Riccardo Cazzin

Appunti di

Calcolo Numerico

Nuova Edizione

estratti dalle lezioni del

Prof. Marco Vianello

Anno Accademico 2019-2020 Università degli Studî di Padova

Ringraziamenti

Questo tentativo di collezionare una buona parte dei temi che il professor Vianello ha esposto nelle sue lezioni non avrebbe potuto essere svolto da una sola persona: troppi gli argomenti, troppi i dettagli importanti per non chiedere aiuto.

Voglio ringraziare, dunque, quanti hanno dato sostegno ad un'impresa cosí ardita come utile a chi voglia approcciarsi al *mare magnum* del Calcolo Numerico (e a passarne l'esame). Ringrazio in particolare Alena Meyer, Adriano Prade, Erika Rampazzo, Emanuele Ronda e Ilenia Zippo per avermi permesso di integrare con i loro appunti quanto avevo raccolto personalmente. Ringrazio, poi, Bryan Harback, che prima di me tentò un'impresa simile, compiendo un ottimo lavoro di sintesi di tutti gli argomenti, e che per questo considero quasi un mentore.

Nella speranza che lo studio e l'esame vi siano lievi e pure piacevoli, vi auguro una buona lettura.

Riccardo Cazzin

Indice

1 Sistema a virgola mobile e propagazione degli errori

1.1 Rappresentazione dei numeri reali; errore di troncamento e arrotondamento

Fissato un numero naturale $b > 1$, chiamato *base*, ogni numero reale $x \in \mathbb{R}$ può essere espresso con la scrittura

$$x = \operatorname{sgn}(x) \left(\sum_{j=0}^{m} c_j b^j + \sum_{j=1}^{\infty} c_{-j} b^{-j} \right)$$

con $c_i \in \{ 0, \dots, b-1 \}$ per ogni i; tale scrittura prende il nome di *rappresentazione a virgola fissa*. La somma $\sum_{j=0}^{m} c_j b^j$ si dice *parte intera* di x, mentre la serie $\sum_{j=1}^{\infty} c_{-j} b^{-j}$ si dice *parte frazionaria* di x.

Perché tale scrittura abbia senso, è necessario che la serie della parte frazionaria converga. Consideriamo preliminarmente la serie $S = \sum_{k=0}^{\infty} \alpha^k$, con $\alpha \in \mathbb{C}$, come limite delle somme parziali $S_n = \sum_{k=0}^{n} \alpha^k$: dal momento che $S_n = n + 1$ per $\alpha = 1$, la serie diverge a ∞ in questo caso; per $\alpha \neq 1$ si ha che $S_n = (1 - \alpha^{n+1})/(1 - \alpha)$: se $|\alpha| < 1$, S_n converge a $1/(1-\alpha)$, se $|\alpha| > 1$, S_n diverge a ∞; se $|\alpha| = 1$ e $\alpha \neq 1$, invece, la successione delle somme parziali S_n non ha limite.

Tornando alla parte frazionaria di x, si verifica che

$$\sum_{j=1}^{\infty} c_{-j} b^{-j} \leq \sum_{j=1}^{\infty} (b-1) b^{-j} = (b-1) \sum_{j=1}^{\infty} b^{-j} = \\ = (b-1) \left(\frac{1}{1 - \frac{1}{b}} - 1 \right) = (b-1) \frac{b - b + 1}{b - 1} = 1$$

da cui segue che tale serie converge ad una quantità nell'intervallo $[\, 0, 1 \,]$ — la chiusura a destra dell'intervallo si verifica proprio nel caso in cui $c_{-j} = b - 1$ per ogni j, come mostrato sopra.

La rappresentazione della parte frazionaria di un numero razionale può essere finita oppure infinita, a seconda della base numerica scelta. Si consideri, ad esempio, la rappresentazione in base 10 della frazione 1/3:

$$\frac{1}{3} = \frac{3}{9} = 3 \left(\frac{1}{1 - \frac{1}{10}} - 1 \right) = 3 \sum_{j=1}^{\infty} 10^{-j} = (0.\overline{3})_{10}$$

scegliendo la base 3, invece, si ha direttamente $1/3 = 1 \cdot 3^{-1} = (0.1)_3$. I numeri irrazionali, per contro, non ammettono rappresentazione finita in alcuna base: se, per assurdo, esistesse una base b tale che un numero irrazionale x abbia rappresentazione finita, esisterebbe $\ell \in \mathbb{N}$ tale che

$$x = \operatorname{sgn} x \left(\sum_{j=0}^{m} c_j b^j + \sum_{j=1}^{\ell} c_{-j} b^{-j} \right) = \operatorname{sgn} x \left(\sum_{j=0}^{m} c_j b^j + \frac{(c_{-1} \dots c_{-\ell})_b}{b^\ell} \right)$$

e ciò è assurdo, perché la parte rappresentata nella parentesi è una somma tra un numero intero ed un numero razionale, che non è irrazionale.

Fissato $n \in \mathbb{N}$, si può definire il *troncamento a n cifre* di un numero $x \in \mathbb{R}$ come

$$x_n \coloneqq \operatorname{sgn} x \left(\sum_{j=0}^{m} c_j b^j + \sum_{j=1}^{n} c_{-j} b^{-j} \right)$$

da cui segue che l'*errore* commesso in funzione di n è stimato da

$$0 \leq |x - x_n| = \sum_{j=n+1}^{\infty} c_{-j} b^{-j} \leq (b-1) \sum_{j=n+1}^{\infty} b^{-j} =$$

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

Perché l'errore commesso a causa del troncamento sia inferiore ad un certo $\varepsilon > 0$, dunque, occorre che $n > -\log_b \varepsilon$.

Fissato $n \in \mathbb{N}$ e scelta una base b pari, si può anche definire l'*arrotondamento a n cifre* di un numero $x \in \mathbb{R}$ come

$$\tilde{x}_n \coloneqq \operatorname{sgn} x \left(\sum_{j=0}^{m} c_j b^j + \sum_{j=1}^{n-1} c_{-j} b^{-j} + \tilde{c}_{-n} b^{-n} \right)$$

ove $\tilde{c}_{-n} = c_{-n}$ se $c_{-n-1} < b/2$ e $\tilde{c}_{-n} = c_{-n} + 1$ se $c_{-n-1} \geq b/2$ — eseguendo il riporto se $c_{-n} = b - 1$. L'*errore* commesso da un arrotondamento si può stimare distinguendo i due casi di arrotondamento: nel caso dell'arrotondamento per difetto si ha

$$0 \leq |x - \tilde{x}_n| = \sum_{j=n+1}^{\infty} c_{-j} b^{-j} =$$

$$= c_{-(n+1)} b^{-(n+1)} + \sum_{j=n+2}^{\infty} c_{-j} b^{-j} \leq \frac{b-1}{2} b^{-(n+1)} + (b-1) \sum_{j=n+2}^{\infty} b^{-j} =$$

$$= \frac{b-1}{2} b^{-(n+1)} + (b-1) \left(\frac{1}{1 - 1/b} - \frac{1 - b^{-(n+2)}}{1 - 1/b} \right) =$$

$$= \frac{b-1}{2} b^{-(n+1)} + b^{-(n+1)} = \frac{b^{-n}}{2}$$

e nel caso dell'arrotondamento per eccesso si ha

$$0 \leq |x - \tilde{x}_n| = b^{-n} - \sum_{j=n+1}^{\infty} c_{-j} b^{-j} =$$

$$= b^{-n} - c_{-(n+1)} b^{-(n+1)} - \sum_{j=n+2}^{\infty} c_{-j} b^{-j} \leq b^{-n} - \frac{b^{-n}}{2} = \frac{b^{-n}}{2}$$

ove la maggiorazione è giustificata dalla rimozione della serie (non sottrarre una quantità positiva maggiora l'espressione in esame) e dalla relazione $c_{-(n+1)} \geq b/2$. A livello grafico-geometrico, ciò si può visualizzare come nella Figura 1.1: fissato $n \in \mathbb{N}$, infatti, ogni numero reale $y \in \mathbb{R}$ è contenuto in uno ed un solo intorno di forma $\left[x - b^{-n}/2,\, x + b^{-n}/2 \right)$, con $x \in \mathbb{R}$ tale che $x = \tilde{x}_n$; l'arrotondamento comporta che $\tilde{y}_n = x$.

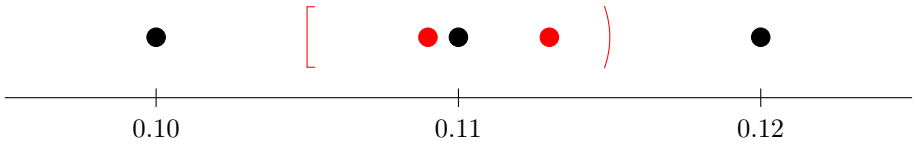

Figura 1.1: Con un arrotondamento a due cifre decimali in base dieci, entrambi i numeri reali 0.109 e 0.113 (punti in rosso) vengono arrotondati a 0.11.

1.2 Sistema *floating-point*

Fissata una base b, ogni numero reale $x \in \mathbb{R}$ può essere espresso con la scrittura

$$x = \operatorname{sgn} x \left(\sum_{j=1}^{\infty} d_j b^{-j} \right) b^p$$

ove $d_j \in \{\, 0, \dots, b-1 \,\}$ per ogni j, ma $d_1 \neq 0$, e $p \in \mathbb{Z}$; tale scrittura prende il nome di *rappresentazione a virgola mobile* o *rappresentazione* floating-point. La serie $\sum_{j=1}^{\infty} d_j b^{-j}$ si dice *mantissa* di x, mentre p si dice *esponente* di x. La condizione $d_1 \neq 0$ è necessaria affinché la scrittura di x in virgola mobile sia unica: se non la si ponesse, infatti, sarebbe possibile aggiungere tante cifre 0 alla mantissa "da sinistra" cambiando p in modo adeguato, e cosí facendo la scrittura di x non sarebbe unica. Scelta la base dieci, sono vere le seguenti uguaglianze:

$$1278.4351\ldots = (0.12784351\ldots) \cdot 10^4$$
$$-0.0003267\ldots = -(0.3267\ldots) \cdot 10^{-3}$$

La mantissa di un numero reale x è un elemento dell'intervallo $[\,0, 1\,]$: si ha, infatti, che

$$0 \leq \sum_{j=1}^{\infty} d_j b^{-j} \leq (b-1) \sum_{j=1}^{\infty} b^{-j} = (b-1) \left(\frac{1}{1 - \frac{1}{b}} - 1 \right) = 1$$

La mantissa di un certo $x \in \mathbb{R}$, tuttavia, non coincide in generale con la parte frazionaria di tale x; si consideri, ad esempio, il numero 102: la sua mantissa è 0.102, in quanto $102 = 0.102 \cdot 10^3$, ma la sua parte frazionaria è nulla, visto che è un numero intero.

Come già dimostrato sopra, i numeri irrazionali hanno parte frazionaria infinita: da ciò segue che anche la loro mantissa è infinita. Se un numero irrazionale avesse mantissa finita $0.d_1 \ldots d_t$, infatti, esso sarebbe il prodotto di una potenza dell'intero b per il numero razionale rapprresentato da $0.d_1 \ldots d_t \in [\,0, 1\,)$, ossia sarebbe un numero razionale — il che è assurdo.

A partire da questa scrittura dei numeri reali si può definire l'*arrotondamento a t cifre di mantissa* di un qualunque $x \in \mathbb{R}$

$$\mathrm{fl}^t(x) \coloneqq \operatorname{sgn} x \, (0.d_1 \ldots \tilde{d}_t) b^p$$

ove $\tilde{d}_t$ è ottenuta per arrotondamento. L'errore commesso in funzione di t è stimato da

$$0 \leq \left| x - \mathrm{fl}^t(x) \right| = b^p \left| 0.d_1 \ldots d_t - 0.d_1 \ldots \tilde{d}_t \right| \leq b^p \frac{b^{-t}}{2} = \frac{b^{p-t}}{2}$$

Questo tipo di approssimazione dei numeri reali porta a definire un nuovo insieme, l'insieme dei *reali macchina*

$$\mathbb{F}(b,t,L,U) \coloneqq \left\{ \mu = \pm(0.\mu_1 \dots \mu_t)_b \, b^p \;\middle|\; \begin{array}{l} \mu_1 \neq 0 \\ \mu_1, \dots, \mu_t \in \{\, 0, \dots, b-1 \,\} \\ p \in [\, L, U \,] \cap \mathbb{Z} \end{array} \right\} \cup \{\, 0 \,\}$$

Per convenzione, con $\mathbb{F}^+$ e $\mathbb{F}^-$ si intende rispettivamente la restrizione di $\mathbb{F}(b,t,L,U)$ ai numeri positivi o negativi, se gli altri parametri si possono evincere dal contesto. Segue uno studio schematico di tale insieme.

- Si ha $\operatorname{Card} \mathbb{F}(b,t,L,U) = 1 + 2(U-L+1)(b-1)b^{t-1}$: ragionando in modo combinatorio, infatti, si nota che si possono scegliere $(b-1)b^{t-1}$ mantisse, ricordando che $d_1 \neq 0$, $U-L+1$ esponenti e due segni; a ciò va aggiunto 0, che si conta a parte.

- Si ha $\min \mathbb{F}^+(b,t,L,U) = b^{L-1}$: scegliendo, infatti, la mantissa piú piccola possibile, ossia $0.10\dots0$, e l'esponente minore, cioè L, si ottiene che $\min \mathbb{F}^+ = (0.10\dots0)b^L = b^{L-1}$.

- Si ha $\max \mathbb{F}^+(b,t,L,U) = b^U(1-b^{-t})$: poiché la mantissa piú grande possibile è data da

$$(b-1)\sum_{j=1}^{t} b^{-j} = (b-1)\left(\frac{1-b^{-(t+1)}}{1-1/b} - 1\right) =$$

$$= (b-1)\frac{\left(1-b^{-(t+1)}\right)b - b + 1}{b-1} = b - b^{-t} - b + 1 =$$

$$= 1 - b^{-t}$$

  e l'esponente maggiore è U per definizione, il numero piú grande ottenibile è $(1-b^{-t})b^U$.

- Per la natura finita delle mantisse di $\mathbb{F}$, tutti i numeri reali macchina sono razionali: per quanto visto sopra, infatti, nessun numero irrazionale ammette rappresentazione finita.

- Il successivo di un qualunque $x = (0.d_1 \dots d_t)b^p \in \mathbb{F}$ si ottiene "aggiungendo 1 all'ultima cifra della mantissa", ossia sommando a x la quantità b^{p-t}; poiché, però, p non è fisso per ogni $x \in \mathbb{F}$, tale differenza assoluta non è costante: da ciò segue che $\mathbb{F}$ è un insieme a *densità variabile*. La densità cambia nei punti del tipo $(0.10\dots0)b^p$, perché il suo precedente deve avere esponente pari a $p-1$ e, dunque, la distanza tra i due numeri è b^{p-t-1}, mentre la distanza di un tale numero col suo successivo è b^{p-t}.

- Dato $a \in \mathbb{R}$, sia $\tilde{a}$ un numero che approssima a; chiamiamo *errore assoluto* la quantità $|a-\tilde{a}|$ ed *errore relativo* la quantità $|a-\tilde{a}|/|a|$. Nel caso di $\mathbb{F}$, $\tilde{a} = \mathrm{fl}^t(a)$: usando la relazione $|x| \geq b^{p-1}$ con p esponente di x, l'errore relativo è stimato da

$$0 \leq \frac{\left|x - \mathrm{fl}^t(x)\right|}{|x|} \leq \frac{b^{p-t}}{2|x|} \leq \frac{b^{p-t+1-p}}{2} = \frac{b^{1-t}}{2} =: \varepsilon_M$$

  Questo valore ε_M, che non dipende da p ma solo da b e t, prende il nome di *precisione di macchina*.

Figura 1.2: Intorno assoluto di 0.1 in $\mathbb{F}(10, 1, -1, 1)$.

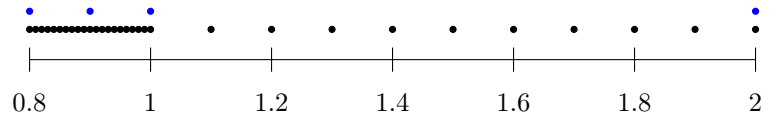

Figura 1.3: Rappresentazione grafica degli insiemi $[\,0.8, 2\,] \cap \mathbb{F}(10, 1, -1, 1)$, in blu, e $[\,0.8, 2\,] \cap \mathbb{F}(10, 2, -2, 2)$, in nero.

- Come già accennato sopra, nei punti del tipo $(0.10\ldots0)b^p$ la densità cambia: da ciò segue che anche l'intorno di arrotondamento non è simmetrico, ma è "decentrato" verso valori piú distanti da 0. Si veda come esempio la Figura 1.2.
- I *numeri rappresentabili* con $\mathbb{F}$, ossia quell'insieme $\mathcal{V} \subseteq \mathbb{R}$ di numeri reali tali che esiste un elemento di $\mathbb{F}$ che li approssimi, è l'insieme

$$\left(-\max\mathbb{F}^+ - \frac{b^{U-t}}{2},\ -\min\mathbb{F}^+ + \frac{b^{L-t}}{2}\right] \cup \left[\min\mathbb{F}^+ - \frac{b^{L-t}}{2},\ \max\mathbb{F}^+ + \frac{b^{U-t}}{2}\right)$$

  a cui va aggiunto lo zero. Per gli scopi del corso, vale la semplificazione $\mathcal{V} = [\,-\max\mathbb{F}^+, -\min\mathbb{F}^+\,] \cup \{\,0\,\} \cup [\,\min\mathbb{F}^+, \max\mathbb{F}^+\,]$.

Nella Figura 1.3 sono disegnate parti $\mathbb{F}(10, 1, -1, 1)$ e di $\mathbb{F}(10, 2, -2, 2)$. Come già visto sopra, 1 è un punto tale che il suo intorno di approssimazione non è simmetrico: nel primo caso, la singola cifra di mantissa fa sí che il precedente di 1 sia 0.9, ma il successivo sia $0.2 \cdot 10 = 2$. Analogamente, nel secondo caso 1 è compreso tra 0.99 e 1.1.

Consideriamo $x = (0.d_1\ldots d_t)b^p \in \mathbb{F}$: qualora si verifichi $p \geq t$, x è necessariamente un intero. Da questa osservazione segue che i numeri reali macchina sono tutti interi a partire da un certo modulo: in particolare, i numeri interi rappresentabili che abbiano piú di t cifre devono essere arrotondati e le loro ultime cifre sono rimpiazzate con 0.

La precisione di macchina ε_M non è il piú piccolo numero macchina rappresentabile: come già notato sopra, il piú piccolo numero positivo rappresentabile è $b^{L-1}(1 - \varepsilon_M)$.

Precisione doppia La codifica standard per approssimare i numeri reali nel calcolatore, e anche in MATLAB, è chiamata *precisione doppia*: in base ad essa, un numero reale (approssimato) può essere rappresentato con 64 bit, di cui uno rappresenta il segno, 52 sono riservati alla mantissa e i restanti 11 all'esponente.[1] Usando la base 2, la mantissa del numero che si vuole rappresentare conta in realtà 53 cifre, perché $d_1 = 1$ per ogni mantissa, non potendo essere 0;

[1] Ciò che segue è una variante facilitata della codifica, perché non tiene conto di alcuni ritocchi dell'implementazione reale, come ad esempio i numeri non normalizzati e la complementazione a due dell'esponente.

per quanto riguarda l'esponente, invece, si usa un bit per il segno e i restanti 10 per il valore assoluto. I numeri macchina secondo questa codifica sono l'insieme $\mathbb{F}(2, 53, -1023, 1023)$. Applicando le osservazioni sopra, si ottiene che $\max \mathbb{F}^+ = 2^{1023}(1 - 2^{-53}) \approx 10^{308}$, $\min \mathbb{F}^+ = b^{L-1} = 2^{-1024} \approx 10^{-308}$, $\varepsilon_M = b^{1-t}/2 = 2^{-53} \approx 10^{-16}$.

1.3 Propagazione degli errori

Siano $x, y \in \mathcal{V}$ e $\cdot$ un'operazione binaria fondamentale. Il computo di una tale operazione in un congruo insieme $\mathbb{F}(b, t, L, U)$ è

$$x \odot y \coloneqq \mathrm{fl}^t\left(\mathrm{fl}^t(x) \cdot \mathrm{fl}^t(y)\right) \tag{1.1}$$

In generale, $\odot$ gode della proprietà commutativa, ma non di quelle associativa e distributiva: si consideri, per esempio, l'insieme $\mathbb{F}(10, 16, -308, 308)$; benché l'operazione $(10^{200} \odot 10^{150}) \odot 10^{-100}$ non possa essere computata per errore di *overflow*, cambiando le parentesi si ottiene $10^{200} \odot (10^{150} \odot 10^{-100}) = 10^{200} \odot 10^{50} = 10^{250}$, come nell'artimetica di $\mathbb{R}$. Nel medesimo $\mathbb{F}$, l'arrotondamento fa sí che $1 \oplus 10^{-16} = 1$: da questo esempio segue che l'elemento neutro delle operazioni macchina non è unico.

Consideriamo l'insieme $\{\, \mu \in \mathbb{F}^+ \mid 1 \oplus \mu > 1 \,\}$ e cerchiamone il minimo. È evidente che il problema sia equivalente a trovare $\min\left\{\, x \in \mathcal{V} \;\middle|\; \mathrm{fl}^t(x) = 1' \,\right\} - 1$, ove $1'$ è il numero successivo a 1 in $\mathbb{F}$; poiché $1' = (0.10\ldots01)b$, esso ha intorno di approssimazione simmetrico di raggio $b^{1-t}/2 = \varepsilon_M$, e in particolare il minimo dell'intorno di approssimazione è $1 - \varepsilon_M$; da ciò segue che $\min\{\, \mu \in \mathbb{F}^+ \mid 1 \oplus \mu > 1 \,\} = \varepsilon_M$.

1.3.1 Stabilità delle operazioni aritmetiche

Dati $x, y \neq 0$, denominiamo gli *errori relativi* su tali quantità

$$\varepsilon_x \coloneqq \frac{|x - \tilde{x}|}{|x|} \qquad\qquad \varepsilon_y \coloneqq \frac{|y - \tilde{y}|}{|y|}$$

Sia, poi, $\star$ un'operazione aritmetica: vogliamo studiare l'errore relativo sull'operazione $x \star y$

$$\varepsilon_{x\star y} \coloneqq \frac{|x \star y - \tilde{x} \star \tilde{y}|}{|x \star y|} \tag{1.2}$$

in funzione di ε_x e ε_y e stabilire se $\star$ è *stabile*, ossia se $\varepsilon_{x\star y}$ ha ordine di grandezza "vicino" a ε_x e ε_y.

Moltiplicazione Si ha

$$\begin{aligned}\varepsilon_{xy} = \frac{|xy - \tilde{x}\tilde{y}|}{|xy|} = \frac{|xy + \tilde{x}y - \tilde{x}y - \tilde{x}\tilde{y}|}{|xy|} = \frac{|(x - \tilde{x})y + (y - \tilde{y})\tilde{x}|}{|xy|} \leq \\ \leq \frac{|(y - \tilde{y})\tilde{x}| + |(x - \tilde{x})y|}{|xy|} = \frac{|x - \tilde{x}|}{|x|\cancel{|y|}}\cancel{|y|} + \frac{|y - \tilde{y}|}{|y|}\frac{|\tilde{x}|}{|x|} = \varepsilon_x + \frac{|\tilde{x}|}{|x|}\varepsilon_y\end{aligned}$$

e, per simmetria, è valida anche la diseguaglianza $\varepsilon_{xy} \leq |\tilde{y}/y|\varepsilon_x + \varepsilon_y$; da ciò segue che $\varepsilon_{xy} \approx \varepsilon_x + \varepsilon_y$ e, dunque, la moltiplicazione è un'operazione stabile.

Divisione Per verificare che la divisione sia un'operazione stabile, consideriamo l'errore relativo sull'inverso di x, supponendo $\varepsilon_x < 1$:

$$\varepsilon_{1/x} = \frac{\left|\frac{1}{x} - \frac{1}{\tilde{x}}\right|}{\left|\frac{1}{x}\right|} = |x|\frac{|\tilde{x} - x|}{|x\tilde{x}|} = \frac{|x|}{|\tilde{x}|}\frac{|x - \tilde{x}|}{|x|} = \frac{|x|}{|\tilde{x}|}\varepsilon_x \leq$$
$$\leq \frac{|x|}{||x - \tilde{x}| - |x||}\varepsilon_x = \left(\frac{|x - \tilde{x}| - |x|}{|x|}\right)^{-1}\varepsilon_x = \frac{\varepsilon_x}{1 - \varepsilon_x}$$

Da ciò segue che $\varepsilon_{x/y} \approx \varepsilon_x + \varepsilon_y$.

Somma algebrica Sotto alcune condizioni, invece, la somma algebrica può essere altamente instabile. La stima dell'errore relativo di tale operazione è data da

$$\varepsilon_{x+y} = \frac{|x + y - \tilde{x} - \tilde{y}|}{|x + y|} = \frac{|x - \tilde{x} + y - \tilde{y}|}{|x + y|} \leq$$
$$\leq \frac{|x - \tilde{x}|}{|x + y|} + \frac{|y - \tilde{y}|}{|x + y|} = \frac{|x|}{|x + y|}\varepsilon_x + \frac{|y|}{|x + y|}\varepsilon_y =: w_1\varepsilon_x + w_2\varepsilon_y$$

Questa stima mostra che, qualora x e y siano concordi, i due "pesi" di ε_x e ε_y sono minori di 1 e, quindi, la somma algebrica è stabile. Qualora, invece, x abbia segno diverso da y, ciò non è garantito: se $|x + y| \ll |x|$ oppure $|x + y| \ll |y|$, infatti, almeno uno dei due pesi è molto maggiore di 1, e ciò "destabilizza" la somma algebrica presa in esame. Si consideri nell'insieme $\mathbb{F}(10, 4, -100, 100)$, ad esempio, la somma algebrica tra $x = 0.10016$ e $y = -0.10012$: la somma in aritmetica di macchina dà come risultato 10^{-4}, mentre il risultato esatto è $4 \cdot 10^{-5}$; l'errore relativo di tale operazione è

$$\varepsilon_{x+y} = \frac{\left|x + y - \mathrm{fl}^4(x) - \mathrm{fl}^4(y)\right|}{|x + y|} = \left|\frac{4 \cdot 10^{-5} - 10^{-4}}{4 \cdot 10^{-5}}\right| = \frac{3}{2}$$

che è inaccettabile perché addirittura maggiore del 100%; il "peso" relativo a ε_x è

$$\left|\frac{x}{x + y}\right| \approx \frac{0.1}{4 \cdot 10^{-5}} = 2500 \approx \left|\frac{y}{x + y}\right|$$

1.3.2 Condizionamento

Analogamente all'errore relativo per le operazioni aritmetiche, possiamo definire l'errore relativo per una certa funzione f

$$\varepsilon_{f(x)} \coloneqq \frac{|f(x) - f(\tilde{x})|}{|f(x)|}$$

supponendo che $f(x) \neq 0$. Qualora f sia anche derivabile, si può applicare la formula di Taylor arrestata al primo ordine

$$f(\tilde{x}) \approx f(x) + f'(x)\,(x - \tilde{x})$$

In base a quest'approssimazione lineare di f si nota che

$$\varepsilon_{f(x)} \approx \frac{|f'(x)||x - \tilde{x}|}{|f(x)|}\frac{|x|}{|x|} = \frac{|f'(x)||x|}{|f(x)|}\varepsilon_x$$

Il "peso" $|x||f'(x)|/|f(x)|$ si indica con $\operatorname{cond} f(x)$ e si dice *indice di condizionamento* di f. Si può scrivere in modo piú compatto, dunque, che

$$\varepsilon_{f(x)} \approx \operatorname{cond}(f(x))\, \varepsilon_x \tag{1.3}$$

Esempio. Si consideri la funzione per ogni $x \neq 0$

$$f(x) \coloneqq \frac{(1+x)-1}{x}$$

Benché $f \equiv 1$, in $\mathbb{F}(2, 53, -1023, 1023)$ – e quindi anche in MATLAB – si ottiene che $f(10^{-15}) = 1.110223024625157$; per contro, $f(2^{-50}) = 1$, nonostante $10^{-15} \approx 2^{-50}$. Si può notare innanzitutto che $10^{-15} \notin \mathbb{F}$: perché un numero della forma $\frac{1}{m}$ con $m \in \mathbb{N}$ ammetta rappresentazione finita in base b, infatti, è necessario che i divisori primi di m siano anche divisori primi di b, e chiaramente non è questo il caso.[2] Calcolando i "pesi" dell'errore relativo, si ha $w_2 = \frac{1}{10^{-15}} = 10^{15} \approx w_1$: da ciò segue che l'errore relativo della sottrazione è circa $10^{15}(\varepsilon_{1+10^{-15}} + \varepsilon_{10^{-15}}) \approx 10^{14}$. Per quanto riguarda 2^{-50}, invece, esso appartiene a $\mathbb{F}$ e cosí anche $1 + 2^{-50}$, perché la mantissa conta al piú 53 cifre; per questo motivo, gli errori relativi ai due operandi della sottrazione sono nulli ed essa restituisce un risultato esatto.

Esempio. Si consideri la funzione per ogni $x \in [\,-1, 1\,]$

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

Il suo condizionamento è

$$\operatorname{cond} f(x) = \left| \frac{x}{\sqrt{1-x^2}} \frac{x}{1-\sqrt{1-x^2}} \right| = \frac{x^2}{\sqrt{1-x^2}\left(1-\sqrt{1-x^2}\right)}$$

che tende a 2 quando $x \to 0$. MATLAB dà che $f(10^{-4}) = 5.000000080634948 \cdot 10^{-9}$, mentre il risultato dovrebbe essere $5.000000012500000 \cdot 10^{-9}$: l'errore relativo è pari a

$$\left| \frac{5.0000000125 \cdot 10^{-9} - 5.000000080634948 \cdot 10^{-9}}{5.0000000125 \cdot 10^{-9}} \right| \approx 10^{-8}$$

ciò è molto maggiore di quanto risulta dalla (1.3), dato che questa dà un risultato dell'ordine di 10^{-17}. Calcolando il peso relativo al secondo operando, si nota che

$$w_2 = \left| \frac{\sqrt{1-x^2}}{1-\sqrt{1-x^2}} \right| = \left| \frac{1+\sqrt{1-x^2}}{x^2} - 1 \right|$$

ed esso tende a $+\infty$ per $x \to 0$: ciò prova l'instabilità di f. Per ovviare a questo problema, occorre *stabilizzare* la funzione:

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

In questo modo si evitano errori relativi grandi per $x \to 0$.

[2] Dimostriamo quanto affermato. Se $\frac{1}{m}$ ammette rappresentazione finita in base b, allora esistono $d_1, \ldots, d_t \in \{\, 0, \ldots, b-1 \,\}$ tali che $\frac{1}{m} = (0.d_1 \ldots d_t)_b\, b^p$; poiché $\frac{1}{m} \leq 1$, si deve avere $p \leq 0$. Agendo sull'ultima eguaglianza si ottiene che dev'essere $\frac{b^{t-p}}{m} = (d_1 \ldots d_t)_b$, ossia che $m | b^{t-p}$. Da questa affermazione si può concludere osservando che i divisori primi di b^{t-p} sono gli stessi di b.

1.3.3 Risoluzione stabilizzata per equazioni di II grado

Data un'equazione di secondo grado $ax^2+bx+c=0$, con $a \neq 0$ e $\Delta = b^2-4ac > 0$ in modo da avere due soluzioni distinte, i due zeri si possono trovare con la formula "classica"; supponendo $b > 0$, si trovano

$$x_- = -\frac{b+\sqrt{\Delta}}{2a} \qquad\qquad x_+ = \frac{-b+\sqrt{\Delta}}{2a}$$

Qualora $\sqrt{\Delta} \approx b$, ossia quando $b^2 \gg 4ac$, tale formula risolutiva è instabile: calcolando i "pesi" della sottrazione al numeratore, infatti, si vede che

$$w_1 = \left|\frac{b}{\sqrt{\Delta}-b}\right| = \frac{b}{\left|b-\sqrt{\Delta}\right|}\frac{\sqrt{\Delta}+b}{\sqrt{\Delta}+b} = \frac{b\Big(\sqrt{\Delta}+b\Big)}{|\Delta-b^2|} \approx \frac{2b^2}{|4ac|} = \frac{b^2}{|2ac|} \approx w_2$$

e, nelle condizioni descritte sopra, questi pesi sono molto maggiori di 1. Questa formula può essere stabilizzata in questo modo:

$$x_+ = \frac{\sqrt{\Delta}-b}{2a} = \frac{\Big(\sqrt{\Delta}-b\Big)\Big(\sqrt{\Delta}+b\Big)}{2a\Big(\sqrt{\Delta}+b\Big)} = \frac{\Delta-b^2}{2a\Big(\sqrt{\Delta}+b\Big)} = \\ = -\frac{4ac}{2a\Big(\sqrt{\Delta}+b\Big)} = -\frac{2c}{\sqrt{\Delta}+b}$$

1.3.4 Approssimazione di π

Volendo calcolare un'approssimazione di π alla precisione di macchina, si può ricorrere alla serie

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$$

da cui segue che

$$\pi = \lim_{n\to\infty}\sqrt{6S_n}, \qquad S_n \coloneqq \sum_{k=1}^{n}\frac{1}{k^2}$$

L'errore compiuto con questo procedimento al passo n-esimo è

$$\left|\sqrt{6S_n}-\pi\right| = \sum_{k=n+1}^{\infty}\frac{1}{k^2} \leq \int_n^{\infty}\frac{1}{x^2}\,\mathrm{d}x = \frac{1}{n}$$

da cui segue che l'errore relativo è asintotico a $\frac{1}{n}$: perché un tale errore sia minore della precisione di macchina ε_M, occorrono circa 10^{16} iterazioni. L'ultimo dato mostra che questo metodo, sebbene sia convergente e stabile, è molto inefficiente.

Un altro metodo convergente a π è la *successione di Archimede*

$$x_2 = 2$$
$$x_{n+1} = 2^{n-\frac{1}{2}}\sqrt{1-\sqrt{1-4^{1-n}x_n^2}}$$

il cui errore relativo decade esponenzialmente in aritmetica a precisione infinita; in aritmetica di macchina, però, la prima sottrazione è instabile e ciò porta

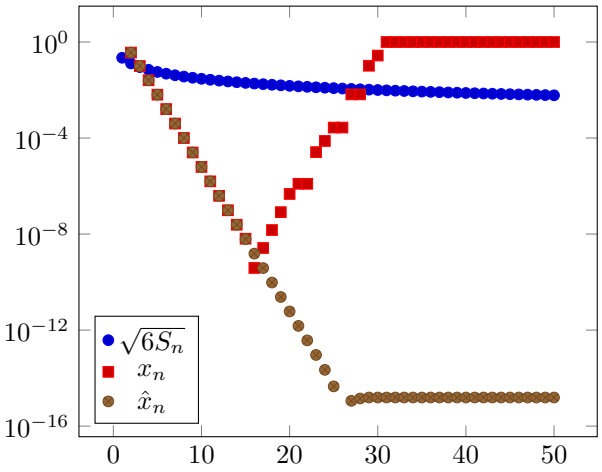

Figura 1.4: Grafico semilogaritmico degli errori relativi per le approssimazioni di π con i metodi visti.

ad un'enorme perdita di precisione dopo poche iterazioni: calcolando il "peso" dell'errore relativo di tale sottrazione, infatti, si ha

$$w_2(n) = \left| \frac{\sqrt{1-4^{1-n}x_n^2}}{1-\sqrt{1-4^{1-n}x_n^2}} \right| = \left| \frac{\sqrt{1-4^{1-n}x_n^2}\left(1+\sqrt{1-4^{1-n}x_n^2}\right)}{4^{1-n}x_n^2} \right| \sim 4^n$$

e ciò mostra l'instabilità della sottrazione. Questa successione si può stabilizzare rimuovendo la sottrazione:

$$x_{n+1} = 2^{n-\frac{1}{2}}\sqrt{1-\sqrt{1-4^{1-n}x_n^2}} = \\ = 2^{n-\frac{1}{2}}\sqrt{\frac{4^{1-n}x_n^2}{1+\sqrt{1-4^{1-n}x_n^2}}} = \frac{\sqrt{2}\,x_n}{\sqrt{1+\sqrt{1-4^{1-n}x_n^2}}}$$

In base a quanto calcolato, si può dare una definizione stabilizzata della successione di Archimede:

$$\hat{x}_2 = 2$$
$$\hat{x}_{n+1} = \frac{\sqrt{2}\,\hat{x}_n}{\sqrt{1+\sqrt{1-4^{1-n}\hat{x}_n^2}}}$$

Per confrontare i due metodi si può usare la stima

$$|\pi - x_n| \leq |\pi - \hat{x}_n| + |x_n - \hat{x}_n|$$

Il primo addendo del membro di destra indica la *convergenza* del metodo, ossia una proprietà teorica; il secondo indica la *stabilità* del metodo, una proprietà che dipende dall'algoritmo e dall'implementazione nel calcolatore. Poiché

$$|\pi - \hat{x}_n| \leq c\,\theta^n \to 0, \qquad \theta \in (0,1)$$
$$|x_n - \hat{x}_n| \lesssim c\,\varepsilon_M 4^n \to +\infty$$

si deduce che il metodo x_n è convergente ma non stabile; oltre un certo n, inoltre, si può verificare che $\theta^n \leq \varepsilon_M 4^n$, ossia che l'errore commesso da x_n cominci a salire: in tal caso l'errore assoluto è asintotico a $\theta^n + O(\varepsilon_M)$.

1.4 Costo computazionale di algoritmi numerici

1.4.1 Algoritmo di Horner

Sia $p(x) = a_0 + \cdots + a_n x^n \in \mathbb{P}_n$ un polinomio: vogliamo calcolarne il valore per un determinato x. L'approccio "classico" a questo problema richiede un numero di operazioni ben determinato:

- occorrono $n-1$ moltiplicazioni per calcolare $x^2, \ldots, x^n$;
- occorrono n moltiplicazioni per calcolare i termini $a_i x^i$;
- occorrono n addizioni per ottenere il risultato finale.

Le operazioni aritmetiche da fare sono, dunque, $3n-1$: segnaliamo tale *costo computazionale* con $C_n^{(1)} \sim 3n$.

Esiste, tuttavia, un metodo con costo computazionale inferiore, ossia lo *schema di Horner*: seguendolo, si può computare un polinomio come

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

con un costo computazionale $C_n^{(2)} \sim 2n$, facendo n prodotti e n somme. Il limite

$$\lim_{n\to\infty} \frac{C_n^{(1)}}{C_n^{(2)}} = \frac{3}{2}$$

indica il fattore di *speed-up* tra i due metodi.

1.4.2 Calcolo di una potenza

Dovendo calcolare la potenza n-esima di un numero a, si effettuano $n-1$ moltiplicazioni col metodo "classico". Esiste, tuttavia, un metodo meno costoso, che si appoggia alla *codifica binaria* dell'esponente $n = \sum_{j=0}^{m} c_j 2^j$ con $m = \lfloor \log_2 n \rfloor$ e $c_j \in \{\, 0, 1 \,\}$ per ogni j. In base a ciò ha senso scrivere

$$a^n = a^{\sum_{j=0}^{m} c_j 2^j} = \prod_{j=0}^{m} a^{c_j 2^j}$$

Seguendo questo algoritmo occorre effettuare al piú $2\lfloor \log_2 n \rfloor$ moltiplicazioni, metà delle quali per calcolare gli a^{2^j} e l'altra metà per ottenere il risultato finale. Lo *speed-up* tra i due metodi è

$$\lim_{n\to\infty} \frac{C_n^{(1)}}{C_n^{(2)}} = +\infty$$

ossia questo secondo metodo è di gran lunga piú veloce del primo.

1.4.3 Calcolo della funzione esponenziale

Supponendo $x > 0$, si verifica che, per ogni $m \in \mathbb{N}$,

$$\mathrm{e}^x = \sum_{k=0}^{m} \frac{x^k}{k!} + h_m(x)$$

ove $h_m(x)$ è il resto, calcolabile mediante l'espansione di Taylor con resto di Lagrange:

$$h_m(x) = \mathrm{e}^{\xi} \frac{x^{m+1}}{(m+1)!}, \qquad \exists \xi \in (\,0, x\,)$$

L'errore relativo commesso considerando solo la somma per l'approssimazione è stimato da

$$\frac{|\mathrm{e}^x - S_m(x)|}{\mathrm{e}^x} \leq \frac{h_m(x)}{\mathrm{e}^x} = \mathrm{e}^{\xi - x} \frac{x^{m+1}}{(m+1)!} < \frac{x^{m+1}}{(m+1)!}$$

ove l'ultima diseguaglianza è giustificata dal fatto che $\xi < x$; tale errore relativo, dunque, è sempre infinitesimo, ma si possono distinguere due casi sull'andamento di tale errore:

- se $x \leq 1$, si ha
$$\frac{x^{m+1}}{(m+1)!} \leq \frac{1}{(m+1)!} < \varepsilon_M$$
non appena $m > 17$; è sufficiente, dunque, calcolare $S_{17}(x)$ per ottenere un'approssimazione alla precisione di macchina;

- se $x > 1$, la stima dell'errore cresce per $m < \lfloor x \rfloor$, poi decresce; ponendo, però, $\kappa \coloneqq \lfloor x \rfloor + 1$ e $y \coloneqq x/\kappa < 1$, vale la relazione
$$\mathrm{e}^x = \left(\mathrm{e}^y\right)^{\kappa} \approx \left(S_{17}(y)\right)^{\kappa}$$
e ci si riconduce al caso precedente, privo di crescita nella stima dell'errore; il costo computazionale è circa $2 \log_2 \kappa + 34$, tenendo conto sia del computo di $S_{17}(y)$ sia dell'elevamento a potenza.

2 Soluzione numerica di equazioni non lineari

Sia $f \colon I \to \mathbb{R}$, con $I = [\,a, b\,]$ intervallo chiuso e limitato, una funzione; in questa sezione esponiamo metodi per risolvere equazioni del tipo $f(x) = 0$ oppure $x = f(x)$ con $x \in I$.

2.1 Metodo di bisezione

Supponiamo che f sia continua in I e che $f(a)\, f(b) < 0$; per il Teorema degli zeri, esiste $\xi \in (\,a, b\,)$ tale che $f(\xi) = 0$. Tale zero è unico se f è strettamente monotona in I; se f è derivabile in I, si ha che lo zero è unico se f' è strettamente positiva o negativa in I.

2.1.1 Algoritmo

Denominando $a_0 = a$, $b_0 = b$, sia $x_0 = \frac{a_0+b_0}{2}$. Se $f(x_0) = 0$, allora si è trovato lo zero; se $f(x_0) > 0$, allora si pongono $a_1 = a_0$ e $b_1 = x_0$; nel caso $f(x_0) < 0$, invece, si pongono $a_1 = x_0$ e $b_1 = b_0$. Il procedimento, se non è stato trovato uno zero, si ripete con l'intervallo $[\, a_1, b_1 \,]$. Si suole denominare con a_n e b_n gli estremi per la $(n+1)$-esima iterazione dell'algoritmo.

2.1.2 Errore *a priori*

Data la successione $(x_n)_{n\in\mathbb{N}}$ dei punti medii degli intervalli ottenuti col metodo di bisezione, si vede che l'errore assoluto, ossia la distanza di un certo x_n da uno zero ξ, è stimato da

$$e_n \coloneqq |x_n - \xi| \leq \frac{b-a}{2^{n+1}} \tag{2.1}$$

in quanto x_n è il punto medio di un intervallo di "lunghezza" $\frac{b-a}{2^n}$ e, per il Teorema degli zeri, ξ deve essere un punto di quell'intervallo. Questa stima mostra che $e_n \to 0$ quando $n \to +\infty$, ossia che il metodo di bisezione è sempre convergente ad uno zero di f. Da questa stima *a priori* di e_n segue che, perché $e_n \leq \varepsilon$, occorre che si verifichi

$$n \geq \log_2 \left(\frac{b-a}{\varepsilon} \right) - 1$$

in questo modo è possibile arrestare l'algoritmo dopo un numero finito di passi se si vuole approssimare ξ con un errore inferiore a ε.

2.1.3 Errore *a posteriori*

Per il Teorema di caratterizzazione sequenziale delle funzioni continue, se $\lim_{n\to\infty} x_n = \xi$, con ξ radice di f, allora $\lim_{n\to\infty} f(x_n) = f(\xi) = 0$. Un parametro che sembra alternativo all'errore assoluto è il *residuo* $|f(x_n)|$ al variare di n; succede, invece, che l'errore è "scorrelato" al residuo: nel caso in cui la crescenza di f in un intorno di ξ sia molto "bassa", infatti, il residuo è una *sottostima* dell'errore; nel caso in cui f cresca molto "rapidamente" in un intorno di ξ, per contro, il residuo è una *sovrastima* dell'errore.

Qualora $f \in C^1(I)$, è possibile costruire un indicatore *a posteriori* piú "attendibile". Se esiste $[\, c, d \,] \subseteq I$ tale che $f'(x) \neq 0$ per ogni $x \in [\, c, d \,]$ e $(x_n)_{n\in\mathbb{N}}$ è definitivamente contenuta in $[\, c, d \,]$, allora ξ è uno *zero semplice*, ossia $f'(\xi) \neq 0$; per il Teorema di Lagrange esiste $z_n \in \operatorname{int}(x_n, \xi)$ tale che $\frac{f(x_n)-f(\xi)}{x_n-\xi} = f'(z_n)$.[3] Poiché ξ è uno zero di f, l'espressione appena scritta giustifica

$$e_n = |x_n - \xi| = \left| \frac{f(x_n)}{f'(z_n)} \right| \qquad \exists z_n \in \operatorname{int}(x_n, \xi) \tag{2.2}$$

Poiché gli z_n non sono prodotti esplicitamente dal Teorema di Lagrange, può sembrare che tale stima non fornisca alcuna informazione. Le ipotesi su $[\, c, d \,]$,

[3] La notazione $\operatorname{int}(x_1, \ldots, x_n)$ indica l'intervallo aperto di estremi $\min\{\, x_1, \ldots, x_n \,\}$ e $\max\{\, x_1, \ldots, x_n \,\}$. Con $\operatorname{int}(a, b)$ si intende l'intervallo $(\, a, b \,)$, se $a < b$, oppure $(\, b, a \,)$, se $b < a$. Si ricorre a questa notazione per non specificare ogni volta quale dei due numeri in esame sia il maggiore.

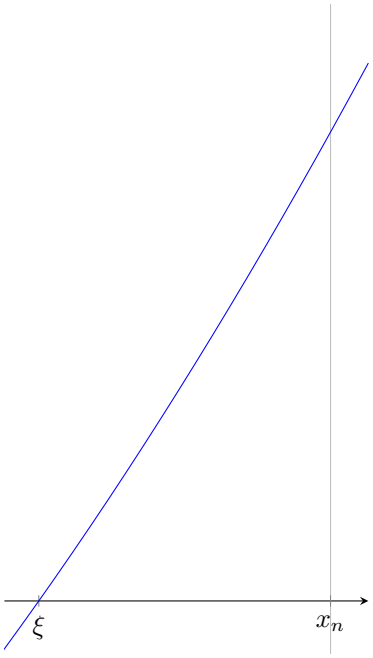

Figura 2.1: Esempio di sovrastima dell'errore da parte del residuo.

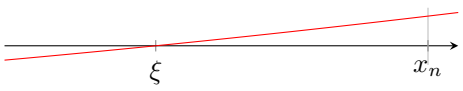

Figura 2.2: Esempio di sottostima dell'errore da parte del residuo.

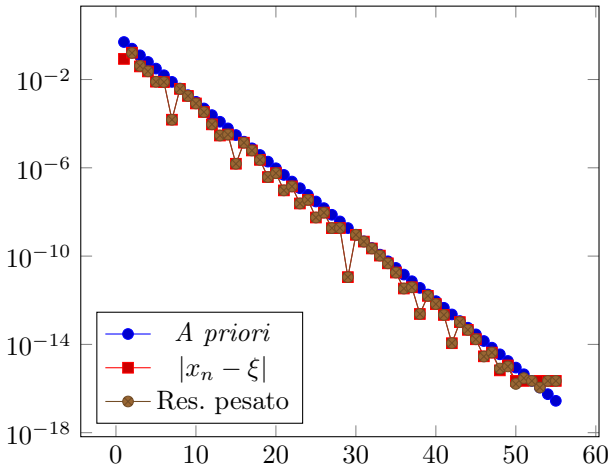

Figura 2.3: Confronto tra gli errori *a priori*, reale e stimato col residuo pesato delle iterazioni per risolvere l'equazione $x^2 - 2 = 0$ col metodo di bisezione.

però, implicano che $|f'(x)| \geq k > 0$ per ogni $x \in [\,c, d\,]$; da ciò segue che

$$e_n = \left| \frac{f(x_n)}{f'(z_n)} \right| \leq \left| \frac{f(x_n)}{k} \right|$$

Poiché, poi, $f \in C^1([\,c, d\,])$, si verifica $|f'(x_n)| \approx |f'(z_n)|$ per n sufficientemente grande; da ciò segue che

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \overset{\text{L}}{=} f'(u_n) \approx f'(x_n) \approx f'(z_n) \qquad \exists u_n \in \text{int}(x_n, x_{n-1}) \subseteq [\,c, d\,]$$

e, per il Teorema dei Carabinieri, le tre derivate tendono tutte a $f'(\xi)$: si può affermare, dunque, che per un n sufficientemente grande è vero

$$e_n \approx |f(x_n)| \left| \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \right| \tag{2.3}$$

Tali affermazioni sono mostrate con il grafico in Figura 2.3.

2.2 Metodo di Newton o delle tangenti

Un altro metodo per trovare uno zero di f, seppur in condizioni molto piú specifiche, è il *metodo di Newton*. A grandi linee, esso consiste nel tracciare la retta tangente alla funzione, trovare l'intersezione di questa retta con l'asse delle ascisse e ripetere il procedimento a partire dal punto della funzione calcolato nell'intersezione trovata prima, come mostrato nella Figura 2.4.

2.2.1 Algoritmo

Le ipotesi per applicare il metodo di Newton sono molto piú forti rispetto a quelle necessarie per il metodo di bisezione: occorre, infatti, che $f \in C^2(I)$, che $f(a)\, f(b) < 0$, che il segno di f'' sia costante e che $f''(x) \neq 0$ per ogni $x \in I$. Diamo le condizioni di convergenza sotto forma di teoremi, distinguendo i due casi di convergenza globale e locale del metodo.

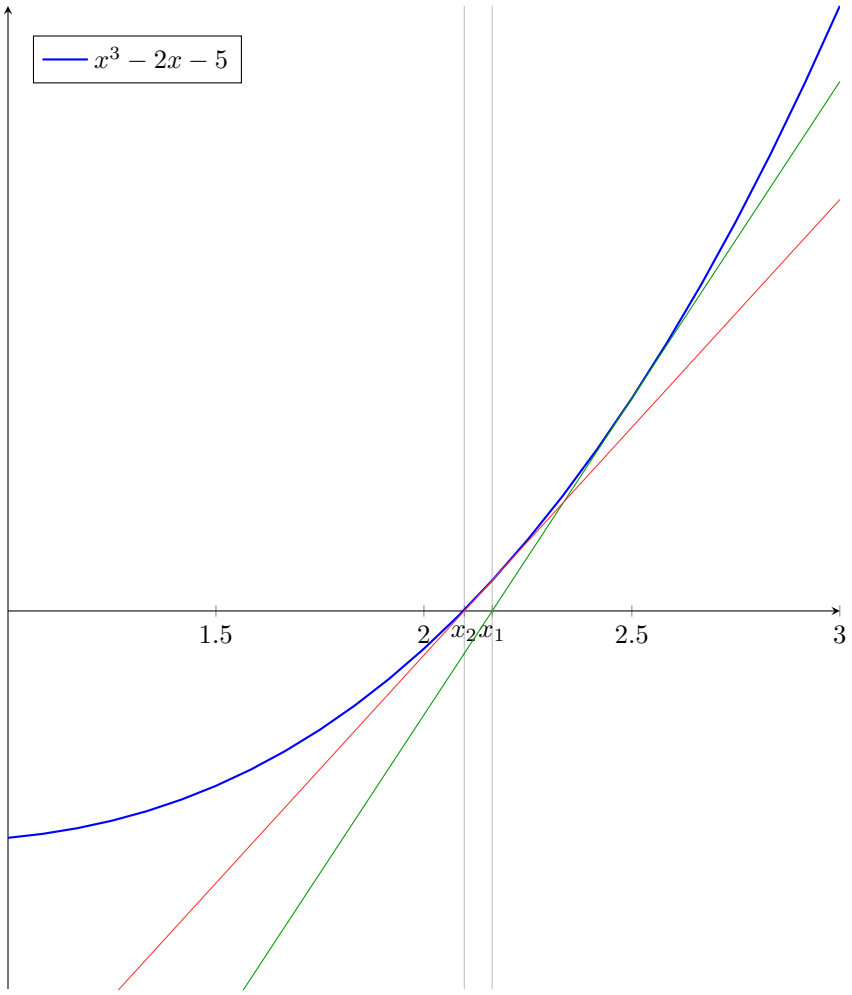

Figura 2.4: Visualizzazione grafica del metodo di Newton per la funzione $x^3 - 2x - 5$, con la quale Newton stesso mostrò il suo metodo. In questo caso, si sceglie $x_0 = 5/2$; la tangente generata dalla prima iterazione è quella verde, mentre la retta rossa è prodotta dalla seconda iterazione. Si noti che x_2 è già molto vicino alla radice.

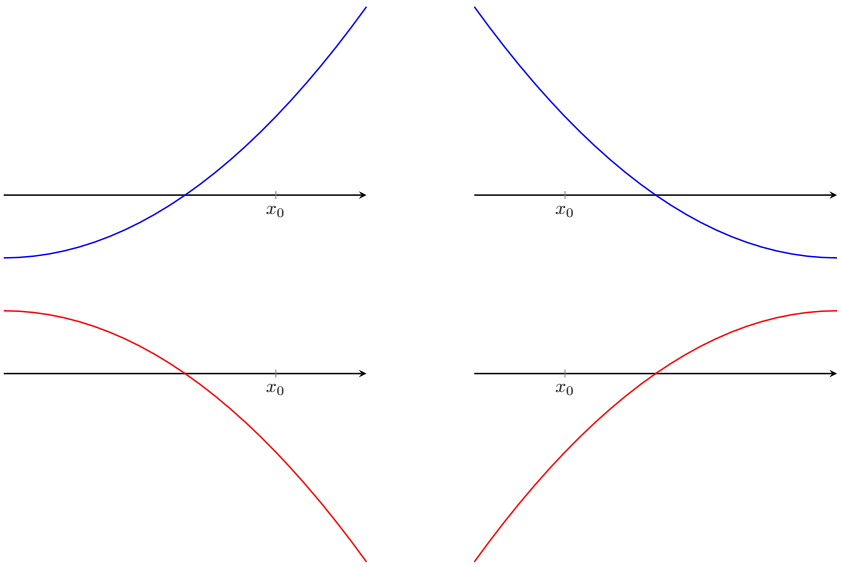

Figura 2.5: Rappresentazione dei quattro casi possibili in cui applicare il metodo di Newton.

Teorema 2.1 (Convergenza globale del metodo di Newton)**.** *Sia $f \in C^2(I)$ tale che $f(a)\,f(b) < 0$ e tale che $f''(x) \geq 0$ (oppure $f''(x) \leq 0$) per ogni $x \in I$. Se esiste $x_0 \in I$ tale che $f''(x_0)\,f(x_0) > 0$, allora la successione*

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.4}$$

è ben definita in I e converge all'unica radice η di f in I.

Dimostrazione. Assumiamo che $f''(x) \geq 0$ per ogni $x \in I$, ovvero che f sia convessa in I: se il Teorema vale per f, allora varrà un analogo per $-f$. Dimostriamo che esiste uno ed un solo zero η per f: per il Teorema degli zeri, f ha almeno uno zero in $[\,a, b\,]$. Per assurdo esistano due zeri di f: poiché f è convessa, il segmento che congiunge η all'altro zero $\tilde{\eta}$ è tutto contenuto nell'epigrafico di f; se f non è identicamente nulla, per quanto appena detto si ha $f(x) < 0$ per ogni $x \in \mathrm{int}(\eta, \tilde{\eta})$. Per il Teorema di Rolle esiste $z \in \mathrm{int}(\eta, \tilde{\eta})$ di derivata nulla; per il Teorema di Weierstrass, z è di minimo assoluto in $\mathrm{int}(\eta, \tilde{\eta})$. Poiché f'' è sempre positiva, però, si ha $f(x) \geq 0$ per ogni $x \in [\,a, b\,] \setminus \mathrm{int}(\eta, \tilde{\eta})$, e ciò è assurdo, perché si avrebbe $f(a)\,f(b) > 0$. Assumiamo, ora, che $x_0 = b > \eta$: se il Teorema vale per $f(x)$, varrà un analogo per $f(-x)$ con $x_0 = a$. Uno specchietto grafico delle quattro situazioni possibili si trova alla Figura 2.5.

Poiché f è convessa, la retta tangente il grafico di f nel punto $(x_0, f(x_0))$ non interseca l'epigrafico di f se non nel punto di tangenza. Applicando la definizione di coefficiente angolare della retta, si trova che

$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1}$$

ove x_1 è l'intersezione della retta tangente con l'asse delle ascisse; da questa formula si trova

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

e tale scrittura ha senso perché f è strettamente crescente tra η e x_0 — e, in particolare, ha derivata in x_0 strettamente positiva. Si ha, inoltre, che $\eta < x_1 < x_0$: la quantità tolta da x_0, infatti, è strettamente positiva per ipotesi; né può essere $x_1 \leq \eta$, dato che, se cosí fosse, $f'(x_0)$ sarebbe minore o uguale al rapporto incrementale tra η e x_0 e la retta tangente dovrebbe intersecare l'epigrafico di f, che è assurdo. La derivata $f'(x_1)$, inoltre, è ancora positiva: se cosí non fosse, si avrebbe che $x_1 < z < \eta$ con z minimo assoluto per f, che contraddice ancora quanto dimostrato.

Ripetendo il procedimento per ogni $n \in \mathbb{N}$ come definito nella (2.4), la successione $(x_n)_{n\in\mathbb{N}}$ è strettamente monotona decrescente ed inferiormente limitata da η: per il Teorema delle successioni monotone, la successione x_n ammette limite $\xi \in [\, a, b\,]$ per $n \to \infty$. Passando al limite nella (2.4), si ha

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \left(x_n - \frac{f(x_n)}{f'(x_n)} \right) \implies \cancel{\xi} = \cancel{\xi} - \frac{f(\xi)}{f'(\xi)} \implies \frac{f(\xi)}{f'(\xi)} = 0$$

e, poiché $f'(\xi) \neq 0$, si ottiene che $f(\xi) = 0$ e, per l'unicità dello zero η, si conclude che $\xi = \eta$. □

Teorema 2.2 (Convergenza locale del metodo di Newton)**.** *Dato $\eta \in \mathbb{R}$, sia $J = [\, \eta - \delta, \eta + \delta\,]$ con $\delta > 0$; sia, poi, $f \in C^2(J)$ tale che $f(\eta) = 0$ e tale che $f'(x) \neq 0$ per ogni $x \in J$. Se esiste $x_0 \in J$ tale che*

$$e_0 \coloneqq |x_0 - \eta| < \min\left\{ \frac{1}{c},\, \delta \right\}$$

ove

$$c \coloneqq \frac{\max\limits_{x\in J} |f''(x)|}{2 \min\limits_{x\in J} |f'(x)|}$$

allora la successione $(x_n)_{n\in\mathbb{N}}$ costruita col metodo di Newton a partire da x_0 per f è ben definita in J (ovvero è tutta contenuta in J) e converge a η; definiti $e_n \coloneqq |x_n - \eta|$, inoltre, vale la diseguaglianza

$$c\, e_n \leq (c\, e_0)^{2^n} \tag{2.5}$$

Dimostrazione. Poiché $f \in C^2(J)$, si può usare lo sviluppo di Taylor con resto di Lagrange centrato in x_n per un certo $n \in \mathbb{N}$:[4]

$$f(\eta) = f(x_n) + f'(x_n)(\eta - x_n) + \frac{f''(\xi)}{2}(\eta - x_n)^2 \qquad \exists \xi \in \operatorname{int}(\eta, x_n)$$

Dal momento che $f(\eta) = 0$, si ricava

$$-\frac{f(x_n)}{f'(x_n)} = \eta - x_n + \frac{f''(\xi)\,(\eta - x_n)^2}{2\, f'(x_n)}$$

[4]Per com'è stato enunciato il Teorema, esiste almeno un $n \in \mathbb{N}$ possibile, ovvero 0. La dimostrazione sottointende, dunque, un ragionamento induttivo su $n \in \mathbb{N}$ per far "partire" il metodo di Newton.

e ciò è lecito perché $f'(x_n) \neq 0$ per ipotesi. Sostituendo quanto appena trovato nella (2.4), si ha

$$x_{n+1} = \cancel{x_n} + \eta - \cancel{x_n} + \frac{f''(\xi)\,(\eta - x_n)^2}{2\,f'(x_n)}$$

da cui segue che

$$e_{n+1} = |x_{n+1} - \eta| = \frac{|f''(\xi)|}{2|f'(x_n)|}(\eta - x_n)^2 \leq c|\eta - x_n|^2 = c\,e_n^2$$

Per ipotesi induttiva, si ottiene $c\,e_{n+1} \leq (c\,e_n)^2 \leq \cdots \leq (c\,e_0)^{2^n}$. Perché la successione $(c\,e_n)_{n\in\mathbb{N}}$ sia infinitesima, ossia perché $x_n \to \eta$, occorre che $c\,e_0 < 1$; per ipotesi del Teorema, si possono distinguere due casi: se $e_0 < 1/c$, allora $c\,e_0 < 1$ e si può concludere; se $e_0 < \delta$, allora $\delta < 1/c$ e ci si riconduce al caso sopra — in particolare, $e_n < \delta$ per ogni $n \in \mathbb{N}$, ossia $x_n \in J$ per ogni $n \in \mathbb{N}$. Si può concludere, dunque, che $x_n \to \eta$ e che $(x_n)_{n\in\mathbb{N}} \subseteq J$. □

2.2.2 Ordine di convergenza

Diamo ora alcune definizioni riguardanti gli errori assoluti e_n per quantificare quanto "velocemente" il metodo di Newton pervenga alla soluzione.

Definizione 2.1. Sia $(e_n)_{n\in\mathbb{N}}$ l'errore assoluto di un metodo convergente. Si dice che tale metodo *ha ordine di convergenza almeno* $p \geq 1$ se esiste $c > 0$ tale che $e_{n+1} \leq c\,e_n^p$; in particolare, se $p = 1$, si deve avere $c \in (\,0, 1\,)$. Si dice, poi, che il metodo *ha ordine di convergenza esattamente* p se

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n^p} = \ell > 0 \tag{2.6}$$

in particolare, se $p = 1$, si deve avere $\ell \in (\,0, 1\,)$.

Se un metodo ha ordine di convergenza 1, tale convergenza si dice *lineare*; se l'ordine di convergenza è esattamente $p > 1$, allora si dice *superlineare*.

Per quanto visto nel Teorema 2.2, il metodo di Newton ha ordine di convergenza almeno 2. Volendo trovare quando l'ordine di convergenza è esattamente 2, notiamo che, se η è uno zero semplice di f,

$$\frac{e_{n+1}}{e_n^2} = \frac{|x_{n+1} - \eta|}{|x_n - \eta|^2} = \frac{\left|x_n - \frac{f(x_n)}{f'(x_n)} - \eta\right|}{|x_n - \eta|^2} = \frac{|f''(\xi)|}{2|f'(x_n)|}\frac{|x_n - \eta|^2}{|x_n - \eta|^2} = \frac{|f''(\xi)|}{2|f'(x_n)|}$$

per un certo $\xi \in \mathrm{int}(\eta, x_n)$. Poiché anche ξ tende a η, l'unico caso in cui il rapporto esplicitato sopra non sia un numero $\ell > 0$ è quello in cui $f''(\xi) \to 0$, ossia se $f''(\eta) = 0$: in questo caso, dunque, l'ordine di convergenza del metodo è maggiore di 2; in tutti gli altri casi, invece, è esattamente 2.

In generale, dato un metodo con ordine di convergenza almeno $p > 1$, esiste $C > 0$ tale che $C\,e_n \leq (C\,e_0)^{p^n}$. Si ha, infatti, che

$$e_n \leq c\,e_{n-1}^p \leq \cdots \leq \left(\prod_{j=0}^{n-1} c^{p^j}\right) e_0^{p^n} =$$

$$= c^{\sum_{j=0}^{n-1} p^j}\, e_0^{p^n} = c^{\frac{p^n-1}{p-1}}\, e_0^{p^n} = c^{-\frac{1}{p-1}} \left(c^{\frac{1}{p-1}}\, e_0\right)^{p^n}$$

Ponendo, quindi, $C = \sqrt[p-1]{c}$, si trova quanto cercato.

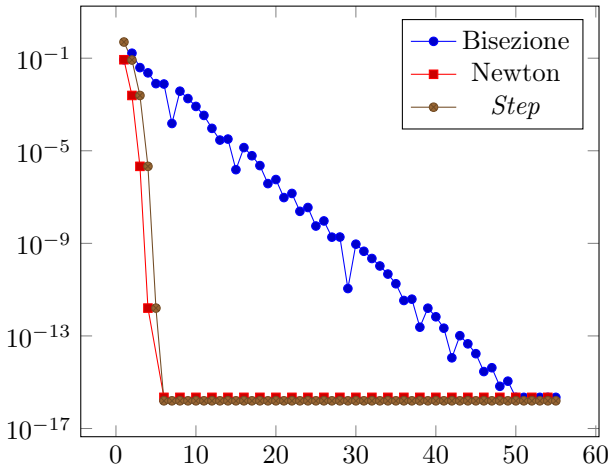

Figura 2.6: Confronto tra gli errori assoluti compiuti dal metodo di bisezione e dal metodo di Newton.

Esempio. Sia $f(x) \coloneqq x^k - a$ con $a > 0$. Col metodo di Newton, scegliendo $x_0 > 0$ tale che $x_0^k > a$, si ha

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^k - a}{k x_n^{k-1}} = \frac{k x_n^k - x_n^k + a}{k x_n^{k-1}} = \frac{k-1}{k} x_n + \frac{a}{k x_n^{k-1}}$$

questa formula è un'estensione dell'algoritmo di Erone per l'estrazione non solo delle radici quadrate ma delle radici k-esime.

2.2.3 Errori ed efficienza

L'algoritmo di Newton ha il vantaggio di convergere in modo particolarmente veloce rispetto al metodo di bisezione. Si consideri, ad esempio, la funzione $f(x) \coloneqq x^2 - 2$ definita nell'intervallo $[\,1, 2\,]$: perché l'errore assoluto commesso col metodo di bisezione sia sotto un certo 10^{-k}, con $k \in \mathbb{N}$, occorre che $n = \left\lfloor \log_2 10^k \right\rfloor$; la successione $(n_k)_{k\in\mathbb{N}}$ è del tipo $(3, 6, 9, 13, 16, 19, 23, \dots)$: è evidente, dunque, che una cifra decimale in piú di precisione corrisponde a 3 oppure 4 iterazioni ulteriori dell'algoritmo. Per il metodo di Newton, invece, occorre innanzitutto vedere che $f'(x) = 2x$ e $f''(x) = 2$: si ottiene che $c = 1$; scegliendo $x_0 = 2$, la successione degli errori è $(0{,}585\,786,\ 0{,}085\,786\,4,\ 0{,}002\,453\,1,\ 2{,}1239 \cdot 10^{-6},\ 1{,}594\,72 \cdot 10^{-12}, \dots)$; si nota che l'ordine di grandezza dell'errore è circa dimezzato ad ogni iterazione, ossia le cifre decimali corrette del risultato approssimato raddoppiano ad ogni iterazione. Una dimostrazione piú rigorosa di questa evidenza si può ottenere considerando l'errore relativo $\rho_n \coloneqq e_n / |\xi|$; studiando questo al passo $n + 1$, infatti, si ha

$$\rho_{n+1} = \frac{e_{n+1}}{|\xi|} \leq \frac{c\, e_n^2}{|\xi|} \frac{|\xi|}{|\xi|} = c\,|\xi| \rho_n^2$$

se $c\,|\xi| \leq 1$ — e ciò è vero scegliendo in modo appropriato il punto iniziale x_0 o, equivalentemente, l'intorno di ξ in cui applicare il metodo.

Per com'è costruita la successione $(x_n)_{n\in\mathbb{N}}$, lo *step* $|x_{n+1} - x_n| = -\frac{f(x_n)}{f'(x_n)}$ è un residuo pesato: analogamente al metodo di bisezione, esso è una buona approssimazione dell'errore.

2.3 Iterazioni di punto fisso

2.3.1 Convergenza

Dal Teorema delle contrazioni si può ottenere un metodo di soluzione approssimata di equazioni del tipo $\varphi(x) = x$, se si prova che φ è una contrazione.

Teorema 2.3 (delle contrazioni)**.** *Sia $\varphi\colon I \to \mathbb{R}$, con $I \subseteq \mathbb{R}$ intervallo chiuso, una funzione derivabile in I tale che $\varphi(I) \subseteq I$ e $|\varphi'(x)| \leq \vartheta < 1$ per ogni $x \in I$. Allora esiste un unico $\eta \in I$ tale che, per ogni $x_0 \in I$, la successione $(x_n)_{n\in\mathbb{N}}$ definita ricorsivamente da $x_{n+1} \coloneqq \varphi(x_n)$ verifichi $\lim_{n\to\infty} x_n = \eta$.*

Dimostrazione. Dalle ipotesi sulla derivata di φ segue che è possibile applicarvi il Teorema di Lagrange:

$$\forall x, y \in I : \exists z \in \operatorname{int}(x, y) : \varphi(x) - \varphi(y) = \varphi'(z)(x - y)$$

Passando ai moduli, si ottiene

$$|\varphi(x) - \varphi(y)| = |\varphi'(z)|\,|x - y| \leq \vartheta\,|x - y| \qquad \exists \vartheta \in (\,0, 1\,)$$

ossia che φ è una contrazione.

Occorre dimostrare, ora, che $(x_n)_{n\in\mathbb{N}}$ è una successione di Cauchy. Per la disuguaglianza triangolare, si ha per ogni $n, k \in \mathbb{N}$

$$\begin{aligned} |x_{n+k} - x_n| &= |x_{n+k} + x_{n+k-1} - x_{n+k-1} + \cdots + x_{n+1} - x_{n+1} - x_n| \leq \\ &\leq \sum_{h=1}^{k} |x_{n+h} - x_{n+h-1}| = \sum_{h=1}^{k} \left|\varphi^{n+h}(x_0) - \varphi^{n+h-1}(x_0)\right| \leq \\ &\leq |\varphi(x_0) - x_0| \sum_{h=1}^{k} \vartheta^{n+h-1} \leq \vartheta^n\,|\varphi(x_0) - x_0| \sum_{h=1}^{\infty} \vartheta^{h-1} \end{aligned}$$

ove con φ^m si intende l'applicazione di φ per m volte; dal momento che $\vartheta^n \to 0$ quando $n \to \infty$ e gli altri fattori dell'ultimo risultato sono limitati, segue che $(x_n)_{n\in\mathbb{N}}$ è di Cauchy e, quindi, ammette un limite $\eta \in I$ — e si verifica

$$\eta = \lim_{n\to\infty} \varphi^n(x_0) = \lim_{n\to\infty} \varphi\big(\varphi^{n-1}(x_0)\big) = \varphi\Big(\lim_{n\to\infty} \varphi^{n-1}(x_0)\Big) = \varphi(\eta)$$

Tale limite è unico: se cosí non fosse ed esistessero $\eta_1, \eta_2 \in I$ che soddisfino alle proprietà richieste, si avrebbe

$$|\eta_1 - \eta_2| = |\varphi(\eta_1) - \varphi(\eta_2)| \leq \vartheta\,|\eta_1 - \eta_2|$$

e ciò è possibile se e solo se $\eta_1 = \eta_2$, perché $0 < \vartheta < 1$. □

Questo Teorema ha senso ove le successioni, se sono di Cauchy, convergono: ciò significa che I può anche essere una semiretta chiusa oppure tutto $\mathbb{R}$, in quanto anch'essi sono spazii metrici completi per la metrica euclidea standard.

Corollario 2.1. *Dato $\xi \in \mathbb{R}$, sia φ una funzione di classe C^1 in un certo intorno di ξ e tale che $|\varphi'(\xi)| < 1$. Se ξ è punto fisso di φ, allora, per ogni x_0 in un opportuno intorno di ξ, la successione $(x_n)_{n\in\mathbb{N}}$ definita ricorsivamente da $x_{n+1} \coloneqq \varphi(x_n)$ converge a ξ.*

Dimostrazione. Poiché $|\varphi'(\xi)| < 1$ e la derivata di φ è continua intorno a ξ, esiste $\delta > 0$ tale che ogni $x \in [\,\xi - \delta, \xi + \delta\,] =: I$ verifichi la proprietà $|\varphi'(x)| < 1$; per il Teorema di Weierstrass, inoltre, esiste $\vartheta \coloneqq \max_{x\in I}|\varphi'(x)| < 1$. Se si mostra che $\varphi(I) \subseteq I$, si può concludere per il Teorema 2.3: dato $x \in I$, si osserva che

$$|\varphi(x) - \xi| = |\varphi(x) - \varphi(\xi)| \leq \vartheta\,|x - \xi| < |x - \xi| \leq \delta$$

da cui segue che $\varphi(x) \in I$. □

2.3.2 Errori

Data una contrazione φ di punto fisso ξ, sia x_0 un punto intorno a ξ tale che valgano le condizioni del Teorema 2.3; costruendo la successione $(x_n)_{n\in\mathbb{N}}$ come nel Teorema, si può notare che l'errore assoluto *a priori* dell'iterazione n-esima è stimato da

$$\begin{aligned} |x_n - \xi| = \\ = |\varphi^n(x_0) - \varphi^n(\xi)| \leq \vartheta\left|\varphi^{n-1}(x_0) - \varphi^{n-1}(\xi)\right| \leq \\ \leq \cdots \leq \vartheta^n\,|x_0 - \xi| \end{aligned} \tag{2.7}$$

Una stima dell'errore assoluto *a posteriori*, invece, si può ricavare dallo *step*

$$x_{n+1} - x_n = x_{n+1} - \xi + \xi - x_n = \varphi'(z_n)\,(x_n - \xi) + \xi - x_n \quad \exists z_n \in \operatorname{int}(x_n, \xi)$$

da cui, riordinando e passando al modulo, si ottiene

$$|x_n - \xi| = \frac{|x_{n+1} - x_n|}{1 - \varphi'(z_n)} \leq \frac{1}{1-\vartheta}|x_{n+1} - x_n| \leq \frac{\vartheta^n}{1-\vartheta}|x_1 - x_0| \tag{2.8}$$

Esempio. Vogliamo risolvere numericamente l'equazione $x - \mathrm{e}^{-\alpha x} = 0$ con $\alpha \in (\,0, 1\,)$. Il problema è equivalente a trovare il punto fisso della funzione $\varphi_\alpha(x) \coloneqq \mathrm{e}^{-\alpha x}$; poiché $\varphi'_\alpha(x) = -\alpha\,\mathrm{e}^{-\alpha x}$, è sufficiente osservare che $|\varphi'_\alpha(x)| < 1$ nell'intervallo $[\,0, +\infty\,)$. Da ciò si può applicare il Teorema 2.3 ed affermare che, per ogni α, l'equazione di partenza ha una ed una sola soluzione, ottenibile con le iterazioni di punto fisso. In base a quanto visto sopra, ponendo ad esempio $\alpha = 1/5$, se si vuole che con l'approssimazione si commetta un errore assoluto inferiore a 10^{-8}, si può usare la (2.8) ed ottenere

$$\frac{\vartheta^n}{1-\vartheta}|x_1 - x_0| < 10^{-8}$$

ove $\vartheta = 1/5$; scegliendo $x_0 = 0$ la relazione sopra diventa

$$\frac{1}{4} \cdot \frac{1}{5^n} < 10^{-8} \implies n > 8\log_5 10 - \log_5 4 \approx 10{,}584$$

2.3.3 Ordine di convergenza

Assumendo che una contrazione φ sia derivabile un certo numero di volte intorno al suo punto fisso ξ, è possibile determinare in modo preciso l'ordine di convergenza del metodo.

Proposizione 2.1. *Sia φ una contrazione di punto fisso ξ e sia $(x_n)_{n\in\mathbb{N}}$ la successione ottenuta iterando a partire da un certo x_0 intorno a ξ. Se φ è di classe C^p, con $p \in \mathbb{N} \setminus \{ 0 \}$, in un certo intorno di ξ, allora*

- *se $p = 1$, il metodo di iterazioni di punto fisso ha ordine di convergenza 1 se e solo se $|\varphi'(\xi)| < 1$;*
- *se $p > 1$, il metodo di iterazioni di punto fisso ha ordine di convergenza p se e solo se $\varphi^{(p)}(\xi) \neq 0$ e $\varphi^{(i)}(\xi) = 0$ per ogni $i \in \{ 1, \dots, p-1 \}$.*

Dimostrazione. Dimostriamo innanzitutto il caso $p = 1$. Per la Definizione 2.1 si ha

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = \ell \in (\,0, 1\,)$$

Per il Teorema di Lagrange si ha

$$e_{n+1} = |x_{n+1} - \xi| = |\varphi(x_n) - \varphi(\xi)| \overset{\text{L}}{=} |\varphi'(z_n)|\,|x_n - \xi| \qquad \exists z_n \in \operatorname{int}(x_n, \xi)$$

e, sostituendo nel limite, si ha

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = \lim_{n\to\infty} |\varphi'(z_n)| = |\varphi'(\xi)| \in (\,0, 1\,)$$

dal momento che l'ultima eguaglianza è ver in ipotesi di convergenza del metodo, si conclude.

Dimostriamo ora il caso $p > 1$.

($\Rightarrow$) Per la Definizione 2.1 si ha

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n^p} = \ell \in \mathbb{R}^+$$

per assurdo esista $k \in \{ 1, \dots, p-1 \}$ tale che $\varphi^{(k)}(\xi) \neq 0$ e sia il minimo per cui ciò accade: per la formula di Taylor con resto di Lagrange centrata in ξ si ha che

$$x_{n+1} - \xi = \varphi(x_n) - \varphi(\xi) = \varphi^{(k)}(z_n)\,(x_n - \xi)^k \qquad \exists z_n \in \operatorname{int}(x_n, \xi)$$

passando al valore assoluto e poi al limite, si ottiene

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n^k} = \lim_{n\to\infty} \frac{\left|\varphi^{(k)}(z_n)\right| \cancel{|x_n - \xi|^k}}{\cancel{|x_n - \xi|^k}} = \varphi^{(k)}(\xi) \in \mathbb{R}^+$$

ossia il metodo converge con ordine esattamente $k < p$, il che è assurdo.

($\Leftarrow$) Applicando la formula di Taylor centrata in ξ con resto di Lagrange si ottiene

$$x_{n+1} - \xi = \varphi(x_n) - \varphi(\xi) = \frac{\varphi^{(p)}(z_n)}{p!}\,(x_n - \xi)^p \qquad \exists z_n \in \operatorname{int}(x_n, \xi)$$

e, inserendo questo risultato nel limite, si conclude che

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n^p} = \lim_{n\to\infty} \frac{\left|\varphi^{(p)}(z_n)\right| \cancel{|x_n - \xi|^p}\cancel{p!}}{\cancel{|x_n - \xi|^p}\cancel{p!}} = \left|\varphi^{(p)}(z_n)\right| \in \mathbb{R}^+ \qquad \square$$

2.3.4 Il metodo di Newton come iterazione di punto fisso

Il metodo di Newton applicato ad una funzione f può essere visto come un'iterazione di punto fisso se si guarda la funzione $\varphi(x) \coloneqq x - \frac{f(x)}{f'(x)}$ con $f'(x) \neq 0$ per ogni x. Derivando φ, si ottiene

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

da cui segue che, se ξ è uno zero semplice di f, $\varphi'(\xi) = 0$. In queste condizioni si può applicare il Corollario 2.1 in un conveniente intorno di ξ, che in base al Teorema 2.3 è anche l'unico punto fisso di φ. Poiché $\varphi'(\xi) = 0$, l'ordine di convergenza del metodo è almeno 2; vogliamo trovare ora i casi in cui l'ordine è esattamente 2: poiché

$$\varphi''(\xi) = \frac{\big(f'(\xi)\, f''(\xi) + \cancel{f(\xi)\, f^{(3)}(\xi)}\big)\big(f'(\xi)\big)^2 - \cancel{2 f(\xi)\, f'(\xi)\big(f''(\xi)\big)^2}}{\big(f'(\xi)\big)^4} = \frac{f''(\xi)}{f'(\xi)}$$

l'ordine di convergenza è esattamente 2 se e solo se $f''(\xi) \neq 0$ e superiore se $f''(\xi) = 0$.

Metodo di Newton per zeri non semplici Sia $f\colon I \to \mathbb{R}$ una funzione di classe C^m, con $m \geq 2$, e sia $\xi \in I$ tale che $f^{(k)}(\xi) = 0$ per ogni $k \in \{\, 0, \ldots, m-1 \,\}$ e $f^{(m)}(\xi) \neq 0$. In questa situazione, quanto visto sopra diviene privo di senso, perché $f'(\xi) = 0$; occorre, dunque, trattare questo caso separatamente. Poiché $f \in C^m(I)$, si possono giustificare le seguenti espansioni di Taylor centrate in ξ:

$$\begin{aligned}
f(x) &= \frac{f^{(m)}(z)}{m!}(x-\xi)^m \qquad \exists z \in \operatorname{int}(x,\xi) \\
f'(x) &= \frac{f^{(m)}(z)}{(m-1)!}(x-\xi)^{m-1} \qquad \exists z \in \operatorname{int}(x,\xi) \\
f''(x) &= \frac{f^{(m)}(z)}{(m-2)!}(x-\xi)^{m-2} \qquad \exists z \in \operatorname{int}(x,\xi)
\end{aligned}$$

ed è sensato usare il medesimo z per tutte le espansioni perché nel primo caso f si può scrivere come serie di potenze in un certo intorno di ξ contenuto in I e in tale intorno f' è uguale alla derivata termine a termine dello sviluppo in serie. Considerando la funzione $\varphi(x) \coloneqq x - \frac{f(x)}{f'(x)}$, si può notare che

$$\begin{aligned}
\lim_{x\to\xi} \varphi'(x) = \lim_{x\to\xi} \frac{f(x)\, f''(x)}{\big(f'(x)\big)^2} &= \\
&= \lim_{x\to\xi} \frac{\left(\frac{f^{(m)}(z)}{m!}(x-\xi)^m\right)\left(\frac{f^{(m)}(z)}{(m-2)!}(x-\xi)^{m-2}\right)}{\left(\frac{f^{(m)}(z)}{(m-1)!}(x-\xi)^{m-1}\right)^2} = \\
&= \frac{m-1}{m} = 1 - \frac{1}{m} \in (\,0,1\,)
\end{aligned}$$

e, per la Proposizione 2.1, si conclude che il metodo ha ordine di convergenza esattamente 1. Se si modifica il metodo di Newton identificandolo con la contrazione $\psi(x) \coloneqq x - m\frac{f(x)}{f'(x)}$, un ragionamento analogo mostra che

$$\lim_{x\to\xi} \psi'(x) = \lim_{x\to\xi} 1 - m\frac{\big(f'(x)\big)^2 - f(x)\,f''(x)}{\big(f'(x)\big)^2} = \\ = 1 - m + m\lim_{x\to\xi} \frac{f(x)\,f''(x)}{\big(f'(x)\big)^2} = 1 - m + \cancel{m}\frac{m-1}{\cancel{m}} = 0$$

e, sempre per la Proposizione 2.1, il metodo ha ordine di convergenza almeno 2.

3 Interpolazione

3.1 Interpolazione polinomiale

Dato un insieme $\{\,(x_i, y_i) \mid i \in \{\,0,\dots,n\,\}, \forall i \neq j : x_i \neq x_j\,\}$, ci si chiede se sia possibile trovare una funzione f_n in uno spazio funzionale di dimensione finita tale che $f_n(x_i) = y_i$ per ogni $i \in \{\,0,\dots,n\,\}$. Poiché, in base al Teorema di densità di Weierstrass, i polinomii sono un insieme denso di $C([\,a,b\,])$ per ogni $[\,a,b\,] \subseteq \mathbb{R}$ compatto, una scelta possibile è lo spazio dei polinomii di grado al piú n, $\mathbb{P}_n \coloneqq \langle 1, x, \dots, x^n\rangle$: in questo caso, $f_n = \Pi_n$ si dice *polinomio interpolatore* e il problema si classifica come *interpolazione polinomiale*.

3.1.1 Costruzione dell'interpolatore

Sia, dunque, $f_n = \Pi_n = a_0 + \cdots + a_n x^n \in \mathbb{P}_n$; imporre che $f_n(x_i) = y_i$ per ogni $i \in \{\,0,\dots,n\,\}$ è equivalente a risolvere il sistema lineare

$$\begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} \tag{3.1}$$

Poiché la matrice a sinistra nella (3.1) è la trasposta di una matrice di Vandermonde, essa ha determinante non nullo e rango $n+1$, ossia il massimo; per il Teorema di Rouché-Capelli, il sistema nella (3.1) ha una ed una sola soluzione — in altre parole, il polinomio interpolatore esiste ed è unico.

Enunciamo questo risultato e diamone una dimostrazione alternativa nella seguente Proposizione.

Proposizione 3.1. *Dato l'insieme* $\{\,(x_i, y_i) \mid i \in \{\,0,\dots,n\,\}, \forall i \neq j : x_i \neq x_j\,\}$, *esiste ed è unico il polinomio* $\Pi_n \in \mathbb{P}_n$ *con* $\deg \Pi_n = n$ *tale che* $\Pi_n(x_i) = y_i$ *per ogni* $i \in \{\,0,\dots,n\,\}$.

Dimostrazione. Definiamo il *polinomio interpolatore di Lagrange*

$$\ell_i(x) \coloneqq \prod_{\substack{j=0 \\ i\neq j}}^{n} \frac{x - x_j}{x_i - x_j}$$

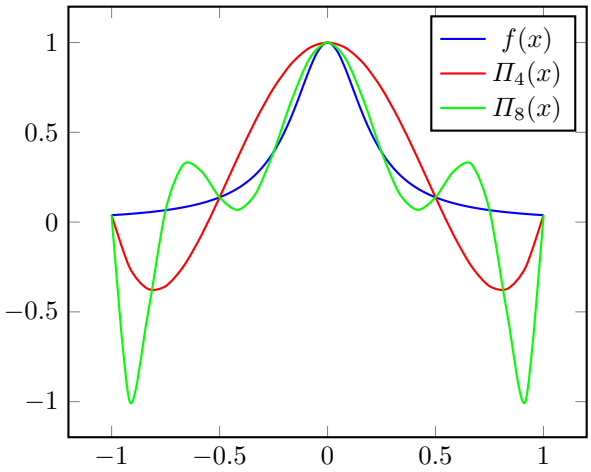

Figura 3.1: Confronto nell'intervallo $[\,-1,1\,]$ tra la funzione di Runge $f(x) \coloneqq \frac{1}{1+25x^2}$ (in blu) e due polinomii interpolatori ottenuti con nodi equispaziati .

esso vale 0 in tutti gli $x_j \neq x_i$ e 1 in x_i: il polinomio $y_i\,\ell_i(x)$, dunque, vale y_i in x_i e 0 negli $x_j \neq x_i$. Sommando tutti i polinomii interpolatori di Lagrange si ottiene

$$\Pi_n(x) = \sum_{i=0}^{n} y_i\,\ell_i(x) \tag{3.2}$$

Questo polinomio soddisfa alle condizioni richieste, perché

$$\Pi_n(x_j) = \sum_{i=0}^{n} y_i\,\ell_i(x_j) = \sum_{i=0}^{n} y_i\,\delta_{i,j} = y_j \qquad \forall j \in \{\,0,\dots,n\,\}$$

Per assurdo il polinomio interpolatore non sia unico: esistono, dunque, $p_1, p_2 \in \mathbb{P}_n$ che interpolano gli stessi (x_i, y_i); da ciò segue che $p_1(x_i) - p_2(x_i) = 0$ per ogni $i \in \{\,0,\dots,n\,\}$. Accade, dunque, che il polinomio $p_1 - p_2 \in \mathbb{P}_n$ ha esattamente $n+1$ zeri: ciò è possibile se e solo se $p_1 - p_2 = 0$, ossia se e solo se $p_1 = p_2$ — il che è assurdo. □

3.1.2 Errori in sup-norma

Benché il polinomio interpolatore sia unico a parità di grado, esso non sempre converge uniformemente alla funzione f: si consideri, ad esempio, la *funzione di Runge* illustrata nella Figura 3.1; per essa si ha addirittura $\|\Pi_n - f\|_\infty \to +\infty$ quando $n \to \infty$. La seguente Proposizione quantifica l'errore in sup-norma tra l'interpolante e la funzione.

Proposizione 3.2. *Data una funzione $f \in C^{n+1}([\,a,b\,])$, con $-\infty < a < b < +\infty$, si consideri l'insieme $\left\{\ (x_i, y_i = f(x_i)) \;\middle|\; i \in \{\,0,\dots,n\,\}, \forall i \neq j : x_i \neq x_j\ \right\}$. Definito $\omega(x) \coloneqq \prod_{i=0}^{n}(x - x_i)$, si ha*

$$E_n(x) \coloneqq f(x) - \Pi_n(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}\omega(x) \qquad \exists \eta \in \operatorname{int}(x, x_0, \dots, x_n) \tag{3.3}$$

Dimostrazione. Fissato $x \in [\,a,b\,] \setminus \{\,x_0,\ldots,x_n\,\}$, definiamo la funzione $G(z) \coloneqq E_n(z) - \omega(z)\frac{E_n(x)}{\omega(x)}$; G è di classe C^{n+1} perché somma di funzioni di classe C^{n+1}; si ha, inoltre, che $G(x) = 0 = G(x_i)$ per ogni $i \in \{\,0,\ldots,n\,\}$: G ha, dunque, $n+2$ zeri in $[\,a,b\,]$. Applicando il Teorema di Rolle ad ognuno degli $n+1$ intervalli chiusi delimitati da due suoi zeri, si deduce che G' si annulla in $n+1$ punti.

Procedendo per induzione sull'ordine k della derivata di G, si ha $G^{(k-1)} \in C^{n-k+1}([\,a,b\,])$; per ipotesi induttiva esistono $n+3-k$ punti in cui $G^{(k-1)}$ si annulla. Da ciò segue, sempre per il Teorema di Rolle, che esistono $n+2-k$ punti in cui $G^{(k)}$ si annulla. Ponendo $k = n+1$, si trova che $G^{(n+1)}$ si annulla in un punto $\eta \in \operatorname{int}(x,x_0,\ldots,x_n)$; sapendo che

$$G^{(k)}(z) = E^{(k)}(z) - \frac{E_n(x)}{\omega(x)}\omega^{(k)}(z) = f^{(k)}(z) - \mathit{\Pi}_n^{(k)}(z) - \frac{E_n(x)}{\omega(x)}\omega^{(k)}(z)$$

e sapendo che $\omega^{(n+1)}(z) = (n+1)!$ e che $\mathit{\Pi}_n^{(n+1)}(z) = 0$ per ogni z, si ottiene

$$G^{(n+1)}(\eta) = f^{(n+1)}(\eta) - \frac{E_n(x)}{\omega(x)}(n+1)! = 0 \implies E_n(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}\omega(x)$$

che è proprio la (3.3). □

Dalla Proposizione 3.2 segue facilmente che l'errore massimo compiuto dall'interpolazione è stimato da

$$\|E_n\|_\infty = \|f - \mathit{\Pi}_n\|_\infty \leq \left\|f^{(n+1)}\right\|_\infty \frac{(b-a)^{n+1}}{(n+1)!} \tag{3.4}$$

Mentre il fattore di destra è infinitesimo, nulla si può dire *a priori* del comportamento di $\left\|f^{(n+1)}\right\|$ senza ulteriori ipotesi sui nodi di interpolazione. Nel caso di nodi equisapziati, ad esempio, chiamato $h \coloneqq |x_1 - x_0|$ il *passo*, si può dimostrare che vale la stima

$$\|f - \mathit{\Pi}_n\|_\infty \leq M_{n+1}\frac{h^{n+1}}{4(n+1)} \tag{3.5}$$

da cui segue che il polinomio interpolatore converge uniformemente a f se e solo se $\left\|f^{(n)}\right\|_\infty \leq M$ per ogni $n \in \mathbb{N}$.

3.1.3 Nodi di Chebyshev

Perché il polinomio interpolatore converga alla funzione, occorre usare i *nodi di Chebyshev*, ossia nodi derivati dalla proiezione sull'asse delle ascisse di punti equidistanti di una semicirconferenza. Dato $n \in \mathbb{N}$, si possono costruire $n+1$ nodi di Chebyshev su un intervallo $[\,a,b\,]$ con la formula

$$x_i = \frac{b-a}{2}\cos\left(\frac{i\,\pi}{n}\right) + \frac{a+b}{2} \qquad \forall i \in \{\,0,\ldots,n\,\} \tag{3.6}$$

Come si può notare anche nella Figura 3.2, i nodi di Chebyshev non sono equispaziati, ma tendono ad infittirsi intorno agli estremi dell'intervallo su cui si costruiscono.

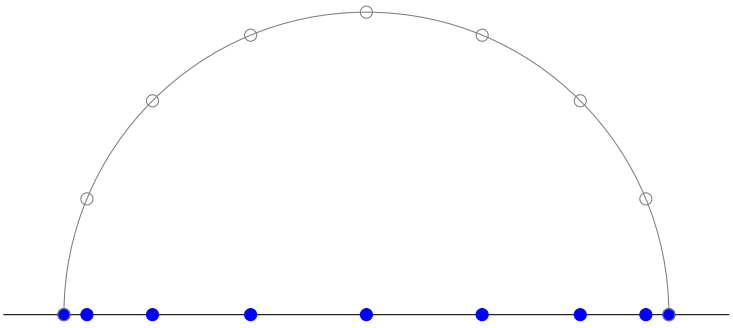

Figura 3.2: Rappresentazione di nove nodi di Chebyshev sull'intervallo $[\,0,1\,]$.

Si può dimostrare che, chiamato $\Pi_n^{\mathrm{Ch}}(x)$ il polinomio interpolatore costruito a partire da $n+1$ nodi di Chebyshev, per ogni funzione $f \in C^k([\,a,b\,])$, con $k > 0$, esiste c_k tale che

$$\left\|\Pi_n^{\mathrm{Ch}} - f\right\|_\infty \leq c_k \frac{\log n}{n^k} \tag{3.7}$$

e, quindi, si ha sempre convergenza uniforme per $n \to \infty$.

3.1.4 Stabilità dell'interpolazione

È interessante studiare come gli errori sperimentali sui dati influiscano sul polinomio interpolatore. Siano, dunque, $f \in C([\,a,b\,])$ una funzione da interpolare, Π_n il polinomio interpolatore e $\tilde{\Pi}_n$ un polinomio che approssimi l'interpolatore; fissato $\varepsilon > 0$, supponiamo che $|y_i - \tilde{y}_i| < \varepsilon$. La sup-norma della differenza tra i due polinomii è stimata da

$$\left\|\Pi_n - \tilde{\Pi}_n\right\|_\infty = \sup_{x\in[\,a,b\,]} \left|\sum_{i=0}^n y_i \ell_i(x) - \sum_{i=0}^n \tilde{y}_i \ell_i(x)\right| \leq \sum_{i=0}^n |y_i - \tilde{y}_i| \, \|\ell_i\|_\infty \leq \varepsilon \Lambda_n$$

ove $\Lambda_n := \sum_{i=0}^n \|\ell_i\|_\infty$ si dice *costante di Lebesgue*. Per com'è definita, essa dipende soltanto dai punti di interpolazione ed è asintotica almeno ad una funzione logaritmica $c \log n$ per $n \to \infty$. Si può dimostrare che $\Lambda_n \sim c\frac{2^n}{n \log n}$ per nodi d'interpolazione equispaziati e che $\Lambda_n \sim c \log n$ per nodi d'interpolazione di Chebyshev.

Esercizio 3.1. Dati $n+1$ nodi distinti in un intervallo $[\,a,b\,]$ si consideri la funzione

$$\begin{array}{rccc} L_n\colon & C([\,a,b\,]) & \to & \mathbb{P}_n \\ & f & \mapsto & \Pi_n \end{array}$$

che manda una funzione nel suo polinomio interpolatore di grado n. Dopo aver verificato che L_n è un operatore lineare, si dimsotri che è continuo verificando la stima

$$\|L_n f\|_\infty \leq \Lambda_n \|f\|_\infty$$

da cui segue che $\|L_n\| \leq \Lambda_n$; si ottenga da ciò la stima dell'errore

$$\|f - L_n f\|_\infty \leq (1 + \Lambda_n)\|f - p_n^*\|_\infty \tag{3.8}$$

ove p_n^* è il polinomio di *migliore approssimazione uniforme* di f in $\mathbb{P}_n$, ossia quel polinomio di grado al piú n tale che $\|f - p_n^*\|_\infty = \inf_{p\in\mathbb{P}_n} \|f - p\|_\infty$.

Soluzione. Chiamati $x_0, \dots, x_n$ i nodi d'interpolazione, siano $f, g \in C([\,a,b\,])$ — e indicheremo con f_n e g_n le loro immagini tramite L_n; poiché $(f+g)(x_i) = f(x_i) + g(x_i)$ per ogni $i \in \{\,0,\dots,n\,\}$, si ha che $(f+g)_n - f_n - g_n$ è un polinomio di grado al piú n con $n+1$ zeri, ossia è il polinomio nullo. Dato, poi, $\alpha \in \mathbb{R}$, è immediato che l'interpolatore della funzione αf è αf_n, visto che $(\alpha f)(x_i) = \alpha f(x_i)$ per ogni $i \in \{\,0,\dots,n\,\}$. Queste considerazioni provano che L_n è lineare.

Per la seconda parte si osservi che

$$\|L_n f\|_\infty = \sup_{x\in[\,a,b\,]} \left| \sum_{i=0}^{n} f(x_i)\,\ell_i(x) \right| \le \sum_{i=0}^{n} \sup_{x\in[\,a,b\,]} |f(x)|\,\|\ell_i\|_\infty = \Lambda_n \|f\|_\infty$$

e, riordinando e passando alla norma operatoriale, si trova $\|L_n\| \le \Lambda_n$.

Per l'ultima parte, si ha

$$\begin{aligned} \|f - L_n f\|_\infty \le \|f - p_n^*\|_\infty + \cancel{\|p_n^* - L_n p_n^*\|_\infty} + \|L_n p_n^* - L_n f\|_\infty = \\ = \|f - p_n^*\|_\infty + \|L_n(p_n^* - f)\|_\infty \le (1+\Lambda_n)\|f - p_n^*\|_\infty \end{aligned}$$

che stima l'errore in sup-norma come si voleva. □

3.2 Interpolazione polinomiale a tratti

Talvolta è piú conveniente interpolare una funzione f con una funzione definita a tratti, cosí da risentire meno di errori. Studiamo il caso in cui l'interpolante è una *polinomiale a tratti*.

3.2.1 Costruzione dell'interpolatore

Sia $f\colon [\,a,b\,] \to \mathbb{R}$ una funzione. Dato $s \in \mathbb{N} \setminus \{\,0\,\}$, sia n multiplo di s; si considerino sull'intervallo $[\,a,b\,]$ i nodi ordinati $a = x_0 < \cdots < x_n = b$; a partire da questi, si considerano i nodi $x_{ks}, \dots, x_{(k+1)s}$ per ogni $k \in \left\{\,0,\dots,\frac{n}{s}-1\,\right\}$ e si costruisce un polinomio interpolatore di grado s di f ristretta all'intervallo $[\,x_{ks}, x_{(k+1)s}\,]$. La funzione che risulta definita a tratti in ognuno di questi intervalli si indica con Π_s^{c}.

La funzione interpolante Π_s^{c} è continua su $[\,a,b\,]$: nei punti del tipo x_{ks}, con $k \in \left\{\,0,\dots,\frac{n}{s}-1\,\right\}$, infatti, si ha che entrambi i polinomii interpolatori "locali" danno lo stesso risultato pari a $f(x_{ks})$, dato che entrambi interpolano f usando x_{ks} come nodo interpolatore. La funzione interpolante, tuttavia, non è in generale derivabile in quei punti, perché non è detto che i due polinomii interpolatori "locali" abbiano uguale derivata in tali punti: in generale, dunque, nei nodi del tipo x_{ks} si trovano punti angolosi di Π_s^{c}.

3.2.2 Convergenza uniforme

Se $s = 1$, l'interpolante si dice *lineare a tratti*. Mostriamo che, in certe condizioni, tale interpolante converge uniformemente a f.

Proposizione 3.3. *Sia $f \in C^2([\,a,b\,])$; chiamati $x_0 < \cdots < x_n$ i nodi d'interpolazione come sopra, sia $h \coloneqq \max\{\,x_i - x_{i-1} \mid i \in \{\,1,\dots,n\,\}\,\}$. Allora la lineare a tratti Π_1^{c} ottenuta a partire da $x_0, \dots, x_n$ converge uniformemente a f con un errore $O(h^2)$.*

Dimostrazione. Siano $I_i \coloneqq [\, x_{i-1}, x_i \,]$ per ogni $i \in \{\, 1, \ldots, n \,\}$. Per definizione di sup-norma, è vero che $\|\Pi_1^{\mathrm{c}} - f\|_\infty = \max_{i \in \{\, 1,\ldots,n \,\}} \|\Pi_{1\,|I_i}^{\mathrm{c}} - f_{|I_i}\|_\infty$. Per ogni I_i, poiché Π_1^{c} interpola f in I_i, si può applicare la (3.3), sicché

$$f(x) - \Pi_1^{\mathrm{c}}(x) = \frac{f''(z)}{2}(x - x_{i-1})(x - x_i) \qquad \exists z \in \mathrm{int}(x, x_{i-1}, x_i)$$

e tale scrittura ha senso perché $f \in C^2([\, a, b \,])$. Da ciò segue che

$$\|f - \Pi_1^{\mathrm{c}}\|_\infty = \max_{i \in \{\, 1,\ldots,n \,\}} \max_{x \in I_i} \frac{f''(z)}{2}(x - x_{i-1})(x - x_i) \leq \frac{\|f''\|_\infty}{2} h^2 \qquad \square$$

Un risultato piú generale e preciso della Proposizione 3.3 è dato dal seguente Teorema.

Teorema 3.1. *Sia $f \in C^{s+1}([\, a, b \,])$ e sia Π_s^{c} una sua interpolante polinomiale a tratti; allora $\|f - \Pi_s^{\mathrm{c}}\|_\infty$ è $O(h^{s+1})$. Nel caso di nodi d'interpolazione equispaziati, inoltre, vale la stima*

$$\|f - \Pi_s^{\mathrm{c}}\|_\infty \leq \frac{\left\|f^{(s+1)}\right\|_\infty}{4(s+1)} h^{s+1} \tag{3.9}$$

Dimostrazione. Proviamo solo il caso di nodi equispaziati, di modo che $h = (b-a)/n$. Siano $I_j \coloneqq [\, x_{js}, x_{(j+1)s} \,]$ per ogni $j \in \left\{\, 0, \ldots, \frac{n}{s} - 1 \,\right\}$. Si ha, usando la (3.5), che

$$\|f - \Pi_s^{\mathrm{c}}\|_\infty = \max_{j \in \left\{\, 0,\ldots,\frac{n}{s}-1 \,\right\}} \max_{x \in I_j} |f(x) - \Pi_s^{\mathrm{c}}(x)| \leq \frac{\left\|f^{(s+1)}\right\|_\infty}{4(s+1)} h^{s+1}$$

e, poiché $f \in C^{s+1}([\, a, b \,])$, il membro a destra della diseguaglianza è limitato. $\square$

Localmente sugli intervalli I_j, inoltre, vale la seguente diseguaglianza per le lineari a tratti:

$$\|f - \Pi_1^{\mathrm{c}}\|_\infty \leq \mathrm{osc}_{I_j}(f, h) \coloneqq \max_{\substack{x,y \in I_j \\ |x-y| \leq h}} |f(x) - f(y)|$$

Posto, poi, che $f \in C^{s+1}([\, a, b \,])$, se alla polinomiale a tratti si sostituisce una sua versione approssimata $\tilde{\Pi}_s^{\mathrm{c}}$, si ottiene

$$\|f - \tilde{\Pi}_s^{\mathrm{c}}\|_\infty \leq \|f - \Pi_s^{\mathrm{c}}\|_\infty + \|\Pi_s^{\mathrm{c}} - \tilde{\Pi}_s^{\mathrm{c}}\|_\infty \leq c_s h^{s+1} + \varepsilon \Lambda_s$$

3.2.3 Interpolazione *spline*

Si può notare, a partire da quanto visto in questa sezione, che l'interpolazione polinomiale a tratti garantisce convergenza uniforme con "pochi" prerequisiti, tra cui $h \to 0$ per $n \to \infty$. Il suo svantaggio piú grande, tuttavia, è la non derivabilità della funzione interpolante nei nodi di interpolazione. Per risolvere questo problema, si ricorre alle interpolazioni *spline*, che uniscono interpolazioni a tratti di modo che nei nodi di interpolazione siano definite le derivate di un certo grado.

Data una funzione f su un intervallo $[\, a, b \,]$, si dice *spline* di grado k una funzione $S_k \in C^{k-1}([\, a, b \,])$ tale che

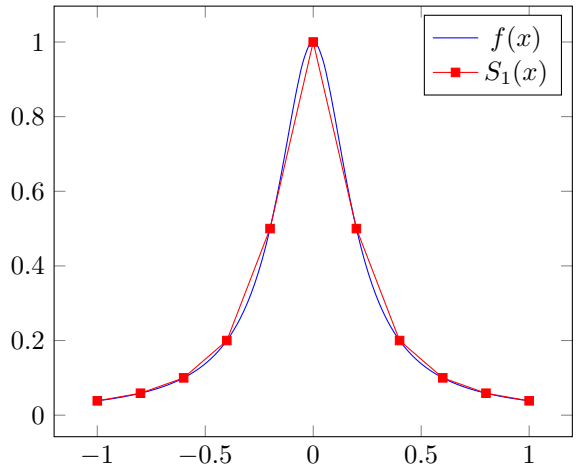

Figura 3.3: Approssimazione lineare a tratti della funzione di Runge.

- $S_k(x_i) = f(x_i)$ per ogni $i \in \{ 0, \dots, n \}$;
- $S_{k|[x_i, x_{i+1}]}$ sia un polinomio di grado al piú k.

Se $k = 1$, le *spline* corrispondenti sono lineari a tratti: esse sono continue, ma non sono derivabili nei nodi di interpolazione. Ponendo $k = 3$, invece, si ottengono le *spline* cubiche, ossia funzioni localmente cubiche con derivate fino al secondo grado nei nodi di interpolazione. Occorre ricordare, però, che, mentre le *spline* sono interpolanti di f, le loro derivate non interpolano le derivate di f, perché non sono costruite in modo tale da garantire l'interpolazione.

Poiché ad ogni intervallo $[\, x_i, x_{i+1} \,]$ si associa una polinomiale di grado k, una *spline* di grado k ha $(k+1)n$ parametri da determinare; le condizioni poste per ipotesi, invece, sono $k(n-1)+2$, perché si impone l'interpolazione di x_0 e x_n e poi si chiede che le polinomiali definite intorno ad un nodo x_i diverso dagli estremi abbiano valore e derivate fino alla $(k-1)$-esima uguali — e ci sono $n-1$ nodi da trattare in questa maniera. In generale, dunque, non esiste una soluzione unica di questo problema senza porre ulteriori condizioni. Nel caso delle *spline* cubiche si hanno $4n$ parametri e $4n-2$ equazioni; solitamente si pone uno di questi due vincoli aggiuntivi (che aggiungono due equazioni):

- $S_3''(x_0) = 0 = S_3''(x_n)$;
- $S_3^{(3)}$ esiste ed è continua in x_1 e x_{n-1}.

Non dimostriamo il fatto che, se $f \in C^4([\, a, b \,])$, allora la *spline* cubica S_3 verifica la proprietà

$$\left\| f^{(j)} - S_3^{(j)} \right\|_\infty = O(h^{4-j}) \qquad \forall j \in \{ 0, 1, 2, 3 \} \tag{3.10}$$

3.3 Approssimazione polinomiale ai minimi quadrati

In alcune situazioni, il numero dei nodi a disposizione è molto alto e, come abbiamo visto, ciò potrebbe compromettere la stabilità dell'interpolazione; in

altri casi, invece, si vuole "regolarizzare" un fenomeno irregolare oppure rimuovere del "rumore" dovuto all'inaccuratezza dei dati sperimentali. In questi casi si ricorre ad un metodo alternativo all'interpolazione e, in un certo senso, "piú generale", ossia l'*approssimazione ai minimi quadrati*.[5]

Data una funzione $f\colon [\,a,b\,] \to \mathbb{R}$ e fissato $N \in \mathbb{N}$, si consdierino le coppie (x_i, y_i), con $i \in \{\,1,\dots,N\,\}$, tali che $x_i \in [\,a,b\,]$ e $y_i = f(x_i)$ per ogni i. Scelto $m \in \mathbb{N}$ tale che $m < N$,[6] cerchiamo un polinomio di grado al piú m, $p \in \mathbb{P}_m$, tale che la *somma degli scarti quadratici*

$$\sum_{i=1}^{N} \big(p(x_i) - y_i\big)^2$$

sia minima. Il problema, dunque, è cercare il valore

$$\min_{p\in\mathbb{P}_m} \sum_{i=1}^{N} \big(p(x_i) - y_i\big)^2 = \min_{a\in\mathbb{R}^{m+1}} \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{m} a_j x_i^j \right)^2 =: \min_{a\in\mathbb{R}^{m+1}} \Phi(a)$$

con Φ polinomio di secondo grado in $m+1$ variabili reali $a_0,\dots,a_m$. Nel caso in cui $m = 1$, $\Phi(a)$ si dice *retta dei minimi quadrati*.

Definendo la matrice

$$V \coloneqq \begin{pmatrix} 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & \cdots & x_N^m \end{pmatrix}$$

si osserva che $\Phi(a) = \|Va - y\|_2^2$, ove y è la colonna degli y_i; da questa nuova prospettiva si nota che, se $m = N-1$, il problema è equivalente ad un problema di interpolazione polinomiale. Poiché trovare una soluzione del sistema lineare $Va - y = 0$ è impossibile se il sistema è sovradeterminato, cerchiamo un $a \in \mathbb{R}^{m+1}$ tale che $\|Va - y\|_2^2$ sia minimo.

Teorema 3.2. *Un vettore $a \in \mathbb{R}^{m+1}$ è di minimo per Φ se e solo se è soluzione del* sistema delle equazioni normali

$$V^{\mathrm{t}} V a = V^{\mathrm{t}} y$$

Dimostrazione. Osserviamo che a è di minimo per Φ se e solo se $\Phi(a+b) \geq \Phi(a)$ per ogni $b \in \mathbb{R}^{m+1}$. Applicando la definizione di norma euclidea, è immediato che $\Phi(x) = \|Vx - y\|_2^2 = \langle Vx - y, Vx - y\rangle$ per ogni $x \in \mathbb{R}^{m+1}$. Da ciò segue che

$$\begin{aligned} \Phi(a+b) &= \\ &= \langle V(a+b) - y, V(a+b) - y\rangle = \langle Va - y + Vb, Va - y + Vb\rangle = \\ &= \langle Va - y, Va - y\rangle + \langle Va - y, Vb\rangle + \langle Vb, Va - y\rangle + \langle Vb, Vb\rangle = \\ &= \Phi(a) + 2\,\langle Vb, Va - y\rangle + \|Vb\|_2^2 = \\ &= \Phi(a) + 2\,\big\langle b, V^{\mathrm{t}}(Va - y)\big\rangle + \|Vb\|_2^2 \end{aligned}$$

Dimostriamo, ora, entrambe le implicazioni.

[5] Nella letteratura scientifica anglofona, questo metodo è spesso indicato con la sigla LS, ovvero *Least Squares*.

[6] In realtà, nelle applicazioni si richiede che $m \ll N$.

($\Leftarrow$) Se $V^{\mathrm{t}}Va = V^{\mathrm{t}}y$, allora $V^{\mathrm{t}}(Va - y) = 0$; da ciò segue che $\Phi(a + b) = \Phi(a) + \|Vb\|_2^2 \geq \Phi(a)$ per ogni $b \in \mathbb{R}^{m+1}$.

($\Rightarrow$) Se a è di minimo per Φ, allora $2\langle b, V^{\mathrm{t}}(Va - y)\rangle + \|Vb\|_2^2 \geq 0$ per ogni $b \in \mathbb{R}^{m+1}$; posto $b = \varepsilon u$, con $\|u\|_2 = 1$ e $\varepsilon > 0$, si ha

$$2\left\langle \varepsilon u, V^{\mathrm{t}}(Va - y)\right\rangle + \|V(\varepsilon u)\|_2^2 \geq 0 \implies$$
$$\implies 2\cancel{\varepsilon}\left\langle u, V^{\mathrm{t}}(Va - y)\right\rangle + \varepsilon^{\cancel{2}}\|Vu\|_2^2 \geq 0$$

Passando al limite per $\varepsilon \to 0^+$, si ottiene $\langle u, V^{\mathrm{t}}(Va - y)\rangle \geq 0$ per ogni versore u; ma anche $-u$ è un versore, perciò anche $\langle -u, V^{\mathrm{t}}(Va - y)\rangle \geq 0$: da ciò segue che $\langle u, V^{\mathrm{t}}(Va - y)\rangle = 0$ per ogni versore u. Si ottiene, quindi, che $V^{\mathrm{t}}(Va - y) = 0$, ossia che a risolve il sistema delle equazioni normali. □

Se la matrice V nel Teorema 3.2 ha rango massimo, la soluzione del sistema delle equazioni normali è unica. Poiché, infatti, $V^{\mathrm{t}}V$ è una matrice reale quadrata simmetrica, si ha $\langle V^{\mathrm{t}}Vu, u\rangle = 0$ se e solo se $u \in \ker V = \langle 0\rangle$. La matrice V ha rango massimo quando almeno $m + 1$ nodi sono distinti, in quanto il minore $(m + 1) \times (m + 1)$ di V ottenuto a partire dalle righe relative a tali punti ha determinante non nullo perché è una matrice di Vandermonde trasposta.

Riportiamo senza dimostrazione che, chiamato L_m l'approssimante ai minimi quadrati di grado m di una certa funzione $f\colon [a, b] \to \mathbb{R}$, se esiste $\vartheta \in (0, 1)$ tale che $h \coloneqq \max_{i\in\{1,\dots,N\}} x_i - x_{i-1} \leq \vartheta(b - a)/m^2$, allora per ogni $k \in \mathbb{N} \setminus \{0\}$ esiste $c_k > 0$ tale che, se $f \in C^k([a, b])$, allora $\|L_m - f\|_\infty \leq c_k m^{1-k}$.

4 Integrazione e derivazione numeriche

4.1 Formule di quadratura, ossia integrazione numerica

Le *formule di quadratura* sono metodi di calcolo approssimato dell'integrale di una funzione f su un intervallo $[a, b]$ ottenuto integrando una funzione $\tilde{f}$ che approssimi f. Ciò risulta agevolato dal fatto che

$$\left|\int_a^b f(x)\,\mathrm{d}x - \int_a^b \tilde{f}(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x) - \tilde{f}(x)|\,\mathrm{d}x \leq$$
$$\leq \|f - \tilde{f}\|_\infty \int_a^b \mathrm{d}x = (b - a)\|f - \tilde{f}\|_\infty \quad (4.1)$$

ossia che l'operatore integrale è stabile se il metodo con cui si costruisce $\tilde{f}$ è convergente e stabile in sup-norma. Se, ad esempio, $(f_n)_{n\in\mathbb{N}}$ è una successione di funzioni Riemann-integrabili che converge uniformemente a f, il membro di destra della diseguaglianza sopra tende a 0 per $n \to \infty$. Ha senso, dunque, studiare formule "facili" per integrare le funzioni interpolanti.

Se Π_n è un polinomio interpolatore di grado n per f, si ha

$$\int_a^b \Pi_n(x)\,\mathrm{d}x = \int_a^b \sum_{i=0}^n f(x_i)\,\ell_i(x)\,\mathrm{d}x = \sum_{i=0}^n f(x_i)\int_a^b \ell_i(x)\,\mathrm{d}x$$

e, indicando con w_i gli integrali $\int_a^b \ell_i(x)\,\mathrm{d}x$, si nota che l'integrale dei polinomii interpolatori ha la forma di una somma pesata. Questo risultato si può facilmente estendere alle polinomiali a tratti, dato che esse sono localmente polinomii interpolatori. Per la (3.9), poi, definito $h \coloneqq \max_{i\in\{\,1,\dots,n\,\}}(x_{i+1}-x_i)$, si ottiene che

$$\int_a^b f(x) - \mathit{\Pi}_s^{\mathrm{c}}(x)\,\mathrm{d}x = O(h^{s+1})$$

se $f \in C^{s+1}([\,a,b\,])$; la formula ha senso se e solo se n è un multiplo di s. Le formule di quadratura ottenute con una polinomiale a tratti a passo costante, ossia usando nodi equispaziati, si dicono *formule di Newton-Côtes*.

4.1.1 Formule dei trapezii e delle parabole

Sia ora $\mathit{\Pi}_1^{\mathrm{c}}$ un'interpolante lineare a tratti che interpola nei nodi equispaziati $a = x_0 < \cdots < x_n = b$ una funzione f. Volendo calcolarne l'integrale, si ha

$$\begin{aligned}
\int_a^b \mathit{\Pi}_1^{\mathrm{c}}(x)\,\mathrm{d}x &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \mathit{\Pi}_1^{\mathrm{c}}(x)\,\mathrm{d}x = \\
&= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}(x-x_i) + f(x_i)\,\mathrm{d}x = \\
= \sum_{i=0}^{n-1} \frac{f(x_{i+1})-f(x_i)}{2(x_{i+1}-x_i)}\big(x_{i+1}^2 - x_i^2\big) &+ \left(f(x_i) - \frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}\right)(x_{i+1}-x_i) = \\
= \sum_{i=0}^{n-1} \frac{1}{2}(x_{i+1}+x_i)\big(f(x_{i+1})-f(x_i)\big) &+ (x_{i+1}-x_i)f(x_i) - x_i\big(f(x_{i+1})-f(x_i)\big) = \\
&= \frac{1}{2}\sum_{i=0}^{n-1}\big(f(x_i)+f(x_{i+1})\big)(x_{i+1}-x_i)
\end{aligned}$$

da cui segue, posto $h \coloneqq x_1 - x_0$, la *formula dei trapezii*

$$\int_a^b \mathit{\Pi}_1^{\mathrm{c}}(x)\,\mathrm{d}x = \frac{\big(f(a)+f(b)\big)h}{2} + \sum_{i=1}^{n-1} h f(x_i) \tag{4.2}$$

Analogamente, sia $\mathit{\Pi}_2^{\mathrm{c}}$ un'interpolante quadratica a tratti che interpola nei nodi equispaziati $x_0,\dots,x_n$, con n pari, una funzione f. Similmente a prima, posto $h \coloneqq x_2 - x_0 = 2(x_1 - x_0)$, si ottiene una formula, detta *formula delle parabole* o *di Cavalieri-Simpson*[7]

$$\int_a^b \mathit{\Pi}_2^{\mathrm{c}}(x)\,\mathrm{d}x = \frac{h}{3}\big(f(a)+f(b)\big) + \sum_{i=1}^{(n-2)/2}\left(\frac{2h}{3}f(x_{2i}) + \frac{4h}{3}f(x_{2i+1})\right) \tag{4.3}$$

4.1.2 Convergenza della quadratura

Dalla stima (4.1) segue che una formula di quadratura è convergente se e solo se è convergente il metodo con cui si crea $\tilde{f}$. Nel caso di polinomii interpolatori

[7] La dimostrazione è facile nei concetti, perché quasi identica a quella dei trapezii, ma le espressioni intermedie da scrivere sono troppo grandi per questo formato.

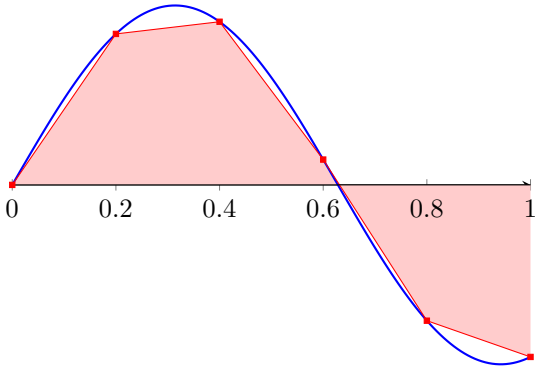

Figura 4.1: Rappresentazione grafica del metodo dei trapezii per il calcolo di $\int_0^1 \sin 5x \, \mathrm{d}x$.

costruiti con nodi di Chebyshev su una funzione di classe almeno $C^1([\,a,b\,])$ la convergenza dell'integrale è garantita, perché

$$\int_a^b f(x)\,\mathrm{d}x - \int_a^b \Pi_n^{\mathrm{Ch}}(x)\,\mathrm{d}x \le \int_a^b \left|f(x) - \Pi_n^{\mathrm{Ch}}(x)\right| \mathrm{d}x \le$$
$$\le \left\|f - \Pi_n^{\mathrm{Ch}}\right\|_\infty \int_a^b \mathrm{d}x = (b-a)\left\|f - \Pi_n^{\mathrm{Ch}}\right\|_\infty$$

e, se $f \in C^1([\,a,b\,])$, tale quantità tende a 0 quando $n \to \infty$. Inoltre, poiché $\left|I(f) - I(\Pi_n^{\mathrm{Ch}})\right|$ è $O(\log n \, \|f - p_n^*\|_\infty)$ e vale la stima $\left\|f - \Pi_n^{\mathrm{Ch}}\right\|_\infty \le (1 + \Lambda_n)\|f - p_n^*\|_\infty$, per una funzione $f \in C^k([\,a,b\,])$ vale che $\left|I(f) - I(\Pi_n^{\mathrm{Ch}})\right|$ è $O(n^{-k}\log n)$. Nel caso in cui $f \notin C^k([\,a,b\,])$ per ogni $k \neq 0$, l'interpolante sui nodi di Chebyshev può non convergere, perché $\|f - p_n^*\|_\infty$ può non avere un ordine di infinitesimo sufficiente ad azzerare l'$O(\log n)$ della costante di Lebesgue, ma l'integrale approssimato converge ugualmente, perché i pesi della corrispodente formula di quadratura sono tutti positivi.

Le formule di quadratura composte sulle polinomiali a tratti sono sempre convergenti: poiché $|I(f) - I(\Pi_s^{\mathrm{c}})| \le (b-a)\|f - \Pi_s^{\mathrm{c}}\|_\infty$, se $f \in C^{s+1}([\,a,b\,])$ allora $|I(f) - I(\Pi_s^{\mathrm{c}})|$ è $O(h^{s+1})$, come la parte destra della stima.

Il seguente Teorema fornisce condizioni necessarie e sufficienti affinché una formula di quadratura sia convergente.

Teorema 4.1 (Polya-Steklov)**.** *Fissata una funzione $f \in C([\,a,b\,])$, siano dati $\{\,(x_i, f(x_i)) \mid i \in \{\,0,\dots,n\,\}\,\}$ con $n \in \mathbb{N}$. L'operatore*

$$\begin{array}{rccc} I_n\colon & C([\,a,b\,]) & \to & \mathbb{R} \\ & f & \mapsto & \displaystyle\sum_{i=0}^{n} w_i\, f(x_i) \end{array}$$

è continuo e lineare; converge, inoltre, all'operatore integrale $I(f) \coloneqq \int_a^b f(x)\,\mathrm{d}x$ se e solo se

1. $\lim_{n\to\infty} I_n(p) = I(p)$ *per ogni polinomio p;*

2. esiste $k > 0$ tale che $S_n \leq k$ per ogni $n \in \mathbb{N}$.

Dimostrazione. La doppia implicazione segue rispettivamente dai Teoremi di densità di Weierstrass ($\Leftarrow$) e di Banach-Steinhaus ($\Rightarrow$).[8] Dimostriamo, invece, che l'operatore I_n è continuo. Per $n \in \mathbb{N}$ fissato, si ha

$$|I_n(f)| = \left| \sum_{i=0}^{n} w_i \, f(x_i) \right| \leq \sum_{i=0}^{n} |w_i| \, |f(x_i)| \leq \|f\|_\infty \sum_{i=0}^{n} |w_i| = \|f\|_\infty \, S_n$$

da cui segue che $\|I_n\| \leq S_n$ per ogni $n \in \mathbb{N}$. Per come è definito I_n, è chiaramente un operatore lineare: da ciò segue che, essendo limitato, è continuo. □

4.1.3 Stabilità della quadratura

Le somme dei pesi delle formule di quadratura $S_n \coloneqq \sum_{i=0}^{n} |w_i|$ svolgono un ruolo analogo alla costante di Lebesgue Λ_n nel regolare la "risposta alle perturbazioni" sui dati sperimentali. Si ha, infatti,

$$\left| \sum_{i=0}^{n} w_i f(x_i) - \sum_{i=0}^{n} w_i \tilde{f}(x_i) \right| \leq \sum_{i=0}^{n} |w_i| \, |f(x_i) - \tilde{f}(x_i)| \leq S_n \|f - \tilde{f}\|_\infty \tag{4.4}$$

Grazie al Teorema 4.1 si può affermare che formule di quadratura a pesi strettamente positivi sono sempre stabili. Dalla (4.4), infatti, si vede che

$$\left| \sum_{i=0}^{n} w_i f(x_i) - \sum_{i=0}^{n} w_i \tilde{f}(x_i) \right| \leq \varepsilon \, S_n$$

ove $\varepsilon \geq \|f - \tilde{f}\|_\infty$; poiché le formule di quadratura, sia algebriche sia composte, commettono un errore nullo sulle funzioni costanti, dall'uguaglianza

$$S_n = \sum_{i=0}^{n} |w_i| = \sum_{i=0}^{n} w_i = \int_a^b \mathrm{d}x = b - a$$

si ottiene che S_n è limitata, ossia la seconda condizione che occorre per applicare il Teorema 4.1. Poiché le formule algebriche e composte ne soddisfano la prima condizione, esse sono convergenti e stabili se hanno solo pesi positivi.[9] Le formule algebriche a passo costante, ossia quelle di Newton-Côtes, possono avere pesi negativi e, per $n \to \infty$, divergono dall'integrale.

Definiamo poi il *grado di esattezza polinomiale* di un metodo di quadratura come

$$m_n \coloneqq \max \{ \, d \in \mathbb{N} \mid \forall p \in \mathbb{P}_d : I_n(p) = I(p) \, \} \tag{4.5}$$

Si ha sempre $n \leq m_n \leq 2n + 1$. La formula dei trapezii ha grado di esattezza pari a 1. Le formule gaussiane hanno il massimo grado di esattezza, pari a $2n+1$.

[8] Il Teorema di densità di Weierstrass asserisce che, data $f \colon [a, b] \to \mathbb{R}$ continua, per ogni $\varepsilon > 0$ esiste un polinomio p tale che $\|f - p\|_\infty < \varepsilon$. Il Teorema di Banach-Steinhaus afferma quanto segue: dati X spazio di Banach e Y spazio normato, sia F una famiglia di operatori lineari continui limitati da X in Y tale che per ogni $x \in X$ si abbia $\sup \{ \, \|Tx\|_Y \mid T \in F \, \} < \infty$; allora $\sup \{ \, \|T\| \mid T \in F \, \} < \infty$, ove $\|T\|$ indica la norma operatoriale di T.

[9] Formule a pesi positivi sono quella dei trapezii, quella di Cavalieri-Simpson, l'integrazione sul polinomio interpolatore costruito su nodi di Chebyshev e le formule gaussiane.

Dalla stima (3.8) segue che

$$|I(f) - I(\Pi_n)| \leq (b-a)(1+\Lambda_n)\|f - p_n^*\|_\infty$$

Essa, tuttavia, è una sovrastima dell'errore di quadratura; per migliorarla occorre dimostrare il seguente Teorema.

Teorema 4.2 (Stieltjes)**.** *Per una formula algebrica di quadratura a pesi positivi vale la stima*

$$|I(f) - I_n(f)| \leq 2(b-a)\|f - p_n^*\|_\infty \tag{4.6}$$

Dimostrazione. Si ha

$$\begin{aligned}
|I(f) - I_n(f)| &= \big|I(f) - I(p_n^*) + \cancel{I(p_n^*)} - \cancel{I_n(p_n^*)} + I_n(p_n^*) - I_n(f)\big| \leq \\
&\leq |I(f-p_n^*)| + |I(f-p_n^*)| = \left|\int_a^b f(x) - p_n^*(x)\,\mathrm{d}x\right| + \left|\sum_{i=0}^n w_i\,(p_n^* - f)(x_i)\right| \leq \\
&\leq \|f-p_n^*\|_\infty \int_a^b \mathrm{d}x + S_n\|p_n^* - f\|_\infty = 2(b-a)\|f-p_n^*\|_\infty
\end{aligned}$$

come richiesto. □

4.2 Calcolo di derivate approssimato, ovvero derivazione numerica

Data una funzione $f \in C^1([\,a,b\,])$, si può mostrare che l'operatore derivata $\mathrm{D}\,f = f'$ è lineare ma non continuo. Si consideri, ad esempio, la successione di funzioni $f_n(x) \coloneqq \sin(nx)/n$ per $x \in [\,0,1\,]$; benché $\lim_{n\to\infty}\|f_n\|_\infty = \lim_{n\to\infty} 1/n = 0$, si ha $\lim_{n\to\infty}\|f_n'\|_\infty = \lim_{n\to\infty}\sup_{x\in[\,0,1\,]}|\cos(nx)| = 1 \neq 0$. Da ciò si nota che funzioni arbitrariamente "vicine" secondo la sup-norma possono avere derivate anche molto "distanti". Ciò prova l'instabilità dell'operatore D.

4.2.1 Rapporto incrementale "classico"

Nel rapporto incrementale "classico"

$$\delta_+(h) = \frac{f(x+h) - f(x)}{h}, \qquad h > 0 \tag{4.7}$$

se f è derivabile in x, esso tende a $f'(x)$ con $h \to 0$. Supponendo f di classe C^2 in un certo intorno di x, vi si ha che $\delta_+(h) = f'(x) + O(h)$, perché espandendo con la formula di Taylor si trova $f(x+h) = f(x) + f'(x)h + \frac{f''(\xi)}{2}h^2$ con $\xi \in \mathrm{int}(x, x+h)$ e, quindi, $\delta_+(h) = f'(x) + \frac{f''(\xi)}{2}h$. Se $\tilde{\delta}_+$ approssima δ_+, si ha la stima dell'errore assoluto

$$|f'(x) - \tilde{\delta}_+(h)| \leq |f'(x) - \delta_+(h)| + |\delta_+(h) - \tilde{\delta}_+(h)|$$

Si ha, poi, che

$$\begin{aligned}
|\delta_+(h) - \tilde{\delta}_+(h)| &= \frac{f(x+h) - f(x) - \tilde{f}(x+h) + \tilde{f}(x)}{h} = \\
&= \frac{1}{h}\Big(\big(f(x+h) - \tilde{f}(x+h)\big) + \big(\tilde{f}(x) - f(x)\big)\Big) \leq \\
&\leq \frac{1}{h}\Big(|f(x+h) - \tilde{f}(x+h)| + |\tilde{f}(x) - f(x)|\Big) \leq \frac{2}{h}\|f - \tilde{f}\|_\infty
\end{aligned}$$

Ponendo $\varepsilon \geq \|f - \tilde{f}\|_\infty$ in un intorno di x, otteniamo che

$$|f'(x) - \tilde{\delta}_+(h)| \leq c\,h + \frac{2\varepsilon}{h} =: E(h), \qquad c = \frac{f''(\xi)}{2}$$

Se $h \to 0$, l'errore $E(h)$ tende a $+\infty$: non potendo azzerare l'errore in questo modo, occorre cercare il *passo ottimale*, ossia il punto h^* di minimo per E. Poiché $E'(h) = c - \frac{2\varepsilon}{h^2}$, il suo punto di minimo è $h^* = \sqrt{\frac{2\varepsilon}{c}} = O(\sqrt{\varepsilon})$; la derivata seconda $E''(h) = 4\varepsilon\, h^{-3} > 0$ mostra che E è strettamente convessa, ossia che l'errore è "grande" per h "distanti" da h^*. L'errore minimo che si può commettere è, dunque,

$$E(h^*) = c\,h^* + \frac{2\varepsilon}{h^*} = c\sqrt{\frac{2\varepsilon}{c}} + \frac{2\varepsilon}{\sqrt{\frac{2\varepsilon}{c}}} = 2\sqrt{2c}\sqrt{\varepsilon} = O(\sqrt{\varepsilon})$$

4.2.2 Rapporto incrementale simmetrico

Sia f una funzione di classe C^3 in un intorno di x e sia $h > 0$. Per le formule di Taylor si ha

$$\begin{cases} f(x+h) = f(x) + f'(x)\,h + \frac{f''(x)}{2}h^2 + \frac{f^{(3)}(\xi)}{6}h^3 & \exists \xi \in \text{int}(x, x+h) \\ f(x-h) = f(x) - f'(x)\,h + \frac{f''(x)}{2}h^2 - \frac{f^{(3)}(\eta)}{6}h^3 & \exists \eta \in \text{int}(x, x-h) \end{cases}$$

la cui differenza è

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

Da questo risultato definiamo il *rapporto incrementale simmetrico*

$$\delta(h) \coloneqq \frac{f(x+h) - f(x-h)}{2h} = f'(x) + O(h^2) \tag{4.8}$$

Come per il rapporto incrementale usuale, sia $\tilde{\delta}$ un approssimante di δ. Se $|f'(x) - \delta(h)| \leq kh^2$, si ottiene la stima

$$|f'(x) - \tilde{\delta}(h)| \leq kh^2 + |\delta(h) - \tilde{\delta}(h)|$$

Dal momento che

$$\begin{aligned} |\delta(h) - \tilde{\delta}(h)| = \left| \frac{f(x+h) - f(x-h)}{2h} - \frac{\tilde{f}(x+h) - \tilde{f}(x-h)}{2h} \right| \leq \\ \leq \frac{1}{2h}|f(x+h) - \tilde{f}(x+h)| + \frac{1}{2h}|f(x-h) - \tilde{f}(x-h)| \leq \frac{\varepsilon}{h} \end{aligned}$$

si ha

$$|f'(x) - \tilde{\delta}(h)| \leq kh^2 + \frac{\varepsilon}{h} =: E(h)$$

Volendo trovare il minimo per E, si osservi che la derivata $E'(h) = 2kh - \frac{\varepsilon}{h^2}$ si annulla in $h^* = \sqrt[3]{\frac{\varepsilon}{2k}} = O(\varepsilon^{1/3})$. In base alla derivata seconda $E''(h) = 2k + 2\varepsilon h^{-3} > 0$, si può dire che E è strettamente convessa, da cui segue che il suo minimo è

$$E(h^*) = k\left(\frac{\varepsilon}{2k}\right)^{2/3} + \frac{\varepsilon}{\sqrt[3]{\frac{\varepsilon}{2k}}} = O(\varepsilon^{2/3})$$

4.3 Estrapolazione, ovvero struttura asintotica dell'errore

Sia $\varphi(h)$ una formula che approssima una certa quantità α con un errore tale che

$$\varphi(h) = \alpha + c\,h^p + O(h^q) \qquad q > p$$

ad esempio, i rapporti incrementali hanno una struttura del tipo

$$\begin{aligned}
\delta_+(h) &= f'(x) + O(h) \qquad f \in C^2 \\
\delta_+(h) &= f'(x) + \frac{f''(x)}{2}h + O(h^2) \qquad f \in C^3 \\
\delta(h) &= f'(x) + O(h) \qquad f \in C^3 \\
\delta(h) &= f'(x) + c\,h^2 + O(h) \qquad f \in C^5
\end{aligned}$$

Da questa scrittura di φ è possibile ricavare una stima *a posteriori* della parte principale dell'errore: poiché $\varphi(h/2) = \alpha + c\,(h/2)^p + O(h^q)$, si ha

$$\varphi(h) - \varphi\left(\frac{h}{2}\right) = 2^p c\,\frac{h^p}{2^p} - c\frac{h^p}{2^p} + O(h^q) = (2^p - 1)\,c\left(\frac{h}{2}\right)^p + O(h^q)$$

da cui segue che

$$\frac{\left|\varphi(h) - \varphi\left(\frac{h}{2}\right)\right|}{2^p - 1} \approx |c|\left(\frac{h}{2}\right)^p \approx \left|\varphi\left(\frac{h}{2}\right) - \alpha\right|$$

In maniera simile, tuttavia, è possibile ottenere una nuova stima dell'errore guadagnando almeno un ordine di infinitesimo, ossia l'*estrapolazione*. Si ha, infatti,

$$\begin{aligned}
2^p\varphi\left(\frac{h}{2}\right) = 2^p\alpha + c\,h^p + O(h^q) \implies \\
\implies 2^p\varphi\left(\frac{h}{2}\right) - \varphi(h) = (2^p - 1)\alpha + O(h^q) \implies \\
\implies \frac{2^p\varphi\left(\frac{h}{2}\right) - \varphi(h)}{2^p - 1} = \alpha + O(h^q) \quad (4.9)
\end{aligned}$$

Un'importante applicazione di questo risultato è il *metodo di Romberg*, che studia l'errore compiuto $T(h)$ nell'applicare la formula dei trapezii con passo costante h. Si ha, per le formule di Eulero-Maclaurin,

$$T(h) = I(f) + c\,h^2 + O(h^4) \qquad f \in C^4$$

e, applicando la regola di estrapolazione appena discussa, si trova

$$\frac{4T\left(\frac{h}{2}\right) - T(h)}{3} = I(f) + O(h^4)$$

L'estrapolazione può essere iterata se la funzione su cui effettuare l'estrapolazione è di forma

$$\varphi(h) = \alpha + \sum_{i=1}^{m} c_i\,h^{p_i} + O(h^{p_{m+1}}) \qquad \forall i < j : p_i < p_j$$

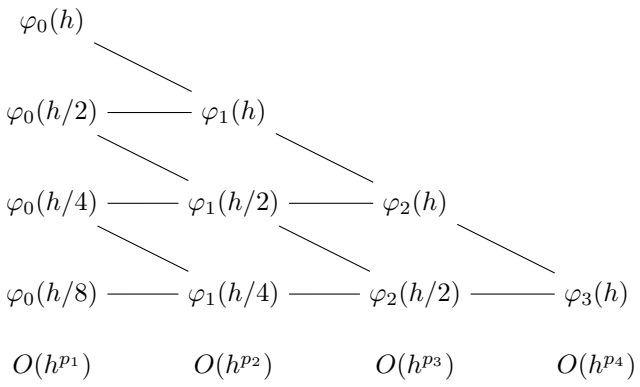

Figura 4.2: Rappresentazione grafica dell'implementazione dell'estrapolazione.

ove tutti i p_i sono strettamente positivi. Costruendo la successione di funzioni

$$\varphi_i(h) \coloneqq \frac{2^{p_i}\varphi_{i-1}\left(\frac{h}{2}\right) - \varphi_{i-1}(h)}{2^{p_i} - 1}$$

e iterando il risultato della (4.9), si ottiene che $\varphi_i(h) = \alpha + O(h^{p_{i+1}})$. Come si può notare in Figura 4.2, per ottenere le successive iterazioni dell'estrapolazione si rende necessario calcolare φ ed applicarvi l'estrapolazione anche per $h/2$, $h/4$ eccetera.

La tabella di estrapolazione si può usare fino al passo m per il rapporto incrementale $\delta_+(h)$ di una funzione f di classe C^{m+1}: in questo caso si ha $p_i = i$ per ogni $i \in \{\, 1, \ldots, m \,\}$. Espandendo f con le formule di Taylor, infatti, si ottiene

$$f(x+h) = f(x) + \sum_{i=1}^{m} \frac{f^{(i)}(x)}{i!}h^i + \frac{f^{(m+1)}(\xi)}{(m+1)!}h^{m+1} \qquad \exists \xi \in \operatorname{int}(x, x+h)$$

da cui segue che

$$\varphi_0(h) \coloneqq \delta_+(h) = f'(x) + \sum_{i=1}^{m} \frac{f^{(i+1)}(x)}{(i+1)!}h^i + O(h^m) = f'(x) + O(h)$$

L'iterazione successiva è data da

$$\varphi_1(h) = f'(x) - \frac{f^{(3)}(x)}{6}\frac{h^2}{2} + \cdots + O(h^m) = f'(x) + O(h^2)$$

che mostra quanto già scritto sopra.

In modo analogo si mostra che per $\varphi(h) = \delta(h)$ con f di classe C^{2m+1} si ha $p_i = 2i$; per il metodo di Romberg applicato ad una funzione di classe C^{2m+2} si ha ancora $p_i = 2i$.

5 Elementi di algebra lineare numerica

Dell'immenso campo di studii dell'algebra lineare numerica (NLA) trattiamo solo la risoluzione di sistemi lineari e l'applicazione di metodi diretti, come ad esempio il metodo di eliminazione gaussiano.

Prima di entrare nel merito, dimostriamo alcuni risultati di algebra lineare.

Proposizione 5.1. *Siano A, B matrici $n \times n$ ad entrate reali e sia $x \in \mathbb{R}^n$. Definita la* norma operatoriale *indotta da una norma vettoriale su $\mathbb{R}^n$*

$$\|A\| \coloneqq \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{5.1}$$

valgono le seguenti proprietà:

1. $\|Ax\| \leq \|A\| \, \|x\|$;
2. $\|AB\| \leq \|A\| \, \|B\|$.

Dimostrazione. Scelto $x \neq 0$, si ha

$$\|Ax\| = \|Ax\| \frac{\|x\|}{\|x\|} \leq \|A\| \, \|x\|$$

e si è dimostrata la prima proprietà.

Scelto, invece, $x \neq 0$ tale che $x \notin \ker B$, si ha

$$\|AB\| = \sup_{x \neq 0} \frac{\|ABx\|}{\|x\|} = \sup_{\substack{x \neq 0 \\ x \notin \ker B}} \frac{\|ABx\| \, \|Bx\|}{\|x\| \, \|Bx\|} \leq$$

$$\leq \left(\sup_{x \notin \ker B} \frac{\|ABx\|}{\|Bx\|} \right) \left(\sup_{x \neq 0} \frac{\|Bx\|}{\|x\|} \right) = \|A\| \, \|B\|$$

e si è arrivati alla seconda proprietà. □

Dalla sup-norma $\|x\|_\infty \coloneqq \max_{i \in \{ 1, \ldots, n \}} |x_i|$ è indotta la norma operatoriale $\|A\|_\infty \coloneqq \max_{i \in \{ 1, \ldots, n \}} \sum_{j=1}^n |a_{i,j}|$. Dalla norma euclidea $\|x\|_2 \coloneqq \left(\sum_{i=1}^n x_i^2 \right)^{1/2}$, invece, è indotta la norma operatoriale $\|A\|_2 \coloneqq \sqrt{\rho(A^{\mathrm{t}} A)}$, ove $\rho(B)$ indica il massimo dei valori assoluti degli autovalori di B.

Teorema 5.1 (Invertibilità di $\mathbf{1} - A$)**.** *Data una norma operatoriale indotta $\|\cdot\|$, se una matrice A soddisfa alla proprietà $\|A\| < 1$, allora $(\mathbf{1} - A) \in \mathrm{GL}_n(\mathbb{R})$ e*

$$(\mathbf{1} - A)^{-1} = \sum_{j=0}^{\infty} A^j \tag{5.2}$$

Vale, inoltre, la stima

$$\left\| (\mathbf{1} - A)^{-1} \right\| \leq \frac{1}{1 - \|A\|} \tag{5.3}$$

Dimostrazione. Ricordiamo innanzitutto che, poiché l'insieme delle matrici dotato della norma operatoriale indotta è uno spazio normato di dimensione finita, esso è uno spazio di Banach, ovvero è completo per la distanza indotta dalla norma. Definiamo la successione $S_n \coloneqq \sum_{j=0}^n A^j$. Per ogni $n, p \in \mathbb{N}$, vale la stima

$$\|S_{n+p} - S_n\| = \left\| \sum_{j=n+1}^{n+p} A^j \right\| \leq \sum_{j=n+1}^{n+p} \left\| A^j \right\| \leq \sum_{j=n+1}^{n+p} \|A\|^j$$

Poiché la serie corrispondente al membro di destra è convergente, il membro di destra tende a 0 per $n \to \infty$; da ciò si può dire che S_n è di Cauchy e, quindi, convergente: esiste, dunque, il limite $S = \lim_{n\to\infty} S_n$. Dall'eguaglianza

$$S(\mathbf{1} - A) = \sum_{j=0}^{\infty} A^j (\mathbf{1} - A) = \sum_{j=0}^{\infty} A^j - \sum_{j=0}^{\infty} A^{j+1} = \mathbf{1}$$

segue che S è l'inversa di $\mathbf{1} - A$. Per le proprietà della serie geometrica, poi, si ha

$$\|S\| = \left\| \sum_{j=0}^{\infty} A^j \right\| \leq \sum_{j=0}^{\infty} \|A\|^j = \frac{1}{1 - \|A\|} \qquad \square$$

Proposizione 5.2 (Localizzazione degli autovalori)**.** *Data una norma indotta $\|\cdot\|$, sia λ un autovalore di una matrice A; allora $|\lambda| \leq \|A\|$.*

Dimostrazione. Preso $v \in \mathbb{R}^n \setminus \{ 0 \}$ autovettore per λ, si ha

$$\|\lambda v\| = |\lambda| \, \|v\| = \|Av\| \leq \|A\| \, \|v\|$$

e, dividendo ambo i membri della diseguaglianza per $\|v\|$, si ottiene quanto richiesto. $\square$

5.1 Condizionamento di matrici e sistemi

Siano $A \in \mathrm{GL}_n(\mathbb{R})$ e $b \in \mathbb{R}^n$; vogliamo studiare l'effetto degli errori sui dati nella risoluzione del sistema lineare

$$Ax = b$$

Dal momento che A è invertibile, la soluzione del sistema esiste unica. In seguito indicheremo i dati "incerti" con la seguente scrittura:

$$\begin{aligned} \tilde{b} &= b + \delta b \\ \tilde{A} &= A + \delta A \\ \tilde{x} &= x + \delta x \end{aligned}$$

Scelta una norma $\|\cdot\|$, l'errore assoluto sarà $\|\delta x\| = \|x - \tilde{x}\|$, mentre l'errore relativo sarà $\frac{\|\delta x\|}{\|x\|} = \frac{\|x - \tilde{x}\|}{\|x\|}$.

5.1.1 Errore su b

Studiamo l'errore che si compie risolvendo il sistema $A\tilde{x} = \tilde{b}$. Si ha

$$A(x + \delta x) = b + \delta b \implies \cancel{Ax} + A\delta x = \cancel{b} + \delta b \implies \delta x = A^{-1}\delta b$$

da cui segue che l'errore assoluto sulla soluzione è $\|\delta x\| = \left\|A^{-1}\delta b\right\| \leq \left\|A^{-1}\right\| \|\delta b\|$. Dato che, poi,

$$\|Ax\| = \|b\| \leq \|A\| \, \|x\| \implies \|x\| \geq \frac{\|b\|}{\|A\|} \implies \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

si ottiene l'errore relativo commesso sulla soluzione

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\left\|A^{-1}\right\| \|A\|}_{\kappa(A)} \frac{\|\delta b\|}{\|b\|} \tag{5.4}$$

La quantità $\kappa(A) \coloneqq \|A^{-1}\| \|A\|$ è detta *indice di condizionamento* della matrice A rispetto alla norma operatoriale indotta $\|\cdot\|$. Dalla diseguaglianza

$$1 = \|\mathbf{1}\| = \left\|A^{-1}A\right\| \leq \left\|A^{-1}\right\| \|A\| = \kappa(A)$$

si ottiene la proprietà che $\kappa(A) \geq 1$ per ogni matrice A.

Proposizione 5.3. *Data una norma operatoriale indotta $\|\cdot\|$, per ogni matrice invertibile A vale*

$$\kappa(A) \geq \frac{|\lambda_M|}{|\lambda_m|} \tag{5.5}$$

ove λ_M e λ_m sono gli autovalori di A (eventualmente complessi) di modulo rispettivamente maggiore e minore. Se, poi, la norma scelta è quella euclidea e A è simmetrica, vale l'uguaglianza.

Dimostrazione. Per la Proposizione 5.2, si ha $\|A\| \geq |\lambda_M|$; poiché gli autovalori di A^{-1} sono i reciproci di quelli di A, l'autovalore di A^{-1} di massimo modulo è $\frac{1}{\lambda_m}$, che è ben definito perché A è invertibile e, quindi, non ha autovalori nulli. Dalla diseguaglianza

$$\kappa(A) = \left\|A^{-1}\right\| \|A\| \geq \left\|A^{-1}\right\| |\lambda_M| \geq \frac{|\lambda_M|}{|\lambda_m|}$$

si ottiene quanto richiesto.

Nel caso in cui A sia simmetrica, si ha $\|A\|_2 = \sqrt{\rho(A^{\mathrm{t}}A)} = \sqrt{\rho(A^2)} = |\lambda_M|$; dato che l'inversa di una matrice simmetrica invertibile è simmetrica anch'essa, si ottiene analogamente $\left\|A^{-1}\right\|_2 = \frac{1}{|\lambda_m|}$ e, sostituendo questi valori nella definizione di $\kappa(A)$, si ottiene l'uguaglianza. □

Se l'indice di condizionamento è "troppo grande" rispetto all'errore relativo sui dati, il sistema si definisce *mal condizionato*; in caso contrario, si dice *ben condizionato.*

Esempio. Vogliamo risolvere il sistema $Ax = b$ nella forma affetta da errore $A\tilde{x} = \tilde{b}$, ove

$$A = \begin{pmatrix} 7 & 10 \\ 5 & 7 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 0{,}7 \end{pmatrix} \qquad \tilde{b} = \begin{pmatrix} 1{,}01 \\ 0{,}69 \end{pmatrix}$$

Scelta la sup-norma, si calcolano $\|\delta b\|_\infty = 10^{-2}$ e $\|b\|_\infty = 1$. Risolvere il sistema esatto produce la soluzione $x = \left(\begin{smallmatrix} 0 \\ 0{,}1 \end{smallmatrix}\right)$, mentre la soluzione prodotta dal sistema inesatto è $\tilde{x} = \left(\begin{smallmatrix} -0{,}17 \\ 0{,}22 \end{smallmatrix}\right)$; l'errore relativo in sup-norma, dunque, è 1,2. Calcolando $\kappa(A) = 289$, si nota che il sistema è malcondizionato, visto che questo valore di condizionamento è molto alto rispetto all'errore relativo sul dato noto b: il sistema, infatti, a fronte di un errore relativo dell'1% restituisce un risultato con un errore relativo del 120% — che è altissimo.

5.1.2 Errore su A

Studiamo l'errore che si compie risolvendo il sistema $\tilde{A}\tilde{x} = b$. Si ha

$$\begin{aligned}(A+\delta A)(x+\delta x) = b \implies \cancel{Ax} + A\,\delta x + \delta A\,x + \delta A\,\delta x = \cancel{b} \implies \\ \implies A\,\delta x = -\delta A(x+\delta x) \implies \delta x = -A^{-1}\delta A(x+\delta x) \implies \\ \implies \|\delta x\| = \left\|A^{-1}\delta A(x+\delta x)\right\| \le \left\|A^{-1}\right\| \|\delta A\| \, \|\tilde{x}\|\end{aligned}$$

e da ciò si ricava un'approssimazione dell'errore relativo sulla soluzione

$$\frac{\|\delta x\|}{\|x\|} \approx \frac{\|\delta x\|}{\|\tilde{x}\|} \le \left\|A^{-1}\right\| \|\delta A\| = \kappa(A)\frac{\|\delta A\|}{\|A\|} \tag{5.6}$$

Si può ottenere una stima esatta nel caso in cui $\kappa(A)\frac{\|\delta A\|}{\|A\|} < 1$: dalla stima precedente, infatti, si vede che

$$\left(1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}\right)\|\delta x\| \le \kappa(A)\frac{\|\delta A\|}{\|A\|}\|x\|$$

e, se vale l'ipotesi detta sopra, si ottiene

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\kappa(A)\frac{\|\delta A\|}{\|A\|}}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}} \tag{5.7}$$

5.1.3 Errore su A e su b

Se il sistema si presenta nella situazione del tipo $\tilde{A}\tilde{x} = \tilde{b}$, si può ottenere una stima dell'errore relativo esatta con la condizione $\kappa(A)\frac{\|\delta A\|}{\|A\|} < 1$. Si ha, infatti,

$$\begin{aligned}(A+\delta A)(x+\delta x) = b + \delta b \implies \cancel{Ax} + \delta A\,x + A\,\delta x + \delta A\,\delta x = \cancel{b} + \delta b \implies \\ \implies (A+\delta A)\delta x = \delta b - \delta A\,x\end{aligned}$$

Verifichiamo che $A + \delta A$ sia invertibile: poiché $A + \delta A = A(\mathbf{1} + A^{-1}\delta A)$, è sufficiente controllare che $\left\|A^{-1}\delta A\right\| < 1$, cosí da poter applicare il Teorema 5.1; e la disuguaglianza

$$\left\|A^{-1}\delta A\right\| \le \left\|A^{-1}\right\| \|\delta A\| = \kappa(A)\frac{\|\delta A\|}{\|A\|} < 1$$

vera per ipotesi, conferma che $A + \delta A$ è invertibile. Si può procedere, dunque, con il conto precedente:

$$\begin{aligned}\delta x = (A+\delta A)^{-1}(\delta b - \delta A\,x) = (\mathbf{1} + A^{-1}\delta A)^{-1}A^{-1}(\delta b - \delta A\,x) \implies \\ \implies \|\delta x\| = \left\|(\mathbf{1} + A^{-1}\delta A)^{-1}A^{-1}(\delta b - \delta A\,x)\right\| \le \\ \le \left\|(\mathbf{1} + A^{-1}\delta A)^{-1}\right\| \left\|A^{-1}\right\| \|\delta b - \delta A\,x\| \le \\ \le \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\delta A\|}\|\delta b - \delta A\,x\| \le \\ \le \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\|\,\|\delta A\|}(\|\delta b\| + \|\delta A\|\,\|x\|)\end{aligned}$$

e, dividendo per $\|x\|$, si ottiene

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right) \tag{5.8}$$

5.2 Metodo di eliminazione gaussiano e fattorizzazione LU

Un sistema lineare $Ax = b$, con A invertibile, può essere risolto con il metodo di eliminazione di Gauss (MEG): esso effettua un numero finito di trasformazioni perché il sistema sia risolubile con un algoritmo diretto uguale per tutti i casi.

5.2.1 Algoritmo per ottenere L e U

L'algoritmo per implementare il MEG è il seguente, supponendo che $A \in \mathrm{GL}_n(\mathbb{R})$:

- per ogni $i \in \{ 1, \dots, n-1 \}$ preso in ordine, la i-esima riga di A è scambiata con una j-esima, di modo che $|a_{j,i}|$ sia il massimo possibile; questa operazione si dice *pivoting* e serve ad evitare che ci siano numeri troppo piccoli sulla diagonale; ogniqualvolta si effettua un tale scambio, si incrementa un contatore s di 1;
- chiamata R_i la riga che ora si trova per i-esima, per ogni riga R_k, con $i < k \leq n$, si effettua la sostituzione $R'_k = R_k - \frac{a_{k,i}}{a_{i,i}} R_i$; il valore $m_{k,i} \coloneqq \frac{a_{k,i}}{a_{i,i}}$ prende il nome di *moltiplicatore*.

La matrice risultante dall'applicazione del MEG si indica con U.

Chiamata $T_{k,i}(\alpha)$ la matrice che alla riga i-esima aggiunge α volte la riga k-esima, se il MEG non deve scambiare l'ordine delle righe per una certa matrice A, si ha $U = T_{n,n-1}(-m_{n,n-1}) \cdots T_{2,1}(-m_{2,1})\, A$, da cui segue che

$$A = T_{2,1}(m_{2,1}) \cdots T_{n,n-1}(m_{n,n-1})\, U = LU$$

con L matrice triangolare inferiore che contiene 1 nelle entrate della diagonale e $m_{k,i}$ nell'entrata (k, i) per ogni $k > i$. La matrice A non necessita di avere l'ordine delle proprie righe cambiato se, ad esempio, è simmetrica e definita positiva, oppure se è a diagonale strettamente dominante, ossia se vale

$$|a_{i,i}| > \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{i,j}|$$

Nel caso in cui, invece, A necessiti di un cambio nel proprio ordine delle righe, il MEG restituisce U in modo tale che $PA = LU$, con P una matrice di permutazione delle righe appropriata.

5.2.2 Risoluzione del sistema lineare

Dopo aver ottenuto P, L e U, il sistema si può risolvere in alcuni passaggi che sfruttano una maggiore "semplicità" dei sistemi lineari con matrici triangolari. Poiché si ha

$$\begin{cases} PA = LU \\ Ax = b \end{cases}$$

è vero che $PAx = Pb$, ovvero che $LUx = Pb$: da ciò si ottiene che

$$\begin{cases} Ly = Pb \\ Ux = y \end{cases} \tag{5.9}$$

Questa coppia di sistemi si risolve con due metodi speciali, basati sul fatto che L e U sono matrici triangolari. Per risolvere il primo sistema, che riscriviamo come $Ly = \beta$, si può usare il metodo della *sostituzione in avanti*

$$\begin{aligned} y_1 &= \frac{\beta_1}{\ell_{1,1}} \\ y_i &= \frac{\beta_i - \sum_{j=1}^{i-1} \ell_{i,j} y_j}{\ell i, i} \end{aligned} \tag{5.10}$$

ove $\ell_{i,j}$ indica l'entrata (i, j) di L. Ottenuto il vettore y, lo si può inserire nel secondo sistema, che darà la soluzione al problema iniziale: questo secondo sistema $Ux = y$ si può risolvere con la *sostituzione all'indietro*

$$\begin{aligned} x_n &= \frac{y_n}{u_{n,n}} \\ x_i &= \frac{y_i - \sum_{j=i+1}^{n} u_{i,j} x_j}{u_{i,i}} \end{aligned} \tag{5.11}$$

ove $u_{i,j}$ indica l'entrata (i, j) di U. Il vettore x cosí ottenuto risolve il sistema $Ax = b$.

5.2.3 Complessità computazionale del MEG

Il costo computazionale del MEG per la i-esima colonna consta circa di $n - i$ addizioni e $n - i$ moltiplicazioni: da ciò segue che il costo computazionale totale è circa pari a

$$\sum_{i=1}^{n-1} 2(n-i)^2 = 2\sum_{i=1}^{n-1} (n-i)^2 = 2\sum_{j=1}^{n-1} j^2$$

Per stimare il comportamento asintotico della somma, usiamo la disuguaglianza

$$\int_0^{n-1} x^2 \, \mathrm{d}x \leq \sum_{j=1}^{n-1} j^2 \leq \int_1^n x^2 \, \mathrm{d}x$$

e, poiché entrambi gli integrali sono asintotici a $n^3/3$, si può concludere che il costo computazionale del calcolo del determinante MEG è asintotico a $2n^3/3$, ossia è $O(n^3)$.

Sia la sostituzione in avanti sia la sostituzione all'indietro compiono $O(n^2)$ operazioni; consideriamo, ad esempio, la sostituzione in avanti: per la riga i-esima (non la prima) occorre effettuare $i - 1$ prodotti, $i - 2$ somme e altre due operazioni, sicché il costo computazionale è pari a

$$2 + \sum_{i=2}^{n} 2i - 1 = n^2 + n - 1 = O(n^2)$$

5.2.4 Calcolo del determinante di una matrice

Per calcolare il determinante di una matrice $A \in M_n(\mathbb{R})$, si può usare la *regola di Laplace*

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

Tabella 5.1: Confronto dei tempi di calcolo di alcune matrici $n \times n$ coi metodi di Laplace e MEG su un calcolatore che effettua un miliardo di operazioni al secondo.

n	Laplace (s)	MEG (s)
10	10^{-3}	10^{-6}
15	2400	$3 \cdot 10^{-6}$
20	$3{,}15 \cdot 10^{9}$	$8 \cdot 10^{-6}$
25	$3{,}15 \cdot 10^{16}$	10^{-5}
100	$3{,}15 \cdot 10^{148}$	10^{-3}

Questo metodo, però, è altamente inefficiente: dalla formula si nota subito che $C_n = n(1+C_{n-1})$, ovvero $C_n = n(1+(n-1)(1+\cdots)) = n(n-1)\cdot\cdots 4\cdot 3\cdot 3 > n!$, da cui segue che $C_n = O(n!)$.

L'altro metodo usato per calcolare i determinanti fa uso del MEG: trovata la matrice U, infatti, il determinante è, a meno del segno dovuto agli scambi di riga, il prodotto delle entrate sulla diagonale di U. Come mostrato nella Tabella 5.1, questo metodo è molto piú efficiente, visto che il suo costo computazionale è $O(n^3)$, lo stesso del solo MEG.

5.2.5 Calcolo della matrice inversa

Data una matrice $n\times n$ invertibile A, le sue colonne $a_1, \ldots, a_n$ sono l'immagine dei versori unitari della base canonica $e_1, \ldots, e_n$. Per ottenere le colonne $c_1, \ldots, c_n$ dell'inversa A^{-1}, dunque, basta risolvere n sistemi lineari del tipo $Ac_i = e_i$. Il costo computazionale di ciò è $O(n^3)$, perché occorre risolvere $2n$ volte un sistema con sostituzione, che ha costo $O(n^2)$, dopo aver applicato il MEG, che ha costo $O(n^3)$.

5.2.6 Malcondizionamento e regolarizzazione con parametro

Per come è costruito l'algoritmo di fattorizzazione LU, esso ha un errore relativo in norma euclidea prossimo alla precisione di macchina quando non calcolato in aritmetica a precisione infinita. Questa accuratezza, tuttavia, non garantisce di risolvere un sistema lineare con la stessa precisione. Definiamo, ad esempio, la n-esima matrice di Hilbert $H_n = \left(\frac{1}{i+j-1}\right)_{1\leq i,j\leq n}$, che è simmetrica e definita positiva — ciò rende superfluo tenere traccia di P, perché è la matrice identica. Applicando il MEG in precisione doppia, si ha comunque

$$\frac{\left\| H - \tilde{L}\tilde{U} \right\|_2}{\|H\|_2} \approx \varepsilon_M$$

ma, usando le due matrici risultanti per risolvere un sistema lineare $\tilde{H}\tilde{x} = \beta$, risulta un errore relativo superiore al 100% se $n \geq 13$. Usando la (5.6), si vede che

$$\frac{\|x - \tilde{x}\|_2}{\|x\|_2} \lesssim \kappa(H_n) \frac{\|\delta H_n\|_2}{\|H_n\|_2} \approx \kappa(H_n)\, \varepsilon_M$$

$\kappa(H_n)$, però, cresce esponenzialmente al variare di n, ed è sufficiente $n = 13$ perché $\kappa(H_n) \gg \varepsilon_M^{-1}$.

Per ovviare a questo problema si considera una famiglia parametrizzata di sistemi del tipo $A_h x_h = b$ tali che $A_h \to A$ e $x_h \to x$ quando $h \to 0$ e tali che $\kappa(A_h) < \kappa(A)$ per ogni h. Studiando i sistemi perturbati $A_h \tilde{x}_h = \tilde{b}$ si ha

$$\|\tilde{x}_h - x\| \leq \|\tilde{x}_h - x_h\| + \underbrace{\|x_h - x\|}_{e(h)} \lesssim \kappa(A_h)\, \varepsilon_M \|x_h\| + e(h)$$

da cui si ricava la stima

$$\frac{\|\delta x\|}{\|x\|} \leq \left(1 + \frac{e(h)}{\|x\|}\right) \kappa(A_h) \frac{\varepsilon_M}{\|x\|} + \frac{e(h)}{\|x\|} \approx \kappa(A_h) \frac{\varepsilon_M}{\|x\|} + \frac{e(h)}{\|x\|}$$

La struttura dell'errore è molto simile a quella della derivazione numerica, perciò risulta naturale minimizzare e in h.

5.3 Sistemi sovradeterminati e fattorizzazione QR

5.3.1 Sistema delle equazioni normali

Dati $m, n \in \mathbb{N}$, $m > n \geq 1$, siano $A \in M_{m \times n}(\mathbb{R})$ e $b \in \mathbb{R}^m$ dati e $x \in \mathbb{R}^n$ incognito. Per il Teorema di Rouché-Capelli, se b non è un elemento dello spazio vettoriale generato dalle colonne di A, il sistema $Ax = b$ non ha soluzione; in tal caso, dunque, una risoluzione "classica" del sistema non dà alcun risultato. Si può, però, cercare $x \in \mathbb{R}^n$ tale che la quantità $\|Ax - b\|_2^2$ sia minima, ossia una *soluzione ai minimi quadrati*. Definito, dunque, $\Phi(x) \coloneqq \|Ax - b\|_2^2 = \langle Ax - b, Ax - b \rangle = \sum_i (Ax - b)_i^2$, esso è un polinomio in n variabili. Come già visto al Teorema 3.2, x è di minimo per Φ se e solo se è soluzione del sistema delle equazioni normali $A^{\mathrm{t}} A x = A^{\mathrm{t}} b$. Benché sia già stato dimostrato con argomenti piú semplici, diamo un'altra dimostrazione del Teorema sfruttando alcuni strumenti di calcolo differenziale in piú variabili nel caso in cui A abbia rango massimo, cioè n.

Dimostrazione. $\Phi(x)$ è un polinomio quadratico in n variabili reali, che può essere scritto come

$$\Phi(x) = \sum_i \left(\left(\sum_j a_{i,j}\, x_j \right) - b_i \right)^2$$

Per cercarne il minimo, occorre porre la condizione $\nabla \Phi(x) = 0$. Le derivate parziali di Φ sono date da

$$\frac{\partial \Phi}{\partial x_k} = \sum_i 2 \left(\sum_j a_{i,j}\, x_j - b_i \right) a_{i,k} = 2\, \mathrm{riga}_k(A^{\mathrm{t}})\, (Ax - b)$$

perciò nei punti critici di Φ vale l'uguaglianza

$$\nabla \Phi(x) = 2A^{\mathrm{t}}(Ax - b) = 0$$

da cui segue facilmente il sistema delle equazioni normali.

Occorre mostrare che la soluzione di tale sistema è di minimo per Φ. Notiamo innanzitutto che $A^{\mathrm{t}} A$ non è singolare, perché A ha rango massimo; è vero, poi,

che $\mathrm{H}\,\Phi(x) = \mathrm{J}\big(\nabla\Phi(x)\big)$ è simmetrica, perché Φ è di classe C^2. Calcolando la matrice hessiana, si ottiene

$$\mathrm{H}\,\Phi(x) = \mathrm{J}\big(2A^{\mathrm{t}}(Ax-b)\big) = 2A^{\mathrm{t}}A$$

Poiché $A^{\mathrm{t}}A$ è definita positiva, per Teorema il punto critico x è di minimo.[10] □

5.3.2 Fattorizzazione QR

Per risolvere il sistema delle equazioni normali in aritmetica di macchina non conviene ricorrere al MEG: nel caso in cui $\kappa(A) \gg 1$, infatti, tale sistema diventa assai malcondizionato: qualora A sia quadrata, ad esempio, si ha $\kappa(A^{\mathrm{t}}A) \approx \kappa(A^2)$ — l'uguaglianza precisa vale nel caso in cui A sia simmetrica.

Per ovviare a questo problema, si decompone A in due matrici:

- una matrice Q risultato dell'ortonormalizzazione di Gram-Schmidt delle colonne di A e tale che $Q^{\mathrm{t}}Q = \mathbf{1}_n$;
- una matrice $R \in \mathrm{GL}_n(\mathbb{R})$ tale che R^{-1} sia triangolare superiore.

In questo modo si ha $Q = AR^{-1}$. Facendo sostituzioni a partire da queste relazioni, si ha

$$A^{\mathrm{t}}A = R^{\mathrm{t}}Q^{\mathrm{t}}QR = R^{\mathrm{t}}R$$
$$A^{\mathrm{t}}b = R^{\mathrm{t}}Q^{\mathrm{t}}b$$

da cui segue che il sistema delle equazioni normali è equivalente al sistema lineare

$$Rx = Q^{\mathrm{t}}b \tag{5.12}$$

Poiché R è costruita a partire da A, ha intuitivamente un condizionamento in norma euclidea molto minore di quello di $A^{\mathrm{t}}A$, e ciò rende vantaggioso l'uso di questo metodo se il condizionamento del sistema di partenza è molto alto.

[10] Il Teorema si trova nelle dispense di Analisi II A del MONTI.