Elementi di Analisi Numerica

Carlo Garoni

30 novembre 2024

Indice

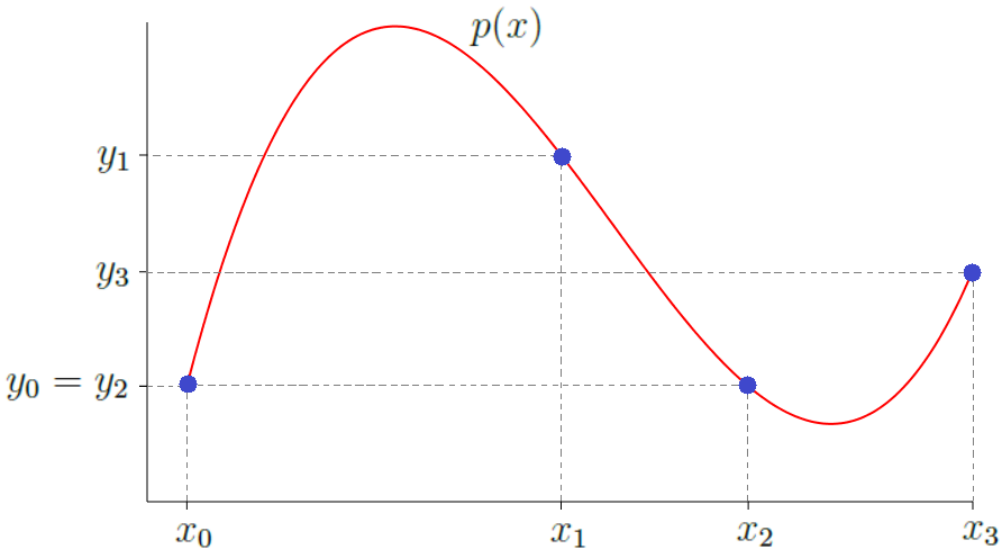

Figura 1.1: Illustrazione del Teorema 1.1 nel caso $n = 3$: esiste un unico polinomio $p(x) \in \mathbb{R}_3[x]$ tale che $p(x_0) = y_0$, $p(x_1) = y_1$, $p(x_2) = y_2$, $p(x_3) = y_3$.

1 Interpolazione polinomiale

1.1 Esistenza, unicità, forma canonica e forma di Lagrange del polinomio d'interpolazione

È data una funzione $f : [a, b] \to \mathbb{R}$ di cui sono noti i valori $f(x_0), f(x_1), \ldots, f(x_n)$ negli $n + 1$ punti distinti $x_0, x_1, \ldots, x_n \in [a, b]$. Si sceglie una classe $\mathcal{C}$ di funzioni definite su $[a, b]$ a valori in $\mathbb{R}$ e si vuole approssimare la funzione $f(x)$ con una funzione $p : [a, b] \to \mathbb{R}$ che appartiene alla classe $\mathcal{C}$ e che nei punti $x_0, x_1, \ldots, x_n$ assume i valori $f(x_0), f(x_1), \ldots, f(x_n)$. Una scelta comune per la sua semplicità è quella di prendere $\mathcal{C}$ come lo spazio vettoriale (reale) dei polinomi di grado $\leq n$, ovvero

$$\mathcal{C} = \mathbb{R}_n[x] = \{a_0 + a_1x + a_2x^2 + \ldots + a_nx^n : a_0, a_1, a_2, \ldots, a_n \in \mathbb{R}\}.$$

In questo caso, esiste un unico polinomio $p(x) \in \mathbb{R}_n[x]$ tale che $p(x_i) = f(x_i)$ per ogni $i = 0, \ldots, n$. Questo fatto è conseguenza del Teorema 1.1.

Teorema 1.1. *Siano $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ tali che $x_0, x_1, \ldots, x_n$ sono tutti distinti. Allora esiste un unico polinomio $p(x) \in \mathbb{R}_n[x]$ tale che $p(x_i) = y_i$ per ogni $i = 0, \ldots, n$.*

La Figura 1.1 illustra l'enunciato del Teorema 1.1.

Dimostrazione. Diamo due dimostrazioni di questo teorema, entrambe istruttive.

1. Un polinomio $p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n \in \mathbb{R}_n[x]$ soddisfa la proprietà che $p(x_i) = y_i$ per ogni $i = 0, \ldots, n$ se e solo se

$$\begin{cases} a_0 + a_1x_0 + a_2x_0^2 + \ldots + a_nx_0^n = y_0 \\ a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n = y_1 \\ a_0 + a_1x_2 + a_2x_2^2 + \ldots + a_nx_2^n = y_2 \\ \ldots \\ a_0 + a_1x_n + a_2x_n^2 + \ldots + a_nx_n^n = y_n \end{cases}$$

cioè se e solo se il suo vettore dei coefficienti $(a_0, a_1, \ldots, a_n)^T$ soddisfa il sistema lineare

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{1.1}$$

La matrice (quadrata)

$$\boxed{V(x_0, x_1, \ldots, x_n) = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}}$$

si chiama *matrice di Vandermonde sui nodi* $x_0, \ldots, x_n$ ed è invertibile perché tra poco dimostreremo che

$$\det[V(x_0, \ldots, x_n)] = \begin{cases} 1 & \text{se } n = 0, \\ \displaystyle\prod_{\substack{i,j=0 \\ j<i}}^{n} (x_i - x_j) = (x_1 - x_0) \cdot (x_2 - x_0)(x_2 - x_1) \cdot \cdots \cdot (x_n - x_0) \cdots (x_n - x_{n-1}) & \text{se } n \geq 1, \end{cases} \tag{1.2}$$

da cui segue che $\det[V(x_0, \ldots, x_n)] \neq 0$ in quanto $x_0, \ldots, x_n$ sono distinti per ipotesi. Quindi esiste un'unica soluzione del sistema (1.1), e questa soluzione è

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = [V(x_0, \ldots, x_n)]^{-1} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{1.3}$$

Ciò significa che esiste un unico polinomio $p(x) \in \mathbb{R}_n[x]$ tale che $p(x_i) = y_i$ per ogni $i = 0, \ldots, n$, e questo polinomio è precisamente quello che ha il vettore dei coefficienti dato dalla (1.3).

Per concludere la dimostrazione ci resta solo da dimostrare la (1.2) e lo facciamo nel caso $n = 3$.[1] Per $i = 1, \ldots, 3$ definiamo $d_i = \det[V(x_0, \ldots, x_i)]$. Il nostro obiettivo è quello di calcolare $d_3 = \det[V(x_0, \ldots, x_3)]$. Per $j = 4, \ldots, 2$, sottraiamo dalla colonna j di $V(x_0, \ldots, x_3)$ la colonna $j-1$ moltiplicata per x_3, dopodiché sviluppiamo il determinante lungo l'ultima riga usando il metodo di Laplace (si veda la Sezione 3.1.1). In questo modo otteniamo

$$d_3 = \begin{vmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \end{vmatrix} = \begin{vmatrix} 1 & x_0 & x_0^2 & x_0^3 - x_0^2 x_3 \\ 1 & x_1 & x_1^2 & x_1^3 - x_1^2 x_3 \\ 1 & x_2 & x_2^2 & x_2^3 - x_2^2 x_3 \\ 1 & x_3 & x_3^2 & 0 \end{vmatrix} = \begin{vmatrix} 1 & x_0 & x_0^2 - x_0 x_3 & x_0^2(x_0 - x_3) \\ 1 & x_1 & x_1^2 - x_1 x_3 & x_1^2(x_1 - x_3) \\ 1 & x_2 & x_2^2 - x_2 x_3 & x_2^2(x_2 - x_3) \\ 1 & x_3 & 0 & 0 \end{vmatrix}$$

[1] La (1.2) vale sicuramente se $n = 0$ perché $V(x_0) = [1]$; la dimostrazione nel caso generale $n \geq 1$ è del tutto analoga a quella che facciamo noi per $n = 3$.

$$= \begin{vmatrix} 1 & x_0 - x_3 & x_0(x_0 - x_3) & x_0^2(x_0 - x_3) \\ 1 & x_1 - x_3 & x_1(x_1 - x_3) & x_1^2(x_1 - x_3) \\ 1 & x_2 - x_3 & x_2(x_2 - x_3) & x_2^2(x_2 - x_3) \\ 1 & 0 & 0 & 0 \end{vmatrix} = (-1)^3 \begin{vmatrix} x_0 - x_3 & x_0(x_0 - x_3) & x_0^2(x_0 - x_3) \\ x_1 - x_3 & x_1(x_1 - x_3) & x_1^2(x_1 - x_3) \\ x_2 - x_3 & x_2(x_2 - x_3) & x_2^2(x_2 - x_3) \end{vmatrix}$$

$$= (-1)^3(x_0 - x_3) \begin{vmatrix} 1 & x_0 & x_0^2 \\ x_1 - x_3 & x_1(x_1 - x_3) & x_1^2(x_1 - x_3) \\ x_2 - x_3 & x_2(x_2 - x_3) & x_2^2(x_2 - x_3) \end{vmatrix}$$

$$= (-1)^3(x_0 - x_3)(x_1 - x_3) \begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ x_2 - x_3 & x_2(x_2 - x_3) & x_2^2(x_2 - x_3) \end{vmatrix}$$

$$= (-1)^3(x_0 - x_3)(x_1 - x_3)(x_2 - x_3) \begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix}$$

$$= (x_3 - x_0)(x_3 - x_1)(x_3 - x_2)d_2.$$

Per ricorrenza, anche d_2 ammette uno sviluppo analogo a d_3 e quindi otteniamo

$$\begin{aligned} d_3 &= (x_3 - x_0)(x_3 - x_1)(x_3 - x_2)d_2 = (x_3 - x_0)(x_3 - x_1)(x_3 - x_2) \cdot (x_2 - x_0)(x_2 - x_1)d_1 \\ &= (x_3 - x_0)(x_3 - x_1)(x_3 - x_2) \cdot (x_2 - x_0)(x_2 - x_1) \cdot (x_1 - x_0), \end{aligned}$$

dove l'ultimo passaggio vale perché $d_1 = \det[V(x_0, x_1)] = x_1 - x_0$.

2. Per ogni $j = 0, \ldots, n$ definiamo il polinomio

$$L_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} = \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)}.$$

Gli $n + 1$ polinomi $L_0(x), L_1(x), \ldots, L_n(x)$ hanno tutti grado n e quindi appartengono a $\mathbb{R}_n[x]$. Mostriamo ora che essi costituiscono una base di $\mathbb{R}_n[x]$ e per fare questo è sufficiente dimostrare che essi sono linearmente indipendenti, in quanto essi sono in numero di $n + 1 = \dim(\mathbb{R}_n[x])$. [2]

Per dimostrare che $L_0(x), L_1(x), \ldots, L_n(x)$ sono linearmente indipendenti, osserviamo che per ogni $i, j = 0, 1, \ldots, n$ si ha

$$L_j(x_i) = \begin{cases} 1 & \text{se } i = j, \\ 0 & \text{se } i \neq j. \end{cases}$$

Se $\alpha_0 L_0(x) + \alpha_1 L_1(x) + \ldots + \alpha_n L_n(x)$ è una combinazione lineare che coincide con il polinomio nullo, cioè tale che $\alpha_0 L_0(x) + \alpha_1 L_1(x) + \ldots + \alpha_n L_n(x) = 0$ per ogni $x \in \mathbb{R}$, allora in particolare deve essere $\alpha_0 L_0(x_i) + \alpha_1 L_1(x_i) + \ldots + \alpha_n L_n(x_i) = 0$ per ogni $i = 0, \ldots, n$, cioè $\alpha_i = 0$ per ogni $i = 0, \ldots, n$. Questo mostra che $L_0(x), L_1(x), \ldots, L_n(x)$ sono linearmente indipendenti e pertanto sono una base di $\mathbb{R}_n[x]$.

Definiamo il polinomio $p(x) = y_0 L_0(x) + y_1 L_1(x) + \ldots + y_n L_n(x)$. $p(x) \in \mathbb{R}_n[x]$ e per ogni $i = 0, \ldots, n$ si ha $p(x_i) = y_0 L_0(x_i) + y_1 L_1(x_i) + \ldots + y_n L_n(x_i) = y_i$. Quindi abbiamo provato l'esistenza di un polinomio in $\mathbb{R}_n[x]$ che nei nodi x_i assume i valori y_i.

[2] Ricordiamo che una base di $\mathbb{R}_n[x]$ è un insieme di elementi $v_1(x), \ldots, v_r(x) \in \mathbb{R}_n[x]$ con le seguenti proprietà:

- sono linearmente indipendenti, cioè l'unica combinazione lineare $\alpha_1 v_1(x) + \ldots + \alpha_r v_r(x)$ che coincide con il polinomio nullo è la combinazione lineare con $\alpha_1 = \ldots = \alpha_r = 0$;
- generano $\mathbb{R}_n[x]$, cioè ogni polinomio $q(x) \in \mathbb{R}_n[x]$ si può scrivere come combinazione lineare $q(x) = \beta_1 v_1(x) + \ldots + \beta_r v_r(x)$.

Ricordiamo inoltre che tutte le basi di $\mathbb{R}_n[x]$ hanno lo stesso numero di elementi, e questo numero comune a ogni base si chiama dimensione di $\mathbb{R}_n[x]$; nel caso di $\mathbb{R}_n[x]$ la dimensione è $r = n + 1$ perché una base di $\mathbb{R}_n[x]$ è quella canonica $1, x, x^2, \ldots, x^n$ che ha $n + 1$ elementi. Ricordiamo infine che se si hanno $n + 1$ elementi in uno spazio vettoriale di dimensione $n + 1$ come $\mathbb{R}_n[x]$, allora questi elementi sono una base di $\mathbb{R}_n[x]$ se e solo se sono linearmente indipendenti.

Supponiamo che $q(x)$ sia un altro polinomio in $\mathbb{R}_n[x]$ che nei nodi x_i assume i valori y_i. Poiché $L_0(x), L_1(x), \ldots, L_n(x)$ è una base di $\mathbb{R}_n[x]$, esistono $\beta_0, \beta_1, \ldots, \beta_n \in \mathbb{R}$ tali che $q(x) = \beta_0 L_0(x) + \beta_1 L_1(x) + \ldots + \beta_n L_n(x)$. Valutando $q(x)$ nei nodi x_i otteniamo che per ogni $i = 0, \ldots, n$ si ha

$$y_i = q(x_i) = \beta_0 L_0(x_i) + \beta_1 L_1(x_i) + \ldots + \beta_n L_n(x_i) = \beta_i,$$

da cui si ricava che $q(x) = y_0 L_0(x) + y_1 L_1(x) + \ldots + y_n L_n(x) = p(x)$. Questo prova che $p(x)$ è l'unico polinomio in $\mathbb{R}_n[x]$ che nei nodi x_i assume i valori y_i. □

Definizione 1.1. Siano $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ con $x_0, x_1, \ldots, x_n$ punti distinti. L'unico polinomio $p(x) \in \mathbb{R}_n[x]$ tale che $p(x_i) = y_i$ per ogni $i = 0, \ldots, n$ si chiama *polinomio d'interpolazione dei dati* $(x_0, y_0), \ldots, (x_n, y_n)$ o anche *polinomio d'interpolazione dei valori* $y_0, \ldots, y_n$ *sui nodi* $x_0, \ldots, x_n$.

La prima dimostrazione del Teorema 1.1 ci dice che $p(x)$ si scrive in *forma canonica* come

$$\boxed{p(x) = a_0 + a_1 x + \ldots + a_n x^n}$$

dove

$$\boxed{\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = [V(x_0, x_1, \ldots, x_n)]^{-1} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}} \tag{1.4}$$

e $V(x_0, x_1, \ldots, x_n)$ è la matrice di Vandermonde sui nodi $x_0, \ldots, x_n$.

La seconda dimostrazione del Teorema 1.1 ci dice che

$$\boxed{p(x) = y_0 L_0(x) + y_1 L_1(x) + \ldots + y_n L_n(x)} \tag{1.5}$$

dove per ogni $j = 0, 1, \ldots, n$,

$$\boxed{L_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} = \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)}}$$

è il j-esimo polinomio di Lagrange relativo ai nodi $x_0, \ldots, x_n$. La (1.5) si chiama *forma di Lagrange* del polinomio d'interpolazione.

Se gli y_i sono i valori nei punti x_i di una funzione $f : [a, b] \to \mathbb{R}$, cioè se risulta $y_i = f(x_i)$ per ogni $i = 0, \ldots, n$, allora il polinomio $p(x)$ si chiama anche *polinomio d'interpolazione della funzione* $f(x)$ *sui nodi* $x_0, \ldots, x_n$.

Esempio 1.1. Della funzione $\sin(x)$ sono noti i valori nei tre punti $x_0 = 0$, $x_1 = \pi/6$, $x_2 = \pi/4$ e questi valori sono rispettivamente $\sin(x_0) = 0$, $\sin(x_1) = 1/2$, $\sin(x_2) = 1/\sqrt{2}$. Scrivere in forma canonica e in forma di Lagrange il polinomio d'interpolazione di $\sin(x)$ sui nodi x_0, x_1, x_2.

Soluzione. Iniziamo dalla forma di Lagrange (1.5), dalla quale si ha immediatamente che il polinomio d'interpolazione di $\sin(x)$ sui nodi x_0, x_1, x_2 è

$$\begin{aligned} p(x) &= \sin(x_0) \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + \sin(x_1) \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + \sin(x_2) \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{1}{2} \frac{x(x - \pi/4)}{-\pi^2/72} + \frac{1}{\sqrt{2}} \frac{x(x - \pi/6)}{\pi^2/48}. \end{aligned}$$

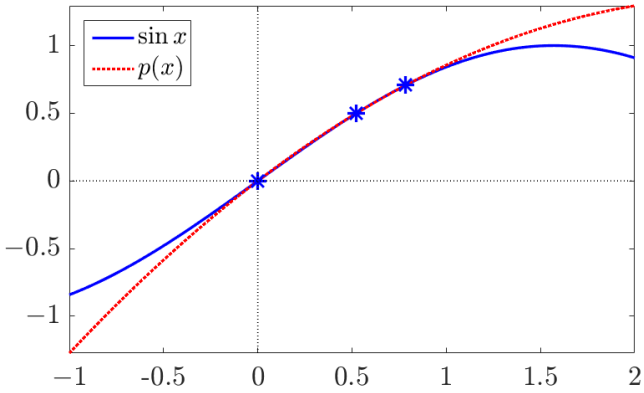

Figura 1.2: Grafici della funzione $\sin(x)$ e del suo polinomio d'interpolazione $p(x)$. I punti d'interpolazione $(x_0, \sin(x_0))$, $(x_1, \sin(x_1))$, $(x_2, \sin(x_2))$ sono segnati con gli asterischi.

Un controllo diretto permette di verificare che effettivamente $p(x_0) = 0$, $p(x_1) = 1/2$, $p(x_2) = 1/\sqrt{2}$. Sviluppando i calcoli possiamo riscrivere il polinomio in forma canonica

$$p(x) = \frac{24\sqrt{2} - 36}{\pi^2}\, x^2 + \frac{9 - 4\sqrt{2}}{\pi}\, x;$$

si veda la Figura 1.2.

Osservazione. Dalla (1.4) sappiamo che il vettore dei coefficienti di $p(x)$ è dato da

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \left[V\left(0, \frac{\pi}{6}, \frac{\pi}{4}\right)\right]^{-1} \begin{pmatrix} 0 \\ 1/2 \\ 1/\sqrt{2} \end{pmatrix}$$

e quindi, senza fare alcun conto, possiamo concludere immediatamente che

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \left[V\left(0, \frac{\pi}{6}, \frac{\pi}{4}\right)\right]^{-1} \begin{pmatrix} 0 \\ 1/2 \\ 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 \\ (9 - 4\sqrt{2})/\pi \\ (24\sqrt{2} - 36)/\pi^2 \end{pmatrix}.$$

L'ultima uguaglianza può essere verificata anche con un calcolo esplicito:

$$V\left(0, \frac{\pi}{6}, \frac{\pi}{4}\right) \begin{pmatrix} 0 \\ (9 - 4\sqrt{2})/\pi \\ (24\sqrt{2} - 36)/\pi^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1/2 \\ 1/\sqrt{2} \end{pmatrix}.$$

Esercizio 1.1. Scrivere in forma canonica e in forma di Lagrange il polinomio d'interpolazione della funzione $\sqrt{x}$ sui nodi $x_0 = 0$, $x_1 = 0.16$, $x_2 = 0.49$, $x_3 = 1$.

Esercizio 1.2. Scrivere in forma canonica e in forma di Lagrange il polinomio che interpola i valori $y_0 = 2$, $y_1 = 2.4$, $y_2 = 2.8$, $y_3 = 3.2$ sui nodi $x_0 = 1$, $x_1 = 1.2$, $x_2 = 1.4$, $x_3 = 1.6$.

Esercizio 1.3. Siano $L_0(x), \ldots, L_n(x)$ i polinomi di Lagrange relativi a $n+1$ nodi distinti $x_0, \ldots, x_n \in \mathbb{R}$. Dimostrare che la loro somma è identicamente uguale a 1:

$$\sum_{i=0}^{n} L_i(x) = 1 \quad \text{per ogni} \quad x \in \mathbb{R}.$$

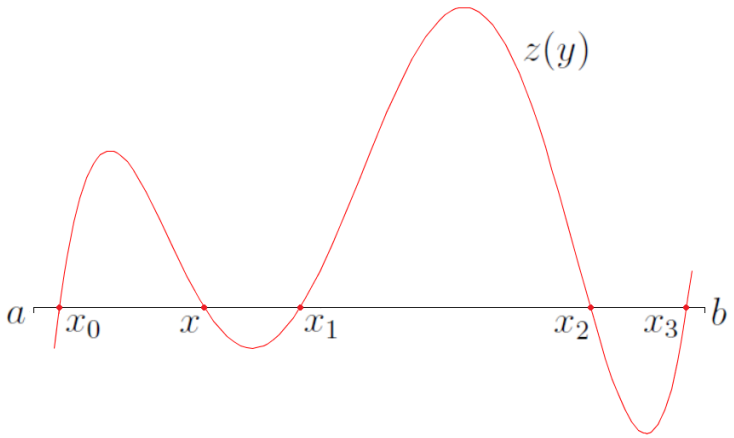

Figura 1.3: Illustrazione di $z(y)$ nel caso $n = 3$.

1.2 Errore o resto dell'interpolazione polinomiale

Teorema 1.2. *Sia $f : [a, b] \to \mathbb{R}$ una funzione di classe $C^{n+1}[a, b]$ e sia $p(x)$ il polinomio d'interpolazione di $f(x)$ sugli $n+1$ nodi distinti $x_0, x_1, \ldots, x_n \in [a, b]$. Allora, per ogni $x \in [a, b]$ esiste un punto $\xi = \xi(x) \in (a, b)$ tale che*

$$\boxed{f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n)} \tag{1.6}$$

Dimostrazione. Sia $x \in [a, b]$. Se x coincide con un nodo x_i allora la (1.6) è verificata con un qualsiasi $\xi \in (a, b)$ (entrambi i membri sono nulli). Supponiamo ora che x non coincida con uno dei nodi x_i. Definiamo $\pi(y) = (y - x_0)(y - x_1)\cdots(y - x_n)$ e $r(y) = f(y) - p(y)$, e consideriamo la funzione

$$z : [a, b] \to \mathbb{R}, \qquad z(y) = r(y) - \frac{r(x)}{\pi(x)}\pi(y);$$

si veda la Figura 1.3. Questa funzione è di classe $C^{n+1}[a, b]$ e si annulla in almeno $n + 2$ punti di $[a, b]$ dal momento che si annulla negli $n+1$ nodi $x_0, x_1, \ldots, x_n$ e anche nel punto x. Pertanto, per il teorema di Rolle, $z'(y)$ si annulla in almeno $n + 1$ punti di (a, b), $z''(y)$ si annulla in almeno n punti di (a, b), $z'''(y)$ si annulla in almeno $n - 1$ punti di (a, b), e così via fino ad avere che $z^{(n+1)}(y)$ si annulla in almeno un punto $\xi \in (a, b)$. Dunque, poiché $p^{(n+1)}(y)$ è identicamente nullo (perché $p(y)$ ha grado $\leq n$) e poiché $\pi^{(n+1)}(y) = (n + 1)!$ (perché $\pi(y)$ ha grado $n + 1$ ed è monico), si ha

$$\begin{aligned} 0 = z^{(n+1)}(\xi) &= r^{(n+1)}(\xi) - \frac{r(x)}{\pi(x)}\pi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - p^{(n+1)}(\xi) - \frac{f(x) - p(x)}{\pi(x)}(n + 1)! \\ &= f^{(n+1)}(\xi) - \frac{f(x) - p(x)}{(x - x_0)(x - x_1)\cdots(x - x_n)}(n + 1)! \end{aligned}$$

□

Esempio 1.2. Fissiamo un punto $t \in [0, 1]$. Stimare l'errore che si commette approssimando $\sin(t)$ con $p(t)$, dove $p(x)$ è il polinomio d'interpolazione della funzione $\sin(x)$ sui nodi $x_0 = 0$, $x_1 = \pi/6$, $x_2 = \pi/4$ che è stato determinato nell'Esempio 1.1.

Soluzione. Si tratta di applicare il Teorema 1.2 con $f(x) = \sin(x) \in C^\infty(\mathbb{R})$ e $n = 2$. Applicheremo il teorema sull'intervallo $[a, b] = [0, 1]$ = il più piccolo intervallo che contiene i nodi x_0, x_1, x_2 e il punto t.

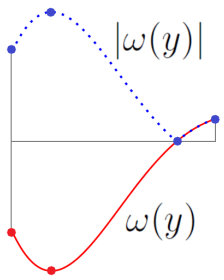

Figura 1.4: Grafici di una funzione $\omega(y)$ e del suo modulo $|\omega(y)|$.

Dalla (1.6) si ha

$$|\sin(t) - p(t)| = \left|\frac{f'''(\xi)}{3!}(t-x_0)(t-x_1)(t-x_2)\right| \qquad (\xi \text{ è un punto in } (0,1))$$
$$= \left|\frac{-\cos(\xi)}{6} t\Big(t-\frac{\pi}{6}\Big)\Big(t-\frac{\pi}{4}\Big)\right| = \frac{\cos(\xi)}{6}\,|t|\,\Big|t-\frac{\pi}{6}\Big|\,\Big|t-\frac{\pi}{4}\Big| \le \frac{1}{6}\cdot 1\cdot\frac{\pi}{6}\cdot\frac{\pi}{4} \approx 0.0685.$$

Volendo ottenere una stima più precisa, si può procedere nel modo seguente.

$$|\sin(t) - p(t)| = \left|\frac{f'''(\xi)}{3!}(t-x_0)(t-x_1)(t-x_2)\right| \qquad (\xi \text{ è un punto in } (0,1))$$
$$= \left|\frac{-\cos(\xi)}{6} t\Big(t-\frac{\pi}{6}\Big)\Big(t-\frac{\pi}{4}\Big)\right| = \frac{\cos(\xi)}{6}\,\Big|t\Big(t-\frac{\pi}{6}\Big)\Big(t-\frac{\pi}{4}\Big)\Big|$$
$$\le \frac{1}{6}\max_{y\in[0,1]}\Big|\underbrace{y\Big(y-\frac{\pi}{6}\Big)\Big(y-\frac{\pi}{4}\Big)}_{\omega(y)}\Big|. \tag{1.7}$$

Andiamo a calcolare il massimo di $|\omega(y)|$ su $[0,1]$. Per farlo, dobbiamo cercare tutti i massimi e i minimi relativi di $\omega(y)$ su $[0,1]$ e scegliere il più grande di essi in modulo; si veda anche la Figura 1.4. Per un teorema dell'analisi matematica, i massimi e i minimi relativi di $\omega(y)$ su $[0,1]$ si trovano o nei punti di bordo dell'intervallo $[0,1]$ oppure nei punti stazionari di $\omega(y)$ in $[0,1]$, cioè quei punti di $[0,1]$ in cui si annulla la derivata $\omega'(y)$. Si ha

$$\omega(y) = y\Big(y-\frac{\pi}{6}\Big)\Big(y-\frac{\pi}{4}\Big) = y^3 - \frac{5\pi}{12}\,y^2 + \frac{\pi^2}{24}\,y,$$
$$\omega'(y) = 3y^2 - \frac{5\pi}{6}\,y + \frac{\pi^2}{24},$$
$$\omega'(y) = 0 \quad\Longleftrightarrow\quad y = y_{1,2} = \frac{\frac{5\pi}{6} \pm \sqrt{\Big(\frac{5\pi}{6}\Big)^2 - \frac{\pi^2}{2}}}{6} = \frac{5\pi}{36} \pm \frac{\sqrt{7}\,\pi}{36}.$$

Siccome $y_{1,2}$ stanno in $[0,1]$, essi sono punti stazionari di $\omega(y)$ in $[0,1]$. Dunque, per quanto detto sopra,

$$\max_{y\in[0,1]} |\omega(y)| = \max\Big(\,|\omega(0)|,\,\Big|\omega\Big(\frac{5\pi}{36}+\frac{\sqrt{7}\,\pi}{36}\Big)\Big|,\,\Big|\omega\Big(\frac{5\pi}{36}-\frac{\sqrt{7}\,\pi}{36}\Big)\Big|,\,|\omega(1)|\,\Big)$$
$$= \max\Big(0, 0.01132..., 0.03790..., 0.10223...\Big) \le 0.103.$$

Dunque, tornando a (1.7), otteniamo

$$|f(x) - p(x)| \le \frac{1}{6}\,0.103 \approx 0.0172.$$

Osservazione 1. Per ottenere la stima dell'errore (sia quella meno precisa che quella più precisa) non è necessario conoscere il polinomio d'interpolazione $p(x)$.

Osservazione 2. Per ottenere la stima più precisa abbiamo determinato i massimi e i minimi di $\omega(y)$. Questo è stato possibile perché $\omega(y)$ è un polinomio di grado 3 e $\omega'(y)$ è un polinomio di grado 2, per cui abbiamo potuto risolvere l'equazione $\omega'(y) = 0$. Se fossimo in presenza di un polinomio di grado maggiore di 3 (come accade quando si hanno più di 3 nodi d'interpolazione), la determinazione dei massimi e dei minimi di $\omega(y)$ potrebbe essere molto complicata. In tal caso, conviene accontentarsi di una stima meno precisa come la prima che abbiamo visto, che è più semplice da ottenere in quanto non richiede di calcolare massimi e minimi.

Esempio 1.3. Si consideri la funzione $f(x) = \mathrm{e}^{x^2}$ e sia $p(x)$ il suo polinomio d'interpolazione sui nodi $x_0 = 0$, $x_1 = \frac{1}{2}$, $x_2 = 1$.

(a) Fornire una stima dell'errore d'interpolazione $|f(x) - p(x)|$ per ogni $x \in [0, 1]$, cioè determinare una costante C tale che $|f(x) - p(x)| \le C$ per ogni $x \in [0, 1]$.

(b) Stimare l'errore che si commette approssimando $\sqrt[9]{\mathrm{e}}$ con $p(\frac{1}{3})$ senza calcolare né $\sqrt[9]{\mathrm{e}}$ né $p(\frac{1}{3})$.

Soluzione.

(a) Applichiamo il Teorema 1.2 con $f(x) = \mathrm{e}^{x^2}$, $[a, b] = [0, 1]$ e $n = 2$. Per ogni $x \in [0, 1]$ si ha

$$|f(x) - p(x)| = \left| \frac{f'''(\xi)}{3!} x\Big(x - \frac{1}{2}\Big)(x - 1) \right| \qquad (\xi \text{ è un punto in } (0,1)) \tag{1.8}$$

Calcoliamo $f'''(x)$:

$$\begin{aligned} f'(x) &= 2x\mathrm{e}^{x^2}, \\ f''(x) &= 2\mathrm{e}^{x^2} + 2x \cdot 2x\mathrm{e}^{x^2} = (2 + 4x^2)\mathrm{e}^{x^2}, \\ f'''(x) &= 8x\mathrm{e}^{x^2} + (2 + 4x^2) \cdot 2x\mathrm{e}^{x^2} = (12x + 8x^3)\mathrm{e}^{x^2}. \end{aligned}$$

Per ogni $x \in [0, 1]$ si ha

$$|f'''(x)| = |(12x + 8x^3)\mathrm{e}^{x^2}| = (12x + 8x^3)\mathrm{e}^{x^2} \le (12 + 8)\mathrm{e} = 20\,\mathrm{e}.$$

Notiamo che questa stima non può essere migliorata perché $20\,\mathrm{e} = |f'''(1)|$. Tornando a (1.8), per ogni $x \in [0, 1]$ si ha

$$|f(x) - p(x)| = \frac{|f'''(\xi)|}{6} |x| \left|x - \frac{1}{2}\right| |x - 1| \le \frac{20\,\mathrm{e}}{6} \cdot 1 \cdot \frac{1}{2} \cdot 1 \approx 4.530.$$

Volendo ottenere una stima più precisa, si può procedere come segue. Per ogni $x \in [0, 1]$, si ha

$$\begin{aligned} |f(x) - p(x)| &= \left| \frac{f'''(\xi)}{3!} x\Big(x - \frac{1}{2}\Big)(x - 1) \right| \qquad (\xi \text{ è un punto in } (0,1)) \\ &= \frac{|f'''(\xi)|}{6} \left| x\Big(x - \frac{1}{2}\Big)(x - 1) \right| \le \frac{20\,\mathrm{e}}{6} \max_{y \in [0,1]} \Big| \underbrace{y\Big(y - \frac{1}{2}\Big)(y - 1)}_{\omega(y)} \Big|. \end{aligned} \tag{1.9}$$

Andiamo a calcolare il massimo di $|\omega(y)|$ su $[0, 1]$. Per farlo, dobbiamo cercare tutti i massimi e i minimi relativi di $\omega(y)$ su $[0, 1]$ e scegliere il più grande di essi in modulo. Per un teorema dell'analisi matematica, i massimi e i minimi relativi di $\omega(y)$ su $[0, 1]$ si trovano o nei punti di bordo dell'intervallo

$[0,1]$ oppure nei punti stazionari di $\omega(y)$ in $[0,1]$, cioè quei punti di $[0,1]$ in cui si annulla la derivata $\omega'(y)$. Si ha

$$\begin{aligned}
&\omega(y) = y\Big(y-\frac{1}{2}\Big)(y-1) = y^3 - \frac{3}{2}y^2 + \frac{1}{2}y,\\
&\omega'(y) = 3y^2 - 3y + \frac{1}{2},\\
&\omega'(y) = 0 \quad \Longleftrightarrow \quad y = y_{1,2} = \frac{3 \pm \sqrt{9-6}}{6} = \frac{1}{2} \pm \frac{\sqrt{3}}{6}.
\end{aligned}$$

Siccome $y_{1,2}$ stanno in $[0,1]$, essi sono punti stazionari di $\omega(y)$ in $[0,1]$. Dunque, per quanto detto sopra,

$$\begin{aligned}
\max_{y\in[0,1]} |\omega(y)| &= \max\Big(|\omega(0)|, \Big|\omega\Big(\frac{1}{2}+\frac{\sqrt{3}}{6}\Big)\Big|, \Big|\omega\Big(\frac{1}{2}-\frac{\sqrt{3}}{6}\Big)\Big|, |\omega(1)| \Big)\\
&= \max\Big(0, \frac{\sqrt{3}}{36}, \frac{\sqrt{3}}{36}, 0\Big) = \frac{\sqrt{3}}{36}.
\end{aligned}$$

Dunque, tornando a (1.9), otteniamo

$$|f(x) - p(x)| \le \frac{20\,\mathrm{e}}{6}\,\frac{\sqrt{3}}{36} \approx 0.436.$$

(b) Siccome $x = \frac{1}{3}$ si trova in $[0,1]$, possiamo usare la stima del punto (a) e concludere subito che

$$\Big|\sqrt[9]{\mathrm{e}} - p\Big(\frac{1}{3}\Big)\Big| = \Big|\mathrm{e}^{(\frac{1}{3})^2} - p\Big(\frac{1}{3}\Big)\Big| \le \frac{20\,\mathrm{e}}{6}\,\frac{\sqrt{3}}{36} \approx 0.436.$$

Volendo ottenere una stima più precisa, si può applicare il Teorema 1.2 direttamente con $x = \frac{1}{3}$. In tal modo, otteniamo

$$\begin{aligned}
\Big|\sqrt[9]{\mathrm{e}} - p\Big(\frac{1}{3}\Big)\Big| = \Big|\mathrm{e}^{(\frac{1}{3})^2} - p\Big(\frac{1}{3}\Big)\Big| &= \Big|\frac{f'''(\xi)}{3!}\,\frac{1}{3}\Big(\frac{1}{3}-\frac{1}{2}\Big)\Big(\frac{1}{3}-1\Big)\Big| \qquad (\xi \text{ è un punto in } (0,1))\\
&= \frac{|f'''(\xi)|}{6}\cdot\frac{1}{3}\cdot\frac{1}{6}\cdot\frac{2}{3} \le \frac{20\,\mathrm{e}}{6}\cdot\frac{1}{3}\cdot\frac{1}{6}\cdot\frac{2}{3} \approx 0.336.
\end{aligned}$$

Esercizio 1.4. Si consideri la funzione $f(x) = \arctan(x)$.

(a) Scrivere in forma di Lagrange e in forma canonica il polinomio d'interpolazione $p(x)$ della funzione $f(x)$ sui nodi $x_0 = 0$, $x_1 = 1$, $x_2 = \sqrt{3}$.

(b) Fissato $t \in [0,2]$, stimare l'errore $|f(t) - p(t)|$ che si commette approssimando $f(t)$ con $p(t)$.

Esercizio 1.5. Per ogni fissato $t \in [0,1]$, stimare l'errore che si commette approssimando $\sin(t)$ con $p(t)$, dove $p(x)$ è il polinomio d'interpolazione della funzione $\sin(x)$ sui nodi $x_0 = 0$, $x_1 = \pi/10$, $x_2 = \pi/6$, $x_3 = \pi/4$, $x_4 = 3\pi/10$. Confrontare la stima ottenuta con quelle dell'Esempio 1.2. Osservare che per risolvere l'esercizio non è necessario conoscere $p(x)$.

Esercizio 1.6. Stimare l'errore che si commette approssimando $\sqrt{2}$ con $p(2)$, dove $p(x)$ è il polinomio d'interpolazione di $\sqrt{x}$ sui nodi 1.69, 1.7689, 1.8769, 1.96, 2.0449, 2.1609, 2.25. Si faccia la stima senza calcolare né $\sqrt{2}$ né $p(2)$.

Esercizio 1.7. Stimare l'errore che si commette approssimando $\cos(1)$ con $p(1)$, dove $p(x)$ è il polinomio d'interpolazione di $\cos(x)$ sui nodi 0, $\frac{\pi}{6}$, $\frac{\pi}{4}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$. Si faccia la stima senza calcolare né $\cos(1)$ né $p(1)$.

1.3 Forma di Newton del polinomio d'interpolazione

Abbiamo visto la forma canonica e la forma di Lagrange per rappresentare il polinomio d'interpolazione. Vediamo ora un'altra forma di rappresentazione, la forma di Newton.

Definizione 1.2. Sia $f : [a, b] \to \mathbb{R}$ una funzione.
- Se $y \in [a, b]$, si definisce differenza divisa di $f(x)$ relativa a y il numero $f[y] = f(y)$.
- Se $y_1, \dots, y_k \in [a, b]$ sono $k \geq 2$ punti distinti, si definisce differenza divisa di $f(x)$ relativa a $y_1, \dots, y_k$ il numero
$$f[y_1, \dots, y_k] = \frac{f[y_1, \dots, y_{k-2}, y_k] - f[y_1, \dots, y_{k-1}]}{y_k - y_{k-1}}.$$
Osserviamo che nel caso $k = 2$ la formula precedente fornisce
$$f[y_1, y_2] = \frac{f[y_2] - f[y_1]}{y_2 - y_1} = \frac{f(y_2) - f(y_1)}{y_2 - y_1},$$
per cui $f[y_1, y_2]$ è il rapporto incrementale di $f(x)$ relativo ai punti y_1, y_2.

Teorema 1.3. *Sia $f : [a, b] \to \mathbb{R}$ una funzione e siano $x_0, x_1, \dots, x_n \in [a, b]$ nodi distinti. Allora il polinomio d'interpolazione di $f(x)$ sui nodi $x_0, x_1, \dots, x_n$ è*

$$\boxed{p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \dots + f[x_0, \dots, x_n](x - x_0) \cdots (x - x_{n-1})} \tag{1.10}$$

La (1.10) si chiama forma di Newton del polinomio d'interpolazione.

Corollario 1.1. *Sia $f : [a, b] \to \mathbb{R}$ una funzione e siano $x_0, x_1, \dots, x_n \in [a, b]$ nodi distinti. Allora $f[x_0, \dots, x_n]$ non cambia se vengono permutati i suoi $n + 1$ argomenti, cioè*

$$f[x_0, \dots, x_n] = f[x_{\sigma(0)}, \dots, x_{\sigma(n)}]$$

per ogni permutazione σ dell'insieme $\{0, \dots, n\}$.

Dimostrazione. Sia σ una permutazione dell'insieme $\{0, \dots, n\}$. Dalla (1.10) applicata prima con i nodi $x_0, \dots, x_n$ e poi con i nodi permutati $x_{\sigma(0)}, \dots, x_{\sigma(n)}$, vediamo che $f[x_0, \dots, x_n]$ e $f[x_{\sigma(0)}, \dots, x_{\sigma(n)}]$ sono entrambi il coefficiente direttore di

$$\begin{aligned} p(x) &= \text{il polinomio d'interpolazione di } f(x) \text{ sui nodi } x_0, \dots, x_n \\ &= \text{il polinomio d'interpolazione di } f(x) \text{ sui nodi } x_{\sigma(0)}, \dots, x_{\sigma(n)}. \end{aligned}$$

Dunque $f[x_0, \dots, x_n] = f[x_{\sigma(0)}, \dots, x_{\sigma(n)}]$. □

Esempio 1.4. Scrivere in forma canonica e in forma di Newton il polinomio d'interpolazione $p(x)$ della funzione $f(x) = \sqrt{x}$ sui nodi $x_0 = 0$, $x_1 = 0.16$, $x_2 = 0.64$, $x_3 = 1$.

Soluzione. Iniziamo dalla forma di Newton, da cui poi otterremo la forma canonica semplicemente sviluppando i calcoli. In base al Teorema 1.3, il polinomio $p(x)$ è dato in forma di Newton dalla formula seguente:

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2). \tag{1.11}$$

L'unica cosa da fare è calcolare le differenze divise che compaiono nella (1.11). A tal fine, si usa la *tabella delle differenze divise* (Tabella 1.1). Si procede calcolando gli elementi della tabella colonna per colonna. Si ha

$$
\begin{aligned}
f[x_0] &= f(x_0) = \boxed{0} \\
f[x_1] &= f(x_1) = 0.4 \\
f[x_2] &= f(x_2) = 0.8 \\
f[x_3] &= f(x_3) = 1 \\
f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0} = 0.4/0.16 = \boxed{5/2} \\
f[x_0, x_2] &= \frac{f[x_2] - f[x_0]}{x_2 - x_0} = 0.8/0.64 = 5/4 \\
f[x_0, x_3] &= \frac{f[x_3] - f[x_0]}{x_3 - x_0} = 1 \\
f[x_0, x_1, x_2] &= \frac{f[x_0, x_2] - f[x_0, x_1]}{x_2 - x_1} = -1.25/0.48 = \boxed{-125/48} \\
f[x_0, x_1, x_3] &= \frac{f[x_0, x_3] - f[x_0, x_1]}{x_3 - x_1} = -1.5/0.84 = -25/14 \\
f[x_0, x_1, x_2, x_3] &= \frac{f[x_0, x_1, x_3] - f[x_0, x_1, x_2]}{x_3 - x_2} = (-25/14 + 125/48)/0.36 = \boxed{6875/3024}
\end{aligned}
$$

Sostituendo i valori trovati nella (1.11) si ottiene la forma di Newton di $p(x)$:

$$p(x) = \frac{5}{2}\,x - \frac{125}{48}\,x(x - 0.16) + \frac{6875}{3024}\,x(x - 0.16)(x - 0.64).$$

Sviluppando i calcoli possiamo riscrivere il polinomio in forma canonica:

$$p(x) = \frac{6875}{3024}\,x^3 - \frac{13375}{3024}\,x^2 + \frac{2381}{756}\,x.$$

Osservazione 1.1. Supponiamo che siano dati i punti $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ con $x_0, x_1, \ldots, x_n$ distinti. I numeri $y_0, y_1, \ldots, y_n$ possono *sempre* essere interpretati come i valori in $x_0, x_1, \ldots, x_n$ di una qualche funzione $f : [a, b] \to \mathbb{R}$ definita su un qualche intervallo $[a, b]$ che contiene i punti $x_0, \ldots, x_n$. Pertanto, ha perfettamente senso parlare di forma di Newton del polinomio d'interpolazione dei dati $(x_0, y_0), \ldots, (x_n, y_n)$, anche qualora non venga specificata alcuna funzione $f(x)$ tale che $f(x_i) = y_i$ per ogni $i = 0, \ldots, n$; in tal caso infatti, basta immaginarsi una qualche funzione $f(x)$ che assume i valori $y_0, \ldots, y_n$ nei nodi $x_0, \ldots, x_n$ e il gioco è fatto.

Esercizio 1.8. Scrivere in forma canonica e in forma di Newton il polinomio d'interpolazione di $f(x) = \cos(\frac{\pi}{2}x)\log_2(x)$ sui nodi $x_0 = 1$, $x_1 = 2$, $x_2 = 4$, $x_3 = 8$.

Esercizio 1.9. Scrivere in forma canonica, in forma di Lagrange e in forma di Newton il polinomio d'interpolazione dei valori $y_0 = 0$, $y_1 = 3$, $y_2 = -3$ sui nodi $x_0 = 0$, $x_1 = 1$, $x_2 = 2$.

1.4 Algoritmo di valutazione del polinomio d'interpolazione in un punto

Grazie alla forma di Newton, è possibile costruire un algoritmo per valutare in un punto il polinomio d'interpolazione che risulta essere molto buono dal punto di vista del costo computazionale. In questa sezione lo descriviamo e ne calcoliamo il costo computazionale.

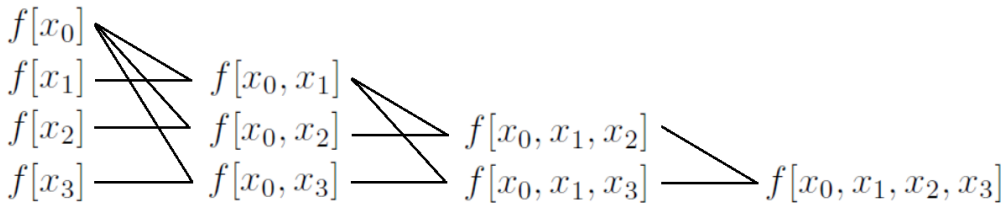

Tabella 1.1: Tabella delle differenze divise.

Sia $f : [a, b] \to \mathbb{R}$, siano $x_0, x_1, \ldots, x_n \in [a, b]$ punti distinti e sia $t \in \mathbb{R}$. Si vuole costruire un algoritmo per calcolare $p(t)$, dove $p(x)$ è il polinomio d'interpolazione di $f(x)$ sui nodi $x_0, x_1, \ldots, x_n$. Per maggiore chiarezza illustriamo l'algoritmo supponendo che sia $n = 3$, cosicché, per la (1.10), è

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2). \quad (1.12)$$

La prima parte dell'algoritmo è *indipendente dal punto* t in cui $p(x)$ deve essere valutato e consiste nel calcolare le differenze divise che compaiono nella (1.12). Questo calcolo viene fatto usando la tabella delle differenze divise (Tabella 1.1), esattamente come abbiamo visto nell'Esempio 1.4.

Una volta calcolate le differenze divise, per calcolare $p(t)$ si usa un algoritmo noto come metodo di Ruffini-Horner: ricordando la (1.12), si scrive $p(t)$ nella forma

$$p(t) = f[x_0] + (t - x_0)\Big(f[x_0, x_1] + (t - x_1)\big(f[x_0, x_1, x_2] + (t - x_2)f[x_0, x_1, x_2, x_3]\big)\Big)$$

e si pone

$$\begin{aligned} h_3 &= f[x_0, x_1, x_2, x_3], \\ h_2 &= f[x_0, x_1, x_2] + (t - x_2)h_3, \\ h_1 &= f[x_0, x_1] + (t - x_1)h_2, \\ h_0 &= f[x_0] + (t - x_0)h_1. \end{aligned}$$

h_0 è proprio $p(t)$.

Valutiamo il costo computazionale dell'algoritmo. Per calcolare le differenze divise della Tabella 1.1 si devono calcolare $6 = \frac{n(n+1)}{2}$ elementi della tabella (tutti meno quelli della prima colonna che sono noti). Il numero di elementi da calcolare coincide con il numero di elementi della parte triangolare inferiore (inclusa la diagonale) di una matrice $n \times n$, ossia $\frac{n^2-n}{2} + n = \frac{n(n+1)}{2}$ ovvero $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$. Per il calcolo di ciascun elemento sono richieste 2 sottrazioni e 1 divisione. Quindi il costo computazionale del calcolo delle differenze divise è il seguente:

- $2 \cdot 6 = n(n + 1)$ sottrazioni,
- $1 \cdot 6 = \frac{n(n+1)}{2}$ divisioni.

Una volta calcolate le differenze divise, restano da calcolare h_2, h_1, h_0 ($h_3 = h_n$ non va calcolato perché è una differenza divisa che è già stata calcolata). Per il calcolo di ciascun h_i sono richieste 1 sottrazione, 1 moltiplicazione e 1 addizione. Quindi il costo computazionale del calcolo di h_2, h_1, h_0 (cioè di $h_{n-1}, \ldots, h_0$) è il seguente:

- $1 \cdot 3 = n$ sottrazioni,
- $1 \cdot 3 = n$ moltiplicazioni,
- $1 \cdot 3 = n$ addizioni.

Notiamo che una sottrazione costa come un'addizione per il computer: la sottrazione $\xi - \eta$ coincide con l'addizione $\xi + (-\eta)$ e per il computer mettere un segno meno davanti a η non costa niente (non deve fare nessuna operazione). Considereremo pertanto le sottrazioni come se fossero delle addizioni. Dunque, indicando con A, M e D rispettivamente le addizioni, le moltiplicazioni e le divisioni,

- il costo del calcolo delle differenze divise è $n(n+1)A + \frac{n(n+1)}{2}D$,
- il costo del calcolo di $h_{n-1}, \dots, h_0$ è $2nA + nM$,
- il costo complessivo dell'algoritmo è

$$c(n) = (n^2 + 3n)A + nM + \left(\frac{n^2}{2} + \frac{n}{2}\right)D \approx n^2 A + \frac{n^2}{2}D,$$

per un totale di $\frac{3}{2}n^2 + \frac{9}{2}n \approx \frac{3}{2}n^2$ operazioni.

Osservazione 1.2. Siccome la prima parte dell'algoritmo consiste nel calcolare le differenze divise ed è indipendente dal punto t in cui $p(x)$ deve essere valutato, per valutare $p(x)$ in m punti $t_1, \dots, t_m \in \mathbb{R}$ si procede nel modo seguente.

- Si calcolano prima le differenze divise come sopra: il costo computazionale è ancora $n(n+1)A + \frac{n(n+1)}{2}D$.
- Si calcolano $p(t_1), \dots, p(t_m)$ allo stesso modo in cui sopra abbiamo calcolato $p(t)$: il costo computazionale è $m(2nA + nM)$.

Quindi il costo complessivo della valutazione di $p(x)$ in m punti è

$$c_m(n) = \left(n^2 + 2mn + n\right)A + mnM + \left(\frac{n^2}{2} + \frac{n}{2}\right)D \approx (n^2 + 2mn)A + mnM + \frac{n^2}{2}D$$

e in totale vengono eseguite $\frac{3}{2}n^2 + 3mn + \frac{3}{2}n \approx \frac{3}{2}n^2 + 3mn$ operazioni.

Esempio 1.5. Consideriamo i dati

$$\begin{aligned}(x_0, y_0) &= (0, 0),\\ (x_1, y_1) &= (1, 3),\\ (x_2, y_2) &= (2, 1),\\ (x_3, y_3) &= (3, 1).\end{aligned}$$

Calcolare mediante l'algoritmo descritto il valore nel punto $t = 2.3$ del polinomio $p(x)$ che interpola i valori y_0, y_1, y_2, y_3 sui nodi x_0, x_1, x_2, x_3.

Soluzione. Sia $f : [0, 3] \to \mathbb{R}$ una qualsiasi funzione tale che $f(x_0) = y_0$, $f(x_1) = y_1$, $f(x_2) = y_2$, $f(x_3) = y_3$. Calcoliamo le differenze divise usando la tabella delle differenze divise (Tabella 1.1). Si ha

$$\begin{aligned}f[x_0] &= f(x_0) = \boxed{0}\\ f[x_1] &= f(x_1) = 3\\ f[x_2] &= f(x_2) = 1\\ f[x_3] &= f(x_3) = 1\\ f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \boxed{3}\\ f[x_0, x_2] &= \frac{f[x_2] - f[x_0]}{x_2 - x_0} = 1/2\\ f[x_0, x_3] &= \frac{f[x_3] - f[x_0]}{x_3 - x_0} = 1/3\\ f[x_0, x_1, x_2] &= \frac{f[x_0, x_2] - f[x_0, x_1]}{x_2 - x_1} = \boxed{-5/2}\\ f[x_0, x_1, x_3] &= \frac{f[x_0, x_3] - f[x_0, x_1]}{x_3 - x_1} = -4/3\end{aligned}$$

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_0, x_1, x_3] - f[x_0, x_1, x_2]}{x_3 - x_2} = \boxed{7/6}$$

Poniamo ora

$$h_3 = f[x_0, x_1, x_2, x_3] = 7/6$$

e calcoliamo

$$\begin{aligned} h_2 &= f[x_0, x_1, x_2] + (t - x_2)h_3 = -43/20, \\ h_1 &= f[x_0, x_1] + (t - x_1)h_2 = 41/200, \\ h_0 &= f[x_0] + (t - x_0)h_1 = 943/2000. \end{aligned}$$

In conclusione, $p(t) = h_0 = 943/2000 = 0.4715$.

Esercizio 1.10. Tramite l'algoritmo descritto, valutare in $t = \frac{1}{2}$ e $t = 4$ il polinomio d'interpolazione della funzione $f(x) = x\sqrt{x}$ sui nodi $x_0 = 0$, $x_1 = 1$, $x_2 = 4$, $x_3 = 9$.

Esercizio 1.11. Scrivere un programma MATLAB che implementa l'algoritmo descritto. Il programma deve:

- prendere in input tre vettori (a componenti reali) $[x_0, x_1, \ldots, x_n]$, $[y_0, y_1, \ldots, y_n]$, $[t_1, \ldots, t_m]$, con $x_0, \ldots, x_n$ tutti distinti;
- restituire in output il vettore $[p(t_1), \ldots, p(t_m)]$ che contiene le valutazioni nei punti $t_1, \ldots, t_m$ del polinomio $p(x)$ interpolante i dati $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$.

Verificare sperimentalmente la correttezza del programma, utilizzandolo in particolare per riottenere il risultato dell'Esempio 1.5 e per risolvere nuovamente l'Esercizio 1.10.

1.5 Aggiunta di un nodo d'interpolazione

La forma di Newton è particolarmente conveniente quando ai dati d'interpolazione $(x_0, y_0), \ldots, (x_n, y_n)$ ne viene aggiunto un altro (x_{n+1}, y_{n+1}) con $x_{n+1} \neq x_0, \ldots, x_n$. Infatti, detta $f(x)$ una qualsiasi funzione tale che $f(x_i) = y_i$ per ogni $i = 0, \ldots, n+1$, il polinomio d'interpolazione dei dati $(x_0, y_0), \ldots, (x_n, y_n)$ si scrive in forma di Newton nel modo seguente:

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1});$$

e il polinomio d'interpolazione dei dati $(x_0, y_0), \ldots, (x_{n+1}, y_{n+1})$ si scrive in forma di Newton nel modo seguente:

$$q(x) = p(x) + f[x_0, \ldots, x_{n+1}](x - x_0) \cdots (x - x_n).$$

Possiamo quindi trarre le seguenti conclusioni. Indichiamo come al solito con A, M e D le addizioni, moltiplicazioni e divisioni.

- Avendo a disposizione $p(x)$ in forma di Newton, sono noti i coefficienti $f[x_0]$, $f[x_0, x_1]$, $\ldots$, $f[x_0, \ldots, x_n]$ e basta quindi calcolare $f[x_0, \ldots, x_{n+1}]$ per ottenere la forma di Newton del nuovo polinomio d'interpolazione $q(x)$. Uno sguardo alla Tabella 1.1 (pensando $n = 2$) mostra che per calcolare $f[x_0, \ldots, x_{n+1}]$ bisogna calcolare solo l'ultima riga della tabella, il che richiede $2(n+1)A + (n+1)D$.
- Avendo a disposizione $p(x)$ in forma di Newton e il suo valore $p(t)$, per calcolare $q(t)$ occorre calcolare $f[x_0, \ldots, x_{n+1}](t - x_0) \cdots (t - x_n)$ e sommarlo a $p(t)$, il che richiede $2(n+1)A + (n+1)D$ per calcolare $f[x_0, \ldots, x_{n+1}]$, più $(n+1)A + (n+1)M$ per calcolare $f[x_0, \ldots, x_{n+1}](t - x_0) \cdots (t - x_n)$, più $1A$ per sommare il risultato a $p(t)$. In totale $(3n+4)A + (n+1)M + (n+1)D$.

Esempio 1.6. Si consideri la funzione $f(x) = \cos(\pi x) + x^2$.

(a) Scrivere il polinomio d'interpolazione $p(x)$ della funzione $f(x)$ sui nodi $x_0 = -1$, $x_1 = 0$, $x_2 = 2$ e calcolarne il valore in $t = \frac{1}{2}$.

(b) Scrivere il polinomio d'interpolazione $q(x)$ della funzione $f(x)$ sui nodi precedenti ai quali si aggiunge un nuovo nodo $x_3 = 1$ e calcolarne il valore in $t = \frac{1}{2}$.

Soluzione. (a) Conviene scrivere il polinomio $p(x)$ in forma di Newton in vista sia della valutazione in $t = \frac{1}{2}$ mediante l'algoritmo conveniente descritto in Sezione 1.4 sia dell'aggiunta di un nodo di cui si parla nel punto (b). Si ha

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1). \tag{1.13}$$

Calcolando le differenze divise seguendo il solito schema (Tabella 1.1 privata dell'ultima riga), si ottiene

$$\begin{aligned}
f[x_0] &= f(x_0) = \boxed{0} \\
f[x_1] &= f(x_1) = 1 \\
f[x_2] &= f(x_2) = 5 \\
f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \boxed{1} \\
f[x_0, x_2] &= \frac{f[x_2] - f[x_0]}{x_2 - x_0} = 5/3 \\
f[x_0, x_1, x_2] &= \frac{f[x_0, x_2] - f[x_0, x_1]}{x_2 - x_1} = \boxed{1/3}
\end{aligned}$$

da cui, sostituendo in (1.13), si ottiene la forma di Newton di $p(x)$:

$$p(x) = (x + 1) + \frac{1}{3}(x + 1)x.$$

Calcoliamo $p(t)$ con $t = \frac{1}{2}$ usando l'algoritmo descritto in Sezione 1.4:

$$\begin{aligned}
h_2 &= f[x_0, x_1, x_2] = 1/3, \\
h_1 &= f[x_0, x_1] + (t - x_1)h_2 = 1 + (1/2) \cdot (1/3) = 7/6, \\
h_0 &= f[x_0] + (t - x_0)h_1 = 0 + (1/2 + 1) \cdot (7/6) = 7/4,
\end{aligned}$$

da cui $p(t) = h_0 = 7/4$.

(b) Scriviamo $q(x)$ in forma di Newton sfruttando il calcolo di $p(x)$ già fatto al punto (a). Si ha

$$q(x) = p(x) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2). \tag{1.14}$$

Per calcolare $f[x_0, x_1, x_2, x_3]$ dobbiamo prima determinare $f[x_3]$, $f[x_0, x_3]$, $f[x_0, x_1, x_3]$, come si desume dal solito schema (Tabella 1.1). Tenendo conto che $f[x_0]$, $f[x_0, x_1]$, $f[x_0, x_1, x_2]$ sono già noti (sono i coefficienti di $p(x)$ nella forma di Newton (1.13)), risulta

$$\begin{aligned}
f[x_3] &= f(x_3) = 0 \\
f[x_0, x_3] &= \frac{f[x_3] - f[x_0]}{x_3 - x_0} = 0 \\
f[x_0, x_1, x_3] &= \frac{f[x_0, x_3] - f[x_0, x_1]}{x_3 - x_1} = -1 \\
f[x_0, x_1, x_2, x_3] &= \frac{f[x_0, x_1, x_3] - f[x_0, x_1, x_2]}{x_3 - x_2} = \boxed{4/3}
\end{aligned}$$

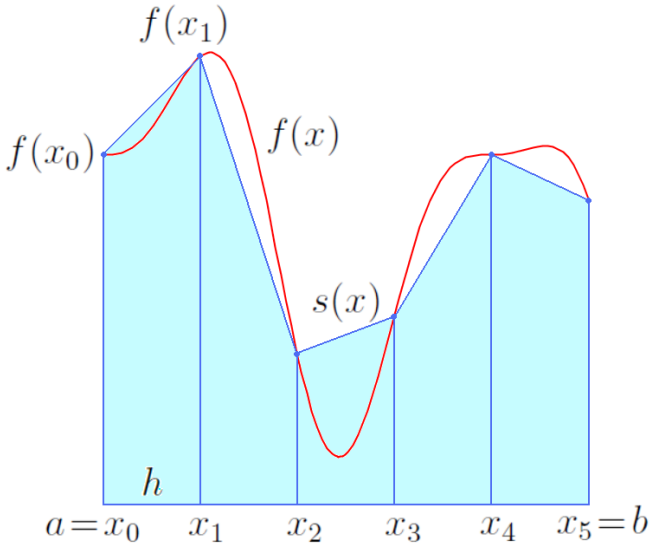

Figura 2.1: Illustrazione della formula dei trapezi per $n = 5$.

da cui, sostituendo in (1.14), si ottiene la forma di Newton di $q(x)$:

$$q(x) = p(x) + \frac{4}{3}(x+1)x(x-2).$$

Poiché $p(t) = 7/4$ è già stato calcolato, otteniamo subito

$$q(t) = p(t) + \frac{4}{3}(t+1)t(t-2) = \frac{7}{4} + \frac{4}{3}\cdot\frac{3}{2}\cdot\frac{1}{2}\Big(-\frac{3}{2}\Big) = \frac{1}{4}.$$

Osservazione. Poiché per $t = \frac{1}{2}$ si ha $f(t) = 1/4$, il polinomio $q(t)$ è un'approssimazione migliore (esatta in questo caso) di $f(t)$ rispetto a $p(t) = 7/4$. L'aggiunta di un nodo ha quindi migliorato l'approssimazione di $f(t)$, come era intuitivo.

Esercizio 1.12. Si consideri la funzione $f(x) = \sin(\pi x)$.

(a) Determinare il polinomio d'interpolazione $p(x)$ della funzione $f(x)$ sui nodi $x_0 = 0$, $x_1 = \frac{1}{2}$, $x_2 = 1$ e calcolare $p(t)$ per $t = \frac{1}{4}$ e $t = \frac{1}{3}$.

(b) Determinare il polinomio d'interpolazione $q(x)$ della funzione $f(x)$ sui nodi precedenti ai quali si aggiunge un nuovo nodo $x_3 = \frac{3}{4}$ e calcolare $q(t)$ per $t = \frac{1}{4}$ e $t = \frac{1}{3}$.

2 Integrazione numerica

2.1 Formula dei trapezi

È data una funzione (integrabile) $f : [a, b] \to \mathbb{R}$ e si vuole calcolare un'approssimazione di $\int_a^b f(x)\mathrm{d}x$. A tal fine si suddivide l'intervallo $[a, b]$ in $n \geq 1$ sottointervalli tutti della stessa ampiezza $h = \frac{b-a}{n}$ e si pone $x_j = a + jh$ per ogni $j = 0, 1, \ldots, n$. Il valore che si prende come approssimazione di $\int_a^b f(x)\mathrm{d}x$ è $\int_a^b s(x)\mathrm{d}x$, dove

$$s : [a, b] \to \mathbb{R}, \quad \begin{cases} s(x) = f(x_j) + \dfrac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j}(x - x_j), \\ \text{per } x \in [x_j, x_{j+1}], \quad j = 0, \ldots, n-1, \end{cases}$$

è la funzione lineare a tratti mostrata in Figura 2.1. Quindi il valore che si prende come approssimazione di $\int_a^b f(x)\mathrm{d}x$ è

$$
\begin{aligned}
I_n &= \int_a^b s(x)\mathrm{d}x = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} s(x)\mathrm{d}x = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} \left[f(x_j) + \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j}(x - x_j) \right] \mathrm{d}x \\
&= \sum_{j=0}^{n-1} \left[f(x_j)(x - x_j) + \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} \frac{(x - x_j)^2}{2} \right]_{x_j}^{x_{j+1}} \\
&= \sum_{j=0}^{n-1} \left[f(x_j)(x_{j+1} - x_j) + \frac{f(x_{j+1}) - f(x_j)}{2}(x_{j+1} - x_j) \right] \\
&= \sum_{j=0}^{n-1} \frac{f(x_j) + f(x_{j+1})}{2} h = \frac{h}{2} \sum_{j=0}^{n-1} \Big[f(x_j) + f(x_{j+1}) \Big] \\
&= \frac{h}{2} \Big[f(x_0) + f(x_1) + f(x_1) + f(x_2) + f(x_2) + f(x_3) + \ldots + f(x_{n-1}) + f(x_n) \Big] \\
&= h \left[\frac{f(a) + f(b)}{2} + \sum_{j=1}^{n-1} f(x_j) \right]
\end{aligned}
$$

cioè

$$
\boxed{I_n = h \left[\frac{f(a) + f(b)}{2} + \sum_{j=1}^{n-1} f(x_j) \right]} \tag{2.1}
$$

La (2.1) prende il nome di *formula dei trapezi di ordine n*. L'ampiezza $h = \frac{b-a}{n}$ si chiama anche *passo (di discretizzazione) della formula I_n*.

2.2 Errore o resto della formula dei trapezi

Vogliamo capire qual è l'errore che si commette approssimando $\int_a^b f(x)\mathrm{d}x$ con I_n. Per far questo utilizzeremo il seguente lemma.

Lemma 2.1. *Siano $\omega, \alpha, \beta : [a,b] \to \mathbb{R}$ funzioni tali che*
- *$\omega(x)$ è continua e ≥ 0 su $[a,b]$,*
- *$\alpha(x)$ e $\beta(x)\omega(x)$ sono continue su $[a,b]$,*
- *$m \leq \beta(x) \leq M$ per ogni $x \in [a,b]$, dove m e M sono rispettivamente il minimo e il massimo di $\alpha(x)$ su $[a,b]$.*

Allora esiste un punto $\eta \in [a,b]$ tale che

$$
\int_a^b \beta(x)\omega(x)\mathrm{d}x = \alpha(\eta) \int_a^b \omega(x)\mathrm{d}x.
$$

Questo lemma è una generalizzazione del teorema della media integrale. Per $\omega(x) = 1$ e $\beta(x) = \alpha(x)$ si ottiene infatti il teorema della media integrale.

Dimostrazione. Poiché $\omega(x) \geq 0$ su $[a,b]$ e $m \leq \beta(x) \leq M$ per ogni $x \in [a,b]$, si ha $m\omega(x) \leq \beta(x)\omega(x) \leq M\omega(x)$ per ogni $x \in [a,b]$ e

$$
m \int_a^b \omega(x)\mathrm{d}x \leq \int_a^b \beta(x)\omega(x)\mathrm{d}x \leq M \int_a^b \omega(x)\mathrm{d}x.
$$

Consideriamo la funzione $z : [a, b] \to \mathbb{R}$,

$$z(y) = \alpha(y) \int_a^b \omega(x)\mathrm{d}x.$$

Questa funzione è continua su $[a, b]$ perché $\alpha(y)$ è continua su $[a, b]$. Quindi, per il teorema dei valori intermedi, $z(y)$ assume su $[a, b]$ tutti i valori compresi tra il suo minimo $m \int_a^b \omega(x)\mathrm{d}x$ e il suo massimo $M \int_a^b \omega(x)\mathrm{d}x$. In particolare $z(y)$ assume il valore $\int_a^b \beta(x)\omega(x)\mathrm{d}x$, ovvero esiste $\eta \in [a, b]$ tale che

$$z(\eta) = \int_a^b \beta(x)\omega(x)\mathrm{d}x. \qquad \square$$

Teorema 2.1. *Sia $f : [a, b] \to \mathbb{R}$ di classe $C^2[a, b]$ e sia I_n la formula dei trapezi di ordine n e passo $h = \frac{b-a}{n}$ per approssimare $\int_a^b f(x)\mathrm{d}x$. Allora esiste un punto $\eta \in [a, b]$ tale che*

$$\boxed{\int_a^b f(x)\mathrm{d}x - I_n = -\frac{(b-a)f''(\eta)}{12}h^2} \tag{2.2}$$

Dimostrazione. Poniamo $x_j = a + jh$ per $j = 0, \ldots, n$ e indichiamo con $s(x)$ la funzione lineare a tratti mostrata in Figura 2.1. Osserviamo che $s(x)$ coincide sull'intervallo $[x_j, x_{j+1}]$ con il polinomio (retta) d'interpolazione di $f(x)$ sui nodi x_j e x_{j+1}. Risulta

$$\int_a^b f(x)\mathrm{d}x - I_n = \int_a^b f(x)\mathrm{d}x - \int_a^b s(x)\mathrm{d}x = \int_a^b [f(x) - s(x)]\mathrm{d}x = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} [f(x) - s(x)]\mathrm{d}x$$

$$= \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} \frac{f''(\xi_j(x))}{2}(x - x_j)(x - x_{j+1})\mathrm{d}x \quad \text{(per il Teorema 1.2 applicato sull'intervallo } [x_j, x_{j+1}];$$
$$\xi_j(x) \text{ è un punto in } (x_j, x_{j+1}))$$

$$= -\sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f''(\xi_j(x))\frac{(x - x_j)(x_{j+1} - x)}{2}\mathrm{d}x$$

$$= -\sum_{j=0}^{n-1} f''(\eta_j) \int_{x_j}^{x_{j+1}} \frac{(x - x_j)(x_{j+1} - x)}{2}\mathrm{d}x \quad \text{(per il Lemma 2.1 applicato sull'intervallo } [x_j, x_{j+1}] \text{ con}$$
$$\alpha(x) = f''(x),\ \beta(x) = f''(\xi_j(x)),\ \omega(x) = \tfrac{1}{2}(x - x_j)(x_{j+1} - x);$$
$$\eta_j \text{ è un punto in } [x_j, x_{j+1}])$$

$$= -\sum_{j=0}^{n-1} f''(\eta_j) \int_0^h \frac{t(h-t)}{2}\mathrm{d}t \quad \text{(cambio variabile } t = x - x_j \iff x = t + x_j \text{ e si ha } \mathrm{d}t = \mathrm{d}x)$$

$$= -\sum_{j=0}^{n-1} f''(\eta_j)\, \frac{1}{2}\left[\frac{h}{2}\,t^2 - \frac{1}{3}\,t^3\right]_0^h = -\sum_{j=0}^{n-1} f''(\eta_j)\, \frac{h^3}{12} = -\frac{h^3 n}{12} \cdot \frac{1}{n}\sum_{j=0}^{n-1} f''(\eta_j) = -\frac{h^2(b-a)}{12} f''(\eta),$$

dove l'ultima uguaglianza vale perché, essendo $f''(x)$ continua su $[a, b]$ per ipotesi ed essendo la media aritmetica $\frac{1}{n}\sum_{j=0}^{n-1} f''(\eta_j)$ un valore compreso tra il minimo e il massimo di $f''(x)$ su $[a, b]$, per il teorema dei valori intermedi esiste sicuramente un $\eta \in [a, b]$ tale che $f''(\eta) = \frac{1}{n}\sum_{j=0}^{n-1} f''(\eta_j)$. $\square$

Esempio 2.1.

(a) Calcolare l'approssimazione di $\int_0^1 \sqrt{\cos x}\,\mathrm{d}x$ con I_{10}, la formula dei trapezi di ordine 10.

(b) Stimare l'errore che si commette approssimando l'integrale $\int_0^1 \sqrt{\cos x}\,\mathrm{d}x$ con I_{10}.

Soluzione.

(a) Per calcolare I_{10} usiamo la formula (2.1). Nel nostro caso si ha $f(x) = \sqrt{\cos x}$, $[a,b] = [0,1]$, $n = 10$, $h = \frac{1}{10}$ e $x_j = jh = \frac{j}{10}$ per $j = 0, \ldots, 10$, per cui

$$\begin{aligned} I_{10} &= \frac{1}{10}\left[\frac{\sqrt{\cos 0} + \sqrt{\cos 1}}{2} + \sum_{j=1}^{9}\sqrt{\cos\frac{j}{10}}\right] \\ &= \frac{1}{10}\left[\frac{1+\sqrt{\cos 1}}{2} + \sqrt{\cos\frac{1}{10}} + \sqrt{\cos\frac{2}{10}} + \ldots + \sqrt{\cos\frac{9}{10}}\right] = 0.9135078... \end{aligned}$$

(b) Osserviamo che $f(x) = \sqrt{\cos x}$ è di classe $C^\infty[0,1]$. Infatti, $\cos x$ non ha zeri su $[0,1]$, essendo una funzione decrescente su $[0,1]$ con $\cos x \in [\cos 1, 1] = [0.54..., 1]$ per ogni $x \in [0,1]$; quindi $f \in C^\infty[0,1]$ come composizione di $\cos x : [0,1] \to [\cos 1, 1]$ e $\sqrt{y} : [\cos 1, 1] \to \mathbb{R}$, entrambe funzioni di classe C^∞ sui loro rispettivi domini $[0,1]$ e $[\cos 1, 1]$.[3] Di conseguenza, per il Teorema 2.1,

$$\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_{10} = -\frac{1\cdot f''(\eta)}{12}\left(\frac{1}{10}\right)^2 = -\frac{1}{1200}f''(\eta),$$

dove $\eta \in [0,1]$. Calcoliamo $f''(x)$:

$$f'(x) = \frac{-\sin x}{2\sqrt{\cos x}},$$

$$f''(x) = \frac{-\cos x \cdot 2\sqrt{\cos x} - (-\sin x)\cdot 2\,\frac{-\sin x}{2\sqrt{\cos x}}}{4\cos x} = -\frac{\sqrt{\cos x}}{2} - \frac{\sin^2 x}{4\sqrt{\cos^3 x}}.$$

Per ogni $x \in [0,1]$ si ha

$$|f''(x)| \le \left|\frac{\sqrt{\cos x}}{2}\right| + \left|\frac{\sin^2 x}{4\sqrt{\cos^3 x}}\right| = \frac{\sqrt{\cos x}}{2} + \frac{\sin^2 x}{4\sqrt{\cos^3 x}} \le \frac{1}{2} + \frac{\sin^2 1}{4\sqrt{\cos^3 1}} \le 0.9458.$$

Dunque,

$$\left|\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_{10}\right| = \frac{1}{1200}|f''(\eta)| \le \frac{1}{1200}\,0.9458 = 0.000788...$$

Esempio 2.2. Fissato $\varepsilon = 10^{-8}$, determinare un n tale che la formula dei trapezi I_n fornisca un'approssimazione di $\int_0^1 \sqrt{\cos x}\,\mathrm{d}x$ con errore $\left|\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_n\right| \le \varepsilon$.

Soluzione. Posto $f(x) = \sqrt{\cos x}$, per il Teorema 2.1 si ha

$$\left|\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_n\right| = \left|-\frac{1}{12n^2}f''(\eta)\right| = \frac{|f''(\eta)|}{12n^2}, \tag{2.3}$$

dove $\eta \in [0,1]$. Nell'Esempio 2.1 abbiamo visto che $|f''(x)| \le 0.9458$ per ogni $x \in [0,1]$ e quindi dalla (2.3) risulta

$$\left|\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_n\right| = \frac{|f''(\eta)|}{12n^2} \le \frac{1}{12n^2}. \tag{2.4}$$

[3] Molto diverso è invece il caso della funzione $\sqrt{\cos x} : [0, \pi/2] \to \mathbb{R}$. In tal caso $\cos x$ ha uno zero su $[0,\pi/2]$ in $x = \pi/2$ e $\sqrt{\cos x} : [0,\pi/2] \to \mathbb{R}$ è composizione di $\cos x : [0,\pi/2] \to [0,1]$ e $\sqrt{y} : [0,1] \to \mathbb{R}$, in cui la prima funzione è $C^\infty[0,\pi/2]$, ma la seconda non è neanche $C^1[0,1]$. Si può verificare che in effetti $\sqrt{\cos x}$ non è neanche $C^1[0,\pi/2]$.

Poiché

$$\frac{1}{12n^2} \le \varepsilon \quad \iff \quad n \ge \sqrt{\frac{1}{12\varepsilon}} = n(\varepsilon),$$

dalla (2.4) si ha $\left|\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_n\right| \le \varepsilon$ per ogni $n \ge n(\varepsilon)$. Nel nostro caso abbiamo $\varepsilon = 10^{-8}$ e quindi, per garantire che $\left|\int_0^1 \sqrt{\cos x}\,\mathrm{d}x - I_n\right| \le 10^{-8}$, basta prendere un qualsiasi $n \ge n(10^{-8}) = 2886.75...$ (ad esempio, $n = 2887$).

Esempio 2.3. Fissato $\varepsilon > 0$, determinare un n tale che la formula dei trapezi I_n fornisca un'approssimazione di $\int_0^1 \frac{x+2}{\log(x+2)}\mathrm{d}x$ con errore $\left|\int_0^1 \frac{x+2}{\log(x+2)}\mathrm{d}x - I_n\right| \le \varepsilon$. Quanto vale n se $\varepsilon = 10^{-8}$?

Soluzione. Posto $f(x) = \frac{x+2}{\log(x+2)}$, per il Teorema 2.1 si ha

$$\left|\int_0^1 \frac{x+2}{\log(x+2)}\mathrm{d}x - I_n\right| = \left|-\frac{1}{12n^2}f''(\eta)\right| = \frac{|f''(\eta)|}{12n^2},$$

dove $\eta \in [0,1]$. Calcoliamo $f''(x)$:

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

$$f''(x) = \frac{\dfrac{1}{x+2}\log^2(x+2) - (\log(x+2)-1)\cdot 2\log(x+2)\dfrac{1}{x+2}}{\log^4(x+2)} = \frac{2-\log(x+2)}{(x+2)\log^3(x+2)}.$$

Per ogni $x \in [0,1]$ si ha

$$|f''(x)| = \left|\frac{2-\log(x+2)}{(x+2)\log^3(x+2)}\right| = \frac{2-\log(x+2)}{(x+2)\log^3(x+2)} \le \frac{2-\log 2}{2\log^3 2} \le 1.97.$$

Dunque,

$$\left|\int_0^1 \frac{x+2}{\log(x+2)}\mathrm{d}x - I_n\right| = \frac{|f''(\eta)|}{12n^2} \le \frac{1.97}{12n^2}. \tag{2.5}$$

Poiché

$$\frac{1.97}{12n^2} \le \varepsilon \quad \iff \quad n \ge \sqrt{\frac{1.97}{12\varepsilon}} = n(\varepsilon),$$

dalla (2.5) si ha $\left|\int_0^1 \frac{x+2}{\log(x+2)}\mathrm{d}x - I_n\right| \le \varepsilon$ per ogni $n \ge n(\varepsilon)$. Nel caso $\varepsilon = 10^{-8}$, per garantire che $\left|\int_0^1 \frac{x+2}{\log(x+2)}\mathrm{d}x - I_n\right| \le 10^{-8}$ basta prendere un qualsiasi $n \ge n(10^{-8}) = 4051.74...$ (ad esempio, $n = 4052$).

Osservazione 2.1. Negli Esempi 2.2 e 2.3, il valore $n(\varepsilon)$—quello che ci garantisce un errore $\le \varepsilon$ se prendiamo $n \ge n(\varepsilon)$—è della forma $C/\sqrt{\varepsilon}$ con C costante. Questo è un fatto generale che vale sempre e non solo per gli Esempi 2.2 e 2.3. Infatti, in base alla (2.2), detta K una costante tale che

$$|f''(x)| \le K \ \text{ per ogni } \ x \in [a,b], \tag{2.6}$$

si ha

$$\left|\int_a^b f(x)\mathrm{d}x - I_n\right| = \left|-\frac{(b-a)f''(\eta)}{12}h^2\right| = \frac{(b-a)^3|f''(\eta)|}{12n^2} \le \frac{(b-a)^3K}{12n^2}$$

e

$$\frac{(b-a)^3K}{12n^2} \le \varepsilon \quad \iff \quad n \ge \sqrt{\frac{(b-a)^3K}{12\varepsilon}} = n(\varepsilon).$$

Dunque risulta garantito che $\left|\int_a^b f(x)\mathrm{d}x - I_n\right| \le \varepsilon$ se prendiamo $n \ge n(\varepsilon) = C/\sqrt{\varepsilon}$ con $C = \sqrt{(b-a)^3K/12}$ e K che soddisfa la (2.6).

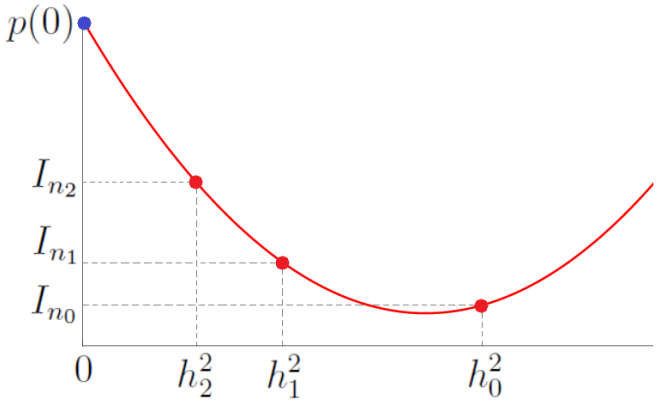

Figura 2.2: Illustrazione del procedimento di estrapolazione nel caso $m = 2$.

Esercizio 2.1. Fissato $\varepsilon > 0$, determinare un n tale che la formula dei trapezi I_n fornisca un'approssimazione di I con errore $|I - I_n| \leq \varepsilon$ nei seguenti casi.

- $I = \int_0^1 \mathrm{e}^{-x^2}\mathrm{d}x$.
- $I = \int_0^2 x\mathrm{e}^x\mathrm{d}x$.
- $I = \int_0^3 \frac{\log(1+x^2)}{\sqrt{2+\cos x}}\mathrm{d}x$.

Esercizio 2.2. Scrivere un programma MATLAB che implementa la formula dei trapezi. Il programma deve:

- prendere in input gli estremi a, b di un intervallo, una funzione $f(x)$ definita su $[a, b]$ e il numero $n \geq 1$ di sottointervalli in cui viene suddiviso $[a, b]$;
- restituire in output I_n, l'approssimazione di $\int_a^b f(x)\mathrm{d}x$ data dalla formula dei trapezi di ordine n.

Verificare sperimentalmente la correttezza del programma usandolo innanzitutto per riottenere il risultato dell'Esempio 2.1(a) e poi ancora nel modo seguente: calcolare l'approssimazione di $\int_0^2 x\mathrm{e}^x\mathrm{d}x$ ottenuta con I_n per $n = 20, 40, 80, 160, 320, 640, 1280, 2560, 5120$ e verificare che al crescere di n l'approssimazione si avvicina sempre più (converge) al valore esatto $\int_0^2 x\mathrm{e}^x\mathrm{d}x = 1 + \mathrm{e}^2 = 8.389056098930650...$

2.3 Estrapolazione

Sia $f : [a, b] \to \mathbb{R}$ (integrabile) e siano $I_{n_0}, I_{n_1}, \ldots, I_{n_m}$ le formule dei trapezi di ordini (distinti) $n_0, n_1, \ldots, n_m$ e passi $h_0 = \frac{b-a}{n_0}, h_1 = \frac{b-a}{n_1}, \ldots, h_m = \frac{b-a}{n_m}$ per approssimare $\int_a^b f(x)\mathrm{d}x$. Sia $p(x) \in \mathbb{R}_m[x]$ il polinomio d'interpolazione dei dati $(h_0^2, I_{n_0}), (h_1^2, I_{n_1}), \ldots, (h_m^2, I_{n_m})$; osserviamo che tale polinomio esiste ed è unico per il Teorema 1.1, in quanto i nodi $h_0^2, h_1^2, \ldots, h_m^2$ sono tutti distinti. In base a risultati nei quali non ci addentriamo, il valore $p(0)$ è un'approssimazione di $\int_a^b f(x)\mathrm{d}x$ *molto più accurata* rispetto alle singole formule dei trapezi $I_{n_0}, I_{n_1}, \ldots, I_{n_m}$. Il procedimento di valutare in 0 il polinomio d'interpolazione $p(x)$ prende il nome di *estrapolazione,* in quanto il polinomio d'interpolazione $p(x)$ viene valutato in un punto, precisamente $x = 0$, che si trova all'esterno del più piccolo intervallo contenente i nodi $h_0^2, h_1^2, \ldots, h_m^2$; si veda la Figura 2.2. Il valore "magico" $p(0)$ che "sconfigge clamorosamente" tutte le approssimazioni $I_{n_0}, I_{n_1}, \ldots, I_{n_m}$ si chiama anche *valore estrapolato.*

Esempio 2.4. Si consideri la funzione $f(x) = x\mathrm{e}^x$. Per ogni $n \geq 1$ indichiamo con I_n la formula dei trapezi di ordine n per approssimare $I = \int_0^2 f(x)\mathrm{d}x$.

(a) Calcolare I.

(b) Calcolare I_{12}, I_{24}, I_{30} e confrontarli con il valore esatto I.

(c) Calcolare $p(0)$, dove $p(x)$ è il polinomio d'interpolazione dei dati (h_0^2, I_{12}), (h_1^2, I_{24}), (h_2^2, I_{30}) e h_0, h_1, h_2 sono i passi di discretizzazione delle formule dei trapezi I_{12}, I_{24}, I_{30}. Confrontare inoltre $p(0)$ con il valore esatto I.

(d) Posto $\varepsilon = |p(0) - I|$, determinare un n in modo tale che la formula dei trapezi I_n fornisca un'approssimazione di I con errore $|I_n - I| \leq \varepsilon$.

Soluzione.

(a) Integrando per parti, si ha

$$\int x\mathrm{e}^x \mathrm{d}x = x\mathrm{e}^x - \int \mathrm{e}^x \mathrm{d}x = x\mathrm{e}^x - \mathrm{e}^x,$$

per cui

$$I = \int_0^2 x\mathrm{e}^x \mathrm{d}x = \left[x\mathrm{e}^x - \mathrm{e}^x\right]_0^2 = 1 + \mathrm{e}^2 = 8.3890560989...$$

(b) Siano

$$h_0 = \frac{2-0}{12} = \frac{1}{6}, \qquad h_1 = \frac{2-0}{24} = \frac{1}{12}, \qquad h_2 = \frac{2-0}{30} = \frac{1}{15}$$

i passi di discretizzazione delle formule I_{12}, I_{24}, I_{30}. Usando la formula (2.1), otteniamo

$$\begin{aligned}
I_{12} &= \frac{1}{6}\left[\frac{0 + 2\mathrm{e}^2}{2} + \sum_{j=1}^{11} \frac{j}{6}\,\mathrm{e}^{j/6}\right] = \frac{1}{6}\left[\mathrm{e}^2 + \frac{1}{6}\,\mathrm{e}^{1/6} + \frac{2}{6}\,\mathrm{e}^{2/6} + \frac{3}{6}\,\mathrm{e}^{3/6} + \ldots + \frac{11}{6}\,\mathrm{e}^{11/6}\right] \\
&= 8.4380178285... \\
I_{24} &= \frac{1}{12}\left[\frac{0 + 2\mathrm{e}^2}{2} + \sum_{j=1}^{23} \frac{j}{12}\,\mathrm{e}^{j/12}\right] = \frac{1}{12}\left[\mathrm{e}^2 + \frac{1}{12}\,\mathrm{e}^{1/12} + \frac{2}{12}\,\mathrm{e}^{2/12} + \frac{3}{12}\,\mathrm{e}^{3/12} + \ldots + \frac{23}{12}\,\mathrm{e}^{23/12}\right] \\
&= 8.4013033444... \\
I_{30} &= \frac{1}{15}\left[\frac{0 + 2\mathrm{e}^2}{2} + \sum_{j=1}^{29} \frac{j}{15}\,\mathrm{e}^{j/15}\right] = \frac{1}{15}\left[\mathrm{e}^2 + \frac{1}{15}\,\mathrm{e}^{1/15} + \frac{2}{15}\,\mathrm{e}^{2/15} + \frac{3}{15}\,\mathrm{e}^{3/15} + \ldots + \frac{29}{15}\,\mathrm{e}^{29/15}\right] \\
&= 8.3968948597...
\end{aligned}$$

Confrontando con il valore esatto I, vediamo che I_{12} e I_{24} non hanno nessuna cifra decimale esatta, mentre I_{30} ha una sola cifra decimale esatta. Si ha

$$\begin{aligned}
|I - I_{12}| &\approx 4.9 \cdot 10^{-2}, \\
|I - I_{24}| &\approx 1.2 \cdot 10^{-2}, \\
|I - I_{30}| &\approx 7.8 \cdot 10^{-3}.
\end{aligned}$$

(c) Usando la forma di Lagrange, si ha

$$\begin{aligned}
p(0) &= I_{12}\frac{(0-h_1^2)(0-h_2^2)}{(h_0^2-h_1^2)(h_0^2-h_2^2)} + I_{24}\frac{(0-h_0^2)(0-h_2^2)}{(h_1^2-h_0^2)(h_1^2-h_2^2)} + I_{30}\frac{(0-h_0^2)(0-h_1^2)}{(h_2^2-h_0^2)(h_2^2-h_1^2)} \\
&= \frac{h_1^2 h_2^2}{(h_0^2-h_1^2)(h_0^2-h_2^2)} I_{12} + \frac{h_0^2 h_2^2}{(h_1^2-h_0^2)(h_1^2-h_2^2)} I_{24} + \frac{h_0^2 h_1^2}{(h_2^2-h_0^2)(h_2^2-h_1^2)} I_{30} \\
&= \frac{4}{63}\, I_{12} - \frac{64}{27}\, I_{24} + \frac{625}{189}\, I_{30} \\
&= 8.3890561002...
\end{aligned}$$

Confrontando con il valore esatto I, si nota che $p(0)$ ha 6 cifre decimali esatte e fornisce quindi un'approssimazione molto migliore rispetto alle singole formule I_{12}, I_{24}, I_{30}. Si ha

$$|I - p(0)| \approx 1.3 \cdot 10^{-9}.$$

(d) Si ha $\varepsilon = |I - p(0)| \approx 1.3 \cdot 10^{-9}$. Per il Teorema 2.1, si ha

$$\left| \int_0^2 x\mathrm{e}^x \mathrm{d}x - I_n \right| = \left| -\frac{2^3}{12n^2} f''(\eta) \right| = \frac{2|f''(\eta)|}{3n^2},$$

dove $\eta \in [0, 2]$. Calcoliamo $f''(x)$:

$$f'(x) = \mathrm{e}^x + x\mathrm{e}^x = (1 + x)\mathrm{e}^x,$$
$$f''(x) = \mathrm{e}^x + (1 + x)\mathrm{e}^x = (2 + x)\mathrm{e}^x.$$

Per ogni $x \in [0, 2]$ si ha

$$|f''(x)| = |(2 + x)\mathrm{e}^x| = (2 + x)\mathrm{e}^x \leq 4\mathrm{e}^2.$$

Dunque,

$$\left| \int_0^2 x\mathrm{e}^x \mathrm{d}x - I_n \right| = \frac{2|f''(\eta)|}{3n^2} \leq \frac{8\mathrm{e}^2}{3n^2}. \tag{2.7}$$

Poiché

$$\frac{8\mathrm{e}^2}{3n^2} \leq \varepsilon \quad \iff \quad n \geq \sqrt{\frac{8\mathrm{e}^2}{3\varepsilon}} = n(\varepsilon),$$

dalla (2.7) si ha $|\int_0^2 x\mathrm{e}^x \mathrm{d}x - I_n| \leq \varepsilon$ per ogni $n \geq n(\varepsilon)$. Nel caso $\varepsilon = 1.3 \cdot 10^{-9}$, per garantire che $|\int_0^2 x\mathrm{e}^x \mathrm{d}x - I_n| \leq \varepsilon$ basta prendere un qualsiasi $n \geq n(1.3 \cdot 10^{-9}) = 123113.92...$

Osservazione. Il risultato ottenuto nel punto (d) mostra che, per garantire mediante la formula dei trapezi I_n un'approssimazione di I con una precisione $\varepsilon = 1.3 \cdot 10^{-9}$ pari a quella fornita da $p(0)$, occorre prendere $n = 123114$ molto grande. Non vale quindi la pena calcolare I_n per un n così grande, considerato che la stessa precisione può essere molto più facilmente ottenuta calcolando prima I_{12}, I_{24}, I_{30} e poi il valore estrapolato $p(0)$ come abbiamo fatto nei punti (b)–(c) di questo esempio. □

Esercizio 2.3. Consideriamo i seguenti due casi:
- $f(x) = \mathrm{e}^{-x}$ e $[a, b] = [0, 1]$;
- $f(x) = \log x$ e $[a, b] = [1, 2]$.

Per ciascuno di questi casi, indichiamo con I_n la formula dei trapezi di ordine n per approssimare $I = \int_a^b f(x)\mathrm{d}x$.

(a) Calcolare I.
(b) Calcolare I_3, I_6, I_{12} e confrontarli con il valore esatto I.
(c) Calcolare $p(0)$, dove $p(x)$ è il polinomio d'interpolazione dei dati (h_0^2, I_3), (h_1^2, I_6), (h_2^2, I_{12}) e h_0, h_1, h_2 sono i passi di discretizzazione delle formule dei trapezi I_3, I_6, I_{12}. Confrontare inoltre $p(0)$ con il valore esatto I.
(d) Posto $\varepsilon = |p(0) - I|$, determinare un n in modo tale che la formula dei trapezi I_n fornisca un'approssimazione di I con errore $|I_n - I| \leq \varepsilon$.

Esercizio 2.4. Usando i programmi creati per risolvere gli Esercizi 1.11 e 2.2, scrivere un programma MATLAB che implementa il metodo di estrapolazione. Il programma deve:
- prendere in input gli estremi a, b di un intervallo, una funzione $f(x)$ definita su $[a, b]$ e un vettore $[n_0, n_1, \ldots, n_m]$ di numeri $n_0, n_1, \ldots, n_m \geq 1$ tutti distinti;

- restituire in output il valore estrapolato $p(0)$, dove $p(x)$ è il polinomio d'interpolazione dei dati (h_0^2, I_{n_0}), (h_1^2, I_{n_1}), …, (h_m^2, I_{n_m}) e h_0, h_1, …, h_m sono i passi di discretizzazione delle formule dei trapezi I_{n_0}, I_{n_1}, …, I_{n_m} per approssimare $\int_a^b f(x)\mathrm{d}x$.

Verificare la correttezza del programma usandolo in particolare per riottenere il risultato dell'Esempio 2.4(c).

3 Analisi di matrici

3.1 Richiami di algebra lineare

3.1.1 Calcolo dei determinanti

Il metodo di Laplace è il metodo "classico" con cui si calcolano i determinanti delle matrici. Per ripassarlo, utilizziamolo per calcolare il determinante della matrice

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 1 & 4 \\ 2 & 5 & 10 \end{bmatrix}.$$

Per prima cosa, ricordiamo come funziona:
- si sceglie una riga oppure una colonna della matrice;
- si sviluppa il determinante lungo quella riga o colonna tenendo conto della cosiddetta regola della scacchiera per la determinazione dei segni:

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}.$$

Scegliendo la prima riga, si ha

$$\begin{aligned} \det(A) = \begin{vmatrix} 1 & 3 & 2 \\ 0 & 1 & 4 \\ 2 & 5 & 10 \end{vmatrix} &= +1 \begin{vmatrix} 1 & 4 \\ 5 & 10 \end{vmatrix} - 3 \begin{vmatrix} 0 & 4 \\ 2 & 10 \end{vmatrix} + 2 \begin{vmatrix} 0 & 1 \\ 2 & 5 \end{vmatrix} \\ &= +1(1 \cdot 10 - 5 \cdot 4) - 3(0 \cdot 10 - 2 \cdot 4) + 2(0 \cdot 5 - 2 \cdot 1) = -10 + 24 - 4 = 10. \end{aligned}$$

Scegliendo la seconda colonna, si ha

$$\begin{aligned} \det(A) = \begin{vmatrix} 1 & 3 & 2 \\ 0 & 1 & 4 \\ 2 & 5 & 10 \end{vmatrix} &= -3 \begin{vmatrix} 0 & 4 \\ 2 & 10 \end{vmatrix} + 1 \begin{vmatrix} 1 & 2 \\ 2 & 10 \end{vmatrix} - 5 \begin{vmatrix} 1 & 2 \\ 0 & 4 \end{vmatrix} \\ &= -3(0 \cdot 10 - 2 \cdot 4) + 1(1 \cdot 10 - 2 \cdot 2) - 5(1 \cdot 4 - 0 \cdot 2) = 24 + 6 - 20 = 10. \end{aligned}$$

Il risultato è sempre lo stesso, qualunque riga o colonna si scelga. Il determinante di A è uguale a 10.

Oltre al metodo di Laplace, un teorema da tener presente per il calcolo dei determinanti è il teorema di Binet. Date due matrici $A, B \in \mathbb{C}^{n\times n}$, il teorema di Binet stabilisce che

$$\det(AB) = \det(A)\det(B).$$

Ricordiamo infine che il determinante di una matrice A è uguale a quello della sua trasposta A^T: per ogni $A \in \mathbb{C}^{n\times n}$ si ha

$$\det(A) = \det(A^T).$$

Esempio 3.1. Sia α un parametro reale e sia

$$U = \begin{bmatrix} 2 & \alpha \\ 1 & 3 \end{bmatrix}.$$

Stabilire per quali valori di α la matrice $A = LU$ ha determinante nullo nei seguenti due casi:

- $L = \begin{bmatrix} 1 & -3 \\ -4 & 12 \end{bmatrix}$;
- $L = \begin{bmatrix} 1 & 0 \\ 5 & 2 \end{bmatrix}$.

Soluzione. In base al teorema di Binet,

$$\det(A) = \det(L)\det(U) = \det(L)(6 - \alpha).$$

- Nel primo caso si ha $\det(L) = 0$ per cui $\det(A) = 0$ e dunque la matrice A ha sempre determinante nullo qualunque sia il valore di α.
- Nel secondo caso si ha $\det(L) = 2$ per cui $\det(A) = 2(6-\alpha)$ e dunque la matrice A ha determinante nullo se e solo se $\alpha = 6$.

3.1.2 Traccia, determinante, raggio spettrale e autovalori

Data una matrice $A \in \mathbb{C}^{n\times n}$ con autovalori $\lambda_1, \lambda_2, \ldots, \lambda_n$ (ciascuno dei quali compare nella sequenza appena scritta un numero di volte pari alla sua molteplicità algebrica come radice del polinomio caratteristico di A), si ha

$$\begin{aligned} \text{traccia}(A) &\stackrel{\text{def}}{=} a_{11} + a_{22} + \ldots + a_{nn} = \lambda_1 + \lambda_2 + \ldots + \lambda_n, \\ \det(A) &= \lambda_1\lambda_2 \cdots \lambda_n, \\ \rho(A) &\stackrel{\text{def}}{=} \text{raggio spettrale di } A \stackrel{\text{def}}{=} \max(|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|). \end{aligned}$$

Esempio 3.2. Sia

$$A = \begin{bmatrix} 2 & 4 & -1 & 0 \\ 0 & 1+\mathrm{i} & 0 & 7 \\ 1 & -1 & 1 & -2\mathrm{i} \\ 3 & -1 & 0 & -12 \end{bmatrix}.$$

(i) Dimostrare che A possiede almeno un autovalore λ non reale (cioè tale che $\mathrm{Im}(\lambda) \neq 0$).
(ii) Dimostrare che A possiede almeno un autovalore μ con parte reale minore o uguale a -2 (cioè tale che $\mathrm{Re}(\mu) \leq -2$).

Soluzione.

(i) Notiamo che $\text{traccia}(A) = -8 + \mathrm{i}$ non è reale. Siccome la traccia è la somma degli autovalori, per forza di cose deve esistere almeno un autovalore λ non reale, perché se tutti gli autovalori fossero reali allora anche la traccia sarebbe reale.
(ii) Deve esistere per forza almeno un autovalore μ con parte reale minore o uguale a -2, perché se tutti gli autovalori $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ avessero parte reale maggiore di -2 allora la traccia avrebbe parte reale maggiore di -8:

$$\begin{aligned} \mathrm{Re}(\text{traccia}(A)) &= \mathrm{Re}(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4) \\ &= \mathrm{Re}(\lambda_1) + \mathrm{Re}(\lambda_2) + \mathrm{Re}(\lambda_3) + \mathrm{Re}(\lambda_4) \\ &> -2 + (-2) + (-2) + (-2) = -8. \end{aligned}$$

Osservazione. Con lo stesso tipo di ragionamento con cui abbiamo dimostrato che A deve avere almeno un autovalore con parte reale ≤ -2, si può dimostrare che A deve avere almeno un autovalore con parte immaginaria $\geq \frac{1}{4}$ (i dettagli della dimostrazione sono lasciati come esercizio al lettore). Quest'ultima è un'informazione più precisa rispetto a quella di dire semplicemente che A deve avere almeno un autovalore non reale.

Esercizio 3.1. Si considerino le matrici

$$A = \begin{bmatrix} \frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & -2 & 5 & 10 \\ 0 & -3 & -7 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

(i) Calcolare gli autovalori di A e B.
(ii) Verificare che per B, come avviene per ogni matrice triangolare, gli autovalori sono gli elementi diagonali.
(iii) Calcolare la traccia, il determinante e il raggio spettrale di A e B.
(iv) Verificare che sia per A che per B la somma degli autovalori è uguale alla traccia e il prodotto degli autovalori è uguale al determinante.

3.1.3 Matrici invertibili

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice invertibile se esiste una matrice $B \in \mathbb{C}^{n\times n}$ tale che $AB = BA = I$.[4] In tal caso, la matrice B è univocamente determinata, prende il nome di matrice inversa di A e viene denotata con A^{-1}. Ricordiamo che una matrice $A \in \mathbb{C}^{n\times n}$ è invertibile se e solo se $\det(A) \neq 0$ se e solo se 0 non è un autovalore di A. Ricordiamo inoltre che il prodotto AB di due matrici $A, B \in \mathbb{C}^{n\times n}$ è invertibile se e solo se A e B sono invertibili;[5] l'inversa in tal caso è $(AB)^{-1} = B^{-1}A^{-1}$ come si può verificare direttamente: $ABB^{-1}A^{-1} = B^{-1}A^{-1}AB = I$.

Esempio 3.3 (**calcolo dell'inversa di una matrice**)**.** Verificare che la matrice

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 1 \\ 1 & 0 & 4 \end{bmatrix}$$

è invertibile e calcolare la sua inversa.

Soluzione. Per verificare che A è invertibile, calcoliamo il suo determinante (con il metodo di Laplace) e verifichiamo che è diverso da 0. Scegliendo l'ultima riga di A (che ha uno zero e quindi semplifica i calcoli), si ha

$$\det(A) = \begin{vmatrix} 1 & 2 & -1 \\ -2 & 1 & 1 \\ 1 & 0 & 4 \end{vmatrix} = +1 \begin{vmatrix} 2 & -1 \\ 1 & 1 \end{vmatrix} - 0 \begin{vmatrix} 1 & -1 \\ -2 & 1 \end{vmatrix} + 4 \begin{vmatrix} 1 & 2 \\ -2 & 1 \end{vmatrix} = 3 + 20 = 23 \neq 0,$$

[4] Useremo i simboli I e O per indicare rispettivamente la matrice identità e la matrice nulla.
[5] Questo lo si può dimostrare anche con il teorema di Binet (lo si faccia per esercizio).

dunque A è invertibile. Ricordiamo ora che l'inversa A^{-1} is calcola usando la seguente formula:

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} \begin{vmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{vmatrix} & \begin{vmatrix} 0 & 2 & -1 \\ 1 & 1 & 1 \\ 0 & 0 & 4 \end{vmatrix} & \begin{vmatrix} 0 & 2 & -1 \\ 0 & 1 & 1 \\ 1 & 0 & 4 \end{vmatrix} \\ \begin{vmatrix} 1 & 1 & -1 \\ -2 & 0 & 1 \\ 1 & 0 & 4 \end{vmatrix} & \begin{vmatrix} 1 & 0 & -1 \\ -2 & 1 & 1 \\ 1 & 0 & 4 \end{vmatrix} & \begin{vmatrix} 1 & 0 & -1 \\ -2 & 0 & 1 \\ 1 & 1 & 4 \end{vmatrix} \\ \begin{vmatrix} 1 & 2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 0 \end{vmatrix} & \begin{vmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \\ 1 & 0 & 0 \end{vmatrix} & \begin{vmatrix} 1 & 2 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{vmatrix} \end{bmatrix} = \frac{1}{23} \begin{bmatrix} 4 & -8 & 3 \\ 9 & 5 & 1 \\ -1 & 2 & 5 \end{bmatrix}.$$

Osservazione. Il metodo di calcolo dell'inversa appena visto coincide sostanzialmente con il cosiddetto metodo dei cofattori.

Esercizio 3.2. Stabilire quali delle seguenti matrici sono invertibili e calcolarne l'inversa.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & -2 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} -2 & 4 & 1 \\ -1 & 2 & 7 \\ 1 & 0 & -1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 4 & 2 & 0 \\ -1 & 2 & 1 & -1 \\ -1 & -3 & -3 & 2 \\ -1 & 3 & 0 & 1 \end{bmatrix}.$$

3.1.4 Matrici diagonalizzabili

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice diagonalizzabile se esistono una matrice invertibile $X \in \mathbb{C}^{n\times n}$ e una matrice diagonale $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{C}^{n\times n}$ tali che

$$A = XDX^{-1}. \tag{3.1}$$

Nella (3.1) c'è scritto che per ogni $i = 1, \ldots, n$ l'elemento diagonale λ_i è un autovalore di A con corrispondente autovettore $\mathbf{x}_i = i$-esima colonna di X. Questo fatto lo si vede moltiplicando a destra per X entrambi i membri della (3.1) ottenendo $AX = XD$; da qui si nota che la colonna i-esima di AX è $A\mathbf{x}_i$ e la colonna i-esima di XD è $\lambda_i\mathbf{x}_i$, per cui $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$. Ricordiamo che ogni matrice $A \in \mathbb{C}^{n\times n}$ che possiede n autovalori *distinti* è diagonalizzabile.

3.1.5 Matrici hermitiane e simmetriche

Data una matrice $A \in \mathbb{C}^{m\times n}$, indichiamo con A^* la trasposta coniugata di A. Se A e B sono matrici moltiplicabili, allora

$$(AB)^T = B^TA^T, \qquad (AB)^* = B^*A^*.$$

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice hermitiana se $A^* = A$. Nel caso in cui le componenti di A sono reali (cioè $A \in \mathbb{R}^{n\times n}$), si ha $A^T = A^*$, per cui dire che A è hermitiana è equivalente a dire che A è simmetrica (cioè $A^T = A$). Gli elementi diagonali di una matrice hermitiana A sono uguali ai loro coniugati e dunque sono reali. Anche gli autovalori di una matrice hermitiana A sono reali. Infatti, se λ è un autovalore di A e indichiamo con $\mathbf{x} \neq \mathbf{0}$ un corrispondente autovettore, allora [6]

$$A\mathbf{x} = \lambda\mathbf{x} \implies \mathbf{x}^*A\mathbf{x} = \mathbf{x}^*(\lambda\mathbf{x}) = \lambda\mathbf{x}^*\mathbf{x} = \lambda\sum_{i=1}^{n}\overline{x_i}x_i = \lambda\sum_{i=1}^{n}|x_i|^2 \implies \lambda = \frac{\mathbf{x}^*A\mathbf{x}}{\sum_{i=1}^{n}|x_i|^2} \in \mathbb{R}$$

[6] Precisiamo che ogni vettore viene sempre pensato come vettore colonna, per cui il suo trasposto coniugato $\mathbf{x}^*$ è un vettore riga le cui componenti sono le coniugate di quelle di $\mathbf{x}$.

perché $\mathbf{x}^*A\mathbf{x}$ è un numero reale essendo uguale al suo complesso coniugato:

$$\overline{\mathbf{x}^*A\mathbf{x}} = (\mathbf{x}^*A\mathbf{x})^* = \mathbf{x}^*A^*(\mathbf{x}^*)^* = \mathbf{x}^*A\mathbf{x}. \tag{3.2}$$

3.2 Matrici definite positive

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice definita positiva se $\mathrm{Re}(\mathbf{x}^*A\mathbf{x}) > 0$ per ogni $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$. Notiamo che, per ogni $A \in \mathbb{C}^{n\times n}$ e ogni $\mathbf{x} \in \mathbb{C}^n$, si ha

$$\mathrm{Re}(\mathbf{x}^*A\mathbf{x}) = \frac{\mathbf{x}^*A\mathbf{x} + \overline{\mathbf{x}^*A\mathbf{x}}}{2} = \frac{\mathbf{x}^*A\mathbf{x} + (\mathbf{x}^*A\mathbf{x})^*}{2} = \frac{\mathbf{x}^*A\mathbf{x} + \mathbf{x}^*A^*\mathbf{x}}{2} = \mathbf{x}^*\Big(\frac{A+A^*}{2}\Big)\mathbf{x} = \mathbf{x}^*\mathrm{Re}(A)\mathbf{x}, \tag{3.3}$$

dove

$$\mathrm{Re}(A) \stackrel{\text{def}}{=} \frac{A+A^*}{2}, \qquad \mathrm{Im}(A) \stackrel{\text{def}}{=} \frac{A-A^*}{2\mathrm{i}}, \qquad A = \mathrm{Re}(A) + \mathrm{i}\,\mathrm{Im}(A).$$

Si osservi che la parte reale $\mathrm{Re}(A)$ e la parte immaginaria $\mathrm{Im}(A)$ di una matrice $A \in \mathbb{C}^{n\times n}$ sono matrici hermitiane.[7] Come conseguenza di (3.3) (oppure della proprietà (3.2) valida per matrici hermitiane), si ha che $\mathbf{x}^*\mathrm{Re}(A)\mathbf{x} \in \mathbb{R}$ per ogni $A \in \mathbb{C}^{n\times n}$ e ogni $\mathbf{x} \in \mathbb{C}^n$. Pertanto,

$$\begin{aligned}
A \text{ è definita positiva} \quad &\iff \quad \mathrm{Re}(\mathbf{x}^*A\mathbf{x}) > 0 \text{ per ogni } \mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\} \\
&\iff \quad \mathbf{x}^*\mathrm{Re}(A)\mathbf{x} > 0 \text{ per ogni } \mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\} \\
&\iff \quad \mathrm{Re}(\mathbf{x}^*\mathrm{Re}(A)\mathbf{x}) > 0 \text{ per ogni } \mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\} \\
&\iff \quad \mathrm{Re}(A) \text{ è definita positiva.}
\end{aligned}$$

Ogni matrice definita positiva $A \in \mathbb{C}^{n\times n}$ è sicuramente invertibile perché i suoi autovalori hanno parte reale positiva (e quindi sono diversi da 0). Infatti, se λ è un autovalore di A e indichiamo con $\mathbf{x} \neq \mathbf{0}$ un corrispondente autovettore, allora

$$\begin{aligned}
A\mathbf{x} = \lambda\mathbf{x} \quad &\Longrightarrow \quad \mathbf{x}^*A\mathbf{x} = \mathbf{x}^*(\lambda\mathbf{x}) = \lambda\mathbf{x}^*\mathbf{x} = \lambda\sum_{i=1}^n \overline{x_i}x_i = \lambda\sum_{i=1}^n |x_i|^2 \\
&\Longrightarrow \quad \lambda = \frac{\mathbf{x}^*A\mathbf{x}}{\sum_{i=1}^n |x_i|^2} \quad \Longrightarrow \quad \mathrm{Re}(\lambda) = \frac{\mathrm{Re}(\mathbf{x}^*A\mathbf{x})}{\sum_{i=1}^n |x_i|^2} > 0.
\end{aligned}$$

Enunciamo senza dimostrazione il seguente importante risultato sulle matrici hermitiane.

Teorema 3.1. *Sia $A \in \mathbb{C}^{n\times n}$ una matrice hermitiana e siano $A_1, A_2, \ldots, A_n$ le sue sottomatrici principali di testa:*

$$A_1 = [a_{11}], \qquad A_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad A_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \qquad \ldots, \qquad A_n = A.$$

Le seguenti condizioni sono equivalenti:
- *A è definita positiva;*
- *$\mathbf{x}^*A\mathbf{x} > 0$ per ogni $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$;*
- *gli autovalori di A sono reali e positivi;*
- *$\det(A_k) > 0$ per ogni $k = 1, \ldots, n$.*

[7] Lo si dimostri per esercizio sfruttando il fatto che $(\alpha B)^* = \overline{\alpha}B^*$ per ogni $\alpha \in \mathbb{C}$ e ogni matrice B.

Esempio 3.4. Dire se la matrice

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

è definita positiva.

Soluzione. Osserviamo che A è hermitiana (reale e simmetrica). Pertanto, grazie al Teorema 3.1, A è definita positiva se e solo se $\det(A_k) > 0$ per ogni $k = 1, 2, 3$. Si ha

$$\begin{aligned} &\det(A_1) = 2 > 0, \\ &\det(A_2) = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3 > 0, \\ &\det(A_3) = \begin{vmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 1 \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 2 & 0 \end{vmatrix} + \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = -2 + 3 = 1 > 0. \end{aligned}$$

Dunque A è definita positiva.

Esempio 3.5. Dire se la matrice

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

è definita positiva.

Soluzione. La matrice A non è hermitiana, dunque non possiamo applicare direttamente il Teorema 3.1 ad A per stabilire se A è definita positiva. Tuttavia, sappiamo che A è definita positiva se e solo se

$$\mathrm{Re}(A) = \frac{A + A^*}{2} = \frac{A + A^T}{2} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

è definita positiva. Siccome $\mathrm{Re}(A)$ è hermitiana, per stabilire se $\mathrm{Re}(A)$ è definita positiva possiamo applicare il Teorema 3.1. Questo è quello che abbiamo fatto nell'Esempio 3.4 e che ci ha permesso di dire che $\mathrm{Re}(A)$ è definita positiva. Dunque A è definita positiva.

Esercizio 3.3. Sia $A \in \mathbb{C}^{n\times n}$ una matrice hermitiana definita positiva. Dimostrare che gli elementi diagonali di A sono tutti positivi, cioè $a_{ii} > 0$ per ogni $i = 1, \ldots, n$.

Suggerimento. Ricordiamo che, per il Teorema 3.1, si ha $\mathbf{x}^* A\mathbf{x} > 0$ per ogni $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$. Indichiamo con $\mathbf{e}_1, \ldots, \mathbf{e}_n$ i vettori della base canonica di $\mathbb{C}^n$, nel senso che $\mathbf{e}_i$ è il vettore con tutte le componenti uguali a 0 tranne quella in posizione i-esima che vale 1. Fissati due indici $i, j \in \{1, \ldots, n\}$, che cos'è $\mathbf{e}_i^* A\mathbf{e}_j$? E che cos'è dunque $\mathbf{e}_i^* A\mathbf{e}_i$?

Esercizio 3.4. Dire quali fra le seguenti matrici sono definite positive.

$$A = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 3 & -4 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}, \qquad C = \begin{bmatrix} 4 & 2+\mathrm{i} & 1 \\ 2-\mathrm{i} & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 2 & 0 & -1 \\ 1 & 1 & -2 \\ -1 & -1 & 1 \end{bmatrix}.$$

3.3 Polinomi di matrici

Se $p(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 + \ldots + a_m\lambda^m$ è un polinomio e $A \in \mathbb{C}^{n\times n}$ è una matrice, definiamo la matrice $p(A) = a_0 I + a_1 A + a_2 A^2 + \ldots + a_m A^m$. Ad esempio, se $p(\lambda) = 1 - 2\lambda^2 + \lambda^3$ allora $p(A) = I - 2A^2 + A^3$. Vale il seguente risultato.

Teorema 3.2. *Sia $p(\lambda)$ un polinomio e sia $A \in \mathbb{C}^{n\times n}$ una matrice con autovalori $\lambda_1, \ldots, \lambda_n$. Allora gli autovalori della matrice $p(A)$ sono $p(\lambda_1), \ldots, p(\lambda_n)$.*

Dimostrazione. Dimostriamo il teorema soltanto in tre casi.

Caso 1. Il polinomio $p(\lambda) = a_0$ è costante. In tal caso, $p(A) = a_0 I$ e i suoi autovalori sono $a_0, \ldots, a_0$ (ripetuto n volte). Dunque gli autovalori di $p(A)$ sono $p(\lambda_1), \ldots, p(\lambda_n)$ e la tesi del teorema vale.

Caso 2. Il polinomio $p(\lambda) = a_0 + a_1\lambda$ ha grado 1. In tal caso, il polinomio caratteristico di $p(A)$ e quello di A sono legati dalla seguente relazione: per ogni $\lambda \in \mathbb{C}$,

$$\begin{aligned} C_{p(A)}(\lambda) &= \det(\lambda I - p(A)) = \det(\lambda I - (a_0 I + a_1 A)) = \det((\lambda - a_0)I - a_1 A) = \det\Big(a_1\Big(\frac{\lambda - a_0}{a_1} I - A\Big)\Big) \\ &= a_1^n \det\Big(\frac{\lambda - a_0}{a_1} I - A\Big) = a_1^n\, C_A\Big(\frac{\lambda - a_0}{a_1}\Big). \end{aligned}$$

Dunque gli autovalori di $p(A)$ sono

$$\begin{aligned} \{\lambda \in \mathbb{C} :\ C_{p(A)}(\lambda) = 0\} &= \Big\{\lambda \in \mathbb{C} :\ C_A\Big(\frac{\lambda - a_0}{a_1}\Big) = 0\Big\} = \Big\{\lambda \in \mathbb{C} :\ \frac{\lambda - a_0}{a_1} = \lambda_1, \ldots, \lambda_n\Big\} \\ &= \Big\{\lambda \in \mathbb{C} :\ \lambda = a_0 + a_1\lambda_1, \ldots, a_0 + a_1\lambda_n\Big\} = \{a_0 + a_1\lambda_1, \ldots, a_0 + a_1\lambda_n\} \\ &= \{p(\lambda_1), \ldots, p(\lambda_n)\}. \end{aligned}$$

Caso 3. La matrice A è diagonalizzabile. In tal caso, esistono una matrice invertibile X e una matrice diagonale $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ (avente sulla diagonale gli autovalori di A) tali che

$$\begin{aligned} A &= XDX^{-1}, \\ A^2 &= XDX^{-1}XDX^{-1} = XD^2X^{-1}, \\ A^3 &= XDX^{-1}XDX^{-1}XDX^{-1} = XD^3X^{-1}, \\ &\vdots \end{aligned}$$

Pertanto, fissato un polinomio $p(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 + \ldots + a_m\lambda^m$, si ha

$$p(A) = a_0 I + a_1 A + a_2 A^2 + \ldots + a_m A^m = X(a_0 I + a_1 D + a_2 D^2 + \ldots + a_m D^m)X^{-1} = Xp(D)X^{-1}, \quad (3.4)$$

dove

$$p(D) = a_0 I + a_1 D + a_2 D^2 + \ldots + a_m D^m = \begin{bmatrix} p(\lambda_1) & & & \\ & p(\lambda_2) & & \\ & & \ddots & \\ & & & p(\lambda_n) \end{bmatrix}. \quad (3.5)$$

Le equazioni (3.4)–(3.5) ci dicono che $p(A)$ è diagonalizzabile con autovalori $p(\lambda_1), \ldots, p(\lambda_n)$ (e con autovettori $\mathbf{x}_1, \ldots, \mathbf{x}_n$ uguali a quelli di A, dati dalle colonne di X). □

Esempio 3.6. Sia

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

Calcolare gli autovalori della matrice $B = 2I + A - 3A^7$.

Soluzione. La matrice B è un polinomio della matrice A, infatti

$$B = 2I + A - 3A^7 = p(A) \text{ con } p(\lambda) = 2 + \lambda - 3\lambda^7.$$

Pertanto, per calcolare gli autovalori di B, non è necessario calcolare esplicitamente B per poi determinarne gli autovalori: basta infatti calcolare gli autovalori di A e applicare il Teorema 3.2. Siccome A è una matrice triangolare superiore, i suoi autovalori sono gli elementi diagonali $1, -3, 1$.[8] Dunque, per il Teorema 3.2, gli autovalori di B sono

$$\begin{aligned} p(1) &= 2 + 1 - 3 \cdot 1^7 = 0, \\ p(-3) &= 2 - 3 - 3 \cdot (-3)^7 = 6560, \\ p(1) &= 2 + 1 - 3 \cdot 1^7 = 0. \end{aligned}$$

Osservazione. B non è invertibile perché 0 è un autovalore di B. Invece A è invertibile perché 0 non è un autovalore di A.

3.4 Matrici irriducibili

Un grafo è un diagramma formato da un certo numero di nodi e da un certo numero di archi. Un arco è semplicemente una freccia che parte da un nodo e arriva a un altro nodo (che può anche coincidere con quello di partenza). Se il grafo possiede n nodi, questi vengono tipicamente indicati con i numeri $1, \dots, n$, mentre l'arco che va dal nodo i al nodo j viene tipicamente indicato con una freccia $i \to j$ che parte da i e arriva a j. Un *cammino* all'interno di un grafo è un percorso che parte da un nodo i e, seguendo gli archi del grafo, arriva a un altro nodo j. Se il nodo di arrivo j coincide con il nodo di partenza i, il cammino si chiama anche *ciclo*. Un grafo si dice *fortemente connesso* se per ogni coppia di nodi i e j esiste un cammino all'interno del grafo che va da i a j. Equivalentemente, un grafo è fortemente connesso se esiste un ciclo nel grafo che tocca tutti i nodi.[9] La Figura 3.1 mostra due grafi, uno fortemente connesso e uno no.

Data una matrice $A \in \mathbb{C}^{n \times n}$, il grafo associato ad A è il grafo così definito:

- i nodi sono $1, \dots, n$;
- gli archi sono le frecce $i \to j$ tali che $a_{ij} \neq 0$.

In pratica quindi, nel grafo di A ho un arco da i a j se e solo se l'elemento di A in posizione (i, j) è diverso da 0. Una matrice $A \in \mathbb{C}^{n \times n}$ si dice *irriducibile* se il suo grafo è fortemente connesso.

Esempio 3.7. Stabilire se le matrici

$$A = \begin{bmatrix} 0 & \mathrm{i} & 0 \\ 1 & 1 & 4 \\ 0 & \mathrm{i} & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & -1 & 0 \\ \mathrm{i} & 2 & -3 \\ 0 & 0 & 0 \end{bmatrix}$$

sono irriducibili oppure no.

[8] Si lascia come esercizio al lettore quello di calcolare gli autovalori di A e verificare che essi sono dati proprio dagli elementi diagonali $1, -3, 1$. Ricordiamo che, come già osservato nell'Esercizio 3.1, gli autovalori di una matrice triangolare sono gli elementi diagonali.

[9] Si dimostri per esercizio questa equivalenza, cioè si dimostri che se un grafo è fortemente connesso allora esiste un ciclo nel grafo che tocca tutti i nodi e, viceversa, se esiste un ciclo nel grafo che tocca tutti i nodi allora il grafo è fortemente connesso.

Figura 3.1: Esempi di grafi. Sinistra: grafo fortemente connesso (il ciclo $1 \to 2 \to 3 \to 2 \to 1$ tocca tutti i nodi per cui da ogni nodo i si può andare in un qualsiasi altro nodo j). Destra: grafo non fortemente connesso (dal nodo 3 non si può raggiungere il nodo 1 (e nemmeno il nodo 2)).

Soluzione. Il grafo della matrice A è così definito:
- i nodi sono $1, 2, 3$;
- gli archi sono $1 \to 2$, $2 \to 1$, $2 \to 2$ (cappio), $2 \to 3$, $3 \to 2$.

Il grafo di A è quello mostrato in Figura 3.1 (sinistra) ed è fortemente connesso, per cui A è irriducibile. Il grafo della matrice B è così definito:
- i nodi sono $1, 2, 3$;
- gli archi sono $1 \to 2$, $2 \to 1$, $2 \to 2$, $2 \to 3$.

Il grafo di B è quello mostrato in Figura 3.1 (destra) e non è fortemente connesso, per cui B non è irriducibile (si dice anche in tal caso che B è riducibile).

Esercizio 3.5. Stabilire quali delle seguenti matrici sono irriducibili.

$$A = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & \mathrm{i} \\ 1+\mathrm{i} & 0 & 0 & -1 \\ -2 & 7 & 10 & 3\mathrm{i} \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 & 3 & 4 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 4 & 7 \\ 0 & -\mathrm{i} & 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

3.5 Localizzazione degli autovalori

Denotiamo con $\mathscr{C}(z_0, r) = \{z \in \mathbb{C} : \ |z - z_0| \leq r\}$ il cerchio nel piano complesso $\mathbb{C}$ di centro z_0 e raggio r. Data una matrice $A \in \mathbb{C}^{n \times n}$, i cerchi di Gershgorin di A sono i cerchi $K_1, \ldots, K_n$ definiti nel modo seguente: per ogni $i = 1, \ldots, n$,

$$K_i = \mathscr{C}(a_{ii}, |a_{i1}| + \ldots + |a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{in}|) = \mathscr{C}\left(a_{ii}, \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\right).$$

Poiché il raggio del cerchio K_i è la somma dei moduli degli elementi sulla riga i-esima di A (escluso l'elemento diagonale a_{ii}), i cerchi di Gershgorin di A vengono chiamati anche cerchi di Gershgorin per riga di A. Questo serve a distinguerli dai cerchi di Gershgorin per colonna di A che sono i cerchi $H_1, \ldots, H_n$ definiti nel modo seguente: per ogni $j = 1, \ldots, n$,

$$H_j = \mathscr{C}(a_{jj}, |a_{1j}| + \ldots + |a_{j-1,j|} + |a_{j+1,j}| + \ldots + |a_{nj}|) = \mathscr{C}\left(a_{jj}, \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|\right).$$

Ad ogni modo, *quando si parla di cerchi di Gershgorin senza altre specificazioni, s'intendono i cerchi di Gershgorin per riga.* Quando si vuole invece parlare dei cerchi di Gershgorin per colonna, questo va specificato ogni volta.

Esempio 3.8. Determinare i cerchi di Gershgorin della matrice A dell'Esercizio 3.5.

Soluzione. Non essendo specificato diversamente, si sta parlando dei cerchi di Gershgorin per riga di A. Tali cerchi sono $K_1 = \mathscr{C}(0,1)$, $K_2 = \mathscr{C}(0,3)$, $K_3 = \mathscr{C}(0, 1+\sqrt{2})$, $K_4 = \mathscr{C}(3\mathrm{i}, 19)$.

Teorema 3.3 (**primo teorema di Gershgorin**). *Gli autovalori di una matrice $A \in \mathbb{C}^{n\times n}$ stanno tutti nell'unione dei cerchi di Gershgorin di A.*

Dimostrazione. Sia λ un generico autovalore di A. Mostriamo che λ appartiene ad almeno un cerchio di Gershgorin di A (e quindi sta nell'unione dei cerchi di Gershgorin di A). Prendiamo $\mathbf{u} \neq \mathbf{0}$ autovettore di A corrispondente a λ. Si ha

$$A\mathbf{u} = \lambda\mathbf{u} \quad \iff \quad (A\mathbf{u})_i = (\lambda\mathbf{u})_i \text{ per ogni } i = 1,\dots,n \quad \iff \quad \sum_{j=1}^{n} a_{ij}u_j = \lambda u_i \text{ per ogni } i = 1,\dots,n.$$

Selezionando un indice i corrispondente a una componente u_i di modulo massimo di $\mathbf{u}$, la precedente equazione i-esima ci dice che

$$\begin{aligned}
\sum_{j=1}^{n} a_{ij}u_j = \lambda u_i \implies (\lambda - a_{ii})u_i &= \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}u_j \\
\implies |\lambda - a_{ii}|\,|u_i| &= \left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}u_j\right| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|u_j| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|u_i| = |u_i| \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \\
\implies |\lambda - a_{ii}| &\leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|.
\end{aligned}$$

Quindi λ appartiene a K_i e dunque appartiene all'unione dei cerchi di Gershgorin di A. □

Teorema 3.4 (**secondo teorema di Gershgorin**). *Supponiamo che l'unione di k cerchi di Gershgorin di $A \in \mathbb{C}^{n\times n}$ sia disgiunta dall'unione degli altri $n-k$ cerchi. Allora k autovalori di A stanno nella prima unione e $n-k$ nella seconda.*

Teorema 3.5 (**terzo teorema di Gershgorin (forte)**). *Supponiamo che $A \in \mathbb{C}^{n\times n}$ sia irriducibile. Allora i punti che stanno sul bordo di quei cerchi di Gershgorin a cui appartengono ma non sul bordo di tutti i cerchi non sono autovalori di A.*

Ad esempio, per una matrice irriducibile $A \in \mathbb{C}^{5\times 5}$ con i cerchi di Gershgorin come in Figura 3.4, l'1 sta sul bordo di quei cerchi a cui appartiene (il solo cerchio grande di destra) ma non sta sul bordo di tutti i cerchi, quindi non può essere autovalore di A. Similmente, lo 0 sta sul bordo di quei cerchi a cui appartiene (i quattro cerchi grandi) ma non sta sul bordo di tutti i cerchi (non sta sul bordo del cerchio piccolo), quindi non può essere autovalore di A; si veda anche l'Esempio 3.11.

Il terzo teorema di Gershgorin si usa per escludere alcuni punti dai possibili autovalori di A. Si tratta quindi di un teorema "esclusivo". Di questo teorema esiste una versione più debole ma più semplice.

Teorema 3.6 (**terzo teorema di Gershgorin debole**). *Supponiamo che $A \in \mathbb{C}^{n\times n}$ sia irriducibile e sia $\mathscr{B}$ il bordo dell'unione dei cerchi di Gershgorin. Allora i punti di $\mathscr{B}$ che non stanno sul bordo di tutti i cerchi non sono autovalori di A.*

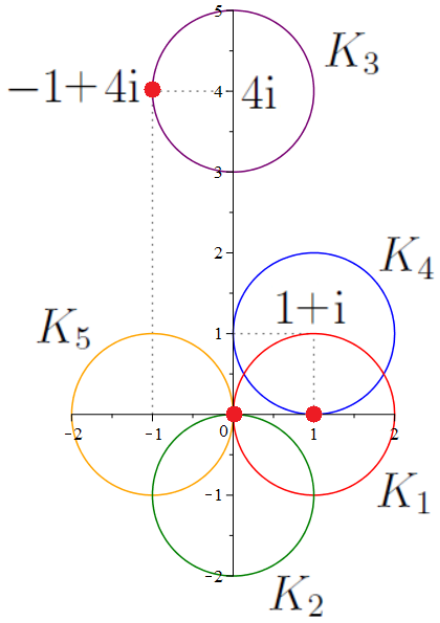

Figura 3.2: Cerchi di Gershgorin della matrice A dell'Esempio 3.9.

Dimostrazione. Ogni punto di $\mathscr{B}$ sta per forza sul bordo di quei cerchi a cui appartiene (non può star dentro a un cerchio altrimenti non sarebbe un punto di $\mathscr{B}$). Pertanto, ogni punto di $\mathscr{B}$ che non sta sul bordo di tutti i cerchi soddisfa le ipotesi del terzo teorema di Gershgorin e quindi va escluso dall'insieme dei possibili autovalori di A. □

Esempio 3.9. Supponiamo di sapere che una matrice $A \in \mathbb{C}^{5\times 5}$ è *irriducibile* e che i suoi cerchi di Gershgorin sono i seguenti:

$$K_1 = \mathscr{C}(1,1), \qquad K_2 = \mathscr{C}(-\mathrm{i},1), \qquad K_3 = \mathscr{C}(4\mathrm{i},1), \qquad K_4 = \mathscr{C}(1+\mathrm{i},1), \qquad K_5 = \mathscr{C}(-1,1).$$

Avendo a disposizione solo queste informazioni, dire se le seguenti affermazioni sono vere o false e motivare la risposta.

1. 0 non è un autovalore di A.
2. A è invertibile.
3. 1 potrebbe essere un autovalore di A.
4. $-1+4\mathrm{i}$ potrebbe essere un autovalore di A.
5. K_3 privato del suo bordo contiene esattamente un autovalore di A.
6. Per il raggio spettrale $\rho(A)$ vale la stima $3 < \rho(A) < 5$.

Soluzione. I cerchi di Gershgorin sono mostrati in Figura 3.2.

1. VERA.
   Infatti, A è irriducibile e 0 sta sul bordo dell'unione dei cerchi di Gershgorin ma non sul bordo di tutti i cerchi, per cui non può essere autovalore di A in base al terzo teorema di Gershgorin debole.
2. VERA.
   Infatti, 0 non è un autovalore di A e quindi A è invertibile.
3. VERA.
   Infatti, non si può escludere che 1 sia un autovalore di A usando il terzo teorema di Gershgorin, perché 1 non soddisfa le ipotesi di tale teorema: 1 appartiene a K_1 ed è interno a K_1, dunque non sta sul bordo di quei cerchi di Gershgorin a cui appartiene.

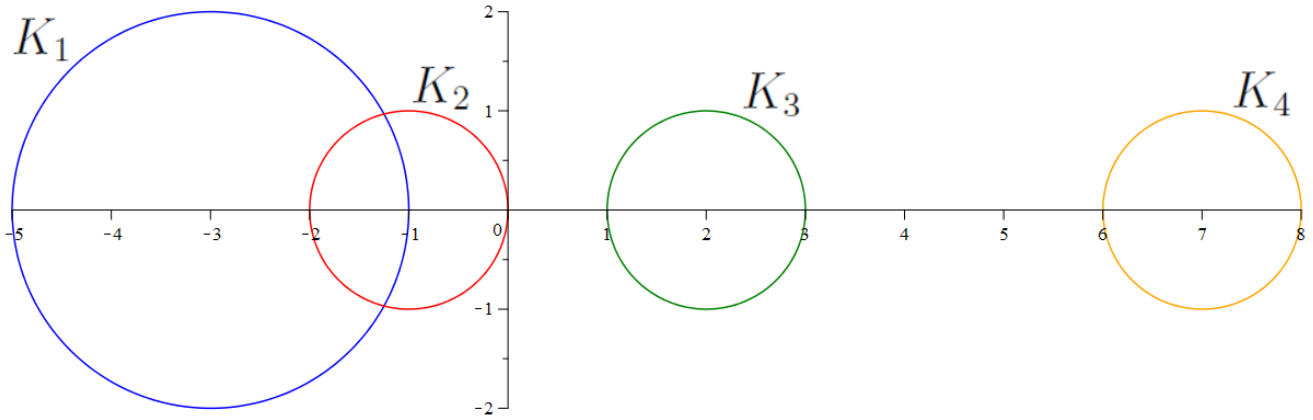

Figura 3.3: Cerchi di Gershgorin della matrice A dell'Esempio 3.10.

4. FALSA.
   Infatti, A è irriducibile e $-1+4\mathrm{i}$ sta sul bordo dell'unione dei cerchi di Gershgorin ma non sul bordo di tutti i cerchi, per cui non può essere autovalore di A in base al terzo teorema di Gershgorin debole.
5. VERA.
   Infatti, K_3 contiene esattamente un autovalore di A per il secondo teorema di Gershgorin. Inoltre, tale autovalore non può stare sul bordo di K_3 perché, per lo stesso motivo visto nel punto precedente, ogni punto del bordo di K_3 non può essere autovalore di A. Dunque K_3 privato del suo bordo contiene esattamente un autovalore di A.
6. VERA.
   Infatti, l'autovalore λ di modulo massimo di A è quello che sta in K_3 privato del suo bordo. Questo perché $|\lambda| = \text{distanza}(\lambda, 0) > 3$ (tutti i punti di K_3 privato del bordo hanno modulo > 3 essendo esterni a $\mathscr{C}(0,3)$), mentre gli altri autovalori hanno modulo < 3 (essendo contenuti in $\mathscr{C}(0,3)$ privato del bordo). Inoltre, $|\lambda| < 5$ perché tutti i punti interni a K_3 hanno questa proprietà essendo contenuti in $\mathscr{C}(0,5)$ privato del bordo.

Esempio 3.10. Supponiamo di sapere che una matrice $A \in \mathbb{R}^{4\times 4}$ (quindi a componenti *reali*) è *irriducibile* e che i suoi cerchi di Gershgorin sono i seguenti:

$$K_1 = \mathscr{C}(-3,2), \qquad K_2 = \mathscr{C}(-1,1), \qquad K_3 = \mathscr{C}(2,1), \qquad K_4 = \mathscr{C}(7,1).$$

Avendo a disposizione solo queste informazioni, dire se le seguenti affermazioni sono vere o false e motivare la risposta.

1. 0 non è autovalore di A e dunque A è invertibile.
2. Un autovalore di A potrebbe trovarsi sul bordo di K_3.
3. Un autovalore di A potrebbe trovarsi sul bordo di K_2.
4. Due autovalori di A si trovano nell'unione $K_1 \cup K_2$ privata del bordo, un autovalore di A si trova nell'intervallo $(1,3)$ e un autovalore di A si trova nell'intervallo $(6,8)$.
5. Per il raggio spettrale $\rho(A)$ vale la stima $6 < \rho(A) < 8$ e inoltre esiste un autovalore di A esattamente uguale al raggio spettrale.

Soluzione. I cerchi di Gershgorin sono mostrati in Figura 3.3.

1. VERA.
   Infatti, A è irriducibile e 0 sta sul bordo dell'unione dei cerchi di Gershgorin ma non sul bordo di tutti i cerchi, per cui non può essere autovalore di A in base al terzo teorema di Gershgorin debole.

2. FALSA.
   Infatti, A è irriducibile e ogni punto del bordo di K_3 sta sul bordo dell'unione dei cerchi di Gershgorin ma non sul bordo di tutti i cerchi, per cui non può essere autovalore di A in base al terzo teorema di Gershgorin debole.
3. VERA.
   Infatti, un autovalore di A potrebbe trovarsi sulla parte del bordo di K_2 interna a K_1. Non potrebbe invece trovarsi sulla parte del bordo di K_2 che non è interna a K_1 perché ciò è escluso dal terzo teorema di Gershgorin debole.
4. VERA.
   Infatti: (a) per il secondo teorema di Gershgorin applicato prima alle unioni $\{K_1 \cup K_2 \cup K_3,\ K_4\}$ e poi alle unioni $\{K_1 \cup K_2 \cup K_4,\ K_3\}$, due autovalori di A stanno in $K_1 \cup K_2$, un autovalore di A sta in K_3 e un autovalore di A sta in K_4; (b) nessun punto sul bordo dell'unione dei cerchi sta sul bordo di tutti i cerchi e dunque nessun punto sul bordo dell'unione dei cerchi può essere un autovalore di A essendo A irriducibile (terzo teorema di Gershgorin debole); (c) se λ è un autovalore di A allora anche $\overline{\lambda}$ è un autovalore di A, perché il polinomio caratteristico di A è a coefficienti reali (essendo A a coefficienti reali) e dunque le sue radici compaiono in coppie complesse coniugate; [10] (d) l'autovalore λ di A che si trova in K_3 è per forza reale (e dunque sta nell'intervallo $(1, 3)$); infatti, se non fosse reale, allora il suo complesso coniugato $\overline{\lambda}$ (= il suo simmetrico rispetto all'asse reale) sarebbe un altro autovalore di A contenuto in K_3, il che è impossibile perché K_3 contiene un unico autovalore di A; (e) l'autovalore λ di A che si trova in K_4 è per forza reale (e dunque sta nell'intervallo $(6, 8)$) per lo stesso motivo esposto al precedente punto (d).
5. VERA.
   Infatti l'autovalore che sta in K_4 (o meglio nell'intervallo $(6, 8)$) è quello di modulo massimo ed essendo positivo coincide con il raggio spettrale. In particolare, vale $6 < \rho(A) < 8$.

Esempio 3.11 (**esempio del quadrifoglio**)**.** Supponiamo che una matrice $A \in \mathbb{C}^{5\times 5}$ sia *irriducibile* e che i suoi cerchi di Gershgorin siano i seguenti:

$$K_1 = \mathscr{C}(1, 1), \qquad K_2 = \mathscr{C}(\mathrm{i}, 1), \qquad K_3 = \mathscr{C}(-1, 1), \qquad K_4 = \mathscr{C}(-\mathrm{i}, 1), \qquad K_5 = \mathscr{C}(1, 1/4).$$

Dimostrare che A è invertibile.

Soluzione. Si tratta di dimostrare che 0 non è un autovalore di A. Osservando la Figura 3.4, si vede che 0 sta sul bordo di quei cerchi di Gershgorin a cui esso appartiene, infatti appartiene a K_1, K_2, K_3, K_4 e sta sul bordo di essi. Tuttavia, 0 non sta sul bordo di tutti i cerchi di Gershgorin perché non sta sul bordo di K_5. Dunque 0 non può essere autovalore di A per il terzo teorema di Gershgorin.

Osservazione. A questa conclusione non si poteva giungere utilizzando il terzo teorema di Gershgorin debole, perché 0 non sta sul bordo dell'unione dei cerchi di Gershgorin bensì all'interno dell'unione.

Osservazione. Se avessimo avuto una matrice $A \in \mathbb{C}^{4\times 4}$ *irriducibile* con i cerchi di Gershgorin K_1, K_2, K_3, K_4 come prima, allora *non* avremmo potuto escludere 0 dai possibili autovalori di A (e quindi non avremmo potuto dire che A è invertibile) nemmeno usando il terzo teorema di Gershgorin (forte). Questo perché 0 non soddisfa le ipotesi di tale teorema in quanto 0 sta sul bordo di tutti i cerchi di Gershgorin (essendo venuto a mancare il cerchio K_5 di prima).

[10] Infatti, se $p(x) = a_0 + a_1 x + \ldots + a_n x^n$ è un polinomio a coefficienti reali e λ è una sua radice allora

$$0 = p(\lambda) = a_0 + a_1\lambda + \ldots + a_n\lambda^n \implies 0 = \overline{p(\lambda)} = \overline{a_0 + a_1\lambda + \ldots + a_n\lambda^n} = a_0 + a_1\overline{\lambda} + \ldots + a_n\overline{\lambda}^n = p(\overline{\lambda}).$$

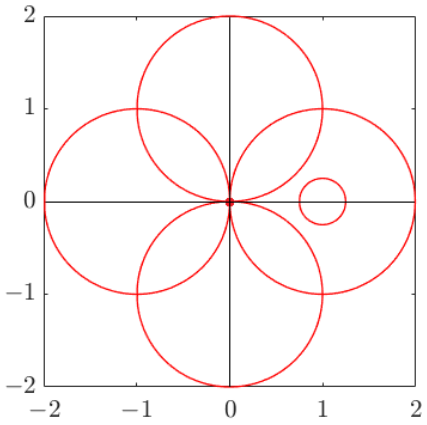

Figura 3.4: Cerchi di Gershgorin della matrice A dell'Esempio 3.11.

Osservazione 3.1. Gli autovalori di una matrice $A \in \mathbb{C}^{n\times n}$ e della sua trasposta A^T coincidono perché i polinomi caratteristici di A e A^T coincidono:

$$C_{A^T}(\lambda) = \det(\lambda I - A^T) = \det((\lambda I - A)^T) = \det(\lambda I - A) = C_A(\lambda).$$

Di conseguenza, possiamo applicare i teoremi di Gershgorin non solo ad A ma anche ad A^T per ottenere localizzazioni migliori degli autovalori di A. In particolare, il primo teorema di Gershgorin applicato ad A e A^T ci dice la cosa seguente.

> *Gli autovalori di una matrice $A \in \mathbb{C}^{n\times n}$ stanno tutti sia nell'unione dei cerchi di Gershgorin $K_1, \ldots, K_n$ di A sia nell'unione dei cerchi di Gershgorin $H_1, \ldots, H_n$ di A^T, per cui stanno nell'intersezione delle due unioni $(K_1 \cup \cdots \cup K_n) \cap (H_1 \cup \cdots \cup H_n)$.*

Notiamo che i cerchi di Gershgorin $H_1, \ldots, H_n$ di A^T sono semplicemente i cerchi di Gershgorin per colonna di A, in quanto le righe di A^T sono le colonne di A. Pertanto, il risultato precedente può anche essere enunciato nel modo seguente.

> *Gli autovalori di una matrice $A \in \mathbb{C}^{n\times n}$ stanno tutti sia nell'unione dei cerchi di Gershgorin per riga $K_1, \ldots, K_n$ di A sia nell'unione dei cerchi di Gershgorin per colonna $H_1, \ldots, H_n$ di A, per cui stanno nell'intersezione delle due unioni $(K_1 \cup \cdots \cup K_n) \cap (H_1 \cup \cdots \cup H_n)$.*

Osserviamo inoltre, in vista dell'applicazione del terzo teorema di Gershgorin, che una matrice A è irriducibile se e solo se la sua trasposta A^T è irriducibile (Esercizio 3.6).

Esercizio 3.6. Sia $A \in \mathbb{C}^{n\times n}$. Dimostrare che A è irriducibile se e solo se A^T è irriducibile.

Suggerimento. Se A è irriducibile allora esiste un ciclo nel grafo di A che tocca tutti i nodi. Dimostrare che nel grafo di A^T vi è lo stesso ciclo percorso al contrario. Ad esempio, se $n = 4$ e nel grafo di A c'è il ciclo $1 \to 2 \to 4 \to 3 \to 4 \to 1$ allora nel grafo di A^T c'è il ciclo $1 \to 4 \to 3 \to 4 \to 2 \to 1$ (vedere anche l'Esempio 3.12).

Esempio 3.12. Si consideri la matrice

$$A = \begin{bmatrix} 7 & 2 & 2 & 2 \\ 0 & -1 & 0 & 1/2 \\ 0 & 0 & 1 & \mathrm{i}/2 \\ 1 & 0 & -1 & -\mathrm{i} \end{bmatrix}.$$

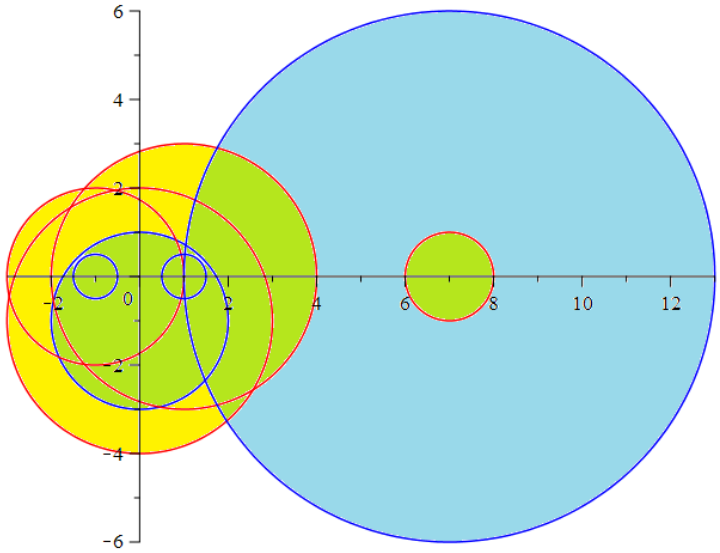

Figura 3.5: Cerchi di Gershgorin per riga K_1, K_2, K_3, K_4 (cerchi azzurri delimitati dalle circonferenze blu) e per colonna H_1, H_2, H_3, H_4 (cerchi gialli delimitati dalle circonferenze rosse) della matrice A dell'Esempio 3.12. L'intersezione $(K_1 \cup K_2 \cup K_3 \cup K_4) \cap (H_1 \cup H_2 \cup H_3 \cup H_4)$ è rappresentata in verde.

(a) Usando i teoremi di Gershgorin (del terzo solo la versione debole), localizzare lo spettro di A nel modo più preciso possibile.
(b) Sulla base delle informazioni spettrali ottenute, dimostrare che la matrice $\alpha I + A$ è invertibile per ogni $\alpha \geq 2$.

Soluzione. (a) Dovendo localizzare lo spettro di A nel modo più preciso possibile, consideriamo sia i cerchi di Gershgorin per riga K_1, K_2, K_3, K_4 che i cerchi di Gershgorin per colonna H_1, H_2, H_3, H_4 della matrice A, che sono dati da

$$\begin{aligned} K_1 &= \mathscr{C}(7, 6) & H_1 &= \mathscr{C}(7, 1) \\ K_2 &= \mathscr{C}(-1, 1/2) & H_2 &= \mathscr{C}(-1, 2) \\ K_3 &= \mathscr{C}(1, 1/2) & H_3 &= \mathscr{C}(1, 3) \\ K_4 &= \mathscr{C}(-\mathrm{i}, 2) & H_4 &= \mathscr{C}(-\mathrm{i}, 3) \end{aligned}$$

Applicazione 1° teorema di Gershgorin (per righe e colonne cioè per A e A^T). Dal primo teorema di Gershgorin (insieme con l'Osservazione 3.1) sappiamo che gli autovalori di A stanno nell'insieme

$$(K_1 \cup K_2 \cup K_3 \cup K_4) \cap (H_1 \cup H_2 \cup H_3 \cup H_4)$$

rappresentato in Figura 3.5.
Applicazione 2° teorema di Gershgorin (per righe e colonne cioè per A e A^T). Osserviamo ora che il cerchio H_1 è disgiunto da $H_2 \cup H_3 \cup H_4$. Quindi, in base al secondo teorema di Gershgorin (applicato ad A^T), un autovalore di A si trova in H_1 e gli altri tre autovalori di A si trovano in $H_2 \cup H_3 \cup H_4$ (più precisamente, si trovano nell'intersezione tra $H_2 \cup H_3 \cup H_4$ e $K_1 \cup K_2 \cup K_3 \cup K_4$).
Applicazione 3° teorema di Gershgorin debole (per righe e colonne cioè per A e A^T). Notiamo ora che il grafo di A contiene il ciclo

$$1 \to 2 \to 4 \to 3 \to 4 \to 1$$

che tocca tutti i nodi. Dunque il grafo di A è fortemente connesso e A è irriducibile. Di conseguenza, anche A^T è irriducibile per l'Esercizio 3.6 (si può verificare che il grafo di A^T contiene il ciclo precedente

percorso al contrario, cioè $1 \to 4 \to 3 \to 4 \to 2 \to 1$). Questo ci permette di applicare il terzo teorema di Gershgorin debole e raffinare la localizzazione degli autovalori: siccome nessun punto del bordo di $K_1 \cup K_2 \cup K_3 \cup K_4$ sta sul bordo di tutti i singoli cerchi K_1, K_2, K_3, K_4 (infatti $\partial K_1 \cap \partial K_2 \cap \partial K_3 \cap \partial K_4 = \emptyset$), concludiamo che nessun punto del bordo di $K_1 \cup K_2 \cup K_3 \cup K_4$ può essere un autovalore di A; analogamente, siccome $\partial H_1 \cap \partial H_2 \cap \partial H_3 \cap \partial H_4 = \emptyset$, nessun punto del bordo di $H_1 \cup H_2 \cup H_3 \cup H_4$ può essere un autovalore di A. Dunque, osservando anche la Figura 3.5, concludiamo che tutti gli autovalori di A stanno nella parte interna dell'intersezione $(K_1 \cup K_2 \cup K_3 \cup K_4) \cap (H_1 \cup H_2 \cup H_3 \cup H_4)$.

(b) Gli autovalori di $\alpha I + A$ sono dati da $\alpha + \lambda_1$, $\alpha + \lambda_2$, $\alpha + \lambda_3$, $\alpha + \lambda_4$ dove λ_1, λ_2, λ_3, λ_4 sono gli autovalori di A (Teorema 3.2). Dal punto (a) sappiamo che gli autovalori di A stanno in

$$\text{parte interna di } (K_1 \cup K_2 \cup K_3 \cup K_4) \cap (H_1 \cup H_2 \cup H_3 \cup H_4)$$

e quindi la parte reale di ogni autovalore di A è strettamente maggiore di -2. Pertanto,

$$\mathrm{Re}(\alpha + \lambda_i) = \alpha + \mathrm{Re}(\lambda_i) > 0$$

per ogni $i = 1, 2, 3, 4$ (essendo $\alpha \geq 2$ per ipotesi) e dunque in particolare $\alpha + \lambda_i \neq 0$ per ogni $i = 1, 2, 3, 4$. In conclusione, $\alpha I + A$ è invertibile perché non ha 0 come autovalore.

Esercizio 3.7. Si consideri la matrice

$$A = \begin{bmatrix} 2 & \mathrm{i} & 0 & 0 & 0 \\ 0 & 2 & \mathrm{i} & 0 & 0 \\ 0 & 0 & 2 & \mathrm{i} & 0 \\ 0 & 0 & 0 & 2 & \mathrm{i} \\ \frac{1}{10} & \frac{1}{2} & \frac{1}{2} & 1 & 2 \end{bmatrix}.$$

(a) Usando i teoremi di Gershgorin (del terzo solo la versione debole), localizzare lo spettro di A nel modo più preciso possibile.
(b) Sulla base delle informazioni spettrali ottenute, è possibile dire se A è invertibile oppure no?
(c) Sulla base delle informazioni spettrali ottenute, fornire un limite superiore (il più preciso possibile) per il raggio spettrale $\rho(A)$.

Esercizio 3.8. Sia data la matrice $A \in \mathbb{C}^{n \times n}$ di elementi $a_{jj} = 3$ per $j = 1, \ldots, n$, $a_{jk} = -1$ per $j, k \geq 2$ con $|j - k| = 1$, $a_{1k} = a_{k1} = 1/2^k$ per $k = 2, \ldots, n$, e $a_{jk} = 0$ per tutte le altre coppie di indici (j, k).

(a) Usando i teoremi di Gershgorin (del terzo solo la versione debole), localizzare lo spettro di A nel modo più preciso possibile.
(b) Sulla base delle informazioni spettrali ottenute, dimostrare che la matrice A è invertibile. È vero che A è definita positiva?

3.6 Matrici a diagonale dominante e a diagonale dominante in senso stretto

Sia $A \in \mathbb{C}^{n \times n}$ una matrice.

- Si dice che A è a diagonale dominante (per righe) se
  - $|a_{ii}| \geq \sum_{j=1,\, j \neq i}^{n} |a_{ij}|$ per ogni $i = 1, \ldots, n$; [11]
  - esiste almeno un indice $k \in \{1, \ldots, n\}$ per il quale vale la disuguaglianza stretta $|a_{kk}| > \sum_{j=1,\, j \neq k}^{n} |a_{kj}|$. [12]

[11] Questa condizione si esprime anche dicendo che per ogni cerchio di Gershgorin di A la distanza del centro dallo 0 è maggiore o uguale al raggio (quindi nessun cerchio di Gershgorin di A può contenere lo 0 al suo interno).

[12] Questa condizione si esprime anche dicendo che deve esistere almeno un cerchio di Gershgorin di A che non contiene lo 0 (cioè tale che la distanza del centro dallo 0 è maggiore del raggio).

- Si dice che A è a diagonale dominante in senso stretto (per righe) se $|a_{ii}| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$ per ogni $i = 1, \ldots, n$.[13]
- Si dice che A è a diagonale dominante per colonne se
  - $|a_{jj}| \geq \sum_{i=1,\, i\neq j}^{n} |a_{ij}|$ per ogni $j = 1, \ldots, n$;
  - esiste almeno un indice $\ell \in \{1, \ldots, n\}$ per il quale vale la disuguaglianza stretta $|a_{\ell\ell}| > \sum_{i=1,\, i\neq \ell}^{n} |a_{i\ell}|$.
- Si dice che A è a diagonale dominante in senso stretto per colonne se $|a_{jj}| > \sum_{i=1,\, i\neq j}^{n} |a_{ij}|$ per ogni $i = 1, \ldots, n$.

Analogamente a quanto avviene per i cerchi di Gershgorin, *quando si parla di dominanza diagonale senza altre specificazioni, s'intende per righe.* Quando si vuole parlare di dominanza diagonale per colonne, questo va specificato ogni volta.

Esempio 3.13. Discutere le proprietà di dominanza diagonale per righe e per colonne della matrice

$$A = \begin{bmatrix} 3 & 1 & -2 \\ 0 & -3 & 1 \\ 1-2\mathrm{i} & -1 & 5 \end{bmatrix}.$$

Soluzione. Poiché

$$\begin{aligned} |3| = |1| + |-2| \qquad &\Longleftrightarrow \qquad 3 = 3 \\ |-3| > |0| + |1| \qquad &\Longleftrightarrow \qquad 3 > 1 \\ |5| > |1-2\mathrm{i}| + |-1| \qquad &\Longleftrightarrow \qquad 5 > \sqrt{5} + 1 \end{aligned}$$

si conclude che A è a diagonale dominante (per righe) ma non a diagonale dominante in senso stretto (per righe). Poiché

$$\begin{aligned} |3| > |1-2\mathrm{i}| \qquad &\Longleftrightarrow \qquad 3 > \sqrt{5} \\ |-3| > |1| + |-1| \qquad &\Longleftrightarrow \qquad 3 > 2 \\ |5| > |-2| + |1| \qquad &\Longleftrightarrow \qquad 5 > 3 \end{aligned}$$

si conclude che A è a diagonale dominante in senso stretto per colonne (e quindi a maggior ragione anche a diagonale dominante per colonne).

Teorema 3.7. *Supponiamo che la matrice $A \in \mathbb{C}^{n\times n}$ soddisfi almeno una delle seguenti condizioni:*
- *A è a diagonale dominante e irriducibile;*
- *A è a diagonale dominante in senso stretto;*
- *A è a diagonale dominante per colonne e irriducibile;*
- *A è a diagonale dominante in senso stretto per colonne.*

Allora A è invertibile.

Dimostrazione. La dimostrazione si basa sui teoremi di Gershgorin. Dimostriamo il teorema sotto l'ipotesi che A sia a diagonale dominante e irriducibile (la dimostrazione del teorema sotto le altre tre ipotesi è simile ed è lasciata come Esercizio 3.9). Mostriamo che 0 non è un autovalore di A usando il terzo teorema di Gershgorin. Poiché A è a diagonale dominante, se 0 appartiene a un cerchio di Gershgorin K_i allora deve stare per forza sul bordo di K_i. Infatti non può stare all'interno, perché per l'ipotesi di dominanza diagonale si ha

$$\text{raggio di } K_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \leq |a_{ii}| = |a_{ii}| = \text{distanza}(a_{ii}, 0) = \text{distanza}(\text{centro di } K_i,\, 0).$$

[13] Questa condizione si esprime anche dicendo che nessuno dei cerchi di Gershgorin di A contiene lo 0.

Dunque 0 sta per forza sul bordo di quei cerchi di Gershgorin a cui esso appartiene (non può stare all'interno). Inoltre, sempre per l'ipotesi di dominanza diagonale, esiste un indice k tale che

$$|a_{kk}| > \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|.$$

Questo significa che 0 non sta sul bordo di K_k (non ci sta proprio in K_k!) e dunque 0 non sta sul bordo di tutti i cerchi di Gershgorin di A. Poiché A è irriducibile per ipotesi, il terzo teorema di Gershgorin ci dice che 0 non può essere un autovalore di A e quindi A è invertibile. □

Osservazione 3.2. Nella dimostrazione del Teorema 3.7 abbiamo dovuto usare la versione forte del terzo teorema di Gershgorin perché quella debole non basta. Infatti, la matrice

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & \mathrm{i} & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -\mathrm{i} & 1 \\ 1/4 & 0 & 0 & 0 & 1 \end{bmatrix}$$

è a diagonale dominante e irriducibile (lo si verifichi per esercizio), e ha i cerchi di Gershgorin (sia per righe che per colonne) mostrati in Figura 3.4, per cui non riusciremmo a dimostrare che è invertibile (cioè che 0 non è un autovalore) usando la sola versione debole del terzo teorema di Gershgorin; si veda anche l'osservazione in fondo alla soluzione dell'Esempio 3.11.

Esercizio 3.9. Dimostrare il Teorema 3.7 sotto le altre tre ipotesi non considerate nella dimostrazione del teorema stesso.

3.7 Norme vettoriali

3.7.1 Il concetto di norma vettoriale

Consideriamo il sistema lineare

$$\begin{bmatrix} 8 & 1 & 1 \\ 1 & 5 & -1 \\ 1 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 26 \\ 7 \\ 7 \end{bmatrix}$$

la cui soluzione è $\mathbf{x} = [3, 1, 1]^T$. Si supponga di aver ottenuto le seguenti approssimazioni della soluzione:

$$\mathbf{y} = [2.99972,\ 1.00023,\ 1.00030]^T, \tag{3.6}$$

$$\mathbf{z} = [3.00027,\ 0.99971,\ 0.99955]^T. \tag{3.7}$$

Come possiamo stabilire quale delle due è più vicina alla soluzione $\mathbf{x}$? Occorre introdurre un concetto di distanza sullo spazio dei vettori e misurare la distanza di $\mathbf{y}$ e $\mathbf{z}$ da $\mathbf{x}$: la soluzione approssimata che dista di meno è quella più vicina. Un ottimo concetto di distanza in uno spazio di vettori è il concetto di norma vettoriale.

Definizione 3.1. Una funzione $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ si dice norma vettoriale se soddisfa le seguenti proprietà:

(a) $\|\mathbf{x}\| \geq 0$ per ogni $\mathbf{x} \in \mathbb{C}^n$ e $\|\mathbf{x}\| = 0$ se e solo se $\mathbf{x} = \mathbf{0}$ [positività];

(b) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ per ogni $\alpha \in \mathbb{C}$ e ogni $\mathbf{x} \in \mathbb{C}^n$ [omogeneità];

(c) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ [disuguaglianza triangolare].

Data una norma vettoriale $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$, definiamo la distanza fra due vettori $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ come $\|\mathbf{x} - \mathbf{y}\|$.

3.7.2 Le norme 1, 2, ∞

Le norme vettoriali più importanti in $\mathbb{C}^n$ sono tre: la norma 1, la norma 2 (o euclidea) e la norma ∞. Esse sono definite nel modo seguente:

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \ldots + |x_n|, \tag{3.8}$$
$$\|\mathbf{x}\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2}, \tag{3.9}$$
$$\|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \ldots, |x_n|). \tag{3.10}$$

Le relative distanze sono definite nel modo seguente:

$$\|\mathbf{x} - \mathbf{y}\|_1 = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|,$$
$$\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + \ldots + |x_n - y_n|^2},$$
$$\|\mathbf{x} - \mathbf{y}\|_\infty = \max(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|).$$

Tornando all'esempio introduttivo, se calcoliamo la distanza dei vettori $\mathbf{y}$ e $\mathbf{z}$ in (3.6)–(3.7) dal vettore soluzione $\mathbf{x} = [3, 1, 1]^T$ usando la $\|\cdot\|_\infty$, otteniamo

$$\mathbf{x} - \mathbf{y} = [0.00028,\ -0.00023,\ -0.00030]^T \qquad \Longrightarrow \qquad \|\mathbf{x} - \mathbf{y}\|_\infty = 0.00030,$$
$$\mathbf{x} - \mathbf{z} = [-0.00027,\ 0.00029,\ 0.00045]^T \qquad \Longrightarrow \qquad \|\mathbf{x} - \mathbf{z}\|_\infty = 0.00045.$$

Quindi rispetto alla $\|\cdot\|_\infty$ il vettore $\mathbf{y}$ è più vicino a $\mathbf{x}$ rispetto al vettore $\mathbf{z}$.

3.7.3 Equivalenza delle norme vettoriali

Teorema 3.8. *Tutte le norme vettoriali in $\mathbb{C}^n$ sono equivalenti, nel senso che se prendiamo due norme vettoriali $\|\cdot\|', \|\cdot\|'' : \mathbb{C}^n \to \mathbb{R}$ allora si ha*

$$\alpha\|\mathbf{x}\|'' \le \|\mathbf{x}\|' \le \beta\|\mathbf{x}\|'' \ \textit{per ogni}\ \mathbf{x} \in \mathbb{C}^n, \tag{3.11}$$

dove $\alpha, \beta > 0$ sono due costanti indipendenti da $\mathbf{x}$.

Verifichiamo ad esempio che $\|\cdot\|_1$ e $\|\cdot\|_\infty$ sono equivalenti. Per ogni $\mathbf{x} \in \mathbb{C}^n$ si ha

$$\max(|x_1|, \ldots, |x_n|) \le |x_1| + \ldots + |x_n| \le n \max(|x_1|, \ldots, |x_n|)$$
$$\Longrightarrow\ \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_1 \le n\|\mathbf{x}\|_\infty\ \Longrightarrow\ \frac{1}{n}\|\mathbf{x}\|_1 \le \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_1.$$

Dunque $\|\cdot\|_1$ e $\|\cdot\|_\infty$ sono equivalenti essendo la (3.11) soddisfatta con $\alpha = 1$ e $\beta = n$ (se si considera $\|\cdot\|' = \|\cdot\|_1$ e $\|\cdot\|'' = \|\cdot\|_\infty$) oppure con $\alpha = \frac{1}{n}$ e $\beta = 1$ (se si considera $\|\cdot\|' = \|\cdot\|_\infty$ e $\|\cdot\|'' = \|\cdot\|_1$). In ogni caso, quello che conta è che α e β sono costanti positive indipendenti da $\mathbf{x}$. Analogamente, possiamo verificare che $\|\cdot\|_2$ e $\|\cdot\|_\infty$ sono equivalenti: per ogni $\mathbf{x} \in \mathbb{C}^n$ si ha

$$\max(|x_1|, \ldots, |x_n|) \le \sqrt{|x_1|^2 + \ldots + |x_n|^2} \le \sqrt{n \max(|x_1|, \ldots, |x_n|)^2} = \sqrt{n} \max(|x_1|, \ldots, |x_n|)$$
$$\Longrightarrow\ \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty\ \Longrightarrow\ \frac{1}{\sqrt{n}}\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2.$$

3.7.4 Successioni di vettori

Una successione di vettori $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... in $\mathbb{C}^n$ si dice convergente al vettore $\mathbf{x} \in \mathbb{C}^n$ rispetto alla norma vettoriale $\|\cdot\|$ se $\|\mathbf{x}^{(k)} - \mathbf{x}\| \to 0$. Poiché tutte le norme vettoriali sono equivalenti per il Teorema 3.8, se una successione di vettori converge a $\mathbf{x}$ rispetto a una norma $\|\cdot\|$ allora converge a $\mathbf{x}$ rispetto a tutte le norme. Infatti, supponiamo che la successione $\{\mathbf{x}^{(k)}\}_{k=0,1,2,\ldots}$ converga a $\mathbf{x}$ rispetto alla norma $\|\cdot\|$ e sia $\|\cdot\|'$ un'altra norma. Poiché $\|\cdot\|$ e $\|\cdot\|'$ sono equivalenti, esistono due costanti $\alpha, \beta > 0$ tali che

$$\alpha\|\mathbf{y}\| \le \|\mathbf{y}\|' \le \beta\|\mathbf{y}\| \text{ per ogni } \mathbf{y} \in \mathbb{C}^n,$$

dunque

$$\alpha\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \|\mathbf{x}^{(k)} - \mathbf{x}\|' \le \beta\|\mathbf{x}^{(k)} - \mathbf{x}\| \text{ per ogni } k = 0, 1, 2, \ldots$$

Siccome $\|\mathbf{x}^{(k)} - \mathbf{x}\| \to 0$ (perché $\mathbf{x}^{(k)} \to \mathbf{x}$ in norma $\|\cdot\|$) concludiamo che $\|\mathbf{x}^{(k)} - \mathbf{x}\|' \to 0$ (cioè $\mathbf{x}^{(k)} \to \mathbf{x}$ in norma $\|\cdot\|'$).

Una successione di vettori $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... in $\mathbb{C}^n$ si dice convergente (componente per componente) al vettore $\mathbf{x} \in \mathbb{C}^n$ se $\mathbf{x}^{(k)} \to \mathbf{x}$ componente per componente, cioè se

$$\begin{cases} x_1^{(k)} \to x_1 \\ x_2^{(k)} \to x_2 \\ \vdots \\ x_n^{(k)} \to x_n \end{cases} \iff \begin{cases} |x_1^{(k)} - x_1| \to 0 \\ |x_2^{(k)} - x_2| \to 0 \\ \vdots \\ |x_n^{(k)} - x_n| \to 0 \end{cases}$$

$$\iff \max(|x_1^{(k)} - x_1|, |x_2^{(k)} - x_2|, \ldots, |x_n^{(k)} - x_n|) \to 0$$

$$\iff \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty \to 0.$$

Vediamo quindi che la convergenza componente per componente altro non è che la convergenza in $\|\cdot\|_\infty$. Pertanto, ricordando l'equivalenza di tutte le norme, dire che $\mathbf{x}^{(k)} \to \mathbf{x}$ componente per componente è lo stesso che dire che $\mathbf{x}^{(k)} \to \mathbf{x}$ in una qualsiasi norma.

3.8 Norme matriciali

3.8.1 Il concetto di norma matriciale

Si vuole introdurre un concetto di distanza sullo spazio delle matrici per misurare la "vicinanza" tra due matrici $A, B \in \mathbb{C}^{n\times n}$. A questo scopo, si può interpretare una matrice $A \in \mathbb{C}^{n\times n}$ come un vettore di n^2 componenti e utilizzare come distanza una delle norme vettoriali già introdotte. Questo procedimento è senz'altro corretto, ma spesso conduce a norme che "non si comportano bene" rispetto al prodotto di matrici e pertanto sono di scarso interesse. Diamo comunque qui la definizione generale di norma matriciale e ci riserviamo di trattare le norme matriciali interessanti nella sezione successiva.

Definizione 3.2. Una funzione $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ si dice norma matriciale se soddisfa le seguenti proprietà:
(a) $\|A\| \ge 0$ per ogni $A \in \mathbb{C}^{n\times n}$ e $\|A\| = 0$ se e solo se $A = O$ [positività];
(b) $\|\alpha A\| = |\alpha|\,\|A\|$ per ogni $\alpha \in \mathbb{C}$ e ogni $A \in \mathbb{C}^{n\times n}$ [omogeneità];
(c) $\|A + B\| \le \|A\| + \|B\|$ per ogni $A, B \in \mathbb{C}^{n\times n}$ [disuguaglianza triangolare].
Data una norma matriciale $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$, definiamo la distanza fra due matrici $A, B \in \mathbb{C}^{n\times n}$ come $\|A - B\|$.

Un esempio di norma matriciale è dato dall'analoga della norma ∞ dei vettori: data $A \in \mathbb{C}^{n\times n}$, s'immagina A come se fosse un vettore di n^2 componenti e si definisce la sua norma come se fosse la norma ∞ del vettore di n^2 componenti:

$$|A|_\infty = \max_{i,j=1,\dots,n} |a_{ij}|.$$

In modo analogo si possono definire le norme $|A|_1$ e $|A|_2$. Purtroppo, la norma $|\cdot|_\infty$ "non si comporta bene" rispetto al prodotto di matrici perché non è submoltiplicativa: date due matrici $A, B \in \mathbb{C}^{n\times n}$, non è detto che risulti $|AB|_\infty \le |A|_\infty |B|_\infty$. Ecco un esempio:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad AB = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

$$|A|_\infty = 1, \qquad |B|_\infty = 1, \qquad |AB|_\infty = 2.$$

3.8.2 Norme matriciali indotte

Vogliamo introdurre una classe di norme matriciali interessanti e ricavarne alcune proprietà.

Definizione 3.3. Data una norma vettoriale $\|\cdot\|$ in $\mathbb{C}^n$ e una matrice $A \in \mathbb{C}^{n\times n}$, definiamo il numero

$$\|A\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{x}\neq\mathbf{0}} \left\| A \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\| = \max_{\|\mathbf{y}\|=1} \|A\mathbf{y}\|.$$

Si può dimostrare che $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ è una norma matriciale che prende il nome di norma matriciale indotta dalla norma vettoriale $\|\cdot\|$.

Si osservi che la norma matriciale indotta dalla norma vettoriale $\|\cdot\|$ viene volutamente denotata con lo stesso simbolo $\|\cdot\|$. Il seguente teorema mette in luce interessanti proprietà delle norme matriciali indotte.

Teorema 3.9. *Sia $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ una norma matriciale indotta dalla norma vettoriale $\|\cdot\|$ e siano $A, B \in \mathbb{C}^{n\times n}$. Allora valgono le seguenti proprietà.*
1. $\|I\| = 1$.
2. $\|A\mathbf{x}\| \le \|A\|\,\|\mathbf{x}\|$ *per ogni* $\mathbf{x} \in \mathbb{C}^n$.
3. $\|A\|$ *è la più piccola costante* C *che soddisfa* $\|A\mathbf{x}\| \le C\|\mathbf{x}\|$ *per ogni* $\mathbf{x} \in \mathbb{C}^n$.
4. $\|AB\| \le \|A\|\,\|B\|$ [submoltiplicatività].
5. $\rho(A) \le \|A\|$.

Dimostrazione. 1. Risulta

$$\|I\| = \max_{\|\mathbf{x}\|=1} \|I\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{x}\| = 1.$$

2. Per ogni $\mathbf{x} \neq \mathbf{0}$ si ha

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \le \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|A\mathbf{y}\|}{\|\mathbf{y}\|} = \|A\| \quad \Longrightarrow \quad \|A\mathbf{x}\| \le \|A\|\,\|\mathbf{x}\|.$$

La disuguaglianza precedente vale ovviamente anche per $\mathbf{x} = \mathbf{0}$.

3. Presa una qualsiasi costante C che soddisfa $\|A\mathbf{x}\| \le C\|\mathbf{x}\|$ per ogni $\mathbf{x} \in \mathbb{C}^n$, si ha

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \le C \text{ per ogni } \mathbf{x} \neq \mathbf{0} \quad \Longrightarrow \quad \|A\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \le C.$$

4. Per ogni $\mathbf{x} \in \mathbb{C}^n$ si ha

$$\|AB\mathbf{x}\| \leq \|A\| \, \|B\mathbf{x}\| \leq \|A\| \, \|B\| \, \|\mathbf{x}\|.$$

Concludiamo che $\|AB\| \leq \|A\| \, \|B\|$ in quanto $\|AB\|$ è la più piccola costante C che soddisfa $\|AB\mathbf{x}\| \leq C\|\mathbf{x}\|$ per ogni $\mathbf{x} \in \mathbb{C}^n$.

5. Sia λ un autovalore di A di modulo massimo e sia $\mathbf{x} \neq \mathbf{0}$ un corrispondente autovettore. Dall'equazione $A\mathbf{x} = \lambda\mathbf{x}$ otteniamo

$$\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda| \, \|\mathbf{x}\| = \rho(A)\|\mathbf{x}\| \quad \Longrightarrow \quad \rho(A) = \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \leq \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|A\mathbf{y}\|}{\|\mathbf{y}\|} = \|A\|. \qquad \square$$

3.8.3 Le norme 1, 2, ∞

Le norme matriciali più importanti per i nostri scopi sono tre: la norma 1, la norma 2 (o euclidea) e la norma ∞. Esse sono semplicemente le norme matriciali indotte dalle norme vettoriali 1, 2 e ∞:

$$\|A\|_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1,$$

$$\|A\|_2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2,$$

$$\|A\|_\infty = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty.$$

Il seguente teorema fornisce delle formule per calcolare le norme 1, 2, ∞ delle matrici. In ciò che segue, se $A \in \mathbb{C}^{n\times n}$, denotiamo con $A_{[1]}, A_{[2]}, \ldots, A_{[n]}$ le righe di A e con $A^{[1]}, A^{[2]}, \ldots, A^{[n]}$ le colonne di A.

Teorema 3.10. *Per ogni $A \in \mathbb{C}^{n\times n}$ valgono le seguenti formule.*

- $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}| = \max(\|A^{[1]}\|_1, \|A^{[2]}\|_1, \ldots, \|A^{[n]}\|_1)$.
- $\|A\|_2 = \sqrt{\rho(A^*A)}$.
- $\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}| = \max(\|A_{[1]}\|_1, \|A_{[2]}\|_1, \ldots, \|A_{[n]}\|_1)$.

Esempio 3.14. Calcolare le norme 1, 2, ∞ della matrice

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

Soluzione. Cominciamo da $\|A\|_1$ e $\|A\|_\infty$ che sono più facili: si ha immediatamente

$$\begin{aligned} \|A^{[1]}\|_1 &= |2| + |-3| = 5 \\ \|A^{[2]}\|_1 &= |-2| + |-1| = 3 \end{aligned} \quad \Longrightarrow \quad \|A\|_1 = 5$$

e

$$\begin{aligned} \|A_{[1]}\|_1 &= |2| + |-2| = 4 \\ \|A_{[2]}\|_1 &= |-3| + |-1| = 4 \end{aligned} \quad \Longrightarrow \quad \|A\|_\infty = 4.$$

Per calcolare $\|A\|_2$ dobbiamo determinare il raggio spettrale di

$$A^*A = \begin{bmatrix} 2 & -3 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} 2 & -2 \\ -3 & -1 \end{bmatrix} = \begin{bmatrix} 13 & -1 \\ -1 & 5 \end{bmatrix}.$$

Il polinomio caratteristico di A^*A è

$$C_{A^*A}(\lambda) = \det(\lambda I - A^*A) = \begin{vmatrix} \lambda - 13 & 1 \\ 1 & \lambda - 5 \end{vmatrix} = (\lambda - 13)(\lambda - 5) - 1 = \lambda^2 - 18\lambda + 64,$$

per cui gli autovalori di A^*A sono

$$\lambda_{1,2} = \frac{18 \pm \sqrt{18^2 - 4 \cdot 64}}{2} = 9 \pm \sqrt{17} \quad \Longrightarrow \quad \|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{9 + \sqrt{17}}.$$

Esercizio 3.10. Calcolare le norme 1, 2, ∞ delle seguenti matrici.

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

Calcolare inoltre $\rho(A)$ e $\rho(B)$, e verificare che

$$\rho(A) \le \|A\|_1,\ \|A\|_2,\ \|A\|_\infty,$$
$$\rho(B) \le \|B\|_1,\ \|B\|_2,\ \|B\|_\infty.$$

3.8.4 Equivalenza delle norme matriciali

Vale per le norme matriciali un teorema di equivalenza identico a quello per le norme vettoriali.

Teorema 3.11. *Tutte le norme matriciali in $\mathbb{C}^{n\times n}$ sono equivalenti, nel senso che se prendiamo due norme matriciali $\|\cdot\|', \|\cdot\|'' : \mathbb{C}^{n\times n} \to \mathbb{R}$ allora si ha*

$$\alpha\|A\|'' \le \|A\|' \le \beta\|A\|'' \ \text{ per ogni } \ A \in \mathbb{C}^{n\times n}, \tag{3.12}$$

dove $\alpha, \beta > 0$ sono due costanti indipendenti da A.

3.8.5 Successioni di matrici

Una successione di matrici $A^{(0)}$, $A^{(1)}$, $A^{(2)}$, ... in $\mathbb{C}^{n\times n}$ si dice convergente alla matrice $A \in \mathbb{C}^{n\times n}$ rispetto alla norma matriciale $\|\cdot\|$ se $\|A^{(k)} - A\| \to 0$. Poiché tutte le norme matriciali sono equivalenti per il Teorema 3.11, se una successione di matrici converge ad A rispetto a una norma $\|\cdot\|$ allora converge ad A rispetto a tutte le norme. La dimostrazione di questo fatto la saltiamo perché è identica a quella che abbiamo fatto per i vettori in Sezione 3.7.4.

Una successione di matrici $A^{(0)}$, $A^{(1)}$, $A^{(2)}$, ... in $\mathbb{C}^{n\times n}$ si dice convergente (componente per componente) alla matrice $A \in \mathbb{C}^{n\times n}$ se $A^{(k)} \to A$ componente per componente, cioè se

$$\begin{aligned} a_{ij}^{(k)} \to a_{ij} \ \text{ per ogni } \ i,j = 1,\dots,n \qquad &\Longleftrightarrow \qquad |a_{ij}^{(k)} - a_{ij}| \to 0 \ \text{ per ogni } \ i,j = 1,\dots,n \\ &\Longleftrightarrow \qquad \max_{i,j=1,\dots,n} |a_{ij}^{(k)} - a_{ij}| \to 0 \\ &\Longleftrightarrow \qquad |A^{(k)} - A|_\infty \to 0. \end{aligned}$$

Vediamo quindi che la convergenza componente per componente altro non è che la convergenza in $|\cdot|_\infty$. Pertanto, ricordando l'equivalenza di tutte le norme, dire che $A^{(k)} \to A$ componente per componente è lo stesso che dire che $A^{(k)} \to A$ in una qualsiasi norma.

Il prossimo teorema è fondamentale per i nostri scopi. Esso ci dice quando la successione delle potenze di una fissata matrice A tende alla matrice nulla.

Teorema 3.12. *Sia $A \in \mathbb{C}^{n\times n}$. Allora*

$$\lim_{k\to\infty} A^k = O \quad \iff \quad \rho(A) < 1.$$

Dimostrazione. Facciamo la dimostrazione soltanto nel caso in cui la matrice A è diagonalizzabile. In tal caso, esistono una matrice invertibile X e una matrice diagonale $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ (avente sulla diagonale gli autovalori di A) tali che

$$\begin{aligned} A &= XDX^{-1}, \\ A^2 &= XDX^{-1}XDX^{-1} = XD^2X^{-1}, \\ A^3 &= XDX^{-1}XDX^{-1}XDX^{-1} = XD^3X^{-1}, \\ &\vdots \\ A^k &= XD^kX^{-1}. \end{aligned}$$

($\Longleftarrow$) Se $\rho(A) < 1$ allora dall'equazione $A^k = XD^kX^{-1}$ e dalla proprietà di submoltiplicatività (proprietà 4 del Teorema 3.9) applicata alla norma $\|\cdot\|_\infty$ si ottiene [14]

$$\|A^k\|_\infty = \|XD^kX^{-1}\|_\infty \le \|X\|_\infty \|D^k\|_\infty \|X^{-1}\|_\infty = \|X\|_\infty \rho(A)^k \|X^{-1}\|_\infty \to 0,$$

per cui $\|A^k\|_\infty \to 0$ e $A^k \to O$.

($\Longrightarrow$) Viceversa, se $A^k \to O$ allora dall'equazione $A^k = XD^kX^{-1}$ si ottiene $D^k = X^{-1}A^kX$ e

$$\rho(A)^k = \|D^k\|_\infty = \|X^{-1}A^kX\|_\infty \le \|X^{-1}\|_\infty \|A^k\|_\infty \|X\|_\infty \to 0,$$

per cui $\rho(A)^k \to 0$ cioè $\rho(A) < 1$. □

4 Metodi iterativi per la risoluzione di sistemi lineari

È dato un sistema lineare

$$A\mathbf{x} = \mathbf{b} \tag{4.1}$$

con $\mathbf{b} \in \mathbb{C}^n$ e $A \in \mathbb{C}^{n\times n}$ invertibile. Tale sistema ha un'unica soluzione $\mathbf{x} = A^{-1}\mathbf{b}$. Ci proponiamo di risolvere (4.1) con un metodo iterativo, cioè un metodo che a partire da un vettore iniziale $\mathbf{x}^{(0)}$ scelto dall'utente costruisce una successione di vettori $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ Vogliamo che tale successione sia "facile da costruire" e converga a $\mathbf{x}$ (componente per componente) [15] qualunque sia il vettore iniziale scelto $\mathbf{x}^{(0)}$.

4.1 Forma generale di un metodo iterativo stazionario e proprietà di convergenza

Per risolvere (4.1) consideriamo solo *metodi iterativi stazionari* cioè metodi iterativi della forma

$$\boxed{\begin{aligned} &\mathbf{x}^{(0)} \in \mathbb{C}^n \text{ dato}, \\ &\mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \mathbf{q}, \quad k = 0, 1, 2, \ldots \end{aligned}} \tag{4.2}$$

dove $P \in \mathbb{C}^{n\times n}$ è una matrice fissata detta *matrice d'iterazione* e $\mathbf{q} \in \mathbb{C}^n$ è un vettore fissato.

[14] Si noti che $\|D^k\|_\infty = \|\mathrm{diag}(\lambda_1^k, \ldots, \lambda_n^k)\|_\infty = \max(|\lambda_1^k|, \ldots, |\lambda_n^k|) = \max(|\lambda_1|^k, \ldots, |\lambda_n|^k) = \max(|\lambda_1|, \ldots, |\lambda_n|)^k = \rho(A)^k$.

[15] Quando si parla di successione di vettori convergente, la convergenza è sempre intesa nel senso componente per componente. Sappiamo dalla Sezione 3.7.4 che la convergenza componente per componente è la stessa cosa della convergenza in una qualsiasi norma vettoriale.

Osservazione 4.1. Se una successione $\{\mathbf{x}^{(k)}\}_{k=0,1,2,\ldots}$ generata dal metodo (4.2) converge a un vettore $\mathbf{x}^{(\infty)}$ allora $\mathbf{x}^{(\infty)}$ soddisfa l'equazione [16]

$$\mathbf{x}^{(\infty)} = \lim_{k\to\infty} \mathbf{x}^{(k+1)} = \lim_{k\to\infty} (P\mathbf{x}^{(k)} + \mathbf{q}) = P\mathbf{x}^{(\infty)} + \mathbf{q} \implies \mathbf{x}^{(\infty)} = P\mathbf{x}^{(\infty)} + \mathbf{q}.$$

Di conseguenza, se la soluzione $\mathbf{x}$ di (4.1) non soddisfa l'equazione $\mathbf{x} = P\mathbf{x} + \mathbf{q}$ allora non c'è speranza che una successione $\{\mathbf{x}^{(k)}\}_{k=0,1,2,\ldots}$ generata dal metodo (4.2) converga a $\mathbf{x}$. Si noti che "soddisfare l'equazione $\mathbf{x} = P\mathbf{x} + \mathbf{q}$" vuol dire "essere punto fisso della funzione $\mathbf{g}(\mathbf{y}) = P\mathbf{y} + \mathbf{q}$".

Definizione 4.1 (**consistenza di un metodo iterativo**)**.** Il metodo iterativo (4.2) si dice consistente con il sistema (4.1) se la soluzione $\mathbf{x}$ di (4.1) soddisfa l'equazione $\mathbf{x} = P\mathbf{x} + \mathbf{q}$.

Definizione 4.2 (**convergenza di un metodo iterativo**)**.** Il metodo iterativo (4.2) per risolvere il sistema (4.1) si dice convergente se per ogni scelta del vettore iniziale $\mathbf{x}^{(0)}$ la successione $\{\mathbf{x}^{(k)}\}_{k=0,1,2,\ldots}$ generata dal metodo a partire da $\mathbf{x}^{(0)}$ converge (componente per componente) alla soluzione $\mathbf{x}$ di (4.1).

Teorema 4.1 (**condizione necessaria e sufficiente di convergenza**)**.** *Supponiamo che il metodo* (4.2) *sia consistente con* (4.1)*. Allora esso è convergente se e solo se* $\rho(P) < 1$.

Dimostrazione. Dimostriamo soltanto che se $\rho(P) < 1$ allora il metodo è convergente. Dobbiamo dimostrare che la successione (4.2) converge alla soluzione $\mathbf{x}$ di (4.1) indipendentemente dalla scelta del vettore iniziale $\mathbf{x}^{(0)}$. Poiché il metodo è consistente per ipotesi, vale l'equazione

$$\mathbf{x} = P\mathbf{x} + \mathbf{q}. \tag{4.3}$$

Inoltre, ovviamente, vale anche l'equazione del metodo, cioè

$$\mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \mathbf{q} \text{ per ogni } k = 0, 1, 2, \ldots \tag{4.4}$$

Sottraendo membro a membro la (4.4) e la (4.3) si ottiene l'equazione dell'errore

$$\mathbf{e}^{(k+1)} = P\mathbf{e}^{(k)} \text{ per ogni } k = 0, 1, 2, \ldots \tag{4.5}$$

dove $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$ è l'errore al passo k. Sviluppando per ricorrenza la (4.5) si ottiene

$$\mathbf{e}^{(k+1)} = P\mathbf{e}^{(k)} = P^2\mathbf{e}^{(k-1)} = P^3\mathbf{e}^{(k-2)} = \ldots = P^{k+1}\mathbf{e}^{(0)} \text{ per ogni } k = 0, 1, 2, \ldots$$

da cui

$$\mathbf{e}^{(k)} = P^k\mathbf{e}^{(0)} \text{ per ogni } k = 0, 1, 2, \ldots \tag{4.6}$$

Siccome stiamo assumendo che $\rho(P) < 1$, il Teorema 3.12 ci dice che $P^k \to O$. Dalla (4.6) si deduce quindi che $\mathbf{e}^{(k)} \to \mathbf{0}$ [17] cioè $\mathbf{x}^{(k)} \to \mathbf{x}$. □

Corollario 4.1 (**condizione sufficiente di convergenza**)**.** *Supponiamo che il metodo* (4.2) *sia consistente con* (4.1)*. Se esiste una norma matriciale indotta* $\|\cdot\|$ *tale che* $\|P\| < 1$ *allora il metodo è convergente.*

[16] Per dimostrare formalmente che $P\mathbf{x}^{(k)} + \mathbf{q} \to P\mathbf{x}^{(\infty)} + \mathbf{q}$, si osservi che

$$(P\mathbf{x}^{(k)} + \mathbf{q})_1 = P_{11}x_1^{(k)} + \ldots + P_{1n}x_n^{(k)} + q_1 \to P_{11}x_1^{(\infty)} + \ldots + P_{1n}x_n^{(\infty)} + q_1 = (P\mathbf{x}^{(\infty)} + \mathbf{q})_1$$

perché $\mathbf{x}^{(k)} \to \mathbf{x}$; similmente, si dimostra che ogni altra componente di $P\mathbf{x}^{(k)} + \mathbf{q}$ converge alla corrispondente componente di $P\mathbf{x}^{(\infty)} + \mathbf{q}$.

[17] Per dimostrarlo formalmente, si osservi che la prima componente di $\mathbf{e}^{(k)}$ è $(P^k)_{11}e_1^{(0)} + (P^k)_{12}e_2^{(0)} + \ldots + (P^k)_{1n}e_n^{(0)}$ e tende a 0 perché $(P^k)_{1j} \to 0$ per ogni $j = 1, 2, \ldots, n$; similmente, si dimostra che ogni altra componente di $\mathbf{e}^{(k)}$ tende a 0.

Dimostrazione. Poiché $\rho(P) \leq \|P\|$ per il Teorema 3.9, la condizione $\|P\| < 1$ implica che $\rho(P) < 1$ e dunque il metodo è convergente per il Teorema 4.1. □

Corollario 4.2 (**condizioni necessarie di convergenza**). *Supponiamo che il metodo* (4.2) *sia consistente con* (4.1).
- *Se* $|\text{traccia}(P)| \geq n$ *allora il metodo non è convergente.*
- *Se* $|\det(P)| \geq 1$ *allora il metodo non è convergente.*

Quindi le condizioni $|\text{traccia}(P)| < n$ *e* $|\det(P)| < 1$ *sono necessarie per la convergenza del metodo* (4.2).

Dimostrazione.
- Se $|\text{traccia}(P)| \geq n$ allora esiste almeno un autovalore di P di modulo ≥ 1. Infatti, se tutti gli autovalori $\lambda_1, \ldots, \lambda_n$ di P fossero di modulo < 1 allora avremmo

$$|\text{traccia}(P)| = |\lambda_1 + \ldots + \lambda_n| \leq |\lambda_1| + \ldots + |\lambda_n| < n.$$

  Poiché esiste almeno un autovalore di P di modulo ≥ 1 si deduce che $\rho(P) = \max(|\lambda_1|, \ldots, |\lambda_n|) \geq 1$ e dunque il metodo non è convergente per il Teorema 4.1.
- Se $|\det(P)| \geq 1$ allora esiste almeno un autovalore di P di modulo ≥ 1. Infatti, se tutti gli autovalori $\lambda_1, \ldots, \lambda_n$ di P fossero di modulo < 1 allora avremmo

$$|\det(P)| = |\lambda_1 \cdots \lambda_n| = |\lambda_1| \cdots |\lambda_n| < 1.$$

  Poiché esiste almeno un autovalore di P di modulo ≥ 1 si deduce che $\rho(P) = \max(|\lambda_1|, \ldots, |\lambda_n|) \geq 1$ e dunque il metodo non è convergente per il Teorema 4.1. □

Osservazione 4.2. Si può dimostrare che se il metodo iterativo (4.2) non è convergente, allora praticamente ogni scelta del vettore $\mathbf{x}^{(0)}$ produce una successione che non converge alla soluzione $\mathbf{x}$ del sistema (4.1). In sostanza, se il metodo iterativo (4.2) non è convergente allora è da buttare.

Esempio 4.1. Consideriamo il sistema lineare

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

e il metodo iterativo

$$\begin{aligned} &\mathbf{x}^{(0)} \in \mathbb{C}^2 \text{ dato,} \\ &\mathbf{x}^{(k+1)} = (I - A)\mathbf{x}^{(k)} + \mathbf{b}, \quad k = 0, 1, 2, \ldots \end{aligned}$$

(i) Stabilire se il metodo è consistente con il sistema dato.
(ii) Stabilire se il metodo è convergente.
(iii) Calcolare le prime 5 iterazioni del metodo partendo dal vettore d'innesco $\mathbf{x}^{(0)} = [0, 0]^T$ e confrontarle con la soluzione esatta.

Soluzione. (i) Il metodo assegnato è della forma (4.2) con $P = I - A$ e $\mathbf{q} = \mathbf{b}$. Sostituendo la soluzione $\mathbf{x}$ del sistema dato nell'equazione del metodo (al posto di $\mathbf{x}^{(k+1)}$ e $\mathbf{x}^{(k)}$) otteniamo

$$\mathbf{x} = (I - A)\mathbf{x} + \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{x} - A\mathbf{x} + \mathbf{b} \quad \Longleftrightarrow \quad A\mathbf{x} = \mathbf{b},$$

che è un'identità verificata. Dunque il metodo è consistente con il sistema dato.

(ii) La matrice d'iterazione del metodo è

$$P = I - A = \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix}.$$

Poiché $|\text{traccia}(P)| = |-1-1| = 2$, non è verificata la condizione necessaria di convergenza $|\text{traccia}(P)| < 2$ espressa nel Corollario 4.2, dunque il metodo non è convergente.

(iii) La soluzione esatta del sistema dato possiamo calcolarla ad esempio con il metodo di Gauss:

$$A\mathbf{x} = \mathbf{b} \quad \Longleftrightarrow \quad \begin{cases} 2x_1 + x_2 = 1 \\ x_1 + 2x_2 = 1 \end{cases} \quad \Longleftrightarrow \quad \begin{cases} 2x_1 + x_2 = 1 \\ \frac{3}{2}x_2 = \frac{1}{2} \end{cases} \quad \Longleftrightarrow \quad \begin{cases} x_1 = \frac{1}{3} \\ x_2 = \frac{1}{3} \end{cases}$$

dunque $\mathbf{x} = \left[\frac{1}{3}, \frac{1}{3}\right]^T$. Calcoliamo le prime 5 iterazioni del metodo partendo da $\mathbf{x}^{(0)} = [0,0]^T$:

$$\begin{aligned}
\mathbf{x}^{(1)} &= \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\
\mathbf{x}^{(2)} &= \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \\
\mathbf{x}^{(3)} &= \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \\
\mathbf{x}^{(4)} &= \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -5 \\ -5 \end{bmatrix}, \\
\mathbf{x}^{(5)} &= \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} -5 \\ -5 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 11 \\ 11 \end{bmatrix}.
\end{aligned}$$

Si nota che le iterazioni non stanno convergendo alla soluzione $\mathbf{x}$.

Esempio 4.2. Consideriamo il sistema lineare

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

e il metodo iterativo

$$\begin{aligned}
&\mathbf{x}^{(0)} \in \mathbb{C}^2 \text{ dato,} \\
&\mathbf{x}^{(k+1)} = \Big(I - \frac{1}{2}A\Big)\mathbf{x}^{(k)} + \frac{1}{2}\mathbf{b}, \quad k = 0, 1, 2, \ldots
\end{aligned}$$

(i) Stabilire se il metodo è consistente con il sistema dato.

(ii) Stabilire se il metodo è convergente.

(iii) Calcolare le prime 5 iterazioni del metodo partendo dal vettore d'innesco $\mathbf{x}^{(0)} = [0,0]^T$ e confrontarle con la soluzione esatta.

Soluzione. (i) Il metodo assegnato è della forma (4.2) con $P = I - \frac{1}{2}A$ e $\mathbf{q} = \frac{1}{2}\mathbf{b}$. Sostituendo la soluzione $\mathbf{x}$ del sistema dato nell'equazione del metodo (al posto di $\mathbf{x}^{(k+1)}$ e $\mathbf{x}^{(k)}$) otteniamo

$$\mathbf{x} = \Big(I - \frac{1}{2}A\Big)\mathbf{x} + \frac{1}{2}\mathbf{b} \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{x} - \frac{1}{2}A\mathbf{x} + \frac{1}{2}\mathbf{b} \quad \Longleftrightarrow \quad A\mathbf{x} = \mathbf{b},$$

che è un'identità verificata. Dunque il metodo è consistente con il sistema dato.

(ii) La matrice d'iterazione del metodo è

$$P = I - \frac{1}{2}A = \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix}.$$

Poiché $\|P\|_\infty = \frac{1}{2} < 1$, è verificata la condizione sufficiente di convergenza espressa nel Corollario 4.1, dunque il metodo è convergente.

(iii) La soluzione esatta del sistema dato l'abbiamo già calcolata nell'Esempio 4.1 ed è $\mathbf{x} = \left[\frac{1}{3}, \frac{1}{3}\right]^T$. Calcoliamo le prime 5 iterazioni del metodo partendo da $\mathbf{x}^{(0)} = [0, 0]^T$:

$$\begin{aligned}
\mathbf{x}^{(1)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \\
\mathbf{x}^{(2)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}, \\
\mathbf{x}^{(3)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{3}{8} \\ \frac{3}{8} \end{bmatrix} = \begin{bmatrix} 0.375 \\ 0.375 \end{bmatrix}, \\
\mathbf{x}^{(4)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{8} \\ \frac{3}{8} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{16} \\ \frac{5}{16} \end{bmatrix} = \begin{bmatrix} 0.3125 \\ 0.3125 \end{bmatrix}, \\
\mathbf{x}^{(5)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{16} \\ \frac{5}{16} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{11}{32} \\ \frac{11}{32} \end{bmatrix} = \begin{bmatrix} 0.34375 \\ 0.34375 \end{bmatrix}.
\end{aligned}$$

Si nota che le iterazioni stanno convergendo alla soluzione $\mathbf{x}$.

Esempio 4.3. Consideriamo il sistema lineare

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

e il metodo iterativo

$$\begin{aligned}
&\mathbf{x}^{(0)} \in \mathbb{C}^2 \text{ dato,} \\
&\mathbf{x}^{(k+1)} = (I - \omega A)\mathbf{x}^{(k)} + \omega \mathbf{b}, \quad k = 0, 1, 2, \ldots
\end{aligned}$$

dove $\omega \in \mathbb{R}$ è un parametro fissato.

(i) Stabilire per quali valori di ω il metodo è consistente con il sistema dato.

(ii) Stabilire per quali valori di ω il metodo è convergente.

Soluzione. (i) Il metodo assegnato è della forma (4.2) con $P = I - \omega A$ e $\mathbf{q} = \omega \mathbf{b}$. Sostituendo la soluzione $\mathbf{x}$ del sistema dato nell'equazione del metodo (al posto di $\mathbf{x}^{(k+1)}$ e $\mathbf{x}^{(k)}$) otteniamo

$$\mathbf{x} = (I - \omega A)\mathbf{x} + \omega \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{x} - \omega A\mathbf{x} + \omega \mathbf{b} \quad \Longleftrightarrow \quad \omega A\mathbf{x} = \omega \mathbf{b},$$

che è un'identità verificata. Dunque il metodo è sempre consistente con il sistema dato qualunque sia il valore di ω.

(ii) La matrice d'iterazione del metodo è

$$P = I - \omega A = \begin{bmatrix} 1-2\omega & -\omega \\ -\omega & 1-2\omega \end{bmatrix}.$$

Per capire quali sono i valori di ω per i quali il metodo è convergente dobbiamo capire quando è soddisfatta la condizione necessaria e sufficiente di convergenza $\rho(P) < 1$ espressa nel Teorema 4.1. In questo caso infatti non possiamo basarci sulle condizioni solo sufficienti o solo necessarie espresse nei Corollari 4.1–4.2. Calcoliamo dunque $\rho(P)$. Il polinomio caratteristico di P è

$$\begin{aligned} C_P(\lambda) = \det(\lambda I - P) &= \begin{vmatrix} \lambda - 1 + 2\omega & \omega \\ \omega & \lambda - 1 + 2\omega \end{vmatrix} \\ &= (\lambda - 1 + 2\omega)^2 - \omega^2 = \lambda^2 + 2(-1+2\omega)\lambda + (-1+2\omega)^2 - \omega^2, \end{aligned}$$

per cui gli autovalori di P sono

$$\lambda_{1,2} = 1 - 2\omega \pm \sqrt{(-1+2\omega)^2 - (-1+2\omega)^2 + \omega^2} = 1 - 2\omega \pm |\omega| = 1 - \omega,\ 1 - 3\omega.$$

Dunque

$$\rho(P) = \max(|1-\omega|, |1-3\omega|).$$

In conclusione, il metodo è convergente per i valori di ω tali che

$$\begin{aligned} \rho(P) < 1 \quad &\iff \quad \begin{cases} |1-\omega| < 1 \\ |1-3\omega| < 1 \end{cases} \quad \iff \quad \begin{cases} -1 < 1-\omega < 1 \\ -1 < 1-3\omega < 1 \end{cases} \quad \iff \quad \begin{cases} 0 < \omega < 2 \\ 0 < \omega < \dfrac{2}{3} \end{cases} \\ &\iff \quad 0 < \omega < \frac{2}{3}. \end{aligned}$$

Il metodo non è invece convergente per gli altri valori di ω.

Osservazione 4.3. In riferimento all'Esempio 4.3, osserviamo quanto segue.

- Se prendiamo $\omega = \frac{3}{4}$ allora

$$\begin{aligned} P &= \begin{bmatrix} -\frac{1}{2} & -\frac{3}{4} \\ -\frac{3}{4} & -\frac{1}{2} \end{bmatrix} \\ |\det(P)| &= \left| \frac{1}{4} - \frac{9}{16} \right| = \left| -\frac{5}{16} \right| = \frac{5}{16} < 1, \\ |\text{traccia}(P)| &= \left| -\frac{1}{2} - \frac{1}{2} \right| = 1 < 2, \end{aligned}$$

per cui sono soddisfatte le condizioni necessarie di convergenza espresse nel Corollario 4.2. Tuttavia, il metodo non è convergente perché $\frac{3}{4} > \frac{2}{3}$. Questo mostra che le condizioni espresse nel Corollario 4.2 sono necessarie ma *non sufficienti* a garantire la convergenza.

- Per ogni valore $\omega \in (0, \frac{2}{3})$ il metodo considerato è convergente. Ci si potrebbe chiedere: qual è il miglior valore di ω che possiamo scegliere? Risposta: il miglior valore di ω è quello che rende minimo il raggio spettrale della matrice d'iterazione P perché questo valore di ω è quello che assicura la maggiore velocità di convergenza (si veda la Sezione 4.2). Nel nostro caso, il miglior valore di ω è quello che rende minimo

$$\rho(P) = \max(|1-\omega|, |1-3\omega|).$$

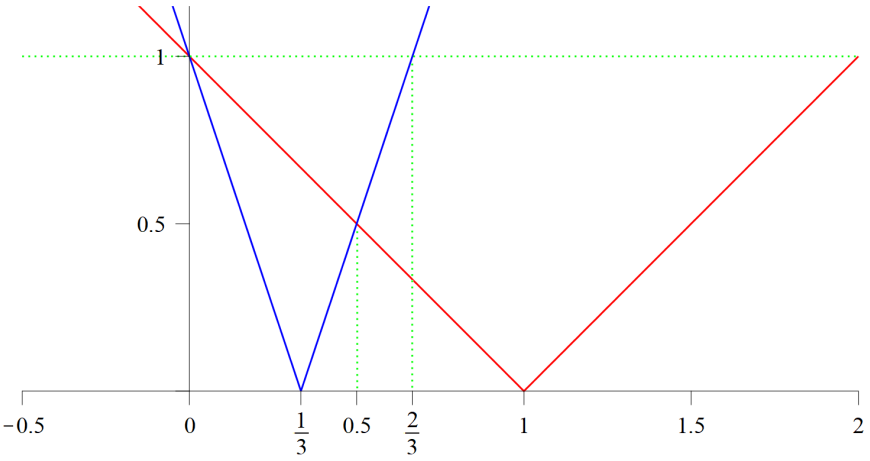

Figura 4.1: Grafici delle funzioni $|1-\omega|$ (rosso) e $|1-3\omega|$ (blu).

La Figura 4.1 mostra i grafici delle funzioni $|1-\omega|$ e $|1-3\omega|$, da cui si vede che il valore di ω che rende minimo $\rho(P)$ è $\omega_{\text{opt}} = \frac{1}{2}$. Il valore ω_{opt} è l'ascissa del punto d'intersezione delle rette di equazione $y = 1-\omega$ e $y = -1+3\omega$, e si ottiene risolvendo l'equazione $1-\omega = -1+3\omega$. Per $\omega = \omega_{\text{opt}} = \frac{1}{2}$ risulta $\rho(P) = \max(|1-\frac{1}{2}|, |1-\frac{3}{2}|) = \frac{1}{2}$ e il metodo che si ottiene è quello visto nell'Esempio 4.2 (esso coincide con il metodo di Jacobi che studieremo in Sezione 4.5).

Esercizio 4.1. Consideriamo il sistema lineare

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

e il metodo iterativo

$$\begin{aligned} &\mathbf{x}^{(0)} \in \mathbb{C}^2 \text{ dato,} \\ &\mathbf{x}^{(k+1)} = (I - \omega A)\mathbf{x}^{(k)} + \omega\mathbf{b}, \quad k = 0, 1, 2, \ldots \end{aligned}$$

dove $\omega \in \mathbb{R}$ è un parametro fissato.

(i) Stabilire per quali valori di ω il metodo è consistente con il sistema dato.
(ii) Stabilire per quali valori di ω il metodo è convergente.
(iii) Determinare il valore ω_{opt} di ω che rende minimo il raggio spettrale della matrice d'iterazione del metodo. Quanto vale il raggio spettrale della matrice d'iterazione per $\omega = \omega_{\text{opt}}$?
(iv) Fissato $\omega = \omega_{\text{opt}}$, calcolare le prime 4 iterazioni del metodo partendo da $\mathbf{x}^{(0)} = [0, 0]^T$ e confrontarle con la soluzione esatta.

4.2 Velocità di convergenza

Consideriamo il metodo (4.2) per risolvere il sistema (4.1) e supponiamo che esso sia convergente (cioè $\mathbf{x} = P\mathbf{x} + \mathbf{q}$ e $\rho(P) < 1$). Sulla base dell'equazione dell'errore $\mathbf{e}^{(k)} = P^k\mathbf{e}^{(0)}$ in (4.6), si possono dimostrare alcuni risultati teorici relativi alla velocità di convergenza del metodo nei quali non ci addentriamo. Per noi è sufficiente sapere questo fatto.

Fissiamo una qualsiasi norma vettoriale $\|\cdot\|$. Per quasi tutti i vettori $\mathbf{x}^{(0)} \in \mathbb{C}^n$, l'errore $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$ commesso dal metodo (4.2) soddisfa

$$\|\mathbf{e}^{(k)}\| \approx Ck^m\rho(P)^k \tag{4.7}$$

per ogni k abbastanza grande (in realtà nella pratica anche per k abbastanza piccolo), dove $0 \leq m \leq n-1$ è un intero che dipende solo da P e C è una costante indipendente da k.[18]

Vediamo quindi che la convergenza delle successioni $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... generate da un metodo della forma (4.2) risulta tanto più veloce quanto più $\rho(P)$ è piccolo. Sulla base di questo fatto, diamo la seguente definizione.

Definizione 4.3. Dati due metodi α e β della forma (4.2) per risolvere (4.1), entrambi convergenti, diremo che α *converge più velocemente di* β *se* $\rho(P_\alpha) < \rho(P_\beta)$, dove P_α e P_β indicano rispettivamente la matrice d'iterazione di α e quella di β.

4.3 Criterio di arresto del residuo

Consideriamo il metodo (4.2) per risolvere il sistema (4.1). La successione di vettori $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... generata dal metodo, anche quando risulta convergente alla soluzione $\mathbf{x}$ del sistema (4.1), deve essere comunque arrestata prima o poi: non si può pretendere di raggiungere esattamente la soluzione $\mathbf{x}$. Il criterio di arresto più utilizzato è quello del residuo: si sceglie una norma vettoriale $\|\cdot\|$ (tipicamente $\|\cdot\|_1$ oppure $\|\cdot\|_2$ oppure $\|\cdot\|_\infty$) e *si arresta la successione al primo vettore* $\mathbf{x}^{(K)}$ *che soddisfa la condizione*

$$\frac{\|\mathbf{r}^{(K)}\|}{\|\mathbf{b}\|} \leq \varepsilon, \tag{4.8}$$

dove $\mathbf{r}^{(K)} = \mathbf{b} - A\mathbf{x}^{(K)}$ è il residuo del sistema (4.1) relativo a $\mathbf{x}^{(K)}$ e $\varepsilon > 0$ è una soglia di precisione prefissata. La condizione (4.8) impone che l'errore relativo $\|A\mathbf{x}^{(K)} - \mathbf{b}\|/\|\mathbf{b}\|$ commesso approssimando $\mathbf{b}$ con $A\mathbf{x}^{(K)}$ sia $\leq \varepsilon$.[19] In tal modo, avremo che l'errore relativo sulla soluzione soddisfa

$$\begin{aligned}
\frac{\|\mathbf{x} - \mathbf{x}^{(K)}\|}{\|\mathbf{x}\|} &= \frac{\|A^{-1}(\mathbf{b} - A\mathbf{x}^{(K)})\|}{\|A^{-1}\mathbf{b}\|} = \frac{\|A^{-1}\mathbf{r}^{(K)}\|}{\|A^{-1}\mathbf{b}\|} \\
&\leq \frac{\|A^{-1}\|\,\|\mathbf{r}^{(K)}\|}{\|A^{-1}\mathbf{b}\|} = \frac{\|A\|\,\|A^{-1}\|\,\|\mathbf{r}^{(K)}\|}{\|A\|\,\|A^{-1}\mathbf{b}\|} \\
&\leq \frac{\|A\|\,\|A^{-1}\|\,\|\mathbf{r}^{(K)}\|}{\|AA^{-1}\mathbf{b}\|} = \frac{\|A\|\,\|A^{-1}\|\,\|\mathbf{r}^{(K)}\|}{\|\mathbf{b}\|} \\
&\leq \mu(A)\,\varepsilon,
\end{aligned}$$

dove $\mu(A) = \|A\|\,\|A^{-1}\|$ si chiama numero di condizionamento della matrice A in norma $\|\cdot\|$.

Osservazione 4.4. La successione di vettori $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... generata dal metodo (4.2), anche quando risulta convergente alla soluzione $\mathbf{x}$ del sistema (4.1), potrebbe impiegare troppo tempo a convergere. In tal caso, potrebbero volerci troppe iterazioni prima che venga soddisfatta la condizione di arresto del residuo (4.8). Per questo motivo è indispensabile, quando si implementa un metodo iterativo come (4.2), fissare sempre un numero massimo di iterazioni consentite. Questo peraltro serve anche ad arrestare le iterazioni quando non c'è convergenza.

[18] Per la precisione, C dipende dalla matrice d'iterazione P, dal vettore iniziale $\mathbf{x}^{(0)}$ e dalla norma utilizzata $\|\cdot\|$, ma non dall'indice d'iterazione k. Per quanto riguarda m che dipende solo da P, risulta $m = 0$ nel caso in cui tutti gli autovalori di P di modulo $\rho(P)$ hanno molteplicità algebrica uguale a quella geometrica; in particolare $m = 0$ se P è diagonalizzabile.

[19] $\|A\mathbf{x}^{(k)} - \mathbf{b}\|/\|\mathbf{b}\|$ rappresenta l'errore relativo con cui il vettore $A\mathbf{x}^{(k)}$ approssima $\mathbf{b} \neq \mathbf{0}$, così come $|\tilde{a} - a|/|a|$ rappresenta l'errore relativo con cui il numero $\tilde{a}$ approssima $a \neq 0$. Si usa l'errore relativo anziché l'errore assoluto perché, ad esempio, se fosse $a = 10000$ e $\tilde{a} = 9999$, diremmo che $\tilde{a}$ è una buona approssimazione di a anche se l'errore assoluto è $|\tilde{a} - a| = 1$ (lontano da 0), visto che comunque l'errore relativo è 10^{-4}, a indicare 4 cifre di approssimazione esatte circa. Similmente, diremmo che $\tilde{a} = 0.05$ non è una buona approssimazione di $a = 0.01$ anche se l'errore assoluto è $|\tilde{a} - a| = 0.04$ (vicino a 0), visto che l'errore relativo è 4; una buona approssimazione di $a = 0.01 = 0.010000$ è invece $\tilde{a} = 0.009999$, per la quale l'errore relativo è 10^{-4}, a indicare 4 cifre di approssimazione esatte circa (esattamente come per $a = 10000$ e $\tilde{a} = 9999$).

4.4 Costruzione di metodi iterativi mediante decomposizione della matrice

Descriviamo ora una procedura generale che consente di costruire un metodo della forma (4.2) per risolvere il sistema (4.1). Si considera una decomposizione di A del tipo

$$A = M - (M - A), \tag{4.9}$$

con $M \in \mathbb{C}^{n\times n}$ invertibile detta *precondizionatore*. Si osserva che

$$\begin{array}{rcl} A\mathbf{x} = \mathbf{b} & \iff & (M-(M-A))\mathbf{x} = \mathbf{b} \iff M\mathbf{x} = (M-A)\mathbf{x} + \mathbf{b} \\ & \iff & \mathbf{x} = M^{-1}(M-A)\mathbf{x} + M^{-1}\mathbf{b} \iff \mathbf{x} = \mathbf{x} + M^{-1}\mathbf{r}(\mathbf{x}), \end{array} \tag{4.10}$$

dove $\mathbf{r}(\mathbf{y}) = \mathbf{b} - A\mathbf{y}$ è il *residuo in* $\mathbf{y}$ del sistema (4.1), e si definisce il metodo

$$\boxed{\begin{array}{l} \mathbf{x}^{(0)} \in \mathbb{C}^n \text{ dato,} \\ \mathbf{x}^{(k+1)} = M^{-1}(M-A)\mathbf{x}^{(k)} + M^{-1}\mathbf{b} = \mathbf{x}^{(k)} + M^{-1}\mathbf{r}^{(k)}, \quad k = 0, 1, 2, \ldots \end{array}} \tag{4.11}$$

dove $\mathbf{r}^{(k)} = \mathbf{r}(\mathbf{x}^{(k)}) = \mathbf{b} - A\mathbf{x}^{(k)}$. Il metodo (4.11) è della forma (4.2) e la sua matrice d'iterazione è $M^{-1}(M-A) = I - M^{-1}A$. Esso è sicuramente consistente con (4.1) grazie alla (4.10). Quindi, per il Teorema 4.1, vale il seguente risultato.

Teorema 4.2. *Il metodo* (4.11) *per risolvere* (4.1) *è convergente se e solo se* $\rho(I - M^{-1}A) < 1$.

Osservazione 4.5. Il polinomio caratteristico di $I - M^{-1}A$ è dato da

$$C_{I-M^{-1}A}(\lambda) = \det(\lambda I - I + M^{-1}A) = \det(M^{-1}(\lambda M - M + A)) = \det(M^{-1})\det(\lambda M + A - M),$$

dove nell'ultima uguaglianza abbiamo usato il teorema di Binet. Pertanto,

$$C_{I-M^{-1}A}(\lambda) = 0 \iff \det(\lambda M + A - M) = 0. \tag{4.12}$$

In conclusione, gli autovalori e il raggio spettrale della matrice d'iterazione $I - M^{-1}A$ del metodo (4.11) possono essere calcolati risolvendo l'equazione a destra della (4.12), senza quindi calcolare esplicitamente né l'inversa M^{-1} né la matrice $I - M^{-1}A$.

Osservazione 4.6. L'iterazione k-esima del metodo (4.11) viene normalmente calcolata mediante la formula $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + M^{-1}\mathbf{r}^{(k)}$ e richiede il calcolo del vettore $\mathbf{z}^{(k)} = M^{-1}\mathbf{r}^{(k)}$ detto *residuo precondizionato*. Tale calcolo viene fatto risolvendo il sistema lineare $M\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ e non calcolando l'inversa M^{-1} (cosa tipicamente sconveniente da un punto di vista computazionale). Il sistema lineare $M\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ deve ovviamente essere più facile/rapido da risolvere rispetto al sistema originario $A\mathbf{x} = \mathbf{b}$, altrimenti non c'è nessun guadagno nel risolvere il sistema originario con il metodo (4.11)!

Osservazione 4.7. Intuitivamente, quanto più il precondizionatore M "è vicino/assomiglia" alla matrice A, tanto più il metodo (4.11) convergerà velocemente. Infatti, se $M \approx A$ allora intuitivamente $M - A \approx O$, $M^{-1}A \approx I$ e $M^{-1}(M-A) = I - M^{-1}A \approx O$, per cui ci si può aspettare che $\rho(I - M^{-1}A)$ sia piccolo. Il caso limite $M = A$ è quello in cui $I - M^{-1}A = O$ e il metodo converge in un'iterazione alla soluzione esatta $\mathbf{x}$, ma al prezzo che tale iterazione costa come la risoluzione del sistema originario $A\mathbf{x} = \mathbf{b}$. Nella scelta del precondizionatore M occorre quindi mediare fra "qualità dell'approssimazione $M \approx A$" e "facilità/rapidità della risoluzione di un sistema lineare di matrice M":

- una buona approssimazione $M \approx A$ generalmente assicura un buona velocità di convergenza;
- la facilità/rapidità della risoluzione di un sistema lineare di matrice M assicura che ogni iterazione del metodo (4.11) è rapida.

4.5 Metodi di Jacobi e Gauss-Seidel

4.5.1 Metodo di Jacobi

Supponiamo che A abbia elementi diagonali non nulli. In tal caso, la parte diagonale di A, cioè la matrice diagonale D ottenuta ricopiando la parte diagonale di A, è invertibile e possiamo definire il metodo di Jacobi, cioè il metodo (4.11) con $M = D$:

$$\boxed{\begin{array}{l}\mathbf{x}^{(0)} \in \mathbb{C}^n \text{ dato,} \\ \mathbf{x}^{(k+1)} = D^{-1}(D-A)\mathbf{x}^{(k)} + D^{-1}\mathbf{b} = \mathbf{x}^{(k)} + D^{-1}\mathbf{r}^{(k)}, \quad k = 0, 1, 2, \ldots\end{array}} \tag{4.13}$$

Il metodo di Jacobi è convergente se e solo se $\rho(J) < 1$, dove $J = D^{-1}(D - A) = I - D^{-1}A$. L'iterazione k-esima del metodo di Jacobi richiede di calcolare il vettore $\mathbf{z}^{(k)} = D^{-1}\mathbf{r}^{(k)}$ risolvendo il sistema diagonale $D\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$, il che è facilissimo:

$$\begin{cases} a_{11}z_1^{(k)} & & & = & r_1^{(k)} \\ & a_{22}z_2^{(k)} & & = & r_2^{(k)} \\ & & \ddots & \vdots & \\ & & a_{nn}z_n^{(k)} & = & r_n^{(k)} \end{cases} \iff \begin{cases} z_1^{(k)} = r_1^{(k)}/a_{11} \\ z_2^{(k)} = r_2^{(k)}/a_{22} \\ \vdots \\ z_n^{(k)} = r_n^{(k)}/a_{nn} \end{cases} \tag{4.14}$$

Il costo del calcolo di $\mathbf{z}^{(k)}$ è nD.[20]

Esercizio 4.2. Scrivere un programma MATLAB che implementa il metodo di Jacobi. Il programma deve:
- prendere in input la matrice A e il termine noto $\mathbf{b}$ del sistema lineare da risolvere $A\mathbf{x} = \mathbf{b}$, una soglia di precisione ε, un vettore d'innesco $\mathbf{x}^{(0)}$ e un numero massimo d'iterazioni consentite $N_{\max}$;
- restituire in output il primo vettore $\mathbf{x}^{(K)}$ calcolato dal metodo di Jacobi (con $0 \le K \le N_{\max}$) che soddisfa la condizione di arresto del residuo $\|\mathbf{r}^{(K)}\|_2 \le \varepsilon\|\mathbf{b}\|_2$, il relativo indice K che conta il numero d'iterazioni effettuate, e la norma $\|\mathbf{r}^{(K)}\|_2$ del residuo a cui ci si arresta. Se nessuno dei vettori $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(N_{\max})}$ soddisfa la condizione di arresto del residuo allora il programma deve restituire in output $\mathbf{x}^{(N_{\max})}$, il relativo indice $N_{\max}$, e la norma $\|\mathbf{r}^{(N_{\max})}\|_2$ dell'ultimo residuo.

4.5.2 Metodo di Gauss-Seidel

Supponiamo che A abbia elementi diagonali non nulli. In tal caso, la parte triangolare inferiore di A, cioè la matrice triangolare inferiore E ottenuta ricopiando la parte triangolare inferiore di A (inclusa la diagonale), è invertibile e possiamo definire il metodo di Gauss-Seidel, cioè il metodo (4.11) con $M = E$:

$$\boxed{\begin{array}{l}\mathbf{x}^{(0)} \in \mathbb{C}^n \text{ dato,} \\ \mathbf{x}^{(k+1)} = E^{-1}(E-A)\mathbf{x}^{(k)} + E^{-1}\mathbf{b} = \mathbf{x}^{(k)} + E^{-1}\mathbf{r}^{(k)}, \quad k = 0, 1, 2, \ldots\end{array}} \tag{4.15}$$

Il metodo di Gauss-Seidel è convergente se e solo se $\rho(G) < 1$, dove $G = E^{-1}(E - A) = I - E^{-1}A$. L'iterazione k-esima del metodo di Gauss-Seidel richiede di calcolare il vettore $\mathbf{z}^{(k)} = E^{-1}\mathbf{r}^{(k)}$ risolvendo il

[20] Indichiamo come al solito con A, M e D le addizioni, moltiplicazioni e divisioni.

sistema triangolare inferiore $E\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$, il che è facile (la soluzione si ottiene per sostituzione in avanti):

$$\begin{cases} a_{11}z_1^{(k)} & & & & & & & = & r_1^{(k)} \\ a_{21}z_1^{(k)} & + & a_{22}z_2^{(k)} & & & & & = & r_2^{(k)} \\ a_{31}z_1^{(k)} & + & a_{32}z_2^{(k)} & + & a_{33}z_3^{(k)} & & & = & r_3^{(k)} \\ \quad\vdots & & & & & & & & \vdots \\ a_{n1}z_1^{(k)} & + & a_{n2}z_2^{(k)} & + & \quad\dots\quad & + & a_{nn}z_n^{(k)} & = & r_n^{(k)} \end{cases}$$

$$\iff \quad \begin{cases} z_1^{(k)} = r_1^{(k)}/a_{11} \\ z_2^{(k)} = (r_2^{(k)} - a_{21}z_1^{(k)})/a_{22} \\ z_3^{(k)} = (r_3^{(k)} - a_{31}z_1^{(k)} - a_{32}z_2^{(k)})/a_{33} \\ \quad\vdots \\ z_n^{(k)} = (r_n^{(k)} - a_{n1}z_1^{(k)} - a_{n2}z_2^{(k)} - \dots - a_{n,n-1}z_{n-1}^{(k)})/a_{nn} \end{cases} \tag{4.16}$$

Per ogni $i = 1, \dots, n$, il costo del calcolo di

$$z_i^{(k)} = \frac{r_i^{(k)} - a_{i1}z_1^{(k)} - a_{i2}z_2^{(k)} - \dots - a_{i,i-1}z_{i-1}^{(k)}}{a_{ii}}$$

è $1D + (i-1)M + (i-1)A$, per cui il costo complessivo del calcolo di $\mathbf{z}^{(k)}$ è

$$\sum_{i=1}^{n}(1D + (i-1)M + (i-1)A) = nD + \frac{n(n-1)}{2}M + \frac{n(n-1)}{2}A.$$

Questo costo può ridursi se la parte triangolare inferiore E di A ha molti zeri.

Esercizio 4.3. Scrivere un programma MATLAB che implementa il metodo di Gauss-Seidel. Il programma deve:

- prendere in input la matrice A e il termine noto $\mathbf{b}$ del sistema lineare da risolvere $A\mathbf{x} = \mathbf{b}$, una soglia di precisione ε, un vettore d'innesco $\mathbf{x}^{(0)}$ e un numero massimo d'iterazioni consentite $N_{\max}$;
- restituire in output il primo vettore $\mathbf{x}^{(K)}$ calcolato dal metodo di Gauss-Seidel (con $0 \le K \le N_{\max}$) che soddisfa la condizione di arresto del residuo $\|\mathbf{r}^{(K)}\|_2 \le \varepsilon\|\mathbf{b}\|_2$, il relativo indice K che conta il numero d'iterazioni effettuate, e la norma $\|\mathbf{r}^{(K)}\|_2$ del residuo a cui ci si arresta. Se nessuno dei vettori $\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(N_{\max})}$ soddisfa la condizione di arresto del residuo allora il programma deve restituire in output $\mathbf{x}^{(N_{\max})}$, il relativo indice $N_{\max}$, e la norma $\|\mathbf{r}^{(N_{\max})}\|_2$ dell'ultimo residuo.

Osservazione 4.8. Confrontando i precondizionatori D ed E dei metodi di Jacobi e Gauss-Seidel, osserviamo quanto segue che si ricollega all'Osservazione 4.7.

- L'approssimazione $E \approx A$ è migliore dell'approssimazione $D \approx A$ perché $E - A$ ha più zeri di $D - A$. Questo spiega perché molto spesso il metodo di Gauss-Seidel converge più velocemente del metodo di Jacobi (cioè $\rho(G) < \rho(J)$, essendo J e G le matrici d'iterazione di Jacobi e Gauss-Seidel).
- La risoluzione di un sistema lineare di matrice E è più costosa della risoluzione di un sistema lineare di matrice D (cfr. (4.16) e (4.14)). Pertanto, un'iterazione di Gauss-Seidel costa di più di un'iterazione di Jacobi.

4.5.3 Teoremi di convergenza

Studiamo in questa sezione i teoremi di convergenza dei metodi di Jacobi e Gauss-Seidel.

Teorema 4.3. *Supponiamo che la matrice $A \in \mathbb{C}^{n\times n}$ soddisfi almeno una delle seguenti condizioni:*
- *A è a diagonale dominante e irriducibile;*
- *A è a diagonale dominante in senso stretto;*
- *A è a diagonale dominante per colonne e irriducibile;*
- *A è a diagonale dominante in senso stretto per colonne.*

Allora i metodi di Jacobi e Gauss-Seidel per risolvere un sistema lineare di matrice A sono convergenti.

Osservazione 4.9. Se $A \in \mathbb{C}^{n\times n}$ soddisfa almeno una delle condizioni del Teorema 4.3, allora:
- A è invertibile per il Teorema 3.7;
- gli elementi diagonali di A sono diversi da 0. Infatti, se per assurdo ce ne fosse uno uguale a 0, allora tutta la corrispondente riga (o colonna) sarebbe nulla in quanto A è a diagonale dominante (o a diagonale dominante per colonne): ciò è impossibile perché A è invertibile e non può avere una riga (o colonna) nulla.

Dunque, se $A \in \mathbb{C}^{n\times n}$ soddisfa almeno una delle condizioni del Teorema 4.3, i metodi di Jacobi e Gauss-Seidel sono applicabili per risolvere un sistema lineare di matrice A.

Dimostrazione del Teorema 4.3. Dimostriamo il teorema per il metodo di Gauss-Seidel sotto l'ipotesi che A sia a diagonale dominante e irriducibile (la dimostrazione del teorema negli altri sette casi è simile ed è lasciata come Esercizio 4.4). Dobbiamo dimostrare che $\rho(G) < 1$, dove $G = I - E^{-1}A$ è la matrice d'iterazione del metodo di Gauss-Seidel. Per l'Osservazione 4.5, gli autovalori di G sono le radici del polinomio [21]

$$\det(\lambda E + A - E) = \begin{vmatrix} \lambda a_{11} & a_{12} & a_{13} & a_{14} \\ \lambda a_{21} & \lambda a_{22} & a_{23} & a_{24} \\ \lambda a_{31} & \lambda a_{32} & \lambda a_{33} & a_{34} \\ \lambda a_{41} & \lambda a_{42} & \lambda a_{43} & \lambda a_{44} \end{vmatrix}.$$

Nessun numero λ di modulo ≥ 1 può essere radice di questo polinomio. Infatti, se $|\lambda| \geq 1$ allora la matrice $\lambda E + A - E$ è a diagonale dominante e irriducibile esattamente come A, per cui è invertibile (Teorema 3.7) e dunque $\det(\lambda E + A - E) \neq 0$. In definitiva, gli autovalori λ di G sono per forza in modulo minori di 1 e perciò $\rho(G) < 1$.

Per concludere, mostriamo più nel dettaglio che se $|\lambda| \geq 1$ allora la matrice $\lambda E + A - E$ è a diagonale dominante e irriducibile esattamente come A. La dimostrazione si basa sulle due osservazioni seguenti.
- $\lambda E + A - E$ è a diagonale dominante come A perché tutti gli elementi diagonali sono stati dilatati di un fattore λ di modulo ≥ 1 mentre gli elementi extradiagonali sono stati alcuni dilatati per il fattore λ (quelli sottodiagonali) e altri lasciati invariati (quelli sopradiagonali). Quindi la condizione di dominanza diagonale per $\lambda E + A - E$ è soddisfatta su ogni riga $i = 1, \ldots, n$:

$$\begin{aligned} |\lambda a_{ii}| &= |\lambda|\,|a_{ii}| \geq |\lambda|(|a_{i1}| + \ldots + |a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{in}|) && \text{(perché } A \text{ è a diagonale dominante)} \\ &= |\lambda|\,|a_{i1}| + \ldots + |\lambda|\,|a_{i,i-1}| + |\lambda|\,|a_{i,i+1}| + \ldots + |\lambda|\,|a_{in}| \\ &\geq |\lambda a_{i1}| + \ldots + |\lambda a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{in}| && \text{(perché } |\lambda| \geq 1) \end{aligned} \tag{4.17}$$

  Inoltre, la disuguaglianza stretta vale per la matrice $\lambda E + A - E$ sulle stesse righe su cui vale per A. Ad esempio, se la disguaglianza stretta valesse per A sulla riga i, allora in (4.17) avremmo il maggiore stretto $>$ anziché il $\geq$.

[21] Scriviamo la matrice $\lambda E + A - E$ nel caso $n = 4$ per maggiore chiarezza, ma il ragionamento vale qualunque sia n.

- $\lambda E + A - E$ ha gli elementi nulli nelle stesse posizioni di A, quindi ha lo stesso grafo di A, quindi è irriducibile come A. □

Esercizio 4.4. Dimostrare il Teorema 4.3 negli altri sette casi non considerati nella dimostrazione del teorema stesso, ovvero: (a) per il metodo di Jacobi sotto l'ipotesi che A sia a diagonale dominante e irriducibile; (b) per i metodi di Jacobi e Gauss-Seidel sotto le altre tre ipotesi (A a diagonale dominante in senso stretto, A a diagonale dominante per colonne e irriducibile, A a diagonale dominante in senso stretto per colonne).

Teorema 4.4. *Supponiamo che $A \in \mathbb{C}^{n\times n}$ sia hermitiana definita positiva. Allora il metodo di Gauss-Seidel per risolvere un sistema lineare di matrice A è convergente.*

Osservazione 4.10. Se $A \in \mathbb{C}^{n\times n}$ è hermitiana definita positiva, allora:
- A è invertibile perché i suoi autovalori sono positivi e dunque 0 non è un autovalore di A;
- gli elementi diagonali di A sono positivi per l'Esercizio 3.3.

Dunque, se $A \in \mathbb{C}^{n\times n}$ è hermitiana definita positiva, i metodi di Jacobi e Gauss-Seidel sono applicabili per risolvere un sistema lineare di matrice A.

Dimostrazione del Teorema 4.4. Dobbiamo dimostrare che $\rho(G) < 1$, dove $G = I - E^{-1}A$ è la matrice d'iterazione del metodo di Gauss-Seidel. La dimostrazione è suddivisa in due parti.

Parte 1. Dimostriamo che $A - G^*AG$ è hermitiana definita positiva. Il fatto che $A - G^*AG$ è hermitiana segue direttamente dall'ipotesi che A è hermitiana e dalla proprietà $(XY)^* = Y^*X^*$ (che vale per ogni coppia di matrici X, Y moltiplicabili):

$$(A - G^*AG)^* = A^* - G^*A^*G^{**} = A - G^*AG.$$

Per dimostrare che $A - G^*AG$ è definita positiva, osserviamo che

$$\begin{aligned}
A - G^*AG &= A - (I - E^{-1}A)^*A(I - E^{-1}A) \\
&= A - (I - F^*)A(I - F) && (F = E^{-1}A\text{; si noti che } F \text{ è invertibile in quanto } E^{-1} \text{ e } A \text{ lo sono,} \\
&&& \text{e si ha } F^{-1} = A^{-1}E) \\
&= A - A + F^*A + AF - F^*AF \\
&= F^*(AF^{-1} + F^{-*}A - A)F && (\text{per } F \text{ come per ogni matrice invertibile vale } (F^{-1})^* = (F^*)^{-1} \\
&&& \text{e si pone per definizione } F^{-*} = (F^{-1})^* = (F^*)^{-1}; \\
&&& \text{per verificare che } (F^{-1})^* = (F^*)^{-1} \text{ basta osservare che} \\
&&& (F^{-1})^*F^* = (FF^{-1})^* = I^* = I \text{ e similmente } F^*(F^{-1})^* = I) \\
&= F^*(E + E^* - A)F \\
&= F^*DF && (A \text{ è hermitiana } \implies E + E^* - A = D = \text{parte diagonale di } A)
\end{aligned}$$

Pertanto, per ogni $\mathbf{y} \neq \mathbf{0}$,

$$\begin{aligned}
\mathbf{y}^*(A - G^*AG)\mathbf{y} &= \mathbf{y}^*F^*DF\mathbf{y} \\
&= (F\mathbf{y})^*D(F\mathbf{y}) \\
&= \mathbf{u}^*D\mathbf{u} && (\mathbf{u} = F\mathbf{y} \neq \mathbf{0} \text{ perché } \mathbf{y} \neq \mathbf{0} \text{ e } F \text{ è invertibile}) \\
&= \sum_{i=1}^{n} a_{ii}|u_i|^2 > 0 && (a_{ii} > 0 \text{ per ogni } i = 1, \ldots, n \text{ perché } A \text{ è hermitiana definita positiva;} \\
&&& \text{vedere Esercizio 3.3})
\end{aligned}$$

per cui $A - G^*AG$ è definita positiva per il Teorema 3.1 (o per definizione).

Parte 2. Dimostriamo che se λ è un autovalore di G allora $|\lambda| < 1$. Una volta fatto questo, avremo $\rho(G) < 1$ e la tesi è dimostrata. Sia dunque λ un autovalore di G e sia $\mathbf{y} \neq \mathbf{0}$ un corrispondente autovettore: $G\mathbf{y} = \lambda\mathbf{y}$. Siccome $A - G^*AG$ è hermitiana definita positiva,

$$\begin{aligned} 0 < \mathbf{y}^*(A - G^*AG)\mathbf{y} &= \mathbf{y}^*A\mathbf{y} - \mathbf{y}^*G^*AG\mathbf{y} = \mathbf{y}^*A\mathbf{y} - (G\mathbf{y})^*A(G\mathbf{y}) = \mathbf{y}^*A\mathbf{y} - (\lambda\mathbf{y})^*A(\lambda\mathbf{y}) \\ &= \mathbf{y}^*A\mathbf{y} - \overline{\lambda}\mathbf{y}^*A(\lambda\mathbf{y}) \qquad \text{(vale in generale } (\alpha B)^* = \overline{\alpha}B^* \text{ per ogni } \alpha \in \mathbb{C} \text{ e ogni matrice } B) \\ &= \mathbf{y}^*A\mathbf{y} - |\lambda|^2\mathbf{y}^*A\mathbf{y} \\ &= (1 - |\lambda|^2)\mathbf{y}^*A\mathbf{y}. \end{aligned}$$

Poiché $\mathbf{y}^*A\mathbf{y} > 0$ per il Teorema 3.1 (essendo A hermitiana definita positiva per ipotesi), deve essere $1 - |\lambda|^2 > 0$ cioè $|\lambda| < 1$. □

Esempio 4.4. Consideriamo il sistema lineare

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 2 & 1 \\ 2 & -2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}.$$

(i) Stabilire se i metodi di Jacobi e Gauss-Seidel applicati al sistema dato sono convergenti e, nel caso lo siano, dire quale dei due converge più velocemente.
(ii) Calcolare le prime 6 iterazioni dei due metodi partendo dal vettore $\mathbf{x}^{(0)} = [0, 0]^T$ e confrontarle con la soluzione esatta.

Soluzione. (i) A è a diagonale dominante e irriducibile (il grafo di A contiene il ciclo $1 \to 2 \to 1$ che tocca tutti i nodi). Dunque i metodi di Jacobi e Gauss-Seidel sono entrambi convergenti per il Teorema 4.3. Per capire quale dei due converge più velocemente, dobbiamo calcolare il raggio spettrale delle rispettive matrici d'iterazione J e G e vedere quale dei due è più piccolo. Per calcolare i raggi spettrali $\rho(J)$ e $\rho(G)$ potremmo sfruttare l'Osservazione 4.5 ed evitare di calcolare esplicitamente J e G. Tuttavia, il calcolo di J e G sarebbe comunque necessario per risolvere il punto (ii), per cui tanto vale farlo subito. Posto

$$D = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} = \text{parte diagonale di } A,$$
$$E = \begin{bmatrix} 2 & 0 \\ 2 & -2 \end{bmatrix} = \text{parte triangolare inferiore di } A,$$

si ha [22]

$$D^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix},$$
$$E^{-1} = \frac{1}{-4}\begin{bmatrix} \begin{vmatrix} 1 & 0 \\ 0 & -2 \end{vmatrix} & \begin{vmatrix} 0 & 0 \\ 1 & -2 \end{vmatrix} \\ \begin{vmatrix} 2 & 1 \\ 2 & 0 \end{vmatrix} & \begin{vmatrix} 2 & 0 \\ 2 & 1 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} \boxed{\frac{1}{2}} & \boxed{0} \\ \frac{1}{2} & \boxed{-\frac{1}{2}} \end{bmatrix},$$

[22] Le componenti riquadrate erano già note senza ricorrere alla formula per l'inversa di una matrice, perché si può dimostrare che in generale *l'inversa di una matrice triangolare inferiore è ancora una matrice triangolare inferiore con elementi diagonali dati dagli inversi di quelli della matrice di partenza.*

per cui [23]

$$J = D^{-1}(D-A) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & -1 \\ -2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix},$$
$$G = E^{-1}(E-A) = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \boxed{0} & -\frac{1}{2} \\ \boxed{0} & -\frac{1}{2} \end{bmatrix}.$$

Il polinomio caratteristico di J è

$$C_J(\lambda) = \det(\lambda I - J) = \begin{vmatrix} \lambda & \frac{1}{2} \\ -1 & \lambda \end{vmatrix} = \lambda^2 + \frac{1}{2},$$

per cui gli autovalori di J sono $\pm\frac{\mathrm{i}}{\sqrt{2}}$ e dunque $\rho(J) = \frac{1}{\sqrt{2}}$. Gli autovalori di G (triangolare superiore) sono gli elementi diagonali $0, -\frac{1}{2}$ per cui $\rho(G) = \frac{1}{2}$. Siccome $\rho(G) = \rho(J)^2 < \rho(J)$, concludiamo che il metodo di Gauss-Seidel converge più velocemente del metodo di Jacobi. [24]

(ii) La soluzione esatta $\mathbf{x}$ del sistema possiamo calcolarla con il metodo di Gauss:

$$A\mathbf{x} = \mathbf{b} \iff \begin{cases} 2x_1 + x_2 = 3 \\ 2x_1 - 2x_2 = 0 \end{cases} \iff \begin{cases} 2x_1 + x_2 = 3 \\ -3x_2 = -3 \end{cases} \iff \begin{cases} x_1 = 1 \\ x_2 = 1 \end{cases}$$

dunque $\mathbf{x} = [1,1]^T$. Calcoliamo le prime 6 iterazioni del metodo di Jacobi partendo da $\mathbf{x}^{(0)} = [0,0]^T$. L'equazione del metodo di Jacobi è la seguente:

$$\mathbf{x}^{(k+1)} = J\mathbf{x}^{(k)} + D^{-1}\mathbf{b} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \mathbf{x}^{(k)} + \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \mathbf{x}^{(k)} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix}.$$

Le prime 6 iterazioni partendo da $\mathbf{x}^{(0)} = [0,0]^T$ sono

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix},$$
$$\mathbf{x}^{(2)} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix},$$
$$\mathbf{x}^{(3)} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{4} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.5 \end{bmatrix},$$
$$\mathbf{x}^{(4)} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{4} \\ \frac{3}{2} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{4} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 0.75 \end{bmatrix},$$
$$\mathbf{x}^{(5)} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{4} \\ \frac{3}{4} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{9}{8} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} 1.125 \\ 0.75 \end{bmatrix},$$

[23] Con riferimento agli elementi riquadrati, la prima colonna della matrice d'iterazione del metodo di Gauss-Seidel è sempre nulla, perché la prima colonna della matrice $E - A$ è sempre nulla. Quindi la matrice d'iterazione del metodo di Gauss-Seidel non è mai invertibile e 0 è sempre un suo autovalore.

[24] Molto spesso accade che quando entrambi i metodi di Jacobi e Gauss-Seidel convergono, il metodo di Gauss-Seidel converge più velocemente di quello di Jacobi; vedere anche l'Osservazione 4.8.

$$\mathbf{x}^{(6)} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{9}{8} \\ \frac{3}{4} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{9}{8} \\ \frac{9}{8} \end{bmatrix} = \begin{bmatrix} 1.125 \\ 1.125 \end{bmatrix}.$$

Calcoliamo le prime 6 iterazioni del metodo di Gauss-Seidel partendo da $\mathbf{x}^{(0)} = [0, 0]^T$. L'equazione del metodo di Gauss-Seidel è la seguente:

$$\mathbf{x}^{(k+1)} = G\mathbf{x}^{(k)} + E^{-1}\mathbf{b} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \mathbf{x}^{(k)} + \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \mathbf{x}^{(k)} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix}.$$

Le prime 6 iterazioni partendo da $\mathbf{x}^{(0)} = [0, 0]^T$ sono

$$\begin{aligned}
\mathbf{x}^{(1)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}, \\
\mathbf{x}^{(2)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{4} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 0.75 \end{bmatrix}, \\
\mathbf{x}^{(3)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{3}{4} \\ \frac{3}{4} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{9}{8} \\ \frac{9}{8} \end{bmatrix} = \begin{bmatrix} 1.125 \\ 1.125 \end{bmatrix}, \\
\mathbf{x}^{(4)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{9}{8} \\ \frac{9}{8} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{15}{16} \\ \frac{15}{16} \end{bmatrix} = \begin{bmatrix} 0.9375 \\ 0.9375 \end{bmatrix}, \\
\mathbf{x}^{(5)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{15}{16} \\ \frac{15}{16} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{33}{32} \\ \frac{33}{32} \end{bmatrix} = \begin{bmatrix} 1.03125 \\ 1.03125 \end{bmatrix}, \\
\mathbf{x}^{(6)} &= \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{33}{32} \\ \frac{33}{32} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{63}{64} \\ \frac{63}{64} \end{bmatrix} = \begin{bmatrix} 0.984375 \\ 0.984375 \end{bmatrix}.
\end{aligned}$$

Si nota che la successione generata dal metodo di Gauss-Seidel converge più velocemente rispetto a quella generata dal metodo di Jacobi. Per l'esattezza, la velocità di convergenza appare doppia: un passo di Gauss-Seidel corrisponde a due passi di Jacobi. Ciò è in accordo con il fatto che $\rho(G) = \rho(J)^2$ e con la (4.7) per $m = 0$ (vedere nota 18 a piè di pag. 55 tenendo conto che sia J che G sono diagonalizzabili avendo entrambe 2 autovalori distinti).

Esempio 4.5. Si consideri la matrice

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

(i) Stabilire se il metodo di Jacobi applicato a un sistema lineare di matrice A è convergente.
(ii) Stabilire se il metodo di Gauss-Seidel applicato a un sistema lineare di matrice A è convergente.

Soluzione. (i) La matrice A non è a diagonale dominante né per righe né per colonne, dunque non abbiamo teoremi che ci permettano di dire che il metodo di Jacobi è convergente. Per capire se il metodo di Jacobi è convergente, calcoliamo il raggio spettrale della sua matrice d'iterazione J. Usiamo due metodi.

Primo metodo. Calcoliamo esplicitamente J. Detta D la parte diagonale di A, si ha

$$J = D^{-1}(D - A) = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} = \underbrace{\frac{1}{2} \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}}_{J'}.$$

Il polinomio caratteristico di J' è dato da

$$C_{J'}(\lambda) = \det(\lambda I - J') = \begin{vmatrix} \lambda & 1 & 1 \\ 1 & \lambda & 1 \\ 1 & 1 & \lambda \end{vmatrix} = \lambda \begin{vmatrix} \lambda & 1 \\ 1 & \lambda \end{vmatrix} - \begin{vmatrix} 1 & 1 \\ 1 & \lambda \end{vmatrix} + \begin{vmatrix} 1 & \lambda \\ 1 & 1 \end{vmatrix}$$
$$= \lambda(\lambda^2 - 1) - (\lambda - 1) + 1 - \lambda = \lambda^3 - 3\lambda + 2.$$

Si nota che 1 è una radice, per cui dividendo $\lambda^3 - 3\lambda + 2$ per $\lambda - 1$ si ottiene

$$C_{J'}(\lambda) = (\lambda - 1)(\lambda^2 + \lambda - 2) \quad \Longrightarrow \quad \lambda_{1,2,3} = 1, \frac{-1 \pm \sqrt{1+8}}{2} = 1, 1, -2.$$

Dunque gli autovalori di $J = \frac{1}{2}J'$ sono $\frac{1}{2}, \frac{1}{2}, -1$ e $\rho(J) = 1$.
Secondo metodo. Sfruttiamo l'Osservazione 4.5. Detta D la parte diagonale di A, calcoliamo gli autovalori di J risolvendo l'equazione $\det(\lambda D + A - D) = 0$. Si ha

$$\det(\lambda D + A - D) = \begin{vmatrix} 2\lambda & 1 & 1 \\ 1 & 2\lambda & 1 \\ 1 & 1 & 2\lambda \end{vmatrix} = 2\lambda \begin{vmatrix} 2\lambda & 1 \\ 1 & 2\lambda \end{vmatrix} - \begin{vmatrix} 1 & 1 \\ 1 & 2\lambda \end{vmatrix} + \begin{vmatrix} 1 & 2\lambda \\ 1 & 1 \end{vmatrix}$$
$$= 2\lambda(4\lambda^2 - 1) - (2\lambda - 1) + (1 - 2\lambda) = 8\lambda^3 - 6\lambda + 2.$$

Si nota che -1 è una radice, per cui dividendo $8\lambda^3 - 6\lambda + 2$ per $\lambda + 1$ si ottiene

$$\det(\lambda D + A - D) = (\lambda + 1)(8\lambda^2 - 8\lambda + 2) \quad \Longrightarrow \quad \lambda_{1,2,3} = -1, \frac{8 \pm \sqrt{64 - 64}}{16} = -1, \frac{1}{2}, \frac{1}{2}.$$

Dunque gli autovalori di J sono $-1, \frac{1}{2}, \frac{1}{2}$ e $\rho(J) = 1$.
In conclusione, entrambi i metodi forniscono, come dev'essere, gli stessi autovalori di J e lo stesso raggio spettrale $\rho(J) = 1$: il metodo di Jacobi non è convergente.

(ii) La matrice A è hermitiana e i determinanti delle sottomatrici principali di testa di A sono

$$\det(A_1) = 2,$$
$$\det(A_2) = 3,$$
$$\det(A_3) = C_{J'}(2) = 4.$$

Siccome tutti i determinanti delle sottomatrici principali di testa di A sono positivi, A è definita positiva per il Teorema 3.1, dunque il metodo di Gauss-Seidel è convergente per il Teorema 4.4.

Esercizio 4.5. Si consideri ancora la matrice A dell'Esempio 4.5. Dimostrare che il metodo di Gauss-Seidel per risolvere un sistema lineare di matrice A è convergente calcolando il raggio spettrale della matrice d'iterazione G del metodo di Gauss-Seidel e verificando che esso risulta < 1. Per calcolare il raggio spettrale $\rho(G)$, si proceda in due modi distinti:
- calcolando esplicitamente G;
- sfruttando l'Osservazione 4.5.

Esempio 4.6. È dato un sistema lineare $A\mathbf{x} = \mathbf{b}$ con

$$A = \begin{bmatrix} \alpha & 1 & 0 \\ 1 & \alpha & 1 \\ 0 & 1 & \alpha \end{bmatrix}, \qquad \alpha \in \mathbb{R}, \qquad \alpha \neq 0.$$

(i) Stabilire se la matrice A è irriducibile.
(ii) Stabilire per quali valori di α la matrice A è a diagonale dominante.
(iii) Determinare per quali valori di α il metodo di Jacobi applicato al sistema dato converge e confrontare la risposta con i valori del punto (ii).
(iv) Stabilire per quali valori di α la matrice A è definita positiva.
(v) Determinare per quali valori di α il metodo di Gauss-Seidel applicato al sistema dato converge e confrontare la risposta con i valori dei punti (ii) e (iv).
(vi) Per i valori di α per i quali sia il metodo di Jacobi che quello di Gauss-Seidel sono convergenti, stabilire quale dei due converge più velocemente.

Soluzione. (i) La matrice A è irriducibile. Infatti, il grafo associato ad A contiene il ciclo $1 \to 2 \to 3 \to 2 \to 1$ che tocca tutti i nodi (osserviamo che tale ciclo è determinato dagli elementi sopradiagonali e sottodiagonali).

(ii) La matrice A è a diagonale dominante per $|\alpha| \geq 2$, mentre per $|\alpha| < 2$ non può essere a diagonale dominante per via della seconda riga.

(iii) Calcoliamo il raggio spettrale della matrice d'iterazione J del metodo di Jacobi e determiniamo i valori di α per cui risulta $\rho(J) < 1$. Usiamo due metodi per calcolare $\rho(J)$. Nel seguito, indichiamo con D la parte diagonale di A.

Primo metodo. Calcoliamo esplicitamente J. Si ha

$$J = D^{-1}(D - A) = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix} = \frac{1}{\alpha} \underbrace{\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}}_{J'}.$$

Il polinomio caratteristico di J' è dato da

$$C_{J'}(\lambda) = \det(\lambda I - J') = \begin{vmatrix} \lambda & 1 & 0 \\ 1 & \lambda & 1 \\ 0 & 1 & \lambda \end{vmatrix} = \lambda(\lambda^2 - 1) - \lambda = \lambda(\lambda^2 - 2),$$

per cui gli autovalori di J' sono $0, \pm\sqrt{2}$, gli autovalori di J sono $0, \pm\frac{\sqrt{2}}{\alpha}$, e $\rho(J) = \frac{\sqrt{2}}{|\alpha|}$.

Secondo metodo. Sfruttiamo l'Osservazione 4.5. Calcoliamo gli autovalori di J risolvendo l'equazione $\det(\lambda D + A - D) = 0$. Si ha

$$\det(\lambda D + A - D) = \begin{vmatrix} \lambda\alpha & 1 & 0 \\ 1 & \lambda\alpha & 1 \\ 0 & 1 & \lambda\alpha \end{vmatrix} = \lambda\alpha(\lambda^2\alpha^2 - 1) - \lambda\alpha = \lambda\alpha(\lambda^2\alpha^2 - 2),$$

per cui gli autovalori di J sono $0, \pm\frac{\sqrt{2}}{\alpha}$ e $\rho(J) = \frac{\sqrt{2}}{|\alpha|}$.

In conclusione, entrambi i metodi forniscono, come dev'essere, gli stessi autovalori di J e lo stesso raggio spettrale $\rho(J) = \frac{\sqrt{2}}{|\alpha|}$. Poiché

$$\rho(J) = \frac{\sqrt{2}}{|\alpha|} < 1 \quad \iff \quad |\alpha| > \sqrt{2},$$

il metodo di Jacobi risulta convergente per $|\alpha| > \sqrt{2}$ e non convergente per $|\alpha| \leq \sqrt{2}$. Si osservi che per $\sqrt{2} < |\alpha| < 2$ (ad esempio $\alpha = \frac{3}{2}$) la matrice A non è a diagonale dominante ma il metodo di Jacobi converge ugualmente: questo mostra che le condizioni del Teorema 4.3 sono sufficienti ma non necessarie per la convergenza del metodo di Jacobi.

(iv) Siccome A è hermitiana, essa è definita positiva se e solo se i determinanti delle sue sottomatrici principali di testa sono tutti positivi (Teorema 3.1). I determinanti delle sottomatrici principali di testa di A sono

$$\det(A_1) = \alpha, \qquad \det(A_2) = \alpha^2 - 1, \qquad \det(A_3) = C_{J'}(\alpha) = \alpha(\alpha^2 - 2),$$

e risultano tutti positivi se e solo se $\alpha > \sqrt{2}$. Dunque A è definita positiva se e solo se $\alpha > \sqrt{2}$.

(v) Calcoliamo il raggio spettrale della matrice d'iterazione G del metodo di Gauss-Seidel e determiniamo i valori di α per cui risulta $\rho(G) < 1$. Usiamo due metodi per calcolare $\rho(G)$. Nel seguito, indichiamo con E la parte triangolare inferiore di A.

Primo metodo. Calcoliamo esplicitamente G. Si ha [25]

$$G = E^{-1}(E - A) = \begin{bmatrix} \alpha & 0 & 0 \\ 1 & \alpha & 0 \\ 0 & 1 & \alpha \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \frac{1}{\alpha^3} \begin{bmatrix} \begin{vmatrix} 1 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 1 & \alpha \end{vmatrix} & \begin{vmatrix} 0 & 0 & 0 \\ 1 & \alpha & 0 \\ 0 & 1 & \alpha \end{vmatrix} & \begin{vmatrix} 0 & 0 & 0 \\ 0 & \alpha & 0 \\ 1 & 1 & \alpha \end{vmatrix} \\ \begin{vmatrix} \alpha & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & \alpha \end{vmatrix} & \begin{vmatrix} \alpha & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & \alpha \end{vmatrix} & \begin{vmatrix} \alpha & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & \alpha \end{vmatrix} \\ \begin{vmatrix} \alpha & 0 & 1 \\ 1 & \alpha & 0 \\ 0 & 1 & 0 \end{vmatrix} & \begin{vmatrix} \alpha & 0 & 0 \\ 1 & \alpha & 1 \\ 0 & 1 & 0 \end{vmatrix} & \begin{vmatrix} \alpha & 0 & 0 \\ 1 & \alpha & 0 \\ 0 & 1 & 1 \end{vmatrix} \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \frac{1}{\alpha^3} \begin{bmatrix} \boxed{\alpha^2} & \boxed{0} & \boxed{0} \\ -\alpha & \boxed{\alpha^2} & \boxed{0} \\ 1 & -\alpha & \boxed{\alpha^2} \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} = \underbrace{\frac{1}{\alpha^3} \begin{bmatrix} \boxed{0} & -\alpha^2 & 0 \\ \boxed{0} & \alpha & -\alpha^2 \\ \boxed{0} & -1 & \alpha \end{bmatrix}}_{G'}.$$

Il polinomio caratteristico di G' è dato da

$$C_{G'}(\lambda) = \det(\lambda I - G') = \begin{vmatrix} \lambda & \alpha^2 & 0 \\ 0 & \lambda - \alpha & \alpha^2 \\ 0 & 1 & \lambda - \alpha \end{vmatrix} = \lambda((\lambda - \alpha)^2 - \alpha^2) = \lambda^2(\lambda - 2\alpha)$$

per cui gli autovalori di G' sono $0, 0, 2\alpha$, gli autovalori di G sono $0, 0, \frac{2}{\alpha^2}$, e $\rho(G) = \frac{2}{\alpha^2}$.

Secondo metodo. Sfruttiamo l'Osservazione 4.5. Calcoliamo gli autovalori di G risolvendo l'equazione $\det(\lambda E + A - E) = 0$. Si ha

$$\det(\lambda E + A - E) = \begin{vmatrix} \lambda\alpha & 1 & 0 \\ \lambda & \lambda\alpha & 1 \\ 0 & \lambda & \lambda\alpha \end{vmatrix} = \lambda\alpha(\lambda^2\alpha^2 - \lambda) - \lambda^2\alpha = \lambda^2\alpha(\lambda\alpha^2 - 2),$$

per cui gli autovalori di G sono $0, 0, \frac{2}{\alpha^2}$ e $\rho(G) = \frac{2}{\alpha^2}$.

[25] Per gli elementi riquadrati, si vedano le note 22 e 23 a piè di pag. 61.

In conclusione, entrambi i metodi forniscono, come dev'essere, gli stessi autovalori di G e lo stesso raggio spettrale $\rho(G) = \frac{2}{\alpha^2}$. Poiché

$$\rho(G) = \frac{2}{\alpha^2} < 1 \quad \iff \quad |\alpha| > \sqrt{2},$$

il metodo di Gauss-Seidel risulta convergente per $|\alpha| > \sqrt{2}$ e non convergente per $|\alpha| \leq \sqrt{2}$. Si osservi che per i valori di α per cui la matrice A è definita positiva ($\alpha > \sqrt{2}$) il metodo di Gauss-Seidel converge in accordo al Teorema 4.4. Si osservi anche che per $-2 < \alpha < -\sqrt{2}$ (ad esempio $\alpha = -\frac{3}{2}$) la matrice A non è né a diagonale dominante né definita positiva ma il metodo di Gauss-Seidel converge ugualmente: questo mostra che le condizioni dei Teoremi 4.3 e 4.4 sono sufficienti ma non necessarie per la convergenza del metodo di Gauss-Seidel.

(vi) I valori di α per cui sia il metodo di Jacobi che quello di Gauss-Seidel convergono sono quelli che soddisfano $|\alpha| > \sqrt{2}$. Per tutti questi valori, si ha

$$\rho(G) = \frac{2}{\alpha^2} = \left(\frac{\sqrt{2}}{|\alpha|}\right)^2 = \rho(J)^2,$$

per cui il metodo di Gauss-Seidel converge più velocemente di quello di Jacobi.

Esercizio 4.6. Si consideri il sistema lineare

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

(i) Stabilire se il metodo di Jacobi applicato al sistema dato converge.
(ii) Calcolare le prime 3 iterazioni del metodo di Jacobi partendo dal vettore $\mathbf{x}^{(0)} = [0, 0, 1]^T$ e confrontare il risultato con la soluzione esatta del sistema.
(iii) Stabilire se il metodo di Gauss-Seidel applicato al sistema dato converge.
(iv) Calcolare le prime 3 iterazioni del metodo di Gauss-Seidel partendo dal vettore $\mathbf{x}^{(0)} = [0, 0, 1]^T$ e confrontare il risultato con la soluzione esatta del sistema.

Esercizio 4.7. Si consideri la matrice

$$A = \begin{bmatrix} -2 & \alpha & 0 \\ \alpha & 2 & \alpha \\ 0 & \alpha & -2 \end{bmatrix}, \qquad \alpha \in \mathbb{R}.$$

(i) Stabilire per quali valori di α la matrice A soddisfa almeno una delle seguenti condizioni: (a) A è a diagonale dominante in senso stretto; (b) A è a diagonale dominante e irriducibile.
(ii) Stabilire per quali valori di α la matrice A è definita positiva.
(iii) Stabilire per quali valori di α il metodo di Jacobi applicato a un sistema lineare di matrice A converge.
(iv) Stabilire per quali valori di α il metodo di Gauss-Seidel applicato a un sistema lineare di matrice A converge.
(v) Per i valori di α per cui entrambi i metodi di Jacobi e Gauss-Seidel convergono, stabilire quale dei due converge più velocemente.

Esercizio 4.8. Si consideri la matrice

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

(i) Stabilire se il metodo di Jacobi applicato a un sistema lineare di matrice A è convergente.
(ii) Stabilire se il metodo di Gauss-Seidel applicato a un sistema lineare di matrice A è convergente.
(iii) Dimostrare che la matrice

$$M = \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 1 & -2 & 0 \\ \hline 0 & 0 & 3 \end{array}\right]$$

è invertibile e verificare che la sua inversa è data da

$$M^{-1} = \left[\begin{array}{c|c} L^{-1} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \hline \begin{matrix} 0 & 0 \end{matrix} & \frac{1}{3} \end{array}\right], \qquad L = \begin{bmatrix} 1 & 2 \\ 1 & -2 \end{bmatrix}.$$

(iv) Stabilire se il metodo iterativo relativo alla decomposizione $A = M - (M - A)$ applicato a un sistema lineare di matrice A è convergente.
(v) Dire quale dei metodi menzionati ai punti (i), (ii), (iv) converge più velocemente.

5 Esercizi di riepilogo risolti

Riportiamo in questo capitolo conclusivo alcuni esercizi di riepilogo (esercizi "tipo esame") con le relative soluzioni.

Esercizio 5.1. Si consideri la funzione $f(x) = \frac{\sin x}{x}$.
(a) Scrivere in forma canonica e in forma di Lagrange il polinomio d'interpolazione $p(x)$ della funzione $f(x)$ sui nodi $x_0 = \frac{\pi}{2}$, $x_1 = \frac{3\pi}{4}$, $x_2 = \pi$.
(b) Dimostrare che

$$f''(x) = -\frac{\sin x}{x} - \frac{2\cos x}{x^2} + \frac{2\sin x}{x^3}, \qquad f'''(x) = -\frac{\cos x}{x} + \frac{3\sin x}{x^2} + \frac{6\cos x}{x^3} - \frac{6\sin x}{x^4}.$$

(c) Fornire una stima dell'errore d'interpolazione $|f(x) - p(x)|$ per ogni $x \in [\frac{\pi}{2}, \pi]$, cioè determinare una costante C tale che $|f(x) - p(x)| \le C$ per ogni $x \in [\frac{\pi}{2}, \pi]$.
(d) Fissato $\varepsilon > 0$, determinare un n tale che la formula dei trapezi I_n di ordine n per approssimare l'integrale $I = \int_{\pi/2}^{\pi} f(x)\mathrm{d}x$ sia affetta da un errore $|I - I_n| \le \varepsilon$. Quanto vale n se $\varepsilon = 10^{-8}$?

Soluzione.
(a) La forma di Lagrange di $p(x)$ è data da

$$\begin{aligned} p(x) &= f(x_0)\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f(x_1)\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f(x_2)\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= \frac{2}{\pi}\,\frac{(x-\frac{3\pi}{4})(x-\pi)}{(\frac{\pi}{2}-\frac{3\pi}{4})(\frac{\pi}{2}-\pi)} + \frac{2\sqrt{2}}{3\pi}\,\frac{(x-\frac{\pi}{2})(x-\pi)}{(\frac{3\pi}{4}-\frac{\pi}{2})(\frac{3\pi}{4}-\pi)} + 0\,\frac{(x-\frac{\pi}{2})(x-\frac{3\pi}{4})}{(\pi-\frac{\pi}{2})(\pi-\frac{3\pi}{4})} \\ &= \frac{2}{\pi}\,\frac{(x-\frac{3\pi}{4})(x-\pi)}{(\pi^2/8)} + \frac{2\sqrt{2}}{3\pi}\,\frac{(x-\frac{\pi}{2})(x-\pi)}{-(\pi^2/16)}. \end{aligned}$$

Sviluppando i calcoli, portiamo il polinomio $p(x)$ in forma canonica:

$$\begin{aligned} p(x) &= \frac{2}{\pi}\,\frac{(x-\frac{3\pi}{4})(x-\pi)}{(\pi^2/8)} + \frac{2\sqrt{2}}{3\pi}\,\frac{(x-\frac{\pi}{2})(x-\pi)}{-(\pi^2/16)} = \frac{16}{\pi^3}\Big(x-\frac{3\pi}{4}\Big)(x-\pi) - \frac{32\sqrt{2}}{3\pi^3}\Big(x-\frac{\pi}{2}\Big)(x-\pi) \\ &= \Big(\frac{16}{\pi^3} - \frac{32\sqrt{2}}{3\pi^3}\Big)x^2 + \Big(\frac{16\sqrt{2}}{\pi^2} - \frac{28}{\pi^2}\Big)x + \frac{12}{\pi} - \frac{16\sqrt{2}}{3\pi}. \end{aligned}$$

Per verificare che questo è effettivamente il polinomio d'interpolazione richiesto (e che quindi non ci sono errori di calcolo), basta calcolare $p(x_0)$, $p(x_1)$, $p(x_2)$ e verificare che si ottengono, rispettivamente, $f(x_0) = \frac{2}{\pi}$, $f(x_1) = \frac{2\sqrt{2}}{3\pi}$, $f(x_2) = 0$.

(b) Usando la formula di derivazione di un quoziente, si ha

$$f'(x) = \frac{(\cos x)x - \sin x}{x^2} = \frac{\cos x}{x} - \frac{\sin x}{x^2},$$

$$\begin{aligned} f''(x) &= \frac{(-\sin x)x - \cos x}{x^2} - \frac{(\cos x)x^2 - (\sin x)(2x)}{x^4} = -\frac{\sin x}{x} - \frac{\cos x}{x^2} - \frac{\cos x}{x^2} + \frac{2\sin x}{x^3} \\ &= -\frac{\sin x}{x} - \frac{2\cos x}{x^2} + \frac{2\sin x}{x^3}, \end{aligned}$$

$$\begin{aligned} f'''(x) &= -\frac{(\cos x)x - \sin x}{x^2} - \frac{(-2\sin x)x^2 - (2\cos x)(2x)}{x^4} + \frac{(2\cos x)x^3 - (2\sin x)(3x^2)}{x^6} \\ &= -\frac{\cos x}{x} + \frac{\sin x}{x^2} + \frac{2\sin x}{x^2} + \frac{4\cos x}{x^3} + \frac{2\cos x}{x^3} - \frac{6\sin x}{x^4} \\ &= -\frac{\cos x}{x} + \frac{3\sin x}{x^2} + \frac{6\cos x}{x^3} - \frac{6\sin x}{x^4}. \end{aligned}$$

(c) In base al teorema sull'errore dell'interpolazione (che è applicabile perché la funzione $f(x) = \frac{\sin x}{x}$ è di classe $C^\infty[\frac{\pi}{2}, \pi]$), per ogni $x \in [\frac{\pi}{2}, \pi]$ si ha

$$\begin{aligned} |f(x) - p(x)| &= \left| \frac{f'''(\xi)}{3!}\Big(x - \frac{\pi}{2}\Big)\Big(x - \frac{3\pi}{4}\Big)(x - \pi) \right| \qquad (\xi \text{ è un punto in } (\tfrac{\pi}{2}, \pi)) \\ &= \frac{1}{6}\left| -\frac{\cos\xi}{\xi} + \frac{3\sin\xi}{\xi^2} + \frac{6\cos\xi}{\xi^3} - \frac{6\sin\xi}{\xi^4} \right| \left| x - \frac{\pi}{2} \right| \left| x - \frac{3\pi}{4} \right| |x - \pi| \\ &\qquad \text{(ricordiamo che il modulo di un prodotto è il prodotto dei moduli} \\ &\qquad \text{così come il modulo di un rapporto è il rapporto dei moduli)} \\ &\le \frac{1}{6}\left(\left|\frac{\cos\xi}{\xi}\right| + \left|\frac{3\sin\xi}{\xi^2}\right| + \left|\frac{6\cos\xi}{\xi^3}\right| + \left|\frac{6\sin\xi}{\xi^4}\right| \right) \frac{\pi}{2} \cdot \frac{\pi}{4} \cdot \frac{\pi}{2} \\ &\qquad \text{(abbiamo usato la disuguaglianza triangolare più il fatto che, per ogni } x \in [\tfrac{\pi}{2}, \pi], \text{ si ha} \\ &\qquad |x - \tfrac{\pi}{2}| \le \tfrac{\pi}{2},\ |x - \tfrac{3\pi}{4}| \le \tfrac{\pi}{4},\ |x - \pi| \le \tfrac{\pi}{2}) \\ &= \frac{1}{6}\left(\frac{|\cos\xi|}{|\xi|} + \frac{3\,|\sin\xi|}{\xi^2} + \frac{6\,|\cos\xi|}{|\xi|^3} + \frac{6\,|\sin\xi|}{\xi^4} \right) \frac{\pi^3}{16} \\ &\le \frac{1}{6}\left(\frac{1}{(\pi/2)} + \frac{3}{(\pi/2)^2} + \frac{6}{(\pi/2)^3} + \frac{6}{(\pi/2)^4} \right) \frac{\pi^3}{16} \\ &\qquad \text{(abbiamo usato il fatto che, qualunque sia } \xi \in (\tfrac{\pi}{2}, \pi), \text{ si ha } |\cos\xi|, |\sin\xi| \le 1 \text{ e } |\xi| \ge \tfrac{\pi}{2}) \\ &\approx 1.4166. \end{aligned}$$

Volendo ottenere una stima più precisa, si può procedere nel modo seguente. Per ogni $x \in [\frac{\pi}{2}, \pi]$ si ha

$$\begin{aligned} |f(x) - p(x)| &= \left| \frac{f'''(\xi)}{3!}\Big(x - \frac{\pi}{2}\Big)\Big(x - \frac{3\pi}{4}\Big)(x - \pi) \right| \qquad (\xi \text{ è un punto in } (\tfrac{\pi}{2}, \pi)) \\ &= \frac{1}{6}\left| -\frac{\cos\xi}{\xi} + \frac{3\sin\xi}{\xi^2} + \frac{6\cos\xi}{\xi^3} - \frac{6\sin\xi}{\xi^4} \right| \left| \Big(x - \frac{\pi}{2}\Big)\Big(x - \frac{3\pi}{4}\Big)(x - \pi) \right| \\ &\le \frac{1}{6}\left(\frac{1}{(\pi/2)} + \frac{3}{(\pi/2)^2} + \frac{6}{(\pi/2)^3} + \frac{6}{(\pi/2)^4} \right) \max_{y \in [\frac{\pi}{2}, \pi]} \Big| \underbrace{\Big(y - \frac{\pi}{2}\Big)\Big(y - \frac{3\pi}{4}\Big)(y - \pi)}_{\omega(y)} \Big|, \end{aligned} \tag{5.1}$$

dove nell'ultima disuguaglianza abbiamo usato i passaggi fatti precedentemente per stimare la quantità

$$\left| -\frac{\cos\xi}{\xi} + \frac{3\sin\xi}{\xi^2} + \frac{6\cos\xi}{\xi^3} - \frac{6\sin\xi}{\xi^4} \right|.$$

Andiamo a calcolare il massimo di $|\omega(y)|$ su $[\frac{\pi}{2}, \pi]$. Per farlo, dobbiamo cercare tutti i massimi e i minimi relativi di $\omega(y)$ su $[\frac{\pi}{2}, \pi]$ e scegliere il più grande di essi in modulo. Per un teorema dell'analisi matematica, i massimi e i minimi relativi di $\omega(y)$ si trovano o nei punti di bordo dell'intervallo $[\frac{\pi}{2}, \pi]$ oppure nei punti stazionari di $\omega(y)$ in $[\frac{\pi}{2}, \pi]$, cioè quei punti di $[\frac{\pi}{2}, \pi]$ in cui si annulla la derivata $\omega'(y)$. Si ha

$$\begin{aligned}
&\omega(y) = \Big(y - \frac{\pi}{2}\Big)\Big(y - \frac{3\pi}{4}\Big)(y - \pi) = y^3 - \frac{9\pi}{4}\, y^2 + \frac{13\pi^2}{8}\, y - \frac{3\pi^3}{8}, \\
&\omega'(y) = 3y^2 - \frac{9\pi}{2}\, y + \frac{13\pi^2}{8}, \\
&\omega'(y) = 0 \quad \Longleftrightarrow \quad y = y_{1,2} = \frac{\frac{9\pi}{2} \pm \sqrt{\Big(\frac{9\pi}{2}\Big)^2 - \frac{39\pi^2}{2}}}{6} = \frac{3\pi}{4} \pm \frac{\sqrt{3}\,\pi}{12}.
\end{aligned}$$

Siccome $y_{1,2}$ stanno in $[\frac{\pi}{2}, \pi]$, essi sono punti stazionari di $\omega(y)$ in $[\frac{\pi}{2}, \pi]$. Dunque, per quanto detto sopra,

$$\begin{aligned}
\max_{y\in[\frac{\pi}{2},\pi]} |\omega(y)| &= \max\bigg(\Big|\omega\Big(\frac{\pi}{2}\Big)\Big|, \Big|\omega\Big(\frac{3\pi}{4} + \frac{\sqrt{3}\,\pi}{12}\Big)\Big|, \Big|\omega\Big(\frac{3\pi}{4} - \frac{\sqrt{3}\,\pi}{12}\Big)\Big|, |\omega(\pi)| \bigg) \\
&= \max\Big(0, \frac{\pi^3\sqrt{3}}{288}, \frac{\pi^3\sqrt{3}}{288}, 0\Big) = \frac{\pi^3\sqrt{3}}{288}.
\end{aligned}$$

Dunque, tornando a (5.1), otteniamo

$$|f(x) - p(x)| \le \frac{1}{6}\left(\frac{1}{(\pi/2)} + \frac{3}{(\pi/2)^2} + \frac{6}{(\pi/2)^3} + \frac{6}{(\pi/2)^4} \right) \frac{\pi^3\sqrt{3}}{288} \approx 0.1363.$$

(d) Sia $\varepsilon > 0$ fissato. In base al teorema sull'errore della formula dei trapezi (che è applicabile perché la funzione $f(x) = \frac{\sin x}{x}$ è $C^\infty[\frac{\pi}{2}, \pi]$ come già osservato), per ogni n si ha

$$\begin{aligned}
|I - I_n| &= \left| -\frac{(\pi - \frac{\pi}{2}) f''(\eta)}{12} \left(\frac{\pi - \frac{\pi}{2}}{n} \right)^2 \right| \qquad (\eta \text{ è un punto in } [\tfrac{\pi}{2}, \pi]) \\
&= \frac{(\frac{\pi}{2})^3 |f''(\eta)|}{12n^2} = \frac{(\frac{\pi}{2})^3}{12n^2} \left| -\frac{\sin\eta}{\eta} - \frac{2\cos\eta}{\eta^2} + \frac{2\sin\eta}{\eta^3} \right| \le \frac{(\frac{\pi}{2})^3}{12n^2} \left(\frac{|\sin\eta|}{|\eta|} + \frac{2\,|\cos\eta|}{\eta^2} + \frac{2\,|\sin\eta|}{|\eta|^3} \right) \\
&\le \frac{(\frac{\pi}{2})^3}{12n^2} \left(\frac{1}{(\pi/2)} + \frac{2}{(\pi/2)^2} + \frac{2}{(\pi/2)^3} \right) = \frac{C}{n^2} \qquad (C = \tfrac{\pi^2}{48} + \tfrac{\pi}{12} + \tfrac{1}{6})
\end{aligned}$$

Imponiamo

$$\frac{C}{n^2} \le \varepsilon \quad \Longleftrightarrow \quad n \ge \sqrt{\frac{C}{\varepsilon}}.$$

Dunque, se prendiamo $n \ge \sqrt{C/\varepsilon}$ siamo sicuri che $|I - I_n| \le \varepsilon$. In particolare, per $\varepsilon = 10^{-8}$ dovremo prendere $n \ge \sqrt{C/10^{-8}} \approx 7962.9317$. □

Esercizio 5.2. Si consideri la funzione

$$f(x) = \frac{1}{\log(x+2)},$$

dove "log" indica il logaritmo naturale. Per ogni intero $n \geq 1$ indichiamo con I_n la formula dei trapezi di ordine n per approssimare $I = \int_0^1 f(x)\mathrm{d}x = 1.118424814...$

(a) Dimostrare che

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

(b) Fissato $\varepsilon > 0$, determinare un n tale che $|I - I_n| \leq \varepsilon$. Quanto vale n se $\varepsilon = 10^{-8}$?

(c) Per $n = 3$, calcolare i valori I_n, I_{2n}.

(d) Per $n = 3$, scrivere in forma di Lagrange il polinomio d'interpolazione $p(x)$ dei dati (h^2, I_n) e $((\frac{h}{2})^2, I_{2n})$, dove $h = \frac{1}{n}$ è il passo di discretizzazione della formula dei trapezi I_n (e $\frac{h}{2}$ quello della formula dei trapezi I_{2n}). Confrontare inoltre i valori I_n, I_{2n}, $p(0)$ con il valore esatto I.

Soluzione.

(a) Notiamo che $f(x) = \log^{-1}(x+2)$. Usando la regola di derivazione della funzione composta per il calcolo di $f'(x)$ e la regola di derivazione di un quoziente per il calcolo di $f''(x)$, otteniamo

$$f'(x) = -\log^{-2}(x+2)\frac{1}{x+2} = -\frac{1}{(x+2)\log^2(x+2)},$$
$$f''(x) = -\frac{-\left(\log^2(x+2) + (x+2)\cdot 2\log(x+2)\frac{1}{x+2}\right)}{(x+2)^2\log^4(x+2)} = \frac{2+\log(x+2)}{(x+2)^2\log^3(x+2)}.$$

(b) Sia $\varepsilon > 0$ fissato. In base al teorema sull'errore della formula dei trapezi (che è applicabile perché la funzione $f(x) = \frac{1}{\log(x+2)}$ è $C^\infty[0,1]$), per ogni n si ha

$$|I - I_n| = \left|-\frac{f''(\eta)}{12n^2}\right| \qquad (\eta \text{ è un punto in } [0,1])$$
$$= \frac{|f''(\eta)|}{12n^2} = \frac{1}{12n^2}\left|\frac{2+\log(\eta+2)}{(\eta+2)^2\log^3(\eta+2)}\right| \leq \frac{1}{12n^2}\left(\frac{2+\log 3}{4\log^3 2}\right) = \frac{C}{n^2} \qquad (C = \tfrac{2+\log 3}{12(4\log^3 2)})$$

Osserviamo che nella disuguaglianza precedente abbiamo usato il fatto che, qualunque sia $\eta \in [0,1]$, si ha $2+\log(\eta+2) \leq 2+\log 3$, $(\eta+2)^2 \geq 4$ e $\log^3(\eta+2) \geq \log^3 2$. Imponiamo

$$\frac{C}{n^2} \leq \varepsilon \quad \iff \quad n \geq \sqrt{\frac{C}{\varepsilon}}.$$

Dunque, se prendiamo $n \geq \sqrt{C/\varepsilon}$ siamo sicuri che $|I - I_n| \leq \varepsilon$. In particolare, per $\varepsilon = 10^{-8}$ dovremo prendere $n \geq \sqrt{C/10^{-8}} \approx 4402.75$.

(c) Per un n generico, la formula dei trapezi in questione I_n è data da

$$I_n = h\left[\frac{f(0)+f(1)}{2} + \sum_{j=1}^{n-1} f(jh)\right], \qquad h = \frac{1}{n}.$$

L'esercizio chiede di calcolare I_3 e I_6. Si ha

$$I_3 = \frac{1}{3}\left[\frac{f(0)+f(1)}{2} + \sum_{j=1}^{2} f\Big(\frac{j}{3}\Big)\right] = \frac{1}{3}\left[\frac{1}{2\log 2} + \frac{1}{2\log 3} + \frac{1}{\log\frac{7}{3}} + \frac{1}{\log\frac{8}{3}}\right]$$
$$= 1.125411694...$$
$$I_6 = \frac{1}{6}\left[\frac{f(0)+f(1)}{2} + \sum_{j=1}^{5} f\Big(\frac{j}{6}\Big)\right] = \frac{1}{6}\left[\frac{1}{2\log 2} + \frac{1}{2\log 3} + \frac{1}{\log\frac{13}{6}} + \frac{1}{\log\frac{14}{6}} + \frac{1}{\log\frac{15}{6}} + \frac{1}{\log\frac{16}{6}} + \frac{1}{\log\frac{17}{6}}\right]$$
$$= 1.120188541...$$

(d) Il polinomio d'interpolazione richiesto $p(x)$ si scrive in forma di Lagrange nel modo seguente:

$$p(x) = I_n \frac{x-(\frac{h}{2})^2}{h^2-(\frac{h}{2})^2} + I_{2n}\frac{x-h^2}{(\frac{h}{2})^2-h^2} = I_3\frac{x-\frac{1}{36}}{\frac{1}{9}-\frac{1}{36}} + I_6\frac{x-\frac{1}{9}}{\frac{1}{36}-\frac{1}{9}} = I_3\Big(12\,x-\frac{1}{3}\Big) + I_6\Big(-12\,x+\frac{4}{3}\Big).$$

Risulta

$$p(0) = -\frac{1}{3}\,I_3 + \frac{4}{3}\,I_6 = 1.118447490...$$

Confrontando I_3, I_6, $p(0)$ con il valore esatto I, si nota che $p(0)$ è molto più vicino a I di I_3 e I_6, mentre I_6 è più vicino a I di I_3 ma non di molto. Infatti,

$$|I_3 - I| \approx 0.0069868 = 6.9868\cdot 10^{-3},$$
$$|I_6 - I| \approx 0.0017637 = 1.7637\cdot 10^{-3},$$
$$|p(0) - I| \approx 0.0000226 = 2.26\cdot 10^{-5}.$$ □

Esercizio 5.3. Si consideri la matrice

$$A = \begin{bmatrix} 1 & 2 & 0 \\ -2 & 4 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

(a) Stabilire se il metodo di Jacobi applicato a un sistema lineare di matrice A è convergente.
(b) Stabilire se il metodo di Gauss-Seidel applicato a un sistema lineare di matrice A è convergente.
(c) Dimostrare che la matrice

$$M = \begin{bmatrix} 1 & 2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

è invertibile e verificare che la sua inversa è data da

$$M^{-1} = \left[\begin{array}{cc|c} & & 0 \\ & L^{-1} & 0 \\ \hline 0 & 0 & 1 \end{array}\right], \qquad L = \begin{bmatrix} 1 & 2 \\ -2 & 4 \end{bmatrix}.$$

(d) Stabilire se il metodo iterativo associato alla decomposizione $A = M-(M-A)$ applicato a un sistema lineare di matrice A è convergente.
(e) Dire quale dei metodi menzionati ai punti (a), (b), (d) converge più velocemente.

Soluzione.

(a) La matrice A non è a diagonale dominante né per righe né per colonne. Per stabilire la convergenza o meno del metodo di Jacobi, calcoliamo il raggio spettrale della sua matrice d'iterazione J e vediamo se è minore di 1 oppure no. Sia

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

la parte diagonale di A. Gli autovalori di $J = D^{-1}(D - A)$ sono le soluzioni dell'equazione $\det(\lambda D + A - D) = 0$. Si ha

$$\det(\lambda D + A - D) = \begin{vmatrix} \lambda & 2 & 0 \\ -2 & 4\lambda & 1 \\ 0 & 1 & \lambda \end{vmatrix} = \lambda(4\lambda^2 - 1) + 2 \cdot 2\lambda = \lambda(4\lambda^2 + 3),$$

per cui gli autovalori di J sono $0, \pm \mathrm{i}\frac{\sqrt{3}}{2}$ e $\rho(J) = \frac{\sqrt{3}}{2} < 1$: il metodo di Jacobi è convergente.

(b) La matrice A non è a diagonale dominante né per righe né per colonne, e non è nemmeno simmetrica (hermitiana). Per stabilire la convergenza o meno del metodo di Gauss-Seidel, calcoliamo il raggio spettrale della sua matrice d'iterazione G e vediamo se è minore di 1 oppure no. Sia

$$E = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 4 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

la parte triangolare inferiore di A. Gli autovalori di $G = E^{-1}(E - A)$ sono le soluzioni dell'equazione $\det(\lambda E + A - E) = 0$. Si ha

$$\det(\lambda E + A - E) = \begin{vmatrix} \lambda & 2 & 0 \\ -2\lambda & 4\lambda & 1 \\ 0 & \lambda & \lambda \end{vmatrix} = \lambda(4\lambda^2 - \lambda) + 2\lambda \cdot 2\lambda = \lambda^2(4\lambda + 3),$$

per cui gli autovalori di G sono $0, 0, -\frac{3}{4}$ e $\rho(G) = \frac{3}{4} < 1$: il metodo di Gauss-Seidel è convergente.

(c) Si ha $\det(L) = 8 \neq 0$ per cui L è invertibile e la sua inversa è

$$L^{-1} = \frac{1}{8} \begin{bmatrix} \begin{vmatrix} 1 & 2 \\ 0 & 4 \end{vmatrix} & \begin{vmatrix} 0 & 2 \\ 1 & 4 \end{vmatrix} \\ \begin{vmatrix} 1 & 1 \\ -2 & 0 \end{vmatrix} & \begin{vmatrix} 1 & 0 \\ -2 & 1 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{8} \end{bmatrix}.$$

Mediante calcolo diretto, si può verificare che

$$\left[\begin{array}{cc|c} & L^{-1} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \hline 0 & 0 & 1 \end{array}\right] M = M \left[\begin{array}{cc|c} & L^{-1} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \hline 0 & 0 & 1 \end{array}\right] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Dunque, per definizione, la matrice M è invertibile perché abbiamo trovato la sua inversa che è

$$M^{-1} = \left[\begin{array}{cc|c} & L^{-1} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \hline 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|c} \frac{1}{2} & -\frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{8} & 0 \\ \hline 0 & 0 & 1 \end{array}\right].$$

(d) Il metodo iterativo associato alla decomposizione $A = M - (M - A)$ per risolvere un sistema lineare di matrice A ha matrice d'iterazione

$$P = M^{-1}(M-A) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{8} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{1}{4} \\ 0 & 0 & -\frac{1}{8} \\ 0 & -1 & 0 \end{bmatrix}.$$

Per stabilire se il metodo è convergente oppure no, calcoliamo il raggio spettrale di P e vediamo se è minore di 1 oppure no. Il polinomio caratteristico di P è

$$C_P(\lambda) = \det(\lambda I - P) = \begin{vmatrix} \lambda & 0 & -\frac{1}{4} \\ 0 & \lambda & \frac{1}{8} \\ 0 & 1 & \lambda \end{vmatrix} = \lambda\Big(\lambda^2 - \frac{1}{8}\Big).$$

Gli autovalori di P sono quindi $0, \pm\frac{1}{\sqrt{8}}$, per cui $\rho(P) = \frac{1}{\sqrt{8}} < 1$: il metodo è convergente.

(e) Il metodo al punto (d) è quello che converge più velocemente in quanto il raggio spettrale della sua matrice d'iterazione è $\frac{1}{\sqrt{8}}$ ed è il più piccolo di tutti. □

Esercizio 5.4. Si consideri la matrice

$$A = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}.$$

(a) Stabilire se i metodi di Jacobi e Gauss-Seidel applicati a un sistema lineare di matrice A sono convergenti.

(b) Sia ω un numero reale positivo fissato, sia

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

la parte diagonale di A, e sia $M = \frac{1}{\omega}D$. Consideriamo il metodo iterativo associato alla decomposizione $A = M - (M - A)$ per risolvere un sistema lineare di matrice A. Si scriva la matrice d'iterazione P di questo metodo e si calcoli il raggio spettrale $\rho(P)$.

(c) Tracciare il grafico della funzione $f(\omega) = \rho(P)^2$ per $\omega \in (0, \infty)$ e stabilire per quali valori di $\omega \in (0, \infty)$ risulta $f(\omega) < 1$.

(d) Stabilire per quali valori di $\omega \in (0, \infty)$ il metodo menzionato al punto (b) risulta convergente.

(e) Determinare il valore di $\omega \in (0, \infty)$ che minimizza il raggio spettrale $\rho(P)$. Quanto vale il raggio spettrale minimo? Qual è il valore di $\omega \in (0, \infty)$ per il quale il metodo menzionato al punto (b) converge più velocemente?

Soluzione.

(a) La matrice A è a diagonale dominante in senso stretto (sia per righe che per colonne) per cui i metodi di Jacobi e Gauss-Seidel applicati a un sistema lineare di matrice A sono entrambi convergenti.

(b) La matrice d'iterazione del metodo iterativo associato alla decomposizione $A = M - (M - A)$ è

$$P = M^{-1}(M-A) = \omega D^{-1}\Big(\frac{1}{\omega}D - A\Big) = \begin{bmatrix} \frac{\omega}{2} & 0 \\ 0 & -\frac{\omega}{2} \end{bmatrix} \begin{bmatrix} \frac{2}{\omega} - 2 & -1 \\ -1 & -\frac{2}{\omega} + 2 \end{bmatrix} = \begin{bmatrix} 1-\omega & -\frac{\omega}{2} \\ \frac{\omega}{2} & 1-\omega \end{bmatrix}.$$

Per calcolare $\rho(P)$, calcoliamo gli autovalori di P. Il polinomio caratteristico di P è dato da

$$C_P(\lambda) = \det(\lambda I - P) = \begin{vmatrix} \lambda - 1 + \omega & \frac{\omega}{2} \\ -\frac{\omega}{2} & \lambda - 1 + \omega \end{vmatrix} = (\lambda - 1 + \omega)^2 + \frac{\omega^2}{4}$$

$$= \lambda^2 + 2(-1+\omega)\lambda + (-1+\omega)^2 + \frac{\omega^2}{4}.$$

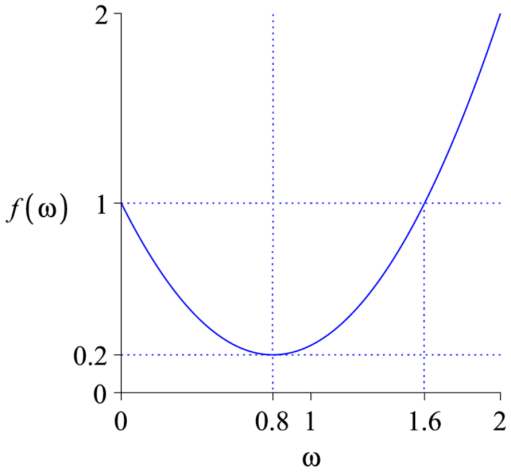

Figura 5.1: Grafico della funzione $f(\omega) = \rho(P)^2 = \dfrac{5\omega^2}{4} - 2\omega + 1$.

Gli autovalori di P sono le soluzioni dell'equazione $C_P(\lambda) = 0$ e quindi sono dati da

$$\lambda_{1,2} = 1 - \omega \pm \sqrt{(-1+\omega)^2 - (-1+\omega)^2 - \frac{\omega^2}{4}} = 1 - \omega \pm \frac{\mathrm{i}\,\omega}{2}.$$

Ricordando che il modulo di un numero complesso $a + \mathrm{i}b$ è $\sqrt{a^2+b^2}$, si ha

$$|\lambda_1| = |\lambda_2| = \sqrt{(1-\omega)^2 + \frac{\omega^2}{4}} = \sqrt{1 + \omega^2 - 2\omega + \frac{\omega^2}{4}} = \sqrt{\frac{5\omega^2}{4} - 2\omega + 1},$$

dunque

$$\rho(P) = \sqrt{\frac{5\omega^2}{4} - 2\omega + 1}.$$

(c) La funzione

$$f(\omega) = \rho(P)^2 = \frac{5\omega^2}{4} - 2\omega + 1$$

è una parabola. Si ha

$$f'(\omega) = \frac{5\omega}{2} - 2 \geq 0 \quad \Longleftrightarrow \quad \omega \geq \frac{4}{5},$$

per cui la funzione $f(\omega)$ è crescente per $\omega \geq \frac{4}{5}$ e decrescente per $\omega < \frac{4}{5}$. Abbiamo così trovato il vertice della parabola senza ricordarne la formula: il vertice V è il punto sulla parabola corrispondente al punto di minimo assoluto $\omega = \frac{4}{5}$ cioè $V = (\frac{4}{5}, f(\frac{4}{5})) = (\frac{4}{5}, \frac{1}{5})$. La parabola volge la concavità verso l'alto essendo positivo il coefficiente $\frac{5}{4}$ del termine di secondo grado ω^2. Inoltre, si ha $f(0) = 1$. Possiamo ora tracciare il grafico della parabola $f(\omega)$ per $\omega \in (0, \infty)$; si veda la Figura 5.1. I valori di $\omega \in (0, \infty)$ per cui si ha $f(\omega) < 1$ sono quelli in $(0, \frac{8}{5})$, come si vede dal grafico tenendo conto che la parabola è simmetrica rispetto al suo asse di simmetria (la retta verticale passante per il vertice). Alternativamente, risolvendo

$$f(\omega) < 1 \quad \Longleftrightarrow \quad \frac{5\omega^2}{4} - 2\omega + 1 < 1 \quad \Longleftrightarrow \quad \omega\Big(\frac{5\omega}{4} - 2\Big) < 0 \quad \Longleftrightarrow \quad 0 < \omega < \frac{8}{5},$$

si scopre ancora che i valori di $\omega \in (0, \infty)$ per cui si ha $f(\omega) < 1$ sono quelli in $(0, \frac{8}{5})$.

(d) Il metodo menzionato al punto (b) converge se e solo se

$$\rho(P) < 1 \quad \Longleftrightarrow \quad \rho(P)^2 < 1 \quad \Longleftrightarrow \quad f(\omega) < 1 \quad \Longleftrightarrow \quad 0 < \omega < \frac{8}{5}.$$

(e) Il valore di ω che minimizza $\rho(P)$ è lo stesso che minimizza $\rho(P)^2 = f(\omega)$ ed è $\omega = \frac{4}{5}$, come abbiamo già visto risolvendo il punto (c). Il raggio spettrale minimo è quindi quello che si ottiene per $\omega = \frac{4}{5}$, cioè

$$\rho(P)|_{\omega=\frac{4}{5}} = \sqrt{f\Big(\frac{4}{5}\Big)} = \sqrt{\frac{1}{5}} = 0.4472135...$$

Il valore di $\omega \in (0, \infty)$ per cui il metodo menzionato al punto (b) converge più velocemente è quello che minimizza il raggio spettrale $\rho(P)$, cioè $\omega = \frac{4}{5}$. □