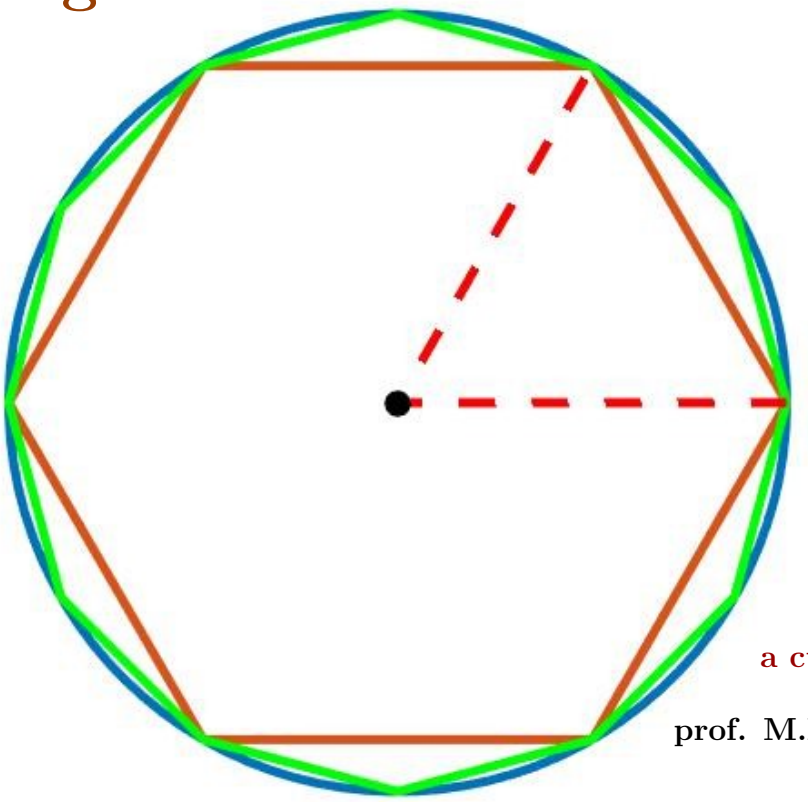
Corso di Laurea in Matematica

a.a. 2021-2022

Dispense del corso di

Laboratorio di Programmazione e Calcolo



a cura di

prof. M.I. Gualtieri

I matematici sono come i francesi:
se dici loro qualche cosa
la traducono nella propria lingua e
diventa subito qualcosa di diverso!

J.W. Goethe (1749-1832)

Questi appunti non hanno pretesa di sistematicità, di rigore e di completezza e non sostituiscono lo studio dei manuali. Sono solo alcune note che affiancano il corso di "Laboratorio di Programmazione e Calcolo", con l'auspicio comunque che la loro lettura possa facilitare la comprensione delle lezioni svolte.
Tali appunti, inoltre, non sono stati ancora sufficientemente riveduti e corretti e, pertanto, possono contenere errori e inesattezze.

Indice

Capitolo 1

Modelli

Una disciplina ha tanto più la dignità della scienza quanto più fa uso dello strumento matematico
Galileo Galilei (1564-1642)

Dare una definizione esauriente e semplice del concetto di modello è difficile in quanto a questo termine vengono spesso attribuiti significati diversi. In generale, un **modello** è *uno schema elaborato per rappresentare gli elementi fondamentali di fenomeni o enti e descrivere quindi quanto osservato.*
Molto noti sono i modelli in scala ridotta, che riproducono qualitativamente un sistema pur riducendone proporzionalmente la dimensione.
Ci soffermiamo sui modelli teorici che, in base all'uso, vengono distinti in

1. **modelli descrittivi** o **statici**: riproducono con eventuali semplificazioni la realtà, sintetizzando in un meccanismo o in un algoritmo i dati osservati relativi ad un fenomeno senza presupporre l'uso che ne verrà fatto e quindi senza tentare di spiegare il meccanismo su cui il fenomeno osservato è basato;
2. **modelli interpretativi**: cercano di spiegare il comportamento di un fenomeno e la sua evoluzione formulando ipotesi ricorrendo a leggi generali; quindi vengono ipotizzate le strutture interne che giustifichino il comportamento esterno e le conseguenza logiche di tali ipotesi;
3. **modelli predittivi**: si propongono di prevedere l'andamento futuro di un fenomeno, almeno entro un dato orizzonte temporale, lasciando spazio ad eventuali scelte.

1.1 Modelli simbolici

Tra tutti i tipi di modelli predittivi, utili nello sviluppo della conoscenza scientifica, un ruolo fondamentale spetta sicuramente alla famiglia dei modelli **simbolici** o **matematici**, che danno una rappresentazione astratta della realtà cui si riferiscono, mediante un insieme di equazioni e/o disequazioni che legano le grandezze coinvolte.
I modelli matematici costituiscono quindi un sottoinsieme proprio dei modelli teorici, caratterizzato dal fatto che le proprietà e le relazioni di una teoria sono espresse nel linguaggio, e seguono la logica, della matematica. Mentre ci sono molte teorie che non si possono tradurre in modelli matematici, al contrario tutti i modelli matematici sono anche, o meglio prima di tutto, modelli teorici.
L'economista francese Edmond Malinvaud (Limoges, 1923 – Parigi, 2015) nel libro *"Méthodes statistiques de l'économetrie"* (Parigi, 1964) dà la seguente definizione di modello matematico

*Un **modello matematico** è la rappresentazione formale di idee o conoscenze relative a un fenomeno.*

Questa definizione contiene una descrizione completa delle caratteristiche fondamentali di un modello matematico

a) un modello matematico è la rappresentazione di un fenomeno. Non si tratta però di una semplice descrizione verbale, ma di una descrizione che mette in luce determinati aspetti caratteristici di un fenomeno;

b) tale rappresentazione non è discorsiva o a parole, ma formale, ossia espressa in linguaggio matematico, il linguaggio formale e astratto per eccellenza;

c) non esiste una via diretta dalla realtà alla sua descrizione matematica, in modo univoco; ciò che invece si fa è di tradurre in formule idee e conoscenze relative al fenomeno.

Malinvaud definisce un modello matematico come la rappresentazione formale *di idee o conoscenze relative a un fenomeno* e non dice semplicemente che è la rappresentazione formale di un fenomeno.
In primo luogo perché la realtà è talmente complessa che è difficile farne una descrizione relativamente semplice e schematica qual è quella matematica. D'altra parte, una descrizione della realtà perfettamente aderente ad essa sarebbe non soltanto impossibile, perché troppo complicata, ma anche inutile. Per descrivere un fenomeno dobbiamo quindi operare delle scelte, selezionare alcuni aspetti del fenomeno osservato e trascurarne altri, che riteniamo secondari, e questa scelta richiede di mettere in campo le nostre idee. Per formarci un'idea e costruire quindi un'ipotesi circa la legge che dovrebbe governare il processo in esame, un ruolo fondamentale spetta all'osservazione empirica e all'esperimento. Quel che noi facciamo quindi è mettere insieme tutte le conoscenze empiriche e le idee che ci siamo fatti del fenomeno, per costruirne una rappresentazione matematica, che si tradurrà per lo più in formule o equazioni di vario tipo. Abbiamo così ottenuto un *modello matematico* del processo reale in oggetto.
Migliorare un modello significa renderlo sempre più simile alla realtà ma, se un modello si complica oltre un certo limite, perde la sua capacità esplicativa. È il cosiddetto *paradosso della modellizzazione*: se è vero che un modello è - nel senso che *deve essere* - diverso dalla realtà, è però altrettanto vero che un modello è tanto più valido, ossia adeguato a descrivere il fenomeno studiato e a prevedere certi effetti, quanto più numerosi sono gli elementi della sua struttura che sono invece aderenti alla realtà modellizzata.
Dal momento che lo stesso fenomeno osservato può essere modellizzato in una infinità di modi diversi, a seconda delle proprietà e delle relazioni che, di volta in volta, si astraggono dalla realtà, un modello non può essere valutato in base ad un criterio di verità, cioè di una sua corrispondenza o perfetta sovrapponibilità alla realtà modellizzata, visto che tutte le idee e le conoscenze che noi abbiamo utilizzato per costruirlo possono aver introdotto delle discrepanze dalla realtà più o meno accettabili. È indispensabile allora la fase di verifica. Per verificare il modello occorrerà dedurre dalla sua struttura matematica la previsione che esso fornisce circa il comportamento del fenomeno; poi confrontare queste previsioni con i dati reali.

1.2 Un po' di storia

Uno dei primi tentativi di applicazione della matematica ad un problema "umano" risale al Settecento, in occasione dell'inoculazione del vaccino del vaiolo. Rispetto all'opportunità di operare tale inoculazione, si crearono due fronti opposti. I sostenitori del vaccino si dichiaravano favorevoli al progresso della scienza al di là dei rischi che questo poteva comportare (come è noto, l'inoculazione del siero attivo provoca una leggera forma della stessa malattia), mentre i contrari sostenevano che inoculare il vaccino rappresentava un'azione contro natura.
Nel 1760 il matematico svizzero *Daniel Bernoulli* (Groninga, 1700 – Basilea, 1782) cercò di dimostrare matematicamente che la vaccinazione era necessaria, calcolando i vantaggi (in termini di vite salvate) dell'essere vaccinati rispetto alla probabilità di morire nel caso non si venisse vaccinati. Uno dei primi a sollevare obiezioni fu *Jean-Baptiste d'Alembert* (Parigi, 1717 – Parigi, 1783 filosofo illuminista) che dedicò a questo argomento addirittura quattro Memorie. D'Alembert sosteneva che i fenomeni umani

sono troppo variabili per essere ridotti a formule ed equazioni. Si tenga conto anche del fatto che i calcoli fatti da Bernoulli erano per lo più di carattere statistico e probabilistico e, a quel tempo, queste due discipline non erano ancora considerate delle vere e proprie scienze.
Ma per avere dei modelli deterministici (anziché probabilistici alla Bernoulli) bisogna aspettare *Thomas Malthus*[1] e *Pierre François Verhulst*[2]. Entrambi si occuparono di dinamica delle popolazioni (all'epoca, Malthus e Verhulst ragionavano in termini di popolazioni umane), che sarebbe diventato in seguito uno dei più importanti campi di applicazione della matematica alla biologia. Malthus arrivò a determinare nel 1798 la famosa legge che da lui prese il nome, ma che in seguito si rivelò poco realistica, e questo portò Verhulst nel 1845 a trovare un nuovo più sofisticato modello oggi noto come equazione logistica.
A parte gli studi di tipo statistico, questi furono gli unici due modelli matematici applicati alla biologia di tutto il diciannovesimo secolo. Sarà il matematico *Vito Volterra* (Ancona, 1860 – Roma, 1940), all'inizio del ventesimo secolo, a spingere perché la biologia accettasse al suo interno l'uso di strumenti di tipo quantitativo. Già nel 1900, in occasione dell'inaugurazione dell'anno accademico dell'Università di Roma, Volterra aveva esordito con una conferenza dal titolo *Sui tentativi di applicazione delle matematiche alle scienze biologiche e sociali*, ma solo nel 1925 entrò nel vivo della questione creando un modello matematico che da allora divenne indispensabile per le scienze biologiche.
A partire dagli anni '20 si assiste ad un'improvvisa accelerata della matematizzazione della scienza che, a partire dalla fisica, ha interessato la chimica, la biologia e anche le cosiddette scienze umane, dalla logica alla sociologia, dall'economia alla scienza dell'informazione. Il ventesimo secolo è, dal punto di vista scientifico, il secolo dei modelli.

1.3 Modellistica matematica

Col termine *modellistica matematica* si intende il processo che si sviluppa attraverso l'interpretazione di un problema reale, la rappresentazione dello stesso problema mediante il linguaggio della matematica, l'analisi di tali equazioni, nonché l'individuazione di metodi idonei a risolverle, ed infine, eventualmente, l'implementazione di tali metodi su calcolatore, dopo averli tradotti in opportuni algoritmi.

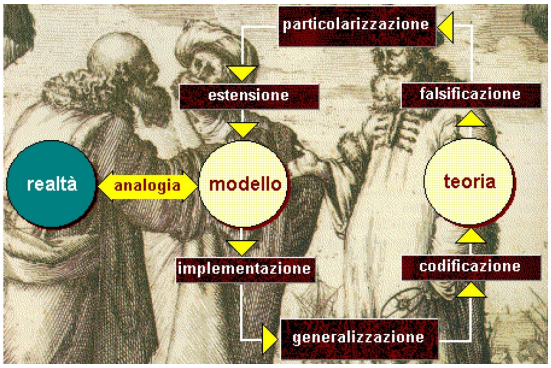


Figura 1.1

I problemi matematici creati nell'ambito della modellistica non sempre sono risolubili per via analitica. I teoremi dell'analisi matematica e della geometria, se pur fondamentali per stabilire l'esistenza e l'unicità della soluzione, spesso non hanno natura costruttiva atta ad indicare un processo di rappresentazione esplicita della soluzione. In tali casi è necessario sviluppare metodologie di approssimazione che conducono ad algoritmi tali da rendere possibile la risoluzione su calcolatore. È anche grazie all'uso di strumenti di calcolo elettronici e l'impiego di software capaci di fornire risposte adeguate in tempi brevi che si è sviluppata la modellizzazione matematica.

[1] Thomas Malthus (The Rookery, 1766 - Bath, 1834) economista e demografo inglese, studioso di problemi sociali.
[2] Pierre François Verhulst (Bruxelles, 1804 - Bruxelles, 1849) matematico belga, considerato uno dei fondatori della moderna statistica.

Prima di proseguire occorre definire alcuni termini.

Col termine ***popolazione*** si indica un insieme di elementi (gli individui) il cui numero (la dimensione) può cambiare nel corso del tempo.

La ***dinamica delle popolazioni*** studia in genere come può cambiare una popolazione al variare del tempo, in termini di età, grandezza e genotipo e quali fattori esterni o interni alla popolazione ne regolano la struttura.

I modelli utilizzati per descrivere l'accrescimento di una popolazione sono suddivisi in due classi

- **modelli stocastici**, per i quali i meccanismi di evoluzione sono influenzati da variazioni dovute al caso
- **modelli deterministici**, ove non interviene il caso.

I modelli deterministici sono basati sull'ipotesi che la popolazione evolva in maniera tale che il suo sviluppo futuro possa essere previsto *esattamente*, una volta specificato lo stato ad un tempo iniziale fissato. Tali modelli non permettono fluttuazioni aleatorie; ossia, i fattori che determinano eventi particolari come nascite e morti sono sufficientemente costanti da considerarsi certi.

Al contrario, i modelli stocastici sono basati sull'ipotesi che l'aumento di una popolazione sia un evento casuale o aleatorio; ossia, un organismo può riprodursi o morire durante un determinato periodo di tempo con una certa probabilità. Nessun evento può accadere con assoluta certezza.

Si può intuire che un modello stocastico è molto più ricco di un modello deterministico, nel senso che, anziché descrivere un comportamento "medio", permette di considerare molti o tutti i casi, anche quelli la cui probabilità è piccola. Naturalmente, la maggiore ricchezza di informazioni di un modello stocastico nei confronti di uno deterministico comporta spesso una maggiore difficoltà di studio e di risoluzione.

La differenza tra i due tipi di modelli è illustrata dal seguente esempio.

Esempio 1.1 *Supponiamo che due persone, indicate rispettivamente con A e B, il primo giorno di ogni mese acquistino alcuni libri. In particolare, A compra un libro al mese, mentre B prima di decidere se acquistare i libri lancia una moneta: se esce testa compra due libri, altrimenti in quel mese non compra libri. Se si suppone che all'inizio dell'osservazione A e B non posseggano alcun libro, si vuole determinare le dimensioni delle rispettive biblioteche dopo n mesi.*
In questo caso la biblioteca costituisce una popolazione i cui individui sono i libri.

Per quanto riguarda A, se si indica con $y(n)$ la dimensione della popolazione all'n-esimo mese, dal momento che ad ogni mese la biblioteca aumenta di un libro, si ha la seguente equazione

$$y(n+1) = y(n) + 1, \qquad y(0) = 0.$$

Si tratta di un'equazione alle differenze lineare e del primo ordine (con la condizione iniziale $y(0) = 0$), la cui soluzione è ovviamente $y(n) = n$, per ogni n intero.

Al contrario, non è possibile determinare esattamente la dimensione della biblioteca di B all'n-esimo mese, in quanto tale numero, che indicheremo con $x(n)$, dipende dai risultati, imprevedibili a priori, degli n lanci della moneta. Infatti, $x(n)$ può assumere un valore compreso tra 0, se esce sempre croce, e $2n$ se esce sempre testa ($x(n)$ è una variabile casuale, che dipende dalla probabilità di ottenere testa nel lancio di una moneta).

1.3.1 Modelli lineari

Si chiama lineare un modello rappresentato da una funzione lineare.

Esempio 1.2 ***Quoziente intellettivo***
Il Q.I. indica l'indice di sviluppo mentale individuale. L'età mentale viene stabilita sottoponendo l'individuo ad una serie di test di difficoltà via via crescente, ciascuno dei quali è risolubile (statisticamente) dalla maggior parte degli individui che hanno la stessa età cronologica.
Se y rappresenta il valore del quoziente intellettivo (Q.I.) di un individuo fino all'età adulta, e_c l'età cronologica, x l'età mentale e si fissa a 100 *il Q.I. di un individuo con sviluppo mentale normale, il Q.I. soddisfa la proporzione*

$$y : 100 = x : e_c$$

da cui la relazione

$$y = ax, \qquad a = \frac{100}{e_c}.$$

Il Q.I. è quindi regolato da un modello ***lineare****.*

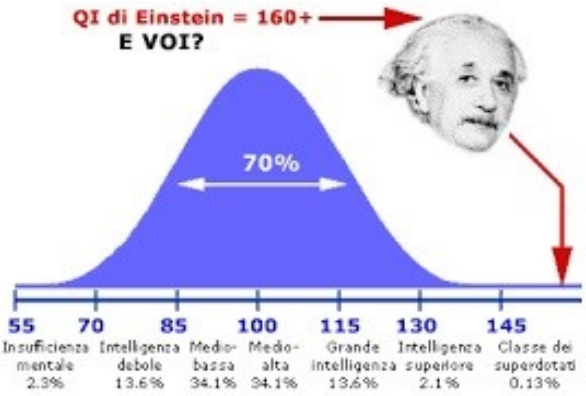


Figura 1.2

Se il Q.I. è uguale a 100 *lo sviluppo mentale dell'individuo è normale; se è superiore o inferiore a* 100*, l'individuo è rispettivamente più o meno sviluppato intellettivamente.*

Esempio 1.3 ***Scale termometriche*** *Dalle relazioni che permettono la conversione delle scale termometriche si può osservare che il modello di conversione è sempre di tipo lineare*

$$y = ax + b,$$

*infatti se consideriamo le quattro scale termometriche centigrada (o Celsius) (**C**), Rèamur (**R**), assoluta o Kelvin (**K**) e Fahrenheit (**F**), si ha*

$\boldsymbol{C} \longleftrightarrow \boldsymbol{R}$	$^\circ C : {}^\circ R = 100 : 180,$	$^\circ C = \frac{100}{180},$	$a = \frac{100}{180},$	$b = 0;$
$\boldsymbol{C} \longleftrightarrow \boldsymbol{K}$	$^\circ K = {}^\circ C + 273.15,$	$^\circ K = {}^\circ C + 273.15,$	$a = 1,$	$b = 273.15;$
$\boldsymbol{C} \longleftrightarrow \boldsymbol{F}$	$^\circ C : (^\circ F - 32) = 100 : 180,$	$^\circ C = \frac{100}{180}{}^\circ F - \frac{3200}{180},$	$a = \frac{100}{180},$	$b = -\frac{3200}{180};$
$\boldsymbol{R} \longleftrightarrow \boldsymbol{F}$	$^\circ R : (^\circ F - 32) = 80 : 180,$	$^\circ R = \frac{80}{180}{}^\circ F - \frac{32 \cdot 80}{180},$	$a = \frac{80}{180},$	$b = -\frac{32 \cdot 80}{180}.$

Esempio 1.4 ***Una situazione problematica (Sistema lineare)***
Un gruppo di amici, volendosi recare a Roma, deve stabilire qual è il mezzo di trasporto più conveniente. Valuta due alternative

a) andare in treno spendendo €35 *a persona;*

b) noleggiare un pullman pagando una quota fissa di €420 *più* €7 *per ogni persona trasportata.*

Naturalmente la convenienza dell'una o dell'altra soluzione dipenderà dal numero dei partecipanti. Indicando con x il numero dei partecipanti e con y il costo del trasporto, le due soluzioni sono

$$\textit{a) (viaggio in treno): } y = 35x \qquad \textit{b) (viaggio in pullman): } y = 7x + 420.$$

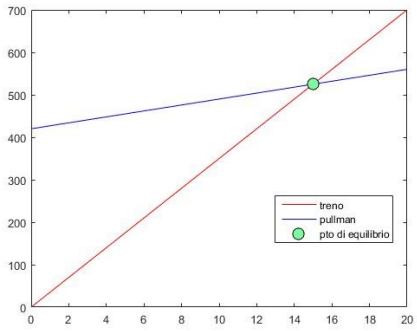


Figura 1.3

Le due rette si intersecano in un punto di ascissa 15. *Si può dedurre che fino a* 15 *partecipanti è conveniente la soluzione a), mentre per un numero maggiore è meno dispendiosa la scelta b).*

1.3.2 Modelli quadratici

Si chiama **quadratico** un modello rappresentato da una funzione quadratica. Negli esempi seguenti si usano funzioni quadratiche per modellizzare problemi reali.

Esempio 1.5 ***Indice di massa corporea***
Se p è il peso di una persona e s la statura, esiste una relazione di questo tipo

$$p = BMI\, s^2 \tag{1.1}$$

dove BMI è l'indice di massa corporea. Dalla (1.1) è evidente che c'è una relazione ***quadratica*** *tra il peso di una persona e la sua statura.*

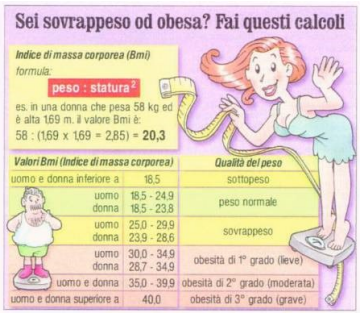


Figura 1.4

Dalla tabella in figura si può osservare per quali valori del BMI un uomo o una donna sono in sottopeso, hanno un peso normale o sono in sovrappeso.

1.3.3 Modelli esponenziali e logaritmici

La funzione **esponenziale**, in particolare di base e, fornisce numerosi modelli matematici che descrivono alcuni importanti fenomeni di natura fisica, economica, biologica, ecc.

Esempio 1.6 ***Decadimento radioattivo***
Le sostanze radioattive sono composti chimici costituiti di atomi che si decompongono spontaneamente in altri atomi non radioattivi. Il fenomeno del decadimento radioattivo è di tipo esponenziale ed è regolato da una legge descritta dall'equazione

$$m_t = m_0 e^{-\lambda t}$$

dove m_t è la massa della sostanza radioattiva al tempo t, m_0 è la massa della sostanza radioattiva presente all'inizio dell'esperimento, cioè al tempo $t = 0$ e λ è una costante detta "costante di decadimento radioattivo" il cui valore è un numero caratteristico di ciascuna sostanza radioattiva e dà la misura della maggiore o minore rapidità con cui avviene il processo di trasformazione.
Più è grande il valore di λ e maggiore è il numero degli atomi radioattivi che si trasformano in atomi non radioattivi nell'unità di tempo e quindi più rapido è il processo di decadimento. Il segno negativo che compare davanti all'esponente suggerisce che la legge di decadimento radioattivo è una legge di tipo esponenziale decrescente, cioè una legge la quale mostra che con il passare del tempo gli elementi presenti all'inizio diminuiscono e non aumentano di numero.

La funzione **logaritmo**, in particolare il logaritmo di base e, si applica in alcuni importanti fenomeni di natura fisica, come la misura dell'intensità del suono o dell'intensità di un terremoto, il calcolo del PH e la misura della magnitudo di una stella.

Esempio 1.7 ***Magnitudo di un terremoto***
Per dare alla classificazione dei terremoti una valenza scientifica fu indispensabile trovare un sistema per misurare l'energia che si libera al momento dell'evento sismico. Allo scopo vennero sistemati, in diversi punti della superficie terrestre, strumenti adeguati in grado di registrare il fenomeno. In seguito a queste misurazioni nacque la cosiddetta "SCALA DELLE MAGNITUDO" ideata dal geofisico americano Charles Francis Richter nel 1935.
Il valore della magnitudo di un terremoto si determina confrontando l'ampiezza delle oscillazioni registrate dal sismografo e quella prodotta, sullo stesso strumento, da un terremoto campione. Oggi, però, ogni stazione sismica è in possesso di una tabella con i valori del terremoto campione già determinati in relazione a diverse distanze, al tipo di terreno e al sismografo operante per semplificare così l'operazione. Per i terremoti a $100km$ di distanza la formula è

$$M_L = \log A$$

dove M_L è appunto la magnitudo Richter, o magnitudo locale, ed A è l'altezza massima della sinusoide da 0 fino al picco in mm.
Poiché l'ampiezza massima registrata sul sismogramma di un forte sisma può essere anche milioni di volte maggiore di quella relativa ad un terremoto debole, al fine di evitare numeri di magnitudo troppo grandi, Richter ritenne opportuno ricorrere al logaritmo in base 10 del rapporto fra l'ampiezza massima A del terremoto, misurata in micrometri, e l'ampiezza A_0 che verrebbe prodotta dal terremoto standard alla stessa distanza epicentrale

$$M = \log \frac{A}{A_0}.$$

La magnitudo di un terremoto è quindi definita come la misura logaritmica dell'energia liberata.

Capitolo 2

Modelli discreti

Finché le leggi della matematica si riferiscono alla realtà, non sono certe, e finché sono certe, non si riferiscono alla realtà

Albert Einstein (1879-1955)

Nel capitolo 1 i modelli matematici sono stati distinti in deterministici e stocastici. Entrambi i tipi di modello possono essere distinti in modelli **a tempo discreto** e **a tempo continuo**.

Un modello *a tempo discreto* è caratterizzato dal fatto che tutte le variabili utilizzate sono funzioni della variabile indipendente temporale che assume solo valori interi. In questo caso i cambiamenti di stato si osservano soltanto a determinati istanti, ad intervalli di tempo regolari (ogni anno, ogni giorno, ogni ora,...), quindi, se T è la variabile temporale, l'insieme dei tempi nei quali si vuole studiare il fenomeno è una successione (finita o infinita) $t_0, t_1, \ldots$. Con un opportuno cambio di scala temporale si può supporre che T sia l'insieme $\mathbb{N} \cup \{0\}$, o un suo sottoinsieme, ossia $t_0 = 0$, $t_1 = 1, \ldots, t_n = n, \ldots$.

Se invece i cambiamenti di stato avvengono in un qualsiasi istante in un intervallo reale (ossia T è la semiretta $[0, +\infty)$ o un intervallo $[a, b] \subset T$), allora il modello è detto *a tempo continuo*.

2.1 Accrescimento di una popolazione

Se y_k indica la dimensione di una popolazione all'istante t_k, e y_{k+1} la dimensione all'istante successivo t_{k+1}, la differenza $\Delta y_k = y_{k+1} - y_k$ rappresenta l'*accrescimento* (positivo o negativo) della popolazione tra gli istanti t_k e t_{k+1}. L'accrescimento può essere imputato alle seguenti cause indipendenti tra loro

- la *nascita* di individui all'interno di una popolazione tra gli istanti t_k e t_{k+1};
- la *morte* di individui appartenenti alla popolazione tra gli istanti t_k e t_{k+1};
- l'*immigrazione* di nuovi individui che entrano a far parte della popolazione tra gli istanti t_k e t_{k+1};
- l'*emigrazione* di individui che abbandonano la popolazione tra gli istanti t_k e t_{k+1}.

Se indichiamo con $n(k)$, $d(k)$, $i(k)$, $e(k)$ rispettivamente il numero di individui che nascono, muoiono, immigrano ed emigrano, l'accrescimento della popolazione può essere scritto nella forma

$$\Delta y_k = y_{k+1} - y_k = \Big[n(k) - d(k)\Big] + \Big[i(k) - e(k)\Big]. \tag{2.1}$$

Se definiamo

$$n^*(k,y) = \frac{n(k)}{y_k} \quad \text{tasso di natalità,} \qquad d^*(k,y) = \frac{d(k)}{y_k} \quad \text{tasso di mortalità,}$$

$$i^*(k,y) = \frac{i(k)}{y_k} \quad \text{tasso di immigrazione,} \qquad e^*(k,y) = \frac{e(k)}{y_k} \quad \text{tasso di emigrazione,}$$

$$r(k,y) = \Big[n^*(k,y) - d^*(k,y)\Big] + \Big[i^*(k,y) - e^*(k,y)\Big] \quad \text{tasso di accrescimento,}$$

l'equazione di accrescimento (2.1) assume la forma $y_{k+1} = y_k + y_k r(k,y)$, cioè

$$y_{k+1} = y_k\Big[1 + r(k,y)\Big]. \tag{2.2}$$

Infatti, dalla (2.1), moltiplicando e dividendo il secondo membro per y_k, si ha

$$y_{k+1} - y_k = y_k \frac{n(k) - d(k) + i(k) - e(k)}{y_k} = y_k\, r(k,y) \qquad \Longrightarrow \qquad y_{k+1} = y_k + y_k\, r(k,y).$$

La (2.2) è un'*equazione alle differenze del primo ordine.*
Nota la funzione $r(k,y)$, la (2.2) permette di determinare, a partire dal valore iniziale y_0, la dimensione della popolazione ad ogni istante t intero.

2.2 Equazioni alle differenze

Definizione 2.1. *Un'**equazione alle differenze** di ordine k è un'espressione del tipo*

$$f(n, y_n, y_{n+1}, \ldots, y_{n+k}) = 0, \qquad f : \mathbb{N} \times \mathbb{R}^{k+1} \to \mathbb{R}. \tag{2.3}$$

Definizione 2.2. *Si dice* **ordine** *di un'equazione alle differenze la differenza tra l'indice più grande e quello più piccolo delle variabili che compaiono nell'equazione. Ad esempio*

$y_{n+2} - 2y_n = 0$ *è un'equazione alle differenze del **secondo** ordine;*

$y_{n+1} = 5n y_{n-5}$ *è un'equazione alle differenze del **sesto** ordine.*

Definizione 2.3. *Una soluzione dell'equazione alle differenze* (2.3) *è una **successione** $\{y_n\}_{n\in\mathbb{N}}$ tale che ciascuna sequenza di $k+1$ termini consecuitivi soddisfa la* (2.3).

Non tutte le equazioni alle differenze ammettono soluzioni reali. Ad esempio l'equazione $y_n^2 + y_{n-1}^2 = -1$ non ammette alcuna soluzione reale.

Una classe di equazioni alle differenze che ammettono sempre soluzioni è data da quelle **lineari a coefficienti costanti**.

Definizione 2.4. *Una equazione alle differenze è* **lineare** *se ha la forma*

$$a_{n,0}y_n + a_{n,1}y_{n+1} + \cdots + a_{n,k}y_{n+k} = g_n \tag{2.4}$$

*ove $\{g_n\}_{n\in\mathbb{N}}$ e $\{a_{n,i}\}_{n\in\mathbb{N}}$, $i = 0, 1, \ldots, k$ sono successioni assegnate. La successione $\{g_n\}_{n\in\mathbb{N}}$ è detta **termine noto**.*

Definizione 2.5. *L'equazione* (2.4) *è detta* **omogenea** *se $g_n = 0$ per ogni n*

$$a_{n,0}y_n + a_{n,1}y_{n+1} + \cdots + a_{n,k}y_{n+k} = 0. \tag{2.5}$$

Definizione 2.6. *I termini delle successioni* $\{g_n\}_{n\in\mathbb{N}}$ *e* $\{a_{n,i}\}_{n\in\mathbb{N}}$, $i = 0, 1, \ldots, k$ *sono detti coefficienti dell'equazione; se non dipendono da* n*, si dice che l'equazione è* **a coefficienti costanti** *e assume la forma*

$$a_0 y_n + a_1 y_{n+1} + \cdots + a_k y_{n+k} = g.$$

Ad esempio l'equazione

- $y_{n+1} - 3y_n + 7y_{n-1} = 10$ è **lineare a coefficienti costanti**;
- $y_{n+1} - 3y_n + ny_{n-1} = 5 \cdot 2^n$ è **lineare a coefficienti non costanti**;
- $y_{n-1} + \sqrt{y_n + 2} = 1$ è **non lineare**.

Teorema 2.1 *Se* $\{z_n\}_{n\in\mathbb{N}}$ *e* $\{w_n\}_{n\in\mathbb{N}}$ *sono soluzioni dell'equazione omogenea* (2.5)*, ogni loro combinazione lineare* $\{\alpha z_n + \beta w_n\}_{n\in\mathbb{N}}$, $\alpha, \beta \in \mathbb{R}$*, è soluzione della* (2.5).

Dimostrazione. Se $\{z_n\}_{n\in\mathbb{N}}$ e $\{w_n\}_{n\in\mathbb{N}}$ soddisfano entrambe l'equazione (2.5), vale

$$a_0 z_n + a_1 z_{n+1} + \cdots + a_k z_{n+k} = 0$$

$$a_0 w_n + a_1 w_{n+1} + \cdots + a_k w z_{n+k} = 0.$$

Moltiplicando la prima per $\alpha \in \mathbb{R}$ e la seconda per $\beta \in \mathbb{R}$

$$a_0 \alpha z_n + a_1 \alpha z_{n+1} + \cdots + a_k \alpha z_{n+k} = 0$$

$$a_0 \beta w_n + a_1 \beta w_{n+1} + \cdots + a_k \beta w_{n+k} = 0,$$

cioè anche $\{\alpha z_n\}_{n\in\mathbb{N}}$ e $\{\beta w_n\}_{n\in\mathbb{N}}$ sono soluzioni della (2.5). Sommando membro a membro si ottiene

$$a_0 (\alpha z_n + \beta w_n) + a_1 (\alpha z_{n+1} + \beta w_{n+1}) + \cdots + a_k (\alpha z_{n+k} + \beta w_{n+k}) = 0$$

quindi anche la successione $\{\alpha z_n + \beta w_n\}_{n\in\mathbb{N}}$ soddisfa l'equazione (2.5).

2.3 Equazioni alle differenze del primo ordine

Data l'equazione lineare del primo ordine **omogenea** $a_1 y_{n+1} + a_0 y_n = 0$, senza perdere in generalità possiamo considerarla espressa nella forma

$$y_{n+1} - a y_n = 0, \qquad a = -\frac{a_0}{a_1}. \tag{2.6}$$

Una soluzione y_n di tale equazione si ottiene facilmente per ricorrenza

$$y_{n+1} = a y_n = a (a y_{n-1}) = \cdots = a^{n+1} y_0 \qquad \Longrightarrow \qquad y_n = a^n y_0.$$

Osserviamo che ogni successione del tipo

$$y_n = c\, a^n \qquad c \in \mathbb{R} \tag{2.7}$$

è soluzione della (2.6). La soluzione risulta univocamente determinata se è noto il suo primo termine y_0.

Consideriamo ora l'equazione **non omogenea**

$$a_1 y_{n+1} + a_0 y_n = \overline{b} \qquad a_1,\ a_0,\ \overline{b} \in \mathbb{R}.$$

Anche in questo possiamo considerare l'equazione espressa nella forma

$$y_{n+1} - a y_n = b, \qquad a = -\frac{a_0}{a_1},\ b = \frac{\overline{b}}{a_1}. \tag{2.8}$$

La *soluzione generale* di tale equazione si ottiene sommando alla (2.7) una soluzione particolare y_n^* della (2.8), detta anche *soluzione di equilibrio*, o *valore di equilibrio* o semplicemente *equilibrio*.
Per determinare una soluzione di equilibrio si procede nel seguente modo:

- se $a \neq 1$ si cerca una soluzione del tipo $y_n^* = k$ con k costante da determinare. Tale k deve soddisfare

$$y_{n+1} - ay_n = k - ak = b \quad \Rightarrow \quad k = \frac{b}{1-a}$$

per cui, se $a \neq 1$, $y_n^* = \dfrac{b}{1-a}$ è una soluzione particolare della (2.8) e la soluzione generale dell'equazione (2.8) è

$$y_n = c\,a^n + \frac{b}{1-a}\,. \tag{2.9}$$

$\dfrac{b}{1-a}$ è quel numero per il quale la popolazione rimane costante nel tempo, ossia quel numero y per il quale risulta $y - ay = b$. Il valore della costante c nella (2.9) è univocamente determinato se si conosce y_0. Infatti, sostituendo nella (2.9), si ha $c = y_0 - \dfrac{b}{1-a}$ e quindi la soluzione è

$$y_n = a^n \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}\,;$$

- se $a = 1$ si cerca una soluzione del tipo $y_n^* = kn$. Sostituendo nella (2.6) si ha

$$y_{n+1} - ay_n = k(n+1) - kn = k \quad \Rightarrow \quad k = b.$$

da cui

$$y_n^* = nb. \tag{2.10}$$

La soluzione generale dell'equazione (2.8) per $a = 1$ pertanto è

$$y_n = c + nb\,,$$

data dalla somma della (2.7), per $a = 1$, e della (2.10).
Sostituendo la condizione iniziale, si ottiene $c = y_0$, pertanto la soluzione è

$$y_n = y_0 + nb\,.$$

Schema di soluzione

$$y_{n+1} - ay_n = b$$

La soluzione è data da

$$y_n = \overline{y}_n + y_n^\star$$

con

$\overline{y}_n$	soluzione dell'equazione omogenea associata
$y_n^\star$	soluzione particolare dell'equazione completa.

$$\overline{y}_n = ca^n$$

$$y_n^\star = \begin{cases} \dfrac{b}{1-a} & a \neq 1 \\ nb & a = 1 \end{cases} \qquad \Longrightarrow \qquad y_n = \begin{cases} ca^n + \dfrac{b}{1-a} & a \neq 1 \\ c + nb & a = 1 \end{cases}$$

Se si impone la condizione iniziale y_0

$$y_n = \begin{cases} a^n \left(y_0 - \dfrac{b}{1-a} \right) + \dfrac{b}{1-a} & a \neq 1 \\ y_0 + nb & a = 1. \end{cases}$$

Esempio 2.1 *Determinare le soluzioni dell'equazione alle differenze*

$$y_{n+1} + 3y_n = 2.$$

Determinare poi la soluzione che si ottiene per $y_0 = 1$.

Le soluzioni dell'equazione omogenea $y_{n+1} + 3y_n = 0$ *sono*

$$y_n = c\,(-3)^n, \qquad c \in \mathbb{R}.$$

Cerchiamo ora una soluzione dell'equazione non omogenea del tipo $y_n^\star = k$, $k \in \mathbb{R}$. *Tale* k *deve soddisfare l'equazione e quindi*

$$k + 3k = 2 \quad \Rightarrow \quad k = \frac{1}{2}.$$

La soluzione generale dell'equazione data si ottiene come somma delle due soluzioni trovate, cioè

$$y_n = c\,(-3)^n + \frac{1}{2}.$$

Per $y_0 = 1$, *si ottiene* $c = \dfrac{1}{2}$, *e quindi*

$$y_n = \frac{1}{2}(-3)^n + \frac{1}{2}.$$

Esempio 2.2 *Determinare le soluzioni dell'equazione alle differenze*

$$y_{n+2} - y_{n+1} = 3.$$

Determinare poi la soluzione che si ottiene per $y_0 = 5$.

Le soluzioni dell'equazione omogenea sono

$$y_n = 1^n c = c.$$

Essendo $a = 1$, $y_n = k$ *con* $k \in \mathbb{R}$ *non può essere una soluzione della non omogenea, quindi si cerca una soluzione del tipo* $y_n = nk$. *Sostituendo nell'equazione*

$$(n+2)\,k - (n+1)\,k = 3 \quad \Rightarrow \quad k = 3.$$

La soluzione generale dell'equazione data quindi è

$$y_n = c + 3n \qquad e\ per\ y_0 = 5 \qquad y_n = 5 + 3n.$$

2.3.1 Comportamento della soluzione

Supposto $y_0 > 0$, i valori di a e di b determinano il comportamento della soluzione al variare di n, ossia

- per $a > 1$, la soluzione

$$y_n = a^n \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}$$

cresce all'infinito al crescere di n, poiché $\lim_{n\to\infty} a^n = +\infty$. In questo caso si dice che la soluzione è *instabile*. Quanto più a è grande, tanto più velocemente la soluzione si allontana dal valore di equilibrio. In questo caso si dice che l'equilibrio è *repulsivo*.

- per $a = 1$ la soluzione

$$y_n = y_0 + nb$$

è tale che

$$\lim_{n\to\infty} y_n = \begin{cases} +\infty & b > 0 \\ y_0 & b = 0 \\ -\infty & b < 0 \end{cases}$$

Quindi, se $b \neq 0$ non vi sono equililibri; se $b = 0$ tutti i punti sono di equilibrio e quindi nè attraggono, nè respingono le soluzioni corrispondenti, ma hanno una proprietà di *stabilità*.

- per $0 < a < 1$ la soluzione

$$y_n = a^n \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}$$

decresce tendendo al valore $\dfrac{b}{1-a}$, poiché $\lim_{n\to\infty} a^n = 0$. In questo caso la soluzione è detta *stabile*. Quanto più a è piccolo, tanto più velocemente a^n tende a 0, la soluzione converge al valore di equilibrio e si dice che l'equilibrio è *attrattivo*;

- per $a = 0$, la soluzione

$$y_n = b$$

è costante;

- per $-1 < a < 0$, a^n oscilla tra valori positivi e negativi, ma tende a zero, quindi la soluzione

$$y_n = a^n \left(y_0 - \frac{b}{1-a} \right) + \frac{b}{1-a}$$

y_n tende a $\dfrac{b}{1-a}$;

- per $a \leq -1$, a^n oscilla come nel caso precedente ma per n che tende all'infinito non ha limite e quindi anche y_n non ha limite .

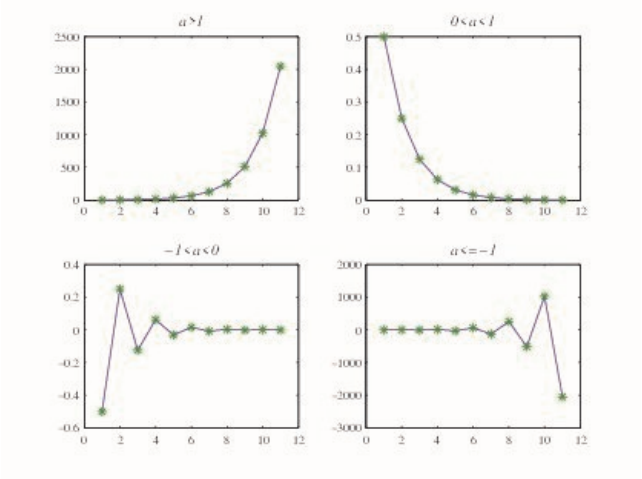


Figura 2.1

Osserviamo che nel caso di una popolazione reale la (2.8) ha senso solo se a è positiva. Il caso $a < 0$ non è realistico; $a = 0$ è poco interessante.

2.3.2 Modelli malthusiani

La prima formulazione di un modello deterministico a tempo discreto della dinamica delle popolazioni è attribuito all'economista inglese *Thomas Malthus* (The Rookery, 1766 – Bath, 1834).[(1)] Tale modello, nato dall'osservazione della popolazione umana, in particolare dai dati dei censimenti della popolazione degli Stati Uniti tra il 1789 e il 1820, si basa su ipotesi che si applicano ad una situazione ideale, riproducibile in laboratorio per alcune specie di organismi molto semplici, che può comunque in alcuni casi ritenersi verificata in natura, almeno per periodi di tempo sufficientemente limitati.
Poche grandezze in natura però possono continuare a crescere in modo esponenziale per periodi di tempo lunghi, in quanto la crescita è di solito limitata da vincoli esterni, quali carestie, malattie, guerre, e soprattutto disponibilità delle risorse alimentari: quando la popolazione cresce, la capacità dell'ambiente di sopportare tale crescita diminuisce. Il modello di Malthus pertanto risulta essere non rappresentativo della realtà e quindi deve essere opportunamente modificato.

Figura 2.2 - T.R. Malthus (1766–1834)

Si chiamano **modelli malthusiani** i modelli nei quali il tasso di accrescimento $r(t, y)$ è costante, ossia non dipende ne' dal tempo t ne' dalla dimensione x della popolazione. Inoltre, per studiare la dinamica di una popolazione secondo il modello di Malthus si deve supporre che

a) la popolazione sia omogenea (gli individui che la compongono si possono considerare identici);
b) la popolazione sia isolata (fertilità e mortalità uniche cause di variazione della popolazione);
c) l'habitat sia invariante (risorse e condizioni di vita non influenzate da fattori esterni).

Sotto queste ipotesi l'equazione (2.2) diventa

$$y_{k+1} = (1+r)y_k \quad \Longrightarrow \quad \text{(equazione (2.6) con } a = 1+r) \quad \Longrightarrow \quad y_k = (1+r)^k y_0$$

essendo y_0 la dimensione della popolazione al tempo $t = 0$.

Si hanno i seguenti casi

- il numero delle nascite e delle immigrazioni è superiore al numero delle morti e delle emigrazioni, e quindi $r > 0$. In questo caso la dimensione della popolazione cresce indefinitamente con il tempo, ossia $y_n \to \infty$ per $n \to \infty$;
- il numero delle morti e delle emigrazioni è superiore al numero delle nascite e delle immigrazioni, e quindi $r < 0$. In questo caso la dimensione della popolazione decresce con il tempo ($y_n \to 0$ per $n \to \infty$), quindi la popolazione è destinata ad estinguersi esponenzialmente;
- il numero delle nascite e delle immigrazioni compensa esattamente il numero delle morti e delle emigrazioni, e quindi $r = 0$. In questo caso la popolazione è *stazionaria*, ossia $y_n = y_0$ per ogni n.

[(1)] Nato nel Surrey (Inghilterra) nel 1766, Thomas Robert Malthus fu Pastore anglicano e vicario di Albury fino al 1803, quando decise di dedicarsi esclusivamente all'economia, divenendo docente all'Hailbury College. Noto per le sue tesi anti-illuministe e fortemente influenzato dalle teorie economiche di Adam Smith, teorizzò (nella sua opera *An assay of the principle of the population as it effects the future improvement of society*) la crescita geometrica della popolazione, in presenza di risorse che invece crescevano in modo aritmetico, quindi molto più lentamente. Forte di queste convinzioni, ritenne opportuno opporsi a qualsiasi forma di ammortizzatore sociale e di intervento pubblico per alleviare le grandi sofferenze delle classi povere nell'Inghilterra del XIX sec., in particolare alla Poor Law, provvedimento legislativo mirante a introdurre un'integrazione al salario dei lavoratori poveri, proveniente dalla tassazione dei fondi agricoli. La teoria demografica di Malthus ispirò la corrente del malthusianesimo, che sostiene il ricorso al controllo delle nascite per impedire l'impoverimento dell'umanità. Oggi vengono chiamate malthusiane le teorie che indicano nella crescita demografica la causa principale della miseria.

2.3.3 Esempi di modelli malthusiani

Esempio 2.3 ***(La divisione cellulare)***. *Nei processi di divisione cellulare,* ***mitosi*** *e* ***meiosi****, si formano rispettivamente* 2 *cellule figlie (aventi lo stesso patrimonio genetico della cellula madre) e* 4 *cellule figlie (dotate di un patrimonio dimezzato di cromosomi).*
Il periodo che intercorre tra la genesi *di una cellula generata da una precedente divisione e il momento in cui questa a sua volta si divide, costituisce il ciclo vitale della cellula.*
Se indichiamo con M_1, M_2, ..., M_n, *il numero di cellule rispettivamente nella prima, nella seconda, ..., nella n-esima generazione, una semplice equazione che collega le successive generazioni è data da*

$$M_{n+1} = 2M_n \qquad \textit{nel caso della mitosi}$$

$$M_{n+1} = 4M_n \qquad \textit{nel caso della meiosi.}$$

Si tratta di due equazioni alle differenze lineari a coefficienti costanti del primo ordine. Se M_0 *è il numero di cellule iniziali, dopo n generazioni la popolazione sarà rispettivamente*

$$M_n = 2^n M_0 \qquad e \qquad M_n = 4^n M_0.$$

Esempio 2.4 *Dopo averli importati e usati come mezzo di trasporto,* 80 *cammelli vengono lasciati liberi nel deserto. Per tale popolazione si ha* $n^* = 0.3$ *e* $d^* = 0.18$ *per stagione (ossia il tasso di natalità è* 30% *e quello di mortalità è* 18%*). Si vuole conoscere la dimensione della popolazione dei cammelli nelle prime* 12 *stagioni?*

In questo caso il modello è dato da

$$x_{k+1} = (1 + 0.3 - 0.18)\, x_k = 1.12 x_k \qquad x_1 = 80.$$

Dobbiamo calcolare x_k *per* $k = 2, ..., 12$. *Nella tabella che segue sono riportati i risultati ottenuti, arrotondati, che indicano una crescita di tipo esponenziale. Ne deriva che il modello matematico utilizzato è inadeguato per lunghi periodi: nessuna popolazione può crescere indefinitamente!*

Anno	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Inizio	*80*	*90*	*100*	*112*	*126*	*141*	*158*	*177*	*198*	*222*	*248*	*278*
Fine	*90*	*100*	*112*	*126*	*141*	*158*	*177*	*198*	*222*	*248*	*278*	*312*

Cosa accade nel modello di base se il tasso di natalità è minore del tasso di mortalità?

Esempio 2.5 *La famiglia degli afidi, meglio noti con il nome di pidocchi delle piante, comprende una varietà estesissima di insetti di piccolissime dimensioni, dotati di un potente apparato succhiatore, grazie al quale surgono la linfa delle piante cui sono parassiti, costituendo un vero pericolo per l'agricoltura e per le aree verdi in generale. Ci limitiamo a considerare gli afidi dei pioppi, in grado di produrre nella corteccia di questi ultimi ulcere profondissime. All'interno di tali ulcere, gli afidi depositano le loro larve, la cui sopravvivenza e successiva crescita dipendono da numerosi fattori, tra cui le condizioni ambientali, la quantità del cibo disponibile, etc.*
Per semplificare, prenderemo in considerazione i seguenti parametri

- a_n *numero di afidi adulti femmina alla n-esima generazione;*
- p_n *numero di discendenti alla n-esima generazione;*
- f *numero dei discendenti per femmina adulta;*
- r *rapporto tra femmine adulte e totale della popolazione adulta.*

Vale

$$p_{n+1} = f\, a_n \tag{2.11}$$

dove p_{n+1} *sono i discendenti alla* $(n+1)$*-esima generazione. Di questi, se* d^* *è il tasso di mortalità, solo* $(1-d^*)p_{n+1}$ *sopravviverà sino all'età adulta. Quindi*

$$a_{n+1} = r(1-d^*)p_{n+1}$$

e sostituendo p_{n+1} *calcolato in* (2.11)*, si ricava*

$$a_{n+1} = f\, r(1-d^*)a_n.$$

La soluzione dell'equazione ottenuta, supponendo che f*,* r*,* d^* *siano costanti, è*

$$a_n = [f\, r(1-d^*)]^n\, a_0$$

essendo a_0 *il numero iniziale di femmine adulte.*

Esempio 2.6 ***(Riproduzione dei globuli rossi)***
Gli eritrociti sono cellule che fungono da vettore per l'ossigeno, grazie alla presenza in esse dell'emoglobina. La loro vita è piuttosto breve e varia da 2-3 settimane ad alcuni mesi, in un continuo ricambio tra cellule distrutte e cellule create. Data la funzione espletata, il livello complessivo di eritrociti deve essere mantenuto costante. Supponiamo che la milza (organo preposto alla distruzione dei globuli rossi senescenti) elimini, giornalmente, un certo numero di cellule e che il midollo spinale produca un numero di cellule proporzionale a quello perduto il giorno precedente. Quanti saranno i globuli rossi prodotti l'n-esimo giorno?
Consideriamo le seguenti quantità

- R_n *numero di globuli rossi in circolazione il giorno n;*
- M_n *numero di globuli rossi prodotti dal midollo il giorno n;*
- f *tasso di "rimozione" di globuli rimossi da parte della milza;*
- γ *costante di produzione (numero di cellule prodotte sul numero di cellule perdute).*

Si ottengono le equazioni

$$R_{n+1} = (1-f)R_n + M_{n+1} \qquad\qquad M_{n+1} = \gamma f R_n$$

da cui $R_{n+1} = (1-f+\gamma f)R_n$*, ossia*

$$R_{n+1} = [1 + f(\gamma - 1)]\, R_n.$$

Si deduce facilmente che per mantenere pressochè costante il numero di globuli deve essere $\gamma \approx 1$*.*

2.3.4 Equazione logistica di Verhulst

Il modello proposto da Malthus prevede una crescita illimitata e non può quindi essere considerato una rappresentazione realistica dell'andamento delle popolazioni a lungo termine. Per esempio una colonia di batteri in abbondante liquido di coltura inizialmente è soggetta ad una crescita di tipo malthusiano; poi, al crescere esponenziale del numero di batteri, l'inquinamento ambientale e la mancanza di risorse modificano, anche radicalmente, i tassi demografici e il modello non è più attendibile. Il modello malthusiano si può applicare alla crescita dei batteri in vitro, ma non ad una popolazione di esseri umani.
In natura, in condizioni di normalità, cioè in assenza di catastrofi particolari, spesso la popolazione di una certa specie è stazionaria o oscilla intorno a un valore stazionario.

Molte popolazioni cominciano a crescere in maniera esponenziale, ma il livello della popolazione non oltrepassa la *capacità* dell'ambiente (o *livello di saturazione* o *portata*), cioè il massimo numero di individui che l'ambiente è in grado di sostenere sul lungo periodo.

Queste considerazioni spinsero il biomatematico *Pierre François Verhulst* (Bruxelles, 1804 – 1849) nel 1845 a trovare un nuovo e più sofisticato modello, oggi noto come *equazione logistica*: ad alte densità, un aumento della popolazione produce un incremento della mortalità e una diminuzione della fertilità.

Figura 2.3 - P.F. Verhulst (1804–1849)

Al modello di Malthus, Verhulst aggiunse alcune limitazioni, ossia un termine che diminuisse il tasso di crescita quando y_t diventava molto grande

$$y_{t+1} - y_t = r y_t \left(1 - \frac{y_t}{L}\right) . \tag{2.12}$$

In tal modo la crescita di una popolazione non è governata solo da un termine di crescita libero malthusiano, ma anche da un meccanismo di regolazione che contrasta la crescita libera. Infatti, quando y_t si avvicina a L, $1 - \dfrac{y_t}{L}$ tende ad avvicinarsi a 0 e quindi la crescita tende a 0 (all'aumentare del numero di individui diventa sempre più difficile la sopravvivenza).

L'equazione non lineare (2.12) può essere scritta nella forma

$$y_{t+1} = (1 + r) y_t - \frac{r}{L} y_t^2 .$$

dove il termine $-\dfrac{r}{L} y_t^2$ funge da "freno" della crescita della popolazione.

Tale equazione non può essere risolta esplicitamente, ma è possibile calcolare i termini y_t ricorsivamente.

2.4 Equazioni alle differenze del secondo ordine

Come nel caso di equazioni del primo ordine, anche la soluzione generale dell'equazione alle differenze lineare a coefficienti costanti di **ordine 2**

$$a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = b, \tag{2.13}$$

si ottiene come somma della soluzione generale dell'equazione omogenea associata

$$a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = 0 \tag{2.14}$$

e di una soluzione particolare della (2.13).

Per risolvere la (2.14), cerchiamo soluzioni del tipo $y_n = \lambda^n$, con λ costante da determinare.

Se $y_n = \lambda^n$ deve essere soluzione della (2.14), deve soddisfare l'equazione, ossia

$$a_2\lambda^{n+2} + a_1\lambda^{n+1} + a_0\lambda^n = \lambda^n\left(a_2\lambda^2 + a_1\lambda + a_0\right) = 0$$

da cui

$$a_2\lambda^2 + a_1\lambda + a_0 = 0. \tag{2.15}$$

La (2.15) prende il nome di **equazione caratteristica** dell'equazione omogenea. Si tratta di un'equazione algebrica di secondo grado, le cui soluzioni, com'è noto, dipendono dal segno del discriminante

$$\Delta = a_1^2 - 4a_0a_2.$$

Si distinguono tre casi

- $\Delta > 0$: il polinomio $a_2\lambda^2 + a_1\lambda + a_0$ ha due radici reali e distinte, che indichiamo con λ_1 e λ_2. Allora $z_n = \lambda_1^n$ e $w_n = \lambda_2^n$ risultano soluzioni dell'equazione omogenea (2.14). Per il teorema 1, tutte le soluzioni dell'equazione omogenea sono date da

  $$y_n = c_1\lambda_1^n + c_2\lambda_2^n \qquad \forall n \in \mathbb{N}$$

  con $c_1, c_2 \in \mathbb{R}$.

- $\Delta = 0$: il polinomio $a_2\lambda^2 + a_1\lambda + a_0$ ha due radici reali coincidenti, che indichiamo con λ. Allora tutte le successioni del tipo $z_n = k\lambda^n$ risultano soluzioni dell'equazione omogenea (2.14).

  In particolare, se scegliamo $z_n = \lambda^n$ e $w_n = n\lambda^n$, per il teorema 1, tutte le soluzioni della (2.14) sono date da

  $$y_n = c_1\lambda^n + c_2n\lambda^n \qquad \forall n \in \mathbb{N}$$

  con $c_1, c_2 \in \mathbb{R}$.

- $\Delta < 0$: il polinomio $a_2\lambda^2 + a_1\lambda + a_0$ ha due radici complesse coniugate $\alpha \pm i\beta$. In questo caso, posto $r = \sqrt{\alpha^2 + \beta^2}$ e θ un argomento di $\alpha \pm i\beta$ ($\tan\theta = \frac{\beta}{\alpha}$), le soluzioni dell'equazione omogenea associata sono date da

  $$y_n = c_1r^n\cos(n\theta) + c_2r^n\sin(n\theta) \qquad \forall n \in \mathbb{N}$$

  con $c_1, c_2 \in \mathbb{R}$.

  In questo caso la soluzione oscilla. r è il fattore di crescita, $\theta/2\pi$ rappresenta la *frequenza delle oscillazioni*. Se $|r| < 1$, le oscillazioni sono smorzate; se $|r| > 1$ le oscillazioni sono esplosive.

Una soluzione y_n dell'equazione (2.14) è univocamente determinata se sono noti i primi due termini y_0 e y_1, o due termini qualunque della successione y_n.

Esempio 2.7 *Determinare tutte le soluzioni dell'equazione alle differenze omogenea*

$$y_{n+2} - 4y_{n+1} + 4y_n = 0.$$

Consideriamo l'equazione caratteristica associata $\lambda^2 - 4\lambda + 4 = 0$. *Essa ha soluzioni* $\lambda_1 = \lambda_2 = 2$, *per cui la soluzione dell'equazione data è*

$$y_n = c_1 2^n + c_2 n 2^n.$$

Esempio 2.8 *Determinare tutte le soluzioni dell'equazione alle differenze omogenea*

$$y_{n+2} - y_n = 0.$$

Determinare poi la soluzione che si ottiene per $y_0 = 0, y_1 = 2$.

Cerchiamo una soluzione del tipo λ^n. *Tale* λ *deve soddisfare l'equazione* $\lambda^2 - 1 = 0$, *che ha due soluzioni reali e distinte* $\lambda_1 = -1$ *e* $\lambda_2 = 1$. *Pertanto tutte le soluzioni dell'equazione sono date da*

$$y_n = c_1(-1)^n + c_2 1^n.$$

Se imponiamo $y_0 = 0$ *e* $y_1 = 2$, *occorre determinare* c_1 *e* c_2 *tali che*

$$\begin{cases} c_1 + c_2 = 0 \\ -c_1 + c_2 = 2 \end{cases}$$

da cui si ricava $c_1 = -1$, $c_2 = 1$. *Quindi la soluzione dell'equazione data è*

$$y_n = (-1)(-1)^n + 1 = (-1)^{n+1} + 1.$$

Esempio 2.9 *Determinare tutte le soluzioni dell'equazione alle differenze omogenea*

$$y_n + y_{n-1} + \frac{1}{2} y_{n-2} = 0.$$

L'equazione caratteristica associata è $\lambda^2 + \lambda + \frac{1}{2} = 0$ *che ha soluzioni* $\lambda_1 = -\frac{1}{2} - \frac{1}{2}i$, $\lambda_2 = -\frac{1}{2} + \frac{1}{2}i$. *In questo caso*

$$r = \sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2} = \frac{\sqrt{2}}{2}, \qquad \theta = arctg(-1) = \frac{3}{4}\pi,$$

per cui la soluzione dell'equazione data è

$$y_n = \left(\frac{\sqrt{2}}{2}\right)^n \left(c_1 \cos\frac{3}{4}\pi n + c_2 \sin\frac{3}{4}\pi n\right).$$

Per calcolare la soluzione dell'equazione non omogenea

$$a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = b. \tag{2.17}$$

dobbiamo determinarne una soluzione particolare.

In questo caso si cerca per la (2.17) una soluzione costante $y_n^* = k$, con k costante da determinare. Sostituendo nella (2.17), si ha

$$a_2 k + a_1 k + a_0 k = b \quad \Leftrightarrow \quad (a_2 + a_1 + a_0)k = b.$$

- Se $a_2 + a_1 + a_0 \neq 0$, si ricava $k = \dfrac{b}{a_2 + a_1 + a_0}$ e quindi

$$y_n^* = \frac{b}{a_2 + a_1 + a_0};$$

- se $a_2+a_1+a_0=0$, si prova se $y_n^* = k\,n$ è soluzione dell'equazione (2.17). Un tale k dovrà verificare

$$a_2k(n+2)+a_1k(n+1)+a_0kn=b$$

ossia $\underbrace{(a_2+a_1+a_0)}_{=0}\,kn+(2a_2+a_1)\,k=b$ e quindi dovrà essere

$$(2a_2+a_1)\,k=b;$$

  - se $2a_2+a_1\neq 0$ si ha $k=\dfrac{b}{2a_2+a_1}$ e quindi

$$y_n^*=\frac{b}{2a_2+a_1}\,n;$$

  - se $2a_2+a_1=0$, si cerca una soluzione particolare del tipo $y_n^*=kn^2$. In questo caso k dovrà verificare

$$a_2k\,(n+2)^2+a_1k\,(n+1)^2+a_0kn^2=b$$

ossia

$$\underbrace{(a_2+a_1+a_0)}_{=0}\,kn^2+2\,\underbrace{(2a_2+a_1)}_{=0}\,kn+4a_2k+a_1k=b$$

da cui

$$(4a_2+a_1)\,k=b \quad\Leftrightarrow\quad k=\frac{b}{4a_2+a_1}$$

e quindi

$$y_n^*=\frac{b}{4a_2+a_1}\,n^2.$$

Si nota esplicitamente che y_n^* è ben definita infatti non può verificarsi $4a_2+a_1=0$. Tale condizione infatti, insieme alla condizione $2a_2+a_1=0$, prevederebbe l'unica soluzione $a_2=a_1=0$ e quindi l'equazione assegnata non avrebbe senso.

La soluzione dell'equazione

$$a_2y_{n+2}+a_1y_{n+1}+a_0y_n=b.$$

è data quindi dalla somma della soluzione (2.16) dell'equazione omogenea (2.14) e della soluzione particolare (2.18) appena ricavata.

Schema di soluzione

$$a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = b$$

La soluzione è data da

$$y_n = \overline{y}_n + y_n^\star$$

con
$\overline{y}_n$ soluzione dell'equazione omogenea associata
$y_n^\star$ soluzione particolare dell'equazione completa.

Si cercano le soluzioni dell'equazione caratteristica

$$a_2\lambda^2 + a_1\lambda + a_0 = 0$$

$$\Delta = a_1^2 - 4a_0a_2, \qquad \begin{array}{llll} 2 \text{ soluzioni} & \lambda_{1,2} = \frac{-a_1 \pm \sqrt{\Delta}}{2a_2} & \Delta > 0 \\ 1 \text{ soluzione} & \lambda = \frac{-a_1}{2a_2} & \Delta = 0 \end{array}$$

(non consideriamo il caso $\Delta < 0$)

$$\overline{y}_n = \begin{cases} c_1\lambda_1^n + c_2\lambda_2^n & \Delta > 0 \\ c_1\lambda^n + c_2 n\lambda^n & \Delta = 0 \end{cases}$$

$$y_n^\star = \begin{cases} \dfrac{b}{a_2 + a_1 + a_0} & a_2 + a_1 + a_0 \neq 0 \\ \dfrac{b}{2a_2 + a_1} n & a_2 + a_1 + a_0 = 0, \quad 2a_2 + a_1 \neq 0 \\ \dfrac{b}{4a_2 + a_1} n^2 & a_2 + a_1 + a_0 = 0, \quad 2a_2 + a_1 = 0 \end{cases}$$

quindi in definitiva

$$y_n = \begin{cases} c_1\lambda_1^n + c_2\lambda_2^n + \dfrac{b}{a_2 + a_1 + a_0} & \Delta > 0, \quad a_2 + a_1 + a_0 \neq 0 \\ c_1\lambda_1^n + c_2\lambda_2^n + \dfrac{b}{2a_2 + a_1} n & \Delta > 0, \quad a_2 + a_1 + a_0 = 0, \quad 2a_2 + a_1 \neq 0 \\ c_1\lambda_1^n + c_2\lambda_2^n + \dfrac{b}{4a_2 + a_1} n^2 & \Delta > 0, \quad a_2 + a_1 + a_0 = 0, \quad 2a_2 + a_1 = 0 \\ & \\ c_1\lambda^n + c_2 n\lambda^n + \dfrac{b}{a_2 + a_1 + a_0} & \Delta = 0, \quad a_2 + a_1 + a_0 \neq 0 \\ c_1\lambda^n + c_2 n\lambda^n + \dfrac{b}{2a_2 + a_1} n & \Delta = 0, \quad a_2 + a_1 + a_0 = 0, \quad 2a_2 + a_1 \neq 0 \\ c_1\lambda^n + c_2 n\lambda^n + \dfrac{b}{4a_2 + a_1} n^2 & \Delta = 0, \quad a_2 + a_1 + a_0 = 0, \quad 2a_2 + a_1 = 0 \end{cases}$$

Le costanti c_1 e c_2 si calcolano imponendo le condizioni iniziali

Esempio 2.10 *Risolvere l'equazione alle differenze*

$$y_{n+2} - 5y_{n+1} + 6y_n = 4.$$

Determinare poi la soluzione che si ottiene per $y_0 = 0, y_1 = 3$.

Le soluzioni sono del tipo $y_n = \overline{y}_n + y_n^*$ *essendo* $\overline{y}_n$ *le soluzioni generali dell'equazione omogenea associata e* y_n^* *la soluzione particolare dell'equazione data.*
Si ricava subito che

$$\overline{y}_n = c_1 2^n + c_2 3^n.$$

Cerhiamo quindi una soluzione particolare del tipo $y_n^* = k$. *Sostituendo nell'equazione, si ha*

$$k - 5k + 6k = 4 \quad \rightarrow \quad (1 - 5 + 6)k = 4 \quad \rightarrow \quad k = 2,$$

pertanto

$$y_n = c_1 2^n + c_2 3^n + 2.$$

Imponendo le condizioni iniziali, c_1 *e* c_2 *si determinano risolvendo il sistema*

$$\begin{cases} c_1 + \quad c_2 + 2 \quad = 0 \\ 2c_1 + \quad 3c_2 + 2 \quad = 3\,. \end{cases}$$

Si ricava $c_1 = -7, c_2 = 5$, *quindi la soluzione cercata è*

$$y_n = -7 \cdot 2^n + 5 \cdot 3^n + 2.$$

Esempio 2.11 *Determinare le soluzioni dell'equazione alle differenze*

$$y_{n+2} - 2y_{n+1} + y_n = 3.$$

Tutte le soluzioni dell'equazione alle differenze omogenea associata sono date da

$$y_n = c_1 1^n + c_2 n 1^n = c_1 + n c_2.$$

Determiniamo ora una soluzione dell'equazione completa, della forma $y_n^* = k$. *Tale costante* k *deve verificare* $k - 2k + k = 3$, *che però non ha soluzioni.*
Proviamo quindi con $y_n^* = kn$ *e troviamo di nuovo un'equazione priva di soluzioni* $2k - 2k = 3$. *Infine, con* $y_n^* = kn^2$ *troviamo* $2k = 3$, *quindi* $y_n^* = \frac{3}{2}n^2$. *Pertanto tutte le soluzioni dell'equazione data sono*

$$y_n = c_1 + n c_2 + \frac{3}{2}n^2 \qquad n \in \mathbb{N}, \quad c_1, c_2 \in \mathbb{R}.$$

Osservazione 1. *La soluzione generale di una equazione alle differenze lineare di ordine superiore si ottiene sempre, come nel caso del primo e del secondo ordine, come somma della soluzione dell'equazione omogenea associata e di una soluzione particolare dell'equazione completa con tecniche analoghe a quelle viste.*

2.4.1 Equilibrio e stabilità

Un *equilibrio* per l'equazione (2.17) è un valore $\alpha \in \mathbb{R}$ tale che la successione costante $y_n = \alpha$ è soluzione dell'equazione.
Da quanto detto nel paragrafo precedente,

- se $a_1 + a_2 + a_3 \neq 0$, allora il valore $\alpha = k = \dfrac{b}{a_1 + a_2 + a_3}$ è l'unico equilibrio per l'equazione (2.13);
- se $a_1 + a_2 + a_3 = 0$ e $b = 0$, allora ogni $\alpha \in \mathbb{R}$ è di equilibrio;
- se $a_1 + a_2 + a_3 = 0$ e $b \neq 0$, allora non esistono equilibri.

Un punto di equilibrio α si dice

- *stabile* se e solo se $|\lambda| \leq 1$ per ogni radice λ dell'equazione caratteristica associata e quelle di modulo 1 ($|\lambda| = 1$) sono semplici (stabilità nel senso di Lyapunov)
- *globalmente stabile* se e solo se $|\lambda| < 1$ per ogni radice λ dell'equazione caratteristica associata. In questo caso α si dice *attrattivo*. Infatti, per ogni coppia di valori iniziali la soluzione corrispondente tende all'equilibrio, ossia
$$\lim_{n\to\infty} y_n = \alpha.$$
- *instabile* in tutti gli altri casi.

Esempio 2.12 *Per gli esempi precedenti vale*

- *negli esempi 2.7 e 2.10 si ha un unico punto di equilibrio instabili (risp. $\alpha = 0$ e $\alpha = 2$);*
- *nell'esempio 2.8 ogni $\alpha \in \mathbb{R}$ è di equilibrio stabile (nel senso di Lyapunov);*
- *nell'esempio 2.9 $\alpha = 0$ è l'unico punto di equilibrio che risulta globalmente stabile;*
- *l'equazione dell'esempio 2.11 non ha equilibri.*

2.4.2 La successione di Fibonacci

Leonardo da Pisa[(2)], noto come *Fibonacci*, è l'autore del *Liber Abaci* (1202), il testo di matematica più diffuso nell'Occidente cristiano sino alla pubblicazione della *Summa de arithmetica, geometria, proportioni et proportionalità* del frate francescano Luca Pacioli (1445-1517). Come dichiara lo storico della Matematica Carl Boyer, "Se una gran parte del Liber Abaci, è di lettura molto noiosa, alcuni problemi, però, sono così vivaci da essere usati da autori posteriori." Il più noto di tali problemi rappresenta uno dei primi esempi di utilizzo di un modello matematico applicato ad una questione di scienze naturali.

Figura 2.4 - Fibonacci (1170–1242)

[(2)] Matematico italiano (Pisa, 1175 ca. - Pisa, 1240 ca.). Insieme al padre Guglielmo dei Bonacci (Fibonacci sta infatti per filius Bonacci), facoltoso mercante pisano e responsabile del commercio pisano presso la colonia di Bugia, in Algeria, trascorse alcuni anni in quella città, dove studiò i procedimenti aritmetici che studiosi musulmani stavano diffondendo nelle varie regioni del mondo arabo. Grazie ai numerosi viaggi a fianco del padre, ebbe occasione di riconoscere i vantaggi offerti dai sistemi matematici localmente in uso. Nel 1202 pubblicò la sua opera fondamentale, il *Liber Abaci* con cui si propose di diffondere nel mondo scientifico occidentale le regole di calcolo note agli Arabi, ovvero il sistema decimale ancora oggi in uso in Europa.

Il problema è il seguente: quante coppie di conigli verranno prodotte in un anno, a partire da un'unica coppia, se ogni mese ciascuna coppia dà alla luce una coppia che diventa produttiva il mese successivo? Si assume che durante l'anno non muoia nessun coniglio.
La soluzione è piuttosto semplice e può essere schematizzata nel modo seguente

Mese	Coppie	
1	1	La coppia originaria appena nata
2	1	La coppia originaria compie 1 mese
3	2	Coppia originaria + nuova prole
4	3	Un'altra coppia della coppia originaria
5	5	Le tre coppie del 4° mese più altre due coppie nate da quelle abbastanza adulte per avere una prole (che hanno cioè compiuto almeno 2 mesi)

La successione delle coppie 1, 1, 2, 3, 5,... dà origine alla successione dei *numeri di Fibonacci*, ove l'n-esimo termine è ottenuto dalla somma dei due termini precedenti

$$u_n = \underbrace{u_{n-1}}_{\substack{\text{tutte le coppie vive}\\ \text{il mese scorso e quindi}\\ \text{ancora vive}}} + \underbrace{u_{n-2}}_{\substack{\text{tutte le coppie nate}\\ \text{dalle coppie vive}\\ \text{due mesi prima}}} \tag{2.19}$$

Inoltre $u_1 = u_2 = 1$.

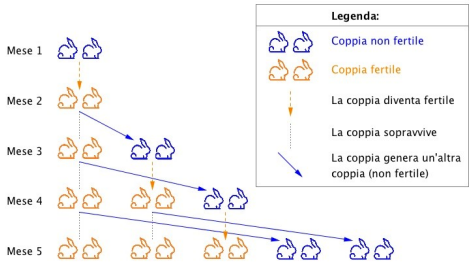


Figura 2.5

I numeri di Fibonacci possiedono eleganti proprietà e, fatto tanto più sorprendente, si ritrovano spesso in natura! Per esempio molte piante da fiore tendono ad avere un numero di petali secondo la successione dei numeri di Fibonacci; i semi del girasole sono disposti su due insiemi di spirali: uno con le spirali avvolte in senso orario e l'altro antiorario. Se contiamo il numero di spirali in un senso o nell'altro otteniamo un numero di Fibonacci; lo stesso accade se contiamo le spirali di una pigna conica; osservando un'arnia notiamo che i percorsi possibili che un'ape può fare spostandosi verso destra e muovendosi una cella per volta seguono la successione di Fibonacci.

La (2.19) è un'equazione alle differenze del secondo ordine. Per determinarne la soluzione consideriamo l'equazione caratteristica associata

$$\lambda^2 - \lambda - 1 = 0$$

le cui soluzioni sono $\dfrac{1 \pm \sqrt{5}}{2}$. Tutte le soluzioni della (2.19) sono quindi del tipo

$$u_n = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^n + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^n, \qquad c_1,\ c_2 \ \text{ costanti da determinare.}$$

Imponendo le condizioni iniziali $u_1 = u_2 = 1$, troviamo che c_1 e c_2 devono essere le soluzioni del sistema

$$\begin{cases} c_1\left(\frac{1+\sqrt{5}}{2}\right) + c_2\left(\frac{1-\sqrt{5}}{2}\right) = 1 \\ c_1\left(\frac{1+\sqrt{5}}{2}\right)^2 + c_2\left(\frac{1-\sqrt{5}}{2}\right)^2 = 1. \end{cases}$$

Risolviamo tale sistema moltiplicando la prima equazione per $\frac{1+\sqrt{5}}{2}$

$$\begin{cases} c_1\left(\frac{1+\sqrt{5}}{2}\right)^2 + c_2\left(\frac{1-\sqrt{5}}{2}\right)\left(\frac{1+\sqrt{5}}{2}\right) = \frac{1+\sqrt{5}}{2} \\ c_1\left(\frac{1+\sqrt{5}}{2}\right)^2 + c_2\left(\frac{1-\sqrt{5}}{2}\right)^2 = 1. \end{cases}$$

e sottraendo membro a membro otteniamo

$$c_2\left(\frac{1-\sqrt{5}}{2}\right)\left(\frac{1+\sqrt{5}}{2} - \frac{1-\sqrt{5}}{2}\right) = \frac{1+\sqrt{5}}{2} - 1$$

$$c_2\left(\frac{1-\sqrt{5}}{2}\right)\sqrt{5} = \frac{-1+\sqrt{5}}{2} = -\frac{1-\sqrt{5}}{2} \quad \Rightarrow \quad c_2 = -\frac{1}{\sqrt{5}}.$$

Sostituendo ora il valore di c_2 nella prima equazione, si ha

$$\begin{aligned} c_1\left(\frac{1+\sqrt{5}}{2}\right) \quad &= 1 - c_2\left(\frac{1-\sqrt{5}}{2}\right) = 1 + \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right) = \\ &= \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right) \quad \Rightarrow \quad c_1 = \frac{1}{\sqrt{5}}. \end{aligned}$$

Quindi la soluzione generale dell'equazione (2.19) è

$$u_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

2.5 Esercizi

2.5.1 Equazioni alle differenze del primo ordine

Esercizio 2.1 *Se inizialmente depositiamo un capitale C e annualmente versiamo una quota b e se il tasso di interesse bancario è r, calcolare l'ammontare totale dopo* 10 *anni.*

Esercizio 2.2 *Una nave che trasporta delle gabbie di topolini bianchi naufraga su un isola e si salvano 1000 topolini. Il tasso di natalità per mese, $n^\star$ è 0.6 (cioè nascono 60 topolini su 100) e il tasso di mortalità, $d^\star$, è inizialmente 0.95 per mese. Si cerchi di prevedere la futura popolazione mensile di topolini bianchi*

Esercizio 2.3 *Dopo aver risolto l'equazione alle differenze del primo ordine*

$$y_{n+1} = 2y_n + 1, \qquad y_0 = 0$$

tracciare in un grafico cartesiano i primi 5 *valori della soluzione.*

Esercizio 2.4 *Dopo aver risolto l'equazione alle differenze del primo ordine*

$$y_{n+1} - \frac{1}{2} y_n = 2, \qquad y_0 = 3,$$

tracciare in un grafico cartesiano i primi 5 *valori della soluzione.*

Esercizio 2.5 *Risolvere l'equazione alle differenze del primo ordine*

$$y_{n+1} - (1-p)\, y_n = b, \quad p \in \mathbb{R} \qquad y_0 = b.$$

Esercizio 2.6 *La popolazione della Bolivia nel 2003 era pari a* 8.6 *milioni di persone e la sua velocità di crescita era approssimativamente dell'*1.8*%. Il numero delle persone che lasciavano il paese eccedeva di* 12 000 *persone il numero di quelli che entravano nel paese.*

1. *Modellizzare la popolazione con un'opportuna equazione alle differenze con condizione iniziale;*
2. *determinare la soluzione generale dell'equazione;*
3. *usare il modello per dare una stima della popolazione nell'anno 2023;*
4. *costruire un opportuno codice MatLab che visualizzi in un grafico l'andamento della popolazione per un periodo di 15 anni e costruisca una tabella che riporti in prima colonna l'anno e in seconda colonna la relativa popolazione.*

Esercizio 2.7 *Una tazza di caffè, alla temperatura iniziale di* $85°C$*, si raffredda fino a* $80°C$ *in un minuto quando viene posta in una stanza alla temperatura costante* $S = 25°C$*. Sia* T_n *la temperatura del caffè dopo n minuti. La differenza tra la temperatura in due istanti successivi,* T_{n+1} *e* T_n*, è direttamente poporzionale alla differenza tra la temperatura* T_n *e la temperatura della stanza* S*.*

1. *Scrivere l'equazione alle differenze a coefficienti costanti che descriva la variazione della temperatura del caffè minuto per minuto;*
2. *fornire la soluzione generale dell'equazione al punto 1;*
3. *noto che, se la temperatura supera i* $35°C$*, la percezione del gusto diminuisce determinare l'istante ottimale dopo cui si può gustare al meglio il caffè;*
4. *stabilire a quale valore tenderà la temperatura del caffè al trascorrere del tempo;*
5. *costruire una function MatLab che riceve in input n e restituisce la temperatura del caffè dopo n minuti. La function deve mostrare in un grafico l'andamento della temperatura al variare del tempo e riportare in una tabella il vettore dei minuti e il vettore delle corrispondenti temperature del caffè.*

Esercizio 2.8 *Una popolazione di lontre sta decrescendo del* 25*% ogni anno a causa della distruzione del loro habitat e di malattie. Il corpo forestale ha deciso di introdurre* 125 *lontre nate in cattività ogni anno per risanare la situazione.*

1. *Se la popolazione corrente di lontre ammonta a* 100 *esemplari, si scriva una equazione alle differenze che modellizzi la situazione;*
2. *costruire un codice MatLab che rappresenti in un grafico l'andamento della popolazione per un periodo di* 12 *anni e costruisca una tabella che riporti in prima colonna l'anno e in seconda colonna il relativo numero di esemplari di lontre;*
3. *da un punto di vista biologico, una popolazione è in equilibrio quando il numero degli esemplari si attesta intorno a un certo valore; determinare il valore di equilibrio per la popolazione di lontre;*
4. *da un punto di vista matematico, un valore di equilibrio per una equazione alle differenze è un valore per* p_{n+1} *pari al valore di* p_n*; determinare algebricamente il punto di equilibrio per la popolazione di lontre e stabilire se il risultato ottenuto è in accordo con quanto osservato al punto 3.*

Esercizio 2.9 *Una popolazione di cinghiali crescerebbe del* 14.5% *all'anno se non vi fossero agenti esterni a perturbare il loro ambiente. Allo scopo di controllare le nascite nel breve periodo, è stato stabilito che i cacciatori possano uccidere al più* 50 *esemplari l'anno durante il periodo venatorio. Sia* c_n *il numero di cinghiali all'*$n-$*esimo anno e si supponga che il loro numero iniziale sia di* 350 *esemplari.*

1. *Scrivere l'equazione alle differenze a coefficienti costanti che descrive la variazione della popolazione di cinghiali di anno in anno, supposto che le morti siano esattamente* 50 *all'anno;*
2. *nelle ipotesi al punto 1, sulla base del modello numerico, fornire informazioni sull'andamento della popolazione nel lungo periodo;*
3. *stabilire cosa accade nel lungo periodo se il numero delle morti è superiore a* 50 *all'anno;*
4. *costruire una function MatLab che, dati in ingresso il numero di anni* N *e il numero di uccisioni per anno* U*, restituisca in output il numero di cinghiali dopo* N *anni, mostri in un grafico l'andamento della popolazione e costruisca una tabella che riporti in prima colonna il vettore degli anni e in seconda colonna il vettore contenente il numero di cinghiali all'*$n-$*esimo anno.*

2.5.2 Equazioni alle differenze del secondo ordine

Esercizio 2.10 *Risolvere le seguenti equazioni alle differenze del secondo ordine omogenee*

1. $y_{n+2} - y_n = 0, \qquad y_0 = 0, \quad y_1 = 2 \qquad \left[y_n = 1 + (-1)^{n+1}\right]$
2. $y_{n+2} - 6y_{n+1} + 8y_n = 0, \qquad y_0 = 3, \quad y_1 = 2 \qquad \left[y_n = 5 \cdot 2^n - 2 \cdot 4^n\right]$
3. $y_{n+1} - 4y_n + 4y_{n-1} = 0, \qquad y_{-1} = 1, \quad y_0 = 3 \qquad \left[y_n = 2^n\,(3+n)\right]$
4. $y_n - y_{n-1} + \frac{1}{4}y_{n-2} = 0, \qquad y_{-1} = 1 \quad y_{-2} = 0 \qquad \left[y_n = \left(\frac{1}{2}\right)^n + n\left(\frac{1}{2}\right)^{n+1}\right]$
5. $3y_{n+2} - 3y_{n+1} - 6y_n = 0, \qquad y_0 = -1 \quad y_1 = 2 \qquad \left[y_n = \frac{1}{3}2^n - \frac{4}{3}\cdot(-1)^n\right]$
6. $\frac{1}{4}y_{n+1} + \frac{1}{2}y_n - 2y_{n-1} = 0, \qquad y_{-1} = 0 \quad y_0 = 2 \qquad \left[y_n = \frac{1}{3}\left(2^{n+1} + (-1)^n\,4^{n+1}\right)\right]$

Esercizio 2.11 *Risolvere le seguenti equazioni alle differenze del secondo ordine non omogenee*

1. $y_n - y_{n-2} = 1, \qquad y_{-1} = 0, \quad y_{-2} = 0 \qquad \left[y_n = \frac{3}{4} + \frac{1}{4}(-1)^n + \frac{1}{2}n\right]$
2. $y_{n+1} - 6y_n + 8y_{n-1} = 2, \qquad y_0 = 2, \quad y_1 = 1 \qquad \left[y_n = \frac{15}{6}2^n - \frac{7}{6}4^n + \frac{2}{3}\right]$
3. $y_{n+2} - 4y_{n+1} + 4y_n = 3, \qquad y_0 = 1, \quad y_1 = 3 \qquad \left[y_n = 2^{n+1}\,(n-1) + 3\right]$
4. $y_n - y_{n-1} + \frac{1}{4}y_{n-2} = 1, \qquad y_{-1} = 2, \quad y_{-2} = 0 \qquad \left[y_n = -\left(\frac{1}{2}\right)^n + 4\right]$
5. $3y_{n+2} - 3y_{n+1} - 6y_n = -1, \qquad y_0 = 1, \quad y_1 = 1 \qquad \left[y_n = \frac{5}{18}(-1)^n + \frac{5}{9}\cdot 2^n + \frac{1}{6}\right]$
6. $\frac{1}{4}y_{n+1} - \frac{1}{2}y_n - 2y_{n-1} = -\frac{3}{4}, \qquad y_{-1} = 0, \quad y_0 = 2 \qquad \left[y_n = (-2)^n + \frac{2}{3}\cdot 4^n + \frac{1}{3}\right]$

Capitolo 3

Metodi numerici

In matematica l'arte di porre problemi deve essere tenuta in maggiore considerazione di quella di risolverli.
Georg Cantor (1845-1918)

3.1 Introduzione

Con il termine "Analisi Numerica" si intende lo sviluppo e lo studio di procedure che permettono di risolvere un problema mediante una sequenza di operazioni aritmetiche.
In genere è coinvolta una gran quantità di calcoli, il che rende di fatto inattuabili le procedure,in assenza di opportuni strumenti di calcolo.
La sequenza ordinata di operazioni che conducono alla soluzione del problema prende il nome di **algoritmo**.
In presenza di un problema che trova origine nella realtà sensibile, il primo passo da affrontare nella ricerca di una soluzione è la riformulazione del problema stesso in "modello matematico".

Dopo aver costruito il modello si procede a definire delle strategie risolventi o degli algoritmi, che a partire dalla situazione iniziale producono una soluzione matematica del problema stesso.
I problemi creati nell'ambito della modellistica matematica non sono quasi mai risolvibili per via analitica, pertanto è necessario formulare per essi opportune procedure che determinano la soluzione per via approssimata.

Nella maggior parte dei casi la soluzione di un modello porta alla stesura di diverse procedure risolventi; l'analista numerico non si limita alla realizzazione dell'algoritmo risolvente ma studia quale tra gli algoritmi sia il più "efficiente". L'efficienza di un algoritmo si stabilisce sulla base del numero di operazioni richieste, del tempo necessario ad arrivare alla soluzione, della quantità di memoria usata (nel caso di uso di un computer), mentre altri importantissimi parametri nella definizione del miglior algoritmo sono la "stabilità" e la "convergenza", parametri strettamente connessi al concetto di "errore"
Metodi numerici sono utilizzati anche nei casi in cui esiste la soluzione analitica del modello formulato, ma risultano particolarmente importanti quando il problema, anche semplice, non ammette una soluzione analitica.
Ad esempio è noto che la lunghezza di una arco della curva $y = \sin(x)$, si ottiene dal calcolo dell'integrale definito

$$\int_0^{\pi} \sqrt{1 + \cos^2(x)}\, dx,$$

per il quale non esiste una soluzione in forma chiusa.

Per calcolare la lunghezza dell'arco si applicano tecniche standard che permettono di calcolare il valore di un integrale definito qualunque sia la funzione integranda, purchè integrabile, e sia noto il valore che essa assume in opportuni punti dell'intervallo di definizione. Le sole operazioni richieste sono le quattro operazioni aritmetiche e precedure di confronto, cioè esattamente le operazioni che un computer è in grado di effettuare.
Una soluzione fornita in analisi numerica è sempre numerica; una soluzione analitica in genere è data in termini di funzioni matematiche di cui è, di solito, immediato analizzare il comportamento; ciò in generale non è vero per soluzioni numeriche.
Una soluzione numerica è accettabile quando teoricamente l'approssimazione può essere resa accurata quanto si vuole e quando è possibile in ogni istante sapere di quanto la soluzione approssimata differisca dalla soluzione esatta.
Normalmente per raggiungere un elevato livello di accuratezza è necessario coinvolgere una gran mole di calcoli; poichè ciascuna operazione effettuata è generalmente affetta da "errore", in eleborazioni molto complesse l'effetto di tali errori può essere pesante e influire sulla bontà della soluzione, pertanto lo studio degli errori e le varie fonti di errori in un calcolo costituiscono un capitolo fondamentale per avvicinarsi alle tecniche dell'Analisi Numerica.

La sequenza di passi che portano dal riconoscimento del problema ad una soluzione "accettabile" può essere sintetizzata nel seguente schema:

- definizione del problema;
- formulazione di un modello matematico;
- stesura di un algoritmo risolvente il modello stesso;
- esecuzione dell'algoritmo (in genere questo passo consiste nella traduzione dell'algoritmo stesso in un linguaggio di programmazione e nell'esecuzione del codice ottenuto);
- analisi e interpretazione dei risultati ottenuti.

Esempio 3.1

- ***Definizione del problema:***
  si vuole calcolare la lunghezza di una arco della curva $y = \sin(x)$*;*
- ***formulazione di un modello matematico:***
  il modello è rappresentato dall'integrale
  $$I = \int_0^{\pi} \sqrt{1 + \cos^2(x)}\, dx;$$
- ***stesura di un algoritmo risolvente il modello:***
  nell'intervallo $[0, \pi]$ *si considerano i punti* $x_j = 0 + jh, \quad j = 0, 1, \ldots, n$*, equidistanti a passo* h*; il valore dell'integrale è approssimato dal valore*
  $$\bar{I} = h \sum_{k=0}^{n-1} \frac{f(x_{k+1}) + f(x_k)}{2} = \frac{h}{2}\Big[f(x_0) + f(x_n) + 2\sum_{k=1}^{n-1} f(x_k)\Big];$$
- ***esecuzione dell'algoritmo:***
  traduzione dell'algoritmo in un linguaggio di programmazione ed esecuzione del codice;
- ***analisi e interpretazione dei risultati ottenuti:***
  si controlla il risultato ottenuto e se ne analizza l'attendibilità. Ad esempio si può pensare di calcolare più valori di $\bar{I}$ *per valori diversi di* h*, sempre più piccoli, e confrontare i risultati ottenuti.*

Lo studio completo di un metodo numerico, oltre ai passi descritti nel paragrafo precedente, deve comprendere uno studio teorico della convergenza. Tale studio è connesso alla definizione di "errore".

3.2 Il calcolo numerico

Scopo del Calcolo Numerico è la conoscenza numerica di **numeri** e **funzioni**, che si acquisisce attraverso **gli strumenti del calcolo numerico** ossia

- metodi numerici o algoritmi (software);
- mezzi di calcolo (hardware).

I due strumenti non sono indipendenti: spesso gli algoritmi sono in funzione dei mezzi di calcolo.

Un **metodo numerico** è un procedimento che, con un numero finito di operazioni elementari, fornisce la soluzione esatta, o la soluzione approssimata con indicazione della precisione, del problema. Gli strumenti alla base di un metodo numerico sono

- formule ricorrenti;
- procedimenti di successive approssimazioni.

3.3 Formule ricorrenti

Una *formula ricorrente* è un legame tra termini successivi di una particolare sequenza di elementi, che consente il calcolo di quantità successive in funzione di quantità precedentemente calcolate.

$$\begin{cases} t_0 \quad \text{assegnato} \\ t_k = f\left(t_{k-1}\right) \end{cases} \qquad \text{ricorrente a un passo}$$

$$\begin{cases} t_0, t_1, \ldots, t_{p-1} \quad \text{assegnati} \\ t_k = f\left(t_{k-1}, t_{k-2}, \ldots, t_{k-p}\right) \quad k = p, p+1, \ldots \end{cases} \qquad \text{ricorrente a p passi}$$

Esempio 3.2 *Formula ricorrente ad un passo*

$$\begin{cases} t_0 = \dfrac{a}{2} \qquad a > 0 \\ t_k = \dfrac{1}{2}\left(t_{k-1} + \dfrac{a}{t_{k-1}}\right) \qquad k = 1, 2, \ldots \qquad \left(t_k \to \sqrt{a}\right) \end{cases}$$

Esempio 3.3 *Formula ricorrente a 2 passi (sequenza di Fibonacci)*

$$\begin{cases} F_0 = F_1 = 1 \\ F_k = F_{k-1} + F_{k-2} \qquad k = 2, 3, \ldots \end{cases}$$

Esempio 3.4 ***(Somma e prodotto di** n **numeri)**. Si vuole calcolare la somma*

$$S = a_1 + a_2 + \ldots + a_n = \sum_{i=1}^{n} a_i \qquad a_i \in R.$$

Si trasforma il problema in forma ricorrente, ponendo

$$\begin{aligned} &S_1 = a_1 \\ &S_2 = a_1 + a_2 = S_1 + a_1 \\ &\vdots \\ &S_k = (a_1 + a_2 + \ldots + a_{k-1}) + a_k = S_{k-1} + a_k, \qquad k = 2, \ldots, n \end{aligned}$$

da cui si ricava

$$S = S_n,$$

si calcola cioè la somma S richiesta mediante la sequenza ricorrente ad un passo

$$\begin{cases} S_1 = a_1 \\ S_k = S_{k-1} + a_k \quad k = 2, \ldots, n \end{cases}$$

Se inoltre si pone $S_0 = 0$, neutro per la somma che non influisce sul calcolo di S_1, la sequenza diventa

$$\begin{cases} S_0 = 0 \\ S_k = S_{k-1} + a_k \quad k = 1, \ldots, n \end{cases}$$

L'algoritmo espresso così in forma ricorrente presenta il vantaggio di essere facilmente programmabile

In maniera del tutto analoga si può definire l'algoritmo per il calcolo del prodotto

$$P = a_1 \cdot a_2 \cdot \ldots \cdot a_n = \prod_{i=1}^{n} a_i \qquad a_i \in R,$$

mediante la sequenza ricorrente ad un passo

$$\begin{cases} P_0 = 1 \\ P_k = P_{k-1} \cdot a_k \quad k = 1, \ldots, n, \end{cases}$$

da cui

$$P = P_n.$$

3.4 Procedimenti di successive approssimazioni

Un *procedimento di successive approssimazioni* consiste nel sostituire il problema originario con una sequenza di problemi approssimanti, facilmente e comunque risolvibili. In sostanza il procedimento di successive approssimazioni è l'analogo del passaggio al limite, quando si lavora con variabile discreta anziché continua. L'enorme mole di calcoli coinvolta rende un procedimento di successive approssimazioni di fatto eseguibile solo con opportuni mezzi di calcolo.

Un esempio di procedimento di successive approssimazioni è dato dal metodo di Archimede per il calcolo di π, che sarà discusso in seguito.

3.5 Conoscenza numerica di numeri

Un numero è conosciuto numericamente, se si dispone della sua scrittura decimale, ossia della sua rappresentazione in base 10. Da questo punto di vista, potrebbero essere noti solo i numeri razionali, per cui, in generale, ci si accontenta della conoscenza di un certo numero di cifre decimali, con indicazione della precisione

$$e = 2.7183 \pm 10^{-4}, \qquad \pi = 3.14 \pm 10^{-2}$$

ma, ricordando che $\pi \approx 3.14159265$, sarà anche vero

$$\pi = 3.142735 \pm 10^{-2}.$$

Diremo quindi che un numero è conosciuto numericamente se è noto il numero di cifre decimali esatte, supponendo di disporre di un procedimento che teoricamente consente di calcolare un numero qualsiasi di cifre decimali con indicazione della precisione raggiunta (nel secondo esempio sei cifre decimali con un errore sulla terza (10^{-2}).

Esempi:

- *Radici di equazioni o di sistemi di equazioni:* data una funzione f, trovare $x \in \mathbb{C}$ tale che $f(x) = 0$;
- *Soluzione di sistemi lineari:* dati $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^n$ trovare $x \in \mathbb{R}^n$ tale che $Ax = b$;
- *Valutazione di polinomi:* dato $\bar{x} \in \mathbb{R}$ calcolare $P(\bar{x}) = a_0\bar{x}^n + a_1\bar{x}^{n-1} + \ldots + a_n$;
- *Calcolo di autovalori:* trovare $\lambda \in \mathbb{C}$ per il quale esiste $x \in \mathbb{R}^n$, $x \neq 0$, tale che $Ax = \lambda x$, $A \in \mathbb{R}^{n\times n}$;
- *Successioni e serie:* $\gamma = \lim_{n\to\infty}\left(1 + \frac{1}{2} + \ldots + \frac{1}{n} - \log n\right)$; $\qquad \alpha = \sum_{n=0}^{\infty}(-1)^n \frac{1}{(2n+1)!}$;
- *Integrali:* $\alpha_1 = \int_0^1 \sqrt{1+x^3}\, dx$; $\qquad \alpha_2 = \int_1^{2.7} e^{-x^2}\, dx$;

3.5.1 Il metodo di Archimede per il calcolo di π

Il metodo di Archimede per il calcolo di π è un esempio di procedimento di successive approssimazioni.

Figura 3.1 - Archimede (287a.C.-212a.C.)

Il simbolo π indica il rapporto costante tra la lunghezza della circonferenza e il suo diametro.
Uno dei primi risultati significativi sul calcolo effettivo di π è dovuto ad Archimede. Egli riuscì a dimostrare che

$$\frac{223}{71} = 3 + \frac{10}{71} < \pi < 3 + \frac{1}{7} = \frac{22}{7} \tag{3.1}$$

approssimando la lunghezza della circonferenza con il perimetro di poligoni regolari iscritti e circoscritti. Il perimetro di un poligono inscritto è sempre minore della lunghezza della circonferenza che a sua volta è minore del perimetro del poligono circoscritto. Se p_1, p_2, C indicano rispettivamente le misure dei perimetri del poligono inscritto, circoscritto e della circonferenza si ha

$$p_1 < C < p_1$$

Sostituendo nella formula $\pi = \frac{C}{2r}$, dove r è la misura del raggio si trova la relazione

$$\pi_1 = \frac{p_1}{2r} < \pi < \frac{p_2}{2r} = \pi_2$$

cioè π_1 è un'approssimazione per difetto di π, così come π_2 lo è per eccesso.
L'approssimazione sarà tanto migliore, quanto maggiore è il numero dei lati del poligono. Il problema quindi si riconduce a calcolare il perimetro dei poligoni inscritti e circoscritti al cerchio, in funzione del raggio.

Teorema 3.1 *Se l_n indica il lato del poligono regolare di n lati inscritto in una circonferenza di raggio 1 (per semplicità) il lato del poligono di $2n$ lati, l_{2n}, è dato da*

$$l_{2n} = \sqrt{2 - \sqrt{4 - l_n^2}}$$

Dimostrazione
In riferimento alla figura 3.2, dato $AB = l_n$ si vuole ricavare $BH = l_{2n}$.

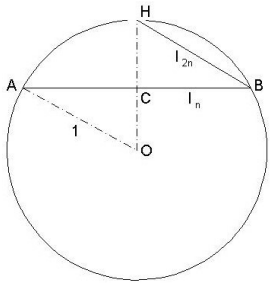


Figura 3.2

Applicando il teorema di Pitagora al triangolo BHC

$$BH^2 = BC^2 + HC^2 \qquad \Rightarrow \qquad l_{2n}^2 = \left(\frac{l_n}{2}\right)^2 + HC^2$$

D'altra parte

$$HC^2 = (OH - OC)^2 = (1 - OC)^2 = 1 + OC^2 - 2OC$$

$$OC^2 = AO^2 - AC^2 = 1 - \left(\frac{l_n}{2}\right)^2$$

quindi

$$\begin{aligned} l_{2n}^2 &= \left(\frac{l_n}{2}\right)^2 + HC^2 = \left(\frac{l_n}{2}\right)^2 + 1 + OC^2 - 2OC = \\ &= \left(\frac{l_n}{2}\right)^2 + 1 + 1 - \left(\frac{l_n}{2}\right)^2 - 2OC = 2 - 2OC = 2 - \sqrt{4 - l_n^2} \end{aligned}$$

da cui la tesi.

Il teorema 3.1, noto il lato di un qualsiasi poligono inscritto (poligono iniziale), consente di calcolare il lato di tutti i poligoni inscritti con numero dei lati multiplo, secondo una potenza di 2, del numero di lati del poligono iniziale. Ricordando che il lato dell'esagono inscritto in una circonferenza è uguale al raggio r della circonferenza stessa, si ha (per semplicità si sceglie di lavorare con $r = 1$)

$$\begin{aligned} l_6 &= 1 \\ l_{12} &= \sqrt{2 - \sqrt{3}} \\ l_{24} &= \sqrt{2 - \sqrt{2 + \sqrt{3}}} \\ l_{48} &= \sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{3}}}}. \end{aligned}$$

Se $\pi_{1,n}$ è l'approssimazione di π relativa al poligono inscritto di n lati si trova

$$\pi_{1,n} = \frac{n l_n}{2} < \pi$$

Per calcolare il lato del poligono circoscritto vale il seguente

Teorema 3.2 *Se l_n indica il lato del poligono regolare di n lati inscritto in una circonferenza di raggio 1, il lato L_n del poligono circoscritto di uguale numero di lati è dato da*

$$L_n = \frac{2l_n}{\sqrt{4 - l_n^2}}$$

Dimostrazione
In riferimento alla figura 3.3, dato $AB = l_n$ si vuole ricavare $TV = L_n$ (si considera sempre la circonferenza di raggio 1).

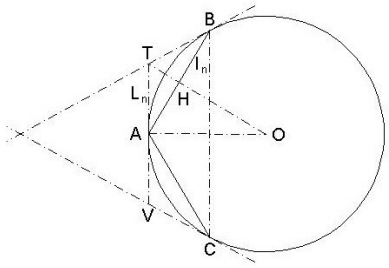


Figura 3.3

Considerando il triangolo OAT si ha

$$AT \cdot OA = AH \cdot OT$$

da cui ponendo $x = AT$ e ricordando che $OA = 1$

$$OT = \frac{AT \cdot OA}{AH} = \frac{x}{\frac{l_n}{2}} = \frac{2x}{l_n}.$$

D'altra parte applicando ancora il teorema di Pitagora al triangolo OAH e osservando che $AH = \frac{l_n}{2}$, si trova

$$OH = \sqrt{1 - \left(\frac{l_n}{2}\right)^2}$$

e quindi

$$TH = OT - OH = \frac{2x}{l_n} - \sqrt{1 - \left(\frac{l_n}{2}\right)^2}.$$

In riferimento ora al triangolo rettangolo ATH

$$\begin{aligned} AT^2 = AH^2 + TH^2 \quad \Rightarrow \quad x^2 &= \left(\frac{l_n}{2}\right)^2 + \left(\frac{2x}{l_n} - \sqrt{1 - \left(\frac{l_n}{2}\right)^2}\right)^2 = \\ &= \frac{l_n^2}{4} + \frac{4x^2}{l_n^2} + \left(1 - \frac{l_n^2}{4}\right) - \frac{4x}{l_n}\sqrt{1 - \left(\frac{l_n}{2}\right)^2} \end{aligned}$$

da cui

$$\left(\frac{4}{l_n^2} - 1\right) x^2 - \frac{2x}{l_n}\sqrt{4 - l_n^2} + 1 = 0$$

Moltiplicando per l_n^2 ambo i membri l'espressione diventa

$$\left(4 - l_n^2\right) x^2 - 2xl_n\sqrt{4 - l_n^2} + l_n^2 = \left(x\sqrt{4 - l_n^2} - l_n\right)^2 = 0$$

da cui

$$x = \frac{l_n}{\sqrt{4 - l_n^2}},$$

ed essendo $L_n = 2x$ la tesi

$$L_n = \frac{2l_n}{\sqrt{4 - l_n^2}}.$$

Segue dal teorema 3.2 che, se $\pi_{2,n}$ è l'approssimazione di π relativa al poligono circoscritto di n lati, si ha

$$\pi < \pi_{2,n} = \frac{nL_n}{2} = \frac{nl_n}{\sqrt{4 - l_n^2}}$$

Dalle espressioni ricavate per $\pi_{1,n}$ e $\pi_{2,n}$ si trova

$$\pi_{1,n} = \frac{nl_n}{2} < \pi < \frac{nl_n}{\sqrt{4 - l_n^2}} = \pi_{2,n}$$

Su tale relazione si può costruire un algoritmo che calcola π, teoricamente con la precisione voluta, ma in pratica dipendente dalla precisione di macchina e da problemi di instabilità numerica.

3.6 Conoscenza numerica di funzioni

Diremo che una funzione è conosciuta numericamente se è noto, nel senso precisato riguardo ai numeri, il valore che essa assume in ogni punto dell'insieme di definizione.
Esistono, nella pratica, tre situazioni familiari:

- *funzione assegnata mediante una tavola o un algoritmo già disponibile in macchina:* è il caso delle funzioni "elementari", x^α $\alpha \in \mathbb{R}$, $\log x$, $\sin x$, $\cos x$, ...e in tal caso si dirà che si dispone di una pseudo-conoscenza esplicita;

- *funzione assegnata mediante uno sviluppo in serie di funzioni elementari:*

$$f(x) = \sum_{n=0}^{\infty} a_n u_n(x)$$

  (es. sviluppo in serie di Taylor se le derivate successive di $f(x)$ sono funzioni elementari oppure sviluppi trigonometrici etc.) allora $f(x)$ è conosciuta numericamente se è noto un numero finito di a_n (supponendo di poterne conoscere un numero qualsiasi).

- *funzione assegnata per punti:* in tal caso per il calcolo della funzione in un punto diverso da quelli noti si procederà per interpolazione.

Accanto a queste funzioni, che possono ritenersi più o meno note esplicitamente, esistono numerosi esempi di funzioni assegnate implicitamente, il cui calcolo è da analizzare:

funzioni speciali:

- funzione errore $erf(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2}\, dt$; funzione $\Gamma(x)$; funzioni di Bessel; funzioni di Legendre, etc.;

soluzione di equazioni differenziali ordinarie e alle derivate parziali;

soluzione di equazioni integrali;

problemi di minimizzazione.

3.6.1 Calcolo di $\sin t$ (Algoritmo A)

Si vuole ora costruire un metodo numerico che dato l'angolo t, calcoli il valore approssimato di $\sin t$. Tale metodo è denominato *algoritmo A* per distinguerlo da un altro algoritmo, presentato in seguito, che risolve lo stesso problema e che sarà indicato come *algoritmo T*.

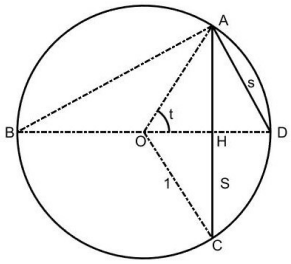


Figura 3.4

Si considera l'arco corrispondente all'angolo t, identificato anche con t; il metodo consiste nel passare dalla corda s sottesa all'arco t, alla corda S sottesa all'arco $2t$. Se si immagina di dividere l'arco $2t$ in 2^n parti uguali, con n sufficientemente grande, si può considerare la lunghezza dell'arco "uguale" alla lunghezza della corda. Usando questa come corda di partenza si itera il procedimento "di raddoppio" fino a calcolare la corda S, che risulta essere $2\sin t$ e che si ottiene esattamente dopo n "raddoppi".
Dalla figura 3.4, applicando il teorema di Pitagora al triangolo BAD, si trova

$$AB = \sqrt{BD^2 - s^2} = \sqrt{4 - s^2},$$

inoltre

$$2\ areaABD = AD \cdot AB = s\sqrt{4 - s^2}$$

(per semplicità di calcolo si considera una circonferenza di raggio 1). D'altra parte

$$2\ areaABD = 2AH = 2\sin t = S,$$

quindi in definitiva

$$S = 2\sin t = s\sqrt{4 - s^2}$$

3.6.2 Calcolo di $\sin t$ (Algoritmo T)

Questo algoritmo, dovuto a Tolomeo, si basa su considerazioni analoghe al precedente.

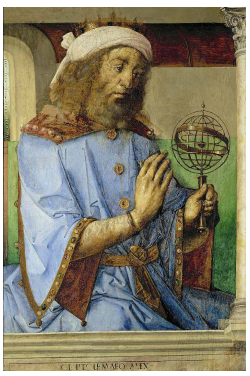


Figura 3.5 - Tolomeo (100–175)

Si considera l'arco θ, corrispondente a un generico angolo θ. Il metodo consiste nell'applicare una formula che permette di passare dalla corda s sottesa all'arco θ, alla corda S sottesa all'arco 3θ.

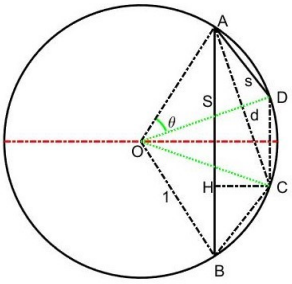


Figura 3.6

Se si immagina ora di dividere l'arco $2t$ in 3^n parti uguali, con n sufficientemente grande, si può considerare la lunghezza dell'arco "uguale" alla lunghezza della corda di partenza. Si itera il procedimento "di triplicazione" fino a calcolare la corda $S \equiv 2\sin t$, che si ottiene dopo n "triplicazioni".

Teorema 3.3 (di Tolomeo) *In un trapezio isoscele il prodotto delle diagonali è pari alla somma dei prodotti dei lati opposti.*

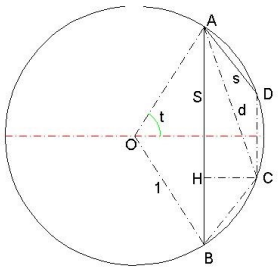


Figura 3.7

In riferimento alla figura 3.7, il teorema di Tolomeo afferma che

$$d^2 = s^2 + sS.$$

Dimostrazione

$$AH = \frac{AB + DC}{2} = \frac{S + s}{2}$$

Applicando il teorema di Pitagora al triangolo AHC

$$d^2 = AC^2 = AH^2 + HC^2 = \left(\frac{S+s}{2}\right)^2 + HC^2,$$

Osservando che

$$HB = \frac{AB - DC}{2} = \frac{S - s}{2}$$

e applicando il teorema di Pitagora al triangolo BHC

$$HC^2 = BC^2 - HB^2 = s^2 - \left(\frac{S-s}{2}\right)^2$$

quindi

$$d^2 = \left(\frac{S+s}{2}\right)^2 + HC^2 \quad = \left(\frac{S+s}{2}\right)^2 + s^2 - \left(\frac{S-s}{2}\right)^2 =$$

$$= \frac{S^2}{4} + \frac{sS}{2} + \frac{s^2}{4} + s^2 - \frac{S^2}{4} + \frac{sS}{2} - \frac{s^2}{4}$$

cioè in definitiva

$$d^2 = s^2 + sS$$

D'altra parte d risulta essere la corda sottesa all'arco $2s$ e quindi, ricordando l'*algoritmo A*,

$$d^2 = s^2 \left(4 - s^2\right).$$

Uguagliando i due risultati si trova

$$s^2 + sS = s^2 \left(4 - s^2\right).$$

cioè

$$S = \frac{3s^2 - s^4}{s} = 3s - s^3$$

che è la formula che permette di passare dalla corda s, sottesa all'arco t, alla corda S, sottesa all'arco $3t$.

Capitolo 4

Teoria degli errori

I numeri governano il mondo.
Platone (428a.C.-348a.C.)

La presenza di errori numerici dipende dall'uso dell'approssimazione per rappresentare operazioni e quantità matematiche "esatte"; ciò è dovuto prevalentemente all'uso di un'aritmetica finita, propria degli strumenti di calcolo.
Altre fonti di errore possono verificarsi a prescindere da quelli dovuti all'uso del calcolatore.

4.1 Sorgenti di errore

I risultati numerici possono essere influenzati da diversi tipi di errori

- **Errori inerenti**:
  - **errori nei dati iniziali**: i dati iniziali di un problema sono generalmente il risultato di misurazioni sperimentali. Essi possono essere dovuti alla sensibilità dello strumento di misurazione o al verificarsi di eventi "imprevedibili" nel corso della misurazione. Possono essere controllati sottoponendo il modello a molti test e verificando la sensibilità della soluzione rispetto a variazioni degli input;
  - **semplificazioni introdotte nel modello:** il modello di norma non rappresenta esattamente la situazione reale.
- **Errori acquisiti**:
  - **errori di arrotondamento:** sono dovuti alla rappresentazione dei numeri in macchina;
  - **errori di propagazione:** sono dovuti alla rappresentazione dei numeri e dipendono dall'accumulo degli errori di arrotondamento nel corso di un processo di calcolo;
  - **errori di troncamento:** sono relativi al metodo di approssimazione usato e generalmente corrispondono al troncamento di un procedimento di successive approssimazioni. In ogni caso vanno analizzati relativamente ad ogni metodo utilizzato.

4.2 Errore assoluto e errore relativo

Dati due numeri reali α e $\widetilde{\alpha}$, se $\widetilde{\alpha}$ è un'approssimazione di α, si definisce in prima istanza come errore la quantità

$$E = |\alpha - \widetilde{\alpha}| .$$

Tale definizione non è però sufficiente a dare un'indicazione della bontà dell'approssimazione ricavata. Infatti siano

a) $\alpha = 2021 \quad \widetilde{\alpha} = 2020 \quad E = |\alpha - \widetilde{\alpha}| = 1$

b) $\alpha = 10 \quad \widetilde{\alpha} = 9 \quad E = |\alpha - \widetilde{\alpha}| = 1.$

In entrambi i casi l'errore vale 1, ma l'approssimazione nel secondo caso è molto meno accettabile che nel primo, (l'errore è dello stesso ordine di grandezza dei dati).
Per avere una stima più significativa si rende necessaria una definizione di errore che tenga conto del rapporto tra l'errore stesso e i dati in esame.

Definizione 4.1. *Siano* $\alpha,\ \widetilde{\alpha} \in \mathbb{R}$ *e* $\widetilde{\alpha} \approx \alpha$. *Si definisce* **errore assoluto** *la quantità*

$$E_\alpha = |\alpha - \widetilde{\alpha}|\,.$$

Si definisce invece **errore relativo** *la quantità*

$$RE_\alpha = \frac{|\alpha - \widetilde{\alpha}|}{|\alpha|}$$

o equivalentemente, poiché in generale non è noto $|\alpha|$ *ed essendo* β *un'approssimazione di* α*, quindi dello stesso ordine di grandezza,*

$$RE_\alpha = \frac{|\alpha - \widetilde{\alpha}|}{|\widetilde{\alpha}|}.$$

Tali definizioni hanno un valore puramente teorico essendo generalmente non noto il valore esatto.

Negli esempi considerati per l'errore relativo si ottengono le quantità

a) $RE = 1/2021 \approx 5 * 10^{-4}$

b) $RE = 1/10 = 1 * 10^{-1}$.

Risulta ora chiaro che nel secondo caso l'errore relativo è molto più grande che nel primo.
Per avere una stima della bontà di un'approssimazione si rendono necessarie ulteriori definizioni.

Definizione 4.2. *Si dice che un numero reale* x *ha* ***ordine di grandezza*** 10^k *con* $k \in \mathbb{Z}$*, se può essere scritto nella forma*

$$x = d \cdot 10^{k+1} \text{ con } 0.1 \leq |d| < 1$$

Ad esempio

$85.7 = 0.857 \cdot 10^2$ ha ordine di grandezza 10

$0.125 = 0.125 \cdot 10^0$ ha ordine di grandezza 10^{-1}.

Il concetto di errore relativo è collegato a quello di cifre significative esatte.

Definizione 4.3. *Se l'errore relativo* $RE = \dfrac{x - \bar{x}}{x}$ *ha ordine di grandezza* 10^{-k} $(k > 0)$ *si dice che* $\bar{x}$ *ha* $k - 1$ ***cifre significative esatte*** *rispetto ad* x.

Esempio 4.1 *Sia*

a) $x = 3.1415926, \qquad \bar{x} = 3.1428571$

b) $y = 0.2138 \cdot 10^{-1}, \qquad \bar{y} = 0.2144 \cdot 10^{-1}.$

Nel caso a)

$$E = x - \bar{x} = -0.12645 \cdot 10^{-2}, \qquad RE = \frac{x - \bar{x}}{x} \approx -0.4025028 \cdot 10^{-3}$$

ordine di grandezza: $10^{-4} \quad \Rightarrow \quad \bar{x}$ *ha* 3 *cifre significative esatte* .

Nel caso b)

$$E = y - \bar{y} = -0.6 \cdot 10^{-4}, \qquad RE = \frac{y - \bar{y}}{y} \approx -0.2806361 \cdot 10^{-2}$$

ordine di grandezza: $10^{-3} \quad \Rightarrow \quad \bar{y}$ *ha* 2 *cifre significative esatte* .

4.3 Numeri e calcoli in virgola mobile

La natura stessa dei numeri impone di operare con un numero finito di cifre. Al fine di poter esprimere la precisione con la quale è stato effettuato un calcolo, conviene fissare a priori l'aritmetica finita, secondo cui operare.
A tal proposito si osserva che ogni numero decimale può essere espresso nella forma

$$x = a \cdot 10^c$$

con $|a| < 10$.

$$354 = 3.54 \cdot 10^2; \qquad 50.78 = 5.078 \cdot 10$$
$$254.7 = 2.547 \cdot 10^2; \qquad 0.0071 = 7.1 \cdot 10^{-3}.$$

Inoltre si può scegliere opportunamente il valore di c in modo tale che risulti

$$10^{-1} = \frac{1}{10} \leq |a| < 10^0 = 1.$$

$$354 = 0.354 \cdot 10^3; \qquad -50.78 = -0.5078 \cdot 10^2$$
$$0.0095 = 0.95 \cdot 10^{-2}; \qquad -0.00295 = -0.295 \cdot 10^{-2}.$$

Tale operazione può essere definita in un sistema di numerazione in base b qualsiasi. Infatti, ogni numero può essere espresso, in base b, sempre nella forma

$$x = a \cdot b^c$$

con $|a| < b$, e si può fare in modo che $b^{-1} \leq |a| < 1$.

Ciò consente di dare la seguente

Definizione 4.4. *Un numero espresso nella forma*

$$x = a \cdot 10^c$$

o più in generale, nella forma

$$x = a \cdot b^c$$

con $|a| < b$*, dicesi in* ***virgola mobile (floating point)****.*
Se inoltre $b^{-1} \leq |a| < 1$ *il numero dicesi in* ***virgola mobile normalizzata****; a si dice* ***mantissa*** *e c* ***caratteristica*** *del numero dato.*

Utilizzando la rappresentazione in virgola mobile, si fissa a priori il numero, m, di cifre della mantissa, ed il numero, n, di cifre per la caratteristica, cioè

$$\begin{aligned} |a| &= 0.\alpha_1\alpha_2\ldots\alpha_m \qquad 0 \le \alpha_i \le 9, \qquad \alpha_1 \neq 0, \qquad \text{in base } b = 10 \\ c &= \beta_1\ldots\beta_k \qquad\qquad k \in \{1,\ldots,n\}\,. \end{aligned}$$

Fissati i numeri m ed n, cioè il numero di cifre della mantissa e della caratteristica, e quindi rappresentando i numeri in virgola mobile, si fissa l'aritmetica finita in cui operare.
Lavorando in aritmetica finita sorge il seguente problema. Sia, ad esempio, $m = 4$, cioè 4 cifre di mantissa, si eseguono i seguenti calcoli (nel sistema decimale)

$$\begin{aligned} &\text{a)} \quad x = 0.4326 \cdot 10^3, \quad y = 0.5129 \cdot 10^2, \qquad z = x \cdot y \quad = \quad 0.22188054 \cdot 10^5 \\ &\text{b)} \quad x = 0.6123 \cdot 10^4, \quad y = 0.5243 \cdot 10^4, \qquad z = x + y \quad = \quad 0.11366 \cdot 10^5. \end{aligned}$$

In entrambi i casi il risultato contiene più di quattro cifre di mantissa. Qual è allora il risultato nell'aritmetica finita considerata ($m = 4$ cifre di mantissa)?

4.3.1 Precisione di macchina

Quando si lavora in aritmetica finita è importante sapere qual è la più piccola differenza tra due numeri che la macchina è in grado di riconoscere, o qual è il più piccolo numero considerato diverso da zero.
Tale quantità è detta ε *macchina* o *precisione di macchina*. Esso fornisce una misura precisa di quante cifre significative sono possibili nella rappresentazione floating-point di un numero.
L'ε *macchina* varia da calcolatore a calcolatore in dipendenza dall'aritmetica finita usata.

4.4 Errori in un metodo numerico

4.4.1 Errore di arrotondamento

Nei calcolatori digitali i numeri sono rappresentati in virgola mobile normalizzata con un numero fissato di cifre per la mantissa e la caratteristica; ne segue che l'insieme dei numeri rappresentabili esattamente in una macchina calcolatrice è finito. Tale insieme sarà indicato con $\mathcal{A}$.
Se un numero reale x non appartiene ad $\mathcal{A}$, occorrerà considerare un valore che sarà indicato con $fl\,(x)$, tale $fl\,(x) \approx x$ e $fl\,(x) \in \mathcal{A}$.

Si suppone di disporre di una macchina con m cifre di mantissa e c di caratteristica. Sia $x \in \mathbb{R}$ e $x \notin \mathcal{A}$ (x ha un espressione decimale con più di m cifre di mantissa). x può essere scritto come

$$x = a \cdot b^t, \qquad\qquad b^{-1} \le |a| < 1$$

e

$$a = \pm 0.\alpha_1\alpha_2\ldots\alpha_m\alpha_{m+1}\ldots \qquad\qquad 0 \le \alpha_i \le b-1, \quad \forall i \text{ e} \quad \alpha_1 \neq 0.$$

Sia a' definito da

$$a' = \begin{cases} \pm 0.\alpha_1\alpha_2\ldots\alpha_m & 0 \le \alpha_{m+1} \le \lfloor \dfrac{b}{2} \rfloor - 1 \\ \pm 0.\alpha_1\alpha_2\ldots\alpha_m + b^{-m} & \alpha_{m+1} \ge \lfloor \dfrac{b}{2} \rfloor \end{cases}$$

cioè si considerano solo le prime m cifre del numero con la m-esima inalterata o aumentata di 1, a seconda del valore della $m+1$-esima.
Ad esempio, in base $b = 10$, l'm-esima cifra di mantissa α_m rimane inalterata se $\alpha_{m+1} < 5$ ed è aumentata di 1 se $\alpha_{m+1} \ge 5$.

Se si pone

$$fl\,(x) = a' \cdot b^t$$

è evidente che $fl\,(x) \in \mathcal{A}$.

Esempio 4.2 *Supponendo di lavorare in virgola mobile con 4 cifre di mantissa*

$$\begin{aligned} x_1 &= 0.789435 & fl(x_1) &= 0.7894 \\ x_2 &= 4.59431 \cdot 10^2 & fl(x_2) &= 0.4594 \cdot 10^3 \\ x_3 &= 0.968781 & fl(x_3) &= 0.9688 \\ x_4 &= 86943.8 & fl(x_4) &= 0.8694 \cdot 10^5 \end{aligned}$$

L'errore che si commette sostituendo al valore esatto x, il valore "arrotondato" ad m cifre di mantissa, $fl(x)$, è detto **errore di arrotondamento**
Ricordando la definizione di errore relativo si verifica che

$$\left|\frac{x - fl(x)}{x}\right| < \frac{\lfloor \frac{b}{2} \rfloor \, b^{-(m+1)}}{|a|} \leq \frac{\lfloor \frac{b}{2} \rfloor \, b^{-(m+1)}}{b^{-1}} = \lfloor \frac{b}{2} \rfloor \, b^{-m}$$

Se si pone $\nu = \lfloor \frac{b}{2} \rfloor \, b^{-m}$, si ottiene

$$fl(x) = x(1+\varepsilon), \qquad \text{con} \quad |\varepsilon| < \nu.$$

Il numero ν è detto **precisione di macchina** e nel sistema decimale vale $5 \cdot 10^{-m}$.
In base 10 quindi l'arrotondamento induce un errore relativo per il quale vale la maggiorazione

$$\left|RE_x\right| = \left|\frac{x - fl(x)}{x}\right| \leq 5 \cdot 10^{-m} \qquad (m \text{ cifre di mantissa})$$

In alcuni casi casi, può verificarsi che anche dopo l'arrotondamento $fl(x) \notin \mathcal{A}$, ad esempio

Esempio 4.3 *Dato $x = 0.999998 \cdot 10^{99}$ (base decimale), si vuole determinare $fl(x)$ in un'aritmetica finita con 5 cifre di mantissa e 2 di caratteristica.*

$$x = 0.999998 \cdot 10^{99} \quad \Longrightarrow \quad fl(x) = 0.10000 \cdot 10^{100} \notin A.$$

$fl(x) \notin A$, essendo espresso con 3 cifre di caratteristica.

In questi casi si ha un fenomeno di sovraesponenziazione detto (**overflow**). In modo analogo si possono avere casi di sottoesponenziazione (**underflow**)

Sull'insieme dei numeri macchina, sottoinsieme dei numeri reali, si devono ridefinire le operazioni aritmetiche, infatti anche se gli operandi sono numeri macchina il risultato può non esserlo.
Definiamo le quattro operazioni elementari macchina

$$x \oplus y = fl(x+y), \qquad x \ominus y = fl(x-y), \qquad x \otimes y = fl(x*y), \qquad x \oslash y = fl(x/y).$$

dove fl indica l'arrotondamento floating-point.
È importante osservare che

Osservazione 2. *Per le operazioni in aritmetica finita (virgola mobile e numero di cifre fissato per la mantissa e la caratteristica) gran parte delle proprietà dell'aritmetica nel campo reale, ad esempio la proprietà associativa o distributiva, possono non essere più valide.*

Esempio 4.4 *Eseguendo i calcoli in virgola mobile non valgono le consuete leggi associative della somma e del prodotto*

$$\begin{aligned}(a+b)+c &\neq a+(b+c)\\ a\cdot(b\cdot c) &\neq (a\cdot b)\cdot c.\end{aligned}$$

Dati $a = 17.6,\ b = 3.81,\ c = 2.736$*, si vuole calcolare*

$$s = a+b+c \qquad e \qquad z = a\cdot b\cdot c$$

con 3 *cifre di mantissa.*

$$a+b = 17.6+3.81 = 21.41 \approx 0.214\cdot 10^2$$
$$s = (\mathbf{a}+\mathbf{b})+\mathbf{c} = 0.214\cdot 10^2 + 0.274\cdot 10 \approx \mathbf{0.241\cdot 10^2}$$

$$b+c = 3.81+2.736 = 6.546 \approx 0.655\cdot 10$$
$$\overline{s} = \mathbf{a}+(\mathbf{b}+\mathbf{c}) = 17.6+6.55 = 24.15 \approx \mathbf{0.242\cdot 10^2}.$$

$$s = 0.241\cdot 10^2 \neq 0.242\cdot 10^2 = \overline{s}.$$

$$a\cdot b = 17.6\cdot 3.81 \approx 0.671\cdot 10^2$$
$$z = (a\cdot b)\cdot c = 67.1\cdot 2.736 \approx \mathbf{0.184\cdot 10^3}$$

$$b\cdot c = 3.81\cdot 2.736 \approx 0.104\cdot 10^2$$
$$\overline{z} = a\cdot(b\cdot c) = 17.6\cdot 10.4 \approx \mathbf{0.183\cdot 10^3}.$$

$$z = 0.184\cdot 10^3 \neq 0.183\cdot 10^3 = \overline{z}.$$

4.4.2 Errore di propagazione

L'errore di propagazione è generato dal "propagarsi" di errori esistenti quando il dato è coinvolto in ulteriori operazioni.
Se un errore sui dati iniziali tende a crescere quando il metodo procede, anche di fronte alla convergenza teorica del metodo, il risultato finale sarà del tutto inattendibile.
Quando un metodo tende ad amplificare gli errori, esso è detto **instabile**, di contro ad un metodo **stabile** che mantiene costante il limite dell'errore nel corso dell'esecuzione.

Si suppone di dover eseguire un calcolo i cui dati di partenza siano affetti da errori inerenti o acquisiti e si analizzano gli effetti di tali errori nel calcolo successivo.

Esempio 4.5 *Si vuole calcolare, arrotondando a quattro cifre decimali, la radice di modulo inferiore dell'equazione*

$$x^2 + 0.4002\ x + 0.00008 = 0$$

usando la classica formula

$$x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}.$$

Essendo $b > 0$*, la radice di modulo inferiore si ottiene prendendo il segno positivo per la radice, cioè*

$$x = \frac{-b + \sqrt{b^2-4ac}}{2a}.$$

Si osserva che $0.00008 = 0.8 \cdot 10^{-4}$*, cioè non deve essere arrotondato; le* 4 *cifre di mantissa si riferiscono al numero scritto in virgola mobile normalizzata.*

$$\begin{aligned} x &= \frac{-b+\sqrt{b^2-4ac}}{2a} = \frac{-0.4002+\sqrt{0.16016004-0.00032}}{2} \approx \\ &\approx \frac{-0.4002+\sqrt{0.1602-0.00032}}{2} = \frac{-0.4002+\sqrt{0.15988}}{2} \approx \\ &\approx \frac{-0.4002+\sqrt{0.1599}}{2} = \frac{-0.4002+0.39987498}{2} \approx \\ &\approx \frac{-0.4002+0.3999}{2} = -\frac{0.0003}{2} = -0.00015 = -0.15 \cdot 10^{-3}. \end{aligned}$$

Il risultato "esatto" è $x = -0.0002 = -0.2 \cdot 10^{-3}$*.*

L'errore percentuale si ottiene come $\left|\frac{x-\bar{x}}{x}\right|$*, se* $x \approx \bar{x}$*, quindi l'errore è*

$$\left|\frac{-0.0002+0.00015}{-0.0002}\right| = \left|\frac{(-0.2+0.15)\cdot 10^{-3}}{-0.2\cdot 10^{-3}}\right| = 0.25 = 25\%.$$

Esempio 4.6 *Dato il sistema lineare*

$$\begin{cases} 5x - 331y = 3.5 \\ 6x - 397y = 5.2 \end{cases}$$

a) la soluzione esatta, calcolata con il metodo di Cramer, è

$$x = 331.7,\ y = 5;$$

b) variando ("perturbando") il termine noto della seconda equazione da 5.2 *a* 5.1*, si ottiene la soluzione*

$$x = 298.6,\ y = 4.5.$$

Un cambiamento percentuale nei dati iniziali di $\frac{5.2-5.1}{5.2} \approx 0.01923 \approx 2\%$*, comporta una variazione nei risultati di*

$$\frac{331.7-298.6}{331.7} \approx 0.09979 \approx 0.1 = 10\%, \quad \frac{5-4.5}{5} = 0.1 = 10\%.$$

c) sostituendo nel sistema i valori $x = 358.173,\ y = 5.4$ *e arrotondando opportunamente i dati al numero di cifre di mantissa fissato,*

$$\begin{aligned} 5 * 358.173 - 331 * 5.4 &= 3.456 \approx 3.5 \\ 6 * 358.173 - 397 * 5.4 &= 5.238 \approx 5.2 \end{aligned}$$

le equazioni sono ancora soddisfatte!

Per formalizzare lo studio e gli effetti dell'errore di propagazione, innanzi tutto si esamina il caso delle **quattro operazioni elementari**: addizione, sottrazione, moltiplicazione, divisione.

Siano x e y due numeri reali e x_1, y_1 le relative approssimazioni e si suppone, inizialmente, che le operazioni $+,-,\cdot,/$ si effettuino senza arrotondamento. Dopo le operazioni, per l'errore assoluto si ha

Addizione $$E_{x+y} = (x+y) - (x_1+y_1) = (x-x_1) + (y-y_1) = E_x + E_y$$

Sottrazione $$E_{x-y} = (x-y) - (x_1-y_1) = (x-x_1) - (y-y_1) = E_x - E_y$$

Moltiplicazione $$E_{xy} = xy - x_1y_1 = (x_1+E_x)(y_1+E_y) - x_1y_1 =$$

$$= x_1y_1 + E_xy_1 + E_yx_1 + E_xE_y - x_1y_1 \cong E_xy_1 + E_yx_1$$

dove, essendo E_x e E_y quantità molto piccole, si suppone che il loro prodotto sia trascurabile rispetto alle quantità stesse

Divisione $$E_{x/y} = \frac{x}{y} - \frac{x_1}{y_1} = \frac{xy_1 - yx_1}{yy_1} = \frac{(x_1+E_x)\,y_1 - (y_1+E_y)\,x_1}{(y_1+E_y)\,y_1} =$$

$$= \frac{x_1y_1 + E_xy_1 - x_1y_1 - x_1E_y}{y_1^2\left(1+\dfrac{E_y}{y_1}\right)} \cong \frac{E_xy_1 - E_yx_1}{y_1^2} \quad \text{se} \quad \frac{E_y}{y_1} \ll 1$$

Molto spesso, però, dopo l'operazione il risultato dovrà essere arrotondato. Basta pensare che lavorando con m cifre di mantissa, il numero di cifre del prodotto di due numeri è compreso tra $2m-1$ e $2m$, e analogamente per le altre operazioni. Pertanto, le formule precedenti devono essere modificate con l'aggiunta di un termine che esprima l'errore di arrotondamento nell'operazione. Se si indica con α, σ, β, δ gli errori di arrotondamento rispettivamente per le operazioni $+,-,\cdot,/$ si ottengono le formule

Addizione	E_{x+y}	$=$	$E_x + E_y + \alpha$
Sottrazione	E_{x-y}	$=$	$E_x - E_y + \sigma$
Moltiplicazione	E_{xy}	$=$	$E_xy_1 + E_yx_1 + \beta$
Divisione	$E_{x/y}$	$=$	$\dfrac{1}{y_1^2}[E_xy_1 - E_yx_1] + \delta.$

È utile osservare che l'errore di arrotondamento non dipende dalla particolare operazione, nel senso che l'arrotondamento deve essere effettuato indipendentemente da come è stato ottenuto il numero.
Poiché l'errore relativo è spesso più significativo dell'errore assoluto, occorre determinarne le formule per le quattro operazioni. Usando la relazione

$$RE_{x\oplus y} = \frac{E_{x\oplus y}}{x_1 \oplus y_1}$$

ove $\oplus$ indica una qualsiasi delle quattro operazioni, che lega l'errore assoluto all'errore relativo, si possono facilmente scrivere le formule per l'errore relativo.

(1)		(2)	(3)		(4)	(5)		(6)	
RE_{x+y}	$=$	RE_x	$\dfrac{x_1}{x_1+y_1}$	$+$	RE_y	$\dfrac{y_1}{x_1+y_1}$	$+$	α	*addizione*
RE_{x-y}	$=$	RE_x	$\dfrac{x_1}{x_1-y_1}$	$+$	RE_y	$\dfrac{-y_1}{x_1-y_1}$	$+$	σ	*sottrazione*
RE_{xy}	$=$	RE_x	1	$+$	RE_y	1	$+$	β	*moltiplicazione*
$RE_{x/y}$	$=$	RE_x	1	$+$	RE_y	(-1)	$+$	δ	*divisione*
		N	**O**		**N**	**O**		**R**	

Le colonne (2), (3), (4), (5), (6), sono state indicate con le lettere N,O,N,O,R per indicare **N**umero, **O**perazione, **N**umero, **O**perazione, **R**ounding (arrotondamento).

- La colonna (1) contiene l'errore nell'operazione $x \oplus y$;
- le colonne (2) e (4) contengono, rispettivamente l'errore relativo su x e su y;
- le colonna (3) e (5) contengono i termini che si riferiscono all'operazione che si sta compiendo (indici di condizionamento);
- la colonna (6) contiene l'errore di arrotondamento relativo all'operazione $x \oplus y$.

Osservando le colonne (3) e (5) si constata che nell'addizione, nella moltiplicazione e nella divisione il fattore moltiplicativo (detto anche indice di condizionamento) non supera in modulo l'unità. Quindi gli errori relativi su x e y non sono amplificati nell'operazione. Ciò non avviene nella sottrazione, in cui almeno uno dei fattori è, in modulo, maggiore di 1. Conseguentemente, uno degli errori relativi sarà aumentato dall'operazione. Tale aumento è tanto maggiore quanto più x è vicino ad y.

Esempio 4.7 *Dati $x = 0.5628 \cdot 10^4$ e $y = 0.5631 \cdot 10^4$, si vuole calcolare $z = x - y$ con quattro cifre di mantissa e dare anche un limite per l'errore relativo RE.*
La soluzione è

$$z = -0.0003 \cdot 10^4$$

Da

$$RE_z = \frac{x}{x-y} RE_x - \frac{y}{x-y} RE_y,$$

passando al valore assoluto si ha

$$|RE_z| \leq |\frac{x}{x-y}||RE_x| + |\frac{y}{x-y}||RE_y|. \tag{4.1}$$

Supponendo x e y arrotondati a cinque cifre di mantissa, cioè,

$$\begin{aligned} \left|RE_x\right| &\leq 0.5 \cdot 10^{-4} = 0.005\% \\ \left|RE_y\right| &\leq 0.5 \cdot 10^{-4} = 0.005\% \end{aligned}$$

e sostituendo nella (4.1) *si trova*

$$\left|RE_z\right| \leq 5 \cdot 10^{-5} \frac{5628 + 5631}{3} \leq 5 \cdot 10^{-5} \cdot 4 \cdot 10^3 = 0.2 = 20\%.$$

Quindi anche se x e y hanno limiti dell'errore relativo molto piccoli, il risultato $z = x - y$ può avere un limite dell'errore relativo molto grande.

Per stimare l'andamento dell'errore di propagazione in una generica operazione si considera un problema tipo definito da $y = \varphi(x)$, con

$$\varphi : D \to R^m, \quad \varphi(x) = \begin{bmatrix} \varphi_1(x_1, \ldots, x_n) \\ \vdots \\ \varphi_m(x_1, \ldots, x_n) \end{bmatrix}, \quad \varphi \in C^1(D), \quad D \in R^n.$$

Se $\tilde{x} \approx x$, cioè $\tilde{x}$ è un'approssimazione di x, si definiscono

$\tilde{y} = \varphi(\tilde{x})$	soluzione sui dati approssimati
$\Delta x_i = \tilde{x}_i - x_i,\ \Delta y_i = \tilde{y}_i - y_i$	errore assoluto su ciascuna componente
$\delta x_i = \dfrac{\tilde{x}_i - x_i}{x_i},\ \delta y_i = \dfrac{\tilde{y}_i - y_i}{y_i}$	errore relativo per ciascuna componente.

Sviluppando in serie con arresto ai termini del primo ordine si trova l'espressione dell'errore assoluto su ciascuna componente della soluzione

$$\Delta y_i = \tilde{y}_i - y_i = \varphi_i(\tilde{x}) - \varphi_i(x) \approx \sum_{j=1}^{n} (\tilde{x}_j - x_j) \frac{\partial \varphi_i(x)}{\partial x_j} = \sum_{j=1}^{n} \frac{\partial \varphi_i(x)}{\partial x_j} \Delta x_j.$$

Le quantità $\dfrac{\partial \varphi_i(x)}{\partial x_j}$ indicano come l'errore sulla componente x_j del dato si amplifica sulla componenete y_i del risultato.
Se $y_i \neq 0, \quad i = 1, \ldots, m$ e $x_j \neq 0, \quad i = 1, \ldots, n$, le quantità

$$\delta y_i = \frac{\Delta y_i}{y_i} = \sum_{j=1}^{n} \frac{\partial \varphi_i(x)}{\partial x_j} \frac{\Delta x_j}{x_j} \frac{x_j}{y_i} \approx \sum_{j=1}^{n} \frac{x_j}{\varphi_i(x)} \frac{\partial \varphi_i(x)}{\partial x_j} \delta x_j.$$

rappresentano gli errori relativi su ciascuna componente della soluzione.
I fattori moltiplicativi presenti nella formula sono detti **indici di condizionamento** del problema.
Se l'indice di condizionamento risulta, in modulo, < 1, l'errore sull'operando non viene amplificato.
Nel caso delle quattro operazioni aritmetiche, gli indici di condizionamento sono sempre < 1, tranne che nella sottrazione.
Questo fenomeno è noto come **cancellazione numerica**, esso rende di fatto la sottrazione l'operazione numericamente più instabile, cioè più pericolosa relativamente alla propagazione dell'errore.
Nella realizzazione di un algoritmo è necessario evitare la sottrazione di numeri quasi uguali, riscrivendo quando possibile le espressioni in una forma, matematicamente equivalente, che non contiene operazioni di tale tipo; ad esempio, se $b \approx c$ e $a \approx b$,

$$\begin{aligned} a\,(b-c) \quad &\Longrightarrow \quad ab - ac \\ \frac{a-b}{c} \quad &\Longrightarrow \quad \frac{a}{c} - \frac{b}{c}. \end{aligned}$$

4.4.3 Propagazione degli errori in semplici espressioni aritmetiche

Si analizza, ora, attraverso degli esempi, la propagazione degli errori in semplici espressioni aritmetiche applicando i risultati del paragrafo precedente.

Esempio 4.8 *Si vuole calcolare la somma*

$$S = a_0 + a_1 + a_2 + a_3 \tag{4.2}$$

dei numeri reali x_0, x_1, x_2, x_3 affetti da errori relativi, rispettivamente, RE_0, RE_1, RE_2, RE_3.
Se si usano le somme parziali

$$S_1 = x_0 + x_1, \quad S_2 = x_0 + x_1 + x_2, \quad S_3 = x_0 + x_1 + x_2 + x_3.$$

la (4.2) *diventa*

$$S = S_3 = S_2 + x_3 = (S_1 + x_2) + x_3 = [(x_0 + x_1) + x_2] + x_3.$$

Se si indica con α_k l'errore di arrotondamento nella $k-$esima somma dallo schema NONOR si ottiene

$$RE_S = RE_{S_3} = \underbrace{RE_{S_2}\frac{S_2}{S} + RE_3\frac{x_3}{S} + \alpha_3}_{N\ O\ N\ O\ R} = \underbrace{\left(\underbrace{RE_{S_1}\frac{S_1}{S_2} + RE_2\frac{x_2}{S_2} + \alpha_2}_{N\ O\ N\ O\ R}\right)\frac{S_2}{S} + RE_3\frac{x_3}{S} + \alpha_3}_{N\ O\ N\ O\ R} =$$

$$= \underbrace{\left[\underbrace{\left(\underbrace{RE_0\frac{x_0}{S_1} + RE_1\frac{x_1}{S_1} + \alpha_1}_{N\ O\ N\ O\ R}\right)\frac{S_1}{S_2} + RE_2\frac{x_2}{S_2} + \alpha_2}_{N\ O\ N\ O\ R}\right]\frac{S_2}{S} + RE_3\frac{x_3}{S} + \alpha_3}_{N\ O\ N\ O\ R} =$$

$$= \frac{RE_0x_0 + RE_1x_1 + RE_2x_2 + RE_3x_3 + \alpha_1S_1 + \alpha_2S_2 + \alpha_3S}{S}$$

Per calcolare l'errore assoluto, tenendo conto che $E_s = S \cdot RE_s$ *e che* $S_3 = S$*, si ottiene*

$$E_s = \sum_{k=0}^{3} RE_k a_k + \sum_{k=1}^{3} \alpha_k S_k \tag{4.3}$$

Se $|RE_k| < k \cdot 10^{-m}$ *con* $k \geq 5$ *e* $|\alpha_k| \leq 5 \cdot 10^{-m}$*, dalla* (4.3)

$$|E_s| \leq k\left(\sum_{k=0}^{3}|x_k| + \sum_{k=1}^{3}|S_k|\right)10^{-m}$$

Generalizzando questo risultato si può affermare che l'errore che si commette calcolando la somma di n *numeri* $S = \sum_{k=0}^{n} x_k$*, con errore relativo* RE_k *nel termine* x_k *ed errore di arrotondamento* α_k *nella* $k-$*esima addizione, è espresso da*

$$E_s = \sum_{k=0}^{n} RE_k x_k + \sum_{k=1}^{n} \alpha_k S_k$$

ove $S_k = x_0 + x_1 + \cdots + x_k$*. Tale errore soddisfa certamente la relazione*

$$|E_s| \leq \sum_{k=0}^{n}|RE_k||x_k| + \sum_{k=1}^{n}|\alpha_k||S_k|. \tag{4.4}$$

Siano i numeri x_k *tutti dello stesso segno e siano ordinati in uno dei seguenti modi*

(1) $|x_0| \leq |x_1| \leq |x_2| \leq \cdots \leq |x_n|$

(2) $|x_0| \geq |x_1| \geq |x_2| \geq \cdots \geq |x_n|$,

allora le somme parziali S_k*, in modulo, sono più piccole per il primo ordinamento rispetto al secondo; conseguentemente anche il limite* (4.4) *è più piccolo per il primo ordinamento. Per questa ragione, dovendo sommare una sequenza di numeri di grandezza variabile (ma dello stesso segno) conviene sommarli in ordine di grandezza dal più piccolo al più grande.*

Esempio 4.9 *Si vuole calcolare l'errore che si commette nel prodotto dei numeri reali a_0, a_1, a_2, a_3 con errori relativi RE_0, RE_1, RE_2, RE_3, e con μ_k errore di arrotondamento nella $k-$esima moltiplicazione.*

$$P = a_0 \cdot a_1 \cdot a_2 \cdot a_3 = (((a_0 \cdot a_1) \cdot a_2) \cdot a_3).$$

La corrispondente espressione per l'errore relativo, secondo lo schema NONOR, è

$$RE_P = \{[(RE_0 \cdot 1 + RE_1 \cdot 1 + \mu_1) \cdot 1 + RE_2 \cdot 1 + \mu_2] \cdot 1 + RE_3 \cdot 1 + \mu_3\}$$

$$\underbrace{\underbrace{\underbrace{N \quad O}_{N} \quad \underbrace{N \quad O}_{O} \quad R}_{N}}_{N} \quad \underbrace{\underbrace{\underbrace{N}_{N} \quad \underbrace{O \quad N}_{O} \quad \underbrace{O \quad R}_{O}}_{O} \quad N \quad O \quad R}_{N}$$

che, con ovvie semplificazioni, diventa:

$$RE_P = RE_0 + RE_1 + \mu_1 + RE_2 + \mu_2 + RE_3 + \mu_3 = \sum_{k=0}^{3} RE_k + \sum_{k=1}^{3} \mu_k$$

Conseguentemente, per l'errore assoluto si ha

$$E_P = P \left[\sum_{k=0}^{3} RE_k + \sum_{k=1}^{3} \mu_k \right]$$

Generalizzando questo risultato, si può affermare che per il prodotto di $n+1$ numeri reali $P = \prod_{k=0}^{n} a_k$, con ovvio significato di simboli, l'errore è

$$E_P = P \left[\sum_{k=0}^{n} RE_k + \sum_{k=1}^{n} \mu_k \right]$$

Se si indica con $R = \max_{0 \le k \le n} |RE_k|$ e $\mu = \max_{0 \le k \le n} |\mu_k|$ si trova che il limite dell'errore nel prodotto è

$$|E_P| \le |P|[(n+1)R + n\mu]$$

Esempio 4.10 *Siano a_0, a_1, a_2 i coefficienti del polinomio*

$$P_2(x) = a_0 x^2 + a_1 x + a_2.$$

Se RE_0, RE_1, RE_2, RE_x, sono rispettivamente gli errori relativi in a_0, a_1, a_2, x,si vuole calcolare l'espressione dell'errore per $P_2(x)$.
Il polinomio $P_2(x)$ può essere scritto nella forma

$$P_2(x) = \{[a_0 x + a_1]x + a_2\} \tag{4.5}$$

Se con α_k e μ_k si indicano rispettivamente gli errori relativi di arrotondamento nella $k-$esima addizione e moltiplicazione, l'espressione dell'errore per la (4.5) secondo lo schema NONOR è

$$\begin{aligned} RE_P \quad &= \left\{ \left[((RE_0 \cdot 1 + RE_x \cdot 1 + \mu_1) \cdot \frac{a_0 x}{a_0 x + a_1} + RE_1 \frac{a_1}{a_0 x + a_1} + \alpha_1) \cdot 1 + \right.\right. \\ &\left.\left. + RE_x \cdot 1 + \mu_2 \right] \frac{(a_0 x + a_1)x}{P_2(x)} + RE_2 \frac{a_2}{P_2(x)} + \alpha_2 \right\} \quad . \end{aligned}$$

Eseguendo i calcoli si trova

$$\begin{aligned} RE_P \quad &= \frac{(RE_0 + RE_x + \mu_1)a_0 x + RE_1 a_1 + (\alpha_1 + RE_x + \mu_2)(a_0 x + a_1)}{a_0 x + a_1} \cdot \\ &\cdot \frac{(a_0 x + a_1)x}{P_2(x)} + \frac{RE_2 a_2 + \alpha_2 P_2(x)}{P_2(x)} \end{aligned}$$

e quindi moltiplicando per $P_2(x)$ e raccogliendo i termini simili

$$E_P = a_0x^2(RE_0 + RE_x + \mu_1 + \alpha_1 + RE_x + \mu_2 + \alpha_2) + a_1x(RE_1 + \alpha_1 + RE_x + \mu_2 + \alpha_2) + \\ + a_2(RE_2 + \alpha_2) \quad .$$

Se $|RE_k|$, $|\alpha_k|$ e $|\mu_k|$ sono limitati da $5 \cdot 10^{-m}$ si ottiene il limite

$$|E_P| \leq \left[7|a_0x^2| + 5|a_1x| + 2|a_2|\right] \cdot 5 \cdot 10^{-m}. \tag{4.6}$$

Si nota che il limite (4.6) *è tanto più piccolo quanto più piccolo è $|x|$.*

4.4.4 Errore di troncamento

Se si indica con α la soluzione del problema originale e con α_n la soluzione dell'n-esimo problema approssimante, un procedimento di successive approssimazioni consiste quindi nell'approssimare α con la successione $\{\alpha_n\}_n$. Perchè un processo di questo tipo abbia significato deve risultare

$$\lim_{n\to\infty} \alpha_n = \alpha.$$

condizione che garantisce la **convergenza teorica** del metodo.
Poiché il calcolo di ciascun α_n, ovvero della soluzione di ciascun problema approssimante, coinvolge un numero finito, ma spesso elevato, di operazioni elementari, e poiché in genere per arrivare ad una soluzione accettabile è necessario calcolare un numero elevato di termini α_n, un procedimento di successive approssimazioni risulta di fatto eseguibile solo con opportuni mezzi di calcolo.
In ogni caso, comunque, non si potrà mai calcolare il limite esatto α quindi, nella pratica, occorre arrestare il procedimento infinito ad un certo valore di n, indicato con $\bar{n}$, e considerare $\alpha_{\bar{n}}$ come soluzione (approssimata) del problema originale. La differenza

$$\alpha - \alpha_{\bar{n}}$$

viene detta **errore di troncamento** del metodo usato.

Esempio 4.11 *Si vuole approssimare e^x mediante il polinomio cubico*

$$p_3(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!}.$$

Il calcolo esatto di e^x richiede il calcolo di una serie infinita

$$e^x = p_3(x) + \sum_{k=4}^{\infty} \frac{x^k}{k!}.$$

L'errore è dovuto al troncamento della serie infinita e può essere ridotto aumentando il numero di termini coinvolti nell'approssimazione, ma mai annullato.

L'errore di troncamento è caratteristico del particolare metodo di approssimazione, e dipende dal troncamento di un processo infinito; il suo studio completo prevede la ricerca di

- sviluppo asintotico $\alpha = \alpha_n + \mathcal{O}\big(\phi(n)\big)$;
- maggiorazione $|\alpha - \alpha_n| < k < \infty$;
- ordine di grandezza $\alpha - \alpha_n = \mathcal{O}\big(\phi(n)\big)$;

Non sempre comunque è agevole ottenere lo studio completo dell'errore di troncamento.

4.4.5 Condizionamento e stabilità

Dato un problema, per quanto riguarda la propagazione degli errori si può distinguere in:

- comportamento del problema;
- comportamento dell'algoritmo risolvente.

La caratterizzazione di un problema rispetto ad un tipo di comportamento è indicata con il termine **condizionamento**; in particolare parleremo di problema *bencondizionato* se le perturbazioni sui dati non influenzano significativamente i risultati, *malcondizionato* altrimenti.

Per caratterizzare il comportamento di un algoritmo invece si usa il termine **stabilità**; un algoritmo sarà detto *stabile* se la successione delle operazioni non amplifica eccessivamente gli errori di arrotondamento, *instabile* altrimenti.
La distinzione tra condizionamento e stabilità è fondamentale:

- per un problema bencondizionato è possibile in generale costruire algoritmi stabili;
- la stabilità di un algoritmo è inutile in presenza di un problema malcondizionato.

Esempio 4.12 *Un tipico problema malcondizionato è il calcolo delle radici di un polinomio*

$$P_n(x) = x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n.$$

Se z è una radice di P_n, si ha, considerando $P_n(z)$ come funzione dei coefficienti a_i

$$P_n(z) = f(a_1, a_2, \ldots, a_n) = z^n + a_1 z^{n-1} + \ldots + a_{n-1}z + a_n = 0.$$

In corrispondenza di una variazione dei coefficienti $\Delta a_i,\ i = 1, 2, \ldots, n$ si avrà

$$\Delta f \approx \sum_{i=1}^{n} \frac{\partial f}{\partial a_i}\Delta a_i + \frac{\partial f}{\partial z}\left(\frac{\partial z}{\partial a_1}\Delta a_1 + \ldots + \frac{\partial z}{\partial a_n}\Delta a_n\right)$$

in cui i coefficienti sono stimati nel punto $\overline{a} = \Big(a_1 + \Delta a_1, \ldots, a_n + \Delta a_n\Big)$. Indicando con

$$\Delta z = \frac{\partial z}{\partial a_1}\Delta a_1 + \ldots + \frac{\partial z}{\partial a_n}\Delta a_n$$

la variazione complessiva della radice, prodotta dall'alterazione dei coefficienti, soddisfa la relazione

$$\sum_{i=1}^{n} z^{n-i}\Delta a_i + \frac{\partial f}{\partial z}\Delta z = 0$$

Se si varia un solo coefficiente, ad es. il $k-$esimo, di una quantità Δa_k, la variazione sulle radici sarà

$$\frac{\Delta z}{z} = A\ \frac{\Delta a_k}{a_k} \qquad con \qquad A = -\frac{a_k z^{n-k}}{z\dfrac{\partial f}{\partial z}}$$

Il coefficiente A può essere grande se, ad esempio,

- *z è grande;*
- *$\dfrac{\partial f}{\partial z}$ è piccolo, come nel caso di radici "vicine" ($\dfrac{\partial f}{\partial z} = 0$ nel caso di radici multiple).*

Nel caso in cui A è grande, a piccole variazioni di un solo coefficiente possono corrispondere grandi variazioni delle radici.

Esempio 4.13 *([Wilkinson, 1959]) Il polinomio di grado* 20

$$P(x) = (x-1)(x-1)\cdots(x-19)(x-20) = x^{20} - 210x^{19} + \ldots + 20!$$

ha radici semplici e ben separate $z_k = k, \quad k = 1, 2, \ldots, 20.$
Se solo il primo coefficiente subisce una variazione $\Delta a_1 = 10^{-7}$ *e quindi*

$$\frac{\Delta a_1}{a_1} \approx 4.8 \cdot 10^{-10},$$

per $z_{20} = 20$ *si ha*

$$A = \frac{210 \cdot 20^{19}}{19!} \approx 0.9 \cdot 10^{-10} \qquad e \qquad \frac{\Delta z_1}{z_1} \approx 4.30 = 430\% \text{ !!}.$$

La variazione più consistente si ha su $z_{16} = 16$*; perturbando il coefficiente* a_5 *di una quantità* $\Delta a_5 = -10^{-10}$ *si ottiene*

$$\frac{\Delta z_{16}}{z_{16}} \approx 5.5 = 550\% \text{ !!}.$$

Per caratterizzare il comportamento di un algoritmo invece si usa il termine **stabilità**; un algoritmo sarà detto *stabile* se la successione delle operazioni non amplifica eccessivamente gli errori di arrotondamento.

Definizione 4.5. *Un procedimento numerico, o algoritmo, si dice numericamente* **stabile** *quando, al crescere del numero dei passi, l'errore si mantiene limitato. Se questo non succede l'algoritmo si dirà (numericamente)* **instabile**

L'instabilità numerica è una caratteristica dell'algoritmo: per la soluzione dello stesso problema possono esistere procedimenti stabili ed altri instabili, come nel caso del metodo di Archimede.

La distinzione tra condizionamento e stabilità è fondamentale:

- per un problema bencondizionato è possibile in generale costruire algoritmi stabili;
- la stabilità di un algoritmo è inutile in presenza di un problema malcondizionato.

4.5 Analisi di un metodo numerico: convergenza, errore di troncamento e di arrotondamento

Lo studio completo di un metodo numerico, oltre alla descrizione ed eventuale codifica dell'algoritmo, necessita, se basato su un procedimento di successive approssimazioni, di una accurata analisi del comportamento. In particolare si devono realizzare:

a) analisi della convergenza teorica;

b) analisi della convergenza numerica;

c) analisi dell'errore di *troncamento*, dovuto al troncamento di un processo di successive approssimazioni. Tale analisi si articola nei seguenti passi fondamentali:

   c1) ordine dell'errore;

   c2) maggiorazione dell'errore;

   c3) sviluppo asintotico dell'errore.

4.5.1 Studio dell'errore nel calcolo numerico di π

Nel caso dell'algoritmo di Archimede per il calcolo approssimato di π, una prima analisi superficiale consente di ottenere facilmente la maggiorazione:

$$\pi - \pi_{1,n} < \pi_{2,n} - \pi_{1,n} = \frac{n l_n (2 - \sqrt{4 - l_n^2})}{2\sqrt{4 - l_n^2}}.$$

Un'analisi più accurata coinvolge tutti gli aspetti legati a convergenza, stabilità e stima dell'errore. Premettiamo i seguenti lemmi

Lemma 4.1 *Sia l_n il lato del poligono regolare di n lati inscritto in una circonferenza di raggio $r = 1$ e sia $\pi_{1,n}$ l'approssimazione per difetto di π ottenuta dal metodo di Archimede. Per $\pi_{1,n}$ vale la relazione*

$$\pi_{1,n} = n \sin \frac{\pi}{n}.$$

Dimostrazione.

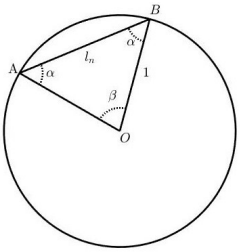


Figura 4.1

Si considera il triangolo isoscele AOB in figura 4.1, costruito su l_n, lato del poligono di n lati inscritto in una circonferenza di raggio 1. Applicando il teorema dei seni al triangolo AOB si trova

$$\frac{l_n}{\sin\beta} = \frac{OA}{\sin\alpha} = \frac{1}{\sin\alpha} \qquad \Longrightarrow \qquad l_n = \frac{\sin\beta}{\sin\alpha}.$$

Essendo l_n il lato di un poligono regolare, vale

$$\beta = \frac{2\pi}{n} \qquad \text{e} \qquad 2\alpha = \pi - \beta$$

da cui

$$2\alpha = \pi - \frac{2\pi}{n} \qquad \Longrightarrow \qquad \alpha = \frac{\pi}{2} - \frac{\pi}{n}.$$

Ricordando che $\sin 2\alpha = 2\sin\alpha\cos\alpha$ e $\sin\left(\frac{\pi}{2} - \alpha\right) = \cos\alpha$

$$l_n = \frac{\sin\beta}{\sin\alpha} = \frac{\sin\frac{2\pi}{n}}{\sin\left(\frac{\pi}{2} - \frac{\pi}{n}\right)} = \frac{2\sin\frac{\pi}{n}\cos\frac{\pi}{n}}{\cos\frac{\pi}{n}} = 2\sin\frac{\pi}{n}. \tag{4.7}$$

e che vale $\pi_{1,n} = \dfrac{n l_n}{2}$, si ha

$$l_n = 2\sin\frac{\pi}{n} \qquad \Longrightarrow \qquad \pi_{1,n} = \frac{n l_n}{2} = n\sin\frac{\pi}{n}.$$

Lemma 4.2 *Per l'approssimazione $\pi_{1,n}$ vale lo sviluppo*

$$\pi_{1,n} = \pi - \frac{1}{3!}\frac{\pi^3}{n^2} + \frac{1}{5!}\frac{\pi^5}{n^4} - \frac{1}{7!}\frac{\pi^7}{n^6} + \ldots = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!}\frac{\pi^{2k+1}}{n^{2k}}. \tag{4.8}$$

Dimostrazione. Sviluppando in serie di Taylor-Mc Laurin (punto iniziale $x_0 = 0$) la funzione $\sin x$, si trova

$$\sin x = \sin 0 + x \sin' 0 + \frac{x^2}{2!}\sin'' 0 + \frac{x^3}{3!}\sin''' 0 + \ldots = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

quindi, da (4.7),

$$\pi_{1,n} = \frac{nl_n}{2} = n \sin\frac{\pi}{n} = n\left(\frac{\pi}{n} - \frac{1}{3!}\left(\frac{\pi}{n}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{n}\right)^5 - \frac{1}{7!}\left(\frac{\pi}{n}\right)^7 + \ldots\right) \tag{4.9}$$

da cui la tesi

Ora possiamo studiare gli aspetti di convergenza e la stima dell'errore nel metodo di Archimede

a) **convergenza teorica**.

È stato già verificato "geometricamente" che

$$\lim_{n\to\infty} \pi_{1,n} = \lim_{n\to\infty} \pi_{2,n} = \pi$$

osservando che per $n \to \infty$ il perimetro dei poligoni inscritti e circoscritti ad una circonferenza tende alla lunghezza della circonferenza stessa.

Per studiare la convergenza analiticamente, dal Lemma 4.2,

$$\lim_{n\to\infty} \pi_{1,n} = \lim_{n\to\infty}\left(\pi - \frac{1}{3!}\frac{\pi^3}{n^2} + \frac{1}{5!}\frac{\pi^5}{n^4} - \frac{1}{7!}\frac{\pi^7}{n^6} + \ldots\right) = \pi$$

che garantisce la convergenza teorica del metodo di Archimede.

b) **convergenza numerica**

Il metodo di Archimede presenta problemi di convergenza numerica. Per analizzare tale fenomeno bisogna studiare la propagazione dell'errore di arrotondamento, da cui dipende la convergenza numerica del metodo. Infatti anche se teoricamente un metodo converge, a causa della propagazione dell'errore di arrotondamento si possono avere risultati falsati. Non sempre è possibile arginare gli effetti di tale fenomeno, mentre in alcuni casi basta modificare opportunamente l'algoritmo sostituendo operazioni computazionalmente "pericolose" con operazioni matematicamente equivalenti ma più "stabili". Per il metodo di Archimede, si osserva che le due formule

$$l_{2n} = \sqrt{2 - \sqrt{4 - l_n^2}}, \qquad l_{2n} = \frac{l_n}{\sqrt{2 + \sqrt{4 - l_n^2}}} \tag{4.10}$$

sono matematicamente equivalenti (la prima si ottiene dalla seconda per razionalizzazione) ma utilizzando la seconda al posto della prima nel calcolo di $\pi_{1,n}$ (o $\pi_{2,n}$) si ottengono i risultati in cui l'effetto dell'errore di arrotondamento risulta limitato. Si può osservare facilmente che per $n \to \infty$, $l_n \to 0$ e nella prima delle (4.10) si genera un problema di cancellazione numerica.

In tabella si può vedere la differenza tra i risultati ottenuti usando le due formule.

$\pi_{1,n}$	$l_{2n} = \sqrt{2 - \sqrt{4 - l_n^2}}$	$l_{2n} = \dfrac{l_n}{\sqrt{2 + \sqrt{4 - l_n^2}}}$
$\pi_{1,6}$	3.(000000000000000)	3.(000000000000000)
$\pi_{1,12}$	3.1(05828541230248)	3.1(05828541230248)
$\pi_{1,1536}$	3.14159(0463228050)	3.14159(0463228049)
$\pi_{1,3072}$	3.141592(105999243)	3.141592(105999271)
$\pi_{1,6144}$	3.141592(516692049)	3.141592(516692156)
$\pi_{1,12288}$	3.1415926(19365120)	3.1415926(19365383)
$\pi_{1,24576}$	3.1415926(45032004)	3.1415926(45033690)
$\pi_{1,49152}$	3.14159265(1438954)	3.14159265(1450767)
$\pi_{1,98304}$	3.141592653(033526)	3.141592653(055036)
$\pi_{1,196608}$	3.141592653(325343)	3.141592653(456103)
$\pi_{1,393216}$	3.141592653(325343)	3.1415926535(56371)
$\pi_{1,786432}$	3.141592653(325343)	3.14159265358(1437)
$\pi_{1,1572864}$	3.1415926(45321215)	3.14159265358(7704)
$\pi_{1,3145728}$	3.1415926(45321215)	3.141592653589(271)
$\pi_{1,6291456}$	3.1415926(45321215)	3.141592653589(663)
$\pi_{1,12582912}$	3.141592(303811737)	3.1415926535897(61)
$\pi_{1,25165824}$	3.1415(89571734577)	3.1415926535897(85)
$\pi_{1,50331648}$	3.1415(75911313137)	3.14159265358979(1)
$\pi_{1,100663296}$	3.1415(43126059329)	3.141592653589793
$\pi_{1,201326592}$	3.141(324548956220)	3.141592653589793
$\pi_{1,402653184}$	3.14(0974793953303)	3.141592653589793
$\pi_{1,805306368}$	3.1(37475099502783)	3.141592653589793
$\pi_{1,1610612736}$	3.1(26249750099949)	3.141592653589793
$\pi_{1,3221225472}$	3.(092329219213245)	3.141592653589793
$\pi_{1,6442450944}$	3.(000000000000000)	3.141592653589793
$\pi_{1,12884901888}$	(0.000000000000000)	3.141592653589793

Tabella 1. *Confronto tra il calcolo di $\pi_{1,n}$ con le formule* (4.10)

Si constata facilmente che a partire da un certo n in poi, con la prima delle (4.10), $\pi_{1,n}$ si discosta anzichè avvicinarsi a π, mentre utilizzando la seconda si ottengono risultati più stabili.

c) **studio dell'errore di troncamento**

c1) $|\pi - \pi_{1,n}| = \mathcal{O}\left(\dfrac{1}{n^\alpha}\right), \quad \alpha \in \mathbb{R}$ **(ordine)**;

Dalla (4.8) si ricava

$$\pi - \pi_{1,n} = \frac{1}{3!}\frac{\pi^3}{n^2} - \frac{1}{5!}\frac{\pi^5}{n^4} + \frac{1}{7!}\frac{\pi^7}{n^6} - \dots$$

cioè

$$\pi - \pi_{1,n} = \mathcal{O}\left(n^{-2}\right)$$

che dà l'**ordine** del metodo di Archimede.

c2) $|\pi - \pi_{1,n}| \le k\ \dfrac{1}{n^\beta}, \quad k \in \mathbb{R}^+,\ \beta \in \mathbb{R}$ **(maggiorazione)**;

Dalla (4.8), essendo $\dfrac{\pi}{n} > 0$,

$$\pi_{1,n} = n \sin\frac{\pi}{n} \ > \ \pi - \frac{\pi^3}{6n^2}$$

da cui

$$\pi - \pi_{1,n} \ < \ \frac{\pi^3}{6n^2} \ < \ \frac{\left(\dfrac{22}{7}\right)^3}{6n^2} \approx \frac{5.17}{n^2}$$

dove si è usata la limitazione di Archimede (3.1). Si ottiene così una **maggiorazione** dell'errore nel metodo di Archimede.

c3) $|\pi - \pi_{1,n}| = g(n) + \mathcal{O}\left(\frac{1}{n^\sigma}\right),\ \sigma \in \mathbb{R},\ g$ nota **(svil. asintot.)**.

Sempre dalla (4.8)

$$\pi_{1,n} = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} \frac{\pi^{2k+1}}{n^{2k}} = \pi + \sum_{k=1}^{\infty} (-1)^k \frac{1}{(2k+1)!} \frac{\pi^{2k+1}}{n^{2k}}$$

da cui

$$\pi - \pi_{1,n} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{(2k+1)!} \frac{\pi^{2k+1}}{n^{2k}}$$

cioè lo **sviluppo asintotico** dell'errore.

Con le informazioni ricavate, è possibile studiare la velocità di convergenza, cioè il comportamento dell'errore al passo $(n+1)$-esimo in relazione all'errore al passo n-esimo.

Dalla (4.9) si ricava

$$\begin{aligned} \pi - \pi_{1,n} &= n\left[\frac{1}{3!}\left(\frac{\pi}{n}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{n}\right)^5 - \frac{1}{7!}\left(\frac{\pi}{n}\right)^7 + \dots\right] \\ \pi - \pi_{1,2n} &= 2n\left[\frac{1}{3!}\left(\frac{\pi}{2n}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{2n}\right)^5 - \frac{1}{7!}\left(\frac{\pi}{2n}\right)^7 + \dots\right] \end{aligned}$$

e quindi

$$\lim_{n\to\infty} \frac{\pi - \pi_{1,2n}}{\pi - \pi_{1,n}} = \lim_{n\to\infty} \frac{2n\left[\frac{1}{3!}\left(\frac{\pi}{2n}\right)^3 - \frac{1}{5!}\left(\frac{\pi}{2n}\right)^5 + \dots\right]}{n\left[\frac{1}{3!}\left(\frac{\pi}{n}\right)^3 - \frac{1}{5!}\left(\frac{\pi}{n}\right)^5 + \dots\right]} = \lim_{n\to\infty} \frac{2\frac{1}{3!}\left(\frac{\pi}{2n}\right)^3\left[1 - \frac{3!}{5!}\left(\frac{\pi}{2n}\right)^2 + \dots\right]}{\frac{1}{3!}\left(\frac{\pi}{n}\right)^3\left[1 - \frac{3!}{5!}\left(\frac{\pi}{n}\right)^2 + \dots\right]} = \mathbf{\frac{1}{4}}$$

cioè

$$\left|\pi - \pi_{1,2n}\right| \approx \frac{1}{4}\left|\pi - \pi_{1,n}\right|$$

che dà la **velocità di convergenza** del metodo di Archimede. La relazione trovata esprime che ad ogni passo l'errore di troncamento si riduce di un fattore $\frac{1}{4}$.

Capitolo 5

Calcolo di Polinomi

"Ovviamente" è la parola più pericolosa in matematica.
Eric Temple Bell (1883-1960)

5.0 Introduzione

I polinomi sono particolarmente importanti in Analisi Numerica, sia perché il loro calcolo effettivo è, come vedremo, agevole, sia perché sono facili da derivare e da integrare. Per il calcolo delle radici di un polinomi esistono algoritmi specifici.
In questo contesto ci limitiamo ad affrontare il problema del calcolo di un polinomio in un punto, introducendo l'algoritmo di Horner, che, rispetto all'usuale metodo del calcolo delle potenze, risulta più efficiente sia rispetto al costo computazionale, sia rispetto alla precisione.

5.1 Algoritmo di Horner

Un polinomio di grado n a coefficienti reali, nella variabile x, è un'espressione del tipo

$$P_n(x) = a_0x^n + a_1x^{n-1} + \ldots\ldots + a_{n-1}x + a_n \equiv \sum_{i=0}^{n} a_i x^{n-i}. \tag{5.1}$$

con $a_i \in \mathbb{R}$ $i = 0, \ldots, n$, $a_0 \neq 0$. Il valore che il polinomio $P(x)$ assume in un punto $z \in \mathbb{R}$ è dato da

$$P_n(z) = a_0z^n + a_1z^{n-1} + \ldots\ldots + a_{n-1}z + a_n \equiv \sum_{i=0}^{n} .a_i z^{n-i};$$

Se ciascuna potenza è calcolata mediante moltiplicazioni successive, cioè $z^k = z \cdot z^{k-1}$, il numero di moltiplicazioni richiesto, per il calcolo di $P_n(z)$, è $2n - 1$. Osserviamo ora che esiste una procedura che consente di ottenere lo stesso risultato con un numero inferiore di moltiplicazioni. Infatti la (5.1), mediante ripetute messe in evidenza parziali, può essere scritta nella forma

$$P_n(x) = a_n + x\left(a_{n-1} + x\left(a_{n-2} + \ldots\ldots + x\left(a_1 + xa_0\right)\ldots\right)\right). \tag{5.2}$$

Ad esempio, per $n = 5$ si ha

$$P_5(x) = a_0x^5 + a_1x^4 + a_2x^3 + a_3x^2 + a_4x + a_5 = a_5 + x\left(a_4 + x\left(a_3 + x\left(a_2 + x\left(a_1 + a_0x\right)\right)\right)\right).$$

Se si vuole ora calcolare $P_n(z)$, utilizzando la (5.2), il calcolo può procedere a partire dalle parentesi più interne, ovvero può essere organizzato come segue:

$$\begin{cases} b_1 = z \cdot a_0 + a_1 \\ b_2 = z \cdot b_1 + a_2 \\ \vdots \\ b_n = z \cdot b_{n-1} + a_n, \end{cases} \tag{5.3}$$

il valore b_n è esattamente il valore richiesto $P_n(z)$. Lo schema può essere scritto nella forma ricorrente

$$\begin{cases} b_0 = a_0 \\ b_k = b_{k-1} \cdot z + a_k \qquad k = 1, \ldots, n \end{cases} \tag{5.4}$$

e prende il nome di "**algoritmo di Horner**"; in sostanza esso si riduce al calcolo di una moltiplicazione ed un'addizione per passo.

Figura 5.1 - William G. Horner (1786–1837)

La (5.4) può essere rappresentata mediante il seguente schema grafico:

$$\begin{array}{ccccccccc} a_0 & & a_1 & & a_2 & & \ldots\ldots & & a_n \\ \updownarrow = & & \downarrow + & & \downarrow + & & & & \downarrow + \\ b_0 & \xrightarrow{*z} & b_1 & \xrightarrow{*z} & b_2 & \xrightarrow{*z} & \ldots\ldots & \xrightarrow{*z} & b_n \end{array}$$

Osserviamo subito che la (5.4) richiede il calcolo di n moltiplicazioni, cioè un numero inferiore a $2n-1$ quando $n \geq 2$.

L'algoritmo con cui si calcola il valore del polinomio ha una considerevole influenza sulla propagazione degli errori inerenti e di arrotondamento. A titolo di esempio vediamo come cambia la stime dell'errore nella valutazione di un polinomio di secondo grado se si usa l'algoritmo di Horner piuttosto che quello classico.

Supponiamo di voler calcolare in un punto z il

$$P_2(x) = ax^2 + bx + c.$$

Per ottenere l'errore relativo propagato utilizziamo lo schema NONOR introdotto nel capitolo 4.

Algoritmo di Horner

$$P_2(z) = az^2 + bz + c = (az + b)\, z + c.$$

Indichiamo con α_1 e α_2 gli errori di arrotondamento nelle due addizioni, con β_1 e β_2 gli errori di arrotondamento nella due moltiplicazioni e con δ_t l'errore relativo nel generico elemento t.

$$\begin{aligned}
\delta_{P(z)} &= \delta_{(az+b)z}\frac{(az+b)\,z}{P\,(z)} + \delta_c\frac{c}{P\,(z)} + \alpha_1 = \\
&= \left(\delta_{az+b} + \delta_z + \beta_1\right)\frac{(az+b)\,z}{P\,(z)} + \delta_c\frac{c}{P\,(z)} + \alpha_1 = \\
&= \left[\left(\delta_{az}\frac{az}{az+b} + \delta_b\frac{b}{az+b} + \alpha_2\right) + \delta_z + \beta_1\right]\frac{(az+b)\,z}{P\,(z)} + \delta_c\frac{c}{P\,(z)} + \alpha_1 = \\
&= \left[\left(\delta_a + \delta_z + \beta_2\right)\frac{az}{az+b} + \delta_b\frac{b}{az+b} + \alpha_2 + \delta_z + \beta_1\right]\frac{(az+b)\,z}{P\,(z)} + \delta_c\frac{c}{P\,(z)} + \alpha_1 = \\
&= \left[\left(\delta_a + \delta_z + \beta_2\right)az + \delta_b b + \left(\alpha_2 + \delta_z + \beta_1\right)\left(az+b\right)\right]\frac{z}{P\,(z)} + \delta_c\frac{c}{P\,(z)} + \alpha_1
\end{aligned}$$

Per ottenere l'errore assoluto $E_{P(z)}$ moltiplichiamo per $P\,(z)$

$$\begin{aligned}
E_{P(z)} &= \left[\left(\delta_a + \delta_z + \beta_2\right)az + \delta_b b + \left(\alpha_2 + \delta_z + \beta_1\right)\left(az+b\right)\right]z + \delta_c c + \alpha_1\left(az^2 + bz + c\right) = \\
&= \left(\delta_a + \delta_z + \beta_2\right)az^2 + \delta_b bz + \left(\alpha_2 + \delta_z + \beta_1\right)\left(az^2 + bz\right) + \delta_c c + \alpha_1\left(az^2 + bz + c\right) = \\
&= \left(\delta_a + \delta_z + \beta_2 + \alpha_2 + \delta_z + \beta_1 + \alpha_1\right)az^2 + \left(\delta_b + \alpha_2 + \delta_z + \beta_1 + \alpha_1\right)bz + \left(\alpha_1 + \delta_c\right)c.
\end{aligned}$$

Usando un'aritmetica finita con m cifre di mantissa in floating-point, se $|z| \leq 1$, abbiamo

$$\left|E_{P(z)}\right| \leq 5 \cdot 10^{-m}\left(7|a| + 5|b| + 2|c|\right) \equiv E_H. \tag{5.5}$$

Algoritmo classico

$$P_2(z) = az^2 + bz + c = a \cdot \left(z \cdot z\right) + \left[\left(b \cdot z\right) + c\right]$$

Indichiamo con α_1 e α_2 gli errori di arrotondamento nelle due addizioni, con β_1, β_2 e β_3 gli errori di arrotondamento nella tre moltiplicazioni e con δ_t l'errore relativo nel generico elemento t.

$$\begin{aligned}
\delta_{P(z)} &= \delta_{az^2}\frac{az^2}{P\,(z)} + \delta_{bz+c}\frac{bz+c}{P\,(z)} + \alpha_1 = \\
&= \left(\delta_a + \delta_{z^2} + \beta_1\right)\frac{az^2}{P\,(z)} + \left(\delta_{bz}\frac{bz}{bz+c} + \delta_c\frac{c}{bz+c} + \alpha_2\right)\frac{bz+c}{P\,(z)} + \alpha_1 = \\
&= \left[\delta_a + \left(\delta_z + \delta_z + \beta_2\right) + \beta_1\right]\frac{az^2}{P\,(z)} + \left(\delta_{bz}\frac{bz}{bz+c} + \delta_c\frac{c}{bz+c} + \alpha_2\right)\frac{bz+c}{P\,(z)} + \alpha_1 = \\
&= \left(\delta_a + 2\delta_z + \beta_2 + \beta_1\right)\frac{az^2}{P\,(z)} + \left[\left(\delta_b + \delta_z + \beta_3\right)\frac{bz}{bz+c} + \delta_c\frac{c}{bz+c} + \alpha_2\right]\frac{bz+c}{P\,(z)} + \alpha_1
\end{aligned}$$

Per ottenere l'errore assoluto $E_{P(z)}$ moltiplichiamo per $P(z)$

$$\begin{aligned} E_{P(z)} &= \Big(\delta_a + 2\delta_z + \beta_2 + \beta_1\Big)az^2 + \Big[\Big(\delta_b + \delta_z + \beta_3\Big)bz + \delta_c c + \alpha_2\Big(bz + c\Big)\Big] + \alpha_1\Big(az^2 + bz + c\Big) = \\ &= \Big(\delta_a + 2\delta_z + \beta_2 + \beta_1 + \alpha_1\Big)az^2 + \Big(\delta_b + \delta_z + \beta_3 + \alpha_2 + \alpha_1\Big)bz + \Big(\delta_c + \alpha_2 + \alpha_1\Big)c \end{aligned}$$

Usando un'aritmetica finita con m cifre di mantissa in floating-point, se $|z| \leq 1$, abbiamo

$$\Big|E_{P(z)}\Big| \leq 5 \cdot 10^{-m}\Big(6|a|+5|b|+3|c|\Big) \equiv E. \tag{5.6}$$

Da (5.5) e (5.6)

$$E - E_H = 5 \cdot 10^{-m}\left(|c| - |a|\right),$$

quindi se $|c| > |a|$ l'algoritmo di Horner ha un limite più piccolo per l'errore di arrotondamento.

Esempio 5.1 *Si vuole dimostrare che, se $|x| \leq 1$, per il polinomio di grado n*

$$P(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

l'errore derivante dall'algoritmo di Horner è limitato da

$$|e_P| \leq 5 \cdot 10^{-m}\left(\sum_{j=0}^{n}(3j+2)\,|a_j| - |a_n|\right). \tag{5.7}$$

Nell'esempio precedente è stato dimostrato che la (5.7) *vale se $P(x)$ è un polinomio di grado* 2*. Supponiamo che essa è vera per un polinomo di grado $n-1$ e dimostriamo che vale per $P_n(x)$ di grado n.*

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n = x\left(a_1 + a_2 x + \ldots + a_n x^{n-1}\right) + a_0,$$

quindi

$$P_n(x) = xQ(x) + a_0$$

con

$$Q(x) = \left(a_1 + a_2 x + \ldots + a_n x^{n-1}\right) \equiv \left(b_0 + b_1 x + \ldots + b_{n-1} x^{n-1}\right).$$

$Q(x)$ è un polinomio di grado $n-1$, dunque vale la (5.7)*, cioè*

$$\begin{aligned} |e_Q| &\leq 5 \cdot 10^{-m}\left(\sum_{j=0}^{n-1}(3j+2)\,|b_j| - |b_{n-1}|\right) \\ &\leq 5 \cdot 10^{-m}\left(\sum_{j=1}^{n}\left[3(j-1)+2\right]|a_j| - |a_n|\right). \end{aligned}$$

Calcoliamo ora l'errore relativo commesso in $P_n(x) = xQ(x) + a_0$. Se Re_x, Re_0 indicano rispettivamente gli errori relativi in x e in a_0, e α, β gli errori di arrotondamento nella moltiplicazione e nella somma, troviamo

$$Re_P = (Re_x + Re_Q + \alpha)\frac{xQ(x)}{P(x)} + Re_0\frac{a_0}{P(x)} + \beta,$$

da cui

$$e_P = (Re_x + \alpha + \beta)\,xQ(x) + xQ(x)Re_Q + (Re_0 + \beta)\,a_0,$$

e quindi

$$|e_P| \leq 3 \cdot 5 \cdot 10^{-m}\,|x|\,|Q(x)| + 2 \cdot 5 \cdot 10^{-m}\,|a_0| + |x|\,|e_Q|.$$

Sostituendo ora l'espressione di $Q(x)$ e la (u4eq1.1.14, e ricordando che $|x| \leq 1$, otteniamo la limitazione

$$|e_P| \leq 3 \cdot 5 \cdot 10^{-m} \sum_{j=1}^{n} |a_j| + 2 \cdot 5 \cdot 10^{-m} |a_0| + 5 \cdot 10^{-m} \left(\sum_{j=1}^{n} [3(j-1)+2] |a_j| - |a_n| \right),$$

cioè

$$|e_P| \leq 5 \cdot 10^{-m} \left\{ 3 \sum_{j=1}^{n} |a_j| + 2|a_0| + \sum_{j=1}^{n} [3(j-1)+2] |a_j| - |a_n| \right\},$$

da cui

$$\begin{aligned} |e_P| &\leq 5 \cdot 10^{-m} \left\{ \sum_{j=1}^{n} [3+3(j-1)+2] |a_j| + 2|a_0| - |a_n| \right\} \\ &\leq 5 \cdot 10^{-m} \left\{ \sum_{j=1}^{n} (3j+2) |a_j| + 2|a_0| - |a_n| \right\} \\ &\leq 5 \cdot 10^{-m} \left\{ \sum_{j=0}^{n} (3j+2) |a_j| - |a_n| \right\}. \end{aligned}$$

Esempio 5.2 *Per evidenziare quanto la propagazione dell'errore possa essere devastante nel calcolo del valore di un polinomio in un punto, e quanto sia sensibile al particolare algoritmo usato, consideriamo i seguenti polinomi*

$$p_1(x) = (x-1)^7 \qquad \longleftrightarrow \qquad p_2(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1.$$

Dal punto di vista dell'algebra le due espressioni sono identiche, ma se calcoliamo numericamente $p_1(x)$ e $p_2(x)$ nell'intervallo $[0.9998, 1.0002]$ e rappresentiamo il grafico otteniamo il seguente risultato!!!

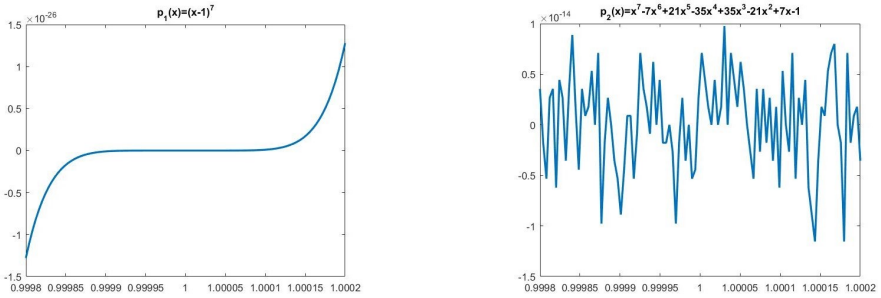


Figura 5.2

Ricordiamo, infine, che l'errore effettivo di propagazione è, salvo particolari situazioni, considerevolmente più piccolo dei limiti sopra determinati.

5.2 Divisioni sintetiche e regola di Ruffini

E' noto l'algoritmo della divisione di due polinomi, ovvero il

Teorema 5.1 *Dati due polinomi $P_n(x)$ e $F_m(x)$, con $n > m$, esistono e sono unici due altri polinomi $Q(x)$ e $R(x)$ tali che*

$$P_n(x) = Q(x)F_m(x) + R(x) \tag{5.8}$$

dove $R(x) \equiv 0$ oppure grado di $R(x) <$ grado di $F_m(x)$. $Q(x)$ è detto **quoziente** *e $R(x)$* **resto**.

Nel caso in cui $F_m(x)$ è un polinomio di primo grado o lineare, ossia

$$x - z,$$

la (5.8) diventa

$$P_n(x) = Q(x)(x - z) + R \tag{5.9}$$

da cui facilmente

$$P_n(z) = R,$$

ovvero il valore del polinomio $P_n(x)$ nel punto z è il resto nella divisione di $P_n(x)$ per $x - z$. Se $R = 0$, z è una radice del polinomio $P_n(x)$.
Si è già dimostrato che $P_n(z)$, ovvero R, uguaglia il termine b_n della sequenza finita

$$\begin{cases} b_0 = a_0 \\ b_k = b_{k-1} \cdot z + a_k \qquad k = 1, \ldots, n. \end{cases}$$

Ci chiediamo, ora, quale sia il significato degli altri termini $b_0, b_1, \ldots, b_{n-1}$ della sequenza. A tale scopo supponiamo che il polinomio $Q(x)$ in (5.9), ovvero il polinomio quoziente, sia scritto nella forma

$$Q(x) = c_0 x^{n-1} + c_1 x^{n-2} + \ldots + c_{n-2} x + c_{n-1};$$

in tal modo per il prodotto $Q(x)(x - z)$ si ha

$$\begin{aligned} Q(x)(x - z) &= \left(c_0 x^{n-1} + c_1 x^{n-2} + \ldots + c_{n-2} x + c_{n-1}\right)(x - z) = \\ &= c_0 x^n + (c_1 - z c_0)\, x^{n-1} + \ldots + (c_{n-1} - z c_{n-2})\, x - z c_{n-1} \end{aligned}$$

e quindi la (5.9) diventa

$$\begin{cases} a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n = \\ = c_0 x^n + (c_1 - z c_0)\, x^{n-1} + \ldots + (c_{n-1} - z c_{n-2})\, x - z c_{n-1} + b_n. \end{cases} \tag{5.10}$$

Applicando il principio di identità dei polinomi, da (5.10) ricaviamo le uguaglianze

$$\begin{cases} a_0 = c_0 \\ a_1 = c_1 - z c_0 \\ a_2 = c_2 - z c_1 \\ \vdots \\ a_{n-1} = c_{n-1} - z c_{n-2} \\ a_n = b_n - z c_{n-1} \end{cases} \qquad \Longrightarrow \qquad \begin{cases} c_0 = a_0 \\ c_1 = z c_0 + a_1 \\ c_2 = z c_1 + a_2 \\ \vdots \\ c_{n-1} = z c_{n-2} + a_{n-1}. \end{cases} \tag{5.11}$$

Dal confronto di (5.11) con (5.3) troviamo

$$\begin{cases} c_0 = a_0 = b_0 \\ c_1 = b_1 \\ c_2 = b_2 \\ \vdots \\ c_{n-1} = b_{n-1}, \end{cases}$$

cioè i numeri $b_0, b_1, \ldots, b_{n-1}$ sono i coefficienti del polinomio quoziente di $P_n(x)$ per $(x - z)$.

Figura 5.3 - Paolo Ruffini (1765–1822)

In questo contesto l'algoritmo di Horner è più noto col nome di ***regola di Ruffini*** e in molti manuali scolastici della scuola inferiore è riportato con lo schema grafico seguente

$$\begin{array}{c|cccc|c} & a_0 & a_1 & \dots & a_{n-1} & a_n \\ z & & & & & \\ \hline & b_0 & b_1 & \dots & b_{n-1} & b_n \end{array}$$

5.2.1 Calcolo delle derivate di un polinomio in un punto

Riprendiamo il teorema 5.1 di fattorizzazione

$$P_n(x) = Q(x)\,(x - z) + R$$

e deriviamo ambo i membri

$$P_n'(x) = Q'(x)\,(x - z) + Q(x). \tag{5.12}$$

Per $x = z$ si ha

$$P_n'(z) = Q(z),$$

ovvero il valore del polinomio quoziente di $P_n(x)$ per $(x - z)$, nel punto $x = z$, uguaglia il valore della derivata prima del polinomio $P_n(x)$, sempre in $x = z$. Il valore $Q(z)$ può, a sua volta, essere calcolato con l'algoritmo di Horner, per cui, avendo dimostrato nel paragrafo precedente che i coefficienti del polinomio $Q(x)$ sono i numeri $b_0, b_1, \dots, b_{n-1}$ calcolati nella sequenza (5.4), si ha

$$\begin{cases} c_0 = b_0 \\ c_k = c_{k-1} \cdot z + b_k \qquad k = 1, \dots, n-1 \end{cases}$$

e

$$P_n'(z) = Q(z) = c_{n-1}.$$

Il calcolo dei coefficienti $c_0, c_1, \dots, c_{n-1}$ può essere rappresentato nello schema

$$\begin{array}{ccccccccc} b_0 & & b_1 & & b_2 & & \dots\dots & & b_{n-1} \\ \updownarrow = & & \downarrow + & & \downarrow + & & & & \downarrow + \\ c_0 & \xrightarrow{*z} & c_1 & \xrightarrow{*z} & c_2 & \xrightarrow{*z} & \dots\dots & \xrightarrow{*z} & c_{n-1} \end{array} \quad = Q(z) = P_n'(z)$$

Derivando ancora la (5.12) si ottiene

$$P_n''(x) = Q''(x)\,(x - z) + 2Q'(x) \tag{5.13}$$

che per $x = z$ diventa

$$P_n''(z) = 2Q'(z).$$

$Q(x)$ è un polinomio di grado $n-1$, quindi $Q'(x)$ nel punto z può essere calcolato con il procedimento su esposto. Ne segue che anche $P_n''(x)$ nel punto z, a meno del fattore costante 2, può essere calcolata con l'algoritmo di Horner. Risulta chiaro perciò lo schema grafico seguente

$$
\begin{array}{ccccccccccc}
a_0 & & a_1 & & \dots & & a_{n-2} & & a_{n-1} & & a_n \\
\updownarrow= & & \downarrow + & & & & \downarrow + & & \downarrow + & & \downarrow + \\
b_0 & \xrightarrow{*z} & b_1 & \xrightarrow{*z} & \dots & \xrightarrow{*z} & b_{n-2} & \xrightarrow{*z} & b_{n-1} & \xrightarrow{*z} & P_n(z) \\
\updownarrow= & & \downarrow + & & & & \downarrow + & & \downarrow + & & \\
c_0 & \xrightarrow{*z} & c_1 & \xrightarrow{*z} & \dots & \xrightarrow{*z} & c_{n-2} & \xrightarrow{*z} & P_n'(z) & & \\
\updownarrow= & & \downarrow + & & & & \downarrow + & & & & \\
d_0 & \xrightarrow{*z} & d_1 & \xrightarrow{*z} & \dots & \xrightarrow{*z} & \dfrac{P_n'(z)}{2} & & & &
\end{array}
$$

Il procedimento si può iterare , infatti la (5.13) può essere ancora derivata e poi calcolata nel punto $x = z$. Si può quindi dimostrare che le sequenze

$$
\begin{cases}
b_k^{(0)} = b_k & k = 0, \dots, n \\
b_0^{(i)} = b_0^{(i-1)} & i = 1, \dots, n \\
b_k^{(i)} = z b_{k-1}^{(i)} + b_k^{(i-1)} & \begin{array}{l} i = 1, \dots, n \\ k = 1, \dots, n-i \end{array}
\end{cases}
\tag{5.14}
$$

forniscono

$$
P_n^{(i)}(z) = i!\; b_{n-i}^{(i)} \qquad i = 1, \dots, n. \tag{5.15}
$$

Infatti le (5.14) esprimono la costruzione della sequenza $b_k^{(i)}$ applicando lo schema di Horner al polinomio di coefficienti $b_k^{(i-1)}$. In queste condizioni, se indichiamo con $Q_i(x)$, $\; i = 0, 1, \dots, n$, il polinomio di grado $n-i-1$ definito dai coefficienti $b_k^{(i)}$, che si trovano sulla riga i dello schema triangolare, possiamo scrivere

$$
\begin{aligned}
& P_n(x) = (x - z)\, Q_0(x) + b_n^{(0)}, \\
& Q_0(x) = (x - z)\, Q_1(x) + b_{n-1}^{(1)}, \\
& \vdots \\
& Q_{i-1}(x) = (x - z)\, Q_i(x) + b_{n-i}^{(i)}, \qquad i = 1, \dots, n-1
\end{aligned}
$$

e

$$
b_{n-i}^{(i)} = Q_{i-1}(z).
$$

Sostituendo l'espressione dei $Q_i(x)$ in $P_n(x)$ troviamo

$$
\begin{aligned}
P_n(x) &= (x - z)\, Q_0(x) + b_n^{(0)} = (x - z)^2\, Q_1(x) + (x - z)\, b_{n-1}^{(1)} + b_n^{(0)} = \\
&= (x - z)^i\, Q_{i-1}(x) + (x - z)^{i-1}\, b_{n-(i-1)}^{(i-1)} + \dots + (x - z)\, b_{n-1}^{(1)} + b_n^{(0)}.
\end{aligned}
$$

Derivando i volte ambo i membri e calcolando le espressioni in z troviamo

$$
P_n^{(i)} = i \cdot (i-1) \cdot \dots \cdot 2 \cdot 1 \cdot Q_{i-1}(z) = i! b_{n-i}^{(i)}.
$$

Le (5.14) possono essere racchiuse nel seguente schema triangolare

$$
\begin{array}{ccccccccc}
a_0 & & a_1 & \dots & a_{n-2} & & a_{n-1} & & a_n \\
\updownarrow= & & \downarrow + & & \downarrow + & & \downarrow + & & \downarrow + \\
b_0^{(0)} & \xrightarrow{*z} & b_1^{(0)} & \dots & b_{n-2}^{(0)} & \xrightarrow{*z} & b_{n-1}^{(0)} & \xrightarrow{*z} & b_n^{(0)} \\
\updownarrow= & & \downarrow + & & \downarrow + & & \downarrow + & & \\
b_0^{(1)} & \xrightarrow{*z} & b_1^{(1)} & \dots & b_{n-2}^{(1)} & \xrightarrow{*z} & b_{n-1}^{(1)} & & \\
\updownarrow= & & \downarrow + & & \downarrow + & & & & \\
b_0^{(2)} & \xrightarrow{*z} & b_1^{(2)} & \dots & b_{n-2}^{(2)} & & & & \\
\vdots & & & \iddots & & & & & \\
\vdots & & \iddots & & & & & & \\
b_0^{(n)} & & & & & & & &
\end{array}
$$

Tralasciando i simboli ed essendo, da (5.15), $b_{n-i}^{(i)} = \dfrac{P_n^{(i)}}{i!} \quad i = 1, \dots, n.$,

$$
\begin{array}{ccccccc}
a_0 & a_1 & \dots & a_{n-2} & a_{n-1} & a_n \\
b_0^{(0)} & b_1^{(0)} & \dots & b_{n-2}^{(0)} & b_{n-1}^{(0)} & \dfrac{P_n(z)}{0!} \\
b_0^{(1)} & b_1^{(1)} & \dots & b_{n-2}^{(1)} & \dfrac{P_n'(z)}{1!} & \\
b_0^{(2)} & b_1^{(2)} & \dots & \dfrac{P_n''(z)}{2!} & & \\
\vdots & & \iddots & & & \\
\vdots & \iddots & & & & \\
\dfrac{P_n^{(n)}(z)}{n!} & & & & &
\end{array}
$$

5.3 Polinomi e Matlab

Per quanto riguarda i polinomi, o in generale le equazioni algebriche, MatLab mette a disposizioni alcune funzioni specifiche. In particolare

roots(p) calcola le radici del polinomio che ha come coefficienti, in ordine decrescente di grado, le componenti del vettore `p`, precedentemente definito.

poly(r) calcola i coefficienti del polinomio le cui radici sono specificate nel vettore `r`

polyval(p,z) calcola il valore del polinomio, con coefficienti specificati nel vettore `p`, nel punto `z`

polyder(p) calcola i coefficienti della derivata prima del polinomio i cui coefficienti sono specificati nel vettore `p`

Capitolo 6

Radici di equazioni non lineari

Le scienze matematiche mostrano ordine, simmetria e limite: e queste sono le più grandi istanze del bello.
Aristotele (384 a.C.–322 a.C)

6.0 Introduzione

6.1 Radici di equazioni

Il calcolo delle radici di un'equazione

$$f(x) = 0 \tag{6.1}$$

è uno dei più antichi problemi matematici. Esso è di notevole importanza perché costituisce il modello matematico di numerosi problemi applicativi.

Esempio 6.1 *Si vuole calcolare a quanti gradi Celsius, C, corrisponde la temperatura di $0°$ gradi Fahrenheit, F. Essendo C e F legate dalla relazione $F = \frac{9}{5}C + 32$, occorre risolvere l'equazione*

$$\frac{9}{5}C + 32 = 0.$$

Esempio 6.2 *Si vuole calcolare la percentuale di ioni di una soluzione $0.0011M$ di acido formico, a $25°C$, la cui equazione di equilibrio è*

$$HCO_2H \rightleftharpoons H^+HCO_2^-.$$

L'equazione è

$$\frac{[H^+][HCO_2^-]}{[HCO_2H]} = K_a$$

dove $K_a = 1.6 \times 10^{-4} moll^{-1}$, e il simbolo [ ] significa "concentrazione di ioni". Ad esempio $[H^+]$ significa concentrazione di ioni di idrogeno.
Se x è la concentrazione di ioni di idrogeno allora

$$[H^+] = x \qquad [HCO_2^-] = x \qquad [HCO_2H] = 0.0011 - x$$

per cui bisogna risolvere l'equazione

$$\frac{x^2}{0.0011 - x} = 1.6 \times 10^{-4} \qquad \Longrightarrow \qquad x^2 + 1.6 \times 10^{-4}x - 1.76 \times 10^{-7} = 0.$$

Esempio 6.3 *L'equazione di Van der Waals per 1 mole di gas è*

$$p = \frac{RT}{V-b} - \frac{a}{V^2}$$

cioè

$$V^3 - (b + \frac{RT}{p})V^2 + \frac{a}{p}V - \frac{ab}{p} = 0$$

ove R è la costante dei gas, a e b sono costanti che dipendono dal gas in esame e p , V , T sono rispettivamente la pressione, il volume e la temperatura del gas.

Esempio 6.4 *Un filo elastico, lungo $6m$, è tale che la sua lunghezza aumenta di $2.5mm$ quando è teso con una forza di $1Kg$. I suoi estremi vengono fissati a due punti distanti esattamente $6m$ e posti alla stessa quota. Nel centro del filo viene applicato, molto lentamente, un peso di $4Kg$*

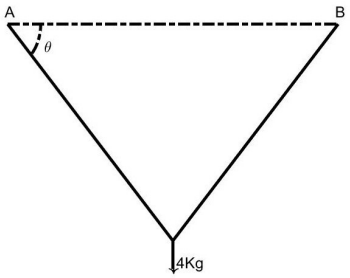


Figura 6.1

Vogliamo determinare la posizione di equilibrio trascurando il peso del filo ed ammettendo che resti valida la legge di Hooke (proporzionalità tra tensione e allungamento). Si ha l'equilibrio quando l'angolo soddisfa la seguente relazione

$$1200 \tan\theta(1 - \cos\theta) - 1 = 0 \quad .$$

Esempio 6.5 *La velocità di caduta di un paracadutista è data dalla formula*

$$v(t) = \frac{g\,m}{c}\left[1 - e^{-\frac{c}{m}t}\right]$$

ove $v(t)$ rappresenta la velocità in funzione del tempo, g è la costante di gravità, c è il coefficiente di resistenza aerodinamica ed m la massa del paracadutista.
Si vuole calcolare il valore del coefficiente di resistenza $\overline{c}$ (ossia scegliere il tipo di paracadute) per un paracadutista di una certa massa, affinchè egli possa avere velocità v_0 al dato istante t_0, cioe

$$v(t_0) = \frac{g\,m}{c}\left[1 - e^{-\frac{c}{m}t_0}\right] = v_0$$

ossia $\overline{c}$ è radice dell'equazione non lineare $f(x) = 0$, essendo

$$f(x) = \frac{gm}{x}\left[1 - e^{-\frac{x}{m}t_0}\right] - v_0.$$

Per calcolare la soluzione, supponendo $g = 980cm/s^2$, $m = 68\,100$ grammi massa, $t = 10s$, $v = 4487.3cm/s$, ccorre applicare un metodo numerico. Usando ad esempio il metodo di bisezione si trova $\overline{c} \approx 12\,500$.

Da questi pochi esempi considerati risulta che le difficoltà del problema, cioè della risoluzione dell'equazione (6.1), dipendono dalla natura di $f(x)$. Se $f(x)$ è algebrica di primo o di secondo grado (esempi 6.1 e 6.2) non sussistono difficoltà, essendo note le formule risolutive. Se $f(x)$ è algebrica di grado superiore al quarto o è trascendente, non esistono formule risolutive. Per $f(x)$ algebrica di terzo e quarto grado esistono formule risolutive per radicali, ma la loro non semplice struttura ne sconsiglia l'uso. Quindi nella maggior parte dei casi la soluzione dell'equazione (6.1) non si esprime in modo esplicito, per cui bisogna utilizzare un metodo numerico.

6.2 Equazioni algebriche

Le equazioni algebriche si ottengono uguagliando a zero un polinomio. Le radici sono calcolabili con formule chiuse solo per equazioni di grado inferiore a 5, ma anche grado 3 e 4 le formule risultano particolarmente complesse per cui in generesi preferisce usare metodi numerici.

6.2.1 Equazioni quadratiche

L'equazione quadratica generale è

$$ax^2 + bx + c = 0 \qquad \text{con} \qquad a, b, c \in \mathbb{R},\ a \neq 0.$$

Come è noto, le sue radici sono espresse dalla formula

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Se $\Delta = b^2 - 4ac \geq 0$ le due radici sono reali (coincidenti se $\Delta = 0$), altrimenti sono complesse coniugate:

$$x_{1,2} = \frac{-b \pm i\sqrt{4ac - b^2}}{2a}.$$

Quindi il problema del calcolo delle radici di un'equazione quadratica, almeno in aritmetica infinita, è completamente risolto. In floating-point, però, sorgono problemi relativi alla propagazione dell'errore di arrotondamento che possono generare algoritmi di calcolo instabili.
La nota formula per il calcolo delle radici di una equazione di secondo grado presenta problemi di stabilità numerica. Infatti se $4ac$ è prossimo a 0 in una delle due radici si verifica un problema di cancellazione numerica. In questo caso si può ricorrere alla formula alternativa

$$x_{1,2} = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}}.$$

Inoltre ricordando che se x_1 e x_2 sono le radici dell'equazione quadratica, valgono le relazioni

$$x_1 + x_2 = -\frac{b}{a}$$
$$x_1 \cdot x_2 = \frac{c}{a},$$

calcolata la radice x_1 con il segno che non comporta cancellazione, x_2 si può calcolare con la formula alternativa $x_2 = \dfrac{c}{ax_1}$. Si può dimostrare che tale procedimento risulta più stabile di quello classico.

6.3 Equazioni non algebriche

Esistono equazioni non algebriche, ovvero che contengono funzioni trigonometriche, logaritmiche o esponenziali. Anche equazioni di questo tipo sono modelli matematici di fenomeni della realtà sensibile, per cui si pone il problema dell'esistenza e del relativo calcolo delle radici. In questa breve sezione, dopo la definizione, accenneremo al problema dell'esistenza di radici.
Le equazioni che non sono algebriche si chiamano trascendenti.

Esempio 6.6 *Sono equazioni trascendenti:*

1) $M = E - e \sin E$ — *equazione di Keplero che lega il moto di due corpi celesti;*

2) $0.1x^2 - x \log x = 0$*;*

3) $e^x - 1.0 - \cos x$*.*

Se

$$f(x) = 0 \tag{6.2}$$

è un'equazione trascendente, un numero α tale che

$$f(\alpha) = 0$$

si chiama *radice* (o *soluzione*) dell'equazione (6.2) o zero della funzione $f(x)$. Anche in questo caso si pone il problema dell'esistenza e del calcolo delle radici.
Per quanto attiene all'esistenza vale il seguente

Teorema 6.1 (Esistenza degli zeri). *Se $f(x)$ è una funzione continua in un intervallo $[a, b]$ tale che $f(a)f(b) < 0$, allora esiste almeno un punto $\alpha \in (a, b)$ tale che $f(\alpha) = 0$.*

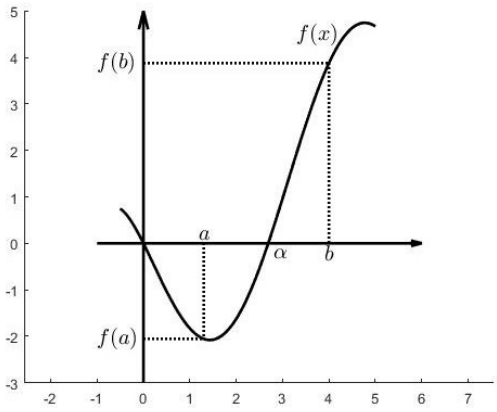


Figura 6.2

Per quanto riguarda il calcolo, anche per le equazioni trascendenti non esistono, in generale, metodi che forniscono la soluzione in forma chiusa, ovvero calcolabile con un numero finito di operazioni elementari (somma, prodotto, estrazione di radice). Occorre, perciò, determinare metodi di approssimazione, ossia procedimenti che calcolano una sequenza di valori $x_1, x_2, \ldots$ tale che

$$\lim_{n\to\infty} x_n = \alpha \tag{6.3}$$

se α è radice dell'equazione (6.2).
I procedimenti numerici che generano la successione $\{x_n\}$, verificante la (6.3), generalmente devono essere inizializzati con un valore x_0, o con più valori, spesso un intervallo, che devono essere approssimazioni della radice cercata. Tali valori possono essere ricavati rappresentando graficamente la funzione $y = f(x)$, se l'equazione è $f(x) = 0$, oppure tramite tabulazione, applicando il teorema 6.1.

Esempio 6.7 *Per determinare un intervallo in cui è compresa la radice della equazione*

$$x - \cos x = 0, \tag{6.4}$$

con un semplice codice di calcolo si può tracciare il grafico della funzione $y = x - \cos x$ da cui risulta l'esistenza di una radice nell'intervallo $(0, 1)$.
Alternativamente la (6.4) può essere scritta nella forma $x = \cos x$, e considerare un intervallo in cui cade l'intersezione della retta $y = x$ e della curva $y = \cos x$.

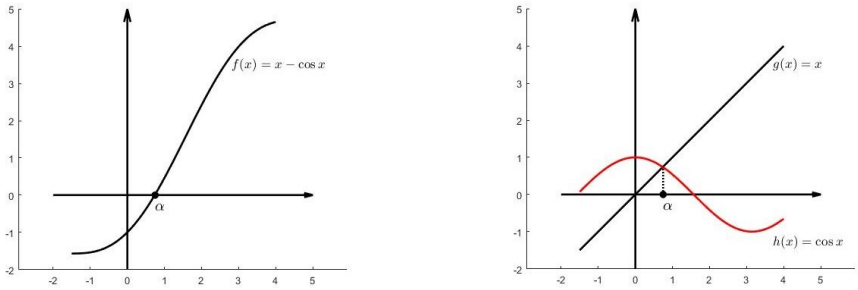


Figura 6.3

6.4 Calcolo approssimato di radici di equazioni non lineari

Come già detto, un metodo numerico per la risoluzione di un'equazione consiste in un procedimento di successive approssimazioni, mediante il quale si calcola una successione di numeri:

$$x_0, x_1, x_2, \ldots, x_n, \ldots \tag{6.5}$$

che converge alla radice, z, dell'equazione.
Riassumendo possiamo dire che per calcolare le radici di un'equazione non lineare

$$f(x) = 0$$

algebrica o trascendente, per la quale cui non esistono formule risolutive, occorre:

a) isolare la radice z, cioè determinare un intervallo $[a, b]$ tale che $z \in]a, b[$;

b) calcolare una sequenza di numeri $x_0, x_1, x_2, \ldots, x_n, \ldots$ che converge alla radice z.

Il punto a) è risolvibile con metodi grafici, cioè tracciando l'andamento della funzione $y = f(x)$, o per tabulazione della funzione f. Per il punto b) occorre individuare un opportuno procedimento di successive approssimazioni.
Anche in assenza di errori di arrotondamento, generalmente, nessuno dei valori della successione (6.5) è la radice della (6.1), cioè per nessun $x_k \in \{x_i\}$ risulta $f(x_k) = 0$, perciò scelta una tolleranza, ϵ, se

$$|x_{i+1} - x_i| < \epsilon$$

assumeremo $x_{i+1} \approx z$, a meno di ϵ.

6.5 Metodo di bisezione

Il metodo di bisezione è il metodo più semplice e intuitivo per localizzare e calcolare, con la precisione richiesta, la radice di un'equazione non lineare e si basa sul teorema di esistenza degli zeri. Data l'equazione non lineare

$$f(x) = 0 \tag{6.6}$$

supponiamo

a) che la funzione $f(x)$ sia continua in$]a, b[$ e dunque $f(x_1) f(x_2) < 0$;

b) di aver determinato, eventualmente con metodi grafici, l'intervallo $]a, b[$ che contiene una sola radice z di (6.6);

c) di aver fissato la tolleranza ϵ.

Se risulta

$$|b-a| < \epsilon \tag{6.7}$$

assumiamo per z il valore del punto medio dell'intervallo $[a,b]$ cioè poniamo

$$z \approx x_m = \frac{a+b}{2}, \tag{6.8}$$

altrimenti inneschiamo il seguente procedimento iterativo: si divide a metà l'intervallo $[a,b]$, ottenendo così due intervalli, uno solo dei quali contiene la radice; scelto quindi quello che la contiene, utilizzando il teorema 6.1, si procede con un nuovo dimezzamento e così via. La successione dei punti medi così costruita converge alla radice z.
Più in dettaglio, se la (6.7) non è verificata, l'intervallo $[a,b]$, che contiene la radice z, è ancora troppo ampio e l'approssimazione (6.8) infatti risulterebbe troppo grossolana. Dobbiamo allora cercare un intervallo più ristretto. In particolare, esaminiamo i due intervalli $[a,x_m]$ e $[x_m,b]$: uno dei due contiene la radice z (vedi fig.).

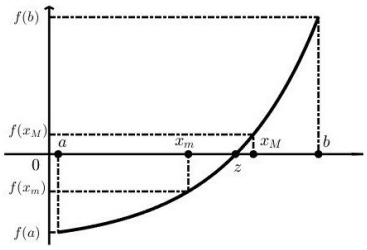


Figura 6.4

Se risulta $f(a)\,f(x_m) < 0$, la radice z appartiene all'intervallo $[a,x_m]$; nel caso contrario risulta $f(x_m)\,f(b) < 0$ e quindi z appartiene all'intervallo $[x_m,b]$.
Supponiamo che la radice z appartenga all'intervallo $[x_m,b]$ (il procedimento è perfettamente uguale nel caso $z \in [a,x_m]$); ripetiamo su di esso le stesse operazioni eseguite sull'intervallo iniziale $[a,b]$, ossia controlliamo l'ampiezza dell'intervallo:

$$|b-x_m|$$

se questa è minore della tolleranza stabilita, ϵ , assumiamo z uguale al punto medio di tale intervallo:

$$z \approx x'_m = \frac{b+x_m}{2}.$$

Se l'intervallo $[x_m,b]$ è ancora troppo ampio, si ripete il ragionamento precedente cercando un intervallo più ristretto che contenga la radice. Questo procedimento di successive approssimazioni porta sicuramente alla radice z con l'approssimazione voluta.
Costruiamo così la successione $x_0 \equiv a,\ x_1 \equiv b,\ , x_2, \ldots, x_k, \ldots$, con x_k punto medio dell'intervallo ottenuto dopo $k-1$ bisezioni,

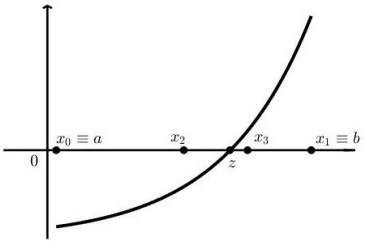


Figura 6.5

si verifica facilemente che

$$|x_k - z| < \frac{b-a}{2^k} \tag{6.9}$$

che dimostra la sicura convergenza del metodo di bisezione per ogni funzione continua. Infatti, per $k \to \infty$, l'ampiezza dell'intervallo in cui si trova la radice, tende a 0, cioè l'intervallo si riduce ad un punto che coincide con la radice cercata. La (6.9) è la stima dell'**errore di troncamento** del metodo di bisezione.

Esempio 6.8 *Supponiamo di voler calcolare una radice dell'equazione*

$$f(x) = x^5 - 3x^3 + 1 = 0$$

col metodo delle bisezioni successive.
Occorre anzitutto determinare l'intervallo $]x_1, x_2[$ *che contiene una radice. A tal fine tabuliamo* $f(x)$ *per diversi valori di* x*, finchè non compare un cambiamento di segno*

x	0.000	0.400	0.800
y	1.000	0.818	−0.208

C'è dunque una radice tra 0.4 *e* 0.8*. Consideriamo il punto medio e valutiamo in esso la funzione* $f(x)$

$$x = \frac{0.4 + 0.8}{2} = 0.6 \qquad f(0.6) = 0.42976.$$

Poichè $f(0.6)$ *è dello stesso segno di* $f(0.4)$*, ma di segno opposto a quello di* (0.8)*, la radice si trova tra* 0.6 *e* 0.8*.*
Consideriamo nuovamente il punto medio e valutiamo la funzione

$$x = \frac{0.6 + 0.8}{2} = 0.7 \qquad f(0.7) = 0.13907.$$

La radice si trova tra 0.7 *e* 0.8

$$x = \frac{0.7 + 0.8}{2} = 0.75 \qquad f(0.75) = -0.02832.$$

Ne segue che la radice è tra 0.7 *e* 0.75 *cioè* $z = 0.7\ldots$
Benchè piuttosto lento, il procedimento porta alla soluzione.

La (6.9) fornisce anche una stima a priori del numero di bisezioni necessarie per ottenere una precisione fissata. Per calcolare nell'intervallo $[a, b]$ uno zero di $f(x)$ con precisione ε, richiediamo che l'amppiezza dell'ultimo sottointervallo che contiene la radice sia $< \varepsilon$.

$$\frac{b-a}{2^n} < \varepsilon \quad \Longrightarrow \quad \frac{b-a}{\varepsilon} < 2^n \quad \Longrightarrow \quad n > \log_2 \frac{b-a}{\varepsilon}.$$

Quindi per ottenere precisione ε è sufficiente ripetere il processo n volte, dove n è il più piccolo intero che soddisfa $n > \log_2 \dfrac{b-a}{\varepsilon}$.

Si osserva che tale stima è del tutto indipendente falla funzione f di cui cerchiamo lo zero. Questo suggerisce che la velocità di convergenza del metodo può essere migliorata apportando opportune modifiche che sfruttano le proprietà della funzione.

6.6 Metodo della falsa posizione

Data l'equazione $f(x) = 0$, sia $[x_0, x_1]$ l'intervallo che contiene la radice z. Consideriamo la retta che unisce i punti $(x_0, f(x_0))$ e $(x_1, f(x_1))$ e sia x_2 l'intersezione con l'asse delle ascisse.

Osserviamo, poi, in quale dei due intervalli $[x_0, x_2]$ o $[x_2, x_1]$ cade la radice z, verificando la condizione $f(x_2)f(\overline{x}) < 0$ con $\overline{x} = x_0$ o $\overline{x} = x_1$; quindi consideriamo la retta che unisce i punti $(x_2, f(x_2))$ e $(\overline{x}, f(\overline{x}))$, e la relativa intersezione, x_3, con l'asse delle x.

Così proseguendo costruiamo una successione $\{x_k\}$ che converge alla radice z dell'equazione $f(x) = 0$ nella sola ipotesi di continuità della funzione $f(x)$.

Il metodo prende il nome di metodo della **falsa posizione**, poichè uno degli estremi della corda può variare ad ogni passo.

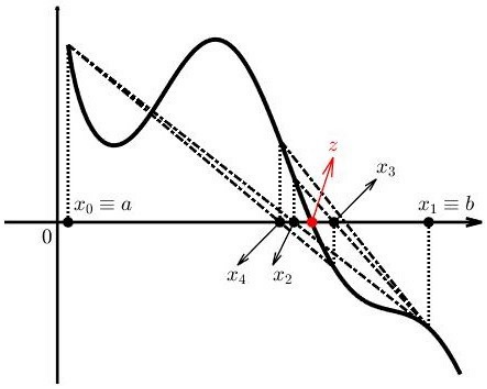


Figura 6.6

Si può dimostrare che almeno da un certo indice, k, in poi l'estremo $\overline{x}$ rimane fisso. L'equazione della retta per $(\overline{x}, f(\overline{x})), (x_k, f(x_k))$ è

$$\frac{f(\overline{x}) - y}{\overline{x} - x} = \frac{f(x_k) - f(\overline{x})}{x_k - \overline{x}}, \qquad k = 2, 3, \ldots$$

ovvero

$$y = f(x_k) - \frac{(f(x_k) - f(\overline{x}))(x_k - x)}{x_k - \overline{x}}. \tag{6.10}$$

L'intersezione della retta (6.10) con l'asse x è data da

$$x = x_k - \frac{(x_k - \overline{x}) f(x_k)}{f(x_k) - f(\overline{x})}$$

quindi la formula iterativa del metodo è

$$x_{k+1} = x_k - \frac{(x_k - \overline{x}) f(x_k)}{f(x_k) - f(\overline{x})}, \qquad k = 1, 2, \ldots$$

Se la funzione $f(x)$ è concava o convessa, cioè $f''(x) \neq 0 \quad \forall x \in [x_0, x_1]$, sin dall'inizio uno degli estremi delle corde rimane fisso e si può verificare che è il punto x_j tale che

$$f(x_j) f''(x_j) > 0 \qquad j = 0 \text{ o } j = 1$$

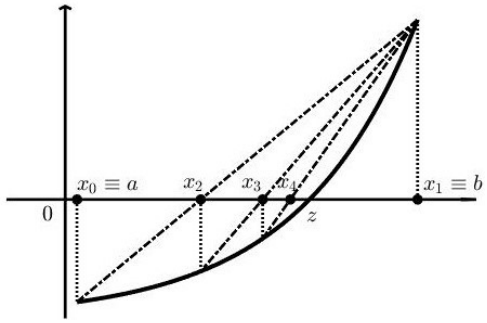


Figura 6.7

L'errore di troncamento del metodo della falsa posizione è dato da

$$z - x_{k+1} = -\frac{f''(\xi_k)}{2f'(\eta_k)}(z - x_k)(z - x_0), \qquad \xi_k \in (x_0, x_k, z), \quad \eta_i \in (x_0, x_k)$$

dove con (x_0, x_i, z) si intende il più piccolo intervallo contenente i punti x_0, x_i, z.

6.7 Metodo della secante o delle corde

Una variante del metodo della falsa posizione, che si può dimostrare, sotto talune ipotesi, avere una convergenza più rapida, è il metodo detto **della secante** o **delle corde**. Esso consiste nel considerare sempre la corda che unisce gli ultimi due punti $(x_{i-1}, f(x_{i-1})), (x_i, f(x_i)), \quad i = 1, 2, \ldots$ e quindi l'intersezione di tale retta con l'asse x come successivo punto della sequenza $\{x_i\}$.

Per ottenere la formula ricorsiva si considera la retta $(x_{i-1}, f(x_{i-1})), (x_i, f(x_i))$, quindi indicata con x_{i+1} l'intersezione di tale retta con l'asse x si ha

$$x_{i+1} = x_i - f(x_i)\frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} \qquad i = 1, 2, \ldots$$

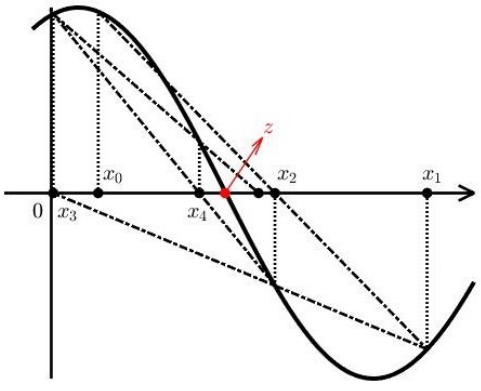


Figura 6.8

Per l'errore di troncamento si può provare la seguente

$$z - x_{k+1} = -\frac{f''(\overline{\xi}_k)}{2f'(\overline{\eta}_k)}(z - x_k)(z - x_{k-1}) \qquad \overline{\xi}_k \in (x_{k-1}, x_k, z), \quad \overline{\eta}_i \in (x_{k-1}, x_k).$$

6.8 Metodo di Newton-Raphson o delle tangenti

Il metodo presentati sono semplici metodo, a sicura convergenza o convergenti sotto opportune ipotesi, ma generalmente molto lenti (occorrono, cioè, molti passi per raggiungere una precisione accettabile).
Il metodo di Newton, o di Newton-Raphson, è un metodo un po' più sofisticato ma più veloce, ossia, un metodo che con pochi passi fornisce buoni risultati.

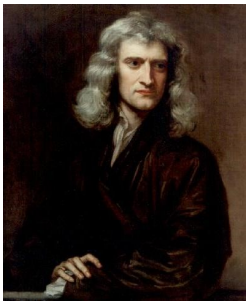


Figura 6.9 - Isaac Newton (1642-1726)

Esso consiste nel sostituire il grafico di $f(x)$ con la retta tangente in un punto opportuno della curva, ed assumere l'intersezione di tale retta con l'asse x, come ulteriore approssimazione della radice.
In altri termini, assegnato il punto x_0, che si suppone abbastanza vicino alla radice, si costruisce la successione $\{x_i\}$, $i = 0, 1, \ldots$ richiedendo che il punto x_{i+1} sia il punto d'intersezione con l'asse x della retta tangente alla curva $y = f(x)$, nel punto $(x_i, f(x_i))$, $\quad i = 0, 1, \ldots$.

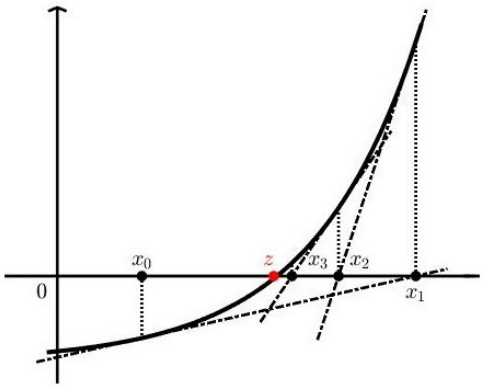


Figura 6.10

Per determinare la formula ricorsiva per il calcolo della successione x_i, ricordiamo che data la curva $y = f(x)$, l'equazione della retta tangente nel punto $(x_0, f(x_1))$ è data da

$$y = f'(x_0)(x - x_0) + f(x_0) \tag{6.11}$$

la cui intersezione con l'asse x è data da (ponendo $y = 0$ nella (6.11) e risolvendo rispetto ad x)

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{6.12}$$

Quindi il secondo punto della successione è

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Iterando questo procedimento si ottiene

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad i = 0, 1, \ldots \tag{6.13}$$

Osserviamo che la (6.13) si può ottenere prescindendo da ogni considerazione geometrica.
Sia data, infatti, l'equazione

$$f(x) = 0 \tag{6.14}$$

ed il valore approssimato, x_0, della sua radice z, consideriamo lo sviluppo in serie di Taylor di $f(x)$ di punto iniziale x_0:

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2!}(x - x_0)^2 f''(x_0) + \dots \tag{6.15}$$

Approssimando la funzione $f(x)$ con il polinomio $y(x)$ che si ottiene considerando i termini di primo grado del secondo membro della (6.15) si ha

$$f(x) \approx y(x) = f(x_0) + (x - x_0) f'(x_0).$$

Quindi sostituendo nella (6.14) $f(x)$ con l'approssimazione $y(x)$ otteniamo l'equazione

$$f(x_0) + (x - x_0) f'(x_0) = 0$$

la cui radice è data esattamente dalla (6.12); segue, poi, ripetendo lo stesso procedimento, la (6.13).

Esempio 6.9 *Vogliamo calcolare $\sqrt{N}, N > 0$ con un metodo di successive approssimazioni. Si tratta di risolvere l'equazione*

$$f(x) = x^2 - N = 0.$$

Risolviamo il problema con il metodo di Newton, assumendo come valore iniziale $x_0 = N$. La formula iterativa (6.13)

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \qquad i = 0, 1, \dots$$

tenendo conto che $f(x) = x^2 - N \quad e \quad f'(x) = 2x$, diventa

$$x_{i+1} = x_i - \frac{x_i^2 - N}{2x_i} \qquad i = 0, 1, \dots$$

o

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{N}{x_i}\right) \qquad i = 0, 1, \dots$$

Esempio 6.10 *Vogliamo calcolare il reciproco di un numero $c \in \mathbb{R}, c > 0$ con un metodo iterativo. Si tratta di risolvere l'equazione*

$$f(x) = \frac{1}{x} - c = 0.$$

Risolviamo il problema con il metodo di Newton, assumendo come valore iniziale $x_0 =$ ..... La formula iterativa (6.13)

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \qquad i = 0, 1, \dots$$

tenendo conto che $f(x) = \dfrac{1}{x} - c \quad e \quad f'(x) = -\dfrac{1}{x^2}$, diventa

$$x_{i+1} = x_i\big(2 - cx_i\big) \qquad i = 0, 1, \dots$$

Si può dimostrare che in questo caso il metodo di Newton converge se si sceglie x_0 tale che $0 < x_0 < \dfrac{1}{c}$.

Per il metodo di Newton si pone il problema della convergenza che dipende fortemente dalla funzione $f(x)$, ma l'argomento esula dal contesto di questo corso.

6.9 Sistemi di equazioni non lineari: il metodo di Newton

Il metodo di Newton può essere esteso alla soluzione dei sistemi di equazioni non lineari.
A tal fine consideriamo per semplicità un sistema non lineare di due equazioni in due incognite

$$\begin{cases} f(x,y) = 0 \\ g(x,y) = 0 \end{cases} \tag{6.16}$$

e così come abbiamo fatto per la singola equazione, sostituiamo le funzioni $f(x,y)$ e $g(x,y)$ con lo sviluppo di Taylor di punto iniziale (x_0, y_0), arrestato al primo termine

$$\begin{cases} f(x,y) \approx f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0) \\ g(x,y) \approx g(x_0,y_0) + g_x(x_0,y_0)(x-x_0) + g_y(x_0,y_0)(y-y_0) \end{cases}$$

Per semplicità nel seguito usiamo il simbolo $=$ al posto di $\approx$ e poniamo $\delta x = x - x_0$ e $\delta y = y - y_0$, al posto di (6.16); risolviamo quindi il sistema lineare

$$\begin{cases} f_x(x_0,y_0)\,\delta x + f_y(x_0,y_0)\,\delta y = -f(x_0,y_0) \\ g_x(x_0,y_0)\,\delta x + g_y(x_0,y_0)\,\delta y = -g(x_0,y_0) \end{cases} \tag{6.17}$$

Se la soluzione fosse esatta e ottenuta dall'intero sviluppo avremmo

$$\begin{cases} x = x_0 + \delta x \\ y = y_0 + \delta y. \end{cases}$$

Consideriamo quindi come nuova stima della soluzione di (6.16)

$$\begin{cases} x_1 = x_0 + \delta x \\ y_1 = y_0 + \delta y. \end{cases}$$

Naturalmente il procedimento si può iterare, e osservando che da (6.17) risulta

$$\begin{aligned} \delta x &= \frac{g(x_0,y_0)\,f_y(x_0,y_0) - f(x_0,y_0)\,g_y(x_0,y_0)}{f_x(x_0,y_0)\,g_y(x_0,y_0) - g_x(x_0,y_0)\,f_y(x_0,y_0)} \\ \delta y &= \frac{f(x_0,y_0)\,g_x(x_0,y_0) - g(x_0,y_0)\,f_x(x_0,y_0)}{f_x(x_0,y_0)\,g_y(x_0,y_0) - g_x(x_0,y_0)\,f_y(x_0,y_0)} \end{aligned}$$

abbiamo la formula iterativa del metodo di Newton per sistemi non lineari di due equazioni in due incognite:

$$\begin{cases} x_{n+1} = x_n - \dfrac{f(x_n,y_n)\,g_y(x_n,y_n) - g(x_n,y_n)\,f_y(x_n,y_n)}{f_x(x_n,y_n)\,g_y(x_n,y_n) - g_x(x_n,y_n)\,f_y(x_n,y_n)} \\ y_{n+1} = y_n - \dfrac{g(x_n,y_n)\,f_x(x_n,y_n) - f(x_n,y_n)\,g_x(x_n,y_n)}{f_x(x_n,y_n)\,g_y(x_n,y_n) - g_x(x_n,y_n)\,f_y(x_n,y_n)} \end{cases} \qquad n = 0, 1, \ldots$$

Anche per il metodo di Newton per sistemi non lineari si pone il problema della convergenza, ma l'argomento esula dal contesto di questo corso.

6.10 Zeri di funzioni e Matlab

Per il calcolo degli zeri di funzioni qualsiasi MatLab dispone della funzione

fzero('funzione',x0).

Essa cerca uno zero della funzione specificata, in prossimità di `x0`, se `x0` è uno scalare. Il valore restituito da `fzero` è prossimo ad un punto in cui la funzione cambia segno. Se la ricerca fallisce, `fzero` restituisce il valore `NaN`. Se `x0` è un vettore di lunghezza 2, possiamo supporre che le componenti di `x0` rappresentino gli estremi di un intervallo tale che il segno della funzione in `x0(1)` sia diverso dal segno in `x0(2)`. Eseguendo `fzero` con tale input si ha la garanzia che restituirà un valore vicino ad un punto in cui la funzione cambia segno.

Capitolo 7

Sistemi lineari

Non preoccuparti delle difficoltà che incontri in matematica, ti posso assicurare che le mie sono ancora più grosse.

Albert Einstein (1879-1955)

7.1 Cenni sulle matrici

Definizione 7.1. *Un insieme di $n \times m$ numeri, $n, m \in \mathbb{N}$, reali o complessi disposti su n righe ed m colonne chiamasi* **matrice** *di ordine $n \times m$, ed in particolare ad n righe ed m colonne. I numeri a_{ij} con $i = 1, \dots, n$ e $j = 1, \dots, m$ dove i indica il numero della riga, e j il numero della colonna, sono gli elementi della matrice, la matrice A di elementi a_{ij} è indicata con $A = (a_{ij})$*

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1m} \\ \vdots & & & \vdots & & \vdots \\ a_{i1} & \cdots & \cdots & a_{ij} & \cdots & a_{im} \\ \vdots & & & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nm} \end{pmatrix} \longleftarrow \text{riga } i$$

$$\uparrow$$

$$\text{colonna } j$$

Dati gli scopi del corso, nel seguito si farà sempre riferimento a sole matrici reali.

Le matrici sono caratterizzate dalla dimensione

- **matrice quadrata** di ordine n si ottiene per $n = m$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{pmatrix};$$

- **matrice rettangolare** si ottiene per $n \neq m$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nm} \end{pmatrix};$$

- **vettore colonna** si ottiene per $m = 1$ $\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$;
- **vettore riga** si ottiene per $n = 1$ $\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \end{pmatrix}$.

Quando non diversamente specificato, i vettori sono da intendersi colonna.

Alcuni vettori e matrici particolari sono caratterizzati dalla natura degli elementi

- vettori e matrici **costanti**: vettori e matrici aventi tutti gli elementi uguali
- la matrice che si ottiene da una matrice A, scambiando le righe con le colonne, è detta matrice **trasposta** e si indica con A^T

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & \dots & & \vdots \\ \vdots & & & \vdots \\ a_{n1} & \dots & \dots & a_{nm} \end{pmatrix}, \qquad A^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & \dots & & \vdots \\ \vdots & & & \vdots \\ a_{1m} & \dots & \dots & a_{nm} \end{pmatrix}.$$

$$A = (a_{ij}) \quad \Rightarrow \quad A^T = (a_{ji}) \qquad \begin{cases} i & = 1, \dots, n \\ j & = 1, \dots, n \end{cases}$$

- la matrice **identità** è la matrice I avente tutti elementi uguali a 1 sulla diagonale e 0 fuori diagonale.

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

7.2 Alcune matrici particolari

Alcune particolari matrici, con elementi predefiniti, sono già disponibili in Matlab (paragrafo 7.13). Si tratta di matrici presenti in diverse applicazioni che suggeriscono spunti interessanti per la realizzazione di algoritmi per la costruzione "manuale". Tra le altre ricordiamo

- **Matrice di Hilbert**

Figura 7.1 - David Hilbert (1862-1943)

$$H_n = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \iddots & & & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \iddots & & & \frac{1}{n} & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & & & \iddots & \frac{1}{n+1} & \frac{1}{n+2} \\ & & & & & & \\ & & & & & & \\ & \iddots & \iddots & & & & \frac{1}{2n-1} \\ \frac{1}{n} & \frac{1}{n+1} & & & & \frac{1}{2n-2} & \frac{1}{2n-1} \end{pmatrix}$$

La matrice di Hilbert è una matrice che presenta lungo le antidiagonali elementi costanti pari ai reciproci dei numeri interi.

In termini di elementi la matrice di Hilbert è definita da

$$h_{ij} = \frac{1}{i+j-1} \qquad i,j = 1,2,\ldots,n$$

La matrice di Hilbert è fortemente malcondizionata, cioè amplifica molto gli errori al crescere della dimensione.

- **Matrice di Pascal**

  La matrice di Pascal è una matrice che "contiene" al suo interno un triangolo di Tartaglia

Figura 7.2 - Blaise Pascal (1623-1662)

$$P = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \ldots \\ 1 & 2 & 3 & 4 & & \\ 1 & 3 & 6 & & & \\ 1 & 4 & & & & \\ 1 & & & & & \\ \vdots & & & & & \end{pmatrix},$$

In termini di elementi

$$p_{ij} = \begin{cases} 1 & i = 1 \qquad j = 1,2,\ldots,n \\ 1 & j = 1 \qquad i = 1,2,\ldots,n \\ p_{i,j-1} + p_{i-1,j} & i,j \neq 1 \end{cases}$$

La matrice di Pascal è una matrice di interi e non presenta particolari problemi di malcondizionamento, anche se, poiché i suoi elementi crescono rapidamente, per grandi dimensioni si può andare incontro a problemi di overflow.

- **Matrici di Toeplitz** Una classe particolare di matrici è data dalle matrici di **Toeplitz**.

  Tali matrici sono caratterizzate dal presentare elementi costanti lungo ciascuna diagonale.

Figura 7.3 - Otto Toeplitz (1881-1940)

$$T = \begin{pmatrix} a_0 & a_1 & a_2 & \ldots & \ldots & a_n \\ b_1 & a_0 & a_1 & & & a_{n-1} \\ b_2 & b_1 & a_0 & & & a_{n-2} \\ \vdots & \ddots & & \ddots & & \vdots \\ \vdots & & & & \ddots & a_1 \\ b_n & \ldots & & b_2 & b_1 & a_0 \end{pmatrix}$$

Una matrice di Toeplitz è completamente definita dai due verrori che fissano, rispettivamente, la prima riga e la prima colonna.

Se la matrice di Toeplitz è simmetrica, essa è completamente definita da un solo vettore, prima riga o prima colonna indifferentemente

$$T = \begin{pmatrix} a_0 & a_1 & a_2 & \ldots & \ldots & a_n \\ a_1 & a_0 & a_1 & & & a_{n-1} \\ a_2 & a_1 & a_0 & & & a_{n-2} \\ \vdots & \ddots & & \ddots & & \vdots \\ \vdots & & & & \ddots & a_1 \\ a_n & \ldots & & a_2 & a_1 & a_0 \end{pmatrix}$$

7.3 Matrici di forma speciale

Alcune matrici sono classificate in base alla particolare struttura degli elementi. Si tratta, in molti casi di matrici quadrate **sparse**, cioè di matrici che presentano un gran numero di elementi pari a 0.
Nel seguito si elencano alcune tra le più comuni matrici cosiddette di forma speciale.

- Matrice **diagonale**: matrice in cui tutti gli elementi fuori diagonale sono nulli

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}, \qquad a_{ij} = 0, \quad i \neq j.$$

  Una particolare matrice diagonale è la matrice **identità** in cui tutti gli elementi sulla diagonale sono uguali a 1

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \qquad a_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j; \end{cases}$$

- matrice **triangolare superiore**: matrice in cui tutti gli elementi sotto diagonale sono nulli

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}, \qquad a_{ij} = 0, \quad i > j;$$

- matrice **triangolare inferiore**: matrice quadrata in cui tutti gli elementi sotto diagonale sono nulli

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \qquad a_{ij} = 0, \quad i < j;$$

- matrice **tridiagonale**: una matrice A si dice **a banda** se esistono due numeri interi positivi p e q tali che tutti gli elementi non nulli appartengono alla diagonale, alle p sovradiagonali e alle q sottodiagonali: $a_{ij} = 0$ se $j - i > p$ oppure $i - j > q$. $q + p + 1$ detto ampiezza della banda.

  Una matrice **tridiagonale** è una particolare matrice a banda in cui gli elementi diversi da 0 si trovano esclusivamente sulla diagonale principale, sulla sottodiagonale e sulla sopradiagonale (ampiezza della banda 3).

$$a_{ij} = 0 \qquad \text{se} \qquad \big|j - i\big| > 1.$$

  Una matrice tridiagonale è rappresentata mediante i tre vettori che rappresentano le tre diagonali non nulle

$$\begin{pmatrix} a_{1,1} & a_{1,2} & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & 0 & a_{n,n-1} & a_{n,n} \end{pmatrix} \longrightarrow \begin{pmatrix} a_1 & c_1 & 0 & \cdots & 0 \\ b_2 & a_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix}.$$

7.4 Operazioni con matrici

Definizione 7.2. *Nell'insieme delle matrici possono essere definite in maniera opportune le classiche operazioni algebriche*

- *date due matrici* $A,\ B \in \mathbb{R}^{n\times m}$ *si definisce* ***matrice somma*** *la matrice* $C \in \mathbb{R}^{n\times m}$

$$C = A \pm B \qquad\qquad c_{ij} = a_{ij} \pm b_{ij} \qquad \begin{matrix} i & = 1,\dots,n \\ j & = 1,\dots,m \end{matrix}$$

- *il prodotto di uno scalare* $\tau \in \mathbb{R}$ *per una matrice* $A \in \mathbb{R}^{n\times m}$ *è la matrice*

$$\tau A = (\tau a_{ij}) \qquad \begin{cases} i & = 1,\dots,n \\ j & = 1,\dots,m \end{cases}$$

- *date due matrici* $A \in \mathbb{R}^{n\times m}$ *e* $B \in \mathbb{R}^{p\times q}$*, l'operazione prodotto è definita solo se* $m = p$ *e si definisce* ***matrice prodotto*** *la matrice* $C \in \mathbb{R}^{n\times q}$

$$C = AB \qquad\qquad c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}, \qquad \begin{matrix} i & = 1,\dots,n \\ j & = 1,\dots,q. \end{matrix}$$

Il prodotto tra matrici non è commutativo, cioè in generale

$$AB \neq BA$$

come si può facilmente verificare con un esempio

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \begin{matrix} AB = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ BA = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \end{matrix} \qquad AB \neq BA$$

Il prodotto righe per colonne soddisfa alle leggi distributive rispetto alla somma di matrici, cioè

$$A(B+C) = AB + AC, \qquad\qquad (B+C)A = BA + CA.$$

Proposizione 7.1 *Siano* $A \in \mathbb{R}^{n\times m}$*,* $c \in \mathbb{R}^m$*,* $a,\ b \in \mathbb{R}^n$

- *il prodotto di una matrice per un vettore colonna è un vettore colonna* $\left(\mathbb{R}^{n\times m} \times \mathbb{R}^{m\times 1} \to \mathbb{R}^{n\times 1}\right)$

$$\begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ \vdots \\ c_m \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m} a_{1j}c_j \\ \sum_{j=1}^{m} a_{2j}c_j \\ \vdots \\ \sum_{j=1}^{m} a_{nj}c_j \end{pmatrix};$$

- *il prodotto di un vettore riga per una matrice è un vettore riga* $\left(\mathbb{R}^{1\times n} \times \mathbb{R}^{n\times m} \to \mathbb{R}^{1\times m}\right)$

$$\begin{pmatrix} c_1 & c_2 & \cdots & c_n \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} c_j a_{j1} & \sum_{j=1}^{n} c_j a_{j2} & \cdots & \sum_{j=1}^{n} c_j a_{jm} \end{pmatrix};$$

- *il prodotto di un vettore riga per un vettore colonna è uno scalare detto* **prodotto scalare** $\left(\mathbb{R}^{1\times n}\times \mathbb{R}^{n\times 1}\to \mathbb{R}^{1\times 1}\equiv \mathbb{R}\right)$

$$\begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \sum_{i=1}^{n} a_i b_i;$$

- *il prodotto di un vettore colonna per un vettore riga è una matrice quadrata della stessa dimensione dei vettori* $\left(\mathbb{R}^{n\times 1}\times \mathbb{R}^{1\times n}\to \mathbb{R}^{n\times m}\right)$

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix} = \begin{pmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & \dots & \dots & a_2b_n \\ \vdots & & & \\ a_nb_1 & \dots & \dots & a_nb_n \end{pmatrix}.$$

Proposizione 7.2 *Il prodotto di due matrici triangolari superiori, inferiori o diagonali è, rispettivamente, una matrice triangolare superiore, inferiore o diagonale.*

7.4.1 Determinante

Ad ogni matrice quadrata A è associato il numero $\left|A\right|$, detto **determinante** di A.

Definizione 7.3. *Data la matrice $A \in \mathbb{R}^{n\times n}$, si definisce* ***complemento algebrico*** *dell'elemento a_{ij}, $i,j = 1,\dots,n$, e si indica con A_{ij}, il determinante della matrice di ordine $n-1$ ottenuta dalla matrice A eliminando la riga i e la colonna j, moltiplicato per $(-1)^{i+j}$*

$$A_{ij} = (-1)^{i+j}\left|\overline{A}_{ij}\right|.$$

Data una matrice $A \in \mathbb{R}^{n\times n}$, di elementi $\left(a_{ij}\right)_{i,j=1,\dots,n}$ per il determinante di valgono le seguenti proprietà:

1. se A ha due righe (o colonne) uguali allora il suo determinante è zero;
2. se A, B, $C \in \mathbb{R}^{n\times n}$ e $C = AB$ allora
$$\left|C\right| = \left|A\right|\left|B\right|;$$
3. se A, $B, \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}$ e $B = cA$ allora
$$\left|B\right| = c^n\left|A\right|;$$
4. se B è la matrice che si ottiene da A cambiando di posto due righe o colonne allora
$$\left|B\right| = -\left|A\right|;$$
5. se $c \in \mathbb{R}$ e B è la matrice che si ottiene da A sostituendo in essa la riga i con la somma della riga i più c per la riga k $(k \neq i)$ allora risulta
$$\left|B\right| = \left|A\right|.$$

7.4.2 Matrice inversa

Definizione 7.4. *Una matrice è detta* **singolare** *se il suo determinante è uguale a zero,* **non singolare** *altrimenti.*

Proposizione 7.3 *Data una matrice $A \in \mathbb{R}^{n\times n}$, non singolare, esiste ed è unica la matrice A^{-1}, detta* **inversa** *di A, tale che*

$$AA^{-1} = A^{-1}A = I.$$

Teorema 7.1 *Se A e B sono due matrici tali che il prodotto AB è definito, vale*

$$(AB)^T = B^T A^T$$

$$(A+B)^T = A^T + B^T.$$

Se inoltre A è invertibile

$$\left(A^{-1}\right)^T = \left(A^T\right)^{-1}.$$

Definizione 7.5. *Una matrice quadrata si dice* **simmetrica** *se*

$$A = A^T.$$

Proposizione 7.4 *Se A è una matrice simmetrica e invertibile, anche la sua inversa è simmetrica*

$$\left(A^{-1}\right)^T = \left(A^T\right)^{-1} = A^{-1}.$$

7.5 Matrici di forma speciale: inversa e determinante

Per matrici di forma particolare il calcolo del determinante è notevolmente semplificato.

- **Matrici diagonali:** data $D \in \mathbb{R}^{n\times n}$, D diagonale vale

$$|D| = \prod_{i=1}^{n} d_{ii}, \tag{7.1}$$

cioè il determinante è il prodotto degli elementi diagonali.

In particolare, per la matrice identità si ha

$$|I| = 1.$$

Sia $D \in \mathbb{R}^{n\times n}$, diagonale e non singolare, la sua inversa $B = D^{-1}$ è diagonale e i suoi elementi sono dati da

$$b_{ij} = \begin{cases} \dfrac{1}{d_{ii}} & i = j \\ 0 & i \neq j \end{cases} \qquad i,j = 1,\ldots,n$$

- **Matrici triangolari:** data $T \in \mathbb{R}^{n\times n}$, T triangolare (inferiore o superiore) vale

$$|T| = \prod_{i=1}^{n} t_{ii}, \tag{7.2}$$

cioè il determinante è il prodotto degli elementi diagonali.

- ▶ Data $U \in \mathbb{R}^{n\times n}$, matrice triangolare superiore, la sua inversa $R = U^{-1}$ è triangolare superiore e i suoi elementi sono dati da

$$r_{ij} = \begin{cases} \dfrac{1}{u_{ii}} & i = j \qquad i = 1,\dots,n \\ -\dfrac{1}{u_{ii}} \displaystyle\sum_{k=i+1}^{j} u_{ik} r_{kj} & i < j \qquad i = n-1, n-2, \dots, 1 \\ 0 & i > j \qquad i = 2,\dots,n; \end{cases}$$

- ▶ data $L \in \mathbb{R}^{n\times n}$, matrice triangolare inferiore, la sua inversa $S = L^{-1}$ è triangolare inferiore e i suoi elementi sono dati da

$$s_{ij} = \begin{cases} \dfrac{1}{l_{ii}} & i = j \qquad i = 1,\dots,n \\ -\dfrac{1}{l_{ii}} \displaystyle\sum_{k=j}^{i-1} l_{ik} s_{kj} & i < j \qquad i = 2,\dots,n \\ 0 & i < j \qquad i = 1,\dots,n-1 \end{cases}$$

- **Matrici tridiagonali:** data $T \in \mathbb{R}^{n\times n}$, T tridiagonale rappresentata come

$$\begin{pmatrix} a_1 & c_1 & 0 & \cdots & 0 \\ b_2 & a_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix},$$

il suo determinante può essere calcolato per ricorrenza secondo la seguente formula: detta T_k la sottomatrice principale di ordine k vale

$$\begin{cases} |T_0| = 1, \qquad |T_1| = a_1 \\ |T_k| = a_k|T_{k-1}| - b_k c_{k-1}|T_{k-2}| \qquad\qquad k = 2,3,\dots,n \end{cases}$$

Per calcolare l'inversa di una matrice tridiagonale A si osserva che questa può essere fattorizzata nel prodotto di due matrici bidiagonali, una inferiore L con 1 sulla diagonale e una superiore U

$$A = LU$$

$$\begin{pmatrix} a_1 & c_1 & 0 & \cdots & 0 \\ b_2 & a_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & 0 & b_n & a_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \beta_2 & 1 & 0 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \beta_{n-1} & 1 & 0 \\ 0 & \dots & 0 & \beta_n & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 & c_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & \alpha_n \end{pmatrix}$$

Gli elementi di L e di U si determinano attraverso l'algoritmo

$$\begin{cases} \alpha_1 = a_1 \\ \\ \beta_k = \dfrac{b_k}{\alpha_{k-1}} & k = 2, \ldots, n. \\ \\ \alpha_k = a_k - \beta_k c_{k-1} \end{cases}$$

L'algoritmo per il calcolo del determinante e della matrice inversa si può ora costruire osservando che

$$\big|\, A \,\big| = \big|\, L \,\big|\, \big|\, U \,\big|, \qquad A^{-1} = U^{-1} L^{-1}$$

e sono note le formule per il calcolo del determinante e della matrice inversa di matrici triangolari.

7.6 Norme

7.6.1 Norme vettoriali

Una **norma** definita in $\mathbb{R}^n$ o in $\mathbb{R}^{n\times n}$ quantifica la "grandezza" dei vettori e delle matrici.

Definizione 7.6. *Una* **norma vettoriale** *è un'applicazione* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ *tale che*

1) $\|\, x \,\| > 0 \qquad \forall\ x \in \mathbb{R}^n,\ x \neq \underline{0}$

2) $\|\, \alpha x \,\| = |\alpha| \|\, x \,\| \qquad \forall\ \alpha \in \mathbb{R},\ \forall\ x \in \mathbb{R}^n$

3) $\|\, x + y \,\| \leq \|\, x \,\| + \|\, y \,\| \qquad \forall\ x, y \in \mathbb{R}^n$

Proposizione 7.5 *Per le norme vettoriali vale*

$$\|\, \underline{0} \,\| = 0, \qquad \|\, x \,\| = \|\, -x \,\| \qquad\qquad \underline{0},\ x \in \mathbb{R}^n$$

Proposizione 7.6

$$\Big|\, \|\, x \,\| - \|\, y \,\| \,\Big| \leq \|\, x - y \,\| \qquad \forall\, x, y \in \mathbb{R}^n.$$

Alcune tra le norme vettoriali più usate sono

a) $\|\, x \,\|_1 = \displaystyle\sum_{i=1}^{n} |x_i|$ (norma 1 o di Manhattan)

b) $\|\, x \,\|_2 = \left(\displaystyle\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$ (norma 2 o euclidea)

c) $\|\, x \,\|_p = \left(\displaystyle\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \qquad p \geq 1$ (norma p)

d) $\|\, x \,\|_\infty = \displaystyle\max_{1\leq i\leq n} |x_i|$ (norma ∞ o di Chebychev)

7.6.2 Norme matriciali

L'insieme $\mathbb{R}^{n\times n}$ delle matrici quadrate di ordine n, ad elementi reali, è isomorfo a $\mathbb{R}^{n^2}$ per cui una norma in $\mathbb{R}^{n\times n}$ è definita in maniera analoga a una norma in $\mathbb{R}^{n^2}$, cioè a una norma vettoriale.

Definizione 7.7. *Una* **norma matriciale** *è un'applicazione* $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ *tale che*

1) $\|A\| > 0 \qquad \forall\ A \in \mathbb{R}^{n\times n},\ A \neq O$

2) $\|\alpha A\| = |\alpha|\|A\| \qquad \forall\ \alpha \in \mathbb{R},\ \forall\ A \in \mathbb{R}^{n\times n}$

3) $\|A + B\| \leq \|A\| + \|B\| \qquad \forall\ A, B \in \mathbb{R}^{n\times n}$

Nella pratica risultano particolarmente usate norme che godono dell'ulteriore proprietà, detta di ***consistenza***

4) $\|A\,B\| \leq \|A\|\,\|B\| \qquad \forall\ A, B \in \mathbb{R}^{n\times n}$

Proposizione 7.7 *Per le norme matriciali vale*

$$\|\underline{O}\| = 0$$

$$\|-A\| = \|A\|$$

$$\Big|\|A\| - \|B\|\Big| \leq \|A - B\| \qquad \forall A, B \in \mathbb{R}^{n\times n}.$$

$$\Big|\|A\| - \|B\|\Big| \leq \|A + B\|$$

Alcune tra le norme matriciali più usate sono

a) $\|A\|_1 = \max\limits_{1\leq k\leq n} \sum\limits_{i=1}^{n} |a_{ik}| \qquad$ (norma 1)

b) $\|A\|_2 = \sqrt{\rho\,(A^T A)} \qquad$ (norma 2)

c) $\|A\|_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{k=1}^{n} |a_{ik}| \qquad$ (norma ∞)

Osservazione 3. *Per ogni norma matriciale,* $\|\cdot\|$*, vale*

$$\|I\| = \|I \cdot I\| \leq \|I\|\|I\| = \big(\|I\|\big)^2 \qquad \Longrightarrow \qquad 1 \leq \|I\|.$$

In particolare per le norme a), b) e c) vale

$$\|I\| = 1.$$

7.6.3 Matrici convergenti

Per studiare la convergenza di alcune procedure iterative occorre studiare il comportamento delle successive potenze di una matrice A; in particolare è importante studiare sotto quali ipotesi si verifica

$$\lim_{m\to\infty} A^m = O.$$

Definizione 7.8. *Sia* $\{A_k\}$ *una successione di matrici in* $\mathbb{R}^{m\times n}$ *con* $A_k \equiv \left(a_{ij}^{(k)}\right)$, $k = 1, 2, \ldots$. *Si dice che la successione converge alla matrice* $A \in \mathbb{R}^{m\times n}$,

$$\lim_{k\to\infty} A_k = A,$$

se e solo se

$$\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij} \qquad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, n. \end{matrix}$$

La nozione di convergenza di matrici può essere caratterizzata in termini di norme

$$\lim_{k\to\infty} A_k = A \quad \Leftrightarrow \quad \lim_{k\to\infty} \| A_k - A \| = 0$$

essendo $\| \cdot \|$ una opportuna norma.

Definizione 7.9. *Una matrice* $A \in \mathbb{R}^{n\times n}$ *per cui vale*

$$\lim_{m\to\infty} A^m = \lim_{m\to\infty} \underbrace{A \cdot A \cdots A}_{m \text{ volte}} = O$$

è detta **matrice convergente**.

Teorema 7.2 *I seguenti enunciati sono equivalenti*

a) A *è convergente;*

b) $\lim_{m\to\infty} \| A^m \| = 0$ *per qualche norma;*

c) $\rho(A) < 1$.

Corollario 7.1 *Una matrice* A *è convergente se per qualche norma risulta*

$$\| A \| < 1.$$

7.7 Operazioni elementari su matrici

Nell'insieme delle matrici sono definite alcune particolari operazioni dette **operazioni elementari**

1) moltiplicare una riga (colonna) per uno scalare non nullo;

2) scambiare fra loro due righe (colonne);

3) sostituire la riga i, con la somma della riga i e della riga k moltiplicata per uno scalare $c \neq 0$.

Definizione 7.10. *Le matrici che si ottengono dalla matrice identità effettuando su di essa una operazione elementare sono dette* **matrici elementari** *e sono indicate con*

$F_k(c)$: *matrice identità con la riga* k *moltiplicata per* c;

F_{jk}: *matrice identità con le righe* j *e* k *scambiate; (matrice* **di permutazione***);*

$F_{jk}(c)$: *matrice identità con la riga* j *sostituita dalla somma della riga* k *moltiplicata per* c *e della riga* j.

Teorema 7.3 *Sia* A *una matrice di ordine* n; *eseguire su* A *una operazione elementare equivale a premoltiplicarla per la corrispondente matrice elementare.*

Esempio 7.1

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 5 \\ 1 & 2 & 1/2 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 5 \\ 9 & 2 & 21/2 \end{pmatrix}$$

La matrice B si ottiene dalla A sostituendo in questa la terza riga con la somma della terza e della seconda moltiplicata per 2.
Se si esegue il prodotto per la corrispondente matrice elementare

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 5 \\ 1 & 2 & 1/2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 5 \\ 9 & 2 & 21/2 \end{pmatrix}$$

Analogamente si possono verificare le altre operazioni elementari.

Se si indica con A' la matrice che si ottiene dopo la premoltiplicazione e con R_i e R'_i le righe $i-$esime di A e A', rispettivamente, le formule che legano gli elementi di A e A' sono riportate nel quadro seguente

Premoltiplicazione	**Relazione sulle righe**		**Relazione sugli elementi**	
$A' = F_i(c)A$	$R'_i = cR_i$	$k \neq i$	$a'_{ij} = ca_{ij}$	$j = 1, \ldots, n$
	$R'_k = R_k$		$a'_{kj} = a_{kj}$	$j = 1, \ldots, n$
$A' = F_{ik}A$	$R'_i = R_k$		$a'_{ij} = a_{kj}$	$j = 1, \ldots, n$
	$R'_k = R_i$		$a'_{kj} = a_{ij}$	$j = 1, \ldots, n$
	$R'_m = R_m$	$m \neq k, i$	$a'_{mj} = a_{mj}$	$j = 1, \ldots, n$
$A' = F_{ik}(c)A$	$R'_i = R_i + cR_k$		$a'_{ij} = a_{ij} + ca_{kj}$	$j = 1, \ldots, n$
	$R'_m = R_m$	$m \neq i$	$a'_{mj} = a_{mj}$	$j = 1, \ldots, n$

Osservazione 4. *Si può facilmente verificare che il prodotto di matrici elementari è ancora una matrice di elementare.*

Osservazione 5. *Analoghe operazioni elementari possono essere definite sulle colonne di una matrice e in questo caso si può dimostrare che eseguire su una matrice A una operazione elementare colonna equivale a postmoltiplicarla per la corrispondente matrice elementare.*

Proposizione 7.8 *Dalla* (7.1)*, dalla* (7.2) *e dalla proprietà 4. dei determinanti (paragrafo 7.4.1) si ottiene facilmente il determinante delle matrici elementari*

$$\big|F_i(c)\big| = c, \qquad \big|F_{ik}\big| = -1, \qquad \big|F_{ik}(c)\big| = 1.$$

7.8 Soluzione di sistemi lineari

In quasi tutte le discipline scientifiche si incontrano problemi che richiedono la risoluzione di sistemi lineari. Ad esempio in Matematica problemi di interpolazione, minimi quadrati, equazioni alle derivate parziali, etc. o in discipline più direttamente rivolte alle applicazioni quali analisi di circuiti elettrici, ingegneria strutturale, etc.

7.8.1 Definizioni e proprietà fondamentali

Un sistema lineare di n equazioni nelle n incognite $x_1, x_2, \ldots, x_n$ si presenta nella forma

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \quad . \end{cases} \tag{7.3}$$

I numeri a_{ij} $i, j = 1, \ldots, n$ sono detti **coefficienti**, e i numeri $b_1, \ldots, b_n$ **termini noti** del sistema.

Definizione 7.11. *Ogni $n-$pla ordinata di numeri $(z_1, \ldots, z_n)$ che soddisfa le (7.3) è detta* **soluzione del sistema**.

Definizione 7.12. *Due sistemi lineari sono* **equivalenti** *se e solo se tutte e sole le soluzioni dell'uno sono anche soluzioni dell'altro.*

Definizione 7.13. *Un sistema lineare è detto* **omogeneo** *se i termini noti sono tutti nulli.*

Definizione 7.14. *Un sistema lineare è detto* **determinato**, **indeterminato** *o* **impossibile** *se ammette, rispettivamente, una, infinite o nessuna soluzione. Un sistema indeterminato o impossibile è detto anche* **singolare**, *mentre un sistema determinato è detto* **non singolare**.

Un sistema lineare è non singolare, cioè ammette una e una sola soluzione, se la matrice del sistema è non singolare ma ciò è vero solo in aritmetica infinita; operando in virgola mobile con m cifre di mantissa, un sistema lineare può essere "quasi" singolare. A titolo di esempio si consideri il sistema

$$\begin{cases} 5x + 7y &= 12 \\ 7x + 10y &= 17 \end{cases} \qquad \Longrightarrow \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{7.4}$$

Sostituendo nel sistema la coppia $x = 2.415$, $y = 0$ si trova

$$\begin{aligned} 5x + 7y &= 5 \times 2.415 + 7 \times 0 = 12.075 \\ 7x + 10y &= 7 \times 2.415 + 10 \times 0 = 16.905 \end{aligned}$$

per cui, operando con due cifre di mantissa, anche la coppia $x = 2.415$, $y = 0$ soddisfa il sistema (7.4). Questo paradosso si spiega osservando che le due rette rappresentate dalle equazioni (7.4) sono "quasi" parallele. Il punto $(2.415, 0)$ pur non appartenendo a nessuna delle due rette è molto "vicino" ad entrambe.

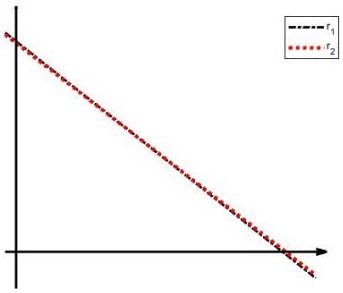


Figura 7.4 - Sistema "quasi" indeterminato

Un sistema lineare è di solito trappresentato in forma compatta: se A è la matrice di elementi (a_{ij}), $i, j = 1, \ldots, n$ e x e b i vettori di componenti rispettive $(x_1, \ldots, x_n)$, $(b_1, \ldots, b_n)$, il sistema può essere scritto nella forma

$$Ax = b. \tag{7.5}$$

La scrittura (7.5), è detta **forma matriciale del sistema** (7.3), la matrice A **matrice dei coefficienti**, e il vettore b **vettore dei termini noti**.

7.8.2 Sistemi non singolari. Regola di Cramer

Teorema 7.4 (Regola di Cramer) *Un sistema lineare di n equazioni in n incognite*

$$Ax = b \tag{7.6}$$

è non singolare (cioè ammette una ed una sola soluzione) se e solo se il determinante della matrice dei coefficienti è diverso da zero

$$\left| A \right| \neq 0.$$

Indicata con A_i $i = 1, \ldots, n$ la matrice che si ottiene da A inserendo la colonna dei termini noti al posto della $i-$esima colonna, la soluzione del sistema è data da

$$x_i = \frac{|A_i|}{|A|} \qquad i = 1, \ldots, n. \tag{7.7}$$

La regola di Cramer non solo fornisce una condizione necessaria e sufficiente per l'esistenza e l'unicità della soluzione del sistema, ma indica anche il modo per calcolarla.
Il problema del calcolo della soluzione dei sistemi lineari sembrerebbe, così completamente risolto. In effetti ciò è vero solo da un punto di vista teorico e non computazionale, infatti per calcolare la soluzione di un sistema lineare secondo la (7.7) occorre calcolare $n+1$ determinanti di ordine n. Si può provare che per il calcolo di un determinante di ordine n, occorrono $n!(n-1)$ moltiplicazioni; quindi per la soluzione di un sistema lineare occorrono

$$(n+1)(n-1)n! = (n^2-1)n! \quad \text{moltiplicazioni.} \tag{7.8}$$

Ad esempio, se $n = 20$, per la (7.8) occorrono

$$399 \cdot 20! \approx 10^{21} \quad \text{moltiplicazioni.}$$

È evidente che la regola di Cramer è di fatto inutilizzabile per sistemi di ordine elevato.

7.8.3 Calcolo effettivo della soluzione di un sistema lineare

Se il sistema (7.6) è non singolare, la matrice dei coefficienti, A, è non singolare, cioè $|A| \neq 0$ e, in tali ipotesi, esiste la matrice inversa, A^{-1}. Moltiplicando a sinistra la (7.6) per A^{-1} si ottiene

$$A^{-1}Ax = A^{-1}b$$

da cui

$$x = A^{-1}b.$$

Tuttavia il calcolo della matrice inversa non è agevole, per cui è preferibile usare metodi che forniscono direttamente la soluzione del sistema. Tali metodi si dividono in due categorie

a) **metodi diretti:** in assenza di errori di arrotondamento, forniscono la soluzione esatta del sistema in un numero finito e predeterminato di passi computazionali. Tra questi i più noti sono i metodi di **Gauss** e di **Gauss-Jordan**;

b) **metodi iterativi:** a partire da una approssimazione iniziale, costruiscono una sequenza di successive approssimazioni, che converge, al limite, alla soluzione del sistema. In questa categoria si considerano i metodi di **Jacobi**, di **Gauss-Seidel**, e i metodi di rilassamento **JOR** e **SOR**.

7.9 Sistemi lineari di forma speciale

Quando un sistema lineare presenta una matrice dei coefficienti di forma particolare è possibile determinare in modo semplice la soluzione, sfruttando la forma stessa della matrice

- **Sistema diagonale**: $a_{ij} = 0$ se $i \neq j$

$$Ax = b \quad \Longleftrightarrow \quad \begin{cases} a_{11}x_1 & & & & = b_1 \\ & a_{22}x_2 & & & = b_2 \\ & & \ddots & & \vdots \\ & & & a_{nn}x_n & = b_n \end{cases}$$

  e la soluzione del sistema è data da

$$x_i = \frac{b_i}{a_{ii}} \qquad i = 1, \ldots, n.$$

- **Sistema triangolare superiore**: $a_{ij} = 0$ se $i > j$

$$Ax = b \quad \Longleftrightarrow \quad \begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & & + & a_{2n}x_n & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & a_{nn}x_n & = & b_n \end{cases} .$$

  La soluzione si ottiene facilmente con una sostituzione all'indietro

$$\begin{cases} x_n & = b_n/a_{nn} \\ x_k & = \dfrac{1}{a_{kk}} \left[b_k - \displaystyle\sum_{j=k+1}^{n} a_{kj}x_j \right] \quad k = n-1, n-2, \ldots, 1 \end{cases} \tag{7.9}$$

- **Sistema triangolare inferiore**: $a_{ij} = 0$ se $i < j$

$$Ax = b \quad \Longleftrightarrow \quad \begin{cases} a_{11}x_1 & & & & & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & & & = & b_2 \\ \vdots & & & \ddots & & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & \ldots & a_{nn}x_n & = & b_n \end{cases} .$$

  La soluzione si ottiene facilmente con una sostituzione in avanti

$$\begin{cases} x_1 & = b_1/a_{11} \\ x_k & = \dfrac{1}{a_{kk}} \left[b_k - \displaystyle\sum_{j=1}^{k-1} a_{kj}x_j \right] \quad k = 2, 3, \ldots, n \end{cases}$$

- **Sistema tridiagonale**: $a_{ij} = 0$ se $\left| i - j \right| \geq 2$. Con un opportuno cambio di notazione

$$Ax = b \quad \Longleftrightarrow \quad \begin{cases} a_1x_1 & + & c_1x_2 & & & & & & & = & d_1 \\ b_2x_1 & + & a_2x_2 & + & c_2x_3 & & & & & = & d_2 \\ & & b_3x_2 & + & a_3x_3 & + & c_3x_4 & & & = & d_3 \\ & & \ddots & & \ddots & & \ddots & & & & \vdots \\ & & & & b_{n-1}x_{n-2} & + & a_{n-1}x_{n-1} & + & c_{n-1}x_n & = & d_{n-1} \\ & & & & & & b_nx_{n-1} & + & a_nx_n & = & d_n. \end{cases} \tag{7.10}$$

La soluzione del sistema, posto $b_1 = c_n = 0$, si ottiene con le seguenti formule ricorsive

$$\begin{cases} f_0 = g_0 = 0 \\ f_r = \dfrac{-c_r}{b_r f_{r-1} + a_r} \\ g_r = \dfrac{d_r - b_r g_{r-1}}{b_r f_{r-1} + a_r} \end{cases} \quad r = 1, \ldots, n \qquad \begin{cases} x_n = g_n \\ x_r = f_r x_{r+1} + g_r, \quad r = n-1, \ldots, 1 \end{cases} \tag{7.11}$$

Dalla prima di (7.10) si ha

$$x_1 = -\frac{c_1}{a_1} x_2 + \frac{d_1}{a_1}$$

quindi vale una relazione ricorsiva del tipo

$$x_r = f_r x_{r+1} + g_r \qquad r = 1, \ldots, n-1. \tag{7.12}$$

Per determinare i coefficienti f_r, g_r $r = 1, \ldots, n$ si considera la $r-$esima equazione di (7.10)

$$b_r x_{r-1} + a_r x_r + c_r x_{r+1} = d_r \tag{7.13}$$

e si applica l'ipotesi induttiva a x_{r-1}

$$x_{r-1} = f_{r-1} x_r + g_{r-1}. \tag{7.14}$$

Sostituendo la (7.14) nella (7.13) si ottiene

$$x_r = \frac{-c_r}{b_r f_{r-1} + a_r} x_{r+1} + \frac{d_r - b_r g_{r-1}}{b_r f_{r-1} + a_r} \qquad r = n-1, n-2, \ldots, 1$$

da cui, confrontando con la (7.12), si ottengono facilmente le seconde di (7.11).

Figura 7.5 - Llewellyn Hilleth Thomas (1903-1992)

L'algoritmo appena descritto è noto come algoritmo di Thomas e si basa sulle formule ricavate per la fattorizzazione di una matrice tridiagonale nel paragrafo 7.5.

7.10 Soluzione numerica di sistemi lineari: metodi diretti

7.10.1 Il metodo di Gauss

Il metodo di Gauss consiste in un processo di trasformazione, mediante opportune operazioni elementari, del sistema originario in un sistema triangolare equivalente.
Dato il sistema lineare non singolare

$$Ax = b \tag{7.15}$$

se si determina una sequenza di matrici elementari $F_m, F_{m-1}, \ldots, F_1$ tali che

$$F_m F_{m-1} \ldots F_1 A = U \tag{7.16}$$

con U triangolare superiore, sostituendo nella (7.15) si ha

$$F_m F_{m-1} \dots F_1 A x = F_m F_{m-1} \dots F_1 b$$

cioè

$$Ux = c \qquad \text{con} \qquad \begin{cases} U &= F_m \dots F_1 A \qquad \text{triangolare superiore} \\ c &= F_m \dots F_1 b. \end{cases} \tag{7.17}$$

I sistemi (7.17) e (7.15) sono ovviamente equivalenti quindi ammettono le stesse soluzioni, ma il sistema (7.17) è di forma triangolare quindi di immediata soluzione.

Nella pratica, la sequenza di matrici $F_m, \dots, F_1$ non è esplicitamente determinata. Il metodo di Gauss si articola in $n-1$ stadi successivi e consiste nel lavorare ad ogni stadio sulle righe di A e sulla corrispondente componente del vettore b con opportune operazioni elementari che azzerano gli elementi fuori diagonale della corrispondente colonna.

Data $A \in \mathbb{R}^{n\times n}$ non singolare, per comodità di notazione, si inserisce la colonna dei termini noti come $(n+1)-$esima colonna della matrice dei coefficienti

$$a_{i,n+1} = b_i \qquad i = 1, \dots, n$$

costruendo quindi la matrice ampliata che, per semplicità chiamiamo ancora A

$$A = \left(\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & a_{1,n+1} \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & a_{n,n+1} \end{array}\right) = \left(\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & b_n \end{array}\right). \tag{7.18}$$

Il metodo di Gauss modifica ad ogni stadio la matrice costruendone una nuova che chiameremo $A^{(j)}$. All'inizio

$$A^{(1)} = \left(\begin{array}{ccc|c} a_{11}^{(1)} & \dots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & \dots & a_{nn}^{(1)} & a_{n,n+1}^{(1)} \end{array}\right) = \left(\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & a_{1,n+1} \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & a_{n,n+1} \end{array}\right) = A.$$

Ad ogni stadio j vengono modificate solo le righe al di sotto della riga j.

Si suppone di aver già operato sulle prime $k-1$ righe e di aver ottenuto quindi la matrice

$$A^{(k-1)} = \begin{pmatrix} a_{1,1}^{(1)} & \dots & \dots & a_{1,k-1}^{(1)} & a_{1,k}^{(1)} & \dots & a_{1,n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{2,2}^{(2)} & \ddots & a_{2,k-1}^{(2)} & a_{2,k}^{(2)} & \dots & a_{2,n}^{(2)} & a_{2,n+1}^{(2)} \\ . & . & \ddots & & . & . & . & \\ . & . & \ddots & a_{k-1,k-1}^{(k-1)} & . & . & . & \\ . & & & 0 & a_{k,k}^{(k-1)} & & & \\ . & . & \dots & \dots & \vdots & . & . & \\ 0 & \dots & \dots & 0 & a_{n,k}^{(k-1)} & \dots & a_{n,n}^{(k-1)} & a_{n,n+1}^{(k-1)} \end{pmatrix}$$

Si procede quindi sulla $k-$esima riga operando come segue

a) se $a_{kk}^{(k-1)} = 0$, si scambia di posto la $k-$esima riga con la riga $j-$esima, $j > k$, che presenta $a_{jk}^{(k-1)} \neq 0$.

   In termini di matrici elementari si premoltiplica la matrice $A^{(k-1)}$ per la matrice elementare F_{jk};

b) si somma alla $i-$esima riga la $k-$esima moltiplicata per $-\dfrac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$, $\quad i = k+1, \ldots, n.$

In termini di matrici elementari si premoltiplica per le matrici $F_{ik}\left(-\dfrac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right)$, $\quad i = k+1, \ldots, n.$

Si ottiene quindi la matrice

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & \ldots & \ldots & a_{1,k-1}^{(1)} & a_{1,k}^{(1)} & \ldots & a_{1,n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{2,2}^{(2)} & \ddots & a_{2,k-1}^{(2)} & a_{2,k}^{(2)} & \ldots & a_{2,n}^{(2)} & a_{2,n+1}^{(2)} \\ \cdot & \cdot & \ddots & & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \ddots & a_{k-1,k-1}^{(k-1)} & \cdot & \cdot & \cdot & \\ \cdot & & & 0 & a_{k,k}^{(k)} & & & \\ \cdot & \cdot & \ldots & \ldots & 0 & \cdot & \cdot & \\ \cdot & \cdot & \ldots & \ldots & \vdots & \cdot & \cdot & \\ 0 & \ldots & \ldots & 0 & 0 & \ldots & a_{n,n}^{(k)} & a_{n,n+1}^{(k)} \end{pmatrix}$$

in cui le righe e gli elementi sono espressi da

$$R_i^{(k)} = R_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} R_k^{(k)}, \qquad a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k)}, \qquad i > k,\ j = k, \ldots, 2n.$$

All'n-esimo passo si ottiene la matrice

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & \ldots & \ldots & a_{1,k-1}^{(1)} & a_{1,k}^{(1)} & \ldots & a_{1,n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{2,2}^{(2)} & \ddots & a_{2,k-1}^{(2)} & a_{2,k}^{(2)} & \ldots & a_{2,n}^{(2)} & a_{2,n+1}^{(2)} \\ \cdot & \cdot & \ddots & & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \ddots & a_{k-1,k-1}^{(k-1)} & \cdot & \cdot & \cdot & \\ \cdot & & & 0 & a_{k,k}^{(k)} & & & \\ \cdot & \cdot & \ldots & \ldots & & \cdot & \cdot & \\ 0 & \ldots & \ldots & 0 & \ldots & 0 & a_{n,n}^{(n-1)} & a_{n,n+1}^{(n-1)} \end{pmatrix}$$

Il metodo di Gauss consente agevolmente anche il calcolo del determinante della matrice, osservando come varia il determinante attraverso le operazioni eseguite.
La generica matrice $A^{(k)}$ si ottiene operando su $A^{(k-1)}$ la sequenza di premoltiplicazioni a), b). Dalle proprietà dei determinanti

$$\left| F_{ik}\left(-\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right) \right| = 1, \qquad \forall i, k$$

quindi in assenza di scambi di righe si ha

$$\big|\, A \,\big| = \big|\, F_1 \,\big|\, F_2 \,\big| \cdots \big|\, F_m \,\big|\, U \,\big| = \big|\, U \,\big|.$$

Se, inoltre, al k-esimo stadio si effettua uno scambio di righe, operazione a), si ha

$$\big|\, A^{(k)} \,\big| = -1 \big|\, A^{(k-1)} \,\big|.$$

Facendo variare $k = 1, \ldots, n$ si ottiene, in assenza di scambi di righe, ed essendo U triangolare superiore,

$$\big|\, U \,\big| = a_{nn}^{(n-1)} \cdot a_{n-1,n-1}^{(n-1)} \cdots a_{11}^{(1)} = \big|\, A \,\big|.$$

Se nel corso del procedimento intervengono m scambi di righe (cioè è necessario applicare l'operazione a) m volte), la formula diventa

$$\left| A \right| = (-1)^m a_{11}^{(1)} a_{22}^{(2)} \dots a_{nn}^{(n-1)}.$$

Ricapitolando, si può affermare che il valore del determinante di una matrice si ottiene moltiplicando tutti i fattori $a_{kk}^{(k-1)} (k = 1, \dots, n)$, detti elementi **pivot**, per $(-1)^m$, se m è il numero di scambi di righe effettuati.

Il metodo di Gauss può essere sintetizzato nel seguente schema. Si considera la matrice A (7.18)
Posto

$$A^{(1)} = A, \qquad a_{ij}^{(1)} = a_{ij}, \quad \begin{matrix} i = 1, \dots, n \\ j = 1, \dots, n+1 \end{matrix}$$

e indicati con $a_{ij}^{(k)}$ gli elementi della matrice dei coefficienti e del vettore dei termini noti al $k-$esimo passo, gli elementi al passo successivo sono dati dalla formule ricorsiva

$$\begin{cases} m_{ik} = -\dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} & \begin{matrix} k = 1, \dots, n-1 \\ i = k+1, \dots, n \end{matrix} \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} + m_{ik} a_{kj}^{(k)} & j = 1, \dots, n+1. \end{cases}$$

Tale algoritmo richiede che ad ogni stadio sia $a_{kk}^{(k)} \neq 0$; in caso contrario si procede con un opportuno scambio di righe per riportare in posizione diagonale un elemento diverso da zero (tecnica del pivot). Se ciò non fosse possibile, cioè se $a_{ik}^{(k)} = 0 \quad i = k, \dots, n$, il sistema sarebbe singolare.

Dopo aver trasformato il sistema, la soluzione si ottiene applicando le (7.9)

$$\begin{cases} x_n = \dfrac{a_{n,n+1}^{(n)}}{a_{nn}^{(n)}} & \\ x_k = \dfrac{1}{a_{kk}^{(n)}} \left(a_{k,n+1}^{(n)} - \displaystyle\sum_{j=k+1}^{n} a_{kj}^{(n)} x_j \right) & k = n-1, \dots, 1. \end{cases}$$

Il metodo di Gauss necessita di circa $n^2/2$ divisioni e $n^3/3$ moltiplicazioni

Osservazione 6. *È facile constatare che le matrici elementari che compaiono nella* (7.16) *sono*

$$F_{ik} \left(\frac{-a_{ik}^{(k)}}{a_{kk}^{(k)}} \right) \qquad i = k+1, \dots, n, \quad k = 1, \dots, n-1.$$

Ciascuna di queste matrici è triangolare inferiore con 1 *sulla diagonale, per cui posto*

$$L = F_{21}^{-1} \left(\frac{-a_{21}^{(1)}}{a_{11}^{(1)}} \right) \dots F_{n,n-1}^{-1} \left(\frac{-a_{n,n-1}^{(n-1)}}{a_{n-1,n-1}^{(n-1)}} \right) \tag{7.19}$$

si ha che L è triangolare inferiore con 1 sulla diagonale. Allora dalla (7.16) *si ha*

$$A = LT,$$

cioè il metodo di Gauss fornisce una fattorizzazione della matrice A in due matrici triangolari, una inferiore con 1 *sulla diagonale e l'altra superiore.*

Ricordando, infine, che

$$m_{ik} = -\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \quad k = 1, \ldots, i-1 \quad i = 2, \ldots, n$$

dalla (7.19) risulta che

$$L = \begin{pmatrix} 1 & & & & \\ -m_{21} & \ddots & & & \\ \vdots & \ddots & \ddots & & \\ \vdots & & & \ddots & \\ -m_{n1} & & & -m_{n,n-1} & 1 \end{pmatrix}.$$

Il teorema successivo dà una condizione sulla matrice A affinchè l'algoritmo di Gauss sia eseguibile.

Teorema 7.5 *L'algoritmo di Gauss è eseguibile, cioè gli elementi* $a_{kk}^{(k)}, \quad k = 1, \ldots, n$ *sono diversi da zero, se e solo se le sottomatrici principali di A,* $A^{[k]}$

$$A^{[k]} = \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ a_{21} & \ldots & \ldots \\ \vdots & & \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}$$

sono non singolari per $k = 1, \ldots, n$.

Dimostrazione. La dimostrazione segue per induzione su k.
Per $k = 1$ il teorema è banalmente vero perchè $A^{[1]} = a_{11}^{(1)}$.
Supponiamo che $A^{[1]}, \ldots, A^{[k-1]}$ siano non singolari, e quindi per ipotesi induttiva $a_{11}^{(1)}, \ldots, a_{k-1,k-1}^{(k-1)}$ siano diversi da zero, e dimostriamo che $A^{[k]}$ è non singolare se e solo se $a_{kk}^{(k)}$ è diverso da zero.
Per l'ipotesi induttiva l'algoritmo di Gauss può essere proseguito fino al $(k-1)-$esimo passo, ottenendo la matrice

$$A_k = F_{k-1} \ldots F_1 A = \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ \\ O & A_{22}^{(k)} \end{pmatrix}$$

ove

- $F_{k-1} \ldots F_1$ sono matrici elementari triangolari inferiori;
- $A_{11}^{(k)}$ è triangolare superiore con elementi diagonali $a_{11}^{(1)}, \ldots, a_{k-1,k-1}^{(k-1)}$;
- $A_{12}^{(k)}$ è una matrice $(k-1) \times (n-k+1)$;
- O è una matrice $(n-k+1) \times (k-1)$ con elementi tutti nulli;
- $A_{22}^{(k)}$ è una matrice $(n-k+1) \times (n-k+1)$.

Segue che $A_k^{[k]}$ è triangolare superiore con elementi diagonali $a_{11}^{(1)}, \ldots, a_{k-1,k-1}^{(k-1)}, a_{kk}^{(k)}$. Essendo $a_{11}^{(1)}, \ldots, a_{k-1,k-1}^{(k-1)} \neq 0$, $A_k^{[k]}$ è non singolare se e solo se $a_{kk}^{(k)}$ è diverso da 0.

Poichè, $F_{k-1}, \ldots, F_1$ sono triangolari inferiori con 1 sulla diagonale si può verificare che

$$A_k^{[k]} = F_{k-1}^{[k]} \ldots F_1^{[k]} A^{[k]}$$

da cui $A^{[k]}$ è non singolare se e solo se $A_k^{[k]}$ è non singolare, che abbiamo visto essere vero se e solo se $a_{kk}^{(k)} \neq 0$.

7.10.2 Il metodo di Gauss-Jordan

Il metodo di Gauss-Jordan è una variante del metodo di Gauss dovuta a Jordan e apparsa nel 1888 nella terza edizione del suo manuale di geodesia "Handbuch der Vermessungskunde".

Figura 7.6 - Carl Friedrich Gauss (1777-1855)

Figura 7.7 - Wilhelm Jordan (1842-1899)

Il metodo propone una trasformazione del sistema originale in un sistema equivalente di forma diagonale e permette di calcolare anche l'inversa della matrice. In termini di coefficienti, il metodo di Gauss-Jordan trasforma la matrice originale in una matrice diagonale che, in particolare, può essere l'identità.
Dato il sistema lineare

$$Ax = b \tag{7.20}$$

non singolare, è possibile determinare una sequenza finita di matrici elementari, $F_1, \ldots, F_m$ tali che

$$F_m F_{m-1} \ldots F_1 A = I$$

e conseguentemente da (7.20)

$$F_m F_{m-1} \ldots F_1 Ax = F_m F_{m-1} \ldots F_1 b \implies x = c \quad \text{con} \quad c = F_m F_{m-1} \ldots F_1 b.$$

Il procedimento per determinare le matrici elementari $F_n, \ldots, F_1$ è simile uguale a quello utilizzato nel metodo di Gauss. Anche in questo caso, per comodità di notazione, si fa uso della matrice aumentata (7.18).
Posto

$$A^{(1)} = A, \qquad a_{ij}^{(1)} = a_{ij}, \qquad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, n+1 \end{matrix}$$

al $k-$esimo stadio ($k = 1, \ldots, n$) gli elementi della matrice sono dati da

$$\begin{cases} a_{kj}^{(k)} = \dfrac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}} & j = k, \ldots, n+1 \\ a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k)} a_{kj}^{(k-1)} & i \neq k \end{cases} \tag{7.21}$$

Dopo n passi (cioè n premoltiplicazioni) la matrice $A^{(n)}$ si presenta nella forma

$$A^{(n)} = \begin{pmatrix} 1 & 0 & \ldots & 0 & a_{1,n+1}^{(n)} \\ 0 & 1 & \ldots & \vdots & a_{2,n+1}^{(n)} \\ \vdots & & \ddots & & \\ 0 & \cdot & \ldots & 1 & a_{n,n+1}^{(n)} \end{pmatrix}.$$

e quindi la soluzione del sistema è data da

$$x_i = a_{i,n+1}^{(n)} \qquad i = 1, \ldots, n.$$

Osservazione 7. *Il metodo di Gauss e il metodo di Gauss-Jordan possono essere usati per risolvere simultaneamente diversi sistemi lineari con la stessa matrice dei coefficienti e diverso termine noto. É sufficiente infatti costruire la matrice aumentata inserendo in coda tuti i termini noti considerati. L'algoritmo trasforma così la matrice una sola volta e contemporaneamente tutti i termini noti.*

Osservazione 8. *Il metodo di Gauss-Jordan può essere utilizzato per calcolare la matrice inversa. Se infatti si risolvono n sistemi con matrice dei coefficienti A e termini noti i vettori della base canonica di $\mathbb{R}^n$, che sono le colonne della matrice identità, i vettori soluzione sono, nell'ordine, le colonne della matrice inversa A^{-1}. In formule, se si risolvono i sistemi*

$$Ax^{(j)} = e_j, \qquad j = 1, \ldots, n,$$

dove e_j è il $j-$esimo vettore della base canonica ($j-$esima colonna della matrice identità), il vettore soluzione $x^{(j)}$ è la $j-$esima colonna della matrice A^{-1}.

Nella pratica si considera la matrice A "aumentata" della matrice identità. Alla fine del processo di Gauss-Jordan nelle ultime n colonne della matrice trasformata si trova la matrice inversa.

Il metodo di Gauss-Jordan e quello di Gauss forniscono la soluzione esatta dopo n stadi ma, naturalmente, questa affermazione è vera in assenza di errori di arrotondamento.

7.10.3 Propagazione dell'errore di arrotondamento

Il metodo di Gauss fornisce la soluzione esatta di un sistema in assenza di errori di arrotondamento; operando con un'aritmetica in virgola mobile e con matrici di dimensioni elevate bisogna tenere conto degli effetti dell'errore di propagazione, che possono essere tutt'altro che trascurabili.
Per studiare tale fenomeno e minimizzare l'errore di propagazione, si considera la relazione che esprime la formula ricorrente del $k-$mo passo del procedimento di Gauss.
Per semplicità di notazioni si pone $a'_{ij} \equiv a^{(k)}_{ij}$ e $a_{ij} \equiv a^{(k-1)}_{ij}$.

$$a'_{ij} = -\frac{a_{ik}}{a_{kk}} a_{kj} + a_{ij} \quad j = 1, \ldots, n, \quad i \neq k.$$

Siano r'_{ij} e r_{ij} l'errore relativo in a'_{ij} e a_{ij} e t, q, s l'errore di arrotondamento rispettivamente nella divisione, moltiplicazione e sottrazione. Applicando lo schema NONOR si ha

$$r'_{ij} = \Big[(r_{ik} \cdot 1 + r_{kk} \cdot (-1) + t) + r_{kj} \cdot 1 + q\Big] \left(\frac{(-a_{ik}/a_{kk})a_{kj}}{a'_{ij}} \right) + r_{ij} \frac{a_{ij}}{a'_{ij}} + s.$$

Moltiplicando l'errore relativo per a'_{ij}, si ottiene l'errore assoluto

$$\begin{aligned} E_{ij} &= \Big[r_{ik} - r_{kk} + t + r_{kj} + q\Big] \frac{-a_{ik}}{a_{kk}} a_{kj} + r_{ij} a_{ij} + s\Big[a_{ij} - \frac{a_{ik}}{a_{kk}} a_{kj}\Big] = \\ &= \Big[r_{ik} - r_{kk} + t + r_{kj} + q + s\Big] \frac{-a_{ik}}{a_{kk}} a_{kj} + (r_{ij} + s) a_{ij}. \end{aligned} \tag{7.22}$$

Se si suppone

$$\big|\, t \,\big|, \big|\, q \,\big|, \big|\, s \,\big| \leq 5 \cdot 10^{-d}, \qquad \big|\, r_{ij} \,\big| \leq m \cdot 10^{-d} \qquad m \geq 5,$$

dalla (7.22) si ottiene il limite superiore dell'errore assoluto per $\big|\, a'_{ij} \,\big|$

$$\big|\, E_{ij} \,\big| \leq \left[3(m+5) \big|\, a_{kj} \,\big| \left| \frac{-a_{ik}}{a_{kk}} \right| + (m+5) \big|\, a_{ij} \,\big| \right] 10^{-d}. \tag{7.23}$$

7.10.4 Pivotaggio della matrice

Se la matrice dei coefficienti è non singolare, gli algoritmi di Gauss e Gauss-Jordan sono applicabili solo se $a_{kk}^{(k)} \neq 0, \quad k = 1, \ldots, n.$
Se per qualche k si ha $a_{kk}^{(k)} = 0$, occorre procedere con un opportuno scambio di righe
Dall'esame della (7.23), si osserva inoltre che il limite dell'errore assoluto per $\left| a'_{ij} \right|$ dipende in qualche misura dal rapporto $\left| \dfrac{a_{ik}}{a_{kk}} \right|$, quindi se si sceglie nella $k-$ma colonna l'elemento di massimo modulo e, con opportuno scambio di righe, lo si porta nella posizione (k, k), vale

$$\left| \frac{a_{ik}}{a_{kk}} \right| \; < \; 1$$

ed il limite (7.23) risulta il più piccolo possibile.
Di conseguenza, nei processo di Gauss e di Gauss-Jordan, conviene, ad ogni stadio, scegliere opportunamente la riga k e portare in posizione diagonale l'elemento di massimo modulo sulla colonna, prima di procedere con la riduzione.
L'elemento di massimo modulo, prende il nome di **pivot**, e l'operazione è chiamata **"pivotaggio della matrice"**.

7.11 Matrici mal condizionate

Esistono molte matrici in cui, a piccole variazioni degli elementi non corrispondono altrettanto piccole variazioni degli elementi dell'inversa e del determinante.

Esempio 7.2 *Per la matrice*

$$A = \begin{pmatrix} 100 & 10 \\ 9.5 & 1 \end{pmatrix}$$

vale

$$\left| A \right| = 5 \quad e \quad A^{-1} = \begin{pmatrix} 0.2 & -2.0 \\ -1.9 & 20.0 \end{pmatrix}.$$

Modificando l'elemento a_{21} *da* 9.5 *a* 9.9 *si ha*

$$A' = \begin{pmatrix} 100 & 10 \\ 9.9 & 1 \end{pmatrix}, \quad \left| A' \right| = 1, \quad (A')^{-1} = \begin{pmatrix} 1 & -10 \\ -9.9 & 100 \end{pmatrix}$$

Sia il valore del determinante che gli elementi della matrice inversa di A'*, sono notevolmente diversi da quelli della matrice* A*, sebbene* A *e* A' *differiscano di 0.4 in un solo elemento.*

Definizione 7.15. *Si definiscono matrici* ***mal condizionate****, quelle in cui piccole variazioni degli elementi comportanto grandi variazioni degli elementi dell'inversa e del determinante.*

Matrici mal condizionate sono molto frequenti nelle applicazioni reali, basta pensare a dati sperimentali in cui piccole variazioni sono molto probabili, e queste possono causare errori notevoli nella soluzione del problema da cui i dati stessi provengono.
Per stabilire a priori se una matrice è mal condizionata sono stati proposti vari criteri, ma nessuno di questi è assoluto, nel senso che forniscono solo delle indicazioni sul condizionamento della matrice. Un criterio consiste nel calcolare il numero, detto **numero di condizionamento**,

$$m\,(A) = \left\| A \right\| \, \left\| A^{-1} \right\|,$$

e giudicare, dalla sua grandezza, il condizionamento della matrice.

In presenza di matrici mal condizionate, gli effetti dell'errore di propagazione sono molto più accentuati; occorre, perciò, studiare accorgimenti appositi che limitino gli effetti dell'errore propagato.

Per chiarire meglio gli effetti della propagazione dell'errore di arrotondamento, si risolve col metodo di Gauss-Jordan il sistema

$$\begin{cases} 15.0x_1 + 15.0x_2 + 14.0x_3 &= 58.0 \\ 15.0x_1 + 14.0x_2 + 13.0x_3 &= 55.0 \\ 14.0x_1 + 13.0x_2 + 12.0x_3 &= 51.0. \end{cases}$$

Applicando le (7.21) per $k = 1, 2, 3$ alla matrice aumentata

$$A^{(0)} = \begin{pmatrix} 15.0 & 15.0 & 14.0 & 58.0 \\ 15.0 & 14.0 & 13.0 & 55.0 \\ 14.0 & 13.0 & 12.0 & 51.0 \end{pmatrix}$$

ed arrotondando a sei cifre decimali si ottengono le matrici

$$A^{(1)} = \begin{pmatrix} 1.0 & 1.0 & 0.933333 & 3.866667 \\ 0.0 & -1.0 & -0.999995 & -3.000005 \\ 0.0 & -1.0 & -1.066662 & -3.133338 \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} 1.0 & 0.0 & -0.066662 & 0.866662 \\ 0.0 & 1.0 & 0.999995 & 3.000005 \\ 0.0 & 0.0 & -0.066667 & -0.133333 \end{pmatrix} \qquad A^{(3)} = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.999985 \\ 0.0 & 1.0 & 0.0 & 1.000030 \\ 0.0 & 0.0 & 1.0 & 1.999985 \end{pmatrix}.$$

La soluzione calcolata col metodo di Gauss-Jordan è

$$x_1 = 0.999985, \quad x_2 = 1.000030, \quad x_3 = 1.999985$$

mentre la soluzione esatta è

$$x_1 = 1, \quad x_2 = 1, \quad x_3 = 2.$$

Dall'esempio precedente si deduce che anche i metodi diretti, a causa dell'errore di arrotondamento, forniscono in generale una soluzione approssimata.

7.12 Soluzione numerica di sistemi lineari: metodi iterativi

Nel caso di sistemi di grandi dimensioni, i metodi diretti risultano inefficienti, pertanto è preferibile usare una diversa classe di metodi, detti **iterativi**, in quanto basati sulla costruzione di una successione di approssimazioni della soluzione che, sotto opportune ipotesi, converge al limite alla soluzione esatta.
Dato il sistema lineare

$$Ax = b, \tag{7.24}$$

un metodo iterativo per la soluzione di (7.24) consiste nel determinare una opportuna funzione vettoriale F per cui valga

$$x = F(x), \qquad \nu = 0, 1, \ldots$$

Scelta quindi un'approssimazione iniziale della soluzione, $x^{(0)}$, si calcola la successione

$$x^{(\nu+1)} = F\left(x^{(\nu)}\right), \qquad \nu = 0, 1, \ldots \tag{7.25}$$

Una volta determinata la funzione di iterazione F, per la sequenza (7.25) si pone il problema della convergenza, cioè sotto quali condizioni

$$\lim_{\nu \to \infty} x^{(\nu)} = x.$$

7.12.1 Il metodo di Jacobi

Dato il sistema lineare

$$Ax = b,$$

il metodo di Jacobi consiste nel risolvere la $j-$esima equazione del sistema rispetto alla $j-$esima componente del vettore soluzione, scrivendo così ciascuna componente in funzione di tutte le altre.

Figura 7.8 - Carl Gustav Jacob Jacobi (1804-1851)

In dettaglio, dato il sistema lineare

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n & = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n & = b_2 \\ \vdots & = \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n & = b_n \end{cases} \tag{7.26}$$

se $a_{ii} \neq 0 \quad i = 1, \ldots, n$, isolando ciascuna componente del vettore soluzione dalla corrispondente equazione si ottiene

$$\begin{cases} x_1 & = \dfrac{1}{a_{11}}\left[b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n\right] \\ x_2 & = \dfrac{1}{a_{22}}\left[b_2 - a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n\right] \\ \vdots & \\ x_n & = \dfrac{1}{a_{nn}}\left[b_n - a_{n1}x_1 - \ldots\cdots - a_{n,n-1}x_{n-1}\right]. \end{cases}$$

Scelto quindi un vettore $x^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^T$ come approssimazione iniziale della soluzione, la procedura iterativa di Jacobi è espressa da

$$\begin{cases} x_1^{(k+1)} & = \dfrac{1}{a_{11}}\left[b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}\right] \\ x_2^{(k+1)} & = \dfrac{1}{a_{22}}\left[b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}\right] \qquad k = 0, 1, 2, \ldots \\ \vdots & \\ x_n^{(k+1)} & = \dfrac{1}{a_{nn}}\left[b_n - a_{n1}x_1^{(k)} - \ldots\cdots - a_{n,n-1}x_{n-1}^{(k)}\right] \end{cases}$$

o, in forma compatta,

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)}\right], \qquad i = 1, 2, \ldots, n, \quad k = 0, 1, 2, \ldots \tag{7.27}$$

Per quanto riguarda la convergenza del metodo di Jacobi vale il seguente

Teorema 7.6 *Condizione sufficiente per la convergenza delle iterate* (7.27) *alla soluzione* x *del sistema* (7.26) *è*

$$\max_{1\le i\le n}\left(\frac{1}{|a_{ii}|}\sum_{\substack{j=1\\ j\neq i}}^{n}|\,a_{ij}\,|\right)<1. \tag{7.28}$$

Dimostrazione. Per la dimostrazione si rimanda alla successiva formulazione generale dei metodi iterativi.

Si osserva esplicitamente che (7.28) è condizione sufficiente, cioè il metodo di Jacobi può convergere anche se la (7.28) non è soddisfatta.

7.12.2 Il metodo di Gauss-Seidel

Nella costruzione del metodo di Gauss-Seidel, nel calcolo della $j-$esima componente della nuova iterata $k+1$, si usano le le prime $j-1$ componenti della nuova iterata stessa che sono già note.

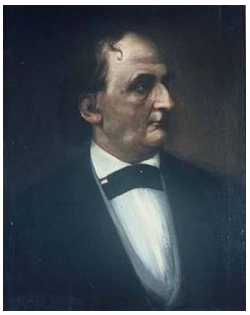


Figura 7.9 - Philipp Ludwig von Seidel (1821-1896)

Procedendo come nel metodo di Jacobi, scelta una approssimazione iniziale $x^{(0)}=\left(x_1^{(0)},\dots,x_n^{(0)}\right)^T$ il metodo di Gauss-Seidel è espresso dalle formule ricorsive

$$\begin{cases}
x_1^{(k+1)} &= \dfrac{1}{a_{11}}\left[b_1-a_{12}x_2^{(k)}-\cdots-a_{1n}x_n^{(k)}\right]\\
x_2^{(k+1)} &= \dfrac{1}{a_{22}}\left[b_2-a_{21}x_1^{(k+1)}-a_{23}x_3^{(k)}-\cdots-a_{2n}x_n^{(k)}\right]\\
\vdots & \\
x_i^{(k+1)} &= \dfrac{1}{a_{ii}}\left[b_i-a_{i1}x_1^{(k+1)}-\cdots-a_{i,i-1}x_{i-1}^{(k+1)}-a_{i,i+1}x_{i+1}^{(k)}-\cdots-a_{in}x_n^{(k)}\right]\\
\vdots & \\
x_n^{(k+1)} &= \dfrac{1}{a_{nn}}\left[b_n-a_{n1}x_1^{(k+1)}-\cdots-a_{n,n-2}x_{n-2}^{(k+1)}-a_{n,n-1}x_{n-1}^{(k+1)}\right].
\end{cases} \qquad k=0,1,2\dots$$

o, in forma compatta,

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\left[b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\right],\qquad i=1,2,\dots,n,\quad k=0,1,2,\dots \tag{7.29}$$

Per la convergenza del metodo di Gauss-Seidel si ha il seguente

Teorema 7.7 *Condizione sufficiente affinchè il metodo di Gauss-Seidel, applicato al sistema lineare $Ax = b$, converga è che*

$$\max_{1\le i\le n}\left(\frac{1}{|\,a_{ii}\,|}\sum_{\substack{j=1\\ j\neq i}}^{n}|\,a_{ij}\,|\right)<1.$$

Dimostrazione. Posti

$$r_i=\sum_{\substack{j=1\\ j\neq i}}^{n}\left|\frac{a_{ij}}{a_{ii}}\right| \qquad i=1,2,\dots,n$$
$$r=\max_{1\le i\le n} r_i$$
$$e^{(\nu)}=x^{(\nu)}-x \qquad \nu=0,1,\dots$$

si dimostra che se $r<1$

$$\left\|e^{(\nu)}\right\|_\infty=\max_{1\le j\le n}\left|e_j^{(\nu)}\right|\le r^\nu\left\|e^{(0)}\right\|_\infty \tag{7.30}$$

da cui la tesi, cioè $\lim_{\nu\to\infty} e^\nu\to\underline{0}$.

Dalla (7.29)

$$e_i^{(\nu)}=-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}e_j^{(\nu)}-\sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}e_j^{(\nu-1)} \qquad i=1,\dots,n.$$

Procedendo per induzione su i si prova che

$$\left|\,e_i^{(\nu)}\,\right|\le r\left\|e^{(\nu-1)}\right\|_\infty \qquad \nu=1,2,\dots, \quad i=1,\dots,n \tag{7.31}$$

Per $i=1$

$$\left|e_1^{(\nu)}\right|\le\sum_{j=2}^{n}\left|\frac{a_{1j}}{a_{11}}\right|\left|\,e_j^{(\nu-1)}\,\right|\le\left\|e^{(\nu-1)}\right\|_\infty\sum_{j=2}^{n}\left|\frac{a_{1j}}{a_{11}}\right|=r_1\left\|e^{(\nu-1)}\right\|_\infty\le r\left\|e^{(\nu-1)}\right\|_\infty.$$

Supposto che $\left|\,e_k^{(\nu)}\,\right|\le r\left\|e^{(\nu-1)}\right\|_\infty$, $\quad k=1,\dots,i-1,\quad$ si ha

$$\left|\,e_i^{(\nu)}\,\right| \le\sum_{j=1}^{i-1}\left|\frac{a_{ij}}{a_{ii}}\right|\left|\,e_j^{(\nu)}\,\right|+\sum_{j=i+1}^{n}\left|\frac{a_{ij}}{a_{ii}}\right|\left|\,e_j^{(\nu-1)}\,\right|\le$$

$$\le\left\|e^{(\nu-1)}\right\|_\infty\left\{\sum_{j=1}^{i-1}r\left|\frac{a_{ij}}{a_{ii}}\right|+\sum_{j=i+1}^{n}\left|\frac{a_{ij}}{a_{ii}}\right|\right\}\le$$

$$\le\left\|e^{(\nu-1)}\right\|_\infty\sum_{\substack{j=1\\ j\neq i}}^{n}\left|\frac{a_{ij}}{a_{ii}}\right|=r_i\left\|e^{(\nu-1)}\right\|_\infty\ \le\ r\left\|e^{(\nu-1)}\right\|_\infty$$

il che prova la (7.31), da cui segue la (7.30).

Osservazione 9. *La condizione di convergenza dei metodi di Jacobi e di Gauss-Seidel è formalmente identica, ma trattandosi di condizione sufficiente la convergenza di uno dei due metodi non implica la convergenza dell'altro.*

7.12.3 Formulazione generale. Splitting di una matrice

Un metodo iterativo per la soluzione del sistema

$$Ax = b \tag{7.32}$$

consiste nella costruzione di una successione di vettori $x^{(\nu)}$ che, sotto opportune ipotesi, converge alla soluzione x di (7.32).
I metodi iterativi di Jacobi e Gauss-Seidel sono due casi particolari di una più ampia classe di metodi, che costruiscono la successione $x^{(\nu)}$ mediante una formula iterativa del tipo

$$x^{(\nu+1)} = Mx^{(\nu)} + c \tag{7.33}$$

essendo M e c rispettivamente una matrice e un vettore opportuni che dipendono da A e da b.
Una possibile tecnica per costruire la matrice M consiste nel realizzare uno "splitting" della matrice A nella forma

$$A = N - P$$

con N matrice non singolare.
In questo modo la (7.32) diventa

$$Nx = Px + b$$

e, se N è non singolare,

$$x = N^{-1}Px + N^{-1}b. \tag{7.34}$$

Posto

$$M = N^{-1}P, \qquad c = N^{-1}b \tag{7.35}$$

la (7.34) diventa

$$x = Mx + c. \tag{7.36}$$

Assegnato un vettore iniziale $x^{(0)}$, si calcola quindi la successione

$$x^{(\nu+1)} = Mx^{(\nu)} + c \tag{7.37}$$

che è del tipo (7.33).

- **Metodo di Jacobi**

  Il metodo di Jacobi si ottiene dalla formulazione generale, ponendo

  $$N = D, \qquad P = N - A = D - A$$

  dove D indica la matrice diagonale che ha come elementi gli elementi diagonali di A e quindi P risulta essere la matrice che ha tutti 0 sulla diagonale principale e fuori diagonale gli elementi di A cambiati di segno.

$$N = \begin{pmatrix} a_{11} & 0 & \dots & & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & & & 0 & a_{n,n} \end{pmatrix}, \qquad P = \begin{pmatrix} 0 & -a_{12} & \dots & & -a_{1n} \\ -a_{21} & 0 & -a_{23} & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -a_{n-1,n} \\ -a_{n,1} & & & -a_{n,n-1} & 0 \end{pmatrix},$$

- **Metodo di Gauss-Seidel**

  Il metodo di Gauss-Seidel si ottiene dalla formulazione generale, ponendo

$$N = T, \qquad P = N - A = T - A$$

  dove T indica la matrice triangolare inferiore che ha come elementi la parte triangolare inferiore di A e quindi P risulta essere la matrice triangolare superiore con tutti 0 sulla diagonale principale e nella parte triangolare alta gli elementi di A cambiati di segno.

$$N = \begin{pmatrix} a_{11} & 0 & \dots & & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ a_{n1} & & & a_{n,n-1} & a_{n,n} \end{pmatrix}, \qquad P = \begin{pmatrix} 0 & -a_{12} & \dots & & -a_{1n} \\ 0 & 0 & -a_{23} & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -a_{n-1,n} \\ 0 & & & 0 & 0 \end{pmatrix},$$

7.12.4 Convergenza dei metodi iterativi con "splitting" della matrice

Come si è detto, il metodo (7.37) è convergente se

$$\lim_{\nu \to \infty} x^{(\nu)} = x$$

o, in modo equivalente, se posto

$$e^{(\nu)} = x^{(\nu)} - x \qquad \nu = 0, \dots, \tag{7.38}$$

si ha

$$\lim_{\nu \to \infty} e^{(\nu)} = \underline{0}.$$

Dalle (7.36), (7.37) e (7.38) risulta facilmente

$$e^{(\nu)} = M e^{(\nu-1)} = M^{\nu} e^{(0)}. \tag{7.39}$$

Dalle proprietà delle matrici convergenti (cfr. paragrafo 7.6.3) segue che una condizione sufficiente per la convergenza di un metodo iterativo è

$$\| M \| < 1. \tag{7.40}$$

In particolare se $M \equiv (m_{i,j})$, considerando $\| \cdot \|_{\infty}$, la (7.40) diventa

$$\| M \|_{\infty} = \max_i \sum_{j=1}^{n} |m_{i,j}| < 1. \tag{7.41}$$

A questo punto è possibile dimostrare la condizione sufficiente di convergenza per il metodo di Jacobi.
Dimostrazione. Data la definizione delle matrici N e P nel metodo di Jacobi, da (7.35)

$$M = N^{-1}P = I - N^{-1}A = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -\frac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\frac{a_{n,1}}{a_{n,n}} & & & -\frac{a_{n,n-1}}{a_{n,n}} & 0 \end{pmatrix},$$

La tesi segue ora facilmente applicando la (7.41).

Osservazione 10. *Nel metodo di Gauss-Seidel la matrice N è triangolare inferiore e $M = N^{-1}P = I - N^{-1}A$, per cui non è immediato verificare la condizione di convergenza*

$$\| M \|_\infty < 1.$$

La (7.39) permette di formulare la condizione necessaria e sufficiente per la convergenza

$$\lim_{\nu \to \infty} M^\nu = O \tag{7.42}$$

cioè la matrice M deve essere convergente. Dal Teorema 7.2 sulle matrici convergenti, la (7.42) equivale a

$$\rho(M) < 1,$$

dove $\rho(M)$ è il **raggio spettrale** della matrice M, cioè il massimo autovalore in modulo di M.

Il risultato appena ottenuto permette di definire un ulteriore parametro che misura la velocità di convergenza di un metodo iterativo.

Definizione 7.16. *Si chiama **raggio di convergenza** del metodo iterativo* (7.37)*, supposto convergente, la quantità*

$$R = -\log \rho(M)$$

ove $\rho(M)$ è il raggio spettrale di M.

Più grande è R, maggiore è la velocità di convergenza del metodo

7.12.5 Metodi di rilassamento

Al fine di meglio definire la matrice di iterazione M della (7.33), si considera ora la matrice A decomposta nella somma delle sue componenti diagonale, triangolare inferiore e superiore

$$A = D - E - F$$

con

$$D = \begin{pmatrix} a_{11} & & & \\ & \ddots & 0 & \\ & 0 & \ddots & \\ & & & a_{11} \end{pmatrix}, \quad -E = \begin{pmatrix} 0 & & & \\ & \ddots & 0 & \\ & a_{ij} & \ddots & \\ & & & 0 \end{pmatrix}, \quad -F = \begin{pmatrix} 0 & & & \\ & \ddots & a_{ij} & \\ & 0 & \ddots & \\ & & & 0 \end{pmatrix}.$$

- **Metodo di Jacobi**

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right]$$

$$N = D, \qquad P = N - A = D - (D - E - F) = E + F.$$

- **Metodo di Gauss-Seidel**

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right]$$

$$N = D - E, \qquad P = N - A = D - E - (D - E - F) = F.$$

Per migliorare le prestazioni dei metodi di Jacobi e di Gauss-Seidel si introduce un parametro ω, che sotto opportune ipotesi accelera la convergenza.

- **Metodo JOR (Jacobi Over-Relaxation)**

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right] - (1-\omega)\, x_i^{(k)}$$

$$N = \frac{1}{\omega} D, \qquad P = N - A = \frac{1}{\omega} D - (D - E - F) = \frac{1}{\omega} \left[(1-\omega)\, D + \omega\, (E+F) \right].$$

Per $\omega = 1$ si ritrova il metodo di Jacobi.

- **Metodo SOR (Successive Over-Relaxation)**

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] - (1-\omega)\, x_i^{(k)}$$

$$N = \frac{1}{\omega} \left(D - \omega E \right), \qquad P = N - A = \frac{1}{\omega} \left(D - \omega E \right) - (D - E - F) = \frac{1}{\omega} \left[(1-\omega)\, D + \omega F \right].$$

Per $\omega = 1$ si ritrova il metodo di Gauss-Seidel.

7.12.6 Alcuni risultati di convergenza

Per particolari classi di matrici è possibile stabilire a priori risultati di convergenza per i metodi finora analizzati.

Teorema 7.8 *Se il metodo di Jacobi converge* ($\omega = 1$)*, allora JOR converge per* $0 < \omega \leq 1$.

Teorema 7.9 *SOR diverge per* $\omega \leq 0$ *e* $\omega \geq 2$.

Definizione 7.17. *Una matrice* $A \in \mathbb{R}^{n \times n}$ *simmetrica si dice* **positiva definita** *se*

$$x \neq 0 \quad \Longrightarrow \quad x^T A x > 0$$

e **semidefinita positiva** *se*

$$x \neq 0 \quad \Longrightarrow \quad x^T A x \geq 0\,.$$

Teorema 7.10 *Se* A *è simmetrica positiva definita allora SOR converge sse* $0 < \omega < 2$*, quindi Gauss-Seidel,* ($\omega = 1$)*, converge.*

Teorema 7.11 *Se* A *è a diagonale dominante stretta per righe, cioè*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad j = 1, \ldots, n,$$

allora

a) *i metodi di Jacobi e di Gauss-Seidel convergono;*

b) *il metodo SOR converge se* $0 < \omega \leq 1$.

7.13 Matrici e Matlab

In Matlab sono predefiniti i comandi per ottenere vettori e matrici di struttura particolare

- vettori e matrici con elementi tutti **nulli** o $= \mathbf{1}$:

```
>> zeros(1,3)
ans =
     0     0     0

>> zeros(3,3)
ans =
     0     0     0
     0     0     0
     0     0     0
```

```
>> ones(1,3)
ans =
     1     1     1

>> ones(3,3)
ans =
     1     1     1
     1     1     1
     1     1     1
```

- vettori e matrici **costanti**: vettori e matrici aventi tutti gli elementi uguali

  Il comando `ones` può essere usato per ottenere array ad elementi costanti pari a k; ad esempio per $k = 4$

```
>> 4*ones(1,3)
ans =
     4     4     4
```

- matrice **trasposta** A^T

  Per ottenere in Matlab la matrice trasposta si usa il carattere ' (apice)

```
>> A
A =
     8     8     9     5     5
     8     1     7     4     5
     4     3     4     8     7
     7     1    10     8     8
     2     1     1     2     8
```

```
>> A'
ans =
     8     8     4     7     2
     8     1     3     1     1
     9     7     4    10     1
     5     4     8     8     2
     5     5     7     8     8
```

- **Matrice di Hilbert**

```
>> H=hilb(4)
H =
    1.0000    0.5000    0.3333    0.2500
    0.5000    0.3333    0.2500    0.2000
    0.3333    0.2500    0.2000    0.1667
    0.2500    0.2000    0.1667    0.1429
```

- **Matrice di Pascal**

```
>> P=pascal(5)
P =
     1     1     1     1     1
     1     2     3     4     5
     1     3     6    10    15
     1     4    10    20    35
     1     5    15    35    70
```

- **Matrice random:** una matrice random, in Matlab, è una matrice di numeri casuali in $[0, 1]$ generati con distribuzione uniforme

  Il comando per ottenere la matrice random di ordine n è `rand(row,col)`. Ad esempio per ottenere una matrice random di 2 righe e 3 colonne

```
>> R=rand(2,3)

R =

    0.2785    0.9575    0.1576
    0.5469    0.9649    0.9706
```

  La matrice random costruita in Matlab è, di solito, ben condizionata.

- **Matrici di Toeplitz:** il comando

```
toeplitz(C,R)
```

  costruisce la matrice di Toeplitz con gli elementi di `C` lungo la prima colonna, e quindi su tutte le sottodiagonali, e quelli di `R` lungo la prima riga, e quindi su tutte le sopradiagonali. La dimensione della matrice è riconosciuta automaticamente dalla lunghezza degli array stessi.

```
>> c=[0:4],r=[0:-1:-4]
c =
     0     1     2     3     4
r =
     0    -1    -2    -3    -4
```

```
>> T=toeplitz(c,r)
T =
     0    -1    -2    -3    -4
     1     0    -1    -2    -3
     2     1     0    -1    -2
     3     2     1     0    -1
     4     3     2     1     0
```

  Se la matrice di Toeplitz è simmetrica, essa è completamente definita da un solo vettore, prima riga o prima colonna indifferentemente. In questo caso il comando `toeplitz` viene richiamato con un solo parametro

```
>> c=[2,-1,zeros(1,2)],r=[2,-1,zeros(1,2)]
c =
     2    -1     0     0
r =
     2    -1     0     0
```

```
>> T=toeplitz(c,r)
T =
     2    -1     0     0
    -1     2    -1     0
     0    -1     2    -1
     0     0    -1     2
```

```
>> T=toeplitz(c)
T =
     2    -1     0     0
    -1     2    -1     0
     0    -1     2    -1
     0     0    -1     2
```

- Matrice **diagonale:** in Matlab è predefinito il comando per costruire la matrice diagonale avente lungo la diagonale gli elementi del vettore `v`

```
diag(v)
```

  Il comando `diag` permette anche di estrarre da una matrice il vettore degli elementi diagonali

```
>> v=[1:5]
v =
     1     2     3
```

```
>> A=diag(v)
A =
     1     0     0
     0     2     0
     0     0     3
```

```
>> w=diag(A)
w =
     1
     2
     3
```

Il comando `diag` permette inoltre di operare su diagonali diverse dalla principale

`diag(v,k)`

se `v` è un vettore di dimensione `n`, costruisce la matrice di dimensione `n+abs(k)` avente gli elementi di `v` lungo la diagonale k-esima

- ▷ $\mathbf{k} = 0$: diagonale principale;
- ▷ $\mathbf{k} > 0$: k-esima diagonale **sopra** la diagonale principale;
- ▷ $\mathbf{k} < 0$: k-esima diagonale **sotto** la diagonale principale;

Se `A` è una matrice dimensione `n`×`n`, il comando

`diag(A,k)`

costruisce il vettore avente come elementi la k-esima diagonale di `A`. In particolare

- ▷ $\mathbf{k} = 0$: $[a_{1,1}, a_{2,2}, \ldots, a_{n,n}]$ (diagonale principale);
- ▷ $\mathbf{k} > 0$: $[a_{1,1+k}, a_{2,2+k}, \ldots, a_{n-k,n}]$;
- ▷ $\mathbf{k} < 0$: $[a_{1+k,1}, a_{2+k,2}, \ldots, a_{n,n-k}]$.

```
>> v=[1:3]
v =
     1     2     3

>> A=diag(v,2)
A =
     0     0     1     0     0
     0     0     0     2     0
     0     0     0     0     3
     0     0     0     0     0
     0     0     0     0     0

>> B=diag(v,-2)
B =
     0     0     0     0     0
     0     0     0     0     0
     1     0     0     0     0
     0     2     0     0     0
     0     0     3     0     0

>> w=diag(B,-2)
w =
     1
     2
     3
```

- matrice **identità**: particolare matrice diagonale con tutti 1 sulla diagonale. Il comando per ottenere la matrice identità di ordine n è `eye(n)`

```
>> I3=eye(3)
I3 =
     1     0     0
     0     1     0
     0     0     1
```

- matrice **triangolare superiore:** il comando Matlab `triu` estrae la parte triangolare superiore di una matrice assegnata. In particolare
  - ▷ `triu(A)` estrae la parte triangolare superiore di `triu(A)`;
  - ▷ `triu(A,k)` estrae la parte triangolare superiore di `triu(A)` al di sopra della diagonale `k`-ma (diagonale compresa).

```
>> A=pascal(5)
A =
     1     1     1     1     1
     1     2     3     4     5
     1     3     6    10    15
     1     4    10    20    35
     1     5    15    35    70
```

```
>> U=triu(A)
U =
     1     1     1     1     1
     0     2     3     4     5
     0     0     6    10    15
     0     0     0    20    35
     0     0     0     0    70
```

```
>> U2=triu(A,2)
U2 =
     0     0     1     1     1
     0     0     0     4     5
     0     0     0     0    15
     0     0     0     0     0
     0     0     0     0     0
```

- matrice **triangolare inferiore**: il comando Matlab `tril` estrae la parte triangolare inferiore di una matrice assegnata. In particolare
  - ▷ `tril(A)` estrae la parte triangolare inferiore di `triu(A)`;
  - ▷ `tril(A,k)` estrae la parte triangolare inferiore di `triu(A)` al di sotto della diagonale `k`-ma (diagonale compresa).

```
>> A=pascal(5)
A =
     1     1     1     1     1
     1     2     3     4     5
     1     3     6    10    15
     1     4    10    20    35
     1     5    15    35    70
```

```
>> L=tril(A)
L =
     1     0     0     0     0
     1     2     0     0     0
     1     3     6     0     0
     1     4    10    20     0
     1     5    15    35    70
```

```
>> L2=tril(A,2)
L2 =
     1     1     1     0     0
     1     2     3     4     0
     1     3     6    10    15
     1     4    10    20    35
     1     5    15    35    70
```

- matrice **inversa**: `inv(A)`

```
>> A=[10,10,10,9
      10,10, 9,8
      10, 9, 8,7
       9, 8, 7,6];
```

```
>> inv(A)
ans =
   -0.0000   -1.0000    2.0000   -1.0000
   -1.0000    2.0000   -1.0000         0
    2.0000   -1.0000  -10.0000   10.0000
   -1.0000         0   10.0000  -10.0000
```

- **determinante** di una matrice: `det(A)`

```
>> A=[10,10,10,9
      10,10, 9,8
      10, 9, 8,7
       9, 8, 7,6];

>> det(A)
ans =
    1.0000
```

- **norma** di una matrice: il comando `norm(A,p)` calcola la norma p di una matrice `A`, dove `p` può assumere i valori `1, 2, Inf, 'Fro'`. Se non specificato, `p` assume per default valore `2`, cioè `norm(A)` restituisce la norma-2 o norma Euclidea. `norm(A,'Fro')` restituisce la norma di Frobenius definita come Le funzioni

$$\| A \|_F = \left(\sum_{i,k=1}^{n} |a_{ik}|^2 \right)^{1/2} .$$

- **numero di condizionamento**: il comando `cond(A,p)` calcola il numero di condizionamento di una matrice `A` in norma `p` con `p` che può assumere i valori `1, 2, Inf, 'Fro'`. Se non specificato, `p` assume per default valore `2`, cioè `cond(A)` restituisce il numero di condizionamento `A` in norma-2 o norma Euclidea.

  Il comando `rcond(A)` restituisce una stima del reciproco del numero di condizionamento di `A` calcolato in norma `1`. Se `A` è ben condizionata allora `rcond(A)` è vicino a 1, se `A` è mal condizionata allora `rcond(A)` è vicino a `eps` cioè allo 0-macchina.

```
>> H=hilb(5)
H =
    1.0000    0.5000    0.3333    0.2500    0.2000
    0.5000    0.3333    0.2500    0.2000    0.1667
    0.3333    0.2500    0.2000    0.1667    0.1429
    0.2500    0.2000    0.1667    0.1429    0.1250
    0.2000    0.1667    0.1429    0.1250    0.1111

>> cond(H)
ans =
   4.7661e+05

>> rcond(H)
ans =
   1.0597e-06
```

- **divisione tra matrici:** in Matlab sono previsti due tipi di divisione
  - ▷ `Z=M/P` equivale a `Z = M*inv(P)`
  - ▷ `Z=M\P` equivale a `Z = inv(M)*P`.

  Pertanto per calcolare la soluzione del sistema lineare

$$Ax = b$$

  si può usare il comando

```
x=A\b
```

Capitolo 8

Codici MatLab

Chi è convinto che i computer siano in grado di soppiantare i matematici non capisce niente né di computer né di matematica. È come credere che i biologi non servono più perché sono stati inventati i microscopi.

Ian Stewart (1945-)

Equazioni alle differenze Ordine 1 Per ricorrenza

```
function y=EqDiffOrd1Ricorr(a0,a1,bt,y0,n)
% Soluzione   a0 y(n+1)+a1 y(n)=bt,    y0 assegnato
% Calcola i primi n+1 termini per ricorrenza
%
% a0 y(n+1)+a1 y(n)=bt <-> y(n+1)-a y(n)=b
a=-a1/a0;
b=bt/a0;
y=[y0];
for k=2:n+1  % si "shifta" l'indice
    y(k)=b+a*y(k-1);
end
```

Equazioni alle differenze Ordine 1 Calcolo diretto

```
function y=EqDiffOrd1(a0,a1,bt,y0,n)
% Soluzione generale  a0 y(n+1)+a1 y(n)=bt,    y0 assegnato
%
% a0 y(n+1)+a1 y(n)=bt <-> y(n+1)-a y(n)=b
a=-a1/a0;
b=bt/a0;
y=[y0];
for k=2:n+1
    if a==1
        y(k)=y0+(k-1)*b;
    else
        y(k)=a^(k-1)*(y0-b/(1-a))+b/(1-a);
    end
end
```

Equazioni alle differenze Ordine 2 Per ricorrenza

```
function y=EqDiffOrd2Ricorr(a0,a1,a2,b,y0,y1,n)
% Soluzione Equazioni alle differenze di ordine 2
%    a2 y(n+2)+a1 y(n+1)+a0 y(n)=b,   y0,y1  assegnati
%
y=[y0,y1];
for k=3:n+2  % si "shifta" l'indice
    y(k)=b-a0*y(k-2)-a1*y(k-1);
end
```

Equazioni alle differenze Ordine 2 Calcolo diretto

```
function yn=EqDiffOrd2(a0,a1,a2,b,y0,y1,n)
% Soluzione Equazioni alle differenze di ordine 2
%    a2 y(n+2)+a1 y(n+1)+a0 y(n)=b,   y0,y1  assegnati
% Calcola il termine n-esimo mediante la formula
%
% Si trova prima la soluzione particolare (costante?) dell'equazione
   den=a2+a1+a0;
   if den~=0
       ystarn=b/den;
       ystar0=b/den;
       ystar1=b/den;
   else
       den1=2*a2+a1;
       if den1~=0
           ystarn=b/den1*n;
           ystar0=0;
           ystar1=b/den1;
       else
           den2=4*a2+a1;
           ystarn=b/den2*n^2;
           ystar0=0;
           ystar1=b/den2;
       end
   end
% Si trova la soluzione dell'equazione omogenea associata
% e si costruiscono matrice e vettore dei termini noti
% del sistema che si ottiene imponendo le condizioni iniziali
   Delta=a1^2-4*a0*a2;
   if Delta~=0
       lambda1=(-a1+sqrt(Delta))/(2*a2);
       lambda2=(-a1-sqrt(Delta))/(2*a2);
       A=[1,1;lambda1,lambda2];
       tn=[y0-ystar0;y1-ystar1];
% Si risolve il sistema e si determina la soluzione
      [c1,c2]=Sistema(A,tn);
       yn=c1*lambda1^n+c2*lambda2^n+ystarn;
   elseif Delta==0
```

```
       lambda=-a1/(2*a2);
       A=[1,0;lambda,lambda];
       tn=[y0-ystar0;y1-ystar1];
% Si risolve il sistema e si determina la soluzione
      [c1,c2]=Sistema(A,tn);
       yn=c1*lambda^n+c2*n*lambda^n+ystarn;
   else
       disp('Delta negativo')
       return
%       error('Delta negativo')     % in alternativa
   end
%%%
function [c1,c2]=Sistema(A,tn);
% Soluzione sistema delle condizioni iniziali
   A1=[tn,A(:,2)];
   A2=[A(:,1),tn];
   denA=det(A);
   c1=det(A1)/denA;
   c2=det(A2)/denA;
```

Metodo di Archimede per il calcolo di π con nro max di lati assegnato

```
function [piginf,pigsup,err]=Pigreco_NmaxAll(nmax)
% Calcolo di pigreco con il metodo di Archimede e numero max di lati fissato
% Restituisce le successive approssimazioni per difetto e per eccesso e gli errori
%
format long;
n=6;
Li=1;
Ls=2*Li/sqrt(4-Li^2);
pigi=n*Li/2;
pigs=n*Ls/2;  %pigs=n*Li/sqrt(4-l^2)   pigs=2*pigi/sqrt(4-Li^2);
piginf=[pigi];
pigsup=[pigs];
while n<nmax
   Li=sqrt(2-(sqrt(4-Li^2)));
%     Li=Li/sqrt(2+(sqrt(4-Li^2)));     % formula stabile
    Ls=2*Li/sqrt(4-Li^2);
    n=2*n;
    pigi=n*Li/2;
    pigs=n*Ls/2;  %pigs=n*Li/sqrt(4-l^2)   pigs=2*pigi/sqrt(4-Li^2);
    piginf=[piginf;pigi];
    pigsup=[pigsup;pigs];
end;
err=pigsup-piginf;
```

Metodo di Archimede per il calcolo di π con precisione assegnata

```
function [pinf,psup,nl]= Pig_TollAll(prec)
% Calcolo di pigreco con il metodo di Archimede e precisione richiesta
% Restituisce le successive approssimazioni per difetto e per eccesso
% e i relativi numeri di lati
%
format long;
n=6;
l=1;
nl=[n];
p1=n*l/2;
p2=2*p1/sqrt(4-l^2);
pinf=[p1];
psup=[p2];
while p2-p1>prec
    l=l/sqrt(2+sqrt(4-l^2));
    n=2*n;
    p1=n*l/2;
    p2=2*p1/sqrt(4-l^2);
    nl=[nl;n];
    pinf=[pinf;p1];
    psup=[psup;p2];
end;
```

Algoritmo A per il calcolo di $\sin t$

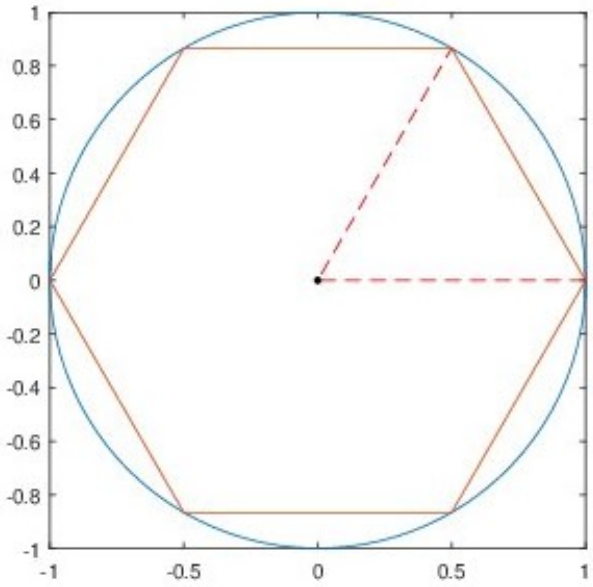
```
function sint = Sin_A(t,n);
% Calcolo di sin(t) con l'algoritmo A
%
format long
s=2*t/2^n;
for k=1:n
    s=s*sqrt(4-s^2);
end
sint=s/2;
```

Algoritmo T per il calcolo di $\sin t$

```
function sint = Sin_T(t,n);
% Calcolo di sin(t) con l'algoritmo T
%
format long
s=2*t/3^n;
for k=1:n
    s=3*s-s^3;
end
sint=s/2;
```

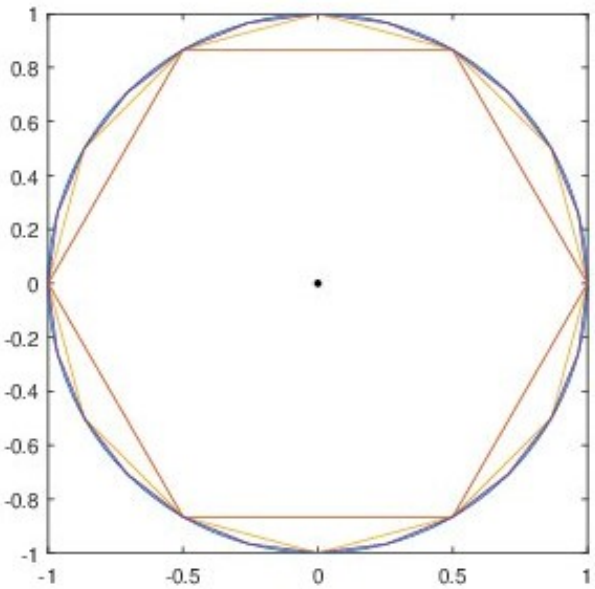
Plot di una circonferenza con esagono inscritto

```
% Disegna una circonferenza, l'esagono regolare inscritto
% e il triangolo determinato dai primi 2 vertici e dal centro
%
t=linspace(0,2*pi);
x0=0;   % ascissa centro
y0=0;   % ordinata centro
r=1;    % r=raggio
x=x0+r*cos(t);   % ascisse dei punti della circonferenza
y=y0+r*sin(t);   % ordinate dei punti della circonferenza
% Disegna la circonferenza
plot(x,y)
axis square
hold on
% Marka il centro
plot(x0,y0,'k.','Markersize',10)
% Disegna l'esagono inscritto
tesa=linspace(0,2*pi,7);
xesa=x0+r*cos(tesa);
yesa=y0+r*sin(tesa);
plot(xesa,yesa)
% Disegna il triangolo
xtria=[xesa(1),x0,xesa(2)];
ytria=[yesa(1),y0,yesa(2)];
plot(xtria,ytria,'r--')
```



Plot di una circonferenza con poligoni inscritti

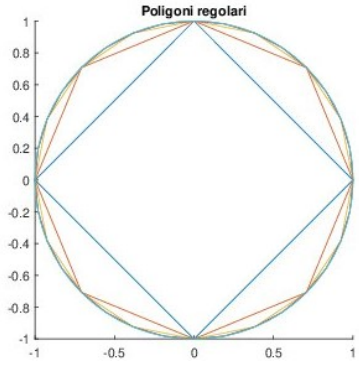
```
% Disegna una circonferenza, e i poligoni regolari inscritti
%
nstart=6;
kend=5;
t=linspace(0,2*pi);
x0=0;   % ascissa centro
y0=0;   % ordinata centro
r=1;    % r=raggio
x=x0+r*cos(t);   % ascisse dei punti della circonferenza
y=y0+r*sin(t);   % ordinate dei punti della circonferenza
% Disegna la circonferenza
plot(x,y)
axis square
hold on
% Marka il centro
plot(x0,y0,'k.','Markersize',10)
% Disegna i poligoni inscritti con nro lati multiplo di nro lati iniziali
for k=0:kend
    pause
    n=nstart*2^k;
    tv=linspace(0,2*pi,n+1);
    xv=x0+r*cos(tv);
    yv=y0+r*sin(tv);
    plot(xv,yv)
end
```



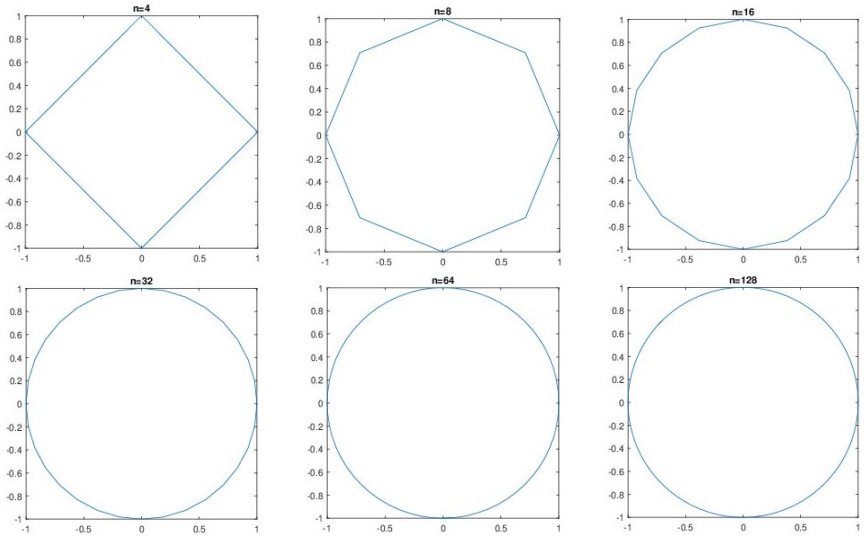
Plot di 6 poligoni regolari

```
% Disegna 6 poligoni regolari con numero lati multipli secondo una potenza di 2
%
nstart=4;
x0=0;   % ascissa centro
y0=0;   % ordinata centro
r=1;    % r=raggio
%%%% Disegna i poligoni sovrapposti sulla stessa finestra grafica
figure(100)
for k=1:6
    hold on
    n=nstart*2^(k-1);
    t1=linspace(0,2*pi,n+1);
    x1=x0+r*cos(t1);
    y1=y0+r*sin(t1);
    plot(x1,y1)
end
    title('Poligoni regolari')
    axis square
hold off
%%%% Disegna i poligoni in 6 finestre grafiche distinte
for k=1:6
    figure(k)
    n=nstart*2^(k-1);
    t2=linspace(0,2*pi,n+1);
    x2=x0+r*cos(t2);
    y2=y0+r*sin(t2);
    plot(x2,y2)
    titolo=['n=',num2str(n)];
    title(titolo)
    axis square
end
%%%% Disegna i poligoni nella stessa finestra grafica ma separati
figure(7)
for k=1:6
    n=4*2^(k-1);
    t3=linspace(0,2*pi,n+1);
    x3=x0+r*cos(t3);
    y3=y0+r*sin(t3);
    subplot(2,3,k)
    plot(x3,y3)
    titolo=['n=',num2str(n)];
    title(titolo)
    axis square
end
```

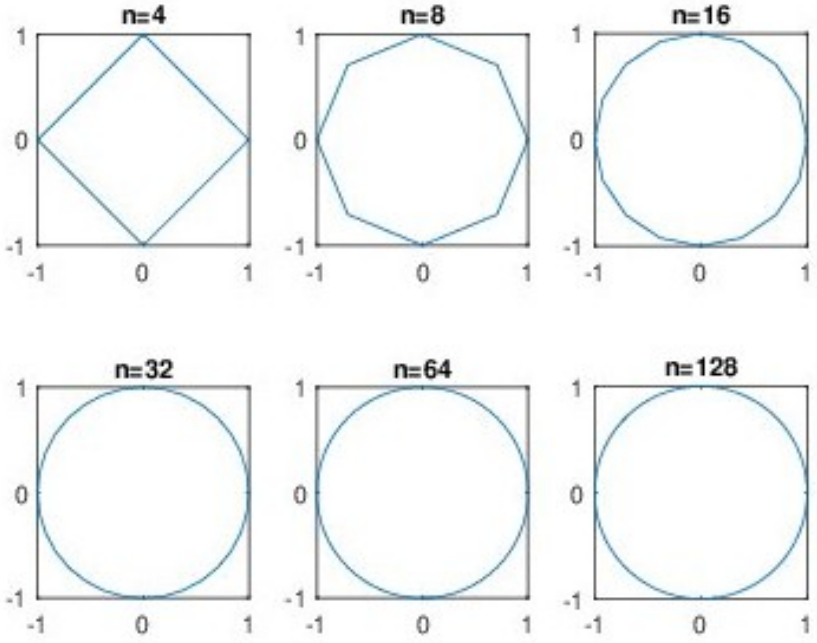
Grafici sovrapposti nella stessa finestra grafica



Grafici in 6 finestre grafiche distinte



Grafici nella stessa finestra grafica ma separati



Algorimo di Horner

```
function [P,PD] = Horner(a,z);
% Calcolo del valore di un polinomio e
% delle successive derivate con l'algoritmo di Horner
%
format long;
P=[];
n=length(a)-1; % grado del polinomio
nr=n+2;
nc=n+1;
Horn=zeros(nr,nc);
Horn(1,:)=a;
Horn(:,1)=a(1);
for i=2:nr
    for j=2:nc-i+2
        Horn(i,j)=Horn(i,j-1)*z+Horn(i-1,j);
    end
end
Horn=Horn(2:nr,:);
for k=0:n
    P(k+1)=Horn(k+1,nc-k)*factorial(k);
end
%%%%  in alternativa
% HornF=fliplr(Horn);
% F=factorial([0:n]);
% P=diag(HornF)'.*F;
%
for k=1:n+1
    PD(k)=polyval(a,z);
    a=polyder(a);
end
```

Metodo di Bisezione

```
function [z,fz,err,it,iflag]=Bisezione(f,a,b,eps,itmax)
f=inline(f);
iflag=1;
err=abs(b-a)/2;
it=0;
fa=f(a);
fb=f(b);
if fa*fb>0
    iflag=0;
else
    while err>eps & it<itmax
        c=(a+b)/2;
        fc=f(c);
        if (fc*fa > 0)
            a=c;
            fa=fc;
        else
            b=c;
        end
        err=abs(b-a)/2;
        it=it+1;
    end
end
z=(a+b)/2;
fz=f(z);
if it>=itmax
    iflag=-1;
end
```

```
     METODO DI BISEZIONE  f(x)=1/x-3
Step 1    [a,b]=[0.10000,0.50000]

 f(0.10000)=7.00000,  f(0.50000)=-1.00000

 c=0.30000,      f(0.30000)=0.33333

Step 2   [a,b]=[0.30000,0.50000]

 f(0.30000)=0.33333,  f(0.50000)=-1.00000

 c=0.40000,      f(0.40000)=-0.50000

Step 3   [a,b]=[0.30000,0.40000]

 f(0.30000)=0.33333,  f(0.40000)=-0.50000

 c=0.35000,      f(0.35000)=-0.14286

Step 4   [a,b]=[0.30000,0.35000]

 f(0.30000)=0.33333,  f(0.35000)=-0.14286

 c=0.32500,      f(0.32500)=0.07692

Step 5   [a,b]=[0.32500,0.35000]

 f(0.32500)=0.07692,  f(0.35000)=-0.14286

 c=0.33750,      f(0.33750)=-0.03704
```

Metodo della Falsa Posizione

```
function [z,fz,k,err]=RegulaFalsi(f,a,b,itmax,eps)
f=inline(f);
iflag=1;   err=eps+1;  it=0;
x=[a,b];
fa=f(a);
fb=f(b);
z=a;
fz=fa;
k=2;
ptofisso=a;
while it<itmax & err>eps
   x(k+1)=x(k)-f(x(k))*(x(k)-ptofisso)/(f(x(k))-f(ptofisso));
   err=abs(x(k+1)-x(k));
   if (f(x(k+1))*fa<0),
      ptofisso=a;
   else
      ptofisso=b;
   end
   k=k+1;
end
z=x(k);
fz=f(z);
if k==itmax
    iflag=0;
end
```

```
     METODO DELLA FALSA POSIZIONE  f(x)=1/x-3
 [x(1),x(2)]=[0.10000,0.50000]
 f(0.10000)=7.00000,  f(0.50000)=-1.00000

 x(3)=0.45000,      f(0.45000)=-0.77778
 x(4)=0.41500,      f(0.41500)=-0.59036
 x(5)=0.39050,      f(0.39050)=-0.43918
 x(6)=0.37335,      f(0.37335)=-0.32155
```

Metodo delle Corde

```
function [z,k,err]=Corde(f,a,b,itmax,eps)
f=inline(f);
iflag=1;
err=eps+1;
it=0;
x=[a,b];
z=a;
fz=f(a);
k=2;
ptofisso=a;
while it<itmax & err>eps
   x(k+1)=x(k)-f(x(k))*(x(k)-x(k-1))/(f(x(k))-f(x(k-1)));
   err=abs(x(k+1)-x(k));
   k=k+1;
end
z=x(k);
if k==itmax
    iflag=0;
end
```

```
     METODO DELLE CORDE  f(x)=1/x-3
 [x(1),x(2)]=[0.10000,0.50000]
 f(0.10000)=7.00000,  f(0.50000)=-1.00000

 x(3)=0.45000,      f(0.45000)=-0.77778
 x(4)=0.27500,      f(0.27500)=0.63636
 x(5)=0.35375,      f(0.35375)=-0.17314
```

Metodo di Newton

```
function [zv,fzv,zerr,it,iflag]=Newton(f,df,zold,itmax,eps)
f=inline(f);
df=inline(df);
iflag=1;   err=eps+1;   it=0;
fzold=f(zold);
zv=[zold];   fzv=[fzold];
zerr=[err];
while it<itmax & err>eps
   dfzold=df(zold);
   if dfzold == 0
       iflag=-1;
       return
   else
       z=zold-fzold/dfzold;
       err=abs(z-zold);
       fz=f(z);
       zv=[zv;z];
       fzv=[fzv;fz];
       zerr=[zerr;err];
       it=it+1;
       zold=z;
       fzold=fz;
   end
end
if it==itmax
    iflag=0;
end
```

```
     METODO Di NEWTON  f(x)=1/x-3
 x(1)=0.40000,  f(0.40000)=-0.50000
 x(2)=0.32000,  f(0.32000)=0.12500
 x(3)=0.33280,  f(0.33280)=0.00481
 x(4)=0.33333,  f(0.33333)=0.00001
```

Metodo di Gauss

```
function [x,flag]=Gauss(A,b);
%%% Metodo di Gauss con pivotaggio per la soluzione simultanea
%%% di più sistemi lineari con la stessa matrice dei coefficienti
%%%
   flag=1;
   n=size(A,2);
   A=[A,b];
   [nr,nc]=size(A);
   ns=nc-n;
   x=zeros(nr,ns);
   for k=1:nr
      [maxk,row]=max(abs(A(k:nr,k)));
      if maxk<eps
         flag=-1;
         return
      end
      row=row+k-1;
      if row~=k
         A([k row],:)=A([row k],:);
      end
%       A(k,k:nc)=A(k,k:nc)/A(k,k);
      for i=k+1:nr
          A(i,k:nc)=A(i,k:nc)-A(i,k)/A(k,k)*A(k,k:nc);
      end
   end
   for j=1:ns
       for k=nr:-1:1
           x(k,j)=(A(k,n+j)-A(k,k+1:n)*x(k+1:nr,j))/A(k,k);
       end
   end
```

Metodo di Gauss-Jordan

```
function [x,flag]=GaussJordan(A,b);
%%% Metodo di Gauss-Jordan con pivotaggio per la soluzione simultanea
%%% di più sistemi lineari con la stessa matrice dei coefficienti
%%%
   flag=1;
   A=[A,b];
   [nr,nc]=size(A);
   x=[];
   for k=1:nr
      [maxk,row]=max(abs(A(k:nr,k)));
      if maxk<eps
         flag=-1;
         return
      end
      row=row+k-1;
      if row~=k
         A([k r],:)=A([r k],:);
      end
      A(k,k:nc)=A(k,k:nc)/A(k,k);
      for i=[1:k-1,k+1:nr]
          A(i,k:nc)=A(i,k:nc)-A(i,k)*A(k,k:nc);
      end
   end
   x=A(:,nr+1:nc);
```