Prova scritta di Calcolo Scientifico

`Udine, 22 giugno 2021`

1. Sia $\mathcal{F} := \mathcal{F}(2, t, e_{\max}, e_{\min})$ l'insieme di numeri di macchina con l'arrotondamento.
   - Determina gli interi $t, e_{\max}, e_{\min}$ in modo che $e_{\max} = e_{\min}$, la precisione di macchina u sia $1/32$ e $realmax/realmin = 31$.
   - Siano dati $x = (1.\overline{0111})_2$ e $y = (10.\overline{0111})_2$. Determina $\tilde{x} = fl(x) \in \mathcal{F}$, $\tilde{y} = fl(y) \in \mathcal{F}$ e $\tilde{z} = \tilde{x} fl(+) \tilde{y}$.
   - ⋆ Scrivi x, y e $\tilde{x}, \tilde{y}$ come frazioni di numeri interi in base 10.
   - Definisci i numeri denormalizzati per $\mathcal{F}$ e determina il numero denormalizzato positivo più piccolo. Giustifica la risposta.
2. Si vuole calcolare la funzione $y = f(x)$.
   - Sia $f(x) = \sqrt{g(x)}$, con g funzioni reale non negativa. Determina la relazione tra il numero di condizionamento di f e quello di g. Studia il condizionamento della funzione $f(x) = \sqrt{\frac{x^2-2}{x+1}}$ con x che varia nel campo di esistenza di f.
   - Sia $p(x) = \sum_{k=0}^{n} x^k$, con $n = 2$ e x numero di macchina. Per calcolare $p(x)$ usa l'algoritmo di Horner e studia la stabilità.
   - Scrivi la pseudocodifica dell'algoritmo di Horner per $p(x) = \sum_{k=0}^{n} a_k x^k$, con n intero qualsiasi. Analizza la sua complessità computazionale.
3. Sia $f(x) = x^4 - 11x^2 + 18x - 8$.
   - Disegna il grafico di f. Determina le radici α, β, γ con $\alpha < \beta < \gamma$ (Suggerimento: valuta $f(1)$).
   - Studia la convergenza del metodo di Newton ad α e a γ.
   - ⋆ Siano $z_1, z_2, z_1 < z_2$, i due punti di flesso della funzione f. Studia la convergenza del metodo di Newton a β quando $x_0 \in (z_1, z_2)$
   - Considera le successioni ottenute con il metodo di Newton con i seguenti valori iniziali

     (a) $x_0 = -5$
     (b) $x_0 = -3$
     (c) $x_0 = 0$
     (d) $x_0 = 3$

     Sono convergenti? Se convergenti, convergono ad α, β o a γ? Qual è l'ordine di convergenza? Giustifica tutte le risposte.
   - Sia $g(x) = x - \frac{f(x)}{m}$. Considera il metodo iterativo $x_{k+1} = g(x_k), k = 0, 1, \ldots$. Determina m in modo che il metodo sia localmente convergente in maniera monotona a α con ordine di convergenza quadratico. La successione ottenuta con $x_0 = -3$ è convergente? Giustifica la risposta.
   - Studia la convergenza locale a γ del metodo iterativo al punto precedente con $m = 120$. La successione ottenuta con $x_0 = 3$ è convergente? Se convergente, qual è l'ordine di convergenza? Giustifica la risposta.
4. Sia data la matrice
$$A = \begin{pmatrix} -\alpha & -9 & \alpha \\ 2 & 3 & 4 \\ 2\alpha & 13 & 2\alpha \end{pmatrix}.$$
   - Calcola la fattorizzazione LU di A. Per quale scelta del parametri α esiste tale fattorizzazione?
   - Studia al variare di α il comportamento del metodo di Gauss con il pivot parziale al primo passo.
   - Sia $\alpha = 4$. Calcola la fattorizzazione $PA = LU$ con la tecnica del pivot parziale.
5. Sia $f(x) = \log_2(1 + 4x^2)$. Dati i punti $P_0 = (-1/2, f(-1/2))$, $P_1 = (0, f(0))$, $P_2 = (1/2, f(1/2))$.
   - Determina il polinomio p che interpola i tre punti nella forma di Newton.
   - Determina il polinomio $\tilde{p}$ che interpola i tre punti e tale che $\tilde{p}'(0) = f'(0)$ nella forma di Newton.
   - Determina il polinomio q di primo grado di miglior approssimazione dei quattro punti P_0, P_1, P_2 e $P_3 = (\sqrt{3}/2, f(\sqrt{3}/2))$ nel senso dei minimi quadrati.

⋆ Scrivi la pseudocodifica dell'algoritmo di eliminazione di Gauss di base. Modificala per applicare la tecnica del pivot parziale.

Esempio di svolgimento della prova scritta di Calcolo scientifico del 22 giugno 2021

Rossana Vermiglio

30 giugno 2021

Esercizio 1.

- Osserviamo che la base è 2 e si usa l'arrotondamento, perciò si ha

$$u = \frac{2^{1-t}}{2} = 2^{-t} = \frac{1}{32} = 2^{-5} \quad \Leftrightarrow \quad t = 5.$$

Inoltre, poiché $e_{\max} = e_{\min} =: e$, si hanno realmin $= 2^{-e_{\min}-1} = 2^{-e-1}$ e realmax $= 2^{e_{\max}}(1 - 2^{-t}) = 2^e(1-2^{-t})$, quindi

$$\frac{\text{realmax}}{\text{realmin}} = \frac{2^e(1-2^{-5})}{2^{-e-1}} = 31 \quad \Leftrightarrow \quad 2^{2e+1} \cdot \frac{31}{32} = 31 \quad \Leftrightarrow \quad 2^{2e+1} = 32 = 2^5$$
$$\Leftrightarrow \quad 2e+1 = 5 \quad \Leftrightarrow \quad e = e_{\max} = e_{\min} = 2.$$

- Ricordando che si usa l'arrotondamento, si hanno

$$\tilde{x} = \text{fl}((1.\overline{0111})_2) = \text{fl}((0.1\overline{0111})_2 \cdot 2) = \text{fl}((0.10111\overline{0111})_2 \cdot 2) = (0.10111)_2 \cdot 2,$$
$$\tilde{y} = \text{fl}((10.\overline{0111})_2) = \text{fl}((0.10\overline{0111})_2 \cdot 2^2) = (0.10100)_2 \cdot 2^2.$$

Inoltre $\tilde{x} + \tilde{y} = (0.10100)_2 \cdot 2^2 + (0.010111)_2 \cdot 2^2 = (0.111111)_2 \cdot 2^2$, da cui $\tilde{z} = \text{fl}(\tilde{x}+\tilde{y}) = (0.10000)_2 \cdot 2^3$. Poiché l'esponente 3 è maggiore di $e_{\max} = 2$, si ha un overflow e $\tilde{z} \notin \mathcal{F}$.

⋆ Si hanno

$$\tilde{x} = 2\left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}\right) = \frac{23}{16}, \qquad \tilde{y} = 2^2\left(\frac{1}{2} + \frac{1}{8}\right) = \frac{5}{2}.$$

Possiamo scrivere $x = 1 + (0.\overline{0111})_2$, $y = 2 + (0.\overline{0111})_2$ e

$$\begin{aligned}(0.\overline{0111})_2 &= (0.0111\,0111\,0111\ldots)_2\\ &= (0.0111)_2 + (0.0000\,0111)_2 + (0.0000\,0000\,0111)_2 + \cdots\\ &= 2^0 \cdot (0.0111)_2 + 2^{-4} \cdot (0.0111)_2 + 2^{-8} \cdot (0.0111)_2 + \cdots.\end{aligned}$$

Osserviamo che $(0.0111)_2 = \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{7}{16}$. Allora

$$(0.\overline{0111})_2 = \frac{7}{16} + \sum_{k=0}^{+\infty}\left(\frac{1}{2^4}\right)^k = \frac{7}{16} \cdot \frac{1}{1-1/2^4} = \frac{7}{16} \cdot \frac{16}{15} = \frac{7}{15},$$

da cui $x = 1 + \frac{7}{15} = \frac{22}{15}$ e $y = 2 + \frac{7}{15} = \frac{37}{15}$.

- I numeri denormalizzati hanno la forma $\pm 2^{-2}(0.0\,d_2\,d_3\,d_4\,d_5)$ con le cifre d_i non tutte nulle. Il più piccolo numero denormalizzato postivo è quindi $+2^{-2}(0.00001) = 2^{-2} \cdot 2^{-5} = 2^{-7}$.

CD Lab – Computational Dynamics Laboratory, Dipartimento di Scienze Matematiche, Informatiche e Fisiche, Università degli Studi di Udine, rossana.vermiglio@uniud.it

Esercizio 2.

- Sia $f(x) = \sqrt{g(x)}$. Si hanno

$$\operatorname{cond}_f(x) = \left|\frac{xf'(x)}{f(x)}\right|, \quad \operatorname{cond}_g(x) = \left|\frac{xg'(x)}{g(x)}\right|, \quad f'(x) = \frac{1}{2}\frac{g'(x)}{\sqrt{g(x)}},$$

da cui

$$\operatorname{cond}_f(x) = \left|\frac{x}{2}\frac{g'(x)}{\sqrt{g(x)}}\frac{1}{\sqrt{g(x)}}\right| = \frac{1}{2}\operatorname{cond}_g(x).$$

Se $g(x) = \frac{x^2-2}{x+1}$, allora

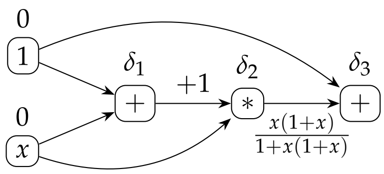
$$g'(x) = \frac{2x(x+1)-(x^2-2)}{(x+1)^2} = \frac{2x^2+2x-x^2+2)}{(x+1)^2} = \frac{x^2+2x+2)}{(x+1)^2},$$

da cui

$$\operatorname{cond}_g(x) = \left|x \cdot \frac{x^2+2x+2)}{(x+1)^2} \cdot \frac{x+1}{x^2-2}\right| = \left|\frac{x(x^2+2x+2))}{(x+1)(x^2-2)}\right|.$$

I limiti per $x \to -1$ e $x \to \pm\sqrt{2}$ sono infiniti, quindi $\operatorname{cond}_g x \gg 1$ per $x \approx -1$ e $x \approx \pm\sqrt{2}$.

- Riscriviamo il polinomio nella forma di Horner: $p(x) = 1 + x + x^2 = (1+x)x + 1$. Ricordiamo i coefficienti di amplificazione della somma e del prodotto. Costruiamo il grafo computazionale.



Per l'errore algoritmico risulta quindi

$$\epsilon_{\text{alg}} = \delta_3 + \frac{x(1+x)}{1+x(1+x)}(\delta_2+\delta_1),$$

da cui, poiché gli errori sulle operazioni sono maggiorati in valore assoluto dalla precisione di macchina,

$$|\epsilon_{\text{alg}}| \le |\delta_3| + \left|\frac{x(1+x)}{1+x(1+x)}\right|(|\delta_2|+|\delta_1|) \le u\left(1+2\left|\frac{x(1+x)}{1+x(1+x)}\right|\right) \le 3u.$$

Infatti, la funzione $x(1+x)$ ha il minimo $-\frac{1}{4}$ per $x = -\frac{1}{2}$, mentre la funzione $\frac{z}{1+z}$ è strettamente crescente per $z > 0$, ha un asintoto orizzontale di ordinata 1 e vale $-\frac{1}{3}$ per $z = -\frac{1}{4}$ (vedi Figura 1); componendo le due funzioni possiamo concludere che la frazione in valore assoluto ha valore minore di 1. L'algoritmo risulta quindi stabile.

Esercizio 3.

- Osserviamo che $f(1) = 0$. Dividendo $f(x)$ per $x-1$ con la regola di Ruffini,

	1	0	-11	18	-8
1		1	1	-10	8
	1	1	-10	8	0

otteniamo $f(x) = (x-1)(x^3+x^2-10x+8)$. Osserviamo che anche il secondo fattore si annulla in 1. Possiamo quindi dividerlo per $x-1$,

	1	1	-10	8
1		1	2	-8
	1	2	-8	0

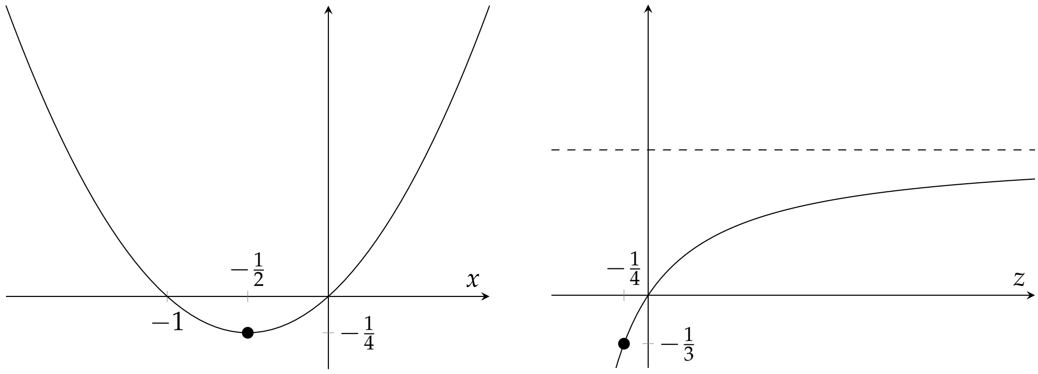


Figura 1: Grafici di $x(1+x)$ e $\frac{z}{1+z}$.

ottenendo $f(x) = (x-1)^2(x^2+2x-8)$. Le radici del fattore di grado 2 sono $-1 \pm \sqrt{1+8}$, ovvero 2 e -4. Si hanno quindi $\alpha = -4$ radice semplice, $\beta = 1$ radice doppia e $\gamma = 2$ radice semplice.

Per tracciare il grafico procediamo allo studio della funzione f. Essendo un polinomio di grado 4 con coefficiente direttivo positivo, i limiti all'infinito sono entrambi $+\infty$. La derivata prima è $f'(x) = 4x^3 - 22x + 18$, che si annulla in 1, essendo questa una radice doppia di f. Dividendo $f'(x)$ per $x-1$,

	4	0	-22	18
1		4	4	-18
	4	4	-18	0

otteniamo $f'(x) = (x-1)(4x^2+4x-18) = 2(x-1)(2x^2+2x-9)$. Le radici del fattore di grado 2 sono $\frac{-1\pm\sqrt{1+18}}{2}$. La derivata seconda è $f''(x) = 12x^2 - 22$, che si annulla in $\pm\sqrt{\frac{11}{6}}$, che sono i punti di flesso $z_1 < z_2$. Un grafico approssimato è rappresentato in Figura 2.

- Per ogni $x < -4$ si hanno $f(x)f''(x) > 0$ e $f'(x) \neq 0$: per il teorema visto a lezione il metodo di Newton converge a α in modo monotono crescente con ordine $p = 2$ per ogni punto iniziale $x_0 < -4$. Dal significato geometrico del metodo di Newton osserviamo che esso converge a α con ordine $p = 2$ in modo monotono crescente dopo la prima iterata anche per ogni $x_0 \in (-4, \frac{-1-\sqrt{19}}{2})$.

  Analogamente, il metodo converge a γ in modo monotono decrescente con ordine $p = 2$ per ogni $x_0 > 2$ e converge a γ con ordine $p = 2$ in modo monotono decrescente dopo la prima iterata anche per ogni $x_0 \in (\frac{-1+\sqrt{19}}{2}, 2)$.

⋆ Per ogni $x \in (z_1, z_2)$ con $x \neq 1$ si hanno $f(x)f''(x) > 0$ e $f'(x) \neq 0$: per il teorema visto a lezione il metodo di Newton converge a β in modo monotono con ordine $p = 1$ e fattore asintotico di riduzione dell'errore $l = \frac{1}{2}$ per ogni punto iniziale $x_0 \in (z_1, z_2)$ con $x_0 \neq 1$.

- Osserviamo che $g'(x) = 1 - \frac{f'(x)}{m}$. Poiché il fattore asintotico di riduzione dell'errore è $l = |g'(x)|$, perché il metodo abbia ordine di convergenza quadratico deve essere $|g'(\alpha)| = 1 - \frac{f'(\alpha)}{m} = 0$. Sostituendo $f'(\alpha) = f'(-4) = -150$ si ottiene $m = -150$.

  Sia $M_1 := \frac{-1+\sqrt{19}}{2}$; osserviamo che $-3 \in (-4, M_1)$. La funzione $g'(x)$ è strettamente crescente in $[-4, M_1]$: infatti $g''(x) = \frac{f''(x)}{150}$ e $f''(x) > 0$ per $x \in [-4, M_1]$. Poiché $f'(M_1) = 0$ si ha $g'(M_1) = 1$, per cui $g'(x) \in [0, 1)$ per ogni $x \in [-4, -3]$. Ricordiamo la relazione tra gli errori di iterate successive

  $$x_{n+1} - \alpha = g'(\xi_n)(x_n - \alpha),$$

  dove ξ_n è compreso tra x_n e α. Allora la successione ottenuta con il punto iniziale $x_0 = -3$ converge a α.

- Con $m = 120$ il fattore asintotico di riduzione dell'errore è $l = |g'(\gamma)| = 1 - \frac{f'(\gamma)}{120} = \frac{19}{20}$, in quanto $f'(\gamma) = f'(2) = 6$.

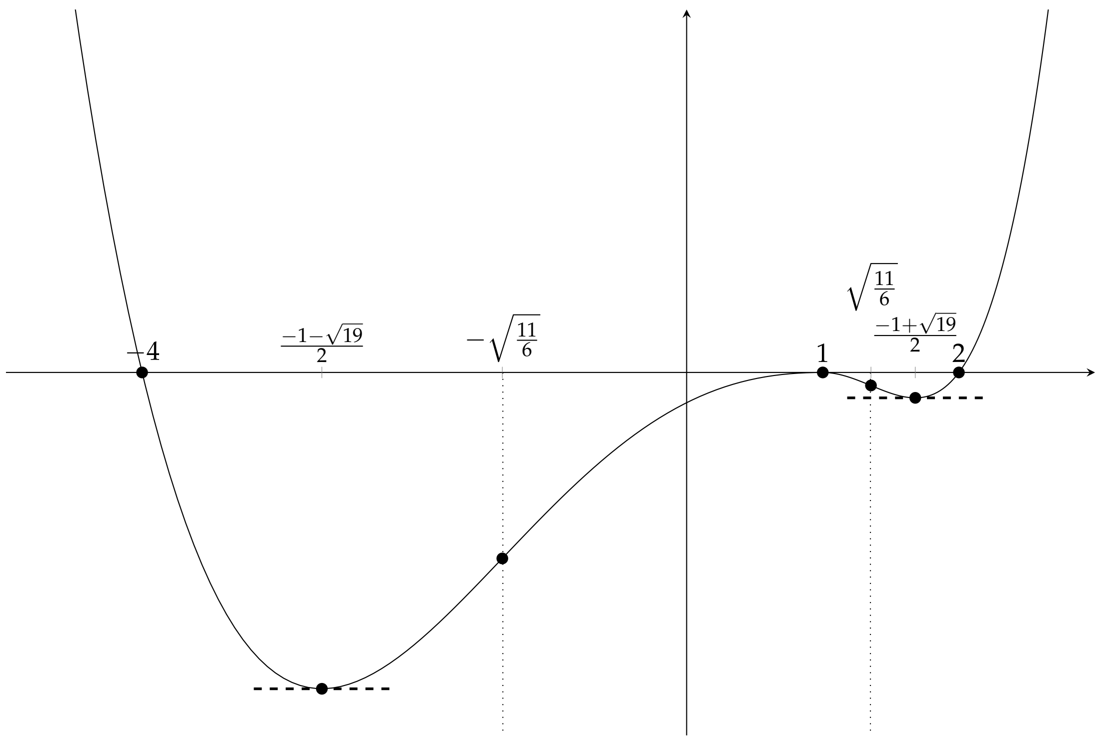


Figura 2: Grafico di $f(x) = x^4 - 11x^2 + 18x - 8$. Le linee tratteggiate sono tangenti al grafico, quelle puntinate indicano i cambi di concavità. Per ragioni grafiche, le scale verticali a sinistra e a destra di $x = 1$ sono diverse.

La funzione $g'(x)$ è strettamente decrescente in $[2, 3]$: infatti $g''(x) = -\frac{f''(x)}{120}$ e $f''(x) > 0$ per $x \in [2, 3]$. Si ha $g'(x) \in [0, 1)$ per ogni $x \in [2, 3]$, poiché $f'(3) = 60$ e quindi $g'(3) = \frac{1}{2}$. Analogamente al punto precedente si conclude che la successione ottenuta con il punto iniziale $x_0 = 3$ converge a γ.

Esercizio 4.

- Applichiamo l'algoritmo di eliminazione di Gauss alla matrice

$$A = \begin{pmatrix} -\alpha & -9 & \alpha \\ 2 & 3 & 4 \\ 2\alpha & 13 & 2\alpha \end{pmatrix}.$$

La matrice elementare di Gauss al primo passo si può definire solo se $\alpha \neq 0$:

$$G_1 = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{\alpha} & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \qquad G_1 A = \begin{pmatrix} -\alpha & -9 & \alpha \\ 0 & -\frac{18}{\alpha} + 3 & 6 \\ 0 & -5 & 4\alpha \end{pmatrix} = \begin{pmatrix} -\alpha & -9 & \alpha \\ 0 & \frac{3(\alpha-6)}{\alpha} & 6 \\ 0 & -5 & 4\alpha \end{pmatrix}.$$

La matrice elementare di Gauss al secondo passo si può definire solo se $\alpha \neq 6$:

$$G_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{5\alpha}{3(\alpha-6)} & 1 \end{pmatrix}.$$

Perciò per $\alpha \in \mathbb{R} \setminus \{0, 6\}$ la fattorizzazione LU esiste ed è

$$L = G_1^{-1} G_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{2}{\alpha} & 1 & 0 \\ -2 & \frac{-5\alpha}{3(\alpha-6)} & 1 \end{pmatrix},$$

$$U = G_2 G_1 A = \begin{pmatrix} -\alpha & -9 & \alpha \\ 0 & -\frac{18}{\alpha} + 3 & 6 \\ 0 & 0 & \frac{10\alpha}{\alpha-6} + 4\alpha \end{pmatrix} = \begin{pmatrix} -\alpha & -9 & \alpha \\ 0 & -\frac{18}{\alpha} + 3 & 6 \\ 0 & 0 & \frac{4\alpha^2 - 14\alpha}{\alpha-6} \end{pmatrix}.$$

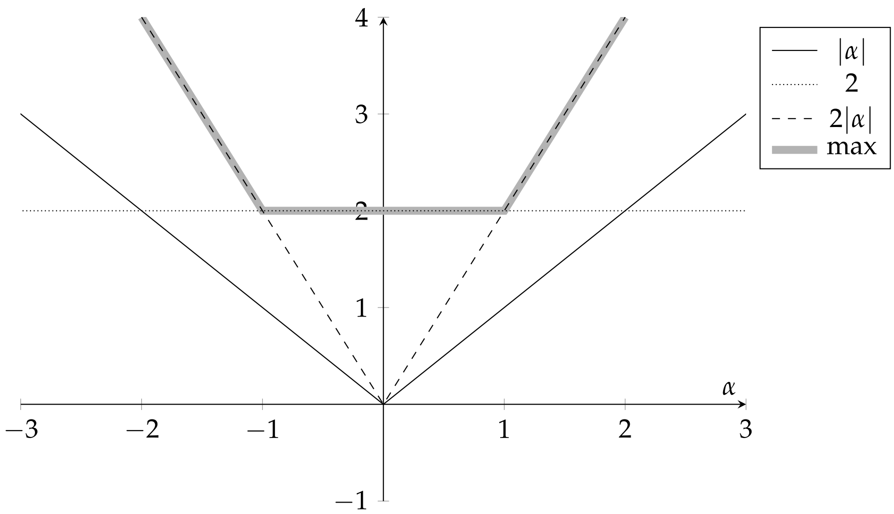


Figura 3: Studio grafico di $\max\{|\alpha|, 2, 2|\alpha|\}$.

- Il comportamento del metodo di Gauss con il pivot parziale al primo passo dipende da quale valore della prima colonna di A è massimo in valore assoluto. Studiamo pertanto $\max\{|\alpha|, 2, 2|\alpha|\}$. Per quanto osserviamo in Figura 3, al primo passo si scambiano la prima e la seconda riga se $|\alpha| \leq 1$ e si scambiano la prima e la terza se $\alpha > 1$.

- Se $\alpha = 4$, la matrice A è
$$A = \begin{pmatrix} -4 & -9 & 4 \\ 2 & 3 & 4 \\ 8 & 13 & 8 \end{pmatrix}.$$
Al primo passo bisogna scambiare la prima e la terza riga, per cui
$$P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad P_1 A = \begin{pmatrix} 8 & 13 & 8 \\ 2 & 3 & 4 \\ -4 & -9 & 4 \end{pmatrix},$$
$$G_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}, \qquad G_1 P_1 A = \begin{pmatrix} 8 & 13 & 8 \\ 0 & -\frac{1}{4} & 2 \\ 0 & -\frac{5}{2} & 8 \end{pmatrix}.$$
Al secondo passo bisogna scambiare la seconda e la terza riga, per cui
$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad P_2 G_1 P_1 A = \begin{pmatrix} 8 & 13 & 8 \\ 0 & -\frac{5}{2} & 8 \\ 0 & -\frac{1}{4} & 2 \end{pmatrix},$$
$$G_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{10} & 1 \end{pmatrix}, \qquad U = G_2 P_2 G_1 P_1 A = \begin{pmatrix} 8 & 13 & 8 \\ 0 & -\frac{5}{2} & 8 \\ 0 & 0 & -\frac{6}{5} \end{pmatrix},$$
$$P = P_2 P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$
$$L = P_2 P_1 P_1^{-1} G_1^{-1} P_2^{-1} G_2^{-1} = P_2 G_1^{-1} P_2 G_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{4} & \frac{1}{10} & 1 \end{pmatrix}.$$

Esercizio 5. Calcoliamo le coordinate dei punti $P_0 = (-1/2, 1)$, $P_1 = (0, 0)$, $P_2 = (1/2, 1)$.

- Calcoliamo le differenze finite.

$$
\begin{array}{lcccc}
 & t_i & f[t_i] & f[t_{i-1}, t_i] & f[t_{i-2}, t_{i-1}, t_i] \\
i=0 & -\frac{1}{2} & 1 & & \\
i=1 & 0 & 0 & \frac{1-0}{-1/2-0} = -2 & \\
i=2 & \frac{1}{2} & 1 & 2 & 4
\end{array}
$$

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

- Calcoliamo le differenze finite imponendo il valore della derivata $\tilde{p}'(0) = f'(0) = 0$.

$$
\begin{array}{ccccc}
-\frac{1}{2} & 1 & & & \\
0 & 0 & -2 & & \\
0 & 0 & 0 & 4 & \\
\frac{1}{2} & 1 & 2 & 4 & 0
\end{array}
$$

Si ha quindi $\tilde{p} = p$.

- Calcoliamo le coordinate del punto $P_3 = (\sqrt{3}/2, 2)$. Per calcolare i coefficienti di $q(x) = a + bx$ risolviamo il sistema

$$
\begin{pmatrix} 1 & 1 & 1 & 1 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}
\begin{pmatrix} 1 & -\frac{1}{2} \\ 1 & 0 \\ 1 & \frac{1}{2} \\ 1 & \frac{\sqrt{3}}{2} \end{pmatrix}
\begin{pmatrix} a \\ b \end{pmatrix} =
\begin{pmatrix} 1 & 1 & 1 & 1 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}
\begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \end{pmatrix},
$$

ovvero

$$
\begin{pmatrix} 4 & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{5}{4} \end{pmatrix}
\begin{pmatrix} a \\ b \end{pmatrix} =
\begin{pmatrix} 4 \\ \sqrt{3} \end{pmatrix},
$$

da cui $a = \frac{14}{17}$ e $b = \frac{8\sqrt{3}}{17}$.