(x)$.

L'equazione (1) deve quindi essere prima trasformata in un'equazione equivalente della forma (6). Uno dei modi con cui questa trasformazione può essere fatta è quello di utilizzare una funzione $h(x)$ tale che l'equazione

$$x = x - \frac{f(x)}{h(x)} \tag{7}$$

abbia le stesse soluzioni della (1) in un opportuno intervallo contenente la soluzione che si vuole approssimare.

3.2 Teorema. *Sia $g(x) \in C[a,b]$ e la successione $\{x_i\}$, definita dalla (5) a partire da un punto iniziale x_0, sia tale che $x_i \in [a,b]$, $i = 0, 1, \ldots$. Se la successione converge, allora il limite*

$$\alpha = \lim_{i\to\infty} x_i$$

è un punto fisso di $g(x)$, e quindi è soluzione della (6).

Dim. Per la continuità di $g(x)$ è

$$\alpha = \lim_{i\to\infty} x_{i+1} = \lim_{i\to\infty} g(x_i) = g(\lim_{i\to\infty} x_i) = g(\alpha). \qquad \blacksquare$$

Il seguente teorema dà delle condizioni sufficienti di convergenza, nel caso che $g(x)$ sia derivabile. Per la convergenza nel caso di ipotesi più deboli si veda l'esercizio 3.2.

3.3 Teorema. *Sia α punto fisso di $g(x)$, $g(x) \in C^1[\alpha-\rho, \alpha+\rho]$, $\rho > 0$. Scelto x_0 tale che sia*

$$|x_0 - \alpha| \le \rho,$$

si consideri la successione $\{x_i\}$, $i = 0, 1, \ldots$, definita dalla (5). Se

$$|g'(x)| < 1, \quad per \quad |x - \alpha| \le \rho, \tag{8}$$

allora $|x_i - \alpha| \le \rho$ per $i = 0, 1, \ldots$, e $\lim_{i\to\infty} x_i = \alpha$.

Dim. Sia $\lambda = \max_{|x-\alpha|\le\rho} |g'(x)|$. Dalla (8) risulta $\lambda < 1$. Si dimostra per induzione che

$$|x_i - \alpha| \le \lambda^i \rho \le \rho. \tag{9}$$

Per $i = 0$ la (9) è vera. Per $i > 0$, per l'ipotesi induttiva è

$$|x_{i-1} - \alpha| \le \lambda^{i-1}\rho \le \rho;$$

per il teorema del valor medio, dalla (5) si ha:

$$x_i - \alpha = g(x_{i-1}) - g(\alpha) = g'(\xi_{i-1})(x_{i-1} - \alpha), \quad |\xi_{i-1} - \alpha| < \rho, \tag{10}$$

e quindi

$$|g'(\xi_{i-1})| \le \lambda < 1.$$

Dalla (10), passando ai moduli, risulta:

$$|x_i - \alpha| = |g'(\xi_{i-1})|\ |x_{i-1} - \alpha| \le \lambda^i \rho,$$

e dalla (9), passando al limite, si ottiene:

$$\lim_{i\to\infty} |x_i - \alpha| = 0. \qquad \blacksquare$$

Un metodo iterativo (5) si dice *(localmente) convergente* ad una soluzione α di (6) se esiste un intervallo $[a, b]$ contenente α tale che per ogni punto $x_0 \in [a, b]$ la successione (5) risulti convergente ad α.

3.4 Teorema. *Sia $\alpha \in [a, b]$ un punto fisso di $g(x) \in C^1[a, b]$. Se $|g'(x)| < 1$ per $x \in [a, b]$, allora α è l'unica soluzione della (6) in $[a, b]$.*

Dim. Sia, per assurdo, $\beta \in [a, b]$, $\beta = g(\beta)$, $\beta \neq \alpha$. Allora esiste $\xi \in [a, b]$ tale che

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| = |g'(\xi)||\alpha - \beta| < |\alpha - \beta|.$$

Ne segue che $\alpha = \beta$. ∎

Del teorema 3.3 si può dare l'interpretazione geometrica illustrata nelle figure 3.2-3.5 per i quattro casi:

$$\begin{aligned} g'(\alpha) &< -1, \\ -1 < g'(\alpha) &< 0, \\ 0 < g'(\alpha) &< 1, \\ 1 < g'(\alpha). & \end{aligned}$$

La soluzione α è l'ascissa del punto di intersezione dei grafici delle funzioni

$$\begin{aligned} y &= x, \\ y &= g(x). \end{aligned}$$

Sull'asse delle ascisse sono riportati i primi punti x_i calcolati. Nei casi illustrati dalle figure 3.3 e 3.4 si ha convergenza, mentre nei casi illustrati dalle figure 3.2 e 3.5 non si ha convergenza.

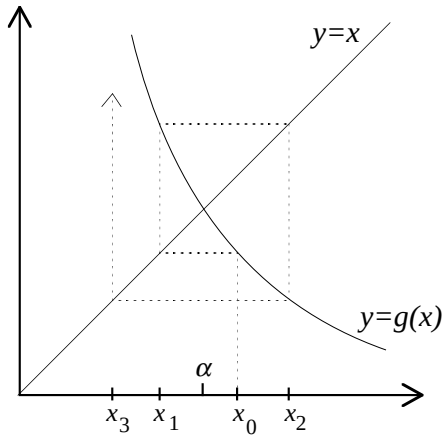


Fig. 3.2 $g'(\alpha) < -1$

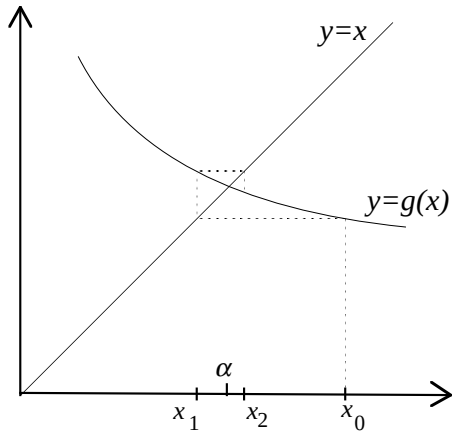


Fig. 3.3 $-1 < g'(\alpha) < 0$

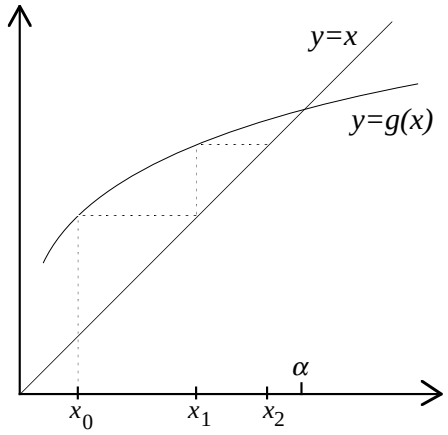


Fig. 3.4 $0 < g'(\alpha) < 1$

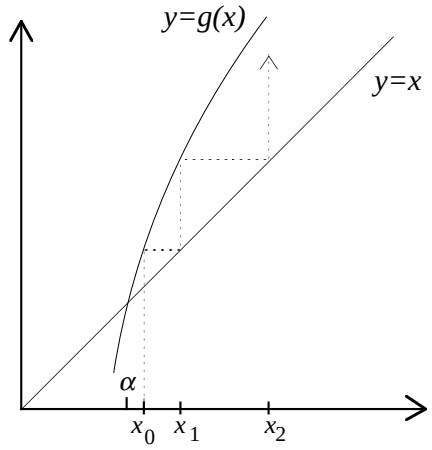


Fig. 3.5 $1 < g'(\alpha)$

Nelle ipotesi del teorema 3.3, se la funzione $g'(x)$ non cambia segno nell'intervallo $[\alpha - \rho, \alpha + \rho]$, si possono verificare i due casi seguenti:

a) se $0 < g'(x) < 1$ (figura 3.4), la successione ottenuta è monotona crescente se $x_0 < \alpha$, decrescente se $x_0 > \alpha$;

b) se $-1 < g'(x) < 0$ (figura 3.3), la successione ottenuta ha elementi alternativamente maggiori e minori di α: ed è quindi costituita da due sottosuccessioni monotone, una crescente e l'altra decrescente e verrà chiamata *successione alternata*.

Se $g'(x)$ cambia segno in $[\alpha-\rho, \alpha+\rho]$ e $g'(\alpha) \neq 0$, si verifica localmente in un intorno di α uno dei due casi a) o b). Se $g'(\alpha) = 0$, il comportamento della successione può non essere né monotono né alternato (si veda l'esercizio 3.3).

Poiché α non è noto, l'ipotesi che la condizione (8) sia verificata in un intorno circolare $[\alpha - \rho, \alpha + \rho]$ di centro α può apparire restrittiva. Se però la condizione (8) è verificata in un intervallo chiuso $[a, b]$ contenente α nella sua parte interna, le ipotesi del teorema 3.3 sono verificate nel massimo intorno circolare chiuso di α contenuto in $[a, b]$, e la convergenza è certamente assicurata se il valore iniziale x_0 è l'estremo dell'intervallo più vicino ad α.

Se nell'intervallo $[a, b]$ è $g'(x) > 0$, poiché in tal caso la successione è monotona, si può scegliere come valore iniziale indifferentemente uno dei due estremi. Inoltre in questo caso è sufficiente che la condizione (8) sia verificata in $[a, \alpha)$, assumendo $x_0 = a$, oppure in $(\alpha, b]$, assumendo $x_0 = b$.

Se nell'intervallo $[a, b]$ è $g'(x) < 0$, la successione risulta alternata e in tal caso, se non è possibile stabilire quale dei due estremi sia il più vicino ad α, si può scegliere come valore iniziale, ad esempio, $x_0 = a$: se $x_1 \in [a, b]$,

la successione converge, se invece $x_1 \notin [a, b]$, basta scegliere $x_0 = b$, per ottenere certamente una successione convergente.

Se nell'intervallo $[a, b]$ la $g'(x)$ cambia segno è opportuno studiare più accuratamente la funzione $g(x)$ (si veda l'esercizio 3.3).

3.5 Esempio. L'equazione $f(x) = x^3 + 4x\cos x - 2 = 0$ dell'esempio 3.1 può essere trasformata in un'equazione equivalente della forma (6) in molti modi diversi. Nell'intervallo $[0, 1]$ si può considerare ad esempio l'equazione

$$x = \frac{2 - x^3}{4\cos x},$$

che corrisponde a porre nella (7)

$$h(x) = 4\cos x.$$

In questo caso risulta

$$g(x) = \frac{2 - x^3}{4\cos x} \quad \text{e} \quad g'(x) = \frac{(2 - x^3)\sin x - 3x^2\cos x}{4\cos^2 x}.$$

Poiché nell'intervallo $[0, 1]$, come si può verificare graficamente, è $|g'(x)| \leq |g'(1)| < 0.7$, in ogni intorno circolare $[\alpha - \rho, \alpha + \rho]$ contenuto nell'intervallo $[0, 1]$ la condizione (8) è verificata. Perciò il metodo iterativo

$$x_{i+1} = \frac{2 - x_i^3}{4\cos x_i}, \quad i = 0, 1, \ldots,$$

è convergente. Poiché $g'(x) > 0$ nell'intervallo $[0, 0.6]$ contenente α, partendo dal punto $x_0 = 0$, si ottiene la successione monotona

i	x_i	i	x_i
1	0.5000000	4	0.5368258
2	0.5341377	5	0.5368375
3	0.5366538	6	0.5368385

Il procedimento è stato arrestato quando in due iterazioni successive risultano uguali le 5 cifre decimali più significative. Scegliendo come valore iniziale $x_0 = 1$ si ottiene $x_1 = 0.4627039$, che è minore di α (infatti $g'(x)$ cambia segno nell'intervallo $[0.6, 1]$). A partire da x_1 la successione risulta monotona.

Se l'equazione data viene trasformata nell'equazione equivalente

$$x = \frac{2 - 4x\cos x}{x^2}, \quad x \neq 0,$$

le condizioni di convergenza del teorema 3.3 non sono verificate. Infatti

$$g(x) = \frac{2 - 4x\cos x}{x^2}\,, \quad g'(x) = \frac{4}{x^3}\,(x\cos x + x^2 \sin x - 1)$$

e $|g'(x)| > 1$ per $x \in (0, 0.7]$, intervallo a cui appartiene α. Partendo da $x_0 = 0.7$ si ottiene la successione

i	x_i	i	x_i
1	-0.2888913	4	265.1611
2	37.23634	5	$-0.4479170\ 10^{-2}$
3	$-0.9468085\ 10^{-1}$	6	100579.1

che non è convergente. ∎

Il metodo iterativo (5) è convergente anche in ipotesi più deboli di quelle richieste dal teorema 3.3 (si veda l'esercizio 3.1). Alla (8) si può infatti sostituire la condizione

$$|g'(x)| < 1, \quad \text{per } 0 < |x - \alpha| < \rho$$

(per continuità nei punti $x = \alpha - \rho$, α e $\alpha + \rho$ è $|g'(x)| \leq 1$). Si noti però che se $|g'(\alpha)|$ è uguale a 1 o ha un valore molto vicino a 1, le successioni generate dal metodo convergono lentamente.

3.6 Esempio. L'equazione

$$x = \sin x$$

ha la soluzione $\alpha = 0$. Il metodo iterativo

$$x_{i+1} = g(x_i) = \sin x_i, \quad i = 0, 1, \ldots,$$

è convergente per ogni x_0 appartenente ad un intorno circolare di α di raggio $\rho \leq \pi$. Si ha infatti $|g'(x)| \leq 1$ per $x \in [-\pi, \pi]$, $g'(\alpha) = 1$ e $g'(-\pi) = g'(\pi) = -1$. Ponendo $x_0 = 1$, la successione $\{x_i\}$ converge molto lentamente, come si vede anche dalla tabella:

i	x_i	i	x_i
1	0.8414710	.	...
2	0.7456242	97	0.1713247
3	0.6784305	98	0.1704878
4	0.6275718	99	0.1696630
5	0.5871809	.	...
.	...	.	...

∎

3. Criteri di arresto

I criteri per decidere a quale iterazione arrestare un metodo iterativo convergente sono vari e dipendono da una tolleranza prefissata $\epsilon > 0$, legata alla precisione con cui si vuole approssimare la soluzione:

$$|x_{i+1} - x_i| < \epsilon, \tag{11}$$

$$\frac{|x_{i+1} - x_i|}{\min(|x_i|, |x_{i+1}|)} < \epsilon, \quad |x_i|, |x_{i+1}| \neq 0, \tag{12}$$

$$|f(x_i)| < \epsilon. \tag{13}$$

Le condizioni (11) e (12) controllano rispettivamente l'errore assoluto e l'errore relativo di x_i rispetto alla soluzione α attraverso la differenza fra le due iterate successive x_i e x_{i+1}, mentre la condizione (13) controlla l'approssimazione della soluzione attraverso i valori della funzione.

A seconda del criterio scelto, si ottengono limitazioni diverse per l'errore commesso. Infatti, se $g(x) \in C^1[a, b]$, per ogni i esiste uno ξ tale che

$$x_{i+1} - \alpha = g'(\xi)(x_i - \alpha), \quad |\xi - \alpha| < |x_i - \alpha|,$$

da cui

$$x_{i+1} - x_i = (x_{i+1} - \alpha) - (x_i - \alpha) = (x_i - \alpha)(g'(\xi) - 1).$$

Se si usa la condizione (11) si ha

$$|x_i - \alpha| = \frac{|x_{i+1} - x_i|}{|g'(\xi) - 1|} < \frac{\epsilon}{|g'(\xi) - 1|}\,; \tag{14}$$

se si usa la condizione (12), si ha

$$\frac{|x_i - \alpha|}{|\alpha|} = \frac{|x_{i+1} - x_i|}{|\alpha||g'(\xi) - 1|} < \frac{\epsilon \min(|x_i|, |x_{i+1}|)}{|\alpha||g'(\xi) - 1|}\,. \tag{15}$$

Da (14) e (15) segue che, per un dato ϵ, l'errore può risultare tanto più grande quanto più $g'(\xi)$ è vicino a 1. Questo si verifica, in particolare, nel caso di soluzioni di molteplicità maggiore di 1 dell'equazione (6), per cui si ha $g'(\alpha) = 1$. Se è $g'(\alpha) < 0$, per i abbastanza grande risulta $g'(\xi) < 0$; allora quando si usa la (11), dalla (14) segue che

$$|x_i - \alpha| < \epsilon,$$

e quando si usa la (12) dalla (15) segue che

$$\frac{|x_i - \alpha|}{|\alpha|} < \epsilon \, \frac{\min(|x_i|, |x_{i+1}|)|}{|\alpha|} \sim \epsilon.$$

Quindi, se la successione $\{x_i\}$ è alternata, è $g'(x) \leq 0$ in un intorno di α e le condizioni (11) e (12) danno dei buoni criteri di arresto. Invece se la successione è monotona e $g'(\alpha)$ è vicino ad 1, non è sempre soddisfacente usare (11) o (12).

Se si usa la (13), poiché per ogni i esiste un η tale che

$$f(x_i) = f(x_i) - f(\alpha) = f'(\eta)(x_i - \alpha), \quad |\eta - \alpha| < |x_i - \alpha|,$$

si ha

$$|x_i - \alpha| = \frac{|f(x_i)|}{|f'(\eta)|} < \frac{\epsilon}{|f'(\eta)|} \, . \tag{16}$$

Dalla (16) segue che l'errore assoluto può risultare tanto più grande quanto più piccolo è $|f'(\eta)|$.

3.7 Esempio. Nel caso dell'equazione $x = \sin x$ dell'esempio 3.6 si ha $g(x) = \sin x$ e $g'(x) = \cos x > 0$ in un intorno di $\alpha = 0$. Se si usa come criterio di arresto la (11) con $\epsilon = 10^{-5}$, l'iterazione si arresta all'indice $i = 1950$ e si ha

$$x_i = 0.03914304,$$
$$x_{i+1} = 0.03913304,$$

mentre $x_i - \alpha = x_i \approx 0.391 \; 10^{-1}$. Ciò è dovuto al fatto che quando x_i è vicino ad α, $g'(x)$ è positiva e molto vicina ad 1. Dal grafico della figura 3.6 risulta chiaramente come la differenza fra x_i e x_{i+1} sia molto piccola pur essendo x_i lontano da α.

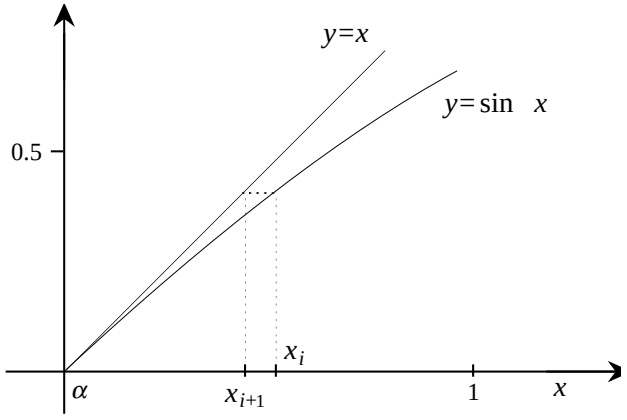


Fig. 3.6 - Interpretazione geometrica del metodo $x_{i+1} = \sin x_i$.

Si osservi che in questo caso, imporre la condizione (13) è del tutto equivalente a imporre la (11), perché

$$f(x_i) = x_i - \sin x_i = x_i - x_{i+1}.$$

Se si considera invece l'equazione

$$x = \cos x,$$

il problema di approssimare la soluzione $\alpha \in \left(0, \frac{\pi}{2}\right)$ è molto più facile da risolvere. Infatti in tale intervallo è $g'(x) = -\sin x$, per cui $-1 < g'(x) < 0$. La successione $\{x_i\}$ definita da

$$\begin{aligned} &x_0 = 1 \\ &x_{i+1} = \cos x_i, \quad i = 0, 1, \ldots \end{aligned}$$

converge ad α (si veda anche la figura 3.7).

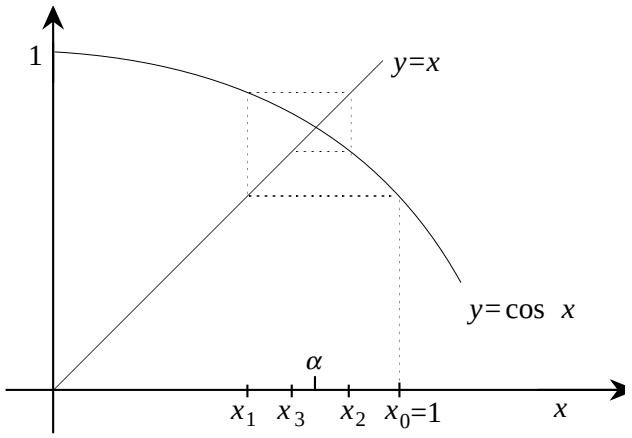


Fig. 3.7 - Interpretazione geometrica del metodo $x_{i+1} = \cos x_i$.

Usando il criterio (11) con $\epsilon = 10^{-5}$, l'iterazione si arresta all'indice $i = 28$ e si ha

$$\begin{aligned} x_i &= 0.7390894, \\ x_{i+1} &= 0.7390823, \end{aligned}$$

mentre in realtà è $|x_{i+1} - \alpha| \approx 0.286\ 10^{-5}$. ∎

Oltre ai criteri descritti, è necessario prevedere un numero massimo di iterazioni consentite, in modo che l'esecuzione di un programma che implementa un metodo iterativo non impieghi eccessive risorse di calcolo nei casi di convergenza lenta. Negli esempi che seguono verrà applicata come condizione di arresto la (12) con $\epsilon = 10^{-5}$, prevedendo un numero massimo di 99 iterazioni.

4. Effetto degli errori di arrotondamento

A causa degli errori di arrotondamento che si commettono nel calcolo di $g(x_i)$ la successione effettivamente calcolata $\{\widetilde{x}_i\}$ può non essere convergente anche quando sono soddisfatte le ipotesi del teorema 3.3. Tuttavia è possibile dimostrare che i valori $\widetilde{x}_i$ effettivamente calcolati appartengono ad intorni di α con raggio via via decrescente.

Sia δ_i l'errore assoluto introdotto nel calcolo effettivo alla i-esima iterazione:

$$\widetilde{x}_i = g(\widetilde{x}_{i-1}) + \delta_i,$$

tale che

$$|\delta_i| \leq \delta, \quad \text{per } i = 0, 1, \ldots$$

Vale il seguente

3.8 Teorema. *Sia α punto fisso di $g(x) \in C^1[\alpha - \rho, \alpha + \rho]$, $\rho > 0$, e sia*

$$\lambda = \max_{|x-\alpha| \leq \rho} |g'(x)| < 1.$$

Posto $\sigma = \dfrac{\delta}{1-\lambda}$, se $\sigma < \rho$, $|x_0 - \alpha| \leq \rho$ e $\widetilde{x}_0 = x_0$, allora si ha

$$|\widetilde{x}_i - \alpha| \leq \sigma + \lambda^i(\rho - \sigma), \quad \textit{per } i = 0, 1, \ldots \tag{17}$$

Dim. Si procede per induzione. Per $i = 0$, la disuguaglianza (17) vale. Per $i > 0$, si ha per l'ipotesi induttiva

$$|\widetilde{x}_{i-1} - \alpha| \leq \sigma + \lambda^{i-1}(\rho - \sigma) \leq \sigma + \rho - \sigma = \rho.$$

Quindi $\widetilde{x}_{i-1} \in [\alpha - \rho, \alpha + \rho]$, e per il teorema del valor medio

$$\begin{aligned} |\widetilde{x}_i - \alpha| &= |g(\widetilde{x}_{i-1}) + \delta_i - g(\alpha)| \leq |g(\widetilde{x}_{i-1}) - g(\alpha)| + \delta \\ &= |g'(\xi)||\widetilde{x}_{i-1} - \alpha| + \delta \end{aligned}$$

e poiché $|\xi - \alpha| < |\widetilde{x}_{i-1} - \alpha| \leq \rho$, si ha $|g'(\xi)| \leq \lambda$ e quindi

$$|\widetilde{x}_i - \alpha| \leq \lambda[\sigma + \lambda^{i-1}(\rho - \sigma)] + \sigma(1 - \lambda) = \sigma + \lambda^i(\rho - \sigma). \qquad \blacksquare$$

Il teorema 3.8 mostra che i valori $\widetilde{x}_i$ effettivamente calcolati possono non avvicinarsi arbitrariamente ad α, ma che, comunque si scelga un intorno di α di raggio maggiore di σ, la successione effettivamente calcolata appartiene definitivamente a tale intorno.

Il teorema permette anche di ottenere una relazione, analoga alla (17), per l'errore relativo. Sia

$$\widetilde{x}_i = g(\widetilde{x}_{i-1})(1+\epsilon_i)$$

e quindi

$$\delta_i = g(\widetilde{x}_{i-1})\epsilon_i\,;$$

se

$$|\epsilon_i| \le \epsilon \quad \text{e} \quad \max_{|x-\alpha|\le\rho} |g(x)| = M,$$

allora, posto $\delta = \epsilon M$, si ha dalla (17)

$$\frac{|\widetilde{x}_i - \alpha|}{|\alpha|} \le \frac{\sigma}{|\alpha|} + \lambda^i\Big(\frac{\rho}{|\alpha|} - \frac{\sigma}{|\alpha|}\Big).$$

Risulta quindi che la quantità

$$\frac{\sigma}{|\alpha|} = \frac{\epsilon M}{|\alpha|(1-\lambda)}$$

è una misura della "incertezza" con cui è possibile determinare la soluzione per effetto degli errori di arrotondamento. Di questo è opportuno tenere conto nella scelta della tolleranza ϵ per le condizioni di arresto, perché se la tolleranza è troppo piccola le condizioni imposte possono non essere verificate per alcun indice di iterazione i.

3.9 Esempio. L'equazione $x = g(x)$, dove

$$g(x) = \frac{\cos \pi x + 1}{(x^2-1)^2} + \frac{1}{4}\,(5x-6), \tag{18}$$

ha una soluzione α reale nell'intervallo (0.9,1), come si vede anche dalla figura 3.8.

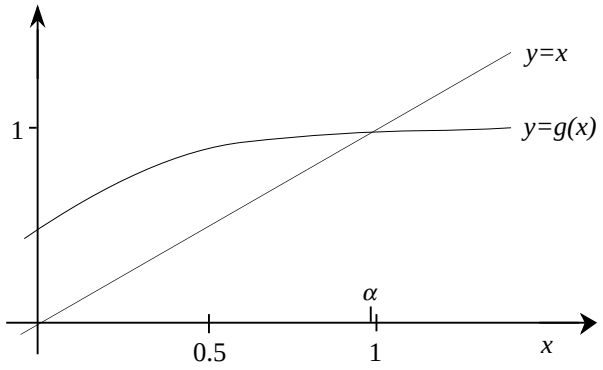


Fig. 3.8 - Grafico della funzione (18).

Poiché nell'intervallo (0.9, 1) è $0 < g'(x) < 1$, la successione ottenuta scegliendo $x_0 = 0.9$ è monotona e convergente ad α. I valori effettivamente calcolati di x_i sono riportati nella seguente tabella.

i	x_i	i	x_i
1	0.9807718	10	0.9834051
2	0.9833469	11	0.9834011
3	0.9833951	12	0.9834473
4	0.9833703	13	0.9833860
5	0.9834147	14	0.9833753
.	...	15	0.9834018

Esaminando la successione ottenuta, appare evidente che la soluzione α non può essere approssimata con molta accuratezza, poiché la quarta cifra continua a variare in tutte le iterate calcolate. Infatti nel calcolo di $g(x_i)$ sono presenti, per x_i vicino a 1, elevati errori di cancellazione, per cui ϵ risulta elevato e quindi anche la quantità $\sigma/|\alpha|$ risulta elevata (dell'ordine di 10^{-4}) nonostante λ sia minore di 0.05. Possono comunque esistere valori $\widetilde{x}_i$ che risultano essere delle buone approssimazioni di α. In questo esempio è $\alpha = 0.9834036$ e il valore fra quelli calcolati che meglio l'approssima è $\widetilde{x}_{10} = 0.9834051$.

Per λ vicino a 1, il fenomeno risulta ancora più evidente, come si rileva nel caso dell'equazione $x = g(x)$, dove

$$g(x) = \frac{\cos \pi x + 1}{(x^2 - 1)^2} + \frac{1}{4}\,(x - 2).$$

Questa equazione ha una soluzione reale in (0.9, 1.1), e in tale intervallo $-0.99 < g'(x) < -0.95$. Anche in questo caso si hanno, per x vicino a 1, elevati errori di cancellazione, e pertanto si giustificano i seguenti risultati calcolati.

i	x_i	i	x_i
1	1.080771	.	...
2	0.9038827	111	1.001607
3	1.077070	112	0.9838431
4	0.9075370	113	0.9995706
5	1.073575	114	0.9626794
.	...	115	1.020272

■

5. Ordine di convergenza

Come si è visto nell'esempio 3.6 la successione $\{x_i\}$, anche se convergente, può convergere così lentamente da essere inutilizzabile per l'approssimazione di α. Per confrontare metodi iterativi diversi che approssimano la stessa soluzione α di una equazione $f(x) = 0$, si può considerare la velocità con cui le successioni ottenute convergono alla soluzione. Lo studio della velocità di convergenza viene generalmente ricondotto a quello dell'ordine di convergenza del metodo.

3.10 Definizione. Sia $\{x_i\}$ una successione convergente ad α e sia $x_i \neq \alpha$ per ogni i. Se esiste un numero reale $p \geq 1$, tale che

$$\lim_{i\to\infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|^p} = \gamma, \quad \text{con} \begin{cases} 0 < \gamma \leq 1 & \text{se } p = 1, \\ \gamma > 0 & \text{se } p > 1, \end{cases} \tag{19}$$

si dice che la successione ha *ordine di convergenza* p. La costante γ è detta *fattore di convergenza*.

Se $p = 1$ e $0 < \gamma < 1$, si dice anche che la convergenza è *lineare*,
se $p = 1$ e $\gamma = 1$, si dice anche che la convergenza è *sublineare*,
se $p > 1$, si dice anche che la convergenza è *superlineare*. ∎

3.11 Esempio. Per la successione dell'esempio 3.6 si ha

$$x_{i+1} = \sin x_i, \quad i = 0, 1, \ldots,$$

e poiché

$$\lim_{i\to\infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|} = \lim_{i\to\infty} \frac{\sin x_i}{x_i} = 1,$$

la convergenza è sublineare.

Invece per la successione dell'esempio 3.7 si ha

$$x_{i+1} = \cos x_i, \quad i = 0, 1, \ldots,$$

e poiché

$$\lim_{i\to\infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|} = \lim_{i\to\infty} \frac{|\cos x_i - \cos \alpha|}{|x_i - \alpha|} = \sin \alpha$$

e $0 < \sin \alpha < 1$, la convergenza è lineare. ∎

Dalla (19) segue che per un metodo lineare o superlineare esiste una costante $\beta > 0$ tale che per i sufficientemente grande vale la relazione

$$|x_{i+1} - \alpha| \leq \beta |x_i - \alpha|^p, \tag{20}$$

con $\beta < 1$ se $p = 1$. Dividendo per $|\alpha|$ si ha

$$\left|\frac{x_{i+1} - \alpha}{\alpha}\right| \leq \beta|\alpha|^{p-1}\left|\frac{x_i - \alpha}{\alpha}\right|^p . \tag{21}$$

Le relazioni (20) e (21) misurano la riduzione ad ogni iterazione degli errori assoluto e relativo, commessi approssimando α con x_{i+1}. È evidente che, quando questi errori sono in modulo minori di 1, tale riduzione è tanto più elevata quanto maggiore è l'ordine di convergenza, e, a parità di ordine di convergenza, quanto minore è il fattore di convergenza.

Esistono però successioni $\{x_i\}$ per cui, pur valendo la (20), non si può determinare l'ordine di convergenza secondo la definizione 3.10. Si dice allora in modo più generale che la successione ha ordine di convergenza p se

$$\max \lim_{i\to\infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|^p} \quad \text{e} \quad \min \lim_{i\to\infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|^p}$$

sono finiti e non nulli (se $p = 1$, i due limiti devono essere minori o uguali a 1). Si dice invece che la successione ha *convergenza di ordine almeno p* se vale la (20).

3.12 Esempio. Si consideri la successione $\{x_i\}$ definita da

$$x_{i+1} = g(x_i), \quad g(x) = \frac{1}{3}\, x^3(\sin\frac{1}{x} + 2),$$

a partire da un punto $x_0 \neq 0$ e tale che $|x_0| < \dfrac{1}{2}$. L'equazione

$$x = \begin{cases} g(x) & \text{per } x \neq 0, \\ \\ 0 & \text{per } x = 0, \end{cases}$$

ha la sola soluzione nulla e per ogni x_0 la successione $\{x_i\}$, formata da termini dello stesso segno di x_0, converge a 0 perché

$$|x_{i+1}| = \frac{1}{3}\,|x_i^3|\,(\sin\frac{1}{x_i} + 2) \leq |x_i^3| < |x_i|, \quad \text{per } |x_i| < \frac{1}{2}\,.$$

Non esiste diverso da 0 il limite (19) per nessun valore di p: infatti per $p < 3$ il limite è nullo, per $p > 3$ il limite è infinito e per $p = 3$ il limite non esiste. Però la successione ha ordine di convergenza 3 secondo la definizione più generale. Assumendo $x_0 = 0.5$ si ottiene la successione

i	x_i	i	x_i
1	0.1212207	3	$0.1802611\ 10^{-8}$
2	$0.1735453\ 10^{-2}$	4	$0.3418121\ 10^{-26}$

Infatti il comportamento della successione è quello di una successione che ha ordine di convergenza 3. ∎

3.13 Teorema. *Sia $\{x_i\}$ una successione convergente ad α, ottenuta con il metodo iterativo (5), in cui la funzione $g(x) \in C^1[\alpha-\rho, \alpha+\rho]$, $\rho > 0$. Se la convergenza della successione $\{x_i\}$ è lineare (rispettivamente sublineare), allora $0 < |g'(\alpha)| < 1$ (rispettivamente $|g'(\alpha)| = 1$).*

Dim. Basta osservare che è

$$\frac{|x_{i+1} - \alpha|}{|x_i - \alpha|} = |g'(\xi)|, \quad |\xi - \alpha| < |x_i - \alpha|, \tag{22}$$

e, passando al limite, risulta $\gamma = |g'(\alpha)|$. ∎

Vale anche, viceversa, il

3.14 Teorema. *Sia $\alpha \in [a, b]$ un punto fisso di $g(x) \in C^1[a, b]$.*

a) Se $0 < |g'(\alpha)| < 1$, esiste $\rho > 0$ tale che per ogni x_0 per cui $0 < |x_0 - \alpha| \leq \rho$ la successione $\{x_i\}$ ottenuta con il metodo (5) è convergente e ha convergenza lineare;

b) se $|g'(\alpha)| = 1$ ed esiste $\rho > 0$ tale che $0 < |g'(x)| < 1$ per $0 < |x - \alpha| < \rho$, allora per ogni x_0 per cui $0 < |x_0 - \alpha| \leq \rho$, la successione $\{x_i\}$ ottenuta con il metodo (5) è convergente e ha convergenza sublineare.

Dim. Per il punto a) la convergenza segue dal teorema 3.3, in quanto se $|g'(\alpha)| < 1$ esiste un intorno di α in tutti i punti del quale è $|g'(x)| < 1$. Per la convergenza nel caso b) si veda l'esercizio 3.1. Inoltre nelle ipotesi fatte, per ogni successione $\{x_i\}$ vale la relazione (22) dalla quale, passando al limite per $i \to \infty$, si ha $\gamma = |g'(\alpha)| = 1$. ∎

3.15 Teorema. *Sia $\alpha \in [a, b]$ un punto fisso di $g(x) \in C^p[a, b]$, con $p \geq 2$ intero. Se per un punto $x_0 \in [a, b]$ la successione $\{x_i\}$ ottenuta con il metodo (5) è convergente con ordine di convergenza p, allora*

$$g'(\alpha) = g''(\alpha) = \ldots = g^{(p-1)}(\alpha) = 0, \quad g^{(p)}(\alpha) \neq 0.$$

Dim. Poiché la successione ha ordine p allora

$$\lim_{i\to\infty} \frac{|x_{i+1}-\alpha|}{|x_i-\alpha|^r} = 0, \quad \text{per } r < p.$$

Si dimostra per induzione su r che $g^{(r)}(\alpha) = 0$, per $r = 1, \ldots, p-1$. Per $r = 1$ dalla formula di Taylor si ha

$$\frac{x_{i+1}-\alpha}{x_i-\alpha} = g'(\alpha) + \frac{1}{2}\, g''(\xi_1)(x_i-\alpha), \quad |\xi_1-\alpha| < |x_i-\alpha|,$$

e poiché per $i \to \infty$ il limite del primo membro è nullo e $g''(\xi_1)$ è limitato, ne segue che $g'(\alpha) = 0$. Per $r > 1$ si supponga che

$$g'(\alpha) = g''(\alpha) = \ldots = g^{(r-1)}(\alpha) = 0, \quad \text{per } r < p.$$

Ne segue che

$$\frac{x_{i+1}-\alpha}{(x_i-\alpha)^r} = \frac{1}{r!}\, g^{(r)}(\alpha) + \frac{1}{(r+1)!}\, g^{(r+1)}(\xi_r)(x_i-\alpha), \quad |\xi_r-\alpha| < |x_i-\alpha|,$$

e, analogamente al caso $r = 1$, si può concludere che $g^{(r)}(\alpha) = 0$. Si ha quindi

$$g'(\alpha) = g''(\alpha) = \ldots = g^{(p-1)}(\alpha) = 0.$$

Inoltre, poiché

$$\frac{x_{i+1}-\alpha}{(x_i-\alpha)^p} = \frac{1}{p!}\, g^{(p)}(\xi), \quad |\xi-\alpha| < |x_i-\alpha|,$$

risulta

$$\lim_{i\to\infty} \frac{|x_{i+1}-\alpha|}{|x_i-\alpha|^p} = \frac{1}{p!}\, |g^{(p)}(\alpha)| \neq 0. \tag{23}$$

∎

Vale anche, viceversa, il

3.16 Teorema. *Sia $\alpha \in [a,b]$ un punto fisso di $g(x) \in C^p[a,b]$, con $p \geq 2$ intero. Se*

$$g'(\alpha) = g''(\alpha) = \ldots = g^{(p-1)}(\alpha) = 0,\ g^{(p)}(\alpha) \neq 0,$$

esiste $\rho > 0$ tale che per ogni x_0 per cui $0 < |x_0-\alpha| \leq \rho$, la successione $\{x_i\}$ ottenuta con il metodo (5) è convergente con ordine di convergenza p.

Dim. Poiché $g'(\alpha) = 0$, esiste $\rho > 0$ tale che $|g'(x)| < 1$ per $|x-\alpha| \leq \rho$, e la convergenza segue al teorema 3.3. Inoltre per ogni successione $\{x_i\}$ ottenuta, dalla (23) si ha che

$$\lim_{i\to\infty} \frac{|x_{i+1}-\alpha|}{|x_i-\alpha|^r} = \frac{1}{r!}\, |g^{(r)}(\alpha)| = 0, \quad \text{per} \quad r < p$$

e

$$\gamma = \frac{1}{p!}\,|g^{(p)}(\alpha)| > 0,$$

e quindi la successione ha ordine di convergenza p. ∎

Si osservi che l'ordine di convergenza p può essere anche un numero non intero: in tal caso, posto $q = \lfloor p \rfloor$, se $g(x) \in C^q[a,b]$, si ha che

$$g'(\alpha) = g''(\alpha) = \ldots = g^{(q)}(\alpha) = 0,$$

e che la funzione $g(x)$ non ha la derivata $(q+1)$-esima continua in α, altrimenti, per il teorema 3.16 tutte le successioni ottenute con il metodo iterativo (5) a partire da un punto iniziale x_0 tale che $0 < |x_0 - \alpha| \leq \rho$, avrebbero ordine almeno $q+1$.

3.17 Esempio. L'equazione

$$x = \sqrt{|x|^3}$$

ha soluzione $\alpha = 0$. La successione $\{x_i\}$ definita da

$$x_{i+1} = \sqrt{|x_i|^3}$$

è tale che $|x_{i+1}| < |x_i|$ per $|x_i| < 1$ e quindi è convergente per $x_0 \in (-1,1)$. Inoltre si ha

$$\frac{|x_{i+1}|}{|x_i|^{3/2}} = 1.$$

Perciò l'ordine di convergenza è $p = 3/2$. ∎

I teoremi 3.13 - 3.16 mettono in relazione l'ordine di convergenza della successione $\{x_i\}$ definita da $x_{i+1} = g(x_i)$ con proprietà della funzione $g(x)$ e in particolare con i valori assunti dalle sue derivate in α. Per questo motivo si può dare la seguente definizione.

3.18 Definizione. Un metodo iterativo convergente ad α si dice di *ordine* p (di *ordine almeno* p) se tutte le successioni ottenute al variare del punto iniziale in un intorno di α convergono con ordine di convergenza p (almeno p). ∎

6. Metodo delle corde

Ponendo nella (7) $h(x) = m$, $m \neq 0$, si ha

$$x = g(x) = x - \frac{f(x)}{m}\,.$$

Il corrispondente metodo è detto *metodo delle corde*:

$$x_{i+1} = x_i - \frac{f(x_i)}{m}\,, \quad i = 0, 1, \ldots \tag{24}$$

Se $f(x) \in C^1[a, b]$, dal teorema 3.3 si ha che la successione $\{x_i\}$ definita dalla (24) è convergente se

$$|g'(x)| = \left|1 - \frac{f'(x)}{m}\right| < 1 \tag{25}$$

in un intorno della soluzione $[\alpha - \rho, \alpha + \rho]$ in cui si sceglie il punto x_0. Dalla (25) si ottengono le seguenti condizioni sufficienti per la convergenza

$$\begin{aligned} &f'(x) \neq 0, \quad \text{per } |x - \alpha| \leq \rho, \\ &mf'(x) > 0, \\ &|m| > \frac{1}{2} \max_{|x-\alpha|\leq\rho} |f'(x)|. \end{aligned}$$

Se $m \neq f'(\alpha)$, il metodo delle corde è del primo ordine e se $m = f'(\alpha)$ il metodo è almeno del primo ordine (se $f(x) \in C^2[a, b]$, il metodo è almeno del secondo ordine).

L'interpretazione geometrica del metodo delle corde è data nella figura 3.9: x_1 è l'ascissa dell'intersezione con l'asse x della retta passante per $(x_0, f(x_0))$, con coefficiente angolare m. I punti successivi si ottengono in modo analogo e quindi le rette che individuano i punti x_i sono tutte parallele.

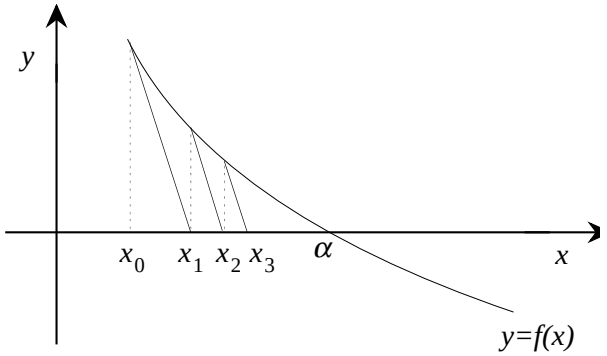


Fig. 3.9 - Interpretazione geometrica del metodo delle corde.

3.19 Esempio. Si applica il metodo delle corde alla risoluzione dell'equazione $f(x) = x^3 + 4x\cos x - 2 = 0$, già studiata negli esempi 3.1 e 3.5, la cui soluzione α appartiene all'intervallo [0,1]. In tale intervallo si ha

$$f'(x) = 3x^2 + 4\cos x - 4x\sin x > 0,$$

quindi occorre scegliere $m > 0$. Poiché

$$\max_{x\in[0,1]} |f'(x)| = f'(0) = 4,$$

per $m > 2$ il metodo è convergente. Il metodo è convergente anche per $m = 2$ perché $|f'(x)| < 4$ per $x \in (0, 1)$. Per $m = 2$ e $m = 3$ è $g'(x) < 0$ nell'intervallo $[0, 0.6]$ che contiene α e le successioni che si ottengono sono alternate quando i punti x_i appartengono a tale intervallo. Per $m \geq 4$ è $g'(x) > 0$ in $(0, 1]$ e le successioni che si ottengono sono monotone. Il numero delle iterazioni necessarie per ottenere 5 cifre decimali corrette varia con m ed è dato da

m	no. iter.
2	25
3	5
4	7
5	11
6	15

Il minor numero di iterazioni richieste per $m = 3$ è dovuto al fatto che $f'(\alpha) \approx 3.20$. Le successioni $\{x_i\}$ che si ottengono per $m = 3$ e per $m = 4$ a partire da $x_0 = 0$ sono

i	$m = 3$	$m = 4$
1	0.6666666	0.5000000
2	0.5360014	0.5299587
3	0.5368957	0.5354853
4	0.5368347	0.5365698
5	0.5368388	0.5367854
6		0.5368283
7		0.5368369

In ogni caso il metodo iterativo è del primo ordine, come sono del primo ordine i metodi usati nell'esempio 3.5. ∎

7. Metodo delle tangenti

Se $f(x)$ è una funzione derivabile, ponendo nella (7) $h(x) = f'(x)$, si ha

$$x = g(x) = x - \frac{f(x)}{f'(x)}\,.$$

Il corrispondente metodo è detto *metodo delle tangenti* o di *Newton*:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}\,, \quad f'(x_i) \neq 0,\ i = 0, 1, \ldots \tag{26}$$

La sua interpretazione geometrica è data nella figura 3.10: x_1 è l'ascissa dell'intersezione con l'asse x della tangente alla curva di equazione $y = f(x)$ in $(x_0, f(x_0))$. I punti successivi si ottengono in modo analogo, cioè le rette che individuano i punti x_i non sono fra di loro parallele, come nel metodo delle corde, ma sono tangenti alla curva.

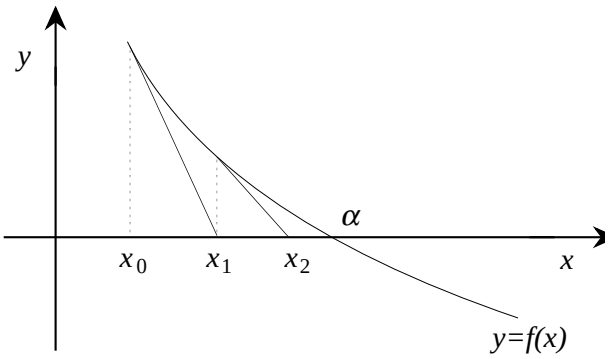


Fig. 3.10 - Interpretazione geometrica del metodo delle tangenti.

Per il metodo delle tangenti, se $f(x)$ è derivabile due volte e $f'(x) \neq 0$, risulta

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}\,. \tag{27}$$

Per esaminare l'ordine di convergenza del metodo delle tangenti, bisogna distinguere il caso in cui α è soluzione multipla dal caso in cui α è soluzione semplice. Se per un intero $r > 0$ è $f(x) \in C^r[a, b]$, una soluzione α si dice di *molteplicità* r se esiste finito e non nullo il

$$\lim_{x \to \alpha} \frac{f(x)}{(x - \alpha)^r}\,.$$

In tal caso risulta

$$f(\alpha) = f'(\alpha) = \ldots = f^{(r-1)}(\alpha) = 0, \quad f^{(r)}(\alpha) \neq 0.$$

Se $f(x) \in C^\infty[a,b]$ e $\lim_{x\to\alpha} \dfrac{f(x)}{(x-\alpha)^r} = 0$ per ogni r, allora la soluzione α si dice di *molteplicità infinita*. In tal caso risulta $f^{(r)}(\alpha) = 0$ per ogni r.

Per una definizione di molteplicità della soluzione nel caso in cui la funzione non sia derivabile un numero sufficiente di volte e per il conseguente comportamento del metodo delle tangenti si veda l'esercizio 3.30.

Si analizza ora l'ordine di convergenza del metodo delle tangenti.

3.20 Teorema. *Sia $\alpha \in [a,b]$ soluzione di $f(x) = 0$, e sia $f'(x) \neq 0$ per $x \in [a,b] - \{\alpha\}$.*

a) Se α ha molteplicità 1 e $f(x) \in C^2[a,b]$, allora il metodo delle tangenti è convergente con ordine almeno 2. In particolare l'ordine è 2 se $f''(\alpha) \neq 0$.

b) Se α ha molteplicità finita $r \geq 2$ e se $f(x) \in C^r[a,b]$, allora il metodo delle tangenti ha convergenza lineare.

Dim. Dalle ipotesi segue che $g'(x)$ è continua per $x \neq \alpha$. Nel caso a) è $f'(\alpha) \neq 0$ e

$$g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{[f'(\alpha)]^2} = 0,$$

e quindi esiste un intorno di α in cui $|g'(x)| < 1$, dove il metodo è convergente. Dalla (26) si ha

$$f'(x_i)(x_{i+1} - \alpha) = f'(x_i)(x_i - \alpha) - f(x_i),$$

e per la formula di Taylor è

$$f(\alpha) = f(x_i) + (\alpha - x_i)f'(x_i) + \frac{1}{2}(\alpha - x_i)^2 f''(\xi), \quad \text{con} \quad |\xi - x_i| < |\alpha - x_i|.$$

Poiché $f(\alpha) = 0$, risulta

$$f'(x_i)(x_{i+1} - \alpha) = \frac{1}{2}(\alpha - x_i)^2 f''(\xi),$$

da cui

$$\frac{x_{i+1} - \alpha}{(x_i - \alpha)^2} = \frac{f''(\xi)}{2f'(x_i)},$$

e passando al limite per $i \to \infty$ si ha

$$\lim_{i\to\infty} \frac{x_{i+1} - \alpha}{(x_i - \alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)},$$

per cui il metodo è almeno del secondo ordine: se $f''(\alpha) \neq 0$ il metodo è del secondo ordine, se $f''(\alpha) = 0$ e $f(x) \in C^3[a,b]$, il metodo è almeno del terzo ordine.

Nel caso b), α è anche soluzione di molteplicità $r-1$ per l'equazione $f'(x) = 0$ (si veda l'esercizio 3.27). Si consideri allora la funzione che individua il metodo delle tangenti in questo caso

$$g(x) = \begin{cases} x - \dfrac{f(x)}{f'(x)}, & \text{per } x \neq \alpha, \\ \alpha, & \text{per } x = \alpha. \end{cases}$$

La funzione $g(x)$ è continua. Infatti dalla formula di Taylor applicata alle funzioni $f(x)$ e $f'(x)$ si ha:

$$f(x) = \frac{f^{(r)}(\xi_1)}{r!}\,(x-\alpha)^r \quad \text{con} \quad |\xi_1 - \alpha| < |x - \alpha|,$$

$$f'(x) = \frac{f^{(r)}(\xi_2)}{(r-1)!}\,(x-\alpha)^{r-1} \quad \text{con} \quad |\xi_2 - \alpha| < |x - \alpha|,$$

dove

$$\lim_{x\to\alpha} f^{(r)}(\xi_1) = \lim_{x\to\alpha} f^{(r)}(\xi_2) = f^{(r)}(\alpha) \neq 0,$$

e quindi

$$\lim_{x\to\alpha} g(x) = \alpha - \lim_{x\to\alpha} \frac{f(x)}{f'(x)} = \alpha - \lim_{x\to\alpha} \frac{x-\alpha}{r}\,\frac{f^{(r)}(\xi_1)}{f^{(r)}(\xi_2)} = \alpha.$$

La funzione $g(x)$ è derivabile in α. Infatti

$$g'(\alpha) = \lim_{h\to 0} \frac{g(\alpha+h) - g(\alpha)}{h} = \lim_{h\to 0} \frac{1}{h}\left[\alpha + h - \frac{f(\alpha+h)}{f'(\alpha+h)} - \alpha\right]$$

$$= \lim_{h\to 0}\left[1 - \frac{1}{h}\,\frac{f^{(r)}(\xi_1)h^r}{r!}\,\frac{(r-1)!}{f^{(r)}(\xi_2)h^{r-1}}\right] = 1 - \frac{1}{r}\,.$$

Inoltre la $g'(x)$ è continua anche in α. Infatti per $r = 2$ dalla (27) è

$$\lim_{x\to\alpha} g'(x) = \lim_{x\to\alpha} \frac{f''(\xi_1)f''(x)}{2[f''(\xi_2)]^2} = \frac{1}{2}\,,$$

e per $r > 2$ dalla formula di Taylor applicata a $f''(x)$ si ha

$$f''(x) = \frac{f^{(r)}(\xi_3)}{(r-2)!}\,(x-\alpha)^{r-2} \quad \text{con} \quad \lim_{x\to\alpha} f^{(r)}(\xi_3) = f^{(r)}(\alpha) \neq 0,$$

e dalla (27) si ottiene

$$\lim_{x\to\alpha} g'(x) = \lim_{x\to\alpha} \frac{[(r-1)!]^2}{r!(r-2)!} \frac{f^{(r)}(\xi_1)f^{(r)}(\xi_3)}{[f^{(r)}(\xi_2)]^2} = 1 - \frac{1}{r}\,.$$

Quindi per $r \geq 2$ risulta che $0 < g'(\alpha) < 1$ e la convergenza lineare segue dal teorema 3.14. ∎

3.21 Esempio. Applicando il metodo delle tangenti alla risoluzione dell'equazione $f(x) = x^3 + 4x\cos x - 2 = 0$, la cui soluzione α appartiene all'intervallo $[0,1]$ (si veda l'esempio 3.19), si ha:

$$f'(x) = 3x^2 + 4\cos x - 4x\sin x \neq 0, \quad \text{per} \quad x \in [0,1],$$

per cui risulta

$$g(x) = x - \frac{x^3 + 4x\cos x - 2}{3x^2 + 4\cos x - 4x\sin x} = \frac{2x^3 - 4x^2\sin x + 2}{3x^2 + 4\cos x - 4x\sin x}\,.$$

Poiché α è soluzione di molteplicità 1 e $f''(\alpha) \neq 0$, il metodo è del secondo ordine. Inoltre per $x \leq \alpha$ è $0 \leq g'(x) < 1$, per cui assumendo $x_0 = 0$ si ottiene una successione monotona crescente. La successione effettivamente calcolata è la seguente:

i	x_i	i	x_i
1	0.5000000	3	0.5368383
2	0.5362971	4	0.5368383

Per mettere in evidenza la differenza di comportamento fra metodi del primo ordine e metodi del secondo, sono stati riportati nella figura 3.11, al variare dell'indice i di iterazione, gli errori relativi effettivamente generati dai vari metodi usati per approssimare la soluzione dell'equazione $x^3 + 4x\cos x - 2 = 0$: il metodo di bisezione utilizzato nell'esempio 3.1, il metodo di iterazione funzionale usato nell'esempio 3.5, il metodo delle corde, con $m = 3$, usato nell'esempio 3.19 e il metodo delle tangenti, usato in questo esempio. Si noti come gli errori relativi decrescano fino a quando essi non diventano dello stesso ordine della precisione di macchina con cui vengono effettuati i calcoli: a questo punto l'errore algoritmico diventa preponderante rispetto all'errore analitico

$$\epsilon_{an} = \frac{x_i - \alpha}{\alpha}\,,$$

e conviene arrestare l'iterazione.

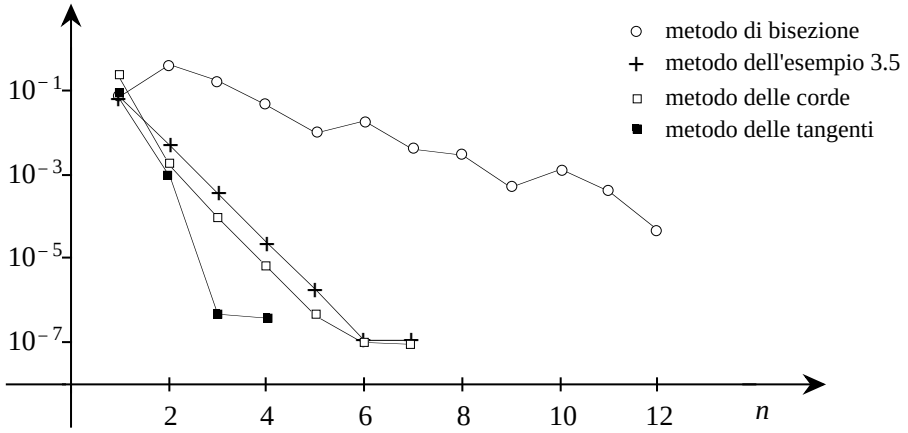


Fig. 3.11 - Errori relativi effettivamente generati nell'approssimazione della soluzione dell'equazione $x^3 + 4x\cos x - 2 = 0$.

3.22 Esempio. L'equazione

$$f(x) = x^n - k = 0, \quad k > 0, \tag{28}$$

ha per n intero e per $x > 0$ una sola soluzione reale $\alpha = \sqrt[n]{k}$, come risulta anche dal grafico della figura 3.12 (per n pari).

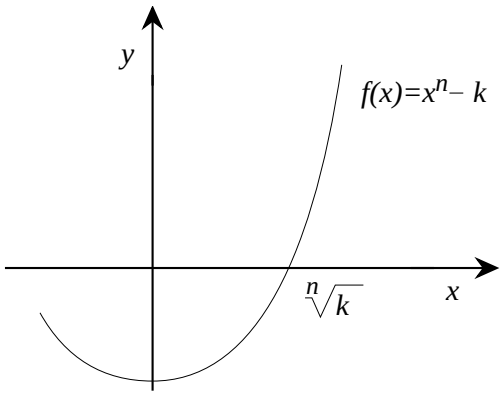


Fig. 3.12 - Grafico della funzione $f(x) = x^n - k = 0$, $k > 0$ (n pari).

Il metodo delle tangenti, applicato all'equazione (28), è:

$$x_{i+1} = \frac{1}{n}\left[(n-1)x_i + \frac{k}{x_i^{n-1}}\right], \quad i = 0, 1, \ldots \tag{29}$$

Poiché α è soluzione di molteplicità 1, e

$$f''(x) = n(n-1)x^{n-2} \neq 0, \quad \text{per ogni } x \neq 0,$$

il metodo delle tangenti è del secondo ordine. Si ha

$$g'(x) = \frac{(n-1)(x^n - k)}{nx^n},$$

ed è $0 \leq g'(x) < 1$ per $x \geq \alpha$, quindi per ogni $x_0 > \alpha$, la successione $\{x_i\}$ è monotona decrescente. Ovviamente il numero delle iterazioni necessarie per approssimare α con una prefissata precisione dipende anche dalla scelta di x_0.

Se n è dispari, $n = 2m + 1$, m intero, l'equazione $f(x) = 0$ può essere trasformata nell'altra equivalente

$$f_1(x) = x^{m+1} - kx^{-m} = 0. \tag{30}$$

Si ha

$$f_1'(x) = (m+1)x^m + kmx^{-m-1}$$
$$f_1''(x) = m(m+1)(x^{m-1} - kx^{-m-2}) = m(m+1)x^{-2}f_1(x).$$

Poiché $f_1'(\alpha) \neq 0$ e $f_1''(\alpha) = 0$, il metodo delle tangenti, applicato all'equazione (30), è almeno del terzo ordine, ed è dato da:

$$x_{i+1} = \frac{mx_i^{n+1} + k(m+1)x_i}{(m+1)x_i^n + km}, \quad i = 0, 1, \ldots \tag{31}$$

In particolare, per calcolare $\sqrt[3]{2}$, soluzione dell'equazione $x^3 - 2 = 0$, dalla (29) si ottiene

$$x_{i+1} = \frac{2}{3}\left(x_i + \frac{1}{x_i^2}\right), \tag{32}$$

mentre dalla (31) si ottiene:

$$x_{i+1} = \frac{x_i^4 + 4x_i}{2(x_i^3 + 1)}. \tag{33}$$

Le successioni effettivamente calcolate a partire dal punto $x_0 = 2$ sono

i	successione (32)	successione (33)
1	1.499999	1.333333
2	1.296295	1.260073
3	1.260931	1.259921
4	1.259921	

Si tenga però presente che 3 iterazioni con la (33) hanno richiesto 21 operazioni, mentre 4 iterazioni con la (32) hanno richiesto 16 operazioni. ∎

Nel caso di soluzioni di molteplicità infinita, il teorema 3.20 non è applicabile. Se però $g'(x)$ è continua in α, allora il metodo delle tangenti può essere convergente, come risulta dall'esempio seguente.

3.23 Esempio. La funzione (di *Brent*)

$$f(x) = \begin{cases} 0 & \text{per } x = 0, \\ x\mathrm{e}^{-x^{-2}} & \text{per } x \neq 0, \end{cases}$$

ha il solo zero $\alpha = 0$, come risulta anche dal grafico di $f(x)$ riportato in figura 3.13.

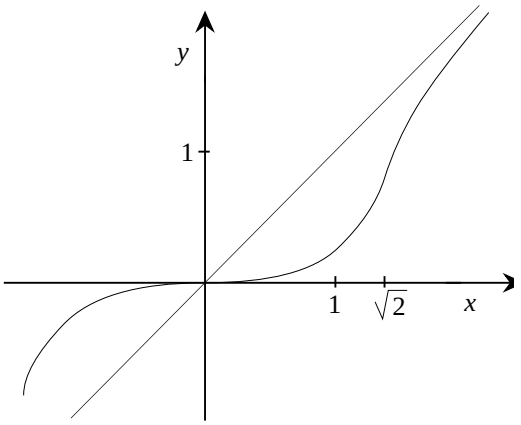


Fig. 3.13 - Grafico della funzione di Brent.

In α la $f(x)$ e tutte le sue derivate risultano continue; inoltre

$$f^{(r)}(\alpha) = 0, \quad \text{per ogni } r,$$

e quindi $\alpha = 0$ è soluzione di molteplicità infinita. Nel caso del metodo delle tangenti si ha

$$g(x) = x - \frac{f(x)}{f'(x)} = \frac{2x}{x^2+2}\,,$$
$$g'(x) = \frac{4-2x^2}{(x^2+2)^2}\,,$$

e quindi

$$g'(0) = 1,$$
$$0 < g'(x) < 1 \quad \text{per} \quad x^2 < 2, \quad x \neq 0.$$

Allora per qualunque x_0 tale che $x_0^2 < 2$, la successione $x_{i+1} = g(x_i)$ è convergente e monotona e per il teorema 3.14 la convergenza è sublineare. Assumendo $x_0 = 1$ si ha infatti:

i	x_i	i	x_i
1	0.6666667	.	...
2	0.5454546	97	0.1003889
3	0.4748201	98	0.09988564
4	0.4267176	99	0.09938985
.	...	.	...

■

Se r è la molteplicità della soluzione α, $r > 1$ intero, e $f(x) \in C^{r+1}[a, b]$, il seguente metodo, ricavato dal metodo delle tangenti,

$$x_{i+1} = x_i - r\,\frac{f(x_i)}{f'(x_i)}\,, \quad f'(x_i) \neq 0,\ i = 0, 1, \ldots. \tag{34}$$

è almeno del secondo ordine. Infatti si ha:

$$g(x) = x - r\,\frac{f(x)}{f'(x)}\,,$$
$$g'(x) = 1 - r + r\,\frac{f(x)f''(x)}{[f'(x)]^2}\,.$$

Procedendo come nella dimostrazione del teorema 3.20, si ha

$$\lim_{x\to\alpha} g'(x) = 1 - r + r\big(1 - \frac{1}{r}\,\big) = 0,$$

e quindi, ponendo $g'(\alpha) = 0$, la $g'(x)$ può essere estesa per continuità in α. Procedendo in modo analogo a quello seguito per determinare l'ordine del metodo delle tangenti nel caso di soluzioni di molteplicità 1, si può verificare che se $f^{(r+1)}(\alpha) \neq 0$, il metodo iterativo (34) è del secondo ordine.

3.24 Esempio. Si applica il metodo delle tangenti alle due equazioni dell'esempio 3.7. L'equazione

$$f(x) = x - \cos x = 0,$$

ha soluzione $\alpha \in [0, 1]$ con molteplicità $r = 1$ (si veda la figura 3.7) e, poiché $f''(\alpha) \neq 0$, il metodo risulta del secondo ordine. Si ha:

$$g(x) = \frac{x\sin x + \cos x}{1 + \sin x}\,,$$
$$g'(x) = \frac{\cos x(x - \cos x)}{(1 + \sin x)^2}\,.$$

Poiché per $x \geq \alpha$ è $0 \leq g'(x) < 1$, se $x_0 > \alpha$ la successione che si ottiene

$$x_{i+1} = \frac{x_i \sin x_i + \cos x_i}{1 + \sin x_i}, \quad i = 0, 1, \ldots$$

è monotona decrescente. Se $x_0 = 1$ risulta

i	x_i	i	x_i
1	0.7503639	3	0.7390853
2	0.7391127	4	0.7390850

L'equazione

$$f(x) = x - \sin x = 0,$$

ha soluzione $\alpha = 0$ con molteplicità $r = 3$ (si veda la figura 3.6) e il metodo delle tangenti risulta del primo ordine. Se $x_0 = 1$ si ottiene la successione:

i	x_i	i	x_i
1	0.6551450	.	...
2	0.4335902	15	$0.1802885\ 10^{-2}$
3	0.2881477	16	$0.1012731\ 10^{-2}$
4	0.1918310	17	$0.4882813\ 10^{-3}$
.	...	18	0

Utilizzando invece il metodo (34):

$$x_{i+1} = x_i - 3\,\frac{x_i - \sin x_i}{1 - \cos x_i}, \quad i = 0, 1, \ldots$$

se $x_0 = 1$ si ottiene la successione:

i	x_i
1	$-0.3456645\ 10^{-1}$
2	0

■

Non è sempre agevole verificare la convergenza del metodo delle tangenti utilizzando la condizione sufficiente $|g'(x)| < 1$, che comunque assicura solo una convergenza locale del metodo. Si possono però individuare condizioni sufficienti di convergenza su intervalli, di facile verifica.

3.25 Teorema. *Sia $f \in C^2(S)$, $S = [\alpha, \alpha + \rho]$, $\rho > 0$, tale che*

$$f(x)f''(x) > 0, \quad f'(x) \neq 0, \quad \textit{per} \quad x \in S - \{\alpha\}.$$

Se $x_0 \in S - \{\alpha\}$, la successione ottenuta con il metodo delle tangenti è decrescente e convergente ad α.

Dim. Per le ipotesi fatte, se $f'(x) > 0$ risulta $f(x) \geq 0$, e se $f'(x) < 0$ risulta $f(x) \leq 0$ e quindi $f(x)/f'(x) > 0$ per $x \in S - \{\alpha\}$. Ne segue che se $x_i \in S$ allora

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} < x_i, \quad \text{per} \quad i = 0, 1, \ldots$$

Per dimostrare che $x_i - \alpha > 0$, si procede per induzione su i. Per $i = 0$ è $x_0 - \alpha > 0$ perché $x_0 \in S - \{\alpha\}$; per $i > 0$ si ha

$$x_{i+1} - \alpha = g'(\xi)(x_i - \alpha), \quad \alpha < \xi < x_i, \tag{35}$$

dove, da (27), è

$$g'(\xi) = \frac{f(\xi)f''(\xi)}{[f'(\xi)]^2}.$$

Poiché per l'ipotesi induttiva risulta $\xi \in S - \{\alpha\}$, è $f(\xi)f''(\xi) > 0$ e $g'(\xi) > 0$, e poiché $x_i - \alpha > 0$, dalla (35) segue che $x_{i+1} - \alpha > 0$. Quindi

$$\alpha < x_{i+1} < x_i, \quad \text{per} \quad i = 0, 1, \ldots$$

Ne segue che la successione è decrescente e inferiormente limitata, quindi convergente ad α, che è l'unica soluzione in S. ∎

Un risultato analogo vale se è $S = [\alpha - \rho, \alpha]$, cioè vale il seguente

3.26 Teorema. *Sia $f \in C^2(S)$, $S = [\alpha - \rho, \alpha]$, $\rho > 0$, tale che*

$$f(x)f''(x) > 0, \quad f'(x) \neq 0, \quad \textit{per} \quad x \in S - \{\alpha\}.$$

Se $x_0 \in S - \{\alpha\}$, la successione ottenuta con il metodo delle tangenti è crescente e convergente ad α. ∎

Condizioni sufficienti per la convergenza del metodo delle tangenti possono essere date anche se la funzione $f(x)$ non ha derivata seconda: in tal caso basta che la funzione sia convessa. Si veda per questo il teorema 3.47.

8. Metodo delle secanti

Il metodo delle tangenti richiede ad ogni iterazione il calcolo di due funzioni, cioè di $f(x_i)$ e di $f'(x_i)$, calcolo che può essere di elevato costo computazionale, in particolare quando $f'(x)$ richiede una maggiore mole di calcolo rispetto a $f(x)$. Può essere perciò conveniente utilizzare metodi, come il metodo delle corde (24), che richiedano ad ogni iterazione solo il calcolo di valori della funzione $f(x)$. Un altro metodo di questo tipo è il *metodo delle secanti*

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - c)}{f(x_i) - f(c)}\,, \quad \text{dove } c \in [a, b], \tag{36}$$

che corrisponde a porre nella (7)

$$h(x) = \frac{f(x) - f(c)}{x - c}\,.$$

La sua interpretazione geometrica è illustrata nella figura 3.14: x_1 è l'ascissa dell'intersezione con l'asse x del segmento che unisce $(c, f(c))$ e $(x_0, f(x_0))$. I punti successivi si ottengono in modo analogo.

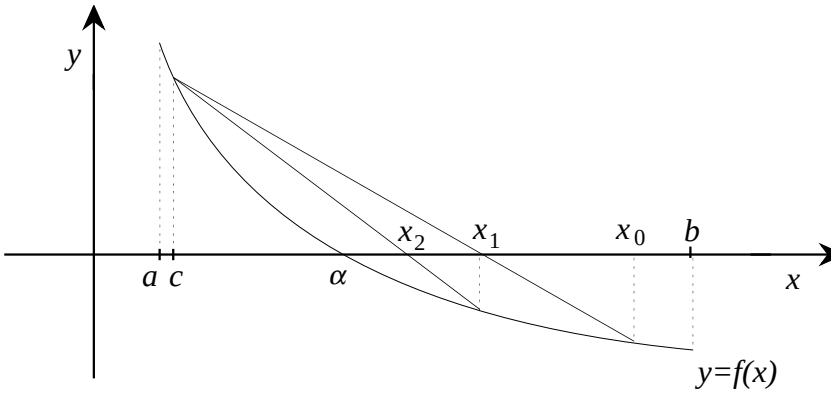


Fig. 3.14 - Interpretazione geometrica del metodo delle secanti.

Poiché

$$g(x) = x - \frac{f(x)(x - c)}{f(x) - f(c)}\,,$$

è

$$g'(x) = f(c)\,\frac{f'(x)(x - c) - f(x) + f(c)}{[f(x) - f(c)]^2}\,, \tag{37}$$

e

$$g'(\alpha) = 1 + \frac{f'(\alpha)}{f(c)}\,(\alpha - c).$$

Se $g'(\alpha) \neq 0$, il metodo delle secanti è del primo ordine. Se c è scelto in modo che il rapporto $f(c)/(c-\alpha)$ abbia lo stesso segno di $f'(\alpha)$ ed inoltre

$$\left|\frac{f(c)}{c-\alpha}\right| > \frac{1}{2}\,|f'(\alpha)|,$$

allora $|g'(\alpha)| < 1$ e quindi si ha la convergenza locale del metodo. Il seguente teorema dà delle condizioni sufficienti di convergenza.

3.27 Teorema. *Sia $\alpha \in [a,b]$ e $f \in C^2[a,b]$ e siano $f'(x) \neq 0$ e $f''(x) \neq 0$ per $x \in (a,b)$. Scelti c tale che $f(c)f''(c) \geq 0$ e x_0 tale che $f(x_0)f''(x_0) \leq 0$, la successione ottenuta con il metodo delle secanti è monotona convergente.*

Dim. Si consideri il caso in cui $f(a)f(c) > 0$, come è nella figura 3.14. Per dimostrare che $\alpha < x_{i+1} < x_i$ si procede per induzione su i. Per $i = 0$ la tesi è ovvia. Per $i > 0$, poiché per ipotesi $f(x_i)$ e $f(c)$ hanno segno opposto e $x_i - c > 0$, ne segue che

$$\frac{f(x_i)(x_i-c)}{f(x_i)-f(c)} > 0,$$

e quindi

$$x_{i+1} = x_i - \frac{f(x_i)(x_i-c)}{f(x_i)-f(c)} < x_i, \quad i = 0,1,\ldots$$

Poiché

$$x_{i+1} - \alpha = g'(\xi)(x_i - \alpha), \quad \text{con} \quad \alpha < \xi < x_i, \tag{38}$$

e per la formula di Taylor risulta

$$f(c) = f(\xi) + f'(\xi)(c-\xi) + \frac{f''(\eta)}{2}\,(c-\xi)^2, \quad \text{con} \quad c < \eta < \xi,$$

sostituendo nella (37), si ha

$$g'(\xi) = f(c)\,\frac{f''(\eta)(c-\xi)^2}{2[f(\xi)-f(c)]^2}\,.$$

Poiché $f''(x)$ non cambia segno in $[a,b]$ ne segue che $f(c)f''(\eta) > 0$, e quindi $g'(\xi) > 0$, e dalla (38) si ha $\alpha < x_{i+1} < x_i$. Ne segue che la successione è monotona decrescente, inferiormente limitata, e quindi convergente ad α che è l'unica soluzione in $[a,b]$. In modo analogo si può condurre la dimostrazione nel caso che sia $f(b)f(c) > 0$. ∎

3.28 Esempio. Si applica il metodo delle secanti alla risoluzione dell'equazione $f(x) = x^3 + 4x\cos x - 2 = 0$, già risolta nell'esempio 3.19 con il

metodo delle corde e nell'esempio 3.21 con il metodo delle tangenti. Poiché la soluzione α appartiene all'intervallo (0,1) in cui

$$f''(x) = 6x - 8\sin x - 4x\cos x < 0,$$

si sceglie $c = 0$ e $x_0 = 1$. La successione effettivamente calcolata è data da

i	x_i	i	x_i
1	0.6326693	5	0.5368798
2	0.5515600	6	0.5368442
3	0.5389310	7	0.5368391
4	0.5371324	8	0.5368387

∎

Se nell'intervallo $[a, b]$ non sono soddisfatte le ipotesi del teorema 3.27, è possibile che x_{i+1} non sia compreso fra c e x_i, e quindi la successione ottenuta potrebbe non convergere. È però possibile ottenere una successione, che è convergente sotto la sola ipotesi di continuità della $f(x)$, introducendo la seguente modifica:

x_0 e c_0 sono scelti in modo tale che $f(x_0)f(c_0) < 0$;

per $i = 0, 1, \ldots$

$$\text{si calcola } x_{i+1} = x_i - \frac{f(x_i)(x_i - c_i)}{f(x_i) - f(c_i)}\,, \tag{39}$$

se $f(x_{i+1})f(c_i) > 0$ si pone $c_{i+1} = x_i$, altrimenti $c_{i+1} = c_i$.

Il metodo così modificato prende il nome di metodo di *falsa posizione (regula falsi)*. L'interpretazione geometrica di questo metodo è riportata nella figura 3.15.

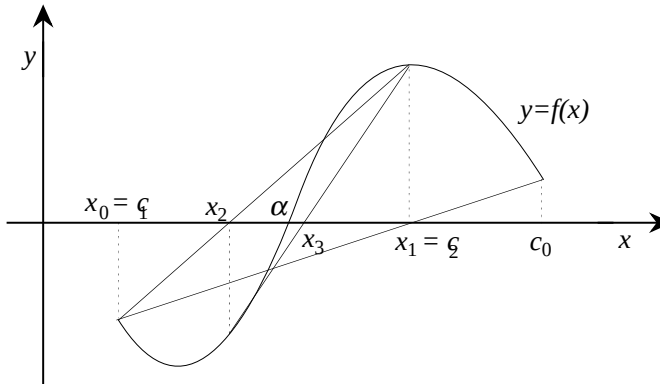


Fig. 3.15 - Interpretazione geometrica del metodo di falsa posizione.

3.29 Teorema. *Sia $f \in C[a,b]$, con $f(a)f(b) < 0$, e sia α l'unica soluzione di (1) in $[a,b]$. Allora il metodo di falsa posizione è convergente.*

Dim. Si supponga per semplicità $x_0 < c_0$. Dalla (39) si ottiene

$$x_{i+1} - c_i = x_i - c_i - \frac{f(x_i)(x_i - c_i)}{f(x_i) - f(c_i)} = \frac{f(c_i)(x_i - c_i)}{f(c_i) - f(x_i)},$$

e passando ai moduli

$$|x_{i+1} - c_i| = \Big|\frac{f(c_i)}{f(c_i) - f(x_i)}\Big| \, |x_i - c_i|. \tag{40}$$

Il metodo individua successivi intervalli di estremi c_i e x_i, contenenti α, ognuno dei quali è incluso nel precedente. Perciò la successione formata dagli estremi sinistri di tali intervalli è non decrescente e la successione formata dagli estremi destri è non crescente. Quindi le due successioni, essendo monotone e limitate, sono convergenti, rispettivamente ai limiti β e γ, con $\gamma \geq \beta$ e $f(\beta)f(\gamma) \leq 0$ e risulta

$$\lim_{i\to\infty} |x_i - c_i| = \gamma - \beta.$$

Escludendo il caso, possibile, che esista un indice j, tale che $x_i = \alpha$ per $i \geq j$, si possono verificare due casi:

a) esiste un indice j tale che $c_i = c_j$ per ogni $i \geq j$ (cioè il metodo coincide con il metodo delle secanti), ed allora

$$\lim_{i\to\infty} |x_{i+1} - c_i| = \gamma - \beta,$$

e dalla (40) segue che

$$\lim_{i\to\infty} \Big|\frac{f(c_i)}{f(c_i) - f(x_i)}\Big| = 1,$$

da cui, poiché $f(x_i)f(c_i) < 0$, si ha

$$\lim_{i\to\infty} f(x_i) = 0.$$

Per la continuità di $f(x)$ nell'intervallo $[a,b]$, poiché α è l'unica soluzione in $[a,b]$, ne segue che

$$\lim_{i\to\infty} x_i = \alpha.$$

b) La successione $\{c_i\}$ non è definitivamente costante e quindi esiste una sottosuccessione della successione $\{x_{i+1} - c_i\}$ che tende a 0, ottenuta

considerando solo elementi della successione degli estremi sinistri. Dalla (40) per la continuità della $f(x)$ deve essere $\beta = \gamma = \alpha$, oppure

$$\lim_{i\to\infty} f(c_i) = 0,$$

in cui i c_i appartengono alla sottosuccessione considerata degli estremi sinistri. Ne segue che $\beta = \alpha$. Poiché il ragionamento può essere ripetuto per gli estremi destri, ne segue che anche $\gamma = \alpha$, e quindi comunque

$$\lim_{i\to\infty} x_i = \alpha. \qquad \blacksquare$$

3.30 Esempio. Per l'equazione $f(x) = x^5 - 5x^3 + 10x + 1 = 0$, che ha una sola soluzione reale α appartenente all'intervallo $(-2, 2)$, la convergenza del metodo delle secanti non è assicurata, in quanto non sono verificate le ipotesi del teorema 3.27 (infatti le derivate prime e seconde di $f(x)$ si annullano in punti interni all'intervallo). Invece la convergenza del metodo di falsa posizione è assicurata a priori, perché sono verificate le ipotesi del teorema 3.29. Applicando i due metodi e arrestando l'iterazione quando $|x_{i+1} - x_i| < 10^{-5}$, con il metodo delle secanti occorrono 20 iterazioni se $x_0 = -2$, $c = 2$ e 29 iterazioni se $x_0 = 2$, $c = -2$; con il metodo di falsa posizione occorrono 7 iterazioni. In tutti e tre i casi si ottiene per α il valore -0.1005067. $\blacksquare$

Un'altra variante del metodo delle secanti consiste nel considerare all'i-esima iterazione la retta che passa per i punti $(x_i, f(x_i))$ e $(x_{i-1}, f(x_{i-1}))$, e il metodo che ne risulta è il seguente:

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}\,. \tag{41}$$

L'interpretazione geometrica di questo metodo è riportata nella figura 3.16.

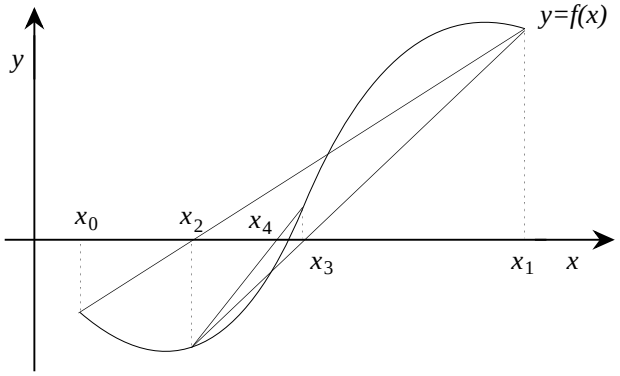


Fig. 3.16 - Interpretazione geometrica del metodo (41).

Poiché questo metodo non appartiene alla classe dei metodi individuati dalla (5), il teorema 3.3 non può essere applicato. È possibile comunque dimostrare che il metodo è localmente convergente. Vale infatti il seguente teorema.

3.31 Teorema. *Sia $f(x) \in C^2[\alpha-\rho, \alpha+\rho]$, $\rho > 0$, e sia $f'(x) \neq 0$ per $|x-\alpha| < \rho$. Allora esiste un intorno di α, in cui il metodo (41) è convergente, con convergenza superlineare.*

Dim. Dalla (41) si ottiene

$$f(x_i) + \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} (x_{i+1} - x_i) = 0. \tag{42}$$

Il polinomio di interpolazione per i punti $(x_i, f(x_i))$ e $(x_{i-1}, f(x_{i-1}))$ è dato da (si veda l'esempio 5.17)

$$p(x) = f(x_i) + (x - x_i) \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}},$$

e per il teorema 5.5 è

$$f(x) = p(x) + (x - x_i)(x - x_{i-1}) \frac{f''(\xi)}{2},$$

dove ξ appartiene all'intervallo aperto di estremi x_i e x_{i-1}. Nel punto $x = \alpha$ è $f(\alpha) = 0$, e quindi

$$-f(x_i) - (\alpha - x_i) \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = (\alpha - x_i)(\alpha - x_{i-1}) \frac{f''(\xi)}{2}.$$

Sommando questa relazione membro a membro con la (42) si ottiene

$$(x_{i+1} - \alpha) \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = (\alpha - x_i)(\alpha - x_{i-1}) \frac{f''(\xi)}{2},$$

e poiché

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = f'(\eta),$$

con η appartenente all'intervallo aperto di estremi x_i e x_{i-1}, si ottiene

$$x_{i+1} - \alpha = (x_i - \alpha)(x_{i-1} - \alpha) \frac{f''(\xi)}{2f'(\eta)}. \tag{43}$$

Ponendo $e_i = x_i - \alpha$, $i = 0, 1, \ldots$ e

$$\mu = \max \left\{ \frac{1}{\rho}, \ \frac{|f''(x)|}{2|f'(y)|}, \ x, y \ \in [\alpha - \rho, \alpha + \rho] \right\},$$

si ha dalla (43)

$$|e_{i+1}| \leq \mu |e_i||e_{i-1}|, \quad i = 1, 2, \ldots \tag{44}$$

Sia U l'intorno di α

$$U = \{ \ x \ : \ |x - \alpha| < \ \frac{1}{\mu} \ \},$$

e siano x_0 e $x_1 \in U$. Si pone

$$\delta = \max \{\mu|e_0|, \mu|e_1|\} < 1$$

e si dimostra per induzione su i la seguente disuguaglianza:

$$\mu|e_i| \leq \delta^i. \tag{45}$$

Per $i = 0$ e per $i = 1$, la (45) è conseguenza della definizione di δ; per $i = 2$, dalla (44) si ha

$$\mu|e_2| \leq \mu|e_1|\mu|e_0| \leq \delta^2;$$

per $i > 2$, per l'ipotesi induttiva risulta $\mu|e_{i-1}| \leq \delta^{i-1} < 1$ e $\mu|e_{i-2}| \leq \delta^{i-2} < 1$ e quindi x_{i-1} e $x_{i-2} \in U$ e dalla (44) si ha

$$\mu|e_i| \leq \mu|e_{i-1}|\mu|e_{i-2}| \leq \delta^{i-1}\delta^{i-2} = \delta^i\delta^{i-3} \leq \delta^i.$$

Dalla (45) segue la convergenza di $\{x_i\}$. Dalla (44), se $x_i \neq \alpha$ per ogni i, si ha:

$$\frac{|e_i|}{|e_{i-1}|} \leq \mu|e_{i-2}|,$$

e, passando al limite

$$\lim_{i \to \infty} \frac{|e_i|}{|e_{i-1}|} = 0. \qquad \blacksquare$$

Più precisamente l'ordine di convergenza della variante (41) del metodo delle secanti è (si veda l'esercizio 4.65)

$$p = \frac{1 + \sqrt{5}}{2} = 1.618034 \quad \text{(sezione aurea)}.$$

3.32 Esempio. Applicando il metodo delle secanti nella variante (41) all'equazione dell'esempio 3.30, e arrestando il procedimento quando $|x_{i+1} - x_i| < 10^{-5}$, le successioni calcolate sia con $x_0 = -2$ e $x_1 = 2$, che con $x_0 = 2$ e $x_1 = -2$, forniscono il valore -0.1005067 in 5 iterazioni. ∎

Confrontando le diverse varianti del metodo delle secanti si nota che:

a) il metodo (36) ha convergenza lineare, e le condizioni sufficienti per la convergenza sono spesso di agevole verifica;

b) il metodo (39) richiede ad ogni iterazione, oltre al calcolo di $f(x)$, anche un confronto, ma la successione $\{x_i\}$ è convergente nella sola ipotesi di continuità della $f(x)$;

c) il metodo (41) ha convergenza superlineare, ma le condizioni sufficienti per la convergenza valgono soltanto localmente.

9. Efficienza di un metodo iterativo

Il concetto di *efficienza* di un metodo iterativo può essere studiato da diversi punti di vista.

Nella pratica si usa legare l'efficienza di un metodo al numero di iterazioni che occorrono per ridurre l'errore iniziale di una quantità prefissata. Il numero di iterazioni dipende dall'ordine di convergenza del metodo. Dalla (20), posto $e_i = x_i - \alpha$, si ha

$$|e_i| \leq \beta |e_{i-1}|^p, \quad \text{con} \quad \beta < 1 \text{ se } p = 1,$$

da cui

$$|e_i| \leq \beta |e_{i-1}|^p \leq \beta\beta^p |e_{i-2}|^{p^2} \leq \ldots \leq \beta\beta^p\beta^{p^2} \ldots \beta^{p^{i-1}} \, |e_0|^{p^i};$$

se $p \neq 1$ risulta

$$|e_i| \leq \frac{1}{\sqrt[p-1]{\beta}} \left(\sqrt[p-1]{\beta} \, |e_0| \right)^{p^i},$$

e se $p = 1$ risulta

$$|e_i| \leq \beta^i \, |e_0|.$$

Quindi, fissato un $\epsilon > 0$, per un metodo del primo ordine risulta

$$\frac{|e_i|}{|e_0|} \leq \epsilon \quad \text{se} \quad i \geq k_1 = \frac{\log \epsilon}{\log \beta},$$

cioè per ridurre l'errore iniziale di ϵ sono sufficienti $\lceil k_1 \rceil$ iterazioni, mentre per un metodo di ordine $p > 1$, se

$$\sqrt[p-1]{\beta} \, |e_0| < 1,$$

sono sufficienti $\lceil k_p \rceil$ iterazioni, dove

$$p^{k_p} = \frac{\log\left(\sqrt[p-1]{\beta}\,|e_0|\epsilon\right)}{\log\left(\sqrt[p-1]{\beta}|e_0|\right)} = 1 + \frac{\log\epsilon}{\log\left(\sqrt[p-1]{\beta}\,|e_0|\right)}\,.$$

Ad esempio, se $p = 2$, $\beta|e_0| = \dfrac{1}{2}$ e $\epsilon = 2^{-31}$, risulta $k_p = 5$, cioè 5 iterazioni sono sufficienti per ridurre l'errore iniziale di $2^{-31} \approx 4.66\ 10^{-10}$. Invece per un metodo del primo ordine per il quale sia $\beta = \dfrac{1}{2}$, il numero delle iterazioni richieste è di 31.

Asintoticamente, cioè per $\epsilon \to 0$, il numero di iterazioni di un metodo del primo ordine cresce come

$$\alpha_1 \log\epsilon^{-1}, \quad \text{dove} \quad \alpha_1 > 0,$$

mentre il numero di iterazioni di un metodo di ordine p cresce come

$$\frac{1}{\log p}\ \log\Big(1 + \frac{\log\epsilon^{-1}}{\alpha_p}\Big), \quad \text{dove} \quad \alpha_p > 0.$$

Perciò, per ottenere la stessa riduzione ϵ di errore, asintoticamente, il numero di iterazioni richieste da un metodo di ordine p è dell'ordine del logaritmo del numero di iterazioni richieste da un metodo del primo ordine, mentre il rapporto fra il numero di iterazioni richieste da un metodo di ordine p e da un metodo di ordine q, con $p, q \neq 1$, è $\log q/\log p$, in quanto

$$\lim_{\epsilon\to 0} \frac{\log\Big(1 + \dfrac{\log\epsilon^{-1}}{\alpha_p}\Big)}{\log\Big(1 + \dfrac{\log\epsilon^{-1}}{\alpha_q}\Big)} = 1.$$

Perciò il metodo di ordine p è preferibile a quello di ordine q se il rapporto fra i numeri di operazioni necessarie ad effettuare un'iterazione con il metodo di ordine p e con il metodo di ordine q è inferiore a $\log p/\log q$.

È quindi opportuno esaminare l'efficienza di un metodo valutandone anche il costo, mediante la relazione fra ordine e numero di valutazioni di funzioni richieste in ogni iterazione.

I metodi della forma (5) possono richiedere ad ogni iterazione il calcolo di valori della funzione $f(x)$, ed eventualmente delle sue derivate, nel solo punto x_i, oppure in punti anche diversi da x_i. Metodi del primo tipo, come i metodi delle corde e delle tangenti, sono detti *metodi ad un punto*. I metodi che ad ogni iterazione richiedono valori della funzione $f(x)$ e delle sue

derivate anche in punti diversi da x_i, valori che non sono già stati calcolati nei passi precedenti, sono detti *metodi a più punti*. Ne sono un esempio il metodo di *Steffensen*

$$x_{i+1} = x_i - \frac{[f(x_i)]^2}{f[x_i + f(x_i)] - f(x_i)},$$

o il metodo delle corde-tangenti

$$x_{i+1} = y_i - \frac{f(y_i)}{f'(x_i)}, \quad y_i = x_i - \frac{f(x_i)}{f'(x_i)}$$

(si vedano gli esercizi 3.35 e 3.36).

Oltre a questi si possono considerare metodi che utilizzano ad ogni iterazione anche valori della funzione e delle derivate già utilizzati nei passi precedenti. Questi metodi vengono detti *metodi con memoria*. Ne è un esempio il metodo iterativo (41), in cui oltre al valore $f(x_i)$ viene utilizzato anche il precedente valore $f(x_{i-1})$.

Per valutare l'efficienza di un metodo si deve innanzi tutto considerare il numero k di nuove informazioni richieste ad ogni iterazione, cioè il numero di nuove valutazioni di funzioni. Ad esempio per il metodo delle corde e delle secanti ad ogni iterazione è richiesta una sola informazione, il valore $f(x_i)$; per il metodo delle tangenti ad ogni iterazione sono richieste due informazioni, i valori di $f(x_i)$ e di $f'(x_i)$; per il metodo di Steffensen ad ogni iterazione sono richieste due informazioni, i valori di $f(x_i)$ e di $f[x_i + f(x_i)]$; per il metodo delle corde-tangenti ad ogni iterazione sono richieste tre informazioni, i valori di $f(x_i)$, di $f'(x_i)$ e di $f(y_i)$. Nel caso del metodo (41) ad ogni iterazione è richiesta una sola informazione, il valore di $f(x_i)$, in quanto il valore $f(x_{i-1})$ è già stato calcolato all'iterazione precedente.

Si definisce allora *efficienza informativa* E di un metodo il rapporto fra l'ordine p e il numero k di nuove informazioni richieste ad ogni iterazione. Quindi, limitandosi al caso di soluzioni di molteplicità 1, per il metodo delle corde, delle tangenti, delle secanti, di Steffensen e delle corde-tangenti risulta $E = 1$. Si può dimostrare [27] che per i metodi iterativi ad un punto è $E \leq 1$ e che se un metodo ad un punto ha ordine p ed efficienza informativa $E = 1$, allora esso dipende esplicitamente da $f(x_i)$ e da $f^{(j)}(x_i)$, per $j = 1, \ldots, p-1$.

Una limitazione così vincolante non vale per i metodi a più punti: sono stati infatti individuati metodi a più punti per cui $E > 1$. Alcuni di questi metodi sono ottenuti approssimando le derivate di $f(x)$ con combinazioni di valori della $f(x)$ in punti opportuni, altri metodi sono ottenuti combinando metodi di ordine più basso.

Ad esempio il metodo di Ostrowski

$$x_{i+1} = y_i - \frac{f(y_i)(y_i - x_i)}{2f(y_i) - f(x_i)}, \quad y_i = x_i - \frac{f(x_i)}{f'(x_i)},$$

ottenuto dal metodo delle tangenti e delle secanti, è del quarto ordine e richiede solo tre valutazioni di funzioni, $f(x_i)$, $f'(x_i)$ e $f(y_i)$. Quindi la sua efficienza è $E = \frac{4}{3}$.

Anche i metodi con memoria possono avere una efficienza informativa E maggiore di 1. I metodi con memoria hanno sempre ordine non intero [27] e, nel caso dei metodi a un punto, sono tali che $1 < E < 2$. Ad esempio il metodo (41) ha ordine

$$\frac{1+\sqrt{5}}{2} \approx 1.62$$

(si veda l'esercizio 4.65) e, poiché viene utilizzata una sola nuova informazione, è $E \approx 1.62$. Una efficienza informativa ancora maggiore, $E \approx 1.84$, ha il metodo

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} + \frac{f(x_i)f(x_{i-1})}{f(x_i) - f(x_{i-2})} \left[\frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} - \frac{x_{i-1} - x_{i-2}}{f(x_{i-1}) - f(x_{i-2})}\right],$$

che si ottiene con una interpolazione inversa della funzione $f(x)$ nei punti x_i, x_{i-1} e x_{i-2} (si veda il paragrafo 8 del capitolo 5). Esistono metodi a più punti con memoria che hanno una più elevata efficienza informativa.

In questa analisi non si è ovviamente tenuto conto dell'effettivo costo computazionale della funzione $f(x)$ e delle sue derivate. La definizione di efficienza informativa si basa sull'ipotesi di un costo computazionale confrontabile per la funzione $f(x)$ e per le sue derivate. Se però il calcolo di $f'(x)$ ha un costo superiore a quello di $f(x)$, conviene utilizzare metodi che non utilizzano derivate, come il metodo (41), il metodo di Steffensen o quello ottenuto con l'interpolazione inversa.

Non è però possibile dare un criterio utilizzabile in ogni caso: esistono infatti casi in cui metodi di elevato ordine danno risultati peggiori di metodi di ordine più basso e addirittura peggiori del metodo di bisezione. Inoltre quando si usa un metodo di ordine più elevato può essere difficile individuare un conveniente punto iniziale. Per questo sono stati studiati metodi a convergenza assicurata sotto ipotesi molto deboli, come ad esempio sotto la sola ipotesi della continuità della $f(x)$ su un intervallo $[a, b]$ di separazione di una soluzione. Il più noto di questi è il metodo di Dekker-Brent [5], che ha i vantaggi del metodo di bisezione e l'elevata efficienza del metodo ottenuto con l'interpolazione inversa (si veda l'esercizio 3.38).

10. Metodo di Aitken

Il metodo di Aitken consente di costruire, a partire da un metodo iterativo della forma (5), una successione che, sotto opportune ipotesi, è convergente anche quando il metodo di partenza non lo è, e che ha ordine di convergenza superiore a quello del metodo di partenza, quando questo è convergente.

Questo metodo si basa sul fatto che se una successione $\{x_i\}$ converge linearmente ad α, allora il suo comportamento in un intorno di α è simile a quello di una successione geometrica di ragione

$$\gamma = \lim_{i\to\infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|}\,.$$

Per la successione geometrica $x_i = \alpha + \gamma^i$ vale

$$x_i - \frac{(x_{i+1} - x_i)^2}{x_{i+2} - 2x_{i+1} + x_i} = \alpha + \gamma^i - \frac{\gamma^{2i}(\gamma-1)^2}{\gamma^i(\gamma^2 - 2\gamma + 1)} = \alpha.$$

Nel caso generale in cui la successione $\{x_i\}$ venga generata per mezzo del metodo iterativo (5), si consideri, a partire da un punto iniziale z_0, la successione $\{z_i\}$ così definita

$$\begin{aligned} \text{per}\quad i = 0, 1, \ldots \qquad & x_i = z_i, \quad x_{i+1} = g(x_i), \quad x_{i+2} = g(x_{i+1}), \\ & z_{i+1} = x_i - \frac{(x_{i+1} - x_i)^2}{x_{i+2} - 2x_{i+1} + x_i}\,. \end{aligned} \tag{46}$$

La successione $\{z_i\}$ è ancora della forma (5), infatti:

$$z_{i+1} = G(z_i),$$

dove

$$G(z) = z - \frac{(g(z) - z)^2}{g(g(z)) - 2g(z) + z}\,, \quad z \neq \alpha. \tag{47}$$

La funzione $G(z)$ non è definita in α, ma può essere estesa per continuità. Infatti vale il

3.33 Teorema. *Sia $g(x) \in C^1[a,b]$; se $g'(\alpha) \neq 1$ (cioè α è soluzione semplice dell'equazione (6)), la funzione definita dalla (47) ed estesa ponendo $G(\alpha) = \alpha$ è continua.*

Dim. Infatti per la regola di L'Hospital è

$$\begin{aligned} \lim_{z\to\alpha} z - G(z) &= \lim_{z\to\alpha} \frac{(g(z) - z)^2}{g(g(z)) - 2g(z) + z} = \lim_{z\to\alpha} \frac{2(g'(z) - 1)(g(z) - z)}{g'(g(z))g'(z) - 2g'(z) + 1} \\ &= \frac{2(g'(\alpha) - 1)(g(\alpha) - \alpha)}{g'(g(\alpha))g'(\alpha) - 2g'(\alpha) + 1} = \frac{2(g'(\alpha) - 1)(\alpha - \alpha)}{(g'(\alpha) - 1)^2} = 0. \end{aligned}$$

∎

Per la convergenza della successione (46) vale il seguente risultato.

3.34 Teorema. *Sia $g(x) \in C^2[a,b]$, se $g'(\alpha) \neq 0$ e $g'(\alpha) \neq 1$, esiste un intorno U di α tale che per ogni $z_0 \in U$ la successione $\{z_i\}$ è convergente, con ordine di convergenza almeno 2.*

Dim. Ponendo $\epsilon = z_i - \alpha$, risulta $z_{i+1} = G(\alpha + \epsilon)$, da cui

$$\frac{z_{i+1} - \alpha}{(z_i - \alpha)^2} = \frac{G(\alpha + \epsilon) - \alpha}{\epsilon^2}\,.$$

Perciò per dimostrare che la successione $\{z_i\}$ è convergente con ordine almeno 2, basta dimostrare che

$$q(\epsilon) = \frac{G(\alpha + \epsilon) - \alpha}{\epsilon^2}$$

è una funzione di ϵ limitata in modulo in un intorno di 0. Per la formula di Taylor è

$$g(\alpha + \epsilon) = \alpha + \theta\epsilon + \sigma\epsilon^2 = \alpha + \delta, \quad \text{dove} \quad \theta = g'(\alpha), \quad \sigma = \frac{g''(\xi_1)}{2}\,,$$

con ξ_1 compreso fra α e $\alpha + \epsilon$, e $\delta = \theta\epsilon + \sigma\epsilon^2$. Analogamente

$$g(\alpha + \delta) = \alpha + \theta\delta + \tau\delta^2, \quad \text{dove} \quad \tau = \frac{g''(\xi_2)}{2}\,,$$

con ξ_2 compreso fra α e $\alpha + \delta$. Per la (47) è

$$G(\alpha + \epsilon) = \alpha + \epsilon - \frac{(\delta - \epsilon)^2}{g(\alpha + \delta) - 2(\alpha + \delta) + \alpha + \epsilon} = \alpha + \epsilon^2 q(\epsilon),$$

dove

$$q(\epsilon) = \frac{(\theta + \sigma\epsilon)(\sigma\tau\epsilon + \theta\tau - \sigma)}{(\theta - 1)^2 + (\theta\sigma - 2\sigma + \theta^2\tau)\epsilon + 2\theta\sigma\tau\epsilon^2 + \sigma^2\tau\epsilon^3}\,.$$

Poiché

$$q(0) = \lim_{\epsilon \to 0} q(\epsilon) = \lim_{\epsilon \to 0} \frac{\theta(\theta\tau - \sigma)}{(\theta - 1)^2} = \frac{g''(\alpha)}{2}\,\frac{g'(\alpha)}{g'(\alpha) - 1}\,,$$

ne segue che $q(\epsilon)$ è limitato in modulo in un intorno dello 0. ∎

Il *metodo di Aitken* è definito dalla (46): nelle condizioni del teorema 3.34 è quindi un metodo almeno del secondo ordine che non richiede il calcolo di derivate prime. Naturalmente il teorema 3.34 garantisce solo la convergenza locale e può essere difficile determinare un opportuno punto iniziale.

3.35 Esempio. Le due funzioni

$$g_1(x) = \frac{2 - x^3}{4\cos x} \quad \text{e} \quad g_2(x) = \frac{2 - 4x\cos x}{x^2}$$

hanno come punto fisso la soluzione dell'equazione dell'esempio 3.1, compresa nell'intervallo (0,1). Come si è visto nell'esempio 3.5, i corrispondenti metodi iterativi

$$x_{i+1} = g_1(x_i) \quad \text{e} \quad x_{i+1} = g_2(x_i)$$

sono il primo convergente ad α con ordine di convergenza 1, e il secondo non convergente. Applicando il metodo di Aitken nel primo caso con $z_0 = 0$ si ottiene la successione

i	z_i
1	0.5366393
2	0.5368390
3	0.5368387

Applicando il metodo di Aitken nel secondo caso si ottiene una successione convergente scegliendo z_0 nell'intervallo $[0.51, 0.6]$. Ad esempio, scegliendo $z_0 = 0.6$ si ottiene la successione:

i	z_i	i	z_i
1	0.5995250	.	. . .
2	0.5990043	23	0.5372088
3	0.5984313	24	0.5368443
.	. . .	25	0.5368385

■

Per i casi non compresi nelle ipotesi del teorema 3.34, valgono i seguenti risultati (per la dimostrazione si veda l'esercizio 3.37). Sia $p \geq 2$ intero e $\rho > 0$:

a) se $g(x) \in C^{2p}[\alpha - \rho, \alpha + \rho]$ e se α è soluzione di molteplicità p dell'equazione $x = g(x)$ (quindi $g'(\alpha) = 1$), allora il metodo di Aitken è convergente con ordine di convergenza 1;

b) se $g(x) \in C^p[\alpha - \rho, \alpha + \rho]$ e se il metodo $x_{i+1} = g(x_i)$ è convergente con ordine di convergenza p (quindi per il teorema 3.15 è $g'(\alpha) = 0$), allora il metodo di Aitken è convergente con ordine di convergenza $2p - 1$.

Si osservi, tuttavia, che in quest'ultimo caso l'aumento dell'ordine di convergenza non corrisponde ad un vantaggio effettivo: infatti, per il metodo di ordine p si ha dalla (20):

$$|g(x_i) - \alpha| \leq \beta |x_i - \alpha|^p$$

$$|g(g(x_i)) - \alpha| \leq |x_{i+2} - \alpha| \leq \beta |x_{i+1} - \alpha|^p \leq \beta^{p+1} |x_i - \alpha|^{p^2},$$

mentre per il metodo di ordine $2p - 1$ si ha:

$$|G(z_i) - \alpha| \leq \beta' |z_i - \alpha|^{2p-1}.$$

Poiché per $p > 1$ fra i due esponenti vale la disuguaglianza $2p - 1 < p^2$, non appare conveniente usare il metodo di Aitken, che per ogni iterazione richiede un costo computazionale superiore a quello richiesto da due iterazioni del metodo di ordine p.

3.36 Esempio. Per calcolare la radice quadrata del numero $k > 0$, si applica il metodo di Aitken al metodo

$$x_{i+1} = g(x_i) = \frac{k}{x_i},$$

che non è convergente, ottenendo

$$G(z) = z - \frac{(g(z) - z)^2}{g(g(z)) - 2g(z) + z} = z - \frac{(k/z - z)^2}{z - 2k/z + z} = \frac{1}{2}\left(z + \frac{k}{z}\right).$$

Si riottiene quindi il metodo delle tangenti (si veda l'esempio 3.22), che è del secondo ordine. Applicando il metodo di Aitken al metodo delle tangenti, si ottiene una successione con ordine di convergenza 3. Nel caso $k = 2$ assumendo come punto iniziale $z_0 = 2$, si ottiene:

i	metodo delle tangenti	metodo di Aitken applicato al metodo delle tangenti
1	1.500000	1.399999
2	1.416666	1.414213
3	1.414215	
4	1.414213	

Si tenga però presente che 4 iterazioni del metodo delle tangenti hanno richiesto 12 operazioni, mentre 2 iterazioni del metodo di Aitken applicato alla successione ottenuta con il metodo delle tangenti hanno richiesto 26 operazioni. In questo caso particolare è possibile ricavare l'espressione esplicita del metodo di Aitken applicato al metodo delle tangenti, ottenendo

$$G(z) = \frac{6z^2 + 4}{z(z^2 + 6)}.$$

Ma anche ricorrendo a questa espressione sarebbero comunque richieste 12 operazioni per le 2 iterazioni. ∎

11. Metodi iterativi per i sistemi non lineari

Gran parte delle considerazioni svolte e dei risultati ottenuti nei paragrafi precedenti riguardo ai metodi iterativi per le singole equazioni possono essere estesi anche al caso di sistemi di equazioni non lineari. In tale generalizzazione saranno utilizzati metodi e concetti dell'algebra lineare, come la norma e il raggio spettrale di una matrice, per i quali si rimanda a [3].

Sia Ω un sottosinsieme aperto di $\mathbf{R}^n$; date le funzioni $\mathbf{f}$, $\mathbf{g} : \Omega \to \mathbf{R}^n$, il sistema da risolvere può essere rappresentato nella forma

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \tag{48}$$

oppure nella forma

$$\mathbf{x} = \mathbf{g}(\mathbf{x}), \tag{49}$$

dove

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}.$$

Se $\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x}) \in C^1(\Omega)$, si considerano le matrici jacobiane $J(\mathbf{x})$ e $H(\mathbf{x})$ rispettivamente di $\mathbf{f}$ e di $\mathbf{g}$:

$$[J(\mathbf{x})]_{rs} = \frac{\partial f_r(\mathbf{x})}{\partial x_s} \quad \text{e} \quad [H(\mathbf{x})]_{rs} = \frac{\partial g_r(\mathbf{x})}{\partial x_s}.$$

Il metodo iterativo (5) si generalizza nella forma

$$\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)}), \quad i = 0, 1, \ldots \tag{50}$$

La successione (50), formata da vettori di $\mathbf{R}^n$, è convergente alla soluzione $\boldsymbol{\alpha} \in \Omega$ se

$$\lim_{i \to \infty} \|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\| = 0,$$

per qualche norma vettoriale, e poiché le norme vettoriali su $\mathbf{R}^n$ sono tutte topologicamente equivalenti, la convergenza non dipende dalla particolare norma considerata.

Anche nel caso dei sistemi di equazioni non lineari, analogamente al teorema 3.2, se $\mathbf{g}(\mathbf{x})$ è continua e se la successione (50) appartiene ad Ω, allora il limite della successione $\{\mathbf{x}^{(i)}\}$, se esiste, è soluzione di (49). Vale il seguente teorema, analogo al teorema 3.3.

3.37 Teorema. *Sia* $\boldsymbol{\alpha}$ *soluzione del sistema (49). Se esiste un intorno* $S \subseteq \Omega$ *di* $\boldsymbol{\alpha}$*:*

$$S = \{\mathbf{x} \in \mathbf{R}^n \ : \ \|\mathbf{x} - \boldsymbol{\alpha}\|_\infty \leq \sigma, \ \sigma > 0\},$$

tale che $\mathbf{g}(\mathbf{x}) \in C^1(S)$ *e*

$$\|H(\mathbf{x})\|_\infty < 1, \quad per \quad \mathbf{x} \in S, \tag{51}$$

allora, scelto $\mathbf{x}^{(0)} \in S$*, la successione* $\{\mathbf{x}^{(i)}\}$ *ottenuta dalla (50) converge ad* $\boldsymbol{\alpha}$*.*

Dim. Posto

$$\lambda = \max_{\mathbf{x} \in S} \|H(\mathbf{x})\|_\infty,$$

dalla (51) risulta $\lambda < 1$. Si dimostra per induzione che

$$\|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty \leq \lambda^i \sigma. \tag{52}$$

Per $i = 0$ la (52) è vera. Per $i > 0$, sviluppando $g_r(\mathbf{x})$, $r = 1, \ldots, n$, in serie di Taylor attorno al punto $\boldsymbol{\alpha}$, si ha:

$$g_r(\mathbf{x}^{(i-1)}) = g_r(\boldsymbol{\alpha}) + \sum_{s=1}^{n} \frac{\partial g_r(\boldsymbol{\xi}_r)}{\partial x_s} \ (x_s^{(i-1)} - \alpha_s),$$

dove $\boldsymbol{\xi}_r$ è un punto del segmento che unisce $\mathbf{x}^{(i-1)}$ e $\boldsymbol{\alpha}$, e passando ai moduli si ha:

$$\begin{aligned} |g_r(\mathbf{x}^{(i-1)}) - g_r(\boldsymbol{\alpha})| &\leq \sum_{s=1}^{n} \left| \frac{\partial g_r(\boldsymbol{\xi}_r)}{\partial x_s} \right| |x_s^{(i-1)} - \alpha_s| \\ &\leq \sum_{s=1}^{n} \left| \frac{\partial g_r(\boldsymbol{\xi}_r)}{\partial x_s} \right| \|\mathbf{x}^{(i-1)} - \boldsymbol{\alpha}\|_\infty \\ &\leq \|H(\boldsymbol{\xi}_r)\|_\infty \|\mathbf{x}^{(i-1)} - \boldsymbol{\alpha}\|_\infty. \end{aligned}$$

Per l'ipotesi induttiva è $\boldsymbol{\xi}_r \in S$ e quindi vale

$$|g_r(\mathbf{x}^{(i-1)}) - g_r(\boldsymbol{\alpha})| \leq \lambda \|\mathbf{x}^{(i-1)} - \boldsymbol{\alpha}\|_\infty, \quad r = 1, \ldots, n,$$

e passando alle norme si ha:

$$\|\mathbf{g}(\mathbf{x}^{(i-1)}) - \mathbf{g}(\boldsymbol{\alpha})\|_\infty \leq \lambda \|\mathbf{x}^{(i-1)} - \boldsymbol{\alpha}\|_\infty.$$

Dalla (50) risulta

$$\|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty \leq \lambda \|\mathbf{x}^{(i-1)} - \boldsymbol{\alpha}\|_\infty,$$

da cui, per l'ipotesi induttiva, segue

$$\|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty \leq \lambda^i \sigma,$$

e, passando al limite, risulta:

$$\lim_{i\to\infty} \|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty = 0. \qquad \blacksquare$$

3.38 Esempio. Si consideri il sistema non lineare

$$\begin{cases} x_1 = g_1(x_1, x_2) = \dfrac{1}{4}\,(x_1^2 + x_2^2) \\ x_2 = g_2(x_1, x_2) = \sin(x_1 + 1). \end{cases}$$

Per individuare intorni delle soluzioni del sistema, si considerino le curve di equazioni $x_1 = g_1(x_1, x_2)$ (la circonferenza di centro (2,0) e raggio 2) e $x_2 = g_2(x_1, x_2)$ (la sinusoide con l'argomento traslato di 1), rappresentate nella figura 3.17.

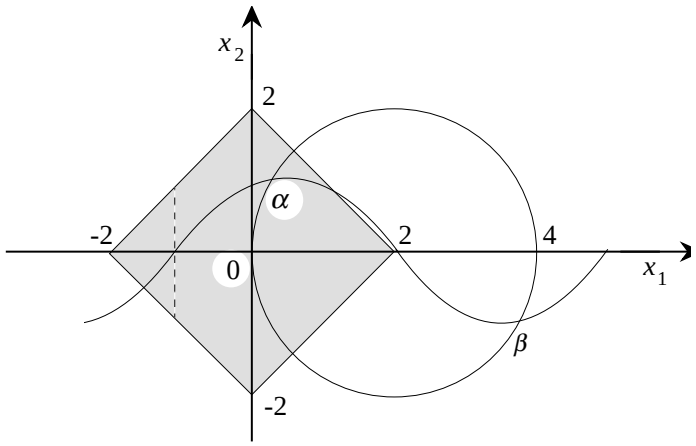


Fig. 3.17 - Grafici relativi al sistema dell'esempio 3.38.

Vi sono quindi due soluzioni reali $\boldsymbol{\alpha} \in [0, 0.5]\times[0.5, 1]$ e $\boldsymbol{\beta} \in [3, 4]\times[-1, -0.5]$. La matrice

$$H(\mathbf{x}) = \begin{bmatrix} x_1/2 & x_2/2 \\ \cos(x_1 + 1) & 0 \end{bmatrix}$$

è tale che $\|H(\mathbf{x})\|_\infty < 1$ se $|x_1|/2 + |x_2|/2 < 1$ e $|\cos(x_1 + 1)| < 1$. Il dominio in cui $\|H(\mathbf{x})\|_\infty < 1$, segnato in grigio sul grafico, è formato dai

punti appartenenti al quadrato di centro l'origine e un vertice nel punto $(0, 2)$, esclusi i punti di ascissa $x_1 = -1$ in cui $\cos(x_1 + 1) = 1$. Esiste allora un intorno S di $\boldsymbol{\alpha}$ con le proprietà riportate nel teorema 3.37, contenuto in tale dominio, e quindi per ogni $\mathbf{x}^{(0)} \in S$ la successione $\{\mathbf{x}^{(i)}\}$ definita da

$$\begin{cases} x_1^{(i+1)} = \dfrac{1}{4}\,({x_1^{(i)}}^2 + {x_2^{(i)}}^2) \\ x_2^{(i+1)} = \sin(x_1^{(i)} + 1) \end{cases}$$

converge ad $\boldsymbol{\alpha}$. Assumendo $\mathbf{x}^{(0)} = (0, 1) \in S$, si ottiene la successione:

i	$x_1^{(i)}$	$x_2^{(i)}$
1	0.2500000	0.8414710
2	0.1926433	0.9489847
3	0.2344208	0.9293481
4	0.2296602	0.9439572
⋮	⋮	⋮
17	0.2372839	0.9448983
18	0.2372841	0.9448983

Però anche partendo da $\mathbf{x}^{(0)} = (2, 0) \notin S$ si ottiene una successione convergente ad $\boldsymbol{\alpha}$. Non è invece possibile assicurare la convergenza a $\boldsymbol{\beta}$. ∎

Se $\mathbf{g}(\mathbf{x}) \in C^1(\Omega)$, allora la funzione $\mathbf{g}$ è *totalmente differenziabile* in Ω, cioè per ogni $\mathbf{x} \in \Omega$ è

$$\lim_{\mathbf{x}' \to \mathbf{x}} \frac{\|\mathbf{g}(\mathbf{x}') - \mathbf{g}(\mathbf{x}) - H(\mathbf{x})(\mathbf{x}' - \mathbf{x})\|}{\|\mathbf{x}' - \mathbf{x}\|} = 0, \tag{53}$$

dove $\|\ .\ \|$ è una qualsiasi norma vettoriale [9]. Il seguente teorema fornisce una condizione di convergenza che utilizza una generica norma matriciale indotta e che è quindi più generale di quella del teorema 3.37.

3.39 Teorema. *Sia $\mathbf{g}(\mathbf{x}) \in C^1(\Omega)$ e sia $\boldsymbol{\alpha}$ una soluzione del sistema (49). Si indichi con $\rho(H(\boldsymbol{\alpha}))$ il raggio spettrale di $H(\boldsymbol{\alpha})$. Se $\rho(H(\boldsymbol{\alpha})) < 1$, esistono una norma vettoriale $\|\ .\ \|$ e un $\sigma > 0$ tali che, posto*

$$S = \{\mathbf{x} \in \mathbf{R}^n \ :\ \|\mathbf{x} - \boldsymbol{\alpha}\| \leq \sigma\},$$

risulta $S \subseteq \Omega$ e la successione (50) converge per ogni $\mathbf{x}^{(0)} \in S$.

Dim. Sia ϵ tale che

$$0 < \epsilon < \frac{1 - \rho(H(\boldsymbol{\alpha}))}{2}\,.$$

Esiste allora una norma matriciale indotta tale che

$$\|H(\boldsymbol{\alpha})\| \leq \rho(H(\boldsymbol{\alpha})) + \epsilon$$

(si veda [3], teorema 3.12). Per la (53), applicata con $\mathbf{x} = \boldsymbol{\alpha}$, esiste $\sigma > 0$ tale che

$$\|\mathbf{g}(\mathbf{x}') - \mathbf{g}(\boldsymbol{\alpha}) - H(\boldsymbol{\alpha})(\mathbf{x}' - \boldsymbol{\alpha})\| \leq \epsilon\|\mathbf{x}' - \boldsymbol{\alpha}\|$$

per ogni $\mathbf{x}' \in S$ e $S \subseteq \Omega$. Quindi per $\mathbf{x}' \in S$ risulta

$$\|\mathbf{g}(\mathbf{x}')-\mathbf{g}(\boldsymbol{\alpha})\| \leq \|\mathbf{g}(\mathbf{x}')-\mathbf{g}(\boldsymbol{\alpha})-H(\boldsymbol{\alpha})(\mathbf{x}'-\boldsymbol{\alpha})\|+\|H(\boldsymbol{\alpha})(\mathbf{x}'-\boldsymbol{\alpha})\| \leq \lambda\|\mathbf{x}'-\boldsymbol{\alpha}\|$$

dove $\lambda = \rho(H(\boldsymbol{\alpha})) + 2\epsilon$. Poiché per la scelta di ϵ è $\lambda < 1$, si dimostra per induzione, analogamente a quanto fatto nel teorema 3.37, che per $\mathbf{x}^{(0)} \in S$ è

$$\|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\| \leq \lambda^i\|\mathbf{x}^{(0)} - \boldsymbol{\alpha}\|,$$

da cui segue che la successione $\{\mathbf{x}^{(i)}\}$ è localmente convergente. ∎

3.40 Esempio. Il sistema non lineare

$$\begin{cases} x_1 = \dfrac{1}{4}\,(x_1^2 + x_2^2) \\ \\ x_2 = -x_1 + 2 \end{cases}$$

ha due soluzioni reali $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$, punti di intersezione della circonferenza di centro (2,0) e raggio 2 con la retta passante per il centro della circonferenza e per il punto (0,2) (si veda la figura 3.18).

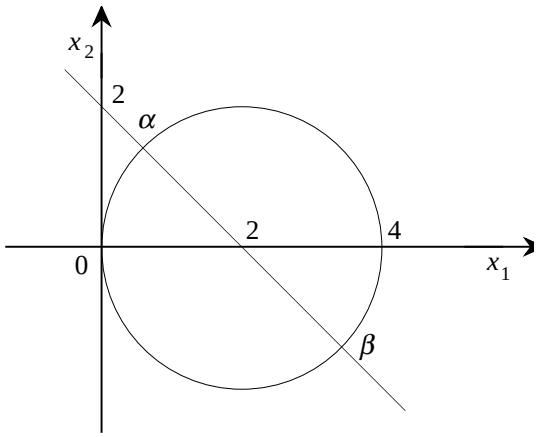


Fig. 3.18 - Grafici relativi al sistema dell'esempio 3.40.

La matrice

$$H(\mathbf{x}) = \begin{bmatrix} x_1/2 & x_2/2 \\ -1 & 0 \end{bmatrix}$$

è tale che $\|H(\mathbf{x})\|_\infty \geq 1$ per ogni $\mathbf{x}$ e nella soluzione è $\|H(\boldsymbol{\alpha})\|_\infty = 1$. Ma

$$\rho(H(\boldsymbol{\alpha})) = \sqrt{\alpha_2/2} < 1, \quad \text{dove} \quad \boldsymbol{\alpha} = (\alpha_1, \alpha_2),$$

e quindi, per opportune scelte di $\mathbf{x}^{(0)}$, il metodo iterativo

$$\begin{cases} x_1^{(i+1)} = \dfrac{1}{4}\,({x_1^{(i)}}^2 + {x_2^{(i)}}^2) \\ x_2^{(i+1)} = -x_1^{(i)} + 2 \end{cases}$$

converge ad $\boldsymbol{\alpha}$. Assumendo $\mathbf{x}^{(0)} = (1, 1)$, si ottiene la successione:

i	$x_1^{(i)}$	$x_2^{(i)}$
1	0.5000000	1.000000
2	0.3125000	1.500000
3	0.5869141	1.687500
4	0.7980311	1.413086
⋮	⋮	⋮
77	0.5857866	1.414213
78	0.5857859	1.414213

∎

12. Metodo di Newton-Raphson

Analogamente al caso delle singole equazioni, da un sistema dato nella forma (48) può essere ottenuto un sistema della forma (49) con le stesse soluzioni ponendo

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - [C(\mathbf{x})]^{-1}\mathbf{f}(\mathbf{x}),$$

dove $C(\mathbf{x})$ è una matrice i cui elementi sono funzioni di $\mathbf{x}$ e il cui determinante non è identicamente nullo. Se $\mathbf{f} \in C^1(\Omega)$, scegliendo $C(\mathbf{x}) = J(\mathbf{x})$ si ottiene il metodo di *Newton-Raphson*, che generalizza ai sistemi di equazioni il metodo delle tangenti:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - [J(\mathbf{x}^{(i)})]^{-1}\mathbf{f}(\mathbf{x}^{(i)}), \quad \det J(\mathbf{x}^{(i)}) \neq 0. \tag{54}$$

Il vettore $\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}$ è allora soluzione del sistema lineare di n equazioni in n incognite

$$J(\mathbf{x}^{(i)})[\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}] = -\mathbf{f}(\mathbf{x}^{(i)}). \tag{55}$$

Ad esempio, per la risoluzione del sistema di 2 equazioni

$$\begin{cases} f_1(x_1, x_2) = 0 \\ \\ f_2(x_1, x_2) = 0, \end{cases} \tag{56}$$

con il metodo di Newton-Raphson si deve risolvere ad ogni iterazione il sistema lineare:

$$\begin{cases} \dfrac{\partial f_1(\mathbf{x}^{(i)})}{\partial x_1}\,\theta_1 + \dfrac{\partial f_1(\mathbf{x}^{(i)})}{\partial x_2}\,\theta_2 = -f_1(\mathbf{x}^{(i)}) \\ \dfrac{\partial f_2(\mathbf{x}^{(i)})}{\partial x_1}\,\theta_1 + \dfrac{\partial f_2(\mathbf{x}^{(i)})}{\partial x_2}\,\theta_2 = -f_2(\mathbf{x}^{(i)}), \end{cases} \tag{57}$$

in cui con θ_1 e θ_2 sono state indicate le componenti del vettore $\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}$. Del metodo di Newton-Raphson si può dare la seguente interpretazione geometrica: le equazioni (56) rappresentano due curve, intersezioni delle superfici $z = f_1(x_1, x_2)$ e $z = f_2(x_1, x_2)$ con il piano $z = 0$; fissato in questo piano il punto $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)})$, si considerano i piani tangenti alle due superfici rispettivamente nei punti $(\mathbf{x}^{(0)}, f_1(\mathbf{x}^{(0)}))$ e $(\mathbf{x}^{(0)}, f_2(\mathbf{x}^{(0)}))$:

$$z = \frac{\partial f_1(\mathbf{x}^{(0)})}{\partial x_1}\,(x_1 - x_1^{(0)}) + \frac{\partial f_1(\mathbf{x}^{(0)})}{\partial x_2}\,(x_2 - x_2^{(0)}) + f_1(\mathbf{x}^{(0)}),$$
$$z = \frac{\partial f_2(\mathbf{x}^{(0)})}{\partial x_1}\,(x_1 - x_1^{(0)}) + \frac{\partial f_2(\mathbf{x}^{(0)})}{\partial x_2}\,(x_2 - x_2^{(0)}) + f_2(\mathbf{x}^{(0)});$$

si considerano poi le rette intersezioni di questi due piani con il piano $z = 0$. Il punto $\mathbf{x}^{(1)} = (x_1^{(1)}, x_2^{(1)})$, soluzione di (57) per $i = 0$, è l'intersezione di queste due rette. Per i passi successivi si procede in modo analogo.

3.41 Teorema. *Sia $\mathbf{f}(\mathbf{x}) \in C^2(\Omega)$ e sia $\boldsymbol{\alpha} \in \Omega$ soluzione del sistema (48). Se $J(\mathbf{x})$ è non singolare in Ω, esiste un intorno $S \subseteq \Omega$ di $\boldsymbol{\alpha}$ tale che, se $\mathbf{x}^{(0)} \in S$, la successione (54) converge ed inoltre per ogni norma vettoriale esiste una costante β tale che*

$$\|\mathbf{x}^{(i+1)} - \boldsymbol{\alpha}\| \leq \beta \|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|^2, \quad i = 0, 1, \ldots \tag{58}$$

Dim. Per il metodo di Newton-Raphson è

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - [J(\mathbf{x})]^{-1}\mathbf{f}(\mathbf{x}),$$

da cui, ponendo $K(\mathbf{x}) = [J(\mathbf{x})]^{-1}$, si ha:

$$g_r(\mathbf{x}) = x_r - \sum_{s=1}^{n} k_{rs}(\mathbf{x}) f_s(\mathbf{x}), \quad r = 1, \ldots, n.$$

Gli elementi della matrice $K(\mathbf{x})$, essendo funzioni razionali degli elementi di $J(\mathbf{x})$, con denominatore uguale a $\det J(\mathbf{x})$, sono derivabili con continuità. Derivando rispetto a x_t, $t = 1, \ldots, n$, si ha

$$\begin{aligned} h_{rt}(\mathbf{x}) = \frac{\partial g_r(\mathbf{x})}{\partial x_t} &= \delta_{rt} - \sum_{s=1}^{n} \frac{\partial k_{rs}(\mathbf{x})}{\partial x_t} f_s(\mathbf{x}) - \sum_{s=1}^{n} k_{rs}(\mathbf{x}) \frac{\partial f_s(\mathbf{x})}{\partial x_t} \\ &= -\sum_{s=1}^{n} \frac{\partial k_{rs}(\mathbf{x})}{\partial x_t} f_s(\mathbf{x}), \quad \text{dove} \quad \delta_{rt} = \begin{cases} 1 & \text{per } r = t, \\ 0 & \text{altrimenti,} \end{cases} \end{aligned}$$

perché

$$\sum_{s=1}^{n} k_{rs}(\mathbf{x}) \frac{\partial f_s(\mathbf{x})}{\partial x_t} = \delta_{rt}.$$

Si ha allora $H(\boldsymbol{\alpha}) = O$ e quindi $\rho(H(\boldsymbol{\alpha})) = 0$, allora per il teorema 3.39 esiste un intorno S di $\boldsymbol{\alpha}$ in cui si ha convergenza. Dalla (55) si ha

$$J(\mathbf{x}^{(i)})\,(\mathbf{x}^{(i+1)} - \boldsymbol{\alpha}) = J(\mathbf{x}^{(i)})\,(\mathbf{x}^{(i)} - \boldsymbol{\alpha}) - \mathbf{f}(\mathbf{x}^{(i)}). \tag{59}$$

Indicate con $S_r(x)$, $r = 1, \ldots, n$, le matrici hessiane delle funzioni $f_r(\mathbf{x})$, i cui elementi sono

$$(S_r(\mathbf{x}))_{st} = \frac{\partial^2 f_r(\mathbf{x})}{\partial x_s \partial x_t}\,, \tag{60}$$

dalla formula di Taylor si ha

$$\mathbf{f}(\boldsymbol{\alpha}) = \mathbf{f}(\mathbf{x}^{(i)}) + J(\mathbf{x}^{(i)})\,(\boldsymbol{\alpha} - \mathbf{x}^{(i)}) + \mathbf{v}(\boldsymbol{\xi}), \tag{61}$$

in cui $\mathbf{v}(\boldsymbol{\xi})$ è un vettore di resti, la cui r-esima componente è

$$v_r(\boldsymbol{\xi}) = \frac{1}{2}\,(\mathbf{x}^{(i)} - \boldsymbol{\alpha})^T S_r(\boldsymbol{\xi})\,(\mathbf{x}^{(i)} - \boldsymbol{\alpha}),$$

e $\boldsymbol{\xi}$ appartiene al segmento di estremi $\boldsymbol{\alpha}$ e $\mathbf{x}^{(i)}$, e quindi è

$$|v_r(\boldsymbol{\xi})| \leq \frac{n}{2}\,\|S_r(\boldsymbol{\xi})\|_\infty\,\|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty^2.$$

Sostituendo la (61) nella (59), poiché $\mathbf{f}(\boldsymbol{\alpha}) = \mathbf{0}$, risulta

$$\mathbf{x}^{(i+1)} - \boldsymbol{\alpha} = K(\mathbf{x}^{(i)})\,\mathbf{v}(\boldsymbol{\xi}),$$

da cui

$$\|\mathbf{x}^{(i+1)} - \boldsymbol{\alpha}\|_\infty \leq \gamma \, \|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty^2,$$

dove

$$\gamma = \frac{n}{2} \max_{\mathbf{x}\in S} \|K(\mathbf{x})\|_\infty \max_{\substack{\mathbf{x}\in S \\ r=1,\ldots,n}} \|S_r(\mathbf{x})\|_\infty.$$

Per l'equivalenza topologica delle norme vettoriali ne segue che esiste una costante β per cui vale la (58) per ogni norma vettoriale. ∎

La (58), che esprime la riduzione dell'errore all'i-esima iterazione, può essere considerata una generalizzazione della (20) per $p = 2$. Analogamente al caso di una singola equazione, se $\det J(\boldsymbol{\alpha}) \neq 0$ si può dire che il metodo di Newton-Raphson per risolvere sistemi di equazioni non lineari è almeno del secondo ordine.

3.42 Esempio. Per il sistema non lineare dell'esempio 3.38 si ha

$$\begin{cases} f_1(x_1, x_2) = x_1 - \dfrac{1}{4}\,(x_1^2 + x_2^2) = 0 \\ \\ f_2(x_1, x_2) = x_2 - \sin(x_1 + 1) = 0, \end{cases}$$

e

$$J(\mathbf{x}) = \begin{bmatrix} 1 - x_1/2 & -x_2/2 \\ \\ -\cos(x_1 + 1) & 1 \end{bmatrix}.$$

Poiché $2 \det J(\mathbf{x}) = 2 - x_1 - x_2 \cos(x_1 + 1)$, la matrice $J(\mathbf{x})$ è non singolare per i punti (x_1, x_2) tali che

$$x_2 \neq \frac{2 - x_1}{\cos(x_1 + 1)},$$

cioè i punti non appartenenti alla curva il cui grafico è rappresentato nella figura 3.19. In particolare risulta $\det J(\mathbf{x}) \neq 0$ in $\boldsymbol{\alpha}$ e in $\boldsymbol{\beta}$.

Con il metodo di Newton-Raphson, assumendo come punto iniziale $\mathbf{x}^{(0)} = (0, 1)$, si ottiene la successione convergente ad $\boldsymbol{\alpha}$:

i	$x_1^{(i)}$	$x_2^{(i)}$
1	0.2339326	0.9678653
2	0.2371014	0.9448433
3	0.2372839	0.9448981
4	0.2372841	0.9448984

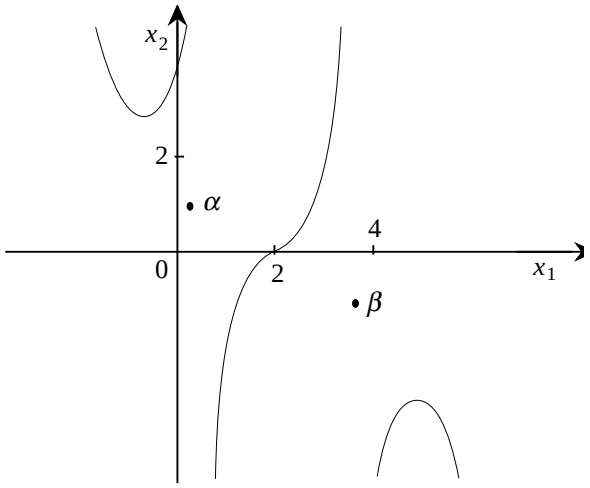


Fig. 3.19 - Grafico della funzione $x_2 = \dfrac{2 - x_1}{\cos(x_1 + 1)}$.

Con il metodo di Newton-Raphson si può approssimare anche la radice $\boldsymbol{\beta}$. Assumendo come punto iniziale $\mathbf{x}^{(0)} = (4, -1)$, si ottiene la successione convergente:

i	$x_1^{(i)}$	$x_2^{(i)}$
1	3.732614	-1.034771
2	3.732522	-0.9997974
3	3.732164	-0.9998045
4	3.732164	-0.9998045

∎

3.43 Esempio. Si consideri il sistema non lineare

$$\begin{cases} x_1 = f_1(x_1, x_2) = \dfrac{1}{3}\,(x_1 - x_2) + x_1^2 = 0 \\ x_2 = f_2(x_1, x_2) = \dfrac{1}{3}\,(-x_1 + x_2) + x_1 x_2 = 0, \end{cases}$$

che ha le due soluzioni $\boldsymbol{\alpha} = (0, 0)$ e $\boldsymbol{\beta} = \Big(-\dfrac{2}{3}, \dfrac{2}{3}\Big)$, come risulta anche dalla figura 3.20, in cui sono rappresentate le curve di equazioni $f_1(\mathbf{x}) = 0$ e $f_2(\mathbf{x}) = 0$.

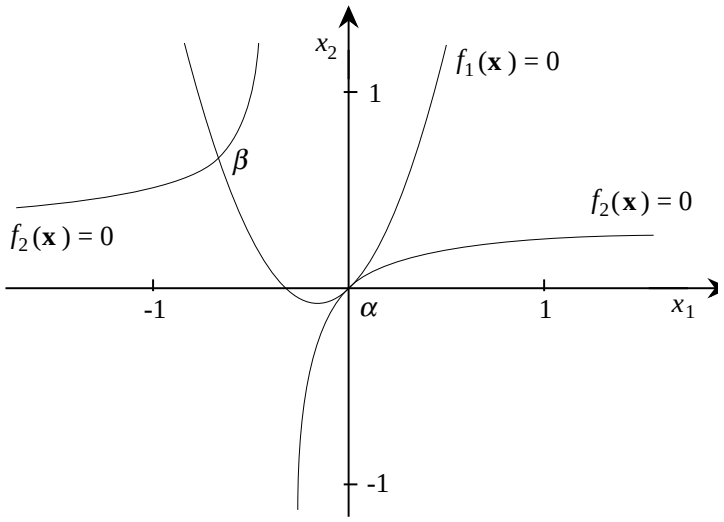


Fig. 3.20 - Grafici relativi al sistema dell'esempio 3.43.

È

$$J(\mathbf{x}) = \begin{bmatrix} \frac{1}{3} + 2x_1 & -\frac{1}{3} \\ -\frac{1}{3} + x_2 & \frac{1}{3} + x_1 \end{bmatrix},$$

quindi $\det J(\boldsymbol{\alpha}) = 0$ e $\det J(\boldsymbol{\beta}) \neq 0$. Applicando il metodo di Newton-Raphson a partire dal punto $\mathbf{x}^{(0)} = (0.5, 0.8)$, si ottiene la successione:

i	$x_1^{(i)}$	$x_2^{(i)}$
1	0.2697367	0.3289472
2	0.1399398	0.1481472
3	$0.7080776\ 10^{-1}$	$0.7151115\ 10^{-1}$
4	$0.3548320\ 10^{-1}$	$0.3551690\ 10^{-1}$
$\vdots$	$\vdots$	$\vdots$
18	$0.2145697\ 10^{-5}$	$0.2145697\ 10^{-5}$
19	$0.1072881\ 10^{-5}$	$0.1072881\ 10^{-5}$

che converge ad $\boldsymbol{\alpha}$. Applicando il metodo a partire dal punto $\mathbf{x}^{(0)} = (-2, 2)$, si ottiene la successione:

i	$x_1^{(i)}$	$x_2^{(i)}$
1	-1.199999	1.199999
2	-0.8307689	0.8307689
3	-0.6937348	0.6937348
4	-0.6676828	0.6676828
5	-0.6666679	0.6666679

che converge ad $\boldsymbol{\beta}$. Mentre la seconda successione converge con ordine almeno 2, la prima converge molto più lentamente (linearmente) poiché per $\boldsymbol{\alpha}$ non sono verificate le ipotesi del teorema 3.41: risulta infatti

$$\|\mathbf{x}^{(i+1)} - \boldsymbol{\alpha}\|_\infty \approx 0.5 \, \|\mathbf{x}^{(i)} - \boldsymbol{\alpha}\|_\infty.$$ ∎

Il teorema 3.41 fornisce condizioni di convergenza locale per il metodo di Newton-Raphson. Però, analogamente al caso di una singola equazione, anche per i sistemi è possibile dare delle condizioni sufficienti di convergenza su insiemi di $\mathbf{R}^n$. Per questo è necessario premettere alcune definizioni e un teorema. Nel seguito le disuguaglianze fra vettori o fra matrici vanno intese componente per componente.

3.44 Definizione. Un insieme D di $\mathbf{R}^n$ si dice *convesso* se

$$\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in D$$

per ogni coppia $\mathbf{x},\ \mathbf{y} \in D$ e ogni $\lambda \in [0,1]$. ∎

3.45 Definizione. Sia D un insieme convesso di $\mathbf{R}^n$. Una funzione $\mathbf{f} : D \to \mathbf{R}^n$ si dice *convessa* su D se

$$\mathbf{f}(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda\mathbf{f}(\mathbf{x}) + (1-\lambda)\mathbf{f}(\mathbf{y}), \tag{62}$$

per ogni coppia $\mathbf{x},\ \mathbf{y} \in D$ e ogni $\lambda \in [0,1]$. ∎

Dalla definizione 3.45 segue che la funzione $\mathbf{f}(\mathbf{x})$ è convessa su D se e solo se tutte le funzioni $f_r(\mathbf{x})$, $r = 1, \ldots, n$, sono convesse su D. Se $\mathbf{f}(\mathbf{x}) \in C^2(D)$, la condizione di convessità definita in 3.45 è equivalente alla condizione che le matrici hessiane $S_r(\mathbf{x})$, $r = 1, \ldots, n$, introdotte in (60), siano semidefinite positive per $\mathbf{x} \in D$ (si veda l'esercizio 3.44).

3.46 Teorema. *Sia D un insieme convesso e $\mathbf{f}(\mathbf{x}) \in C^1(D)$, allora $\mathbf{f}(\mathbf{x})$ è convessa su D se e solo se*

$$\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) \geq J(\mathbf{u})(\mathbf{v} - \mathbf{u}) \tag{63}$$

per ogni $\mathbf{u},\ \mathbf{v} \in D$.

Dim. Siano $\mathbf{x},\ \mathbf{y} \in D$, $\lambda \in [0,1]$ e $\mathbf{z}(\lambda) = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. Poiché D è convesso, $\mathbf{z}(\lambda) \in D$. Dalla (63), si ha:

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{z}(\lambda)) \geq J(\mathbf{z}(\lambda))(\mathbf{x} - \mathbf{z}(\lambda)),$$

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z}(\lambda)) \geq J(\mathbf{z}(\lambda))(\mathbf{y} - \mathbf{z}(\lambda)).$$

Moltiplicando entrambi i membri di queste due disuguaglianze rispettivamente per λ e $(1-\lambda)$, e sommando membro a membro, si ottiene

$$\lambda\mathbf{f}(\mathbf{x}) + (1-\lambda)\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z}(\lambda)) \geq J(\mathbf{z}(\lambda))(\lambda\mathbf{x} + (1-\lambda)\mathbf{y} - \mathbf{z}(\lambda)) = \mathbf{0},$$

da cui segue la disuguaglianza (62). Viceversa, se $\mathbf{f}(\mathbf{x})$ è convessa su D, da (62) si ha per $\mathbf{u},\ \mathbf{v} \in D$, $\lambda \in [0,1]$ e $\mathbf{w}(\lambda) = \lambda\mathbf{v} + (1-\lambda)\mathbf{u} \in D$:

$$\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) \geq \frac{1}{\lambda}[\mathbf{f}(\mathbf{w}(\lambda)) - \mathbf{f}(\mathbf{u})]. \tag{64}$$

Poiché $\lim_{\lambda\to 0} \mathbf{w}(\lambda) = \mathbf{u}$, dalla (53) segue che

$$\lim_{\lambda\to 0} \frac{\|\mathbf{f}(\mathbf{w}(\lambda)) - \mathbf{f}(\mathbf{u}) - J(\mathbf{u})(\mathbf{w}(\lambda) - \mathbf{u})\|}{\|\mathbf{w}(\lambda) - \mathbf{u}\|} = 0,$$

e poiché $\|\mathbf{w}(\lambda) - \mathbf{u}\| = |\lambda|\|\mathbf{v} - \mathbf{u}\|$, si ottiene

$$\lim_{\lambda\to 0} \frac{1}{\lambda} \|\mathbf{f}(\mathbf{w}(\lambda)) - \mathbf{f}(\mathbf{u}) - \lambda J(\mathbf{u})(\mathbf{v} - \mathbf{u})\| = 0,$$

e quindi

$$\lim_{\lambda\to 0} \frac{1}{\lambda} [\mathbf{f}(\mathbf{w}(\lambda)) - \mathbf{f}(\mathbf{u})] = J(\mathbf{u})(\mathbf{v} - \mathbf{u}). \tag{65}$$

Da (64) e (65) si ottiene

$$\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) \geq J(\mathbf{u})(\mathbf{v} - \mathbf{u}). \qquad \blacksquare$$

Per il metodo di Newton-Raphson vale allora il seguente teorema.

3.47 Teorema. *Siano $a_i, b_i \in \mathbf{R}$ per $i = 1, \ldots, n$, e sia D l'insieme*

$$D = \prod_{j=1}^{n} [a_j, b_j]$$

(cioè D è un intervallo di $\mathbf{R}^n$). Siano $\mathbf{f}(\mathbf{x}) \in C^1(D)$ una funzione convessa su D e $\boldsymbol{\alpha} \in D$ la soluzione del sistema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Se per $\mathbf{x} \in D$, $J(\mathbf{x})$ è non singolare e $[J(\mathbf{x})]^{-1} \geq O$, per ogni scelta di $\mathbf{x}^{(0)} \in D$ tale che $\mathbf{f}(\mathbf{x}^{(0)}) \geq \mathbf{0}$, la successione (54) converge ad $\boldsymbol{\alpha}$ ed $\boldsymbol{\alpha}$ è l'unica soluzione del sistema nell'insieme D.

Dim. Si dimostra per induzione su i che

$$\boldsymbol{\alpha} \leq \mathbf{x}^{(i+1)} \leq \mathbf{x}^{(i)} \quad \text{e} \quad \mathbf{f}(\mathbf{x}^{(i+1)}) \geq \mathbf{0}, \quad i = 0, 1, \ldots \tag{66}$$

Per $i = 0$, per l'ipotesi di convessità si ha

$$\mathbf{f}(\boldsymbol{\alpha}) - \mathbf{f}(\mathbf{x}^{(0)}) \geq J(\mathbf{x}^{(0)})(\boldsymbol{\alpha} - \mathbf{x}^{(0)}),$$

da cui si ha

$$-[J(\mathbf{x}^{(0)})]^{-1}\mathbf{f}(\mathbf{x}^{(0)}) \geq \boldsymbol{\alpha} - \mathbf{x}^{(0)},$$

e quindi, poiché $\mathbf{f}(\mathbf{x}^{(0)}) \geq \mathbf{0}$ per ipotesi, ne segue

$$\mathbf{x}^{(0)} \geq \mathbf{x}^{(1)} = \mathbf{x}^{(0)} - [J(\mathbf{x}^{(0)})]^{-1}\mathbf{f}(\mathbf{x}^{(0)}) \geq \boldsymbol{\alpha},$$

quindi $\mathbf{x}^{(1)} \in D$. Inoltre per l'ipotesi di convessità,

$$\mathbf{f}(\mathbf{x}^{(1)}) - \mathbf{f}(\mathbf{x}^{(0)}) \geq J(\mathbf{x}^{(0)})(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = -\mathbf{f}(\mathbf{x}^{(0)}),$$

per cui $\mathbf{f}(\mathbf{x}^{(1)}) \geq \mathbf{0}$. Per $i > 0$, per l'ipotesi di convessità si ha

$$\mathbf{f}(\boldsymbol{\alpha}) - \mathbf{f}(\mathbf{x}^{(i)}) \geq J(\mathbf{x}^{(i)})(\boldsymbol{\alpha} - \mathbf{x}^{(i)}),$$

da cui si ha

$$-[J(\mathbf{x}^{(i)})]^{-1}\mathbf{f}(\mathbf{x}^{(i)}) \geq \boldsymbol{\alpha} - \mathbf{x}^{(i)},$$

e quindi, poiché $\mathbf{x}^{(i)} \in D$ e $\mathbf{f}(\mathbf{x}^{(i)}) \geq \mathbf{0}$ per l'ipotesi induttiva, segue che

$$\mathbf{x}^{(i)} \geq \mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - [J(\mathbf{x}^{(i)})]^{-1}\mathbf{f}(\mathbf{x}^{(i)}) \geq \boldsymbol{\alpha}.$$

Quindi $\mathbf{x}^{(i+1)} \in D$. Inoltre per l'ipotesi di convessità si ha:

$$\mathbf{f}(\mathbf{x}^{(i+1)}) - \mathbf{f}(\mathbf{x}^{(i)}) \geq J(\mathbf{x}^{(i)})(\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}) = -\mathbf{f}(\mathbf{x}^{(i)}),$$

per cui $\mathbf{f}(\mathbf{x}^{(i+1)}) \geq \mathbf{0}$. Dalla (66) segue allora che le successioni $\{x_r^{(i)}\}$, $r = 1, \ldots, n$ sono monotone e convergenti, e posto

$$\boldsymbol{\beta} = \lim_{i \to \infty} \mathbf{x}^{(i)},$$

poiché

$$\mathbf{f}(\mathbf{x}^{(i)}) = J(\mathbf{x}^{(i)})[\mathbf{x}^{(i)} - \mathbf{x}^{(i+1)}],$$

risulta $\mathbf{f}(\boldsymbol{\beta}) = \mathbf{0}$. Essendo anche $\mathbf{f}(\boldsymbol{\alpha}) = \mathbf{0}$, dalla (63) si ha

$$\mathbf{0} = \mathbf{f}(\boldsymbol{\beta}) - \mathbf{f}(\boldsymbol{\alpha}) \geq J(\boldsymbol{\alpha})(\boldsymbol{\beta} - \boldsymbol{\alpha}),$$

$$\mathbf{0} = \mathbf{f}(\boldsymbol{\alpha}) - \mathbf{f}(\boldsymbol{\beta}) \geq J(\boldsymbol{\beta})(\boldsymbol{\alpha} - \boldsymbol{\beta}),$$

e poiché $[J(\boldsymbol{\alpha})]^{-1} \geq 0$ e $[J(\boldsymbol{\beta})]^{-1} \geq 0$, ne segue che

$$\mathbf{0} \geq \boldsymbol{\beta} - \boldsymbol{\alpha} \quad \text{e} \quad \mathbf{0} \geq \boldsymbol{\alpha} - \boldsymbol{\beta},$$

cioè $\boldsymbol{\alpha} = \boldsymbol{\beta}$. ∎

Il teorema 3.47, nel caso particolare $n = 1$, assicura la convergenza del metodo delle tangenti nell'ipotesi che la funzione $f(x) \in C^1[a, b]$ sia convessa, $\alpha \in [a, b]$, e si scelga un punto x_0 tale che $f(x_0) > 0$. Se $f(x) \in C^2[a, b]$, allora la condizione di convessità è equivalente a $f''(x) > 0$, e le ipotesi del teorema 3.47 si riducono a quelle del teorema 3.25.

Teoremi analoghi al teorema 3.47 per la convergenza del metodo di Newton-Raphson possono essere dimostrati sotto le seguenti ipotesi:

1) $\mathbf{f}(\mathbf{x}) \in C^1(D)$ e convessa, $[J(\mathbf{x})]^{-1} \leq O$ per $\mathbf{x} \in D$. Si ha convergenza per ogni scelta di $\mathbf{x}^{(0)} \in D$ per cui $\mathbf{f}(\mathbf{x}^{(0)}) \geq \mathbf{0}$.
2) $\mathbf{f}(\mathbf{x}) \in C^1(D)$ e concava (cioè $-\mathbf{f}(\mathbf{x})$ convessa), $[J(\mathbf{x})]^{-1} \leq O$ oppure $[J(\mathbf{x})]^{-1} \geq O$ per $\mathbf{x} \in D$. Si ha convergenza per ogni scelta di $\mathbf{x}^{(0)} \in D$ per cui $\mathbf{f}(\mathbf{x}^{(0)}) \leq \mathbf{0}$.

3.48 Esempio. Il sistema non lineare

$$\begin{cases} f_1(x_1, x_2) = x_1^2 - \sin x_2 - 25 = 0 \\ \\ f_2(x_1, x_2) = -\cos x_1 + x_2 = 0 \end{cases}$$

per $x_1 > 0$ ha una sola soluzione $\boldsymbol{\alpha}$ appartenente all'insieme $D = [4.6, 5.2] \times [0, 0.8]$, come risulta anche nella figura 3.21. Poiché

$$J(\mathbf{x}) = \begin{bmatrix} 2x_1 & -\cos x_2 \\ \\ \sin x_1 & 1 \end{bmatrix},$$

è $\det(J(\mathbf{x})) = 2x_1 + \sin x_1 \cos x_2 \neq 0$, e quindi $J(\mathbf{x})$ è non singolare per $\mathbf{x} \in D$.

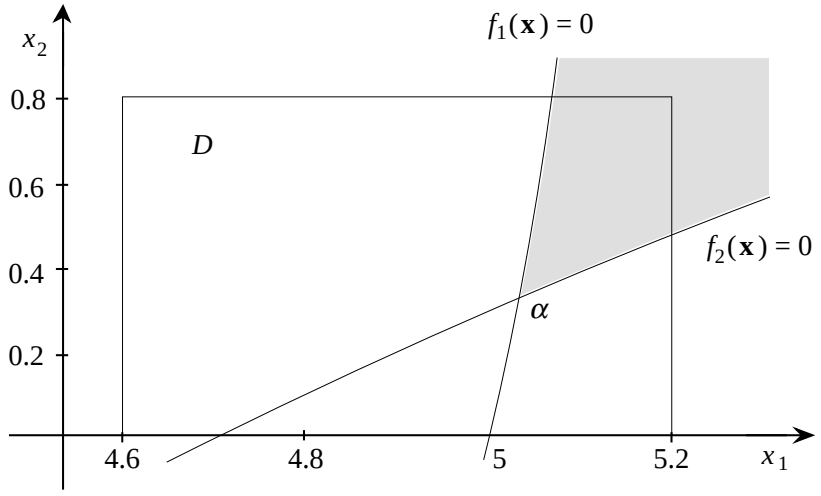


Fig. 3.21 - Grafici relativi al sistema dell'esempio 3.48.

Risulta

$$[J(\mathbf{x})]^{-1} = \frac{1}{\det(J(\mathbf{x}))} \begin{bmatrix} 1 & \cos x_2 \\ -\sin x_1 & 2x_1 \end{bmatrix} > O.$$

Inoltre le matrici hessiane sono

$$S_1 = \begin{bmatrix} 2 & 0 \\ 0 & \sin x_2 \end{bmatrix}, \quad S_2 = \begin{bmatrix} \cos x_1 & 0 \\ 0 & 0 \end{bmatrix}$$

e risultano semidefinite positive in D. Perciò in D la funzione $\mathbf{f}(\mathbf{x})$ è totalmente differenziabile e convessa e verifica quindi le ipotesi del teorema 3.47. Scegliendo come $\mathbf{x}^{(0)}$ un qualunque punto di D per cui $\mathbf{f}(\mathbf{x}^{(0)}) > \mathbf{0}$, si ottiene una successione monotona convergente ad $\boldsymbol{\alpha}$. Nell'insieme indicato in grigio nella figura è $\mathbf{f}(\mathbf{x}) > \mathbf{0}$. Scegliendo $\mathbf{x}^{(0)} = (5.2, 0.8)$, si ottiene la successione convergente

i	$x_1^{(i)}$	$x_2^{(i)}$
1	5.041220	0.3282420
2	5.030712	0.3129920
3	5.030693	0.3129568
4	5.030694	0.3129578

∎

13. Condizionamento e localizzazione degli zeri di un polinomio

Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R}, \quad a_0 a_n \neq 0, \tag{67}$$

un polinomio di grado n a coefficienti reali. Per studiare il condizionamento del problema del calcolo degli zeri del polinomio (67), si esaminano gli effetti indotti sugli zeri da piccole perturbazioni introdotte sui coefficienti di (67). In tal modo si misura di quanto gli zeri siano sensibili agli errori da cui sono affetti i dati del problema.

Si considerano un polinomio

$$s(x) = \sum_{i=0}^{n} s_i x^i,$$

una perturbazione relativa ϵ e il polinomio perturbato

$$\widetilde{p}(x) = p(x) + \epsilon s(x) = \sum_{i=0}^{n} \widetilde{a}_i x^i. \tag{68}$$

Il seguente teorema stabilisce come variano gli zeri del polinomio perturbato rispetto a quelli di $p(x)$.

3.49 Teorema. *Sia α uno zero di $p(x)$. Allora*

1) *se α ha molteplicità 1, esiste un intorno U di 0 ed una funzione analitica $\alpha(\epsilon) : U \to \mathbf{C}$ tale che*
   $\alpha(\epsilon)$ è uno zero di molteplicità 1 di $\widetilde{p}(x)$,
   $$\alpha(0) = \alpha, \tag{69}$$
   $$\alpha'(0) = -\frac{s(\alpha)}{p'(\alpha)}, \tag{70}$$
   cioè, a meno di termini di ordine superiore in ϵ, è
   $$\alpha(\epsilon) \doteq \alpha - \epsilon \frac{s(\alpha)}{p'(\alpha)}. \tag{71}$$

2) *Se α ha molteplicità $m > 1$, esiste un intorno U di 0 ed una funzione analitica $h(t) : U \to \mathbf{C}$ tale che*
   $\alpha(\epsilon) = \alpha + h(\epsilon^{1/m})$ e $\alpha(\epsilon)$ è uno zero di $\widetilde{p}(x)$,
   $\alpha(0) = \alpha$,

$$h'(0) = -\left(\frac{m!s(\alpha)}{p^{(m)}(\alpha)}\right)^{1/m}, \tag{72}$$

cioè, a meno di termini di ordine superiore in ϵ*, è*

$$\alpha(\epsilon) \doteq \alpha - \epsilon^{1/m}\left(\frac{m!s(\alpha)}{p^{(m)}(\alpha)}\right)^{1/m}. \tag{73}$$

Dim. L'esistenza della funzione $\alpha(\epsilon)$ segue dal fatto che gli zeri di un polinomio sono funzioni continue dei coefficienti del polinomio [12]; inoltre dalla teoria delle funzioni di variabile complessa segue che la funzione $\alpha(\epsilon)$ risulta essere analitica, e quindi può essere rappresentata come una serie di potenze della forma

$$\alpha(\epsilon) = \alpha + \sum_{i=1}^{\infty} c_i \epsilon^i. \tag{74}$$

Dalla (74) segue la (69) per $\epsilon = 0$.

Poiché $\alpha(\epsilon)$ è uno zero di $\widetilde{p}(x)$, sostituendo nella (68) si ottiene

$$p(\alpha(\epsilon)) + \epsilon s(\alpha(\epsilon)) = 0, \tag{75}$$

e derivando rispetto a ϵ risulta

$$p'(\alpha(\epsilon))\alpha'(\epsilon) + s(\alpha(\epsilon)) + \epsilon s'(\alpha(\epsilon))\alpha'(\epsilon) = 0,$$

da cui

$$\alpha'(\epsilon) = -\frac{s(\alpha(\epsilon))}{p'(\alpha(\epsilon)) + \epsilon s'(\alpha(\epsilon))}.$$

Ponendo $\epsilon = 0$ si ottiene la (70).

Sostituendo la (70) nella (74), poiché $c_1 = \alpha'(0)$, risulta

$$\alpha(\epsilon) = \alpha - \epsilon\,\frac{s(\alpha)}{p'(\alpha)} + \sum_{i=2}^{\infty} c_i \epsilon^i,$$

da cui si ottiene la (71) con un'approssimazione al primo ordine.

La dimostrazione per il caso che α abbia molteplicità $m > 1$ è analoga a quella precedente: al posto della (75) si considera la relazione

$$p(\alpha + h(t)) + t^m s(\alpha + h(t)) = 0, \quad \text{dove} \quad t = \epsilon^{1/m},$$

e si deriva m volte rispetto a t; tenendo conto del fatto che

$$p(\alpha) = p'(\alpha) = \ldots = p^{(m-1)}(\alpha) = 0, \quad p^{(m)}(\alpha) \neq 0,$$

si ottiene la (72). ∎

Si consideri in particolare il caso in cui si introduca una perturbazione relativa ϵ su un solo coefficiente a_k del polinomio, cioè

$$s(x) = a_k x^k$$

per un indice k. Allora per (71) e (73) la variazione indotta su α risulta

$$|\alpha(\epsilon) - \alpha| \doteq |\epsilon| \left|\frac{a_k \alpha^k}{p'(\alpha)}\right|, \quad \text{se } \alpha \text{ ha molteplicità } 1, \tag{76}$$

$$|\alpha(\epsilon) - \alpha| \doteq |\epsilon|^{1/m} \left|\frac{m! a_k \alpha^k}{p^{(m)}(\alpha)}\right|^{1/m}, \quad \text{se } \alpha \text{ ha molteplicità } m > 1. \tag{77}$$

Quindi il problema del calcolo di uno zero α di molteplicità 1 del polinomio $p(x)$ è malcondizionato se il coefficiente a_k è elevato o se $p'(\alpha)$ è piccolo, ciò che accade ad esempio se $p(x)$ ha due zeri vicini ad un punto stazionario. Se α è uno zero di molteplicità $m > 1$, il problema è in generale malcondizionato.

3.50 Esempio. Il polinomio

$$p(x) = (x-1)^n$$

ha lo zero $\alpha = 1$ di molteplicità n. Se si perturba di $\epsilon > 0$ il termine noto, si ottiene il polinomio perturbato

$$\widetilde{p}(x) = (x-1)^n - \epsilon,$$

che ha gli n zeri (nel campo complesso)

$$\alpha(\epsilon) = 1 + \omega_n \sqrt[n]{\epsilon},$$

dove con ω_n si intende una radice n-esima dell'unità. Quindi la variazione indotta sugli zeri ha modulo $\sqrt[n]{\epsilon}$. Se ad esempio $n = 6$ e $\epsilon = 10^{-6}$, la variazione indotta ha modulo 10^{-1}. ∎

3.51 Esempio. Sia $\epsilon = 0.01$.

a) Il polinomio

$$p(x) = 24.5\, x^3 - 490.99\, x^2 + 19.81\, x - 0.2$$

ha i tre zeri $\alpha_1 = 20$, $\alpha_2 = \dfrac{1}{49} = 0.02040816$, $\alpha_3 = 0.02$, in corrispondenza ai quali risulta

$$p'(\alpha_1) = 9780.21, \quad p'(\alpha_2) = -0.1997959, \quad p'(\alpha_3) = 0.1998.$$

Se si altera di $-\epsilon$ il coefficiente di x^2 si ottiene il polinomio perturbato

$$\widetilde{p}(x) = 24.5\,x^3 - 491\,x^2 + 19.81\,x - 0.2,$$

i cui zeri sono

$$\widetilde{\alpha}_1 = 20.00041, \quad \widetilde{\alpha}_2 = 0.02038618, \quad \widetilde{\alpha}_3 = 0.02002116,$$

e quindi

$$\left|\frac{\widetilde{\alpha}_1 - \alpha_1}{\alpha_1}\right| \approx 0.204\ 10^{-4}, \ \left|\frac{\widetilde{\alpha}_2 - \alpha_2}{\alpha_2}\right| \approx 0.108\ 10^{-2}, \ \left|\frac{\widetilde{\alpha}_3 - \alpha_3}{\alpha_3}\right| \approx 0.106\ 10^{-2}.$$

Si noti come la variazione sia più elevata per gli zeri α_2 e α_3, in cui la derivata di $p(x)$ assume un valore più piccolo in modulo.

b) Il polinomio

$$p(x) = 0.2\,x^3 - 19.81\,x^2 + 490.99\,x - 24.5$$

è ottenuto dal precedente invertendo l'ordine dei coefficienti, quindi (si veda l'esercizio 3.50) ha i tre zeri $\alpha_1 = 50$, $\alpha_2 = 49$, $\alpha_3 = 0.05$, in corrispondenza ai quali risulta

$$p'(\alpha_1) = 9.99, \quad p'(\alpha_2) = -9.79, \quad p'(\alpha_3) = 489.0105.$$

Se si altera di ϵ il coefficiente di x si ottiene il polinomio perturbato

$$\widetilde{p}(x) = 0.2\,x^3 - 19.81\,x^2 + 491\,x - 24.5,$$

i cui zeri sono

$$\widetilde{\alpha}_1 = 49.94716, \quad \widetilde{\alpha}_2 = 49.05284, \quad \widetilde{\alpha}_3 = 0.04999898,$$

e quindi

$$\left|\frac{\widetilde{\alpha}_1 - \alpha_1}{\alpha_1}\right| \approx 0.106\ 10^{-2}, \ \left|\frac{\widetilde{\alpha}_2 - \alpha_2}{\alpha_2}\right| \approx 0.108\ 10^{-2}, \ \left|\frac{\widetilde{\alpha}_3 - \alpha_3}{\alpha_3}\right| \approx 0.204\ 10^{-4}.$$

Come nel caso a) l'elevata variazione degli zeri α_1 e α_2 è dovuta principalmente al fatto che la derivata di $p(x)$ assume un valore più piccolo in modulo.

c) Il polinomio

$$p(x) = 25\,x^3 - 501\,x^2 + 20.01\,x - 0.2$$

ha lo zero $\alpha_1 = 20$ di molteplicità 1 e lo zero $\alpha_2 = 0.02$ di molteplicità 2, e risulta

$$p'(\alpha_1) = 9980.01, \quad p''(\alpha_2) = -999.$$

Se si altera di ϵ il coefficiente di x si ottiene il polinomio perturbato

$$\widetilde{p}(x) = 25\,x^3 - 501\,x^2 + 20.02\,x - 0.2,$$

i cui zeri sono

$$\widetilde{\alpha}_1 = 19.99998, \quad \widetilde{\alpha}_2 = 0.02064287, \quad \widetilde{\alpha}_3 = 0.01937717,$$

e quindi

$$\left|\frac{\widetilde{\alpha}_1 - \alpha_1}{\alpha_1}\right| \approx 0.1\ 10^{-5}, \ \left|\frac{\widetilde{\alpha}_2 - \alpha_2}{\alpha_2}\right| \approx 0.321\ 10^{-1}, \ \left|\frac{\widetilde{\alpha}_3 - \alpha_2}{\alpha_2}\right| \approx 0.311\ 10^{-1}.$$

In questo caso l'elevata variazione da α_2 ad $\widetilde{\alpha}_2$ e $\widetilde{\alpha}_3$ è dovuta al fatto che α_2 ha molteplicità 2.

Se il coefficiente di x venisse alterato di $-\epsilon$, il polinomio $\widetilde{p}(x)$ avrebbe un solo zero reale.

d) Il polinomio

$$p(x) = 0.2\,x^3 - 20.01\,x^2 + 501\,x - 25$$

è ottenuto dal precedente invertendo l'ordine dei coefficienti, quindi ha lo zero $\alpha_1 = 50$ di molteplicità 2 e lo zero $\alpha_2 = 0.05$ di molteplicità 1, e risulta

$$p''(\alpha_1) = 19.98, \quad p'(\alpha_2) = 499.0005.$$

Se si altera di ϵ il coefficiente di x^2 si ottiene il polinomio perturbato

$$\widetilde{p}(x) = 0.2\,x^3 - 20.02\,x^2 + 501\,x - 25,$$

i cui zeri sono

$$\widetilde{\alpha}_1 = 51.60713, \quad \widetilde{\alpha}_2 = 48.44287, \quad \widetilde{\alpha}_3 = 0.05000005,$$

e quindi

$$\left|\frac{\widetilde{\alpha}_1-\alpha_1}{\alpha_1}\right| \approx 0.321\ 10^{-1},\ \left|\frac{\widetilde{\alpha}_2-\alpha_1}{\alpha_1}\right| \approx 0.311\ 10^{-1},\ \left|\frac{\widetilde{\alpha}_3-\alpha_2}{\alpha_2}\right| \approx 0.1\ 10^{-5}.$$

Anche in questo caso l'elevata variazione dei primi due zeri è dovuta al fatto che essi provengono da uno zero di molteplicità 2. ∎

Per la (76) e la (77) anche polinomi le cui radici sono ben separate possono risultare malcondizionati se il coefficiente a_k è di modulo elevato, come nel seguente classico esempio di Wilkinson [29].

3.52 Esempio. Sia

$$p(x) = \prod_{i=1}^{20}(x-i) = \sum_{i=0}^{20} a_i x^i.$$

Se si perturba il coefficiente a_k di $p(x)$ della quantità ϵ, per lo zero $\alpha_i = i$ dalla (76) si ha

$$\left|\frac{\alpha_i(\epsilon)-\alpha_i}{\alpha_i}\right| \doteq |\gamma(k,i)|\,|\epsilon|, \quad \gamma(k,i) = \frac{a_k i^{k-1}}{p'(i)}, \quad \begin{matrix} k=0,\ldots,20, \\ i=1,\ldots,20. \end{matrix}$$

I coefficienti a_i hanno modulo crescente al diminuire dell'indice fino a $i=2$ e il coefficiente di massimo modulo è $a_2 = 0.1380376\ 10^{20}$. Per quanto i valori di $|p'(i)|$ siano compresi fra 10^{12} e 10^{18}, alcuni valori di $|\gamma(k,i)|$ sono molto elevati. Per esempio per $i=14$ si ha $|\gamma(k,14)| > 10^{12}$ per $k=8,\ldots,16$, con un massimo di 10^{13} per $|\gamma(12,14)|$. ∎

In molti casi è utile poter dare una limitazione superiore del modulo degli zeri di un polinomio. Le limitazioni fornite dal seguente teorema hanno fra l'altro il pregio di richiedere poche operazioni.

3.53 Teorema. *Per gli zeri α_i, $i=1,\ldots,n$, del polinomio (67) valgono le seguenti limitazioni (per le limitazioni c) e d) si suppone anche $a_i \neq 0$ per $i=1,\ldots,n-1$):*

a) $$|\alpha_i| \le \max\left\{\left|\frac{a_0}{a_n}\right|,\ 1+\left|\frac{a_1}{a_n}\right|,\ \ldots,\ 1+\left|\frac{a_{n-1}}{a_n}\right|\right\},$$

b) $$|\alpha_i| \le \max\left\{1,\ \sum_{i=0}^{n-1}\left|\frac{a_i}{a_n}\right|\right\},$$

c) $$|\alpha_i| \le \max\left\{\left|\frac{a_0}{a_1}\right|,\ 2\left|\frac{a_1}{a_2}\right|,\ \ldots,\ 2\left|\frac{a_{n-1}}{a_n}\right|\right\},$$

d) $$|\alpha_i| \le \sum_{i=0}^{n-1}\left|\frac{a_i}{a_{i+1}}\right|,$$

e) $$|\alpha_i| \le 2\max\left\{\left|\frac{a_{n-1}}{a_n}\right|,\ \sqrt{\left|\frac{a_{n-2}}{a_n}\right|},\ \sqrt[3]{\left|\frac{a_{n-3}}{a_n}\right|},\ \ldots,\ \sqrt[n]{\left|\frac{a_0}{a_n}\right|}\right\}.$$

Dim. Il polinomio $(-1)^n \dfrac{p(\lambda)}{a_n}$ è il polinomio caratteristico della matrice (detta di *Frobenius*)

$$F = \begin{bmatrix} 0 & 0 & \dots & 0 & -\dfrac{a_0}{a_n} \\ 1 & 0 & \dots & 0 & -\dfrac{a_1}{a_n} \\ & \ddots & \ddots & \vdots & \vdots \\ & & 1 & 0 & -\dfrac{a_{n-2}}{a_n} \\ & & & 1 & -\dfrac{a_{n-1}}{a_n} \end{bmatrix}.$$

Le limitazioni a) e b) si ottengono applicando il primo teorema di Gerschgorin alle matrici F e F^T (si veda [3], p. 81).

Le limitazioni c) e d) si ottengono applicando il primo teorema di Gerschgorin alla matrice $D^{-1}FD$, dove

$$D = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix},$$

Risulta infatti

$$D^{-1}FD = \begin{bmatrix} 0 & 0 & \dots & 0 & -\dfrac{a_0}{a_1} \\ \dfrac{a_1}{a_2} & 0 & \dots & 0 & -\dfrac{a_1}{a_2} \\ & \ddots & \ddots & \vdots & \vdots \\ & & \dfrac{a_{n-2}}{a_{n-1}} & 0 & -\dfrac{a_{n-2}}{a_{n-1}} \\ & & & \dfrac{a_{n-1}}{a_n} & -\dfrac{a_{n-1}}{a_n} \end{bmatrix}.$$

Per ottenere la limitazione e) si ponga

$$\sigma = \max\left\{ \left|\frac{a_{n-1}}{a_n}\right|, \sqrt{\left|\frac{a_{n-2}}{a_n}\right|}, \sqrt[3]{\left|\frac{a_{n-3}}{a_n}\right|}, \dots, \sqrt[n]{\left|\frac{a_0}{a_n}\right|} \right\};$$

risulta quindi

$$\left|\frac{a_i}{a_n}\right| \leq \sigma^{n-i}.$$

Un qualsiasi zero α del polinomio (67) soddisfa alla relazione

$$\alpha^n = -\sum_{i=0}^{n-1} \frac{a_i}{a_n}\, \alpha^i,$$

e quindi, passando ai moduli, per $\sigma \neq 0$, si ha

$$|\alpha|^n \leq \sum_{i=0}^{n-1} \left|\frac{a_i}{a_n}\right| |\alpha|^i \leq \sum_{i=0}^{n-1} \sigma^{n-i}|\alpha|^i = \sigma^n \sum_{i=0}^{n-1} \frac{|\alpha|^i}{\sigma^i}\,. \tag{78}$$

Se tutti gli zeri α del polinomio (67) sono tali che $|\alpha| \leq \sigma$, la limitazione e) è verificata, altrimenti per gli zeri α per cui $|\alpha| > \sigma$, poiché dalla (78) risulta

$$\frac{|\alpha|^n}{\sigma^n} \leq \frac{|\alpha|^n/\sigma^n - 1}{|\alpha|/\sigma - 1}\,,$$

si ha

$$\Big(\frac{|\alpha|}{\sigma} - 1\Big)\frac{|\alpha|^n}{\sigma^n} \leq \frac{|\alpha|^n}{\sigma^n} - 1 < \frac{|\alpha|^n}{\sigma^n}\,,$$

e quindi

$$\frac{|\alpha|}{\sigma} < 2.$$

∎

Dal teorema 3.53 si possono ottenere anche delle limitazioni inferiori per i moduli degli zeri di un polinomio: infatti, essendo $a_0 \neq 0$, il polinomio

$$q(x) = \sum_{i=0}^{n} a_{n-i} x^i$$

ha per zeri i reciproci di quelli di $p(x)$ e quindi se per gli zeri $\dfrac{1}{\alpha_i}$ di $q(x)$ vale la limitazione $\dfrac{1}{|\alpha_i|} \leq \beta$, per $i = 1, \ldots, n$, per gli zeri α_i di $p(x)$ vale la limitazione $|\alpha_i| \geq \dfrac{1}{\beta}$, per $i = 1, \ldots, n$.

3.54 Esempio. Per gli zeri del polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096$$

dal teorema 3.53 si ottengono le seguenti limitazioni

$$\begin{array}{llllll} a) & |\alpha_i| \le 1281, & b) & |\alpha_i| \le 3639.077, & c) & |\alpha_i| \le 56, \\ d) & |\alpha_i| \le 42.20293, & e) & |\alpha_i| \le 56, & \text{per } i = 1, \dots, 6. \end{array}$$

Applicando il teorema 3.53 al polinomio che ha per zeri i reciproci degli α_i si ottengono le limitazioni

$$\begin{array}{llllll} a) & |\alpha_i| \ge 0.1975309, & b) & |\alpha_i| \ge 0.09475998, & c) & |\alpha_i| \ge 0.1538462, \\ d) & |\alpha_i| \ge 0.1801029, & e) & |\alpha_i| \ge 0.1538462, & \text{per } i = 1, \dots, 6. \end{array}$$

Quindi ogni zero α_i del polinomio verifica la limitazione

$$0.1975309 \le |\alpha_i| \le 42.20293 \quad \text{per} \quad i = 1, \dots, 6.$$

Nell'esempio 3.62 si verificherà che gli zeri di $p(x)$ appartengono in effetti all'intervallo $[1, 18]$. ∎

Sfruttando la limitazione e) del teorema 3.53 si può ottenere un teorema di perturbazione in cui la maggiorazione della variazione di ogni singolo zero è indipendente dal particolare zero considerato.

3.55 Teorema *(di Ostrowski). Siano $p(x)$ e $\widetilde{p}(x)$ i polinomi definiti in (67) e (68), e sia $s_n \neq 0$. Allora, posto*

$$\gamma = 2 \max \left\{ \sqrt[j]{\left| \frac{a_{n-j}}{a_n} \right|}, \sqrt[j]{\left| \frac{\widetilde{a}_{n-j}}{\widetilde{a}_n} \right|}, \quad j = 0, \dots, n \right\},$$

esiste un ordinamento degli zeri α_i di $p(x)$ e un ordinamento degli zeri $\alpha_i(\epsilon)$ di $\widetilde{p}(x)$, per cui

$$|\alpha_i - \alpha_i(\epsilon)| < 2n \sqrt[n]{\epsilon \sum_{j=1}^{n} \left| \frac{s_j}{s_n} \right| \gamma^j}, \quad i = 1, \dots, n.$$

Per la dimostrazione si veda [19]. ∎

14. Successione di Sturm

Il numero delle radici reali di un polinomio e la determinazione dei relativi intervalli di separazione possono essere ottenuti per mezzo delle successioni di Sturm.

3.56 Definizione. Una successione di polinomi $p_i(x),\ i = 0, 1, \ldots, m$, che verifica le seguenti proprietà:

1) $p_m(x)$ non cambia segno,

2) se in un punto ξ è $p_i(\xi) = 0$, allora $p_{i-1}(\xi)p_{i+1}(\xi) < 0$, per $i = 1, 2, \ldots, m-1$,

3) se in un punto ξ è $p_0(\xi) = 0$, allora $p_0'(\xi)p_1(\xi) < 0$ (e quindi $p_0(x)$ ha tutti zeri di molteplicità 1),

è detta *successione di Sturm* del polinomio $p_0(x)$. ∎

Una successione di Sturm di un polinomio $p_0(x) = p(x)$ può essere costruita con il procedimento analogo a quello di Euclide delle divisioni successive per la determinazione del massimo comun divisore di $p(x)$ e $p'(x)$: posto

$$r_0(x) = p(x), \quad r_1(x) = -p'(x),$$

si calcolano i quozienti $q_i(x)$ e i resti cambiati di segno $r_i(x)$, delle divisioni di $r_{i-2}(x)$ per $r_{i-1}(x)$, cioè

$$r_{i-2}(x) = q_i(x)r_{i-1}(x) - r_i(x), \quad \text{per} \quad i = 2, \ldots, m+1,$$

finché

$$r_{m+1}(x) = 0.$$

I polinomi $r_i(x)$ hanno grado decrescente e l'ultimo polinomio costruito $r_m(x)$ è, a meno del segno, il massimo comun divisore di $p(x)$ e $p'(x)$. Se $r_m(x)$ ha grado zero, i polinomi $p(x)$ e $p'(x)$ non hanno zeri in comune e quindi $p(x)$ ha tutti zeri di molteplicità 1. In tal caso la successione di Sturm di $p(x)$ è

$$p_i(x) = r_i(x), \quad \text{per} \quad i = 0, \ldots, m.$$

Se $r_m(x)$ ha grado maggiore di zero, esso divide tutti i polinomi $r_i(x)$, $i = 0, \ldots, m$, ed ha per zeri gli zeri di molteplicità maggiore di 1 di $p(x)$. In tal caso la successione di Sturm di $p(x)$ è

$$p_i(x) = \frac{r_i(x)}{r_m(x)}, \quad \text{per} \quad i = 0, \ldots, m.$$

È facile verificare che la successione così costruita soddisfa le tre proprietà della definizione 3.56.

3.57 Esempio. a) La successione di Sturm, costruita con il procedimento sopra descritto, del polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096$$

(già considerato nell'esempio 3.54) è data da

$$\begin{aligned}
p_0(x) &= 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096,\\
p_1(x) &= -26\,(3\,x^5 - 70\,x^4 + 448\,x^3 - 1152\,x^2 + 1280\,x - 512),\\
p_2(x) &= \frac{4}{9}\,(1001\,x^4 - 9152\,x^3 + 27456\,x^2 - 33280\,x + 14080),\\
p_3(x) &= -\,\frac{2304}{539}\,(143\,x^3 - 702\,x^2 + 1080\,x - 528),\\
p_4(x) &= \frac{784}{99}\,(51\,x^2 - 136\,x + 88),\\
p_5(x) &= -\,\frac{4096}{833}\,(19\,x - 22),\\
p_6(x) &= \frac{3136}{361}\,.
\end{aligned}$$

b) Per il polinomio

$$p(x) = x^5 - 4\,x^4 + 4\,x^3 - 17\,x^2 + 10\,x - 24$$

si ha

$$\begin{aligned}
p_0(x) &= x^5 - 4\,x^4 + 4\,x^3 - 17\,x^2 + 10\,x - 24,\\
p_1(x) &= -5\,x^4 + 16\,x^3 - 12\,x^2 + 34\,x - 10,\\
p_2(x) &= \frac{1}{25}\,(24\,x^3 + 207\,x^2 - 64\,x + 560),\\
p_3(x) &= \frac{25}{192}\,(4111\,x^2 - 2368\,x + 10672),\\
p_4(x) &= -\,\frac{9856}{422508025}\,(1535\,x - 22723),\\
p_5(x) &= -\,\frac{688400775293}{6031936}\,.
\end{aligned}$$

∎

Si consideri, in un punto ξ, la successione $p_0(\xi), p_1(\xi), \ldots, p_m(\xi)$ (se fosse $p_i(\xi) = 0$ per un indice $i < m$, si attribuisca a tale valore il segno di $p_{i+1}(\xi)$) e si indichi con $w(\xi)$ il numero di *cambiamenti di segno* di tale successione. Vale il seguente teorema.

3.58 Teorema. *Se $p_i(x)$, $i = 0, 1, \ldots, m$, è una successione di Sturm di $p_0(x)$, il numero $w(b)-w(a)$ è uguale al numero di zeri di $p_0(x)$ appartenenti all'intervallo $[a, b)$.*

Dim. Si faccia variare x con continuità da a verso b. Si può avere una variazione nel numero $w(x)$ solo quando x incontra uno zero di uno dei polinomi $p_i(x)$. Si consideri perciò un valore ξ tale che $p_i(\xi) = 0$ per un indice i. Per la proprietà 1) della definizione 3.56 deve essere $i < m$. Si distinguono allora i due casi:

a) $i > 0$. In questo caso, per la proprietà 2) della definizione 3.56 si ha

$$p_{i-1}(\xi)p_{i+1}(\xi) < 0.$$

Esiste perciò un numero h tale che nell'intervallo $[\xi - h, \xi + h]$ è ancora

$$p_{i-1}(x)p_{i+1}(x) < 0$$

e

$$p_i(x) \neq 0,$$

eccetto che nel punto ξ. Poiché per ogni $x \in [\xi - h, \xi + h]$ i due polinomi $p_{i-1}(x)$ e $p_{i+1}(x)$ hanno segno discorde, $p_i(x)$ deve avere in questo intervallo segno concorde con uno dei due e discorde con l'altro. Quindi nella sequenza $p_{i-1}(x), p_i(x), p_{i+1}(x)$ vi è una sola variazione di segno in tutto l'intervallo $[\xi - h, \xi + h]$. Cioè il fatto che $p_i(x)$ si annulli in ξ non comporta variazioni del numero $w(x)$ nell'intervallo $[\xi - h, \xi + h]$.

b) $i = 0$. In questo caso, poiché per la proprietà 3) della definizione 3.56 il polinomio $p_0(x)$ ha radici semplici, la sua derivata $p_0'(x)$ non si annulla in ξ ed esiste un numero h tale che nell'intervallo $[\xi - h, \xi + h]$ $p_0'(x)$ ha lo stesso segno che $p_0(x)$ ha in $\xi + h$ e segno opposto a quello che $p_0(x)$ ha in $\xi - h$. Se h è tale che nell'intervallo $[\xi - h, \xi + h]$ anche $p_1(x)$ non si annulla, poiché per la proprietà 3) della definizione 3.56 $p_1(x)$ ha segno opposto a quello di $p_0'(x)$ per $x \in [\xi - h, \xi + h]$, la sequenza $p_0(\xi+h)$, $p_1(\xi+h)$ presenta una variazione di segno, mentre la sequenza $p_0(\xi - h)$, $p_1(\xi - h)$ non presenta alcuna variazione di segno. Cioè il fatto che $p_0(x)$ si annulli in ξ comporta una variazione del numero $w(x)$ nell'intervallo $[\xi - h, \xi + h]$.

Se ne conclude che il numero di variazioni di segno in tutta la sequenza $p_0(x), p_1(x), \ldots, p_m(x)$ può cambiare solo nei punti in cui si annulla $p_0(x)$, ed aumenta di 1 ogni volta che si annulla $p_0(x)$.

Nella tesi del teorema l'intervallo $[a, b)$ è aperto a destra perché se fosse $p_0(b) = 0$, poiché a $p_0(b)$ viene assegnato lo stesso segno assunto in b da

$p_1(x)$, che è diverso da zero in un intorno sinistro di b, $w(x)$ non cambia in tale intorno. Perciò la radice b non altera il numero di variazioni di segno.

∎

Gli estremi a e b dell'intervallo a cui si riferisce il teorema 3.58 possono essere anche $-\infty$ e $+\infty$: in tal caso si assume come segno dell'i-esimo polinomio quello del corrispondente limite. È così possibile determinare il numero delle radici reali di un polinomio.

Utilizzando il teorema 3.58 si può realizzare un procedimento di bisezione per calcolare il k-esimo zero α_k di un polinomio $p(x)$ che abbia tutti zeri reali e distinti.

1) Sia $[a_0, b_0)$ tale che $w(b_0) - w(a_0) = n$,

2) per $i = 1, 2, \ldots$, sia $c_i = \dfrac{1}{2}(a_{i-1} + b_{i-1})$,
se $p(c_i) = 0$ e $w(c_i) - w(a_0) = k - 1$, allora stop,
altrimenti se $w(c_i) - w(a_0) \geq k$, allora $a_i = a_{i-1}$ e $b_i = c_i$,
altrimenti $a_i = c_i$ e $b_i = b_{i-1}$.

Poiché questo questo procedimento converge molto lentamente, in generale conviene utilizzarlo per separare gli zeri e non per approssimarli.

3.59 Esempio. Nel caso della successione di Sturm del polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096,$$

dall'esempio 3.57 risulta

x	$p_0(x)$	$p_1(x)$	$p_2(x)$	$p_3(x)$	$p_4(x)$	$p_5(x)$	$p_6(x)$	$w(x)$
0	+	+	+	+	+	+	+	0
1	+	+	+	+	+	+	+	0
2	−	0	+	+	+	−	+	3
3	+	−	−	−	+	−	+	4
5	−	+	+	−	+	−	+	5
20	+	−	+	−	+	−	+	6

Dall'ultima colonna risulta che tutti gli zeri sono reali e positivi, che ve ne sono tre nell'intervallo (1,2), uno nell'intervallo (2,3), uno nell'intervallo (3,5), uno nell'intervallo (5,20). Poiché inoltre $w(1.1) = 1$ e $w(1.2) = 2$, risulta

$$1 < \alpha_6 < 1.1 < \alpha_5 < 1.2 < \alpha_4 < 2 < \alpha_3 < 3 < \alpha_2 < 5 < \alpha_1 < 20.$$

Per ridurre l'intervallo di separazione di α_2, si può applicare l'algoritmo di bisezione all'intervallo [3,5], ottenendo per c_i e $w(c_i)$ successivamente i valori

c_i	$w(c_i)$
4	4
4.5	4
4.75	5
4.625	4
4.6875	5

da cui si ha che

$$4.625 < \alpha_2 < 4.6875.$$

Per l'altro polinomio di cui si è calcolata la successione di Sturm nell'esempio 3.57

$$p(x) = x^5 - 4\,x^4 + 4\,x^3 - 17\,x^2 + 10\,x - 24$$

si ha

x	$p_0(x)$	$p_1(x)$	$p_2(x)$	$p_3(x)$	$p_4(x)$	$p_5(x)$	$w(x)$
$-\infty$	$-$	$-$	$-$	$+$	$+$	$-$	2
0	$-$	$-$	$+$	$+$	$+$	$-$	2
$+\infty$	$+$	$-$	$+$	$+$	$-$	$-$	3

Ne segue che il polinomio ha una sola radice reale, che risulta essere positiva.

∎

15. Metodo di Newton per il calcolo degli zeri di un polinomio

Il metodo di Newton è particolarmente vantaggioso per il calcolo degli zeri di un polinomio: infatti il calcolo di $p(\xi)$ e $p'(\xi)$ in un punto ξ può essere eseguito in modo efficiente per mezzo del procedimento di divisione sintetica di Ruffini, nel modo seguente. Siano $q(x)$ e r quoziente e resto della divisione di $p(x)$ per $x-\xi$ e $q_1(x)$ e r_1 quoziente e resto della divisione di $q(x)$ per $x-\xi$ (si veda l'esercizio 3.47):

$$\begin{aligned} p(x) &= q(x)(x-\xi) + r \\ q(x) &= q_1(x)(x-\xi) + r_1, \end{aligned}$$

allora è

$$p(\xi) = r \quad \text{e} \quad p'(\xi) = q(\xi) = r_1.$$

In modo ricorrente r e r_1 si calcolano mediante le relazioni

$$\begin{aligned}
&b_n = a_n, \quad c_n = b_n,\\
&b_i = b_{i+1}\xi + a_i, \quad c_i = c_{i+1}\xi + b_i, \quad i = n-1, \ldots, 1,\\
&p'(\xi) = c_1, \quad p(\xi) = b_1\xi + a_0.
\end{aligned}$$

Se si sceglie un opportuno punto x_0 complesso e si opera in aritmetica complessa, è possibile con il metodo di Newton approssimare anche gli zeri complessi di un polinomio.

3.60 Esempio. L'equazione

$$x^3 - 1 = 0$$

ha una radice reale e due radici complesse coniugate. Scegliendo come punto iniziale $x_0 = \boldsymbol{i}$ e operando in aritmetica complessa si ottiene la successione di numeri complessi

i	x_i
1	$-0.3333333 + \boldsymbol{i}\, 0.6666667$
2	$-0.5822216 + \boldsymbol{i}\, 0.9244434$
3	$-0.5087910 + \boldsymbol{i}\, 0.8681659$
4	$-0.5000684 + \boldsymbol{i}\, 0.8659816$
5	$-0.4999998 + \boldsymbol{i}\, 0.8660252$
6	$-0.4999999 + \boldsymbol{i}\, 0.8660254$

∎

Nel caso dell'esempio 3.60 si può dimostrare che per quasi ogni punto iniziale del piano complesso la successione generata dal metodo di Newton è convergente. Questo però non è vero in generale: esistono casi in cui l'insieme dei punti del piano complesso, a partire dai quali la successione ottenuta con il metodo di Newton non converge, è un insieme di misura non nulla; un caso semlice è quello del seguente esempio.

3.61 Esempio (Smale [26]). Si consideri il polinomio

$$p(x) = \frac{1}{2}\, x^3 - x + 1.$$

Il metodo di Newton è definito dalla funzione di iterazione

$$g(x) = \frac{2x^3 - 2}{3x^2 - 2}\,.$$

Poiché $g(0) = 1,\ g(1) = 0,\ g'(0) = 0$, la funzione $\phi(x) = g(g(x))$ è tale che

$$\phi(0) = 0, \quad \phi'(0) = 0,$$

cioè 0 è punto fisso di $\phi(x)$. Per il teorema 3.16 la successione generata dalla ricorrenza $z_{i+1} = \phi(z_i)$, a partire da un punto z_0 appartenente ad un opportuno intorno U dello 0, converge a 0 con ordine almeno due. Quindi la successione $x_{i+1} = g(x_i)$ ottenuta a partire da $x_0 \in U$ è tale che $\lim_{i\to\infty} x_{2i} = 0$, mentre $p(0) \neq 0$. ∎

Nel caso però di polinomi con zeri tutti reali è stato dimostrato [2] che sulla retta reale l'insieme dei punti a partire dai quali il metodo di Newton non converge è un insieme di misura nulla.

Se il polinomio (67) ha zeri $\alpha_1 > \alpha_2 > \ldots > \alpha_n$ tutti reali, scegliendo come punto iniziale un punto $x_0 > \alpha_1$, per il teorema 3.25 la successione ottenuta con il metodo delle tangenti è monotona decrescente. È possibile accelerare la convergenza utilizzando la tecnica del *passo doppio*:

$$x_{i+1} = x_i - 2\,\frac{p(x_i)}{p'(x_i)}\,, \tag{79}$$

fintanto che

$$p(x_{i+1})p(x_0) > 0. \tag{80}$$

È possibile dimostrare (si veda l'esercizio 3.62) che con il metodo a passo doppio la condizione (80) implica che $x_{i+1} > \alpha_2$, cioè che non è possibile con questo metodo oltrepassare più di una radice: al primo indice i per cui la (80) non è verificata si prosegue con il metodo di Newton a passo semplice, a partire dal punto x_i perché il punto x_{i+1} calcolato con la (79) ha oltrepassato lo zero α_1.

In teoria, una volta calcolato uno zero del polinomio, ad esempio α_1, sarebbe possibile ridurre il grado del polinomio, dividendolo per $x - \alpha_1$, e procedere al calcolo dello zero successivo. In pratica però in questo processo, detto di *deflazione*, può essere consistente l'influenza dell'errore con cui è stato approssimato α_1 e degli errori generati durante la divisione eseguita con un'aritmetica finita: i coefficienti del polinomio quoziente effettivamente calcolato possono essere sensibilmente diversi da quelli teorici e, conseguentemente, gli zeri di tale polinomio possono essere molto diversi da quelli del polinomio $p(x)$. È addirittura possibile che il polinomio quoziente effettivamente calcolato abbia un diverso numero di zeri reali.

Se si procede approssimando gli zeri in ordine di modulo crescente, la situazione può non essere così critica: Wilkinson ha infatti dimostrato [30] che in questo caso, se gli zeri vengono calcolati con la massima precisione possibile compatibilmente con la precisione di macchina, allora gli

errori generati dalla deflazione sono trascurabili. Il processo di deflazione può essere applicato al polinomio $p\left(\frac{1}{x}\right)$, ottenuto invertendo l'ordine dei coefficienti, se gli zeri vengono calcolati in ordine di modulo decrescente.

3.62 Esempio. Il polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096$$

del quale sono stati separati gli zeri utilizzando la successione di Sturm nell'esempio 3.59, ha gli zeri

$$\alpha_{n-i} = \frac{1}{\cos^2 \dfrac{(2i+1)\pi}{26}}\,, \quad i = 0, \ldots, 5.$$

Approssimando gli zeri in ordine decrescente con il metodo di Newton a passo doppio e con la deflazione, per gli errori relativi ϵ_i da cui sono affetti i risultati $\widetilde{\alpha}_i$, risulta

i	$\widetilde{\alpha}_i$	ϵ_i
1	17.46054	$0.1328950\ 10^{-6}$
2	4.631017	$0.1496924\ 10^{-3}$
3	2.252462	$0.9520976\ 10^{-2}$
4	1.577625	$0.6852945\ 10^{-1}$
5	1.283130	0.1217839
6	1.076390	$0.6075129\ 10^{-1}$

Invertendo l'ordine dei coefficienti e approssimando quindi i reciproci degli zeri in ordine crescente, i corrispondenti errori relativi risultano

i	$\widetilde{\alpha}_i$	ϵ_i
1	17.46054	$0.1328950\ 10^{-6}$
2	4.630325	$0.3693404\ 10^{-6}$
3	2.274032	$0.3629274\ 10^{-4}$
4	1.476577	$0.8907815\ 10^{-4}$
5	1.143682	$0.1302795\ 10^{-3}$
6	1.014814	$0.7004611\ 10^{-4}$

■

Il procedimento di deflazione può essere realizzato implicitamente, evitando gli inconvenienti connessi con l'ordinamento degli zeri, con *il metodo*

di Maehly che si basa sulla seguente osservazione: una volta calcolati, secondo un ordinamento qualsiasi, k zeri di $p(x)$, che saranno indicati con $\alpha_1, \ldots, \alpha_k$, il polinomio

$$q(x) = \frac{p(x)}{\prod_{j=1}^{k} (x - \alpha_j)}$$

ha come zeri i rimanenti zeri $\alpha_{k+1}, \ldots, \alpha_n$ di $p(x)$, e risulta

$$q'(x) = \frac{1}{\prod_{j=1}^{k} (x - \alpha_j)} \left[p'(x) - p(x) \sum_{j=1}^{k} \frac{1}{x - \alpha_j} \right].$$

Il metodo di Newton può essere applicato al polinomio $q(x)$, senza doverne calcolare effettivamente i coefficienti, mediante la relazione

$$x_{i+1} = x_i - \frac{q(x_i)}{q'(x_i)} = x_i - \frac{1}{\dfrac{p'(x_i)}{p(x_i)} - \displaystyle\sum_{j=1}^{k} \dfrac{1}{x_i - \alpha_j}}. \tag{81}$$

Il metodo di Maehly converge localmente agli zeri di $q(x)$ ed ha ordine di convergenza 2 per gli zeri di molteplicità 1.

In pratica, anche se i valori $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_k$ effettivamente calcolati sono delle approssimazioni degli zeri $\alpha_1, \ldots, \alpha_k$ di $p(x)$, la funzione razionale

$$\widetilde{q}(x) = \frac{p(x)}{\prod_{j=1}^{k} (x - \widetilde{\alpha}_j)}$$

ha ancora come zeri gli zeri $\alpha_1, \ldots, \alpha_n$ di $p(x)$. Quindi la precisione con cui vengono calcolati gli zeri successivi non dipende dalla precisione con cui sono stati calcolati $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_k$.

Con questo metodo esiste comunque il rischio di approssimare uno degli zeri già calcolati, se la precisione con cui si opera non è sufficiente, specie nel caso di zeri molto vicini fra loro.

Poiché gli zeri $\alpha_1, \ldots, \alpha_k$ già calcolati non possono essere assunti come punti iniziali per calcolare α_{k+1}, si può usare, quando si procede al calcolo degli zeri da destra verso sinistra, una buona strategia per la scelta del punto iniziale:

a) per calcolare il primo zero α_1 si utilizza il metodo di Newton scegliendo $x_0 > \alpha_1$ e procedendo con passo doppio fino a quando non si ottiene un

punto $\beta = x_{i+1}$ che scavalca α_1 e successivamente si procede a passo semplice a partire da x_i;

b) calcolato α_1, si sceglie $x_0 = \beta$ per approssimare α_2 con la (81) prima a passo doppio e poi a passo semplice, e così via.

3.63 Esempio. Del polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096$$

studiato nell'esempio precedente si calcolano gli zeri con il metodo di Maehly, in ordine di modulo decrescente. Nella seguente tabella sono riportate le approssimazioni calcolate e i corrispondenti errori relativi.

i	$\widetilde{\alpha}_i$	ϵ_i
1	17.46054	$0.1328950\ 10^{-6}$
2	4.630327	$0.7812660\ 10^{-6}$
3	2.274094	$0.9034285\ 10^{-5}$
4	1.476436	$0.6518889\ 10^{-5}$
5	1.143648	$0.1594609\ 10^{-3}$
6	1.014803	$0.5970810\ 10^{-4}$

■

16. Metodo di Bairstow

È possibile evitare l'utilizzazione di un'aritmetica complessa per il calcolo degli zeri complessi di polinomi a coefficienti reali, sviluppando metodi, come il metodo di *Bairstow*, che consentono, mediante l'uso dell'aritmetica reale, di calcolare simultaneamente uno zero complesso α e il suo coniugato $\overline{\alpha}$, tramite il calcolo del fattore quadratico

$$x^2 + (\alpha + \overline{\alpha})x + \alpha\overline{\alpha}.$$

Se $p(x)$è diviso per un generico polinomio di secondo grado $x^2 + bx + c$ si ha

$$p(x) = (x^2 + bx + c)q(x) + rx + s, \tag{82}$$

dove $r = r(b, c)$ e $s = s(b, c)$. Il polinomio $x^2 + bx + c$ risulta quindi fattore di $p(x)$ se $r = 0$ e $s = 0$. Il metodo di Bairstow consiste nella risoluzione del sistema

$$\begin{cases} r(b, c) = 0 \\ \\ s(b, c) = 0, \end{cases}$$

applicando il metodo di Newton-Raphson, la cui i-esima iterazione comporta la risoluzione del sistema lineare

$$\begin{bmatrix} \dfrac{\partial r(b_i,c_i)}{\partial b} & \dfrac{\partial r(b_i,c_i)}{\partial c} \\ \dfrac{\partial s(b_i,c_i)}{\partial b} & \dfrac{\partial s(b_i,c_i)}{\partial c} \end{bmatrix} \begin{bmatrix} b_{i+1}-b_i \\ c_{i+1}-c_i \end{bmatrix} = - \begin{bmatrix} r(b_i,c_i) \\ s(b_i,c_i) \end{bmatrix}. \tag{83}$$

Le derivate parziali dello jacobiano possono essere rappresentate in funzione di b, c e dei coefficienti a_i, $i = 0, \ldots, n$. Infatti dividendo il polinomio $q(x)$ per $x^2 + bx + c$ si ha

$$q(x) = (x^2 + bx + c)v(x) + tx + u, \tag{84}$$

e sostituendo nella (82) si ha

$$p(x) = (x^2 + bx + c)^2 v(x) + (x^2 + bx + c)(tx + u) + rx + s.$$

Derivando rispetto a b e a c e calcolando le derivate parziali in un punto z tale che

$$z^2 + bz + c = 0,$$

poiché $p(x)$ non dipende da b e da c si ha

$$\begin{cases} 0 = \dfrac{\partial p(z)}{\partial b} = z(tz+u) + z\,\dfrac{\partial r}{\partial b} + \dfrac{\partial s}{\partial b} \\ 0 = \dfrac{\partial p(z)}{\partial c} = tz + u + z\,\dfrac{\partial r}{\partial c} + \dfrac{\partial s}{\partial c}\,. \end{cases}$$

Inoltre, essendo $z^2 = -bz - c$, si ottiene il sistema

$$\begin{cases} z\Big(\dfrac{\partial r}{\partial b} - bt + u\Big) + \Big(\dfrac{\partial s}{\partial b} - ct\Big) = 0 \\ z\Big(\dfrac{\partial r}{\partial c} + t\Big) + \Big(\dfrac{\partial s}{\partial c} + u\Big) = 0\,. \end{cases} \tag{85}$$

Nell'ipotesi che il polinomio $x^2 + bx + c$ non sia un quadrato perfetto, le equazioni (85) sono soddisfatte per due valori distinti di z e quindi deve risultare

$$\begin{aligned} \frac{\partial r}{\partial b} &= bt - u, & \frac{\partial r}{\partial c} &= -t, \\ \frac{\partial s}{\partial b} &= ct, & \frac{\partial s}{\partial c} &= -u. \end{aligned} \tag{86}$$

I valori delle derivate di r e di s rispetto a b e a c sono ricavati per mezzo dei coefficienti dei resti delle due divisioni (82) e (84) e possono essere ottenuti mediante le formule ricorrenti (si veda l'esercizio 3.45)

$$\begin{aligned} &q_n = q_{n-1} = 0, \\ &q_i = a_{i+2} - bq_{i+1} - cq_{i+2}, \quad i = n-2, \dots, -2, \\ &r = q_{-1}, \quad s = q_{-2} + br, \\ &v_{n-2} = v_{n-3} = 0, \\ &v_i = q_{i+2} - bv_{i+1} - cv_{i+2}, \quad i = n-4, \dots, -2, \\ &t = v_{-1}, \quad u = v_{-2} + bt. \end{aligned} \tag{87}$$

Il metodo di Bairstow è quindi un metodo iterativo: fissati per b e c i valori iniziali b_0 e c_0, si calcolano r, s, t, u mediante le (87), si calcolano le derivate mediante le (86) e si risolve il sistema (83) per ottenere la successiva approssimazione b_{i+1} e c_{i+1}.

Come condizione di arresto si può controllare il resto $rx + s$ della divisione (82), arrestando il procedimento quando r ed s sono sufficientemente piccoli in modulo.

3.64 Esempio. Si applica il procedimento di Bairstow al polinomio

$$p(x) = x^5 - 4\,x^4 + 4\,x^3 - 17\,x^2 + 10\,x - 24,$$

che, come risulta dall'esempio 3.59, ha un solo zero reale. Assumendo come valori iniziali $b_0 = 1$ e $c_0 = 1$, si ottiene la successione

i	b_i	c_i
1	0.5319927	1.720292
2	0.4989448	3.124874
3	0.9052511	2.514734
4	0.9544439	3.014984
5	0.9998856	2.995825
6	0.9999961	2.999997
7	0.9999995	3.000000

Quindi il polinomio $p(x)$ ha il fattore quadratico

$$x^2 + 0.9999995\,x + 3,$$

da cui si possono ricavare le approssimazioni dei due zeri complessi $-\dfrac{1}{2} \pm \boldsymbol{i}\,\dfrac{\sqrt{11}}{2}$ di $p(x)$. Assumendo invece come valori iniziali $b_0 = 1$ e $c_0 = -1$,

si ottiene la successione

i	b_i	c_i
1	0.4312270	0.2193222
2	0.05509335	1.400377
3	−0.6232033	3.059950
4	−0.6130959	1.547807
5	−1.058820	2.125072
6	−1.001415	2.003648
7	−1.000000	2.000003

Quindi il polinomio $p(x)$ ha il fattore quadratico

$$x^2 - x + 2.000003,$$

da cui si possono ricavare le approssimazioni degli altri due zeri complessi $\frac{1}{2} \pm \boldsymbol{i}\,\frac{\sqrt{7}}{2}$ di $p(x)$.

3.65 Esempio. Se si applica il procedimento di Bairstow al polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096,$$

che, come risulta dall'esempio 3.59, non ha zeri complessi coniugati, si ottengono dei fattori quadratici, da cui si possono ricavare delle approssimazioni di zeri reali. Assumendo come valori iniziali $b_0 = -2$ e $c_0 = 2$, si ottiene la successione

i	b_i	c_i
1	−2.393842	1.838592
2	−2.593541	1.809242
3	−2.654349	1.774718
4	−2.640789	1.722891
5	−2.623610	1.693897
6	−2.620420	1.689018
7	−2.620276	1.688779

Quindi il polinomio $p(x)$ ha il fattore quadratico

$$x^2 - 2.620276\,x + 1.688779,$$

da cui si possono ricavare le approssimazioni

$$1.476519, \quad 1.143757$$

degli zeri α_4 e α_5. Assumendo invece come valori iniziali $b_0 = -20$ e $c_0 = 10$, si ottiene la successione

i	b_i	c_i
1	-18.63109	9.301706
2	-18.07014	9.030700
3	-17.98471	9.112988
.	...	...
.	...	...
11	-18.47520	17.71652
12	-18.47523	17.71707
13	-18.47520	17.71654

Quindi il polinomio $p(x)$ ha il fattore quadratico

$$x^2 - 18.47520\,x + 17.71654,$$

da cui si possono ricavare le approssimazioni

$$17.46054, \quad 1.014662$$

degli zeri α_1 e α_6. ∎

17. Metodo di Bernoulli e metodo qd

È possibile associare al polinomio (67)

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_0 a_n \neq 0,$$

l'equazione alle differenze (si veda il capitolo 4)

$$y_k = -\frac{1}{a_n}\,(a_{n-1}y_{k-1} + a_{n-2}y_{k-2} + \ldots + a_0 y_{k-n}), \quad k \geq n, \qquad (88),$$

che, assegnata una n-upla di valori iniziali

$$y_0,\ y_1, \ldots, y_{n-1},$$

permette di costruire univocamente la successione $\{y_k\}$.

Si suppone che il polinomio abbia n zeri $\alpha_1, \alpha_2, \ldots, \alpha_n$ reali e distinti, e che $|\alpha_1| > |\alpha_2| \geq \ldots \geq |\alpha_n|$; allora la soluzione generale dell'equazione (88) (si veda il teorema 4.13) è della forma

$$y_k = \sum_{i=1}^{n} c_i \alpha_i^k, \quad c_i \in \mathbf{R}, \quad \text{per} \quad i = 1, \ldots, n, \tag{89}$$

e si suppone che $y_k \neq 0$ per ogni k. Se $c_1 \neq 0$, si ha

$$\frac{y_{k+1}}{y_k} = \frac{\sum\limits_{i=1}^{n} c_i \alpha_i^{k+1}}{\sum\limits_{i=1}^{n} c_i \alpha_i^k} = \alpha_1 \frac{c_1 + \sum\limits_{i=2}^{n} c_i \Big(\dfrac{\alpha_i}{\alpha_1}\Big)^{k+1}}{c_1 + \sum\limits_{i=2}^{n} c_i \Big(\dfrac{\alpha_i}{\alpha_1}\Big)^{k}},$$

e al tendere di k all'infinito, poiché $\left|\dfrac{\alpha_i}{\alpha_1}\right| < 1$ per $i = 2, \ldots, n$, risulta

$$\lim_{k \to \infty} \frac{y_{k+1}}{y_k} = \alpha_1$$

e

$$\left| \frac{y_{k+1}}{y_k} - \alpha_1 \right| = O\Big(\Big| \frac{\alpha_2}{\alpha_1} \Big|^k \Big).$$

Scelti in modo arbitrario gli n valori iniziali $y_0, y_1, \ldots, y_{n-1}$, il metodo di Bernoulli consiste nel calcolare la successione degli y_k con la (88) e i rapporti y_{k+1}/y_k. Il metodo è arrestato quando i rapporti y_{k+1}/y_k si stabilizzano su un valore, che viene assunto come approssimazione di α_1.

Il metodo di Bernoulli corrisponde all'applicazione del metodo delle potenze [3] per il calcolo dell'autovalore di modulo massimo della matrice di Frobenius

$$F = \begin{bmatrix} 0 & 1 & & & \\ 0 & 0 & \ddots & & \\ \vdots & & \ddots & 1 & \\ 0 & & & 0 & 1 \\ -\dfrac{a_0}{a_n} & -\dfrac{a_1}{a_n} & \ldots & \ldots & -\dfrac{a_{n-1}}{a_n} \end{bmatrix}.$$

Fissato un vettore iniziale $\mathbf{t}_0$, il metodo delle potenze è definito dalle relazioni

$$\mathbf{t}_i = F\mathbf{t}_{i-1} = \ldots = F^i \mathbf{t}_0.$$

Nel caso del metodo di Bernoulli è

$$\mathbf{t}_0 = [y_0, y_1, \ldots, y_{n-1}]^T \quad \text{e} \quad \mathbf{t}_i = [y_i, y_{i+1}, \ldots, y_{n+i-1}]^T,$$

infatti risulta

$$F\mathbf{t}_i = [y_{i+1}, y_{i+2}, \ldots, y_{n+i}]^T.$$

Il metodo di Bernoulli ha convergenza lineare, ma è possibile ottenere una convergenza quadratica con la modifica descritta nell'esercizio 3.63.

Il vettore iniziale $\mathbf{t}_0$ deve essere scelto in modo tale che $c_1 \neq 0$. In pratica, anche se questa condizione non è verificata, per la presenza degli errori di arrotondamento dopo qualche passo le componenti y_k effettivamente calcolate risultano comunque esprimibili nella forma (89), con $c_1 \neq 0$. Per evitare errori di overflow generati dalla crescita delle componenti del vettore $\mathbf{t}_i$, conviene ad ogni passo normalizzare $\mathbf{t}_i$ dividendo le componenti per quella di massimo modulo.

3.66 Esempio. Si applica il metodo di Bernoulli al polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096,$$

di cui sono stati calcolati gli zeri negli esempi 3.62 e seguenti. Assumendo come vettore iniziale $\mathbf{t}_0 = [1, 1, 1, 1, 1, 1]^T$ e normalizzando ad ogni passo con la componente di massimo modulo, si ottiene la successione:

k	y_{k+1}/y_k	k	y_{k+1}/y_k
5	0.9230771	.	...
6	− 1.333353	14	17.46086
7	36.00020	15	17.46062
8	18.66676	16	17.46056
.	...	17	17.46054

Il procedimento è stato arrestato quando la differenza fra due valori successivi è diventata minore di 10^{-5} e il valore 17.46054 viene assunto come approssimazione dello zero di massimo modulo del polinomio. La velocità di convergenza, abbastanza alta, è determinata dal fattore

$$\left|\frac{\alpha_2}{\alpha_1}\right| \approx 0.265,$$

che in questo caso è piuttosto piccolo.

Applicando il metodo di Bernoulli al polinomio

$$p(x) = x^5 - 4\,x^4 + 4\,x^3 - 17\,x^2 + 10\,x - 24,$$

di cui sono stati calcolati gli zeri nell'esempio 3.64, e procedendo come nel caso precedente si ottiene la successione:

k	y_{k+1}/y_k	k	y_{k+1}/y_k
4	31.00000	.	...
5	4.870968	17	3.999973
6	3.384106	18	4.000020
7	3.876713	19	3.999998
.	...	20	3.999996

In questo caso la velocità di convergenza è determinata dal fattore

$$\left|\frac{\alpha_2}{\alpha_1}\right| = \frac{\sqrt{3}}{4} \approx 0.433. \qquad \blacksquare$$

Applicando il metodo di Bernoulli al polinomio $p_1(x)$, ottenuto per trasformazione a radici reciproche (si veda l'esercizio 3.50), si può approssimare lo zero di minimo modulo α_n di $p(x)$. Dopo che è stato calcolato questo zero con la tecnica della deflazione si può ridurre il grado del polinomio $p(x)$ e riapplicare il metodo di Bernoulli, e così via.

3.67 Esempio. Si applica il metodo di Bernoulli per il calcolo dello zero di minimo modulo dei polinomi dell'esempio 3.66. Nel primo caso il polinomio ottenuto per trasformazione a radici reciproche è

$$p_1(x) = 4096\,x^6 - 13312\,x^5 + 16640\,x^4 - 9984\,x^3 + 2912\,x^2 - 364\,x + 13,$$

e applicando il metodo di Bernoulli come nell'esempio precedente si ottiene la successione dei rapporti:

k	y_{k+1}/y_k	k	y_{k+1}/y_k
5	0.9997559	.	...
6	0.9992063	48	0.9855116
7	0.9984114	49	0.9855004
8	0.9974642	50	0.9854895
.	...	51	0.9854796

Si assume quindi il valore $1.014734 = \dfrac{1}{0.9854796}$ come approssimazione della soluzione di minimo modulo della prima equazione. La velocità di convergenza, piuttosto bassa, è determinata dal fattore

$$\left|\frac{1/\alpha_5}{1/\alpha_6}\right| \approx 0.887,$$

che è abbastanza elevato. Nel secondo caso il polinomio ottenuto per trasformazione a radici reciproche è

$$p_1(x) = 24\,x^5 - 10\,x^4 + 17\,x^3 - 4\,x^2 + 4\,x - 1.$$

e applicando il metodo di Bernoulli come nell'esempio precedente si ottiene la successione dei rapporti:

k	y_{k+1}/y_k	k	y_{k+1}/y_k
4	− 0.2500000	.	...
5	3.0833330	97	1.0944510
6	0.1328828	98	0.04315011
7	− 3.2867231	99	− 11.08747
.	...	.	...

Questa successione non converge, e ciò è dovuto al fatto che il polinomio $p(x)$ ha due zeri diversi (complessi) di modulo minimo. ∎

Il metodo di Bernoulli può essere modificato in modo da approssimare anche zeri diversi di modulo massimo, come nel caso di zeri complessi coniugati. Si veda per questo il metodo delle potenze per il calcolo degli autovalori di una matrice in [3].

Derivato dal metodo di Bernoulli, il metodo *qd* (abbreviazione di *quotient difference*) permette di effettuare la deflazione in modo implicito. Nel metodo di Bernoulli si calcolano i quozienti

$$q_k = \frac{y_{k+1}}{y_k}\,,$$

che convergono allo zero di modulo massimo del polinomio. Gli elementi di questa successione, indicati con $q_k^{(1)}$, formano la prima colonna di una tabella di elementi $e_k^{(i)}$, $q_k^{(i)}$ che vengono generati alternando differenze e quozienti come segue

$$\left.\begin{aligned} e_{k-1}^{(i)} &= (q_k^{(i)} - q_{k-1}^{(i)}) + e_{k-1}^{(i-1)}, \\ q_k^{(i+1)} &= \frac{e_k^{(i)}}{e_{k-1}^{(i)}}\, q_k^{(i)}, \end{aligned}\right\} \quad \begin{aligned} &i = 1, \ldots, n-1, \\ &k = i, i+1, \ldots \end{aligned} \tag{90}$$

dove si assume $e_k^{(0)} = 0$, per $k \geq 0$. Per $n = 4$ ad esempio lo schema generale

del qd è dato da

$$
\begin{array}{ccccccccc}
 & q_0^{(1)} & & & & & & & \\
0 & & e_0^{(1)} & & & & & & \\
 & q_1^{(1)} & & q_1^{(2)} & & & & & \\
0 & & e_1^{(1)} & & e_1^{(2)} & & & & \\
 & q_2^{(1)} & & q_2^{(2)} & & q_2^{(3)} & & & \\
0 & & e_2^{(1)} & & e_2^{(2)} & & e_2^{(3)} & & \\
 & q_3^{(1)} & & q_3^{(2)} & & q_3^{(3)} & & q_3^{(4)} & \\
0 & & e_3^{(1)} & & e_3^{(2)} & & e_3^{(3)} & & \\
 & q_4^{(1)} & & q_4^{(2)} & & q_4^{(3)} & & q_4^{(4)} & \\
\vdots & & \vdots & & \vdots & & \vdots & & \\
 & \vdots & & \vdots & & \vdots & & \vdots &
\end{array}
$$

Lo schema ovviamente non è applicabile se per qualche k è $y_k = 0$ o $e_k^{(i)} = 0$ per qualche $i > 0$. È però possibile dare delle condizioni sufficienti perché ciò non accada. Ad esempio se gli zeri di $p(x)$ sono di modulo distinto, cioè $|\alpha_1| > |\alpha_2| > \ldots > |\alpha_n|$, e quindi tutti reali, per valori abbastanza elevati di i risulta $e_k^{(i)} \neq 0$ e

$$\lim_{k\to\infty} q_k^{(i)} = \alpha_i \quad \text{e} \quad \lim_{k\to\infty} e_k^{(i)} = 0. \tag{91}$$

Inoltre è

$$
\begin{aligned}
&|q_k^{(1)} - \alpha_1| = O\Big(\Big|\frac{\alpha_2}{\alpha_1}\Big|^k\Big),\\
&|q_k^{(i)} - \alpha_i| = O\Big(\Big|\frac{\alpha_i}{\alpha_{i-1}}\Big|^k + \Big|\frac{\alpha_{i+1}}{\alpha_i}\Big|^k\Big), \quad i = 2,\ldots,n-1,\\
&|q_k^{(n)} - \alpha_n| = O\Big(\Big|\frac{\alpha_n}{\alpha_{n-1}}\Big|^k\Big),\\
&|e_k^{(i)}| = O\Big(\Big|\frac{\alpha_{i+1}}{\alpha_i}\Big|^k\Big), \quad i = 1,\ldots,n-1,
\end{aligned}
\tag{92}
$$

(si veda [13]).

Il metodo qd ottenuto mediante le relazioni (90), calcolando gli elementi della tabella da sinistra a destra, è però numericamente instabile a causa del fenomeno della cancellazione che si ha nel calcolo di $q_k^{(i)} - q_{k-1}^{(i)}$, cioè nella sottrazione di due elementi consecutivi di una stessa successione. È possibile però calcolare le stesse quantità in forma diversa, riga per riga

anziché colonna per colonna, mediante le seguenti relazioni ricavate dalle (90)

$$\left.\begin{aligned} q_{k+1}^{(i)} &= (e_k^{(i)} - e_k^{(i-1)}) + q_k^{(i)}, \quad i = 1, \ldots, n, \\ e_{k+1}^{(i)} &= \frac{q_{k+1}^{(i+1)}}{q_{k+1}^{(i)}}\, e_k^{(i)}, \quad i = 1, \ldots, n-1, \end{aligned}\right\} \quad k = 0, 1, \ldots \tag{93}$$

Per poter applicare queste relazioni occorre assegnare le prime due righe della tabella, e ciò può essere fatto, se $a_i \neq 0$, nel modo seguente

$$q_0^{(1)} = -\,\frac{a_{n-1}}{a_n}\,, \quad q_0^{(j)} = 0,\ j = 2, \ldots, n,$$
$$e_0^{(j)} = \frac{a_{n-j-1}}{a_{n-j}}\,,\ j = 1, 2, \ldots, n-1.$$

Inoltre si assume

$$e_j^{(0)} = e_j^{(n)} = 0,\ j = 0, 1, \ldots$$

La tabella dei $q_k^{(i)}$ ed $e_k^{(i)}$ così costruita coincide con quella che si otterrebbe utilizzando le (90) a partire da un opportuno vettore iniziale $[y_0, y_1, \ldots, y_{n-1}]$. Valgono perciò le relazioni (91) e (92). A differenza del caso precedente, in questo caso il calcolo di $q_{k+1}^{(i)}$ è meno influenzato dal fenomeno della cancellazione, perché gli elementi $e_k^{(i)}$ ed $e_k^{(i-1)}$ appartengono a due successioni diverse, che per la (92) tendono a zero con un comportamento diverso.

3.68 Esempio. Si applica il metodo qd al polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096,$$

di cui nell'esempio 3.66 è stato approssimato lo zero di modulo massimo con il metodo di Bernoulli. Si costruisce con le (93) la seguente tabella dei $q_k^{(i)}$ (per esigenze di spazio non si riportano le colonne degli $e_k^{(i)}$)

k	$q_k^{(1)}$	$q_k^{(2)}$	$q_k^{(3)}$	$q_k^{(4)}$	$q_k^{(5)}$	$q_k^{(6)}$
0	28.00000	0.	0.	0.	0.	0.
1	20.00000	4.571429	1.761904	0.8666661	0.4923077	0.3076923
2	18.17141	5.078573	2.263512	1.232046	0.7544380	0.4999999
3	17.66035	5.000667	2.406236	1.400005	0.9052602	0.6274508
.	...	...	...	...	...	...
38	17.46045	4.630318	2.274118	1.476439	1.146059	1.012456
39	17.46045	4.630318	2.274117	1.476419	1.145822	1.012712
40	17.46045	4.630318	2.274116	1.476403	1.145611	1.012938

Come si vede, la prima e la seconda colonna non vengono più modificate alla 40-esima iterazione (in realtà i valori della prima colonna si stabilizzano alla 15-esima iterazione e quelli della seconda alla 26-esima iterazione). La convergenza agli zeri di modulo minore è invece più lenta, perché gli zeri sono più vicini. Se per costruire la tabella, anziché le (93) fossero state usate le (90), si sarebbe avuta convergenza solo a valori della prima colonna. Nella colonna delle $q_k^{(2)}$ infatti, dopo un'iniziale tendenza alla convergenza, dalla 15-esima iterazione si hanno i valori

k	$q_k^{(2)}$
15	4.631174
16	4.866145
17	4.108379
18	4.365189
19	17.46054

Quindi la successione converge ad α_1 perché la successione $q_k^{(1)}$ ha raggiunto il suo limite e quindi le $e_k^{(1)}$ risultano costanti per $k \geq 19$. La situazione è ancora peggiore nelle colonne successive. ∎

Il metodo qd in questa seconda versione stabile può essere utilizzato anche per calcolare dei fattori del polinomio $p(x)$. Infatti una volta calcolati $q_k^{(i)}$ ed $e_k^{(i)}$ si costruiscono i polinomi $p_k^{(i)}(x)$ definiti dalle relazioni

$$p_k^{(i+1)}(x) = xp_k^{(i)}(x) - q_k^{(i+1)}p_{k-1}^{(i)}(x),\ i = 0, 1, \ldots, n-1,\ k = 1, 2, \ldots, \quad (94)$$

a partire da

$$p_k^{(0)}(x) = 1.$$

È possibile dimostrare [13] che se

$$|\alpha_n| < |\alpha_{n-1}| \leq \ldots \leq |\alpha_{i+1}| < |\alpha_i| \leq \ldots \leq |\alpha_1|,$$

allora esiste il limite

$$\lim_{k\to\infty} p_k^{(i)}(x) = p_i(x),$$

dove

$$p_i(\alpha_j) = 0, \quad \text{per} \quad j = 1, \ldots, i,$$

cioè il polinomio $p_i(x)$ è un fattore di $p(x)$, ed inoltre i coefficienti del polinomio $|p_k^{(i)}(x) - p_i(x)|$ tendono a zero come $\left|\dfrac{\alpha_{i+1}}{\alpha_i}\right|^k$.

In questo modo è possibile calcolare i fattori di $p(x)$ anche nel caso di zeri distinti e con lo stesso modulo. La seguente variante di Householder permette il calcolo dei fattori lineari di $p(x)$, corrispondenti a zeri che hanno modulo distinto, e di fattori di grado opportuno, i cui zeri hanno tutti lo stesso modulo.

1. Si calcolano gli elementi $q_k^{(i)}$, $e_k^{(i)}$;
2. si calcolano, mediante la (94), i coefficienti di $p_k^{(i)}(x)$;
3. se per un indice i o k si ha $|e_k^{(i)}| < \epsilon$, dove ϵ è un valore prefissato, si considera ottenuto $p_k^{(i)}(x)$ e si prosegue sostituendo $p_k^{(i)}(x)$ con 1.

I polinomi così calcolati sono approssimazioni di fattori di $p(x)$ che hanno zeri di ugual modulo. Per polinomi a coefficienti reali è possibile estrarre i fattori quadratici.

3.69 Esempio. Per il polinomio

$$p(x) = 13\,x^6 - 364\,x^5 + 2912\,x^4 - 9984\,x^3 + 16640\,x^2 - 13312\,x + 4096,$$

a cui è stato applicato il metodo del qd nell'esempio 3.68, si costruiscono i polinomi $p_k^{(i)}(x)$, per mezzo della (94). Risulta

alla 14^a iterazione $\quad p_1(x) = x - 17.46045,$

alla 26^a iterazione $\quad p_2(x) = x^2 - 22.09076\,x + 80.84744,$

alla 40^a iterazione $\quad p_3(x) = x^3 - 24.36487\,x^2 + 131.0844\,x - 183.8565,$

alla 61^a iterazione $\quad p_4(x) = x^4 - 25.84120\,x^3 + 167.0552\,x^2 - 377.3813\,x + 271.4346,$

alla 108^a iterazione $\quad p_5(x) = x^5 - 26.98508\,x^4 + 196.6144\,x^3 - 568.4722\,x^2 + 703.1130\,x - 310.4883,$

alla 108^a iterazione $\quad p_6(x) = x^6 - 27.99974x^5 + 223.9952x^4 - 767.9697x^3 + 1279.922x^2 - 1023.912x + 315.0413.$

$p_1(x)$ ha come zero un'approssimazione di α_1, $p_2(x)$ ha come zeri delle approssimazioni di α_1 e α_2, …, $p_6(x)$ ha come zeri delle approssimazioni di α_1, $\alpha_2, \ldots, \alpha_6$, ed infatti, a meno degli errori di arrotondamento, $p_6(x)$ è il polinomio monico equivalente a $p(x)$.

Lo stesso procedimento si applica al polinomio

$$p(x) = x^5 - 4\,x^4 + 4\,x^3 - 17\,x^2 + 10\,x - 24,$$

a cui è stato applicato il metodo di Bernoulli negli esempi 3.66 e 3.67. Lo zero di modulo minimo non può essere approssimato con il metodo di Bernoulli perché è complesso. Per le successioni $\{p_k^{(1)}\}$, $\{p_k^{(3)}\}$, $\{p_k^{(5)}\}$ risulta

alla 18^a iterazione	$p_1(x) = x - 3.999991,$
alla 27^a iterazione	$p_3(x) = x^3 - 2.995079\,x^2 - 1.018381\,x - 12.00499,$
alla 27^a iterazione	$p_5(x) = x^5 - 3.999995\,x^4 + 3.998352\,x^3 - 16.99072\,x^2$ $+9.990860\,x - 24.00362.$

Il polinomio $p_1(x)$ ha come zero un'approssimazione di α_1, il polinomio $p_3(x)$ ha come zeri delle approssimazioni di α_1, α_2 e α_3 e il polinomio $p_5(x)$ coincide, a meno degli errori di arrotondamento, con $p(x)$. Le successioni $\{p_k^{(2)}\}$, $\{p_k^{(4)}\}$ non sono convergenti, perché il polinomio $p(x)$ ha quattro zeri complessi $\alpha_2 = \overline{\alpha}_3$ e $\alpha_4 = \overline{\alpha}_5$. ∎

Esercizi proposti

3.1 Si dimostri che, se $g(x) \in C^1[\alpha - \rho, \alpha + \rho]$, dove $\alpha = g(\alpha)$ e $\rho > 0$ e

$$|g'(x)| < 1, \quad \text{per } 0 < |x - \alpha| < \rho,$$

allora la successione $\{x_i\}$ ottenuta dalla (5) a partire da un punto iniziale x_0 tale che $|x_0 - \alpha| \leq \rho$, è convergente ad α.

(Traccia: se $|g'(\alpha)| < 1$ e $|x_0 - \alpha| < \rho$, vale il teorema 3.3 riferito all'intervallo $\{x \ : \ |x - \alpha| \leq |x_0 - \alpha|\}$.
Se $|g'(\alpha)| < 1$ e $|x_0 - \alpha| = \rho$, si osservi che nella (10) per $i = 1$ è $|\xi_0 - \alpha| < \rho$, quindi è possibile condurre la dimostrazione del teorema 3.3 ponendo però

$$\lambda = \max_{|x-\alpha| \leq |\xi_0 - \alpha|} |g'(x)|.$$

Se $|g'(\alpha)| = 1$, da (10) segue che la successione $|x_i - \alpha|$ è monotona decrescente e quindi convergente. Si verifichi allora che se fosse

$$\lim_{i \to \infty} |x_i - \alpha| > 0,$$

esisterebbe un punto fisso di $g(x)$ in $[\alpha - \rho, \alpha + \rho]$ diverso da α, e ciò non è possibile.)

3.2 Si dice che una funzione $g(x)$ continua in un intervallo $[a, b]$ *soddisfa alla condizione di Lipschitz con costante* λ se

$$|g(x) - g(x')| \leq \lambda |x - x'|, \quad \text{per ogni } x, x' \in [a, b].$$

Si dimostri che se in un intorno di un punto fisso α dell'equazione $x = g(x)$ la $g(x)$ soddisfa alla condizione di Lipschitz con costante $\lambda < 1$, allora la

successione definita dal metodo $x_{i+1} = g(x_i)$, a partire da un punto iniziale dell'intorno, è convergente.

(Traccia: si segua la dimostrazione del teorema 3.3)

3.3 Sia $g(x) \in C[a,b]$ e sia α l'unico punto fisso di $g(x)$ in $[a,b]$. Si verifichi che

se $x < g(x) < \alpha$ per $a \le x < \alpha$ vi è convergenza monotona per $a \le x_0 < \alpha$,
se $\alpha < g(x) < x$ per $\alpha < x \le b$ vi è convergenza monotona per $\alpha < x_0 \le b$,
se $\left\{ \begin{array}{l} \alpha < g(x) < 2\alpha - x \le b \text{ per } a \le x < \alpha \\ a \le 2\alpha - x < g(x) < \alpha \text{ per } \alpha < x \le b \end{array} \right\}$ vi è convergenza alternata per $a \le x_0 \le b$.

Si dia un esempio di una funzione $g(x)$ tale che $g'(\alpha) = 0$ e (1) la convergenza è monotona, (2) la convergenza è alternata, (3) la convergenza non è né monotona né alternata.

(Traccia: (1) $g(x) = x^3$, per $|x| \le \frac{1}{2}$, (2) $g(x) = -x^3$, per $|x| \le \frac{1}{2}$, (3) $g(x) = x^2 \sin \frac{1}{x}$, per $|x| \le \frac{1}{2}$.)

3.4 Per $0 < k < 1$ i tre metodi iterativi

$$(1) \quad x_{i+1} = (1-k)x_i + 1\,, \quad (2) \quad x_{i+1} = (2 - kx_i)x_i\,,$$

$$(3) \quad x_{i+1} = [3(1 - kx_i) + (kx_i)^2]x_i\,,$$

convergono alla radice dell'equazione

$$x - \frac{1}{k} = 0$$

(consentendo così di calcolare l'inverso di un numero k senza eseguire divisioni). Per ciascuno dei tre metodi si dica qual è l'ordine e come deve essere scelto x_0.

(Traccia: (1) ordine 1, convergenza per ogni x_0; (2) ordine 2, dalla condizione del teorema 3.3 segue che il metodo converge per $\frac{1}{2k} < x_0 < \frac{3}{2k}$, ma tenendo conto del fatto che si tratta del metodo delle tangenti applicato all'equazione $\frac{1}{x} - k = 0$, vi è convergenza per $0 < x_0 < \frac{2}{k}$; (3) è il metodo delle tangenti applicato all'equazione

$$\frac{kx - 1}{\sqrt{x(2 - kx)}} = 0,$$

l'ordine è 3, vi è convergenza per $0 < x_0 < \dfrac{2}{k}$.)

3.5 Si studi la convergenza dei tre metodi iterativi, per $k > 2$:

$$(1) \quad x_{i+1} = \frac{1}{k - x_i}, \quad (2) \quad x_{i+1} = \frac{1 + x_i^2}{k}, \quad (3) \quad x_{i+1} = k - \frac{1}{x_i}.$$

(Traccia: siano α e β le due soluzioni dell'equazione $x^2 - kx + 1$, con $\alpha < \dfrac{k}{2} < \beta$. Per (1) si ha convergenza ad α, scegliendo $x_0 < \dfrac{k}{2}$. Per (2) si ha convergenza ad α scegliendo $-\dfrac{k}{2} < x_0 < \dfrac{k}{2}$. Per (3) si ha convergenza a β scegliendo $x_0 > \dfrac{k}{2}$.)

3.6 Per l'equazione

$$x^3 - 3x^2 - 4x + 1 = 0$$

si dica quante sono le soluzioni, se ne diano degli intervalli di separazione e si esamini la convergenza dei metodi iterativi $x_{i+1} = g(x_i)$, dove

$$(1) \quad g(x) = \frac{3x^2 + 4x - 1}{x^2}, \qquad (2) \quad g(x) = \frac{x^3 - 3x^2 + 1}{4},$$
$$(3) \quad g(x) = \frac{3x^2 - 1}{x^2 - 4}.$$

(Traccia: tre soluzioni, $-2 < \alpha < -1$, $0 < \beta < 0.5$, $3 < \gamma < 4$. Il metodo (1) converge alla sola γ, assumendo $x_0 \in (2, 4)$. Il metodo (2) converge alla sola β, assumendo $x_0 \in (0, 1.5)$. Il metodo (3) converge a β, assumendo $x_0 \in (0, 0.5)$ e a γ, assumendo $x_0 \in (3.6, 4)$. In ogni caso l'ordine è 1. Invece sfruttando le condizioni date nell'esercizio 3.3 risultano degli intervallli di convergenza maggiori: per (1) $x_0 > \beta$, per (2) $x_0 \in (\alpha, \gamma)$, per (3) se $x_0 \in \left(\alpha, \dfrac{\sqrt{3}}{3}\right)$ si ha convergenza a β, se $x_0 > 2$ si ha convergenza a γ.)

3.7 Per le seguenti equazioni $x = g(x)$, dove

(1) $$g(x) = k(1 + \mathrm{e}^{-x}), \quad k > 0,$$

(2) $$g(x) = \begin{cases} 1 + \mathrm{e}^{-\frac{1}{|x-1|}}, & \text{per } x \neq 1, \\ 1 & \text{per } x = 1, \end{cases}$$

si dica quante sono le soluzioni, se ne diano degli intervalli di separazione e si esamini la convergenza del metodo iterativo

$$x_{i+1} = g(x_i).$$

(Traccia: (1) una sola radice $\alpha \in (k, k(1 + \mathrm{e}^{-k}))$; il metodo è convergente per ogni $x_0 > 0$; (2) la sola radice $\alpha = 1$; il metodo è convergente per ogni punto iniziale x_0 e non ha ordine finito, in quanto $g^{(r)}(1) = 0$ per ogni intero r.)

3.8 Per le seguenti equazioni $f(x) = 0$, dove

(1) $$f(x) = \log x - x^2 + 1,$$

(2) $$f(x) = 2x^2 + \log(kx), \quad k > 0,$$

(3) $$f(x) = 3x^3 + \mathrm{e}^x - 4,$$

si dica quante sono le soluzioni, se ne diano degli intervalli di separazione e esamini la convergenza di vari metodi iterativi.

(Traccia: (1) due radici reali $\alpha \in (\mathrm{e}^{-1}, 0.5)$ e $\beta = 1$; il metodo iterativo

$$x_{i+1} = g(x_i), \quad g(x) = \sqrt{1 + \log x}$$

è convergente alla radice β per $x_0 > \alpha$. Il metodo delle tangenti converge alla soluzione α se $x_0 \in (0, \gamma)$, dove $\gamma \approx 0.653$ è la soluzione dell'equazione $x^2 + \log x = 0$, e alla soluzione β se $x_0 > \dfrac{\sqrt{2}}{2}$.

(2) Una sola radice $\alpha \in \left(0, \dfrac{1}{k}\right)$; il metodo iterativo

$$x_{i+1} = g(x_i), \quad g(x) = \frac{1}{k}\, \mathrm{e}^{-2x^2}$$

è convergente per $k > \dfrac{2}{\sqrt{\mathrm{e}}}$ e per ogni $x_0 > 0$. Il metodo iterativo

$$x_{i+1} = g(x_i), \quad g(x) = -\,\frac{\log(kx)}{2x}$$

non è convergente. Il metodo delle tangenti converge ad α per $x_0 > 0$.

(3) Una sola radice, $\alpha \in \left(\dfrac{1}{2}, 1\right)$. Il metodo iterativo

$$x_{i+1} = g(x_i), \quad g(x) = \log(4 - 3x^3)$$

non è convergente. Il metodo delle tangenti converge ad α per ogni scelta di x_0.)

3.9 Si dica per quali valori del parametro k l'equazione

$$x^3 - x + k = 0$$

ha una radice negativa α e due radici positive β e γ. Per tali valori di k si studi la convergenza dei metodi iterativi

$$x_{i+1} = x_i^3 + k, \tag{1}$$

$$x_{i+1} = \frac{k}{1 - x_i^2}\,, \tag{2}$$

(Traccia: $k \in \left(0, \dfrac{2}{3\sqrt{3}}\right)$; entrambi i metodi convergono alla sola soluzione β; se si sfruttano le condizioni date nell'esercizio 3.3, risulta che vi è convergenza, scegliendo per (1) $\alpha < x_0 < \gamma$, e per (2) $-\gamma < x_0 < \gamma$.)

3.10 Sia

$$f(x) = (1-x)^p + x^q - 1, \quad p, q \text{ interi positivi.}$$

Si dica quante sono le radici dell'equazione $f(x) = 0$ al variare di p e q, e si studi la convergenza dei due metodi iterativi $x_{i+1} = g(x_i)$, dove

$$(1)\ g(x) = (1 - (1-x)^p)^{1/q}, \quad (2)\ g(x) = 1 - (1 - x^q)^{1/p}.$$

(Traccia: escluso il caso banale $p = q = 1$, se il grado dell'equazione è pari, vi sono solo le due radici $\alpha = 0$ e $\beta = 1$; se il grado è dispari, oltre ad α e β esiste un'altra radice γ, $\gamma \le -1$ se $p < q$ e $\gamma \ge 2$ se $p > q$. Per il metodo (1) si ha convergenza di ordine p a β, non si ha convergenza ad α, si ha convergenza di ordine 1 a γ se $p < q$, non si ha convergenza a γ se $p > q$. Risultati opposti valgono per il metodo (2).)

3.11 Si vuole determinare il minimo valore positivo α in cui la curva di equazione $y = x^3$ è tangente a una delle curve della famiglia

$$y = \mathrm{e}^x + k, \quad k \in \mathbf{R}.$$

a) Si scriva un'equazione $f(x) = 0$ di cui α è la minima soluzione positiva;

b) si dica quante soluzioni reali ha l'equazione $f(x) = 0$ e se ne diano degli intervalli di separazione;

c) si esaminino vari metodi iterativi per il calcolo di α.

(Traccia: a) $f(x) = 3x^2 - \mathrm{e}^x$; b) $\dfrac{3}{4} < \alpha < 1$, $-\dfrac{1}{2} < \beta < 0$, $\dfrac{7}{2} < \gamma < 4$;
c) (1) posto

$$g(x) = \frac{\mathrm{e}^x}{3x}\,,$$

si ha

$$|g'(x)| < 1 \quad \text{per } \frac{3}{4} < x < 1.$$

Scegliendo x_0 in tale intervallo la successione $x_{i+1} = g(x_i)$ converge ad α con ordine di convergenza 1. Sfruttando le condizioni date nell'esercizio 3.3, risulta che vi è convergenza per $x_0 \in (\alpha, \gamma)$.
(2) posto $g(x) = \log 3x^2$, è $g'(x) > 1$ nell'intervallo di separazione e il metodo non è convergente.
(3) nell'intervallo $S = (\alpha, \log 6)$ è $f(x) > 0$, $f'(x) \neq 0$ e $f''(x) > 0$. Quindi il metodo delle tangenti converge ad α per ogni $x_0 \in S$, con ordine di convergenza 2. Si scelga ad esempio $x_0 = 1$.)

3.12 Si determini il numero delle radici reali dell'equazione

$$f(x) = kx, \quad \text{dove} \quad f(x) = \mathrm{e}^{1-x^2} - 1, \quad k > 0,$$

e si dica per quali valori di k il metodo delle corde

$$x_{i+1} = g(x_i), \quad g(x) = \frac{f(x)}{k}$$

converge alla soluzione positiva α dell'equazione.

(Traccia: indicate con $\overline{\alpha} \approx 1.71$ la soluzione positiva dell'equazione

$$\mathrm{e}^{1-x^2} = \frac{1}{1+2x^2}\,,$$

e con

$$\overline{k} = \frac{2\overline{\alpha}}{1+2\overline{\alpha}^{\,2}} \approx 0.5,$$

vi sono una sola soluzione positiva oppure una positiva e due negative a seconda che k sia maggiore o minore di $\overline{k}$. Per ogni k è $\alpha \in (0,1)$. Per $x \in (0,1)$ risulta $|g'(x)| < 1$ se $k > \sqrt{2\mathrm{e}} \approx 2.33$.)

3.13 Si verifichi che la funzione

$$f(x) = k \log x - \mathrm{e}^{-x}, \quad k > 0$$

ha una sola soluzione reale α al variare di k e se ne dia un intervallo di separazione. Si determini l'intero m tale che il metodo delle corde converga il più rapidamente possibile.

(Traccia: è $\alpha \in (1, \mathrm{e}^{1/(k\mathrm{e})})$. Si scelga m fra gli interi $\lfloor k \rfloor$ (se $k \geq 1$), $\lceil k \rceil$ e $\lceil k + \frac{1}{\mathrm{e}} \rceil$.)

3.14 Sia $f(x) = x^3 - 3x^2 + 6x - 1$. Si dica per quali valori interi di k il metodo delle corde

$$x_{i+1} = x_i - \frac{f(x_i)}{k}$$

converge alla radice reale dell'equazione $f(x) = 0$ e fra questi si scelga il valore di k per cui la velocità di convergenza è massima.

(Traccia: una soluzione reale $\alpha \in \left(\frac{1}{6}, \frac{1}{3}\right)$; per $k \geq 3$ il metodo converge e la velocità di convergenza è massima per $k = 5$.)

3.15 Si verifichi che per un valore opportuno del parametro k la funzione

$$f(x) = \mathrm{e}^x - k(\log^2 x + x)$$

ha un unico punto stazionario di flesso α che può essere approssimato con il metodo delle tangenti, applicato scegliendo come x_0 un qualsiasi punto dell'intervallo $(0, \mathrm{e})$.

(Traccia: i punti stazionari di flesso sono le soluzioni dell'equazione

$$\phi(x) = \log x - \frac{2 - x^2}{2x + 2} = 0.$$

È

$$\phi'(x) > 0 \quad \text{e} \quad \phi''(x) < 0 \quad \text{per ogni } x,$$

quindi l'equazione $\phi(x) = 0$ ha un'unica soluzione reale $\alpha \in (1, \sqrt{2})$, e il metodo delle tangenti converge per ogni $x_0 \in (0, \alpha)$. Però scegliendo $x_0 \in (\alpha, \mathrm{e})$ risulta $x_1 \in (0, \alpha)$, e quindi il metodo delle tangenti converge anche in tale intervallo.)

3.16 Sia m il minimo della funzione $f(x) = x^4 - x^3$. Si scriva l'equazione $\phi(x) = 0$ le cui soluzioni sono le ascisse dei punti di tangenza alla curva di equazione $y = f(x)$ delle rette passanti per il punto $(0, m)$. Si dica quante soluzioni reali ha l'equazione $\phi(x) = 0$ e si studi la convergenza del metodo delle tangenti a tali soluzioni.

(Traccia: è $\phi(x) = 3x^4 - 2x^3 - \dfrac{27}{256}$. Vi sono due radici reali, una negativa, che può essere approssimata scegliendo $x_0 < 0$, e una positiva uguale a $\dfrac{3}{4}$, che può essere approssimata scegliendo $x_0 > \dfrac{1}{2}$.)

3.17 Si determini il valore di k per cui la soluzione positiva α dell'equazione

$$f(x) = \mathrm{e}^{-x^2} - k(x-1) - 1 = 0$$

coincide con un punto di flesso, e per tale valore di k si studi la convergenza del metodo delle tangenti ad α.

(Traccia: è $k = \left(\mathrm{e}^{-1/2} - 1\right) / \left(\dfrac{1}{\sqrt{2}} - 1\right) \approx 1.34$. Il metodo è del terzo ordine e converge per ogni x. Infatti, posto $g(x) = x - \dfrac{f(x)}{f'(x)}$ e $\overline{k} = 1 - \dfrac{1}{k}$, si verifichi che $g(x) > \overline{k}$ per ogni x, che $g(x) > \alpha$ per $\overline{k} < x < \alpha$, che $\overline{k} < g(x) < \alpha$ per $x > \alpha$ e che $|g'(x)| < 1$ per $x > \overline{k}$.)

3.18 Si dica quante soluzioni ha l'equazione

$$f(x) = \cos x + (x+1)\,\mathrm{e}^x,$$

si determini un intervallo di separazione per la soluzione α di minimo modulo e si studi il comportamento del metodo delle tangenti quando si sceglie $x_0 = 0$.

(Traccia: l'equazione ha infinite soluzioni, tutte negative. È $\alpha \in \left(-\dfrac{\pi}{2}, -1\right)$. Nell'intervallo $(\alpha, 0]$ è $f(x) > 0$, $f'(x) > 0$, $f''(x) > 0$. Quindi il metodo delle tangenti converge scegliendo $x_0 = 0$, con ordine di convergenza 2.)

3.19 Il metodo delle tangenti viene applicato all'equazione

$$f(x) = k, \quad \text{dove} \quad f(x) = \begin{cases} \cos x & \text{per } |x| \le 1, \\ \cos x + (x^2-1)^2 & \text{per } |x| > 1. \end{cases}$$

Si determini il valore $\overline{k}$ di k per cui si ha $x_i = (-1)^i$, quando $x_0 = 1$, e si studi, al variare di k, la convergenza del metodo alle diverse soluzioni dell'equazione.

(Traccia: la funzione $f(x)$ è continua e derivabile anche per $|x| = 1$. È $\overline{k} = 2\sin 1 + \cos 1$. Si determini un punto di minimo m di $f(x)$ e si ponga $k_0 = f(m)$. Per $x > 0$ e $k_0 < k < 1$ vi sono due soluzioni separate da m:

per $x_0 = 1$ si ha convergenza alla minore, per $x_0 > m$ si ha convergenza alla maggiore. Per $x > 0$ e $k > 1$ una sola soluzione, il metodo converge per $x_0 > m$.)

3.20 Sia $f(x) = x^n - k$, con $n > 0$ pari, $k > 0$. Si determini p in modo che il metodo delle tangenti $x_{i+1} = g(x_i)$, applicato all'equazione $x^p f(x) = 0$, abbia ordine quanto più elevato possibile, e si verifichi che la funzione $g(x)$ è razionale per ogni p. Si applichi al caso particolare del calcolo di $\sqrt{k}$.

(Traccia: per $p = \dfrac{1-n}{2}$ il metodo è del terzo ordine, e si ha

$$g(x) = x\,\frac{(n-1)x^n + (n+1)k}{(n+1)x^n + (n-1)k}\,.$$

Il metodo è convergente per ogni $x_0 > 0$. Nel caso particolare

$$g(x) = \frac{x^3 + 3kx}{3x^2 + k}\,.\ )$$

3.21 Si dimostri che il teorema 3.25 vale anche nell'ipotesi che sia soltanto $f(x) \in C^1(S)$ e $f(x) \in C^2(S - \{\alpha\})$.

3.22 Sia $p(x)$ il polinomio monico (cioè con primo coefficiente uguale a 1) di secondo grado, avente due zeri distinti non nulli α e β. Si determini un polinomio $h(x)$ di grado non superiore al primo tale che il metodo iterativo

$$x_{i+1} = g(x_i), \quad g(x) = x - \frac{xp(x)}{h(x)}$$

converga con ordine 2 ad α e β. Si esamini in particolare il caso del polinomio $p(x) = x^2 - k$, $k > 0$.

(Traccia: si impongano le condizioni $g'(\alpha) = g'(\beta) = 0$. Si ottiene

$$h(x) = (\alpha + \beta)x - 2\alpha\beta.$$

Nel caso particolare è $h(x) = 2k$ e

$$g(x) = \frac{3kx - x^3}{2k} = \frac{3}{2}\,x - \frac{x^3}{2k}\,.\ )$$

3.23 Sia $f(x) \in C^2[a,b]$ e $x_i \in [a,b]$. Dalla formula di Taylor

$$f(x_{i+1}) = f(x_i) + (x_{i+1} - x_i)f'(x_i) + \frac{(x_{i+1} - x_i)^2}{2}\,f''(x_i) + \ldots$$

troncata al terzo termine e in cui si pone $f(x_{i+1}) = 0$, si ricava

$$x_{i+1} = x_i - \frac{f'(x_i) \pm \sqrt{[f'(x_i)]^2 - 2f(x_i)f''(x_i)}}{f''(x_i)} .$$

Questa relazione può essere usata come metodo iterativo per approssimare una soluzione α dell'equazione $f(x) = 0$, scegliendo la determinazione positiva della radice se $f'(x) < 0$ e quella negativa se $f'(x) > 0$. Si studi la convergenza di tale metodo, che va sotto il nome di metodo di *Newton di secondo grado* e se ne dia l'interpretazione geometrica. Si applichi il metodo in particolare al calcolo della radice cubica di un numero k.

(Traccia: nell'ipotesi che $f'(x) \neq 0$ e $f''(x) \neq 0$ per $|x - \alpha| \leq \rho,\ \rho > 0$, il metodo è almeno del terzo ordine. Geometricamente il punto x_1 è dato dall'intersezione con l'asse x della parabola che nel punto $(x_0, f(x_0))$ ha in comune con la funzione $f(x)$ la derivata prima e seconda. Nel caso particolare

$$x_{i+1} = \frac{3x_i^2 + \sqrt{12kx_i - 3x_i^4}}{6x_i} . \ )$$

3.24 Sia $f(x) \in C^p[a, b]$, $p \geq 2$ intero e sia $\alpha \in [a, b]$ soluzione di molteplicità 1 dell'equazione $f(x) = 0$. Si consideri il metodo iterativo

$$x_{i+1} = g(x_i), \quad g(x) = x - \phi(x) + h(x),$$

$$h(x) = \sum_{i=2}^{p-1} c_i(x)\phi^i(x), \quad \phi(x) = \frac{f(x)}{f'(x)} ,$$

in cui le $c_i(x)$, $i = 2, \ldots, p-1$ sono delle opportune funzioni. È possibile determinare le $c_i(x)$ in modo che il metodo sia di ordine p. Si determinino $c_2(x)$ e $c_3(x)$ in modo da ottenere metodi di ordine 3 e 4. Si applichi al caso particolare del calcolo di $\sqrt{k}$.

(Traccia: si impongano le condizioni $g''(\alpha) = g^{(3)}(\alpha) = 0$. Si ottiene

$$c_2(x) = -\frac{f''(x)}{2f'(x)} , \quad c_3(x) = \frac{f^{(3)}(x)}{3!f'(x)} - \frac{[f''(x)]^2}{2[f'(x)]^2} .$$

Nel caso particolare si ha

$$\phi(x) = \frac{x^2 - k}{2x} , \quad c_2 = -\frac{1}{2x} , \quad c_3 = -\frac{1}{2x^2} ,$$

da cui si ottiene per il metodo del terzo ordine

$$g(x) = \frac{3x^4 + 6kx^2 - k^2}{8x^3}\,,$$

e per il metodo del quarto ordine

$$g(x) = \frac{5x^6 + 15kx^4 - 5k^2x^2 + k^3}{16x^5}\,.\ )$$

3.25 Sia $f(x) \in C^4[a,b]$ e $\alpha \in [a,b]$ soluzione di molteplicità 1 dell'equazione $f(x) = 0$, tale che $f'(\alpha) > 0$. Si determini una funzione $\phi(x) \neq 0$ in $[a,b]$, tale che il metodo delle tangenti applicato all'equazione $\phi(x)f(x) = 0$ converga ad α con ordine 3. Si applichi al caso particolare del calcolo di $\sqrt{k}$.

(Traccia: posto

$$g(x) = x - \frac{y(x)}{y'(x)}, \quad y(x) = \phi(x)f(x),$$

si imponga la condizione che $g''(\alpha) = 0$. Si ottiene

$$\frac{\phi'(\alpha)}{\phi(\alpha)} = -\frac{1}{2}\,\frac{f''(\alpha)}{f'(\alpha)}\,.$$

È quindi sufficiente scegliere

$$\phi(x) = \frac{1}{\sqrt{f'(x)}}\,,$$

e il metodo delle tangenti risulta

$$x_{i+1} = x_i - \frac{2f(x_i)f'(x_i)}{2[f'(x_i)]^2 - f(x_i)f''(x_i)}\,.$$

Nel caso particolare è

$$g(x) = \frac{x^3 + 3kx}{3x^2 + k}$$

e l'ordine di convergenza è 3.)

3.26 Sia $f(x) \in C^5[a,b]$ e $\alpha \in [a,b]$ soluzione di molteplicità 1 dell'equazione $f(x) = 0$. Si determini l'ordine del metodo iterativo

$$x_{i+1} = \frac{1}{2}\,\big[g_1(x_i) + g_2(x_i)\big], \quad g_1(x) = x - \frac{f(x)}{f'(x)}\,,$$

$$g_2(x) = x - \frac{\phi(x)}{\phi'(x)}, \quad \phi(x) = \frac{f(x)}{f'(x)}.$$

Si applichi al caso particolare del calcolo di $\sqrt{k}$.

(Traccia: terzo ordine; nel caso particolare è

$$g(x) = \frac{x^4 + 6kx^2 + k^2}{4x(x^2 + k)}$$

e l'ordine di convergenza è 4.)

3.27 Sia $f(x) \in C^r[a,b]$ e $\alpha \in [a,b]$ soluzione di molteplicità finita r dell'equazione $f(x) = 0$.

a) Si dica se α è anche soluzione (e di quale molteplicità) delle equazioni $f'(x) = 0$ e $\phi(x) = \dfrac{f(x)}{f'(x)} = 0$;

b) si scriva il metodo delle tangenti per l'equazione $\phi(x) = 0$, e se ne determini l'ordine;

c) si applichi in particolare alle equazioni $f(x) = 0$, dove

(1) $f(x) = x^4 - 4x^2 + 4$, (2) $f(x) = x^{5/2} + x^2$, (3) $f(x) = x^4 + x^2$.

(Traccia: a) dalle relazioni

$$f(x) = (x-\alpha)^r \frac{f^{(r)}(\xi_1)}{r!} \quad \text{e} \quad f'(x) = (x-\alpha)^{r-1} \frac{f^{(r)}(\xi_2)}{(r-1)!}$$

segue che

$$\lim_{x\to\alpha} \frac{f'(x)}{(x-\alpha)^{r-1}} = \frac{f^{(r)}(\alpha)}{(r-1)!} \quad \text{e} \quad \lim_{x\to\alpha} \frac{\phi(x)}{x-\alpha} = \frac{1}{r}.$$

Quindi le equazioni $f'(x) = 0$ e $\phi(x) = 0$ hanno la soluzione α di molteplicità rispettivamente $r-1$ e 1; b) il metodo è

$$x_{i+1} = g(x_i), \quad g(x) = x - \frac{f(x)f'(x)}{[f'(x)]^2 - f(x)f''(x)},$$

e l'ordine di convergenza è almeno 2 se $f(x) \in C^{r+1}[a,b]$; c) nei casi particolari è

(1) $$g(x) = \frac{4x}{x^2+2}, \quad r = 2, \quad \text{ordine } 2,$$

(2) $$g(x) = -\frac{x^{3/2}}{10\,x + 17\,x^{1/2} + 8}, \quad r = 2, \quad \text{ordine } \frac{3}{2},$$

(3) $$g(x) = -\frac{2x^3}{2x^4 + x^2 + 1}, \quad r = 2, \quad \text{ordine } 3.\ )$$

3.28 Siano $g_1(x),\ g_2(x) \in C^2[a,b]$ tali che i due metodi iterativi

$$x_{i+1} = g_1(x_i), \quad x_{i+1} = g_2(x_i)$$

convergano entrambi alla stessa soluzione $\alpha \in [a,b]$ e siano del primo ordine, ed inoltre sia $g_1'(x) \neq g_2'(x)$ in tutto $[a,b]$. Si determini una funzione $\rho(x)$ tale che il metodo iterativo

$$x_{i+1} = h(x_i), \quad h(x) = \rho(x)g_1(x) + [1-\rho(x)]g_2(x)$$

converga ad α con ordine 2. Si verifichi che nel caso particolare in cui i due metodi del primo ordine siano i due metodi delle corde

$$x_{i+1} = x_i - \frac{f(x_i)}{m_1}, \quad x_{i+1} = x_i - \frac{f(x_i)}{m_2}, \quad m_1 \neq m_2,$$

si ottiene il metodo delle tangenti per la risoluzione dell'equazione $f(x) = 0$.

(Traccia: si imponga la condizione $h'(\alpha) = 0$. Risulta

$$\rho(\alpha) = \frac{g_2'(\alpha)}{g_2'(\alpha) - g_1'(\alpha)}\,.$$

È quindi sufficiente scegliere

$$\rho(x) = \frac{g_2'(x)}{g_2'(x) - g_1'(x)}\,.\ )$$

3.29 Sia α soluzione dell'equazione $f(x) = 0$. Si dica se sono ben condizionati i problemi del calcolo di $f(x)$, $\phi(x) = \dfrac{f(x)}{f'(x)}$ e $g(x) = x - \phi(x)$ (funzione di iterazione del metodo delle tangenti), per x in un intorno di α.

(Traccia: sia r la molteplicità di α. Si verifichi che in un intorno di α risulta: per $r = 1$

$$\epsilon_f \doteq \frac{x}{x-\alpha}\,\epsilon_x, \quad \epsilon_\phi \doteq \frac{x}{x-\alpha}\,\epsilon_x, \quad \epsilon_g \doteq (x-\alpha)\,\frac{2f''(\alpha)}{f'(\alpha)}\,\epsilon_x;$$

per $r > 1$

$$\epsilon_f \doteq \frac{rx}{x-\alpha}\,\epsilon_x, \quad \epsilon_\phi \doteq \frac{x}{x-\alpha}\,\epsilon_x, \quad \epsilon_g \doteq \left(1 - \frac{1}{r}\right)\epsilon_x.$$

Quindi il calcolo del metodo delle tangenti, applicato nella forma

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}\,,$$

è sempre ben condizionato, mentre il calcolo di $f(x_i)$ e $\phi(x_i)$ sono malcondizionati in un intorno di α.)

3.30 Sia α una soluzione dell'equazione $f(x) = 0$; si definisce *molteplicità* di α il numero $r > 0$ per cui esiste un intorno U di α tale che

$$m \leq \frac{|f(x)|}{|x-\alpha|^r} \leq M, \quad m,\, M > 0, \text{ per } x \in U - \{\alpha\}.$$

Si noti che esistono funzioni $f(x)$ e soluzioni α per cui la molteplicità non è definibile con questa definizione.

a) Si verifichi che se $f(x) \in C^r(U)$ per r intero, tale definizione coincide con quella data nel paragrafo 7;

b) si dica qual è la molteplicità di $\alpha = 0$ per l'equazione $f(x) = 0$ e si studi la convergenza del metodo delle tangenti per le seguenti funzioni

(1) $f(x) = x\big(1 + |x|^{\frac{1}{2}}\big)$,

(2) $f(x) = x\big(1 + |x|^{\frac{3}{2}}\big)$,

(3) $f(x) = x + |x|^{\frac{1}{2}}$,

(4) $$f(x) = \begin{cases} 0 & \text{per } x = 0 \\ x^2\big(2 + x \sin \dfrac{1}{x}\big) & \text{altrimenti,} \end{cases}$$

(5) $$f(x) = \begin{cases} 0 & \text{per } x = 0 \\ x\big(2 + x^4 \sin \dfrac{1}{x}\big) & \text{altrimenti,} \end{cases}$$

(6) $$f(x) = \begin{cases} 0 & \text{per } x = 0 \\ x^2\big(2 + x^2 \sin \dfrac{1}{x}\big) & \text{altrimenti,} \end{cases}$$

(7) $f(x) = \sqrt[5]{\arctan x - x}$

(8) $f(x) = \sqrt[3]{\dfrac{\sin^2 x}{x}}$

(9) $$f(x) = \begin{cases} 0 & \text{per } x = 0 \\ \dfrac{1}{\log |x|} & \text{altrimenti,} \end{cases}$$

(10) $$f(x) = \begin{cases} 0 & \text{per } x = 0 \\ \dfrac{x^2}{\log |x|} & \text{altrimenti.} \end{cases}$$

(Risposta: (1) la molteplicità è $r = 1$, la funzione è derivabile una sola volta per $x = 0$, il metodo è convergente con ordine $\dfrac{3}{2}$;

(2) la molteplicità è $r = 1$, è $f''(0) = 0$, ma la funzione è derivabile due sole volte per $x = 0$, il metodo è convergente con ordine $\frac{5}{2}$;
(3) la molteplicità è $r = \frac{1}{2}$, il metodo non è convergente alla soluzione $\alpha = 0$;
(4) la molteplicità è $r = 2$, la funzione è derivabile una sola volta per $x = 0$, il metodo è convergente con ordine 1 perché $|\frac{g(x)}{x}| < 1$;
(5) la molteplicità è $r = 1$, la funzione è derivabile due sole volte per $x = 0$ e $f''(0) = 0$; poiché $|\frac{f(x)}{x^4}| < 1$, il metodo è convergente con ordine 4;
(6) la molteplicità è $r = 2$, il metodo è convergente con ordine 1, malgrado la funzione $f(x)$ sia derivabile una sola volta, in quanto $\lim_{x\to 0} |\frac{g(x)}{x}| = \frac{1}{2}$;
(7) la molteplicità è $r = \frac{3}{5}$, il metodo è convergente con ordine 1, malgrado la funzione $f(x)$ non sia derivabile in 0, perché $\lim_{x\to 0} |\frac{g(x)}{x}| = \frac{2}{3}$;
(8) la molteplicità è $r = \frac{1}{3}$, il metodo non è convergente perché $\lim_{x\to 0} |\frac{g(x)}{x}| = 2$;
(9) la molteplicità non è definita, il metodo non è convergente perché $|\frac{g(x)}{x}| > 1$ in un intorno dello zero;
(10) la molteplicità non è definita, il metodo è convergente con ordine 1 perché $\lim_{x\to 0} |\frac{g(x)}{x}| = \frac{1}{2}$.)

3.31 Per il calcolo di $\sqrt{k}$, $k > 0$, si considerino i seguenti metodi e se ne studi la convergenza:

(1) metodo delle corde, applicato all'equazione $f(x) = x^2 - k$;
(2) metodo delle tangenti, applicato all'equazione $f(x) = x^2 - k$;
(3) metodo delle corde, applicato all'equazione $f(x) = x^3 - kx$;
(4) metodo di *Dedekind*, ricavato negli esercizi 3.20 e 3.25;
(5) e (6) metodi ricavati nell'esercizio 3.24;
(7) metodo ricavato nell'esercizio 3.26.

(Traccia:
(1) $x_{i+1} = x_i - \frac{x_i^2 - k}{m}$, dove m è costante. Una ragionevole scelta di m è 2α, dove α è un'approssimazione di $\sqrt{k}$; ordine del metodo 1; si scelga $0 < x_0 < 2\alpha$;

(2) $x_{i+1} = \dfrac{x_i^2 + k}{2x_i} = \dfrac{1}{2}\left(x_i + \dfrac{k}{x_i}\right)$; ordine del metodo 2; si scelga $x_0 > 0$;

(3) $x_{i+1} = \dfrac{3kx_i - x_i^3}{2k} = \dfrac{3}{2}\, x_i - \dfrac{x_i^3}{2k}$, si è scelta come costante $f'(\alpha)$; questo metodo, se è disponibile il valore $\dfrac{1}{2k}$, consente si eseguire il calcolo con sole moltiplicazioni; il metodo è stato ricavato anche nell'esercizio 3.22; ordine del metodo 2; si scelga $0 < x_0 < \sqrt{3k}$;

(4) $x_{i+1} = \dfrac{3kx_i + x_i^3}{3x_i^2 + k}$; ordine del metodo 3; si scelga $x_0 > 0$;

(5) $x_{i+1} = \dfrac{3x_i^4 + 6kx_i^2 - k^2}{8x_i^3}$; ordine del metodo 3; si scelga $x_0 > \sqrt{\dfrac{k}{5}}$;

(6) $x_{i+1} = \dfrac{5x_i^6 + 15kx_i^4 - 5k^2x_i^2 + k^3}{16x_i^5}$; ordine del metodo 4; si scelga $x_0 > 0$;

(7) $x_{i+1} = \dfrac{x_i^4 + 6kx_i^2 + k^2}{4x_i(x_i^2 + k)}$; ordine del metodo 4; si scelga $x_0 > 0$.)

3.32 Per approssimare $\sqrt{k}$, $k > 0$, si può costruire un metodo di ordine prefissato p intero nel modo seguente: si sviluppi la potenza p-esima del binomio $x - \sqrt{k}$ e si raccolgano separatamente i termini di posto dispari e quelli di posto pari, nel modo seguente

$$(x - \sqrt{k})^p = r(x,k) - \sqrt{k}\; s(x,k),$$

dove $r(x,k)$ e $s(x,k)$ sono polinomi in x e k con coefficienti tutti positivi; il metodo iterativo cercato è

$$x_{i+1} = \frac{r(x_i,k)}{s(x_i,k)}\,.$$

Si dimostri che tale metodo converge a $\sqrt{k}$ ed è effettivamente di ordine p. Si costruiscano metodi di ordine 2, 3, 4, e 5.

(Traccia: risulta

$$x_{i+1} - \sqrt{k} = \frac{r(x_i,k)}{s(x_i,k)} - \sqrt{k} = \frac{r(x_i,k) - \sqrt{k}\, s(x_i,k)}{s(x_i,k)} = \frac{(x_i - \sqrt{k})^p}{s(x_i,k)}\,,$$

da cui

$$\lim_{i\to\infty} \frac{x_{i+1} - \sqrt{k}}{(x_i - \sqrt{k})^p} = \frac{1}{s(\sqrt{k},k)}\,, \quad s(\sqrt{k},k) \neq 0.$$

Per $p = 2$ si ottiene il metodo delle tangenti, per $p = 3$ si ottiene il metodo di Dedekind, per $p = 4$ si ottiene il metodo (7) dell'esercizio 3.31, per $p = 5$ si ottiene il metodo

$$x_{i+1} = \frac{x_i^5 + 10kx_i^3 + 5k^2x_i}{5x_i^4 + 10kx_i^2 + k^2} \,.\ )$$

3.33 a) Si dimostri che ogni metodo iterativo della forma $x_{i+1} = g(x_i)$, dove $g(x) \in C[a, b]$, può essere visto come un metodo delle tangenti applicato a un'equazione $f(x) = 0$, le cui soluzioni $\alpha \in [a, b]$ sono punti fissi della funzione $g(x)$.

b) Si scrivano i metodi dell'esercizio 3.31 come metodi delle tangenti.

c) Si scriva il metodo

$$x_{i+1} = \frac{x_i(n + 1 - kx_i^n)}{n}\,, \quad k > 0, \quad n \geq 1,$$

come metodo delle tangenti e se ne studi la convergenza.

d) Si scriva il seguente metodo di *Halley*

$$x_{i+1} = x_i - \frac{2f(x_i)f'(x_i)}{2[f'(x_i)]^2 - f(x_i)f''(x_i)}$$

come metodo delle tangenti e se ne studi la convergenza ad una soluzione α di molteplicità 1 dell'equazione $f(x) = 0$.

(Traccia: a) si impone che

$$x - \frac{f(x)}{f'(x)} = g(x),$$

da cui

$$\frac{f'(x)}{f(x)} = \frac{1}{x - g(x)}\,.$$

Integrando si ha

$$\int_{x_0}^{x} \frac{f'(t)}{f(t)}\, dt = \int_{x_0}^{x} \frac{1}{t - g(t)}\, dt,$$

con x_0 costante, da cui

$$|f(x)| = |f(x_0)| \exp\Big(\int_{x_0}^{x} \frac{1}{t - g(t)}\, dt\Big).$$

Sia la costante $f(x_0)$ che il segno sono irrilevanti in quanto si deve considerare il rapporto fra $f(x)$ e $f'(x)$.

b) (1) $f(x) = \Big(\dfrac{x-\sqrt{k}}{x+\sqrt{k}}\Big)^{\frac{m}{2\sqrt{k}}}$, (3) $f(x) = 1 - \dfrac{k}{x^2}$,

(4) $f(x) = x^{\frac{3}{2}} - \dfrac{k}{x^{\frac{1}{2}}}$, (5) $f(x) = \dfrac{x^2-k}{(5x^2-k)^{\frac{1}{5}}}$,

(6) $f(x) = \dfrac{x^2-k}{11(11x^4-4kx^2+k^2)^{\frac{3}{22}}}\ \exp\Big[\dfrac{5}{11\sqrt{7}}\ \arctan\dfrac{11x^2-2k}{\sqrt{7}k}\Big]$,

(7) $f(x) = \dfrac{x^2-k}{(3x^2+k)^{\frac{1}{3}}}$.

c) $f(x) = k - \dfrac{1}{x^n}$; il metodo converge a $k^{-\frac{1}{n}}$, con ordine 2, per $0 < x_0 < \sqrt[n]{\dfrac{n+1}{k}}$.

d) il metodo di Halley è il metodo delle tangenti applicato alla funzione

$$\varphi(x) = \frac{f(x)}{\sqrt{f'(x)}},$$

come già visto nell'esercizio 3.25. Se $f(x) \in C^4[a,b]$, poiché $\varphi(\alpha) = 0$, $\varphi'(\alpha) = \sqrt{f'(\alpha)}$, $\varphi''(\alpha) = 0$, $\varphi'''(\alpha) \neq 0$ in generale, l'ordine è 3.)

3.34 Sia $f(x) \in C[a,b]$ e α soluzione di $f(x) = 0$, e sia $f(a)f(b) < 0$. Il seguente metodo per il calcolo di α, detto metodo *Illinois*, è una variante del metodo di falsa posizione:

siano x_0 e c_0 tali che $f(x_0)f(c_0) < 0$; sia $\gamma = 1$;
per $i = 0, 1, \ldots$
si calcola $x_{i+1} = x_i - \dfrac{f(x_i)(x_i - c_i)}{f(x_i) - \gamma f(c_i)}$;
se $f(x_{i+1})f(c_i) > 0$ si pone $c_{i+1} = x_i,\ \gamma = 1$,
altrimenti si pone $c_{i+1} = c_i,\ \gamma = \dfrac{1}{2}\gamma$.

Lo scopo dell'introduzione del parametro γ, quando $\gamma < 1$, è quello di determinare un punto x_{i+2} più vicino a c_{i+1} di quanto si avrebbe se fosse $\gamma = 1$. In tal modo non potrà verificarsi che da un certo indice in poi $f(x_{i+2})$ abbia sempre lo stesso segno di $f(x_{i+1})$. Quando accade che $f(x_{i+1})f(x_{i+2}) < 0$, la retta successiva viene tracciata fra i punti $(x_{i+1}, f(x_{i+1}))$ e $(x_{i+2}, f(x_{i+2}))$

come nella variante (41) del metodo delle corde, che ha una maggiore velocità di convergenza. È infatti possibile dimostrare [6], sotto ipotesi abbastanza generali, che per il metodo Illinois vale la relazione

$$\lim_{i\to\infty} \frac{x_{i+3}-\alpha}{(x_i-\alpha)^3} = \delta, \quad \delta \text{ costante non nulla.}$$

Si dimostri che il metodo Illinois è convergente sotto la sola ipotesi di continuità della funzione $f(x)$.

(Traccia: si proceda in modo analogo alla dimostrazione del teorema 3.29, notando però che per il metodo Illinois il caso a) non si può verificare.)

3.35 Siano $f(x) \in C^3[a,b]$ e α unica soluzione dell'equazione $f(x)=0$ in $[a,b]$. Nell'ipotesi che $f'(\alpha) \neq 0$, si determini l'ordine di convergenza del metodo iterativo

$$x_{i+1} = y - \frac{f(y)}{f'(x_i)}, \quad y = x_i - \frac{f(x_i)}{f'(x_i)}.$$

Si applichi al caso particolare dell'equazione $f(x) = x^2 - k = 0$, $k > 0$.

(Traccia: posto

$$g(x) = x - \frac{f(x)}{f'(x)} \quad \text{e} \quad \phi(x) = g(x) - \frac{f(g(x))}{f'(x)},$$

risulta

$$g(x) - \alpha = (x-\alpha)^2 \frac{f''(\alpha)}{2f'(\alpha)} + O((x-\alpha)^3),$$

$$\phi(x) - \alpha = g(x) - \alpha - \frac{f(g(x))}{f'(x)} = (x-\alpha)^3 \frac{[f''(\alpha)]^2}{2[f'(\alpha)]^2} + O((x-\alpha)^4),$$

e quindi il metodo è del terzo ordine. Nel caso particolare si riottiene il metodo (7) dell'esercizio 3.31.)

3.36 Sia $f(x) \in C^2[a,b]$ e α soluzione di $f(x)=0$. Il metodo iterativo

$$x_{i+1} = x_i - \frac{[f(x_i)]^2}{f[x_i + f(x_i)] - f(x_i)} \tag{95}$$

è detto metodo di *Steffensen* e richiede ad ogni iterazione due valutazioni della funzione $f(x)$.

a) Si dimostri che se α ha molteplicità 1 e se $f'(\alpha) \neq -1$ e $f''(\alpha) \neq 0$, il metodo è del secondo ordine e che se $f(x) \in C^3[a,b]$ e $f'(\alpha) = -1$ o $f''(\alpha) = 0$, il metodo è del terzo ordine.

b) Si verifichi che il metodo di Aitken è equivalente al metodo di Steffensen applicato all'equazione $f(z) = g(z) - z$.

(Traccia: a) si proceda come nella dimostrazione del teorema 3.20, caso a): posto $\beta_i = f(x_i)$, è

$$f(x_i + \beta_i) = f(x_i) + \beta_i f'(x_i) + \frac{\beta_i^2}{2}\, f''(\xi), \quad \text{con} \quad |\xi - x_i| < |\beta_i|,$$

per cui

$$f[x_i + f(x_i)] - f(x_i) = f(x_i)\big[f'(x_i) + \frac{f(x_i)}{2}\, f''(\xi)\big],$$

e sostituendo nella (95) si ha

$$\big[f'(x_i) + \frac{f(x_i)}{2}\, f''(\xi)\big](x_{i+1} - \alpha) = \big[f'(x_i) + \frac{f(x_i)}{2}\, f''(\xi)\big](x_i - \alpha) - f(x_i).$$

Dalla formula di Taylor si ha

$$0 = f(\alpha) = f(x_i) + (\alpha - x_i) f'(x_i) + \frac{1}{2}(\alpha - x_i)^2 f''(\xi_1), \quad \text{con} \quad |\xi_1 - x_i| < |x_i - \alpha|,$$

da cui risulta

$$\begin{aligned}\big[f'(x_i) + \frac{f(x_i)}{2}\, f''(\xi)\big](x_{i+1} - \alpha) &= \frac{1}{2}(\alpha - x_i)^2 f''(\xi_1) \\ + \frac{f(x_i)}{2}\, f''(\xi)(x_i - \alpha) &= \frac{1}{2}(\alpha - x_i)^2 \big[f''(\xi_1) + f'(\xi_2) f''(\xi)\big], \\ &\text{con} \quad |\xi_2 - \alpha| < |x_i - \alpha|,\end{aligned}$$

e quindi

$$\lim_{i \to \infty} \frac{x_{i+1} - \alpha}{(x_i - \alpha)^2} = \frac{f''(\alpha)(1 + f'(\alpha))}{2 f'(\alpha)}\,. \;)$$

3.37 Sia $p \geq 2$ intero e $\rho > 0$. Si dimostri che

a) se $g(x) \in C^{2p}[\alpha - \rho, \alpha + \rho]$ e se α è soluzione di molteplicità p dell'equazione $x = g(x)$, il metodo di Aitken è convergente con ordine di convergenza 1;

b) se $g(x) \in C^p[\alpha - \rho, \alpha + \rho]$ e se il metodo $x_{i+1} = g(x_i)$ è convergente con ordine di convergenza p, il metodo di Aitken è convergente con ordine di convergenza $2p - 1$.

(Traccia: si proceda in modo analogo a quanto fatto nella dimostrazione del teorema 3.34. a) Poiché

$$g(\alpha) = \alpha, \quad g'(\alpha) = 1, \quad g^{(r)}(\alpha) = 0, \text{ per } r = 2, \ldots, p-1, \quad g^{(p)}(\alpha) \neq 0,$$

si pone

$$g(\alpha + \epsilon) = \alpha + \epsilon + \sigma\epsilon^p = \alpha + \delta, \quad \text{e} \quad g(\alpha + \delta) = \alpha + \delta + \tau\delta^p,$$

dove

$$\sigma = \sum_{i=0}^{p-1} \gamma_i \epsilon^i + O(\epsilon^p), \quad \tau = \sum_{i=0}^{p-1} \gamma_i \delta^i + O(\delta^p), \quad \gamma_i = \frac{g^{(p+i)}(\alpha)}{(p+i)!},$$

si verifichi che

$$G(\alpha + \epsilon) = \alpha + \epsilon - \frac{\sigma^2 \epsilon^p}{(\tau - \sigma) + p\tau\sigma\epsilon^{p-1} + O(\epsilon^{2p-2})},$$

dove

$$\tau - \sigma = \sum_{i=0}^{p-1} \gamma_i \left[(\epsilon + \sigma\epsilon^p)^i - \epsilon^i \right] + O(\epsilon^p) = O(\epsilon^p),$$

e quindi

$$\frac{G(\alpha + \epsilon) - \alpha}{\epsilon} = 1 - q(\epsilon), \quad \text{dove } \lim_{\epsilon \to 0} q(\epsilon) = \frac{1}{p}.$$

b) Poiché

$$g(\alpha) = \alpha, \quad g^{(r)}(\alpha) = 0, \text{ per } r = 1, \ldots, p-1, \quad g^{(p)}(\alpha) \neq 0,$$

posto

$$g(\alpha + \epsilon) = \alpha + \delta = \alpha + \sigma\epsilon^p, \quad \text{e} \quad g(\alpha + \delta) = \alpha + \tau\delta^p,$$

dove

$$\sigma = \frac{g^{(p)}(\xi)}{p!}, \quad \text{e} \quad \tau = \frac{g^{(p)}(\xi')}{p!},$$

si verifichi che

$$G(\alpha + \epsilon) = \alpha + \epsilon^{2p-1} \frac{-\sigma^2 + \epsilon^{(p-1)^2} \sigma^p \tau}{1 - 2\epsilon^{p-1}\sigma + \epsilon^{p^2-1}\sigma^p\tau},$$

e quindi

$$\frac{G(\alpha + \epsilon) - \alpha}{\epsilon^{2p-1}} = -q(\epsilon), \quad \text{dove } \lim_{\epsilon \to 0} q(\epsilon) = \left[\frac{g^{(p)}(\alpha)}{p!} \right]^2. \;)$$

3.38 Il seguente metodo di *Dekker-Brent* combina i vantaggi del metodo di bisezione, del metodo delle secanti e del metodo ottenuto con l'interpolazione quadratica inversa (si veda il paragrafo 8 del capitolo 5). Sia $\alpha \in [a, b]$ una soluzione di $f(x) = 0$, con $f(a)f(b) < 0$. Al primo passo il metodo utilizza i due punti $(x_0, f(x_0))$ e $(x_1, f(x_1))$, $x_0 = a$, $x_1 = b$, per generare un terzo punto $x_2 \in [a, b]$ con il metodo delle secanti. All'i-esimo passo, $i \geq 2$, il metodo utilizza i valori nei tre punti x_{i-2}, x_{i-1}, x_i per generare con la (33, cap. 5) un punto x_i che viene sostituito a uno dei tre utilizzati, in modo che il nuovo intervallo contenga ancora la soluzione. Si trascriva il metodo e si dica quali controlli devono essere effettuati perché il metodo sia efficace e in quale situazione è invece più conveniente utilizzare il metodo di bisezione e il metodo delle secanti.

(Traccia: indicati con

$$f_i = f(x_i), \quad r_i = \frac{f_{i-1}}{f_i}, \quad s_i = \frac{f_{i-1}}{f_{i-2}}, \quad t_i = \frac{f_{i-2}}{f_i},$$

$p_i = s_i\big[t_i(r_i - t_i)(x_i - x_{i-1}) - (1 - r_i)(x_{i-1} - x_{i-2})\big]$, $q_i = (r_i - 1)(s_i - 1)(t_i - 1)$,

si pone

$$x_2 = x_1 - \frac{(x_1 - x_0)f_1}{f_1 - f_0}, \quad x_{i+1} = x_{i-1} + \frac{p_i}{q_i}\,.$$

I controlli devono accertare che la funzione $f(x)$ sia invertibile nell'intervallo contenente i tre punti x_{i-2}, x_{i-1} e x_i, e quindi monotona, e che il denominatore q_i non sia di modulo troppo piccolo, allo scopo di garantire che il punto x_{i+1} appartenga all'intervallo. Se la funzione non è monotona, il punto successivo viene calcolato con il metodo delle secanti, mentre se q_i è in modulo troppo piccolo, cioè $|p_i| \geq \frac{1}{2}\,|q_i|e_i$, dove e_i è l'ampiezza dell'intervallo individuato dai punti x_{i-2}, x_{i-1} e x_i, allora si utilizza la bisezione.)

3.39 Si dimostri, sfruttando il metodo delle tangenti, che le operazioni di moltiplicazione, divisione ed estrazione di radice quadrata hanno la stessa complessità computazionale in termini di operazioni sulle singole cifre, cioè indicato con $M(n)$ il numero di operazioni sulle singole cifre richiesto per moltiplicare due numeri di n cifre, si dimostri che il numero delle operazioni richiesto per calcolare n cifre del quoziente di due numeri di n cifre o della radice quadrata di un numero di n cifre è dato da $O(M(n))$. Si sfruttino le seguenti ipotesi di regolarità

(1) $M(n)$ è una funzione crescente con n,

(2) $2M(n) \leq M(2n) \leq 4M(n)$.

Si supponga inoltre che n sia potenza di 2, cioè $n = 2^m$, m intero.

(Traccia: si consideri prima il calcolo di $\dfrac{1}{k}$, dove $k \neq 0$ è un numero di n cifre. Il metodo delle tangenti applicato all'equazione $\dfrac{1}{x} - k = 0$ è dato da $x_{i+1} = (2 - kx_i)x_i$. Supponendo che x_0 sia tale che $\left|x_0 - \dfrac{1}{k}\right| < 10^{-1}$, dopo m iterazioni si ha $\left|x_m - \dfrac{1}{k}\right| < 10^{-2^m} = 10^{-n}$. Se i calcoli alla i-esima iterazione si eseguono con numeri di 2^i cifre, ottenendo un risultato di 2^{i+1} cifre, la i-esima iterazione richiede 2 moltiplicazioni fra numeri di 2^i cifre (la sottrazione richiede un numero di operazioni fra cifre che è trascurabile). Quindi per m iterazioni si ha

$$2\sum_{i=1}^{m} M(2^i) \leq 2M(2^{m+1}) \leq 8M(n),$$

perché

$$M(2) < M(2^2) \quad \text{e} \quad 2M(2^i) \leq M(2^{i+1}).$$

Poiché $\dfrac{y}{k} = y\,\dfrac{1}{k}$, la complessità della divisione fra y e k è di $9M(n)$. Si proceda in modo analogo per la radice quadrata, sfruttando il metodo delle tangenti applicato all'equazione $x^2 - k = 0$, o un qualunque altro metodo che garantisca convergenza almeno quadratica.)

3.40 Si approssimi la soluzione $\boldsymbol{\alpha}$, appartenente al primo quadrante, del sistema

$$\begin{cases} x_1 = \sin(x_1 + x_2) \\ \\ x_2 = \cos(x_1 - x_2). \end{cases}$$

(Traccia: $\boldsymbol{\alpha} \in [0.6, 1] \times [0.6, 1]$. Si verifichi che per $\mathbf{x}$ in tale intervallo è $\|H(\mathbf{x})\|_\infty < 1$.)

3.41 Si dica se è convergente il metodo iterativo

$$\begin{cases} x_1^{(i+1)} = (x_1^{(i)})^2 - x_2(i) \\ \\ x_2^{(i+1)} = \dfrac{1}{2}\left[x_1^{(i)} - (x_1^{(i)})^2\right]. \end{cases}$$

(Traccia: il sistema

$$\begin{cases} x_1 = x_1^2 - x_2 \\ \\ x_2 = \dfrac{1}{2}\,(x_1 - x_1^2) \end{cases}$$

ha le due soluzioni $\boldsymbol{\alpha} = (0,0)$ e $\boldsymbol{\beta} = (1,0)$. Risulta $\|H(\boldsymbol{\alpha})\|_\infty = 1$, ma $\rho(H(\boldsymbol{\alpha})) < 1$, quindi il metodo converge ad $\boldsymbol{\alpha}$, mentre $\|H(\boldsymbol{\beta})\|_\infty = 3$, e $\rho(H(\boldsymbol{\beta})) > 1$.)

3.42 Si studi la convergenza del metodo di Newton-Raphson alle soluzioni del sistema

$$\begin{cases} 4x_2^2 - \log(x_1+1) = 0 \\ \\ x_1^2 - 2x_2 \cos^2 x_1 + x_2^2 = 0. \end{cases}$$

(Traccia: il sistema ha la soluzione $\boldsymbol{\alpha} = (0,0)$ e una soluzione $\boldsymbol{\beta} \in D = [0.4, 0.73] \times [0.3, 0.5]$. Si verifichi che in D il sistema soddisfa alle condizioni sufficienti del teorema 3.47, mentre in un intorno di $\boldsymbol{\alpha}$ le condizioni non sono soddisfatte.)

3.43 Sia $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, dove $\mathbf{f} : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^n$, un sistema di n equazioni in n incognite. Si consideri il metodo iterativo

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - C^{-1}\mathbf{f}(\mathbf{x}^{(i)}),$$

dove $C \in \mathbf{R}^{n\times n}$ è una matrice non singolare. Tale metodo è una generalizzazione del metodo delle corde. Si verifichi che il metodo, se convergente, converge ad una soluzione del sistema e si diano delle condizioni sufficienti di convergenza, nell'ipotesi che $\mathbf{f}(\mathbf{x}) \in C^1(\Omega)$. Si indichi come può essere costruita la matrice C e si consideri anche il caso particolare in cui C sia una matrice diagonale. Si applichi al sistema dell'esempio 3.42.

(Traccia: il sistema $\mathbf{x} = \mathbf{x} - C^{-1}\mathbf{f}(\mathbf{x})$ è equivalente al sistema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Condizione sufficiente di convergenza è che

$$\|I - C^{-1}J(\mathbf{x})\|_\infty < 1 \quad \text{o che} \quad \rho(I - C^{-1}J(\mathbf{x})) < 1$$

per ogni $\mathbf{x} \in \Omega$, o in un intorno della soluzione. La matrice C può essere scelta come $J(\mathbf{y})$, dove $\mathbf{y}$ è un punto opportuno, sufficientemente vicino alla soluzione. La scelta di una matrice C diagonale semplifica il calcolo di C^{-1}. Nel caso particolare è

$$J(\mathbf{x}) = \begin{bmatrix} 1 - \dfrac{x_1}{2} & -\dfrac{x_2}{2} \\ -\cos(x_1+1) & 1 \end{bmatrix}, \quad \text{e posto} \quad C = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix},$$

è

$$I - C^{-1}J(\mathbf{x}) = \begin{bmatrix} 1 - \dfrac{2-x_1}{2c_1} & \dfrac{x_2}{2c_1} \\ \dfrac{\cos(x_1+1)}{c_2} & 1 - \dfrac{1}{c_2} \end{bmatrix}.$$

Assumendo $c_1 = c_2 = 1$, cioè $C = I$, si ottiene il metodo iterativo dell'esempio 3.38, che converge alla soluzione $\boldsymbol{\alpha}$. Assumendo $c_1 = -2$, $c_2 = 1$, si ottiene

$$\|I - C^{-1}J(\mathbf{x})\|_\infty = \max\left\{\left|\frac{2c_1 - 2 + x_1}{2c_1}\right| + \left|\frac{x_2}{2c_1}\right|, \left|\frac{\cos(x_1+1)}{c_2}\right| + \left|\frac{c_2 - 1}{c_2}\right|\right\}$$
$$= \max\left\{\frac{|6 - x_1|}{4} + \frac{|x_2|}{4}, |\cos(x_1+1)|\right\}$$

e poiché $\boldsymbol{\beta} \in [3.5, 4] \times [-1, -0.5]$, scegliendo un opportuno punto iniziale in tale intervallo si ha convergenza a $\boldsymbol{\beta}$.)

3.44 Sia D un sottoinsieme convesso di $\mathbf{R}^n$ e sia $f : D \to \mathbf{R}$, $f \in C^2(D)$. Si dimostri che f è convessa se e solo se la matrice hessiana $S(\mathbf{x})$ di f è semidefinita positiva per ogni $\mathbf{x} \in D$.

(Traccia: siano $\mathbf{x}, \mathbf{y} \in D$ e $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$, con $0 \le \lambda \le 1$. Per la formula di Taylor si ha

$$f(\mathbf{x}) = f(\mathbf{z}) + \mathbf{h}(\mathbf{z})^T(\mathbf{x} - \mathbf{z}) + \frac{1}{2}(\mathbf{x} - \mathbf{z})^T S(\boldsymbol{\xi}_1)(\mathbf{x} - \mathbf{z}),$$
$$f(\mathbf{y}) = f(\mathbf{z}) + \mathbf{h}(\mathbf{z})^T(\mathbf{y} - \mathbf{z}) + \frac{1}{2}(\mathbf{y} - \mathbf{z})^T S(\boldsymbol{\xi}_2)(\mathbf{y} - \mathbf{z}),$$

dove $\mathbf{h}(\mathbf{z})$ è il vettore la cui i-esima componente è $\dfrac{\partial f(\mathbf{z})}{\partial x_i}$ e $\boldsymbol{\xi}_1$ e $\boldsymbol{\xi}_2$ sono due punti appartenenti ai segmenti da $\mathbf{x}$ a $\mathbf{z}$ e da $\mathbf{y}$ a $\mathbf{z}$, rispettivamente. Poiché

$$\lambda(\mathbf{x} - \mathbf{z}) + (1-\lambda)(\mathbf{y} - \mathbf{z}) = \mathbf{0},$$

risulta

$$\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) = f(\mathbf{z}) + \frac{1}{2}\big[\lambda(\mathbf{x} - \mathbf{z})^T S(\boldsymbol{\xi}_1)(\mathbf{x} - \mathbf{z}) + (1-\lambda)(\mathbf{y} - \mathbf{z})^T S(\boldsymbol{\xi}_2)(\mathbf{y} - \mathbf{z})\big].$$

Se $S(\mathbf{x})$ è semidefinita positiva per $\mathbf{x} \in D$, ne segue che

$$(\mathbf{x} - \mathbf{z})^T S(\boldsymbol{\xi}_1)(\mathbf{x} - \mathbf{z}) \ge 0, \quad (\mathbf{y} - \mathbf{z})^T S(\boldsymbol{\xi}_2)(\mathbf{y} - \mathbf{z}) \ge 0,$$

e quindi

$$\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \ge f(\mathbf{z}) \quad \text{per} \quad 0 \le \lambda \le 1.$$

Viceversa, se in un punto $\mathbf{u} \in D$ la matrice $S(\mathbf{u})$ non fosse semidefinita positiva, esisterebbe un vettore $\mathbf{w}$ tale che $\mathbf{w}^T S(\mathbf{u})\mathbf{w} < 0$ ed un intorno convesso U di $\mathbf{u}$ tale che $\mathbf{w}^T S(\mathbf{v})\mathbf{w} < 0$ per ogni $\mathbf{v} \in U$. Si scelgono allora $\mathbf{x}$ e $\mathbf{y} \in U$ tali che $\mathbf{x} - \mathbf{y} = \gamma\mathbf{w}$, $\gamma \ne 0$. Si verifichi che in tal caso risulta

$$\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) < f(\mathbf{z}).\ )$$

3.45 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R},$$

un polinomio di grado n.

a) Si descriva una generalizzazione del metodo di Ruffini-Horner per calcolare i coefficienti del polinomio quoziente e il resto della divisione di $p(x)$ per un fattore quadratico

$$p(x) = q(x)(x^2 + bx + c) + rx + s,$$

e si dica quante operazioni aritmetiche sono richieste.

b) Si dica quante operazioni aritmetiche reali sono richieste dal metodo di Ruffini-Horner per calcolare il valore di $p(x)$ in un punto complesso $u + \boldsymbol{i}\, v$, tenendo conto del fatto che il prodotto di due numeri complessi richiede 4 moltiplicazioni e 2 addizioni reali.

c) Per calcolare il valore di $p(x)$ nel punto $u + \boldsymbol{i}\, v$ si può anche procedere nel modo seguente: posto $b = -2u$ e $c = u^2 + v^2$, si calcolano come al punto a) il quoziente e il resto della divisione di $p(x)$ per

$$x^2 + bx + c = [x - (u + \boldsymbol{i}\, v)]\, [x - (u - \boldsymbol{i}\, v)].$$

Risulta quindi

$$p(x) = q(x)[x - (u + \boldsymbol{i}\, v)]\, [x - (u - \boldsymbol{i}\, v)] + rx + s$$

e

$$p(u + \boldsymbol{i}\, v) = r(u + \boldsymbol{i}\, v) + s.$$

Il vantaggio di questo procedimento consiste, oltre che in un minor numero di operazioni aritmetiche, nel fatto che è richiesto un solo prodotto per numero complesso alla fine. Si dica quante operazioni aritmetiche reali sono richieste.

(Traccia: a) posto

$$q(x) = \sum_{i=0}^{n-2} q_i x^i,$$

si sviluppi il prodotto e si uguaglino i coefficienti. Si ottiene

$$\begin{aligned}
&q_{n-2} = a_n, \qquad q_{n-3} = a_{n-1} - q_{n-2} b, \\
&q_i = a_{i+2} - q_{i+2} c - q_{i+1} b, \quad i = n-4, \ldots, 0, \\
&r = a_1 - q_1 c - q_0 b, \qquad s = a_0 - q_0 c.
\end{aligned}$$

A meno di termini di ordine inferiore, sono richieste $2n$ moltiplicazioni e $2n$ addizioni per a) e per c), $4n$ moltiplicazioni e $3n$ addizioni per b).)

3.46 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{C},$$

un polinomio di grado n a coefficienti complessi e siano $\alpha_1, \alpha_2, \ldots, \alpha_n$ i suoi zeri. Si consideri il polinomio a coefficienti coniugati

$$\overline{p}(x) = \sum_{i=0}^{n} \overline{a}_i x^i.$$

Si verifichi che il polinomio $p(x)\ \overline{p}(x)$, che ha grado $2n$ e si annulla negli stessi zeri di $p(x)$, ha coefficienti reali.

(Traccia: si verifichi che $\overline{p}(x)$ ha per zeri $\overline{\alpha}_1, \overline{\alpha}_2, \ldots, \overline{\alpha}_n$ e che $p(x)\ \overline{p}(x)$ ha per zeri i numeri coniugati $\alpha_1,\ \overline{\alpha}_1, \alpha_2,\ \overline{\alpha}_2, \ldots, \alpha_n,\ \overline{\alpha}_n$.)

3.47 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R},$$

un polinomio di grado n. Fissato un punto ξ si costruisca con un algoritmo delle divisioni successive la successione $\{q_i(x)\}$ di polinomi nel modo seguente:

$$\begin{aligned} &q_0(x) = p(x) \\ &q_i(x) = q_{i+1}(x)(x - \xi) + r_i, \quad i = 0, \ldots, n-1, \\ &r_n = q_n(x). \end{aligned}$$

Si verifichi che $r_0 = p(\xi)$ e $r_i = \dfrac{p^{(i)}(\xi)}{i!}$, per $i = 1, \ldots, n$.

(Traccia: risulta

$$p(x) = \sum_{i=0}^{n} r_i (x - \xi)^i;$$

d'altra parte, poiché $p(x)$ è un polinomio di grado n, per la formula di Taylor è

$$p(x) = p(\xi) + \sum_{i=1}^{n} \frac{p^{(i)}(\xi)}{i!}\ (x - \xi)^i.\ )$$

3.48 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R}, \quad a_n \neq 0,$$

un polinomio di grado n e siano $\alpha_1, \alpha_2, \ldots, \alpha_n$ i suoi zeri. Si dimostrino le seguenti relazioni (*funzioni simmetriche elementari* degli zeri)

$$s_1(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sum_{i=1}^{n} \alpha_i = -\frac{a_{n-1}}{a_n},$$

$$s_2(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sum_{i,j=1}^{n} \alpha_i \alpha_j = \frac{a_{n-2}}{a_n},$$

$$\ldots$$

$$s_j(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} \alpha_{i_1} \alpha_{i_2} \ldots \alpha_{i_j} = (-1)^j \frac{a_{n-j}}{a_n},$$

$$\ldots$$

$$s_n(\alpha_1, \alpha_2, \ldots, \alpha_n) = \prod_{i=1}^{n} \alpha_i = (-1)^n \frac{a_0}{a_n}.$$

(Traccia: si sfrutti la relazione $p(x) = a_n(x-\alpha_1)(x-\alpha_2)\ldots(x-\alpha_n)$.)

3.49 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R},$$

un polinomio di grado n, siano $\alpha_1, \alpha_2, \ldots, \alpha_n$ i suoi zeri e sia

$$S_j = \sum_{i=1}^{n} \alpha_i^j, \quad j = 1, 2, \ldots$$

Si dimostrino le seguenti relazioni (di *Girard-Newton*):

a) $$\sum_{j=0}^{i-1} a_{n-j} S_{i-j} = -i a_{n-i}, \quad \text{per} \quad i = 1, \ldots, n-1,$$

b) $$\sum_{j=0}^{n} a_{n-j} S_{i-j} = 0, \quad \text{per} \quad i \ge n.$$

(Traccia: a) per $x \neq \alpha_i$ si ha

$$p(x) = (x-\alpha_i)\big[a_n x^{n-1} + (a_{n-1} + a_n\alpha_i)x^{n-2} + \ldots + (a_1 + a_2\alpha_i + \ldots + a_{n-1}\alpha_i^{n-2} + a_n\alpha_i^{n-1})\big]$$

e quindi

$$\sum_{i=1}^{n} \frac{p(x)}{x-\alpha_i} = n a_n x^{n-1} + (n a_{n-1} + a_n S_1)x^{n-2} + \ldots + (n a_1 + a_2 S_1 + \ldots + a_{n-1} S_{n-2} + a_n S_{n-1}).$$

D'altra parte, poiché $p(x) = a_n(x-\alpha_1)\dots(x-\alpha_n)$, è

$$\sum_{i=1}^{n} \frac{p(x)}{x-\alpha_i} = p'(x) = na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \dots + a_1.$$

Dal confronto fra le due espressioni seguono le relazioni a).
b) Dalla relazione

$$\sum_{j=0}^{n} a_{n-j}\alpha_k^{n-j} = 0, \quad \text{per} \quad k = 1, \dots, n,$$

segue che

$$\sum_{j=0}^{n} a_{n-j}\alpha_k^{i-j} = 0, \quad \text{per} \quad i \geq n,$$

e quindi, sommando rispetto a k, seguono le relazioni b).)

3.50 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R},$$

un polinomio di grado n e siano $\alpha_1, \alpha_2, \dots, \alpha_n$ i suoi zeri. Si costruiscano i polinomi

a) $p_1(x)$, trasformato a *radici opposte*, che ha per zeri $-\alpha_1, -\alpha_2, \dots, -\alpha_n$;
b) $p_2(x)$, trasformato a *radici traslate*, che ha per zeri $\alpha_1 + h,\ \alpha_2 + h,\ \dots,$ $\alpha_n + h$, dove h è una costante;
c) $p_3(x)$, trasformato a *radici reciproche*, che ha per zeri $\dfrac{1}{\alpha_1}, \dfrac{1}{\alpha_2}, \dots, \dfrac{1}{\alpha_n}$, $\alpha_j \neq 0$, per $j = 1, \dots, n$;
d) $p_4(x)$, che ha per zeri $\alpha_1^2, \alpha_2^2, \dots, \alpha_n^2$.

(Traccia:

a) $$p_1(x) = p(-x) = \sum_{i=0}^{n} (-1)^i a_i x^i,$$

i coefficienti di $p_1(x)$ si ottengono da quelli di $p(x)$ cambiandoli alternativamente di segno;

b) $$p_2(x) = p(x-h) = \sum_{i=0}^{n} a_i(x-h)^i = \sum_{i=0}^{n} b_i x^i,$$

si verifichi che i coefficienti b_i si possono ricavare come resti delle divisioni successive di $p(x)$ per $x+h$ come nell'esercizio 3.47, in cui si ponga $\xi = -h$;

c) $$p_3(x) = x^n p\Big(\frac{1}{x}\Big) = \sum_{i=0}^{n} a_{n-i}\, x^i,$$

i coefficienti di $p_3(x)$ si ottengono da quelli di $p(x)$ invertendone l'ordine;

d) $$p_4(x) = p(\sqrt{x})p(-\sqrt{x}) = \sum_{i=0}^{n} b_i x^i, \quad b_i = \sum_{j=\max(0,2i-n)}^{\min(2i,n)} (-1)^j a_j a_{2i-j},$$

l'i-esimo coefficiente b_i di $p_4(x)$ si ottiene moltiplicando a coppie i coefficienti di $p(x)$ di indici tali che la loro somma sia uguale a $2i$ e sommando a segni alterni. Per il calcolo effettivo del polinomio prodotto ci si riferisca all'esercizio 1.6)

3.51 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{Z},$$

un polinomio di grado n a coefficienti interi.

a) Si verifichi che ogni polinomio a coefficienti razionali può essere scritto in forma equivalente come polinomio a coefficienti interi.

b) Si dimostri che se $\dfrac{\alpha}{\beta}$ è uno zero razionale di $p(x)$, con $\alpha,\ \beta \in \mathbf{Z}$, $\alpha,\ \beta$ primi fra loro, allora α divide a_0 e β divide a_n.

c) Si dimostri che se $\dfrac{\alpha}{\beta}$ è uno zero razionale di $p(x)$, con $\alpha,\ \beta \in \mathbf{Z}$, $\alpha,\ \beta$ primi fra loro, allora tutti i coefficienti del polinomio quoziente $q(x)$ di $p(x)$ per $x - \dfrac{\alpha}{\beta}$

$$q(x) = p(x) \Big/ \Big(x - \frac{\alpha}{\beta}\Big)$$

sono interi.

d) Sia α un intero tale che $p(\alpha) \neq 0$. Si dimostri che l'intero β può essere zero di $p(x)$ solo se $\beta - \alpha$ divide $p(\alpha)$.

e) Si determinino le radici razionali dell'equazione

$$p(x) = 2\,x^5 - 21\,x^4 + 38\,x^3 + 87\,x^2 + 36\,x + 108 = 0.$$

(Traccia: a) sia

$$q(x) = \sum_{i=0}^{n} \frac{b_i}{c_i}\, x^i, \quad b_i,\ c_i \in \mathbf{Z},$$

e sia

$$c = \text{m.c.m. } \{c_i,\ i = 0, \dots, n\};$$

si consideri il polinomio $p(x) = cq(x)$.
b) È

$$\beta^n p\Big(\frac{\alpha}{\beta}\Big) = \beta^n \sum_{i=0}^{n} a_i \frac{\alpha^i}{\beta^i} = \sum_{i=0}^{n} a_i \alpha^i \beta^{n-i} = a_n \alpha^n + \beta\Big(\sum_{i=0}^{n-1} a_i \alpha^i \beta^{n-i-1}\Big).$$

Poiché $p\Big(\frac{\alpha}{\beta}\Big) = 0$, e β è primo con α, ne segue che β deve dividere a_n. In modo analogo si dimostri che α divide a_0.
c) Analogamente a quanto fatto al punto b), si ha per $0 < j < n$

$$\alpha^j \Big(\sum_{i=j}^{n} a_i \alpha^{i-j} \beta^{n-i}\Big) = -\beta^{n-j+1}\Big(\sum_{i=0}^{j-1} a_i \alpha^i \beta^{j-i-1}\Big).$$

Poiché

$$\sum_{i=j}^{n} a_i \alpha^{i-j} \beta^{n-i} \quad \text{e} \quad \sum_{i=0}^{j-1} a_i \alpha^i \beta^{j-i-1}$$

sono numeri interi e β^{n-j}, per $n - j > 0$, non divide α^j, ne segue che β^{n-j} deve dividere

$$\sum_{i=j}^{n} a_i \alpha^{i-j} \beta^{n-i},$$

cioè il numero

$$\sum_{i=j}^{n} a_i \frac{\alpha^{i-j}}{\beta^{i-j}}, \quad \text{per } j = 1, \dots, n-1,$$

che è il coefficiente di x^{j-1} del polinomio $q(x)$, deve essere intero. Inoltre il primo coefficiente di $q(x)$ è uguale al primo coefficiente di $p(x)$ e quindi è intero per ipotesi.
d) È $p(x) = (x - \alpha)q(x) + p(\alpha)$. Se $p(\beta) = 0$, risulta $(\beta - \alpha)q(\beta) = -p(\alpha)$, dove $p(\alpha)$ e $q(\beta)$ sono interi.
e) Considerando i fattori di 108 e quelli di 2, le eventuali radici razionali vanno cercate fra

$$\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 9, \pm 12, \pm 18, \pm 27, \pm 36, \pm 54, \pm 108, \pm\frac{1}{2}, \pm\frac{3}{2}, \pm\frac{9}{2}, \pm\frac{27}{2}.$$

Inoltre $p(1) = 250$, per cui vanno esclusi gli interi $-2, -3, 4, -6, 9, \pm 12, \pm 18, \pm 27, \pm 36, \pm 54, \pm 108$, le cui differenze con 1 non dividono 250. Si ha poi

$p(-1) = 98$, per cui vanno esclusi gli interi $2, 3, -4, -9$, le cui differenze con -1 non dividono 98. L'unico intero rimasto nell'insieme è 6, che infatti è radice. Risulta

$$p(x) = (x-6)q(x) = (x-6)(2x^4 - 9x^3 - 16x^2 - 9x - 18).$$

Si ripete il ragionamento per $q(x)$. Tenendo conto dei risultati precedenti, risulta che le radici razionali di $q(x)$ vanno ricercate fra

$$6, \pm\frac{1}{2}, \pm\frac{3}{2}, \pm\frac{9}{2}.$$

In conclusione l'equazione data ammette come radici razionali 6 (radice doppia) e $-\dfrac{3}{2}$.)

3.52 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R},$$

un polinomio di grado n.

a) Si dimostri che se $a_i \geq 0$ per ogni i, allora il polinomio non ha zeri reali positivi;

b) indicati con $q(x)$ ed r il quoziente e il resto della divisione di $p(x)$ per $x - \beta$, $\beta > 0$:

$$p(x) = q(x)(x - \beta) + r,$$

si dimostri che se i coefficienti di $q(x)$ ed r sono non negativi, allora non esistono zeri reali di $p(x)$ maggiori di β;

c) sia $a_n > 0$, indicati con

$$j = \max\{i \ : \ a_i < 0\}, \quad \gamma = \max\{|a_i| \ : \ a_i < 0\},$$

si dimostri che non esistono zeri reali di $p(x)$ maggiori di

$$1 + \left(\frac{\gamma}{a_n}\right)^{\frac{1}{n-j}}.$$

d) Per l'equazione del punto e) dell'esercizio 3.51 si determinino i numeri γ_1, γ_2, γ_3, γ_4, tali che le radici reali negative appartengano all'intervallo (γ_1, γ_2) e le radici reali positive appartengano all'intervallo (γ_3, γ_4).

(Traccia: c) sia α il massimo zero reale di $p(x)$ e si supponga che $\alpha > 1$. Si ha

$$0 = p(\alpha) = a_n\ \alpha^n + \sum_{i=0}^{j} a_i \alpha^i + \sum_{i=j+1}^{n} a_i \alpha^i,$$

da cui

$$a_n\ \alpha^n = -\sum_{i=0}^{j} a_i \alpha^i - \sum_{i=j+1}^{n} a_i \alpha^i \leq -\sum_{i=0}^{j} a_i \alpha^i \leq \gamma \sum_{i=0}^{j} \alpha^i = \gamma\ \frac{\alpha^{j+1} - 1}{\alpha - 1}\,,$$

e quindi

$$\alpha^n \leq \frac{\gamma}{a_n}\ \frac{\alpha^{j+1} - 1}{\alpha - 1} < \frac{\gamma}{a_n}\ \frac{\alpha^{j+1}}{\alpha - 1}\,.$$

D'altra parte

$$(\alpha - 1)^{n-j} = (\alpha - 1)(\alpha - 1)^{n-j-1} < (\alpha - 1)\alpha^{n-j-1} < \frac{\gamma}{a_n}\,.$$

d) Per il punto b) è $\gamma_4 = 10.5$ (per il punto c) sarebbe $\gamma_4 = 1 + \dfrac{21}{2} = 11.5$); il polinomio trasformato a radici opposte è

$$p_1(x) = 2\,x^5 + 21\,x^4 + 38\,x^3 - 87\,x^2 + 36\,x - 108$$

e per il punto b) è $\gamma_1 = -2$ (per il punto c) sarebbe $\gamma_1 = -(1 + \sqrt[3]{54}) \approx -4.78$); il polinomio trasformato a radici reciproche è

$$p_2(x) = 108\,x^5 + 36\,x^4 + 87\,x^3 + 38\,x^2 - 21\,x + 2$$

e per il punto b) è $\gamma_3 = 3$ (per il punto c) sarebbe $\gamma_3 = 1\Big/(1 + \sqrt[4]{\dfrac{21}{108}}) \approx 0.601$); il polinomio trasformato a radici reciproche e opposte è

$$p_3(x) = 108\,x^5 - 36\,x^4 + 87\,x^3 - 38\,x^2 - 21\,x - 2$$

e per il punto b) è $\gamma_2 = -1$ (per il punto c) sarebbe $\gamma_2 = -\ \dfrac{54}{73} \approx -0.74.$)

3.53 Sia

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad a_i \in \mathbf{R},$$

un polinomio di grado n e sia γ l'unica radice reale positiva dell'equazione

$$|a_n| x^n - \sum_{i=0}^{n-1} |a_i| x^i = 0. \tag{96}$$

a) Si dimostri che tutti gli zeri del polinomio $p(x)$ appartengono al cerchio del piano complesso di centro l'origine e raggio γ.

b) Si determini una limitazione inferiore e superiore dei moduli delle radici dell'equazione del punto e) dell'esercizio 3.51.

(Traccia: a) si dimostri prima che l'equazione (96) ha una sola soluzione reale positiva, notando che il secondo membro della relazione

$$|a_n| = \sum_{i=0}^{n-1} |a_i| \, \frac{1}{x^{n-i}}$$

è una combinazione a coefficienti positivi delle funzioni $\dfrac{1}{x^{n-i}}$ e quindi per $x > 0$ è una funzione monotona decrescente che assume tutti i valori positivi. Risulta poi che $|x| > \gamma$ che

$$|p(x)| = \Big| \sum_{i=0}^{n} a_i x^i \Big| \geq |a_n| \, |x|^n - \sum_{i=0}^{n-1} |a_i| \, |x|^i,$$

e poiché il polinomio

$$|a_n| \, z^n - \sum_{i=0}^{n-1} |a_i| \, z^i$$

non ha zeri maggiori di γ, ne risulta che per $|x| > \gamma$ è $|p(x)| > 0$.
b) L'equazione

$$p(x) = 2\,x^5 - 21\,x^4 - 38\,x^3 - 87\,x^2 - 36\,x - 108 = 0$$

ha la soluzione positiva compresa nell'intervallo $(12.3, 12.4)$, mentre l'equazione

$$p(x) = 108\,x^5 - 36\,x^4 - 87\,x^3 - 38\,x^2 - 21\,x - 2 = 0$$

ha la soluzione positiva compresa nell'intervallo $(1.2, 1.3)$. Quindi i moduli delle radici sono compresi fra 0.77 e 12.4.)

3.54 Si determini una maggiorazione del modulo degli zeri del polinomio

$$p(x) = 10x^{10} + x^5 + x - 1.$$

(Traccia: sfruttando il teorema 3.53 punto b), risulta $|\alpha_i| < 1$, invece dall'esercizio 3.53 risulta $|\alpha_i| < 0.9$, infatti, posto $q(x) = 10x^{10} - x^5 - x - 1$, è $q(0.9) > 0$.)

3.55 L'equazione

$$x^3 - 1.56\,x^2 + 0.8111\,x - 0.140556 = 0$$

ha tre radici reali positive, di cui la minima è 0.51. Risolvendo invece l'equazione

$$x^3 - 1.56\,x^2 + 0.8111\,x - 0.14 = 0$$

si trova la sola radice reale 0.4373657. Si spieghi perché una così piccola perturbazione dell'ultimo coefficiente provoca una variazione così elevata negli zeri del polinomio.

(Traccia: si utilizzi la (76).)

3.56 Sia

$$p(x) = x^3 + a_2x^2 + a_1x + a_0 = 0 \tag{97}$$

un'equazione di terzo grado.

a) Si determini un numero h tale che posto $x = y + h$ nell'equazione a radici traslate

$$q(y) = p(y + h) = 0$$

manchi il termine di secondo grado e si scriva l'equazione così ottenuta nella forma

$$q(y) = y^3 + 3\alpha y - 2\beta = 0. \tag{98}$$

b) Si verifichi che se si indicano con s^3 e t^3 le due soluzioni dell'equazione di secondo grado

$$z^2 - 2\beta z - \alpha^3 = 0,$$

e si scelgono per s e t le determinazioni tali che $st = -\alpha$, allora $y_1 = s+t$ è soluzione dell'equazione (98).

c) Si calcolino anche le altre soluzioni y_2 e y_3 dell'equazione (98).

d) Si dia la formula risolutiva (di *Cardano*) dell'equazione di terzo grado (97).

e) Si dica, nel caso che i coefficienti siano reali, quando le soluzioni sono tutte reali.

(Traccia: a) si ha

$$p(y + h) = (y + h)^3 + a_2(y + h)^2 + \ldots = y^3 + (3h + a_2)y^2 + \ldots,$$

da cui $$h = -\frac{a_2}{3}, \quad \alpha = \frac{3a_1 - a_2^2}{9}, \quad \beta = \frac{9a_1a_2 - 27a_0 - 2a_2^3}{54}.$$

b) È $$s^3 + t^3 = 2\beta \quad \text{e} \quad st = -\alpha,$$

e risulta

$$q(s+t) = (s+t)^3 + 3\alpha(s+t) - 2\beta = s^3 + t^3 + 3(st+\alpha)(s+t) - 2\beta = 0.$$

c) Si ha
$$q(y) = [y - (s+t)]\,[y^2 + (s+t)y + 3\alpha + (s+t)^2],$$

da cui

$$y_{2,3} = -\frac{s+t}{2} \pm \frac{\sqrt{d}}{2}, \quad \text{dove} \quad d = (s+t)^2 - 12\alpha - 4(s+t)^2 = -3(s-t)^2,$$

e quindi
$$y_{2,3} = -\frac{s+t}{2} \pm \frac{1}{2}\,\boldsymbol{i}\sqrt{3}(s-t).$$

d) Calcolati α e β e posto

$$s = \sqrt[3]{\beta + \sqrt{\alpha^3 + \beta^2}}, \qquad t = \sqrt[3]{\beta - \sqrt{\alpha^3 + \beta^2}},$$

le soluzioni di (97) sono

$$\begin{aligned} x_1 &= s + t - \frac{a_2}{3}, \\ x_2 &= -\frac{s+t}{2} - \frac{a_2}{3} + \frac{1}{2}\,\boldsymbol{i}\sqrt{3}(s-t), \\ x_3 &= -\frac{s+t}{2} - \frac{a_2}{3} - \frac{1}{2}\,\boldsymbol{i}\sqrt{3}(s-t). \end{aligned}$$

e) Si verifichi che è sempre possibile scegliere s e t in modo che la soluzione reale sia la x_1. Se $D = \alpha^3 + \beta^2 < 0$, allora $s - t$ è un numero immaginario puro, e quindi x_2 e x_3 sono reali.)

3.57 Sia
$$p(x) = x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 = 0 \tag{99}$$

un'equazione di quarto grado.

a) Si determinino α, β, γ tali che

$$p(x) = \left(x^2 + \frac{1}{2}\,a_3x + \frac{1}{4}\,\alpha\right)^2 - \frac{1}{4}\,(\beta x + \gamma)^2.$$

b) Si determinino le soluzioni dell'equazione (99).

(Traccia: a) α è una soluzione reale dell'equazione di terzo grado

$$y^3 - a_2 y^2 + (a_1 a_3 - 4a_0)y + (4a_0 a_2 - a_1^2 - a_0 a_3^2) = 0,$$

e $$\beta^2 = 4\alpha - 4a_2 + a_3^2, \qquad \gamma^2 = \alpha^2 - 4a_0,$$

scegliendo per β e γ due determinazioni tali che $\beta\gamma = a_3\alpha - 2a_1$.
b) Dall'equazione

$$\left(x^2 + \frac{1}{2}\, a_3 x + \frac{1}{2}\, \alpha\right)^2 = \frac{1}{4}\, (\beta x + \gamma)^2$$

segue che le soluzioni di (99) sono le soluzioni delle equazioni di secondo grado

$$x^2 + \frac{1}{2}\, a_3 x + \frac{1}{2}\, \alpha = \pm\, \frac{1}{2}\, (\beta x + \gamma),$$

cioè delle equazioni

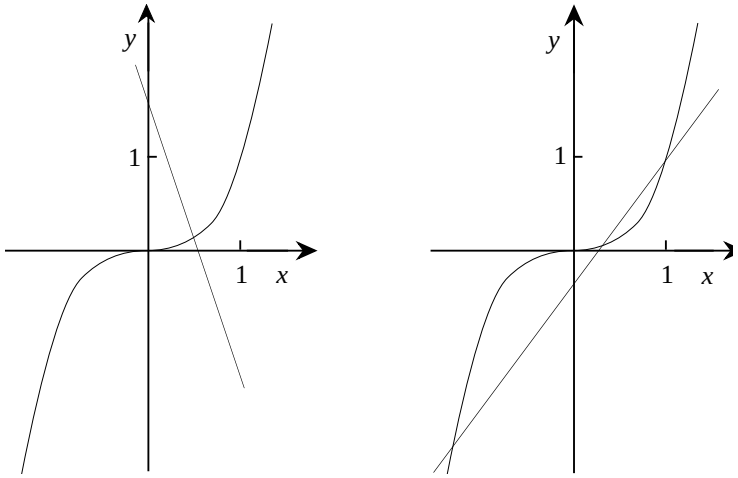
$$x^2 + \frac{1}{2}\, (a_3 \pm \beta) + \frac{1}{2}\, (\alpha \pm \gamma) = 0.\ )$$

3.58 Sia

$$x^3 + a_1 x + a_0 = 0$$

un'equazione di terzo grado (ogni equazione di terzo grado può essere ricondotta a questa forma, si veda l'esercizio 3.56). Usando una tecnica grafica, si dica quante soluzioni reali ha l'equazione e per quali valori dei coefficienti vi sono soluzioni reali di molteplicità due.

(Traccia: si disegnino i grafici delle due funzioni $y = x^3$ e $y = -a_1 x - a_0$, come in figura 3.22, in cui sono distinti i casi $a_1 > 0$ e $a_1 < 0$.



$a_1 > 0$ $\qquad\qquad$ $a_1 < 0$

Fig. 3.22 - Grafico delle funzioni $y = x^3$ e $y = -a_1 x - a_0$.

Le ascisse delle intersezioni danno gli zeri cercati. Se $a_1 > 0$ vi è un solo zero reale qualunque sia a_0, se $a_1 < 0$ gli zeri reali possono essere 1 o 3, a seconda del valore di a_0. In questo caso la retta è tangente alla cubica in un punto $\overline{x}$ se

$$\overline{x}^{\,3} = -a_1\overline{x} - a_0, \quad 3\overline{x}^{\,2} = -a_1,$$

per cui vi sono zeri multipli se $a_1 < 0$ e

$$|a_0| = \frac{2}{3}\sqrt{\frac{|a_1|^3}{3}}\,.$$

Per valori di $|a_0|$ minori vi sono tre zeri reali.)

3.59 Sia $p(x)$ un polinomio di grado n; si consideri la successione di polinomi

$$p(x), p'(x), \ldots, p^{(n)}(x),$$

e per un x si definisca $w(x)$ il numero di variazioni di segno della successione.

a) Si dimostri che (*regola di Budan-Fourier*) se $\alpha < \beta$, $p(\alpha)p(\beta) \neq 0$ e m è il numero delle radici di $p(x)$ nell'intervallo $[\alpha, \beta]$, contate con la loro molteplicità, allora

$$w(\alpha) - w(\beta) \geq m,$$

e il numero $w(\alpha) - w(\beta) - m$ è pari (eventualmente zero).

b) Se nella successione dei coefficienti

$$a_n, a_{n-1}, \ldots, a_0$$

del polinomio $p(x)$ ci sono w variazioni di segno e se $p(x)$ ha m zeri positivi, allora il numero $w - m$ è non negativo e pari (*regola dei segni di Cartesio*).

(Traccia: a) si supponga di far variare x da α a β; un cambiamento nel numero di variazioni di segno si può avere soltanto se si annulla uno dei polinomi per qualche ξ, $\alpha \leq \xi \leq \beta$. Se ξ è zero di $p(x)$ di molteplicità k, si verifichi che per $\epsilon > 0$ abbastanza piccolo, limitatamente alla successione dei primi $k+1$ polinomi, risulta $w(\xi - \epsilon) - w(\xi + \epsilon) = k$, tenendo conto che è

$$\left.\begin{aligned} p^{(s)}(\xi + \epsilon) &= \frac{\epsilon^{k-s} p^{(k)}(\xi)}{(k-s)!} + \ldots \\ p^{(s)}(\xi - \epsilon) &= \frac{(-\epsilon)^{k-s} p^{(k)}(\xi)}{(k-s)!} + \ldots \end{aligned}\right\} \qquad s = 0, \ldots, k-1.$$

In modo analogo si verifichi che se ξ è zero di molteplicità k di $p^{(r)}(x)$, con $1 \leq r \leq n$ e $p^{(r-1)}(\xi) \neq 0$, allora per la successione $p^{(r-1)}(x), \ldots, p^{(r+k)}(x)$ risulta che $w(\xi - \epsilon) - w(\xi + \epsilon)$ è non negativo e pari.

b) La proprietà è conseguenza della regola di Budan-Fourier applicata all'intervallo $[0, +\infty]$.)

3.60 Per le seguenti equazioni

a) $$2x^3 + 3x^2 - kx - 1 = 0,$$

b) $$3x^3 - kx^2 + x + 1 = 0,$$

c) $$2x^4 - kx^3 + 5x + 4 = 0,$$

si dica, facendo uso della successione di Sturm, quante sono le radici reali positive e negative, al variare del parametro k.

(Traccia: si ha

a)
$$\begin{aligned} p_0(x) &= 2x^3 + 3x^2 - kx - 1, \\ p_1(x) &= -6x^2 - 6x + k, \\ p_2(x) &= \frac{1}{6}\,[(4k+6)x - k + 6], \\ p_3(x) &= -\,\frac{k(8k^2 + 9k + 108)}{2(3+2k)^2}\,. \end{aligned}$$

Per $k < 0$ c'è una sola radice positiva, per $k = 0$ una radice positiva e una doppia negativa, per $k > 0$ due radici negative e una positiva.

b)
$$\begin{aligned} p_0(x) &= 3x^3 - kx^2 + x + 1, \\ p_1(x) &= -9x^2 + 2kx - 1, \\ p_2(x) &= \frac{1}{27}\,[2k^2x - 18x - (k+27)], \\ p_3(x) &= -\,\frac{27(k-5)(4k^2 + 21k + 51)}{4(k^2-9)^2}\,. \end{aligned}$$

Per $k < 5$ c'è una sola radice negativa, per $k = 5$ una radice negativa e una doppia positiva, per $k > 5$ una radice negativa e due positive.

c)
$$\begin{aligned} p_0(x) &= 2x^4 - kx^3 + 5x + 4, \\ p_1(x) &= -8x^3 + 3kx^2 - 5, \\ p_2(x) &= \frac{1}{32}\,[3k^2x^2 - 120x - (5k+128)], \\ p_3(x) &= \frac{64}{3k^4}\,[(600 + 16k^2 - 5k^3)x + (640 + 25k - 6k^3)], \\ p_4(x) &= -\,\frac{k^4}{32}\,\frac{(k+1)(4k-23)(27k^2 + 97k + 691)}{(5k^3 - 16k^2 - 600)^2}\,. \end{aligned}$$

Per $k < -1$ vi sono due radici negative, per $k = -1$ una radice doppia negativa, per $-1 < k < \dfrac{23}{4}$ nessuna radice reale, per $k = \dfrac{23}{4}$ una radice doppia positiva, due radici positive per $k > \dfrac{23}{4}$.

3.61 Facendo uso della successione di Sturm, si verifichi che l'equazione

$$x^3 - 3x^2 + (3 - 2k^2)x - (1 - 2k^2) = 0$$

ha per ogni k tre soluzioni reali nell'intervallo $[1 - 2|k|, 1 + 2|k|]$.

(Traccia: si ha

$$\begin{aligned} p_0(x) &= x^3 - 3x^2 + (3 - 2k^2)x - (1 - 2k^2), \\ p_1(x) &= -3x^2 + 6x - (3 - 2k^2), \\ p_2(x) &= \frac{4k^2}{3}\,(x - 1), \\ p_3(x) &= -2k^2. \end{aligned}$$

Per $k > 0$ risulta $w(1 - 2k) = 0$ e $w(1 + 2k) = 3$; per $k < 0$ risulta $w(1 + 2k) = 0$ e $w(1 - 2k) = 3$.)

3.62 Sia $p(x)$ un polinomio di grado $n \geq 3$ con tutti zeri reali $\alpha_1 > \alpha_2 > \ldots > \alpha_n$. Si dimostri che se $x_i > \alpha_1$, posto

$$x_{i+1} = x_i - 2\,\frac{p(x_i)}{p'(x_i)}\,,$$

risulta $x_{i+1} > \alpha_1$ oppure $\alpha_2 < x_{i+1} < \alpha_1$.

(Traccia: poiché

$$\frac{p'(x)}{p(x)} = \sum_{j=1}^{n} \frac{1}{x - \alpha_j}\,,$$

per $x_i > \alpha_1$ è

$$\frac{p'(x_i)}{p(x_i)} > \frac{1}{x_i - \alpha_1} + \frac{1}{x_i - \alpha_2}\,.$$

D'altra parte è

$$\frac{p'(x_i)}{p(x_i)} = \frac{2}{x_i - x_{i+1}}\,,$$

e quindi

$$\frac{2}{x_i - x_{i+1}} > \frac{1}{x_i - \alpha_1} + \frac{1}{x_i - \alpha_2}\,.$$

Se $x_{i+1} < \alpha_1$, è $\dfrac{1}{x_i - \alpha_1} > \dfrac{1}{x_i - x_{i+1}}$, per cui $\dfrac{1}{x_i - x_{i+1}} > \dfrac{1}{x_i - \alpha_2}$ e quindi $x_{i+1} > \alpha_2$.)

3.63 Sia $p(x)$ un polinomio di grado $n \geq 2$ con n zeri reali e distinti e tali che $|\alpha_1| > |\alpha_2| \geq \ldots \geq |\alpha_n|$, e sia F la matrice di Frobenius definita per il metodo di Bernoulli. Si calcolino i coefficienti del polinomio $\phi_k(x)$ di grado minore di n

$$\phi_k(x) = x^{2^k} \bmod p(x),$$

mediante le relazioni

$$\begin{aligned} &\phi_0(x) = x \\ &\phi_{i+1}(x) = \phi_i^2(x) \bmod p(x), \quad i = 0, \ldots, k-1, \end{aligned}$$

calcolando il resto della divisione di $\phi_i^2(x)$ per $p(x)$.

a) Si dimostri che per i vettori $\mathbf{t}_i = F^i \mathbf{t}_0$ vale la relazione

$$\mathbf{t}_{2^k} = \phi_k(F)\mathbf{t}_0;$$

b) si dimostri che la radice α_1 risulta approssimata dal rapporto fra l'ultima e la penultima componente di $\mathbf{t}_{2^k}$, a meno di un errore dell'ordine di

$$O\Big(\Big|\frac{\alpha_2}{\alpha_1}\Big|^{2^k}\Big);$$

c) posto

$$\phi_k(x) = \sum_{i=0}^{n-1} \sigma_i x^i,$$

si valuti il numero di operazioni aritmetiche sufficienti a calcolare i coefficienti σ_i, $i = 0, \ldots, n-1$, e a costruire il vettore $\mathbf{t}_{2^k}$, calcolando $\phi_k(F)\mathbf{t}_0$ con il metodo di Ruffini-Horner

$$\phi_k(F)\mathbf{t}_0 = F(\ldots(F(\sigma_{n-1}F\mathbf{t}_0 + \sigma_{n-2}\mathbf{t}_0) + \sigma_{n-3}\mathbf{t}_0) + \ldots) + \sigma_0\mathbf{t}_0.$$

(Traccia: a) si usi il fatto che $p(F) = O$ e che

$$x^{2^k} = p(x)q(x) + \phi_k(x),$$

per dimostrare che $\phi_k(F) = F^{2^k}$; b) si proceda come per il metodo di Bernoulli; c) si utilizzino i risultati degli esercizi 1.6 e 1.10 per il calcolo del quadrato e del resto della divisione e si tenga conto che il vettore $\mathbf{y} = F\mathbf{x}$ può essere calcolato con n moltiplicazioni ed $n-1$ addizioni.)

3.64 Sia $p(x)$ un polinomio di grado n, di zeri α_1, $\alpha_2, \ldots, \alpha_n$ reali e distinti.

a) Si verifichi che per $x \neq \alpha_i$, $i = 1, \ldots, n$, è

$$\frac{p'(x)}{p(x)} = \sum_{i=1}^{n} \frac{1}{x - \alpha_i}\,, \quad \frac{[p'(x)]^2 - p(x)p''(x)}{[p(x)]^2} = \sum_{i=1}^{n} \frac{1}{(x - \alpha_i)^2}\,.$$

b) Posto

$$u(x) = \frac{p(x)}{p'(x)} \quad \text{e} \quad v(x) = \frac{p''(x)}{p'(x)}\,,$$

si verifichi che la funzione

$$h(x) = (n-1)^2 - n(n-1)u(x)v(x)$$

è non negativa.

Il metodo iterativo

$$x_{i+1} = x_i - \frac{nu(x_i)}{1 + \sqrt{h(x_i)}}$$

è detto metodo di *Laguerre*, e richiede ad ogni passo il calcolo di $p(x_i)$, $p'(x_i)$ e $p''(x_i)$.

c) Si dimostri che l'ordine del metodo è 3 e che se $\alpha_1 > \alpha_2 > \ldots > \alpha_n$, e il valore iniziale x_0 è tale che $p'(x_0) \neq 0$ e $x_0 \in (\alpha_{k+1}, \alpha_k)$, $1 \leq k \leq n-1$, allora il metodo converge ad α_k o ad α_{k+1}; se $x_0 < \alpha_n$, il metodo converge ad α_n e se $x_0 > \alpha_1$, il metodo converge ad α_1.

Quindi il metodo è globalmente convergente. Una delle difficoltà del metodo è che, nel caso che gli zeri non siano tutti reali, possono venire generati punti complessi anche se il polinomio è a coefficienti reali e la successione converge ad uno zero reale. Ciò può accadere, a causa degli errori di arrotondamento, anche se gli zeri sono tutti reali, ma non tutti distinti o ben separati.

(Traccia: a) la seconda relazione si ottiene derivando la prima. b) Risulta

$$\begin{aligned} h(x) &= (n-1)\Big(\frac{p(x)}{p'(x)}\Big)^2 \Big[n\, \frac{[p'(x)]^2 - p(x)p''(x)}{[p(x)]^2} \; - \; \Big(\frac{p'(x)}{p(x)}\Big)^2 \Big] \\ &= (n-1)\Big(\frac{p(x)}{p'(x)}\Big)^2 \Big[n \sum_{i=1}^{n} y_i^2 - \Big(\sum_{i=1}^{n} y_i \Big)^2 \Big], \quad y_i = \frac{1}{x - \alpha_i}\,, \end{aligned}$$

e per la disuguaglianza di Cauchy-Schwartz ([3], pag. 5) è

$$n \sum_{i=1}^{n} y_i^2 \geq \Big(\sum_{i=1}^{n} y_i \Big)^2 .$$

c) La funzione di iterazione è

$$g(x) = x - \frac{nu(x)}{1 + (n-1)\sqrt{1 - \dfrac{n}{n-1}\,u(x)v(x)}}\,.$$

Per quanto riguarda l'ordine del metodo, si può verificare direttamente che $g'(\alpha_i) = g''(\alpha_i) = 0$, per $i = 1, \ldots, n$, oppure sviluppando in serie di potenze di u si ha

$$g(x) = x - u - \frac{u^2 v}{2} + \ldots = x - \frac{p(x)}{p'(x)} - \frac{[p(x)]^2 p''(x)}{2[p'(x)]^3} + \ldots$$

Sfruttando i primi termini della serie di Taylor di $p(x)$ e $p'(x)$, si verifichi che esiste una funzione $\phi(x)$ tale che

$$g(x) - \alpha_i = (x - \alpha_i)^3 \phi(\alpha_i) + O((x - \alpha_i)^4).$$

Poiché le radici sono reali e distinte, in ogni intervallo (α_{k+1}, α_k), $1 \le k \le n-1$, esiste un punto ξ_k tale che $u(x) > 0$ per $\alpha_{k+1} < x < \xi_k$ e $u(x) < 0$ per $\xi_k < x < \alpha_k$. Quindi se $x_0 < \xi_k$, la successione converge a α_{k+1}, se $x_0 > \xi_k$, la successione converge a α_k. Analogamente per $x > \alpha_1$ o $x < \alpha_n$.)

3.65 Sia $p(x)$ un polinomio di grado n, di zeri α_1, $\alpha_2, \ldots, \alpha_n$. Il metodo di *Gräffe* consiste nel costruire una successione di polinomi di grado n, $p_j(x)$, $j = 0, 1, \ldots$, tali che

$p_1(x) = p(x)$,

$p_{j+1}(x)$ è il polinomio i cui zeri sono i quadrati degli zeri di $p_j(x)$ (per la costruzione si veda l'esercizio 3.50).

Si indichino con $a_i^{(j)}$, $i = 0, \ldots, n$ i coefficienti del polinomio $p_j(x)$. Si dimostri che

a) gli zeri di $p_j(x)$ sono $\alpha_i^{2^j}$, $i = 1, \ldots, n$;

b) se $|\alpha_1| > |\alpha_2| > \ldots > |\alpha_n|$, vale

$$\left|\frac{a_{n-1}^{(j)}}{a_n^{(j)}}\right| = |\alpha_1|^{2^j}\left[1 + O\Big(\Big|\frac{\alpha_2}{\alpha_1}\Big|^{2^j}\Big)\right],$$

e quindi

$$|\alpha_1| = \lim_{j\to\infty} \left|\frac{a_{n-1}^{(j)}}{a_n^{(j)}}\right|^{1/2^j}.$$

Inoltre

$$\Big|\frac{a_{n-i}^{(j)}}{a_n^{(j)}}\Big| = |\alpha_1\alpha_2\ldots\alpha_i|^{2^j}\Big[1+O\Big(\Big|\frac{\alpha_{i+1}}{\alpha_i}\Big|^{2^j}\Big)\Big],$$

e

$$\Big|\frac{a_{n-i+1}^{(j)}}{a_n^{(j)}}\Big| = |\alpha_1\alpha_2\ldots\alpha_{i-1}|^{2^j}\Big[1+O\Big(\Big|\frac{\alpha_i}{\alpha_{i-1}}\Big|^{2^j}\Big)\Big],$$

per cui

$$|\alpha_i| = \lim_{j\to\infty}\ \Big|\frac{a_{n-i}^{(j)}}{a_{n-i+1}^{(j)}}\Big|^{1/2^j}.$$

Il metodo di Gräffe permette quindi di approssimare i moduli degli zeri di un polinomio. Nell'implementazione del metodo conviene arrestare le iterazioni quando durante la costruzione dei coefficienti del polinomio $p_{j+1}(x)$ i doppi prodotti risultano trascurabili rispetto ai quadrati. Il metodo è piuttosto sensibile agli errori di arrotondamento, inoltre è necessario implementare delle tecniche per contenere gli errori di overflow, che altrimenti non consentirebbero di operare per valori elevati di j.

(Traccia: b) si sfruttino le relazioni dell'esercizio 3.48. È

$$-\frac{a_{n-1}^{(j)}}{a_n^{(j)}} = \alpha_1^{2^j}+\alpha_2^{2^j}+\ldots+\alpha_n^{2^j} = \alpha_1^{2^j}\Big[1+\Big(\frac{\alpha_2}{\alpha_1}\Big)^{2^j}+\ldots+\Big(\frac{\alpha_n}{\alpha_1}\Big)^{2^j}\Big],$$

e analogamente per le altre relazioni.)

3.66 Per approssimare gli zeri $\alpha_1,\ \alpha_2,\ldots,\alpha_n$ di un polinomio $p(x)$ si risolva con il metodo di Newton-Raphson il sistema non lineare

$$s_j(x_1,x_2,\ldots,x_n) = (-1)^j\ \frac{a_{n-j}}{a_n}, \quad j=1,2\ldots,n,$$

dove $s_j(x_1,x_2,\ldots,x_n)$ sono le funzioni simmetriche elementari definite nell'esercizio 3.48. Infatti $\alpha_1,\ \alpha_2,\ldots,\alpha_n$ sono soluzioni di tale sistema. Si dimostri che

a) il metodo ottenuto è individuato dalla seguente relazione

$$x_i^{(k+1)} = x_i^{(k)} - \frac{p(x_i^{(k)})}{a_n\ \prod\limits_{\substack{j=1\\ j\neq i}}^{n}(x_i^{(k)}-x_j^{(k)})}, \quad i=1,2,\ldots,n;$$

b) il metodo è localmente convergente del secondo ordine se $\alpha_i\neq\alpha_j$, per $i\neq j$.

Tale metodo è noto in letteratura come metodo di *Durand-Kerner*. È stata fatta la congettura che questo metodo sia globalmente convergente.

(Traccia: a) posto $\mathbf{x} = (x_1, \ldots, x_n)$, dalla relazione

$$\prod_{j=1}^{n}(z - x_j) = z^n + \sum_{j=1}^{n}(-1)^j s_j(\mathbf{x}) z^{n-j}, \tag{100}$$

derivando rispetto a x_k si ottiene

$$-\prod_{\substack{j=1\\ j\neq k}}^{n}(z - x_j) = \sum_{j=1}^{n}(-1)^j \frac{\partial s_j(\mathbf{x})}{\partial x_k}\ z^{n-j}.$$

Si sfrutti questa relazione per dimostrare che, indicate con $J(\mathbf{x})$ la matrice il cui elemento (j,k)-esimo è dato da $\dfrac{\partial s_j(\mathbf{x})}{\partial x_k}$, con $K(\mathbf{x})$ la matrice il cui elemento (j,i)-esimo è dato da $(-1)^j x_i^{n-j}$ e con $D(\mathbf{x})$ la matrice diagonale il cui i-esimo elemento principale è dato da

$$-\prod_{\substack{j=1\\ j\neq i}}^{n}(x_i - x_j),$$

risulta

$$J^T(\mathbf{x})K(\mathbf{x}) = D(\mathbf{x}),$$

da cui

$$J^{-1}(\mathbf{x}) = D^{-1}(\mathbf{x})K^T(\mathbf{x}).$$

Perciò indicato con $\mathbf{f}(\mathbf{x})$ il vettore la cui j-esima componente è

$$f_j(\mathbf{x}) = s_j(\mathbf{x}) - (-1)^j\ \frac{a_{n-j}}{a_n}\,,$$

il metodo di Newton-Raphson è dato da

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - J^{-1}(\mathbf{x}^{(k)})\mathbf{f}(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} - D^{-1}(\mathbf{x}^{(k)})K^T(\mathbf{x}^{(k)})\mathbf{f}(\mathbf{x}^{(k)}).$$

La i-esima componente del vettore $K^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ è data da

$$\begin{aligned}\sum_{j=1}^{n}(-1)^j x_i^{n-j}\Big(s_j(\mathbf{x}) - (-1)^j\ \frac{a_{n-j}}{a_n}\Big) &= -\ \frac{1}{a_n}\ \Big[\sum_{j=1}^{n} a_{n-j}x_i^{n-j} + a_n x_i^n\Big]\\ &= -\ \frac{1}{a_n}\ p(x_i),\end{aligned}$$

poiché per la (100) è

$$\sum_{j=1}^{n}(-1)^j s_j(\mathbf{x})x_i^{n-j} = \prod_{j=1}^{n}(x_i - x_j) - x_i^n.$$

b) La convergenza locale segue dalla convergenza del metodo di Newton-Raphson, essendo le funzioni simmetriche elementari differenziabili in un intorno delle soluzioni.)

3.67 Sia $p(x)$ un polinomio di grado n, di zeri distinti α_1, $\alpha_2, \ldots, \alpha_n$.

a) Si verifichi che $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ è, a meno di permutazioni, l'unica soluzione del sistema non lineare

$$p[x_1, x_2, \ldots, x_i] = 0, \quad i = 1, \ldots, n, \quad x_j \neq x_k \text{ per } j \neq k \qquad (101)$$

(per la definizione di differenza divisa di una funzione, si veda 5.14).

b) Si verifichi che il metodo di Newton-Raphson applicato al sistema (101) è definito da

$$x_i^{(k+1)} = x_i^{(k)} - d_i(\mathbf{x}^{(k)}), \quad i = 1, \ldots, n, \quad \text{in cui}$$

$$d_i(\mathbf{x}) = \frac{p[x_1, \ldots, x_i] - \sum_{j=1}^{i-1} d_j(\mathbf{x}) p[x_1, \ldots, x_j, x_j, \ldots, x_i]}{p[x_1, \ldots, x_i, x_i]},$$

dove il risultato della sommatoria è da intendersi uguale a zero se il secondo estremo risulta nullo.

Tale metodo è noto in letteratura come metodo di *Pasquini-Trigiante*. È stato dimostrato che questo metodo è convergente del secondo ordine per quasi ogni scelta del vettore iniziale (cioè per tutti i vettori iniziali, escluso al più un insieme di misura nulla di vettori), nell'ipotesi che gli zeri siano tutti reali e distinti.

(Traccia: a) $p[x_1, \ldots, x_i]$ è combinazione lineare dei valori $p(x_1), \ldots, p(x_i)$; b) si verifichi, facendo riferimento alla (29, cap. 5), che l'elemento (i, j)-esimo della matrice $J(\mathbf{x})$ è dato da

$$\frac{\partial p[x_1, \ldots, x_i]}{\partial x_j} = \begin{cases} p[x_1, \ldots, x_j, x_j, \ldots, x_i], & \text{per } j = 1, \ldots, i, \\ 0, & \text{per } j = i+1, \ldots, n. \end{cases}$$

Quindi la matrice $J(\mathbf{x})$ è triangolare inferiore. Si determini l'i-esimo elemento del vettore $\mathbf{d}(\mathbf{x})$, soluzione del sistema $J(\mathbf{x})\mathbf{d}(\mathbf{x}) = \mathbf{f}(\mathbf{x})$, in cui $f_i(\mathbf{x}) = p[x_1, \ldots, x_i]$.)

3.68 Sia $p(x)$ un polinomio di grado n, di zeri distinti $\alpha_1, \alpha_2, \ldots, \alpha_n$. Si consideri il sistema non lineare $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, dove

$$f_i(\mathbf{x}) = \frac{p(x_i)}{\prod_{j=1}^{i-1}(x_i - x_j)},$$

dove il risultato della produttoria è da intendersi uguale a uno se il secondo estremo risulta nullo.

a) Si verifichi che $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ è una soluzione del sistema;

b) indicata con $C(\mathbf{x})$ la matrice diagonale il cui i-esimo elemento principale è $\dfrac{\partial f_i(\mathbf{x})}{\partial x_i}$, si verifichi che $C(\boldsymbol{\alpha})$ è non singolare e che il metodo iterativo

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [C(\mathbf{x}^{(k)})]^{-1}\mathbf{f}(\mathbf{x}^{(k)})$$

può essere scritto nella forma

$$x_i^{(k+1)} = x_i^{(k)} - \frac{1}{\dfrac{p'(x_i^{(k)})}{p(x_i^{(k)})} - \displaystyle\sum_{j=1}^{i-1} \frac{1}{x_i^{(k)} - x_j^{(k)}}}, \quad i = 1, \ldots, n,$$

dove il risultato della sommatoria è da intendersi uguale a zero se il secondo estremo risulta nullo, e se ne analizzi la convergenza locale.

Tale metodo può essere utilizzato per il calcolo simultaneo degli zeri di $p(\mathbf{x})$ e può essere ottenuto anche applicando una tecnica simile a quella della deflazione implicita di Maehly, in cui al posto degli zeri già calcolati si usano le loro approssimazioni. Del metodo si può dare la seguente modifica

$$x_i^{(k+1)} = x_i^{(k)} - \frac{1}{\dfrac{p'(x_i^{(k)})}{p(x_i^{(k)})} - \displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{x_i^{(k)} - x_j^{(k)}}}, \quad i = 1, \ldots, n,$$

che ovvia all'inconveniente, presentato dalla prima formulazione, di convergenza di più componenti allo stesso α_i. Tale metodo è noto in letteratura come metodo di *Aberth*.

c) Si verifichi che l'ordine di convergenza è 3, cioè che esiste una costante β tale che

$$\|\mathbf{x}^{(k+1)} - \boldsymbol{\alpha}\| \leq \beta \|\mathbf{x}^{(k)} - \boldsymbol{\alpha}\|^3,$$

per ogni norma.

(Traccia: b) si verifichi che, indicata con $\mathbf{g}(\mathbf{x}) = \mathbf{x} - [C(\mathbf{x})]^{-1}\mathbf{f}(\mathbf{x})$ la funzione di iterazione, risulta

$$\frac{\partial g_i(\boldsymbol{\alpha})}{\partial x_j} = 0, \quad \text{per} \quad i, j = 1, \ldots, n.$$

D'altra parte per $i = 1$ è

$$g_1(\mathbf{x}) = x_1 - \frac{p(x_1)}{p'(x_1)},$$

per cui in generale risulta

$$\frac{\partial^2 g_1(\boldsymbol{\alpha})}{\partial x_1^2} = \frac{p''(\alpha_1)}{p'(\alpha_1)} \neq 0,$$

e quindi il metodo è localmente del secondo ordine.

c) Si verifichi che in questo caso sono nulle in $\boldsymbol{\alpha}$ anche le derivate parziali seconde di $\mathbf{g}(\mathbf{x})$, in quanto risulta

$$\frac{\partial^2 g_i(\mathbf{x})}{\partial x_j \partial x_r} = p(x_i) q_{jr}(\mathbf{x}), \quad \text{per } j \text{ e } r \text{ non entrambi uguali ad } i,$$

dove $q_{jr}(\boldsymbol{\alpha})$ è limitato, e

$$\frac{\partial^2 g_i(\mathbf{x})}{\partial x_i^2} = \Big[p''(x_i) - 2p'(x_i) \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{x_i - x_j} \Big] q_{ii}(\mathbf{x}) + p(x_i) \widehat{q}_{ii}(\mathbf{x}),$$

dove $q_{ii}(\boldsymbol{\alpha})$ e $\widehat{q}_{ii}(\boldsymbol{\alpha})$ sono limitati. Si verifichi che anche questa espressione è nulla per $\mathbf{x} = \boldsymbol{\alpha}$.)

3.69 a) Sia $f(x) = (x - \alpha)\, q(x)$, $q(\alpha) \neq 0$, una funzione razionale della x con coefficienti in un insieme $\mathcal{F}$, con struttura di anello. Si dimostri che per la successione $\{x_i\}$ generata dal metodo di Newton applicato ad $f(x)$ vale la relazione

$$x_{i+1} - \alpha = (x_i - \alpha)^2 \, \frac{q'(x_i)}{q(x_i) + (x_i - \alpha) q'(x_i)}.$$

b) Dato il polinomio a coefficienti reali

$$p(z) = \sum_{i=0}^{n} a_i \, z^i, \quad a_0 \neq 0,$$

si calcolino con il metodo di Newton i primi k coefficienti della serie

$$s(z) = \sum_{i=0}^{\infty} \sigma_i z^i, \quad \text{tale che} \quad s(z)p(z) = 1,$$

cioè i coefficienti del polinomio

$$s_k(z) = \sum_{i=0}^{k-1} \sigma_i z^i, \quad \text{tale che} \quad s_k(z)p(z) = 1 + O(z^k).$$

c) Per lo stesso polinomio si calcolino con il metodo di Newton i primi k coefficienti della serie

$$t(z) = \sum_{i=0}^{\infty} \tau_i z^i, \quad \text{tale che} \quad t^2(z) = p(z).$$

d) Si determinino i primi 8 coefficienti delle serie $s(z)$ e $t(z)$ nel caso in cui $p(z) = 1 + z$.

(Traccia: b) sia $f(x) = x^{-1} - p(z)$ una funzione razionale con coefficienti nell'anello $\mathcal{F}$ dei polinomi in z a coefficienti reali. Il metodo di Newton, applicato all'equazione $f(x) = 0$, genera la successione di polinomi

$$x_{i+1}(z) = 2x_i(z) - x_i^2(z)p(z), \quad \text{con} \quad x_0(z) = \frac{1}{a_0}\,.$$

Posto $q(x) = -\dfrac{p(z)}{x}$, dal punto a) segue che

$$x_{i+1}(z) - \alpha = -(x_i(z) - \alpha)^2 p(z), \quad \text{con} \quad \alpha = s(z).$$

Notando che $x_0(z) = \sigma_0$, ne segue che

$$x_i(z) = \sum_{j=0}^{2^i-1} \sigma_j z^j + O\big(z^{2^i}\big), \quad i \geq 1.$$

c) Si proceda come nel caso b). Conviene considerare la funzione razionale $f(y) = 1 - y^{-2}p(z)$. Il metodo di Newton genera la successione

$$y_{i+1}(z) = \frac{1}{2}\left(3y_i(z) - \frac{y_i^3(z)}{p(z)}\right), \quad \text{con} \quad y_0(z) = \sqrt{a_0},$$

che ha convergenza quadratica anche quando al posto di $\dfrac{1}{p(z)}$ si utilizza il polinomio di grado $k-1$ ricavato al punto b). Si verifichi che la convergenza è quadratica anche se all'i-esimo passo si opera con i soli 2^{i+1} coefficienti dei termini di grado più basso, cioè con la successione ottenuta nel modo seguente

$$y_{i+1}(z) = \frac{1}{2}\left(3y_i(z) - y_i^3(z)x_{i+1}(z)\right) \bmod z^{2^{i+1}}.$$

d) Nel caso particolare

$$\begin{aligned}
&x_0(z) = 1, \quad x_1(z) = 1 - z,\\
&x_2(z) = 1 - z + z^2 - z^3, \quad x_3(z) = 1 - z + z^2 - z^3 + z^4 - z^5 + z^6 - z^7,\\
&y_0(z) = 1, \quad y_1(z) = 1 + \frac{z}{2},\\
&y_2(z) = 1 + \frac{z}{2} - \frac{z^2}{8} + \frac{z^3}{16},\\
&y_3(z) = 1 + \frac{z}{2} - \frac{z^2}{8} + \frac{z^3}{16} - \frac{5z^4}{128} + \frac{7z^5}{256} - \frac{21z^6}{1024} + \frac{33z^7}{2048}.\ )
\end{aligned}$$

Commento bibliografico

Il problema del calcolo delle soluzioni di equazioni è stato uno dei problemi centrali affrontati dai matematici di ogni tempo. Il primo esempio noto di metodo iterativo è riportato in un papiro egiziano risalente al 1650 a.C.: si tratta di un metodo per risolvere un'equazione lineare. L'algoritmo per il calcolo della radice quadrata di un numero k, che oggi si ricava applicando il metodo delle tangenti all'equazione $x^2 - k$, è noto come formula di Erone, matematico alessandrino del 2^o secolo, ma era in realtà già stato utilizzato dai matematici babilonesi, che erano in grado di determinare le soluzioni reali delle equazioni di secondo grado e di alcune equazioni di terzo grado. Anche Archimede risolveva equazioni di terzo grado con metodi geometrici.

Il metodo di falsa posizione fu usato dai primi algebristi italiani. Il metodo delle tangenti è originariamente di Newton e si trova in uno scritto che avrebbe dovuto essere pubblicato nel 1671, ma non fu stampato fino al 1736, per cui la prima versione stampata del metodo risulta quella di Wallis del 1685. In realtà il procedimento proposto da Newton è un po' diverso da quello attuale che è dovuto a Raphson (1690), quindi sarebbe più corretto chiamare il metodo delle tangenti come metodo di Newton-Raphson (come si è fatto per il caso dei sistemi di equazioni). Il fatto che il metodo sia indicato con il solo nome di Newton è dovuto a Fourier, che se ne occupò

estesamente e nel 1818 ne determinò l'ordine di convergenza; le condizioni sufficienti di convergenza del metodo delle tangenti e delle secanti sono note come condizioni di Fourier. Dal confronto fra i valori ottenuti applicando contemporaneamente il metodo delle tangenti ed il metodo delle secanti si ottiene una limitazione, dovuta a Dandelin, dell'errore commesso con l'approssimazione [19].

La definizione di ordine di un metodo iterativo $x_{i+1} = g(x_i)$ tramite le derivate della funzione g fu data da Schröder nel 1870. A Schröder è dovuta anche la modifica del metodo delle tangenti che dà convergenza quadratica per il calcolo delle soluzioni di molteplicità maggiore di uno. Successivamente sono stati costruiti metodi di ordine comunque elevato. Le definizioni di ordine e di fattore di convergenza date nel paragrafo 5 sono molto restrittive. Per definizioni più generali si veda [18], in cui si definiscono il fattore del quoziente e il fattore della radice, definizioni che si estendono più facilmente al caso dei metodi iterativi per i sistemi di equazioni.

Il metodo di Aitken fu presentato nel 1926, originariamente per accelerare la convergenza delle successioni ottenute con il metodo di Bernoulli per equazioni algebriche. La versione qui descritta fu proposta da Steffensen nel 1933, come applicazione del metodo delle secanti all'equazione $x - g(x) = 0$. Per altre tecniche di accelerazione della convergenza si vedano i libri di Ostrowski [19] e Traub [27], nei quali viene anche introdotta la nozione di efficienza di un metodo iterativo e vengono studiati molti metodi.

Negli anni 1960 fu costituito all'università di Amsterdam un gruppo del quale facevano parte, fra gli altri, van Wijngaarden e Dekker, per la realizzazione di un metodo per risolvere equazioni sotto ipotesi il più possibile generali. Il metodo sviluppato da Dekker nel 1969 [7], poi migliorato da Brent nel 1973 [5], è noto come metodo di Dekker-Brent (si veda l'esercizio 3.38) e combina il metodo di bisezione con metodi di ordine superiore.

Oltre al metodo delle tangenti, molti altri metodi possono essere generalizzati per i sistemi, si veda [19] e [27]. Il metodo delle secanti può essere generalizzato in vari modi: si possono ad esempio, partendo dal metodo di Newton-Raphson, sostituire le derivate parziali con approssimazioni alle differenze, oppure si possono sostituire alle funzioni f_i delle funzioni interpolanti (una tale generalizzazione è stata utilizzata da Gauss nel 1809). Alcune delle generalizzazioni non sono state ancora completamente analizzate dal punto di vista della stabilità e dell'ordine di convergenza, si veda [18]. Poiché il problema della risoluzione di un sistema non lineare è equivalente ad un problema di minimizzazione di un funzionale, in generale non quadratico, per risolvere un sistema si può fare ricorso anche ai metodi di minimizzazione, quali i metodi del gradiente [18].

La condizione sufficiente di convergenza basata sul raggio spettrale della

matrice jacobiana nella soluzione è stata enunciata nel 1957 da Ostrowski. Si tratta però di un risultato che può essere collegato ad un altro ottenuto da Perron nel 1929. La convergenza quadratica del metodo di Newton-Raphson fu stabilita da Runge nel 1899. La condizione sufficiente di convessità per la convergenza del metodo di Newton-Raphson è stata data da Baluev nel 1952, che dimostrò il risultato in spazi più generali di $\mathbf{R}^n$.

La formula risolutiva dell'equazione di terzo e quarto grado (si vedano gli esercizi 3.56 e 3.57) fu pubblicata da Cardano nel 1545, anche se vi sono dubbi sulla sua effettiva paternità. Le relazioni fra le radici e i coefficienti delle equazioni algebriche è dovuta a Girard nel 1629. È di Girard anche la scoperta che il numero delle radici di un'equazione algebrica è pari al grado, anche se per una dimostrazione rigorosa si deve attendere fino al 1799, anno in cui Gauss nella sua tesi di laurea dimostrò che i precedenti tentativi di dimostrazione, compresi quelli di Eulero e di Lagrange, non erano esenti da errori. Nel 1826 Abel dimostrò che in generale, eseguendo solamente operazioni aritmetiche ed estrazioni di radici quadrate, non è possibile determinare le radici di un'equazione algebrica di grado maggiore di 4, quindi per gradi superiori a 4 si deve ricorrere a metodi iterativi.

Nel 1924 Weyl diede una dimostrazione costruttiva del teorema fondamentale dell'algebra, dimostrando sostanzialmente che gli zeri di un polinomio complesso dipendono ricorsivamente dai suoi coefficienti. Alcuni dei più efficienti algoritmi per la determinazione degli zeri di un polinomio esistenti in letteratura sono tuttora basati sulla tecnica di esclusione di Weyl [12], come ad esempio l'algoritmo di Lehmer [16], che opera isolando le radici in cerchi del piano complesso e poi restringendo il raggio dei cerchi. Questi algoritmi presentano in generale una convergenza assai regolare (convergenza globale, insensibilità alle radici multiple), ma estremamente lenta. Per quanto riguarda lo studio della complessità computazionale di questo problema in ambiente di calcolo sequenziale e parallelo si veda [4], [20] e [25].

Per altre limitazioni superiori delle radici delle equazioni algebriche si veda [17] e [28]. Per il teorema di Ostrowski e per altre limitazioni dell'errore provocato sulle radici da perturbazioni dei coefficienti si veda [19]. Per la convergenza in caso di deflazione si veda [30]. Per la convergenza globale del metodo delle tangenti applicato alle equazioni algebriche si veda [2] e [26].

Il metodo di Bairstow è stato descritto da Bairstow nel 1914, in un articolo in cui si studia la stabilità del volo degli aerei. Il metodo di Bernoulli è stato descritto da Daniel Bernoulli nel 1728, ed è noto anche con il nome di *metodo dei momenti*. Aitken nel 1926 dimostrò per le equazioni algebriche che con i rapporti di particolari determinanti formati con i coefficienti si possono approssimare anche radici non di massimo modulo. Da questa idea

si sviluppò, ad opera di Rutishauser nel 1954 [23] il metodo qd. La versione del metodo qd descritta in questo libro è quella di Householder, particolarmente adatta ad una implementazione in ambiente di calcolo parallelo. Però di questo metodo esiste anche un'altra versione, in cui la tabella delle quantità $q_k^{(i)}$ e $e_k^{(i)}$ è ottenuta seguendo un diverso ordinamento. Tale variante, data da Rutishauser nel 1962, corrisponde al metodo LR per il calcolo degli autovalori di matrici tridiagonali e permette di ottenere una convergenza del secondo ordine se si utilizza con un'opportuna tecnica di *shift* [24]. Il metodo qd presenta molte connessioni insospettabili con una quantità di settori diversi della matematica teorica e applicata. Se ne vedano alcuni esempi di applicazioni al calcolo delle serie reciproche, dei poli di funzioni meromorfe, delle frazioni continue, di integrali di Laplace e di polinomi ortogonali in [11].

La possibilità di utilizzare le funzioni simmetriche elementari dei quadrati delle radici di un'equazione algebrica fu suggerita da Dandelin nel 1826 e ripresa da Gräffe nel 1837, che la utilizzò per l'algoritmo che porta il suo nome (si veda l'esercizio 3.65). Lo stesso algoritmo è noto anche con il nome di Lobachevsky, che nel 1834, in modo indipendente da Dandelin suggerì lo stesso procedimento. L'effettiva implementazione del metodo di Gräffe deve prevedere delle tecniche per evitare che si verifichino errori di overflow o di underflow [10]. Per una descrizione del metodo di Gräffe, con le tecniche per determinare anche il segno (o l'argomento per gli zeri complessi) e la molteplicità degli zeri, si veda il libro di Householder [13].

Molti algoritmi, alcuni dei quali usati frequentemente, approssimano le radici una alla volta, ricorrendo alla deflazione. A questa categoria, ad esempio, appartiene il metodo di Jenkins e Traub [14], che è praticamente diventato lo standard ed è incluso nella libreria di programmi del IMSL, si veda [22]. Questo modo di procedere presenta un notevole inconveniente quando non è richiesto di determinare le radici con grande accuratezza (come accade spesso nelle applicazioni pratiche), in quanto il processo di deflazione può amplificare notevolmente l'errore.

Oltre a questi esistono metodi che consentono di approssimare simultaneamente tutte le radici di un'equazione algebrica: i più usati sono il metodo di Durand e Kerner [8], [15], il metodo di Pasquini e Trigiante [21] e il metodo di Aberth [1]. Il metodo di Durand e Kerner (si veda l'esercizio 3.66) trae origine dalla dimostrazione che Weierstrass diede nel 1903 del teorema fondamentale dell'algebra, e per questo il metodo è talvolta indicato come metodo W. Il metodo presenta un basso costo computazionale per passo e convergenza localmente quadratica nel caso di radici distinte; non esitono però teoremi che assicurino la convegenza globale ed inoltre, nel caso di radici addensate, la convergenza è molto lenta. Il metodo di Pasquini e Trigiante (si veda l'esercizio 3.67) migliora, con un aumento del

costo computazionale ad ogni passo, la convergenza del metodo di Durand e Kerner: il metodo presenta una convergenza globale quadratica se le radici sono tutte reali e semplici, mentre se le radici non sono tutte distinte è possibile applicare una tecnica di accelerazione. Anche il metodo di Aberth (si veda l'esercizio 3.68) è derivato dal metodo di Durand e Kerner ed ha ordine 3. I metodi di Aberth e di Durand e Kerner possono essere applicati con una tecnica di tipo Gauss-Seidel, cioè utilizzando nel calcolo della i-esima componente le componenti di indice minore di i calcolate nella stessa iterazione; i metodi risultano allora avere ordine di convergenza più elevato.

Bibliografia

[1] O. Aberth, "Iteration Methods for Finding All Zeros of a Polynomial Simultaneously", *Math. Comp.*, 27, 1973, pp. 339-344.

[2] B. Barna, "Über die Divergenzpunkte des Newtonsches Verfahrens zur Bestimmung von Wurzeln algebraischen Gleichungen II", *Publicationes Mathematicae*, Debrecen, 4, 1956, pp. 384-397.

[3] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, Bologna, 1988.

[4] D. Bini, "Complexity of Parallel Polynomial Computations", *Proc. of International Meeting on Parallel Computing Methods, Algorithms and Applications*, Verona 28-30 Settembre 1988, D. J. Evans, C. Sutti Eds., Adam Hilger, Bristol, 1989.

[5] R. P. Brent, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, N. J., 1973.

[6] G. Dahlquist, Å. Björk, N. Anderson, *Numerical Methods*, Prentice Hall, Englewood Cliffs, N. J., 1974.

[7] T. J. Dekker, "Finding a zero by means of Successive Linear Interpolation", in B. Dejon, P. Henrici (Eds.), *Constructive Aspects of the Fundamental Theorem of Algebra*, Wiley Interscience, New York, 1969.

[8] E. Durand, *Solutions Numériques des Équations Algébriques, Tome 1*, Masson, Paris, 1960.

[9] A. Ghizzetti, *Lezioni di analisi matematica*, vol. I, Ed. Veschi, Roma, 1971.

[10] A. A. Grau, "On the Reduction of Number Range in the Use of the Gräffe Process", *J. ACM*, 10, 1963, pp. 538-544.

[11] P. Henrici, "Some Applications of the Quotient-Difference Algorithm", *Proc. of Symposium in Applied Math., A.M.S.*, 15, 1963, pp. 159-183.

[12] P. Henrici, *Applied and Computational Complex Analysis, vol. 1*, J. Wiley, New York, 1974.

[13] A. S. Householder, *The Numerical Treatment of a Single Non-linear Equation*, McGraw-Hill, New York, 1970.

[14] M. A. Jenkins, J. F. Traub, "A Three-Stage Variable-Shift Iteration for Polynomial Zeros and its Relation to Generalized Rayleigh Iteration", *Numer. Math.*, 14, 1970, pp. 252-263.

[15] I. O. Kerner, "Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen", *Numer. Math.*, 8, 1966, pp. 290-294.

[16] D. H. Lehmer, "A Machine Method for Solving Polynomial Equations", *J. ACM*, 8, 1961, pp. 151-162.

[17] M. Marden, *The Geometry of the Zeros of a Polynomial in a Complex Variable*, AMS Math. Surv. III, New York, 1949.

[18] J. M. Ortega, W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[19] A. M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1960.

[20] V. Pan, "Sequential and Parallel Complexity of Approximate Evaluation of Polynomial Zeros", *Computers and Math. (with Applications)*, 14, 1987, pp. 591-622.

[21] L. Pasquini, D. Trigiante, "A Globally Convergent Method for Simultaneously Finding Polynomial Roots", *Math. Comp.*, 44, 1985, pp. 135-149.

[22] A. Ralston, P. Rabinowitz, *A First Course in Numerical Analysis, 2nd Ed.*, Mc Graw-Hill, New York, 1978.

[23] H. Rutishauser, "Der Quotienten-Differenzen Algorithmus", *Z. Angew. Math. Phys.*, 5, 1954, pp. 233-251.

[24] H. Rutishauser, "On a Modification of the QD-Algorithm with Gräffe-Type Convergence", *Z. Angew. Math. Phys.*, 13, 1962, pp. 493-496.

[25] S. Smale, "The Fundamental Theorem of Algebra and Complexity Theory", *Bull. Amer. Math. Soc.*, 4, 1981, pp. 1-36.

[26] S. Smale, "On the Efficiency of Algorithms of Analysis", *Bull. Amer. Math. Soc.*, 13, 1985, pp. 87-121.

[27] J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, N. J., 1964.

[28] A. Van der Sluis, "Upperbounds for Roots of Polynomials", *Numer. Math.*, 15, 1970, pp. 250-262.

[29] J. H. Wilkinson, "The Evaluation of the Zeros of Ill-Conditioned Polynomials", *Numer. Math.*, 1, 1959, pp. 150-180.

[30] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, H. M. Stationary Office, London, 1963.

Capitolo 4

CALCOLO DELLE DIFFERENZE

1. Somme e serie

L'uso sempre crescente dei calcolatori ha richiesto uno studio sempre più approfondito delle tecniche matematiche che stanno alla base dei processi di discretizzazione. Le differenze finite, introdotte nel 17^o secolo proprio per il calcolo di funzioni, si prestano molto bene ad essere utilizzate in procedimenti e problemi discreti: da una parte le equazioni alle differenze sono fra le tecniche più usate per approssimare le soluzioni delle equazioni differenziali, dall'altra molti fenomeni naturali sono essenzialmente discreti e quindi sono meglio descritti da equazioni alle differenze che da equazioni differenziali.

Un semplice problema discreto, che si presenta in diversi campi scientifici, è quello della somma, finita o infinita, dei termini di una successione. Per questo problema l'utilizzazione del calcolatore può essere molto efficace, anche se si possono presentare delle sorprese.

4.1 Esempio. Un'approssimazione del valore di π può essere ottenuta sommando un certo numero di termini della serie di Taylor dell'arcotangente

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots,$$

per $x = 1$, cioè della serie numerica

$$\pi = 4 \sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1}\,. \tag{1}$$

Questa serie non è però adatta al calcolo effettivo di π, perché converge troppo lentamente. Infatti per ottenere un errore analitico inferiore a 10^{-6} sarebbe necessario sommare $\frac{1}{2}\ 10^6$ termini. Troppi, tenendo anche conto dell'accumulazione dell'errore di troncamento. Ad esempio, indicata con s_n la n-esima somma parziale della (1), i valori che si ottengono sono

n	s_n
10^1	3.232315
10^2	3.151484
10^3	3.142484
10^4	3.140576
10^5	3.130428
10^6	3.029699

Un'altra serie che non è opportuno utilizzare è

$$\log 2 = \sum_{i=0}^{\infty} \frac{(-1)^i}{i+1} \,. \tag{2}$$

Infatti con questa serie, anche scegliendo valori di n superiori a 10^6 si possono ottenere solo casualmente risultati affetti da errori inferiori a 10^{-3}.

∎

È quindi fondamentale fare uno studio preliminare della velocità di convergenza della serie che si intende utilizzare. Se la somma da calcolare è ottenuta troncando una serie di Taylor, la velocità di convergenza, che dipende dal punto in cui la serie viene calcolata, può essere valutata tramite il resto. Spesso è possibile ottenere una migliore velocità di convergenza sfruttando alcune proprietà della funzione che si vuole approssimare.

4.2 Esempio. Per approssimare il valore di $\sin x$ si possono sommare n termini della serie di Maclaurin

$$\sum_{i=0}^{n} (-1)^i \, \frac{x^{2i+1}}{(2i+1)!} \,. \tag{3}$$

Nella figura 4.1 sono riportati per l'intervallo $[0, \pi/2]$ i grafici dei moduli degli errori assoluti da cui sono affetti i valori effettivamente calcolati per $n = 0, 1, 2, 3$.

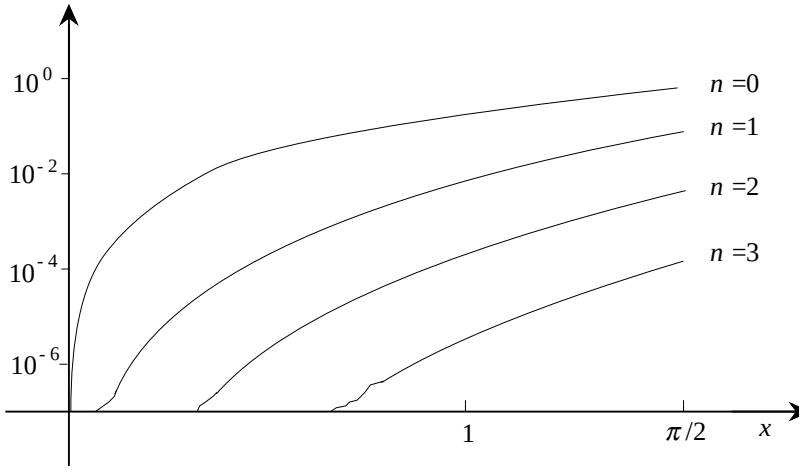


Fig. 4.1 - Errori dell'approssimazione di $\sin x$ con la (3) al variare di x.

Si noti come l'approssimazione, buona per x piccolo, non sia assolutamente accettabile per x grande. Naturalmente poiché la serie di Maclaurin di $\sin x$ converge per ogni x reale, aumentando n si possono ottenere resti comunque piccoli. Dal punto di vista pratico però, ad un elevato valore di n non solo corrisponde un elevato volume di calcolo, ma anche un'elevata propagazione degli errori di arrotondamento.

Nella figura 4.2 è riportato il grafico del modulo dell'errore relativo effettivamente generato nel punto $x = 6.5$ al variare di n. Dopo una diminuzione iniziale dell'errore dovuta alla diminuzione dell'errore analitico, per valori di $n > 12$, a causa della propagazione degli errori di arrotondamento, l'errore relativo si mantiene costante su un valore più elevato della precisione di macchina, che è di circa 10^{-6}.

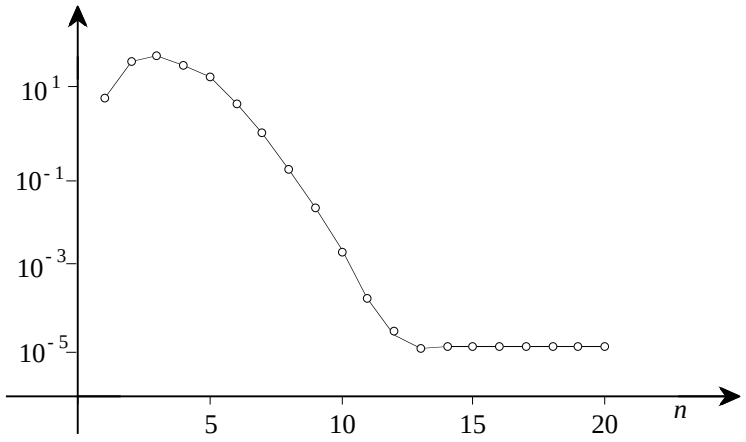


Fig. 4.2 - Errori dell'approssimazione di $\sin x$ per $x = 6.5$ con la (3) al variare di n.

Per calcolare il valore di $\sin x$ per x grande è però possibile sfruttare la periodicità della funzione, riconducendo il calcolo ad un argomento compreso fra 0 e $\dfrac{\pi}{2}$. In questo caso si ha

$$\sin 6.5 = \sin(6.5 - 2\pi) = \sin 0.2168146.$$

Per ottenere una buona approssimazione di questo valore basta ora sommare 3 termini della serie di Maclaurin, il valore che si ottiene è affetto da un errore di circa $0.179\ 10^{-6}$. Nel risultato ottenuto è presente anche un errore dovuto alla riduzione dell'argomento. ∎

In altri casi è possibile, con una trasformazione elementare dell'argomento, mettere in relazione la serie da calcolare con altre serie note o

più rapidamente convergenti. Ad esempio la serie

$$\sum_{i=1}^{\infty} \frac{1}{i^2+1}$$

converge molto lentamente, infatti per approssimarla con un errore minore di 10^{-4} occorre sommare 10^4 termini. Si può però utilizzare la relazione

$$\sum_{i=1}^{\infty} \frac{1}{i^2+1} = \sum_{i=1}^{\infty} \frac{1}{i^2} - \sum_{i=1}^{\infty} \frac{1}{i^2(i^2+1)} = \frac{\pi^2}{6} - \sum_{i=1}^{\infty} \frac{1}{i^2(i^2+1)}$$

(si veda il paragrafo 5), e quindi ricondurre il calcolo della serie data ad un'altra che converge assai più velocemente. Infatti con quest'ultima serie si ottiene un'approssimazione di 10^{-4} sommando 15 termini.

In alcuni casi, se la serie che si vuole calcolare converge troppo lentamente, la velocità di convergenza può essere aumentata con tecniche di *accelerazione*, come quella di Aitken, esposta nel capitolo 3. Altre tecniche di accelerazione sono quella di Eulero, presentata nel paragrafo 6 di questo capitolo e quella dell'economizzazione, descritta nel paragrafo 7 del capitolo 6.

Uno strumento classico per il calcolo di serie, soprattutto se con termini formati da funzioni razionali o elementari, è il *calcolo delle differenze*.

2. Operatore differenza

Sia $Y = \{y \ : \ \mathbf{Z} \ \rightarrow \mathbf{R}\}$ l'insieme delle funzioni di variabile intera a valori reali. Si definiscono l'operatore *differenza finita* $\Delta \ : \ Y \ \rightarrow \ Y$

$$\Delta y_k = y_{k+1} - y_k, \quad y_k \in Y,$$

e l'operatore *traslazione* $E \ : \ Y \ \rightarrow \ Y$

$$E y_k = y_{k+1}, \quad y_k \in Y.$$

I due operatori sono legati dalla relazione

$$\Delta y_k = E y_k - y_k.$$

Indicando con I l'operatore *identità*, formalmente si scrive

$$\Delta = E - I \tag{4}$$

o anche

$$E = \Delta + I.$$

Si definisce poi l'operatore *differenza finita di ordine r*

$$\Delta^r y_k = \Delta(\Delta^{r-1} y_k), \quad y_k \in Y,$$

intendendo che $\Delta^0 = I$ e $\Delta^1 = \Delta$. Anche l'operatore E può essere applicato ripetutamente e risulta

$$E^r y_k = E(E^{r-1} y_k) = \ldots = y_{k+r}, \quad y_k \in Y.$$

Gli operatori Δ e E sono lineari, cioè per $\alpha, \beta \in \mathbf{R}$ e $u_k, v_k \in Y$ risulta

$$\begin{aligned} \Delta(\alpha u_k + \beta v_k) &= \alpha \Delta u_k + \beta \Delta v_k, \\ E(\alpha u_k + \beta v_k) &= \alpha E u_k + \beta E v_k. \end{aligned} \tag{5}$$

Inoltre soddisfano le proprietà (si veda l'esercizio 4.1)

$$\begin{aligned} &\Delta E = E \Delta, \\ &E^r E^s = E^{r+s}, \\ &\Delta^r \Delta^s = \Delta^{r+s}. \end{aligned}$$

Poiché l'operatore identità commuta con qualsiasi operatore su Y, per la (4) è

$$\Delta^r = (E - I)^r = \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} E^i,$$

e analogamente

$$E^r = (\Delta + 1)^r = \sum_{i=0}^{r} \binom{r}{i} \Delta^i,$$

e quindi

$$\begin{aligned} \Delta^r y_k &= \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} E^i y_k = \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} y_{k+i}, \\ y_{k+r} &= E^r y_k = \sum_{i=0}^{r} \binom{r}{i} \Delta^i y_k. \end{aligned} \tag{6}$$

Applicando l'operatore Δ al prodotto $u_k v_k$, si ha

$$\begin{aligned} \Delta(u_k v_k) &= u_{k+1} v_{k+1} - u_k v_k = u_{k+1} v_{k+1} - u_k v_{k+1} + u_k v_{k+1} - u_k v_k \\ &= (\Delta u_k) v_{k+1} + u_k (\Delta v_k), \end{aligned} \tag{7}$$

e applicando Δ ad un quoziente si ha

$$\Delta \frac{u_k}{v_k} = \frac{u_{k+1}}{v_{k+1}} - \frac{u_k}{v_k} = \frac{(\Delta u_k)v_k - u_k(\Delta v_k)}{v_k v_{k+1}}\,. \tag{8}$$

Data una funzione $f : \mathbf{R} \to \mathbf{R}$ e fissato un numero reale h, è possibile associare ad f la funzione $y \in Y$, definita da

$$y_k = f(x + kh), \tag{9}$$

e quindi è possibile applicare alla funzione $f(x + kh)$ gli operatori Δ ed E. Nel seguito, per semplicità, si userà spesso il valore $h = 1$

$$Ef(x) = f(x + 1), \quad \Delta f(x) = f(x + 1) - f(x).$$

Dalle proprietà (5), (7) e (8) scaturisce una certa analogia fra l'operatore differenza Δ e l'operatore di derivazione. Tale analogia viene ancora più evidenziata se si considerano le *potenze fattoriali* di x, così definite per m intero positivo:

$$\begin{aligned} x^{(0)} &= 1, \\ x^{(m)} &= (x - m + 1)x^{(m-1)}, \\ x^{(-m)} &= \frac{x^{(-m+1)}}{(x + m)}\,. \end{aligned} \tag{10}$$

Si ha quindi

$$x^{(m)} = x(x-1)\dots(x-m+1), \quad x^{(-m)} = \frac{1}{(x+1)(x+2)\dots(x+m)}\,.$$

Risulta

$$\begin{aligned} \Delta x^{(m)} &= (x+1)^{(m)} - x^{(m)} \\ &= (x+1)x\dots(x-m+2) - x(x-1)\dots(x-m+1) \\ &= (x+1)x^{(m-1)} - x^{(m-1)}(x-m+1) = mx^{(m-1)}, \end{aligned} \tag{11}$$

e analogamente

$$\Delta x^{(-m)} = -mx^{(-m-1)}. \tag{12}$$

Dalla (11) si ottiene

$$\Delta^m x^{(m)} = m! \tag{13}$$

e quindi

$$\Delta^{m+1} x^{(m)} = 0. \tag{14}$$

Le potenze fattoriali intervengono quando si calcolano le derivate dei polinomi:

$$\frac{d^m}{dx^m}\, x^k = k^{(m)} x^{k-m}. \tag{15}$$

Poiché le potenze fattoriali $x^{(i)}$, $i = 0, 1, \ldots$, costituiscono una base per lo spazio dei polinomi in x, ogni polinomio $p_n(x)$ di grado n in x può essere rappresentato come combinazione lineare delle potenze fattoriali $x^{(i)}$, con $0 \le i \le n$:

$$p_n(x) = \sum_{i=0}^{n} a_i x^i = \sum_{i=0}^{n} b_i x^{(i)}. \tag{16}$$

I coefficienti a_i possono essere ricavati dai coefficienti b_i, e viceversa, con vari metodi.

a) Per mezzo dei *numeri di Stirling*. Dalle (10) si ottiene

$$x^{(m)} = \sum_{i=1}^{m} s_i^{(m)} x^i, \quad m \ge 1. \tag{17}$$

I numeri $s_i^{(m)}$, $i = 1, \ldots, m$ sono detti *numeri di Stirling di prima specie*. Ponendo

$$s_1^{(1)} = 1, \quad s_0^{(m)} = 0 \quad \text{e} \quad s_{m+1}^{(m)} = 0 \quad \text{per} \quad m = 1, 2, \ldots,$$

vale la formula di ricorrenza (si veda l'esercizio 4.9):

$$s_i^{(m+1)} = s_{i-1}^{(m)} - m s_i^{(m)} \quad i = 1, \ldots, m+1, \quad m = 1, 2, \ldots \tag{18}$$

Applicando la (18) si ottiene la tabella dei numeri di Stirling di prima specie per $m \le 5$, riportati in figura 4.3.

m | $i =$	1	2	3	4	5
1	1				
2	−1	1			
3	2	−3	1		
4	−6	11	−6	1	
5	24	−50	35	−10	1

Fig. 4.3 - Numeri di Stirling di prima specie.

Viceversa, dalla (17) si possono ricavare le potenze x^i dalle potenze fattoriali $x^{(m)}$: si ottiene

$$x^i = \sum_{m=1}^{i} S_m^{(i)} x^{(m)}, \quad i \ge 1. \tag{19}$$

I numeri $S_m^{(i)}$, $m = 1, \ldots, i$ sono detti *numeri di Stirling di seconda specie*. Ponendo

$$S_1^{(1)} = 1, \quad S_0^{(i)} = 0 \quad \text{e} \quad S_{i+1}^{(i)} = 0 \quad \text{per} \quad i = 1, 2, \ldots,$$

vale la formula di ricorrenza (si veda l'esercizio 4.9):

$$S_m^{(i+1)} = S_{m-1}^{(i)} + m S_m^{(i)}, \quad m = 1, \ldots, i+1, \quad i = 1, 2, \ldots \tag{20}$$

Applicando la (20) si ottiene la tabella dei numeri di Stirling di seconda specie per $i \leq 5$, riportata in figura 4.4. Si noti che $S_m^{(m)} = 1$ per ogni m, e quindi nella (16) risulta $b_n = a_n$.

i	$m =$	1	2	3	4	5
1		1				
2		1	1			
3		1	3	1		
4		1	7	6	1	
5		1	15	25	10	1

Fig. 4.4 - Numeri di Stirling di seconda specie.

b) Applicando il metodo di Ruffini-Horner: poiché

$$\sum_{i=0}^{n} b_i x^{(i)} = [\ldots[b_n(x-n+1) + b_{n-1}](x-n+2) + \ldots]x + b_0,$$

i coefficienti b_i si ottengono come resti delle divisioni del polinomio $p_n(x)$ della (16) e dei successivi quozienti per $x - i$, $i = 0, 1, \ldots, n-1$, nel modo seguente:

b_i è il resto della divisione fra $p_{n-i}(x)$ e $x - i$,

$p_{n-i-1}(x)$ è il quoziente della divisione fra $p_{n-i}(x)$ e $x - i$,

e risulta $b_n = p_0(x)$.

4.3 Esempio. Per trasformare il polinomio

$$p_3(x) = 5x^3 - 2x^2 + 3x + 2$$

in combinazione di potenze fattoriali, se si utilizzano i numeri di Stirling di seconda specie si ha

$$\begin{aligned} p_3(x) &= 5\big[x^{(3)} + 3x^{(2)} + x^{(1)}\big] - 2\big[x^{(2)} + x^{(1)}\big] + 3x^{(1)} + 2 \\ &= 5x^{(3)} + 13x^{(2)} + 6x^{(1)} + 2x^{(0)}. \end{aligned}$$

Se invece si applica il metodo di Ruffini-Horner si ha

$p_3(x) \rightarrow$		5	-2	3	2	
	0		0	0	0	
$p_2(x) \rightarrow$		5	-2	3	2	$\leftarrow$ primo resto $= b_0$
	1		5	3		
$p_1(x) \rightarrow$		5	3	6		$\leftarrow$ secondo resto $= b_1$
	1		10			
		5	13			$\leftarrow$ terzo resto $= b_2$
		$\uparrow$				
		$p_0(x) = b_3$				

da cui

$$p_3(x) = 5x^{(3)} + 13x^{(2)} + 6x^{(1)} + 2x^{(0)}. \qquad \blacksquare$$

Dalla (16), per la linearità dell'operatore Δ, per un polinomio $p_n(x)$ di grado n si ha

$$\Delta^m p_n(x) = \sum_{i=0}^{n} b_i \Delta^m x^{(i)},$$

e per le (13) e (14)

$$\Delta^n p_n(x) = b_n n!$$

e

$$\Delta^{n+1} p_n(x) = 0, \tag{21}$$

cioè la differenza finita di ordine m di un polinomio di grado n è un polinomio di grado $n - m$ se $n \geq m$, ed è nulla se $m > n$.

3. Operatore somma

L'operatore Δ non è iniettivo, infatti

$$\Delta[y_k + c_k] = \Delta y_k, \quad \text{se} \quad c_{k+1} = c_k,$$

cioè $c_k \in Y$ è una funzione costante. Quindi in generale l'operatore Δ non può essere invertito. È possibile però definire, a meno di una funzione costante arbitraria c_k, detta *costante periodica*, un operatore $\sum$ tale che

$$y_k = \sum z_k, \quad \text{se} \quad z_k = \Delta y_k.$$

Il termine di costante periodica è motivato dal fatto che per la (9) una funzione c_k costante di Y corrisponde ad una funzione $f(x)$ periodica di periodo h.

L'operatore $\sum$ viene detto operatore *antidifferenza* o operatore *somma*. Alla naturale analogia fra l'operatore Δ e l'operatore di derivazione corrisponde una simile analogia fra l'operatore $\sum$ e l'operatore di integrazione. Questo è messo in evidenza dalle relazioni che seguono e che possono essere verificate direttamente.

a) $\sum$ è un operatore lineare, cioè

$$\sum(\alpha y_k + \beta v_k) = \alpha \sum y_k + \beta \sum v_k$$

b) vale la relazione, analoga alla relazione di integrazione per parti, di *somma per parti*

$$\sum(y_k \Delta v_k) = y_k v_k - \sum(v_{k+1} \Delta y_k), \tag{22}$$

che si ricava dalla (7);

c) per m intero, $m \neq -1$, è

$$\sum x^{(m)} = \frac{x^{(m+1)}}{m+1}\,; \tag{23}$$

d) in analogia al teorema fondamentale del calcolo integrale, vale per $n > m$

$$\sum_{k=m}^{n-1} y_k = \sum y_k \Big|_m^n\,, \tag{24}$$

dove la *sommatoria*

$$\sum_{k=m}^{n-1} y_k = y_m + y_{m+1} + \ldots + y_{n-1}$$

è detta anche *somma definita*. Per dimostrare la (24), sia z_k tale che

$$\Delta z_k = y_k, \quad \text{cioè} \quad z_k = \sum y_k + c_k;$$

allora sostituendo nel primo membro della (24) si ha

$$\begin{aligned}\sum_{k=m}^{n-1} y_k &= \sum_{k=m}^{n-1} \Delta z_k = \Delta z_m + \Delta z_{m+1} + \ldots + \Delta z_{n-1} \\ &= z_{m+1} - z_m + z_{m+2} - z_{m+1} + \ldots + z_n - z_{n-1} = z_n - z_m.\end{aligned}$$

4.4 Esempio. Per calcolare la

$$\sum_{k=1}^{n} k^2,$$

si esprime il polinomio k^2 come combinazione di potenze fattoriali. Si ha

$$k^2 = k^{(2)} + k^{(1)},$$

e per la linearità di $\sum$ e per la (23) risulta

$$\begin{aligned}\sum k^2 &= \sum k^{(2)} + \sum k^{(1)} = \frac{k^{(3)}}{3} + \frac{k^{(2)}}{2} \\ &= \frac{k(k-1)(k-2)}{3} + \frac{k(k-1)}{2} = \frac{k(k-1)(2k-1)}{6}\,.\end{aligned}$$

Dalla (24) segue

$$\sum_{k=1}^{n} k^2 = \frac{k(k-1)(2k-1)}{6}\Big|_1^{n+1} = \frac{n(n+1)(2n+1)}{6}\,. \qquad \blacksquare$$

4.5 Esempio. Per calcolare

$$\sum_{k=0}^{n-1} \alpha^k, \quad \alpha \neq 0, 1,$$

si determina una successione z_k tale che

$$\Delta z_k = \alpha^k.$$

Poiché

$$\alpha^{k+1} - \alpha^k = \alpha^k(\alpha - 1),$$

si pone

$$z_k = \frac{\alpha^k}{\alpha - 1}\,, \tag{25}$$

per cui

$$\sum \alpha_k = \frac{\alpha^k}{\alpha - 1} + c_k$$

e

$$\sum_{k=0}^{n-1} \alpha_k = \frac{\alpha^n - 1}{\alpha - 1}\,. \qquad \blacksquare$$

4.6 Esempio. Per calcolare

$$\sum_{k=1}^{n} k\alpha^k, \quad \alpha \neq 0, 1,$$

si fa uso della relazione (22), ponendo

$$y_k = k, \quad \Delta v_k = \alpha^k,$$

e quindi per la (25)

$$v_k = \frac{\alpha^k}{\alpha - 1}\,.$$

Si ottiene

$$\begin{aligned}\sum k\alpha^k &= k\frac{\alpha^k}{\alpha - 1} - \sum \frac{\alpha^{k+1}}{\alpha - 1} = \frac{1}{\alpha - 1}[k\alpha^k - \alpha \sum \alpha^k] \\ &= \frac{1}{\alpha - 1}\big[k\alpha^k - \frac{\alpha^{k+1}}{\alpha - 1}\big]\end{aligned}$$

per cui

$$\sum_{k=1}^{n} k\alpha^k = \frac{1}{\alpha - 1}\left[k\alpha^k - \frac{\alpha^{k+1}}{\alpha - 1}\right]_1^{n+1} = \frac{n\alpha^{n+1}}{\alpha - 1} - \alpha\,\frac{\alpha^n - 1}{(\alpha - 1)^2}\,. \qquad \blacksquare$$

4. Funzione gamma

Per generalizzare la funzione $k^{(m)}$ al caso in cui m non è un numero intero, si introduce la *funzione gamma (di Eulero)*, definita da

$$\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}dt, \quad x \text{ reale, } x \neq 0, -1, -2, \ldots \tag{26}$$

Applicando alla (26) la formula di integrazione per parti si ha

$$\Gamma(x+1) = \int_0^\infty t^x\mathrm{e}^{-t}dt = -t^x\mathrm{e}^{-t}\Big|_0^\infty + x\int_0^\infty t^{x-1}\mathrm{e}^{-t}dt = x\Gamma(x)$$

e quindi per la funzione $\Gamma(x)$ vale la relazione di ricorrenza

$$\Gamma(x+1) = x\Gamma(x). \tag{27}$$

Poiché

$$\Gamma(1) = 1,$$

ne segue che per $x = k$ intero positivo è

$$\Gamma(k+1) = k!$$

Il grafico della funzione $\Gamma(x)$ per $-5 \leq x \leq 5$ è riportato nella figura 4.5.

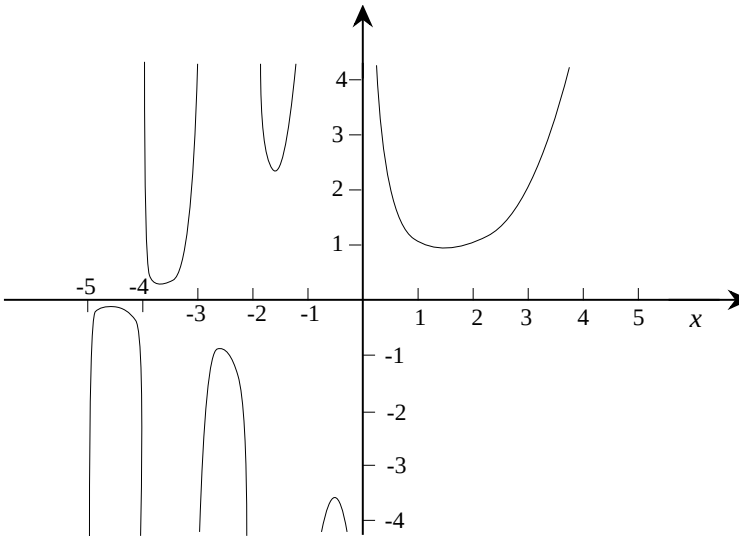


Fig. 4.5 - Grafico della funzione $\Gamma(x)$.

Applicando ripetutamente la (27) si ha

$$\begin{aligned}\Gamma(x+1) &= x\Gamma(x) = x(x-1)\Gamma(x-1) = \dots \\ &= x(x-1)\dots(x-m+1)\Gamma(x-m+1),\end{aligned}$$

da cui

$$\frac{\Gamma(x+1)}{\Gamma(x-m+1)} = x(x-1)\dots(x-m+1) \tag{28}$$

e la (28) può essere assunta come definizione di $x^{(m)}$ per m non intero, cioè

$$x^{(m)} = \frac{\Gamma(x+1)}{\Gamma(x-m+1)}\,. \tag{29}$$

La (29) vale anche per m negativo.

La (11) e la (12) valgono anche nel caso che m non sia intero. Infatti per la (11) si ha

$$\begin{aligned}\Delta x^{(m)} &= (x+1)^{(m)} - x^{(m)} = \frac{\Gamma(x+2)}{\Gamma(x-m+2)} - \frac{\Gamma(x+1)}{\Gamma(x-m+1)} \\ &= \frac{(x+1)\Gamma(x+1)}{\Gamma(x-m+2)} - \frac{(x-m+1)\Gamma(x+1)}{\Gamma(x-m+2)} = \frac{m\Gamma(x+1)}{\Gamma(x-m+2)} \\ &= \frac{m\Gamma(x+1)}{\Gamma(x-(m-1)+1)} = mx^{(m-1)},\end{aligned}$$

e quindi risulta dimostrata la (11); analogamente si ricava la (12). Ne segue che la (23) vale per ogni m reale diverso da -1.

Per $m = -1$, si consideri la funzione

$$\Psi(x) = \frac{d}{dx}\,\log\Gamma(x+1) = \frac{\Gamma'(x+1)}{\Gamma(x+1)}\,,$$

detta *funzione digamma*. Si ha allora

$$\begin{aligned}\Delta\Psi(x) = \Psi(x+1) - \Psi(x) &= \frac{d}{d(x+1)}\log\Gamma(x+2) - \frac{d}{dx}\log\Gamma(x+1) \\ &= \frac{d}{dx}\log\Gamma(x+2) - \frac{d}{dx}\log\Gamma(x+1) = \frac{d}{dx}\log\frac{\Gamma(x+2)}{\Gamma(x+1)} \\ &= \frac{d}{dx}\log(x+1) = \frac{1}{x+1} = x^{(-1)},\end{aligned}$$

da cui segue che, a meno di una costante periodica, è

$$\sum x^{(-1)} = \Psi(x).$$

Della funzione digamma si possono dare diverse rappresentazioni integrali, che consentono di ricavarne molte proprietà. Una delle più importanti è *la formula di Gauss* (per la dimostrazione si veda [6])

$$\Psi(x) = \int_0^\infty \left[\ \frac{\mathrm{e}^{-t}}{t} \ - \ \frac{\mathrm{e}^{-t(x+1)}}{1-\mathrm{e}^{-t}}\ \right] dt, \quad x > -1.$$

Il grafico della funzione $\Psi(x)$ è riportato nella figura 4.6 per $-2 < x < 5$. Tratteggiato nella stessa figura è il grafico della funzione $\log x$, che per x grande ha lo stesso andamento della funzione $\Psi(x)$ (si veda l'esercizio 4.24). La quantità

$$\gamma = -\Gamma'(1) = -\int_0^\infty \mathrm{e}^{-t}\log t\ dt = -\Psi(0)$$

è detta *costante di Eulero* e riveste notevole importanza nella teoria dell'approssimazione di funzioni. Il valore approssimato di γ è $0.5772156\ldots$

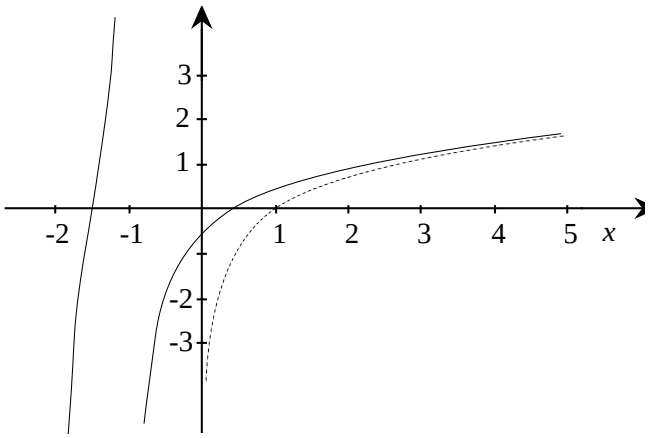


Fig. 4.6 - Grafico della funzione $\Psi(x)$.

5. Polinomi di Bernoulli

Con i polinomi e i numeri di Bernoulli si possono risolvere molti problemi di calcolo di somme: in particolare le somme di termini polinomiali e razionali; inoltre i numeri di Bernoulli compaiono come coefficienti di alcune serie di Taylor di funzioni elementari.

La funzione di t

$$f(t;x) = \begin{cases} \dfrac{t\mathrm{e}^{xt}}{\mathrm{e}^t - 1}, & \text{per } t \neq 0, \\ 1 & \text{per } t = 0, \end{cases}$$

è sviluppabile in serie di potenze in un intorno di $t = 0$ e si ha

$$f(t;x) = \sum_{i=0}^{\infty} B_i(x)\, \frac{t^i}{i!}\,. \tag{30}$$

Le funzioni $B_i(x)$ sono polinomi di grado i, detti *polinomi di Bernoulli*, e i numeri $B_i = B_i(0)$ sono detti *numeri di Bernoulli*. Per calcolare i numeri di Bernoulli si procede con il metodo seguente: ponendo $x = 0$ nella (30) si ha per $|t| < 2\pi$

$$\frac{t}{\mathrm{e}^t - 1} = \sum_{i=0}^{\infty} B_i\, \frac{t^i}{i!}\,; \tag{31}$$

dalla formula di Maclaurin per e^t è

$$t = \left[\sum_{j=1}^{\infty} \frac{t^j}{j!}\right] \left[\sum_{i=0}^{\infty} B_i\, \frac{t^i}{i!}\right] = \sum_{k=1}^{\infty} c_k\, \frac{t^k}{k!}\,,$$

dove

$$c_k = \sum_{i=0}^{k-1} \frac{B_i\ k!}{i!(k-i)!} = \sum_{i=0}^{k-1} \binom{k}{i} B_i, \tag{32}$$

e uguagliando i coefficienti delle successive potenze t^k si ha

$$\begin{aligned} c_1 &= B_0 = 1 \\ c_k &= \sum_{i=0}^{k-1} \binom{k}{i} B_i = 0, \quad \text{per } k = 2, 3, \dots \end{aligned} \tag{33}$$

I B_i si ottengono quindi risolvendo per sostituzione il sistema lineare triangolare (33). I primi valori sono dati da

$$B_0 = 1, \quad B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_4 = -\frac{1}{30}, \quad B_6 = \frac{1}{42}, \quad \dots,$$
$$B_{2i+1} = 0 \quad \text{per } i = 1, 2, \dots$$

Sostituendo le (31) nella (30), dalla formula di Maclaurin di e^{xt} si ha

$$\left[\sum_{j=0}^{\infty} B_j\ \frac{t^j}{j!}\right]\ \left[\sum_{i=0}^{\infty} \frac{x^i t^i}{i!}\right] = \sum_{r=0}^{\infty} B_r(x)\ \frac{t^r}{r!},$$

e ponendo $r = i + j$, si ha

$$\sum_{r=0}^{\infty} \left[\sum_{i=0}^{r} B_{r-i}\ \frac{x^i}{i!(r-i)!}\right] t^r = \sum_{r=0}^{\infty} B_r(x)\ \frac{t^r}{r!}.$$

Tenendo conto che

$$\frac{1}{i!(r-i)!} = \frac{1}{r!} \binom{r}{i},$$

si ottiene la relazione

$$B_r(x) = \sum_{i=0}^{r} \binom{r}{i} B_{r-i}\ x^i, \tag{34}$$

che esprime i coefficienti dei polinomi di Bernoulli per mezzo dei numeri di Bernoulli. I primi polinomi di Bernoulli sono

$$\begin{aligned} B_0(x) &= 1, \\ B_1(x) &= x - \frac{1}{2}, \\ B_2(x) &= x^2 - x + \frac{1}{6} \\ B_3(x) &= x^3 - \frac{3}{2}\,x^2 + \frac{1}{2}\,x, \\ B_4(x) &= x^4 - 2x^3 + x^2 - \frac{1}{30}, \\ B_5(x) &= x^5 - \frac{5}{2}\,x^4 + \frac{5}{3}\,x^3 - \frac{1}{6}\,x. \end{aligned}$$

Per ogni x i polinomi di Bernoulli verificano la relazione

$$B_r(x+1) - B_r(x) = rx^{r-1}. \tag{35}$$

Si ha infatti dalla (34)

$$\begin{aligned} B_r(x+1) &= \sum_{i=0}^{r} \binom{r}{i} B_{r-i}(x+1)^i = \sum_{i=0}^{r} \binom{r}{i} B_{r-i} \sum_{j=0}^{i} \binom{i}{j} x^j \\ &= \sum_{i=0}^{r} \binom{r}{i} B_{r-i}\, x^i + \sum_{i=1}^{r} \binom{r}{i} B_{r-i} \sum_{j=0}^{i-1} \binom{i}{j} x^j \\ &= B_r(x) + \sum_{j=0}^{r-1} \left\{ \binom{r}{j} x^j \left[\sum_{i=j+1}^{r} \binom{r-j}{r-i} B_{r-i} \right] \right\}, \end{aligned} \tag{36}$$

in quanto

$$\binom{r}{i}\binom{i}{j} = \binom{r}{j}\binom{r-j}{r-i}.$$

Poiché per la (32) e la (33) è

$$\sum_{i=j+1}^{r} \binom{r-j}{r-i} B_{r-i} = \sum_{s=0}^{r-j-1} \binom{r-j}{s} B_s = c_{r-j} = \begin{cases} 1 & \text{per } j = r-1, \\ 0 & \text{per } j = 0, \ldots, r-2, \end{cases}$$

dalla (36) segue la (35).

Dalla (35) si ricavano le relazioni seguenti:

$$\sum_{k=1}^{n} k^r = \frac{B_{r+1}(n+1) - B_{r+1}(0)}{r+1}, \tag{37}$$

$$\sum_{k=1}^{\infty} \frac{1}{k^{2i}} = \frac{(-1)^{i+1}(2\pi)^{2i} B_{2i}}{2\,(2i)!}. \tag{38}$$

Per dimostrare la (37) si ha dalla (35)

$$\Delta B_{r+1}(x) = (r+1)x^r,$$

e quindi, a meno di una costante periodica, è

$$\sum x^r = \frac{1}{r+1}\, B_{r+1}(x), \tag{39}$$

e dalla (24) si ha

$$\sum_{k=0}^{n} k^r = \frac{1}{r+1}\, B_{r+1}(k)\Big|_0^{n+1} = \frac{B_{r+1}(n+1) - B_{r+1}(0)}{r+1}\,.$$

In particolare, per $r = 1, 2, 3$

$$\sum_{k=1}^{n} k = \frac{B_2(n+1) - B_2(0)}{2} = \frac{1}{2}\,[(n+1)^2 - (n+1)] = \frac{1}{2}\,n(n+1),$$

$$\sum_{k=1}^{n} k^2 = \frac{B_3(n+1) - B_3(0)}{3} = \frac{1}{3}\,\Big[(n+1)^3 - \frac{3}{2}\,(n+1)^2 + \frac{1}{2}\,(n+1)\Big]$$

$$= \frac{1}{3}\,n(n+1)\big(n+\frac{1}{2}\big),$$

$$\sum_{k=1}^{n} k^3 = \frac{B_4(n+1) - B_4(0)}{4} = \frac{1}{4}[(n+1)^4 - 2(n+1)^3 + (n+1)^2]$$

$$= \frac{1}{4}\,n^2(n+1)^2.$$

Per dimostrare la (38) si esprime la funzione $B_r(x)$, per $0 \le x < 1$, in serie di Fourier nel modo seguente (si veda l'esercizio 4.33)

$$B_r(x) = \begin{cases} (-1)^{r/2+1}\dfrac{2\,r!}{(2\pi)^r}\displaystyle\sum_{k=1}^{\infty}\frac{\cos 2\pi kx}{k^r} & \text{per } r \text{ pari,} \\[2ex] (-1)^{(r+1)/2}\dfrac{2\,r!}{(2\pi)^r}\displaystyle\sum_{k=1}^{\infty}\frac{\sin 2\pi kx}{k^r} & \text{per } r \text{ dispari.} \end{cases}$$

Allora per r pari, $r = 2i$, si ha in $x = 0$

$$B_{2i} = B_{2i}(0) = (-1)^{i+1}\frac{2\,(2i)!}{(2\pi)^{2i}}\sum_{k=1}^{\infty}\frac{1}{k^{2i}}\,,$$

da cui si ricava la (38). In particolare si ha

$$\text{per } i = 1 \qquad \sum_{k=1}^{\infty}\frac{1}{k^2} \;=\; \frac{\pi^2}{6}\,, \tag{40}$$

$$\text{per } i = 2 \qquad \sum_{k=1}^{\infty}\frac{1}{k^4} \;=\; \frac{\pi^4}{90}\,.$$

6. Trasformazione di Eulero

Per accelerare la convergenza di una serie che converge lentamente si possono usare varie tecniche; nel capitolo 3 si è esaminata la trasformazione di Aitken, in questo paragrafo si esamina la *trasformazione di Eulero*, che si applica a serie a segni alterni e che consente di trasformare una serie convergente in un'altra convergente allo stesso limite, che sotto opportune condizioni (si veda l'esercizio 4.39) ha una maggiore velocità di convergenza. La trasformazione di Eulero consiste nella sostituzione dei termini della serie con differenze finite di ordine crescente, e si basa sul seguente teorema.

4.7 Teorema. *Siano*

$$\sigma_n = \sum_{k=0}^{n-1} (-1)^k a_k \tag{41}$$

e

$$\tau_n = \frac{1}{2} \sum_{k=0}^{n-1} \big(-\frac{1}{2}\big)^k \Delta^k a_0. \tag{42}$$

Se la prima serie è convergente, anche la seconda è convergente e hanno lo stesso limite.

Dim. Si dimostra prima per induzione su n che

$$\tau_n = \frac{1}{2^n} \sum_{i=0}^{n} \binom{n}{i} \sigma_i, \tag{43}$$

in cui $\sigma_0 = \tau_0 = 0$. La (43) è ovvia per $n = 0$. Per $n > 0$ si ha per la (6)

$$\tau_n = \tau_{n-1} + \frac{1}{2} \big(-\frac{1}{2}\big)^{n-1} \Delta^{n-1} a_0 = \tau_{n-1} + \frac{1}{2^n} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i a_i.$$

Per l'ipotesi induttiva risulta allora

$$\begin{aligned}
\tau_n &= \frac{1}{2^{n-1}} \sum_{i=0}^{n-1} \binom{n-1}{i} \sigma_i + \frac{1}{2^n} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i a_i \\
&= \frac{1}{2^n} \left[2 \sum_{i=0}^{n-1} \binom{n-1}{i} \sigma_i + \sum_{i=0}^{n-1} \binom{n-1}{i} (\sigma_{i+1} - \sigma_i) \right] \\
&= \frac{1}{2^n} \left[\sum_{i=0}^{n-1} \binom{n-1}{i} \sigma_i + \sum_{i=0}^{n-1} \binom{n-1}{i} \sigma_{i+1} \right] \\
&= \frac{1}{2^n} \left\{ \sigma_0 + \sum_{i=1}^{n-1} \left[\binom{n-1}{i} + \binom{n-1}{i-1} \right] \sigma_i + \sigma_n \right\} = \frac{1}{2^n} \sum_{i=0}^{n} \binom{n}{i} \sigma_i.
\end{aligned}$$

La (43) risulta quindi dimostrata. La convergenza della (43) per $n \to \infty$ allo stesso limite della (41) segue da un teorema generale sulla convergenza delle serie (si veda l'esercizio 4.38). ∎

Con la trasformazione di Eulero la somma (41) viene calcolata per mezzo della (42). In molti casi in cui la (41) converge più lentamente di una serie geometrica di ragione 1/2, la (42) converge più rapidamente (si veda l'esercizio 4.39).

4.8 Esempio. Applicando la trasformazione di Eulero alla serie (1) per il calcolo di π, in cui

$$a_k = \frac{4}{2k+1}\,,$$

si ha

$$\pi = \frac{1}{2}\left[a_0 - \frac{1}{2}\,\Delta a_0 + \frac{1}{4}\,\Delta^2 a_0 - \frac{1}{8}\,\Delta^3 a_0 + \frac{1}{16}\,\Delta^4 a_0 + \ldots\right],$$

dove

$$\begin{aligned}
&a_0 = 4,\\
&\Delta a_0 = a_1 - a_0 = \frac{4}{3} - 4 = -\frac{8}{3}\,,\\
&\Delta^2 a_0 = \Delta a_1 - \Delta a_0 = \left(\frac{4}{5} - \frac{4}{3}\right) + \frac{8}{3} = \frac{32}{15}\,.\\
&\ldots
\end{aligned}$$

Si dispone il calcolo nel modo seguente

i	a_i	Δa_i	$\Delta^2 a_i$	$\Delta^3 a_i$	$\Delta^4 a_i$
0	4				
		$-\frac{8}{3}$			
1	$\frac{4}{3}$		$\frac{32}{15}$		
		$-\frac{8}{15}$		$-\frac{64}{35}$	
2	$\frac{4}{5}$		$\frac{32}{105}$		$\frac{512}{315}$
		$-\frac{8}{35}$		$-\frac{64}{315}$	
3	$\frac{4}{7}$		$\frac{32}{315}$		
		$-\frac{8}{63}$			
4	$\frac{4}{9}$				

per cui

$$\pi = 2\left[1 + \frac{1}{2}\,\frac{2}{3} + \frac{1}{4}\,\frac{8}{15} + \frac{1}{8}\,\frac{16}{35} + \frac{1}{16}\,\frac{128}{315} + \ldots\right]$$

$$= 2\left[1 + \frac{1}{3} + \frac{1\cdot 2}{3\cdot 5} + \frac{1\cdot 2\cdot 3}{3\cdot 5\cdot 7} + \frac{1\cdot 2\cdot 3\cdot 4}{3\cdot 5\cdot 7\cdot 9} + \ldots\right].$$

Questa serie converge molto più rapidamente della (1), consentendo di ottenere un'approssimazione di π con un errore analitico inferiore a 10^{-6} con soli 18 termini. ∎

Talvolta è conveniente applicare la trasformazione di Eulero a partire non dal primo termine ma da uno successivo nel modo seguente:

$$\sigma_n = \sigma_r + (-1)^r \sum_{k=0}^{n-1-r} (-1)^k a_{k+r},$$

trasformando solo la seconda sommatoria. Il numero r viene detto *ritardo*.

4.9 Esempio. La serie

$$\log 2 = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1}$$

converge lentamente. Applicando la trasformazione di Eulero con ritardo r si ottiene un'approssimazione di $\log 2$ con un errore inferiore a 10^{-6} con N termini, come riportato nella seguente tabella, comprendendo in N anche gli r termini richiesti dal calcolo di σ_r.

r	N	r	N
0	17	6	12
1	13	7	13
2	13	8	14
3	12	9	14
4	12	10	15
5	12		

∎

7. Equazioni alle differenze lineari

Sia $f : \mathbf{N} \times \mathbf{R}^{n+1} \to \mathbf{R}$; si definisce *equazione alle differenze di ordine* n una relazione del tipo

$$f(k, y_k, y_{k+1}, \ldots, y_{k+n}) = 0. \tag{44}$$

Una funzione di variabile intera $y : Z' \to \mathbf{R}$, $Z' \subseteq \mathbf{Z}$, è una soluzione dell'equazione (44) se y_k verifica la (44) per ogni $k \in Z'$. Se la funzione f è lineare rispetto a y_i, $i = k, k+1, \ldots, k+n$, l'equazione si dice *lineare*, e può essere scritta nella forma

$$\sum_{j=0}^{n} a_j(k)\ y_{k+j} = b(k), \tag{45}$$

in cui i coefficienti $a_j(k)$, $j = 0, 1, \ldots, n$ e $b(k)$ sono funzioni di k e $a_0(k)$ e $a_n(k)$ non sono identicamente nulli.

Per semplicità nel seguito si considera solo il caso in cui la soluzione y_k è definita nell'insieme $Z' = \{\ k \in \mathbf{Z},\ k \geq k_0\ \}$. I risultati che si ottengono valgono anche nel caso in cui l'insieme Z' sia costituito dagli interi minori o uguali a k_0. Si assume che le funzioni $a_j(k)$, $j = 0, \ldots, n$, e $b(k)$ siano definite su Z'. Fissato k_0, una soluzione dell'equazione alle differenze di ordine n è individuata imponendo n *condizioni iniziali*, cioè assegnando n valori y_i, $i = k_0, k_0+1, \ldots, k_0+n-1$. Per le equazioni alle differenze lineari vale infatti il seguente teorema.

4.10 Teorema. *Sia $a_n(k) \neq 0$ per ogni $k \geq k_0$; assegnate n condizioni iniziali, la soluzione di (45) esiste ed è unica.*

Dim. Dalla (45), poiché $a_n(k) \neq 0$ per ogni $k \geq k_0$, si ha

$$y_{k+n} = \frac{1}{a_n(k)} \left[b(k) - \sum_{j=0}^{n-1} a_j(k)\ y_{k+j} \right]. \tag{46}$$

Dalla (46) si ricavano in modo unico i valori y_i per $i \geq k_0 + n$. ∎

4.11 Esempio. Per $k \geq 0$ l'integrale

$$y_k = \int_0^1 x^k\, \mathrm{e}^{x-1}\, dx$$

soddisfa la relazione ricorrente

$$y_k + k y_{k-1} = 1, \tag{47}$$

che si può ricavare integrando per parti

$$\int_0^1 x^k \, \mathrm{e}^{x-1} \, dx = x^k \, \mathrm{e}^{x-1} \Big|_0^1 - k \int_0^1 x^{k-1} \, \mathrm{e}^{x-1} \, dx.$$

La (47) è un'equazione alle differenze del primo ordine. Poiché vale la condizione iniziale

$$y_0 = \int_0^1 \mathrm{e}^{x-1} \, dx = 1 - \frac{1}{\mathrm{e}} \,,$$

dalla (47) si possono ricavare i successivi valori della soluzione

$$y_1 = \frac{1}{\mathrm{e}} \,, \quad y_2 = 1 - \frac{2}{\mathrm{e}} \,, \quad y_3 = -2 + \frac{6}{\mathrm{e}} \,, \ldots \qquad \blacksquare$$

La possibilità di calcolare in modo ricorrente con la (46) la soluzione di un'equazione alle differenze lineare fa sì che tali equazioni siano molto utilizzate nella matematica computazionale. Ad esempio equazioni alle differenze lineari sono usate nella teoria dell'approssimazione, per calcolare le soluzioni delle equazioni differenziali, nell'algebra lineare numerica.

4.12 Esempio. Si consideri la matrice tridiagonale A_k i cui elementi per $i,\ j = 1, 2, \ldots, k$ sono dati da

$$a_{ij} = \begin{cases} \alpha_i & \text{se } i - j = 0, \\ \beta_{i-1} & \text{se } i - j = 1, \\ \gamma_i & \text{se } i - j = -1, \\ 0 & \text{altrimenti.} \end{cases}$$

e sia d_k il determinante di A_k. Applicando il teorema di Laplace si ha

$$\det \begin{bmatrix} \alpha_1 & \gamma_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix} = \alpha_k \det \begin{bmatrix} \alpha_1 & \gamma_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_{k-2} \\ & & \beta_{k-2} & \alpha_{k-1} \end{bmatrix}$$

$$-\beta_{k-1} \det \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \gamma_{k-3} & \\ & & \ddots & \alpha_{k-2} & 0 \\ & & & \beta_{k-2} & \gamma_{k-1} \end{bmatrix},$$

da cui

$$d_k = \alpha_k d_{k-1} - \beta_{k-1}\gamma_{k-1} d_{k-2}. \tag{48}$$

Quindi la successione $\{d_k\}_{k=1,2,\ldots}$ è soluzione dell'equazione alle differenze del secondo ordine lineare (48). Tenendo conto che

$$d_1 = \alpha_1 \quad \text{e} \quad d_2 = \alpha_1\alpha_2 - \beta_1\gamma_1,$$

i valori di d_k per $k = 3, 4, \ldots$, si possono calcolare applicando in modo ricorrente la (48). Si ottiene così

$$\begin{aligned} d_3 &= \alpha_3 d_2 - \beta_2\gamma_2 d_1 = \alpha_1(\alpha_2\alpha_3 - \beta_2\gamma_2) - \beta_1\gamma_1\alpha_3 \\ d_4 &= \alpha_4 d_3 - \beta_3\gamma_3 d_2 = (\alpha_1\alpha_2 - \beta_1\gamma_1)(\alpha_3\alpha_4 - \beta_3\gamma_3) - \alpha_1\beta_2\gamma_2\alpha_4 \\ &\ldots \end{aligned}$$

∎

Calcolando la soluzione in modo ricorrente con la (46), in generale non si ottengono informazioni sul comportamento analitico della soluzione. Inoltre, come si vedrà nel paragrafo 8, possono sorgere problemi di instabilità numerica, per cui può essere necessario determinare la soluzione in forma analitica. La conoscenza della forma analitica della soluzione permette di risolvere un'equazione alle differenze anche quando vengono assegnate condizioni di tipo più generale di quelle iniziali.

Se nell'equazione (45) è $b(k) = 0$ per ogni k, allora l'equazione si dice *omogenea* ed ha la forma

$$\sum_{j=0}^{n} a_j(k) y_{k+j} = 0. \tag{49}$$

n soluzioni $z_k^{(1)}, \ldots, z_k^{(n)}$ della (49) si dicono *linearmente indipendenti* su Z' se la relazione

$$\sum_{j=1}^{n} \alpha_j z_k^{(j)} = 0$$

è soddisfatta se e solo se $\alpha_j = 0$ per $j = 1, \ldots, n$.

4.13 Teorema. *Sia $a_n(k) \neq 0$ per ogni $k \geq k_0$. Si considerino le n soluzioni linearmente indipendenti $z_k^{(j)}$, $j = 1, \ldots, n$, della (49) corrispondenti alle condizioni iniziali $z_{k_0+i-1}^{(j)} = \delta_{ij}$, $i, j = 1, \ldots, n$. Allora ogni combinazione lineare*

$$y_k = \alpha_1 z_k^{(1)} + \alpha_2 z_k^{(2)} + \ldots + \alpha_n z_k^{(n)} \tag{50}$$

è soluzione della (49) e, viceversa, ogni soluzione della (49) è esprimibile nella forma (50).

Dim. La (50) è soluzione della (49) per la linearità dell'equazione. Viceversa, sia z_k una soluzione di (49) e si consideri la successione

$$w_k = z_k - \sum_{j=1}^{n} \alpha_j z_k^{(j)}, \quad \alpha_j = z_{k_0+j-1}. \tag{51}$$

La (51), che è soluzione della (49), in quanto differenza di due soluzioni, è tale che

$$w_k = 0 \quad \text{per} \quad k = k_0, \ldots, k_0 + n - 1,$$

e poiché per il teorema 4.10 a condizioni iniziali nulle corrisponde la sola soluzione nulla, risulta

$$w_k = 0 \quad \text{per ogni} \quad k \geq k_0,$$

e quindi

$$z_k = \sum_{j=1}^{n} x_j z_k^{(j)}, \quad \text{per} \quad k = k_0, k_0 + 1, \ldots \qquad \blacksquare$$

La indipendenza lineare delle soluzioni $z_k^{(j)}$, $j = 1, \ldots, n$ del teorema 4.13 è essenzialmente dovuta alle condizioni iniziali imposte, e non implica in generale che le soluzioni siano ancora linearmente indipendenti se si fa variare l'indice k in un sottoinsieme di Z', ad esempio per $k \geq k_1$, con $k_1 > k_0$. Si dimostra però (si veda l'esercizio 4.48) che, se oltre ad essere $a_n(k) \neq 0$ è anche $a_0(k) \neq 0$ per ogni $k \geq k_0$, allora per ogni $k_1 \geq k_0$ i vettori di n componenti $(z_{k_1}^{(j)}, \ldots, z_{k_1+n-1}^{(j)})$ per $j = 1, \ldots, n$ sono linearmente indipendenti, e quindi l'insieme delle soluzioni della (49) è uno spazio vettoriale di dimensione n. Nel seguito si supporrà sempre che $a_0(k) \neq 0$ e $a_n(k) \neq 0$ per ogni $k \geq k_0$; senza ledere la generalità si porrà $a_0(k) = 1$.

Una combinazione delle soluzioni $z_k^{(1)}, \ldots, z_k^{(n)}$ è detta *soluzione generale* della (49). Il termine generale sta ad indicare che tale soluzione può essere determinata da un qualunque insieme di condizioni iniziali. Una soluzione di (49) determinata da particolari condizioni iniziali è detta *soluzione particolare.*

4.14 Teorema. *Sia z_k una soluzione particolare dell'equazione (45). Allora ogni soluzione dell'equazione (45) è esprimibile nella forma $v_k = z_k - y_k$, dove y_k è la soluzione generale dell'equazione (49), e quindi la soluzione generale dell'equazione (45) ha la forma*

$$v_k = z_k + \alpha_1 z_k^{(1)} + \alpha_2 z_k^{(2)} + \ldots + \alpha_n z_k^{(n)}, \tag{52}$$

dove $z_k^{(1)}, \ldots, z_k^{(n)}$ *sono soluzioni linearmente indipendenti della (49).*

Dim. La tesi è conseguenza immediata della linearità dell'equazione. ∎

Si studia ora in dettaglio il caso delle equazioni alle differenze lineari *a coefficienti costanti*, cioè

$$\sum_{j=0}^{n} a_j y_{k+j} = b(k), \quad a_0 \neq 0, \quad a_n = 1, \tag{53}$$

caso che può essere trattato in modo esauriente con tecniche elementari. In alcuni casi, con opportune trasformazioni, si riesce a ricondurre un'equazione alle differenze non lineare oppure lineare, ma a coefficienti non costanti ad una a coefficienti costanti (si vedano gli esercizi da 4.50 a 4.54).

L'equazione algebrica

$$P(x) = \sum_{j=0}^{n} a_j x^j = 0 \tag{54}$$

è detta *equazione caratteristica*. Indicate con $x_1, x_2, \ldots, x_r$, le soluzioni distinte dell'equazione (54), di molteplicità rispettivamente $m_1, m_2, \ldots, m_r$, tali che

$$\sum_{j=1}^{r} m_j = n,$$

si considerano le successioni $\{z_k^{(i,s)}\}_{k \in \mathbf{N}}$ così definite

$$z_k^{(i,s)} = k^s x_i^k, \quad i = 1, \ldots, r, \ s = 0, \ldots, m_i - 1. \tag{55}$$

4.15 Teorema. *Le successioni* $\{z_k^{(i,s)}\}_{k \in \mathbf{N}}$ *definite in (55) sono soluzioni dell'equazione omogenea associata alla (53)*

$$\sum_{j=0}^{n} a_j y_{k+j} = 0, \quad a_0 \neq 0, \quad a_n = 1. \tag{56}$$

Dim. Fissato un valore dell'indice i, se $m_i = 1$ si ha

$$\sum_{j=0}^{n} a_j \, z_{k+j}^{(i,0)} = \sum_{j=0}^{n} a_j \, x_i^{k+j} = x_i^k \sum_{j=0}^{n} a_j \, x_i^j,$$

e tale espressione è nulla per la (54). Se $m_i > 1$, si considerano i polinomi

$$\begin{aligned}
&P_0(x) = P(x),\\
&P_1(x) = xP_0'(x) = xP'(x)\\
&P_2(x) = xP_1'(x) = xP'(x) + x^2P''(x)\\
&\vdots\\
&P_{m_i-1}(x) = xP_{m_i-2}'(x) = xP'(x) + \ldots + x^{m_i-1}P^{(m_i-1)}(x).
\end{aligned}$$

È facile verificare per induzione che

$$P_s(x) = \sum_{j=0}^{n} a_j j^s x^j, \qquad s = 0, \ldots, m_i - 1. \tag{57}$$

Poiché $P'(x_i) = P''(x_i) = \ldots = P^{(m_i-1)}(x_i) = 0$, ne segue che

$$P_s(x_i) = 0, \qquad s = 0, \ldots, m_i - 1,$$

e per la (57) è

$$\sum_{j=0}^{n} a_j j^s x_i^j = 0, \qquad s = 0, \ldots, m_i - 1. \tag{58}$$

Risulta quindi

$$\begin{aligned}
\sum_{j=0}^{n} a_j \, z_{k+j}^{(i,s)} &= \sum_{j=0}^{n} a_j (k+j)^s x_i^{k+j} = \sum_{j=0}^{n} a_j \sum_{r=0}^{s} \binom{s}{r} k^{s-r} j^r x_i^{k+j}\\
&= \sum_{r=0}^{s} \binom{s}{r} k^{s-r} x_i^k \sum_{j=0}^{n} a_j \, j^r x_i^j.
\end{aligned} \tag{59}$$

Per la (58) se $s \le m_i - 1$, la (59) è nulla per ogni k. ∎

Le successioni $z_k^{(i,s)}$ sono linearmente indipendenti (si veda l'esercizio 4.55), e la base da esse formata si dice *sistema fondamentale di soluzioni* dell'equazione (56). Per il teorema 4.13 la soluzione generale dell'equazione (56) è una combinazione lineare delle soluzioni $z_k^{(i,s)} = k^s x_i^k$, $i = 1, \ldots, r$, $s = 0, \ldots, m_i - 1$.

4.16 Esempio. Si consideri l'equazione lineare omogenea del secondo ordine:

$$y_{k+2} = y_k + y_{k+1} \tag{60}$$

con le condizioni iniziali $y_0 = 0$ e $y_1 = 1$. La soluzione di (60) è la successione

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$$

i cui elementi sono detti *numeri di Fibonacci*. Un sistema fondamentale di soluzioni della (60) è

$$z_k^{(1)} = x_1^k, \quad z_k^{(2)} = x_2^k,$$

dove x_1 e x_2 sono le due soluzioni distinte dell'equazione caratteristica

$$x^2 - x - 1 = 0,$$

quindi

$$z_k^{(1)} = \Big(\frac{1-\sqrt{5}}{2}\Big)^k, \quad z_k^{(2)} = \Big(\frac{1+\sqrt{5}}{2}\Big)^k.$$

La soluzione generale della (60) è dunque della forma

$$\alpha_1 z_k^{(1)} + \alpha_2 z_k^{(2)},$$

dove α_1 e α_2 sono dei parametri che si determinano in base alle condizioni iniziali. Per ottenere la soluzione che soddisfa le condizioni $y_0 = 0$ e $y_1 = 1$, si determinano α_1 e α_2 tali che

$$\alpha_1 + \alpha_2 = 0, \qquad \alpha_1 \, \frac{1-\sqrt{5}}{2} \, + \alpha_2 \, \frac{1+\sqrt{5}}{2} \, = 1,$$

da cui risulta

$$\alpha_1 = -\alpha_2 = -\,\frac{1}{\sqrt{5}}\,.$$

I numeri di Fibonacci sono quindi

$$y_k = \frac{1}{\sqrt{5}} \left[\Big(\frac{1+\sqrt{5}}{2}\Big)^k - \Big(\frac{1-\sqrt{5}}{2}\Big)^k \right]. \tag{61}$$ ∎

Se non tutte le soluzioni $x_1, x_2, \ldots, x_r$ della (54) sono reali, si può ugualmente trovare una base di soluzioni reali per la (56). Infatti poiché i coefficienti dell'equazione (54) sono reali, per ogni soluzione complessa esiste anche la soluzione coniugata. Siano ad esempio

$$x_1 = \rho(\cos\phi + \boldsymbol{i}\sin\phi), \quad x_2 = \rho(\cos\phi - \boldsymbol{i}\sin\phi)$$

due soluzioni complesse coniugate della (54) con molteplicità m_1. Le $2m_1$ soluzioni della base

$$k^s x_1^k = k^s \rho^k (\cos k\phi + \boldsymbol{i} \sin k\phi), \quad k^s x_2^k = k^s \rho^k (\cos k\phi - \boldsymbol{i} \sin k\phi),$$

$$s = 0, \dots, m_1 - 1$$

possono essere sostituite dalle loro combinazioni lineari, ancora soluzioni della (56),

$$\begin{aligned} -\frac{\boldsymbol{i}}{2}\, k^s x_1^k + \frac{\boldsymbol{i}}{2}\, k^s x_2^k &= k^s \rho^k \sin k\phi, \\ \frac{1}{2}\, k^s x_1^k + \frac{1}{2}\, k^s x_2^k &= k^s \rho^k \cos k\phi. \end{aligned}$$

4.17 Esempio. Se la matrice A_k dell'esempio 4.12 ha uguali fra di loro gli elementi posti su ogni diagonale, cioè

$$\alpha_i = \alpha \text{ per } i = 1, \dots, k, \quad \beta_i = \beta \text{ e } \gamma_i = \gamma \text{ per } i = 1, \dots, k-1,$$

l'equazione (48) è a coefficienti costanti

$$d_k - \alpha d_{k-1} + \beta\gamma d_{k-2} = 0.$$

La sua soluzione generale è data da

$$d_k = a_1 x_1^k + a_2 x_2^k,$$

dove $a_1,\ a_2$ sono dei parametri, se le radici x_1 e x_2 dell'equazione caratteristica sono distinte, e da

$$d_k = (a_1 + a_2 k) x_1^k,$$

se l'equazione caratteristica ha una sola radice x_1 di molteplicità 2. I parametri a_1 e a_2 vengono determinati imponendo le condizioni iniziali

$$d_1 = \alpha \quad \text{e} \quad d_2 = \alpha^2 - \beta\gamma.$$

Se $\alpha = 2$, $\beta = 3$, $\gamma = 1/4$, allora l'equazione caratteristica è

$$4x^2 - 8x + 3 = 0,$$

per cui

$$d_k = a_1 \big(\frac{3}{2}\big)^k + a_2 \big(\frac{1}{2}\big)^k.$$

Imponendo le condizioni iniziali si ottengono per i parametri i valori

$$a_1 = \frac{3}{2}\,, \quad a_2 = -\,\frac{1}{2}\,,$$

per cui il determinante di A_k è dato da

$$d_k = \frac{1}{2}\,\left[3\,(\frac{3}{2})^k - (\frac{1}{2})^k\right].$$

Se $\alpha = 2$, $\beta = 1$, $\gamma = 1$, allora l'equazione caratteristica

$$x^2 - 2x + 1 = 0$$

ha la sola soluzione $x_1 = 1$, di molteplicità 2, per cui

$$d_k = a_1 + a_2 k.$$

Imponendo le condizioni iniziali si ottengono per i parametri i valori

$$a_1 = a_2 = 1,$$

per cui il determinante di A_k è dato da

$$d_k = 1 + k.$$

Se $\alpha = 2$, $\beta = 2$, $\gamma = 2$, allora l'equazione caratteristica

$$x^2 - 2x + 4 = 0$$

ha le due soluzioni complesse

$$x_1 = 2\Big(\cos\,\frac{\pi}{3} + \boldsymbol{i}\sin\,\frac{\pi}{3}\Big) \quad \text{e} \quad x_2 = 2\Big(\cos\,\frac{\pi}{3} - \boldsymbol{i}\sin\,\frac{\pi}{3}\Big)$$

per cui

$$d_k = 2^k\Big(a_1\cos\,\frac{k\pi}{3} + a_2\sin\,\frac{k\pi}{3}\Big).$$

Imponendo le condizioni iniziali si ottengono per i parametri i valori

$$a_1 = 1, \quad a_2 = \frac{1}{\sqrt{3}}\,,$$

per cui il determinante di A_k è dato da

$$d_k = 2^k\Big(\cos\frac{k\pi}{3} + \frac{1}{\sqrt{3}}\,\sin\frac{k\pi}{3}\Big).$$

∎

4.18 Esempio. Si vogliono determinare autovalori e autovettori della matrice tridiagonale simmetrica

$$A_n = \begin{bmatrix} a & -1 & & \\ -1 & a & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & a \end{bmatrix}, \quad a \in \mathbf{R}.$$

Per i teoremi di Gerschgorin [2] gli autovalori λ di A_n sono tali che

$$|a - \lambda| < 2. \tag{62}$$

Dall'esempio 4.12 segue che il polinomio caratteristico $p_n(\lambda) = \det(A_n - \lambda I)$ soddisfa, per ogni λ, l'equazione alle differenze lineare omogenea del secondo ordine

$$p_n(\lambda) = (a - \lambda)p_{n-1}(\lambda) - p_{n-2}(\lambda), \tag{63}$$

con condizioni iniziali

$$p_1(\lambda) = a - \lambda, \quad p_2(\lambda) = (a - \lambda)^2 - 1. \tag{64}$$

Perciò per calcolare $p_n(\lambda)$ si considera l'equazione caratteristica associata alla (63)

$$x^2 - (a - \lambda)x + 1 = 0,$$

in cui per la (62) si può porre $a - \lambda = 2\cos\phi$, $0 < \phi < \pi$. Le soluzioni sono

$$x_{1,2} = \cos\phi \pm \boldsymbol{i}\sin\phi,$$

da cui si ottiene la soluzione generale della (63)

$$p_n(\lambda) = c_1 \cos n\phi + c_2 \sin n\phi.$$

Imponendo le condizioni iniziali (64) si ottiene

$$p_n(\lambda) = \frac{\sin(n+1)\phi}{\sin\phi}.$$

Poiché $p_n(\lambda) = 0$ se ϕ è soluzione dell'equazione

$$\sin(n+1)\phi = 0,$$

gli autovalori λ_i sono della forma

$$\lambda_i = a - 2\cos\frac{i\pi}{n+1}, \quad i = 1, \ldots, n.$$

Indicato poi con $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})^T$ un autovettore di A_n, corrispondente all'autovalore λ_i, dalla relazione $A_n\mathbf{x}^{(i)} = \lambda_i\mathbf{x}^{(i)}$ si ha

$$(a - \lambda_i)x_1^{(i)} - x_2^{(i)} = 0, \tag{65}$$

$$-x_k^{(i)} + (a - \lambda_i)x_{k+1}^{(i)} - x_{k+2}^{(i)} = 0, \quad k = 1, \ldots, n-2, \tag{66}$$

$$-x_{n-1}^{(i)} + (a - \lambda_i)x_n^{(i)} = 0.$$

Poiché λ_i è autovalore, il determinante di questo sistema lineare è nullo, e quindi ogni equazione è combinazione lineare delle altre. Se si elimina l'ultima equazione e si assegna ad $x_1^{(i)}$ un valore arbitrario, si possono calcolare le incognite $x_2^{(i)}, \ldots, x_n^{(i)}$. Posto $x_1^{(i)} = 1$, dalla (65) si ha $x_2^{(i)} = a - \lambda_i$, e quindi la (66) risulta essere un'equazione alle differenze con le condizioni iniziali

$$x_1^{(i)} = 1, \quad x_2^{(i)} = a - \lambda_i. \tag{67}$$

Risolvendo la (66) con le condizioni iniziali (67) in modo analogo a quanto fatto per la (63), si ottiene la k-esima componente dell'autovettore $\mathbf{x}^{(i)}$, normalizzato in modo che sia $x_1^{(i)} = 1$

$$x_k^{(i)} = \sin\frac{ik\pi}{n+1} \Big/ \sin\frac{i\pi}{n+1}\,. \qquad \blacksquare$$

Per il teorema 4.14, la determinazione della soluzione generale della (53) è ricondotta alla individuazione di una soluzione particolare z_k. Si può dimostrare [15] che se sono note n soluzioni linearmente indipendenti $z_k^{(1)}, \ldots, z_k^{(n)}$ dell'equazione omogenea (56), una soluzione particolare della (53) è data da

$$z_k = \alpha_1(k)z_k^{(1)} + \ldots + \alpha_n(k)z_k^{(n)},$$

dove $\alpha_1(k), \ldots, \alpha_n(k)$ sono delle opportune funzioni di k. Per determinare $\alpha_1(k), \ldots, \alpha_n(k)$ si può usare il metodo della *variazione dei parametri* (descritto nell'esercizio 4.70 per il caso dei coefficienti costanti, ma estendibile anche al caso dei coefficienti non costanti). Un metodo più semplice per determinare una soluzione della (53) è il seguente *metodo dei coefficienti indeterminati*.

Se $b(k)$ ha un'espressione compresa fra quelle elencate nella prima colonna della tabella di seguito riportata, una soluzione particolare z_k della (53) è della corrispondente forma indicata nella seconda colonna. I parametri α, β, α_i, β_i, vengono determinati imponendo che z_k soddisfi la (53).

$b(k)$	z_k	note
αx^k	βx^k	(1)
$\alpha_0+\alpha_1 k+\ldots+\alpha_m k^m$	$\beta_0+\beta_1 k+\ldots+\beta_m k^m$	
$x^k(\alpha_0+\alpha_1 k+\ldots+\alpha_m k^m)$	$x^k(\beta_0+\beta_1 k+\ldots+\beta_m k^m)$	(1)
$\sin\alpha k$ opp. $\cos\alpha k$	$\beta_0\sin\alpha k+\beta_1\cos\alpha k$	(2)
$x^k\sin\alpha k$ opp. $x^k\cos\alpha k$	$x^k(\beta_0\sin\alpha k+\beta_1\cos\alpha k)$	(1) (2)

Note: (1) se x è radice di molteplicità m dell'equazione caratteristica (54), si deve porre $k^m x^k$ al posto di x^k; (2) se $\cos\alpha+\boldsymbol{i}\sin\alpha$ è radice di molteplicità m dell'equazione caratteristica (54), si deve porre $k^m(\beta_0\sin\alpha k+\beta_1\cos\alpha k)$ al posto di $\beta_0\sin\alpha k+\beta_1\cos\alpha k$.

4.19 Esempio. Si vuole calcolare il determinante d_k della matrice di ordine k *ad albero*

$$A_k=\det\begin{bmatrix} a & & & b\\ & a & & \vdots\\ & & \ddots & b\\ b & \ldots & b & a\end{bmatrix},\quad a,b\in\mathbf{C}.$$

Applicando la regola di Laplace alla prima riga si ha per $k\geq 2$

$$d_k=ad_{k-1}+(-1)^{k-1}b\,\det\begin{bmatrix} 0 & a & & \\ & \ddots & \ddots & \\ & & 0 & a\\ b & \ldots & b & b\end{bmatrix}=ad_{k-1}-b^2a^{k-2},\quad(68)$$

ed inoltre

$$d_1=a\qquad(69)$$

Il determinante di A_k è dato dalla soluzione dell'equazione alle differenze lineare (68) che soddisfa la condizione iniziale (69). L'equazione omogenea associata alla (68) ha la soluzione generale

$$y_k=\alpha a^k.$$

Una soluzione particolare z_k dell'equazione completa (68) è della forma

$$z_k=k\beta a^{k-2},$$

dove il parametro β viene determinato sostituendo z_k nell'equazione (68). Risulta che deve essere $\beta=-b^2$, per cui la soluzione generale della (68) è

$$v_k=\alpha a^k-kb^2a^{k-2}.$$

Imponendo che tale soluzione verifichi la condizione iniziale (69) si ricava per α il valore

$$\alpha = \frac{a^2 + b^2}{a^2},$$

a cui corrisponde la soluzione particolare cercata

$$d_k = a^{k-2}[a^2 + (1-k)b^2]. \qquad \blacksquare$$

8. Stabilità del calcolo delle formule ricorrenti

Nella risoluzione di equazioni alle differenze mediante l'uso di relazioni ricorrenti come la (46) si possono presentare problemi di instabilità numerica.

Si suppone che nell'equazione (45) sia $a_n(k) = 1$ e $a_0(k) \neq 0$ per ogni $k \geq k_0$, per cui la relazione ricorrente (46) risulta

$$y_{k+n} = b(k) - \sum_{j=0}^{n-1} a_j(k) y_{k+j}. \tag{70}$$

Una soluzione y_k della (45) è detta *minimale* se esiste una soluzione z_k dell'equazione omogenea (49) tale che

$$\lim_{k\to\infty} \frac{y_k}{z_k} = 0. \tag{71}$$

Nel calcolo di una soluzione minimale, mediante la (70), a partire da condizioni iniziali assegnate, gli errori di arrotondamento possono accumularsi in modo tale da rendere completamente inutilizzabile l'algoritmo.

4.20 Esempio. La successione

$$y_k = \int_0^1 x^k \mathrm{e}^{x-1}\, dx$$

è per $k \geq 0$ monotona decrescente al crescere di k e risulta

$$\lim_{k\to\infty} y_k = 0.$$

Come si è visto nell'esempio 4.11, la successione y_k costituisce una soluzione particolare dell'equazione alle differenze del primo ordine

$$y_k + k y_{k-1} = 1,$$

corrispondente alla condizione iniziale $y_0 = 1 - \dfrac{1}{\mathrm{e}} = 0.6321206$. L'equazione omogenea associata è

$$y_k + k y_{k-1} = 0,$$

la cui soluzione generale è

$$\alpha_1 z_k^{(1)} = \alpha_1 (-1)^k \, k!.$$

Quindi la soluzione y_k che si vuole calcolare è minimale. Calcolando y_k per mezzo della relazione ricorrente

$$y_k = 1 - k y_{k-1},$$

a partire dal valore $\widetilde{y}_0 = 0.6321206$, si ottiene la successione

k	$\widetilde{y}_k$	k	$\widetilde{y}_k$
1	0.3678795	6	0.1267529
2	0.2642410	7	0.1127300
3	0.2072771	8	0.09815979
4	0.1708918	9	0.1165619
5	0.1455412	10	− 0.1656189

I valori ottenuti per $\widetilde{y}_k$, con $k \geq 9$ sono evidentemente affetti da errori molto elevati. L'algoritmo utilizzato è quindi instabile, e l'instabilità è dovuta al fatto che la soluzione y_k che si vuole calcolare è minimale. ∎

Per studiare la propagazione dell'errore che si produce quando si calcola una soluzione minimale y_k, si indichi con $\widetilde{y}_k$ la soluzione ottenuta applicando la (70) alle condizioni iniziali $\widetilde{y}_0, \ldots, \widetilde{y}_{n-1}$. Indicata con $z_k^{(1)}$ la soluzione z_k della equazione omogenea (49) per cui vale la (71), se $n \geq 2$ si considerino altre $n-1$ soluzioni $z_k^{(2)}, \ldots, z_k^{(n)}$ della (49), in modo che le soluzioni $z_k^{(i)}$, $i = 1, \ldots, n$ siano linermente indipendenti. Poichè y_k è una soluzione particolare della (70), per la (52) esistono $\widetilde{\alpha}_j$, $j = 1, \ldots, n$, per cui

$$\widetilde{y}_k = y_k + \sum_{j=1}^{n} \widetilde{\alpha}_j z_k^{(j)}.$$

Se $y_k \neq 0$ per ogni k, l'errore relativo e_k di $\widetilde{y}_k$ rispetto a y_k è

$$e_k = \frac{\widetilde{y}_k - y_k}{y_k} = \frac{1}{y_k} \sum_{j=1}^{n} \widetilde{\alpha}_j z_k^{(j)} = \sum_{j=1}^{n} \widetilde{\alpha}_j \frac{z_k^{(j)}}{y_k},$$

e dalla (71) segue che $|e_k|$ non è limitabile superiormente per ogni k e per ogni scelta delle condizioni iniziali $\widetilde{y}_0, \ldots, \widetilde{y}_{n-1}$. Quindi la soluzione y_k della (70), considerata come funzione delle condizioni iniziali $y_0, \ldots, y_{n-1}$, è mal condizionata, in quanto a piccole perturbazioni $\delta_i = \widetilde{y}_i - y_i$, $i = 0 \ldots, n-1$, sui dati, corrisponde una perturbazione sulla soluzione non limitabile superiormente. Nel calcolo della soluzione con la (70) ad ogni passo si utilizzano gli n valori, precedentemente calcolati, che sono affetti da errore. Quindi, dal mal condizionamento del problema che deve risolto ad ogni passo, discende che l'algoritmo non è stabile.

Poiché a priori non è facile stabilire se la soluzione y_k da calcolare è o no minimale, l'utilizzazione della (70) deve essere sempre accompagnata da un esame del problema, volto a determinare, per quanto possibile, qual è l'andamento della soluzione.

4.21 Esempio. Per $k \geq 0$ l'integrale

$$y_k = \int_0^\pi \frac{\cos kx}{5 - 3\cos x}\, dx$$

soddisfa la relazione ricorrente

$$y_{k+2} - \frac{10}{3}\, y_{k+1} + y_k = 0, \tag{72}$$

che si può ricavare dalla seguente uguaglianza, ottenuta applicando note identità trigonometriche

$$\begin{aligned} \frac{\cos(k+2)x}{5 - 3\cos x} &= \frac{2\cos(k+1)x\cos x - \cos kx}{5 - 3\cos x} \\ &= -\frac{2}{3}\, \cos(k+1)x + \frac{10}{3}\, \frac{\cos(k+1)x}{5 - 3\cos x} - \frac{\cos kx}{5 - 3\cos x}\,, \end{aligned}$$

e tenendo conto del fatto che

$$\int_0^\pi \cos(k+1)x\, dx = 0.$$

Inoltre valgono le condizioni iniziali $y_0 = \pi/4$ e $y_1 = \pi/12$, come si può vedere per calcolo diretto. L'equazione caratteristica della (72) ha le due radici $x_1 = 3$ e $x_2 = 1/3$, e quindi la soluzione generale è data da

$$y_k = \alpha_1 3^k + \alpha_2 \big(\frac{1}{3}\big)^k.$$

Imponendo che valgano le condizioni iniziali, si ricavano per i parametri i valori

$$\alpha_1 = 0 \quad \text{e} \quad \alpha_2 = \frac{\pi}{4}\,,$$

da cui si ottiene

$$y_k = \frac{\pi}{4}\left(\frac{1}{3}\right)^k. \tag{73}$$

La (73) rappresenta quindi una soluzione minimale dell'equazione alle differenze (72). Nella seguente tabella sono riportati i valori y_k effettivamente calcolati per mezzo della (73) e i valori v_k effettivamente calcolati applicando la (70) con le condizioni iniziali, cioè

$$v_0 = \frac{\pi}{4}\,, \quad v_1 = \frac{\pi}{12}\,, \quad v_{k+2} = \frac{10}{3}\,v_{k+1} - v_k, \text{ per } k = 0, 1, \ldots$$

k	y_k	v_k
2	$0.8726645\ 10^{-1}$	$0.8726627\ 10^{-1}$
3	$0.2908881\ 10^{-1}$	$0.2908808\ 10^{-1}$
4	$0.9696271\ 10^{-2}$	$0.9693980\ 10^{-2}$
5	$0.3232090\ 10^{-2}$	$0.3225181\ 10^{-2}$
6	$0.1077363\ 10^{-2}$	$0.1056623\ 10^{-2}$
7	$0.3591210\ 10^{-3}$	$0.2968940\ 10^{-3}$
8	$0.1197070\ 10^{-3}$	$-0.6697630\ 10^{-4}$
9	$0.3990233\ 10^{-4}$	$-0.5201141\ 10^{-3}$

Quindi i valori v_k risultano del tutto inaccettabili già per $k = 7$. ∎

Il calcolo di una soluzione minimale y_k dell'equazione (45) può essere affrontato con un metodo iterativo che genera, all'N-esima iterazione, una soluzione particolare $y_k^{(N)}$ dell'equazione, in modo da garantire la convergenza ad y_k per $N \to \infty$. La trattazione sarà limitata ai soli casi delle equazioni del primo ordine

$$y_{k+1} + a_0(k)\, y_k = b(k), \quad a_0(k) \neq 0 \quad \text{per} \quad k \geq 0, \tag{74}$$

o del secondo ordine

$$y_{k+2} + a_1(k)\, y_{k+1} + a_0(k)\, y_k = b(k), \quad a_0(k) \neq 0 \quad \text{per} \quad k \geq 0, \tag{75}$$

e si basa sul seguente teorema di convergenza.

4.22 Teorema.

a) Si considerino l'equazione (74) con la condizione iniziale y_0, tale che la corrispondente soluzione y_k sia minimale e, per ogni intero $N \geq 1$, la soluzione $y_k^{(N)}$ di (74) che soddisfa la condizione

$$y_N^{(N)} = 0. \tag{76}$$

Sia z_k una soluzione dell'equazione omogenea associata alla (74); se $z_0 \neq 0$, fissato un valore dell'indice k, è

$$\lim_{N\to\infty} y_k^{(N)} = y_k.$$

b) Si considerino l'equazione (75) con le condizioni iniziali y_0 e y_1, tali che la corrispondente soluzione y_k sia minimale e, per ogni intero $N \geq 1$, la soluzione $y_k^{(N)}$ di (75) che soddisfa le condizioni (dette condizioni al contorno)

$$y_0^{(N)} = y_0, \qquad y_N^{(N)} = 0. \tag{77}$$

Se l'equazione omogenea associata alla (75), oltre alla soluzione z_k per cui vale la (71), ha anche una soluzione v_k, tale che

$$v_0 \neq 0 \quad e \quad \lim_{k\to\infty} \frac{v_k}{z_k} = 0,$$

fissato un valore dell'indice k, è

$$\lim_{N\to\infty} y_k^{(N)} = y_k.$$

Dim. Nel caso a), poiché y_k è una soluzione particolare dell'equazione (74), risulta

$$y_k^{(N)} = y_k + \alpha^{(N)} z_k,$$

e imponendo la condizione (76) si ha che

$$\alpha^{(N)} = -\frac{y_N}{z_N},$$

dove $z_N \neq 0$ perché $z_0 \neq 0$. Poiché la soluzione y_k è minimale e l'equazione è del primo ordine, la (71) vale per ogni soluzione z_k non nulla dell'equazione omogenea, e quindi segue che

$$\lim_{N\to\infty} \alpha^{(N)} = 0.$$

Nel caso b), si indichino con $z_k^{(1)} = z_k$ e $z_k^{(2)} = v_k$ le due soluzioni dell'equazione omogenea associata alla (75) che sono linearmente indipendenti (si veda l'esercizio 4.49), e poiché y_k è una soluzione particolare dell'equazione (75), risulta

$$y_k^{(N)} = y_k + \alpha_1^{(N)} z_k^{(1)} + \alpha_2^{(N)} z_k^{(2)}.$$

Poiché $\lim_{k\to\infty} |z_k^{(1)}/z_k^{(2)}| = \infty$, esiste un indice m tale che per $N \geq m$ è

$$\frac{z_N^{(1)}}{z_N^{(2)}} \neq \frac{z_0^{(1)}}{z_0^{(2)}},$$

essendo $z_0^{(2)} = v_0 \neq 0$. Imponendo le condizioni (77) si ha che

$$\alpha_1^{(N)} = \frac{-z_0^{(2)} y_N}{z_0^{(1)} z_N^{(2)} - z_0^{(2)} z_N^{(1)}} = \frac{-z_0^{(2)} \dfrac{y_N}{z_N^{(1)}}}{z_0^{(1)} \dfrac{z_N^{(2)}}{z_N^{(1)}} - z_0^{(2)}},$$

$$\alpha_2^{(N)} = \frac{z_0^{(1)} y_N}{z_0^{(1)} z_N^{(2)} - z_0^{(2)} z_N^{(1)}} = \frac{z_0^{(1)} \dfrac{y_N}{z_N^{(1)}}}{z_0^{(1)} \dfrac{z_N^{(2)}}{z_N^{(1)}} - z_0^{(2)}},$$

e quindi, per le ipotesi fatte, è

$$\lim_{N\to\infty} \alpha_1^{(N)} = \lim_{N\to\infty} \alpha_2^{(N)} = 0. \qquad \blacksquare$$

Per il calcolo della soluzione minimale y_k, per $k = 1, \ldots, m$, si procede quindi calcolando, per $N = m+1, m+2, \ldots$ la soluzione $y_k^{(N)}$ dell'equazione (74) con la condizione (76), nel caso del primo ordine, o dell'equazione (75) con la condizione (77) nel caso del secondo ordine, e arrestando il procedimento ad un valore di N per cui

$$\max_{k=0,\ldots,m} |y_k^{(N+1)} - y_k^{(N)}| < \epsilon, \tag{78}$$

dove ϵ è una costante prefissata. Poiché nel teorema 4.22 non è richiesta la condizione che la soluzione y_k tenda a zero per $k \to \infty$, la convergenza di $y_k^{(N)}$ a y_k in generale non può essere che puntuale, e non è quindi possibile imporre che la disuguaglianza (78) valga per ogni $k \geq 0$.

Il criterio di arresto (78) non è sempre adeguato, in particolare quando le grandezze dei valori della successione differiscono molto fra di loro. In tal caso, se $y_k^{(N)} \neq 0$ per $k = 0, \ldots, m$, un criterio migliore è

$$\max_{k=0,\ldots,m} \left| \frac{y_k^{(N+1)} - y_k^{(N)}}{y_k^{(N)}} \right| < \epsilon,$$

la cui applicazione però non consente di sfruttare le stime a priori dell'errore (79) e (86), che saranno successivamente ottenute.

Nel caso che l'equazione alle differenze sia del primo ordine, il metodo iterativo è il metodo di *Miller*, che consiste nell'applicare l'equazione alle differenze *all'indietro*, cioè

$$y_N^{(N)} = 0, \qquad y_k^{(N)} = \frac{b(k) - y_{k+1}^{(N)}}{a_0(k)}, \quad \text{per} \quad k = N-1, \ldots, 1.$$

Così facendo, l'equazione alle differenze cambia, e quindi è possibile che la soluzione che si vuole calcolare non sia più minimale.

Per determinare il valore di N per cui vale la (78) si può evitare il calcolo della successione $y_k^{(N)}$ per valori crescenti di N, notando che se si aumenta di 1 il valore di N e si impone che la successione $y_k^{(N+1)}$ soddisfi la condizione $y_{N+1}^{(N+1)} = 0$, si ha

$$y_k^{(N+1)} = \frac{b(k) - y_{k+1}^{(N+1)}}{a_0(k)}\,,$$

per cui

$$y_k^{(N+1)} - y_k^{(N)} = -\,\frac{1}{a_0(k)}\,(y_{k+1}^{(N+1)} - y_{k+1}^{(N)}),$$

ed essendo

$$y_N^{(N+1)} - y_N^{(N)} = \frac{b(N)}{a_0(N)}\,,$$

risulta

$$\left|y_k^{(N+1)} - y_k^{(N)}\right| = \frac{|b(N)|}{\left|\prod_{j=k}^{N} a_0(j)\right|}\,. \tag{79}$$

Il valore di N opportuno può quindi essere determinato confrontando con ϵ il massimo valore delle (79) per $k = 0, \ldots, m$.

4.23 Esempio. Si applica il metodo di Miller al calcolo della successione y_k definita nell'esempio 4.20, per $k = 1, \ldots, 10$. La relazione ricorrente all'indietro è

$$y_N^{(N)} = 0, \qquad y_{k-1}^{(N)} = \frac{1}{k}\,(1 - y_k^{(N)}), \quad k = N, \ldots, 1.$$

Fissato $\epsilon = 10^{-5}$, per determinare un N per cui vale la (78), si utilizza la (79). Risulta

$$\max_{k=0,\ldots,10} |y_k^{(N+1)} - y_k^{(N)}| = \frac{1}{\prod_{j=10}^{N} j} = \frac{10!}{N!}\,.$$

Il minimo N per cui $N! > 10!\ 10^5$ è $N = 15$. Si calcola quindi $y_k^{(16)}$

k	$y_k^{(16)}$	k	$y_k^{(16)}$
15	0.06250000	9	0.09161228
14	0.06250000	8	0.1009319
⋮	⋮	⋮	⋮
10	0.08387703	1	0.3678795

Il valore $y_9^{(16)} = 0.09161228$ così calcolato è molto più preciso di quello ottenuto nell'esempio 4.20. ∎

Nel caso che l'equazione sia del secondo ordine, un metodo iterativo basato sul teorema 4.22 richiede, ad ogni iterazione la risoluzione del sistema triangolare

$$\begin{cases} a_1(0)\, y_1^{(N)} + y_2^{(N)} = b(0) - a_0(0)\, y_0, & (80) \\ a_0(k-1)\, y_{k-1}^{(N)} + a_1(k-1)\, y_k^{(N)} + y_{k+1}^{(N)} = b(k-1),\ k = 2, \ldots, N-2, \\ a_0(N-2)\, y_{N-2}^{(N)} + a_1(N-2)\, y_{N-1}^{(N)} = b(N-2). \end{cases}$$

Se l'equazione è omogenea, il termine noto di (80) è un vettore con tutte le componenti nulle, eccetto la prima. Per risolvere tale sistema *Miller* ha proposto il metodo seguente: si elimina la prima equazione, si assegna un valore arbitrario a $y_{N-1}^{(N)}$ (non nullo, altrimenti darebbe luogo ad una soluzione tutta nulla), e si ricavano gli y_k^N con la sostituzione all'indietro

$$y_k^{(N)} = -\frac{1}{a_0(k)} \left[y_{k+2}^{(N)} + a_1(k)\, y_{k+1}^{(N)} \right], \quad \text{per} \quad k = N-2, \ldots, 0.$$

In generale però la successione $y_k^{(N)}$ così ottenuta è tale che

$$y_0^{(N)} \neq y_0.$$

Sfruttando il fatto che l'equazione è omogenea e quindi se $y_k^{(N)}$ è soluzione, anche $\alpha y_k^{(N)}$ è soluzione per ogni α costante, si costruisce la successione

$$w_k = y_k^{(N)}\, \frac{y_0}{y_0^{(N)}}, \quad k = 0, \ldots, N,$$

che soddisfa le condizioni $w_0 = y_0$ e $w_N = 0$.

Uno studio dell'errore del metodo di Miller per le equazioni del secondo ordine risulta assai più complicato di quello sviluppato per le equazioni del primo ordine, si veda [19]. In questo caso non è in generale possibile determinare un valore di N per cui valga la (78) senza calcolare effettivamente i valori della successione.

4.24 Esempio. Si applica il metodo di Miller all'equazione alle differenze del secondo ordine omogenea dell'esempio 4.21. Fissato $N = 20$ e $y_{19}^{(N)} = 0.1$, la soluzione viene calcolata con la relazione

$$y_k^{(N)} = \frac{10}{3}\, y_{k+1}^{(N)} - y_{k+2}^{(N)}, \quad k = N-2, \ldots, 0,$$

e poi moltiplicata per $y_0/y_0^{(N)}$. Si ottiene la seguente successione che soddisfa la (78) con $\epsilon = 10^{-6}$.

k	w_k	k	w_k
1	0.2617994	6	$0.1077367\ 10^{-2}$
2	$0.8726650\ 10^{-1}$	7	$0.3591222\ 10^{-3}$
3	$0.2908884\ 10^{-1}$	8	$0.1197075\ 10^{-3}$
4	$0.9696282\ 10^{-2}$	9	$0.3990250\ 10^{-4}$
5	$0.3232099\ 10^{-2}$	10	$0.1330085\ 10^{-4}$

∎

Uno dei più importanti campi di applicazione del metodo di Miller è quello della tabulazione di funzioni speciali che soddisfano equazioni del secondo ordine omogenee. Il metodo si può applicare anche quando sono assegnate condizioni più generali di quelle iniziali.

4.25 Esempio. La funzione di Bessel di primo tipo di ordine n è definita da

$$J_n(x) = \frac{1}{\pi} \int_0^{\pi} \cos(n\theta - x \sin\theta)\ d\theta,$$

e soddisfa l'equazione alle differenze omogenea

$$J_{k+1}(x) - \frac{2k}{x}\, J_k(x) + J_{k-1}(x) = 0. \tag{81}$$

Fissato un valore di x, per esempio $x = 1$, e determinate, per altra via, le due condizioni iniziali $J_0(1)$ e $J_1(1)$, il calcolo delle successive $J_k(1)$, $k = 2, 3, \ldots$, può essere fatto utilizzando la (81). Poiché la soluzione cercata è minimale, i valori ottenuti risultano affetti da errori che crescono al crescere di k. Per $k = 9$ nessuna cifra ottenuta è esatta. Si applica invece il metodo di

Miller. Ottenuta la successione $y_k^{(N)}$, si sfrutta la seguente condizione di normalizzazione verificata dalle funzioni di Bessel

$$J_0^2(x) + 2J_1^2(x) + 2J_2^2(x) + \ldots = 1,$$

scalando la successione $y_k^{(N)}$ del fattore

$$\gamma = \Big[(y_0^{(N)})^2 + 2\sum_{j=1}^{N-1}(y_j^{(N)})^2\Big]^{1/2}.$$

Ad esempio, fissando $N = 20$ e $y_{19}^{(N)} = 0.1$, si ottiene la successione

k	$y_k^{(N)}$	k	$y_k^{(N)}$
0	0.7651978	5	$0.2497581\ 10^{-3}$
1	0.4400510	6	$0.2093839\ 10^{-4}$
2	0.1149036	7	$0.1502330\ 10^{-5}$
3	$0.1956338\ 10^{-1}$	8	$0.9422365\ 10^{-7}$
4	$0.2476643\ 10^{-2}$	9	$0.5249266\ 10^{-8}$

Si noti che in questo caso il metodo di Miller non solo ha consentito di calcolare in modo stabile la funzione cercata, ma sfruttando una condizione di normalizzazione nota teoricamente, non ha richiesto la determinazione dei due valori iniziali $J_0(1)$ e $J_1(1)$. ∎

Se l'equazione del secondo ordine non è omogenea, il metodo di Miller non è applicabile, e si ricorre allora al metodo di *Olver*, che è un po' più complesso di quello di Miller, ma consente di valutare in modo semplice l'errore al crescere di N. Con opportune combinazioni lineari delle righe, il sistema (80) viene trasformato nel sistema con matrice dei coefficienti bidiagonale superiore

$$\begin{cases} c_{k+1}y_k^{(N)} - c_k y_{k+1}^{(N)} = d_k, \quad k = 1, \ldots, N-2, \\ c_N y_{N-1}^{(N)} = d_{N-1}, \end{cases} \tag{82}$$

dove

$$\begin{aligned} &c_1 = 1, \qquad c_2 = -a_1(0), \\ &c_k = -\big[a_1(k-2)\ c_{k-1} + a_0(k-2)\ c_{k-2}\big], \quad k = 3, \ldots, N, \\ &d_1 = a_0(0)\ y_0 - b(0), \\ &d_k = a_0(k-1)\ d_{k-1} - b(k-1)\ c_k, \quad k = 2, \ldots, N-1. \end{aligned} \tag{83}$$

Dall'ultima equazione del sistema (82) si ottiene

$$y_{N-1}^{(N)} = \frac{d_{N-1}}{c_N}\,.$$

Si aumenta adesso di 1 il valore di N e si impone che la soluzione $y_k^{(N+1)}$ soddisfi, al posto delle (77), le condizioni iniziali

$$y_0^{(N+1)} = y_0, \qquad y_{N+1}^{(N+1)} = 0.$$

Allora $y_k^{(N+1)}$ è soluzione di un sistema lineare di N equazioni, le cui prime $N-2$ equazioni coincidono con quelle del sistema (80), mentre le ultime due sono

$$\begin{cases} a_0(N-2)\; y_{N-2}^{(N+1)} + a_1(N-2)\; y_{N-1}^{(N+1)} + y_N^{(N)} = b(N-2), \\ \\ a_0(N-1)\; y_{N-1}^{(N+1)} + a_1(N-1)\; y_N^{(N+1)} = b(N-1). \end{cases}$$

Per la riduzione in forma bidiagonale superiore di questo sistema si usano ancora le (83), aggiungendo le ulteriori relazioni

$$\begin{aligned} c_{N+1} &= -\big[a_1(N-1)\; c_N + a_0(N-1)\; c_{N-1}\big], \\ d_N &= a_0(N-1)\; d_{N-1} - b(N-1)\; c_N. \end{aligned} \tag{84}$$

Si ottiene così il sistema

$$\begin{cases} c_{k+1} y_k^{(N+1)} - c_k y_{k+1}^{(N+1)} = d_k, \quad k = 1, \ldots, N-1, \\ \\ c_{N+1} y_N^{(N+1)} = d_N, \end{cases} \tag{85}$$

dalla cui ultima equazione si ricava

$$y_N^{(N+1)} = \frac{d_N}{c_{N+1}}\,.$$

Confrontando fra di loro i valori delle due successioni $\{y_k^{(N)}\}$ e $\{y_k^{(N+1)}\}$ si può dare una stima dell'errore. Da (82) e (85) si ricava

$$y_k^{(N+1)} - y_k^{(N)} = \frac{c_k}{c_{k+1}}\,(y_{k+1}^{(N+1)} - y_{k+1}^{(N)}), \quad k = 1, \ldots, N-1,$$

e, posto

$$e_k^{(N)} = y_k^{(N+1)} - y_k^{(N)} \quad \text{e} \quad e_N^{(N)} = \frac{d_N}{c_{N+1}}\,,$$

si ottiene la relazione

$$e_k^{(N)} = \frac{c_k}{c_{k+1}}\, e_{k+1}^{(N)} = \frac{c_k}{c_N}\, e_N^{(N)} = \frac{c_k d_N}{c_N c_{N+1}}\,, \tag{86}$$

che ci consente di determinare $e_k^{(N)}$ senza dover calcolare effettivamente le componenti delle due soluzioni $y_k^{(N)}$ e $y_k^{(N+1)}$, sfruttando solo i coefficienti c_k e d_k, calcolati con le (83) e (84). Quindi il metodo di Olver per calcolare la soluzione minimale y_k, $k = 1, \ldots, m$, assegnata la condizione iniziale y_0 e fissata una tolleranza ϵ, opera nel modo seguente:

1. per $N = m + 1$, si calcolano c_k, $k = 1, \ldots, N$, e d_k, $k = 1, \ldots, N - 1$, con le (83) e $\gamma = \max\limits_{k=1,\ldots,m} |c_k|$;
2. si calcolano c_{N+1} e d_N con le (84) e
$$|\epsilon^{(N)}| = \max_{k=1,\ldots,m} |e_k^{(N)}| = \gamma \left| \frac{d_N}{c_N c_{N+1}} \right|,$$
3. se $|\epsilon^{(N)}| > \epsilon$, si pone $N = N + 1$ e si torna al punto 2. Altrimenti il valore di N è quello richiesto e dal sistema (85), con una sostituzione all'indietro, si calcolano i valori
$$y_k^{(N+1)} = \frac{1}{c_{k+1}}\, (d_k + c_k y_{k+1}^{(N+1)}), \quad \text{per} \quad k = N - 1, \ldots, 1.$$

Anche in questo caso il calcolo effettivo della soluzione viene fatto una sola volta, dopo che è stato determinato il valore di N opportuno.

Il metodo di Olver può essere ovviamente applicato anche alle equazioni omogenee: il costo computazionale del calcolo di $y_k^{(N)}$ è quasi doppio di quello del metodo di Miller. Se però si considera anche il costo della determinazione del valore di N per mezzo della (78), il metodo di Olver è più conveniente.

4.26 Esempio. Per $k \geq 0$ l'integrale

$$y_k = \int_1^\infty \frac{dx}{x^k(2x^2 + 5x + 2)}$$

soddisfa la relazione ricorrente

$$y_{k+2} + \frac{5}{2}\, y_{k+1} + y_k = \frac{1}{2(k+1)}\,, \tag{87}$$

come si può verificare integrando per parti. Poiché la successione y_k è decrescente, mentre l'equazione omogenea associata ha una soluzione crescente, il valore cercato dell'integrale rappresenta una soluzione minimale

della (87). Per calcolare i valori y_k, $k = 1, \ldots, 20$, si applica il metodo di Olver, utilizzando la condizione iniziale $y_0 = \dfrac{1}{3} \log 2$. Fissato $\epsilon = 10^{-5}$ risulta $|\epsilon^{(N)}| < \epsilon$ per $N = 29$. Con la sostituzione all'indietro si ottiene

k	$y_k^{(30)}$	k	$y_k^{(30)}$
20	$0.5551159\ 10^{-2}$	.	...
19	$0.5839907\ 10^{-2}$	4	$0.2707386\ 10^{-1}$
18	$0.6164853\ 10^{-2}$	3	$0.3546477\ 10^{-1}$
17	$0.6525729\ 10^{-2}$	2	$0.5093088\ 10^{-1}$
.	...	1	$0.8720797\ 10^{-1}$

■

La trasformazione del sistema (80) nel sistema (82) con le (83) consiste essenzialmente nel metodo di eliminazione di Gauss applicato senza pivot, e può essere instabile quando $|a_1|$ è molto più piccolo di $|a_0| + 1$. Se ciò accade, è necessario adottare delle tecniche di risoluzione del sistema lineare più stabili.

Il metodo di Olver può essere applicato con le opportune varianti anche quando la condizione iniziale è sostituita da una condizione di normalizzazione, mentre non si presta facilmente ad essere esteso al caso di equazioni di ordine superiore al secondo. Esistono altri algoritmi iterativi che possono essere applicati in questo caso, si veda [5].

Esercizi proposti

4.1 Si verifichino le seguenti relazioni formali

$$\begin{aligned} &\text{a)} \quad \Delta E = E\Delta \\ &\text{b)} \quad E^r E^s = E^{r+s} \\ &\text{c)} \quad \Delta^r \Delta^s = \Delta^{r+s}. \end{aligned}$$

(Traccia: a) $\Delta E y_k = \Delta y_{k+1} = y_{k+2} - y_{k+1} = E(y_{k+1} - y_k) = E\Delta y_k$;
b) $E^r E^s y_k = E^r y_{k+s} = y_{k+r+s} = E^{r+s} y_k$;
c) si dimostri che $(E - I)^r (E - I)^s y_k = (E - I)^{r+s} y_k$.)

4.2 Si verifichi che per $\alpha, \beta \in \mathbf{R}$ ($\alpha > 0$ dove occorre) è

$$\begin{aligned} &\Delta\alpha = 0, \\ &\Delta x^{(m)} = m x^{(m-1)}, \end{aligned}$$

$$\Delta\binom{x}{m} = \binom{x}{m-1}, \quad \text{dove} \quad \binom{x}{m} = \frac{x^{(m)}}{m!}\,,$$

$$\Delta x^m = \sum_{i=0}^{m-1} \binom{m}{i} x^i,$$

$$\Delta\alpha^x = \alpha^x(\alpha - 1),$$

$$\Delta \mathrm{e}^{\alpha x} = \mathrm{e}^{\alpha x}(\mathrm{e}^{\alpha} - 1),$$

$$\Delta \log x = \log(1 + \frac{1}{x}),$$

$$\Delta \log_\alpha x = \log_\alpha(1 + \frac{1}{x}),$$

$$\Delta \sin \alpha x = 2 \sin \frac{\alpha}{2}\ \cos[\alpha(x + \frac{1}{2})],$$

$$\Delta \cos \alpha x = -2 \sin \frac{\alpha}{2}\ \sin[\alpha(x + \frac{1}{2})],$$

$$\Delta \tan x = \frac{\sec^2 x \tan 1}{1 - \tan x \tan 1}\,,$$

$$\Delta \arctan x = \arctan \frac{1}{x^2 + x + 1}\,.$$

(Traccia: ad esempio

$$\Delta \sin \alpha x = \sin[\alpha(x+1)] - \sin \alpha x = 2 \sin \frac{\alpha}{2}\ \cos \frac{2\alpha x + \alpha}{2}\,.\ )$$

4.3 Si dimostri che

$$\Delta^i x^m = \begin{cases} m! & \text{se } i = m, \\ \\ 0 & \text{se } i > m; \end{cases}$$

$$\Delta^i x^{(m)} = \frac{m!}{(m-i)!}\ x^{(m-i)} = m^{(i)} x^{(m-i)};$$

$$\Delta^i x^{(-m)} = (-1)^i\ \frac{(m+i-1)!}{(m-1)!}\ x^{(-m-i)}.$$

(Traccia: si sfruttino le (11), (12), (13), (14) e (16).)

4.4 Le potenze fattoriali possono essere definite anche per funzioni $f(x)$ nel modo seguente per $m > 0$

$$(f(x))^{(0)} = 1, \quad (f(x))^{(m)} = f(x - m + 1)(f(x))^{(m-1)},$$

$$(f(x))^{(-m)} = \frac{(f(x))^{(-m+1)}}{f(x+m)}\,.$$

Si verifichi che

$$\Delta^i(\alpha x+\beta)^{(m)} = \frac{\alpha^i\, m!}{(m-i)!}\,(\alpha x+\beta)^{(m-i)},$$
$$\Delta^i(\alpha x+\beta)^{(-m)} = (-1)^i\,\frac{\alpha^i(m+i-1)!}{(m-1)!}\,(\alpha x+\beta)^{(-m-i)}.$$

(Traccia: si verifichi che

$$\Delta(\alpha x+\beta)^{(m)} = m\alpha(\alpha x+\beta)^{(m-1)}$$

e che

$$\Delta(\alpha x+\beta)^{(-m)} = -\alpha m(\alpha x+\beta)^{(-m-1)},$$

e si proceda per induzione.)

4.5 Si verifichi che

$$\Delta^m \sin(\alpha x+\beta) = \big(2\sin\frac{\alpha}{2}\big)^m \sin[\alpha x+\beta+\frac{m}{2}(\alpha+\pi)],$$
$$\Delta^m \cos(\alpha x+\beta) = \big(2\sin\frac{\alpha}{2}\big)^m \cos[\alpha x+\beta+\frac{m}{2}(\alpha+\pi)].$$

(Traccia: si proceda per induzione utilizzando la relazione

$$\Delta \sin(\alpha x+\beta) = 2\sin\frac{\alpha}{2}\sin[\alpha x+\beta+\frac{1}{2}(\alpha+\pi)].\ )$$

4.6 Si verifichi che

a) $$\sum_{i=0}^{n}(-1)^i\binom{n}{i}\frac{1}{i+1} = \frac{1}{n+1}\,;$$

b) $$\sum_{i=0}^{n}(-1)^i\binom{n}{i}\frac{1}{2i+1} = \frac{2^{2n}\ (n!)^2}{(2n+1)!}\,;$$

c) $$\sum_{i=0}^{n}(-1)^i\binom{n}{i}i^n = (-1)^n n!.$$

(Traccia: applicando la (6) per $k=0$ si ottiene

$$\Delta^n y_0 = (-1)^n\sum_{i=0}^{n}(-1)^i\binom{n}{i}y_i.$$

Per a) si scelga $y_i = i^{(-1)}$ e si tenga presente che per l'esercizio 4.3 è

$$\Delta^n y_k = (-1)^n n!\, k^{(-n-1)}, \qquad k^{(-n-1)}\Big|_{k=0} = \frac{1}{(n+1)!}\,.$$

Per b) si scelga $y_i = (2i-1)^{(-1)}$ e si tenga presente che per l'esercizio 4.4 è

$$\Delta^n y_k = (-1)^n 2^n n!\, (2k-1)^{(-n-1)},$$
$$(2k-1)^{(-n-1)}\Big|_{k=0} = \frac{1}{1\cdot 3\cdot 5\cdots(2n+1)} = \frac{2^n n!}{(2n+1)!}\,.$$

Per c) si scelga $y_i = i^n$ e si tenga presente che per l'esercizio 4.3 è $\Delta^n y_k = n!$.)

4.7 Si verifichi che per le potenze fattoriali non vale l'uguaglianza

$$x^{(m)}x^{(n)} = x^{(m+n)}.$$

4.8 Si esprimano le seguenti funzioni razionali come combinazioni di potenze fattoriali

a) $(2x+1)^2$, b) $4x^2-1$, c) $(3x-1)(3x+2)$, d) $\dfrac{1}{x(x+1)}$,

e) $\dfrac{1}{4x^2-1}$, f) $\dfrac{1}{x(x+3)}$, g) $\dfrac{1-3x}{x(x+1)(x+2)}$.

(Risposta: a) $4x^{(2)} + 8x^{(1)} + x^{(0)}$; b) $4x^{(2)} + 4x^{(1)} - x^{(0)} = (2x+1)^{(2)}$;
c) $9x^{(2)} + 12x^{(1)} - 2x^{(0)} = (3x+2)^{(2)}$; d) $(x-1)^{(-2)}$;
e) $(2x-3)^{(-2)}$; f) $2(x-1)^{(-4)} - 2(x-1)^{(-3)} + (x-1)^{(-2)}$;
g) $(x^{(0)} - 3x^{(1)})(x-1)^{(-3)}$.)

4.9 Si dimostrino le formule (18) e (20) di ricorrenza per i numeri di Stirling di prima e di seconda specie.

(Traccia: si proceda per induzione su m nella (18) e su i nella (20). Dalla (17) si ha per $m+1$

$$x^{(m+1)} = \sum_{i=1}^{m+1} s_i^{(m+1)} x^i.$$

D'altra parte è

$$x^{(m+1)} = x^{(m)}(x-m) = \left[\sum_{i=1}^{m} s_i^{(m)} x^i\right](x-m).$$

Confrontando le due espressioni si ottiene la (18). Per la (20) si proceda in modo analogo.)

4.10 Si dimostri che per $m > 1$ è

a) $$\sum_{i=1}^{m} s_i^{(m)} = 0,$$

b) $$\sum_{i=1}^{m} (-1)^i s_i^{(m)} = (-1)^m m!,$$

c) $$\sum_{i=1}^{m} |s_i^{(m)}| = m!,$$

d) $$s_i^{(m)} = \frac{1}{i!} \frac{d^i}{dx^i} x^{(m)} \Big|_{x=0}.$$

e) Siano $A, B \in \mathbf{R}^{n\times n}$ definite da

$$a_{m,i} = \begin{cases} s_i^{(m)} & \text{per } m \geq i, \\ 0 & \text{altrimenti,} \end{cases} \qquad b_{i,m} = \begin{cases} S_m^{(i)} & \text{per } i \geq m, \\ 0, & \text{altrimenti.} \end{cases}$$

Si dimostri che $B = A^{-1}$.

(Traccia: a) si proceda per induzione, oppure dalla (17) per $x = 1$ si ha

$$\sum_{i=1}^{m} s_i^{(m)} = 1^{(m)} = 0, \quad \text{per } m > 1;$$

b) in modo analogo; c) si ottiene da b) verificando che i segni dei numeri di Stirling di prima specie sono alternati; d) si derivi la (17); e) segue dalla linearità delle (17) e (19).)

4.11 Sia $p_n(x)$ un polinomio di grado n. Si dimostri la seguente formula di *Gregory-Newton*

$$p_n(x) = \sum_{i=0}^{n} \frac{x^{(i)}}{i!} \Delta^i p_n(0).$$

(Traccia: posto per la (16)

$$p_n(x) = \sum_{i=0}^{n} b_i x^{(i)},$$

poiché $x^{(0)} = 1$, si ha applicando la (11)

$$
\begin{aligned}
&p_n(0) = b_0,\\
&\Delta p_n(x) = \sum_{i=1}^{n} b_i i x^{(i-1)}, \qquad \Delta p_n(0) = b_1,\\
&\Delta^2 p_n(x) = \sum_{i=2}^{n} b_i i(i-1) x^{(i-2)}, \qquad \Delta^2 p_n(0) = 2b_2,\\
&\qquad\qquad \dots\\
&\Delta^n p_n(x) = b_n n!\, x^{(0)}, \qquad\qquad \Delta^n p_n(0) = n!\, b_n. \ )
\end{aligned}
$$

4.12 Si dimostri la seguente *regola di Leibniz* per le differenze

$$\Delta^n(fg) = \sum_{i=0}^{n} \binom{n}{i} (\Delta^i f)(\Delta^{n-i} E^i g);$$

si applichi in particolare al caso $\Delta^n(x\alpha^x)$.

(Traccia: si proceda per induzione oppure si considerino gli operatori E_1 e E_2 definiti da

$$E_1[f(x)g(x)] = f(x+1)g(x) \quad \text{e} \quad E_2[f(x)g(x)] = f(x)g(x+1),$$

e gli operatori $\Delta_1 = E_1 - I$ e $\Delta_2 = E_2 - I$. Allora

$$\Delta = E - I = E_1 E_2 - I = \Delta_2 + \Delta_1 E_2,$$

si verifichi che Δ_2 e $\Delta_1 E_2$ commutano e si sviluppi

$$\Delta^n(fg) = (\Delta_2 + \Delta_1 E_2)^n (fg);$$

si tenga conto del fatto che

$$\Delta_2^{n-i} \Delta_1^i E_2^i (fg) = (\Delta^i f)(\Delta^{n-i} E^i g).$$

Nel caso particolare, poiché $\Delta x = 1$ e $\Delta^2 x = 0$, è

$$\Delta^n(x\alpha^x) = x\Delta^n \alpha^x + n\Delta^{n-1} E\alpha^x = x\alpha^x(\alpha-1)^n + n\alpha^{x+1}(\alpha-1)^{n-1}. \ )$$

4.13 Si verifichi che

a) gli operatori Δ e Σ non commutano;

b) vale la relazione

$$(E - I)\sum_{k=1}^{n} y_k = (E^n - I)y_1.$$

(Traccia: a) è $\Delta\Sigma y_k = y_k$ e $\Sigma\Delta y_k = y_k + c_k$, con c_k costante periodica;
b) si applichi la (4) e si sfrutti la linearità di Δ, si ottiene

$$\sum_{k=1}^{n} \Delta y_k = y_k \Big|_1^{n+1} .\ )$$

4.14 Si verifichi che per α, $\beta \in \mathbf{R}$ ($\alpha > 0$ dove occorre) è, a meno di una costante periodica

$$\sum \alpha = \alpha x,$$
$$\sum (\alpha x + \beta)^{(m)} = \frac{(\alpha x + \beta)^{(m+1)}}{\alpha(m+1)}\,,\ m \text{ intero},\ m \neq -1,$$
$$\sum \alpha^x = \frac{\alpha^x}{\alpha - 1}\,,$$
$$\sum \mathrm{e}^{\alpha x} = \frac{\mathrm{e}^{\alpha x}}{\mathrm{e}^{\alpha} - 1}\,,$$
$$\sum \sin \alpha x = -\,\frac{\cos[\alpha(x - \frac{1}{2})]}{2\sin\dfrac{\alpha}{2}}\,,$$
$$\sum \cos \alpha x = \frac{\sin[\alpha(x - \frac{1}{2})]}{2\sin\dfrac{\alpha}{2}}\,.$$

(Traccia: si faccia riferimento agli esercizi 4.2 e 4.4.)

4.15 Si calcoli $\sum x \cos \alpha x$.

(Traccia: con la relazione di somma per parti si ha, a meno di una costante periodica

$$\sum x \cos \alpha x = \frac{x \sin[\alpha(x - \frac{1}{2})]}{2\sin\dfrac{\alpha}{2}} \;+\; \frac{\cos \alpha x}{4\sin^2\dfrac{\alpha}{2}}\ .\ )$$

4.16 Si verifichi che

a) $1^2 + 3^2 + 5^2 + \ldots + (2n-1)^2 = \dfrac{n}{3}\,(4n^2 - 1)$,

b) $2^2 + 5^2 + 8^2 + \ldots + (3n-1)^2 = \dfrac{n}{2}\,(6n^2 + 3n - 1)$,

c) $1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \ldots + n(n+1) = \dfrac{n}{3}\,(n+1)(n+2)$,

d) $1 \cdot 3 + 3 \cdot 5 + 5 \cdot 7 + \ldots + (2n-1)(2n+1) = \dfrac{n}{3}\,(4n^2+6n-1)$,

e) $2 \cdot 5 + 5 \cdot 8 + 8 \cdot 11 + \ldots + (3n-1)(3n+2) = n(3n^2+6n+1)$,

f) $1^2 \cdot 2 + 2^2 \cdot 3 + 3^2 \cdot 4 + \ldots + n^2(n+1) = \dfrac{n}{12}\,(n+1)(n+2)(3n+1)$

g) $1 \cdot 2^1 + 2 \cdot 2^2 + 3 \cdot 2^3 + \ldots + n\,2^n = 2 + 2^{n+1}(n-1)$,

h) $1^2 \cdot 2 + 2^2 \cdot 2^2 + 3^2 \cdot 2^3 + \ldots + n^2 2^n = 2^{n+1}(n^2-2n+3) - 6$,

i) $\dfrac{1}{2} + \dfrac{2}{2^2} + \dfrac{3}{2^3} + \ldots + \dfrac{n}{2^n} = 2 - \dfrac{n+2}{2^n}$,

l) $1{\cdot}2{\cdot}3 + 2{\cdot}3{\cdot}4 + 3{\cdot}4{\cdot}5 + \ldots + n(n+1)(n+2) = \dfrac{n}{4}\,(n+1)(n+2)(n+3)$,

m) $1 \cdot 3 \cdot 5 + 3 \cdot 5 \cdot 7 + 5 \cdot 7 \cdot 9 + \ldots + (2n-1)(2n+1)(2n+3)$
$= n(n+2)(2n^2+4n-1)$,

n) $1{\cdot}3{\cdot}5{\cdot}7 + 3{\cdot}5{\cdot}7{\cdot}9 + 5{\cdot}7{\cdot}9{\cdot}11 + \ldots + (2n-1)(2n+1)(2n+3)(2n+5)$
$= \dfrac{21}{2} + \dfrac{1}{10}(2n-1)(2n+1)(2n+3)(2n+5)(2n+7)$,

o) $\dfrac{1}{1\cdot 2} + \dfrac{1}{2\cdot 3} + \dfrac{1}{3\cdot 4} + \ldots + \dfrac{1}{n(n+1)} = \dfrac{n}{n+1}$,

p) $\dfrac{1}{1\cdot 3} + \dfrac{1}{3\cdot 5} + \dfrac{1}{5\cdot 7} + \ldots + \dfrac{1}{(2n-1)(2n+1)} = \dfrac{n}{2n+1}$,

q) $\dfrac{1}{1\cdot 4} + \dfrac{1}{2\cdot 5} + \dfrac{1}{3\cdot 6} + \ldots + \dfrac{1}{n(n+3)} = \dfrac{n}{18}\,\dfrac{11n^2+48n+49}{(n+1)(n+2)(n+3)}$,

r) $\dfrac{1}{1\cdot 2\cdot 3} + \dfrac{1}{2\cdot 3\cdot 4} + \dfrac{1}{3\cdot 4\cdot 5} + \ldots + \dfrac{1}{n(n+1)(n+2)}$
$= \dfrac{n(n+3)}{4(n+1)(n+2)}$,

s) $\dfrac{1}{1\cdot 3\cdot 5} + \dfrac{1}{3\cdot 5\cdot 7} + \dfrac{1}{5\cdot 7\cdot 9} + \ldots + \dfrac{1}{(2n-1)(2n+1)(2n+3)}$
$= \dfrac{n(n+2)}{3(2n+1)(2n+3)}$,

t) $\dfrac{1}{1\cdot 4\cdot 7} + \dfrac{1}{4\cdot 7\cdot 10} + \dfrac{1}{7\cdot 10\cdot 13} + \ldots + \dfrac{1}{(3n-2)(3n+1)(3n+4)}$
$= \dfrac{n(3n+5)}{8(3n+1)(3n+4)}$,

u) $\dfrac{1}{1\cdot 2\cdot 3} + \dfrac{4}{2\cdot 3\cdot 4} + \dfrac{7}{3\cdot 4\cdot 5} + \ldots + \dfrac{3n-2}{n(n+1)(n+2)}$
$= \dfrac{n^2}{(n+1)(n+2)}$,

v) $$\frac{1}{2\cdot 3\cdot 4}+\frac{3}{3\cdot 4\cdot 5}+\frac{5}{4\cdot 5\cdot 6}+\ldots+\frac{2n-1}{(n+1)(n+2)(n+3)} = \frac{5n^2+n}{12(n+2)(n+3)},$$

z) $$\frac{5}{1\cdot 3\cdot 5}+\frac{6}{3\cdot 5\cdot 7}+\frac{7}{5\cdot 7\cdot 9}+\ldots+\frac{n+4}{(2n-1)(2n+1)(2n+3)} = \frac{n(11n+19)}{6(2n+1)(2n+3)},$$

(Traccia: a) $\sum(2x-1)^2=\frac{x-1}{3}\,[4(x-1)^2-1]$,

b) $\sum(3x-1)^2=\frac{x}{2}\,(6x^2-15x+5)$, c) $\sum x(x+1)=\frac{x}{3}\,(x^2-1)$,

d) $\sum(2x-1)(2x+1)=\frac{x}{3}\,(4x^2-6x-1)$,

e) $\sum(3x-1)(3x+2)=x(3x^2-3x-2)$,

f) $\sum x^2(x+1)=\frac{x(x-1)}{12}\,(3x^2+x-2)$

g) $\sum x2^x=(x-2)2^x$, h) $\sum x^2 2^x=(x^2-4x+6)2^x$,

i) $\sum x2^{-x}=-(x+1)2^{-x+1}$, l) $\sum(x+2)^{(3)}=\frac{1}{4}\,(x+2)^{(4)}$,

m) $\sum(2x+3)^{(3)}=\frac{1}{8}\,(2x+3)^{(4)}$, n) $\sum(2x+5)^{(4)}=\frac{1}{10}(2x+5)^{(5)}$,

o) $\sum(x-1)^{(-2)}=-(x-1)^{(-1)}$, p) $\sum(2x-3)^{(-2)}=-\frac{1}{2}\,(2x-3)^{(-1)}$,

q) $\sum\frac{1}{x(x+3)}=-\frac{1}{3}\,(3x^2+6x+2)(x-1)^{(-3)}$,

r) $\sum(x-1)^{(-3)}=-\frac{1}{2}\,(x-1)^{(-2)}$,

s) $\sum(2x-3)^{(-3)}=-\frac{1}{4}\,(2x-3)^{(-2)}$,

t) $\sum(3x-5)^{(-3)}=-\frac{1}{6}\,(3x-5)^{(-2)}$,

u) $\sum(3x-2)(x-1)^{(-3)}=-(3x-1)(x-1)^{(-2)}$,

v) $\sum(2x-1)x^{(-3)}=-\frac{1}{2}(4x+1)x^{(-2)}$,

z) $\sum(x+4)(2x-3)^{(-3)}=-\frac{1}{8}\,\frac{4x+7}{4x^2-1}$.)

4.17 Si verifichi che

a) $\sum_{k=1}^{\infty} k2^{-k}=2$,

b) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1,$

c) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{4k^2-1} = \frac{1}{2},$

d) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+3)} = \frac{11}{18},$

e) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)} = \frac{1}{4},$

f) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{(2k-1)(2k+1)(2k+3)} = \frac{1}{12},$

g) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{(3k-2)(3k+1)(3k+4)} = \frac{1}{24},$

h) $\displaystyle\sum_{k=1}^{\infty} \frac{3k-2}{k(k+1)(k+2)} = 1,$

i) $\displaystyle\sum_{k=1}^{\infty} \frac{2k-1}{(k+1)(k+2)(k+3)} = \frac{5}{12},$

l) $\displaystyle\sum_{k=1}^{\infty} \frac{n+4}{(2k-1)(2k+1)(2k+3)} = \frac{11}{24}.$

4.18 Si verifichi che

a) $\displaystyle\sum_{k=1}^{n} \sin k\alpha = \frac{\cos\frac{\alpha}{2} - \cos\left(n+\frac{1}{2}\right)\alpha}{2\sin\frac{\alpha}{2}},$

b) $\displaystyle\frac{1}{2} + \sum_{k=1}^{n} \cos k\alpha = \frac{\sin\left(n+\frac{1}{2}\right)\alpha}{2\sin\frac{\alpha}{2}},$

c) $\displaystyle\sum_{k=1}^{n} \beta^k \sin k\alpha = \frac{\beta^{n+2}\sin n\alpha - \beta^{n+1}\sin(n+1)\alpha + \beta\sin\alpha}{\beta^2 - 2\beta\cos\alpha + 1},$

d) $\displaystyle\frac{1}{2} + \sum_{k=1}^{n} \beta^k \cos k\alpha = \frac{\beta^{n+2}\cos n\alpha - \beta^{n+1}\cos(n+1)\alpha + \frac{1}{2}(1-\beta^2)}{\beta^2 - 2\beta\cos\alpha + 1},$

e) $\displaystyle\sum_{k=0}^{n} \frac{1}{\sin 2^k\alpha} = \frac{\cos\frac{\alpha}{2}}{\sin\frac{\alpha}{2}} - \frac{\cos 2^n\alpha}{\sin 2^n\alpha},$

f) $\displaystyle\sum_{k=1}^{n} (-1)^k \cos\frac{kj\pi}{n} = \begin{cases} 0 & \text{se } n+j \text{ pari,} \\ -1 & \text{se } n+j \text{ dispari.} \end{cases}$

(Traccia: a), b) si veda l'esercizio 4.14,

c) $\sum \beta^k \sin k\alpha = \dfrac{\beta^{k+1}\sin(k-1)\alpha - \beta^k \sin k\alpha}{\beta^2 - 2\beta\cos\alpha + 1}$,

d) $\sum \beta^k \cos k\alpha = \dfrac{\beta^{k+1}\cos(k-1)\alpha - \beta^k \cos k\alpha}{\beta^2 - 2\beta\cos\alpha + 1}$,

e) $\sum \dfrac{1}{\sin 2^k\alpha} = -\dfrac{\cos 2^{k-1}\alpha}{\sin 2^{k-1}\alpha}$,

f) si applichi la d) con $\beta = -1$.)

4.19 Si dimostri la seguente relazione, detta *trasformazione di Abel*

$$\sum_{k=0}^{n-1} y_k z_k = y_n \sum_{k=0}^{n-1} z_k - \sum_{k=0}^{n-1} \left[\Delta y_k \sum_{i=0}^{k} z_i \right].$$

Si applichi al caso delle somme g), h) e i) dell'esercizio 4.16.

(Traccia: dalla relazione di somma per parti (22) per il teorema fondamentale del calcolo delle somme (24) si ha

$$\begin{aligned}\sum_{k=0}^{n-1}(y_k \Delta v_k) &= \Big[y_k v_k - \sum (v_{k+1}\Delta y_k) \Big]_0^n \\ &= y_n v_n - y_0 v_0 - \sum_{k=0}^{n-1}(v_{k+1}\Delta y_k);\end{aligned}$$

inoltre

$$\begin{aligned}\sum_{k=0}^{n-1}(v_{k+1}\Delta y_k) &= \sum_{k=0}^{n-1}\big[(v_{k+1} - v_0)\Delta y_k\big] + v_0 \sum_{k=0}^{n-1} \Delta y_k \\ &= \sum_{k=0}^{n-1}\Big[\Delta y_k \sum_{i=0}^{k} \Delta v_i\Big] + v_0(y_n - y_0).\end{aligned}$$

Sostituendo si ha

$$\begin{aligned}\sum_{k=0}^{n-1}(y_k \Delta v_k) &= y_n(v_n - v_0) - \sum_{k=0}^{n-1}\Big[\Delta y_k \sum_{i=0}^{k} \Delta v_i\Big] \\ &= y_n \sum_{k=0}^{n-1} \Delta v_k - \sum_{k=0}^{n-1}\Big[\Delta y_k \sum_{i=0}^{k} \Delta v_i\Big].\end{aligned}$$

Si ponga $\Delta v_k = z_k$.)

4.20 Si calcoli

a) $\sum_{k=1}^{n} k \sin k\alpha$ b) $\sum_{k=1}^{n} k \cos k\alpha$

c) $\sum_{k=1}^{n} \sin^2 k\alpha$ d) $\sum_{k=1}^{n} \cos^2 k\alpha$

e) $\sum_{k=1}^{n} k^2 \sin k\alpha$ f) $\sum_{k=1}^{n} k^2 \cos k\alpha$.

(Traccia: si applichi la trasformazione di Abel dell'esercizio precedente.)

4.21 Sia $z > -1$ ed m un intero tale che $0 \le m < z+1$. Si dimostri che

a)
$$\frac{d^m}{dx^m}\,(1+x)^z = \frac{\Gamma(z+1)}{\Gamma(z-m+1)}\,(1+x)^{z-m}$$
$$\frac{d^m}{dx^m}\,(1-x)^z = (-1)^m\,\frac{\Gamma(z+1)}{\Gamma(z-m+1)}\,(1-x)^{z-m}$$

b)
$$\frac{\Gamma(2z+1)}{\Gamma(2z-m+1)} = \sum_{i=0}^{m} \binom{m}{i} \frac{[\Gamma(z+1)]^2}{\Gamma(z-i+1)\,\Gamma(z-m+i+1)}\,.$$

(Traccia: a) si utilizzino la (15) e la (29); b) si applichi la formula di Leibniz

$$\begin{aligned}\frac{d^m}{dx^m}\,(1+x)^{2z} &= \frac{d^m}{dx^m}\,\left[(1+x)^z(1+x)^z\right] \\ &= \sum_{i=0}^{m} \binom{m}{i} \left[\frac{d^i}{dx^i}\,(1+x)^z\right] \left[\frac{d^{m-i}}{dx^{m-i}}\,(1+x)^z\right],\end{aligned}$$

e si esprimano le derivate al primo e al secondo membro per mezzo della formula a).)

4.22 a) Si dimostri che per $z > -1$ è

$$\int_{-1}^{1} (1-x^2)^z\,dx = 2^{2z+1}\,\frac{[\Gamma(z+1)]^2}{\Gamma(2z+2)}\,.$$

b) Si dimostri che $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

c) Si calcoli $\Gamma\left(n+\frac{1}{2}\right)$, per n intero.

(Traccia: a) ponendo $t = x^2$ si ha dalla (26)

$$\Gamma(z+1) = \int_0^\infty t^z \mathrm{e}^{-t}\,dt = 2\int_0^\infty x^{2z+1}\mathrm{e}^{-x^2}\,dx,$$

e quindi

$$\begin{aligned}\big[\Gamma(z+1)\big]^2 &= 4\Big[\int_0^\infty x^{2z+1}\mathrm{e}^{-x^2}\,dx\Big]\,\Big[\int_0^\infty y^{2z+1}\mathrm{e}^{-y^2}\,dy\Big]\\ &= 4\int_0^\infty\int_0^\infty (xy)^{2z+1}\mathrm{e}^{-(x^2+y^2)}\,dx\,dy\Big].\end{aligned}$$

Ponendo $x=\rho\cos\theta$, $y=\rho\sin\theta$, si ha

$$\big[\Gamma(z+1)\big]^2 = 4\Big[\int_0^\infty \rho^{4z+3}\mathrm{e}^{-\rho^2}\,d\rho\Big]\,\Big[\int_0^{\pi/2}(\cos\theta\sin\theta)^{2z+1}\,d\theta\Big]$$

e ponendo $\rho^2=u$ e $\sin^2\theta=v$, si ha

$$\begin{aligned}\big[\Gamma(z+1)\big]^2 &= \Big[\int_0^\infty u^{2z+1}\mathrm{e}^{-u}\,du\Big]\,\Big[\int_0^1 v^z(1-v)^z\,dv\Big]\\ &= \Gamma(2z+2)\int_0^1 v^z(1-v)^z\,dv.\end{aligned}$$

D'altra parte, ponendo $x=2v-1$, si ha

$$\int_{-1}^1 (1-x^2)^z\,dx = 2^{2z+1}\int_0^1 v^z(1-v)^z\,dv.$$

b) Ponendo $z=-\dfrac{1}{2}$ segue da a)

$$\Big[\Gamma\big(\frac{1}{2}\big)\Big]^2 = \int_{-1}^1 (1-x^2)^{-1/2}\,dx = \arcsin x\Big|_{-1}^1 = \pi.$$

c) Per $n\geq 1$ si ha

$$\begin{aligned}\Gamma\Big(n+\frac{1}{2}\Big) &= \Big(n-\frac{1}{2}\Big)\Gamma\Big(n-\frac{1}{2}\Big) = \Big(n-\frac{1}{2}\Big)\Big(n-\frac{3}{2}\Big)\cdots\frac{1}{2}\,\Gamma\Big(\frac{1}{2}\Big)\\ &= \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^n}\,\Gamma\Big(\frac{1}{2}\Big),\end{aligned}$$

e poiché

$$(2n)! = 1\cdot 3\cdot 5\cdots(2n-1)\cdot n!\,2^n,$$

risulta

$$\Gamma\Big(n+\frac{1}{2}\Big) = \frac{(2n)!}{n!\,2^{2n}}\,\sqrt{\pi}.$$

Per $n \le -1$ è

$$\Gamma\Big(n+\frac{1}{2}\Big) = \frac{1}{n+\frac{1}{2}}\,\Gamma\Big[(n+1)+\frac{1}{2}\Big] = \frac{1}{\big(n+\frac{1}{2}\big)\,\big(n+\frac{3}{2}\big)\ldots\big(-1+\frac{1}{2}\big)}\,\Gamma\Big(\frac{1}{2}\Big)$$

$$= \frac{(-1)^n(-n)!\,2^{2n}}{(-2n)!}\,\sqrt{\pi}.\ )$$

4.23 Per k intero si calcolino gli integrali

$$\text{a)}\ \int_0^\infty x^k \mathrm{e}^{-2x}\,dx, \qquad \text{b)}\ \int_0^\infty x^k \mathrm{e}^{-x^2}\,dx, \qquad \text{c)}\ \int_0^\infty x^k \mathrm{e}^{-x^3}\,dx.$$

(Risposta: a) $\dfrac{1}{2^{k+1}}\,\Gamma(k+1)$, b) $\dfrac{1}{2}\,\Gamma\Big(\dfrac{k+1}{2}\Big)$, c) $\dfrac{1}{3}\,\Gamma\Big(\dfrac{k+1}{3}\Big)$.)

4.24 Si dimostri che

$$\text{a)}\ \ \Psi(x) = \lim_{n\to\infty}\Big[\log n - \frac{1}{x+1} - \frac{1}{x+2} - \ldots - \frac{1}{x+n+1}\Big], \quad x > -1,$$

per cui

$$\gamma = \lim_{n\to\infty}\Big[1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n+1} - \log n\Big];$$

$$\text{b)} \qquad \Psi(x) - \log x = \int_0^\infty \mathrm{e}^{-xt}\Big[\frac{1}{t} - \frac{\mathrm{e}^{-t}}{1-\mathrm{e}^{-t}}\Big]\,dt, \quad x > 0,$$

per cui

$$\Psi(x) > \log x \quad \text{e} \quad \lim_{x\to\infty}\big[\Psi(x) - \log x\big] = 0. \tag{88}$$

(Traccia: a) si verifichi prima che

$$\int_0^\infty \mathrm{e}^{-(x+m)t}\,dt = \frac{1}{x+m} \quad \text{e} \quad \int_0^\infty \frac{1}{t}\,\big(\mathrm{e}^{-t} - \mathrm{e}^{-nt}\big)\,dt = \log n. \tag{89}$$

Si ha

$$\lim_{n\to\infty}\Big[\log n - \frac{1}{x+1} - \frac{1}{x+2} - \ldots - \frac{1}{x+n+1}\Big]$$
$$= \lim_{n\to\infty}\int_0^\infty \Big[\frac{1}{t}\,\big(\mathrm{e}^{-t} - \mathrm{e}^{-nt}\big) - \mathrm{e}^{-(x+1)t} - \mathrm{e}^{-(x+2)t} - \ldots - \mathrm{e}^{-(x+n+1)t}\Big]\,dt$$
$$= \lim_{n\to\infty}\Big\{\int_0^\infty \Big[\frac{\mathrm{e}^{-t}}{t} - \frac{\mathrm{e}^{-(x+1)t}}{1-\mathrm{e}^{-t}}\Big]\,dt - \int_0^\infty \mathrm{e}^{-nt}\Big[\frac{1}{t} - \frac{\mathrm{e}^{-(x+2)t}}{1-\mathrm{e}^{-t}}\Big]\,dt\Big\},$$

in quanto

$$-\big(1+\mathrm{e}^{-t}+\mathrm{e}^{-2t}+\ldots+\mathrm{e}^{-nt}\big)=\frac{\mathrm{e}^{-(n+1)t}-1}{1-\mathrm{e}^{-t}}\,.$$

Il secondo integrale tende a 0 per $n\to\infty$, il primo non dipende da n. Si sfrutti quindi la formula di Gauss. b) Si sfrutti la formula di Gauss e l'espressione per $\log x$ ottenuta dalla (89).)

4.25 Si dimostri che per n intero positivo e $x>-1$ vale

a) $$\frac{1}{x+1}\;+\;\frac{1}{x+2}\;+\ldots+\;\frac{1}{x+n}\;=\Psi(x+n)-\Psi(x),$$

b) $$\Psi(x)=\Psi(x-1)+\;\frac{1}{x}\,,$$

c) $$\Psi(x+n)-\Psi(x)-\Psi(n)=\gamma+\sum_{k=1}^{n}\Big(\frac{1}{k+x}\;-\;\frac{1}{k}\Big),$$

d) $$\lim_{x\to\infty}\big[\Psi(x+n)-\Psi(x)\big]=0.$$

(Traccia: a) si verifichi che $\sum\dfrac{1}{x+k}\;=\Psi(x+k-1)$, a meno di una costante periodica; c) e d) si sfrutti la a).)

4.26 Si calcoli

$$\sum_{k=1}^{\infty}\frac{1}{(k+a)(k+b)}\,.$$

(Traccia: poiché

$$\frac{1}{(k+a)(k+b)}\;=\;\frac{1}{b-a}\,\Big[\frac{1}{k+a}\;-\;\frac{1}{k+b}\Big],$$

risulta per l'esercizio 4.25

$$\sum_{k=1}^{n}\frac{1}{(k+a)(k+b)}\;=\;\frac{1}{b-a}\,\big[\Psi(n+a)-\Psi(n+b)+\Psi(b)-\Psi(a)\big].$$

Si verifichi che per la (88) è

$$\lim_{n\to\infty}\big[\Psi(n+a)-\Psi(n+b)\big]=0,$$

risulta

$$\sum_{k=1}^{\infty} \frac{1}{(k+a)(k+b)} \;=\; \frac{\Psi(b)-\Psi(a)}{b-a}\,.\;)$$

4.27 Si verifichi che

a) $$\Psi(x) = \sum_{k=1}^{\infty} \frac{x}{k(k+x)} \;-\gamma;$$

b) $$\Psi^{(n)}(x) = (-1)^{n+1} n! \sum_{k=1}^{\infty} \frac{1}{(k+x)^{n+1}}\,;$$

c) $$\Psi^{(n)}(x+m) - \Psi^{(n)}(x) = (-1)^{n} n! \sum_{k=1}^{m} \frac{1}{(k+x)^{n+1}}\,;$$

d) si calcolino $\Psi(0), \Psi(1), \ldots, \Psi(n),\ \Psi'(0), \Psi'(1), \ldots, \Psi'(n)$.

(Traccia: a) si sfrutti l'esercizio 4.26 con $a = 0$ e $b = x$; b) per $n = 1$ si derivi la b) dell'esercizio 4.25 e si sommi, per $n > 1$ si derivi termine a termine la serie precedente; c) si sfrutti la b); d) per l'esercizio 4.25 a) è

$$1 + \;\frac{1}{2}\; + \;\frac{1}{3}\; + \ldots + \;\frac{1}{n} = \Psi(n) + \gamma,$$

e quindi

$$\Psi(0) = -\gamma,\ \Psi(1) = 1 - \gamma,\ \ldots,\ \Psi(n) = 1 + \;\frac{1}{2}\; + \;\frac{1}{3}\; + \ldots + \;\frac{1}{n} - \gamma.$$

Per la b) e la (40) è

$$\Psi'(0) = \sum_{k=1}^{\infty} \frac{1}{k^2} \;=\; \frac{\pi^2}{6}\,,$$

$$\ldots$$

$$\Psi'(n) = \sum_{k=1}^{\infty} \frac{1}{(k+n)^2} \;=\; \sum_{k=n+1}^{\infty} \frac{1}{k^2} \;=\; \frac{\pi^2}{6} - 1 - \;\frac{1}{2^2}\; - \ldots - \;\frac{1}{n^2}\,.$$

4.28 Si calcoli

$$\text{a)}\quad \sum_{k=1}^{\infty} \frac{2k+1}{k(k+1)^2}\,, \qquad \text{b)}\quad \sum_{k=1}^{\infty} \frac{1}{1^2 + 2^2 + \ldots + k^2}\,.$$

(Traccia: a) poiché

$$\frac{2k+1}{k(k+1)^2} = \frac{1}{k(k+1)} + \frac{1}{(k+1)^2},$$

segue dagli esercizi 4.26 e 4.27 che

$$\sum_{k=1}^{\infty} \frac{2k+1}{k(k+1)^2} = \Psi(1) - \Psi(0) + \Psi'(1) = \frac{\pi^2}{6}.$$

b) Poiché

$$\frac{1}{1^2+2^2+\ldots+k^2} = \frac{6}{k(k+1)(2k+1)} = 6\Big[\frac{1}{k(k+1/2)} - \frac{1}{k(k+1)}\Big],$$

segue dall'esercizio 4.26 che

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{1}{1^2+2^2+\ldots+k^2} &= 12\Big(\Psi\big(\frac{1}{2}\big) - \Psi(0)\Big) - 6\big(\Psi(1) - \Psi(0)\big) \\ &= 12\Psi\big(\frac{1}{2}\big) + 12\gamma - 6.\ )\end{aligned}$$

4.29 Si dimostri che

$$\frac{d}{dx}\, B_r(x) = rB_{r-1}(x).$$

(Traccia: si derivi la (34) e si tenga conto del fatto che

$$i\binom{r}{i} = r\binom{r-1}{i-1}.\ )$$

4.30 Si dimostri che

a) $$B_r(x) = \frac{d}{dx} \sum_{k=1}^{r} S_k^{(r)}\, \frac{x^{(k+1)}}{k+1},$$

b) $$B_r = \sum_{k=1}^{r} S_k^{(r)}\, \frac{(-1)^k k!}{k+1}.$$

(Traccia: a) dalla (39) si ha a meno di una costante periodica

$$B_{r+1}(x) = (r+1)\sum x^r,$$

e per la (19) e la linearità dell'operatore $\sum$ è

$$B_{r+1}(x) = (r+1)\sum_{k=1}^{r} S_k^{(r)} \sum x^{(k)}.$$

Si sfruttino poi la (23) e l'esercizio 4.29; b) si verifichi che

$$\frac{d}{dx}\, x^{(k+1)}\Big|_{x=0} = (-1)^k k!\ )$$

4.31 Si dimostri che

a) $\sum_{i=0}^{k-1}(-1)^i \binom{k}{i} B_i = k,$

b) $B_r(x) = (-1)^r B_r(1-x),$

c) $B_r(0) = B_r(1), \quad \text{per } r \neq 1,$

d) $B_{2r+1}(\frac{1}{2}) = 0, \quad \text{per } r \geq 0.$

(Traccia: a) si sfrutti la (33) tenendo conto del fatto che tutti i B_i con indice dispari sono nulli, eccetto $B_1 = -\frac{1}{2}$; b) dalla (34) si ha

$$B_r(1-x) = \sum_{i=0}^{r} \binom{r}{i} (1-x)^i B_{r-i}.$$

Si dimostri, in modo analogo a quanto fatto per dimostrare la (35), che

$$B_r(1-x) = B_r(-x) + r(-x)^{r-1},$$

e si verifichi che

$$B_r(-x) = (-1)^r [B_r(x) + r x^{r-1}],$$

procedendo come in a); c) per r pari, $r \geq 2$, la proprietà deriva da b), per r dispari, $r \geq 3$, la proprietà deriva dal fatto che $B_r(0) = B_r = 0$; d) da b) si ha

$$B_{2r+1}\Big(\frac{1}{2}\Big) = -B_{2r+1}\Big(\frac{1}{2}\Big).\ )$$

4.32 Si dimostri che

$$\int_0^1 B_n(x)dx = \begin{cases} 1 & \text{se } n = 0, \\ 0 & \text{se } n > 0. \end{cases}$$

(Traccia: per $n > 0$ si sfrutti l'esercizio 4.29 e il punto c) dell'esercizio 4.31.)

4.33 Si esprima $B_r(x)$ in serie di Fourier nell'intervallo $0 \le x < 1$.

(Traccia: il polinomio $B_r(x)$ è funzione simmetrica o antisimmetrica rispetto al punto $\frac{1}{2}$ a seconda che r sia pari o dispari (si veda l'esercizio 4.31). Perciò la serie di Fourier di $B_r(x)$ è una serie di soli coseni o di soli seni a seconda che r sia pari o dispari. Per $r = 1$ si ha

$$B_1(x) = \sum_{k=1}^{\infty} b_k \sin 2k\pi x,$$

dove

$$b_k = 2\int_0^1 B_1(x) \sin 2k\pi x \; dx = 2\int_0^1 \Big(x - \frac{1}{2}\Big) \sin 2k\pi x \; dx = -\frac{1}{k\pi}\,,$$

e quindi

$$B_1(x) = -\frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\sin 2k\pi x}{k}\,.$$

Per $r = 2$ si ha

$$B_2(x) = \frac{1}{2}\, a_0 + \sum_{k=1}^{\infty} a_k \cos 2k\pi x,$$

in cui

$$a_0 = 2\int_0^1 B_2(x) \; dx = 0$$

(si veda l'esercizio 4.32). Per ottenere a_k, anziché calcolare

$$2\int_0^1 B_2(x) \cos 2k\pi x \; dx,$$

si può tenere conto che $\dfrac{d}{dx} B_r(x) = r B_{r-1}(x)$ (si veda l'esercizio 4.29), e integrare da 0 a t termine a termine la serie di Fourier di $B_1(x)$, ottenendo

$$B_2(t) - B_2(0) = \frac{1}{\pi^2} \sum_{k=1}^{\infty} \frac{\cos 2k\pi t}{k^2} \; - \frac{1}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2}\,.$$

Poiché

$$B_2(0) - \frac{1}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{2}\, a_0 = 0,$$

risulta

$$B_2(x) = \frac{1}{\pi^2} \sum_{k=1}^{\infty} \frac{\cos 2k\pi x}{k^2} \,.$$

Si ripeta il ragionamento per $r = 3, 4, \dots$)

4.34 a) Si verifichi che

$$\sum_{k=1}^{\infty} \frac{(-1)^k}{k^{2i}} = (-1)^{i+1} \, \frac{(2\pi)^{2i}}{2(2i)!} \, B_{2i}\Big(\frac{1}{2}\Big), \quad i = 1, 2, \dots$$

b) si calcolino

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2} \quad \text{e} \quad \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^4} \,.$$

(Traccia: a) si utilizzi la serie di Fourier di $B_{2i}(x)$ e si ponga $x = \dfrac{1}{2}$;

b) $\displaystyle\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2} = \frac{\pi^2}{12}$, $\displaystyle\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^4} = \frac{7\pi^4}{720}$.)

4.35 a) Si dimostri che

$$\lim_{r\to\infty} |B_{2r}| = \infty,$$

e si determinino r_1 e r_2 tali che

$$\begin{aligned} |B_{2r}| &> 1 \qquad \text{per} \quad r \geq r_1, \\ |B_{2r}| &> 10^6 \qquad \text{per} \quad r \geq r_2; \end{aligned}$$

b) Si determini il raggio di convergenza della serie

$$\frac{x}{\mathrm{e}^x - 1} = \sum_{i=0}^{\infty} B_i \, \frac{x^i}{i!} \,.$$

(Traccia: a) dalla (38) risulta

$$|B_{2r}| = \frac{2\,(2r)!}{(2\pi)^{2r}} \sum_{k=1}^{\infty} \frac{1}{k^{2r}} > \frac{2(2r)!}{(2\pi)^{2r}} \,.$$

È $r_1 = 7$, $r_2 = 13$. b) Posto $y = x^2$, risulta

$$\sum_{i=0}^{\infty} B_i \, \frac{x^i}{i!} = B_1 x + \sum_{i=0}^{\infty} B_{2i} \, \frac{y^i}{(2i)!} \, .$$

Applicando il criterio del rapporto risulta che la serie

$$\sum_{i=0}^{\infty} B_{2i} \, \frac{y^i}{(2i)!}$$

ha raggio r, dove

$$r = \lim_{i\to\infty} \frac{|B_{2i}|(2i+2)!}{(2i)!|B_{2i+2}|} = (2\pi)^2 \lim_{i\to\infty} \left(\sum_{i=1}^{n} \frac{1}{k^{2i}} \right) \Big/ \left(\sum_{i=1}^{n} \frac{1}{k^{2i+2}} \right).$$

Si verifichi che quest'ultimo limite è 1, pertanto $r = (2\pi)^2$ e la serie data converge per $|x| < 2\pi$.)

4.36 Si dimostri che

a) $\dfrac{x}{2} \coth \dfrac{x}{2} = 1 + \dfrac{B_2 x^2}{2!} + \dfrac{B_4 x^4}{4!} + \ldots \quad$ per $|x| < 2\pi$,

b) $\dfrac{x}{2} \tanh \dfrac{x}{2} = (2^2 - 1) \dfrac{B_2 x^2}{2!} + (2^4 - 1) \dfrac{B_4 x^4}{4!} + \ldots \quad$ per $|x| < \pi$,

c) $\dfrac{x}{\sinh x} = 1 - (2^2 - 2) \dfrac{B_2 x^2}{2!} - (2^4 - 2) \dfrac{B_4 x^4}{4!} - \ldots \quad$ per $|x| < \pi$,

da cui si ottiene

$$\coth x = \frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \frac{2x^5}{945} + \ldots \quad \text{per } 0 < |x| < \pi,$$

$$\tanh x = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \ldots \quad \text{per } |x| < \frac{\pi}{2} \, ,$$

$$\operatorname{csch} x = \frac{1}{x} - \frac{x}{6} + \frac{7x^3}{360} - \frac{31x^5}{15120} + \ldots \quad \text{per } 0 < |x| < \pi.$$

(Traccia: a) è

$$\frac{x}{2} \coth \frac{x}{2} = \frac{x}{2} \, \frac{\mathrm{e}^{x/2} + \mathrm{e}^{-x/2}}{\mathrm{e}^{x/2} - \mathrm{e}^{-x/2}} = \frac{x}{2} + \frac{x}{\mathrm{e}^x - 1} \, .$$

Si utilizzi la (31) e si tenga conto del fatto che i B_i con indice dispari sono nulli, eccetto $B_1 = -\dfrac{1}{2}$; per b) e c) si sfruttino le relazioni

$$\frac{x}{2} \tanh \frac{x}{2} = x \coth x - \frac{x}{2} \coth \frac{x}{2} \, , \quad \frac{1}{\sinh x} = \coth x - \tanh \frac{x}{2} \, .$$

Per i raggi di convergenza si veda l'esercizio 4.35.)

4.37 Si dimostri che

$$\text{a)}\quad \cot x = \frac{1}{x} + \sum_{k=1}^{\infty} (-1)^k\, 2^{2k}\, B_{2k}\, \frac{x^{2k-1}}{(2k)!}$$
$$= \frac{1}{x} - \frac{x}{3} - \frac{x^3}{45} - \frac{2x^5}{945} - \ldots \quad \text{per } 0 < |x| < \pi,$$

$$\text{b)}\quad \tan x = \sum_{k=1}^{\infty} (-1)^{k-1}\, 2^{2k}\, (2^{2k}-1)\, B_{2k}\, \frac{x^{2k-1}}{(2k)!}$$
$$= x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \ldots \quad \text{per } |x| < \frac{\pi}{2}\,,$$

$$\text{c)}\quad \csc x = \frac{1}{x} + \sum_{k=1}^{\infty} (-1)^{k-1}\, (2^{2k}-2)\, B_{2k}\, \frac{x^{2k-1}}{(2k)!}$$
$$= \frac{1}{x} + \frac{x}{6} + \frac{7x^3}{360} + \frac{31x^5}{15120} + \ldots \quad \text{per } 0 < |x| < \pi.$$

(Traccia: a) dalla relazione

$$\cot x = \boldsymbol{i} + \frac{1}{x}\, \frac{2\boldsymbol{i}x}{\mathrm{e}^{2\boldsymbol{i}x} - 1}\,,$$

si applichi la (31) con $t = 2\boldsymbol{i}x$; per b) e c) si sfruttino le relazioni

$$\tan x = \cot x - 2\cot 2x, \quad \csc x = \cot x + \tan \frac{x}{2}\,.\ )$$

4.38 Sia $\{\beta_{nj}\}$, per $n, j = 0, 1, \ldots$, una successione doppia, tale che

$$a)\quad \beta_{nj} \geq 0 \quad \text{per ogni} \quad n, j, \quad \text{e} \quad \beta_{nj} = 0 \quad \text{per} \quad j > n;$$

$$b)\quad \lim_{n\to\infty} \beta_{nj} = 0; \qquad c)\quad \sum_{j=0}^{n} \beta_{nj} = 1.$$

Si dimostri che se $\sigma = \lim_{n\to\infty} \sigma_n$, allora la serie

$$\tau_n = \sum_{j=0}^{n} \beta_{nj}\sigma_j$$

è convergente e $\lim_{n\to\infty} \tau_n = \sigma$. Se ne deduca la convergenza della serie ottenuta con la trasformazione di Eulero.

(Traccia: si ha

$$\tau_n - \sigma = \sum_{j=0}^{n} \beta_{nj}(\sigma_j - \sigma).$$

Per ogni $\epsilon > 0$ esiste $\overline{n}$ tale che $|\sigma_j - \sigma| < \dfrac{\epsilon}{2}$ per $j > \overline{n}$ e quindi per ogni $n > \overline{n}$ è

$$|\tau_n - \sigma| \le \Big|\sum_{j=0}^{\overline{n}} \beta_{nj}(\sigma_j - \sigma)\Big| + \sum_{j=\overline{n}+1}^{n} \beta_{nj}|\sigma_j - \sigma|$$

$$< \Big|\sum_{j=0}^{\overline{n}} \beta_{nj}(\sigma_j - \sigma)\Big| + \frac{\epsilon}{2}\,.$$

Si dimostri poi che esiste m tale che

$$\Big|\sum_{j=0}^{\overline{n}} \beta_{nj}(\sigma_j - \sigma)\Big| < \frac{\epsilon}{2} \quad \text{per ogni } n > m.$$

Nel caso della trasformazione di Eulero è

$$\beta_{nj} = \frac{1}{2^n}\binom{n}{j} \quad \text{e si ha} \quad \frac{1}{2^n}\sum_{j=1}^{n}\binom{n}{j} = 1.\ )$$

4.39 Si dimostri che se la serie

$$\sigma_n = \sum_{k=0}^{n-1} (-1)^k a_k$$

è convergente a σ e

$$(-1)^k \Delta^k a_i \ge 0, \quad \text{per ogni } k \text{ e } i \ge 0,$$

(per $k = 0$ e $k = 1$ equivale a dire che $a_i \ge 0$ e $a_i \ge a_{i+1}$ per $i \ge 0$) e

$$\frac{a_{i+1}}{a_i} \ge \rho > \frac{1}{2}\,, \quad \text{per ogni } i \ge 0,$$

allora, indicati con $r_n = \sigma - \sigma_n$ e $r'_n = \sigma - \tau_n$, dove τ_n è la serie ottenuta con la trasformazione di Eulero, vale

$$\left|\frac{r'_n}{r_n}\right| < \frac{1}{(2\rho)^n} < 1,$$

cioè la serie τ_n converge più rapidamente della serie σ_n.

(Traccia: è

$$\begin{aligned} r_n &= \sum_{i=n}^{\infty}(-1)^i a_i = (-1)^n(a_n - a_{n+1} + a_{n+2} - a_{n+3} + \dots) \\ &= (-1)^n(-\Delta a_n - \Delta a_{n+2} - \dots) \end{aligned}$$

Utilizzando la relazione $\Delta^2 a_i \geq 0$, si verifichi che

$$-\Delta a_i \geq -\frac{\Delta a_i + \Delta a_{i+1}}{2} \geq 0,$$

e quindi

$$|r_n| \geq \frac{1}{2}(-\Delta a_n - \Delta a_{n+1} - \Delta a_{n+2} - \dots) = \frac{1}{2}\, a_n.$$

Si verifichi poi che per ogni i la successone di elementi positivi $\{(-1)^k \Delta^k a_i\}_k$ è monotona non crescente, infatti

$$(-1)^k \Delta^k a_i = (-1)^{k-1}\Delta^{k-1} a_i - (-1)^{k-1}\Delta^{k-1} a_{i+1},$$

dove $(-1)^k \Delta^k a_i \geq 0$, $(-1)^{k-1}\Delta^{k-1} a_i \geq 0$ e $(-1)^{k-1}\Delta^{k-1} a_{i+1} \geq 0$ per ipotesi. Inoltre è

$$-\Delta a_0 \leq \frac{1}{2}\, a_0,$$

quindi

$$(-1)^k \Delta^k a_0 \leq \frac{1}{2}\, a_0$$

e

$$|r'_n| = \frac{1}{2}\sum_{k=n}^{\infty}\Big(-\frac{1}{2}\Big)^k \Delta^k a_0 \leq \frac{a_0}{2}\sum_{k=n}^{\infty}\frac{1}{2^{k+1}} = \frac{a_0}{2^{n+1}}\,.$$

Ne segue che

$$\left|\frac{r'_n}{r_n}\right| < \frac{1}{(2\rho)^n} < 1.\ )$$

4.40 Si applichi la trasformazione di Eulero alla serie (2) per il calcolo di $\log 2$.

(Traccia: si ha

$$a_k = \frac{1}{k+1} = k^{(-1)}$$

e dalla (12)

$$\Delta a_k = -k^{(-2)}, \quad \Delta^2 a_k = 2k^{(-3)}, \ldots, \quad \Delta^i a_k = (-1)^i \, i! \, k^{(-i-1)},$$

da cui

$$\Delta^i a_0 = (-1)^i \, \frac{1}{i+1}$$

e

$$\tau_n = \sum_{k=0}^{n-1} \frac{1}{2^{k+1} \, (k+1)} \, . \,)$$

4.41 Si applichi la trasformazione di Eulero alla serie geometrica a segni alterni

$$\sum_{k=0}^{\infty} \Big(- \, \frac{1}{\alpha} \Big)^k, \quad \alpha > 1,$$

e si dica per quali valori di α si ottiene una serie che converge più rapidamente di quella data. Si esamini anche il caso in cui $0 < \alpha \leq 1$ e in particolare il caso $\alpha = \dfrac{1}{2}$.

(Traccia: si ha $a_k = \dfrac{1}{\alpha^k}$ e dalla (6)

$$\Delta^i a_0 = \sum_{j=0}^{i} \binom{i}{j} \, (-1)^{i-j} \, \frac{1}{\alpha^j} \, = (-1)^i \Big(\frac{\alpha - 1}{\alpha} \Big)^i,$$

da cui

$$\tau_n = \frac{1}{2} \, \sum_{k=0}^{n-1} \Big(\frac{\alpha - 1}{2\alpha} \Big)^k.$$

Quindi una serie geometrica a segni alterni di ragione $\dfrac{1}{\alpha}$ viene trasformata in una serie a termini positivi, ancora geometrica di ragione $\dfrac{\alpha - 1}{2\alpha}$. Se $\alpha = 3$ la nuova serie ha la stessa velocità di convergenza, se $\alpha < 3$ la nuova serie converge più velocemente, se $\alpha > 3$ converge più lentamente. Se $0 \leq \alpha < 1$, la serie data non è convergente, mentre la serie trasformata è a segni alterni ed è convergente per $\alpha > \dfrac{1}{3}$.)

4.42 a) Sia $f(x) \in C^{2m+1}[x_0, x_n]$ e si indichi $f_i = f(x_i)$, $x_i = x_0 + ih$, $i = 0, \ldots, n$; si dimostri la seguente formula di *Eulero-Maclaurin*

$$\begin{aligned}\sum_{i=0}^{n-1} f_i &= \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx + \sum_{i=1}^{2m} \frac{B_i}{i!}\, h^{i-1}\left[f^{(i-1)}(x_n) - f^{(i-1)}(x_0)\right] + O(h^{2m}) \\ &= \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx - \frac{1}{2}\left[f(x_n) - f(x_0)\right] + \frac{h}{12}\left[f'(x_n) - f'(x_0)\right] \\ &+ \frac{h^3}{720}\left[f^{(3)}(x_n) - f^{(3)}(x_0)\right] + \frac{h^5}{30240}\left[f^{(5)}(x_n) - f^{(5)}(x_0)\right] \\ &+ \ldots + O(h^{2m})\end{aligned}$$

che, per quanto non sempre convergente per $m \to \infty$, viene spesso usata per stimare le somme quando si è in grado di calcolare l'integrale per altra via.

b) Con la formula di Eulero-Maclaurin si calcoli

$$\sum_{i=0}^{n} i^k, \quad k \le 1 \text{ intero},$$

e nel caso particolare $k = 5$.

(Traccia: a) è

$$\int_{x_0}^{x_n} f(x)\,dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,dx.$$

Con la trasformazione di variabile $x = x_i + th$ si ha

$$\frac{1}{h}\int_{x_i}^{x_{i+1}} f(x)\,dx = \int_0^1 y_i(t)\,dt, \quad y_i(t) = f(x_i + th).$$

Si calcoli

$$\int_0^1 y_i(t)\,dt$$

sfruttando la relazione

$$B_k'(t) = kB_{k-1}(t)$$

(si veda l'esercizio 4.29), con successive integrazioni per parti, nel modo

seguente

$$\begin{aligned}\int_0^1 y_i(t)\ dt &= \int_0^1 y_i(t)B_0(t)\ dt = \Big[y_i(t)B_1(t)\Big]_0^1 - \int_0^1 y_i'(t)B_1(t)\ dt \\ &= \Big[y_i(t)B_1(t)\Big]_0^1 - \frac{1}{2}\ \Big[y_i'(t)B_2(t)\Big]_0^1 + \frac{1}{2}\ \int_0^1 y_i''(t)B_2(t)\ dt \\ &= \Big[y_i(t)B_1(t)\Big]_0^1 - \frac{1}{2}\ \Big[y_i'(t)B_2(t)\Big]_0^1 + \dots \\ &\quad - \frac{1}{(2m)!}\ \Big[y_i^{(2m-1)}(t)B_{2m}(t)\Big]_0^1 + \frac{1}{(2m+1)!}\ \Big[y_i^{(2m)}(t)B_{2m+1}(t)\Big]_0^1 \\ &\quad - \frac{1}{(2m+1)!}\ \int_0^1 y_i^{(2m+1)}(t)B_{2m+1}(t)\ dt.\end{aligned}$$

Si tenga conto del fatto che

$$B_1(0) = -B_1(1) = B_1 = -\frac{1}{2}, \quad B_i(1) = B_i(0) = B_i, \quad \text{per } i > 1$$

$$\text{e} \quad B_i = 0 \quad \text{per} \quad i > 2 \quad \text{dispari;}$$

risulta

$$\begin{aligned}\int_0^1 y_i(t)\ dt &= \frac{1}{2}\ \big[y_i(1) + y_i(0)\big] - \frac{1}{2}\ B_2\big[y_i'(1) - y_i'(0)\big] - \dots \\ &\quad - \frac{B_{2m}}{(2m)!}\ \big[y_i^{(2m-1)}(1) - y_i^{(2m-1)}(0)\big] + e_i \\ &= y_i(0) - \sum_{j=1}^{2m} \frac{B_j}{j!}\ \big[y_i^{(j-1)}(1) - y_i^{(j-1)}(0)\big] + e_i,\end{aligned}$$

dove

$$e_i = -\frac{1}{(2m+1)!}\ \int_0^1 y_i^{(2m+1)}(t)B_{2m+1}(t)\ dt.$$

Poiché

$$y^{(2m+1)} = h^{2m+1} f^{(2m+1)}(x),$$

esiste una costante H tale che $|e_i| \le Hh^{2m+1}$, per cui

$$\sum_{i=0}^{n-1} |e_i| \le nHh^{2m+1} = H(x_n - x_0)h^{2m}.$$

Nell'ipotesi che $\lim_{m\to\infty} e_i = 0$, la formula di Eulero-Maclaurin si potrebbe scrivere

$$\sum_{i=0}^{n-1} f_i = \frac{1}{h}\ \int_{x_0}^{x_n} f(x)\ dx + \sum_{i=1}^{\infty} \frac{B_i}{i!}\ h^{i-1}\big[f^{(i-1)}(x_n) - f^{(i-1)}(x_0)\big].$$

b) Per $2m \geq k+1$ si ha per la (15)

$$\sum_{i=0}^{n} i^k = \int_0^{n+1} x^k \, dx + \sum_{i=1}^{k+1} \frac{B_i}{i!} \, k^{(i-1)} x^{k-i+1} \Big|_0^{n+1}$$
$$= \frac{(n+1)^{k+1}}{k+1} + \sum_{i=1}^{k} \frac{B_i}{i} \binom{k}{i-1} (n+1)^{k-i+1}.$$

Per $k = 5$ è

$$\sum_{i=0}^{n} i^5 = \frac{(n+1)^6}{6} - \frac{(n+1)^5}{2} + \frac{5(n+1)^4}{12} - \frac{(n+1)^2}{12}$$
$$= \frac{(n+1)^2 n^2 (2n^2+2n-1)}{12} \, . \,)$$

4.43 a) Si verifichi, per mezzo della formula di Wallis [1]

$$\lim_{n\to\infty} \frac{2 \cdot 2 \cdot 4 \cdot 4 \cdots 2n \cdot 2n}{1 \cdot 3 \cdot 3 \cdot 5 \cdots (2n-1) \cdot (2n+1)} = \frac{\pi}{2} \, ,$$

che

$$\lim_{n\to\infty} [4n \log 2 + 4 \log n! - \log(2n+1) - 2 \log(2n)!] = \log \frac{\pi}{2} \, .$$

b) Si scriva $\log n! = \sum_{i=1}^{n} \log i$ applicando la formula di Eulero-Maclaurin alla funzione $f(x) = \log(x+1)$,

c) si dimostri la formula di *Stirling* per l'approssimazione di $n!$ per n grande

$$n! \sim \sqrt{2\pi} \; n^{n+\frac{1}{2}} \; \mathrm{e}^{-n};$$

d) si verifichi l'accuratezza della formula di Stirling per $n = 10$ e $n = 20$.
(Traccia: a) si passi ai logaritmi; b) posto $x_0 = 0$ e $h = 1$, dalla formula di Eulero-Maclaurin per $n-2$ si ha

$$\sum_{i=0}^{n-2} \log(i+1) = \int_0^{n-1} \log(x+1) \, dx + B_1 \log n$$
$$+ \sum_{i=2}^{2m} \frac{B_i}{i!} \, (-1)^i (i-2)! \, \Big[\frac{1}{n^{i-1}} - 1\Big] + \dots$$

Tenendo conto che $B_i = 0$ per i dispari, $i > 2$, si ha

$$\sum_{i=1}^{n-1} \log i - n \log n + n - 1 + \frac{1}{2} \log n = \sum_{i=2}^{2m} \frac{B_i}{i(i-1)} \left[\frac{1}{n^{i-1}} - 1 \right] + \dots \quad (90)$$

I coefficienti $\dfrac{B_i}{i(i-1)}$ per i pari hanno segno alterno e decrescono in modulo fino a $i = 8$ (è $\left|\dfrac{B_8}{7 \cdot 8}\right| < 0.0006$), poi diventano crescenti e tendono in modulo ad ∞ (si veda l'esercizio 4.35). Quindi la serie

$$\sum_{i=2}^{\infty} \frac{B_i}{i(i-1)} \left[\frac{1}{n^{i-1}} - 1 \right]$$

è non convergente, ma è possibile verificare che il primo membro della (90) è compreso fra due somme parziali pari consecutive e che l'errore che si commette approssimando il primo membro con una somma parziale pari è limitato dal modulo del primo termine di indice pari trascurato. Si pone allora, sommando e sottraendo $\log n$,

$$\log n! \sim c + \left(n + \frac{1}{2}\right) \log n - n + \sum_{i=2}^{2m} \frac{B_i}{i(i-1)n^{i-1}} \quad (91)$$

(è opportuno scegliere $m \leq 4$), in cui c è una costante che non dipende da n. Per determinare c si scriva la (91) nella forma

$$2 \log n! \sim 2c + (2n+1) \log n - 2n + 2s_n,$$

da cui

$$2 \log(2n)! \sim 2c + (4n+1) \log(2n) - 4n + 2s_{2n}.$$

Sostituendo nella relazione in a) e notando che

$$\lim_{n \to \infty} s_n = 0 \quad \text{e} \quad \lim_{n \to \infty} [\log n - \log(2n+1)] = -\log 2,$$

risulta

$$2c - 2 \log 2 = \log \frac{\pi}{2},$$

da cui segue che

$$c = \log \sqrt{2\pi}.$$

Si ha perciò

$$\log n! \sim \log \left[\sqrt{2\pi}\, n^{n+\frac{1}{2}}\, \mathrm{e}^{-n} \right] + \sum_{i=2}^{2m} \frac{B_i}{i(i-1)n^{i-1}}.$$

c) La formula di Stirling viene ottenuta trascurando la sommatoria e commettendo su $\log n!$ un errore minore di $\dfrac{B_2}{2n} = \dfrac{1}{12n}$.

d) Si ha

$$10! = 36288\ 10^2,$$
$$20! = 243290200817664\ 10^4.$$

Per l'implementazione su calcolatore si consiglia di trasformare la formula di Stirling nel modo seguente

$$n! \sim \sqrt{2n\pi}\ \left(\frac{n}{\mathrm{e}}\right)^n,$$

perché in tal modo si riduce il rischio che si verifichi un errore di overflow per n grande; risulta

$$10! \sim \sqrt{20\pi}\ \left(\frac{10}{\mathrm{e}}\right)^{10} = 3.598696\ 10^6,$$
$$20! \sim \sqrt{40\pi}\ \left(\frac{20}{\mathrm{e}}\right)^{20} = 2.422787\ 10^{18}.\ )$$

4.44 I problemi di accrescimento sono tipici del mondo naturale. Un esempio classico è il seguente: da un seme piantato nasce un fiore che produce un seme alla fine del primo anno e un altro seme alla fine del secondo anno. Supponendo che ogni seme ottenuto venga immediatamente ripiantato, si determini il numero di semi ottenuti dopo k anni a partire da un solo seme.

(Traccia: indicato con y_k il numero di semi ottenuti dopo k anni, risulta $y_k = y_{k-1} + y_{k-2}$, con $y_1 = 1$ e $y_2 = 1$.)

4.45 Si determini in quante parti viene diviso il piano se su di esso si tracciano k rette in modo che non ve ne siano due parallele né tre passanti per lo stesso punto (Steiner, 1826).

(Traccia: indicato con y_k il numero di parti del piano individuate da k rette, la $k+1$-esima retta interseca le altre k rette, ciascuna in un solo punto, dividendo $k+1$ parti esistenti del piano in due, e aggiungendo quindi $k+1$ parti. Risulta perciò

$$y_{k+1} = y_k + k + 1, \quad y_1 = 2.\ )$$

4.46 Ad ogni passo di un gioco un giocatore scommette 1 lira contro un avversario ed ha probabilità p di vincere. Inizialmente il giocatore ha N lire e l'avversario ha M lire, e il gioco termina quando il giocatore o l'avversario

perde totalmente la somma posseduta. Si determini quale probabilità ha il giocatore di vincere.

(Traccia: indicata con y_k la probabilità del giocatore di vincere quando ha k lire, è

$$y_k = py_{k+1} + (1-p)y_{k-1}, \quad y_0 = 0,\ y_{N+M} = 1,$$

infatti il giocatore può vincere se, avendo k lire, vince (con probabilità p) al passo corrente e poi vince il gioco (con probabilità y_{k+1}, avendo $k+1$ lire), oppure perde (con probabilità $1-p$) al passo corrente e poi vince (con probabilità y_{k-1}, avendo $k-1$ lire) il gioco.

4.47 Si risolva l'equazione alle differenze lineare del primo ordine

$$\Delta y_k = b(k), \quad k \geq m,$$

assegnato un valore iniziale y_m, $m \geq 0$.

(Traccia: per la (24), oppure direttamente, è

$$y_n = y_m + \sum_{k=m}^{n-1} b(k), \quad \text{per} \quad n > m. \ )$$

4.48 Siano $z_k^{(j)}$ per $j = 1, \ldots, n$, soluzioni della (49), e per ogni $k \geq k_0$ si consideri la matrice (detta *di Casorati*) di ordine n

$$C(k) = \begin{bmatrix} z_k^{(1)} & z_k^{(2)} & \cdots & z_k^{(n)} \\ z_{k+1}^{(1)} & z_{k+1}^{(2)} & \cdots & z_{k+1}^{(n)} \\ \vdots & \vdots & & \vdots \\ z_{k+n-1}^{(1)} & z_{k+n-1}^{(2)} & \cdots & z_{k+n-1}^{(n)} \end{bmatrix}.$$

Si dimostri che se $a_0(k)a_n(k) \neq 0$ per ogni $k \geq k_0$, allora

$$\det C(k+1) = (-1)^n \, \frac{a_0(k)}{a_n(k)} \, \det C(k),$$

e quindi se, come nel caso del teorema 4.13, è $C(k_0) = I$, allora $\det C(k) \neq 0$ per ogni $k \geq k_0$.

(Traccia: l'ultima riga di $C(k+1)$ è data da

$$[z_{k+n}^{(1)},\ z_{k+n}^{(2)},\ \ldots,\ z_{k+n}^{(n)}]$$

e per la (49) risulta uguale a

$$-\frac{1}{a_n(k)}\Big[\sum_{j=0}^{n-1} a_j(k)z_{k+j}^{(1)},\ \sum_{j=0}^{n-1} a_j(k)z_{k+j}^{(2)},\ \ldots,\ \sum_{j=0}^{n-1} a_j(k)z_{k+j}^{(n)}\Big].$$

Si applichino i teoremi sui determinanti e si verifichi che

$$C(k+1) = -\frac{a_0(k)}{a_n(k)}\begin{bmatrix} z_{k+1}^{(1)} & z_{k+1}^{(2)} & \cdots & z_{k+1}^{(n)} \\ \vdots & \vdots & & \vdots \\ z_{k+n-1}^{(1)} & z_{k+n-1}^{(2)} & \cdots & z_{k+n-1}^{(n)} \\ z_k^{(1)} & z_k^{(2)} & \cdots & z_k^{(n)} \end{bmatrix}. \ )$$

4.49 Siano $z_k^{(j)}$ per $j = 1, \ldots, n$, soluzioni della equazione omogenea (49). Si dimostri che se

$$\lim_{k\to\infty} \frac{z_k^{(j+1)}}{z_k^{(j)}} = 0, \quad \text{per} \quad j = 1, \ldots, n-1, \tag{92}$$

allora le soluzioni sono linearmente indipendenti.

(Traccia: si proceda per induzione su n. Per $n = 2$, si verifichi che se vale la (92), non è possibile che esista $\alpha \neq 0$ tale che $z_k^{(2)} = \alpha z_k^{(1)}$. Per $n > 2$, si verifichi che se $z_k^{(1)}, \ldots, z_k^{(n-1)}$ sono linearmente indipendenti, e se vale la (92), non è possibile che

$$z_k^{(n)} = \alpha_1 z_k^{(1)} + \ldots + \alpha_{n-1} z_k^{(n-1)},$$

per qualche $\alpha_j \neq 0$, $j = 1, \ldots, n-1$. Risulterebbe infatti

$$\lim_{k\to\infty} \frac{z_k^{(n)}}{z_k^{(i)}} = \alpha_i, \quad \text{dove} \quad i = \min\{j \ :\ \alpha_j \neq 0\},$$

mentre per la (92) tale limite dovrebbe essere nullo.)

4.50 Si trasformi l'equazione alle differenze lineare a coefficienti non costanti

$$y_{k+1} - a(k)y_k = b(k), \quad k \geq 0, \quad a(k) \neq 0 \text{ per ogni } k \geq 0,$$

in una equazione lineare a coefficienti costanti e si risolva. Si applichi al caso particolare delle seguenti equazioni:

a) $(k+1)y_{k+1} - ky_k = 1;$

b) $y_{k+1} - a^k y_k = 0;$

c) $(k+2)y_{k+1} - ky_k = k.$

(Traccia: si ponga

$$d(k) = \prod_{i=0}^{k-1} a(i), \quad k \geq 1,$$

dividendo l'equazione data per $d(k+1)$, si ha

$$\frac{y_{k+1}}{d(k+1)} - \frac{y_k}{d(k)} = \frac{b(k)}{d(k+1)},$$

e quindi si ottiene l'equazione lineare

$$z_{k+1} - z_k = c(k), \quad \text{dove} \quad z_k = \frac{y_k}{d(k)}, \quad c(k) = \frac{b(k)}{d(k+1)}.$$

Si risolva come nell'esercizio 4.47 con $m = 1$. Casi particolari:

a) si pone $z_k = ky_k$, risulta $y_k = \dfrac{\alpha_1 + k}{k}$;

b) si pone $z_k = \dfrac{y_k}{a^{k(k-1)/2}}$, risulta $y_k = \alpha_1 a^{k(k-1)/2}$;

c) si pone $z_k = \dfrac{k(k+1)y_k}{2}$, risulta $y_k = \dfrac{\alpha_1}{k(k+1)} + \dfrac{k-1}{3}$,

dove α_1 è un parametro.)

4.51 Si trasformi l'equazione alle differenze non lineare

$$y_{k+1} - y_k + a(k)y_k y_{k+1} = 0, \quad k \geq 0,$$

in una equazione lineare e si risolva. Si consideri il caso particolare dell'equazione

$$y_{k+1} - y_k + ky_k y_{k+1} = 0.$$

(Traccia: si divida per $y_k y_{k+1}$ e si ponga $z_k = \dfrac{1}{y_k}$. Si ottiene l'equazione $z_{k+1} - z_k = a(k)$. Si risolva come nell'esercizio 4.47. Nel caso particolare risulta

$$z_k = \frac{\alpha + k(k-1)}{2}, \quad \text{da cui} \quad y_k = \frac{2}{\alpha + k(k-1)},$$

dove α è un parametro.)

4.52 Si trasformi l'equazione alle differenze non lineare

$$y_{k+1}^2 = a(k)y_k, \quad a(k) > 0 \quad \text{per} \quad k \geq 0,$$

in una equazione lineare e si risolva. Si consideri il caso particolare dell'equazione

$$y_{k+1}^2 = \frac{k(k+1)^2}{2(k-1)}\, y_k.$$

(Traccia: si passi ai logaritmi e si ponga $z_k = \log y_k$. Si ottiene l'equazione

$$2z_{k+1} - z_k = b(k), \quad \text{dove} \quad b(k) = \log a(k),$$

da cui

$$z_k = \left(\frac{1}{2}\right)^k \alpha_1 + \sum_{i=0}^{k-1} \left(\frac{1}{2}\right)^{k-i} b(i),$$

e quindi

$$y_k = \alpha_2^{1/2^k} \prod_{i=0}^{k-1} a(i)^{1/2^{k-i}},$$

in cui α_1 e α_2 sono due parametri. Nel caso particolare risulta

$$y_k = \alpha_2^{1/2^k}\ \frac{k(k-1)}{2}\,.\ )$$

4.53 Si trasformi l'equazione alle differenze non lineare

$$y_k y_{k+1} y_{k+2} = y_k + y_{k+1} + y_{k+2}, \quad k \geq 0,$$

in una equazione lineare e si risolva.

(Traccia: si ponga $y_k = \tan z_k$. Tenendo conto che fra gli angoli α, β e γ di un triangolo vale la relazione

$$\tan\alpha + \tan\beta + \tan\gamma = \tan\alpha \tan\beta \tan\gamma,$$

si ottiene l'equazione lineare

$$z_k + z_{k+1} + z_{k+2} = 2\pi,$$

da cui

$$z_k = \frac{2\pi}{3}\ + \alpha_1 \cos\frac{2k\pi}{3}\ + \alpha_2 \sin\frac{2k\pi}{3}\,,$$

dove α_1 e α_2 sono due parametri.)

4.54 Si consideri la matrice tridiagonale

$$A_k = \begin{bmatrix} \alpha_1 & \gamma_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix}.$$

Nell'ipotesi che esista la fattorizzazione LU di A_k, si ha

$$A_k = LU = \begin{bmatrix} 1 & & & & \\ b_1 & 1 & & & \\ & b_2 & \ddots & & \\ & & \ddots & \ddots & \\ & & & b_{k-1} & 1 \end{bmatrix} \begin{bmatrix} a_1 & c_1 & & & \\ & a_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & c_{k-1} \\ & & & & a_k \end{bmatrix}.$$

Per determinare a_i, $i = 1, \ldots, k$ e b_i, c_i, $i = 1, \ldots, k-1$, si uguagliano gli elementi di A_k e della matrice prodotto LU e si ottengono le relazioni

$$\begin{aligned} &c_i = \gamma_i, \quad i = 1, \ldots, k-1, \qquad a_1 = \alpha_1, \\ &b_i = \frac{\beta_i}{a_i}, \quad i = 1, \ldots, k-1, \\ &a_i = \alpha_i - b_{i-1}c_{i-1}, \quad i = 2, \ldots, k. \end{aligned}$$

Quindi gli a_i soddisfano l'equazione alle differenze non lineare

$$a_i = \alpha_i - \frac{\beta_{i-1}\gamma_{i-1}}{a_{i-1}},$$

con la condizione iniziale $a_1 = \alpha_1$. Si trasformi questa equazione in una lineare.

(Traccia: si ponga

$$z_i = \prod_{j=1}^{i} a_j,$$

risulta

$$z_i = \alpha_i z_{i-1} - \beta_{i-1}\gamma_{i-1}z_{i-2},$$

con le condizioni iniziali

$$z_1 = \alpha_1, \quad z_2 = \alpha_1\alpha_2 - \beta_1\gamma_1. \ )$$

4.55 a) Siano $p(k)$ un polinomio di grado al più n e $\alpha \neq 0, 1$. Si verifichi

che per ogni s è

$$\Delta^s\big[p(k)\alpha^k\big] = q_s(k)\alpha^k,$$

dove i $q_s(k)$ sono polinomi di grado al più n, identicamente nulli se e solo se $p(k)$ è identicamente nullo.

b) Dati r numeri $x_1, \dots, x_r$ distinti ($r \geq 1$, se $r = 1$ sia $x_1 \neq 0$), si dimostri che le successioni $\{z_k^{(i,s)}\}_{k\in N}$ definite da

$$z_k^{(i,s)} = k^s x_i^k, \quad i = 1, 2, \dots, r, \ s = 0, 1, \dots, t,$$

sono linearmente indipendenti per ogni t intero, $t \geq 0$.

(Traccia: a) si dimostri per induzione su s. Per $s = 1$ è

$$\Delta\big[p(k)\alpha^k\big] = p(k+1)\alpha^{k+1} - p(k)\alpha^k = \big[p(k+1)\alpha - p(k)\big]\alpha^k = q_1(k)\alpha^k.$$

Per $s > 1$ si proceda in modo analogo.

b) Indicata con

$$c(k) = \sum_{i=1}^{r}\sum_{s=0}^{t} \alpha_{i,s} z_k^{(i,s)} = \sum_{i=1}^{r} x_i^k \sum_{s=0}^{t} \alpha_{i,s} k^s$$

una qualunque combinazione lineare di $z_k^{(i,s)}$ e con

$$p_i(k) = \sum_{s=0}^{t} \alpha_{i,s} k^s,$$

risulta

$$c(k) = \sum_{i=1}^{r} p_i(k)\ x_i^k,$$

in cui $p_i(k)$ è un polinomio in k di grado al più t. Si dimostri per induzione su r che se la successione $c(k)$ è nulla per ogni k, allora tutti i polinomi $p_i(k)$ sono identicamente nulli. La tesi è ovvia per $r = 1$. Si supponga che la tesi sia vera fino all'indice $r - 1$, si ha per l'indice r

$$c(k) = \sum_{i=1}^{r-1} p_i(k) x_i^k + p_r(k) x_r^k.$$

Se $x_r = 0$, la tesi segue direttamente dall'ipotesi induttiva. Se $x_r \neq 0$, si ha

$$c(k) = x_r^k \left[\sum_{i=1}^{r-1} p_i(k)\big(\frac{x_i}{x_r}\big)^k + p_r(k)\right].$$

Se $c(k)$ è identicamente nulla, allora è

$$d(k) = \sum_{i=1}^{r-1} p_i(k)\Big(\frac{x_i}{x_r}\Big)^k + p_r(k) = 0, \tag{93}$$

e quindi anche le differenze finite di $d(k)$ di qualunque ordine sono nulle. In particolare è

$$\Delta^{t+1} d(k) = 0.$$

Poiché il grado di $p_r(k)$ è minore o uguale a t, si ha per la (21)

$$\Delta^{t+1} p_r(k) = 0$$

e per quanto dimostrato al punto a) risulta

$$\Delta^{t+1} d(k) = \sum_{i=1}^{r-1} \Delta^{t+1}\left[p_i(k)\Big(\frac{x_i}{x_r}\Big)^k\right] = \sum_{i=1}^{r-1} \Big(\frac{x_i}{x_r}\Big)^k w_i(k) = 0,$$

per $i = 1, \ldots, r-1$, dove il polinomio $w_i(k)$ ha lo stesso grado di $p_i(k)$ ed è identicamente nullo se e solo se lo è $p_i(k)$. Per l'ipotesi induttiva, ne segue che $w_i(k) = 0$ per $i = 1, \ldots, r-1$, e quindi $p_i(k) = 0$, per $i = 1, \ldots, r-1$. Dalla (93) segue infine che $p_r(k) = 0$.)

4.56 Si determini la soluzione generale delle seguenti equazioni alle differenze omogenee

a) $\quad 4y_{k+3} - 15y_{k+1} + 2y_k = 0,$

b) $\quad y_{k+3} - 4y_{k+2} + 5y_{k+1} - 2y_k = 0,$

c) $\quad y_{k+4} + 4y_k = 0,$

d) $\quad y_{k+4} + 4y_{k+2} + 4y_k = 0.$

(Risposta:

a) $\quad y_k = \alpha_1(-2)^k + \alpha_2\Big(\frac{2-\sqrt{3}}{2}\Big)^k + \alpha_3\Big(\frac{2+\sqrt{3}}{2}\Big)^k,$

b) $\quad y_k = \alpha_1 + \alpha_2 k + \alpha_3 2^k,$

c) $\quad y_k = \big(\sqrt{2}\big)^k\Big[\alpha_1 \cos\frac{k\pi}{4} + \alpha_2 \sin\frac{k\pi}{4} + \alpha_3 \cos\frac{3k\pi}{4} + \alpha_4 \sin\frac{3k\pi}{4}\Big],$

d) $\quad y_k = \big(\sqrt{2}\big)^k\Big[(\alpha_1 + \alpha_2 k)\cos\frac{k\pi}{2} + (\alpha_3 + \alpha_4 k)\sin\frac{k\pi}{2}\Big].$

4.57 Si calcoli la soluzione dell'equazione alle differenze

$$y_{k+3} = 3y_{k+1} - 2y_k$$

che soddisfa le condizioni iniziali $y_0 = 0$, $y_1 = 1$, $y_2 = 2$.

(Risposta: $y_k = k$.)

4.58 Si determini la soluzione generale delle seguenti equazioni alle differenze complete

$$y_{k+2} - 2y_{k+1} + y_k = b(k),$$

dove

a) $b(k) = k + 1$; b) $b(k) = k(k + 2)$; c) $b(k) = \sin k$; d) $b(k) = a^k$;
e) $b(k) = ka^k$.

(Risposta: $y_k = \alpha_1 + \alpha_2 k + z_k$, dove

$$\text{a) } z_k = \frac{k^3}{6}\,; \quad \text{b) } z_k = \frac{k^2}{12}\,(k^2 - 7); \quad \text{c) } z_k = \frac{\sin(k-1)}{2(\cos 1 - 1)}\,;$$

$$\text{d) se } a \neq 1,\ \text{è } z_k = \frac{a^k}{(a-1)^2}\,,\ \text{se } a = 1,\ \text{è } z_k = \frac{k^2}{2}\,;$$

$$\text{e) se } a \neq 1,\ \text{è } z_k = \frac{a^k}{(a-1)^2}\,\Big[k - \frac{2a}{a-1}\Big],\ \text{se } a = 1,\ \text{è } z_k = \frac{k^2}{6}\,(k-3).\ )$$

4.59 Sia y_k una soluzione dell'equazione alle differenze

$$y_{k+2} + ay_{k+1} + by_k = c,$$

dove a, b e c sono costanti. Si dimostri che y_k è limitata qualunque siano le condizioni iniziali se e solo se

$$-1 \leq b < 1 \quad \text{e} \quad -1 - b \leq a \leq 1 + b,$$

oppure

$$b = 1 \quad \text{e} \quad -2 < a < 2,$$

e si dica quando

$$\lim_{k\to\infty} y_k = \frac{c}{1 + a + b}\,.$$

(Traccia: si verifichi che la soluzione dell'equazione omogenea associata è limitata se e solo se le soluzioni dell'equazione $x^2 + ax + b = 0$ hanno modulo minore di 1, oppure uguale a 1 ma sono distinte, e la soluzione dell'equazione completa è limitata. Se i moduli sono minori di 1, allora y_k tende ad una soluzione particolare dell'equazione completa.)

4.60 Si risolva l'equazione alle differenze lineare del primo ordine

$$y_{k+1} + \alpha y_k = b(k),$$

in cui α è costante e si verifichi che se $|\alpha| < 1$ e $\lim_{k\to\infty} b(k) = \beta$, allora

$$\lim_{k\to\infty} y_k = \frac{\beta}{1+\alpha}\,,$$

qualunque sia la condizione iniziale.

(Traccia: fissata una condizione iniziale y_m, $m \geq 0$, risulta

$$y_k = (-\alpha)^{k-m} y_m + \sum_{i=m}^{k-1} (-\alpha)^{k-i-1} b(i);$$

inoltre, poiché

$$\sum_{i=m}^{k-1} (-\alpha)^{k-i-1}(1+\alpha) = 1 - (-\alpha)^{k-m},$$

si ha

$$y_k - \frac{\beta}{1+\alpha} = (-\alpha)^{k-m}\big(y_m - \frac{\beta}{1+\alpha}\big) + \sum_{i=m}^{k-1} (-\alpha)^{k-i-1}(b(i)-\beta).$$

Si verifichi che per ogni ϵ e per k sufficientemente elevato è possibile determinare m tale che

$$\big|y_k - \frac{\beta}{1+\alpha}\big| < \epsilon.\ )$$

4.61 Si calcoli la soluzione generale dell'equazione

$$\Delta^n y_k = 0, \quad n \geq 1.$$

(Traccia: per la (6) l'equazione caratteristica è

$$\sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} x^j = (x-1)^n = 0.$$

La soluzione generale è quindi

$$y_k = \sum_{i=0}^{n-1} \alpha_i k^i,$$

dove $\alpha_0, \ldots, \alpha_{n-1}$ sono n parametri.)

4.62 Si verifichi che il numero

$$y_k = \frac{\sqrt{3}}{6}\left[(1+\sqrt{3})^{k+1} - (1-\sqrt{3})^{k+1}\right]$$

è intero per ogni $k \geq 0$.

(Traccia: y_k è della forma $y_k = \alpha_1 x_1^k + \alpha_2 x_2^k$ e quindi è soluzione dell'equazione alle differenze lineare omogenea del secondo ordine

$$y_{k+2} - (x_1 + x_2)y_{k+1} + x_1 x_2 y_k = 0.$$

Si verifichi che i coefficienti di tale equazione e le condizioni iniziali sono numeri interi.)

4.63 Si determini la soluzione dell'equazione alle differenze

$$2y_{k+2} - 5y_{k+1} + 2y_k = 2^{-k},$$

tale che $y_0 = 1$ e $\lim_{k\to\infty} y_k = 0$.

(Traccia: la soluzione generale è

$$y_k = \alpha_1 2^k + \alpha_2\left(\frac{1}{2}\right)^k - \frac{2}{3}\,\frac{k}{2^k}\,.$$

Imponendo le condizioni si ha $\alpha_1 = 0$, $\alpha_2 = 1$.)

4.64 a) Siano $y_k,\ k = 0, 1, \ldots$, i numeri di Fibonacci determinati nell'esempio 4.16. Si verifichi che

$$s = \lim_{k\to\infty} \frac{y_{k+1}}{y_k} = \frac{1+\sqrt{5}}{2}\,.$$

Il numero s era chiamato dai greci *sezione aurea* e rivestiva una notevole importanza in campo artistico. Si verifichi che

b) $$\sum_{k=1}^{n} y_k = y_{n+2} - 1;$$

c) i numeri di Fibonacci sono anche soluzione dell'equazione non lineare

$$y_k^2 + y_{k+1}^2 = y_{2k+1};$$

d) le soluzioni z_k dell'equazione

$$z_{k+1} = \beta z_k z_{k-1}, \quad \beta > 0,$$

sono tali che il

$$\lim_{k\to\infty} \frac{z_{k+1}}{z_k^s}$$

dove s è definito al punto a), è finito e non nullo.

(Traccia: a) segue dalla (61); b) si sfrutti l'equazione (60) e il fatto che $y_1 = 1$; c) si verifichi per induzione su j, sfruttando la (60), che

$$y_{2k+1} = y_{2k+1-j}\ y_{j+1} + y_{2k-j}\ y_j;$$

d) si ponga $v_k = \log z_k$, risulta

$$v_{k+1} - v_k - v_{k-1} = \log\beta,$$

da cui

$$v_k = \alpha_1\Big(\frac{1-\sqrt{5}}{2}\Big)^k + \alpha_2\Big(\frac{1+\sqrt{5}}{2}\Big)^k - \log\beta$$

e

$$\frac{z_{k+1}}{z_k^s} = \exp\Big[-\alpha_1\sqrt{5}\Big(\frac{1-\sqrt{5}}{2}\Big)^k + \frac{\sqrt{5}-1}{2}\ \log\beta\Big],$$

da cui

$$\lim_{k\to\infty} \frac{z_{k+1}}{z_k^s} = \beta^{(\sqrt{5}-1)/2}.\ )$$

4.65 Indicato con $e_k = |x_k - \alpha|$ l'errore alla k-esima iterazione della variante (41, cap. 3) del metodo delle secanti, dalla (43, cap. 3) segue che

$$e_{k+1} = e_k e_{k-1}\ \delta_k, \quad \text{dove} \quad \delta_k = \Big|\frac{f''(\xi_k)}{2f'(\eta_k)}\Big|,$$

in cui ξ_k e η_k appartengono ad un intorno di α di raggio tendente a zero per $k \to \infty$ ed è

$$\lim_{k\to\infty} \delta_k = \delta = \Big|\frac{f''(\alpha)}{2f'(\alpha)}\Big|,$$

in quanto la successione e_k converge a zero. Si dimostri che se $e_k \neq 0$ per ogni k, allora

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^s}, \quad \text{dove} \quad s = \frac{1+\sqrt{5}}{2},$$

è finito e diverso da zero.

(Traccia: si ponga $v_k = \log e_k$, risulta

$$v_{k+1} - v_k - v_{k-1} = \log \delta_k.$$

Si verifichi che tale equazione è equivalente all'equazione

$$z_{k+1} + (s-1)z_k = \log \delta_k, \quad \text{dove} \quad z_k = v_k - sv_{k-1},$$

la cui soluzione, per quanto visto nell'esercizio 4.60, è tale che

$$\lim_{k\to\infty} z_k = \frac{\log \delta}{s},$$

per ogni condizione iniziale. Ne segue che

$$\log \sqrt[s]{\delta} = \lim_{k\to\infty} v_k - sv_{k-1} = \lim_{k\to\infty} \log \frac{e_k}{e_{k-1}^s}. \ )$$

4.66 Si determinino i coefficienti dello sviluppo in serie di potenze

$$\frac{\beta x - 1}{x^3 - x^2 - 4x + 4} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots$$

nei due casi $\beta = \frac{1}{2}$, $\beta = 1$.

(Traccia: moltiplicando per il denominatore e uguagliando i termini dello stesso grado in x si ottengono le relazioni

$$\begin{aligned} &4a_0 = -1, \\ &4a_1 - 4a_0 = \beta, \\ &4a_2 - 4a_1 - a_0 = 0, \\ &4a_{k+3} - 4a_{k+2} - a_{k+1} + a_k = 0, \quad k = 0, 1, \dots \end{aligned} \tag{94}$$

La (94) è un'equazione alle differenze del terzo ordine a cui devono soddisfare i coefficienti cercati, mentre le prime tre relazioni forniscono le condizioni iniziali. La soluzione generale della (94) è

$$a_k = \alpha_1 + \alpha_2 \big(\frac{1}{2}\big)^k + \alpha_3 \big(-\frac{1}{2}\big)^k.$$

Imponendo le condizioni iniziali si ottiene il sistema

$$\begin{cases} \alpha_1 + \alpha_2 + \alpha_3 = -\frac{1}{4}, \\ \alpha_2 + 3\alpha_3 = -\frac{\beta}{2}, \\ \alpha_1 + 2\alpha_2 - 2\alpha_3 = 0. \end{cases}$$

Le soluzioni particolari cercate sono

$$\text{per} \quad \beta = \frac{1}{2}\,, \quad a_k = \frac{1}{6}\Big[\Big(-\frac{1}{2}\Big)^{k+1} - 1\Big],$$

$$\text{per} \quad \beta = 1, \quad a_k = \begin{cases} 0 & \text{per } k \text{ dispari,} \\ -\dfrac{1}{2^{k+2}} & \text{per } k \text{ pari.)} \end{cases}$$

4.67 Si risolva il sistema lineare $A\mathbf{x} = \mathbf{b}$, di ordine 100, in cui gli elementi della matrice tridiagonale simmetrica A e del vettore $\mathbf{b}$ sono

$$a_{ij} = \begin{cases} 2 & \text{se } i = j, \\ -1 & \text{se } |i-j| = 1, \\ 0 & \text{altrimenti,} \end{cases} \qquad b_i = i.$$

(Traccia: le componenti del vettore $\mathbf{x}$ soddisfano l'equazione alle differenze

$$-x_{k+1} + 2x_k - x_{k-1} = k, \quad 1 \le k \le 100,$$

con le condizioni al contorno $x_0 = 0$ e $x_{101} = 0$. La soluzione generale è $x_k = \alpha_1 + \alpha_2 k - \dfrac{k^3}{6}$, la soluzione particolare è $x_k = \dfrac{k}{6}\left(101^2 - k^2\right)$.)

4.68 Per $x \in \mathbf{R}$, $|x| < 1$, si consideri l'equazione alle differenze

$$y_{k+2} - 2xy_{k+1} + y_k = 0. \tag{95}$$

a) Posto $x = \cos\theta$, si determinino le soluzioni particolari
   (1) $T_k(x)$ che soddisfano le condizioni iniziali $T_0(x) = 1$, $T_1(x) = x$,
   (2) $U_k(x)$ che soddisfano le condizioni iniziali $U_{-1}(x) = 0$, $U_0(x) = 1$.

Le soluzioni $T_k(x)$ e $U_k(x)$ sono polinomi in x e sono detti *polinomi di Chebyshev* di 1^a e 2^a specie (si veda il capitolo 6).

b) Si dimostri che ogni soluzione dell'equazione (95) può essere espressa come
$$y_k = \alpha_1 T_k(x) + \alpha_2 U_{k-1}(x), \quad \alpha_1,\ \alpha_2 \in \mathbf{R}.$$

c) Si determinino le soluzioni particolari dell'equazione (95) che soddisfano le seguenti condizioni al contorno

(1) $y_0 = y_N = 0$,
(2) $y_0 = y_N, \quad y_1 = y_{N+1}$,
(3) $y_0 = 0, \quad y_{N-1} - xy_N = 0$.

d) Si dimostri che le funzioni

$$f_k(\theta) = \int_0^\pi \frac{\cos kt - \cos k\theta}{\cos t - \cos \theta}\, dt$$

soddisfano l'equazione (95) e si determini un'espressione esplicita per $f_k(\theta)$.

(Traccia: a) l'equazione caratteristica $z^2 - 2z\cos\theta + 1 = 0$ ha le due soluzioni $z_{1,2} = \cos\theta \pm \boldsymbol{i}\sin\theta$, quindi la soluzione generale della (95) è della forma

$$y_k = \beta_1 \cos k\theta + \beta_2 \sin k\theta, \quad \beta_1,\ \beta_2 \in \mathbf{R}.$$

Risulta

$$T_k(\theta) = \cos k\theta, \quad \text{e} \quad U_k(\theta) = \frac{\sin(k+1)\theta}{\sin\theta}\,.$$

b) Si dimostri che i polinomi $T_k(x)$ e $U_{k-1}(x)$ sono linearmente indipendenti per ogni x. c) Nel caso (1) vi sono soluzioni non banali se $\sin N\theta = 0$, e sono date da

$$y_k = \beta_2 \sin\, \frac{kn\pi}{N}\,, \quad n = 1, \dots, N-1;$$

nel caso (2), imponendo le condizioni si ottiene il sistema lineare omogeneo

$$\begin{cases} \beta_1(1 - \cos N\theta) - \beta_2 \sin N\theta = 0, \\ \beta_1\big[\cos\theta - \cos(N+1)\theta\big] + \beta_2\big[\sin\theta - \sin(N+1)\theta\big] = 0, \end{cases}$$

il cui determinante è $2\sin\theta(1 - \cos N\theta)$. Quindi vi sono soluzioni non banali se $\cos N\theta = 1$, e sono date da

$$y_k = \beta_1 \cos\, \frac{2kn\pi}{N} + \beta_2 \sin\, \frac{2kn\pi}{N}\,, \quad n = 0, \dots, N-1.$$

Nel caso (3) imponendo le condizioni si ha

$$\beta_1 = 0, \quad \beta_2 \sin\theta \cos N\theta = 0.$$

Quindi vi sono soluzioni non banali se $\cos N\theta = 0$, e sono date da

$$y_k = \beta_2 \sin\, \frac{k(2n+1)\pi}{2N}\,, \quad n = 0, \dots, N-1.$$

d) Si tenga conto della relazione

$$\cos(k+2)x = 2\cos(k+1)x\cos x - \cos kx.$$

Le condizioni iniziali sono $f_0(\theta) = 0$ e $f_1(\theta) = \pi$, da cui si ricava la soluzione particolare

$$f_k(\theta) = \frac{\pi \sin k\theta}{\sin \theta} \,.\ )$$

4.69 Si dica se la soluzione dell'equazione alle differenze

$$y_{k+2} - \frac{3}{2}\, y_{k+1} + \frac{1}{2}\, y_k = \frac{1}{k+1}$$

che soddisfa le condizioni iniziali $y_0 = 0$, $y_1 = 0$, è limitata.

(Traccia: tenendo conto degli esercizi 4.25 e 4.27, si verifichi che la funzione $z_k = \Psi(k) + \gamma - 1$ è tale che

$$z_0 < y_0, \quad z_1 = y_1, \quad \text{e per } k \geq 2 \quad z_k < y_k \text{ e } \Delta z_k < \Delta y_k$$

per induzione su k, notando che

$$\begin{aligned} y_{k+2} &= y_{k+1} + \frac{1}{2}\, \Delta y_k + \frac{1}{k+1}\,, \\ z_{k+2} &= z_{k+1} + \frac{1}{2}\, \Delta z_k + \frac{k}{2(k+1)(k+2)}\,. \end{aligned}$$

Poiché $\{z_k\}$ non è limitata, ne segue che $\{y_k\}$ non è limitata.

4.70 Sia

$$y_k = \sum_{i=1}^{n} \alpha_i z_k^{(i)}$$

la soluzione generale dell'equazione omogenea (56). Per determinare una soluzione particolare dell'equazione completa (53) della forma

$$z_k = \sum_{i=1}^{n} \alpha_i(k) z_k^{(i)}, \tag{96}$$

si può procedere nel seguente modo: poiché

$$z_{k+1} = \sum_{i=1}^{n} \alpha_i(k+1) z_{k+1}^{(i)} = \sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+1}^{(i)} + \sum_{i=1}^{n} \alpha_i(k) z_{k+1}^{(i)},$$

imponendo che

$$\sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+1}^{(i)} = 0,$$

risulta

$$z_{k+1} = \sum_{i=1}^{n} \alpha_i(k) z_{k+1}^{(i)};$$

poiché

$$z_{k+2} = \sum_{i=1}^{n} \alpha_i(k+1) z_{k+2}^{(i)} = \sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+2}^{(i)} + \sum_{i=1}^{n} \alpha_i(k) z_{k+2}^{(i)},$$

imponendo che

$$\sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+2}^{(i)} = 0,$$

risulta

$$z_{k+2} = \sum_{i=1}^{n} \alpha_i(k) z_{k+2}^{(i)};$$

e si ripete fino ad ottenere z_{k+n-1}. Si impongono cioè le $n-1$ condizioni

$$\sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+j}^{(i)} = 0, \quad \text{per} \quad j = 1, \ldots, n-1, \tag{97}$$

ottenendo le $n-1$ relazioni

$$z_{k+j} = \sum_{i=1}^{n} \alpha_i(k) z_{k+j}^{(i)}, \quad \text{per} \quad j = 1, \ldots, n-1. \tag{98}$$

Risulta in tal modo che

$$z_{k+n} = \sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+n}^{(i)} + \sum_{i=1}^{n} \alpha_i(k) z_{k+n}^{(i)}. \tag{99}$$

Si dimostri che sostituendo le (98) e (99) nell'equazione alle differenze (53) si ottiene la relazione

$$a_n \sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+n}^{(i)} = b(k). \tag{100}$$

Il sistema lineare formato dalle (97) e dalla (100) consente di ricavare le differenze $\Delta\alpha_i(k)$ per $i = 1, \ldots, n$, e quindi di determinare i parametri $\alpha_i(k)$ della soluzione particolare (96). Questo metodo per determinare una soluzione particolare dell'equazione (53) è detto *metodo della variazione*

dei parametri. Con tale metodo si determini una soluzione particolare dell'equazione

$$y_{k+2} - 7y_{k+1} + 12y_k = 6 \cdot 2^k.$$

(Traccia: sostituendo le (98) e (99) nella (53) risulta

$$\begin{aligned} b(k) &= \sum_{j=0}^{n} a_j z_{k+j} = \sum_{j=0}^{n} a_j \sum_{i=1}^{n} \alpha_i(k) z_{k+j}^{(i)} + a_n \sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+n}^{(i)} \\ &= \sum_{i=1}^{n} \alpha_i(k) \sum_{j=0}^{n} a_j z_{k+j}^{(i)} + a_n \sum_{i=1}^{n} \Delta\alpha_i(k) z_{k+n}^{(i)}. \end{aligned}$$

La (100) segue dal fatto che le $z_k^{(i)}$ sono soluzioni della (56). Nel caso particolare, l'equazione omogenea associata ha la soluzione generale

$$y_k = \alpha_1 3^k + \alpha_2 4^k.$$

Si cerca una soluzione particolare dell'equazione completa della forma

$$z_k = \alpha_1(k)\ 3^k + \alpha_2(k)\ 4^k.$$

Da (97) e (100) si ha

$$\begin{cases} \Delta\alpha_1(k)\ 3^{k+1} + \Delta\alpha_2(k)\ 4^{k+1} = 0 \\ \Delta\alpha_1(k)\ 3^{k+2} + \Delta\alpha_2(k)\ 4^{k+2} = 6 \cdot 2^k, \end{cases}$$

da cui si ricava

$$\Delta\alpha_1(k) = -2\ \big(\frac{2}{3}\big)^k \quad \text{e} \quad \Delta\alpha_2(k) = \frac{3}{2}\ \big(\frac{1}{2}\big)^k,$$

e quindi

$$\alpha_1(k) = 6\ \big(\frac{2}{3}\big)^k, \qquad \alpha_2(k) = -3\ \big(\frac{1}{2}\big)^k.\ )$$

4.71 Siano $a_0, \ldots, a_n \in \mathbf{R}$, si definisce l'*operatore lineare alle differenze*

$$\mathcal{L} = a_n E^n + a_{n-1} E^{n-1} + \ldots + a_1 E + a_0 1.$$

Con questa definizione l'equazione alle differenze (53) può essere scritta $\mathcal{L} y_k = b(k)$.

a) Si dimostri che se $\mathcal{L}_1$ e $\mathcal{L}_2$ sono due operatori lineari alle differenze, allora $\mathcal{L}_1 \mathcal{L}_2 = \mathcal{L}_2 \mathcal{L}_1$;

b) se $\mathcal{L}_{11}$, $\mathcal{L}_{12}$, $\mathcal{L}_{21}$, $\mathcal{L}_{22}$ sono operatori lineari alle differenze, si dimostri che le soluzioni $\{u_k, v_k\}$ del sistema alle differenze

$$\begin{cases} \mathcal{L}_{11}u_k + \mathcal{L}_{12}v_k = b(k) \\ \\ \mathcal{L}_{21}u_k + \mathcal{L}_{22}v_k = c(k) \end{cases}$$

sono anche soluzioni del sistema

$$\begin{cases} \mathcal{M}u_k = \mathcal{L}_{22}b(k) - \mathcal{L}_{12}c(k) \\ \\ \mathcal{M}v_k = \mathcal{L}_{11}c(k) - \mathcal{L}_{21}b(k), \end{cases}$$

dove $\mathcal{M} = \mathcal{L}_{11}\mathcal{L}_{22} - \mathcal{L}_{12}\mathcal{L}_{21}$;

c) si calcoli il rapporto u_k/v_k, dove $\{u_k, v_k\}$ è la soluzione del sistema alle differenze omogeneo

$$\begin{cases} u_{k+1} - u_k - xv_k = 0 \\ \\ v_{k+1} + xu_k - v_k = 0, \end{cases} \qquad u_0 = 0, \quad v_0 = 1, \quad x \in \mathbf{R}.$$

(Traccia: a) si sfrutti la linearità dell'operatore E; b) si applichi l'operatore $\mathcal{L}_{22}$ alla prima equazione e $\mathcal{L}_{12}$ alla seconda, poi si sottragga e si sfrutti la commutatività; si proceda in modo analogo con la seconda equazione; si noti che non tutte le soluzioni del secondo sistema sono anche soluzioni del primo sistema; c) si risolve il sistema

$$\begin{cases} u_{k+2} - 2u_{k+1} + (1+x^2)u_k = 0 \\ \\ v_{k+2} - 2v_{k+1} + (1+x^2)v_k = 0. \end{cases}$$

La soluzione generale è

$$u_k = (1+x^2)^{k/2}(\alpha_1 \cos k\theta + \alpha_2 \sin k\theta),$$
$$v_k = (1+x^2)^{k/2}(\beta_1 \cos k\theta + \beta_2 \sin k\theta),$$
$$\theta = \arctan |x|.$$

Sostituendo nel sistema dato, risulta che $\alpha_2 = \beta_1$ e $\alpha_1 = -\beta_2$, e imponendo le condizioni iniziali risulta che $\alpha_2 = 1$ e $\alpha_1 = 0$. Quindi

$$\frac{u_k}{v_k} = \frac{(1+x^2)^{k/2} \sin k\theta}{(1+x^2)^{k/2} \cos k\theta} = \tan k\theta. \ )$$

4.72 Si consideri l'equazione alle differenze

$$\sum_{i=0}^{n} a_i y_{k+i} = 0, \quad \text{con} \quad a_n = 1.$$

Si verifichi che, posto

$$\mathbf{y}^{(k)} = (y_k, y_{k+1}, \ldots, k_{k+n-1})^T,$$

vale la relazione

$$\mathbf{y}^{(k+1)} = F\mathbf{y}^{(k)}, \quad \text{dove} \quad F = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_0 & -a_1 & \ldots & -a_{n-2} & -a_{n-1} \end{bmatrix},$$

e quindi

$$\mathbf{y}^{(k)} = F^k \mathbf{y}^{(0)}.$$

Si determini il costo computazionale del calcolo di $\mathbf{y}^{(k)}$ quando si utilizzi un algoritmo analogo a quello descritto nell'esercizio 3.63.)

4.73 Posto $\phi = \dfrac{\sqrt{5}-1}{2}$, si verifichi che le potenze n-esime di ϕ verificano la relazione

$$\phi^n = \phi^{n-2} - \phi^{n-1}.$$

Si dica se tale relazione è stabile per calcolare ϕ^n, per $n \geq 2$, a partire dalle due condizioni iniziali $\phi^0 = 1$ e $\phi^1 = \phi$.

(Traccia: si verifichi che ϕ^n rappresenta una soluzione minimale dell'equazione alle differenze.)

4.74 Si dimostri che gli integrali

a) $$y_k = \int_0^1 \mathrm{e}^{x/3} x^{k+3}\, dx, \quad k \geq -3,$$

b) $$y_k = \int_1^\infty \frac{\sqrt{3x+1}}{x^k}\, dx, \quad k \geq 2,$$

c) $$y_k = \int_0^1 x^k (2x+1)^k\, dx, \quad k \geq 0,$$

d) $$y_k = \int_1^\infty \frac{dx}{\left(x^2 + \frac{1}{2}\right)^k}, \quad k \geq 1,$$

soddisfano equazioni alle differenze del primo ordine e per ciascuno di essi si determini la condizione iniziale e si dica se la relazione ricorrente, ottenuta dall'equazione e applicata in avanti, è stabile.

(Traccia:

a) $$y_{k+1} + 3(k+4)y_k = 3\,\mathrm{e}^{1/3}, \quad \text{con} \quad y_{-3} = 3(\mathrm{e}^{1/3} - 1);$$

la soluzione cercata è minimale.

b) $$y_{k+1} + \Big(3 - \frac{9}{2k}\Big)y_k = \frac{8}{k}\,, \quad \text{con} \quad y_2 = 2 + \frac{3}{2}\,\log 3;$$

la soluzione cercata è minimale.

c) $$y_{k+1} + \frac{k+1}{4(2k+3)}\,y_k = \frac{5}{4}\,\frac{3^{k+1}}{2k+3}\,, \quad \text{con} \quad y_0 = 1;$$

la soluzione cercata non è minimale.

d) $$y_{k+1} - \big(2 - \frac{1}{k}\big)y_k = -\frac{1}{k}\,\Big(\frac{2}{3}\Big)^k, \quad \text{con} \quad y_1 = \sqrt{2}\Big(\frac{\pi}{2} - \arctan\sqrt{2}\Big);$$

la soluzione cercata è minimale.)

4.75 Si dimostri che gli integrali

a) $$y_k = \int_1^2 \frac{\sqrt{(x-1)^k}}{x}\,dx, \quad k \geq 0,$$

b) $$y_k = \int_0^1 \frac{x^k}{x^2 + x + 1}\,dx, \quad k \geq 0,$$

c) $$y_k = \int_0^1 \frac{x^k}{(x^2 + x + 1)^2}\,dx, \quad k \geq 2,$$

d) $$y_k = \int_0^{\pi/4} x^k \sin x\,dx, \quad k \geq 0,$$

soddisfano equazioni alle differenze del secondo ordine e per ciascuno di essi si determinino le condizioni iniziali e si dica se la relazione ricorrente, ottenuta dall'equazione e applicata in avanti, è stabile.

(Traccia:

a) $$y_{k+2} + y_k = \frac{2}{k+2}\,, \quad \text{con} \quad y_0 = \log 2, \quad y_1 = 2 - \frac{\pi}{2}\,;$$

la soluzione cercata è minimale.

b) $$y_{k+2} + y_{k+1} + y_k = \frac{1}{k+1}\,,$$
$$\text{con} \quad y_0 = \frac{\pi}{3\sqrt{3}}\,, \quad y_1 = \frac{1}{2}\left(\log 3 - \frac{\pi}{3\sqrt{3}}\right);$$
la soluzione cercata è minimale.

c) $$y_{k+2} + \frac{k}{k-1}\,y_{k+1} + \frac{k+1}{k-1}\,y_k = \frac{1}{3(k-1)}\,,$$
$$\text{con} \quad y_2 = \frac{1}{3}\left(\frac{2\pi}{3\sqrt{3}} - 1\right), \quad y_3 = \frac{1}{2}\left(\log 3 - \frac{5\pi}{9\sqrt{3}}\right);$$
la soluzione cercata è minimale, si verifichi infatti che una soluzione particolare dell'equazione omogenea associata è
$$z_k = \sqrt{3}(k-1)\cos\frac{2k\pi}{3} + \sin\frac{2k\pi}{3}\,.$$

d) $$y_{k+2} + (k+2)(k+1)y_k = \left(\frac{\pi}{4}\right)^{k+1}\frac{\sqrt{2}}{2}\left(k+2-\frac{\pi}{4}\right),$$
$$\text{con} \quad y_0 = 1 - \frac{\sqrt{2}}{2}\,, \quad y_1 = \frac{\sqrt{2}}{2}\left(1 - \frac{\pi}{4}\right);$$
la soluzione cercata è minimale.)

4.76 Si supponga che la soluzione generale dell'equazione omogenea (49) sia della forma
$$z_k = \sum_{j=1}^{n} \alpha_j \rho_j^k, \quad \text{con} \quad |\rho_1| \le \ldots \le |\rho_{n-1}| < |\rho_n|,$$
e si faccia l'ipotesi che per assegnate condizioni iniziali $y_0, \ldots, y_{n-1}$ sia $\alpha_n = 0$. Si dimostri che il problema del calcolo di z_k è mal condizionato.

(Traccia: indicato con c_i il coefficiente di amplificazione di z_k rispetto a y_i, si ha
$$c_i = \frac{\partial z_k}{\partial y_i}\,\frac{y_i}{z_k} = \frac{y_i}{z_k}\sum_{j=1}^{n}\frac{\partial z_k}{\partial \alpha_j}\,\frac{\partial \alpha_j}{\partial y_i} = \frac{y_i}{z_k}\sum_{j=1}^{n}\rho_j^k\,\frac{\partial \alpha_j}{\partial y_i}\,.$$
Poiché $\dfrac{\partial \alpha_j}{\partial y_i}$ non dipende da k e
$$\lim_{k\to\infty}\frac{z_k}{\rho_n^k} = \lim_{k\to\infty}\sum_{s=1}^{n-1}\alpha_s\,\frac{\rho_s^k}{\rho_n^k} = 0,$$
il coefficiente c_i, come funzione di k, non è superiormente limitato in modulo. Inoltre
$$|c_i| = O\Big(\Big|\frac{\rho_n}{\rho_{n-1}}\Big|^k\Big).\ )$$

Commento bibliografico

I metodi alle differenze finite furono introdotti nel 1600 da Harriot e successivamente da Briggs, per la compilazione delle tabelle dei logaritmi e delle funzioni circolari, usate per le misurazioni geometriche nel campo dell'astronomia per usi nautici. È dallo studio dei procedimenti matematici che stanno alla base della compilazione delle tabelle dei logaritmi che trae origine l'analisi numerica [10]. Il simbolo Δ fu utilizzato da Bernoulli nel 1706 per indicare la differenza finita del primo ordine; adottato poi da Eulero è giunto fino ai nostri tempi senza modifiche. I simboli E, ∇, δ, μ, ecc. introdotti da autori diversi per indicare gli altri operatori lineari, hanno subito vari cambiamenti di significato nel tempo. Una trattazione teorica di questi operatori è stata fatta da Boole nel 1860 [3] e da Milne-Thomson [15]. Si veda anche il più recente libro di Jordan [12].

La funzione $\Gamma(z)$ (la notazione $\Gamma(z)$ fu introdotta da Legendre nel 1814) fu definita da Eulero nel 1729 come

$$\Gamma(z) = \lim_{n\to\infty} \frac{(n-1)!}{z(z+1)\dots(z+n-1)}\, n^z;$$

da questa definizione può essere ottenuta le relazione con l'integrale infinito che viene ora assunta come definizione. La funzione $\Gamma(z)$ potrebbe essere definita anche tramite una qualunque delle relazioni che la coinvolgono e che sono state travate da matematici dell'800 (si vedano [6] e [18]). Weierstrass definiva la funzione $\Gamma(z)$ per mezzo di un prodotto infinito, contenente anche la funzione esponenziale e la costante γ, che si presta bene a valutare prodotti infiniti di funzioni razionali. Per un un teorema formulato da Weierstrass, la funzione $\Gamma(z)$ non soddisfa alcuna equazione differenziale a coefficienti razionali. La costante γ è nota anche come costante di Eulero-Mascheroni (quest'ultimo fu un matematico italiano della seconda metà del 700).

I numeri di Bernoulli furono introdotti da Jakob Bernoulli nel suo famoso libro *Ars conjectandi*, pubblicato postumo nel 1713: i primi 66 numeri di Bernoulli furono calcolati da Adams nel 1877. I polinomi di Bernoulli furono invece introdotti da Raabe nel 1851.

Pur mancando di una rigorosa base teorica, il calcolo delle serie è stato uno dei problemi che più hanno affascinato i matematici: già Archimede aveva calcolato la somma della serie geometrica di ragione 1/4. La serie dell'arcotangente, che tanta importanza ha avuto per il calcolo di π, fu trovata da Leibniz e da Gregory nel 1688. Successivamente con la serie di Taylor le serie sono diventate strumento di uso corrente nella matematica. La formula di Eulero-Maclaurin fu trovata da Eulero nel 1732 e pubblicata nel 1738. In quello stesso periodo era stata trovata anche da Maclaurin che la pubblicò nel 1742. Nell'800 il problema della somma di funzioni

fu nuovamente affrontato da molti matematici, fra cui Abel e Cauchy e soprattutto Nörlung, che poté sfruttare la teoria della variabile complessa introdotta da Cauchy.

Già nel 1789 Lagrange usò uno schema basato su una ricorrenza non lineare per calcolare un integrale ellittico. Oggi le equazioni alle differenze sono applicate nei più diversi campi: nello studio dei modelli economici, nelle simulazioni del traffico, nella teoria delle code, nella dinamica delle popolazioni. Per applicazioni delle equazioni alle differenze nella realizzazione di modelli di fenomeni del mondo reale si veda [13]. Per applicazioni alla risoluzione di equazioni differenziali si veda [9], [4] e [11].

Molte delle funzioni speciali soddisfano a relazioni di ricorrenza lineari del secondo ordine. Il primo algoritmo efficace per la tabulazione di funzioni corrispondenti a soluzioni minimali è stato proposto da Miller nel 1952 [14] per le funzioni di Bessel modificate: Miller notò che per calcolare una soluzione minimale di un'equazione alle differenze del secondo ordine, è sufficiente un solo valore iniziale. In realtà la tecnica usata da Miller si basa su alcune osservazioni fatte da Lord Rayleigh nel 1910. Sull'algoritmo di Miller è basato l'algoritmo di Olver che nel 1967 [16] osservò l'equivalenza dell'algoritmo di Miller con il metodo di sostituzione per la risoluzione di un sistema lineare tridiagonale. Per lo studio della instabilità del calcolo delle soluzioni minimali per mezzo delle relazioni di ricorrenza, per la descrizione del metodo di Olver, per l'estensione di questo al caso di più equazioni e alle relazioni ricorrenti di ordine superiore al secondo e alla risoluzione numerica delle equazioni differenziali ordinarie, si vedano i libri di Cash [5] e di Wimp [19]. Nel libro di Abramowitz e Stegun [1] sono indicate quali formule ricorrenti sono stabili e quali no. Per una tecnica euristica che consente di stabilire se una relazione ricorrente è stabile, si veda [17]. Per uno studio sistematico delle ricorrenze in avanti e all'indietro nel calcolo di molte funzioni speciali e di integrali si vedano gli articoli di Gautschi, ad esempio [7], [8].

Bibliografia

[1] M. Abramowitz, I. A. Stegun, editors, *Handbook of Mathematical Functions*, National Bureau of Standards, U. S. Gov't Printing Office, 1964.

[2] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, Bologna, 1988.

[3] G. Boole, *Finite Differences*, 3^o Ed., G. E. Stechert & Co., New York, 1931 (1^a Ed. Cambridge, 1860).

[4] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, J. Wiley & Sons, New York, 1987.

[5] J. R. Cash, *Stable Recursions*, Academic Press, New York, 1979.

[6] A. Erdélyi, W. Magnus, F. Oberhettinger, F. G. Tricomi, *Higher Trascendental Functions*, vol. 1, McGraw-Hill, New York, 1953.

[7] W. Gautschi, "Computational Aspects of Three Terms Recurrence Relations", *SIAM Rev.*, 9, 1967, pp. 24-82.

[8] W. Gautschi, "Minimal Solutions of Three Terms Recurrence Relations and Orthogonal Polynomials", *Math. Comput.*, 36, 1981, pp. 547-554.

[9] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N. J., 1971.

[10] H. H. Goldstine, *A History of Numerical Analysis from the 16th through the 19th Century*, Springer-Verlag, New York, 1977.

[11] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, J. Wiley & Sons, New York, 1962.

[12] C. Jordan, *Calculus of Finite Differences*, Chelsea, New York, 1950.

[13] V. Lakshmikantham, D. Trigiante, *Theory of Difference Equations: Numerical Methods and Applications*, Academic Press, New York, 1988.

[14] J. C. P. Miller, *Bessel Functions, Part II*, Mathematical Tables, British Association for the Advancement of Sciences, Vol. 10, Cambridge Univ. Press, 1952.

[15] L. M. Milne-Thomson, *The Calculus of Finite Differences*, Macmillan and Co., London, 1933.

[16] F. W. J. Olver, "Numerical Solution of Second Order Linear Difference Equations", *Journal of Research, N.B.S.*, 71B, 1967, pp. 111-129.

[17] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes. The Art of Scientific Computing*, Cambridge Univ. Press, Cambridge, 1986.

[18] E. T. Whittaker, G. N. Watson, *A Course of Modern Analysis*, Cambridge Univ. Press, 1963.

[19] J. Wimp, *Computation with Recurrence Relations*, Pitman, 1984.

Capitolo 5

INTERPOLAZIONE

1. Il problema dell'interpolazione

Un classico problema di approssimazione di funzioni è il problema dell'*interpolazione*, che in generale viene formulato nel modo seguente:

a) di una funzione $f(x)$ sono noti i valori $y_i = f(x_i)$ per $i = 0, 1, \ldots, n$; i punti x_i, appartenenti ad un intervallo $[a, b]$, sono detti *nodi* dell'interpolazione;

b) è fissato un insieme di $n + 1$ funzioni $\phi_j(x)$, $j = 0, \ldots, n$, definite su $[a, b]$ e ivi *linearmente indipendenti*, cioè tali che una loro combinazione lineare a coefficienti reali

$$\sum_{j=0}^{n} c_j \phi_j(x)$$

è identicamente nulla su $[a, b]$ se e solo se $c_j = 0$, $j = 0, \ldots, n$;

si tratta di determinare una funzione

$$g(x) = \sum_{j=0}^{n} \alpha_j \phi_j(x)$$

che assuma gli stessi valori y_i nei punti x_i, cioè $g(x_i) = y_i$, $i = 0, \ldots, n$.

Un importante aspetto di questo problema consiste nell'individuare la classe delle funzioni $\phi_j(x)$ o, come si dice, di scegliere il *modello* della approssimazione. La scelta di questa classe di funzioni deve tenere conto di specifiche proprietà della funzione $f(x)$. Inoltre si richiede che le funzioni $\phi_j(x)$ siano facilmente calcolabili e dotate di buone proprietà di regolarità. Due sono le classi di funzioni più usate: la classe delle funzioni razionali (fra cui i polinomi) e la classe delle funzioni trigonometriche.

Si esamina dapprima il caso in cui $\phi_j(x) = x^j$ (*interpolazione polinomiale*). Sotto opportune ipotesi è possibile dimostrare che la funzione $g(x)$ esiste ed è unica.

5.1 Teorema. *Se (x_i, y_i), $i = 0, \ldots, n$, sono $n + 1$ punti tali che $x_i \neq x_j$ per $i \neq j$, esiste ed è unico il polinomio $p_n(x)$ di grado al più n, tale che*

$$p_n(x_i) = y_i, \quad i = 0, 1, \ldots, n. \tag{1}$$

Tale polinomio viene detto polinomio di interpolazione della funzione $f(x)$ sui nodi x_i, $i = 0, \ldots, n$.

Dim. Si considerino il vettore $\mathbf{a} = [a_0, a_1, \ldots, a_n]^T$ dei coefficienti del polinomio

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0, \tag{2}$$

il vettore $\mathbf{y} = [y_0, y_1, \ldots, y_n]^T$ e la matrice V, detta matrice di *Vandermonde*, così definita

$$v_{ij} = x_{i-1}^{j-1}, \quad i = 1, \ldots n+1, \ j = 1, \ldots, n+1.$$

Imponendo che $p_n(x)$ verifichi le $n+1$ condizioni (1), si ottiene il sistema lineare di $n+1$ equazioni in $n+1$ incognite

$$V\mathbf{a} = \mathbf{y}. \tag{3}$$

La matrice di Vandermonde è non singolare, in quanto (si veda l'esercizio 5.1)

$$\det\ V = \prod_{\substack{i,j=0 \\ i>j}}^{n} (x_i - x_j)$$

e i punti x_i sono per ipotesi a due a due distinti. Ne segue che il sistema (3) ha una e una sola soluzione e quindi il polinomio $p_n(x)$ esiste ed è unico. ∎

Solo raramente del polinomio di interpolazione sono richiesti i coefficienti, in generale ciò che si vuole è il valore di $p_n(x)$ in uno o più punti. Per calcolare $p_n(x)$ non è però conveniente risolvere il sistema (3) perché ciò richiederebbe un numero di operazioni moltiplicative dell'ordine di $n^3/3$. Inoltre la matrice di Vandermonde può essere malcondizionata: in tal caso il calcolo di $p_n(x)$ attraverso la risoluzione del sistema (3) sarebbe numericamente instabile.

5.2 Esempio. Per $n \geq 4$ il polinomio di interpolazione della funzione che nei punti $x_i = i/n$, $i = 0, \ldots, n$ assume i valori $y_i = x_i^4 + 0.1$ è $p_n(x) = x^4 + 0.1$. Per determinare i coefficienti di tale polinomio si scrive e si risolve il sistema (3). La seguente tabella riporta al crescere di n il numero di condizionamento $\mu_\infty(V)$ della matrice di Vandermonde e il massimo errore assoluto ϵ da cui sono affetti i coefficienti effettivamente calcolati utilizzando

il metodo di Gauss.

n	$\mu_\infty(V)$	ϵ
5	$1.04\ 10^3$	$6.40\ 10^{-6}$
6	$8.06\ 10^3$	$4.37\ 10^{-5}$
7	$6.30\ 10^4$	$1.07\ 10^{-4}$
8	$4.96\ 10^5$	$1.64\ 10^{-3}$
9	$3.92\ 10^6$	$3.31\ 10^{-2}$
10	$3.11\ 10^7$	$1.24\ 10^{-1}$

È evidente che già per $n = 10$ i coefficienti ottenuti sono privi di significato. Cambiando i punti x_i si possono ottenere, a parità di n, dei risultati migliori, anche se non soddisfacenti. Ad esempio se i punti x_i sono i seguenti

$$x_i = \frac{1}{2}\left[1 - \cos\frac{(2i-1)\pi}{2(n+1)}\right], \quad i = 0, \ldots, n,$$

per $n = 10$ risulta $\mu_\infty(V) = 6.19\ 10^6$ e $\epsilon = 2.92\ 10^{-2}$. ∎

Il polinomio di interpolazione, pur essendo unico, può essere però rappresentato in diverse forme più convenienti sia dal punto di vista del costo computazionale che della stabilità. Si studieranno qui la forma di Lagrange e la forma di Newton. Il costo computazionale verrà dato come funzione di n, riportando solo i termini di ordine superiore in n.

2. Polinomio di Lagrange

Per scrivere il polinomio di interpolazione nella forma di Lagrange si utilizzano come funzioni $\phi_j(x)$ i polinomi di grado n

$$L_j(x) = \prod_{\substack{i=0 \\ i\neq j}}^{n} \frac{x - x_i}{x_j - x_i}, \quad j = 0, \ldots, n, \tag{4}$$

che sono linearmente indipendenti in quanto

$$L_j(x_i) = \begin{cases} 1 & \text{se } i = j, \\ \\ 0 & \text{se } i \neq j. \end{cases}$$

Allora il polinomio

$$p_n(x) = \sum_{j=0}^{n} y_j L_j(x), \tag{5}$$

che ha grado al più n, è tale che

$$p_n(x_i) = \sum_{j=0}^{n} y_j L_j(x_i) = y_i, \quad i = 0, \ldots, n,$$

e quindi $p_n(x)$ soddisfa alle (1).

5.3 Definizione. Il polinomio $p_n(x)$, rappresentato nella forma (5), è il polinomio di interpolazione nella *forma di Lagrange*. ∎

5.4 Esempio (*interpolazione lineare*). Nella forma di Lagrange, il polinomio di grado al più 1 che assume nei punti x_0, x_1, distinti tra loro, rispettivamente i valori y_0, y_1 è il seguente

$$p_1(x) = \frac{x - x_1}{x_0 - x_1}\, y_0 + \frac{x - x_0}{x_1 - x_0}\, y_1.$$

Nella figura 5.1 è riportato il grafico del polinomio $p_1(x)$ di interpolazione lineare (linea tratteggiata) di una funzione $f(x)$ (linea continua) che nei punti x_0 e x_1 assume rispettivamente i valori y_0 e y_1.

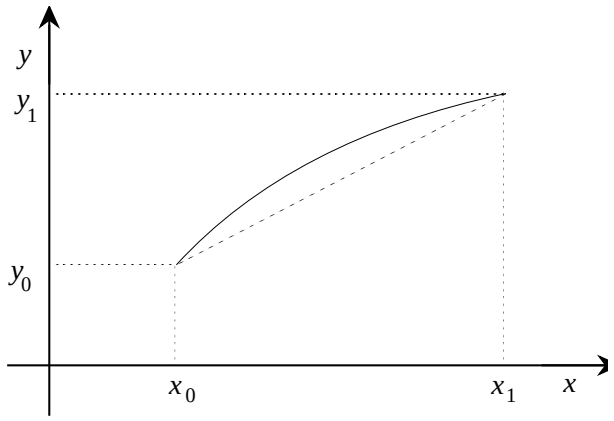


Fig. 5.1 - Interpolazione lineare.

L'interpolazione lineare può essere usata anche per determinare il valore che una funzione tabulata assume in un punto diverso da quelli riportati nelle tavole. Ad esempio, nelle tavole delle funzioni trigonometriche si trova

x	$\tan x$
1.35	4.4552
1.36	4.6734

Per approssimare il valore di $\tan 1.354$ si può fare un'interpolazione lineare, utilizzando i nodi $x_0 = 1.35$ e $x_1 = 1.36$. Il valore $p_1(1.354) = 4.54248$ si assume come valore approssimato di $\tan 1.354$. ∎

Utilizzando la (5) il polinomio di interpolazione potrebbe essere trasformato nella forma (2), per calcolarne il valore in un punto x con il metodo di Ruffini-Horner, ma in questo modo sono richieste, per il solo calcolo dei coefficienti del polinomio, un numero di operazioni additive e moltiplicative superiore a $6n^2$. Un valore $p_n(x)$ per $x \neq x_j$, $j = 0, \ldots, n$, si ottiene invece con un minor numero di operazioni procedendo nel modo seguente: posto

$$\pi_n(x) = (x - x_0) \ldots (x - x_n), \tag{6}$$

si rappresenta il polinomio nella forma

$$p_n(x) = \pi_n(x) \sum_{j=0}^{n} \frac{z_j}{x - x_j}\,, \tag{7}$$

dove

$$z_j = \frac{y_j}{\prod\limits_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i)}\,.$$

Le operazioni richieste sono, a meno di temini di ordine inferiore:

n sottrazioni per calcolare $x - x_i$, $i = 0, \ldots, n$;

$n^2/2$ sottrazioni per calcolare $x_j - x_i$, $j = 0, \ldots, n$, $\;i = j + 1, \ldots, n$;

n^2 moltiplicazioni per calcolare z_j, $j = 0, \ldots, n$;

n addizioni e $2n$ moltiplicazioni per calcolare $p_n(x)$.

In totale il calcolo di $p_n(x)$ in un punto x con il polinomio di Lagrange nella forma (7) richiede $n^2/2$ addizioni e n^2 moltiplicazioni.

Nel caso in cui si voglia calcolare il valore dello stesso polinomio di interpolazione in un altro punto, conviene utilizzare ancora il polinomio di Lagrange. Infatti le quantità z_j, $j = 0, \ldots, n$, che dipendono solo dagli x_i e non dal punto x, vengono calcolate una sola volta, e il calcolo del nuovo valore del polinomio richiede solo altre $2n$ addizioni e $2n$ moltiplicazioni.

Inoltre, una volta calcolato in un punto il polinomio di interpolazione di Lagrange sui nodi (x_i, y_i), $i = 0, \ldots, n$, è possibile calcolare nello stesso punto il polinomio di interpolazione di Lagrange sui nodi (x_i, y_i), $i = 0, \ldots, n + 1$, utilizzando le quantità $z_j/(x - x_j)$, $j = 0, \ldots, n$ e il valore di $\pi_n(x)$ già calcolati, con solo altre $2n$ addizioni e $2n$ moltiplicazioni.

Infine cambiando i valori y_j ma lasciando inalterati i nodi x_j, con la (7) si calcola $p_n(x)$ con solo altre n addizioni e $2n$ moltiplicazioni.

Per mezzo della (6) si può dare di $L_j(x)$ un'altra espressione più compatta che non viene in generale usata per il calcolo effettivo del polinomio di interpolazione, ma che sarà utilizzata in seguito. Poiché

$$\pi_n'(x_j) = \prod_{\substack{i=0 \\ i\neq j}}^{n} (x_j - x_i),$$

sostituendo nella (4) si ha

$$L_j(x) = \frac{\pi_n(x)}{(x - x_j)\pi_n'(x_j)}\,. \tag{8}$$

Il polinomio di Lagrange assume un'espressione particolarmente semplice quando i punti x_i sono *equidistanti di passo* h, cioè

$$x_i = x_0 + ih, \quad i = 0, \ldots, n, \quad h > 0.$$

In questo caso, eseguendo il cambiamento di variabile

$$x = x_0 + th, \tag{9}$$

e introducendo il polinomio di grado $n+1$

$$\tau_n(t) = t(t-1)\ldots(t-n), \tag{10}$$

si ha dalla (7)

$$p_n(x_0 + th) = \tau_n(t) \sum_{j=0}^{n} \frac{z_j}{t - j}\,,$$

dove

$$z_j = \frac{y_j}{\displaystyle\prod_{\substack{i=0 \\ i\neq j}}^{n} (j - i)}\,.$$

Poiché

$$\prod_{\substack{i=0 \\ i\neq j}}^{n} (j-i) = (-1)^{n-j} \prod_{i=0}^{j-1} (j-i) \prod_{i=j+1}^{n} (i-j) = (-1)^{n-j} j!(n-j)!$$

risulta

$$p_n(x_0 + th) = \tau_n(t) \sum_{j=0}^{n} \frac{(-1)^{n-j} y_j}{j!(n-j)!(t-j)} = \frac{\tau_n(t)}{n!} \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} \frac{y_j}{t-j}\,.$$

Il costo computazionale per il calcolo di $p_n(x_0+th)$ è allora di $n^2/4$ addizioni e $4n$ moltiplicazioni. Gran parte delle operazioni additive sono quelle richieste dal calcolo dei coefficienti binomiali, che conviene costruire, sfruttando le simmetrie, con il triangolo di Tartaglia. Se si considerano già noti i valori di $n!$ e dei coefficienti binomiali, il costo computazionale scende a $2n$ addizioni e $3n$ moltiplicazioni.

3. Resto nell'interpolazione polinomiale

Sia $p_n(x)$ il polinomio di interpolazione della funzione $f(x)$ sui nodi $x_0, \ldots, x_n$, a due a due distinti. Si definisce *resto* dell'interpolazione di $f(x)$ con il polinomio $p_n(x)$ la funzione

$$r(x) = f(x) - p_n(x),$$

che assume il valore zero nei punti x_i. Posto

$$a = \min_{i=0,\ldots,n} x_i, \quad b = \max_{i=0,\ldots,n} x_i,$$

se in $[a, b]$ la funzione $f(x)$ soddisfa a opportune ipotesi di regolarità, allora è possibile trovare un'espressione di $r(x)$ per $x \in [a, b]$.

5.5 Teorema. *Se $f(x) \in C^{n+1}[a, b]$, esiste un punto $\xi = \xi(x) \in (a, b)$, tale che*

$$r(x) = \pi_n(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{11}$$

dove $\pi_n(x)$ è il polinomio di grado $n+1$ definito in (6).

Dim. Per $x = x_i$, $i = 0, \ldots, n$, è

$$r(x_i) = \pi_n(x_i) = 0,$$

e quindi la (11) è verificata. Per $x \neq x_i$ sia

$$s(x) = \frac{r(x)}{\pi_n(x)}, \tag{12}$$

e si consideri la funzione della variabile y

$$z(y) = r(y) - s(x)\pi_n(y).$$

La funzione $z(y)$ si annulla almeno negli $n+2$ punti distinti x_i, $i = 0, \ldots, n$ e x. Inoltre, poiché $z(y)$ è derivabile almeno $n+1$ volte, in quanto $f(y) \in C^{n+1}[a, b]$, per il teorema di Rolle la funzione $z'(y)$ si annulla in almeno $n+1$ punti distinti interni ad $[a, b]$, la $z''(y)$ si annulla in almeno n punti distinti interni ad $[a, b]$, ..., la $z^{(n+1)}(y)$ si annulla in almeno un punto $\xi \in (a, b)$. Allora si ha:

$$0 = z^{(n+1)}(\xi) = r^{(n+1)}(\xi) - (n+1)!\, s(x) = f^{(n+1)}(\xi) - (n+1)!\, s(x), \tag{13}$$

in quanto il polinomio $p_n(x)$ ha la derivata $(n+1)$-esima identicamente nulla. Ricavando $s(x)$ dalla (13) e sostituendo nella (12), si ottiene la (11) per $x \neq x_i$, $i = 0, \ldots, n$. ∎

5.6 Esempio. Il resto dell'interpolazione lineare (esempio 5.4) è dato da

$$r(x) = (x - x_0)(x - x_1)\frac{f''(\xi)}{2}, \quad \xi \in (x_0, x_1).$$

Poiché

$$\max_{x \in (x_0, x_1)} |(x - x_0)(x - x_1)| = \frac{|x_1 - x_0|^2}{4},$$

posto

$$M_2 = \max_{x \in (x_0, x_1)} |f''(x)|,$$

si ha

$$|r(x)| \leq \frac{M_2 |x_1 - x_0|^2}{8}.$$

Per $f(x) = \tan x$ è $f''(x) = \dfrac{2 \sin x}{\cos^3 x}$, e per $x \in [1.35, 1.36]$ è $|r(x)| \leq 0.267\ 10^{-2}$. ∎

5.7 Esempio. Per approssimare la radice quadrata di un numero x, che non sia un quadrato perfetto, si può procedere nel modo seguente: si considerano le radici quadrate dei tre numeri che sono quadrati perfetti e che sono più vicini a x e si costruisce il polinomio di interpolazione che nei punti x_i assume i valori $\sqrt{x_i}$, $i = 0, 1, 2$. La radice cercata viene approssimata con il valore assunto dal polinomio di interpolazione in x. Ad esempio, se $x = 0.6$, conviene considerare

$$\begin{aligned} &x_0 = 0.49, \quad x_1 = 0.64, \quad x_2 = 0.81 \\ &y_0 = \sqrt{0.49} = 0.7, \quad y_1 = \sqrt{0.64} = 0.8, \quad y_2 = \sqrt{0.81} = 0.9. \end{aligned}$$

Dal polinomio di Lagrange si ha

$$\begin{aligned} p_2(x) &= \frac{0.7}{0.048}\,(x - 0.64)(x - 0.81) - \frac{0.8}{0.0255}\,(x - 0.49)(x - 0.81) \\ &\quad + \frac{0.9}{0.0544}\,(x - 0.49)(x - 0.64) \\ &= -0.245098\,x^2 + 0.9436275\,x + 0.2964706, \end{aligned}$$

per cui $p_2(0.6) = 0.7744118$. Per $x \in [0.49, 0.81]$ il resto dell'interpolazione è dato da

$$r(x) = \frac{(x-0.49)(x-0.64)(x-0.81)}{16\sqrt{\xi^5}}, \quad \xi \in (0.49, 0.81).$$

Si ha perciò

$$|r(0.6)| < \frac{0.93\ 10^{-3}}{16\sqrt{0.49^5}} \approx 0.346\ 10^{-3}.$$

In realtà risulta

$$\max_{x\in[0.49,0.81]} |\sqrt{x} - p_2(x)| \approx 0.264\ 10^{-3}$$

e

$$|\sqrt{0.6} - p_2(0.6)| \approx 0.185\ 10^{-3},$$

per cui le 3 cifre più significative di $p_2(0.6)$ sono esatte. Nella figura 5.2 è riportato il grafico della maggiorazione di $|r(x)|$ ottenuta sostituendo 0.49 al posto di ξ.

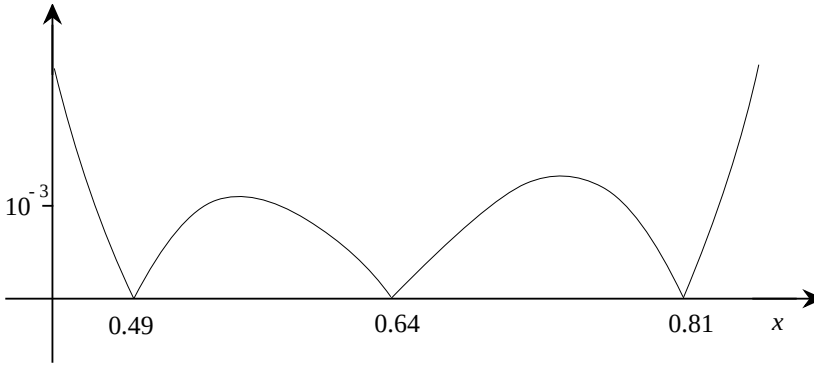


Fig. 5.2 - Maggiorazione di $|r(x)|$ per la funzione $f(x) = \sqrt{x}$.

Un risultato più preciso si ottiene aumentando di uno il grado del polinomio di interpolazione. Assumendo come nodi i 4 punti

$$x_0 = 0.36, \quad x_1 = 0.49, \quad x_2 = 0.64, \quad x_3 = 0.81,$$

si ottiene

$$p_3(x) = 0.2693385\, x^3 - 0.7676147\, x^2 + 1.274618\, x + 0.2280543,$$

per cui $p_3(0.6) = 0.7746606$. In questo caso risulta

$$\max_{x\in[0.36,0.81]} |\sqrt{x} - p_3(x)| \approx 0.165\ 10^{-3}$$

e

$$|\sqrt{0.6} - p_3(0.6)| \approx 0.639\ 10^{-4}. \qquad \blacksquare$$

Nel caso di punti equidistanti il resto del polinomio di interpolazione assume una forma più semplice.

5.8 Teorema. *Siano $x_i = x_0 + ih$, $i = 0, \ldots, n$, e sia $f(x) \in C^{n+1}[a, b]$. Allora esiste un punto $\xi = \xi(x) \in (a, b)$ tale che*

$$r(x_0 + th) = \tau_n(t) h^{n+1} \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

dove $\tau_n(t)$ è il polinomio in t di grado $n+1$ definito in (10).

Dim. La dimostrazione segue subito dal teorema 5.5 e dalla (9). $\blacksquare$

Le espressioni del resto del polinomio di interpolazione date nei teoremi 5.5 e 5.8 permettono di valutare di quanto il polinomio di interpolazione differisce dalla funzione $f(x)$ nell'intervallo $[a, b]$. Posto

$$M_{n+1} = \max_{x\in[a,b]} |f^{(n+1)}(x)|,$$

si ha

$$r_{\max} = \max_{x\in[a,b]} |r(x)| \le \max_{x\in[a,b]} |\pi_n(x)| \, \frac{M_{n+1}}{(n+1)!}. \tag{14}$$

Se i punti x_i sono equidistanti, si ha

$$|r(x_0 + th)| \le |\tau_n(t)| h^{n+1} \frac{M_{n+1}}{(n+1)!},$$

e quindi

$$r_{\max} \le \max_{t\in[0,n]} |\tau_n(t)| h^{n+1} \frac{M_{n+1}}{(n+1)!}.$$

Nel caso di punti equidistanti si può avere una valutazione del comportamento di $r(x_0 + th)$ al variare di t nell'intervallo $[0, n]$, studiando il comportamento del polinomio $\tau_n(t)$. Tale polinomio gode delle seguenti proprietà:

a) $\tau_n(t)$ ha come zeri gli $n+1$ punti $0, \ldots, n$;

b) il punto $\frac{n}{2}$ è punto di simmetria per $|\tau_n(t)|$. Infatti

$$|\tau_n(\tfrac{n}{2}+t)| = |\prod_{i=0}^{n}(\tfrac{n}{2}+t-i)| = |\prod_{i=0}^{n}(i-t-\tfrac{n}{2})|,$$

e con il cambiamento di variabile $j = n - i$, risulta

$$|\tau_n(\tfrac{n}{2}+t)| = |\prod_{j=0}^{n}(\tfrac{n}{2}-t-j)| = |\tau_n(\tfrac{n}{2}-t)|.$$

c) per t non intero, $t \leq \frac{n}{2}$, è $|\tau_n(t-1)| > |\tau_n(t)|$. Infatti

$$|\tau_n(t-1)| = |\prod_{i=0}^{n}(t-1-i)| = |\prod_{j=1}^{n+1}(t-j)| = |\tau_n(t)|\left|\frac{t-n-1}{t}\right|;$$

ed essendo $t \leq \frac{n}{2}$, risulta $|t-n-1| = |n+1-t| > \frac{n}{2}$, e quindi

$$\left|\frac{t-n-1}{t}\right| > 1,$$

da cui la tesi.

d) per la simmetria di $|\tau_n(t)|$ rispetto al punto $\frac{n}{2}$ e per la c), si ha che per t non intero, $t \geq \frac{n}{2}$, risulta $|\tau_n(t)| < |\tau_n(t+1)|$.

Dalle due proprietà c) e d) segue che i massimi relativi di $|\tau_n(t)|$ in ogni intervallo $(i, i+1)$, crescono quando ci si allontana dal centro dell'intervallo $[0, n]$ verso gli estremi. Nelle figure 5.3 e 5.4 sono riportati i grafici di $\tau_n(t)$ per $n = 7$ e per $n = 10$.

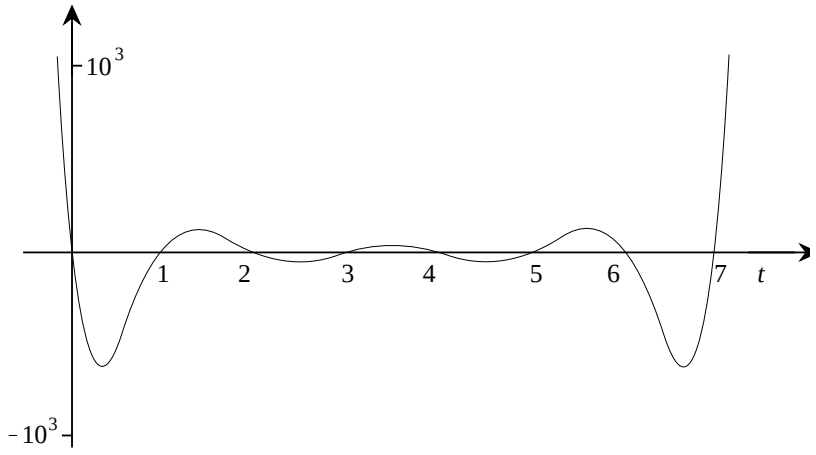


Fig. 5.3 - Grafico di $\tau_7(t)$.

Dallo studio del comportamento di $\tau_n(t)$ si può dedurre che se $f^{(n+1)}(x)$ non varia molto nell'intervallo $[x_0, x_n]$, il resto risulta minore nella parte centrale dell'intervallo.

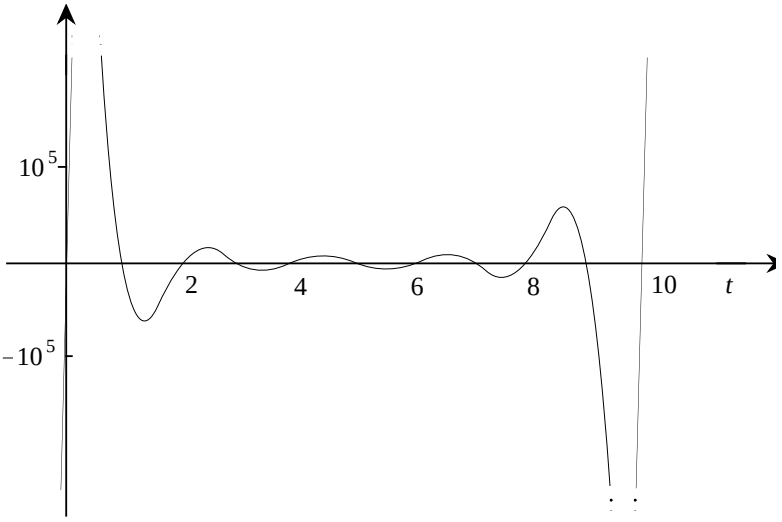


Fig. 5.4 - Grafico di $\tau_{10}(t)$.

Un'altra proprietà rilevante dei polinomi $\tau_n(t)$ è quella della rapida crescita, al crescere di n, del massimo di $|\tau_n(t)|$ in $[0, n]$, che viene assunto nel primo e nell'ultimo sottointervallo. Si ha

$$\begin{aligned}\max_{t\in[0,n]} |\tau_n(t)| &= \max_{t\in[0,1]} |\tau_n(t)| \geq \left|\tau_n\left(\frac{1}{2}\right)\right| = \left|\prod_{j=0}^{n} \left(\frac{1}{2} - j\right)\right| \\ &= \frac{1}{2^{n+1}}\left[1 \cdot 3 \cdot 5 \cdots (2n-1)\right] = \frac{(2n-1)!}{2^{n+1}[2 \cdot 4 \cdots (2n-2)]} \\ &= \frac{(2n-1)!}{2^{2n}(n-1)!}\,.\end{aligned}$$

È opportuno rilevare che, anche se la funzione $f(x) \in C^\infty[a, b]$, 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 che tende all'infinito.

5.9 Esempio. Per il polinomio di interpolazione in n punti equidistanti della funzione di *Runge*

$$f(x) = \frac{1}{1 + x^2}\,,$$

definita sull'intervallo $[a, b] = [-5, 5]$, si può dimostrare [13] che al crescere di n la successione dei polinomi di interpolazione sui nodi $x_i = a+i(b-a)/n$, $i = 0, \ldots, n$, non converge puntualmente a $f(x)$ e che i corrispondenti resti diventano in modulo arbitrariamente grandi in punti dell'intervallo $[-5, 5]$. Le figure 5.5 e 5.6 illustrano l'andamento del polinomio di interpolazione $p_n(x)$ (tabulato in punti equidistanti, linea più sottile) nei due casi $n = 9$ e $n = 15$, rispetto alla funzione $f(x)$ (linea spessa).

∎

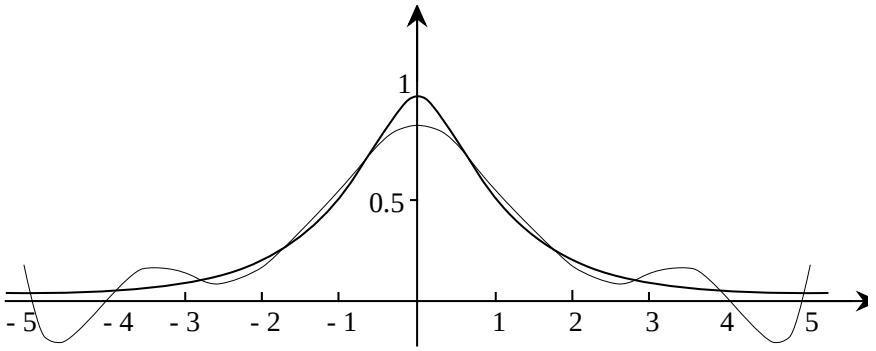


Fig. 5.5 - Polinomio di interpolazione di grado 9 della funzione di Runge.

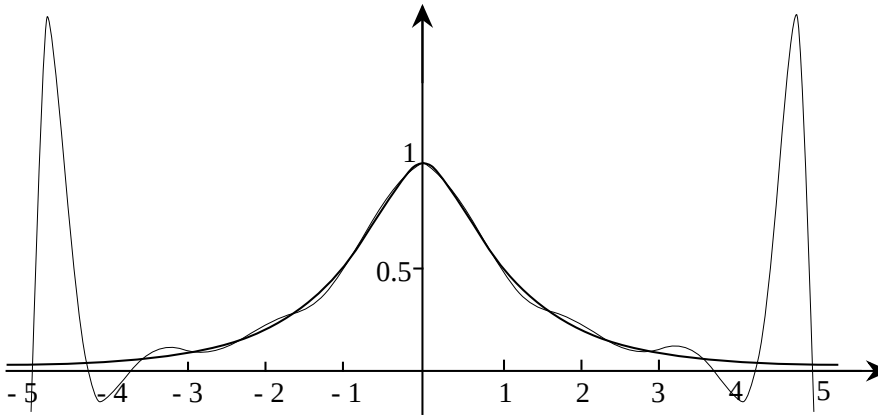


Fig. 5.6 - Polinomio di interpolazione di grado 15 della funzione di Runge.

Perché la successione $\{p_n(x)\}$ converga uniformemente ad $f(x)$ per $x \in [a, b]$, è sufficiente che la funzione $f(x)$ sia analitica in una regione del piano complesso contenente $[a, b]$ e abbastanza ampia. Ad esempio, è possibile dimostrare [7] che se i nodi dell'interpolazione appartengono all'intervallo $I = [a, b]$, allora la successione $\{p_n(x)\}$ converge uniformemente ad $f(x)$ in tutto I se $f(x)$ è analitica nella regione del piano complesso

$$D = \{z \in \mathbf{C} \ : \ |z - x| \leq b - a, \ \text{per ogni } x \in I\}.$$

L'insieme D è illustrato nella figura 5.7.

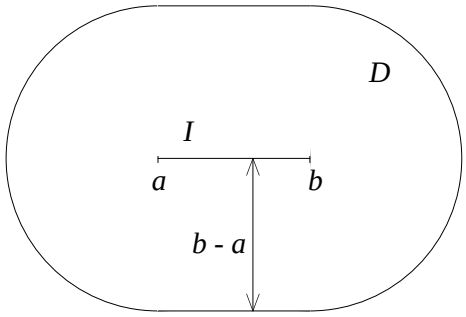


Fig. 5.7 - Regione D del piano complesso.

La condizione precedente è una condizione sufficiente, e quindi non si può dedurre che la successione $\{p_n(x)\}$ sia divergente se la funzione $f(x)$ ha un polo in D. Però se il polo è molto vicino ad I, accade spesso che la successione non sia convergente in qualche punto di I. Nel caso della funzione di Runge, per la presenza dei due punti singolari $\pm\boldsymbol{i}$ non si ha convergenza della successione dei polinomi di interpolazione per $|x| > 3.63\ldots$

Il risultato precedente non fa riferimento alla particolare distribuzione dei nodi nell'intervallo $[a, b]$. Esistono dei risultati più raffinati validi quando come nodi si scelgono i *punti di Chebyshev*

$$x_i = \frac{a+b}{2} - \frac{b-a}{2} \cos \frac{(2i+1)\pi}{2(n+1)}, \quad i = 0, \ldots, n.$$

In tal caso, se la funzione $f(z)$ nel campo complesso ha un numero finito di poli, esterni all'intervallo $[a, b]$, e soddisfa alla condizione

$$\lim_{|z|\to\infty} \frac{f(z)}{z^k} = 0 \quad \text{per qualche } k$$

(tutte le funzioni razionali il cui denominatore non si annulla in $[a, b]$ verificano queste condizioni), allora la successione dei polinomi $\{p_n(x)\}$ converge uniformemente ad $f(x)$ in $[a, b]$. Infatti nel caso della funzione di Runge dell'esempio 5.9, la successione $\{p_n(x)\}$ che si ottiene scegliendo come nodi i punti di Chebyshev converge uniformemente nell'intervallo $[-5, 5]$ (si veda l'esempio 6.44).

La scelta dei punti di Chebyshev come nodi dell'interpolazione è quella che minimizza la quantità $\pi_n(x)$ (si veda il teorema 6.19) e quindi in generale, se la $f^{(n+1)}(x)$ non varia molto in $[a, b]$, dà un resto minore rispetto alla scelta dei nodi equidistanti.

4. Polinomi osculatori

Una generalizzazione del polinomio di interpolazione è rappresentata dai *polinomi osculatori*, cioè polinomi $p(x)$ che nei nodi x_i, $i = 0, \ldots, n$, soddisfano a condizioni più generali delle (1), che coinvolgono anche valori delle derivate fino ad un ordine prefissato, nel modo seguente:

$$p^{(k)}(x_i) = f^{(k)}(x_i), \quad (i,k) \in K_n, \tag{15}$$

dove K_n è un sottoinsieme di $\{0, \ldots, n\} \times \{0, \ldots, n\}$.

Il modo più semplice per risolvere un tale problema è quello di considerare un polinomio di grado opportuno (di solito il numero dei coefficienti da determinare è uguale al numero delle condizioni date) e imporre che soddisfi le condizioni (15). I coefficienti del polinomio risultano allora soluzione di un sistema lineare.

5.10 Esempio. Per determinare il polinomio $p(x)$ di grado minimo che soddisfa alle seguenti condizioni

$$x_0 = 0, \quad x_1 = 1, \quad p(x_0) = 0, \quad p(x_1) = 1, \quad p'(x_1) = 0, \quad p''(x_0) = -1,$$

si considera un generico polinomio di terzo grado (le condizioni da imporre sono quattro)

$$p_3(x) = a_3x^3 + a_2x^2 + a_1x + a_0;$$

i coefficienti risultano soluzione del sistema lineare

$$\begin{aligned} p_3(0) &= a_0 = 0 \\ p_3(1) &= a_3 + a_2 + a_1 + a_0 = 1 \\ p_3'(1) &= 3a_3 + 2a_2 + a_1 = 0 \\ p_3''(0) &= 2a_2 = -1, \end{aligned}$$

da cui si ha $a_3 = -1/4$, $a_2 = -1/2$, $a_1 = 7/4$, $a_0 = 0$. Il polinomio cercato è quindi dato da

$$p_3(x) = -\frac{1}{4}\,(x^3 + 2x^2 - 7x). \qquad \blacksquare$$

A seconda delle condizioni imposte, il problema può avere o meno soluzione e tale soluzione può essere o no unica.

5.11 Esempio. I coefficienti del polinomio

$$p_2(x) = a_2x^2 + a_1x + a_0$$

che soddisfa alle condizioni

$$p_2(x_0) = f(x_0), \quad p_2(x_2) = f(x_2), \quad p_2'(x_1) = f'(x_1),$$

in cui $f(x_0)$, $f(x_2)$, $f'(x_1)$ sono valori dati e $x_0 < x_1 < x_2$, sono soluzione del sistema lineare

$$\begin{bmatrix} x_0^2 & x_0 & 1 \\ x_2^2 & x_2 & 1 \\ 2x_1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_2) \\ f'(x_1) \end{bmatrix}.$$

Tale sistema ha una e una sola soluzione se $x_1 \neq \dfrac{x_0 + x_2}{2}$. Se è $x_1 = \dfrac{x_0 + x_2}{2}$, il sistema ha infinite soluzioni se $f'(x_1) = \dfrac{f(x_2) - f(x_0)}{x_2 - x_0}$, non ne ha alcuna se $f'(x_1) \neq \dfrac{f(x_2) - f(x_0)}{x_2 - x_0}$. ∎

Un caso particolarmente importante di polinomio osculatore è quello del polinomio di Hermite : dati $n+1$ nodi x_i, $i = 0, \ldots, n$, il polinomio osculatore di Hermite è un polinomio $p(x)$ di grado al più $2n+1$ tale che

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad i = 0, \ldots, n. \tag{16}$$

Analogamente al polinomio di Lagrange, il polinomio di Hermite può essere rappresentato nella forma

$$p(x) = \sum_{j=0}^{n} U_j(x) f(x_j) + \sum_{j=0}^{n} V_j(x) f'(x_j), \tag{17}$$

dove i polinomi $U_j(x)$ e $V_j(x)$ sono funzioni dei polinomi $L_j(x)$ definiti in (4)

$$U_j(x) = [1 - 2L_j'(x_j)(x - x_j)]\, L_j^2(x),$$
$$V_j(x) = (x - x_j)\, L_j^2(x).$$

Come si può facilmente verificare,

a) i polinomi $U_j(x)$ e $V_j(x)$ hanno grado $2n+1$;

b) valgono le relazioni:

$$U_j(x_k) = \begin{cases} 1 & \text{se } k = j, \\ 0 & \text{se } k \neq j, \end{cases}$$
$$V_j(x_k) = 0 \quad \text{per ogni } k,$$
$$U_j'(x_k) = 0 \quad \text{per ogni } k,$$
$$V_j'(x_k) = \begin{cases} 1 & \text{se } k = j, \\ 0 & \text{se } k \neq j. \end{cases}$$

Ne segue che il polinomio (17) ha grado minore od uguale a $2n+1$ e soddisfa le condizioni (16).

5.12 Teorema. *Siano*

$$a = \min_{i=0,\dots,n} x_i, \quad b = \max_{i=0,\dots,n} x_i$$

e sia $f(x) \in C^{2n+2}[a,b]$. *Allora esiste un punto* $\xi = \xi(x) \in [a,b]$ *tale che il resto* $r(x)$ *del polinomio di Hermite è dato da*

$$r(x) = \pi_n^2(x)\, \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\,,$$

dove $\pi_n(x)$ *è il polinomio di grado* $n+1$ *definito in (6).*

Dim. La dimostrazione segue la traccia di quella del teorema 5.5. Per $x \neq x_i$, $i = 0, 1, \dots, n$, si considerano le funzioni

$$s(x) = \frac{r(x)}{\pi_n^2(x)}$$

e

$$z(y) = r(y) - s(x)\pi_n^2(y), \quad y \in [a,b].$$

La funzione $z(y)$ si annulla nei punti x_i, $i = 0, \dots, n$, e nel punto x. Per il teorema di Rolle quindi $z'(y)$ si annulla in $n+1$ punti distinti, diversi dagli x_i e x. Però

$$z'(y) = r'(y) - 2s(x)\pi_n(y)\pi_n'(y),$$

e poiché

$$r'(x_i) = f'(x_i) - p'(x_i) = 0, \quad \text{per} \quad i = 0, \dots, n,$$

$z'(y)$ si annulla anche nei punti x_i, $i = 0, \dots, n$. Complessivamente quindi $z'(y)$ si annulla in $2n+3$ punti distinti. La parte rimanente della dimostrazione è analoga a quella del teorema 5.5. ∎

5.13 Esempio. Il polinomio $p_3(x)$ di Hermite di grado al più 3 che nei punti x_0 e x_1, con $x_0 < x_1$, soddisfa alle condizioni

$$\begin{aligned} p_3(x_0) &= f(x_0), \quad p_3(x_1) = f(x_1), \\ p_3'(x_0) &= f'(x_0), \quad p_3'(x_1) = f'(x_1), \end{aligned}$$

per la (17) è dato da

$$p_3(x) = U_0(x)f(x_0) + U_1(x)f(x_1) + V_0(x)f'(x_0) + V_1(x)f'(x_1),$$

dove, posto $h = x_1 - x_0$, è

$$\begin{aligned}
U_0(x) &= \frac{1}{h^2}\,\Big(1 + 2\,\frac{x - x_0}{h}\Big)(x - x_1)^2,\\
U_1(x) &= \frac{1}{h^2}\,\Big(1 - 2\,\frac{x - x_1}{h}\Big)(x - x_0)^2,\\
V_0(x) &= \frac{1}{h^2}\,(x - x_0)(x - x_1)^2,\\
V_1(x) &= \frac{1}{h^2}\,(x - x_1)(x - x_0)^2.
\end{aligned}$$

Il resto per il teorema 5.12 è dato da

$$r(x) = \frac{1}{4!}\,(x - x_0)^2(x - x_1)^2 f^{(4)}(\xi), \quad \xi \in (x_0, x_1).$$

Per il caso particolare della funzione $f(x) = \sqrt{x}$, si considera il polinomio di Hermite che nei punti $x_0 = 0.49$ e $x_1 = 0.64$ soddisfa alle condizioni

$$\begin{aligned}
&p_3(0.49) = \sqrt{0.49} = 0.7, \quad p_3(0.64) = \sqrt{0.64} = 0.8,\\
&p_3'(0.49) = \frac{1}{2\sqrt{0.49}} = \frac{1}{1.4}\,, \quad p_3'(0.64) = \frac{1}{2\sqrt{0.64}} = \frac{1}{1.6}\,.
\end{aligned}$$

Tale polinomio è dato da

$$\begin{aligned}
p_3(x) = \frac{1}{0.0225}\,\Big[&0.7\Big(1 + 2\,\frac{x - 0.49}{0.15}\Big)(x - 0.64)^2\\
&+ 0.8\Big(1 - 2\,\frac{x - 0.64}{0.15}\Big)(x - 0.49)^2\\
&+ (x - 0.49)(x - 0.64)\Big(\frac{x - 0.64}{1.4} + \frac{x - 0.49}{1.6}\Big)\Big]\\
= {}& 0.2645503\,x^3 - 0.7460317\,x^2 + 1.254841\,x + 0.2331259.
\end{aligned}$$

Nel punto $x = 0.6$ si ottiene il valore $p_3(0.6) = 0.774602$. Per $x \in [0.49, 0.64]$ il resto è dato da

$$r(x) = -(x - 0.49)^2(x - 0.64)^2\,\frac{15}{4!\;16\sqrt{\xi^7}}\,,$$

in cui $\xi \in (0.49, 0.64)$. Si ha perciò

$$|r(0.6)| < 0.918\;10^{-5}.$$

In realtà risulta

$$\max_{x\in[0.49,0.64]} |\sqrt{x} - p_3(x)| \approx 0.914\ 10^{-5}$$

e

$$|\sqrt{0.6} - p_3(0.6)| \approx 0.528\ 10^{-5}.$$

Si confronti il valore qui ottenuto con il valore 0.7746606 ottenuto nell'esempio 5.7 con il polinomio di interpolazione di terzo grado della stessa funzione, assumendo come nodi i quattro punti

$$x_0 = 0.36, \quad x_1 = 0.49, \quad x_2 = 0.64, \quad x_3 = 0.81,$$

e che risulta affetto da un errore di $0.639\ 10^{-4}$. ∎

5. Polinomio di Newton

Un altro modo per rappresentare il polinomio di interpolazione è fornito dal *polinomio di Newton* che consente di calcolare, fra l'altro, il valore di $p_n(x)$ in un punto con un numero di operazioni moltiplicative inferiore a quello richiesto dal polinomio di Lagrange, richiedendo però un corrispondente aumento delle operazioni additive. Inoltre il calcolo di $p_n(x)$ con il polinomio di Newton è in certi casi più stabile del calcolo effettuato con il polinomio di Lagrange. Per rappresentare il polinomio di Newton è necessario introdurre le differenze divise. Si suppone che gli $n+1$ nodi x_i, $i = 0, \ldots, n$, siano distinti.

5.14 Definizione. Si chiama *differenza divisa* di ordine k della funzione $f(x)$ relativa ai punti $x_0, \ldots, x_{k-1}$, la funzione $[x_0, x_1, \ldots, x_{k-1}, x]$ definita per $x \neq x_i$, $i = 0, \ldots, k-1$, ricorsivamente nel modo seguente

per $k = 0$ $\quad f[x] = f(x)$

per $k = 1$ $\quad f[x_0, x] = \dfrac{f[x] - f[x_0]}{x - x_0}$

per $k \geq 2$

$$f[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] = \frac{f[x_0, x_1, \ldots, x_{k-2}, x] - f[x_0, x_1, \ldots, x_{k-2}, x_{k-1}]}{x - x_{k-1}}. \tag{18}$$

∎

Per il calcolo delle differenze divise di ordine k conviene utilizzare la seguente tabella

$$
\begin{array}{c|cccccc}
x_0 & f[x_0] & & & & & \\
 & & f[x_0,x_1] & & & & \\
x_1 & f[x_1] & & f[x_0,x_1,x_2] & & & \\
 & & f[x_0,x_2] & & f[x_0,x_1,x_2,x_3] & & \\
x_2 & f[x_2] & & f[x_0,x_1,x_3] & & \ddots & \\
 & & f[x_0,x_3] & & \vdots & & f[x_0,x_1,\ldots,x_n] \\
x_3 & f[x_3] & & \vdots & & \iddots & \\
 & & \vdots & & f[x_0,x_1,x_2,x_n] & & \\
\vdots & \vdots & & f[x_0,x_1,x_n] & & & \\
 & & f[x_0,x_n] & & & & \\
x_n & f[x_n] & & & & &
\end{array}
$$

Gli elementi della tabella, considerati come elementi di una matrice A, triangolare inferiore di ordine $n+1$, vengono generati mediante le relazioni

$$
\begin{aligned}
a_{i1} &= f(x_{i-1}), \quad i = 1,\ldots,n+1,\\
a_{ij} &= \frac{a_{i,j-1} - a_{j-1,j-1}}{x_{i-1} - x_{j-2}}, \quad j = 2,\ldots,n+1, \quad i = j,\ldots,n+1.
\end{aligned}
$$

Poiché per ogni elemento sono richieste due sottrazioni e una divisione, la costruzione della tabella richiede, a meno di termini di ordine inferiore, n^2 addizioni e $n^2/2$ moltiplicazioni.

5.15 Teorema. *Per la funzione $f(x)$ vale la seguente relazione*

$$
\begin{aligned}
f(x) =& f[x_0] + (x-x_0)f[x_0,x_1] + (x-x_0)(x-x_1)f[x_0,x_1,x_2] \\
&+ \ldots + (x-x_0)(x-x_1)\ldots(x-x_{n-1})f[x_0,x_1,\ldots,x_n] \\
&+ (x-x_0)(x-x_1)\ldots(x-x_{n-1})(x-x_n)f[x_0,x_1,\ldots,x_n,x].
\end{aligned} \tag{19}
$$

Dim. Si procede per induzione. Per $n = 0$ si ha

$$
f(x) = f[x_0] + (x-x_0)f[x_0,x],
$$

come segue direttamente dalla definizione 5.14. Per $n > 0$, si supponga che la (19) valga fino all'indice $n-1$, cioè

$$
\begin{aligned}
f(x) &= f[x_0] + (x-x_0)f[x_0,x_1] + (x-x_0)(x-x_1)f[x_0,x_1,x_2] \\
&+ \ldots + (x-x_0)(x-x_1)\ldots(x-x_{n-2})(x-x_{n-1})f[x_0,x_1,\ldots,x_{n-1},x].
\end{aligned} \tag{20}
$$

Per la (18) è

$$f[x_0, x_1, \ldots, x_{n-1}, x] = f[x_0, x_1, \ldots, x_{n-1}, x_n] + (x - x_n) f[x_0, x_1, \ldots, x_n, x],$$

e sostituendo nella (20), ne segue la (19). ∎

Le differenze divise consentono di rappresentare in modo molto semplice il polinomio di interpolazione.

5.16 Teorema. *Il polinomio*

$$\begin{aligned} p_n(x) = & f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ & + \ldots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f[x_0, x_1, \ldots, x_n] \end{aligned} \tag{21}$$

di grado al più n, è tale che $p_n(x_i) = f(x_i)$, per $i = 0, \ldots, n$, e pertanto è il polinomio di interpolazione. Il polinomio (21) è detto polinomio di Newton.

Dim. Si procede per induzione su n. Per $n = 0$ è $p_0(x) = f[x_0]$, e quindi $p_0(x_0) = f(x_0)$. Per $n > 0$ è

$$p_n(x) = p_{n-1}(x) + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f[x_0, x_1, \ldots, x_n].$$

Si suppone che $p_{n-1}(x_i) = f(x_i)$, per $i = 0, 1, \ldots, n-1$, e si dimostra che $p_n(x_i) = f(x_i)$, per $i = 0, 1, \ldots, n$. Infatti per $i = 0, 1, \ldots, n-1$ è

$$\begin{aligned} p_n(x_i) &= p_{n-1}(x_i) + (x_i - x_0)(x_i - x_1) \ldots (x_i - x_{n-1}) f[x_0, x_1, \ldots, x_n] \\ &= p_{n-1}(x_i) = f(x_i). \end{aligned}$$

Per $i = n$ si ha dalla (21)

$$\begin{aligned} p_n(x_n) = & f[x_0] + (x_n - x_0) f[x_0, x_1] + (x_n - x_0)(x_n - x_1) f[x_0, x_1, x_2] \\ & + \ldots + (x_n - x_0)(x_n - x_1) \ldots (x_n - x_{n-1}) f[x_0, x_1, \ldots, x_n] \end{aligned}$$

e per la (20) risulta $p_n(x_n) = f(x_n)$. ∎

5.17 Esempio. Il polinomio di Newton per l'interpolazione lineare è

$$p_1(x) = f(x_0) + (x - x_0) \, \frac{f(x_1) - f(x_0)}{x_1 - x_0} \, . \qquad ∎$$

5.18 Esempio. Della funzione $f(x) = \dfrac{\tan x}{x}$ sono noti i seguenti valori

x	0.3	0.6	1.0	1.4
$\dfrac{\tan x}{x}$	1.031120	1.140227	1.557407	4.141337

Per calcolare un'approssimazione di $f(1.2)$, si costruisce la tabella delle differenze divise

0.3	1.031120			
		0.3636900		
0.6	1.140227		0.9703712	
		0.7518385		5.273385
1.0	1.557407		3.079725	
		2.827470		
1.4	4.141337			

Il corrispondente polinomio di Newton è dato da:

$$\begin{aligned} p_3(x) =& 1.03112 + 0.36369\,(x - 0.3) + 0.9703712\,(x - 0.3)\,(x - 0.6) \\ &+ 5.273385\,(x - 0.3)\,(x - 0.6)\,(x - 1) \end{aligned}$$

e si ha $p_3(1.2) = 2.451967$. Nella figura 5.8 sono illustrati l'andamento della funzione $f(x)$ (con linea spessa) e del polinomio $p_3(x)$ di interpolazione (linea più sottile) nell'intervallo [0.3, 1.4]. ∎

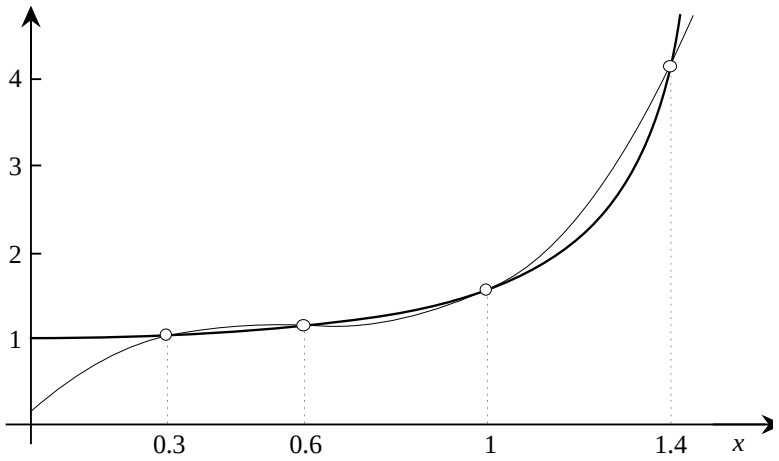


Fig. 5.8 - Polinomio di interpolazione della funzione $f(x) = \tan x/x$.

Il polinomio (21) può essere posto nella forma

$$\begin{aligned} p_n(x) =& \{\dots\{f[x_0, x_1, \dots, x_n](x - x_{n-1}) + f[x_0, x_1, \dots, x_{n-1}]\}(x - x_{n-2}) \\ &+ \dots + f[x_0, x_1]\}(x - x_0) + f[x_0], \end{aligned}$$

che consente il calcolo di $p_n(x)$, con una tecnica analoga a quella del metodo di Ruffini-Horner

$$\begin{aligned} &h_n = f[x_0, x_1, \dots, x_n], \\ &h_i = h_{i+1}(x - x_i) + f[x_0, x_1, \dots, x_i], \quad \text{per } i = n-1, \dots, 1, \\ &p_n(x) = h_1(x - x_0) + f[x_0]. \end{aligned}$$

Le operazioni richieste per il calcolo del polinomio di interpolazione in un punto x, una volta che sia stata costruita la tabella delle differenze divise, sono quindi $2n$ addizioni e n moltiplicazioni.

Una volta valutato in un punto il polinomio di Newton di interpolazione sui nodi $(x_i, f(x_i))$, $i = 0, \dots, n$, è possibile, utilizzando i calcoli già fatti, valutare nello stesso punto il polinomio di Newton di interpolazione sui nodi $(x_i, f(x_i))$, $i = 0, \dots, n+1$. Infatti la generazione di un'altra riga della tabella delle differenze divise richiede $2n$ addizioni e n moltiplicazioni e quindi il calcolo del nuovo polinomio di interpolazione nello stesso punto richiede solo $2n$ moltiplicazioni e $2n$ addizioni aggiuntive.

Nella seguente tabella, in cui con A si sono indicate le "addizioni" e con M le "moltiplicazioni", sono messi a confronto i costi computazionali relativi ai polinomi di Newton e di Lagrange.

	polinomio di Lagrange	polinomio di Newton
calcolo in un punto	$\frac{n^2}{2}A + n^2M$	$n^2A + \frac{n^2}{2}M$
calcolo in m punti	$\left(\frac{n^2}{2} + 2mn\right)A$ $+(n^2 + 2mn)M$	$(n^2 + 2mn)A$ $+\left(\frac{n^2}{2} + mn\right)M$
aggiunta di un nodo	$2nA + 2nM$	$2nA + 2nM$

Dai teoremi 5.15 e 5.16 discendono interessanti proprietà delle differenze divise.

5.19 Teorema. *Per $x \neq x_i$, $i = 0, 1, \ldots, n$, è*

$$r(x) = f(x) - p_n(x) = \pi_n(x) f[x_0, x_1, \ldots, x_n, x], \tag{22}$$

dove $\pi_n(x)$ è il polinomio di grado $n+1$ definito in (6), e quindi se $f(x) \in C^{n+1}[a, b]$ è

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{23}$$

dove $\xi = \xi(x) \in (a, b)$.

Dim. La (22) si ottiene sottraendo la (21) dalla (19), la (23) si ottiene per confronto con la (11). ∎

La relazione (22) fornisce un'espressione del resto del polinomio di interpolazione senza ipotesi di regolarità della funzione $f(x)$. Invece la (11), da cui deriva la (23), è stata ricavata nel teorema 5.5 sotto l'ipotesi di derivabilità fino all'ordine $n+1$ della funzione $f(x)$.

5.20 Teorema. *La differenza divisa di ordine k*

$$f[x_0, x_1, \ldots, x_k]$$

è funzione simmetrica dei suoi argomenti $x_0, x_1, \ldots, x_k$, cioè è invariante comunque vengano permutati i suoi argomenti.

Dim. Si consideri il polinomio di interpolazione $p_n(x)$ tale che $p_n(x_i) = f(x_i)$, $i = 0, \ldots, k$. Dalla (21) risulta che il coefficiente del termine di grado più elevato di tale polinomio è $f[x_0, x_1, \ldots, x_k]$. Poiché il polinomio di interpolazione non dipende dall'ordine dei nodi, il coefficiente del termine di grado più elevato è invariante comunque vengano permutati i nodi. ∎

Quando i punti x_i sono equidistanti di passo $h > 0$, vi è una stretta relazione fra le differenze divise e le differenze finite, studiate nel capitolo 4 e così definite

$$\Delta f(x_i) = f(x_{i+1}) - f(x_i), \quad i = 0, \ldots, n-1,$$
$$\Delta^k f(x_i) = \Delta^{k-1} f(x_{i+1}) - \Delta^{k-1} f(x_i), \ i = 0, \ldots, n-k, \ k = 1, \ldots, n.$$

5.21 Teorema. *Se i punti x_i, $i = 0, 1, \ldots, n$, sono equidistanti di passo h, allora*

$$\Delta^k f(x_i) = k! h^k f[x_i, x_{i+1}, \ldots, x_{i+k}], \quad i = 0, \ldots, n-k, \quad k = 1, \ldots, n. \tag{24}$$

Dim. Si procede per induzione su k. Per $k = 1$ si ha

$$\Delta f(x_i) = f(x_{i+1}) - f(x_i) = hf[x_i, x_{i+1}], \quad i = 0, \ldots, n-1.$$

Per $k > 1$ si ha

$$\begin{aligned}\Delta^{k+1} f(x_i) &= \Delta^k f(x_{i+1}) - \Delta^k f(x_i) \\ &= k! h^k (f[x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}] - f[x_i, x_{i+1}, \ldots, x_{i+k}]);\end{aligned}$$

per il teorema 5.20 nella seconda differenza divisa si possono permutare gli argomenti, per cui

$$\begin{aligned}\Delta^{k+1} f(x_i) &= k! h^k (f[x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}] - f[x_{i+1}, \ldots, x_{i+k}, x_i]) \\ &= k! h^k (x_{i+k+1} - x_i) f[x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}, x_i].\end{aligned}$$

Applicando di nuovo il teorema 5.20 si ha

$$\Delta^{k+1} f(x_i) = k! h^k (k+1) h f[x_i, x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}],$$

da cui segue la tesi. ∎

Con il cambiamento di variabile (9), sostituendo le (24) nel polinomio (21), si ottiene il *polinomio di Newton con le differenze finite* per l'interpolazione nel caso di punti equidistanti

$$\begin{aligned}p_n(x_0 + th) = f(x_0) + t\Delta f(x_0) + \frac{t(t-1)}{2}\, \Delta^2 f(x_0) \\ + \ldots + \frac{t(t-1)\ldots(t-n+1)}{n!}\, \Delta^n f(x_0)\end{aligned}$$

e utilizzando il polinomio $\tau_k(t)$ definito in (10), si ha

$$p_n(x_0 + th) = f(x_0) + \sum_{k=1}^{n} \frac{\tau_{k-1}(t)}{k!}\, \Delta^k f(x_0). \tag{25}$$

Con la convenzione

$$\binom{t}{k} = \frac{\tau_{k-1}(t)}{k!}$$

anche per t non intero, la (25) può essere scritta nella forma

$$p_n(x_0 + th) = \sum_{k=0}^{n} \binom{t}{k}\, \Delta^k f(x_0), \tag{26}$$

e se $f(x) \in C^{n+1}[a,b]$ il resto (22) risulta

$$r(x_0+th) = \binom{t}{n+1} h^{n+1} f^{(n+1)}(\xi), \quad \xi \in (x_0, x_n).$$

In pratica conviene calcolare una tabella delle differenze finite nel modo seguente

$$\begin{array}{lllll}
f(x_0) & & & & \\
 & \Delta f(x_0) & & & \\
f(x_1) & & \Delta^2 f(x_0) & & \\
 & \Delta f(x_1) & & \Delta^3 f(x_0) & \\
f(x_2) & & \Delta^2 f(x_1) & & \ddots \\
 & \Delta f(x_2) & & & \Delta^n f(x_0) \\
f(x_3) & & \vdots & & \iddots \\
 & \vdots & & \Delta^3 f(x_{n-3}) & \\
\vdots & & \Delta^2 f(x_{n-2}) & & \\
 & \Delta f(x_{n-1}) & & & \\
f(x_n) & & & &
\end{array}$$

in cui ogni elemento è dato dalla differenza dei due elementi contigui della colonna precedente, ed eseguire il calcolo della (25) con una tecnica simile a quella del metodo di Ruffini-Horner

$$\begin{aligned}
&d_n = \frac{\Delta^n f(x_0)}{n} \\
&d_i = \frac{d_{i+1}(t-i) + \Delta^i f(x_0)}{i}, \quad i = n-1, \ldots, 1, \\
&p_n(x_0+th) = d_1 t + f(x_0).
\end{aligned}$$

Le operazioni richieste per il calcolo del polinomio di interpolazione in un punto x, una volta che sia stata costruita la tabella delle differenze finite, sono quindi $2n$ addizioni e $2n$ moltiplicazioni. Tenendo conto anche delle operazioni richieste per la costruzione della tabella, il calcolo di $p_n(x_0+th)$ richiede $n^2/2$ addizioni e $2n$ moltiplicazioni.

5.22 Esempio. Si vuole costruire il polinomio di grado 4 che assume gli stessi valori di $f(x) = |x|$ nei punti $x_i = -1 + i/2$, $i = 0, 1, \ldots, 4$. Quindi è

$h = 0.5$. La tabella delle differenze finite è

$$\begin{array}{ccccc}
1 & & & & \\
 & -0.5 & & & \\
0.5 & & 0 & & \\
 & -0.5 & & 1 & \\
0 & & 1 & & -2 \\
 & 0.5 & & -1 & \\
0.5 & & 0 & & \\
 & 0.5 & & & \\
1 & & & &
\end{array}$$

per cui è

$$p_n\big(-1+\frac{t}{2}\big) = 1 - \frac{t}{2} + \frac{1}{6}\,t(t-1)(t-2) - \frac{1}{12}\,t(t-1)(t-2)(t-3).$$

Sostituendo $t = 2(x+1)$, si ha

$$p_n(x) = \frac{x^2}{3}\,(7-4x^2).$$

Nella figura 5.9 è illustrato il comportamento del polinomio di interpolazione $p_n(x)$ (linea più sottile) rispetto alla funzione $f(x)$ (linea spessa). Il resto è

$$r(x) = |x| - p_n(x) = |x| - \frac{x^2}{3}\,(7-4x^2)$$

e per la simmetria di $f(x)$ e di $p_n(x)$ risulta

$$\max_{x\in[-1,1]} |r(x)| = \max_{x\in[0,1]} \Big|x - \frac{x^2}{3}\,(7-4x^2)\Big| \approx 0.147. \qquad \blacksquare$$

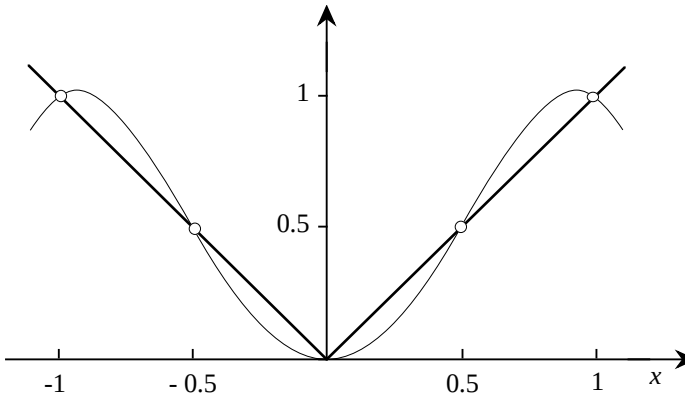


Fig. 5.9 - Polinomio di interpolazione della funzione $f(x) = |x|$.

6. Errori di arrotondamento del polinomio di interpolazione

Nell'esame dell'accuratezza dell'approssimazione di $f(x)$ con un polinomio di interpolazione $p_n(x)$, occorre tener conto, oltre che dell'errore analitico

$$\epsilon_{an} = -\frac{r(x)}{f(x)}, \quad f(x) \neq 0,$$

anche dell'errore algoritmico generato nel calcolo di $p_n(x)$, che può diventare preponderante rispetto all'errore analitico e che dipende dalla particolare espressione del polinomio di interpolazione che viene usata.

5.23 Esempio. Nelle figure 5.10 e 5.11 sono riportati i grafici dei moduli degli errori relativi effettivamente generati nel calcolo dei polinomi di interpolazione di Lagrange e di Newton per la funzione $f(x) = \log x$ nell'intervallo [10,11], assumendo come nodi i punti equidistanti $x_i = 10 + ih$, $h = 1/29$, $i = 0, \ldots, 29$.

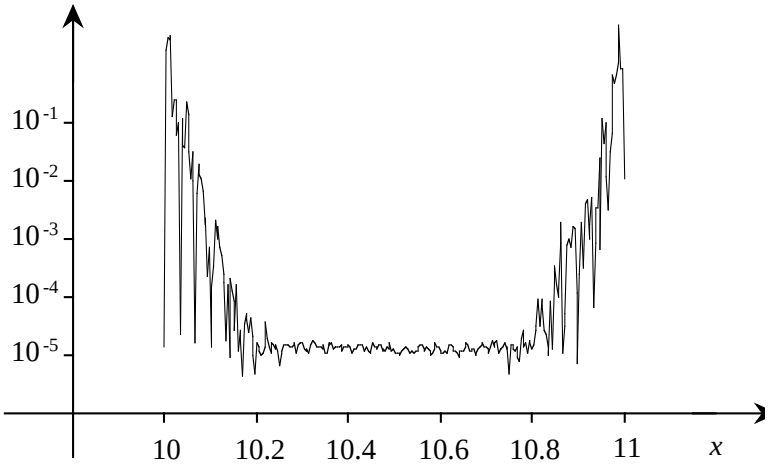


Fig. 5.10 - Errore relativo del polinomio di interpolazione di Lagrange di $f(x) = \log x$.

Nell'intervallo $[10, 11]$ l'errore analitico, per la (14), può essere così maggiorato

$$|\epsilon_{an}| = \frac{|r(x)|}{|f(x)|} \leq \frac{|b-a|^{n+1}}{\min\limits_{x\in[10,11]} \log x} \frac{M_{n+1}}{(n+1)!} = \frac{1}{(n+1)\log 10} \max_{x\in[10,11]} \frac{1}{x^{n+1}}$$
$$< \frac{10^{-(n+1)}}{2.3(n+1)},$$

e per $n = 29$ è $|\epsilon_{an}| < 10^{-31}$. Poiché tale quantità è molto minore della precisione di macchina, si può assumere che gli errori effettivamente generati siano sostanzialmente dovuti all'errore algoritmico. ∎

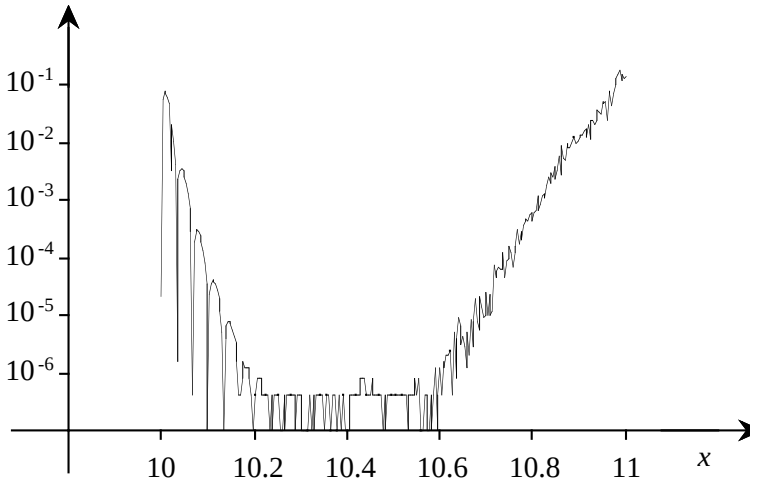


Fig. 5.11 - Errore relativo del polinomio di interpolazione di Newton di $f(x) = \log x$.

Per il polinomio di Lagrange l'errore algoritmico è essenzialmente prodotto dal calcolo della somma

$$p(x) = \sum_{j=0}^{n} \alpha_j, \quad \alpha_j = L_j(x) f(x_j),$$

quando i termini non sono tutti dello stesso segno e hanno modulo maggiore di quello del risultato, per cui si verifica un errore di cancellazione. Infatti gli errori relativi ϵ_j, presenti nei termini α_j, contribuiscono all'errore relativo della somma con la quantità

$$\epsilon_p = \frac{1}{p(x)} \sum_{j=0}^{n} \alpha_j \epsilon_j.$$

Posto $\epsilon = \max_{j=0,n} |\epsilon_j|$, si ha

$$|\epsilon_p| < \lambda_n(x)\, \epsilon \max_{x \in [a,b]} \left| \frac{f(x_j)}{p(x)} \right|, \quad \text{dove} \quad \lambda_n(x) = \sum_{j=0}^{n} |L_j(x)|.$$

Se $f(x)$ non varia molto nell'intervallo $[a, b]$, il fattore $\lambda_n(x)$ indica quanto influiscono gli errori ϵ_j sul risultato e quindi dà una misura di quanto può essere instabile il calcolo del polinomio di Lagrange nel punto x.

Se n è grande, nei punti vicini agli estremi dell'intervallo di interpolazione esistono valori di $\lambda_n(x)$ molto elevati, in particolare quando i nodi sono equidistanti. Nell'esempio 5.23, in cui la funzione $f(x) = \log x$ varia poco in $[10, 11]$, nel punto $x = 10.008$ l'errore effettivo è circa 0.25, e infatti in tale punto risulta $\lambda_n(x) \approx 10^6$, mentre nel punto $x = 10.52$, in cui $\lambda_n(x) \approx 1$, l'errore effettivo è circa $0.851\ 10^{-5}$.

Per il polinomio di Newton l'errore algoritmico è principalmente generato dalla propagazione degli errori nel calcolo della tabella delle differenze divise o finite (si veda l'esercizio 5.24 per la propagazione dell'errore assoluto in una tabella delle differenze finite). Se si considera l'errore relativo, la propagazione dell'errore risulta particolarmente elevata quando esiste una derivata di $f(x)$ che nell'intervallo è piccola rispetto ai valori della funzione (ciò accade in particolare quando $f(x)$ è ben approssimabile con un polinomio di grado inferiore ad n). Poiché per la (23) a tale derivata corrisponde un valore piccolo della tabella delle differenze divise (o finite se i punti sono equidistanti), che risulta calcolato mediante sottrazione a partire da valori più elevati della funzione, è possibile che nella costruzione della tabella si verifichino elevati errori di cancellazione.

Nel caso della funzione $f(x) = \log x$ dell'esempio 5.23, i valori effettivamente calcolati per $x_0 = 10$ sono

$$f(x_0) = 2.302585 \qquad \Delta f(x_0) = 3.442764\ 10^{-3}$$

$$\Delta^2 f(x_0) = -1.239777\ 10^{-5} \qquad \Delta^3 f(x_0) = 9.536743\ 10^{-7}.$$

Il valore calcolato per $\Delta^2 f(x_0)$ ha una sola cifra esatta, mentre il valore di $\Delta^3 f(x_0)$ non ne ha nessuna. Proseguendo nella costruzione della tabella, gli altri valori calcolati non hanno più alcuna relazione con quelli esatti. Risulta ad esempio $\Delta^{29} f(x_0) = 2.131082$, mentre il valore esatto è minore di 10^{-40}. Poiché nella (25) queste differenze sono moltiplicate per i polinomi $\tau_k(t)$, per valori di k grandi e di t vicini agli estremi dell'intervallo $[0, n]$, per cui i polinomi $\tau_k(t)$ hanno valori alti, i risultati che si ottengono sono affetti da errori elevati. Invece nella parte centrale dell'intervallo di interpolazione, in cui i polinomi $\tau_k(t)$ assumono valori praticamente nulli, l'influenza delle differenze di ordine alto è minima e l'errore è più contenuto.

L'ordinamento dei nodi può ovviamente infuenzare l'errore del risultato, come risulta anche dal seguente esempio.

5.24 Esempio. Nella figura 5.12 è riportato il grafico dei moduli degli errori relativi effettivamente generati nel calcolo del polinomio di interpolazione di Newton della funzione $f(x) = \log x$ nell'intervallo [10,11], sui nodi equidistanti $x_i = 11 + ih$, $h = -1/29, i = 0, \ldots, 29$. Il polinomio è quindi lo stesso di quello a cui si riferiscono le figure 5.10 e 5.11.

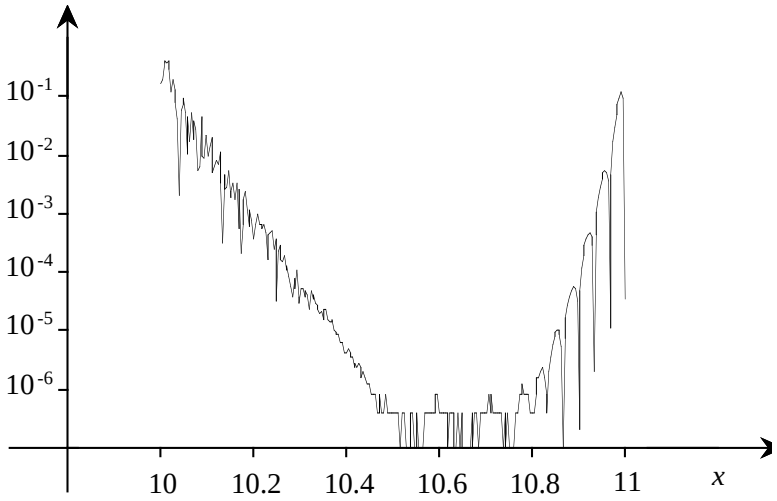


Fig. 5.12 - Errore relativo del polinomio di interpolazione di Newton di $f(x) = \log x$ con nodi in ordine decrescente.

Confrontando il grafico di figura 5.12 con quello di figura 5.11 (entrambi si riferiscono al polinomio di Newton relativo agli stessi nodi, ma calcolato con due ordinamenti diversi), si nota come l'errore algoritmico si accumuli di più nella parte destra dell'intervallo quando i nodi sono presi in ordine crescente, nella parte sinistra quando i nodi sono presi in ordine decrescente. ∎

In generale nei punti vicini agli estremi dell'intervallo l'errore algoritmico cresce con il grado del polinomio, in quanto crescono i gradi dei polinomi $L_j(x)$ o dei polinomi $\tau_j(x)$ e quindi i loro massimi. Per il polinomio di Lagrange si definisce la quantità

$$\Lambda_n = \max_{x \in [a,b]} \lambda_n(x),$$

detta *costante di Lebesgue* di ordine n, che dipende da come sono distribuiti i nodi in $[a, b]$, e che misura la instabilità numerica sull'intervallo $[a, b]$ del calcolo del polinomio di Lagrange costruito sui nodi $x_0, \ldots, x_n$. La crescita di Λ_n è stata studiata per diverse scelte dei nodi: se i nodi sono equidistanti la costante Λ_n cresce in maniera esponenziale (si veda l'esercizio 5.22), mentre se i nodi sono i punti di Chebyshev in $[a, b]$ la costante Λ_n cresce solo in maniera logaritmica [18]. Nella zona centrale dell'intervallo di interpolazione questo fenomeno si presenta in maniera molto più ridotta.

5.25 Esempio. Nella figura 5.13 sono riportati i grafici dei moduli degli errori relativi effettivamente generati nel calcolo dei polinomi di interpolazione

su $n+1$ nodi equidistanti nell'intervallo $[2,3]$ della funzione $f(x) = \log x$ nel punto 2.008, al crescere di n. I pallini individuano gli errori del polinomio di Newton, i quadratini neri individuano gli errori del polinomio di Lagrange. Si noti come al crescere di n gli errori inizialmente decrescano, mettendo in evidenza la maggiore influenza dell'errore analitico, che decresce al crescere di n, e successivamente crescano, con il crescere dell'errore algoritmico, che diventa preponderante rispetto all'errore analitico. In generale l'errore relativo del polinomio di Newton risulta inferiore a quello di Lagrange.

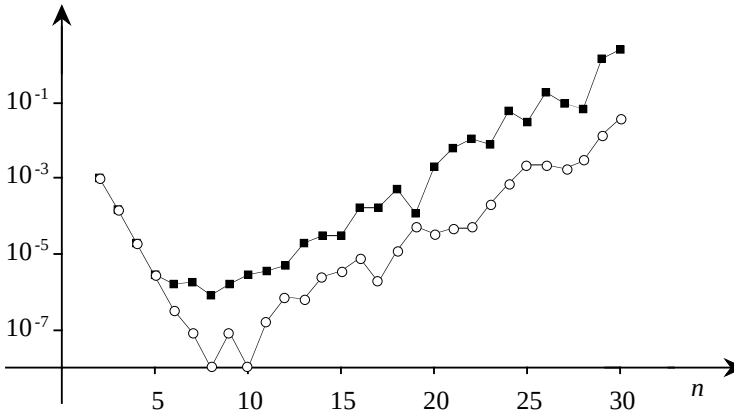


Fig. 5.13 - Errori relativi dei polinomi di interpolazione di Newton e di Lagrange di $f(x) = \log x$ in $x = 2.008$.

Nella figura 5.14 sono riportati gli analoghi grafici nel punto $x = 2.52$ (nella parte centrale dell'intervallo di interpolazione). Si noti come anche in questo caso il polinomio di Newton sia affetto da un errore algoritmico minore di quello del polinomio di Lagrange. ∎

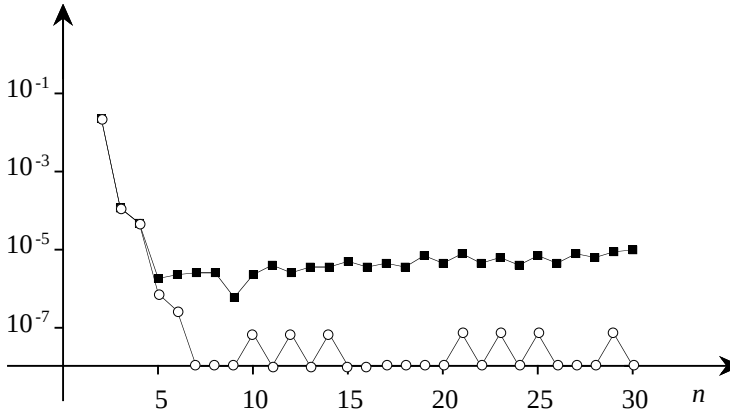


Fig. 5.14 - Errori relativi dei polinomi di interpolazione di Newton e di Lagrange di $f(x) = \log x$ in $x = 2.52$.

Poiché l'errore algoritmico tende a crescere con n, non è conveniente scegliere polinomi di grado troppo elevato. Una tecnica facilmente applicabile con il polinomio di Newton, consiste nell'arrestare la costruzione della tabella delle differenze divise o finite quando tutte le differenze di una colonna sono piccole, dell'ordine della precisione di macchina. Se ciò accade per le differenze di ordine k, si utilizza il polinomio di interpolazione di grado $k-1$.

Questo procedimento non è applicabile con il polinomio di Lagrange, in cui la propagazione dell'errore può però essere controllata tramite il fattore $\lambda_n(x)$. Comunque, nel caso di nodi equidistanti, se è possibile scegliere l'intervallo di interpolazione, è opportuno farlo in modo che il punto x venga a trovarsi nella parte centrale dell'intervallo.

7. Proprietà delle differenze divise

5.26 Teorema. *Sia*

$$f(x) = a_k x^k + a_{k-1}x^{k-1} + \ldots + a_0$$

un polinomio di grado k. Per $n \leq k$ la differenza divisa $f[x_0, \ldots, x_{n-1}, x]$ di ordine n è un polinomio di grado $k-n$, avente a_k come primo coefficiente, e per $n > k$ è $f[x_0, \ldots, x_{n-1}, x] = 0$.

Dim. Per $n \leq k$ si procede per induzione su n. Se $n = 0$ la tesi segue subito in quanto $f[x] = f(x)$. Se $n > 0$ per l'ipotesi induttiva la differenza divisa di ordine $n-1$ è un polinomio di grado $k-n+1$, con primo coefficiente a_k e può essere così rappresentato

$$f[x_0, \ldots, x_{n-2}, x] = \sum_{j=0}^{k-n+1} b_j x^j, \quad b_{k-n+1} = a_k,$$

da cui

$$f[x_0, \ldots, x_{n-1}, x] = \frac{f[x_0, \ldots, x_{n-2}, x] - f[x_0, \ldots, x_{n-2}, x_{n-1}]}{x - x_{n-1}}$$
$$= \frac{1}{x - x_{n-1}} \sum_{j=0}^{k-n+1} b_j (x^j - x_{n-1}^j) = \sum_{j=1}^{k-n+1} b_j \sum_{i=0}^{j-1} x^i x_{n-1}^{j-i-1} = \sum_{j=0}^{k-n} c_j x^j,$$

dove $c_{k-n} = b_{k-n+1} = a_k$. Quindi $f[x_0, \ldots, x_{n-1}, x]$ è un polinomio di grado $k-n$, con primo coefficiente a_k. Poiché per $k = n$ risulta

$$f[x_0, \ldots, x_{n-1}, x] = a_k,$$

ne segue che la differenza divisa di ordine maggiore di k è nulla. ∎

I seguenti teoremi forniscono due rappresentazioni diverse delle differenze divise di ordine n, mediante le differenze del primo ordine e mediante integrali.

5.27 Teorema. *Per $x \neq x_i$, $i = 0, \ldots, n$, è*

$$f[x_0, x_1, \ldots, x_n, x] = \sum_{i=0}^{n} \frac{f[x_i, x]}{\pi_n'(x_i)}\,.$$

Dim. Sostituendo nella (22) l'espressione del polinomio di interpolazione di Lagrange e utilizzando la (8) si ha:

$$f(x) - \pi_n(x) \sum_{i=0}^{n} \frac{f(x_i)}{(x - x_i)\pi_n'(x_i)} = \pi_n(x) f[x_0, x_1, \ldots, x_n, x]. \tag{27}$$

Poiché

$$\pi_n(x) \sum_{i=0}^{n} \frac{1}{(x - x_i)\pi_n'(x_i)} = 1,$$

(si veda l'esercizio 5.9 b)), dalla (27) si ottiene

$$\pi_n(x) \sum_{i=0}^{n} \frac{f(x) - f(x_i)}{(x - x_i)\pi_n'(x_i)} = \pi_n(x) f[x_0, x_1, \ldots, x_n, x],$$

da cui la tesi. ∎

5.28 Teorema. *Sia $f(x) \in C^n[a, b]$, si consideri la funzione*

$$g_n(x_0, x_1, \ldots, x_n)$$
$$= \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} f^{(n)}\big((x_n - x_{n-1})t_n + \ldots + (x_1 - x_0)t_1 + x_0\big)\, dt_n.$$

Allora per tutti i punti $(x_0, \ldots, x_n)$ tali che $x_i \neq x_j$ per $i \neq j$, vale

$$g_n(x_0, x_1, \ldots, x_n) = f[x_0, x_1, \ldots, x_n].$$

Cioè la funzione $g_n(x_0, x_1, \ldots, x_n)$, che è continua anche nel caso che i punti x_i non siano tutti distinti, è quella che estende per continuità la $f[x_0, x_1, \ldots, x_n]$ su tutto $[a, b]^{n+1}$.

Dim. Si procede per induzione su n. Per $n = 1$ è

$$g_1(x_0, x_1) = \int_0^1 f'\big((x_1 - x_0)t_1 + x_0\big)\, dt_1.$$

Ponendo $s = (x_1 - x_0)t_1 + x_0$, si ha

$$g_1(x_0, x_1) = \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} f'(s)\, ds = f[x_0, x_1].$$

In generale, se $g_{n-1}(x_0, x_1, \ldots, x_{n-1}) = f[x_0, x_1, \ldots, x_{n-1}]$, posto

$$s = (x_n - x_{n-1})t_n + \ldots + (x_1 - x_0)t_1 + x_0,$$

si ha

$$\begin{aligned}&\int_0^{t_{n-1}} f^{(n)}\big((x_n - x_{n-1})t_n + \ldots + (x_1 - x_0)t_1 + x_0\big)\, dt_n \\ &= \frac{1}{x_n - x_{n-1}} \int_{s_0}^{s_1} f^{(n)}(s)\, ds = \frac{f^{(n-1)}(s_1) - f^{(n-1)}(s_0)}{x_n - x_{n-1}}\,,\end{aligned}$$

dove si è posto

$$\begin{aligned}s_0 &= (x_{n-1} - x_{n-2})t_{n-1} + \ldots + (x_1 - x_0)t_1 + x_0, \\ s_1 &= (x_n - x_{n-2})t_{n-1} + (x_{n-2} - x_{n-3})t_{n-2} + \ldots + (x_1 - x_0)t_1 + x_0.\end{aligned}$$

Da cui, per l'ipotesi induttiva

$$\begin{aligned}&g_n(x_0, x_1, \ldots, x_n) \\ &= \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-2}} \frac{f^{(n-1)}(s_1) - f^{(n-1)}(s_0)}{x_n - x_{n-1}}\, dt_{n-1} \\ &= \frac{g_{n-1}(x_0, x_1, \ldots, x_{n-2}, x_n) - g_{n-1}(x_0, x_1, \ldots, x_{n-2}, x_{n-1})}{x_n - x_{n-1}} \\ &= \frac{f[x_0, x_1, \ldots, x_{n-2}, x_n] - f[x_0, x_1, \ldots, x_{n-2}, x_{n-1}]}{x_n - x_{n-1}} = f[x_0, x_1, \ldots, x_n].\end{aligned}$$

∎

Dal teorema 5.28 segue, come si è notato, che ha senso scrivere differenze divise con argomenti ripetuti. Si ha allora per $k = 1, \ldots, n$

$$\begin{aligned}g_k(\underbrace{x, \ldots, x}_{k+1 \text{ volte}}) &= \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{k-1}} f^{(k)}(x)\, dt_k \\ &= f^{(k)}(x) \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{k-1}} dt_k = \frac{f^{(k)}(x)}{k!}\,,\end{aligned}$$

da cui

$$f[\underbrace{x,\dots,x}_{k+1 \text{ volte}}] = \frac{1}{k}\,\frac{d}{dx}\,f[\underbrace{x,\dots,x}_{k \text{ volte}}] = \frac{f^{(k)}(x)}{k!}\,, \quad k = 1,\dots,n, \tag{28}$$

e in modo analogo se $x_i \neq x_j$ per $i \neq j$ e $x \neq x_0,\dots,x_m$, è

$$\begin{aligned} f[x_0,x_1,\dots,x_m,\underbrace{x,\dots,x}_{k+1 \text{ volte}}] &= \frac{1}{k}\,\frac{d}{dx}\,f[x_0,x_1,\dots,x_m,\underbrace{x,\dots,x}_{k \text{ volte}}] \\ &= \frac{1}{k!}\,\frac{d^k}{dx^k}\,f[x_0,x_1,\dots,x_m,x]. \end{aligned} \tag{29}$$

Relazioni analoghe alle (28) e (29) valgono nel caso in cui vi siano più gruppi di variabili ripetute. Ad esempio, se x e y sono punti distinti di $[a,b]$, è

$$f[\underbrace{x,\dots,x}_{k+1 \text{ volte}},\underbrace{y,\dots,y}_{j+1 \text{ volte}}] = \frac{1}{k!\,j!}\,\frac{d^k}{dx^k}\,\frac{d^j}{dy^j}\,f[x,y], \quad \text{per} \quad 1 \leq k,\, j \leq n.$$

Il teorema 5.28, oltre a estendere la definizione di differenza divisa al caso di punti ripetuti, permette anche di scrivere i polinomi osculatori, sfruttando i valori delle derivate della $f(x)$. Per mezzo della (28) e della (29) si può infatti costruire la tabella delle differenze divise, oltre che con i valori della $f(x)$, anche con i valori delle derivate.

5.29 Esempio. Se della funzione $f(x)$ si conoscono i valori $f(x_0)$, $f(x_1)$, $f'(x_0)$, $f'(x_1)$, $f''(x_1)$, si può, sfruttando la proprietà di simmetria, costruire la seguente tabella delle differenze divise:

$$\begin{array}{c|ccccc}
x_0 & f[x_0] & & & & \\
 & & f[x_0,x_0] & & & \\
x_0 & f[x_0] & & f[x_0,x_0,x_1] & & \\
 & & f[x_0,x_1] & & f[x_0,x_0,x_1,x_1] & \\
x_1 & f[x_1] & & f[x_0,x_1,x_1] & & f[x_0,x_0,x_1,x_1,x_1] \\
 & & f[x_1,x_1] & & f[x_0,x_1,x_1,x_1] & \\
x_1 & f[x_1] & & f[x_0,x_1,x_1] & & \\
 & & f[x_1,x_1] & & & \\
x_1 & f[x_1] & & & &
\end{array}$$

tenendo conto che

$$\begin{aligned} f[x_0,x_0] &= f'(x_0), \\ f[x_1,x_1] &= f'(x_1), \\ f[x_1,x_1,x_1] &= \frac{f''(x_1)}{2}\,. \end{aligned}$$

Si noti che la tabella è stata costruita in modo che le differenze di ordine superiore al secondo risultino calcolabili con la definizione (18), a partire dagli elementi contigui della colonna precedente. Si ottiene così il polinomio di grado 4

$$\begin{aligned} p(x) = {} & f(x_0) + (x - x_0) f'(x_0) + (x - x_0)^2 f[x_0, x_0, x_1] \\ & + (x - x_0)^2 (x - x_1) f[x_0, x_0, x_1, x_1] \\ & + (x - x_0)^2 (x - x_1)^2 f[x_0, x_0, x_1, x_1, x_1], \end{aligned}$$

che è tale che $p(x_i) = f(x_i)$ e $p'(x_i) = f'(x_i)$, per $i = 0, 1$, e $p''(x_1) = f''(x_1)$.

Nel caso che della funzione $f(x)$ si conoscano il valore in un punto x_0 e i valori delle derivate fino all'ordine k in x_0, la tabella delle differenze divise risulta:

$$\begin{array}{c|cccccc} x_0 & f[x_0] & & & & & \\ & & f[x_0, x_0] & & & & \\ x_0 & f[x_0] & & f[x_0, x_0, x_0] & & & \\ & & f[x_0, x_0] & & \ddots & & \\ x_0 & f[x_0] & & \vdots & & f[\underbrace{x_0, \ldots, x_0}_{k+1 \text{ volte}}] & \\ & & \vdots & & \iddots & & \\ \vdots & \vdots & & f[x_0, x_0, x_0] & & & \\ & & f[x_0, x_0] & & & & \\ x_0 & f[x_0] & & & & & \end{array}$$

dove

$$f[x_0, x_0] = f'(x_0),$$

$$f[x_0, x_0, x_0] = \frac{f''(x_0)}{2},$$

$$\ldots$$

$$f[\underbrace{x_0, \ldots, x_0}_{k+1 \text{ volte}}] = \frac{f^{(k)}(x_0)}{k!},$$

e quindi il polinomio di Newton coincide, in questo caso, con il polinomio di grado k ottenuto dalla formula di Taylor. ∎

5.30 Esempio. Ripetendo opportunamente i punti x_i, si può costruire con le differenze divise il polinomio osculatore di Hermite, cioè il polinomio $p(x)$ di grado al più $2n+1$, che negli $n+1$ punti distinti x_i, $i = 0, \dots, n$, soddisfa alle $2n+2$ condizioni

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad i = 0, \dots, n.$$

Costruendo la tabella delle differenze divise nel modo seguente

$$\begin{array}{c|cccccc}
x_0 & f[x_0] & & & & & \\
 & & f[x_0,x_0] & & & & \\
x_0 & f[x_0] & & f[x_0,x_0,x_1] & & & \\
 & & f[x_0,x_1] & & \ddots & & \\
x_1 & f[x_1] & & f[x_0,x_1,x_1] & & & \\
 & & f[x_1,x_1] & & & & \\
x_1 & f[x_1] & & \vdots & & f[x_0,x_0,\dots,x_n,x_n] & \\
 & & \vdots & & & & \\
\vdots & \vdots & & f[x_0,x_0,x_n] & & & \\
 & & f[x_0,x_n] & & \ddots & & \\
x_n & f[x_n] & & f[x_0,x_n,x_n] & & & \\
 & & f[x_n,x_n] & & & & \\
x_n & f[x_n] & & & & &
\end{array}$$

dove $f[x_i, x_i] = f'(x_i)$, $i = 0, \dots, n$, si ottiene il polinomio osculatore di Hermite (17). ∎

Il seguente teorema estende la (23) al caso in cui i nodi non siano a due a due distinti.

5.31 Teorema. *Se $f(x) \in C^{n+1}[a,b]$, esiste un punto $\xi = \xi(x) \in (a,b)$ tale che*

$$f[x_0, x_1, \dots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,. \tag{30}$$

Se $f(x) \in C^{n+2}[a,b]$, esiste un punto $\eta = \eta(x) \in (a,b)$ tale che

$$\frac{d}{dx}\, f[x_0, x_1, \dots, x_n, x] = \frac{f^{(n+2)}(\eta)}{(n+2)!}\,. \tag{31}$$

In entrambi i casi, se i punti $x_0, \dots, x_n, x$ sono tutti coincidenti, allora $\xi = \eta = x$.

Dim. Applicando il teorema della media integrale alla funzione $g_{n+1}(x_0, \ldots, x_n, x)$ del teorema 5.28 si ottiene

$$g_{n+1}(x_0, \ldots, x_n, x) = f^{(n+1)}(\xi) \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_n} dt_n,$$

da cui segue la (30). La (31) si ottiene in modo analogo, essendo per la (29)

$$\frac{d}{dx}\, f[x_0, x_1, \ldots, x_n, x] = f[x_0, x_1, \ldots, x_n, x, x]. \qquad \blacksquare$$

8. Interpolazione inversa

Mentre nel problema dell'interpolazione è richiesta l'approssimazione del valore che una funzione $f(x)$ assume in un punto x, nel problema dell'*interpolazione inversa* è richiesta l'approssimazione del valore x tale che $f(x) = y$, dove y è assegnato. Questo problema viene di solito affrontato invertendo fra di loro i ruoli della variabile dipendente con quella indipendente, cioè considerando la funzione $x = g(y)$ tale che

$$g(y_i) = x_i, \quad i = 0, 1, \ldots, n,$$

e determinando il valore $g(y)$. Poiché ciò corrisponde ad interpolare la funzione inversa $x = f^{-1}(y)$, il procedimento richiede che la funzione $f(x)$ sia invertibile nell'intervallo $[a, b]$ contenente i punti x_i e quindi sia monotona.

Si possono usare le tecniche normali dell'interpolazione. L'unico inconveniente è rappresentato dal fatto che, anche se in origine i nodi x_i sono equidistanti, in generale i punti y_i non lo sono e quindi per la $g(y)$ non è possibile usare il polinomio di Newton con le differenze finite.

5.32 Esempio. L'interpolazione inversa viene frequentemente usata per approssimare uno zero di una funzione invertibile di cui siano noti i valori $y_i = f(x_i)$, per $i = 0, \ldots, n$: si ricerca cioè il punto

$$\alpha = f^{-1}(0).$$

Nel caso più semplice si assume $n = 1$, $x_0 < x_1$, $\alpha \in [x_0, x_1]$ e $f(x_0)f(x_1) < 0$. Con l'*interpolazione lineare inversa* si ottiene il polinomio

$$p_1(y) = x_0\, \frac{y - y_1}{y_0 - y_1} + x_1\, \frac{y - y_0}{y_1 - y_0}\,.$$

Ponendo $y = 0$ si ha

$$x = p_1(0) = x_0 - \frac{y_0(x_1 - x_0)}{y_1 - y_0}\,.$$

Assumere il valore x così calcolato come approssimazione di α corrisponde ad eseguire un passo del metodo delle secanti (si confronti con la (41, cap. 3).

Se $n = 2$ e $x_0 < x_1 < x_2$, $\alpha \in [x_0, x_2]$, $f(x_0)f(x_2) < 0$, con l'*interpolazione quadratica inversa* si ottiene il polinomio

$$p_2(y) = x_0 \frac{(y-y_1)(y-y_2)}{(y_0-y_1)(y_0-y_2)} + x_1 \frac{(y-y_0)(y-y_2)}{(y_1-y_0)(y_1-y_2)} + x_2 \frac{(y-y_0)(y-y_1)}{(y_2-y_0)(y_2-y_1)}.$$

Ponendo $y = 0$ si ha

$$x = p_2(0) = x_1 + \frac{y_1\big[y_0(y_1-y_0)(x_2-x_1) - y_2(y_2-y_1)(x_1-x_0)\big]}{(y_1-y_2)(y_1-y_0)(y_0-y_2)}. \quad (32)$$

Dalla (32), se y_0 e y_2 sono non nulli si ottiene la formula più compatta

$$x = x_1 + \frac{p}{q}, \quad \begin{aligned} p &= s\big[t(r-t)(x_2-x_1) - (1-r)(x_1-x_0)\big], \\ q &= (r-1)(s-1)(t-1), \end{aligned} \quad (33)$$

dove

$$r = \frac{y_1}{y_2}, \quad s = \frac{y_1}{y_0}, \quad t = \frac{y_0}{y_2},$$

usata nel metodo di Dekker-Brent (si veda l'esercizio 3.38) per approssimare le soluzioni delle equazioni.

Si consideri ad esempio l'equazione

$$f(x) = \sin^2 x - \cos x = 0,$$

che ha una sola soluzione nell'intervallo $\left[0, \frac{\pi}{2}\right]$. Fissati i punti $x_0 = 0$, $x_1 = \frac{\pi}{4}$, $x_2 = \frac{\pi}{2}$, in cui la funzione $f(x)$ ha i valori $y_0 = -1$, $y_1 = -0.2071067$, $y_2 = 1$, risulta

$$r = -s = -0.2071067, \quad t = -1 \quad \text{e} \quad x = p_2(0) = 0.9553491,$$

che si può assumere come approssimazione della soluzione cercata (la soluzione dell'equazione data è 0.9045569, quindi l'errore del valore ottenuto con l'interpolazione quadratica inversa è circa $0.5\ 10^{-1}$). ∎

9. Interpolazione razionale

La classe delle funzioni razionali è più ampia di quella dei polinomi e consente di approssimare meglio funzioni che alternano brusche variazioni a un comportamento quasi rettilineo.

Siano $p(x)$ un polinomio di grado al più m e $q(x)$ un polinomio di grado al più n e si consideri la funzione razionale di grado al più $m+n$

$$w(x) = \frac{p(x)}{q(x)} = \frac{a_m x^m + a_{m-1}x^{m-1} + \ldots + a_0}{b_n x^n + b_{n-1}x^{n-1} + \ldots + b_0}. \tag{34}$$

Dati $m+n+1$ punti distinti x_i, $i = 0, \ldots, m+n$, e i corrispondenti valori $f(x_i)$ della funzione $f(x)$, si dice che la funzione razionale $w(x)$ è *di interpolazione* della $f(x)$ se

$$w(x_i) = f(x_i), \quad i = 0, \ldots, m+n. \tag{35}$$

Si impongono quindi $m+n+1$ condizioni, cioè quante se ne potrebbero imporre con un polinomio di grado $m+n$.

Moltiplicando entrambi i membri della (35) per $q(x_i)$ si ottiene il sistema lineare omogeneo

$$a_m x_i^m + a_{m-1}x_i^{m-1} + \ldots + a_0 - f(x_i)\,(b_n x_i^n + b_{n-1}x_i^{n-1} + \ldots + b_0) = 0, \quad \text{per} \quad i = 0, \ldots, m+n, \tag{36}$$

di $m+n+1$ equazioni nelle $m+n+2$ incognite $a_m, a_{m-1}, \ldots, a_0$ e $b_n, b_{n-1}, \ldots, b_0$. Le soluzioni del sistema (36) per cui $b_n x_i^n + b_{n-1}x_i^{n-1} + \ldots + b_0 \neq 0$, per $i = 0, \ldots, m+n$, sono anche soluzioni del sistema (35), però in generale i due sistemi possono non essere equivalenti.

Si noti che non può esistere una soluzione non nulla di (36) in cui i coefficienti b_i, $i = 0, \ldots, n$, siano tutti nulli, perché se per assurdo ciò fosse vero, ne seguirebbe che il polinomio $p(x)$ di grado al più m avrebbe $m+n+1$ zeri distinti. Perciò il polinomio $q(x)$ al denominatore non è identicamente nullo.

Se la matrice del sistema (36) ha rango massimo, la soluzione di (36) è unica a meno di una costante moltiplicativa e la funzione $w(x)$ risulta determinata da soli $m+n+1$ parametri, in quanto il numeratore e il denominatore possono essere divisi per una stessa costante non nulla, senza che si alteri il valore di $w(x)$. Per questa ragione si considerano spesso funzioni razionali della forma (34), in cui uno dei due polinomi $p(x)$ o $q(x)$ è *monico* (cioè con il coefficiente del termine di grado massimo uguale a 1). Se invece la matrice del sistema (36) non ha rango massimo, vi sono infinite soluzioni.

Inoltre, poiché i due sistemi (35) e (36) possono non essere equivalenti, a differenza di quanto accade per l'interpolazione polinomiale, non è detto che il problema dell'interpolazione razionale abbia sempre una soluzione.

5.33 Esempio. La funzione $f(x)$ assume i valori

x	-1	0	1	2	3
$f(x)$	-4	-3/2	2	-1/4	0

Volendo costruire la funzione razionale $w(x)$ di grado 4 della forma

$$w(x) = \frac{a_2x^2 + a_1x + a_0}{b_2x^2 + b_1x + b_0}\,,$$

si risolve il sistema lineare omogeneo (36) $M\mathbf{a} = \mathbf{0}$, dove

$$M = \begin{bmatrix} 1 & -1 & 1 & 4 & -4 & 4 \\ 0 & 0 & 1 & 0 & 0 & 3/2 \\ 1 & 1 & 1 & -2 & -2 & -2 \\ 4 & 2 & 1 & 1 & 1/2 & 1/4 \\ 9 & 3 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_2 \\ a_1 \\ a_0 \\ b_2 \\ b_1 \\ b_0 \end{bmatrix}$$

Con opportune combinazioni lineari e scambi di righe, si vede che questo sistema è equivalente al sistema $M'\mathbf{a} = \mathbf{0}$, dove

$$M' = \begin{bmatrix} 1 & -1 & 1 & 4 & -4 & 4 \\ 0 & 1 & 0 & -3 & 1 & -3 \\ 0 & 0 & 1 & 0 & 0 & 3/2 \\ 0 & 0 & 0 & 1 & 7/2 & 9/4 \\ 0 & 0 & 0 & 0 & 1 & 1/2 \end{bmatrix}.$$

Ne risulta che la matrice M ha rango massimo. Ponendo $b_2 = 1$, si ottiene $a_2 = 1$, $a_1 = -4$, $a_0 = 3$, $b_1 = 1$, $b_0 = -2$, per cui

$$w(x) = \frac{x^2 - 4x + 3}{x^2 + x - 2}\,. \tag{37}$$

Il polinomio $q(x)$ al denominatore si annulla per $x = 1$, cioè $q(x)$ ha come fattore il polinomio $x-1$, che è anche fattore di $p(x)$. Riducendo la frazione $w(x)$ ai minimi termini, si ha che

$$w(x) = \frac{x - 3}{x + 2}\,. \tag{38}$$

Però la funzione $w(x)$ non soddisfa alle condizioni date perché $w(1) = -2/3$. Non esiste quindi alcuna funzione razionale, in cui numeratore e denominatore abbiano grado non superiore a 2, di interpolazione della funzione data.

Se la condizione nel punto $x = 1$ fosse sostituita dalla seguente $f(1) = -2/3$, allora la funzione (38) sarebbe soluzione del problema. Anche la (37), prolungata con continuità con il valore $-2/3$ nel punto 1, si potrebbe assumere come soluzione del problema. ∎

Per evitare che la funzione $w(x)$ possa avere più di una rappresentazione, si richiederà che i due polinomi $p(x)$ e $q(x)$ non abbiano come fattori comuni dei polinomi di grado maggiore o uguale a 1, cioè che $w(x)$ sia *irriducibile*.

5.34 Teorema. *Fissati i due interi m e n, siano $p_1(x)$ e $p_2(x)$ due polinomi di grado m, $q_1(x)$ e $q_2(x)$ due polinomi di grado n e siano*

$$w_1(x) = \frac{p_1(x)}{q_1(x)} \quad \text{e} \quad w_2(x) = \frac{p_2(x)}{q_2(x)}$$

due funzioni razionali ottenute risolvendo il sistema (36), tali che $q_1(x_i)$, $q_2(x_i) \neq 0$ per $i = 0, \ldots, m+n$. Allora $w_1(x)$ e $w_2(x)$ soddisfano alla (35), e se $w_1(x)$ e $w_2(x)$ sono irriducibili, allora $p_1(x) = \alpha p_2(x)$, $q_1(x) = \alpha q_2(x)$ per una opportuna costante $\alpha \neq 0$, cioè la soluzione del problema di interpolazione razionale è unica, a meno di costanti comuni a numeratore e denominatore.

Dim. Poiché

$$\frac{p_1(x_i)}{q_1(x_i)} = \frac{p_2(x_i)}{q_2(x_i)} = f(x_i), \quad i = 0, \ldots, m+n,$$

il polinomio, di grado non superiore a $m+n$,

$$t(x) = p_1(x)q_2(x) - p_2(x)q_1(x),$$

è tale che $t(x_i) = 0$, per $i = 0, \ldots, m+n$, e quindi è identicamente nullo. Se $w_1(x)$ e $w_2(x)$ sono irriducibili, segue che esiste $\alpha \neq 0$ per cui

$$p_1(x) = \alpha p_2(x), \qquad q_1(x) = \alpha q_2(x). \qquad ∎$$

Il problema dell'interpolazione razionale può avere soluzione per certe scelte di m e n, e non avere soluzione per altre scelte di m e n.

5.35 Esempio. Per la funzione $f(x)$ dell'esempio 5.33 non esiste una funzione razionale interpolante data come rapporto di due polinomi di grado 2. Se invece si cerca una funzione razionale della forma

$$w(x) = \frac{a_3x^3 + a_2x^2 + a_1x + a_0}{b_1x + b_0},$$

risolvendo il sistema lineare (36), si ottiene

$$a_3 = 2, \quad a_2 = -12, \quad a_1 = 21, \quad a_0 = -9, \quad b_1 = -5, \quad b_0 = 6,$$

per cui

$$w(x) = \frac{2x^3 - 12x^2 + 21x - 9}{-5x + 6}.$$

La funzione $w(x)$ è di interpolazione in quanto il denominatore non si annulla in alcuno dei nodi. ∎

Prima di proseguire nello studio dell'interpolazione razionale, è opportuno vedere come ogni funzione razionale possa essere espressa per mezzo di una frazione continua finita.

10. Frazioni continue

Una *frazione continua finita di ordine k* è un'espressione della forma

$$w(x) = d_0 + \cfrac{c_1}{d_1 + \cfrac{c_2}{d_2 + \cfrac{c_3}{\ddots + \cfrac{c_k}{d_k}}}}, \tag{39}$$

in cui i c_i e d_i sono polinomi a coefficienti reali, $c_i \neq 0$ per $i = 1, \ldots, k$, e $d_k \neq 0$. Per semplicità si usa la notazione

$$w(x) = d_0 + \frac{c_1}{d_1} + \frac{c_2}{d_2} + \cdots + \frac{c_k}{d_k}.$$

Indicata con w_i, $i \leq k$, la i-esima *frazione parziale*, cioè

$$w_i = d_0 + \frac{c_1}{d_1} + \frac{c_2}{d_2} + \cdots + \frac{c_i}{d_i},$$

è facile verificare (si veda l'esercizio 5.35) che w_i è una funzione razionale della forma

$$w_i = \frac{p_i}{q_i},$$

dove p_i e q_i soddisfano le relazioni ricorrenti

$$\left.\begin{array}{l} p_{-1} = 1, \quad q_{-1} = 0, \quad p_0 = d_0, \quad q_0 = 1, \\ p_i = d_i p_{i-1} + c_i p_{i-2} \\ \\ q_i = d_i q_{i-1} + c_i q_{i-2} \end{array}\right\} \quad \text{per } i = 1, \ldots, k. \tag{40}$$

Quindi i p_i e i q_i sono soluzioni della stessa equazione alle differenze, con condizioni iniziali diverse. Poiché

$$w(x) = w_k = \frac{p_k}{q_k},$$

$w(x)$ risulta una funzione della forma (34). Indicato con z_i l'i-esimo *residuo*

$$z_i = d_i + \frac{c_{i+1}}{d_{i+1}} + \cdots + \frac{c_k}{d_k}, \quad \text{per} \quad i = 0, \ldots, k-1, \quad \text{e} \quad z_k = d_k,$$

vale la relazione

$$z_i = d_i + \frac{c_{i+1}}{z_{i+1}}, \quad i = 0, \ldots, k-1, \tag{41}$$

che può essere utilizzata come algoritmo di calcolo della funzione $w(x) = z_0$, a partire da $z_k = d_k$, facendo variare l'indice i da $k-1$ fino a 0.

La (41) presenta però lo svantaggio che non è possibile calcolare w_{i+1} a partire da w_i, cosa che è invece possibile fare utilizzando le (40), però con un costo superiore (si veda l'esercizio 5.37). Altri algoritmi si ottengono sfruttando il fatto che il valore di una frazione continua può essere ricavato per mezzo di un opportuno sistema lineare (si veda l'esercizio 5.39)

Una funzione razionale $w(x)$ può essere rappresentata in vari modi sotto forma di frazione continua. Il seguente teorema mostra come $w(x)$ sia rappresentabile nella forma (39) in cui i c_i, $i = 1, \ldots, k$, sono delle costanti. Con questa rappresentazione il calcolo di $w(x)$ risulta in generale più efficiente. Successivamente si vedrà come $w(x)$ sia rappresentabile nella forma (39) in cui i d_i, $i = 1, \ldots, k$, sono delle costanti. Questa forma è anche quella che si ottiene quando $w(x)$ è costruita con una tecnica di interpolazione razionale.

5.36 Teorema. *Sia*

$$w(x) = \frac{p(x)}{q(x)}$$

una funzione razionale irriducibile, in cui $q(x)$ è un polinomio monico di grado $n \geq 1$. Allora $w(x)$ può essere espressa sotto forma di frazione continua finita di ordine $k \leq n$

$$w(x) = s_0(x) + \frac{\alpha_1}{s_1(x)} + \frac{\alpha_2}{s_2(x)} + \cdots + \frac{\alpha_k}{s_k(x)}, \tag{42}$$

dove gli α_i, per $i = 1, \ldots, k$, sono delle costanti e gli $s_i(x)$, per $i = 0, \ldots, k$, sono dei polinomi. In particolare, se m è il grado di $p(x)$, allora $s_0(x)$ ha grado $m - n$ se $m \geq n$ ed è identicamente nullo se $m < n$, e gli $s_i(x)$, per $i = 1, \ldots, k$, sono polinomi monici di grado almeno 1.

Dim. Se $m \geq n$, si ponga $t_0(x) = p(x)$ e $t_1(x) = q(x)$. Si applichi il seguente algoritmo di Euclide modificato

$$t_i(x) = t_{i+1}(x)s_i(x) + \alpha_{i+1}t_{i+2}(x), \quad \text{per } i = 0, \ldots, k-1, \tag{43}$$

dove $s_i(x)$ è il quoziente e $\alpha_{i+1}t_{i+2}(x)$ è il resto della divisione di $t_i(x)$ per $t_{i+1}(x)$, e α_{i+1} è scelto in modo che $t_{i+2}(x)$ sia monico. L'algoritmo, che calcola il massimo comun divisore fra $t_0(x)$ e $t_1(x)$, termina dopo $k \leq n$ passi, dove k è il minimo intero per cui $t_{k+1}(x)$ è un polinomio di grado 0. L'ipotesi di irriducibilità fatta su $w(x)$ assicura che $t_{k+1}(x)$ è una costante non nulla e quindi uguale a 1. Poiché il grado del polinomio $t_{i+2}(x)$ è minore del grado di $t_{i+1}(x)$, il quoziente $s_{i+1}(x)$ della divisione di $t_{i+1}(x)$ per $t_{i+2}(x)$ ha grado almeno 1, e poiché i primi coefficienti di $t_{i+1}(x)$ e $t_{i+2}(x)$ sono uguali a 1, il primo coefficiente di $s_{i+1}(x)$ è uguale a 1. Si pone infine $s_k(x) = t_k(x)$. Dividendo $t_i(x)$ per $t_{i+1}(x)$ e ponendo

$$z_i(x) = \frac{t_i(x)}{t_{i+1}(x)}\,,$$

dalla (43) si ha

$$z_i(x) = s_i(x) + \frac{\alpha_{i+1}}{z_{i+1}(x)}\,.$$

Confrontando con la (41), ne segue la (42), poiché

$$w(x) = \frac{t_0(x)}{t_1(x)}\,.$$

Se $m < n$, si ponga $t_1(x) = q(x)$ e $\alpha_1 t_2(x) = p(x)$, dove α_1 è tale che $t_2(x)$ sia monico, e si proceda in modo analogo applicando la (43) a partire dall'indice $i = 1$. L'unica differenza con il caso precedente è quindi l'assenza del primo termine, cioè $s_0(x) \equiv 0$. ∎

Si considerino i seguenti casi particolarmente importanti:

a) se $m = n$, allora $s_0(x)$ è una costante;

b) se $m = n + 1$, allora $s_0(x)$ è un polinomio di grado 1.

Sia nel caso a) che nel caso b) se il procedimento (43) determina come quozienti dei polinomi $s_i(x)$ di grado 1, allora $k = n$. Quindi la frazione continua è della forma

$$w(x) = \alpha_0 x + \beta_0 + \frac{\alpha_1}{x + \beta_1} + \frac{\alpha_2}{x + \beta_2} + \cdots + \frac{\alpha_n}{x + \beta_n}\,,$$

dove $\alpha_0 = 0$ se $m = n$. Il valore di $w(x)$ viene allora calcolato con l'algoritmo

$$\begin{aligned} &z_n = x + \beta_n, \\ &z_i = x + \beta_i + \frac{\alpha_{i+1}}{z_{i+1}}, \quad \text{per } i = n-1, \ldots, 1, \\ &w(x) = \alpha_0 x + \beta_0 + \frac{\alpha_1}{z_1}. \end{aligned}$$

È ovviamente possibile che il procedimento richieda un numero minore di passi e che risulti $k < n$.

Il calcolo di $w(x)$, dati i coefficienti dei polinomi $p(x)$ e $q(x)$, è in generale più costoso del calcolo della stessa funzione, dati i coefficienti della frazione continua, in quanto il numero delle operazioni moltiplicative richieste può risultare anche doppio nel primo caso.

5.37 Esempi. Per la funzione razionale

$$w(x) = \frac{x^4 - 15x^2 + x + 30}{x^3 + 2x^2 - 14x - 30}, \tag{44}$$

applicando la (43) si ha

$$\begin{aligned} x^4 - 15x^2 + x + 30 &= (x^3 + 2x^2 - 14x - 30)(x - 2) + 3(x^2 + x - 10) \\ x^3 + 2x^2 - 14x - 30 &= (x^2 + x - 10)(x + 1) - 5(x + 4) \\ x^2 + x - 10 &= (x + 4)(x - 3) + 2, \end{aligned}$$

per cui

$$w(x) = x - 2 + \frac{3}{x+1} + \frac{-5}{x-3} + \frac{2}{x+4}. \tag{45}$$

In questo caso risulta quindi $k = n$. Calcolare un valore di $w(x)$ usando la forma (44) richiede 6 operazioni additive e 5 moltiplicative, mentre se si usa la forma (45) occorrono 7 operazioni additive e 3 moltiplicative. Per la funzione razionale

$$w(x) = \frac{x^4 - 3x^3 + 3x^2 - 5x + 2}{x^3 - 2x^2 + x - 6}, \tag{46}$$

applicando la (43) si ha

$$\begin{aligned} x^4 - 3x^3 + 3x^2 - 5x + 2 &= (x^3 - 2x^2 + x - 6)(x - 1) + 2(x - 2) \\ x^3 - 2x^2 + x - 6 &= (x - 2)(x^2 + 1) - 4, \end{aligned}$$

per cui

$$w(x) = x - 1 + \frac{2}{x^2+1} + \frac{-4}{x-2}. \tag{47}$$

In questo caso risulta quindi $k = n - 1$. Calcolare un valore di $w(x)$ usando la forma (46) richiede 7 operazioni additive e 6 moltiplicative, mentre se si usa la forma (47) occorrono 5 operazioni additive e 3 moltiplicative. ∎

Se $w(x)$ è una funzione pari, nella corrispondente frazione continua i polinomi $s_i(x)$ sono tutti pari e se $w(x)$ è funzione dispari, tutti gli $s_i(x)$ sono dispari.

5.38 Esempi. Per la funzione razionale pari

$$w(x) = \frac{2x^6 - x^4 + 6x^2 + 3}{x^4 - x^2 + 4},$$

applicando la (43) si ha

$$\begin{aligned} 2x^6 - x^4 + 6x^2 + 3 &= (x^4 - x^2 + 4)(2x^2 + 1) - (x^2 + 1) \\ x^4 - x^2 + 4 &= (x^2 + 1)(x^2 - 2) + 6, \end{aligned}$$

per cui

$$w(x) = 2x^2 + 1 + \frac{-1}{x^2 - 2} + \frac{6}{x^2 + 1}.$$

Per la funzione razionale dispari

$$w(x) = \frac{3x^5 - 4x^3 + 2x}{x^4 + 2},$$

applicando la (43) si ha

$$\begin{aligned} 3x^5 - 4x^3 + 2x &= (x^4 + 2)3x - 4(x^3 + x) \\ x^4 + 2 &= (x^3 + x)x - (x^2 - 2) \\ x^3 + x &= (x^2 - 2)x + 3x \\ x^2 - 2 &= xx - 2, \end{aligned}$$

per cui

$$w(x) = 3x + \frac{-4}{x} + \frac{-1}{x} + \frac{3}{x} + \frac{-2}{x}.$$ ∎

La rappresentazione di una funzione razionale con una frazione continua della forma (42) può non risultare la migliore, dal punto di vista computazionale, come risulta dal seguente esempio.

5.39 Esempio. Si consideri la funzione razionale

$$w(x) = \frac{2x^4 - x^3 - 6x^2 - 3x - 4}{x^4 - 2x^2 - x - 1}.$$

Applicando la (43) si ha

$$\begin{aligned}
2x^4 - x^3 - 6x^2 - 3x - 4 &= (x^4 - 2x^2 - x - 1)2 - (x^3 + 2x^2 + x + 2)\\
x^4 - 2x^2 - x - 1 &= (x^3 + 2x^2 + x + 2)(x - 2) + x^2 - x + 3\\
x^3 + 2x^2 + x + 2 &= (x^2 - x + 3)(x + 3) + x - 7\\
x^2 - x + 3 &= (x - 7)(x + 6) + 45,
\end{aligned}$$

da cui si ottiene per $w(x)$ la frazione continua

$$w(x) = 2 + \frac{-1}{x-2} + \frac{1}{x+3} + \frac{1}{x+6} + \frac{45}{x-7},$$

che richiede per il calcolo 8 operazioni additive e 4 moltiplicative. Procedendo invece così

$$\begin{aligned}
2x^4 - x^3 - 6x^2 - 3x - 4 &= (x^4 - 2x^2 - x - 1)2 - (x + 2)(x^2 + 1)\\
x^4 - 2x^2 - x - 1 &= (x^2 + 1)(x^2 - 3) - (x - 2),
\end{aligned}$$

si ottiene la frazione continua

$$w(x) = 2 - \frac{x+2}{x^2-3} + \frac{-x+2}{x^2+1}, \tag{48}$$

che richiede per il calcolo 6 operazioni additive e 3 moltiplicative. Naturalmente la (48) non è una frazione continua della forma (42). ∎

Dal punto di vista dell'accuratezza del risultato, il calcolo di una funzione razionale utilizzando una frazione continua può essere numericamente instabile.

5.40 Esempio. Dalla funzione razionale pari

$$w(x) = \frac{x^4 + x^2 + 1}{\gamma x^4 + x^2 + 1}$$

si ottiene, applicando la (43), la frazione continua

$$w(x) = \beta_0 - \frac{\alpha_1}{x^2 + \beta_1} + \frac{\alpha_2}{x^2 + \beta_2},$$

dove

$$\beta_0 = \frac{1}{\gamma}, \quad \alpha_1 = \frac{1-\gamma}{\gamma^2}, \quad \beta_1 = \frac{1-\gamma}{\gamma}, \quad \alpha_2 = \beta_2 = 1.$$

Fissato un valore di γ e determinati i coefficienti della frazione continua, il calcolo di $w(x)$ per mezzo della frazione continua richiede 4 operazioni

additive e 3 operazioni moltiplicative, mentre il calcolo di $w(x)$ come rapporto di polinomi richiede 4 operazioni additive e 4 operazioni moltiplicative. Assegnando a γ un valore compreso fra 0 e 1, risulta $1 \leq w(x) \leq 1.5$ per $x \in [0,1]$. Poiché il primo coefficiente della frazione continua è $\beta_0 = \dfrac{1}{\gamma}$, se γ è molto piccolo, il valore di $w(x)$ risulta differenza di due numeri molto grandi e vicini fra loro, con la conseguente possibilità che si verifichino errori di cancellazione. Infatti per valori decrescenti di γ si ha la seguente tabella, in cui ϵ_1 e ϵ_2 sono i massimi moduli, per $x \in [0,1]$, degli errori relativi effettivamente generati calcolando $w(x)$ rispettivamente come rapporto di polinomi e per mezzo della frazione continua.

γ	ϵ_1	ϵ_2
10^{-1}	$0.128\ 10^{-5}$	$0.268\ 10^{-5}$
10^{-2}	$0.160\ 10^{-5}$	$0.476\ 10^{-4}$
10^{-3}	$0.160\ 10^{-5}$	$0.118\ 10^{-2}$
10^{-4}	$0.126\ 10^{-5}$	$0.672\ 10^{-2}$
10^{-5}	$0.132\ 10^{-5}$	$0.870\ 10^{-1}$

■

Per trasformare una funzione razionale in frazione continua si può usare anche una tecnica diversa da quella descritta nella dimostrazione del teorema 5.36. Con questa tecnica, che va sotto il nome di *metodo di Viskovatov*, si ottiene una frazione continua della forma

$$w(x) = \frac{\alpha_1 x^{i_1}}{1} + \frac{\alpha_2 x^{i_2}}{1} + \cdots + \frac{\alpha_k x^{i_k}}{1}, \tag{49}$$

se nella (34) è $b_0 \neq 0$ (se fosse $b_0 = 0$, poiché $w(x)$ è irriducibile, dovrebbe essere $a_0 \neq 0$ e quindi il metodo potrebbe essere applicato alla funzione $\dfrac{1}{w(x)}$).

Si supponga perciò che sia $b_0 = 1$; posto $t_1(x) = q(x)$, si indichi con $t_2(x)$ il polinomio tale che

$$p(x) = \alpha_1 x^{i_1} t_2(x),$$

in cui $\alpha_1 x^{i_1}$ è il termine non nullo di grado minimo di $p(x)$ (se $a_0 \neq 0$ è $\alpha_1 = a_0$ e $i_1 = 0$). Quindi il termine di grado minimo di $t_2(x)$ è uguale a 1 e si ha

$$\frac{p(x)}{q(x)} = \frac{\alpha_1 x^{i_1}}{\dfrac{t_1(x)}{t_2(x)}} = \frac{\alpha_1 x^{i_1}}{1 + \dfrac{t_1(x) - t_2(x)}{t_2(x)}}.$$

Il polinomio $t_1(x) - t_2(x)$ è divisibile per x e si può scrivere

$$t_1(x) - t_2(x) = \alpha_2 x^{i_2} t_3(x),$$

in cui $\alpha_2 x^{i_2}$ è il termine non nullo di grado minimo di $t_1(x) - t_2(x)$. Procedendo in questo modo fino ad un indice k per cui $t_k(x) = t_{k+1}(x)$, si ottiene la (49).

Se $i_1 = 0$ e $i_2 = \ldots = i_k = 1$, si può verificare che $k = m + n + 1$, per cui la (49) risulta

$$w(x) = \frac{\alpha_1}{1} + \frac{\alpha_2 x}{1} + \cdots + \frac{\alpha_k x}{1}\,. \tag{50}$$

Frazioni continue della forma (50) vengono dette *S-frazioni*.

5.41 Esempio. Per la funzione razionale

$$w(x) = \frac{x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!}}{1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!}}$$

si ha

$$w(x) = \frac{x t_2(x)}{t_1(x)}\,, \quad t_1(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!}\,, \quad t_2(x) = 1 - \frac{x^2}{3!} + \frac{x^4}{5!}\,,$$

$$\frac{t_1(x)}{t_2(x)} = 1 - \frac{x^2 t_3(x)}{3\, t_2(x)}\,, \quad t_3(x) = 1 - \frac{x^2}{10}\,,$$

$$\frac{t_2(x)}{t_3(x)} = 1 - \frac{x^2 t_4(x)}{15\, t_3(x)}\,, \quad t_4(x) = 1 - \frac{x^2}{8}\,,$$

$$\frac{t_3(x)}{t_4(x)} = 1 + \frac{x^2 t_5(x)}{40\, t_4(x)}\,, \quad t_5(x) = 1,$$

$$\frac{t_4(x)}{t_5(x)} = 1 - \frac{x^2 t_6(x)}{8\, t_5(x)}\,, \quad t_6(x) = 1,$$

da cui

$$w(x) = \frac{x}{1} - \frac{x^2/3}{1} - \frac{x^2/15}{1} + \frac{x^2/40}{1} - \frac{x^2/8}{1}\,. \qquad \blacksquare$$

Mentre l'algoritmo usato nella dimostrazione del teorema 5.36 non consente di trasformare in frazione continua i polinomi, questo è possibile con il metodo di Viskovatov.

5.42 Esempio. Per il polinomio

$$p(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$$

si ha

$$\begin{aligned}
&p(x) = \frac{t_2(x)}{t_1(x)}\,, \quad t_1(x) = 1, \quad t_2(x) = p(x),\\
&\frac{t_1(x)}{t_2(x)} = 1 - \frac{x\,t_3(x)}{t_2(x)}\,, \quad t_3(x) = 1 + \frac{x}{2} + \frac{x^2}{6}\,,\\
&\frac{t_2(x)}{t_3(x)} = 1 + \frac{x\,t_4(x)}{2\,t_3(x)}\,, \quad t_4(x) = 1 + \frac{2x}{3} + \frac{x^2}{3}\,,\\
&\frac{t_3(x)}{t_4(x)} = 1 - \frac{x\,t_5(x)}{6\,t_4(x)}\,, \quad t_5(x) = 1 + x,\\
&\frac{t_4(x)}{t_5(x)} = 1 - \frac{x\,t_6(x)}{3\,t_5(x)}\,, \quad t_6 = 1 - x,\\
&\frac{t_5(x)}{t_6(x)} = 1 + \frac{2x}{t_6(x)} = 1 - \frac{2x}{1-x}\,,
\end{aligned}$$

da cui

$$w(x) = \frac{1}{1} - \frac{x}{1} + \frac{x/2}{1} - \frac{x/6}{1} - \frac{x/3}{1} + \frac{2x}{1} - \frac{x}{1}\,. \qquad \blacksquare$$

Con piccole varianti del metodo di Viskovatov è possibile scrivere $w(x)$ in frazione continua in un modo ancora diverso. Ad esempio si applica il metodo alla funzione $w(x) - w(0)$, ottenendo

$$w(x) = w(0) + \frac{\alpha_1 x^{i_1}}{1} + \frac{\alpha_2 x^{i_2}}{1} + \cdots + \frac{\alpha_k x^{i_k}}{1}\,,$$

in cui $i_1 \geq 1$, oppure nel caso $m \geq n$ si può prima calcolare il quoziente $s(x)$ e il resto $r(x)$ della divisione di $p(x)$ per $q(x)$, ottenendo

$$w(x) = s(x) + \frac{r(x)}{q(x)} = s(x) + \frac{\alpha_1 x^{i_1}}{1} + \frac{\alpha_2 x^{i_2}}{1} + \cdots + \frac{\alpha_k x^{i_k}}{1}\,.$$

In generale le diverse frazioni continue corrispondenti ad una stessa funzione razionale $w(x)$ non sono *equivalenti*, dove per *equivalenza* di due frazioni continue si intende che le successioni delle frazioni parziali coincidono.

11. Differenze inverse

I coefficienti della frazione continua di una funzione razionale $w(x)$ che interpola la funzione $f(x)$ sui nodi x_i, $i = 0, \dots, m+n$, possono essere ricavati direttamente per mezzo di un procedimento analogo a quello con cui si costruisce il polinomio di interpolazione di Newton, usando al posto delle differenze divise le differenze inverse.

5.43 Definizione. Si chiama *differenza inversa* di ordine k della funzione $f(x)$ relativa ai punti $x_i, i = 0, \dots, k-1$, a due a due distinti, la funzione $\phi[x_0, x_1, \dots, x_{k-1}, x]$ definita ricorsivamente nel modo seguente

$$\begin{aligned}
&\text{per } k = 0 \quad && \phi[x] = f(x), \\
&\text{per } k = 1 \quad && \phi[x_0, x] = \frac{x - x_0}{\phi[x] - \phi[x_0]}\,, \\
&\text{per } k \geq 2 \quad && \phi[x_0, x_1, \dots, x_{k-2}, x_{k-1}, x] \qquad (51) \\
& && = \frac{x - x_{k-1}}{\phi[x_0, x_1, \dots, x_{k-2}, x] - \phi[x_0, x_1, \dots, x_{k-2}, x_{k-1}]}\,. \qquad \blacksquare
\end{aligned}$$

Oltre che nel punto x_{k-1} la funzione $\phi[x_0, x_1, \dots, x_{k-2}, x_{k-1}, x]$ risulta non definita anche in ogni altro punto x tale che $\phi[x_0, x_1, \dots, x_{k-2}, x] = \phi[x_0, x_1, \dots, x_{k-2}, x_{k-1}]$.

Per il calcolo delle differenze inverse si usa una tabella simile a quella delle differenze divise:

$$\begin{array}{c|ccccccc}
x_0 & \phi[x_0] & & & & & & \\
 & & \phi[x_0, x_1] & & & & & \\
x_1 & \phi[x_1] & & \phi[x_0, x_1, x_2] & & & & \\
 & & \phi[x_0, x_2] & & & & & \\
x_2 & \phi[x_2] & & \vdots & & \dots & \phi[x_0, x_1, \dots, x_{m+n}] & \\
 & & \vdots & & & & & \\
\vdots & \vdots & & \phi[x_0, x_1, x_{m+n}] & & & & \\
 & & \phi[x_0, x_{m+n}] & & & & & \\
x_{m+n} & \phi[x_{m+n}] & & & & & &
\end{array}$$

Gli elementi della tabella, considerati come elementi di una matrice A, triangolare inferiore di ordine $m+n+1$, vengono costruiti mediante le relazioni

$$\begin{aligned}
&a_{i1} = f(x_{i-1}), \quad i = 1, \dots, m+n+1 \\
&a_{ij} = \frac{x_{i-1} - x_{j-2}}{a_{i,j-1} - a_{j-1,j-1}}\,, \quad j = 2, \dots, m+n+1, \quad i = j, \dots, m+n+1.
\end{aligned}$$

Poiché per ogni elemento sono richieste due sottrazioni e una divisione, la costruzione della tabella richiede, a meno di termini di ordine inferiore, $(m+n)^2$ addizioni e $(m+n)^2/2$ moltiplicazioni.

Il seguente teorema mostra che la funzione $f(x)$ è esprimibile come frazione continua mediante le sue differenze inverse.

5.44 Teorema. *Se le differenze inverse $\phi[x_0, x_1, \ldots, x_k]$ sono definite per $k = 0, \ldots, m+n$, allora vale la seguente relazione*

$$f(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x_1]} + \frac{x - x_1}{\phi[x_0, x_1, x_2]} + \cdots + \frac{x - x_{m+n-1}}{\phi[x_0, x_1, \ldots, x_{m+n}]} + \frac{x - x_{m+n}}{\phi[x_0, \ldots, x_{m+n}, x]}, \tag{52}$$

in ogni punto x per cui è definita $\phi[x_0, x_1, \ldots, x_{m+n}, x]$.

Dim. Dal fatto che $\phi[x_0, x_1, \ldots, x_{m+n}, x]$ è definita, segue che in x sono definite anche tutte le differenze inverse di ordine inferiore. Si procede quindi per induzione. Per $m = n = 0$, poiché $\phi[x_0, x] \neq 0$ per $x \neq x_0$, si ha

$$f(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x]},$$

come segue direttamente dalla definizione 5.43 per $k = 1$. Per $m + n > 0$ si supponga che la (52) valga fino all'indice $m + n - 1$, cioè

$$f(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x_1]} + \frac{x - x_1}{\phi[x_0, x_1, x_2]} + \cdots + \frac{x - x_{m+n-1}}{\phi[x_0, \ldots, x_{m+n-1}, x]}, \tag{53}$$

in cui $\phi[x_0, x_1, \ldots, x_{m+n-1}, x] \neq 0$ per $x \neq x_{m+n-1}$. Ma per la (51) è per $k = m + n + 1$

$$\phi[x_0, \ldots, x_{m+n-1}, x] = \phi[x_0, \ldots, x_{m+n-1}, x_{m+n}] + \frac{x - x_{m+n}}{\phi[x_0, \ldots, x_{m+n}, x]},$$

e sostituendo nella (53), ne segue subito la (52). ∎

5.45 Teorema. *Nell'ipotesi del teorema 5.44, si considerino la frazione continua di ordine $m + n$*

$$w(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x_1]} + \frac{x - x_1}{\phi[x_0, x_1, x_2]} + \cdots + \frac{x - x_{m+n-1}}{\phi[x_0, x_1, \ldots, x_{m+n}]} \tag{54}$$

e i suoi residui per $i = 0, \ldots, m + n + 1$

$$z_i(x) = \phi[x_0, x_1 \ldots, x_i] + \frac{x - x_i}{\phi[x_0, x_1, \ldots, x_{i+1}]} + \cdots + \frac{x - x_{m+n-1}}{\phi[x_0, x_1, \ldots, x_{m+n}]}.$$

Allora la (54) verifica la condizione $w(x_k) = f(x_k)$ *per* $k = m+n-1$, *per* $k = m+n$ *e per quei* $k = 0, \ldots, m+n-2$, *per cui* $z_{k+1}(x_k) \neq 0$. *La (54), detta frazione continua di Thiele, è funzione razionale di interpolazione della* $f(x)$ *sui nodi* x_i, $i = 0, \ldots, m+n$.

Dim. Dalla (52) segue che

$$f(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x_1]} + \frac{x - x_1}{\phi[x_0, x_1, x_2]} + \cdots + \frac{x - x_{m+n-1}}{\phi[x_0, \ldots, x_{m+n-1}, x]},$$

e quindi $f(x_{m+n}) = w(x_{m+n})$. Posto $z_{m+n}(x) = \phi[x_0, x_1, \ldots, x_{m+n}]$, per $k = 0, \ldots, m+n-1$ risulta

$$w(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x_1]} + \cdots + \frac{x - x_{k-1}}{\phi[x_0, x_1, \ldots, x_k]} + \frac{x - x_k}{z_{k+1}(x)}.$$

Per $k = m+n-1$ è $z_{m+n}(x) \neq 0$ perché $x_{m+n-1} \neq x_{m+n}$, e quindi $f(x_{m+n-1}) = w(x_{m+n-1})$.
Per $k = 0, \ldots, m+n-2$, se $z_{k+1}(x_k) \neq 0$, risulta per la (52) che

$$w(x_k) = \phi[x_0] + \frac{x_k - x_0}{\phi[x_0, x_1]} + \cdots + \frac{x_k - x_{k-1}}{\phi[x_0, x_1, \ldots, x_k]} = f(x_k). \qquad \blacksquare$$

5.46 Esempio. La tabella delle differenze inverse della funzione $f(x)$ dell'esempio 5.33 è data da

-1	-4				
		$\frac{2}{5}$			
0	$-\frac{3}{2}$		-15		
		$\frac{1}{3}$		$\frac{1}{20}$	
1	2		5		20
		$\frac{4}{5}$		$\frac{1}{10}$	
2	$-\frac{1}{4}$		5		
		1			
3	0				

La funzione razionale che si ottiene

$$w(x) = -4 + \frac{x+1}{2/5} + \frac{x}{-15} + \frac{x-1}{1/20} + \frac{x-2}{20}$$

non è di interpolazione perché risulta $w(1) \neq 2$. Infatti è

$$z_3(x) = \frac{1}{20} + \frac{x-2}{20}, \quad \text{e} \quad z_3(1) = 0.$$

Riducendo a frazione unica, a seconda che si eseguano o meno semplificazioni, si ottiene per $w(x)$ l'espressione (38) oppure la (37). ∎

In molti casi le funzioni razionali di interpolazione sono migliori dei polinomi di interpolazione dello stesso grado, anche perché consentono di interpolare funzioni che hanno delle singolarità in punti dell'intervallo di interpolazione o vicini ad esso. Inoltre, mentre è impossibile costruire un polinomio di interpolazione di una funzione limitata su un intervallo illimitato, ciò è possibile con le funzioni razionali di interpolazione.

5.47 Esempio. Utilizzando gli stessi valori della funzione $f(x) = \dfrac{\tan x}{x}$ usati nell'esempio 5.18 per ottenere il polinomio di interpolazione, si costruisce la tabella delle differenze inverse

0.3	1.031120			
		2.749594		
0.6	1.140227		−0.2817849	
		1.330073		−7.675201
1.0	1.557407		−0.3339008	
		0.353673		
1.4	4.141337			

da cui si ottiene

$$w(x) = 1.031120 + \frac{x-0.3}{2.749594} + \frac{x-0.6}{-0.2817849} + \frac{x-1}{-7.675201}.$$

Il valore $f(1.2)$ viene approssimato con $w(1.2) = 2.155350$. Poiché $f(1.2) = 2.143460$, il valore ottenuto con l'interpolazione razionale è migliore del valore $p_3(1.2) = 2.451967$, ottenuto con l'interpolazione polinomiale nell'esempio 5.18. Nella figura 5.15 sono illustrati l'andamento della funzione $f(x)$ (con linea più spessa) e della funzione razionale $w(x)$ di interpolazione (linea sottile) nell'intervallo $[0,\ 3]$. Si noti come la funzione razionale abbia un andamento che riproduce bene quello della $f(x)$, sia perché nell'intervallo $[0, 1.5]$ i grafici delle due funzioni sono sovrapposti, sia perché la $f(x)$ ha una singolarità in $\dfrac{\pi}{2} = 1.570796$, e la funzione razionale ha una singolarità in 1.584013. Questo comportamento non è ovviamente possibile per il polinomio di interpolazione dell'esempio 5.18. ∎

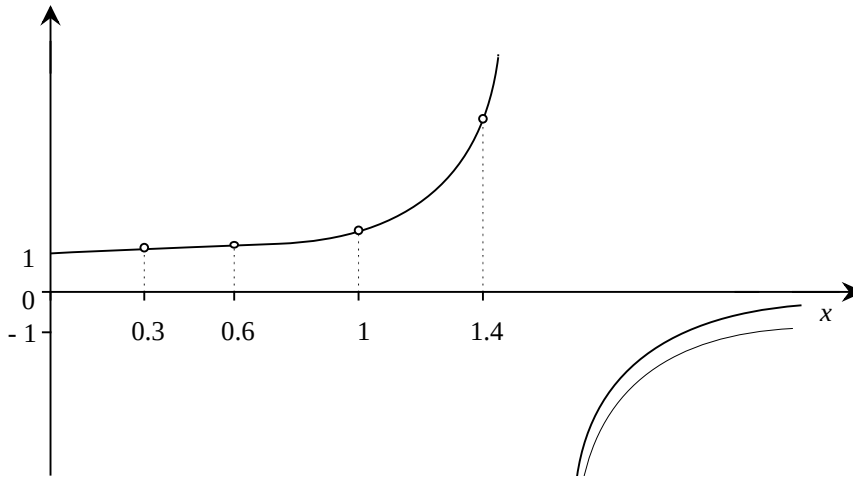


Fig. 5.15 - Funzione razionale di interpolazione di $f(x) = \dfrac{\tan x}{x}$.

Se per un indice k accade che

$$\phi[x_0, x_1, \ldots, x_{k-2}, x_k] = \phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}], \tag{55}$$

allora la differenza inversa $\phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x_k]$ non può essere calcolata e quindi la funzione razionale di interpolazione non può essere rappresentata nella forma (55). Se $k < n$ è talvolta possibile, con un opportuno riordinamento dei punti, evitare che ciò accada.

Dal teorema 5.45 segue che la i-esima frazione parziale della (54)

$$w_i(x) = \phi[x_0] + \frac{x - x_0}{\phi[x_0, x_1]} + \frac{x - x_1}{\phi[x_0, x_1, x_2]} + \cdots + \frac{x - x_{i-1}}{\phi[x_0, x_1, \ldots, x_i]}$$

è funzione razionale di interpolazione della $f(x)$ sui nodi $x_0, \ldots, x_{i-1}$. Dalla (40) si ha che, indicata con

$$w_i(x) = \frac{p_i(x)}{q_i(x)} \tag{56}$$

la i-esima frazione parziale, i polinomi $p_i(x)$ e $q_i(x)$ soddisfano alle relazioni ricorrenti

$$\begin{aligned} &p_{-1}(x) = 1, \quad q_{-1}(x) = 0, \quad p_0(x) = f(x_0), \quad q_0(x) = 1, \\ &p_i(x) = \phi[x_0, x_1, \ldots, x_i]\, p_{i-1}(x) + (x - x_{i-1}) p_{i-2}(x), \\ &q_i(x) = \phi[x_0, x_1, \ldots, x_i]\, q_{i-1}(x) + (x - x_{i-1}) q_{i-2}(x). \end{aligned} \tag{57}$$

Se non intervengono elisioni dei termini di grado massimo, dalle (57) segue che

$$\begin{aligned} &\text{se } i = 2s \qquad \text{grado di } p_i(x) = \text{ grado di } q_i(x) = s, \\ &\text{se } i = 2s + 1 \qquad \text{grado di } p_i(x) = s + 1 \text{ e grado di } q_i(x) = s, \end{aligned} \tag{58}$$

per cui la funzione razionale di interpolazione che si ottiene per mezzo della frazione continua di Thiele è quella che ha uguali i gradi del numeratore e del denominatore se n è pari, ed ha il grado del numeratore di 1 superiore a quello del denominatore se n è dispari.

Nel calcolo di $p_i(x)$ con i pari e di $q_i(x)$ con i dispari può accadere che i termini di grado massimo si elidano, per cui il grado di $p_i(x)$ o di $q_i(x)$ sia minore di s. In generale però i gradi dei polinomi $p_j(x)$ e $q_j(x)$ con $j > i$ non risultano alterati.

5.48 Esempio. Per determinare la funzione razionale

$$w(x) = \frac{a_3x^3 + a_2x^2 + a_1x + a_0}{b_3x^3 + b_2x^2 + b_1x + b_0}$$

di interpolazione della funzione $f(x)$ che assume i valori

x	0	1	2	3	4	5	6
$f(x)$	0	$\frac{1}{2}$	$\frac{2}{9}$	$\frac{3}{28}$	$\frac{4}{65}$	$\frac{5}{126}$	$\frac{6}{217}$

si costruisce la tabella delle differenze inverse

0	0						
		2					
1	$\frac{1}{2}$		$\frac{1}{7}$				
		9		$-\frac{91}{6}$			
2	$\frac{2}{9}$		$\frac{1}{13}$		$-\frac{6}{35}$		
		28		-21		$\frac{1435}{6}$	
3	$\frac{3}{28}$		$\frac{1}{21}$		$-\frac{48}{287}$		$\frac{1}{35}$
		65		$-\frac{217}{8}$		$\frac{1645}{6}$	
4	$\frac{4}{65}$		$\frac{1}{31}$		$-\frac{54}{329}$		
		126		$-\frac{301}{9}$			
5	$\frac{5}{126}$		$\frac{1}{43}$				
		217					
6	$\frac{6}{217}$						

da cui si ottiene

$$w(x) = \frac{x}{2} + \frac{x-1}{1/7} + \frac{x-2}{(-91/6)} + \frac{x-3}{(-6/35)} + \frac{x-4}{1435/6} + \frac{x-5}{1/35}\,.$$

Le successive frazioni parziali sono

$$w_1(x) = \frac{x}{2}\,, \quad \text{che interpola su 0 e 1,}$$

$$w_2(x) = \frac{x}{7x-5}\,, \quad \text{che interpola su } 0, 1 \text{ e } 2,$$

$$w_3(x) = \frac{-6x^2+25x}{79x-41}\,, \quad \text{che interpola su } 0, 1, 2 \text{ e } 3,$$

$$w_4(x) = \frac{x^2-10x}{-35x^2+51x-34}\,, \quad \text{che interpola su } 0, 1, 2, 3 \text{ e } 4,$$

$$w_5(x) = \frac{x^3-15x^2+85x}{226x^2-289x+205}\,, \quad \text{che interpola su } 0, 1, 2, 3, 4 \text{ e } 5,$$

$$w(x) = w_6(x) = \frac{x}{x^3+1}\,, \quad \text{che interpola su } 0, 1, 2, 3, 4, 5 \text{ e } 6. \qquad \blacksquare$$

12. Differenze reciproche

L'analogia fra le differenze divise e quelle inverse si perde quando si passa a considerare la proprietà di simmetria espressa dal teorema 5.20: nel caso delle differenze inverse la simmetria riguarda solo gli ultimi due argomenti, ma in generale non gli altri. La proprietà di simmetria vale invece per le differenze reciproche della $f(x)$, così definite.

5.49 Definizione. Si chiama *differenza reciproca* di ordine k della funzione $f(x)$ relativa ai punti $x_i, i = 0, \ldots, k-1$, a due a due distinti, la funzione $\rho[x_0, x_1, \ldots, x_{k-1}, x]$ definita ricorsivamente nel modo seguente

$$\begin{aligned}
&\text{per } k = 0 \quad && \rho[x] = f(x),\\
&\text{per } k = 1 \quad && \rho[x_0, x] = \frac{x - x_0}{\rho[x] - \rho[x_0]}\,,\\
&\text{per } k \geq 2 \quad && \rho[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] = \rho[x_0, x_1, \ldots, x_{k-2}] \qquad (59)\\
& && \quad + \frac{x - x_{k-1}}{\rho[x_0, x_1, \ldots, x_{k-2}, x] - \rho[x_0, x_1, \ldots, x_{k-2}, x_{k-1}]}\,. \qquad \blacksquare
\end{aligned}$$

Dal confronto con la definizione 5.43 risulta che $\rho[x] = \phi[x]$ e $\rho[x_0, x] = \phi[x_0, x]$. Il seguente teorema dà la relazione che lega fra di loro le differenze inverse e le differenze reciproche di ordine più elevato.

5.50 Teorema. *Per $k \geq 2$ vale la relazione*

$$\begin{aligned}\phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] = \rho[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] \\ - \rho[x_0, x_1, \ldots, x_{k-2}].\end{aligned} \tag{60}$$

Dim. Si procede per induzione su k: per $k = 2$ si ha

$$\begin{aligned}\rho[x_0, x_1, x] &= \rho[x_0] + \frac{x - x_1}{\rho[x_0, x] - \rho[x_0, x_1]} = \rho[x_0] + \frac{x - x_1}{\phi[x_0, x] - \phi[x_0, x_1]} \\ &= \rho[x_0] + \phi[x_0, x_1, x].\end{aligned}$$

Per $k > 2$ per l'ipotesi induttiva si ha

$$\rho[x_0, \ldots, x_{k-3}, x_{k-2}, x] = \rho[x_0, \ldots, x_{k-3}] + \phi[x_0, x_1, \ldots, x_{k-2}, x], \tag{61}$$

per cui

$$\begin{aligned}\rho[x_0, \ldots, x_{k-3}, x_{k-2}, x_{k-1}] = \rho[x_0, \ldots, x_{k-3}] \\ + \phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}].\end{aligned} \tag{62}$$

Sottraendo la (62) dalla (61) e utilizzando la definizione 5.49 si ha

$$\begin{aligned}\rho[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] = \rho[x_0, x_1, \ldots, x_{k-2}] \\ + \frac{x - x_{k-1}}{\phi[x_0, x_1, \ldots, x_{k-2}, x] - \phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}]},\end{aligned}$$

da cui segue la (60) tenendo conto della definizione 5.43. ∎

5.51 Teorema. *La differenza reciproca di ordine k*

$$\rho[x_0, x_1, \ldots, x_k]$$

è funzione simmetrica dei suoi argomenti $x_0, x_1, \ldots, x_k$, cioè è invariante comunque vengano permutati i suoi argomenti.

Dim. La dimostrazione viene qui fatta sotto l'ipotesi semplificativa che nei nodi $x_0, \ldots, x_k$ esista la funzione razionale di interpolazione della $f(x)$ e che nella costruzione dei polinomi $p_i(x)$ e $q_i(x)$ della (56) non vi siano elisioni dei termini di grado massimo. Per la dimostrazione nel caso generale si veda [16].

Indicati con a_i e b_i i coefficienti dei termini di grado massimo dei polinomi $p_i(x)$ e $q_i(x)$, si dimostra per induzione che, se non si moltiplicano numeratore e denominatore per costante, allora

$$\begin{aligned}&\text{per } i = 2s \quad \text{è} \quad a_i = \rho[x_0, x_1, \ldots, x_i] \text{ e } b_i = 1, \\ &\text{per } i = 2s + 1 \quad \text{è} \quad a_i = 1 \text{ e } b_i = \rho[x_0, x_1, \ldots, x_i].\end{aligned} \tag{63}$$

Infatti per $i = 0$ e $i = 1$ queste relazioni sono immediate. Per $i = 2s > 1$ si ha dalla (58) che i polinomi $p_{i-1}(x)$ e $(x - x_{i-1})p_{i-2}(x)$ hanno lo stesso grado minore o uguale a s, e quindi per la (57) e per l'ipotesi induttiva il primo coefficiente di $p_i(x)$ è dato da

$$a_i = \phi[x_0, x_1, \ldots, x_i]a_{i-1} + a_{i-2} = \phi[x_0, x_1, \ldots, x_i] + \rho[x_0, x_1, \ldots, x_{i-2}],$$

da cui per la (60) si ottiene la prima delle (63). Le altre relazioni si verificano in modo del tutto analogo.

Permutando in un qualsiasi modo i punti $x_0, x_1, \ldots, x_i$, la i-esima frazione parziale $w_i(x)$ in (56), che è la funzione razionale di interpolazione di $f(x)$ nei punti $x_0, x_1, \ldots, x_i$, non varia. Ne segue che anche i primi coefficienti dei due polinomi $p_i(x)$ e $q_i(x)$ non variano permutando i punti $x_0, x_1, \ldots, x_i$ e quindi non variano le differenze reciproche $\rho[x_0, x_1, \ldots, x_i]$. ∎

La simmetria delle differenze reciproche consentirebbe di rappresentare la funzione razionale di interpolazione non tenendo conto dell'ordinamento dei punti nella costruzione della tabella delle differenze, purché nella (54) si esprimano i coefficienti mediante le differenze reciproche con la (60). In pratica, però, le differenze reciproche, data la maggiore complessità di costruzione, non vengono di solito usate nei casi in cui i punti x_i, $i = 0, \ldots, m + n$, sono distinti. Ma se alcuni o tutti i punti x_i coincidono, allora la proprietà di simmetria delle differenze reciproche consente di rappresentare ugualmente la funzione razionale di interpolazione, in modo analogo a quanto fatto nel paragrafo 8 nel caso delle differenze divise.

Per semplicità si considera solo il caso in cui tutti i punti x_i siano coincidenti nell'unico punto x, ma sviluppi analoghi a quelli che si otterranno valgono anche quando solo alcuni dei punti x_i coincidono. Qui però, a differenza del caso polinomiale, anche se $f(x) \in C^k[a, b]$, non è detto in generale che la funzione

$$\rho[\underbrace{x, \ldots, x}_{k+1 \text{ volte}}] = \lim_{\substack{x_i \to x \\ i=0,\ldots,k}} \rho[x_0, \ldots, x_k]$$

sia definita per tutti i punti dell'intervallo $[a, b]$: infatti la funzione può non esistere nei punti in cui si annulla almeno un denominatore. Negli sviluppi che seguono per ogni k si indicherà con Ω_k il massimo sottoinsieme aperto, che si suppone esista non vuoto, di $[a, b]^{k+1}$ in cui la funzione $\rho[x_0, \ldots, x_k]$ è estendibile per continuità ed ha derivate prime continue rispetto a tutti gli argomenti. Se

$$(\underbrace{x, \ldots, x}_{k+1 \text{ volte}}) \in \Omega_k,$$

per semplicità si userà la notazione

$$\rho^{(0)}[x] = \rho[x] \quad \text{e} \quad \rho^{(k)}[x] = \rho[\underbrace{x, \ldots, x}_{k+1 \text{ volte}}].$$

5.52 Teorema. *Sia $f(x) \in C^k[a,b]$, per un intero $k \geq 1$.*

a) Se per $k = 1$ è $f'(x) \neq 0$, allora $[x,x] \in \Omega_1$ e vale $\rho^{(1)}[x] = 1/f'(x)$.

b) Se per $k > 1$ è

$$(\underbrace{x, \ldots, x}_{k-1 \text{ volte}}) \in \Omega_{k-2}, \quad (\underbrace{x, \ldots, x}_{k \text{ volte}}) \in \Omega_{k-1}, \quad \text{e} \quad \frac{d}{dx}\rho^{(k-1)}[x] \neq 0,$$

allora

$$(\underbrace{x, \ldots, x}_{k+1 \text{ volte}}) \in \Omega_k$$

e vale

$$\rho^{(k)}[x] = \rho^{(k-2)}[x] + \frac{k}{\dfrac{d}{dx}\rho^{(k-1)}[x]}\,. \tag{64}$$

Dim. a) Per la (59) per $|\epsilon| \neq 0$ e sufficientemente piccolo, è $[x, x+\epsilon] \in \Omega_1$ e

$$\rho[x, x+\epsilon] = \frac{\epsilon}{f(x+\epsilon) - f(x)}\,.$$

Per ipotesi il limite per $\epsilon \to 0$ del secondo membro esiste ed è

$$\rho^{(1)}[x] = \lim_{\epsilon \to 0} \rho[x, x+\epsilon] = \lim_{\epsilon \to 0} \frac{\epsilon}{f(x+\epsilon) - f(x)} = \frac{1}{f'(x)}\,.$$

b) Per $|\epsilon| \neq 0$ e sufficientemente piccolo è

$$(\underbrace{x, \ldots, x}_{k-1 \text{ volte}}, x+\epsilon) \in \Omega_{k-1}$$

e

$$\lim_{\epsilon \to 0} \frac{\rho[\overbrace{x, \ldots, x}^{k-1 \text{ volte}}, x+\epsilon] - \rho[\overbrace{x, \ldots, x}^{k \text{ volte}}]}{\epsilon} = \left.\frac{\partial \rho[x_0, \ldots, x_{k-1}]}{\partial x_{k-1}}\right|_{x_0, \ldots, x_{k-1} = x} \tag{65}$$

Poiché $\rho[x_0, \ldots, x_{k-1}]$ è funzione simmetrica dei suoi argomenti, vale

$$\begin{aligned}\frac{d}{dx}\,\rho[\underbrace{x, \ldots, x}_{k \text{ volte}}] &= \sum_{j=0}^{k-1} \frac{\partial \rho[x_0, \ldots, x_{k-1}]}{\partial x_j}\,\frac{dx_j}{dx}\bigg|_{x_0, \ldots, x_{k-1} = x} \\ &= k\,\frac{\partial \rho[x_0, \ldots, x_{k-1}]}{\partial x_{k-1}}\bigg|_{x_0, \ldots, x_{k-1} = x},\end{aligned}$$

in quanto $\dfrac{dx_j}{dx} = 1$ perché $x_j(x) = x$. Sostituendo nella (65) si ha per la (59)

$$\begin{aligned}\rho[\underbrace{x, \ldots, x}_{k+1 \text{ volte}}] &= \rho[\underbrace{x, \ldots, x}_{k-1 \text{ volte}}] + \lim_{\epsilon \to 0} \frac{\epsilon}{\rho[\underbrace{x, \ldots, x}_{k-1 \text{ volte}}, x + \epsilon] - \rho[\underbrace{x, \ldots, x}_{k \text{ volte}}]} \\ &= \rho[\underbrace{x, \ldots, x}_{k-1 \text{ volte}}] + \frac{k}{\dfrac{d}{dx}\,\rho[\underbrace{x, \ldots, x}_{k \text{ volte}}]},\end{aligned}$$

da cui si ottiene la (64). ∎

La (64) consente quindi, a partire da

$$\rho^{(0)}[x] = f(x) \quad \text{e} \quad \rho^{(1)}[x] = \frac{1}{f'(x)},$$

di costruire le differenze reciproche di ordine più elevato. Inoltre, per la (60), se $\rho^{(k-2)}[x]$ e $\rho^{(k)}[x]$ esistono, allora esiste anche

$$\phi^{(k)}[x] = \phi[\underbrace{x, \ldots, x}_{k+1 \text{ volte}}] = \lim_{\substack{x_i \to x \\ i=0,\ldots,k}} \phi[x_0, \ldots, x_k]$$

e vale

$$\phi^{(k)}[x] = \rho^{(k)}[x] - \rho^{(k-2)}[x] = \frac{k}{\dfrac{d}{dx}\,\rho^{(k-1)}[x]}. \tag{66}$$

È così possibile scrivere la frazione continua (54) nel caso in cui tutti i punti x_i coincidono con il punto x_0. La funzione razionale che si ottiene, per il teorema 5.45, diventa osculatoria.

5.53 Teorema. *Sia $x_0 \in [a,b]$. Se le differenze inverse $\phi^{(k)}[x_0]$ esistono finite per $k = 0, \dots, m+n$, allora la frazione continua di Thiele di ordine $m+n$*

$$w[x] = f(x_0) + \frac{x - x_0}{\phi^{(1)}[x_0]} + \cdots + \frac{x - x_0}{\phi^{(m+n)}[x_0]} \tag{67}$$

è la funzione razionale che nel punto x_0 assume il valore $f(x_0)$ e le cui derivate fino all'ordine $m+n$ assumono i valori delle corrispondenti derivate della $f(x)$ in x_0. ∎

Si osservi che l'ipotesi di non annullamento dei residui fatta nel teorema 5.45 è qui soddisfatta in quanto $\phi^{(k)}[x_0] \neq 0$ per ogni k.

5.54 Esempio. Per la funzione $f(x) = \mathrm{e}^x$ si ottiene

$$\begin{aligned}
\rho^{(0)}[x] &= \mathrm{e}^x & \phi^{(1)}[x] &= \mathrm{e}^{-x} \\
\rho^{(1)}[x] &= \mathrm{e}^{-x} & \phi^{(2)}[x] &= -2\,\mathrm{e}^x \\
\rho^{(2)}[x] &= -\,\mathrm{e}^x & \phi^{(3)}[x] &= -3\,\mathrm{e}^{-x} \\
\rho^{(3)}[x] &= -2\,\mathrm{e}^{-x} & \phi^{(4)}[x] &= 2\,\mathrm{e}^x \\
&\cdots
\end{aligned}$$

Continuando ricorsivamente si vede che le differenze $\rho^{(k)}[x]$ e $\phi^{(k)}[x]$, ad eccezione di $\phi^{(0)}[x]$, seguono la regola

$$\left.\begin{aligned}
\rho^{(2s)}[x] &= (-1)^s \mathrm{e}^x, \quad & \rho^{(2s+1)}[x] &= (-1)^s (s+1) \mathrm{e}^{-x}, \\
\phi^{(2s)}[x] &= (-1)^s 2\mathrm{e}^x, \quad & \phi^{(2s+1)}[x] &= (-1)^s (2s+1) \mathrm{e}^{-x},
\end{aligned}\right\} \; s = 0, 1, \dots$$

Si ha infatti per induzione dalle (64) e (66)

$$\begin{aligned}
\rho^{(2s+2)}[x] &= \rho^{(2s)}[x] + \frac{2s+2}{\dfrac{d}{dx}\rho^{(2s+1)}[x]} = (-1)^s \mathrm{e}^x + \frac{2s+2}{(-1)^{s+1}(s+1)\mathrm{e}^{-x}} \\
&= (-1)^{s+1} \mathrm{e}^x,
\end{aligned}$$

$$\phi^{(2s+2)}[x] = \frac{2s+2}{\dfrac{d}{dx}\rho^{(2s+1)}[x]} = \frac{2s+2}{(-1)^{s+1}(s+1)\mathrm{e}^{-x}} = (-1)^{s+1} 2\mathrm{e}^x,$$

$$\begin{aligned}
\rho^{(2s+3)}[x] &= \rho^{(2s+1)}[x] + \frac{2s+3}{\dfrac{d}{dx}\rho^{(2s+2)}[x]} = (-1)^s (s+1) \mathrm{e}^{-x} + \frac{2s+3}{(-1)^{s+1}\mathrm{e}^x} \\
&= (-1)^{s+1} (s+2) \mathrm{e}^{-x},
\end{aligned}$$

$$\phi^{(2s+3)}[x] = \frac{2s+3}{\dfrac{d}{dx}\rho^{(2s+2)}[x]} = \frac{2s+3}{(-1)^{s+1}\mathrm{e}^{x}} = (-1)^{s+1}(2s+3)\mathrm{e}^{-x}.$$

Arrestando il calcolo al $(2s+1)$-esimo termine e assumendo $x_0 = 0$, si ha dalla (67) la frazione continua di Thiele di ordine $2s+1$ di e^x

$$w(x) = 1 + \frac{x}{1} + \frac{x}{(-2)} + \frac{x}{(-3)} + \frac{x}{2} + \cdots + \frac{x}{(-1)^s 2} + \frac{x}{(-1)^s(2s+1)},$$

o, tenendo conto dei segni dei coefficienti,

$$w(x) = 1 + \frac{x}{1} - \frac{x}{2} + \frac{x}{3} - \frac{x}{2} + \cdots - \frac{x}{2} + \frac{x}{2s+1}. \tag{68}$$

Nella figura 5.16 sono riportati, per $x \in [0,1]$, i grafici degli errori assoluti

$$|\mathrm{e}^x - w(x)|$$

nei casi in cui $w(x)$ è di ordine $k = 1, 2, 3, 4$.

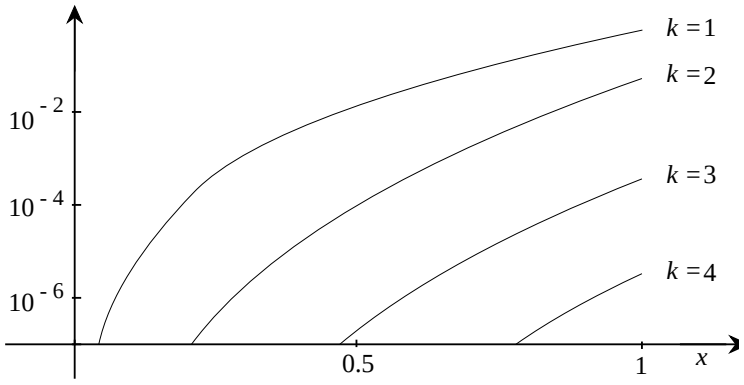


Fig. 5.16 - Grafici degli errori assoluti della formula di Thiele di e^x.

In modo analogo si procede per la funzione $f(x) = \log(1+x)$, per la quale, ad eccezione di $\phi^{(0)}[x] = \log(1+x)$, è

$$\phi^{(2s)}[x] = \frac{2}{s}, \quad \phi^{(2s+1)}[x] = (2s+1)(1+x), \quad s = 0, 1, \ldots$$

Assumendo $x_0 = 0$, si ottiene la frazione continua di Thiele di ordine $2s+1$ di $\log(1+x)$

$$w(x) = \frac{x}{1} + \frac{x}{2} + \frac{x}{3} + \frac{x}{2/2} + \frac{x}{5} + \frac{x}{2/3} + \cdots + \frac{x}{2/s} + \frac{x}{2s+1}. \qquad \blacksquare$$

13. Interpolazione trigonometrica e trasformata discreta di Fourier

Le funzioni trigonometriche $\sin kx$ e $\cos kx$, con k intero, pur non essendo funzioni razionali, vengono ugualmente utilizzate per l'interpolazione in quanto facilmente calcolabili. Le loro proprietà di periodicità e di ortogonalità ne fanno la classe ideale per approssimare funzioni periodiche. Inoltre esse hanno in comune con la classe delle funzioni razionali la proprietà, utilissima per le applicazioni, che le loro derivate e le loro primitive sono ancora funzioni della stessa classe.

5.55 Definizione. Una funzione della forma

$$F(x) = \sum_{j=0}^{m} (\alpha_j \cos jx + \beta_j \sin jx) \tag{69}$$

è detta *polinomio trigonometrico* di grado m. ∎

Data una funzione $f(x)$, reale e definita nell'intervallo $[0, 2\pi)$, il problema dell'*interpolazione trigonometrica* consiste nel determinare il polinomio trigonometrico $F(x)$ di minimo grado tale che

$$F(x_k) = y_k, \quad \text{per} \quad k = 0, \ldots, n-1, \tag{70}$$

dove $y_k = f(x_k)$ sono i valori assunti da $f(x)$ negli n punti equidistanti $x_k = 2k\pi/n$, $k = 0, \ldots, n-1$.

Il problema dell'interpolazione trigonometrica è riconducibile a un problema di interpolazione polinomiale sul cerchio unitario del piano complesso. A questo scopo si introducono le radici n-esime dell'unità.

5.56 Definizione. Sia n un intero. Si definisce *radice n-esima dell'unità* ogni numero complesso z tale che $z^n = 1$. Una radice n-esima ω è detta *primitiva* se l'insieme $\{\omega^i,\ i = 0, \ldots, n-1\}$ è costituito da n elementi distinti. In particolare, indicata con $\boldsymbol{i}$ l'unità immaginaria ($\boldsymbol{i}^2 = -1$), il numero complesso

$$\omega_n = \mathrm{e}^{\boldsymbol{i} 2\pi/n}$$

è una radice primitiva n-esima dell'unità. ∎

Poiché

$$\omega_n^r = \mathrm{e}^{\boldsymbol{i} 2\pi r/n},$$

le successive potenze di ω_n sono nell'ordine tutte e sole le radici n-esime dell'unità, e

$$\omega_n^k = \omega_n^p \quad \text{se e solo se} \quad k \equiv p \pmod{n}.$$

5.57 Teorema. *Vale la relazione di ortogonalità:*

$$\sum_{j=0}^{n-1} \omega_n^{jk} = \begin{cases} n & \text{se } k \equiv 0 \pmod{n}, \\ 0 & \text{altrimenti.} \end{cases}$$

Dim. Se $k \equiv 0 \pmod{n}$, esiste un intero s per cui $k = sn$, e quindi è $\omega_n^{jk} = 1$. Altrimenti, se $k \not\equiv 0 \pmod{n}$, è $\omega_n^k \neq 1$; ponendo $x = \omega_n^k$ e utilizzando la nota relazione

$$(1 + x + x^2 + \ldots + x^{n-1})(1 - x) = 1 - x^n,$$

si ottiene

$$(\sum_{j=0}^{n-1} \omega_n^{jk})(1 - \omega_n^k) = 1 - \omega_n^{nk},$$

e, ricordando che $\omega_n^k \neq 1$ e che $\omega_n^{nk} = 1$, ne segue la tesi. ∎

Si considera ora un problema di interpolazione polinomiale nel campo complesso che, come si vedrà nel teorema 5.59, è equivalente al problema dell'interpolazione trigonometrica: si tratta di determinare i numeri complessi z_j, $j = 0, \ldots, n-1$, coefficienti del polinomio di grado al più $n-1$

$$p(w) = \sum_{j=0}^{n-1} z_j \; w^j, \tag{71}$$

tale che

$$p(\omega_n^k) = y_k, \quad \text{per} \quad k = 0, \ldots, n-1. \tag{72}$$

I coefficienti z_j si ottengono risolvendo il sistema (3)

$$V\mathbf{z} = \mathbf{y},$$

dove $\mathbf{y} = [y_0, y_1, \ldots, y_{n-1}]^T$, $\mathbf{z} = [z_0, z_1, \ldots, z_{n-1}]^T$ e V è la matrice di Vandermonde di ordine n, i cui elementi sono

$$v_{kj} = \omega_n^{kj}, \quad k, j = 0, \ldots, n-1.$$

Per il teorema 5.57 la matrice V soddisfa alla proprietà (si veda l'esercizio 5.54) $V^H V = nI$ (dove con V^H si è indicata la matrice trasposta della matrice i cui elementi sono i coniugati degli elementi di V). Ne segue che $V^{-1} = \dfrac{1}{n} V^H$, e quindi

$$\mathbf{z} = \frac{1}{n} V^H \mathbf{y}. \tag{73}$$

5.58 Definizione. L'applicazione che al vettore $\mathbf{y}$ associa il vettore $\mathbf{z}$ definito in (73) è detta *trasformata discreta di Fourier* e viene generalmente indicata con la sigla DFT, mentre il vettore $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$ è detto *trasformata discreta di Fourier* del vettore $\mathbf{y}$ e verifica la relazione

$$z_j = \frac{1}{n} \sum_{k=0}^{n-1} y_k \, \omega_n^{-jk}, \quad j = 0, \ldots, n-1. \tag{74}$$

L'applicazione che al vettore $\mathbf{z}$ associa il vettore $\mathbf{y}$ è detta *trasformata discreta inversa di Fourier* e viene generalmente indicata con la sigla IDFT, mentre il vettore $\mathbf{y} = \mathrm{IDFT}(\mathbf{z})$ è detto *trasformata discreta inversa di Fourier* del vettore $\mathbf{z}$ e verifica la relazione

$$y_k = \sum_{j=0}^{n-1} z_j \, \omega_n^{jk}, \quad k = 0, \ldots, n-1. \tag{75}$$

∎

Se il vettore $\mathbf{y}$ ha componenti reali, le componenti del vettore $\mathbf{z}$ sono tali che

$$z_0 \text{ è reale e } \quad \overline{z}_j = z_{n-j}, \quad \text{per } j = 1, \ldots, n-1. \tag{76}$$

Infatti dalla (74), essendo $\omega_n^{-nk} = 1$ per k intero, si ha che

$$z_0 = \frac{1}{n} \sum_{k=0}^{n-1} y_k \quad \text{e} \quad z_{n-j} = \frac{1}{n} \sum_{k=0}^{n-1} y_k \, \omega_n^{-(n-j)k} = \frac{1}{n} \sum_{k=0}^{n-1} y_k \, \omega_n^{jk} = \overline{z}_j.$$

Se n è pari, $n = 2m$, anche z_m è reale, in quanto $\overline{z}_m = z_m$.

5.59 Teorema. *Il polinomio trigonometrico di interpolazione della funzione $f(x)$ negli n punti x_k, $k = 0, \ldots, n-1$, è*

$$F_n(x) = \begin{cases} \dfrac{\alpha_0}{2} + \displaystyle\sum_{j=1}^{m-1} (\alpha_j \cos jx + \beta_j \sin jx), & \text{se } n = 2m-1, \\ \dfrac{\alpha_0}{2} + \displaystyle\sum_{j=1}^{m-1} (\alpha_j \cos jx + \beta_j \sin jx) + \dfrac{\alpha_m}{2} \cos mx, & \text{se } n = 2m, \end{cases} \tag{77}$$

in cui i coefficienti α_j, β_j sono reali e sono dati da:

$$\begin{aligned} \alpha_j &= 2 \, \mathrm{Re}(z_j) = \frac{2}{n} \sum_{k=0}^{n-1} f(x_k) \cos jx_k, \\ \beta_j &= -2 \, \mathrm{Im}(z_j) = \frac{2}{n} \sum_{k=0}^{n-1} f(x_k) \sin jx_k, \end{aligned} \tag{78}$$

dove $\mathbf{z}$ è la DFT del vettore $\mathbf{y} = [f(x_0), f(x_1), \ldots, f(x_{n-1})]^T$; in particolare se $n = 2m$, è

$$\alpha_m = 2z_m = \frac{2}{n} \sum_{k=0}^{n-1} (-1)^k f(x_k);$$

e tale polinomio è unico.

Dim. Si esamina dapprima il caso $n = 2m - 1$. In questo caso la (75) per la (76) si può scrivere

$$\begin{aligned} y_k &= \sum_{j=0}^{2m-2} z_j \, \omega_n^{jk} = \sum_{j=0}^{m-1} z_j \, \omega_n^{jk} + \sum_{j=m}^{2m-2} z_j \, \omega_n^{jk} \\ &= \sum_{j=0}^{m-1} z_j \, \omega_n^{jk} + \sum_{j=m}^{2m-2} \overline{z}_{n-j} \, \omega_n^{jk} = \sum_{j=0}^{m-1} z_j \, \omega_n^{jk} + \sum_{p=1}^{m-1} \overline{z}_p \, \omega_n^{(n-p)k} \\ &= z_0 + \sum_{j=1}^{m-1} (z_j \, \omega_n^{jk} + \overline{z}_j \, \omega_n^{-jk}). \end{aligned}$$

Poiché

$$\omega_n^{jk} = \mathrm{e}^{\boldsymbol{i} j x_k} = \cos jx_k + \boldsymbol{i} \sin jx_k,$$

risulta

$$y_k = z_0 + \sum_{j=1}^{m-1} \big[(z_j + \overline{z}_j) \cos jx_k + \boldsymbol{i}(z_j - \overline{z}_j) \sin jx_k \big],$$

e ponendo

$$\alpha_j = z_j + \overline{z}_j, \quad \beta_j = \boldsymbol{i}(z_j - \overline{z}_j), \quad j = 0, \ldots, n-1,$$

si ha

$$y_k = \frac{\alpha_0}{2} + \sum_{j=1}^{m-1} (\alpha_j \cos jx_k + \beta_j \sin jx_k).$$

Ne segue che il polinomio trigonometrico della forma (69)

$$F_n(x) = \frac{\alpha_0}{2} + \sum_{j=1}^{m-1} (\alpha_j \cos jx + \beta_j \sin jx)$$

soddisfa alle (70). I coefficienti α_j e β_j, calcolati per mezzo delle (74), sono

$$\alpha_j = z_j + \overline{z}_j = \frac{1}{n}\sum_{k=0}^{n-1} y_k(\omega_n^{-jk} + \omega_n^{jk}) = \frac{1}{n}\sum_{k=0}^{n-1} y_k(\mathrm{e}^{-\boldsymbol{i}jx_k} + \mathrm{e}^{\boldsymbol{i}jx_k})$$
$$= \frac{2}{n}\sum_{k=0}^{n-1} y_k \cos jx_k, \quad \text{per } j = 0, \ldots, n-1,$$
$$\beta_0 = 0, \quad \beta_j = \boldsymbol{i}(z_j - \overline{z}_j) = \frac{2}{n}\sum_{k=0}^{n-1} y_k \sin jx_k \quad \text{per } j = 1, \ldots, n-1.$$

Nel caso in cui $n = 2m$, procedendo in modo analogo, dalla (75) si ha

$$y_k = \sum_{j=0}^{2m-1} z_j\, \omega_n^{jk} = z_0 + \sum_{j=1}^{m-1}(z_j\, \omega_n^{jk} + \overline{z}_j\, \omega_n^{-jk}) + z_m\, \omega_n^{mk}.$$

Poiché $\omega_n^{mk} = \cos mx_k$, il polinomio trigonometrico cercato è della forma

$$F_n(x) = \frac{\alpha_0}{2} + \sum_{j=1}^{m-1}(\alpha_j \cos jx + \beta_j \sin jx) + \frac{\alpha_m}{2}\cos mx.$$

L'unicità del polinomio di interpolazione trigonometrico segue dall'unicità del polinomio (71) che verifica le (72) e dal fatto che la relazione che lega i coefficienti α_j e β_j con gli z_j è biunivoca. ∎

5.60 Esempio. Sia $f(x)$ la funzione continua a tratti ottenuta per estensione periodica di periodo 2π della funzione $f(x) = x$, per $0 \le x < 2\pi$. Assumendo $n = 3$, si ha

x	0	$\frac{2\pi}{3}$	$\frac{4\pi}{3}$
$f(x)$	0	$\frac{2\pi}{3}$	$\frac{4\pi}{3}$

Il polinomio trigonometrico di interpolazione è dato da

$$F_3(x) = \frac{\alpha_0}{2} + \alpha_1 \cos x + \beta_1 \sin x$$

dove

$$\alpha_0 = \frac{2}{3}\left(\frac{2\pi}{3} + \frac{4\pi}{3}\right) = \frac{4\pi}{3},$$
$$\alpha_1 = \frac{2}{3}\left(\frac{2\pi}{3}\cos\frac{2\pi}{3} + \frac{4\pi}{3}\cos\frac{4\pi}{3}\right) = -\frac{2\pi}{3},$$
$$\beta_1 = \frac{2}{3}\left(\frac{2\pi}{3}\sin\frac{2\pi}{3} + \frac{4\pi}{3}\sin\frac{4\pi}{3}\right) = -\frac{2\pi\sqrt{3}}{9},$$

per cui

$$F_3(x) = \frac{2\pi}{3}\left(1 - \cos x - \frac{\sqrt{3}}{3}\,\sin x\right).$$

Nella figura 5.17 sono riportati il grafico su 3 periodi della funzione $f(x)$ (linea spessa) e il grafico del polinomio trigonometrico di interpolazione $F_3(x)$ (linea sottile).

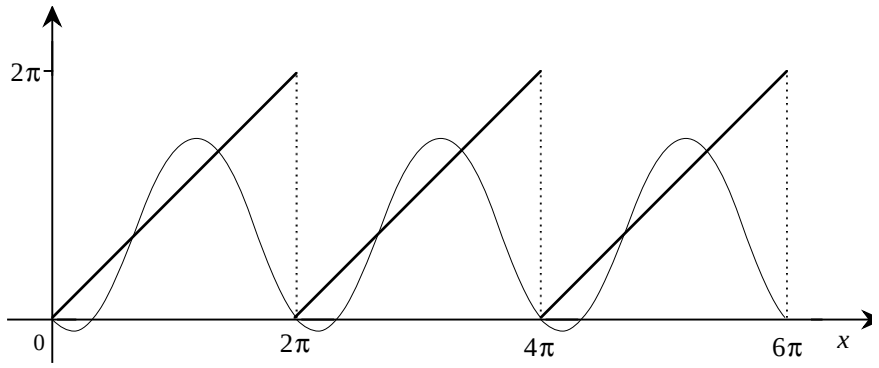


Fig. 5.17 - Grafico del polinomio trigonometrico di interpolazione di $f(x) = x,\ 0 \le x < 2\pi$, per $n = 3$.

Nella figura 5.18 è riportato il grafico del polinomio trigonometrico ottenuto con $n = 8$, i cui coefficienti sono

$$\alpha_0 = \frac{7\pi}{4}\,, \quad \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = -\,\frac{\pi}{4}\,,$$
$$\beta_1 = -\,\frac{\pi}{4}\,(1+\sqrt{2}), \quad \beta_2 = -\,\frac{\pi}{4}\,, \quad \beta_3 = \frac{\pi}{4}\,(1-\sqrt{2}).$$

∎

Se nell'intervallo $[0, 2\pi)$ la funzione $f(x)$ è simmetrica, oppure antisimmetrica, rispetto al punto π i coefficienti α_j e β_j del polinomio trigonometrico di interpolazione si semplificano nel modo seguente:

a) se $f(x)$ è simmetrica, allora

$$\alpha_j = \begin{cases} \dfrac{2}{n}\left[f(x_0) + 2\displaystyle\sum_{k=1}^{m-1} f(x_k)\cos jx_k\right], & \text{se } n = 2m-1, \\[2ex] \dfrac{2}{n}\left[f(x_0) + (-1)^j f(x_m) + 2\displaystyle\sum_{k=1}^{m-1} f(x_k)\cos jx_k\right], & \text{se } n = 2m, \end{cases}$$

$$\beta_j = 0,$$

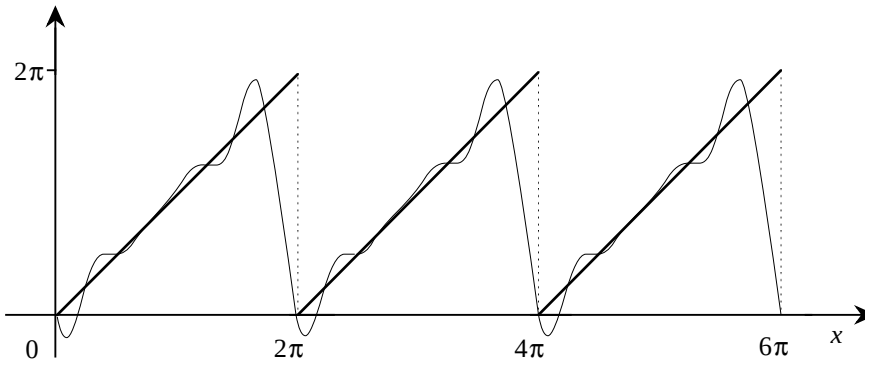


Fig. 5.18 - Grafico del polinomio trigonometrico di interpolazione di $f(x) = x,\ 0 \le x < 2\pi$, per $n = 8$.

b) se $f(x)$ è antisimmetrica (e quindi è $f(\pi) = 0$), allora

$$\alpha_j = \frac{2}{n}\, f(x_0),$$
$$\beta_j = \frac{4}{n} \sum_{k=1}^{m-1} f(x_k) \sin jx_k.$$

La dimostrazione delle relazioni precedenti tiene conto del fatto che, se $f(x)$ è simmetrica, allora $f(x_k) = f(x_{n-k})$ e se $f(x)$ è antisimmetrica, allora $f(x_k) = -f(x_{n-k})$ ed inoltre del fatto che

$$\cos jx_k = \cos jx_{n-k} \quad \text{e} \quad \sin jx_k = -\sin jx_{n-k}.$$

Per esempio se la funzione $f(x)$ è simmetrica e $n = 2m - 1$, allora

$$\begin{aligned}\beta_j &= \frac{2}{n}\left[\sum_{k=1}^{m-1} f(x_k) \sin jx_k + \sum_{k=m}^{2m-2} f(x_k) \sin jx_k\right] \\ &= \frac{2}{n}\left[\sum_{k=1}^{m-1} f(x_k) \sin jx_k + \sum_{k=1}^{m-1} f(x_{n-k}) \sin jx_{n-k}\right] = 0.\end{aligned}$$

Le altre relazioni si dimostrano in modo analogo.

5.61 Esempio. Sia $f(x)$ la funzione ottenuta per estensione periodica di periodo 2π della funzione $f(x) = x(2\pi - x)$, per $0 \le x < 2\pi$. Assumendo $n = 3$, si ha

x	0	$\frac{2\pi}{3}$	$\frac{4\pi}{3}$
$f(x)$	0	$\frac{8\pi^2}{9}$	$\frac{8\pi^2}{9}$

Poiché la funzione $f(x)$ è simmetrica rispetto al punto π il polinomio trigonometrico di interpolazione è dato da

$$F_3(x) = \frac{\alpha_0}{2} + \alpha_1 \cos x$$

dove

$$\alpha_0 = 2\,\frac{2}{3}\,\frac{8\pi^2}{9} = \frac{32\pi^2}{27}\,, \quad \alpha_1 = 2\,\frac{2}{3}\,\frac{8\pi^2}{9}\,\cos\,\frac{2\pi}{3} = -\,\frac{16\pi^2}{27}\,,$$

per cui

$$F_3(x) = \frac{16\pi^2}{27}(1 - \cos x).$$

Nella figura 5.19 sono riportati il grafico su 3 periodi della funzione $f(x)$ (linea spessa) e il grafico del polinomio trigonometrico di interpolazione $F_3(x)$ (linea sottile).

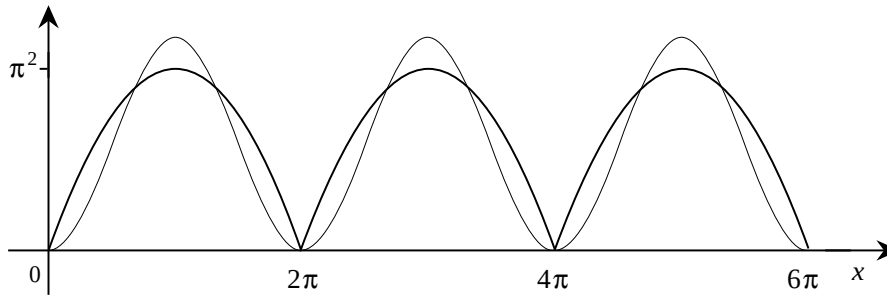


Fig. 5.19 - Grafico del polinomio trigonometrico di interpolazione di $f(x) = x(2\pi - x),\ 0 \le x < 2\pi$, per $n = 3$.

Nella figura 5.20 è riportato il grafico del polinomio trigonometrico ottenuto con $n = 8$, i cui coefficienti sono

$$\alpha_0 = \frac{21\pi^2}{16}\,, \quad \alpha_1 = -\,\frac{\pi^2}{8}\,(2+\sqrt{2}), \quad \alpha_2 = -\,\frac{\pi^2}{8}\,, \quad \alpha_3 = -\,\frac{\pi^2}{8}\,(2-\sqrt{2}),$$
$$\alpha_4 = -\,\frac{\pi^2}{16}\,.$$

∎

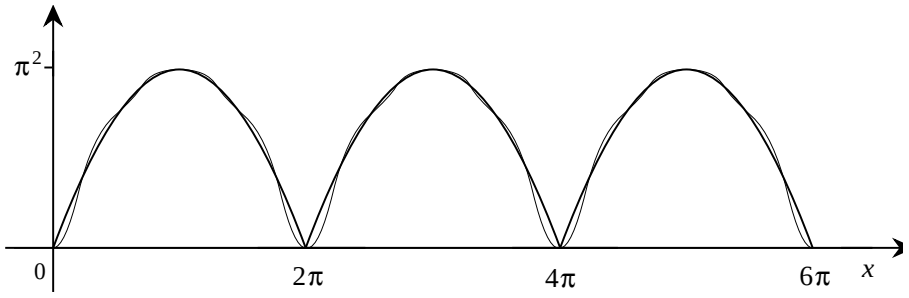


Fig. 5.20 - Grafico del polinomio trigonometrico di interpolazione di $f(x) = x(2\pi - x),\ 0 \le x < 2\pi$, per $n = 8$.

Se la funzione $f(x)$ è definita sull'intervallo $[0, \pi]$, è sempre possibile estenderla all'intervallo $[0, 2\pi)$ in modo da ottenere una funzione simmetrica rispetto al punto π e un corrispondente polinomio trigonometrico di interpolazione di soli coseni. Se $y_k = f(x_k)$ per $x_k = \dfrac{k\pi}{n+1}$, $k = 0, \ldots, n+1$, si estende la funzione all'intervallo $[0, 2\pi)$ definendo i valori $f(x_k) = y_{2n+2-k}$ per $k = n+2, \ldots, 2n+1$. Dalla (77) risulta che il polinomio di interpolazione è

$$F_{2n+2}(x) = \frac{\alpha_0}{2} + \sum_{j=1}^{n} \alpha_j \cos jx + \frac{\alpha_{n+1}}{2} \cos(n+1)x, \tag{79}$$

con

$$\alpha_j = \frac{1}{n+1} \left[y_0 + 2 \sum_{k=1}^{n} y_k \cos jx_k + (-1)^j y_{n+1} \right], \quad j = 0, \ldots, n+1. \tag{80}$$

Un'importante proprietà di cui gode l'interpolazione trigonometrica è quella della convergenza, per n che tende all'infinito, del polinomio $F_n(x)$ alla funzione $f(x)$, se sono soddisfatte certe ipotesi di regolarità. Vale infatti il seguente teorema, per la cui dimostrazione si veda l'esercizio 5.64.

5.62 Teorema. *Sia $f(x)$ una funzione periodica di periodo 2π e derivabile due volte con continuità. Allora per ogni $\epsilon > 0$ esiste un polinomio trigonometrico di interpolazione $F_n(x)$ tale che*

$$|f(x) - F_n(x)| \le \epsilon \quad \textit{per ogni} \quad x. \qquad \blacksquare$$

Il calcolo del valore $F_n(x)$ in un punto x di un polinomio trigonometrico, di cui siano noti i coefficienti α_j e β_j, sembra a prima vista assai più complesso di quello di una funzione razionale. In realtà un'espressione

della forma (77) può essere calcolata con una sola valutazione di funzione trigonometrica. Poiché valgono le seguenti relazioni:

$$\begin{aligned}
&\cos^2 x = 1 - \sin^2 x,\\
&\sin jx = \sin x \cos(j-1)x + \cos x \sin(j-1)x,\\
&\cos jx = \cos x \cos(j-1)x - \sin x \sin(j-1)x,
\end{aligned}$$

nel caso di n dispari, $n = 2m-1$, il valore di $F_n(x)$ può essere calcolato con il seguente algoritmo, in cui il segno di c_1 viene scelto in base al valore di x:

$$\begin{aligned}
&s_1 = \sin x, \quad c_1 = \pm\sqrt{1 - s_1^2}, \quad F_1 = \frac{\alpha_0}{2} + \alpha_1 c_1 + \beta_1 s_1,\\
&\left.\begin{aligned}
&s_j = s_1 c_{j-1} + c_1 s_{j-1},\\
&c_j = c_1 c_{j-1} - s_1 s_{j-1},\\
&F_j = F_{j-1} + \alpha_j c_j + \beta_j s_j,
\end{aligned}\right\} \quad \text{per } j = 2, 3. \ldots, m-1,\\
&F_n(x) = F_{m-1},
\end{aligned}$$

il cui costo computazionale, a meno di costanti additive, è di $4m$ addizioni, $6m$ moltiplicazioni, più una estrazione di radice quadrata e una valutazione di $\sin x$. Il calcolo di queste funzioni trascendenti può essere trascurato nel conto totale delle operazioni, perché esse richiedono un numero di operazioni aritmetiche indipendente da m. Nel caso di n pari, l'algoritmo è assai simile e ha lo stesso costo computazionale. Infine nel caso che il polinomio di interpolazione sia di soli coseni, il costo computazionale può essere ulteriormente ridotto (si veda l'esercizio 5.61).

I coefficienti α_j e β_j, dati dalle (78), nella pratica sono calcolati per mezzo della DFT del vettore $\mathbf{y}$, il cui calcolo, per la particolare struttura della matrice V, richiede meno delle n^2 operazioni sufficienti per la moltiplicazione di una matrice di ordine n per un vettore. Per il calcolo della IDFT e della DFT esistono infatti algoritmi molto efficienti, detti *algoritmi FFT (Fast Fourier Transform)*, che hanno un costo computazionale dell'ordine di $n \log_2 n$. Per semplicità si esamina il caso in cui n è potenza di 2.

5.63 Teorema. *Sia $n = 2^s$; il costo computazionale del calcolo della IDFT di un vettore $\mathbf{z}$ di ordine n o del calcolo della DFT di un vettore $\mathbf{y}$ di ordine n, a meno di termini di ordine inferiore, è di $(n/2)\log_2 n$ moltiplicazioni fra numeri complessi e $n \log_2 n$ addizioni fra numeri complessi, non contando il calcolo delle n radici n-esime dell'unità.*

Dim. Posto $n = 2m$, per la IDFT($\mathbf{z}$) si ha dalla (75)

$$\begin{aligned} y_k &= \sum_{j=0}^{n-1} z_j\, \omega_n^{jk} = \sum_{j \text{ pari}} z_j\, \omega_n^{jk} + \sum_{j \text{ dispari}} z_j\, \omega_n^{jk} \\ &= \sum_{p=0}^{m-1} z_{2p}\, \omega_n^{2kp} + \sum_{p=0}^{m-1} z_{2p+1}\, \omega_n^{k(2p+1)}. \end{aligned}$$

Ponendo $z'_p = z_{2p}$ e $z''_p = z_{2p+1}$, $p = 0, 1, \ldots, m-1$, si ha

$$y_k = \sum_{p=0}^{m-1} z'_p\, \omega_n^{2kp} + \sum_{p=0}^{m-1} z''_p\, \omega_n^{k(2p+1)}.$$

Tenendo presente che $\omega_n^{2p} = (\omega_{n/2})^p = \omega_m^p$, è

$$y_k = \sum_{p=0}^{m-1} z'_p\, \omega_m^{kp} + \omega_n^k \sum_{p=0}^{m-1} z''_p\, \omega_m^{kp}, \quad k = 0, \ldots, n-1. \tag{81}$$

Posto $\mathbf{y}' = \text{IDFT}(\mathbf{z}')$ e $\mathbf{y}'' = \text{IDFT}(\mathbf{z}'')$, cioè

$$y'_q = \sum_{p=0}^{m-1} z'_p\, \omega_m^{qp}, \quad y''_q = \sum_{p=0}^{m-1} z''_p\, \omega_m^{qp}, \quad q = 0, \ldots, m-1,$$

dalla (81) segue che i primi m elementi della trasformata sono dati da

$$y_q = y'_q + \omega_n^q y''_q, \quad q = 0, \ldots, m-1. \tag{82}$$

Per calcolare i rimanenti m elementi della trasformata, poiché $\omega_n^m = -1$ e $\omega_m^{q+m} = \omega_m^q$, dalla (81) segue

$$\begin{aligned} y_{q+m} &= \sum_{p=0}^{m-1} z'_p\, \omega_m^{(q+m)p} + \omega_n^{q+m} \sum_{p=0}^{m-1} z''_p\, \omega_m^{(q+m)p} \\ &= \sum_{p=0}^{m-1} z'_p\, \omega_m^{qp} + \omega_n^{q+m} \sum_{p=0}^{m-1} z''_p\, \omega_m^{qp} \\ &= y'_q - \omega_n^q y''_q, \quad q = 0, \ldots, m-1. \end{aligned} \tag{83}$$

La trasformata di ordine n può quindi essere effettuata con 2 trasformate di ordine $n/2$ più $n/2$ moltiplicazioni e n addizioni. Poiché la trasformata di ordine 1 non richiede operazioni, si possono scrivere le seguenti relazioni

di ricorrenza per il numero di addizioni $A(n)$ e di moltiplicazioni $M(n)$ occorrenti per il calcolo della trasformata di ordine n

$$A(1) = 0, \quad A(n) = 2A(n/2) + n, \tag{84}$$

$$M(1) = 0, \quad M(n) = 2M(n/2) + n/2.$$

Posto $t_s = A(n)$, dove $s = \log_2 n$, dalla (84) si ottiene l'equazione alle differenze

$$t_0 = 0, \quad t_s = 2t_{s-1} + 2^s,$$

la cui soluzione è data da

$$t_s = s2^s,$$

da cui $A(n) = n \log_2 n$. Analogamente si ottiene $M(n) = n/2 \log_2 n$. Si procede nello stesso modo per la DFT($\mathbf{y}$), eseguendo solo al termine le divisioni per n. ∎

Delle m radici n-esime dell'unità presenti nelle (82) e (83), basta calcolare le prime $m/4$, in quanto le altre si ottengono per simmetria.

L'algoritmo per il calcolo della IDFT($\mathbf{z}$) che si ricava dalla dimostrazione del teorema 5.63 non è direttamente implementabile con un linguaggio che non ammetta la ricorsione, e comunque richiede un notevole ingombro di memoria a causa del procedimento ricorsivo. È però possibile costruire algoritmi non ricorsivi che sfruttano ancora le (82) e (83). Fra questi il più noto (per n potenza di 2) è quello di *Cooley e Tukey*, che verrà ora esposto.

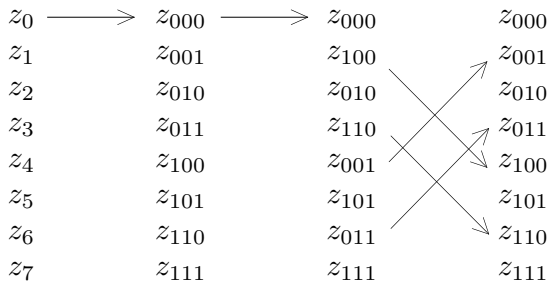
Dalla dimostrazione del teorema 5.63 risulta che la IDFT di ordine n viene calcolata mediante due IDFT di ordine $n/2$, che a loro volta vengono calcolate mediante IDFT di ordine $n/4$, e così via fino alle trasformate di ordine 2. Percorrendo il cammino inverso, si parte dalle IDFT di ordine 2 e si calcolano successivamente le IDFT di ordine $4, \ldots, n/2, n$. La difficoltà del procedimento consiste nel determinare il giusto ordinamento delle componenti del vettore $\mathbf{z}$, a cui applicare le IDFT di ordine più basso. Ad esempio, nel caso $n = 8$ i vettori che vengono considerati successivamente possono essere rappresentati secondo lo schema

$$\begin{array}{cccccccc}
 & & & [z_0, z_1, z_2, z_3, z_4, z_5, z_6, z_7] & & & & \\
 & & & \swarrow \quad \searrow & & & & \\
 & & [z_0, z_2, z_4, z_6] & & & & [z_1, z_3, z_5, z_7] & \\
 & & \swarrow \quad \searrow & & & & \swarrow \quad \searrow & \\
 & [z_0, z_4] & & [z_2, z_6] & & [z_1, z_5] & & [z_3, z_7]
\end{array}$$

Quindi per applicare correttamente il procedimento a partire dalle trasformate di ordine 2, è necessario riordinare il vettore $\mathbf{z}$ nel vettore

$$\widehat{\mathbf{z}} = [z_0, z_4, z_2, z_6, z_1, z_5, z_3, z_7],$$

in modo da considerare ad ogni passo sottosequenze formate da elementi con indice pari o sottosequenze formate da elementi con indice dispari. Si può facilmente verificare che il riordinamento richiesto può essere realizzato mediante la seguente procedura, indicata come *bit reversal*: gli indici j delle componenti z_j vengono rappresentati con notazione in base 2, utilizzando $\log_2 n$ bit, poi questi indici vengono scritti in ordine inverso e il vettore $\mathbf{z}$ riordinato secondo questi nuovi indici. Nel caso visto precedentemente di $n = 8$ risulta



Vengono così scambiati di posto fra loro z_1 e z_4, z_3 e z_6, ottenendo il vettore $\widehat{\mathbf{z}}$ (nell'implementazione gli scambi non vengono fatti sulle componenti del vettore $\mathbf{z}$, ma sulle componenti di un vettore ausiliario di indici). Riapplicando la procedura bit reversal al vettore $\widehat{\mathbf{z}}$ si riottiene il vettore $\mathbf{z}$. In altri termini la permutazione di indici descritta è *involutoria*.

Il passo successivo consiste nell'applicazione delle (82) e (83). Le radici n-esime dell'unità possono essere calcolate tutte all'inizio, a partire da $\omega_n = \cos\dfrac{2\pi}{n} + \boldsymbol{i}\sin\dfrac{2\pi}{n}$, sfruttando le formule trigonometriche di addizione. Questo procedimento richiede però la memorizzazione di $n/8$ numeri complessi. In pratica conviene calcolare ogni volta le radici n-esime che servono, così è sufficiente memorizzare solo il vettore $\mathbf{z}$, che viene modificato durante l'esecuzione e che al termine è uguale a IDFT($\mathbf{z}$). Si tenga conto però del fatto che il vettore $\mathbf{z}$ deve essere definito per numeri complessi anche se i dati iniziali sono reali.

5.64 Algoritmo FFT di Cooley e Tukey ($n = 2^s$, il vettore $\mathbf{z}$ di n componenti complesse deve essere già riordinato con il bit reversal, il risultato IDFT($\mathbf{z}$) viene ancora memorizzato in $\mathbf{z}$).

$m := 1;$
for $t := 1$ **to** s **do begin**
$m := 2m;\ \theta := \dfrac{2\pi}{m};\ c := \cos\theta;\ s := \sin\theta;\ \omega_{re} := 1;\ \omega_{im} := 0;$
for $j := 0$ **to** $\dfrac{m}{2} - 1$ **do begin**
for $h := j$ **to** $n-1$ **step** m **do begin**
$k := h + \dfrac{m}{2};\ \rho := (\omega_{re} + \boldsymbol{i}\omega_{im})z_k;$
$z_k := z_h - \rho;\ z_h := z_h + \rho$
end;
$\sigma := \omega_{re};\ \omega_{re} := \sigma c - \omega_{im}s;\ \omega_{im} := \omega_{im}c + \sigma s$
end
end;

Per il calcolo della DFT si può usare un programma analogo, in cui si sostituisce $\overline{\omega}$ ad ω (cioè $\omega_{re} - \boldsymbol{i}\omega_{im}$ al posto di $\omega_{re} + \boldsymbol{i}\omega_{im}$) e si dividono per n le componenti ottenute. ∎

Ovviamente, nel caso che si debbano calcolare più trasformate dello stesso ordine, conviene calcolare le radici n-esime dell'unità una sola volta.

Si osservi che la struttura ricorsiva della DFT permette di calcolare i coefficienti del polinomio trigonometrico di grado $2n$ utilizzando integralmente i calcoli fatti per il polinomio trigonometrico di grado n.

Esistono altri modi di definire un algoritmo FFT, ad esempio è possibile applicare la trasformazione bit reversal al termine, utilizzando l'*algoritmo di Sande e Tukey*, che sfrutta il fatto che la matrice V è simmetrica (si veda l'esercizio 5.58).

In altre applicazioni è possibile utilizzare metodi che non effettuano trasformazioni bit reversal: ad esempio nel calcolo di un prodotto di convoluzione (si veda l'esercizio 5.60)

$$\mathbf{y} = \mathbf{u} * \mathbf{v}, \quad \text{eseguito con} \quad \mathbf{y} = \mathrm{IDFT}\,(\mathrm{DFT}(\mathbf{u}) \ \cdot\ \mathrm{DFT}(\mathbf{v})),$$

dove il segno $\cdot$ indica il prodotto componente per componente, si possono calcolare le due trasformate interne con l'algoritmo di Sande e Tukey e la trasformata esterna con l'algoritmo di Cooley e Tukey, non effettuando le permutazioni perché il bit reversal è involutorio.

Il costo computazionale di un algoritmo di FFT può essere leggermente ridotto se, anziché iniziare il calcolo con le trasformate di ordine 2, si inizia direttamente con trasformate di ordine 4 (che non richiedono moltiplicazioni in quanto le radici quarte dell'unità sono $1, \boldsymbol{i}, -1, \boldsymbol{i}$), o con le trasformate di ordine 8 (che possono essere calcolate con un ridotto numero di moltiplicazioni in quanto le radici ottave di 1 che non siano già radici quarte hanno parte reale e immaginaria di modulo uguale a $\sqrt{2}/2$).

Con una DFT di ordine n di un vettore complesso è possibile calcolare due DFT di ordine n di vettori reali (si veda l'esercizio 5.52), riducendo alla metà il costo computazionale. Inoltre, con la stessa riduzione del costo, è possibile calcolare la DFT di ordine n di un vettore reale con una DFT di ordine $n/2$ di un vettore complesso (si veda l'esercizio 5.53).

Se n non è potenza di 2 ma è fattorizzabile, è possibile individuare metodi per il calcolo delle DFT e IDFT, che si basano su proprietà analoghe a quelle usate per costruire l'algoritmo 5.64. Si supponga per semplicità che $n = n_1 n_2$, con $n_1 = 3$; procedendo in modo analogo a quanto fatto nella dimostrazione del teorema 5.63, si ha

$$\begin{aligned} y_k &= \sum_{k=0}^{n-1} z_j\, \omega_n^{jk} = \sum_{p=0}^{n_2-1} z_{3p}\, \omega_n^{3kp} + \sum_{p=0}^{n_2-1} z_{3p+1}\, \omega_n^{k(3p+1)} + \sum_{p=0}^{n_2-1} z_{3p+2}\, \omega_n^{k(3p+2)} \\ &= \sum_{p=0}^{n_2-1} z_{3p}\, \omega_{n_2}^{kp} + \omega_n^k \sum_{p=0}^{n_2-1} z_{3p+1}\, \omega_{n_2}^{kp} + \omega_n^{2k} \sum_{p=0}^{n_2-1} z_{3p+2}\, \omega_{n_2}^{kp}. \end{aligned}$$

Quindi la trasformata di ordine $n_1 n_2$ può essere calcolata mediante n_1 trasformate di ordine n_2.

Si possono avere notevoli riduzioni del costo computazionale se n è scomponibile in un elevato numero di fattori primi. Va però osservato che le permutazioni degli indici dei vettori, conseguenti a tali fattorizzazioni, sono, a differenza del bit reversal, assai più difficilmente implementabili. Esistono altri algoritmi, individuati da Winograd per alcuni valori primi di n, che sfruttano le proprietà di struttura della matrice V [25].

Gli algoritmi per il calcolo della trasformata discreta di Fourier, sia quello che implementa direttamente la (74) o la (75), sia l'algoritmo FFT sono numericamente stabili. Vale infatti il seguente teorema per la cui dimostrazione si rimanda a [10].

5.65 Teorema *(di Gentleman e Sande). Sia $n = 2^s$ e sia $\widetilde{\mathbf{z}}$ il vettore effettivamente calcolato operando in un'aritmetica con precisione u. Se il calcolo viene eseguito mediante la (74) risulta*

$$\frac{\|\mathbf{z} - \widetilde{\mathbf{z}}\|_2}{\|\mathbf{z}\|_2} \leq k_1 u (2n)^{3/2},$$

mentre se il calcolo viene eseguito mediante l'algoritmo FFT di Cooley e Tukey risulta

$$\frac{\|\mathbf{z} - \widetilde{\mathbf{z}}\|_2}{\|\mathbf{z}\|_2} \leq k_2 u \sqrt{n} \log_2 n,$$

in cui k_1 e k_2 sono due costanti (se l'aritmetica utilizzata è in base 2 è $k_1 = 1.06$ e $k_2 = 8.5$). ∎

Una tipica applicazione della FFT riguarda il *filtraggio digitale* di un segnale.

5.66 Esempio. Sia

$$s(x) = \frac{\alpha_0}{2} + \sum_{j=1}^{\infty} \big(\alpha_j \cos jx + \beta_j \sin jx\big)$$

una funzione periodica di periodo T che rappresenta un *segnale*. Il j-esimo termine della serie, detto *j-esima componente o armonica* del segnale, ha modulo massimo $\rho_j = \sqrt{\alpha_j^2 + \beta_j^2}$, detto *ampiezza* della j-esima componente. Si considera una funzione $r(x)$ che rappresenta un *rumore*. Nella figura 5.21 si riporta il grafico della funzione definita fra $[0, 2\pi]$ da

$$s(x) = \Big((x - 1.2)\sin 3x + \frac{1}{x+1}\ \sin(1 + x^2)\Big) \sin\frac{x}{2}$$

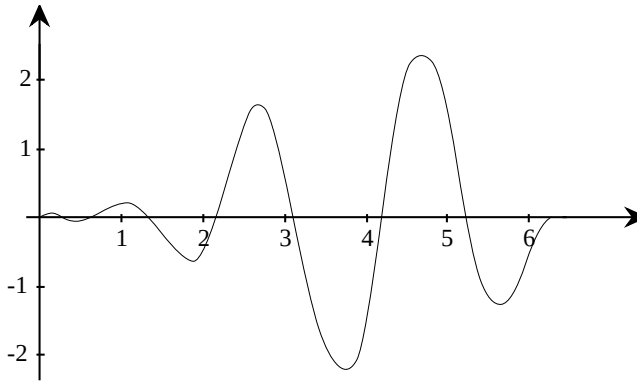


Fig. 5.21 - Segnale $s(x)$.

ed estesa per periodicità con periodo $T = 2\pi$, nella figura 5.22 si riporta il grafico della funzione $t(x) = s(x) + r(x)$, dove la funzione rumore è

$$r(x) = 0.1 \sin 7x + 2 \sin 23x\ \cos 31x\ \sin(1 - 19x).$$

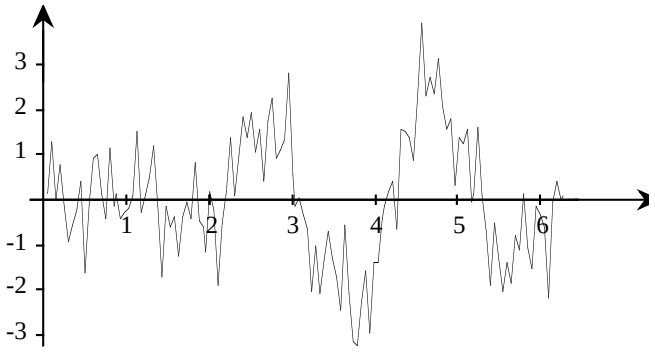


Fig. 5.22 - Segnale + rumore.

A partire dalla funzione $t(x)$, mediante un operazione di filtraggio è possibile recuperare la maggior parte dell'informazione contenuta nella funzione $s(x)$. Infatti, costruendo attraverso la FFT il polinomio trigonometrico $F_m(x)$ della forma (69) della funzione $t(x)$ relativo a 128 punti, ed arrestando la sommatoria ad $m = 8$ (*eliminazione delle alte frequenze*), si ottiene un nuovo polinomio trigonometrico *filtrato* $F_8(x)$, il cui grafico è riportato nella figura 5.23.

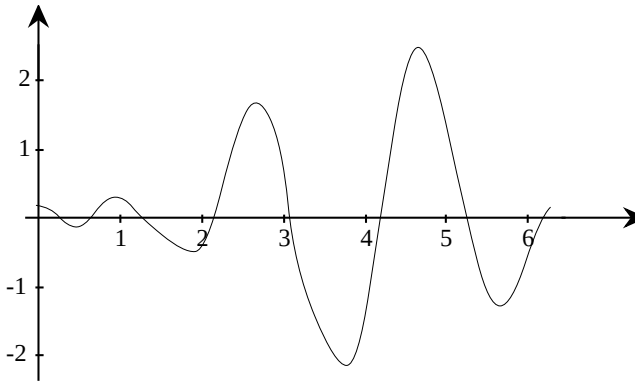


Fig. 5.23 - Polinomio trigonometrico filtrato $F_8(x)$.

In questo caso la forte rassomiglianza delle funzioni $s(x)$ e $F_8(x)$ è dovuta al fatto che le ampiezze massime ρ_j, con $j > 8$, delle componenti di $s(x)$ sono trascurabili, mentre la funzione $r(x)$ ha componenti nulle per $j < 7$, come illustrato nella figura 5.24 (quadratini neri per le ampiezze delle componenti di $s(x)$ e pallini per le ampiezze delle componenti di $r(x)$) ∎.

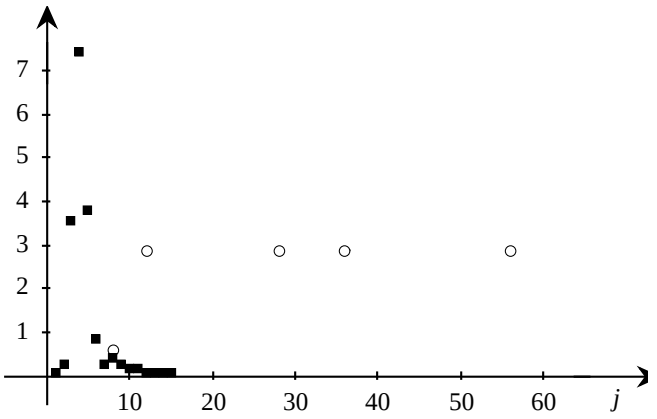


Fig. 5.24 - Ampiezze massime delle componenti di $s(x)$ e di $r(x)$.

14. Funzioni spline

A causa del comportamento oscillante dei polinomi di grado elevato spesso non è possibile utilizzare la tecnica dell'interpolazione per approssimare le funzioni. Polinomi di grado più basso si possono ottenere con le tecniche di approssimazione, ma in tal caso nei nodi i valori del polinomio approssimante non sono uguali a quelli della funzione. Se invece l'uguaglianza dei valori nei nodi è fondamentale, come ad esempio nella grafica, si possono utilizzare funzioni che coincidono a tratti con polinomi di grado basso.

L'intervallo $[a, b]$ viene diviso in n sottointervalli con $n + 1$ nodi x_i, $i = 0, \ldots, n$, tali che

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Una funzione $g(x)$ *polinomiale a tratti* su $[a, b]$ è una funzione che sull'i-esimo sottointervallo $[x_i, x_{i+1}]$ coincide con un polinomio di grado k_i. Di solito $k_i = k$, $i = 0, \ldots, n-1$, cioè i polinomi usati nei diversi sottointervalli hanno sempre lo stesso grado k. La funzione $g(x)$ viene quindi rappresentata per mezzo di una matrice contenente nell'i-esima riga, $i = 0, \ldots, n-1$, i coefficienti

$$a_{i,k}, \quad a_{i,k-1}, \quad \ldots \quad a_{i,0}$$

dell'i-esimo polinomio

$$p_i(y) = a_{i,k}y^k + a_{i,k-1}y^{k-1} + \ldots + a_{i,0},$$

dove $y = x - x_i$, cioè con la variabile traslata rispetto al punto x_i.

I casi più semplici di funzioni polinomiali a tratti sono i seguenti.

a) Polinomiale *lineare a tratti*: nell'i-esimo sottointervallo la funzione $g(x)$ coincide con il polinomio di interpolazione della $f(x)$ sui nodi x_i e x_{i+1}, cioè dall'esempio 5.17

$$p_i(x) = \frac{f(x_{i+1}) - f(x_i)}{h_i}\,(x - x_i) + f(x_i),$$

dove $h_i = x_{i+1} - x_i$, $i = 0, \ldots, n-1$. Per la sua semplicità questo metodo è usato spesso nella pratica. Inoltre se $f''(x)$ è limitata in $[a, b]$, al tendere a zero del massimo degli h_i, la funzione $g(x)$ tende alla funzione $f(x)$. Però questo metodo non è adatto per una buona rappresentazione grafica della funzione: poiché non vi è alcuna condizione sulle derivate dei due polinomi, nei nodi x_i il raccordo fra due diversi polinomi lineari presenta in generale un punto angoloso.

5.67 Esempio. I seguenti dati si riferiscono alla portata d'acqua di un fiume italiano, misurata mensilmente in m^3/sec.

mese	G	F	M	A	M	G
portata	12.51	13.05	11.7	9.26	8.3	6.25

mese	L	A	S	O	N	D
portata	5.34	4.59	5.14	6.36	10.31	13.88

Nella figura 5.25 sono riportati i grafici della polinomiale lineare a tratti (linea più spessa) e il grafico del polinomio di interpolazione di grado 11 su tutti i punti (linea sottile). Dalla figura risulta che l'andamento del polinomio di interpolazione non rende accettabile l'approssimazione. ∎

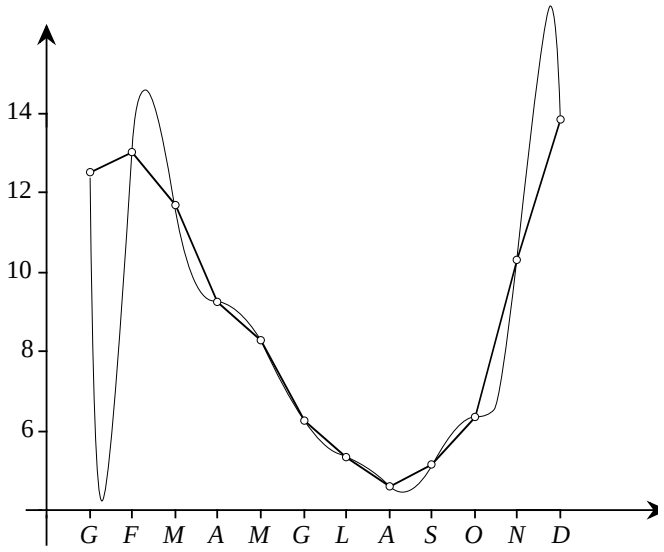


Fig. 5.25 - Grafici del polinomio lineare a tratti e del polinomio di interpolazione.

b) Polinomiale *cubica a tratti di Hermite*: nell'i-esimo sottointervallo la funzione $g(x)$ coincide con il polinomio di Hermite di grado al più 3 che assume nei punti x_i e x_{i+1} gli stessi valori della $f(x)$ e la cui derivata assume nei punti x_i e x_{i+1} gli stessi valori della $f'(x)$, cioè dall'esempio 5.13:

$$p_i(x) = \frac{1}{h_i^3}\Big[2\big(f(x_i) - f(x_{i+1})\big) + h_i\big(f'(x_i) + f'(x_{i+1})\big)\Big]y_i^3$$
$$- \frac{1}{h_i^2}\Big[3\big(f(x_i) - f(x_{i+1})\big) + h_i\big(2f'(x_i) + f'(x_{i+1})\big)\Big]y_i^2 + f'(x_i)\,y_i + f(x_i),$$

dove $y_i = x - x_i$, $h_i = x_{i+1} - x_i$, $i = 0, \ldots, n-1$. Queste polinomiali forniscono una migliore approssimazione della funzione e quindi consentono una migliore rappresentazione grafica, perché non danno luogo a punti angolosi nei nodi, ma non sono utilizzabili nel caso, frequente nella pratica, in cui i valori di $f'(x_i)$ non sono noti. Inoltre nei punti di raccordo i polinomi hanno la stessa pendenza, ma non la stessa concavità, per cui nei nodi si può presentare un andamento distorto.

c) Altre polinomiali a tratti possono essere costruite interpolando la $f(x)$ su più di due punti consecutivi, ad esempio considerando le cubiche che interpolano su quattro nodi. È anche possibile sfruttare altre condizioni che caso per caso possono essere fornite dal problema.

Fra le funzioni polinomiali a tratti quelle più usate nella pratica, anche perché consentono di ottenere ottimi risultati dal punto di vista grafico, sono le polinomiali cubiche ottenute senza utilizzare i valori, in generale non disponibili, delle derivate, e imponendo invece condizioni di continuità delle derivate prima e seconda.

5.68 Definizione. Siano $x_0, \ldots x_n$, $n+1$ punti distinti di $[a, b]$ tali che

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Una funzione reale $s(x) \in C^2[a, b]$ viene chiamata *spline cubica* per l'approssimazione della $f(x)$ se

a) in ogni sottointervallo $[x_i, x_{i+1}]$, $i = 0, \ldots, n-1$, $s(x)$ coincide con un polinomio di grado al piú 3;

b) $s(x_i) = f(x_i)$, per $i = 0, \ldots, n$. ∎

Indicando con $s_i(x)$, $i = 0, \ldots, n-1$, il polinomio che coincide con la $s(x)$ nel sottointervallo $[x_i, x_{i+1}]$, dalla definizione precedente si ottengono le $4n-2$ condizioni

a) $s_i(x_i) = f(x_i), \quad s_i(x_{i+1}) = f(x_{i+1}), \quad i = 0, \ldots, n-1,$

b) $s'_{i-1}(x_i) = s'_i(x_i), \quad i = 1, \ldots, n-1,$

c) $s''_{i-1}(x_i) = s''_i(x_i), \quad i = 1, \ldots, n-1.$

Poiché i coefficienti dei polinomi $s_i(x)$ sono $4n$, occorre imporre due condizioni aggiuntive, che vengono scelte in modo da fornire una buona approssimazione. Vari sono i criteri che possono essere seguiti per individuare queste due condizioni: per esempio

d′) $s''_0(x_0) = s''_{n-1}(x_n) = 0,$

oppure, se sono noti i valori di $f'(a)$ e $f'(b)$,

d″) $s'_0(x_0) = f'(a), \quad s'_{n-1}(x_n) = f'(b),$

oppure, se la funzione $f(x)$ è periodica di periodo $b-a$, cioè $f(a) = f(b)$,
d‴) $\quad s_0'(x_0) = s_{n-1}'(x_n), \quad s_0''(x_0) = s_{n-1}''(x_n)$.
Una spline cubica viene detta *spline naturale* se verifica le condizioni d′), *spline completa* se verifica le condizioni d″) e *spline periodica* se verifica le condizioni d‴).

Per determinare i coefficienti dei polinomi $s_i(x)$ si potrebbero sfruttare direttamente le condizioni a) - c) e le condizioni aggiuntive scelte, risolvendo un sistema lineare di $4n$ equazioni in $4n$ incognite. È possibile però ridurre il numero delle equazioni necessarie considerando come incognite le quantità, dette *momenti*,

$$\mu_i = s_i''(x_i), \quad i = 0, \ldots, n-1,$$
$$\mu_n = s_{n-1}''(x_n).$$

Infatti, poiché $s_i(x)$ per $x \in [x_i, x_{i+1}]$ è un polinomio di grado al più 3, $s_i''(x)$ è di grado al più 1 e può essere così rappresentata con la formula di Lagrange

$$s_i''(x) = \mu_{i+1} \frac{x - x_i}{h_i} - \mu_i \frac{x - x_{i+1}}{h_i}, \tag{85}$$

dove $h_i = x_{i+1} - x_i$. Integrando due volte si ottiene per $i = 0, \ldots, n-1$,

$$\begin{aligned} s_i'(x) &= \mu_{i+1} \frac{(x-x_i)^2}{2h_i} - \mu_i \frac{(x-x_{i+1})^2}{2h_i} + \alpha_i, \\ s_i(x) &= \mu_{i+1} \frac{(x-x_i)^3}{6h_i} - \mu_i \frac{(x-x_{i+1})^3}{6h_i} + \alpha_i(x-x_i) + \beta_i, \end{aligned} \tag{86}$$

e le costanti α_i e β_i vengono determinate imponendo le condizioni a)

$$\begin{cases} \mu_i \dfrac{h_i^2}{6} + \beta_i = f(x_i) \\ \mu_{i+1} \dfrac{h_i^2}{6} + \alpha_i h_i + \beta_i = f(x_{i+1}), \end{cases}$$

da cui

$$\begin{cases} \beta_i = f(x_i) - \mu_i \dfrac{h_i^2}{6} \\ \alpha_i = \dfrac{f(x_{i+1}) - f(x_i)}{h_i} - \dfrac{h_i}{6}(\mu_{i+1} - \mu_i). \end{cases}$$

Restano quindi da calcolare i μ_i, $i = 0, \ldots, n$. Dalle (86), imponendo le condizioni b) e sostituendo α_{i-1} e α_i, si ottengono le $n-1$ relazioni

$$\begin{aligned} &\frac{h_{i-1}}{6}\mu_{i-1} + \frac{h_{i-1}+h_i}{3}\mu_i + \frac{h_i}{6}\mu_{i+1} \\ &\qquad = \frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{f(x_i) - f(x_{i-1})}{h_{i-1}}, \quad i = 1, \ldots, n-1. \end{aligned} \tag{87}$$

Con le notazioni delle differenze divise (paragrafo 5), si ha

$$\frac{f(x_{i+1})-f(x_i)}{h_i} - \frac{f(x_i)-f(x_{i-1})}{h_{i-1}} = f[x_i,x_{i+1}]-f[x_{i-1},x_i]$$
$$= f[x_{i-1},x_i,x_{i+1}](h_{i-1}+h_i)$$

e quindi la (87) può essere rappresentata come

$$\frac{h_{i-1}}{h_{i-1}+h_i}\,\mu_{i-1}+2\mu_i+\frac{h_i}{h_{i-1}+h_i}\,\mu_{i+1}=6f[x_{i-1},x_i,x_{i+1}],$$

da cui si ottiene

$$\gamma_i\mu_{i-1}+2\mu_i+\delta_i\mu_{i+1}=6f[x_{i-1},x_i,x_{i+1}],\quad i=1,\dots,n-1, \tag{88}$$

dove $\gamma_i=\dfrac{h_{i-1}}{h_{i-1}+h_i}>0,\quad \delta_i=\dfrac{h_i}{h_{i-1}+h_i}>0,\quad \gamma_i+\delta_i=1.$

Altre due relazioni si ottengono tramite le condizioni aggiuntive. Per la spline naturale, dalle d′) si ha

$$\mu_0=0,\qquad \mu_n=0. \tag{89}$$

Per la spline completa, dalle d″) si ha:

$$s_0'(x_0)=-\mu_0\,\frac{h_0}{3}-\mu_1\,\frac{h_0}{6}+\frac{f(x_1)-f(x_0)}{h_0}=f'(x_0),$$
$$s_{n-1}'(x_n)=\mu_{n-1}\,\frac{h_{n-1}}{6}+\mu_n\,\frac{h_{n-1}}{3}+\frac{f(x_n)-f(x_{n-1})}{h_{n-1}}=f'(x_n),$$

da cui

$$\begin{cases}\dfrac{h_0}{3}\,\mu_0+\dfrac{h_0}{6}\,\mu_1=\dfrac{f(x_1)-f(x_0)}{h_0}-f'(x_0)\\[2ex] \dfrac{h_{n-1}}{6}\,\mu_{n-1}+\dfrac{h_{n-1}}{3}\,\mu_n=f'(x_n)-\dfrac{f(x_n)-f(x_{n-1})}{h_{n-1}}\,;\end{cases} \tag{90}$$

con le notazioni delle differenze divise, tenendo conto che per la (28) è

$$f[x_0,x_0]=f'(x_0),\quad f[x_n,x_n]=f'(x_n),$$

si ha

$$2\mu_0+\mu_1=6f[x_0,x_0,x_1],\qquad \mu_{n-1}+2\mu_n=6f[x_{n-1},x_n,x_n]. \tag{91}$$

Per la spline periodica, dalle d‴) si ha:

$$\begin{cases} \dfrac{h_{n-1}+h_0}{3}\,\mu_0 + \dfrac{h_0}{6}\,\mu_1 + \dfrac{h_{n-1}}{6}\,\mu_{n-1} \\ \qquad\qquad = \dfrac{f(x_1)-f(x_0)}{h_0} - \dfrac{f(x_n)-f(x_{n-1})}{h_{n-1}}\,, \\ \mu_0 = \mu_n, \end{cases}$$

tenendo conto che in questo caso è $f(x_n) = f(x_0)$, con l'introduzione del punto x_{n+1} tale che $x_{n+1} - x_n = x_1 - x_0$ e $f(x_{n+1}) = f(x_1)$, si ha

$$\gamma_n\mu_{n-1} + 2\mu_n + \delta_n\mu_1 = 6f[x_{n-1}, x_n, x_{n+1}], \tag{92}$$

mentre la prima delle (88) può essere scritta

$$2\mu_1 + \delta_1\mu_2 + \gamma_1\mu_n = 6f[x_0, x_1, x_2].$$

In ogni caso i $\mu_0, \ldots, \mu_n$ sono soluzione di un sistema lineare, ottenuto associando alle (88) le (89) o le (91) o le (92), a seconda che debbano essere verificate le condizioni d′), o d″), o d‴).

5.69 Teorema. *Siano $x_0, \ldots x_n$, $n+1$ punti distinti di $[a, b]$ tali che*

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Allora esiste ed è unica la spline cubica che approssima la $f(x)$ e verifica una delle condizioni d′), o d″), o d‴).

Dim. Nel primo caso, tenendo conto che $\mu_0 = \mu_n = 0$, il sistema ottenuto dalle (88) è

$$\begin{bmatrix} 2 & \delta_1 & & & \\ \gamma_2 & 2 & \delta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-2} & 2 & \delta_{n-2} \\ & & & \gamma_{n-1} & 2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{n-2} \\ \mu_{n-1} \end{bmatrix} = 6 \begin{bmatrix} f[x_0, x_1, x_2] \\ f[x_1, x_2, x_3] \\ \vdots \\ f[x_{n-3}, x_{n-2}, x_{n-1}] \\ f[x_{n-2}, x_{n-1}, x_n] \end{bmatrix}.$$

Nel secondo caso vengono aggiunte una prima e un'ultima equazione al sistema che diventa

$$\begin{bmatrix} 2 & 1 & & & \\ \gamma_1 & 2 & \delta_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-1} & 2 & \delta_{n-1} \\ & & & 1 & 2 \end{bmatrix} \begin{bmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_{n-1} \\ \mu_n \end{bmatrix} = 6 \begin{bmatrix} f[x_0, x_0, x_1] \\ f[x_0, x_1, x_2] \\ \vdots \\ f[x_{n-2}, x_{n-1}, x_n] \\ f[x_{n-1}, x_n, x_n] \end{bmatrix}.$$

Nel terzo caso si ha

$$\begin{bmatrix} 2 & \delta_1 & & & \gamma_1 \\ \gamma_2 & 2 & \delta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-1} & 2 & \delta_{n-1} \\ \delta_n & & & \gamma_n & 2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{n-1} \\ \mu_n \end{bmatrix} = 6 \begin{bmatrix} f[x_0, x_1, x_2] \\ f[x_1, x_2, x_3] \\ \vdots \\ f[x_{n-2}, x_{n-1}, x_n] \\ f[x_{n-1}, x_n, x_{n+1}] \end{bmatrix}.$$

Poiché le matrici di questi sistemi hanno predominanza diagonale in senso stretto e quindi sono non singolari, i sistemi hanno una e una sola soluzione [2], che può essere calcolata con il metodo di Gauss senza scambi di righe. Nei primi due casi la matrice è tridiagonale, e quindi il metodo di Gauss ha un basso costo computazionale, dell'ordine di n. Anche nel terzo caso è possibile ricondurre il problema alla risoluzione di opportuni sistemi lineari con matrice tridiagonale, ad esempio usando la formula di Woodbury. ∎

5.70 Esempio. Nella figura 5.26 è riportato il grafico della spline cubica naturale che approssima la funzione dell'esempio 5.67. Dal confronto con la figura 5.25 risulta che l'approssimazione ottenuta con le spline è migliore di quelle ottenute con la polinomiale lineare a tratti e con il polinomio di interpolazione. ∎

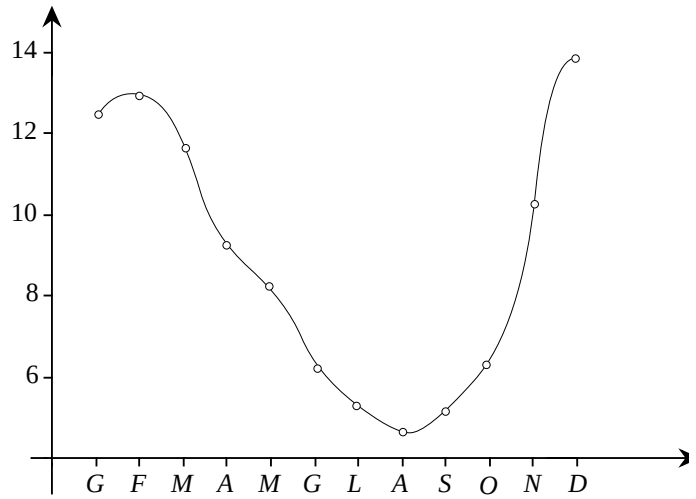


Fig. 5.26 - Grafico della spline cubica.

Se i punti x_i sono equidistanti, cioè $h_i = h$, per $i = 0, \ldots, n-1$, la matrice del sistema risulta molto semplice. La (88) infatti si può scrivere,

tenendo conto della (24)

$$\mu_{i-1} + 4\mu_i + \mu_{i+1} = \frac{6}{h^2}\ \Delta^2 f(x_{i-1}),$$

e quindi per le spline naturali il sistema diventa

$$\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{n-2} \\ \mu_{n-1} \end{bmatrix} = \frac{6}{h^2} \begin{bmatrix} \Delta^2 f(x_0) \\ \Delta^2 f(x_1) \\ \vdots \\ \Delta^2 f(x_{n-3}) \\ \Delta^2 f(x_{n-2}) \end{bmatrix}.$$

Le spline cubiche sono molto usate nella grafica perché fra le funzioni con derivata seconda continua che interpolano la funzione $f(x)$ nei nodi x_i, $i = 0, \dots, n$, sono quelle che hanno minima curvatura, cioè che oscillano meno, come risulta dal seguente teorema.

5.71 Teorema. *Fra tutte le funzioni $g(x) \in C^2[a,b]$, tali che $g(x_i) = f(x_i)$, $i = 0, \dots, n$, la spline cubica naturale $s(x)$ è quella che minimizza l'integrale*

$$\int_a^b [g''(x)]^2\ dx. \tag{93}$$

Dim. Si ha

$$\begin{aligned} 0 &\le \int_a^b [g''(x) - s''(x)]^2\ dx \\ &= \int_a^b [g''(x)]^2\ dx - 2\int_a^b [g''(x) - s''(x)]s''(x)\ dx - \int_a^b [s''(x)]^2\ dx. \end{aligned} \tag{94}$$

Per ogni sottointervallo $[x_i, x_{i+1}]$ si ha, integrando due volte per parti,

$$\begin{aligned} \int_{x_i}^{x_{i+1}} [g''(x) - s''(x)]s''(x)\ dx &= \Big[\big[g'(x) - s'(x)\big]s''(x)\Big]_{x_i}^{x_{i+1}} \\ &\quad - \Big[\big[g(x) - s(x)\big]s^{(3)}(x)\Big]_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} \big[g(x) - s(x)\big]s^{(4)}(x)\ dx. \end{aligned}$$

Poiché $s(x)$ sull'intervallo $[x_i, x_{i+1}]$ coincide con un polinomio di grado al più 3, è $s^{(4)}(x) = 0$. Inoltre $s(x_i) = g(x_i)$, $s(x_{i+1}) = g(x_{i+1})$, per cui

$$\begin{aligned} \int_a^b [g''(x) - s''(x)]s''(x)\ dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} [g''(x) - s''(x)]s''(x)\ dx \\ &= \sum_{i=0}^{n-1} \Big[\big[g'(x) - s'(x)\big]s''(x)\Big]_{x_i}^{x_{i+1}} = \Big[\big[g'(x) - s'(x)\big]s''(x)\Big]_a^b \end{aligned}$$

e tale espressione è nulla in quanto $s''(a) = s''(b) = 0$. Dalla (94) segue che

$$\int_a^b [s''(x)]^2 \, dx \leq \int_a^b [g''(x)]^2 dx$$

per ogni funzione $g(x)$ a derivata seconda continua tale che $g(x_i) = f(x_i)$.∎

La $g''(x)$ è legata alla curvatura della funzione $g(x)$ nel punto x, definita come il reciproco del raggio del cerchio osculatore in x, e data dall'espressione

$$c(x) = |g''(x)|(1 + [g'(x)]^2)^{-3/2};$$

l'integrale (93) può allora essere assunto come una misura della *curvatura globale* della funzione $g(x)$, se $|g'(x)|$ è piccolo rispetto ad 1. Dal teorema 5.71 risulta quindi che la spline cubica naturale è quella che minimizza la curvatura globale. Dal teorema 5.71 segue anche che se $f(x) \in C^2[a, b]$, allora

$$\int_a^b [s''(x)]^2 \, dx \leq \int_a^b [f''(x)]^2 dx. \tag{95}$$

A differenza di quanto accade per i polinomi che interpolano la $f(x)$ su tutto l'intervallo $[a, b]$, le funzioni spline convergono alla $f(x)$ quando si infittiscono i nodi x_i , $i = 0, \ldots, n$. Valgono infatti i seguenti teoremi.

5.72 Teorema. *Sia $f(x) \in C^4[a, b]$, con*

$$M_4 = \max_{x \in [a,b]} |f^{(4)}(x)|,$$

e sia

$$H = \max_{i=0,\ldots,n-1} h_i.$$

Allora per i momenti della spline cubica $s(x)$ completa vale la relazione

$$\max_{i=0,\ldots,n} |\mu_i - f''(x_i)| \leq \frac{3}{4} \, M_4 H^2. \tag{96}$$

Dim. Dalla formula di Taylor si ha

$$\begin{aligned}
\frac{f(x_{i+1}) - f(x_i)}{h_i} &= f'(x_i) + \frac{h_i}{2} \, f''(x_i) + \frac{h_i^2}{3!} \, f''''(x_i) + \frac{h_i^3}{4!} \, f^{(4)}(\xi_{i,1}), \\
\frac{f(x_i) - f(x_{i-1})}{h_{i-1}} &= f'(x_i) - \frac{h_{i-1}}{2} \, f''(x_i) + \frac{h_{i-1}^2}{3!} \, f'''(x_i) \\
&\quad - \frac{h_{i-1}^3}{4!} \, f^{(4)}(\xi_{i,2}), \\
f''(x_{i-1}) &= f''(x_i) - h_{i-1} f'''(x_i) + \frac{h_{i-1}^2}{2} \, f^{(4)}(\xi_{i,3}), \\
f''(x_{i+1}) &= f''(x_i) + h_i f'''(x_i) + \frac{h_i^2}{2} \, f^{(4)}(\xi_{i,4}),
\end{aligned}$$

in cui $\xi_{i,j} \in [a,b]$, per $j = 1, \ldots, 4$. Posto per $i = 1, \ldots, n-1$

$$g_i = \frac{h_{i-1}}{6}\,[\mu_{i-1} - f''(x_{i-1})] + \frac{h_{i-1}+h_i}{3}\,[\mu_i - f''(x_i)] + \frac{h_i}{6}\,[\mu_{i+1} - f''(x_{i+1})],$$

sostituendo le relazioni precedenti nella (87) si ha

$$g_i = \frac{h_i^3}{12}\,\left[\frac{1}{2}\,f^{(4)}(\xi_{i,1}) - f^{(4)}(\xi_{i,4})\right] + \frac{h_{i-1}^3}{12}\,\left[\frac{1}{2}\,f^{(4)}(\xi_{i,2}) - f^{(4)}(\xi_{i,3})\right].$$

In modo analogo, posto

$$g_0 = \frac{h_0}{3}\,[\mu_0 - f''(x_0)] + \frac{h_0}{6}\,[\mu_1 - f''(x_1)],$$
$$g_n = \frac{h_{n-1}}{6}\,[\mu_{n-1} - f''(x_{n-1})] + \frac{h_{n-1}}{3}\,[\mu_n - f''(x_n)],$$

dalle (90) si ottiene

$$g_0 = \frac{h_0^3}{12}\,\left[\frac{1}{2}\,f^{(4)}(\xi_{0,1}) - f^{(4)}(\xi_{0,4})\right],$$
$$g_n = \frac{h_{n-1}^3}{12}\,\left[\frac{1}{2}\,f^{(4)}(\xi_{n,2}) - f^{(4)}(\xi_{n,3})\right].$$

Passando ai moduli si ha

$$\begin{aligned} &|g_0| \le \frac{M_4}{8}\,h_0^3, \\ &|g_i| \le \frac{M_4}{8}\,(h_{i-1}^3 + h_i^3), \text{ per } i = 1, \ldots, n-1, \\ &|g_n| \le \frac{M_4}{8}\,h_{n-1}^3. \end{aligned} \tag{97}$$

Sia ora k l'indice per cui

$$|\mu_k - f''(x_k)| = \max_{i=0,\ldots,n} |\mu_i - f''(x_i)|. \tag{98}$$

Se $1 \le k \le n-1$, si ha

$$\begin{aligned} g_k = {} & \frac{h_{k-1}}{6}\,\left[\mu_{k-1} - f''(x_{k-1})\right] + \frac{h_{k-1}+h_k}{3}\,\left[\mu_k - f''(x_k)\right] \\ & + \frac{h_k}{6}\,\left[\mu_{k+1} - f''(x_{k+1})\right] = \frac{h_{k-1}+h_k}{6}\,\left[\mu_k - f''(x_k)\right] \\ & + \frac{h_{k-1}}{6}\,\left[\mu_k - f''(x_k) + \mu_{k-1} - f''(x_{k-1})\right] \\ & + \frac{h_k}{6}\,\left[\mu_k - f''(x_k) + \mu_{k+1} - f''(x_{k+1})\right]. \end{aligned}$$

Per la (98) l'espressione $\mu_k - f''(x_k) + \mu_i - f''(x_i)$, per $i = k-1$ e $k+1$, ha lo stesso segno di $\mu_k - f''(x_k)$, per cui

$$|g_k| \geq \frac{h_{k-1} + h_k}{6} \left|\mu_k - f''(x_k)\right|$$

e quindi

$$|\mu_k - f''(x_k)| \leq \frac{6\,|g_k|}{h_{k-1} + h_k}\,. \tag{99}$$

Per la (97) è allora

$$|\mu_k - f''(x_k)| \leq \frac{3}{4}\, M_4\, \frac{h_{k-1}^3 + h_k^3}{h_{k-1} + h_k} = \frac{3}{4}\, M_4\, (h_{k-1}^2 - h_{k-1}h_k + h_k^2) \leq \frac{3}{4}\, M_4 H^2.$$

Se invece $k = 0$ oppure $k = n$, al posto della (99) si ottiene

$$|\mu_0 - f''(x_0)| \leq \frac{6|g_0|}{h_0}\,,$$
$$|\mu_n - f''(x_n)| \leq \frac{6|g_n|}{h_{n-1}}\,,$$

e in entrambi i casi segue la tesi. ∎

5.73 Teorema. *Nelle ipotesi del teorema 5.72, indicato con*

$$h = \min_{i=0,\ldots,n} h_i,$$

per la spline completa valgono le limitazioni

$$|f'''(x) - s_i'''(x)| \leq 2M_4\, \frac{H^2}{h}\,, \quad \textit{per} \;\; x \in [x_i, x_{i+1}], \quad i = 0, \ldots, n-1,$$
$$|f''(x) - s''(x)| \leq \frac{7}{4}\, M_4\, \frac{H^3}{h}\,,$$
$$|f'(x) - s'(x)| \leq \frac{7}{4}\, M_4\, \frac{H^4}{h}\,,$$
$$|f(x) - s(x)| \leq \frac{7}{8}\, M_4\, \frac{H^5}{h}\,, \quad \textit{per} \;\; x \in [a, b].$$

Dim. Dalla (85) si ha che per $x \in [x_i, x_{i+1}]$ è

$$s_i'''(x) = \frac{\mu_{i+1} - \mu_i}{h_i}\,,$$

da cui

$$s_i'''(x) - f'''(x) = \frac{\mu_{i+1} - \mu_i}{h_i} - f'''(x)$$
$$= \frac{[\mu_{i+1} - f''(x_{i+1})] - [\mu_i - f''(x_i)]}{h_i}$$
$$+ \frac{[f''(x_{i+1}) - f''(x)] - [f''(x_i) - f''(x)]}{h_i} - f'''(x).$$

Per la formula di Taylor e per la (96) è

$$\left|s_i'''(x) - f'''(x)\right| \leq \frac{3}{2h_i}\, M_4 H^2 + \frac{1}{2h_i}\, \left|(x_{i+1}-x)^2 f^{(4)}(\xi_1) - (x_i - x)^2 f^{(4)}(\xi_2)\right|,$$

con $\xi_1, \xi_2 \in [x_i, x_{i+1}]$, da cui, poiché

$$(x_{i+1} - x)^2 + (x_i - x)^2 \leq (x_{i+1} - x_i)^2 = h_i^2,$$

segue che per $x \in [x_i, x_{i+1}]$ vale

$$|s_i'''(x) - f'''(x)| \leq \frac{3}{2h_i}\, M_4 H^2 + \frac{1}{2}\, h_i M_4 \leq 2M_4\, \frac{H^2}{h_i}\,. \tag{100}$$

Per la seconda disuguaglianza, se x coincide con uno dei nodi la maggiorazione discende subito dalla (96); se x non coincide con uno dei nodi, si consideri un indice i tale che

$$|x - x_i| \leq \frac{H}{2}\,, \tag{101}$$

e tale che non vi siano nodi nell'intervallo di estremi x e x_i. Allora

$$\int_{x_i}^{x} [s'''(t) - f'''(t)]\, dt = [s''(x) - f''(x)] - [s''(x_i) - f''(x_i)],$$

da cui

$$s''(x) - f''(x) = s''(x_i) - f''(x_i) + \int_{x_i}^{x} [s'''(t) - f'''(t)]\, dt.$$

Tenendo conto della (101), per le (96) e (100) risulta

$$|s''(x) - f''(x)| \leq \frac{3}{4}\, M_4 H^2 + 2M_4\, \frac{H^2}{h}\, \frac{H}{2} \leq \frac{7}{4}\, M_4\, \frac{H^3}{h}\,.$$

Per ricavare la terza disuguaglianza, poiché per $i = 0, \ldots, n$, è

$$f(x_i) = s(x_i),$$

per il teorema di Rolle in ogni intervallo $[x_i, x_{i+1}]$, $i = 0, \ldots, n-1$, esiste un punto ξ_i, tale che

$$f'(\xi_i) = s'(\xi_i). \tag{102}$$

Quindi per ogni $x \in [a, b]$, esiste uno ξ_i, con

$$|\xi_i - x| \leq H,$$

per cui vale la (102), e quindi

$$\int_{\xi_i}^{x} [s''(t) - f''(t)]\, dt = s'(x) - f'(x).$$

Passando ai moduli si ha

$$|s'(x) - f'(x)| \leq \frac{7}{4}\, M_4\, \frac{H^3}{h}\, H = \frac{7}{4}\, M_4\, \frac{H^4}{h}\,.$$

In modo analogo si ricava la quarta disuguaglianza, tenendo conto che per ogni $x \in [a, b]$ esiste un indice i per cui vale la (101). ∎

Dal teorema 5.73 segue che per una funzione $f(x)$ derivabile con continuità fino al quarto ordine, se si infittiscono i nodi in modo regolare, cioè in modo che il rapporto H/h sia sempre limitato, allora si ha convergenza della spline e delle sue derivate fino al terzo ordine rispettivamente alla $f(x)$ e alle sue derivate. In particolare se i nodi rimangono equidistanti, allora $H/h = 1$ e la convergenza è molto rapida, perché

$$|f(x) - s(x)| \leq M_4 H^4.$$

Dalla (95) segue poi che se i nodi vengono infittiti aggiungendo altri nodi alla precedente suddivisione dell'intervallo, la successione degli integrali

$$\int_a^b [s''(x)]^2 dx$$

risulta non decrescente (si veda l'esercizio 5.79) e convergente a

$$\int_a^b [f''(x)]^2 dx.$$

Le spline cubiche studiate in questo paragrafo sono quelle più usate. Non esistono comunque difficoltà a una generalizzazione, considerando spline di ordine dispari maggiore di 3 (si veda l'esercizio 5.82).

Esercizi proposti

5.1 Siano $x_0, x_1, \ldots, x_n$, $n+1$ numeri e sia V la matrice di Vandermonde i cui elementi sono

$$v_{i,j} = x_{i-1}^{j-1}, \quad i,\ j = 1, \ldots, n+1.$$

a) Si dimostri che

$$\det V = \prod_{\substack{i,j=0,n \\ j>i}} (x_j - x_i),$$

e quindi $\det V \neq 0$ se e solo se i numeri x_i sono a due a due distinti;

b) si dica quanto vale $\det V$ nel caso che i punti x_i siano equidistanti.

(Traccia: a) si dimostri per induzione che, detto d_n il determinante della matrice V di ordine $n+1$, vale la relazione

$$d_n = (x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}) d_{n-1}, \quad d_0 = 1.$$

Per questo, conviene sottrarre dalla j-esima colonna di V la $(j-1)$-esima moltiplicata per x_n, per $j = 2, \ldots, n+1$; b) se $x_i = x_0 + ih$, posto $m = n(n+1)/2$, è

$$\det V = h^m \prod_{\substack{i,j=0,n \\ j>i}} (j - i) = h^m 1!\, 2! \ldots n!.\ )$$

5.2 Si scrivano i polinomi di interpolazione parabolica (di grado al più 2, sui nodi x_0, x_1 e x_2) e cubica (di grado al più 3, sui nodi x_0, x_1, x_2 e x_3), i relativi resti e le maggiorazioni dei resti, in modo analogo a quanto fatto negli esempi 5.4 e 5.6 per il caso lineare.

5.3 Si determinino i punti x_0, $x_1 \in [-1, 1]$ tali che

$$\max_{x \in [-1,1]} |\pi_1(x)|$$

sia più piccolo possibile. Questa scelta di x_0 e x_1 minimizza sull'intervallo $[-1, 1]$ il resto dell'interpolazione lineare di una funzione $f(x)$ la cui derivata seconda vari poco nell'intervallo.

(Risposta: $x_1 = -x_0 = \dfrac{1}{\sqrt{2}}$.)

5.4 Sia $p(x) = \displaystyle\sum_{k=0}^{n} a_k x^k$ un polinomio di grado n e siano

a) p_{n-1} il polinomio di interpolazione di $p(x)$ sui nodi x_i, $i = 0, \ldots, n-1$;

b) p_{n-2} il polinomio di interpolazione di $p(x)$ sui nodi x_i, $i = 0, \ldots, n-2$.

Si scrivano i due resti.

(Risposta: a) $r(x) = a_n \pi_{n-1}(x)$; b) $r(x) = \pi_{n-2}(x)\Big[a_n\Big(x + \sum_{i=0}^{n-2} x_i\Big) + a_{n-1}\Big]$.)

5.5 Della funzione $f(x)$ sono noti i valori

x	1	2	3	4
$f(x)$	1	-1	1	-1

Per approssimare $f(2.5)$ si utilizzino

a) il polinomio di interpolazione su 2 e 3,

b) il polinomio di interpolazione su 1, 2 e 3,

c) il polinomio di interpolazione su 2, 3 e 4,

d) il polinomio di interpolazione su 1, 2, 3 e 4.

Se $f(x) = \cos \pi(x-1)$, quale dei valori ottenuti è il migliore?

(Traccia: a) 0, b) $-\dfrac{1}{2}$, c) $\dfrac{1}{2}$, d) 0, i polinomi che danno il valore migliore sono il primo e il quarto.)

5.6 Sia $f(x) = \sin \dfrac{\pi x}{2}$ e sia $p_3(x)$ il polinomio di interpolazione di $f(x)$ sui nodi $x_i = i$, $i = 0, \ldots, 3$. Si verifichi che $p_3(4) = f(4)$, per cui il resto $r(x) = f(x) - p_3(x)$ può essere scritto nella forma

$$r(x) = \left(\frac{\pi}{2}\right)^5 \pi_4(x)\, \frac{\cos \dfrac{\pi}{2}\,\xi}{5!}\,, \quad \xi \in (0,4).$$

5.7 Si supponga che la funzione $f(x)$, definita in $[x_0, x_2]$, con $x_i = x_0 + ih$, $i = 0, 1, 2$, $h > 0$, abbia il minimo $f_{\min}$ in un punto di (x_0, x_2). Allora $f_{\min}$ può essere approssimato con il minimo $p_{\min}$ assunto in $x_{\min}$ dal polinomio di grado al più 2 di interpolazione di $f(x)$ sui nodi x_0, x_1 e x_2. Si calcolino $x_{\min}$ e $p_{\min}$.

(Risposta: posto $\alpha = f(x_2) - f(x_0)$, $\quad \beta = f(x_2) - 2f(x_1) + f(x_0)$, è

$$x_{\min} = x_1 - \frac{h\alpha}{2\beta}\,, \quad p_{\min} = f(x_1) - \frac{\alpha^2}{8\beta}\,.\ )$$

5.8 Sia $f(x) = \dfrac{\sin x}{x}$. Per approssimare un valore di $f(x)$, dove $x \in [0.1, 0.4]$, si hanno a disposizione le seguenti tabelle

x	$\sin x$
0.1	0.09983342
0.2	0.1986693
0.3	0.2955202
0.4	0.3894183

x	$f(x)$
0.1	0.9983342
0.2	0.9933467
0.3	0.9850674
0.4	0.9735459

Si approssimi $f(0.25)$ con i seguenti procedimenti:

a) interpolando direttamente dalla seconda tabella;

b) approssimando $\sin 0.25$ con interpolazione nella prima tabella, e poi dividendo per 0.25.

Si dica quale dei due procedimenti è più conveniente dal punto di vista dell'errore analitico.

(Traccia: indicato con p_3 il valore ottenuto in a), con q_3 l'approssimazione di $\sin 0.25$, gli errori analitici sono

$$\epsilon_a = \frac{p_3 - \dfrac{\sin 0.25}{0.25}}{\dfrac{\sin 0.25}{0.25}} = -\frac{0.25\ \tau_3(1.5)}{4!\ 10^4\ \sin 0.25}\ \frac{d^4}{dx^4}\ \frac{\sin x}{x}\Big|_{x=\xi_1},$$

$$\epsilon_b = \frac{\dfrac{q_3}{0.25} - \dfrac{\sin 0.25}{0.25}}{\dfrac{\sin 0.25}{0.25}} = -\frac{\tau_3(1.5)}{4!\ 10^4\ \sin 0.25}\ \sin \xi_2,$$

dove ξ_1 e $\xi_2 \in (0.1, 0.4)$. Si verifichi che per $x \in [0.1, 0.4]$ la $\dfrac{d^4}{dx^4}\ \dfrac{\sin x}{x}$ è decrescente e che

$$\max_{x\in[0.1,0.4]} \Big|\frac{d^4}{dx^4}\ \frac{\sin x}{x}\Big| = \Big|\frac{d^4}{dx^4}\ \frac{\sin x}{x}\Big|_{x=0.1} < 0.2.$$

Poiché $\tau_3(1.5) = 0.5625$, risulta

$$|\epsilon_a| < 0.474\ 10^{-6}, \quad |\epsilon_b| < 0.369\ 10^{-5}.\ )$$

5.9 Si verifichi che per $x \neq x_i$, $i = 0, \ldots, n$, è

a) $$\sum_{j=0}^{n} x_j^k \, L_j(x) = x^k, \quad \text{per} \quad k = 0, \ldots, n;$$

b) $$\pi_n(x) \sum_{j=0}^{n} \frac{1}{(x - x_j)\pi_n'(x_j)} = 1;$$

c) $$\sum_{j=0}^{n} L_j(0) x_j^k = \begin{cases} 1 & \text{se } k = 0, \\ 0 & \text{se } k = 1, \ldots, n, \\ (-1)^n \prod_{j=0}^{n} x_j & \text{se } k = n+1, \\ (-1)^n \left(\prod_{j=0}^{n} x_j \right) \left(\sum_{j=0}^{n} x_j \right) & \text{se } k = n+2; \end{cases}$$

d) $$\sum_{j=0}^{n} \frac{x_j^k}{\prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i)} = \begin{cases} 0 & \text{se } k = 0, \ldots, n-1, \\ 1 & \text{se } k = n, \\ \sum_{j=0}^{n} x_j & \text{se } k = n+1. \end{cases}$$

(Traccia: a) si scriva il polinomio di Lagrange per le funzioni $f(x) = x^k$. Per $k = 0, \ldots, n$, è $f(x) = p_n(x)$; b) per $k = 0$ si esprima $L_j(x)$ mediante la (8); c) per $k \leq n$ è $p_n(x) = x^k$, per $k = n+1$ e $k = n+2$ si utilizzi l'esercizio 5.4; d) segue dal punto c), tenendo conto che per $x_j \neq 0$ è

$$L_j(0) = \frac{(-1)^{n+1} x_0 x_1 \ldots x_n}{\prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i)(-x_j)} \, . \,)$$

5.10 Si consideri il polinomio di interpolazione della funzione e^x sui nodi equidistanti $x_i = \dfrac{i}{n}$, $i = 0, \ldots, n$. Si determini n affinché $|r(x)| \leq 10^{-6}$ per $x \in [0, 1]$.

5.11 Per la funzione $f(x) = \dfrac{1}{x}$ si determinino i polinomi $p(x)$ di grado minimo tali che

a) $$p(1) = f(1), \quad p(2) = f(2), \quad p(3) = f(3),$$

b) $$p(1) = f(1), \quad p'(1) = f'(1), \quad p''(1) = f''(1),$$

c) $p(1) = f(1), \quad p(3) = f(3), \quad p'(3) = f'(3),$

d) $p(2) = f(2), \quad p'(2) = f'(2), \quad p'(3) = f'(3),$

e) $p(1) = f(1), \quad p'(1) = f'(1), \quad p(2) = f(2).$

Si disegnino i grafici dei resti e si dica quale dei polinomi approssima meglio la funzione nel punto $x = 1.5$.

Risposta: a) $\dfrac{x^2}{6} - x + \dfrac{11}{6}\,,$

b) $x^2 - 3x + 3,$

c) $\dfrac{x^2}{9} - \dfrac{7}{9}\,x + \dfrac{5}{3}\,,$

d) $\dfrac{5}{72}\,x^2 - \dfrac{19}{36}\,x + \dfrac{23}{18}\,,$

e) $\dfrac{x^2}{2} - 2x + \dfrac{5}{2}\,;$

il polinomio che approssima meglio $f(x)$ in 1.5 è il d).)

5.12 Si determini il polinomio di grado minimo che nei punti x_0 e x_1 assume i valori y_0 e y_1 e la cui derivata in x_1 è uguale a z_1.

(Traccia: il polinomio è in generale di secondo grado, ma per certi valori dei dati può essere di grado inferiore.)

5.13 Si dimostri che il polinomio osculatore $p(x)$ di grado al più $2n + 1$ tale che

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad i = 0, \ldots, n,$$

è unico.

(Traccia: se vi fossero due polinomi diversi $p(x)$ e $q(x)$, di grado al più $2n+1$ che soddisfano alle condizioni, $q(x)$ risulterebbe il polinomio osculatore di $p(x)$. Per il teorema 5.12 sarebbe

$$p(x) - q(x) = \pi_n^2(x)\,\frac{p^{(2n+2)}(\xi)}{(2n+2)!}\,,$$

ma $p^{(2n+2)}(x) = 0$ per ogni x.)

5.14 Si dimostri che se $f^{(2n+2)}(x) \geq 0$ per $x \in [a, b]$, il grafico del polinomio osculatore è sempre al di sotto di quello della funzione.

5.15 Per la funzione $f(x) = \sin \pi x$

a) si scriva il polinomio $p(x)$ di grado minimo tale che

$$p(0) = f(0), \quad p(1) = f(1), \quad p'(0) = f'(0), \quad p'(1) = f'(1);$$

b) si determini $\epsilon = \max_{x\in[0,1]} |f(x) - p(x)|$ e si confronti ϵ con la limitazione ottenuta maggiorando il modulo dell'espressione del resto data nel teorema 5.12.

(Traccia: a) $p(x) = \pi x(1-x)$; b) è $\epsilon = 1 - \dfrac{\pi}{4} \approx 0.215$, mentre dal teorema risulta

$$\max_{x\in[0,1]} |r(x)| < \frac{\pi^4}{4!\ 16} \approx 0.254.\ )$$

5.16 Si determini il numero di operazioni richieste per calcolare il valore del polinomio di Hermite in un punto.

(Traccia: si verifichi che

$$L'_j(x_j) = \sum_{\substack{i=0 \\ i\neq j}}^{n} \frac{1}{x_j - x_i}\ ;$$

quindi, a meno di termini di ordine inferiore, il calcolo degli $L'_j(x_j)$ per $j = 0, \ldots, n$, richiede n^2 addizioni e $n^2/2$ moltiplicazioni (in quanto le differenze $x_j - x_i$ vengono contate per il calcolo di $L_j(x)$). In totale risultano quindi $3n^2/2$ addizioni e $3n^2/2$ moltiplicazioni.)

5.17 Si dimostri che $f[x_0, x_1]$ è indipendente da x_0 e x_1 se e solo se $f(x)$ è un polinomio di grado minore o uguale a 1.

(Traccia: se $f(x) = a_1 x + a_2$, è $f'(x) = f[x_0, x_1] = a_1$ e viceversa.)

5.18 Si dimostri che se $f(x) = u(x)v(x)$, allora

$$f[x_0, x_1] = u(x_0)v[x_0, x_1] + u[x_0, x_1]v(x_1).$$

5.19 Sia $f(x) = \pi_n(x)$. Si verifichi che

a) $f[x_0, x_1, \ldots, x_k] = 0 \quad \text{per} \quad k = 0, \ldots, n,$

b) $f[x_0, x_1, \ldots, x_n, x] = 1 \quad \text{per ogni } x,$

c) $f[x_0, x_1, \ldots, x_n, x, y] = 0 \quad \text{per ogni } x \text{ e } y.$

(Traccia: b) e c) discendono dal teorema 5.19.)

5.20 Sia $f(x) = \dfrac{1}{x}$ e $x_0, \ldots, x_n$ distinti e non nulli. Si verifichi che

$$f[x_0, \ldots, x_n] = (-1)^n \Big/ \prod_{i=0}^{n} x_i .$$

(Traccia: si proceda per induzione.)

5.21 Si verifichi che

a) $$f[x_0, x_1, \ldots, x_n] = \sum_{j=0}^{n} \frac{f(x_j)}{\displaystyle\prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i)} = \sum_{j=0}^{n} \frac{f(x_j)}{\pi_n'(x_j)} ;$$

b) se i punti x_i sono equidistanti di passo h, allora

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{h^n \; n!} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f(x_j).$$

(Traccia: a) si uguaglino i coefficienti dei termini di grado più elevato dei polinomi di Lagrange e di Newton; b) in modo analogo, utilizzando il polinomio di Lagrange scritto per punti equidistanti.)

5.22 Si determini una limitazione inferiore per la costante di Lebesgue Λ_n quando i nodi x_i, $i = 0, \ldots, n$, sono equidistanti.

(Traccia: si verifichi, procedendo come nel paragrafo 2, che

$$L_i(x_0 + th) = (-1)^{n-i} \; \frac{\tau_n(t)}{n!} \; \binom{n}{i} \frac{1}{t - i} ,$$

quindi

$$\lambda_n(t) = \sum_{i=0}^{n} |L_i(x_0 + th)| = \frac{|\tau_n(t)|}{n!} \; \sum_{i=0}^{n} \binom{n}{i} \frac{1}{|t - i|} ,$$

e risulta

$$\Lambda_n = \max_{0 \le t \le n} \lambda_n(t) \ge \lambda_n\Big(\frac{1}{2}\Big) = \frac{2}{n!} \left|\tau_n\Big(\frac{1}{2}\Big)\right| \Big(1 + \sum_{i=1}^{n} \binom{n}{i} \frac{1}{2i - 1}\Big).$$

Per quanto visto nel paragrafo 3 è

$$\left|\tau_n\Big(\frac{1}{2}\Big)\right| = \frac{(2n-1)!}{2^{2n} \; (n-1)!} ,$$

e si ha

$$1 + \sum_{i=1}^{n} \binom{n}{i} \frac{1}{2i - 1} > 1 + \sum_{i=1}^{\lfloor n/2 \rfloor} \binom{n}{i} \frac{1}{2i - 1} > \frac{1}{n} \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{i} \ge \frac{2^{n-1}}{n} .$$

Perciò

$$\Lambda_n \geq \frac{(2n-1)!}{2^n\ (n!)^2}\,.$$

Si verifichi, usando la formula di Stirling (esercizio 4.43) che asintoticamente per $n \to \infty$ risulta

$$\frac{(2n-1)!}{2^n\ (n!)^2} \sim \frac{2^{n-1}}{\sqrt{\pi}\ n^{3/2}}\,.)$$

5.23 a) Si dimostri che se $f(x) \in C^n[x_0, x_n]$, $x_i = x_0 + ih$, $h > 0$, $i = 0, \ldots, n$, esiste $\xi \in (x_0, x_n)$ tale che

$$\Delta^n f(x_0) = h^n f^{(n)}(\xi);$$

b) per $f(x) = \log x$, $x_0 \geq 1$, si dimostri che la successione $\Delta^n f(x_0)$ ha i termini di segno alterno e di modulo decrescente.

(Traccia: a) si sfruttino i teoremi 5.21 e 5.19; b) essendo per $n \geq 1$

$$f^{(n)}(x) = (-1)^{n-1}\,\frac{(n-1)!}{x^n}\,,$$

per la a) il segno di $\Delta^n f(x_0)$ è negativo per $n \geq 2$ pari e positivo per n dispari. Inoltre

$$\begin{aligned}\left|\Delta^n f(x_0)\right| - \left|\Delta^{n+1} f(x_0)\right| &= \operatorname{sgn}\left(\Delta^n f(x_0)\right)\left[\Delta^n f(x_0) + \Delta^{n+1} f(x_0)\right] \\ &= \operatorname{sgn}\left(\Delta^n f(x_0)\right)\Delta^n f(x_1) = \left|\Delta^n f(x_1)\right| > 0,\end{aligned}$$

e quindi $\left|\Delta^n f(x_0)\right| > \left|\Delta^{n+1} f(x_0)\right|$.)

5.24 Si supponga che i valori $f(x_i)$ effettivamente calcolati della funzione $f(x)$ siano affetti da errori assoluti δ_i, cioè

$$\widetilde{f}(x_i) = f(x_i) + \delta_i, \quad \text{con} \quad |\delta_i| \leq \delta, \quad \text{per } i = 0, \ldots, n.$$

Si dia una maggiorazione dell'errore assoluto da cui può essere affetto il valore $\Delta^n f(x_0)$.

(Traccia: per $\delta_i = (-1)^i \delta$ si ha il caso peggiore. La propagazione di tali errori nella costruzione della tabella delle differenze finite di $f(x)$ avviene

secondo lo schema

δ_i	$\Delta\delta_i$	$\Delta^2\delta_i$	$\Delta^3\delta_i$	$\dots$	$\Delta^n\delta_i$
δ					
	-2δ				
$-\delta$		4δ			
	2δ		-8δ		
δ		-4δ		$\ddots$	
	-2δ				$(-1)^n 2^n\delta$
$-\delta$		$\vdots$			
	$\vdots$			$\iddots$	
$\vdots$		$(-1)^n 4\delta$			
	$(-1)^n 2\delta$				
$(-1)^n\delta$					

5.25 Molti polinomi di interpolazione con le differenze finite possono essere scritti utilizzando un metodo di costruzione che sfrutta il *diagramma a losanghe* riportato nella figura 5.27.

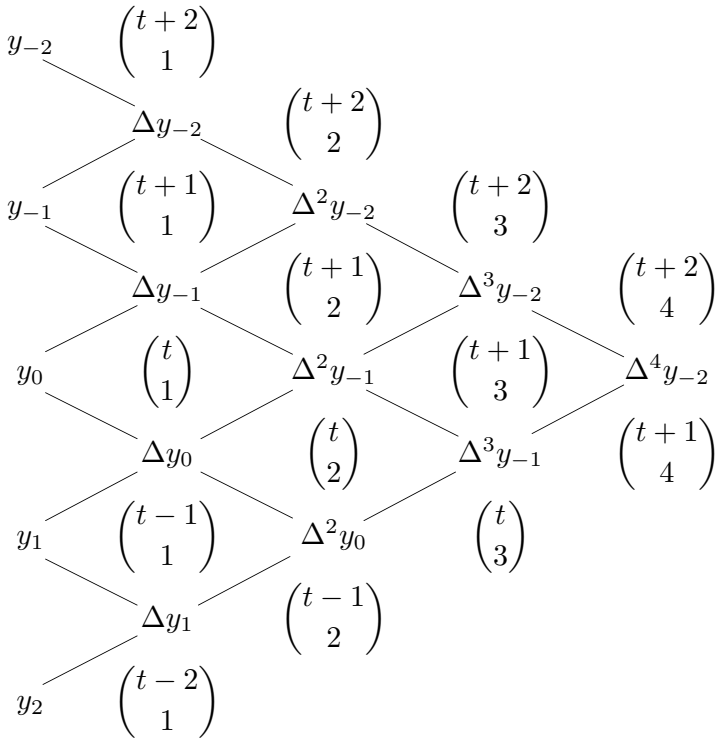


Fig. 5.27 - Diagramma a losanghe.

Per costruire una formula di interpolazione si segue sul diagramma un cammino a zig zag che parte da un punto y_r della prima colonna a sinistra e raggiunge una differenza $\Delta^n y_i$ dell'ultima colonna a destra, toccando tutte le colonne intermedie, in corrispondenza ad una differenza o ad un coefficiente binomiale. Fissato l'indice r, si costruisce il polinomio di grado al più n in t della forma

$$q(t) = t_0 + t_1 + t_2 + \ldots + t_n,$$

dove t_j, per $j = 0, \ldots, n$, è il termine individuato dall'incontro del cammino con la j-esima colonna del diagramma ed è formato nel modo seguente:

$t_0 = y_r$,

$t_j = \binom{t-k-1}{j} \Delta^j y_k$ se nella j-esima colonna si è raggiunta la differenza $\Delta^j y_k$ provenendo dalla riga inferiore (il coefficiente binomiale è quello scritto sotto alla differenza),

$t_j = \binom{t-k}{j} \Delta^j y_k$ se nella j-esima colonna si è raggiunta la differenza $\Delta^j y_k$ provenendo dalla riga superiore (il coefficiente binomiale è quello scritto sopra alla differenza),

$t_j = \binom{t-k}{j} \frac{1}{2} \left[\Delta^j y_{k-1} + \Delta^j y_k\right]$ se nella j-esima colonna si è raggiunto il coefficiente binomiale $\binom{t-k}{j}$ provenendo dalla stessa riga (si moltiplica per la media delle differenze, scritte sopra e sotto il coefficiente binomiale),

$t_j = \frac{1}{2} \left[\binom{t-k-1}{j} + \binom{t-k}{j}\right] \Delta^j y_k$ se nella j-esima colonna si è raggiunta la differenza $\Delta^j y_k$ provenendo dalla stessa riga (si moltiplica per la media dei coefficienti binomiali sopra e sotto la differenza).

Si costruiscano in particolare i polinomi di:

a) *Newton discendente*, ottenuto partendo da y_0 e toccando le differenze discendenti Δy_0, $\Delta^2 y_0$, $\Delta^3 y_0$, $\Delta^4 y_0$, ...;

b) *Newton ascendente*, ottenuto partendo da y_0 e toccando le differenze ascendenti Δy_{-1}, $\Delta^2 y_{-2}$, $\Delta^3 y_{-3}$, $\Delta^4 y_{-4}$, ...;

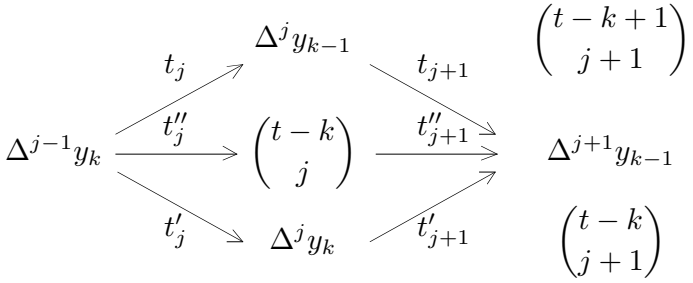
c) *Gauss* 1^o, ottenuto partendo da y_0 e toccando le differenze a zig-zag Δy_0, $\Delta^2 y_{-1}$, $\Delta^3 y_{-1}$, $\Delta^4 y_{-2}$, ...;

d) *Gauss* 2^o, ottenuto partendo da y_0 e toccando le differenze a zig-zag Δy_{-1}, $\Delta^2 y_{-1}$, $\Delta^3 y_{-2}$, $\Delta^4 y_{-2}$, ...;

e) *Stirling*, ottenuto partendo da y_0, con un cammino orizzontale che tocca alternativamente i coefficienti binomiali e le differenze $\binom{t}{1}$, $\Delta^2 y_{-1}$, $\binom{t+1}{3}$, $\Delta^4 y_{-2}$, ...

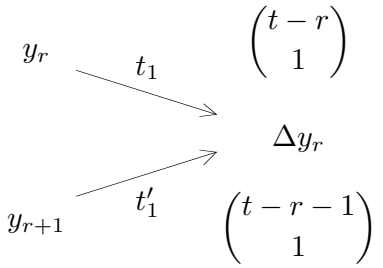
Si dimostri che

f) indicati con t_j e t_{j+1} i termini della formula dovuti al cammino che percorre i due lati superiori di una losanga, con t'_j e t'_{j+1} i termini della formula dovuti al cammino che percorre i due lati inferiori della losanga e con t''_j e t''_{j+1} i termini della formula dovuti al cammino che attraversa la losanga



allora $t_j + t_{j+1} = t'_j + t'_{j+1} = t''_j + t''_{j+1}$;

g) indicati con t_0 e t_1 i termini della formula dovuti al cammino da y_r a Δy_r e con t'_0 e t'_1 i termini della formula dovuti al cammino da y_{r+1} a Δy_r



si ha $t_0 + t_1 = t'_0 + t'_1$;

h) due cammini che arrivano alla stessa differenza $\Delta^n y_i$ sono equivalenti, nel senso che producono lo stesso polinomio, pur espresso come somma di termini diversi.

Quindi tutti i polinomi ottenuti con cammini che arrivano alla differenza $\Delta^n y_i$ sono uguali al polinomio di Newton discendente il cui primo termine è y_i, che per la (26) coincide con il polinomio $p_n(x_i + th)$ di interpolazione sui nodi $x_i, \ldots, x_{i+n}$.

i) Per i polinomi ottenuti in a), b), c) e d) si dica su quali nodi essi sono polinomi di interpolazione e se ne diano i resti.

(Traccia:

a) polinomio di Newton discendente

$$q_{Nd}(t) = y_0 + \binom{t}{1}\Delta y_0 + \binom{t}{2}\Delta^2 y_0 + \binom{t}{3}\Delta^3 y_0 + \binom{t}{4}\Delta^4 y_0 + \ldots + \binom{t}{n}\Delta^n y_0,$$

b) polinomio di Newton ascendente

$$q_{Na}(t) = y_0 + \binom{t}{1}\Delta y_{-1} + \binom{t+1}{2}\Delta^2 y_{-2} + \binom{t+2}{3}\Delta^3 y_{-3} + \binom{t+3}{4}\Delta^4 y_{-4} + \ldots + \binom{t+n-1}{n}\Delta^n y_{-n},$$

c) polinomio di Gauss 1^o

$$q_{G1}(t) = y_0 + \binom{t}{1}\Delta y_0 + \binom{t}{2}\Delta^2 y_{-1} + \binom{t+1}{3}\Delta^3 y_{-1} + \binom{t+1}{4}\Delta^4 y_{-2} + \ldots + \binom{t+i}{2i+1}\Delta^{2i+1} y_{-i} + \binom{t+i}{2i+2}\Delta^{2i+2} y_{-i-1} \quad \text{per} \quad n = 2i+2;$$

d) polinomio di Gauss 2^o

$$q_{G2}(t) = y_0 + \binom{t}{1}\Delta y_{-1} + \binom{t+1}{2}\Delta^2 y_{-1} + \binom{t+1}{3}\Delta^3 y_{-2} + \binom{t+2}{4}\Delta^4 y_{-2} + \ldots + \binom{t+i}{2i+1}\Delta^{2i+1} y_{-i-1} + \binom{t+i+1}{2i+2}\Delta^{2i+2} y_{-i-1} \quad \text{per} \quad n = 2i+2;$$

e) polinomio di Stirling

$$q_S(t) = y_0 + \binom{t}{1}\frac{1}{2}\left[\Delta y_{-1} + \Delta y_0\right] + \frac{1}{2}\left[\binom{t+1}{2} + \binom{t}{2}\right]\Delta^2 y_{-1} + \binom{t+1}{3}\frac{1}{2}\left[\Delta^3 y_{-2} + \Delta^3 y_{-1}\right] + \ldots + \frac{1}{2}\left[\binom{t+i+1}{2i+2} + \binom{t+i}{2i+2}\right]\Delta^{2i+2} y_{-i-1}, \quad \text{per } n = 2i+2;$$

f) risulta

$$\begin{aligned} t_j + t_{j+1} &= \binom{t-k}{j} \Delta^j y_{k-1} + \binom{t-k+1}{j+1} \Delta^{j+1} y_{k-1} \\ &= \binom{t-k}{j} \Delta^j y_{k-1} + \binom{t-k+1}{j+1} \left[\Delta^j y_k - \Delta^j y_{k-1}\right] \\ &= \binom{t-k+1}{j+1} \Delta^j y_k - \left[\binom{t-k+1}{j+1} - \binom{t-k}{j}\right] \Delta^j y_{k-1}, \end{aligned}$$

$$\begin{aligned} t'_j + t'_{j+1} &= \binom{t-k}{j} \Delta^j y_k + \binom{t-k}{j+1} \Delta^{j+1} y_{k-1} \\ &= \binom{t-k}{j} \Delta^j y_k + \binom{t-k}{j+1} \left[\Delta^j y_k - \Delta^j y_{k-1}\right] \\ &= \left[\binom{t-k}{j} + \binom{t-k}{j+1}\right] \Delta^j y_k - \binom{t-k}{j+1} \Delta^j y_{k-1}, \end{aligned}$$

da cui segue che

$$t_j + t_{j+1} = t'_j + t'_{j+1},$$

perché

$$\binom{t-k}{j} + \binom{t-k}{j+1} = \binom{t-k+1}{j+1}.$$

Inoltre

$$\begin{aligned} t''_j + t''_{j+1} &= \frac{1}{2} \binom{t-k}{j} \left[\Delta^j y_k + \Delta^j y_{k-1}\right] \\ &\quad + \frac{1}{2} \left[\binom{t-k}{j+1} + \binom{t-k+1}{j+1}\right] \Delta^{j+1} y_{k-1} \\ &= \frac{1}{2} \left[\binom{t-k}{j} + \binom{t-k}{j+1} + \binom{t-k+1}{j+1}\right] \Delta^j y_k \\ &\quad - \frac{1}{2} \left[\binom{t-k}{j+1} + \binom{t-k+1}{j+1} - \binom{t-k}{j}\right] \Delta^j y_{k-1} \\ &= t_j + t_{j+1}; \end{aligned}$$

g) è

$$t_0 + t_1 = y_r + \binom{t-r}{1} \Delta y_r = (t-r) y_{r+1} - (t-r-1) y_r,$$

$$t'_0 + t'_1 = y_{r+1} + \binom{t-r-1}{1} \Delta y_r = (t-r) y_{r+1} - (t-r-1) y_r;$$

h) si combinino i risultati dei punti f) e g), considerando una successione di cammini che fa passare dal primo al secondo dei due cammini dati e che differiscono l'uno dall'altro solo per il punto di partenza o per gli elementi di una losanga.

i) Per $h > 0$, $0 \le t \le n$ (n pari per le formule di Gauss e di Stirling), i polinomi ottenuti sono di interpolazione sui nodi $x_k, \dots, x_{k+n}$, dove k è l'indice dell'n-esima differenza che compare nella formula. Il relativo resto è

$$r(x_0 + th) = \binom{t-k}{n+1} h^{n+1} f^{(n+1)}(\xi), \quad \xi \in (x_k, x_{k+n}).$$

Quindi per la formula di Newton discendente è $k = 0$, per la formula di Newton ascendente è $k = -n$, per le formule di Gauss e per la formula di Stirling è $k = -\dfrac{n}{2}$.)

5.26 Siano

$p_{0,\dots,k-1}(x)$ il polinomio di interpolazione di $f(x)$ sui punti $x_0, \dots, x_{k-1}$,

$p_{1,\dots,k}(x)$ il polinomio di interpolazione di $f(x)$ sui punti $x_1, \dots, x_k$.

Si dimostri che il polinomio

$$p_{0,\dots,k}(x) = \frac{(x - x_0)p_{1,\dots,k}(x) - (x - x_k)p_{0,\dots,k-1}(x)}{x_k - x_0}$$

è il polinomio di interpolazione su $x_0, \dots, x_k$. Su questa proprietà è basato l'algoritmo di *Neville* per la costruzione del polinomio di interpolazione su $x_0, \dots, x_n$: procedendo per colonne si costruiscono i polinomi della seguente tabella

$$\begin{array}{c|cccccc}
x_0 & y_0 & & & & & \\
 & & p_{0,1}(x) & & & & \\
x_1 & y_1 & & p_{0,1,2}(x) & & & \\
 & & p_{1,2}(x) & & p_{0,1,2,3}(x) & & \\
x_2 & y_2 & & p_{1,2,3}(x) & & \ddots & \\
 & & p_{2,3}(x) & & & & p_{0,1,\dots,n}(x) \\
x_3 & y_3 & \vdots & & & & \\
 & & & \vdots & & \iddots & \\
\vdots & \vdots & & p_{n-2,n-1,n}(x) & & & \\
 & & p_{n-1,n}(x) & & & & \\
x_n & y_n & & & & &
\end{array}$$

Il vantaggio di tale metodo è che il polinomio di interpolazione viene costruito mediante una successione di polinomi di interpolazione parziale e di

grado crescente, per cui il procedimento si arresta quando è stata raggiunta l'approssimazione richiesta. Si determini il costo computazionale di questo metodo che è superiore a quello del polinomio di Newton.

(Traccia: $p_{0,\ldots,k-1}(x)$ e $p_{1,\ldots,k}(x)$ hanno grado al più $k-1$, per cui $p_{0,\ldots,k}(x)$ ha grado al più k, inoltre $p_{0,\ldots,k}(x_i) = f(x_i)$, per $i = 0, \ldots, k$. Il costo computazionale per il calcolo del polinomio in un punto è di n^2 addizioni e $3n^2/2$ moltiplicazioni.)

5.27 Si dimostri direttamente, senza ricorrere al teorema 5.28, che se i punti x_i sono distinti per $i = 0, \ldots, n$, e se la funzione $f(x)$ è derivabile, allora

a) la funzione $f[x_0, x_1, \ldots, x_n, x]$ è continua per $x \neq x_i$;

b) $\lim_{x_0 \to x} f[x_0, x] = f'(x)$;

c) $\lim_{x_n \to x} f[x_0, \ldots, x_{n-1}, x_n, x] = \dfrac{d}{dx}\, f[x_0, \ldots, x_{n-1}, x]$.

(Traccia: a) per la differenza del primo ordine si ha:

$$f[x_i, x] = \frac{f(x) - f(x_i)}{x - x_i}\,,$$

e la continuità di $f[x_i, x]$ per $x \neq x_i$ segue dalla continuità della $f(x)$. Se $n > 0$, la continuità di $f[x_0, x_1, \ldots, x_n, x]$ segue, per il teorema 5.27, dalla continuità delle differenze $f[x_i, x]$.

b) per le differenze del primo ordine si ha per $\epsilon \neq 0$

$$f[x + \epsilon, x] = \frac{f(x + \epsilon) - f(x)}{\epsilon}\,,$$

si passi al limite per $\epsilon \to 0$, tenendo conto che $f(x)$ è derivabile.

c) per le differenze del secondo ordine, sia $\epsilon \neq 0$ tale che $x + \epsilon \neq x_i$, per $i = 0, \ldots, n$; applicando la proprietà di simmetria, si ha

$$f[x_i, x, x + \epsilon] = f[x + \epsilon, x_i, x] = \frac{f[x + \epsilon, x] - f[x + \epsilon, x_i]}{x - x_i}\,.$$

Poiché $x \neq x_i$ e il limite di $f[x + \epsilon, x]$ per $\epsilon \to 0$ esiste finito, allora esiste finito anche il limite di $f[x_i, x, x + \epsilon]$ per $\epsilon \to 0$. Inoltre, poiché

$$\lim_{\epsilon \to 0} \frac{f[x_i, x + \epsilon] - f[x_i, x]}{\epsilon} = \lim_{\epsilon \to 0} f[x_i, x, x + \epsilon],$$

ne segue che la differenza $f[x_i, x]$ è derivabile in x e

$$\lim_{\epsilon \to 0} f[x_i, x, x+\epsilon] = \frac{d}{dx}\, f[x_i, x].$$

Per le differenze di ordine superiore al secondo, si applichi il teorema 5.27.)

5.28 Per ogni n si considerino i nodi $x_i = \dfrac{1}{i+1}$, $i = 0, \ldots, n$, e si costruisca il polinomio $p_n(x)$ di interpolazione della funzione $f(x) = x \sin \dfrac{\pi}{x}$. Si dica se per $n \to \infty$ la successione dei polinomi converge alla funzione $f(x)$.

(Risposta: i polinomi sono tutti identicamente nulli e quindi la successione converge al polinomio nullo e non a $f(x)$.)

5.29 Sia $\sigma \neq -1$. Per ogni n si considerino i nodi $x_i = i$, per $i = 0, \ldots, n$ e si costruisca il polinomio $p_n(x)$ di interpolazione della funzione $f(x) = (1+\sigma)^x$. Si dica per quali valori di σ la successione dei polinomi converge alla funzione $f(x)$ per $n \to \infty$.

(Traccia: Si verifichi prima che

$$\Delta^n (1+\sigma)^x = \sigma^n (1+\sigma)^x,$$

e quindi il polinomio di interpolazione di Newton con le differenze finite è

$$p_n(x) = 1 + \sigma x + \frac{\sigma^2}{2}\, x(x-1) + \ldots + \frac{\sigma^n}{n!}\, x(x-1)\ldots(x-n+1).$$

Se $x = k$ intero, $f(x)$ è un polinomio di grado k in σ e quindi $p_n(x) = f(x)$ per $n \geq k$. Se x non è intero, $p_n(x)$ costituisce la somma dei primi $n+1$ termini dello sviluppo di Maclaurin della funzione $g(\sigma) = (1+\sigma)^x$, che è convergente per $|\sigma| < 1$.)

5.30 Si dia l'espressione del resto dell'interpolazione lineare inversa della funzione $y = f(x)$ nell'intervallo $[x_0, x_1]$.

(Risposta: sia $f'(x) \neq 0$ per $x \in [x_0, x_1]$, allora

$$r(y) = -\big[y - f(x_0)\big]\ \big[y - f(x_1)\big]\ \frac{f''(\xi)}{2[f'(\xi)]^3}\ , \quad \xi \in (x_0, x_1).\ )$$

5.31 Con il termine *interpolazione multidimensionale di grado n* si intende il seguente problema:

siano k e n interi, $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(n)} \in \mathbf{R}^k$ e $y^{(0)}, \ldots, y^{(n)} \in \mathbf{R}$. Si vuole determinare un polinomio in k variabili $p : \mathbf{R}^k \to \mathbf{R}$ di grado al più m in ogni variabile, tale che $p(\mathbf{x}^{(i)}) = y^{(i)}$, $i = 0, \ldots, n$. Anche quando il numero dei parametri da determinare è uguale al numero di condizioni imposte, il problema dell'interpolazione multidimensionale può non avere soluzione.

a) Si cerchi di determinare il polinomio in due variabili $p(x_1, x_2)$ di grado al più 2 in ciascuna variabile, che soddisfa alle condizioni $p(x_1^{(i)}, x_2^{(i)}) = y^{(i)}$, per $i = 0, \ldots, 6$, dove

i	$\mathbf{x}^{(i)}$	$y^{(i)}$
0	(0 , 0)	0
1	(0 , 1)	-1
2	(0 , 3)	0
3	(1 , 0)	1
4	(2 , 0)	3
5	(2 , 1)	2
6	(3 , 2)	2
7	(4 , 3)	1
8	(6 , 5)	8

b) Si cerchi di determinare il polinomio in due variabili $p(x_1, x_2)$ di grado al più 2 in ciascuna variabile, che soddisfa alle condizioni $p(x_1^{(i)}, x_2^{(i)}) = y^{(i)}$, per $i = 0, \ldots, 6$, dove

i	$\mathbf{x}^{(i)}$	$y^{(i)}$
0	(0 , 0)	0
1	(0 , 1)	-1
2	(0 , 3)	-9
3	(1 , 0)	1
4	(2 , 0)	2
5	(2 , 1)	13
6	(3 , 2)	47
7	(4 , 3)	91
8	(6 , 5)	41

c) Si consideri il caso particolare in cui $k = 2$ e i nodi $\mathbf{x}^{(i)}$ sono scelti su una griglia rettangolare di $r \times s$ punti. Pertanto $n = rs - 1$. Per $r = s = 2$, come nella figura 5.28, si determini il polinomio di interpolazione *bilineare*, cioè di grado al più 1 in x e in y

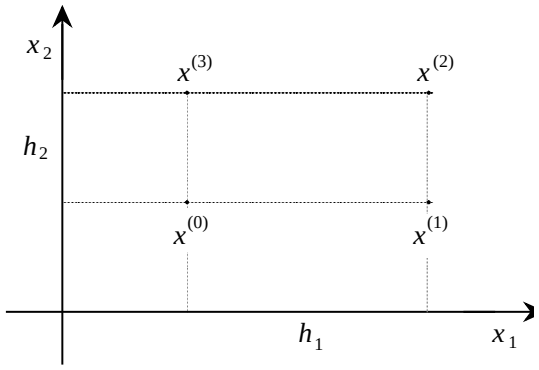


Fig. 5.28 - Interpolazione bilineare.

usando le tecniche seguenti:

(1) si consideri il polinomio lineare nelle variabili x_1 e x_2

$$p_1(\mathbf{x}) = \alpha_{0,0} + \alpha_{1,0}x_1 + \alpha_{0,1}x_2 + \alpha_{1,1}x_1x_2$$

e si determinino i coefficienti in modo che il polinomio soddisfi le condizioni date;

(2) si scrivano i due polinomi $q_1(x_1)$ e $q_2(x_1)$ lineari nella x_1, che interpolano separatamente lungo l'asse x_1, cioè tali che

$$q_1(x_1^{(0)}) = y^{(0)}, \quad q_1(x_1^{(1)}) = y^{(1)},$$
$$q_2(x_1^{(3)}) = y^{(3)}, \quad q_2(x_1^{(2)}) = y^{(2)}.$$

Si scriva poi il polinomio di interpolazione lineare lungo l'asse x_2, che interpola fra i valori $q_1(x_1)$ e $q_2(x_1)$.

Si verifichi che con entrambi i metodi si ottiene lo stesso polinomio.

d) Si generalizzi il procedimento (2) al caso in cui $r, s \geq 2$.

(Traccia: a) non esiste alcun polinomio della forma

$$p(\mathbf{x}) = \alpha_{0,0} + \alpha_{1,0}x_1 + \alpha_{0,1}x_2 + \alpha_{1,1}x_1x_2 + \alpha_{2,0}x_1^2 + \alpha_{0,2}x_2^2 + \alpha_{2,1}x_1^2x_2 + \alpha_{1,2}x_1x_2^2 + \alpha_{2,2}x_1^2x_2^2,$$

che soddisfa le condizioni date; b) esistono infiniti polinomi della forma

$$p(\mathbf{x}) = x_1 + (k-4)x_1x_2 - x_2^2 + (6-k)x_1^2x_2 + kx_1x_2^2 - x_1^2x_2^2,$$

in cui k è un parametro, che soddisfano le condizioni date; c) risulta

$$p(x_1, x_2) = (1-t_1)(1-t_2)y^{(0)} + t_1(1-t_2)y^{(1)} + (1-t_1)t_2y^{(3)} + t_1t_2y^{(2)},$$

dove $x_1 = x_1^{(0)} + t_1h_1$ e $x_2 = x_2^{(0)} + t_2h_2$; d) per $r, s \geq 2$ si considerino gli s polinomi di grado al più $r-1$ di interpolazione sui nodi collocati lungo le parallele all'asse x_1 e poi si interpolino gli s valori dei polinomi così ottenuti lungo l'asse x_2.)

5.32 La funzione $f(x)$ assume i valori

x	0	1	2	3	4
$f(x)$	4	2	1	-1	-4

Si scrivano le funzioni razionali di interpolazione della $f(x)$ della forma $w(x) = \dfrac{p(x)}{q(x)}$, dove $p(x)$ e $q(x)$ sono polinomi di gradi al più m e n, nei seguenti casi:

$$(1) \quad m = n = 2; \qquad (2) \quad m = 3,\ n = 1; \qquad (3) \quad m = 1,\ n = 3.$$

Risposta: (1) $$w(x) = 4\,\frac{3x^2 - 10x + 6}{x^2 - 9x + 6}\,;$$

(2) $w(x)$ non esiste, infatti risulta

$$w(x) = \frac{x^3 - x^2 - 4x}{-2x} = -\,\frac{x^2}{2} + \frac{x}{2} + 2$$

che non soddisfa a tutte le condizioni;

(3) $$w(x) = 8\,\frac{-8x + 21}{3x^3 - 20x^2 + 27x + 42}\,.\ )$$

5.33 Si verifichi che non sempre è possibile trovare delle funzioni razionali della forma

$$w(x) = \frac{a_1 x + a_0}{b_1 x + b_0}\,, \quad b_1 \neq 0,$$

che interpolano una funzione $f(x)$ su tre nodi distinti.

(Traccia: si determini il rango della matrice del sistema (36).)

5.34 Sia $f(x) \in C^{m+n+1}[a, b]$, dove

$$a = \min_{i=0,\ldots,m+n} x_i, \qquad b = \max_{i=0,\ldots,m+n} x_i,$$

e sia $w(x) = \dfrac{p(x)}{q(x)}$ la funzione razionale di interpolazione di $f(x)$ sui nodi x_i, $i = 0, \ldots, m + n$, con $q(x) \neq 0$ per $x \in [a, b]$. Si dimostri che esiste $\xi = \xi(x) \in (a, b)$, tale che

$$f(x) - w(x) = \frac{\pi_{m+n}(x)}{(m + n + 1)!}\,\big[f^{(m+n+1)}(\xi) - w^{(m+n+1)}(\xi)\big].$$

(Traccia: si proceda come per la dimostrazione del teorema 5.5.)

5.35 Si dimostri che, posto

$$w_i = d_0 + \frac{c_1}{d_1} + \frac{c_2}{d_2} + \cdots + \frac{c_i}{d_i},$$

risulta

$$w_i = \frac{p_i}{q_i},$$

dove p_i e q_i soddisfano alle relazioni ricorrenti

a) $p_i = d_i\, p_{i-1} + c_i\, p_{i-2}$,
$q_i = d_i\, q_{i-1} + c_i\, q_{i-2}$,
con $p_{-1} = 1, \quad q_{-1} = 0, \quad p_0 = d_0, \quad q_0 = 1, \quad$ per $\quad i \geq 1$;

b) $p_i = (c_i + d_{i-1}d_i + \dfrac{d_i}{d_{i-2}}\, c_{i-1})\, p_{i-2} - \dfrac{d_i}{d_{i-2}}\, c_{i-1}c_{i-2}\, p_{i-4}$,
$q_i = (c_i + d_{i-1}d_i + \dfrac{d_i}{d_{i-2}}\, c_{i-1})\, q_{i-2} - \dfrac{d_i}{d_{i-2}}\, c_{i-1}c_{i-2}\, q_{i-4}$,
con $\quad p_{-2} = 0, \quad q_{-2} = \dfrac{1}{c_0}, \quad$ per $\quad i \geq 2.$

(Traccia: a) si proceda per induzione; per $i = 0,\, 1$ si verifichi direttamente, per $i \geq 2$, supposte vere le relazioni fino all'indice $i-1$, si ha

$$\begin{aligned} w_i &= d_0 + \frac{c_1}{d_1} + \cdots + \frac{c_{i-2}}{d_{i-2}} + \frac{c_{i-1}}{d_{i-1}} + \frac{c_i}{d_i} \\ &= d_0 + \frac{c_1}{d_1} + \cdots + \frac{c_{i-2}}{d_{i-2}} + \frac{c_{i-1}d_i}{d_{i-1}d_i + c_i} = \frac{\widehat{p}_{i-1}}{\widehat{q}_{i-1}}, \end{aligned}$$

dove

$$\begin{aligned} &\widehat{p}_{i-1} = (d_{i-1}d_i + c_i)p_{i-2} + c_{i-1}d_i p_{i-3} \\ &= d_i(d_{i-1}p_{i-2} + c_{i-1}p_{i-3}) + c_i p_{i-2} = d_i p_{i-1} + c_i p_{i-2}, \end{aligned}$$

e analogamente per $\widehat{q}_{i-1}$; b) si sfruttino le a).)

5.36 Indicata con $w_i = \dfrac{p_i}{q_i}$ la i-esima frazione parziale della frazione continua

$$w = d_0 + \frac{c_1}{d_1} + \frac{c_2}{d_2} + \cdots + \frac{c_k}{d_k},$$

e supposto che $q_i \neq 0$ per $i = 1, \ldots, k$, si verifichi che per $i \geq 1$ vale

a) $$p_i q_{i-1} - q_i p_{i-1} = (-1)^{i-1} c_1 c_2 \ldots c_i;$$

b) $$w_i - w_{i-1} = (-1)^{i-1}\, \frac{c_1 c_2 \ldots c_i}{q_{i-1} q_i};$$

c) $$w = w_k = d_0 + \sum_{i=1}^{k} (-1)^{i-1}\, \frac{c_1 c_2 \ldots c_i}{q_{i-1} q_i};$$

d) Si verifichi che se i c_i e i d_i sono tutti numeri positivi, allora per le frazioni parziali di indice dispari si ha

$$w_1 > w_3 > w_5 > \dots,$$

e per quelle di indice pari si ha

$$w_0 < w_2 < w_4 < \dots$$

(Traccia: a) si proceda per induzione utilizzando le (40); b) e c) si sfrutti la a); d) dalla b) e dalla (40) segue che

$$\begin{aligned} w_i - w_{i-2} &= (-1)^{i-2}\, \frac{c_1 c_2 \dots c_{i-1}}{q_{i-1}} \left(\frac{1}{q_{i-2}} - \frac{c_i}{q_i} \right) \\ &= (-1)^i\, \frac{c_1 c_2 \dots c_{i-1}}{q_{i-1}}\, \frac{d_i q_{i-1}}{q_{i-2} q_i} = (-1)^i\, \frac{c_1 c_2 \dots c_{i-1} d_i}{q_{i-2} q_i}\,, \end{aligned}$$

e poiché i c_i e i d_i, e quindi anche i q_i, sono positivi, ne segue che

per i pari $\qquad w_i - w_{i-2} > 0,$

per i dispari $\qquad w_i - w_{i-2} < 0.$)

5.37 Si dica quante operazioni additive e moltiplicative sono richieste per il calcolo della frazione continua (39), utilizzando la (41) oppure utilizzando la (40).

(Risposta: con la (41) k operazioni additive e k operazioni moltiplicative; con la (40) $2k$ operazioni additive e $4k$ operazioni moltiplicative.)

5.38 Dati $k-1$ numeri $\alpha_1, \dots, \alpha_{k-1}$, si consideri la frazione continua (39) con

$$d_0 = 0,\ c_1 = d_1 = 1,\ c_i = -\alpha_{i-1} \text{ e } d_i = 1 + \alpha_{i-1},\ \text{per } i = 2, \dots, k.$$

Si verifichi che la i-esima frazione parziale è data da

$$w_i = 1 + \alpha_1 + \alpha_1\alpha_2 + \dots + \alpha_1\alpha_2 \dots \alpha_{i-1}.$$

(Traccia: si verifichi che posto $w_i = \dfrac{p_i}{q_i}$, è $p_1 = q_1 = q_2 = 1$ e $p_2 = 1 + \alpha_1$, e si proceda per induzione, sfruttando le (40).)

5.39 Sia $w = w_k = \dfrac{p_k}{q_k}$ la frazione continua (39), e si considerino le due matrici tridiagonali $A_k \in \mathbf{R}^{k\times k}$ e $B_k \in \mathbf{R}^{(k+1)\times(k+1)}$:

$$A_k = \begin{bmatrix} d_1 & -1 & & \\ c_2 & d_2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & c_k & d_k \end{bmatrix}, \quad B_k = \begin{bmatrix} d_0 & -1 & & \\ c_1 & d_1 & \ddots & \\ & \ddots & \ddots & -1 \\ & & c_k & d_k \end{bmatrix}.$$

Si dimostri che

a) $q_k = \det A_k$ e $p_k = \det B_k$;

b) $w_k - d_0$ è uguale alla prima componente x_1 del vettore $\mathbf{x}$ soluzione del sistema lineare

$$A_k \mathbf{x} = c_1 \mathbf{e}_1, \quad \text{dove} \quad \mathbf{e}_1 = [1, 0, \ldots, 0]^T \in \mathbf{R}^k.$$

Perciò per calcolare w ci si può servire di uno qualsiasi dei metodi di risoluzione dei sistemi lineari tridiagonali.

(Traccia: a) si verifichi che

$$\det A_k = d_k \det A_{k-1} + c_k \det A_{k-2}, \quad \det A_1 = d_1, \quad \det A_2 = d_1 d_2 + c_2,$$

e si confronti con la (40). Si proceda analogamente con p_k. b) Si verifichi prima che $p_k = d_0 q_k + c_1 \det D_k$, dove $D_k \in \mathbf{R}^{(k-1)\times(k-1)}$ è ottenuta da A_k cancellando la prima riga e la prima colonna. Con la regola di Cramer si ha $x_1 = \dfrac{c_1 \det D_k}{q_k}$.)

5.40 Siano a e b due interi il cui massimo comun divisore è r. Si dimostri che l'equazione di Diofanto

$$ax + by = r$$

ha soluzione, con x e y interi e primi fra loro.

(Traccia: si applichi l'algoritmo delle divisioni successive alla coppia di numeri a e b nel modo seguente

$$t_0 = a, \quad t_1 = b,$$
$$t_i = t_{i+1} d_i + t_{i+2}, \ i = 0, \ldots, k, \ \text{dove } d_k = r \text{ e } t_{k+2} = 0.$$

Risulta quindi

$$\frac{a}{b} = d_0 + \cfrac{1}{d_1 + \cdots + \cfrac{1}{d_{k-1} + \cfrac{1}{r}}}.$$

Indicata con $w_i = \dfrac{p_i}{q_i}$ la i-esima somma parziale di questa frazione continua, si ha per l'esercizio 5.36 a)

$$p_k \, q_{k-1} - q_k \, p_{k-1} = (-1)^{k-1},$$

quindi p_k e q_k sono primi fra loro, perché un divisore comune di p_k e q_k dovrebbe dividere 1. Essendo

$$\frac{p_k}{q_k} = \frac{a}{b},$$

deve essere $a = rp_k$ e $b = rq_k$ e risulta

$$a\,q_{k-1} - b\,p_{k-1} = (-1)^{k-1} r.$$

Ne segue che q_{k-1} e p_{k-1} sono, a meno del segno, soluzione dell'equazione data.)

5.41 Si trasformino in frazione continua della forma (42) le seguenti funzioni razionali

$$(1) \quad \frac{3x^2 + 5x + 1}{x^2 + 2x + 1}\,, \qquad (2) \quad \frac{3x^2 + 6x + 1}{x^2 + 2x + 1}\,.$$

(Risposta:

$$(1) \quad 3 - \cfrac{1}{x + \cfrac{1}{x+2}}\,, \qquad (2) \quad 3 - \cfrac{2}{x^2 + 2x + 1}\,.$$

5.42 Si trasformi in frazione continua della forma (42) la funzione razionale

$$w(x) = \frac{x(x^3 + x^2 + x + 1) + 1}{x^3(x+2)}\,,$$

e si valuti il costo computazionale delle due espressioni.

(Risposta:

$$w(x) = 1 + \cfrac{-1}{x+3 + \cfrac{4}{x-2 + \cfrac{1/4}{x-1 + \cfrac{11/4}{x+2}}}}\,.$$

Per la funzione razionale sono richieste 5 operazioni additive e 5 moltiplicative, per la frazione continua sono richieste 8 operazioni additive e 4 moltiplicative.)

5.43 Si verifichi che

a) la differenza inversa della funzione $f(x) + g(x)$ in generale non è uguale alla somma delle differenze inverse di $f(x)$ e di $g(x)$;

b) moltiplicando $f(x)$ per uno scalare c, la differenza inversa risulta moltiplicata per c se è di ordine pari, divisa per c se è di ordine dispari;

c) aggiungendo ad $f(x)$ uno scalare, la differenza inversa non cambia.

5.44 Si dimostri che

a) la differenza inversa del primo ordine di un polinomio $p(x)$ di grado $m \geq 2$ è il reciproco di un polinomio di grado $m - 1$ e la differenza del

secondo ordine è data dal rapporto di un polinomio di grado $m-1$ per uno di grado $m-2$;

b) se $f(x)$ è una funzione razionale, allora esiste un intero $k > 0$ tale che la corrispondente differenza inversa $\phi[x_0, \ldots, x_{k-1}, x]$ di ordine k è costante;

c) si scriva la frazione continua di Thiele di

$$f(x) = x^2 \quad \text{e di} \quad f(x) = x^{-2} \quad \text{con} \quad x_i = i + 1.$$

(Traccia: a) si noti che $p(x) - p(x_0)$ è divisibile per $x - x_0$; b) sia $f(x) = \dfrac{p(x)}{q(x)}$, con $p(x)$ e $q(x)$ polinomi di grado rispettivamente m e n. Si verifichi che la funzione razionale $\phi[x_0, x]$ è data dal rapporto di due polinomi di grado n e $\max(n, m) - 1$. Procedendo in questo modo si verifichi che se $m > n$ dopo $2(m-1)$ passi la differenza inversa diventa un polinomio di grado 1, mentre se $m \leq n$ ciò accade dopo $2n - 1$ passi.

c) $$x^2 = 1 + \frac{x-1}{1/3} + \frac{x-2}{-12} + \frac{x-3}{-1/3};$$
$$x^{-2} = 1 + \frac{x-1}{-4/3} + \frac{x-2}{-12/11} + \frac{x-3}{154/3} + \frac{x-4}{1/11}. \ )$$

5.45 Sia $w(x)$ la frazione continua di Thiele (54) della funzione $f(x)$ e sia

$$w_i(x) = \frac{p_i(x)}{q_i(x)}, \quad i = 1, \ldots, m+n,$$

la i-esima frazione parziale di $w(x) = w_{m+n}(x)$. Si dimostri che, posto

$$c(x) = \frac{(x - x_{m+n})\, q_{m+n-1}(x)}{\phi[x_0, \ldots, x_{m+n}, x]\, q_{m+n}(x)},$$

risulta

$$f(x) - w(x) = -\frac{c(x)}{1 + c(x)} \left[w(x) - w_{m+n-1}(x)\right]$$
$$= \frac{(-1)^{m+n} \pi_{m+n}(x)}{\left\{\phi[x_0, \ldots, x_{m+n}, x]\, q_{m+n}(x) + (x - x_{m+n})\, q_{m+n-1}(x)\right\} q_{m+n}(x)}.$$

(Traccia: dalle (57) e (40) risulta che

$$f(x) = \frac{\phi[x_0, \ldots, x_{m+n}, x]\, p_{m+n}(x) + (x - x_{m+n})\, p_{m+n-1}(x)}{\phi[x_0, \ldots, x_{m+n}, x]\, q_{m+n}(x) + (x - x_{m+n})\, q_{m+n-1}(x)}.$$

Per la seconda relazione si tenga conto dell'esercizio 5.36 a).)

5.46 Si verifichi che

per k pari è $\rho[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] = \phi[x_0] + \phi[x_0, x_1, x_2]$
$+\phi[x_0, x_1, x_2, x_3, x_4] + \ldots + \phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x]$,

per k dispari è $\rho[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x] = \phi[x_0, x_1] + \phi[x_0, x_1, x_2, x_3]$
$+\phi[x_0, x_1, x_2, x_3, x_4, x_5] + \ldots + \phi[x_0, x_1, \ldots, x_{k-2}, x_{k-1}, x]$.

(Traccia: si proceda per induzione: per $k = 0$ e per $k = 1$ la relazione è immediata. Per $k \geq 2$ pari, per l'ipotesi induttiva è

$$\rho[x_0, \ldots, x_{k-2}] = \phi[x_0] + \phi[x_0, x_1, x_2] + \ldots + \phi[x_0, x_1, \ldots, x_{k-2}],$$

e si applichi il teorema 5.50. Si proceda in modo analogo per k dispari.)

5.47 Si scrivano le frazioni continue di Thiele di ordine n delle funzioni

$$(1) \quad f(x) = \sqrt{x + k^2}, \qquad (2) \quad f(x) = \frac{1}{\sqrt{x + k^2}}\,, \quad k > 0,$$

per $x_0 = 0$.

(Traccia: (1) è

$$\rho^{(i)}[x] = (i+1)(x+k^2)^{1/2} \quad \text{e per} \quad i \geq 1 \quad \phi^{(i)}[x] = 2(x+k^2)^{1/2},$$

quindi

$$w(x) = k + \underbrace{\frac{x}{2k} + \frac{x}{2k} + \frac{x}{2k} + \cdots + \frac{x}{2k}}_{n \text{ volte}};$$

(2) è

$$\rho^{(2s)}[x] = \frac{(x+k^2)^{-1/2}}{2s+1}\,,$$
$$\rho^{(2s+1)}[x] = -\frac{2}{3}\,(s+1)(2s+1)(2s+3)(x+k^2)^{3/2},$$
$$\phi^{(2s)}[x] = -\frac{2(x+k^2)^{-1/2}}{4s^2-1}\,,$$
$$\phi^{(2s+1)}[x] = -2(2s+1)^2(x+k^2)^{3/2}, \quad \text{per } s \geq 1,$$

quindi

$$w(x) = \frac{1}{k} - \frac{x}{2k^3} + \frac{x}{2/(3k)} + \frac{x}{18k^3} + \frac{x}{2/(15k)}$$
$$+ \cdots + \frac{x}{2/[(4s^2-1)k]} + \frac{x}{2(2s+1)^2k^3}\,, \quad n = 2s+1. \ )$$

5.48 Si scriva la frazione continua di Thiele di ordine 5 della funzione

$$f(x) = \arctan x, \quad \text{per} \quad x_0 = 1.$$

(Risposta:

$$w(x) = \frac{\pi}{4} + \frac{x-1}{2} + \frac{x-1}{1} + \frac{x-1}{-6} + \frac{x-1}{-1/4} + \frac{x-1}{40}\,.\ )$$

5.49 Si dimostri che

a) se la funzione

$$f(x) = a_0 + a_1 \cos x + a_2 \cos 2x + \ldots + a_n \cos nx$$

si annulla in $n+1$ punti $0 \le x_0 < x_1 < \ldots < x_n < \pi$, allora si annulla identicamente;

b) se la funzione

$$f(x) = b_1 \sin x + b_2 \sin 2x + \ldots + b_n \sin nx$$

si annulla in n punti $0 < x_1 < \ldots < x_n < \pi$, allora si annulla identicamente.

(Traccia: a) poiché

$$\cos jx = \frac{1}{2}\left(\mathrm{e}^{\boldsymbol{i}jx} + \mathrm{e}^{-\boldsymbol{i}jx}\right),$$

è

$$f(x) = \sum_{j=0}^{n} a_j \cos jx = a_0 + \sum_{j=1}^{n} \frac{a_j}{2}\ \mathrm{e}^{\boldsymbol{i}jx} + \sum_{j=1}^{n} \frac{a_j}{2}\ \mathrm{e}^{-\boldsymbol{i}jx} = \sum_{j=-n}^{n} b_j\, \mathrm{e}^{\boldsymbol{i}jx},$$

dove

$$b_j = \begin{cases} a_{-j}/2, & \text{per } -n \le j \le -1, \\ a_0 & \text{per } j = 0, \\ a_j/2, & \text{per } 1 \le j \le n. \end{cases}$$

Si verifichi che, posto $z = \mathrm{e}^{\boldsymbol{i}x}$, è $f(x) = z^{-n}p(z)$, in cui $p(z)$ è un polinomio di grado $2n$, e che se $f(x_k) = 0$, allora $p(z_k) = p(\overline{z}_k) = 0$, dove $z_k = \mathrm{e}^{\boldsymbol{i}x_k}$. Si proceda analogamente per b).)

5.50 Si determini il polinomio trigonometrico $F_3(x)$ di primo grado che assume i valori y_0, y_1, y_2 nei nodi x_0, x_1, x_2.

(Traccia: si verifichi che la funzione

$$F_3(x) = \frac{\sin\frac{1}{2}(x-x_1)\sin\frac{1}{2}(x-x_2)}{\sin\frac{1}{2}(x_0-x_1)\sin\frac{1}{2}(x_0-x_2)}\,y_0 + \frac{\sin\frac{1}{2}(x-x_0)\sin\frac{1}{2}(x-x_2)}{\sin\frac{1}{2}(x_1-x_0)\sin\frac{1}{2}(x_1-x_2)}\,y_1 + \frac{\sin\frac{1}{2}(x-x_0)\sin\frac{1}{2}(x-x_1)}{\sin\frac{1}{2}(x_2-x_0)\sin\frac{1}{2}(x_2-x_1)}\,y_2$$

assume i valori prescritti nei nodi e che può essere scritta nlla forma (77).)

5.51 In questo e negli esercizi seguenti che riguardano la trasformata discreta di Fourier, con la lettera $\mathbf{v}$, oltre che il vettore $\mathbf{v} = [v_0, \ldots, v_{n-1}]^T$, si indicherà anche la successione $\{v_j\}_{j=-\infty,\infty}$, ottenuta estendendo il vettore $\mathbf{v}$ per periodicità, cioè ponendo $v_j = v_k$, se $j \equiv k \pmod n$, $0 \le k < n$.

Si verifichino le seguenti proprietà della DFT.

a) linearità:

$$\mathrm{DFT}(\alpha\mathbf{y}' + \beta\mathbf{y}'') = \alpha\,\mathrm{DFT}(\mathbf{y}') + \beta\,\mathrm{DFT}(\mathbf{y}''), \quad \alpha, \beta \in \mathbf{C},\ \mathbf{y}', \mathbf{y}'' \in \mathbf{C}^n;$$

b) decomposizione: se $\mathbf{y} \in \mathbf{R}^n$, posto $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$ e

$$z_j^{(r)} = \mathrm{Re}(z_j), \quad z_j^{(i)} = \mathrm{Im}(z_j), \quad \text{per} \quad j = 0, \ldots, n-1,$$

e

$$y'_k = \frac{1}{2}\left[y_k + y_{n-k}\right], \quad y''_k = \frac{1}{2}\left[y_k - y_{n-k}\right], \quad k = 0, \ldots, n-1,$$

risulta

$$\mathbf{z}^{(r)} = \mathrm{DFT}(\mathbf{y}'), \quad \boldsymbol{i}\,\mathbf{z}^{(i)} = \mathrm{DFT}(\mathbf{y}'');$$

c) traslazione degli indici: se $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$ e $\mathbf{y}' \in \mathbf{C}^n$ è tale che $y'_k = y_{k+k_0}$, posto $\mathbf{z}' = \mathrm{DFT}(\mathbf{y}')$, risulta

$$z'_j = \omega_n^{jk_0} z_j, \quad j = 0, \ldots, n-1;$$

d) traslazione degli indici nella trasformata inversa: se $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$ e $\mathbf{z}' \in \mathbf{C}^n$ è tale che $z'_j = z_{j+j_0}$, posto $\mathbf{y}' = \mathrm{IDFT}(\mathbf{z}')$, risulta

$$y'_k = \omega_n^{-j_0 k} y_k, \quad k = 0, \ldots, n-1;$$

e) simmetria: se $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$ e $\mathbf{w} = \mathrm{DFT}(\mathbf{z})$, allora

$$w_j = \frac{1}{n}\,y_{-j};$$

f) trasformata di un vettore reale simmetrico rispetto a $\dfrac{n}{2}$: se $y_k = y_{n-k}$ per $k = 1, \ldots, \left\lfloor \dfrac{n}{2} \right\rfloor$ e $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$, allora anche $\mathbf{z}$ è reale e simmetrico rispetto a $\dfrac{n}{2}$.

(Traccia: per a), c), d) si applichino le (74) e (75); b) si noti che per la (76) è

$$\mathrm{Re}(z_j) = \frac{1}{2}\,(z_j + \overline{z}_j) = \frac{1}{2}\,(z_j + z_{n-j}),$$

e si sfrutti la (74); analogamente per la parte immaginaria; e) si ha

$$\begin{aligned} w_j &= \frac{1}{n}\sum_{k=0}^{n-1} \omega_n^{-jk} z_k = \frac{1}{n^2}\sum_{k=0}^{n-1} \omega_n^{-jk} \sum_{h=0}^{n-1} \omega_n^{-hk} y_h \\ &= \frac{1}{n^2}\sum_{h=0}^{n-1} \Big[\sum_{k=0}^{n-1} \omega_n^{-(j+h)k}\Big] y_h, \end{aligned}$$

e si applichi il teorema 5.57; f) segue da b) che $y''_k = 0$ per $k = 0, \ldots, n-1$, e quindi $\mathbf{z}^{(i)} = \mathbf{0}$, e la simmetria segue dalla (76).)

5.52 Siano $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$. Si verifichi che, posto $\mathbf{w} = \mathbf{u} + \boldsymbol{i}\mathbf{v}$, $\mathbf{x} = \mathrm{DFT}(\mathbf{u})$, $\mathbf{y} = \mathrm{DFT}(\mathbf{v})$, $\mathbf{z} = \mathrm{DFT}(\mathbf{w})$, risulta

$$\begin{aligned} &x_0 = \mathrm{Re}(z_0), \quad y_0 = \mathrm{Im}(z_0), \\ &x_k = \frac{1}{2}\,(z_k + \overline{z}_{n-k}), \quad y_k = -\frac{\boldsymbol{i}}{2}\,(z_k - \overline{z}_{n-k}), \quad k = 1, \ldots, n-1. \end{aligned}$$

(Traccia: essendo $\mathbf{u}$ e $\mathbf{v}$ ad elementi reali, è

$$x_k = \overline{x}_{n-k}, \quad y_k = \overline{y}_{n-k}, \quad k = 1, \ldots, n-1.$$

Per la linearità è $z_k = x_k + \boldsymbol{i}y_k$, per cui

$$z_k + \overline{z}_{n-k} = x_k + \boldsymbol{i}y_k + \overline{x}_{n-k} - \boldsymbol{i}\overline{y}_{n-k} = 2x_k;$$

analogamente per y_k.)

5.53 Sia $\mathbf{v} \in \mathbf{R}^n$, n pari. Si consideri il vettore $\mathbf{w} \in \mathbf{C}^{n/2}$ definito da

$$w_k = v_{2k} + \boldsymbol{i}v_{2k+1}, \quad k = 0, \ldots, \frac{n}{2} - 1.$$

Posto $\mathbf{z} = \mathrm{DFT}(\mathbf{w})$ e $\mathbf{u} = \mathrm{DFT}(\mathbf{v})$, si dimostri che vale

$$\left.\begin{aligned} u_0 &= \frac{1}{2}\left[\mathrm{Re}(z_0) + \mathrm{Im}(z_0)\right], \qquad u_{n/2} = \frac{1}{2}\left[\mathrm{Re}(z_0) - \mathrm{Im}(z_0)\right], \\ u_k &= \frac{1}{4}\left[(z_k + \overline{z}_{n/2-k}) - \boldsymbol{i}\omega_n^{-k}(z_k - \overline{z}_{n/2-k})\right] \\ u_{n/2+k} &= \overline{u}_{n/2-k} \end{aligned}\right\}, \quad k = 1, \ldots, \frac{n}{2} - 1.$$

(Traccia: scrivendo le analoghe della (82) e della (83) per la $\mathrm{DFT}(\mathbf{v})$, si ha

$$\left.\begin{aligned} u_k &= \frac{1}{2}\left(u'_k + \omega_n^{-k} u''_k\right) \\ u_{n/2+k} &= \frac{1}{2}\left(u'_k - \omega_n^{-k} u''_k\right) \end{aligned}\right\}, \quad k = 0, \ldots, \frac{n}{2} - 1,$$

in cui $\mathbf{u}' = \mathrm{DFT}(\mathbf{v}')$, $\mathbf{u}'' = \mathrm{DFT}(\mathbf{v}'')$ e $v'_k = v_{2k}$, $v''_k = v_{2k+1}$, per $k = 0, \ldots, \dfrac{n}{2} - 1$. Si tenga conto che per l'esercizio 5.52 si ha

$$u'_k = \frac{1}{2}\left(z_k + \overline{z}_{n/2-k}\right) \quad \text{e} \quad u''_k = -\frac{\boldsymbol{i}}{2}\left(z_k - \overline{z}_{n/2-k}\right), \quad \text{per} \quad k = 1, \ldots, \frac{n}{2} - 1.\ )$$

5.54 Sia $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$ e sia $V \in \mathbf{C}^{n\times n}$ la matrice della IDFT tale che $V\mathbf{z} = \mathbf{y}$. Si verifichi che

a) $V = V^T$;

b) $V^H V = V V^H = nI$;

c) $V^2 = n\Pi$, dove $\Pi \in \mathbf{R}^{n\times n}$ è la matrice di permutazione

$$\Pi = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ & & & 1 \\ & & \iddots & \\ 0 & 1 & & \end{bmatrix};$$

d) $$\sum_{j=0}^{n-1} |z_j|^2 = \frac{1}{n} \sum_{k=0}^{n-1} |y_k|^2.$$

e) Si consideri la matrice di permutazione $\Pi' \in \mathbf{R}^{n\times n}$

$$\Pi' = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ 1 & & & & 0 \end{bmatrix}.$$

Si verifichi che

$$V^H \Pi' V = nD,$$

dove D è la matrice diagonale

$$D = \begin{bmatrix} \omega_n^0 & & & \\ & \omega_n^1 & & \\ & & \ddots & \\ & & & \omega_n^{n-1} \end{bmatrix}.$$

f) Si scriva la matrice V per $n = 8$.

(Traccia: b) gli elementi di V sono $v_{kj} = \omega_n^{kj}$, $k, j = 0, \ldots, n-1$. Quindi l'elemento (j,k) di $V^H V$ è

$$\sum_{s=0}^{n-1} \overline{v}_{sj} v_{sk} = \sum_{s=0}^{n-1} \omega_n^{-sj} \omega_n^{sk} = \sum_{s=0}^{n-1} \omega_n^{s(k-j)},$$

e si applichi il teorema 5.57; c) si proceda come per b);

d) $$\sum_{j=0}^{n-1} |z_j|^2 = \mathbf{z}^H \mathbf{z} = \frac{1}{n^2} \mathbf{y}^H V V^H \mathbf{y} = \frac{1}{n} \mathbf{y}^H \mathbf{y};$$

e) l'elemento (j,k) di $V^H \Pi' V$, è

$$\sum_{s=0}^{n-1} \omega_n^{-js} \omega_n^{(s+1)k} = \omega_n^k \sum_{s=0}^{n-1} \omega_n^{s(k-j)},$$

e si applichi il teorema 5.57; f) posto

$$S = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 2 & 4 & 6 & 0 & 2 & 4 & 6 \\ 0 & 3 & 6 & 1 & 4 & 7 & 2 & 5 \\ 0 & 4 & 0 & 4 & 0 & 4 & 0 & 4 \\ 0 & 5 & 2 & 7 & 4 & 1 & 6 & 3 \\ 0 & 6 & 4 & 2 & 0 & 6 & 4 & 2 \\ 0 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix},$$

gli elementi di V sono dati da

$$v_{jk} = \omega_8^{s_{jk}}, \quad j, k = 0, \ldots, 7.$$

Quindi

$$V = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \rho(1+\boldsymbol{i}) & \boldsymbol{i} & \rho(-1+\boldsymbol{i}) & -1 & \rho(-1-\boldsymbol{i}) & -\boldsymbol{i} & \rho(1-\boldsymbol{i}) \\ 1 & \boldsymbol{i} & -1 & -\boldsymbol{i} & 1 & \boldsymbol{i} & -1 & -\boldsymbol{i} \\ 1 & \rho(-1+\boldsymbol{i}) & -\boldsymbol{i} & \rho(1+\boldsymbol{i}) & -1 & \rho(1-\boldsymbol{i}) & \boldsymbol{i} & \rho(-1-\boldsymbol{i}) \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & \rho(-1-\boldsymbol{i}) & \boldsymbol{i} & \rho(1-\boldsymbol{i}) & -1 & \rho(1+\boldsymbol{i}) & -\boldsymbol{i} & \rho(-1+\boldsymbol{i}) \\ 1 & -\boldsymbol{i} & -1 & \boldsymbol{i} & 1 & -\boldsymbol{i} & -1 & \boldsymbol{i} \\ 1 & \rho(1-\boldsymbol{i}) & -\boldsymbol{i} & \rho(-1-\boldsymbol{i}) & -1 & \rho(-1+\boldsymbol{i}) & \boldsymbol{i} & \rho(1+\boldsymbol{i}) \end{bmatrix},$$

dove $\rho = \dfrac{\sqrt{2}}{2}$.)

5.55 Si calcoli la DFT del vettore

$$\mathbf{y} = \left[0,\, \frac{1}{2}\,, 1,\, \frac{1}{2}\,, 0, -\frac{1}{2}\,, -1, -\frac{1}{2}\right]^T,$$

usando la matrice V dell'esercizio 5.54 oppure la FFT. Si scriva poi il polinomio trigonometrico di interpolazione $F_8(x)$ tale che $F_8(x_k) = y_k$, $x_k = \pi k/4$, $k = 0, \ldots, 7$.

(Risposta: $\mathbf{z} = \dfrac{\boldsymbol{i}}{4}\,[0, -1-\rho, 0, 1-\rho, 0, -1+\rho, 0, 1+\rho]^T$, dove $\rho = \dfrac{\sqrt{2}}{2}$. Risulta $\beta_1 = \dfrac{1}{2}\,(1+\rho)$, $\beta_3 = \dfrac{1}{2}\,(-1+\rho)$, gli altri coefficienti sono nulli.)

5.56 Si determinino i polinomi trigonometrici di interpolazione delle funzioni

$$\text{a)} \quad f(x) = 1 + \cos^2 x, \qquad \text{b)} \quad f(x) = 1 + \cos^6 x,$$

nei punti $x_k = \pi k/4$, $k = 0, \ldots, 7$.

(Risposta: a) risulta

$$\mathbf{y} = \left[2,\, \frac{3}{2}\,, 1,\, \frac{3}{2}\,, 2,\, \frac{3}{2}\,, 1,\, \frac{3}{2}\right]^T$$

e

$$\mathbf{z} = \mathrm{DFT}(\mathbf{y}) = \frac{1}{4}\,[6, 0, 1, 0, 0, 0, 1, 0]^T, \quad \alpha_0 = 3, \quad \alpha_2 = \frac{1}{2}\,,$$

gli altri coefficienti sono nulli. Si ha quindi

$$F_8(x) = \frac{3}{2} + \frac{1}{2}\cos 2x.$$

b) Si proceda in modo analogo. Risulta

$$F_8(x) = \frac{21}{16} + \frac{1}{2}\cos 2x + \frac{3}{16}\cos 4x. \ )$$

5.57 Si scrivano i polinomi trigonometrici di interpolazione delle funzioni

$$(1) \quad f(x) = \begin{cases} \pi & \text{per } 0 \le x < \pi, \\ -\pi & \text{per } \pi \le x < 2\pi, \end{cases}$$

$$(2) \quad f(x) = \begin{cases} \pi & \text{per } 0 \le x < \dfrac{\pi}{2} \text{ e } \dfrac{3\pi}{2} < x < 2\pi, \\ -\pi & \text{per } \dfrac{\pi}{2} \le x \le \dfrac{3\pi}{2}, \end{cases}$$

per $n = 3$ e per $n = 8$.

(Risposta: (1) per $n = 3$ si ha $\alpha_0 = \alpha_1 = \dfrac{2\pi}{3}$, $\beta_1 = \dfrac{2\sqrt{3}\pi}{3}$; per $n = 8$ si ha $\alpha_0 = \alpha_2 = \alpha_4 = \beta_2 = 0$, $\alpha_1 = \alpha_3 = \dfrac{\pi}{2}$, $\beta_1 = \dfrac{\pi}{2}(1+\sqrt{2})$, $\beta_3 = \dfrac{\pi}{2}(-1+\sqrt{2})$; (2) per $n = 3$ si ha $\alpha_0 = -\dfrac{2\pi}{3}$, $\alpha_1 = \dfrac{4\pi}{3}$, $\beta_1 = 0$; per $n = 8$ si ha $\alpha_0 = \alpha_4 = -\dfrac{\pi}{2}$, $\alpha_1 = \dfrac{\pi}{2}(1+\sqrt{2})$, $\alpha_2 = \dfrac{\pi}{2}$, $\alpha_3 = \dfrac{\pi}{2}(1-\sqrt{2})$, $\beta_1 = \beta_2 = \beta_3 = 0$.)

5.58 Siano V_n e $V_{n/2}$ le matrici della IDFT di ordine n e $n/2$, con n pari. Si dimostri che

a) $$V_n = \begin{bmatrix} I_{n/2} & I_{n/2} \\ I_{n/2} & -I_{n/2} \end{bmatrix} \begin{bmatrix} I_{n/2} & \\ & D \end{bmatrix} \begin{bmatrix} V_{n/2} & \\ & V_{n/2} \end{bmatrix} \Pi,$$

b) $$V_n = \Pi^T \begin{bmatrix} V_{n/2} & \\ & V_{n/2} \end{bmatrix} \begin{bmatrix} I_{n/2} & \\ & D \end{bmatrix} \begin{bmatrix} I_{n/2} & I_{n/2} \\ I_{n/2} & -I_{n/2} \end{bmatrix},$$

dove $I_{n/2}$ è la matrice identica di ordine $n/2$, D è la matrice diagonale il cui k-esimo elemento principale è ω_n^k, $k = 0, \ldots, \dfrac{n}{2} - 1$, e $\Pi \in \mathbf{R}^{n\times n}$ è la matrice di permutazione che trasforma il vettore $[0, 1, 2, 3, \ldots, n-1]^T$ nel vettore $[0, 2, 4, \ldots, n-2, 1, 3, 5, \ldots, n-1]^T$. Mentre l'algoritmo di Cooley e Tukey per la IDFT si basa sulla fattorizzazione a) della matrice V_n, l'algoritmo di Sande e Tukey si basa sulla fattorizzazione b).

(Traccia: a) si verifichi prima che la matrice $V_n \Pi^T$ è della forma

$$V_n \Pi^T = \begin{bmatrix} A & B \\ A & C \end{bmatrix}, \quad A, B, C \in \mathbf{R}^{n/2 \times n/2},$$

dove

$$a_{kj} = \omega_n^{2kj} = \omega_{n/2}^{kj}, \quad b_{kj} = \omega_n^{k(2j+1)} = \omega_n^k \omega_n^{2kj} = \omega_n^k \omega_{n/2}^{kj},$$
$$c_{kj} = \omega_n^{(n/2+k)(2j+1)} = -\omega_n^k \omega_n^{2kj} = -\omega_n^k \omega_{n/2}^{kj},$$

e quindi $A = V_{n/2}$, $B = DA$, $C = -DA$; b) si tenga conto del fatto che $V = V^T$.)

5.59 Sia $n = 3m$, m intero, e siano V_n e V_m le matrici della IDFT di ordine n e m. Si dimostri che

$$V_n = \begin{bmatrix} I_m & I_m & I_m \\ I_m & \omega_3 I_m & \omega_3^2 I_m \\ I_m & \omega_3^2 I_m & \omega_3 I_m \end{bmatrix} \begin{bmatrix} I_m & & \\ & D & \\ & & D^2 \end{bmatrix} \begin{bmatrix} V_m & & \\ & V_m & \\ & & V_m \end{bmatrix} \Pi,$$

dove I_m è la matrice identica di ordine m, D è la matrice diagonale il cui k-esimo elemento principale è ω_n^k, $k = 0, \ldots, m-1$, e $\Pi \in \mathbf{R}^{n \times n}$ è la matrice di permutazione che trasforma il vettore $[0, 1, 2, 3, \ldots, n-1]^T$ nel vettore $[0, 3, \ldots, n-3, 1, 4, \ldots, n-2, 2, 5, \ldots, n-1]^T$.

(Traccia: si verifichi prima che la matrice $V_n \Pi^T$ è della forma

$$V_n \Pi^T = \begin{bmatrix} A & B & C \\ A & \omega_3 B & \omega_3^2 C \\ A & \omega_3^2 B & \omega_3 C \end{bmatrix}, \quad A, B, C \in \mathbf{R}^{m \times m},$$

dove

$$a_{kj} = \omega_n^{3kj} = \omega_m^{kj}, \quad b_{kj} = \omega_n^{k(3j+1)} = \omega_n^k \omega_n^{3kj} = \omega_n^k \omega_m^{kj},$$
$$c_{kj} = \omega_n^{k(3j+2)} = \omega_n^{2k} \omega_n^{3kj} = \omega_n^{2k} \omega_m^{kj}. \ )$$

5.60 Siano $\mathbf{u}$ e $\mathbf{v}$ due successioni periodiche di periodo n. Si definisce *convoluzione circolare discreta* di $\mathbf{u}$ e $\mathbf{v}$ la successione $\mathbf{y} = \mathbf{u} * \mathbf{v}$ data da

$$y_j = \frac{1}{n} \sum_{k=0}^{n-1} u_k v_{j-k}, \quad j = \ldots, -1, 0, 1, \ldots$$

Si verifichi che
a) $\mathbf{y}$ è periodica di periodo n;
b) se $\mathbf{z} = \mathrm{DFT}(\mathbf{y})$, $\mathbf{x} = \mathrm{DFT}(\mathbf{u})$, $\mathbf{w} = \mathrm{DFT}(\mathbf{v})$, risulta

$$z_h = x_h w_h,$$

e quindi

$$\mathbf{y} = \mathrm{IDFT}\,(\mathrm{DFT}(\mathbf{u}) \ \cdot \ \mathrm{DFT}(\mathbf{v})),$$

in cui con il segno $\cdot$ si indica il prodotto componente per componente.

(Traccia: b) risulta

$$z_h = \frac{1}{n^2}\sum_{j=0}^{n-1}\sum_{k=0}^{n-1}\omega_n^{-jh}u_k v_{j-k} = \frac{1}{n^2}\sum_{k=0}^{n-1}\alpha(k,h)u_k,$$

dove

$$\begin{aligned}\alpha(k,h) &= \sum_{j=0}^{n-1}\omega_n^{-jh}v_{j-k} = \sum_{j=0}^{k-1}\omega_n^{-jh}v_{j-k} + \sum_{j=k}^{n-1}\omega_n^{-jh}v_{j-k}\\ &= \sum_{j=0}^{k-1}\omega_n^{-jh}v_{n+j-k} + \sum_{j=k}^{n-1}\omega_n^{-jh}v_{j-k}\\ &= \sum_{r=n-k}^{n-1}\omega_n^{-(r+k-n)h}v_r + \sum_{r=0}^{n-k-1}\omega_n^{-(r+k)h}v_r = \sum_{r=0}^{n-1}\omega_n^{-(r+k)h}v_r.\ )\end{aligned}$$

5.61 a) Si verifichi la validità e si determini il costo computazionale del seguente algoritmo per il calcolo in un punto x del polinomio trigonometrico di soli coseni (79), una volta che siano noti i coefficienti α_j, $j = 0, \ldots, n+1$:

$$\begin{aligned}&c_0 = 1, \quad c_1 = \cos x, \quad f_1 = \frac{\alpha_0}{2} + \alpha_1 c_1,\\ &\left.\begin{aligned}c_j &= 2c_1c_{j-1} - c_{j-2},\\ f_j &= f_{j-1} + \alpha_j c_j,\end{aligned}\right\}, \quad \text{per } j = 2, 3. \ldots, n,\\ &c_{n+1} = 2c_1c_n - c_{n-1}, \quad F_{2n+2}(x) = f_n + \frac{\alpha_{n+1}}{2}\,c_{n+1}.\end{aligned}$$

b) Si dica come si possono calcolare i coefficienti (80) del polinomio trigonometrico (79) facendo uso della trasformata discreta di Fourier e quale risulta il costo computazionale.

(Traccia: a) si verifichi che

$$\cos jx = 2\cos x\cos(j-1)x - \cos(j-2)x;$$

il costo computazionale è, a meno di costanti additive, di $2n$ addizioni e $2n$ moltiplicazioni. b) Si consideri il vettore di $2(n+1)$ componenti

$$\mathbf{v} = \big[y_0,\, y_1, \ldots, y_n, y_{n+1}, y_n, \ldots, y_1\big]^T.$$

Per quanto visto nell'esercizio 5.51f), il vettore $\mathbf{u} = \mathrm{DFT}(\mathbf{v})$ di $2(n+1)$ componenti è reale, quindi $\alpha_j = 2u_j$. È possibile calcolare α_j senza raddoppiare la dimensione, procedendo in modo simile a quanto fatto nell'esercizio 5.53: si considerino i vettori $\mathbf{v}'$, $\mathbf{v}'' \in \mathbf{R}^{n+1}$ e $\mathbf{w} \in \mathbf{C}^{n+1}$, così definiti

$$v'_k = v_{2k}, \quad v''_k = (v_{2k-1} + v_{2k+1}), \quad w_k = v'_k + \boldsymbol{i} v''_k, \quad k = 0, \ldots, n$$

(in cui si è posto $v_{-1} = y_1$). Il vettore $\mathbf{w}$ qui costruito sfrutta, a differenza di quello dell'esercizio 5.53, la simmetria di $\mathbf{v}$. I vettori $\mathbf{v}'$ e $\mathbf{v}''$ sono simmetrici rispetto a $n/2$, quindi le trasformate $\mathbf{u}' = \mathrm{DFT}(\mathbf{v}')$ e $\mathbf{u}'' = \mathrm{DFT}(\mathbf{v}'')$ sono reali e simmetriche (si veda l'esercizio 5.51 f)) e per la trasformata $\mathbf{z} = \mathrm{DFT}(\mathbf{w})$ risulta

$$\mathrm{Re}(\mathbf{z}) = \mathbf{u}' \quad \text{e} \quad \mathrm{Im}(\mathbf{z}) = \mathbf{u}''.$$

Per $k = 0, \ldots, n, \;\; k \neq \dfrac{n+1}{2}$, risulta (si vedano gli esercizi 5.51 c) e 5.53)

$$u_k = \frac{1}{2}\left(u'_k + \frac{\omega_{2n+2}^{-k}}{1+\omega_{n+1}^{-k}}\, u''_k\right) = \frac{1}{2}\left(\mathrm{Re}(z_k) + \frac{1}{2}\,\sec\frac{k\pi}{n+1}\,\mathrm{Im}(z_k)\right),$$

ed inoltre è

$$u_{n+1} = \frac{1}{2}\left(\mathrm{Re}(z_0) - \frac{1}{2}\,\mathrm{Im}(z_0)\right), \quad u_{(n+1)/2} = \frac{1}{2}\,\mathrm{Re}(z_{(n+1)/2}), \text{ se } n \text{ è dispari.}$$

Se $n+1$ è potenza di 2 e si utilizza un algoritmo FFT per il calcolo del vettore $\mathbf{z}$, il costo computazionale risulta dell'ordine di $n \log_2 n$.)

5.62 Si verifichi che con una IDFT di ordine $2n$ è possibile calcolare le *trasformate di seni e di coseni*

$$z_j = \sum_{k=1}^{n-1} \sin\frac{jk\pi}{n}\, y_k, \quad v_j = \sum_{k=0}^{n-1} \cos\frac{jk\pi}{n}\, y_k, \quad j = 0, \ldots, n-1.$$

(Traccia: si consideri il vettore $\mathbf{w} = [y_0, y_1, \ldots, y_{n-1}, 0, \ldots, 0]^T$ di $2n$ componenti e si calcoli $\mathbf{u} = \mathrm{IDFT}(\mathbf{w})$. Risulta dalla (75)

$$v_j = \mathrm{Re}(u_j) \quad \text{e} \quad z_j = \mathrm{Im}(u_j) \quad \text{per} \quad j = 0, \ldots, n-1.$$

Lo stesso risultato può essere ottenuto senza raddoppiare la dimensione, si veda [19].)

5.63 Sia $f(x)$ una funzione periodica di periodo 2π e derivabile due volte con continuità. Si verifichi che i coefficienti α_j e β_j del polinomio trigonometrico di interpolazione $F_n(x)$ di $f(x)$ sui nodi $x_k = \dfrac{2k\pi}{n}$, $k = 0, \dots, n-1$, verificano la relazione

$$|\alpha_j|,\ |\beta_j| \le \frac{\pi^2 M_2}{2j^2}, \quad \text{dove} \quad M_2 = \max_{x\in[0,2\pi]} |f''(x)|.$$

(Traccia: posto $h = \dfrac{2\pi}{n}$, si verifichi che

$$G_n(x) = \frac{1}{h^2}\big[F_n(x-h) - 2F_n(x) + F_n(x+h)\big] \tag{103}$$

è il polinomio trigonometrico di interpolazione della funzione

$$g(x) = \frac{1}{h^2}\big[f(x-h) - 2f(x) + f(x+h)\big],$$

sui nodi x_k, $k = 0, \dots, n-1$. Si supponga per semplicità che n sia dispari, $n = 2m-1$, sostituendo la (77) nella (103) risulta

$$G_n(x) = \frac{1}{h^2}\sum_{j=1}^{m-1}\big\{\alpha_j\big[\cos j(x-h) - 2\cos jx + \cos j(x+h)\big] + \beta_j\big[\sin j(x-h) - 2\sin jx + \sin j(x+h)\big]\big\},$$

e poiché

$$\begin{aligned}\cos(\theta-\varphi) - 2\cos\theta + \cos(\theta+\varphi) &= -4\sin^2\frac{\varphi}{2}\cos\theta,\\ \sin(\theta-\varphi) - 2\sin\theta + \sin(\theta+\varphi) &= -4\sin^2\frac{\varphi}{2}\sin\theta,\end{aligned} \qquad \text{per ogni } \theta, \varphi,$$

ne segue che

$$G_n(x) = \frac{\widehat{\alpha}_0}{2} + \sum_{j=1}^{m-1}(\widehat{\alpha}_j\cos jx + \widehat{\beta}_j\sin jx),$$

dove $\widehat{\alpha}_0 = 0$ e

$$\widehat{\alpha}_j = -\left(\frac{2}{jh}\sin\frac{jh}{2}\right)^2 j^2\alpha_j, \quad \text{per} \quad 0 < j \le m-1,$$

e analogamente per $\widehat{\beta}_j$. Poiché $\sin\theta \geq \dfrac{2\theta}{\pi}$ per $0 \leq \theta \leq \dfrac{\pi}{2}$, risulta che

$$|\widehat{\alpha}_j| \geq \frac{4}{\pi^2}\, j^2 |\alpha_j|, \quad \text{per} \quad 1 \leq j \leq m-1.$$

In modo analogo si dimostri che

$$|\widehat{\beta}_j| \geq \frac{4}{\pi^2}\, j^2 |\beta_j|, \quad \text{per} \quad 1 \leq j \leq m-1.$$

D'altra parte per la (78) è

$$|\widehat{\alpha}_j| = \frac{2}{n} \left| \sum_{k=0}^{n-1} g(x_k) \cos j x_k \right| \leq 2 \max_{x\in[0,2\pi]} |g(x)|,$$

ed essendo per la formula di Taylor

$$g(x) = \frac{1}{2}\left[f''(\xi_1) + f''(\xi_2)\right], \quad \xi_1, \xi_2 \in (x-h, x+h),$$

ne segue che $|g(x)| \leq M_2$.)

5.64 Sia $f(x)$ una funzione periodica di periodo 2π, derivabile due volte con continuità e siano n ed r due interi positivi, con $r+2 \leq \sqrt{n}$. Si considerino

(1) la somma parziale r-esima della serie di Fourier di $f(x)$

$$\sigma_r(x) = \frac{a_0}{2} + \sum_{j=1}^{r} (a_j \cos jx + b_j \sin jx),$$

dove

$$a_j = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos jx \; dx, \qquad b_j = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin jx \; dx;$$

(2) la somma parziale r-esima del polinomio trigonometrico di interpolazione $F_n(x)$ di $f(x)$ sui nodi $x_k = 2k\pi/n$, $k = 0, \ldots, n-1$

$$F_{n,r}(x) = \frac{\alpha_0}{2} + \sum_{j=1}^{r} (\alpha_j \cos jx + \beta_j \sin jx),$$

dove

$$\alpha_j = \frac{2}{n} \sum_{k=0}^{n-1} f(x_k) \cos j x_k, \qquad \beta_j = \frac{2}{n} \sum_{k=0}^{n-1} f(x_k) \sin j x_k.$$

Si verifichi che

a) α_j e β_j sono le approssimazioni di a_j e b_j che si otterrebbero utilizzando sui nodi x_k, $k = 0, \ldots, n$, la formula dei trapezi per il calcolo degli integrali (si veda il capitolo 7);

b) esiste una costante γ per cui

$$|a_j - \alpha_j| \le \frac{\gamma(j+1)^2}{n^2} \quad \text{e} \quad |b_j - \beta_j| \le \frac{\gamma(j+1)^2}{n^2} \quad \text{per} \quad j = 0, \ldots, r;$$

c) $|\sigma_r(x) - F_{n,r}(x)| < \dfrac{\gamma}{\sqrt{n}}$ per ogni x.

d) Si dimostri che per ogni $\epsilon > 0$ esiste un n tale che

$$|f(x) - F_n(x)| < \epsilon \quad \text{per ogni} \quad x.$$

(Traccia: a) applicando la formula dei trapezi al calcolo di a_j si ottiene l'approssimazione

$$\frac{1}{n}\left[f(x_0)\cos jx_0 + 2\sum_{k=1}^{n-1} f(x_k)\cos jx_k + f(x_n)\cos jx_n\right].$$

Si tenga conto del fatto che per la periodicità è $f(x_0) = f(x_n)$ e che $\cos jx_0 = \cos jx_n$. Si proceda in modo analogo per b_j.

b) Dalla (35, cap. 7) risulta

$$|a_j - \alpha_j| = \frac{1}{12\pi}\left(\frac{2\pi}{n}\right)^3 n|h_j''(\xi_j)|, \quad \xi_j \in (x_0, x_n),$$

dove $h_j(x) = f(x)\cos jx$. Risulta

$$|h_j''(x)| \le \max_{x\in[0,2\pi]} \left(|f''(x)| + 2j|f'(x)| + j^2|f(x)|\right) \le \gamma'(j+1)^2$$

per un γ' opportuno. Si ponga $\gamma = \dfrac{2}{3}\,\pi^2\gamma'$.

c) È

$$\begin{aligned}|\sigma_r - F_{n,r}(x)| &\le \frac{|a_0 - \alpha_0|}{2} + \sum_{j=1}^{r}(|a_j - \alpha_j|\,|\cos jx| + |b_j - \beta_j|\,|\sin jx|) \\ &\le \frac{2\gamma}{n^2}\sum_{j=0}^{r}(j+1)^2 < \frac{2\gamma(r+2)^3}{3n^2} < \frac{\gamma}{\sqrt{n}}\,.\end{aligned}$$

d) Siano ν un intero per cui

$$|f(x) - \sigma_\nu(x)| < \frac{\epsilon}{3}$$

(sotto le ipotesi fatte la serie di Fourier della $f(x)$ converge assolutamente e uniformemente), e tale che

$$\sum_{j=\nu+1}^{\infty} \frac{1}{j^2} < \frac{\epsilon}{3\pi^2 M_2}, \quad \text{dove} \quad M_2 = \max_{x\in[0,2\pi]} |f''(x)|,$$

e $n \geq (\nu+2)^2$ un intero per cui

$$\frac{\gamma}{\sqrt{n}} < \frac{\epsilon}{3}\,.$$

Allora risulta

$$|f(x) - F_n(x)| \leq |f(x) - \sigma_\nu(x)| + |\sigma_\nu(x) - F_{n,\nu}(x)| + |F_{n,\nu}(x) - F_n(x)|.$$

Poiché (supponendo per semplicità n dispari, $n = 2m-1$) è

$$\begin{aligned} |F_{n,\nu}(x) - F_n(x)| &= \Big| \sum_{j=\nu+1}^{m-1} (\alpha_j \cos jx + \beta_j \sin jx) \Big| \\ &\leq \sum_{j=\nu+1}^{\infty} (|\alpha_j| + |\beta_j|) \leq \pi^2 M_2 \sum_{j=\nu+1}^{\infty} \frac{1}{j^2} < \frac{\epsilon}{3} \end{aligned}$$

(si veda l'esercizio precedente), risulta

$$|f(x) - F_n(x)| < \epsilon.\ )$$

5.65 Si consideri un insieme $\mathbf{Z}_p$ (anello degli interi modulo p) e si supponga che valgano le seguenti proprietà:

(1) esiste $n \in \mathbf{Z}_p$ *invertibile*, cioè tale che esiste il numero n^{-1}, detto *reciproco* di n, per cui $n^{-1}n \equiv 1 \pmod p$;

(2) esiste $q \in \mathbf{Z}_p$ tale che $q^n \equiv 1 \pmod p$ e $(q^r - 1) \pmod p$ è invertibile per ogni r, $0 < r < n$ (poiché $q^r \not\equiv 1 \pmod p$ per $0 < r < n$, q è radice n-esima primitiva dell'unità).

a) Si verifichi che q^k è invertibile per ogni k, con $1 \leq k < n$.

Quindi si possono definire le matrici V e W di ordine n a elementi in $\mathbf{Z}_p$ nel modo seguente

$$v_{jk} \equiv q^{jk} (\text{mod } p), \quad w_{kj} \equiv q^{-jk} (\text{mod } p), \quad k, j = 0, \ldots, n-1.$$

b) Si dimostri che vale la relazione

$$WV \equiv nI \,(\mathrm{mod}\ p).$$

Perciò su $\mathbf{Z}_p$ è definibile una trasformata DFT di ordine n, in cui la matrice W svolge il ruolo che sui complessi è svolto da V^H: dato un vettore $\mathbf{y}$ di n elementi in $\mathbf{Z}_p$, si definisce $\mathbf{z} = \mathrm{DFT}(\mathbf{y}) = n^{-1}W\mathbf{y}$ il vettore di componenti

$$z_j \equiv n^{-1} \sum_{k=0}^{n-1} y_k q^{-jk} (\mathrm{mod}\ p), \quad j = 0, \ldots, n-1.$$

Si esaminino i seguenti casi particolari:

c) Se $p = 2^n - 1$, con n primo (p è detto numero di *Mersenne*), si verifichi che il numero

$$s = p - \frac{p-1}{n}$$

appartiene a $\mathbf{Z}_p$ ed è il reciproco di n e che la scelta $q = 2$ soddisfa la (2). Per il caso $n = 5$, e quindi $p = 31$, si scrivano le matrici V e W, e si costruisca la DFT del vettore $\mathbf{y} = [1, 2, 4, 4, 2]^T$.

d) Se $p = 2^{hn/2} + 1$, con n potenza di 2, cioè $n = 2^m$, e $h \geq 2$ intero, si verifichi che $n^{-1} \equiv 2^{nh-m} \pmod{p}$ e che la scelta $q = 2^h$ soddisfa la (2). Per il caso $n = 4$ e $h = 3$, e quindi $p = 65$, si scrivano V e W e si costruisca la DFT del vettore $\mathbf{y} = [7, 15, 15, 7]^T$.

e) Se $p = 17$ è possibile definire trasformate DFT per diversi n e per $n > 2$ con diverse radici primitive dell'unità. In particolare si verifichi che

per $n = 2$ è $n^{-1} = 9$ e si può scegliere $q = 16$;

per $n = 4$ è $n^{-1} = 13$ e si può scegliere $q = 4, 13$;

per $n = 8$ è $n^{-1} = 15$ e si può scegliere $q = 2, 8, 9, 15$;

per $n = 16$ è $n^{-1} = 16$ e si può scegliere $q = 3, 5, 6, 7, 10, 11, 12, 14$.

(Traccia: a) è $q^{-k} = q^{n-k}$, infatti $q^{-k}q^k = q^n \equiv 1 \pmod{p}$. b) Si verifichi che

$$\sum_{h=0}^{n-1} w_{kh} v_{hj} \equiv \begin{cases} n \pmod{p} & \text{se } k = j, \\ 0 \pmod{p} & \text{se } k \neq j, \end{cases} \qquad k, j = 0, \ldots, n-1.$$

Infatti per $k = j$ è

$$\sum_{h=0}^{n-1} w_{jh} v_{hj} \equiv \sum_{h=0}^{n-1} q^{-jh} q^{jh} \equiv \sum_{h=0}^{n-1} 1 \pmod{p} = n,$$

mentre per $k \neq j$ è

$$\sum_{h=0}^{n-1} w_{kh} v_{hj} \equiv \sum_{h=0}^{n-1} q^{h(j-k)} \equiv \left(q^{n(j-k)} - 1\right) \left(q^{j-k} - 1\right)^{-1} \equiv 0 \pmod{p},$$

perché $0 < |j-k| < n$ e $q^{n(j-k)} \equiv 1 \pmod{p}$.
c) Il numero s è intero e minore di p in quanto n è primo. Infatti per un teorema di Fermat è $a^n \equiv a \pmod{n}$ per ogni a e quindi $p - 1 = 2^n - 2 \equiv 0 \pmod{n}$. Si verifichi poi che $s\,n \equiv 1 \pmod{p}$. Risulta $2^n \equiv 1 \pmod{p}$, infatti $2^n = p + 1$; per verificare che $2^r - 1$ è invertibile in $\mathbf{Z}_p$, cioè che non ha fattori comuni con $2^n - 1$, per $0 < r < n$, si dimostri prima, facendo riferimento all'algoritmo di Euclide, che

$$\operatorname{mcd}(2^n - 1, 2^r - 1) = 2^d - 1, \quad \text{dove} \quad d = \operatorname{mcd}(n, r)$$

e si noti che $\operatorname{mcd}(n, r) = 1$. Per $n = 5$ è $n^{-1} = 25$,

$$V = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 \\ 1 & 4 & 16 & 2 & 8 \\ 1 & 8 & 2 & 16 & 4 \\ 1 & 16 & 8 & 4 & 2 \end{bmatrix}, \quad W = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 16 & 8 & 4 & 2 \\ 1 & 8 & 2 & 16 & 4 \\ 1 & 4 & 16 & 2 & 8 \\ 1 & 2 & 4 & 8 & 16 \end{bmatrix},$$

$$\mathrm{DFT}(\mathbf{y}) = [15, 17, 7, 7, 17]^T.$$

d) è $2^{nh} \equiv 1 \pmod{p}$, infatti $2^{nh} = p(p-2) + 1$; per verificare che $2^{hr} - 1$ è invertibile in $\mathbf{Z}_p$, cioè che non ha fattori comuni con $2^{hn/2} + 1$, per $0 < r < n$, si noti che per quanto visto alla traccia del punto c) è

$$\operatorname{mcd}(2^{hr} - 1, 2^{hn} - 1) = 2^d - 1, \quad \text{dove} \quad d = h \operatorname{mcd}(r, n)$$
$$\operatorname{mcd}(2^{hr} - 1, 2^{hn/2} - 1) = 2^e - 1, \quad \text{dove} \quad e = h \operatorname{mcd}(r, n/2),$$

e poiché

$$\operatorname{mcd}(2^{hr} - 1, 2^{hn} - 1) = \operatorname{mcd}(2^{hr} - 1, 2^{hn/2} - 1) \operatorname{mcd}(2^{hr} - 1, 2^{hn/2} + 1),$$

ne segue che

$$\operatorname{mcd}(2^{hr} - 1, 2^{hn/2} + 1) = \frac{2^d - 1}{2^e - 1} = 1,$$

in quanto $\operatorname{mcd}(r, n) = \operatorname{mcd}(r, n/2)$, perché n è potenza di 2 e $r < n$. Si verifichi poi che $n n^{-1} \equiv 1 \pmod{p}$. Nel caso particolare è $n^{-1} = 49$,

$$V = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 8 & 64 & 57 \\ 1 & 64 & 1 & 64 \\ 1 & 57 & 64 & 8 \end{bmatrix}, \quad W = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 57 & 64 & 8 \\ 1 & 64 & 1 & 64 \\ 1 & 8 & 64 & 57 \end{bmatrix},$$

$$\mathrm{DFT}(\mathbf{y}) = [11, 47, 0, 14]^T. \ )$$

5.66 Sia n potenza di 2, e sia $p(x)$ un polinomio di grado al più $n-1$. Si verifichi che

a) dati i coefficienti di $p(x)$, è possibile calcolare i valori che $p(x)$ assume nelle n radici n-esime dell'unità con $O(n \log_2 n)$ operazioni;

b) dati i valori che $p(x)$ assume nelle n radici n-esime dell'unità, è possibile calcolare i coefficienti di $p(x)$ con $O(n \log_2 n)$ operazioni.

(Traccia: indicata con V la matrice di Vandermonde di ordine n il cui elemento (k, j)-esimo è $v_{kj} = \omega_n^{kj}$, con $\mathbf{y}$ il vettore dei coefficienti del polinomio e con $\mathbf{z}$ il vettore dei valori che $p(x)$ assume nelle radici dell'unità, è $V\mathbf{y} = \mathbf{z}$. Quindi, quanto richiesto al punto a) viene realizzato con una IDFT di ordine n e quanto richiesto al punto b) con una DFT di ordine n.)

5.67 Sia n potenza di 2, e siano $p(x)$ e $q(x)$ due polinomi di grado n_1 e n_2, con $n_1 + n_2 < n$. Si verifichi che, dati i coefficienti di $p(x)$ e $q(x)$, è possibile calcolare i coefficienti del polinomio prodotto $s(x) = p(x)q(x)$ con $O(n \log_2 n)$ operazioni, procedendo con il seguente algoritmo:

a) si calcolano i valori

$$\alpha_i = p(\omega_n^i), \quad \beta_i = q(\omega_n^i), \quad i = 0, \ldots, n-1;$$

b) si calcolano i prodotti

$$\gamma_i = \alpha_i \beta_i, \quad i = 0, \ldots, n-1;$$

c) si calcolano i coefficienti del polinomio $s(x)$ tale che

$$s(\omega_n^i) = \gamma_i, \quad i = 0, \ldots, n-1.$$

(Traccia: per quanto visto nell'esercizio precedente, il punto a) richiede due IDFT di ordine minore o uguale ad n, il punto b) n prodotti, il punto c) una DFT di ordine n.)

5.68 Sono assegnati i coefficienti del polinomio

$$p(z) = \sum_{i=0}^{n} a_i z^i, \quad a_0, a_n \neq 0.$$

a) Si dimostri che il calcolo dei primi r coefficienti della serie

$$s(z) = \sum_{i=0}^{\infty} \sigma_i z^i \quad \text{tale che} \quad s(z)p(z) = 1,$$

può essere fatto con $O(r \log_2 r)$ operazioni aritmetiche.

b) Si utilizzi tale risultato per calcolare quoziente e resto della divisione di due polinomi $p(z)$ e $q(z)$.

(Traccia: a) per quanto visto nell'esercizio 3.69 b) la successione

$$x_{i+1}(z) = 2x_i(z) - x_i^2(z)p(z), \quad x_0(z) = \frac{1}{a_0},$$

consente di determinare i primi coefficienti di $s(z)$ con convergenza quadratica. Si verifichi che questa proprietà vale ancora quando all'i-esimo passo si operi con i soli 2^{i+1} coefficienti dei termini di grado più basso, cioè con la successione $x_i(z)$ ottenuta nel modo seguente

$$\begin{aligned} &p_i(z) \equiv p(z) \ \left(\mathrm{mod}\ z^{2^{i+1}}\right), \\ &y_{i+1}(z) = 2x_i(z) - x_i^2(z)p_i(z), \\ &x_{i+1}(z) \equiv y_{i+1}(z) \ \left(\mathrm{mod}\ z^{2^{i+1}}\right). \end{aligned}$$

Si effettuino i prodotti di polinomi al passo i-esimo con il metodo descritto nell'esercizio 5.67. Il costo totale risulta allora (si veda l'esercizio 4.16 g))

$$O\Big(\sum_{i=0}^{\log_2 r} i2^i\Big) = O(r \log_2 r).$$

b) Siano n ed $m \le n$ i gradi di $p(z)$ e $q(z)$, e sia $p(z) = q(z)t(z) + r(z)$, con $t(z)$ quoziente della divisione. Indicati con $\widetilde{p}(z)$, $\widetilde{q}(z)$ e $\widetilde{t}(z)$ i polinomi a radici reciproche di $p(z)$, $q(z)$ e $t(z)$, si verifichi che

$$\widetilde{t}(z) \equiv \widetilde{p}(z)\widetilde{s}(z) \ (\mathrm{mod}\ z^{n-m+1}), \quad \text{dove} \quad \widetilde{s}(z)\widetilde{q}(z) = 1. \ )$$

5.69 Sia $n = 2^r$, r intero, e siano $p(x)$ un polinomio di grado al più $n-1$ e $x_0, x_1, \ldots, x_{n-1}$, n punti distinti. Si verifichi che, dati i coefficienti di $p(x)$, è possibile calcolare i valori $p(x_0), p(x_1), \ldots, p(x_{n-1})$, con $O(n \log_2^2 n)$ operazioni, procedendo con il seguente algoritmo:

a) posto

$$s_i^{(0)}(x) = x - x_i, \quad \text{per} \quad i = 0, \ldots, n-1,$$

si calcolano i coefficienti dei polinomi

$$s_i^{(j)}(x) = s_{2i}^{(j-1)}(x)s_{2i+1}^{(j-1)}(x), \quad \text{per} \quad i = 0, \ldots, \frac{n}{2^j} - 1, \ j = 1, \ldots, r-1;$$

b) posto

$$t_0^{(0)}(x) = p(x),$$

si calcolano per $i = 0, \ldots, 2^{j-1} - 1$, $j = 1, \ldots, r$, i coefficienti del polinomio $t_{2i}^{(j)}(x)$, resto della divisione di $t_i^{(j-1)}(x)$ per $s_{2i}^{(r-j)}(x)$, e del polinomio $t_{2i+1}^{(j)}(x)$, resto della divisione di $t_i^{(j-1)}(x)$ per $s_{2i+1}^{(r-j)}(x)$.

Si verifichi che per $i = 0, \ldots, n-1$, i polinomi $t_i^{(r)}(x)$ hanno grado 0 e che vale $p(x_i) = t_i^{(r)}(x)$.

(Traccia: i polinomi $s_i^{(j)}(x)$ hanno grado 2^j, quindi per $j \geq 1$ i polinomi $t_{2i}^{(j)}(x)$ e $t_{2i+1}^{(j)}(x)$ hanno grado minore o uguale a $2^{r-j} - 1$. Per verificare che $p(x_i) = t_i^{(r)}(x)$, si usi induttivamente il fatto che $t_{2i}^{(j)}(x_k) = t_i^{(j-1)}(x_k)$, per gli indici k per cui $s_{2i}^{(r-j)}(x_k) = 0$, e analogamente per $t_{2i+1}^{(j)}(x)$. Esaminando il costo computazionale del passo a), per un valore j dell'indice si devono calcolare $\dfrac{n}{2^j}$ prodotti di polinomi di grado 2^{j-1}, producendo polinomi di grado 2^j; se i prodotti di polinomi vengono eseguiti con il metodo descritto nell'esercizio 5.67, ogni prodotto richiede $O(j\, 2^j)$ operazioni, e si hanno

$$O\Big(\sum_{j=1}^{r-1} \frac{n}{2^j}\; j\, 2^j\Big) = O\Big(\sum_{j=1}^{r-1} jn\Big) = O\Big(n\, \frac{r^2}{2}\Big) \quad \text{operazioni.}$$

Quindi per il passo a) sono sufficienti $O(n \log_2^2 n)$ operazioni. In modo analogo si tratta il costo computazionale del passo b), tenendo conto dell'esercizio 5.68.)

5.70 Sia n potenza di 2 e siano $p(x)$ un polinomio di grado al più $n-1$ e $x_0, x_1, \ldots, x_{n-1}$, n punti distinti. Si verifichi che, dati i coefficienti di $p(x)$, è possibile calcolare le differenze divise $b_i = p[x_0, \ldots, x_i]$ per $i = 0, \ldots, n-1$, con $O(n \log_2^2 n)$ operazioni.

(Traccia: posto $m = \dfrac{n}{2}$, per la (21) è $p(x) = s(x)q(x) + r(x)$, dove

$$\begin{aligned} r(x) &= b_0 + (x - x_0)b_1 + \ldots + (x - x_0)\ldots(x - x_{m-2})b_{m-1}, \\ s(x) &= b_m + (x - x_m)b_{m+1} + \ldots + (x - x_m)\ldots(x - x_{n-2})b_{n-1}, \\ q(x) &= (x - x_0)\ldots(x - x_{m-1}). \end{aligned}$$

Si applichi ricorsivamente questo fatto, utilizzando gli esercizi 5.69 per il calcolo dei divisori $(x - x_0)\ldots(x - x_i)$ e 5.68 per il calcolo dei polinomi quoziente e resto.)

5.71 Sia n potenza di 2, e siano $p(x)$ e $q(x)$ due polinomi di grado al più $n-1$. Dati i coefficienti di $p(x)$ e $q(x)$, si dica come si possono calcolare i coefficienti del polinomio $s(x) = p(q(x))$ con $O(n^2 \log_2^2 n)$ operazioni.

(Traccia: il polinomio $s(x)$ ha grado minore di n^2; si calcolano i valori $\alpha_i = q(\omega_{n^2}^i)$ (si veda l'esercizio 5.66) e $\beta_i = p(\alpha_i)$ (si veda l'esercizio 5.69) per $i = 0, \ldots, n^2 - 1$, poi si calcolano i coefficienti del polinomio $s(x)$ tale che $s(\omega_{n^2}^i) = \beta_i$ (si veda l'esercizio 5.66).)

5.72 Sia $n = 2^r$, r intero. Sono dati n punti distinti $x_0, x_1, \ldots, x_{n-1}$, e gli n valori $y_0, y_1, \ldots, y_{n-1}$. Si verifichi che è possibile calcolare i coefficienti del polinomio $p(x)$ di grado al più $n-1$, tale che $p(x_i) = y_i$, per $i = 0, \ldots, n-1$, con $O(n \log_2^2 n)$ operazioni, utilizzando il polinomio di Lagrange

$$p(x) = \pi_{n-1}(x) \sum_{j=0}^{n-1} \frac{y_j}{(x - x_j)\pi'_{n-1}(x_j)}$$

(si veda (5) e (8)), e procedendo con il seguente algoritmo:

a) si calcolano i coefficienti di $\pi_{n-1}(x) = s_0^{(r)}(x)$ eseguendo il passo a) dell'esercizio 5.69, per $j = 1, \ldots, r$;

b) si calcolano i coefficienti di $\pi'_{n-1}(x)$;

c) si calcolano i valori $\pi'_{n-1}(x_j)$ mediante l'algoritmo dell'esercizio 5.69;

d) si calcolano i coefficienti del numeratore della funzione razionale

$$\sum_{j=0}^{n-1} \frac{y_j}{(x - x_j)\pi'_{n-1}(x_j)} = \frac{\theta_0^{(r)}(x)}{s_0^{(r)}(x)},$$

nel modo seguente:

$$\theta_i^{(0)}(x) = \theta_i^{(0)} = \frac{y_i}{\pi'_{n-1}(x_i)}, \quad i = 0, \ldots, n-1, \tag{1}$$

$$\theta_i^{(k)}(x) = s_{2i+1}^{(k-1)}(x)\theta_{2i}^{(k-1)}(x) + s_{2i}^{(k-1)}(x)\theta_{2i+1}^{(k-1)}(x), \tag{2}$$
$$i = 0, \ldots, \frac{n}{2^k} - 1, \; k = 1, \ldots, r;$$

e) risulta quindi $p(x) = \theta_0^{(r)}(x)$.

(Traccia: i costi computazionali dei singoli passi sono: per a) $O(n \log_2^2 n)$, per b) $O(n)$, per c) $O(n \log_2^2 n)$; per d), per il valore k dell'indice vengono effettuate $2\,\dfrac{n}{2^k}$ moltiplicazioni di polinomi, ottenendo polinomi di grado minore o uguale a $2^k - 1$; tali prodotti vengono complessivamente calcolati

come nell'esercizio 5.69 con $O\big(2\,\frac{n}{2^k}\;k2^k\big) = O(kn)$ operazioni; perciò per tutto il passo d) sono sufficienti $O(nr^2) = O(n\log_2^2 n)$ operazioni.)

5.73 Si verifichi che il calcolo in un punto $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ delle n funzioni simmetriche elementari introdotte nell'esercizio 3.48 può essere effettuato con $O(n\log_2^2 n)$ operazioni.

(Traccia: tali funzioni sono, a parte il segno, i coefficienti del polinomio

$$\prod_{i=1}^{n}(x-\alpha_i).\ )$$

5.74 a) Si verifichi che le funzioni $\phi_j(x) = \mathrm{e}^{jx}$, $j = 0, \ldots, n$, (e analogamente le funzioni $\phi_j(x) = \mathrm{e}^{-jx}$, $j = 0, \ldots, n$) possono essere usate come base per l'interpolazione.
b) Si dica come si può risolvere il seguente problema di *interpolazione esponenziale*: dati $n+1$ punti distinti $x_0, \ldots, x_n$ e i corrispondenti valori $y_0, \ldots, y_n$, si determinino gli $n+1$ coefficienti della combinazione lineare

$$q_n(x) = \sum_{j=0}^{n} \alpha_j \phi_j(x),$$

tali che $q_n(x_i) = y_i$, per $i = 0, \ldots, n$. Si risolva il problema dell'interpolazione esponenziale per la funzione

x	0.1	0.2	0.3	0.4
$f(x)$	0.76	0.58	0.44	0.35

(Traccia: a) fissati $n+1$ punti distinti $x_0, \ldots, x_n$, la matrice V i cui elementi sono $v_{ij} = \mathrm{e}^{jx_i}$ è una matrice di Vandermonde formata con i numeri $\beta_i = \mathrm{e}^{x_i}$, $i = 0, \ldots, n$, distinti. b) Si costruisca il polinomio di interpolazione che nei punti β_i, $i = 0, \ldots, n$, assume i valori y_i. Nel caso particolare, poiché la funzione è decrescente, si sceglie la base e^{-jx}, e risulta

$$\beta_0 = 0.9048374,\ \beta_1 = 0.8187308,\ \beta_2 = 0.7408183,\ \beta_3 = 0.6703200,$$

e il polinomio di interpolazione risulta

$$p(x) = -7.316491x^3 + 19.82035x^2 - 15.75648x + 4.209712.$$

Quindi la funzione esponenziale cercata è

$$q_3(x) = -7.316491\mathrm{e}^{-3x} + 19.82035\mathrm{e}^{-2x} - 15.75648\mathrm{e}^{-x} + 4.209712.\ )$$

5.75 Si costruisca la spline cubica naturale che approssima la funzione $f(x) = \sin x \cos x$ nei nodi 0, 1, 1.5, 2, 3. Si valuti l'errore effettivamente commesso nell'approssimazione e si confronti con la maggiorazione teorica.

(Risposta: risulta

$$\mu_0 = \mu_4 = 0, \quad \mu_1 = -2.382718, \quad \mu_2 = -0.3775958, \quad \mu_3 = 2.336162,$$
$$\alpha_0 = 0.8517682, \quad \alpha_1 = -0.9352708, \quad \alpha_2 = -1.124068, \quad \alpha_3 = 0.6280535,$$
$$\beta_0 = 0, \quad \beta_1 = 0.5539286, \quad \beta_2 = 0.08629310, \quad \beta_3 = -0.7677614,$$

ed è $|f(x) - s(x)| < 0.076$, mentre, essendo

$$M_4 = \max_{x\in[0,3]} 8|\sin 2x| = 8, \quad H = 1, \quad h = \frac{1}{2},$$

la maggiorazione teorica di $|f(x) - s(x)|$ è 14.)

5.76 Si costruisca la spline cubica naturale che approssima la funzione di Runge $f(x) = \dfrac{1}{1+x^2}$ nell'intervallo $[-5, 5]$ usando i punti $x_0 = -5, \ldots,$ $x_n = 5$ equidistanti, nei due casi $n = 10$ e $n = 16$. Si confrontino i risultati ottenuti con quelli dell'esempio 5.9, si valuti l'errore effettivamente commesso nell'approssimazione e si confronti con la maggiorazione teorica.

(Risposta: per $n = 10$ è $h_j = 1$, $j = 0, \ldots, 10$, e risulta

$$\mu_0 = \mu_{10} = 0, \quad \mu_1 = \mu_9 = 0.01640203, \quad \mu_2 = \mu_8 = 0.05927858,$$
$$\mu_3 = \mu_7 = 0.09942490, \quad \mu_4 = \mu_6 = 0.7430215, \quad \mu_5 = -1.871511,$$
$$\alpha_0 = -\alpha_9 = 0.01762832, \ \alpha_1 = -\alpha_8 = 0.3403035, \ \alpha_2 = -\alpha_7 = 0.09330893,$$
$$\alpha_3 = -\alpha_6 = 0.1927340, \quad \alpha_4 = -\alpha_5 = 0.9357551,$$
$$\beta_0 = 0.03846154, \quad \beta_1 = \beta_9 = 0.05608985, \quad \beta_2 = \beta_8 = 0.09012020,$$
$$\beta_3 = \beta_7 = 0.1834291, \quad \beta_4 = \beta_6 = 0.3761631, \quad \beta_5 = 1.311918,$$

ed è $|f(x) - s(x)| < 0.022$, mentre, essendo

$$M_4 = \max_{x\in[-5,5]} \frac{|24 - 240x^2 + 120x^4|}{(1+x^2)^5} = 24,$$

la maggiorazione teorica di $|f(x) - s(x)|$ è 21.)

5.77 Sia $f(x) \in C^2[a, b]$ e sia $s(x)$ la polinomiale lineare a tratti che approssima la $f(x)$ nei nodi $a = x_0 < x_1 < \ldots < x_n = b$. Si ricavino delle maggiorazioni per

$$|f'(x) - s'(x)| \quad \text{e} \quad |f(x) - s(x)|,$$

analoghe a quelle del teorema 5.73 per le spline cubiche.

(Traccia: posto

$$M_2 = \max_{x\in[a,b]} |f''(x)|, \quad H = \max_{i=0,\dots,n-1} h_i, \quad \text{e} \quad h = \min_{i=0,\dots,n-1} h_i,$$

in ogni intervallo $[x_i, x_{i+1}]$ dal resto dell'interpolazione lineare (si veda l'esempio 5.6) si ha

$$|f(x) - s(x)| \le \frac{1}{8}\, M_2\, (x_{i+1} - x_i)^2.$$

Quindi su $[a, b]$ risulta

$$|f(x) - s(x)| \le \frac{1}{8}\, M_2\, H^2.$$

Dalla formula di Taylor si ha

$$\begin{aligned} f(x_{i+1}) &= f(x) + (x_{i+1} - x)f'(x) + \frac{1}{2}\,(x_{i+1} - x)^2 f''(\xi_i), \\ f(x_i) &= f(x) + (x_i - x)f'(x) + \frac{1}{2}\,(x_i - x)^2 f''(\eta_i), \end{aligned}$$

da cui sottraendo risulta

$$\frac{f(x_{i+1}) - f(x_i)}{h_i} - f'(x) = \frac{1}{2h_i}\,\big[(x_{i+1} - x)^2 f''(\xi_i) - (x_i - x)^2 f''(\eta_i)\big].$$

Notando che nell'intervallo $[x_i, x_{i+1}]$ è

$$s(x) = \frac{f(x_{i+1}) - f(x_i)}{h_i}\,(x - x_i) + f(x_i),$$

e che

$$(x_{i+1} - x)^2 + (x_i - x)^2 \le (x_{i+1} - x_i)^2,$$

risulta

$$|f'(x) - s'(x)| \le \frac{1}{2h_i}\, M_2\, (x_{i+1} - x_i)^2.$$

Quindi su $[a, b]$ risulta

$$|f'(x) - s'(x)| \le \frac{1}{2}\, M_2\, \frac{H^2}{h}\,.\ )$$

5.78 Sia $f(x) \in C^4[a,b]$ e sia $s(x)$ la polinomiale cubica a tratti di Hermite che approssima la $f(x)$ nei nodi $a = x_0 < x_1 < \ldots < x_n = b$. Si ricavi una maggiorazione per

$$|f(x) - s(x)|.$$

(Traccia: posto

$$M_4 = \max_{x\in[a,b]} |f^{(4)}(x)| \quad \text{e} \quad H = \max_{i=0,\ldots,n-1} h_i,$$

in ogni intervallo $[x_i, x_{i+1}]$ dal resto del polinomio osculatore di Hermite (si veda l'esempio 5.13) si ha

$$|f(x) - s(x)| \leq \frac{1}{4!}\, M_4\, (x - x_i)^2 (x - x_{i+1})^2,$$

e poiché

$$\max_{x\in[x_i,x_{i+1}]} (x - x_i)^2 (x - x_{i+1})^2 = \frac{1}{16}\, (x_{i+1} - x_i)^4,$$

risulta su $[a,b]$ che

$$|f(x) - s(x)| \leq \frac{1}{384}\, M_4\, H^4.$$

5.79 Siano m e n due interi tali che $n \leq m$ e si considerino due suddivisioni dell'intervallo $[a,b]$

$$a = x_0 < x_1 < \ldots < x_n = b,$$
$$a = y_0 < y_1 < \ldots < y_m = b,$$

in cui tutti i nodi della prima suddivisione sono anche nodi della seconda. Indicata con $s(x)$ la spline cubica naturale di una funzione $f(x) \in C^2[a,b]$ relativa alla suddivisione con i punti x_i, e con $t(x)$ la spline cubica naturale relativa alla suddivisione con i punti y_i, si verifichi che

$$\int_a^b [s''(x)]^2\, dx \leq \int_a^b [t''(x)]^2\, dx \leq \int_a^b [f''(x)]^2\, dx.$$

(Traccia: segue dal teorema 5.71.)

5.80 Sia $s(x)$ la spline cubica completa per l'approssimazione di una funzione $f(x)$.

a) Si calcoli
$$S_{n+1} = \int_a^b s(x)\ dx;$$

b) si valuti l'errore che si commette se si utilizza S_{n+1} come approssimazione di
$$\int_a^b f(x)\ dx;$$

c) si scriva S_{n+1} nel caso che i punti x_i siano equidistanti e si confronti con la formula di Eulero-Maclaurin (esercizio 4.42);

d) si utilizzi questa tecnica per approssimare
$$\int_{1/2}^{3/2} \mathrm{e}^{-x^2}\ dx.$$

(Traccia: a) risulta
$$\begin{aligned} S_{n+1} &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s_i(x)\ dx \\ &= \sum_{i=0}^{n-1} \frac{h_i}{2}\ \big[f(x_i) + f(x_{i+1})\big] - \sum_{i=0}^{n-1} \frac{h_i^3}{24}\ (\mu_i + \mu_{i+1}). \end{aligned}$$

La prima sommatoria fornisce lo stesso valore della formula dei trapezi (si veda il capitolo 7). b) Se $f(x) \in C^4[a,b]$, dal teorema 5.73 segue che
$$\left| \int_a^b \Big[f(x) - s(x)\Big]\ dx \right| \le \int_a^b \big|f(x) - s(x)\big|\ dx \le \frac{7}{8}\, M_4\, \frac{H^5}{h}\, (b-a) \le \frac{7n}{8}\, M_4\, \frac{H^6}{h}.$$

c) Se $h_i = h$ per $i = 0, \ldots, n-1$, è
$$S_{n+1} = \frac{h}{2} \sum_{i=0}^{n-1} \big[f(x_i) + f(x_{i+1})\big] - \frac{h^2}{12}\ \big(f'(b) - f'(a)\big).\ )$$

5.81 Per definire le spline cubiche, come condizioni ausiliarie, oltre alle d′), d″) e d‴) del paragrafo 14, si possono imporre anche altre condizioni. Ad esempio:

e′) $s_0''(x_0) = \sigma_0, \quad s_{n-1}''(x_n) = \sigma_n$, dove σ_0 e σ_n sono valori assegnati;

e″) derivata seconda costante vicino agli estremi dell'intervallo $[a,b]$, per cui si assume $s_0''(x_0) = s_0''(x_1)$ e $s_{n-1}''(x_{n-1}) = s_{n-1}''(x_n)$;

e$'''$) derivata seconda lineare vicino agli estremi dell'intervallo $[a, b]$, cioè

$$s_0''(x_0) = \sigma_0 s_0''(x_1) + \sigma_1 s_1''(x_2),$$
$$s_{n-1}''(x_n) = \tau_0 s_{n-1}''(x_{n-1}) + \tau_1 s_{n-2}''(x_{n-2}),$$

dove σ_0, σ_1, τ_0, τ_1 sono valori assegnati.

Si dica come viene modificato il sistema lineare del teorema 5.69 nei diversi casi e se sono verificate le condizioni di esistenza e unicità della spline.

(Traccia: per le e$'$) la prima e l'ultima delle (88) risultano

$$2\mu_1 + \delta_1\mu_2 = 6f[x_0, x_1, x_2] - \gamma_1\sigma_0,$$
$$\gamma_{n-1}\mu_{n-2} + 2\mu_{n-1} = 6f[x_{n-2}, x_{n-1}, x_n] - \delta_{n-1}\sigma_n;$$

per le e$''$) la prima e l'ultima delle (88) risultano

$$(2 + \gamma_1)\mu_1 + \delta_1\mu_2 = 6f[x_0, x_1, x_2],$$
$$\gamma_{n-1}\mu_{n-2} + (2 + \delta_{n-1})\mu_{n-1} = 6f[x_{n-2}, x_{n-1}, x_n];$$

per le e$'''$) la prima e l'ultima delle (88) risultano

$$(2 + \sigma_0\gamma_1)\mu_1 + (\delta_1 + \sigma_1\gamma_1)\mu_2 = 6f[x_0, x_1, x_2],$$
$$(\gamma_{n-1} + \tau_1\delta_{n-1})\mu_{n-2} + (2 + \tau_0\delta_{n-1})\mu_{n-1} = 6f[x_{n-2}, x_{n-1}, x_n].$$

In quest'ultimo caso non è garantita l'esistenza e l'unicità della spline. Per questo è opportuno imporre le condizioni sufficienti

$$|2 + \sigma_0\gamma_1| > |\delta_1 + \sigma_1\gamma_1| \quad \text{e} \quad |2 + \tau_0\delta_{n-1}| > |\gamma_{n-1} + \tau_1\delta_{n-1}|$$

che assicurano la predominanza diagonale della matrice.)

5.82 Siano $a = x_0 < x_1 < \ldots < x_n = b$ e sia $2 \le m \le n + 1$. La spline di ordine $2m - 1$ per l'approssimazione di una funzione $f(x)$ viene definita in uno dei due modi seguenti.

Definizione variazionale: si definisce *spline di ordine* $2m - 1$ una funzione reale $s(x)$ tale che

(1) $s(x_i) = f(x_i)$, per $i = 0, \ldots, n$,

(2) $s(x) \in C^{m-1}[a, b]$,

(3) l'integrale

$$\int_a^b [s^{(m)}(x)]^2 \, dx$$

esiste ed è quello minimo fra gli integrali di tutte le funzioni che soddisfano le condizioni (1) e (2).

Definizione descrittiva: si definisce *spline di ordine* $2m-1$ una funzione reale $s(x)$ tale che

(1) $s(x_i) = f(x_i)$, per $i = 0, \ldots, n$,

(2) $s(x)$ è un polinomio di grado al più $2m-1$ in ogni intervallo $[x_i, x_{i+1}]$, $i = 0, \ldots, n-1$,

(3) $s(x) \in C^{2m-2}[a,b]$,

(4) $$s^{(m)}(a) = s^{(m+1)}(a) = \ldots = s^{(2m-2)}(a) = 0,$$
$$s^{(m)}(b) = s^{(m+1)}(b) = \ldots = s^{(2m-2)}(b) = 0,$$

(la definizione 5.68 con la condizione d′) corrisponde alla definizione descrittiva con $m = 2$). Si verifichi che

a) la definizione descrittiva individua univocamente la spline;

b) le due definizioni variazionale e descrittiva sono equivalenti.

(Traccia: a) siano $s_1(x)$ e $s_2(x)$ due spline che soddisfano la definizione descrittiva. Allora la funzione $s(x) = s_1(x) - s_2(x)$ verifica le (1) − (4) della definizione descrittiva relativamente alla funzione $f(x)$ identicamente nulla. Poiché in ogni intervallo $s(x)$ coincide con un polinomio di grado al più $2m-1$, essa è individuata da $2mn$ coefficienti. Le (1), (3) e (4) forniscono esattamente $2mn$ condizioni lineari sui coefficienti. Quindi i coefficienti di $s(x)$ soddisfano il sistema lineare $A\mathbf{x} = \mathbf{0}$ di ordine $2mn$. Per verificare che tale sistema ha un'unica soluzione, si applichi ripetutamente la formula di integrazione per parti:

$$\begin{aligned}
\int_a^b [s^{(m)}(x)]^2 \, dx &= \Big[s^{(m-1)}(x) s^{(m)}(x) \Big]_a^b - \int_a^b s^{(m-1)}(x) s^{(m+1)}(x) \, dx \\
&= \ldots = (-1)^m \int_a^b s''(x) s^{(2m-2)}(x) \, dx \\
&= (-1)^m \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s''(x) s^{(2m-2)}(x) \, dx \\
&= (-1)^m \sum_{i=0}^{n-1} \Big[s'(x) s^{(2m-2)}(x) - s(x) s^{(2m-1)}(x) \Big]_{x_i}^{x_{i+1}},
\end{aligned}$$

in quanto $s^{(2m)}(x) \equiv 0$ su ogni intervallo, e poiché $s(x_i) = 0$ per $i = 0, \ldots, n$ e $s^{(2m-2)}(a) = s^{(2m-2)}(b) = 0$, ne segue che

$$\int_a^b [s^{(m)}(x)]^2 \, dx = 0.$$

Quindi $s(x)$, per le condizioni di continuità, deve essere un polinomio di grado $m-1$, e poiché si annulla in $n+1$ punti, con $n+1 > m-1$, ne segue che è identicamente nulla la spline che approssima la funzione identicamente nulla.
b) Sia $s(x)$ la spline che soddisfa le (1) − (4) della definizione descrittiva per l'approssimazione di una funzione $f(x)$ e $t(x)$ una funzione che soddisfa le (1) e (2) della definizione variazionale, allora

$$\begin{aligned}\int_a^b [t^{(m)}(x)]^2\ dx - \int_a^b [s^{(m)}(x)]^2\ dx &= \int_a^b [t^{(m)}(x) - s^{(m)}(x)]^2\ dx \\ &+ 2\int_a^b [t^{(m)}(x) - s^{(m)}(x)]s^{(m)}(x)\ dx \\ &\geq 2\int_a^b [t^{(m)}(x) - s^{(m)}(x)]s^{(m)}(x)\ dx,\end{aligned}$$

in cui il segno di uguaglianza vale solo se $s(x)$ e $t(x)$ coincidono identicamente. Si dimostri, procedendo come per la dimostrazione del teorema 5.71, che l'ultimo integrale è nullo. Ne segue che

$$\int_a^b [t^{(m)}(x)]^2\ dx \geq \int_a^b [s^{(m)}(x)]^2\ dx$$

e che il minimo viene assunto solo dalla $s(x)$.)

5.83 Si definisca e si costruisca la spline quadratica per l'approssimazione di una funzione $f(x)$. La spline quadratica viene usata raramente nella pratica, perché meno efficace di quella cubica dal punto di vista grafico, ed inoltre perché, per certe scelte dell'unica condizione ausiliaria, è instabile. Si individui la causa di tale instabilità.

(Traccia: si definisce *spline quadratica* una funzione reale $s(x) \in C^1[a,b]$, che in ogni intervallo coincide con un polinomio $s_i(x)$ di grado al più 2 e tale che $s(x_i) = f(x_i)$, per $i = 0, \ldots, n$. Procedendo come per le spline cubiche, si ponga $\nu_i = s_i'(x_i)$, per $i = 0, \ldots, n-1$ e $\nu_n = s_{n-1}'(x_n)$; risulta quindi

$$s_i'(x) = \nu_{i+1}\ \frac{x - x_i}{h_i}\ - \nu_i\ \frac{x - x_{i+1}}{h_i}\,, \quad h_i = x_{i+1} - x_i.$$

Integrando si ha

$$s_i(x) = \nu_{i+1}\ \frac{(x - x_i)^2}{2h_i}\ - \nu_i\ \frac{(x - x_{i+1})^2}{2h_i}\ + \alpha_i,$$

e imponendo le condizioni che $s(x_i) = f(x_i)$, per $i = 0 \ldots, n$, si ottiene

$$\begin{cases} -\nu_i \dfrac{h_i}{2} + \alpha_i = f(x_i) \\ \nu_{i+1} \dfrac{h_i}{2} + \alpha_i = f(x_{i+1}), \end{cases}$$

da cui

$$\nu_i + \nu_{i+1} = 2f[x_i, x_{i+1}], \quad \text{per} \quad i = 0, \ldots, n-1.$$

Questa equazione alle differenze del primo ordine consente di calcolare i ν_i, purché sia assegnato un valore iniziale $\nu_0 = s_0'(x_0)$ oppure $\nu_n = s_{n-1}'(x_n)$. Ad esempio con la condizione $\nu_0 = 0$ si ottiene

$$\nu_i = 2 \sum_{j=0}^{i-1} (-1)^{i+j-1} f[x_j, x_{j+1}].$$

Gli α_i, $i = 0, \ldots, n-1$ vengono poi calcolati per sostituzione. L'instabilità è causata dal fatto che si risolve un'equazione alle differenze la cui equazione omogenea associata ha la soluzione $\nu_i = (-1)^i \nu_0$, mentre la soluzione dell'equazione completa potrebbe essere decrescente. Si noti anche che se viene modificato un valore della funzione $f(x_k)$, la soluzione varia solo nelle componenti ν_i, con $i \geq k$, diversamente da quanto accade per le spline cubiche, in cui la variazione di un valore della funzione si ripercuote su tutti i nodi, con intensità decrescente con la distanza.)

5.84 Si dica qual è il costo computazionale

a) della costruzione della spline cubica per n nodi (si consideri anche il caso dei nodi equidistanti);

b) del calcolo del valore della spline in un punto x, una volta che sia stato individuato l'indice i tale che $x \in (x_i, x_{i+1})$;

c) se n è potenza di 2, si dica quanti confronti sono richiesti al più per determinare l'indice i tale che $x \in (x_i, x_{i+1})$ con l'algoritmo banale (cioè confrontando x successivamente con x_j, $j = 1, \ldots, n-1$) e con l'algoritmo di bisezione (cioè confrontando x con $x_{n/2}$, se x è minore di $x_{n/2}$ si confronta x con $x_{n/4}$, se x è maggiore di $x_{n/2}$ si confronta x con $x_{3n/4}$, e così via).

(Traccia: a) si calcola il costo computazionale, a meno di termini costanti rispetto ad n. Indicando con A le operazioni additive e con M le operazioni moltiplicative, per la costruzione degli h_i, degli $h_i/6$ e delle differenze divise

$[f(x_i) - f(x_{i-1})]/h_{i-1}$ sono richieste $2n$A e $2n$M. Per la costruzione dei coefficienti e dei termini noti del sistema lineare

$$\frac{h_{i-1}}{6}\,\mu_{i-1}+2\big(\frac{h_{i-1}}{6}+\frac{h_i}{6}\big)\mu_i+\frac{h_i}{6}\,\mu_{i+1}=\frac{f(x_{i+1})-f(x_i)}{h_i}-\frac{f(x_i)-f(x_{i-1})}{h_{i-1}}$$

sono richieste $2n$A e nM. Per la risoluzione del sistema tridiagonale sono richieste $3n$A e $5n$M. Per il calcolo degli α_i e β_i sono richieste $3n$A e $3n$M. Quindi in totale la costruzione della spline richiede $10n$A e $11n$M. Nel caso dei punti equidistanti sono richieste $2n$A per la costruzione dei termini noti del sistema, $3n$A e $5n$M per la risoluzione del sistema lineare, $3n$A e $3n$M per il calcolo degli α_i e β_i. Quindi in totale $8n$A e $8n$M.
b) Sono richieste 5A e 9M.
c) $n-1$ confronti con l'algoritmo banale, $\log_2 n$ con l'algoritmo di bisezione.)

5.85 Facendo un'analisi dell'errore all'indietro (si veda il paragrafo 10 del capitolo 2) si verifichi che i μ_i, $i = 1, \ldots, n-1$, di una spline naturale, effettivamente calcolati risolvendo il sistema lineare (88), sono soluzione di un sistema con matrice e termini noti ottenuti perturbando la matrice e i termini noti del sistema (88) con perturbazioni maggiorabili in modulo da quantità indipendenti da n e quindi il calcolo dei μ_i è stabile.

(Traccia: con il metodo di Gauss il sistema $A\boldsymbol{\mu} = \mathbf{b}$, dove

$$A=\begin{bmatrix} 2 & \delta_1 & & & \\ \gamma_2 & 2 & \delta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-2} & 2 & \delta_{n-2} \\ & & & \gamma_{n-1} & 2 \end{bmatrix}, \qquad \mathbf{b} = 6\begin{bmatrix} f[x_0,x_1,x_2] \\ f[x_1,x_2,x_3] \\ \vdots \\ f[x_{n-3},x_{n-2},x_{n-1}] \\ f[x_{n-2},x_{n-1},x_n] \end{bmatrix},$$

si trasforma nel sistema equivalente $U\boldsymbol{\mu} = \mathbf{c}$, dove

$$U=\begin{bmatrix} d_1 & \delta_1 & & & \\ & d_2 & \delta_2 & & \\ & & \ddots & \ddots & \\ & & & d_{n-2} & \delta_{n-2} \\ & & & & d_{n-1} \end{bmatrix}$$

e

$$\left.\begin{aligned} d_1 &= 2, \quad c_1 = b_1 \\ d_{i+1} &= 2 - \frac{\gamma_{i+1}\delta_i}{d_i} \\ c_{i+1} &= b_{i+1} - \frac{\gamma_{i+1}c_i}{d_i} \end{aligned}\right\}, \quad i = 1, \ldots, n-2.$$

Con l'analisi dell'errore all'indietro si ha

$$\begin{aligned}\widetilde{d}_{i+1} &= \Big[2 - \frac{\gamma_{i+1}\delta_i(1+\epsilon_{i+1})}{\widetilde{d}_i}\,(1+\eta_{i+1})\Big](1+\zeta_{i+1}) \\ &= 2(1+\zeta_{i+1}) - \frac{\gamma_{i+1}}{\widetilde{d}_i}\,\delta_i(1+\xi_{i+1}),\end{aligned}$$

dove

$$|\epsilon_{i+1}|,\ |\eta_{i+1}|,\ |\zeta_{i+1}| < u, \quad |\xi_{i+1}| \dot{<} 3u, \quad \text{e } u \text{ è la precisione di macchina.}$$

Quindi i $\widetilde{d}_i$ effettivamente calcolati sono quelli che si otterrebbero operando in modo esatto su una matrice perturbata $A + \Delta A$, con $|\Delta A| \dot{<} 3u|A|$. Un analogo risultato si ottiene per il vettore $\mathbf{c}$ e per la soluzione del sistema $\widetilde{U}\boldsymbol{\mu} = \widetilde{\mathbf{c}}$. Le maggiorazioni che si ottengono non dipendono dalla dimensione n del sistema.)

5.86 Siano $a = x_0 < x_1 < \ldots < x_n = b$, con $n \geq 4$. Considerati i punti ausiliari $x_{-3} < x_{-2} < x_{-1} < a$ e $b < x_{n+1} < x_{n+2} < x_{n+3}$, si definiscono *B-spline* (o *spline fondamentali) cubiche normalizzate* le funzioni $S_i(x)$, $i = -1, \ldots, n+1$, dotate delle proprietà

(1) $S_i(x) \in C^2[x_{-3}, x_{n+3}]$;

(2) in ogni intervallo $S_i(x)$ coincide con un polinomio di terzo grado;

(3) $S_i(x) \equiv 0$ per $x \leq x_{i-2}$ e $x \geq x_{i+2}$;

(4) $S_i(x_i) = 1$.

La condizione (4) può essere sostituita da una qualunque altra condizione che faccia sì che la spline non sia identicamente nulla, ad esempio

$$\int_{x_{i-2}}^{x_{i+2}} S_i(x)\ dx = 1.$$

Si verifichi che le B-spline così definite

a) esistono e sono uniche,

b) $S_i(x) > 0$ per $x \in (x_{i-2}, x_{i+2})$,

c) sono linearmente indipendenti.

d) Si verifichi che ogni spline cubica $s(x)$ approssimante una funzione $f(x)$ su $[a, b]$ nei nodi $x_0, \ldots, x_n$ può essere scritta come

$$s(x) = \sum_{i=-1}^{n+1} \alpha_i S_i(x), \tag{104}$$

dove gli α_i sono degli opportuni coefficienti.

e) Nel caso particolare dei nodi equidistanti si esprimano $S_i(x)$, $S_i'(x)$, $S_i''(x)$ per $i = -1, \ldots, n+1$.

Si osservi che se uno dei coefficienti α_i della (104) viene modificato, la spline $s(x)$ risulta modificata solo nell'intervallo (x_{i-2}, x_{i+2}), e quindi non è necessario ricalcolarla tutta. Questo è particolarmente vantaggioso nelle applicazioni grafiche.

(Traccia: a) i coefficienti della spline sono soluzione di un sistema lineare e l'esistenza e unicità vengono dimostrate verificando la non singolarità della matrice del sistema. Per questo basta dimostrare che se una funzione $t_i(x)$ soddisfa le (1), (2), (3) ed inoltre è tale che $t_i(x_i) = 0$, allora $t_i(x) \equiv 0$. Infatti $t_i(x)$ si annulla in almeno 3 punti nell'intervallo $[x_{i-2}, x_{i+2}]$, quindi $t_i'(x)$ si annulla in almeno due punti interni a tale intervallo, e poiché $t_i'(x_{i-2}) = t_i'(x_{i+2}) = 0$, $t_i'(x)$ si annulla in almeno 4 punti di $[x_{i-2}, x_{i+2}]$. In modo analogo si vede che $t_i''(x)$ si annulla in almeno 5 punti di $[x_{i-2}, x_{i+2}]$, di cui 3 interni, e questo è possibile solo se $t_i(x) \equiv 0$, perché $t_i(x)$ coincide con 4 polinomi di terzo grado sui 4 intervalli $[x_i, x_{i+1}]$, $i = i-2, \ldots, i+1$. b) Si proceda come in a). c) Si consideri una combinazione nulla delle B-spline

$$\phi(x) = \sum_{i=-1}^{n+1} \alpha_i S_i(x) = 0.$$

Poiché $S_i(x_{-2}) = 0$ per $i \geq 0$, risulta $\phi(x_{-2}) = \alpha_{-1} S_{-1}(x_{-2})$, e dovendo essere $\phi(x_{-2}) = 0$, ne segue che $\alpha_{-1} = 0$. Si proceda considerando i valori $\phi(x_k)$, $k = -1, \ldots, n$. Risulta così che i coefficienti della combinazione sono tutti nulli.

d) Per ogni intervallo (x_i, x_{i+1}), $i = 0, \ldots, n-1$, esistono quattro B-spline linearmente indipendenti e non nulle con cui è possibile esprimere qualunque polinomio di terzo grado.

e) Indicato con h il passo costante e posto $y_i = \dfrac{x - x_i}{h}$, risulta

$$S_i(x) = \begin{cases} (y_i^3 + 6y_i^2 + 12y_i + 8)/4 & \text{per } -2 \leq y_i \leq -1 \\ (-3y_i^3 - 6y_i^2 + 4)/4 & \text{per } -1 \leq y_i \leq 0 \\ (3y_i^3 - 6y_i^2 + 4)/4 & \text{per } 0 \leq y_i \leq 1 \\ (-y_i^3 + 6y_i^2 - 12y_i + 8)/4 & \text{per } 1 \leq y_i \leq 2 \\ 0 & \text{per } y_i \leq -2 \text{ e } y_i \geq 2. \end{cases}$$

Si veda la figura 5.29.)

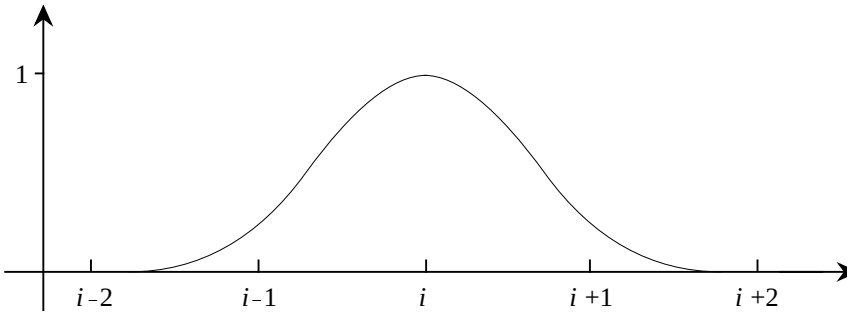


Fig. 5.29 - B-spline

Commento bibliografico

L'avvio degli studi nel campo dell'interpolazione coincise, a partire dal 17^o secolo, con il rapido sviluppo della matematica dovuto alla definitiva acquisizione del simbolismo matematico e alla necessità di compilare delle valide tabelle di valori delle funzioni trigonometriche e dei logaritmi, per rendere più agevole il calcolo e le misurazioni nel campo della geometria e dell'astronomia e per gli usi nautici. Lo studio dei procedimenti che stanno alla base della compilazione delle tabelle fu il primo passo nello sviluppo della teoria dell'interpolazione: già prima del 1611 Harriot studiò i polinomi di interpolazione e nel 1617 Briggs pubblicò le sue tavole dei logaritmi, per la cui compilazione aveva fatto uso di tecniche di calcolo con le differenze finite. Anche l'interesse di Newton per la teoria dell'interpolazione nacque dallo studio dei logaritmi: molte delle formule che oggi sono conosciute con il nome di matematici posteriori (come la formula di Bessel) furono per la prima volta descritte da Newton stesso. A Newton si può far risalire la nozione di differenza divisa, anche se il termine oggi usato fu introdotto da De Morgan nel 1842. I matematici della scuola inglese del 17^o secolo, Gregory, Taylor, Halley, Stirling, Maclaurin, hanno poi contribuito allo sviluppo della teoria dell'intepolazione. Per una storia dettagliata dell'analisi numerica di quel periodo si veda il libro di Goldstine [11].

La formula di interpolazione nota con il nome di Lagrange appare effettivamente nell'opera di Lagrange nel 1795, ma già prima altri matematici, come Eulero nel 1755 e Waring nel 1779, l'avevano utilizzata. Prima dell'avvento dei calcolatori erano stati estensivamente tabulati i coefficienti delle formule di interpolazione di Lagrange con nodi equidistanti e anche con alcuni nodi non equidistanti. Oltre che l'interpolazione polinomiale, Lagrange studiò anche l'interpolazione trigonometrica, di cui si era occupato anche Clairaut nel 1759. Gauss si occupò di interpolazione polinomiale, razionale e trigonometrica. Mise in evidenza la connessione tra polinomio di Lagrange e polinomio di Newton e trovò le relazioni fra i coefficienti

dei polinomi trigonometrici, che però furono trascurate fino a quando non furono riscoperte da Cooley e Tukey nel 1965.

Il resto del polinomio di interpolazione dato in questo testo è quello di Cauchy. Vi sono altre forme in cui il resto può essere espresso, come ad esempio la forma di Peano e la forma di Kowalewski (si veda [7]).

L'algoritmo di Neville (si veda l'esercizio 5.26) è del 1934 ed è derivato da un altro metodo descritto da Aitken nel 1932. Algoritmi di tipo Neville possono essere impiegati anche nell'interpolazione razionale, si veda [22]. Per l'interpolazione in più variabili si veda [23].

Uno dei fatti più sorprendenti dell'interpolazione polinomiale, anche alla luce del teorema di Weierstrass sull'approssimazione, è che aumentando il grado del polinomio non sempre si ha convergenza alla funzione. La prima indicazione di questo fatto fu data alla fine del secolo scorso, quando Runge scoprì che se si interpola la funzione $f(x) = 1/(1 + x^2)$ in nodi equidistanti dell'intervallo $[-5, 5]$ vi è convergenza solo per $|x| < 3.63\ldots$. La funzione $f(x)$ è analitica su tutto l'asse reale, però la divergenza è indotta dalla singolarità in $\pm\boldsymbol{i}$. Nel 1912 Bernstein dimostrò che l'interpolazione della funzione $f(x) = |x|$ su nodi equidistanti dell'intervallo $[-1, 1]$ diverge per ogni $x \neq -1, 0, 1$. Risultati analoghi di non convergenza valgono anche nel caso che i nodi non siano equidistanti: anche assumendo come nodi gli zeri dei polinomi di Chebyshev si ottengono in taluni casi successioni di polinomi di interpolazione che non convergono. Nel 1937 Marcinkiewicz costruì una funzione continua il cui polinomio di interpolazione sui nodi di Chebyshev diverge in ogni punto di $(-1, 1)$.

Il risultato fondamentale per la non convergenza è quello ottenuto simultaneamente da Bernstein e Faber nel 1914, per cui se i nodi dell'interpolazione sono assegnati, è sempre possibile determinare una funzione continua per cui il processo di interpolazione su questi nodi non converge uniformemente alla funzione. Nel 1916 Bernstein dimostrò che l'interpolazione nei nodi di Chebyshev è convergente se la funzione $f(x)$ è continua ed inoltre gode di certe proprietà di regolarità. Quindi perché vi sia convergenza vi deve essere uno stretto legame fra la distribuzione dei nodi e le proprietà di crescenza della funzione. Se l'interpolazione è eseguita su un intervallo reale limitato per la convergenza è sufficiente supporre che la funzione sia analitica in opportune regioni del piano complesso contenenti l'intervallo. Per una trattazione completa del problema della convergenza si veda il libro di Davis [7].

Il primo uso delle frazioni continue sembra che sia stato fatto nel 16^o secolo per esprimere mediante frazioni con numeratori e denominatori piccoli delle frazioni con numeratori e denominatori grossi, che venivano rappresentate con difficoltà nel periodo iniziale dell'uso della rappresentazione posizionale dei numeri. Ben presto però le frazioni continue trovarono ap-

plicazione nell'approssimazione dei numeri irrazionali: nel 1572 Bombelli dette l'approssimazione

$$3 + \cfrac{4}{6 + \cfrac{4}{6}}$$

per $\sqrt{13}$ e nel 1613 Cataldi dette l'approssimazione

$$4 + \cfrac{2}{8 + \cfrac{2}{8 + \cfrac{2}{8}}}$$

per $\sqrt{18}$. L'uso sistematico delle proprietà delle frazioni continue si diffuse nel 1700 ad opera di Eulero, che le utilizzò come valido strumento nello studio della teoria dei numeri e dell'analisi matematica. Successivamente se ne interessarono, fra gli altri, Lagrange, Gauss, Galois, Liouville e Stieltjes. Nel 1803 Viskovatov pubblicò un articolo in cui si illustrava un metodo per trasformare le serie di potenze in frazioni continue, nel 1906 Thiele introdusse una nuova classe di frazioni continue, legate alle differenze inverse, che sono alla base dell'interpolazione razionale. I libri più importanti sull'argomento sono quelli di Perron [17], di Wall [24] e più recentemente di Jones e Thron [14], in cui è riportata anche un'ampia introduzione storica. Per una trattazione più elementare si veda [6].

La trasformata discreta di Fourier, oltre che per il calcolo dei coefficienti del polinomio trigonometrico di interpolazione, è utilizzata nella risoluzione di moltissimi problemi della matematica applicata, in particolare come approssimazione della trasformata continua di Fourier. Un uso frequente della DFT viene fatto attualmente nell'analisi e nel filtraggio dei segnali. La tecnica FFT per il calcolo della trasformata discreta di Fourier, che viene fatta risalire a Gauss e a Runge, si basa essenzialmente sul teorema 5.63, scoperto da Danielson e Lanczos nel 1942. In realtà l'algoritmo FFT oggi usato è stato elaborato, in modo indipendente dai lavori precedenti, da Cooley e Tuckey nel 1965. In [9] e [4] sono riportate le proprietà della DFT e sono descritte le molte tecniche di calcolo della FFT. Per l'uso della FFT nei problemi concernenti i polinomi e le serie di potenze, si veda [3]. Trasformate di Fourier su campi finiti sono state usate in [20], come base per l'algoritmo di moltiplicazione di interi con d cifre di costo $O(d \log_2 d \log_2 \log_2 d)$. Per la storia dell'FFT si veda [5].

Il termine inglese di *spline*, che si potrebbe tradurre in italiano con *curvilinea*, indicava uno strumento, usato un tempo nella progettazione navale per tracciare linee curve e formato da una sottile striscia di materiale flessibile, legno o metallo, che veniva bloccata con dei chiodi in corrispondenza ai punti per cui doveva passare la linea. La forma assunta dalla striscia è quella che minimizza l'energia potenziale causata dalla deformazione del materiale e che è inversamente proporzionale in ogni punto al raggio di curvatura.

L'uso delle funzioni polinomiali a tratti risale al 1943 quando Courant se ne servì per la formulazione variazionale di problemi di vibrazioni. Da allora tali metodi sono stati adottati con successo nella tecnica detta *degli elementi finiti*. La definizione originale di spline cubica venne data da Schoenberg nel 1946, ma lo studio approfondito prese l'avvio con il teorema di minimo dato da Hollady nel 1957. Nel 1964 Birkhoff e deBoor dimostrarono che in norma ∞ per le spline cubiche vi è convergenza fino alla derivata quarta, almeno quando il rapporto fra la massima e la minima delle lunghezze degli intervalli è limitato. Un testo classico sulle spline è [1]. Per limitazioni migliori dell'errore della spline cubica si vedano [21] e [12]. Interessanti considerazioni, oltre che sull'uso pratico, sulla complessità e il condizionamento del calcolo delle splines possono essere trovate in [8]. Per l'uso delle spline nella risoluzione delle equazioni alle differenze e delle equazioni differenziali si veda [15].

Bibliografia

[1] J. H. Ahlberg, E. N. Nilson, J. L. Walsh. *The Theory of Splines and their Applications*, Academic Press, New York, 1967.

[2] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, Bologna, 1988.

[3] D. Bini, V. Pan, *Numerical and Algebraic Computations with Matrices and Polynomials, vol.1: Fundamental Algorithms*, Birkhäuser-Verlag, Boston, 1991.

[4] E. O. Brigham, *The Fast Fourier Transform*, Prentice Hall, Englewood Cliffs, N. J., 1974.

[5] J. W. Cooley, P. A. Lewis, P. D. Welch, "History of the Fast Fourier Transform", *Proc. IEEE*, 55, 1967, pp. 1675-1677.

[6] A. Cuyt, L. Wuytack, *Nonlinear Methods in Numerical Analysis*, North-Holland, Amsterdam, 1987.

[7] P. J. Davis, *Interpolation and Approximation*, Blaisdell, Pub. Co., New York, 1963.

[8] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[9] D. F. Elliott, K. R. Rao, *Fast Transforms Algorithms, Analyses, Applications*, Academic Press, New York, 1982.

[10] W. M. Gentleman, G. Sande, "Fast Fourier Transform - for Fun and Profit", *Proc. AFIPS 1966 Fall Joint Comput. Conf.*, 29, Spartan Books, Washington, 1966, pp. 563-578.

[11] H. H. Goldstine, *A History of Numerical Analysis from the 16th through the 19th Century*, Springer-Verlag, New York, 1977.

[12] C. A. Hall, W. W. Meyer, "Optimal Error Bounds for Cubic Spline Interpolation", *J. of Appr. Theory*, 16, 1976, pp. 105-122.

[13] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[14] W. B. Jones, W. J. Thron, *Continued Fractions, Analytic Theory and Applications*, Encyclopedia of Mathematics and its Applications, vol. 11, Addison-Wesley, Reading, 1980.

[15] G. I. Marchuk, *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1975.

[16] L. M. Milne-Thomson, *The Calculus of Finite Differences*, Macmillan and Co., London, 1933.

[17] O. Perron, *Die Lehre von den Kettenbrüchen*, Band I, II, Teubner, Stuttgart, 1954, 1957.

[18] M. J. D. Powell, "On the Maximum Errors of Polynomial Approximations Defined by Interpolation and by Least Squares Criteria", *Comput. J.*, 9, 1967, pp. 404-407.

[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes. The Art of Scientific Computing*, Cambridge Univ. Press, Cambridge, 1986.

[20] A. Schönhage, V. Strassen, "Schnelle Multiplikation Grosser Zahlen", *Computing* 7, 1971, pp. 281-292.

[21] M. H. Shultz, *Spline Analysis*, Prentice Hall, Englewood Cliffs, N. J., 1973.

[22] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[23] H. C. Thacher, W. E. Milne, "Interpolation in Several Variables", *Siam J.*, 8, 1960, pp. 33-42.

[24] H. S. Wall, *Analytic Theory of Continued Fractions*, Van Nostrand, New York, 1948.

[25] S. Winograd, *Arithmetic Complexity of Computations*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, Penn., 1980.

Capitolo 6

APPROSSIMAZIONE

1. Introduzione

Il problema dell'interpolazione esaminato nel capitolo precedente per i modelli polinomiale, razionale e trigonometrico, può essere visto come un caso particolare del più ampio problema dell'*approssimazione* di funzioni. A grandi linee si possono distinguere due casi in cui è richiesta l'approssimazione di una funzione $f(x)$:

a) La funzione $f(x)$ è una funzione matematica, spesso non razionale, di cui sono conosciute proprietà di continuità e di derivabilità, che possono essere sfruttate per scriverne la formula di Taylor, ed eventualmente proprietà specifiche, come ad esempio proprietà di simmetria e di periodicità, che servono per semplificare i calcoli. In questo caso l'approssimazione della $f(x)$ ha lo scopo di produrre una funzione $g(x)$ più semplice, cioè più facilmente "trattabile" (ad esempio più facilmente calcolabile, derivabile o integrabile) della $f(x)$.

b) La funzione $f(x)$ è nota su un insieme di punti x_i, $i = 0, \ldots, m$, detti *nodi*. In questo caso il problema consiste nel costruire una funzione $g(x)$ che consenta di approssimare altri valori della funzione $f(x)$. Questo è un caso caratteristico nell'analisi di dati sperimentali e presenta una grande variabilità: è infatti possibile che il numero $m+1$ di dati disponibili sia piccolo, come nello studio di certi fenomeni fisici o biologici, o molto elevato, come di solito accade nelle analisi statistiche. I valori $f(x_i)$ possono essere affetti da errori di misura limitati e con una distribuzione abbastanza regolare, come accade di solito per i dati rilevati in laboratorio, ma possono anche essere affetti da errori molto elevati e senza alcuna distribuzione regolare, come può accadere per esempio per un non corretto funzionamento degli strumenti di laboratorio o quando in una indagine economica o demografica i dati vengono rilevati in modo errato.

La differenza fondamentale fra i due casi è che nel caso a), che verrà indicato come *caso continuo*, per la costruzione della $g(x)$ si possono sfruttare, oltre a tutte le informazioni di carattere globale che si hanno sulla $f(x)$, anche i valori che la $f(x)$ assume in un insieme opportuno di punti. Questo insieme è legato al modello scelto per l'approssimazione e non ai dati del problema, e si assume che in tali punti i valori della $f(x)$ siano disponibili con una precisione arbitraria. Nel caso b), indicato come *caso discreto*, invece, la precisione dei dati (nodi e valori della funzione) è assegnata. Anche

se il problema è espresso in forma più semplice, le tecniche di risoluzione sono più complicate, dovendo tenere conto dei vincoli imposti dal problema.

In questo capitolo verrà dato più spazio al caso continuo, riservando il solo paragrafo 13 al caso discreto. Per il caso continuo verrà prima trattato il modello polinomiale e successivamente quello razionale.

Per l'approssimazione con polinomi è fondamentale il seguente teorema, di cui esistono varie dimostrazioni: si veda nell'esercizio 6.1 la dimostrazione data da Bernstein e nell'esercizio 6.56 la dimostrazione data da Fejér.

6.1 Teorema *(di Weierstrass).* *Sia $f(x) \in C[a,b]$. Per ogni $\epsilon > 0$, esiste un polinomio $p(x)$ tale che per ogni $x \in [a,b]$ è*

$$|f(x) - p(x)| \leq \epsilon. \qquad \blacksquare$$

Questo teorema in pratica assicura che l'insieme dei polinomi è denso nell'insieme delle funzioni continue, cioè che per ogni funzione continua esiste sempre un polinomio che la approssima con errore arbitrariamente piccolo. Nel teorema però non viene fornita alcuna indicazione sul grado del polinomio approssimante, che per un ϵ prefissato potrebbe anche essere estremamente elevato.

Per studiare la velocità di convergenza di una successione di polinomi ad una funzione $f(x) \in C[a,b]$, si definisce la *distanza* di $f(x)$ dal sottospazio dei polinomi di grado al più n come

$$d_n(f) = \min_{[a_0,\ldots,a_n]} \max_{x\in[a,b]} \Big| f(x) - \sum_{i=0}^{n} a_i x^i \Big|.$$

Ebbene, è possibile dimostrare che per ogni successione $\{\epsilon_n\}$ monotona, decrescente e convergente a zero, esiste una $h(x) \in C[a,b]$ tale che $d_n(h) \geq \epsilon_n$ per ogni n (si veda l'esercizio 6.55). Però, se la funzione $f(x)$, oltre ad essere continua, ha altre proprietà di regolarità, la situazione, come risulta dal teorema 6.35, non è così pessimistica. Comunque, anche se la classe dei polinomi sembra la scelta più conveniente, se non è possibile ottenere l'approssimazione con un polinomio di grado basso, per problemi di complessità computazionale e di stabilità conviene utilizzare altri modelli.

Un metodo classico per ottenere approssimazioni polinomiali di funzioni $f(x) \in C^n[a,b]$ è fornito dalla formula di Taylor troncata all'n-esimo termine

$$p(x) = f(x_0) + \sum_{i=1}^{n} (x - x_0)^i \, \frac{f^{(i)}(x_0)}{i!}, \quad x_0 \in (a,b).$$

Il resto dell'approssimazione, dato da $r(x) = f(x) - p(x)$, può essere rappresentato in vari modi, a seconda delle proprietà della $f(x)$. Ad esempio,

se $f(x) \in C^{n+1}[a,b]$, allora

$$r(x) = (x - x_0)^{n+1} \, \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad \text{dove} \quad |\xi - x_0| < |x - x_0|.$$

Però la formula di Taylor, utilissimo strumento teorico, raramente è utilizzabile in pratica, perché sfrutta le informazioni che si hanno sulla $f(x)$ nel solo punto x_0. Quindi, oltre a fornire un'approssimazione accettabile solo in un intorno piccolo di x_0, può richiedere il calcolo di un numero eccessivo di termini (come si è visto nell'esempio 4.1). È allora opportuno studiare altri metodi.

2. Il problema dell'approssimazione lineare

Nello spazio vettoriale $\mathcal{F}$ delle funzioni reali di variabile reale si sceglie un insieme di funzioni $\phi_i(x)$, $i = 0, 1, \ldots$, linearmente indipendenti e, fissato un intero n, si considera l'insieme delle combinazioni lineari

$$\sum_{i=0}^{n} \alpha_i \phi_i(x),$$

in cui i coefficienti α_i dovranno essere determinati per ottenere la funzione approssimante $g(x)$.

Nel caso dell'approssimazione la scelta di un *modello* appropriato, cioè delle funzioni $\phi_i(x)$, è ancora più importante di quanto non lo fosse per l'interpolazione, perché se il modello prescelto non si adatta al comportamento della funzione $f(x)$, non è possibile ottenere una buona approssimazione neppure ampliando a dismisura il numero dei coefficienti. Ad esempio, se la funzione $f(x)$ ha un asintoto, un modello polinomiale non è utilizzabile e conviene sceglierne uno razionale, e se $f(x)$ è periodica, una buona approssimazione su un intervallo che contenga più di un periodo può essere ottenuta solo con un modello trigonometrico. Quindi è particolarmente importante determinare un modello che si adatti bene al comportamento della funzione.

Definito il modello, è necessario introdurre una *norma* e quindi una metrica con la quale si possa misurare la distanza fra funzioni.

6.2 Definizione. Sia $\mathcal{G} \subset \mathcal{F}$ un sottospazio vettoriale e sia $f \in \mathcal{G}$. Una funzione da $\mathcal{G}$ in $\mathbf{R}^+$

$$f \to \|f\|$$

che verifica le seguenti proprietà

a) $\|f\| \geq 0$ e $\|f\| = 0$ se e solo se $f = 0$;

b) $\|\alpha f\| = |\alpha|\|f\|$, per ogni $\alpha \in \mathbf{R}$;

c) $\|f + g\| \leq \|f\| + \|g\|$, per ogni $g \in \mathcal{G}$ (*disuguaglianza triangolare*),

è detta *norma*. ∎

La norma è una funzione continua. Infatti per la c) si ha

$$\|f\| = \|f - g + g\| \leq \|f - g\| + \|g\|,$$

da cui

$$\|f\| - \|g\| \leq \|f - g\|,$$

e analogamente per la b)

$$\|g\| - \|f\| \leq \|g - f\| = \|f - g\|;$$

ne segue che

$$\big|\|f\| - \|g\|\big| \leq \|f - g\|. \tag{1}$$

È ora possibile definire il seguente problema.

6.3 Problema dell'approssimazione lineare. Date nello spazio $\mathcal{G}$ una funzione $f(x)$ e $n+1$ funzioni $\phi_i(x)$, $i = 0, \ldots, n$, linearmente indipendenti, e fissata in $\mathcal{G}$ una norma, si determinino $n + 1$ numeri reali $\alpha_0^*, \ldots, \alpha_n^*$ tali che

$$\|f - \sum_{i=0}^{n} \alpha_i^* \phi_i\| = \min_{\boldsymbol{\alpha}} \|f - \sum_{i=0}^{n} \alpha_i \phi_i\|$$

dove $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_n)$. La funzione

$$g_n(x) = \sum_{i=0}^{n} \alpha_i^* \phi_i(x) \tag{2}$$

è detta *funzione di (migliore) approssimazione* rispetto alla norma fissata, la funzione

$$r_n(x) = f(x) - g_n(x)$$

è detta *resto dell'approssimazione* e la quantità

$$\delta_n = \|f - g_n\|$$

è detta *errore assoluto in norma*.

6.4 Teorema.

a) Il problema 6.3 dell'approssimazione lineare ha soluzione;

b) le soluzioni del problema 6.3 formano un insieme convesso;

c) la successione $\{\delta_n\}_{n\in\mathbf{N}}$ *è monotona non crescente, e quindi esiste*

$$\lim_{n\to\infty} \delta_n \geq 0.$$

Dim. a) Poiché la norma è una funzione continua, le funzioni di $\boldsymbol{\alpha}$

$$c(\boldsymbol{\alpha}) = \|\sum_{i=0}^{n} \alpha_i \phi_i\| \quad \text{e} \quad d(\boldsymbol{\alpha}) = \|f - \sum_{i=0}^{n} \alpha_i \phi_i\|$$

sono continue. Quindi indicando con

$$m(\boldsymbol{\alpha}) = \max_{i=0,\ldots,n} |\alpha_i|,$$

la funzione $c(\boldsymbol{\alpha})$ ha minimo sull'insieme compatto

$$S = \{\boldsymbol{\alpha} \in \mathbf{R}^{n+1} \ : \ m(\boldsymbol{\alpha}) = 1\},$$

e chiamato γ tale minimo, che non è nullo per la indipendenza lineare delle $\phi_i(x)$, si ha

$$c(\boldsymbol{\alpha}) = \|\sum_{i=0}^{n} \alpha_i \phi_i\| = m(\boldsymbol{\alpha}) \, \|\sum_{i=0}^{n} \frac{\alpha_i}{m(\boldsymbol{\alpha})} \, \phi_i\| \geq \gamma \, m(\boldsymbol{\alpha}),$$

in quanto $\boldsymbol{\alpha}/m(\boldsymbol{\alpha}) \in S$, e per la (1)

$$d(\boldsymbol{\alpha}) = \|f - \sum_{i=0}^{n} \alpha_i \phi_i\| \geq \big|\|f\| - c(\boldsymbol{\alpha})\,\big| \geq c(\boldsymbol{\alpha}) - \|f\| \geq \gamma \, m(\boldsymbol{\alpha}) - \|f\|.$$

Posto $t = \inf_{\boldsymbol{\alpha}} d(\boldsymbol{\alpha})$, si consideri una costante k tale che

$$k \geq \frac{t + \|f\|}{\gamma}\,.$$

Si ha allora che per $\boldsymbol{\alpha}$, tale che $m(\boldsymbol{\alpha}) > k$, è $d(\boldsymbol{\alpha}) > \gamma k - \|f\| \geq t$ e quindi

$$t = \inf_{\boldsymbol{\alpha}} d(\boldsymbol{\alpha}) = \inf_{m(\boldsymbol{\alpha})\leq k} d(\boldsymbol{\alpha}),$$

ma l'insieme $\{\boldsymbol{\alpha} \in \mathbf{R}^{n+1} \ : \ m(\boldsymbol{\alpha}) \leq k\}$ è un insieme compatto, per cui la funzione $d(\boldsymbol{\alpha})$ ha in esso minimo, e tale minimo è proprio t.

b) Siano $g_n(x)$ e $\overline{g}_n(x)$ due soluzioni del problema 6.3, quindi

$$\|f - g_n\| = \|f - \overline{g}_n\|$$

e sia λ un numero reale tale che $0 \leq \lambda \leq 1$; allora

$$\begin{aligned}\|f - \lambda g_n - (1-\lambda)\overline{g}_n\| &= \|\lambda(f - g_n) + (1-\lambda)(f - \overline{g}_n)\| \\ &\leq \lambda\|f - g_n\| + (1-\lambda)\|f - \overline{g}_n\| = \|f - g_n\|,\end{aligned}$$

perciò ogni combinazione lineare convessa di $g_n(x)$ e $\overline{g}_n(x)$ è funzione di migliore approssimazione.

c) È $$\delta_n = \min_{\boldsymbol{\alpha}} \|f - \sum_{i=0}^{n} \alpha_i \phi_i\| \geq \min_{\boldsymbol{\beta}} \|f - \sum_{i=0}^{n+1} \beta_i \phi_i\|,$$

dove $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_n, \beta_{n+1})$. Quindi $\delta_n \geq \delta_{n+1}$. ∎

Dal punto b) del teorema segue che il problema 6.3 o ha una sola soluzione o ne ha infinite. Si tratta quindi di stabilire quali ipotesi assicurano l'unicità della soluzione. Inoltre, perché la funzione $g_n(x)$ possa essere considerata una buona approssimazione della $f(x)$, dovrebbe accadere che

$$\lim_{n\to\infty} \delta_n = 0,$$

cioè che al crescere di n la successione delle approssimazioni *converga in media* alla funzione $f(x)$, mentre il punto c) del teorema assicura solo l'esistenza del limite, ma non che tale limite sia nullo.

I due problemi, dell'unicità e della convergenza in media, non possono essere affrontati utilizzando una generica norma: è necessario specificare la particolare norma rispetto alla quale si vuole minimizzare (si veda l'esercizio 6.4). Entrambi i problemi possono essere risolti molto semplicemente se si opera in uno *spazio di Hilbert*, in cui la norma viene introdotta per mezzo di un *prodotto scalare*. Poiché in uno spazio di Hilbert si può dare l'espressione esplicita dei coefficienti α_i^* di (2), è opportuno fare una breve digressione per illustrarne gli aspetti fondamentali.

6.5 Definizioni. Sia $\mathcal{G}$ uno spazio vettoriale sul campo $\mathbf{R}$ dei reali. Per ogni coppia ordinata $x, y \in \mathcal{G}$ si definisce *prodotto scalare* e si indica con (x, y) un'applicazione da $\mathcal{G} \times \mathcal{G}$ in $\mathbf{R}$ che gode delle seguenti proprietà

a) $(x, y) = (y, x)$,

b) $(x + z, y) = (x, y) + (z, y), \quad z \in \mathcal{G}$,

c) $\alpha(x, y) = (\alpha x, y), \quad \alpha \in \mathbf{R}$,

d) $(x, x) \geq 0$ e $(x, x) = 0$ se e solo se $x = 0$.

Si definisce poi la norma

$$\|x\| = (x, x)^{1/2}$$

(sono soddisfatte le proprietà a), b) e c) della definizione 6.2).

Sia $\{x_i\}_{i\in\mathbf{N}}$ una successione di elementi di $\mathcal{G}$. Si dice che la successione è *di Cauchy* se per ogni $\epsilon > 0$ esiste un i_0 tale che per ogni $j, i > i_0$ vale

$$\|x_i - x_j\| < \epsilon.$$

Una successione si dice *convergente* ad un limite $x \in \mathcal{G}$ se

$$\lim_{i\to\infty} \|x_i - x\| = 0.$$

Lo spazio $\mathcal{G}$, su cui si è definito un prodotto scalare, è detto *spazio di Hilbert* se ogni successione di Cauchy di elementi di $\mathcal{G}$ è convergente ad un elemento di $\mathcal{G}$. ∎

In uno spazio di Hilbert il problema 6.3 ha un'unica soluzione. Vale infatti il seguente teorema.

6.6 Teorema. *Siano $f \in \mathcal{G}$ e ϕ_i, $i = 0, \ldots, n$ elementi linearmente indipendenti di $\mathcal{G}$. Allora il problema 6.3 ha un'unica soluzione. Il sistema lineare, detto sistema normale,*

$$A\boldsymbol{\alpha} = \mathbf{b}, \tag{3}$$

dove

$$a_{ij} = (\phi_i, \phi_j) \quad \text{e} \quad b_i = (f, \phi_i), \quad i, j = 0, \ldots, n,$$

ha un'unica soluzione $\boldsymbol{\alpha}^* = [\alpha_0^*, \ldots, \alpha_n^*]^T$ e la combinazione lineare

$$g_n = \sum_{i=0}^{n} \alpha_i^* \phi_i$$

è l'unica soluzione del problema 6.3.

Per la dimostrazione si veda [42]. ∎

Una notevole semplificazione nella risoluzione del sistema (3) si ha se gli elementi ϕ_i, $i = 0, \ldots, n$, sono *ortogonali*, cioè se

$$(\phi_i, \phi_j) = 0 \quad \text{per} \quad i \neq j.$$

In tal caso infatti la matrice A risulta diagonale con gli elementi principali $a_{ii} = (\phi_i, \phi_i)$ e la soluzione $\boldsymbol{\alpha}^*$ può essere calcolata direttamente: si ha

$$\alpha_i^* = \frac{b_i}{a_{ii}} = \frac{(f, \phi_i)}{(\phi_i, \phi_i)}, \quad i = 0, \ldots, n \quad (\textit{coefficienti di Fourier}), \tag{4}$$

e quindi

$$g_n = \sum_{i=0}^{n} \frac{(f, \phi_i)}{(\phi_i, \phi_i)} \, \phi_i,$$

da cui, per la linearità del prodotto scalare,

$$\|g_n\|^2 = \sum_{i=0}^{n} \frac{(f, \phi_i)^2}{(\phi_i, \phi_i)}$$

e

$$\delta_n^2 = \|f - g_n\|^2 = \|f\|^2 - \sum_{i=0}^{n} \frac{(f, \phi_i)^2}{(\phi_i, \phi_i)} \, . \tag{5}$$

Dalla (5) segue la *disuguaglianza di Bessel*

$$\|f\|^2 \geq \sum_{i=0}^{n} \frac{(f, \phi_i)^2}{(\phi_i, \phi_i)} \, .$$

Si esamina adesso la convergenza a zero, quando $n \to \infty$, dell'errore assoluto in norma δ_n.

6.7 Definizione. Un insieme $\{\phi_i\}_{i \in \mathbf{N}}$ di elementi linearmente indipendenti di $\mathcal{G}$ si dice *completo* se non esiste in $\mathcal{G}$ alcun elemento x tale che l'insieme $x \cup \{\phi_i\}_{i \in \mathbf{N}}$ sia linearmente indipendente. Un tale insieme viene anche detto *base* di $\mathcal{G}$. ∎

6.8 Teorema. *Sia $\{\phi_i\}_{i \in \mathbf{N}}$ una base ortogonale, cioè un insieme completo di elementi ortogonali di $\mathcal{G}$. Allora per ogni $f \in \mathcal{G}$ risulta*

$$\|f\|^2 = \sum_{i=0}^{\infty} \frac{(f, \phi_i)^2}{(\phi_i, \phi_i)} \quad (\textit{uguaglianza di Parseval}),$$

e quindi $\lim_{n \to \infty} \delta_n = 0.$

Per la dimostrazione si veda [42]. ∎

Nello spazio $C[a, b]$ delle funzioni continue su un intervallo $[a, b]$ si può definire il prodotto scalare

$$(f, h) = \int_a^b f(x) h(x) \, dx,$$

o, più in generale

$$(f, h) = \int_a^b \omega(x) f(x) h(x) \, dx, \tag{6}$$

dove $\omega(x) > 0$ per $x \in (a, b)$ è una funzione continua, detta *funzione peso*, tale che l'integrale

$$\int_a^b \omega(x) f(x)\ dx$$

esista per ogni $f(x) \in C[a, b]$. È facile verificare che la (6) soddisfa la definizione 6.5. Questo prodotto scalare induce sullo spazio $C[a, b]$ la norma

$$\|f\|_2 = \left[\int_a^b \omega(x) f^2(x)\ dx \right]^{1/2},$$

che viene detta *norma 2*. L'approssimazione in norma 2 viene anche detta *approssimazione ai minimi quadrati*.

L'introduzione del peso $\omega(x)$ nella norma 2 ha lo scopo di ottenere approssimazioni più precise nelle zone dell'intervallo $[a, b]$ in cui $\omega(x)$ ha valore maggiore. In tal modo, scegliendo opportunamente il peso $\omega(x)$ si può porre il problema dell'approssimazione lineare 6.3 anche quando l'intervallo di definizione non è limitato.

Lo spazio delle funzioni continue su un intervallo $[a, b]$ con la struttura indotta dal prodotto scalare (6) non è uno spazio di Hilbert, perché non tutte le successioni di Cauchy di funzioni continue convergono a funzioni continue. Tuttavia, se si considera lo spazio $L^2[a, b]$ delle funzioni a quadrato sommabile rispetto all'integrale di Lebesgue, in cui si identificano due funzioni che coincidono quasi ovunque, questo è uno spazio di Hilbert, e lo spazio delle funzioni continue è denso in esso [42]. Poiché in $L^2[a, b]$ si può scegliere una base di polinomi, ne segue che ogni funzione continua può essere approssimata in norma 2 con polinomi. La scelta più immediata è quella della base $\phi_i(x) = x^i$, $i = 0, 1, \ldots$

6.9 Esempio. Siano $[a, b] = [0, 1]$ e $\omega(x) = 1$. Con la base $\phi_i(x) = x^i$, $i = 0, 1, \ldots, n$, gli elementi della matrice A del sistema normale (3) sono

$$a_{ij} = \int_0^1 x^{i+j}\ dx = \frac{1}{i + j + 1}, \quad i, j = 0, \ldots, n, \tag{7}$$

e i termini noti sono

$$b_i = \int_0^1 x^i f(x)\ dx, \quad i = 0, \ldots, n.$$

Il sistema (3) ha un'unica soluzione (si veda l'esercizio 6.6). Se ad esempio si vuole approssimare la funzione $f(x) = \sin \pi x$ con un polinomio di secondo

grado della forma $\alpha_2 x^2 + \alpha_1 x + \alpha_0$, si ha

$$b_0 = \int_0^1 \sin \pi x \; dx = \frac{2}{\pi}\,, \quad b_1 = \int_0^1 x \sin \pi x \; dx = \frac{1}{\pi}\,,$$
$$b_2 = \int_0^1 x^2 \sin \pi x \; dx = \frac{\pi^2 - 4}{\pi^3}\,,$$

e il sistema normale è

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{2}{\pi} \\ \frac{1}{\pi} \\ \frac{\pi^2 - 4}{\pi^3} \end{bmatrix},$$

da cui si ottiene la soluzione

$$\begin{aligned} \boldsymbol{\alpha}^* &= \frac{12}{\pi^3} \left[\pi^2 - 10, 60 - 5\pi^2, 5\pi^2 - 60\right]^T \\ &= [-0.0504655, 4.122512, -4.122512]^T. \end{aligned}$$

Il polinomio $g_2(x)$ di secondo grado di approssimazione ai minimi quadrati di $f(x) = \sin \pi x$ su $[0, 1]$ è allora

$$g_2(x) = -4.122512\, x^2 + 4.122512\, x - 0.0504655.$$

Nella figura 6.1 è riportato il grafico del resto $r_2(x) = f(x) - g_2(x)$.

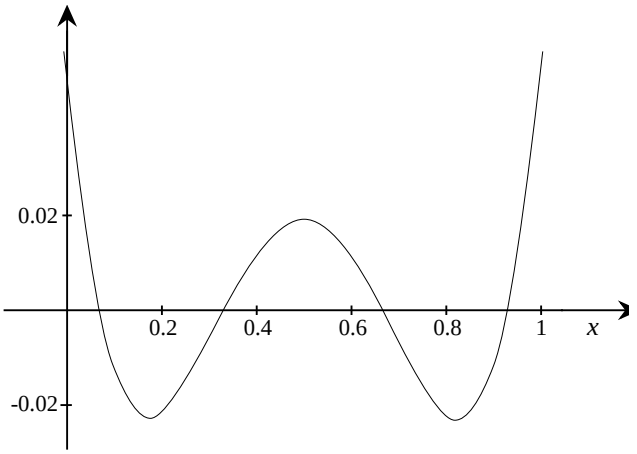


Fig. 6.1 - Resto nell'approssimazione ai minimi quadrati della funzione $f(x) = \sin \pi x$.

La matrice (7), detta matrice di *Hilbert*, risulta fortemente malcondizionata anche per valori bassi di n (si veda la tabella dei numeri di condizionamento delle matrici di Hilbert nell'esempio 4.2 di [7]). Inoltre, se si

vuole aumentare il grado n del polinomio di approssimazione, è necessario risolvere un nuovo sistema (3), in quanto non è possibile sfruttare alcuno dei coefficienti calcolati per valori inferiori di n. Per questi motivi la base $\phi_i(x) = x^i$ non viene generalmente usata.

Il calcolo di $g_2(x)$ è molto più agevole se si usano i polinomi

$$\phi_0(x) = 1, \quad \phi_1(x) = 2x - 1, \quad \phi_2(x) = 6x^2 - 6x + 1, \quad \dots$$

che sono ortogonali sull'intervallo $[0, 1]$ rispetto al peso $\omega(x) = 1$. Per la (4) si ha

$$\alpha_0^* = \frac{\int_0^1 f(x)\phi_0(x)\,dx}{\int_0^1 \phi_0^2(x)\,dx} = \int_0^1 \sin \pi x\; dx = \frac{2}{\pi}\,,$$

$$\alpha_1^* = \frac{\int_0^1 f(x)\phi_1(x)\,dx}{\int_0^1 \phi_1^2(x)\,dx} = 3\int_0^1 (2x-1)\sin \pi x\; dx = 0,$$

$$\alpha_2^* = \frac{\int_0^1 f(x)\phi_2(x)\,dx}{\int_0^1 \phi_2^2(x)\,dx} = 5\int_0^1 (6x^2-6x+1)\sin \pi x\; dx = \frac{5}{\pi^3}\,(2\pi^2 - 24),$$

da cui si ottiene

$$g_2(x) = \alpha_2^*\phi_2(x) + \alpha_0^*\phi_0(x) = \frac{60}{\pi^3}\,(\pi^2 - 12)(x^2 - x) + \frac{12}{\pi^3}\,(\pi^2 - 10).$$

Risulta inoltre

$$\delta_2^2 = \|f\|_2^2 - \sum_{i=0}^{2} \frac{(f, \phi_i)^2}{(\phi_i, \phi_i)} = \frac{1}{2} - \frac{4}{\pi^2} - 5\,\frac{(2\pi^2 - 24)^2}{\pi^6} \approx 0.298\ 10^{-3}. \quad \blacksquare$$

Poiché per il teorema 6.8 è $\lim_{n\to\infty} \delta_n = 0$, è possibile, fissato ϵ, determinare n in modo che il polinomio approssimante ai minimi quadrati $g_n(x)$ soddisfi la condizione $\delta_n < \epsilon$. Questo però non garantisce che $|r_n(x)| < \epsilon$ per ogni $x \in [a, b]$. Cioè con la norma 2 si ottiene un'approssimazione in media sull'intervallo, che può essere buona in certi punti e meno in altri. Se invece è richiesto di determinare un'approssimazione che soddisfi la condizione $|r_n(x)| < \epsilon$ in ogni punto $x \in [a, b]$, come generalmente avviene

quando si calcolano funzioni matematiche con un calcolatore, si deve usare la *norma* ∞, definita da

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|.$$

L'approssimazione in norma ∞ viene detta anche *approssimazione minimax o di Chebyshev*. Tuttavia, poiché non esistono, come si vedrà, metodi espliciti per il calcolo dell'approssimazione in norma ∞, spesso si preferisce operare con la norma 2, che consente di calcolare l'approssimazione esplicitamente.

6.10 Esempio. Si vogliono determinare i polinomi di primo grado di migliore approssimazione alla funzione $f(x) = x^3$ sull'intervallo $[-1, 1]$ rispetto alle norme 2 e ∞. È intuitivo che in entrambi i casi la retta di migliore approssimazione debba tagliare il grafico di x^3 in tre punti dell'intervallo (cioè deve essere del tipo riportato nella figura 6.2), e quindi il suo primo coefficiente debba essere positivo e minore di 1.

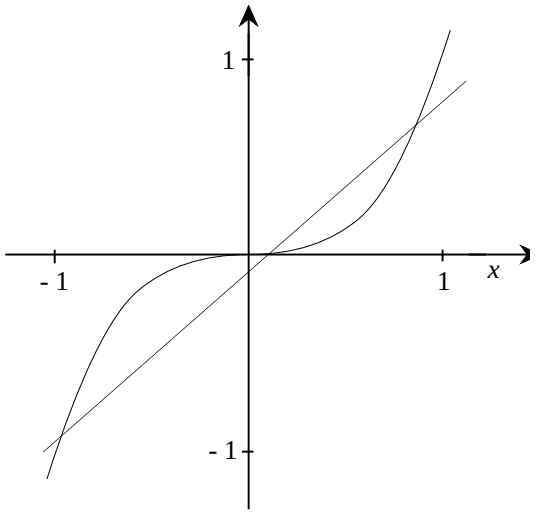


Fig. 6.2 - Approssimazione di x^3 con una retta.

a) norma 2: si determinano i coefficienti α_1^* e α_0^* tali che

$$\|x^3 - \alpha_1^* x - \alpha_0^*\|_2 = \min_{(\alpha_0,\alpha_1)} \|x^3 - \alpha_1 x - \alpha_0\|_2.$$

Con il peso $\omega(x) = 1$ i polinomi $\phi_0(x) = 1$ e $\phi_1(x) = x$ sono ortogonali sull'intervallo $[-1, 1]$. Risulta quindi

$$\alpha_0^* = \frac{\int_{-1}^{1} f(x)\phi_0(x)\,dx}{\int_{-1}^{1} \phi_0^2(x)\,dx} = 0, \quad \alpha_1^* = \frac{\int_{-1}^{1} f(x)\phi_1(x)\,dx}{\int_{-1}^{1} \phi_1^2(x)\,dx} = \frac{3}{5},$$

e il polinomio di approssimazione ai minimi quadrati di primo grado è

$$g^{(2)}(x) = \frac{3}{5}\, x.$$

Inoltre è

$$\delta_2 = \frac{2}{5}\sqrt{\frac{2}{7}}\,, \quad \text{e} \quad \max_{x\in[-1,1]} |x^3 - g^{(2)}(x)| = \frac{2}{5}\,.$$

b) norma ∞: si determinano i coefficienti α_1^* e α_0^* tali che

$$\|x^3 - \alpha_1^* x - \alpha_0^*\|_\infty = \min_{(\alpha_0,\alpha_1)} \|x^3 - \alpha_1 x - \alpha_0\|_\infty,$$

cioè il punto di minimo di

$$m(\alpha_0,\alpha_1) = \max_{x\in[-1,1]} |x^3 - \alpha_1 x - \alpha_0|.$$

$m(\alpha_0,\alpha_1)$ va ricercato fra i valori che $|x^3-\alpha_1 x-\alpha_0|$ assume nei due estremi -1 e 1 e i valori che assume nei due punti stazionari di $x^3 - \alpha_1 x - \alpha_0$, che sono $\pm\sqrt{\alpha_1/3}$ interni a $[-1,1]$. Perciò

$$m(\alpha_0,\alpha_1) = \max\big\{\beta+\alpha_0,\ \beta-\alpha_0,\ \gamma-\alpha_0,\ \gamma+\alpha_0\big\},$$

dove

$$\beta = 1-\alpha_1, \quad \gamma = -\sqrt{\frac{\alpha_1^3}{3^3}} + \alpha_1\sqrt{\frac{\alpha_1}{3}}\,.$$

Nel punto di minimo di $m(\alpha_0,\alpha_1)$ è necessariamente $\alpha_0^* = 0$. Perciò

$$\min_{(\alpha_0,\alpha_1)} m(\alpha_0,\alpha_1) = \min_{\alpha_1} \max\{\beta,\gamma\},$$

e quindi α_1^* è il valore per cui $\beta = \gamma$, cioè α_1^* è la soluzione dell'equazione

$$2\alpha_1\sqrt{\alpha_1} = (1-\alpha_1)3\sqrt{3},$$

da cui $\alpha_1^* = \dfrac{3}{4}$. Il polinomio di approssimazione minimax è quindi

$$g^{(\infty)}(x) = \frac{3}{4}\, x,$$

e risulta

$$\max_{x\in[-1,1]} |x^3 - g^{(\infty)}(x)| = \frac{1}{4}\,.$$

$g^{(\infty)}(x)$ è un'approssimazione di $f(x)$ migliore su tutto l'intervallo, mentre $g^{(2)}(x)$ è un'approssimazione migliore nella zona centrale dell'intervallo e peggiore ai due estremi -1 e 1. ∎

L'approssimazione in norma 2 verrà esaminata nel paragrafo 4, quella in norma ∞ verrà esaminata nel paragrafo 5. Prima di trattare l'approssimazione in norma 2 è opportuno introdurre nel paragrafo 3 i polinomi ortogonali e studiarne le principali proprietà.

3. Polinomi ortogonali

In questo paragrafo e nel successivo, in cui si tratterà di polinomi ortogonali, si prenderanno in considerazione anche funzioni definite su intervalli non limitati, cioè intervalli della retta *reale estesa* $\mathbf{R}\cup\{-\infty,+\infty\}$. Pertanto con la notazione $[a,b]$ potranno essere indicati anche gli intervalli $[-\infty,+\infty]$ e $[0,+\infty]$.

Fissata una funzione peso $\omega(x)$, si considera un insieme $\{p_i(x)\}_{i\in\mathbf{N}}$ di polinomi, in cui $p_i(x)$ è di grado i, *ortogonali sull'intervallo* $[a,b]$ rispetto al prodotto scalare (6), cioè

$$(p_i,p_j)=\int_a^b \omega(x)\,p_i(x)p_j(x)\,dx=0 \quad \text{per} \quad i\neq j.$$

Indicata con

$$h_i=(p_i,p_i) \tag{8}$$

la *costante di normalizzazione*, dall'insieme $\{p_i(x)\}_{i\in\mathbf{N}}$ si ottiene l'insieme di polinomi *ortonormali*

$$\frac{1}{\sqrt{h_i}}\,p_i(x)\,.$$

L'insieme $\{p_i(x)\}_{i\in\mathbf{N}}$ costituisce una base dello spazio dei polinomi (si veda l'esercizio 6.8), quindi ogni polinomio $q(x)$ di grado n può essere espresso come combinazione lineare dei polinomi $p_i(x)$, $i=0,\dots,n$.

6.11 Teorema. *Sia $q(x)$ un polinomio di grado n. Allora*

$$(p_i,q)=0 \quad \textit{per} \quad i>n.$$

Dim. Posto

$$q(x)=\sum_{j=0}^{n}\gamma_j p_j(x),$$

risulta

$$(p_i,q)=\sum_{j=0}^{n}\gamma_j(p_i,p_j)=0 \quad \text{per} \quad i>n. \qquad \blacksquare$$

6.12 Teorema. *Gli zeri di $p_i(x)$ sono tutti reali, semplici e sono interni ad $[a,b]$.*

Dim. Indicati con $x_1,\dots,x_j$ gli zeri reali e distinti di $p_i(x)$ che cadono in (a,b), si suppone per assurdo che $j<i$. Si suppone inoltre che k zeri, fra quelli che cadono in (a,b) abbiano molteplicità dispari (senza violare la

generalità, si può supporre che siano i primi k). Si considera poi il polinomio di grado k

$$q(x) = \begin{cases} 1 & \text{se } k = 0, \\ \prod_{r=1}^{k} (x - x_r) & \text{se } k > 0. \end{cases}$$

Il polinomio $p_i(x)q(x)$ ha tutti gli zeri in (a, b) di molteplicità pari, e pertanto non cambia segno in (a, b). Allora

$$\int_a^b \omega(x)\, p_i(x)q(x)\, dx \neq 0,$$

e questo, per il teorema 6.11, è assurdo, perché $q(x)$ ha grado minore di i.

∎

6.13 Teorema. *Tre polinomi ortogonali consecutivi soddisfano una relazione ricorrente a tre termini della forma*

$$p_{i+1}(x) = (A_i x + B_i)p_i(x) - C_i p_{i-1}(x), \quad i \geq 1, \tag{9}$$

in cui i coefficienti A_i e C_i sono non nulli per ogni i. Più precisamente, indicati con a_i e b_i rispettivamente i coefficienti dei termini di grado i e $i - 1$ in $p_i(x)$, è

$$A_i = \frac{a_{i+1}}{a_i}, \quad B_i = \frac{a_{i+1}}{a_i}\left(\frac{b_{i+1}}{a_{i+1}} - \frac{b_i}{a_i}\right), \quad C_i = \frac{a_{i+1}a_{i-1}h_i}{a_i^2 h_{i-1}},$$

in cui h_i è definita in (8).

Dim. Il polinomio

$$q(x) = p_{i+1}(x) - A_i x p_i(x), \quad \text{con} \quad A_i = \frac{a_{i+1}}{a_i}, \tag{10}$$

ha grado minore od uguale ad i e quindi può essere scritto come

$$q(x) = \sum_{j=0}^{i} \gamma_j p_j(x), \tag{11}$$

da cui per $n \leq i$ risulta

$$(q, p_n) = \sum_{j=0}^{i} \gamma_j (p_j, p_n) = \gamma_n (p_n, p_n).$$

Dalla (10) si ha per $n \leq i-2$

$$\gamma_n(p_n, p_n) = (q, p_n) = (p_{i+1}, p_n) - A_i(xp_i, p_n) = 0,$$

in quanto è $(p_{i+1}, p_n) = 0$ per l'ortogonalità e $(xp_i, p_n) = (p_i, xp_n) = 0$ per il teorema 6.11 perché $xp_n(x)$ è un polinomio di grado minore del grado di $p_i(x)$. Quindi nella (11) resta

$$q(x) = \gamma_i p_i(x) + \gamma_{i-1} p_{i-1}(x),$$

e sostituendo nella (10) si ha

$$p_{i+1}(x) = (A_i x + \gamma_i) p_i(x) + \gamma_{i-1} p_{i-1}(x). \tag{12}$$

Per determinare l'espressione di $C_i = -\gamma_{i-1}$ dalla (12) si ha

$$\gamma_{i-1}(p_{i-1}, p_{i-1}) = -A_i(xp_i, p_{i-1}) = -A_i(p_i, xp_{i-1}). \tag{13}$$

Il polinomio $xp_{i-1}(x)$ ha grado i e primo coefficiente uguale ad a_{i-1}, perciò può essere scritto come

$$xp_{i-1}(x) = \frac{a_{i-1}}{a_i}\, p_i(x) + r(x),$$

dove $r(x)$ è un polinomio di grado al più $i-1$. Allora si ha

$$(p_i, xp_{i-1}) = \frac{a_{i-1}}{a_i}\, (p_i, p_i) + (p_i, r) = \frac{a_{i-1}}{a_i}\, (p_i, p_i) = \frac{a_{i-1}}{a_i}\, h_i,$$

in quanto $(p_i, r) = 0$, da cui, sostituendo nella (13), risulta

$$\gamma_{i-1} h_{i-1} = -A_i \frac{a_{i-1}}{a_i}\, h_i = -\frac{a_{i+1} a_{i-1}}{a_i^2}\, h_i,$$

e

$$C_i = -\gamma_{i-1} = \frac{a_{i+1} a_{i-1} h_i}{a_i^2 h_{i-1}}\,.$$

Per determinare l'espressione di $B_i = \gamma_i$, si ha dalla (12)

$$\gamma_i p_i(x) = p_{i+1}(x) - A_i x p_i(x) + C_i p_{i-1}(x),$$

per cui, considerando i coefficienti dei termini di grado i, risulta

$$\gamma_i a_i = b_{i+1} - A_i b_i$$

e

$$B_i = \frac{b_{i+1} - A_i b_i}{a_i} = \frac{1}{a_i}\left(b_{i+1} - \frac{a_{i+1}}{a_i}\, b_i\right) = \frac{a_{i+1}}{a_i}\left(\frac{b_{i+1}}{a_{i+1}} - \frac{b_i}{a_i}\right). \qquad \blacksquare$$

Dalla (9) si ricava un'altra relazione che verrà sfruttata nello studio di un'importante classe di formule di integrazione approssimata.

6.14 Teorema. *Con le notazioni del teorema 6.13, per $\xi \in [a,b]$, vale la seguente formula di Christoffel-Darboux*

$$(x-\xi)\sum_{i=0}^{n}\frac{1}{h_i}\,p_i(x)p_i(\xi)=\frac{a_n}{a_{n+1}h_n}\,\big[p_{n+1}(x)p_n(\xi)-p_{n+1}(\xi)p_n(x)\big].$$

Dim. Per $n=0$, poiché $p_0(x)=a_0$ e $p_1(x)=a_1x+b_1$, si ha

$$\frac{a_0}{a_1h_0}\,\big[p_1(x)p_0(\xi)-p_1(\xi)p_0(x)\big]=\frac{a_0^2}{h_0}\,(x-\xi)=(x-\xi)\frac{1}{h_0}\,p_0(x)p_0(\xi)\quad(14)$$

e quindi la tesi. Per $n\geq 1$, moltiplicando la (9) per $p_i(\xi)$ si ottiene

$$(A_ix+B_i)p_i(x)p_i(\xi)=p_{i+1}(x)p_i(\xi)+C_ip_{i-1}(x)p_i(\xi).\qquad(15)$$

Scambiando fra loro le due variabili x e ξ, la (15) diventa

$$(A_i\xi+B_i)p_i(\xi)p_i(x)=p_{i+1}(\xi)p_i(x)+C_ip_{i-1}(\xi)p_i(x);\qquad(16)$$

sottraendo la (16) dalla (15) e dividendo per A_ih_i, si ha

$$\begin{aligned}(x-\xi)\frac{1}{h_i}\,p_i(x)p_i(\xi)=&\frac{a_i}{a_{i+1}h_i}\,\big[p_{i+1}(x)p_i(\xi)-p_{i+1}(\xi)p_i(x)\big]\\&+\frac{a_{i-1}}{a_ih_{i-1}}\,\big[p_{i-1}(x)p_i(\xi)-p_{i-1}(\xi)p_i(x)\big].\end{aligned}$$

Sommando per $i=1,2,\dots,n$ e tenendo conto del fatto che i termini al secondo membro si elidono a due a due, si ha

$$\begin{aligned}(x-\xi)\sum_{i=1}^{n}\frac{1}{h_i}\,p_i(x)p_i(\xi)=&\frac{a_n}{a_{n+1}h_n}\,\big[p_{n+1}(x)p_n(\xi)-p_{n+1}(\xi)p_n(x)\big]\\&+\frac{a_0}{a_1h_0}\,\big[p_0(x)p_1(\xi)-p_0(\xi)p_1(x)\big],\end{aligned}$$

e portando al primo membro il secondo termine del secondo membro, per la (14) segue la tesi. ∎

6.15 Teorema. *Per i intero, $i\geq 1$, sia $s(x)\in C^i[a,b]$ tale che per le sue derivate di ordine k, con $k=0,\dots,i-1$, valga*

$$s^{(k)}(a)=s^{(k)}(b)=0.$$

La funzione $t(x) = \dfrac{s^{(i)}(x)}{\omega(x)}$ *è ortogonale sull'intervallo* $[a, b]$*, rispetto al prodotto scalare (6), ad ogni polinomio di grado minore di* i*.*

Dim. Sia $q(x)$ un qualunque polinomio di grado $k \leq i-1$. Si ha

$$(t, q) = \int_a^b \omega(x) q(x) \, \frac{s^{(i)}(x)}{\omega(x)} \, dx = \int_a^b q(x) s^{(i)}(x) \, dx,$$

e, integrando per parti, si ha

$$(t, q) = \Big[q(x) s^{(i-1)}(x) \Big]_a^b - \int_a^b q'(x) s^{(i-1)}(x) \, dx.$$

Per l'ipotesi fatta, il primo termine è nullo. Integrando i volte per parti, poiché il primo termine è sempre nullo, si ha

$$(t, q) = (-1)^i \int_a^b q^{(i)}(x) s(x) \, dx.$$

Poiché $q(x)$ ha grado inferiore ad i, è $q^{(i)}(x) = 0$ e quindi $(t, q) = 0$. ∎

I polinomi ortogonali che si esamineranno sono i cosiddetti *polinomi ortogonali classici*, che possono essere rappresentati per mezzo della *formula di Rodrigues*

$$p_i(x) = \frac{\beta_i}{\omega(x)} \, \frac{d^i}{dx^i} \, s_i(x), \quad i = 0, 1, \ldots, \tag{17}$$

dove per ogni i, β_i è costante e $s_i(x) \in C^i[a, b]$ è tale che la funzione $\dfrac{1}{\omega(x)} \, \dfrac{d^i}{dx^i} \, s_i(x)$ sia un polinomio di grado i. Secondo la notazione consueta si indicherà $\dfrac{d^k}{dx^k} \, s_i(x) = s_i^{(k)}(x)$.

Se per ogni i le funzioni $s_i(x)$ soddisfano le ipotesi del teorema 6.15, l'insieme dei polinomi $p_i(x)$ definiti nella (17) è un insieme di polinomi ortogonali sull'intervallo $[a, b]$ rispetto al peso $\omega(x)$. Indicato con a_i il coefficiente del termine di grado i di $p_i(x)$, risulta

$$p_i(x) = a_i x^i + q(x),$$

dove $q(x)$ è un polinomio di grado minore o uguale a $i-1$, per cui dalla (8) è

$$h_i = (a_i x^i + q(x), p_i) = a_i (x^i, p_i) + (q, p_i).$$

Per il teorema 6.15, è $(q, p_i) = 0$ e quindi

$$h_i = a_i(x^i, p_i) = a_i \int_a^b \omega(x) x^i \frac{\beta_i s_i^{(i)}(x)}{\omega(x)}\, dx = a_i \beta_i \int_a^b x^i s_i^{(i)}(x)\, dx,$$

e, integrando per parti, si ha

$$\begin{aligned} h_i &= a_i \beta_i \Big[x^i s_i^{(i-1)}(x) \Big]_a^b - a_i \beta_i\, i \int_a^b x^{i-1} s_i^{(i-1)}(x)\, dx \\ &= -a_i \beta_i\, i \int_a^b x^{i-1} s_i^{(i-1)}(x)\, dx = \ldots = (-1)^i a_i \beta_i\, i! \int_a^b s_i(x)\, dx. \end{aligned} \quad (18)$$

La (18) dà l'espressione della costante di normalizzazione dei polinomi (17).

I polinomi ortogonali che si studieranno sono i seguenti

$[a, b]$	$\omega(x)$	$s_i(x)$	nome
$[-1, 1]$	1	$(1 - x^2)^i$	pol. di Legendre
$[-1, 1]$	$(1 - x^2)^{-1/2}$	$(1 - x^2)^{i-1/2}$	pol. di Chebyshev di 1^a specie
$[-1, 1]$	$(1 - x^2)^{1/2}$	$(1 - x^2)^{i+1/2}$	pol. di Chebyshev di 2^a specie
$[0, \infty]$	e^{-x}	$\mathrm{e}^{-x} x^i$	pol. di Laguerre
$[-\infty, \infty]$	e^{-x^2}	e^{-x^2}	pol. di Hermite

I polinomi di Legendre e quelli di Chebyshev possono essere considerati come un caso particolare dei *polinomi ultrasferici*, ortogonali sull'intervallo $[-1, 1]$ rispetto al peso

$$\omega(x) = (1 - x^2)^\alpha, \quad \alpha > -1.$$

Per i polinomi ultrasferici e, separatamente, per quelli di Laguerre e di Hermite, si vedrà che la funzione $s_i(x)$ è tale che $\dfrac{s_i^{(i)}(x)}{\omega(x)}$ è un polinomio e che sono verificate le ipotesi del teorema 6.15. Si determineranno inoltre le costanti h_i di normalizzazione date dalla (18) e le relazioni ricorrenti a tre termini (9).

6.16 Teorema. *Posto* $\omega(x) = (1-x^2)^\alpha$ *e* $s_i(x) = (1-x^2)^{\alpha+i}$, *con* $\alpha > -1$, *la funzione*

$$f_{(i,\alpha)}(x) = \frac{s_i^{(i)}(x)}{\omega(x)} = \frac{1}{(1-x^2)^\alpha}\ \frac{d^i}{dx^i}\ [(1-x^2)^{\alpha+i}], \tag{19}$$

è un polinomio in x *di grado* i, *la cui espressione esplicita è*

$$f_{(i,\alpha)}(x) = \sum_{j=0}^{i}(-1)^j \binom{i}{j}\ \frac{[\Gamma(\alpha+i+1)]^2}{\Gamma(\alpha+j+1)\Gamma(\alpha+i-j+1)}\ (1-x)^{i-j}(1+x)^j, \tag{20}$$

dove la funzione $\Gamma(x)$ *è definita nel paragrafo 4 del capitolo 4. Inoltre per* $i \geq 1$ *è*

$$s_i^{(k)}(-1) = s_i^{(k)}(1) = 0, \quad \textit{per} \quad k = 0, \dots, i-1. \tag{21}$$

Dim. Per $i = 0$ si ha

$$f_{(0,\alpha)}(x) = \frac{1}{(1-x^2)^\alpha}\ (1-x^2)^\alpha = 1.$$

Per $i > 0$, posto $z = \alpha + i$, dalla formula di Leibniz si ha:

$$\begin{aligned}\frac{d^k}{dx^k}\ \left[(1-x^2)^z\right] &= \frac{d^k}{dx^k}\ [(1-x)^z(1+x)^z] \\ &= \sum_{j=0}^{k}\binom{k}{j}\frac{d^j}{dx^j}\ (1-x)^z\frac{d^{k-j}}{dx^{k-j}}\ (1+x)^z.\end{aligned}$$

Dall'esercizio 4.21 a) segue che

$$\begin{aligned}&\frac{d^k}{dx^k}\ [(1-x^2)^z] \\ &= \sum_{j=0}^{k}(-1)^j\binom{k}{j}\ \frac{\Gamma(z+1)}{\Gamma(z-j+1)}\ (1-x)^{z-j}\ \frac{\Gamma(z+1)}{\Gamma(z-k+j+1)}\ (1+x)^{z-k+j} \\ &= (1-x^2)^{z-k}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\ \frac{[\Gamma(z+1)]^2\ (1-x)^{k-j}(1+x)^j}{\Gamma(z-j+1)\Gamma(z-k+j+1)}.\end{aligned} \tag{22}$$

Per $k = 0, \dots, i-1$, è $z - k \geq \alpha + 1 > 0$, e quindi vale la (21). Per $k = i$ risulta

$$\frac{(1-x^2)^{z-i}}{\omega(x)} = \frac{1}{(1-x^2)^\alpha}\ (1-x^2)^{z-i} = 1.$$

Dalla (22) segue quindi che la funzione $f_{(i,\alpha)}(x)$, data nella (19), è un polinomio di grado i, la cui espressione esplicita è la (20). ∎

I polinomi ultrasferici sono dati da

$$p_i^{(\alpha)}(x) = \beta_{(i,\alpha)} f_{(i,\alpha)}(x)$$

dove $\beta_{(i,\alpha)}$ sono delle costanti. Qui si studieranno solo i polinomi di Legendre e di Chebyshev, corrispondenti ai casi particolari $\alpha = 0$, $\alpha = -1/2$ e $\alpha = 1/2$.

Dalla (21) segue, per il teorema 6.15, che i polinomi ultrasferici sono ortogonali sull'intervallo $[-1, 1]$ rispetto al peso $\omega(x) = (1 - x^2)^\alpha$. Dalla (20) si ricava il coefficiente a_i del termine di grado massimo del polinomio $p_i^{(\alpha)}(x)$. Si ha infatti:

$$a_i = (-1)^i \beta_{(i,\alpha)} \sum_{j=0}^{i} \binom{i}{j} \frac{[\Gamma(\alpha + i + 1)]^2}{\Gamma(\alpha + j + 1)\Gamma(\alpha + i - j + 1)},$$

e, per quanto dimostrato nell'esercizio 4.21 b), è

$$a_i = (-1)^i \beta_{(i,\alpha)} \frac{\Gamma(2\alpha + 2i + 1)}{\Gamma(2\alpha + i + 1)}. \tag{23}$$

Per la (18) la costante di normalizzazione è

$$\begin{aligned} h_i^{(\alpha)} &= (p_i^{(\alpha)}, p_i^{(\alpha)}) = (-1)^i a_i \beta_{(i,\alpha)} i! \int_{-1}^{1} (1 - x^2)^{\alpha+i}\, dx \\ &= (-1)^i a_i \beta_{(i,\alpha)} i!\, 2^{2\alpha+2i+1} \frac{[\Gamma(\alpha + i + 1)]^2}{\Gamma(2\alpha + 2i + 2)} \end{aligned} \tag{24}$$

(per il calcolo dell'integrale si veda l'esercizio 4.22 a)). Le relazioni a tre termini verranno determinate caso per caso per i singoli polinomi.

È immediato verificare che

$$f_{(i,\alpha)}(-x) = (-1)^i f_{(i,\alpha)}(x),$$

cioè i polinomi ultrasferici sono, a seconda che il grado sia pari o dispari, funzioni pari o dispari: se i è pari, tutti i coefficienti dei termini di grado dispari sono nulli, mentre se i è dispari, tutti i coefficienti dei termini di grado pari sono nulli. Quindi i coefficienti b_i dei termini di grado $i - 1$ dei polinomi ultrasferici di grado i sono nulli.

3.1 Polinomi di Legendre

I polinomi di Legendre $\{P_i(x)\}_{i\in\mathbf{N}}$ sono i polinomi ortogonali sull'intervallo $[-1,1]$ rispetto al peso $\omega(x)\equiv 1$, che si ottengono dai polinomi ultrasferici ponendo

$$\alpha = 0 \quad \text{e} \quad \beta_i = \beta_{(i,0)} = \frac{(-1)^i}{2^i i!}\,, \quad \text{cioè} \quad P_i(x) = \beta_{(i,0)} f_{(i,0)}(x),$$

la cui corrispondente formula di Rodrigues per la (17) è

$$P_i(x) = \frac{(-1)^i}{2^i i!}\,\frac{d^i}{dx^i}\,[(1-x^2)^i].$$

Dalla (23) si ottiene l'espressione del primo coefficiente

$$a_i = \frac{1}{2^i i!}\,\frac{\Gamma(2i+1)}{\Gamma(i+1)} = \frac{(2i)!}{2^i (i!)^2}\,,$$

e quindi, per la (24), la costante di normalizzazione è

$$h_i = h_i^{(0)} = \frac{2(2i)!}{(i!)^2}\,\frac{[\Gamma(i+1)]^2}{\Gamma(2i+2)} = \frac{2}{2i+1}\,.$$

Poiché $b_i = 0$, per il teorema 6.13 si ha:

$$A_i = \frac{2i+1}{i+1}\,, \quad B_i = 0, \quad C_i = \frac{i}{i+1}\,.$$

La relazione ricorrente a tre termini (9) è allora:

$$(i+1)P_{i+1}(x) = (2i+1)xP_i(x) - iP_{i-1}(x),$$

e, a partire dai primi due polinomi $P_0(x) = 1$, $P_1(x) = x$, consente di costruire facilmente i polinomi di Legendre.

6.17 Esempio. I primi 7 polinomi di Legendre sono

$$\begin{aligned}
P_0(x) &= 1\\
P_1(x) &= x\\
P_2(x) &= \frac{1}{2}\,(3x^2-1)\\
P_3(x) &= \frac{1}{2}\,(5x^3-3x)\\
P_4(x) &= \frac{1}{8}\,(35x^4-30x^2+3)\\
P_5(x) &= \frac{1}{8}\,(63x^5-70x^3+15x)\\
P_6(x) &= \frac{1}{16}\,(231x^6-315x^4+105x^2-5).
\end{aligned}$$

I grafici dei polinomi di Legendre $P_i(x)$, per $i = 1, \ldots, 5$, sono rappresentati nella figura 6.3. ∎

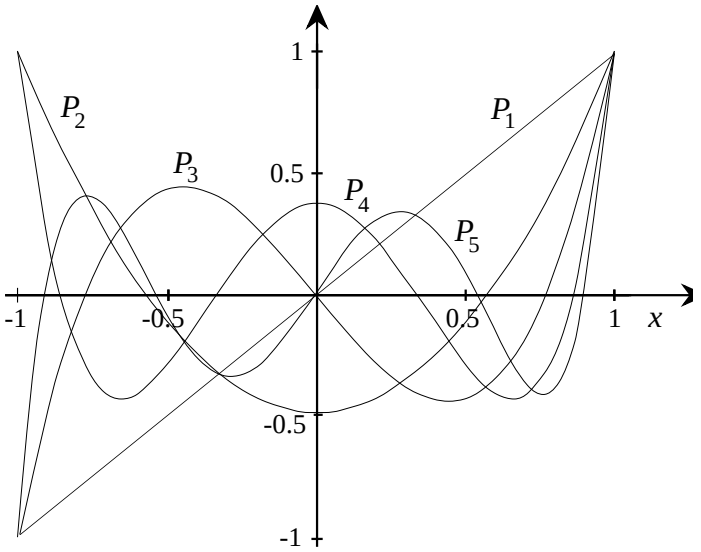


Fig. 6.3 - Grafici dei polinomi di Legendre $P_i(x),\ i = 1, \ldots, 5$.

3.2 Polinomi di Chebyshev di 1^a specie

I polinomi di Chebyshev di 1^a specie $\{T_i(x)\}_{i\in\mathbf{N}}$ sono i polinomi ortogonali sull'intervallo $[-1, 1]$ rispetto al peso $\omega(x) = (1 - x^2)^{-1/2}$, che si ottengono dai polinomi ultrasferici ponendo

$$\alpha = -\frac{1}{2} \text{ e } \beta_i = \beta_{(i,-1/2)} = \frac{(-1)^i 2^i i!}{(2i)!}\,, \text{ cioè } T_i(x) = \beta_{(i,-1/2)} f_{(i,-1/2)}(x)$$

(per $i = 0$ è $\beta_{(0,-1/2)} = 1$). Dalla (17) si ha la formula di Rodrigues per i polinomi di Chebyshev di 1^a specie

$$T_i(x) = \frac{(-1)^i 2^i i!}{(2i)!}\,(1 - x^2)^{1/2}\,\frac{d^i}{dx^i}\,\big[(1 - x^2)^{i-1/2}\big].$$

Quindi risulta $a_0 = 1$, mentre per $i \geq 1$ dalla (23) si ottiene

$$a_i = \frac{2^i i!}{(2i)!}\,\frac{\Gamma(2i)}{\Gamma(i)} = 2^{i-1}.$$

Dalla (24), per $i = 0$ la costante di normalizzazione è

$$h_0 = h_0^{(-1/2)} = \big[\Gamma\big(\frac{1}{2}\big)\big]^2 = \pi,$$

mentre per $i \geq 1$ è

$$h_i = h_i^{(-1/2)} = \frac{(i!)^2\, 2^{4i-1}}{(2i)!}\, \frac{\left[\Gamma\left(i+\frac{1}{2}\right)\right]^2}{\Gamma(2i+1)} = \frac{(i!)^2\, 2^{4i-1}}{(2i)!}\, \frac{[(2i)!]^2\pi}{2^{4i}(i!)^2(2i)!} = \frac{\pi}{2}$$

(si veda l'esercizio 4.22 c)). Poiché $b_i = 0$, per il teorema 6.13 si ha:

$$A_i = 2, \quad B_i = 0, \quad C_i = 1.$$

La relazione ricorrente a tre termini (9) è allora:

$$T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x), \tag{25}$$

e, a partire dai primi due polinomi $T_0(x) = 1$, $T_1(x) = x$, consente di costruire facilmente i polinomi di Chebyshev di 1^a specie.

6.18 Esempio. I primi 7 polinomi di Chebyshev di 1^a specie sono

$$\begin{aligned}
T_0(x) &= 1\\
T_1(x) &= x\\
T_2(x) &= 2x^2 - 1\\
T_3(x) &= 4x^3 - 3x\\
T_4(x) &= 8x^4 - 8x^2 + 1\\
T_5(x) &= 16x^5 - 20x^3 + 5x\\
T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1.
\end{aligned}$$

I grafici dei polinomi di Chebyshev di 1^a specie $T_i(x)$, per $i = 1, \ldots, 5$, sono rappresentati nella figura 6.4. ∎

Dei polinomi di Chebyshev di 1^a specie si può dare un'espressione esplicita molto semplice. Ponendo $x = \cos\theta$, $0 \leq \theta \leq \pi$, si ha

$$T_i(\cos\theta) = \cos i\theta, \quad i = 0, 1, \ldots \tag{26}$$

come si può vedere per induzione applicando la (25); infatti

$$\begin{aligned}
T_{i+1}(\cos\theta) &= 2\cos\theta\cos i\theta - \cos(i-1)\theta\\
&= 2\cos\theta\cos i\theta - \cos\theta\cos i\theta - \sin\theta\sin i\theta\\
&= \cos\theta\cos i\theta - \sin\theta\sin i\theta = \cos(i+1)\theta.
\end{aligned}$$

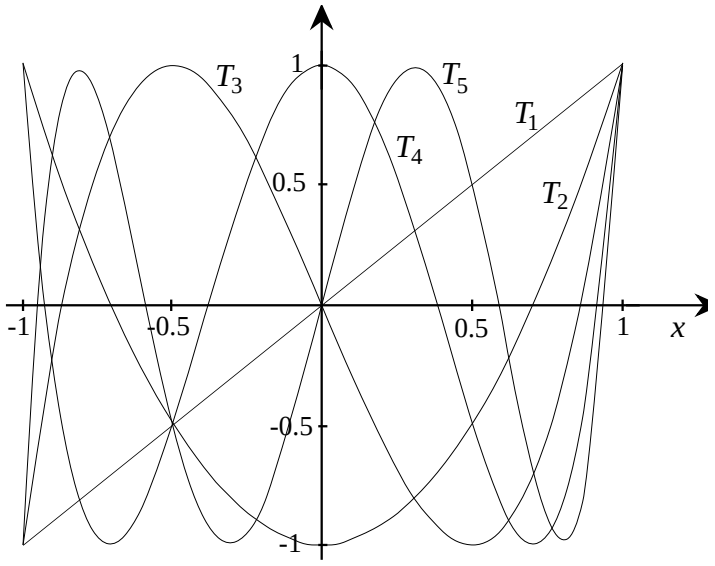


Fig. 6.4 - Grafici dei polinomi di Chebyshev di 1^a specie $T_i(x),\ i = 1, \ldots, 5$.

Dalla (26) è facile determinare gli zeri $x_k^{(i)}$, $k = 1, \ldots, i$, dell'i-esimo polinomio di Chebyshev di 1^a specie:

$$x_k^{(i)} = \cos \frac{(2k-1)\pi}{2i}\,, \quad k = 1, \ldots, i.$$

Il seguente teorema dà un'importante proprietà dei polinomi di Chebyshev di 1^a specie, che sarà assai utile in seguito.

6.19 Teorema. *Fra tutti i polinomi monici di grado $n \geq 1$, il polinomio*

$$\frac{1}{2^{n-1}}\, T_n(x)$$

è quello che ha la minima norma ∞ sull'intervallo $[-1, 1]$.

Dim. Dalla (26) si ha che nell'intervallo $[-1, 1]$ il polinomio $T_n(x)$ assume alternativamente, complessivamente $n+1$ volte, il valore 1 come massimo e il valore -1 come minimo; ne segue che $\|T_n(x)\|_\infty = 1$ per ogni n e quindi

$$\|\frac{1}{2^{n-1}}\, T_n(x)\|_\infty = \frac{1}{2^{n-1}}\,.$$

Si suppone ora per assurdo che esista un polinomio monico $p(x)$ di grado n, tale che

$$\|p(x)\|_\infty < \frac{1}{2^{n-1}}\,.$$

Quindi i massimi e i minimi di $p(x)$ avrebbero modulo minore di $1/2^{n-1}$ e il seguente polinomio, non identicamente nullo,

$$q(x) = p(x) - \frac{1}{2^{n-1}}\, T_n(x),$$

di grado minore o uguale a $n-1$, in quanto differenza di due polinomi monici di grado n, assumerebbe valore negativo nei punti in cui $T_n(x)$ ha massimo e valore positivo nei punti in cui $T_n(x)$ ha minimo. Complessivamente $q(x)$ assumerebbe valori alternativamente positivi e negativi in $n+1$ punti, e quindi dovrebbe annullarsi in almeno n punti, il che è assurdo. ∎

Il teorema 6.19 può essere espresso anche nella forma equivalente: fra tutti i polinomi di grado n che hanno norma ∞ uguale a 1 nell'intervallo $[-1, 1]$, il polinomio di Chebyshev di 1^a specie è quello che ha il primo coefficiente più elevato.

3.3 Polinomi di Chebyshev di 2^a specie

I polinomi di Chebyshev di 2^a specie $\{U_i(x)\}_{i\in\mathbf{N}}$ sono i polinomi ortogonali sull'intervallo $[-1, 1]$ rispetto al peso $\omega(x) = (1-x^2)^{1/2}$, che si ottengono dai polinomi ultrasferici ponendo

$$\alpha = \frac{1}{2} \text{ e } \beta_i = \beta_{(i,1/2)} = \frac{(-1)^i 2^i (i+1)!}{(2i+1)!}\,, \text{ cioè } U_i(x) = \beta_{(i,1/2)} f_{(i,1/2)}(x)$$

(per $i = 0$ è $\beta_{(0,1/2)} = 1$). Dalla (17) si ha la formula di Rodrigues per i polinomi di Chebyshev di 2^a specie

$$U_i(x) = \frac{(-1)^i 2^i (i+1)!}{(2i+1)!}\,(1-x^2)^{-1/2}\,\frac{d^i}{dx^i}\,\big[(1-x^2)^{i+1/2}\big].$$

Dalla (23) si ottiene

$$a_i = \frac{2^i (i+1)!}{(2i+1)!}\,\frac{\Gamma(2i+2)}{\Gamma(i+2)} = 2^i.$$

Dalla (24) si ottiene la costante di normalizzazione

$$\begin{aligned} h_i = h_i^{(1/2)} &= \frac{i!\,(i+1)!\,2^{4i+2}}{(2i+1)!}\,\frac{\big[\Gamma\big(i+\frac{3}{2}\big)\big]^2}{\Gamma(2i+3)} \\ &= \frac{i!\,(i+1)!\,2^{4i+2}}{(2i+1)!}\,\frac{[(2i+2)!]^2\pi}{2^{4i+4}[(i+1)!]^2(2i+2)!} = \frac{\pi}{2} \end{aligned}$$

(si veda l'esercizio 4.22 c)). Poiché $b_i = 0$, per il teorema 6.13 si ha:

$$A_i = 2, \quad B_i = 0, \quad C_i = 1.$$

La relazione ricorrente a tre termini (9) è allora:

$$U_{i+1}(x) = 2xU_i(x) - U_{i-1}(x), \tag{27}$$

e, a partire dai primi due polinomi $U_0(x) = 1$, $U_1(x) = 2x$, consente di costruire facilmente i polinomi di Chebyshev di 2^a specie.

6.20 Esempio. I primi 7 polinomi di Chebyshev di 2^a specie sono

$$\begin{aligned}
U_0(x) &= 1 \\
U_1(x) &= 2x \\
U_2(x) &= 4x^2 - 1 \\
U_3(x) &= 8x^3 - 4x \\
U_4(x) &= 16x^4 - 12x^2 + 1 \\
U_5(x) &= 32x^5 - 32x^3 + 6x \\
U_6(x) &= 64x^6 - 80x^4 + 24x^2 - 1.
\end{aligned}$$

I grafici dei polinomi di Chebyshev di 2^a specie $U_i(x)$, per $i = 1, \ldots, 5$, sono rappresentati nella figura 6.5. ∎

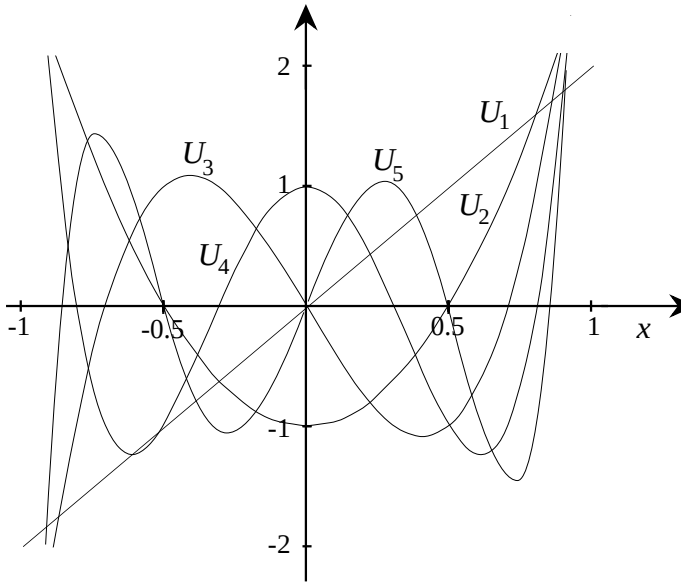


Fig. 6.5 - Grafici dei polinomi di Chebyshev di 2^a specie $U_i(x),\ i = 1, \ldots, 5$.

Anche per i polinomi di Chebyshev di 2^a specie si può dare un'espressione esplicita molto semplice. Ponendo $x = \cos\theta$, si ha

$$U_i(\cos\theta) = \frac{\sin(i+1)\theta}{\sin\theta}\,, \quad i = 0, 1, \ldots \tag{28}$$

come si può vedere per induzione applicando la (27); infatti

$$\begin{aligned} U_{i+1}(\cos\theta) &= 2\cos\theta\,\frac{\sin(i+1)\theta}{\sin\theta} - \frac{\sin i\theta}{\sin\theta} \\ &= \frac{1}{\sin\theta}\left\{2\cos\theta\sin(i+1)\theta - \sin[(i+1)\theta - \theta]\right\} \\ &= \frac{1}{\sin\theta}\left[2\cos\theta\sin(i+1)\theta - \sin(i+1)\theta\cos\theta + \cos(i+1)\theta\sin\theta\right] \\ &= \frac{1}{\sin\theta}\left[\cos\theta\sin(i+1)\theta + \cos(i+1)\theta\sin\theta)\right] = \frac{\sin(i+2)\theta}{\sin\theta}\,. \end{aligned}$$

Dalla (28) è facile determinare gli zeri $x_k^{(i)}$, $k = 1, \ldots, i$, dell'i-esimo polinomio di Chebyshev di 2^a specie:

$$x_k^{(i)} = \cos\,\frac{k\pi}{i+1}\,, \quad k = 1, \ldots, i.$$

Inoltre, dalla (26) risulta che

$$\frac{dT_i(x)}{dx} = \frac{dT_i(\cos\theta)}{d\theta}\,\frac{d\theta}{dx} = \frac{d\cos i\theta}{d\theta}\Big/\frac{d\cos\theta}{dx} = i\,\frac{\sin i\theta}{\sin\theta}\,,$$

da cui, confrontando la (28), si ottiene una relazione che lega fra loro i polinomi di Chebyshev di 1^a specie e quelli di 2^a specie:

$$T_i'(x) = i\,U_{i-1}(x), \quad i = 1, 2, \ldots \tag{29}$$

3.4 Polinomi di Laguerre

I polinomi di Laguerre $\{L_i(x)\}_{i\in\mathbf{N}}$ sono i polinomi ortogonali sull'intervallo $[0, +\infty]$ rispetto al peso $\omega(x) = \mathrm{e}^{-x}$, che si ottengono ponendo nella (17) $s_i(x) = \mathrm{e}^{-x}x^i$ e $\beta_i = \dfrac{1}{i!}$. Infatti la funzione

$$\frac{s_i^{(i)}(x)}{\omega(x)} = \frac{1}{\mathrm{e}^{-x}}\,\frac{d^i}{dx^i}\,[\mathrm{e}^{-x}x^i]$$

è un polinomio di grado i, come si vede applicando la regola di Leibniz:

$$\begin{aligned}\frac{d^i}{dx^i}\,[\mathrm{e}^{-x}x^i] &= \sum_{j=0}^{i}\binom{i}{j}\,\frac{d^j}{dx^j}\,\mathrm{e}^{-x}\,\frac{d^{i-j}}{dx^{i-j}}\,x^i\\ &= \mathrm{e}^{-x}\sum_{j=0}^{i}(-1)^j\binom{i}{j}\,i(i-1)\cdots(j+1)x^j\\ &= \mathrm{e}^{-x}\sum_{j=0}^{i}(-1)^j\,\frac{(i!)^2}{(j!)^2(i-j)!}\,x^j.\end{aligned}$$

La formula di Rodrigues per i polinomi di Laguerre è

$$L_i(x) = \frac{1}{i!}\,\mathrm{e}^x\,\frac{d^i}{dx^i}\,[\mathrm{e}^{-x}x^i],$$

e la loro espressione esplicita è data da:

$$L_i(x) = \sum_{j=0}^{i}(-1)^j\,\frac{i!}{(j!)^2(i-j)!}\,x^j.$$

I coefficienti a_i e b_i dei termini di grado i e $i-1$ sono dati rispettivamente da

$$a_i = \frac{(-1)^i}{i!}\,,\quad b_i = (-1)^{i-1}\,\frac{i!}{((i-1)!)^2} = (-1)^{i-1}\,\frac{i}{(i-1)!}\,.$$

L'ortogonalità dei polinomi di Laguerre discende dal teorema 6.15. Infatti in modo analogo a quanto fatto sopra, si ha che

$$\frac{d^k}{dx^k}\,[\mathrm{e}^{-x}x^i] = \mathrm{e}^{-x}\sum_{j=0}^{k}(-1)^j\,\frac{i!\;k!}{j!\;(k-j)!\;(i-k+j)!}\,x^{i-k+j},$$

e quindi

$$\frac{d^k}{dx^k}\,[\mathrm{e}^{-x}x^i]\Big|_{x=0} = 0 \quad\text{e}\quad \lim_{x\to\infty}\frac{d^k}{dx^k}\,[\mathrm{e}^{-x}x^i] = 0$$

per ogni i e $k = 0,\ldots,i-1$. Inoltre dalla (18) segue

$$h_i = \frac{1}{i!}\int_0^\infty \mathrm{e}^{-x}x^i\,dx = \frac{1}{i!}\,\Gamma(i+1) = 1.$$

Quindi i polinomi di Laguerre sono normalizzati. Dal teorema 6.13 si ricava la relazione ricorrente a tre termini (9):

$$\begin{aligned}A_i &= -\,\frac{1}{i+1}\,,\quad B_i = \frac{(i+1)^2-i^2}{i+1} = \frac{2i+1}{i+1}\,,\\ C_i &= \frac{(i!)^2}{(i+1)!\;(i-1)!} = \frac{i}{i+1}\,,\end{aligned}$$

da cui

$$(i+1)L_{i+1}(x) = (2i+1-x)L_i(x) - iL_{i-1}(x).$$

A partire dai primi due polinomi $L_0(x) = 1$, $L_1(x) = -x + 1$, con questa relazione si costruiscono i polinomi di Laguerre.

6.21 Esempio. I primi 7 polinomi di Laguerre sono

$$\begin{aligned}
L_0(x) &= 1 \\
L_1(x) &= -x + 1 \\
L_2(x) &= \frac{1}{2}\,(x^2 - 4x + 2) \\
L_3(x) &= -\frac{1}{6}\,(x^3 - 9x^2 + 18x - 6) \\
L_4(x) &= \frac{1}{24}\,(x^4 - 16x^3 + 72x^2 - 96x + 24) \\
L_5(x) &= -\frac{1}{120}\,(x^5 - 25x^4 + 200x^3 - 600x^2 + 600x - 120) \\
L_6(x) &= \frac{1}{720}\,(x^6 - 36x^5 + 450x^4 - 2400x^3 + 5400x^2 - 4320x + 720).
\end{aligned}$$

I grafici dei polinomi di Laguerre $L_i(x)$, per $i = 1, \ldots, 5$, sono rappresentati nella figura 6.6. ∎

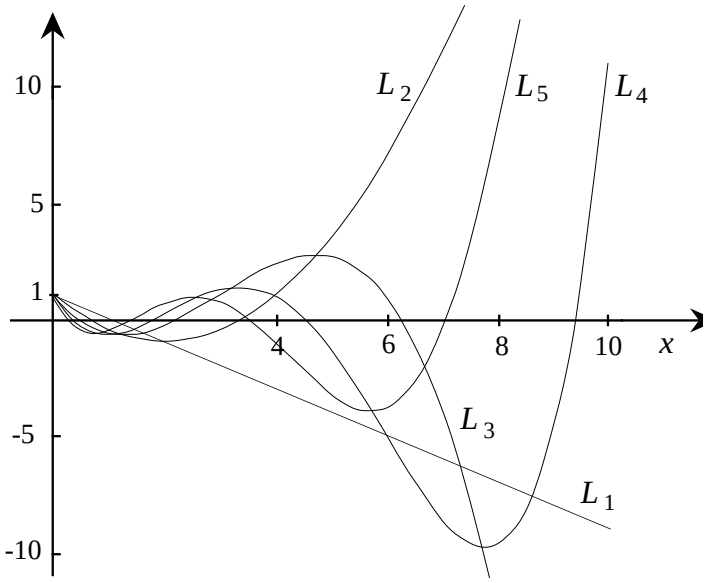


Fig. 6.6 - Grafici dei polinomi di Laguerre $L_i(x),\ i = 1, \ldots, 5$.

3.5 Polinomi di Hermite

I polinomi di Hermite $\{H_i(x)\}_{i\in\mathbf{N}}$ sono i polinomi ortogonali sull'intervallo $[-\infty,+\infty]$ rispetto al peso $\omega(x) = \mathrm{e}^{-x^2}$, che si ottengono ponendo nella (17) $s_i(x) = \mathrm{e}^{-x^2}$ e $\beta_i = (-1)^i$. Poiché

$$\frac{d^i}{dx^i}\ \mathrm{e}^{-x^2} = \mathrm{e}^{-x^2} q_i(x),$$

dove

$$q_i(x) = q'_{i-1}(x) - 2xq_{i-1}(x), \quad \text{e} \quad q_0(x) = 1,$$

la funzione

$$\mathrm{e}^{x^2}\ \frac{d^i}{dx^i}\ \mathrm{e}^{-x^2}$$

è in effetti un polinomio di grado i. La formula di Rodrigues per i polinomi di Hermite è allora

$$H_i(x) = (-1)^i \mathrm{e}^{x^2}\ \frac{d^i}{dx^i}\ \mathrm{e}^{-x^2}.$$

Inoltre, poiché $q_0(x) = 1$, ad ogni derivazione il coefficiente del termine di grado più elevato risulta moltiplicato per -2. Quindi

$$a_i = 2^i.$$

Poiché la funzione e^{-x^2} è pari, le sue derivate sono funzioni pari o dispari a seconda dell'ordine di derivazione. Ne segue che per i polinomi di Hermite vale la relazione

$$H_i(-x) = (-1)^i H_i(x),$$

e quindi, se i è pari, tutti i coefficienti dei termini di grado dispari sono nulli, se i è dispari, tutti i coefficienti dei termini di grado pari sono nulli, perciò $b_i = 0$ per ogni i. L'ortogonalità dei polinomi di Hermite discende immediatamente dal teorema 6.15. Per la (18) la costante di normalizzazione è

$$h_i = 2^i i!\ \int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\ dx = 2^i i!\ \sqrt{\pi}.$$

Dal teorema 6.13 si ricava la relazione ricorrente a tre termini (9):

$$A_i = 2, \quad B_i = 0, \quad C_i = 2i,$$

da cui

$$H_{i+1}(x) = 2xH_i(x) - 2iH_{i-1}(x).$$

A partire dai primi due polinomi $H_0(x) = 1$, $H_1(x) = 2x$, con questa relazione si costruiscono i polinomi di Hermite.

6.22 Esempio. I primi 7 polinomi di Hermite sono

$$
\begin{aligned}
H_0(x) &= 1 \\
H_1(x) &= 2x \\
H_2(x) &= 4x^2 - 2 \\
H_3(x) &= 8x^3 - 12x \\
H_4(x) &= 16x^4 - 48x^2 + 12 \\
H_5(x) &= 32x^5 - 160x^3 + 120x \\
H_6(x) &= 64x^6 - 480x^4 + 720x^2 - 120.
\end{aligned}
$$

I grafici dei polinomi $H_i(x)/2^i$, per $i = 1, \ldots, 5$, sono rappresentati nella figura 6.7. ∎

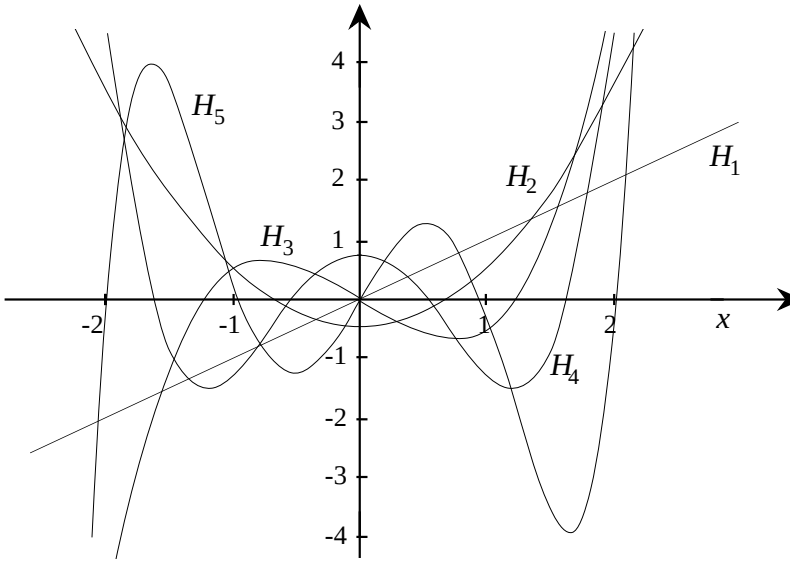


Fig. 6.7 - Grafici dei polinomi di Hermite $H_i(x)/2^i,\ i = 1, \ldots, 5$.

4. Approssimazione ai minimi quadrati

Per quanto visto nel paragrafo 1, per costruire il polinomio $g_n(x)$ di approssimazione ai minimi quadrati di una funzione, conviene scegliere una base ortogonale di polinomi $\{p_i(x)\}_{i \in \mathbf{N}}$. Per la (4) risulta allora

$$
g_n(x) = \sum_{i=0}^{n} \alpha_i^* p_i(x), \tag{30}
$$

dove

$$\alpha_i^* = \frac{1}{h_i} \int_a^b \omega(x) f(x) p_i(x)\, dx. \tag{31}$$

Per la (5) l'errore assoluto in norma 2 è

$$\delta_n^2 = \|f\|_2^2 - \sum_{i=0}^{n} h_i (\alpha_i^*)^2, \tag{32}$$

e dall'uguaglianza di Parseval risulta $\lim_{n\to\infty} \delta_n = 0$.

I polinomi ortogonali più frequentemente usati sono quelli del paragrafo precedente. Se i due estremi dell'intervallo non sono limitati si usano i polinomi di Hermite, se uno dei due estremi è limitato e l'altro no si usano i polinomi di Laguerre, se l'intervallo è limitato si usano i polinomi di Legendre e di Chebyshev. Nel caso di un intervallo limitato da una sola parte, è sempre possibile ricondursi all'intervallo $[0, +\infty]$ con una traslazione ed eventualmente un cambio di segno, mentre nel caso di un intervallo limitato $[a, b]$ si effettua il cambiamento di variabile

$$x = \frac{1}{2}\,[(b-a)y + (a+b)], \tag{33}$$

che fa corrispondere all'intervallo $[a, b]$ l'intervallo $[-1, 1]$.

6.23 Esempio. La funzione $f(x) = |x|$ può essere approssimata sull'intervallo $[-1, 1]$ con la combinazione lineare (30) usando i polinomi di Legendre (o anche i polinomi di Chebyshev, si veda l'esempio 6.25). Poiché la funzione $f(x)$ è pari e i polinomi di Legendre sono funzioni pari oppure dispari a seconda che il grado sia pari o dispari, i coefficienti α_i^* con indice i dispari sono nulli.

Assumendo $n = 4$ dalla (31) si ha

$$\alpha_0^* = \frac{1}{h_0} \int_{-1}^{1} |x|\, P_0(x)\, dx = \frac{1}{2} \int_{-1}^{1} |x|\, dx = \frac{1}{2}\,, \qquad \alpha_1^* = 0,$$

$$\alpha_2^* = \frac{1}{h_2} \int_{-1}^{1} |x|\, P_2(x)\, dx = \frac{5}{2} \int_{-1}^{1} |x|\, \frac{3x^2-1}{2}\, dx = \frac{5}{8}\,, \qquad \alpha_3^* = 0,$$

$$\alpha_4^* = \frac{1}{h_4} \int_{-1}^{1} |x|\, P_4(x)\, dx = \frac{9}{2} \int_{-1}^{1} |x|\, \frac{35x^4 - 30x^2 + 3}{8}\, dx = -\,\frac{3}{16}\,,$$

e quindi

$$g_4(x) = \frac{1}{2}\, P_0(x) + \frac{5}{8}\, P_2(x) - \frac{3}{16}\, P_4(x) = \frac{15}{128}\, (-7x^4 + 14x^2 + 1),$$

con un errore assoluto in norma che, per la (32), è dato da

$$\delta_4 = \Big[\|f\|_2^2 - h_0(\alpha_0^*)^2 - h_2(\alpha_2^*)^2 - h_4(\alpha_4^*)^2\Big]^{1/2} \approx 0.510\ 10^{-1}.$$

Il massimo modulo del resto sull'intervallo $[-1, 1]$ risulta

$$\max_{x\in[-1,1]} \big|\ |x| - g_4(x)\big| = g_4(0) = \frac{15}{128} \approx 0.117.$$

Si noti come questo valore sia minore del valore 0.147 ottenuto nell'esempio 5.22 interpolando $f(x)$ nello stesso intervallo in 5 punti equidistanti. Nella figura 6.8 sono riportati il grafico di $g_4(x)$ (linea più sottile) e il grafico della funzione $f(x)$ (linea spessa). È interessante confrontare questo grafico con quello della figura 5.9 relativo al polinomio di interplazione. ∎

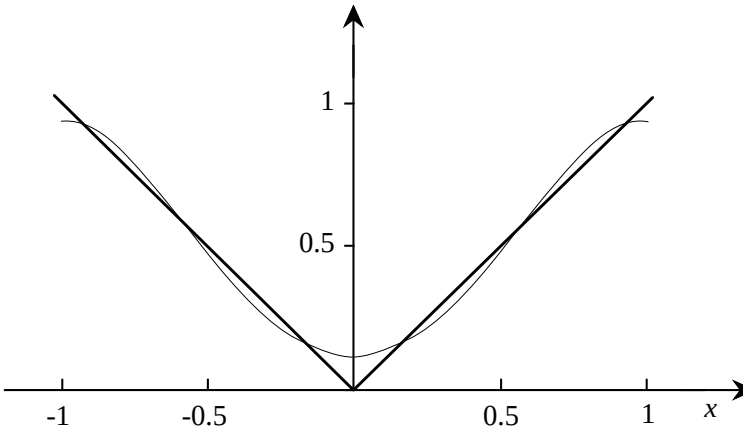


Fig. 6.8 - Approssimazione della funzione $f(x) = |x|$ con polinomi di Legendre.

6.24 Esempio. Con il cambiamento di variabile (33) $x = \dfrac{\pi}{4}(y+1)$, la funzione $f(x) = \sin x$, definita sull'intervallo $\big[0, \dfrac{\pi}{2}\big]$, viene trasformata nella funzione

$$\psi(y) = \sin\frac{y+1}{\rho}, \quad \rho = \frac{4}{\pi},$$

definita nell'intervallo $[-1, 1]$. Per determinare l'approssimazione di $\psi(y)$ con i primi 5 polinomi di Legendre, dalla (31) si ha

$$\alpha_0^* = \frac{1}{h_0}\int_{-1}^{1} \psi(y)P_0(y)\,dy = \frac{1}{2}\int_{-1}^{1} \sin\frac{y+1}{\rho}\,dy = \frac{\rho}{2} = 0.6366198,$$

$$\alpha_1^* = \frac{1}{h_1}\int_{-1}^{1} \psi(y)P_1(y)\,dy = \frac{3}{2}\int_{-1}^{1} y\sin\frac{y+1}{\rho}\,dy = \frac{3}{2}\,\rho(\rho-1) = 0.5218491,$$

$$\alpha_2^* = \frac{1}{h_2}\int_{-1}^{1} \psi(y)P_2(y)\,dy = \frac{5}{2}\int_{-1}^{1} \frac{3y^2-1}{2}\,\sin\,\frac{y+1}{\rho}\,dy$$
$$= \frac{5}{2}\,\rho\,(-3\rho^2+3\rho+1) = -0.1390956,$$

$$\alpha_3^* = \frac{1}{h_3}\int_{-1}^{1} \psi(y)P_3(y)\,dy = \frac{7}{2}\int_{-1}^{1} \frac{5y^3-3y}{2}\,\sin\,\frac{y+1}{\rho}\,dy$$
$$= \frac{7}{2}\,\rho\,(-15\rho^3+15\rho^2+6\rho-1) = -0.02206651,$$

$$\alpha_4^* = \frac{1}{h_4}\int_{-1}^{1} \psi(y)P_4(y)\,dy = \frac{9}{2}\int_{-1}^{1} \frac{35y^4-30y^2+3}{8}\,\sin\,\frac{y+1}{\rho}\,dy$$
$$= \frac{9}{2}\,\rho\,(105\rho^4-105\rho^3-45\rho^2+10\rho+1) = 0.002491448.$$

Si ha quindi

$$g_4(y) = \sum_{i=0}^{4} \alpha_i^* P_i(y),$$

dove $y = \dfrac{4}{\pi}\,x - 1$, con un errore assoluto in norma che per la (32) è dato da

$$\delta_4 = \Big[\|f\|_2^2 - \sum_{i=0}^{4} h_i(\alpha_i^*)^2\Big]^{1/2} = \Big[\rho\int_0^{\pi/2} \sin^2 x\,dx - \sum_{i=0}^{4} h_i(\alpha_i^*)^2\Big]^{1/2}$$
$$\approx 0.119\ 10^{-3}.$$

Il massimo modulo del resto sull'intervallo $[-1,1]$ risulta

$$\max_{y\in[-1,1]} \big|\psi(y) - g_4(y)\big| = |g_4(-1)| \approx 0.233\ 10^{-3}.$$

Nella figura 6.9 è riportato il grafico del resto. ∎

Se al posto dei polinomi di Legendre si usano i polinomi di Chebyshev, si ottengono approssimazioni più accurate nelle zone vicine agli estremi dell'intervallo e meno accurate nella zona centrale dell'intervallo con i polinomi di 1^a specie, e viceversa con i polinomi di 2^a specie; infatti per i polinomi di 1^a specie è $\omega(x) = (1-x^2)^{-1/2}$, che ha minimo in $x = 0$ e tende all'infinito per $x \to \pm 1$, mentre per i polinomi di 2^a specie è $\omega(x) = (1-x^2)^{1/2}$, che ha minimo in $x = \pm 1$ e massimo in $x = 0$.

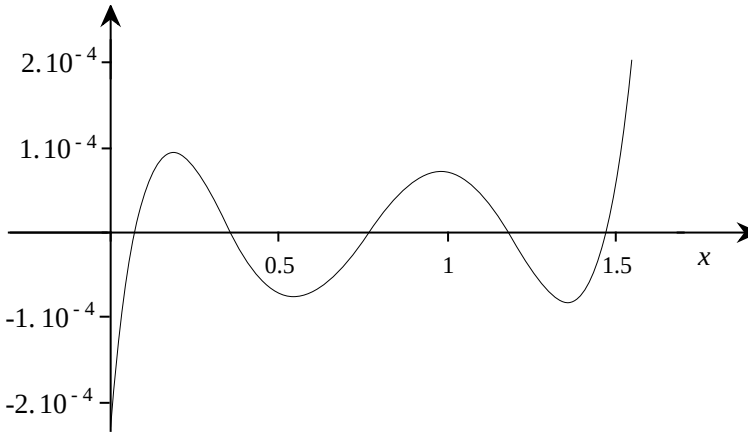


Fig. 6.9 - Resto nell'approssimazione della funzione $f(x) = \sin x$ con polinomi di Legendre.

Nel caso in cui i polinomi usati siano quelli di Chebyshev di 1^a specie, conviene utilizzare la rappresentazione (26). Ponendo $x = \cos\theta$, $0 \le \theta \le \pi$, si ha dalla (31)

$$\alpha_i^* = \frac{1}{h_i} \int_{-1}^{1} (1-x^2)^{-1/2} f(x) T_i(x)\, dx = \frac{1}{h_i} \int_0^{\pi} f(\cos\theta) \cos i\theta \, d\theta,$$

dove

$$h_0 = \pi, \quad h_i = \frac{\pi}{2} \quad \text{per} \quad i = 1, 2, \ldots$$

Per uniformità di notazione si usa raddoppiare il coefficiente α_0^*, per cui l'approssimazione ai minimi quadrati con polinomi di Chebyshev di 1^a specie diventa

$$g_n(x) = \frac{\alpha_0^*}{2} + \sum_{i=1}^{n} \alpha_i^* T_i(x), \tag{34}$$

o in forma di polinomio trigonometrico

$$g_n(\cos\theta) = \frac{\alpha_0^*}{2} + \sum_{i=1}^{n} \alpha_i^* \cos i\theta, \tag{35}$$

dove

$$\alpha_i^* = \frac{2}{\pi} \int_0^{\pi} f(\cos\theta) \cos i\theta \, d\theta. \tag{36}$$

Dalla (32) risulta poi che

$$\delta_n = \left[\|f\|_2^2 - \frac{\pi}{4} (\alpha_0^*)^2 - \frac{\pi}{2} \sum_{i=1}^{n} (\alpha_i^*)^2 \right]^{1/2}.$$

Se la funzione $f(x)$ è simmetrica in $[-1,1]$, è $\alpha_i^* = 0$ per i dispari, se $f(x)$ è antisimmetrica in $[-1,1]$, è $\alpha_i^* = 0$ per i pari.

Il polinomio (35) con i coefficienti (36) coincide con il polinomio trigonometrico che si ottiene troncando la serie di Fourier (di soli coseni) della funzione ottenuta estendendo per simmetria la $f(\cos\theta)$ all'intervallo $[-\pi,\pi]$ (si veda l'esercizio 6.35).

6.25 Esempio. Come nell'esempio 6.23 sia $f(x) = |x|$, $-1 \le x \le 1$. Usando i polinomi di Chebyshev di 1^a specie, per $n = 4$ si ha

$$\begin{aligned}
\alpha_0^* &= \frac{2}{\pi}\int_0^\pi |\cos\theta|\, d\theta = \frac{4}{\pi}\,, \quad \alpha_1^* = 0,\\
\alpha_2^* &= \frac{2}{\pi}\int_0^\pi |\cos\theta|\cos 2\theta\, d\theta = \frac{4}{3\pi}\,, \quad \alpha_3^* = 0,\\
\alpha_4^* &= \frac{2}{\pi}\int_0^\pi |\cos\theta|\cos 4\theta\, d\theta = -\,\frac{4}{15\pi}\,,
\end{aligned}$$

e quindi il polinomio di approssimazione è

$$g_4^{(T)}(x) = \frac{2}{\pi}\,[1 + \frac{2}{3}\,T_2(x) - \;\frac{2}{15}\,T_4(x)] = \frac{2}{15\pi}\,(-16x^4 + 36x^2 + 3),$$

con l'errore di approssimazione $\delta_4^{(T)} \approx 0.575\ 10^{-1}$. Il massimo modulo del resto sull'intervallo $[-1,1]$ è

$$\max_{x\in[-1,1]} \big|\, |x| - g_4^{(T)}(x)\big| = g_4^{(T)}(0) = \frac{2}{5\pi} \approx 0.127.$$

Usando i polinomi di Chebyshev di 2^a specie risulta

$$\alpha_0^* = \frac{4}{3\pi}\,, \quad \alpha_1^* = 0, \quad \alpha_2^* = \frac{4}{5\pi}\,, \quad \alpha_3^* = 0, \quad \alpha_4^* = -\,\frac{4}{21\pi}\,,$$

e quindi il polinomio di approssimazione è

$$g_4^{(U)}(x) = \frac{4}{3\pi}\,[1 + \frac{3}{5}\,U_2(x) - \;\frac{1}{7}\,U_4(x)] = \frac{4}{105\pi}\,(-80x^4 + 144x^2 + 9),$$

con l'errore di approssimazione $\delta_4^{(U)} \approx 0.468\ 10^{-1}$. Il massimo modulo del resto sull'intervallo $[-1,1]$ è

$$\max_{x\in[-1,1]} \big|\, |x| - g_4^{(U)}(x)\big| = 1 - g_4^{(U)}(1) \approx 0.115.$$

Nella figura 6.10 sono riportati sull'intervallo $[0, 1]$ (il resto è una funzione simmetrica nell'intervallo $[-1, 1]$) i grafici dei resti dei polinomi di approssimazione ottenuti nell'esempio 6.23 con i polinomi di Legendre (linea continua) e in questo esempio con i polinomi di Chebyshev di 1^a specie (linea a puntini) e con i polinomi di Chebyshev di 2^a specie (linea tratteggiata).

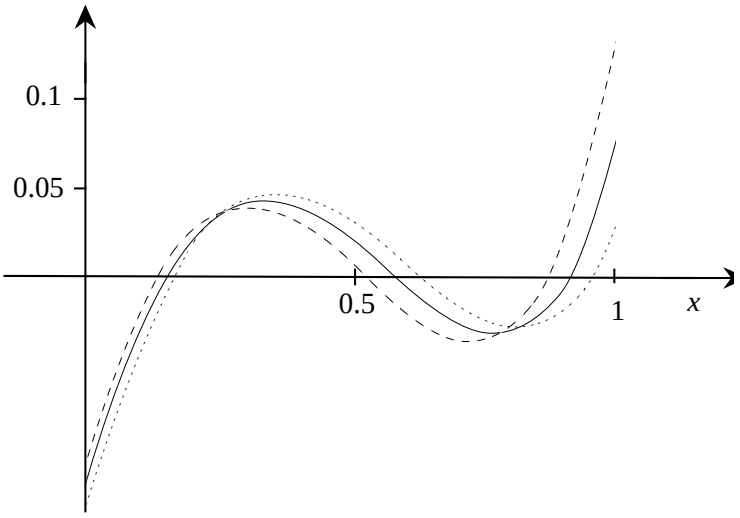


Fig. 6.10 - Resti delle approssimazioni di Legendre e di Chebyshev di $f(x) = |x|$.

Si noti come rispetto all'approssimazione di Legendre, l'approssimazione di Chebyshev di 1^a specie sia più accurata vicino agli estremi dell'intervallo e meno al centro, e l'approssimazione di Chebyshev di 2^a specie sia più accurata nel centro e meno agli estremi. Infatti, indicati con $r_L(x)$, $r_T(x)$ e $r_U(x)$ i resti delle approssimazioni di Legendre, di Chebyshev di 1^a specie e di Chebyshev di 2^a specie, risulta

$$|r_L(\pm 1)| = 1 - \frac{15}{16} = 0.0625, \qquad |r_L(0)| = \frac{15}{128} = 0.117,$$
$$|r_T(\pm 1)| = 1 - \frac{46}{15\pi} \approx 0.0238, \quad |r_T(0)| = \frac{2}{5\pi} \approx 0.127,$$
$$|r_U(\pm 1)| = 1 - \frac{292}{105\pi} \approx 0.115, \quad |r_U(0)| = \frac{12}{35\pi} \approx 0.109,$$

per cui

$$|r_U(0)| < |r_L(0)| < |r_T(0)| \quad \text{e} \quad |r_T(\pm 1)| < |r_L(\pm 1)| < |r_U(\pm 1)|.$$

In questo caso particolare risulta

$$\|r_U\|_\infty < \|r_L\|_\infty < \|r_T\|_\infty,$$

e quindi l'approssimazione di Chebyshev di 2^a specie è globalmente la migliore. ∎

Il calcolo dei coefficienti α_i^* tramite le (31) può essere troppo complicato, o addirittura irrealizzabile quando non è possibile valutare l'integrale in termini di funzioni elementari. Quindi spesso si deve ricorrere a metodi approssimati per ricavare i coefficienti, seguendo due possibili strade: o si passa a un problema discretizzato (si veda il paragrafo 13) o si approssima l'integrale (31) con una formula di quadratura (si veda il capitolo 7). Si genera così un errore nel calcolo dei coefficienti, che non dovrà superare il modulo del resto.

Nel caso dei polinomi di Chebyshev di 1^a specie conviene approssimare i coefficienti α_i^* utilizzando la formula dei trapezi. Fissato un valore N, si scelgono come nodi della formula i punti $\theta_j = j\pi/N$, $j = 0, \ldots, N$ e si ottiene per α_i^* il valore approssimato

$$\alpha_i = \frac{1}{N}\left[y_0 + (-1)^i y_N + 2\sum_{j=1}^{N-1} y_j \cos i\theta_j\right], \quad y_j = f(\cos\theta_j).$$

Il calcolo può essere fatto usando un algoritmo veloce per la trasformata discreta di coseni (si veda l'esercizio 5.62). Naturalmente per approssimare l'integrale (36) si possono usare formule con un ordine più elevato, e quindi più precise, di quella dei trapezi.

Una volta calcolati i coefficienti α_i^*, per calcolare il valore della (30) in un punto x si possono utilizzare due tecniche:

a) determinare prima i coefficienti a_k del polinomio

$$g_n(x) = \sum_{k=0}^{n} a_k x^k \tag{37}$$

e poi calcolare il valore in x con il metodo di Ruffini-Horner. I coefficienti a_k si ottengono dagli α_i^* in modo molto semplice: sia infatti

$$p_i(x) = \sum_{k=0}^{i} t_{ik} x^k, \quad i = 0, 1, \ldots, n,$$

l'i-esimo polinomio ortogonale. Allora

$$g_n(x) = \sum_{i=0}^{n} \alpha_i^* \sum_{k=0}^{i} t_{ik} x^k = \sum_{k=0}^{n} \Big[\sum_{i=k}^{n} \alpha_i^* t_{ik}\Big] x^k,$$

e confrontando con la (37) si ha che

$$a_k = \sum_{i=k}^{n} \alpha_i^* t_{ik}, \quad k = 0, 1, \ldots, n. \tag{38}$$

I coefficienti t_{ik} sono riportati in apposite tabelle, e comunque si possono calcolare ricorsivamente applicando la relazione (9). Nell'ipotesi che $p_0(x) = 1$, ipotesi che è verificata per tutti i polinomi qui studiati, si ottiene

$$\begin{aligned} &t_{00} = 1, \\ &t_{ik} = A_{i-1} t_{i-1,k-1} + B_{i-1} t_{i-1,k} - C_{i-1} t_{i-2,k}, \quad k = 0, \ldots, i, \quad i = 1, 2, \ldots \end{aligned} \tag{39}$$

considerando nulli i t_{ik} per cui $k < 0$ oppure $i < 0$ oppure $k > i$. Noti i coefficienti t_{ik}, il numero delle operazioni richieste per calcolare per mezzo delle (38) gli a_k, $k = 0, 1, \ldots, n$, è di $n^2/2$ addizioni e $n^2/2$ moltiplicazioni. Per i polinomi di Legendre, di Chebyshev e di Hermite, in cui metà dei coefficienti sono nulli, tale numero scende a $n^2/4$ addizioni e $n^2/4$ moltiplicazioni.

Nel caso particolare in cui i polinomi $p_i(x)$ siano quelli di Chebyshev di 1^a specie, i coefficienti t_{ik}, come si ricava dall'esempio 6.18, sono

	$k=0$	1	2	3	4
$i=0$	1				
1	0	1			
2	-1	0	2		
3	0	-3	0	4	
4	1	0	-8	0	8

e si possono ottenere utilizzando invece della (39) la relazione ricorrente

$$t_{ik} = 2t_{i-1,k-1} - t_{i-2,k}, \quad \text{per} \quad k = 0, \ldots, i.$$

Viceversa per la (38) i coefficienti α_i^* possono essere ricavati dai coefficienti a_k, calcolando le soluzioni di un sistema lineare a matrice triangolare, con $n^2/2$ addizioni e $n^2/2$ moltiplicazioni. Così un polinomio invece che come somma di monomi viene espresso come combinazione di polinomi ortogonali.

b) Calcolare direttamente la (30), sfruttando la relazione ricorrente a tre

termini (9). Per $n \geq 3$ si ha:

$$\begin{aligned} g_n(x) &= \sum_{i=0}^{n-3} \alpha_i^* p_i(x) + \alpha_{n-2}^* p_{n-2}(x) + \alpha_{n-1}^* p_{n-1}(x) \\ &\quad + \alpha_n^* \big[(A_{n-1}x + B_{n-1})p_{n-1}(x) - C_{n-1}p_{n-2}(x)\big] \\ &= \sum_{i=0}^{n-3} \alpha_i^* p_i(x) + (\alpha_{n-2}^* - \alpha_n^* C_{n-1})p_{n-2}(x) \\ &\quad + \big[\alpha_{n-1}^* + \alpha_n^*(A_{n-1}x + B_{n-1})\big]p_{n-1}(x) \\ &= \sum_{i=0}^{n-1} \beta_i p_i(x), \end{aligned}$$

dove

$$\beta_i = \alpha_i^*, \quad \text{per} \quad i = 0, \ldots, n-3,$$

$$\beta_{n-2} = \alpha_{n-2}^* - \alpha_n^* C_{n-1}, \quad \beta_{n-1} = \alpha_{n-1}^* + \alpha_n^*(A_{n-1}x + B_{n-1}),$$

cioè $g_n(x)$ può essere calcolato come combinazione lineare con coefficienti β_i dei polinomi ortogonali fino al grado $n-1$. Da questa relazione scaturisce il seguente algoritmo, detto *di Clenshaw*

$$\begin{aligned} &q_n = \alpha_n^* \\ &q_{n-1} = \alpha_{n-1}^* + q_n(A_{n-1}x + B_{n-1}) \\ &q_i = \alpha_i^* + q_{i+1}(A_i x + B_i) - q_{i+2}C_{i+1}, \quad \text{per} \quad i = n-2, \ldots, 1 \\ &g_n(x) = \alpha_0^* + q_1 p_1(x) - q_2 C_1, \end{aligned} \tag{40}$$

(si è supposto che $p_0(x) = 1$), con il quale il calcolo di $g_n(x)$ in un punto x richiede $3n$ addizioni e $3n$ moltiplicazioni. Per i polinomi ortogonali studiati, ad eccezione dei polinomi di Laguerre, i B_i sono tutti nulli e quindi il numero delle operazioni additive è di $2n$. Per i polinomi di Chebyshev si ha

$$\begin{aligned} &y = 2x \\ &q_n = \alpha_n^* \\ &q_{n-1} = \alpha_{n-1}^* + yq_n \\ &q_i = \alpha_i^* + yq_{i+1} - q_{i+2}, \quad \text{per} \quad i = n-2, \ldots, 1, \\ &g_n(x) = \begin{cases} \alpha_0^*/2 + xq_1 - q_2, & \text{per i pol. di Chebyshev di } 1^a \text{ specie,} \\ \alpha_0^* + yq_1 - q_2, & \text{per i pol. di Chebyshev di } 2^a \text{ specie,} \end{cases} \end{aligned}$$

(per i polinomi di 1^a specie, si è usata la notazione (34)). In questo caso l'algoritmo è particolarmente efficiente in quanto permette di calcolare il valori di $g_n(x)$ con $2n$ addizioni e n moltiplicazioni. In generale, anche se il numero di operazioni richieste per valutare un polinomio espresso come combinazione di polinomi ortogonali è superiore al numero di operazioni richieste dalla regola di Ruffini-Horner, la trasformazione del polinomio scritto in somma di monomi è conveniente solo se è richiesta la valutazione in molti punti. Questo vale in particolare se si usano i polinomi di Chebyshev.

Dal punto di vista della stabilità numerica, l'algoritmo (40) dà in generale buoni risultati. Infatti per $i = 0, \ldots, n$, è

$$q_i = \sum_{j=i}^{n} \alpha_j^* p_j(x),$$

e quindi

$$\|q_i\|_2^2 = (q_i, q_i) = \sum_{j=i}^{n} (\alpha_j^*)^2 h_j \leq \sum_{j=0}^{n} (\alpha_j^*)^2 h_j = (g_n, g_n) = \|g_n\|_2^2. \qquad (41)$$

La limitazione (41) assicura l'esistenza di un'ampia regione $D \subset [a, b]$ di stabilità dell'algoritmo di Clenshaw, nel senso che per $x \in D$ nel calcolo dei q_i non vi possono essere sottrazioni di numeri significativamente più grandi di $\|g_n\|_2^2$, che potrebbero provocare perdite di precisione.

6.26 Esempio. La funzione $f(x) = \mathrm{e}^x$ definita sull'intervallo [0,1], con il cambiamento di variabile $x = (y + 1)/2$, viene trasformata nella funzione

$$\psi(y) = \mathrm{e}^{(y+1)/2},$$

definita nell'intervallo $[-1, 1]$. L'approssimazione (34) con polinomi di Chebyshev di 1^a specie ha i coefficienti

$$\begin{aligned}
&\alpha_0^* = 3.506775, && \alpha_1^* = 0.8503917, && \alpha_2^* = 0.1052087,\\
&\alpha_3^* = 0.008722105, && \alpha_4^* = 0.000543437, && \alpha_5^* = 0.000027115,\\
&\alpha_6^* = 0.000001128, && \alpha_7^* = 0.000000040, && \alpha_8^* = 0.000000001.
\end{aligned}$$

Indicato con $h_1(y)$ il valore calcolato di $g_n(y)$ con l'algoritmo di Clenshaw e con $h_2(y)$ il valore calcolato trasformando la (34) nella forma (37), cioè in somma di monomi, e utilizzando poi la regola di Ruffini-Horner, per $y \in [-1, 1]$, si ottengono gli errori relativi

$$\epsilon_j = \frac{|h_j(y) - g_n(y)|}{|g_n(y)|}, \quad j = 1, 2.$$

Dalla figura 6.11, dove con i quadratini sono indicati gli errori ϵ_1 e con i pallini gli errori ϵ_2, risulta che i primi sono sempre minori dei secondi. Ovviamente occorre tenere conto che i coefficienti a_k delle (37) sono affetti anche dagli errori di arrotondamento dovuti alle combinazioni (38), in cui i t_{ik} non nulli sono alternativamente positivi e negativi. ∎

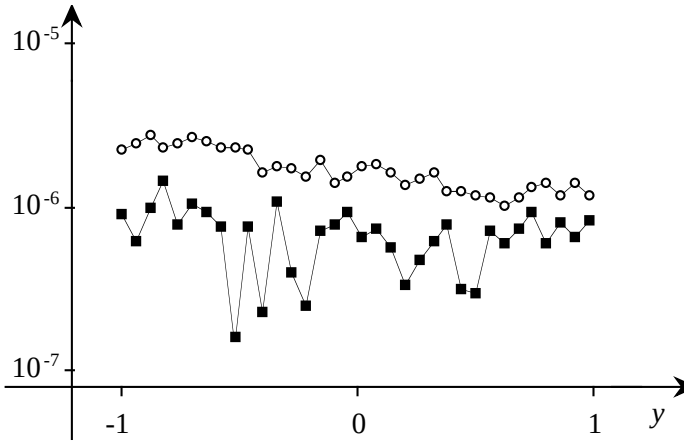


Fig. 6.11 - Errori relativi generati nel calcolo delle combinazioni di polinomi di Chebyshev e di monomi.

Poiché $\lim_{n\to\infty} \delta_n = 0$, al crescere di n la successione $\{g_n(x)\}_{n\in\mathbf{N}}$ converge in norma 2 alla funzione $f(x)$, ma ciò non assicura in generale la convergenza in norma ∞ della successione (si veda l'esercizio 6.36). Però se la funzione $f(x)$, oltre ad essere continua, soddisfa certe ipotesi di regolarità, si può dimostrare che la successione $\{g_n(x)\}_{n\in\mathbf{N}}$ converge in norma ∞ alla funzione $f(x)$ per i diversi polinomi ortogonali studiati [28]. In particolare nel caso dei polinomi di Chebyshev di 1^a specie, vale il seguente teorema.

6.27 Teorema. *Per $n \geq 1$ sia*

$$g_n(x) = \frac{\alpha_0^*}{2} + \sum_{i=1}^{n} \alpha_i^* T_i(x)$$

il polinomio di approssimazione ai minimi quadrati di una funzione $f(x) \in C[-1,1]$ ottenuto con i polinomi di Chebyshev di 1^a specie. Se $f(x)$ è lipschitziana su $[-1,1]$, allora per $n \to \infty$ la successione $\{g_n(x)\}_{n\in\mathbf{N}}$ converge uniformemente su $[-1,1]$ a $f(x)$, cioè

$$\lim_{n\to\infty} \|r_n\|_\infty = 0.$$

Se inoltre $f(x) \in C^k[-1,1]$, $k \geq 1$, *allora esiste una costante* γ *indipendente da* n, *tale che*

$$\|r_n\|_\infty \leq \frac{\gamma \log n}{n^k}. \tag{42}$$

Se la successione $g_n(x)$ *converge, risulta*

$$f(x) = \frac{\alpha_0^*}{2} + \sum_{i=1}^{\infty} \alpha_i^* T_i(x),$$

e questa espressione viene detta espansione in serie di Chebyshev della $f(x)$.

Per una dimostrazione di convergenza nel caso che $f(x) \in C^1[-1,1]$ si veda l'esercizio 6.23; per la dimostrazione di un risultato più debole della (42) nel caso che $f(x) \in C^2[-1,1]$, si veda l'esercizio 6.34; per la dimostrazione completa del teorema si veda [30] e [39]. ∎

5. Approssimazione minimax polinomiale

L'approssimazione di una funzione in norma ∞, detta *approssimazione minimax* o *di Chebyshev*, è quella generalmente usata per calcolare le funzioni matematiche mediante calcolatore.

La norma ∞ non è indotta da nessun prodotto scalare (si veda l'esercizio 6.37), pertanto la teoria dell'approssimazione sviluppata per la norma 2 con gli spazi di Hilbert non può essere utilizzata nel caso della norma ∞. In questo paragrafo viene esaminato il problema dell'approssimazione lineare in norma ∞ quando l'insieme $\mathcal{G}$ delle funzioni approssimanti è l'insieme dei polinomi; nel paragrafo 10 si esaminerà il caso in cui $\mathcal{G}$ è l'insieme delle funzioni razionali; per classi più generali di funzioni si veda l'esercizio 6.38.

Sia $\mathcal{P}_n$ la classe dei polinomi di grado minore o uguale ad n e sia $f(x)$ una funzione continua su un intervallo limitato $[a,b]$. Il teorema 6.4 assicura l'esistenza di polinomi di *approssimazione minimax*, cioè polinomi $p_n^*(x) \in \mathcal{P}_n$, tali che

$$\|f - p_n^*\|_\infty = \min_{p_n \in \mathcal{P}_n} \|f - p_n\|_\infty. \tag{43}$$

Una caratteristica di tali polinomi è, come si vedrà, quella di equioscillare attorno alla funzione $f(x)$ e tale proprietà è proprio quella che consentirà di stabilire l'unicità del polinomio di approssimazione minimax.

Posto

$$r^*(x) = f(x) - p_n^*(x) \quad \text{e} \quad r^* = \|f - p_n^*\|_\infty$$

dalla (43) segue che

$$r^* \leq \|f - p_n\|_\infty,$$

per ogni polinomio $p_n(x) \in \mathcal{P}_n$. Il seguente teorema consente di dare delle limitazioni inferiori per r^*.

6.28 Teorema *(de la Vallée-Poussin).* *Siano $f(x) \in C[a,b]$ e $p_n(x) \in \mathcal{P}_n$ tali che*

$$f(x_i) - p_n(x_i) = (-1)^i d_i, \qquad i = 0, 1, \ldots, n+1,$$

in cui

$$a \le x_0 < x_1 < \ldots < x_{n+1} \le b,$$

e tutti i d_i sono non nulli e hanno lo stesso segno. Allora

$$\min_i |d_i| \le r^*.$$

Dim. Per semplicità, si suppone $d_i > 0$ per ogni i, ma la dimostrazione è perfettamente analoga nel caso opposto. Si suppone per assurdo che esista un polinomio $q(x) \in \mathcal{P}_n$ tale che

$$\|f - q\|_\infty < \min_i d_i. \tag{44}$$

Il polinomio $s(x) = p_n(x) - q(x) \in \mathcal{P}_n$ è tale che per $i = 0, 1, \ldots, n+1$

$$\begin{aligned} s(x_i) &= p_n(x_i) - q(x_i) \\ &= [f(x_i) - q(x_i)] - [f(x_i) - p_n(x_i)] = [f(x_i) - q(x_i)] - (-1)^i d_i. \end{aligned}$$

Dalla (44), per gli indici i pari, si ha

$$s(x_i) = [f(x_i) - q(x_i)] - d_i < 0,$$

mentre per gli indici i dispari, si ha

$$s(x_i) = [f(x_i) - q(x_i)] + d_i > 0.$$

Perciò il polinomio $s(x)$ di grado al più n assume $n+2$ volte valori di segno opposto, e quindi ha almeno $n+1$ zeri. Ne segue che $s(x)$ deve essere identicamente nullo e quindi $p_n(x) \equiv q(x)$, cioè $|f(x_i) - q(x_i)| = d_i$, e questo è assurdo per la (44). ∎

6.29 Definizione. Sia $p_n(x) \in \mathcal{P}_n$. I punti

$$x_0, x_1, \ldots, x_k, \quad \text{con} \quad a \le x_0 < x_1 < \ldots < x_k \le b,$$

sono detti di *equioscillazione* per $p_n(x)$ se

$$f(x_i) - p_n(x_i) = (-1)^i d, \quad \text{per } i = 0, \ldots, k, \quad \text{dove} \quad |d| = \|f - p_n\|_\infty. \ ∎$$

È ora possibile dare il teorema di unicità del polinomio di approssimazione minimax, tramite una sua forte caratterizzazione.

6.30 Teorema *(di equioscillazione di Chebyshev).* *Sia $f(x) \in C[a,b]$. Un polinomio $p_n^*(x) \in \mathcal{P}_n$ è di approssimazione minimax di $f(x)$ se e solo se esistono almeno $n+2$ punti $x_0^*, x_1^*, \ldots, x_{n+1}^* \in [a,b]$ di equioscillazione per $p_n^*(x)$. Inoltre il polinomio di approssimazione minimax è unico.*

Dim. Sia $p_n(x) \in \mathcal{P}_n$ un polinomio con almeno $n+2$ punti di equioscillazione. Per il teorema 6.28 si ha

$$|d| \leq r^* \leq \|f - p_n\|_\infty,$$

e poiché $|d| = \|f - p_n\|_\infty$, ne segue la sufficienza della condizione. Per dimostrare la necessità della condizione, si fa vedere che se per assurdo la funzione $r^*(x)$ assumesse il valore $\pm d$ con segno alternato in un numero k di punti $x_0, x_1, \ldots, x_{k-1}$, con $k \leq n+1$, allora esisterebbe un polinomio $q(x)$ tale che $p_n^*(x) + q(x) \in \mathcal{P}_n$ e

$$\|f(x) - \left[p_n^*(x) + q(x)\right]\|_\infty < r^*.$$

Per semplicità si suppone che $d > 0$, ma la dimostrazione è perfettamente analoga nel caso opposto.

Se fosse $k = 1$, si considerino

$$M = \max_{x \in [a,b]} r^*(x) = d \quad \text{e} \quad m = \min_{x \in [a,b]} r^*(x),$$

e si definisca $q(x) = \dfrac{M - |m|}{2}$ (si veda la figura 6.12).

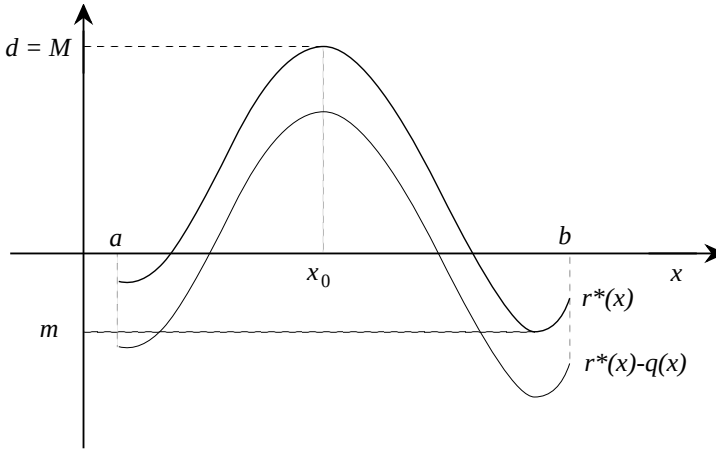


Fig. 6.12 - Caso $k = 1$.

Quindi $p_n^*(x) + q(x) \in \mathcal{P}_n$. Poiché

$$m \leq r^*(x) \leq M = d,$$

si ha

$$-d < m - \frac{M - |m|}{2} \leq r^*(x) - q(x) < d,$$

e quindi

$$\|f(x) - (p_n^*(x) + q(x))\|_\infty = \|r^*(x) - q(x))\|_\infty < d.$$

Se fosse $2 \leq k \leq n+1$, per la continuità della funzione $r^*(x)$ esisterebbero $k+1$ punti $\xi_0, \ldots, \xi_k$, tali che

$$\xi_0 = a \leq x_0 < \xi_1 < x_1 < \ldots < \xi_{k-1} < x_{k-1} \leq b = \xi_k,$$

in cui $r^*(\xi_i) = 0$ per $i = 1, \ldots, k-1$. Posto

$$M_i = \max_{x \in [\xi_i, \xi_{i+1}]} r^*(x) \quad \text{e} \quad m_i = \min_{x \in [\xi_i, \xi_{i+1}]} r^*(x), \quad i = 0, 1, \ldots, k-1,$$

e

$$\epsilon = \min_i \frac{|\,|M_i| - |m_i|\,|}{2},$$

si consideri il polinomio di grado $k-1$

$$s(x) = \prod_{i=1}^{k-1} (\xi_i - x),$$

e siano

$$H = \max_{x \in [a,b]} |s(x)| \quad \text{e} \quad q(x) = \frac{\epsilon\, s(x)}{H}.$$

Il polinomio $q(x)$ (si veda la figura 6.13) è tale che $|q(x)| \leq \epsilon$ e $p_n^*(x) + q(x) \in \mathcal{P}_n$ per $k \leq n+1$.

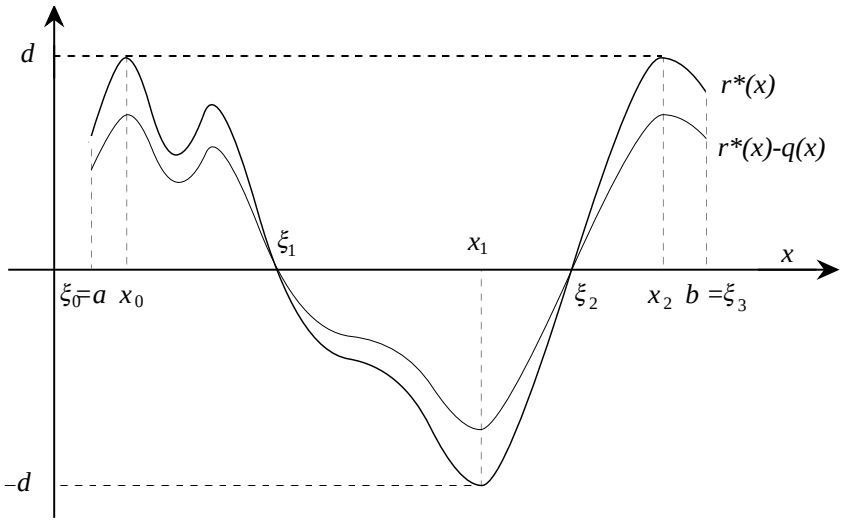


Fig. 6.13 - Caso $k \geq 2$.

Negli intervalli $[\xi_i, \xi_{i+1}]$ "pari", cioè quelli con i pari, si ha

$$0 \le q(x) \le \epsilon, \quad m_i \le r^*(x) \le M_i = d,$$

e quindi

$$m_i - \epsilon \le r^*(x) - q(x) < M_i = d,$$

in quanto $q(x) = 0$ solo nei punti in cui anche $r^*(x) = 0$. In tali intervalli è $-d < m_i \le 0$, per cui

$$m_i - \epsilon \ge m_i - \frac{M_i + m_i}{2} = -\frac{d - m_i}{2} > -d.$$

Ne segue che per x appartenente agli intervalli "pari" è

$$|r^*(x) - q(x)| = |f(x) - (p_n^*(x) + q(x))| < d. \tag{45}$$

Per gli intervalli $[\xi_i, \xi_{i+1}]$ "dispari" la (45) può essere dimostrata in modo del tutto analogo. Allora

$$\|f(x) - (p_n^*(x) + q(x))\|_\infty < d,$$

da cui l'assurdo.

Per dimostrare l'unicità del polinomio di approssimazione minimax, si suppone per assurdo che ne esistano due $p_n(x)$ e $\widetilde{p}_n(x)$. Allora per il teorema 6.4 b) anche il polinomio

$$q(x) = \frac{1}{2}\,(p_n(x) + \widetilde{p}_n(x))$$

è di approssimazione minimax. Indicato con y uno degli $n+2$ punti di $[a, b]$ di equioscillazione per $q(x)$, è

$$\begin{aligned} r^* &= |f(y) - \frac{1}{2}\,(p_n(y) + \widetilde{p}_n(y))| \\ &\le \frac{1}{2}\left[|f(y) - p_n(y)| + |f(y) - \widetilde{p}_n(y)|\right] \le \frac{1}{2}\,(r^* + r^*) = r^*, \end{aligned}$$

per cui

$$f(y) - p_n(y) = f(y) - \widetilde{p}_n(y),$$

in quanto $|f(y) - p_n(y)|$ e $|f(y) - \widetilde{p}_n(y)|$ non possono diventare più grandi di r^*. Perciò $p_n(x)$ e $\widetilde{p}_n(x)$ sono polinomi di grado al più n che coincidono in almeno $n+2$ punti e quindi sono identicamente uguali. ∎

Il teorema di equioscillazione 6.30 è fondamentale per la costruzione dell'approssimazione minimax. Poiché il resto $r^*(x)$ assume $n+2$ massimi

o minimi di segno opposto, esso si annulla, per il teorema di Rolle, in almeno $n+1$ punti $\zeta_0, \ldots, \zeta_n$, quindi il polinomio $p_n^*(x)$ e la funzione $f(x)$ assumono lo stesso valore negli $n+1$ punti ζ_i, $i = 0, 1, \ldots, n$. Perciò il polinomio $p_n^*(x)$ è il polinomio di interpolazione di grado al più n della funzione $f(x)$ nei punti ζ_i, $i = 0, 1, \ldots, n$, che però non sono noti a priori. Ad esso si possono applicare i risultati ottenuti nella teoria dell'interpolazione: in particolare se $f(x) \in C^{n+1}[a, b]$, allora

$$r^*(x) = (x - \zeta_0) \ldots (x - \zeta_n) \, \frac{f^{(n+1)}(\eta)}{(n+1)!}, \quad \eta \in (a, b).$$

Ne segue che se $f^{(n+1)}(x) \neq 0$ per $x \in (a, b)$, la funzione $r^*(x)$ ha esattamente $n + 2$ punti di equioscillazione, compresi gli estremi a e b.

6.31 Esempio *(approssimazione minimax lineare).* Sia $f(x) \in C^2[a, b]$ e sia $f''(x) \neq 0$ in (a, b). Il polinomio $p_1^*(x) = a_1 x + a_0$ di approssimazione minimax è tale che $r^*(x)$ assume massimo e minimo in 3 punti distinti x_0^*, x_1^*, x_2^* di $[a, b]$. Poiché $f''(x) \neq 0$ per $x \in (a, b)$, risulta

$$a = x_0^* < x_1^* < x_2^* = b,$$

e x_1^* è tale che

$$(r^*)'(x_1^*) = 0. \tag{46}$$

Poiché $r^*(x) = f(x) - a_1 x - a_0$, dalla (46) si ottiene l'equazione

$$f'(x_1^*) = a_1.$$

Imponendo poi le condizioni di equioscillazione nei tre punti a, x_1^*, b si ottengono le altre 3 equazioni

$$\begin{cases} f(a) - a_1 a - a_0 = d \\ f(x_1^*) - a_1 x_1^* - a_0 = -d \\ f(b) - a_1 b - a_0 = d, \end{cases}$$

che insieme alla (46) danno per x_1^*, a_1, a_0 e d le espressioni

$$\begin{aligned} f'(x_1^*) &= \frac{f(b) - f(a)}{b - a}, \qquad (47) \\ a_1 &= \frac{f(b) - f(a)}{b - a}, \\ a_0 &= \frac{f(a) + f(x_1^*)}{2} - \frac{f(b) - f(a)}{2(b - a)} \, (a + x_1^*), \\ d &= \frac{f(a) - f(x_1^*)}{2} - \frac{f(b) - f(a)}{2(b - a)} \, (a - x_1^*). \end{aligned}$$

Una volta determinato dalla (47) il punto x_1^* in cui la funzione $f(x)$ ha la derivata uguale al valore del rapporto incrementale dal punto a al punto b, è facile ricavare i coefficienti a_1 e a_0 del polinomio cercato. Poiché $f''(x) \neq 0$ in (a, b), il punto x_1^* esiste ed è unico nell'intervallo (a, b); la sua determinazione può non essere facile e può richiedere un metodo approssimato di risoluzione di equazioni.

Se $f(x) = \mathrm{e}^x$, $a = 0$, $b = 1$, il punto x_1^* è la soluzione dell'equazione

$$\mathrm{e}^x = \mathrm{e} - 1,$$

per cui

$$x_1^* = 0.5413249, \quad a_1^* = 1.718282, \quad a_0^* = 0.8940666, \quad d = 0.1059334.$$

Il polinomio minimax di grado 1 per la funzione $f(x) = \mathrm{e}^x$, $x \in [0, 1]$, è allora

$$p_1^*(x) = 1.718282\, x + 0.8940666.$$

Nella figura 6.14 è riportato il grafico del resto $r^*(x)$.

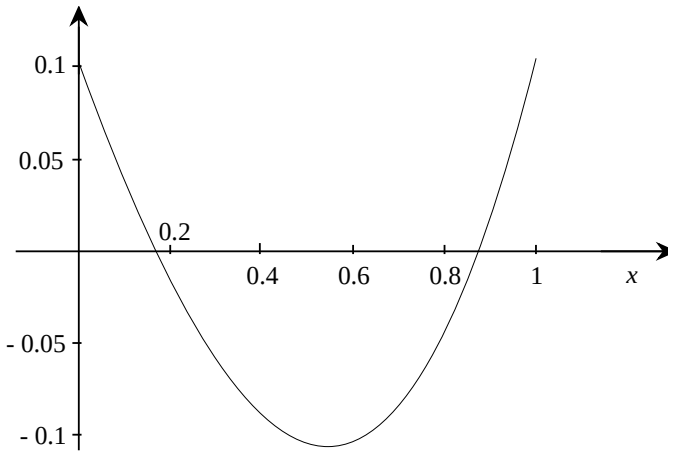


Fig. 6.14 - Resto dell'approssimazione lineare minimax della funzione $f(x) = \mathrm{e}^x$.

Se $f(x) = \sqrt{x}$, $a = 1/16$, $b = 1$, il punto x_1^* è la soluzione dell'equazione

$$\frac{1}{2\sqrt{x}} = \frac{1 - 1/4}{1 - 1/16},$$

per cui $x_1^* = 25/64 = 0.390625$, $a_1^* = 4/5 = 0.8$, $a_0^* = 41/160 = 0.25625$ e $d = -0.05625$.

Il polinomio minimax di grado 1 per la funzione $f(x) = \sqrt{x}$, $x \in [1/16, 1]$ è allora

$$p_1^*(x) = 0.8\,x + 0.25625.$$

Nella figura 6.15 è riportato il grafico del resto $r^*(x)$. ∎

Fig. 6.15 - Resto dell'approssimazione lineare minimax della funzione $f(x) = \sqrt{x}$.

6.32 Definizione. Si dice che il resto $r^*(x)$ del polinomio di approssimazione minimax è una funzione *standard* quando ha esattamente $n + 2$ punti, compresi gli estremi a e b dell'intervallo, di massimo o minimo locale. ∎

Per quanto visto, se $f(x) \in C^{n+1}[a, b]$ e $f^{(n+1)}(x) \neq 0$ in $[a, b]$, allora il resto risulta standard, come nel caso delle due funzioni dell'esempio 6.31. Se invece $f^{(n+1)}(x) = 0$ in almeno un punto di $[a, b]$, allora è possibile che il resto non sia standard, e questo può creare delle complicazioni nella determinazione dell'approssimazione minimax.

6.33 Esempio. La funzione $f(x) = x^3$ è tale che $f''(x) = 0$ in un punto dell'intervallo $[-1, 1]$. Non è quindi possibile dire se il resto $r^*(x)$ dell'approssimazione minimax lineare di $f(x)$ su $[-1, 1]$ è una funzione standard oppure no. Supponendo che il resto sia standard, cioè che i punti di massimo o minimo locale di $r^*(x)$ siano tali che

$$a = x_0^* < x_1^* < x_2^* = b,$$

si applica la tecnica dell'esempio 6.31, ottenendo il sistema non lineare

$$\begin{cases} 3(x_1^*)^2 = a_1 \\ -1 + a_1 - a_0 = d \\ (x_1^*)^3 - a_1 x_1^* - a_0 = -d \\ 1 - a_1 - a_0 = d \end{cases}$$

da cui si ricavano le due soluzioni:

$$x_1^* = \frac{1}{\sqrt{3}}\,, \quad a_1 = 1, \quad a_0 = -\frac{1}{3\sqrt{3}}\,, \quad d = \frac{1}{3\sqrt{3}}\,,$$

e

$$x_1^* = -\frac{1}{\sqrt{3}}\,, \quad a_1 = 1, \quad a_0 = \frac{1}{3\sqrt{3}}\,, \quad d = -\frac{1}{3\sqrt{3}}\,.$$

Poiché il polinomio di approssimazione minimax è unico, le due soluzioni trovate non possono riferirsi ad esso e il resto non è una funzione standard. Si suppone allora che dei tre punti di massimo o minimo locale due siano interni all'intervallo, cioè ad esempio

$$a < x_0^* < x_1^* < x_2^* = b.$$

Al posto della (46) si ottengono allora le due equazioni

$$3(x_0^*)^2 = a_1 \quad \text{e} \quad 3(x_1^*)^2 = a_1,$$

da cui segue che $x_0^* = -x_1^*$. Imponendo le condizioni di equioscillazione si hanno le tre equazioni

$$\begin{cases} (x_0^*)^3 - a_1 x_0^* - a_0 = d \\ (x_1^*)^3 - a_1 x_1^* - a_0 = -d \\ 1 - a_1 - a_0 = d. \end{cases}$$

Dalla prima e seconda equazione segue che $a_0 = 0$. Per sostituzione risulta che x_1^* è la soluzione dell'equazione

$$2x^3 + 3x^2 - 1 = 0$$

che appartiene all'intervallo $[0, 1]$, cioè

$$x_1^* = \frac{1}{2}\,,$$

per cui $a_1^* = \dfrac{3}{4}$, $a_0^* = 0$, $x_0^* = -\dfrac{1}{2}$, $d = \dfrac{1}{4}$, e quindi

$$p_1^*(x) = \frac{3}{4}\,x.$$

È inoltre facile verificare che $r^*(-1) = -d$, cioè i punti di massimo o minimo locale in questo esempio sono 4, anziché 3, due punti interni e due estremi dell'intervallo, come risulta dalla figura 6.16.

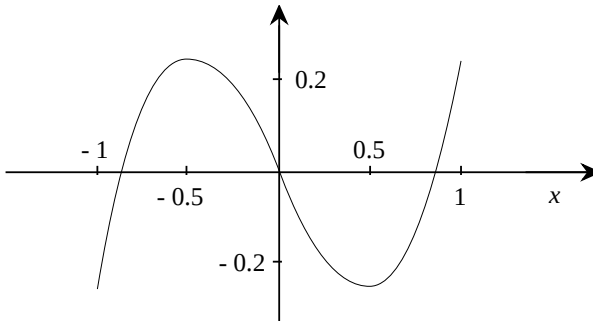


Fig. 6.16 - Resto dell'approssimazione lineare minimax della funzione $f(x) = x^3$.

Volendo invece determinare il polinomio di approssimazione minimax di grado al più 2 per la stessa funzione $f(x) = x^3$ nello stesso intervallo $[-1, 1]$, non vi sono difficoltà, perché $f'''(x) \neq 0$ in tutto l'intervallo, e quindi il resto è una funzione standard. D'altra parte il polinomio di primo grado sopra determinato è tale che $p_1^*(x) \in \mathcal{P}_2$, ha 4 punti di oscillazione compresi gli estremi. Per l'unicità del polinomio di migliore approssimazione è $p_2^*(x) = p_1^*(x)$. ∎

Perciò, a parte il caso in cui $f^{(n+1)}(x) \neq 0$, per $x \in [a, b]$, non è possibile dire a priori se il resto sarà una funzione standard. Il seguente teorema assicura la convergenza ad $f(x)$ del polinomio di approssimazione minimax al crescere del grado n.

6.34 Teorema. *Sia* $f(x) \in C[a, b]$. *Posto*

$$r_n^* = \|f - p_n^*\|_\infty,$$

la successione $\{r_n^*\}, n \in \mathbf{N}$ *è monotona non crescente e*

$$\lim_{n\to\infty} r_n^* = 0.$$

Dim. La monotonia della successione $\{r_n^*\}, n \in \mathbf{N}$ discende direttamente dal teorema 6.4. Per il teorema di Weierstrass si ha che, comunque si fissi una costante $\epsilon > 0$, esistono un intero m e un polinomio $q_m(x)$ di grado al più m, tale che

$$|f(x) - q_m(x)| \leq \epsilon$$

per ogni $x \in [a, b]$. Si ha quindi

$$r_m^* = \|f(x) - p_m^*(x)\|_\infty \leq \|f(x) - q_m(x)\|_\infty \leq \epsilon.$$

Per la monotonia della successione r_n^*, ne segue che per ogni $\epsilon > 0$, esiste un intero m tale che per ogni $n \geq m$ è $r_n^* \leq \epsilon$. ∎

La proprietà di convergenza illustrata nel teorema 6.34 non dà però alcuna informazione sulla velocità con cui la successione degli r_n^* tende a zero.

Se $f(x) = \arcsin x$ e $[a, b] = [-1, 1]$, risulta che r_n^* converge più lentamente di $1/n^2$ (si veda l'esercizio 6.52), per cui per ottenere un'approssimazione dell'ordine di 10^{-6} si deve determinare un polinomio di grado maggiore di 10^3. In questo caso quindi, per la lentezza della convergenza della successione degli r_n^*, non conviene approssimare la funzione in questo modo. Se la funzione $f(x)$ ha una maggiore regolarità, la successione degli r_n^* converge assai più rapidamente, come risulta dal seguente teorema, per la cui dimostrazione si rimanda a [30].

6.35 Teorema *(di Jackson).* *Se $f(x) \in C^k[a, b]$ e se $n > k$, esiste una costante γ, indipendente da n, tale che*

$$r_n^* \leq \frac{\gamma \|f^{(k)}\|_\infty}{n^k} .$$ ∎

Si noti che se la funzione $f(x)$ è analitica (cioè infinitamente derivabile e sviluppabile in serie di potenze), la successione degli r_n^* converge più rapidamente della successione $\dfrac{1}{n^k}$ per ogni $k \geq 1$.

Il seguente teorema dà una limitazione inferiore e superiore di r_n^* nel caso in cui $f^{(n+1)}(x)$ non cambi segno in $[a, b]$.

6.36 Teorema. *Sia $f(x) \in C^{n+1}[a, b]$; se esistono due costanti m e M, con $0 \leq m < M$, tali che*

$$m \leq f^{(n+1)}(x) \leq M, \quad \textit{oppure} \quad m \leq -f^{(n+1)}(x) \leq M, \quad \textit{per } x \in [a, b],$$

allora

$$\frac{m(b-a)^{n+1}}{2^{2n+1}(n+1)!} \leq r_n^* \leq \frac{M(b-a)^{n+1}}{2^{2n+1}(n+1)!} .$$

Per la dimostrazione si veda l'esercizio 6.42. ∎.

Il teorema 6.36, consente di determinare con sufficiente precisione il grado del polinomio che fornisce l'approssimazione minimax con accuratezza prefissata.

6.37 Esempio. Si applica il teorema 6.36 per determinare il grado del polinomio di approssimazione minimax della funzione $f(x) = \mathrm{e}^x$ per $x \in [0,1]$, tale che $r_n^* \leq 0.5\ 10^{-6}$. Poiché $m = 1$, $M = \mathrm{e}$, risulta

$$\frac{m(b-a)^{n+1}}{2^{2n+1}(n+1)!} \approx 0.678\ 10^{-6} \quad \text{per} \quad n = 5,$$
$$\frac{M(b-a)^{n+1}}{2^{2n+1}(n+1)!} = \frac{\mathrm{e}}{2^{2n+1}(n+1)!} \leq 0.5\ 10^{-6} \quad \text{per} \quad n = 6.$$

Occorre quindi fissare $n = 6$ e risulta

$$0.242\ 10^{-7} \leq r_6^* \leq 0.658\ 10^{-7}. \qquad \blacksquare$$

Dal teorema 6.36 risulta che per un fissato n, al diminuire dell'ampiezza dell'intervallo, segue una drastica riduzione dell'errore. Quindi per ottenere polinomi di approssimazione di grado basso occorre ridurre l'intervallo. Ad esempio, per ottenere il polinomio di approssimazione minimax della funzione $f(x) = \sin x$ tale che $r_n^* \leq 0.5\ 10^{-6}$, deve essere $n = 7$ se $[a,b] = [0, \pi/2]$ e $n = 5$ se $[a,b] = [0, \pi/4]$. Si vedano gli esercizi 6.81 - 6.86 per la riduzione degli intervalli nel caso di alcune funzioni fra le più comuni.

6. Algoritmo di Remez

La determinazione effettiva di un polinomio di approssimazione minimax potrebbe in via teorica essere ottenuta con un procedimento analogo a quello seguito nell'esempio 6.31 per il caso $n = 1$. Si dovrebbero però risolvere sistemi di equazioni non lineari più complicate delle (47), che richiedono, salvo casi particolari, l'utilizzazione di metodi iterativi assai onerosi. D'altra parte, a differenza di quanto accade per i polinomi di approssimazione in norma 2, non esiste alcuna procedura diretta che fornisca i coefficienti del polinomio di approssimazione minimax e non vi è alcuna relazione fra i coefficienti dei polinomi minimax $p_n^*(x)$ e $p_{n+1}^*(x)$. Sono stati perciò studiati specifici metodi iterativi: uno di tali metodi, che va sotto il nome di *(secondo) algoritmo di Remez*, costruisce, a partire da un vettore iniziale $\mathbf{x}^{(0)}$, una successione di vettori

$$\mathbf{x}^{(k)}, \quad k = 1, 2, \ldots,$$

di $n+2$ componenti che converge al vettore dei punti di equioscillazione. In teoria il vettore $\mathbf{x}^{(0)}$ può essere arbitrario, in pratica conviene sceglierlo opportunamente. La descrizione del metodo è articolata nelle seguenti parti:

a) descrizione del procedimento con cui si costruisce il vettore $\mathbf{x}^{(k+1)}$ a partire dal vettore $\mathbf{x}^{(k)}$,
b) scelta del vettore iniziale $\mathbf{x}^{(0)}$,
c) criteri di arresto,
d) convergenza del metodo,
e) note di implementazione.

a) Sia $\mathbf{x}^{(k)} = (x_0^{(k)}, x_1^{(k)}, \ldots, x_{n+1}^{(k)})$ un vettore di $n+2$ componenti, tale che

$$a \leq x_0^{(k)} < x_1^{(k)} < \ldots < x_{n+1}^{(k)} \leq b$$

(se il resto è standard, conviene porre subito $x_0^{(k)} = a$ e $x_{n+1}^{(k)} = b$). Si risolva il sistema lineare nelle $n+2$ incognite $a_0^{(k)}, a_1^{(k)}, \ldots, a_n^{(k)}, d^{(k)}$

$$\sum_{j=0}^{n} a_j^{(k)} \left(x_i^{(k)}\right)^j + (-1)^i d^{(k)} = f(x_i^{(k)}), \quad i = 0, \ldots, n+1, \tag{48}$$

la cui soluzione è unica (si veda l'esercizio 6.46). I numeri $a_0^{(k)}, a_1^{(k)}, \ldots, a_n^{(k)}$ così determinati sono i coefficienti del polinomio $p_n^{(k)}(x) \in \mathcal{P}_n$

$$p_n^{(k)}(x) = a_n^{(k)} x^n + a_{n-1}^{(k)} x^{n-1} + \ldots + a_0^{(k)},$$

tale che, indicato con

$$r^{(k)}(x) = f(x) - p_n^{(k)}(x)$$

il resto corrispondente, è

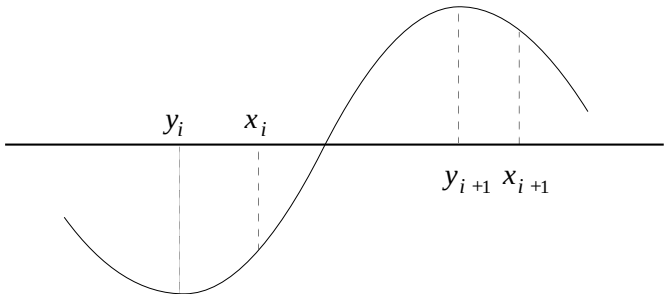
$$r^{(k)}(x_i^{(k)}) = f(x_i^{(k)}) - p_n^{(k)}(x_i^{(k)}) = (-1)^i d^{(k)}.$$

Il polinomio $p_n^{(k)}(x)$ oscilla quindi $n+2$ volte in $[a, b]$. Se i punti $x_i^{(k)}$, $i = 0, \ldots, n+1$ fossero tutti punti di massimo o di minimo di $r^{(k)}(x)$, allora $p_n^{(k)}(x)$ sarebbe il polinomio di migliore approssimazione. Naturalmente questo in generale non sarà vero, per cui è possibile determinare $n+2$ punti di massimo o minimo locale

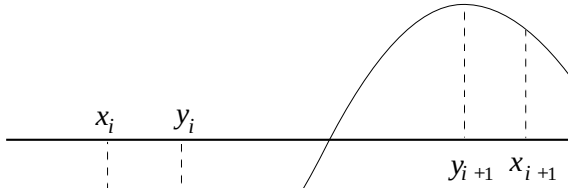
$$a \leq y_0 < y_1 < \ldots < y_{n+1} \leq b$$

di $r^{(k)}(x)$ in $[a, b]$, tali che $r^{(k)}(y_i)$ abbia lo stesso segno di $r^{(k)}(x_i^{(k)})$. Questa condizione è fondamentale perché garantisce che il resto nei punti y_i , $i = 0, 1, \ldots, n+1$, abbia ancora segno alternato. Le seguenti figure illustrano il procedimento di scelta dei punti y_i (per semplicità si è omesso l'indice k).

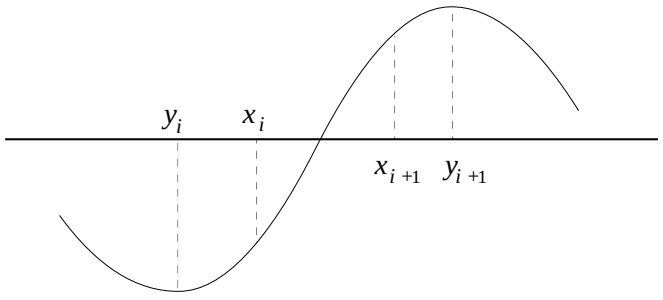
1^o caso: vi è un solo punto di massimo o minimo locale di $r(x)$ compreso fra x_i e x_{i+1}.



2^o caso: vi sono due punti di massimo o minimo locale di $r(x)$ compresi fra x_i e x_{i+1}.



3^o caso: non vi sono punti di massimo o minimo locale di $r(x)$ compresi fra x_i e x_{i+1}.



Se fra x_i e x_{i+1} vi sono più di due punti di massimo o minimo locale, si scelgono i punti corrispondenti a massimi più alti o a minimi più bassi. I punti estremi $x_0 = a$ e $x_{n+1} = b$ possono essere sostituiti da $y_0 \neq a$ e $y_{n+1} \neq b$ solo nel caso che il resto non sia standard.

Si ottiene così il vettore

$$\mathbf{x}^{(k+1)} = (y_0, y_1, \ldots, y_{n+1}).$$

b) Come risulta dal successivo teorema di convergenza 6.38, il vettore iniziale $\mathbf{x}^{(0)}$ può essere arbitrario, purché risulti $d^{(0)} \neq 0$ nel sistema (48). Si possono scegliere $n+2$ punti equidistanti nell'intervallo, con $x_0^{(0)} = a$ e $x_{n+1}^{(0)} = b$, ma come si vedrà nel prossimo paragrafo, una scelta migliore dei punti $x_i^{(0)}$ è data da

$$x_i^{(0)} = \frac{b-a}{2} \cos \frac{(n+1-i)\pi}{n+1} + \frac{a+b}{2}, \quad i = 0, 1, \ldots, n+1.$$

Questi punti $x_i^{(0)}$ sono, in ordine crescente, i punti di massimo o minimo dell'$(n+1)$-esimo polinomio di Chebyshev di 1^a specie, definito nell'intervallo $[a, b]$, estremi compresi.

c) Poiché i punti y_i sono di massimo o di minimo locale di $r^{(k)}(x)$, ne segue che

$$|r^{(k)}(y_i)| \geq |d^{(k)}|, \quad i = 0, \ldots, n+1,$$

in quanto $d^{(k)}$ è un valore assunto da $r^{(k)}(x)$ nei punti $x_i^{(k)}$. La condizione di oscillazione di segno di $r^{(k)}(x)$ imposta nei punti y_i, è tale che i punti soddisfino le ipotesi del teorema 6.28 (de la Vallée-Poussin). Quindi indicato con

$$m^{(k)} = \min_{i=0,\ldots,n+1} |r^{(k)}(y_i)|,$$

si ha che $m^{(k)} \leq r^*$. D'altra parte, posto

$$M^{(k)} = \max_{i=0,\ldots,n+1} |r^{(k)}(y_i)|,$$

si ha

$$M^{(k)} = \|f - p_n^{(k)}\|_\infty \geq r^*.$$

Un criterio di arresto per l'iterazione è che il rapporto $M^{(k)}/m^{(k)}$ sia sufficientemente vicino a 1, cioè

$$\left| \frac{M^{(k)}}{m^{(k)}} - 1 \right| < \epsilon,$$

dove ϵ è una quantità piccola prefissata. È opportuno anche fissare un numero massimo di iterazioni. Si possono utilizzare altri criteri di arresto, sfruttando ad esempio la differenza fra i vettori $\mathbf{x}^{(k)}$ e $\mathbf{x}^{(k+1)}$ oppure fra i coefficienti di due successivi polinomi $p_n^{(k)}(x)$ calcolati.

d) I seguenti teoremi per la cui dimostrazione si vedano gli esercizi 6.47 e 6.48, garantiscono la convergenza del metodo di Remez, che sotto ipotesi assai generali, è molto rapida, addirittura quadratica.

6.38 Teorema. *Se $f(x) \in C[a,b]$, il metodo di Remez è convergente per ogni scelta del punto iniziale $\mathbf{x}^{(0)}$, purché nel sistema (48) risulti $d^{(0)} \neq 0$.* ∎

Il teorema seguente valuta la velocità di convergenza nell'ipotesi che il resto $r^*(x)$ sia standard. Un teorema analogo vale anche sotto ipotesi meno rigide [30].

6.39 Teorema. *Sia $f(x) \in C^2[a,b]$. Se $r^*(x)$ è standard e tale che*

$$\frac{d^2 r^*(x)}{dx^2}\Big|_{x=x_i^*} \neq 0, \quad per \quad i = 0, 1, \ldots, n+1,$$

esiste una costante $\gamma \neq 0$, tale che

$$\frac{|\ \|r^{(k+1)}(x)\|_\infty - r^*\ |}{|\ \|r^{(k)}(x)\|_\infty - r^*\ |^2} \leq \gamma, \quad per \quad k = 1, 2, \ldots$$ ∎

e) La matrice del sistema (48) è, a meno di una colonna, una matrice di Vandermonde, e quindi può essere malcondizionata. Per questo motivo e per ridurre il costo computazionale, non conviene determinare $p_n^{(k)}(x)$ risolvendo il sistema (48) ma sfruttare le tecniche di costruzione dei polinomi di interpolazione nel modo seguente: siano $q(x)$ e $s(x) \in \mathcal{P}_{n+1}$ tali che

$$q(x_i^{(k)}) = f(x_i^{(k)}), \quad s(x_i^{(k)}) = (-1)^i, \quad \text{per } i = 0, 1, \ldots, n+1.$$

Indicato con $d^{(k)}$ il rapporto fra i coefficienti di grado massimo di $q(x)$ e di $s(x)$, si ha

$$p_n^{(k)}(x) = q(x) - d^{(k)} s(x).$$

Infatti $p_n^{(k)}(x) \in \mathcal{P}_n$ ed è tale che

$$p_n^{(k)}(x_i^{(k)}) + (-1)^i d^{(k)} = f(x_i^{(k)}).$$

Se per $k = 0$ il polinomio $q(x)$ fosse di grado minore di $n+1$, risulterebbe $d^{(0)} = 0$, e sarebbe necessario scegliere punti $x_i^{(0)}$ diversi.

Il principale problema del metodo di Remez consiste nella determinazione dei punti y_i di massimo o di minimo di $r^{(k)}(x)$. Una possibile via da seguire è quella di approssimare gli zeri della derivata di $r^{(k)}(x)$ se questa non è una funzione troppo complicata, applicando un metodo iterativo a convergenza garantita, come il metodo di bisezione o quello di falsa posizione. Questo modo di procedere è facilitato dal fatto che dopo le prime iterazioni i punti y_i non potranno trovarsi molto distanti dai punti $x_i^{(k)}$.

Fissata una tolleranza per la condizione di arresto, il numero delle iterazioni richieste dal metodo di Remez varia a seconda del criterio di arresto scelto e della precisione con cui sono determinati i punti $y_0, \ldots, y_{n+1}$.

6.40 Esempio. Si vuole determinare con il metodo di Remez il polinomio di grado al più 3 di approssimazione minimax per la funzione $f(x) = \mathrm{e}^x$ nell'intervallo $[0, 1]$. Dal teorema 6.36 si ha che

$$0.325\ 10^{-3} \leq r_3^* \leq 0.885\ 10^{-3}.$$

Si considerano inizialmente i punti

$$\begin{aligned}
x_0^{(0)} &= \frac{1}{2}\,(1 + \cos\pi) = 0,\\
x_1^{(0)} &= \frac{1}{2}\,\Big(1 + \cos\frac{3}{4}\,\pi\Big) = 0.1464466,\\
x_2^{(0)} &= \frac{1}{2}\,\Big(1 + \cos\frac{1}{2}\,\pi\Big) = 0.5,\\
x_3^{(0)} &= \frac{1}{2}\,\Big(1 + \cos\frac{1}{4}\,\pi\Big)0.8535534,\\
x_4^{(0)} &= \frac{1}{2}\,(1 + \cos 0) = 1.
\end{aligned}$$

Poiché $f^{(4)}(x) \neq 0$ in $[0, 1]$, il resto risulta una funzione standard: perciò si assumerà in ogni iterazione $x_0^{(k)} = 0$ e $x_4^{(k)} = 1$. Determinando $p_3^{(0)}(x)$ per mezzo del polinomio di interpolazione di Newton, si ha

$$a_3^{(0)} = 0.2799751,\ a_2^{(0)} = 0.4217160,\ a_1^{(0)} = 1.016591,\ a_0^{(0)} = 0.9994565$$

e $d^{(0)} = 0.5434368\ 10^{-3}$, da cui risulta

$$r^{(0)}(x) = \mathrm{e}^x - 0.2799751x^3 - 0.4217160x^2 - 1.016591x - 0.9994565.$$

I punti di massimo o minimo di $r^{(0)}(x)$ sono

$$y_1 = 0.1525799,\ y_2 = 0.5124530,\ y_3 = 0.8598699,$$

e si ottiene

$$\Big|\frac{M^{(0)}}{m^{(0)}} - 1\Big| \approx 0.497\ 10^{-2}.$$

Assumendo $x_1^{(1)} = y_1$, $x_2^{(1)} = y_2$, $x_3^{(1)} = y_3$, e ripetendo il calcolo si ottiene

$$a_3^{(1)} = 0.2799765,\ a_2^{(1)} = 0.4217030,\ a_1^{(1)} = 1.016602,\ a_0^{(1)} = 0.9994552$$

e $d^{(1)} = 0.5447914\ 10^{-3}$. I punti di massimo o minimo di $r^{(1)}(x)$ sono

$$y_1 = 0.1526980,\ y_2 = 0.5124711,\ y_3 = 0.8597686,$$

e si ottiene

$$\left|\frac{M^{(1)}}{m^{(1)}} - 1\right| \approx 0.833\ 10^{-6}.$$

Le successive iterazioni non modificano ulteriormente le prime 7 cifre dei coefficienti ottenuti. Il polinomio $p_3^*(x)$ di approssimazione minimax di e^x nell'intervallo $[0, 1]$ è perciò

$$p_3^*(x) = 0.2799765x^3 + 0.421703x^2 + 1.016602x + 0.9994552,$$

e risulta $r_3^* = 0.545\ 10^{-3}$. Nella figura 6.17 è riportato il grafico del resto $f(x) - p_3^*(x)$. ∎

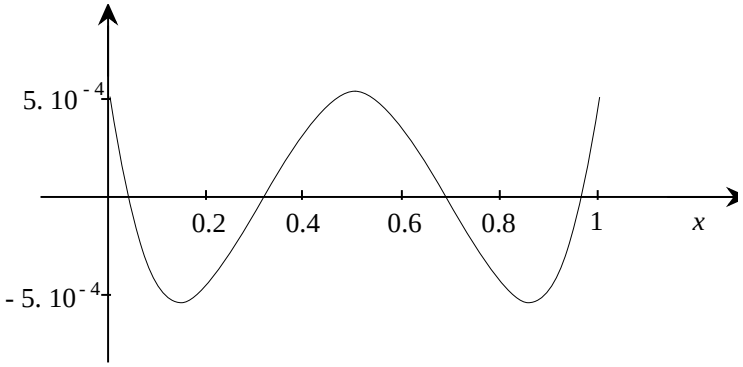


Fig. 6.17 - Resto dell'approssimazione minimax di terzo grado della funzione $f(x) = \mathrm{e}^x$.

Il polinomio di approssimazione minimax $p_n^*(x)$ potrebbe essere espresso, invece che come combinazione lineare dei polinomi di base $\phi_j(x) = x^j$, $j = 0, 1, \ldots, n$, come combinazione degli elementi di una qualsiasi base ortogonale per lo spazio dei polinomi. In particolare si potrebbe scegliere la base dei polinomi di Chebyshev di 1^a specie:

$$p_n^*(x) = \frac{\beta_0^*}{2} + \sum_{j=1}^{n} \beta_j^* T_j(x).$$

Per determinare i coefficienti β_j^* si può usare il metodo di Remez sostituendo al posto del sistema (48), il sistema, solitamente meglio condizionato

$$\frac{\beta_0^{(k)}}{2} + \sum_{j=1}^{n} \beta_j^{(k)} T_j(x_i^{(k)}) + (-1)^i d^{(k)} = f(x_i^{(k)}), \quad i = 0, 1, \ldots, n+1. \quad (49)$$

7. Approssimazione quasi minimax

Il calcolo del polinomio $p_n^*(x)$ di approssimazione minimax richiede, come si è visto, un notevole costo computazionale. Per questo motivo sono stati studiati altri metodi che sfruttano le proprietà dei polinomi di Chebyshev e che con un costo computazionale assai inferiore consentono di determinare polinomi $p_n(x)$ che sono delle stime ragionevoli di $p_n^*(x)$.

Questi metodi, detti di approssimazione *quasi minimax*, sono fondamentalmente quattro:

a) economizzazione,

b) serie di Chebyshev troncata,

c) interpolazione nei nodi di Chebyshev,

d) arresto al primo passo del metodo di Remez.

Se l'intervallo $[a, b]$ su cui si cerca l'approssimazione non coincide con l'intervallo $[-1, 1]$, si applica prima la trasformazione di variabile (33).

a) Il metodo di economizzazione, che viene di solito applicato ad approssimazioni polinomiali ottenute troncando serie di Taylor, si basa sulla seguente proprietà.

Siano $p_n(x)$ un polinomio monico di grado n e $p_{n-1}(x) \in \mathcal{P}_{n-1}$ il polinomio di approssimazione minimax di $p_n(x)$. Poiché il resto $r(x) = p_n(x) - p_{n-1}(x)$ è un polinomio monico di grado n, per il teorema 6.19 deve essere

$$r(x) = \frac{1}{2^{n-1}}\, T_n(x),$$

dove $T_n(x)$ è l'n-esimo polinomio ortogonale di Chebyshev di 1^a specie, per cui risulta

$$p_{n-1}(x) = p_n(x) - \frac{1}{2^{n-1}}\, T_n(x).$$

Sia ora $p_n(x)$ un polinomio di approssimazione di grado n di una funzione $f(x)$ sull'intervallo $[-1, 1]$, tale che

$$\|f(x) - p_n(x)\|_\infty \leq \eta,$$

e sia $\epsilon > \eta$ l'approssimazione richiesta. Il metodo di economizzazione consiste nel sostituire al posto di $p_n(x)$ il polinomio $p_{n-1}(x)$ di minimax di $p_n(x)$ che, per quanto visto sopra, è dato da

$$p_{n-1}(x) = p_n(x) - \frac{a_n}{2^{n-1}}\, T_n(x),$$

dove a_n è il primo coefficiente di $p_n(x)$. Si ha allora

$$\|p_n(x) - p_{n-1}(x)\|_\infty \leq \frac{|a_n|}{2^{n-1}}\,,$$

e

$$\|f(x) - p_{n-1}(x)\|_\infty \leq \eta + \frac{|a_n|}{2^{n-1}}\,.$$

Perciò se

$$\eta + \frac{|a_n|}{2^{n-1}} \leq \epsilon,$$

il polinomio $p_{n-1}(x)$ soddisfa l'approssimazione richiesta.

Questo procedimento si applica soprattutto a serie che convergono lentamente, e può essere riapplicato abbassando ogni volta il grado di 1. Se la funzione $f(x)$ è pari oppure dispari, il procedimento determina l'abbassamento del grado di 2.

6.41 Esempio. Applicando il metodo di economizzazione al polinomio

$$p_{2n+1}(x) = \sum_{j=0}^{n} (-1)^j \, \frac{x^{2j+1}}{(2j+1)!}\,,$$

che approssima la funzione $f(x) = \sin x$, si ottiene il polinomio

$$q_{2n-1}(x) = \sum_{j=0}^{n} (-1)^j \, \frac{x^{2j+1}}{(2j+1)!} \; - \; \frac{(-1)^n}{(2n+1)!\,2^{2n}} \; T_{2n+1}(x).$$

Per $n = 1$ si ha

$$p_3(x) = x - \frac{x^3}{3!}\,,$$

da cui si ottiene l'approssimazione lineare di $\sin x$

$$q_1(x) = p_3(x) + \frac{1}{3!\;4}\; T_3(x) = \frac{7}{8}\; x.$$

Per $x \in [-1, 1]$ risulta

$$\|f(x) - p_3(x)\|_\infty \approx 0.814\; 10^{-2},$$
$$\|f(x) - q_1(x)\|_\infty \approx 0.419\; 10^{-1}.$$

Per $n = 2$ si ha

$$p_5(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}\,,$$
$$q_3(x) = p_5(x) - \frac{1}{5!\;16}\; T_5(x) = \frac{1}{384}\; (383x - 60x^3).$$

Per $x \in [-1, 1]$ risulta

$$\|f(x) - p_5(x)\|_\infty \approx 0.196\; 10^{-3},$$
$$\|f(x) - q_3(x)\|_\infty \approx 0.568\; 10^{-3},$$

e quindi nell'intervallo $[-1,1]$ $q_3(x)$ risulta un'approssimazione migliore in norma ∞ di $p_3(x)$.

Per $n = 3$ si ha

$$p_7(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!},$$

da cui applicando successivamente due volte il procedimento si ottiene

$$q_5(x) = p_7(x) + \frac{1}{7!\ 64}\, T_7(x) = \frac{1}{46080}\,(46079x - 7672x^3 + 368x^5),$$
$$t_3(x) = q_5(x) - \frac{23}{2880}\,\frac{1}{16}\, T_5(x) = \frac{1}{11520}\,(11491x - 1803x^3).$$

Per $x \in [-1,1]$ risulta

$$\|f(x) - p_7(x)\|_\infty \approx 0.273\ 10^{-5},$$
$$\|f(x) - q_5(x)\|_\infty \approx 0.424\ 10^{-5},$$
$$\|f(x) - t_3(x)\|_\infty \approx 0.502\ 10^{-3},$$

e quindi nell'intervallo $[-1,1]$ risulta che $q_5(x)$ è un'approssimazione migliore in norma ∞ di $p_5(x)$ e che $t_3(x)$ è migliore di $q_3(x)$. Nella figura 6.18 sono riportati nell'intervallo $[0,1]$ i grafici dei resti dei tre polinomi di terzo grado ottenuti: $f(x) - p_3(x)$ (con i pallini), $f(x) - q_3(x)$ (con i quadratini neri) e $f(x) - t_3(x)$ (con linea continua). Come si può notare, i grafici di questi ultimi due resti sono praticamente coincidenti. ∎

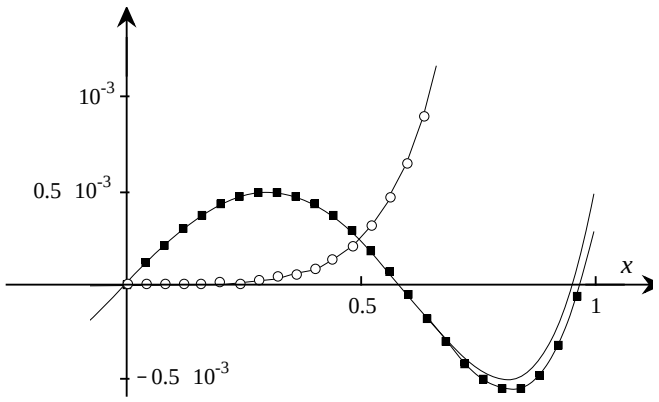


Fig. 6.18 - Resti dei polinomi di terzo grado ottenuti dalla serie di Taylor della funzione $f(x) = \sin x$.

b) La tecnica dei minimi quadrati consente di determinare esplicitamente i coefficienti dei polinomi $p_n(x)$ che minimizzano l'errore in norma 2. Tali

polinomi ovviamente non coincidono con quelli che minimizzano l'errore in norma ∞, ma possono rappresentare anch'essi delle valide approssimazioni. In particolare conviene usare i polinomi ottenuti troncando la serie di Chebyshev, sia per la maggiore semplicità di calcolo, sia perché per il teorema 6.27 i polinomi di Chebyshev convergono rapidamente, oltre che in norma 2, anche in norma ∞ quando la funzione $f(x)$ è derivabile fino ad un ordine elevato.

Posto $x = \cos\theta$, $0 \le \theta \le \pi$, per le (35) e (36) il polinomio $p_n^C(x)$ ottenuto troncando all'$(n+1)$-esimo termine l'espansione di $f(x)$ in polinomi di Chebyshev di 1^a specie è

$$p_n^C(x) = p_n^C(\cos\theta) = \frac{\alpha_0}{2} + \sum_{j=1}^{n} \alpha_j \cos j\theta, \tag{50}$$

dove

$$\alpha_j = \frac{2}{\pi} \int_0^{\pi} f(\cos\theta) \cos j\theta \; d\theta. \tag{51}$$

Se la convergenza è rapida, al crescere del grado i coefficienti dell'espansione tendono rapidamente a zero, per cui il coefficiente del primo termine trascurato può ragionevolmente stimare l'errore commesso.

c) Se gli integrali (51) non possono essere ottenuti per via analitica, occorre approssimarli con una formula di quadratura. Usando la formula dei punti di mezzo con $n+1$ nodi (si veda il paragrafo 4 del capitolo 7), si ottiene il polinomio

$$p_n^+(x) = p_n^+(\cos\theta) = \frac{\alpha_0^+}{2} + \sum_{j=1}^{n} \alpha_j^+ \cos j\theta, \tag{52}$$

dove

$$\alpha_j^+ = \frac{2}{n+1} \sum_{i=0}^{n} f(\cos\theta_i) \cos j\theta_i, \quad \theta_i = \frac{(2i+1)\pi}{2(n+1)}, \quad i = 0, \ldots, n. \tag{53}$$

Sostituendo le (53) nella (52) risulta

$$p_n^+(\cos\theta) = \sum_{i=0}^{n} L_i(\cos\theta) f(\cos\theta_i),$$

dove

$$L_i(\cos\theta) = \frac{2}{n+1} \left[\frac{1}{2} + \sum_{j=1}^{n} \cos j\theta_i \cos j\theta \right].$$

Poiché (si veda l'esercizio 6.23 a))

$$L_i(\cos\theta_k) = \delta_{i,k} \quad \text{per} \quad i, k = 0, \ldots, n,$$

si ha

$$p_n^+(\cos\theta_k) = f(\cos\theta_k),$$

per cui $p_n^+(x)$ è il polinomio di interpolazione di $f(x)$ nei nodi $x_i = \cos\theta_i$, $i = 0, \ldots, n$, detti *punti di Chebyshev*. Il polinomio $p_n^+(x)$ rappresenta una valida alternativa al polinomio $p_n^*(x)$ ed è molto usato per la sua semplicità. Inoltre per il teorema 5.5 è

$$f(x) - p_n^+(x) = \pi_n(x)\ \frac{f^{(n+1)}(\xi)}{(n+1)!}\,, \quad \xi \in (-1, 1),$$

e la scelta dei punti di Chebyshev come nodi è quella che, per il teorema 6.19, minimizza la norma ∞ di $\pi_n(x)$ e si ha

$$\|f - p_n^+\|_\infty \leq\ \frac{\|f^{(n+1)}(x)\|_\infty}{2^n\ (n+1)!}\,.$$

d) Se per approssimare gli integrali (51) si usa la formula di quadratura dei trapezi con $n+2$ nodi (si veda il paragrafo 3 del capitolo 7), si ottiene il polinomio trigonometrico

$$p_n^R(x) = p_n^R(\cos\theta) = \frac{\beta_0}{2}\ + \sum_{j=1}^{n} \beta_j \cos j\theta, \tag{54}$$

dove

$$\begin{gathered}\beta_j = \frac{1}{n+1}\Big[f(\cos\phi_0) + 2\sum_{i=1}^{n} f(\cos\phi_i)\cos j\phi_i + (-1)^j f(\cos\phi_{n+1})\Big], \\ \phi_i = \frac{i\pi}{n+1}\,, \quad i = 0, \ldots, n+1.\end{gathered} \tag{55}$$

Per le (79) e (80) del capitolo 5 il polinomio trigonometrico

$$F(\cos\theta) = p_n^R(\cos\theta) + \frac{\beta_{n+1}}{2}\ \cos(n+1)\theta,$$

dove β_{n+1} è anch'esso definito dalla (55), è di interpolazione nei nodi ϕ_i, $i = 0, \ldots, n+1$, cioè

$$\frac{\beta_0}{2}\ + \sum_{j=1}^{n} \beta_j \cos j\theta_i + \frac{\beta_{n+1}}{2}\ \cos(n+1)\phi_i = f(\cos\phi_i).$$

Si pone $x_i = \cos\phi_i$, ed essendo $\cos(n+1)\phi_i = (-1)^i$, si ha

$$\frac{\beta_0}{2}\ + \sum_{j=1}^{n} \beta_j T_j(x_i) + (-1)^i\ \frac{\beta_{n+1}}{2} = f(x_i), \quad i = 0, \ldots, n+1. \tag{56}$$

Se si pone $k = 0$, $x_i^{(0)} = x_i$, $\beta_j^{(0)} = \beta_j$ per $j = 0, \ldots, n$ e $d^{(0)} = \dfrac{\beta_{n+1}}{2}$, la (56) coincide con la (49), cioè il polinomio $p_n^R(x)$ è quello ottenuto con l'applicazione di un passo dell'algoritmo di Remez scegliendo come punti iniziali gli x_i, $i = 0, \ldots, n$ ed esprimendo il polinomio come combinazione di polinomi di Chebyshev (si noti però che i nodi x_i sono ordinati in modo decrescente, invece che crescente come nell'ordinamento consueto).

Anche il polinomio $p_n^R(x)$ rappresenta una valida alternativa al polinomio $p_n^*(x)$, anche perché i coefficienti β_j possono essere calcolati con algoritmi veloci per il calcolo della trasformata discreta di Fourier (si veda l'esercizio 5.61).

Anche tenendo conto della rapidità di convergenza dell'algoritmo di Remez, il costo computazionale del calcolo dei polinomi $p_n^C(x)$, $p_n^+(x)$ e $p_n^R(x)$ è molto minore di quello del polinomio $p_n^*(x)$. Inoltre i resti dei polinomi ottenuti con uno di questi metodi sono dell'ordine del resto del polinomio di approssimazione minimax, come risulta dal seguente teorema, per la cui dimostrazione si vedano [33] e [44].

6.42 Teorema. *Per $[a, b] = [-1, 1]$, valgono le seguenti maggiorazioni:*

$$\|f - p_n^C\|_\infty \le r^* u_n, \quad \|f - p_n^+\|_\infty \le r^* v_n, \quad \|f - p_n^R\|_\infty \le r^* w_n,$$

dove u_n, v_n e w_n sono funzioni crescenti con n, e asintoticamente è

$$u_n \sim \frac{4}{\pi^2}\ \log n, \quad v_n \sim \frac{2}{\pi}\ \log n, \quad w_n \sim \frac{2}{\pi}\ \log n. \qquad \blacksquare$$

Nella seguente tabella sono riportati i valori di u_n, v_n e w_n per alcuni valori di n.

n	u_n	v_n	w_n
1	2.436	2.414	2.5
2	2.642	2.667	2.667
5	2.961	3.104	3.094
10	3.223	3.489	3.489
50	3.860	4.466	4.466
100	4.139	4.901	4.901

6.43 Esempio. Per calcolare approssimazioni quasi minimax di grado 3 della funzione e^x nell'intervallo $[0, 1]$ si fa il cambiamento di variabile $x = (y + 1)/2$ ottenendo la funzione

$$f(y) = \mathrm{e}^{(y+1)/2}, \quad y \in [-1, 1].$$

a) La serie di Taylor di $f(y)$ è

$$\mathrm{e}^{(y+1)/2} = \sum_{j=0}^{\infty} \frac{(y+1)^j}{2^j\ j!}\,.$$

Troncando al quinto termine ed applicando il metodo di economizzazione si ottiene il polinomio

$$\begin{aligned} q(y) &= \sum_{j=0}^{4} \frac{(y+1)^j}{2^j\ j!} - \frac{1}{3072}\, T_4(y) \\ &= 0.03125\, y^3 + 0.2057292\, y^2 + 0.8229167\, y + 1.648112, \end{aligned}$$

da cui

$$p_3(x) = q(2x-1) = 0.25\, x^3 + 0.4479167\, x^2 + 1.010417\, x + 0.9996744,$$

e risulta

$$\max_{x\in[0,1]} |\mathrm{e}^x - p_3(x)| \approx 0.103\ 10^{-1}.$$

Troncando al sesto termine ed applicando il metodo di economizzazione per due volte si ottiene il polinomio

$$\begin{aligned} \overline{q}(y) &= \sum_{j=0}^{5} \frac{(y+1)^j}{2^j\ j!} - \frac{1}{61440}\, T_5(y) - \frac{1}{2048}\, T_4(y) \\ &= 0.03417969\, y^3 + 0.2096354\, y^2 + 0.8241374\, y + 1.64821, \end{aligned}$$

da cui

$$\overline{p}_3(x) = \overline{q}(2x-1) = 0.2734375\, x^3 + 0.4283854\, x^2 + 1.014811\, x + 0.9995280,$$

e risulta

$$\max_{x\in[0,1]} |\mathrm{e}^x - \overline{p}_3(x)| \approx 0.212\ 10^{-2}.$$

b) Si calcolano i coefficienti α_j, $j = 0, \ldots, 3$, dello sviluppo di $f(y)$ in serie di polinomi di Chebyshev di 1^a specie. Dalla (51), posto $y = \cos\theta$, si ha

$$\alpha_j = \frac{2}{\pi} \int_0^{\pi} \mathrm{e}^{(\cos\theta+1)/2} \cos j\theta\ d\theta, \quad j = 0, \ldots, 3,$$

e si ottengono i valori

$$\alpha_0 = 3.506775,\ \alpha_1 = 0.8503917,\ \alpha_2 = 0.1052087,\ \alpha_3 = 0.008722105.$$

Facendo nella (50) la sostituzione $\theta = \arccos(2x-1)$, risulta

$$\begin{aligned} &p_3^C(x) = 0.2791074\, x^3 + 0.4230086\, x^2 + 1.016112\, x + 0.9994824, \\ &\|f - p_3^C(x)\|_\infty \approx 0.572\ 10^{-3}. \end{aligned}$$

c) Posto $\theta_i = \dfrac{(2i+1)\pi}{8}$, $i = 0, \ldots, 3$, dalla (53) si ha

$$\alpha_0^+ = 3.506774,\ \alpha_1^+ = 0.8503904,\ \alpha_2^+ = 0.1052073,\ \alpha_3^+ = 0.008694202,$$

e facendo nella (52) la sostituzione $\theta = \arccos(2x-1)$, risulta

$$p_3^+(x) = 0.2782145\,x^3 + 0.4243367\,x^2 + 1.015618\,x + 0.9995097,$$
$$\|f - p_3^+\|_\infty \approx 0.603\ 10^{-3}.$$

d) Fissati i punti $\phi_i = \dfrac{i\pi}{4}$, $i = 0, \ldots, 4$, dalla (55) si ha

$$\beta_0 = 3.506775, \quad \beta_1 = 0.8503917, \quad \beta_2 = 0.1052098, \quad \beta_3 = 0.00874922,$$

e facendo nella (54) la sostituzione $\theta = \arccos(2x-1)$, risulta

$$p_3^R(x) = 0.2799751\,x^3 + 0.4217160\,x^2 + 1.016591\,x + 0.9994565,$$
$$\|f - p_3^R\|_\infty \approx 0.547\ 10^{-3}$$

(questo polinomio coincide ovviamente con quello ottenuto al primo passo nell'esempio 6.40).

I massimi moduli dei resti dei polinomi di grado 3 ottenuti con i metodi quasi minimax per l'approssimazione di e^x sull'intervallo $[0, 1]$ sono riportati nella seguente tabella, a confronto con il valore corrispondente del polinomio minimax dell'esempio 6.40.

metodo	resto in norma ∞
minimax	$0.545\ 10^{-3}$
serie di Taylor + 1 procedimento di economizzazione	$0.103\ 10^{-1}$
serie di Taylor + 2 procedimenti di economizzazione	$0.212\ 10^{-2}$
serie di Chebyshev troncata	$0.572\ 10^{-3}$
interpolazione nei nodi di Chebyshev	$0.603\ 10^{-3}$
un passo del metodo di Remez	$0.547\ 10^{-3}$

■

Come conseguenza del teorema 6.42 si ha che se $f(x) \in C^1[a, b]$, allora la successione dei polinomi di interpolazione di grado n, costruiti assumendo come nodi gli zeri dei polinomi di Chebyshev di 1^a specie, converge alla

funzione $f(x)$ su tutto l'intervallo $[-1,1]$. Infatti combinando il teorema di Jackson 6.35 e il teorema 6.42, asintoticamente risulta

$$\|f - p_n^+\|_\infty \leq \gamma \|f'\|_\infty \, \frac{v_n}{n}\,, \quad v_n \sim \frac{2}{\pi}\, \log n.$$

Se invece $f(x) \notin C^1[a,b]$ è solo continua, la convergenza della successione dei polinomi di interpolazione non è garantita (mentre risulta garantita la convergenza in media, cioè in norma 2, si veda l'esercizio 6.54).

6.44 Esempio. Si considera il polinomio di interpolazione della funzione di Runge

$$f(x) = \frac{1}{1+x^2}\,,$$

definita sull'intervallo $[-5,5]$. Come si è visto nell'esempio 5.9, assumendo come nodi dei punti equidistanti, al crescere di n il polinomio di interpolazione approssima sempre peggio la funzione $f(x)$. Assumendo invece come nodi i punti di Chebyshev, che nell'intervallo $[-5,5]$ sono

$$x_i = 5\cos\,\frac{(2i+1)\pi}{2(n+1)}\,, \quad i = 0,1,\ldots,n,$$

si ottiene una successione di polinomi che converge alla funzione $f(x)$. Nella figura 6.19 sono riportati, al variare di n, i valori della norma $r_n = \|f - p_n\|_\infty$ nell'intervallo $[-5,5]$ per i polinomi di interpolazione $p_n(x)$ di grado n, assumendo come nodi punti equidistanti (linea con i pallini) e punti di Chebyshev (linea con i quadratini neri). Mentre nel caso dei nodi equidistanti risulta chiaramente la divergenza della successione r_n, nel caso dei nodi di Chebyshev i resti decrescono per $n \leq 20$. Per valori di n più elevati gli errori di arrotondamento distruggono il carattere di convergenza della successione. ∎

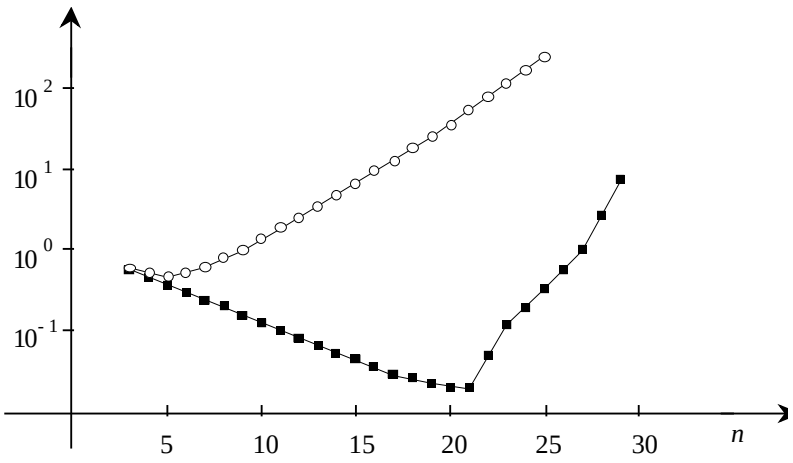


Fig. 6.19 - Errori dei polinomi di interpolazione per la funzione di Runge.

8. Approssimazione minimax rispetto all'errore relativo

La bontà dell'approssimazione $\widetilde{x}$ di un numero reale $x \neq 0$, rappresentato in virgola mobile (*floating point*) viene misurata più frequentemente per mezzo dell'errore relativo

$$\epsilon = \frac{\widetilde{x} - x}{x}$$

che dell'errore assoluto $\widetilde{x} - x$. Così facendo si prende in considerazione il numero delle cifre significative di x, indipendentemente dalla grandezza di x, come si è già visto nel capitolo 2. Le stesse considerazioni si possono applicare all'approssimazione di funzioni. L'approssimazione minimax consente di determinare, con poche modifiche, anche approssimazioni rispetto all'errore relativo.

Il teorema 6.30 di equioscillazione di Chebyshev vale anche quando come funzione di errore si considera il resto relativo

$$r(x) = \frac{f(x) - p_n(x)}{f(x)}\,, \tag{57}$$

dove $f(x) \neq 0$, per $x \in [a, b]$. La dimostrazione è analoga a quella svolta per l'errore assoluto.

Ovviamente il polinomio di approssimazione minimax rispetto all'errore relativo non è lo stesso di quello rispetto all'errore assoluto, anzi i due polinomi possono essere abbastanza diversi.

6.45 Esempio. Nelle stesse ipotesi dell'esempio 6.31, se $f(x) \neq 0$ per $x \in [a, b]$ il polinomio

$$p_1^*(x) = a_1 x + a_0$$

di approssimazione minimax lineare rispetto all'errore relativo è tale che

$$r^*(x) = \frac{f(x) - p_1^*(x)}{f(x)} \tag{58}$$

assume massimo e minimo locale in tre punti distinti di $[a, b]$

$$a = x_0^* < x_1^* < x_2^* = b.$$

Dovendo essere

$$(r^*)'(x_1^*) = 0,$$

ne segue che x_1^* è soluzione dell'equazione

$$(p_1^*)'(x) f(x) - p_1^*(x) f'(x) = 0,$$

cioè

$$a_1[f(x) - xf'(x)] - a_0 f'(x) = 0.$$

Ne segue che x_1^* è tale che

$$\frac{a_0}{a_1} = \frac{f(x_1^*)}{f'(x_1^*)} - x_1^*. \tag{59}$$

Imponendo le condizioni di equioscillazione nei tre punti a, x_1^*, b, si ottengono altre 3 equazioni

$$\begin{cases} f(a) - a_1 a - a_0 = d\ f(a) \\ f(x_1^*) - a_1 x_1^* - a_0 = -d\ f(x_1^*) \\ f(b) - a_1 b - a_0 = d\ f(b). \end{cases} \tag{60}$$

Ad esempio, nel caso della funzione $f(x) = \sqrt{x}$ per $x \in [1/16, 1]$, il sistema formato dalle (59) e (60) è

$$\begin{cases} x_1^* = a_0/a_1 \\ 4 - a_1 - 16a_0 = 4d \\ \sqrt{x_1^*} - a_1 x_1^* - a_0 = -d\sqrt{x_1^*} \\ 1 - a_1 - a_0 = d, \end{cases}$$

la cui soluzione è $x_1^* = 1/4$, $a_0 = 2/9$, $a_1 = 8/9$, $d = -1/9$. Il polinomio minimax di grado 1 rispetto all'errore relativo della funzione $f(x) = \sqrt{x}$ nell'intervallo $[1/16, 1]$ è allora dato da

$$p_1^*(x) = \frac{1}{9}\ (8x + 2).$$

Nella figura 6.20 è riportato il corrispondente andamento del resto relativo (58). Si noti come il polinomio trovato in questo esempio sia alquanto diverso da quello trovato nell'esempio 6.31 rispetto all'errore assoluto. ∎

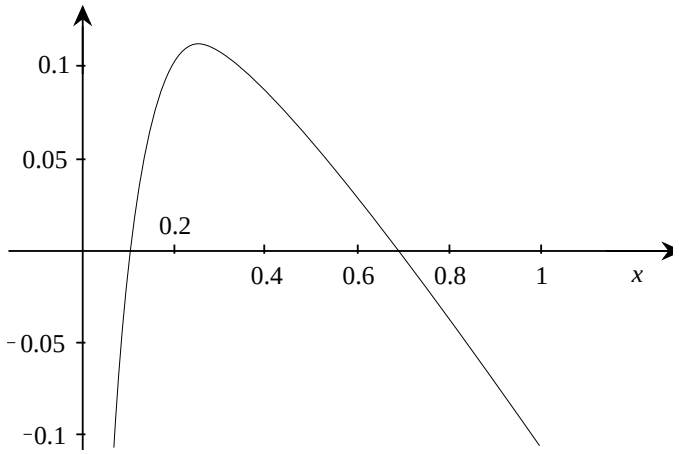


Fig. 6.20 - Resto relativo dell'approssimazione minimax lineare rispetto all'errore relativo della funzione $f(x) = \sqrt{x}$.

Anche l'algoritmo di Remez non richiede modifiche sostanziali. Basta sostituire la (48) con la

$$\sum_{j=0}^{n} a_j^{(k)} \big(x_i^{(k)}\big)^j + (-1)^i f(x_i^{(k)}) d^{(k)} = f(x_i^{(k)}), \quad i = 0, \ldots, n+1, \qquad (61)$$

per ottenere i coefficienti del polinomio $p_n^{(k)}(x) \in \mathcal{P}_n$ tale che

$$\frac{f(x_i) - p_n^{(k)}(x_i)}{f(x_i)} = (-1)^i d^{(k)}.$$

Inoltre i punti $y_i,\ i = 1, \ldots, n$, devono essere i punti di massimo o minimo locale del resto relativo (57). La dimostrazione della convergenza del metodo di Remez così modificato è sostanzialmente uguale a quella del teorema 6.38 (si veda l'esercizio 6.47).

6.46 Esempio. Il polinomio di approssimazione minimax rispetto all'errore relativo di grado al più 3 della funzione $f(x) = \mathrm{e}^x$ nell'intervallo $[0, 1]$ può essere determinato con il metodo di Remez, analogamente a quanto fatto nell'esempio 6.40. Assumendo gli stessi 5 punti iniziali

$$a = x_0^{(0)} < x_1^{(0)} < x_2^{(0)} < x_3^{(0)} < x_4^{(0)} = b,$$

si ottiene dal sistema (61) la soluzione

$$a_3^{(0)} = 0.2715409,\ a_2^{(0)} = 0.4341063,\ a_1^{(0)} = 1.012102,\ a_0^{(0)} = 0.9996901,$$

e

$$d^{(0)} = 0.3099354\ 10^{-3}.$$

I punti di massimo o minimo del resto relativo della prima iterazione

$$r^{(0)}(x) = \frac{\mathrm{e}^x - 0.2715409x^3 - 0.4341063x^2 - 1.012102x - 0.9996901}{\mathrm{e}^x}$$

interni all'intervallo $[0, 1]$ sono

$$y_1 = 0.1284527,\ y_2 = 0.4665217,\ y_3 = 0.8378632.$$

Assumendo $x_1^{(1)} = y_1$, $x_2^{(1)} = y_2$, $x_3^{(1)} = y_3$ e ripetendo il calcolo, dopo altre due iterazioni si ottiene

$$a_3^{(3)} = 0.2713634,\ a_2^{(3)} = 0.4342106,\ a_1^{(3)} = 1.012156,\ a_0^{(3)} = 0.9996789,$$

e

$$d^{(3)} = 0.3211005\ 10^{-3}.$$

Le successive iterazioni non modificano tali coefficienti, per cui si assume che il polinomio di terzo grado di approssimazione minimax rispetto all'errore relativo di $f(x) = \mathrm{e}^x$ nell'intervallo $[0, 1]$ è dato da

$$p_3^*(x) = 0.2713634x^3 + 0.4342106x^2 + 1.012156x + 0.9996789.$$

Come si vede, questo polinomio non differisce molto da quello ottenuto nell'esempio 6.40. ∎

La condizione che $f(x)$ non debba annullarsi nell'intervallo $[a, b]$ è piuttosto pesante, perché esclude la possibilità di approssimazione relativa di funzioni per altri aspetti del tutto regolari, come ad esempio la funzione $\log x$ in un qualunque intervallo che contenga il punto 1. Questa condizione può essere indebolita: infatti si può approssimare con il minimax rispetto all'errore relativo anche una funzione $f(x)$ che nell'intervallo $[a, b]$ abbia un solo zero α di molteplicità 1, cioè tale che il

$$\lim_{x\to\alpha} \frac{f(x)}{x - \alpha}$$

sia finito e diverso da zero. In tal caso si considera la funzione

$$g(x) = \begin{cases} \dfrac{f(x)}{x - \alpha}, & \text{se } x \neq \alpha, \\ \displaystyle\lim_{t\to\alpha} \frac{f(t)}{t - \alpha}, & \text{se } x = \alpha, \end{cases}$$

e si determina il polinomio $p^*_{n-1}(x)$ di migliore approssimazione per $g(x)$. Poiché per $x \neq \alpha$ è

$$\frac{\big|g(x) - p^*_{n-1}(x)\big|}{\big|g(x)\big|} = \Big|\frac{f(x) - (x-\alpha)p^*_{n-1}(x)}{f(x)}\Big|,$$

ne segue che $p^*_n(x) = (x-\alpha)p^*_{n-1}(x)$ è il polinomio di migliore approssimazione minimax di $f(x)$ rispetto all'errore relativo in $[a,b] - \{\alpha\}$.

6.47 Esempio. Nell'intervallo $[1,2]$ la funzione $\log x$ ha uno zero di molteplicità 1. Per determinare il polinomio di secondo grado di approssimazione minimax rispetto all'errore relativo, si considera la funzione

$$g(x) = \begin{cases} \dfrac{\log x}{x-1}, & \text{se } x \neq 1, \\ 1, & \text{se } x = 1. \end{cases}$$

Con il metodo di Remez si ottiene il polinomio lineare

$$p^*_1(x) = -0.3002439\,x + 1.278706$$

di approssimazione minimax rispetto all'errore relativo della $g(x)$ e quindi il polinomio

$$p^*_2(x) = (x-1)p^*_1(x) = -0.3002439\,x^2 + 1.578949\,x - 1.278706$$

è il polinomio cercato e risulta

$$r^* \approx 0.380\ 10^{-1}.$$ ∎

9. Approssimazione minimax con vincoli

In alcuni casi le approssimazioni cercate devono soddisfare alcuni vincoli. I casi più frequenti sono:

a) la funzione approssimante deve assumere in uno o più punti gli stessi valori della funzione data $f(x)$, oppure avere in uno o più punti gli stessi valori e le stesse derivate della funzione $f(x)$ fino ad un ordine prefissato;

b) la funzione approssimante appartiene ad una sottoclasse della classe dei polinomi, come ad esempio nel caso in cui alcuni coefficienti debbano assumere valori prefissati.

Approssimazioni di questo genere possono essere utilissime, quando si richiede ad esempio che la funzione approssimante abbia lo stesso andamento della $f(x)$, oppure conservi le stesse simmetrie.

In generale, ben poco si può dire, dal punto di vista teorico, per quanto riguarda l'esistenza e l'unicità delle soluzioni di tali problemi. Per alcuni casi particolari, come ad esempio per il caso in cui la funzione approssimante è un polinomio con alcuni coefficienti prefissati, esiste una generalizzazione del teorema 6.30 di Chebyshev, in cui il numero dei punti di equioscillazione del resto risulta diminuito, per tener conto del minor numero di gradi di libertà.

6.48 Esempio. L'approssimazione minimax di $f(x) = \sqrt{x}$ rispetto all'errore relativo con un polinomio della forma

$$p_1^*(x) = x + a_0$$

nell'intervallo $[1/16, 1]$, è tale che il resto relativo

$$r^*(x) = 1 - \frac{x + a_0}{\sqrt{x}}$$

ha due punti di equioscillazione. Poiché $(r^*)'(x)$ può annullarsi in un punto solo, non è possibile che i due punti di equioscillazione siano entrambi interni. Supponendo che essi siano gli estremi dell'intervallo, si impongono le condizioni $r^*(1/16) = d$, $r^*(1) = -d$ e si risolve il sistema lineare di due equazioni nelle incognite a_0 e d che così si ottiene. La soluzione è $a_0 = d = 3/20$. Però il polinomio $p(x) = x + 3/20$ non è l'approssimazione minimax cercata, perché il resto $r^*(x)$ risulta avere un punto di massimo interno all'intervallo $[1/16, 1]$, esattamente il punto $x = a_0 = 3/20$, in cui ha il valore $r^*(3/20) = 0.2254033$, superiore a d. Di conseguenza uno dei due punti di equioscillazione deve essere interno all'intervallo e l'altro deve coincidere con uno dei due estremi.

Ponendo

$$a < x_0^* < x_1^* = b,$$

si ricava

$$x_0^* = a_0 = d = 3 - 2\sqrt{2} = 0.1715729 = r^*;$$

ponendo invece

$$a = x_0^* < x_1^* < b,$$

si ricava

$$x_1^* = a_0 = \frac{9 - 4\sqrt{2}}{16} = 0.2089466 \quad \text{e} \quad d = -\frac{3 - 2\sqrt{2}}{2} = -0.08578644,$$

ma $|r^*(1)| = 0.2089466$.

Confrontando i valori ottenuti risulta che il polinomio che fornisce l'approssimazione minimax cercata è

$$p_1^*(x) = x + 3 - 2\sqrt{2} = x + 0.1715729. \qquad \blacksquare$$

Nel caso in cui il polinomio cercato debba avere alcuni coefficienti prefissati, il metodo di Remez può essere applicato senza grosse modifiche, sostituendo nella (48) i valori prefissati e diminuendo di conseguenza il numero delle equazioni. Ad esempio, se il polinomio $p_n^*(x)$ deve essere pari, i coefficienti a_j per j dispari devono essere nulli e la (48) assume la forma

$$\sum_{\substack{j=0 \\ j \text{ pari}}}^{n} a_j^{(k)} \big(x_i^{(k)}\big)^j + (-1)^i d^{(k)} = f(x_i^{(k)}), \quad i = 0, \dots, \frac{n}{2} + 1.$$

6.49 Esempio. Per determinare il polinomio di approssimazione minimax di $\cos x$ della forma

$$p_2^*(x) = 1 + a_0 x^2 + a_1 x^4$$

nell'intervallo $\big[0, \dfrac{\pi}{2}\big]$, con il metodo di Remez, al posto del sistema (48) si risolve il sistema

$$1 + a_0^{(k)} \big(x_i^{(k)}\big)^2 + a_1^{(k)} \big(x_i^{(k)}\big)^4 + (-1)^i d^{(k)} = f(x_i^{(k)}), \quad i = 0, 1, 2.$$

Non si può scegliere $x_0^{(0)} = 0$ perché risulterebbe $d^{(0)} = 0$. Scegliendo invece ad esempio

$$x_0^{(0)} = \frac{\pi}{6}, \quad x_1^{(0)} = \frac{\pi}{3}, \quad x_2^{(0)} = \frac{\pi}{2},$$

dopo 3 iterazioni si ottiene il polinomio

$$p_2^*(x) = 1 - 0.4966048\, x^2 + 0.03713169\, x^4,$$

per il quale risulta $r^* \approx 0.737\ 10^{-3}$. $\blacksquare$

10. Approssimazione minimax razionale

Quando l'approssimazione minimax polinomiale non dà buoni risultati perché la successione r_n^* converge lentamente, si può prendere in considerazione un'approssimazione con funzioni razionali. Il problema che si presenta non può essere espresso nei termini del problema 6.3 in quanto non è

lineare; quindi l'esistenza di una soluzione non segue dal teorema 6.4, ma richiede un'ulteriore indagine.

Siano $p(x)$ e $q(x)$ due polinomi appartenenti rispettivamente a $\mathcal{P}_m$ e $\mathcal{P}_n$, con $q(x) \neq 0$ in $[a, b]$. Si considera la classe $\mathcal{R}_{m,n}$ delle funzioni razionali

$$w(x) = \frac{p(x)}{q(x)} = \frac{a_m x^m + a_{m-1} x^{m-1} + \ldots + a_0}{b_n x^n + b_{n-1} x^{n-1} + \ldots + b_0}\,, \tag{62}$$

in cui $p(x)$ e $q(x)$ sono primi fra loro, cioè $w(x)$ è irriducibile. Poiché $w(x)$ non cambia se si dividono numeratore e denominatore per una stessa costante non nulla, si fissa di solito $b_0 = 1$ se $b_0 \neq 0$, altrimenti $a_0 = 1$. Teoremi analoghi a quelli del caso polinomiale valgono anche in questo caso.

6.50 Teorema. *Sia $f(x) \in C[a, b]$. Il problema dell'approssimazione razionale in norma ∞, cioè di determinare una funzione $w^*(x) \in \mathcal{R}_{m,n}$ tale che*

$$\|f(x) - w^*(x)\|_\infty = \min_{w(x) \in \mathcal{R}_{m,n}} \|f(x) - w(x)\|_\infty,$$

ha soluzione. ∎

Anche il teorema di equioscillazione di Chebyshev, su cui si basa tutta la teoria del minimax, ha un corrispondente per l'approssimazione razionale.

6.51 Teorema. *Una funzione razionale $w^*(x) \in \mathcal{R}_{m,n}$ è di approssimazione minimax se e solo se, detti m' e n' i gradi effettivi di $p(x)$ e $q(x)$ e*

$$N = \max\{m + n', n + m'\},$$

esistono almeno $N + 2$ punti

$$a \leq x_0^* < x_1^* < \ldots < x_{N+1}^* \leq b,$$

per cui è

$$f(x_i^*) - w^*(x_i^*) = (-1)^i d, \quad i = 0, \ldots, N + 1,$$

dove

$$|d| = \|f - w^*\|_\infty.$$

Inoltre la funzione razionale di approssimazione minimax è unica. ∎

La dimostrazione dei teoremi 6.50 e 6.51 è alquanto più complicata di quella del caso polinomiale: si veda [2], [30], [37] e [39].

La presenza nella definizione di N dei gradi effettivi m' e n' dei polinomi $p(x)$ e $q(x)$ rende più complicata che nel caso polinomiale la ricerca della funzione razionale di migliore approssimazione minimax.

6.52 Esempio. Nell'esempio 6.33 si è visto che il polinomio $p_1^*(x) = \frac{3}{4}\,x$ è l'approssimazione minimax della funzione $f(x) = x^3$ su $[-1, 1]$ e che il resto $f(x) - p_1^*(x)$ ha 4 punti di equioscillazione. Poiché $\mathcal{R}_{1,0} = \mathcal{P}_1$, si può anche dire che $w^*(x) = \frac{3}{4}\,x$ è l'approssimazione minimax di $f(x)$ in $\mathcal{R}_{1,0}$. Ma $w^*(x) = \frac{3}{4}\,x$ è l'approssimazione minimax di $f(x)$ anche in $\mathcal{R}_{1,1}$, in $\mathcal{R}_{2,0}$, in $\mathcal{R}_{2,1}$. Infatti risulta $m' = 1$ e $n' = 0$ e

$$\begin{aligned}
&\text{in } \mathcal{R}_{1,1} \qquad m = 1,\ n = 1,\ N = \max\{1, 2\},\\
&\text{in } \mathcal{R}_{2,0} \qquad m = 2,\ n = 0,\ N = \max\{2, 1\},\\
&\text{in } \mathcal{R}_{2,1} \qquad m = 2,\ n = 1,\ N = \max\{2, 2\}.
\end{aligned}$$

Quindi in ognuno dei tre casi risulta $N = 2$, e secondo il teorema 6.51 la funzione razionale di approssimazione minimax deve avere 4 punti di equioscillazione, come in effetti ha $w^*(x) = \frac{3}{4}\,x$.

Questa funzione non è però l'approssimazione di $f(x) = x^3$ in $\mathcal{R}_{1,2}$, perché in questo caso risulterebbe $N = 3$ e vi dovrebbero essere almeno 5 punti di equioscillazione. Per la funzione razionale di approssimazione in $\mathcal{R}_{1,2}$ non è possibile che $m' = 1$ e che $n' < 2$, perché tale funzione dovrebbe stare anche in $\mathcal{R}_{m,n'}$ e quindi dovrebbe coincidere con $\frac{3}{4}\,x$. In $\mathcal{R}_{1,2}$ esiste perciò una funzione razionale di approssimazione minimax della forma (62) con $q(x)$ di grado 2. D'altra parte, tenendo conto che la funzione $f(x)$ è antisimmetrica su $[-1, 1]$, anche $w(x)$ deve essere antisimmetrica e quindi deve essere della forma

$$w(x) = \frac{a_1 x}{b_2 x^2 + 1}\,.$$

Inoltre, poiché $f(0) - w(0) = 0$, il punto 0 non può essere di equioscillazione, e quindi per l'antisimmetria devono esistere 6 punti di equioscillazione in $[-1, 1]$. Si supponga ad esempio che tali punti siano

$$-1 = x_0^* < x_1^* < x_2^* < 0 < x_3^* < x_4^* < x_5^* = 1, \quad \text{con} \quad x_3^* = -x_2^*,\ x_4^* = -x_1^*.$$

Posto $r(x) = x^3 - w(x)$, si risolve il sistema

$$r(x_i) = (-1)^i d \quad \text{per } i = 3, 4, 5 \quad \text{e} \quad r'(x_i) = 0 \quad \text{per } i = 3, 4,$$

e si ottiene

$$\begin{aligned}
&x_3^* = 0.3730518, \quad x_4^* = 0.8730500, \quad a_1^* = 0.3081020, \quad b_1^* = -0.7135783,\\
&|d^*| = 0.07569391.
\end{aligned}$$

L'approssimazione razionale così ottenuta è anche l'approssimazione minimax in $\mathcal{R}_{2,2}$, in $\mathcal{R}_{1,3}$, in $\mathcal{R}_{2,3}$. Infatti in questi tre casi si ha $N = 4$. ∎

Sono quindi i casi di degenerazione, cioè quelli per cui $m' < m$ e $n' < n$, che rendono più complicata l'approssimazione razionale. Esclusi questi casi (si veda l'esercizio 6.59), per determinare la funzione razionale di approssimazione minimax si può usare ancora il metodo di Remez, con una variante.

Scelto un vettore iniziale $\mathbf{x}^{(0)}$ di $N+2$ componenti $x_i^{(0)}$, tali che

$$a \leq x_0^{(0)} < x_1^{(0)} < \ldots < x_{N+1}^{(0)} \leq b,$$

si risolve al k-esimo passo il sistema non lineare nelle incognite $a_0^{(k)}, a_1^{(k)}, \ldots, a_m^{(k)}, b_1^{(k)}, \ldots, b_n^{(k)}$ e $d^{(k)}$

$$\begin{aligned} \sum_{j=0}^{m} a_j^{(k)} \big(x_i^{(k)}\big)^j + \big[(-1)^i d^{(k)} - f(x_i^{(k)})\big] \sum_{j=0}^{n} b_j^{(k)} \big(x_i^{(k)}\big)^j = 0, \\ b_0^{(k)} = 1, \qquad i = 0, \ldots, N+1, \end{aligned} \tag{63}$$

ottenendo così i coefficienti della funzione razionale

$$w^{(k)}(x) = \frac{\displaystyle\sum_{j=0}^{m} a_j^{(k)} x^j}{\displaystyle\sum_{j=0}^{n} b_j^{(k)} x^j},$$

tale che i punti $x_i^{(k)}$, $i = 0, \ldots, N+1$, sono di equioscillazione per il resto $r^{(k)}(x) = f(x) - w^{(k)}(x)$.

L'algoritmo procede poi come nel caso polinomiale, determinando i punti $x_i^{(k+1)}$, $i = 0, \ldots, N+1$, di massimo o minimo per il resto $r^{(k)}(x)$. Le difficoltà, teoriche e pratiche, che si possono presentare nel caso razionale sono molte: la scelta del vettore iniziale $\mathbf{x}^{(0)}$ è critica, perché in questo caso non esiste un teorema che garantisca la convergenza del metodo qualunque sia la scelta dei punti iniziali. Esiste comunque un teorema analogo al 6.38, che assicura che sotto certe ipotesi il metodo è localmente convergente, richiedendo quindi che i punti $x_i^{(0)}$, $i = 0, \ldots, N+1$, iniziali siano abbastanza vicini agli x_i^*, $i = 0, \ldots, N+1$ [35]. In analogia con il caso polinomiale si possono scegliere come punti iniziali, oltre ad a e b se il resto è standard, i punti di Chebyshev. Poiché però in questo caso non vi è alcun teorema analogo al 6.19, che dia una giustificazione teorica a questa scelta, qualunque altra scelta, come ad esempio quella dei punti equidistanti, può andare ugualmente bene.

La seconda difficoltà è rappresentata dal fatto che il sistema (63) non è lineare. Quindi per la risoluzione si deve fare ricorso a metodi iterativi, per i quali si può assicurare la convergenza solo se i valori iniziali sono scelti

sufficientemente vicini alla soluzione. Dopo la prima iterazione del metodo di Remez, dei buoni valori iniziali per risolvere il sistema (63) sono quelli ottenuti all'iterazione precedente.

Infine resta la difficoltà di determinare i punti di massimo e minimo della funzione $r^{(k)}(x) = f(x) - w^{(k)}(x)$.

6.53 Esempio. Con il termine di *approssimazione iperbolica minimax* di una funzione $f(x)$ si intende l'approssimazione razionale minimax della $f(x)$ della forma

$$w(x) = \frac{a_1 x + a_0}{b_1 x + 1}\,. \tag{64}$$

Sia ad esempio $f(x) = \sqrt{x}$, $1/16 \le x \le 1$. Si considerano inizialmente i punti

$$x_0^{(0)} = a = 0.0625, \quad x_1^{(0)} = \frac{b-a}{2}\,\cos\frac{2\pi}{3} + \frac{a+b}{2} = 0.2968750,$$

$$x_2^{(0)} = \frac{b-a}{2}\,\cos\frac{\pi}{3} + \frac{a+b}{2} = 0.7656250, \quad x_3^{(0)} = b = 1.$$

Risolvendo il sistema (63) con il metodo iterativo di Newton-Raphson, si ottengono i valori

$$a_1^{(0)} = 1.799039,\ a_0^{(0)} = 0.1595804,\ b_1^{(0)} = 0.9713378,\ d^{(0)} = 0.0064515,$$

da cui

$$r^{(0)}(x) = \sqrt{x} - \frac{1.799039x + 0.1595804}{0.9713378x + 1}\,.$$

I punti di massimo o minimo di $r^{(0)}(x)$ interni all'intervallo $[1/16, 1]$ sono

$$y_1 = 0.1710588, \qquad y_2 = 0.6406236,$$

e si ha

$$\left|\frac{M^{(0)}}{m^{(0)}} - 1\right| \approx 0.993.$$

Assumendo $x_1^{(1)} = y_1$, $x_2^{(1)} = y_2$, e ripetendo il calcolo si ottiene

$$a_1^{(1)} = 1.828341,\ a_0^{(1)} = 0.1613202,\ b_1^{(1)} = 1.008246,\ d^{(1)} = 0.0092545.$$

Alla terza iterazione infine si ottiene

$$a_1^{(3)} = 1.826492,\ a_0^{(3)} = 0.1614652,\ b_1^{(3)} = 1.006633,\ d^{(3)} = 0.0093067,$$

e poiché

$$\left|\frac{M^{(3)}}{m^{(3)}} - 1\right| \approx 0.130\ 10^{-10},$$

si assume come approssimazione iperbolica di $\sqrt{x}$ la funzione razionale

$$w^*(x) = \frac{1.826492\,x + 0.1614652}{1.006633\,x + 1}\,,$$

che viene anche scritta nella forma

$$w^*(x) = 1.814457 - \frac{1.651079}{x + 0.9934107}\,.$$

Risulta $r^* \approx 0.931\ 10^{-2}$. ∎

Anche per quanto riguarda l'approssimazione razionale non vi sono difficoltà teoriche nel considerare il resto relativo anziché quello assoluto. Il teorema di equioscillazione e l'algoritmo di Remez ne risultano modificati di conseguenza.

6.54 Esempio. Per determinare l'approssimazione minimax iperbolica rispetto all'errore relativo per la funzione $f(x) = \sqrt{x}$ nell'intervallo $[1/16, 1]$, si procede in modo analogo a quanto fatto nell'esempio 6.53. Assumendo come punti iniziali gli stessi $x_i^{(0)}$, $i = 0, \ldots, 3$, lì considerati, si determinano i coefficienti della funzione razionale di equioscillazione per il resto

$$r^{(0)}(x) = \frac{f(x) - w^{(0)}(x)}{f(x)}\,,$$

dove $w(x)$ è della forma (64). Si ottiene

$$r^{(0)}(x) = 1 - \frac{1.872796\,x + 0.1514012}{(1.040949\,x + 1)\sqrt{x}}\,,$$

i cui punti di massimo o minimo interni all'intervallo $[1/16, 1]$ sono

$$y_1 = 0.1326446. \quad y_2 = 0.5854902.$$

Sono sufficienti quattro iterazioni per ottenere

$$\left|\frac{M^{(4)}}{m^{(4)}} - 1\right| \approx 0.237\ 10^{-7},$$

per cui si assume come approssimazione iperbolica dell'errore relativo di $\sqrt{x}$ la funzione razionale

$$w^*(x) = \frac{1.999654\,x + 0.1485957}{1.188972\,x + 1}\,,$$

che viene anche scritta nella forma

$$w^*(x) = 1.681835 - \frac{1.289551}{x + 0.8410629}\,.$$

Risulta

$$r^* \approx 0.186\ 10^{-1}.$$ ∎

6.55 Esempio. L'approssimazione minimax iperbolica $w(x)$ di $f(x) = \sqrt{x}$ rispetto all'errore relativo nell'intervallo $[1/16, 1]$, con il vincolo che $w(1) = f(1)$, è data da

$$w(x) = \frac{a_1 x + a_0}{b_1 x + 1}\,,$$

in cui i coefficienti devono soddisfare la relazione

$$a_1 + a_0 = b_1 + 1. \tag{65}$$

Poiché in b il resto è zero, si scelgono i nodi

$$a = x_0^* < x_1^* < x_2^* < b.$$

Posto

$$r(x) = \frac{f(x) - w(x)}{f(x)}\,,$$

deve essere

$$\begin{cases} r(x_i^*) = (-1)^i d, \quad i = 0, 1, 2, \\ r'(x_i^*) = 0, \quad i = 1, 2. \end{cases} \tag{66}$$

Il sistema formato dalla (65) e dalle (66) ha come soluzione

$$a_1 = 1.751081, \qquad a_0 = 0.1599500, \quad b_1 = 0.9110308,$$

$$x_1^* = 0.1485144, \quad x_2^* = 0.6751124, \quad d = -0.1951885\ 10^{-1}.$$

L'approssimazione iperbolica cercata è quindi

$$w^*(x) = \frac{1.751081\,x + 0.1599500}{0.9110308\,x + 1}\,,$$

che viene anche scritta nella forma

$$w^*(x) = 1.922088 - \frac{1.762137}{0.9110308\,x + 1}\,.$$ ∎

Lo studio della velocità di convergenza dell'approssimazione razionale minimax al crescere del grado $m+n$ è molto più complesso che nel caso polinomiale, e ciò è dovuto anche alla possibile esistenza di punti di singolarità di $f(x)$ e di $w^*_{m,n}(x)$ al di fuori dell'intervallo $[a,b]$. Stime asintotiche dell'errore del minimax razionale di funzioni analitiche e con un numero finito di poli sono riportate in [24]. Sono state individuate classi di funzioni per le quali l'approssimazione razionale converge più rapidamente, al crescere del grado, di quella polinomiale [32]. Ad esempio, se $f(x) \in C^k[a,b]$ e $f^{(k)}(x)$ è assolutamente continua, esiste una costante M per cui

$$\|f - w^*_{n,n}\|_\infty \leq \frac{M}{n^{k+1}}$$

(si confronti con la maggiorazione fornita dal teorema di Jackson 6.35).

Nel caso particolare della funzione $f(x) = \mathrm{e}^x$ sull'intervallo $[-1,1]$ è asintoticamente per $m,\, n \to \infty$

$$\|f - p^*_{m+n}\|_\infty \sim \frac{1}{2^{m+n}(m+n+1)!}\,,$$

$$\|f - w^*_{m,n}\|_\infty \sim \frac{m!\,n!}{2^{m+n}(m+n)!\,(m+n+1)!}\,,$$

dove $p^*_{m+n} \in \mathcal{P}_{m+n}$ e $w^*_{m,n} \in \mathcal{R}_{m,n}$ [32]. In questo caso quindi, a parità di grado, l'approssimazione razionale è migliore di quella polinomiale e i valori minimi dell'errore vengono ottenuti quando i gradi del numeratore e del denominatore coincidono o differiscono di 1.

6.56 Esempio. Nella tabella sono riportate, al variare dei gradi del numeratore e del denominatore, le norme ∞ dei resti delle approssimazioni minimax razionali di grado 6 della funzione $f(x) = \mathrm{e}^x$ per $x \in [-1,1]$. Posto

$$r^*_{m,n} = \|f - w^*_{m,n}\|_\infty, \quad m+n=6,$$

è

m	n	$r^*_{m,n}$
6	0	$0.321\ 10^{-5}$
5	1	$0.532\ 10^{-6}$
4	2	$0.210\ 10^{-6}$
3	3	$0.155\ 10^{-6}$
2	4	$0.219\ 10^{-6}$
1	5	$0.490\ 10^{-6}$
0	6	$0.283\ 10^{-5}$

■

6.57 Esempio. La velocità di convergenza della successione delle approssimazioni minimax polinomiali alla funzione $f(x) = \arcsin x$, $x \in [-1,1]$,

è molto bassa. Infatti (si veda l'esercizio 6.52) asintoticamente, per $n \to \infty$, risulta

$$r_n^* = \|f - p_n^*\|_\infty \geq \gamma_n, \quad \text{dove} \quad \gamma_n \sim \frac{\pi}{2} \, \frac{1}{n \log n} \, .$$

Ciò è dovuto principalmente al fatto che la funzione non è derivabile per $x = \pm 1$.

La successione delle approssimazioni minimax razionali però converge più rapidamente, come risulta dal grafico della figura 6.21 in cui con i pallini sono indicati i resti polinomiali r_n^* per $n = 1, \ldots, 17$, n dispari, e con i quadratini neri i resti razionali

$$r_{m,m+1}^* = \|f - w_{m,m+1}^*\|_\infty, \quad m = \frac{n-1}{2} \, .$$

La funzione $f(x)$ è antisimmetrica in $[-1, 1]$, quindi l'approssimazione $w_{m,m+1}^*(x)$ in $\mathcal{R}_{m,m+1}$ per m dispari è anche l'approssimazione in $\mathcal{R}_{m+1,m+2}$ (si veda l'esercizio 6.60). Nella figura risultano infatti uguali i resti $r_{m,m+1}^*$ e $r_{m+1,m+2}^*$ per m dispari. ∎

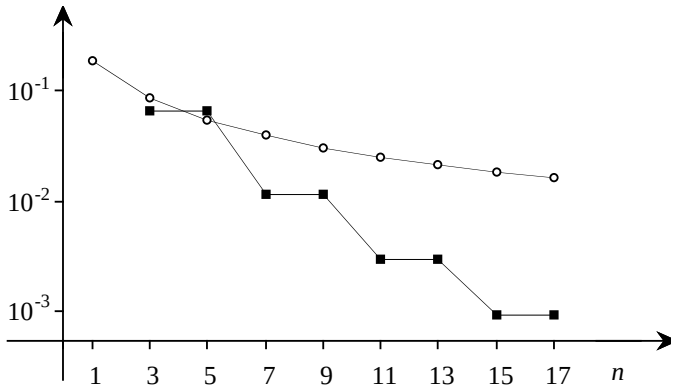


Fig. 6.21 - Resti delle approssimazioni polinomiali e razionali della funzione $f(x) = \arcsin x$, $x \in [-1, 1]$.

11. Approssimazione razionale con frazioni continue infinite

Le frazioni continue finite considerate per l'interpolazione razionale possono essere ottenute come frazioni parziali di frazioni continue con infiniti termini, dette *frazioni continue infinite*, che sono della forma

$$w = d_0 + \frac{c_1}{d_1 \; + \; \dfrac{c_2}{d_2 \; + \; \ldots}} \tag{67}$$

Se $c_i = 0$ per un indice i, la frazione continua è in realtà finita. Indicata con

$$w_k = d_0 + \frac{c_1}{d_1} \; + \; \frac{c_2}{d_2} \; + \; \ldots \; + \; \frac{c_k}{d_k}\,,$$

la k-esima frazione parziale, se esiste finito il

$$\lim_{k\to\infty} w_k, \tag{68}$$

si dice che la frazione continua infinita (67) *converge* e w è il valore del limite (68). Per la convergenza di una frazione continua infinita valgono i seguenti teoremi.

6.58 Teorema (di Seidel). *Se $d_i \geq 0$ per $i \geq 1$, allora la frazione continua*

$$w = d_0 + \frac{1}{d_1} \; + \; \frac{1}{d_2} \; + \; \ldots$$

converge se e solo se la serie $\sum_{i=1}^{\infty} d_i$ diverge. ∎

6.59 Teorema. *Se $c_i \neq 0$ per $i \geq 1$, allora la frazione continua*

$$w = d_0 + \frac{c_1}{d_1} \; + \; \frac{c_2}{d_2} \; + \; \ldots$$

converge se $|d_i| \geq |c_i| + 1$ per $i \geq 1$. ∎

Per la dimostrazione di 6.58 e 6.59 si vedano gli esercizi 6.61 e 6.62.

L'approssimazione razionale corrispondente alla formula di Taylor di una funzione $f(x)$ troncata al k-esimo termine è, come si è visto nel capitolo 5, la frazione continua di Thiele, che può essere vista come la k-esima frazione parziale dell'espansione di $f(x)$ in frazione continua.

L'insieme di convergenza della frazione continua di una funzione $f(x)$ può essere molto più grande dell'insieme di convergenza della corrispondente formula di Taylor, che deve comunque essere un cerchio complesso non contenente poli di $f(x)$. Ad esempio, nel caso in cui sia

$$w(x) = d_0 + \frac{c_1}{1} \; + \; \frac{c_2 x}{1} \; + \; \frac{c_3 x}{1} \; + \; \ldots\,, \tag{69}$$

dove i coefficienti c_i sono reali e diversi da zero, se $\lim_{i\to\infty} c_i = 0$, allora la frazione continua (69) converge uniformemente a $f(x)$ in ogni sottoinsieme compatto del piano complesso che non contenga poli di $f(x)$; se $\lim_{i\to\infty} c_i = c \neq 0$, allora la frazione continua (69) converge uniformemente

a $f(x)$ in ogni sottoinsieme compatto del piano complesso che non contenga poli di $f(x)$ ed inoltre non contenga punti della semiretta $x = -\dfrac{t}{4c}\,,\ t \geq 1$ [26].

6.60 Esempio. Dall'esempio 5.54 si ricavano l'espansione in frazione continua della funzione e^x

$$\mathrm{e}^x = 1 + \frac{x}{1} \underset{-}{} \frac{x}{2} \underset{+}{} \frac{x}{3} \underset{-}{} \frac{x}{2} \underset{+}{} \frac{x}{5} \underset{-}{} \frac{x}{2} \underset{+}{} \frac{x}{7} \underset{-}{} \frac{x}{2} \underset{+}{} \ldots \tag{70}$$

e l'espansione in frazione continua della funzione $\log(1+x)$

$$\log(1+x) = \frac{x}{1} \underset{+}{} \frac{x}{2/1} \underset{+}{} \frac{x}{3} \underset{+}{} \frac{x}{2/2} \underset{+}{} \frac{x}{5} \underset{+}{} \frac{x}{2/3} \underset{+}{} \frac{x}{7} \underset{+}{} \ldots \tag{71}$$

La (70), tenendo conto che $\mathrm{e}^x = 1/\mathrm{e}^{-x}$, può essere scritta anche nella forma equivalente

$$\mathrm{e}^x = \frac{1}{1} \underset{+}{} \frac{-x}{1} \underset{+}{} \frac{x/2}{1} \underset{+}{} \frac{-x/6}{1} \underset{+}{} \frac{x/6}{1} \underset{+}{} \frac{-x/10}{1} \underset{+}{} \frac{x/10}{1} \underset{+}{} \ldots$$

e, poiché $\lim\limits_{i\to\infty} c_i = 0$ e la funzione e^x non ha poli, la (70) converge per ogni x complesso (e quindi anche reale).

La (71), per $x \neq 0$, può essere scritta anche nella forma equivalente

$$\frac{\log(1+x)}{x} = \frac{1}{1} \underset{+}{} \frac{x/2}{1} \underset{+}{} \frac{x/6}{1} \underset{+}{} \frac{x/3}{1} \underset{+}{} \frac{x/5}{1} \underset{+}{} \frac{3x/10}{1} \underset{+}{} \frac{3x/14}{1} \underset{+}{} \ldots$$

e, poiché $\lim\limits_{i\to\infty} c_i = \dfrac{1}{4}$ e la funzione $\log x$ non ha poli, la (71) converge per ogni x complesso, esclusi i punti della semiretta reale $x \leq -1$. Si noti come in questo caso risulti ampliato l'insieme di convergenza rispetto a quello della serie di Taylor di $\log(1+x)$, che è un cerchio di centro 1 e raggio 1. Ponendo nella (71) $x = 1$ si ottiene l'espansione in frazione continua di $\log 2$

$$\log 2 = \frac{1}{1} \underset{+}{} \frac{1}{2/1} \underset{+}{} \frac{1}{3} \underset{+}{} \frac{1}{2/2} \underset{+}{} \frac{1}{5} \underset{+}{} \frac{1}{2/3} \underset{+}{} \frac{1}{7} \underset{+}{} \ldots$$

Le successive frazioni parziali sono

$$w_1 = 1, \quad w_2 = \frac{2}{3} = 0.6666667, \quad w_3 = \frac{7}{10} = 0.7,$$
$$w_4 = \frac{9}{13} = 0.6923077, \quad w_5 = \frac{52}{75} = 0.6933333, \quad w_6 = \frac{131}{189} = 0.6931217,$$
$$w_7 = \frac{1073}{1548} = 0.6931525, \quad w_8 = \frac{445}{642} = 0.6931464.$$

w_8 approssima $\log 2$ con un errore di circa $0.763\ 10^{-6}$. L'approssimazione di $\log 2$ con la serie di Taylor e errore dello stesso ordine di grandezza avrebbe invece richiesto un numero molto più elevato di termini (si veda l'esempio 4.1). ∎

Non sempre è possibile dare delle espressioni semplici per i coefficienti degli sviluppi in frazione continua delle funzioni non razionali più comuni, come ad esempio le funzioni $\sin x$ e $\cos x$. In alcuni casi si può applicare il metodo di Viskovatov, descritto nel paragrafo 10 del capitolo 5.

6.61 Esempio. La funzione $f(x) = \tan x$ è dispari. Per ottenere il suo sviluppo in frazione continua infinita, non è conveniente partire dalla formula di Maclaurin che ha dei coefficienti espressi in termini dei numeri di Bernoulli, ma conviene applicare il metodo di Viskovatov agli sviluppi delle due funzioni $\cos x$ e $\sin x$, nel modo seguente

$$\tan x = \frac{\sin x}{\cos x} = \frac{x \displaystyle\sum_{i=0}^{\infty} (-1)^i \, \frac{x^{2i}}{(2i)!(2i+1)}}{\displaystyle\sum_{i=0}^{\infty} (-1)^i \, \frac{x^{2i}}{(2i)!}} .$$

Le funzioni

$$t_j(x) = \sum_{i=0}^{\infty} (-1)^i \, \beta_{j,i} \, \frac{x^{2i}}{(2i)!}, \quad \text{dove } \beta_{j,i} = \begin{cases} 1 & \text{per } j = 1, \\ \displaystyle\prod_{r=1}^{j-1} \frac{1}{2(i+r)-1}, & \text{per } j \geq 2, \end{cases}$$

soddisfano la relazione

$$x^2 t_{j+2}(x) = (2j-1) t_{j+1}(x) - t_j(x), \quad j = 1, 2, \ldots \tag{72}$$

Infatti

$$\begin{aligned} &(2j-1) t_{j+1}(x) - t_j(x) = \\ &\quad = \sum_{i=0}^{\infty} (-1)^i \left[(2j-1)\beta_{j+1,i} - \beta_{j,i} \right] \frac{x^{2i}}{(2i)!} \\ &\quad = \sum_{i=1}^{\infty} \left[(-1)^i \, \frac{x^{2i}}{(2i)!} \, (-2i) \prod_{r=1}^{j} \frac{1}{2(i+r)-1} \right] \\ &\quad = \sum_{p=0}^{\infty} \left[(-1)^p \, \frac{x^{2p+2}}{(2p+2)!} \, (2p+2) \prod_{r=1}^{j} \frac{1}{2(p+1+r)-1} \right] = x^2 t_{j+2}(x). \end{aligned}$$

Dalla (72) si ottiene allora

$$\frac{t_j(x)}{t_{j+1}(x)} = (2j-1) - \frac{x^2 t_{j+2}(x)}{t_{j+1}(x)} = (2j-1) - \cfrac{x^2}{\cfrac{t_{j+1}(x)}{t_{j+2}(x)}}, \quad j = 1, 2, \ldots,$$

e quindi

$$\tan x = \frac{x t_2(x)}{t_1(x)} = \cfrac{x}{\cfrac{t_1(x)}{t_2(x)}} = \cfrac{x}{1} - \cfrac{x^2}{\cfrac{t_2(x)}{t_3(x)}} = \cfrac{x}{1} - \cfrac{x^2}{3} - \cfrac{x^2}{\cfrac{t_3(x)}{t_4(x)}}$$
$$= \ldots = \cfrac{x}{1} - \cfrac{x^2}{3} - \cfrac{x^2}{5} - \cfrac{x^2}{7} - \ldots, \quad \text{per} \quad x \neq \frac{\pi}{2} + k\pi, \quad k \text{ intero.}$$

In modo analogo si procede per determinare lo sviluppo in frazione continua per la funzione $f(x) = \tanh x$. Tenendo conto del fatto che gli sviluppi di Maclaurin delle funzioni $\sinh x$ e $\cosh x$ sono, a parte il segno, gli stessi di $\sin x$ e $\cos x$, si ottiene

$$\tanh x = \cfrac{x}{1} + \cfrac{x^2}{3} + \cfrac{x^2}{5} + \cfrac{x^2}{7} + \ldots, \quad \text{per} \quad x \neq \boldsymbol{i}\Big(\frac{\pi}{2} + k\pi\Big), \quad k \text{ intero.} \ \blacksquare$$

12. Approssimazione di Padé

Un'approssimazione razionale di una funzione $f(x)$ sviluppabile in serie di Maclaurin può essere ottenuta anche con il metodo di *Padé*. La funzione razionale

$$R_{m,n}(x) = \frac{p_m(x)}{q_n(x)},$$

dove $p_m(x)$ e $q_n(x)$ sono due polinomi di grado rispettivamente minore o uguale a m e a n e $q_n(0) \neq 0$, è detta *approssimante di Padé* di ordine $m+n$ della funzione $f(x)$ se

$$R_{m,n}^{(k)}(0) = f^{(k)}(0), \quad k = 0, \ldots, m+n, \tag{73}$$

cioè se gli sviluppi di Maclaurin di $f(x)$ e di $R_{m,n}(x)$ coincidono fino al termine $(m+n+1)$-esimo.

Nel caso particolare $n = 0$ si ha

$$R_{m,0}(x) = p_m(x) = \sum_{i=0}^{m} \alpha_i x^i,$$

cioè $R_{m,0}(x)$ è il polinomio ottenuto troncando all'$(m+1)$-esimo termine la serie di Maclaurin di $f(x)$. Se $n \geq 1$, posto

$$r(x) = f(x) - \frac{p_m(x)}{q_n(x)}$$

e

$$s(x) = r(x)q_n(x) = f(x)q_n(x) - p_m(x),$$

poiché

$$s^{(k)}(0) = 0, \quad \text{per} \quad k = 0, \ldots, m+n,$$

la (73) è equivalente alla condizione che i primi $m+n+1$ termini della serie di Maclaurin di $s(x)$ siano nulli. Indicati con $a_0, \ldots, a_m$ gli $m+1$ coefficienti di $p_m(x)$ e con $b_0, \ldots, b_n$ gli $n+1$ coefficienti di $q_n(x)$ e con

$$f(x) = \sum_{i=0}^{\infty} \alpha_i x^i$$

lo sviluppo di Maclaurin della $f(x)$, è

$$s(x) = \Big(\sum_{i=0}^{\infty} \alpha_i x^i \Big) \, (b_n x^n + b_{n-1} x^{n-1} + \ldots + b_0) - (a_m x^m + a_{m-1} x^{m-1} + \ldots + a_0).$$

I primi $m+n+1$ coefficienti di $s(x)$ sono

$$s_k = \begin{cases} \displaystyle\sum_{j=0}^{\min(k,n)} \alpha_{k-j} b_j - a_k, & \text{per } k = 0, \ldots, m, \\ \displaystyle\sum_{j=0}^{\min(k,n)} \alpha_{k-j} b_j, & \text{per } k = m+1, \ldots, m+n. \end{cases}$$

Imponendo che tali coefficienti siano nulli si ottengono le relazioni

$$a_k = \sum_{j=0}^{\min(k,n)} \alpha_{k-j} b_j, \quad k = 0, \ldots, m, \tag{74}$$

$$\sum_{j=0}^{\min(k,n)} \alpha_{k-j} b_j = 0, \quad k = m+1, \ldots, m+n. \tag{75}$$

I coefficienti b_j, $j = 0, \ldots, n$, si ricavano dal sistema lineare (75), omogeneo di n equazioni. Posto $b_0 = 1$, se il sistema è risolubile, per $n \geq 1$ gli altri coefficienti b_j si ricavano risolvendo il sistema di ordine n

$$\begin{bmatrix} \alpha_m & \alpha_{m-1} & \ldots & \alpha_{m-n+1} \\ \alpha_{m+1} & \alpha_m & \ldots & \alpha_{m-n+2} \\ \vdots & & \ddots & \vdots \\ \alpha_{m+n-1} & & \ldots & \alpha_m \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = - \begin{bmatrix} \alpha_{m+1} \\ \alpha_{m+2} \\ \vdots \\ \alpha_{m+n} \end{bmatrix}, \tag{76}$$

in cui $\alpha_i = 0$ per $i < 0$. Poi si ricavano gli a_k, $k = 0, \dots, m$, per mezzo delle (74).

La matrice del sistema (76) è una matrice di *Toeplitz*, cioè i suoi elementi sono funzione solo della differenza degli indici. Esistono metodi particolarmente efficienti per risolvere un sistema con matrice di Toeplitz. I primi metodi, ad esempio quello di Trench [47], richiedevano $O(n^2)$ operazioni aritmetiche; più recentemente sono stati sviluppati metodi "superfast" che richiedono $O(n \log^2 n)$ operazioni aritmetiche [3] e [8].

Calcolati i coefficienti b_j, $j = 0, \dots, n$, si ottiene per il resto l'espressione

$$r_{m,n}(x) = f(x) - R_{m,n}(x) = \frac{1}{q_n(x)} \sum_{k=m+n+1}^{\infty} s_k x^k, \text{ dove } s_k = \sum_{j=0}^{\min(k,n)} \alpha_{k-j} b_j,$$

e se $q_n(0) = b_0 = 1$ e $s_{m+n+1} \neq 0$, in un intorno di 0 è

$$r_{m,n}(x) = s_{m+n+1} x^{m+n+1} + O(x^{m+n+2}).$$

Le approssimanti di Padé di una funzione $f(x)$ possono essere ottenute anche riducendo in frazione unica le frazioni continue di Thiele (67, cap. 5) di $f(x)$ di ordine $m + n$, riferite al punto $x_0 = 0$. In questo modo si ottengono in generale apppprossimanti con $m = n$ o $m = n + 1$.

6.62 Esempio. La serie di Maclaurin della funzione $f(x) = \mathrm{e}^x$ è

$$\mathrm{e}^x = \sum_{i=0}^{\infty} \alpha_i \, x^i, \quad \text{dove} \quad \alpha_i = \frac{1}{i!}\,.$$

Per $m = 3, n = 2$, le (74) e (75) risultano

$$\begin{cases} a_0 = b_0 \\ a_1 = b_1 + b_0 \\ a_2 = b_2 + b_1 + \dfrac{1}{2}\, b_0 \\ a_3 = b_2 + \dfrac{1}{2}\, b_1 + \dfrac{1}{6}\, b_0, \end{cases} \qquad \begin{cases} \dfrac{1}{2}\, b_2 + \dfrac{1}{6}\, b_1 + \dfrac{1}{24}\, b_0 = 0 \\ \dfrac{1}{6}\, b_2 + \dfrac{1}{24}\, b_1 + \dfrac{1}{120}\, b_0 = 0. \end{cases}$$

Ponendo $b_0 = 1$, si ricava

$$b_1 = -\frac{2}{5}, \quad b_2 = \frac{1}{20}\,, \quad a_0 = 1, \quad a_1 = \frac{3}{5}\,, \quad a_2 = \frac{3}{20}\,, \quad a_3 = \frac{1}{60}\,.$$

L'approssimante di Padé di ordine 5 della funzione e^x risulta quindi

$$R_{3,2}(x) = \frac{1 + \dfrac{3}{5}\,x + \dfrac{3}{20}\,x^2 + \dfrac{1}{60}\,x^3}{1 - \dfrac{2}{5}\,x + \dfrac{1}{20}\,x^2}\,,$$

e il resto è dato da

$$r_{3,2}(x) = \sum_{j=0}^{2} \frac{b_j}{(6-j)!}\; x^6 + O(x^7) = \frac{1}{7200}\; + O(x^7).$$

La stessa approssimante si ottiene riducendo a frazione unica la frazione continua di Thiele (68, cap. 5) di e^x di ordine 5

$$R_{3,2}(x) = 1 + \frac{x}{1\;-}\;\frac{x}{2\;+}\;\frac{x}{3\;-}\;\frac{x}{2\;+}\;\frac{x}{5}\,. \qquad \blacksquare$$

Poiché per un dato valore di $m+n$, vi sono più approssimanti di Padé di ordine $m+n$ di una stessa funzione $f(x)$, è interessante confrontare fra loro le approssimazioni che si ottengono con diverse approssimanti dello stesso ordine. Dall'esperienza numerica risulta che l'approssimazione è tanto migliore quanto più vicini sono i gradi dei due polinomi.

6.63 Esempio. Le 6 approssimanti di Padé di ordine 5 di e^x sono

$$R_{5,0}(x) = 1 + x + \frac{1}{2}\,x^2 + \frac{1}{6}\,x^3 + \frac{1}{24}\,x^4 + \frac{1}{120}\,x^5, \quad r_{5,0}(x) = \frac{1}{720}\,x^6 + O(x^7),$$

$$R_{4,1}(x) = \frac{1 + \dfrac{4}{5}\,x + \dfrac{3}{10}\,x^2 + \dfrac{1}{15}\,x^3 + \dfrac{1}{120}\,x^4}{1 - \dfrac{1}{5}\,x}\,,\; r_{4,1}(x) = -\,\frac{1}{3600}\,x^6 + O(x^7),$$

$$R_{3,2}(x) = \frac{1 + \dfrac{3}{5}\,x + \dfrac{3}{20}\,x^2 + \dfrac{1}{60}\,x^3}{1 - \dfrac{2}{5}\,x + \dfrac{1}{20}\,x^2}\,, \quad r_{3,2}(x) = \frac{1}{7200}\,x^6 + O(x^7),$$

$$R_{2,3}(x) = \frac{1}{R_{3,2}(-x)}\,, \quad r_{2,3}(x) = -r_{3,2}(x) + O(x^7),$$

$$R_{1,4}(x) = \frac{1}{R_{4,1}(-x)}\,, \quad r_{1,4}(x) = -r_{4,1}(x) + O(x^7),$$

$$R_{0,5}(x) = \frac{1}{R_{5,0}(-x)}\,, \quad r_{0,5}(x) = -r_{5,0}(x) + O(x^7).$$

Nella tabella sono riportati i massimi moduli $e_{m,n}$ degli errori effettivamente generati da queste approssimanti nell'intervallo $[-0.4, 0.4]$.

m	n	$e_{m,n}$
5	0	$0.954\ 10^{-5}$
4	1	$0.477\ 10^{-5}$
3	2	$0.381\ 10^{-5}$
2	3	$0.286\ 10^{-5}$
1	4	$0.477\ 10^{-5}$
0	5	$0.124\ 10^{-4}$

Risulta evidente che l'errore più piccolo si ha per $n = m + 1$. ∎

Le approssimanti di Padé di una funzione $f(x)$ vengono in generale disposte nella tabella, detta *tabella di Padé* della $f(x)$

$$\begin{array}{cccc} R_{0,0}(x) & R_{0,1}(x) & R_{0,2}(x) & \dots \\ R_{1,0}(x) & R_{1,1}(x) & R_{1,2}(x) & \dots \\ R_{2,0}(x) & R_{2,1}(x) & R_{2,2}(x) & \dots \\ \dots & \dots & & \end{array}$$

Una parte della tabella di Padé di e^x è riportata nella figura 6.22. Si noti come in questo caso dalla proprietà

$$\mathrm{e}^x = \frac{1}{\mathrm{e}^{-x}}\,,$$

segua che (si veda l'esercizio 6.72)

$$R_{m,n}(x) = \frac{1}{R_{n,m}(-x)}\,.$$

1	$\dfrac{1}{1-x}$	$\dfrac{1}{1-x+\frac{1}{2}\,x^2}$	$\dfrac{1}{1-x+\frac{1}{2}\,x^2-\frac{1}{6}\,x^3}$
$1+x$	$\dfrac{1+\frac{1}{2}\,x}{1-\frac{1}{2}\,x}$	$\dfrac{1+\frac{1}{3}\,x}{1-\frac{2}{3}\,x+\frac{1}{6}\,x^2}$	$\dfrac{1+\frac{1}{4}\,x}{1-\frac{3}{4}\,x+\frac{1}{4}\,x^2-\frac{1}{24}}$
$1+x+\frac{1}{2}\,x^2$	$\dfrac{1+\frac{2}{3}\,x+\frac{1}{6}\,x^2}{1-\frac{1}{3}\,x}$	$\dfrac{1+\frac{1}{2}\,x+\frac{1}{12}\,x^2}{1-\frac{1}{2}\,x+\frac{1}{12}\,x^2}$	$\dfrac{1+\frac{2}{5}\,x+\frac{1}{20}\,x^2}{1-\frac{3}{5}\,x+\frac{3}{20}\,x^2-\frac{1}{60}}$
$1+x+\frac{1}{2}\,x^2+\frac{1}{6}\,x^3$	$\dfrac{1+\frac{3}{4}\,x+\frac{1}{4}\,x^2+\frac{1}{24}\,x^3}{1-\frac{1}{4}\,x}$	$\dfrac{1+\frac{3}{5}\,x+\frac{3}{20}\,x^2+\frac{1}{60}\,x^3}{1-\frac{2}{5}\,x+\frac{1}{20}\,x^2}$	$\dfrac{1+\frac{1}{2}\,x+\frac{1}{10}\,x^2+\frac{1}{12}}{1-\frac{1}{2}\,x+\frac{1}{10}\,x^2-\frac{1}{12}}$

Fig. 6.22 - Tabella di Padé di $f(x)=\mathrm{e}^x$.

Come nel caso dell'interpolazione razionale, anche nella costruzione delle approssimanti di Padé i due polinomi $p_m(x)$ e $q_n(x)$ ottenuti risolvendo i sistemi (74) e (75) possono avere in comune fattori di grado maggiore o uguale a 1. Semplificando i fattori comuni, le funzioni razionali che si ottengono possono non essere approssimanti dell'ordine cercato. In questo caso, convenzionalmente, la frazione ottenuta semplificando i fattori comuni a numeratore e denominatore viene considerata approssimante di Padé di ordine $m+n$, anche se non verifica la condizione (73). Pertanto si richiederà che le approssimanti di Padé siano frazioni irriducibili, e sotto questa ipotesi di irriducibilità si dimostra che per m ed n interi non negativi l'approssimante di Padé $R_{m,n}(x)$ di una funzione $f(x)$ è unica (si veda l'esercizio 6.70).

6.64 Esempio. Dalla serie di Maclaurin della funzione $f(x) = 1 + \sin x$

$$1 + \sin x = 1 + x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

per $m = 2, n = 1$, da (74) e (75) si ricava

$$\begin{cases} a_0 = b_0 \\ a_1 = b_1 + b_0 \qquad \dfrac{1}{3!}\, b_0 = 0. \\ a_2 = b_1, \end{cases}$$

Una soluzione è data da

$$a_0 = b_0 = 0, \quad a_1 = a_2 = b_1 = 1.$$

La funzione razionale $\dfrac{x + x^2}{x}$, che è un'approssimante di ordine 3, è riducibile, mentre la funzione $1 + x$, che si ottiene dalla precedente dividendo numeratore e denominatore per x, è un'approssimante di ordine solo 2. Ciononostante tale funzione viene considerata l'approssimante di Padé $R_{2,1}(x)$ di ordine 3 di $f(x) = 1 + \sin x$ e risulta

$$R_{1,0}(x) = R_{2,0}(x) = R_{1,1}(x) = R_{2,1}(x) = 1 + x. \qquad \blacksquare$$

Una tabella di Padé viene detta *normale* se tutte le approssimanti vi compaiono una sola volta. È normale, ad esempio, la tabella di e^x, ma non quella di $1 + \sin x$. Se la tabella è normale allora

$$R_{m,n}(x) = \frac{p_m(x)}{q_n(x)},$$

in cui $p_m(x)$ e $q_n(x)$ sono polinomi di grado m ed n. Infatti se vi fossero dei fattori comuni fra i due polinomi ottenuti risolvendo i sistemi (74) e (75), e risultasse

$$R_{m,n}(x) = \frac{p_{m'}(x)}{q_{n'}(x)}\,, \quad \text{con} \quad m' < m \quad \text{e} \quad n' < n,$$

ne seguirebbe che i primi $m' + n' + 1$ termini della serie di Maclaurin di

$$f(x)q_{n'}(x) - p_{m'}(x)$$

sarebbero nulli e quindi, poiché

$$R_{m',n'}(x) = \frac{p_{m'}(x)}{q_{n'}(x)}\,,$$

la tabella di Padé non potrebbe essere normale.

Fissato un intero $j \geq 0$ si può percorrere nella tabella di Padé un *cammino discendente a gradini* che tocca successivamente le approssimanti (per semplicità di notazione sarà omessa la variabile x)

$$R_{j,0},\ R_{j+1,0},\ R_{j+1,1},\ R_{j+2,1},\ R_{j+2,2},\ R_{j+3,2},\ \ldots$$

Graficamente un tale cammino si rappresenta così

$$\begin{array}{ccccc} R_{j,0} & & & & \\ \downarrow & & & & \\ R_{j+1,0} & \longrightarrow & R_{j+1,1} & & \\ & & \downarrow & & \\ & & R_{j+2,1} & \longrightarrow & R_{j+2,2} \\ & & & & \downarrow \\ & & & & R_{j+3,2} \quad \longrightarrow \end{array}$$

Vale il seguente teorema che sarà dimostrato nell'ipotesi di una tabella di Padé normale, ma che risulta valido anche in ipotesi più deboli.

6.65 Teorema. *Per $j \geq 0$ sia*

$$C_j = \big(R_{j,0},\ R_{j+1,0},\ R_{j+1,1},\ R_{j+2,1},\ R_{j+2,2},\ \ldots\big)$$

un cammino discendente a gradini della tabella di Padé di una funzione $f(x)$, sviluppabile in serie di Maclaurin. Se la tabella è normale, esiste una frazione continua della forma

$$w^{(j)}(x) = R_{j,0} + \cfrac{c_1 x^{j+1}}{1 + \cfrac{c_2 x}{1 + \cfrac{c_3 x}{1 + \ldots}}} \tag{77}$$

la cui successione delle somme parziali coincide con C_j, cioè è tale che

$$w_{i+l}^{(j)} = R_{j+i,l}.$$

Dim. Posto

$$R_{j+i,l} = \frac{P_{i+l}}{Q_{i+l}}$$

(gli indici dei due polinomi P_{i+l} e Q_{i+l} stanno qui ad indicare la posizione nella successione C_j e non il grado, che è $j+i$ per P_{i+l} ed l per Q_{i+l}), si consideri la frazione continua

$$v = \delta_0 + \cfrac{\gamma_1}{\delta_1 + \cfrac{\gamma_2}{\delta_2 + \dots}} \tag{78}$$

dove $\delta_0 = R_{j,0}$, $\gamma_1 = R_{j+1,0} - R_{j,0}$, $\delta_1 = 1$ e per $h \geq 2$

$$\gamma_h = \frac{P_{h-1}Q_h - P_hQ_{h-1}}{P_{h-1}Q_{h-2} - P_{h-2}Q_{h-1}},$$

$$\delta_h = \frac{P_hQ_{h-2} - P_{h-2}Q_h}{P_{h-1}Q_{h-2} - P_{h-2}Q_{h-1}}.$$

La k-esima frazione parziale di v è

$$v_k = \delta_0 + \cfrac{\gamma_1}{\delta_1 + \dots + \cfrac{\gamma_k}{\delta_k}} = \frac{P_k}{Q_k}.$$

Infatti per $k = 0$ e $k = 1$ si ha

$$v_0 = \delta_0 = R_{j,0} = \frac{P_0}{Q_0}, \quad v_1 = \delta_0 + \frac{\gamma_1}{\delta_1} = R_{j+1,0} = \frac{P_1}{Q_1}.$$

Procedendo per induzione, per $k > 1$ si suppone che

$$v_{k-2} = \frac{P_{k-2}}{Q_{k-2}} \quad \text{e} \quad v_{k-1} = \frac{P_{k-1}}{Q_{k-1}},$$

e per le relazioni ricorrenti (40, cap. 5) risulta

$$v_k = \frac{\delta_kP_{k-1} + \gamma_kP_{k-2}}{\delta_kQ_{k-1} + \gamma_kQ_{k-2}} = \frac{P_k}{Q_k}.$$

Si può dimostrare (si veda l'esercizio 6.71) che il numeratore di γ_h è un monomio di grado $j+h$, mentre il denominatore di γ_h e δ_h e il numeratore

di δ_h sono monomi di grado $j+h-1$. Quindi γ_h è un monomio di grado 1 e δ_h è costante.

La frazione continua (78) può equivalentemente essere scritta nella forma

$$v = \delta_0 + \cfrac{\gamma_1/\delta_1}{1} + \cfrac{\gamma_2/\delta_1\delta_2}{1} + \cfrac{\gamma_3/\delta_2\delta_3}{1} + \dots$$

e quindi assumere la forma (77) in cui

$$c_1\, x^{j+1} = \alpha_{j+1} x^{j+1} \quad \text{e} \quad c_h x = \frac{\gamma_h}{\delta_{h-1}\delta_h}\,, \quad h = 2, 3, \dots \qquad \blacksquare$$

Il teorema 6.65 consente di calcolare i coefficienti della frazione continua (77) note le approssimanti di Padé. È però possibile sfruttare il teorema nel verso opposto, cioè calcolare le approssimanti di Padé dopo che sono stati costruiti i coefficienti c_i, $i = 2, 3, \dots$ Si può infatti dimostrare che i coefficienti c_i possono essere ottenuti per mezzo del *metodo qd* (si veda l'applicazione di tale metodo al calcolo delle radici delle equazioni algebriche nel capitolo 3). Per questo la (77) viene scritta rinominandone i coefficienti nel modo seguente

$$w^{(j)}(x) = R_{j,0} + \cfrac{\alpha_{j+1}x^{j+1}}{1} - \cfrac{q^{(1)}_{j+1}x}{1} - \cfrac{e^{(1)}_{j+1}x}{1} - \cfrac{q^{(2)}_{j+1}x}{1} - \cfrac{e^{(2)}_{j+1}x}{1} - \dots \tag{79}$$

e si nota che vi sono delle approssimanti in comune nelle successioni C_{j-1} e C_j, come illustrato nella figura 6.23, in cui il cammino relativo a C_{j-1} è indicato con linea continua e quello relativo a C_j è indicato con linea tratteggiata. Si ha infatti

$$C_{j-1} = \big(\dots,\ R_{(j-1)+i,i-1},\ R_{(j-1)+i,i},\ R_{(j-1)+i+1,i},\ R_{(j-1)+i+1,i+1},\ R_{(j-1)+i+2,i+1},\ \dots\big),$$

$$C_j = \big(\dots,\ R_{j+i-1,i-1},\ R_{j+i,i-1},\ R_{j+i,i},\ R_{j+i+1,i},\ R_{j+i+1,i+1},\ \dots\big),$$

e risulta

$$R_{j+i,i} = R_{(j-1)+i+1,i} \quad \text{per ogni } i.$$

Accanto ad ogni segmento del cammino è indicato il coefficiente del termine da sommare nella (79).

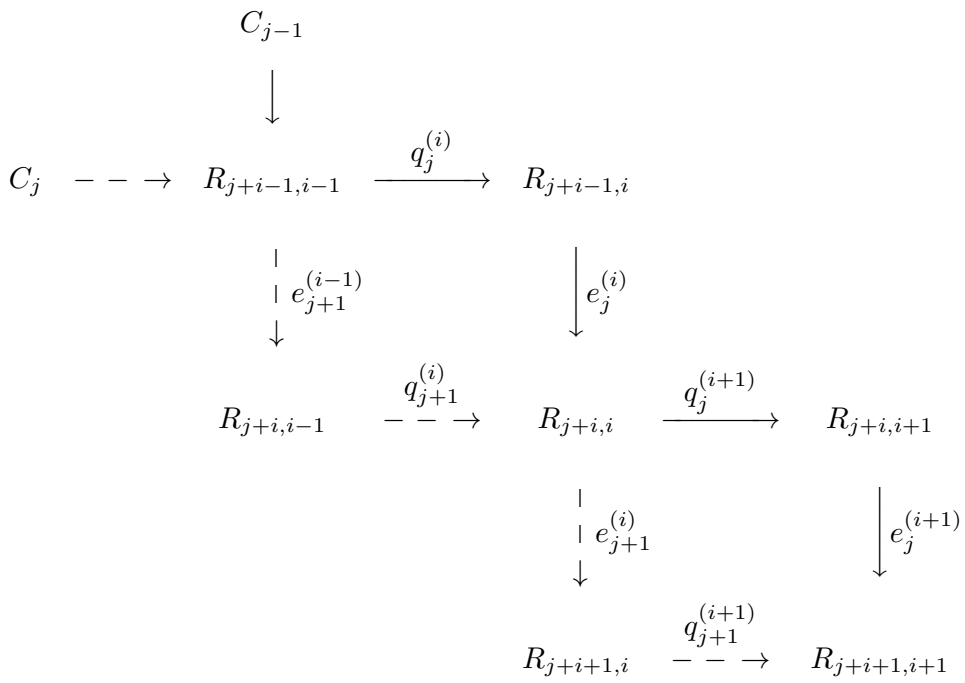


Fig. 6.23 - Cammino discendente a gradini in una tabella di Padé.

Seguendo il cammino C_{j-1} si ha

$$w^{(j-1)}(x) = \ldots \underset{-}{} \frac{q_j^{(i)} x}{1} \underset{-}{} \frac{e_j^{(i)} x}{1} \underset{-}{} \frac{q_j^{(i+1)} x}{1} \underset{-}{} \frac{e_j^{(i+1)} x}{1} \underset{-}{} \ldots$$

mentre seguendo il cammino C_j si ha

$$w^{(j)}(x) = \ldots \underset{-}{} \frac{e_{j+1}^{(i-1)} x}{1} \underset{-}{} \frac{q_{j+1}^{(i)} x}{1} \underset{-}{} \frac{e_{j+1}^{(i)} x}{1} \underset{-}{} \frac{q_{j+1}^{(i+1)} x}{1} \underset{-}{} \ldots$$

Deve risultare

$$w_{2i-1}^{(j-1)} = R_{j+i-1,i-1} = w_{2i-2}^{(j)}$$

$$w_{2i+1}^{(j-1)} = R_{j+i,i} = w_{2i}^{(j)}$$

$$w_{2i+3}^{(j-1)} = R_{j+i+1,i+1} = w_{2i+2}^{(j)}.$$

Poiché

$$R_{j+i-1,i-1} = \frac{P_{2i-2}}{Q_{2i-2}}\,, \quad R_{j+i,i} = \frac{P_{2i}}{Q_{2i}}\,, \quad R_{j+i+1,i+1} = \frac{P_{2i+2}}{Q_{2i+2}}\,,$$

se si segue il cammino C_{j-1}, sfruttando la relazione b) dell'esercizio 5.35, si ottiene

$$P_{2i+2} = \big[1 - (e_j^{(i+1)} + q_j^{(i+1)})x\big] P_{2i} - q_j^{(i+1)} e_j^{(i)} x^2 P_{2i-2},$$

mentre se si segue il cammino C_j si ottiene

$$P_{2i+2} = \big[1 - (q_{j+1}^{(i+1)} + e_{j+1}^{(i)})x\big] P_{2i} - e_{j+1}^{(i)} q_{j+1}^{(i)} x^2 P_{2i-2}.$$

Imponendo che le due espressioni siano uguali si ottengono le due relazioni

$$\begin{aligned} &e_j^{(i+1)} + q_j^{(i+1)} = q_{j+1}^{(i+1)} + e_{j+1}^{(i)}, \\ &q_j^{(i+1)} e_j^{(i)} = e_{j+1}^{(i)} q_{j+1}^{(i)}. \end{aligned} \tag{80}$$

Considerando invece i passi iniziali dei due cammini C_{j-1} e C_j, come illustrato nella figura 6.24, si ha

$$\begin{array}{ccc} R_{j-1,0} & & \\ \big\downarrow & & \\ R_{j,0} & \xrightarrow{\;q_j^{(1)}\;} & R_{j,1} \\ \big\downarrow & & \big\downarrow e_j^{(1)} \\ R_{j+1,0} & \xrightarrow{\;q_{j+1}^{(1)}\;} & R_{j+1,1} \end{array}$$

Fig. 6.24 - Cammino discendente a gradini in una tabella di Padé.

da cui

$$w^{(j-1)}(x) = R_{j-1,0} + \cfrac{\alpha_j x^j}{1} \; - \cfrac{q_j^{(1)} x}{1} \; - \cfrac{e_j^{(1)} x}{1} \; - \ldots$$

$$w^{(j)}(x) = R_{j,0} + \cfrac{\alpha_{j+1} x^{j+1}}{1} \; - \cfrac{q_{j+1}^{(1)} x}{1} \; - \ldots$$

Imponendo che la terza frazione parziale di $w^{(j-1)}$ sia uguale alla seconda di $w^{(j)}$ si ottiene

$$R_{j-1,0} + \cfrac{\alpha_j x^j}{1} \; - \cfrac{q_j^{(1)} x}{1} \; - \cfrac{e_j^{(1)} x}{1} = R_{j,0} + \cfrac{\alpha_{j+1} x^{j+1}}{1} \; - \cfrac{q_{j+1}^{(1)} x}{1},$$

da cui si ricavano le relazioni

$$\begin{aligned} \alpha_{j+1} &= \alpha_j q_j^{(1)}, \\ e_j^{(1)} + q_j^{(1)} &= q_{j+1}^{(1)}, \end{aligned}$$

che possono essere usate come condizioni iniziali per il calcolo dei coefficienti con relazioni ricorrenti ricavate dalla (80). I coefficienti vengono quindi calcolati con la regola

$$\left.\begin{aligned} &e_k^{(0)} = 0, \quad q_k^{(1)} = \frac{\alpha_{k+1}}{\alpha_k}, \quad k = 1, 2, \dots \\ &e_k^{(l)} = q_{k+1}^{(l)} - q_k^{(l)} + e_{k+1}^{(l-1)}, \quad k = 1, 2, \dots \\ &q_k^{(l+1)} = \frac{e_{k+1}^{(l)}}{e_k^{(l)}} q_{k+1}^{(l)}, \quad k = 1, 2, \dots \end{aligned}\right\}, \quad l = 1, 2, \dots \tag{81}$$

costruendo per colonna la seguente tabella

$$\begin{array}{ccccc} e_1^{(0)} & & & & \\ & q_1^{(1)} & & & \\ e_2^{(0)} & & e_1^{(1)} & & \\ & q_2^{(1)} & & q_1^{(2)} & \\ e_3^{(0)} & & e_2^{(1)} & & e_1^{(2)} \\ & q_3^{(1)} & & q_2^{(2)} & \\ e_4^{(0)} & & e_3^{(1)} & & \vdots \\ & q_4^{(1)} & & \vdots & \\ e_5^{(0)} & & \vdots & & \\ & \vdots & & & \\ \vdots & & & & \end{array} \tag{82}$$

Poiché per un fissato $j \geq 0$, i coefficienti $q_{j+1}^{(1)}$, $e_{j+1}^{(1)}$, $q_{j+1}^{(2)}$, $e_{j+1}^{(2)}$, $\dots$, $q_{j+1}^{(i)}$ che servono per costruire $R_{j+i,i}$ con la (79) si trovano lungo la $(j+1)$-esima diagonale della tabella (82), è sufficiente applicare le (81) con $l = 1, \dots, i-1$ e con $k = j+1, \dots, j+2(i-l)$ per il calcolo degli $e_k^{(l)}$ e con $k = j+1, \dots, j+2(i-l)-1$ per il calcolo dei $q_k^{(l+1)}$, con un costo computazionale di $2i^2$ operazioni additive e altrettante moltiplicative (a meno di termini di ordine inferiore). Quindi per calcolare con il qd i coefficienti dell'approssimante di Padé $R_{m,n}(x)$, scritta come frazione continua, sono richieste $O(n^2)$ operazioni aritmetiche.

6.66 Esempio. Per la funzione $f(x) = \mathrm{e}^x$ la tabella del qd è

$$\begin{array}{ccccc}
0 & & & & \\
 & \frac{1}{2} & & & \\
0 & & -\frac{1}{6} & & \\
 & \frac{1}{3} & & \frac{1}{6} & \\
0 & & -\frac{1}{12} & & -\frac{1}{10} \\
 & \frac{1}{4} & & \frac{3}{20} & \\
0 & & -\frac{1}{20} & & \vdots \\
 & \frac{1}{5} & & \vdots & \\
0 & & \vdots & & \\
 & \vdots & & & \\
\vdots & & & &
\end{array}$$

Per $j = 0$ si ha

$$w^{(0)} = 1 + \cfrac{x}{1 - \cfrac{1/2\,x}{1 + \cfrac{1/6\,x}{1 - \cfrac{1/6\,x}{1 + \cfrac{1/10\,x}{1 + \dots}}}}}$$

da cui, per esempio, si ricava

$$R_{2,2} = 1 + \cfrac{x}{1 - \cfrac{1/2\,x}{1 + \cfrac{1/6\,x}{1 - \cfrac{1/6\,x}{1}}}} = \frac{1 + \frac{1}{2}\,x + \frac{1}{12}\,x^2}{1 - \frac{1}{2}\,x + \frac{1}{12}\,x^2}\,.$$

Per $j = 1$ si ha

$$w^{(1)} = 1 + x + \cfrac{1/2\,x^2}{1 - \cfrac{1/3\,x}{1 + \cfrac{1/12\,x}{1 - \cfrac{3/20\,x}{1 + \dots}}}}$$

da cui, per esempio, si ricava

$$\begin{aligned} R_{3,2} &= 1 + x + \cfrac{1/2\,x^2}{1 - \cfrac{1/3\,x}{1 + \cfrac{1/12\,x}{1 - \cfrac{3/20\,x}{1}}}} \\ &= \frac{1 + \frac{3}{5}\,x + \frac{3}{20}\,x^2 + \frac{1}{60}\,x^3}{1 - \frac{2}{5}\,x + \frac{1}{20}\,x^2}\,. \end{aligned}$$

∎

La tabella (82) consente di calcolare approssimanti di Padé $R_{m,n}(x)$ con $m \geq n$. Per il calcolo di approssimanti con primo indice minore del secondo è ancora possibile usare la tabella (82), opportunamente estesa nel triangolo superiore (si veda [13]). Oppure si può calcolare l'approssimante di Padé di $1/f(x)$ e sfruttare la proprietà di reciprocità (si veda l'esercizio 6.72).

Dal punto di vista del costo computazionale, per calcolare una stessa approssimante in più punti conviene prima calcolarne esplicitamente i coefficienti e utilizzare la regola di Ruffini-Horner separatamente per numeratore e denominatore. Se invece è richiesto il calcolo in uno stesso punto di una successione di approssimanti, conviene costruire la tabella del qd. È anche possibile trasformare le approssimanti di Padé in frazioni continue della forma (42, cap. 5), che hanno un costo computazionale più basso, prestando però attenzione alla possibilità che si ottengano espressioni numericamente instabili.

6.67 Esempio. Le seguenti approssimazioni razionali di $\tan x$ (si veda l'esempio 6.61)

$$x(x) = \frac{x}{1} \underset{-}{} \frac{x^2}{3} \underset{-}{} \frac{x^2}{5} \underset{-}{} \frac{x^2}{7} \underset{-}{} \frac{x^2}{9} \underset{-}{} \frac{x^2}{11} \underset{-}{} \frac{x^2}{13},$$

$$R_{7,6}(x) = \frac{x\Big(1 - \dfrac{5\,x^2}{39} + \dfrac{2\,x^4}{715} - \dfrac{x^6}{135135}\Big)}{1 - \dfrac{6\,x^2}{13} + \dfrac{10\,x^4}{429} - \dfrac{4\,x^6}{19305}},$$

$$z(x) = x\Big(0.03571429 - \frac{9.482143}{x^2 - 55.63559} \underset{-}{} \frac{1426.990}{x^2 - 41.11098} \underset{-}{} \frac{156.8337}{x^2 - 15.75343}\Big),$$

sono espressioni diverse della stessa funzione razionale, e quindi sono equivalenti dal punto di vista dell'errore analitico, a meno, per $z(x)$, dell'arrotondamento dei coefficienti. La prima e la seconda, previa riduzione dei denominatori, richiedono 6 operazioni additive e 8 moltiplicative, la terza richiede 6 operazioni additive e 5 moltiplicative. I moduli degli errori assoluti effettivamente generati per $x \in \big[0, \dfrac{\pi}{4}\big]$ sono minori di

$0.119\ 10^{-6}$ per $w(x)$, $\qquad 0.775\ 10^{-6}$ per $R_{7,6}(x)$, $\qquad 0.429\ 10^{-5}$ per $z(x)$.

Il risultato peggiore, nel terzo caso, si spiega con l'arrotondamento dei coefficienti e con la maggiore cancellazione. ∎

Lo studio della convergenza delle successioni di approssimanti di Padé si presenta molto complicato: nel caso più generale i teoremi esistenti, anziché

individuare classi di funzioni per cui le successioni convergono, dimostrano che sotto opportune ipotesi le regioni di non convergenza tendono a diventare arbitrariamente piccole. Un'ipotesi abbastanza forte, sotto cui è possibile dimostrare la convergenza, è quella della equilimitatezza.

6.68 Teorema. *Sia $f(x)$ una funzione analitica nell'origine e sia $\{R_{m,n}(x)\}_{(m,n)}$ una successione di approssimanti di Padé di $f(x)$, per cui esista una costante $M > 0$ tale che $|R_{m,n}(x)| \leq M$ per ogni m, n e per ogni x appartenente ad un dominio D connesso e limitato del piano complesso contenente l'origine. Allora*

$$\lim_{m,\, n\to\infty} R_{m,n}(x) = f(x)$$

uniformemente su ogni compatto contenuto in D.
Per la dimostrazione si veda [5]. ∎

Per quanto riguarda le successioni formate con approssimanti che si trovano sulla diagonale principale della tabella di Padé, esiste una congettura per cui se $f(x)$ è analitica nell'origine e meromorfa (cioè ha al più un numero finito di poli) in un cerchio D del piano complesso contenente l'origine, allora esistono sottosuccessioni che convergono puntualmente alla $f(x)$ in ogni sottoinsieme compatto di D che non contenga poli di $f(x)$.

13. Approssimazione nel discreto

Se della funzione $f(x)$ non è nota l'espressione analitica ma solo i valori assunti nei punti distinti x_i, $i = 0, \ldots, m$, non è possibile in generale affrontare il problema dell'approssimazione come si è fatto nei paragrafi precedenti. In questo caso, al posto delle norme nel continuo considerate nel paragrafo 1 si considerano le norme nel discreto

a) la norma 2

$$\|f\|_2 = \left[\sum_{i=0}^{m} [f(x_i)]^2\right]^{1/2}, \tag{83}$$

b) la norma ∞

$$\|f\|_\infty = \max_{i=0,\ldots,m} |f(x_i)|.$$

Poiché i valori $f(x_i)$ provengono di solito dal rilevamento di dati sperimentali, e quindi sono affetti da errori casuali, viene spesso usata la norma 2, che consente un'approssimazione in media. L'approssimazione in norma ∞ nel discreto viene talvolta usata anche quando della funzione $f(x)$ è nota l'espressione analitica, perché gli algoritmi che si utilizzano sono più semplici

che nel caso continuo. In questo modo si ottengono delle approssimazioni quasi minimax di $f(x)$.

Nel discreto l'uso della norma pesata serve anche per ridurre l'influenza di alcuni dati, meno significativi o affetti da errori elevati, e accentuare invece l'influenza di altri dati; se $\omega_i \geq 0$ per $i = 0, \ldots, m$, si considera quindi per la norma 2

$$\|f\|_2 = \left[\sum_{i=0}^{m} \omega_i \, [f(x_i)]^2 \right]^{1/2}. \tag{84}$$

La norma 2 definita in (83) e (84) risulta indotta dai due prodotti scalari

$$\langle f, g \rangle = \sum_{i=0}^{m} f(x_i) g(x_i), \tag{85}$$

$$\langle f, g \rangle = \sum_{i=0}^{m} \omega_i f(x_i) g(x_i). \tag{86}$$

Poiché le (85) e (86) non sono delle buone definizioni di prodotto scalare sull'insieme delle funzioni continue, in quanto esistono funzioni $f(x)$ non nulle per cui $\langle f, f \rangle = 0$, sarebbe più corretto chiamare i prodotti (85) e (86) *pseudoscalari* e le norme (83) e (84) *pseudonorme*.

Nella norma 2 il problema 6.3 dell'approssimazione lineare di una funzione $f(x)$ di cui sono noti i valori $f(x_i)$, $i = 0, \ldots, m$, si pone nel modo seguente: fissate le $n + 1$ funzioni linearmente indipendenti $\phi_j(x)$, $j = 0, \ldots, n$, $n \leq m$, e indicati con $\mathbf{f}$ il vettore di componenti $f(x_i)$, $i = 0, \ldots, m$ e con Φ la matrice i cui elementi sono $\phi_j(x_i)$, $i = 0, \ldots, m$, $j = 0, \ldots, n$, si deve determinare il vettore $\boldsymbol{\alpha}^{(m)} \in \mathbf{R}^{n+1}$ tale che

$$\|\mathbf{f} - \Phi \boldsymbol{\alpha}^{(m)}\|_2 = \min_{\boldsymbol{\alpha} \in \mathbf{R}^{n+1}} \|\mathbf{f} - \Phi \boldsymbol{\alpha}\|_2. \tag{87}$$

Questo problema viene detto *dei minimi quadrati nel discreto* e la funzione

$$g_n^{(m)}(x) = \sum_{j=0}^{n} \alpha_j^{(m)} \phi_j(x) \tag{88}$$

è l'approssimazione *ai minimi quadrati nel discreto*.

Se la norma 2 usata è quella non pesata, cioè la (83), il problema (87) è quello considerato in [7, cap. 7], per il quale vale il seguente teorema.

6.69 Teorema. *Le soluzioni del problema (87) sono anche soluzioni del sistema lineare, detto sistema normale,*

$$\Phi^T\Phi\boldsymbol{\alpha} = \Phi^T\mathbf{f}. \tag{89}$$

Se la matrice Φ ha rango massimo, allora il problema (87) ha un'unica soluzione. ∎

Per la dimostrazione del teorema 6.69 e per le tecniche effettive di calcolo di $\boldsymbol{\alpha}^{(m)}$ si veda [7].

Se come funzioni $\phi_j(x)$ si scelgono i monomi x^j, la matrice Φ è la matrice di Vandermonde (in generale non quadrata) di elementi

$$\phi_{ij} = x_i^j, \quad i = 0, \dots, m,\ j = 0, \dots, n,$$

che ha rango massimo. Il problema (87) viene quindi trascritto nella forma

$$\min_{\boldsymbol{\alpha}\in\mathbf{R}^{n+1}} \big[\varphi(\boldsymbol{\alpha})\big]^{1/2}, \quad \text{dove} \quad \varphi(\boldsymbol{\alpha}) = \sum_{i=0}^{m} \Big[f(x_i) - \sum_{j=0}^{n} \alpha_j x_i^j\Big]^2,$$

e il sistema normale (89) è

$$\sum_{j=0}^{n} \Big[\sum_{i=0}^{m} x_i^{k+j}\Big]\alpha_j = \sum_{i=0}^{m} f(x_i)x_i^k, \quad k = 0, \dots, n.$$

6.70 Esempio. Si vuole costruire il polinomio di grado al più 3 di migliore approssimazione ai minimi quadrati della seguente funzione $f(x)$

x	1	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	12.51	13.05	11.7	9.26	8.3	6.25	5.34	4.59	5.14	6.36	10.31	13.88

(si tratta della funzione dell'esempio 5.67). Per $n = 3$ si ottiene

$$\Phi^T\Phi = \begin{bmatrix} 12 & 78 & 650 & 6084 \\ 78 & 650 & 6084 & 60710 \\ 650 & 6084 & 60710 & 630708 \\ 6084 & 60710 & 630708 & 6735950 \end{bmatrix}, \quad \Phi^T\mathbf{f} = \begin{bmatrix} 106.69 \\ 653.68 \\ 5604.66 \\ 55408.96 \end{bmatrix}$$

e

$$\boldsymbol{\alpha}^{(11)} = [12.39141, 1.099278, -0.6152425, 0.04473323]^T.$$

Il polinomio cercato risulta quindi

$$g_3^{(11)}(x) = 0.04473323\,x^3 - 0.6152425\,x^2 + 1.099278\,x + 12.39141.$$

Nella figura 6.25 è riportato il grafico di $g_3^{(11)}(x)$. Si confronti con i grafici del polinomio di interpolazione della stessa funzione riportato nella figura 5.25 e con il grafico della spline cubica riportato nella figura 5.26. ∎

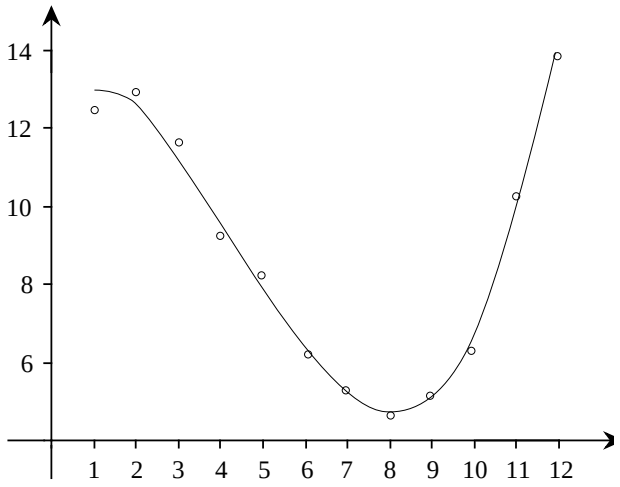


Fig. 6.25 - Approssimazione ai minimi quadrati nel discreto.

Se si usa la norma 2 pesata, definita in (84), è facile vedere che, indicata con W la matrice diagonale di ordine $m+1$ i cui elementi principali sono ω_i, la soluzione $\boldsymbol{\alpha}^{(m)}$ può essere calcolata per mezzo del sistema normale

$$\Phi^T W \Phi \boldsymbol{\alpha} = \Phi^T W \mathbf{f}. \tag{90}$$

Come nel caso continuo, anche nel caso discreto, allo scopo di ottenere sistemi (89) o (90) con matrice diagonale, si cercano funzioni $\phi_j(x)$ ortogonali rispetto ai prodotti scalari (85) e (86), cioè

$$\langle \phi_k, \phi_j \rangle = \sum_{i=0}^{m} \omega_i \phi_k(x_i)\phi_j(x_i) = 0, \quad \text{per} \quad k \neq j, \tag{91}$$

con $\omega_i = 1$ se il prodotto scalare è non pesato.

Nel caso generale non vi è alcuna relazione fra le funzioni $\phi_j(x)$ che sono ortogonali secondo la (91) con le corrispondenti funzioni ortogonali nel caso continuo. Vi è però un caso particolare in cui è possibile stabilire una relazione fra il caso discreto e quello continuo.

6.71 Teorema. *Sia $\{p_j(x)\}_{j\in\mathbf{N}}$ un insieme di polinomi a coefficienti reali, ortogonali sull'intervallo $[a,b]$ rispetto al peso $\omega(x)$. Fissati due interi n ed m, con $m \geq n$, siano x_i, $i = 0,\dots,m$ gli zeri, interni ad $[a,b]$, di $p_{m+1}(x)$ e si considerino gli $m+1$ numeri*

$$\omega_i = \int_a^b \omega(x)\ \frac{p_{m+1}(x)}{(x-x_i)p'_{m+1}(x_i)}\ dx, \quad i = 0,\dots,m.$$

Allora gli $n+1$ polinomi $p_0(x),\dots,p_n(x)$ sono ortogonali nei nodi x_i, rispetto ai pesi ω_i, cioè verificano la relazione

$$\sum_{i=0}^{m} \omega_i p_k(x_i)p_j(x_i) = 0 \quad \text{per} \quad j,k = 0,\dots,n \text{ e } j \neq k.$$

Dim. La formula

$$S_{m+1} = \sum_{i=0}^{m} \omega_i f(x_i)$$

è una formula di integrazione approssimata sull'intervallo $[a,b]$ con grado di precisione almeno $2m$ (si veda il paragrafo 1 del capitolo 7), cioè integra esattamente polinomi fino al grado $2m$. Quindi, essendo $p_k(x)p_j(x)$ un polinomio di grado minore od uguale a $2m$ per $k,j \leq n$, risulta

$$\sum_{i=0}^{m} \omega_i p_k(x_i)p_j(x_i) = \int_a^b \omega(x)p_k(x)p_j(x)\ dx = 0, \quad \text{per} \quad k \neq j.$$

Inoltre per il teorema 7.8 risulta $\omega_i > 0$ per $i = 0,\dots,m$. ∎

Dal teorema 6.71 segue che se i nodi x_i sono scelti opportunamente, cioè sono gli zeri del polinomio $p_{m+1}(x)$, allora i polinomi $p_j(x)$ per $j = 0,\dots,n$ possono essere usati per risolvere il problema (87) per mezzo del sistema (90) in cui la matrice W ha come elementi principali i pesi ω_i e la matrice $\Phi^T W \Phi$ è diagonale. I coefficienti della (88) risultano quindi

$$\alpha_j^{(m)} = \frac{\sum_{i=0}^{m} \omega_i p_j(x_i) f(x_i)}{\sum_{i=0}^{m} \omega_i p_j^2(x_i)}, \quad j = 0,\dots,n.$$

Nel caso particolare che $[a,b] = [-1,1]$ e $p_j(x) = T_j(x)$, cioè i polinomi siano quelli di Chebyshev di 1^a specie, per la (26) i nodi x_i sono

$$x_i = \cos\theta_i, \quad \theta_i = \frac{(2i+1)\pi}{2(m+1)}, \quad i = 0,\dots,m,$$

e i pesi ω_i risultano (si veda l'esempio 7.21)

$$\omega_i = \frac{\pi}{m+1}\,.$$

Si dimostra con lo stesso ragionamento della dimostrazione del teorema 6.71 che

$$\sum_{i=0}^{m} \omega_i T_j^2(x_i) = \int_{-1}^{1} \frac{T_j^2(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} \pi & \text{se } j = 0, \\ \dfrac{\pi}{2} & \text{se } j \geq 1, \end{cases}$$

per cui, in analogia a quanto fatto nel caso continuo, si pone $x = \cos\theta$ e si considera, al posto della (88), il polinomio trigonometrico

$$g_n^{(m)}(\cos\theta) = \frac{\alpha_0^{(m)}}{2} + \sum_{j=1}^{n} \alpha_j^{(m)} \cos j\theta, \tag{92}$$

dove

$$\alpha_j^{(m)} = \frac{2}{m+1} \sum_{i=0}^{m} f(\cos\theta_i) \cos j\theta_i. \tag{93}$$

La (93) rappresenta l'approssimazione dell'integrale (36) con la formula dei punti di mezzo applicata su $m+1$ nodi (si veda il paragrafo 4 del capitolo 7). Ne segue che per $m \to \infty$ i coefficienti $\alpha_j^{(m)}$ tendono ai coefficienti α_j^* del polinomio (35) di approssimazione ai minimi quadrati nel continuo.

Se i nodi x_i sono equidistanti nell'intervallo $[a, b]$ utilizzando il metodo di Gram-Schmidt [25], è possibile costruire una successione di polinomi, detti *polinomi di Gram*, che soddisfa la (91). Anche in questo caso è possibile dimostrare che per $m \to \infty$ la soluzione del problema discreto tende alla soluzione del problema continuo.

In norma ∞ il problema dell'approssimazione nel discreto richiede di determinare un polinomio $q_n^{(m)}(x) \in \mathcal{P}_n$

$$q_n^{(m)}(x) = c_n^{(m)} x^n + c_{n-1}^{(m)} x^{n-1} + \ldots + c_0^{(m)},$$

tale che

$$r_n^{(m)} = \max_{i=0,\ldots,m} |f(x_i) - q_n^{(m)}(x_i)| = \min_{p_n \in \mathcal{P}_n} \max_{i=0,\ldots,m} |f(x_i) - p_n(x_i)|. \tag{94}$$

Anche nel caso discreto [24] vale il teorema di equioscillazione di Chebyshev, per cui se i punti x_i sono in ordine crescente, è

$$f(x_{j_i}) - q_n^{(m)}(x_{j_i}) = (-1)^i d, \quad |d| = r_n^{(m)}, \quad \text{per} \quad j_i \in J,$$

dove $J = \{j_0, j_1, \ldots, j_{n+1}\}$ è un opportuno sottoinsieme ordinato dell'insieme degli interi $\{0, 1, \ldots, m\}$. Per l'effettiva determinazione dei coefficienti $c_0^{(m)}, \ldots, c_n^{(m)}$ si possono usare degli *algoritmi di scambio*, versioni discretizzate degli algoritmi di Remez, in cui con un numero finito di passi si individua il giusto sottoinsieme di indici (si vedano gli esercizi 6.77 e 6.78).

Un altro modo di procedere è quello di trasformare il problema (94) in un problema di programmazione lineare. Sia $V \in \mathbf{R}^{(m+1)\times(n+1)}$ la matrice di Vandermonde (in generale non quadrata) di elementi

$$v_{ij} = x_i^j, \quad i = 0, \ldots, m, \ j = 0, \ldots, n,$$

e siano

$$\mathbf{c} = [c_0, c_1, \ldots, c_n]^T \quad \text{e} \quad \mathbf{f} = [f(x_0), f(x_1), \ldots, f(x_m)]^T$$

i vettori rispettivamente dei coefficienti del polinomio $p_n(x)$ e dei valori assunti dalla funzione $f(x)$ nei nodi, allora il problema (94) è equivalente al seguente problema

$$\min_{\mathbf{c}\in\mathbf{R}^{n+1}} \|\mathbf{f} - V\mathbf{c}\|_\infty. \tag{95}$$

Introducendo la variabile reale z e il vettore $\mathbf{u} \in \mathbf{R}^{m+1}$, $\mathbf{u} = [1, 1, \ldots, 1]^T$, il problema (95) è equivalente al seguente problema di programmazione lineare

$$\begin{array}{ll} \min & z, \\ {\scriptstyle V\mathbf{c} - z\mathbf{u} \le \mathbf{f}} & \\ {\scriptstyle -V\mathbf{c} - z\mathbf{u} \le -\mathbf{f}} & \end{array} \tag{96}$$

che può essere risolto ad esempio con il metodo del simplesso. Per la risoluzione del problema (95) sono stati anche studiati algoritmi efficienti, che utilizzano trasformazioni ortogonali, basati su metodi di *steepest descent* (metodo di Cline) [6], [12].

6.72 Esempio. Per $m = 5$ si consideri la seguente funzione $f(x)$

x	0	0.2	0.4	0.6	0.8	1
$f(x)$	1	1.221403	1.491825	1.822119	2.225541	2.718282

(i valori $f(x_i)$, $i = 0, 1, \ldots, 5$, sono ottenuti dalla funzione $f(x) = \mathrm{e}^x$). Per

$n = 3$ la matrice V è data da

$$V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{1}{5} & \frac{1}{25} & \frac{1}{125} \\ 1 & \frac{2}{5} & \frac{4}{25} & \frac{8}{125} \\ 1 & \frac{3}{5} & \frac{9}{25} & \frac{27}{125} \\ 1 & \frac{4}{5} & \frac{16}{25} & \frac{64}{125} \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Il problema (96) è in questo caso

$$\min_{M\mathbf{y} \leq \mathbf{g}} z, \tag{97}$$

dove

$$M = \begin{bmatrix} V & -\mathbf{u} \\ -V & -\mathbf{u} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{c} \\ z \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} \mathbf{f} \\ -\mathbf{f} \end{bmatrix}.$$

Risolvendo il problema (97) con il metodo del simplesso si ottiene la soluzione

$$\mathbf{c} = \begin{bmatrix} 0.9995455 \\ 1.015769 \\ 0.4230042 \\ 0.2795258 \end{bmatrix}$$

e

$$r_3^{(5)} = \min z \approx 0.454\ 10^{-3}.$$

Si noti come il polinomio così ottenuto

$$q_3^{(5)}(x) = 0.2795258\, x^3 + 0.4230042\, x^2 + 1.015769\, x + 0.9995455$$

sia molto vicino a quello determinato nell'esempio 6.40 con il minimax nel caso continuo. Risulta inoltre nell'intervallo [0,1]

$$\|\mathrm{e}^x - q_3^{(5)}(x)\|_\infty \approx 0.602\ 10^{-3}. \qquad \blacksquare$$

Una questione interessante è quella della convergenza, al crescere del numero m dei nodi, dei polinomi di minimax di grado n ottenuti nel discreto, al polinomio di minimax dello stesso grado ottenuto nel continuo. Valgono, a questo proposito, i seguenti teoremi.

6.73 Teorema. *Siano $f(x)$ continua in $[a,b]$, $p_n^*(x)$ il polinomio di approssimazione minimax di $f(x)$ su $[a,b]$ e $q_n^{(m)}(x)$ il polinomio di approssimazione minimax di $f(x)$ sui nodi $x_i \in [a,b]$ per $i = 0,\dots,m$, ordinati in modo crescente. Posto*

$$\delta_m = \max_{i=0,\dots,m+1} \big(x_i - x_{i-1}\big), \quad \textit{dove} \quad x_{-1} = a,\ x_{m+1} = b,$$

se $\lim_{m\to\infty} \delta_m = 0$, *allora*

$$\lim_{m\to\infty} c_j^{(m)} = a_j^*, \quad j = 0,1,\dots,n, \tag{98}$$

dove gli a_j^* sono i coefficienti di $p_n^*(x)$ e i $c_j^{(m)}$ sono i coefficienti di $q_n^{(m)}(x)$.

Dim. Posto

$$r(x) = f(x) - q_n^{(m)}(x) \quad \text{e} \quad r_n^{(m)} = \max_{i=0,\dots,m} |r(x_i)|,$$

dal teorema 6.28 (de La Vallèe-Poussin) segue che

$$r_n^{(m)} \le r_n^* = \|f - p_n^*\|_\infty. \tag{99}$$

Si consideri $x \in [a,b]$ e sia x_i il nodo più vicino a x; è

$$|r(x)| \le |r(x_i)| + |\omega_m(x)|, \quad \text{dove} \quad \omega_m(x) = r(x) - r(x_i).$$

Poiché $|r(x_i)| \le r_n^{(m)}$, risulta

$$\max_{x\in[a,b]} |r(x)| \le r_n^{(m)} + \max_{x\in[a,b]} |\omega_m(x)|,$$

ed essendo

$$\max_{x\in[a,b]} |r(x)| \ge r_n^*,$$

ne segue che

$$r_n^* \le r_n^{(m)} + \max_{x\in[a,b]} |\omega_m(x)|.$$

Poiché la funzione $r(x)$ è continua e $|x - x_i| < \delta_m$, per ogni ϵ esiste m tale che

$$\max_{x\in[a,b]} |\omega_m(x)| < \epsilon,$$

e quindi

$$r_n^* \le r_n^{(m)} + \epsilon.$$

Confrontando con la (99) ne segue che $\lim_{m\to\infty} r_n^{(m)} = r_n^*$ e, per l'unicità del polinomio di approssimazione minimax, ne segue la (98). ∎

6.74 Teorema. *Siano* $f(x) \in C^2[a,b]$ *e* $a = x_0 < \ldots < x_m = b$, *con* $\delta_m = \max_{i=1,\ldots,m} \big(x_i - x_{i-1}\big)$. *Se* $\lim_{m\to\infty} \delta_m = 0$, *allora esiste una costante* $\gamma \geq 0$, *indipendente da* m, *tale che*

$$\|f - q_n^{(m)}\|_\infty - r_n^* \leq \gamma \delta_m^2. \tag{100}$$

Dim. Posto

$$r(x) = f(x) - q_n^{(m)}(x) \quad \text{e} \quad r_n^{(m)} = \max_{i=0,\ldots,m} |r(x_i)|,$$

sia $y \in [a,b]$ tale che

$$|r(y)| = \|f - q_n^{(m)}\|_\infty.$$

Dalla (99) segue che

$$|r(y)| \geq r_n^* \geq r_n^{(m)} \geq |r(x_i)|, \quad \text{per} \quad i = 0, \ldots, m. \tag{101}$$

Se y coincide con un nodo x_i, allora

$$\|f - q_n^{(m)}\|_\infty = r_n^*,$$

e la (100) vale con $\gamma = 0$. Se invece y non coincide con alcun nodo, sia x_i un nodo tale che $|y - x_i| < \delta_m$. Essendo y un punto interno ad $[a,b]$, è $r'(y) = 0$ e dalla formula di Taylor, per un opportuno ξ_m, si ha

$$r(x_i) = r(y) + \frac{(x_i - y)^2}{2}\, r''(\xi_m),$$

da cui segue che

$$|r(y)| - |r(x_i)| \leq \frac{M_m}{2}\, \delta_m^2, \quad \text{dove} \quad M_m = \max_{x\in[a,b]} |r''(x)|,$$

e per la (101) è

$$\|f - q_n^{(m)}\|_\infty - r_n^* \leq \frac{M_m}{2}\, \delta_m^2.$$

Resta da dimostrare che la successione $\{M_m\}$ è superiormente limitata: per la (98) le successioni dei coefficienti di $q_n^{(m)}(x)$ sono superiormente limitate al crescere di m e quindi lo sono anche le successioni dei coefficienti delle derivate seconde di $q_n^{(m)}(x)$; ne segue che esiste γ tale che per ogni m è

$$\gamma \geq \max_{x\in[a,b]} \left\{ |f''(x)| + \left| \frac{d^2}{dx^2}\, q_n^{(m)}(x) \right| \right\}. \qquad \blacksquare$$

Dal confronto dell'errore ottenuto nell'esempio 6.72 con quelli riportati nella tabella dell'esempio 6.43, risulta che se è nota l'espressione analitica della funzione $f(x)$, non appare conveniente, nel caso di una funzione di una sola variabile, utilizzare queste tecniche di approssimazione nel discreto che sono più costose delle tecniche per il quasi minimax nel continuo. Però nel caso di funzioni di più variabili, in cui approssimazioni quasi minimax nel continuo possono essere assai più difficili da determinare, le tecniche di approssimazione nel discreto sono in generale preferibili.

Esercizi proposti

6.1 Sia $f(x) \in C[0,1]$. L'n-esimo *polinomio di Bernstein* di $f(x)$ è così definito

$$B_n(f;x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}.$$

Si verifichi che

a) $B_n(f;0) = f(0), \qquad B_n(f;1) = f(1),$

b) $B_n(1;x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = 1,$

c) $B_n(x;x) = \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} x^k (1-x)^{n-k} = x,$

d) $B_n(x^2;x) = \sum_{k=0}^{n} \left(\frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k} = \frac{x}{n}\left[1 + (n-1)x\right],$

e) $\sum_{k=0}^{n} \left(\frac{k}{n} - x\right)^2 \binom{n}{k} x^k (1-x)^{n-k} = \frac{x(1-x)}{n};$

f) si dimostri che

$$\lim_{n\to\infty} B_n(f;x) = f(x),$$

uniformemente per $x \in [0,1]$;

g) si approssimi con i polinomi di Bernstein la funzione

$$f(x) = |x|, \quad \text{per} \quad x \in [-1,1].$$

Quindi, per la f), i polinomi di Bernstein forniscono una dimostrazione del teorema di Weierstrass. Inoltre consentono di approssimare contemporaneamente le derivate della funzione $f(x)$, cioè se $f(x) \in C^p[0,1]$, allora $B_n^{(p)}(f;x) \to f^{(p)}(x)$ uniformemente [14]. Se $f(x)$ è convessa, i polinomi $B_n(f;x)$ sono convessi e la convergenza è monotona in ogni punto. Tuttavia i polinomi di Bernstein per la lentezza della loro convergenza non vengono usati nella pratica. Risulta ad esempio da d) che

$$B_n(x^2;x) - x^2 = \frac{x(1-x)}{n},$$

cioè l'errore di approssimazione tende a 0 come $1/n$ (e quindi la convergenza è lentissima).

(Traccia: b) si scriva l'espansione del binomio di Newton per la funzione

$$\big[x + (1-x)\big]^n = 1,$$

c) $$B_n(x;x) = \sum_{k=1}^{n} \binom{n-1}{k-1} x^k (1-x)^{n-k}$$
$$= \sum_{j=0}^{n-1} \binom{n-1}{j} x^{j+1}(1-x)^{n-1-j} = xB_{n-1}(1;x) = x,$$

d) si verifichi che

$$\frac{k^2}{n^2}\binom{n}{k} = \frac{n-1}{n}\,\frac{k-1}{n-1}\binom{n-1}{k-1} + \frac{1}{n}\binom{n-1}{k-1},$$

per cui

$$B_n(x^2;x) = \frac{n-1}{n}\,xB_{n-1}(x;x) + \frac{1}{n}\,xB_{n-1}(1;x),$$

e) poiché

$$\left(\frac{k}{n} - x\right)^2 = \left(\frac{k}{n}\right)^2 - 2x\left(\frac{k}{n}\right) + x^2,$$

l'espressione da calcolare è data da

$$B_n(x^2;x) - 2xB_n(x;x) + x^2 B_n(1;x);$$

f) sia $\epsilon > 0$, poiché $f(x)$ è uniformemente continua su $[0,1]$, esiste un δ tale che per ogni x, y, con $|x-y| < \delta$, è $|f(x) - f(y)| < \epsilon$. Inoltre, essendo $f(x)$ limitata, esiste un M tale che $|f(x) - f(y)| \le M$ per ogni coppia di punti x e $y \in [0,1]$.

Sia $x \in [0,1]$; si consideri l'insieme degli indici

$$\{k,\ k = 1, \ldots, n\} = K_1 \cup K_2,$$

dove

$$K_1 = \left\{k\ :\ \left|\frac{k}{n} - x\right| < \delta\right\}, \qquad K_2 = \left\{k :\ \left|\frac{k}{n} - x\right| \ge \delta\right\}.$$

Dalla b) segue che

$$f(x) = \sum_{k=0}^{n} f(x)\binom{n}{k} x^k (1-x)^{n-k},$$

e quindi

$$f(x) - B_n(f;x) = \sum_{k=0}^{n} \left[f(x) - f\left(\frac{k}{n}\right) \right] \binom{n}{k} x^k (1-x)^{n-k} = S_1 + S_2,$$

dove

$$S_1 = \sum_{k \in K_1} \left[f(x) - f\left(\frac{k}{n}\right) \right] \binom{n}{k} x^k (1-x)^{n-k},$$

$$S_2 = \sum_{k \in K_2} \left[f(x) - f\left(\frac{k}{n}\right) \right] \binom{n}{k} x^k (1-x)^{n-k}.$$

Per $k \in K_2$ è $\dfrac{1}{\delta^2} \left(\dfrac{k}{n} - x \right)^2 \geq 1$, per cui

$$|S_2| \leq M \sum_{k \in K_2} \binom{n}{k} x^k (1-x)^{n-k} \leq \frac{M}{\delta^2} \sum_{k \in K_2} \left(\frac{k}{n} - x \right)^2 \binom{n}{k} x^k (1-x)^{n-k}$$
$$\leq \frac{M}{\delta^2} \sum_{k=0}^{n} \left(\frac{k}{n} - x \right)^2 \binom{n}{k} x^k (1-x)^{n-k} = \frac{M}{\delta^2} \, \frac{x(1-x)}{n}$$

per la e), e poiché $x(1-x) \leq \dfrac{1}{4}$ per $x \in [0,1]$, è

$$|S_2| \leq \frac{M}{4n\delta^2} \, .$$

Inoltre

$$|S_1| \leq \epsilon \sum_{k \in K_1} \binom{n}{k} x^k (1-x)^{n-k} \leq \epsilon \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = \epsilon$$

per la b). Quindi

$$|f(x) - B_n(f;x)| \leq \epsilon + \frac{M}{4n\delta^2} < 2\epsilon$$

per n sufficientemente grande. Poiché la relazione vale indipendentemente dal particolare x scelto, la convergenza è uniforme.

g) si faccia prima la trasformazione di variabile $y = \dfrac{1}{2}(x+1)$, per far corrispondere l'intervallo $[0,1]$ all'intervallo $[-1,1]$, e si costruiscano i polinomi di Bernstein di $g(y) = |2y-1|$. Si ottiene per $n = 1, \ldots, 4$

$$B_1(g;y) = 1, \qquad B_2(g;y) = 1 - 2y + 2y^2,$$

$$B_3(g;y) = B_2(g;y), \quad B_4(g;y) = 1 - 2y + 4y^3 - 2y^4,$$

da cui

$$B_1(f;x) = 1, \qquad B_2(f;x) = \frac{1}{2}\,(x^2+1),$$
$$B_3(f;x) = B_2(f;x), \quad B_4(f;x) = \frac{1}{8}\,(3+6x^2-x^4),$$

(si veda la figura 6.26) e risulta

$$\max_{x\in[-1,1]} |f(x) - B_4(f;x)| = B_4(f;0) = \frac{3}{8}\,.\ )$$

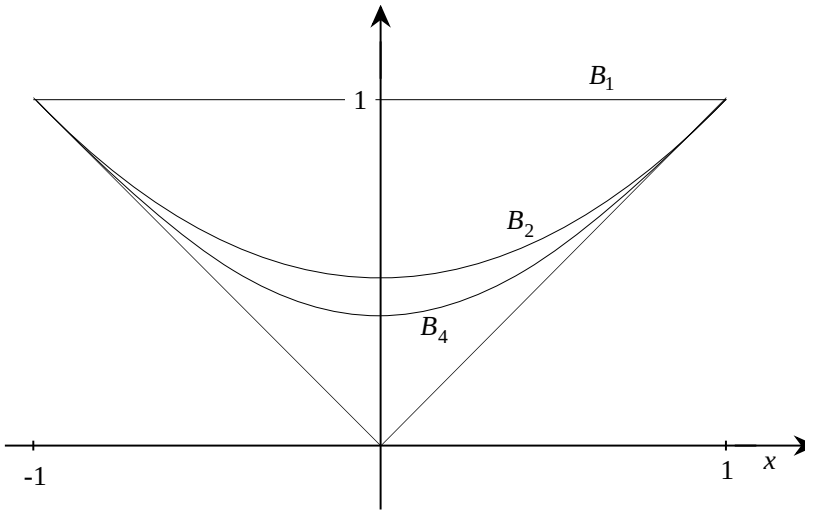


Fig. 6.26 - Polinomi di Bernstein di $|x|$.

6.2 Si scriva la formula di Taylor in un intorno del punto x_0 per le funzioni

a) $f(x) = \sqrt{1+x}, \qquad x_0 = 0,$

b) $f(x) = \dfrac{1}{x}\,, \qquad x_0 = 1,$

c) $f(x) = \dfrac{1}{\sqrt{1+x}}\,, \qquad x_0 = 0,$

d) $f(x) = \log(1+x), \qquad x_0 = 0.$

Si dica quale grado deve avere il polinomio ottenuto troncando ciascuna formula in modo da approssimare $\sqrt{1.1}$, $\dfrac{1}{1.1}$, $\dfrac{1}{\sqrt{1.1}}$ e $\log 1.1$ con un errore relativo non superiore a 10^{-4}.

(Risposta:

a) $f(x) = 1 + \dfrac{1}{2}\,x - \dfrac{1}{2\cdot 4}\,x^2 + \dfrac{1\cdot 3}{2\cdot 4\cdot 6}\,x^3 - \dfrac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 8}\,x^4 + \ldots,$

b) $f(x) = 1 - (x-1) + (x-1)^2 - (x-1)^3 + (x-1)^4 - \cdots,$

c) $f(x) = 1 - \dfrac{1}{2}\,x + \dfrac{1\cdot 3}{2\cdot 4}\,x^2 - \dfrac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\,x^3 + \dfrac{1\cdot 3\cdot 5\cdot 7}{2\cdot 4\cdot 6\cdot 8}\,x^4 - \ldots,$

d) $f(x) = x - \dfrac{x^2}{2} + \dfrac{x^3}{3} - \dfrac{x^4}{4}\ \ldots;$

i gradi dei polinomi sono 2 per a), 3 per b) e c), 4 per d).)

6.3 Quanti termini della serie di Maclaurin di $f(x) = \sin x$ sono richiesti per approssimare la funzione $f(x)$ nell'intervallo $[0, 2\pi]$ con un errore assoluto minore di 10^{-4}? Tenendo conto della relazione

$$\sin x = -\sin(x - \pi), \quad \text{per} \quad \pi \le x \le 2\pi,$$

basta determinare l'approssimazione nell'intervallo $[0, \pi]$. Quanti termini sono allora richiesti? Infine tenendo conto della relazione

$$\sin x = \sin(\pi - x), \quad \text{per} \quad \frac{\pi}{2} \le x \le \pi,$$

basta determinare l'approssimazione nell'intervallo $[0, \pi/2]$. Quanti termini sono allora richiesti?

(Risposta: 11 termini (grado 21), 7 termini (grado 13), 5 termini (grado 9).)

6.4 Sia $\mathcal{G}$ un sottospazio vettoriale dello spazio $\mathcal{F}$ delle funzioni reali di variabile reale, e sia $\|\ .\ \|$ una norma definita su $\mathcal{G}$. Si dice che $\mathcal{G}$ è *strettamente convesso* rispetto alla norma ivi definita, se per $f, g \in \mathcal{G}$, con $f \neq g$, $\|f\| \le m$ e $\|g\| \le m$, risulta $\|f + g\| < 2m$.

a) Si verifichi che $C[a, b]$ con la norma 2 è strettamente convesso, mentre con la norma ∞ non lo è.

b) Si dimostri che nello spazio vettoriale $\mathcal{G}$ strettamente convesso rispetto a una norma $\|\ .\ \|$, il problema 6.3 dell'approssimazione lineare ha un'unica soluzione.

Da questa proprietà segue l'unicità della funzione di migliore approssimazione quando si usi la norma 2.

(Traccia: a) per la norma 2, siano f, g tali che $\|f\|_2 = \|g\|_2 = 1$. Per una costante $\lambda \in \mathbf{R}$ risulta

$$\|f + \lambda g\|_2^2 = (f + \lambda g, f + \lambda g) = \|f\|_2^2 + \lambda^2 \|g\|_2^2 + 2\lambda(f, g) = 1 + \lambda^2 + 2\lambda(f, g).$$

Poiché $\|f + \lambda g\|_2^2 \geq 0$ per ogni λ, ne segue che $(f,g)^2 - 1 \leq 0$, e quindi $|(f,g)| \leq 1$. Se fosse $(f,g) = 1$, allora sarebbe $(f-g, f-g) = 0$ e quindi $f = g$, caso che è stato escluso. Se fosse $(f,g) = -1$, allora sarebbe $(f+g, f+g) = 0$ e quindi $f = -g$. In ogni altro caso è $-1 < (f,g) < 1$ e

$$\|f+g\|_2^2 = \|f\|_2^2 + \|g\|_2^2 + 2(f,g) < 4.$$

Per la norma ∞, siano $[a,b] = [0,1]$, $f(x) = 1$, $g(x) = 1 - x$, risulta $f + g = 2 - x$, $\|f\|_\infty = 1$, $\|g\|_\infty = 1$, $\|f+g\|_\infty = 2$.

b) Se vi fossero due diverse funzioni di migliore approssimazione $g_n(x)$ e $\overline{g}_n(x)$, risulterebbe

$$\|f - g_n\| = \|f - \overline{g}_n\| = \delta_n,$$

e per la stretta convessità

$$\|f - g_n + f - \overline{g}_n\| < 2\delta_n,$$

quindi la funzione $\dfrac{1}{2}\,(g_n(x) + \overline{g}_n(x))$ sarebbe tale che

$$\|f - \frac{1}{2}\,(g_n + \overline{g}_n)\| < \delta_n,$$

che è assurdo.)

6.5 Si determinino i polinomi di grado 0 di migliore approssimazione della funzione $f(x) = \mathrm{e}^x$ sull'intervallo $[0,1]$ per la norma 2 e per la norma ∞.

(Risposta: per la norma 2, $g_0(x) = \alpha_0^* = \mathrm{e} - 1$; per la norma ∞, $g_0(x) = \alpha_0^* = \dfrac{1}{2}\,(\mathrm{e}+1)$.)

6.6 Sia $f(x) \in C[a,b]$, con $a \neq b$. Si verifichi che il sistema

$$A\boldsymbol{\alpha} = \mathbf{b},$$

dove

$$a_{ik} = \int_a^b x^{i+k}\,dx, \quad b_i = \int_a^b x^i f(x)\,dx, \quad i,k = 0,\ldots,n,$$

ha un'unica soluzione $\boldsymbol{\alpha}^*$.

(Traccia: si verifichi che se $f(x)$ è la funzione identicamente nulla, allora il sistema ha la sola soluzione $\boldsymbol{\alpha}^* = \mathbf{0}$. Infatti dalla relazione

$$\sum_{k=0}^n a_{ik}\alpha_k = \sum_{k=0}^n \alpha_k \int_a^b x^{i+k}\,dx = 0, \quad i = 0,\ldots,n,$$

si ottiene

$$\sum_{i=0}^{n} \alpha_i \sum_{k=0}^{n} \alpha_k \int_a^b x^{i+k}\, dx = \int_a^b \sum_{i=0}^{n} \sum_{k=0}^{n} \alpha_i \alpha_k x^{i+k}\, dx = \int_a^b p^2(x)\, dx = 0,$$

in cui si è posto $p(x) = \sum_{i=0}^{n} \alpha_i x^i$. Ne segue che il polinomio $p(x)$ è identicamente nullo e quindi sono nulli i suoi coefficienti.)

6.7 Sia $f(x) \in C[a,b]$.

a) Si verifichi che la funzione

$$f \to \int_a^b |f(x)|\, dx$$

è una norma. Tale norma, detta *norma 1*, viene indicata con $\|f\|_1$ e misura lo *scostamento medio* della funzione $f(x)$ rispetto allo zero. La soluzione del problema 6.3, formulato con la norma 1, viene detta *approssimazione di minima deviazione*.

b) Si determini il polinomio di primo grado di minima deviazione per la funzione $f(x) = x^3$ sull'intervallo $[-1, 1]$.

c) Si verifichi che lo spazio $C[a,b]$ con la norma 1 non è strettamente convesso (si veda l'esercizio 6.4). Ne segue che l'unicità della soluzione del problema 6.3 con la norma 1 deve essere dimostrata per altra via.

d) Siano $f(x) \in C[a,b]$ e $p(x)$ un polinomio di grado n tale che $f(x) - p(x)$ abbia un numero finito di zeri in (a,b). Si dimostri che $p(x)$ è il polinomio di minima deviazione per $f(x)$ su $[a,b]$ se e solo se per ogni polinomio $q(x)$ di grado al più n è

$$\int_a^b q(x)\, \mathrm{sgn}[f(x) - p(x)]\, dx = 0, \tag{102}$$

dove con $\mathrm{sgn}[w]$ si indica la funzione che assume i valori $+1$, 0, -1 a seconda che $w > 0$, $w = 0$, $w < 0$.

e) Si verifichi che l'approssimazione polinomiale di minima deviazione è unica.

(Traccia: b) si determinino i coefficienti α_1^* e α_0^* tali che

$$\|x^3 - \alpha_1^* x - \alpha_0^*\|_1 = \min_{(\alpha_0, \alpha_1)} \|x^3 - \alpha_1 x - \alpha_0\|_1,$$

cioè si determini il punto di minimo di

$$I_1 = \int_{-1}^{1} |x^3 - \alpha_1 x - \alpha_0| \, dx.$$

Come nell'esempio 6.10 si suppone che le soluzioni x_1, x_2, x_3 dell'equazione

$$u(x) = x^3 - \alpha_1 x - \alpha_0 = 0$$

siano distinte e interne a $[-1, 1]$. Tenendo conto del segno della funzione $u(x)$, si ha

$$\begin{aligned} I_1 &= \int_{-1}^{x_1} -u(x) \, dx + \int_{x_1}^{x_2} u(x) \, dx + \int_{x_2}^{x_3} -u(x) \, dx + \int_{x_3}^{1} u(x) \, dx \\ &= \alpha_1[x_1^2 - x_2^2 + x_3^2 - 1] + 2\alpha_0[x_1 - x_2 + x_3] - \frac{1}{2}\,[x_1^4 - x_2^4 + x_3^4 - 1]. \end{aligned}$$

Si verifichi che derivando I_1 rispetto a α_1 e α_0 risulta

$$\frac{\partial I_1}{\partial \alpha_1} = x_1^2 - x_2^2 + x_3^2 - 1, \quad \frac{1}{2}\,\frac{\partial I_1}{\partial \alpha_0} = x_1 - x_2 + x_3;$$

α_0^* e α_1^* sono i punti stazionari di I_1 e quindi

$$\begin{cases} x_2 = x_1 + x_3 \\ x_1 x_3 = -\dfrac{1}{2}\,. \end{cases}$$

Tenendo conto che $u(x_i) = 0$ per $i = 1, 2, 3$, ne segue che $\alpha_0^* = 0$, $\alpha_1^* = \dfrac{1}{2}$, e quindi il polinomio di minima deviazione per la funzione $f(x) = x^3$ su $[-1, 1]$ è

$$g_1(x) = \frac{1}{2}\, x.$$

c) Siano $[a, b] = [0, 1]$, $f(x) = 1$, $g(x) = 2 - 2x$, risulta $f + g = 3 - 2x$, $\|f\|_1 = 1$, $\|g\|_1 = 1$, $\|f + g\|_1 = 2$.

d) Si supponga che $p(x)$ sia il polinomio di minima deviazione per $f(x)$ su $[a, b]$ e si supponga per assurdo che non valga la (102) e che esista un polinomio $q(x)$ di grado al più n tale che

$$\int_a^b q(x) \, \mathrm{sgn}[g(x)] \, dx > 0, \quad \text{dove} \quad g(x) = f(x) - p(x)$$

(la dimostrazione sarebbe analoga nel caso che l'integrale fosse negativo). Siano x_i, $i = 1, \ldots, k$ gli zeri di $g(x)$ in (a, b), con

$$a = x_0 < x_1 < \ldots < x_k < x_{k+1} = b,$$

e fissato un $\epsilon > 0$, si considerino gli insiemi

$$A = \bigcup_{i=0}^{k} [x_i + \epsilon, x_{i+1} - \epsilon], \quad B = [a, b] - A,$$

e si scelga ϵ così piccolo che

$$\int_{x \in A} q(x) \operatorname{sgn}[g(x)]\, dx > \int_{x \in B} |q(x)|\, dx. \tag{103}$$

Poiché A è chiuso e non contiene zeri di $g(x)$, è

$$m = \min_{x \in A} |g(x)| > 0,$$

ed esiste un c tale che

$$0 \leq c\, |q(x)| < m, \quad \text{per ogni} \quad x \in A,$$

e quindi

$$\operatorname{sgn}[g(x) - cq(x)] = \operatorname{sgn}[g(x)], \quad \text{per} \quad x \in A,$$

e

$$\begin{aligned}
\int_{x \in A} |g(x) - cq(x)|\, dx &= \int_{x \in A} \big[g(x) - cq(x)\big] \operatorname{sgn}[g(x)]\, dx \\
&= \int_{x \in A} |g(x)|\, dx - c \int_{x \in A} q(x) \operatorname{sgn}[g(x)]\, dx \\
&= \int_a^b |g(x)|\, dx - \int_{x \in B} |g(x)|\, dx - c \int_{x \in A} q(x) \operatorname{sgn}[g(x)]\, dx.
\end{aligned}$$

Risulta quindi

$$\begin{aligned}
\int_a^b |g(x) - cq(x)|\, dx &= \int_{x \in A} |g(x) - cq(x)|\, dx + \int_{x \in B} |g(x) - cq(x)|\, dx \\
&\leq \int_a^b |g(x)|\, dx - c \int_{x \in A} q(x) \operatorname{sgn}[g(x)]\, dx + c \int_{x \in B} |q(x)|\, dx \\
&< \int_a^b |g(x)|\, dx,
\end{aligned}$$

per la (103), e questo è assurdo perché risulterebbe

$$\int_a^b |f(x) - [p(x) + cq(x)]|\ dx < \int_a^b |f(x) - p(x)|\ dx,$$

dove $p(x) + cq(x)$ è un polinomio di grado al più n, cioè $p(x)$ non sarebbe polinomio di minima deviazione.
Viceversa, si supponga che $p(x)$ soddisfi la condizione (102). Per ogni polinomio $p_1(x)$ di grado al più n si ha

$$\begin{aligned}\int_a^b |f(x) - p_1(x)|\ dx &\geq \int_a^b \big[f(x) - p_1(x)\big]\ \mathrm{sgn}\big[g(x)\big]\ dx \\ &= \int_a^b g(x)\ \mathrm{sgn}\big[g(x)\big]\ dx + \int_a^b \big[p(x) - p_1(x)\big]\ \mathrm{sgn}\big[g(x)\big]\ dx.\end{aligned}$$

Poiché $p(x) - p_1(x)$ è un polinomio di grado al più n, per la (102) il secondo integrale è nullo e ne segue che

$$\int_a^b |f(x) - p_1(x)|\ dx \geq \int_a^b |f(x) - p(x)|\ dx.$$

Quindi $p(x)$ è un polinomio di minima deviazione per $f(x)$ su $[a, b]$.
e) Si supponga per assurdo che i due polinomi $p_1(x)$ e $p_2(x)$ di grado al più n siano soluzioni del problema 6.3. Per la convessità dell'insieme delle soluzioni anche $p(x) = \dfrac{1}{2}\,[p_1(x) + p_2(x)]$ è soluzione e si ha

$$\int_a^b |f(x) - p(x)|\ dx = \int_a^b |f(x) - p_1(x)|\ dx = \int_a^b |f(x) - p_2(x)|\ dx$$

da cui

$$\int_a^b |f(x) - p(x)|\ dx - \frac{1}{2}\int_a^b |f(x) - p_1(x)|\ dx - \frac{1}{2}\int_a^b |f(x) - p_2(x)|\ dx = 0.$$

D'altra parte è

$$|f(x) - p(x)| - \frac{1}{2}\,|f(x) - p_1(x)| - \frac{1}{2}\,|f(x) - p_2(x)| \leq 0,$$

e quindi

$$|f(x) - p(x)| - \frac{1}{2}\,|f(x) - p_1(x)| - \frac{1}{2}\,|f(x) - p_2(x)| = 0.$$

Si distinguono due casi:
(1) la funzione $f(x)-p(x)$ ha almeno $n+1$ zeri in (a,b), allora $p_1(x)$ e $p_2(x)$ assumono gli stessi valori in almeno $n+1$ punti e quindi coincidono;
(2) la funzione $f(x)-p(x)$ ha al più n zeri in (a,b), allora per quanto visto al punto d) è

$$\int_a^b x^j \operatorname{sgn}[f(x)-p(x)]\,dx = 0, \quad \text{per} \quad j=0,\dots,n.$$

Indicati con $a = x_0 < x_1 < \ldots < x_n < x_{n+1} = b$ dei punti di $[a,b]$ fra i quali vi sono tutti gli zeri di $f(x)-p(x)$, si ha per $j=0,\dots,n$

$$\begin{aligned} 0 &= \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} x^j \operatorname{sgn}[f(x)-p(x)]\,dx \\ &= \sum_{i=0}^{n} \sigma_i \int_{x_i}^{x_{i+1}} x^j\,dx, \quad \sigma_i = \operatorname{sgn}[f(x)-p(x)] \quad \text{per} \quad x \in [x_i, x_{i+1}]. \end{aligned}$$

Ne segue che la matrice i cui elementi sono

$$\int_{x_i}^{x_{i+1}} x^j\,dx, \quad i,j=0,\dots,n,$$

è singolare e quindi esiste una combinazione lineare con coefficienti non nulli

$$\sum_{j=0}^{n} c_j \int_{x_i}^{x_{i+1}} x^j\,dx = 0, \quad \text{per ogni} \quad i=0,\dots,n,$$

cioè il polinomio

$$q(x) = \sum_{j=0}^{n} c_j x^j,$$

di grado al più n, è tale che

$$\int_{x_i}^{x_{i+1}} q(x)\,dx = 0$$

per ogni $i=0,\dots,n$, ciò che è assurdo perché implicherebbe l'esistenza di uno zero di $q(x)$ in $n+1$ intervalli distinti.)

6.8 Sia $\{q_i(x)\}_{i\in\mathbf{N}}$ una qualsiasi successione di polinomi, in cui $q_i(x)$ è di grado i. Si verifichi che i polinomi della successione

a) sono linearmente indipendenti,

b) formano una base dello spazio dei polinomi.

(Traccia: a) se fosse $\sum_{i=0}^{n} \alpha_i q_i(x) = 0$, con $\alpha_n \neq 0$, allora la combinazione lineare sarebbe un polinomio di grado n e non potrebbe essere uguale a 0; b) si verifichi che ogni polinomio $p(x)$ di grado n può essere espresso come combinazione lineare dei polinomi $q_i(x)$ per $i = 0, \ldots, n$. Si indichi con a_i il primo coefficiente di $q_i(x)$ e con α il primo coefficiente di $p(x)$ e si proceda per induzione. Se $n = 0$, è $p(x) = \alpha$, e quindi

$$p(x) = \frac{\alpha}{a_0}\, q_0(x).$$

Se $n > 0$, il polinomio

$$r(x) = p(x) - \frac{\alpha}{a_n}\, q_n(x)$$

ha al più grado $n - 1$ e per l'ipotesi induttiva è

$$r(x) = \sum_{j=0}^{n-1} \gamma_j q_j(x).$$

Allora

$$p(x) = \frac{\alpha}{a_n}\, q_n(x) + \sum_{j=0}^{n-1} \gamma_j q_j(x).\ )$$

6.9 Assegnato un prodotto scalare (p, q) sui polinomi e i due polinomi

$$p_0(x) = 1, \quad \text{e} \quad p_1(x) = x - \gamma_1, \quad \gamma_1 = \frac{(x, 1)}{(1, 1)}\,,$$

si considerino i polinomi costruiti con il seguente algoritmo

$$\begin{aligned} &p_i(x) = (x - \gamma_i)p_{i-1}(x) - \delta_i p_{i-2}(x), \\ &\gamma_i = \frac{(xp_{i-1}, p_{i-1})}{(p_{i-1}, p_{i-1})}, \quad \delta_i = \frac{(xp_{i-1}, p_{i-2})}{(p_{i-2}, p_{i-2})}\,, \quad i = 2, 3, \ldots \end{aligned}$$

Si verifichi che

a) $p_i(x)$ è ben definito e ha grado i,

b) i polinomi $p_i(x)$ sono, a meno di un fattore, i polinomi ortogonali studiati nel paragrafo 3, a seconda del prodotto scalare scelto.

(Traccia: a) si noti che i fattori ai denominatori di γ_i e δ_i non sono nulli; b) si verifichi che $(p_i, p_j) = 0$ per $j = 0, \ldots, i-1$, procedendo per induzione su i.)

6.10 Si dimostri che l'insieme dei polinomi $\{x^i\}$, $i = 0, 1, \ldots$ non può essere un insieme ortogonale secondo il prodotto scalare (6) qualunque sia la funzione peso.

(Traccia: si verifichi che per tale insieme non può valere la relazione ricorrente a tre termini (9).)

6.11 Sia $\{p_i(x)\}_{i \in \mathbf{N}}$ una successione di polinomi ortogonali. Se una funzione $f(x) \in C[a, b]$ è ortogonale a $p_0(x), \ldots, p_{n-1}(x)$, allora esistono almeno n punti distinti di (a, b) in cui la $f(x)$ cambia segno.

(Traccia: si proceda come nella dimostrazione del teorema 6.12 e si ottenga la relazione

$$\int_a^b \omega(x) f(x) q(x)\, dx \neq 0,$$

in cui $q(x)$ è un polinomio di grado $k < n$. Si tenga conto del fatto che $q(x)$ è uguale a una combinazione di polinomi di grado minore di n, a cui $f(x)$ è ortogonale.)

6.12 Sia $\{p_i(x)\}_{i \in \mathbf{N}}$ un insieme di polinomi ortogonali, $f(x) \in C[a, b]$ e sia

$$g_n(x) = \sum_{j=0}^{n} \alpha_j^* p_j(x)$$

la sua approssimazione ai minimi quadrati. Posto $r_n(x) = f(x) - g_n(x)$, si verifichi che

a) $(r_n, p_i) = 0$ per $i = 0, \ldots, n$,

b) $r_n(x)$ si annulla in almeno $n + 1$ punti di (a, b).

(Traccia: a) $$\begin{aligned}(r_n, p_i) = (f - g_n, p_i) &= (f, p_i) - \sum_{j=0}^{n} \alpha_j^* (p_i, p_j) \\ &= (f, p_i) - \alpha_i^* h_i = (f, p_i) - (f, p_i) = 0;\end{aligned}$$

b) segue dall'esercizio 6.11.)

6.13 Sia $\{p_i(x)\}_{i \in \mathbf{N}}$ un insieme di polinomi ortogonali sull'intervallo $[-a, a]$, $a > 0$, rispetto al peso $\omega(x)$, e sia $\omega(-x) = \omega(x)$. Si verifichi che

$$p_i(-x) = (-1)^i p_i(x).$$

(Traccia: riferendosi all'esercizio 6.9, si dimostri che $\gamma_i = 0$, per $i = 1, 2, \ldots$, verificando per induzione che se $p_{i-1}(x)$ è pari oppure dispari, allora la funzione $x\omega(x)p_{i-1}^2(x)$ è dispari.)

6.14 Sia $\{p_i(x)\}_{i\in\mathbf{N}}$ un insieme di polinomi ortogonali sull'intervallo $[a, b]$.

a) Si verifichi che per ogni n è

$$\frac{a_n}{a_{n+1}}\left[p'_{n+1}(x)p_n(x) - p'_n(x)p_{n+1}(x)\right] > 0, \quad \text{per} \quad x \in [a, b],$$

e quindi $p_n(x)$ e $p_{n+1}(x)$ non possono avere zeri comuni.

b) Indicati con $x_1, x_2, \ldots, x_n$ gli zeri di $p_n(x)$, ordinati in modo che $x_1 < x_2 < \ldots < x_n$, e posto $x_0 = a$ e $x_{n+1} = b$, si verifichi che ogni intervallo (x_j, x_{j+1}), per $j = 0, \ldots, n$ contiene esattamente uno zero di $p_{n+1}(x)$ (*proprietà di separazione* degli zeri dei polinomi ortogonali).

(Traccia: a) si derivi rispetto ad x la formula di Christoffel-Darboux, si ponga $\xi = x$ e si tenga conto che $h_i > 0$ per $i = 0, \ldots, n$.
b) per $j = 1, \ldots, n-1$ risulta $p'_n(x_j)p'_n(x_{j+1}) < 0$ e dalla a) segue che $p_{n+1}(x_j)p_{n+1}(x_{j+1}) < 0$, cioè vi è almeno uno zero di $p_{n+1}(x)$ in (x_j, x_{j+1}). Inoltre è $a_n p'_n(x_n) > 0$ e dalla a) segue che $a_{n+1}p_{n+1}(x_n) < 0$ e quindi vi è almeno uno zero di $p_{n+1}(x)$ a destra di x_n. In modo analogo si vede che vi è almeno uno zero di $p_{n+1}(x)$ a sinistra di x_1. Poiché $p_{n+1}(x)$ ha esattamente $n + 1$ zeri, fra due zeri consecutivi di $p_n(x)$ vi deve essere un solo zero di $p_{n+1}(x)$.)

6.15 Si verifichi che per i polinomi di Legendre $P_i(x)$ valgono le seguenti relazioni

a) $$P'_{i+1}(x) - xP'_i(x) = (i+1)P_i(x);$$

b) $$xP'_i(x) - P'_{i-1}(x) = iP_i(x);$$

c) $$P'_{i+1}(x) - P'_{i-1}(x) = (2i+1)P_i(x);$$

d) $$(x^2-1)P'_i(x) = ixP_i(x) - iP_{i-1}(x);$$

e) $$P_i(0) = \begin{cases} 0 & \text{se } i \text{ è dispari,} \\ (-1)^{i/2}\,\dfrac{1\cdot 3\cdot 5\cdots(i-1)}{2\cdot 4\cdot 6\cdots i} & \text{se } i \geq 2 \text{ è pari;} \end{cases}$$

f) $$P_i(1) = 1,\ P_i(-1) = (-1)^i,$$

$$P'_i(1) = \frac{1}{2}\,i(i+1),\ P'_i(-1) = \frac{(-1)^{i+1}}{2}\,i(i+1);$$

g) $$P_i^{(k)}(x) = xP_{i-1}^{(k)}(x) + (i+k-1)P_{i-1}^{(k-1)}(x) \quad \text{per } k \geq 1,$$

$$P_i^{(k)}(1) = \frac{(i+k)!}{(i-k)!\,2^k k!}\,,\ P_i^{(k)}(-1) = (-1)^{i+k}P_i^{(k)}(1);$$

h) $$\int_{-1}^{x} P_i(t)\ dt = \frac{P_{i+1}(x) - P_{i-1}(x)}{2i+1}\,;$$

i) $$\int_{-1}^{1} x^k P_i(x)\ dx = \begin{cases} 0 & \text{per } k = 0, \ldots, i-1, \\ \dfrac{2^{i+1}(i!)^2}{(2i+1)!} & \text{per } k = i; \end{cases}$$

j) $$\int_{-1}^{1} \frac{P_i(x)}{x - x_k}\ dx = \frac{-2}{(i+1)P_{i+1}(x_k)} = \frac{2}{iP_{i-1}(x_k)}\,,$$
dove x_k è uno zero di $P_i(x)$;

k) $$\int_{-1}^{1} \frac{P_i(x) + P_{i+1}(x)}{x+1}\ dx = \frac{(-1)^i 2}{(i+1)}\,;$$

l) $$\int_{-1}^{1} (x^2-1)P_i'(x)P_k'(x)\ dx = \begin{cases} 0 & \text{per } i \neq k, \\ -\dfrac{2(i+i^2)}{2i+1} & \text{per } i = k; \end{cases}$$

m) $P_i(x)$ è soluzione dell'equazione differenziale

$$(1-x^2)y'' - 2xy' + i(i+1)y = 0;$$

n) $$\|P_i\|_\infty = P_i(1) = 1;$$

o) $$\|P_i'\|_\infty = P_i'(1) = \frac{1}{2}\,i(i+1).$$

(Traccia: a) dalla formula di Rodrigues si ha

$$P_{i+1}'(x) = \frac{1}{2^{i+1}(i+1)!}\ \frac{d^{i+2}}{dx^{i+2}}(x^2-1)^{i+1} = \frac{1}{2^i i!}\ \frac{d^{i+1}}{dx^{i+1}}\ x(x^2-1)^i.$$

Si applichi poi la formula di Leibniz; b) si sfruttino la a) e la relazione a tre termini; c) e d) si ottengono da a) e b); e) si proceda per induzione, sfruttando la relazione a tre termini; f) per induzione sfruttando la relazione a tre termini e la a); g) per induzione su k, sfruttando la a) e la f) per il passo iniziale, con $k = 1$; h) si ottiene da c); i) per $k = 0, \ldots, i-1$ discende dal teorema 6.11, per $k = i$ poiché

$$P_i(x) = a_i x^i + \sum_{j=0}^{i-1} \alpha_j x^j,$$

risulta

$$\int_{-1}^{1} x^i P_i(x)\ dx = \frac{1}{a_i}\int_{-1}^{1} P_i^2(x)\ dx - \sum_{j=0}^{i-1} \frac{\alpha_j}{a_i}\ \int_{-1}^{1} x^j P_i(x)\ dx = \frac{h_i}{a_i}\,;$$

j) si veda la dimostrazione del teorema 7.17; k) si utilizzi la d) per verificare che

$$\frac{P_i(x) + P_{i+1}(x)}{x+1} = P_{i+1}(x) - \frac{x-1}{i+1} P'_{i+1}(x),$$

e si integri

$$\int_{-1}^{1} (x-1) P'_{i+1}(x)\, dx$$

per parti; l) si applichi la d), si integri per parti e si applichi la b); m) si derivi la d) e si sfrutti la b); n) si consideri il polinomio

$$q(x) = P_i^2(x) + \frac{1-x^2}{i(i+1)} [P'_i(x)]^2 > 0 \quad \text{per} \quad |x| \le 1.$$

Nei punti $x = \pm 1$ e nei punti x tali che $P'_i(x) = 0$ risulta $q(x) = P_i^2(x)$. Derivando si ha

$$q'(x) = \frac{2P'_i(x)}{i(i+1)} \left[i(i+1)P_i(x) - xP'_i(x) + (1-x^2)P''_i(x) \right].$$

Tenendo conto che $P_i(x)$ soddisfa l'equazione differenziale del punto m), risulta

$$q'(x) = \frac{2x[P'_i(x)]^2}{i(i+1)},$$

per cui $q'(x) \le 0$ per $x < 0$ e $q'(x) \ge 0$ per $x > 0$, cioè il polinomio $q(x)$ è non crescente per $x < 0$ e non decrescente per $x > 0$. Ne segue che il massimo valore di $q(x)$ per $|x| \le 1$ viene assunto in ± 1 e vale $q(\pm 1) = 1$, e quindi nei punti x in cui $P'_i(x) = 0$ risulta $P_i^2(x) = q(x) \le q(\pm 1) = 1$. o) Si proceda per induzione sfruttando la a), la f) e la n).)

6.16 Si esprimano le potenze x^k, $k = 0, 1, \ldots$, come combinazione dei polinomi di Legendre.

(Traccia: fissato n e considerati i vettori

$$\mathbf{p} = \left[P_0(x), P_1(x), \ldots, P_n(x) \right]^T,$$
$$\mathbf{x} = [1, x, x^2, \ldots, x^n]^T,$$

dall'esempio 6.17 risulta che $\mathbf{p} = M\,\mathbf{x}$, dove

$$M = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ -\frac{1}{2} & 0 & \frac{3}{2} & & & \\ 0 & -\frac{3}{2} & 0 & \frac{5}{2} & & \\ \frac{3}{8} & 0 & -\frac{15}{4} & 0 & \frac{35}{8} & \\ \vdots & & & & & \ddots \end{bmatrix} \in \mathbf{R}^{(n+1)\times(n+1)}.$$

La matrice M è non singolare e quindi $\mathbf{x} = M^{-1}\mathbf{p}$, dove

$$M^{-1} = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ \frac{1}{3} & 0 & \frac{2}{3} & & & \\ 0 & \frac{3}{5} & 0 & \frac{2}{5} & & \\ \frac{1}{5} & 0 & \frac{4}{7} & 0 & \frac{8}{35} & \\ \vdots & & & & & \ddots \end{bmatrix},$$

da cui si ha

$$\begin{aligned} x^0 &= P_0(x) \\ x^1 &= P_1(x) \\ x^2 &= \frac{1}{3}\big[2P_2(x) + P_0(x)\big] \\ x^3 &= \frac{1}{5}\big[2P_3(x) + 3P_1(x)\big] \\ x^4 &= \frac{1}{35}\big[8P_4(x) + 20P_2(x) + 7P_0(x)\big] \\ &\dots \end{aligned}$$

6.17 Si verifichi che per i polinomi di Chebyshev di 1^a specie $T_i(x)$ e di 2^a specie $U_i(x)$ valgono le seguenti relazioni

a) $T_{k+i}(x) + T_{k-i}(x) = 2T_k(x)T_i(x), \quad \text{per } k \geq i;$

b) $T_{2i}(x) = 2T_i^2(x) - 1$;

c) $\dfrac{1}{i}\,T_i'(x) = 2T_{i-1}(x) + \dfrac{1}{i-2}\,T_{i-2}'(x), \quad \text{per } i \geq 3$;

d) $xT_i'(x) = iT_i(x) + \dfrac{i}{i-1}\,T_{i-1}'(x), \quad \text{per } i \geq 2$;

e) $U_{k+i}(x) + U_{k-i}(x) = 2U_k(x)T_i(x), \quad \text{per } k \geq i$;

f) $U_{2i}(x) = 2U_i(x)T_i(x) - 1$;

g) $T_i(x) = U_i(x) - xU_{i-1}(x)$;

h) $(1 - x^2)U_{i-1}(x) = xT_i(x) - T_{i+1}(x)$;

i) $U_i'(x) = (i+1)U_{i-1}(x) + xU_{i-1}'(x)$;

j) $U_i'(x) = 2iU_{i-1}(x) + U_{i-2}'(x)$;

k) $T_i(0) = U_i(0) = \begin{cases} 0 & \text{se } i \text{ è dispari,} \\ (-1)^{i/2} & \text{se } i \text{ è pari;} \end{cases}$

l) $T_i(1) = 1,\ T_i(-1) = (-1)^i,\ U_i(1) = i+1,\ U_i(-1) = (-1)^i(i+1)$;

m) $\displaystyle\int_{-1}^{x} T_0(t)\,dt = T_1(x) + 1, \quad \int_{-1}^{x} T_1(t)\,dt = \frac{1}{4}\big[T_2(x) - 1\big],$

$$\int_{-1}^{x} T_i(t)\,dt = \frac{1}{2}\left[\frac{T_{i+1}(x)}{i+1} - \frac{T_{i-1}(x)}{i-1}\right] - \frac{(-1)^i}{i^2-1} \quad \text{per} \quad i > 1,$$

e quindi $\displaystyle\int_{-1}^{1} T_i(x)\,dx = \begin{cases} 0 & \text{se } i \text{ è dispari,} \\ -\dfrac{2}{i^2-1} & \text{se } i \text{ è pari;} \end{cases}$

n) $\displaystyle\int_{-1}^{x} U_0(t)\,dt = \frac{1}{2}\,U_1(x) + 1,$

$$\int_{-1}^{x} U_i(t)\,dt = \frac{1}{i+1}\left[\frac{U_{i+1}(x) - U_{i-1}(x)}{2} + (-1)^i\right] \quad \text{per} \quad i \geq 1,$$

e quindi $\displaystyle\int_{-1}^{1} U_i(x)\,dx = \begin{cases} 0 & \text{se } i \text{ è dispari,} \\ \dfrac{2}{i+1} & \text{se } i \text{ è pari;} \end{cases}$

o) $T_i(x)$ è soluzione dell'equazione differenziale

$$(1 - x^2)y'' - xy' + i^2 y = 0;$$

p) $U_i(x)$ è soluzione dell'equazione differenziale

$$(1 - x^2)y'' - 3xy' + i(i+2)y = 0;$$

q) $\|U_i\|_\infty = U_i(1) = i+1$;

r) $\|T_i'\|_\infty = T_i'(1) = i^2$;

$$= \frac{1}{2}\left[(2x)^i - i(2x)^{i-2} + \frac{i}{2}\binom{i-3}{1}(2x)^{i-4} - \dots\right];$$

s) $$2\sum_{i=0}^{n} T_{2i}(x) = 1 + \frac{T_{2n+1}'(x)}{2n+1} = 1 + U_{2n}(x),$$

$$2\sum_{i=0}^{n-1} T_{2i+1}(x) = \frac{T_{2n}'(x)}{2n} = U_{2n-1}(x);$$

t) $$2x\sum_{i=0}^{n}(-1)^i T_{2i}(x) = x + (-1)^n T_{2n+1}(x),$$

$$2x\sum_{i=0}^{n-1}(-1)^i T_{2i+1}(x) = 1 - (-1)^n T_{2n}(x);$$

u) $$2(1-x^2)\sum_{i=0}^{n} U_{2i}(x) = 1 - T_{2n+2}(x),$$

$$2(1-x^2)\sum_{i=0}^{n-1} U_{2i+1}(x) = x - T_{2n+1}(x);$$

v) $$2x\sum_{i=0}^{n}(-1)^i U_{2i}(x) = (-1)^n U_{2n+1}(x),$$

$$2x\sum_{i=0}^{n-1}(-1)^i U_{2i+1}(x) = 1 - (-1)^n U_{2n}(x);$$

w) $$4\sum_{i=0}^{n} T_i^2(x) = 2n + 3 + U_{2n}(x).$$

(Traccia: a) − l) si sfruttino la (26) e la (28), la relazione a tre termini e varie formule di trigonometria; m) dalla c) in quanto $T_i(-1) = (-1)^i$; n) dalla j); o) posto $x = \cos\theta$ e $y = \cos i\theta$, risulta

$$y' = \frac{i\sin i\theta}{\sin\theta} \quad \text{e} \quad y'' = \frac{i\sin i\theta\cos\theta - i^2\cos i\theta\sin\theta}{\sin^3\theta};$$

p) si derivi l'equazione in o) e si sfrutti la (29); q) si applichi lo stesso ragionamento del punto n) dell'esercizio 6.15, con

$$q(x) = U_i^2(x) + \frac{1-x^2}{i(i+2)}\,[U_i'(x)]^2,$$

per dimostrare che $U_i^2(x)$ assume il massimo per $|x| = 1$, e si usi la l); r) si sfruttino la (29), la l) e la q); s) si sfrutti la c); t) e v) si sfrutti la relazione a tre termini; u) si sfruttino le h), t) e g); w) si sfruttino la b) e la s).)

6.18 Si verifichi che per i polinomi di Laguerre $L_i(x)$ valgono le seguenti relazioni

a) $$L_i'(x) = L_{i-1}'(x) - L_{i-1}(x);$$

b) $$xL_i'(x) = i\,[L_i(x) - L_{i-1}(x)];$$

c) $$L_i(0) = 1, \quad L_i'(0) = -i;$$

d) $$\int_0^x L_i(t)\,dt = L_i(x) - L_{i+1}(x);$$

e) $$\int_0^\infty x^k \mathrm{e}^{-x} L_i(x)\,dx = \begin{cases} 0 & \text{se } k < i, \\ (-1)^i\, i! & \text{se } k = i; \end{cases}$$

f) $L_i(x)$ è soluzione dell'equazione differenziale

$$xy'' + (1-x)y' + iy = 0.$$

(Traccia: a) si sfrutti la formula di Rodrigues; b) si derivi la relazione a tre termini e si sfrutti la a); c) dall'espressione esplicita; d) da a) e c); e) se $k < i$ deriva dal teorema 6.11, se $k = i$ si può esprimere

$$x^i = (-1)^i\, i!\, L_i(x) + \sum_{j=0}^{i-1} \gamma_j L_j(x),$$

e quindi

$$\int_0^\infty x^i \mathrm{e}^{-x} L_i(x)\,dx = (-1)^i i!\,(L_i, L_i);$$

f) si derivi la b) e si sfrutti la a).)

6.19 Si verifichi che per i polinomi di Hermite $H_i(x)$ valgono le seguenti relazioni

a) $$H_i'(x) = 2iH_{i-1}(x);$$

b) $$H_i(0) = \begin{cases} 1 & \text{se } i = 0, \\ 0 & \text{se } i \text{ è dispari}, \\ (-1)^{i/2} 2^{i/2} 1 \cdot 3 \cdot 5 \cdots (i-1) & \text{se } i \geq 2 \text{ è pari}; \end{cases}$$

c) $$\int_0^x H_i(t)\ dt = \frac{H_{i+1}(x) - H_{i+1}(0)}{2(i+1)}\,;$$

d) $H_i(x)$ è soluzione dell'equazione differenziale

$$y'' - 2xy' + 2iy = 0.$$

(Traccia: a) si sfruttino la formula di Rodrigues e la relazione a tre termini; b) si proceda per induzione sfruttando la relazione a tre termini; c) si integri la a); d) si derivi la a) e si sfrutti la relazione a tre termini.)

6.20 Si individui un metodo iterativo per il calcolo degli zeri dei polinomi ortogonali di Legendre, Laguerre e Hermite, che sfrutti le relazioni caratteristiche dei singoli polinomi, senza determinarne esplicitamente i coefficienti. Si dica qual è il costo computazionale per iterazione.

(Traccia: si usi il metodo di Newton, approssimando gli zeri positivi (per i polinomi di Legendre e di Hermite gli zeri negativi sono simmetrici). Per un'iterazione del metodo

$$x_{k+1} = x_k - \frac{p_n(x_k)}{p_n'(x_k)}\,,$$

si calcoli $p_n(x_k)$ con la relazione a tre termini e $p_n'(x_k)$ con una delle relazioni degli esercizi 6.15, 6.18 e 6.19. In particolare per i polinomi di Legendre conviene scrivere

$$\begin{aligned} &P_0(x) = 1, \quad P_1(x) = x, \\ &P_i(x) = \frac{i-1}{i}\,\big[xP_{i-1}(x) - P_{i-2}(x)\big] + xP_{i-1}(x), \quad i = 2, \ldots, n, \end{aligned}$$

per cui il calcolo di $P_n(x_k)$ richiede $3(n-1)$ moltiplicazioni e $2(n-1)$ addizioni, e dalla 6.15 d) risulta subito

$$P_n'(x) = \frac{n\big[xP_n(x) - P_{n-1}(x)\big]}{x^2 - 1}\,.$$

Si proceda in modo analogo per i polinomi di Laguerre e di Hermite, sfruttando la 6.18 b) e la 6.19 a) per il calcolo delle derivate. Poiché per la 6.15 m) è

$$P_n''(x) = \frac{2xP_n'(x) - n(n+1)P_n(x)}{1 - x^2}\,,$$

si potrebbe usare anche, con lo stesso costo computazionale per passo, un metodo di ordine più elevato, che faccia intervenire esplicitamente il valore della derivata seconda, come il metodo di Halley (esercizio 3.33) o il metodo

di Laguerre (esercizio 3.64). Per approssimare il k-esimo zero positivo dei polinomi di Legendre, una buona scelta del punto iniziale è quella del k-esimo zero del polinomio di Chebyshev di 1^a specie dello stesso grado, cioè

$$x_0^{(k)} = \cos\ \frac{(2k-1)\pi}{2n}\,, \quad k = 1, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor.$$

Una scelta migliore, proposta in [34], è quella degli zeri con indice dispari del polinomio di Chebyshev di 1^a specie di grado $2n+1$, cioè

$$x_0^{(k)} = \cos\ \frac{(4k-1)\pi}{2(2n+1)}\,, \quad k = 1, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor. \ )$$

6.21 Sia $\{p_i(x)\}_{i\in\mathbf{N}}$ un sistema di polinomi ortogonali che soddisfano la relazione a tre termini (9).

a) Si verifichi che gli zeri di $p_n(x)$ sono gli autovalori della matrice

$$J = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{bmatrix}, \quad \text{dove} \quad \begin{cases} \alpha_i = \dfrac{b_{i-1}}{a_{i-1}} - \dfrac{b_i}{a_i}\,, \\ \beta_i = \dfrac{a_{i-1}}{a_i}\sqrt{\dfrac{h_i}{h_{i-1}}}\,. \end{cases}$$

b) Si costruisca la matrice J per i diversi polinomi ortogonali.

c) Indicato con

$$\mathbf{y}^{(j)} = [y_0^{(j)}, \ldots, y_{n-1}^{(j)}]^T$$

l'autovettore di J corrispondente a x_j, normalizzato in modo che $y_0^{(j)} = \dfrac{a_0}{\sqrt{h_0}}$, si verifichi che

$$\|\mathbf{y}^{(j)}\|_2^2 = \frac{a_{n-1}}{a_n h_{n-1}}\, p_n'(x_j) p_{n-1}(x_j).$$

(Traccia: dalla (9) si ha

$$\frac{1}{A_{i-1}}\, p_i(x) = \Big(x + \frac{B_{i-1}}{A_{i-1}}\Big) p_{i-1}(x) - \frac{C_{i-1}}{A_{i-1}}\, p_{i-2}(x).$$

Si verifichi che i polinomi normalizzati

$$q_i(x) = \frac{p_i(x)}{\sqrt{h_i}}$$

soddisfano la relazione

$$xq_{i-1}(x) = \beta_i q_i(x) + \alpha_i q_{i-1}(x) + \beta_{i-1} q_{i-2}(x), \quad i = 1, \ldots, n,$$

dove $q_{-1}(x) = 0$. Uno zero x_j di $p_n(x)$ è tale che $q_n(x_j) = 0$, e quindi

$$x_j q_{n-1}(x_j) = \alpha_n q_{n-1}(x_j) + \beta_{n-1} q_{n-2}(x_j).$$

Indicato con $\mathbf{q}(x)$ il vettore

$$\mathbf{q}(x) = [q_0(x), q_1(x), \ldots, q_{n-1}(x)]^T \quad \in \mathbf{R}^n,$$

si ha

$$x_j \mathbf{q}(x_j) = J\mathbf{q}(x_j),$$

e quindi x_j è autovalore di J e $\mathbf{q}(x_j)$ è l'autovettore corrispondente, normalizzato in modo che $y_0^{(j)} = \dfrac{a_0}{\sqrt{h_0}}$. Quindi è possibile calcolare gli zeri dei polinomi ortogonali per mezzo degli autovalori di una matrice tridiagonale simmetrica. Per far questo un ottimo metodo è il QR (si veda [7], pag. 353).
b) Per i polinomi di Legendre è

$$\alpha_i = 0, \quad \beta_i = \frac{i}{\sqrt{4i^2 - 1}},$$

per i polinomi di Chebyshev di 1^a specie è

$$\alpha_i = 0, \quad \beta_1 = \sqrt{\frac{1}{2}}, \quad \beta_i = \frac{1}{2} \quad \text{per} \quad i \geq 2,$$

per i polinomi di Chebyshev di 2^a specie è

$$\alpha_i = 0, \quad \beta_i = \frac{1}{2},$$

per i polinomi di Laguerre è

$$\alpha_i = 2i - 1, \quad \beta_i = -i,$$

per i polinomi di Hermite è

$$\alpha_i = 0, \quad \beta_i = \frac{\sqrt{2i}}{2}.$$

c) Dalla formula di Christoffel-Darboux si ha

$$\sum_{i=0}^{n-1} \frac{p_i(x)p_i(x_j)}{h_i} = \frac{a_{n-1}}{a_n h_{n-1}} \frac{p_n(x)p_{n-1}(x_j) - p_n(x_j)p_{n-1}(x)}{x - x_j}.$$

Poiché $p_n(x_j) = 0$, passando al limite per $x \to x_j$ risulta

$$\sum_{i=0}^{n-1} q_i^2(x_j) = \frac{a_{n-1}}{a_n h_{n-1}} p_n'(x_j)p_{n-1}(x_j). \ )$$

6.22 Data una funzione $f(x) \in C[-1,1]$, si dimostri che

a) $$\lim_{j\to\infty} \int_{-1}^{1} f(x)P_j(x)\ dx = 0,$$

b) $$\lim_{j\to\infty} \int_{-1}^{1} f(x)\ \frac{T_j(x)}{\sqrt{1-x^2}}\ dx = 0,$$

dove $P_j(x)$ e $T_j(x)$ sono i j-esimi polinomi ortogonali di Legendre e di Chebyshev di 1^a specie.

(Traccia: a) dall'uguaglianza di Parseval segue che

$$\lim_{j\to\infty} \frac{(f,P_j)^2}{(P_j,P_j)} = 0,$$

cioè

$$\lim_{j\to\infty} \frac{1}{h_j} \Big[\int_{-1}^{1} f(x)P_j(x)\ dx \Big]^2 = 0,$$

dove $h_j = \dfrac{2}{2j+1}$; b) in modo analogo.)

6.23 a) Si verifichi che per i polinomi di Chebyshev di 1^a specie valgono le relazioni

$$1 + 2\sum_{i=1}^{n} T_i(x)T_i(\xi) = \frac{T_{n+1}(x)T_n(\xi) - T_n(x)T_{n+1}(\xi)}{x-\xi}\ ,\ \text{per } x \neq \xi,\ n \geq 1;$$

$$\frac{2}{n+1}\left[\frac{1}{2} + \sum_{i=1}^{n} \cos i\theta_j\ \cos i\theta_k\right] = \delta_{j,k},\ \text{per } \theta_j = \frac{(2j+1)\pi}{2(n+1)},\ j,k = 0,\dots,n.$$

b) Sia $f(x) \in C^1[-1,1]$. Indicato con $g_n(x)$ il polinomio di approssimazione ai minimi quadrati di $f(x)$ con polinomi di Chebyshev di 1^a specie, si verifichi che per $n \to \infty$ la successione $g_n(x)$ converge a $f(x)$ su $[-1,1]$.

(Traccia: a) per la prima relazione si utilizzi la formula di Christoffel-Darboux; la seconda relazione per $j \neq k$, segue dalla precedente, per $j = k$ segue dall'esercizio 6.17 w). b) Fissato $\xi \in [-1,1]$ è

$$g_n(\xi) = \frac{1}{\pi} \int_{-1}^{1} \omega(x)f(x)\big[1 + 2\sum_{i=1}^{n} T_i(x)T_i(\xi)\big]\ dx,$$

e poiché

$$\int_{-1}^{1} \omega(x)T_i(x)T_i(\xi)\,dx = 0 \quad \text{per } i \geq 1, \quad \text{e} \quad \int_{-1}^{1} \omega(x)\,dx = \pi,$$

risulta

$$r_n(\xi) = f(\xi) - g_n(\xi) = \frac{1}{\pi}\int_{-1}^{1} \omega(x)\big[f(\xi] - f(x)\big]\,\big[1 + 2\sum_{i=1}^{n} T_i(x)T_i(\xi)\big]\,dx,$$

e per la a)

$$\begin{aligned} r_n(\xi) &= \frac{1}{\pi}\int_{-1}^{1} \omega(x)\,\frac{f(\xi)-f(x)}{\xi - x}\,\big[T_n(x)T_{n+1}(\xi) - T_{n+1}(x)T_n(\xi)\big]\,dx \\ &= \frac{1}{\pi}\big[T_{n+1}(\xi)(v, T_n) - T_n(\xi)(v, T_{n+1})\big], \end{aligned}$$

dove la funzione $v(x) = \dfrac{f(\xi)-f(x)}{\xi - x}$ è estendibile per continuità su $C[-1,1]$. Dall'esercizio 6.22 segue che (v, T_n) tende a zero per $n \to \infty$. Poiché $\|T_n\|_\infty = 1$, ne segue che

$$\lim_{n\to\infty} |r_n(\xi)| = 0, \quad \text{per ogni } \xi. \ )$$

6.24 Si costruisca il polinomio di primo grado di approssimazione ai minimi quadrati con i polinomi di Legendre per le funzioni

a) $f(x) = x^3 - 2x^2$ su $[0,1]$,
b) $f(x) = \dfrac{1}{x}$ su $[1,2]$,
c) $f(x) = \cos \pi x$ su $[0,1]$,
d) $f(x) = \mathrm{e}^{-x}$ su $[-1,1]$,
e) $f(x) = \log x$ su $[1,2]$.

(Risposta:

a) $g(x) = -\dfrac{11}{10}\,x + \dfrac{2}{15}$; b) $g(x) = 6(2 - 3\log 2)x + 2(14\log 2 - 9)$;
c) $g(x) = \dfrac{12}{\pi^2}\,(-2x + 1)$; d) $g(x) = \dfrac{1}{\mathrm{e}}\,\big(-3x + \dfrac{\mathrm{e}^2 - 1}{2}\big)$;
e) $g(x) = -3(4\log 2 - 3)x + \big(20\log 2 - \dfrac{29}{2}\big)$.)

6.25 Si determini il polinomio $g_2(x)$ di grado al più 2 di approssimazione ai minimi quadrati con i polinomi di Legendre della funzione $f(x) = \mathrm{e}^x$ nell'intervallo $[-1, 1]$, si dica quanto vale δ_2 e si determini il massimo modulo del resto.

(Traccia: si ha

$$\alpha_0^* = \frac{1}{2}\int_{-1}^{1} \mathrm{e}^x\, dx = \frac{\mathrm{e} - \mathrm{e}^{-1}}{2}\,, \quad \alpha_1^* = \frac{3}{2}\int_{-1}^{1} x\mathrm{e}^x\, dx = 3\mathrm{e}^{-1},$$
$$\alpha_2^* = \frac{5}{2}\int_{-1}^{1} \frac{3x^2-1}{2}\,\mathrm{e}^x\, dx = \frac{5(\mathrm{e} - 7\mathrm{e}^{-1})}{2}\,,$$
$$\delta_2^2 = \int_{-1}^{1} \mathrm{e}^{2x}\, dx - 2(\alpha_0^*)^2 - \frac{2}{3}\,(\alpha_1^*)^2 - \frac{2}{5}\,(\alpha_2^*)^2$$
$$= \frac{-5\mathrm{e}^4 + 72\mathrm{e}^2 - 259}{2\mathrm{e}^2} \approx 0.00144,$$

$$\max_{x\in[-1,1]} |f(x) - g_2(x)| \approx 0.0816, \quad \text{per} \quad x = 1.\ )$$

6.26 Si determini la retta di approssimazione ai minimi quadrati di $f(x) = x^2$ nell'intervallo $[0, 1]$ con i polinomi di Legendre e con i polinomi di Chebyshev di 1^a e 2^a specie. Si determini l'errore assoluto in norma 2 e il resto nei tre casi.

(Traccia: si esegua il cambiamento di variabile $y = 2x - 1$. Con i polinomi di Legendre risulta $\alpha_0^* = \dfrac{1}{3}$ e $\alpha_1^* = \dfrac{1}{2}$, da cui

$$g_1^{(L)}(y) = \frac{1}{3} + \frac{1}{2}\,y, \quad \text{quindi} \quad g_1^{(L)}(x) = x - \frac{1}{6}\,,$$

$$(\delta_1^{(L)})^2 = \int_{-1}^{1} [f(y)]^2\, dy - 2(\alpha_0^*)^2 - \frac{2}{3}\,(\alpha_1^*)^2 = \frac{1}{90} \approx 0.0111;$$

con i polinomi di Chebyshev di 1^a specie da (35) e (36) risulta $\alpha_0^* = \dfrac{3}{4}$ e $\alpha_1^* = \dfrac{1}{2}$, da cui

$$g_1^{(T)}(y) = \frac{3}{8} + \frac{1}{2}\,y, \quad \text{quindi} \quad g_1^{(T)}(x) = x - \frac{1}{8}\,,$$

$$(\delta_1^{(T)})^2 = \int_0^{\pi} \Big[f\Big(\frac{1+\cos\theta}{2}\Big)\Big]^2\, d\theta - \frac{\pi}{4}\,(\alpha_0^*)^2 - \frac{\pi}{2}\,(\alpha_1^*)^2 = \frac{\pi}{128} \approx 0.0245;$$

con i polinomi di Chebyshev di 2^a specie risulta $\alpha_0^* = \dfrac{5}{16}$ e $\alpha_1^* = \dfrac{1}{4}$, da cui

$$g_1^{(U)}(y) = \frac{5}{16} + \frac{1}{4}\,2y, \quad \text{quindi} \quad g_1^{(U)}(x) = x - \frac{3}{16}\,,$$

$$(\delta_1^{(U)})^2 = \int_{-1}^{1} f^2(y)(1-y^2)^{1/2}\,dy - \frac{\pi}{2}\left[(\alpha_0^*)^2 + (\alpha_1^*)^2\right] = \frac{\pi}{512} \approx 0.00614.$$

I tre resti sono

$$r_L(x) = x^2 - g_1^{(L)}(x) = x^2 - x + \frac{1}{6}, \quad \|r_L\|_\infty = \frac{1}{6}\,,$$

$$r_T(x) = x^2 - g_1^{(T)}(x) = x^2 - x + \frac{1}{8}\,, \quad \|r_T\|_\infty = \frac{1}{8}\,,$$

$$r_U(x) = x^2 - g_1^{(U)}(x) = x^2 - x + \frac{3}{16}\,, \quad \|r_U\|_\infty = \frac{3}{16}\,,$$

quindi

$$\|r_T\|_\infty < \|r_L\|_\infty < \|r_U\|_\infty,$$

ma nel punto di mezzo dell'intervallo risulta

$$\left|r_U\left(\frac{1}{2}\right)\right| < \left|r_L\left(\frac{1}{2}\right)\right| < \left|r_T\left(\frac{1}{2}\right)\right|. \ )$$

6.27 a) Si scriva, usando i polinomi di Chebyshev di 1^a specie, il polinomio di grado al più n di approssimazione ai minimi quadrati di

$$(1) \quad f(x) = \arcsin x, \quad (2) \quad f(x) = \arccos x, \quad \text{per} \quad x \in [-1, 1].$$

b) Si dica quanto valgono

$$\sum_{i=0}^{\infty} \frac{T_{2i+1}(x)}{(2i+1)^2} \quad \text{per} \quad x \in [-1, 1]$$

e in particolare

$$\sum_{i=0}^{\infty} \frac{1}{(2i+1)^2}\,.$$

(Traccia: a) (1) risulta $\alpha_i^* = 0$ per i pari, $\alpha_i^* = \dfrac{4}{\pi i^2}$ per i dispari, quindi il polinomio di approssimazione ai minimi quadrati di grado n dispari, è

$$g_n(x) = \frac{4}{\pi}\left[T_1(x) + \frac{1}{3^2}\,T_3(x) + \frac{1}{5^2}\,T_5(x) + \ldots + \frac{1}{n^2}\,T_n(x)\right],$$

(2) risulta $\alpha_0^* = \pi$, $\alpha_i^* = 0$ per $i \geq 2$ pari, $\alpha_i^* = -\dfrac{4}{\pi i^2}$ per i dispari, quindi

$$g_n(x) = \frac{\pi}{2} - \frac{4}{\pi}\Big[T_1(x) + \frac{1}{3^2}\,T_3(x) + \frac{1}{5^2}\,T_5(x) + \ldots + \frac{1}{n^2}\,T_n(x)\Big].$$

b) Poiché $|T_{2i+1}(x)| \leq 1$ per $x \in [-1,1]$, la serie è convergente e

$$\sum_{i=0}^{\infty} \frac{T_{2i+1}(x)}{(2i+1)^2} = \frac{\pi}{4}\,\arcsin x \quad \text{e} \quad \sum_{i=0}^{\infty} \frac{1}{(2i+1)^2} = \frac{\pi^2}{8}\,.\ )$$

6.28 Si scriva, usando i polinomi di Chebyshev di 1^a specie, il polinomio di grado al più n di approssimazione ai minimi quadrati di $f(x) = |x|$ per $-1 \leq x \leq 1$ e se ne studi la convergenza per $n \to \infty$. Si ripeta lo studio per i polinomi di Legendre, stimando i fattoriali con la formula di Stirling (si veda l'esercizio 4.43).

(Traccia: per la simmetria risulta $\alpha_i^* = 0$ per i dispari.

Con i polinomi di Chebyshev è $\alpha_0^* = \dfrac{4}{\pi}$ e per $i = 2k \geq 2$, risulta

$$\begin{aligned}\alpha_{2k}^* &= \frac{2}{\pi}\int_0^{\pi} |\cos\theta| \cos 2k\theta \; d\theta = \frac{4}{\pi}\int_0^{\pi/2} \cos\theta \cos 2k\theta \; d\theta \\ &= \frac{4}{\pi}\left[\frac{\sin(2k-1)\theta}{2(2k-1)} + \frac{\sin(2k+1)\theta}{2(2k+1)}\right]_0^{\pi/2} = \frac{4}{\pi}\,\frac{(-1)^{k+1}}{4k^2-1}\,,\end{aligned}$$

per cui

$$g_{2n}(x) = \frac{2}{\pi} - \frac{4}{\pi}\sum_{k=1}^{n} \frac{(-1)^k}{4k^2-1}\,T_{2k}(x).$$

La funzione $f(x) = |x|$ è lipschitziana in $[-1,1]$, quindi per il teorema 6.27 vi è convergenza. La convergenza è però molto lenta: infatti per ogni x, $|x| \leq 1$, tenendo conto che $|T_{2k}(x)| \leq 1$, è (si vedano gli esercizi 4.17 c) e 4.16 p))

$$\begin{aligned}\big|\,|x| - g_{2n}(x)\big| &= \frac{4}{\pi}\Big|\sum_{k=n+1}^{\infty} \frac{(-1)^k}{4k^2-1}\,T_{2k}(x)\Big| \leq \frac{4}{\pi}\sum_{k=n+1}^{\infty} \frac{1}{4k^2-1} \\ &= \frac{4}{\pi}\left[\frac{1}{2} - \frac{n}{2n+1}\right] = \frac{2}{(2n+1)\pi}\,.\end{aligned}$$

Poiché $T_{2k}(0) = (-1)^k$, risulta in particolare

$$\max_{x\in[-1,1]} \big|\,|x| - g_{2n}(x)\big| = |g_{2n}(0)| = \frac{2}{(2n+1)\pi}\,.$$

Con i polinomi di Legendre è $\alpha_0^* = \dfrac{1}{2}$, per $i = 2k \geq 2$ risulta

$$\begin{aligned}\alpha_{2k}^* &= \frac{4k+1}{2} \int_{-1}^{1} |x| P_{2k}(x)\, dx = -(4k+1) \int_{-1}^{0} x P_{2k}(x)\, dx \\ &= -(4k+1) \left[xq(x) \Big|_{-1}^{0} - \int_{-1}^{0} q(x)\, dx \right],\end{aligned}$$

dove $q(x)$ è una primitiva di $P_{2k}(x)$. Per la h) dell'esercizio 6.15 è

$$q(x) = \frac{P_{2k+1}(x) - P_{2k-1}(x)}{4k+1}, \quad \text{quindi} \quad q(-1) = 0 \text{ e } q(0) = 0.$$

Inoltre una primitiva di $q(x)$ è

$$\frac{1}{4k+1} \left[\frac{P_{2k+2}(x) - P_{2k}(x)}{4k+3} - \frac{P_{2k}(x) - P_{2k-2}(x)}{4k-1} \right].$$

Si verifichi, utilizzando la e) dell'esercizio 6.15, che

$$\int_{-1}^{0} q(x)\, dx = (-1)^{k+1} \frac{(2k-2)!}{2^{2k}\,(k-1)!\,(k+1)!},$$

per cui

$$\alpha_{2k}^* = (-1)^{k+1} \frac{(2k-2)!\,(4k+1)}{2^{2k}\,(k-1)!\,(k+1)!},$$

e quindi

$$g_{2n}(x) = \frac{1}{2} - \sum_{k=1}^{n} (-1)^k \frac{(2k-2)!(4k+1)}{2^{2k}\,(k-1)!\,(k+1)!}\, P_{2k}(x).$$

Con la formula di Stirling per k grande si ha

$$\begin{aligned}|\alpha_{2k}^*| &\sim \frac{\sqrt{2\pi}(2k-2)^{2k-3/2}\mathrm{e}^{-(2k-2)}(4k+1)}{2^{2k}\sqrt{2\pi}(k-1)^{k-1/2}\mathrm{e}^{-(k-1)}\sqrt{2\pi}(k+1)^{k+3/2}\mathrm{e}^{-(k+1)}} \\ &= \frac{\mathrm{e}^2}{4\sqrt{\pi}} \left(\frac{k-1}{k+1}\right)^k \frac{4k+1}{(k-1)(k+1)^{3/2}}.\end{aligned}$$

Poiché per k grande è

$$\left(\frac{k-1}{k+1}\right)^k \sim \mathrm{e}^{-2} \quad \text{e} \quad \frac{4k+1}{(k-1)(k+1)^{3/2}} \sim \frac{4}{k^{3/2}},$$

risulta

$$|\alpha_{2k}^*| \sim \frac{1}{\sqrt{\pi}\ k^{3/2}}\,.$$

Ne segue la convergenza di $g_n(x)$ a $f(x)$.)

6.29 Si scriva l'espansione in serie di Chebyshev della funzione

$$f(x) = \frac{1}{a+x}\,, \quad a > 1, \quad x \in [-1, 1].$$

Nel caso particolare di $a = 3$ si scriva l'espansione troncata al quarto termine e si dica quant'è il massimo modulo del resto.

(Traccia: posto

$$\frac{1}{a+x} = \frac{\alpha_0^*}{2} + \sum_{i=1}^{\infty} \alpha_i^* T_i(x),$$

si imponga che

$$1 = (a+x)\Big(\frac{\alpha_0^*}{2} + \sum_{i=1}^{\infty} \alpha_i^* T_i(x)\Big),$$

cioè

$$T_0(x) = a\,\frac{\alpha_0^*}{2}\,T_0(x) + \sum_{i=1}^{\infty} a\alpha_i^* T_i(x) + \frac{\alpha_0^*}{2}\,x + \sum_{i=1}^{\infty} \alpha_i^* x T_i(x).$$

Tenendo conto che per la (25) è

$$x = T_1(x) \quad \text{e} \quad xT_i(x) = \frac{1}{2}\Big[T_{i-1}(x) + T_{i+1}(x)\Big], \quad i = 1, 2, \ldots,$$

si verifichi che i coefficienti α_i^* devono soddisfare alle relazioni

$$\begin{aligned} &a\alpha_0 + \alpha_1 = 2 \\ &\alpha_i + 2a\alpha_{i+1} + \alpha_{i+2} = 0, \quad \text{per} \quad i = 0, 1, \ldots \end{aligned} \tag{104}$$

La (104) è una equazione alle differenze omogenea del secondo ordine, la cui soluzione generale è

$$\alpha_i = \gamma\rho_1^i + \delta\rho_2^i, \quad \rho_1 = \sqrt{a^2-1} - a, \quad \rho_2 = -\sqrt{a^2-1} - a, \quad \gamma,\, \delta \in \mathbf{R}.$$

Per $a > 1$ risulta $|\rho_1| < 1$ e $|\rho_2| > 1$, quindi vi potrà essere convergenza solo escludendo la presenza di ρ_2 nella soluzione. Si pone quindi $\delta = 0$.

Per determinare γ ci si serve della relazione $a\alpha_0 + \alpha_1 = 2$, ottenendo $\gamma = \dfrac{2}{\sqrt{a^2-1}}$. Quindi si ha

$$\alpha_i^* = \frac{2}{\sqrt{a^2-1}} \left(\sqrt{a^2-1} - a\right)^i.$$

Nel caso particolare è $\alpha_i^* = \dfrac{(2\sqrt{2}-3)^i}{\sqrt{2}}$,

$$\begin{aligned} g_4(x) &= \frac{\sqrt{2}}{2}\left[\frac{1}{2} + \sum_{i=1}^{4}(2\sqrt{2}-3)^i T_i(x)\right] \\ &= \sqrt{2}\left[\frac{1121}{4} - 198\sqrt{2} + (147 - 104\sqrt{2})x - (2291 - 1620\sqrt{2})x^2 \right. \\ &\quad \left. - (198 - 140\sqrt{2})x^3 + (2308 - 1632\sqrt{2})x^4\right], \end{aligned}$$

e risulta

$$\begin{aligned} \max_{x\in[-1,1]} \left|f(x) - g_4(x)\right| &= f(-1) - g_4(-1) \\ &= \frac{1}{2} - \frac{\sqrt{2}}{2}\left[\frac{1}{2} + \sum_{i=1}^{4}(-1)^i(2\sqrt{2}-3)^i\right] \approx 0.127\ 10^{-3}. \end{aligned}$$

6.30 Si scriva, usando i polinomi di Laguerre, il polinomio di grado al più n di approssimazione ai minimi quadrati di

$$f(x) = \mathrm{e}^{-ax}, \quad a > 0,$$

sull'intervallo $[0, +\infty]$. In particolare per $n = 3$ si considerino i casi $a = 1$ e $a = 2$ e si determini l'errore assoluto in norma 2.

(Traccia: si ha

$$\alpha_i^* = \int_0^\infty \mathrm{e}^{-x} f(x) L_i(x)\, dx = \frac{1}{i!} \int_0^\infty \mathrm{e}^{-ax} \frac{d^i}{dx^i}\left(\mathrm{e}^{-x} x^i\right) dx.$$

Integrando successivamente per parti e notando che

$$\left.\frac{d^k}{dx^k}\left(\mathrm{e}^{-x} x^i\right)\right|_0^\infty = 0, \quad \text{per} \quad k = 0, 1, \ldots, i-1,$$

risulta

$$\alpha_i^* = \frac{a^i}{i!} \int_0^\infty \mathrm{e}^{-ax}\mathrm{e}^{-x} x^i\, dx = \frac{a^i}{(a+1)^{i+1}}.$$

Si ha quindi

$$g_n(x) = \frac{1}{a+1} \sum_{i=0}^{n} \Big(\frac{a}{a+1}\Big)^i L_i(x).$$

Casi particolari: se $a = 1$, è $f(x) = \mathrm{e}^{-x}$ e

$$\begin{aligned} g_3(x) &= \frac{1}{2}\, L_0(x) + \frac{1}{4}\, L_1(x) + \frac{1}{8}\, L_2(x) + \frac{1}{16}\, L_3(x) \\ &= \frac{1}{96}\,(-x^3 + 15x^2 - 66x + 90), \end{aligned}$$

con l'errore

$$\delta_3^2 = \int_0^\infty \mathrm{e}^{-3x}\, dx - \sum_{i=0}^{3} (\alpha_i^*)^2 = \frac{1}{768} \approx 0.130\ 10^{-2},$$

se $a = 2$ è $f(x) = \mathrm{e}^{-2x}$ e

$$\begin{aligned} g_3(x) &= \frac{1}{3}\, L_0(x) + \frac{2}{9}\, L_1(x) + \frac{4}{27}\, L_2(x) + \frac{8}{81}\, L_3(x) \\ &= \frac{1}{243}\,(-4x^3 + 54x^2 - 198x + 195), \end{aligned}$$

con l'errore

$$\delta_3^2 = \int_0^\infty \mathrm{e}^{-5x}\, dx - \sum_{i=0}^{3} (\alpha_i^*)^2 = \frac{256}{32805} \approx 0.780\ 10^{-2}.)$$

6.31 Si scriva, usando i polinomi di Hermite, il polinomio di grado al più n di approssimazione ai minimi quadrati di $f(x) = |x|$ per $-\infty < x < \infty$. Si disegni il grafico di $g_6(x)$.

(Traccia: per la simmetria è $\alpha_i^* = 0$ per i dispari. Per i pari risulta

$$\begin{aligned} \alpha_0^* &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} |x|\, \mathrm{e}^{-x^2} H_0(x)\, dx = \frac{2}{\sqrt{\pi}} \int_0^\infty x\, \mathrm{e}^{-x^2}\, dx = \frac{1}{\sqrt{\pi}}\,, \\ \alpha_{2k}^* &= \frac{1}{h_{2k}} \int_{-\infty}^{\infty} |x|\, \mathrm{e}^{-x^2} H_{2k}(x)\, dx = \frac{2}{h_{2k}} \int_0^\infty x\, \frac{d^{2k}}{dx^{2k}}\, \mathrm{e}^{-x^2}\, dx \\ &= \frac{2}{h_{2k}} \left[x\, \frac{d^{2k-1}}{dx^{2k-1}}\, \mathrm{e}^{-x^2} - \frac{d^{2k-2}}{dx^{2k-2}}\, \mathrm{e}^{-x^2} \right]_0^\infty = \frac{2}{h_{2k}}\, \frac{d^{2k-2}}{dx^{2k-2}}\, \mathrm{e}^{-x^2} \Big|_{x=0} \\ &= \frac{2}{h_{2k}}\, H_{2k-2}(0), \end{aligned}$$

e per l'esercizio 6.19 b) risulta che

$$\alpha^*_{2k} = 2\,\frac{(-1)^{k+1}2^{k-1}1\cdot 3\cdots(2k-3)}{2^{2k}(2k)!\,\sqrt{\pi}} = \frac{(-1)^{k+1}}{\sqrt{\pi}\,2^{2k}k!\,(2k-1)}\,.$$

Si ha quindi

$$g_{2n}(x) = \frac{1}{\sqrt{\pi}}\Big[1 - \sum_{k=1}^{n} \frac{(-1)^k}{2^{2k}k!\,(2k-1)}\; H_{2k}(x)\Big].$$

Nella figura 6.27 sono riportati i grafici di $g_6(x)$ e di $|x|$.

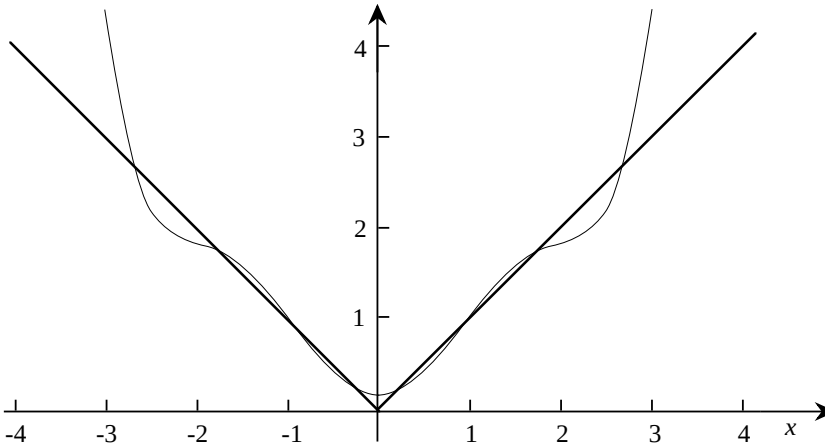


Fig. 6.27 - Approssimazione di $|x|$ con i polinomi di Hermite.

6.32 Si scriva, usando i polinomi di Hermite, il polinomio di grado al più n di approssimazione ai minimi quadrati di $f(x) = \mathrm{e}^{ax}$, con $a > 0$, per $-\infty < x < \infty$. Si consideri il caso particolare di $a = 2$.

(Traccia: si ha

$$\alpha^*_i = \frac{1}{h_i}\;\int_{-\infty}^{\infty} \mathrm{e}^{-x^2} f(x) H_i(x)\; dx = \frac{(-1)^i}{h_i}\;\int_{-\infty}^{\infty} \mathrm{e}^{ax}\;\frac{d^i}{dx^i}\,\mathrm{e}^{-x^2}\;dx.$$

Integrando successivamente per parti e notando che

$$\frac{d^k}{dx^k}\,\mathrm{e}^{-x^2}\Big|_{-\infty}^{\infty} = 0, \quad \text{per} \quad k = 0,\ldots,i-1,$$

risulta

$$\alpha^*_i = \frac{a^i}{h_i}\;\int_{-\infty}^{\infty} \mathrm{e}^{-x^2+ax}\;dx = \frac{a^i}{h_i}\;\sqrt{\pi}\;\mathrm{e}^{a^2/4} = \frac{a^i}{2^i i!}\;\mathrm{e}^{a^2/4}.$$

Si ha quindi

$$g_n(x) = \mathrm{e}^{a^2/4} \sum_{i=0}^{n} \frac{a^i}{2^i i!} H_i(x).$$

Nel caso particolare

$$g_n(x) = \mathrm{e} \sum_{i=0}^{n} \frac{1}{i!} H_i(x). \ )$$

6.33 Sia $n = 2^h,\ h > 1$ intero, e sia

$$p(x) = \sum_{i=0}^{n-1} \alpha_i T_i(x) = \sum_{k=0}^{n-1} a_k x^k.$$

Si verifichi che

a) dati i coefficienti α_i, $i = 0, \ldots, n-1$, è possibile calcolare i valori che $p(x)$ assume negli n punti

$$x_j = \cos \frac{j\pi}{n}, \quad j = 0, \ldots, n-1,$$

con $O(n \log_2 n)$ operazioni;

b) dati i coefficienti α_i, $i = 0, \ldots, n-1$, è possibile calcolare i coefficienti a_k, $k = 0, \ldots, n-1$, con $O(n \log_2^2 n)$ operazioni;

c) dati i coefficienti α_i, $i = 0, \ldots, n-1$, è possibile calcolare i valori che $p(x)$ assume in n punti distinti $y_0, \ldots, y_{n-1}$ con $O(n \log_2^2 n)$ operazioni.

(Traccia: a) risulta per la (26)

$$p(x_j) = \sum_{i=0}^{n-1} \alpha_i \cos \frac{ij\pi}{n}, \quad j = 0, \ldots, n-1.$$

Il vettore $[p(x_0), \ldots, p(x_{n-1})]^T$ può quindi essere calcolato (si veda l'esercizio 5.62) per mezzo di una trasformata di coseni di ordine $2n$ con $O(n \log_2 n)$ operazioni; b) si calcolino prima i valori $p(x_i)$, $i = 0, \ldots, n-1$, come sopra, e si applichi poi l'esercizio 5.72; c) si proceda come in b), poi si applichi l'esercizio 5.69.)

6.34 Sia $f(x) \in C^2[-1, 1]$ e sia

$$g_n(x) = \frac{\alpha_0^*}{2} + \sum_{i=1}^{n} \alpha_i^* T_i(x).$$

Si verifichi che

$$|f(x) - g_n(x)| \leq \frac{4(M_1 + M_2)}{n\pi},$$

dove

$$M_1 = \max_{x\in[-1,1]} |f'(x)|, \quad M_2 = \max_{x\in[-1,1]} |f''(x)|.$$

(Traccia: per $x \in [-1, 1]$ è

$$|f(x) - g_n(x)| = \Big| \sum_{i=n+1}^{\infty} \alpha_i^* T_i(x) \Big| \leq \sum_{i=n+1}^{\infty} |\alpha_i^*|,$$

e integrando 2 volte per parti si ha

$$\begin{aligned}\alpha_i^* &= \frac{2}{\pi}\int_0^\pi f(\cos\theta)\cos i\theta \; d\theta = \frac{2}{i\pi}\int_0^\pi f'(\cos\theta)\sin i\theta \sin\theta \; d\theta \\ &= \frac{2}{i^2\pi}\int_0^\pi \left[f'(\cos\theta)\cos\theta - f''(\cos\theta)\sin^2\theta\right]\cos i\theta \; d\theta,\end{aligned}$$

da cui

$$|\alpha_i^*| \leq \frac{2}{i^2\pi}\Big[M_1\int_0^\pi |\cos\theta| \; d\theta + M_2\int_0^\pi \sin^2\theta \; d\theta\Big] \leq \frac{4(M_1 + M_2)}{i^2\pi},$$

e (si vedano gli esercizi 4.17 c) e 4.16 p))

$$\begin{aligned}\sum_{i=n+1}^{\infty} |\alpha_i^*| &\leq \frac{4(M_1 + M_2)}{\pi} \sum_{i=n+1}^{\infty} \frac{1}{i^2} \leq \frac{16(M_1 + M_2)}{\pi} \sum_{i=n+1}^{\infty} \frac{1}{4i^2 - 1} \\ &= \frac{8(M_1 + M_2)}{\pi(2n + 1)} \leq \frac{4(M_1 + M_2)}{n\pi}.\end{aligned}$$

Questa maggiorazione è ovviamente peggiore di quella del teorema 6.27 nelle stesse ipotesi.)

6.35 a) Si verifichi che le funzioni

$$\cos ix, \quad \text{per } i = 0, 1, \ldots, \quad \sin ix, \quad \text{per } i = 1, 2, \ldots$$

costituiscono una base ortogonale dello spazio delle funzioni continue e periodiche su $[-\pi, \pi]$ con il prodotto scalare

$$(f, g) = \int_{-\pi}^{\pi} f(x)g(x) \; dx.$$

b) Per una funzione $f(x) \in C[-\pi, \pi]$ si determinino i coefficienti $\alpha_i^*,\ i = 0, 1, \ldots$ e $\beta_i^*,\ i = 1, 2, \ldots$ tali che il polinomio trigonometrico

$$g_n(x) = \frac{\alpha_0^*}{2} + \sum_{i=1}^{n} \left(\alpha_i^* \cos ix + \beta_i^* \sin ix\right)$$

sia di migliore approssimazione *trigonometrica* rispetto alla norma indotta dal prodotto scalare. Si verifichi che $g_n(x)$ è uguale alla somma parziale n-esima della serie di Fourier di $f(x)$. Si dica quant'è l'errore assoluto in norma δ_n e sotto quale ipotesi vale $\lim_{n\to\infty} \delta_n = 0$.

c) Si verifichi che se $f(x)$ è una funzione pari, la $g_n(x)$ ha un'espansione in soli coseni, e se $f(x)$ è una funzione dispari, la $g_n(x)$ ha un'espansione in soli seni. Poiché una funzione $f(x) \in C[0, \pi]$ può essere prolungata sull'intervallo $[-\pi, \pi]$ in modo da ottenere una funzione pari oppure, se $f(0) = 0$, una funzione dispari, si possono ottenere approssimazioni $g_n(x)$ con espansioni in soli coseni oppure in soli seni.

d) Si determini la migliore approssimazione trigonometrica per le funzioni

$$(1)\quad f(x) = |x|, \quad (2)\quad f(x) = |\sin x|, \quad (3)\quad f(x) = x^2,$$

sull'intervallo $[-\pi, \pi]$.

(Traccia: a) si verifichi che

$$\int_{-\pi}^{\pi} \sin ix\ dx = \int_{-\pi}^{\pi} \cos ix\ dx = 0, \quad \text{se} \quad i \neq 0,$$

$$\int_{-\pi}^{\pi} \sin ix \sin kx\ dx = \int_{-\pi}^{\pi} \cos ix \cos kx\ dx = \begin{cases} 0 & \text{se } i \neq k, \\ \pi & \text{se } i = k \neq 0, \end{cases}$$

$$\int_{-\pi}^{\pi} \sin ix \cos kx\ dx = 0.$$

Per la completezza si veda [14, pag. 267]. b) Risulta $h_0 = 2\pi$ e $h_i = \pi$, per $i = 1, 2, \ldots$, quindi

$$\alpha_i^* = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos ix\ dx, \quad \beta_i^* = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin ix\ dx,$$

e

$$\delta_n^2 = \int_{-\pi}^{\pi} f^2(x)\ dx - \frac{\pi}{2}\,(\alpha_0^*)^2 - \pi \sum_{i=1}^{n} (\alpha_i^*)^2.$$

La convergenza a zero di δ_n vale quando la serie di Fourier di $f(x)$ è convergente, ad esempio quando la $f(x)$ è lipschitziana in $[-\pi, \pi]$.
c) Se la funzione $f(x)$ è pari risulta

$$\beta_i^* = 0 \quad \text{e} \quad \alpha_i^* = \frac{2}{\pi} \int_0^\pi f(x) \cos ix \; dx \quad \text{per} \quad i = 0, 1, \ldots$$

se la funzione $f(x)$ è dispari risulta

$$\alpha_i^* = 0 \quad \text{e} \quad \beta_i^* = \frac{2}{\pi} \int_0^\pi f(x) \sin ix \; dx \quad \text{per} \quad i = 1, 2, \ldots$$

d) Nei casi particolari si ha

$$(1) \quad g_{2n+1}(x) = \frac{\pi}{2} - \frac{4}{\pi} \left(\frac{\cos x}{1^2} + \frac{\cos 3x}{3^2} + \ldots + \frac{\cos(2n+1)x}{(2n+1)^2} \right),$$

$$(2) \quad g_{2n}(x) = \frac{2}{\pi} - \frac{4}{\pi} \left(\frac{\cos 2x}{1 \cdot 3} + \frac{\cos 4x}{3 \cdot 5} + \ldots + \frac{\cos(2n)x}{(2n-1) \cdot (2n+1)} \right),$$

$$(3) \quad g_n(x) = \frac{\pi^2}{3} - 4 \left(\frac{\cos x}{1^2} - \frac{\cos 2x}{2^2} + \ldots + (-1)^{n-1} \frac{\cos nx}{n^2} \right). \;)$$

6.36 Si verifichi che per ogni intero $n \geq 1$ è

a) $$s_n = \int_0^\pi \left| \frac{\sin(n + 1/2)\theta}{\sin \theta/2} \right| \; d\theta > \frac{4}{\pi} \log n;$$

b) si dimostri che esiste una funzione continua su $[-1, 1]$ per cui la successione $\{g_n(x)\}$ delle approssimazioni ai minimi quadrati con polinomi di Chebyshev di 1^a specie non converge in norma ∞.

(Traccia: a) poiché $\sin \dfrac{\theta}{2} \leq \dfrac{\theta}{2}$ per $\theta \in [0, \pi]$, è

$$s_n \geq 2 \int_0^\pi \frac{\left| \sin(n + 1/2)\theta \right|}{\theta} \; d\theta.$$

Con il cambiamento di variabile $\theta = \dfrac{\pi x}{n + 1/2}$ risulta

$$s_n \geq 2 \int_0^{n+1/2} \frac{\left| \sin \pi x \right|}{x} \; dx > 2 \int_0^n \frac{\left| \sin \pi x \right|}{x} \; dx = 2 \sum_{i=0}^{n-1} \int_i^{i+1} \frac{\left| \sin \pi x \right|}{x} \; dx$$

$$= 2 \int_0^1 \left[\sin \pi x \sum_{i=0}^{n-1} \frac{1}{x + i} \right] dx > 2 \sum_{i=0}^{n-1} \frac{1}{i+1} \int_0^1 \sin \pi x \; dx > \frac{4}{\pi} \log n,$$

in quanto è

$$\sum_{i=0}^{n-1} \frac{1}{i+1} > \int_0^n \frac{dx}{x+1} > \log n.$$

b) Da (35) e (36) si ha

$$g_n(1) = \frac{\alpha_0^*}{2} + \sum_{j=1}^{n} \alpha_j^* = \frac{2}{\pi} \int_0^\pi \Big[\frac{1}{2} + \sum_{j=1}^{n} \cos j\theta\Big] f(\cos\theta)\ d\theta.$$

Per l'esercizio 4.18 b) è

$$g_n(1) = \frac{1}{\pi} \int_0^\pi \frac{\sin(n+1/2)\theta}{\sin\theta/2}\ f(\cos\theta)\ d\theta.$$

Se per ogni $f(x)$ continua $g_n(1)$ fosse limitato, per il teorema della limitatezza uniforme di Banach-Steinhaus [41], esisterebbe una costante M tale che

$$\int_0^\pi \left|\frac{\sin(n+1/2)\theta}{\sin\theta/2}\right| d\theta \le M \quad \text{per ogni } n,$$

che è assurdo per il punto a). Ne segue che esiste $f(x) \in C[-1,1]$ tale che

$$\max \lim_{n\to\infty} \big|f(1) - g_n(1)\big| = \infty.\ )$$

6.37 Sia (f,h) un prodotto scalare definito su $C[a,b]$ e sia $\|f\| = \sqrt{(f,f)}$ la norma da esso indotta.

a) Si verifichi che vale la *legge del parallelogramma*

$$\|f+h\|^2 + \|f-h\|^2 = 2\big(\|f\|^2 + \|h\|^2\big), \quad f,h \in C[a,b];$$

b) si verifichi che la norma ∞ non soddisfa la legge del parallelogramma, e pertanto non è indotta da alcun prodotto scalare.

(Traccia: a) si applichino le proprietà a) e b) della definizione 6.5 a $(f+h, f+h) + (f-h, f-h)$; b) siano ad esempio $f(x) = x$, $h(x) = 1$, e $[a,b] = [0,1]$. Risulta $\|f+h\|_\infty = 2$, $\|f-h\|_\infty = 1$, $\|f\|_\infty = 1$, $\|h\|_\infty = 1$.)

6.38 Siano $\phi_0(x), \ldots, \phi_n(x)$, $n+1$ funzioni continue definite su un intervallo $[a,b]$.

a) Si verifichi che le seguenti condizioni, dette *condizioni di Haar*, sono equivalenti:

(1) la matrice $\Phi \in \mathbf{R}^{(n+1)\times(n+1)}$, il cui elemento (i,j)-esimo è $\phi_j(x_i)$, per $i,j = 0,\ldots,n$, è non singolare per ogni n-upla $x_0,\ldots,x_n$ di punti distinti di $[a,b]$;

(2) nessuna combinazione lineare

$$\sum_{j=0}^{n} \alpha_j \phi_j(x)$$

che non sia identicamente nulla, ha più di n zeri in $[a,b]$;

(3) c'è un'unica combinazione lineare delle $\phi_j(x)$, $j = 0,\ldots,n$, che interpola una qualsiasi funzione definita su $[a,b]$ in $n+1$ punti distinti $x_0,\ldots,x_n \in [a,b]$.

Un insieme di $n+1$ funzioni continue che soddisfano le condizioni di Haar su un intervallo $[a,b]$ viene detto *insieme di Chebyshev* su $[a,b]$.

b) Si verifichi che l'insieme $\phi_j(x) = x^j$, $j = 0,\ldots,n$, è un insieme di Chebyshev su ogni intervallo $[a,b]$;

c) si dica quale dei seguenti insiemi di funzioni è un insieme di Chebyshev:

(i) $\{1, x^2, x^4\}$ su $[0,1]$,

(ii) $\{1, x^2, x^4\}$ su $[-1,1]$,

(iii) $\{\dfrac{1}{x}, \dfrac{1}{x+1}, \dfrac{1}{x+2}\}$ su $[1,2]$,

(iv) $\{1, \mathrm{e}^x, \mathrm{e}^{2x}\}$ su $[0,1]$;

d) Si verifichi che il teorema di de la Vallée-Poussin e il teorema di equioscillazione di Chebyshev valgono anche quando al posto dello spazio $\mathcal{P}_n$ dei polinomi si consideri lo spazio $\mathcal{G}$ delle combinazioni lineari

$$g_n(x) = \sum_{j=0}^{n} \alpha_j \phi_j(x),$$

in cui le funzioni $\phi_0(x),\ldots,\phi_k(x)$ formano un insieme di Chebyshev per $k = 0,\ldots,n$.

e) Si determini l'approssimazione minimax di x^2 per $0 \le x \le 2$, utilizzando i due insiemi di funzioni

$$(v) \quad \{1, \mathrm{e}^x\}, \qquad (vi) \quad \{x, \mathrm{e}^x\}.$$

(Traccia: a) si esprimano le condizioni (2) e (3) mediante sistemi lineari la cui matrice dei coefficienti sia Φ; b) la matrice Φ è in questo caso una matrice di Vandermonde; c) (i) sì, (ii) no, (iii) sì, (iv) sì.

d) Si proceda in modo analogo alle dimostrazioni dei teoremi 6.28 e 6.30. Per il teorema di de la Vallée-Poussin si sostituiscano al polinomio $p_n(x)$ la combinazione $g_n(x)$ e al polinomio $q(x)$ una combinazione

$$\sum_{j=0}^{n} \beta_j \phi_j(x).$$

La conclusione segue dalla condizione (2) di Haar. Per il teorema di equioscillazione di Chebyshev si sostituisca al polinomio $p_n^*(x)$ la combinazione lineare

$$g_n^*(x) = \sum_{j=0}^{n} \alpha_j^* \phi_j(x).$$

Per il caso $k = 1$, sia

$$H = \max_{x \in [a,b]} \frac{\phi_0(x)}{\phi_0(a)},$$

(si noti che $\phi_0(x) \neq 0$ per $x \in [a, b]$) e si sostituisca al posto di $q(x)$ la funzione

$$\frac{M - |m|}{2H} \frac{\phi_0(x)}{\phi_0(a)}.$$

Per il caso $k > 1$ si sostituisca al polinomio $s(x)$ una combinazione lineare non nulla

$$s(x) = \sum_{i=0}^{k-1} \beta_i \phi_i(x)$$

che si annulli nei $k - 1$ punti $\xi_1, \ldots, \xi_{k-1}$ e tale che $s(a) = 1$ (l'esistenza di $s(x)$ è garantita dalla condizione (3) di Haar). Per l'unicità si utilizzi la condizione (2) di Haar.

e) (v) è un insieme di Chebyshev, infatti l'equazione $\mathrm{e}^x + c = 0$, con c costante reale, non ha più di uno zero; per determinare $g_1^*(x) = \alpha_0^* + \alpha_1^* \mathrm{e}^x$, si impone che vi siano tre punti di equioscillazione. Poiché il resto

$$r^*(x) = x^2 - \alpha_0^* - \alpha_1^* \mathrm{e}^x$$

ha la derivata seconda

$$[r^*(x)]'' = 2 - \alpha_1^* \mathrm{e}^x$$

che si può annullare in un punto interno all'intervallo $[0, 2]$, $r^*(x)$ può non essere standard. Si assumono come punti di equioscillazione $0 = x_0^* < x_1^* < x_2^* \leq 2$ e si ottiene il sistema

$$\begin{cases} -\alpha_0 - \alpha_1 = d \\ x_1^2 - \alpha_0 - \alpha_1 \mathrm{e}^{x_1} = -d \\ x_2^2 - \alpha_0 - \alpha_1 \mathrm{e}^{x_2} = d \\ 2x_1 - \alpha_1 \mathrm{e}^{x_1} = 0 \\ 2x_2 - \alpha_1 \mathrm{e}^{x_2} = 0, \end{cases}$$

la cui soluzione è

$$\alpha_0^* = -0.7339367,\ \alpha_1^* = 0.6476102,\ x_1^* = 0.5760462,\ x_2^* = 1.593624,$$
$$|d| = 0.08632648.$$

(*vi*) non è un insieme di Chebyshev, infatti l'equazione $\mathrm{e}^x + cx = 0$ ha due zeri nell'intervallo $[0, 2]$ per $c = -3$, quindi non è detto che l'approssimazione cercata, della forma $g_1^*(x) = \alpha_0^* x + \alpha_1^* \mathrm{e}^x$, goda della proprietà di equioscillazione. Infatti, imponendo che vi siano tre punti di equioscillazione $0 = x_0^* < x_1^* < x_2^* = 2$, si ottiene il sistema

$$\begin{cases} -\alpha_1 = d \\ x_1^2 - \alpha_0 x_1 - \alpha_1 \mathrm{e}^{x_1} = -d \\ 4 - 2\alpha_0 - \alpha_1 \mathrm{e}^2 = d \\ 2x_1 - \alpha_0 - \alpha_1 \mathrm{e}^{x_1} = 0, \end{cases}$$

la cui soluzione è

$$\alpha_0^* = 8.465577,\ \alpha_1^* = -2.023954,\ x_1^* = 1.122734,\ |d| = 2.023954.$$

Il resto presenta due punti di massimo, 0 e 2, e un punto di minimo x_1^*, ma la funzione così ottenuta non è la migliore approssimazione possibile. Imponendo invece che vi siano due punti di equioscillazione $0 < x_0^* < x_1^* = 2$ e si ottiene il sistema

$$\begin{cases} x_0^2 - \alpha_0 x_0 - \alpha_1 \mathrm{e}^{x_0} = d \\ 4 - 2\alpha_0 - \alpha_1 \mathrm{e}^2 = -d \\ 2x_0 - \alpha_0 - \alpha_1 \mathrm{e}^{x_0} = 0. \end{cases}$$

Ricavando d in funzione del parametro x_0 e minimizzando si ottiene

$$\alpha_0^* = 0.1842325,\ \alpha_1^* = 0.4186312,\ x_0^* = 0.4063757,\ |d| = 0.5382453.$$

Questo valore di $|d|$ è molto minore di quello ottenuto con tre punti di equioscillazione.)

6.39 Sia $p_n^*(x)$ il polinomio di approssimazione minimax di una funzione $f(x)$ sull'intervallo $[-a, a]$, $a > 0$. Si verifichi che se $f(x)$ è simmetrica (antisimmetrica) anche $p_n^*(x)$ è simmetrico (antisimmetrico).

(Traccia: se $f(x)$ è simmetrica, si ha

$$\max_{x\in[-a,a]} |f(x)-p_n^*(x)| = \max_{x\in[-a,a]} |f(-x)-p_n^*(-x)| = \max_{x\in[-a,a]} |f(x)-p_n^*(-x)|.$$

Quindi $p_n^*(x)$ e $p_n^*(-x)$ sono entrambi polinomi di approssimazione minimax, e per l'unicità deve essere $p_n^*(x) = p_n^*(-x)$. Si proceda in modo analogo per il caso antisimmetrico.)

6.40 Sia $f(x) \in C[a, b]$. Si verifichi che l'approssimazione minimax costante di $f(x)$ su $[a, b]$ è

$$p_0^* = \frac{1}{2}\left[\max_{x\in[a,b]} f(x) + \min_{x\in[a,b]} f(x)\right]$$

e risulta

$$r^* = \|f - p_0^*\|_\infty = \frac{1}{2}\left[\max_{x\in[a,b]} f(x) - \min_{x\in[a,b]} f(x)\right].$$

(Traccia: si verifichi che $f(x) - p_0^*$ ha massimo r^* e minimo $-r^*$.)

6.41 Si determinino le approssimazioni minimax di grado al più 3 e di grado al più 5 della funzione $f(x) = |x|$ per $|x| \le 1$. Si confronti il secondo polinomio ottenuto con quelli degli esempi 6.23 e 6.25.

(Traccia: nel caso $n = 3$ per simmetria si assume che vi siano 5 punti di equioscillazione

$$-1 = x_0^* < x_1^* < x_2^* = 0 < x_3^* < x_4^* = 1, \quad \text{e} \quad x_1^* = -x_3^*.$$

Poiché il polinomio deve essere pari, si pone $p_3(x) = a_2x^2 + a_0$ e si ottiene il sistema

$$\begin{cases} -a_0 = d \\ x_3 - a_2x_3^2 - a_0 = -d \\ 1 - a_2 - a_0 = d \\ 1 - 2a_2x_3 = 0, \end{cases}$$

da cui si ha

$$p_3^*(x) = x^2 + \frac{1}{8}, \quad x_3^* = \frac{1}{2}, \quad |d| = \frac{1}{8}.$$

Nel caso $n = 5$ si assume analogamente

$$-1 = x_0^* < x_1^* < x_2^* < x_3^* = 0 < x_4^* < x_5^* < x_6^* = 1, \quad x_2^* = -x_4^*,\ x_1^* = -x_5^*.$$

Si pone $p_5(x) = a_4x^4 + a_2x^2 + a_0$ e si ottiene il sistema

$$\begin{cases} -a_0 = d \\ x_4 - a_4x_4^4 - a_2x_4^2 - a_0 = -d \\ x_5 - a_4x_5^4 - a_2x_5^2 - a_0 = d \\ 1 - a_4 - a_2 - a_0 = -d \\ 1 - 4a_4x_4^3 - 2a_2x_4 = 0 \\ 1 - 4a_4x_5^3 - 2a_2x_5 = 0, \end{cases}$$

da cui si ha

$$p_5^*(x) = -1.065541x^4 + 1.930299x^2 + 0.0676209,$$
$$x_4^* = 0.2844316, \quad x_5^* = 0.7770815, \quad |d| = 0.0676209. \ )$$

6.42 a) Siano $-1 \le x_0 < x_1 < \ldots < x_n \le 1$. Si dimostri che non esiste alcun polinomio monico $p_n(x)$ di grado n tale che

$$p_n(x_i) = (-1)^i d_i, \quad |d_i| > \frac{1}{2^{n-1}}, \quad i = 0, \ldots, n.$$

b) Sia $f(x) \in C^{n+1}[a,b]$ e siano $0 \le m \le M$ tali che

$$m \le f^{(n+1)}(x) \le M, \quad \text{oppure} \quad m \le -f^{(n+1)}(x) \le M, \quad \text{per } x \in [a,b].$$

Si verifichi che

$$\frac{m(b-a)^{n+1}}{2^{2n+1}\,(n+1)!} \le r_n^* \le \frac{M(b-a)^{n+1}}{2^{2n+1}\,(n+1)!}\,.$$

(Traccia: a) si proceda in modo analogo a quanto fatto nella dimostrazione del teorema 6.19. Il polinomio

$$p_n(x) - \frac{1}{2^{n-1}}\,T_n(x),$$

di grado al più $n-1$ dovrebbe annullarsi in n punti distinti. b) Si trasformi l'intervallo $[a,b]$ in $[-1,1]$, ponendo $x = \frac{1}{2}\big[(b-a)y + (a+b)\big]$, e sia $g(y) = f(x(y))$. Indicato con $q_n^*(y)$ il polinomio di approssimazione minimax di $g(y)$ su $[-1,1]$, per il teorema 6.30 esistono almeno $n+1$ punti $\xi_0, \ldots, \xi_n \in [-1,1]$ tali che $g(\xi_i) = q_n^*(\xi_i)$. Per il teorema 5.5 è

$$g(y) - q_n^*(y) = s(y)\,\frac{g^{(n+1)}(\xi)}{(n+1)!}\,, \quad s(y) = (y-\xi_0)\ldots(y-\xi_n), \quad \xi \in (-1,1).$$

Si verifichi che se $f^{(n+1)}(x) \geq 0$ è

$$m\left(\frac{b-a}{2}\right)^{n+1} \leq g^{(n+1)}(\xi) \leq M\left(\frac{b-a}{2}\right)^{n+1},$$

e si applichi il teorema 6.19 al polinomio $s(y)$ per ottenere la limitazione inferiore. Per la limitazione superiore si noti che se fosse

$$\|g - q_n^*\|_\infty > \frac{1}{2^n\,(n+1)!} \max_{x\in[-1,1]} |g^{(n+1)}(x)|,$$

il polinomio $s(y)$ avrebbe in $[-1,1]$ $n+2$ punti distinti di oscillazione, in cui assumerebbe valori di segno alterno e modulo maggiore di $\frac{1}{2^n}$, ciò che è assurdo per a).)

6.43 Si approssimi la funzione $f(x) = \dfrac{1}{x+3}$ sull'intervallo $[-1,1]$ con un polinomio $p(x)$ di grado minimo in modo che

$$\max_{|x|\leq 1} |f(x) - p(x)| \leq 0.5\ 10^{-2}.$$

(Traccia: si verifichi che il polinomio di grado 1 di approssimazione minimax non soddisfa la limitazione. Il polinomio di grado 2, $p_2^*(x) = a_2^* x^2 + a_1^* x + a_0^*$ si ottiene imponendo che vi siano 4 punti di equioscillazione $-1 = x_0^* < x_1^* < x_2^* < x_3^* = 1$ e risolvendo il sistema

$$\begin{cases} \dfrac{1}{2} - a_2 + a_1 - a_0 = d \\ \dfrac{1}{x_1+3} - a_2 x_1^2 - a_1 x_1 - a_0 = -d \\ \dfrac{1}{x_2+3} - a_2 x_2^2 - a_1 x_2 - a_0 = d \\ \dfrac{1}{4} - a_2 - a_1 - a_0 = -d \\ -\dfrac{1}{(x_1+3)^2} - 2a_2 x_1 - a_1 = 0 \\ -\dfrac{1}{(x_2+3)^2} - 2a_2 x_2 - a_1 = 0. \end{cases}$$

Si ottiene

$$a_2^* = \frac{3-2\sqrt{2}}{4},\ a_1^* = \frac{4-3\sqrt{2}}{2},\ a_0^* = \frac{4\sqrt{2}-3}{8},$$
$$x_1^* = \sqrt{2}-2,\ x_2^* = \sqrt{2}-1,\ |d| = \frac{17-12\sqrt{2}}{8} < 0.5\ 10^{-2}.\ )$$

6.44 Si approssimi la funzione

$$f(x) = \frac{1}{x} \int_0^x \frac{1 - \mathrm{e}^{-t^2}}{t^2}\, dt$$

sull'intervallo $[-1, 1]$ con un polinomio $p(x)$ di grado minimo in modo che

$$\max_{|x|\le 1} |f(x) - p(x)| \le 0.5\ 10^{-2}.$$

(Traccia: la funzione $f(x)$ è simmetrica, quindi anche il polinomio cercato è pari. Si tenga inoltre conto che il resto dei polinomi pari non è standard, per cui vi saranno $n + 3$ punti di equioscillazione. Si verifichi che il polinomio di grado 0 di approssimazione minimax non soddisfa la limitazione. Il polinomio di grado 2, $p_2^*(x) = a_2^* x^2 + a_0^*$ si ottiene imponendo che vi siano 5 punti di equioscillazione $-1 = x_0^* < x_1^* < x_2^* = 0 < x_3^* < x_4^* = 1$, con $x_1^* = -x_3^*$, e risolvendo il sistema (si noti che $f(0) = 1$)

$$\begin{cases} 1 - a_0 = d \\ \dfrac{1}{x_3} \displaystyle\int_0^{x_3} \frac{1 - \mathrm{e}^{-t^2}}{t^2}\, dt - a_2 x_3^2 - a_0 = -d \\ \displaystyle\int_0^1 \frac{1 - \mathrm{e}^{-t^2}}{t^2}\, dt - a_2 - a_0 = d \\ -\dfrac{1}{x_3^2} \displaystyle\int_0^{x_3} \frac{1 - \mathrm{e}^{-t^2}}{t^2}\, dt + \frac{1 - \mathrm{e}^{-x_3^2}}{x_3^3} - 2a_2 x_3 = 0. \end{cases}$$

Si approssimino gli integrali con la formula dei trapezi (si veda il capitolo 7). Dalla prima e terza equazione si ricavi a_2, si sostituisca nell'ultima equazione, che va risolta con un metodo iterativo. Si ottiene

$$a_2^* = -0.1384723,\ a_0^* = 0.9967672,\ x_3^* = -x_1^* = 0.6918517,$$
$$|d| = 0.0032328 < 0.5\ 10^{-2}.\ )$$

6.45 Si costruiscano con il metodo di Remez i polinomi di grado $n = 2, 3, 4$ di approssimazione minimax per la funzione $f(x) = \dfrac{1}{x^2}$ sull'intervallo $[1, 2]$.

(Risposta: per $n = 2$ risulta

$$p_2^*(x) = 0.7573984\, x^2 - 2.979224\, x + 3.20034, \quad r_2^* \approx 0.215\ 10^{-1};$$

per $n = 3$ risulta

$$p_3^*(x) = -0.6863154\,x^3 + 3.809156\,x^2 - 7.373259\,x + 5.245841,$$
$$r_3^* \approx 0.458\ 10^{-2};$$

per $n = 4$ risulta

$$p_4^*(x) = 0.5852931\,x^4 - 4.168019\,x^3 + 11.42908\,x^2 - 14.63864x + 7.791341,$$
$$r_4^* \approx 0.945\ 10^{-3}.\ )$$

6.46 Sia $f(x) \in C[a,b]$ e siano $x_0, \ldots, x_{n+1}$, $n+2$ punti distinti di $[a,b]$ tali che

$$a \le x_0 < x_1 < \ldots < x_{n+1} \le b.$$

Si considerino la $(n+2)$-upla $(a_0, \ldots, a_n, d)$, soluzione del sistema

$$\sum_{j=0}^{n} x_i^j\, a_j + (-1)^i d = f(x_i), \quad \text{per} \quad i = 0, \ldots, n+1, \tag{105}$$

e la $(n+2)$-upla $\lambda_0, \ldots, \lambda_{n+1}$, soluzione del sistema

$$\begin{cases} \displaystyle\sum_{i=0}^{n+1} x_i^j\, \lambda_i = 0, \quad \text{per} \quad j = 0, \ldots, n, & (106) \\ \displaystyle\sum_{i=0}^{n+1} |\lambda_i| = 1, \quad \text{con} \quad \lambda_0 > 0. & (107) \end{cases}$$

Si verifichi che

a) il sistema (105) e il sistema (106), (107) hanno una e una sola soluzione;

b) per ogni polinomio $p(x) \in \mathcal{P}_n$ è

$$\sum_{i=0}^{n+1} \lambda_i\, p(x_i) = 0;$$

c) tutti i λ_i, per $i = 0, \ldots, n+1$, sono non nulli e i loro segni sono alternati, cioè

$$\lambda_i = (-1)^i\, |\lambda_i|, \quad \text{per} \quad i = 0, 1, \ldots, n+1;$$

d) si verifichi che assumendo λ_0 come parametro, la soluzione del sistema (106) è

$$\lambda_s = -\lambda_0 \prod_{\substack{i=1 \\ i \ne s}}^{n} \frac{x_0 - x_i}{x_s - x_i}\,;$$

e) per ogni polinomio $p(x) \in \mathcal{P}_n$ è

$$d = \sum_{i=0}^{n+1} \lambda_i f(x_i) = \sum_{i=0}^{n+1} \lambda_i \big[f(x_i) - p(x_i) \big].$$

(Traccia: a) per il sistema (105) si noti che l'ultima colonna della matrice del sistema è formata da elementi che sono alternativamente uguali a $+1$ e a -1. Pertanto sviluppando il determinante rispetto a questa colonna, esso risulta uguale in modulo alla somma di determinanti di matrici di Vandermonde, che sono positivi perché $x_j > x_i$ per $j > i$ (si veda l'esercizio 5.1); per il sistema (106), (107) si noti che la matrice di (106) è di rango massimo e quindi il sistema (106) ha infinite soluzioni non nulle, che dipendono da un parametro moltiplicativo. Assumendo $\lambda_0 > 0$ come tale parametro, dalla condizione (107) segue che la $(n+2)$-upla λ_i, $i = 0, 1, \ldots, n+1$, è unica.

b) Segue dalla (106).

c) Fissato un indice j, $1 \le j \le n-1$, si consideri il polinomio di grado n

$$q(x) = (x - x_0) \ldots (x - x_{j-1})(x - x_{j+2}) \ldots (x - x_{n+1}).$$

Poiché nell'intervallo (x_{j-1}, x_{j+2}) il polinomio $q(x)$ non si annulla, è

$$\operatorname{sgn}(q(x_j)) = \operatorname{sgn}(q(x_{j+1})).$$

Per la b) è

$$0 = \sum_{i=0}^{n+1} \lambda_i q(x_i) = \lambda_j q(x_j) + \lambda_{j+1} q(x_{j+1}) \tag{108}$$

Per l'arbitrarietà dell'indice j, segue che se un λ_j fosse nullo, lo sarebbero anche λ_{j-1} e λ_{j+1} e quindi tutti i λ_i, contraddicendo la (107). Inoltre dalla (108) segue che λ_j e λ_{j+1} hanno segno opposto.

d) Si scriva il sistema (106) nella forma

$$\sum_{i=1}^{n+1} x_i^j \, \lambda_i = -x_0^j \, \lambda_0, \quad j = 0, \ldots, n,$$

e si utilizzi il metodo di Cramer; per esprimere i determinanti che si ottengono si veda l'esercizio 5.1.

e) Posto

$$s(x) = \sum_{j=0}^{n} a_j x^j,$$

per la (105) è

$$f(x_i) - s(x_i) = (-1)^i\, d,$$

e dalla (107) e da c) segue che

$$\sum_{i=0}^{n+1} \lambda_i [f(x_i) - s(x_i)] = \sum_{i=0}^{n+1} (-1)^i\, \lambda_i d = \sum_{i=0}^{n+1} |\lambda_i| d = d.$$

Si tenga poi conto della b).)

6.47 Sia $f(x) \in C[a,b]$ e sia $\{\mathbf{x}^{(k)}\}$ la successione di vettori generata applicando l'algoritmo di Remez ad un vettore $\mathbf{x}^{(0)}$ arbitrario. Riferendosi alle notazioni usate nel paragrafo 6, e considerando inoltre alla k-esima iterazione la $(n+2)$-upla $\lambda_i^{(k)}$, $i = 0, 1, \ldots, n+1$, definita dal sistema

$$\begin{cases} \displaystyle\sum_{i=0}^{n+1} \lambda_i^{(k)} \big(x_i^{(k)}\big)^j = 0, \quad \text{per} \quad j = 0, 1, \ldots, n, \\ \displaystyle\sum_{i=0}^{n+1} |\lambda_i^{(k)}| = 1, \end{cases}$$

si dimostri che se esiste $\delta > 0$ tale che

$$|d^{(k)}| \geq \delta, \quad \text{per ogni} \quad k, \tag{109}$$

allora

a) esiste $\xi > 0$ tale che $|x_{i+1}^{(k)} - x_i^{(k)}| \geq \xi$, per ogni i e k;

b) esiste $\gamma > 0$ tale che $|\lambda_i^{(k)}| \geq \gamma$, per ogni i e k;

c) $|d^{(k+1)}| \geq |d^{(k)}|$, per ogni k;

d) $|d^{(k)}| \leq r^*$, per ogni k;

e) per ogni ϵ esiste un k_0 tale che per $k \geq k_0$ è

$$r^* \leq \|f - p_n^{(k)}\|_\infty \leq r^* + \epsilon; \tag{110}$$

f) per $k \to \infty$ la successione $\{p_n^{(k)}(x)\}$ converge uniformemente su $[a,b]$ a $p_n^*(x)$.

(Traccia: a) si supponga per assurdo che esista un indice j, $0 \leq j \leq n$, per cui

$$\min_{k\to\infty} \lim (x_{j+1}^{(k)} - x_j^{(k)}) = 0,$$

e si consideri il polinomio $q_n(x) \in \mathcal{P}_n$, tale che

$$q_n(x_i^{(k)}) = f(x_i^{(k)}), \quad \text{per} \quad i = 0, \ldots, n+1, \ i \neq j.$$

Per il punto f) dell'esercizio precedente si avrebbe

$$\begin{aligned} |d^{(k)}| &= \Big| \sum_{i=0}^{n+1} \lambda_i^{(k)} [f(x_i^{(k)}) - q_n(x_i^{(k)})] \Big| = |\lambda_j^{(k)}| \ |f(x_j^{(k)}) - q_n(x_j^{(k)})| \\ &\leq |\lambda_j^{(k)}| \ \Big[|f(x_j^{(k)}) - f(x_{j+1}^{(k)})| + |q_n(x_{j+1}^{(k)}) - q_n(x_j^{(k)})| \Big]. \end{aligned}$$

Per continuità, per ogni ϵ esisterebbe un k per cui

$$|f(x_j^{(k)}) - f(x_{j+1}^{(k)})| \leq \epsilon \quad \text{e} \quad |q_n(x_{j+1}^{(k)}) - q_n(x_j^{(k)})| \leq \epsilon,$$

e quindi, poiché $|\lambda_j^{(k)}| < 1$, risulterebbe

$$|d^{(k)}| \leq 2\epsilon |\lambda_j^{(k)}| < 2\epsilon$$

in contrasto con l'ipotesi (109).
b) Segue dal punto a) e dal punto d) dell'esercizio precedente.
c) Alla $(k+1)$-esima iterazione, poiché $x_i^{(k+1)} = y_i$, $i = 0, 1, \ldots, n+1$, per il punto e) dell'esercizio precedente è

$$\begin{aligned} d^{(k+1)} &= \sum_{i=0}^{n+1} \lambda_i^{(k+1)} [f(y_i) - p_n^{(k)}(y_i)] = \sum_{i=0}^{n+1} \lambda_i^{(k+1)} r^{(k)}(y_i) \\ &= \sum_{i=0}^{n+1} \lambda_i^{(k+1)} |r^{(k)}(y_i)| \operatorname{sgn}(r^{(k)}(y_i)). \end{aligned}$$

Poiché sia i $\lambda_i^{(k+1)}$ (per il punto c) dell'esercizio precedente) che gli $r^{(k)}(y_i)$ hanno segno alternato, i fattori

$$\lambda_i^{(k+1)} \operatorname{sgn}(r^{(k)}(y_i))$$

hanno tutti lo stesso segno e quindi

$$|d^{(k+1)}| = \sum_{i=0}^{n+1} |\lambda_i^{(k+1)}| \ |r^{(k)}(y_i)|. \tag{111}$$

Poiché

$$|r^{(k)}(y_i)| \geq |r^{(k)}(x_i^{(k)})| = |d^{(k)}|, \quad \text{per} \quad i = 0, \ldots, n+1,$$

risulta

$$|d^{(k+1)}| \geq |d^{(k)}|.$$

d) Si utilizzi il teorema 6.28 per ogni k.

e) Da c) e d) segue che la successione $\{|d^{(k)}|\}$ è monotona non decrescente, limitata da r^*, e quindi convergente. Posto

$$|d^{(k+1)}| - |d^{(k)}| = \epsilon^{(k)} \geq 0, \tag{112}$$

risulta

$$\lim_{k\to\infty} \epsilon^{(k)} = 0.$$

Dalla (111) si ha

$$|d^{(k+1)}| - |d^{(k)}| = \sum_{i=0}^{n+1} |\lambda_i^{(k+1)}|\ |r^{(k)}(y_i)| - |d^{(k)}|,$$

e poiché

$$|d^{(k)}| = \sum_{i=0}^{n+1} |\lambda_i^{(k+1)}|\ |d^{(k)}|,$$

si ha

$$|d^{(k+1)}| - |d^{(k)}| = \sum_{i=0}^{n+1} |\lambda_i^{(k+1)}| \Big[|r^{(k)}(y_i)| - |d^{(k)}| \Big].$$

Poiché i punti y_i sono tutti punti di massimo o minimo, esiste un indice j (dipendente da k) tale che

$$|r^{(k)}(y_j)| = \|r^{(k)}\|_\infty,$$

per cui, posto $\Lambda^{(k)} = |\lambda_j^{(k+1)}|$, è

$$|d^{(k+1)}| - |d^{(k)}| \geq \Lambda^{(k)} \Big[\|r^{(k)}\|_\infty - |d^{(k)}| \Big],$$

e dalla (112) segue

$$\|r^{(k)}\|_\infty \leq |d^{(k)}| + \frac{\epsilon^{(k)}}{\Lambda^{(k)}} \leq r^* + \frac{\epsilon^{(k)}}{\Lambda^{(k)}}.$$

D'altra parte

$$r^* \leq \|r^{(k)}\|_\infty,$$

da cui

$$r^* \leq \|f - p_n^{(k)}\|_\infty \leq r^* + \frac{\epsilon^{(k)}}{\Lambda^{(k)}}.$$

La e) segue dal fatto che i $\lambda_i^{(k)}$ sono limitati inferiormente in modulo.
f) Si dimostra che

$$\lim_{k\to\infty} \|p_n^{(k)} - p_n^*\|_\infty = 0.$$

Infatti, se ciò non fosse, esisterebbe una sottosuccessione di polinomi $\{p_n^{(k_m)}(x)\}$ di $\{p_n^{(k)}(x)\}$ tale che

$$\|p_n^{(k_m)} - p_n^*\|_\infty \geq M, \quad M > 0. \tag{113}$$

Per il teorema di Bolzano-Weierstrass, la successione $\{p_n^{(k_m)}(x)\}$ ammette una sottosuccessione uniformemente convergente $\{p_n^{(k_q)}(x)\}$. Indicato con $q(x)$ il limite di tale sottosuccessione, si ha dalla (110) che

$$\|f - q\|_\infty = r^*.$$

Ma dalla (113) segue che $q(x) \neq p_n^*(x)$, ciò che contraddice la proprietà di unicità del polinomio di migliore approssimazione.)

6.48 Sia $f(x) \in C^2[a,b]$ e sia $r^*(x)$ standard (quindi ad ogni passo del metodo di Remez è $x_0^{(k)} = a$ e $x_{n+1}^{(k)} = b$). Si suppone inoltre che

$$\left.\frac{d^2 r^*(x)}{dx^2}\right|_{x = x_i^*} \neq 0, \quad \text{per} \quad i = 0, 1, \ldots, n+1.$$

a) Si verifichi che esiste un intorno U del punto $(a_1^*, \ldots, a_n^*)$ ed esistono n funzioni

$$x_i = \phi_i(a_1, \ldots, a_n), \quad \text{per} \quad i = 1, \ldots, n,$$

definite in U, tali che

$$x_i^* = \phi_i(a_1^*, \ldots, a_n^*), \quad \text{per} \quad i = 1, \ldots, n,$$

e per cui valgono le relazioni

$$f'(x_i) - \sum_{j=1}^{n} a_j \; j \; x_i^{j-1} = 0, \quad i = 1, \ldots, n, \tag{114}$$

per ogni $(a_1, \ldots, a_n) \in U$.
b) Si consideri il sistema non lineare

$$F_i(a_0, a_1, \ldots, a_n, d) = f(x_i) - \sum_{j=0}^{n} a_j x_i^j - (-1)^i d = 0, \; i = 0, \ldots, n+1, \tag{115}$$

in cui le $x_i = \phi_i(a_1, \ldots, a_n)$ sono le funzioni definite in U implicitamente dalle (114), e si scriva il metodo iterativo di Newton-Raphson applicato al sistema (115).

c) Si verifichi che la successione $\{a_0^{(k)}, \ldots, a_n^{(k)}\}$ generata con il metodo di Remez coincide con quella generata dal metodo di Newton-Raphson applicato al sistema (115). Pertanto l'ordine di convergenza della successione $\{a_0^{(k)}, \ldots, a_n^{(k)}\}$ ai coefficienti $a_0^*, \ldots, a_n^*$ di $p_n^*(x)$ è due.

(Traccia: a) i punti x_i^*, $i = 1, \ldots, n$, sono stazionari per

$$r^*(x) = f(x) - \sum_{j=0}^{n} a_j^* \, x^j,$$

quindi risulta

$$f'(x_i^*) - \sum_{j=1}^{n} a_j^* \, j \, \left(x_i^*\right)^{j-1} = 0, \quad i = 1, \ldots, n.$$

Poiché $(r^*(x))''|_{x = x_i^*} \neq 0$, l'esistenza dell'intorno U e delle funzioni $x_i = \phi_i(a_1, \ldots, a_n)$ segue dal teorema del Dini applicato a $(r^*(x))'$.
b) Applicando il metodo di Newton-Raphson al sistema (115) si ha per la k-esima iterazione

$$\sum_{m=0}^{n} \frac{\partial F_i}{\partial a_m}\bigg|_{\mathbf{a}^{(k-1)}} \left[a_m^{(k)} - a_m^{(k-1)}\right] + \frac{\partial F_i}{\partial d}\bigg|_{\mathbf{a}^{(k-1)}} \left[d^{(k)} - d^{(k-1)}\right] = -F_i\Big|_{\mathbf{a}^{(k-1)}}, \tag{116}$$

dove

$$\mathbf{a}^{(k-1)} = (a_0^{(k-1)}, \ldots, a_n^{(k-1)}, d^{(k-1)}),$$

e si indichi

$$x_i^{(k)} = \phi_i(a_1^{(k-1)}, \ldots, a_n^{(k-1)}), \quad \text{per} \quad i = 1, \ldots, n.$$

Dalla (114) risulta quindi che

$$\sum_{j=1}^{n} a_j^{(k-1)} j \left(x_i^{(k)}\right)^{j-1} = f'(x_i^{(k)}), \quad i = 1, \ldots, n. \tag{117}$$

Poiché per $m = 1, \ldots, n$ e $i = 1, \ldots, n$, è

$$\frac{\partial F_i}{\partial a_m}\bigg|_{\mathbf{a}^{(k-1)}} = -\left(x_i^{(k)}\right)^m + \left[f'(x_i^{(k)}) - \sum_{j=1}^{n} a_j^{(k-1)} j \left(x_i^{(k)}\right)^{j-1}\right] \frac{\partial \phi_i}{\partial a_m}\bigg|_{\mathbf{a}^{(k-1)}},$$

dalla (117) segue che

$$\frac{\partial F_i}{\partial a_m}\bigg|_{\mathbf{a}^{(k-1)}} = -\big(x_i^{(k)}\big)^m.$$

Si verifichi che tale relazione vale anche per $i = 0, n+1$ e $m = 0$. Quindi la (116) diventa

$$-\sum_{m=0}^{n}\big(x_i^{(k)}\big)^m\big[a_m^{(k)} - a_m^{(k-1)}\big] - (-1)^i\big[d^{(k)} - d^{(k-1)}\big]$$
$$= -\left[f(x_i^{(k)}) - \sum_{j=0}^{n} a_j^{(k-1)}\big(x_i^{(k)}\big)^j - (-1)^i d^{(k-1)}\right].$$

Semplificando si ottiene

$$\sum_{j=0}^{n} a_j^{(k)}\big(x_i^{(k)}\big)^j + (-1)^i d^{(k)} = f(x_i^{(k)}),\ i = 0, \ldots, n+1. \tag{118}$$

c) I punti $x_i^{(k)}$ definiti implicitamente in (117) sono i punti stazionari di $r^{(k-1)}(x)$, mentre la (118) coincide con la (48). Per il teorema 6.38 il metodo di Remez è convergente, quindi esiste un indice k tale che

$$(a_1^{(k-1)}, \ldots, a_n^{(k-1)}) \in U,$$

per il quale le (117) sono esplicitabili.)

6.49 Si verifichi direttamente mediante la equioscillazione del resto che il polinomio di grado 2 di approssimazione minimax di

$$f(x) = T_0(x) + T_1(x) + T_2(x) + T_3(x), \quad -1 \le x \le 1,$$

è

$$p_2(x) = T_0(x) + T_1(x) + T_2(x),$$

mentre il polinomio di grado 1 di approssimazione minimax *non* è

$$p_1(x) = T_0(x) + T_1(x).$$

(Traccia: è $r_2(x) = f(x) - p_2(x) = T_3(x)$ e

$$T_3(-1) = -T_3\big(-\frac{1}{2}\big) = T_3\big(\frac{1}{2}\big) = -T_3(1) = -1;$$

è $\quad r_1(x) = f(x) - p_1(x) = T_2(x) + T_3(x) = 4x^3 + 2x^2 - 3x - 1$ e

$$r_1'(x) = 0 \text{ in } x_{1,2} = -\frac{1}{6} \pm \frac{\sqrt{10}}{6}\,,$$

ma

$$r_1(-1) = 0, \quad r_1(x_1) \neq -r_1(x_2) \neq r_1(1).\ )$$

6.50 Si approssimi nell'intervallo $[0,1]$ il polinomio $p(x) = 1 - x + x^2 - x^3 + x^4 - x^5 + x^6$ con un polinomio $q(x)$ di grado minore, tale che

$$\|p - q\|_\infty < 0.5\ 10^{-2}.$$

Si verifichi che il grado minimo di tale polinomio è 4.

(Traccia: si faccia prima la trasformazione di variabile $x = \frac{1}{2}(y+1)$ ottenendo

$$p(y) = \frac{1}{64}\left[y^6 + 4y^5 + 9y^4 + 8y^3 + 11y^2 - 12y + 43\right].$$

Esprimendo $p(y)$ come combinazione di polinomi di Chebyshev di 1^a specie, si ha

$$p(y) = \frac{1}{2048}\left[T_6(y) + 8T_5(y) + 42T_4(y) + 104T_3(y) + 335T_2(y) - 112T_1(y) + 1670T_0(y)\right].$$

Il polinomio di approssimazione minimax di 3^o grado non soddisfa la condizione data, mentre il polinomio

$$q(y) = \frac{1}{2048}\left[42T_4(y) + 104T_3(y) + 335T_2(y) - 112T_1(y) + 1670T_0(y)\right],$$

pur non essendo il polinomio di approssimazione minimax di quarto grado, è tale che

$$\max_{y\in[-1,1]} |p(y) - q(y)| \approx 0.439\ 10^{-2},$$

per cui il polinomio cercato è

$$q(x) = 2.625\,x^4 - 3.625\,x^3 + 2.152344\,x^2 - 1.160156\,x + 1.003418.\ )$$

6.51 Si confrontino i massimi errori assoluti generati nelle due seguenti approssimazioni di

$$f(x) = \frac{\arctan x}{x}\,, \quad x \in [-1,1]:$$

a) serie di Maclaurin fino al termine di grado 4;

b) economizzazione al termine di grado 4 della serie di Maclaurin scritta fino al termine di grado 8.

(Traccia: a) si ha

$$p_4(x) = 1 - \frac{x^2}{3} + \frac{x^4}{5} \quad \text{e} \quad \max_{x\in[-1,1]} |f(x) - p_4(x)| \approx 0.813\ 10^{-1};$$

b) si ha

$$q_4(x) = \frac{8077}{8064} - \frac{353x^2}{1008} + \frac{227x^4}{1260} \quad \text{e} \quad \max_{x\in[-1,1]} |f(x) - q_4(x)| \approx 0.462\ 10^{-1}.\ )$$

6.52 Siano $f(x) = \arcsin x$ e $p_n^*(x)$ l'approssimazione minimax di grado al più n di $f(x)$ su $[-1, 1]$. Si verifichi che asintoticamente per $n \to \infty$ è

$$\|f - p_n^*\|_\infty \geq \gamma_n, \quad \text{dove} \quad \gamma_n \sim \frac{\pi}{2}\,\frac{1}{n \log n}\,.$$

(Traccia: dall'esercizio 6.27 risulta che

$$f(x) - p_{2n-1}^C(x) = \frac{4}{\pi}\sum_{i=n}^{\infty} \frac{T_{2i+1}(x)}{(2i+1)^2}\,,$$

dove $p_{2n-1}^C(x)$ è il polinomio di grado $2n-1$ di approssimazione ai minimi quadrati di $f(x)$, usando i polinomi di Chebyshev di 1^a specie. Quindi

$$\|f - p_{2n-1}^C\|_\infty = f(1) - p_{2n-1}^C(1) = \frac{4}{\pi}\sum_{i=n}^{\infty} \frac{1}{(2i+1)^2}\,.$$

Poiché

$$\frac{1}{2}\int_{2n+1}^{\infty} \frac{dx}{x^2} < \sum_{i=n}^{\infty} \frac{1}{(2i+1)^2} < \frac{1}{2}\int_{2n-1}^{\infty} \frac{dx}{x^2}\,,$$

risulta

$$\frac{1}{2(2n+1)} < \sum_{i=n}^{\infty} \frac{1}{(2i+1)^2} < \frac{1}{2(2n-1)}\,.$$

Quindi

$$\frac{2}{\pi(n+2)} < \|f - p_n^C\|_\infty < \frac{2}{\pi n}\,,$$

e asintoticamente

$$\|f - p_n^C\|_\infty \sim \frac{2}{\pi n}.$$

Si applichi poi il teorema 6.42.)

6.53 Sia $f(x) \in C[a,b]$ e sia $p_n^*(x)$ il polinomio di grado al più n di approssimazione minimax di $f(x)$ su $[a,b]$. Si verifichi che

$$\lim_{n\to\infty} \|f - p_n^*\|_2 = 0,$$

in cui $\| \ . \ \|_2$ è la norma 2 rispetto ad una qualunque funzione peso $\omega(x)$ sull'intervallo $[a,b]$.

(Traccia: si ha

$$\|f - p_n^*\|_2^2 = \int_a^b \omega(x)\big[f(x) - p_n^*(x)\big]^2\, dx \le \big(r_n^*\big)^2 \int_a^b \omega(x)\, dx,$$

e si applichi il teorema 6.34.)

6.54 Sia $f(x) \in C[-1,1]$ e sia $p_n(x)$ il polinomio di interpolazione di $f(x)$ nei nodi di Chebyshev

$$x_i = \cos\theta_i, \quad \theta_i = \frac{(2i+1)\pi}{2(n+1)}, \quad i = 0, \dots, n,$$

cioè

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x),$$

dove per la (8, cap. 5) è

$$L_i(x) = \frac{T_{n+1}(x)}{(x - x_i) T'_{n+1}(x_i)}.$$

Si verifichi che, posto $\omega(x) = (1-x^2)^{-1/2}$, è

a) $$(L_i, L_j) = \int_{-1}^{1} \omega(x) L_i(x) L_j(x)\, dx = 0, \quad \text{per} \quad i \neq j;$$

b) $$\sum_{i=0}^{n} \|L_i\|_2^2 = \sum_{i=0}^{n} \int_{-1}^{1} \omega(x) L_i^2(x)\, dx = \int_{-1}^{1} \omega(x)\, dx = \pi;$$

c) $$\lim_{n\to\infty} \|f - p_n\|_2^2 = \lim_{n\to\infty} \int_{-1}^{1} \omega(x)\big[f(x) - p_n(x)\big]^2\, dx = 0.$$

Tale proprietà di convergenza può non valere per la norma ∞.

(Traccia: a) è

$$\int_{-1}^{1} \omega(x) L_i(x) L_j(x)\, dx$$
$$= \frac{1}{T'_{n+1}(x_i) T'_{n+1}(x_j)} \int_{-1}^{1} \omega(x) T_{n+1}(x)\, \frac{T_{n+1}(x)}{(x - x_i)\,(x - x_j)}\, dx$$

e la funzione $\dfrac{T_{n+1}(x)}{(x - x_i)\,(x - x_j)}$ è un polinomio di grado minore di $n + 1$.
b) Per l'esercizio 5.9 è

$$\left[\sum_{i=0}^{n} L_i(x) \right]^2 = 1,$$

e quindi

$$\int_{-1}^{1} \omega(x) \left[\sum_{i=0}^{n} L_i(x) \right]^2 dx = \int_{-1}^{1} \omega(x)\, dx,$$

si sfrutti poi la proprietà di ortogonalità del punto a).
c) Indicato con $p_n^*(x)$ il polinomio di approssimazione minimax di grado n della $f(x)$ su $[-1, 1]$, è

$$p_n^*(x) = \sum_{i=0}^{n} p_n^*(x_i) L_i(x),$$

quindi per il punto a) risulta

$$\|p_n - p_n^*\|_2^2 = \sum_{i=0}^{n} \left[f(x_i) - p_n^*(x_i) \right]^2 \|L_i(x)\|_2^2 \leq \left(r_n^* \right)^2 \sum_{i=0}^{n} \|L_i(x)\|_2^2.$$

Dal punto b) e dal teorema 6.34 segue che

$$\lim_{n \to \infty} \|p_n - p_n^*\|_2^2 = 0.$$

Si applichi poi l'esercizio 6.53 alla relazione

$$\|f - p_n\|_2 \leq \|p_n - p_n^*\|_2 + \|f - p_n^*\|_2. \ )$$

6.55 a) Sia $\sum_{k=0}^{\infty} \alpha_k$ convergente, con $\alpha_k > 0$ per ogni k, e si consideri la funzione

$$f(x) = \sum_{k=0}^{\infty} \alpha_k T_{3^k}(x), \quad x \in [-1, 1].$$

Si verifichi che per ogni intero n l'approssimazione minimax di grado al più 3^n di $f(x)$ per $x \in [-1,1]$ è data da

$$g_{3^n}(x) = \sum_{k=0}^{n} \alpha_k T_{3^k}(x).$$

b) Si dimostri che per ogni successione $\{\epsilon_n\}$ monotona, decrescente e convergente a zero, esiste una funzione $f(x) \in C[-1,1]$ tale che $d_n(f) \geq \epsilon_n$, dove $d_n(f)$ è la *distanza* di $f(x)$ dal sottospazio dei polinomi di grado al più n (per la definizione di distanza si veda il paragrafo 1).

(Traccia: a) si verifichi che i punti

$$x_i = \cos \frac{i\pi}{3^{n+1}}, \quad i = 0, \ldots, 3^{n+1},$$

sono di equioscillazione per la funzione

$$r_{3^n}(x) = f(x) - g_{3^n}(x) = \sum_{k=n+1}^{\infty} \alpha_k T_{3^k}(x).$$

Si ha infatti per $k \geq n+1$ che $T_{3^k}(x_i) = (-1)^i$, e quindi

$$r_{3^n}(x_i) = (-1)^i \delta, \quad i = 0, \ldots, 3^{n+1}, \quad \text{dove} \quad \delta = \sum_{k=n+1}^{\infty} \alpha_k.$$

D'altra parte, poiché $|T_i(x)| \leq 1$ per $x \in [-1,1]$ e per ogni i, risulta

$$|r_{3^n}(x)| \leq \sum_{k=n+1}^{\infty} \alpha_k = \delta.$$

Quindi la funzione $r_{3^n}(x)$ ha $3^{n+1}+1$ punti di equioscillazione, e $3^{n+1}+1 > 3^n+2$ per $n \geq 1$.
b) Si ponga $\alpha_k = \epsilon_k - \epsilon_{k+1} > 0$, $k = 0, 1, \ldots$ Poiché $\sum_{k=0}^{n} \alpha_k = \epsilon_0 - \epsilon_{n+1}$, la serie $\sum_{k=0}^{\infty} \alpha_k$ è convergente. Una funzione che soddisfa la condizione posta è la funzione $f(x)$ definita in a). Si ha infatti $\delta = \sum_{k=n+1}^{\infty} \alpha_k = \epsilon_{n+1}$, e quindi $d_{3^n}(f) = \epsilon_{n+1}$. Poiché $d_{n+1}(f) \geq d_{3^n}(f)$, ne segue che $d_{n+1}(f) \geq \epsilon_{n+1}$.)

6.56 Sia $f(x) \in C[-1,1]$ e per ogni n siano

$$x_i = \cos \frac{(2i+1)\pi}{2(n+1)}, \quad i = 0, \ldots, n,$$

i nodi di Chebyshev di 1^a specie.

a) Si scriva il polinomio $p_{2n+1}(x)$ di osculazione di Hermite di grado $2n+1$ tale che

$$p_{2n+1}(x_i) = f(x_i), \quad p'_{2n+1}(x_i) = 0, \quad \text{per} \quad i = 0, \dots, n.$$

b) Si verifichi che

$$\lim_{n\to\infty} |f(x) - p_{2n+1}(x)| = 0 \quad \text{uniformemente per } x \in [-1, 1].$$

Dal punto b) segue che ogni funzione continua può essere approssimata uniformemente su un intervallo con una successione di polinomi. Questa proprietà rappresenta anche una dimostrazione alternativa del teorema di Weierstrass 6.1.

(Traccia: a) per la (17, cap. 5) si ha

$$p_{2n+1}(x) = \sum_{j=0}^{n} U_j(x) f(x_j),$$

dove

$$U_j(x) = \big[1 - 2L'_j(x_j)(x - x_j)\big] L_j^2(x),$$
$$L_j(x) = \frac{\pi_n(x)}{\pi'_n(x_j)(x - x_j)}, \quad \pi_n(x) = (x - x_0) \dots (x - x_n).$$

Poiché $\pi_n(x)$ è monico, risulta $\pi_n(x) = \dfrac{T_{n+1}(x)}{2^n}$. Si verifichi che

$$L'_j(x_j) = \frac{\pi''_n(x_j)}{2\pi'_n(x_j)},$$

e quindi

$$2L'_j(xj) = \frac{T''_{n+1}(x_j)}{T'_{n+1}(x_j)}.$$

D'altra parte, per la o) dell'esercizio 6.17 è

$$(1 - x^2)T''_{n+1}(x) - xT'_{n+1}(x) + (n+1)^2 T_{n+1}(x) = 0$$

e, tenendo conto che $T_{n+1}(x_j) = 0$, si ha

$$\frac{T''_{n+1}(x_j)}{T'_{n+1}(x_j)} = \frac{x_j}{1 - x_j^2},$$

da cui

$$1 - 2L_j'(x_j)(x - x_j) = \frac{1 - xx_j}{1 - x_j^2}\,.$$

Per la (29) è

$$T_{n+1}'(x) = (n+1)\,\frac{\sin(n+1)\theta}{\sin\theta}\,, \quad \cos\theta = x,$$

per cui

$$T_{n+1}'(x_j) = \frac{n+1}{\sqrt{1 - x_j^2}}\,\sin\,\frac{(2j+1)\pi}{2}\,=\,\frac{(-1)^j(n+1)}{\sqrt{1 - x_j^2}}\,,$$

e

$$L_j^2(x) = \frac{T_{n+1}^2(x)(1 - x_j^2)}{(n+1)^2(x - x_j)^2}\,.$$

Si è così ottenuto

$$p_{2n+1}(x) = \sum_{j=0}^{n} f(x_j)U_j(x), \quad \text{dove} \quad U_j(x) = \frac{T_{n+1}^2(x)(1 - xx_j)}{(n+1)^2(x - x_j)^2}\,.$$

b) Poiché per $f(x) \equiv 1$ è $p_{2n+1}(x) \equiv 1$, si ha

$$\sum_{j=0}^{n} U_j(x) = 1 \quad \text{per ogni } x,$$

e quindi, tenendo conto che $1 - xx_j > 0$ per $x \in [-1, 1]$, risulta $U_j(x) \geq 0$, per cui

$$|f(x) - p_{2n+1}(x)| \leq \sum_{j=0}^{n} |f(x) - f(x_j)|\, U_j(x).$$

Poiché $f(x)$ è continua, per ogni $\epsilon > 0$ esiste $\delta > 0$ tale che per ogni x e y, con $|x - y| < \delta$, è $|f(x) - f(y)| < \epsilon/2$. Sia $x \in [-1, 1]$; per un n fissato, sia J l'insieme degli indici j tali che $|x - x_j| < \delta$, risulta

$$\sum_{j \in J} |f(x) - f(x_j)|\, U_j(x) < \frac{\epsilon}{2}\,\sum_{j=0}^{n} U_j(x) = \frac{\epsilon}{2}\,.$$

Per gli indici $j \notin J$ si ha

$$|x - x_j| \geq \delta, \quad 0 < 1 - xx_j < 2, \quad \left|T_{n+1}(x)\right| \leq 1,$$

per cui

$$U_j(x) < \frac{2}{(n+1)^2\delta^2}\,,$$

ed essendo $|f(x) - f(x_j)| < M$, per un'opportuna costante M, risulta

$$\sum_{j\notin J} |f(x) - f(x_j)|\, U_j(x) \le \sum_{j=0}^{n} \frac{2M}{(n+1)^2\delta^2} = \frac{2M}{(n+1)\delta^2}\,.$$

Scegliendo n tale che $\dfrac{2M}{(n+1)\delta^2} < \dfrac{\epsilon}{2}$, risulta

$$|f(x) - p_{2n+1}(x)| < \epsilon \quad \text{uniformemente per } x \in [-1,1].\ )$$

6.57 Siano $p_n^*(x)$ e $q_n^*(x)$ i polinomi di approssimazione minimax di una funzione $f(x)$ su $[a,b]$ rispetto all'errore assoluto e rispetto all'errore relativo. Si verifichi che se $0 < f(x) \le 1$ per $x \in [a,b]$, allora

$$\max_{x\in[a,b]} \left| \frac{f(x) - q_n^*(x)}{f(x)} \right| \ge \max_{x\in[a,b]} \big|f(x) - p_n^*(x)\big|.$$

(Traccia: si tenga conto del fatto che

$$\max_{x\in[a,b]} \big|f(x) - q_n^*(x)\big| \ge \max_{x\in[a,b]} \big|f(x) - p_n^*(x)\big|.\ )$$

6.58 Sia $f(x)$ tale che $f(0) = 0$, con $a \le 0 \le b$ e siano $p_n^*(x)$ e $q_n^*(x)$ i polinomi di approssimazione minimax su $[a,b]$ rispettivamente senza vincoli e con il vincolo che $q_n^*(0) = 0$. Si verifichi che

$$\max_{x\in[a,b]} \big|f(x) - p_n^*(x)\big| \le \max_{x\in[a,b]} \big|f(x) - q_n^*(x)\big| \le 2 \max_{x\in[a,b]} \big|f(x) - p_n^*(x)\big|.$$

(Traccia: la prima disuguaglianza è ovvia; per la seconda si consideri il polinomio $s_n(x) = p_n^*(x) - p_n^*(0)$, per cui è

$$\big|f(x) - s_n(x)\big| \le \big|f(x) - p_n^*(x)\big| + \big|f(0) - p_n^*(0)\big| \le 2 \max_{x\in[a,b]} \big|f(x) - p_n^*(x)\big|,$$

e si noti che $s_n(0) = 0$.)

6.59 Sia $f(x) \in [a,b]$ una funzione non razionale, tale che per ogni coppia di polinomi $u(x)$ di grado i e $v(x)$ di grado j non identicamente nulli la funzione $u(x) + v(x)f(x)$ abbia al più $i + j + 1$ zeri distinti in $[a,b]$.

a) Si verifichi che per l'approssimazione razionale minimax

$$w^*(x) = \frac{p^*(x)}{q^*(x)} \in \mathcal{R}_{m,n}$$

di $f(x)$ su $[a, b]$ è $m' = m$ e $n' = n$, dove m' e n' sono i gradi effettivi di $p^*(x)$ e di $q^*(x)$.

b) Si verifichi che la funzione $f(x) = \mathrm{e}^x$ soddisfa le ipotesi del punto a) su ogni intervallo $[a, b]$, per cui non si possono presentare casi di degenerazione nell'approssimazione minimax razionale di e^x.

(Traccia: se fosse ad esempio $m' < m$ e $n' = n$, il resto

$$f(x) - \frac{p^*(x)}{q^*(x)}$$

avrebbe almeno $m + n + 2$ punti di equioscillazione e quindi $m + n + 1$ zeri distinti in $[a, b]$. Ma per l'ipotesi la funzione $q^*(x)f(x) - p^*(x)$ non dovrebbe avere più di $m' + n + 1$ zeri distinti in $[a, b]$. Si proceda in modo analogo per i casi $m' = m$, $n' < n$ e $m' < m$, $n' < n$.
b) Si supponga per assurdo che esistano due polinomi $u(x)$ di grado i e $v(x)$ di grado j, tali che la funzione $z(x) = u(x) + v(x)\mathrm{e}^x$ abbia almeno $i + j + 2$ zeri distinti in $[a, b]$. Per il teorema di Rolle la derivata $(i+1)$-esima di $z(x)$ dovrebbe avere almeno $j + 1$ zeri distinti in $[a, b]$. Ma

$$z^{(i+1)}(x) = \frac{d^{i+1}}{dx^{i+1}} \left[v(x)\mathrm{e}^x \right] = \mathrm{e}^x t(x),$$

dove $t(x)$ è un polinomio di grado j.)

6.60 Sia $f(x)$ simmetrica (antisimmetrica) su $[-a, a]$, $a > 0$.

a) Si verifichi che anche le funzioni razionali di approssimazione minimax su $[-a, a]$ sono simmetriche (antisimmetriche).

b) Si dispongano le approssimazioni minimax razionali di $f(x)$ nella tabella

$$\begin{array}{cccc} w^*_{0,0}(x) & w^*_{0,1}(x) & w^*_{0,2}(x) & \dots \\ w^*_{1,0}(x) & w^*_{1,1}(x) & w^*_{1,2}(x) & \dots \\ w^*_{2,0}(x) & w^*_{2,1}(x) & w^*_{2,2}(x) & \dots \\ \dots & \dots & \dots & \end{array}$$

Si verifichi che risulta

$$w^*_{m,n}(x) = w^*_{m,n+1}(x) = w^*_{m+1,n}(x) = w^*_{m+1,n+1}(x)$$

per m, n pari se $f(x)$ è simmetrica,

per m dispari, n pari se $f(x)$ è antisimmetrica,

e quindi nella tabella compaiono dei blocchi quadrati di quattro elementi uguali.

(Traccia: a) si proceda come per l'esercizio 6.39 e si tenga conto del fatto che l'approssimazione razionale è irriducibile e quindi unica; b) si tenga conto del punto a) e si noti che se la funzione è simmetrica (antisimmetrica), il numero dei punti di equioscillazione deve essere dispari (pari).)

6.61 Si verifichi che la frazione continua

$$w = d_0 + \cfrac{1}{d_1 + \cfrac{1}{d_2 + \ldots}}, \qquad d_i \geq 0 \quad \text{per} \quad i \geq 1,$$

converge se e solo se la serie $\displaystyle\sum_{i=1}^{\infty} d_i$ diverge.

(Traccia: sia h il più piccolo indice dispari per cui $d_h \neq 0$; indicata con $w_k = \dfrac{p_k}{q_k}$ la k-esima frazione parziale, si verifichi, sfruttando la (40, cap. 5), che

$$q_k \geq \theta = \min\{1, d_h\} > 0, \quad \text{per ogni } k \geq h,$$

per cui risulta

$$q_k q_{k-1} = (d_k q_{k-1} + q_{k-2}) q_{k-1} = d_k q_{k-1}^2 + q_{k-1} q_{k-2} \geq d_k \theta^2 + q_{k-1} q_{k-2}.$$

Posto $r_k = q_k q_{k-1}$, risulta

$$r_k \geq r_{k-1} + \theta^2 d_k \quad \text{e quindi} \quad r_k \geq \theta^2 \sum_{i=1}^{k} d_i,$$

da cui segue che se la serie $\displaystyle\sum_{i=1}^{\infty} d_i$ diverge, allora $\displaystyle\lim_{k\to\infty} r_k = \infty$. D'altra parte per l'esercizio 5.36 c) è

$$w_k - d_0 = \sum_{i=1}^{k} \frac{(-1)^{i-1}}{q_i q_{i-1}} = \sum_{i=1}^{k} \frac{(-1)^{i-1}}{r_i}.$$

Viceversa si verifichi, sfruttando la (40, cap. 5), che

$$q_k \leq \prod_{i=1}^{k} (1 + d_i),$$

per cui, essendo $1 + x \le \mathrm{e}^x$ per $x \ge 0$, risulta

$$q_k \le \prod_{i=1}^{k} \mathrm{e}^{d_i} = \exp\Big(\sum_{i=1}^{k} d_i\Big).$$

Ne segue che

$$r_k = q_k q_{k-1} \le \mathrm{e}^{2\sigma}, \quad \text{dove} \quad \sigma = \sum_{i=1}^{\infty} d_i.$$

Quindi se la serie converge, gli r_k sono superiormente limitati e la successione $\{w_k\}$ non converge.)

6.62 Si verifichi che

a) la frazione continua

$$w = d_0 + \cfrac{c_1}{d_1 + \cfrac{c_2}{d_2 + \ldots}}, \qquad |d_i| \ge |c_i| + 1 \quad \text{per} \quad i \ge 1,$$

è convergente;

b) per la k-esima frazione parziale vale $|w_k| < 1 + |d_0|$, per $k \ge 1$.

(Traccia: a) indicata con $w_k = \dfrac{p_k}{q_k}$ la k-esima frazione parziale, risulta, per la (40, cap. 5), che

$$\begin{aligned} |q_k| &= |d_k q_{k-1} + c_k q_{k-2}| \ge |d_k|\ |q_{k-1}| - |c_k|\ |q_{k-2}| \\ &\ge \big(|c_k| + 1\big)\,|q_{k-1}| - |c_k|\ |q_{k-2}|, \end{aligned}$$

da cui

$$|q_k| - |q_{k-1}| \ge |c_k|\,\big(|q_{k-1}| - |q_{k-2}|\big) \ge \prod_{i=1}^{k} |c_i|.$$

Quindi la successione dei $|q_k|$ è non decrescente e vale la relazione

$$\frac{\displaystyle\prod_{i=1}^{k} |c_i|}{|q_k q_{k-1}|} \le \frac{|q_k| - |q_{k-1}|}{|q_k q_{k-1}|} = \frac{1}{|q_{k-1}|} - \frac{1}{|q_k|}\,.$$

Ne segue che la serie

$$\sum_{k=1}^{j} \frac{\displaystyle\prod_{i=1}^{k} |c_i|}{|q_k q_{k-1}|} \le \frac{1}{|q_0|} - \frac{1}{|q_j|} = 1 - \frac{1}{|q_j|}$$

è assolutamente convergente, e poiché per l'esercizio 5.36 c) è

$$w_j - d_0 = \sum_{k=1}^{j} \frac{(-1)^{k-1} \displaystyle\prod_{i=1}^{k} |c_i|}{q_k q_{k-1}}\,,$$

anche la successione dei w_j è convergente. b) Posto

$$s_1 = \frac{c_k}{d_k}\,, \quad s_{i+1} = \frac{c_{k-i}}{d_{k-i} + s_i}\,,$$

risulta $w_k = d_0 + s_k$. Si dimostri per induzione che

$$|s_i| < 1 \quad \text{per} \quad i = 1, \ldots, k.)$$

6.63 Sia

$$w = d_0 + \frac{c_1}{d_1} \; + \; \frac{c_2}{d_2} \; + \ldots \quad .$$

a) Si verifichi che per ogni successione di numeri non nulli $\{\alpha_i\}$, con $\alpha_0 = 1$, la frazione continua

$$v = \alpha_0 d_0 + \frac{\alpha_0\alpha_1 c_1}{\alpha_1 d_1} \; + \; \frac{\alpha_1\alpha_2 c_2}{\alpha_2 d_2} \; + \ldots$$

è equivalente alla w (cioè tutte le frazioni parziali di w e v coincidono).

b) Facendo riferimento al punto a), si determini una frazione continua della forma

$$u = e_0 + \frac{1}{e_1} \; + \; \frac{1}{e_2} \; + \ldots$$

equivalente alla w.

c) Si verifichi che se $c_i,\ d_i > 0$ per $i \geq 1$, la w converge se la serie

$$\sum_{i=1}^{\infty} \sqrt{\frac{d_{i-1} d_i}{c_i}}$$

diverge.

d) In particolare si dica per quali valori di k converge la frazione continua

$$\frac{1^k}{\beta} \; + \; \frac{2^k}{\beta} \; + \; \frac{3^k}{\beta} \; + \ldots \quad , \quad \beta > 0.$$

(Traccia: b) si scelgano gli α_i in modo che $\alpha_{i-1}\alpha_i c_i = 1$, $\alpha_0 = 1$. Risulta $e_i = \alpha_i d_i$. c) Si sfrutti il teorema 6.58, notando che se la serie $\sum\limits_{i=1}^{\infty} \sqrt{e_{i-1} e_i}$ diverge, allora la serie $\sum\limits_{i=1}^{\infty} e_i$ diverge. d) La frazione continua è convergente per $k \leq 2$.)

6.64 Sia

$$w = d_0 + \cfrac{1}{d_1 + \cfrac{1}{d_2 + \dots}}, \qquad d_i \geq 0 \quad \text{per} \quad i \geq 1,$$

una frazione continua infinita e sia $w_k = \dfrac{p_k}{q_k}$ la k-esima frazione parziale di w. Se la frazione continua converge ad α, allora

$$\left| \alpha - \frac{p_k}{q_k} \right| \leq \frac{1}{q_k q_{k-1}} .$$

(Traccia: si sfruttino il punto d) e il punto b) dell'esercizio 5.36.)

6.65 Si verifichi che la frazione continua

$$w(x) = x + \cfrac{1}{x + \cfrac{1}{x + \dots}}$$

converge per ogni numero reale $x \neq 0$ e se ne dia il limite. Si deduca, come caso particolare, che

$$1 + \cfrac{1}{1 + \cfrac{1}{1 + \dots}} = \frac{1+\sqrt{5}}{2} \quad (\textit{sezione aurea}).$$

(Traccia: posto $x_0 = x$ e $x_{i+1} = x + \dfrac{1}{x_i}$, si verifichi che il metodo iterativo converge per ogni x a

$$\frac{x + \operatorname{sgn}(x)\sqrt{x^2+4}}{2} . \ )$$

6.66 Si scrivano le espansioni in frazione continua delle funzioni:

$$\text{a)} \quad f(x) = \arctan x, \qquad \text{b)} \quad \frac{\arcsin x}{\sqrt{1-x^2}} .$$

c) Si ricavi da a) una frazione continua per il calcolo di π e si dica quanti termini vanno considerati per ottenere un valore approssimato con un errore minore di 10^{-6}. Si confronti con gli errori che si ottengono con la serie di Taylor (esempi 4.1 e 4.8).

(Traccia: a) posto

$$t_1(x) = 1, \quad t_2(x) = \sum_{i=0}^{\infty} (-1)^i \, \frac{x^{2i}}{2i+1} ,$$

si verifichi che le funzioni

$$t_{k+2}(x) = \frac{t_k(x) - (2k-1)t_{k+1}(x)}{k^2x^2}\,, \quad k = 1, 2, \ldots$$

sono serie di potenze pari in x. Con il metodo di Viskovatov si ottiene allora

$$\begin{aligned}\arctan x = \frac{xt_2(x)}{t_1(x)} &= \frac{x}{\dfrac{t_1(x)}{t_2(x)}} = \frac{x}{1} + \frac{x^2t_3(x)}{t_2(x)} = \frac{x}{1} + \frac{x^2}{\dfrac{t_2(x)}{t_3(x)}} \\ &= \frac{x}{1} + \frac{x^2}{3} + \frac{4x^2t_4(x)}{t_3(x)} \\ &= \ldots = \frac{x}{1} + \frac{x^2}{3} + \frac{4x^2}{5} + \ldots + \frac{k^2x^2}{2k+1} + \ldots\end{aligned}$$

L'espansione converge per ogni x reale.
b) Si proceda in modo analogo, a partire da

$$\begin{aligned}t_1(x) &= \sqrt{1-x^2} = 1 - \frac{x^2}{2} - \frac{x^4}{2\cdot 4} - \frac{1\cdot 3x^6}{2\cdot 4\cdot 6} - \ldots \\ t_2(x) &= \frac{\arcsin x}{x} = 1 + \frac{x^2}{2\cdot 3} + \frac{1\cdot 3x^4}{2\cdot 4\cdot 5} + \frac{1\cdot 3\cdot 5x^6}{2\cdot 4\cdot 6\cdot 7} + \ldots\end{aligned}$$

verificando che le funzioni

$$t_{k+2}(x) = \frac{t_k(x) - t_{k+1}(x)}{\alpha_k x^2}\,, \text{ dove } \alpha_k = \begin{cases} -\dfrac{k(k-1)}{(2k-1)(2k+1)} & \text{se } k \text{ pari,} \\ -\dfrac{k(k+1)}{(2k-1)(2k+1)} & \text{se } k \text{ dispari,}\end{cases}$$

sono serie di potenze pari in x. Si ottiene l'espansione

$$\begin{aligned}\frac{\arcsin x}{\sqrt{1-x^2}} = \frac{x}{1} &- \frac{1\cdot 2x^2/(1\cdot 3)}{1} - \frac{1\cdot 2x^2/(3\cdot 5)}{1} - \frac{3\cdot 4x^2/(5\cdot 7)}{1} \\ &- \frac{3\cdot 4x^2/(7\cdot 9)}{1} - \frac{5\cdot 6x^2/(9\cdot 11)}{1} - \ldots\end{aligned}$$

che nel campo reale converge per $|x| < 1$.
c) Ponendo $x = 1$ si ha

$$\pi = 4\arctan 1 = \frac{4}{1} + \frac{1}{3} + \frac{4}{5} + \frac{9}{7} + \frac{16}{9} + \ldots$$

Le successive frazioni parziali sono

$$w_1 = 4, \quad w_2 = 3, \quad w_3 = \frac{19}{6} = 3.166667, \quad w_4 = \frac{160}{51} = 3.137255,$$
$$w_5 = \frac{1744}{555} = 3.142342, \quad w_6 = \frac{644}{205} = 3.141463.$$

Si ottiene un risultato affetto da un errore minore di 10^{-6} considerando w_9.)

6.67 Un metodo per trasformare in frazione continua funzioni sviluppabili in serie di Taylor si basa sulla seguente relazione

$$\sum_{k=1}^{n} \frac{1}{a_k} = \frac{1}{a_1} \; - \; \frac{a_1^2}{a_1 + a_2} \; - \ldots \; - \; \frac{a_{n-1}^2}{a_{n-1} + a_n} \, .$$

a) Si verifichi tale relazione;
b) si applichi la relazione al caso della funzione e^x e si utilizzi l'espansione ottenuta con $x = -1$ per esprimere $\dfrac{1}{\mathrm{e} - 1}$.

(Traccia: a) si proceda per induzione su n: ponendo

$$\frac{1}{b} = \frac{1}{a_n} \; + \; \frac{1}{a_{n+1}} \, ,$$

si ha

$$\sum_{k=1}^{n+1} \frac{1}{a_k} = \sum_{k=1}^{n-1} \frac{1}{a_k} \; + \; \frac{1}{b} = \frac{1}{a_1} \; - \ldots \; - \; \frac{a_{n-1}^2}{a_{n-1} + b} \, ,$$

e si noti che

$$b = \frac{a_n a_{n+1}}{a_n + a_{n+1}} = a_n - \; \frac{a_n^2}{a_n + a_{n+1}} \, ;$$

b) si ha

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots = 1 + \frac{1}{x^{-1}} + \frac{1}{2x^{-2}} + \frac{1}{3!\, x^{-3}} + \frac{1}{4!\, x^{-4}} + \ldots$$

da cui

$$\begin{aligned} \mathrm{e}^x &= \frac{1}{1} \; - \; \frac{1}{1 + x^{-1}} \; - \; \frac{x^{-2}}{x^{-1} + 2x^{-2}} \; - \ldots \; - \; \frac{(k!)^2 x^{-2k}}{k!\, x^{-k} + (k+1)!\, x^{-(k+1)}} \; - \ldots \\ &= \frac{1}{1} \; - \; \frac{x}{x+1} \; - \; \frac{x}{x+2} \; - \ldots \; - \; \frac{(k!)^2 x}{k!\, x + (k+1)!} \; - \ldots \\ &= \frac{1}{1} \; - \; \frac{x}{x+1} \; - \; \frac{x}{x+2} \; - \ldots \; - \; \frac{kx}{x+k+1} \; - \ldots \end{aligned}$$

e quindi

$$\frac{1}{\mathrm{e}-1} = \frac{1}{\dfrac{1}{\mathrm{e}^{-1}}-1} = \frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \ldots + \frac{k}{k} + \ldots \,.\ )$$

6.68 Si determini l'approssimante di Padé $R_{4,4}(x)$ della funzione $f(x) = \mathrm{e}^x$ e se ne valuti il resto $r_{4,4}(x) = f(x) - R_{4,4}(x)$ in un intorno dello zero.

(Risposta: $$R_{4,4}(x) = \frac{1 + \dfrac{1}{2}\,x + \dfrac{3}{28}\,x^2 + \dfrac{1}{84}\,x^3 + \dfrac{1}{1680}\,x^4}{1 - \dfrac{1}{2}\,x + \dfrac{3}{28}\,x^2 - \dfrac{1}{84}\,x^3 + \dfrac{1}{1680}\,x^4}\,,$$

$$r_{4,4}(x) = \frac{x^9}{25401600} + O(x^{10}).\ )$$

6.69 Si determinino le approssimanti di Padé $R_{m,n}(x)$ di ordine $m+n = 5$, con $m \geq 1$, della funzione $f(x) = \sin x$ e si diano delle stime dei resti $r_{m,n}(x) = f(x) - R_{m,n}(x)$ in un intorno dello zero.

(Risposta:

$$R_{5,0}(x) = x - \frac{x^3}{6} + \frac{x^5}{120}\,, \quad r_{5,0}(x) = -\frac{x^7}{5040} + O(x^9);$$

$$R_{4,1}(x) = x - \frac{x^3}{6}\,, \quad r_{4,1}(x) = \frac{x^5}{120} + O(x^7);$$

$$R_{3,2}(x) = \frac{x - \dfrac{7}{60}\,x^3}{1 + \dfrac{1}{20}\,x^2}\,, \quad r_{3,2}(x) = \frac{11x^7}{50400} + O(x^9);$$

$$R_{2,3}(x) = \frac{x}{1 + \dfrac{1}{6}\,x^2}\,, \quad r_{2,3}(x) = -\frac{7x^5}{360} + O(x^7);$$

$$R_{1,4}(x) = \frac{x}{1 + \dfrac{1}{6}\,x^2 + \dfrac{7}{360}\,x^4}\,, \quad r_{1,4}(x) = -\frac{31x^7}{15120} + O(x^9).\ )$$

6.70 Siano $R_{m,n}(x) = \dfrac{p_m(x)}{q_n(x)}$ e $\widetilde{R}_{m,n}(x) = \dfrac{\widetilde{p}_m(x)}{\widetilde{q}_n(x)}$ due approssimanti di Padé di ordine $m+n$ di una funzione $f(x)$. Si verifichi che

$$p_m(x)\widetilde{q}_n(x) = \widetilde{p}_m(x)q_n(x),$$

da cui segue che, fissati m e n, l'approssimante di Padé $R_{m,n}(x)$ di una funzione $f(x)$ è unica.

(Traccia: si sfrutti la relazione

$$\begin{aligned} p_m(x)\widetilde{q}_n(x) - \widetilde{p}_m(x)q_n(x) = {} & \big[f(x)\widetilde{q}_n(x) - \widetilde{p}_m(x)\big]q_n(x) \\ & - \big[f(x)q_n(x) - p_m(x)\big]\widetilde{q}_n(x), \end{aligned}$$

e si tenga conto che il polinomio al primo membro ha grado $k \le m+n$, mentre lo sviluppo di Maclaurin al secondo membro ha nulli i primi $k+1$ termini.)

6.71 Siano

$$\begin{aligned} R_{m,n}(x) &= \frac{p_m(x)}{q_n(x)} \quad R_{m,n+1}(x) = \frac{\widetilde{p}_m(x)}{q_{n+1}(x)} \\ R_{m+1,n}(x) &= \frac{p_{m+1}(x)}{\widetilde{q}_n(x)} \quad R_{m+1,n+1}(x) = \frac{\widetilde{p}_{m+1}(x)}{\widetilde{q}_{n+1}(x)} \end{aligned}$$

quattro approssimanti contigue di una tabella di Padé normale. Si dimostri che esistono delle costanti non nulle β_i, $i = 1, \ldots, 6$, tali che

$$\begin{aligned} & p_m(x)\widetilde{q}_n(x) - p_{m+1}(x)q_n(x) = \beta_1 x^{m+n+1} \\ & \widetilde{p}_m(x)\widetilde{q}_{n+1}(x) - \widetilde{p}_{m+1}(x)q_{n+1}(x) = \beta_2 x^{m+n+2} \\ & p_m(x)q_{n+1}(x) - \widetilde{p}_m(x)q_n(x) = \beta_3 x^{m+n+1} \\ & p_{m+1}(x)\widetilde{q}_{n+1}(x) - \widetilde{p}_{m+1}(x)\widetilde{q}_n(x) = \beta_4 x^{m+n+2} \\ & p_m(x)\widetilde{q}_{n+1}(x) - \widetilde{p}_{m+1}(x)q_n(x) = \beta_5 x^{m+n+1} \\ & p_{m+1}(x)q_{n+1}(x) - \widetilde{p}_m(x)\widetilde{q}_n(x) = \beta_6 x^{m+n+2}. \end{aligned}$$

(Traccia: il polinomio

$$t(x) = p_m(x)\widetilde{q}_n(x) - p_{m+1}(x)q_n(x)$$

ha grado $m+n+1$, ed inoltre

$$t(x) = \big[f(x)\widetilde{q}_n(x) - p_{m+1}(x)\big]q_n(x) - \big[f(x)q_n(x) - p_m(x)\big]\widetilde{q}_n(x),$$

in cui $f(x)\widetilde{q}_n(x) - p_{m+1}(x)$ ha nulli i primi $m+n+2$ termini, e $f(x)q_n(x) - p_m(x)$ ha nulli i primi $m+n+1$ termini. Le altre relazioni si dimostrano in modo analogo.)

6.72 Si verifichi che se $R_{m,n}(x) = \dfrac{p_m(x)}{q_n(x)}$ è l'approssimante di Padé di ordine $m+n$ di una funzione $f(x)$ tale che $f(0) \neq 0$, allora

$$\frac{q_n(x)/f(0)}{p_m(x)/f(0)}$$

è l'approssimante di Padé di ordine $m+n$ di $\dfrac{1}{f(x)}$.

(Traccia: dallo sviluppo di Maclaurin

$$f(x)q_n(x) - p_m(x) = \sum_{i=1}^{\infty} \beta_i \, x^{m+n+i}$$

si ottiene

$$\frac{1}{f(x)} \, p_m(x) \, - q_n(x) = - \frac{1}{f(x)} \, \sum_{i=1}^{\infty} \beta_i \, x^{m+n+i}.$$

Si verifichi che lo sviluppo di Maclaurin della funzione a secondo membro è della forma

$$- \frac{1}{f(x)} \, \sum_{i=1}^{\infty} \beta_i \, x^{m+n+i} = \sum_{i=1}^{\infty} \gamma_i \, x^{m+n+i},$$

dove $\gamma_1 = - \dfrac{\beta_1}{f(0)}$.)

6.73 Sia

$$r_{m,m}(x) = \frac{p_m(x)}{q_m(x)} = \frac{a_m x^m + a_{m-1}x^{m-1} + \ldots + a_0}{b_m x^m + b_{m-1}x^{m-1} + \ldots + b_0}$$

una funzione razionale irriducibile. Si verifichi che $r_{m,m}(x) = -r_{m,m}\left(\dfrac{1}{x}\right)$ se e solo se

$$a_j = \alpha a_{m-j} \quad \text{e} \quad b_j = -\alpha b_{m-j}, \quad j = 0, \ldots, m,$$

dove $\alpha = 1$ oppure $\alpha = -1$. Questa proprietà vale per le approssimanti di Padé $R_{m,m}(x)$ di $\log x$ con centro nel punto $x = 1$ (si ottengono dalle approssimanti di $\log(1+y)$ con la trasformazione $y = x - 1$), infatti per il

logaritmo vale la proprietà $\log x = -\log \dfrac{1}{x}$. Si verifichi in particolare per il caso $m = 3$.

(Traccia: dalla condizione $r_{m,m}(x) = -r_{m,m}\Big(\dfrac{1}{x}\Big)$ segue che

$$p_m(x) = \alpha x^m p_m\Big(\frac{1}{x}\Big) \quad \text{e} \quad q_m(x) = -\alpha x^m q_m\Big(\frac{1}{x}\Big),$$

in cui $\alpha = 1$ oppure $\alpha = -1$, e viceversa. Per $f(x) = \log x$ si ha

$$R_{3,3}(x) = \frac{\dfrac{11}{3}\,x^3 + 9x^2 - 9x - \dfrac{11}{3}}{x^3 + 9x^2 + 9x + 1}\,.\ )$$

6.74 a) Si verifichi che se la funzione $f(x)$ è pari o è dispari, la sua tabella di Padé non è normale.

b) In tal caso conviene porre $z = x^2$ nello sviluppo di Maclaurin e costruire la tabella rispetto a z, poi fare la sostituzione inversa. Si costruisca in questo modo la tabella di $\cos x$.

(Traccia: a) se $f(x)$ è pari, lo sviluppo di Maclaurin è della forma

$$\sum_{i=0}^{\infty} \alpha_{2i} x^{2i},$$

e sia $R_{m,n}(x) = \dfrac{p_m(x)}{q_n(x)}$, m, n pari, un'approssimante di Padé tale che i polinomi $p_m(x)$ e $q_n(x)$ siano di grado m ed n. Si verifichi che anche $p_m(x)$ e $q_n(x)$ sono funzioni pari, per cui la funzione $s(x) = f(x)q_n(x) - p_m(x)$ è pari. Ne segue che il primo termine non nullo dello sviluppo di Maclaurin di $s(x)$ è di grado $m+n+2$, per cui $R_{m,n}(x)$ è anche approssimante di Padé di ordine $m + n + 1$, cioè

$$R_{m,n}(x) = R_{m+1,n}(x) = R_{m,n+1}(x).$$

Inoltre è possibile che un'approssimante sia rapporto di due polinomi dispari in quanto

$$\frac{p_{m+1}(x)}{q_{n+1}(x)} = \frac{x p_m(x)}{x q_n(x)}\,,$$

e riducendo si ha

$$R_{m,n}(x) = R_{m+1,n+1}(x).$$

Si proceda in modo analogo per le funzioni dispari. b) Per la funzione

$$\cos\sqrt{z} = 1 - \frac{z}{2} + \frac{z^2}{4!} - \frac{z^3}{6!} + \dots$$

è

$$R_{0,0}(z) = 1, \quad R_{1,0}(z) = 1 - \frac{z}{2}, \quad R_{2,0}(z) = 1 - \frac{z}{2} + \frac{z^2}{4!}, \quad \dots$$

$$R_{0,1}(z) = \frac{1}{1 + \frac{1}{2}z}, \quad R_{1,1}(z) = \frac{1 - \frac{5}{12}z}{1 + \frac{1}{12}z},$$

$$R_{2,1}(z) = \frac{1 - \frac{7}{15}z + \frac{1}{40}z^2}{1 + \frac{1}{30}z}, \quad \dots$$

da cui si ottiene, per la funzione $f(x) = \cos x$

$$R_{0,0}(x) = R_{1,0}(x) = R_{0,1}(x) = R_{1,1}(x) = 1,$$
$$R_{2,0}(x) = R_{2,1}(x) = R_{3,0}(x) = R_{3,1}(x) = 1 - \frac{x^2}{2},$$
$$R_{0,2}(x) = R_{1,2}(x) = R_{0,3}(x) = R_{1,3}(x) = \frac{1}{1 + \frac{1}{2}x^2},$$
$$R_{2,2}(x) = R_{2,3}(x) = R_{3,2}(x) = R_{3,3}(x) = \frac{1 - \frac{5}{12}x^2}{1 + \frac{1}{12}x^2},$$
$$R_{4,2}(x) = R_{4,3}(x) = R_{5,2}(x) = R_{5,3}(x) = \frac{1 - \frac{7}{15}x^2 + \frac{1}{40}x^4}{1 + \frac{1}{30}x^2}. \ )$$

6.75 Si determini l'approssimazione lineare ai minimi quadrati per la funzione $f(x)$ di cui sono noti i valori $f(x_i) = y_i$, $i = 0, 1, 2$. Si applichi al caso particolare della funzione

x	0	1	2
$f(x)$	0	1	1

(Traccia: i coefficienti $\alpha_0^{(2)}$, $\alpha_1^{(2)}$ del polinomio cercato $p_1^{(2)}(x) = \alpha_0^{(2)} + \alpha_1^{(2)} x$ sono dati dalla soluzione del sistema

$$\begin{bmatrix} 3 & x_0 + x_1 + x_2 \\ x_0 + x_1 + x_2 & x_0^2 + x_1^2 + x_2^2 \end{bmatrix} \begin{bmatrix} \alpha_0^{(2)} \\ \alpha_1^{(2)} \end{bmatrix} = \begin{bmatrix} y_0 + y_1 + y_2 \\ x_0 y_0 + x_1 y_1 + x_2 y_2 \end{bmatrix}.$$

Nel caso particolare $\alpha_0^{(2)} = \dfrac{1}{6}$, $\alpha_1^{(2)} = \dfrac{1}{2}$.)

6.76 Siano $P_i = (x_i, y_i)$, $i = 0, 1, 2$, tre punti di $\mathbf{R}^2$ tali che $x_1 < x_2 < x_3$. Si verifichi che esiste una e una sola retta $s(x)$, la cui distanza dai tre punti, misurata parallelamente all'asse y, è la stessa, e tale che due dei punti giacciono da parte opposta della retta rispetto al terzo. Tale retta è detta *retta di Chebyshev* e rappresenta l'approssimazione minimax lineare della funzione discreta $f(x_i) = y_i$, $i = 0, 1, 2$. Si determini la retta di Chebyshev dei punti $(0, 0)$, $(1, 1)$, $(2, 1)$, e si confronti con l'approssimazione lineare ai minimi quadrati ottenuta nell'esercizio precedente.

(Traccia: si verifichi che, posto $s(x) = ax + b$ e $h_i = y_i - s(x_i)$, per $i = 0, 1, 2$, è possibile determinare a e b in modo che sia $h_0 = -h_1 = h_2$; risulta infatti

$$a = \frac{y_2 - y_0}{x_2 - x_0}, \quad b = \frac{y_1}{2} + \frac{y_0(x_1 + x_2) - y_2(x_0 + x_1)}{2(x_2 - x_0)},$$
$$h = |h_i| = \left| \frac{y_1}{2} + \frac{y_0(x_1 - x_2) + y_2(x_0 - x_1)}{2(x_2 - x_0)} \right|.$$

Nel caso particolare si ha

$$a = \frac{1}{2}, \quad b = \frac{1}{4}, \quad h = \frac{1}{4}, \quad \text{quindi} \quad \max_{i=0,1,2} |y_i - s(x_i)| = \frac{1}{4},$$

mentre nel caso della retta ottenuta con i minimi quadrati risulta

$$p_1^{(2)}(x) = \frac{1}{2} x + \frac{1}{6}, \quad \text{quindi} \quad \max_{i=0,1,2} |y_i - p_1^{(2)}(x_i)| = \frac{1}{3}.)$$

6.77 Per determinare l'approssimazione minimax lineare della funzione $f(x)$ di cui sono noti i valori $f(x_i) = y_i$, $i = 0, \ldots, m$, si proceda con il seguente *algoritmo di scambio*:

(1) si scelga inizialmente una terna di punti,

(2) si determini la retta di Chebyshev $s(x)$ relativa alla terna scelta e si determini il corrispondente h,

(3) si calcoli $H = \max_{i=0,\ldots,m} |y_i - s(x_i)|$. Se $H = h$, allora $s(x)$ è l'approssimazione cercata, altrimenti si aggiunga ai punti della terna un punto x_k tale che $|y_k - s(x_k)| = H$, e si scarti uno dei punti precedenti, scelto in modo che i tre rimasti abbiano errori di segno alterno; si ripeta da (2).

Si verifichi che questo algoritmo termina in un numero finito di passi e che la retta $s(x)$ ottenuta corrisponde effettivamente all'approssimazione minimax cercata. Si applichi al caso particolare della funzione

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$f(x)$	1	2	2	3	1	2	3	2	3	2	4	3	4	5	4	5	5

Un algoritmo, in generale più veloce, si ha se in (3), anziché scambiare un solo punto se ne scambiano due o tre, scelti in modo che gli errori abbiano segno alterno e modulo il più elevato possibile.

(Traccia: si può supporre, senza ledere la generalità, che all'i-esimo passo dell'algoritmo la terna scelta sia quella formata da x_0, x_1, x_2 e per semplicità che al passo successivo sia x_0, x_1, x_3, quindi $|y_3 - ax_3 - b| = H > h$, dove a, b e h sono dati nella traccia dell'esercizio precedente. Si verifichi che, posto

$$\overline{h} = \left| \frac{y_1}{2} + \frac{y_0(x_1 - x_3) + y_3(x_0 - x_1)}{2(x_3 - x_0)} \right|,$$

risulta $\overline{h} > h$. Perciò la successione degli h determinati dall'algoritmo è crescente e non sarà possibile scegliere una terna già scelta in un passo precedente, cioè non saranno possibili cicli. Nel caso particolare si ha

terna	$(0,1,2)$	$(0,1,16)$	$(0,1,9)$	$(1,9,13)$	$(3,9,13)$
a, b	$\frac{1}{2}, \frac{5}{4}$	$\frac{1}{4}, \frac{11}{8}$	$\frac{1}{19}, \frac{13}{9}$	$\frac{1}{4}, \frac{3}{4}$	$\frac{1}{5}, \frac{13}{10}$
h, H	$\frac{1}{4}, \frac{17}{4}$	$\frac{3}{8}, \frac{13}{8}$	$\frac{4}{9}, \frac{19}{9}$	$1, \frac{3}{2}$	$\frac{11}{10}, \frac{11}{10}$

Quindi l'approssimazione minimax cercata è

$$q_1^{(20)}(x) = \frac{1}{5}\,(x+13).$$

Usando l'algoritmo più veloce vengono scelte le terne $(0,1,2)$, $(0,1,16)$, $(3,9,13)$.)

6.78 a) Negli $n+2$ punti x_i, $i=0,\ldots,n+1$, ordinati in modo crescente, sono noti i valori $f(x_i)$. Si costruiscano i polinomi $t(x)$ e $s(x) \in \mathcal{P}_{n+1}$ tali che

$$t(x_i) = f(x_i) \quad \text{e} \quad s(x_i) = (-1)^i, \quad \text{per} \quad i = 0,\ldots,n+1,$$

e siano t_{n+1} e s_{n+1} i coefficienti dei termini di grado $n+1$ di $t(x)$ e $s(x)$. Si verifichi che il polinomio

$$q_n^{(n+1)}(x) = t(x) - \frac{t_{n+1}}{s_{n+1}}\, s(x)$$

è l'approssimazione minimax discreta di grado n della funzione $f(x)$ sui nodi x_i, $i=0,\ldots,n+1$.

b) Si generalizzi l'algoritmo di scambio dell'esercizio precedente al caso dell'approssimazione minimax di grado n della funzione $f(x)$ di cui sono noti i valori $f(x_i)$, $i=0,\ldots,m$, con $m \geq n+1$.

(Traccia: a) se $t_{n+1}=0$, allora $q_n^{(n+1)}(x) = t(x)$ è il polinomio di interpolazione di $f(x)$ nei nodi e

$$\max_{i=0,\ldots,n+1} |f(x_i) - q_n^{(n+1)}(x_i)| = 0.$$

Se $t_{n+1} \neq 0$, allora $q_n^{(n+1)}(x) \in \mathcal{P}_n$ ed è tale che

$$f(x_i) - q_n^{(n+1)}(x_i) = (-1)^i d, \quad d = \frac{t_{n+1}}{s_{n+1}}, \quad \text{per } i = 0,\ldots,n+1,$$

e quindi $q_n^{(n+1)}(x)$ è il polinomio di equioscillazione.
b) Si ha:

(1) si scelga inizialmente un sottoinsieme S di $n+2$ nodi;

(2) si determini, come indicato in a), il polinomio $q_n(x)$ di approssimazione minimax di grado n su tale sottoinsieme;

(3) si calcoli $q_n(x_i)$ per $i=0,\ldots,m$, e si controlli se

$$\max_{i=0,\ldots,m} |f(x_i) - q_n(x_i)| = |d|.$$

In caso affermativo è $q_n^{(n+1)}(x) = q_n(x)$, altrimenti si scambino uno o più nodi di S con altri nodi non di S, scelti in modo che gli errori abbiano segno alterno e modulo maggiore o uguale a $|d|$, e si ripeta da (2).)

6.79 Per approssimare $f(x) = \sqrt{x}$ con un polinomio di grado 2 nell'intervallo $[1/16, 1]$, si seguano le vie seguenti:

a) si costruisca il polinomio di approssimazione quasi minimax con un passo del metodo di Remez, scegliendo come punti iniziali

$$x_0 = \frac{1}{16}\,, \quad x_1 = \frac{1}{4}\,, \quad x_2 = \frac{9}{16}\,, \quad x_3 = 1;$$

b) si costruisca il polinomio di approssimazione minimax;

c) si costruisca il polinomio di approssimazione minimax nel discreto per la funzione $f(x)$ tabulata negli stessi punti di a);

d) si costruisca il polinomio di approssimazione minimax nel discreto per la funzione $f(x)$ tabulata nei punti

$$x_0 = 0.09, \quad x_1 = 0.16, \quad x_2 = 0.25, \quad x_3 = 0.36, \quad x_4 = 0.49,$$
$$x_5 = 0.64, \quad x_6 = 0.81, \quad x_7 = 1.$$

Si confrontino fra di loro i massimi moduli dei resti ottenuti. Si confronti anche con il risultato dell'esempio 5.7. Si verifichi che il polinomio ottenuto in c) è migliore in norma $\| \ .\|_\infty$ di quello ottenuto in d), nonostante sia costruito su un minor numero di punti. Si dia una spiegazione di questo risultato.

(Traccia: a) si ottiene

$$p_2(x) = -\frac{8}{15}\,x^2 + \frac{4}{3}\,x + \frac{59}{320} = -0.5333333\,x^2 + 1.333333\,x + 0.184375,$$
$$\text{con} \quad \|f - p_2\|_\infty \approx 0.198\ 10^{-1}.$$

b) Direttamente o con il metodo di Remez si ottiene

$$p_2^*(x) = -0.5237170\,x^2 + 1.319420\,x + 0.1869398,$$
$$\text{con} \quad \|f - p_2^*\|_\infty \approx 0.174\ 10^{-1}.$$

c) Si ottiene lo stesso polinomio che in a); d) si ottiene

$$q_2^{(7)}(x) = -0.4661961\,x^2 + 1.2494135\,x + 0.2040562,$$
$$\text{con} \quad \|f - q_2^{(7)}\|_\infty \approx 0.303\ 10^{-1};$$

il resto dell'approsimazione nel continuo è standard, fra i nodi del polinomio in c) sono compresi entrambi gli estremi, mentre fra i nodi del polinomio in d) non è compreso il primo estremo.)

6.80 Per costruire l'approssimazione minimax razionale $w(x) \in \mathcal{R}_{0,1}$ della funzione discreta definita dalla tabella

x	x_0	x_1	x_2
$f(x)$	y_0	y_1	y_2

con $x_0 < x_1 < x_2$,

si considera la funzione

$$w(x) = \frac{a_0}{b_1 x + 1}, \quad b_1 x + 1 \neq 0 \quad \text{per} \quad x \in [x_0, x_2],$$

e si risolve il sistema

$$y_i - w(x_i) = (-1)^i d, \quad i = 0, 1, 2. \tag{119}$$

Si verifichi che il sistema (119) può avere due soluzioni reali diverse, a cui corrispondono due diverse funzioni razionali di equioscillazione, e che in generale non è detto che la soluzione corrispondente al minimo $|d|$ sia l'approssimazione minimax, né che lo sia una delle due soluzioni. Quindi nel discreto non valgono teoremi analoghi al 6.50 e 6.51. Si considerino i seguenti casi particolari

a)

x	0	$\frac{1}{2}$	1
$f(x)$	$\frac{3}{4}$	$-\frac{7}{4}$	$-\frac{3}{4}$

b)

x	0	$\frac{1}{2}$	1
$f(x)$	-1	$\frac{1}{2}$	1

(Traccia: procedendo per sostituzione nella risoluzione del sistema (119), e sostituendo d e a_0, b_1 risulta soluzione di un'equazione di 2^0 grado. Per uno o entrambi i valori di b_1 che si ottengono è possibile che il denominatore $b_1 x + 1$ si annulli nell'intervallo $[x_0, x_2]$, pertanto la corrispondente soluzione non può essere considerata di minimax. Casi particolari:
a) dal sistema (119) si ottengono le due soluzioni

$$a_0 = -\frac{3}{8}, \quad b_1 = -\frac{4}{5}, \quad d = \frac{9}{8},$$
$$a_0 = 1, \quad b_1 = -3, \quad d = -\frac{1}{4}.$$

Entrambe le funzioni sono di equioscillazione, e la seconda con un minore $|d|$. Però la funzione

$$w(x) = \frac{1}{-3x+1}$$

non è definita in $\frac{1}{3} \in [0,1]$, e quindi non può essere considerata di minimax.
b) Dal sistema (119) si ottengono le due soluzioni

$$a_0 = \frac{-7+\sqrt{33}}{8}, \quad b_1 = \frac{-9+\sqrt{33}}{3}, \quad d = \frac{-1-\sqrt{33}}{8},$$
$$a_0 = \frac{-7-\sqrt{33}}{8}, \quad b_1 = \frac{-9-\sqrt{33}}{3}, \quad d = \frac{-1+\sqrt{33}}{8}.$$

Entrambe le funzioni sono di equioscillazione, ma entrambe non sono definite in un punto di $[x_0, x_2]$ e quindi non sono di minimax.)

6.81 a) Si verifichi che il calcolo di $\sqrt{x}$ per ogni $x > 0$ può essere ricondotto al calcolo di

$$\sqrt{x} \quad \text{per} \quad x \in [\beta^{-1}, 1],$$

in cui β è la base usata nella rappresentazione dei numeri.

b) L'approssimazione di $\sqrt{x}$ viene calcolata con 2 iterazioni del metodo delle tangenti (si veda l'esempio 3.22), a partire da un punto iniziale x_0 sufficientemente vicino. Per $\beta = 16$, ad esempio, tale approssimazione può essere calcolata con la funzione razionale dell'esempio 6.54

$$x_0 = 1.681835 - \frac{1.289551}{x + 0.8410629}.$$

Si stimi il massimo errore assoluto e relativo commesso dopo due iterazioni del metodo delle tangenti.

(Traccia: a) sia

$$x = m\,\beta^p, \quad \text{con} \quad \beta^{-1} \le m < 1.$$

Se $p = 2n$, n intero, è

$$\sqrt{x} = \sqrt{m}\,\beta^n, \quad \text{con} \quad \beta^{-1} < \sqrt{m} < 1.$$

Se $p = 2n+1$, è

$$\sqrt{x} = \sqrt{m\beta^{-1}}\,\beta^{n+1}, \quad \text{con} \quad \beta^{-1} \le \sqrt{m\beta^{-1}} < 1.$$

b) I massimi errori assoluto e relativo effettivamente generati risultano minori in modulo di $0.155\ 10^{-7}$.)

6.82 a) Si verifichi che il calcolo di $\sin x$ e $\cos x$ per ogni x reale può essere ricondotto al calcolo di $\sin x$ e $\cos x$ per $x \in \left[0, \frac{\pi}{4}\right]$.

b) Si determini, con il metodo di economizzazione, un polinomio $p_7(x)$ di grado 7 che approssimi $\sin x$ per $x \in \left[0, \frac{\pi}{4}\right]$, tale che

(1) $p_7(x)$ sia una funzione dispari, e quindi $p_7(0) = 0$,

(2) $\lim_{x\to 0} \frac{p_7(x)}{x} = 1$,

e se ne stimino i massimi errori analitici assoluti e relativi.

c) Si determini, con il metodo di economizzazione, un polinomio $p_6(x)$ di grado 6 che approssimi $\cos x$ per $x \in \left[0, \frac{\pi}{4}\right]$, tale che $p_6(x)$ sia una funzione pari, e valga $p_6(0) = 1$. Se ne stimino i massimi errori analitici assoluti e relativi.

d) Si costruiscano i polinomi di approssimazione minimax di grado 7 per la funzione $\sin x$ e di grado 6 per la funzione $\cos x$, in modo che valgano le proprietà dei punti b) e c).

e) Si esamini se le proprietà di crescenza della funzione $\sin x$ e di decrescenza della funzione $\cos x$ sull'intervallo $[0, \pi/2]$ sono verificate anche dalle approssimazioni minimax.

(Traccia: a) si tenga conto della periodicità e della simmetria delle funzioni e del fatto che

$$\sin\left(\frac{\pi}{4} \pm x\right) = \cos\left(\frac{\pi}{4} \mp x\right).$$

b) Si considerino i primi 4 termini della serie di Maclaurin di

$$\frac{1}{x^2}\left(\frac{\sin x}{x} - 1\right), \quad x \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right].$$

Si ottiene il polinomio

$$-\frac{1}{3!} + \frac{x^2}{5!} - \frac{x^4}{7!} + \frac{x^6}{9!}.$$

Con la trasformazione $x = \frac{\pi}{4}\, t$ e il metodo di economizzazione si determina un polinomio $q(t)$ di quarto grado, da cui si ottiene

$$p_7(x) = x + x^3 q\left(\frac{4}{\pi}\, x\right) = x - 0.1666666\, x^3 + 0.008332744\, x^5 - 0.0001958629\, x^7,$$

e risulta

$$\max_{x\in[0,\pi/4]} |\sin x - p_7(x)| \approx 0.240\ 10^{-8},\quad \max_{x\in[0,\pi/4]} \left|\frac{\sin x - p_7(x)}{\sin x}\right| \approx 0.364\ 10^{-8}.$$

c) Si proceda come in b), partendo dallo sviluppo di

$$\frac{\cos x - 1}{x^2},\quad x \in \Big[-\frac{\pi}{4}, \frac{\pi}{4}\Big].$$

Si ottiene il polinomio

$$p_6(x) = 1 - 0.4999998\, x^2 + 0.04166132\, x^4 - 0.001365941\, x^6,$$

e risulta

$$\max_{x\in[0,\pi/4]} |\cos x - p_6(x)| \approx 0.924\ 10^{-7},\quad \max_{x\in[0,\pi/4]} \left|\frac{\cos x - p_6(x)}{\cos x}\right| \approx 0.129\ 10^{-6}.$$

d) Per la funzione $\sin x$ si applichi l'algoritmo di Remez a un polinomio della forma

$$p_7(x) = x + \sum_{i=1}^{3} a_i x^{2i+1},$$

scegliendo come nodi iniziali 3 punti in $(0, \pi/4)$ e il punto $\pi/4$. Non si include fra i nodi iniziali il punto 0, per cui risulterebbe $d^{(0)} = 0$. Si ottiene

$$p_7(x) = x - 0.1666665\, x^3 + 0.008331967\, x^5 - 0.0001949568\, x^7,$$

e risulta

$$\max_{x\in[0,\pi/4]} |\sin x - p_7(x)| \approx 0.212\ 10^{-8},\quad \max_{x\in[0,\pi/4]} \left|\frac{\sin x - p_7(x)}{\sin x}\right| \approx 0.637\ 10^{-8}.$$

Per la funzione $\cos x$ si applichi l'algoritmo di Remez a un polinomio della forma

$$p_6(x) = 1 + \sum_{i=1}^{3} a_i x^{2i}.$$

scegliendo i nodi iniziali come sopra. Si ottiene

$$p_6(x) = 1 - 0.4999989\, x^2 + 0.04165621\, x^4 - 0.001359781\, x^6,$$

e risulta

$$\max_{x\in[0,\pi/4]} |\cos x - p_6(x)| \approx 0.346\ 10^{-7},\quad \max_{x\in[0,\pi/4]} \left|\frac{\cos x - p_6(x)}{\cos x}\right| \approx 0.489\ 10^{-7}.$$

e) $p_7(x)$ è crescente e $p_6(x)$ è decrescente per $x \in \big[0, \frac{\pi}{4}\big]$, ma la funzione

$$p(x) = \begin{cases} p_7(x), & \text{per } x \in \big[0, \frac{\pi}{4}\big], \\ p_6(\frac{\pi}{2} - x), & \text{per } x \in \big(\frac{\pi}{4}, \frac{\pi}{2}\big] \end{cases}$$

non è né continua né crescente in $\frac{\pi}{4}$, infatti

$$p_7\big(\frac{\pi}{4}\big) - p_6\big(\frac{\pi}{4}\big) \approx 0.325\ 10^{-7}.\ )$$

6.83 a) Si verifichi che il calcolo di $\tan x$ per ogni x reale può essere ricondotto al calcolo di $\tan x$ per $x \in \big[0, \frac{\pi}{4}\big]$.

b) Si determini il polinomio $p_{11}(x)$ di grado 11 di approssimazione minimax di $\tan x$, $x \in \big[0, \frac{\pi}{4}\big]$, tale che

(1) $p_{11}(x)$ sia una funzione dispari, e quindi $p_{11}(0) = 0$,

(2) $\lim\limits_{x \to 0} \dfrac{p_{11}(x)}{x} = 1$.

c) Si determini il polinomio $p_{13}(x)$ di grado 13 di approssimazione minimax di $\tan x$, $x \in \big[0, \frac{\pi}{4}\big]$, come in b). Si noti come sia lenta la convergenza delle approssimazioni minimax di $\tan x$ al crescere del grado, a causa della singolarità nel punto $\frac{\pi}{2}$.

d) Si scrivano le formule di Padé $R_{3,4}(x)$ e $R_{5,4}(x)$ per $\tan x$.

e) Si determini la funzione razionale della forma

$$w(x) = x\left(1 + \frac{x^2}{p_4(x)}\right),$$

in cui $p_4(x)$ è un polinomio pari di grado 4, di approssimazione minimax di $\tan x$ per $x \in \big[0, \frac{\pi}{4}\big]$ (si noti che una funzione della forma di $w(x)$ soddisfa le condizioni poste in b)).

f) Si determini la funzione razionale della forma

$$w_{5,4}(x) = \frac{p_5(x)}{q_4(x)},$$

in cui $p_5(x)$ e $q_4(x)$ sono polinomi rispettivamente dispari di grado 5 e pari di grado 4, di approssimazione minimax di $\tan x$ per $x \in \big[0, \frac{\pi}{4}\big]$, e che soddisfi le condizioni poste in b).

g) Per ogni approssimazione trovata si stimino i massimi errori analitici assoluti e relativi e si determini quale approssimazione conviene usare anche sulla base del costo computazionale.

(Traccia: a) si tenga conto della periodicità, dell'antisimmetria e della relazione

$$\tan\left(x+\frac{\pi}{4}\right)=\frac{1}{\tan\left(\frac{\pi}{4}-x\right)}\,.$$

b) Si applichi l'algoritmo di Remez a un polinomio della forma

$$p_{11}(x)=x+\sum_{i=1}^{5}a_i x^{2i+1}.$$

scegliendo come nodi iniziali 5 punti in $(0,\pi/4)$ e il punto $\pi/4$. Non si include fra i nodi iniziali il punto 0, per cui risulterebbe $d^{(0)}=0$. Si ottiene

$$\begin{aligned}p_{11}(x)=x&+0.3333689\,x^3+0.1326939\,x^5+0.05791018\,x^7+0.01118992\,x^9\\&+0.02124745\,x^{11},\end{aligned}$$

e risulta

$$\max_{x\in[0,\pi/4]}|\tan x-p_{11}(x)|=\max_{x\in[0,\pi/4]}\left|\frac{\tan x-p_{11}(x)}{\tan x}\right|\approx 0.152\ 10^{-6}.$$

c) Procedendo come in b) si ottiene

$$\begin{aligned}p_{13}(x)=x&+0.3333295\,x^3+0.1334274\,x^5+0.05315180\,x^7+0.02520298\,x^9\\&+0.002051160\,x^{11}+0.009943453\,x^{13},\end{aligned}$$

e risulta

$$\max_{x\in[0,\pi/4]}|\tan x-p_{13}(x)|=\max_{x\in[0,\pi/4]}\left|\frac{\tan x-p_{13}(x)}{\tan x}\right|\approx 0.161\ 10^{-7}.$$

d) Risulta

$$R_{3,4}(x)=\frac{x(105-10x^2)}{105-45x^2+x^4}\,,$$

$$R_{5,4}(x)=\frac{x(945-105x^2+x^4)}{945-420x^2+15x^4}\,,$$

con gli errori

$$\max_{x\in[0,\pi/4]}|\tan x-R_{3,4}(x)|=\max_{x\in[0,\pi/4]}\left|\frac{\tan x-R_{3,4}(x)}{\tan x}\right|\approx 0.213\ 10^{-5},$$

e

$$\max_{x\in[0,\pi/4]} |\tan x - R_{5,4}(x)| = \max_{x\in[0,\pi/4]} \left|\frac{\tan x - R_{5,4}(x)}{\tan x}\right| \approx 0.135\ 10^{-7}.$$

e) Si applichi l'algoritmo di Remez alla funzione

$$w(x) = x\left(1 + \frac{x^2}{\sum_{i=0}^{2} a_i x^{2i}}\right).$$

Si ottiene

$$w(x) = x\left(1 + \frac{x^2}{2.999982 - 1.199857\,x^2 - 0.006059484\,x^4}\right),$$

e risulta

$$\max_{x\in[0,\pi/4]} |\tan x - w(x)| = \max_{x\in[0,\pi/4]} \left|\frac{\tan x - w(x)}{\tan x}\right| \approx 0.278\ 10^{-7}.$$

f) Si applichi l'algoritmo di Remez alla funzione

$$w_{5,4}(x) = \frac{\sum_{i=0}^{2} a_i x^{2i+1}}{\sum_{i=0}^{2} b_i x^{2i}}, \quad \text{con} \quad a_0 = b_0.$$

Si ottiene

$$w_{5,4}(x) = \frac{((x^2 - 103.5172)x^2 + 929.4733)x}{(14.85069\,x^2 - 413.3416)x^2 + 929.4733},$$

e risulta

$$\max_{x\in[0,\pi/4]} |\tan x - w_{5,4}(x)| \approx 0.428\ 10^{-8},$$

$$\max_{x\in[0,\pi/4]} \left|\frac{\tan x - w_{5,4}(x)}{\tan x}\right| \approx 0.514\ 10^{-8}.$$

g) Indicando con A il numero di operazioni additive e con M il numero di operazioni moltiplicative, il calcolo in un punto del polinomio ottenuto in b) richiede 5A+7M, di quello ottenuto in c) richiede 6A+8M,il calcolo delle approssimanti di Padé richiede 3A+5M per la prima, 4A+6M per la seconda, dell'approssimazione razionale ottenuta in e) richiede 3A+5M, di

quella ottenuta in f) richiede 4A+6M. Tenendo conto anche degli errori analitici trovati conviene usare l'approssimazione razionale ottenuta in f).)

6.84 a) Si verifichi che il calcolo di $\arctan x$ per ogni x reale può essere ricondotto al calcolo di

$$\arctan x \quad \text{per} \quad |x| \leq 2 - \sqrt{3} = 0.2679492.$$

b) Si determini, con il metodo di economizzazione, un polinomio $p_7(x)$ di grado 7 che approssimi $\arctan x$ per $|x| \leq 2 - \sqrt{3}$, tale che

(1) $p_7(x)$ sia una funzione dispari, e quindi $p_7(0) = 0$,

(2) $\lim_{x \to 0} \dfrac{p_7(x)}{x} = 1$,

c) Si determini, con il metodo di economizzazione, un polinomio $p_9(x)$ di grado 9 che approssimi $\arctan x$ per $|x| \leq 2 - \sqrt{3}$, come in b).

d) Si scrivano le formule di Padé $R_{3,4}(x)$ e $R_{5,4}(x)$ per $\arctan x$.

e) Si determini la funzione razionale della forma

$$w_{3,4}(x) = \frac{p_3(x)}{q_4(x)}\,,$$

in cui $p_3(x)$ e $q_4(x)$ sono polinomi rispettivamente dispari di grado 3 e pari di grado 4, di approssimazione minimax di $\arctan x$ per $|x| \leq 2-\sqrt{3}$ e che soddisfi le condizioni poste in b).

f) Per ogni approssimazione trovata si stimino i massimi errori analitici assoluti e relativi e si determini quale approssimazione conviene usare anche sulla base del costo computazionale.

(Traccia: a) per la riduzione all'intervallo $[0, 1]$ si tenga conto delle relazioni

$$\arctan(-x) = -\arctan(x),$$
$$\arctan x = \frac{\pi}{2} - \arctan\frac{1}{x}\,.$$

Per l'ulteriore riduzione si tenga conto della relazione

$$\arctan x = \frac{\pi}{6} + \arctan\frac{x\sqrt{3} - 1}{x + \sqrt{3}}\,.$$

b) Si considerino i primi 4 termini della serie di Maclaurin di

$$\frac{1}{x^2}\left(\frac{\arctan x}{x} - 1\right), \quad |x| \leq 2 - \sqrt{3}.$$

Si ottiene il polinomio

$$-\frac{1}{3}+\frac{x^2}{5}-\frac{x^4}{7}+\frac{x^6}{9}\,.$$

Con la trasformazione $x=(2-\sqrt{3})t$ e il metodo di economizzazione si determina un polinomio $q(t)$ di quarto grado, da cui si ottiene

$$p_7(x)=x+x^3q\big(\frac{x}{2-\sqrt{3}}\big)=x-0.3333321\,x^3+0.1996777\,x^5-0.1308920\,x^7,$$

e risulta

$$\max_{|x|\le 2-\sqrt{3}}|\arctan x-p_7(x)|\approx 0.178\ 10^{-7},$$
$$\max_{|x|\le 2-\sqrt{3}}\Big|\frac{\arctan x-p_7(x)}{\arctan x}\Big|\approx 0.129\ 10^{-6}.$$

c) Procedendo come in b) si ottiene

$$p_9(x)=x-0.3333333\,x^3+0.1999916\,x^5-0.1422714\,x^7+0.09805715\,x^9,$$

e risulta

$$\max_{|x|\le 2-\sqrt{3}}|\arctan x-p_9(x)|\approx 0.200\ 10^{-8},$$
$$\max_{|x|\le 2-\sqrt{3}}\Big|\frac{\arctan x-p_9(x)}{\arctan x}\Big|\approx 0.765\ 10^{-8}.$$

d) Risulta

$$R_{3,4}(x)=\frac{x(11.66667+6.111111\,x^2)}{11.66667+10\,x^2+x^4}\,,$$

$$R_{5,4}(x)=\frac{x(4.2+3.266667x^2+0.2844444\,x^4)}{4.2+4.666667\,x^2+x^4}\,,$$

con gli errori

$$\max_{|x|\le 2-\sqrt{3}}|\arctan x-R_{3,4}(x)|\approx 0.339\ 10^{-7},$$
$$\max_{|x|\le 2-\sqrt{3}}\Big|\frac{\arctan x-R_{3,4}(x)}{\arctan x}\Big|\approx 0.129\ 10^{-6},$$
$$\max_{|x|\le 2-\sqrt{3}}|\arctan x-R_{5,4}(x)|\approx 0.638\ 10^{-9},$$
$$\max_{|x|\le 2-\sqrt{3}}\Big|\frac{\arctan x-R_{5,4}(x)}{\arctan x}\Big|\approx 0.244\ 10^{-8}.$$

e) Si applichi l'algoritmo di Remez alla funzione

$$w_{3,4}(x) = \frac{\sum_{i=0}^{1} a_i x^{2i+1}}{\sum_{i=0}^{1} b_i x^{2i} + x^4}, \quad \text{con} \quad a_0 = b_0.$$

Si ottiene

$$w_{3,4}(x) = \frac{(12.07176 + 6.218014\, x^2)x}{12.07176 + 10.24193\, x^2 + x^4},$$

e risulta

$$\max_{|x| \le 2-\sqrt{3}} |\tan x - w_{3,4}(x)| \approx 0.697\ 10^{-9},$$

$$\max_{|x| \le 2-\sqrt{3}} \left| \frac{\tan x - w_{3,4}(x)}{\tan x} \right| \approx 0.266\ 10^{-8}.$$

f) Indicando con A il numero di operazioni additive e con M il numero di operazioni moltiplicative, il calcolo in un punto del polinomio ottenuto in b) richiede 3A+5M, di quello ottenuto in c) richiede 4A+6M, il calcolo delle approssimanti di Padé richiede 3A+5M per la prima, 4A+6M per la seconda, dell'approssimazione razionale ottenuta in e) richiede 3A+5M. Tenendo conto anche degli errori analitici trovati conviene usare l'approssimazione razionale ottenuta in e).)

6.85 a) Si verifichi che il calcolo di e^x per ogni x reale può essere ricondotto al calcolo di

$$2^x \quad \text{per} \quad x \in [-\beta^{-1}, 0],$$

dove β è la base usata nella rappresentazione dei numeri.

b) Per il caso $\beta = 2$, si determini il polinomio $p_5(x)$ di grado 5 di approssimazione minimax di 2^x per $x \in [-0.5, 0]$, tale che $p_5(0) = 1$.

c) Si scriva la formula di Padé $R_{3,3}(x)$ per 2^x.

d) Si determini la funzione razionale della forma

$$w_{2,2}(x) = \frac{p_2(x)}{q_2(x)},$$

in cui $p_2(x)$ e $q_2(x)$ sono due polinomi di grado 2, di approssimazione minimax di 2^x per $x \in [-0.5, 0]$ e tale che $w_{2,2}(0) = 1$.

e) Per ogni approssimazione trovata si stimino i massimi errori analitici assoluti e relativi e si determini quale approssimazione conviene usare anche sulla base del costo computazionale.

(Traccia: a) posto

$$\mathrm{e}^x = 2^{z/\beta}, \quad \text{con } z = \frac{\beta x}{\log 2} = r + s, \quad r = \lceil z \rceil,\ -1 < s \le 0,$$

risulta

$$\frac{z}{\beta} = u - v + \frac{s}{\beta},$$

dove $u = \left\lceil \dfrac{r}{\beta} \right\rceil$ è intero, $v = \dfrac{k}{\beta}$ per un intero k, $0 \le k \le \beta - 1$, e $\dfrac{s}{\beta} \in (-\beta^{-1}, 0]$. Quindi

$$\mathrm{e}^x = 2^u 2^{-v} 2^y, \quad y \in (-\beta^{-1}, 0].$$

Se $\beta = 2$, questa riduzione dell'intervallo richiede la memorizzazione delle costanti $\log 2$ e $2^{-1/2}$, che devono essere fornite con sufficiente precisione. La moltiplicazione per 2^u non viene effettivamente eseguita, perché comporta la sola modifica dell'esponente. Se $\beta \neq 2$, il numero delle costanti da memorizzare è maggiore.
b) Si applichi l'algoritmo di Remez a un polinomio della forma

$$p_5(x) = 1 + \sum_{i=1}^{5} a_i x^i,$$

scegliendo come nodi iniziali il punto -0.5 e 5 punti in $(-0.5, 0)$. Non si include fra i nodi iniziali il punto 0, per cui risulterebbe $d^{(0)} = 0$. Si ottiene

$$\begin{aligned} p_5(x) = 1 &+ 0.6931471\,x + 0.2402237\,x^2 + 0.0554773\,x^3 + 0.009508167\,x^4 \\ &+ 0.001126223\,x^5, \end{aligned}$$

e risulta

$$\max_{x \in [-0.5,0]} |2^x - p_5(x)| \approx 0.272\ 10^{-8}, \quad \max_{x \in [-0.5,0]} \left| \frac{2^x - p_5(x)}{2^x} \right| \approx 0.385\ 10^{-8}.$$

c) È

$$R_{3,3}(x) = \frac{120 + 60 \log 2\, x + 12 \log^2 2\, x^2 + \log^3 2\, x^3}{120 - 60 \log 2\, x + 12 \log^2 2\, x^2 - \log^3 2\, x^3},$$

che conviene riscrivere così

$$R_{3,3}(x) = \frac{(20.81369 + x^2) + x(7.213475 + 0.05776227\,x^2)}{(20.81369 + x^2) - x(7.213475 + 0.05776227\,x^2)}.$$

Risulta

$$\max_{x\in[-0.5,0]} |2^x - R_{3,3}(x)| \approx 0.698\ 10^{-8},\ \max_{x\in[-0.5,0]} \left|\frac{2^x - R_{3,3}(x)}{2^x}\right| \approx 0.924\ 10^{-8}.$$

d) Si applichi l'algoritmo di Remez alla funzione

$$w_{2,2}(x) = \frac{\sum_{i=0}^{1} a_i x^i + x^2}{\sum_{i=0}^{2} b_i x^i}, \quad \text{con} \quad b_0 = a_0.$$

Si ottiene

$$w_{2,2}(x) = \frac{(x + 9.141081)x + 27.12576}{(1.180681\, x - 9.66104)x + 27.12576},$$

e risulta

$$\max_{x\in[-0.5,0]} |2^x - w_{2,2}(x)| \approx 0.158\ 10^{-7},\ \max_{x\in[-0.5,0]} \left|\frac{2^x - w_{2,2}(x)}{2^x}\right| \approx 0.215\ 10^{-7}.$$

e) Indicando con A il numero di operazioni additive e con M il numero di operazioni moltiplicative, il calcolo in un punto del polinomio ottenuto in b) richiede 5A+5M, il calcolo sia della approssimante di Padé ottenuta in c) che dell'approssimazione razionale ottenuta in d) richiede 4A+4M. Tenendo conto anche degli errori analitici trovati conviene usare l'approssimante di Padé ottenuta in c).)

6.86 a) Si verifichi che il calcolo di $\log x$ per ogni $x > 0$ può essere ricondotto al calcolo di

$$f(x) = \log \frac{1+x}{1-x} \quad \text{per} \quad |x| \le \delta, \quad \delta = \frac{\sqrt{2}-1}{\sqrt{2}+1} = 0.1715729.$$

b) Si determini, con il metodo di economizzazione, un polinomio $p_7(x)$ di grado 7 che approssimi

$$\log \frac{1+x}{1-x} \quad \text{per} \quad |x| \le \delta,$$

tale che

(1) $p_7(x)$ sia una funzione dispari, e quindi $p_7(0) = 0$,

(2) $\lim_{x\to 0} \dfrac{p_7(x)}{x} = 2.$

c) Si determini il polinomio $p_5(x)$ di grado 5 di approssimazione minimax rispetto all'errore assoluto di $f(x)$, $|x| \le \delta$, che verifichi le condizioni del punto b).

d) Si determini il polinomio $q_5(x)$ di grado 5 di approssimazione minimax rispetto all'errore relativo di $f(x)$, $|x| \le \delta$, che verifichi le condizioni del punto b).

e) Si determini il polinomio $q_7(x)$ di grado 7 di approssimazione minimax rispetto all'errore relativo di $f(x)$, $|x| \le \delta$, che verifichi le condizioni del punto b).

f) Si determini la funzione razionale di approssimazione minimax rispetto all'errore assoluto di $f(x)$, $|x| \le \delta$, della forma

$$w_{3,2}(x) = \frac{p_3(x)}{q_2(x)}\,,$$

in cui $p_3(x)$ e $q_2(x)$ sono polinomi rispettivamente dispari di grado 3 e pari di grado 2, e tale che verifichi le condizioni del punto b).

g) Si determini la funzione razionale di approssimazione minimax rispetto all'errore relativo di $f(x)$, $|x| \le \delta$, della forma

$$v_{3,2}(x) = \frac{s_3(x)}{t_2(x)}\,,$$

in cui $s_3(x)$ e $t_2(x)$ sono polinomi rispettivamente dispari di grado 3 e pari di grado 2, e tale che verifichi le condizioni del punto b).

h) Per ogni approssimazione trovata si stimino i massimi errori analitici e si determini quale approssimazione conviene usare anche sulla base del costo computazionale.

(Traccia: a) Sia $x = \beta^p m$, dove β è la base usata nella rappresentazione dei numeri, p è l'esponente e m è la mantissa, con $\beta^{-1} \le m < 1$. Si supponga β potenza di 2 e si determini un intero dispari k tale che

$$2^{-(k+1)/2} \le m < 2^{-(k-1)/2},$$

e si ponga

$$z = \frac{2^{k/2}m - 1}{2^{k/2}m + 1}\,.$$

Risulta

$$-\frac{\sqrt{2}-1}{\sqrt{2}+1} \le z < \frac{\sqrt{2}-1}{\sqrt{2}+1}\,, \quad \text{e} \quad m = 2^{-k/2}\,\frac{1+z}{1-z}\,.$$

Se $\beta = 2^n$, è $\log x = \big(np - \dfrac{k}{2}\big)\log 2 + \log\dfrac{1+z}{1-z}$. Questa riduzione dell'intervallo risulta di facile applicazione se si può operare con istruzioni di macchina sulla prima cifra di m rappresentata nella base β. È richiesta la memorizzazione di alcune costanti: se $\beta = 2$ è $k = 1$ e occorre fornire $\sqrt{2}$ e $\log 2$, se $\beta = 16$ è $k \in \{1, 3, 5, 7\}$ e occorre fornire anche $\sqrt{8}$, $\sqrt{32}$, e $\sqrt{128}$, con sufficiente precisione.
b) Si considerino i primi 4 termini della serie di Maclaurin di

$$\frac{1}{x^2}\Big(\frac{f(x)}{2x} - 1\Big), \quad |x| \le \delta.$$

Si ottiene il polinomio

$$p(x) = \frac{1}{3} + \frac{x^2}{5} + \frac{x^4}{7} + \frac{x^6}{9}.$$

Con la trasformazione $x = t\,\delta$ e il metodo di economizzazione si determina un polinomio $q(t)$ di quarto grado, da cui si ottiene

$$p_7(x) = 2x + 2x^3 q\Big(\frac{x}{\delta}\Big) = 2x + 0.6666668\,x^3 + 0.3998917\,x^5 + 0.2955267\,x^7,$$

e risulta

$$\max_{|x|\le\delta} |f(x) - p_7(x)| \approx 0.189\ 10^{-8},\ \max_{|x|\le\delta}\Big|\frac{f(x) - p_7(x)}{f(x)}\Big| \approx 0.525\ 10^{-8}.$$

c) Si applichi l'algoritmo di Remez a un polinomio della forma

$$p_5(x) = 2x + a_0 x^3 + a_1 x^5,$$

scegliendo come nodi iniziali 2 punti in $(0, \delta)$ e il punto δ. Non si include fra i nodi iniziali il punto 0, per cui risulterebbe $d^{(0)} = 0$. Si ottiene

$$p_5(x) = 2x + 0.6665343\,x^3 + 0.4128748\,x^5,$$

e risulta

$$\max_{|x|\le\delta} |f(x) - p_5(x)| \approx 0.344\ 10^{-7}.$$

d) Procedendo come in c) si ottiene

$$q_5(x) = 2x + 0.6665562\,x^3 + 0.4120199\,x^5,$$

e risulta

$$\max_{|x|\le\delta}\Big|\frac{f(x) - q_5(x)}{f(x)}\Big| \approx 0.146\ 10^{-6}.$$

e) Procedendo come in d) si ottiene

$$q_7(x) = 2x + 0.6666678\,x^3 + 0.3997747\,x^5 + 0.2986992\,x^7,$$

e risulta

$$\max_{|x|\le\delta} \Big| \frac{f(x) - q_7(x)}{f(x)} \Big| \approx 0.863\ 10^{-9}.$$

f) Si applichi l'algoritmo di Remez alla funzione

$$w_{3,2}(x) = \frac{2a_0x + a_1x^3}{a_0 + x^2}\,.$$

Si ottiene

$$w_{3,2}(x) = \frac{(0.8945395\,x^2 - 3.316486)x}{x^2 - 1.658243}\,,$$

e risulta

$$\max_{|x|\le\delta} |f(x) - w_{3,2}(x)| \approx 0.555\ 10^{-8}.$$

g) Procedendo come al punto f) si ottiene

$$v_{3,2}(x) = \frac{(0.8941712\,x^2 - 3.317574)x}{x^2 - 1.658787}\,,$$

e risulta

$$\max_{|x|\le\delta} \Big| \frac{f(x) - v_{3,2}(x)}{f(x)} \Big| \approx 0.238\ 10^{-7}.$$

h) Indicando con A il numero di operazioni additive e con M il numero di operazioni moltiplicative, il calcolo in un punto del polinomio ottenuto in b) richiede 3A+5M, di quelli ottenuti in c) e d) richiede 2A+4M, di quello ottenuto in e) richiede 3A+5M, delle approssimazioni razionali ottenute in f) e g) richiede 2A+4M. Tenendo conto anche degli errori analitici trovati, conviene usare l'approssimazione razionale ottenuta in f) se si considerano gli errori assoluti e l'approssimazione razionale ottenuta in g) se si considerano gli errori relativi.)

Commento bibliografico

Il teorema di Weierstrass, fondamentale nella teoria dell'approssimazione polinomiale, è stato enunciato e dimostrato da Weierstrass nel 1885. Delle molte dimostrazioni fatte, la più elegante è considerata quella data da Bernstein nel 1912, in cui viene effettivamente costruita una successione di polinomi approssimanti (si veda l'esercizio 6.1). Un'altra dimostrazione

interessante è quella di Fejér del 1930, basata su una successione di polinomi di osculazione di Hermite (si veda l'esercizio 6.56).

Uno dei primi problemi affrontati nella teoria dell'approssimazione riguardava la risoluzione di un sistema lineare inconsistente: nel 1799 Laplace suggerì di risolverlo minimizzando il massimo modulo degli scarti, ma il metodo proposto era poco adatto per il calcolo. Nel 1804 Legendre affrontò il problema, minimizzando la somma dei quadrati degli scarti, secondo il metodo oggi detto dei minimi quadrati. Gauss, che rivendicò la paternità del metodo perché lo aveva usato con successo nel 1801 per determinare l'orbita del pianeta Cerere, fu il primo a collegare il metodo dei minimi quadrati alla teoria della probabilità. In norma 1 invece il problema dell'approssimazione lineare fu considerato da Chebyshev nel 1854 e non trovò soluzione fino al 1898, quando fu affrontato da Markoff.

Gli spazi di Hilbert sono la naturale generalizzazione dello spazio $\mathbf{R}^n$ ad un numero infinito di dimensioni. La loro introduzione permette di trattare in modo unitario un gran numero di casi particolari della teoria dell'approssimazione. Un'esposizione moderna e completa delle proprietà degli spazi di Hilbert è data nel libro di Rudin [42]. Una trattazione approfondita dell'approssimazione di Fourier in spazi di Hilbert è data nei libri di Achieser [2], Davis [14] e Zygmund [50]. Nel libro di Laurent [27] sono esaminate in dettaglio le connessioni fra la teoria dell'approssimazione e l'ottimizzazione negli spazi di Hilbert.

I polinomi ortogonali di Legendre, di Chebyshev, di Laguerre e di Hermite, introdotti fra la fine del 1700 e l'inizio del 1800, non portano tutti il nome del matematico che li ha introdotti per primo, ma in generale il nome di quello che li ha più estensivamente usati o ne ha studiato alcune delle proprietà: ad esempio i polinomi di Hermite erano già stati usati da Laplace e quelli di Laguerre da Lagrange. Molte proprietà dei polinomi di Legendre e di Chebyshev derivano dalle proprietà dei polinomi ultrasferici, che sono un caso particolare dei polinomi introdotti da Jacobi nel 1826 per studiare le formule di quadratura gaussiane. Altre proprietà, come quelle della relazione a tre termini e quella degli zeri, sono state trovate nel caso generale da Stieltjes nel 1884. Il termine "ortogonale" sembra sia stato usato per la prima volta da Schmidt nel 1905. Il testo classico sui polinomi ortogonali è quello di Szegö [46]. Utili e sintetiche descrizioni delle proprietà dei polinomi ortogonali più usati in pratica si possono trovare nei libri di Abramowitz, Stegun [1] e Davis [14]. Due testi di carattere generale che includono i polinomi ortogonali sono quelli di Atkinson [4] e Isaacson, Keller [25]. Una consistente raccolta di materiale grafico sui polinomi ortogonali può essere trovata nei libri di Abramowitz, Stegun [1] e di Gatteschi [21]. Per le proprietà dei polinomi di Chebyshev, si veda il libro di Rivlin [40]; per il calcolo numerico delle combinazioni lineari di polinomi ortogonali si

veda Rice [38].

Nel libro di Brezinsky [9], a seguito di un'ampia trattazione della teoria dei polinomi ortogonali generalizzati, sono descritte applicazioni anche nei settori dell'algebra lineare (metodi di Lanczos e del gradiente coniugato), del calcolo delle approssimanti di Padé e dei metodi di accelerazione della convergenza. Infine nei lavori di Canuto, Quarteroni [10] e di Dupont, Scott [17] vengono dati dei risultati, utili per la risoluzione di equazioni differenziali, dell'approssimazione con polinomi ortogonali in spazi di Sobolev.

Il problema dell'approssimazione minimax di una funzione fu inizialmente studiato da Chebyshev nel 1854, ma il teorema di equioscillazione che prende il suo nome fu in realtà completamente dimostrato solo da Borel nel 1905. De la Vallée-Poussin dimostrò l'omonimo teorema nel 1910. Trattazioni classiche della teoria del minimax sono quelle di Achieser [2], di Cheney [11], di Davis [14], di Meinardus [30] e di Rice [37]. La maggior parte dei testi che trattano dell'approssimazione minimax considerano uno spazio di funzioni continue che soddisfano a condizioni introdotte da Haar nel 1918, in cui vale ancora il teorema di equioscillazione (si veda l'esercizio 6.38). Esposizioni elementari possono essere trovate in Atkinson [4], Fike [18], Rivlin [39] e Snyder [45].

Il campo di applicazione del minimax, in particolare delle applicazioni delle proprietà di minimo dei polinomi di Chebyshev, si estende anche alla costruzione di metodi iterativi per il calcolo della soluzione di sistemi lineari: si veda Varga [48], Golub, Varga [22] e Hageman, Young [23]. Approssimazioni minimax polinomiali e razionali sono state inoltre utilizzate per la costruzione di algoritmi efficienti per la risoluzione di equazioni differenziali: si veda ad esempio Fox, Parker [19]. Per le relazioni che esistono con la teoria dell'ottimizzazione si vedano i libri di Laurent [27] e Dem'yanov, Malozemov [16]. Il classico algoritmo di Remez [36], descritto nel 1934 per il caso polinomiale, è stato successivamente generalizzato in varie direzioni: si veda Meinardus [30] e Ralston [35] per il caso razionale, Laurent [27] per il caso con vincoli e De Boor, Rice [15] per il caso in cui l'intervallo di definizione non è connesso. La dimostrazione che il metodo di Remez è del secondo ordine (si veda l'esercizio 6.48) è stata data da Veidinger nel 1960. Algoritmi per il calcolo del minimax diversi dall'algoritmo di Remez sono stati proposti da Bartels, Conn, Charalambous [6], Cline [12] e Dem'yanov, Malozemov [16]. Un'ampia esposizione dei metodi numerici per il calcolo del minimax e del quasi minimax polinomiale è riportata in Fraser [20]. Per il calcolo dell'approssimazione minimax razionale si può ricorrere al metodo di Maehly [29]. È anche possibile ottenere approssimazioni quasi minimax con un metodo di economizzazione analogo a quello del caso polinomiale [24] e [18]. In [1] si possono trovare le approssimazioni minimax delle funzioni più comuni. Con il nome di Jackson vengono indicati vari teoremi

che legano alle proprietà della $f(x)$ la velocità di convergenza in norma ∞ delle approssimazioni. Questi teoremi sono stati ottenuti da Jackson, da Bernstein e da altri a partire dal 1911. In Powell [33] e Shampine [44] sono dimostrate le maggiorazioni del teorema 6.42.

La prima espansione in frazione continua infinita è probabilmente quella data, senza dimostrazione, nel 1659 da Lord Brouncker per l'approssimazione di π. Nel 18^o secolo Eulero e poi Lambert gettarono le basi della teoria analitica delle frazioni continue; Lambert in particolare discusse anche problemi di convergenza. Successivamente vennero sviluppate nuove espansioni in frazioni continue in diversi campi della matematica, dalla teoria dei polinomi ortogonali alle formule di quadratura, allo studio del movimento dei pianeti. La prima definizione moderna di convergenza per frazioni continue fu data da Seidel nel 1846 (teorema 6.58). Nel 1898 Pringsheim dimostrò il teorema 6.59. Alla fine del 1800 Stieltjes, prendendo lo spunto da alcuni lavori di Chebyshev, trovò un numero considerevole di frazioni continue ottenute da serie di potenze e per il calcolo di integrali, e nel 1894 pubblicò una memoria in cui veniva studiata la convergenza di un particolare tipo di frazione continua, che prende il nome da lui. Successivamente Van Vleck nel 1903 e Hamburger nel 1920 estesero la teoria al caso di altri tipi di frazioni continue. Nel 1914 Hellinger e Toeplitz presentarono la teoria matriciale delle frazioni continue. I testi classici sulla teoria delle frazioni continue sono quello di Perron [31], nelle sue tre edizioni dal 1913 al 1957, e quello di Wall [49] del 1948.

L'approssimazione di Padé è stata in realtà studiata da Jacobi nel 1846 e da Frobenius nel 1881, anche se prende il nome di Padé che nel 1892 ebbe l'idea di costruire la tabella delle approssimanti e ne studiò le proprietà nella sua tesi. L'applicazione del metodo del qd al calcolo delle approssimanti di Padé è stata proposta da Rutishauser nel 1956 [43]. Molti metodi per calcolare le approssimazioni di Padé possono essere derivati dalla teoria dei polinomi ortogonali, si veda [9]. La complessità del calcolo delle approssimanti di Padé è stata studiata in [8]. Recentemente nella collana Encyclopedia of Mathematics and its Applications sono stati pubblicati i libri di Jones e Thron [26] e di Baker e Graves-Morris [5], che rappresentano lo stato dell'arte sulle frazioni continue e sull'approssimazione di Padé, e che elencano in bibliografia alcune centinaia di articoli sulla materia.

Bibliografia

[1] M. Abramowitz, I. A. Stegun, editors, *Handbook of Mathematical Functions*, National Bureau of Standards, U. S. Gov't Printing Office, 1964.

[2] N. I. Achieser, *Theory of Approximation*, Unger, New York, 1956.

[3] G. S. Ammar, W. E. Gragg, "Superfast Solution of Real Positive Definite Toeplitz Systems", *SIAM J. Matrix Analysis Appl.*, 9, 1988, pp. 61-76.

[4] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & Sons, New York, 1978.

[5] G. A. Baker, P. Graves-Morris, *Padé Approximants*, Encyclopedia of Mathematics and its Applications, vol. 13, 14, Addison-Wesley, Reading, 1981.

[6] R. H. Bartels, A. R. Conn, C. Charalambous, "On Cline's Direct Method for Solving Overdetermined Linear Systems in the l_∞ Sense", *Siam J. Numer. Anal.*, 15, 1978, pp. 255-270.

[7] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, Bologna, 1988.

[8] R. P. Brent, F. G. Gustavson, D. Y. Y. Yun, "Fast Solution of Toeplitz Systems of Equations and Computation of Padé Approximants", *J. of Algorithms*, 1, 1980, pp. 259-295.

[9] C. Brezinski, *Padé-Type Approximation and General Orthogonal Polynomials*, Birkhäuser Verlag, Basel, 1980.

[10] C. Canuto, A. Quarteroni, "Approximation Results for Orthogonal Polynomials in Sobolev Spaces", *Math. Comput.*, 38, 1982, pp. 67-86.

[11] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[12] A. K. Cline, "A Descent Method for the Uniform Solution to Overdetermined Systems of Linear Equations", *Siam J. Numer. Anal.*, 13, 1976, pp. 293-309.

[13] A. Cuyt, L. Wuytack, *Nonlinear Methods in Numerical Analysis*, North-Holland, Amsterdam, 1987.

[14] P. J. Davis, *Interpolation and Approximation*, Blaisdell, Pub. Co., New York, 1963.

[15] C. De Boor, J. R. Rice, "Extremal Polynomials with Application to Richardson Iteration for Indefinite Linear Systems", *SIAM J. Stat. and Sci. Comp.*, 3, 1982, pp. 47-57.

[16] V. F. Dem'yanov, V. N. Malozemov, *Introduction to Minimax*, John Wiley & Sons, New York, N. Y., 1974.

[17] T. Dupont, R. Scott, "Polynomial Approximation of Functions in Sobolev Spaces", *Math. Comput.*, 34, 1980, pp. 441-463.

[18] C. T. Fike, *Computer Evaluation of Mathematical Functions*, Prentice Hall, Englewood Cliffs, N. J., 1968.

[19] L. Fox, I. B. Parker, *Chebyshev Polynomials in Numerical Analysis*, Oxford Univ. Press, London, 1968.

[20] W. Fraser, "A Survey of Methods of Computing Minimax and Near-Minimax Polynomial Approximations for Functions of a Single Indipendent Variable", *J. ACM*, 12, 1965, pp. 295-314.

[21] L. Gatteschi, *Funzioni speciali*, UTET, Torino, 1973.

[22] G. H. Golub, R. S. Varga, "Chebyshev Semi-iterative Methods, Successive Overrelaxation Iterative Methods, and Second Order Richardson Iterative Methods, Parts I and II", *Numer. Math.*, 3, 1961, pp. 147-168.

[23] L. A. Hageman, D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[24] D. C. Handscomb, *Methods of Numerical Approximation*, Pergamon Press, Oxford, 1966.

[25] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[26] W. B. Jones, W. J. Thron, *Continued Fractions, Analytic Theory and Applications*, Encyclopedia of Mathematics and its Applications, vol. 11, Addison-Wesley, Reading, 1980.

[27] P. J. Laurent, *Approximation et Optimization*, Hermann, Paris, 1972.

[28] N. N. Lebedev, *Special Functions and Their Applications*, Dover Publ., New York, 1972.

[29] H. J. Maehly, "Methods for Fitting Rational Approximations. Part I: Telescoping Procedures for Continued Fractions", *J. ACM*, 7, 1960, pp. 150-162.

[30] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer-Verlag, Berlin, 1967.

[31] O. Perron, *Die Lehre von den Kettenbrüchen*, Band I, II, Teubner, Stuttgart, 1954, 1957.

[32] P. P. Petrushev, V. A. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, 1987.

[33] M. J. D. Powell, "On the Maximum Errors of Polynomial Approximations Defined by Interpolation and by Least Squares Criteria", *Comput. J.*, 9, 1967, pp. 404-407.

[34] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986.

[35] A. Ralston, "Rational Chebyshev Approximation by Remes' Algorithms", *Numer. Math.*, 7, 1965, pp. 322-330.

[36] E. J. Remes, "Sur le calcul effectif des polynomes d'approximation de Tchebichef", *Compt. Rend. Acad. Sci. Paris*, 1 99, 1934, pp. 337-340.

[37] J. R. Rice, *The Approximation of Functions, vol. 1. Linear Theory*, Addison-Wesley, Reading, Mass., 1964.

[38] J. R. Rice, "On the Conditioning of Polynomial and Rational Forms", *Numer. Math.*, 7, 1965, pp. 426-435.

[39] T. J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, New York, 1969.

[40] T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley & Sons, New York, 1974.

[41] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[42] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.

[43] H. Rutishauser, *Der Quotienten-Differenzen-Algorithmus*, Birkhäuser Verlag, Basel, 1956.

[44] L. F. Shampine, "Efficiency of a Procedure for Near-Minimax Approximation", *J. ACM*, 17, 1970, pp. 655-660.

[45] M. A. Snyder, *Chebyshev Methods in Numerical Approximation*, Prentice Hall, Englewood Cliffs, N. J., 1966.

[46] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc., Providence, R. I., 1959.

[47] W. F. Trench, "An Algorithm for the Inversion of Finite Toeplitz Matrices", *SIAM J.*, 12, 1964, pp. 515-522.

[48] R. S. Varga, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, N. J., 1962.

[49] H. S. Wall, *Analytic Theory of Continued Fractions*, Van Nostrand, New York, 1948.

[50] A. Zygmund, *Trigonometric Series*, Cambridge Press, Cambridge, 1959.

Capitolo 7

INTEGRAZIONE E DERIVAZIONE APPROSSIMATE

1. Formule di quadratura interpolatorie

Sia $f(x)$ una funzione reale definita su un intervallo $[a, b]$. Il problema che si vuole studiare è quello della approssimazione dell'integrale

$$S = \int_a^b f(x)\ dx. \tag{1}$$

Nel caso in cui $f(x)$ sia una funzione continua, il teorema fondamentale del calcolo integrale assicura l'esistenza su $[a, b]$ di una funzione $F(x)$, detta *primitiva* della $f(x)$ tale che

$$S = F(b) - F(a). \tag{2}$$

Non sempre però la $F(x)$ è esprimibile in termini di funzioni elementari, e anche quando questo è possibile, il calcolo di $F(a)$ e $F(b)$ nella (2) può essere difficoltoso. Per questo è importante disporre di tecniche numeriche per il calcolo approssimato della (1). Con le formule di integrazione approssimata è possibile trattare, oltre al caso delle funzioni continue su un intervallo limitato, anche i casi di funzioni con singolarità o di funzioni su intervalli illimitati o di funzioni i cui valori sono noti solo in un insieme discreto di punti.

Poiché le tecniche numeriche utilizzano in generale solo i valori della funzione su un insieme finito di punti dell'intervallo, senza usare altre informazioni sul comportamento analitico della $f(x)$, non è possibile stimare in alcun modo la bontà dell'approssimazione ottenuta. Come esempio della difficoltà del problema si veda la figura 7.1, in cui sono tracciati i grafici di diverse funzioni $f(x)$, tutte passanti per lo stesso insieme di punti.

7.1 Definizioni. Le formule di *integrazione approssimata* o di *quadratura* che si considerano sono in generale della forma

$$S_{n+1} = \sum_{i=0}^{n} w_i f(x_i), \tag{3}$$

dove gli $n + 1$ punti $x_i \in [a, b]$ per $i = 0, \ldots, n$, sono i *nodi* della formula e i numeri w_i, $i = 0, \ldots, n$, sono i *coefficienti* (o *pesi*) della formula.

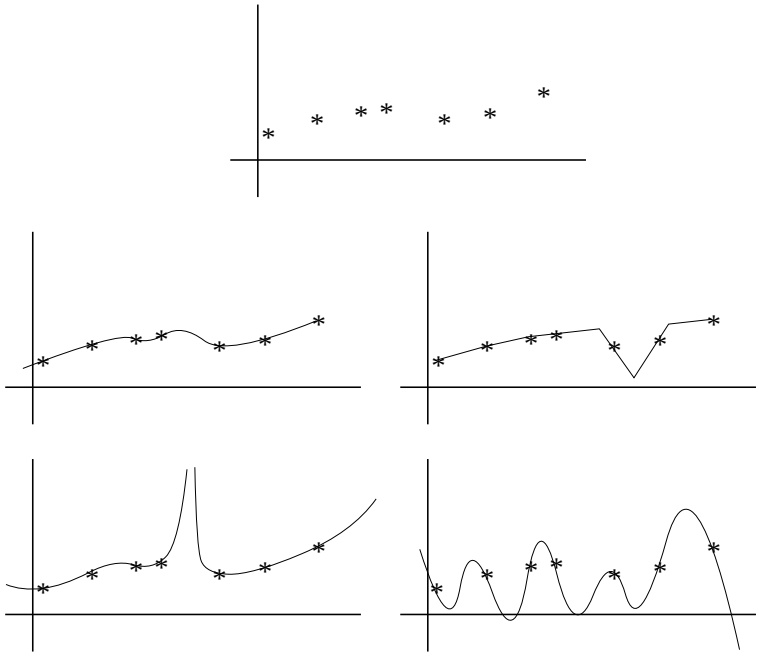


Fig. 7.1 - Funzioni passanti per gli stessi punti.

I coefficienti w_i non dipendono dalla funzione $f(x)$, ma dipendono oltre che da n, anche dalla scelta dei nodi. La quantità

$$r_{n+1} = S - S_{n+1},$$

che rappresenta l'errore analitico assoluto, è detta *resto* della formula di quadratura. Si dice che la formula di quadratura (3) ha *grado di precisione* k se risulta $r_{n+1} = 0$ quando $f(x) = x^j$, per $j = 0, \ldots, k$, e $r_{n+1} \neq 0$ quando $f(x) = x^{k+1}$. ∎

Per la linearità dell'operatore integrale, una formula con grado di precisione k dà il risultato esatto ($r_{n+1} = 0$) quando è applicata ad una funzione che sia un polinomio di grado minore o uguale a k.

Una delle tecniche più usate per ricavare formule di quadratura (3) consiste nel sostituire alla funzione $f(x)$ un polinomio $p(x)$ che l'approssimi e quindi considerare l'integrale

$$\int_a^b p(x)\ dx$$

come una approssimazione di S. Fissati $n + 1$ punti distinti $x_0, \ldots, x_n$ dell'intervallo $[a, b]$, se si considera come polinomio approssimante la $f(x)$

il polinomio di interpolazione di Lagrange

$$p(x) = \sum_{i=0}^{n} L_i(x) f(x_i),$$

risulta

$$\begin{aligned} S_{n+1} = \int_a^b p(x)\ dx = \int_a^b \sum_{i=0}^{n} L_i(x) f(x_i)\ dx = \sum_{i=0}^{n} w_i f(x_i), \\ \text{dove} \quad w_i = \int_a^b L_i(x)\ dx = \int_a^b \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\ dx. \end{aligned} \tag{4}$$

La (4) è una formula di quadratura della forma (3). Poiché per la (8, cap. 5), è

$$L_i(x) = \frac{\pi_n(x)}{(x - x_i)\pi_n'(x_i)}\,,$$

risulta anche

$$w_i = \frac{1}{\pi_n'(x_i)} \int_a^b \frac{\pi_n(x)}{x - x_i}\ dx, \quad i = 0, \ldots, n. \tag{5}$$

7.2 Definizione. Siano $x_0, \ldots, x_n$, $n+1$ punti distinti in $[a, b]$. Una formula di quadratura è detta *formula interpolatoria* se i suoi coefficienti w_i, $i = 0, \ldots, n$, sono della forma (5). ∎

Fissati gli $n+1$ nodi distinti $x_0, \ldots, x_n$ in $[a, b]$, i coefficienti di formule di quadratura (3) possono essere ricavati anche con il *metodo dei coefficienti indeterminati*, imponendo la condizione che la formula cercata abbia grado di precisione k, cioè che sia tale che

$$\sum_{i=0}^{n} w_i x_i^j = \int_a^b x^j\ dx = \frac{b^{j+1} - a^{j+1}}{j+1}\,, \quad j = 0, \ldots, k. \tag{6}$$

I coefficienti w_i sono quindi soluzione di un sistema lineare la cui matrice dei coefficienti ha gli elementi

$$x_i^j, \quad \text{per} \quad i = 0, \ldots, n,\ j = 0, \ldots, k.$$

Se $k = n$, la matrice risulta una matrice di Vandermonde e poiché gli x_i sono distinti, la matrice è non singolare, e quindi, comunque siano stati scelti i nodi x_i, $i = 0, \ldots, n$, esiste sempre una e una sola formula di precisione almeno n costruita con quei nodi.

7.3 Teorema. *La formula di quadratura S_{n+1} della forma (3), i cui coefficienti w_i sono ottenuti risolvendo il sistema lineare (6) per $k = n$, e che ha quindi grado di precisione almeno n, coincide con la (5) e cioè è interpolatoria.*

Dim. Se la formula (3) è interpolatoria si ha

$$r_{n+1} = \int_a^b f(x)\ dx - \int_a^b p(x)\ dx = \int_a^b [f(x) - p(x)]\ dx. \tag{7}$$

Se $f(x) = x^j$, $j = 0, \dots, n$, il polinomio $p(x)$ di grado minore od uguale a n coincide con la $f(x)$ e quindi $r_{n+1} = 0$. Ne segue che la formula (3) con i coefficienti dati dalla (5) ha grado di precisione almeno n. D'altra parte, poiché la matrice del sistema (6) è non singolare per $k = n$, il sistema (6) ha una sola soluzione, le cui componenti w_i coincidono con le (5). ∎

Dal teorema 7.3 segue che una formula di quadratura S_{n+1} interpolatoria ha ordine di precisione almeno n.

7.4 Teorema. *Se i nodi $x_0, \dots, x_n$ di una formula di quadratura interpolatoria S_{n+1} sono disposti in modo simmetrico rispetto al punto di mezzo dell'intervallo $[a, b]$, cioè*

$$x_i + x_{n-i} = a + b, \quad i = 0, \dots, n, \tag{8}$$

allora i coefficienti relativi a nodi simmetrici sono uguali tra di loro, cioè

$$w_i = w_{n-i}, \quad i = 0, \dots, n.$$

Dim. Per la (8) si ha

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - (a + b - x_{n-j})}{x_i - (a + b - x_{n-j})}$$

$$= \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(a + b - x) - x_{n-j}}{(a + b - x_i) - x_{n-j}},$$

e ponendo $k = n - j$, risulta

$$L_i(x) = \prod_{\substack{k=0 \\ k \neq n-i}}^{n} \frac{(a + b - x) - x_k}{x_{n-i} - x_k} = L_{n-i}(a + b - x).$$

Quindi

$$w_i = \int_a^b L_i(x)\ dx = \int_a^b L_{n-i}(a + b - x)\ dx,$$

e con la trasformazione di variabile $y = a + b - x$, risulta

$$w_i = -\int_b^a L_{n-i}(y)\ dy = \int_a^b L_{n-i}(y)\ dy = w_{n-i}.$$ ∎

Utilizzando l'espressione del resto del polinomio di interpolazione è possibile esprimere il resto della formula di quadratura interpolatoria.

7.5 Teorema. *Per una formula di quadratura interpolatoria S_{n+1} si ha:*

$$r_{n+1} = \int_a^b \pi_n(x) f[x_0, \ldots, x_n, x]\ dx, \tag{9}$$

dove $\pi_n(x) = (x - x_0) \ldots (x - x_n)$ e $f[x_0, \ldots, x_n, x]$ è la differenza divisa di ordine $n + 1$ definita nella 5.14, e se $f(x) \in C^{n+1}[a, b]$, allora

$$r_{n+1} = \frac{1}{(n+1)!} \int_a^b \pi_n(x) f^{(n+1)}(\xi)\ dx, \quad \xi = \xi(x) \in (a, b). \tag{10}$$

Dim. La (9) e la (10) seguono immediatamente sostituendo nella (7) le relazioni (22) e (23) del teorema 5.19. ∎

Nel caso che la funzione $f(x)$ sia un polinomio di grado j, la differenza divisa $f[x_0, \ldots, x_n, x]$ è, per il teorema 5.26, un polinomio $p_{j-n-1}(x)$ di grado $j - n - 1$ se $j > n$ e se $j \leq n$ tale differenza è nulla. In particolare se $f(x) = x^j$ allora il polinomio $p_{j-n-1}(x)$ ha primo coefficiente uguale a 1 e

$$r_{n+1} = \begin{cases} 0 & \text{se } j \leq n, \\ \displaystyle\int_a^b \pi_n(x) p_{j-n-1}(x)\ dx & \text{se } j > n. \end{cases} \tag{11}$$

7.6 Teorema. *Il massimo grado di precisione di una formula interpolatoria S_{n+1} è $2n + 1$. Una tale formula può essere costruita solamente scegliendo come nodi gli $n + 1$ zeri dell'$(n + 1)$-esimo polinomio ortogonale, rispetto alla funzione peso $\omega(x) = 1$, nell'intervallo $[a, b]$*

Dim. Sia k il grado di precisione della formula di quadratura, per cui $r_{n+1} = 0$ per $f(x) = x^j$, $j = 0, \ldots, k$. Se $k > n$, dalla (11) si ha

$$\int_a^b \pi_n(x) p_{j-n-1}(x)\ dx = 0, \quad \text{per} \quad j = n + 1, \ldots, k, \tag{12}$$

cioè il polinomio $\pi_n(x)$ risulta ortogonale, nell'intervallo $[a, b]$, ai polinomi $p_r(x)$, $r = 0, \ldots, k - n - 1$, che sono linearmente indipendenti. Perciò $\pi_n(x)$ risulta ortogonale a tutti i polinomi di grado r minore o uguale a $k - n - 1$. Poiché $\pi_n(x)$ ha grado $n + 1$, il massimo valore che $k - n - 1$ può assumere è n, e quindi il massimo valore di k per cui può valere la (12) è $2n + 1$. Il polinomio $\pi_n(x)$ deve coincidere, a meno di un fattore γ costante, con l'$(n + 1)$-esimo polinomio ortogonale nell'intervallo $[a, b]$. Indicato con $p_{n+1}(x)$ tale polinomio, si ha infatti dal teorema 6.11 che

$$\int_a^b p_{n+1}(x) q(x)\ dx = 0,$$

per ogni polinomio $q(x)$ di grado minore o uguale ad n. Quindi gli zeri di $\pi_n(x)$, cioè i nodi $x_0, \ldots, x_n$, devono coincidere con gli zeri di $p_{n+1}(x)$. ∎

A differenza di quanto accade per i polinomi di interpolazione, per le formule di quadratura si può dare un semplice teorema di *convergenza* al crescere del numero dei nodi.

7.7 Teorema. *Sia $f \in C[a,b]$ e sia $\{S_{n+1}\}$ una successione di formule di quadratura interpolatorie, con*

$$S_{n+1} = \sum_{i=0}^{n} w_i^{(n)} f(x_i^{(n)}),$$

tali che esista una costante H per cui

$$\sum_{i=0}^{n} |w_i^{(n)}| < H, \quad \textit{per ogni} \quad n,$$

allora

$$\lim_{n\to\infty} r_{n+1} = 0.$$

Dim. Per ogni $\epsilon > 0$ per il teorema di Weierstrass 6.1 esiste un polinomio $p(x)$ tale che

$$|f(x) - p(x)| \le \epsilon, \quad \text{per ogni} \quad x \in [a,b]. \tag{13}$$

Si ha

$$\begin{aligned} r_{n+1} &= \int_a^b f(x)\ dx - S_{n+1} \\ &= \int_a^b [f(x) - p(x)]\ dx + \int_a^b p(x)\ dx - \sum_{i=0}^{n} w_i^{(n)} f(x_i^{(n)}). \end{aligned} \tag{14}$$

Se m è il grado di $p(x)$, per tutti gli $n \ge m$ la formula di quadratura S_{n+1} ha grado di precisione almeno m, per cui si ha

$$\int_a^b p(x)\ dx = \sum_{i=0}^{n} w_i^{(n)} p(x_i^{(n)}),$$

e, sostituendo nella (14), è

$$r_{n+1} = \int_a^b [f(x) - p(x)]\ dx + \sum_{i=0}^{n} w_i^{(n)} [p(x_i^{(n)}) - f(x_i^{(n)})].$$

Passando ai moduli, per la (13) è

$$|r_{n+1}| \leq \epsilon\Big[\int_a^b dx + \sum_{i=0}^{n} |w_i^{(n)}|\Big],$$

e quindi per ogni $n \geq m$ vale $|r_{n+1}| \leq \epsilon[b-a+H]$, da cui la tesi. ∎

Importante conseguenza del teorema 7.7 è che ogni successione $\{S_{n+1}\}$ di formule di quadratura a coefficienti positivi è convergente. Infatti poiché la formula ha grado di precisione almeno 0, applicando la (3) alla funzione $f(x) = 1$, si ha

$$\sum_{i=0}^{n} |w_i| = \sum_{i=0}^{n} w_i = \int_a^b dx = b - a.$$

Il teorema seguente mostra come le formule ad alto grado di precisione abbiano coefficienti positivi e quindi godano della proprietà della convergenza.

7.8 Teorema. *I coefficienti di una formula interpolatoria S_{n+1}, avente grado di precisione almeno $2n$, sono tutti positivi.*

Dim. Siano $x_0, \ldots, x_n$ i nodi della formula di quadratura interpolatoria S_{n+1}. Per $i = 0, \ldots, n$ si consideri il polinomio di grado n

$$q_i(x) = \prod_{\substack{r=0 \\ r \neq i}}^{n} (x - x_r).$$

Poiché la formula S_{n+1} ha grado di precisione almeno $2n$, essa integra esattamente i polinomi $q_i^2(x)$, $i = 0, \ldots, n$, di grado $2n$, cioè

$$\sum_{j=0}^{n} w_j q_i^2(x_j) = \int_a^b q_i^2(x)\, dx, \quad i = 0, \ldots, n.$$

Poiché

$$q_i^2(x_j) = \begin{cases} 0 & \text{se } i \neq j, \\ \displaystyle\prod_{\substack{r=0 \\ r \neq i}}^{n} (x_i - x_r)^2 & \text{se } j = i, \end{cases}$$

ne segue che

$$w_i \prod_{\substack{r=0 \\ r \neq i}}^{n} (x_i - x_r)^2 = \int_a^b q_i^2(x)\, dx, \quad i = 0, \ldots, n,$$

e quindi $w_i > 0$ per $i = 0, \ldots, n$. ∎

Dal teorema 7.7 risulta perciò che esistono successioni doppie $\{x_i^{(n)}, w_i^{(n)}\}_{(i,n)\in\mathbf{N}}$ di nodi distinti e di pesi, tali che

$$\lim_{n\to\infty} \sum_{i=0}^{n} w_i^{(n)} f(x_i^{(n)}) = \int_a^b f(x)\, dx$$

per ogni $f(x) \in C[a, b]$. Dal punto di vista computazionale sarebbe molto più conveniente se una simile proprietà di convergenza valesse anche per successioni di nodi e pesi dipendenti da un solo indice. Questo però non è vero, in quanto non esiste alcuna successione $\{x_i, w_i\}_{i\in\mathbf{N}}$ per cui sia

$$\lim_{n\to\infty} \sum_{i=0}^{n} w_i f(x_i) = \int_a^b f(x)\, dx$$

per ogni $f(x) \in C[a, b]$ (si veda l'esercizio 7.52).

Le formule di quadratura interpolatorie vengono di solito divise nel modo seguente:

a) i punti $x_0, \ldots, x_n$ sono prefissati nell'intervallo $[a, b]$. A questa classe appartengono le *formule newtoniane* o *di Newton-Cotes*, che si ottengono scegliendo i nodi x_i equidistanti. Queste formule, che hanno grado di precisione n oppure $n+1$ e coefficienti facilmente ricavabili ed espressi con semplici numeri razionali, hanno lo svantaggio che per $n \geq 8$ i coefficienti non sono tutti dello stesso segno.

b) i punti $x_0, \ldots, x_n$ non sono prefissati, i coefficienti e i nodi vengono ricavati in modo da massimizzare il grado di precisione che risulta $2n+1$. Queste formule, dette *gaussiane*, presentano rispetto alle formule newtoniane, oltre all'elevato grado di precisione, il vantaggio di avere coefficienti sempre positivi, anche se nodi e coefficienti sono espressi con numeri non razionali.

c) alcuni nodi sono prefissati, i nodi rimanenti e i coefficienti vengono ricavati in modo da massimizzare il grado di precisione. Per certe formule di questa classe continua a valere la proprietà che i coefficienti sono positivi per ogni n.

Altre formule, dette formule *a coefficienti uniformi*, vengono costruite in modo da avere coefficienti tutti uguali. Esse presentano dei vantaggi dal punto di vista della propagazione degli errori di arrotondamento, però come le formule della classe b), hanno nodi espressi con numeri non razionali. Inoltre queste formule non possono essere ricavate per $n = 8$ e $n \geq 10$ e non è nota un'espressione generale del resto.

2. Formule di Newton-Cotes

Posto $h = (b-a)/n$, siano $x_i = a + ih$, $i = 0, \ldots, n$, i nodi, equidistanti di passo h nell'intervallo $[a, b]$, sui quali si costruisce la formula di quadratura interpolatoria S_{n+1}, detta *formula di Newton-Cotes degli* $n+1$ *punti*. Per determinare i coefficienti conviene eseguire il cambiamento di variabile

$$x = a + th, \quad 0 \le t \le n. \tag{15}$$

Si ha allora dalla (4)

$$w_i = h\alpha_i,$$

dove

$$\alpha_i = \int_0^n \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t-j}{i-j}\, dt, \tag{16}$$

per cui la formula di quadratura di Newton-Cotes può essere così scritta

$$S_{n+1} = h \sum_{i=0}^{n} \alpha_i f(x_i),$$

dove i coefficienti α_i, $i = 0, \ldots, n$, dati dalla (16), dipendono solo da i e da n, ma non dai nodi e quindi possono essere tabulati. Per il teorema 7.4, poiché i nodi x_i sono simmetrici rispetto al punto di mezzo dell'intervallo $[a, b]$, i coefficienti α_i sono tali che $\alpha_i = \alpha_{n-i}$.

7.9 Esempio. Per $n = 1$, posto $x_0 = a$ e $x_1 = b$, $h = b - a$, si ha dalla (16)

$$\alpha_0 = \int_0^1 (1-t)\, dt = \frac{1}{2}, \quad \alpha_1 = \alpha_0,$$

da cui si ottiene la formula di quadratura dei *due punti*

$$S_2 = \frac{h}{2}\, [f(x_0) + f(x_1)] \tag{17}$$

per l'approssimazione dell'integrale (1). L'interpretazione geometrica della (17) è illustrata nella figura 7.2: il valore dell'integrale è approssimato con quello dell'area (in grigio) del trapezio di basi $f(a)$ e $f(b)$ e altezza h. Il resto dell'approssimazione è dato dall'area della superficie compresa tra il grafico della funzione $f(x)$ e il segmento che ha per estremi $(a, f(a))$ e $(b, f(b))$.

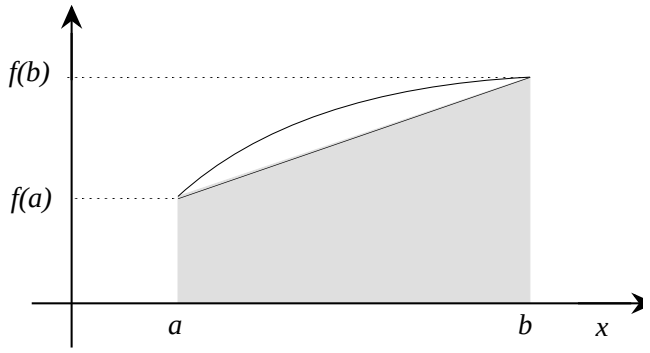


Fig. 7.2 - Approssimazione con la formula dei due punti.

Per $n = 2$, posto $x_0 = a$ e $x_2 = b$, $h = (b-a)/2$, si ha dalla (16)

$$\alpha_0 = \int_0^2 \frac{1}{2}\,(t-1)(t-2)\,dt = \frac{1}{3}, \quad \alpha_1 = \int_0^2 t(2-t)\,dt = \frac{4}{3}, \quad \alpha_2 = \alpha_0,$$

da cui si ottiene la formula di quadratura dei *tre punti*

$$S_3 = \frac{h}{3}\,[f(x_0) + 4f(x_1) + f(x_2)] \tag{18}$$

per l'approssimazione dell'integrale (1). ∎

Poiché le formule di Newton-Cotes sono interpolatorie, per esse vale l'espressione generale del resto data nel teorema 7.5. In questo caso poiché i punti $x_0, \ldots, x_n$ sono equidistanti, si può ottenere un'espressione del resto assai più semplice.

7.10 Teorema. *Sia S_{n+1} una formula di Newton-Cotes, allora*

a) se n è pari e $f(x) \in C^{n+2}[a,b]$, allora esiste un punto $\xi \in (a,b)$ tale che:

$$r_{n+1} = \frac{f^{(n+2)}(\xi)}{(n+2)!} \int_a^b x\pi_n(x)\,dx, \tag{19}$$

b) se n è dispari e $f(x) \in C^{n+1}[a,b]$, allora esiste un punto $\xi \in (a,b)$ tale che:

$$r_{n+1} = \frac{f^{(n+1)}(\xi)}{(n+1)!} \int_a^b \pi_n(x)\,dx. \tag{20}$$

Dim. Si esamina dapprima il caso in cui n è pari; con il cambiamento di variabile (15) si ha

$$\pi_n(x) = h^{n+1}\tau_n(t),$$

dove $\tau_n(t) = t(t-1)\dots(t-n)$. Del polinomio $\tau_n(t)$ di grado $n+1$ sono state studiate nel paragrafo 3 del capitolo 5 alcune proprietà che saranno utili in questa dimostrazione:

1) $\tau_n(t)$ è antisimmetrico rispetto al punto $n/2$,

2) per $t \le n/2$ è $|\tau_n(t-1)| > |\tau_n(t)|$. (21)

Indicata allora con

$$\sigma_n(x) = \int_a^x \pi_n(u)\ du,$$

una primitiva della $\pi_n(x)$, è

$$\sigma_n(x) = \sigma_n(a+th) = h^{n+2} \int_0^t \tau_n(v)\ dv.$$

Il polinomio $\sigma_n(x)$ di grado $n+2$ è simmetrico rispetto al punto $\dfrac{a+b}{2}$ e quindi

$$\sigma_n(a) = \sigma_n(b) = 0.$$

Inoltre per $a < x < b$ è

$$\sigma_n(x) > 0. \tag{22}$$

La (22) segue dal fatto che per $0 < t < 1$ è $\tau_n(t) > 0$ e quindi

$$\int_0^t \tau_n(v)\ dv > 0,$$

e che il contributo dei successivi intervalli è per la (21) sempre decrescente in modulo al crescere di t fino ad $n/2$. Per $t > n/2$, la permanenza del segno è assicurata dalla simmetria della $\tau_n(t)$. La figura 7.3 riporta il grafico della funzione $\int_0^t \tau_{10}(u)\ du$ (si confronti con il grafico di $\tau_{10}(t)$ riportato nella figura 5.4).

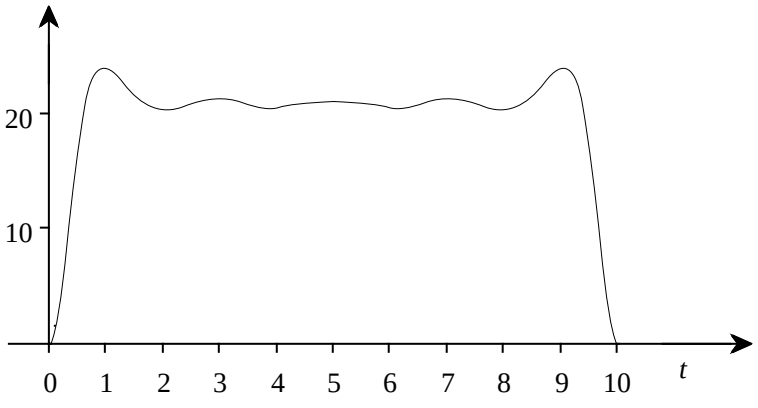


Fig. 7.3 - Grafico della funzione $\int_0^t \tau_{10}(v)\ dv$

Si può ora passare alla dimostrazione della (19). Dalla (9) si ha:

$$r_{n+1} = \int_a^b \pi_n(x) f[x_0, \ldots, x_n, x]\ dx,$$

in cui $f[x_0, \ldots, x_n, x]$ è derivabile con continuità per il teorema 5.28. Integrando per parti si ha

$$r_{n+1} = \Big[\sigma_n(x) f[x_0, \ldots, x_n, x]\Big]_a^b - \int_a^b \sigma_n(x) \frac{d}{dx}\ f[x_0, \ldots, x_n, x]\ dx.$$

Per quanto visto sopra è $\sigma_n(a) = \sigma_n(b) = 0$ e per il teorema 5.31 è

$$\frac{d}{dx}\ f[x_0, \ldots, x_n, x] = \frac{f^{(n+2)}(\zeta)}{(n+2)!}\,,$$

dove $\zeta = \zeta(x) \in (a, b)$, per cui

$$r_{n+1} = -\int_a^b \sigma_n(x)\ \frac{f^{(n+2)}(\zeta)}{(n+2)!}\ dx. \tag{23}$$

Per la (22) $\sigma_n(x)$ non cambia segno in $[a, b]$, è quindi possibile applicare alla (23) il teorema della media integrale, ottenendo

$$r_{n+1} = -\frac{f^{(n+2)}(\xi)}{(n+2)!}\ \int_a^b \sigma_n(x)\ dx, \tag{24}$$

dove $\xi \in (a, b)$. Inoltre integrando per parti si ottiene la relazione

$$\int_a^b \sigma_n(x)\ dx = \Big[x\sigma_n(x)\Big]_a^b - \int_a^b x\ \frac{d\sigma_n(x)}{dx}\ dx = -\int_a^b x\pi_n(x)\ dx,$$

che sostituita nella (24) dà la (19).

Se n è dispari si può considerare l'intervallo $[a, b]$ come unione dei due intervalli $[a, b-h]$ e $[b-h, b]$, dove $b - h = x_{n-1}$, per cui

$$\begin{aligned} r_{n+1} &= \int_a^b \pi_n(x) f[x_0, \ldots, x_n, x]\ dx \\ &= \int_a^{b-h} \pi_n(x) f[x_0, \ldots, x_n, x]\ dx + \int_{b-h}^b \pi_n(x) f[x_0, \ldots, x_n, x]\ dx. \end{aligned} \tag{25}$$

Per quanto riguarda il secondo integrale della (25), per il teorema 5.31 è

$$f[x_0, \ldots, x_n, x] = \frac{f^{(n+1)}(\zeta)}{(n+1)!}\,,$$

dove $\zeta = \zeta(x) \in (b-h, b)$. Poiché $\pi_n(x)$ nell'intervallo $[b-h, b]$ non cambia segno, per il teorema della media integrale è

$$\int_{b-h}^{b} \pi_n(x) f[x_0, \ldots, x_n, x]\ dx = \frac{f^{(n+1)}(\eta)}{(n+1)!}\ \int_{b-h}^{b} \pi_n(x)\ dx, \tag{26}$$

dove $\eta \in (b-h, b)$ Per quanto riguarda il primo integrale della (25), per la definizione di differenza divisa si ha

$$\begin{aligned}\pi_n(x) f[x_0, \ldots, x_n, x] &= \pi_n(x)\ \frac{f[x_0, \ldots, x_{n-1}, x] - f[x_0, \ldots, x_{n-1}, x_n]}{x - x_n} \\ &= \pi_{n-1}(x)\big\{f[x_0, \ldots, x_{n-1}, x] - f[x_0, \ldots, x_{n-1}, x_n]\big\};\end{aligned}$$

poiché $n-1$ è pari si può procedere come nel caso precedente, ottenendo

$$\begin{aligned}\int_a^{b-h} \pi_n(x) f[x_0, \ldots, x_n, x]\ dx &= \int_a^{b-h} \pi_{n-1}(x) f[x_0, \ldots, x_{n-1}, x]\ dx \\ &\quad - f[x_0, \ldots, x_{n-1}, x_n] \int_a^{b-h} \pi_{n-1}(x)\ dx \\ &= -\frac{f^{(n+1)}(\theta)}{(n+1)!}\ \int_a^{b-h} \sigma_{n-1}(x)\ dx, \end{aligned} \tag{27}$$

con $\theta \in (a, b-h)$, dato che

$$\int_a^{b-h} \pi_{n-1}(x)\ dx = 0,$$

per l'antisimmetria di $\pi_{n-1}(x)$ rispetto al punto di mezzo dell'intervallo $[a, b-h]$.

Sostituendo le (26) e (27) nella (25) si ha

$$r_{n+1} = \frac{1}{(n+1)!}\ [\alpha f^{(n+1)}(\theta) + \beta f^{(n+1)}(\eta)], \tag{28}$$

dove

$$\alpha = -\int_a^{b-h} \sigma_{n-1}(x)\ dx \quad \text{e} \quad \beta = \int_{b-h}^{b} \pi_n(x)\ dx.$$

Poiché $\sigma_{n-1}(x) \geq 0$ per $a \leq x \leq b-h$ e $\pi_n(x) \leq 0$ per $b-h \leq x \leq b$, ne segue che le due costanti α e β sono entrambe negative, per cui per la continuità di $f^{(n+1)}(x)$ esiste un punto ξ con $\theta \leq \xi \leq \eta$, tale che

$$\alpha f^{(n+1)}(\theta) + \beta f^{(n+1)}(\eta) = f^{(n+1)}(\xi)(\alpha + \beta) \tag{29}$$

(si veda l'esercizio 7.1). D'altra parte

$$\begin{aligned}\int_a^b \pi_n(x)\ dx &= \int_a^{b-h} \pi_n(x)\ dx + \int_{b-h}^b \pi_n(x)\ dx \\ &= \int_a^{b-h} \pi_{n-1}(x)(x - x_n)\ dx + \beta \\ &= \Big[\sigma_{n-1}(x)(x - x_n)\Big]_a^{b-h} - \int_a^{b-h} \sigma_{n-1}(x)\ dx + \beta \\ &= \alpha + \beta, \end{aligned} \tag{30}$$

in quanto

$$\sigma_{n-1}(a) = \sigma_{n-1}(b - h) = 0.$$

Quindi sostituendo la (30) nella (29), dalla (28) segue la (20). ∎

Con la trasformazione di variabile (15), gli integrali nei resti (19) e (20) possono essere scritti nel modo seguente:
per n pari

$$\int_a^b x\pi_n(x)\ dx = h^{n+2} \int_0^n (x_0 + th)\tau_n(t)\ dt = h^{n+3} \int_0^n t\tau_n(t)\ dt,$$

in quanto

$$\int_0^n \tau_n(t)\ dt = 0,$$

per n dispari

$$\int_a^b \pi_n(x)\ dx = h^{n+2} \int_0^n \tau_n(t)\ dt,$$

per cui le due espressioni del resto possono essere scritte nella forma

$$r_{n+1} = \gamma_n h^{s+1}\ \frac{f^{(s)}(\xi)}{s!}\,, \tag{31}$$

dove
$$\begin{cases} s = n + 2 \quad \text{e} \quad \gamma_n = \displaystyle\int_0^n t\tau_n(t)\ dt, \quad \text{per } n \text{ pari}, \\ s = n + 1 \quad \text{e} \quad \gamma_n = \displaystyle\int_0^n \tau_n(t)\ dt, \quad \text{per } n \text{ dispari}. \end{cases}$$

Dall'esame della (31) risulta quindi che le formule di Newton-Cotes hanno grado di precisione $n + 1$ se n è pari e n se n è dispari. Quindi due formule corrispondenti ad n pari e al successivo $n+1$ dispari hanno lo stesso

grado di precisione: per questa ragione per $n > 1$ è più conveniente usare formule con n pari.

Nella figura 7.4 sono riportati i coefficienti α_i, $i = 0, \ldots, \lfloor n/2 \rfloor$ e i resti delle formule di Newton-Cotes per $n = 1, \ldots, 7$. Le formule di Newton-Cotes per $n \geq 8$ hanno dei coefficienti negativi e non verificano le ipotesi del teorema 7.7. Esistono infatti delle funzioni, anche derivabili per ogni ordine, per cui la successione dei resti delle formule di Newton-Cotes diverge al crescere di n. Si possono dare teoremi di convergenza [5], ma le ipotesi che la garantiscono sono alquanto restrittive e richiedono che la funzione $f(x)$ sia analitica in una opportuna regione del piano complesso contenente l'intervallo $[a, b]$.

n	α_0	α_1	α_2	α_3	resto
1	$\frac{1}{2}$				$-\frac{1}{12}\, h^3 f''(\xi)$
2	$\frac{1}{3}$	$\frac{4}{3}$			$-\frac{1}{90}\, h^5 f^{(4)}(\xi)$
3	$\frac{3}{8}$	$\frac{9}{8}$			$-\frac{3}{80}\, h^5 f^{(4)}(\xi)$
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$		$-\frac{8}{945}\, h^7 f^{(6)}(\xi)$
5	$\frac{95}{288}$	$\frac{375}{288}$	$\frac{250}{288}$		$-\frac{275}{12096}\, h^7 f^{(6)}(\xi)$
6	$\frac{41}{140}$	$\frac{216}{140}$	$\frac{27}{140}$	$\frac{272}{140}$	$-\frac{9}{1400}\, h^9 f^{(8)}(\xi)$
7	$\frac{5257}{17280}$	$\frac{25039}{17280}$	$\frac{9261}{17280}$	$\frac{20923}{17280}$	$-\frac{8183}{518400}\, h^9 f^{(8)}(\xi)$

Fig. 7.4 - Coefficienti e resti delle formule di Newton-Cotes per $n = 1, \ldots, 7$.

7.11 Esempio. Per approssimare l'integrale

$$S = \int_0^1 \mathrm{e}^{-x^2}\, dx,$$

con un resto in modulo non superiore a $\epsilon = 0.5\ 10^{-3}$, si determina per diversi valori di n una maggiorazione del modulo del resto.

$n = 1$

$f''(x) = 2(2x^2 - 1)\mathrm{e}^{-x^2}$, $\quad \max\limits_{x\in(0,1)} |f''(x)| = 2$,

$h = 1,\ |r_2| \leq \dfrac{1}{6}$;

$n = 2$

$f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)\mathrm{e}^{-x^2}$, $\quad \max\limits_{x\in(0,1)} |f^{(4)}(x)| = 12$,

$h = \dfrac{1}{2},\ |r_3| \leq \dfrac{1}{240}$;

$n = 4$

$f^{(6)}(x) = 8(8x^6 - 60x^4 + 90x^2 - 15)\mathrm{e}^{-x^2}$, $\quad \max\limits_{x\in(0,1)} |f^{(6)}(x)| = 120$,

$h = \dfrac{1}{4},\ |r_5| \leq \dfrac{1}{16128}$.

Poiché $|r_5| < 0.5\ 10^{-3}$, utilizzando la formula di Newton-Cotes con $n = 4$, per cui sono richieste 5 valutazioni della funzione, si ottiene un valore che approssima S per meno di ϵ. Risulta

$$S_5 = \frac{1}{180}\left[14 + 64\ \mathrm{e}^{-1/16} + 24\ \mathrm{e}^{-1/4} + 64\ \mathrm{e}^{-9/16} + 14\ \mathrm{e}^{-1}\right] = 0.7468337,$$

e dato che $S = 0.7468241$, l'errore effettivamente generato è di circa $0.958\ 10^{-5}$. Nella figura 7.5 sono riportati i moduli degli errori relativi effettivamente generati dal calcolatore quando si utilizzano le formule di Newton-Cotes per valori di n compresi tra 1 e 10.

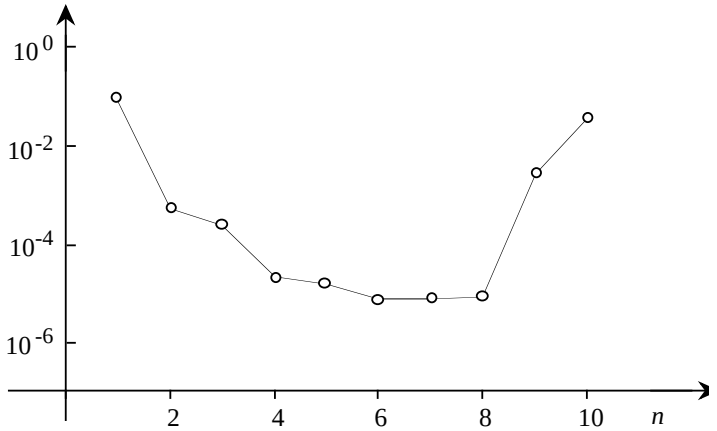


Fig. 7.5 - Errore relativo nel calcolo delle formule di Newton-Cotes per l'approssimazione di $\int_0^1 \mathrm{e}^{-x^2}\ dx$.

Si può notare che all'aumentare di n vi è una iniziale diminuzione dell'errore, dovuta ad una diminuzione del resto, fino a $n = 8$, mentre per

$n > 8$ l'errore cresce rapidamente. Poiché la funzione in esame è analitica su tutto il piano complesso, è assicurata la convergenza dei resti, quindi la crescita degli errori per $n > 8$ è dovuta alla instabilità numerica delle formule, nelle quali sono presenti dei coefficienti negativi di modulo crescente al crescere di n. Infatti S è minore di 1, mentre per $n = 10$ risulta

$$\sum_{i=0}^{n} \alpha_i = 10, \quad \sum_{i=0}^{n} |\alpha_i| \approx 10^7,$$

e quindi si verificano elevati errori di cancellazione. ∎

3. Formule newtoniane composte

Il procedimento seguito nell'esempio 7.11 per la determinazione della formula di Newton-Cotes che approssima l'integrale dato con un resto limitato in modulo da una quantità prefissata, non è applicabile nel caso in cui la funzione $f(x)$ non sia sufficientemente regolare nell'intervallo $[a, b]$. Inoltre, se il valore di n richiesto risulta maggiore di 7, in generale si presentano fenomeni di instabilità numerica, dovuta alla presenza di coefficienti negativi e positivi di modulo elevato. Per queste ragioni si preferisce utilizzare le formule *composte* che si ottengono mediante applicazione ripetuta delle formule di quadratura interpolatorie studiate.

L'intervallo $[a, b]$ viene diviso in N sottointervalli mediante i punti equidistanti z_k, $k = 0, \ldots, N$, tali che $a = z_0$, $b = z_N$. Per la proprietà di additività degli integrali si ha

$$\int_a^b f(x)\, dx = \sum_{k=0}^{N-1} \int_{z_k}^{z_{k+1}} f(x)\, dx,$$

da cui si ottiene la formula di quadratura

$$J_{n+1}^{(N)} = \sum_{k=0}^{N-1} S_{n+1}^{(k)}, \tag{32}$$

dove $S_{n+1}^{(k)}$ è la formula di Newton-Cotes S_{n+1} della funzione $f(x)$ nell'intervallo $[z_k, z_{k+1}]$.

Per $n = 1$ la formula composta si ottiene applicando la (17) ad ogni sottointervallo $[z_k, z_{k+1}]$ di ampiezza $h = (b - a)/N$, cioè

$$S_2^{(k)} = \frac{b-a}{2N}\, [f(z_k) + f(z_{k+1})].$$

Per la (32) è

$$J_2^{(N)} = \frac{b-a}{2N} \sum_{k=0}^{N-1} [f(z_k) + f(z_{k+1})]$$
$$= \frac{b-a}{2N} [f(z_0) + f(z_1) + f(z_1) + f(z_2) + \ldots + f(z_{N-1}) + f(z_N)],$$

da cui si ha la formula *dei trapezi*

$$J_2^{(N)} = \frac{b-a}{2N} \Big[f(z_0) + 2 \sum_{k=1}^{N-1} f(z_k) + f(z_N) \Big].$$

L'interpretazione geometrica di tale formula è illustrata dalla figura 7.6.

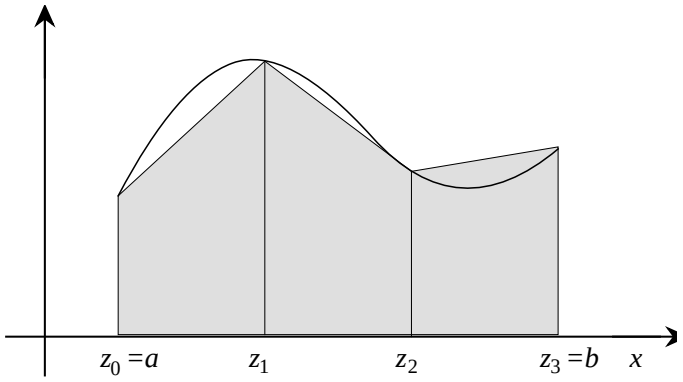


Fig. 7.6 - Formula dei trapezi.

Per $n = 2$ la formula composta si ottiene applicando la (18) ad ogni sottointervallo $[z_k, z_{k+1}]$ di ampiezza $h = (b-a)/(2N)$

$$S_3^{(k)} = \frac{b-a}{6N} [f(z_k) + 4f\Big(\frac{z_k + z_{k+1}}{2}\Big) + f(z_{k+1})].$$

Per la (32) è

$$J_3^{(N)} = \frac{b-a}{6N} \sum_{k=0}^{N-1} [f(z_k) + 4f\Big(\frac{z_k + z_{k+1}}{2}\Big) + f(z_{k+1})],$$

da cui si ha la formula di *Cavalieri-Simpson*

$$J_3^{(N)} = \frac{b-a}{6N} \Big[f(z_0) + 2 \sum_{k=1}^{N-1} f(z_k) + 4 \sum_{k=0}^{N-1} f\Big(\frac{z_k + z_{k+1}}{2}\Big) + f(z_N) \Big].$$

In modo analogo si possono ottenere le formule composte per $n > 2$.

Una formula composta è quindi del tipo

$$J_{n+1}^{(N)} = \sum_{i=0}^{m} \alpha_i f(x_i), \quad \text{con} \quad m = nN, \tag{33}$$

ed ha lo stesso grado di precisione della corrispondente formula di Newton-Cotes. Se $f(x) \in C^s[a,b]$, dove $s = n+2$ per n pari e $s = n+1$ per n dispari, il resto della formula $S_{n+1}^{(k)}$ usata per costruire la formula composta è

$$r_{n+1}^{(k)} = \gamma_n h^{s+1} \, \frac{1}{s!} \, f^{(s)}(\xi_k),$$

dove $\xi_k \in (z_k, z_{k+1})$, e $h = (z_{k+1} - z_k)/n$, per cui il resto della formula composta è dato da

$$R_{n+1}^{(N)} = \sum_{k=0}^{N-1} r_{n+1}^{(k)} = \gamma_n h^{s+1} \, \frac{1}{s!} \, \sum_{k=0}^{N-1} f^{(s)}(\xi_k) = \gamma_n N h^{s+1} \, \frac{f^{(s)}(\xi)}{s!}, \ \xi \in (a,b)$$

(si veda l'esercizio 7.1), e quindi

$$R_{n+1}^{(N)} = \gamma_n \, \frac{(b-a)^{s+1}}{n^{s+1} N^s} \, \frac{f^{(s)}(\xi)}{s!} \, . \tag{34}$$

In particolare per le formule dei trapezi e di Cavalieri-Simpson il resto è rispettivamente

$$R_2^{(N)} = -\frac{(b-a)^3}{12 \ N^2} \, f''(\xi), \tag{35}$$

$$R_3^{(N)} = -\frac{(b-a)^5}{2880 \ N^4} \, f^{(4)}(\xi). \tag{36}$$

Poiché $f(x) \in C^s[a,b]$, risulta che $|f^{(s)}(x)|$ è limitata in $[a,b]$, e quindi

$$\lim_{N \to \infty} |R_{n+1}^{(N)}| = 0,$$

e, fissato un $\epsilon > 0$, è possibile determinare un N tale che

$$|R_{n+1}^{(N)}| \leq \epsilon.$$

7.12 Esempio. Si determini il numero N di sottointervalli in cui dividere l'intervallo $[0,1]$ affinché l'integrale

$$\int_0^1 \mathrm{e}^{-x^2} \, dx$$

sia approssimato con la formula dei trapezi con un resto minore in modulo di $0.5\ 10^{-3}$. Essendo $\max_{x\in[0,1]} |f''(x)| = 2$, (si veda l'esempio 7.11), dalla (35) risulta

$$|R_2^{(N)}| \leq \frac{1}{6\ N^2}\,.$$

Imponendo che

$$\frac{1}{6\ N^2}\ \leq 0.5\ 10^{-3},$$

ne segue che l'approssimazione richiesta può essere ottenuta con $N \geq 19$. Risulta $J_2^{(19)} = 0.7466516$, con un errore di circa $0.173\ 10^{-3}$.

Con la formula di Cavalieri-Simpson, essendo $\max_{x\in[0,1]} |f^{(4)}(x)| = 12$, dalla (36) risulta

$$|R_3^{(N)}| \leq \frac{1}{240\ N^4}\,.$$

Imponendo che

$$\frac{1}{240\ N^4}\ \leq 0.5\ 10^{-3},$$

ne segue che l'approssimazione richiesta può essere ottenuta con $N \geq 2$. Risulta $J_3^{(2)} = 0.7468553$, con un errore di circa $0.312\ 10^{-4}$. Nel primo caso occorrono 20 valutazioni della $f(x)$, nel secondo caso ne bastano 5. ∎

Naturalmente il valore di ϵ non può essere scelto in modo del tutto arbitrario in quanto, oltre all'errore analitico

$$\epsilon_{an} = -\frac{R_{n+1}^{(N)}}{S}$$

del procedimento di integrazione approssimata, si deve considerare l'errore di arrotondamento commesso nel calcolo della formula stessa di integrazione. Questo errore dipende sia dalla precisione con cui nella (33) vengono calcolati i valori della funzione nei nodi, sia dal numero m dei termini che vengono sommati. Per questo motivo, è opportuno scegliere ϵ in modo che l'errore analitico non sia inferiore all'errore algoritmico e comunque che non risulti molto inferiore alla precisione di macchina usata nei calcoli.

7.13 Esempio. Nella figura 7.7 sono riportati gli errori relativi effettivamente generati, al crescere di N, dalle formule dei trapezi (linea con i pallini) e di Cavalieri-Simpson (linea con i quadratini neri) per il calcolo di

$$\int_0^1 \sin x\ dx.$$

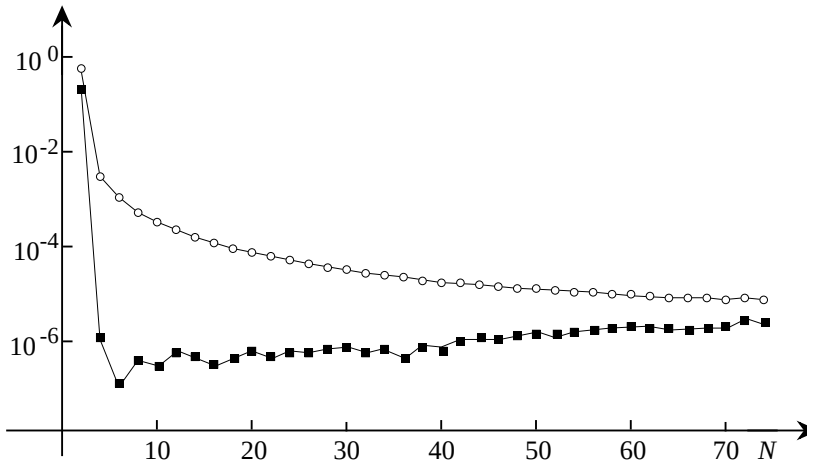


Fig. 7.7 - Errori relativi nel calcolo di $\int_0^1 \sin x \; dx$ con le formule dei trapezi e di Cavalieri-Simpson.

Per $N \leq 65$ per la formula dei trapezi e per $N \leq 6$ per la formula di Cavalieri-Simpson l'errore analitico prevale su quello di arrotondamento, mentre per valori più grandi di N l'errore totale si mantiene più o meno costante attorno ad un valore determinato dalla sola precisione di macchina (circa 10^{-6}) con tendenza ad aumentare con N. ∎

Il procedimento usato nell'esempio 7.12 per la determinazione di un valore adeguato di N si basa sull'espressione del resto della formula e richiede lo studio preliminare, che può essere pesante, di una derivata opportuna di $f(x)$. È possibile anche dare una stima del resto in modo automatico, confrontando fra loro le approssimazioni dell'integrale ottenute con due diversi valori di N. Di solito si confrontano i valori ottenuti per N e per $2N$, in modo da poter sfruttare nel secondo calcolo i valori della funzione già utilizzati per il primo. Indicati con $J^{(N)}$ e $J^{(2N)}$ le approssimazioni così ottenute, per la (34) si ha

$$S - J^{(N)} = \frac{\delta_N}{N^s}\,, \quad S - J^{(2N)} = \frac{\delta_{2N}}{2^s N^s}\,,$$

in cui δ_N e δ_{2N} differiscono per la presenza del fattore $f^{(s)}(\xi)$, che viene calcolato in punti generalmente diversi nei due casi. Nell'ipotesi che $f^{(s)}(x)$ vari di poco al variare di x, si può supporre che $\delta_N \approx \delta_{2N} = \delta$, per cui

$$J^{(2N)} - J^{(N)} \approx \frac{\delta}{2^s N^s}\,(2^s - 1),$$

e si ottiene la seguente stima del resto

$$S - J^{(2N)} \approx \frac{J^{(2N)} - J^{(N)}}{2^s - 1}\,.$$

Si può quindi procedere con successivi raddoppi di N fino a quando

$$\left| \frac{J^{(2N)} - J^{(N)}}{2^s - 1} \right| < \epsilon, \tag{37}$$

e correggere l'ultimo valore così ottenuto, assumendo

$$J^{(2N)} + \frac{J^{(2N)} - J^{(N)}}{2^s - 1} = \frac{2^s J^{(2N)} - J^{(N)}}{2^s - 1} \tag{38}$$

come approssimazione dell'integrale. Questo procedimento va sotto il nome di *estrapolazione di Richardson*. Naturalmente non vi è alcuna garanzia che l'errore dell'approssimazione sia effettivamente minore di ϵ, vista l'ipotesi che è stata fatta su $f^{(s)}(x)$.

Questa tecnica automatica di stimare l'errore e di correggere il valore ottenuto sarà poi perfezionata nello schema di Romberg (si veda il paragrafo 8).

7.14 Esempio. Si applica l'estrapolazione di Richardson al calcolo di

$$\int_0^1 \mathrm{e}^{-x^2}\, dx$$

con la formula dei trapezi. In questo caso è $s = 2$ e partendo con il valore $N = 2$ si ottiene

N	$J^{(N)}$
2	0.7313700
4	0.7429838
8	0.7458653
16	0.7465825

Per $N = 8$ risulta verificata la condizione (37) con $\epsilon = 0.5\ 10^{-3}$ e si assume il valore 0.7468214 ottenuto con la (38) come approssimazione dell'integrale. In realtà l'errore effettivo è di circa $0.273\ 10^{-5}$. Confrontando con i risultati dell'esempio 7.12, si noti come con l'estrapolazione di Richardson si sia potuta ottenere un'approssimazione migliore con un minor numero di valutazioni della $f(x)$.

Con la formula di Cavalieri-Simpson ($s = 4$) si ottiene

N	$J^{(N)}$
2	0.7468553
4	0.7468255

Per $N = 2$ risulta verificata la condizione (37) con $\epsilon = 0.5\ 10^{-3}$ e si assume il valore 0.7468235 ottenuto con la (38), come approssimazione dell'integrale. In realtà l'errore effettivo è di circa $0.633\ 10^{-6}$. ∎

Anche se la funzione $f(x)$ non è sufficientemente regolare, per cui la (34) non è applicabile, vale comunque un risultato di convergenza (si veda l'esercizio 7.3). In tal caso però la convergenza a zero del resto può essere più lenta di quanto risulta dalla (34).

7.15 Esempio. Si applica la formula di Cavalieri-Simpson al calcolo dei due integrali

$$\int_0^1 \sqrt{x}\ dx \quad \text{e} \quad \int_{0.1}^{1.1} \sqrt{x}\ dx,$$

con valori crescenti di N. Indicati con $err_N^{(1)}$ e $err_N^{(2)}$ gli errori assoluti effettivamente generati nel calcolo del primo e del secondo integrale, si ha

N	$err_N^{(1)}$	$err_N^{(2)}$
2	$0.101\ 10^{-1}$	$0.671\ 10^{-3}$
4	$0.359\ 10^{-2}$	$0.832\ 10^{-4}$
8	$0.127\ 10^{-2}$	$0.900\ 10^{-5}$
16	$0.449\ 10^{-3}$	
32	$0.161\ 10^{-3}$	
64	$0.599\ 10^{-4}$	

La convergenza nel primo caso è molto più lenta che nel secondo, e questo dipende dalla singolarità delle derivate di $f(x)$ nel punto 0. ∎

Nel caso in cui i nodi x_i, $i = 0, \ldots, m$ non sono equidistanti, ad esempio nel caso di funzioni $f(x)$ note in un insieme assegnato di punti, è ancora possibile applicare con piccole modifiche le formule di quadratura studiate. In particolare, per la formula dei trapezi, posto $h_i = x_{i+1} - x_i$, si ha

$$J_2^{(m)} = \sum_{i=0}^{m-1} \frac{h_i}{2} \left[f(x_i) + f(x_{i+1}) \right].$$

Una formula più precisa si ottiene integrando una spline cubica che approssima la $f(x)$, come nell'esercizio 5.80. Ad esempio, se sono noti anche i valori $f'(a)$ e $f'(b)$, si può usare la formula

$$\overline{J}_2^{(m)} = \sum_{i=0}^{m-1} \frac{h_i}{2} \left[f(x_i) + f(x_{i+1}) \right] - \sum_{i=0}^{m-1} \frac{h_i^3}{24} (\mu_i + \mu_{i+1}),$$

in cui μ_i, $i = 0, \ldots, m$ sono ricavati risolvendo il sistema (88), (89) del capitolo 5.

7.16 Esempio. Si calcola

$$\int_{1/\pi}^{5\pi} \sin \frac{1}{x}\, dx,$$

usando la formula dei trapezi con i nodi $x_i = (i+1)/\pi$ per $i = 0, \ldots, 4$ e $x_i = (i-4)\pi$ per $i = 5, \ldots, 9$. Il valore ottenuto 3.282565 è affetto da un errore di circa $0.316\ 10^{-2}$. Integrando invece la spline cubica costruita sugli stessi nodi si ottiene il valore 3.250903, che è affetto da un errore di circa $0.547\ 10^{-4}$. ∎

4. Formule gaussiane

Le formule gaussiane sono formule interpolatorie ad alto grado di precisione. Siano $x_0, \ldots, x_n$ gli $n+1$ zeri dell'$(n+1)$-esimo polinomio ortogonale $p_{n+1}(x)$ nell'intervallo $[a, b]$ rispetto al peso $\omega(x) \equiv 1$. La formula interpolatoria S_{n+1} costruita su tali nodi è detta *formula gaussiana* e ha, per il teorema 7.6, grado di precisione $2n+1$ (cioè il massimo possibile) e coefficienti positivi, per il teorema 7.8. Ciò per il teorema 7.7 assicura la convergenza di queste formule.

I coefficienti delle formule gaussiane possono essere dati in forma esplicita, come risulta dal seguente teorema (per le notazioni si veda il paragrafo 3 del capitolo 6).

7.17 Teorema. *I coefficienti della formula gaussiana S_{n+1} sono dati da*

$$w_i = \frac{a_{n+1} h_n}{a_n p'_{n+1}(x_i) p_n(x_i)}, \quad i = 0, \ldots, n, \tag{39}$$

dove $p_n(x)$ è l'n-esimo polinomio ortogonale nell'intervallo $[a, b]$ rispetto al peso $\omega(x) = 1$, a_n è il suo primo coefficiente e h_n è la costante di normalizzazione.

Dim. Poiché

$$p_{n+1}(x) = a_{n+1} \pi_n(x), \tag{40}$$

dalla (5) si ha che i coefficienti di S_{n+1} sono dati da

$$w_i = \frac{1}{p'_{n+1}(x_i)} \int_a^b \frac{p_{n+1}(x)}{x - x_i}\, dx, \quad i = 0, \ldots, n. \tag{41}$$

Per la formula di Christoffel-Darboux (teorema 6.14) ponendo $\xi = x_i$, si ha

$$\frac{a_{n+1}}{a_{n+2}h_{n+1}} \frac{1}{x - x_i} \left[p_{n+2}(x)p_{n+1}(x_i) - p_{n+2}(x_i)p_{n+1}(x)\right] = \sum_{j=0}^{n+1} \frac{1}{h_j} p_j(x)p_j(x_i),$$

e poiché $p_{n+1}(x_i) = 0$, integrando si ha

$$-\frac{a_{n+1}}{a_{n+2}h_{n+1}} p_{n+2}(x_i) \int_a^b \frac{p_{n+1}(x)}{x - x_i}\, dx = \sum_{j=0}^{n+1} \frac{1}{h_j} p_j(x_i) \int_a^b p_j(x)\, dx. \quad (42)$$

I polinomi $p_j(x)$ sono ortogonali nell'intervallo $[a, b]$, quindi

$$\int_a^b p_j(x)\, dx = (p_j, 1) = 0, \quad \text{per} \quad j \geq 1,$$

e per $j = 0$ è $p_0(x) = p_0(x_i) = a_0$ e

$$\frac{1}{h_0} p_0(x_i) \int_a^b p_0(x)dx = \frac{1}{h_0} (p_0, p_0) = 1.$$

Sostituendo nella (42), si ha

$$\int_a^b \frac{p_{n+1}(x)}{x - x_i}\, dx = -\frac{h_{n+1}a_{n+2}}{a_{n+1}p_{n+2}(x_i)},$$

e dalla (41) segue che

$$w_i = -\frac{h_{n+1}a_{n+2}}{a_{n+1}p'_{n+1}(x_i)p_{n+2}(x_i)}, \quad i = 0, \ldots, n. \quad (43)$$

Per il teorema 6.13, tenendo conto che $p_{n+1}(x_i) = 0$, si ha

$$p_{n+2}(x_i) = -C_{n+1}p_n(x_i) = -\frac{a_{n+2}a_n h_{n+1}}{a_{n+1}^2 h_n} p_n(x_i)$$

e sostituendo nella (43) ne segue la (39). ∎

7.18 Teorema. *Se $f(x) \in C^{2n+2}[a,b]$, il resto r_{n+1} della formula gaussiana S_{n+1} è*

$$r_{n+1} = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \pi_n^2(x)\ dx = \frac{h_{n+1}}{(2n+2)!\, a_{n+1}^2}\ f^{(2n+2)}(\xi), \quad \xi \in (a,b). \tag{44}$$

Dim. Siano $\zeta_0, \ldots, \zeta_n$, $n+1$ punti dell'intervallo $[a,b]$ distinti dagli x_i, $i = 0, \ldots, n$. Si considera il polinomio $q(x)$ di grado al più $2n+1$ che interpola la $f(x)$ nei $2n+2$ punti x_i, ζ_i, $i = 0, \ldots, n$. Indicato con $r(x)$ il resto dell'interpolazione, è

$$f(x) = q(x) + r(x)$$

e

$$S = \int_a^b f(x)\ dx = \int_a^b q(x)\ dx + \int_a^b r(x)\ dx. \tag{45}$$

La formula di quadratura S_{n+1} ha grado di precisione $2n+1$, quindi

$$\int_a^b q(x)\ dx = \sum_{i=0}^{n} w_i\ q(x_i)$$

e, poiché $q(x_i) = f(x_i)$, $i = 0, \ldots, n$, dalla (45) segue che

$$S = \sum_{i=0}^{n} w_i\ f(x_i) + \int_a^b r(x)\ dx.$$

Perciò il resto della formula di quadratura S_{n+1} è

$$r_{n+1} = \int_a^b r(x)\ dx,$$

in cui per il teorema 5.19 è

$$r(x) = \prod_{i=0}^{n}(x - x_i) \prod_{i=0}^{n}(x - \zeta_i) f[x_0, \ldots, x_n, \zeta_0, \ldots, \zeta_n, x]. \tag{46}$$

Poiché $f(x)$ è derivabile con continuità fino all'ordine $2n+2$ e il numero degli argomenti della differenza divisa $f[x_0, \ldots, x_n, \zeta_0, \ldots, \zeta_n, x]$ è $2n+3$, tale differenza, per il teorema 5.28, è funzione continua rispetto a tutti i suoi argomenti; facendo tendere ζ_i ad x_i per $i = 0, \ldots, n$, dalla (46) si ha per continuità

$$r_{n+1} = \int_a^b [(x - x_0) \ldots (x - x_n)]^2 f[x_0, \ldots, x_n, x_0, \ldots, x_n, x]\ dx.$$

Poiché la funzione $\pi_n^2(x) = [(x - x_0) \dots (x - x_n)]^2$ non cambia segno, per il teorema della media integrale esiste $\theta \in (a, b)$ tale che

$$r_{n+1} = f[x_0, \dots, x_n, x_0, \dots, x_n, \theta] \int_a^b \pi_n^2(x)\ dx,$$

inoltre per il teorema 5.19 esiste $\xi \in (a, b)$ tale che

$$f[x_0, \dots, x_n, x_0, \dots, x_n, \theta] = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\,.$$

Ne segue che

$$r_{n+1} = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \pi_n^2(x)\ dx.$$

Poiché per la (40) è

$$\pi_n(x) = \frac{1}{a_{n+1}}\ p_{n+1}(x),$$

risulta

$$\int_a^b \pi_n^2(x)\ dx = \frac{1}{a_{n+1}^2} \int_a^b p_{n+1}^2(x)\ dx = \frac{h_{n+1}}{a_{n+1}^2}\,,$$

da cui segue la (44). ∎

Se $a = -1$ e $b = 1$, i polinomi ortogonali usati per costruire le formule gaussiane sono i polinomi di Legendre, descritti nel paragrafo 3 del capitolo 6. Per tale motivo queste formule di quadratura sono anche dette di *Gauss-Legendre*. Poiché l'n-esimo polinomio di Legendre $P_n(x)$ ha primo coefficiente

$$a_n = \frac{(2n)!}{2^n\ (n!)^2}$$

e costante di normalizzazione

$$h_n = \frac{2}{2n+1}\,,$$

i coefficienti della formula di Gauss-Legendre S_{n+1} per la (39) sono

$$w_i = \frac{2}{(n+1)P'_{n+1}(x_i)P_n(x_i)}\,, \quad i = 0, \dots, n, \tag{47}$$

e il resto per la (44) è

$$r_{n+1} = \frac{2^{2n+3}\ [(n+1)!]^4}{[(2n+2)!]^3\ (2n+3)}\ f^{(2n+2)}(\xi). \tag{48}$$

Per $n = 0$, l'$(n + 1)$-esimo polinomio di Legendre è

$$P_1(x) = x,$$

che si annulla nel punto $x_0 = 0$. Dalla (47), poiché

$$P_1'(x_0)P_0(x_0) = 1,$$

si ha

$$w_0 = 2,$$

per cui la corrispondente formula di Gauss-Legendre è data da

$$S_1 = 2f(0) \tag{49}$$

ed ha grado di precisione 1.

Per $n = 1$, l'$(n + 1)$-esimo polinomio di Legendre è

$$P_2(x) = \frac{1}{2}\,(3x^2 - 1),$$

che si annulla nei punti $x_0 = -1/\sqrt{3}$ e $x_1 = 1/\sqrt{3}$. Dalla (47), poiché

$$P_2'(x_0)P_1(x_0) = P_2'(x_1)P_1(x_1) = 1,$$

si ha

$$w_0 = w_1 = 1,$$

per cui la corrispondente formula di Gauss-Legendre è data da

$$S_2 = f\Big(-\frac{1}{\sqrt{3}}\Big) + f\Big(\frac{1}{\sqrt{3}}\Big) \tag{50}$$

ed ha grado di precisione 3 (cioè quanto la formula di Newton-Cotes (18), che però richiede il calcolo della $f(x)$ in tre punti).

Per $n = 2$, l'$(n + 1)$-esimo polinomio di Legendre è

$$P_3(x) = \frac{1}{2}\,(5x^3 - 3x)$$

che si annulla nei punti $x_0 = -\sqrt{3/5}$, $x_1 = 0$ e $x_2 = \sqrt{3/5}$. Dalla (47), poiché

$$P_3'(x_0)P_2(x_0) = P_3'(x_2)P_2(x_2) = \frac{6}{5}\,, \quad P_3'(x_1)P_2(x_1) = \frac{3}{4}\,,$$

si ha

$$w_0 = w_2 = \frac{5}{9}\,, \quad w_1 = \frac{8}{9}\,,$$

per cui la corrispondente formula di Gauss-Legendre è data da

$$S_3 = \frac{1}{9}\left[5f\Big(-\sqrt{\frac{3}{5}}\Big) + 8f(0) + 5f\Big(\sqrt{\frac{3}{5}}\Big)\right] \tag{51}$$

ed ha grado di precisione 5 (cioè quanto la formula di Newton-Cotes che richiede il calcolo della $f(x)$ in 5 punti).

Per valori di n più elevati gli zeri x_i, $i = 0, \ldots, n$, di $P_{n+1}(x)$ vengono calcolati con metodi iterativi (si veda l'esercizio 7.15). È anche possibile calcolare contemporaneamente gli x_i e i w_i mediante gli autovalori e gli autovettori di una matrice tridiagonale simmetrica (si veda l'esercizio 7.17).

Nella tabella di figura 7.8 sono riportati i nodi, i coefficienti e i resti delle formule di Gauss-Legendre S_{n+1} per $n = 0, \ldots, 6$.

n	x_i	w_i	r_{n+1}
0	0	2	$0.333\ f''(\xi)$
1	± 0.5773502692	1	$0.741\ 10^{-2} f^{(4)}(\xi)$
2	± 0.7745966692 0	0.5555555556 0.8888888889	$0.635\ 10^{-4} f^{(6)}(\xi)$
3	± 0.8611363116 ± 0.3399810436	0.3478548451 0.6521451549	$0.288\ 10^{-6} f^{(8)}(\xi)$
4	± 0.9061798459 ± 0.5384693101 0	0.2369268851 0.4786286705 0.5688888889	$0.808\ 10^{-9} f^{(10)}(\xi)$
5	± 0.9324695142 ± 0.6612093865 ± 0.2386191861	0.1713244924 0.3607615730 0.4679139346	$0.154\ 10^{-11} f^{(12)}(\xi)$
6	± 0.9491079123 ± 0.7415311856 ± 0.4058451514 0	0.1294849662 0.2797053915 0.3818300505 0.4179591837	$0.213\ 10^{-14} f^{(14)}(\xi)$

Fig. 7.8 - Nodi, coefficienti e resti delle formule di Gauss-Legendre S_{n+1}, $n = 0, \ldots, 6$.

7.19 Esempio. Per applicare le formule di Gauss-Legendre all'approssimazione dell'integrale $S = \int_0^1 \mathrm{e}^{-x^2}\ dx$, già esaminato nell'esempio 7.11, occorre fare la trasformazione di variabile $x = (t+1)/2$ con cui si ottiene

$$S = \frac{1}{2}\ \int_{-1}^{1} g(t)\ dt, \quad \text{con} \quad g(t) = \exp\Big(-\ \frac{(t+1)^2}{4}\Big).$$

Il problema è quindi ricondotto all'approssimazione di quest'ultimo integrale, con un resto in modulo non superiore a $\epsilon = 0.5\ 10^{-3}$. Notando che

$$\max_{t\in[-1,1]} |g^{(k)}(t)| = 2^{-k} \max_{x\in[0,1]} |f^{(k)}(x)|,$$

dove le derivate di ordine pari della $f(x)$ e i corrispondenti massimi in modulo sono riportati nell'esempio 7.11, si ha che

$$\max_{t\in[-1,1]} |g^{(6)}(t)| = \frac{15}{8}\,.$$

Perciò per $n = 2$ il resto della formula di Gauss-Legendre è

$$|r_3| \approx 0.635\ 10^{-4} |g^{(6)}(\xi)| < \epsilon,$$

e si ottiene

$$\begin{aligned}\frac{1}{2}\,S_3 = \frac{1}{18}\ \Big[5\ \exp\Big(-\ \frac{2}{5}\ +\ \frac{1}{2}\,\sqrt{\frac{3}{5}}\Big) + 8\ \exp\Big(-\ \frac{1}{4}\Big)\\ + 5\ \exp\Big(-\ \frac{2}{5}\ -\ \frac{1}{2}\,\sqrt{\frac{3}{5}}\Big)\Big] = 0.7468145.\end{aligned}$$

L'errore assoluto effettivamente generato è

$$|S - \frac{1}{2}\,S_3| \approx 0.963\ 10^{-5}.$$

Quindi con una formula gaussiana con tre valutazioni della funzione integranda si è ottenuto un risultato con un errore dello stesso ordine di quello ottenuto nell'esempio 7.11 con una formula newtoniana che richiede 5 valutazioni di $f(x)$.

Nella figura 7.9 sono riportati i moduli degli errori relativi da cui sono affetti i risultati ottenuti applicando alla funzione $f(x)$ la formula di Gauss-Legendre per valori di n compresi tra 1 e 10. Si confronti con la figura 7.5 che riporta i corrispondenti errori per le formule newtoniane. ∎

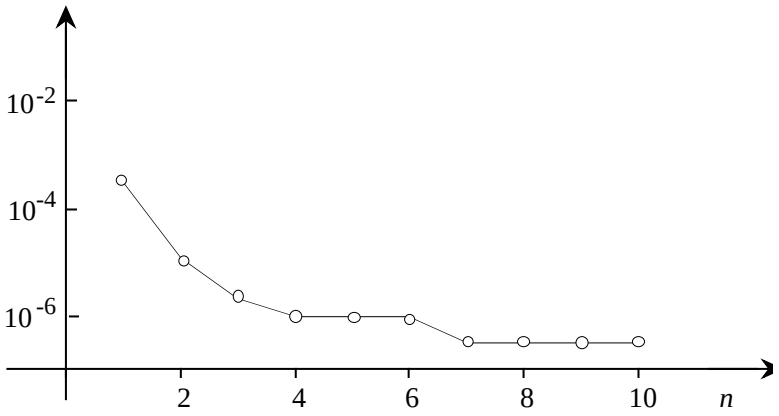


Fig. 7.9 - Errore relativo nel calcolo di $S = \int_0^1 \mathrm{e}^{-x^2}\, dx$ con le formule di Gauss-Legendre.

Una tecnica di composizione, analoga a quella descritta nel paragrafo 3, è valida anche nel caso delle formule gaussiane. Come per la (32) la formula gaussiana composta è data da

$$G_{n+1}^{(N)} = \sum_{k=0}^{N-1} S_{n+1}^{(k)},$$

dove $S_{n+1}^{(k)}$ è la formula gaussiana S_{n+1} applicata alla funzione $f(x)$ nell'intervallo $[z_k, z_{k+1}]$. Posto

$$h = \frac{b-a}{2N}\,, \quad z_k = a + 2kh, \quad k = 0, \ldots, N,$$

la trasformazione di variabile

$$x = ht + \frac{z_k + z_{k+1}}{2}$$

riconduce il calcolo dell'integrale dall'intervallo $[z_k, z_{k+1}]$ all'intervallo $[-1, 1]$. Quindi

$$G_{n+1}^{(N)} = h \sum_{j=0}^{n} w_j \sum_{k=0}^{N-1} f\Big(ht_j + \frac{z_k + z_{k+1}}{2}\Big),$$

dove t_j, $j = 0, \ldots, n$, sono gli zeri dell'$(n+1)$-esimo polinomio ortogonale. Dalla (48) si ha

$$R_{n+1}^{(N)} = \sum_{k=0}^{N-1} \gamma_{n+1} h^{2n+3} f^{(2n+2)}(\xi_k), \quad \gamma_{n+1} = \frac{2^{2n+3}\,[(n+1)!]^4}{[(2n+2)!]^3\,(2n+3)}\,,$$

dove $\xi_k \in (z_k, z_{k+1})$, e quindi esiste un punto $\xi \in (a, b)$ (si veda l'esercizio 7.1) tale che

$$R_{n+1}^{(N)} = \gamma_{n+1} h^{2n+3} N f^{(2n+2)}(\xi) = \gamma_{n+1} \frac{(b-a)^{2n+3}}{2^{2n+3} N^{2n+2}} f^{(2n+2)}(\xi).$$

7.20 Esempio. Dalla formula di Gauss-Legendre (49) si ottiene la formula composta

$$G_1^{(N)} = \frac{b-a}{N} \sum_{k=0}^{N-1} f\Big(\frac{z_k + z_{k+1}}{2}\Big), \quad (\textit{formula dei punti di mezzo})$$

la cui interpretazione geometrica è illustrata nella figura 7.10. Dalla formula di Gauss-Legendre (50) si ottiene la formula composta

$$G_2^{(N)} = \frac{b-a}{2N} \sum_{k=0}^{N-1} \Big[f\Big(z_k + \frac{3-\sqrt{3}}{3}\, h\Big) + f\Big(z_k + \frac{3+\sqrt{3}}{3}\, h\Big)\Big]. \qquad (52)$$

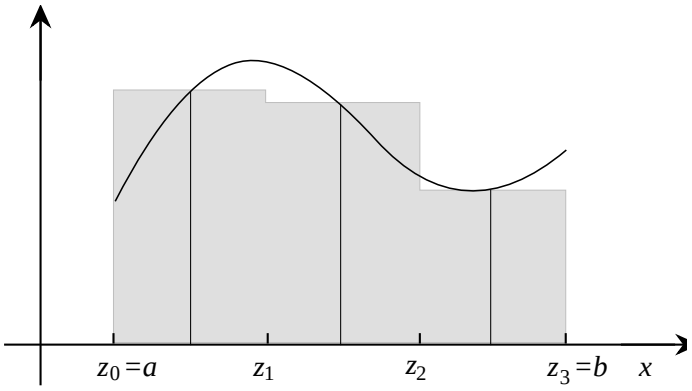


Fig. 7.10 - Formula dei punti di mezzo.

I corrispondenti resti sono

$$R_1^{(N)} = \frac{(b-a)^3}{24N^2} f''(\xi),$$

$$R_2^{(N)} = \frac{(b-a)^5}{4320N^4} f^{(4)}(\xi).$$

Quindi non vi sono differenze rilevanti fra questi resti e quelli delle formule di Newton-Cotes composte dati in (35) e (36). Anche il numero di

valutazioni di funzione richieste da queste formule è all'incirca lo stesso di quelle richieste dalle formule dei trapezi e di Cavalieri-Simpson. ∎

Confrontando tra loro le formule newtoniane e quelle gaussiane, si possono fare queste considerazioni:

a) a parità di numero di nodi le formule gaussiane hanno un grado di precisione che è circa il doppio di quelle newtoniane;

b) se gli estremi dell'intervallo di integrazione sono dati come numeri razionali, le formule newtoniane hanno come nodi e coefficienti dei numeri razionali, mentre quelle gaussiane li hanno irrazionali; in passato ciò rappresentava la principale obiezione all'uso di formule gaussiane con n elevato, in quanto la limitata precisione con cui venivano determinate le radici dei polinomi ortogonali e quindi i coefficienti delle formule, poteva generare un'elevata propagazione degli errori di arrotondamento;

c) i coefficienti delle formule gaussiane sono sempre positivi, mentre quelli delle formule newtoniane lo sono solo per $n \leq 7$, e quindi le formule gaussiane convergono per $n \to \infty$, mentre le formule newtoniane possono non convergere;

d) i nodi di una formula gaussiana S_{n+1} per $n \geq 1$ non sono un sottoinsieme dei nodi di alcuna formula gaussiana S_{m+1}, con $m > n$.

Con la diffusione dell'uso dei calcolatori, la considerazione b) ha perso molta della sua importanza: si sono ottenute, con l'uso di aritmetiche ad elevata precisione, formule gaussiane di ordine molto elevato (oltre $n = 1000$); inoltre anche nella rappresentazione di macchina di un numero razionale viene generalmente commesso un errore maggiorato dalla precisione di macchina, così come accade per i numeri irrazionali. La considerazione d) rappresenta invece un notevole inconveniente per le formule gaussiane, in quanto è impossibile applicare la formula gaussiana S_{n+k} con $k > 1$, utilizzando i valori della funzione precedentemente calcolati per applicare la formula S_{n+1}, cosa che invece è possibile fare con le formule newtoniane. Applicazioni delle formule gaussiane nella quadratura automatica saranno discusse nel paragrafo 8.

5. Formule gaussiane pesate

Delle formule gaussiane si può dare una generalizzazione che, sfruttando i polinomi ortogonali rispetto ad un peso $\omega(x) \not\equiv 1$, consente di approssimare, per mezzo della formula di quadratura (3), integrali della forma

$$\int_a^b \omega(x) f(x)\, dx,$$

in cui $\omega(x)$ è la stessa funzione peso dei polinomi ortogonali. Tali formule si chiamano *formule gaussiane pesate* e assumono di volta in volta il nome della classe di polinomi ortogonali usati per costruirle.

Tutti i risultati ottenuti nel caso delle formule gaussiane possono essere riscritti per le formule pesate, con minime modifiche. Anche le dimostrazioni dei teoremi sono analoghe a quelle già viste.

7.21 Teorema. *Fissata una funzione peso $\omega(x)$ sull'intervallo $[a,b]$, la formula di quadratura gaussiana pesata S_{n+1}, costruita scegliendo come nodi gli $n+1$ zeri dell'$(n+1)$-esimo polinomio $p_{n+1}(x)$ ortogonale nell'intervallo $[a,b]$ rispetto alla funzione $\omega(x)$, ha grado di precisione $2n+1$. I coefficienti di S_{n+1} sono tutti positivi e sono dati dalla (39). Per il resto r_{n+1}, se $f(x) \in C^{2n+2}[a,b]$, vale la relazione*

$$r_{n+1} = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \omega(x)\, \pi_n^2(x)\, dx = \frac{h_{n+1}}{(2n+2)!\, a_{n+1}^2}\, f^{(2n+2)}(\xi), \quad (53)$$

dove $\xi \in (a,b)$. ∎

Se $a=-1$ e $b=1$ la formula gaussiana pesata rispetto al peso $\omega(x) = (1-x^2)^{-1/2}$ è la formula di Gauss-Chebyshev e consente di approssimare integrali del tipo

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx,$$

scegliendo come nodi gli zeri del polinomio $T_{n+1}(x)$ di Chebyshev di prima specie descritto nel paragrafo 3 del capitolo 6

$$x_i = \cos\theta_i, \quad \theta_i = \frac{(2i+1)\pi}{2(n+1)}, \quad i = 0, \dots, n. \quad (54)$$

I coefficienti, per la (39), sono

$$\begin{cases} w_0 = \pi & \text{per } n = 0, \\ w_i = \dfrac{\pi}{T'_{n+1}(x_i) T_n(x_i)}, \ i = 0, \dots, n, & \text{per } n \geq 1. \end{cases} \quad (55)$$

Ponendo $x = \cos\theta$, si ha $T_n(x) = \cos n\theta$,

$$T'_{n+1}(x) = \frac{d}{dx}\, T_{n+1}(x) = \frac{d}{d\theta}\, T_{n+1}(\cos\theta)\, \frac{d\theta}{dx} = (n+1)\, \frac{\sin(n+1)\theta}{\sin\theta}.$$

Poiché per $i = 0, \dots, n$ è $\cos(n+1)\theta_i = 0$, risulta

$$\cos n\theta_i \cos\theta_i = \sin n\theta_i \sin\theta_i, \quad (56)$$

e quindi

$$T'_{n+1}(x_i)T_n(x_i) = (n+1)\,\frac{\sin(n+1)\theta_i \cos n\theta_i}{\sin\theta_i}$$
$$= (n+1)\,\frac{\sin\theta_i \cos^2 n\theta_i + \cos\theta_i \sin n\theta_i \cos n\theta_i}{\sin\theta_i}\,,$$

e per la (56) è

$$T'_{n+1}(x_i)T_n(x_i) = (n+1)(\cos^2 n\theta_i + \sin^2 n\theta_i) = n+1.$$

Sostituendo nella (55) si ha allora per $n \geq 1$ che

$$w_i = \frac{\pi}{n+1}\,, \quad i = 0, \ldots, n.$$

Le formule di Gauss-Chebyshev sono perciò date da

$$S_{n+1} = \frac{\pi}{n+1}\,\sum_{i=0}^{n} f(x_i),$$

dove i nodi x_i sono dati dalla (54). Risulta quindi che le formule di Gauss-Chebyshev sono formule a coefficienti uniformi. Per quanto riguarda il resto, dalla (53) si ha che

$$r_{n+1} = \frac{\pi}{2^{2n+1}(2n+2)!}\, f^{(2n+2)}(\xi). \tag{57}$$

Per $n = 0$ si ha $x_0 = 0$; la corrispondente formula di Gauss-Chebyshev è

$$S_1 = \pi f(0),$$

ed ha grado di precisione 1. Per $n = 1$ è $x_0 = -x_1 = -1/\sqrt{2}$; la corrispondente formula di Gauss-Chebyshev è

$$S_2 = \frac{\pi}{2}\,\Big[f\Big(-\frac{1}{\sqrt{2}}\Big) + f\Big(\frac{1}{\sqrt{2}}\Big)\Big],$$

ed ha grado di precisione 3.

7.22 Esempio. Si applicano le formule di Gauss-Chebyshev al calcolo dell'integrale

$$S = \int_{-1}^{1} \frac{\cos x}{\sqrt{1-x^2}}\; dx,$$

richiedendo che il resto sia minore in modulo di $\epsilon = 0.5\ 10^{-3}$. Poiché $f(x) = \cos x$ e per ogni k è $|f^{(k)}(x)| \leq 1$, dalla (57) si ha che per $n = 2$ è $|r_{n+1}| \leq \epsilon$ e la formula da usare è

$$S_3 = \frac{\pi}{3} \sum_{i=0}^{2} \cos x_i = \frac{\pi}{3} \Big[\cos\Big(-\ \frac{\sqrt{3}}{2} \Big) + \cos 0 + \cos\Big(\frac{\sqrt{3}}{2} \Big) \Big]$$
$$= \frac{\pi}{3} \Big[1 + 2 \cos\Big(\frac{\sqrt{3}}{2} \Big) \Big] = 2.404070.$$

Poiché $S = 2.403939$, l'errore assoluto è di circa $0.131\ 10^{-3}$. ∎

Le formule gaussiane pesate possono essere usate anche per approssimare integrali su un intervallo infinito. Se $a = 0$ e $b = \infty$, la formula gaussiana pesata rispetto al peso $\omega(x) = \mathrm{e}^{-x}$ è la formula di Gauss-Laguerre e consente di approssimare integrali del tipo

$$\int_0^\infty \mathrm{e}^{-x} f(x)\ dx,$$

scegliendo come nodi gli zeri x_i, $i = 0, \ldots, n$, del polinomio $L_{n+1}(x)$ di Laguerre descritto nel paragrafo 3 del capitolo 6. I coefficienti sono dati da

$$w_i = -\ \frac{1}{(n+1) L'_{n+1}(x_i) L_n(x_i)}\,,$$

il resto per la (53) è dato da

$$r_{n+1} = \frac{((n+1)!)^2}{(2n+2)!}\ f^{(2n+2)}(\xi). \tag{58}$$

Per $n = 0$ si ha $L_1(x) = -(x-1)$ e quindi

$$x_0 = 1, \quad L'_1(x_0) = -1, \quad L_0(x_0) = 1,$$

per cui $w_0 = 1$. La corrispondente formula di Gauss-Laguerre è

$$S_1 = f(1)$$

ed ha grado di precisione 1. Per $n = 1$ si ha

$$L_2(x) = \frac{1}{2}\ (x^2 - 4x + 2),$$

da cui

$$x_0 = 2 - \sqrt{2}, \quad x_1 = 2 + \sqrt{2}, \quad L'_2(x_0) = -\sqrt{2} = -L'_2(x_1),$$
$$L_1(x_0) = \sqrt{2} - 1, \quad L_1(x_1) = -\sqrt{2} - 1,$$

e quindi

$$w_0 = \frac{1}{4}\,(2+\sqrt{2}), \quad w_1 = \frac{1}{4}\,(2-\sqrt{2}).$$

La corrispondente formula di Gauss-Laguerre è

$$S_2 = \frac{1}{4}\,[(2+\sqrt{2})f(2-\sqrt{2}) + (2-\sqrt{2})f(2+\sqrt{2})]$$

ed ha grado di precisione 3. Nella tabella di figura 7.11 sono riportati i nodi, i coefficienti e i resti delle formule di Gauss-Laguerre S_{n+1}, per $n = 0, \ldots, 5$.

n	x_i	w_i	r_{n+1}
0	1	1	$0.500\ f''(\xi)$
1	0.5857864376 3.414213562	0.8535533906 0.1464466094	$0.167\ f^{(4)}(\xi)$
2	0.4157745568 2.29428036 6.289945083	0.7110930099 0.2785177336 $0.103892565\ 10^{-1}$	$0.500\ 10^{-1} f^{(6)}(\xi)$
3	0.3225476896 1.745761101 4.536620297 9.395070912	0.6031541043 0.3574186924 $0.3888790852\ 10^{-1}$ $0.5392947056\ 10^{-3}$	$0.143\ 10^{-1} f^{(8)}(\xi)$
4	0.2635603197 1.413403059 3.596425771 7.085810006 12.64080084	0.5217556106 0.3986668111 $0.7594244968\ 10^{-1}$ $0.3611758680\ 10^{-2}$ $0.2336997239\ 10^{-4}$	$0.397\ 10^{-2} f^{(10)}(\xi)$
5	0.2228466042 1.188932102 2.992736326 5.775143569 9.837467418 15.98287398	0.4589646739 0.4170008308 0.1133733821 $0.1039919745\ 10^{-1}$ $0.2610172028\ 10^{-3}$ $0.8985479064\ 10^{-6}$	$0.108\ 10^{-2} f^{(12)}(\xi)$

Fig. 7.11 - Nodi, coefficienti e resti delle formule di Gauss-Laguerre S_{n+1}, $n = 0, \ldots, 5$.

7.23 Esempio. Si applicano le formule di Gauss-Laguerre al calcolo dell'integrale

$$S = \int_0^\infty \mathrm{e}^{-x} \cos x \; dx,$$

richiedendo che il resto sia minore in modulo di $\epsilon = 0.5\ 10^{-3}$. Poiché $f(x) = \cos x$ e per ogni k è $|f^{(k)}(x)| \leq 1$, dalla (58) si ha che per $n = 6$ è $|r_{n+1}| \leq \epsilon$ (per i coefficienti e i nodi della formula di Gauss-Laguerre con $n = 6$ si veda l'esercizio 7.17). Il valore ottenuto è $S_7 = 0.5000424$, e poiché $S = 0.5$, l'errore assoluto è di circa $0.424\ 10^{-4}$. ∎

Se $a = -\infty$ e $b = +\infty$, la formula gaussiana pesata rispetto al peso $\omega(x) = \mathrm{e}^{-x^2}$ è la formula di Gauss-Hermite e consente di approssimare integrali del tipo

$$\int_{-\infty}^\infty \mathrm{e}^{-x^2} f(x) \; dx,$$

scegliendo come nodi gli zeri $x_i, i = 0, \ldots, n$, del polinomio $H_{n+1}(x)$ di Hermite descritto nel paragrafo 3 del capitolo 6. I coefficienti sono dati da

$$w_i = \frac{2^{n+1} n! \sqrt{\pi}}{H'_{n+1}(x_i) H_n(x_i)},$$

il resto è dato da

$$r_{n+1} = \frac{(n+1)! \sqrt{\pi}}{2^{n+1}(2n+2)!} \; f^{(2n+2)}(\xi). \tag{59}$$

Per $n = 0$ si ha $H_1(x) = 2x$ e quindi

$$x_0 = 0, \quad H'_1(x_0) = 2, \quad H_0(x_0) = 1,$$

per cui $w_0 = \sqrt{\pi}$. La corrispondente formula di Gauss-Hermite è

$$S_1 = \sqrt{\pi} f(0)$$

ed ha grado di precisione 1. Per $n = 1$ si ha

$$H_2(x) = 4x^2 - 2,$$

da cui

$$x_0 = -\frac{1}{\sqrt{2}} = -x_1, \quad H'_2(x_0) = -H'_2(x_1) = -\frac{8}{\sqrt{2}},$$

$$H_1(x_0) = -H_1(x_1) = -\frac{2}{\sqrt{2}},$$

e quindi

$$w_0 = w_1 = \frac{\sqrt{\pi}}{2}\,.$$

La corrispondente formula di Gauss-Hermite è

$$S_2 = \frac{\sqrt{\pi}}{2}\,\Big[f\Big(-\frac{1}{\sqrt{2}}\Big) + f\Big(\frac{1}{\sqrt{2}}\Big)\Big]$$

ed ha grado di precisione 3. Nella tabella di figura 7.12 sono riportati i nodi, i coefficienti e i resti delle formule di Gauss-Hermite S_{n+1} per $n = 0, \ldots, 6$.

n	x_i	w_i	r_{n+1}
0	0	1.772453851	$0.443\ f''(\xi)$
1	$\pm$0.7071067812	0.8862269255	$0.369\ 10^{-1} f^{(4)}(\xi)$
2	$\pm$1.224744871 0	0.2954089752 1.181635901	$0.185\ 10^{-2} f^{(6)}(\xi)$
3	$\pm$1.650680124 $\pm$0.5246476233	$0.8131283545\ 10^{-1}$ 0.8049140900	$0.659\ 10^{-4} f^{(8)}(\xi)$
4	$\pm$2.02018287 $\pm$0.9585724646 0	$0.1995324206\ 10^{-1}$ 0.3936193232 0.9453087205	$0.183\ 10^{-5} f^{(10)}(\xi)$
5	$\pm$2.350604974 $\pm$1.335849074 $\pm$0.4360774119	$0.4530009906\ 10^{-2}$ 0.1570673203 0.7246295952	$0.416\ 10^{-7} f^{(12)}(\xi)$
6	$\pm$2.651961357 $\pm$1.673551629 $\pm$0.8162878829 0	$0.9717812451\ 10^{-3}$ $0.5451558282\ 10^{-1}$ 0.4256072526 0.8102646176	$0.801\ 10^{-9} f^{(14)}(\xi)$

Fig. 7.12 - Nodi, coefficienti e resti delle formule di Gauss-Hermite S_{n+1}, $n = 0, \ldots, 6$.

7.24 Esempio. Si applicano le formule di Gauss-Hermite al calcolo dell'integrale

$$S = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2} \cos x\ dx,$$

richiedendo che il resto sia minore in modulo di $\epsilon = 0.5\ 10^{-3}$. Poiché $f(x) = \cos x$ e per ogni k è $|f^{(k)}(x)| \leq 1$, dalla (59) si ha che per $n = 3$ è $|r_{n+1}| \leq \epsilon$. Il valore ottenuto è $S_4 = 1.380329$, e poiché $S = \sqrt{\pi}\mathrm{e}^{-1/4} = 1.380388$, l'errore assoluto è di circa $0.594\ 10^{-4}$. ∎

6. Integrali impropri

Il calcolo dell'integrale (1) quando $f(x)$ ha una singolarità, oppure quando l'intervallo di integrazione non è limitato, presenta spesso qualche difficoltà.

Si esamina dapprima il caso in cui l'intervallo $[a, b]$ è limitato e la funzione $f(x)$ è singolare, ad esempio, nel punto a. In tal caso supporremo che esista il

$$\lim_{\eta \to 0} \int_{a+\eta}^{b} f(x)\ dx. \tag{60}$$

Talvolta è possibile, con un opportuno cambiamento di variabile, trasformare un integrale improprio di questo tipo in un integrale proprio. Ad esempio, ponendo $x = t^n$, si ha

$$\int_0^1 \frac{1}{\sqrt[n]{x}}\ dx = n \int_0^1 t^{n-2}\ dt, \quad \text{per } n \geq 2.$$

Più spesso però non si riesce a trovare una trasformazione che elimini la singolarità e lasci limitato l'intervallo di integrazione. Altre volte è possibile ricondurre il calcolo di (1) al calcolo di integrali noti o di integrali propri. Per esempio, ponendo

$$f(x) = g(x) + h(x),$$

in modo che la funzione $g(x)$ abbia un punto di singolarità, ma l'integrale $\int_a^b g(x)\ dx$ sia noto e l'integrale $\int_a^b h(x)\ dx$ sia proprio (metodo di *sottrazione della singolarità*).

7.25 Esempio. Per calcolare l'integrale

$$S = \int_0^1 \frac{\cos x^2}{\sqrt{x}}\ dx,$$

conviene utilizzare la relazione

$$\int_0^1 \frac{\cos x^2}{\sqrt{x}}\ dx = \int_0^1 \frac{1}{\sqrt{x}}\ dx + \int_0^1 \frac{\cos x^2 - 1}{\sqrt{x}}\ dx.$$

Il primo integrale può essere calcolato direttamente, in quanto $2\sqrt{x}$ è una primitiva di $\dfrac{1}{\sqrt{x}}$, il secondo integrale è proprio, in quanto la funzione integranda può essere estesa per continuità in 0 essendo

$$\lim_{x\to 0} \frac{\cos x^2 - 1}{\sqrt{x}} = 0.$$

Il calcolo dell'integrale può quindi essere fatto con uno dei metodi esposti. Poiché però

$$h(x) = \frac{\cos x^2 - 1}{\sqrt{x}} \notin C^4[0,1],$$

il calcolo ad esempio con il metodo di Cavalieri-Simpson potrebbe richiedere un elevato numero di valutazioni della funzione. Conviene allora sottrarre altri termini, ottenendo

$$\int_0^1 \frac{\cos x^2}{\sqrt{x}}\, dx = \int_0^1 \frac{1}{\sqrt{x}}\, dx - \frac{1}{2}\int_0^1 x^{7/2}\, dx + \int_0^1 \frac{\cos x^2 - 1 + x^4/2}{\sqrt{x}}\, dx, \tag{61}$$

e risulta

$$h(x) = \frac{\cos x^2 - 1 + x^4/2}{\sqrt{x}} \in C^4[0,1].$$

Applicando il metodo di Cavalieri-Simpson e l'estrapolazione di Richardson con $\epsilon = 0.5\ 10^{-3}$, si ottiene per $N = 2$ il valore 0.004793137 come approssimazione dell'ultimo integrale della (61). Il valore

$$2 - \frac{1}{2}\,\frac{2}{9} + 0.004793137 = 1.893682$$

è un'approssimazione di S, con un errore assoluto di circa $0.772\ 10^{-6}$. ∎

Un altro modo di procedere è implicitamente indicato nella (60): si determina η in modo che

$$\left| \int_a^{a+\eta} f(x)\, dx \right|$$

sia minore della precisione richiesta per il risultato e si approssima

$$\int_a^b f(x)\, dx \quad \text{con} \quad \int_{a+\eta}^b f(x)\, dx.$$

7.26 Esempio. Per l'integrale dell'esempio 7.25

$$S = \int_0^1 \frac{\cos x^2}{\sqrt{x}}\, dx,$$

si ha

$$\left| \int_0^\eta \frac{\cos x^2}{\sqrt{x}}\, dx \right| < \int_0^\eta \frac{1}{\sqrt{x}}\, dx = 2\sqrt{\eta}.$$

Quindi per approssimare S con un errore assoluto minore di $\epsilon = 0.5\ 10^{-3}$, basta scegliere η in modo che

$$2\sqrt{\eta} \leq \epsilon, \quad \text{cioè} \quad \eta = 6.25\ 10^{-8}.$$

Dal punto di vista numerico questo modo di procedere presenta dei grossi inconvenienti: infatti il calcolo di un integrale proprio con una singolarità della funzione integranda appena fuori dell'intervallo di integrazione può richiedere, come si vedrà più avanti, un numero molto elevato di nodi. ∎

Un'approssimazione affetta da un errore dello stesso ordine di grandezza può essere ottenuta calcolando l'integrale fra a e b della funzione $g(x)$ ottenuta per estensione della $f(x)$ in a con il valore 0 ed applicando una formula composta i cui primi due punti siano a e $a + \eta$. Su questo si basa la tecnica che consiste nell'*ignorare la singolarità*, utilizzando formule di quadratura che non fanno intervenire il valore $f(a)$, come ad esempio la formula dei punti di mezzo, oppure utilizzando formule nelle quali il valore $f(a)$ è sostituito con 0. A seconda del tipo di singolarità è possibile stimare l'errore che viene commesso. In alcuni casi, come ad esempio per le funzioni oscillanti, non è possibile dare limitazioni superiori all'errore.

7.27 Esempio. Per calcolare

$$S = \int_0^1 \frac{1}{\sqrt[3]{x}}\, dx$$

si utilizza la formula dei punti di mezzo con valori crescenti di N, ottenendo le approssimazioni $G_1^{(N)}$

N	$G_1^{(N)}$
2	1.344021
4	1.400460
8	1.436967
.	...
256	1.493670
512	1.495948
1024	1.497396

Poiché $S = 1.5$, l'errore effettivamente commesso per $N = 1024$ è di circa $0.260\ 10^{-2}$. Utilizzando invece la formula dei trapezi per $N = 1024$ si ottiene $J_2^{(N)} = 1.490360$ con un errore assoluto di circa $0.964\ 10^{-2}$. Nel caso dell'integrale

$$S = \int_0^1 \frac{1}{x} \sin \frac{1}{x}\, dx,$$

il cui valore è $S = 0.6247133$, si ottiene

N	$J_2^{(N)}$
2	1.119664
4	0.1270092
8	1.535175
.	...
128	1.757367
256	−0.0168677
512	0.9092879

Chiaramente in questo caso il metodo non è convergente (si veda anche l'esercizio 7.32). ∎

È possibile dimostrare [5] che se in un intorno di un punto di singolarità la funzione $f(x)$ è maggiorabile in modulo con una funzione monotona integrabile, la successione ottenuta al crescere del numero N dei nodi di una formula composta è convergente al valore dell'integrale.

Un altro metodo consiste nello sviluppare la funzione integranda, o parte di essa, in serie e integrare poi termine a termine. Questo metodo è efficace se la convergenza è sufficientemente rapida.

7.28 Esempio. Per il calcolo di

$$S = \int_0^1 \frac{\cos x^2}{\sqrt{x}}\, dx,$$

si considera lo sviluppo in serie di $\cos x^2$, ottenendo

$$S = \int_0^1 \frac{1}{\sqrt{x}}\, dx - \frac{1}{2} \int_0^1 x^{7/2}\, dx + \frac{1}{4!} \int_0^1 x^{15/2}\, dx - \dots \tag{62}$$

Poiché

$$\frac{1}{k!} \int_0^1 x^{2k-1/2}\, dx = \frac{2}{k!\,(4k+1)},$$

scegliendo $\epsilon = 0.5\ 10^{-3}$ risulta $\dfrac{2}{k!\,(4k+1)} < \epsilon$ per $k = 6$. Quindi si può ottenere un'approssimazione di S con un errore assoluto minore di ϵ sommando tre soli termini della (62). Il valore così ottenuto, 1.893791, è in realtà affetto da un errore assoluto di circa $0.110\ 10^{-3}$. ∎

Si possono infine utilizzare formule gaussiane pesate, in particolare la formula di Gauss-Chebyshev, ponendo

$$\int_{-1}^{1} f(x)\ dx = \int_{-1}^{1} \frac{g(x)}{\sqrt{1-x^2}}\ dx,$$

nel caso che la funzione $g(x) = \sqrt{1-x^2}f(x)$ non abbia singolarità. Naturalmente questa tecnica è applicabile anche per integrali su intervalli limitati, diversi da $[-1, 1]$, purché si faccia l'opportuna trasformazione di variabile.

7.29 Esempio. Per l'integrale

$$S = \int_{0}^{1} \frac{\cos x^2}{\sqrt{x}}\ dx,$$

si ponga $y = 2x - 1$. Si ottiene

$$S = \frac{\sqrt{2}}{2} \int_{-1}^{1} \frac{1}{\sqrt{1+y}}\ \cos\ \frac{(1+y)^2}{4}\ dy = \frac{\sqrt{2}}{2} \int_{-1}^{1} \frac{\sqrt{1-y}}{\sqrt{1-y^2}}\ \cos\ \frac{(1+y)^2}{4}\ dy.$$

Poiché la funzione $g(y) = \sqrt{1-y}\ \cos\dfrac{(1+y)^2}{4}$ non ha singolarità nell'intervallo $[-1, 1]$, si può applicare una formula di Gauss-Chebyshev. Posto $\epsilon = 0.5\ 10^{-3}$, non è possibile, servendosi della (57), determinare un valore di n per cui $|r_{n+1}| < \epsilon$ perché la funzione $g(y)$ non è sufficientemente derivabile per $y = 1$. Per valori crescenti di n si ha

n	S_{n+1}	err_{n+1}
2	1.903627	$0.995\ 10^{-2}$
4	1.897875	$0.419\ 10^{-2}$
8	1.895028	$0.135\ 10^{-2}$
16	1.894061	$0.381\ 10^{-3}$
32	1.893775	$0.944\ 10^{-4}$

in cui err_{n+1} è il modulo dell'errore assoluto di S_{n+1} effettivamente calcolato. ∎

La presenza di singolarità della funzione $f(x)$ fuori dall'intervallo di integrazione, ma vicino agli estremi può avere effetti negativi sull'approssimazione dell'integrale, in quanto per ottenere lo stesso resto possono essere richiesti valori di n più elevati che nel caso di funzioni senza singolarità.

7.30 Esempio. Si calcola l'integrale

$$\int_{-1}^{1} \frac{dx}{x^2 - \alpha}\, dx$$

per i valori di $\alpha = 1.01$ e $\alpha = 2$ con la formula di Cavalieri-Simpson. Posto $\epsilon = 0.5\ 10^{-3}$, con l'estrapolazione di Richardson

risulta $N = 256$, per $\alpha = 1.01$,

risulta $N = 4$, per $\alpha = 2$.

L'elevato valore di N per $\alpha = 1.01$ è causato dalla presenza dei punti di singolarità ± 1.004988 della $f(x)$ fuori dall'intervallo $[-1, 1]$. Lo stesso inconveniente si presenta anche nel caso dell'integrale

$$\int_{-1}^{1} \frac{dx}{x^2 + \alpha}\,,$$

in cui la $f(x)$ non ha punti reali di singolarità. Infatti

risulta $N = 2$, per $\alpha = 1$,

risulta $N = 32$, per $\alpha = 0.01$.

L'elevato valore di N per $\alpha = 0.01$ è causato dalla presenza di una singolarità nel campo complesso molto vicina all'intervallo di integrazione. ∎

Si esamina adesso il caso in cui l'intervallo di integrazione non è limitato e la funzione $f(x)$ non ha singolarità. Se l'intervallo non è limitato da una sola parte, si supporrà che esso sia $[0, +\infty]$ e che esista il

$$\lim_{t\to\infty} \int_0^t f(x)\, dx,$$

altrimenti si supporrà che esista il

$$\lim_{s\to\infty} \int_{-s}^0 f(x)\, dx + \lim_{t\to\infty} \int_0^t f(x)\, dx.$$

Anche in questo caso è talvolta possibile fare dei cambiamenti di variabile che trasformano l'integrale improprio in uno proprio, come ad esempio la

trasformazione $y = \mathrm{e}^{-x}$, che riduce l'intervallo $[0, +\infty]$ in un intervallo finito, però queste trasformazioni introducono in molti casi delle singolarità.

La tecnica numerica più semplice è quella di troncare l'intervallo di integrazione, approssimando l'integrale su $[0, +\infty]$ o su $[-\infty, +\infty]$ con

$$\int_0^t f(x)\ dx \quad \text{o} \quad \int_{-s}^t f(x)\ dx,$$

per opportuni valori positivi e grandi di t e di s. Naturalmente questa tecnica dovrebbe essere accompagnata da una stima dell'errore che si commette, cioè una stima di

$$\left| \int_t^{+\infty} f(x)\ dx \right| \quad \text{o} \quad \left| \int_{-\infty}^{-s} f(x)\ dx \right| + \left| \int_t^{+\infty} f(x)\ dx \right|$$

(si veda ad esempio l'esercizio 7.36). Una stima ragionevole può essere ottenuta, nel caso di funzioni $f(x)$ che godono di certe proprietà di regolarità, procedendo nel modo seguente. Per l'intervallo $[0, +\infty]$, fissato $t_0 = 0$, si considera una successione $\{t_i\}$ monotona, crescente e tendente a $+\infty$, e si valutano gli integrali

$$\int_0^{t_n} f(x)\ dx = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} f(x)\ dx,$$

interrompendo il procedimento quando

$$\left| \int_{t_n}^{t_{n+1}} f(x)\ dx \right|$$

è minore di una tolleranza prefissata.

Si può dimostrare [5] che se la funzione $f(x)$ è monotona decrescente e se la successione $\{t_i\}$ è scelta in modo opportuno, ad esempio $t_i = 2^i$, per $i \geq 1$, questo procedimento consente di ottenere una buona approssimazione dell'integrale.

7.31 Esempio. Per calcolare

$$S = \int_0^{\infty} \frac{\cos x}{1 + x^2}\ dx,$$

si considera la successione

$$t_0 = 0, \quad t_i = 2^i, \quad i = 1, 2, \ldots,$$

e si calcolano gli integrali

$$\phi_i = \int_{t_i}^{t_{i+1}} f(x)\ dx$$

con la formula di Cavalieri-Simpson e l'estrapolazione di Richardson. Si ottiene

i	$[t_i, t_{i+1}]$	ϕ_i
0	$[0, 2]$	0.7285330
1	$[2, 4]$	-0.1741832
2	$[4, 8]$	$0.3810822\ 10^{-1}$
3	$[8, 16]$	$-0.1524590\ 10^{-1}$
4	$[16, 32]$	$0.1120762\ 10^{-2}$
5	$[32, 64]$	$-0.2611752\ 10^{-3}$

Arrestando il calcolo ad $i = 5$ si ottiene il valore 0.5780724. Poiché $S = \dfrac{\pi}{2e} = 0.5778637$, l'errore assoluto è di circa $0.209\ 10^{-3}$. ∎

Se si applica la formula dei trapezi, conviene fissare h e quindi i punti $z_k = kh$, e procedere nel modo seguente:

$$J^{(0)} = \frac{h}{2}\ \big[f(0) + f(z_1)\big],$$
$$J^{(k)} = J^{(k-1)} + \frac{h}{2}\ \big[f(z_k) + f(z_{k+1})\big], \quad k = 1, 2, \ldots$$

Si può dimostrare [5] che se $f(x) \in C^3[0, +\infty]$, se $|f'''(x)|$ è integrabile su $[0, +\infty]$, se $f'(0) = 0$ e $\lim\limits_{x\to+\infty} f'(x) = 0$, allora

$$\lim_{h\to 0} \lim_{k\to\infty} J^{(k)} = S$$

e $|S - \lim\limits_{k\to\infty} J^{(k)}|$ tende a 0 come h^3. Se la funzione $f(x)$ ha derivate di ordine più elevato, integrabili in modulo e che si annullano in 0 e in $+\infty$, allora la convergenza è ancora più rapida.

È anche possibile utilizzare le formule gaussiane pesate, di Gauss-Laguerre o di Gauss-Hermite, nel modo seguente

$$\int_0^{+\infty} f(x)\ dx = \int_0^{+\infty} \mathrm{e}^{-x} g(x)\ dx, \quad \text{con} \quad g(x) = \mathrm{e}^{x} f(x),$$
$$\int_{-\infty}^{+\infty} f(x)\ dx = \int_{-\infty}^{+\infty} \mathrm{e}^{-x^2} g(x)\ dx, \quad \text{con} \quad g(x) = \mathrm{e}^{x^2} f(x),$$

nei casi, ad esempio, in cui la funzione $g(x)$ abbia un comportamento di tipo polinomiale. Vale infatti il seguente teorema di convergenza, per la cui dimostrazione si rimanda a [22].

7.32 Teorema. *Se esistono $\rho > 1$ e $\sigma > 0$ tali che*

a) $$|f(x)| \leq \frac{1}{x^\rho} \quad per \quad x > \sigma,$$

allora

$$\lim_{n\to\infty} S_{n+1} = \int_0^{+\infty} f(x)\ dx,$$

in cui S_{n+1} è il valore ottenuto applicando la formula di Gauss-Laguerre con n nodi alla funzione $g(x) = e^x f(x)$;

b) $$|f(x)| \leq \frac{1}{|x|^\rho} \quad per \quad |x| > \sigma,$$

allora

$$\lim_{n\to\infty} S_{n+1} = \int_{-\infty}^{+\infty} f(x)\ dx,$$

in cui S_{n+1} è il valore ottenuto applicando la formula di Gauss-Hermite con n nodi alla funzione $g(x) = e^{x^2} f(x)$. ∎

7.33 Esempio. Per il calcolo di

$$S = \int_0^{\infty} \frac{x^3 - 10}{\mathrm{e}^x + 1}\ dx,$$

si pone

$$S = \int_0^{\infty} \mathrm{e}^{-x} g(x)\ dx, \quad \text{dove} \quad g(x) = \frac{\mathrm{e}^x (x^3 - 10)}{\mathrm{e}^x + 1}\,,$$

e si applicano le formule di Gauss-Laguerre per diversi valori di n. Poiché $S = \dfrac{7\pi^4}{120} - 10 \log 2 = -1.249275$, risulta

n	S_{n+1}	err_{n+1}
3	-1.260805	$0.115\ 10^{-1}$
4	-1.252348	$0.307\ 10^{-2}$
5	-1.249070	$0.205\ 10^{-3}$

in cui err_{n+1} è il modulo dell'errore assoluto di S_{n+1} effettivamente calcolato. Nel caso dell'integrale

$$S = \int_0^\infty \frac{\cos x}{1+x^2}\, dx$$

dell'esempio precedente, applicando le formule di Gauss-Laguerre si ottiene ancora una successione di approssimazioni che convergono a S, ma molto più lentamente. Per $n = 5$ risulta $S_{n+1} = 0.5297295$ con un errore assoluto di circa $0.481\ 10^{-1}$. ∎

7. Formule gaussiane con nodi prefissati

Delle formule gaussiane si può dare una generalizzazione che include fra i nodi, oltre agli zeri di opportuni polinomi ortogonali, anche punti prefissati, come ad esempio gli estremi dell'intervallo di integrazione.

7.34 Definizione. Sia

$$S = \int_a^b \omega(x) f(x)\, dx;$$

una formula di quadratura per il calcolo di S del tipo

$$S_{m+n+2} = \sum_{k=0}^{m} \alpha_k f(y_k) + \sum_{j=0}^{n} w_j f(x_j), \tag{63}$$

in cui i nodi y_k, $k = 0, \ldots, m$, sono prefissati e i nodi x_j, $j = 0, 1, \ldots, n$, e i pesi α_k, $k = 0, \ldots, m$, e w_j, $j = 0, 1, \ldots, n$, sono determinati in modo da ottenere il massimo grado di precisione possibile, è detta *formula gaussiana con nodi prefissati.* ∎

Il seguente teorema, per la cui dimostrazione si rimanda a [5], permette di determinare il grado di precisione della formula (63).

7.35 Teorema. *Siano $r(x)$ e $s(x)$ i polinomi*

$$r(x) = \prod_{k=0}^{m} (x - y_k), \quad s(x) = \prod_{j=0}^{n} (x - x_j).$$

La formula (63) ha grado di precisione almeno $m + 2n + 2$ se e solo se:

a) è esatta per tutti i polinomi di grado minore o uguale a $m + n + 2$;

b) vale

$$\int_a^b \omega(x)r(x)s(x)p(x)\ dx = 0$$

per ogni polinomio $p(x)$ di grado al più n. ∎

In pratica, dati il peso $\omega(x)$ e il polinomio $r(x)$, si tratta di determinare una famiglia di polinomi $s(x)$, non necessariamente monici, ortogonali su $[a, b]$ rispetto al peso $\omega(x)r(x)$. Le formule di questo tipo più usate sono le seguenti:

formula di Radau

$m = 0,\ y_0 = a,\ \omega(x) \equiv 1,$ grado di precisione $2n + 2,$

formula di Lobatto

$m = 1,\ y_0 = a,\ y_1 = b,\ \omega(x) \equiv 1,$ grado di precisione $2n + 3.$

Si fissi per semplicità l'intervallo $[-1, 1]$. Per costruire le formule di Radau si devono determinare i polinomi ortogonali su $[-1, 1]$ rispetto al peso $r(x) = x + 1$. Poiché per l'i-esimo polinomio di Legendre è $P_i(-1) = (-1)^i$, si ha che $P_i(-1) + P_{i+1}(-1) = 0$, quindi $P_i(x) + P_{i+1}(x)$ è divisibile per $x + 1$ e la funzione

$$s_i(x) = \frac{P_i(x) + P_{i+1}(x)}{x + 1}$$

è un polinomio di grado i. Inoltre la successione $\{s_i(x)\}_{i\in\mathbf{N}}$ è costituita da polinomi ortogonali rispetto al peso $r(x)$ (si veda l'esercizio 7.40). Si scelgono quindi i nodi x_i, $i = 0, \ldots, n$, come zeri del polinomio $s_{n+1}(x)$. I coefficienti sono (si veda l'esercizio 7.40)

$$\alpha_0 = \frac{2}{(n+2)^2}, \quad w_i = \frac{1 - x_i}{(n+2)^2 P_{n+1}^2(x_i)}, \quad i = 0, \ldots, n.$$

La formula di Radau risulta allora

$$S_{n+2} = \frac{1}{(n+2)^2}\ \Big[2f(-1) + \sum_{i=0}^{n} \frac{1 - x_i}{P_{n+1}^2(x_i)}\ f(x_i)\Big], \tag{64}$$

e il resto è (si veda l'esercizio 7.40)

$$r_{n+2} = \frac{2^{2n+3}(n+2)[(n+1)!]^4}{[(2n+3)!]^3}\ f^{(2n+3)}(\xi), \quad \xi \in (-1, 1). \tag{65}$$

Nella tabella di figura 7.13 sono riportati i nodi, i coefficienti e i resti delle formule di Radau del tipo S_{n+2}.

n	y_0, x_i	α_0, w_i	r_{n+2}
0	-1 0.3333333333	0.5 1.5	$0.741\ 10^{-1} f^{(3)}(\xi)$
1	-1 -0.2898979486 0.6898979486	0.2222222222 1.024971652 0.7528061254	$0.889\ 10^{-3} f^{(5)}(\xi)$
2	-1 -0.5753189235 0.1810662711 0.822824081	0.125 0.65768864 0.7763869377 0.4409244224	$0.518\ 10^{-5} f^{(7)}(\xi)$
3	-1 -0.7204802713 -0.1671808647 0.4463139727 0.8857916078	0.08 0.4462078022 0.623653046 0.5627120303 0.2874271216	$0.178\ 10^{-7} f^{(9)}(\xi)$
4	-1 -0.8029298284 -0.3909285467 0.1240503795 0.6039731643 0.9203802859	0.05555555556 0.3196407532 0.4853871885 0.5209267832 0.4169013343 0.2015883853	$0.401\ 10^{-10} f^{(11)}(\xi)$

Fig. 7.13 - Nodi, coefficienti e resti delle formule di Radau del tipo S_{n+2}, $n = 0, 1, \ldots, 4$.

Per costruire le formule di Lobatto si devono determinare i polinomi ortogonali su $[-1, 1]$ rispetto al peso $r(x) = x^2 - 1$. Poiché per $i \neq j$ è (si veda l'esercizio 6.15 l))

$$\int_{-1}^{1} (x^2 - 1) P_i'(x) P_j'(x)\ dx = 0,$$

i polinomi $\{P_i'(x)\}_{i \in \mathbf{N}}$ sono ortogonali rispetto al peso $r(x)$ e si può assumere $s_i(x) = P_{i+1}'(x)$. Si scelgono quindi i nodi x_i, $i = 0, \ldots, n$, come zeri di $s_{n+1} = P_{n+2}'(x)$. I coefficienti sono (si veda l'esercizio 7.41)

$$\alpha_0 = \alpha_1 = \frac{2}{(n+2)(n+3)}, \quad w_i = \frac{2}{(n+2)(n+3) P_{n+2}^2(x_i)}, \quad i = 0, \ldots, n.$$

La formula di Lobatto risulta allora

$$S_{n+3} = \frac{2}{(n+2)(n+3)} \Big[f(-1) + f(1) + \sum_{i=0}^{n} \frac{f(x_i)}{P_{n+2}^2(x_i)} \Big], \qquad (66)$$

e il resto è (si veda l'esercizio 7.41)

$$r_{n+3} = \frac{-(n+3)(n+2)^3 2^{2n+5}[(n+1)!]^4}{(2n+5)[(2n+4)!]^3} \, f^{(2n+4)}(\xi), \quad \xi \in (-1,1). \qquad (67)$$

Nella tabella di figura 7.14 sono riportati i nodi, i coefficienti e i resti delle formule di Lobatto del tipo S_{n+3}.

n	$y_{0,1}, x_i$	$\alpha_{0,1}, w_i$	r_{n+1}
0	± 1 0	0.3333333333 1.333333333	$-0.111\ 10^{-1} f^{(4)}(\xi)$
1	± 1 ± 0.4472135955	0.1666666667 0.8333333333	$-0.847\ 10^{-4} f^{(6)}(\xi)$
2	± 1 ± 0.6546536707 0	0.1 0.5444444444 0.7111111111	$-0.360\ 10^{-6} f^{(8)}(\xi)$
3	± 1 ± 0.7650553239 ± 0.2852315165	0.06666666667 0.3784749563 0.554858377	$-0.970\ 10^{-9} f^{(10)}(\xi)$
4	± 1 ± 0.8302238963 ± 0.4688487935 0	0.04761904762 0.2768260474 0.4317453812 0.4876190476	$-0.180\ 10^{-11} f^{(12)}(\xi)$
5	± 1 ± 0.8717401485 ± 0.5917001814 ± 0.2092992179	0.03571428571 0.2107042271 0.3411226925 0.4124587947	$-0.243\ 10^{-14} f^{(14)}(\xi)$

Fig. 7.14 - Nodi, coefficienti e resti delle formule di Lobatto del tipo S_{n+3}, $n = 0, 1, \ldots, 5$.

Le formule di Radau e di Lobatto sono usate in particolare quando i valori della $f(x)$ in uno o in entrambi gli estremi dell'intervallo sono noti

o facilmente calcolabili (come ad esempio nel caso in cui la $f(x)$ ha una singolarità apparente agli estremi).

Se $f(-1) = 0$ la formula di Radau si riduce a

$$S_{n+2} = \sum_{i=0}^{n} w_i f(x_i)$$

e ha grado di precisione $2n+2$, cioè 1 in più della formula di Gauss-Legendre con lo stesso numero di nodi. Analogamente se $f(\pm 1) = 0$, la formula di Lobatto ha grado di precisione $2n+3$, cioè 2 in più della formula di Gauss-Legendre con lo stesso numero di nodi.

7.36 Esempio. Si applicano le formule di Gauss-Legendre e di Lobatto al calcolo di

$$S = \int_{-1}^{1} \frac{(\mathrm{e}^{x^2} - \mathrm{e})^2}{x^2 - 1}\, dx,$$

in cui la $f(x)$ ha una singolarità apparente per $x = \pm 1$, dove può essere prolungata per continuità con il valore 0 (è $S = -4.778502$). Al crescere di n si ottengono i valori

	Lobatto		Gauss-Legendre	
n	S_{n+3}	err_{n+3}	S_{n+1}	err_{n+1}
0	-3.936654	$0.842\ 10^{0}$	-5.904984	$0.113\ 10^{1}$
1	-4.668013	$0.110\ 10^{0}$	-5.248358	$0.470\ 10^{0}$
2	-4.767337	$0.112\ 10^{-1}$	-4.855290	$0.768\ 10^{-1}$
3	-4.777578	$0.924\ 10^{-3}$	-4.787048	$0.855\ 10^{-2}$
4	-4.778434	$0.687\ 10^{-4}$	-4.779243	$0.740\ 10^{-3}$

Si noti come la formula di Lobatto S_{n+3} con $n = 3$ (e quindi con 4 valutazioni della funzione) risulti altrettanto accurata quanto la formula di Gauss-Legendre S_{n+1} con $n = 4$ (e quindi con 5 valutazioni della funzione). ∎

Altre importanti formule gaussiane con nodi prefissati sono quelle di *Kronrod*, in cui gli y_k, $k = 0, \ldots, m$, sono i nodi dell'$(m+1)$-esimo polinomio di Legendre e $n = m + 1$. Si può dimostrare [5] che i nodi x_j, $j = 0, \ldots, n$, appartengono all'intervallo $(-1, 1)$, che sono separati dai punti y_k, che i pesi α_k, $k = 0, \ldots, m$, e w_j, $j = 0, \ldots, n$, sono positivi e che il grado di precisione è $3m + 4$. Per i nodi e i pesi delle formule di Kronrod con $7, 10, 15, 20, 25, 30$ nodi si veda [16]. Queste formule consentono di superare una delle principali obiezioni all'uso delle formule gaussiane, e cioè che nel passare da una

formula con n nodi ad una con $m > n$ nodi non è possibile sfruttare i calcoli fatti perché non vi sono nodi comuni. Sulle formule di Kronrod è infatti basato lo *schema di Patterson* per la quadratura automatica.

8. Quadratura automatica

Problemi di quadratura numerica si presentano nei più disparati settori scientifici, e così accade che chi deve calcolare numericamente degli integrali molto spesso non è un analista numerico esperto. Per questo negli ultimi anni è stato sviluppato software che consente un agevole calcolo numerico di integrali anche ad utenti privi di nozioni specialistiche.

Per *programma di quadratura automatica* si intende un programma che tipicamente:

a) riceve come dati in ingresso la funzione $f(x)$, l'intervallo $[a, b]$ di solito finito, una tolleranza ϵ_{tol} sull'errore e il limite massimo N di valutazioni di funzione ammesse;

b) fornisce come risultati il valore dell'integrale calcolato S_c, una stima ϵ_{st} dell'errore e, talvolta, il numero delle valutazioni di funzione M effettivamente compiute.

Il programma cerca di calcolare il valore S dell'integrale (1) con un errore stimato $\epsilon_{st} \leq \epsilon_{tol}$ e con un numero di valutazioni $M \leq N$. Se ciò non è possibile, l'esecuzione viene interrotta quando il numero di valutazioni supera N, con un appropriato messaggio.

La stima dell'errore effettuata dal programma è una stima euristica che vale sotto determinate condizioni sulla funzione integranda. Per ampie classi di funzioni la condizione $\epsilon_{st} \leq \epsilon_{tol}$ implica che

$$\big|S - S_c\big| \leq \epsilon_{tol}, \quad \text{o che} \quad \frac{\big|S - S_c\big|}{\big|S\big|} \leq \epsilon_{tol},$$

a seconda che si consideri l'errore assoluto o quello relativo. Questo però non vale in generale, anzi, per ogni programma di quadratura automatica è possibile costruire funzioni per cui l'errore effettivamente commesso è arbitrariamente più grande della tolleranza ϵ_{tol} e dell'errore stimato ϵ_{st}.

Alcuni programmi di quadratura automatica forniscono in uscita degli indicatori dell'affidabilità dei risultati, ma in ogni caso è fortemente da sconsigliare una utilizzazione non ragionata di questi programmi di quadratura automatica.

7.37 Esempio. Si consideri la funzione definita nell'intervallo $[a, b]$ e dipendente dai parametri α, β e δ

$$f(x) = \begin{cases} \alpha & \text{se } |x - \beta| < \delta, \\ \\ 0 & \text{altrimenti,} \end{cases}$$

dove $\beta \in [a, b]$. L'integrale esatto è $S = 2\alpha\delta$ e, fissato δ, cresce a piacere con α. Se δ è molto piccolo, scegliendo opportunamente β in modo che il programma non valuti $f(x)$ in alcun punto dell'intervallo $[\beta - \delta, \beta + \delta]$, si ottiene verosimilmente il valore $S_c = 0$. ∎

Un programma di quadratura automatica viene realizzato utilizzando uno *schema di quadratura automatica* e una *strategia di comportamento*. Lo schema di quadratura automatica è costituito dai seguenti elementi:

a) una successione di formule di quadratura che comportano un numero sempre crescente di valutazioni di funzione;
b) un criterio per determinare il valore S_c che può essere quello ottenuto dalla formula con il maggior numero di valutazioni, oppure dall'applicazione di tecniche di estrapolazione (si veda ad esempio l'estrapolazione di Richardson nel paragrafo 3);
c) un criterio per determinare la stima dell'errore ϵ_{st}.

Le strategie di comportamento si distinguono fondamentalmente in *adattive* e *non adattive*. In una strategia non adattiva i punti in cui si valuta la funzione $f(x)$ sono scelti senza tenere conto del comportamento della $f(x)$; può accadere allora che se la funzione ha un comportamento molto irregolare in una parte dell'intervallo di integrazione, questo comporti un numero elevato di nodi su tutto l'intervallo, anche dove la funzione ha un comportamento più regolare. Invece in una strategia adattiva il numero dei nodi viene scelto in base al comportamento della funzione. Si suddivide l'intervallo di integrazione in sottointervalli e si applica ricorsivamente a questi l'algoritmo di quadratura automatica, tenendo conto di opportune condizioni di arresto. La funzione integranda viene così valutata in pochi punti nei sottointervalli in cui ha un andamento regolare e in tanti punti negli intervalli dove sono presenti irregolarità della funzione. Nel seguito sono presentati alcuni dei più comuni schemi di quadratura automatica.

Lo schema di *Patterson* è basato su una famiglia di formule con 3, 7, 15, 31, 63, 127, 255 nodi, ricavate da un procedimento simile a quello di Kronrod: si applica prima la formula di Gauss-Legendre su 3 punti, poi formule con nodi prefissati, aggiungendo agli n nodi già utilizzati altri $n+1$ nodi; il grado di precisione della formula con $2n+1$ nodi è $3n+1$. Il criterio di arresto adottato è

$$\frac{|S_{2n+1} - S_n|}{|S_{2n+1}|} \leq \epsilon_{tol}, \quad \text{per} \quad 2n + 1 \leq 255.$$

Lo schema di *Romberg* è basato sulla applicazione della estrapolazione di Richardson alla formula dei trapezi

$$J_2^{(N)} = \frac{h}{2}\,[f(a) + 2\sum_{i=1}^{N-1} f(x_i) + f(b)], \quad h = \frac{b-a}{N}\,, \quad x_i = a + ih.$$

L'idea base del metodo consiste nel calcolare $J_2^{(N)}$ per diversi valori di N (e perciò di h) e quindi utilizzare questi valori per ottenere una più accurata approssimazione dell'integrale mediante estrapolazione.

Data la successione $N_k = 2^k$, $k = 0, 1, \ldots$, si considerino i valori così definiti

$$\begin{aligned} T_0^{(k)} &= J_2^{(N)}, \quad \text{con} \quad N = N_k,\ k = 0, 1, \ldots, \\ T_j^{(k)} &= \frac{4^j T_{j-1}^{(k+1)} - T_{j-1}^{(k)}}{4^j - 1}\,, \quad k = 0, 1, \ldots,\ j = 1, 2, \ldots, \end{aligned} \tag{68}$$

con cui si costruisce la seguente tabella:

$$\begin{array}{llll} T_0^{(0)} & & & \\ T_0^{(1)} & T_1^{(0)} & & \\ T_0^{(2)} & T_1^{(1)} & T_2^{(0)} & \\ T_0^{(3)} & T_1^{(2)} & T_2^{(1)} & T_3^{(0)} \\ \vdots & & & \ddots \end{array}$$

Se $f(x) \in C^{2m+1}[a, b]$ dalla formula di Eulero-Maclaurin (si veda l'esercizio 4.42) risulta che

$$J_2^{(N)} - S = c_2 h^2 + c_4 h^4 + \ldots + c_{2m} h^{2m} + O\big(h^{2m+1}\big), \quad h = \frac{b-a}{2^k}\,,$$

in cui per $2 \le i \le 2m$ è

$$c_i = \frac{B_i}{i!}\,\big[f^{(i-1)}(b) - f^{(i-1)}(a)\big] = \frac{B_i(b-a)}{i!}\,f^{(i)}(\xi_i), \quad \xi_i \in (a, b).$$

È possibile dimostrare (si veda l'esercizio 7.45) che per $j < m$ è

$$T_j^{(k)} - S = c_{2(j+1)}\delta_{j,j+1} h^{2(j+1)} + \ldots + c_{2m}\delta_{j,m} h^{2m} + O\big(h^{2m+1}\big), \tag{69}$$

con

$$\lim_{j\to\infty} \delta_{j,j+1} = 0,$$

per cui ogni $T_j^{(k)}$ rappresenta un'approssimazione di S ottenuta con una formula avente grado di precisione $2j+1$. Se $f(x)$ è analitica in una regione

contenente $[a, b]$ si può dimostrare [5] che la successione $\{T_j^{(0)}\}$ sulla diagonale converge ad S in modo superlineare. Il criterio di arresto adottato è

$$\epsilon_{st} = \frac{|T_k^{(0)} - T_{k-1}^{(0)}|}{|T_k^{(0)}|} < \epsilon_{tol}.$$

Vi sono delle varianti di questo schema che consistono in scelte diverse della successione N_k, in un diverso procedimento di estrapolazione sui valori $T_0^{(k)}$, nell'applicazione di criteri diversi di arresto. Inoltre è possibile, studiando opportunamente l'andamento delle differenze tra i valori della tabella, ottenere delle stime sull'affidabilità del risultato ottenuto [5].

7.38 Esempio. Si applica lo schema di Romberg al calcolo dell'integrale

$$S = \int_0^1 \mathrm{e}^{-x^2}\, dx.$$

Si sceglie la successione $N_k = 2^k$ e $\epsilon_{tol} = 0.5\ 10^{-4}$. Poiché il calcolo di $T_0^{(k)}$ sfrutta tutti i valori calcolati per $T_0^{(k-1)}$, il numero totale di valutazioni di $f(x)$ è $1+2^m$, dove m è il valore dell'indice k a cui si arresta il procedimento. La tabella che si ottiene è la seguente (sono riportate la prima colonna e la diagonale):

k	$T_0^{(k)}$	$T_k^{(0)}$
0	0.6839395	0.6839395
1	0.7313700	0.7471801
2	0.7429836	0.7468331
3	0.7458652	0.7468236

Sono state richieste 9 valutazioni della $f(x)$ e si assume il valore $T_3^{(0)} = 0.7468236$ come approssimazione di S. L'errore assoluto effettivo è di circa $0.534\ 10^{-6}$. Per ottenere la stessa precisione con la sola applicazione della formula dei trapezi e l'estrapolazione di Richardson sarebbero necessarie 257 valutazioni della funzione. ∎

Lo schema di Romberg può essere combinato con una strategia adattiva. Si consideri ad esempio il seguente algoritmo:

a) si applica il metodo di Romberg all'integrale sull'intervallo $[a, b]$ per al più 4 passi; se la stima dell'errore è inferiore alla tolleranza si ottiene la stima definitiva dell'integrale sull'intervallo;

b) se la stima dell'errore non è inferiore alla tolleranza, l'intervallo viene diviso in due parti uguali alle quali viene riapplicato il procedimento descritto nel punto a).

Ad ogni passo è possibile sfruttare i valori della funzione calcolati precedentemente. Per una efficiente utilizzazione di questi valori è necessaria una oculata gestione della memoria.

7.39 Esempio. Si applica lo schema di Romberg con la strategia adattiva al calcolo dell'integrale

$$S = \int_0^1 1 - \sqrt{1-x}\; dx.$$

Nella figura 7.15 è riportato il grafico della funzione $1 - \sqrt{1-x}$ sull'intervallo $[0, 1]$; in ascissa sono indicati i nodi in cui la funzione viene valutata dall'algoritmo per ottenere una stima dell'integrale con un'errore $\epsilon_{st} \leq \epsilon_{tol} = 0.5\ 10^{-4}$. L'integrale esatto è $1/3$, l'integrale stimato 0.3333502, quindi l'errore assoluto effettivo è circa $0.169\ 10^{-4}$. Il calcolo ha richiesto 28 valutazioni della funzione. Per ottenere la stessa precisione con la formula dei trapezi sarebbero state necessarie 513 valutazioni della funzione e con il metodo di Romberg 257 valutazioni. ∎

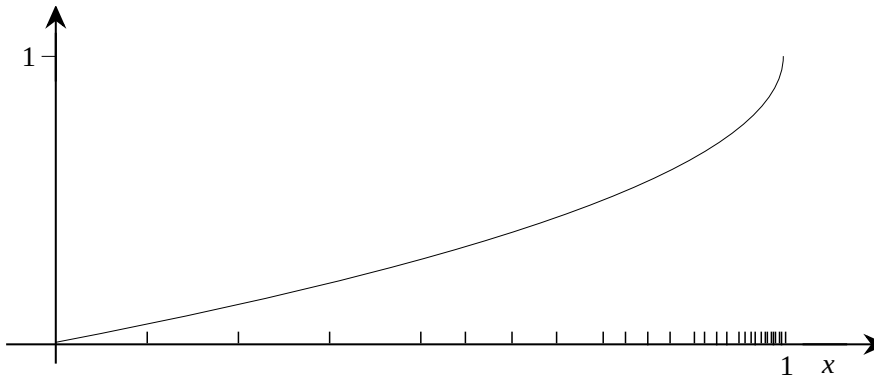


Fig. 7.15 - Grafico della funzione $1 - \sqrt{1-x}$ sull'intervallo $[0, 1]$ e nodi dove essa viene valutata dall'algoritmo di quadratura adattiva.

Uno dei programmi attualmente più usati per l'integrazione automatica è il CADRE (=Cautious Adaptive Romberg Extrapolation), realizzato da De Boor [6]. È un programma molto efficiente, che consente anche di riconoscere le singolarità agli estremi dell'intervallo.

Lo schema di *Clenshaw-Curtis* utilizza formule con cui si calcolano successive approssimazioni dell'integrale mediante trasformate di coseni, ed è quindi possibile sfruttare gli algoritmi veloci per il calcolo della trasformata

discreta di Fourier. Le formule di Clenshaw-Curtis sono formule interpolatorie che approssimano l'integrale

$$\int_{-1}^{1} f(x)\ dx$$

nei nodi i punti $x_i = \cos\theta_i$, $\theta_i = \dfrac{i\pi}{n}$ $i = 0, \ldots, n$. Supponendo per semplicità n pari, la formula di Clenshaw-Curtis è data da (si veda l'esercizio 7.48)

$$S_{n+1} = \sum_{i=0}^{n} w_i f(x_i), \qquad w_0 = w_n = \frac{1}{n^2-1}\,,$$

$$w_i = w_{n-i} = \frac{2}{n}\left[1 + 2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n-2} \frac{\cos j\theta_i}{1-j^2} \;+\; \frac{(-1)^i}{1-n^2}\right], \quad \text{per} \quad i = 1, \ldots, \frac{n}{2}\,.$$

Si ha quindi, cambiando l'ordine delle sommatorie,

$$S_{n+1} = \frac{1}{n^2-1}\,\big[f(x_0) + f(x_n)\big] + b_0 + 2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n-2} \frac{b_j}{1-j^2} \;+\; \frac{b_n}{1-n^2}\,,$$

dove

$$b_j = \frac{2}{n}\sum_{i=1}^{n-1} f(\cos\theta_i)\cos j\theta_i, \quad j = 0, 2, \ldots, n-2.$$

I coefficienti b_j vengono calcolati mediante una trasformata di coseni (si veda l'esercizio 5.62). Per l'analisi dell'errore si usa la stima

$$\epsilon_{st} = \frac{|S_{cn+1} - S_{n+1}|}{|S_{cn+1}|}\,,$$

dove c è un intero piccolo, in pratica $c = 2$ o $c = 3$).

L'algoritmo di quadratura automatica è quindi così strutturato:

a) si fissa un valore pari piccolo di n (per esempio $n = 4$) e si calcola S_{n+1};
b) si assegna ad n un nuovo valore cn e si calcola S_{n+1}, utilizzando tutti i valori già calcolati della funzione e i risultati precedentemente ottenuti;
c) si calcola ϵ_{st}; se $\epsilon_{st} \le \epsilon_{tol}$ oppure $n > N$, il procedimento si arresta con S_{cn+1} uguale all'ultimo valore calcolato, altrimenti si ripete il punto b).

QUADPACK[16] è un pacchetto di programmi per la quadratura automatica che utilizza molte delle formule studiate. È formato da una dozzina di programmi, fra cui due, di uso molto generale, QAG e QAGS, utilizzano formule di Kronrod. Il secondo è particolarmente adatto per il caso di funzioni con singolarità agli estremi dell'intervallo.

9. Integrazione in più dimensioni

L'integrazione approssimata in più dimensioni presenta difficoltà molto maggiori che quella in una dimensione: una valutazione di una funzione in più variabili richiede di solito più tempo; il numero di valutazioni richieste per ottenere una ragionevole approssimazione aumenta consistentemente con l'aumentare del numero delle dimensioni; il comportamento di una funzione in più variabili può presentare irregolarità molto maggiori che possono essere di tipo diverso per ogni singola variabile; l'insieme di integrazione può assumere una quantità di forme diverse e irregolari.

Ci si limiterà qui a considerare il caso più semplice dell'approssimazione di

$$S = \int_{a_1}^{b_1} \dots \int_{a_r}^{b_r} f(x_1, \dots, x_r)\ dx_1 \dots\ dx_r.$$

Per semplicità di notazione si considera il caso $r = 2$, in cui

$$S = \int_a^b \int_c^d f(x, y)\ dx\ dy,$$

ma la tecnica si estende facilmente al caso $r > 2$.

Fissati due interi m e n e scelte due formule di quadratura

$$S_{m+1} = \sum_{i=0}^{m} w_i g(x_i) \quad \text{e} \quad S_{n+1} = \sum_{i=0}^{n} z_i g(x_i)$$

per il calcolo dell'integrale di una funzione $g(x)$, una formula di quadratura per l'integrale in due dimensioni si ottiene come "prodotto" di S_{m+1} e S_{n+1} nel modo seguente

$$S_{m+1} \times S_{n+1} = \sum_{i=0}^{m} \sum_{j=0}^{n} w_i z_j f(x_i, y_j).$$

Vale il seguente teorema.

7.40 Teorema. *Se S_{m+1} ha grado di precisione k_1 e S_{n+1} ha grado di precisione k_2, allora $S_{m+1} \times S_{n+1}$ integra esattamente i monomi $x^s y^t$, con $s \le k_1$ e $t \le k_2$.*

Dim. Per $s \le k_1$ e $t \le k_2$ è

$$\int_a^b \int_c^d x^s y^t\ dx\ dy = \int_a^b x^s\ dx\ \int_c^d y^t\ dy = \sum_{i=0}^{m} w_i x_i^s \sum_{j=0}^{n} z_j y_j^t$$

$$= \sum_{i=0}^{m} \sum_{j=0}^{n} w_i z_j x_i^s y_j^t. \qquad \blacksquare$$

Ad esempio, per costruire la formula prodotto di due formule di Newton-Cotes S_3 (18) si pone

$$h = \frac{b-a}{2}, \quad k = \frac{d-c}{2}, \quad x_i = a + ih, \quad y_i = c + ik.$$

Si ottiene

$$\begin{aligned} S_3 \times S_3 = \frac{hk}{9} \big[& f(x_0, y_0) + 4f(x_1, y_0) + f(x_2, y_0) \\ & + 4f(x_0, y_1) + 16f(x_1, y_1) + 4f(x_2, y_1) \\ & + f(x_0, y_2) + 4f(x_1, y_2) + f(x_2, y_2) \big]. \end{aligned}$$

Nella figura 7.16 sono indicati, accanto ai nodi, i corrispondenti pesi moltiplicati per 9.

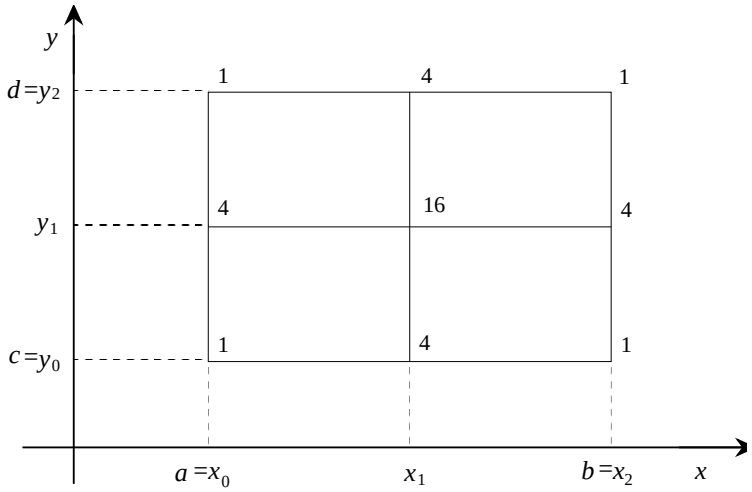


Fig. 7.16 - Nodi e pesi della formula $S_3 \times S_3$

Per quanto riguarda il resto, si dimostra (si veda l'esercizio 7.50) che se $f(x, y) \in C^4(D)$, dove $D = (a, b) \times (c, d)$, vale la relazione

$$\begin{aligned} r = S - S_3 \times S_3 = -\frac{hk}{45} \Big[h^4 \frac{\partial^4 f}{\partial x^4}(\xi_1, \eta_1) + k^4 \frac{\partial^4 f}{\partial y^4}(\xi_2, \eta_2) \Big], \\ (\xi_1, \eta_1),\ (\xi_2, \eta_2) \in D. \end{aligned} \tag{70}$$

Nello stesso modo si possono costruire le formule prodotto delle formule composte. Ad esempio, per la formula prodotto di Cavalieri-Simpson, posto

$$h = \frac{b-a}{M}, \quad k = \frac{d-c}{N}, \quad x_i = a + ih, \quad y_j = c + jk,$$

e

$$\sigma(y) = f(x_0, y) + 2\sum_{i=1}^{M-1} f(x_i, y) + 4\sum_{i=0}^{M-1} f\big(\frac{x_i + x_{i+1}}{2}, y\big) + f(x_M, y),$$

si ha

$$J_3^{(M)} \times J_3^{(N)} = \frac{hk}{36}\Big[\sigma(y_0) + 2\sum_{j=1}^{N-1}\sigma(y_j) + 4\sum_{j=0}^{N-1}\sigma\big(\frac{y_j + y_{j+1}}{2}\big) + \sigma(y_N)\Big].$$

Nella figura 7.17 sono indicati i nodi e i pesi (moltiplicati per 36) della formula per $M = 3$ e $N = 2$.

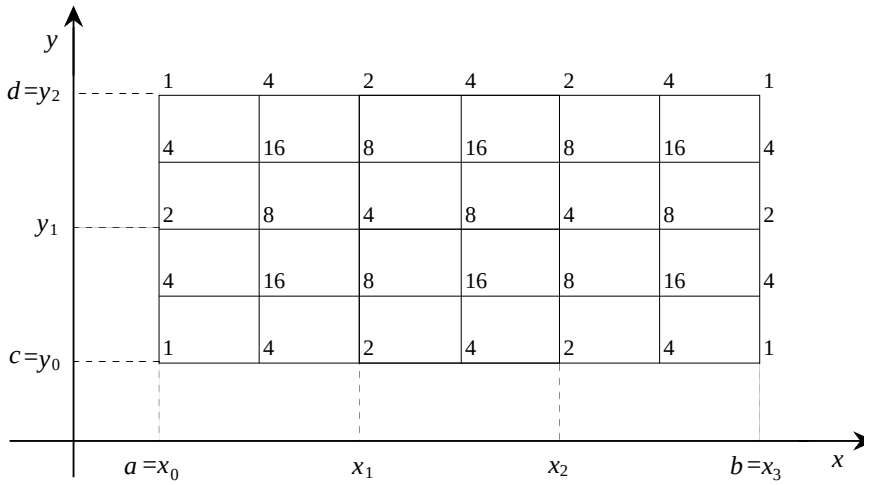


Fig. 7.17 - Nodi e pesi della formula $J_3^{(M)} \times J_3^{(N)}$

Il numero di valutazioni della funzione $f(x, y)$ richieste è $(2M+1)(2N+1)$. Il resto è

$$r = S - J_3^{(M)} \times J_3^{(N)} = -\frac{(b-a)(d-c)}{2880}\Big[h^4\frac{\partial^4 f}{\partial x^4}(\xi_1, \eta_1) + k^4\frac{\partial^4 f}{\partial y^4}(\xi_2, \eta_2)\Big],$$
$$(\xi_1, \eta_1),\ (\xi_2, \eta_2) \in D.$$

7.41 Esempio. Si approssima con la formula prodotto di Cavalieri-Simpson l'integrale

$$S = \int_0^{\pi/2}\int_0^{\pi/2} (x\cos y + y\cos x)\ dx\ dy,$$

assumendo $M = N$, e quindi $h = k = \dfrac{\pi}{2M}$. Posto $D = \big(0, \frac{\pi}{2}\big) \times \big(0, \frac{\pi}{2}\big)$, è

$$\max_{(x,y)\in D} \Big|\frac{\partial^4 f}{\partial x^4}\,(x,y)\Big| = \max_{(x,y)\in D} \Big|\frac{\partial^4 f}{\partial y^4}\,(x,y)\Big| = \frac{\pi}{2},$$

e si ha

$$|r| < \frac{(b-a)(d-c)}{2880}\; 2h^4\; \frac{\pi}{2} = \big(\frac{\pi}{2}\big)^7\; \frac{1}{1440M^4}\,.$$

Al variare di M si ottengono i valori elencati nella seguente tabella (in cui err_M sono gli errori effettivamente generati, essendo $S = \pi^2/4 = 2.467401$)

M	$J_3^{(M)} \times J_3^{(M)}$	err_M
2	2.467731	$0.330\ 10^{-3}$
4	2.467416	$0.153\ 10^{-4}$
8	2.467395	$0.572\ 10^{-5}$

Il numero di valutazioni di $f(x,y)$ richieste per $M = 8$ è $17^2 = 289$. Con la formula dei trapezi con $M = 128$, cioè con 16641 valutazioni della funzione $f(x,y)$, si ottiene un'approssimazione affetta da un errore di circa $0.5\ 10^{-4}$. Per valori più grandi di M l'errore effettivamente generato è maggiore, a causa dell'elevato errore di arrotondamento prodotto dalla somma di un così gran numero di termini. ∎

7.42 Esempio. Ponendo

$$x_m = \frac{a+b}{2}\,, \quad y_m = \frac{c+d}{2}\,, \quad h = \sqrt{\frac{3}{5}}\;\frac{b-a}{2}\,, \quad k = \sqrt{\frac{3}{5}}\;\frac{d-c}{2}\,,$$

dalla (51) si ricava la formula prodotto di Gauss-Legendre $S_3 \times S_3$

$$\begin{aligned} S_3 \times S_3 = \frac{(b-a)(d-c)}{324} \Big\{ & 25\big[f(x_m-h, y_m-k) + f(x_m+h, y_m-k) \\ & \quad + f(x_m-h, y_m+k) + f(x_m+h, y_m+k)\big] \\ & + 40\big[f(x_m, y_m-k) + f(x_m-h, y_m) \\ & \quad + f(x_m+h, y_m) + f(x_m, y_m+k)\big] \\ & + 64 f(x_m, y_m)\Big\}. \end{aligned}$$

Applicando questa formula al calcolo dell'integrale

$$S = \int_0^{\pi/2} \int_0^{\pi/2} (x\cos y + y\cos x)\; dx\; dy$$

dell'esempio precedente si ottiene, con solo 9 valutazioni della funzione $f(x, y)$, il valore 2.467418 che è affetto da un errore di circa $0.169\ 10^{-4}$. In modo analogo si costruisce la formula prodotto di Gauss-Legendre $S_4 \times S_4$, con la quale si ottiene, con solo 16 valutazioni della funzione $f(x, y)$, il valore 2.467395 che è affetto da un errore di circa $0.610\ 10^{-5}$. ∎

10. Metodo Monte Carlo

Un approccio completamente diverso al problema dell'integrazione numerica si ottiene con il cosiddetto *metodo di integrazione Monte Carlo* che consiste nel simulare un processo statistico in cui il valore atteso del risultato finale è il valore dell'integrale che si vuole calcolare.

Si consideri una variabile casuale ξ uniformemente distribuita nell'intervallo $[a, b]$ e la variabile casuale $\delta = (b-a)f(\xi)$. È facile verificare che δ ha media

$$\mu = \frac{1}{b-a} \int_a^b (b-a)f(x)\ dx = S$$

e varianza

$$\sigma^2 = \frac{1}{b-a} \int_a^b [S-(b-a)f(x)]^2 dx = (b-a)\int_a^b f(x)^2\ dx - S^2.$$

Il metodo Monte Carlo consiste dei seguenti passi:

a) si generano n valori $x_1, \ldots, x_n$ della variabile casuale ξ;

b) si calcola la media

$$S_n = \frac{b-a}{n} \sum_{i=1}^{n} f(x_i),$$

che si assume come approssimazione di S.

Il problema di generare n valori $x_1, \ldots, x_n$ di una variabile casuale uniforme ξ viene affrontato comunemente utilizzando un *generatore di numeri pseudocasuali*, che costruisce una sequenza di numeri equidistribuiti e scarsamente correlati tra loro, e che quindi gode di buone proprietà statistiche. Uno degli schemi più usati per generare sequenze pseudocasuali è quello delle sequenze lineari generate a partire da un intero x_0 assegnato con il metodo *congruenziale*

$$x_{i+1} \equiv \big[\alpha x_i + \beta\big] \pmod{m}, \tag{71}$$

dove α, β, m sono interi assegnati con $0 < \alpha, \beta < m$. I numeri così generati sono interi minori di m e quindi la sequenza $x_1, x_2, \ldots$ è periodica di periodo minore od uguale ad m. I numeri α e β vengono scelti in modo che il periodo

risulti quanto più lungo possibile. In [14] viene dimostrato che il periodo è esattamente m se e solo se

a) β è primo con m,
b) $\alpha \equiv 1 \pmod{p}$ per ogni fattore primo p di m,
c) $\alpha \equiv 1 \pmod{4}$ se m è multiplo di 4

(si veda l'esercizio 7.53 per il caso $m = 2^e$, con e intero positivo). Queste condizioni però non garantiscono che i numeri generati con la (71) non siano in qualche modo correlati fra di loro. È quindi necessario controllare che le sequenze generate soddisfino anche dei test statistici (si veda [14]). In [17], nella tabella di pag. 198, sono elencate alcune scelte delle costanti m, α e β che soddisfano tali test. I valori di m ivi riportati sono però abbastanza piccoli, in quanto scelti in modo da non generare errori di overflow nel calcolo della (71).

Un modo classico di implementare la (71) è quello di scegliere come m il minimo intero positivo non rappresentabile. Questo naturalmente dipende dal particolare calcolatore che si sta usando: ad esempio per un calcolatore che utilizza un'aritmetica intera in base 2 con 31 cifre, si sceglie $m = 2^{31}$. Così, se si verifica overflow, i risultati di operazioni modulo m sono ottenuti immediatamente con la perdita delle cifre più significative. Il procedimento risulta quindi molto più rapido. In [8] viene dato un programma per il calcolo di α e β, scelte secondo le indicazioni di [14]. In particolare, per $m = 2^{31}$ è $\alpha = 843314861$ e $\beta = 453816693$.

La (71), dividendo per m, genera una sequenza uniformemente distribuita nell'intervallo $[0, 1)$.

Per poter valutare la bontà del metodo di integrazione Monte Carlo è necessario stimare l'errore che si commette approssimando S con S_n. Il valore S_n è un valore assunto dalla variabile casuale

$$y_n = \frac{b-a}{n} \sum_{i=1}^{n} f(\xi_i),$$

dove $\xi_1, \ldots, \xi_n$ sono n variabili casuali indipendenti con la stessa distribuzione di probabilità di ξ. Per il teorema centrale di convergenza (si veda il paragrafo 11 del capitolo 2), la variabile y_n, per valori elevati di n, ha una distribuzione approssimativamente normale, con media $\mu = S$ e varianza σ^2, e quindi

$$\text{la probabilità che} \quad \big|S_n - S\big| < k\,\frac{\sigma}{\sqrt{n}} \quad \text{è data da} \quad \operatorname{erf}\big(\frac{k}{\sqrt{2}}\big), \tag{72}$$

dove $\operatorname{erf}(x)$ è la funzione errore. Risulta in particolare che

$$|S_n - S| \leq 1.645\frac{\sigma}{\sqrt{n}} \quad \text{con probabilità} \quad 0.9,$$
$$|S_n - S| \leq 2.576\frac{\sigma}{\sqrt{n}} \quad \text{con probabilità} \quad 0.99,$$
$$|S_n - S| \leq 3.291\frac{\sigma}{\sqrt{n}} \quad \text{con probabilità} \quad 0.999.$$

Ciò significa che fissato un valore di k la limitazione dell'errore decresce, al crescere di n, come $\sigma/\sqrt{n}$.

Da un semplice confronto con l'errore delle formule classiche di quadratura risulta che il metodo Monte Carlo non è conveniente: se n cresce di un fattore 100 la precisione aumenta solo di un fattore 10. Tuttavia questo metodo può essere utilizzato per il calcolo degli integrali multidimensionali, e in questo caso può divenire molto competitivo. Sia da calcolare l'integrale

$$S = \int_D f(\mathbf{x})\ d\mathbf{x}, \tag{73}$$

dove D è una regione finita in $\mathbf{R}^r$ e $d\mathbf{x}$ è l'elemento di volume. Sia P un parallelepipedo in r dimensioni, contenente D

$$P = \{\mathbf{x} \in \mathbf{R}^r \ : \ a_i \leq x_i \leq b_i,\ i = 1, \ldots, r\}, \quad D \subseteq P.$$

La funzione $f(\mathbf{x})$ può essere estesa in P assumendo $f(\mathbf{x}) = 0$ per $\mathbf{x} \in P - D$ e l'integrale (73) diviene

$$S = \int_P f(\mathbf{x})\ d\mathbf{x}.$$

Il calcolo del valore approssimato dell'integrale con il metodo Monte Carlo avviene nel modo seguente:

a) si generano n vettori $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (valori della variabile casuale vettoriale $\boldsymbol{\xi}$), uniformemente distribuiti in P;

b) si calcola la media

$$S_n = \frac{1}{n}\ \prod_{i=1}^{r}(b_i - a_i) \sum_{j=1}^{n} f(\mathbf{x}_j),$$

che si assume come approssimazione di S.

I vettori $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vengono generati utilizzando r sequenze pseudocasuali, tra loro non correlate, uniformemente distribuite negli intervalli $[a_i, b_i]$, $i = 1, \ldots, r$.

La (72) vale indipendentemente dalla dimensione r [5], e dal punto di vista computazionale questa tecnica di integrazione diviene tanto più vantaggiosa quando più r è elevato (nelle applicazioni il metodo Monte Carlo viene usualmente applicato quando r è superiore a 10).

7.43 Esempio. Si approssima con il metodo Monte Carlo l'integrale

$$S = \int_0^{\pi/2} \int_0^{\pi/2} (x \cos y + y \cos x)\, dx\, dy,$$

dell'esempio 7.41. Nella figura 7.18 sono riportati, in scala logaritmica, gli errori assoluti delle medie S_n per valori di n compresi fra 1 e 10^6 e il grafico di una retta di coefficiente angolare $-\dfrac{1}{2}$, che mostra come l'errore tenda a diminuire come $n^{-1/2}$.

Per $n = 10^6$, cioè con 10^6 valutazioni di $f(x, y)$, si ottiene il valore 2.466504, che ha un errore assoluto di circa $0.898\ 10^{-3}$ (si confronti con i valori ottenuti negli esempi 7.41 e 7.42). ∎

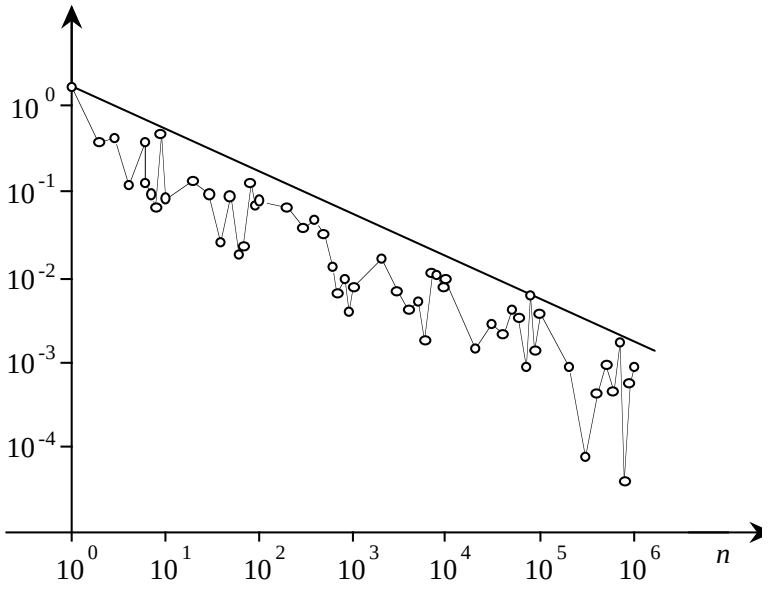


Fig. 7.18 - Errori nell'integrazione con il metodo Monte Carlo

La convergenza del metodo Monte Carlo risulta quindi molto lenta. È però possibile, utilizzando varie tecniche, diminuire il numero di valutazioni della funzione. Queste tecniche in generale riducono la varianza σ (si veda ad esempio l'esercizio 7.55), approssimando il problema dato con un opportuno problema modificato: riducendo la varianza di 100 volte, ad esempio, il numero di valutazioni di funzione può essere ridotto di 10 volte.

11. Approssimazione delle derivate

Le formule per l'approssimazione delle derivate vengono usate soprattutto in questi casi:

a) funzioni che sono note solo in un insieme discreto di punti,

b) discretizzazione di un problema differenziale, cioè trasformazione di un problema continuo in un problema discreto, con l'utilizzazione di tecniche alle differenze finite.

Nella maggior parte dei casi le formule che si usano sono interpolatorie, cioè ottenute derivando dei polinomi di interpolazione, e quindi esatte se applicate a polinomi di grado opportuno.

Derivando il polinomio di interpolazione di Lagrange (5, cap. 5), si ottengono le formule di derivazione approssimata

$$p_n'(x) = \sum_{i=0}^{n} L_i'(x) f(x_i),$$
$$p_n''(x) = \sum_{i=0}^{n} L_i''(x) f(x_i),$$
$$\dots \qquad \dots$$

Le formule più comunemente usate sono quelle che si ottengono quando i nodi sono equidistanti e il punto x coincide con uno di essi.

7.44 Esempio. Per $n = 2$, $x_1 = x_0 + h$, $x_2 = x_0 + 2h$, e $x = x_0$ si ha

$$L_0'(x_0) = \frac{2x_0 - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)} = -\frac{3}{2h},$$
$$L_1'(x_0) = \frac{x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)} = \frac{2}{h},$$
$$L_2'(x_0) = \frac{x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)} = -\frac{1}{2h},$$

da cui

$$p_2'(x_0) = \frac{1}{h}\Big[-\frac{3}{2}f(x_0) + 2f(x_1) - \frac{1}{2}f(x_2)\Big].$$

In modo analogo si ottengono le formule

$$p_2'(x_1) = \frac{1}{2h}\big[-f(x_0) + f(x_2)\big],$$
$$p_2'(x_2) = \frac{1}{h}\Big[\frac{1}{2}f(x_0) - 2f(x_1) + \frac{3}{2}f(x_2)\Big].$$

Per la derivata seconda si ha, per ogni x,

$$L_0''(x) = \frac{1}{h^2}\,, \quad L_1''(x) = -\frac{2}{h^2}\,, \quad L_2''(x) = \frac{1}{h^2}\,,$$

da cui

$$p_2''(x_0) = p_2''(x_1) = p_2''(x_2) = \frac{1}{h^2}\left[f(x_0) - 2f(x_1) + f(x_2)\right]. \qquad \blacksquare$$

In generale, se $p_n(x)$ è il polinomio di interpolazione di grado n della funzione $f(x)$ nei nodi distinti $x_0, \ldots, x_n$, si assume $p_n^{(k)}(x)$, per $k \le n$, come approssimazione di $f^{(k)}(x)$. Se il punto x coincide con uno dei nodi, ad esempio con x_j, la *formula di derivazione approssimata di ordine k* assume la forma

$$p_n^{(k)}(x_j) = \sum_{i=0}^{n} \omega_i f(x_i), \tag{74}$$

in cui i coefficienti ω_i dipendono da k e da j. Fissati k e j, per ricavare i coefficienti ω_i, anziché derivare il polinomio di interpolazione, si può fare ricorso al *metodo dei coefficienti indeterminati*, imponendo che la formula (74) abbia *grado di precisione* n, cioè che sia esatta per i polinomi di grado minore od uguale ad n. Per la linearità basterà imporre che la (74) sia esatta per le funzioni

$$f(x) = (x - x_j)^r, \quad \text{per} \quad r = 0, \ldots, n,$$

e poiché

$$f^{(k)}(x_j) = \begin{cases} 0 & \text{se } r \neq k, \\ k! & \text{se } r = k, \end{cases}$$

ne segue che gli ω_i, $i = 0, \ldots, n$, devono soddisfare il sistema lineare

$$\sum_{i=0}^{n} \omega_i (x_i - x_j)^r = \begin{cases} 0 & \text{se } r \neq k, \\ k! & \text{se } r = k, \end{cases} \qquad r = 0, \ldots, n. \tag{75}$$

Poiché la matrice del sistema (75) è di Vandermonde e i punti x_i sono distinti, si possono ricavare gli ω_i $i = 0, \ldots, n$.

Dalla prima equazione (75), poiché $k \ge 1$, risulta che $\sum_{i=0}^{n} \omega_i = 0$, cioè la somma dei coefficienti di una formula di derivazione approssimata è nulla.

Se i nodi x_i sono equidistanti, cioè $x_i = x_0 + ih$, $i = 0, \ldots, n$, si può porre $\alpha_i = h^k \omega_i$, dove gli α_i sono soluzione del sistema lineare

$$\sum_{i=0}^{n} \alpha_i (i - j)^r = \begin{cases} 0 & \text{se } r \neq k, \\ k! & \text{se } r = k, \end{cases} \qquad r = 0, \ldots, n. \tag{76}$$

In tal caso la formula di derivazione approssimata assume la forma

$$p_n^{(k)}(x_j) = \frac{1}{h^k} \sum_{i=0}^{n} \alpha_i f(x_i), \tag{77}$$

Il resto di una formula di derivazione di ordine k è

$$r_n^{(k)}(x) = f^{(k)}(x) - p_n^{(k)}(x).$$

7.45 Teorema. *Sia $f(x) \in C^{n+1}[a,b]$, siano $x_0 < \ldots < x_n \in [a,b]$ e sia $p_n(x)$ il polinomio di interpolazione. Allora esistono n punti distinti $y_i \in (x_i, x_{i+1})$, $i = 0, \ldots, n-1$, indipendenti da x, e un punto $\xi = \xi(x)$ tali che*

$$r_n'(x) = \prod_{i=0}^{n-1} (x - y_i)\, \frac{f^{(n+1)}(\xi)}{n!}\,. \tag{78}$$

Dim. La funzione $r_n(x) = f(x) - p_n(x)$, derivabile con continuità, si annulla negli $n+1$ punti $x_0, \ldots, x_n$ e quindi per il teorema di Rolle la funzione $r_n'(x)$ ha n zeri $y_0, \ldots, y_{n-1}$, che dipendono dalla funzione $f(x)$ e dai nodi dell'interpolazione, e tali che $y_i \in (x_i, x_{i+1})$, $i = 0, \ldots, n-1$. Quindi la (78) è verificata nei punti $x = y_i$. Per $x \neq y_i$ sia

$$s(x) = \frac{r_n'(x)}{\prod_{i=0}^{n-1} (x - y_i)}\,, \tag{79}$$

e si consideri la funzione della variabile y

$$z(y) = r_n'(y) - s(x) \prod_{i=0}^{n-1} (y - y_i).$$

La funzione $z(y)$ si annulla almeno negli $n+1$ punti distinti y_i, $i = 0, 1, \ldots$, $n-1$, e x. Proseguendo come nella dimostrazione del teorema 5.5, risulta che la $z^{(n)}(y)$ si annulla in almeno un punto $\xi \in (a,b)$, per il quale si ha:

$$0 = z^{(n)}(\xi) = r_n^{(n+1)}(\xi) - n!\; s(x) = f^{(n+1)}(\xi) - n!\; s(x),$$

da cui, sostituendo nella (79), si ottiene la (78) per $x \neq y_i$, $i = 0, \ldots, n-1$. ∎

In modo analogo si può ottenere il resto per la derivata k-esima, $k \leq n$,

$$r_n^{(k)}(x) = f^{(k)}(x) - p_n^{(k)}(x) = \prod_{i=0}^{n-k} (x - y_i)\, \frac{f^{(n+1)}(\xi)}{(n+1-k)!}\,,$$

dove $y_i \in (x_i, x_{i+k})$, per $i = 0, \dots, n-k$.

Per valutare il resto dell'approssimazione della derivata quando i nodi x_i sono equidistanti di passo h e il punto x coincide con uno dei nodi, conviene utilizzare la forma (22, cap. 5) del resto dell'interpolazione, che è derivabile per il teorema 5.28 se $f(x) \in C^{n+1}[x_0, x_n]$. Per la (29, cap. 5) si ha

$$r_n'(x) = \pi_n'(x) f[x_0, \dots, x_n, x] + \pi_n(x)\, f[x_0, \dots, x_n, x, x].$$

Poiché $f[x_0, \dots, x_n, x, x]$ è limitata in $[x_0, x_n]$ e $\pi_n(x_j) = 0$, risulta

$$r_n'(x_j) = \pi_n'(x_j) f[x_0, \dots, x_n, x_j],$$

e per il teorema 5.31 è

$$r_n'(x_j) = \pi_n'(x_j)\, \frac{f^{(n+1)}(\xi_j)}{(n+1)!}\,, \quad \text{dove} \quad \xi_j \in (x_0, x_n).$$

Con il cambiamento di variabile $x = x_0 + th$ è

$$\pi_n(x) = h^{n+1} \tau_n(t), \quad \text{dove} \quad \tau_n(t) = t(t-1)\dots(t-n),$$

e quindi

$$\pi_n'(x_j) = h^n \tau_n'(j).$$

Ne segue che

$$r_n'(x_j) = \gamma_j^{(1)} h^n f^{(n+1)}(\xi_j), \quad \text{dove} \quad \gamma_j^{(1)} = \frac{\tau_n'(j)}{(n+1)!}\,. \tag{80}$$

In modo analogo si ha

$$\begin{aligned} r_n^{(k)}(x_j) = \gamma_j^{(k)} h^{n-k+1} f^{(n+1)}(\xi_j^{(k)}) + O\big(h^{n-k+2}\big), \\ \text{dove} \quad \gamma_j^{(k)} = \frac{\tau_n^{(k)}(j)}{(n+1)!}. \end{aligned} \tag{81}$$

Quando $j = n/2$ e k è pari, risulta $\tau_n^{(k)}(j) = 0$, e se $f(x) \in C^{n+2}[x_0, x_n]$, si ha

$$r_n^{(k)}(x_j) = \delta_j^{(k)} h^{n-k+2} f^{(n+2)}(\xi_j^{(k)}) + O\big(h^{n-k+3}\big), \quad \text{dove } \delta_j^{(k)} = k\, \frac{\tau_n^{(k-1)}(j)}{(n+2)!}.$$

Le formule per cui $j = n/2$, cioè x_j è il punto centrale dell'intervallo $[x_0, x_n]$, sono dette *centrali* o *simmetriche*. Poiché tali formule hanno di solito coefficienti del resto più bassi, quando è possibile, conviene usarle.

7.46 Esempio. Sia $x_i = x_0 + ih$, $i = 0, \dots, n$. Formule per la derivata prima si ottengono ponendo $k = 1$. Per $n = 1$, $x = x_0$ dalla (76) risulta

$$\alpha_0 + \alpha_1 = 0 \quad \text{e} \quad \alpha_1 = 1,$$

da cui si ottiene

$$p_1'(x_0) = \frac{1}{h}\,\big((f(x_1) - f(x_0)\big), \tag{82}$$

che rappresenta l'approssimazione della derivata con il rapporto incrementale. Dalla (80) si ha

$$r_1'(x_0) = -\,\frac{h}{2}\,f''(\xi), \quad \xi \in (x_0, x_1).$$

Una formula centrale si ottiene ponendo $n = 2$ e $x = x_1$; dalla (76) risulta

$$p_2'(x_1) = \frac{1}{2h}\,\big(f(x_2) - f(x_0)\big). \tag{83}$$

Dalla (80) si ha

$$r_2'(x_1) = -\,\frac{h^2}{6}\,f'''(\xi), \quad \xi \in (x_0, x_2).$$

Una formula centrale per la derivata seconda si ottiene ponendo $n = 2$, $k = 2$ e $x = x_1$; dalla (76) risulta

$$p_2''(x_1) = \frac{1}{h^2}\,\big(f(x_2) - 2f(x_1) + f(x_0)\big)\,.$$

Poiché $\tau_2''(1) = 0$, il resto di questa formula viene espresso mediante la derivata quarta della $f(x)$

$$r_2''(x_1) = -\,\frac{h^2}{12}\,f^{(4)}(\xi), \quad \xi \in (x_0, x_2). \qquad \blacksquare$$

Le formule di derivazione approssimata con nodi x_i equidistanti possono essere espresse anche mediante l'operatore Δ di differenza finita. Dalla (25, cap. 5) il polinomio di interpolazione di Newton risulta

$$p_n(x_0 + th) = f(x_0) + \sum_{i=1}^{n} \frac{\tau_{i-1}(t)}{i!}\,\Delta^i f(x_0).$$

Per l'approssimazione della derivata prima, derivando rispetto a t e ponendo ad esempio $t = 0$ si ha

$$p'_n(x_0) = \frac{1}{h} \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i} \Delta^i f(x_0). \tag{84}$$

Derivando più volte si ottengono formule approssimate per le derivate di ordine superiore.

7.47 Esempio. Della funzione $f(x) = \sin x$ si suppongono noti i valori per $x_i = ih$, per $h = 0.1$ e $i = 0, \ldots, 4$. Per approssimare le derivate di $f(x)$ in $x = 0$ si costruisce la tabella delle differenze finite

x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$
0	0				
		0.09983342			
0.1	0.09983342		−0.00099754		
		0.09883588		−0.00098744	
0.2	0.1986693		−0.00198498		0.00001962
		0.0968509		−0.00096782	
0.3	0.2955202		−0.0029528		
		0.0938981			
0.4	0.3894183				

Per la (84) la derivata prima è approssimata da

$$p'_n(0) = \frac{1}{h} \left[\Delta f(0) - \frac{\Delta^2 f(0)}{2} + \frac{\Delta^3 f(0)}{3} + \ldots + (-1)^{n-1} \frac{\Delta^n f(0)}{n} \right].$$

Al crescere di n si ottengono i valori

$$p'_1(0) = 0.9983342,\ p'_2(0) = 1.003322,\ p'_3(0) = 1.00003,\ p'_4(0) = 0.9999813$$

(il valore esatto della derivata prima in 0 è 1). Per approssimare la derivata seconda si utilizza la formula

$$p''_n(0) = \frac{1}{h^2} \left[\Delta^2 f(0) - \Delta^3 f(0) + \frac{11}{12} \Delta^4 f(0) - \ldots \right],$$

ottenuta derivando due volte il polinomio di interpolazione di Newton. Al crescere di n si ottengono i valori

$$p''_2(0) = -0.099754,\ p''_3(0) = -0.00101,\ p''_4(0) = 0.0007885$$

(il valore esatto della derivata seconda in 0 è 0). Per approssimare la derivata terza si utilizza la formula

$$p_n'''(0) = \frac{1}{h^3}\left[\Delta^3 f(0) - \frac{3}{2}\Delta^4 f(0) + \dots\right],$$

ottenuta derivando tre volte il polinomio di interpolazione di Newton. Al crescere di n si ottengono i valori

$$p_3'''(0) = -0.9874401,\ p_4'''(0) = -1.01687$$

(il valore esatto della derivata terza in 0 è -1). Infine per la derivata quarta si può assumere l'approssimazione (il valore esatto è 0)

$$p_4^{(4)}(0) = \frac{1}{h^4}\Delta^4 f(0) = 0.1962. \qquad \blacksquare$$

Le formule di derivazione approssimata sono in generale molto meno stabili di quelle di interpolazione (si veda l'esercizio 7.58) e di quadratura, e l'approssimazione peggiora al crescere dell'ordine di derivazione. Infatti, esaminando per semplicità il caso di formule con nodi equidistanti, per la (77) i coefficienti di $p_n^{(k)}(x_j)$ sono della forma $\dfrac{\alpha_i}{h^k}$, con α_i indipendente da h, mentre per la (81) il resto $r_n^{(k)}(x_j)$ è dell'ordine di h^{n-k+1}. Perciò, se per diminuire l'errore analitico si scelgono valori di h piccoli, aumentano i coefficienti della formula, e poiché la somma dei coefficienti è nulla e i valori $f(x_i)$ di solito non differiscono molto fra loro, si generano inevitabilmente forti errori di cancellazione.

Se si indica con ϵ_i l'errore da cui è affetto $f(x_i)$, per la (77) l'errore inerente ϵ_p di $p_n^{(k)}(x_j)$ rispetto ai valori $f(x_i)$ è

$$\epsilon_p \doteq \frac{1}{h^k p_n^{(k)}(x_j)} \sum_{i=0}^{n} \alpha_i f(x_i)\epsilon_i,$$

da cui

$$|\epsilon_p| \dot{<} \frac{A\epsilon}{h^k}, \quad \text{dove} \quad \epsilon = \max_{i=0,n} |\epsilon_i|$$

e A è indipendente da h e da ϵ_i, mentre per l'errore analitico relativo ϵ_{an} è

$$|\epsilon_{an}| \dot{<} Bh^{n-k+1},$$

dove B è indipendente da h. Trascurando gli errori algoritmici nel calcolo della (77), l'errore risulta maggiorato in modulo dalla funzione di h

$$\varphi(h) = \frac{A\epsilon}{h^k} + Bh^{n-k+1},$$

che assume il valore minimo per h tale che $\varphi'(h) = 0$, cioè per

$$h = \overline{h} = \sqrt[n+1]{\frac{kA}{(n-k+1)B}}\ \sqrt[n+1]{\epsilon}.$$

Conviene quindi scegliere h dello stesso ordine di $\sqrt[n+1]{\epsilon}$. Se ϵ è dell'ordine della precisione di macchina, cioè $\epsilon = \beta^{-(t-1)}$, dove β è la base e t è il numero delle cifre significative dell'aritmetica di macchina, conviene allora scegliere h dell'ordine di $\beta^{-(t-1)/(n+1)}$ e in ogni caso il numero di cifre significative esatte è ridotto asintoticamente del fattore $1 - k/(n+1)$.

7.48 Esempio. Per calcolare la derivata seconda si possono usare le formule con i nodi equidistanti $x_i = x_0 + ih$, $i = 0, \ldots, 4$

$$\text{(a)}\quad p_4''(x_1) = \frac{1}{12h^2}\big(11f(x_0) - 20f(x_1) + 6f(x_2) + 4f(x_3) - f(x_4)\big),$$

$$\text{(b)}\quad p_4''(x_2) = \frac{1}{12h^2}\big(-f(x_0) + 16f(x_1) - 30f(x_2) + 16f(x_3) - f(x_4)\big)$$

(ricavate nell'esercizio 7.57). I resti, a meno di termini di ordine superiore in h, sono rispettivamente

$$\frac{h^3}{12}\,f^{(5)}(\xi) \quad \text{e} \quad \frac{h^4}{90}\,f^{(6)}(\xi).$$

Per la funzione $f(x) = \sin x$, supponendo che l'errore con cui sono calcolati i valori $f(x_i)$ sia dell'ordine di $\epsilon = 16^{-5}$, gli errori sono così maggiorati

$$\text{(a)}\quad p_4''(x_1)\varphi(h) \dot{<} \frac{42\epsilon}{12\,h^2} + \frac{1}{12}\,h^3, \qquad \text{(b)}\quad p_4''(x_1)\varphi(h) \dot{<} \frac{64\epsilon}{12\,h^2} + \frac{1}{90}\,h^4.$$

I valori di h che rendono minime queste maggiorazioni sono

$$\text{(a)}\quad \overline{h} = \sqrt[5]{28}\ 16^{-1} \approx 0.121, \qquad \text{(b)}\quad \overline{h} = \sqrt[6]{240}\ 16^{-5/6} \approx 0.247.$$

Al variare di h si ottengono i seguenti errori effettivi

$$e_1 = |f''(x_1) - p_4''(x_1)| \quad \text{e} \quad e_2 = |f''(x_2) - p_4''(x_2)|,$$

per $x_0 = 1$.

h	e_1	e_2
0.05	$0.404\ 10^{-3}$	$0.186\ 10^{-3}$
0.10	$0.123\ 10^{-3}$	$0.199\ 10^{-4}$
0.15	$0.567\ 10^{-4}$	$0.109\ 10^{-4}$
0.20	$0.169\ 10^{-3}$	$0.463\ 10^{-4}$
0.25	$0.200\ 10^{-3}$	$0.573\ 10^{-4}$
0.30	$0.174\ 10^{-3}$	$0.991\ 10^{-4}$

■

Esercizi proposti

7.1 Sia $f(x) \in C[a,b]$ e siano $x_1, \ldots, x_n \in [a,b]$. Si verifichi che se $\alpha_1, \ldots, \alpha_n$ sono numeri tutti dello stesso segno, allora esiste un punto $\xi \in (a,b)$ tale che:

$$\sum_{i=1}^{n} \alpha_i f(x_i) = f(\xi) \sum_{i=1}^{n} \alpha_i.$$

(Traccia: escluso il caso $\alpha_i = 0$ per ogni i, in cui la relazione è banalmente vera, si ponga

$$m = \min_{x \in [a,b]} f(x), \qquad M = \max_{x \in [a,b]} f(x),$$

per ogni i risulta $m \le f(x_i) \le M$. Supponendo che $\alpha_i \ge 0$, per $i = 1, \ldots, n$, è $m\,\alpha_i \le \alpha_i f(x_i) \le M\,\alpha_i$, e sommando membro a membro rispetto ad i è

$$m \sum_{i=1}^{n} \alpha_i \le \sum_{i=1}^{n} \alpha_i f(x_i) \le M \sum_{i=1}^{n} \alpha_i,$$

da cui

$$m \le \frac{1}{\sum_{i=1}^{n} \alpha_i} \sum_{i=1}^{n} \alpha_i f(x_i) \le M.$$

Si tenga poi conto del fatto che $f(x)$ assume tutti i valori compresi fra m e M. Si proceda in modo analogo se $\alpha_i \le 0$ per $i = 1, \ldots, n$.)

7.2 Sia $f(x) \in C^1[a,b]$ e $S = \int_a^b f(x)\,dx$.

a) Si scriva la formula di quadratura interpolatoria S_1, che si ottiene sostituendo a $f(x)$ il polinomio costante $f(a)$, e il resto;

b) si scriva la corrispondente formula composta $J_1^{(N)}$ (detta *formula dei rettangoli*) e il resto e se ne dia l'interpretazione geometrica;

c) si dia una maggiorazione del resto della formula dei rettangoli quando $f(x)$ è solo continua in $[a,b]$.

(Traccia:

a) $S_1 = hf(x_0), \quad x_0 = a, \quad h = b - a, \quad r_1 = \dfrac{h^2}{2} f'(\xi), \quad \xi \in (a,b);$

b) $J_1^{(N)} = h \sum_{k=0}^{N-1} f(z_k), \quad z_k = a + kh, \quad h = \dfrac{b-a}{N},$

$$R_1^{(N)} = \frac{(b-a)^2}{2N} f'(\xi), \quad \xi \in (a,b),$$

e quindi

$$|R_1^{(N)}| \le \frac{M_1(b-a)^2}{2N}\,, \quad M_1 = \max_{x\in[a,b]} |f'(x)|.$$

c) $$R_1^{(N)} = \int_a^b f(x)\,dx - h\sum_{k=0}^{N-1} f(z_k) = \sum_{k=0}^{N-1}\Big[\int_{z_k}^{z_{k+1}} f(x)\,dx - hf(z_k)\Big]$$
$$= \sum_{k=0}^{N-1}\int_{z_k}^{z_{k+1}} \big[f(x) - f(z_k)\big]\,dx.$$

Se $f(x)\in C[a,b]$, sia

$$\omega(\delta) = \max\,\big\{|f(x_1)-f(x_2)|\ :\ x_1,x_2\in[a,b],\ |x_1-x_2|\le\delta\big\}$$

il *modulo di continuità* di $f(x)$ su $[a,b]$. Allora

$$|R_1^{(N)}| \le (b-a)\omega(h).\ )$$

7.3 Sia $S_{n+1} = \sum_{i=0}^{n} w_i f(x_i)$ una formula di quadratura interpolatoria per il calcolo di $\int_0^1 f(x)\,dx$.

a) Si scriva la corrispondente formula composta $J_{n+1}^{(N)}$ per il calcolo di

$$\int_a^b f(x)\,dx;$$

b) si verifichi che se $f(x)\in C[a,b]$ e la formula ha grado di precisione almeno 0, allora

$$\lim_{N\to\infty} J_{n+1}^{(N)} = \int_a^b f(x)\,dx.$$

(Traccia:

a) $$J_{n+1}^{(N)} = h\sum_{i=0}^{n} w_i \sum_{k=0}^{N-1} f(z_k + x_i h), \quad z_k = a+kh, \quad h = \frac{b-a}{N}\,;$$

b) per $i=0,\dots,n$, si ha (per la definizione di modulo di continuità $\omega(h)$ si veda la traccia dell'esercizio 7.2c)

$$\Big|\int_a^b f(x)\,dx - h\sum_{k=0}^{N-1} f(z_k+x_ih)\Big| \le \sum_{k=0}^{N-1}\int_{z_k}^{z_{k+1}} \big|f(x) - f(z_k+x_ih)\big|\,dx$$
$$\le (b-a)\omega(h).$$

Poiché $f(x) \in C[a,b]$, è $\lim_{h\to 0} \omega(h) = 0$ e quindi

$$\lim_{h\to 0}\ h \sum_{k=0}^{N-1} f(z_k + x_i h) = \int_a^b f(x)\ dx.$$

Inoltre è $\sum_{i=0}^{n} w_i = 1$.)

7.4 Si approssimi l'integrale

$$\int_0^1 f(x)\ dx$$

per le seguenti funzioni

$$(1)\quad f(x) = x^{10}, \qquad (2)\quad f(x) = \log(x+1), \qquad (3)\quad f(x) = \arcsin x,$$

con le formule dei trapezi e di Cavalieri-Simpson, per valori crescenti di N. Si confrontino i valori ottenuti con l'integrale esatto e si dica come tendono a zero i resti al crescere di N. Si dia una spiegazione del diverso comportamento nei tre casi.

(Traccia: con la formula dei trapezi gli errori assoluti risultano

N	(1)	(2)	(3)
2	$0.160\ 10^{0}$	$0.103\ 10^{-1}$	$0.837\ 10^{-1}$
4	$0.484\ 10^{-1}$	$0.260\ 10^{-2}$	$0.316\ 10^{-1}$
8	$0.128\ 10^{-1}$	$0.651\ 10^{-3}$	$0.117\ 10^{-1}$
16	$0.324\ 10^{-2}$	$0.163\ 10^{-3}$	$0.427\ 10^{-2}$
32	$0.813\ 10^{-3}$	$0.4121\ 10^{-4}$	$0.154\ 10^{-2}$

con la formula di Cavalieri-Simpson gli errori assoluti risultano

N	(1)	(2)	(3)
2	$0.114\ 10^{-1}$	$0.350\ 10^{-4}$	$0.143\ 10^{-1}$
4	$0.903\ 10^{-3}$	$0.256\ 10^{-5}$	$0.506\ 10^{-2}$
8	$0.598\ 10^{-4}$	$0.656\ 10^{-6}$	$0.179\ 10^{-2}$
16	$0.376\ 10^{-5}$		$0.635\ 10^{-3}$
32	$0.179\ 10^{-6}$		$0.225\ 10^{-3}$

Il diverso comportamento è spiegabile in termini delle derivate seconde e quarte delle funzioni.)

7.5 Si dica in quanti sottointervalli deve essere suddiviso l'intervallo $[0,1]$ affinché sia minore di $0.5\ 10^{-3}$ l'errore analitico relativo che si commette approssimando con le formule dei trapezi e di Cavalieri-Simpson l'integrale

$$S = \int_0^1 f(x)\ dx$$

per le seguenti funzioni

$$(1)\quad f(x) = \frac{1}{1+x}\,, \qquad (2)\quad f(x) = x\sqrt{1+x^2}, \qquad (3)\quad f(x) = x \sin \pi x,$$

$$(4)\quad f(x) = x^2 \mathrm{e}^x, \qquad (5)\quad \frac{1}{1+x}\ \cos\frac{\pi x}{2}\,, \qquad (6)\quad f(x) = \exp\big(\sin\frac{\pi x}{2}\big).$$

(Traccia: siano $M_2 = \max\limits_{x\in[0,1]} |f''(x)|$ e $M_4 = \max\limits_{x\in[0,1]} |f^{(4)}(x)|$; risulta

$$(1)\qquad f''(x) = \frac{2}{(1+x)^3}, \quad M_2 = |f''(0)| = 2,$$
$$f^{(4)}(x) = \frac{24}{(1+x)^5}, \quad M_4 = |f^{(4)}(0)| = 24.$$

Poiché

$$|f(x)| \geq \frac{1}{2}\,, \quad \text{è} \quad |S| > \frac{1}{2}\,, \quad \Big|\frac{R_2^{(N)}}{S}\Big| < \frac{1}{3N^2}\,, \quad \Big|\frac{R_3^{(N)}}{S}\Big| < \frac{1}{60N^4}\,,$$

e quindi $N = 26$ per la formula dei trapezi e $N = 3$ per la formula di Cavalieri-Simpson.

$$(2)\qquad f''(x) = \frac{3x + 2x^3}{\sqrt{(1+x^2)^3}}, \quad M_2 = |f''(1)| = \frac{5}{2\sqrt{2}}\,,$$
$$f^{(4)}(x) = \frac{-15x}{\sqrt{(1+x^2)^7}}, \quad M_4 < 4.$$

Poiché

$$|f(x)| \geq x, \quad \text{è} \quad |S| > \frac{1}{2}\,, \quad \Big|\frac{R_2^{(N)}}{S}\Big| < \frac{5}{12\sqrt{2}N^2}\,, \quad \Big|\frac{R_3^{(N)}}{S}\Big| < \frac{1}{360N^4}\,,$$

e quindi $N = 25$ per la formula dei trapezi e $N = 2$ per la formula di Cavalieri-Simpson.

(3) $$\begin{aligned} &f''(x) = 2\pi\cos\pi x - \pi^2 x\sin\pi x, \quad M_2 < 10,\\ &f^{(4)}(x) = -4\pi^3\cos\pi x + \pi^4 x\sin\pi x, \quad M_4 < 150. \end{aligned}$$

Poiché

$$|f(x)| \ge (x-1)(0.6-3x), \quad \text{per } x \in [0.2,1], \quad \text{è} \quad |S| > \frac{1}{4}\,,$$

$$\Big|\frac{R_2^{(N)}}{S}\Big| < \frac{10}{3N^2}\,, \quad \Big|\frac{R_3^{(N)}}{S}\Big| < \frac{5}{24N^4}\,,$$

e quindi $N = 82$ per la formula dei trapezi e $N = 5$ per la formula di Cavalieri-Simpson.

(4) $$\begin{aligned} &f''(x) = \mathrm{e}^x(2+4x+x^2), \quad M_2 = |f''(1)| = 7\mathrm{e},\\ &f^{(4)}(x) = \mathrm{e}^x(12+8x+x^2), \quad M_4 = |f^{(4)}(1)| = 21\mathrm{e}. \end{aligned}$$

Poiché

$$|f(x)| \ge x^2 + x^3, \quad \text{è} \quad |S| > \frac{7}{12}\,, \quad \Big|\frac{R_2^{(N)}}{S}\Big| < \frac{\mathrm{e}}{N^2}\,, \quad \Big|\frac{R_3^{(N)}}{S}\Big| < \frac{\mathrm{e}}{80N^4}\,,$$

e quindi $N = 74$ per la formula dei trapezi e $N = 3$ per la formula di Cavalieri-Simpson.

(5) $$\begin{aligned} f''(x) &= \frac{1}{4(1+x)^3}\,\Big[\cos\frac{\pi x}{2}\,(8-\pi^2(1+x)^2) + 4\pi\sin\frac{\pi x}{2}\,(1+x)\Big],\\ f^{(4)}(x) &= \frac{1}{16(1+x)^5}\,\Big[\cos\frac{\pi x}{2}(384 - 48\pi^2(1+x)^2 + \pi^4(1+x)^4)\\ &\qquad + \sin\frac{\pi x}{2}\,(192\pi(1+x) - 8\pi^3(1+x)^3)\Big],\\ M_2 &= |f''(1)| = \frac{\pi}{4}\,, \quad M_4 = |f^{(4)}(1)| = \frac{\pi^3 - 6\pi}{8} < 2. \end{aligned}$$

Poiché

$$|f(x)| \ge \frac{1-x}{2}\,, \quad \text{è} \quad |S| > \frac{1}{4}\,, \quad \Big|\frac{R_2^{(N)}}{S}\Big| < \frac{\pi}{12N^2}\,, \quad \Big|\frac{R_3^{(N)}}{S}\Big| < \frac{1}{360N^4}\,,$$

e quindi $N = 23$ per la formula dei trapezi e $N = 2$ per la formula di Cavalieri-Simpson.

$$
\begin{aligned}
(6) \qquad f''(x) &= \frac{1}{4}\,\pi^2 \exp\big(\sin\frac{\pi x}{2}\big)\ \big(1 - \sin\frac{\pi x}{2}\ - \sin^2\frac{\pi x}{2}\big),\\
f^{(4)}(x) &= \frac{1}{16}\,\pi^4 \exp\big(\sin\frac{\pi x}{2}\big)\ \big(-3 - 5\sin\frac{\pi x}{2}\\
&\qquad + 5\sin^2\frac{\pi x}{2} + 6\sin^3\frac{\pi x}{2}\ + \sin^4\frac{\pi x}{2}\big)\\
M_2 = |f''(1)| &= \frac{\mathrm{e}\pi^2}{4} < 7, \quad M_4 = |f^{(4)}(1)| = \frac{\mathrm{e}\pi^4}{4} < 67.
\end{aligned}
$$

Poiché

$$
|f(x)| \geq 1 + x, \quad \text{è} \quad |S| > \frac{3}{2}\,, \quad \Big|\frac{R_2^{(N)}}{S}\Big| < \frac{7}{18N^2}\,, \quad \Big|\frac{R_3^{(N)}}{S}\Big| < \frac{1}{64N^4}\,,
$$

e quindi $N = 28$ per la formula dei trapezi e $N = 3$ per la formula di Cavalieri-Simpson.)

7.6 Si approssimi l'integrale

$$
\int_{21.4}^{21.5} \log x\ dx
$$

a) per mezzo di una primitiva,

b) con la formula dei due punti,

operando in entrambi i casi con 6 cifre significative. Si confrontino gli errori.

(Traccia: a) una primitiva è $F(x) = x\log x - x$ e operando con 6 cifre si ha

$$
F(21.4) = 44.1565, \quad F(21.5) = 44.4631, \quad F(21.5) - F(21.4) = 0.3066;
$$

b) con la formula dei due punti, operando con 6 cifre si ha

$$
S_2 = \frac{0.1}{2}\,(\log 21.5 + \log 21.4) = 0.306572.
$$

Il valore ottenuto in a) è affetto da un errore di circa $0.276\ 10^{-4}$, mentre quello ottenuto in b) è affetto da un errore di circa $0.374\ 10^{-6}$.)

7.7 Si studi l'errore algoritmico che si genera nel calcolo della formula di quadratura

$$
S_{n+1} = \sum_{i=0}^{n} w_i f(x_i), \quad w_i > 0,
$$

trascurando gli errori da cui sono affetti gli $f(x_i)$.

(Traccia: dall'esempio 2.28 segue che

$$|\epsilon_{alg}| < \frac{u}{|S_{n+1}|} \sum_{i=1}^{n} \left\{ \left| \sum_{j=0}^{i} w_j f(x_j) \right| + \left| w_i f(x_i) \right| \right\},$$

in cui u è la precisione di macchina usata. Se $f(x_i) > 0$ per $i = 0, \ldots, n$, è $|\epsilon_{alg}| < (n+1)u$ e l'algoritmo è stabile, altrimenti l'algoritmo può essere instabile se a valori elevati di $f(x_i)$ corrisponde un risultato di modulo piccolo.)

7.8 Per approssimare $\log 2$ si possono usare le seguenti formule:

a) la formula di Taylor di $\log(1+x)$, troncata ad un opportuno termine;

b) formule di quadratura per valutare

$$\log 2 = \int_1^2 \frac{1}{x}\, dx.$$

Si dica quanti termini sono richiesti con la formula di Taylor (anche con la trasformazione di Eulero) e con le formule dei trapezi e di Cavalieri-Simpson con estrapolazione di Richardson, per ottenere un'approssimazione di $\log 2$ con un errore relativo minore di 10^{-5}.

(Traccia: Con la formula di Taylor sommando 7687 termini si ottiene il valore 0.6931541, con un errore relativo di circa $0.998\ 10^{-5}$ (per il valore calcolato con la trasformazione di Eulero si veda l'esempio 4.9, in cui si ottiene l'approssimazione con un errore in modulo minore di 10^{-6} sommando 12 termini). Con la formula dei trapezi e l'estrapolazione di Richardson con $N = 128$ (quindi sommando 129 termini) si ottiene 0.6931403, con un errore relativo di circa $0.993\ 10^{-5}$. Con la formula di Cavalieri-Simpson e l'estrapolazione di Richardson con $N = 4$ (quindi sommando 9 termini) si ottiene 0.6931472, con un errore relativo di circa $0.280\ 10^{-7}$.)

7.9 Si calcoli

$$S = \int_0^1 y(x)\, dx,$$

dove la funzione $y(x)$ è definita implicitamente da $x = ye^y$, con un errore minore di 10^{-4}.

(Traccia: la funzione $y(x)$ è la funzione inversa della funzione $f(y) = ye^y$ ed è tale che $f(0) = 0$ e $f(y) > 0$ per $y > 0$. Quindi

$$\int_0^b f^{-1}(x)\, dx = bf^{-1}(b) - \int_0^{f^{-1}(b)} f(y)\, dy.$$

In questo caso è $b = 1$ e $f^{-1}(b) = 0.5671433$. Si calcoli l'integrale con la formula di Cavalieri-Simpson e l'estrapolazione di Richardson. Si ottiene $S = 0.3303661$.)

7.10 Si esprimano le formule di quadratura di Newton-Cotes per mezzo delle differenze finite della funzione $f(x)$ nei nodi equidistanti $x_i = x_0 + ih$, $i = 0, \dots, n$, con $x_0 = a$ e $x_n = b$.

(Traccia: posto $x = x_0 + th$, è per la (26, cap. 5)

$$p(x_0 + th) = f(x_0) + \sum_{i=1}^{n} \binom{t}{i} \Delta^i f(x_0),$$

da cui

$$S_{n+1} = \int_{x_0}^{x_n} p(x)\ dx = h \sum_{i=0}^{n} \alpha_{n,i}\ \frac{\Delta^i f(x_0)}{i!}\,,$$

dove

$$\alpha_{n,i} = \begin{cases} n & \text{per } i = 0, \\ \displaystyle\int_0^n t(t-1)\dots(t-i+1)\ dt & \text{per } i \geq 1. \end{cases}$$

Per $n = 1$ si ha

$$S_2 = h\big[f(x_0) + \frac{1}{2}\Delta f(x_0)\big],$$

per $n = 2$ si ha

$$S_3 = h\big[2f(x_0) + 2\Delta f(x_0) + \frac{1}{3}\Delta^2 f(x_0)\big],$$

per $n = 3$ si ha

$$S_4 = h\big[3f(x_0) + \frac{9}{2}\Delta f(x_0) + \frac{9}{4}\Delta^2 f(x_0) + \frac{3}{8}\Delta^3 f(x_0)\big].\ )$$

7.11 Posto $h = \dfrac{b-a}{n}$, $x_i = a + ih$, $i = 0, \dots, n$, si costruiscano le formule di quadratura interpolatorie assumendo come nodi i punti x_i, $i = 1, \dots, n-1$. Tali formule sono dette *di Newton-Cotes di tipo aperto* e non fanno intervenire i valori della funzione $f(x)$ nei punti a e b. Si dia l'espressione del resto e si costruiscano le formule per $n = 2, \dots, 5$.

(Traccia: indicando con T_{n-1} la formula di Newton-Cotes di tipo aperto con $n-1$ nodi, si ha

$$T_{n-1} = h \sum_{i=1}^{n-1} \alpha_i f(x_i), \quad \alpha_i = \int_0^n \prod_{\substack{j=1 \\ j \neq i}}^{n-1} \frac{t-j}{i-j}\ dt.$$

Per il resto si proceda come nella dimostrazione del teorema 7.10. È

$$r_{n-1} = \int_a^b (x - x_1) \dots (x - x_{n-1}) f[x_1, \dots, x_{n-1}, x] \, dx.$$

Per n pari si ponga

$$\sigma_{n-2}(x) = \int_a^x (u - x_1) \dots (u - x_{n-1}) \, du,$$

e si verifichi che

$$\sigma_{n-2}(a) = \sigma_{n-2}(b) = 0, \qquad \sigma_{n-2}(x) < 0 \quad \text{per} \quad a < x < b.$$

Risulta

$$r_{n-1} = \gamma_n h^{s+1} \, \frac{f^{(s)}(\xi)}{s!}, \quad \xi \in (a, b),$$

dove
$$\begin{cases} s = n \quad \text{e} \quad \gamma_n = \displaystyle\int_0^n \tau_{n-1}(t) \, dt, & \text{per } n \text{ pari}, \\ s = n - 1 \quad \text{e} \quad \gamma_n = \displaystyle\int_0^n \tau_{n-2}(t - 1) \, dt, & \text{per } n \text{ dispari}. \end{cases}$$

Per $n = 2, \dots, 5$ si ha

$$T_1 = 2hf(x_1), \quad r_1 = \frac{1}{3} h^3 f''(\xi),$$
$$T_2 = \frac{3h}{2} \big[f(x_1) + f(x_2)\big], \quad r_2 = \frac{3}{4} h^3 f''(\xi),$$
$$T_3 = \frac{4h}{3} \big[2f(x_1) - f(x_2) + 2f(x_3)\big], \quad r_3 = \frac{14}{45} h^5 f^{(4)}(\xi),$$
$$T_4 = \frac{5h}{24} \big[11f(x_1) + f(x_2) + f(x_3) + 11f(x_4)\big], \quad r_4 = \frac{95}{144} h^5 f^{(4)}(\xi).$$

Si noti che T_1 coincide con la formula (49) e che non tutti i coefficienti di T_3 sono positivi.)

7.12 Si verifichi che

a) il coefficiente γ_n del resto delle formule di Newton-Cotes (31) è negativo per ogni n;

b) l'analogo coefficiente per le formule di Newton-Cotes di tipo aperto (si veda l'esercizio 7.11) è positivo per ogni n.

(Traccia: si proceda come per la dimostrazione del teorema 7.10 per verificare che, posto

$$\sigma_n(t) = \int_0^t \tau_n(u) \, du,$$

per n pari, poiché $\sigma_n(0) = \sigma_n(n) = 0$, è

$$\int_0^n \sigma_n(t)\ dt = -\int_0^n t\tau_n(t)\ dt,$$

e che

per n pari è $\sigma_n(t) > 0$ per $0 < t < n$,
per n dispari è $\sigma_n(t) < 0$ per $0 < t < n$.

b) si proceda in modo analogo.)

7.13 Sia $f(x) \in C^2[a,b]$, con $f''(x) > 0$ per $x \in [a,b]$. Fissato N, siano $J_2^{(N)}$ e $G_2^{(N)}$ le approssimazioni di $\int_a^b f(x)\ dx$ ottenute con la formula dei trapezi e con la formula dei punti di mezzo. Si verichi che

$$G_2^{(N)} \le \int_a^b f(x)\ dx \le J_2^{(N)}.$$

(Traccia: si tenga conto dei segni dei corrispondenti resti.)

7.14 Si dica in quanti sottointervalli deve essere suddiviso l'intervallo $[0,1]$ affinché sia minore di $0.5\ 10^{-3}$ l'errore analitico relativo che si commette approssimando con le formule composte di Gauss-Legendre (formula dei punti di mezzo e formula (52)) l'integrale

$$S = \int_0^1 f(x)\ dx$$

per le seguenti funzioni

$$(1)\quad f(x) = \frac{1}{1+x}\,, \qquad (2)\quad f(x) = x\sqrt{1+x^2}, \qquad (3)\quad f(x) = x\sin \pi x,$$

$$(4)\quad f(x) = x^2 \mathrm{e}^x, \qquad (5)\quad \frac{1}{1+x}\ \cos\frac{\pi x}{2}\,, \qquad (6)\quad f(x) = \exp\big(\sin\frac{\pi x}{2}\big)$$

(le funzioni sono le stesse dell'esercizio 7.5).

(Traccia: si veda l'esercizio 7.5 per le maggiorazioni delle derivate; indicando con N_1 il numero dei punti richiesto dalla formula dei punti di mezzo e con N_2 il numero dei punti richiesto dalla formula (52), si ha:

$$(1)\quad \Big|\frac{R_1^{(N)}}{S}\Big| < \frac{1}{6N^2}\,, \quad \Big|\frac{R_2^{(N)}}{S}\Big| < \frac{1}{90N^4}\,, \quad N_1 = 19, \quad N_2 = 3;$$

(2) $\left|\dfrac{R_1^{(N)}}{S}\right| < \dfrac{5}{24\sqrt{2}N^2}\,, \quad \left|\dfrac{R_2^{(N)}}{S}\right| < \dfrac{1}{540N^4}\,, \quad N_1 = 18, \quad N_2 = 2;$

(3) $\left|\dfrac{R_1^{(N)}}{S}\right| < \dfrac{5}{3N^2}\,, \quad \left|\dfrac{R_2^{(N)}}{S}\right| < \dfrac{5}{36N^4}\,, \quad N_1 = 58, \quad N_2 = 5;$

(4) $\left|\dfrac{R_1^{(N)}}{S}\right| < \dfrac{\mathrm{e}}{2N^2}\,, \quad \left|\dfrac{R_2^{(N)}}{S}\right| < \dfrac{\mathrm{e}}{120N^4}\,, \quad N_1 = 53, \quad N_2 = 3;$

(5) $\left|\dfrac{R_1^{(N)}}{S}\right| < \dfrac{\pi}{24N^2}\,, \quad \left|\dfrac{R_2^{(N)}}{S}\right| < \dfrac{1}{540N^4}\,, \quad N_1 = 17, \quad N_2 = 2;$

(6) $\left|\dfrac{R_1^{(N)}}{S}\right| < \dfrac{7}{36N^2}\,, \quad \left|\dfrac{R_2^{(N)}}{S}\right| < \dfrac{1}{96N^4}\,, \quad N_1 = 20, \quad N_2 = 3.\ )$

7.15 a) Si verifichi che i coefficienti w_i delle formule di Gauss-Legendre possono essere espressi anche nel modo seguente

$$w_i = \frac{2(1-x_i^2)}{(n+1)^2 P_n^2(x_i)} = \frac{2}{(1-x_i^2)[P'_{n+1}(x_i)]^2}\,.$$

b) Si scriva una procedura per il calcolo dei nodi e dei coefficienti della formula di quadratura di Gauss-Legendre che utilizzi il metodo di Newton come descritto nell'esercizio 6.20.

(Traccia: a) si applichi l'esercizio 6.15 d) alla (47).

b)

$m := \left\lceil \dfrac{n+1}{2} \right\rceil$;
for $i := 1$ **to** m **do begin**
  $z := \cos \dfrac{(4i-1)\pi}{4n+6}$;
  repeat
    $p := 1;\ p_1 := 0;$
    for $j := 1$ **to** $n+1$ **do begin**
      $p_2 := p_1;\ p_1 := p;\ p := zp_1;$
      $p := \dfrac{j-1}{j}\,(p - p_2) + p$
    end;
    $q := \dfrac{(n+1)(zp - p_1)}{z^2 - 1}$;
    $z_1 := z;\ z := z_1 - \dfrac{p}{q}$
  until $(|z - z_1| < \epsilon)$;
  $x_{i-1} := -z;\ x_{n+1-i} := z;$
  $w_{i-1} := \dfrac{2}{(1-z^2)q^2};\ w_{n+1-i} := w_{i-1}$
end;

7.16 Si ricavino le formule di Gauss-Legendre con il metodo dei coefficienti indeterminati. Si applichi al caso particolare di $n = 1$.

(Traccia: dalla (6) segue che w_i e x_i, $i = 0, \ldots, n$, soddisfano il sistema non lineare

$$\sum_{i=0}^{n} w_i x_i^j = m_j, \quad j = 0, \ldots, 2n+1, \tag{85}$$

dove

$$m_j = \begin{cases} \dfrac{2}{j+1} & \text{per } j \text{ pari,} \\ 0 & \text{per } j \text{ dispari.} \end{cases}$$

Indicato con

$$p(x) = \prod_{i=0}^{n}(x - x_i) = \sum_{j=0}^{n+1} a_j x^j, \quad a_{n+1} = 1,$$

si ha per $k = 0, \ldots, n$

$$\sum_{j=0}^{n+1} a_j m_{j+k} = \sum_{i=0}^{n} w_i x_i^k \sum_{j=0}^{n+1} a_j x_i^j = \sum_{i=0}^{n} w_i x_i^k p(x_i),$$

e poiché $p(x_i) = 0$, ne segue che

$$\sum_{j=0}^{n+1} a_j m_{j+k} = 0, \quad k = 0, \ldots, n. \tag{86}$$

Risolvendo questo sistema lineare di $n+1$ equazioni nelle $n+1$ incognite $a_0, \ldots, a_n$, si ottengono i coefficienti di $p(x)$. Si noti che, poiché $m_j = 0$ per j dispari, risulta che gli a_j, con indice j pari se n è pari e con indice j dispari se n è dispari, sono nulli. Poiché la matrice del sistema è fortemente malcondizionata, questo metodo è computazionalmente meno efficiente di quello dell'esercizio precedente. I nodi x_i vengono calcolati come zeri di $p(x)$ e i pesi w_i vengono calcolati risolvendo il sistema (85), che è diventato lineare, per $j = 0, \ldots, n$. Per $n = 1$ si ha

$$\begin{cases} w_0 + w_1 = 2 \\ w_0 x_0 + w_1 x_1 = 0 \\ w_0 x_0^2 + w_1 x_1^2 = \dfrac{2}{3} \\ w_0 x_0^3 + w_1 x_1^3 = 0, \end{cases}$$

mentre dal sistema (86) si ottiene

$$a_0 = -\frac{1}{3}, \quad a_1 = 0, \quad a_2 = 1.$$

Il polinomio $p(x) = x^2 - \frac{1}{3}$ ha gli zeri $x_0 = -\frac{1}{\sqrt{3}}$, $x_1 = \frac{1}{\sqrt{3}}$, e risolvendo il sistema

$$\begin{cases} w_0 + w_1 = 2 \\ -\frac{1}{\sqrt{3}}\, w_0 + \frac{1}{\sqrt{3}}\, w_1 = 0, \end{cases}$$

si ha $w_0 = w_1 = 1$.)

7.17 Facendo riferimento all'esercizio 6.21, si dica come si calcolano i nodi e i coefficienti delle formule gaussiane per mezzo degli autovalori e autovettori di matrici tridiagonali simmetriche. In particolare si calcolino i nodi e i coefficienti delle formule di Gauss-Laguerre per $n = 6$.

(Traccia: si costruisca la matrice

$$J = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_n \\ & & \beta_n & \alpha_{n+1} \end{bmatrix},$$

dove

$$\alpha_i = 0, \quad \beta_i = \frac{i}{\sqrt{4i^2 - 1}}, \quad \text{per le formule di Gauss-Legendre,}$$
$$\alpha_i = 2i - 1, \quad \beta_i = -i, \quad \text{per le formule di Gauss-Laguerre,}$$
$$\alpha_i = 0, \quad \beta_i = \frac{\sqrt{2i}}{2}, \quad \text{per le formule di Gauss-Hermite.}$$

Gli autovalori x_i, $i = 0, \ldots, n$, di J sono gli zeri dell'$(n+1)$-esimo polinomio ortogonale, e quindi sono i nodi della formula gaussiana S_{n+1}. Indicato con $\mathbf{y}^{(j)}$ l'autovettore di J corrispondente a x_j, normalizzato in modo che $y_0^{(j)} = \frac{1}{\sqrt{h_0}}$, cioè in modo che

$$y_0^{(j)} = \frac{1}{\sqrt{2}}, \quad \text{per le formule di Gauss-Legendre,}$$
$$y_0^{(j)} = 1, \quad \text{per le formule di Gauss-Laguerre,}$$
$$y_0^{(j)} = \frac{1}{\sqrt[4]{\pi}}, \quad \text{per le formule di Gauss-Hermite,}$$

risulta $w_j = \dfrac{1}{\|\mathbf{y}^{(j)}\|_2^2}$. Nel caso delle formule di Gauss-Laguerre per $n = 6$ si ha

$$J = \begin{bmatrix} 1 & -1 & & & & & \\ -1 & 3 & -2 & & & & \\ & -2 & 5 & -3 & & & \\ & & -3 & 7 & -4 & & \\ & & & -4 & 9 & -5 & \\ & & & & -5 & 11 & -6 \\ & & & & & -6 & 13 \end{bmatrix},$$

da cui si ottengono gli autovalori

$$x_0 = 0.1930437,\ x_1 = 1.026665,\ x_2 = 2.567877,\ x_3 = 4.900353,$$
$$x_4 = 8.182153,\ x_5 = 12.73418,\ x_6 = 19.39573$$

e la matrice degli autovettori corrispondente

$$Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.807 & -0.02666 & -1.568 & -3.9 & -7.182 & -11.73 & -18.4 \\ 0.6325 & -0.5263 & -0.8388 & 3.206 & 18.11 & 56.61 & 150.3 \\ 0.4756 & -0.6793 & 0.3653 & 2.707 & -14.42 & -138.1 & -709 \\ 0.3349 & -0.6197 & 1.034 & -0.9837 & -9.32 & 155.5 & 2084 \\ 0.2094 & -0.4447 & 1.038 & -2.972 & 10.01 & -5.67 & -3767 \\ 0.09811 & -0.2229 & 0.5968 & -2.202 & 12.47 & -128 & 3533 \end{bmatrix},$$

e risulta

$$w_0 = 0.4093190,\ w_1 = 0.4218313,\ w_2 = 0.1471263,\ w_3 = 0.2063351\ 10^{-1},$$
$$w_4 = 0.1074010\ 10^{-2},\ w_5 = 0.1586546\ 10^{-4},\ w_6 = 0.3170315\ 10^{-7}.\ )$$

7.18 Si determinino i parametri delle formule

a) $T_1 = \alpha f(a) + \beta f(b) + \gamma f'(a) + \delta f'(b)$,

b) $\overline{T}_1 = \alpha f(a) + \beta f\big(\dfrac{a+b}{2}\big) + \gamma f(b) + \delta f'(a) + \epsilon f'\big(\dfrac{a+b}{2}\big) + \zeta f'(b)$,

c) $T_2 = \alpha f(a) + \beta f(b) + \gamma f'(a) + \delta f'(b) + \epsilon f''(a) + \zeta f''(b)$,

per l'approssimazione dell'integrale

$$S = \int_a^b f(x)\ dx,$$

in modo che il grado di precisione sia il più elevato possibile. Si scrivano poi le corrispondenti formule composte e si applichino in particolare al calcolo di

$$(1)\quad \int_{-1}^{1} \frac{1}{1+x^2}\,dx, \qquad (2)\quad \int_{0}^{\pi/2} \sin x\,dx.$$

(Traccia: a) con il metodo dei coefficienti indeterminati si ha

$$\begin{cases} \alpha+\beta = b-a \\ \alpha a+\beta b+\gamma+\delta = \dfrac{b^2-a^2}{2} \\ \alpha a^2+\beta b^2+2\gamma a+2\delta b = \dfrac{b^3-a^3}{3} \\ \alpha a^3+\beta b^3+3\gamma a^2+3\delta b^2 = \dfrac{b^4-a^4}{4}\,, \end{cases}$$

da cui, posto $h=b-a$, si ottiene

$$\alpha=\beta=\frac{h}{2}\,, \qquad \gamma=-\delta=\frac{h^2}{12}\,,$$

grado di precisione 3; in modo analogo si ottiene per b)

$$\alpha=\gamma=\frac{7h}{30}\,, \qquad \beta=\frac{16h}{30}\,, \qquad \delta=-\zeta=\frac{h^2}{60}\,, \qquad \epsilon=0,$$

grado di precisione 5; per c)

$$\alpha=\beta=\frac{h}{2}\,, \qquad \gamma=-\delta=\frac{h^2}{10}\,, \qquad \epsilon=\zeta=\frac{h^3}{120}\,,$$

grado di precisione 5. Posto

$$h=\frac{b-a}{N}\,, \quad z_k=a+kh, \quad k=0,\dots,N,$$

le formule composte sono

$$T_1^{(N)} = \frac{h}{2}\,\Big[f(a)+f(b)+2\sum_{k=1}^{N-1} f(z_k)\Big] + \frac{h^2}{12}\,\big[f'(a)-f'(b)\big];$$

$$\overline{T}_1^{(N)} = \frac{h}{30}\,\Big[7f(a)+7f(b)+14\sum_{k=1}^{N-1} f(z_k)+16\sum_{k=0}^{N-1} f\big(\frac{z_k+z_{k+1}}{2}\big)\Big] + \frac{h^2}{60}\,\big[f'(a)-f'(b)\big];$$

$$T_2^{(N)} = \frac{h}{2}\,\Big[f(a)+f(b)+2\sum_{k=1}^{N-1} f(z_k)\Big] + \frac{h^2}{10}\,\big[f'(a)-f'(b)\big] + \frac{h^3}{120}\,\Big[f''(a)+f''(b)+2\sum_{k=1}^{N-1} f''(z_k)\Big]$$

(si confronti $T_1^{(N)}$ con la formula di Eulero-Maclaurin dell'esercizio 4.42). Nei due casi particolari per $T_1^{(N)}$ al valore ottenuto con la formula dei trapezi risulta aggiunto $\dfrac{h^2}{12}$. Nel caso (1) si ottiene $T_1^{(4)} = 1.570833$ con un errore di circa $0.367\ 10^{-4}$. Per $T_2^{(N)}$ nel caso (2) si tenga conto che $f''(z_k) = -f(z_k)$.)

7.19 Si determinino i parametri α_i della formula

$$T_n = \sum_{i=0}^{n} \alpha_i \big[f^{(i)}(-1) + (-1)^i f^{(i)}(1) \big]$$

per l'approssimazione dell'integrale

$$S = \int_{-1}^{1} f(x)\ dx,$$

in modo che il grado di precisione sia il più elevato possibile. Si determini anche il resto. Si ricavino, come caso particolare, le formule a) e c) dell'esercizio 7.18.

(Traccia: si verifichi, integrando ripetutamente per parti, che per una funzione $g(x)$ derivabile $n+1$ volte è

$$\int_{-1}^{1} f(x) g^{(n+1)}(x)\ dx = \sum_{i=0}^{n} (-1)^i \Big[f^{(i)}(x) g^{(n-i)}(x) \Big]_{-1}^{1} + (-1)^{n+1} \int_{-1}^{1} f^{(n+1)}(x) g(x)\ dx.$$

Per $g(x) = P_{n+1}(x)$, l'$(n+1)$-esimo polinomio di Legendre, è

$$g^{(n+1)}(x) = (n+1)!\ a_{n+1} = \frac{(2n+2)!}{2^{n+1}(n+1)!}$$

e per l'esercizio 6.15 g) è

$$g^{(n-i)}(-1) = (-1)^{i-1}\ \frac{(2n-i+1)!}{2^{n-i}(i+1)!\,(n-i)!}\,,$$
$$g^{(n-i)}(1) = \frac{(2n-i+1)!}{2^{n-i}(i+1)!\,(n-i)!}\,,$$

da cui

$$\int_{-1}^{1} f(x)\ dx = \sum_{i=0}^{n} \alpha_i \big[f^{(i)}(-1) + (-1)^i f^{(i)}(1) \big] + r_n,$$

dove

$$\alpha_i = \frac{2^{n+1}(n+1)!}{(2n+2)!}\ \frac{(2n-i+1)!}{2^{n-i}(i+1)!\,(n-i)!} = \frac{2^{i+1}(2n-i+1)!}{(2n+2)!}\ \binom{n+1}{i+1},$$

$$r_n = \frac{(-1)^{n+1}2^{n+1}(n+1)!}{(2n+2)!}\ \int_{-1}^{1} f^{(n+1)}(x)P_{n+1}(x)\ dx.$$

Tenendo conto che

$$P_{n+1}(x) = \frac{1}{2^{n+1}(n+1)!}\ \frac{d^{n+1}}{dx^{n+1}}\ (x^2-1)^{n+1},$$

si ha

$$r_n = \frac{(-1)^{n+1}}{(2n+2)!}\ \int_{-1}^{1} f^{(n+1)}(x)\ \frac{d^{n+1}}{dx^{n+1}}\ (x^2-1)^{n+1}\ dx,$$

e integrando ripetutamente per parti si ha

$$r_n = \frac{1}{(2n+2)!}\ \int_{-1}^{1} f^{(2n+2)}(x)\ (x^2-1)^{n+1}\ dx,$$

e poiché $(x^2-1)^{n+1}$ ha segno costante per $x \in [-1,1]$, applicando il teorema della media integrale si ha

$$\begin{aligned} r_n &= \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\ \int_{-1}^{1} (x^2-1)^{n+1}\ dx \\ &= \frac{(-1)^{n+1}2^{2n+3}[(n+1)!]^2}{[(2n+2)!]^2(2n+3)}\ f^{(2n+2)}(\xi), \quad \xi \in (-1,1).\ ) \end{aligned}$$

7.20 Facendo riferimento al polinomio osculatore di Hermite, studiato nel paragrafo 4 del capitolo 5, e assegnati in $[a,b]$ $n+1$ nodi x_i, $i = 0,\dots,n$,

a) si costruiscano le formule di *quadratura di Hermite* della forma

$$H_{n+1} = \sum_{i=0}^{n} w_i f(x_i) + \sum_{i=0}^{n} z_i f'(x_i),$$

per l'approssimazione dell'integrale

$$\int_a^b f(x)\ dx$$

(le formule a) e b) ottenute nell'esercizio 7.18 ne sono un esempio);

b) si determini il corrispondente resto;

c) si dica come si riduce la formula quando $a = -1$, $b = 1$ e come nodi x_i si scelgono gli zeri dell'$(n+1)$-esimo polinomio di Legendre.

(Traccia: a) indicato con $q_{2n+1}(x)$ il polinomio osculatore di Hermite di grado $2n+1$ tale che

$$q_{2n+1}(x_i) = f(x_i), \qquad q'_{2n+1}(x_i) = f'(x_i), \qquad i = 0, \dots, n,$$

si pone

$$H_{n+1} = \int_a^b q_{2n+1}(x)\ dx.$$

Per la (17, cap. 5) è

$$w_i = \int_a^b \big[1 - 2L'_i(x_i)(x - x_i)\big] L_i^2(x)\ dx, \quad z_i = \int_a^b (x - x_i) L_i^2(x)\ dx,$$

dove $L_i(x)$ è l'i-esimo polinomio di Lagrange.
b) per il teorema 5.12 risulta

$$r_{n+1} = \int_a^b \frac{\pi_n^2(x)}{(2n+2)!}\ f^{(2n+2)}(\eta)\ dx, \quad \eta \in (a, b),$$

e per il teorema della media integrale, essendo $\pi_n^2(x) \geq 0$ in $[a, b]$, esiste $\xi \in (a, b)$ tale che

$$r_{n+1} = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \pi_n^2(x)\ dx.$$

c) per la (8, cap. 5) è

$$(x - x_i) L_i(x) = \frac{\pi_n(x)}{\pi'_n(x_i)}\,,$$

per cui

$$z_i = \frac{1}{\pi'_n(x_i)} \int_{-1}^1 \pi_n(x) L_i(x)\ dx = 0,$$

in quanto $\pi_n(x)$ è, a meno di un fattore, uguale al polinomio di Legendre $P_{n+1}(x)$ e quindi ortogonale su $[-1, 1]$ a $L_i(x)$, che è un polinomio di grado n. Inoltre

$$\begin{aligned} w_i &= \int_{-1}^1 L_i^2(x)\ dx - 2 z_i L'_i(x_i) = \int_{-1}^1 L_i^2(x)\ dx \\ &= \frac{1}{\pi'_n(x_i)} \int_{-1}^1 L_i(x)\ \frac{\pi_n(x)}{x - x_i}\ dx \\ &= \frac{1}{\pi'_n(x_i)} \left\{ \int_{-1}^1 \big[L_i(x) - 1\big]\ \frac{\pi_n(x)}{x - x_i}\ dx + \int_{-1}^1 \frac{\pi_n(x)}{x - x_i}\ dx \right\}. \end{aligned}$$

Il primo integrale è nullo perché il polinomio $L_i(x)-1$ è divisibile per $x-x_i$ (infatti $L_i(x_i)=1$) e $\pi_n(x)$ è ortogonale al polinomio $\big[L_i(x)-1\big]/(x-x_i)$ di grado $n-1$. Quindi

$$w_i = \frac{1}{\pi_n'(x_i)} \int_{-1}^{1} \frac{\pi_n(x)}{x-x_i}\, dx,$$

e per la (41) tali coefficienti coincidono con quelli della formula di Gauss-Legendre. La formula di quadratura di Hermite H_{n+1} si riduce quindi a quella di Gauss-Legendre quando i nodi x_i sono gli zeri di $P_{n+1}(x)$.)

7.21 Si costruisca la formula di quadratura a coefficienti uniformi del tipo

$$S_3 = \alpha\big[f(x_0)+f(x_1)+f(x_2)\big]$$

per l'approssimazione di

$$\int_{-1}^{1} f(x)\, dx.$$

Si dica qual è il grado di precisione della formula.

(Traccia: si determinino α, x_0, x_1 e x_2 in modo che la formula abbia il massimo grado di precisione. Con il metodo dei coefficienti indeterminati si ha

$$\begin{cases} 3\alpha = 2 \\ \alpha(x_0+x_1+x_2) = 0 \\ \alpha(x_0^2+x_1^2+x_2^2) = \dfrac{2}{3} \\ \alpha(x_0^3+x_1^3+x_2^3) = 0, \end{cases}$$

da cui si ricava

$$\alpha = \frac{2}{3}\,, \quad x_0 = -x_2 = \frac{1}{\sqrt{2}}\,, \quad x_1 = 0.$$

Per determinare il grado di precisione si applichi la formula alla funzione $f(x)=x^4$. Poiché

$$\int_{-1}^{1} x^4\, dx = \frac{2}{5}\,, \quad S_3 = \frac{1}{3}\,,$$

la formula ha grado di precisione 3.)

7.22 Sia a una costante assegnata. Si costruisca una formula di quadratura per il calcolo dell'integrale

$$\int_{-1}^{1} (a-x) f(x)\, dx$$

della forma

$$S_3 = \alpha_0 f(-x_0) + \alpha_1 f(0) + \alpha_2 f(x_0).$$

Si dica qual è il grado di precisione della formula. Si applichi al caso particolare

$$S = \int_{-1}^{1} \frac{1-x}{1+x^2}\, dx.$$

(Traccia: si determinino x_0, α_0, α_1 e α_2 in modo che la formula abbia il massimo grado di precisione, usando il metodo dei coefficienti indeterminati. Risulta

$$x_0 = \sqrt{\frac{3}{5}}\,, \quad \alpha_0 = \frac{5}{9}\,(a + \sqrt{\frac{3}{5}}), \quad \alpha_1 = \frac{8}{9}\,a, \quad \alpha_2 = \frac{5}{9}\,(a - \sqrt{\frac{3}{5}}).$$

Per determinare il grado di precisione si applichi la formula alle funzioni $f(x) = x^4$ e $f(x) = x^5$. Si verifichi che nel primo caso la formula fornisce il risultato corretto, mentre nel secondo si ha

$$\int_{-1}^{1} (a-x)x^5\, dx = -\frac{2}{7}\,, \quad S_3 = -\frac{6}{25}\,.$$

Quindi la formula ha grado di precisione 4. Nel caso particolare risulta $S_3 = \dfrac{19}{12}$, mentre $S = \dfrac{\pi}{2}$, per cui $|S - S_3| \approx 0.125\ 10^{-1}$.)

7.23 Si costruiscano le formule di quadratura per il calcolo dell'integrale

$$\int_0^1 \frac{f(x)}{\sqrt{x(1-x)}}\, dx$$

della forma

a) $$S_3 = \alpha_0 f(0) + \alpha_1 f(\frac{1}{2}) + \alpha_2 f(1),$$

b) $$S_3 = \alpha_0 f(x_0) + \alpha_1 f(\frac{1}{2}) + \alpha_2 f(1-x_0).$$

Si determinino anche i gradi di precisione. Si applichi al caso particolare

$$S = \int_0^1 \frac{\log(1+x)}{\sqrt{x(1-x)}}\, dx.$$

(Traccia: a) si determinino α_0, α_1 e α_2 in modo che la formula abbia il massimo grado di precisione, usando il metodo dei coefficienti indeterminati. Risulta

$$\alpha_0 = \alpha_2 = \frac{\pi}{4} \quad \alpha_1 = \frac{\pi}{2}\,.$$

Per determinare il grado di precisione si applichi la formula alle funzioni $f(x) = x^3$ e $f(x) = x^4$. Risulta che la formula ha grado di precisione 3.
b) Si proceda come in a), risulta che

$$x_0 = \frac{2-\sqrt{3}}{4}\,, \quad \alpha_0 = \alpha_1 = \alpha_2 = \frac{\pi}{3}\,,$$

e che il grado di precisione è 5.

Nel caso particolare con la prima formula si ha $S_3 = 1.181300$, e poiché $S = 1.182661$ è $|S - S_3| \approx 0.136\ 10^{-2}$; con la seconda si ha $S_3 = 1.182688$, ed è $|S - S_3| \approx 0.270\ 10^{-4}$.)

7.24 Si applichino le formule di Gauss-Chebyshev al calcolo degli integrali

$$(1) \quad \int_{-1}^{1} \frac{x^4}{\sqrt{1-x^2}}\,dx, \qquad (2) \quad \int_{-1}^{1} \frac{\sqrt{1+x^2}}{\sqrt{1-x^2}}\,dx.$$

Si dica come deve essere scelto n nei due casi affinché il resto risulti minore in modulo di 10^{-5}.

(Traccia: (1) per $n = 2$ l'integrale è calcolato esattamente e risulta

$$S_3 = \frac{\pi}{3}\,\big[f\big(-\frac{\sqrt{3}}{2}\big) + f(0) + f\big(\frac{\sqrt{3}}{2}\big)\big] = \frac{3}{8}\,\pi;$$

(2) posto $f(x) = \sqrt{1+x^2}$, le maggiorazioni dei resti dati nella (57) risultano maggiori di 10^{-5} per $n < 6$. Per $n = 6$ è

$$M_{14} = \max_{x\in[-1,1]} |f^{(14)}(x)| < 1.5\ 10^9 \quad \text{e} \quad |r_7| < \frac{\pi M_{14}}{2^{13}\ 14!} < 10^{-5}.$$

Posto $x_j = \cos\dfrac{(2j+1)\pi}{14}$, $j = 0, \ldots, 6$, risulta

$$S_7 = \frac{\pi}{7}\sum_{j=0}^{6} f(x_j) = 3.820197.$$

Poiché il valore esatto dell'integrale è 3.820198, l'errore assoluto risulta minore di 10^{-5}.)

7.25 Si verifichi che per $k,\ j = 1, \ldots, n$ è

$$\frac{2}{n+1}\ \sum_{i=0}^{n} \cos k\theta_i\ \cos j\theta_i = \delta_{kj}, \quad \text{dove} \quad \theta_i = \frac{(2i+1)\pi}{2(n+1)}\,, \quad i = 0, \ldots, n.$$

(Traccia: si applichi la formula S_{n+1} di Gauss-Chebyshev con $n+1$ nodi all'integrale

$$S = \int_{-1}^{1} \frac{T_k(x)T_j(x)}{\sqrt{1-x^2}}\, dx = \int_0^{\pi} \cos k\theta \, \cos j\theta = \frac{\pi}{2}\, \delta_{kj}.$$

Poiché il grado del polinomio $T_k(x)T_j(x)$ è minore o uguale a $2n$ e la formula ha grado di precisione $2n+1$, si ha $S_{n+1} = S$, quindi

$$S_{n+1} = \frac{\pi}{n+1} \sum_{i=0}^{n} T_k(x_i)T_j(x_i) = \frac{\pi}{2}\, \delta_{kj}.\ )$$

7.26 Per calcolare l'integrale

$$S = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx$$

a) si aplichi la formula di Gauss-Chebyshev con $n+1$ nodi;
b) si faccia il cambiamento di variabile $x = \cos\theta$, e poi si applichi la formula dei punti di mezzo, dividendo l'intervallo di integrazione in $n+1$ sottointervalli.

Si confrontino le due formule.

(Traccia: in entrambi i casi si ottiene

$$\frac{\pi}{n+1} \sum_{i=0}^{n} f(\cos\theta_i), \quad \text{dove } \theta_i = \frac{(2i+1)\pi}{2(n+1)}, \quad \text{per } i = 0, \ldots, n.\ )$$

7.27 Si applichino le formule di Gauss-Laguerre al calcolo degli integrali

$$(1) \quad \int_0^{\infty} \mathrm{e}^{-x}(x - \sin x)\, dx, \qquad (2) \quad \int_0^{\infty} \mathrm{e}^{-x} \cos^2 \frac{x}{4}\, dx.$$

Si dica come deve essere scelto n nei due casi affinché il resto risulti minore di 10^{-4}.

(Traccia: (1) risulta $|f^{(2n+2)}(x)| \le 1$, e dalla (58) segue che $n = 7$; (2) per $n \le 3$ risulta $|f^{(2n+2)}(x)| \le \dfrac{1}{2^{2n+3}}$, e dalla (58) segue che $n = 3$.)

7.28 Si applichino le formule di Gauss-Hermite al calcolo degli integrali

$$(1) \quad \int_{-\infty}^{\infty} x^4 \mathrm{e}^{-x^2}\, dx, \qquad (2) \quad \int_{-\infty}^{\infty} \mathrm{e}^{-x^2} \sin^2 x\, dx.$$

Si dica come deve essere scelto n nei vari casi affinché il resto risulti minore di 10^{-4}.

(Traccia: (1) per $n = 2$ l'integrale è calcolato esattamente; (2) risulta $|f^{(2n+2)}(x)| \le 2^{2n+1}$, e dalla (59) segue che $n = 5$.)

7.29 Si trasformino i seguenti integrali impropri in integrali propri

$$(1)\ \int_1^2 \frac{dx}{\sqrt{x(2-x)}}\,, \qquad (2)\ \int_0^\infty \frac{dx}{(1+x^2)^n}\,,$$
$$(3)\ \int_1^\infty \frac{1}{x^2}\,\sin\frac{1}{x^2}\,dx\,, \quad (4)\ \int_0^\infty (1+x^2)^{-4/3}\,dx.$$

(Traccia: si facciano i cambiamenti di variabile (1) $x = 2-y^2$, (2) $x = \tan y$, (3) $x = \dfrac{1}{y}$, (4) $x = \dfrac{1}{y^3} - 1$.)

7.30 Si applichi il metodo di sottrazione della singolarità a

$$(1)\quad \int_0^1 \frac{\sin x}{\sqrt{x^3}}\,dx, \qquad (2)\quad \int_0^1 \frac{\mathrm{e}^{-x}}{\sqrt{1-x}}\,dx.$$

(Traccia:

$$(1)\qquad \int_0^1 \frac{x}{\sqrt{x^3}}\,dx + \int_0^1 \frac{\sin x - x}{\sqrt{x^3}}\,dx = 2 + \int_0^1 \frac{\sin x - x}{\sqrt{x^3}}\,dx$$

$$(2)\quad \int_0^1 \frac{\mathrm{e}^{-1}}{\sqrt{1-x}}\,dx + \int_0^1 \frac{\mathrm{e}^{-x} - \mathrm{e}^{-1}}{\sqrt{1-x}}\,dx = \frac{2}{\mathrm{e}} + \int_0^1 \frac{\mathrm{e}^{-x} - \mathrm{e}^{-1}}{\sqrt{1-x}}\,dx.\ )$$

7.31 Si calcoli l'integrale improprio

$$S = \int_0^{\pi/2} \frac{\cos x \log(\sin x)}{\sin^2 x + 1}\,dx.$$

(Traccia: si faccia prima la trasformazione di variabile $y = \sin x$ e si integri per parti. Si ottiene

$$S = \int_0^1 \frac{\log y}{1+y^2}\,dy = -\int_0^1 \frac{1}{y}\,\arctan y\,dy.$$

L'ultimo integrale è proprio e può essere calcolato con una qualsiasi formula di quadratura. Ad esempio con la formula di Cavalieri-Simpson e

l'estrapolazione di Richardson si ottiene con $N = 8$ l'approssimazione -0.9159659, affetta da un errore assoluto di circa $0.535\ 10^{-6}$.)

7.32 Per la funzione

$$f(x) = \frac{1}{x}\ \sin\ \frac{1}{x}$$

a) si determini una successione di punti $\{\eta_k\}$ tale che

$$\lim_{k\to\infty} \eta_k = 0 \quad \text{e} \quad f(\eta_k) = \frac{1}{\eta_k}\,;$$

b) si verifichi che se si ignora la singolarità nel calcolo di

$$\int_0^{2/\pi} f(x)\ dx,$$

applicando la formula dei trapezi su N punti, si ottiene una successione non convergente per $N \to \infty$.

(Traccia: a) è $\eta_k = \dfrac{2}{\pi(1+4k)}$; b) per $N = 1 + 4k$ l'ampiezza dei sottointervalli è η_k, quindi il contributo a $J_2^{(N)}$ del primo sottointervallo è $\dfrac{1}{2}\,\eta_k f(\eta_k) = \dfrac{1}{2}$ per ogni k, mentre $\lim\limits_{k\to\infty} \displaystyle\int_0^{\eta_k} f(x)\ dx = 0$, perché la funzione $f(x)$ è integrabile.)

7.33 Si verifichi, applicando il procedimento di integrazione per serie, che

$$\int_0^1 \frac{\log x}{1-x}\ dx = -\,\frac{\pi^2}{6}\,, \tag{1}$$

$$\int_0^1 \frac{\log x}{1+x}\ dx = -\,\frac{\pi^2}{12}\,. \tag{2}$$

(Traccia: (1) è

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i, \quad \int_0^1 x^i \log x\ dx = -\,\frac{1}{(i+1)^2}\,,$$

e

$$\sum_{i=0}^{\infty} \frac{1}{(i+1)^2} = \frac{\pi^2}{6}$$

per la (40, cap. 4); (2) in modo analogo, tenendo conto dell'esercizio 4.34 b).)

7.34 Si applichi il procedimento di integrazione per serie al calcolo di

$$(1)\quad \int_0^1 \frac{\mathrm{e}^x}{\sqrt{x}}\,dx, \qquad (2)\quad \int_0^1 \frac{\sin x}{\sqrt{x^3}}\,dx, \qquad (3)\quad \int_0^1 \cos x \log x\,dx,$$
$$(4)\quad \int_0^1 \mathrm{e}^{-x}\log x\,dx, \qquad (5)\quad \int_0^\infty \frac{x-\sin x}{x^2}\,\mathrm{e}^{-x}\,dx.$$

Si dica quanti termini vanno sommati per ottenere un errore assoluto minore in modulo di 10^{-5}.

(Traccia:

$$(1)\qquad \int_0^1 \Big[\frac{1}{\sqrt{x}} + \sum_{k=0}^{\infty} \frac{x^{k+1/2}}{(k+1)!}\Big]\,dx = 2 + \sum_{k=0}^{\infty} \frac{2}{(k+1)!\,(2k+3)}\,,$$

si devono sommare i termini fino a $k = 6$;

$$(2)\qquad \int_0^1 \Big[\frac{1}{\sqrt{x}} + \sum_{k=1}^{\infty} \frac{(-1)^k x^{2k-1/2}}{(2k+1)!}\Big]\,dx = 2 + \sum_{k=1}^{\infty} \frac{2(-1)^k}{(2k+1)!\,(4k+1)}\,,$$

si devono sommare i termini fino a $k = 3$;

$$(3)\qquad \int_0^1 \Big[\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}\Big]\log x\,dx = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{(2k)!\,(2k+1)^2}\,,$$

si devono sommare i termini fino a $k = 3$;

$$(4)\qquad \int_0^1 \Big[\sum_{k=0}^{\infty} \frac{(-1)^k x^k}{k!}\Big]\log x\,dx = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!\,(k+1)^2}\,,$$

si devono sommare i termini fino a $k = 6$;

$$(5)\qquad \int_0^\infty \Big[\sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^{2k-1}}{(2k+1)!}\Big]\mathrm{e}^{-x}\,dx = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{2k\,(2k+1)}\,,$$

si devono sommare i termini fino a $k = 157$; si suggerisce di applicare la trasformazione di Eulero (par. 6 cap. 4).)

7.35 Si approssimino con le formule di Gauss-Chebyshev gli integrali

$$(1) \quad \int_0^1 \frac{\mathrm{e}^x}{\sqrt{x}}\, dx, \qquad (2) \quad \int_0^1 \frac{\mathrm{e}^{-x}}{\sqrt{1-x}}\, dx.$$

(Traccia: si faccia prima il cambiamento di variabile $x = \dfrac{y+1}{2}$. Si ottiene

$$(1) \qquad \frac{\sqrt{2}}{2} \int_{-1}^1 \frac{g(y)}{\sqrt{1-y^2}}\, dy, \quad \text{dove} \quad g(y) = \sqrt{1-y}\, \mathrm{e}^{(y+1)/2},$$

$$(2) \qquad \frac{\sqrt{2}}{2} \int_{-1}^1 \frac{g(y)}{\sqrt{1-y^2}}\, dy, \quad \text{dove} \quad g(y) = \sqrt{1+y}\, \mathrm{e}^{-(y+1)/2}.$$

Per entrambe le funzioni le derivate non sono limitate in $[-1, 1]$. Per $n = 32$ si ottiene in (1) il valore 2.925806, affetto da un errore assoluto di circa $0.503\ 10^{-3}$, in (2) il valore 1.076346, affetto da un errore assoluto di circa $0.187\ 10^{-3}$.)

7.36 Siano

$$f(x) = \frac{1}{\sqrt{1+x^2}}\, \sin\frac{\pi x}{4}\,, \quad g(x) = \frac{1}{x}\, \sin\frac{\pi x}{4}\,,$$

e $\epsilon = 10^{-4}$. Per calcolare

$$S = \int_0^\infty f(x)\, dx$$

si può procedere nei due modi seguenti:

a) troncando l'intervallo di integrazione e calcolando

$$\int_0^t f(x)\, dx,$$

dove t è scelto in modo che

$$\left| \int_t^\infty f(x)\, dx \right| < \epsilon;$$

b) ponendo

$$S = \int_0^\infty g(x)\, dx - \int_0^\infty \big[g(x) - f(x)\big]\, dx,$$

sfruttando il fatto che

$$\int_0^\infty g(x)\ dx = \frac{\pi}{2}\,,$$

e poi troncando l'intervallo di integrazione e calcolando

$$\int_0^t \big[g(x) - f(x)\big]\ dx,$$

dove t è scelto in modo tale che

$$\left| \int_t^\infty \big[g(x) - f(x)\big]\ dx \right| \ < \ \epsilon.$$

Si dica quanto deve valere t nei due casi e si calcoli S.

(Traccia: a) posto $t = 4k$, k intero, si ha

$$\int_t^\infty f(x)\ dx = \sum_{i=k}^{\infty} u_i, \quad \text{dove} \quad u_i = \int_{4i}^{4(i+1)} f(x)\ dx.$$

Poiché $u_i > 0$ per i pari e $u_i < 0$ per i dispari, e $|u_i| > |u_{i+1}|$ per ogni i, risulta per k pari

$$\left| \int_t^\infty f(x)\ dx \right| < u_k < \frac{1}{\sqrt{1+(4k)^2}} \int_0^4 \sin\frac{\pi x}{4}\ dx = \frac{8}{\pi\sqrt{1+(4k)^2}}\,.$$

Imponendo che tale quantità sia minore di 10^{-4} risulta $4k = 25470$.
b) È

$$g(x) - f(x) = \sin\frac{\pi x}{4}\ \frac{1}{x\sqrt{1+x^2}\big(x+\sqrt{1+x^2}\big)}\,,$$

e procedendo come nel caso precedente, per k pari si ha

$$\left| \int_t^\infty \big[g(x) - f(x)\big]\ dx \right| \ < \frac{8}{4k\pi\sqrt{1+(4k)^2}\big[4k+\sqrt{1+(4k)^2}\big]}\,,$$

da cui risulta $4k = 24$. Quindi conviene approssimare S nel secondo modo e si ottiene

$$\frac{\pi}{2} - \int_0^{24} \big[g(x) - f(x)\big]\ dx = 0.9820044;$$

l'errore assoluto risulta circa $0.453\ 10^{-4}$.)

7.37 Si approssimino con le formule di Gauss-Laguerre gli integrali

$$(1)\quad \int_0^\infty \frac{\mathrm{e}^{-x^2}-\mathrm{e}^{-x}}{x}\,dx, \qquad (2)\quad \int_0^\infty \frac{x^4}{\mathrm{e}^x+1}\,dx,$$
$$(3)\quad \int_0^\infty \frac{\sin(x/\pi)}{\mathrm{e}^x-1}\,dx.$$

(Traccia: (1) si ponga

$$f(x)=\frac{\mathrm{e}^{-x^2+x}-1}{x}\,,$$

per $n=5$ si ottiene il valore 0.2882506 affetto da un errore assoluto di circa $0.357\ 10^{-3}$; (2) si ponga

$$f(x)=\frac{x^4\mathrm{e}^x}{\mathrm{e}^x+1}\,,$$

per $n=6$ si ottiene il valore 23.33113 affetto da un errore assoluto di circa $0.256\ 10^{-3}$; (3) si ponga

$$f(x)=\frac{\mathrm{e}^x\sin(x/\pi)}{\mathrm{e}^x-1}\,,$$

per $n=4$ si ottiene il valore 0.4917089 affetto da un errore assoluto di circa $0.578\ 10^{-5}$.)

7.38 Si approssimino con le formule di Gauss-Hermite gli integrali

$$(1)\quad \int_{-\infty}^\infty \frac{x^4}{\mathrm{e}^{x^2}-1}\,dx, \qquad (2)\quad \int_{-\infty}^\infty 3^{-x^2}\,dx.$$

(Traccia: (1) si ponga

$$f(x)=\frac{x^4\mathrm{e}^{x^2}}{\mathrm{e}^{x^2}-1}\,,$$

per $n=6$ si ottiene il valore 1.783133 affetto da un errore assoluto di circa $0.160\ 10^{-3}$; (2) si ponga

$$f(x)=\left(\frac{\mathrm{e}}{3}\right)^{x^2},$$

per $n=6$ si ottiene il valore 1.691026 affetto da un errore assoluto di circa $0.944\ 10^{-5}$.)

7.39 Si verifichi che per $k>1$ è

$$\int_0^\infty \frac{dx}{1+x^k} \;=\; \int_0^1 \frac{dy}{\sqrt[k]{1-y^k}}\,.$$

Si dica se è più conveniente approssimare il primo integrale con le formule di Gauss-Laguerre oppure il secondo integrale con l'estrapolazione di Richardson applicata alla formula di Cavalieri o con le formule di Gauss-Chebyshev. Si confrontino i risultati ottenuti con il valore dell'integrale che è $\dfrac{\pi}{k \sin(\pi/k)}$.

(Traccia: si ponga $x^k = \dfrac{y^k}{1-y^k}$. Non è conveniente approssimare il primo integrale con le formule di Gauss-Laguerre, perché le derivate della funzione integranda presentano vicino a zero delle oscillazioni che aumentano di numero e di ampiezza al crescere dell'ordine. Ad esempio la derivata decima ha modulo massimo superiore a 10^8 per $k = 5$ e a 10^{11} per $k = 9$. Problemi analoghi si presentano con il secondo integrale. Ad esempio, per $k = 5$ si ottiene il valore 1.067966 (errore assoluto di circa $0.993\ 10^{-3}$) con l'estrapolazione di Richardson con 2^{11} valutazioni di funzione e il valore 1.0689253 (errore assoluto di circa $0.343\ 10^{-4}$) con le formule di Gauss-Chebyshev con 2^{10} valutazioni di funzione. Risultati migliori si ottengono per il primo integrale decomponendo l'intervallo, ad esempio nei due sottointervalli $[0, 2]$ e $[2, \infty]$, per il secondo integrale utilizzando una strategia adattiva (si veda l'esercizio 7.47).)

7.40 a) Si verifichi che l'insieme dei polinomi di grado i

$$s_i(x) = \frac{P_i(x) + P_{i+1}(x)}{x+1},$$

dove $P_i(x)$ è l'i-esimo polinomio di Legendre, è costituito da polinomi ortogonali sull'intervallo $[-1, 1]$ rispetto al peso $r(x) = x + 1$.

b) Indicati con $x_i,\ i = 0, \ldots, n$ gli zeri di $s_{n+1}(x)$, e posto

$$\pi_{n+1}(x) = (x+1)(x-x_0)\ldots(x-x_n),$$

è

$$a_{n+2}\pi_{n+1}(x) = (x+1)s_{n+1}(x),$$

dove a_{n+2} è il primo coefficiente di $P_{n+2}(x)$; si calcolino

$$\alpha_0 = \frac{1}{\pi'_{n+1}(-1)} \int_{-1}^{1} \frac{\pi_{n+1}(x)}{x+1}\, dx,$$

$$w_i = \frac{1}{\pi'_{n+1}(x_i)} \int_{-1}^{1} \frac{\pi_{n+1}(x)}{x-x_i}\, dx, \quad i = 0, \ldots, n.$$

La formula di quadratura i cui nodi sono -1 e $x_i,\ i = 0, \ldots, n$ e i cui coefficienti sono α_0 e $w_i,\ i = 0, \ldots, n$, è la formula di Radau S_{n+2} data in (64).

c) Si verifichi che se $f(x) \in C^{2n+3}[-1,1]$, il resto della formula di Radau è dato dalla (65).

d) Si calcolino le formule di Radau per $n = 0, 1$.

(Traccia: a) si ha

$$\int_{-1}^{1} r(x)s_i(x)s_j(x)\ dx = \int_{-1}^{1} \left[P_i(x) + P_{i+1}(x)\right]\ \frac{P_j(x) + P_{j+1}(x)}{x+1}\ dx.$$

Se $i > j$, il polinomio $\dfrac{P_j(x) + P_{j+1}(x)}{x+1}$ ha grado minore di i e quindi è ortogonale sia a $P_i(x)$ che a $P_{i+1}(x)$.
b) Si ha

$$a_{n+2}\pi_{n+1}(x) = P_{n+1}(x) + P_{n+2}(x),$$

e per l'esercizio 6.15 f) è

$$\begin{aligned} a_{n+2}\pi'_{n+1}(-1) &= (-1)^{n+2}\frac{(n+1)(n+2)}{2}\ + (-1)^{n+3}\frac{(n+2)(n+3)}{2} \\ &= (-1)^{n+1}(n+2); \end{aligned}$$

quindi per l'esercizio 6.15 k) è

$$\alpha_0 = \frac{(-1)^{n+1}}{n+2}\int_{-1}^{1} \frac{P_{n+1}(x) + P_{n+2}(x)}{x+1}\ dx = \frac{2}{(n+2)^2}\,.$$

Per l'esercizio 6.15 d), sfruttando la relazione a tre termini e il fatto che $P_{n+1}(x_i) = -P_{n+2}(x_i)$, si ha

$$a_{n+2}\pi'_{n+1}(x_i) = P'_{n+1}(x_i) + P'_{n+2}(x_i) = 2(n+2)\,\frac{P_{n+1}(x_i)}{1-x_i}.$$

Procedendo come nella dimostrazione del teorema 7.17, si verifichi che

$$\int_{-1}^{1} \frac{P_{n+1}(x) + P_{n+2}(x)}{x - x_i}\ dx = \frac{a_{n+2}h_{n+1}}{a_{n+1}P_{n+1}(x_i)} = \frac{2}{(n+2)P_{n+1}(x_i)}\,,$$

e quindi

$$w_i = \frac{1-x_i}{(n+2)^2 P_{n+1}^2(x_i)}\,.$$

c) Si proceda come per la dimostrazione del teorema 7.18. Risulta

$$r_{n+2} = \frac{f^{(2n+3)}(\xi)}{(2n+3)!}\ \int_{-1}^{1} \frac{\pi_{n+1}^2(x)}{x+1}\ dx, \quad \xi \in (-1,1).$$

Per il teorema 6.11 e per l'esercizio 6.15 i) risulta

$$\int_{-1}^{1} \frac{\pi_{n+1}^2(x)}{x+1}\,dx = \frac{1}{a_{n+2}^2}\int_{-1}^{1} \frac{\big[P_{n+1}(x)+P_{n+2}(x)\big]^2}{(x+1)}\,dx$$
$$= \frac{1}{a_{n+2}^2}\int_{-1}^{1} P_{n+1}(x)\frac{P_{n+1}(x)+P_{n+2}(x)}{(x+1)}\,dx = \frac{1}{a_{n+2}}\int_{-1}^{1} P_{n+1}(x)x^{n+1}\,dx$$
$$= \frac{1}{a_{n+2}}\,\frac{2^{n+2}\big[(n+1)!\big]^2}{(2n+3)!}\,.$$

d) Per $n=0$ si ha

$$s_1(x) = \frac{3}{2}\,x - \frac{1}{2}\,, \quad \text{da cui} \quad x_0 = \frac{1}{3}\,, \quad \alpha_0 = \frac{1}{2}\,, \quad w_0 = \frac{3}{2}\,,$$

e quindi

$$S_2 = \frac{1}{2}\,f(-1) + \frac{3}{2}\,f\big(\frac{1}{3}\big).$$

Per $n=1$ si ha

$$s_2(x) = \frac{5}{2}\,x^2 - x - \frac{1}{2}\,, \quad \text{da cui} \quad x_0 = \frac{1-\sqrt{6}}{5}, \quad x_1 = \frac{1+\sqrt{6}}{5}\,,$$

$$\alpha_0 = \frac{2}{9}\,, \quad w_0 = \frac{16+\sqrt{6}}{18}\,, \quad w_1 = \frac{16-\sqrt{6}}{18}$$

e quindi

$$S_3 = \frac{2}{9}\,f(-1) + \big[\frac{16+\sqrt{6}}{18}\,f\big(\frac{1-\sqrt{6}}{5}\big) + \frac{16-\sqrt{6}}{18}\,f\big(\frac{1+\sqrt{6}}{5}\big)\big].\ )$$

7.41 a) Si verifichi che l'insieme dei polinomi di grado i

$$s_i(x) = P'_{i+1}(x),$$

dove $P_i(x)$ è l'i-esimo polinomio di Legendre, è costituito da polinomi ortogonali sull'intervallo $[-1,1]$ rispetto al peso $r(x) = x^2 - 1$.

b) Indicati con $x_i,\ i = 0, \ldots, n$ gli zeri di $s_{n+1}(x)$, e posto

$$\pi_{n+2}(x) = (x^2-1)(x-x_0)\ldots(x-x_n),$$

è

$$(n+2)a_{n+2}\pi_{n+2}(x) = (x^2-1)s_{n+1}(x),$$

dove a_{n+2} è il primo coefficiente di $P_{n+2}(x)$; si calcolino

$$\alpha_0 = \frac{1}{\pi'_{n+2}(-1)} \int_{-1}^{1} \frac{\pi_{n+2}(x)}{x+1}\, dx, \quad \alpha_1 = \frac{1}{\pi'_{n+2}(1)} \int_{-1}^{1} \frac{\pi_{n+2}(x)}{x-1}\, dx,$$
$$w_i = \frac{1}{\pi'_{n+2}(x_i)} \int_{-1}^{1} \frac{\pi_{n+2}(x)}{x-x_i}\, dx, \quad i = 0, \ldots, n.$$

La formula di quadratura i cui nodi sono ± 1 e x_i, $i = 0, \ldots, n$ e i cui coefficienti sono α_0, α_1 e w_i, $i = 0, \ldots, n$, è la formula di Lobatto S_{n+3} data in (66).

c) Si verifichi che se $f(x) \in C^{2n+4}[-1, 1]$, il resto della formula di Lobatto è dato dalla (67).

d) Si calcolino le formule di Lobatto per $n = 0, 1$.

(Traccia: a) per l'esercizio 6.15 l) è

$$\int_{-1}^{1} r(x) s_i(x) s_j(x)\, dx = \int_{-1}^{1} (x^2-1) P'_{i+1}(x) P'_{j+1}(x)\, dx = 0, \quad i \neq j.$$

b) Si ha

$$(n+2) a_{n+2} \pi_{n+2}(x) = (x^2-1) P'_{n+2}(x),$$
$$(n+2) a_{n+2} \pi'_{n+2}(x) = 2x P'_{n+2}(x) + (x^2-1) P''_{n+2}(x),$$

e per l'esercizio 6.15 f) è

$$(n+2) a_{n+2} \pi'_{n+2}(-1) = -2P'_{n+2}(-1) = (-1)^n (n+2)(n+3);$$

quindi

$$\alpha_0 = \frac{(-1)^n}{(n+2)(n+3)} \int_{-1}^{1} \frac{(x^2-1) P'_{n+2}(x)}{x+1}\, dx.$$

Poiché

$$\int_{-1}^{1} \frac{(x^2-1) P'_{n+2}(x)}{x+1}\, dx = \int_{-1}^{1} (x-1) P'_{n+2}(x)\, dx$$
$$= \Big[(x-1) P_{n+2}(x) \Big]_{-1}^{1} - \int_{-1}^{1} P_{n+2}(x)\, dx = 2P_{n+2}(-1) = (-1)^n 2,$$

risulta

$$\alpha_0 = \frac{2}{(n+2)(n+3)};$$

e analogamente

$$\alpha_1 = \frac{2}{(n+2)(n+3)}\,.$$

Per l'esercizio 6.15 m) è

$$(n+2)a_{n+2}\pi'_{n+2}(x_i) = (x_i^2-1)P''_{n+2}(x_i) = (n+2)(n+3)P_{n+2}(x_i),$$

e per la 6.15 d) si ha

$$\begin{aligned}\frac{(x^2-1)P'_{n+2}(x)}{x-x_i} &= (n+2)\,\frac{xP_{n+2}(x)-P_{n+1}(x)}{x-x_i}\\ &= (n+2)\left[P_{n+2}(x)+\,\frac{x_iP_{n+2}(x)-P_{n+1}(x)}{x-x_i}\right].\end{aligned}$$

Procedendo come nella dimostrazione del teorema 7.17 e tenendo conto che $P'_{n+2}(x_i)=0$, e quindi per la 6.15 d) $x_iP_{n+2}(x_i) = P_{n+1}(x_i)$, si verifichi che

$$\int_{-1}^{1}\frac{x_iP_{n+2}(x)-P_{n+1}(x)}{x-x_i}\,dx = \frac{a_{n+2}h_{n+1}}{a_{n+1}P_{n+2}(x_i)} = \frac{2}{(n+2)P_{n+2}(x_i)}\,,$$

e quindi

$$\begin{aligned}w_i &= \frac{1}{(n+2)(n+3)P_{n+2}(x_i)}\int_{-1}^{1}(x^2-1)\,\frac{P'_{n+2}(x)}{x-x_i}\,dx\\ &= \frac{2}{(n+2)(n+3)P^2_{n+2}(x_i)}\,.\end{aligned}$$

c) Si proceda come per la dimostrazione del teorema 7.18. Risulta

$$r_{n+3} = \frac{f^{(2n+4)}(\xi)}{(2n+4)!}\int_{-1}^{1}\frac{\pi^2_{n+2}(x)}{x^2-1}\,dx,\quad \xi\in(-1,1).$$

Per l'esercizio 6.15 l) risulta

$$\begin{aligned}\int_{-1}^{1}\frac{\pi^2_{n+2}(x)}{x^2-1}\,dx &= \frac{1}{(n+2)^2a^2_{n+2}}\int_{-1}^{1}(x^2-1)[P'_{n+2}(x)]^2\,dx\\ &= -\,\frac{2(n+3)}{(n+2)a^2_{n+2}(2n+5)}\,.\end{aligned}$$

d) per $n=0$ si ha

$$s_1(x) = 3x,\quad \text{da cui}\quad x_0=0,\quad \alpha_0=\alpha_1=\frac{1}{3}\,,\quad w_0=\frac{4}{3}\,,$$

e quindi

$$S_3 = \frac{1}{3}\left[f(-1) + 4f(0) + f(1)\right]$$

(si noti che questa formula coincide con la formula di Newton-Cotes dei tre punti). Per $n = 1$ si ha

$$s_2(x) = \frac{15}{2}\,x^2 - \frac{3}{2}\,, \quad \text{da cui} \quad x_0 = \frac{-\sqrt{5}}{5}\,, \quad x_1 = \frac{\sqrt{5}}{5}\,,$$

$$\alpha_0 = \alpha_1 = \frac{1}{6}\,, \quad w_0 = w_1 = \frac{5}{6}$$

e quindi

$$S_3 = \frac{1}{6}\left[f(-1) + 5f\big(\frac{-\sqrt{5}}{5}\big) + 5f\big(\frac{\sqrt{5}}{5}\big) + f(1)\right].\ )$$

7.42 Sia $f(x) \in C^{2n+2}[-1,1]$, tale che $f^{(2n+2)}(x)$ non cambi segno in $[-1,1]$, e sia

$$S = \int_{-1}^{1} f(x)\ dx.$$

Indicati con G_{n+1} il valore ottenuto applicando la formula di Gauss-Legendre con $n+1$ nodi e con L_{n+2} il valore ottenuto applicando la formula di Lobatto con n nodi liberi (più i nodi ± 1), si verifichi che il valore di S è compreso fra G_{n+1} e L_{n+2} e che se $f^{(2n+2)}(x)$ varia di poco al variare della x, si può estrapolare il valore

$$\frac{(n+2)G_{n+1} + (n+1)L_{n+2}}{2n+3}$$

come migliore approssimazione di S.

(Traccia: si confrontino le formule dei resti (48) per G_{n+1} e (67) per L_{n+2}; in quest'ultima si sostituisca $n-1$ al posto di n.)

7.43 Siano $T_j^{(k)}$, $k, j = 0, 1, \ldots$, gli elementi della tabella di Romberg ottenuta con la successione $N_k = 2^k$. Si verifichi che

a) i valori $T_1^{(k)}$ della seconda colonna sono le approssimazioni dell'integrale ottenute con la formula di Cavalieri-Simpson su N_k nodi, cioè

$$T_1^{(k)} = J_3^{(N)}, \quad N = N_k;$$

b) i valori $T_2^{(k)}$ della terza colonna sono le approssimazioni dell'integrale ottenute con la formula composta corrispondente alla formula di Newton-Cotes S_5 su N_k nodi, cioè

$$T_2^{(k)} = J_5^{(N)}, \quad N = N_k;$$

c) i valori $T_3^{(k)}$ della quarta colonna *non* sono le approssimazioni dell'integrale ottenute con le formule composte corrispondenti alle formule di Newton-Cotes S_7 e S_9 su N_k nodi.

(Traccia: a) posto per semplicità $a = 0$, $b = 1$, risulta per $h = \dfrac{1}{N} = \dfrac{1}{N_k}$

$$T_1^{(k)} = \frac{4T_0^{(k+1)} - T_0^{(k)}}{3} = \frac{4J_2^{(2N)} - J_2^{(N)}}{3}$$
$$= \frac{4}{3}\,\frac{1}{4N}\Big[f(0) + 2\sum_{j=1}^{2N-1} f\big(\frac{j}{2N}\big) + f(1)\Big]$$
$$- \frac{1}{3}\,\frac{1}{2N}\Big[f(0) + 2\sum_{j=1}^{N-1} f\big(\frac{j}{N}\big) + f(1)\Big].$$

Si tenga poi conto del fatto che

$$\sum_{j=1}^{2N-1} f\big(\frac{j}{2N}\big) = \sum_{\substack{j=2\\ j \text{ pari}}}^{2N-2} f\big(\frac{j}{2N}\big) + \sum_{\substack{j=1\\ j \text{ dispari}}}^{2N-1} f\big(\frac{j}{2N}\big)$$
$$= \sum_{j=1}^{N-1} f\big(\frac{j}{N}\big) + \sum_{j=0}^{N-1} f\big[\frac{1}{2}\big(\frac{j}{N} + \frac{j+1}{N}\big)\big].$$

Per b) e c) si proceda in modo analogo.)

7.44 Si applichi il metodo di Romberg per calcolare un'approssimazione di

$$\frac{\pi}{4} = \int_0^1 \frac{dx}{1+x^2}$$

con un errore relativo minore di 10^{-5}. Si dica quante valutazioni della funzione $f(x) = \dfrac{1}{1+x^2}$ sono richieste, e si confronti tale numero con il numero di valutazioni che sarebbero richieste per ottenere la stessa precisione con la formula dei trapezi. Si confronti anche con il numero di termini richiesti per ottenere la stessa precisione dalla formula di Taylor con la trasformazione di Eulero (esempio 4.8).

(Traccia: la prima colonna e la diagonale della tabella di Romberg sono

k	$T_0^{(k)}$	$T_k^{(0)}$
0	0.7500000	0.7500000
1	0.7749996	0.7833328
2	0.7827938	0.7855290
3	0.7847468	0.7853955
4	0.7852349	0.7853974

Il valore 0.7853974 così ottenuto è affetto da un errore assoluto di circa $0.763\ 10^{-6}$ ed ha richiesto 17 valutazioni di $f(x)$. Con la formula dei trapezi si sarebbe ottenuto il valore 0.7853892 affetto da un errore assoluto di circa $0.896\ 10^{-5}$ con 257 valutazioni di $f(x)$.)

7.45 Siano $T_j^{(k)}$, $k, j = 0, 1, \ldots$, gli elementi della tabella di Romberg ottenuta con la successione $N_k = 2^k$. Si verifichi che per i $\delta_{j,i}$ definiti nella (69) è

$$\delta_{0,i} = 1 \quad \text{e} \quad \delta_{j,i} = \prod_{r=1}^{j} \frac{4^{r-i} - 1}{4^r - 1} \quad \text{per } 0 < j \leq i \leq m,$$

per cui $\delta_{j,j} = 0$, e che

$$\delta_{j,j+1} = \frac{(-1)^j}{4^{j(j+1)/2}}, \quad \text{per} \quad j \geq 1.$$

(Traccia: si proceda per induzione su j, notando che

$$\delta_{j,i} = \frac{4^{j-i} - 1}{4^j - 1}\ \delta_{j-1,i}.$$

Inoltre si verifichi che

$$\delta_{j,j+1} = -\frac{1}{4^j}\ \delta_{j-1,j}.\ )$$

7.46 Per calcolare l'integrale

$$S = \int_{-1}^{1} f(x)\ dx$$

si può applicare la formula dei trapezi dopo aver fatto il cambiamento di variabile $x = \cos\theta$

$$S = \int_0^{\pi} f(\cos\theta)\sin\theta\ d\theta = \int_0^{\pi} g(\theta)\ d\theta.$$

In tal modo la funzione $f(x)$ risulta valutata, anziché in punti equidistanti, nei nodi $x_i = \cos\theta_i$, $\theta_i = \dfrac{i\pi}{N}$, per $i = 0, \ldots, N$.

a) Si verifichi che se $f(x) \in C^{2m+1}[-1, 1]$, risulta

$$J_2^{(N)} - S = -\frac{h^2}{12}\ \big[f(-1) + f(1)\big] + \sum_{j=2}^{m-1} c_{2j} h^{2j} + O(h^{2m+1}),$$

dove $h = \dfrac{\pi}{N}$.

b) Posto

$$T_1^{(k)} = J_2^{(N)} + \frac{h^2}{12}\,\big[f(-1) + f(1)\big], \quad N = N_k = 2^k,$$

si costruisca la tabella di Romberg, applicando la (68) per $j = 2, 3, \ldots$ Si analizzi per quale classe di funzioni questo metodo presenta un vantaggio rispetto all'applicazione alla funzione $f(x)$ del metodo di Romberg.

(Traccia: a) dalla formula di Eulero-Maclaurin (si veda l'esercizio 4.42) risulta

$$J_2^{(N)} - S = \sum_{j=1}^{m-1} c_{2j} h^{2j} + O(h^{2m+1}),$$

in cui

$$c_2 = \frac{1}{12}\,\big[g'(\pi) - g'(0)\big] = -\frac{1}{12}\,\big[f(-1) + f(1)\big].$$

b) Si verifichi che le derivate di ordine $i+1$ dispari della funzione $g(\theta)$ nei punti 0 e π possono essere espresse come combinazioni lineari delle derivate di ordine minore o uguale a $i/2$ della funzione $f(x)$ nei punti 1 e -1. Perciò questo metodo è vantaggioso nel caso di funzioni le cui derivate agli estremi crescono molto all'aumentare dell'ordine.)

7.47 Si calcoli con un errore minore di 10^{-4} l'integrale

$$\int_0^1 \frac{dy}{\sqrt[5]{1 - y^5}}$$

con la formula di Cavalieri-Simpson, utilizzando una strategia adattiva. Si dica in quanti punti viene valutata la funzione integranda. Si confronti tale numero con il numero di valutazioni riportate nella traccia dell'esercizio 7.39.

(Traccia: usando la formula di Cavalieri-Simpson con l'estrapolazione di Richardson si ottiene il valore 1.068868, affetto da un errore minore in modulo di 10^{-4} con 105 valutazioni della funzione.)

7.48 a) Per $n \geq 2$ pari, si costruiscano le formule di *Clenshaw-Curtis*, che sono formule interpolatorie del tipo

$$S_{n+1} = \sum_{i=0}^{n} w_i f(x_i),$$

in cui i nodi sono

$$x_i = \cos\theta_i, \quad \theta_i = \frac{i\pi}{n}, \quad i = 0, \dots, n.$$

b) Si verifichi che i coefficienti delle formule di Clenshaw-Curtis sono positivi.

c) Si verifichi come le stesse formule possono essere ottenute integrando il polinomio trigonometrico di interpolazione di soli coseni della funzione $f(\cos\theta)$ nei nodi θ_i.

(Traccia: si noti che i θ_i sono gli zeri del polinomio di Chebyshev di 2^a specie $U_{n-1}(x)$, oltre ai punti ± 1, per cui $\pi_n(x) = 2^{1-n}(x^2-1)U_{n-1}(x)$. Dalla (5) si ha, per la simmetria dei nodi,

$$w_i = w_{n-i} = \frac{1}{2x_iU_{n-1}(x_i) + (x_i^2-1)U'_{n-1}(x_i)} \int_{-1}^{1} \frac{(x^2-1)U_{n-1}(x)}{x - x_i}\, dx.$$

Per $i = 0$ si ha

$$w_0 = \frac{1}{2U_{n-1}(1)} \int_{-1}^{1} (x+1)U_{n-1}(x)\, dx.$$

Calcolando l'integrale per parti e applicando le l) e n) dell'esercizio 6.17, si ha per n pari

$$w_0 = w_n = \frac{1}{n^2-1}.$$

Per $i = 1, \dots, \dfrac{n}{2}$ si ha

$$w_i = \frac{1}{(x_i^2-1)U'_{n-1}(x_i)} \int_{-1}^{1} \frac{(x^2-1)U_{n-1}(x)}{x - x_i}\, dx.$$

Per l'esercizio 6.17 h) è

$$\frac{(x^2-1)U_{n-1}(x)}{x-x_i} = \frac{T_{n+1}(x) - xT_n(x)}{x-x_i} = -T_n(x) + \frac{T_{n+1}(x) - x_iT_n(x)}{x-x_i}.$$

Per lo stesso esercizio, poiché $U_{n-1}(x_i) = 0$ per $i = 1, \dots, \dfrac{n}{2}$, risulta

$$x_iT_n(x_i) = T_{n+1}(x_i),$$

e per la formula di Christoffel-Darboux è

$$\frac{T_{n+1}(x) - x_iT_n(x)}{x-x_i} = \frac{1}{T_n(x_i)} \Big[1 + 2\sum_{j=1}^{n} T_j(x_i)T_j(x)\Big],$$

da cui per l'esercizio 6.17 m) per n pari è

$$\begin{aligned}
&\int_{-1}^{1} \frac{(x^2-1)U_{n-1}(x)}{x-x_i}\,dx \\
&\quad = -\int_{-1}^{1} T_n(x)\,dx + \frac{2}{T_n(x_i)}\Big[1+\sum_{j=1}^{n} T_j(x_i)\int_{-1}^{1} T_j(x)\,dx\Big] \\
&\quad = \frac{2}{n^2-1} + \frac{2}{T_n(x_i)}\Big[1-2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n} \frac{T_j(x_i)}{j^2-1}\Big] \\
&\quad = \frac{2}{T_n(x_i)}\Big[1-2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n-2} \frac{T_j(x_i)}{j^2-1} - \frac{T_n(x_i)}{n^2-1}\Big].
\end{aligned}$$

Inoltre

$$(x_i^2-1)U'_{n-1}(x_i) = n\cos i\pi,$$

e quindi

$$w_i = \frac{2}{n}\Big[1+2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n-2} \frac{\cos j\theta_i}{1-j^2} + \frac{(-1)^i}{1-n^2}\Big], \quad i=1,\dots,\frac{n}{2}\,.$$

b) Per $i=1,\dots,\dfrac{n}{2}$ si ha

$$w_i = \frac{2}{n}\big[1-\rho_i\big],$$

dove per l'esercizio 4.16 p) è

$$\begin{aligned}
|\rho_i| &= \Big|2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n-2} \frac{\cos j\theta_i}{j^2-1} + \frac{(-1)^i}{n^2-1}\Big| \\
&< 2\sum_{\substack{j=2\\ j\ \text{pari}}}^{n} \frac{1}{j^2-1} = 2\sum_{j=1}^{n/2} \frac{1}{4j^2-1} = \frac{n}{n+1} < 1.
\end{aligned}$$

Quindi $w_i > 0$ per $i=1,\dots,\dfrac{n}{2}$.

c) Usando per semplicità la notazione compatta

$$\sum_{i=0}^{n}{}'' y_i = \frac{y_0}{2} + \sum_{i=1}^{n-1} y_i + \frac{y_n}{2}\,,$$

per le (79) e (80) del capitolo 5, il polinomio trigonometrico è

$$F(\theta) = \sum_{j=0}^{n}{}'' \alpha_j \cos j\theta, \quad \text{dove} \quad \alpha_j = \frac{2}{n} \sum_{i=0}^{n}{}'' f(\cos\theta_i) \cos j\theta_i.$$

Poiché

$$\int_{-1}^{1} f(x)\ dx = \int_{0}^{\pi} f(\cos\theta) \sin\theta\ d\theta,$$

si pone

$$S_{n+1} = \int_0^\pi F(\theta) \sin\theta\ d\theta = \sum_{j=0}^{n}{}'' \alpha_j \int_0^\pi \cos j\theta \sin\theta\ d\theta,$$

ed essendo

$$\int_0^\pi \cos j\theta \sin\theta\ d\theta = \begin{cases} 0 & \text{per } j \text{ dispari,} \\ \dfrac{2}{1-j^2} & \text{per } j \text{ pari,} \end{cases}$$

risulta

$$S_{n+1} = 2 \sum_{\substack{j=0 \\ j \text{ pari}}}^{n}{}'' \frac{\alpha_j}{1-j^2},$$

in cui i coefficienti α_j possono essere calcolati con una trasformata di coseni. Sostituendo si ha

$$\begin{aligned} S_{n+1} &= \frac{4}{n} \sum_{\substack{j=0 \\ j \text{ pari}}}^{n}{}'' \frac{1}{1-j^2} \sum_{i=0}^{n}{}'' f(\cos\theta_i) \cos j\theta_i \\ &= \frac{4}{n} \sum_{i=0}^{n}{}'' f(\cos\theta_i) \sum_{\substack{j=0 \\ j \text{ pari}}}^{n}{}'' \frac{\cos j\theta_i}{1-j^2} = \sum_{i=0}^{n} \beta_i f(\cos\theta_i), \end{aligned}$$

in cui

$$\begin{aligned} \beta_0 &= \frac{2}{n}\Big[\frac{1}{2} + \frac{1}{2(1-n^2)} + \sum_{\substack{j=2 \\ j \text{ pari}}}^{n-2} \frac{1}{1-j^2}\Big] \\ &= \frac{1}{n}\Big[1 + \frac{1}{1-n^2} - 2 \sum_{j=1}^{n/2-1} \frac{1}{4j^2-1}\Big]. \end{aligned}$$

Per l'esercizio 4.16 p) è

$$-2 \sum_{j=1}^{n/2-1} \frac{1}{4j^2-1} = \frac{2-n}{n-1},$$

e quindi

$$\beta_0 = \frac{1}{n^2-1} = w_0,$$

e in modo analogo $\beta_n = w_n$. Per $i = 1, \ldots, n-1$ è

$$\beta_i = \frac{4}{n} \sum_{\substack{j=0 \\ j \text{ pari}}}^{n}{}'' \frac{\cos j\theta_i}{1-j^2} = w_i. \ )$$

7.49 Si scriva la formula prodotto dei trapezi $J_2^{(M)} \times J_2^{(N)}$ per il calcolo dell'integrale

$$\int_a^b \int_c^d f(x,y)\, dx\, dy;$$

in particolare si approssimi

$$\int_0^1 \int_0^1 \frac{dx\, dy}{1+x+y}$$

scegliendo $M = N$, per diversi valori di M, e si confronti con il risultato esatto. Si dica per quale valore di M l'errore assoluto risulta in modulo minore di $0.5\ 10^{-3}$.

(Traccia: posto

$$h = \frac{b-a}{M}, \quad k = \frac{d-c}{N}, \quad x_i = a + ih, \quad y_j = c + jk,$$

$$\sigma(y) = f(x_0, y) + 2 \sum_{i=1}^{M-1} f(x_i, y) + f(x_M, y),$$

è

$$J_2^{(M)} \times J_2^{(N)} = \frac{hk}{4} \Big[\sigma(y_0) + 2 \sum_{j=1}^{N-1} \sigma(y_j) + \sigma(y_N) \Big].$$

Nel caso particolare risulta $M = 512$.)

7.50 Si determini il resto (70) della formula $S_3 \times S_3$, se $f(x,y) \in C^4(D)$, dove $D = (a,b) \times (c,d)$.

(Traccia: per la formula di quadratura dei tre punti applicata alla funzione $f(x,y)$ di x, con y prefissato, è

$$g(y) = \int_a^b f(x,y)\,dx = \frac{h}{3}\big[f(x_0,y)+4f(x_1,y)+f(x_2,y)\big] - \frac{h^5}{90}\,\frac{\partial^4 f}{\partial x^4}\,(\xi(y),y),$$

dove

$$h = \frac{b-a}{2}\,, \quad x_i = a+ih, \quad \xi(y) \in (a,b).$$

Si ha poi

$$S = \int_c^d g(y)\,dy = \frac{k}{3}\big[g(y_0)+4g(y_1)+g(y_2)\big] - \frac{k^5}{90}\,g^{(4)}(\eta_2),$$

dove

$$k = \frac{d-c}{2}\,, \quad y_i = c+ik, \quad \eta_2 \in (c,d).$$

Quindi

$$r = S - S_3 \times S_3 = -\frac{1}{90}\Big\{ + \frac{kh^5}{3}\Big[\frac{\partial^4 f}{\partial x^4}\,(\xi(y_0),y_0) + 4\,\frac{\partial^4 f}{\partial x^4}\,(\xi(y_1),y_1) + \frac{\partial^4 f}{\partial x^4}\,(\xi(y_2),y_2)\Big] + k^5 g^{(4)}(\eta_2)\Big\}.$$

Per l'esercizio 7.1 esiste un punto $\eta_1 \in (c,d)$ tale che

$$\frac{\partial^4 f}{\partial x^4}\,(\xi(y_0),y_0) + 4\,\frac{\partial^4 f}{\partial x^4}\,(\xi(y_1),y_1) + \frac{\partial^4 f}{\partial x^4}\,(\xi(y_2),y_2) = 6\,\frac{\partial^4 f}{\partial x^4}\,(\xi(\eta_1),\eta_1).$$

Inoltre per ogni y esiste un punto $\xi_2(y) \in (a,b)$ tale che

$$g^{(4)}(y) = \int_a^b \frac{\partial^4 f}{\partial y^4}\,(x,y)\,dx = (b-a)\,\frac{\partial^4 f}{\partial y^4}\,(\xi_2(y),y),$$

per cui

$$r = -\frac{1}{90}\Big[2kh^5\frac{\partial^4 f}{\partial x^4}\,(\xi_1(\eta_1),\eta_1) + k^5(b-a)\frac{\partial^4 f}{\partial y^4}\,(\xi_2(\eta_2),\eta_2)\Big].$$

Ponendo $\xi_1 = \xi_1(\eta_1)$, $\xi_2 = \xi_2(\eta_2)$, ne segue la (70).)

7.51 Si scriva la formula di quadratura $S_3 \times S_3$ per il calcolo di

$$\int_a^b \int_{c(x)}^{d(x)} f(x,y)\ dy\ dx.$$

(Traccia: si ponga

$$h = \frac{b-a}{2}, \quad x_i = a + ih, \quad i = 0,1,2,$$
$$k_i = \frac{d(x_i) - c(x_i)}{2}, \quad y_{i,j} = c(x_i) + jk_i, \quad j = 0,1,2,$$

risulta

$$\begin{aligned} S_3 \times S_3 = \frac{h}{9}\Big\{ & k_0\big[f(x_0,y_{0,0}) + 4f(x_0,y_{0,1}) + f(x_0,y_{0,2})\big] \\ & + 4k_1\big[f(x_1,y_{1,0}) + 4f(x_1,y_{1,1}) + f(x_1,y_{1,2})\big] \\ & + k_2\big[f(x_2,y_{2,0}) + 4f(x_2,y_{2,1}) + f(x_2,y_{2,2})\big]\Big\}. \ ) \end{aligned}$$

7.52 Si dimostri che non esistono successioni $\{x_i\}$ di nodi distinti e $\{w_i\}$ di pesi per cui si abbia

$$\lim_{n\to\infty} \sum_{i=0}^{n} w_i f(x_i) = \int_a^b f(x)\ dx, \tag{87}$$

per ogni $f(x) \in C[a,b]$.

(Traccia: si supponga per assurdo che esistano $\{x_i\}$ e $\{w_i\}$ par cui vale la (87). Allora la successione $\{x_i\}$ sarebbe densa in $[a,b]$. Infatti se esistesse un intervallo $[c,d] \subset [a,b]$ in cui non vi fossero punti x_i, si potrebbe costruire una funzione $f(x)$ continua, nulla su $[a,b] - [c,d]$, e tale che

$$\int_c^d f(x)\ dx \neq 0.$$

Per tale funzione la formula di quadratura darebbe valore nullo per ogni n e quindi non potrebbe esservi convergenza. Inoltre per il teorema della limitatezza uniforme di Banach-Steinhaus (si veda [19]) esisterebbe una costante M tale che $\sum_{i=0}^{n} |w_i| \leq M$ per ogni n, quindi la serie $\sum_{i=0}^{n} w_i$ sarebbe uniformemente convergente. Per ogni ϵ si scelgano un nodo x_k per cui $w_k \neq 0$ e un η, con $0 < \eta \leq \epsilon$, tali che l'intorno $I(\eta)$ di x_k di raggio η stia tutto in $[a,b]$

e che i nodi che cadono in $I(\eta)$ abbiano tutti indice $i \geq \overline{k} > k$, e si consideri la funzione continua

$$f_\epsilon(x) = \begin{cases} 0 & \text{per } x \notin I(\eta), \\ 1 - \dfrac{|x - x_k|}{\eta} & \text{per } x \in I(\eta). \end{cases}$$

Allora è

$$\int_a^b f_\epsilon(x)\, dx = \eta \quad \text{e} \quad \sum_{i=0}^{\infty} w_i f_\epsilon(x_i) = w_k + \sum_{i=\overline{k}}^{\infty} w_i f_\epsilon(x_i).$$

Poiché si è supposto che vi sia convergenza, risulta

$$\sum_{i=0}^{\infty} w_i f_\epsilon(x_i) = \eta,$$

e quindi

$$|\eta - w_k| = \left| \sum_{i=\overline{k}}^{\infty} w_i f_\epsilon(x_i) \right| \leq \sum_{i=\overline{k}}^{\infty} |w_i|.$$

Per $\epsilon \to 0$ risulta $\eta \to 0$, e quindi il primo membro tende a $|w_k| \neq 0$, mentre $\overline{k} \to \infty$, e quindi l'ultimo membro tende a 0, il che è assurdo.)

7.53 Sia $r > 1$ intero; assegnato un intero $x_0 \in [0, 2^r - 1]$, si consideri la successione

$$x_{i+1} \equiv [\alpha x_i + \beta] \ (\text{mod} 2^r), \quad 0 < \alpha, \beta < 2^r,$$

e si definisca *periodo* p della successione il più piccolo intero n per cui esiste x_i tale che $x_{i+n} = x_i$.

a) Si verifichi che se β è dispari, allora p è indipendente dalla scelta di x_0, e che se $p > 1$ è sempre $x_p = x_0$, mentre se $p = 1$ può essere $x_1 \neq x_0$ (cioè il comportamento periodico della successione si instaura dopo uno o più termini).

b) Nel caso $r = 5$, $\beta = 7$, si determini $\alpha \neq 1$ in modo che p sia uno dei seguenti: $1, 2, 4, 8, 16, 32$. Se $p = 1$ si indichi x_0 tale che $x_1 = x_0$.

c) Si verifichi che se $\alpha \equiv 3 \ (\text{mod} 4)$, allora

$$\sum_{i=0}^{2^k - 1} \alpha^i \equiv 0 \ (\text{mod} 2^{k+1}), \quad \text{per ogni} \quad k \geq 1.$$

d) Si verifichi che $p = 2^r$ se e solo se β è dispari e $\alpha \equiv 1 \pmod 4$.

(Traccia: a) per $\alpha = 1$ è

$$x_n \equiv [x_0 + n\beta] \pmod{2^r}$$

e risulta $x_n = x_0$ per n tale che $n \equiv 0 \pmod{2^r}$, quindi $p = 2^r$. Per $\alpha > 1$ è

$$x_n \equiv [\alpha^n x_0 + \beta(1 + \alpha + \ldots + \alpha^{n-1})] \pmod{2^r}.$$

Sapendo che per ogni α vale

$$\alpha^n = 1 + (1 + \alpha + \ldots + \alpha^{n-1})(\alpha - 1),$$

ne segue che

$$x_n \equiv [x_0 + (1 + \alpha + \ldots + \alpha^{n-1})(x_0(\alpha - 1) + \beta)] \pmod{2^r},$$

e se $x_n = x_0$, è

$$(1 + \alpha + \ldots + \alpha^{n-1})\big[x_0(\alpha - 1) + \beta\big] \equiv 0 \pmod{2^r}.$$

Se α è dispari, $x_0(\alpha - 1) + \beta$ è dispari e non ha fattori comuni con 2^r, per cui deve essere

$$1 + \alpha + \ldots + \alpha^{n-1} \equiv 0 \pmod{2^r}.$$

Quindi p è il minimo di tali n, non dipende da x_0 e non può essere $p = 1$. Se α è pari, $1 + \alpha + \ldots + \alpha^{n-1}$ è dispari e non ha fattori comuni con 2^r, per cui deve essere

$$x_0(\alpha - 1) + \beta \equiv 0 \pmod{2^r},$$

cioè

$$x_1 \equiv [x_0\alpha + \beta] \pmod{2^r} \equiv x_0 \pmod{2^r};$$

il periodo p è uguale a 1 e la relazione è verificata solo per particolari x_0.

b) $p = 1$ per $\alpha = 4$ e $x_0 = 19$; $p = 2$ per $\alpha = 31$; $p = 4$ per $\alpha = 15$; $p = 8$ per $\alpha = 7$; $p = 16$ per $\alpha = 3$; $p = 32$ per $\alpha = 5$.

c) Si proceda per induzione: per $k = 1$ si ha $1 + \alpha$, che è divisibile per 4; per $k > 1$ è

$$\sum_{i=0}^{2^k-1} \alpha^i = (1 + \alpha^{2^{k-1}}) \sum_{i=0}^{2^{k-1}-1} \alpha^i,$$

in cui al secondo membro il primo fattore è divisibile per 2 in quanto α è dispari e il secondo fattore è divisibile per 2^k per l'ipotesi induttiva.

d) Per β pari, se x_0 è pari risulta x_i pari per ogni i, quindi non vengono generati numeri dispari e il periodo non può essere massimo. Ne segue che se il periodo è massimo, β è dispari e per quanto visto in a), α è dispari, e viceversa esistono valori dispari di α per cui il periodo è massimo qualunque sia β dispari. Un numero α dispari può soddisfare una sola delle due condizioni

$$\alpha \equiv 1 \ (\mathrm{mod}4) \quad \text{oppure} \quad \alpha \equiv 3 \ (\mathrm{mod}4).$$

Nel secondo caso per quanto visto in c) il periodo non può superare 2^{r-1}. Viceversa, se β è dispari e $\alpha \equiv 1$ (mod4), il periodo è massimo. Si dimostri infatti che per $\alpha \neq 1$ e per $r \geq 1$ è

$$(1) \quad \sum_{i=0}^{2^r-1} \alpha^i \equiv 0 \ (\mathrm{mod}2^r), \qquad (2) \quad \sum_{i=0}^{2^k-1} \alpha^i \not\equiv 0 \ (\mathrm{mod}2^r) \quad \text{per} \quad k < r,$$

procedendo per induzione su r nel modo seguente: per verificare la (1), posto $\alpha = 1 + 4q$, $q \neq 0$, si ha per $r = 1$

$$1 + \alpha = 2(1 + 2q) \equiv 0 \ (\mathrm{mod}2) \quad \text{e} \quad 1 \ \not\equiv 0 \ (\mathrm{mod}2);$$

e per $r > 1$

$$\sum_{i=0}^{2^r-1} \alpha^i = (1 + \alpha^{2^{r-1}}) \sum_{i=0}^{2^{r-1}-1} \alpha^i,$$

in cui al secondo membro il primo fattore è pari e il secondo fattore è divisibile per 2^{r-1} per l'ipotesi induttiva. Per verificare la (2), poiché

$$1 + \alpha \not\equiv 0 \ (\mathrm{mod}4)$$

e

$$\sum_{i=0}^{2^k-1} \alpha^i = (1 + \alpha^{2^{k-1}}) \sum_{i=0}^{2^{k-1}-1} \alpha^i,$$

si tenga conto del fatto che ogni numero della forma $1+\alpha^t$, in cui α è dispari e t è pari, è divisibile per 2 ma non per 4.)

7.54 Un metodo molto usato in passato per la generazione di numeri pseudocasuali per un calcolatore con aritmetica intera in base 2 e 31 cifre, si basa sulla formula

$$x_{i+1} \equiv (2^{16} + 3)x_i \ (\mathrm{mod}2^{31}), \quad x_0 \text{ dispari}$$

(implementata nel programma RANDU distribuito dall'IBM). Si verifichi che

a) la successione $\{x_i\}$ ha periodo 2^{29};

b) tre elementi consecutivi della successione verificano la relazione

$$x_{i+2} \equiv \left[6x_{i+1} - 9x_i\right] \ (\mathrm{mod} 2^{31}) \quad \text{per ogni} \quad i,$$

e quindi i numeri generati sono correlati, per cui la successione non può essere considerata una buona successione pseudocasuale.

(Traccia: a) posto $\gamma = 2^{14} + 1$, è $2^{16} + 3 = 4\gamma - 1$, per cui

$$x_{i+1} \equiv (4\gamma - 1)x_i \ (\mathrm{mod} 2^{31}),$$

e risulta

$$(-1)^{i+1}x_{i+1} - 1 \equiv \left[(1 - 4\gamma)[(-1)^i x_i - 1] - 4\gamma\right] \ (\mathrm{mod} 2^{31}).$$

Si verifichi che, se per esempio $x_0 - 1$ è divisibile per 4, allora tutti i numeri $[(-1)^i x_i - 1] \ (\mathrm{mod} 2^{31})$ sono divisibili per 4, e posto

$$4y_i \equiv [(-1)^i x_i - 1] \ (\mathrm{mod} 2^{31}),$$

risulta

$$4y_{i+1} \equiv \left[(1 - 4\gamma)4y_i - 4\gamma\right] \ (\mathrm{mod} 2^{31}),$$

da cui

$$y_{i+1} \equiv \left[(1 - 4\gamma)y_i - \gamma\right] \ (\mathrm{mod} 2^{29}).$$

La successione y_i risulta così generata con la (71), in cui $\beta = -\gamma = -2^{14} + 1$ è primo con $m = 2^{29}$ e $\alpha = -(2^{16} + 3)$ è tale che $\alpha \equiv 1 \ (\mathrm{mod} 4)$. Quindi la successione y_i ha periodo 2^{29}.
b) È

$$\begin{aligned} x_{i+2} &\equiv (2^{16} + 3)^2 x_i \ (\mathrm{mod} 2^{31}) \equiv (2^{32} + 6 \cdot 2^{16} + 9)x_i \ (\mathrm{mod} 2^{31}) \\ &\equiv (6 \cdot 2^{16} + 9)x_i \ (\mathrm{mod} 2^{31}) \equiv \left[6(2^{16} + 3)x_i - 9x_i\right] \ (\mathrm{mod} 2^{31}). \ ) \end{aligned}$$

7.55 Una delle tecniche di riduzione della varianza σ per il metodo Monte Carlo applicato al calcolo di

$$S = \int_a^b f(x) \ dx$$

consiste nel determinare una funzione $g(x)$ di cui sia noto l'integrale $\int_a^b g(x)\,dx$ e calcolare S nel modo seguente

$$S = \int_a^b g(x)\,dx + \int_a^b [f(x) - g(x)]\,dx,$$

applicando il metodo Monte Carlo al secondo integrale. Se $g(x)$ è tale che $|f(x) - g(x)| < \epsilon$ per $x \in [a, b]$, per la varianza di $(b-a)[f(x) - g(x)]$ risulta

$$\sigma_1^2 = (b-a)\int_a^b [f(x) - g(x)]^2\,dx - \left[\int_a^b [f(x) - g(x)]\,dx\right]^2 < \epsilon^2(b-a),$$

e per ϵ abbastanza piccolo σ_1 risulta minore di σ. La tecnica può essere estesa al calcolo di integrali multipli.

a) Si calcoli la varianza nel caso dell'integrale

$$S = \int_0^{\pi/2}\int_0^{\pi/2} f(x,y)\,dx\,dy, \quad f(x,y) = x\cos y + y\cos x;$$

b) Si determini un polinomio $g(x, y)$ di secondo grado in x e y che approssimi $f(x, y)$ e lo si utilizzi per ridurre la varianza.

(Traccia: a) è

$$\sigma^2 = \frac{\pi^2}{4}\int_0^{\pi/2}\int_0^{\pi/2} f(x,y)^2\,dx\,dy - \left[\int_0^{\pi/2}\int_0^{\pi/2} f(x,y)\,dx\,dy\right]^2 = 0.5269674;$$

b) un'approssimazione (ottenuta con tecniche minimax) di $\cos x$ sull'intervallo $[0, \pi/2]$ è data da

$$p(x) = 1.01 - 0.134\,x - 0.331\,x^2,$$

da cui si ottiene il polinomio

$$g(x,y) = x\,p(y) + y\,p(x) = 1.01(x+y) - 0.268\,xy - 0.331\,xy(x+y),$$

e si ha

$$\sigma_1^2 = \frac{\pi^2}{4}\int_0^{\pi/2}\int_0^{\pi/2} [f(x,y) - g(x,y)]^2\,dx\,dy - \left[\int_0^{\pi/2}\int_0^{\pi/2} [f(x,y) - g(x,y)]\,dx\,dy\right]^2 = 0.002626804.\ )$$

7.56 Si dia un'interpretazione geometrica delle formule di derivazione approssimata (82) e (83).

(Traccia: ci si riferisca alla figura 7.19.)

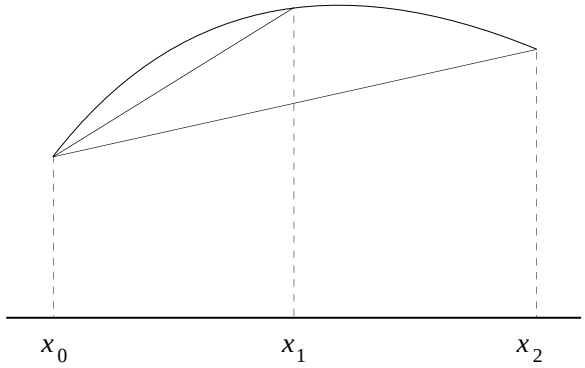


Fig. 7.19 - Approssimazione della derivata.

7.57 Si scrivano le formule di derivazione approssimata con nodi equidistanti $p_n^{(k)}(x_i)$, per $n = 4$ e $i, k = 1, \ldots, n$. Si indichino anche i corrispondenti resti $r_n^{(k)}(x_i)$ e si dica quali sono le formule con più basso errore analitico.

(Traccia: siano $x_i = x_0 + ih$, $i = 0, \ldots, 4$. Risolvendo il sistema (76), per la derivata prima si ottiene

$$p_4'(x_j) = \frac{1}{12h} \sum_{i=0}^{4} \alpha_i f(x_i), \quad r_4'(x_j) = \gamma_j h^4 f^{(5)}(\xi_j), \quad \xi_j \in (x_0, x_4),$$

dove i coefficienti α_i della formula e γ_j del resto sono

j	α_0	α_1	α_2	α_3	α_4	γ_j
0	-25	48	-36	16	-3	$\frac{1}{5}$
1	-3	-10	18	-6	1	$-\frac{1}{20}$
2	1	-8	0	8	-1	$\frac{1}{30}$
3	-1	6	-18	10	3	$-\frac{1}{20}$
4	3	-16	36	-48	25	$\frac{1}{5}$

Per la derivata seconda si ottiene

$$p_4''(x_j) = \frac{1}{12h^2} \sum_{i=0}^{4} \alpha_i f(x_i), \quad r_4''(x_j) = \gamma_j h^3 f^{(5)}(\xi_j), \quad \xi_j \in (x_0, x_4),$$

dove i coefficienti α_i della formula e γ_j del resto sono

j	α_0	α_1	α_2	α_3	α_4	γ_j
0	35	-104	114	-56	11	$-\frac{5}{6}$
1	11	-20	6	4	-1	$\frac{1}{12}$
2	-1	16	-30	16	-1	$(*)$
3	-1	4	6	-20	11	$-\frac{1}{12}$
4	11	-56	114	-104	35	$\frac{5}{6}$

$(*)$ in cui il resto della formula $p_4''(x_2)$ è $\frac{1}{90}\, h^4 f^{(6)}(\xi_2)$.

Per la derivata terza si ottiene

$$p_4'''(x_j) = \frac{1}{2h^3} \sum_{i=0}^{4} \alpha_i f(x_i), \quad r_4'''(x_j) = \gamma_j h^2 f^{(5)}(\xi_j), \quad \xi_j \in (x_0, x_4),$$

dove i coefficienti α_i della formula e γ_j del resto sono

j	α_0	α_1	α_2	α_3	α_4	γ_j
0	-5	18	-24	14	-3	$\frac{7}{4}$
1	-3	10	-12	6	-1	$\frac{1}{4}$
2	-1	2	0	-2	1	$-\frac{1}{4}$
3	1	-6	12	-10	3	$\frac{1}{4}$
4	3	-14	24	-18	5	$\frac{7}{4}$

Per la derivata quarta si ottiene

$$p_4^{(4)}(x_j) = \frac{1}{h^4}\left(f(x_0) - 4f(x_1) + 6f(x_2) - 4f(x_3) + f(x_4)\right), \quad j = 0, \dots, 4,$$
$$r_4^{(4)}(x_2) = -\frac{1}{6}h^2 f^{(6)}(\xi_2), \quad \xi_2 \in (x_0, x_4),$$
e $\quad r_4^{(4)}(x_j) = O(h) \quad$ per $\quad j = 0, 1, 3, 4.$)

7.58 Sia $f(x) = \dfrac{1}{k}\sin k^2 x$, per $x \in [0, \pi]$ e $k > 1$. Si stimino gli errori commessi approssimando $f(x)$ con il polinomio di interpolazione e $f'(x)$ con le formule di derivazione approssimata, usando come nodi i punti $x_i = \dfrac{i\pi}{k^2}$, $i = 0, \dots, 4$ e si confrontino per valori grandi di k.

(Traccia: è $f(x_i) = 0$, quindi $p_4(x) = 0$ e $p_4'(x) = 0$; quindi

$$\max_{x\in[0,\pi]} |f(x) - p_4(x)| = \frac{1}{k} \quad \text{e} \quad \max_{x\in[0,\pi]} |f'(x) - p_4'(x)| = k. \ )$$

7.59 a) Il procedimento di estrapolazione di Richardson può essere applicato anche all'approssimazione della derivata. Seguendo la traccia del procedimento applicato nel paragrafo 3 al calcolo dell'integrale, si indichi come procedere in questo caso.

b) Anche lo schema di Romberg può essere applicato all'approssimazione della derivata. Si indichi come procedere.

(Traccia: a) si applichi la (77), una volta con passo h e una volta con passo $h/2$, e siano $P(h)$ e $P(h/2)$ i due valori ottenuti. Dalla (81) risulta che

$$f^{(k)}(x_j) - P(h) = h^{n-k+1}\delta_1(h),$$
$$f^{(k)}(x_j) - P(h/2) = \frac{h^{n-k+1}}{2^{n-k+1}}\,\delta_2(h/2),$$

dove le funzioni $\delta_1(h)$ e $\delta_2(h/2)$ sono limitate in modulo nell'intervallo. Se $f^{(n+1)}(x)$ varia di poco nell'intervallo considerato, risulta $\delta_1(h) \approx \delta_2(h/2) = \delta$. Combinando le due relazioni si ha

$$P(h/2) - P(h) \approx \frac{2^{n-k+1} - 1}{2^{n-k+1}}\, h^{n-k+1}\delta,$$

da cui si ottiene la stima

$$f^{(k)}(x_j) - P(h/2) \approx \frac{P(h/2) - P(h)}{2^{n-k+1} - 1}\,.$$

Si può così correggere l'ultimo valore ottenuto con

$$P(h/2) + \frac{P(h/2) - P(h)}{2^{n-k+1} - 1} = \frac{2^{n-k+1} P(h/2) - P(h)}{2^{n-k+1} - 1} .$$

b) Fissati un punto x e un passo h si costruisca una tabella $T_j^{(s)}$ nel modo seguente: gli elementi della prima colonna sono ottenuti applicando la formula centrale (83) nella forma

$$T_0^{(s)} = \frac{f(x + h_s) - f(x - h_s)}{2h_s} , \quad h_s = \frac{h}{2^s} , \quad s = 0, 1, \ldots ;$$

gli elementi delle altre colonne sono ottenuti con la relazione

$$T_j^{(s)} = \frac{4^j T_{j-1}^{(s+1)} - T_{j-1}^{(s)}}{4^j - 1} , \quad s = 0, 1, \ldots , \ j = 1, 2, \ldots$$

Si verifichi, usando la formula di Taylor, che gli elementi della tabella sono delle approssimazioni di $f'(x)$ con errori analitici dell'ordine di $h^{2(j+1)}$.)

Commento bibliografico

Le formule ora note come formule di Newton-Cotes comparvero per la prima volta in una lettera di Newton a Leibniz del 1676, in cui Newton suggeriva di approssimare l'area sotto una curva con la corrispondente area del polinomio di interpolazione. Cotes ne calcolò poi i coefficienti fino all'ordine $n = 11$. Casi particolari delle formule però erano conosciuti già prima. La regola dei trapezi si può far risalire ai babilonesi. La regola di Simpson, era già stata pubblicata da Cavalieri nel 1639 e da Gregory nel 1668, mentre Simpson la pubblicò solo nel 1743. Il primo trattamento analitico del resto delle formule di Newton-Cotes è stato fatto da Steffensen nel 1921 e Walther nel 1925.

Nel 1814 Gauss dimostrò, facendo ricorso a frazioni continue associate a serie ipergeometriche, che era possibile ottenere formule con un maggior grado di precisione scegliendo come nodi punti non equidistanti. Si trattava delle formule oggi note come Gauss-Legendre. Nel 1826 Jacobi ottenne lo sesso risultato basandosi solamente su argomenti di ortogonalità dei polinomi. Il lavoro fondamentale sulle formule gaussiane è di Christoffel nel 1858. Gli integrali con pesi diversi furono introdotti nel 1864 da Mehler. Il teorema sulla convergenza delle formule gaussiane fu dato da Stieltjes nel 1884. L'estensione all'intervallo seminfinito con il peso e^{-x} è di Radau nel 1883. L'estensione all'intervallo infinito con il peso e^{-x^2} è di Gourier nel

1883. Radau, a cui si devono molte formule, introdusse il termine "grado di precisione" nel 1880.

La trattazione più esauriente sull'integrazione numerica è il libro di Davis e Rabinowitz [5], in cui, oltre alla parte teorica sono riportate formule per ogni tipo di problema, programmi FORTRAN dei metodi più importanti e una bibliografia di un migliaio di titoli. Più recente è il libro di Engels [7]. Ottime trattazioni più elementari si trovano in [11] e in [1]. Una trattazione specifica per le formule gaussiane è fatta in [20], in cui sono riportati anche i nodi e i coefficienti delle formule più comuni con le stime degli errori.

Nell'ultimo mezzo secolo, partendo dal principio dell'estrapolazione al limite, introdotto da Richardson nel 1927, si è sviluppato un particolare indirizzo di ricerca nel campo dell'integrazione approssimata, quello del controllo automatico della precisione dei risultati. Lo schema più significativo basato sulle formule newtoniane è quello di Romberg del 1955, di cui si può trovare un'ampia discussione in [2]. Allo schema di Romberg sono state proposte molte variazioni; si veda ad esempio [3]. Formule di integrazione automatica basate sui polinomi di Chebyshev sono invece quelle di Clenshaw-Curtis [4], rese particolarmente efficienti dall'uso della FFT proposto da Gentleman [9]. Lo schema più significativo basato sulle formule gaussiane è quello di Patterson [15], che utilizza una famiglia di formule con nodi prefissati, costruite secondo il principio considerato da Kronrod nel 1964.

Con l'introduzione di potenti sistemi di calcolo si è poi aperto il nuovo settore del software per la quadratura automatica: sono stati creati dei sistemi di sottoprogrammi che con raffinate tecniche adattive sono in grado di dare delle ottime approssimazioni con le corrispondenti stime degli errori. Il primo programma adattivo, basato sulla formula di Cavalieri-Simpson, è stato pubblicato da McKeeman nel 1962. Successivamente de Boor ha sviluppato il programma CADRE [6], che usa modelli polinomiali di ordine differente sui diversi intervalli. Un altro sistema di programmi molto usato è QUADPACK [16]. Un'interessante discussione sul software per l'integrazione, con confronti e suggerimenti si trova in [18].

Per le tecniche di integrazione approssimata in più dimensioni si veda [21]. Algoritmi adattivi per la quadratura automatica in più dimensioni si trovano in [12] e [13].

I metodi Monte Carlo, che si basano su esperimenti con numeri casuali, sono stati utilizzati, a partire dagli anni 50, per affrontare problemi difficilmente trattabili per altra via, come ad esempio nel campo della ricerca operativa e della fisica nucleare. Il nome e lo sviluppo sistematico risalgono al 1944, quando il metodo venne utilizzato da Von Neumann, Ulam e Fermi come strumento di ricerca nella seconda guerra mondiale, per simulare problemi probabilistici legati alla diffusione dei neutroni nel materiale fissile.

Attualmente la ricerca sul metodo Monte Carlo comprende il riconoscimento dei problemi nei quali il metodo rappresenta la migliore, se non l'unica tecnica disponibile. Per una trattazione del metodo Monte Carlo si veda [10], per la generazione dei numeri casuali si veda anche [14], [17], [8].

Bibliografia

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & Sons, New York, 1978.

[2] F. L. Bauer, H. Rutishauser, E. L. Stiefel, "New Aspects in Numerical Quadrature", *Proc. of Symp. in Appl. Math.*, vol. 15: *High Speed Computing and Experimental Arithmetic*, pp. 199-218, Amer. Math. Soc., Providence, R. I., 1963.

[3] R. Bulirsch, J. Stoer, "Handbook Series Numerical Integration: Numerical Quadrature by Extrapolation", *Numer. Math.*, 9, 1967, pp. 271-278.

[4] C. W. Clenshaw, A. R. Curtis, "A Method for Numerical Integration on an Automatic Computer", *Numer. Math.*, 2, 1960, pp. 197-205.

[5] P. J. Davis, P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, New York, 1975.

[6] C. de Boor, "CADRE: An Algorithm for Numerical Quadrature", in *Mathematical Software*, ed. J. R. Rice, Academic Press, New York, 1971, pp. 417-449.

[7] H. Engels, *Numerical Quadrature and Cubature*, Academic Press, New York, 1980.

[8] G. E. Forsythe, M. A. Malcom, C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, 1977.

[9] W. M. Gentleman, "Implementing Clenshaw-Curtis Quadrature", *Comm. ACM*, 15, 1972, pp. 337-346.

[10] J. M. Hammersley, D. C. Handscomb, *Monte Carlo Methods*, Methuen's Monograph on Applied Probability and Statistics, London, 1964.

[11] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[12] D. K. Kahaner, O. W. Rechard, "TWODQD an Adaptive Routine for Two-Dimensional Integration", *J. Comput. Appl. Math.*, 17, 1987, pp. 215-234.

[13] D. K. Kahaner, M. B. Wells, "An Experimental Algorithm for N-Dimensional Adaptive Quadrature", *ACM Trans. Math. Soft.*, 5, 1979, pp. 86-96.

[14] D. E. Knuth, *The Art of Computer Programming, vol. 2, Seminumerical Algorithms*, Addison-Wesley, Reading, Mass., 1969.

[15] T. N. L. Patterson, "The Optimum Addition of Points to Quadrature Formulae", *Math. Comp.*, 22, 1968, pp. 847-856.

[16] R. Piessens, E. de Donker-Kapenga, C. W. Überhuber, D. K. Kahaner, *QUADPACK: A Subroutine Package for Automatic Integration*, Springer-Verlag, Berlin, 1983.

[17] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986.

[18] J. R. Rice, *Numerical Methods, Software and Analysis*, McGraw-Hill, New York, 1983.

[19] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[20] A. H. Stroud, D. H. Secrest, *Gaussian Quadrature Formulas*, Prentice-Hall, Englewood Cliffs, 1966.

[21] A. H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall, Englewood Cliffs, 1971.

[22] J. V. Uspensky, "On the Convergence of Quadrature Formulas Related to an Infinite Interval", *Trans. AMS*, 30, 1928, pp. 542-559.

Bibliografia generale

N. I. Achieser, *Theory of Approximation*, Unger, New York, 1956.

K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & Sons, New York, 1978.

G. A. Baker, P. Graves-Morris, *Padé Approximants*, Encyclopedia of Mathematics and its Applications, vol. 13, 14, Addison-Wesley, Reading, 1981.

D. Bini, M. Capovani, G. Lotti, F. Romani, *Complessità Numerica*, Boringhieri, Torino, 1981.

A. Borodin, I. Munro, *The Computational Complexity of Algebraic and Numeric Problems*, Elsevier Computer Science Library, New York, 1975.

E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

G. Dahlquist, Å. Björk, N. Anderson, *Numerical Methods*, Prentice Hall, Englewood Cliffs, N. J., 1974.

P. J. Davis, *Interpolation and Approximation*, Dover Inc., New York, 1975.

P. J. Davis, P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, New York, 1975.

C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

D. F. Elliott, K. R. Rao, *Fast Transforms Algorithms Analises, Applications*, Academic Press, New York, 1982.

H. Engels, *Numerical Quadrature and Cubature*, Academic Press, New York, 1980.

C. T. Fike, *Computer Evaluation of Mathematical Functions*, Prentice Hall, Englewood Cliffs, N. J., 1968.

G. E. Forsythe, M. A. Malcom, C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, 1977.

J. M. Hammersley, D. C. Handscomb, *Monte Carlo Methods*, Methuen & Co., Ltd, London, 1964.

D. C. Handscomb, *Methods of Numerical Approximation*, Pergamon Press, Oxford, 1966.

A. S. Householder, *The Numerical Treatment of a Single Nonlinear Equation*, Mac Graw-Hill, New York, 1970.

E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

W. B. Jones, W. J. Thron, *Continued Fractions, Analytic Theory and Applications*, Encyclopedia of Mathematics and its Applications, vol. 11, Addison-Wesley, Reading, 1980.

D. E. Knuth, *The Art of Computer Programming, vol. 2, Seminumerical Algorithms*, Addison-Wesley, Reading, Mass., 1969.

L. Kronsjö, *Algorithms, their Complexity and Efficiency*, John Wiley & Sons, New York, 1979.

P. J. Laurent, *Approximation et Optimization*, Hermann, Paris, 1972.

G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer-Verlag, Berlin, 1967.

L. M. Milne-Thomson, *The Calculus of Finite Differences*, Macmillan and Co., London, 1933.

J. M. Ortega, W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

A. M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1960.

P. P. Petrushev, V. A. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, 1987.

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes. The Art of Scientific Computing*, Cambridge Univ. Press, 1986.

J. R. Rice, *The Approximation of Functions, vol. 1. Linear Theory*, Addison-Wesley, Reading, Mass., 1964.

J. R. Rice, *Numerical Methods, Software, and Analysis*. Mc Graw-Hill, New York, 1983.

T. J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, New York, 1969.

T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley & Sons, New York, 1974.

P. H. Sterbenz, *Floating-Point Computation*. Prentice-Hall, Englewood Cliffs, N. J., 1974.

J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc., Providence, R. I., 1967.

J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, N. J., 1964.

J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, N. J., 1963.

J. Wimp, *Computation with Recurrence Relations*, Pitman, 1984.

Abel, trasformazione di, 308
Aberth, metodo di, 243
adattiva, strategia, 774, 777, 831
addizione
  di macchina, algoritmo per, 46, 47, 84
  di n numeri, errore nella, 60, 72, 76
  in parallelo, algoritmo di, 61, 92
Aitken, metodo di, 146-149, 216
al primo ordine, analisi dell'errore, 51
albero, matrice ad, 285
algoritmico, errore, 9, 50, 67, 72, 380
algoritmo, 7, 55
  di addizione in parallelo, 61, 92
  di Clenshaw, 549
  di Cooley e Tukey, 428, 478
  di Csanky, 19
  di Euclide, 397
  di Neville, 460
  di Remez, 563, 574, 581, 585, 588, 666, 702, 709
  di Sande e Tukey, 430, 478
  di scambio, 618, 696
  di Strassen, 14
  di Winograd, 21, 431
  FFT, 17, 24, 26, 426, 430
  Las Vegas, 21
  per l'addizione di macchina, 46, 47, 84
  stabilità di un, 7, 52
all'indietro, analisi dell'errore, 71-75
ampiezza di un'armonica, 432
amplificazione, coefficiente di, 52, 94, 95
analisi
  automatica dell'errore, 79-82
  di Fourier, 7
  di significatività, 80
  statistica dell'errore, 75-79, 96
analisi dell'errore
  al primo ordine, 51
  all'indietro, 71-75
  in avanti, 71
analitico, errore, 9, 50, 53, 67, 379, 739
antidifferenza, operatore, 262, 304
approssimanti di Padé, 597, 703, 706, 708
  convergenza di, 612
approssimazione
  ai minimi quadrati, 517, 540-552
  ai minimi quadrati, convergenza, 551
  delle derivate, 787-794, 843
  di $\arcsin x$, 562, 592, 648, 676
  di Chebyshev, ved. approssimazione minimax
  di minima deviazione, 628-632
  di Padé, 597-612
  di $|x|$, 377, 542, 545, 624, 649, 653, 663
  lineare, 512
  nel discreto, 612-621, 695
  quasi minimax, 570-578, 613
  resto della, 512, 552, 557, 634
  trigonometrica, 657
approssimazione minimax, 520, 552-569
  con vincoli, 583, 682, 703
  convergenza della, 562
  iperbolica, 589
  lineare, 557
  nel discreto, 617-621
  razionale, 585-593, 683, 703, 706, 711
  relativa, 579, 682, 711
$\arcsin x$, approssimazione di, 562, 592, 648, 676
$\arctan x$, calcolo di, 253, 706
aritmetica
  di macchina, 45
  finita, 2
  in virgola mobile, 45
  noisy-mode, 80
aritmetica degli intervalli, 80
armonica di un segnale, 432
arresto, criteri di, 104, 112
arrotondamento, 39, 115, 431
ascendente, polinomio di Newton, 456
automatica dell'errore, analisi, 79-82
automatica, quadratura, 773-778

B-spline cubica, 502
Bairstow, metodo di, 184-188
base
  conversione di, 34, 83
  di una rappresentazione, 30
  ortogonale, 516
Bernoulli
  metodo di, 188-192, 237
  numeri di, 267, 314
  polinomi di, 267, 314
Bernstein, polinomio di, 622

Bessel
disuguaglianza di, 516
funzioni di, 294
bilineare, interpolazione, 463
bisezione, metodo di, 104, 178
bit reversal, 429
Brent, funzione di, 131
Budan-Fourier, regola di, 234
byte, 38

calcolo
complessità di, 12
di $\arctan x$, 253, 706
di $\cos x$, 701
di e^x, 708
di $\log x$, 710
di π, 20, 253, 272, 688
di $\sin x$, 701
di $\tan x$, 703
di $\sqrt{x}$, 700
di x^n, 22, 92
parallelo, 18
sequenziale, 18
calcolo della DFT, 419, 426-431
costo computazionale del, 426
errore di arrotondamento nel, 431
calcolo di un polinomio
complessità del, 374, 488, 489, 655
errore nel, 63-67
in un punto, 222
trigonometrico, complessità del, 480
calcolo di una funzione, errore nel, 50-53
cambiamenti di segno, 176
cammino discendente a gradini, 604
cancellazione numerica, 54
caratteristica, 37
caratteristica, equazione, 278
Cardano, formula di, 231
Cartesio, regola dei segni di, 234
Casorati, matrice di, 328
Cavalieri-Simpson, formula di, 737, 780, 798, 836
Chebyshev
approssimazione di, ved. approssimazione minimax
disuguaglianza di, 76
insieme di, 660
nodi di, 566, 573, 578, 677, 680
polinomi di, 340, 527, 531, 534, 543, 551, 569, 570, 616, 638, 645, 753, 832
punti di, 365, 382
retta di, 695
serie di, 572
teorema di, 554, 579, 584, 586, 660, 617
Christoffel-Darboux, formula di, 525, 644, 744, 832
cifre
di una rappresentazione, 33
significative, 44
Clenshaw, algoritmo di, 549
Clenshaw-Curtis, formule di, 777, 831
Cline, metodo di, 618
coefficiente di amplificazione, 52, 94, 95
coefficienti
costanti, equazione a, 278
di Fourier, 515
di una formula di quadratura, 720, 726
indeterminati, metodo dei, 284, 722, 788, 809
non costanti, equazione a, 329
uniformi, formule a, 727, 813
complessi, zeri, 180, 184
complessità computazionale, 14, 218
complessità di calcolo, 12
del polinomio di interpolazione, 356, 374
del prodotto di due polinomi, 488
del quoziente di due polinomi, 488
di un polinomio, 374, 488, 489, 655
di un polinomio trigonometrico, 480
completa, spline, 437, 444, 495
completo, insieme, 516
composte, formule, 736-743, 750
computazionale
complessità, 14, 218
costo, 7, 19, 22, 426, 501
condizionamento degli zeri di un polinomio, 166
condizioni
al contorno, 290
di Haar, 659
iniziali, 274
congruenziale, metodo, 783, 838
continua, frazione, 395-403, 405, 466
contorno, condizioni al, 290
convergente, metodo, 108
convergenza
del metodo delle corde, 123
del metodo delle secanti, 136, 338
del metodo delle tangenti, 134
dell'algoritmo di Remez, 567, 588, 669, 672

dell'interpolazione polinomiale, 364, 462
dell'interpolazione trigonometrica, 425, 483
di approssimanti di Padé, 612
di espansioni in frazioni continue, 594
di formule di quadratura, 725, 796, 737
di ordine p, 118-122
in media, 514
lineare, 118
locale, 108
sublineare, 118, 132
superlineare, 118, 140
teorema centrale di, 78
convergenza dell'approssimazione
ai minimi quadrati, 551
minimax, 562
minimax nel discreto, 620
minimax razionale, 592
conversione di base, 34, 83
convessa, funzione, 161, 221
convesso, insieme, 161, 221, 513
convoluzione circolare discreta, 479
Cooley e Tukey, algoritmo di, 428, 478
corde, metodo delle, 123
$\cos x$, calcolo di, 701
coseni, trasformata di, 481, 547, 834
costante
di Lebesgue, 382, 453
di normalizzazione, 522, 527, 743
costanti, equazione a coefficienti, 278
costo computazionale, 7, 19, 22
del calcolo della DFT, 426
della costruzione di una spline, 501
Cramer, metodo di, 13, 468
criteri di arresto, 104, 112
Csanky, algoritmo di, 19
cubica a tratti di Hermite, 435, 495
cubica, spline, 436-446, 502, 742
curvatura globale, 442

de la Vallée-Poussin, teorema, 553, 660
Dedekind, metodo di, 211
deflazione di un polinomio, 181
Dekker-Brent, metodo di, 145, 218, 391
densità di probabilità, 77
derivata
approssimazione della, 787-794, 843
errore nel calcolo della, 5, 793
derivazione approssimata, resto di una formula di, 789, 845
DFT, calcolo della, 419, 426-431
diagramma a losange, 455
differenze
divise, 370, 384, 452, 461, 490, 731, 745
divise, simmetria delle, 375, 461
equazione alle, 274-286, 327
finite, 256, 257, 298, 375, 454
inverse, 404, 469
operatore lineare alle, 344
reciproche, 410
reciproche, simmetria delle, 411
digamma, funzione, 266, 311
Diofanto, equazione di, 468
discendente
cammino a gradini, 604
polinomio di Newton, 456
discretizzazione, processo di, 8
discreto, approssimazione nel, 612-621, 695
distanza da un sottospazio di polinomi, 510, 679
distribuzione logaritmica, 99
disuguaglianza
di Bessel, 516
di Chebyshev, 76
triangolare, 512
divisa, differenza, 370, 375, 384, 452, 461, 490, 731, 745
divisione di polinomi, 26, 488
divisione sintetica, 16, 179
due punti, formula dei, 728
Durand-Kerner, metodo di, 241

economizzazione, 570, 701, 706, 710
efficienza
di un metodo iterativo, 142
informativa, 144
equazione
a coefficienti costanti, 278
a coefficienti non costanti, 329
alle differenze, 274-286, 327
caratteristica, 278
di Diofanto, 468
di Poisson, 7
di quarto grado, 232
di secondo grado, 2, 54
di terzo grado, 231, 233
equazioni non lineari, 104-155, 166-197
equioscillazione
teorema di, 554, 579, 584, 586, 660, 617
punti di, 553, 586
equivalenti, frazioni continue, 403

errore
algoritmico, 9, 50, 67, 72, 380
analisi automatica, 79-82
analisi statistica, 75-79, 96
analitico, 9, 50, 53, 67, 379, 739
assoluto, 40
assoluto in norma, 512
complementare, funzione, 78
di arrotondamento nel calcolo della DFT, 431
di rappresentazione, 42, 57
funzione, 77, 96, 784
inerente, 9, 50
locale, 45, 53, 56
nel prodotto di n numeri, 90
nell'addizione di n numeri, 60, 72, 76
nelle operazioni di macchina, 53
relativo, 40
totale, 50
errore nel calcolo
del logaritmo, 68
dell'esponenziale, 4, 70
della derivata, 5, 793
della differenza finita, 454
di un polinomio, 63-67
di una funzione, 50-53
errori di arrotondamento, 115
del polinomio di interpolazione, 379
espansione in serie di Chebyshev, 552
esponente, 33
esponenziale, interpolazione, 492
estrapolazione di Richardson, 741, 760, 818, 845,
Euclide, algoritmo di, 397
Eulero
costante di, 266, 313
funzione gamma di, 264, 309
trasformazione di, 271, 319, 819
Eulero-Maclaurin, formula di, 97, 323, 496, 775, 831
e^x, calcolo di, 708
errore nel, 4, 70

falsa posizione, metodo di, 137
fattoriale, potenza, 258, 299, 301
FFT, algoritmo, 17, 24, 26, 426, 430
Fibonacci, numeri di, 280, 337
filtraggio digitale, 432
finita
aritmetica, 2
differenza, 256, 257, 298, 375, 454
formula
dei due punti, 728
dei punti di mezzo, 751, 804
dei rettangoli, 795
dei trapezi, 484, 496, 547, 737, 798, 835
dei tre punti, 739
di Cardano, 231
di Cavalieri-Simpson, 737, 780, 798, 836
di Christoffel-Darboux, 525, 644, 744, 832
di Eulero-Maclaurin, 97, 323, 496, 775, 831
di Gauss, 266, 312
di Gauss-Chebyshev, 754, 763, 815
di Gauss-Hermite, 756, 767, 807, 816, 822
di Gauss-Laguerre, 755, 767, 807, 816, 822
di Gauss-Legendre, 746, 772, 782, 805, 813
di Gregory-Newton, 302
di Lobatto, 769, 772, 826
di Radau, formula di, 769, 823
di Rodrigues, 526, 530, 531, 534, 537, 539
di Stirling, 325, 454, 650
di Wallis, 325
di Woodbury, 440
interpolatoria, 722
formule
a coefficienti uniformi, 727, 813
composte, 736-743, 750
di Clenshaw-Curtis, 777, 831
di derivazione approssimata, 787-794, 843
di Kronrod, 772
di Newton-Cotes, 727, 728
di Newton-Cotes di tipo aperto, 803
gaussiane, 727, 743-752
gaussiane con nodi prefissati, 768-773
gaussiane pesate, 752-759
newtoniane, 727, 728-736, 752
prodotto, 779-783, 835
ricorrenti, stabilità delle, 286, 346
formule di quadratura, 720-772
coefficienti di, 720, 726
convergenza delle, 725, 796, 737
di Hermite, 811
Fourier
analisi di, 7

coefficienti di, 515
serie di, 270, 316, 483, 545, 657
trasformata discreta di, 419, 426-431, 473
frazione continua, 395-403, 405, 466
di Thiele, 406, 415, 470, 599
infinita, 593
residuo di una, 396
frazione parziale, 395
frazioni continue
convergenza di espansioni in, 594
equivalenti, 403
Frobenius, matrice di, 172, 189, 237
funzione
a quadrato sommabile, 517
convessa, 161, 221
cubica a tratti di Hermite, 435, 495
di Brent, 131
di Runge, 363, 578
digamma, 266, 311
errore, 77, 96, 784
errore complementare, 78
errore nel calcolo di una, 50-53
gamma di Eulero, 264, 309
lineare a tratti, 434, 493
peso, 517
polinomiale a tratti, 434
totalmente differenziabile, 153
funzioni
di Bessel, 294
linearmente indipendenti, 352, 512
simmetriche elementari, 224, 492

gamma di Eulero, funzione, 264, 309
Gauss
formula di, 266, 312
metodo (di eliminazione) di, 13, 19, 440
polinomi di, 456
Gauss-Chebyshev, formula di, 754, 763, 815
Gauss-Hermite, formula di, 756, 767, 807, 816, 822
Gauss-Laguerre, formula di, 755, 767, 807, 816, 822
Gauss-Legendre, formula di, 746, 772, 782, 805, 813
gaussiane, formule, 727, 743-752
con nodi prefissati, 768-773
pesate, 752-759
generale, soluzione, 277, 335
Gentleman e Sande, teorema di, 431
Girard-Newton, relazioni di, 224
gradini, cammino discendente a, 604
grado di precisione, 721, 726, 745, 768, 788
Gräffe, metodo di, 239
grafo, 56
Gram, polinomi di, 617
Gregory-Newton, formula di, 302

Haar, condizioni di, 659
Halley, metodo di, 213, 642
hardware, 14
Hermite
formule di quadratura di, 811
funzione cubica a tratti di, 435, 495
polinomi di, 527, 539, 641, 757
polinomio osculatore di, 367-370, 389, 452, 680, 811
Hilbert
matrice di, 518
spazio di, 514

identità, operatore, 256
Illinois, metodo, 214
in avanti, analisi dell'errore, 71
indeterminati, metodo dei coefficienti, 284, 722, 788, 809
inerente, errore, 9, 50
infinita
frazione continua, 593
molteplicità, 126
informativa, efficienza, 144
iniziali, condizioni, 274
insieme
completo, 516
convesso, 161, 221, 513
di Chebyshev, 660
strettamente convesso, 626
integrali impropri, 759-768, 817
integrazione in più dimensioni, 779-783, 785, 835, 837, 842
interi, prodotto di, 24
interpolatoria, formula, 722
interpolazione, 352-446
bilineare, 463
complessità di calcolo, 356, 374
esponenziale, 492
inversa, 145, 390
lineare, 355, 359, 372, 447
multidimensionale, 462
nei nodi di Chebyshev, 365, 573, 578
osculatoria, 367-370

osculatoria, resto della, 368
parabolica, 447
polinomiale, convergenza della, 364, 462
polinomiale, 352-391
polinomiale, resto della, 358, 361, 375, 448
razionale, 392, 395, 406, 465
trigonometrica, 417-433
trigonometrica, convergenza della, 425, 483
intervalli, aritmetica degli, 80
intrinsecamente complesso, problema, 17
inversa
differenza, 404, 469
interpolazione, 145, 390
iperbolica, approssimazione minimax, 589
iterazione funzionale, metodo di, 106

Jackson, teorema di, 562

Kronrod, formule di, 772

Lagrange, polinomio di, 354-357, 380, 450, 491, 573
Laguerre
metodo di, 238, 643
polinomi di, 527, 536, 641, 755
Las Vegas, algoritmo, 21
Lebesgue, costante di, 382, 453
Legendre, polinomi di, 527, 530, 541, 635, 645, 746, 812, 823
legge del parallelogramma, 659
Leibniz, regola di, 303
lineare
a tratti, funzione, 434, 493
approssimazione, 512
approssimazione minimax, 557
convergenza, 118
equazione alle differenze, 274-286, 327
interpolazione, 355, 359, 372, 447
triangolare, sistema, 73
linearmente indipendenti
funzioni, 352, 512
soluzioni, 276, 278, 329, 333
Lipschitz, condizione di, 197
Lobatto, formula di, 769, 772, 826
locale
convergenza, 108
errore, 45, 53, 56
localizzazione degli zeri di un polinomio, 171, 228
$\log x$, calcolo di, 710
logaritmica, distribuzione, 99
logaritmo, errore nel calcolo del, 68
losange, diagramma a, 455
lunghezza di una rappresentazione, 33

macchina
aritmetica di, 45
numero di, 36
precisione di, 42
Maehly, metodo di, 183
mal condizionamento di un sistema lineare, 10
mal condizionato, problema, 10, 52, 348
mantissa, 33
marching, metodo del, 7
matrice
ad albero, 285
di Casorati, 328
di Frobenius, 172, 189, 237
di Hilbert, 518
di Toeplitz, 24, 25, 599
di Vandermonde, 353, 418, 447, 492, 614, 618, 722, 788
tridiagonale, 276, 283, 332, 340, 439, 467, 643, 807
matrici, prodotto di, 22
memoria, metodo con, 144
Mersenne, numero di, 486
metodo
a più punti, 144
ad un punto, 143
con memoria, 144
congruenziale, 783, 838
convergente, 108
dei coefficienti indeterminati, 284, 722, 788, 809
del marching, 7
della variazione dei parametri, 284, 343
delle corde, 123
delle potenze, 189
delle secanti, 135-142, 338
delle tangenti, 125-134, 700
di Aberth, 243
di Aitken, 146-149, 216
di Bairstow, 184-188
di Bernoulli, 188-192, 237
di bisezione, 104, 178
di Cline, 618
di Cramer, 13, 468

di Dedekind, 211
di Dekker-Brent, 145, 218, 391
di Durand-Kerner, 241
di eliminazione di Gauss, 13, 19, 440
di falsa posizione, 137
di Gräffe, 239
di Halley, 213, 642
di iterazione funzionale, 106
di Laguerre, 238, 643
di Maehly, 183
di Miller, 291
di Newton, 125, 179, 805
di Newton di secondo grado 206
di Newton-Raphson, 155, 220, 240, 673
di Olver, 295
di Pasquini-Trigiante, 242
di Ruffini-Horner, 16, 22, 64, 93, 222, 237, 260, 377, 547, 550
di sostituzione, 73
di Steffensen, 144, 215
di Viskovatov, 401, 596, 687
Illinois, metodo, 214
Monte Carlo, 20, 783-786, 842
qd, 192-197, 606-610
metodo iterativo, efficienza di un, 142
metodi iterativi per,
equazioni non lineari, 104-155
sistemi non lineari, 150-165
Miller, metodo di, 291
minima deviazione, approssimazione di, 628-632
minimale, soluzione, 286, 289, 347
minimax, approssimazione, 520, 552-569
con vincoli, 583, 682, 703
convergenza della, 562
iperbolica, 589
lineare, 557
nel discreto, 617-621
razionale, 585-593, 683, 703, 706, 711
relativa, 579, 682, 711
minimi quadrati, approssimazione ai, 517, 540-552
modello, 352, 511
probabilistico, 20
modulo e segno, rappresentazione in, 37
molteplicità, 125, 166, 208, 210, 278
infinita, 126
momento, 437
Monte Carlo, metodo, 20, 783-786, 842
multidimensionale, interpolazione, 462
naturale, spline, 437, 441, 493, 495
Neville, algoritmo di, 460
Newton con le differenze finite, polinomio di, 376, 791, 802
ascendente, 456
discendente, 456
Newton, metodo di, 125, 179, 805
di secondo grado 206
Newton, polinomio di, 370, 372, 462,
Newton-Cotes, formule di, 727, 728
di tipo aperto, 803
Newton-Raphson, metodo di, 155, 220, 240, 673
newtoniane, formule, 727, 728-736, 752
nodi
dell'interpolazione, 352
di Chebyshev, 566, 573, 578, 677, 680
di una formula di quadratura, 720
noisy-mode, aritmetica, 80
non lineare, equazione alle differenze, 330
norma, 511
errore assoluto in, 512
norma 1, 628
norma 2, 517, 612
norma infinito, 520, 612
normale, sistema, 515, 614, 627
normalizzata, rappresentazione, 33
normalizzazione, costante di, 522, 527, 743
numeri
di Bernoulli, 267, 314
di Fibonacci, 280, 337
di Stirling, 259, 301
pseudocasuali, 783, 840
numero
di macchina, 36
di Mersenne, 486

Olver, metodo di, 295
omogenea, equazione alle differenze, 276, 329, 334, 348, 651
operatore
antidifferenza, 262, 304
identità, 256
lineare alle differenze, 344
somma, 261, 304
traslazione, 256, 298
operazioni di macchina, 44
errore nelle, 53
ordine almeno p, 119
ordine di convergenza, 118-122, 124, 126, 148

ortogonale, base, 516
ortogonali, polinomi, 522, 616, 743
osculatore, polinomio, 366-370, 451
Ostrowski, teorema di, 174
overflow, 1, 39, 41, 49, 88, 93

Padé
  approssimanti di, 597, 703, 706, 708
  approssimazione di, 597-612
  tabella di, 601, 603, 690, 691, 693
parabolica, interpolazione, 447
parallelo
  algoritmo di addizione in, 61, 92
  calcolo, 18
parallelogramma, legge del, 659
Parseval, uguaglianza di, 516, 541
particolare, soluzione, 277, 290, 334
parziale, frazione, 395
Pasquini-Trigiante, metodo di, 242
passo doppio, tecnica del, 181
Patterson, schema di, 773
periodica, spline, 437
pesi di una formula di quadratura, 720
peso, funzione, 517
più dimensioni, integrazione in, 779-783, 785, 835, 837, 842
più punti, metodo a, 144
Poisson, equazione di, 7
polinomi
  di Bernoulli, 267, 314
  di Gauss, 456
  di Gram, 617
  divisione di, 26, 488
  prodotto di, 23, 488
polinomi ortogonali, 522, 616, 743
  di Chebyshev, 340, 527, 531, 534, 543, 551, 569, 570, 616, 638, 645, 753, 832
  di Hermite, 527, 539, 641, 757
  di Laguerre, 527, 536, 641, 755
  di Legendre, 527, 530, 541, 635, 645, 746, 812, 823
  ultrasferici, 527, 529
  zeri di, 522, 642, 724
polinomiale a tratti, funzione, 434
polinomiale, interpolazione, 352-391
polinomio
  complessità di calcolo, 374, 488, 489, 655
  condizionamento degli zeri di un, 166
  deflazione di un, 181
  di Bernstein, 622
  di Lagrange, 354-357, 380, 450, 491, 573
  di Newton, 370, 372, 376, 462, 791, 802
  di Newton ascendente, 456
  di Newton discendente, 456
  di Stirling, 457
  errore nel calcolo di un, 63-67
  in un punto, calcolo di un, 222
  localizzazione degli zeri di un, 171, 228
  osculatore, 366-370, 451
  osculatore di Hermite, 367-370, 389, 452, 680, 811
  radice quadrata di un, 245
  reciproco di un, 244, 489
  trigonometrico, 417, 472, 480
  trigonometrico, convergenza del, 425, 483
polinomio di interpolazione
  complessità del, 356, 374
  errori di arrotondamento del, 379
post-normalizzazione, 46, 48, 99
potenza fattoriale, 258, 299, 301
potenze, metodo delle, 189
PRAM, 18
precisione
  di macchina, 42
  doppia, 12, 38
  grado di, 721, 726, 745, 768, 788
  multipla, 12
  semplice, 12, 38
  variabile, 82
precondizionamento, 16
probabilistico, modello, 20
probabilità, densità di, 77
problema
  intrinsecamente complesso, 17
  mal condizionato, 10, 52, 348
processo di discretizzazione, 8
processori, 18
prodotto
  di interi, 24
  di matrici, 22
  di n numeri, errore nel, 90
  di numeri complessi, 17
  di polinomi, 23, 488
  formule, 779-783, 835
  scalare, 21, 514, 516, 613
pseudocasuali, numeri, 783, 840
punti
  di Chebyshev, 365, 382
  di equioscillazione, 553, 586

di mezzo, formula dei, 751, 804
equidistanti, 357, 361, 375, 728
punto fisso, 106
π, calcolo di, 20, 253, 272, 688

qd, metodo, 192-197, 606-610
quadratica, spline, 499
quadrato sommabile, funzione a, 517
quadratura
automatica, 773-778
formule di, 720-772
quasi minimax, approssimazione, 570-578, 613

Radau, formula di, 769, 823
radice n-esima, 129
dell'unità, 417, 488
primitiva dell'unità, 417
radice quadrata, 149, 211, 212
di un polinomio, 245
radici
opposte, 225
reciproche, 225
traslate, 225
rappresentazione
cifre di una, 33
errore di, 42, 57
in base, 30, 33
in modulo e segno, 37
in traslazione, 37
in virgola mobile, 42
lunghezza di una, 33
normalizzata, 33
per arrotondamento, errore statistico nella, 97
per troncamento, errore statistico nella, 97
razionale,
approssimazione minimax, 585-593, 683, 703, 706, 711
interpolazione, 392, 395, 406, 465
razionali di un polinomio, zeri, 226
reciproca, differenza, 410
reciproco
calcolo del, 198
di un polinomio, 244, 489
regola
dei segni di Cartesio, 234
di Budan-Fourier, 234
di Leibniz, 303
regula falsi, 137
relazione ricorrente a tre termini, 523, 548
relazioni di Girard-Newton, 224
Remez, algoritmo di, 563, 574, 581, 585, 588, 666, 702, 709
convergenza, 567, 588, 669, 672
residuo di una frazione continua, 396
resto
dell'approssimazione, 512, 552, 557, 634
dell'interpolazione osculatoria, 368
dell'interpolazione polinomiale, 358, 361, 375, 448
di una formula di derivazione approssimata, 789, 845
di una formula di quadratura, 721, 729, 745, 753, 803
standard, 559, 567, 672
retta di Chebyshev, 695
rettangoli, formula dei, 795
Richardson, estrapolazione di, 741, 760, 818, 845,
riduzione della varianza, tecnica di, 786, 841
Rodrigues, formula di, 526, 530, 531, 534, 537, 539
Romberg, schema di, 741, 774, 828, 845
Ruffini-Horner, metodo di, 16, 22, 64, 93, 222, 237, 260, 377, 547, 550
rumore, 432
Runge, funzione di, 363, 578

S-frazione, 402
Sande e Tukey, algoritmo di, 430, 478
scalare, prodotto, 21, 514, 516, 613
scambio, algoritmo di, 618, 696
schema
di Patterson, 773
di Romberg, 741, 774, 828, 845
scostamento medio, 628
secanti, metodo delle, 135-142
condizioni sufficienti di convergenza, 136
variante del, 139, 338
segnale, 432
segno, cambiamenti di, 176
Seidel, teorema di, 594, 684
seni, trasformata di, 481
separazione degli zeri, proprietà di, 635
sequenziale, calcolo, 18
serie
di Chebyshev, espansione in, 552

di Fourier, 270, 316, 483, 545, 657
sezione aurea, 141, 337, 687
significatività, analisi di, 80
$\sin x$, calcolo di, 701
simmetria delle differenze
divise, 375, 461
reciproche, 411
simmetriche elementari, funzioni, 224, 492
singolarità, sottrazione della, 759, 817
sistema
fondamentale di soluzioni, 279
lineare triangolare, 73
mal condizionamento di un, 10
normale, 515, 614, 627
sistemi non lineari, metodi iterativi per, 150-165
software, 14
soluzione
minimale, 286, 289, 347
particolare, 277, 290, 334
soluzioni
linearmente indipendenti, 276, 278, 329, 333
sistema fondamentale di, 279
somma definita, 262, 307
somma, operatore, 261, 304
sostituzione, metodo di, 73
sottospazio di polinomi, distanza da un, 510, 679
sottrazione della singolarità, 759, 817
spazio di Hilbert, 514
spline, 434
completa, 437, 444, 495
cubica, 436-446, 742
di ordine $2m-1$, 497
naturale, 437, 441, 493, 495
periodica, 437
quadratica, 499
costo computazionale, 501
stabilità
delle formule ricorrenti, 286, 346
di un algoritmo, 7, 52
standard, resto, 559, 567, 672
statistica, analisi dell'errore, 75-79, 96
Steffensen, metodo di, 144, 215
Stirling
formula di, 325, 454, 650
numeri di, 259, 301
polinomio di, 457
Strassen, algoritmo di, 14
strategia adattiva, 774, 777, 831
strettamente convesso, insieme, 626
Sturm, successione di, 175, 235
sublineare, convergenza, 118, 132
successione
alternata, 109
di Sturm, 175, 235
monotona, 109, 136
superlineare, convergenza, 118, 140

tabella di Padé, 601, 603, 690, 691, 693
$\tan x$, calcolo di, 703
tangenti, metodo delle, 125-134, 700
condizioni sufficienti di convergenza, 134
ordine del, 126,132, 210
tecnica
del passo doppio, 181
di riduzione della varianza, 786, 841
teorema
centrale di convergenza, 78
de la Vallée-Poussin, 553, 660
di Chebyshev, 554, 579, 584, 586, 660, 617
di Gentleman e Sande, 431
di Jackson, 562
di Ostrowski, 174
di Seidel, 594, 684
di Weierstrass, 510, 622, 680
Thiele, frazione continua di, 406, 415, 470, 599
Toeplitz, matrice di, 24, 25, 599
totalmente differenziabile, funzione, 153
trapezi, formula dei, 484, 496, 547, 737, 798, 835
trasformata di coseni, 481, 547, 834
trasformata di seni, 481
trasformata discreta di Fourier, 419, 426-431, 473
trasformata discreta inversa di Fourier, 419, 426-431, 475
trasformazione
a radici opposte, 225
a radici reciproche, 225
a radici traslate, 225
di Abel, 308
di Eulero, 271, 319, 819
traslazione
operatore, 256, 298
rappresentazione in, 37
tre punti, formula dei, 739
triangolare, disuguaglianza, 512

tridiagonale, matrice, 276, 283, 332, 340, 439, 467, 643, 807
trigonometrica
  approssimazione, 657
  interpolazione, 417-433
trigonometrico, polinomio, 417, 472
  complessità del calcolo di un, 480
troncamento, 39

uguaglianza di Parseval, 516, 541
ultrasferici, polinomi, 527, 529
un punto, metodo ad, 143
underflow, 2, 39, 41, 44, 49, 88
uniformi, formule a coefficienti, 727, 813

Vandermonde, matrice di, 353, 418, 447, 492, 614, 618, 722, 788
variabile casuale normale, 77
variazione dei parametri, metodo della, 284, 343
vincoli, approssimazione minimax con, 583, 682, 703
virgola mobile, aritmetica in, 45
Viskovatov, metodo di, 401, 596, 687

Wallis, formula di, 325
Weierstrass, teorema di, 510, 622, 680
Winograd, algoritmo di
  per la FFT, 431
  per il prodotto scalare, 21
Woodbury, formula di, 440

$\sqrt{x}$, calcolo di, 700
x^n, calcolo di, 22, 92
$|x|$, approssimazione di, 377, 542, 545, 624, 649, 653, 663

zeri
  complessi, 180, 184
  di polinomi ortogonali, 522, 642, 724
  razionali di un polinomio, 226
  proprietà di separazione degli, 635
zeri di un polinomio
  condizionamento degli, 166
  localizzazione degli, 171, 228

Gli autori

Roberto Bevilacqua, Dario Bini, Milvio Capovani e Ornella Menchi, docenti di discipline di Matematica Numerica presso i Corsi di Laurea in Matematica e Scienze dell'Informazione della Facoltà di Scienze Matematiche, Fisiche e Naturali dell'Università di Pisa, hanno svolto e stanno svolgendo ricerche in vari settori della Matematica Computazionale e, in paticolare, nel settore dell'Analisi e della sintesi di algoritmi numerici sequenziali e paralleli. Sono autori di "Introduzione alla matematica computazionale" (Zanichelli, 1987). Dario Bini, Milvio Capovani e Ornella Menchi sono autori di "Metodi numerici per l'algebra lineare" (Zanichelli, 1988). Dario Bini e Milvio Capovani sono autori, con Grazia Lotti e Francesco Romani, di "Complessità numerica" (Boringhieri, 1981).

L'opera

La matematica nella società moderna ha assunto una crescente importanza: ormai si può affermare che il livello di cultura matematica è una misura del progresso scientifico e tecnologico di un paese. Nell'analisi dei problemi del mondo reale la matematica svolge un ruolo determinante. In ogni disciplina scientifica e in ogni settore della tecnologia i modelli matematici che approssimano l'evolversi dell'evento oggetto di studio consentono di simulare, e quindi prevedere, lo sviluppo del fenomeno senza dover effettuare fisicamente esperimenti complessi, costosi e in alcuni casi anche pericolosi.
In generale le soluzioni dei modelli matematici non sono esprimibili in forma esplicita. Si presenta quindi la necessità di risolvere algoritmicamente il problema matematico, cioè di ottenere mediante un numero finito di operazioni aritmetiche e/o logiche un'adeguata approssimazione della soluzione. L'analisi numerica è la disciplina che sviluppa e studia tutti quegli strumenti matematici atti ad individuare e analizzare questi metodi di risoluzione numerica.
In questo libro sono esposti e analizzati i principali metodi dell'analisi numerica. I metodi descritti sono accompagnati da esempi, direttamente sperimentati al calcolatore, con la presentazione di tabelle e grafici, e da numerosi esercizi. È riportata anche un'ampia bibliografia e ogni capitolo è chiuso da note bibliografiche e da cenni storici, che possono guidare i lettori interessati in indagini più approfondite.
Il testo è rivolto principalmente agli studenti dei corsi di laurea in Matematica, Fisica, Ingegneria e Scienze dell'Informazione, ed ai ricercatori che operano nel settore del calcolo scientifico.