Corso di Calcolo Numerico e Programmazione

——————— ~ ♢ ~ ———————

Eserciziario ragionato

——————— ~ ♢ ~ ———————

Claudia Fassino
(fassino@dima.unige.it)

Contents

Premessa

Queste dispense sono relative al corso di Calcolo Numerico e Programmazione, all'interno del corso di Laurea Triennale in Scienza dei Materiali e del corso di Laurea Triennale in Chimica e Tecnologie Chimiche (Università di Genova). Le dispense sono principalmente rivolte a tutti gli studenti che, pur non frequentando il corso di Laurea in Matematica, devono sostenere un esame di Calcolo Numerico. Tali appunti sono stati scritti con l'intento di fornire agli studenti un aiuto per la preparazione della prova scritta d'esame e non vogliono e non possono in alcun modo sostituire i tanti libri sull'argomento presenti in letteratura, quali ad esempio "Introduzione alla matematica computazionale" di Bevilacqua, Bini, Capovani, Menchi (Zanichelli) per quanto riguarda l'errore, le equazioni non lineari e l'interpolazione e "Metodi numerici per l'algebra lineare" di Bini, Capovani, Menchi (Zanichelli) per quanto l'algebra lineare.

Ogni capitolo è dedicato a un singolo argomento e contiene, oltre a un breve richiamo dei principali risultati utili per svolgere gli esercizi, una rassegna di esercizi svolti e una raccolta di esercizi proposti di cui si fornisce il solo risultato in fondo al capitolo stesso. Gli esercizi con l'asterisco $(*)$ non sono di base; essi presentano un livello di difficoltà più alto e sono rivolti a coloro che vogliono approfondire (seppur leggermente) la parte teorica o "cimentarsi" nella soluzione di problemi piú complessi.

La soluzione proposta degli esercizi svolti è piuttosto schematica e deve servire soprattutto da traccia: per utilizzare il testo in modo proficuo è necessario, dopo aver compreso la parte teorica, svolgere i singoli esercizi in modo autonomo. Solo a questo punto, confrontando l'intero procedimento e non solo il risultato, lo studente potrà avere un riscontro sull'effettiva comprensione delle nozioni utilizzate. Al fine di una buona comprensione della materia, non è necessario svolgere centinaia di esercizi, ma piuttosto comprendere a fondo quelli svolti.

Un ulteriore suggerimento è quello di risolvere alcuni esercizi anche (ma ovviamente non solo) utilizzando MatLab per poter avere un riscontro "pratico" della soluzione teorica trovata.

Vorrei, infine, scusarmi per gli inevitabili errori di stampa/calcolo presenti nel testo e ringrazio in anticipo tutti coloro che mi segnaleranno sviste ed errori e daranno suggerimenti per rendere migliori queste dispense.

Chapter 1

Aritmetica floating point ed errori

In questo capitolo vengono presentati esercizi relativi alla rappresentazione in base 2 dei numeri reali, alla rappresentazione floating point (virgola mobile) e agli errori che perturbano il valore di una funzione quando viene calcolata a partire da dati perturbati e quando le operazioni vengono eseguite da un computer, utilizzando l'aritmetica floating point.

1.1 Aritmetica in base 2 e floating point

- Scelta una base β, ogni numero in tale base è rappresentato da una sequenza di cifre che rappresentano i coefficienti delle potenze della base. Le cifre dopo il punto (la virgola) sono i coefficienti delle potenze negative e $\beta^0 = 1$. Ad esempio $d_1d_2d_3.d_4d_5d_6$ in base β corrisponde al numero, espresso in base 10,

$$d_1\beta^2 + d_2\beta^1 + d_3\beta^0 + d_4\beta^{-1} + d_5\beta^{-2} + d_6\beta{-}2$$

  Le cifre $d_i \in \{0, 1, \ldots, d-1\}$. In base 2 si ha $\beta = 2$ e si utilizzano solo le cifre 0 e 1.

- La rappresentazione floating point di un qualunque numero in base β consiste nello scrivere il numero nella forma $\pm 0.d_1d_2 \ldots d_td_{t+1} \cdots * \beta^p$. Ogni numero può essere scritto in questa forma.

- Un computer usa la rappresentazione floating point con t cifre, t fissato, e $\beta = 2$.

- Poiché un computer usa solo un numero t finito di cifre, ogni numero in input viene sporcato da errore in quanto vengono eliminate le cifre che non possono essere memorizzate.

  Dato $x = \pm 0.d_1 d_2 \dots d_t d_{t+1} \cdots * \beta^p$, con il **troncamento** viene memorizzato il numero $\widetilde{x} = \pm 0.d_1 d_2 \dots d_t * \beta^p$ e con l'**arrotondamento** viene memorizzato il numero $\widetilde{x} = \pm 0.d_1 d_2 \dots \widetilde{d_t} * \beta^p$, dove $\widetilde{d_t} = d_t$ se $d_{t+1} < \beta/2$ e $\widetilde{d_t} = d_t + 1$ altrimenti.

- Si ha che $|\widetilde{x} - x|/|x| < \mathbf{u}$, dove $\mathbf{u}$ è detta precisione di macchina e $\mathbf{u} = \beta^{1-t}$ usando il troncamento e $\mathbf{u} = \beta^{1-t}/2$ usando l'arrotondamento.

- Se a e b sono numeri di macchina rispetto a t cifre, cioè la loro rappresentazione floating point con t cifre non introduce errore, allora indicato con $\diamond$ una delle 4 operazioni aritmetiche e con $fl(a \diamond b)$ il risultato calcolato in aritmetica floating point, si ha che

$$\frac{|fl(a \diamond b) - (a \diamond b)|}{|a \diamond b|} < \mathbf{u} \tag{1.1}$$

Esercizio 1.1.1. $(*)$ **Rappresentazione numeri in base 2.**

- Rappresentare in base 10 i seguenti numeri espressi in base 2: 101, 10111, -1010, 11.1.

$$\begin{aligned} 101 &=_{10} 2^2 + 2^0 &&= 5 \\ 10111 &=_{10} 2^4 + 2^2 + 2^1 + 2^0 &&= 23 \\ -1010 &=_{10} -(2^3 + 2^1) &&= -10 \\ 11.1 &=_{10} 2^2 + 2^1 + 2^{-1} &&= 5 + \tfrac{1}{2} = 5.5 \end{aligned}$$

- Scrivere la tabella dei primi 10 numeri interi in base 2.

Base 10	Base 2	Base 10	Base 2
0	0000	5	0101
1	0001	6	0110
2	0010	7	0111
3	0011	8	1000
4	0100	9	1001

- Calcolare le seguenti somme di numeri binari $101 + 1010$ e $101 + 1001$. La somma si calcola come in base 10 cifra per cifra, ricordando che $0 + 0 = 0$, $0 + 1 = 1$, mentre $1 + 1 = 0$ con riporto di 1 sulla cifra successiva perché in base 2 si ha $1 + 1 = 10$.

$$\begin{array}{r} 101 \quad + \\ 1010 \\ \hline 1111 \end{array} =_{10} 15 \qquad \begin{array}{r} 101 \quad + \\ 1001 \\ \hline 1110 \end{array} =_{10} 14$$

Esercizio 1.1.2. $(*)$ **Notazione floating point, aritmetica finita.**

- Trasformare in base 10 i seguenti numeri frazionari espressi in aritmetica floating point con base $\beta = 2$:

$$0.101 \cdot 2^2 \qquad 0.100 \cdot 2^2 \qquad 0.110 \cdot 2^2 \qquad 0.111 \cdot 2$$

In questo caso ogni cifra è il coefficiente di una potenza negativa di 2 e il numero finale si ottiene moltiplicando il risultato cos ottenuto per la potenza di 2 indicata.

$$\begin{array}{lcll} .101\,2^2 & =_{10} & (1/2 + 1/8)4 & = \quad 5/2 \\ .100\,2^2 & =_{10} & (1/2)4 & = \quad 2 \\ .110\,2^2 & =_{10} & (1/2 + 1/4)4 & = \quad 3 \\ .111\,2 & =_{10} & (1/2 + 1/4 + 1/8)2 & = \quad 7/4 \end{array}$$

- Calcolare le seguenti somme in aritmetica floating point, utilizzando il troncamento e $t = 3$ cifre:

$$s_1 = 0.101 \cdot 2^2 + 0.100 \cdot 2^2 \qquad s_2 = 0.110 \cdot 2^2 + 0.111 \cdot 2$$

Somma esatta $s_1 = .101\,2^2 + .100\,2^2 = 1.001\,2^2 =_{10} 4 * (1 + 2^{-3}) = 4.5$.

Somma floating point $fl(s_1)$.
Innanzitutto si esegue la somma cifra per cifra, poi si attua la normalizzazione scrivendo il risultato in forma floating point e si eliminano le cifre "eccedenti" mediante il troncamento del risultato.

$$\begin{array}{rl} .101 \quad 2^2 \qquad + & \\ .100 \quad 2^2 & \\ \hline 1.001 \quad 2^2 & = .1001 \cdot 2^3 \qquad \text{normalizzazione.} \\ & \approx .100 \cdot 2^3 \quad \text{troncamento fl. point.} \end{array}$$

Si ha quindi che $fl(s_1) = 0.100 \cdot 2^3 =_{10} 0.5 * 8 = 4$.

Si verifica che vale la formula (1.1):

$$\frac{|fl(s_1) - s_1|}{|s_1|} = \frac{4.5 - 4}{4.5} = \frac{1}{9} < \mathbf{u} = 2^{1-3} = \frac{1}{4}$$

Somma esatta $s_2 = .110 \cdot 2^2 + .111 \cdot 2 = 1.0011 \cdot 2^2 =_{10} 4.75$.

Somma floating point $fl(s_2)$.

Allineamento della mantissa del secondo addendo: $0.111 \cdot 2 = .0111 \cdot 2^2$.

Troncamento floating point del secondo addendo: $.0111 \cdot 2^2 \approx .011 \cdot 2^2$.

$$\begin{array}{rcll} .110 & 2^2 & + & \\ .011 & 2^2 & & \\ \hline 1.001 & 2^2 & = .1001 \quad 2^3 & \text{normalizzazione.} \\ & & \approx .100 \quad 2^3 & \text{troncamento fl. point.} \end{array}$$

Si ha quindi che $fl(s_2) = 0.100 \cdot 2^3 =_{10} 0.5 * 8 = 4$.
Si verifica che vale la formula (1.1):

$$\frac{|fl(s_2) - s_2|}{|s_2|} = \frac{4.75 - 4}{4.75} = \frac{3}{19} < \mathbf{u} = 2^{1-3} = \frac{1}{4}$$

Esercizio 1.1.3. $(*)$ **Controesempio per la proprietà associativa.**

- Provare che non vale la proprietà associativa per la somma floating point dei numeri $a = .100 \cdot 2^2$, $b = .100 \cdot 2^{-1}$, $c = .101 \cdot 2^1$.

  Si deve provare che $(a + b) + c \neq a + (b + c)$; si utilizzi, ad esempio, l'aritmetica floating point con $t = 3$ cifre e l'approssimazione tramite troncamento. Nelle somme seguenti vengono eseguiti l'allineamento delle mantisse, la normalizzazione e il troncamento (come nelle somme dell'esercizio 1.1.1), ma non vengono evidenziati.

$$\begin{array}{lcl} (a+b)+c: & \neq & (b+c)+a: \\ \\ \begin{array}{rcl} .100 & \cdot 2^2 & + \\ .000 & \cdot 2^2 & \\ \hline .100 & \cdot 2^2 & + \\ .010 & \cdot 2^2 & \\ \hline .110 & \cdot 2^2 & \end{array} & & \begin{array}{rcl} .001 & \cdot 2^1 & + \\ .101 & \cdot 2^1 & \\ \hline .011 & \cdot 2^2 & + \\ .100 & \cdot 2^2 & \\ \hline .111 & \cdot 2^2 & \end{array} \end{array}$$

Esercizio 1.1.4. $(*)$ **Proprietà della precisione di macchina.**
Verificare le seguenti proprietà della precisione di macchina $u = \beta^{1-t}$, utilizzando l'aritmetica floating point con troncamento, $\beta = 2$ e numero di cifre $t = 4$:

- $fl(1+u) > 1$. Nel caso considerato si ha che $u = \beta^{1-t} =_{10} 2^{-3} =_2 .0010$ e $1 = .1000 \cdot 2^1$. Somma floating point: $fl(.1000 \cdot 2^1 + .0010 \cdot 2^0)$.

  Allineamento della mantissa del secondo addendo: $.0010 \cdot 2^0 = .0001 \cdot 2^1$.

$$\begin{array}{llll} .1000 & \cdot 2^1 & + & \\ .0001 & \cdot 2^1 & & \\ \hline .1001 & \cdot 2^1 & = 1.001 \cdot 2^0 > & 1 \end{array}$$

  Questo risultato non stupisce perché se si somma a 1 un numero positivo non nullo si ottiene un numero maggiore di 1.

- $fl(1 + u/2) = 1$. Somma floating point: $fl(.1000 \cdot 2^1 + .0001 \cdot 2^0)$.

  Allineamento della mantissa del secondo addendo: $.0001 = .00001 \cdot 2^1$.

  Troncamento float. point del secondo addendo: $.00001 \cdot 2^1 \approx .0000 \cdot 2^1 = 0$.

$$\begin{array}{llll} .1000 & \cdot 2 & + & \\ .0000 & \cdot 2^0 & & \\ \hline .1000 & \cdot 2 & = & 1 \end{array}$$

Questo risultato stupisce perché si somma a 1 un numero positivo non nullo e si ottiene 1.

Esercizio 1.1.5. $(*)$ **errori di rappresentazione dei dati.**

- Rappresentare come numeri di macchina i seguenti numeri reali utilizzando l'aritmetica floating point con $\beta = 2$ e numero di cifre $t = 3$. Utilizzare il troncamento per $.1011 \cdot 2^0$ e $.11001 \cdot 2^2$, e l'arrotondamento per $.10101 \cdot 2^0$ e $.1101 \cdot 2^3$.

- Verificare inoltre, per ciascuno dei numeri precedenti, che $\frac{|fl(x)-x|}{|x|} \leq \mathbf{u}$, dove $\mathbf{u}$ è la precisione di macchina.

Troncamento: $\mathbf{u} = \beta^{1-t} = 1/4$.

$$x = .1011 \cdot 2^0 =_{10} 11/16 \qquad fl(x) = .101 \cdot 2^0 =_{10} 5/8$$
$$\frac{|fl(x) - x|}{|x|} = 1/11 \leq 1/4.$$

$$x = .11001 \cdot 2^2 =_{10} 25/8 \qquad fl(x) = .110 \cdot 2^2 =_{10} 3$$
$$\frac{|fl(x) - x|}{|x|} = 1/25 \leq 1/4.$$

Arrotondamento: $\mathbf{u} = .5\beta^{1-t} = 1/8$.

$$x = .10101 \cdot 2^0 =_{10} 21/32 \qquad fl(x) = .101 \cdot 2^0 =_{10} 5/8$$
$$\frac{|fl(x) - x|}{|x|} = 1/21 \leq 1/8.$$

$$x = .1101 \cdot 2^3 =_{10} 13/2 \qquad fl(x) = .111 \cdot 2^3 =_{10} 7$$
$$\frac{|fl(x) - x|}{|x|} = 1/13 \leq 1/8.$$

1.2 Errore inerente ed errore algoritmo

In questo paragrafo vengono descritti gli errori che perturbano il valore di una funzione

$$f : \mathbb{R}^n \to \mathbb{R}$$

Vengono richiamati nel seguito alcune definizioni e alcuni risultati di base.

- Se $\widetilde{z} \in \mathbb{R}$ è una perturbazione di $z \in \mathbb{R}$, si ha che

$$e_z = \widetilde{z} - z \quad \text{viene detto} \quad \text{errore assoluto}$$
$$\epsilon_z = \frac{\widetilde{z} - z}{z} \quad \text{viene detto} \quad \text{errore relativo}$$

- In parole povere l'**errore inerente** è l'errore relativo che perturba il risultato causato dall'errore sui dati. Non dipende dal metodo di calcolo scelto per ottenere il risultato. Più formalmente, dati un valore esatto $x \in \mathbb{R}^n$ e un valore perturbato $\widetilde{x} \in \mathbb{R}^n$, l'errore inerente ϵ_{in} associato a una funzione f è l'errore relativo che perturba il valore $f(\widetilde{x})$ rispetto a $f(x)$. In formule

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

- Nel caso di una funzione in una sola variabile l'errore inerente di può approssimare nel modo seguente, se si trascurano le potenze dell'errore su x superiori a 1.

$$\epsilon_{in} \approx \epsilon_x x \frac{f'(x)}{f(x)}$$

dove ϵ_x è l'errore relativo sul dato x.

- Un problema (o una funzione) si dice **ben condizionato** se l'errore inerente è basso, altrimenti si dice **mal condizionato**.

- L'**errore algoritmico** è l'errore relativo che perturba il valore $f(x)$, quando tale valore viene calcolato in aritmetica floating point utilizzando un numero finito di cifre, partendo da un numero di macchina x, cioè da un valore non perturbato da errore se espresso con t cifre. Se si indica con $fl(f(x))$ il valore calcolato con l'aritmetica floating point, si ha che l'errore algoritmico ϵ_{alg} è dato da

$$\epsilon_{alg} = \frac{fl(f(x)) - f(x)}{f(x)}$$

- L'errore algoritmo dipende dal metodo di calcolo e un algoritmo si dice **stabile** se l'errore algoritmico introdotto non è elevato, altrimenti si dice **instabile**. Differenti algoritmi per calcolare la stessa funzione (o risolvere lo stesso problema) introducono, in generale, differenti errori algoritmici. Per risolvere uno stesso problema alcuni algoritmi possono risultare stabili e altri instabili.

- Per la stima dell'errore algoritmo si sfrutta la relazione (1.1)

$$\frac{|fl(a \diamond b) - (a \diamond b)|}{|a \diamond b|} < \mathbf{u},$$

dove $\diamond$ è una singola operazione aritmetica. Da tale relazione si ha che, posto $\epsilon_1 = \frac{|fl(a \diamond b) - (a \diamond b)|}{|a \diamond b|}$, allora

$$fl(a \diamond b) = (1 + \epsilon_1)(a \diamond b) \qquad \text{con} \qquad |\epsilon_1| < \mathbf{u}$$

- Si ha che l'**errore totale** ϵ_{tot} dovuto sia all'effetto dell'errore sui dati sia all'uso dell'aritmetica floating point è approssimato da

$$\epsilon_{tot} \approx \epsilon_{in} + \epsilon_{alg}$$

dove sono stati trascurati gli errori eletti a potenze maggiori di 1.

- **Importante:** nello studio dell'errore (inerente, algoritmo, totale) si trascurano sempre gli errori elevati a potenze maggiori di 1. Infatti, essendo un errore ϵ in generale molto minore di 1, si ha che ϵ^k è trascurabile rispetto a ϵ, per $k > 1$. Per esempio, se $\epsilon = 0.1$ allora $\epsilon^2 = 0.01$ e $\epsilon^3 = 0.001$.

Esercizio 1.2.1. Errore inerente.
Stimare l'errore inerente introdotto dal calcolo della funzione $f(x) = (x-2)^2$ e dire se il problema è ben condizionato. In caso contrario, dire quali valori di x, se perturbati, possono introdurre elevati errori inerenti.

Sia $\widetilde{x}$ il valore di x soggetto ad una perturbazione, e sia $\epsilon_x = \frac{\widetilde{x}-x}{x}$, cioè $\widetilde{x} = x(1+\epsilon_x)$. L'errore inerente, dovuto alla perturbazione di x, può essere valutato usando due diverse strategie, che portano alla stessa stima dell'errore inerente.
1) **Stima diretta**. Si calcola il valore $\frac{f(\widetilde{x})-f(x)}{f(x)}$, trascurando gli errori elevati a potenza maggiore di 1.

$$f(\widetilde{x}) - f(x) = f(x + x\epsilon_x) - f(x) = (x + x\epsilon_x - 2)^2 - (x-2)^2 = (x-2)^2 + (x\epsilon_x)^2 + 2(x\epsilon_x)(x-2) - (x-2)^2 \approx 2x\epsilon_x(x-2)$$

da cui si ha

$$\frac{f(\widetilde{x}) - f(x)}{f(x)} \approx \frac{2x}{x-2}\epsilon_x.$$

2) **Stima teorica** (da utilizzarsi per funzioni in una variabile). Si stima l'errore inerente mediante la formula

$$\epsilon_{in} \approx \frac{f'(x)x}{f(x)}\epsilon_x,$$

cioè

$$\epsilon_{in} \approx \frac{2(x-2)}{(x-2)^2}x\epsilon_x = \frac{2x}{x-2}\epsilon_x.$$

L'errore inerente può essere elevato, cioè il problema è mal condizionato, se la funzione viene valutata in valori vicini a 2, in quanto in tal caso il coefficiente $\frac{2x}{(x-2)}$ tende all'infinito.

Importante: prima di stimare l'errore è necessario semplificare il coefficiente $\frac{xf'(x)}{f(x)}$ il più possibile per non introdurre "falsi" valori per cui il problema è mal condizionato.

Osservazione. Esprimendo la funzione in altra forma l'errore inerente rimane inalterato. Ad esempio, poiché $f(x) = (x^2+4) - 4x$ si ottiene

$$\epsilon_{in} \approx \frac{f'(x)x}{f(x)}\epsilon_x = \frac{2x-4}{(x-2)^2}x\epsilon_x = \frac{2x}{x-2}\epsilon_x.$$

Esercizio 1.2.2. Errore inerente. Le funzioni $f(x) = x^2 - 2x + 1$ e $g(x) = \dfrac{5x^2+2}{5x+224}$ assumono lo stesso valore su $x = 1.2$. Sapendo che x è perturbato da errore, quale delle due funzioni mèglio usare per non perturbare troppo il risultato?

Per rispondere alla domanda è sufficiente studiare il condizionamento delle due funzioni in un intorno di $x = 1.2$, cioè studiare l'errore inerente associato a f e g. Essendo entrambe funzioni in una variabile è possobile stimare l'errore inerente nel modo seguente:

$$\begin{aligned}
\epsilon_{in}^{(f)} &\approx x\frac{f'(x)}{f(x)}\epsilon_x = x\frac{2x-2}{x^2-2x+1}\epsilon_x = 2x\frac{x-1}{(x-1)^2}\epsilon_x = 2x\frac{1}{(x-1)}\epsilon_x \\
\epsilon_{in}^{(g)} &\approx x\frac{g'(x)}{g(x)}\epsilon_x = x\frac{10x(5x+224)-5(5x^2+2)}{(5x+224)^2}\frac{5x+224}{5x^2+2}\epsilon_x \\
&= \frac{25x^3+2240x^2-10x}{(5x+224)(5x^2+2)}\epsilon_x
\end{aligned}$$

La funzione f è mal condizionata se $x \approx 1$ mentre la funzione g è mal condizionata se $x \approx -224/5 = -44.8$. Poiché x ha un valore vicino a 1 è preferibile utilizzare la funzione g perché è ben condizionata in un intervallo che contiene x.

Esercizio 1.2.3. Errore algoritmico. Stimare l'errore algoritmico introdotto dal calcolo della funzione $f(x) = (x-2)^2$ utilizzando i due algoritmi seguenti:

- $(x-2)^2$;
- $(x^2+4) - 4x$.

e dire se sono stabili.

Osservazione Per ogni operazione $\diamond$ si deve utilizzare la formula $fl(a \diamond b) = ((a \diamond b)(1+\epsilon)$ con $|\epsilon| < \mathbf{u}$, dove $\mathbf{u}$ è la precisione di macchina. Si noti che NON si possono analizzare due operazioni insieme.

Primo algoritmo: $(x-2)^2$.

Innanzitutto è bene osservare che NON vale $fl((x-2)^2) = (x-2)^2(1+\epsilon)$ perché in tale espressione ci sono 2 operazioni, la differenza e il quadrato. Si deve applicare la formula $fl(a \diamond b) = ((a \diamond b)(1+\epsilon)$ analizzando le operazioni nell'ordine in cui vengono eseguite, una alla volta.

Nell'algoritmo in esame la prima operazione è la differenza e quindi, ponendo $\diamond = -$, si ha

$$fl(x-2) = (x-2)(1+\epsilon_1) \qquad \text{con} \qquad |\epsilon_1| < \mathbf{u}$$

È importante osservare che, a questo punto, l'algoritmo non può calcolare il quadrato di $(x-2)$ in quanto tale valore non è stato calcolato. L'algoritmo calcola, quindi, il quadrato di $fl(x-2)$, cioè calcola $fl(x-2) * fl(x-2)$.

Il risultato calcolato è $fl(fl(x-2) * fl(x-2))$ e quindi, posto $\diamond = *$, si ha

$$fl(fl(x-2) * fl(x-2)) = fl(x-2) * fl(x-2)(1+\epsilon_2) \qquad \text{con} \qquad |\epsilon_2| < \mathbf{u}$$

Sostituendo, nella formula precedente l'espressione che coinvolge ϵ_1 calcolata per $fl(x-2)$, si ha

$$fl\left((x-2)^2\right) \;=\; fl(x-2) * fl(x-2)(1+\epsilon_2) = (x-2)^2(1+\epsilon_1)^2(1+\epsilon_2)$$

Trascurando gli errori elevati a una potenza maggiore di 1 o il prodotto di errori, si ottiene

$$fl\left((x-2)^2\right) \approx (x-2)^2(1+2\epsilon_1+\epsilon_2).$$

L'errore algoritmico introdotto è quindi approssimato da

$$\frac{fl\left((x-2)^2\right) - (x-2)^2}{(x-2)^2} \approx 2\epsilon_1 + \epsilon_2,$$

ed in valore assoluto si ottiene

$$\frac{|fl\left((x-2)^2\right) - (x-2)^2|}{|(x-2)^2|} \approx |2\epsilon_1 + \epsilon_2| < 3\mathbf{u}.$$

Dalle stime precedenti si può concludere che l'algoritmo analizzato è stabile.

Secondo algoritmo: $(x^2+4) - 4x$.

Come nel primo caso, è bene osservare che NON vale $fl((x^2+4) - 4x) = ((x^2+4) - 4x)(1+\epsilon)$ perché in tale espressione ci sono 4 operazioni, un quadrato, una somma, un prodotto e una differenza. Si deve applicare la

formula $fl(a \diamond b) = ((a \diamond b)(1+\epsilon)$ analizzando le operazioni nell'ordine in cui vengono eseguite, una alla volta.

Nell'algoritmo in esame la prima operazione è il quadrato e quindi, ponendo $\diamond = *$, si ha

$$fl(x * x) = (x * x)(1+\epsilon_1) \qquad \text{con} \qquad |\epsilon_1| < \mathbf{u}$$

È importante osservare che, a questo punto, l'algoritmo non può calcolare (x^2+4) in quanto tale valore non è stato calcolato. L'algoritmo calcola, quindi, $fl(x^2)+4$ e quindi

$$fl\left(fl(x^2)+4\right) = \left(fl(x^2)+4\right)(1+\epsilon_2) = (x^2(1+\epsilon_1)+4)(1+\epsilon_2) \quad \text{con} \quad |\epsilon_2| < \mathbf{u}$$

e trascurando gli errori elevati a potenza maggiore di 1 o il prodotto di errori si ha

$$fl\left(fl(x^2)+4\right) \approx x^2 + 4 + x^2\epsilon_1 + (x^2+4)\epsilon_2$$

La terza operazione è il calcolo di $4 * x$ e si ottiene

$$fl(4x) = 4x(1+\epsilon_3) \qquad \text{con} \qquad |\epsilon_3| < \mathbf{u}$$

L'ultima operazione da eseguire è la differenza tra $fl(fl(x^2)+4)$ e $fl(4x)$, cioè

$$\begin{aligned} &fl\left(fl(fl(x^2)+4) - fl(4x)\right) = \left(fl(fl(x^2)+4) - fl(4x)\right)(1+\epsilon_4) \\ &\approx \left(x^2+4+x^2\epsilon_1 + (x^2+4)\epsilon_2 - 4x(1+\epsilon_3)\right)(1+\epsilon_4) \\ &\approx x^2+4+x^2\epsilon_1 + (x^2+4)\epsilon_2 - 4x(1+\epsilon_3) + (x^2+4-4x)\epsilon_4. \left(x^2+4+x^2\epsilon_1 + (x^2+4)\epsilon_3 - 4x(1+\right. \end{aligned}$$

con $|\epsilon_4| < \mathbf{u}$. L'errore algoritmico introdotto, a meno di errori di ordine superiore al primo, è quindi dato da

$$\begin{aligned} &\frac{fl\left((x^2+4)-4x\right) - (x^2+4-4x)}{(x^2+4)-4x} \\ &\approx \frac{x^2}{(x^2+4)-4x}\epsilon_1 + \frac{(x^2+4)}{(x^2+4)-4x}\epsilon_2 - \frac{4x}{(x^2+4)-4x}\epsilon_3 + \epsilon_4 \end{aligned}$$

ed in valore assoluto si ottiene

$$\begin{aligned} &\frac{|fl\left((x^2+4)-4x\right) - (x^2+4) + 4x|}{|(x^2+4)-4x|} \\ &< \frac{x^2}{(x-2)^2}\mathbf{u} + \frac{(x^2+4)}{(x-2)^2}\mathbf{u} + \frac{|4x|}{(x-2)^2}\mathbf{u} + \mathbf{u}. \end{aligned}$$

Si noti che al denominatore $(x^2+4)-4x=(x-2)^2$.

Dalle stime precedenti si può concludere che l'algoritmo analizzato non è stabile e si possono avere elevati errori algoritmici nel caso di valori in input vicini a 2.

Si noti che l'algoritmo è stabile se $x<0$. Infatti, in tal caso, si ha che $|4x|=-4x$ e quindi

$$\begin{aligned}&\frac{|fl\left((x-2)^2\right)-(x-2)^2|}{(x-2)^2}\\&<\frac{x^2}{(x-2)^2}\mathbf{u}+\frac{(x^2-4x+4)}{(x-2)^2}\mathbf{u}+\mathbf{u}<3\mathbf{u}.\end{aligned}$$

Esercizio 1.2.4. errore totale. Stimare l'errore totale introdotto dal calcolo della funzione $f(x)=(x-2)^2$, a partire da dati perturbati, utilizzando i due algoritmi seguenti:

- $(x-2)^2$;
- $(x^2+4)-4x$.

Dalle stime presentate negli esercizi 1.2.1 e 1.2.3, risulta che, a meno di termini di ordine superiore al primo, l'errore totale può essere stimato come segue.

Primo algoritmo.

$$\frac{|fl\left((x-2)^2\right)-(x-2)^2|}{|(x-2)^2|}\approx|\epsilon_{in}+\epsilon_{alg}|<\frac{|2x|}{|x-2|}|\epsilon_x|+3\mathbf{u}.$$

Come già sottolineato negli esercizi precedenti, il problema risulta mal condizionato per valori di x vicini a 2, ma l'algoritmo risulta stabile.

Secondo algoritmo.

$$\begin{aligned}&\frac{|fl\left(x^2+4-4x\right)-(x-2)^2|}{|(x-2)^2|}\approx|\epsilon_{in}+\epsilon_{alg}|\\&<\frac{|2x|}{|x-2|}|\epsilon_x|+\frac{2x^2+4+|4x|}{(x-2)^2}\mathbf{u}+\mathbf{u}.\end{aligned}$$

Come già sottolineato negli esercizi precedenti, il problema risulta mal condizionato e l'algoritmo risulta instabile per valori di x vicini a 2.

Esercizio 1.2.5. Data la funzione $f(x,y) = 2y+5xy$, studiare l'errore totale provocato dalle perturbazioni relative ϵ_x e ϵ_y degli argomenti x e y e dall'uso dell'aritmetica floating point con un numero finito di cifre.

L'errore totale è approssimato dalla somma dell'errore inerente e dell'errore algoritmo. È necessario quindi stimare separatamente tali errori.

Errore inerente. Poiché la funzione da analizzare dipende da due variabili, l'errore viene stimato direttamente, a partire dalla definizione.

Siano $\widetilde{x} = x(1+\epsilon_x)$ il valore di x perturbato e $\widetilde{y} = y(1+\epsilon_y)$ il valore di y perturbato. Si ha che

$$f(\widetilde{x},\widetilde{y}) = 2\widetilde{y}+5\widetilde{x}\widetilde{y} = 2y(1+\epsilon_y)+5x(1+\epsilon_x)y(1+\epsilon_y) \approx 2y(1+\epsilon_y)+5xy(1+\epsilon_x+\epsilon_y)$$

da cui segue che

$$\begin{aligned}\epsilon_{in} &= \frac{f(\widetilde{x},\widetilde{y})-f(x,y)}{f(x,y)} \approx \frac{2y(1+\epsilon_y)+5xy(1+\epsilon_x+\epsilon_y)-2y-5xy}{2y+5xy} \\ &= \frac{2y\epsilon_y+5xy(\epsilon_x+\epsilon_y)}{2y+5xy} = \frac{5xy}{2y+5xy}\epsilon_x + \frac{2y+5xy}{2y+5xy}\epsilon_y = \frac{5x}{2+5x}\epsilon_x+\epsilon_y\end{aligned}$$

Il problema è mal condizionato quando il denominatore del coefficiente di ϵ_x tende a zero, cioè quando $x \approx -2/5$. Si noti che, prima di analizzare il condizionamento del problema, i coefficienti di ϵ_x e di ϵ_y sono stati semplificati.

Errore algoritmico. Questo algoritmo richiede 3 prodotti e una somma, che vanno analizzati separatamente. Si ha che

$$\begin{aligned}&fl(2y) = 2y(1+\epsilon_1) \\ &fl(5x) = 5x(1+\epsilon_2) \\ &fl(fl(5x)y) = fl(5x)y(1+\epsilon_3) = 5x(1+\epsilon_2)y(1+\epsilon_3) \approx 5xy(1+\epsilon_2+\epsilon_3) \\ &fl\left(fl(2y)+fl(fl(5x)y)\right) = \left(fl(2y)+fl(fl(5x)y)\right)(1+\epsilon_4) \\ &\qquad \approx \left(2y(1+\epsilon_1)+5xy(1+\epsilon_2+\epsilon_3)\right)(1+\epsilon_4) \\ &\qquad \approx 2y(1+\epsilon_1)+5xy(1+\epsilon_2+\epsilon_3)+(2y+5xy)\epsilon_4\end{aligned}$$

con $|\epsilon_1| < \mathbf{u}$, $|\epsilon_2| < \mathbf{u}$, $|\epsilon_3| < \mathbf{u}$, e $|\epsilon_4| < \mathbf{u}$. Dalla formula precedente si può

ricavare un'approssimazione dell'errore algoritmo

$$\begin{aligned}
\epsilon_{alg} &= \frac{fl(f(x,y)) - f(x,y)}{f(x,y)} \\
&\approx \frac{2y(1+\epsilon_1) + 5xy(1+\epsilon_2+\epsilon_3) + (2y+5xy)\epsilon_4 - 2y - 5xy}{2y+5xy} \\
&= \frac{2y}{2y+5xy}\epsilon_1 + \frac{5xy}{2y+5xy}\epsilon_2 + \frac{5xy}{2y+5xy}\epsilon_3 + \frac{2y+5xy}{2y+5xy}\epsilon_4 \\
&= \frac{2}{2+5x}\epsilon_1 + \frac{5x}{2+5x}\epsilon_2 + \frac{5x}{2+5x}\epsilon_3 + \epsilon_4
\end{aligned}$$

L'algoritmo risulta instabile quando il denominatore dei coefficienti di ϵ_i, $i = 1, 2, 3, 4$, tende a zero, cioè per $x \approx -2/5$.

Errore totale. L'errore totale viene approssimato dalla somma dell'errore inerente e dell'errore algoritmo e quindi

$$\epsilon_{tot} \approx \frac{5x}{2+5x}\epsilon_x + \epsilon_y + \frac{2}{2+5x}\epsilon_1 + \frac{5x}{2+5x}\epsilon_2 + \frac{5x}{2+5x}\epsilon_3 + \epsilon_4$$

Tale errore è elevato per $x \approx -2/5$.

Si noti che, se $x > 0$ allora i denominatori sono maggiori di 2 e quindi l'errore totale non è elevato.

1.3 Esercizi proposti

Esercizio 1.3.1. Siano $\widetilde{x}$ il valore del numero x perturbato da un errore relativo e_x e u la precisione di macchina. Quali delle seguenti relazioni e/o affermazioni sono vere?

1. $e_x = \widetilde{x} - x$
2. $e_x = (\widetilde{x} - x)/x$
3. $\widetilde{x} = x(1 + e_x)$
4. $x = \widetilde{x}e_x$
5. $\widetilde{x} = x(1 - e_x)$
6. $e_x = x/(\widetilde{x} - x)$
7. $|e_x| < u$

8. $\widetilde{x} - x = xe_x$

9. Un problema è ben condizionato quando l'errore sul risultato, causato dall'errore sui dati, è piccolo.

10. Un problema è ben condizionato quando l'errore dovuto all'aritmetica finita è piccolo.

11. Un algoritmo è stabile quando l'errore sul risultato, causato dall'errore sui dati, è piccolo.

12. Un algoritmo è stabile quando l'errore sul risultato, causato dall'aritmetica finita, è piccolo.

13. Un problema è stabile quando l'errore sui dati è piccolo.

14. Se una variazione dei dati del 2 % provoca una variazione dei risultati del 40 %, il problema è ben condizionato.

15. Se un problema è ben condizionato piccole variazioni dei dati provocano piccole variazioni sui risultati.

16. L'errore algoritmico è causato da perturbazioni sui dati.

17. L'errore algoritmico è causato dall'uso dell'aritmetica finita.

18. Un algoritmo è stabile quando l'errore sui dati è piccolo.

Esercizio 1.3.2. Data la funzione $f(x) = x^2 - x + 1$, valutare l'errore inerente provocato dalla perturbazione relativa ϵ_x dell'argomento x e dire per quali valori di x il problema è mal condizionato.

Esercizio 1.3.3. Data una funzione $f(x) = (5 - x)^2 + \sqrt{x}$, $x > 0$, calcolare l'errore inerente introdotto da una perturbazione relativa ϵ_x sul dato x.

Esercizio 1.3.4. Stimare l'errore inerente introdotto dal calcolo delle seguenti funzioni, a partire da dati perturbati:

$$f(x) = x\sqrt{x^2 + 1} \qquad \text{e} \qquad g(x) = \frac{1}{99 - 70x}$$

Dire se sono problemi ben condizionati e, in caso contrario, dire quali sono i valori di x che, se perturbati, possono introdurre elevati errori inerenti.

Esercizio 1.3.5. Sia f una funzione tale che

$$f(x, y) = x^2y + xy^2,$$

stimare l'errore inerente relativo introdotto se x e y vengono perturbati, rispettivamente, dagli errori relativi ϵ_x e ϵ_y. Stimare, inoltre, il valore assoluto dell'errore inerente relativo, sapendo che $|\epsilon_x| < \mathbf{u}$ e $|\epsilon_y| < \mathbf{u}$, $\mathbf{u}$ precisione di macchina, e dire per quali scelte degli argomenti x e y il problema è mal condizionato.

Esercizio 1.3.6. Data una funzione tale che $f(x) = x^2-7x$, calcolare l'errore algoritmico introdotto da tale calcolo.

Esercizio 1.3.7. Data la funzione $f(x) = x^3+1$, valutare l'errore algoritmico introdotto e dire per quali valori di x l'algoritmo risulta instabile.

Esercizio 1.3.8. Data la funzione $f(x) = x^2+6x+9$, calcolare l'errore algoritmico introdotto dall'algoritmo x^2+6x+9 e l'errore algoritmico introdotto dall'algoritmo $(x + 3)^2$. Quale algoritmo è stabile?

Esercizio 1.3.9. Stimare l'errore algoritmico introdotto dalle due seguenti funzioni, equivalenti matematicamente, viste come 2 differenti algoritmi:

$$f(x, y) = x^2 - y^2 \qquad \text{e} \qquad g(x, y) = (x + y)(x - y).$$

Dire se gli algoritmi sono stabili e, in caso contrario, dire per quali valori di x e y si possono ottenere elevati errori algoritmici.

Esercizio 1.3.10. Calcolare l'errore totale introdotto dal calcolo $x^3 + x^2$, a partire da x' valore di x perturbato. Se x è un numero positivo l'errore totale può essere elevato?

Esercizio 1.3.11. $(*)$ Studiare l'errore inerente introdotto dal calcolo della funzione $f(x) = \sin x-\cos x$, in caso di input perturbato da un errore relativo ϵ_x, individuando i casi mal condizionati. Valutare inoltre l'errore algoritmico per calcolare $f(x)$, supponendo che $fl(s \sin x) = \sin x(1 + \epsilon_1)$ e $fl(\cos x) = \cos x(1 + \epsilon_2)$.

Esercizio 1.3.12. Studiare l'errore inerente introdotto dal calcolo della funzione $f(x, y) = x^2 - 4y$, in caso di input perturbato da errore, individuando i casi mal condizionati. Valutare inoltre l'errore algoritmico per calcolare $f(x, y)$.

Soluzioni degli esercizi proposti.

1.3.1 Falso, Vero, Vero, Falso, Falso, Falso, Vero, Vero, Vero, Falso, Falso, Vero, Falso, Falso, Vero, Falso, Vero, Falso.

1.3.2 $\epsilon_{in} = \epsilon_x \frac{x(2x-1)}{x^2-x+1}$, ben condizionato $\forall x$ perché $x^2 - x + 1 \geq 0.75$.

1.3.3 $\epsilon_{in} = \epsilon_x \frac{\sqrt{x}\left(1-4\sqrt{x}(5-x)\right)}{2\left(\sqrt{x}+(5-x)^2\right)}$, sempre ben condizionato.

1.3.4 $\epsilon_{in}^f = \epsilon_x \frac{2x^2+1}{x^2+1}$, $\epsilon_{in}^g = \epsilon_x \frac{70x}{99-70x}$. La funzione f è sempre ben condizionata, la funzione g è mal condizionata per $x \approx 99/70$.

1.3.5 $\epsilon_{in} \approx \frac{2x+y}{x+y}\epsilon_x + \frac{x+2y}{x+y}\epsilon_y$. Il problema è mal condizionato per $x \approx -y$. In valore assoluto $|\epsilon_{in}| \leq \frac{|2x+y|+|x+2y|}{|x+y|}\mathbf{u}$.

1.3.6 $\epsilon_{alg} = \epsilon_1 \frac{x}{x-7} - \epsilon_2 \frac{7}{x-7} + \epsilon_3$. L'algoritmo è instabile per $x \approx 7$.

1.3.7 $\epsilon_{alg} = \frac{x^3}{x^3+1}(\epsilon_1 + \epsilon_2) + \epsilon_3$. L'algoritmo è instabile per $x \approx -1$.

1.3.8 Algoritmo 1: $\epsilon_{alg}^{(1)} = \epsilon_1 \frac{x^2}{x^2+6x+9} + \epsilon_2 \frac{6x}{x^2+6x+9} + \epsilon_3 \frac{x^2+6x}{x^2+6x+9} + \epsilon_4$; instabile se $x \approx -3$. Algoritmo 2: $\epsilon_{alg}^{(2)} = 2\delta_1 + \delta_2$; sempre stabile.

1.3.9 Algoritmo 1: $\epsilon_{alg}^{(1)} = \epsilon_1 \frac{x^2}{x^2-y^2} - \epsilon_2 \frac{y^2}{x^2-y^2} + \epsilon_3$; instbile se $x \approx \pm y$.
Algoritmo 2: $\epsilon_{alg}^{(2)} = \delta_1 + \delta_2 + \delta_3$; sempre stabile.

1.3.10 $\epsilon_{tot} = \epsilon_x \frac{3x+2}{x+1} + \epsilon_1 + \epsilon_3 + \epsilon_2 \frac{x}{x+1}$. Il problema è ma condizionato e l'algoritmo è instabile se $x \approx -1$. Se $x > 0$ l'errore totale non è elevato.

1.3.11 $\epsilon_{in} = x\frac{\cos x+\sin x}{\sin x-\cos x}\epsilon_x = x\frac{1+\tan x}{\tan x-1}\epsilon_x$; mal condizionato per $x \approx \frac{\pi}{4} + k\pi$.
$\epsilon_{alg} = \frac{\sin x}{\sin x-\cos x}\epsilon_1 - \frac{\cos x}{\sin x-\cos x}\epsilon_2 + \epsilon_3 = \frac{\tan x}{\tan x-1}\epsilon_1 - \frac{1}{\tan x-1}\epsilon_2 + \epsilon_3$; instabile per $x \approx \frac{\pi}{4} + k\pi$.

1.3.12 $\epsilon_{in} = \frac{x^2}{x^2-4y}\epsilon_x - \frac{4y}{x^2-4y}\epsilon_y$; il problema è mal condizionato se $x^2 \approx 4y$.
$\epsilon_{alg} = \frac{x^2}{x^2-4y}\epsilon_1 - \frac{4y}{x^2-4y}\epsilon_2 + \epsilon_3$; l'algoritmo è instabile se $x^2 \approx 4y$.

Chapter 2

Equazioni non lineari

In questo capitolo vengono descritti alcuni metodi per approssimare gli zeri di un'equazione non lineare. Data una funzione $f : \mathbb{R} \to \mathbb{R}$, sia $\alpha \in \mathbb{R}$ uno zero di f, cioè $f(\alpha) = 0$; i metodi descritti nel seguito permettono di approssimare α.

Metodo di bisezione. Il metodo di bisezione si può applicare ad ogni funzione continua f tale che $f(a)f(b) < 0$. Tale condizione, per il teorema degli zeri, garantisce che nell'intervallo $[a,\ b]$ è contenuto almeno uno zero di f.

Ad ogni passo del metodo di bisezione l'intervallo considerato viene dimezzato in modo che la funzione f assuma valori discordi sugli estremi del nuovo intervallo. Più in dettaglio, dato l'intervallo $[a,\ b]$ e f continua tale che $f(a)f(b) < 0$,

1. si calcola il punto medio c di $[a,\ b]$;

2. se $f(c) = 0$ l'algoritmo termina e c è lo zero cercato; altrimenti

3. se $f(a) * f(c) < 0$ si pone $b = c$, mentre se $f(a) * f(c) > 0$ (cioè $f(b) * f(c) < 0$) si pone $a = c$;

4. se $b - a < S$, dove S è una soglia prefissata, l'algoritmo termina e c è un'approssimazione dello zero cercato; altrimenti si riparte dal passo 1.

Al passo 4 viene utilizzato un criterio d'arresto che si basa sull'ampiezza dell'intervallo. Il metodo termina quando si ottiene un intervallo minore di una soglia S prefissata. Gli estremi a e b dell'ultimo intervallo calcolato approssimano α con un errore minore di S.

Si può calcolare a priori il minimo numero N di passi al fine di ottenere un intervallo si ampiezza minore di S. Il numero N deve soddisfare la relazione

$$N > log_2 \left(\frac{b-a}{S} \right) \tag{2.1}$$

Generalità sui metodi di iterazione funzionale. I metodi di iterazione funzionale permettono di approssimare uno zero α di una funzione continua f trasformando il problema del calcolo degli zeri in un problema di punto fisso.

- A partire dalla funzione f si costruisce una funzione g definita da

  $$g(x) = x - \frac{f(x)}{h(x)}$$

  dove $h(x)$ è una qualsiasi funzione che non si annulla sull'intervallo $[a,\ b]$ dentro al quale si sta cercando lo zero di f.

- Si ha che α è zero di f se e solo se è punto fisso di g, cioè $f(\alpha) = 0$ se e solo se $g(\alpha) = \alpha$.

- Un metodo di iterazione funzionale è definito dalla scelta della funzione h e, a partire da un valore x_0 costruisce una successione $x_{k+1} = g(x_k)$.

- Si dice che il metodo è convergente se, per ogni $x_0 \in [a,\ b]$ la successione $\{x_k\}$ ha un limite finito per $k \to \infty$.

- Se il metodo è convergente allora la successione $\{x_k\}$ converge ad α zero di f e punto fisso di g per ogni punto iniziale x_0.

- I seguente teorema dimostra quali sono le condizioni sufficienti per la convergenza del metodo.

  Teorema di convergenza locale.
  Siano g una funzione e ρ un numero reale positivo tali che g è continua su $I_\rho = [x^* - \rho, x^* + \rho]$.
  Se $|g'(x)| < 1$ per ogni $x \in I$, allora, per ogni $x_0 \in I$, la successione $\{x_{k+1} = g(x_k)\}$ è contenuta in I_ρ e converge a x^* per $k \to \infty$. Inoltre x^* è l'unico punto fisso di g in I_ρ.

- La velocità di convergenza del metodo è la velocità con cui il metodo approssima la soluzione.

- **Criterio d'arresto.** Fissata una soglia S il metodo si interrompe se $|x_{i+1} - x_i| < S$ e x_{i+1} fornisce un'approssimazione di α. Analogamente, si può scegliere come criterio d'arresto $|f(x_{i+1})| < S$.

Metodo delle corde. Il metodo delle corde è un particolare metodo di iterazione funzionale ottenuto scegliendo $h(x) = m$, cioè è definito dalla relazione

$$x_{i+1} = x_i - \frac{f(x_i)}{m}$$

Condizioni sufficienti per la convergenza del metodo sono le seguenti.
Sia $\rho > 0$ tale che

$$\begin{cases} f \text{ e } f' \text{ continue} & \forall x \in [\alpha - \rho, \alpha + \rho] \\ f'(x) \neq 0 & \forall x \in [\alpha - \rho, \alpha + \rho] \\ m * f'(x) > 0 & \forall x \in [\alpha - \rho, \alpha + \rho] \\ |m| > \frac{1}{2} \max_{x \in [\alpha - \rho, \alpha + \rho]} |f'(x)| \end{cases}$$

allora il metodo è convergente per ogni scelta di $x_0 \in [\alpha - \rho, \alpha + \rho]$.

Metodo delle tangenti o di Newton. Il metodo delle tangenti è un particolare metodo di iterazione funzionale ottenuto scegliendo $h(x) = f'(x)$, cioè definito dalla relazione

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Condizioni sufficienti per la convergenza del metodo sono le seguenti.
Sia $\rho > 0$ tale che

$$\begin{cases} f,\ f' \text{e } f'' \text{ continue} & \forall x \in [\alpha - \rho, \alpha] \\ f'(x) \neq 0 & \forall x \in [\alpha - \rho, \alpha \\ f(x) * f''(x) > 0 & \forall x \in [\alpha - \rho, \alpha] \end{cases}$$

allora il metodo è convergente per ogni scelta di $x_0 \in [\alpha - \rho, \alpha]$.

Vale un teorema analogo in cui si sostituisce l'intervallo $\alpha - \rho, \alpha]$ con $[\alpha, \alpha + \rho]$.

2.1 Metodo di bisezione

Esercizio 2.1.1. metodo di bisezione. Con il metodo di bisezione, se applicabile, approssimare lo zero della funzione non lineare $f(x) = xe^{-x} + e$, nell'intervallo $[-2, 1]$.

Per motivi didattici, in particolare per calcolare l'errore commesso, nel seguito si usa il fatto che $\alpha = -1$ è la soluzione esatta dell'equazione $f(x) = 0$, cioè $f(\alpha) = 0$.

Volendo approssimare il valore di α utilizzando il metodo di bisezione, come prima cosa è necessario verificare l'applicabilità del metodo nell'intervallo $[-2, 1]$. Le condizioni sono verificate perché la funzione f è continua ed inoltre $f(-2) = -2e^2 + e < 0$ mentre $f(1) = e^{-1} + e > 0$.
Nel seguito vengono riportate in dettaglio le prime iterazioni. Siano $a_0 = -2$ e $b_0 = 1$ gli estremi dell'intervallo iniziale la cui ampiezza è data da $b_0 - a_0 = 3$.

Prima iterazione. Viene calcolato il punto medio c_0 dell'intervallo $[a_0,\ b_0]$

$$c_0 = \frac{a_0 + b_0}{2} = -0.5$$

Poiché $f(c_0) = -0.5e^{0.5} + e > 0$, si considera come nuovo intervallo $[a_0,\ c_0]$, e quindi si pone

$$\begin{array}{rcl} a_1 & = & a_0 \\ b_1 & = & c_0 \end{array} \quad = \quad \begin{array}{l} -2 \\ -0.5 \end{array} .$$

Ampiezza nuovo intervallo: $b_1 - a_1 = 1.5$.
La stima dell'errore se si approssima α con l'estremo a_1 o con l'estremo b_1 è data dall'ampiezza di $[a_1,\ b_1]$, e cioè $b_1 - a_1 = 1.5$.
L'errore vero commesso se si sceglie come stima a_1 è dato da $|a_1 - \alpha| = 1$.
L'errore vero commesso se si sceglie come stima b_1 è dato da $|b_1 - \alpha| = 0.5$.
Entrambi gli errori sono minori della stima dell'errore.

Seconda iterazione. Viene calcolato il punto medio c_1 dell'intervallo $[a_1,\ b_1]$

$$c_1 = \frac{a_1 + b_1}{2} = -1.25$$

Poiché $f(c_1) = -1.25e^{1.25} + e < 0$, si considera come nuovo intervallo $[c_1,\ b_1]$, e quindi si pone

$$\begin{array}{rcl} a_2 & = & c_1 \\ b_2 & = & b_1 \end{array} \quad = \quad \begin{array}{l} -1.25 \\ -0.5 \end{array} .$$

Ampiezza nuovo intervallo: $b_2 - a_2 = 0.75$.
La stima dell'errore se si approssima α con l'estremo a_2 o con l'estremo b_2 è data dall'ampiezza di $[a_2,\ b_2]$, e cioè $b_2 - a_2 = 0.75$.
L'errore vero commesso se si sceglie come stima a_2 è dato da $|a_2 - \alpha| = 0.25$.
L'errore vero commesso se si sceglie come stima b_2 è dato da $|b_2 - \alpha| = 0.5$.
Entrambi gli errori sono minori della stima dell'errore.

Terza iterazione. Viene calcolato il punto medio c_2 dell'intervallo $[a_2,\ b_2]$

$$c_2 = \frac{a_2 + b_2}{2} = -0.875$$

Poiché $f(c_2) = -0.875e^{0.875} + e > 0$, si considera come nuovo intervallo $[a_2,\ c_2]$, e quindi si pone

$$\begin{array}{rcccl} a_3 & = & a_2 & = & -1.25 \\ b_3 & = & c_2 & = & -0.875 \end{array} .$$

Ampiezza nuovo intervallo: $b_3 - a_3 = 0.375$.

La stima dell'errore se si approssima α con l'estremo a_3 o con l'estremo b_3 è data dall'ampiezza di $[a_3,\ b_3]$, e cioè $b_3 - a_3 = 0.375$.

L'errore vero commesso se si sceglie come stima a_3 è dato da $|a_3 - \alpha| = 0.25$.
L'errore vero commesso se si sceglie come stima b_3 è dato da $|b_3 - \alpha| = 0.125$.
Entrambi gli errori sono minori della stima dell'errore.

Calcolando in modo analogo ulteriori iterazioni si costruisce una successione di valori c_i, $i = 0, 1, \ldots$, che tende ad α. Calcolando (con MATLAB) 14 iterazioni si ottengono i seguenti valori c_i, $i = 0, \ldots, 13$,

c_0	c_1	c_2	c_3	c_4	c_5	c_6
-0.5000	-1.2500	-0.8750	-1.0625	-0.9688	-1.0156	-0.9922
c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}
-1.0039	-0.9980	-1.0010	-0.9995	-1.0002	-0.9999	-1.0001

la cui rappresentazione su un grafico è riportata in figura 2.1.

Si noti che, ad ogni iterazione, l'ampiezza del nuovo intervallo calcolato si dimezza. La formula (2.1) permette di stimare "a priori" il numero di passi necessari per ottenere un intervallo la cui ampiezza è minore di una soglia prefissata S. Ad esempio, scelta la soglia $S = 0.001$, il minimo numero N di iterazioni per ottenere un intervallo di ampiezza minore di S è dato da

$$N > log_2(\frac{b_0 - a_0}{S}) = log_2(\frac{3}{0.001}) = log_2(3000) \approx 11.5507$$

cioè si deve scegliere $N \geq 12$.

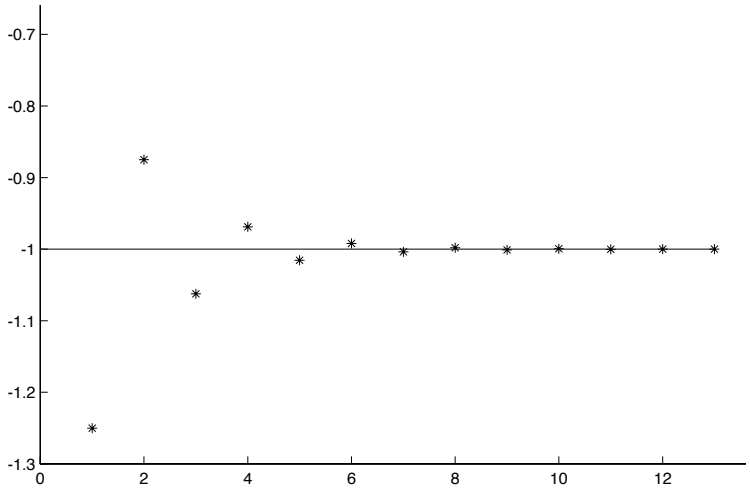

Figure 2.1: Approssimazioni ottenute con il metodo di bisezione.

2.2 Metodi di iterazione funzionale

Esercizio 2.2.1. Un metodo di iterazione funzionale.

Data la funzione $f(x) = x^3 - 4x$, analizzare il comportamento, al variare di $k > 0$, del metodo di iterazione funzionale

$$\begin{cases} x_{i+1} &= g(x_i) \\ g_1(x) &= x - \frac{f(x)}{kx} \end{cases}$$

per approssimare la radice di f appartenente all'intervallo $[1.5, 2.5]$.

I metodi di iterazione funzionale definiti precedentemente sono convergenti se esiste $\rho > 0$ tale che $|g'(x)| < 1$ per ogni $x \in [\alpha - \rho, \alpha + \rho] \subseteq [1.5, 2.5]$, dove g' è la derivata prima di g. Poiché non si conosce α, si studia il comportamento di $|g'(x)|$ su tutto l'intervallo $[1.5, 2.5]$. Poiché

$$g(x) = x - \frac{x^3 - 4x}{kx} = x - \frac{x^2 - 4}{k},$$

si ha

$$g'(x) = 1 - \frac{2x}{k},$$

da cui si ottiene

$$|g'(x)| < 1 \quad \Longleftrightarrow -1 < 1 - \frac{2x}{k} < 1 \quad \forall x \in [1.5, 2.5].$$

Essendo $k > 0$ e $x \in [1.5, 2.5]$, la condizione $1 - \frac{2x}{k} < 1$, equivalente a $-\frac{2x}{k} < 0$, è verificata per ogni k. Si consideri la relazione $-1 < 1 - \frac{2x}{k}$, che corrisponde a $k > x$. Tale relazione è verificata per ogni $x \in [1.5, 2.5]$ se si sceglie $k > 1.5$.

In conclusione, scegliendo $k > 1.5$ il metodo che si ottiene è convergente a partire da ogni x_0 appartenente a un intervallo $[\alpha - \rho, \alpha + \rho]$ contenuto in $[1.5,\ 2.5]$. Poiché non si conosce il valore di α, non si riesce a conoscere l'intervallo $[\alpha - \rho, \alpha + \rho]$ in cui scegliere x_0. Poiché però tale intervallo contiene l'estremo sinistro 1.5 o l'estremo destro 2.5 dell'intervallo in cui si cerca α, si sceglie come x_0 uno dei due estremi e, se per qualche indice i il valore $x_i \notin [1.5, 2.5]$, allora si sceglie come valore iniziale l'altro estremo dell'intervallo.

Vengono analizzati tre casi: $k = 4$ e $k = 8$, che generano metodi convergenti e $k = 1$ associato a un metodo che non converge.
Con i$k = 4$ si sceglie $x_0 = 1.5$ e si ottengono

$$\begin{aligned} x_1 &= x_0 - \frac{x_0^2}{4} + 1 = \frac{31}{16} \approx 1.9375 \\ x_2 &= x_1 - \frac{x_1^2}{4} + 1 = \frac{2047}{1024} \approx 1.9990 \\ x_3 &= x_2 - \frac{x_2^2}{4} + 1 = \frac{8388607}{4194304} \approx 1.99999998 \end{aligned}$$

cioè la successione sta convergendo a 2.
Con $k = 8$ si sceglie $x_0 = 1.5$ e si ottiene

$$x_1 = x_0 - \frac{x_0^2}{8} + \frac{1}{2} = \frac{55}{32} \approx 1.71878$$

e, proseguendo in modo analogo, si ha

x_2	x_3	x_4	x_5	x_6	x_7
1.8495	1.9219	1.9602	1.9799	1.9899	1.9949
x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1.9975	1.9987	1.9994	1.9997	1.9998	1.9999

cioè, anche in questo caso, la successione sta convergendo a 2.

Sebbene entrambi i metodi siano convergenti alla radice $\alpha = 2$, il primo è più veloce rispetto al secondo, cioè sono necessari meno passi per ottenere la stessa precisione. Infatti con $k = 4$ sono sufficienti 3 passi per approssimare $\alpha = 2$ con il numero $x_3 = 1.99999998$, mentre con $k = 8$ dopo 12 passi α è approssimata con 1.9998. In figura 2.2 vengono riportate le prime 10 iterazioni di entrambi i metodi per sottolinearne il diverso comportamento.

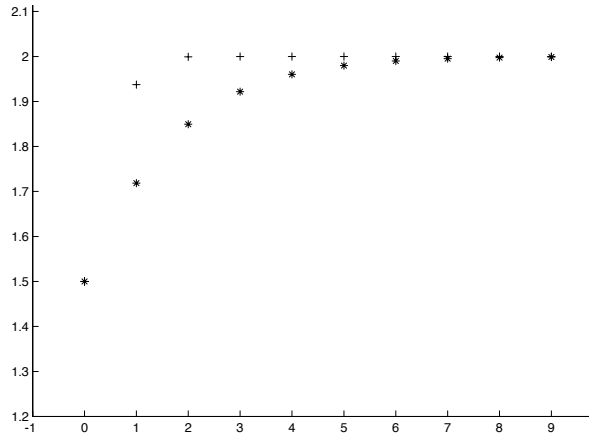

Figure 2.2: I metodi di iterazione funzionale con '*' per il caso $k = 8$ e '+' per il caso $k = 4$.

Se si sceglie $k = 1$ il metodo non converge. Infatti si ottengono le seguenti successioni:

x_0	x_1	x_2	x_3	x_4	x_5	x_6
1.5	3.2500	-3.3125	-10.2852	-112.0696	$-1.2668 * 10^4$	$-1.6048 * 10^8$
2.5	0.2500	4.1875	-9.3477	-92.7263	$-8.6869 * 10^3$	$-7.5471 * 10^7$

Esercizio 2.2.2. Il metodo delle corde.
Siano $f(x) = \log(1 + x^2) + 3x$ e $I = [-1,\ 2]$.

- Calcolare un valore del parametro m che garantisca la convergenza del metodo delle corde a partire da un estremo di I.

  Le condizioni sufficienti per la convergenza del metodo sono le seguenti:

  - $f'(x) \neq 0$, in $[\alpha - \rho, \alpha + \rho] \subseteq [-1, 2]$;
  - $m * f'(x) > 0$, in $[\alpha - \rho, \alpha + \rho] \subseteq [-1, 2]$;
  - $m > 0.5 \max_{[\alpha-\rho,\alpha+\rho]} f'(x)$, e quindi non conoscendo in genrale α è sufficiente imporre $|m| > 0.5 \max_{[-1,2]} |f'(x)|$.

  Nel caso in esame si ha che

$$f'(x) = \frac{3x^2 + 2x + 3}{1 + x^2},$$

di cui si deve studiare il segno. Il denominatore è sempre positivo e quindi è sufficiente studiare il segno di $3x^2+2x+3$. Poiché il discriminante di $3x^2+2x+3=0$ è negativo, si ha che $3x^2+2x+3$ non cambia segno e, poiché in $x=0$ assume valore positivo, allora $3x^2+2x+3>0$ per ogni $x \in \mathbb{R}$. Ne segue che si deve scegliere $m>0$.

Il valore di $|m|$ è legato al massimo di f' sull'intervallo $[-1,2]$. Per calcolare il massimo di una funzione è necessario calcolare gli zeri della sua derivata prima. In questo caso la derivata prima di f' è la derivata seconda di f, cioè

$$f''(x) = \frac{2(1-x^2)}{(1+x^2)^2},$$

e quindi $f''(x) \geq 0$ per $-1 \leq x \leq 1$, cioè la funzione f' cresce in [-1,1], ha massimo per $x=1$ e decresce in $[1,\ 2]$. Allora i valori di m che garantiscono la convergenza del metodo sono tali che

$$m > \frac{1}{2} f'(1) = 2.$$

- Approssimare lo zero di f nell'intervallo I mediante il metodo delle corde con il valore di m precedentemente calcolato. Valutare l'errore commesso, sapendo che la soluzione esatta dell'equazione $f(x)=0$ è $\alpha = 0$.

  In base all'analisi precedente si sceglie $m=2.5$. Il metodo delle corde costruisce una successione di valori x_i, fissato x_0, nel modo seguente:

$$\begin{aligned} x_{i+1} &= x_i - \frac{f(x_i)}{m}, \quad \text{e} \quad \text{quindi} \\ x_{i+1} &= x_i - \frac{\log(1+x_i^2)+3x_i}{2.5}. \end{aligned}$$

  Se si scelgono, ad esempio, $x_0 = 0.5$ come punto iniziale, $S = 10^{-3}$ come valore soglia e $|x_{i+1} - x_i| < S$ come criterio d'arresto, mediante il metodo delle corde con $m=2.5$ si costruisce la seguente successione:

x_0	x_1	x_2	x_3	x_4
0.500	$0.107 * 10^{-1}$	$0.210 * 10^{-2}$	$0.418 * 10^{-3}$	$0.836 * 10^{-4}$

  che converge ad $\alpha = 0$.

  Ponendo come condizione d'arresto $|x_{i+1} - x_i| < 10^{-3}$ si ottiene $x_4 = 0.836 * 10{-}4$ e quindi l'errore di approssimazione è dato da $|x_4 - 0| = 0.836 * 10^{-4}$; in questo caso l'errore è inferiore alla soglia S utilizzata nel criterio d'arresto.

- Studiare il comportamento dell'algoritmo se si scelgono valori di m che non soddisfano le condizioni sufficienti per la convergenza.

  Scegliendo un parametro m che non soddisfa le condizioni sufficienti per la convergenza non si può concludere nulla riguardo la convergenza del metodo. Infatti, la scelta $m = 2$ genera una successione convergente e la scelta $m = 1$ genera una successione non convergente.

  Caso $m = 2$.

x_0	x_1	x_2	x_3	x_4	x_5
0.5	-0.3243	0.1288	-0.0699	0.0333	-0.0173
x_6	x_7	x_8	x_9	x_{10}	x_{11}
0.0084	-0.0042	0.0021	-0.0010	0.0005	-0.0002

  Caso $m = 1$.

x_0	x_1	x_2	x_3	x_4
0.5	-1.0743	1.8929	-4.2933	7.5977
x_5	x_6	x_7	x_8	x_9
-16.5530	31.2338	-64.7623	126.7441	-256.7163

Esercizio 2.2.3. Il metodo delle tangenti. Approssimare, con il metodo delle tangenti, la radice α della funzione $f(x) = 0.5x - e^{x-2}$, nell'intervallo $I = [1.5, 3]$, scegliendo il valore iniziale x_0 in modo tale da garantire la convergenza del metodo.

Per scegliere il valore iniziale x_0 che garantisca la convergenza è necessario studiare il comportamento di $f'(x)$ e di f'' e applicare il teorema sulla convergenza. Innanzitutto si ha che $f \in C^\infty(I)$ e $f'(x) \neq 0$ su I. Inoltre, si ha che $f'(x) = 0.5 - e^{x-2}$ e, poiché $e^{x-2} > e > 0.5$ per $x \in I$, allora $f'(x) < 0$ su I. Infine $f''(x) = -e^{x-2} < 0$ perché l'esponenziale è una funzione sempre positiva.

Poiché su I la funzione f' è negativa, allora su I la funzione f è decrescente e quindi, essendo $f(\alpha) = 0$, si ha che $f(x) < 0$ su $(\alpha, 3]$. si conclude che su $(\alpha, 3]$ $f(x)f''(x) > 0$.

Poiché le condizioni sufficienti per la convergenza sono soddisfatte su $(\alpha, 3]$, si sceglie $x_0 \in (\alpha, 3]$ e, poiché non è nota la posizione di α, si sceglie $x_0 = 3$, poiché è l'unico punto di cui si è certi che appartenga a $(\alpha, 3]$.

Il metodo delle tangenti costruisce, a partire da x_0, una successione di valori nel modo seguente:

$$\begin{aligned} x_{i+1} &= x_i - \frac{f(x_i)}{f'(x_i)}, \quad \text{e} \quad \text{quindi} \\ x_{i+1} &= \frac{e^{x_i-2}(1-x_i)}{0.5 - e^{x_i-2}}. \end{aligned}$$

Scegliendo $x_0 = 3$, le prime 5 iterate del metodo delle tangenti sono le seguenti:

x_0	x_1	x_2	x_3	x_4	x_5
3	2.4508	2.1290	2.0142	2.0002	2.0000

Il metodo sta convergendo a 2, soluzione di $f(x) = 0$. In figura 2.3 è riportato il grafico delle prime 10 iterate.

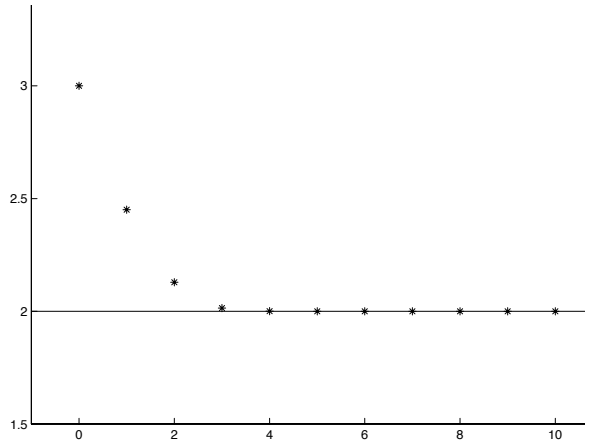

Figure 2.3: metodo delle tangenti.

Esercizio 2.2.4. Confronto tra i metodi. Approssimare, con il metodo delle tangenti e il metodo delle corde, la radice α della funzione non lineare $f(x) = x^6 - 64$ nell'intervallo $I = [-1, 9]$, scegliendo il parametro m del metodo delle corde e il valore iniziale x_0 in modo tale da garantire la convergenza dei metodi.

Per poter utilizzare le condizioni sufficienti per la convergenza dei metodi è necessario che $f'(x)$ non si annulli su tutto l'intervallo in esame. In questo

caso $f'(x) = 6x^5$ e si annulla in $x = 0 \in [-1, 9]$. È quindi opportuno applicare, a partire da I, alcuni passi del metodo di bisezione, per ridurre l'ampiezza dell'intervallo.

Si noti che il metodo di bisezione è applicabile in quanto f è continua su I e $f(-1) < 0$ e $f(9) > 0$.
Siano $a_0 = -1$ e $b_0 = 9$; essendo $c_0 = \frac{a_0+b_0}{2} = 4$ e $f(4) > 0$, la radice α appartiene all'intervallo $[-1, 4]$.
Posti $a_1 = a_0 = -1$ e $b_1 = 4$, un'ulteriore iterazione del metodo di bisezione permette di ottenere $c_1 = \frac{a_+b_1}{2} = 1.5$, e poiché $f(1.5) < 0$ si può considerare $J = [1.5, 4]$ come intervallo di partenza su cui applicare i metodi delle corde e delle tangenti.

Metodo delle corde.

Essendo $f'(x) = 6x^5$ una funzione positiva e crescente sull'intervallo J, per ottenere un algoritmo convergente qualunque sia il punto $x_0 \in [\alpha - \rho, \alpha + \rho]$, è sufficiente scegliere il parametro $m > 0$ in modo tale che $m > 0.5 * \max_J f'(x) = 0.5 * 6 * (4)^5 = 3072$ (vedi esercizio 2.2.2). Si sceglie $m = 3073$. Poiché non si conosce la posizione di α, si sceglie $x_0 = 4$ e, se la successione non converge, allora si sceglie $x_0 = 1.5$.

Metodo delle tangenti.

Essendo $f'(x) = 6 * x^5$ positiva sull'intervallo J, la funzione funzione f è crescente su J e quindi $f(x) > 0 \ \forall x > \alpha$. Inoltre, poiché $f''(x) = 30x^4 > 0$ $\forall x$, si ha che le condizioni sufficienti per la convergenza del metodo delle tangenti sono verificate sull'intervallo $[\alpha, 4]$ e si può scegliere come punto iniziale $x_0 = 4$.

Le prime 40 iterazioni dei due metodi (calcolate con Matlab) sono rappresentate in figura 2.4, ove sono evidenziate le differenti velocità di convergenza: il metodo delle code ('*') converge più lentamente alla soluzione rispetto al metodo delle tangenti ('+'), cioè richiede più iterazioni per avvicinarsi alla soluzione.

Si noti che dopo circa 7 iterazioni il metodo delle tangenti approssima $\alpha = 2$ con un errore minore di 10^{-4}, mentre la stessa precisione è ottenuta con più di 40 passi del metodo delle corde.

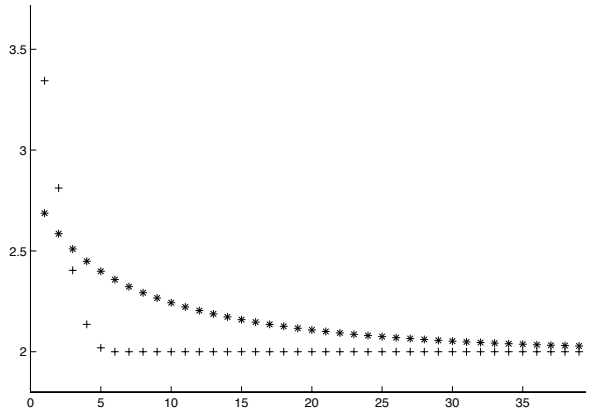

Figure 2.4: metodo delle corde e delle tangenti.

2.3 Esercizi proposti

Esercizio 2.3.1. Sia α lo zero di una funzione continua f sull'intervallo $[a, b]$. Dire quali delle seguenti affermazioni sono vere:

1. La funzione in α è nulla
2. α è un punto di $[a, b]$
3. $f(a) * f(b) < 0$
4. $\alpha = a$ oppure $\alpha = b$
5. $|f(\alpha)| < S$ dove $S > 0$ è una soglia prefissata.
6. $\alpha = 0$
7. $f(\alpha) = 0$
8. $\alpha = \frac{a+b}{2}$

Esercizio 2.3.2. Sia f una funzione continua su $[a,\ b]$. Il metodo di bisezione si può applicare se

1. $f(a) > 0$ e $f(b) < 0$
2. $f(a) * f(b) < 0$
3. $f(a)$ e $f(b)$ sono concordi
4. $f(a) > 0$

Esercizio 2.3.3. Data una soglia S, se a_k e b_k sono gli estremi dell'intervallo al k-esimo passo del metodo di bisezione, si può scegliere come criterio d'arresto

1. $|a_k - a| < S$
2. $|a_k - b_k| < S$
3. $\frac{|b-a|}{2^k} < S$
4. $|a_k + b_k|/2 < S$

Esercizio 2.3.4. Dati la funzione $f(x) = (x + 2)(x^2 + 3)$ e gli intervalli $I = [3,\ 5]$ e $J = [-5,\ 0]$, dire se il metodo di bisezione è applicabile a I e/o J e perché. Inoltre, calcolare 3 passi del metodo di bisezione a partire dall'intervallo o dagli intervalli su cui il metodo è applicabile.

Esercizio 2.3.5. Dati la funzione $f(x) = \frac{x^3+3}{x^2+1}$, l'intervallo $I = [-2,\ 0]$ e la soglia $S = 0.1$, calcolare a priori il numero N di passi del metodo di bisezione affinché l'ampiezza del nuovo intervallo sia minore di S. Inoltre, calcolare N passi del metodo di bisezione.

Esercizio 2.3.6. Dati la funzione $f(x) = x^4 - 3x^3 - 2x^2$ e il metodo di iterazione funzionale

$$x_{i+1} = g(x_i) \quad \text{con} \quad g(x) = x - \frac{2}{5}\frac{f(x)}{x^2}$$

trovare un intervallo I su cui $|g'(x)| < 1$. Applicare due passi del metodo a partire da entrambi gli estremi di tale intervallo.

Esercizio 2.3.7. $(*)$ Sia f una funzione tale che $f(x) = x^2 + x - 6$ con $x \in I = [-5, -1]$. Dato il metodo di iterazione funzionale

$$x_{i+1} = g(x_i) \quad \text{con} \quad g(x) = x - \frac{f(x)}{p(x-2)}$$

per approssimare la radice α di f appartenente a I, trovare i valori del parametro p per i quali sono soddisfatte le condizioni sufficienti per la convergenza di tale metodo sull'intervallo I. Con p che garantisce la convergenza, applicare 2 passi del metodo a partire da $x_0 = -1$ per stimare α.

Esercizio 2.3.8. Sia α lo zero di una funzione f sull'intervallo $[a, b]$. Quale delle seguenti condizioni devono essere verificate per garantire la convergenza del metodo delle corde?

1. $f(x) * f''(x) > 0$, $m = 0.5 * max|f'(x)|$ su $[\alpha - \rho, \alpha + \rho]$
2. $f'(x) \neq 0$, $mf'(x) > 0$ $|m| > 0.5 * max|f'(x)|$ su $[\alpha - \rho, \alpha + \rho]$
3. $f'(x) \neq 0$, $mf'(x) > 0$ $m > 0.5 * maxf'(x)$ su $[\alpha - \rho, \alpha + \rho]$
4. $|1 - f'(x)/m| < 1$ su $[\alpha - \rho, \alpha + \rho]$
5. $f(x) * f'(x) > 0$ su $[\alpha - \rho, \alpha + \rho]$

Esercizio 2.3.9. $(*)$ Data la funzione $f(x) = 3e^{-2x}x^2 - 2$ dire (giustificando la risposta) quali dei seguenti valori di m soddisfano le condizioni sufficienti di convergenza del metodo delle corde sull'intervallo $[-2, 0)$:

$$m = -13,\ \ m = 13,\ \ m = -984,\ \ m = 984,\ \ m = -1123,\ \ m = 1123$$

Esercizio 2.3.10. Sia $f(x) = x^3 - 3x^2 + 2$. Scegliere tra i valori $\{-2, -1, 0, 1, 2\}$ il valore del parametro m che garantisca la convergenza del metodo delle corde per approssimare la radice di f sull'intervallo $[0.5,\ 1.5]$. Calcolare 3 iterate del metodo delle corde con tale valore di m, a partire da $x_0 = 0$, per approssimare la radice di f.

Esercizio 2.3.11. Applicare 4 passi del metodo delle corde per approssimare la radice della funzione $f(x) = x^3 + x^2 - 2$, nell'intervallo $[.5, 1.5]$, partendo da $x_0 = .5$ e scegliendo come valore della costante m il minimo intero che garantisce la convergenza del metodo. A quale valore sembra convergere la successione?

Esercizio 2.3.12. Data una funzione continua f sull'intervallo $[a, b]$ dire se:

1. Per applicare il metodo delle tangenti è necessario calcolare l'equazione della tangente.
2. Il metodo delle tangenti richiede il calcolo della derivata di f.
3. Il metodo delle tangenti impiega (a parità di soglia) meno iterazioni, in generale, del metodo di bisezione.
4. Converge qualunque sia il primo punto scelto.
5. Il metodo delle tangenti serve per calcolare l'equazione della retta tangente.
6. Converge solo se $f(a) * f(b) < 0$.
7. Il metodo delle tangenti permette di calcolare una radice di f.

Esercizio 2.3.13. Sia $f(x) = 3x^4 + x^2 - 2$ e sia α l'unica radice di f in $[-1,\ -0.5]$. Scegliendo $x_0 = -1$, approssimare α mediante alcune iterazioni del metodo delle tangenti.

Esercizio 2.3.14. Approssimare, con il metodo delle tangenti, la radice α della funzione $f(x) = \log_e(x) - x^2 + 1$, nell'intervallo $[0.75, 3]$, scegliendo il valore iniziale x_0 in modo tale da garantire la convergenza del metodo.

Esercizio 2.3.15. Approssimare con il metodo delle tangenti lo zero della funzione $f(x) = e^x(2x - 1)$ nell'intervallo $[0,\ 3]$, scegliendo il punto iniziale $x_0 = 3$. Giustificare perché con tale scelta il metodo converge.

Esercizio 2.3.16. $(*)$ Sia $f(x) = \sqrt{x}(2x^2 - 1)$, che ammette una radice nell'intervallo $I = [0.5, 3.5]$. Mediante il metodo di bisezione trovare un sotto intervallo J di I, contenente la radice, di ampiezza minore di 0.8. Applicare il metodo delle tangenti, scegliendo un punto $x_0 \in J$ che garantisca la convergenza del metodo.

Esercizio 2.3.17. Sia data la funzione $f(x) = x^3 - 3x^2 + 2x$, applicare un passo del metodo di bisezione, se possibile, a partire da $I = [-1, 0.5]$ o da $J = [-2, -1]$. Sia K l'intervallo così ottenuto. Scelto $x_0 \in K$ per il quale il metodo delle tangenti converge, calcolare 2 iterate di tale metodo.

Esercizio 2.3.18. $(*)$ Data la funzione $f(x) = (1-x^2)(x^2-4) = -x^4+5x^2-4$ sull'intervallo $I = [0.5, 1.5]$:

a) calcolare il parametro m affinché il metodo delle corde converga su I;

b) applicare 3 passi del metodo delle corde a partire da $x_0 = 0.5$ utilizzando i minimo intero$m$ che garantisca la convergenza;

c) applicare 3 passi del metodo delle tangenti a partire da $x_0 = 0.5$;

d) sapendo che la soluzione approssimata è $\alpha = 1$ stimare l'errore commesso con il metodo delle corde e delle tangenti.

Esercizio 2.3.19. $(*)$ Confrontare il diverso comportamento del metodo delle corde e del metodo delle tangenti applicati alla funzione $f(x) = 3x^3 - 9x^2 + x - 3$, nell'intervallo $[2, 5]$.

Soluzioni degli esercizi proposti.

2.3.1 Vero, Vero, Falso, Falso, Vero, Falso, Vero, Falso.

2.3.2 Vero, Vero, Falso, Falso.

2.3.3 Falso, Vero, Vero, Falso.

2.3.4 Si può solo partire da J e si ottiene $[-5/2,\ -15/8]$.

2.3.5 $N = 5$, e gli intervalli calcolati sono $[-2,\ -1]$, $[-3/2,\ -1]$, $[-3/2,\ -5/4]$, $[-3/2,\ -11/8]$, $[-3/2,\ -23/16]$.

2.3.6 $I = [1.5,\ 4]$. Sia con $x_0 = 1.5$ che con $x_0 = 4$ si ottengono $x_1 = 3.2$ e $x_2 = 3.7440$.

2.3.7 $(*)$ $p > 0.5$. Con $p = 1$, per ogni $i > 0$, $x_i = -3$ che è la soluzione cercata, qualunque sia x_0. Per $p = 2$ si ha $x_0 = -1$, $x_1 = -5/2$ e $x_2 = -11/4$.

2.3.8 Falso, Vero, Falso, Vero, Falso.

2.3.9 $(*)$ $m = -984$; $m = -1123$.

2.3.10 $m = -2$; $x_0 = 0$, $x_1 = 1$, $x_2 = 1$...

2.3.11 $m = 5$; $x_0 = 0.5$, $x_1 \simeq 0.8250$, $x_2 \simeq 0.9766$, $x_3 \simeq 0.9996$, $x_4 = 1.0000$. La successione tende a 1.

2.3.12 Falso, Vero, Vero, Falso, Falso, Falso, Vero.

2.3.13 $x_1 \simeq -0.8571$, $x_2 \simeq -0.8190$, $x_3 \simeq -0.8165$, $x_4 \simeq -0.8165$.

2.3.14 $x_0 = 3$, $x_1 \simeq 1.7821$, $x_2 \simeq 1.25$, $x_3 \simeq 1.0504$, $x_4 = 1.0032$.

2.3.15 $x_0 = 3$, $x_1 \simeq 2.2857$, $x_2 \simeq 1.6447$, $x_3 \simeq 1.0504$, $x_4 = 0.7317$.

2.3.16 $(*)$ $I = [0.5,\ 1.25]$. $x_0 = 1.25$, $x_1 = 0.8250$, $x_2 \simeq 0.7155$, $x_3 \simeq 0.7072$, $x_a \simeq 0.7071 \simeq 1/\sqrt{2}$.

2.3.17 Bisezione si può applicare solo a partire da I. $K = [-0.25,\ 0.5]$. Il metodo delle tangenti converge si $[-0.25,\ \alpha]$ e quindi si sceglie $x_0 = -0.25$, da cui si ha $x_1 = -0.0385$, $x_2 = -0.0011$, $x_3 = -8.6549 * 10^{-7}$, $x_4 = -5.6181 * 10^{-13}$. La successione tende a 0.

2.3.18 $(*)$ a) $m > 3.0429$. b) Scelto $m = 4$ si ha $x_1 = 1.2031$, $x_2 = 0.9176$, $x_3 = 1.0424$. c) $x_1 = 1.125$, $x_2 = 0.9942$, $x_3 = 1.0000$. d) Errore relativo metodo delle corde $\simeq 0.0424$. Errore relativo metodo delle tangenti $< 10^{-4}$.

2.3.19 $(*)$ Il metodo delle corde converge per $m > 68$. Scelto $m = 70$ si ha $x_0 = 5$, $x_1 = 2.8286$, $x_2 = 2.8898$, $x_3 = 2.9308 \ldots x_{18} = 3.0000$. Il metodo delle tangenti converge su $[\alpha, 5]$ e si ha $x_0 = 5$, $x_1 = 3.8824$, $x_2 = 3.2716$, $x_3 = 3.0377$, $x_4 = 3.0009$, $x_5 = 3.0000$.

Chapter 3

Interpolazione polinomiale

Dato un insieme di $(N+1)$ punti $P_0 = (x_0,\ y_0)$, $P_1 = (x_1,\ y_1) \dots P_N = (x_N,\ y_N)$, si vuole calcolare il **polinomio interpolatore**, cioè un polinomio p di grado minore o uguale a N tale che $p(x_i) = y_i$, $i = 0, \dots, N$. I valori x_i, $i = 0 \dots, N$ sono detti nodi. Valgono le seguenti proprietà.

- Il polinomio interpolatore esiste sempre ed è **unico**.
- Esistono diversi modi per calcolare il polinomio interpolatore. Poiché tale polinomio è unico tutti i metodi permettono di calcolare lo stesso polinomio. In queste dispense vengono utilizzati il metodo di **Lagrange** e la matrice di **Vandermonde**.
- Metodo di **Lagrange**. Per ogni $i = 0, \dots, N$ si definiscono le funzioni (sono polinomi di grado N)

  $$L_i(x) = \prod_{j=0, j\neq i}^{N} \frac{x - x_j}{x_i - x_j} = \frac{x - x_0}{x_i - x_0} \cdots \frac{x - x_{i-1}}{x_i - x_{i-1}} \frac{x - x_{i+1}}{x_i - x_{i+1}} \cdots \frac{x - x_N}{x_i - x_N}$$

  e si calcola il polinomio interpolatore $p(x)$ nel modo seguente:

  $$p(x) = \sum_{i=0}^{N} y_i L_i(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_N L_N(x)$$

- Metodo con la matrice di **Vandermonde**. Si costruiscono la matrice $N+1 \times (N+1)$ di Vandermonde V e il vettore $N \times 1$ y nel modo seguente

  $$V = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^N \\ 1 & x_1 & x_1^2 & \cdots & x_1^N \\ \vdots & & & & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^N \end{pmatrix} \qquad \text{e} \qquad y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Se il polinomio interpolatore viene scritto nella forma

$$p(x) = a_N x^N + a_{N-1}x^{N-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

allora il vettore dei coefficienti $a = [a_0,\ a_1\ \ldots,\ a_N]^t$ è la soluzione del sistema $Va = y$. Purtroppo, in generale, tale sistema è mal condizionato.

- Il polinomio interpolatore serve per approssimare una funzione f non nota di cui si conoscono i valori y_i, $i = 0, \ldots, N$ sui nodi. La funzione f e il polinomio p coincidono sui nodi. La differenza $r(x) = f(x) - p(x)$ per x appartenente a un intervallo $[a,\ b]$ che contiene i nodi viene detta **errore** o **resto** dell'interpolazione. Si ha che

$$r(x) = \frac{f^{(N+1)}(z)}{(N+1)!}(x - x_0)(x - x_1)\cdots(x - x_N)$$

dove $z \in [x_0,\ x_N]$ è un valore non noto e $f^{(N+1)}$ è la derivata $(N+1)$-esima di f. Se $\max_{[x_0,x_N]} |f^{(N+1)}(z)| < M$ allora

$$|r(x)| < \frac{M}{(N+1)!}|(x - x_0)(x - x_1)\cdots(x - x_N)|$$

3.1 Calcolo del polinomio interpolatore

Esercizio 3.1.1. Forma di Lagrange e di Vandermonde. Si calcoli il polinomio che interpola i punti $P_0 = (-2,\ 0)$, $P_1 = (-1,\ 2)$, $P_2 = (0,\ -2)$, $P_3 = (1, -6)$.

Poiché $N = 3$ si deve ottenere un polinomio di grado minore o uguale a 3.

Forma di Lagrange. Il polinomio è della forma

$$p(x) = y_0 L_0 + y_1 L_1 + y_2 L_2 + y_3 L_3 = 0 * L_0 + 2L_1 - 2L_2 - 6L_3 = 2L_1 - 2L_2 - 6L_3$$

e quindi si devono calcolare solo L_1, L_2 e L_3. Si ha che

$$\begin{aligned} L_1 &= \frac{(x-x_0)}{(x_1-x_0)}\frac{(x-x_2)}{(x_1-x_2)}\frac{(x-x_3)}{(x_1-x_3)} = \frac{(x+2)x(x-1)}{(-1+2)(-1)(-1-1)} \\ &= \frac{1}{2}(x^3+x^2-2x) \\ L_2 &= \frac{(x-x_0)}{(x_2-x_0)}\frac{(x-x_1)}{(x_2-x_1)}\frac{(x-x_3)}{(x_2-x_3)} = \frac{(x+2)(x+1)(x-1)}{(2)(1)(-1)} \\ &= -\frac{1}{2}(x^3+2x^2-x-2) \\ L_3 &= \frac{(x-x_0)}{(x_3-x_0)}\frac{(x-x_1)}{(x_3-x_1)}\frac{(x-x_2)}{(x_3-x_2)} = \frac{(x+2)(x+1)x}{(1+2)(1+1)(1)} \\ &= \frac{1}{6}(x^3+3x^2+2x). \end{aligned}$$

Si può verificare di aver calcolato i polinomi L_i in modo corretto valutando $L_i(x_j)$, $i,j = 1,2,3$. Si deve ottenere che $L_i(x_i) = 1$ e $L_i(x_j) = 0$ se $j \neq i$ (verificare per esercizio).

Il polinomio interpolatore è dato da

$$p(x) = 2L_1 - 2L_2 - 6L_3 = (x^3+2x^2-x-2)+(x^3+2x^2-x-2)-(x^3+3x^2+2x),$$

cioè $p(x) = x^3 - 5x - 2$. Il polinomio calcolato è corretto se il suo grado è minore o uguale a N e se $p(x_i) = y_i$, $i = 0,\ldots,N$ (verificare per esercizio).

Matrice di Vandermonde. Il polinomio è della forma $p(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, dove i coefficienti sono la soluzione del seguente sistema lineare $Va = y$, cioè

$$\begin{pmatrix} 1 & -2 & (-2)^2 & (-2)^3 \\ 1 & -1 & (-1)^2 & (-1)^3 \\ 1 & 0 & 0^2 & 0^3 \\ 1 & 1 & 1^2 & 1^3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ -2 \\ -6 \end{pmatrix}$$

Si deve quindi risolvere

$$\begin{pmatrix} 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ -2 \\ -6 \end{pmatrix}$$

la cui soluzione è data da $a = [-2,-5,0,1]^t$.

Esercizio 3.1.2. Forma di Lagrange e resto. Si consideri la funzione $f(x) = \frac{(x+2)^2}{(x-3)}$ nell'intervallo $[-1, 2]$. Si trovi il polinomio interpolatore di f, nella forma di Lagrange, sui nodi $x_0 = -1$, $x_1 = 0$, $x_2 = 0.5$ e $x_3 = 2$. Si studi, inoltre, il grafico del resto dell'interpolazione.

Per calcolare il polinomio interpolatore è necessario conoscere il valore della funzione nei nodi. In questo caso si ha che

$$y_0 = f(-1) = -\frac{1}{4},\ y_1 = f(0) = -\frac{4}{3},\ y_2 = f(0.5) = -\frac{5}{2},\ y_3 = f(2) = -16$$

Il polinomio interpolatore $p(x)$ è di grado al più $N = 3$ e viene espresso nella forma di Lagrange nel modo seguente:

$$p(x) = \sum_{i=0}^{3} L_i(x) f(x_i) = -\frac{1}{4}L_0(x) - \frac{4}{3}L_1(x) - \frac{5}{2}L_2(x) - 16L_3(x)$$

dove

$$\begin{aligned}
L_0 &= \frac{(x-x_1)}{(x_0-x_1)}\frac{(x-x_2)}{(x_0-x_2)}\frac{(x-x_3)}{(x_0-x_3)} = \frac{x(x-0.5)(x-2)}{(-1)(-1.5)(-3)} \\
&= -\frac{2}{9}\left(x^3 - \frac{5}{2}x^2 + x\right) \\
L_1 &= \frac{(x-x_0)}{(x_1-x_0)}\frac{(x-x_2)}{(x_1-x_2)}\frac{(x-x_3)}{(x_1-x_3)} = \frac{(x+1)(x-0.5)(x-2)}{(1)(-0.5)(-2)} \\
&= x^3 - \frac{3}{2}x^2 - \frac{3}{2}x + 1 \\
L_2 &= \frac{(x-x_0)}{(x_2-x_0)}\frac{(x-x_1)}{(x_2-x_1)}\frac{(x-x_3)}{(x_2-x_3)} = \frac{(x+1)(x)(x-2)}{(1.5)(0.5)(-1.5)} \\
&= -\frac{8}{9}(x^3 - x^2 - 2x) \\
L_3 &= \frac{(x-x_0)}{(x_3-x_0)}\frac{(x-x_1)}{(x_3-x_1)}\frac{(x-x_2)}{(x_3-x_2)} = \frac{(x+1)(x)(x-0.5)}{(3)(2)(1.5)} \\
&= \frac{1}{9}\left(x^3 + \frac{1}{2}x^2 - \frac{1}{2}x\right).
\end{aligned}$$

Si ottiene allora che il polinomio interpolatore è dato da

$$p(x) = \frac{1}{18}\left(x^3 - \frac{5}{2}x^2 + x\right) - \frac{4}{3}\left(x^3 - \frac{3}{2}x^2 - \frac{3}{2}x + 1\right) + \frac{20}{9}(x^3 - x^2 - 2x) - \frac{16}{9}\left(x^3 + \frac{1}{2}x^2 - \frac{1}{2}x\right) = -\frac{5}{6}x^3 - \frac{5}{4}x^2 - \frac{3}{2}x - \frac{4}{3}.$$

(Verificare per esercizio le proprietà di p e L_i, $i = 0, \ldots, 3$).

Il resto è dato da

$$r(x) = f(x) - p(x) = \frac{(x+2)^2}{x-3} + \frac{5}{6}x^3 + \frac{5}{4}x^2 + \frac{3}{2}x + \frac{4}{3}.$$

In figura 3.1 vengono riportati, a sinistra, i grafici della funzione f (linea tratteggiata) e del polinomio interpolatore p (linea continua) e, a destra, il grafico del resto dell'interpolazione r (che si annulla, ovviamente, nei nodi).

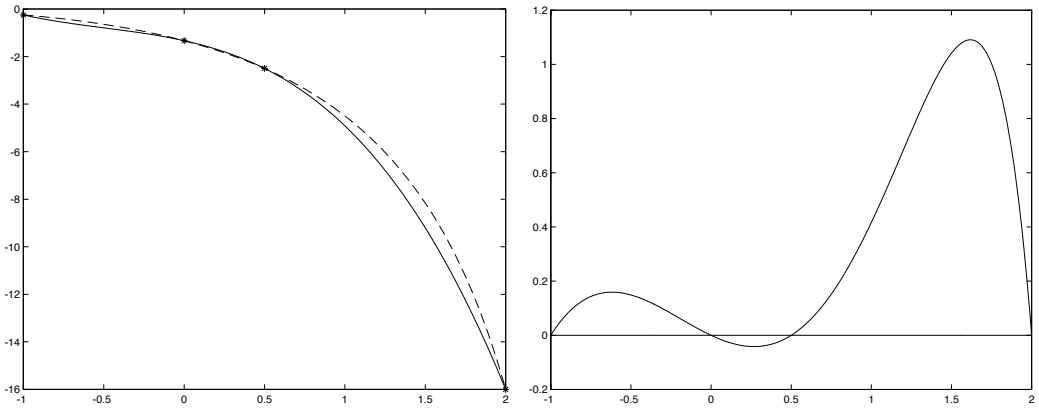

Figure 3.1: A sinistra i gafici della funzione f (linea tratteggiata) e del polinomio p che la interpola (linea continua) e, a destra, il grafico del resto r.

Esercizio 3.1.3. Polinomio interpolatore con grado strettamente minore di N.

Si consideri la funzione $f(x) = x\sin(\pi x)$ nell'intervallo $[-0.5, 1.5]$. Si trovi il polinomio che interpola f, nella forma di Lagrange, sui nodi $x_0 = -0.5$, $x_1 = 0$ e $x_2 = 1.5$.

Per calcolare il polinomio interpolatore è necessario conoscere il valore della funzione nei nodi. In questo caso si ha che

$$y_0 = f(-0.5) = \frac{1}{2},\ y_1 = f(0) = 0,\ y_2 = f(1.5) = -\frac{3}{2}$$

Il polinomio interpolatore $p(x)$ è di grado al più $N = 2$ e viene espresso nella forma di Lagrange nel modo seguente:

$$p(x) = \sum_{i=0}^{2} L_i(x) f(x_i) = \frac{1}{2}L_0(x) + 0L_1(x) - \frac{3}{2}L_2(x) = \frac{1}{2}L_0(x) - \frac{3}{2}L_2(x)$$

Poiché $f(x_1) = 0$, non è necessario calcolare $L_1(x)$. Si ha che

$$\begin{aligned} L_0 &= \frac{(x-x_1)}{(x_0-x_1)}\frac{(x-x_2)}{(x_0-x_2)} = \frac{x(x-1.5)}{(-0.5)(-2)} = x^2 - \frac{3}{2}x \\ L_2 &= \frac{(x-x_0)}{(x_2-x_0)}\frac{(x-x_1)}{(x_2-x_1)} = \frac{x(x+0.5)}{(1.5)(2)} = \frac{1}{3}x^2 + \frac{1}{6}x \end{aligned}$$

Si ottiene allora che il polinomio interpolatore è dato da $p_2(x) = -x$ (semplice verifica), cioè è una retta e quindi di grado strettamente minore di 1; questo fatto si verifica perché i valori della funzione f sui nodi sono allineati.

Esercizio 3.1.4. utilizzo matrice di Vandermonde.
Si consideri la funzione f tale che $f(x) = \sin(\pi x) + 2 * \cos(\pi x)$ nell'intervallo $[0.5, 1.5]$. Si calcolino, utilizzando la matrice di Vandermonde, i coefficienti del polinomio interpolatore di f sui nodi $x_0 = 0.5$, $x_1 = 0.75$ e $x_2 = 1.5$.

Per calcolare il polinomio interpolatore è necessario conoscere il valore della funzione nei nodi. In questo caso si ha che

$$y_0 = f(x_0) = 1 \quad y_1 = f(x_1) = -\sqrt{2}/2, \quad f(x_2) = -1.$$

Il polinomio interpolatore p è di grado al più 2 ed è quindi della forma $p_2(x) = a_0 + a_1 x + a_2 x^2$. I coefficienti a_0, a_1 e a_2 sono la soluzione di un sistema di Vandermonde $Va = y$, dove

$$V = \begin{pmatrix} 1 & 0.5 & 0.25 \\ 1 & .75 & 0.5625 \\ 1 & 1.5 & 2.25 \end{pmatrix}, \qquad a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}, \qquad y = \begin{bmatrix} 1 \\ -\sqrt{2}/2 \\ -1 \end{bmatrix}.$$

la cui soluzione è data da

$$a = \left[4 + 2\sqrt{2},\ -\frac{22 + 16\sqrt{2}}{3},\ \frac{8 + 8\sqrt{2}}{3}\right]^t.$$

Il polinomio interpolatore della funzione f sui tre nodi precedenti è quindi:

$$\frac{8 + 8\sqrt{2}}{3}x^2 - \frac{22 + 16\sqrt{2}}{3}x + 4 + 2\sqrt{2}.$$

Per verificare che il polinomio ottenuto sia veramente il polinomio interpolatore è sufficiente verificare che $p(x_i) = f(x_i)$, $i = 0, 1, 2$, (semplice verifica lasciata al lettore).

La figura 3.1.4 rappresenta il grafico di f (linea tratteggiata) e di p (linea continua) nell'intervallo $[0.5, 1.5]$, evidenziandone il diverso comportamento.

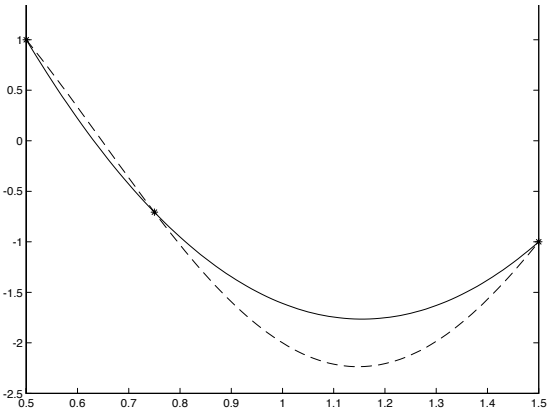

Figure 3.2: Interpolazione su 3 punti

Per quanto riguarda il numero di condizionamento della matrice V, scelta ad esempio la norma 1, si ha che

$$K_1(V) = \|V\|_1 \|V^{-1}\|_1 = \frac{980}{16} \approx 61.2$$

La matrice V è quindi abbastanza mal condizionata.

3.2 Esercizi proposti

Esercizio 3.2.1. Dati 5 nodi $x_0 \dots x_4$ e i corrispondenti valori $y_0, \dots y_4$, dire quale/i delle seguenti affermazioni è/sono vere:

- il polinomio $p(x)$ che li interpola ha sempre grado 4;
- il polinomio $p(x)$ che li interpola ha al più grado 4;
- il polinomio $p(x)$ che li interpola ha al più grado 5
- il resto r si annulla sui nodi.

Esercizio 3.2.2. Dati i punti $P_0 = (1, 1)$, $P_1 = (2, 4)$, $P_2 = (3, 5)$, dire quale/i delle seguenti affermazioni è/sono vere:

- il polinomio $p(x)$ che li interpola ha al più grado 2;
- il polinomio $p(x)$ che li interpola è tale che $p(3) = 5$;
- il polinomio $p(x)$ che li interpola è tale che $p(5) = 3$.

Esercizio 3.2.3. Sia f una funzione tale che $f(0) = 0$, $f(1) = 2$, $f(2) = 5$ e $f(3) = 0$. Calcolare il polinomio interpolatore nella forma di Lagrange e valutarlo nel punto $\sqrt{2}$.

Esercizio 3.2.4. Data la funzione $f(x) = \cos(\pi x) + \sin(\pi x)$, trovare il polinomio che interpola f sui nodi $x_0 = 1$,$x_1 = 1.5$ e $x_2 = 1.75$. Calcolare esattamente il resto nel punto $\hat{x} = 0.75$. In base a tale risultato qual è il polinomio interpolatore per f passante per i nodi $\hat{x}$ e x_0, x_1 e x_2? Giustificare la risposta (non sono necessari calcoli ulteriori).

Esercizio 3.2.5. Dati i nodi $x_0 = 0$, $x_1 = 1$, $x_2 = 3$, $x_3 = 4$ e i valori: $f(x_0) = f(x_3) = 3$, $f(x_1) = f(x_2) = 0$ calcolare il polinomio che li interpola e commentare il grado del polinomio ottenuto.

Esercizio 3.2.6. Sia a un parametro appartenente all'intervallo $[3, 9]$. Calcolare il polinomio interpolatore $p(x)$ della funzione $f(x) = \sin(\pi x) + a\cos(\pi x)$ utilizzando i nodi $x_0 = -1$, $x_1 = 0$ e $x_2 = 0.5$. Scegliere, inoltre, il valore di $a \in [3, 9]$ affinché il valore di $p(-0.5) - f(-0.5)$ (dipendente da a) sia minimo.

Esercizio 3.2.7. $(*)$ Sia $f(x) = x^5 - 3x^4 + 2x^3 + 3$. Calcolare il polinomio interpolatore sui nodi $x_0 = 0$, $x_1 = 2$ e $x_2 = 3$ e valutare, mediante la stima teorica, il valore assoluto dell'errore commesso in $x = 1$.

Esercizio 3.2.8. Sia $f(x) = 2^x + x^2 - 8$. Calcolare il polinomio interpolatore $p(x)$ sui nodi $x_0 = -1$, $x_1 = 1$, $x_2 = 2$, $x_3 = 3$; calcolare, inoltre, il valore assoluto dell'errore commesso approssimando il valore $f(0)$ con $p(0)$.

Esercizio 3.2.9. Interpolare tra $[0, 1.5]$ la funzione $f(x) = e^x(-x^2 + 6x - 9)$ sui nodi $x_0 = 0$, $x_1 = 1$, $x_2 = 1.5$. Scelto $\hat{x} = 0.5$ calcolare, il valore assoluto dell'errore commesso approssimando $f(\hat{x})$ con il valore del polinomio interpolatore in $\hat{x}$.

Esercizio 3.2.10. $(*)$ Sia $f(x) = \frac{x+2}{x}$; verificare che il polinomio

$$q(x) = -\frac{1}{15}x^3 + \frac{7}{10}x^2 - \frac{13}{5}x + \frac{149}{30}$$

è interpolatore per f nei punti $x_0 = 1$, $x_1 = 2.5$, $x_2 = 3$, $x_3 = 4$. Valutare, mediante la stima teorica del resto, il valore assoluto dell'errore commesso nel punto $x = 0.5$ approssimando $f(0.5)$ con $q(0.5)$.

Esercizio 3.2.11. $(*)$ Sia f tale che $f(x) = 2e^{8x}$ nell'intervallo $[0, 0.5]$. Si trovi il polinomio interpolatore di f, nella forma di Lagrange, su nodi equidistanti con passo $h = 1/4$ e calcoli la stima teorica del resto nel punto $x = 1/8$.

Esercizio 3.2.12. $(*)$ Interpolare $f(x) = 3\sin^2(x)\cos(x) - \sin(x) + 2$, sui nodi $x_0 = 0$, $x_1 = \pi/6$ e $x_2 = \pi/2$. Valutare l'errore commesso in $x = \pi/3$ utilizzando la stima teorica del valore assoluto del resto.

Esercizio 3.2.13. Si consideri la funzione $f(x) = \cos^2(\pi x) - x^3$ nell'intervallo $[-1, 1]$. Si calcolino, utilizzando la matrice di Vandermonde, i coefficienti del polinomio interpolatore di f sui nodi $x_0 = -1$, $x_1 = 1/4$ e $x_2 = 1$. Si calcoli, inoltre, il numero di condizionamento in norma 1 della matrice di Vandermonde.

Soluzione esercizi proposti

3.2.1 Falso, Vero, Falso, Vero.

3.2.2 Falso, Vero, Falso.

3.2.3 $p(x) = -1.5x^3 + 5x^2 - 1.5x$; $p(\sqrt{2}) = 10 - 4.5\sqrt{2}$.

3.2.4 $p(x) = \frac{16}{3}x^2 - \frac{40}{3}x + 7$; $p(0.75) = 0 = f(0.75$, cioè $r(0.75) = 0$ e p interpola i punti $\hat{x}$, x_0, x_1, x_2.

3.2.5 $p(x) = x^2 - 4x + 3$, il grado è strettamente minore di $N = 3$.

3.2.6 $p(x) = \frac{4}{3}(1-2a)x^2 - \frac{2}{3}(a-2)x + a$; $a = 3$.

3.2.7 $(*)$ $p(x) = 18x^2 - 36x + 3$; $|r(1)| < 232$.

3.2.8 $p(x) = (7x^3 + 54x^2 + 29x - 330)/48$; $|r(0)| = 1/8$.

3.2.9 $p(x) = x^2(8e - 3e^{1.5} - 6) + x(15 - 12e + 3e^{1.5}) - 9$; $|r(0.5)| = 0.2074$.

3.2.10 $(*)$ $|r(0.5)| < 17.5$.

3.2.11 $(*)$ $p(x) = 16(1-2e^2+e^4)x^2+4(-3+4e^2-e^4)x+2$. $|r(1/8)| = 7.2525$. $|r(1/8)| < e^4 = 54.59$.

3.2.12 $(*)$ $p(x) = \frac{x^2}{\pi^2}(3 - \frac{27}{4}\sqrt{3}) + \frac{x}{\pi}(-\frac{7}{2} + \frac{27}{8}\sqrt{3}) + 2$; $|r(\pi/3)| = 0.6593$, $|r(\pi/3)| < 0.159\pi^3 \approx 4.9285$.

3.2.13 $a_0 = 43/60$, $a_1 = -1$, $a_2 = 17/60$. $K_1(V) = 32/5$.

Chapter 4

Algebra lineare numerica

Un vettore colonna v con n componenti

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

viene indicato nel seguito con $v \in \mathbb{R}^n$ oppure descritto come un vettore $n \times 1$; un generico elemento di v viene indicato con v_i.

Una matrice A con m righe e n colonne

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \\ a_{2,1} & \cdots & a_{2,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}$$

viene indicata nel seguito con $A \in \mathbb{R}^{m \times n}$ oppure descritta come una matrice $m \times n$; un generico elemento di A viene indicato con $a_{i,j}$. Una matrice con $n = 1$ colonne è un vettore colonna.

4.1 Operazioni vettoriali e matriciali

Le possibili operazioni su matrici e vettori sono le seguenti.

- Date due matrici A, $B \in \mathbb{R}^{m \times n}$, la matrice $C = A + B \in \mathbb{R}^{m \times n}$ è tale che $c_{i,j} = a_{i,j} + b_{i,j}$. Inoltre, dato $\lambda \in \mathbb{R}$, $D = \lambda A \in \mathbb{R}^{m \times n}$ è tale che $d_{i.j} = \lambda a_{i,j}$. Le stesse operazioni valgono sui vettori colonna scegliendo $n = 1$ o sui vettori riga scegliendo $m = 1$.

- Dati due vettori v e w di n componenti, si definisce **prodotto scalare** di v e w il seguente valore

$$S = \sum_{i=1}^{n} v_i w_i = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n$$

- Date due matrici $A \in \mathbb{R}^{m \times p}$ e $B \in \mathbb{R}^{q \times n}$, il prodotto $C = AB$ è calcolabile se e solo se $p = q$ ed è definito nel modo seguente. L'elemento $c_{i,j}$ di C è dato dal prodotto scalare della riga i della matrice A con la colonna j della matrice B. La matrice C è una matrice $m \times n$. La stessa regola vale anche nel caso in cui $m = 1$ e/o $n = 1$.

- **NON** è possibile eseguire una divisione tra matrici e/o vettori.

Esercizio 4.1.1. Operazioni vettoriali. Dalti i vettori v e w

$$v = \begin{pmatrix} 1 \\ -3 \\ 6 \end{pmatrix} \qquad w = \begin{pmatrix} -2 \\ 4 \\ 7 \end{pmatrix}$$

e i numeri reali $\alpha = 3$ e $\beta = -4$, calcolare il vettore $z = \alpha v + \beta z$ e calcolare il prodotto scalare tra v e w.

Si ha che

$$\alpha v + \beta w = 3 \begin{pmatrix} 1 \\ -3 \\ 6 \end{pmatrix} - 4 \begin{pmatrix} -2 \\ 4 \\ 7 \end{pmatrix} = \begin{pmatrix} 3 \\ -9 \\ 18 \end{pmatrix} + \begin{pmatrix} 8 \\ -16 \\ -28 \end{pmatrix} = \begin{pmatrix} 11 \\ -25 \\ -10 \end{pmatrix}$$

e

$$S = (1)(-2) + (-3)(4) + (6)(7) = -2 - 12 + 42 = 28$$

Esercizio 4.1.2. Operazioni matriciali. Date le matrici A, B e C e dato il vettore v

$$A = \begin{pmatrix} 1 & -2 & 3 \\ 0 & 2 & 6 \end{pmatrix} \qquad B = \begin{pmatrix} -1 & 3 \\ 10 & 2 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0 \\ 3 & -4 \\ -1 & 5 \end{pmatrix} \qquad v = \begin{pmatrix} -2 \\ 4 \\ 7 \end{pmatrix}$$

eseguire, se possibile, le seguenti operazioni:

(1) AB (2) AC (3) Av (4) Bv (5) vA (6) CB (7) BB

1. AB non è calcolabile perché A è 2×3 e B è 2×2, e quindi il numero delle colonne di A è diverso dal numero delle righe di B.

2. $D = AC$ è calcolabile ed è una matrice 2×2.
   L'elemento $d_{1,1}$ è dato dal prodotto scalare della riga 1 di A e della colonna 1 di C, cioè $d_{1,1} = (1)(1) + (-2)(3) + (3)(-1) = -8$. Analogamente si ha

$$\begin{aligned} d_{1,2} &= (1)(0) + (-2)(-4) + (3)(5) = 23 \\ d_{2,1} &= (0)(1) + (2)(3) + (6)(-1) = 0 \\ d_{2,2} &= (0)(0) + (2)(-4) + (6)(5) = 22 \end{aligned}$$

   da cui si ha che

$$D = \begin{pmatrix} -8 & 23 \\ 0 & 22 \end{pmatrix}$$

3. $w = Av$ è calcolabile ed è un vettore di 2 componenti.
   L'elemento w_1 è dato dal prodotto scalare della riga 1 di A e dell'unica colonna di v, cioè $w_1 = (1)(-2)+(-2)(4)+(3)(7) = 11$. Analogamente si ha $w_2 = (0)(-2) + (2)(4) + (6)(7) = 50$, da cui si ha che

$$w = \begin{pmatrix} 11 \\ 50 \end{pmatrix}$$

4. Bv non è calcolabile perché B è 2×2 e v è 3×1, e quindi il numero delle colonne di B è diverso dal numero delle righe di v.

5. vA non è calcolabile perché v è 3×1 e A è 2×3, e quindi il numero delle colonne di v è diverso dal numero delle righe di A.

6. $E = CB$ è calcolabile ed è una matrice 3×2.
   L'elemento $e_{1,1}$ è dato dal prodotto scalare della riga 1 di C e della colonna 1 di B, cioè $e_{1,1} = (1)(-1) + (0)(10) = -1$. Analogamente si ha

$$\begin{aligned} e_{1,2} &= (1)(3) + (0)(2) = 3 \\ e_{2,1} &= (3)(-1) + (-4)(10) = -43 \\ e_{2,2} &= (3)(3) + (-4)(2) = 1 \\ e_{3,1} &= (-1)(-1) + (5)(10) = 51 \\ e_{3,2} &= (-1)(3) + (5)(2) = 7 \end{aligned}$$

da cui si ha che

$$E = \begin{pmatrix} -1 & 3 \\ -43 & 1 \\ 51 & 7 \end{pmatrix}$$

7. $F = BB$ ed è una matrice 2×2.
   L'elemento $f_{1,1}$ è dato dal prodotto scalare della riga 1 di B e della colonna 1 di B, cioè $f_{1,1} = (-1)(-1) + (3)(10) = 31$. Analogamente si ha

$$\begin{aligned} f_{1,2} &= (-1)(3) + (3)(2) = 3 \\ f_{2,1} &= (10)(-1) + (2)(10) = 10 \\ f_{2,2} &= (10)(3) + (2)(2) = 34 \end{aligned}$$

   da cui si ha che

$$F = \begin{pmatrix} 31 & 3 \\ 10 & 34 \end{pmatrix}$$

4.2 Norme vettoriali e matriciali

La norma di un vettore v viene denotata con $\|v\|$ ed è una sorta di misura della "lunghezza" di un vettore. Una norma è un funzione che ad ogni vettore associa un numero reale positivo e che soddisfa le seguenti proprietà. Dati $v, w \in \mathbb{R}^n$,

1. $\|v\| \geq 0$ e $\|v\| = 0$ se e solo se $v = [0, \ldots, 0]^t$;
2. $\|\alpha v\| = |\alpha| \|v\|$, $\forall \alpha \in \mathbb{R}$;
3. $\|v + w\| \leq \|v\| + \|w\|$.

Se $v = [v_1, \ldots, v_n]^t$, alcune norme standard sono:

$$\|v\|_1 = \sum_{i=1}^{n} |v_i| \qquad \|v\|_\infty = \max_{i=1,\ldots,n} |v_i| \qquad \|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$$

La norma di una matrice A viene denotata con $\|A\|$ ed è una sorta di misura della "grandezza" di una matrice. Una norma è un funzione che ad ogni matrice associa un numero reale positivo e che soddisfa le seguenti proprietà. Date $A, B \in \mathbb{R}^{n \times n}$:

1. $\|A\| \geq 0$ e $\|A\| = 0$ se e solo se A è la matrice formata da tutti zero.
2. $\|\alpha A\| = |\alpha|\|A\|$, $\forall \alpha \in \mathbb{R}$;
3. $\|A + B\| \leq \|A\| + \|B\|$;
4. $\|AB\| \leq \|A\|\|B\|$

Particolari norme matriciali sono le **norme matriciali indotte** (da norme vettoriali). Data una norma vettoriale $\|\|$ la norma matriciale indotta di una matrice A $n \times n$ è definita da

$$\|A\| = \max_{v \in \mathbb{R}^n, v \neq 0} \|Av\|.$$

Si può dimostrare che le norme matriciali indotte rispettivamente dalle norme vettoriali 1 e infinito sono:

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{i,j}| \qquad \|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{i,j}|$$

La norma di Frobenious definita da $\|A\|_F = \sqrt{\sum_{i,j=1}^{n} a_{i,j}^2}$ non è indotta.

Data una norma vettoriale $\|\|$ e la corrispondente norma matriciale indotta, si ha che $\|Av\| \leq \|A\|\|v\|$ per ogni $A \in \mathbb{R}^{n \times n}$ e per ogni $v \in \mathbb{R}^n$. Inoltre $\|I\| = 1$, done I è la matrice identità.

Esercizio 4.2.1. Norme vettoriali. Calcolare le norme 1, 2 e ∞ dei seguenti vettori:

$$\begin{aligned} x &= (1, 4, -8) \qquad y = (-3, -21, -9), \\ z &= (0.5, 0, 0, -0.875) \qquad t = (-0.5, 0, 0, 0.875) \end{aligned}$$

Norma 1: se v è un vettore n-dimensionale, $\|v\|_1 = \sum_{i=1}^{n} |v_i|$.

$$\begin{aligned} \|x\|_1 &= 13 \qquad \|y\|_1 = 33 \\ \|z\|_1 &= 1.375 \qquad \|t\|_1 = 1.375 \end{aligned}$$

Norma 2: se v è un vettore n-dimensionale, $\|v\|_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2}$.

$$\begin{aligned} \|x\|_2 &= 9 \qquad \|y\|_2 = \sqrt{531} \\ \|z\|_2 &= \sqrt{65}/8 \qquad \|t\|_2 = \sqrt{65}/8 \end{aligned}$$

Norma ∞: se v è un vettore n-dimensionale, $\|v\|_\infty = \max_{i=1,n} |v_i|$.

$$\begin{aligned} \|x\|_\infty &= 8 \qquad \|y\|_\infty = 21 \\ \|z\|_\infty &= 0.875 \qquad \|t\|_\infty = 0.875 \end{aligned}$$

♣

Esercizio 4.2.2. Norme matriciali. Calcolare le norme 1, ∞ e di Frobenius delle seguenti matrici:

$$A = \begin{pmatrix} 1 & -1 \\ 5 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} -15 & 68 \\ 12 & -3 \end{pmatrix}$$
$$C = \begin{pmatrix} -7 & 3 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 4 \end{pmatrix} \qquad D = \begin{pmatrix} 1 & -4.5 & 16 \\ -4.5 & 3 & 11 \\ 16 & 11 & -3 \end{pmatrix}$$

Norma 1: se M è una matrice $n \times n$, $\|M\|_1 = \max_j \sum_{i=1}^{n} |M_{ij}|$.

$$\|A\|_1 = 6 \qquad \|B\|_1 = 71$$
$$\|C\|_1 = 7 \qquad \|D\|_1 = 30$$

Norma ∞: se M è una matrice $n \times n$, $\|M\|_\infty = \max_i \sum_{j=1}^{n} |M_{ij}|$.

$$\|A\|_\infty = 5 \qquad \|B\|_\infty = 83$$
$$\|C\|_\infty = 10 \qquad \|D\|_\infty = 30$$

Norma di Frobenius: se M è una matrice $n \times n$, $\|M\|_F = \sqrt{\sum_{i,j=1}^{n} |M_{ij}|^2}$.

$$\|A\|_F = \sqrt{27} \qquad \|B\|_F = \sqrt{5002} \approx 70.72$$
$$\|C\|_F = \sqrt{79} \qquad \|D\|_F = \sqrt{813.5} \approx 28.52$$

♣

4.3 Soluzione sistemi triangolari

Una matrice T $n \times n$ si dice triangolare superiore (inferiore) se $a_{i,j} = 0$ quando $i > j$ ($i < j$). Un sistema lineare $Tx = b$ si dice triangolare se la matrice dei coefficienti T è triangolare.

- Il determinante di una matrice **triangolare** è dato dal prodotto degli elementi diagonali, cioè da $t_{1,1} \cdots t_{n,n}$. **Importante.** Questa proprietà vale solo per le matrici triangolari.

- Un sistema triangolare ammette un'unica soluzione se e solo se il determinante della matrice dei coefficienti è non nullo. Se il determinante è nullo, il sistema può avere infinite soluzioni o non ammetterne nessuna (Tale proprietà vale anche per sistemi quadrati qualsiasi).

- Un sistema con matrice dei coefficienti triangolare superiore può essere risolto, se il determinante di T è non nullo, con il metodo di **sostituzione all'indietro** come segue

$$\begin{aligned} x_n &= \frac{b_n}{t_{n,n}} \\ x_k &= \frac{b_k - \sum_{j=k+1}^{n} a_{k,j} x_j}{t_{k,k}} \end{aligned}$$

  In parole povere, si ricava x_n dall'ultima equazione e ogni componente x_k si calcola sostituendo i valori noti nella k-esima equazione dal basso.

- Se il determinante di T è uguale a zero, allora per almeno un valore x_k non è possibile dividere per $t_{k,k}$. Per tale k si ottiene un'uguaglianza che può essere sempre vera (e quindi tale equazione non fornisce informazioni su x_k) o sempre falsa (e allora il sistema non ammette soluzione).

Esercizio 4.3.1. Sistema triangolare con unica soluzione.
Dati la matrice T e il vettore b,

$$T = \begin{pmatrix} -2 & 2 & 7 \\ 0 & 4 & 2 \\ 0 & 0 & 0.5 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix},$$

calcolare il $\det T$ e risolvere, con il metodo di sostituzione all'indietro, il sistema $Tx = b$.

Il determinante di una matrice triangolare è dato dal prodotto degli elementi della diagonale principale e quindi si ha che $\det T = -2 * 4 * 0.5 = -4$. Ne segue che il sistema ha un'unica soluzione.

Per risolvere il sistema, se $x = [x_1,\ x_2,\ x_3]^t$, si ha che

$$\begin{aligned} x_3 &= \frac{3}{0.5} = 6 \\ x_2 &= \frac{-2 - 2 * x_3}{4} = -3.5 \\ x_1 &= \frac{1 - 2 * x_2 - 7 * x_3}{-2} = 17 \end{aligned}$$

Per verificare che $x = [17,\ ,-3.5,\ 6]^t$ è la soluzione del sistema è sufficiente calcolare il prodotto Tx e verificare che sia uguale a b.

Esercizio 4.3.2. Sistemi triangolari con nessuna o infinite soluzioni. Dati la matrice T e i vettori b_1 e b_2

$$T = \begin{pmatrix} 5 & 2 & 4 \\ 0 & 0 & 2 \\ 0 & 0 & -3 \end{pmatrix}, \qquad b_1 = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}, \qquad b_2 = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix},$$

calcolare il $\det T$ e risolvere, con il metodo di sostituzione all'indietro, i sistemi $Tx = b_1$ e $Tx = b_2$.

Il determinante di una matrice triangolare è dato dal prodotto degli elementi della diagonale principale e quindi si ha che $\det T = 5 * 0 * (-3) = 0$. Non si otterrà quindi un'unica soluzione in nessuno dei due casi.

Per risolvere il sistema $Tx = b_1$, indicato con $x = (x_1,\ x_2,\ x_3)^t$, a partire dall'ultima equazione si ha che $x_3 = \frac{3}{-3} = -1$ e sostituendo nella penultima equazione si ha $0x_2 + 2(-1) = -2$ che è verificata per ogni valore di x_2. Dalla prima equazione si ottiene quindi

$$x_1 = \frac{1 - 2 * x_2 - 4 * x_3}{5} \quad \Longrightarrow \quad x_1 = \frac{5 - 2 * x_2}{5} = 1 - \frac{2}{5}x_2$$

La soluzione del primo sistema è data da $[1 - \frac{2}{5}x_2,\ x_2,\ -1]^t$, cioè dipende da x_2 e quindi il sistema ha infinite soluzioni.

Per risolvere il sistema $Tx = b_2$, indicato con $x = (x_1,\ x_2,\ x_3)^t$, a partire dall'ultima equazione si ha che $x_3 = \frac{3}{-3} = -1$ e sostituendo nella penultima equazione si ha $0x_2 + 2(-1) = 1$, cioè $-2 = 1$ che non è verificata per nessun valore di x_2. Si conclude che il sistema non ammette soluzioni.

4.4 Sistemi con matrice quadrata: metodo di Gauss

Il metodo di eliminazione di Gauss permette di calcolare, se esiste, la soluzione x di un qualsiasi sistema lineare $Ax = b$, $A \in \mathbb{R}^{n \times n}$ trasformando il sistema

originale in un sistema lineare $Tx = c$, con T triangolare superiore, equivalente al primo, cioè che ammette la stessa soluzione x. La soluzione x può quindi essere calcolata con il metodo di sostituzione all'indietro applicato a $Tx = b$.

A partire da A e da b si effettuano delle operazioni che, pur mantenendo costante la soluzione dei sistemi che via via si costruiscono, trasformano A in una matrice T triangolare superiore, annullando, ad ogni passaggio gli elementi di una colonna di A che stanno sotto l'elemento diagonale della colonna stessa.

Importante Oltre ad elaborare la matrice A è necessario elaborare nello stesso modo anche gli elementi di b. Per tale motivo si considera la matrice $[A \,|\, b]$ ottenuta da A aggiungendo come ultima colonna il vettore b.
Al passo j, $j = 1, \ldots, (n-1)$, indicate con r_j e r_k rispettivamente la j-esima e la k-esima riga della matrice $\widehat{A}_j$ da elaborare, si lasciano inalterate le prime j righe di $\widehat{A}_j$ ed ogni riga r_k, $k > j$ viene sostituita $r_k + \alpha r_j$ con α scelto in modo tale che il j-esimo elemento della nuova riga r_k sia nullo. Il valore di α è dato da $\alpha = -\frac{a_{k,j}^{(j)}}{a_{j,j}^{(j)}}$, dove $a_{j,j}^{(j)}$ è il j-esimo elemento di riga r_j (detto pivot) e $a_{k,j}^{(j)}$ è il j-esimo elemento di riga r_k.

Se l'elemento pivot è nullo, prima di eseguire i calcoli relativi al passo j, si effettua uno scambio tra r_j e una riga r_k, $k > j$ in modo da ottenere un elemento pivot non nullo. Se non ci sono "candidati" pivot non nulli, allora vuol dire che gli elementi sotto la diagonale principale della j-esima colonna sono nulli e si può passare al passo successivo.

Dopo $(n-1)$ passi si ottiene la matrice $[T \,|\, c]$ con T triangolare superiore tale che $Tx = c$ è un sistema equivalente a $Ax = b$.

Inoltre $\det T = (-1)^p \det A$, con p numero di scambi di righe necessari per terminare la triangolarizzazione. Da ciò segue che il metodo di Gauss permette di calcolare il $\det A$ con un metodo molto più veloce del classico metodo di Laplace.

Infine, si ha che $\det A \neq 0$ se e solo se $\det T \neq 0$, cioè esiste unica soluzione di $Ax = b$ se e solo se esiste unica soluzione di $Tx = b$.

Esercizio 4.4.1. Il metodo di eliminazione di Gauss. Calcolare, con il metodo di Gauss, la soluzione del sistema $Ax = b$, dove

$$A = \begin{pmatrix} 2 & -4 & 1 \\ 6 & -14 & 8 \\ -2 & 0 & 6 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

Vengono elaborati contemporaneamente la matrice A e il vettore dei ter-

mini noti.

$$[A|b] = \left(\begin{array}{ccc|c} 2 & -4 & 1 & 1 \\ 6 & -14 & 8 & -1 \\ -2 & 0 & 6 & 1 \end{array}\right)$$

Primo passo: $j = 1$. Si eseguono le somme delle righe $(r_1 - 3r_1)$ e delle righe $(r - 3 + r_1)$ e si ottiene

$$\left(\begin{array}{ccc|c} 2 & -4 & 1 & 1 \\ 0 & -2 & 5 & -4 \\ 0 & -4 & 7 & 2 \end{array}\right)$$

Secondo passo: $j = 2$. Si esegue la somma delle righe $(r - 3 - 2r_2)$ e si ottiene

$$\left(\begin{array}{ccc|c} 2 & -4 & 1 & 1 \\ 0 & -2 & 5 & -4 \\ 0 & 0 & -3 & 10 \end{array}\right).$$

Risolvendo il sistema triangolare così ottenuto si ha

$$\begin{aligned} x_3 &= -\frac{10}{3} \\ x_2 &= -\frac{-4 - 50/3}{2} = -\frac{19}{3} \\ x_1 &= \frac{1 + 10/3 - 76/3}{2} = -\frac{21}{2} \end{aligned}$$

cioè

$$x = \left(\begin{array}{c} \frac{21}{2} \\ \frac{19}{3} \\ -\frac{10}{3} \end{array}\right).$$

Per verificare di aver ottenuto la soluzione esatta è sufficiente calcolare il prodotto Ax e verificare che sia uguale a b.

Il determinante della matrice A è dato da $\det A = (-1)^p \det T = (-1)^0(2) * (-2) * (-3) = 12$, poiché non ci sono stati scambi di righe, cioè $p = 0$ e il determinante di una matrice triangolare è dato dal prodotto dei suoi elementi diagonali.

Esercizio 4.4.2. Gauss con scambio di righe e $\det A \neq 0$**.**

Risolvere, con il metodo di Gauss, il sistema lineare $Ax = b$, con

$$A = \begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -3 \\ 4 & -7 & 3 \end{pmatrix} \qquad b = \begin{pmatrix} \frac{1}{12} \\ -\frac{5}{12} \\ \frac{1}{4} \end{pmatrix}$$

e calcolare il determinante di A.

La matrice $[A \,|\, b]$ viene elaborata nel modo seguente.

$$[A \,|\, b] = \left(\begin{array}{ccc|c} 1 & -2 & 1 & \frac{1}{12} \\ -2 & 4 & -3 & -\frac{5}{12} \\ 4 & -7 & 3 & \frac{1}{4} \end{array}\right) \quad (r_2 + 2r_1), \quad (r_3 - 4r_1) \Longrightarrow$$
$$\left(\begin{array}{ccc|c} 1 & -2 & 1 & \frac{1}{12} \\ 0 & 0 & -1 & -\frac{1}{4} \\ 0 & 1 & -1 & -\frac{1}{12} \end{array}\right)$$

Poiché l'elemento pivot di posto $(2, 2)$ è nullo, si effettua uno scambio della seconda e terza riga e si ottiene

$$\left(\begin{array}{ccc|c} 1 & -2 & 1 & \frac{1}{12} \\ 0 & 1 & -1 & -\frac{1}{12} \\ 0 & 0 & -1 & -\frac{1}{4} \end{array}\right).$$

Dopo lo scambio tra la terza e la seconda riga, dovuto al fatto di aver ottenuto il pivot nullo, la matrice così ottenuta risulta triangolare superiore e si può risolvere il sistema con il metodo di sostituzione all'indietro:

$$\begin{aligned} x_3 &= \frac{1}{4} \\ x_2 &= -\frac{1}{12} + x_3 = -\frac{1}{12} + \frac{1}{4} = \frac{1}{3} \\ x_1 &= \frac{1}{12} - x_3 + 2x_2 = \frac{1}{12} - \frac{1}{4} + \frac{2}{3} = \frac{1}{2} \end{aligned}$$

cioè

$$x = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{4} \end{pmatrix}.$$

Per verificare di aver ottenuto la soluzione esatta è sufficiente calcolare il prodotto Ax e verificare che sia uguale a b.

Poiché è stato effettuato uno scambio di righe si ha che $\det A = (-1)^1 \det T = -(1) * (1) * (-1) = 1$. Essendo $\det A \neq 0$ allora il sistema $Ax = b$ ammette unica soluzione.

♣

Esercizio 4.4.3. Gauss con $\det A = 0$**.**

Risolvere, con il metodo di Gauss, i sistemi lineari $Ax = b_1$ e $Ax = b_2$ con

$$A = \begin{pmatrix} 3 & -2 & 1 \\ 1 & 4 & 7 \\ 2 & -6 & -6 \end{pmatrix} \qquad b_1 = \begin{pmatrix} 0 \\ -20 \\ 20 \end{pmatrix} \qquad b_2 = \begin{pmatrix} 1 \\ -5 \\ 10 \end{pmatrix}$$

e calcolare il determinante di A.

Poiché si hanno due sistemi con la stessa matrice dei coefficienti e poiché la triangolarizzazione è"governata" dalla matrice A, si costruisce aa matrice $[A \,|\, b_1\, b_2]$ e si attua una solo triangolarizzazione nel modo seguente.

$$[A \,|\, b_1\, b_2] = \left(\begin{array}{ccc|cc} 3 & -2 & 1 & 0 & 1 \\ 1 & 4 & 7 & -20 & -5 \\ 2 & -6 & -6 & 20 & 10 \end{array}\right) \quad (r_2 - r_1/3), \quad (r_3 - 2r_1/3) \Longrightarrow$$

$$\left(\begin{array}{ccc|cc} 3 & -2 & 1 & 0 & 1 \\ 0 & \frac{14}{3} & \frac{20}{3} & -20 & -\frac{16}{3} \\ 0 & -\frac{14}{3} & -\frac{20}{3} & 20 & \frac{28}{3} \end{array}\right) \quad (r_3 + r_2) \Longrightarrow$$

$$\left(\begin{array}{ccc|cc} 3 & -2 & 1 & 0 & 1 \\ 0 & \frac{14}{3} & \frac{20}{3} & -20 & -\frac{16}{3} \\ 0 & 0 & 0 & 0 & 4 \end{array}\right)$$

La matrice così ottenuta risulta triangolare superiore con determinante uguale a 0. Da ciò segue che $\det A = 0$ e quindi si possono avere infinite soluzioni o nessuna soluzione.

Si consideri il primo sistema triangolare dato da:

$$\begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{14}{3} & \frac{20}{3} \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -20 \\ 0 \end{pmatrix}$$

La terza equazione $0 = 0$ è soddisfatta da qualunque valore di x_1, x_2 e x_3. Mediante il metodo di sostituzione all'indietro si ottiene

$$\begin{aligned} x_2 &= \frac{3}{14}\left(-20 - \frac{20}{3}x_3\right) = -\frac{30}{7} - \frac{10}{7}x_3 \\ x_1 &= \frac{1}{3}(2x_2 - x_3) = -\frac{9}{7}x_3 - \frac{20}{7} \end{aligned}$$

Il sistema ha infinite soluzioni, una per ciascun valore di x_3 scelto.

Si consideri il secondo sistema triangolare dato da:

$$\begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{14}{3} & \frac{20}{3} \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{16}{3} \\ 4 \end{pmatrix}$$

La terza equazione $0 = 4$ e quindi non è soddisfatta da nessun valore di x_1, x_2 e x_3. Il sistema non ha quindi soluzione.

4.5 Sistemi con matrice quadrata: il metodo di Jacobi

Il metodo di Jacobi permette di approssimare la soluzione di un sistema $Ax = b$ con n equazioni ed n incognite costruendo una successione (sequenza) di vettori tali che, scelto un primo vettore $x_0 \in \mathbb{R}^n$, si ha

$$x_{k+1} = Jx_k + q, \qquad \text{con} \quad J = D^{-1}(D - A) \quad \text{e} \quad q = D^{-1}b$$

dove D è la matrice diagonale formata dagli elementi diagonali di A, cioè D ha tutti gli elementi nulli tranne $d_{i,i} = a_{i,i}$, $i = 1, \ldots, n$. Essendo D diagonale la sua inversa è una matrice che ha tutti gli elementi nulli tranne gli elementi sulla diagonale dati da $1/a_{i,i}$, $i = 1, \ldots, n$. (**Importante** Tale regola per calcolare l'inversa vale solo per le matrici diagonali).

Se la successione $\{x_k\}_{k=1,2,\ldots}$ tende a un vettore y per k che tende all'infinito, allora y è soluzione del sistema lineare, cioè $Ay = b$.

Scelto un qualsiasi vettore x_0 la convergenza del metodo non è garantita. Vale però la seguente **condizione sufficiente per la convergenza del metodo di Jacobi**: se la matrice A è a predominanza diagonale per righe, cioè

$$|a_{k,k}| > \sum_{j=1, j\neq k}^{n} |a_{k,j}|, \qquad \forall k = 1, \ldots, n$$

oppure se la matrice A è a predominanza diagonale per colonne, cioè

$$|a_{k,k}| > \sum_{i=1, i\neq k}^{n} |a_{i,k}|, \qquad \forall k = 1, \ldots, n$$

allora, per ogni scelta di x_0, la successione $\{x_k\}_{k=1,2,\ldots}$ costruita con il metodo di Jacobi tende alla soluzione del sistema $Ax = b$.

Esercizio 4.5.1. Il metodo di Jacobi. Approssimare, con il metodo di Jacobi, la soluzione del sistema $Ax = b$, dove

$$A = \begin{pmatrix} 12 & -4 & 1 \\ 6 & -16 & 8 \\ -2 & 0 & 6 \end{pmatrix} \qquad b = \begin{pmatrix} 18 \\ -20 \\ -16 \end{pmatrix}$$

Innanzi tutto si noti che la matrice A non è a predominanza diagonale per colonne. Infatti per quanto riguarda l'ultima colonna $|6|$ non è maggiore di $|1| + |8|$. Tuttavia la matrice A è a predominanza diagonale per righe poiché

$$|12| > |-4| + |1| \qquad |-16| > |6| + |8| \qquad |6| > |-2| + |0|$$

e quindi il metodo di Jacobi converge per ogni scelta del vettore x_0.

La matrice di iterazione J è data da

$$J = D^{-1}(D - A) = \begin{pmatrix} 12 & 0 & 0 \\ 0 & -16 & 0 \\ 0 & 0 & 6 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 4 & -1 \\ -6 & 0 & -8 \\ 2 & 0 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} 1/12 & 0 & 0 \\ 0 & -1/16 & 0 \\ 0 & 0 & 1/6 \end{pmatrix} \begin{pmatrix} 0 & 4 & -1 \\ -6 & 0 & -8 \\ 2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1/3 & -1/12 \\ 3/8 & 0 & 1/2 \\ 1/3 & 0 & 0 \end{pmatrix},$$

e il vettore q è dato da

$$q = D^{-1}b = \begin{pmatrix} 1/12 & 0 & 0 \\ 0 & -1/16 & 0 \\ 0 & 0 & 1/6 \end{pmatrix} \begin{pmatrix} 18 \\ -20 \\ -16 \end{pmatrix} = \begin{pmatrix} 3/2 \\ 5/4 \\ -8/3 \end{pmatrix}.$$

Sia $x_0 = (1,\ 0,\ -1)^t$. Si ottiene che

$$x_1 = Jx_0 + q = \begin{pmatrix} 0 & 1/3 & -1/12 \\ 3/8 & 0 & 1/2 \\ 1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 3/2 \\ 5/4 \\ -8/3 \end{pmatrix}$$

$$= \begin{pmatrix} 1/12 \\ -1/8 \\ 1/3 \end{pmatrix} + \begin{pmatrix} 3/2 \\ 5/4 \\ -8/3 \end{pmatrix} = \begin{pmatrix} 1/12 + 3/2 \\ -1/8 + 5/4 \\ 1/3 - 8/3 \end{pmatrix}$$

$$= \begin{pmatrix} 19/12 \\ 9/8 \\ -7/3 \end{pmatrix} \approx \begin{pmatrix} 1.5833 \\ 0.6250 \\ -2.3333 \end{pmatrix}$$

Analogamente per i passi successivi

$$x_2 = Jx_1 + q = \begin{pmatrix} 41/72 + 3/2 \\ -55/96 + 5/4 \\ 19/36 - 8/3 \end{pmatrix} = \begin{pmatrix} 149/72 \\ 65/96 \\ -77/36 \end{pmatrix} \approx \begin{pmatrix} 2.0694 \\ 0.6771 \\ -2.1389 \end{pmatrix}$$

$$x_3 = Jx_2 + q = \begin{pmatrix} 349/864 + 3/2 \\ -169/576 + 5/4 \\ 149/216 - 8/3 \end{pmatrix} = \begin{pmatrix} 1645/864 \\ 571/576 \\ -427/216 \end{pmatrix} \approx \begin{pmatrix} 1.9039 \\ 0.9566 \\ -1.9769 \end{pmatrix}.$$

Le iterazioni precedenti sembrano approssimare, come soluzione del sistema, il vettore $[2,\, 1,\, -2]^t$. Per ottenere una valutazione della precisione di tale scelta si può verificare se $Ax = b$. In questo caso l'uguaglianza è verificata e quindi il vettore x è la soluzione del sistema.

Esercizio 4.5.2. Confronto tra i metodi di Gauss e di Jacobi.

Dati la matrice A e il vettore b

$$A = \begin{pmatrix} 4 & -1 & 1 & 1 \\ 1 & 2 & 0 & 0 \\ -1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 5 \\ -1 \\ -1 \end{pmatrix}$$

Calcolare la soluzione del sistema $Ax = b$ utilizzando il metodo di Gauss e approssimarla utilizzando il metodo di Jacobi.

Con il metodo di Gauss la matrice $[A \,|\, b]$ viene elaborata nel modo seguente.

$$[A \,|\, b] = \left(\begin{array}{cccc|c} 4 & -1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 & 5 \\ -1 & 0 & 2 & 0 & -1 \\ 1 & 0 & 0 & 2 & -1 \end{array}\right) \quad (r_2 - 1/4r_1), \quad (r_3 + 1/4r_1) \quad (r_4 - 1/4r_1)$$

$$\Longrightarrow \quad \left(\begin{array}{cccc|c} 4 & -1 & 1 & 1 & 1 \\ 0 & 9/4 & -1/4 & -1/4 & 19/4 \\ 0 & -1/4 & 9/4 & 1/4 & -3/4 \\ 0 & 1/4 & -1/4 & 7/4 & -5/4 \end{array}\right) \quad (r_3 + 1/9r_2), \quad (r_4 - 1/9r_2)$$

$$\Longrightarrow \quad \left(\begin{array}{cccc|c} 4 & -1 & 1 & 1 & 1 \\ 0 & 9/4 & -1/4 & -1/4 & 19/4 \\ 0 & 0 & 20/9 & 2/9 & -2/9 \\ 0 & 0 & -2/9 & 16/9 & -16/9 \end{array}\right) (r_4 + 1/10r_3)$$

$$\Longrightarrow \quad \left(\begin{array}{cccc|c} 4 & -1 & 1 & 1 & 1 \\ 0 & 9/4 & -1/4 & -1/4 & 19/4 \\ 0 & 0 & 20/9 & 2/9 & -2/9 \\ 0 & 0 & 0 & 9/5 & -9/5 \end{array}\right).$$

Risolvendo il sistema triangolare così ottenuto si ha

$$\begin{aligned} x_4 &= -1 \\ x_3 &= \frac{9}{20}(-\frac{2}{9} - \frac{2}{9}x_4) = 0 \\ x_2 &= \frac{4}{9}(\frac{19}{4} + \frac{1}{4}x_4 + \frac{1}{4}x_3) = 2 \\ x_1 &= \frac{1}{4}(1 - x_4 - x_3 + x_2) = 1 \end{aligned}$$

La soluzione del sistema è data da $[1,\ 2,\ 0,\ -1]^t$.

Per applicare il metodo di Jacobi, è necessario innanzi tutto calcolare la matrice di iterazione $J = D^{-1}(D - A)$ e il vettore $q = D^{-1}b$, cioè

$$\begin{aligned} J &= \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 0 & 1 & -1 & -1 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} = \\ & \begin{pmatrix} 0 & 1/4 & -1/4 & -1/4 \\ -1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 0 \end{pmatrix} \\ q &= \begin{pmatrix} 1/4 \\ 5/2 \\ -1/2 \\ -1/2 \end{pmatrix} \end{aligned}$$

Si noti che la matrice triangolare superiore ottenuta con il metodo di Gauss non conserva, nella parte superiore, gli zeri originali, mentre la matrice di iterazione J conserva gli stessi zeri di A, cioè la sua la stessa struttura.

Scegliendo $x_0 = [0,\ 0,\ 0,\ 0]^t$ si ottiene $x_1 = Jx_0 + q = q$ e

$$
\begin{aligned}
x_2 &= Jx_1 + q = \begin{pmatrix} 9/8 \\ 19/8 \\ -3/8 \\ -5/8 \end{pmatrix} \approx \begin{pmatrix} 1.1250 \\ 2.3750 \\ -0.3750 \\ -0.6250 \end{pmatrix} \\
x_3 &= Jx_1 + q = \begin{pmatrix} 35/32 \\ 31/16 \\ 1/16 \\ -17/16 \end{pmatrix} \approx \begin{pmatrix} 1.0938 \\ 1.9375 \\ 0.0625 \\ -1.0625 \end{pmatrix} \\
x_4 &= Jx_1 + q = \begin{pmatrix} 63/64 \\ 125/64 \\ 3/64 \\ -67/64 \end{pmatrix} \approx \begin{pmatrix} 0.9844 \\ 1.9531 \\ 0.0469 \\ -1.0469 \end{pmatrix} \\
x_5 &= Jx_1 + q = \begin{pmatrix} 253/256 \\ 257/128 \\ -1/128 \\ -127/128 \end{pmatrix} \approx \begin{pmatrix} 0.9883 \\ 2.0078 \\ -0.0078 \\ -0.9922 \end{pmatrix}
\end{aligned}
$$

La matrice A è a predominanza diagonale, il metodo di Jacobi converge ed effettivamente i vettori x_0, x_1, x_2, x_3, x_4 e x_5 si stanno "avvicinando" alla soluzione $x = [1,\ 2,\ 0,\ -1]^t$ del sistema.

4.6 Numero di condizionamento

Il **numero di condizionamento** $K(A)$ di una matrice A $n \times n$ rispetto a una data una norma matriciale $\|\ \ \|$ è definito da $K(A) = \|A\|\|A^{-1}\|$, dove A^{-1} è la matrice **inversa** di A.

Valgono le seguenti proprietà relative alla matrice A^{-1} inversa di A.

- L'inversa di A esiste se e solo se $\det A \neq 0$ e, se esiste, è unica.
- L'inversa di A soddisfa la relazione $AA^{-1} = A^{-1}A = I$ dove I è la matrice identità, cioè una matrice diagonale $n \times n$ con gli elementi sulla diagonale uguali a 1.
- La matrice inversa di A^{-1} è la matrice A.
- La matrice A^{-1} può essere calcolata utilizzando il metodo di Gauss risolvendo n sistemi lineari del tipo $Ay^{(i)} = e_i$ con e_i i-esima colonna di I. Il vettore soluzione $y(i)$ dell'i-esimo sistema lineare coincide con la i-esima colonna di A^{-1}. Poiché tutti i sistemi in questione hanno A come

matrice dei coefficienti si possono calcolare tutte le traingolarizzazioni contemporaneamente, cioè si applica Gauss alla matrice $[A \mid I]$, dove sono state aggiunte alle colonne di A le colonne della matrice identità. Si ottiene una matrice $[T \mid c^{(1)}, \ldots, c^{(n)}]$ e le colonne di A^{-1} sono quindi soluzioni dei sistemi lineari

$$Ty^{(1)} = c^{(1)} \qquad \ldots \qquad Ty^{(n)} = c^{(n)}.$$

Il numero di condizionamento di una matrice permette di valutare di quanto varia la soluzione del sistema lineare $Ax = b$ quando si perturbano gli elementi della matrice A e del vettore b.

Se si suppone di perturbare il solo vettore b aggiungendo un vettore δb si ha che

$$\frac{\|\delta x\|}{\|x\|} \leq K(A)\frac{\|\delta b\|}{\|b\|}$$

dove x è la soluzione del sistema non perturbato $Ax = b$ e $(x + \delta x)$ è la soluzione del sistema perturbato $A(x + \delta x) = (b + \delta b)$.

Se la norma usata è una norma matriciale indotta, allora $K(A) \geq 1$. Se $K(A)$ è un valore basso, allora il sistema si dice **ben condizionato**, cioè a piccole variazioni del vettore b corrispondono piccole variazioni della soluzione x. Viceversa, se $K(A)$ è un valore elevato, allora il sistema si dice **mal condizionato**, cioè a piccole variazioni del vettore b possono corrispondere grandi variazioni della soluzione x.

Una matrice $n \times n$ è ben condizionata se il suo numero di condizionamento non si discosta troppo da n.

Esercizio 4.6.1. L'inversa di una matrice. Calcolare, con il metodo di Gauss, l'inversa della matrice

$$A = \begin{pmatrix} 5 & -1 & 2 \\ 3 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Per calcolare l'inversa della matrice si risolvono tre sistemi i cui termini noti sono costituiti dalle colonne della matrice identica. Si noti che non è necessario triangolarizzare tre volte la matrice, ma vengono elaborati contemporaneamente la matrice e le tre colonne dei termini noti.

$$(A|I) = \left(\begin{array}{ccc|ccc} 5 & -1 & 2 & 1 & 0 & 0 \\ 3 & -1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \quad (r_2 - 3/5r_1), \quad (r_3 \text{ inalterata}) \Longrightarrow$$

$$\left(\begin{array}{ccc|ccc} 5 & -1 & 2 & 1 & 0 & 0 \\ 0 & -2/5 & -6/5 & -3/5 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \quad (r_3 + 5/2 r_2) \Longrightarrow$$

$$\left(\begin{array}{ccc|ccc} 5 & -1 & 2 & 1 & 0 & 0 \\ 0 & -2/5 & -6/5 & -3/5 & 1 & 0 \\ 0 & 0 & -2 & -3/2 & 5/2 & 1 \end{array}\right).$$

Poiché la matrice triangolare ottenuta ha lo stesso determinante di A, si ha che $\det A = 4c \neq 0$ e quindi esiste unica l'inversa della matrice A. Risolvendo i tre sistemi lineari così ottenuti

$$\left\{\begin{array}{rcr} 5x_1 + x_2 + 2x_3 &=& 1 \\ -2/5x_2 - 6/5x_3 &=& -3/5 \\ -2x_3 &=& -3/2 \end{array}\right. \qquad \left\{\begin{array}{rcr} 5x_1 + x_2 + 2x_3 &=& 0 \\ -2/5x_2 - 6/5x_3 &=& 1 \\ -2x_3 &=& 5/2 \end{array}\right.$$

$$\left\{\begin{array}{rcr} 5x_1 + x_2 + 2x_3 &=& 0 \\ -2/5x_2 - 6/5x_3 &=& 0 \\ -2x_3 &=& 1 \end{array}\right.$$

si ottengono rispettivamente la prima, la seconda e la terza colonna di A^{-1}, cioè

$$A^{-1} = \left(\begin{array}{ccc} -0.25 & 0.75 & 0.5 \\ -0.75 & 1.25 & 1.5 \\ 0.75 & -1.25 & -0.5 \end{array}\right).$$

Per verificare di aver ottenuto la corretta matrice inversa è sufficiente calcolare il prodotto AA^{-1} e controllare che coincida con la matrice identità.

Esercizio 4.6.2. Numero di condizionamento. Calcolare il numero di condizionamento della matrice A dell'esercizio precedente

$$A = \left(\begin{array}{ccc} 5 & -1 & 2 \\ 3 & -1 & 0 \\ 0 & 1 & 1 \end{array}\right)$$

utilizzando la norma 1.

Poiché l'inversa della matrice A è data da (vedere esercizio precedente)

$$A^{-1} = \left(\begin{array}{ccc} -0.25 & 0.75 & 0.5 \\ -0.75 & 1.25 & 1.5 \\ 0.75 & -1.25 & -0.5 \end{array}\right).$$

per calcolare $K_1(A)$ è sufficiente calcolare $\|A\|_1$ e $\|A^{-1}\|_1$. Si ha che

$$\|A\|_1 = \max\{8,\, 3,\, 3\} = 8 \quad \text{e} \quad \|A^{-1}\|_1 = \max\{1.75,\, 3.25,\, 2.5\} = 3.25$$

e quindi $K_1(A) = 8 * 3.25 = 26$.

Esercizio 4.6.3. Numero di condizionamento. Discutere, al variare del parametro $\alpha > 0$, il numero di condizionamento, in norma ∞, della matrice

$$A = \begin{pmatrix} 1+\alpha & 1 \\ -1 & -1 \end{pmatrix}.$$

La matrice inversa di A è data da

$$A^{-1} = \begin{pmatrix} \frac{1}{\alpha} & \frac{1}{\alpha} \\ -\frac{1}{\alpha} & -(\frac{1}{\alpha}+1) \end{pmatrix},$$

da cui si ottiene che

$$K_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = \max\{2+\alpha, 2\} \max\{\frac{2}{\alpha}, \frac{2}{\alpha}+1\} = (2+\alpha)(\frac{2}{\alpha}+1)$$

Studiando la funzione $f(\alpha) = (2+\alpha)(\frac{2}{\alpha}+1) = \alpha + 4 + \frac{4}{\alpha}$ si ottiene che tale funzione tende a $+\infty$ per α che tende a 0 e per α che tende a $+\infty$ e in $\alpha = 2$ assume il valore minimo pari a 8. Si può concludere che il condizionamento ottimo si ha per $\alpha = 2$ e che la matrice A è ben condizionata se α assume valori non troppo vicini a 0 e non troppo elevati. Ad esempio:

- per $\alpha = 8$ si ha $K_\infty(A) = 12.5$ e la matrice ha il numero di condizionamento minimo;
- per $\alpha = 0.1$ si ha $K_\infty(A) = 44.1$ e per $\alpha = 10$ si ha $K_\infty(A) = 14.4$ (matrice ben condizionata);
- per $\alpha = 0.001$ si ha $K_\infty(A) = 4004.001$ e per $\alpha = 1000$ si ha $K_\infty(A) = 1004.004$ (matrice mal condizionata).

Esercizio 4.6.4. Studio dell'errore inerente per i sistemi lineari. Siano dati la matrice A, il vettore b e la perturbazione δb di b

$$A = \begin{pmatrix} 1.01 & 1 \\ -1 & -1 \end{pmatrix} \qquad b = \begin{pmatrix} 1.1 \\ -1 \end{pmatrix} \qquad \delta b = \begin{pmatrix} 0.1 \\ 0 \end{pmatrix}$$

Calcolare le soluzioni dei sistemi $Ax = b$ e $A(x + \delta x)x = b + \delta b$; inoltre, calcolare direttamente e valutare con la stima teorica l'errore inerente causato dalla perturbazione δb.

La matrice A di questo esercizio coincide con la matrice A dell'esercizio 4.6.3 ponendo $\alpha = 0.01$, e quindi $K_\infty(A) = 2.01 * 201 = 404.01$, cioè la matrice A è mal condizionata. La soluzione del sistema $Ax = b$ è data da $[10,\ -9]^t$ (si può ottenere con il metodo di Gauss).

Poiché $\widetilde{b} = (b + \delta b) = [1.2,\ -1]^t$, posto $\widetilde{x} = (x + \delta x)$, la soluzione del sistema perturbato $A\widetilde{x} = \widetilde{b}$ è data da $\widetilde{x} = [20,\ -19]^t$.

È quindi possibile calcolare direttamente l'errore inerente, usado ad esempio la norma infinito, scelta per valutare il numero di condizionamento:

$$\frac{\|\widetilde{x} - x\|_\infty}{\|x\|_\infty} = \frac{\|(x + \delta x) - x\|_\infty}{\|x\|_\infty} = \frac{\|\delta x\|_\infty}{\|x\|_\infty} = \frac{\|[20,\ -19]^t - [10,\ -9]^t\|_\infty}{\|[10,\ -9]^t\|_\infty} = \frac{10}{10} = 1$$

Si ha quindi che, in questo caso, poiché la matrice è mal condizionata, l'errore relativo sulla soluzione è del 100%.

Si consideri, adesso, la valutazione fornita dalla stima teorica dell'errore:

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \leq K_\infty(A)\frac{\|\delta b\|_\infty}{\|b\|_\infty} = 404.01\frac{0.1}{1.1} = 404.01\frac{1}{11} \approx 36.7282$$

Si noti che la stima teorica dell'errore inerente risulta essere eccessivamente elevata.

4.7 Esercizi proposti

Esercizio 4.7.1. Calcolare (se possibile) i prodotti AB e BA, dove

(1) $$A = \begin{pmatrix} 1 & -1 & 0 \\ 2 & -3 & 1 \\ -1 & 0 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}.$$

(2) $$A = \begin{pmatrix} -1 & 3 & 2 \\ 2 & 0.5 & -1 \\ 2 & 1 & -2 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & -2 & 1 \\ 1 & -3 & 4 \\ -1 & 1 & -1 \end{pmatrix}.$$

(3) $$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 2 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 5 & -1 \\ -1 & 1 & -2 \\ -2 & 0 & 1 \end{pmatrix}.$$

(4) $$A = \begin{pmatrix} -1 & 1 \\ 2 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 5 \\ 2 & -1 \\ 3 & 0 \end{pmatrix}.$$

(5) $$A = \begin{pmatrix} -1 & 1 & 0 \\ 2 & -3 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 5 \\ 3 & -1 \\ 3 & 2 \end{pmatrix}.$$

(6) $$A = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ -2 & -6 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 5 & -7 \\ 0 & 1 & 3 \end{pmatrix}.$$

(7) $$A = \begin{pmatrix} 1 & -1 & -1 & -1 \\ -2 & 1 & -3 & -1 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 3 & 3 \\ 2 & 3 \\ 0 & 1 \\ 1 & 3 \end{pmatrix}.$$

Esercizio 4.7.2. Calcolare le norme 1, 2 e ∞ dei seguenti vettori:

$$x = [9,\ -7]; \qquad y = [8,\ 8,\ -17];$$
$$z = [7,\ -11,\ 7,\ -9]; \qquad t = [-0.2,\ 6.4,\ 4.7,\ -1.7]$$

Esercizio 4.7.3. Calcolare le norme 1, ∞ e di Frobenius delle seguenti matrici:

$$A = \begin{pmatrix} -6.1 & 3.9 \\ 3.6 & -2.4 \end{pmatrix} \qquad B = \begin{pmatrix} 6.9 & -9.6 \\ -1 & 0.5 \end{pmatrix}$$

$$C = \begin{pmatrix} -5 & -9.6 & -1.6 \\ -4.5 & 8.6 & -5 \\ -6 & -0.6 & 3.4 \end{pmatrix} \qquad D = \begin{pmatrix} -4.9 & 7.8 & -0.6 \\ 7.5 & -6 & -8.7 \\ 4.7 & -4 & 9 \end{pmatrix}$$

Esercizio 4.7.4. Dati la matrice T e il vettore b,

$$T = \begin{pmatrix} 7 & 2 & 0.5 \\ 0 & 2 & 4 \\ 0 & 0 & -2 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix},$$

Calcolare il determinante di T e risolvere il sistema $Tx = b$.

Esercizio 4.7.5. Dati la matrice T e i vettori b_1 e b_2

$$T = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 2 & -4 \\ 0 & 0 & 6 \end{pmatrix}, \qquad b_1 = \begin{pmatrix} 1 \\ 1 \\ -12 \end{pmatrix}, \qquad b_2 = \begin{pmatrix} -7/3 \\ 1 \\ -4 \end{pmatrix}$$

Calcolare il determinante di T e risolvere i sistemi $Tx = b_1$ e $Tx = b_2$.

Esercizio 4.7.6. Calcolare, con il metodo di eliminazione di Gauss, la soluzione dei sistemi lineari $Ax = b$ e verificare di aver trovato la soluzione esatta, utilizzando le seguenti matrici e i seguenti vettori. Calcolare inoltre il determinante di A.

1.
$$A = \begin{pmatrix} 5 & -1 & 2 \\ 3 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

2.
$$A = \begin{pmatrix} 20 & -5 & -1 \\ 20 & -7 & -4 \\ -40 & 20 & 9 \end{pmatrix} \qquad b = \begin{pmatrix} -2 \\ 6 \\ -3 \end{pmatrix}$$

Esercizio 4.7.7. Calcolare, con il metodo di eliminazione di Gauss, l'inversa della matrice A

$$A = \begin{pmatrix} 3 & -1 & 2 \\ 12 & 1 & 0 \\ 2 & -1 & 2 \end{pmatrix}.$$

Esercizio 4.7.8. Date la matrice A e la sua inversa A^{-1},

$$A = \begin{pmatrix} 1 & 2 & 1 \\ -1 & 0 & -1 \\ 2 & 2 & 1 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} -1 & 0 & 1 \\ 0.5 & 0.5 & 0 \\ 1 & -1 & -1 \end{pmatrix},$$

calcolare la soluzione del sistema $Ax = b$, dato il vettore $b = (4,\ -2,\ 5)^t$; calcolare inoltre, con il metodo di Gauss, la soluzione del sistema $A(x + \delta x) = b + \delta b$, con $\delta b = (-1,\ -1,\ -1)^t$ e il valore

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty}.$$

Esercizio 4.7.9. Data la matrice A

$$A = \begin{pmatrix} 2 & 9 & 0 \\ -4 & 4 & 1 \\ 1 & 5 & 0 \end{pmatrix}.$$

risolvere, con il metodo di Gauss, i tre sistemi lineari $Ax^{(i)} = e_i$, $i = 1, \ldots, 3$, dove e_i è la i-esima colonna della matrice identità e $x^{(i)}$ è la i-esima colonna della matrice A^{-1}. Calcolare, inoltre, il numero di condizionamento della matrice A in norma 1.

3.

$$A = \begin{pmatrix} 1 & -2 & 3 \\ 2 & 2 & 3 \\ 0 & 1 & 1 \end{pmatrix} \quad b = \begin{pmatrix} -3 \\ 1 \\ 1 \end{pmatrix}$$

4.

$$A = \begin{pmatrix} -1 & -4 & 0 \\ 2 & 8 & 1 \\ 2 & -3 & 1 \end{pmatrix} \quad b = \begin{pmatrix} 2 \\ -4 \\ 7 \end{pmatrix}$$

5.

$$A = \begin{pmatrix} -1 & 8 & 0 \\ 2 & -3 & 1 \\ 2 & -16 & 0 \end{pmatrix} \quad b = \begin{pmatrix} 6 \\ 2 \\ 8 \end{pmatrix} \text{ oppure } \quad b = \begin{pmatrix} 6 \\ 2 \\ -12 \end{pmatrix}$$

6. $(*)$

$$A = \begin{pmatrix} 1 & -4 & 0 & 0 \\ 2 & 8 & 1 & 1 \\ 3 & -3 & 5 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \quad b = \begin{pmatrix} -1 \\ 4 \\ -7/2 \\ 3/2 \end{pmatrix}$$

Esercizio 4.7.10. Approssimare la soluzione del sistema lineare $Ax = b$, con A e b definiti nel seguito, utilizzando alcuni passi del metodo iterativo di Jacobi. Dire se (e perchè) il metodo converge e cercare di dedurre, dalle iterazioni fatte, quale sia la soluzione del sistema.

1.
$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & -1 \\ -1 & 0 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 5 \\ -3 \end{pmatrix} \quad \text{e} \quad x_0 = \begin{pmatrix} 7/4 \\ 7/4 \\ -7/4 \end{pmatrix}$$

2.
$$A = \begin{pmatrix} 4 & -1 & 0 \\ 1 & 4 & -2 \\ 0 & 1 & 2 \end{pmatrix}, b = \begin{pmatrix} 2 \\ 7 \\ 4 \end{pmatrix} \quad \text{e} \quad x_0 = \begin{pmatrix} 0.5 \\ 2.3 \\ 0.8 \end{pmatrix}$$

3.
$$A = \begin{pmatrix} -4 & 1 & 1 \\ 0 & 2 & 1 \\ -1 & -2 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 5 \\ -1 \end{pmatrix} \quad \text{e} \quad x_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

4.
$$A = \begin{pmatrix} 20 & -5 & -1 \\ -7 & 20 & -4 \\ 9 & 20 & -40 \end{pmatrix}, \qquad b = \begin{pmatrix} -2 \\ 6 \\ -3 \end{pmatrix} \quad \text{e} \quad x_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

5.
$$A = \begin{pmatrix} 8 & -3 & 2 \\ 0 & 5 & -1 \\ 0 & 0 & 3 \end{pmatrix}, \quad b = \begin{pmatrix} -7 \\ -4 \\ -3 \end{pmatrix} \quad \text{e} \quad x_0 = \begin{pmatrix} -1.5 \\ -0.5 \\ -0.5 \end{pmatrix}$$

Esercizio 4.7.11. Scegliere tra le due seguenti matrici

$$A = \begin{pmatrix} 5 & -1 & 2 \\ 1 & -4 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad B = \begin{pmatrix} -1 & 0 & 3 \\ 8 & 7 & -1 \\ -4 & -1 & 10 \end{pmatrix}$$

quella con norma 1 minore. Calcolare due iterate del metodo di Jacobi per approssimare la soluzione di un sistema lineare con la matrice prescelta come matrice dei coefficienti e termine noto $b = (6, -2, 1)^t$. Scegliere come vettore iniziale $x_0 = (0.5, 0.5, 1)^t$.

Esercizio 4.7.12. Calcolare $\|J\|_1$, dove J è la matrice di Jacobi costruita a partire dalla matrice B dell'esercizio precedente.

Esercizio 4.7.13. Dati la matrice A e il vettore b, dove

$$A = \begin{pmatrix} 2 & 1 & a \\ 2 & 4 & -1 \\ -1 & 0 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 13 \\ 4 \\ 7 \end{pmatrix},$$

scegliere quale valore dare al parametro a tra i seguenti $\{-2, 0.5, 1.5\}$ affinché sia soddisfatta la condizione sufficiente per la convergenza del metodo di Jacobi. Usando tale valore di a, calcolando solo 3 iterazioni e scegliendo, come vettore iniziale, $x_0 = (4, 1, 5)^T$, dire qual è la soluzione del sistema.

Esercizio 4.7.14. $(*)$ Dato il sistema $Ax = b$, dove:

$$A = \begin{pmatrix} -3 & 1 & \alpha^2 \\ -1 & 2 & 0.5 \\ 1 & -0.5 & 2 \end{pmatrix} \quad b = \begin{pmatrix} 4 \\ 3 \\ -1.5 \end{pmatrix}$$

trovare un valore di α affinché sia soddisfatta la condizione sufficiente per la convergenza del metodo di Jacobi e (contemporaneamente) la norma infinito della matrice J sia minima. Inoltre, scelto un valore di α fra quelli che soddisfano le condizioni precedenti, calcolare due passi del metodo di Jacobi a partire dal vettore iniziale $x_0 = (-0.5\ 0.5\ 0.5)^t$.

Esercizio 4.7.15. Verificare se le eseguenti matrici sono invertibili e, se possibile, calcolarne l'inversa con il metodo di Gauss.

$$A = \begin{pmatrix} 7 & -2 & -3 \\ -21 & -2 & 14 \\ -28 & -16 & 25 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ 1 & -1 & 1 \end{pmatrix} \qquad C = \begin{pmatrix} 10 & 3 & -2 \\ 1 & -2 & 4 \\ 8 & 7 & -10 \end{pmatrix}$$

Numero di condizionamento

Esercizio 4.7.16. Data una matrice A, con det $\neq 0$, dire quali delle seguenti affermazioni sono vere:

- $K(A) = \|A\|_1 \|A^{-1}\|_\infty$
- $K(A)_\infty = \|A\|_\infty \|A^{-1}\|_\infty$
- $K(A)$ permette di valutare il condizionamento del sistema lineare da risolvere

- $K(A) < 0.5$.

Se $K(A) = 156$ si ha che:

- il sistema $Ax = b$ è ben condizionato e la matrice A è mal condizionata
- la matrice è mal condizionata
- la matrice è stabile

Esercizio 4.7.17. Sia A una matrice con numero di condizionamento in norma 1 pari a $K_1(A) = 13$. Supponendo di perturbare il termine noto b del sistema lineare $Ax = b$ con una perturbazione relativa in norma minore di 10^{-2}, stimare la perturbazione relativa della soluzione x.

Esercizio 4.7.18. Calcolare il numero di condizionamento delle seguenti matrici, usando sia la norma 1 che la norma infinito. Utilizzare il metodo di Gauss per calcolare l'inversa delle matrici.

$$A = \begin{pmatrix} 1 & 10 \\ 5 & 8 \end{pmatrix} \qquad B = \begin{pmatrix} -1 & 2 \\ 3 & 50 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 4 \\ 2 & -1 & 1 \end{pmatrix} \qquad D = \begin{pmatrix} 3 & -2 & 1 \\ 2 & 10 & 1 \\ -4 & 6 & 0 \end{pmatrix}$$

Esercizio 4.7.19. $(*)$ Data la matrice:

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & 1 \\ 3 & \alpha & 4 \end{pmatrix},$$

con $2 \leq \alpha \leq 4$, calcolare $K_\infty(A)$ e il valore di $\alpha \in [2, 4]$ che lo minimizza. Usare il metodo di Gauss per calcolare l'inversa della matrice A.

Esercizio 4.7.20. Calcolare, usando il metodo di Gauss, l'inversa della matrice A

$$A = \begin{pmatrix} 0 & -1 & 2 \\ 1 & 2 & 5 \\ 1 & 1 & 1 \end{pmatrix},$$

e calcolare il numero di condizionamento di A in norma 1. Dati, inoltre, $b = (1, 1, 1)^t$ e $\|\delta b\|_1 = .001$, trovare una maggiorazione di $\frac{\|\delta x\|_1}{\|x\|_1}$, (senza calcolare esplicitamente x e δx), dove x e δx sono, rispettivamente, soluzione di $Ax = b$ e $A(x + \delta x) = b + \delta b$.

Esercizio 4.7.21. $(*)$ Data la matrice A_α, dipendente dal parametro α, $0 < \alpha \leq 2$,

$$A_\alpha = \begin{pmatrix} 1 & 1 \\ 0 & \alpha \end{pmatrix}$$

trovare il valore di α affinché il condizionamento di A_α sia minimo in norma 1. Per tale valore di α risolvere i sistemi

$$A_\alpha x = b \text{ e } A_\alpha \tilde{x} = \tilde{b}$$

con

$$b = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad \tilde{b} = \begin{pmatrix} 3.01 \\ 1.01 \end{pmatrix}$$

Calcolare direttamente il rapporto $\frac{\|\tilde{x}-x\|_1}{\|x\|_1}$ e maggiorare tale valore con la stima teorica.

Soluzione esercizi proposti

4.7.1

(1) $AB = \begin{pmatrix} 2 \\ 7 \\ 3 \end{pmatrix}$ BA non calcolabile

(2) $AB = \begin{pmatrix} -1 & -5 & 9 \\ 5.5 & -6.5 & 5 \\ 7 & -9 & 8 \end{pmatrix}$ $BA = \begin{pmatrix} -4 & 6 & 4 \\ 1 & 5.5 & -3 \\ 1 & -3.5 & -1 \end{pmatrix}$

(3) $AB = \begin{pmatrix} 2 & 4 & 1 \\ -4 & 2 & -3 \end{pmatrix}$ BA non calcolabile

(4) AB non calcolabile $BA = \begin{pmatrix} 9 & 1 \\ -4 & 2 \\ -3 & 3 \end{pmatrix}$

(5) $AB = \begin{pmatrix} 2 & -6 \\ -10 & 11 \end{pmatrix}$ $BA = \begin{pmatrix} 9 & -14 & -5 \\ -5 & 6 & 1 \\ 1 & -3 & -2 \end{pmatrix}$

(6) AB non calcolabile $BA = \begin{pmatrix} -1 & -2 \\ 25 & 58 \\ -4 & -15 \end{pmatrix}$

(7) $AB = \begin{pmatrix} 0 & -4 \\ -5 & -9 \\ -3 & -3 \end{pmatrix}$ BA non calcolabile

4.7.2

$$\begin{array}{lll} \|x\|_1 = 16 & \|x\|_2 \simeq 11.40 & \|x\|_\infty = 9 \\ \|y\|_1 = 33 & \|y\|_2 \simeq 20.42 & \|y\|_\infty = 17 \\ \|z\|_1 = 34 & \|z\|_2 \simeq 17.32 & \|z\|_\infty = 11 \\ \|t\|_1 = 13 & \|t\|_2 \simeq 8.12 & \|t\|_\infty = 6.4 \end{array}$$

4.7.3

$$\begin{array}{lll} \|A\|_1 = 9.7 & \|A\|_\infty = 10 & \|A\|_F \simeq 8.43 \\ \|B\|_1 = 10.1 & \|B\|_\infty = 16.5 & \|B\|_F \simeq 11.87 \\ \|C\|_1 = 18.8 & \|C\|_\infty = 18.1 & \|C\|_F \simeq 16.93 \\ \|D\|_1 = 18.3 & \|D\|_\infty = 21.3.2 & \|D\|_F \simeq 19.29 \end{array}$$

4.7.4 $\det T = -28$ $x = [17/28,\ -5/2,\ 3/2]^t$.

4.7.5 $\det T = 0$. il primo sistema non ha soluzioni. Il secondo sistema ha infinite soluzioni del tipo $x = [x_1\ -5/6,\ -2/3]^t$.

$$\begin{array}{ll} K_1(A) = 39/7 & K_F(A) = 95/21 \\ K_1(B) = 689/14 & K_F(B) = 1257/28 \\ K_1(C) = 252.7 & K_F(C) = 250.76 \end{array}$$

$$T^{-1} = \begin{bmatrix} -0.5 & 0 & 0 \\ 0.25 & 0.25 & 0 \\ 6 & -1 & 2 \end{bmatrix}.$$

4.7.6

(1) $\det A = 4$ $\quad x = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$

(2) $\det A = 320$ $\quad x = \begin{pmatrix} 77/320 \\ 35/16 \\ -33/8 \end{pmatrix}$

(3) $\det A = 9$ $\quad x = \begin{pmatrix} -4/9 \\ 10/9 \\ -1/9 \end{pmatrix}$

(4) $\det A = 11$ $\quad x = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}$

(5) $\det A = 0$ il primo sistema non ha soluzioni,
il secondo ne ha infinite del tipo $x = \begin{pmatrix} -\frac{8}{13}x_3 + \frac{34}{13} \\ \frac{14}{13} - \frac{1}{13}x_3 \\ x_3 \end{pmatrix}$

(6∗) $\det A = -90$ $\quad x = \begin{pmatrix} 1 \\ \frac{1}{2} \\ -1 \\ 0 \end{pmatrix}$

4.7.7 Le matrici sono tutte a predominanza diagonale e quindi il metodo

converge in ogni caso.

$$(1)\quad x_1 = \begin{pmatrix} 9/8 \\ 13/8 \\ -5/8 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 19/16 \\ 35/16 \\ -15/16 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$$

$$(2)\quad x_1 = \begin{pmatrix} 43/40 \\ 81/40 \\ 17/20 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 161/160 \\ 61/32 \\ 79/80 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

$$(3)\quad x_1 = \begin{pmatrix} 3/4 \\ 2 \\ 1/2 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 7/8 \\ 9/4 \\ 15/16 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

$$(4)\quad x_1 = \begin{pmatrix} -1/10 \\ 3/10 \\ 3/40 \end{pmatrix} \qquad x_2 = \begin{pmatrix} -17/800 \\ 7/25 \\ 81/400 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} 0 \\ 7/20 \\ 1/4 \end{pmatrix}$$

$$(5)\quad x_1 = \begin{pmatrix} -15/16 \\ -9/10 \\ -1 \end{pmatrix} \qquad x_2 = \begin{pmatrix} -77/80 \\ -1 \\ -1 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$$

4.7.8 Si sceglie la matrice A.

$$x_1 = \begin{pmatrix} 9/10 \\ 7/8 \\ 1 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 39/40 \\ 39/40 \\ 1 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

4.7.9 $\|J\|_1 = \frac{22}{7}$.

4.7.10 $a = 0.5$.

$$x_1 = \begin{pmatrix} 19/4 \\ 1/4 \\ 11/2 \end{pmatrix} \quad x_2 = \begin{pmatrix} 5 \\ 0 \\ 47/8 \end{pmatrix} \quad x_3 = \begin{pmatrix} 161/32 \\ -1/32 \\ 6 \end{pmatrix} \quad \to_{k\to\infty} \begin{pmatrix} 5 \\ 0 \\ 6 \end{pmatrix}$$

4.7.11 $a = 0$.

$$x_1 = \begin{pmatrix} -7/6 \\ 9/8 \\ -3/8 \end{pmatrix} \qquad x_2 = \begin{pmatrix} -23/24 \\ 97/96 \\ 11/96 \end{pmatrix} \qquad \to_{k\to\infty} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

4.7.12

$$\begin{aligned} A^{-1} &= \begin{pmatrix} 87/56 & 49/56 & -17/56 \\ 19/16 & 13/16 & -5/16 \\ 5/2 & 3/2 & -1/2 \end{pmatrix} \\ B^{-1} &= \begin{pmatrix} 3/5 & 1/5 & 2/5 \\ 1/5 & 2/5 & -1/5 \\ -2/5 & 1/5 & 2/5 \end{pmatrix} \\ C &= \text{non invertibile} \end{aligned}$$

4.7.13 F V V F F - dipende dalla dimensione di A - F.

4.7.14 $\frac{\|\delta x\|_1}{\|x\|_1} \leq 13 \cdot 10^{-2}$.

4.7.15

$$A^{-1} = \frac{1}{21}\begin{pmatrix} -8 & 10 \\ 5 & -1 \end{pmatrix} \qquad K_1(A) = \frac{117}{21} \quad K_\infty(A) = \frac{117}{21}$$

$$B^{-1} = \frac{1}{56}\begin{pmatrix} -50 & 2 \\ 3 & 1 \end{pmatrix} \qquad K_1(B) = \frac{689}{14} \quad K_\infty(B) = \frac{689}{14}$$

$$C^{-1} = \frac{1}{14}\begin{pmatrix} 4 & -4 & 4 \\ 9 & -1 & -5 \\ 1 & 3 & 1 \end{pmatrix} \qquad K_1(C) = 6 \quad K_\infty(C) = \frac{75}{14}$$

$$D^{-1} = \frac{1}{42}\begin{pmatrix} -6 & 6-12 & \\ -4 & 4 & -1 \\ 52 & = 10 & 34 \end{pmatrix} \qquad K_1(D) = \frac{187}{7} \quad K_\infty(D) = \frac{208}{7}$$

4.7.16 $(*)$

$$A^{-1} = \frac{1}{3\alpha+5}\begin{pmatrix} -\alpha & 2\alpha+4 & -1 \\ -5 & -2 & 3 \\ 2\alpha & -\alpha-3 & 2 \end{pmatrix} \qquad K_\infty(A) = 7+\alpha$$

Per $\alpha \in [2, 4]$ il numero di condizionamento è minimo se $\alpha = 2$.

4.7.17

$$A^{-1} = \frac{1}{6}\begin{pmatrix} 3 & -3 & 9 \\ -4 & 2 & -2 \\ 1 & 1 & -1 \end{pmatrix} \qquad K_1(A) = 16 \qquad \frac{\|\delta x\|_1}{\|x\|_1} \leq \frac{16}{3000}$$

4.7.18$(*)$

$$A_\alpha^{-1} = \begin{pmatrix} 1 & -\frac{1}{\alpha} \\ 0 & \frac{1}{\alpha} \end{pmatrix} \qquad K_1(A) = 3$$

$$x = \begin{pmatrix} 2.5 \\ 0.5 \end{pmatrix} \qquad \widetilde{x} = \begin{pmatrix} 2.5050 \\ 0.5050 \end{pmatrix} \qquad \frac{\|\delta x\|_1}{\|x\|_1} = \frac{1}{300} \leq \frac{3}{200}$$

Chapter 5

Sistemi sovradeterminati e soluzione ai minimi quadrati

Sia A una matrice $m \times n$ e b un vettore di m componenti, con $m > n$. Il sistema lineare $Ax = b$ ha n incognite e $m > n$ equazioni e viene detto sistema **sovradeterminato**. In generale non esiste un vettore $x \in \mathbb{R}^n$ tale che Ax sia uguale a b come evidenziato dal seguente esempio. Siano A e b tali che

$$A = \begin{pmatrix} 2 & 4 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

e si consideri il sistema $Ax = b$. Dall'ultima equazione, per soddisfare l'uguaglianza, si ha $x_3 = 3$, dalla seconda $x_2 = 2$, ma se si sostituiscono tali valori nella prima equazione l'uguaglianza non è verificata perché si ottiene $6 = 1$.

Si introduce quindi il concetto di **soluzione ai minimi quadrati**. Dato il sistema lineare $Ax = b$, con $A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$, $m > n$, si calcola il vettore $x \in \mathbb{R}^n$ tale che

$$\|Ax - b\|_2^2 = \min_{z \in \mathbb{R}^n} \|Az - b\|_2^2$$

Il vettore x è la soluzione ai minimi quadrati del sistema sovradeterminato $Ax = b$. Tale soluzione può essere calcolata risolvendo le equazioni normali, cioè calcolando la soluzione del sistema lineare $A^tAx = A^tb$, che ha n equazioni e n incognite. Essendo un sistema quadrato può essere risolto utilizzando, ad esempio, il metodo di Gauss.

Calcolata la soluzione x ai minimi quadrati, il vettore $\rho = Ax - b$ viene detto **residuo** e si dimostra che $A^t\rho$ è il vettore nullo. Dal punto di vista

didattico tale proprietà può essere sfruttata per verificare la correttezza dei risultati ottenuti.

Retta di regressione. Dato un insieme di s punti $(x_1, y_1), \ldots, (x_s, y_s)$, si vuole trovare la retta che passa più vicino ai punti, cioè si vuole trovare la retta $y = mx + q$ tale che $\sum_{i=1}^{s}(mx_i + q - y_i)^2$ è minima. Tale problema si risolve impostando il sistema sovradeterminato

$$Ax = b \qquad \Longleftrightarrow \qquad \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix} \begin{pmatrix} q \\ m \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}$$

e calcolando la soluzione delle equazioni normali

$$A^tAx = A^tb \quad \Longleftrightarrow \quad \begin{pmatrix} s & \sum_{i=1}^{s} x_i \\ \sum_{i=1}^{s} x_i & \sum_{i=1}^{s} x_i^2 \end{pmatrix} \begin{pmatrix} q \\ m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{s} y_i \\ \sum_{i=1}^{s} x_i y_i \end{pmatrix}$$

Parabola ai minimi quadrati. Analogamente al caso della retta di regressione, dato un insieme di s punti $(x_1, y_1), \ldots, (x_s, y_s)$, si vuole trovare la parabola che passa più vicino ai punti, cioè si vuole trovare la parabola $y = a_2x^2 + a_1x + a_0$ tale che $\sum_{i=1}^{s}(a_2x_i^2 + a_1x_i + a_0 - y_i)^2$ è minima. Tale problema si risolve impostando il sistema sovradeterminato

$$Ax = b \qquad \Longleftrightarrow \qquad \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_s & x_s^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix}$$

e calcolando la soluzione delle equazioni normali

$$A^tAx = A^tb \iff \begin{pmatrix} s & \sum_{i=1}^{s} x_i & \sum_{i=1}^{s} x_i^2 \\ \sum_{i=1}^{s} x_i & \sum_{i=1}^{s} x_i^2 & \sum_{i=1}^{s} x_i^3 \\ \sum_{i=1}^{s} x_i^2 & \sum_{i=1}^{s} x_i^3 & \sum_{i=1}^{s} x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{s} y_i \\ \sum_{i=1}^{s} x_i y_i \\ \sum_{i=1}^{s} x_i^2 y_i \end{pmatrix}$$

5.1 Sistemi sovradeterminati

Esercizio 5.1.1. Sistema sovradeterminato.
Risolvere il sistema sovradeterminato $Ax = b$ con

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 0 & -1 \\ 0 & -2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ -2 \\ 0 \\ 1 \end{pmatrix} \quad \text{e} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

e calcolare la norma del residuo ρ.

La soluzione ai minimi quadrati di un sistema sovrdeterminato $Ax = b$ si calcola risolvendo le equazioni normali, cioè $A^tAx = A^tb$. In questo caso si ha

$$\begin{aligned} A^tA &= \begin{pmatrix} 2 & 1 & 0 & 1 \\ -1 & 0 & -2 & 1 \\ 1 & -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 1 \\ 1 & 0 & -1 \\ 0 & -2 & 3 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 6 & -1 & 2 \\ -1 & 6 & -6 \\ 2 & -6 & 12 \end{pmatrix} \\ A^tb &= \begin{pmatrix} 2 & 1 & 0 & 1 \\ -1 & 0 & -2 & 1 \\ 1 & -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} \end{aligned}$$

Risolvendo con il metodo di Gauss il sistema $A^tAx = A^tb$ si ha

$$\begin{aligned} [A^tA \,|\, A^tb] &= \left(\begin{array}{ccc|c} 6 & -1 & 2 & 1 \\ -1 & 6 & -6 & 0 \\ 2 & -6 & 12 & 4 \end{array}\right) \qquad r_2 + \frac{1}{6}r_1 \quad r_3 - \frac{1}{3}r_1 \\ &= \left(\begin{array}{ccc|c} 6 & -1 & 2 & 1 \\ 0 & 35/6 & -17/3 & 1/6 \\ 0 & -17/3 & 34/3 & 11/3 \end{array}\right) \qquad r_3 - \frac{34}{35}r_2 \\ &= \left(\begin{array}{ccc|c} 6 & -1 & 2 & 1 \\ 0 & 35/6 & -17/3 & 1/6 \\ 0 & 0 & 204/35 & 134/35 \end{array}\right) \end{aligned}$$

Da cui si ottiene

$$x = \frac{1}{102} \begin{pmatrix} 6 \\ 68 \\ 67 \end{pmatrix}$$

Poiché x è la soluzione delle equazioni normali, si ha che $A^tAx = A^tb$, cioè $A^tAx - A^tb = 0$. Tuttavia, essendo x la soluzione ai minimi quadrati del sistema sovradeterminato $Ax = b$ si ha che $Ax - b \neq 0$. Si definisce residuo il vettore $\rho = Ax - b$. In questo caso si ha che

$$\rho = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 0 & -1 \\ 0 & -2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 6/102 \\ 68/102 \\ 67/102 \end{pmatrix} - \begin{pmatrix} 1 \\ -2 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 11/102 & -1 \\ -61/102 & +2 \\ 65/102 & -0 \\ 141/102 & -1 \end{pmatrix} = \begin{pmatrix} -91/102 \\ 143/102 \\ 65/102 \\ 39/102 \end{pmatrix}$$

Con semplici calcoli si può verificare che $A^t\rho = 0$.

♣

Esercizio 5.1.2. Elaborazione dati sperimentali.

(Esempio tratto da: Dalhquist-Bjork, Numerical Methods.)

Usando i pesi molecolari dei sei ossidi di azoto seguenti, calcolare i pesi atomici di azoto e ossigeno con 4 cifre decimali:

NO	N_2O	NO_2	N_2O_3	N_2O_5	N_2O_4
30.006	44.013	46.006	76.012	108.010	92.011

Indicando con n il peso atomico dell'azoto e con o il peso atomico dell'ossigeno, dai dati precedenti si ricava il sistema sovradeterminato $Ax = b$ tale che

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2 \\ 2 & 3 \\ 2 & 5 \\ 2 & 4 \end{pmatrix} \begin{bmatrix} n \\ o \end{bmatrix} = \begin{bmatrix} 30.006 \\ 44.013 \\ 46.006 \\ 76.012 \\ 108.010 \\ 92.011 \end{bmatrix}.$$

Passando alle equazioni normali, considerando cioè il sistema quadrato $A^TAx = A^Tb$, si ottiene

$$\begin{pmatrix} 18 & 29 \\ 29 & 56 \end{pmatrix} \begin{bmatrix} n \\ o \end{bmatrix} = \begin{bmatrix} 716.104 \\ 1302.161 \end{bmatrix},$$

la cui soluzione fornisce i valori $n = 14.0069$ (valore vero 14.00674) e 15.9993 (valore vero 15.9994). Il residuo $\rho = Ax - b$ è

$$\rho = 10^{-3} \cdot \begin{pmatrix} 0.2096 \\ 0.1257 \\ -0.4970 \\ -0.2874 \\ 0.2994 \\ 0.0060 \end{pmatrix} \qquad \text{che ha norma pari a} \quad \|\rho\|_2 = 6.9213 \cdot 10^{-4}$$

Si noti che un numero minore di equazioni porta al calcolo di valori meno accurati. Se, ad esempio si considera il sistema

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix} \begin{bmatrix} n \\ o \end{bmatrix} = \begin{bmatrix} 30.006 \\ 44.013 \end{bmatrix},$$

si ottengono $n = 14.007$ e $o = 15.999$.

♣

5.2 Approssimazione polinomiale.

Esercizio 5.2.1. Retta di regressione.

Determinare la retta di regressione per i dati:

x	1	2	3	4	5
y	0.5	0	-0.5	1	0.75

Si vuole trovare la retta $y = mx + q$ i cui coefficienti m e q sono soluzione ai minimi quadrati del sistema sovradeterminato $y_i = q + mx_i$, $i = 1, \ldots, 5$. Si ottiene allora

$$Ax = b \quad \Longleftrightarrow \quad \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{bmatrix} q \\ m \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ -0.5 \\ 1 \\ 0.75 \end{bmatrix}.$$

Passando alle equazioni normali si ottiene il sistema quadrato non singolare

$$A^t Ax = A^t b \quad \Longleftrightarrow \quad \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix} \begin{bmatrix} q \\ m \end{bmatrix} = \begin{bmatrix} 1.75 \\ 6.75 \end{bmatrix}.$$

la cui soluzione è data da $q = -0.1$ e $m = 0.15$; il grafico della retta di regressione $y = 0.15x - 0.1$ e dei punti che essa approssima è riportato in figura 5.1.

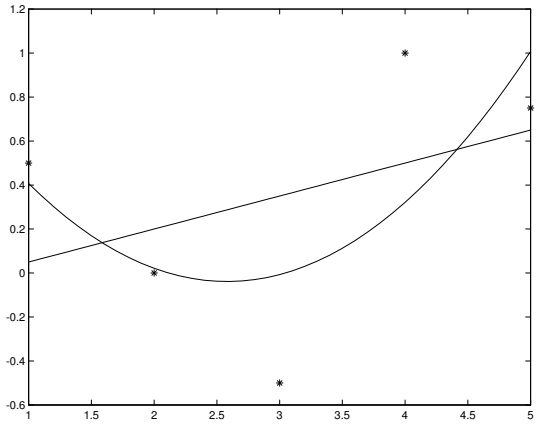

Figure 5.1: Retta di regrassione e parabola ai minimi quadrati

♣

Esercizio 5.2.2. Approssimazione parabolica. Determinare la parabola dei minimi quadrati per i dati (gli stessi dell'esercizio precedente):

x	1	2	3	4	5
y	0.5	0	-0.5	1	0.75

Si vuole trovare la parabola $y = a_2x^2 + a_1x + a_0$ i cui coefficienti a_0, a_1 e a_2 sono soluzione ai minimi quadrati del sistema sovradeterminato la cui i-esima equazione è data da $y_i = a_0 + a_1x_i + a_2x_i^2$, $i = 1, \ldots, 5$. Si ottiene allora

$$Ax = b \quad \Longleftrightarrow \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{pmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ -0.5 \\ 1 \\ 0.75 \end{bmatrix}.$$

Passando alle equazioni normali si ottiene il sistema quadrato non singolare

$$A^tx = A^tb \quad \Longleftrightarrow \quad \begin{pmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1.75 \\ 6.75 \\ 30.75 \end{bmatrix}.$$

la cui soluzione è data da $a_0 = 23/20$, $a_1 = -129/140$ e $a_2 = 4895/27412$. Il grafico della parabola dei minimi quadrati $y = 4895/27412x^2 - 129/140x + 23/20$ e dei punti approssimati è riportato in figura 5.1.

5.3 Esercizi proposti

Esercizio 5.3.1. Dati una matrice A $m \times n$, $m > n$ e un vettore $b \in \mathbb{R}^m$, si consideri il sistema sovradeterminato $Ax = b$. Dire se sono vere le seguenti affermazioni:
a) il residuo è sempre nullo; b) il residuo è dato da $\|x\|$; c) il residuo è dato da $Ax - b$.

Inoltre, risolvere le equazioni normali relative al problema ai minimi quadrati $Ax = b$, dove

$$A = \begin{pmatrix} -1 & -1 \\ 0 & 3 \\ 1 & 2 \\ -1 & -1 \end{pmatrix} \quad b = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

Esercizio 5.3.2. Siano dati la matrice A e il vettore b, dove

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 0 & 1 \\ 3 & 0 & -2 \\ 0 & 1 & 2 \\ 0 & 1 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 6 \\ -3 \\ -2 \\ 1 \\ 1 \end{pmatrix}.$$

Trovare la soluzione x ai minimi quadrati del problema $Ax = b$ e calcolare la norma del residuo $\|Ax - b\|_2$.

Esercizio 5.3.3. Dato il sistema sovradeterminato $Ax = b$, dove

$$A = \begin{pmatrix} 1 & 5 \\ 2 & 1 \\ -3 & -1 \\ 0 & 3 \\ 1 & 0 \end{pmatrix} \quad b = \begin{pmatrix} -2 \\ 4 \\ -8 \\ -2 \\ 2 \end{pmatrix},$$

calcolarne la soluzione ai minimi quadrati e calcolare la norma 2 del residuo.

Esercizio 5.3.4. Sia dato il sistema sovradeterminato $Ax = b$,

$$A = \begin{pmatrix} 1 & -0.5 \\ 2 & -1 \\ -1 & 0 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad b = \begin{pmatrix} \alpha \\ 0 \\ -1 \end{pmatrix}.$$

Determinare α in modo tale che, risolvendo il sistema con i minimi quadrati, la norma del residuo $\|Ax - b\|_2$ sia nulla.

Esercizio 5.3.5. Dato il seguente sistema sovradeterminato $Ax = b$:

$$\begin{pmatrix} 0 & \alpha \\ -1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

trovare il valore di $\alpha \in [0, 1]$ affinché il condizionamento di A^tA sia minimo in norma 1. Inoltre, per tale valore di α risolvere il problema ai minimi quadrati $Ax = b$.

Esercizio 5.3.6. Dati i punti $P_0 = (1, 1)$, $P_1 = (1.5, 2)$, $P_2 = (2, 1)$, $P_3 = (2.5, 3)$, $P_4 = (3, 2.5)$, $P_5 = (4, 4.5)$ trovare la retta di regressione.

Esercizio 5.3.7. Dati i punti $(1, 1)$, $(2, 0.5)$, $(3, 1)$, $(4, \alpha)$, calcolare per quali valori di α la retta di regressione individuata da tali punti ha coefficiente angolare positivo e interseca l'asse y in un punto di ordinata positiva.

Esercizio 5.3.8. Determinare la parabola dei minimi quadrati per i dati:

x	-1	0	1	2
y	-5.5	-0.5	-2	-8.5

Esercizio 5.3.9. A partire dai dati:

x	1	1.5	2	2.5	3
y	1	5	16	20	80

si calcolino

- la retta di regressione;
- la parabola dei minimi quadrati;
- la cubica dei minimi quadrati (utilizzando ad esempio la funzione polyfit di MatLab);
- si confonti il grafico delle 3 curve precedenti discutendone il comportamento (utilizzando ad esempio la funzione plot di MatLab).

Soluzione esercizi proposti

5.3.1 a) Falso, b) Falso, c) Vero. $x = (-39/29,\ 22/29)^t$.

5.3.2 $x = (1,\ -1,\ 1)^t$; $\|Ax - b\|_2 = 6$.

5.3.3 $x = (149/55,\ -19/22)^t$; $\|Ax - b\|_2 \simeq 1.3618$.

5.3.4 Soluzione ai minimi quadrati $x = (1,\ 2 - (2/5)\alpha)^t$, residuo $\rho = (-(4/5)\alpha,\ (2/5)\alpha,\ 0)^t$, $\alpha = 0$.

5.3.5 $K_1(A^tA) = \frac{25}{2+3\alpha^2}$, $\alpha = 1$ e $x = (-0.6,\ 1.4)^t$.

5.3.6 Retta di regressione: $y = -0.2 + 38/35x$.

5.3.7 Per $5/6 < \alpha < 5/2$.

5.3.8 Parabola: y $= -2.875x^2 + 1.8250x - 0.725$.

5.3.9 Retta: $y = -44.8 + 34.6x$. Parabola: $y = 30x^2 - 85.4x + 60.2$. Cubica: $y = 32.6x^3 - 166x^2 + 278.83x - 145.6$.