Appunti di Calcolo Numerico - a.a. 2021/2022

Corso di Laurea in Ingegneria Elettronica

Maurizio Ciampa

Dipartimento di Matematica - Università di Pisa

Indice

Premessa

In questi appunti affronteremo alcuni problemi classici di Analisi Matematica ed Algebra Lineare, dal punto di vista del Calcolo Numerico. Precisamente studieremo i problemi seguenti:

P1: Data una funzione $f : \mathbb{R} \to \mathbb{R}$, determinare un numero reale α tale che $f(\alpha) = 0$.

P2: Date la matrice $A \in \mathbb{R}^{n\times n}$ e la colonna $b \in \mathbb{R}^n$, determinare $x^* \in \mathbb{R}^n$ tale che $Ax^* = b$.

P3: Date le coppie di numeri reali $(x_0, y_0), \ldots, (x_k, y_k)$ e le funzioni $f_0, \ldots, f_k : \mathbb{R} \to \mathbb{R}$, determinare numeri reali $a_0, \ldots, a_k$ tali che, posto $f(x) = a_0 f_0(x) + \cdots + a_k f_k(x)$ si abbia $f(x_0) = y_0, \ldots, f(x_k) = y_k$.

P4: Dati $A \in \mathbb{R}^{n\times k}$ con $n > k$ e $b \in \mathbb{R}^n$, determinare $x^* \in \mathbb{R}^k$ che rende minimo il valore della funzione $\mathrm{SQ} : \mathbb{R}^k \to \mathbb{R}$ definita da $\mathrm{SQ}(x) = \| Ax - b \|^2$.

Si osservi che in tutti questi problemi si richiede di determinare uno o più *numeri reali*. Nel Calcolo Numerico si cercano (a) *procedure*, da eseguire utilizzando un calcolatore, che determinano *scritture posizionali finite* (usualmente in base dieci) di *approssimazioni* dei numeri richiesti e (b) informazioni sull'*errore* commesso utilizzando le scritture ottenute per approssimare i numeri reali richiesti.

Ad esempio, data la funzione $f(x) = x^2 - 2$, si consideri il problema P1. Come noto, $f(\sqrt{2}) = 0$. La risposta (ottenuta, ad esempio, con una procedura che utilizza l'usuale formula risolutiva delle equazioni di secondo grado):

$$\alpha = \sqrt{2}$$

non è soddisfacente per il Calcolo Numerico perché, pur indicando un ben preciso numero reale, non ne fornisce una scrittura posizionale finita. In questo caso, ma è *quasi sempre* così, la richiesta di una scrittura posizionale finita può essere soddisfatta *solo* se si accetta di ottenere quella di un numero reale che *approssima* il numero richiesto. Ad esempio, scritture accettabili per il Calcolo Numerico, ma risposte non ancora soddisfacenti, sono:

$$\xi = 1 \quad , \quad \xi = 1.41421356237309514547462185873882845044136047363281 25$$

Per renderle risposte soddisfacenti occorre dare informazioni sull'errore. Come vedremo, *un modo* per misurare l'errore commesso approssimando un numero reale $\alpha \neq 0$ con il numero ξ è l'*errore relativo*:

$$\epsilon = \frac{\xi - \alpha}{\alpha}$$

Risposte soddisfacenti sono allora:

$$\xi = 1 \quad , \quad |\,\epsilon\,| < 0.5$$

e:

$$\xi = 1.41421356237309514547462185873882845044136047363281 25 \quad , \quad |\,\epsilon\,| < 2^{-53} \approx 10^{-16}$$

La seconda risposta fornisce una limitazione sull'errore relativo *più stringente* della prima e questo la rende *preferibile*. Si osservi però che le stime *non consentono* di decidere quale delle due approssimazioni dia luogo ad un errore relativo più piccolo – ovvero quale delle due risposte sia più *accurata*.

In questi appunti le procedure sono descritte utilizzando un linguaggio, inventato e di immediata comprensione, che consente di usare un tipo "ideale" di dato numerico elementare: il *numero reale*. Gli *oggetti* del tipo *numero reale* sono gli elementi di $\mathbb{R}$ e le *funzioni utilizzabili* per operare su tali oggetti sono le *operazioni aritmetiche*, le *funzioni elementari* (funzioni trigonometriche, funzione esponenziale, logaritmica, radice n–esima, ...) ed i *confronti*.

Nel discutere l'uso del calcolatore per eseguire una procedura, faremo l'ipotesi che sia *sufficiente* studiare l'effetto della *sostituzione*, nella procedura in esame, del tipo – praticamente non realizzabile – *numero reale* con il tipo – praticamente realizzabile – *numero in virgola mobile e precisione finita*.[1] Nel Capitolo 0 si descrive il tipo *numero in virgola mobile e precisione finita* ed un procedimento per

[1] Questo tipo di dato corrisponde, concettualmente, ad uno dei *formati base* descritti nel documento *IEEE Standard for Floating-Point Arithmetic* (IEEE Std 754-2019) che prescrive regole – ampiamente condivise – per eseguire calcoli in virgola mobile in modo che il risultato sia *indipendente* dal dispositivo di calcolo utilizzato.

effettuare la sostituzione. I quattro capitoli successivi saranno dedicati, uno ciascuno, ai problemi P1 – P4 menzionati sopra.

Gli esercizi contrassegnati dal simbolo ★ sono leggermente più astratti rispetto agli altri. Gli esercizi contrassegnati dal simbolo **S** sono risolti, e la soluzione si trova nell'appendice del capitolo di pertinenza. Quelli contrassegnati dal simbolo ♠ richiedono direttamente, o comunque riguardano, l'uso del calcolatore. A chi legge si raccomanda di riprodurre al calcolatore i "dialoghi" con *Scilab* proposti e di prendere spunto da essi per crearne di nuovi (per ottenere *Scilab* visitare la pagina `https://www.scilab.org/`).

0 Il tipo *numero in virgola mobile e precisione finita*

In questo capitolo descriviamo il tipo *numero in virgola mobile e precisione finita*, il procedimento per trasformare una procedura che usa il tipo *numero reale* in una che usa il tipo *numero in virgola mobile e precisione finita*, e come studiare gli effetti numerici della trasformazione. Il capitolo è suddiviso in quattro sezioni: nella prima si definisce l'insieme M dei *numeri in virgola mobile e precisione finita*, ovvero l'insieme degli *oggetti* del tipo *numero in virgola mobile e precisione finita*; nella seconda si descrive la *funzione arrotondamento* utilizzata per approssimare elementi di $\mathbb{R}$ con elementi di M; nella terza si descrive l'insieme delle *funzioni predefinite*: le *funzioni* che il tipo mette a disposizione per operare sugli elementi di M. Infine, nella quarta sezione si descrive il procedimento di trasformazione e si mostra, usando alcuni semplici esempi, come ottenere informazioni sull'errore commesso utilizzando i valori numerici generati dalla procedura che usa il tipo *numero in virgola mobile e precisione finita* per approssimare quelli generati dalla procedura che usa il tipo *numero reale*.

0.1 Numeri in virgola mobile e precisione finita

Per definire l'insieme dei numeri in virgola mobile e precisione finita, è utile ricordare alcune nozioni riguardanti la "rappresentazione scientifica" di un numero reale.

0.1.1 Definizione (esponente e frazione di un numero reale non nullo)

Siano x un numero reale *diverso da zero* e β un numero intero maggiore o uguale a due, detto *base*. È *univocamente determinato* un numero intero b tale che, posto:

$$g = \frac{|\,x\,|}{\beta^b}$$

si ha:

$$\beta^{-1} \leqslant g < 1$$

ovvero: esiste *un solo modo* di scrivere x nella forma:

$$x = (-1)^s\,\beta^b\,g \qquad \text{con} \qquad s \in \{\,0,1\,\} \quad , \quad b \in \mathbb{Z} \quad , \quad \frac{1}{\beta} \leqslant g < 1$$

s è il *segno* di x, b e g – che *dipendono da* β – sono, rispettivamente, l'*esponente* e la *frazione* di x (in base β).

– *Dimostrazione*: Sia b l'*unico* numero intero tale che $\beta^{b-1} \leqslant |\,x\,| < \beta^b$. Allora:

$$\beta^{-1} \leqslant \frac{|\,x\,|}{\beta^b} < 1$$

0.1.2 Esempio

Sia $x = \sqrt{5}$.

Per $\beta = 10$ si ottiene: $s = 0$ (x è positivo) e, poiché $10^0 \leqslant \sqrt{5} < 10^1$:

$$b = 1 \quad \text{e} \quad g = \frac{\sqrt{5}}{10}$$

Per $\beta = 2$ si ottiene ancora $s = 0$ (il segno di x non dipende dal valore di β) e poi, poiché $2^1 \leqslant \sqrt{5} < 2^2$:

$$b = 2 \quad \text{e} \quad g = \frac{\sqrt{5}}{4}$$

0.1.3 Osservazione

Alcuni numeri reali ammettono *due* scritture posizionali (ad esempio, in base dieci, le possibili scritture posizionali di *un decimo* sono: 0.1 e $0.0\overline{9}$). In tal caso, delle due si considera quella *finita*. Con questa precisazione, la condizione $\beta^{-1} \leqslant g < 1$ è equivalente a:

la scrittura posizionale di g in base β ha la forma $0.c_1c_2\cdots$ con $c_1 \neq 0$

Le cifre $c_1, c_2, \ldots$ della scrittura posizionale di g in base β si possono ottenere, una alla volta, con la procedura seguente:[2]

- Passo 1: $i = 1$; $t_i = g$; $\quad (t_1 = 0.c_1c_2\cdots)$
- Passo 2: $c_i = \lfloor \beta t_i \rfloor$; $\quad (\beta t_i = c_i.c_{i+1}c_{i+2}\cdots)$
- Passo 3: $t_{i+1} = \beta t_i - \lfloor \beta t_i \rfloor$; $\quad (t_{i+1} = 0.c_{i+1}c_{i+2}\cdots)$
- Passo 4: Se $t_{i+1} = 0$ allora STOP, altrimenti $i = i + 1$; VAI AL Passo 2.

0.1.4 Esempio

Sia $x = \frac{1}{10}$.
Per $\beta = 10$ si ottiene: $s = 0$ e, poiché $10^{-1} \leqslant x < 10^0$:

$$b = 0 \quad \text{e} \quad g = \frac{1}{10} = 0.1 \quad \text{ovvero} \quad x = (-1)^0\, 10^0\, 0.1$$

Per $\beta = 2$ si ottiene ancora $s = 0$ e poi, poiché $2^{-4} \leqslant x < 2^{-3}$:

$$b = -3 \quad \text{e} \quad g = \frac{8}{10} = \frac{4}{5} = 0.\overline{1100} \quad \text{ovvero} \quad x = (-1)^0\, 2^{-3}\, 0.\overline{1100}$$

Si osservi che in base dieci la frazione ha scrittura posizionale di *lunghezza uno* mentre in base due la frazione ha scrittura posizionale di *lunghezza infinita*.

0.1.5 Esempio

La procedura descritta nell'Osservazione precedente si può applicare in modo abbastanza semplice anche in alcuni casi in cui x non è un numero razionale. Sia, ad esempio, $x = \sqrt{2}$.
Per $\beta = 2$ si ottiene: $s = 0$ e, poiché $2^0 \leqslant x < 2^1$:

$$b = 1 \quad \text{e} \quad g = \tfrac{\sqrt{2}}{2}$$

Poiché x non è un numero razionale, la scrittura posizionale di g in base due ha *certamente* lunghezza infinita *e* non è periodica. Le prime cifre della scrittura posizionale si possono ottenere, con la procedura descritta nell'Osservazione precedente, come segue.
Posto $i = 1$ e $t_1 = g = \frac{\sqrt{2}}{2}$ si ha:

$$c_1 = \lfloor 2t_1 \rfloor = \lfloor \sqrt{2} \rfloor = 1$$

Si pone poi:

$$t_2 = 2t_1 - c_1 = \sqrt{2} - 1$$

Essendo $t_2 \neq 0$, si pone $i = 2$ e si ottiene:

$$c_2 = \lfloor 2t_2 \rfloor = \lfloor 2\sqrt{2} - 2 \rfloor$$

ovvero: se $2\sqrt{2} - 2 \geqslant 1$ allora $c_2 = 1$, altrimenti $c_2 = 0$. Sfruttando la monotonia della funzione $t \mapsto t^2$, si constata facilmente che

$$2\sqrt{2} - 2 \geqslant 1 \qquad \text{se e solo se} \qquad 2 \geqslant \tfrac{9}{4}$$

dunque: $c_2 = 0$. Procedendo analogamente si pone:

$$t_3 = 2t_2 - c_2 = 2\sqrt{2} - 2$$

Essendo $t_3 \neq 0$, si pone $i = 3$ e si ottiene:

$$c_3 = \lfloor 2t_3 \rfloor = \lfloor 4\sqrt{2} - 4 \rfloor$$

ovvero: se $4\sqrt{2} - 4 \geqslant 1$ allora $c_3 = 1$, altrimenti $c_3 = 0$. Si constata facilmente che

$$4\sqrt{2} - 4 \geqslant 1 \qquad \text{se e solo se} \qquad 2 \geqslant \tfrac{25}{16}$$

[2] Se x è un numero reale positivo, si indica con $\lfloor x \rfloor$ la *parte intera di* x, ovvero il più grande numero intero minore o uguale ad x.

dunque: $c_3 = 1$.

Si ottiene cioè:

$$x = \sqrt{2} = 2^1 \cdot 0.101\cdots$$

0.1.6 Osservazione

Una procedura alternativa a quella descritta nell'Osservazione precedente per determinare, una alla volta, le cifre in base $\beta = 2$ della scittura posizionale di $g \in [\frac{1}{2}, 1)$, è la seguente:

- Passo 1: $i = 1$; $c_i = 1$;
- Passo 2: se $g = 0.c_1 \cdots c_i$ allora STOP;
- Passo 3: se $g \geqslant 0.c_1 \cdots c_i 1$ allora $c_{i+1} = 1$, altrimenti $c_{i+1} = 0$; $i = i + 1$; VAI AL Passo 2.

Sia, ad esempio, $x = \log_2 3$.

Per $\beta = 2$ si ottiene: $s = 0$ e, poiché $2^0 \leqslant x < 2^1$:

$$b = 1 \quad \text{e} \quad g = \tfrac{\log_2 3}{2} \in [\tfrac{1}{2}, 1)$$

Poiché g non è un numero razionale, la sua scrittura posizionale in base due ha certamente lunghezza infinita e non è periodica. Le prime quattro cifre della scrittura posizionale si possono ottenere, con la procedura descritta sopra, come segue.

Posto $i = 1$ e $c_1 = 1$ si ha:

$$g \neq 0.1$$

Si ha poi:

$$\frac{\log_2 3}{2} > 0.11 = \tfrac{3}{4}$$

infatti, per la monotonia della funzione $t \mapsto 2^t$:

$$\frac{\log_2 3}{2} > \tfrac{3}{4} \quad \Leftrightarrow \quad 2\log_2 3 > 3 \quad \Leftrightarrow \quad \log_2 3^2 > 3 \quad \Leftrightarrow \quad 3^2 > 2^3 \quad \Leftrightarrow \quad 9 > 8$$

Dunque: $c_2 = 1$. Posto $i = 2$ si ha poi, procedendo allo stesso modo:

$$g \neq 0.11$$

e:

$$\frac{\log_2 3}{2} < 0.111 = \tfrac{7}{8}$$

e quindi $c_3 = 0$. Posto $i = 3$ si ha infine:

$$g \neq 0.110$$

e:

$$\frac{\log_2 3}{2} < 0.1101 = \tfrac{13}{16}$$

e quindi $c_4 = 0$.

Si ottiene cioè:

$$x = \log_2 3 = 2^1 \cdot 0.1100\cdots$$

0.1.7 Definizione (numeri in virgola mobile, precisione)

Siano β un numero intero maggiore o uguale a due ed m un numero intero positivo. L'insieme:

$$F(\beta, m) = \{\,0\,\} \cup \Big\{\, x \in \mathbb{R} \text{ tali che } x = (-1)^s\, \beta^b\, 0.c_1 \cdots c_m \\ \text{con } s \in \{\,0, 1\,\},\ b \in \mathbb{Z},\ c_1, \ldots, c_m \text{ cifre in base } \beta \text{ e } c_1 \neq 0 \,\Big\}$$

si chiama *insieme dei numeri in virgola mobile* (normalizzati) *in base β e precisione m.*

L'insieme $F(\beta, m)$ contiene dunque zero e *tutti i numeri reali per i quali in base β la frazione ha scrittura posizionale di lunghezza non superiore* a m.

0.1.8 Esempio

Si consideri $F(10, 1)$.

– Poiché $\frac{1}{100} = 10^{-1}\,0.1$ allora $\frac{1}{100} \in F(10,1)$. Invece: $\frac{11}{100} \notin F(10,1)$ perché $\frac{11}{100} = 10^{0}\,0.11$ e la scrittura posizionale della frazione *non è compatibile* con la precisione.

– Se $x \in F(10,1)$ allora $-x \in F(10,1)$: l'insieme $F(10,1)$ è *simmetrico* rispetto a zero.

– Le possibili scritture posizionali (in base dieci) della frazione di un elemento non nullo di $F(10,1)$ sono:

$$0.1\,,\ 0.2\,,\ \ldots\,,\ 0.9$$

Allora: per ogni numero intero b l'insieme degli elementi positivi di $F(10,1)$ con esponente b è:

$$B_b = \{\, 10^b\,0.1\,,\ 10^b\,0.2\,,\ \ldots\,,\ 10^b\,0.9\,\}$$

Gli insiemi B_b sono "ordinati:" se c, d sono numeri interi tali che $c < d$ allora $\max B_c < \min B_d$. Graficamente questo significa che rappresentando gli elementi di B_c e B_d sulla retta reale, i punti che rappresentano gli elementi di B_c sono *tutti* a sinistra del punto che rappresenta $\min B_d$ e quelli che rappresentano gli elementi di B_d sono *tutti* a destra del punto che rappresenta $\max B_c$.

– Infine:[3]

$$F(10,1) = [\cup_{b\in\mathbb{Z}}\,(-1)B_b]\ \cup\ \{\,0\,\}\ \cup\ [\cup_{b\in\mathbb{Z}}\,B_b]$$

e $F(10,1)$ ha infiniti elementi.

0.1.9 Esercizio

Si consideri $F(10,1)$.

Rappresentare sulla retta reale (non in scala) gli insiemi B_0, B_1 e B_{-1}. Determinare la distanza tra due elementi consecutivi in B_0, in B_1 e in B_{-1}. Determinare infine la distanza tra $\max B_{-1}$ e $\min B_0$ e tra $\max B_0$ e $\min B_1$.

In generale si ha: dato $b \in \mathbb{Z}$ la distanza tra due elementi consecutivi in B_b è 10^{b-1}.

0.1.10 Osservazione (Proprietà di $F(\beta, m)$)

Si ha:

(1) L'insieme $F(\beta,m)$ è un *sottoinsieme proprio* di $\mathbb{Q}$.

Infatti: $\xi = (-1)^s\beta^b\,0.c_1\cdots c_m = (-1)^s\beta^{b-m}\,c_1\cdots c_m \in \mathbb{Q}$ e il numero razionale $1+\beta^{-m}$ *non appartiene* ad $F(\beta,m)$ perché la scrittura posizionale della frazione ha lunghezza maggiore della precisione.

(2) Per quanto detto al punto precedente l'insieme $F(\beta,m)$ è *numerabile* ed *ordinato*.

(3) L'insieme $F(\beta,m)$ è *simmetrico rispetto a zero*.

(4) Zero è (l'unico) *punto di accumulazione* di $F(\beta,m)$.

Esercizio: Determinare una successione ξ_k di elementi positivi di $F(\beta,m)$ tale che $\lim \xi_k = 0$.

(5) $\sup F(\beta,m) = +\infty, \inf F(\beta,m) = -\infty$.

Esercizio: Determinare una successione $\xi_k \in F(\beta,m)$ tale che $\lim \xi_k = +\infty$.

0.1.11 Definizione (Funzioni successore e predecessore)

Si consideri la rappresentazione degli elementi di $F(\beta,m)$ sulla retta reale e sia ξ un elemento *non nullo* di $F(\beta,m)$. Il *successore* di ξ, che si indica con $\sigma(\xi)$, è "il primo elemento di $F(\beta,m)$ *a destra* di ξ." Il *predecessore* di ξ, che si indica con $\pi(\xi)$, è "il primo elemento di $F(\beta,m)$ *a sinistra* di ξ." Le funzioni σ e π, definite per ogni elemento non nullo di $F(\beta,m)$, si chiamano, rispettivamente, *funzione successore* e *funzione predecessore* e sono *una l'inversa dell'altra*.[4]

0.1.12 Esempio

Si consideri $F(10,3)$.

[3] Se $B \subset \mathbb{R}$ e $a \in \mathbb{R}$ allora: $aB = \{\, ax, x \in B\,\}$, ovvero aB è l'insieme che si ottiene moltiplicando ciascuno degli elementi di B per a.

[4] Più formalmente: il primo elemento di $F(\beta,m)$ *a destra* di ξ è *il più piccolo* elemento di $F(\beta,m)$ *maggiore* di ξ; il primo elemento di $F(\beta,m)$ *a sinistra* di ξ è *il più grande* elemento di $F(\beta,m)$ *minore* di ξ.

– Per $\xi = 10^{-2}\,0.501$ si ha $\sigma(\xi) = 10^{-2}\,0.502$ e $\pi(\xi) = 10^{-2}\,0.500$. Infatti: $\xi \in B_{-2}$, il primo elemento a destra di ξ in B_{-2} è quello con frazione 0.502 ed il primo elemento a sinistra è quello con frazione 0.500.

– Per $\xi = 10^{4}\,0.100$ si ha $\sigma(\xi) = 10^{4}\,0.101$ e $\pi(\xi) = 10^{3}\,0.999$. Il successore si ottiene ragionando come nel caso precedente. Per il predecessore si osserva che ξ è *il primo* elemento di B_4 e quindi il primo elemento a sinistra di ξ è *l'ultimo* elemento di B_3, quello con frazione 0.999.

– *Esercizio*: Sia b un numero intero. Determinare $\sigma(10^{b}\,0.999)$ e $\pi(10^{b+1}\,0.100)$.

– *Esercizio*: Determinare $\sigma(\max B_2)$ e $\pi(\min B_{-1})$.

– *Esercizio*: Sia $\xi \in (-1)B_3$. Dimostrare che $\sigma(\xi) = -\pi(-\xi)$ e $\pi(\xi) = -\sigma(-\xi)$.

0.1.13 Teorema (distribuzione degli elementi di $F(\beta, m)$)
Si consideri $F(\beta, m)$ e sia $\xi = \beta^b\, g$ un suo elemento *positivo*. Allora:

$$\sigma(\xi) - \xi = \beta^{b-m} \qquad \text{e} \qquad \frac{\sigma(\xi) - \xi}{\beta^b} = \beta^{-m}$$

La distanza tra elementi positivi consecutivi di $F(\beta, m)$ *aumenta* proporzionalmente all'ordine di grandezza β^b del primo elemento e, quindi, il rapporto tra la distanza e l'ordine di grandezza è un valore *costante* dipendente solo da β e m.

– *Dimostrazione*: La prima uguaglianza si ottiene considerando che, in ogni caso:

$$\sigma(\xi) = \beta^b\,(g + \beta^{-m})$$

La seconda uguaglianza si ottiene dalla prima.

0.1.14 Definizione (numeri in virgola mobile con esponente limitato ed elementi denormalizzati)
Siano β un numero intero maggiore di uno, m un numero intero positivo, $b_{\min}$ e $b_{\max}$ numeri interi tali che $b_{\min} < b_{\max}$.

Il sottoinsieme di $F(\beta, m)$ costituito da 0 e dagli elementi con esponente b *limitato*, $b_{\min} \leqslant b \leqslant b_{\max}$, si indica con:

$$F(\beta, m, b_{\min}, b_{\max})$$

e si chiama insieme dei numeri in virgola mobile (normalizzati) in base β e precisione m *con esponente limitato* tra $b_{\min}$ e $b_{\max}$.

Il sottoinsieme di $F(\beta, m)$ costituito dagli elementi con esponente b *limitato*, $b_{\min} \leqslant b \leqslant b_{\max}$, e da tutti i numeri reali x tali che:

$$x = (-1)^s \beta^{b_{\min}}\, 0.0 c_2 \cdots c_m$$

con $s \in \{\,0, 1\,\}$ e $c_2, \ldots, c_m$ cifre in base β, si indica con:

$$F_d(\beta, m, b_{\min}, b_{\max})$$

Gli elementi non nulli con esponente minore di $b_{\min}$ di dicono *denormalizzati*, e $F_d(\beta, m, b_{\min}, b_{\max})$ si chiama insieme dei numeri in virgola mobile in base β e precisione m con esponente limitato tra $b_{\min}$ e $b_{\max}$ *ed elementi denormalizzati*.

0.1.15 Osservazione
(1) L'insieme $F(\beta, m, b_{\min}, b_{\max})$ si ottiene da $F(\beta, m)$ *eliminando* gli elementi con esponente b maggiore di $b_{\max}$ e quelli con esponente b minore di $b_{\min}$. L'insieme $F(\beta, m, b_{\min}, b_{\max})$ ha allora un numero *finito* di elementi.

L'insieme $F_d(\beta, m, b_{\min}, b_{\max})$ si ottiene da $F(\beta, m, b_{\min}, b_{\max})$ *aggiungendo* gli elementi denormalizzati. Gli elementi denormalizzati sono un numero finito: anche l'insieme $F_d(\beta, m, b_{\min}, b_{\max})$ ha un numero *finito* di elementi.

Inoltre:

$$F(\beta, m, b_{\min}, b_{\max}) \subset F_d(\beta, m, b_{\min}, b_{\max}) \subset F(\beta, m)$$

(2) Sia $\xi \in F_d(\beta, m, b_{\min}, b_{\max})$. Se ξ ha esponente maggiore o uguale a $b_{\min}$ allora $0.c_1 \cdots c_m$ è la scrittura posizionale (in base β) della frazione di ξ. Se invece ξ ha esponente minore di $b_{\min}$

– ovvero ξ è un elemento denormalizzato – allora $0.0c_2 \cdots c_m$ *non è* la scrittura posizionale della frazione di ξ.

(3) L'insieme $F_d(\beta, m, b_{\min}, b_{\max})$ *include zero* perché:

$$0 = (-1)^s \beta^{b_{\min}}\, 0.0 \cdots 0$$

ovvero si ottiene zero scegliendo $c_1 = c_2 = \cdots = c_m = 0$.

0.1.16 Esempio

Per $F(10, 4, -99, 99)$ si ha:

- È *simmetrico rispetto a zero.*
- È *limitato*, $\xi_{\max} = \max F(10, 4, -99, 99) = 10^{99}\, 0.9999$ e la funzione successore *non è definita* in $\xi_{\max}$.[5]
- Zero *non è punto di accumulazione*, le funzioni successore e predecessore sono definite *anche* in zero e $\xi_{\min} = \sigma(0) = 10^{-99}\, 0.1000$. Quest'ultimo è *il più piccolo* elemento *positivo* dell'insieme considerato.
- *Esercizio*: Dimostrare che $F(10, 4, -99, 99)$ ha $199 \cdot 9000 = 1\,791\,000$ elementi *positivi.*

Per $F_d(10, 4, -99, 99)$ si ha:

- È *simmetrico rispetto a zero.*
- È *limitato*, $\xi_{\max} = \max F_d(10, 4, -99, 99) = \max F(10, 4, -99, 99) = 10^{99}\, 0.9999$ e la funzione successore *non è definita* in $\xi_{\max}$.[5]
- Zero *non è punto di accumulazione*, le funzioni successore e predecessore sono definite *anche* in zero e $\xi_{\min} = \sigma(0) = 10^{-99}\, 0.0001 = 10^{-102}\, 0.1$. Quest'ultimo è *il più piccolo* elemento *positivo* dell'insieme considerato, ed è un elemento denormalizzato. Il più piccolo elemento positivo *normalizzato* dell'insieme è $\xi^*_{\min} = 10^{-99}\, 0.1000$.
- *Esercizio*: Dimostrare che $F_d(10, 4, -99, 99)$ ha $199 \cdot 9000 + 999 = 1\,791\,999$ elementi *positivi.*

0.1.17 Osservazione (l'insieme M)

Abbiamo introdotto diversi insiemi di numeri in virgola mobile e precisione finita. Perché l'ipotesi che la sostituzione del tipo *numero reale* con il tipo *numero in virgola mobile e precisione finita* sia sufficiente per discutere l'uso del calcolatore, saranno opportune *scelte diverse* di M in contesti diversi.

Ad esempio:

- Nella discussione della realizzazione di una procedura in *Scilab* (*Matlab*, *Octave*) è opportuno scegliere $M = F_d(2, 53, -1021, 1024)$[6] perché questi sono gli oggetti del tipo di dato numerico che *Scilab* (*Matlab*, *Octave*) consente di usare.[7] Qualora nella discussione si ritenga *trascurabile* l'effetto della limitazione sull'esponente, si sceglierà $M = F(2, 53)$.
- I linguaggi *Matlab* e *Octave* realizzano anche il tipo di dato numerico `single` per il quale $M = F_d(2, 24, -125, 128)$.[8]
- Nella discussione della realizzazione di una procedura nel linguaggio della calcolatrice tascabile *HP 49G* è opportuno scegliere $M = F(10, 12, -498, 498)$ perché questi sono gli oggetti del tipo di dato numerico che la calcolatrice *HP 49G* consente di usare. Qualora nella discussione si ritenga *trascurabile* l'effetto della limitazione sull'esponente, si sceglierà $M = F(10, 12)$.

[5] Analogamente, la funzione predecessore *non è definita* in $\min F(10, 4, -99, 99) = -\xi_{\max}$.
[6] Questo è il formato "binary64" dello IEEE Standard for Floating-Point Arithmetic.
[7] Nei linguaggi *Matlab* e *Octave* questo tipo di dato si chiama `double`.
[8] Questo è il formato "binary32" dello IEEE Standard for Floating-Point Arithmetic.

E1 Determinare l'esponente e la frazione di *due quinti* in base tre.

E2 Operando come nell'Osservazione 0.1.6, e sfruttando la monotonia della funzione $t \mapsto t^3$, calcolare l'esponente e le prime tre cifre della scrittura posizionale della frazione di $\sqrt[3]{3}$, in base due.

E3 **S** (Tecnica di *riduzione dell'argomento* per la radice quadrata.) Siano x un numero reale positivo e β un numero intero maggiore o uguale a due. Mostrare che per calcolare $\sqrt{x}$ è sufficiente saper calcolare $\sqrt{y}$ per ogni $y \in [1/\beta, \beta)$.

E4 Indicare quali dei seguenti numeri reali appartengono ad $F(2,3)$: *uno*, *un terzo*, *meno un sedicesimo*, *tre sedicesimi*, *zero*, π.

E5 Determinare il numero di elementi dell'insieme:

$$\{\, \xi \in F(10,3) \text{ tali che } -10^{-6}\, 0.311 \leqslant \xi \leqslant -10^{-9}\, 0.581 \,\}$$

E6 Dimostrare che $F(2,2) \subset F(2,3)$ e che $F(10,1) \subset F(10,2)$. In generale:

$$n < m \qquad \Rightarrow \qquad F(\beta, n) \subset F(\beta, m)$$

La relazione tra insiemi di numeri in virgola mobile *in basi diverse* è meno semplice: si veda il prossimo esercizio.

E7 ★ Siano F_2 un insieme di numeri in virgola mobile e base due e F_{10} un insieme di numeri in virgola mobile e base dieci.

(a) Mostrare che $\frac{1}{10} \in F_{10}$ ma $\frac{1}{10} \notin F_2$ (si ricordi quanto stabilito nell'Esempio 0.1.4) e dedurne che sono falsi gli asserti $F_2 \supset F_{10}$ e $F_2 = F_{10}$.

(b) Mostrare che *per ogni intero positivo* k, 2^k *non è divisibile per* 10 (e quindi che la cifra delle unità dell'espansione decimale di 2^k è sempre non zero) e che *per ogni intero positivo* n *esiste* k *tale che* $2^k > 10^n$.

(c) Mostrare che, assegnata una precisione m, *tutti gli elementi di* $F(10,m)$ *maggiori od uguali a* $10^m = 10^{m+1}\, 0.1$ (ovvero tutti gli elementi positivi con esponente maggiore di m) *sono divisibili per dieci*. Questo asserto, insieme a quelli mostrati nel punto (b), provano che per k sufficientemente grande si ha $2^k \notin F_{10}$, e quindi che è falso anche l'asserto $F_2 \subset F_{10}$.

E8 ★ La dimostrazione dell'asserto (1) dell'Osservazione 0.1.10 prova che: *se* ξ *è un elemento positivo di* $F(\beta,m)$ *allora* $\xi = N/\beta^k$ *con* N *numero intero positivo e* k *numero intero non negativo.*

Utilizzare questo asserto per verificare che per ogni numero intero $m > 1$ si ha: un decimo non appartiene a $F(2,m)$ e un terzo non appartiene a $F(10,m)$.

E9 Sia $x = 3.7$ (scrittura in base dieci). Decidere se $x \in F(2,8)$.

E10 Mostrare che tutti gli elementi positivi di $F(2,4)$ con esponente maggiore o uguale a 4 sono interi, e poi determinare:

$$\max\{\, \xi \in F(2,4) \text{ tali che } \xi > 0 \text{ e } \xi \notin \mathbb{Z} \,\} \quad \text{e} \quad \min\{\, \alpha \in \mathbb{N} \text{ tali che } \alpha \notin F(2,4) \,\}$$

E11 ★ Siano $\mathtt{esp}, \mathtt{fraz} : F(\beta,m) \setminus \{\, 0 \,\} \to \mathbb{R}$ le funzioni definite da:

$$\mathtt{esp}(\xi) = \text{esponente di } \xi \quad , \quad \mathtt{fraz}(\xi) = \text{frazione di } \xi$$

Mostrare che per ogni elemento non nullo $\xi \in F(\beta,m)$ si ha $\mathtt{fraz}(\xi) \in F(\beta,m)$, ma che `esp` non ha la stessa proprietà. Per ciascuna di tali funzioni, decidere se sia monotona.

E12 Posto $\xi = 2^{-3}\ 0.1101 \in F(2,4)$, indicare per quali numeri interi n si ha $4^n\,\xi \in F(2,4)$.

E13 Si consideri $F(2,10)$. Determinare il numero di elementi positivi con esponente -6, ovvero il numero di elementi dell'insieme B_{-6}.

E14 Si consideri $F(2,3)$. Determinare:

$$\sigma(2^{-3}\,0.101) \quad , \quad \pi(2^{-3}\,0.101) \quad \text{e} \quad \sigma(2^{4}\,0.100) \quad , \quad \pi(2^{4}\,0.100)$$

Determinare poi:

$$\sigma(2^{-1}\,0.110) \quad , \quad \pi(-2^{-1}\,0.101)$$

e verificare che $\pi(-2^{-1}\,0.101) = -\sigma(2^{-1}\,0.101)$. Determinare infine:

$$\max B_{-2} \quad \text{e} \quad \min B_7$$

E15 Assegnate una base β ed una precisione m, dimostrare che:

$$\textit{per ogni } \xi \textit{ elemento non nullo di } F(\beta,m) \textit{ si ha: } \pi(-\xi) = -\sigma(\xi)$$

E16 Si consideri $F(2,3,-7,7)$. Determinare:

$$\sigma(1) \quad , \quad \pi(1) \quad , \quad \sigma(0) \quad , \quad \pi(0) \quad , \quad \sigma(2^{7}\,0.111) \quad , \quad \pi(2^{-7}\,0.100)$$

Determinare poi $\xi_{\max}$ e $\xi_{\min}$.

E17 Si consideri $F_d(2,3,-7,7)$. Determinare:

$$\sigma(1) \quad , \quad \pi(1) \quad , \quad \sigma(0) \quad , \quad \pi(0) \quad , \quad \sigma(2^{7}\,0.111) \quad , \quad \pi(2^{-7}\,0.100)$$

Determinare poi $\xi_{\max}$, $\xi_{\min}$ e $\xi^*_{\min}$ e di ciascuno indicare l'esponente e la frazione (in base due).

E18 ★ Sia ϕ la funzione definita, per ogni elemento non nullo di $F(\beta,m)$, da $\phi(\xi) = \sigma(\xi) - \xi$. Mostrare che per ogni ξ si ha $\phi(\xi) \in F(\beta,m)$. Discutere la monotonia della funzione ϕ.

E19 ♠ Utilizzare la funzione `number_properties` per verificare che in *Scilab* è opportuno scegliere $M = F_d(2,53,-1021,1024)$ e per determinare $\xi_{\max}$, $\xi_{\min}$ e $\xi^*_{\min}$.

E20 Sia $M = F(\beta,m)$. Discutere i seguenti asserti:

(1) Se $\xi \in M$, anche $\beta^2\xi \in M$;

(2) Gli intervalli $[\beta,\beta^2]$ e $[\beta^{10},\beta^{11}]$ contengono lo stesso numero di elementi di M.

E21 **S** Siano $\xi = \beta^b 0.c_1\cdots c_m$ un elemento positivo di $F(\beta,m)$, e x un numero reale positivo. Si ha, ovviamente:

$$\textit{se } x = \beta^b 0.c_1\cdots c_m\cdots \textit{ allora } \xi \le x < \sigma(\xi)$$

Dimostrare che:

$$\textit{se } \xi \le x < \sigma(\xi) \textit{ allora } x = \beta^b 0.c_1\cdots c_m\cdots$$

Unendo i due asserti si ottiene: x e ξ hanno lo *stesso esponente* b e le *stesse prime* m *cifre* $c_1,\ldots,c_m$ della scrittura posizionale della frazione in base β *se e solo se* $\xi \leqslant x < \sigma(\xi)$.

0.2 Funzione arrotondamento

Gli elementi di M sono utilizzati per *approssimare numeri reali*. L'approssimazione è realizzata tramite la funzione arrotondamento, descritta in questa sezione.

0.2.1 Definizione (Elementi di M adiacenti ad un numero reale).
Siano M un insieme di numeri in virgola mobile e precisione finita, ed x un numero reale *non* in M. Se M è un insieme con esponente limitato, sia anche $|x| < \xi_{\max} = \max M$. Si dicono *adiacenti ad* x i due elementi consecutivi di M tra i quali è compreso x.

0.2.2 Esempio
Si consideri $M = F(\beta, m)$ e sia $x \notin M$ un numero reale positivo. Se, in base β, $x = \beta^b\, 0.c_1c_2\cdots$ allora, posto $\xi_- = \beta^b\, 0.c_1\cdots c_m$ (l'elemento di M ottenuto da x *troncando* la scrittura della frazione alla m–esima cifra) e $\xi_+ = \sigma(\xi_-)$ si ha:

$$\xi_- < x < \xi_+$$

ovvero ξ_- e ξ_+ sono gli elementi di M adiacenti ad x.
Esercizio: Determinare gli elementi adiacenti ad $x = \sqrt{2} = 1.4142\cdots$ in $F(10, 3)$.

0.2.3 Definizione (Funzione arrotondamento).
Sia x un numero reale. L'*arrotondato* di x in M, che si indica con $\mathrm{rd}(x)$, è *l'elemento di M più vicino ad x*. Questa definizione è però *ambigua* in tutti i casi in cui $x \notin M$ è *equidistante* dai due elementi di M ad esso adiacenti. L'ambiguità è risolta operando una delle due seguenti scelte mutuamente esclusive:

(a) In tutti i casi di ambiguità, dette β la base e m la precisione dell'insieme M, si sceglie come arrotondato di x quello tra i due elementi adiacenti ad x che ha *cifra meno significativa pari*[9]; se questo non è possibile[10] si sceglie come arrotondato di x quello tra i due elementi adiacenti ad x *più lontano da zero* — questa scelta, sarà indicata con la sigla RTTE[11] ed è quella da operare quando si discute la realizzazione di una procedura in *Scilab* (*Matlab*, *Octave*);

(b) In tutti i casi di ambiguità si sceglie come arrotondato di x quello tra i due elementi adiacenti ad x *più lontano da zero* — questa scelta sarà indicata con la sigla RTTA ed è quella da operare quando si discute la realizzazione di una procedura nel linguaggio della calcolatrice tascabile *HP 49G*.

La funzione $\mathrm{rd} : \mathbb{R} \to M$ così definita si chiama *funzione arrotondamento* in M.

0.2.4 Esempio
Si consideri $M = F(2, 2)$ e sia $x = \frac{1}{10}$. Allora $\mathrm{rd}(x) = 2^{-3}\, 0.11 = \frac{3}{32}$. Infatti: come sappiamo (Esempio 0.1.4) si ha:

$$x = 2^{-3}\, 0.\overline{1100}$$

dunque $x \notin M$. Inoltre, come mostrato nell'Esempio 0.2.2, gli elementi adiacenti ad x sono: $\xi_- = 2^{-3}\, 0.11$ (l'elemento di M ottenuto da x *troncando* la scrittura della frazione alla seconda cifra) e $\xi_+ = \sigma(\xi_-) = 2^{-2}\, 0.10$. Poiché il punto medio del segmento di estremi ξ_-, ξ_+ è $2^{-3}\, 0.111 > x$ allora l'elemento di M più vicino ad x è ben definito: ξ_-.

0.2.5 Osservazione (Proprietà della funzione arrotondamento)
Siano M l'insieme dei numeri in virgola mobile e precisione finita ed $\mathrm{rd} : \mathbb{R} \to M$ la funzione arrotondamento in M scelti.

- La funzione rd *non è invertibile*. Ad esempio, se $\xi = \mathrm{rd}(x) \neq 0$ ($|x| < \xi_{\max}$ se M è un insieme con esponente limitato) allora, detto m_- il punto medio del segmento di estremi $\pi(\xi), \xi$ ed m_+ il punto medio del segmento di estremi $\xi, \sigma(\xi)$, l'insieme delle $y \in \mathbb{R}$ tali che $\mathrm{rd}(y) = \xi$ include l'intervallo non vuoto (m_-, m_+).
- La funzione rd è *dispari*: $\mathrm{rd}(-x) = -\mathrm{rd}(x)$. *Esercizio*: Verificare aiutandosi con un disegno!

[9] Per la definizione di *cifra meno significativa* si fa riferimento: nel caso di un elemento *normalizzato*, alla scrittura $(-1)^s\, \beta^b\, 0.c_1\cdots c_m$, con $c_1 \neq 0$; nel caso di un elemento *denormalizzato* alla scrittura $(-1)^s\, \beta^{b_{\min}}\, 0.0c_2\cdots c_m$ — in entrambi i casi la cifra meno significativa è c_m. La cifra meno significativa di *zero* è, per definizione, 0.
[10] Ad esempio, in $F(3, 2)$ sia $\xi = 3^2 \cdot 0.12$ che il suo successore $\sigma(\xi) = 3^2 \cdot 0.20$ hanno c_m pari.
[11] Le sigle RTTE e RTTA sono abbreviazioni, rispettivamente, dei termini *round ties to even* e *round ties to away* utilizzati nello standard IEEE Std 754-2019.

– La funzione rd è *non decrescente*: $x < y \Rightarrow \mathrm{rd}(x) \leqslant \mathrm{rd}(y)$. Infatti, detto I l'insieme dei numeri reali t tali che $\mathrm{rd}(t) = \mathrm{rd}(x)$ si ha: se $y \in I$ allora $\mathrm{rd}(x) = \mathrm{rd}(y)$, altrimenti $\mathrm{rd}(x) < \mathrm{rd}(y)$.

– $\mathrm{rd}(x) = x \Leftrightarrow x \in M$.

– Se $M = F(\beta, m)$ allora: $\mathrm{rd}(x) = 0 \Leftrightarrow x = 0$.

Esercizi

E22 Calcolare l'arrotondato di $\frac{1}{4}$ in $F(3, 2)$.

E23 ★ Sia $\xi = 3^b\, 0.c_1c_2c_3 \in F(3, 3)$. Detto m il punto medio del segmento di estremi ξ e $\sigma(\xi)$, mostrare (aiutandosi con la rappresentazione grafica di tutti i numeri considerati) che:

$$3^b\, 0.c_1c_2c_31 < m < 3^b\, 0.c_1c_2c_32 \quad , \quad 3^b\, 0.c_1c_2c_311 < m < 3^b\, 0.c_1c_2c_312 \quad , \quad \ldots$$

e quindi che:

$$m = 3^b\, 0.c_1c_2c_3\bar{1}$$

E24 Calcolare l'arrotondato di $2^2\, 0.1011$ in $F(10, 2)$.

E25 Calcolare l'arrotondato di $\frac{1}{2}\,\xi_{\min}$ in $F(2, 5, -9, 9)$.

E26 Calcolare l'arrotondato di $\frac{1}{2}\,\xi_{\min}$ in $F_d(2, 5, -9, 9)$.

E27 Sia rd la funzione arrotondamento in $F(10, 3)$ con RTTE. Determinare tutti gli $x \in \mathbb{R}$ tali che $\mathrm{rd}(x) = 642$.

E28 Sia rd la funzione arrotondamento in $F(10, 3)$ con RTTE. Determinare:

$$\max\{\, y \in \mathbb{R} \text{ tale che } \mathrm{rd}(314 + y) = 314 \,\}$$

E29 Sia M un insieme di numeri in virgola mobile con esponente limitato (ma *senza* elementi denormalizzati). Determinare l'arrotondato in M di $\frac{1}{2}\xi_{\min}$ con RTTE.

E30 Sia $M = F(10, 1)$. Determinare l'arrotondato in M di $x = \frac{19}{20}$ con RTTE.

E31 Sia $M = F(3, 2)$. Determinare l'arrotondato in M di $x = \frac{11}{2}$ con RTTE.

Si è detto che gli elementi di M sono utilizzati per *approssimare numeri reali*, e che l'approssimazione è realizzata dalla funzione arrotondamento. Per studiare *quantitativamente* l'approssimazione, introduciamo *misure* dell'errore commesso.

0.2.6 Definizione (funzioni errore)

Siano M l'insieme dei numeri in virgola mobile e precisione finita e rd la funzione arrotondamento in M scelti. La funzione δ tale che:

$$\delta(x) = \mathrm{rd}(x) - x$$

si chiama *funzione errore assoluto* ed è definita per ogni $x \in \mathbb{R}$. Le funzioni ϵ e η tali che:

$$\epsilon(x) = \frac{\mathrm{rd}(x) - x}{x} = \frac{\delta(x)}{x} \quad , \quad \eta(x) = \frac{\mathrm{rd}(x) - x}{\mathrm{rd}(x)} = \frac{\delta(x)}{\mathrm{rd}(x)}$$

si chiamano *funzioni errore relativo* e sono definite, rispettivamente, per ogni numero reale $x \neq 0$ e per ogni numero reale x tale che $\mathrm{rd}(x) \neq 0$.

La funzione errore assoluto è *dispari*, quelle errore relativo *pari*.

0.2.7 Esercizio

Sia $x = \frac{1}{3}$. Determinare l'errore assoluto $\delta(x)$ e gli errori relativi $\epsilon(x)$ e $\eta(x)$ commessi approssimando x con l'arrotondato di x in $F(10, 3)$.

0.2.8 Teorema (stime delle funzioni errore in $F(\beta, m)$)

Sia $M = F(\beta, m)$ e $x = \beta^b\, g$ un numero reale positivo. Si ha:

$$|\,\delta(x)\,| \leqslant \tfrac{1}{2}\,\beta^{b-m} \qquad , \qquad |\,\epsilon(x)\,| \leqslant \tfrac{1}{2}\,\beta^{1-m} \qquad , \qquad |\,\eta(x)\,| \leqslant \tfrac{1}{2}\,\beta^{1-m}$$

(*Infatti*: x è un numero reale positivo con esponente b dunque $\beta^b\,\beta^{-1} \leqslant x < \beta^{b+1}\,\beta^{-1}$; la disuguaglianza relativa alla funzione δ si ottiene immediatamente dal Teorema 0.1.13. Le altre disuguaglianze si ottengono utilizzando quella relativa a δ e considerando che il valore minimo per x e per $\mathrm{rd}(x)$ è $\beta^b\,\beta^{-1}$.)

La validità delle stime si estende per simmetria al caso $x < 0$.

0.2.9 Osservazione (stime in insiemi con esponente limitato ed elementi denormalizzati)

Siano assegnate la base β, la precisione m ed i valori minimo $b_{\min}$ e massimo $b_{\max}$ dell'esponente. Detti $\xi^*_{\min}$ il più piccolo elemento positivo di $F(\beta, m, b_{\min}, b_{\max})$ e $\xi_{\max}$ il più grande elemento di $F(\beta, m, b_{\min}, b_{\max})$ si ha:

$$[\xi^*_{\min}, \xi_{\max}] \cap F(\beta, m) = [\xi^*_{\min}, \xi_{\max}] \cap F(\beta, m, b_{\min}, b_{\max}) = [\xi^*_{\min}, \xi_{\max}] \cap F_d(\beta, m, b_{\min}, b_{\max})$$

Indicando con rd la funzione arrotondamento in $F(\beta, m)$, con rd_ℓ quella in $F(\beta, m, b_{\min}, b_{\max})$ e con rd_d quella in $F_d(\beta, m, b_{\min}, b_{\max})$ si ottiene allora:

$$\text{se} \quad \xi^*_{\min} \leqslant x \leqslant \xi_{\max} \quad \text{allora} \quad \mathrm{rd}(x) = \mathrm{rd}_\ell(x) = \mathrm{rd}_d(x)$$

Dunque: se $\xi^*_{\min} \leqslant x \leqslant \xi_{\max}$ allora le stime riportate nel Teorema 0.2.8 per le funzioni errore sussistono anche quando M è un insieme di numeri *con esponente limitato*. Se, invece, M è un insieme di numeri *con esponente limitato* ed x è un numero reale al di fuori dell'intervallo indicato, allora gli errori *possono non rispettare le limitazioni riportate*.

0.2.10 Definizione (precisione di macchina)

Sia M un insieme di numeri in virgola mobile e precisione finita. Si chiama *precisione di macchina* in M la quantità (determinata *solo* dalla base e dalla precisione dell'insieme dei numeri in virgola mobile):

$$u = \tfrac{1}{2}\,\beta^{1-m}$$

In termini di precisione di macchina, le stime riportate nel Teorema 0.2.8 si esprimono:

$$|\epsilon(x)| \leqslant u \qquad , \qquad |\eta(x)| \leqslant u$$

e, quindi:

$$|\delta(x)| \leqslant u\,|\,x\,| \qquad \text{oppure} \qquad |\delta(x)| \leqslant u\,|\,\mathrm{rd}(x)\,|$$

0.2.11 Esempio (precisione di macchina in $F(2, 53)$ e $F(10, 12)$)

In $F(2, 53)$ si ha $u = 2^{-53} \approx 10^{-16}$, in $F(10, 12)$ si ha: $u = 5 \cdot 10^{-12}$.

Il valore della precisione di macchina è *significativo* nel contesto dell'uso di elementi di $F(\beta, m)$ per approssimare numeri reali: tanto più *piccolo* è il valore della precisione di macchina quanto più *stringente* è, in base al Teorema 0.2.8, la limitazione dell'errore relativo commesso arrotondando numeri reali. Per i due insiemi in esame si ha:

$$\text{precisione di macchina in } F(2, 53) < \text{precisione di macchina in } F(10, 12)$$

dunque la limitazione dell'errore relativo commesso arrotondando numeri reali in $F(2, 53)$ è più stringente della limitazione dell'errore relativo commesso arrotondando numeri reali in $F(10, 12)$.

Ad esempio:

$$\text{in } F(2, 53)\text{: } \mathrm{rd}(\pi) = 3.14159265358979311599796346854418516159057617187 5$$

e:

$$\text{in } F(10, 12)\text{: } \mathrm{rd}(\pi) = 3.14159265359$$

Considerando che $\pi = 3.1415926535897932...$ si ottiene:

$$\text{in } F(2, 53)\text{: } |\,\epsilon(\pi)\,| < 10^{-16}$$

$$\text{in } F(10,12)\text{: } |\,\epsilon(\pi)\,| > 2\cdot 10^{-13}$$

e l'errore relativo in $F(2,53)$ è *minore* di quello in $F(10,12)$. Però:

$$\text{in } F(2,53)\text{: } \mathrm{rd}(0.1) = 0.1000000000000000055511151231257827021181583404541015625$$

e:

$$\text{in } F(10,12)\text{: } \mathrm{rd}(0.1) = 0.1$$

In questo caso:

$$\text{in } F(2,53)\text{: } |\,\epsilon(0.1)\,| = 0.55\cdots 10^{-16}$$

$$\text{in } F(10,12)\text{: } |\,\epsilon(0.1)\,| = 0$$

e l'errore relativo in $F(2,53)$ è *maggiore* di quello in $F(10,12)$.

Questo risultato non deve sorprendere: la precisione di macchina è soltanto una *limitazione superiore* per l'errore relativo.

0.2.12 Osservazione

Sia $M = F(\beta, m)$. Le funzioni errore relativo sono *limitate*: per ogni numero reale x non nullo l'errore relativo commesso approssimando x con $\mathrm{rd}(x)$ non supera la precisione di macchina, quantità *indipendente da* x. La funzione errore assoluto, invece, *non è limitata*. Questa differenza, *importante*, è conseguenza della struttura dell'insieme $F(\beta, m)$ e rende *naturale* misurare l'errore commesso approssimando un numero reale con un numero in virgola mobile e precisione finita con una funzione errore *relativo*.

– *Esercizio.*

Sia rd una funzione arrotondamento in $F(\beta, m)$. Disegnare il grafico delle funzioni $x \mapsto u$ e $x \mapsto u\,|\,x\,|$. Discutere il legame tra i grafici disegnati e quelli delle funzioni $x \mapsto |\,\epsilon(x)\,|$, $x \mapsto |\,\eta(x)\,|$ e $x \mapsto |\,\delta(x)\,|$.

0.2.13 Teorema (arrotondamento e perturbazioni)

Sia rd una funzione arrotondamento in $F(\beta, m)$ ed x un numero reale.

– Esiste un numero reale d tale che:

$$\mathrm{rd}(x) = x + d \qquad \text{e} \qquad |\,d\,| \leqslant u\,|\,x\,|$$

In questo caso si interpreta $\mathrm{rd}(x)$ come *perturbazione additiva* di x.

– Esiste un numero reale d tale che:

$$x = \mathrm{rd}(x) + d \qquad \text{e} \qquad |\,d\,| \leqslant u\,|\,x\,|$$

In questo caso si interpreta x come *perturbazione additiva* di $\mathrm{rd}(x)$.

– Esiste un numero reale e tale che:

$$\mathrm{rd}(x) = (1+e)\,x \qquad \text{e} \qquad |\,e\,| \leqslant u$$

In questo caso si interpreta $\mathrm{rd}(x)$ come *perturbazione moltiplicativa* di x.

– Esiste un numero reale t tale che:

$$x = (1+t)\,\mathrm{rd}(x) \qquad \text{e} \qquad |\,t\,| \leqslant u$$

In questo caso si interpreta x come *perturbazione moltiplicativa* di $\mathrm{rd}(x)$.

(*Infatti*: $|\,d\,| = |\,\delta(x)\,|$; $e = \epsilon(x)$ per $x \neq 0$, $e = 0$ per $x = 0$; $t = 0$ per $\mathrm{rd}(x) = 0$, $t = \eta(x)$ per $\mathrm{rd}(x) \neq 0$. Le limitazioni seguono dal Teorema 0.2.8.)

0.2.14 Osservazione

La stima della funzione errore relativo ϵ fornita nel Teorema 0.2.8 non è *ottima*, nel senso che non esiste $y \in \mathbb{R}$ tale che $\epsilon(y) = u$.

Una stima ottima per la funzione $\epsilon(x)$ è invece:

$$|\,\epsilon(x)\,| \leqslant \frac{u}{1+u}$$

(*Infatti*: x è un numero reale positivo con esponente b dunque $\beta^b\,\beta^{-1} \leqslant x < \beta^{b+1}\,\beta^{-1}$, e quindi $\mathrm{rd}(x) \geqslant \beta^{b-1}$. Se $\mathrm{rd}(x) = \beta^{b-1}$, allora: $|x| - \beta^{b-1} = |x - \mathrm{rd}(x)| = |\delta(x)|$. Se, invece, $\mathrm{rd}(x) > \beta^{b-1}$, allora: $|x| - \beta^{b-1} \geqslant \frac{1}{2}\,\beta^{b-m} \geqslant |\delta(x)|$. Dunque, in ogni caso si ha:

$$|x| - \beta^{b-1} \geqslant |\delta(x)| \quad \text{ovvero} \quad |x| \geqslant \beta^{b-1} + |\delta(x)|$$

Ne segue:

$$|\,\epsilon(x)\,| = \left|\,\frac{\delta(x)}{x}\,\right| \leqslant \frac{|\delta(x)|}{\beta^{b-1} + |\delta(x)|}$$

da cui, essendo $|\,\delta(x)\,| \leqslant \frac{1}{2}\,\beta^{b-m}$, si ottiene:

$$|\,\epsilon(x)\,| \leqslant \frac{\frac{1}{2}\,\beta^{b-m}}{\beta^{b-1} + \frac{1}{2}\,\beta^{b-m}} = \frac{u}{1+u}$$

Si osservi poi che, quale che sia la funzione arrotondamento utilizzata:

$$\epsilon(1+u) = \frac{u}{1+u}$$

e quindi la stima è ottima.

Esercizi

E32 Siano $x = \frac{5}{4}$ e rd la funzione arrotondamento in $F(2,2)$ con RTTE. Determinare $\mathrm{rd}(x)$ e gli errori assoluto e relativo commessi approssimando x con il suo arrotondato. Infine, verificare le limitazioni date degli errori nel Teorema 0.2.8 e le tesi del Teorema 0.2.13.

E33 ♠ Utilizzare la funzione `number_properties` per ottenere, da *Scilab*, la precisione di macchina e verificare, utilizzando la funzione `log2`, che tale precisione di macchina è 2^{-53}.

E34 Dimostrare che, detta u la precisione di macchina in $M = F(\beta, m)$, si ha:

$$\sigma(1) = 1 + 2u \qquad , \qquad \pi(1) = 1 - \frac{2u}{\beta}$$

E35 Sia $M = F(\beta, m)$. Discutere ciascuno dei seguenti asserti:

(1) l'errore relativo commesso approssimando $x \in \mathbb{R}$ con $\mathrm{rd}(x)$ è minore o uguale ad u;

(2) l'errore assoluto commesso approssimando $x \in \mathbb{R}$ con $\mathrm{rd}(x)$ è minore o uguale ad 1;

(3) ★ se $x \in \mathbb{R}$ e $\xi \in M$ sono tali che $\mathrm{rd}(x) = \xi$ allora $\mathrm{rd}(\beta^{12}x) = \beta^{12}\xi$.

E36 Siano $M = F(\beta, m)$, σ la funzione successore in M e u la precisione di macchina in M. Si constati che $\sigma(1) = 1 + 2u$, e quindi che $1 + u \notin M$. Verificare poi che:

(1) se rd è *una* funzione arrotondamento in M, si ha:

$$\epsilon(1+u) = \frac{u}{1+u}$$

(2) se β è un numero intero positivo pari e rd la funzione arrotondamento in M con RTTE, si ha:

$$\eta(1+u) = u$$

I risultati mostrano che la stima per la funzione η ottenuta nel Teorema 0.2.8 e quella per la funzione ϵ ottenuta nell'Osservazione 0.2.14, sono *ottime*.

0.3 Funzioni predefinite

Le *funzioni predefinite* sono le *funzioni* che il tipo *numero in virgola mobile e precisione finita* mette a disposizione per operare sugli elementi di M, gli *oggetti* del tipo.

Siano M un insieme di numeri in virgola mobile e precisione finita e rd una funzione arrotondamento in M.

0.3.1 Definizione (funzioni predefinite)
L'insieme delle *funzioni predefinite* è l'unione dei seguenti tre sottoinsiemi di funzioni su M:

- *Funzioni predefinite corrispondenti alle operazioni aritmetiche*

$$\oplus, \ominus, \otimes : M \times M \to M \qquad \text{tali che} \qquad \xi_1 \circledast \xi_2 = \mathrm{rd}(\xi_1 * \xi_2)$$

  e:

$$\oslash : M \times M \setminus \{\, 0 \,\} \to M \qquad \text{tale che} \qquad \xi_1 \oslash \xi_2 = \mathrm{rd}(\xi_1 / \xi_2)$$

- *Funzioni predefinite corrispondenti alle funzioni elementari*

  Sia $f : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}$, una *funzione elementare* (una funzione trigonometrica, esponenziale, logaritmica, radice n–esima, ...). *La* funzione predefinita corrispondente ad f è la funzione $\mathtt{F} : \Omega \cap M \to M$ definita da:

$$\mathtt{F}(\xi) = \mathrm{rd}\big(\, f(\xi) \,\big)$$

- *Funzioni predefinite corrispondenti ai confronti*

$$<, \leqslant, =, \neq, \geqslant, >: M \times M \to \{\, \mathtt{V}, \mathtt{F} \,\}$$

  Sono le *restrizioni* ad $M \times M$ delle corrispondenti funzioni sui numeri reali.[12]

Si osservi che anche in queste definizioni la funzione arrotondamento è utilizzata per approssimare un numero reale con un elemento di M. Inoltre le funzioni predefinite sono definite *nel modo migliore possibile* nel senso che "il valore di una funzione predefinita è l'elemento di M che *dista meno* dal risultato esatto."[13]

0.3.2 Esempio (Proprietà delle funzioni predefinite)
Le funzioni predefinite *non hanno* le stesse proprietà delle corrispondenti funzioni sui reali. Ad esempio, sia $M = F(10, 2)$. Si ha allora:

(A.1) $\oplus$ *è simmetrica* (per ogni $\xi_1, \xi_2 \in M$ si ha $\xi_1 \oplus \xi_2 = \xi_2 \oplus \xi_1$)

(A.2) $\oplus$ *non è associativa*: con $\xi_1 = 10^2\, 0.10$ e $\xi_2 = \xi_3 = 10^0\, 0.38$ si ha

$$(\xi_1 \oplus \xi_2) \oplus \xi_3 \neq \xi_1 \oplus (\xi_2 \oplus \xi_3)$$

(A.3) $\oplus$ *è debolmente monotona* (per ogni $\xi_1, \xi_2, \alpha \in M$ si ha $\xi_1 > \xi_2 \Rightarrow \xi_1 \oplus \alpha \geqslant \xi_2 \oplus \alpha$).

(A.4) "l'elemento zero non è unico:" esiste *un* solo elemento $\alpha \in M$ tale che per ogni $\xi \in M$ si ha $\xi \oplus \alpha = \xi$, precisamente $\alpha = 0$. Ma: per ogni $\xi \neq 0$ esiste $\alpha \neq 0$ tale che $\xi \oplus \alpha = \xi$ (ad esempio: $10^2\, 0.67 \oplus 10^{-2}\, 0.11 = 10^2\, 0.67$).

(A.5) per ogni $\xi \in M$ si ha $\xi \oplus (-\xi) = 0$, e "l'opposto è unico."

(M.1) $\otimes$ *è simmetrica* (per ogni $\xi_1, \xi_2 \in M$ si ha $\xi_1 \otimes \xi_2 = \xi_2 \otimes \xi_1$)

(M.2) $\otimes$ *non è associativa*: con $\xi_1 = 10^0\, 0.20, \xi_2 = 10^1\, 0.51$ e $\xi_3 = 10^1\, 0.76$ si ha

$$(\xi_1 \otimes \xi_2) \otimes \xi_3 \neq \xi_1 \otimes (\xi_2 \otimes \xi_3)$$

[12] I valori V e F sono codificati, rispettivamente, dagli elementi 1 e 0 di M. Dunque anche i confronti sono funzioni a valori in M.

[13] Le definizioni date delle funzioni predefinite corrispondenti alle operazioni aritmetiche, la funzione radice quadrata e quelle dei confronti rispecchiano fedelmente la realtà (lo standard IEEE Std 754–2019 le *impone*). Invece, le definizioni date delle funzioni predefinite corrispondenti alle rimanenti funzioni elementari possono essere *troppo stringenti* (lo standard le *raccomanda* – ma non *impone*): in casi concreti le funzioni predefinite corrispondenti alle funzioni elementari diverse dalla radice quadrata possono essere definite in modo leggermente diverso, quindi "peggiore" (si veda l'Esempio 0.3.4).

(M.3) $\otimes$ *è debolmente monotona* (per ogni $\xi_1, \xi_2, \alpha \in M$ con $\alpha > 0$, si ha $\xi_1 > \xi_2 \Rightarrow \xi_1 \otimes \alpha \geqslant \xi_2 \otimes \alpha$).

(M.4) "l'elemento unità non è unico:" per ogni $\xi \in M$ si ha $\xi \otimes 1 = \xi$, ma per ogni $\xi \neq 0$ esiste $\alpha \neq 1$ tale che $\xi \otimes \alpha = \xi$ (ad esempio, per $\xi = 10^0\, 0.49$ si ha: $\xi \otimes 10^0\, 0.99 = \xi$).

(M.5) sia $\xi \in M$ non zero: l'insieme degli elementi inversi di ξ

$$\{\, \theta \in M \text{ tali che } \xi \otimes \theta = 1 \,\}$$

può essere vuoto o avere più di un elemento: "l'elemento inverso può non esistere o non essere unico" (ad esempio, se $\xi = 10^0\, 0.20$ si ha: $\xi \otimes 10^1\, 0.50 = 1$ e $\xi \otimes 10^1\, 0.51 = 1$, ovvero ξ ha due elementi inversi; se $\xi = 10^1\, 0.89$ si ha: $\xi \otimes 10^0\, 0.11 = 10^0\, 0.98 < 1$ e $\xi \otimes 10^0\, 0.12 = 10^1\, 0.11 > 1$ e quindi, per la monotonia di $\otimes$ — (M.3) —, ξ non ha elemento inverso).

(F.1) La funzione predefinita `SEN`, corrispondente alla funzione elementare sen, *ha un solo zero*: $\xi = 0$ (infatti: l'uguaglianza $\texttt{SEN}(\xi) = 0$ equivale a $\mathrm{rd}(\operatorname{sen} \xi) = 0$ ovvero $\operatorname{sen} \xi = 0$, e $\xi = 0$ è l'unico elemento di M che la verifica).

(F.2) Il *Teorema di esistenza degli zeri* non si estende alle funzioni predefinite: Se $\phi : M \to M$ è una funzione predefinita corrispondente ad una funzione elementare *continua*, $\phi(\xi) < 0$ e $\phi(\theta) > 0$, *non è detto* che esista α tale che $\phi(\alpha) = 0$ (ad esempio: $1 \in M, 4 \in M$, $\texttt{SEN}(1) > 0$ e $\texttt{SEN}(4) < 0$ ma per ogni $\alpha \in M$ compreso tra 1 e 4 si ha $\texttt{SEN}(\alpha) \neq 0$).

0.3.3 Osservazione (errore relativo per le funzioni predefinite)

Siano $x \neq 0$ il risultato di una operazione aritmetica tra elementi di M o il valore di una funzione elementare in un elemento di M, e $\xi \in M$ il valore della corrispondente funzione predefinita. Se $M = F(\beta, m)$ allora *il valore assoluto dell'errore relativo commesso approssimando x con ξ non supera la precisione di macchina u.* Infatti:

$$\left| \frac{\xi - x}{x} \right| = \left| \frac{\mathrm{rd}(x) - x}{x} \right|$$

e, per il Teorema 0.2.8, l'ultima quantità non supera la precisione di macchina.

Lo stesso risultato vale se M è un insieme di numeri in virgola mobile e precisione finita *con esponente limitato* e $\xi^*_{\min} \leqslant |\, x \,| \leqslant \xi_{\max}$.

0.3.4 Esempio (funzioni predefinite in *Scilab* e nella calcolatrice *HP 49G*)

Si consideri la funzione elementare *radice quadrata.*

Nel linguaggio della calcolatrice tascabile *HP 49G* è disponibile la funzione predefinita $\sqrt{}$ e si ottiene, ad esempio:

$$\sqrt{}\texttt{2 = 1.41421356237}$$

che coincide ($\sqrt{2} = 1.41421356237\ 3095\ 04880 \cdots$) con l'arrotondato di $\sqrt{2}$ in $F(10, 12)$.

Nel linguaggio *Scilab* è disponibile la funzione predefinita `sqrt` e si ottiene, ad esempio:

```
sqrt(2) = 1.414213562373095 145474621858738828450413604736328125
```

che coincide con l'arrotondato di $\sqrt{2}$ in $F(2, 53)$. Infatti, esprimendo le frazioni in base due si ha:

$$\sqrt{2} = 2^1\, 0.10110101000001001111001100110011111111001110111100 1100 1001 \cdots$$

e:

$$\texttt{sqrt(2)} = 2^1\, 0.10110101000001001111001100110011111111001110111100 1101$$

In questi casi la Definizione 0.3.1 rispecchia la realtà.

Si consideri, invece, la funzione elementare *logaritmo in base dieci.*

Nel linguaggio *Scilab* è disponibile la funzione predefinita corrispondente `log10` ma, ad esempio, si ottiene (si osservi che $10^{15} \in F(2, 53)$):

```
log10(10^15) = 14.999999999999998223643160599749535322189331054687 5
```

che *non* coincide con l'arrotondato di $\log_{10} 10^{15}$ in $F(2, 53)$ – infatti: $\mathrm{rd}(\log_{10} 10^{15}) = 15$. La definizione della funzione predefinita è quindi *diversa* da quella della Definizione 0.3.1.

Si ha inoltre:

$$\sigma\big(\texttt{log10(10\^{}15)}\big) = 15 = \mathrm{rd}(\log_{10} 10^{15})$$

e per l'errore relativo commesso approssimando $\log_{10} 10^{15}$ con `log10(10^15)`, detta u la precisione di macchina in $F(2, 53)$ si ha:

$$\left| \frac{\pi(15) - 15}{15} \right| = \frac{2^{-49}}{15} = \tfrac{16}{15}\, u$$

Questo valore è *leggermente più grande* del massimo conseguente alla Definizione 0.3.1.

Esercizi

E37 Sia $M = F(10, 2)$. Dimostrare, utilizzando le proprietà della funzione rd che:

(1) Per ogni ξ si ha: $\xi \oplus (-\xi) = 0$;

(2) Per ogni ξ esiste un solo α tale che: $\xi \oplus \alpha = 0$.

E38 ★ Sia $M = F(\beta, m)$. Discutere ciascuno dei seguenti asserti:

(1) Se ξ ed α sono due elementi positivi di M allora $\xi \oplus \alpha > \xi$;

(2) La funzione predefinita `COS`, corrispondente alla funzione elementare cos, *non ha zeri*.

0.4 Il procedimento di trasformazione e lo studio dell'errore

In questa sezione descriviamo il procedimento per trasformare una procedura che usa il tipo *numero reale* in una che usa il tipo *numero in virgola mobile e precisione finita* e mostriamo, in alcuni semplici casi, *come ottenere informazioni sull'errore* commesso approssimando i valori delle variabili nella procedura che usa il tipo *numero reale* con i valori delle variabili nella procedura, ottenuta dal procedimento di trasformazione, che usa il tipo *numero in virgola mobile e precisione finita*.

A Il procedimento di trasformazione

Siano M un insieme di numeri in virgola mobile e precisione finita ed rd una funzione arrotondamento in M. Il procedimento di trasformazione di una procedura che usa il tipo *numero reale* in una che usa il tipo *numero in virgola mobile e precisione finita* consiste in:

(a) Sostituire a ciascuna costante a valore in $\mathbb{R}$ il suo arrotondato in M;

(b) Sostituire a ciascuna operazione aritmetica o funzione elementare la corrispondente funzione predefinita aggiungendo, se è il caso, *opportune precedenze tra operatori*.

0.4.1 Esempio

(1) Si consideri la procedura seguente, che usa il tipo *numero reale*:

```
x = π;
per i = 1,...,3 ripeti:
    x = x / i;
    y = sen(x) cos(x);
fine
```

Il procedimento di trasformazione fornisce quest'altra procedura che usa il tipo *numero in virgola mobile e precisione finita*:

```
x = rd(π);
per i = rd(1),...,rd(3) ripeti:
    x = x ⊘ i;
    y = SEN(x) ⊗ COS(x);
fine
```

Si osservi che *la descrizione* della procedura trasformata *non dipende* dalla scelta di M ed rd, ma ne dipende *l'effetto dell'esecuzione*. Ad esempio, il valore della variabile `x` dopo il primo assegnamento è diverso a seconda se $M = F(2, 53)$ oppure $M = F(10, 12)$ – si veda l'Esempio 0.2.11. Analogamente, dopo l'esecuzione della procedura in *Scilab* si ottiene:

$$\texttt{y = 0.43301270189 22192 98294151030859211459755897521972656 25}$$

mentre dopo l'esecuzione con la calcolatrice *HP 49G* si ha:

$$\texttt{y = 0.433012701893}$$

Il valore di `y` dopo l'esecuzione della procedura originale è:

$$\texttt{y} = \operatorname{sen}\tfrac{\pi}{6}\,\cos\tfrac{\pi}{6} = \tfrac{\sqrt{3}}{4} = 0.43301270189221923\cdots$$

(2) Si consideri la procedura seguente, che usa il tipo *numero reale*:

$\texttt{x} = \sqrt{2}$

Il procedimento di trasformazione fornisce quest'altra procedura che usa il tipo *numero in virgola mobile e precisione finita*:

`x = SQRT(`rd(2)`)`

Tenuto conto che $2 \in F(2, 53)$, il valore di `x` dopo l'esecuzione in *Scilab* è `sqrt(2)` ovvero, si veda l'Esempio 0.3.4:

$$\texttt{x = 1.41421356237309514547462185873882845044136047363 28125}$$

Analogamente, tenuto conto che $2 \in F(10, 12)$, il valore di `x` dopo l'esecuzione con la calcolatrice tascabile *HP 49G* è $\sqrt{2}$ ovvero, si veda ancora l'Esempio 0.3.4:

$$\texttt{x = 1.41421356237}$$

(3) Si consideri la procedura seguente, che usa il tipo *numero reale*:

$\texttt{x} = \log_{10} 10^{15}$

Il procedimento di trasformazione fornisce quest'altra procedura che usa il tipo *numero in virgola mobile e precisione finita*:

`x = LOG10(`rd(10^{15})`)`

Tenuto conto che $10^{15} \in F(2, 53)$, il valore di `x` dopo l'esecuzione in *Scilab* è `log10(`10^{15}`)` ovvero, si veda l'Esempio 0.3.4:

$$\texttt{x = 14.999999999999998223643160599749535322189331054 6875}$$

Analogamente, tenuto conto che $10^{15} \in F(10, 12)$, il valore di `x` dopo l'esecuzione con la calcolatrice tascabile *HP 49G* è `LOG(`10^{15}`)` ovvero:

$$\texttt{x = 15}$$

(4) Si consideri la procedura seguente, che usa il tipo *numero reale*:

$\texttt{u} = 2^{-53}\texttt{;}$
`a = -u;`
`b = u;`
`x = a + b + 1;`
`y = a + (b + 1);`

Il procedimento di trasformazione fornisce quest'altra procedura che usa il tipo *numero in virgola mobile e precisione finita*:

```
u = rd(rd(2)^rd(−53));
a = -u;
b = u;
x = (a ⊕ b) ⊕ rd(1);
y = a ⊕ (b ⊕ rd(1));
```

In questo caso, nell'assegnamento che definisce il valore di `x`, il procedimento di trasformazione, oltre a sostituire le operazioni di somma (associativa) con i corrispondenti operatori di pseudo-somma (*non* associativa: asserto(A.2) dell'Esempio 0.3.2) *deve* aggiungere una precedenza tra i due operatori. Quale precedenza sia opportuno adottare dipende dal contesto. Nel caso in esame si è adottata la precedenza (implicitamente) usuale nella discussione della realizzazione della procedura in *Scilab*. Dopo l'esecuzione della procedura in *Scilab* si ha poi:

```
x = 1   ,   y = 0.99999999999999988897769753748434595763683319091796875
```

ovvero `x` $\neq$ `y`.

– *Esercizio*
Verificare, utilizzando la funzione `nearfloat`, che `y` $= \pi(1)$.

B Studio dell'errore

In questa sezione consideriamo il caso elementare e frequente in cui la procedura consista nell'assegnamento `y` $= f(x)$ quando utilizza il tipo *numero reale* e nell'assegnamento `y` $= \phi(x)$ quando utilizza il tipo *numero in virgola mobile e precisione finita*, con f e ϕ funzioni opportune e x valore assegnato dell'argomento.

Più precisamente, scelti un insieme di numeri in virgola mobile M (con precisione di macchina u) ed una funzione arrotondamento rd, e date una funzione f da $\Omega \subset \mathbb{R}^n$ in $\mathbb{R}$ ed una funzione ϕ da Ω in M (detta *algoritmo*) tali che:

- esiste una sequenza *finita* di funzioni predefinite $\mathrm{fp}_1, \ldots, \mathrm{fp}_j$ tale che:

  $$\phi = \mathrm{fp}_j \circ \cdots \circ \mathrm{fp}_1 \circ \mathrm{rd}$$

  nel senso che per ogni $x \in \Omega$, il numero $\phi(x)$ è ottenuto arrotondando le componenti $x_1, \ldots, x_n$ ed utilizzando poi opportunamente, nell'ordine, le funzioni predefinite $\mathrm{fp}_1, \ldots, \mathrm{fp}_j$

- detta $f_1, \ldots, f_j$ la sequenza di funzioni elementari o operazioni aritmetiche corrispondente alla sequenza $\mathrm{fp}_1, \ldots, \mathrm{fp}_j$ si ha:

  $$f = f_j \circ \cdots \circ f_1$$

  nel senso che per ogni $x \in \Omega$, il numero $f(x)$ è ottenuto utilizzando opportunamente, nell'ordine, le funzioni $f_1, \ldots, f_j$

si considera il seguente problema: *per ogni $x \in \Omega$ tale che $f(x) \neq 0$, determinare informazioni sull'errore commesso approssimando $f(x)$ con $\phi(x)$*, ovvero sulla quantità:

$$e_t = \frac{\phi(x) - f(x)}{f(x)}$$

L'errore e_t, che dipende da x, si chiama *errore totale* commesso approssimando $f(x)$ con $\phi(x)$.

Dopo aver introdotto la nozione di *algoritmo accurato* utilizzeremo alcuni semplici esempi per discutere le nozioni di *algoritmo stabile* e *calcolo ben condizionato* e per mostrare come ottenere informazioni sull'errore.

Per semplicità, assumeremo che M sia un insieme di numeri in virgola mobile e precisione finita *con esponente non limitato*.

La nozione di algoritmo accurato è la formalizzazione dell'idea di "algoritmo che fornisce una buona approssimazione."

0.4.2 Definizione (*qualitativa* di algoritmo accurato)
Sia x un elemento di Ω tale che $f(x) \neq 0$.
L'algoritmo ϕ *è* accurato *quando utilizzato per approssimare* f *in* x *se, posto*:

$$\phi(x) = (1 + e_t)\, f(x) \qquad \textit{ovvero} \qquad e_t = \frac{\phi(x) - f(x)}{f(x)}$$

l'errore relativo e_t *risulta* piccolo*, ovvero se* $\phi(x)$ *è una* piccola *perturbazione moltiplicativa di* $f(x)$.
Si osservi che:

- Se $f(x) = 0$ e $\phi(x) \neq 0$ *non è possibile* interpretare $\phi(x)$ come perturbazione *moltiplicativa* di $f(x)$. In questo caso la nozione di *accuratezza* va definita interpretando $\phi(x)$ come perturbazione *additiva* di $f(x)$.
- Se $f(x) = 0$ e $\phi(x) = 0$ la relazione:

  $$\phi(x) = (1 + e_t)\, f(x)$$

  è verificata per $e_t = 0$ e $\phi(x)$ *è* una piccola perturbazione moltiplicativa di $f(x)$. In questo caso si può estendere la definizione e ritenere ϕ un algoritmo accurato quando utilizzato per approssimare f in x.

0.4.3 Osservazione
La definizione di algoritmo accurato è *qualitativa* perché non si è dato un significato quantitativo all'aggettivo *piccolo*. Tenuto conto che la migliore approssimazione di $f(x)$ in M è l'arrotondato $\mathrm{rd}\big(f(x)\big)$ e che per il Teorema 0.2.13 si ha:

$$\mathrm{rd}\big(f(x)\big) = (1 + e_t)\, f(x) \qquad \text{con} \qquad |e_t| \leqslant u$$

l'*unità di misura* da usare per stabilire se l'errore e_t risulta piccolo è la *precisione di macchina* u. Dunque, nella definizione precedente, l'errore e_t risulta piccolo se in valore assoluto non supera un multiplo *non troppo grande* di u.

0.4.4 Esempio
Si consideri la procedura che, assegnati numeri reali positivi r ed h, determina la superficie del cilindro circolare retto in cui r è il raggio della base e h l'altezza. La procedura consiste nel semplice assegnamento:

`y =` $f(r, h)$

con:

$$f(R, H) = 2\pi R^2 + 2\pi RH = 2\pi R\,(R + H)$$

Si *scelga* come algoritmo per approssimare il valore $f(r, h)$ la funzione da $\mathbb{R}^2$ in M definita da:

$$\phi(R, H) = 2 \overset{4}{\otimes} \mathrm{rd}(\pi) \overset{3}{\otimes} \mathrm{rd}(R) \overset{2}{\otimes} (\mathrm{rd}(R) \overset{1}{\oplus} \mathrm{rd}(H))$$

Posto $\mathrm{rd}(r) = \hat{r}$ e $\mathrm{rd}(h) = \hat{h}$, ricordando la Definizione 0.3.1 di funzioni predefinite corrispondenti alle operazioni aritmetiche ed utilizzando ripetutamente il Teorema 0.2.13 che consente di scrivere l'arrotondato di un numero reale come un'opportuna perturbazione moltiplicativa del numero reale, si riscrive:

$$\mathrm{rd}(\pi) = (1 + \theta)\pi \qquad \text{con} \qquad |\theta| \leqslant u$$

e:

$$\phi(r, h) = (1 + e_4)(1 + e_3)(1 + e_2)(1 + e_1)(1 + \theta)\, 2\pi\, \hat{r}\, (\hat{r} + \hat{h}) \qquad \text{con} \qquad |e_k| \leqslant u \text{ per } k = 1, \ldots, 4$$

Posto:

$$(1 + e_4)(1 + e_3)(1 + e_2)(1 + e_1)(1 + \theta) = 1 + e_v$$

si ottiene:

$$\phi(r,h) = (1+e_v)\, f(\hat{r},\hat{h}) \qquad \text{e} \qquad |e_v| \leqslant 5u + \cdots \approx 5u$$

Quest'ultima uguaglianza consente di interpretare $\phi(r,h)$ come *approssimazione accurata del valore di f in un punto vicino a (r,h).*

0.4.5 Definizione (*qualitativa* di algoritmo stabile)

Sia x un elemento di Ω diverso da zero.[14]

L'algoritmo ϕ è stabile *quando utilizzato per approssimare f in x se esistono numeri reali* piccoli e_v, e_a (dipendenti da x) *tali che*:

$$\phi(x) = (1+e_v)\, f\big((1+e_a)\, x\big)$$

*ovvero se $\phi(x)$ è un'*approssimazione accurata *del valore di f in un punto* vicino *ad x.*[15]

Si osservi che:

- Se $x = 0$ la proprietà di *stabilità* coincide con quella di *accuratezza*. Per ottenere una nozione più utile la *stabilità* va in questo caso riformulata introducendo una perturbazione *additiva* di x.
- Se $x = 0$, $f(0) = 0$ e $\phi(0) = 0$ la relazione:

  $$\phi(0) = (1+e_v)\, f\big((1+e_a)0\big)$$

  è verificata per $e_a = e_v = 0$, cioè: $\phi(0)$ *è* un'approssimazione accurata del valore di f in un punto vicino a 0. In questo caso si può estendere la definizione e ritenere ϕ un algoritmo stabile quando utilizzato per approssimare f in $x = 0$.

0.4.6 Osservazione

La definizione di algoritmo stabile, anch'essa *qualitativa* perché non è dato un significato quantitativo all'aggettivo *piccolo*, formalizza l'idea di "algoritmo buono."

Si osservi che, assegnati $f : \mathbb{R} \to \mathbb{R}$ ed $x \in \mathbb{R}$, la migliore approssimazione di $f(x)$ in M è $\mathrm{rd}\big(f(x)\big)$ *ma non è ragionevole* sperare di ottenere, utilizzando il calcolatore, un'approssimazione migliore di:

$$\mathrm{rd}\Big(f\big(\mathrm{rd}(x)\big)\Big)$$

ovvero *dell'elemento di M più vicino al valore di f nel punto di M più vicino ad x.* Dunque, è un "buon algoritmo" quello che restituisce *una buona approssimazione* del valore di f in un punto *vicino* ad x.

Tenuto conto che, utilizzando il Teorema 0.2.13:

$$\mathrm{rd}\Big(f\big(\mathrm{rd}(x)\big)\Big) = (1+e_v)\, f\big((1+e_a)\, x\big) \qquad \text{con} \qquad |e_v| \leqslant u \text{ e } |e_a| \leqslant u$$

anche in questo caso l'*unità di misura* da usare per stabilire se le perturbazioni e_v ed e_a risultano piccole è la *precisione di macchina* u. Dunque, nella definizione precedente, le perturbazioni risultano piccole se ciascuna in valore assoluto non supera un multiplo "non troppo grande" di u.

0.4.7 Esempio (continuazione)

Si è mostrato che:

$$\phi(r,h) = (1+e_v)\, f(\hat{r},\hat{h}) \qquad \text{con} \qquad |e_v| \leqslant 5u + \cdots \approx 5u$$

dunque, tenuto conto che:

$$\hat{r} = \mathrm{rd}(r) = (1+\rho)r \qquad \text{con} \qquad |\rho| \leqslant u$$

e:

$$\hat{h} = \mathrm{rd}(h) = (1+\omega)h \qquad \text{con} \qquad |\omega| \leqslant u$$

l'algoritmo ϕ, quando utilizzato per approssimare f in (r,h), è *stabile.*

[14] La definizione è data nel caso di f funzione di *una* variabile. Le modifiche da apportare nel caso generale sono ovvie.

[15] Il pedice v ricorda che e_v si riferisce al *valore* di f, il pedice a che e_a si riferisce all'*argomento* di f.

Per decidere se sia anche accurato occorre indagare se $f(\hat{r},\hat{h})$ sia una approssimazione accurata di $f(r,h)$, ovvero se *esiste un numero reale* e_p^f *piccolo tale che*:[16]

$$f(\hat{r},\hat{h}) = (1+e_p^f)\, f(r,h)$$

Si ottiene:

$$f(\hat{r},\hat{h}) = 2\pi\,\hat{r}\,(\hat{r}+\hat{h}) = 2\pi\,(1+\rho)r\,\big((1+\rho)r+(1+\omega)h\big)$$

da cui, introdotto l'errore relativo e_p^s commesso approssimando $r+h$ con $\hat{r}+\hat{h} = (1+\rho)r+(1+\omega)h$:

$$e_p^s = \frac{(1+\rho)r+(1+\omega)h-(r+h)}{r+h} = \frac{r}{r+h}\,\rho + \frac{h}{r+h}\,\omega$$

ovvero:

$$\hat{r}+\hat{h} = (1+\rho)r+(1+\omega)h = (1+e_p^s)\,(r+h)$$

si ottiene:

$$f(\hat{r},\hat{h}) = (1+\rho)(1+e_p^s)\,2\pi\,r\,(r+h) = (1+\rho)(1+e_p^s)\,f(r,h)$$

Tenuto conto delle limitazioni per ρ e ω e che r ed h sono numeri positivi si ottiene poi:

$$|e_p^s| = \left|\frac{r}{r+h}\,\rho + \frac{h}{r+h}\,\omega\right| \leqslant \left|\frac{r}{r+h}\right|\,|\rho| + \left|\frac{h}{r+h}\right|\,|\omega| \leqslant |\rho|+|\omega| \leqslant 2u$$

e infine, posto:

$$1+e_p^f = (1+\rho)(1+e_p^s) \qquad \text{ovvero} \qquad e_p^f = \rho + e_p^s + \rho e_p^s$$

si conclude:

$$f(\hat{r},\hat{h}) = (1+e_p^f)\,f(r,h) \qquad \text{con} \qquad |e_p^f| \leqslant 3u+2u^2 \approx 3u$$

dunque $f(\hat{r},\hat{h})$ è un'approssimazione accurata di $f(r,h)$.

Utilizzando i risultati ottenuti:

$$\phi(r,h) = (1+e_v)\,f(\hat{r},\hat{h}) = (1+e_v)(1+e_p^f)\,f(r,h)$$

e, posto:

$$1+e_t = (1+e_v)(1+e_p^f) \qquad \text{ovvero} \qquad e_t = e_v + e_p^f + e_v e_p^f$$

risulta:

$$\phi(r,h) = (1+e_t)\,f(r,h) \qquad \text{con} \qquad |e_t| \leqslant 8u+\cdots \approx 8u$$

ovvero: l'algoritmo ϕ, quando utilizzato per approssimare f in (r,h), è *accurato*.

Nell'esempio si è mostrato che $f(\hat{r},\hat{h})$ è una approssimazione accurata di $f(r,h)$. Questo è un caso particolare di una *proprietà locale* di f formalizzata dalla definizione seguente:

0.4.8 Definizione (*qualitativa* di calcolo ben condizionato)

Sia x un elemento di Ω diverso da zero e $f(x) \neq 0$.[17]

Il calcolo di f *in* x *è* ben condizionato *se per ogni numero reale* e_a piccolo, *posto*:

$$f\big((1+e_a)x\big) = (1+e_p^f)\,f(x) \qquad \textit{ovvero} \qquad e_p^f = \frac{f\big((1+e_a)x\big)-f(x)}{f(x)}$$

l'errore relativo e_p^f (dipendente sia da x che da e_a) *risulta* piccolo, *ovvero se* in ogni punto vicino *ad* x *il valore di* f *è un'*approssimazione accurata *di* $f(x)$.

Si osservi che se x è uno zero isolato di f *non è possibile* interpretare $f\big((1+e_a)x\big)$ come perturbazione *moltiplicativa* di $f(x)$. In questo caso la nozione di *calcolo ben condizionato* va definita interpretando $f\big((1+e_a)x\big)$ come perturbazione *additiva* di $f(x)$.

---

[16] L'errore e_p^f si chiama *errore propagato* da f: è l'errore sul valore di f causato dall'errore presente sull'argomento di f.

[17] La definizione è data nel caso di f funzione di *una* variabile. Le modifiche da apportare nel caso generale sono ovvie.

0.4.9 Osservazione

La definizione di calcolo ben condizionato, anch'essa *qualitativa* perché non è dato un significato quantitativo all'aggettivo *piccolo*, è simile a quella di funzione continua ed individua le funzioni f per le quali "il valore di f è poco sensibile a piccole variazioni dell'argomento intorno ad x."

Le tre nozioni sono legate dal seguente asserto, che formalizza il procedimento in due passi seguito negli Esempi 0.4.4 e 0.4.7.

0.4.10 Teorema (stabilità + buon condizionamento ⇒ accuratezza)

Sia x un elemento di Ω diverso da zero e tale che $f(x) \neq 0$.

Se ϕ è un algoritmo stabile quando utilizzato per approssimare f in x e il calcolo di f in x è ben condizionato, allora ϕ è accurato quando utilizzato per approssimare f in x.

(*Dimostrazione*: Per la stabilità si ha: esistono numeri reali e_v, e_a piccoli tali che:

$$\phi(x) = (1 + e_v)\, f\big((1 + e_a)x\big)$$

Poiché il calcolo di f in x è ben condizionato, posto:

$$e_p^f = \frac{f\big((1 + e_a)x\big) - f(x)}{f(x)} \qquad \text{ovvero} \qquad f\big((1 + e_a)x\big) = (1 + e_p^f)\, f(x)$$

l'errore relativo e_p^f risulta piccolo. Posto infine:

$$1 + e_t = (1 + e_v)(1 + e_p^f) \qquad \text{ovvero} \qquad e_t = e_v + e_p^f + e_v e_p^f$$

si ottiene:

$$\phi(x) = (1 + e_v)(1 + e_p^f) f(x) = (1 + e_t) f(x)$$

ed e_t risulta piccolo. Dunque l'algoritmo è accurato.)

0.4.11 Osservazione (condizionamento delle funzioni regolari)

Siano Ω un intervallo di $\mathbb{R}$, $f : \Omega \to \mathbb{R}$ una funzione regolare (ovvero: sufficientemente derivabile) e $x \in \Omega$ un numero reale diverso da zero e tale che $f(x) \neq 0$. Per ogni numero reale e_a tale che $(1 + e_a)x \in \Omega$ si ponga:

$$f\big((1 + e_a)x\big) = (1 + e_p^f)\, f(x) \qquad \text{ovvero} \qquad e_p^f = \frac{f\big((1 + e_a)x\big) - f(x)}{f(x)}$$

Per il Teorema di Lagrange, esiste un numero reale y tra x e $(1 + e_a)x$ tale che:

$$f\big((1 + e_a)x\big) - f(x) = f'(y)\, e_a x$$

dunque:

$$e_p^f = \frac{f'(y)\, e_a x}{f(x)}$$

Una *stima* di e_p^f si ottiene, nel caso in cui e_a è piccolo, ponendo $y = x$:

$$e_p^f \approx \frac{f'(x)}{f(x)}\, x\, e_a$$

Introdotto il *numero di condizionamento* del calcolo di f in x:

$$c_f(x) = \left| \frac{f'(x)}{f(x)}\, x \right|$$

si ottiene infine:

$$|e_p^f| \approx c_f(x)\, |e_a|$$

Lo studio del condizionamento del calcolo di $f(x)$ si riduce, in questi casi, allo studio di $c_f(x)$.

0.4.12 Esempio

Siano:

$$f(x) = \operatorname{sen} x \qquad , \qquad \phi(x) = \texttt{SEN(rd}(x)\texttt{)}$$

e $x \in (0, \frac{\pi}{2})$. Discutiamo stabilità e accuratezza dell'algoritmo ϕ quando utilizzato per approssimare i valori di f.

- *Stabilità dell'algoritmo ϕ quando utilizzato per approssimare $f(x)$*:

  Per il Teorema 0.2.13 esistono numeri reali e_a e e_v, entrambi in valore assoluto minori od uguali ad u, tali che

  $$\phi(x) = (1+e_v)\,\mathrm{sen}\,\big((1+e_a)x\big) = (1+e_v)f\big((1+e_a)x\big)$$

  L'algoritmo ϕ è dunque stabile per ogni $x \in (0, \frac{\pi}{2})$.

- *Condizionamento del calcolo di $f(x)$*:

  Sia e_a un numero reale piccolo. Poiché per ogni x la funzione f è regolare, per quanto detto nell'Osservazione 0.4.11, essendo:

  $$c_f(x) = \Big|\,\frac{x}{\tan x}\,\Big|$$

  si ha:

  $$f\big((1+e_a)x\big) = (1+e_p^f)\,f(x) \qquad \text{con} \qquad |e_p^f| \approx c_f(x)\,|e_a|$$

  Per giudicare il condizionamento del calcolo di $f(x)$ si studia la funzione $c_f(x)$. Per ogni $x \in (0, \frac{\pi}{2})$ si ha:

  $$|c_f(x)| = \Big|\frac{x}{\tan x}\Big| < 1$$

  dunque il calcolo di $f(x)$ è ben condizionato per ogni $x \in (0, \frac{\pi}{2})$.

- *Accuratezza dell'algoritmo ϕ quando utilizzato per approssimare $f(x)$*:

  In base al Teorema 0.4.10, l'algoritmo ϕ è accurato. Informazioni quantitative sull'errore commesso approssimando $f(x)$ con $\phi(x)$ si possono ottenere procedendo come nella dimostrazione del Teorema 0.4.10. Si ha:

  $$\phi(x) = (1+e_v)\,f\big((1+e_a)x\big) = (1+e_v)(1+e_p^f)\,f(x) = (1+e_t)\,f(x)$$

  e, utilizzando le limitazioni

  $$|e_v| \leqslant u \qquad , \qquad |e_p^f| \approx c_f(x)|e_a| \leqslant u$$

  ottenute nello studio della stabilità e del condizionamento, si ricava che, *approssimativamente*:

  $$|e_t| \leqslant 2u + u^2 \approx 2u$$

– *Esercizio*

  La funzione numero di condizionamento del calcolo di $\mathrm{sen}\,x$:

  $$c_f(x) = \Big|\,\frac{x}{\tan x}\,\Big|$$

  è definita per ogni $x \in \mathbb{R}$ non multiplo intero di π. Per ogni numero intero k diverso da zero si ha:

  $$\lim_{x \to k\pi} c_f(x) = +\infty$$

  Utilizzare *Scilab* per ottenere il (più correttamente: un'approssimazione del) grafico della funzione $c_f(x)$ per $x \in (0,\pi) \cup (\pi, 2\pi)$ e dedurne che il calcolo di $\mathrm{sen}\,x$ risulta ragionevolmente ben condizionato (e quindi, per il Teorema 0.4.10, l'algoritmo ϕ risulta accurato) per $x \in (0, \pi - h) \cup (\pi + h, 2\pi - h)$ con h non troppo piccolo.

0.4.13 Osservazione (condizionamento delle operazioni aritmetiche)

Sia $*$ un'operazione aritmetica e x_1, x_2 numeri reali tali che $x_1 * x_2 \neq 0$. Assegnati numeri reali e_1, e_2 si ponga:

$$(1+e_1)x_1 * (1+e_2)x_2 = (1+e_p^*)(x_1 * x_2) \quad \text{ovvero} \quad e_p^* = \frac{\big((1+e_1)x_1 * (1+e_2)x_2\big) - (x_1 * x_2)}{(x_1 * x_2)}$$

Con semplici passaggi si ottiene, per la *somma*:

$$e_p^s = \frac{x_1}{x_1+x_2}\, e_1 + \frac{x_2}{x_1+x_2}\, e_2$$

per la *moltiplicazione*:

$$e_p^m = e_1 + e_2 + e_1 e_2$$

e per la *divisione*:

$$e_p^d = \frac{e_1 - e_2}{1+e_2}$$

In base alla Definizione 0.4.8, il calcolo della moltiplicazione e della divisione è *sempre ben condizionato*. Infatti, per e_1 ed e_2 piccoli, per la moltiplicazione si ha:

$$|e_p^m| \leqslant |e_1| + |e_2| + |e_1|\,|e_2| \approx |e_1| + |e_2|$$

e per la divisione:

$$|e_p^d| \leqslant \frac{|e_1| + |e_2|}{1 - |e_2|} \approx |e_1| + |e_2|$$

Per il calcolo della somma, invece, il condizionamento del calcolo *dipende dagli addendi*:

– *Se gli addendi hanno lo stesso segno il calcolo è ben condizionato.* Infatti in tal caso si ha:

$$|e_p^s| \leqslant \left|\frac{x_1}{x_1+x_2}\right| |e_1| + \left|\frac{x_2}{x_1+x_2}\right| |e_2| \leqslant \max\{\,|e_1|, |e_2|\,\} \leqslant |e_1| + |e_2|$$

– *Se gli addendi hanno segno opposto, il condizionamento del calcolo può essere tanto peggiore quanto più il rapporto x_2/x_1 è vicino a -1.* Infatti, posto:

$$\frac{x_2}{x_1} = -1 + h$$

si ha:

$$\frac{x_1}{x_1+x_2} = \frac{1}{h} \qquad , \qquad \frac{x_2}{x_1+x_2} = 1 - \frac{1}{h}$$

e quindi:

$$\lim_{h\to 0}\left|\frac{x_1}{x_1+x_2}\right| = \lim_{h\to 0}\left|\frac{x_2}{x_1+x_2}\right| = +\infty$$

Ad esempio, siano $x_1 = 1 + 6\cdot 10^{-12}$ e $x_2 = -1$. Detta rd la funzione arrotondamento in $F(10,12)$ si approssima $x_1 + x_2$ con $\mathrm{rd}(x_1) + \mathrm{rd}(x_2)$. Si ottiene:

$$e_1 = \frac{\mathrm{rd}(x_1) - x_1}{x_1} = \frac{4\cdot 10^{-12}}{1 + 6\cdot 10^{-12}} \approx 4\cdot 10^{-12} \qquad , \qquad e_2 = \frac{\mathrm{rd}(x_2) - x_2}{x_2} = 0$$

e:

$$\frac{x_1}{x_1+x_2} = \frac{1 + 6\cdot 10^{-12}}{6\cdot 10^{-12}} \approx \tfrac{1}{6}\cdot 10^{12}$$

Infine:

$$x_1 + x_2 = 6\cdot 10^{-12} \qquad , \qquad \mathrm{rd}(x_1) + \mathrm{rd}(x_2) = 10\cdot 10^{-12}$$

e:

$$|e_p^s| = \left|\frac{10\cdot 10^{-12} - 6\cdot 10^{-12}}{6\cdot 10^{-12}}\right| = \tfrac{2}{3}$$

L'errore $|e_p^s|$ è *molto maggiore* dell'errore sui singoli addendi: il calcolo non è ben condizionato.

0.4.14 Osservazione (stabilità delle funzioni predefinite)

Siano f da $\Omega \subset \mathbb{R}$ in $\mathbb{R}$ una funzione elementare e fp la funzione predefinita corrispondente ad f. Per la Definizione 0.3.1, per ogni $\xi \in \Omega \cap M$ si ha: $\mathrm{fp}(\xi) = \mathrm{rd}\big(f(\xi)\big)$.

Il procedimento utilizzato nell'Esempio 0.4.12 per mostrare la stabilità prova che l'algoritmo ϕ definito da $\phi(x) = \mathrm{fp}(\mathrm{rd}(x))$ — definito nell'insieme $\Omega^* \subset \Omega$ dei punti $x \in \Omega$ tali che $\mathrm{rd}(x) \in \Omega$ — è *stabile* quando utilizzato per approssimare f per *ogni* $x \in \Omega^*$.

Siano ora $*$ un'operazione aritmetica, f da $\Omega \subset \mathbb{R}^2$ in $\mathbb{R}$ la funzione definita da $f(x_1, x_2) = x_1 * x_2$ e $\circledast$ la funzione predefinita (ovvero la pseudo-operazione aritmetica) corrispondente a $*$.

L'algoritmo ϕ definito da $\phi(x_1, x_2) = \mathrm{rd}(x_1) \circledast \mathrm{rd}(x_2)$ — definito nell'insieme $\Omega^* \subset \Omega$ dei punti $(x_1, x_2) \in \Omega$ tali che $\big(\mathrm{rd}(x_1), \mathrm{rd}(x_2)\big) \in \Omega$ — è *stabile* quando utilizzato per approssimare f per *ogni* $(x_1, x_2) \in \Omega^*$.

Infatti: per il Teorema 0.2.13 si ha che per ogni $(x_1, x_2) \in \Omega^*$ esistono numeri reali e_1, e_2 ed e_3 tali che:

$$\phi(x_1, x_2) = (1 + e_3)\big((1 + e_1)x_1 * (1 + e_2)x_2\big) = (1 + e_3) f\big((1 + e_1)x_1, (1 + e_2)x_2\big)$$

e:

$$|e_1| \leqslant u \quad , \quad |e_2| \leqslant u \quad , \quad |e_3| \leqslant u$$

Dunque: $\phi(x_1, x_2)$ è una piccola perturbazione moltiplicativa del valore di f in un punto vicino a (x_1, x_2). Quanto scritto costituisce precisamente l'estensione della definizione di stabilità di un algoritmo al caso di funzioni di più variabili.

Salvo casi particolarmente semplici, un algoritmo è definito *componendo* più funzioni predefinite. L'osservazione precedente mostra che gli "algoritmi elementari" che utilizzano una sola funzione predefinita *sono stabili*. La prossima osservazione ed il successivo esempio mostrano invece che la composizione di algoritmi stabili *non necessariamente* genera algoritmi a loro volta stabili e chiarisce perché ciò accade.

0.4.15 Osservazione (algoritmi *non* stabili)

Siano $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ due funzioni e ϕ_1, ϕ_2 due algoritmi *stabili* quando utilizzati per approssimare, rispettivamente, i valori di f_1 e f_2 per ogni x. Assegnato $x \in \mathbb{R}$, si vuole studiare la stabilità dell'algoritmo $\gamma = \phi_2 \circ \phi_1$ quando utilizzato per approssimare i valori della funzione $g = f_2 \circ f_1$ in x.

Tenuto conto della stabilità di ϕ_1 quando utilizzato per approssimare f_1 in x, esistono numeri reali e_{v1}, e_{a1} tali che:

$$\phi_1(x) = (1 + e_{v1})\, f_1\big((1 + e_{a1})x\big) \quad \text{con} \quad e_{v1} \text{ ed } e_{a1} \text{ piccoli}$$

Tenuto conto della stabilità di ϕ_2 quando utilizzato per approssimare f_2 in $\phi_1(x)$, esistono numeri reali e_{v2}, e_{a2} tale che:

$$\phi_2\big(\phi_1(x)\big) = (1 + e_{v2})\, f_2\big((1 + e_{a2})\, \phi_1(x)\big) \quad \text{con} \quad e_{v2} \text{ ed } e_{a2} \text{ piccoli}$$

Dunque esistono numeri reali $e_{v2}, e_{a2}, e_{v1}, e_{a1}$ piccoli tali che:

$$\gamma(x) = (1 + e_{v2})\, f_2\Big((1 + e_{a2})(1 + e_{v1})\, f_1\big((1 + e_{a1})x\big)\Big)$$

Per leggere $\gamma(x)$ come perturbazione moltiplicativa del valore di g in un opportuno punto si riscrive:

$$f_2\Big((1 + e_{a2})(1 + e_{v1})\, f_1\big((1 + e_{a1})x\big)\Big) = (1 + e_p^{f_2})\, f_2\Big(f_1\big((1 + e_{a1})x\big)\Big)$$

con $e_p^{f_2}$ numero reale opportuno, certamente esistente se $f_2\Big(f_1\big((1 + e_{a1})x\big)\Big) \neq 0$, cosicché:

$$\gamma(x) = (1 + e_{v2})(1 + e_p^{f_2})\, f_2\Big(f_1\big((1 + e_{a1})x\big)\Big) = (1 + e_{v2})(1 + e_p^{f_2})\, g\big((1 + e_{a1})x\big)$$

Infine, ponendo $(1 + e_{v2})(1 + e_p^{f_2}) = 1 + e_v$ ovvero $e_v = e_{v2} + e_p^{f_2} + e_p^{f_2}\, e_{v2}$ si ottiene:

$$\gamma(x) = (1 + e_v)\, g\big((1 + e_{a1})x\big)$$

Per giudicare la stabilità di γ occorre decidere se e_v, ovvero $e_p^{f_2}$, sia piccolo. In altri termini occorre indagare il *condizionamento* del calcolo di f_2 in $f_1\big((1 + e_{a1})x\big)$:

– Se il calcolo di f_2 in $f_1\big((1+e_{a1})x\big)$ è *ben condizionato* allora $e_p^{f_2}$ risulta piccolo. Dunque anche e_v lo è e l'algoritmo γ è *stabile*.

– Se il calcolo di f_2 in $f_1\big((1+e_{a1})x\big)$ *non è* ben condizionato allora l'algoritmo γ *può* risultare *non stabile.*

0.4.16 Esempio

Si consideri $M = F(2, 53)$, e siano:

$$f(x) = 1 - \cos x \qquad , \qquad \phi(x) = 1 \overset{2}{\ominus} \overset{1}{\texttt{COS}}\big(\mathrm{rd}(x)\big)$$

e $\xi = 2^k (\in M)$ con $k \in \mathbb{Z}$ tale che $\texttt{COS}(\xi) = \mathrm{rd}(\cos\xi) = 1$.[18] Si utilizzi ϕ per approssimare il valore di f in ξ.

Si ha: $\phi(\xi) = 0$. Se l'algoritmo ϕ fosse stabile quando utilizzato per approssimare il valore di f in ξ, esisterebbero due numeri reali e_v, e_a *piccoli* (in particolare: $|e_v| < 1, |e_a| < 1$) tali che:

$$0 = \phi(\xi) = (1+e_v)\Big(1 - \cos\big((1+e_a)\xi\big)\Big)$$

Poiché $|e_v| < 1$, non può essere $1 + e_v = 0$. Allora dovrebbe essere $1 - \cos\big((1+e_a)\xi\big) = 0$, ovvero $\cos\big((1+e_a)\xi\big) = 1$. Poiché $|e_a| < 1$ si ha: $0 < 1 + e_a < 2$ e quindi: $0 < (1+e_a)\xi < 2\xi < 2\pi$, dunque $\cos\big((1+e_a)\xi\big) \neq 1$. Se ne deduce che *non possono esistere* e_v, e_a con le proprietà richieste, ovvero che *l'algoritmo ϕ non è stabile quando utilizzato per approssimare il valore di f in ξ.*

Il risultato è coerente con l'osservazione precedente. Infatti: il calcolo di $f_2(y) = 1 - y$ in $y = \cos\xi$ è *mal condizionato.* Per dimostrarlo, basta constatare che per il numero di condizionamento di f_2 in $\cos\xi$ si ha:

$$\left|\frac{\cos\xi}{1-\cos\xi}\right| > \frac{4}{u} - 1 \approx 3 \cdot 10^{16}$$

Esercizi

E39 Tenuto conto che 2^{-53} è la precisione di macchina in $F(2, 53)$, spiegare i risultati del punto (4) dell'Esempio 0.4.1.

E40 ★ Realizzando la procedura del punto (4) dell'Esempio 0.4.1 con la calcolatrice *HP 49G* si ottiene `x` = `y` = 1. Spiegare questi risultati e poi indicare come modificare l'assegnamento che definisce il valore di `u` in modo da ottenere anche in questo caso `x` ≠ `y`.

E41 ♠**S** L'insieme M in *Scilab* è $F_d(2, 53, -1021, 1024)$ (Osservazione 0.1.17), dunque il massimo di M è $2^{1024}(1 - 2^{-53})$. Spiegare perché, eseguendo in *Scilab* la procedura:

$$\texttt{x = } 2^{1024}(1 - 2^{-53})$$

il valore di `x` dopo l'assegnamento è `Inf`.

E42 Per ogni $x > 0$ sia $f(x) = 1/\sqrt{x}$. Determinare il *numero di condizionamento* del calcolo di f in $x > 0$ e discutere il condizionamento del calcolo al variare di x.

E43 Per ogni $x > 0$ sia $f(x) = 1/\sqrt{x}$. Determinare l'insieme di definizione e discutere stabilità ed accuratezza dell'algoritmo:

$$\phi(x) = 1 \oslash \texttt{SQRT(}\mathrm{rd}(x)\texttt{)}$$

quando utilizzato per approssimare i valori di f.

E44 Per ogni $x > 0$ sia $f(x) = \sqrt{x}/x = 1/\sqrt{x}$. Determinare l'insieme di definizione e discutere stabilità e accuratezza dell'algoritmo:

$$\phi(x) = \texttt{SQRT(}\mathrm{rd}(x)\texttt{)} \oslash \mathrm{rd}(x)$$

quando utilizzato per approssimare i valori di f.

[18] Un numero intero k che verifica la proprietà richiesta esiste certamente. Infatti: si consideri la successione $\xi_n = 2^{-n}$ di elementi di M. Si ha: $\lim_{n\to\infty}\xi_n = 0$ e quindi, per la continuità della funzione coseno: $\lim_{n\to\infty}\cos(\xi_n) = 1$. Allora esiste certamente un numero intero N tale che, per $n > N$, si ha: $\cos\xi_n \in (m_-, 1)$, dove $m_- = 1 - u/4$ è il punto medio del segmento di estremi $\pi(1), 1$. Allora, per $n > N$, si ha: $\texttt{COS}(\xi_n) = \mathrm{rd}(\cos\xi_n) = 1$.

E45 ★ Siano $f : \mathbb{R} \to \mathbb{R}$ una funzione con derivata prima continua tale che per ogni $x \in \mathbb{R}$ si abbia $|\, f'(x)\,| > L > 0$ ed $\alpha \neq 0$ l'unico zero di f. Mostrare che per il numero di condizionamento del calcolo di f in x si ha:

$$\lim_{x \to \alpha} c_f(x) = +\infty$$

E46 Si consideri l'Esempio 0.4.12. Tenuto conto che in *Scilab* si ha: `%pi` $= \mathrm{rd}(\pi) < \pi$ e $\phi(\pi) =$ `SEN(%pi)` > 0:

(1) Mostrare che per ogni $x \in (\texttt{\%pi}, \pi)$ si ha $\mathrm{rd}(x) =$ `%pi` e quindi $\phi(x) =$ `SEN(%pi)`.

(2) Mostrare che, posto per ogni $x \in (\texttt{\%pi}, \pi)$:

$$e(x) = \frac{\phi(x) - f(x)}{f(x)}$$

si ha:

$$\lim_{x \to \pi -} e(x) = +\infty$$

ovvero: per x vicino a π l'algoritmo ϕ *non è accurato* quando utilizzato per approssimare sen x.

E47 Si consideri l'Esempio 0.4.16. Tenuto conto che per ogni $x \in \mathbb{R}$ si ha: $\cos x = \cos(\frac{1}{2}x + \frac{1}{2}x)$, dimostrare che:

$$f(x) = 2\left(\mathrm{sen}(x/2)\right)^2$$

Siano poi M un insieme di numeri in virgola mobile e precisione finita, `SQR` la funzione predefinita corrispondente alla funzione quadrato e $\psi : \mathbb{R} \to M$ l'algoritmo definito da:

$$\psi(x) = 2 \otimes \texttt{SQR}\Big(\texttt{SEN}\big(\mathrm{rd}(x) \oslash 2\big)\Big)$$

Dimostrare che per ogni $x \in \mathbb{R}$, l'algoritmo ψ è stabile quando utilizzato per approssimare il valore di f in x.

E48 Siano $f : \mathbb{R} \to \mathbb{R}$ una funzione regolare, $x \in \mathbb{R}$ e $c_f(x)$ il numero di condizionamento del calcolo di f in x. Siano poi M un insieme di numeri in virgola mobile ed esponente non limitato e rd la funzione arrotondamento in M. Per approssimare $f(x)$ si utilizza l'algoritmo "ideale": $\phi(x) = \mathrm{rd}\Big(f\big(\mathrm{rd}(x)\big)\Big)$.

Mostrare che per l'errore relativo e_t commesso approssimando $f(x)$ con $\phi(x)$ si ha:

$$|e_t| \leqslant \cdots \approx u + c_f(x)\,(u + u^2)$$

Mostrare poi che l'errore relativo commesso approssimando $u + c_f(x)\,(u + u^2)$ con $u + c_f(x)u$ è minore di u.

0.5 Appendice

Questa Appendice contiene le soluzioni degli esercizi contrassegnati da S.

- Esercizio *E3*

  Sia g la frazione di x in base β, ovvero:

  $$x = \beta^b g \quad , \quad g \in [1/\beta, 1)$$

  Se b è pari, $b = 2n$, allora:

  $$\sqrt{x} = \sqrt{\beta^{2n} g} = \sqrt{\beta^{2n}}\sqrt{g} = \beta^n \sqrt{g}$$

  e per calcolare $\sqrt{x}$ è sufficiente saper calcolare $\sqrt{y}$ per ogni $y \in [1/\beta, 1)$.

Se b è dispari, $b = 2n + 1$, allora:

$$\sqrt{x} = \sqrt{\beta^{2n+1} g} = \sqrt{\beta^{2n}}\sqrt{(\beta g)} = \beta^n \sqrt{\beta g}$$

e per calcolare $\sqrt{x}$ è sufficiente saper calcolare $\sqrt{y}$ per ogni $y \in [1, \beta)$.
Ad esempio: siano $\beta = 2$ e $x = 6$. Allora: $6 = 2^3(6/8) = 2^3(3/4)$ e quindi:

$$\sqrt{6} = \sqrt{2^3(3/4)} = \sqrt{2^2(3/2)} = \sqrt{2^2}\sqrt{3/2} = 2\sqrt{3/2}$$

Si osservi che $3/2 \in [1/2, 2)$.

- Esercizio *E21*

  Poiché l'esponente di ξ in base β è b, allora:

  (A) $\beta^{b-1} \leqslant \xi < \beta^b$
  (B) $\sigma(\xi) \leqslant \beta^b$

  Per le ipotesi su x si ha allora:

  $$\beta^{b-1} \leqslant \xi \leqslant x < \sigma(\xi) \leqslant \beta^b \qquad (*)$$

  ovvero:

  $$\beta^{b-1} \leqslant x < \beta^b$$

  e l'esponente di x in base β coincide con quello, b, di ξ.
  Per quanto appena mostrato si può scrivere:

  $$x = \beta^b \, 0.d_1 \cdots d_m d_{m+1} \cdots$$

  (si osservi che le cifre $d_{m+1}, d_{m+2}, \ldots$ *non* sono tutte uguali a 1) e quindi:

  $$x - \xi = \beta^b(0.d_1 \cdots d_m - 0.c_1 \cdots c_m) + \beta^{b-m} \, 0.d_{m+1} \cdots = \beta^{b-m}(d_1 \cdots d_m - c_1 \cdots c_m + 0.d_{m+1} \cdots)$$

  Ma, dalle disuguaglianze $(*)$:

  $$0 \leqslant x - \xi < \sigma(\xi) - \xi = \beta^{b-m}$$

  Allora:

  (C) $0 \leqslant \beta^{b-m}(d_1 \cdots d_m - c_1 \cdots c_m + 0.d_{m+1} \cdots)$
  (D) $\beta^{b-m}(d_1 \cdots d_m - c_1 \cdots c_m + 0.d_{m+1} \cdots) < \beta^{b-m}$.

  Dall'asserto (C) si ottiene:

  $$0 \leqslant d_1 \cdots d_m - c_1 \cdots c_m + 0.d_{m+1} \cdots$$

  e dall'asserto (D):

  $$d_1 \cdots d_m - c_1 \cdots c_m + 0.d_{m+1} \cdots < 1$$

  Poiché $0.d_{m+1} \cdots < 1$, e $d_1 \cdots d_m - c_1 \cdots c_m$ è un numero intero:

  $$d_1 \cdots d_m = c_1 \cdots c_m$$

- Esercizio *E41*

  La procedura eseguita da *Scilab* è quella che si ottiene applicando il procedimento di trasformazione alla procedura data assumendo $M = F_d(2, 53, -1021, 1024)$. La procedura trasformata è (si osservi che $1, -53$ e 1024 sono in M):

  $$\texttt{x =} \ \mathrm{rd}(2^{1024}) \otimes \left(1 \ominus \mathrm{rd}(2^{-53})\right)$$

  La sequenza di operazioni eseguite da *Scilab* è allora:

  (1) $\xi_1 = \mathrm{rd}(2^{1024})$

(2) $\xi_2 = \mathrm{rd}(2^{-53}) = 2^{-53}$

(3) $\xi_3 = \mathrm{rd}(1 - \xi_2) = 1 - 2^{-53}$

(4) $\mathtt{x} = \mathrm{rd}(\xi_1\, \xi_3)$

Poiché:

$$2^{1024} = 2^{1025}\, 0.1$$

si ha, come richiesto dallo *IEEE Standard for Floating-Point Arithmetic* (IEEE Std 754-2019), paragrafo 7.4 - Overflow, $\xi_1 =$ `Inf`. Come richiesto dal documento citato, paragrafo 6.1 - Infinity Arithmetic, si ha: $\mathrm{rd}(\xi_1\, \xi_3) =$ `Inf`, dunque il valore di `x` ottenuto è `Inf`.

1 Zeri di funzione

Sia $f : [a, b] \to \mathbb{R}$ una funzione *continua* ed $\alpha \in [a, b]$ uno zero di f. In questo Capitolo affrontiamo il problema di *determinare un'approssimazione accurata di* α.

Una *condizione sufficiente* per l'esistenza di *almeno uno* zero di f è data dal seguente:

1.0.1 Teorema (di esistenza degli zeri)
Sia $f : [a, b] \to \mathbb{R}$ una funzione *continua*. Se $f(a)\, f(b) < 0$ allora esiste $\alpha \in (a, b)$ zero di f.

– *Esempio*

Sia $f : \mathbb{R} \to \mathbb{R}$ definita da:

$$f(x) = \begin{cases} 1/x + x - 3 & \text{per } x \neq 0 \\ 1 & \text{per } x = 0 \end{cases}$$

La funzione è continua su $[1, 3]$ e $f(1) = -1 < 0$, $f(3) = \frac{1}{3} > 0$: il Teorema di esistenza degli zeri assicura l'esistenza di *almeno uno* zero di f in $(1, 3)$.

La funzione è continua su $[\frac{1}{3}, 3]$ ma $f(\frac{1}{3}) = \frac{1}{3} > 0$ e $f(3) > 0$: il Teorema di esistenza degli zeri *non è applicabile* e quindi non fornisce informazioni sull'esistenza di zeri di f in $(\frac{1}{3}, 3)$. Ovviamente, per quanto detto prima, f ha almeno uno zero in $(\frac{1}{3}, 3)$.

Si ha infine: $f(-\frac{1}{3}) < 0$ e $f(\frac{1}{3}) > 0$, ma la funzione *non* è continua su $[-\frac{1}{3}, \frac{1}{3}]$: il Teorema di esistenza degli zeri *non è applicabile* e quindi non fornisce informazioni sull'esistenza di zeri di f in $(-\frac{1}{3}, \frac{1}{3})$.

1.1 Metodo di bisezione

Sia $f : [a, b] \to \mathbb{R}$ una funzione *continua* tale che $f(a)\, f(b) < 0$. Il Teorema precedente assicura l'*esistenza* di almeno uno zero di f in (a, b). Il primo metodo che consideriamo per approssimare uno di questi zeri è il *metodo di bisezione*, basato sul Teorema appena enunciato. Si tratta di un metodo *iterativo*, ovvero che determina l'oggetto cercato costruendo *una successione*. La procedura seguente, descritta in un linguaggio che utilizza il tipo *numero reale*, realizza il metodo:

$z = \mathit{Bisezione}\,(f, a, b)$

// $f : [a, b] \to \mathbb{R}$ *continua* tale che $f(a) f(b) < 0$.

// k è il contatore delle iterazioni eseguite.

$k = 0$;

$a_0 = a; b_0 = b; x_0 = (a_0 + b_0)/2$;

`ripeti`:

> `se` $f(x_k) = 0$ `allora` esci dal ciclo;
> `se` $f(x_k) f(b_k) < 0$ `allora` $a_{k+1} = x_k; b_{k+1} = b_k$;
> `se` $f(a_k) f(x_k) < 0$ `allora` $a_{k+1} = a_k; b_{k+1} = x_k$;
> $x_{k+1} = (a_{k+1} + b_{k+1})/2$;
> $k = k + 1$;

$z = x_k$

La procedura opera in questo modo: Se per qualche k si ha $f(x_k) = 0$, allora essa *termina* e restituisce uno zero di f. Se, invece, per ogni k si ha $f(x_k) \neq 0$, allora essa *non termina* e genera *due* successioni: la successione di intervalli $I_k = [a_k, b_k]$ e la successione di numeri reali x_k, *punti medi* degli intervalli I_k.

Per ciascun k si ha:

∗ I_k contiene, per costruzione, almeno uno zero di f

$*$ $I_{k+1} \subset I_k$

$*$ $\operatorname{mis} I_k = \dfrac{\operatorname{mis} I_0}{2^k}$

Dalla terza proprietà segue:

$$\lim_{k\to\infty} \operatorname{mis} I_k = 0$$

dunque la successione di intervalli "individua con incertezza tendente a zero" uno zero di f.

Si ha inoltre:

1.1.1 Osservazione (convergenza delle successioni)

Le successioni a_k, b_k ed x_k sono convergenti ad uno stesso limite α tale che $f(\alpha) = 0$.

Infatti: Per costruzione la successione a_k risulta *monotona non decrescente* e *superiormente limitata* (da b), dunque convergente: $\lim a_k = A$. Analogamente: la successione b_k risulta *monotona non crescente* e *inferiormente limitata* (da a), dunque convergente: $\lim b_k = B$. La successione $\operatorname{mis} I_k = b_k - a_k$ è allora differenza di successioni convergenti e quindi:

$$0 = \lim_{k\to\infty} \operatorname{mis} I_k = \lim_{k\to\infty} (b_k - a_k) = B - A \qquad \text{dunque} \qquad A = B$$

Posto $\alpha = A$, poiché $a_k < x_k < b_k$ si ha anche $\lim x_k = \alpha$.

Infine, sia ad esempio: $f(a) < 0$ e $f(b) > 0$. Per ogni k si ha, per costruzione: $f(a_k) < 0$ e $f(b_k) > 0$. Tenuto conto della continuità di f e della convergenza delle successioni a_k e b_k:

$$\lim_{k\to\infty} f(a_k) = f(\alpha) \leqslant 0 \qquad \text{e} \qquad \lim_{k\to\infty} f(b_k) = f(\alpha) \geqslant 0$$

e quindi $f(\alpha) = 0$.

1.1.2 Esercizio

Sia:

$$f(x) = \frac{1}{x - \sqrt{2}}$$

Discutere l'assegnamento $z = \mathit{Bisezione}\,(f, 0, 2)$.

La funzione f è definita e continua sull'unione $\Omega = [0, \sqrt{2}) \cup (\sqrt{2}, 2]$ e *non*, come richiesto dal commento della procedura *Bisezione*, su $[0, 2]$. Però per ogni k si ha: a_k, b_k e perciò x_k sono numeri razionali in $[0, 2]$ dunque in Ω. Allora la procedura *non termina* (si ha sempre $f(x_k) \neq 0$, infatti f *non ha zeri* in Ω). Le successioni a_k, b_k e x_k che la procedura costruisce sono ancora convergenti ad un limite comune α (come mostra la prima parte della dimostrazione dell'Osservazione precedente). Inoltre, se fosse $\alpha \neq \sqrt{2}$ la funzione f sarebbe continua in α e quindi si avrebbe $f(\alpha) = 0$. Ma, come già detto, f non ha zeri in Ω. *La procedura individua il punto in cui f "cambia segno".*

L'Osservazione 1.1.1 mostra che la procedura *Bisezione* (come tutte le procedure che realizzano metodi iterativi) determina uno zero di f come *limite* di una successione. Come abbiamo detto nella parte introduttiva del Capitolo 0, le procedure descritte saranno eseguite da un calcolatore. Una procedura che costruisce *tutta* una successione *non è accettabile* in questo contesto perché il calcolatore impiegherebbe un *tempo infinito* per eseguirla (il calcolatore impiega un tempo *non infinitesimo* per calcolare *ciascun elemento* della successione). Per rendere *finito* in ogni caso il tempo di esecuzione, è necessario *interrompere* la costruzione della successione. Così facendo la procedura determinerà, con l'ultimo elemento calcolato della successione, solo *un'approssimazione* di uno zero di f. La costruzione della successione deve essere interrotta *quando l'ultimo elemento costruito approssima lo zero di f con sufficiente accuratezza.* A questo scopo si introduce nella procedura un *criterio d'arresto.*

1.1.3 Esempio (criterio d'arresto di tipo assoluto)

Assegnato un numero reale positivo δ, un comune esempio di criterio d'arresto è:

$$\textit{se } \operatorname{mis} I_k < \delta \textit{ allora} \text{ arresta la costruzione}$$

ovvero: "arresta la costruzione se l'ultimo intervallo calcolato è sufficientemente piccolo." Il criterio d'arresto è introdotto nella procedura *Bisezione* modificandola come segue:

$z = \mathit{Bisezione}\,(f, a, b, \delta)$

// $f : [a, b] \to \mathbb{R}$ continua tale che $f(a)f(b) < 0$, δ numero reale positivo.

// k è il contatore delle iterazioni eseguite.

$k = 0$;

$a_0 = a;\ b_0 = b;\ x_0 = (a_0 + b_0)/2$;

`ripeti`:

`se` $\big(f(x_k) = 0$ `oppure` $b_k - a_k < \delta\big)$ `allora` esci dal ciclo;
`se` $f(x_k)f(b_k) < 0$ `allora` $a_{k+1} = x_k; b_{k+1} = b_k$;
`se` $f(a_k)f(x_k) < 0$ `allora` $a_{k+1} = a_k; b_{k+1} = x_k$;
$x_{k+1} = (a_{k+1} + b_{k+1})/2$;
$k = k + 1$;

$z = x_k$

Un criterio d'arresto è in generale definito da *un'opportuna condizione* sugli elementi della successione calcolati dalla procedura. La costruzione della successione verrà interrotta *appena* e *solo se* la condizione risulterà soddisfatta.

1.1.4 Osservazione (proprietà di un criterio d'arresto)
La condizione che definisce il criterio d'arresto deve avere le proprietà seguenti:

- Essere *calcolabile*: ad ogni iterazione la procedura *deve* essere in grado di verificare se la condizione è soddisfatta.
- Essere *efficace*: in ogni caso la condizione *deve* essere soddisfatta dopo un numero *finito* di iterazioni.
- Quando la condizione è soddisfatta la procedura *deve* restituire un elemento che *approssima l'oggetto cercato con l'accuratezza richiesta* dall'utilizzatore.

Il criterio d'arresto proposto nell'Esempio 1.1.3 *soddisfa* le tre proprietà: *è calcolabile*, infatti a ciascuna iterazione la procedura conosce a_k e b_k, può calcolare $\operatorname{mis} I_k = b_k - a_k$ e verificare se è minore del valore δ fornito dall'utilizzatore; *è efficace*, infatti si ha $\lim \operatorname{mis} I_k = 0$ e per ogni $\delta > 0$ la disuguaglianza $\operatorname{mis} I_k < \delta$ è *certamente* soddisfatta dopo un numero *finito* di iterazioni. Infine, quando il criterio di arresto è soddisfatto la procedura restituisce x_k, punto medio dell'ultimo intervallo calcolato I_k, e tale intervallo, per costruzione, contiene almeno uno zero α di f. Si ha allora:

$$|\, x_k - \alpha \,| \leqslant \frac{\operatorname{mis} I_k}{2} < \tfrac{1}{2}\,\delta < \delta$$

ovvero la procedura restituisce un valore che approssima uno zero di f con *errore assoluto* minore di δ. Il criterio verifica dunque la terza proprietà a patto che l'utilizzatore misuri l'accuratezza con l'errore assoluto. Per questo motivo il criterio d'arresto proposto è classificato *di tipo assoluto.*

- *Esempi*

  Un criterio d'arresto calcolabile ed efficace ma che *non necessariamente* restituisce un valore che approssima uno zero di f con l'accuratezza richiesta è il seguente. Sia δ un numero reale positivo:

  $$\textit{se } |\, f(x_k) \,| < \delta \textit{ allora} \text{ arresta la costruzione}$$

  Supponiamo che f sia una funzione con derivata prima continua e non nulla in $[a, b]$, α lo zero di f in $[a, b]$ e x_k tale che $|\, f(x_k) \,| = \frac{1}{2}\,\delta$. Per il Teorema di Lagrange esiste θ tra x_k ed α tale che:

  $$|\, f(x_k) \,| = |\, f(x_k) - f(\alpha) \,| = |\, f'(\theta) \,|\, |x_k - \alpha \,|$$

dunque:

$$|\, x_k - \alpha \,| = \frac{|\, f(x_k) \,|}{|\, f'(\theta) \,|} = \frac{\delta}{2\,|\, f'(\theta) \,|}$$

Il valore x_k approssima α con l'accuratezza richiesta *se e solo se* $|\, f'(\theta) \,| > \frac{1}{2}$.

Un criterio d'arresto che è efficace e restituisce un valore che approssima uno zero di f con l'accuratezza richiesta *ma non è calcolabile*, è il seguente. Siano α uno zero di f in $[a, b]$ e δ un numero reale positivo:

$$\textit{se } |\, x_k - \alpha \,| < \delta \textit{ allora} \text{ arresta la costruzione}$$

La procedura *non conosce* α e quindi non può verificare se la condizione è soddisfatta.

1.1.5 Esercizio

Sia $f : [a, b] \to \mathbb{R}$ una funzione continua tale che $f(a)f(b) < 0$ e δ un numero reale positivo. La procedura:

$$z = \textit{Bisezione}\,(f, a, b, \delta)$$

restituisce un'approssimazione di uno zero di f in $[a, b]$ dopo aver eseguito k iterazioni. Si vuole determinare k.

Se la procedura termina trovando uno zero di f, il numero di iterazioni eseguite è in generale imprevedibile. Se invece la procedura termina perchè l'ultimo intervallo costruito ha misura minore di δ allora:

$$\operatorname{mis} I_k = \frac{\operatorname{mis} I_0}{2^k} < \delta \quad \Rightarrow \quad k > \log_2 \operatorname{mis} I_0 - \log_2 \delta$$

La procedura si arresta dopo aver eseguito:

$$k = \lfloor \log_2 \operatorname{mis} I_0 - \log_2 \delta \rfloor + 1$$

iterazioni.[19]

Ad esempio, se $\operatorname{mis} I_0 = 2$ e $\delta = 10^{-10}$ si ha:

$$k = \lfloor 1 + 10 \log_2 10 \rfloor + 1 = 35$$

Inoltre, fissato $\operatorname{mis} I_0$, per $\delta \to 0$ il valore di k tende a infinito come $|\log_2 \delta|$.

In generale: *tanto più accurata* l'utilizzatore vuole che sia l'approssimazione richiesta, *tante più iterazioni* deve eseguire la procedura (cioè: *tanto più impegnativo* è ottenere l'approssimazione).

1.1.6 Esempio (criterio d'arresto di tipo relativo)

Il criterio d'arresto utilizzato nella procedura $\textit{Bisezione}\,(f, a, b, \delta)$ è stato classificato di *tipo assoluto* perchè l'ultimo elemento calcolato della successione approssima uno zero di f con l'accuratezza richiesta *a patto* che l'utilizzatore misuri l'accuratezza con l'*errore assoluto*. Un criterio di *tipo relativo*, adatto quindi se l'utilizzatore misura l'accuratezza con l'*errore relativo*, è il seguente. Dato un numero reale positivo ϵ, che misurerà l'accuratezza richiesta dall'utilizzatore, e posto $m_k = \min\{\,|\, a_k \,|, |\, b_k \,|\,\}$:

$$\textit{se } \frac{\operatorname{mis} I_k}{m_k} < \epsilon \textit{ allora} \text{ arresta la costruzione}$$

Il criterio, in quanto di tipo relativo, è utilizzabile *solo* quando la procedura approssima uno zero *non nullo* di f e in tal caso si può supporre che sia:

$$0 \notin I_0 = [a, b]$$

e quindi $0 \notin I_k$ (che assicura $m_k \neq 0$) per ogni k. Il criterio d'arresto è introdotto nella procedura *Bisezione* modificandola come segue:

[19] Se t è un numero reale positivo, $\lfloor t \rfloor$ è la *parte intera* di t: il più grande intero minore o uguale di t. Dunque $\lfloor t \rfloor + 1$ è il più piccolo intero maggiore di t.

$z = \mathit{Bisezione}\,(f, a, b, \epsilon)$

// $[a, b] \not\ni 0, f : [a, b] \to \mathbb{R}$ continua e tale che $f(a)f(b) < 0$, ϵ numero reale positivo.

// k è il contatore delle iterazioni eseguite.

$k = 0;$

$a_0 = a;\ b_0 = b;\ x_0 = (a_0 + b_0)/2;\ m_0 = \min\{\,|\,a_0\,|, |\,b_0\,|\,\};$

`ripeti`:

`se` $\big(f(x_k) = 0$ `oppure` $(b_k - a_k)/m_k < \epsilon\big)$ `allora` esci dal ciclo;
`se` $f(x_k)f(b_k) < 0$ `allora` $a_{k+1} = x_k; b_{k+1} = b_k;$
`se` $f(a_k)f(x_k) < 0$ `allora` $a_{k+1} = a_k; b_{k+1} = x_k;$
$x_{k+1} = (a_{k+1} + b_{k+1})/2;$
$m_{k+1} = \min\{\,|\,a_{k+1}\,|, |\,b_{k+1}\,|\,\};$
$k = k + 1;$

$z = x_k$

Il criterio è *calcolabile*: a ciascuna iterazione la procedura conosce a_k e b_k, sa calcolare m_k e verificare se la disuguaglianza è soddisfatta. Il criterio è anche *efficace*, infatti:

$$\operatorname{mis} I_k \to 0 \quad \text{e} \quad m_k \geqslant m_0 > 0 \qquad \Rightarrow \qquad \lim_{k\to\infty} \frac{\operatorname{mis} I_k}{m_k} = 0$$

dunque per ogni $\epsilon > 0$ la disuguaglianza è certamente verificata dopo un numero finito di iterazioni. Infine, quando il criterio di arresto è verificato la procedura restituisce x_k, punto medio dell'ultimo intervallo calcolato I_k. Tale intervallo, per costruzione, contiene almeno uno zero $\alpha \neq 0$ di f e si ha:

$$|\,\alpha\,| > m_k \qquad \text{e quindi} \qquad \frac{|\,x_k - \alpha\,|}{|\,\alpha\,|} < \frac{\operatorname{mis} I_k}{m_k} < \epsilon$$

ovvero la procedura restituisce un valore che approssima uno zero di f con *errore relativo* minore di ϵ.

1.2 Uso del tipo *numero in virgola mobile e precisione finita* nel metodo di bisezione

In questa sezione si discute l'esecuzione in *Scilab* della procedura definita nell'Esempio 1.1.3.

Si assume, per semplicità, $M = F(2, 53)$ — si veda l'Osservazione 0.1.17 — e si indicano, come usuale, con rd e u, rispettivamente, la funzione arrotondamento e la precisione di macchina in M.

1.2.1 Teorema (stabilità dell'algoritmo di bisezione)
Siano: $a < b$ due numeri reali positivi tali che $\mathrm{rd}(a) \neq \mathrm{rd}(b)$, J_0 l'intervallo $[\mathrm{rd}(a), \mathrm{rd}(b)]$, $f : J_0 \to \mathbb{R}$ una funzione continua, $\phi : J_0 \to M$ l'algoritmo utilizzato per approssimare i valori di f e δ un numero reale positivo.

Se l'algoritmo ϕ è *uniformemente accurato* quando utilizzato per approssimare f in $J_0 \cap M$, ovvero:

esiste un numero reale d_ϕ *piccolo* tale che per ogni $\theta \in J_0 \cap M$ si ha: $\quad |\phi(\theta) - f(\theta)| \leqslant d_\phi$

e la procedura $\mathit{Bisezione}\,(f, a, b, \delta)$ eseguita in *Scilab* definisce un elemento $\xi \in M$, *allora* si ha:

$$|\,\xi - \alpha^*| < \delta$$

dove α^* è uno zero di una funzione continua $g : J_0 \to \mathbb{R}$ *vicina* ad f nel senso che:

$$\text{per ogni } x \in J_0 \text{ si ha:} \quad |f(x) - g(x)| \leqslant d_\phi$$

Dimostrazione. La trasformazione descritta nella Sezione A produce:

$z = \mathit{Bisezione}^*(\phi, \mathrm{rd}(a), \mathrm{rd}(b), \mathrm{rd}(\delta))$

// L'algoritmo ϕ sia tale che: $\phi\big(\mathrm{rd}(a)\big)\,\phi\big(\mathrm{rd}(b)\big) < 0$.

// k è il contatore delle iterazioni eseguite.

k = 0;

α_0 = rd(a); β_0 = rd(b); ξ_0 = ($\alpha_0 \oplus \beta_0$) $\oslash$ 2;

ripeti:

se ($\phi(\xi_k)$ = 0 oppure $\beta_k \ominus \alpha_k < \mathrm{rd}(\delta)$) allora esci dal ciclo;
se $\phi(\xi_k) \otimes \phi(\beta_k) < 0$ allora α_{k+1} = ξ_k; β_{k+1} = β_k;
se $\phi(\alpha_k) \otimes \phi(\xi_k) < 0$ allora α_{k+1} = α_k; β_{k+1} = ξ_k;
ξ_{k+1} = ($\alpha_{k+1} \oplus \beta_{k+1}$) $\oslash$ 2;
k = $k \oplus 1$;

$z = \xi_k$

Se la procedura *Bisezione*(f, a, b, δ) eseguita in *Scilab* definisce un elemento $\xi \in M$ allora: esiste un numero intero non negativo k tale che alla k-esima iterazione il criterio d'arresto è verificato, cioè:

$$\phi(\xi_k) = 0 \qquad \text{oppure} \qquad \beta_k \ominus \alpha_k < \mathrm{rd}(\delta)$$

e si ha: $\xi = \xi_k \in M \cap J_0$.

Siano adesso $p : J_0 \to \mathbb{R}$ la funzione continua il cui grafico è la spezzata di vertici i punti di coordinate $\big(\xi, \phi(\xi) - f(\xi)\big)$, $\xi \in M \cap J_0$, ordinati per ascisse crescenti, e $g : J_0 \to \mathbb{R}$ la funzione definita da $g(x) = f(x) + p(x)$. Allora:

(i) Per ogni $\xi \in M \cap J_0$ si ha: $p(\xi) = \phi(\xi) - f(\xi)$ e quindi $g(\xi) = \phi(\xi)$.

(ii) Dall'ipotesi di uniforme accuratezza dell'algoritmo ϕ si ottiene che (si ricordi che p è una funzione continua il cui grafico è una spezzata) per ogni $x \in J_0$ sussiste la limitazione $|p(x)| \leqslant d_\phi$ e quindi:

$$\text{per ogni } x \in J_0 : \quad |g(x) - f(x)| = |p(x)| \leqslant d_\phi$$

ovvero: *la funzione g è vicina ad f.*

(iii) La funzione g è *continua.*

Si osservi infine che:

– Se la procedura si arresta perchè $\phi(\xi_k) = 0$, l'asserto (i) garantisce che si ha anche $g(\xi_k) = 0$. Posto $\alpha^* = \xi_k$ si ha: α^* è *uno zero* della funzione g, vicina ad f per l'asserto (ii), e $|\xi - \alpha^*| = 0 < \delta$.

– Se la procedura si arresta perchè $\beta_k \ominus \alpha_k < \mathrm{rd}(\delta)$, poiché per la monotonia della funzione rd (Osservazione 0.2.5) si ha:

$$\beta_k \ominus \alpha_k < \mathrm{rd}(\delta) \quad \Rightarrow \quad \beta_k - \alpha_k < \delta$$

allora l'ultimo intervallo costruito, $J_k = [\alpha_k, \beta_k]$, ha misura minore di δ. Inoltre, per costruzione, si ha: $\phi(\alpha_k)\,\phi(\beta_k) < 0$ dunque, utilizzando l'asserto (i):

$$g(\alpha_k)\,g(\beta_k) = \phi(\alpha_k)\,\phi(\beta_k) < 0$$

Per il Teorema di esistenza degli zeri e la continuità di g, asserto (iii), esiste allora $\alpha^* \in J_k$ zero di g e, ricordando che per costruzione è $\xi = \xi_k \in J_k$, si ha:

$$|\xi - \alpha^*| \leqslant \mathrm{mis}\, J_k = \beta_k - \alpha_k < \delta$$

Il teorema è dimostrato.

1.2.2 Osservazione (efficacia del criterio d'arresto)

La procedura introdotta nell'Esempio 1.1.3 definisce *in ogni caso* un numero reale perchè, per la convergenza a zero della successione $\operatorname{mis} I_k$, il criterio d'arresto utilizzato è *efficace*. Invece, la successione $\operatorname{mis} J_k$ delle misure degli intervalli generati dalla procedura *Bisezione** è solamente *non crescente*. Infatti: poiché per ogni k si ha:[20]

$$\xi_k = \mathrm{rd}\Big(\frac{\alpha_k + \beta_k}{2}\Big)$$

ovvero ξ_k è l'*arrotondato del punto medio* dell'intervallo $J_k = [\alpha_k, \beta_k]$, allora $\alpha_k \leqslant \xi_k \leqslant \beta_k$, e per $k \geqslant 1$ è $J_k \subset J_{k-1}$. Come l'esempio seguente mostra, la successione $\operatorname{mis} J_k$, salvo casi particolari, *non tende a zero* e la procedura *Bisezione*$\,(f, a, b, \delta)$ eseguita in *Scilab può non definire* un elemento $\xi \in M$ perchè il criterio d'arresto *può risultare non efficace*.

1.2.3 Esempio

Sia $f(x) = x^2 - 2$. Se, posto $\delta = 10^{-16}$ e scelto l'algoritmo ingenuo per approssimare i valori di f, si esegue l'assegnamento:

$$\texttt{z = } \textit{Bisezione}\,(f, 0, 2, \delta)$$

con *Scilab*, la procedura *non termina*: il criterio d'arresto risulta *non efficace*.

Il problema è questo (si ricordi che si è scelto $M = F(2, 53)$): la procedura *Bisezione** cerca di determinare un intervallo *ad estremi elementi di* M*, contenente* $\sqrt{2}$ e *di misura minore di* δ. Ma *il più piccolo* intervallo che ha le prime due delle proprietà richieste è quello che ha per estremi i due elementi di M *adiacenti* a $\sqrt{2}$. Poiché l'esponente di $\sqrt{2}$ in base due è *uno*, la misura di tale intervallo (la distanza tra i due numeri di macchina) è $\beta^{b-m} = 2^{1-53} \approx 2.22\,10^{-16}$, *maggiore* di δ.

La procedura *non può* trovare un intervallo sufficientemente piccolo e quindi *non termina.* In generale, detto b l'esponente dello zero di f che si vuole approssimare, l'utilizzatore *deve* assegnare al parametro δ un valore maggiore di $2^{b-53} = 2^b\,u$.

1.2.4 Osservazione (accuratezza dell'algoritmo di bisezione)

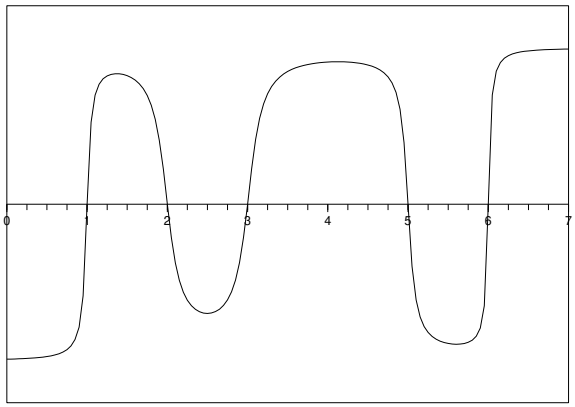
Il Teorema 1.2.1 stabilisce che, sotto opportune ipotesi, la procedura *Bisezione** determina un'elemento $\xi \in M$ approssimazione *accurata* di uno zero di una funzione *vicina* ad f, ovvero che "l'algoritmo di bisezione è *stabile*." Per decidere se ξ è anche un'approssimazione accurata di *uno zero di* f, ovvero se "l'algoritmo di bisezione è *accurato*," occorre studiare il *condizionamento* del calcolo di uno zero di f: *quanto grande può essere rispetto a* d_ϕ la distanza tra uno zero di g e lo zero di f, argomento della Sezione 1.6.

Esercizi

E1 Sia $f(x) = 1/x$, definita per $x \neq 0$. La funzione è *continua* nel suo insieme di definizione e $f(-1) < 0$, $f(1) > 0$. Perché non possiamo concludere, in base al Teorema di esistenza degli zeri, che f ha almeno uno zero in $(-1, 1)$?

E2 Il grafico della funzione $f : [0, 7] \to \mathbb{R}$ è rappresentato nella figura seguente.

[20] Vedere gli Esercizi *E4*-*E6*.



Sia x_k la successione costruita dalla procedura *Bisezione*$(f, 0, 7)$. Determinare $\lim_{k\to\infty} x_k$.

E3 Sia $f(x) = x^3 - 2$.

(1) Determinare analiticamente gli zeri di f.

(2) Determinare *Bisezione*$(f, 0, 2, \frac{1}{2})$.

E4 Siano $M = F(2, m)$ e $\alpha, \omega \in M$. Dimostrare che:

$$\xi = (\alpha \oslash 2) \oplus (\omega \oslash 2) = \mathrm{rd}\Big(\frac{\alpha + \omega}{2}\Big)$$

Dunque ξ è *l'arrotondato del punto medio* dell'intervallo $[\alpha, \omega]$. Dimostrare che, allora:

$$\alpha \leqslant \xi \leqslant \omega$$

E5 ★ Siano β un numero intero pari e rd una funzione arrotondamento in $F(\beta, m)$. Si supponga noto che per ogni $x \in \mathbb{R}$ e $b \in \mathbb{Z}$ si ha:

$$\mathrm{rd}(\beta^b x) = \beta^b \,\mathrm{rd}(x)$$

Siano $M = F(2, m)$ e $\alpha, \omega \in M$. Dimostrare che:

$$\theta = (\alpha \oplus \omega) \oslash 2 = \mathrm{rd}\Big(\frac{\alpha + \omega}{2}\Big)$$

ovvero che θ è *l'arrotondato del punto medio* dell'intervallo $[\alpha, \omega]$.

E6 Siano $M = F(10, 6)$, $\alpha = 0.742531$ e $\omega = 0.742533$. Calcolare:

$$\gamma = (\alpha \oplus \omega) \oslash 2$$

e constatare che $\gamma < \alpha$.

E7 ★ Siano β un numero intero pari, x un numero reale positivo e b un numero intero. Dimostrare che, detta rd la funzione arrotondamento in $F(\beta, m)$, si ha:

$$\mathrm{rd}(\beta^b x) = \beta^b \,\mathrm{rd}(x) \qquad (*)$$

Soluzione.

Se $x \in F(\beta, m)$ anche $\beta^b x \in F(\beta, m)$ e l'uguaglianza $(*)$ è verificata: $\mathrm{rd}(\beta^b x) = \beta^b x = \beta^b \mathrm{rd}(x)$.

Se $x \notin F(\beta, m)$, siano ξ e $\sigma(\xi)$ gli elementi di $F(\beta, m)$ *adiacenti* ad x. Detti n l'esponente e γ la frazione di ξ si ha:

(a) $\beta^b\xi < \beta^b x < \beta^b\sigma(\xi) = \beta^b\sigma(\beta^n\gamma) = \beta^{b+n}(\gamma + \beta^{-m}) = \sigma(\beta^b\xi)$

(b) $|\,x-\xi\,| \gtreqless |\,x-\sigma(\xi)\,|$ se e solo se $|\,\beta^b x-\beta^b\xi\,| \gtreqless |\,\beta^b x-\sigma(\beta^b\xi)\,|\,\big(=|\,\beta^b x-\beta^b\sigma(\xi)\,|\,\big)$

(c) Per ogni $\theta \in F(\beta, m)$ la frazione di $\beta^b\theta$ è uguale a quella di θ.

L'asserto (a) significa che $\beta^b\xi$ e $\sigma(\beta^b\xi)$ sono gli elementi di $F(\beta, m)$ *adiacenti* a $\beta^b x$; l'asserto (b) dimostra l'uguaglianza $(*)$ nel caso di elementi adiacenti *non* equidistanti e l'asserto (c) nel caso di elementi adiacenti equidistanti.

E8 Siano $f : \mathbb{R} \to \mathbb{R}$ una funzione, $x \in \mathbb{R}$ tale che $f(x) \neq 0$ e ϕ l'algoritmo utilizzato per approssimare il valore di f in x. Sia infine e l'errore relativo commesso approssimando $f(x)$ con $\phi(x)$. Dimostrare che $f(x)$ e $\phi(x)$ hanno lo *stesso segno* se e solo se $e > -1$. In particolare: se $|e| < 1$ allora $f(x)$ e $\phi(x)$ hanno lo stesso segno.

E9 ♠ Sia `Bisezione` la procedura *Scilab* realizzata nella prima parte dell'Esercitazione 3 e f la funzione definita da $f(x) = \text{sen } x$.

(1) Dopo aver definito la funzione di intestazione:

```
function y = S(x)
```

che realizza f ed assegnato alla variabile `u` il valore della precizione di macchina, constatare che dopo l'assegnamento:

```
[z,v] = Bisezione(S,2,4,5*u)
```

il valore di `z` è $\text{rd}(\pi)$.

(2) Spiegare perché l'esecuzione dell'assegnamento precedente *termina* mentre quella dell'assegnamento:

```
[z,v] = Bisezione(S,2,4,4*u)
```

non termina.

E10 Si consideri la procedura *Bisezione* descritta nell'Esempio 1.1.6. Assegnata una funzione continua $f : [a, b] \to \mathbb{R}$ tale che $0 \notin [a, b]$ e $f(a)f(b) < 0$, l'assegnamento:

$$z = \mathit{Bisezione}\,(f, a, b, \epsilon)$$

restituisce un'approssimazione di uno zero di f in $[a, b]$, con $f(z) \neq 0$, dopo aver eseguito k iterazioni. Determinare una limitazione superiore per k in termini di a, b ed ϵ.

1.3 Metodi ad un punto

Sia $h : [a, b] \to \mathbb{R}$ una funzione *continua*. La procedura seguente, descritta in un linguaggio che utilizza il tipo *numero reale*, realizza il *metodo iterativo ad un punto definito da h*:

$z = \mathit{MetodoUnPunto}\,(h, \gamma)$

// $h : [a, b] \to \mathbb{R}$ funzione *continua*, $\gamma \in [a, b]$.

$x_0 = \gamma$;

$k = 0$;

`ripeti`:

`se` $x_k \notin [a, b]$ `allora` esci dal ciclo;

$x_{k+1} = h(x_k)$;

$k = k + 1$;

$z = x_k$

La procedura opera in questo modo: Se per qualche k si ha $x_k \notin [a,b]$ allora essa *termina*. Se, invece, per ogni k si ha $x_k \in [a,b]$ allora essa *non termina* e costruisce una successione di numeri reali $x_k \in [a,b]$. Inoltre:

1.3.1 Osservazione

Se la procedura MetodoUnPunto(h,γ) genera una successione x_k convergente, allora il limite della successione è un punto unito di h in $[a,b]$.[21]

Dimostrazione: Sia α il limite della successione x_k. La successione $h(x_0), h(x_1), \ldots$, per la continuità di h, è convergente e $\lim h(x_k) = h(\alpha)$. Ma le successioni $x_1, x_2, \ldots$ e $h(x_0), h(x_1), \ldots$ sono *identiche* e quindi hanno lo stesso limite, ovvero $\alpha = h(\alpha)$.

Dunque: *il metodo ad un punto definito da h determina i punti uniti di h generando successioni ad essi convergenti.*

Sia f la funzione della quale si vuole approssimare uno zero. *Se* una funzione continua h è tale che:

$$\text{insieme degli } \textit{zeri} \text{ di } f \quad = \quad \text{insieme dei } \textit{punti uniti} \text{ di } h$$

allora è *ragionevole* tentare di utilizzare il metodo ad un punto definito da h per approssimare gli zeri di f.

Assegnata f esistono *infinite* funzioni h che hanno la proprietà richiesta.

1.3.2 Esempio

Sia $f : [a,b] \to \mathbb{R}$ una funzione *continua*.

- La funzione $h : [a,b] \to \mathbb{R}$ definita da: $h(x) = x - f(x)$ è continua (perchè lo è f) e si ha:

$$f(\alpha) = 0 \Rightarrow h(\alpha) = \alpha - f(\alpha) = \alpha$$

  e:

$$\alpha = h(\alpha) \Rightarrow \alpha = \alpha - f(\alpha) \Rightarrow f(\alpha) = 0$$

- Sia $g : [a,b] \to \mathbb{R}$ una funzione continua tale che:

$$\text{per ogni } x \in [a,b] \text{ si ha } g(x) \neq 0$$

  Allora la funzione $h : [a,b] \to \mathbb{R}$ definita da: $h(x) = x - g(x)f(x)$ è continua e $\alpha \in [a,b]$ è punto unito di h se e solo se è zero di f. (*Esercizio*: dimostrare l'asserto.)

Una volta scelta la funzione h, ci si domanda se esista, ed eventualmente come individuare, qualche valore di γ a partire dal quale la successione generata dal metodo iterativo definito da h risulti *convergente*. Si osservi che se α è punto unito di h allora la successione generata a partire da $\gamma = \alpha$ è *costante* (per ogni k si ha $x_k = \alpha$) e quindi convergente. Ma la scelta $\gamma = \alpha$ *non è praticamente ragionevole*, dunque dalla ricerca dei valori di γ da cui partire si devono *escludere* i punti uniti di h.

Il Teorema seguente fornisce *condizioni sufficienti* affinché la procedura *MetodoUnPunto*(h,γ) generi una successione convergente.

1.3.3 Teorema (di convergenza)

Siano $[a,b]$ un intervallo non degenere, $h : [a,b] \to \mathbb{R}$ una funzione *con derivata prima continua* e γ un elemento di $[a,b]$ tali che:

(1) esiste α punto unito di h in $[a,b]$;

(2) esiste $L \in [0,1)$ tale che per ogni $x \in [a,b]$ si ha: $|\, h'(x)\,| \leqslant L$;

(3) la procedura *MetodoUnPunto*(h,γ) genera una successione x_k in $[a,b]$.

[21] Si ricordi che un *punto unito* di una funzione $h : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}$, è un numero reale $\alpha \in \Omega$ che verifica la relazione: $\alpha = h(\alpha)$.

Allora: (i) α è l'*unico* punto unito di h in $[a,b]$ e (ii) la successione x_k è *convergente* ad α.

Dimostrazione. (i) Per assurdo: se $\beta \neq \alpha$ è un altro punto unito di h in $[a,b]$ si ha:

$$\beta - \alpha = h(\beta) - h(\alpha)$$

Per il Teorema di Lagrange esiste un numero reale θ compreso tra α e β, e quindi $\theta \in [a,b]$, tale che:

$$h(\beta) - h(\alpha) = h'(\theta)(\beta - \alpha)$$

ovvero:

$$\beta - \alpha = h'(\theta)(\beta - \alpha)$$

Essendo $\beta - \alpha \neq 0$, l'uguaglianza precedente sussiste *se e solo se* $h'(\theta) = 1$. Questo contraddice l'ipotesi (2).

(ii) Dimostriamo che la successione $x_k - \alpha$ converge a zero. Sia k un intero positivo. Allora:

$$x_k - \alpha = h(x_{k-1}) - h(\alpha)$$

Per il Teorema di Lagrange esiste un numero reale θ_{k-1} compreso tra x_{k-1} e α, e quindi $\theta_{k-1} \in [a,b]$, tale che:

$$h(x_{k-1}) - h(\alpha) = h'(\theta_{k-1})(x_{k-1} - \alpha)$$

ovvero:

$$x_k - \alpha = h'(\theta_{k-1})(x_{k-1} - \alpha)$$

Allora, utilizzando l'ipotesi (2):

$$|\, x_k - \alpha \,| = |\, h'(\theta_{k-1}) \,|\, |\, x_{k-1} - \alpha \,| \leqslant L \,|\, x_{k-1} - \alpha \,|$$

Ripetendo il ragionamento a partire da $x_{k-1} - \alpha$ si ottiene:

$$|\, x_{k-1} - \alpha \,| \leqslant L \,|\, x_{k-2} - \alpha \,|$$

e quindi:

$$|\, x_k - \alpha \,| \leqslant L^2 \,|\, x_{k-2} - \alpha \,|$$

Iterando all'indietro si ha infine:

$$0 \leqslant |\, x_k - \alpha \,| \leqslant L^k \,|\, x_0 - \alpha \,|$$

Poiché $L < 1$ la successione $L^k \,|\, x_0 - \alpha \,|$, e quindi $|\, x_k - \alpha \,|$, tende a zero. Il Teorema è dimostrato.

Il ruolo del numero reale γ (il *valore iniziale* della successione) nel Teorema precedente è di garantire il sussistere dell'ipotesi (3), ovvero che il metodo definito da h generi una successione in $[a,b]$. L'Osservazione che segue fornisce, sotto opportune ipotesi, *un valore* che soddisfa la richiesta.

1.3.4 Osservazione (Criterio di scelta del valore iniziale per metodi ad un punto)

Siano $[a,b]$ un intervallo ed $h : [a,b] \to \mathbb{R}$ una funzione con derivata prima continua che verificano le ipotesi (1) e (2) del Teorema di convergenza. Allora, detto α il punto unito di h in $[a,b]$, l'elemento:

$$\gamma = \text{l'estremo di } [a,b] \text{ più vicino ad } \alpha$$

rende soddisfatta l'ipotesi (3) del Teorema di convergenza.

Inoltre, se per ogni $x \in [a,b]$ si ha anche $h'(x) \geqslant 0$, allora: *qualunque* $\gamma \in [a,b]$ rende soddisfatta l'ipotesi (3) del Teorema di convergenza.

Dimostrazione. Sia $d = |\, \gamma - \alpha \,|$ e I l'intorno chiuso di centro α e raggio d. Per come definito γ si ha $I \subset [a,b]$. Sia ora $x \in I$. Allora: $|\, h(x) - \alpha \,| = |\, h(x) - h(\alpha) \,|$ e, utilizzando il Teorema di Lagrange: esiste un numero reale θ compreso tra x e α, e quindi $\theta \in [a,b]$, tale che: $h(x) - h(\alpha) = h'(\theta)(x - \alpha)$, dunque $|\, h(x) - \alpha \,| = |\, h'(\theta) \,|\, |\, x - \alpha \,|$. Utilizzando l'ipotesi (2): $|\, h(x) - \alpha \,| \leqslant L \,|\, x - \alpha \,| < |\, x - \alpha \,| \leqslant d$, ovvero $h(x) \in I$. Ne segue che se $x_0 \in I \subset [a,b]$ allora per ogni numero intero positivo k si ha: $x_k = h(x_{k-1}) \in I \subset [a,b]$.

Per dimostrare l'asserto finale si osservi che, per ogni $x \in [a,\alpha)$, si ha (utilizzando, come sopra, il Teorema di Lagrange): $\alpha - h(x) = h(\alpha) - h(x) = h'(\theta)(\alpha - x)$. Ma, per ipotesi, $0 \leqslant h'(\theta) < 1$ e $\alpha - x > 0$, quindi: $0 \leqslant \alpha - h(x) < \alpha - x$, ovvero $h(x) \in (x,\alpha] \subset [a,b]$. Analogamente si dimostra che: per ogni $x \in (\alpha,b]$ si ha $h(x) \in [\alpha,x) \subset [a,b]$. Dunque, come sopra: se $x_0 \in [a,b]$ allora per ogni numero intero positivo k si ha: $x_k = h(x_{k-1}) \in [a,b]$.

L'esempio che segue mostra l'uso del Teorema di convergenza e del Criterio di scelta del valore iniziale.

1.3.5 Esempio

Sia f la funzione definita, per ogni $x > 0$, da: $f(x) = x + \log x$. Poiché per ogni $x > 0$ si ha $f'(x) = 1 + 1/x > 0$, la funzione f ha *al più* uno zero. L'*esistenza* di uno zero si ottiene osservando che:

$$\lim_{x\to 0+} f(x) = -\infty \qquad \text{e} \qquad \lim_{x\to+\infty} f(x) = +\infty$$

Infine, essendo $f(1) = 1 > 0$, l'intervallo $(0, 1)$ *separa* lo zero di f.[22]

Sia α lo zero di f. Per approssimare α si considerano i metodi ad un punto definiti dalle funzioni (continue):

$$h_1(x) = -\log x \qquad , \qquad h_2(x) = e^{-x} \qquad , \qquad h_3(x) = \frac{x + e^{-x}}{2}$$

Si verifica facilmente (esercizio!) che i punti uniti di ciascuna di esse sono *tutti e soli* gli zeri di f. Dunque ciascuna ha *un solo* punto unito in $(0, 1)$.

Per ciascuno dei tre metodi ci si domanda se sia *utilizzabile*, ovvero se sia possibile *determinare un intervallo che, insieme alla funzione che definisce il metodo, soddisfa le ipotesi (1) e (2) del Teorema di convergenza*. Se il metodo risulta utilizzabile, *si utilizza il Criterio di scelta del punto iniziale* per determinare un valore a partire dal quale la successione generata dal metodo ad un punto risulta *convergente* ad α.

- *Metodo definito da h_1.*

  La funzione h_1 ha derivata prima continua. L'ipotesi (1) del Teorema di convergenza richiede un intervallo *chiuso* su cui h_1 è definita e che include il punto unito. L'intervallo $[0, 1]$ non è utilizzabile: la funzione h_1 non è definita in 0. Un intervallo che soddisfa le richieste è $[\frac{1}{2}, 1]$, ottenuto constatando che nel punto medio dell'intervallo $[0, 1]$ la funzione f assume valore *negativo* ed utilizzando il Teorema di esistenza degli zeri.

  Scelto l'intervallo, studiamo la *derivata prima* di h_1. Per ogni $x \in [\frac{1}{2}, 1]$ si ha:

  $$|\, h_1'(x)\,| = \frac{1}{x} \geqslant 1$$

  dunque l'ipotesi (2) *non è verificata*. In questo caso *non esiste* un intervallo che verifica le ipotesi (1) e (2) perché essendo $\alpha \in (\frac{1}{2}, 1)$ si ha certamente:

  $$|\, h_1'(\alpha)\,| > 1$$

  Il metodo è *non utilizzabile*.

- *Metodo definito da h_2.*

  La funzione h_2 ha derivata prima continua. L'intervallo $[\frac{1}{2}, 1]$ verifica l'ipotesi (1). Inoltre per ogni $x \in [\frac{1}{2}, 1]$ si ha:

  $$|\, h_2'(x)\,| = e^{-x} \leqslant L_2 = \frac{1}{\sqrt{e}} < 1$$

  dunque è verificata anche l'ipotesi (2): il metodo è *utilizzabile*. Poiché $f(\frac{3}{4}) > 0$, per il Criterio di scelta del punto iniziale la successione x_k generata a partire da $\gamma = \frac{1}{2}$ è convergente ad α.

  Essendo $h_2'(x) < 0$ per ogni $x \in [\frac{1}{2}, 1]$, utilizzando il Teorema di Lagrange si può dedurre la seguente *proprietà qualitativa* della successione: per ogni k le differenze $x_k - \alpha$ e $x_{k+1} - \alpha$ sono non nulle ed *hanno segno opposto*, ovvero: x_k ed x_{k+1} sono "da parti opposte" rispetto ad α.

- *Metodo definito da h_3.*

  La funzione h_3 ha derivata prima continua e l'intervallo $[\frac{1}{2}, 1]$ verifica l'ipotesi (1). Inoltre per ogni $x \in [\frac{1}{2}, 1]$ si ha:

  $$|\, h_3'(x)\,| = \frac{1 - e^{-x}}{2} \leqslant L_3 = \frac{1 - 1/e}{2} < 1$$

  dunque è verificata anche l'ipotesi (2): il metodo è *utilizzabile*. Come già stabilito studiando il metodo definito da h_2, per il Criterio di scelta del punto iniziale la successione x_k generata a partire da $\gamma = \frac{1}{2}$ è convergente ad α.

[22] Ovvero: è un intervallo *non degenere e limitato* che include *un solo* zero di f.

Essendo $h_3'(x) > 0$ per ogni $x \in [\frac{1}{2}, 1]$, utilizzando il Teorema di Lagrange si può dedurre la seguente *proprietà qualitativa* della successione generata a partire da *qualsiasi* $\gamma \in [\frac{1}{2}, 1]$: per ogni k le differenze $x_k - \alpha$ sono non nulle ed *hanno lo stesso segno*, ovvero: x_k ed x_{k+1} sono "dalla stessa parte" rispetto ad α. Allora: poiché la successione delle *distanze* $|\, x_k - \alpha \,|$ è, come sappiamo dalla dimostrazione del Teorema di convergenza, *decrescente*, si conclude che la successione x_k è *monotona* (crescente se $\gamma < \alpha$, decrescente se $\gamma > \alpha$). Nel caso in esame, $\gamma = \frac{1}{2} < \alpha$ e la successione è monotona crescente.

1.3.6 Osservazione (metodo utilizzabile per approssimare un punto unito)

Si è scelto di dichiarare un metodo *utilizzabile* (*per approssimare un punto unito* α) quando *è possibile determinare un intervallo che, insieme alla funzione che definisce il metodo, soddisfi le ipotesi (1) e (2) del Teorema di convergenza.*

Un metodo è certamente utilizzabile se è definito da una funzione h con derivata prima continua e nel punto unito in esame si ha $|\, h'(\alpha) \,| < 1$. In tal caso, infatti, la continuità di $h'(x)$ garantisce l'esistenza di un intervallo chiuso che contiene α e in tutti i punti x del quale di ha $|\, h'(x) \,| < 1$. Osservando che se l'ipotesi (2) del Teorema di convergenza è soddisfatta allora si ha $|\, h'(\alpha) \,| < 1$, si conclude che:

un metodo è utilizzabile per approssimare il punto unito α se e solo se $|\, h'(\alpha) \,| < 1$

Una *condizione sufficiente* di *non* utilizzabilità di un metodo è che esso sia definito da una funzione h con derivata prima continua e che nel punto unito in esame si abbia $|\, h'(\alpha) \,| > 1$ (è la situazione incontrata analizzando il metodo definito da h_1). La *non utilizzabiltà* del metodo in questo caso è motivata dall'osservazione che si ha: *Se* x_k è una successione generata dal metodo ad un punto definito da h *allora*:

$$x_k \text{ è definitivamente uguale a } \alpha \qquad \text{oppure} \qquad x_k \text{ non converge ad } \alpha \qquad (*)$$

(*Dimostrazione.* Supponiamo che per ogni k sia $x_k \neq \alpha$. Dobbiamo dimostrare che, allora, x_k non converge ad α.

Si osservi, preliminarmente, che poiché h' è una funzione continua e $|\, h'(\alpha) \,| > 1$, esistono due numeri reali positivi ρ e δ tali che: $|\, h'(x) \,| > 1 + \delta$ per ogni x nell'intorno $I_\rho(\alpha)$ di centro α e raggio ρ.

Adesso, procedendo *per assurdo*, supponiamo che $\lim x_k = \alpha$. Allora esiste un numero intero positivo n tale che $x_k \in I_\rho(\alpha)$ per ogni $k \geqslant n$. Sia poi m un numero intero tale che:

$$m > n \qquad \text{e} \qquad (1+\delta)^m > \frac{\rho}{|\, x_0 - \alpha \,|}$$

Utilizzando ripetutamente il Teorema di Lagrange si ottiene che esistono $\theta_{m-1}, \ldots, \theta_0 \in I_\rho(\alpha)$ tali che:

$$|\, x_m - \alpha \,| = |\, h'(\theta_{m-1}) \,| \cdots |\, h'(\theta_0) \,|\, |\, x_0 - \alpha \,|$$

Ma per ogni $j = 0, \ldots, m-1$ si ha: $|\, h'(\theta_j) \,| > 1 + \delta$ e quindi:

$$|\, x_m - \alpha \,| = |\, h'(\theta_{m-1}) \,| \cdots |\, h'(\theta_0) \,|\, |\, x_0 - \alpha \,| > (1+\delta)^m \,|\, x_0 - \alpha \,| > \rho$$

ovvero $x_m \notin I_\rho(\alpha)$. Questo è assurdo perchè, essendo $m > n$, si ha $x_m \in I_\rho(\alpha)$.)

Dunque: anche se $|\, h'(\alpha) \,| > 1$, il metodo *può* generare successioni convergenti ad α (se ne ottiene una, ad esempio, scegliendo come valore iniziale α) ma *non è ragionevole* supporre di poter ottenere un valore iniziale *praticamente utilizzabile* per la costruzione di una successione convergente.

Anche la condizione $|\, h'(\alpha) \,| = 1$ è *sufficiente* per dichiarare il metodo *non* utilizzabile, ma in questo caso non necessariamente sussiste l'asserto $(*)$. Ritorneremo a discutere questa condizione dopo aver introdotto la nozione di *ordine di convergenza* di un metodo.

Esercizi

E11 Sia $h(x) = \frac{1}{2} \cos x$.

(1) Dimostrare che l'intervallo $[0, \pi/2]$ *separa* un punto unito, α, di h.

(2) Constatare che le ipotesi (1) e (2) del Teorema di convergenza sono verificate con $[a, b] = [0, \pi/2]$.

(3) Dimostrare che *se* $x \in [0, \pi/2]$ *allora* $h(x) \in [0, \pi/2]$.

(4) Determinare *tutti* i valori $\gamma \in [0, \pi/2]$ a partire dai quali la successione generata dal metodo definito da h risulta convergente ad α.

E12 Sia $h : (0, +\infty) \to \mathbb{R}$ la funzione definita da $h(x) = 3 - \frac{1}{2}x$. Dimostrare che h ha derivata prima continua e che le ipotesi (1) e (2) del Teorema di convergenza sono verificate con $[a, b] = [1, 7]$. Discutere gli assegnamenti $z = \mathit{MetodoUnPunto}\,(h, 7)$ e $z = \mathit{MetodoUnPunto}\,(h, 1)$.

E13 Dimostrare la *versione lipschitziana* del Teorema di convergenza:

Siano $[a, b]$ un intervallo, $h : [a, b] \to \mathbb{R}$ una funzione continua e γ un elemento di $[a, b]$ tali che:

(1) esiste α punto unito di h in $[a, b]$;

(2) esiste $L \in [0, 1)$ tale che per ogni $x, y \in [a, b]$ si ha: $|\, h(x) - h(y)\,| \leqslant L\,|\, x - y\,|$;[23]

Allora: (i) α è l'*unico* punto unito di h in $[a, b]$, (ii) la procedura $\mathit{MetodoUnPunto}\,(h, \gamma)$ genera una successione x_k in $[a, b]$ e (iii) la successione x_k è *convergente* ad α.

E14 ★ Si consideri una funzione continua $f : [a, b] \to \mathbb{R}$ con derivata seconda su (a, b).

(1) Dimostrare che *se* $\alpha < \beta < \gamma$ sono tre zeri di f in $[a, b]$ *allora* esiste $c \in (a, b)$ tale che $f''(c) = 0$. (Suggerimento: applicare il Teorema di Rolle[24] prima alla funzione f poi ad f'.)

(2) Dedurne che: *se* per ogni $x \in (a, b)$ si ha $f''(x) \neq 0$ *allora* f ha *al più* due zeri in $[a, b]$.

In generale si ha: *se* la funzione continua $f : [a, b] \to \mathbb{R}$ ha derivata k-esima su (a, b) e per ogni $x \in (a, b)$ si ha $f^{(k)}(x) \neq 0$ *allora* f ha *al più* k zeri in $[a, b]$.

E15 Sia $f : \mathbb{R} \to \mathbb{R}$ una funzione tale che: per ogni x si ha $f^{(3)}(x) \neq 0$, per ogni $x < 0$ si ha $f'(x) \neq 0$, $f(-1) > 0$ e $f(0) < 0$. Cosa si può dedurre riguardo agli zeri di f?

E16 ★ Sia $h : [a, b] \to \mathbb{R}$ una funzione continua. Applicare i risultati dell'Esercizio *E14* alla funzione f definita da $f(x) = x - h(x)$ e dedurne condizioni sufficienti affinché h abbia *al più* uno o, rispettivamente, *al più* due punti uniti in $[a, b]$.

Nell'Esempio 1.3.5 si sono trovati *due* metodi utilizzabili per approssimare lo zero α di f. Per decidere se uno dei due metodi sia da preferirsi rispetto all'altro studiamo la *rapidità di convergenza* ad α delle successioni generate.

1.3.7 Definizione (ordine di convergenza di un metodo ad un punto)

Siano $[a, b]$, h e γ che verificano le ipotesi del Teorema di convergenza e supponiamo che per la successione x_k, convergente al punto unito $\alpha \in [a, b]$, si abbia $x_k \neq \alpha$ per ogni k.

– *Sia* $h'(\alpha) \neq 0$. Per il Teorema di Lagrange, per ogni k esiste θ_k tra x_k ed α tale che:

$$|x_{k+1} - \alpha| = |h(x_k) - h(\alpha)| = |h'(\theta_k)|\,|x_k - \alpha|$$

Tenuto conto che h' è una funzione continua e che $\lim x_k = \alpha \;\Rightarrow\; \lim \theta_k = \alpha$ si ha:

$$\lim_{k\to\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|} = |h'(\alpha)| \in (0, 1)$$

ovvero: per ogni $\epsilon > 0$ esiste un indice n tale che:

$$\text{se } k \geqslant n \text{ allora} \qquad |h'(\alpha)| - \epsilon \leqslant \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|} \leqslant |h'(\alpha)| + \epsilon$$

[23] Una funzione che verifica questa proprietà si chiama *contrazione* su $[a, b]$. La disuguaglianza significa, infatti, che h "contrae" la distanza tra x ed y.

[24] Se la funzione continua $f : [a, b] \to \mathbb{R}$ è derivabile su (a, b) e $f(a) = f(b)$ allora esiste $c \in (a, b)$ tale che $f'(c) = 0$.

e quindi tale che:

$$\text{se } k \geqslant n \text{ allora} \qquad \big(|h'(\alpha)| - \epsilon\big)|x_k - \alpha| \leqslant |x_{k+1} - \alpha| \leqslant \big(|h'(\alpha)| + \epsilon\big)|x_k - \alpha|$$

Allora, scelto ϵ sufficientemente piccolo (precisamente: $\epsilon < \min\{\, |h'(\alpha)|, 1 - |h'(\alpha)| \,\}$), esiste un numero intero positivo n tale che:

$$\text{per ogni } k \geqslant n \text{ si ha:} \quad (|\, h'(\alpha)\,| - \epsilon)^{k-n} |\, x_n - \alpha\,| \leqslant |\, x_k - \alpha\,| \leqslant (|\, h'(\alpha)\,| + \epsilon)^{k-n} |\, x_n - \alpha\,|$$

ovvero: la successione $x_k - \alpha$ tende a zero *almeno* rapidamente come $(|\, h'(\alpha)\,| + \epsilon)^k$ ma *non più* rapidamente di $(|\, h'(\alpha)\,| - \epsilon)^k$.

– Sia $h'(\alpha) = 0$ e la funzione h abbia *derivata seconda continua* con $h''(\alpha) \neq 0$. Poiché h ha derivata seconda continua, sussiste lo sviluppo di Taylor in α con resto in forma di Lagrange: per ogni $x \in (a, b)$ esiste τ tra x ed α tale che:

$$h(x) = h(\alpha) + h'(\alpha)(x - \alpha) + \tfrac{1}{2} h''(\tau)(x - \alpha)^2$$

ovvero, essendo $h'(\alpha) = 0$:

$$h(x) = h(\alpha) + \tfrac{1}{2} h''(\tau)(x - \alpha)^2$$

Dunque: per ogni k esiste τ_k tra x_k ed α tale che:

$$|x_{k+1} - \alpha| = |h(x_k) - h(\alpha)| = |\tfrac{1}{2} h''(\tau_k)|\, |x_k - \alpha|^2$$

Tenuto conto che $\lim \tau_k = \alpha$ si ha:

$$\lim_{k \to \infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = |\tfrac{1}{2} h''(\alpha)| \neq 0$$

Da questa relazione si deduce (dimostrazione riportata nell'Appendice 1) che: per ogni $\theta > 0$ si ha:

$$\lim_{k \to \infty} \frac{|x_k - \alpha|}{\theta^k} = 0$$

ovvero: la successione $x_k - \alpha$ tende a zero più rapidamente di qualsiasi successione di tipo esponenziale.

Si chiama *ordine di convergenza* del metodo ad un punto definito da h quando utilizzato per approssimare il punto unito α: *il più piccolo numero intero q tale che $h^{(q)}(\alpha) \neq 0$.*

Si ha dunque (si ricordi che si stanno considerando solo le successioni x_k tali che $x_k \neq \alpha$ per ogni k):

– Se $h'(\alpha) \neq 0$, l'ordine di convergenza è *uno* e *per tutte* le successioni convergenti x_k generate dal metodo la distanza $|\, x_k - \alpha\,|$ tende a zero *sostanzialmente* come $|\, h'(\alpha)\,|^k$.

– Se $h'(\alpha) = 0$, l'ordine è *almeno due* e per *tutte* le successioni convergenti x_k generate dal metodo la distanza $|\, x_k - \alpha\,|$ tende a zero più rapidamente di qualsiasi successione di tipo esponenziale. Dunque: *qualunque successione convergente generata da un metodo di ordine due converge più rapidamente di qualunque successione convergente generata da un metodo di ordine uno.*

– In generale: *qualunque successione convergente generata da un metodo di ordine q converge più rapidamente di qualunque successione convergente generata da un metodo di ordine minore di q.*

1.3.8 Osservazione

Siano h una funzione con derivata prima continua, α un punto unito di h, e $|\, h'(\alpha)\,| = 1$. Sia infine x_k una successione generata dal metodo iterativo definito da h. *Se* $\lim x_k = \alpha$ e per ogni k si ha $x_k \neq \alpha$ *allora*:[25]

$$\textit{per ogni } \theta \in (0, 1) \textit{ si ha:} \qquad \lim_{k \to +\infty} \frac{|\, x_k - \alpha\,|}{\theta^k} = +\infty$$

[25] La dimostrazione dell'asserto è riportata nell'Appendice 2 di fine capitolo.

ovvero: la successione $x_k - \alpha$ tende a zero *più lentamente* di qualsiasi successione di tipo esponenziale.

1.3.9 Esempio (continuazione)

Per i due metodi utilizzabili individuati nell'Esempio 1.3.5 si ha:

$$|\, h_2'(\alpha)\,| = e^{-\alpha} \neq 0 \qquad \text{e} \qquad |\, h_3'(\alpha)\,| = \frac{1-e^{-\alpha}}{2} \neq 0$$

dunque entrambi hanno ordine di convergenza *uno*. Essendo poi:

$$|\, h_2'(\alpha)\,| = e^{-\alpha} > \frac{1-e^{-\alpha}}{2} = |\, h_3'(\alpha)\,|$$

si conclude che il metodo definito da h_3 genera una successione che tende ad α *più rapidamente* del metodo definito da h_2.

1.3.10 Osservazione (Studio grafico di un metodo ad un punto)

Si suppongano rappresentati, in uno stesso piano cartesiano, *i grafici* della funzione h e quello della funzione identità, entrambi su un intervallo (limitato) $[a, b]$.

- *Ricerca dei punti uniti di h in $[a, b]$.*

  I punti uniti di h in $[a, b]$ sono *le ascisse dei punti di intersezione* dei due grafici. Infatti, se $A \equiv (\bar{x}, \bar{y})$[26] è uno dei punti di intersezione si ha: $\bar{y} = \bar{x}$ (perché A fa parte del grafico della funzione identità) e $\bar{y} = h(\bar{x})$ (perché A fa parte del grafico della funzione h) e quindi $\bar{x} = h(\bar{x})$.

- *Costruzione di un elemento della successione generata dal metodo.*

  Assegnato un elemento x in $[a, b]$ è possibile rappresentare $h(x)$ sull'asse delle ascisse con la costruzione seguente:

  (1) Si disegna la retta *verticale* passante per il punto $P \equiv (x, 0)$ e si individua il punto $Q \equiv (x, h(x))$ intersezione della retta con il grafico di h.

  (2) Si disegna la retta *orizzontale* passante per il punto Q e si individua il punto $R \equiv (h(x), h(x))$ intersezione della retta con il grafico della funzione identità.

  (3) Si disegna la retta *verticale* passante per il punto R e si individua il punto $S \equiv (h(x), 0)$ intersezione della retta con l'asse delle ascisse.

- *Studio dell'utilizzabilità del metodo.*

  Sia $A \equiv (\alpha, \alpha)$ un punto di intersezione dei due grafici. Per studiare l'utilizzabilità del metodo per approssimare α:

  (i) Si considerano la retta t tangente al grafico di h in A, la retta b grafico della funzione identità (già presente nel disegno) e la retta p grafico della funzione $x \mapsto \alpha - x$, e

  (ii) Si *ruota* la retta b intorno al punto A in senso *orario*.

  Si ha: $|\, h'(\alpha)\,| < 1$ *se e solo se* $t \neq b, t \neq p$ e nella rotazione b si sovrappone *prima* a t e *poi* a p.

Esercizi

E17 ★ Si consideri la funzione h_3 definita nell'Esempio 1.3.5. Per ogni $x \in [\frac{1}{2}, 1]$ si ha $h_3'(x) > 0$. Dimostrare che:

(1) *se* $x \in [\frac{1}{2}, \alpha)$ *allora* $h_3(x) \in (x, \alpha)$.

(2) *se* $x \in (\alpha, 1]$ *allora* $h_3(x) \in (\alpha, x)$.

[26] Il simbolo $\equiv$ si legge: "di coordinate."

Dedurne che *se* $x \in [\frac{1}{2}, 1]$ *allora* $h(x) \in [\frac{1}{2}, 1]$ e quindi che *per ogni* $\gamma \in [\frac{1}{2}, 1]$ la successione generata dal metodo definito da h_3 *converge* ad α.

E18 ★ Siano $[a, b]$, h e γ che verificano le ipotesi del Teorema di convergenza. Inoltre, per ogni $x \in [a, b]$ sia:

$$\lambda \leqslant |\, h'(x)\,| \leqslant L$$

Dimostrare che, allora:

$$\lambda^k |\, x_0 - \alpha\,| \leqslant |\, x_k - \alpha\,| \leqslant L^k |\, x_0 - \alpha\,|$$

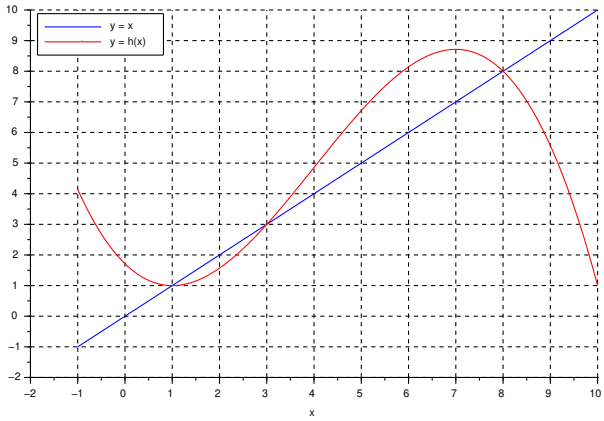
E19 Sia f la funzione definita, per ogni $x \in \mathbb{R}$ da $f(x) = x - e^{x-2}$.

(1) Dimostrare che f ha *due* zeri e *separarli*.

(2) Dimostrare che i punti uniti della funzione h definita da $h(x) = e^{x-2}$ sono *tutti e soli* gli zeri di f.

(3) Dimostrare, prima *graficamente* poi *analiticamente*, che il metodo ad un punto definito da h è *utilizzabile* per approssimare uno degli zeri (ed il metodo risulta di *ordine uno* quando utilizzato per approssimare tale zero) e *non utilizzabile* per l'altro.

E20 Nella figura seguente, generata da *Scilab*, sono rappresentati, sull'intervallo $I = [-1, 10]$, il grafico della funzione $h : I \to \mathbb{R}$ (in rosso) e quello della funzione identità (in blu).



Individuare i punti uniti di h e, per ciascuno di essi: decidere se il metodo iterativo definito da h sia utilizzabile per l'approssimazione ed eventualmente indicare l'ordine di convergenza.

1.4 Metodo di Newton

Tra tutti i metodi ad un punto, il *metodo di Newton* è di uso particolarmente frequente.

1.4.1 Definizione (metodo di Newton)
Sia $f : [a, b] \to \mathbb{R}$ con derivata *prima* continua e per ogni $x \in [a, b]$ sia $f'(x) \neq 0$. Il *metodo di Newton* (*applicato ad* f) è il metodo ad un punto definito dalla funzione:

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

Siano dunque $f : [a, b] \to \mathbb{R}$ una funzione con derivata *prima* continua tale che $f'(x) \neq 0$ per ogni $x \in [a, b]$ e h la funzione che definisce il metodo di Newton.

1.4.2 Osservazione (utilizzabilità e ordine di convergenza del Metodo di Newton)
Per quanto mostrato nell'Esempio 1.3.2, *i punti uniti di* h *sono* tutti e soli *gli zeri di* f.

Inoltre: Se f ha derivata *seconda* continua si ha:

$$h'(x) = \frac{f''(x)f(x)}{\big(f'(x)\big)^2}$$

e quindi, detto α uno zero di f:

$$h'(\alpha) = 0$$

Si deduce che:

- Per quanto detto nell'Osservazione 1.3.6, la condizione $|\, h'(\alpha)\,| = 0 < 1$ è *sufficiente* per poter affermare che *il metodo di Newton è utilizzabile per approssimare* α.
- Il metodo ha *ordine di convergenza almeno due* quando utilizzato per approssimare α.

1.4.3 Osservazione (Interpretazione geometrica del Metodo di Newton: metodo delle tangenti)
Si rappresenti su un piano cartesiano il grafico della funzione f su $[a, b]$. Assegnato $z \in [a, b]$ il valore $h(z)$ si determina con la seguente costruzione grafica:

- Si disegna la retta t tangente al grafico di f nel punto $P \equiv (z, f(z))$;
- Si determina i punto $Q \equiv (\bar{z}, 0)$ intersezione di t con l'asse delle ascisse (l'intersezione è un punto perché, essendo $f'(z) \neq 0$, la retta t non è orizzontale): si ha $\bar{z} = h(z)$.

Infatti: L'equazione della retta tangente t è:

$$y = f'(z)(x - z) + f(z)$$

da cui si ricava l'ascissa di Q:

$$0 = f'(z)(x - z) + f(z) \quad \Rightarrow \quad \bar{z} = z - \frac{f(z)}{f'(z)} = h(z)$$

1.4.4 Osservazione (Criterio di scelta del valore iniziale per il metodo di Newton)
Siano f con derivata *seconda* continua ed I un intervallo contenente α zero di f e tale che:

$$\text{per ogni } x \in I \text{ si ha } f'(x) \neq 0 \text{ e } f''(x) \neq 0$$

Sia infine γ un elemento di I, certamente esistente (perché ?), tale che:

$$f(\gamma)f''(\gamma) > 0$$

Allora: la successione generata dal metodo di Newton a partire da γ *è convergente* ad α e *monotona*.
Dimostrazione: Per via grafica, in un caso particolare, si dimostra che la successione è monotona e limitata, dunque *convergente*. Il limite della successione è uno zero di f, ovvero un punto unito della funzione h, perché la successione è generata da un metodo ad un punto definito da una funzione *h continua*.

1.4.5 Esempio
Sia $f(x) = x + \log x$, definita per ogni $x > 0$. Sappiamo già che f ha un solo zero, α, separato dall'intervallo $[\frac{1}{2}, 1]$. La funzione f ha derivata prima *sempre positiva* e derivata seconda continua, dunque *il metodo di Newton è utilizzabile* per approssimare α. Inoltre, la derivata seconda è *sempre negativa*, dunque il criterio di scelta del valore iniziale per il metodo di Newton è utilizzabile e stabilisce che *per ogni* $\gamma \in [\frac{1}{2}, \alpha)$ la successione generata dal metodo di Newton converge allo zero ed è monotona crescente. Si osservi che, non essendo noto il valore di α, l'unico punto accessibile di quest'ultimo intervallo è $\frac{1}{2}$.

1.5 Criteri d'arresto per metodi ad un punto

I criteri d'arresto studiati per il metodo di bisezione (calcolabili, efficaci e che consentono in ogni caso di arrestare la costruzione della successione non appena si è trovata un'approssimazione sufficientemente accurata di uno zero di f) si basano sulla costruzione di una successione di intervalli che racchiudono uno zero di f. I metodi ad un punto, in particolare il metodo di Newton, *non* costruiscono successioni di intervalli, dunque quei criteri d'arresto non sono utilizzabili.

1.5.1 Definizione (criteri d'arresto di tipo assoluto)

Siano $f : \mathbb{R} \to \mathbb{R}$ la funzione continua della quale si vuole approssimare uno zero, α uno zero di f, x_k una successione convergente ad α e $s : \mathbb{R} \to \mathbb{R}$ una funzione con *derivata prima continua* e tale che $s(\alpha) = 0$.

Allora:

(a) Per la continuità di s si ha: $\lim s(x_k) = s(\alpha) = 0$.

(b) Per il Teorema di Lagrange, esiste θ_k tra x_k ed α tale che:

$$s(x_k) = s(x_k) - s(\alpha) = s'(\theta_k)(x_k - \alpha)$$

e l'errore relativo E_k commesso approssimando $\delta_k = |x_k - \alpha|$ con $S_k = |s(x_k)|$ è:

$$E_k = \frac{S_k - \delta_k}{\delta_k} = |s'(\theta_k)| - 1$$

Inoltre, per la continuità di s' e la convergenza della successione θ_k ad α si ha: $\lim E_k = |s'(\alpha)| - 1$.

(c) Per ogni k, x_k è zero della funzione continua $s_k^*(x) = s(x) - s(x_k)$ tale che:

$$\text{per ogni } x \text{ si ha:} \quad |s_k^*(x) - s(x)| \leqslant |s(x_k)|$$

Si considerino adesso la procedura iterativa che costruisce la successione x_k e, dato un numero reale positivo δ, il seguente criterio d'arresto:

$$\textit{se } |s(x_k)| < \delta \textit{ allora} \text{ esci dal ciclo}$$

Il criterio d'arresto è introdotto nella procedura *MetodoUnPunto* modificandola come segue:

$z = \textit{MetodoUnPunto}\,(h, \gamma, s, \delta)$

// $h : [a, b] \to \mathbb{R}$ funzione *continua*, $\gamma \in [a, b]$,

// $s : [a, b] \to \mathbb{R}$ funzione, *con derivata prima continua*, tale che:

// α punto unito di $h \Leftrightarrow s(\alpha) = 0$,

// δ numero reale *positivo*

$x_0 = \gamma$;

$k = 0$;

`ripeti`:

`se` $\big(x_k \notin [a, b]$ `oppure` $|s(x_k)| < \delta\big)$ `allora` esci dal ciclo;
$x_{k+1} = h(x_k)$;
$k = k + 1$;

$z = x_k$

Si ha:

(1) Il criterio è *calcolabile*.

(2) Per quanto mostrato nel punto (a), il criterio è *efficace*.

(3) Per quanto mostrato nel punto (b), *il criterio utilizza* $|s(x_k)|$ *per stimare l'errore assoluto* $|x_k - \alpha|$. Si ha:

- Se $|s'(\alpha)| = 1$, possiamo ritenere la stima, per k sufficientemente elevato, *buona*, ed il criterio arresta la costruzione non appena $|x_k - \alpha| < \delta$.
- Se $|s'(\alpha)| > 1$, per k sufficientemente elevato si ha $E_k > 0$ e quindi $|x_k - \alpha| < |s(x_k)|$. Il criterio arresta la costruzione non appena $|s(x_k)| < \delta$ e in tal caso x_k è un'approssimazione sufficientemente accurata di α, *ma* la condizione $|x_n - \alpha| < \delta$ potrebbe essere stata già verificata per $n < k$: il criterio rischia di accorgersi *in ritardo* che l'approssimazione è sufficientemente accurata.
- Se $|s'(\alpha)| < 1$, per k sufficientemente elevato si ha $E_k < 0$ e quindi $|x_k - \alpha| > |s(x_k)|$. Il criterio rischia di arrestare la costruzione quando $|x_k - \alpha| > \delta$, dunque con un'approssimazione *non* sufficientemente accurata.

(4) Per quanto mostrato al punto (c), quando il criterio è verificato, la procedura restituisce un elemento x_k tale che $|s(x_k)| < \delta$, e sussiste la seguente *interpretazione*: x_k è zero della funzione continua $s^*(x) = s(x) - s(x_k)$ tale che:

$$\text{per ogni } x \text{ si ha:} \quad |s^*(x) - s(x)| < \delta$$

Questa interpretazione *non fornisce direttamente informazioni sull'accuratezza* di x_k come approssimazione di α. Per averle occorre studiare quanto distante può essere lo zero x_k di s^* dallo zero α di s in termini di δ, ovvero occorre studiare il *condizionamento del calcolo di uno zero* di s. Come vedremo nella Sezione 1.6, *nel peggiore dei casi* si ha:

$$|x_k - \alpha| \approx \frac{\delta}{|s'(\alpha)|}$$

1.5.2 Esempio

Siano $h, [a, b]$ e γ che verificano le ipotesi del Teorema di convergenza e x_k la successione generata dal metodo ad un punto definito da h a partire da γ, convergente al punto unito α.

Assegnato un numero reale positivo δ, un criterio d'arresto comunemente utilizzato è il seguente:

$$\textit{se } |\, h(x_k) - x_k \,| < \delta \textit{ allora} \text{ esci dal ciclo}$$

che corrisponde alla scelta $s(x) = h(x) - x$. Con questa scelta, s è una funzione con derivata prima continua e con zeri coincidenti con i punti uniti di h.

Per quanto mostrato nella definizione precedente, il criterio è calcolabile ed efficace, e l'accuratezza di x_k come approssimazione di α dipende da $s'(\alpha) = h'(\alpha) - 1$. Infine, quando il criterio è verificato, x_k risulta essere uno zero della funzione $s^*(x) = s(x) - s(x_k) = h(x) - x - (h(x_k) - x_k)$.

I questo caso, è interessante anche interpretare x_k come un *punto unito* della funzione $h^*(x) = h(x) - (h(x_k) - x_k)$, *vicina* ad h nel senso che: per ogni x si ha $|h^*(x) - h(x)| < \delta$. Come già osservato, anche questa interpretazione *non fornisce direttamente informazioni sull'accuratezza* di x_k come approssimazione di α. Per averle occorre studiare quanto distante può essere il punto unito x_k di h^* dal punto unito α di h in termini di δ, ovvero occorre studiare il *condizionamento del calcolo di un punto unito* di h. Come vedremo nella Sezione 1.6, *nel peggiore dei casi* si ha ancora:

$$|x_k - \alpha| \approx \frac{\delta}{1 - h'(\alpha)}$$

1.5.3 Esempio

Siano $f : \mathbb{R} \to \mathbb{R}$ la funzione *con derivata seconda continua e derivata prima mai nulla* della quale si vuole approssimare uno zero, α uno zero di f, x_k una successione convergente ad α e

$$s(x) = \frac{f(x)}{f'(x)}$$

La funzione s ha derivata prima continua e $s(\alpha) = 0$.

Inoltre, l'errore relativo E_k commesso approssimando $\delta_k = |x_k - \alpha|$ con $S_k = |s(x_k)|$ è (si veda il punto (b) della Definizione 1.5.1):

$$E_k = \frac{S_k - \delta_k}{\delta_k} = |s'(\theta_k)| - 1 = -\frac{f''(\theta_k) f(\theta_k)}{\big(f'(\theta_k)\big)^2}$$

e, per la continuità di f, f' e f'', e la convergenza delle successioni x_k e θ_k ad α si ha: $\lim E_k = 0$.

Si considerino ancora la procedura iterativa che costruisce la successione x_k e, dato un numero reale positivo δ, il seguente criterio d'arresto:

$$\textit{se } |f(x_k)/f'(x_k)| < \delta \textit{ allora} \text{ esci dal ciclo}$$

Il criterio d'arresto è introdotto nella procedura *MetodoUnPunto* modificandola come segue:

$z = \mathit{MetodoUnPunto}\,(h, \gamma, f, f', \delta)$

// $h : [a, b] \to \mathbb{R}$ funzione *continua*, $\gamma \in [a, b]$,

// $f : [a, b] \to \mathbb{R}$ funzione *con derivata prima continua e mai nulla*,

// δ numero reale *positivo*

$x_0 = \gamma$;

$k = 0$;

`ripeti`:

`se` $\big(x_k \notin [a, b]$ `oppure` $|f(x_k)/f'(x_k)| < \delta \big)$ `allora` esci dal ciclo;
$x_{k+1} = h(x_k)$;
$k = k + 1$;

$z = x_k$

Si ha in questo caso:

(1) Il criterio è *calcolabile* e *efficace*.

(2) *Il criterio utilizza* $|f(x_k)/f'(x_k)|$ *per stimare l'errore assoluto* $|x_k - \alpha|$. Poiché, come mostrato sopra, l'errore relativo E_k commesso approssimando $\delta_k = |x_k - \alpha|$ con $S_k = |s(x_k)|$ tende a zero al crescere di k, possiamo ritenere la stima, per k sufficientemente elevato, *buona*, ed il criterio arresta la costruzione non appena $|x_k - \alpha| < \delta$.

(3) Per quanto mostrato al punto (c) della Definizione1.5.1, quando il criterio è verificato, la procedura restituisce un elemento x_k tale che $|s(x_k)| < \delta$, e sussiste la seguente *interpretazione*: x_k è zero della funzione continua

$$s^*(x) = \frac{f(x)}{f'(x)} - \frac{f(x_k)}{f'(x_k)}$$

tale che

$$\text{per ogni } x \text{ si ha:} \quad |s^*(x) - s(x)| < \delta$$

Di nuovo, questa interpretazione *non fornisce direttamente informazioni sull'accuratezza* di x_k come approssimazione di α. Per averle occorre studiare quanto distante può essere lo zero x_k di s^* dallo zero α di s in termini di δ, ovvero occorre studiare il *condizionamento del calcolo di uno zero* di s. Come vedremo nella Sezione 1.6, *nel peggiore dei casi* si ha (essendo, in questo caso, $s'(\alpha) = 1$):

$$|x_k - \alpha| \approx \delta$$

1.6 Condizionamento del calcolo di uno zero o di un punto unito di una funzione

Sia $f : \mathbb{R} \to \mathbb{R}$ una funzione continua ed α uno zero *isolato* di f. Siano poi $[a, b]$ un intervallo che separa α e $f^* : [a, b] \to \mathbb{R}$ una funzione continua *vicina* ad f nel senso che:

$$\text{esiste un numero reale } \delta > 0 \text{ } \textit{piccolo} \text{ tale che: per ogni } x \in [a, b] \text{ si ha: } |\, f^*(x) - f(x)\,| \leqslant \delta$$

Lo studio del *condizionamento* del calcolo di α consiste nel determinare quanto lontano da α può essere uno zero di f^*, rispetto a δ.

In termini grafici, la relazione tra f e f^* si rilegge:

$$\text{per ogni } x \in [a, b] \text{ si ha: } f(x) - \delta \leqslant f^*(x) \leqslant f(x) + \delta$$

dunque: *il grafico di f^* giace nella parte di piano compresa tra il grafico di $f(x) - \delta$ ed il grafico di $f(x) + \delta$.*

Consideriamo alcuni casi in cui f è *sufficientemente regolare*.

– Siano $f'(\alpha) \neq 0$, $[a, b]$ un intorno di α in cui sia ragionevole l'approssimazione:

$$f(x) \approx f'(\alpha)(x - \alpha)$$

e δ tale che $|\, f(a)\,|, |\, f(b)\,| > \delta$. In queste ipotesi f^* ha certamente qualche zero in $[a, b]$. Si consideri, ad esempio, la situazione rappresentata a sinistra in Figura 1, in cui è riportato a tratteggio nero il grafico di $f(x)$, in rosso quello di $f(x) + \delta$ e in blu quello di $f(x) - \delta$.

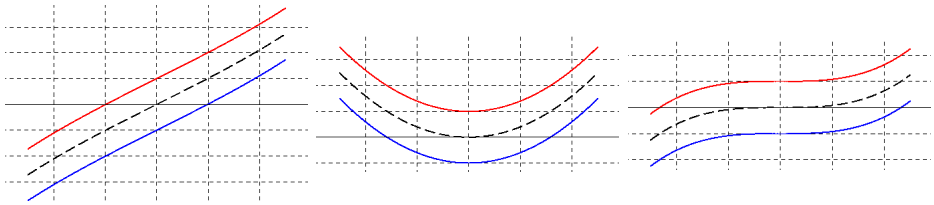


Figura 1: Grafici di f (nero), $f + \delta$ (rosso) e $f - \delta$ (blu).

Come graficamente evidente, il più piccolo intervallo che certamente contiene gli zeri di f^* è quello di estremi le intersezioni con l'asse delle ascisse delle curve rossa e blu. Dunque, se α^* è uno zero di f^*, *nel peggiore dei casi* si ha:

$$|\, \alpha^* - \alpha\,| \approx k\delta \quad \text{con} \quad k = 1/|\, f'(\alpha)\,|$$

Lo scostamento è quindi *propozionale a* δ ed il condizionamento è *tanto peggiore* quanto più $f'(\alpha)$ è *vicino a zero*:

$$\lim_{\delta \to 0} \frac{|\, \alpha^* - \alpha\,|}{\delta} = \frac{1}{|\, f'(\alpha)\,|}$$

– Siano $f'(\alpha) = 0$, $f''(\alpha) \neq 0$ e $[a, b]$ un intorno di α in cui sia ragionevole l'approssimazione:

$$f(x) \approx \tfrac{1}{2}\, f''(\alpha)(x - \alpha)^2$$

In questo caso, schematizzato al centro in Figura 1, il più piccolo intervallo che certamente contiene gli zeri di f^* è quello di estremi le intersezioni con l'asse delle ascisse della curva blu. Dunque, se α^* è uno zero di f^*, *nel peggiore dei casi* si ha:

$$|\, \alpha^* - \alpha\,| \approx k\sqrt{\delta} \quad \text{con} \quad k = \sqrt{2/|\, f''(\alpha)\,|}$$

Lo scostamento è proporzionale alla *radice quadrata* di δ e il calcolo di α è *mal condizionato*:

$$\lim_{\delta \to 0} \frac{|\, \alpha^* - \alpha\,|}{\delta} = +\infty$$

Si osservi anche che in questo caso, per quanto piccolo sia δ, il grafico di f^* potrebbe essere compreso tra le curve nera e rossa e f^* *non avere zeri* in $[a, b]$.

– Siano $f'(\alpha) = f''(\alpha) = 0, f^{(3)}(\alpha) \neq 0$, $[a, b]$ un intorno di α in cui sia ragionevole l'approssimazione:

$$f(x) \approx \tfrac{1}{6} f^{(3)}(\alpha)(x - \alpha)^3$$

e δ tale che $|\, f(a)\,|, |\, f(b)\,| > \delta$. In questo caso, schematizzato a destra in Figura 1, f^* ha certamente qualche zero in $[a, b]$ e il più piccolo intervallo che certamente contiene gli zeri di f^* è quello di estremi le intersezioni con l'asse delle ascisse delle curve rossa e blu. Dunque, se α^* è uno zero di f^*, *nel peggiore dei casi* si ha:

$$|\, \alpha^* - \alpha\,| \approx k \sqrt[3]{\delta} \quad \text{con} \quad k = \sqrt[3]{6/|\, f^{(3)}(\alpha)\,|}$$

Lo scostamento è proporzionale alla *radice cubica* di δ e il calcolo di α è *mal condizionato*:

$$\lim_{\delta \to 0} \frac{|\, \alpha^* - \alpha\,|}{\delta} = +\infty$$

Da questi esempi si deduce che *l'unico caso in cui il calcolo di α è ben condizionato è quello in cui $f'(\alpha)$ è non troppo piccolo.*

1.6.1 Osservazione (condizionamento per funzioni dipendenti da un parametro)

Sia $f(x; t)$ una funzione *regolare* della variabile x e del parametro reale t. Sia poi $\alpha \in \mathbb{R}$ tale che: $f(\alpha; 0) = 0$. Se per la derivata parziale di f rispetto ad x si ha:

$$\partial_x f(\alpha; 0) \neq 0$$

allora (Teorema delle funzioni implicite[27]) esiste una funzione regolare $z(t)$, definita in un intorno I di 0, tale che:

$$\text{(a)} \quad z(0) = \alpha \qquad \text{e} \qquad \text{(b)} \quad \text{per ogni } t \in I \text{ si ha} \quad f\big(z(t); t\big) = 0$$

La funzione z descrive quindi *come varia* lo zero in funzione di t.

La regolarità di z consente di ottenere *un'approssimazione dello scostamento dello zero da α per t piccolo*:

$$z(t) \approx z(0) + z'(0)\, t \quad \text{ovvero} \quad z(t) - \alpha \approx z'(0)\, t$$

Per l'uguaglianza (b) e per la regolarità di f e z si ottiene:

$$\frac{d}{dt} f\big(z(t); t\big)\Big|_{t=0} = \partial_x f\big(z(0); 0\big)\, z'(0) + \partial_t f\big(z(0); 0\big) = 0$$

e quindi:

$$z'(0) = -\frac{\partial_t f\big(z(0); 0\big)}{\partial_x f\big(z(0); 0\big)}$$

dunque:

$$z(t) - \alpha \approx -\frac{\partial_t f\big(z(0); 0\big)}{\partial_x f\big(z(0); 0\big)}\, t$$

1.6.2 Esempio

Sia:

$$f(x; t) = (x - \tfrac{1}{10})(x - 10) + t = x^2 - (10 + \tfrac{1}{10})x + t$$

Posto $\alpha_1 = \frac{1}{10}$ e $\alpha_2 = 10$ si ha:

$$f(\alpha_1; 0) = 0 \quad \text{e} \quad f(\alpha_2; 0) = 0$$

Inoltre:

$$\partial_x f(x; t) = 2x - (10 + \tfrac{1}{10}) \qquad , \qquad \partial_t f(x; t) = 1$$

e quindi:

$$\partial_x f(\alpha_1; 0) = -\tfrac{99}{10} \neq 0 \quad \text{e} \quad \partial_x f(\alpha_2; 0) = \tfrac{99}{10} \neq 0$$

[27] Si veda ad esempio: `https://it.wikipedia.org/wiki/Teorema_delle_funzioni_implicite`.

Allora:

$$\frac{\partial_t f(\alpha_1;0)}{\partial_x f(\alpha_1;0)} = \frac{10}{99} \quad \text{e} \quad \frac{\partial_t f(\alpha_2;0)}{\partial_x f(\alpha_2;0)} = -\frac{10}{99}$$

Misurando lo scostamento degli zeri con l'errore *assoluto* si ha:

$$z(t) - \alpha_1 \approx \tfrac{10}{99}\, t \quad \text{e} \quad z(t) - \alpha_2 \approx -\tfrac{10}{99}\, t$$

e gli zeri subiscono uno scostamento, in valore assoluto, circa uguale.

Se si sceglie di misurare lo scostamento degli zeri con l'errore *relativo* si ha:

$$\frac{z(t) - \alpha_1}{\alpha_1} \approx \tfrac{100}{99}\, t \quad \text{e} \quad \tfrac{z(t)-\alpha_2}{\alpha_2} \approx -\tfrac{1}{99}\, t$$

In questo caso lo zero α_1 subisce uno scostamento, in valore assoluto, *cento volte* maggiore di quello subito dallo zero α_2.

Consideriamo adesso il condizionamento del calcolo di un *punto unito* di una funzione.

Siano h una funzione *sufficientemente regolare* ed α un punto unito *isolato* di h tali che $|h'(\alpha)| < 1$. Siano poi $[a, b]$ un intorno di α in cui sia ragionevole l'approssimazione:

$$h(x) \approx \alpha + h'(\alpha)(x - \alpha)$$

e $h^* : [a, b] \to \mathbb{R}$ una funzione continua *vicina* ad h nel senso che:

esiste un numero reale $\delta > 0$ *piccolo* tale che: per ogni $x \in [a, b]$ si ha: $|\, h^*(x) - h(x)\,| \leqslant \delta$

Lo studio del *condizionamento* del calcolo di α consiste nel determinare quanto lontano da α può essere un punto unito di h^*, rispetto a δ.

Procedendo come nel caso del condizionamento del calcolo di uno zero si ottiene: Se α^* è un punto unito di h^*, *nel peggiore dei casi* si ha:

$$|\, \alpha^* - \alpha\,| \approx k\delta \quad \text{con} \quad k = 1/\big(1 - h'(\alpha)\big)$$

Lo scostamento è quindi *propozionale a* δ ed il condizionamento è *tanto peggiore* quanto più $h'(\alpha)$ è *vicino a uno*:

$$\lim_{\delta \to 0} \frac{|\, \alpha^* - \alpha\,|}{\delta} = \frac{1}{1 - h'(\alpha)}$$

1.7 Uso del tipo *numero in virgola mobile e precisione finita* nei metodi ad un punto

In questa sezione si discute l'esecuzione in *Scilab* della procedura definita nella Definizione 1.5.1.

Si assume, per semplicità, $M = F(2, 53)$ — si veda l'Osservazione 0.1.17 — e si indicano, come usuale, con rd e u, rispettivamente, la funzione arrotondamento e la precisione di macchina in M.

Siano h, $[a, b]$ e $\gamma \in M$ che verificano le ipotesi del Teorema di convergenza. Detto α il punto unito di h in $[a, b]$, la successione di numeri reali x_k generata dal metodo ad un punto definito da h a partire da $x_0 = \gamma$ è convergente ad α e $x_k \in [a, b]$ per ogni k. Sia poi $\phi : [a, b] \to M$ un algoritmo *uniformemente accurato* quando utilizzato per approssimare i valori di h in $[a, b] \cap M$, ovvero:

esiste un numero reale d_ϕ *piccolo* tale che per ogni $\theta \in [a, b] \cap M$ si ha: $\quad |\phi(\theta) - h(\theta)| \leqslant d_\phi$

e tale che la successione ξ_k di elementi di M definita da

$$\xi_0 = \gamma \qquad \text{e} \qquad \xi_{k+1} = \phi(\xi_k) \quad , \quad k = 0, 1, 2, \ldots$$

è contenuta nell'intervallo $[a, b]$.

1.7.1 Teorema (stabilità dei metodi ad un punto)

Sia $f(x) = h(x) - x$. *Se* l'istruzione `MetodoUnPunto(`h, γ, f, δ`)` eseguita in *Scilab* definisce un elemento $\xi \in M$ tale che:

$$|\, \phi(\xi) \ominus \xi\,| < \mathrm{rd}(\delta)$$

allora ξ è un punto unito di una funzione h^*: $[a,b] \to \mathbb{R}$ *vicina* ad h nel senso che:

$$\text{per ogni } x \in [a,b] \text{ si ha:} \quad |h^*(x) - h(x)| \leqslant d_\phi + \delta$$

Sia f una funzione con *derivata prima continua* e tale che $f(\alpha) = 0$. *Se* $\psi : [a,b] \to M$ è un algoritmo *uniformemente accurato* quando utilizzato per approssimare i valori di f in $[a,b] \cap M$, ovvero:

$$\text{esiste un numero reale } d_\psi \textit{ piccolo} \text{ tale che per ogni } \theta \in [a,b] \cap M \text{ si ha:} \quad |\psi(\theta) - f(\theta)| \leqslant d_\psi$$

e l'istruzione `MetodoUnPunto(`h, γ, f, δ`)` eseguita in *Scilab* definisce un elemento $\xi \in M$ tale che:

$$|\,\psi(\xi)\,| < \mathrm{rd}(\delta)$$

allora ξ è uno zero di una funzione $f^* : [a,b] \to \mathbb{R}$ *vicina* ad f nel senso che:

$$\text{per ogni } x \in [a,b] \text{ si ha:} \quad |f^*(x) - f(x)| \leqslant d_\psi + \delta$$

Dimostrazione. Nel primo caso, definito per ogni $x \in [a,b]$:

$$h^*(x) = h(x) - \big(h(\xi) - \xi\big)$$

si ha: $h^*(\xi) = h(\xi) - h(\xi) + \xi = \xi$, ovvero ξ è un punto unito di h^*, e per ogni $x \in [a,b]$:

$$|h^*(x) - h(x)| = |h(\xi) - \xi| = |h(\xi) - \phi(\xi) + \phi(\xi) - \xi| \leqslant |h(\xi) - \phi(\xi)| + |\phi(\xi) - \xi|$$

Il primo addendo, per l'uniforme accuratezza di ϕ, è minore di d_ϕ. Il secondo addendo è minore di δ perché, per la monotonia della funzione rd (Osservazione 0.2.5), si ha:

$$|\,\phi(\xi) \ominus \xi\,| < \mathrm{rd}(\delta) \qquad \Rightarrow \qquad |\,\phi(\xi) - \xi\,| < \delta$$

Nel secondo caso, definito per ogni $x \in [a,b]$:

$$f^*(x) = f(x) - f(\xi)$$

si ha: $f^*(\xi) = f(\xi) - f(\xi) = 0$, ovvero ξ è uno zero di f^*, e per ogni $x \in [a,b]$:

$$|f^*(x) - f(x)| = |f(\xi)| = |f(\xi) - \psi(\xi) + \psi(\xi)| \leqslant |f(\xi) - \psi(\xi)| + |\psi(\xi)|$$

Il primo addendo, per l'uniforme accuratezza di ψ, è minore di d_ψ. Il secondo addendo è minore di δ perché, per la monotonia della funzione rd, si ha:

$$|\psi(\xi)| < \mathrm{rd}(\delta) \qquad \Rightarrow \qquad |\psi(\xi)| < \delta$$

Il teorema è dimostrato.

1.7.2 Osservazione (efficacia del criterio d'arresto)

La procedura introdotta nella Definizione 1.5.1 definisce *in ogni caso* un numero reale perchè il criterio d'arresto utilizzato è *efficace*. Infatti, per la continuità di h ed f:

$$\lim_{k\to\infty} x_k = \alpha \qquad \Rightarrow \qquad \lim_{k\to\infty} |\,f(x_k)\,| = 0$$

L'esempio seguente considera sia il caso particolare di $f(x) = h(x) - x$ che il caso generale e mostra che utilizzando il tipo *numero in virgola mobile e precisione finita* in entrambi i casi il criterio d'arresto *può risultare non efficace.*

1.7.3 Esempio

Siano $[a,b]$ un intervallo *non contenente zero*, $\phi : [a,b] \to M$ l'algoritmo scelto per approssimare h, $\gamma \in [a,b] \cap M$ e ξ_k la successione di elementi di $[a,b] \cap M$ definita da:

$$\xi_0 = \gamma \qquad \text{e} \qquad \xi_{k+1} = \phi(\xi_k) \quad , \quad k = 0, 1, 2, \ldots$$

– Se ϕ non ha punti uniti in $[a,b]\cap M$ allora, detta $\Delta>0$ la minima distanza tra due elementi consecutivi di $[a,b]\cap M$, si ha:[28]

$$\text{per ogni } k:\quad |\,\phi(\xi_k)-\xi_k\,| = |\xi_{k+1}-\xi_k| \geqslant \Delta$$

e quindi:

$$\text{per ogni } k:\quad |\,\phi(\xi_k)\ominus\xi_k\,| \geqslant \mathrm{rd}(\Delta) > 0$$

– Se l'algoritmo $\psi:[a,b]\to M$ utilizzato per approssimare f non ha zeri in $[a,b]\cap M$ allora, detto $\Delta>0$ il minimo valore di $|\psi(\xi)|$ per $\xi\in[a,b]\cap M$ si ha:[29]

$$\text{per ogni } k:\quad |\,\psi(\xi_k)\,| \geqslant \Delta$$

e quindi:

$$\text{per ogni } k:\quad |\,\psi(\xi_k)\,| \geqslant \mathrm{rd}(\Delta) > 0$$

In entrambi i casi, scelto $0<\delta<\mathrm{rd}(\Delta)$ l'istruzione `MetodoUnPunto(`h,γ,f,δ`)` eseguita in *Scilab* *non* definisce un elemento $\xi\in M$ perchè il criterio d'arresto *non è efficace*.

I teoremi seguenti studiano la successione ξ_k e contengono informazioni riguardanti l'*efficacia* dei criteri d'arresto.

1.7.4 Teorema (uso del tipo *numero in virgola mobile e precisione finita* nei metodi ad un punto)
Si ha:

(A) per ogni $\xi\in[a,b]\cap M$:

$$|\,\xi-\alpha\,| > \frac{d_\phi}{1-L} \qquad\Rightarrow\qquad |\,\phi(\xi)-\alpha\,| < |\,\xi-\alpha\,|$$

(B) per ogni k si ha:

$$|\,\xi_k-x_k\,| \leqslant \frac{1-L^k}{1-L}\,d_\phi$$

(C) per ogni k si ha:

$$|\,\xi_k-\alpha\,| \leqslant \frac{1-L^k}{1-L}\,d_\phi + L^k|\,\xi_0-\alpha\,| = \frac{d_\phi}{1-L} + L^k\left(|\,\xi_0-\alpha\,| - \frac{d_\phi}{1-L}\right)$$

Dimostrazione. Per ogni $\xi\in[a,b]\cap M$ si ha, utilizzando l'uniforme accuratezza di ϕ:

$$|\,\phi(\xi)-\alpha\,| \leqslant |\,\phi(\xi)-h(\xi)\,| + |\,h(\xi)-h(\alpha)\,| \leqslant d_\phi + |\,h(\xi)-h(\alpha)\,|$$

Per il Teorema di Lagrange esiste un numero reale θ tra ξ ed α tale che:

$$|\,h(\xi)-h(\alpha)\,| = |\,h'(\theta)\,|\,|\,\xi-\alpha\,|$$

e quindi, essendo $\theta\in[a,b]$:

$$|\,h(\xi)-h(\alpha)\,| \leqslant L\,|\,\xi-\alpha\,|$$

Dunque:

$$|\,\phi(\xi)-\alpha\,| \leqslant d_\phi + L\,|\,\xi-\alpha\,|$$

Siccome:

$$|\,\xi-\alpha\,| > \frac{d_\phi}{1-L} \qquad\Rightarrow\qquad d_\phi < (1-L)|\,\xi-\alpha\,|$$

si ottiene l'asserto (A).

Si ha poi:

$$|\,\xi_k-x_k\,| = |\,\phi(\xi_{k-1})-h(x_{k-1})\,| \leqslant |\,\phi(\xi_{k-1})-h(\xi_{k-1})\,| + |\,h(\xi_{k-1})-h(x_{k-1})\,|$$

[28] La minima distanza tra due elementi consecutivi di $[a,b]\cap M$ è ben definita perché l'insieme $[a,b]\cap M$ è *finito*. Inoltre $\phi(\xi_k)-\xi_k\neq 0$ perché ϕ non ha punti uniti in $[a.b]\cap M$.

[29] Il minimo valore di $|\psi(\xi)|$ per $\xi\in[a,b]\cap M$ è ben definito perché l'insieme $[a,b]\cap M$ è *finito*; tale minimo è positivo perché ψ non ha zeri in $[a,b]\cap M$.

da cui, utilizzando ancora l'uniforme accuratezza di ϕ ed il Teorema di Lagrange:

$$|\,\xi_k - x_k\,| \leqslant d_\phi + L|\,\xi_{k-1} - x_{k-1}\,|$$

Iterando, e ricordando che $x_0 = \xi_0$ si ottiene:

$$|\,\xi_k - x_k\,| \leqslant \left(1 + L + \cdots + L^{k-1}\right) d_\phi = \frac{1 - L^k}{1 - L}\, d_\phi$$

ovvero l'asserto (B).

L'asserto (C) si ottiene immediatamente dall'asserto (B):

$$|\,\xi_k - \alpha\,| \leqslant |\,\xi_k - x_k\,| + |\,x_k - \alpha\,| \leqslant \frac{1 - L^k}{1 - L}\, d_\phi + L^k|\,\xi_0 - \alpha\,|$$

Il teorema è dimostrato.

1.7.5 Osservazione

Si osservi che:

- L'asserto (A) garantisce che *la successione delle distanze* $|\,\xi_k - \alpha\,|$ *è decrescente finché* ξ_k *non entra nell'intorno chiuso di centro* α *e raggio* $d_\phi/(1 - L)$*, dopodiché nulla si può dire.* In particolare: *non è garantita la convergenza della successione* ξ_k*.*
- L'asserto (B) afferma che *le successioni* ξ_k *ed* x_k *non sono mai troppo lontane.*
- L'asserto (C) traduce in termini di distanza di ξ_k da α quanto mostrato dall'asserto (A).

1.7.6 Teorema (uso del tipo *numero in virgola mobile e precisione finita*, continuazione)

Si ha anche:

(D) Se:

$$\textit{se } |\,\phi(\xi_k) \ominus \xi_k\,| < \mathrm{rd}(\delta) \textit{ allora} \text{ esci dal ciclo}$$

è la realizzazione del criterio d'arresto, si ha:

- Il criterio è *calcolabile.*
- Per decidere l'efficacia si studia la successione $|\,\phi(\xi_k) - \xi_k\,|$. Si ha:

$$|\,\phi(\xi_k) - \xi_k\,| \leqslant |\,\phi(\xi_k) - h(\xi_k)\,| + |\,h(\xi_k) - h(\xi_{k-1})\,| + |\,h(\xi_{k-1}) - \phi(\xi_{k-1})\,|$$

Utilizzando l'uniforme accuratezza dell'algoritmo ϕ ed il Teorema di Lagrange si ottiene:

$$|\,\phi(\xi_k) - \xi_k\,| \leqslant d_\phi + L\,|\,\phi(\xi_{k-1}) - \xi_{k-1}\,| + d_\phi = L\,|\,\phi(\xi_{k-1}) - \xi_{k-1}\,| + 2\,d_\phi$$

Allora:

(a) iterando all'indietro:

$$|\,\phi(\xi_k) - \xi_k\,| \leqslant L^k\,|\,\phi(\xi_0) - \xi_0\,| + 2\,\frac{1 - L^k}{1 - L}\,d_\phi = \frac{2\,d_\phi}{1 - L} + L^k\left(|\,\phi(\xi_0) - \xi_0\,| - \frac{2\,d_\phi}{1 - L}\right)$$

(b) se $|\,\phi(\xi_{k-1}) - \xi_{k-1}\,| > \dfrac{2\,d_\phi}{1 - L}$ allora $|\,\phi(\xi_k) - \xi_k\,| < |\,\phi(\xi_{k-1}) - \xi_{k-1}\,|$.[30]

Se ne deduce che *la successione* $|\,\phi(\xi_k) - \xi_k\,|$ *è decrescente finché:*

$$|\,\phi(\xi_k) - \xi_k\,| > \frac{2\,d_\phi}{1 - L}$$

dopodiché nulla si può dire. Dunque *il criterio può risultare non efficace.* In base ai risultati ottenuti, una *condizione sufficiente* per l'efficacia del criterio è:

$$\delta > \frac{2\,d_\phi}{1 - L}$$

[30] Infatti, posto $\Delta_k = |\phi(\xi_k) - \xi_k|$ si ha: $\Delta_{k-1} > 2\,d_\phi/(1 - L) \Rightarrow (1 - L)\,\Delta_{k-1} > 2\,d_\phi \Rightarrow \Delta_{k-1} > L\,\Delta_{k-1} + 2\,d_\phi$ e quindi: $\Delta_k \leqslant L\,\Delta_{k-1} + 2\,d_\phi < \Delta_{k-1}$.

– Sia k tale che:

$$|\,\phi(\xi_k) \ominus \xi_k\,| < \mathrm{rd}(\delta)$$

Allora, per il Teorema 1.7.1, ξ_k è punto unito di una funzione $h^* : [a,b] \to \mathbb{R}$ tale che:

$$\text{per ogni } x \in [a,b]: \quad |\,h^*(x) - h(x)\,| < \delta + d_\phi$$

Per quanto detto nella Sezione 1.6 sul condizionamento del calcolo di un punto unito si ottiene:

$$|\xi_k - \alpha| < \frac{\delta + d_\phi}{1 - h'(\alpha)}$$

risultato simile a quello ottenuto utilizzando il tipo *numero reale*, e quindi soggetto alle stesse critiche (vedere la Definizione 1.5.1 ed il successivo Esempio 1.5.2).

(E) Se:

$$\textit{se } |\,\psi(\xi_k)\,| < \mathrm{rd}(\delta) \textit{ allora} \text{ arresta la costruzione}$$

è la realizzazione del criterio d'arresto, si ha:

– Il criterio è *calcolabile*.

– Per decidere l'efficacia si studia la successione $|\,\psi(\xi_k)\,|$. Si riscrive:

$$\psi(\xi_k) = \big(\psi(\xi_k) - f(\xi_k)\big) + \big(f(\xi_k) - f(\alpha)\big)$$

Per il Teorema di Lagrange esiste θ_k tra ξ_k e α tale che:

$$\psi(\xi_k) = \big(\psi(\xi_k) - f(\xi_k)\big) + f'(\theta_k)(\xi_k - \alpha)$$

Utilizzando l'uniforme accuratezza dell'algoritmo ψ e quanto osservato riguardo all'asserto (C), detto M_1 il massimo valore di $|\,f'(x)\,|$ nell'intorno chiuso di centro α e raggio $d_\phi/(1-L)$, per k sufficientemente grande si ottiene:

$$|\,\psi(\xi_k)\,| \leqslant d_\psi + M_1 \frac{d_\phi}{1-L}$$

e *nulla si può dire sulla convergenza a zero della successione.* Dunque *il criterio può risultare non efficace.* In particolare, supponendo $M_1 \approx |\,f'(\alpha)\,|$, *non è ragionevole aspettarsi di ottenere:*

$$|\,\psi(\xi_k)\,| < d_\psi + |\,f'(\alpha)\,| \frac{d_\phi}{1-L}$$

È perciò opportuno che l'utilizzatore scelga:

$$\delta > d_\psi + |\,f'(\alpha)\,| \frac{d_\phi}{1-L}$$

– Sia k tale che:

$$|\,\psi(\xi_k)\,| < \mathrm{rd}(\delta)$$

Allora, per il Teorema 1.7.1, ξ_k è zero di una funzione $f^* : [a,b] \to \mathbb{R}$ tale che:

$$\text{per ogni } x \in [a,b]: \quad |\,f^*(x) - f(x)\,| < \delta + d_\psi$$

Per quanto detto nella Sezione 1.6 sul condizionamento del calcolo di uno zero si ottiene:

$$|\xi_k - \alpha| < \frac{\delta + d_\psi}{|\,f'(\alpha)\,|}$$

risultato simile a quello ottenuto utilizzando il tipo *numero reale*, e quindi soggetto alle stesse critiche (vedere la Definizione 1.5.1).

E21 ♠ Sia f la funzione definita da $f(x) = x + x^3$.

(1) Dimostrare che il metodo di Newton *è utilizzabile* per approssimare lo zero di f.

(2) Dimostrare che *non è possibile* utilizzare il criterio di scelta del valore iniziale per il metodo di Newton.

(3) Calcolare la funzione h che definisce il metodo di Newton applicato ad f e la derivata prima $h'(x)$. Utilizzare poi *Scilab* per disegnare il grafico di $h'(x)$ sull'intervallo $[-2, 2]$.

(4) Con il grafico disegnato al punto precedente determinare un intervallo che, insieme ad h, verifica le ipotesi (1) e (2) del Teorema di convergenza ed utilizzare poi il criterio di scelta del valore iniziale per metodi ad un punto.

E22 Sia $f(x) = e^x + x - 3$.

(1) Determinare il numero di zeri di f e separarli.

(2) Per ciascuno degli zeri di f decidere se il metodo ad un punto definito dalla funzione:

$$h(x) = 3 - e^x$$

sia utilizzabile per approssimare lo zero e, eventualmente, indicare un valore a partire dal quale la successione generata è convergente.

(3) Per ciascuno degli zeri di f decidere se il metodo di Newton sia utilizzabile per approssimare lo zero e, eventualmente, indicare un valore a partire dal quale la successione generata è convergente.

E23 Sia $f(x;t) = x^2 - (10 + \frac{1}{10} + t)x + 1$. Stimare lo scostamento dei rispettivi zeri di $f(x;0)$ e $f(x;0.1)$.

1.8 Appendice 1

Siano $h : [a, b] \to \mathbb{R}$ una funzione con derivata seconda continua ed $\alpha \in (a, b)$ un punto unito di h. Se $h'(\alpha) = 0$ il metodo ad un punto definito da h è utilizzabile per approssimare α. Sia allora x_k una successione generata dal metodo e convergente ad α. In questa Appendice si dimostra che se $h''(\alpha) \neq 0$ allora per ogni $\theta \in (0, 1)$ si ha:

$$\lim_{k\to\infty} \frac{|\, x_k - \alpha \,|}{\theta^k} = 0$$

Dalla relazione:

$$\lim_{k\to\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = |\tfrac{1}{2}h''(\alpha)|$$

posto:

$$\lambda = |\tfrac{1}{2}h''(\alpha)| \neq 0$$

si ha, per definizione di limite: per ogni numero reale $\epsilon \in (0, \lambda)$ esiste un numero intero positivo n tale che:

$$\text{per ogni } k \geqslant n \text{ si ha: } \quad (\lambda - \epsilon)|x_k - \alpha|^2 \leqslant |x_{k+1} - \alpha| \leqslant (\lambda + \epsilon)|x_k - \alpha|^2$$

Sia poi $m \geqslant n$ tale che:

$$(\lambda + \epsilon)|x_m - \alpha| < 1$$

(un tale m esiste certamente perchè $\lim |x_k - \alpha| = 0$). Iterando all'indietro, si ottiene, per ogni $k \geqslant m$:

$$(\lambda - \epsilon)^{2^{k-m}-1}|x_m - \alpha|^{2^{k-m}} \leqslant |x_k - \alpha| \leqslant (\lambda + \epsilon)^{2^{k-m}-1}|x_m - \alpha|^{2^{k-m}}$$

Queste disuguaglianze provano che la successione $|x_k - \alpha|$ tende a zero *più rapidamente* di $(\lambda+\epsilon)|x_m-\alpha|^{2^{k-m}}$ ma *meno rapidamente* di $(\lambda-\epsilon)|x_m-\alpha|^{2^{k-m}}$.

Per ogni $\theta \in (0,1)$, dalla seconda disuguaglianza si deduce:

$$\frac{|x_k-\alpha|}{\theta^k} \leqslant \frac{((\lambda+\epsilon)|x_m-\alpha|)^{2^{k-m}}}{(\lambda+\epsilon)\,\theta^k}$$

da cui, osservato che:

$$\log \frac{((\lambda+\epsilon)|x_m-\alpha|)^{2^{k-m}}}{\theta^k} = 2^{k-m}\log((\lambda+\epsilon)|x_n-\alpha|) - k\log\theta$$

e quindi (si ricordi che $(\lambda+\epsilon)|x_m-\alpha| \in (0,1)$):

$$\lim_{k\to\infty} 2^{k-m}\log((\lambda+\epsilon)|x_n-\alpha|) - k\log\theta = -\infty$$

si conclude:

$$\lim_{k\to\infty} \frac{((\lambda+\epsilon)|x_m-\alpha|)^{2^{k-m}}}{(\lambda+\epsilon)\,\theta^k} = 0$$

L'asserto è dimostrato.

1.9 Appendice 2

Siano $h : \mathbb{R} \to \mathbb{R}$ una funzione con derivata prima continua, α l'unico punto unito di h e $|\,h'(\alpha)\,| = 1$. Sia infine x_k una successione generata dal metodo iterativo definito da h. In questa appendice si dimostra che se $\lim x_k = \alpha$ e per ogni k si ha $x_k \neq \alpha$ allora:

$$\textit{per ogni } \theta \in (0,1) \textit{ si ha:} \qquad \lim \frac{|\,x_k-\alpha\,|}{\theta^k} = +\infty$$

Sia $x_0 \neq \alpha$. Per ogni j numero intero non negativo, il Teorema di Lagrange assicura l'esistenza di un numero reale t_j tra x_{j+1} e α tale che:

$$|\,x_{j+1}-\alpha\,| = |\,h'(t_j)\,|\,|\,x_j-\alpha\,|$$

Poichè $\lim x_k = \alpha$ si ha anche $\lim t_k = \alpha$ e quindi $\lim |\,h'(t_k)\,| = 1$. Scelto $\theta \in (0,1)$, sia n un numero intero tale che:

$$\text{per ogni } k \geqslant n: \quad |\,h'(t_k)\,| \geqslant \frac{1+\theta}{2}$$

Sia adesso $k > n$. Si ha:

$$\frac{|\,x_k-\alpha\,|}{\theta^k} = \frac{|\,h'(t_{k-1})\,|}{\theta} \cdots \frac{|\,h'(t_n)\,|}{\theta}\,\frac{|\,h'(t_{n-1})\,|}{\theta} \cdots \frac{|\,h'(t_0)\,|}{\theta}\,|\,x_0-\alpha\,|$$

Posto:

$$\Gamma = \frac{|\,h'(t_{n-1})\,|}{\theta} \cdots \frac{|\,h'(t_0)\,|}{\theta}\,|\,x_0-\alpha\,|$$

e constatato che per $k \geqslant n$ si ha:

$$\frac{|\,h'(t_k)\,|}{\theta} \geqslant \frac{1+\theta}{2\theta}$$

si ottiene:

$$\frac{|\,x_k-\alpha\,|}{\theta^k} \geqslant \Big(\frac{1+\theta}{2\theta}\Big)^{k-n}\Gamma$$

Ma:

$$\theta < 1 \quad \Rightarrow \quad \frac{1+\theta}{2\theta} > 1$$

dunque:

$$\lim \Big(\frac{1+\theta}{2\theta}\Big)^{k-n}\Gamma = +\infty$$

da cui segue l'asserto.

2 Sistemi di equazioni lineari

Siano $A \in \mathbb{R}^{n\times n}$ una matrice *invertibile*, b una colonna di $\mathbb{R}^n$ e x^* *l'unica* colonna di $\mathbb{R}^n$ soluzione del *sistema di equazioni lineari* $Ax = b$.

I metodi per determinare la soluzione di un sistema di equazioni lineri si suddividono in *diretti* e *iterativi*. Un metodo diretto determina la soluzione del sistema con un *numero finito* di operazioni elementari su numeri reali (operazioni aritmetiche e calcolo di radici quadrate). Un metodo iterativo determina con un numero finito di operazioni elementari su numeri reali *un elemento* di una successione che converge alla soluzione del sistema.

In questo Capitolo affrontiamo il problema di *determinare un'approssimazione accurata di* x^* utilizzando alcuni metodi diretti.

Si ricordi che l'asserto "la colonna y di $\mathbb{R}^n$ è soluzione del sistema $Ax = b$" significa che $Ay = b$ e che l'asserto "$A \in \mathbb{R}^{n\times n}$ è invertibile" significa che *esiste* $B \in \mathbb{R}^{n\times n}$ *tale che*: $AB = BA = I$, con $I \in \mathbb{R}^{n\times n}$ matrice identità di colonne $e_1, \ldots, e_n$. Proprietà *equivalenti* all'invertibilità di A sono:

* $\det A \neq 0$;
* $Ax = 0 \Rightarrow x = 0$, ovvero $\ker A = \{\, 0 \,\}$;
* Le colonne (righe) di A sono elementi linearmente indipendenti, dunque una *base*, di $\mathbb{R}^n$;
* *Per ogni* $b \in \mathbb{R}^n$ il sistema di equazioni $Ax = b$ ha *una sola soluzione*.

Se M è una matrice $n \times n$ e v una colonna di n numeri, indichiamo come usuale con m_{ij} e v_i $(i, j = 1, \ldots, n)$ gli elementi di M e quelli di v. Invece, se k è un numero intero e M_k una matrice $n \times n$, indichiamo i suoi elementi con la notazione $M_k(i, j)$.[31]

2.1 Casi semplici

Sia $A \in \mathbb{R}^{n\times n}$. Elenchiamo un insieme di casi particolari in cui la verifica dell'invertibilità di A ed il calcolo di x^* sono particolarmente *semplici*.

(D) A *diagonale*, ovvero: per ogni i, j si ha $i \neq j \Rightarrow a_{ij} = 0$.

La matrice è *invertibile se e solo se* $a_{kk} \neq 0, k = 1, \ldots, n$; una volta verificata l'invertibilità, per le componenti della soluzione si ha:

$$x^*_k = b_k/a_{kk} \quad , \quad k = 1, \ldots, n$$

Il numero di operazioni aritmetiche richiesto dal calcolo di x^* è: n (precisamente: n divisioni).

(T) A *triangolare*, ovvero: per ogni i, j si ha $i > j \Rightarrow a_{ij} = 0$ (triangolare *superiore*) oppure per ogni i, j si ha $i < j \Rightarrow a_{ij} = 0$ (triangolare *inferiore*).

Anche in questo caso la matrice è *invertibile se e solo se* $a_{kk} \neq 0, k = 1, \ldots, n$; una volta verificata l'invertibilità, se la matrice è triangolare superiore le componenti della soluzione si determinano con la procedura di *Sostituzione all'Indietro*:

$x = \mathrm{SI}\,(T, c)$

// T matrice $n \times n$ triangolare superiore invertibile, c colonna di n numeri reali;

// x verifica la relazione: $Tx = c$.

$x_n = c_n/t_{nn}$;

`per` $k = n-1, \ldots, 1$ `ripeti`:
    $s_k = c_k - (t_{k,k+1}x_{k+1} + \cdots + t_{kn}x_n)$;
    $x_k = s_k/t_{kk}$

Se la matrice è triangolare inferiore la soluzione si calcola con l'analoga procedura di *Sostituzione in Avanti* (Esercizio *E1*).

Il numero di operazioni aritmetiche richiesto dal calcolo di x^* è: n^2 (precisamente: n divisioni, $\frac{1}{2}\,n(n-1)$ moltiplicazioni ed altrettante somme).

[31] Questa notazione, poco usuale in matematica, è invece usuale in *Scilab*.

(O) A matrice *ortogonale*, ovvero che verifica una delle tre proprietà equivalenti:

- Le colonne (righe) di A sono una *base ortonormale* di $\mathbb{R}^n$ rispetto al prodotto scalare canonico ($a \cdot b = a_1b_1 + \cdots + a_nb_n = b^\mathsf{T}a$);
- $A^\mathsf{T}A = I$;
- A è invertibile e $A^{-1} = A^\mathsf{T}$.

La matrice è *certamente invertibile*; per la soluzione si ha: il sistema $Ax = b$ è equivalente al sistema $A^\mathsf{T}Ax = A^\mathsf{T}b$, a sua volta equivalente a: $x = A^\mathsf{T}b$, dunque: $x^* = A^\mathsf{T}b$.

Il numero di operazioni aritmetiche richiesto dal calcolo di x^* è quello richiesto dal prodotto di una matrice $n \times n$ per una colonna di n componenti: $2n^2 - n$ (precisamente: n^2 moltiplicazioni e $n(n-1)$ somme).

(P) A matrice *di permutazione*, ovvero le cui colonne (righe) sono *una permutazione* di quelle della matrice identità. Si osservi che, se A è una matrice di permutazione allora:

- Le colonne (righe) di A sono una *base ortonormale* di $\mathbb{R}^n$ rispetto al prodotto scalare canonico, dunque: *le matrici di permutazione sono particolari matrici ortogonali*;
- Se $v \in \mathbb{R}^n$ allora le componenti di Av si ottengono *permutando* quelle di v come indicato da A, in particolare il numero di operazioni aritmetiche richiesto dal calcolo di Av è *zero*;
- Anche A^T è di permutazione.

La matrice è *certamente invertibile*; per la soluzione si ha: il sistema $Ax = b$ è equivalente ai sistema $A^\mathsf{T}Ax = A^\mathsf{T}b$, a sua volta equivalente a: $x = A^\mathsf{T}b = x^*$, dunque: $x^* = A^\mathsf{T}b$.

Il numero di operazioni aritmetiche richiesto dal calcolo di x^* è quello richiesto dal prodotto di una matrice $n \times n$ *di permutazione* per una colonna di n componenti: *zero*.

Esercizi

E1 Descrivere la procedura di *Sostituzione in Avanti* di intestazione:

$$x = \mathrm{SA}(T, c)$$

che determina, dati una matrice $n \times n$ triangolare *inferiore* invertibile e una colonna c di n numeri reali, la colonna x che verifica: $Tx = c$. Verificare anche che il numero di operazioni aritmetiche richiesto dal calcolo di $x = \mathrm{SA}(T, c)$ è lo stesso di quello riportato per il calcolo della soluzione di un sistema nel caso di matrice triangolare superiore con la procedura SI.

E2 Sia $A \in \mathbb{R}^{n\times n}$. Verificare che: Le colonne di A sono una base ortonormale di $\mathbb{R}^n$ rispetto al prodotto scalare canonico *se e solo se* $A^\mathsf{T}A = I$.

E3 Sia:

$$v = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix}$$

Determinare la matrice di permutazione $P \in \mathbb{R}^{3\times 3}$ tale che:

$$Pv = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$$

E4 Sia $P_{23} \in \mathbb{R}^{3\times 3}$ la matrice di permutazione "che scambia la seconda e la terza riga," ovvero tale che per ogni r_1, r_2, r_3 in $\mathbb{R}^{1\times 3}$:

$$P_{23} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_3 \\ r_2 \end{bmatrix}$$

Verificare che per ogni c_1, c_2, c_3 in $\mathbb{R}^3$:

$$(c_1, c_2, c_3)P_{23}^\mathsf{T} = (c_1, c_3, c_2)$$

2.2 Caso generale

Sia $A \in \mathbb{R}^{n\times n}$ una matrice *non* diagonale, triangolare, ortogonale o di permutazione. Per verificare se A è invertibile ed eventualmente calcolare la soluzione del sistema $Ax = b$ si procede come segue:

– *Passo 1*:

Si fattorizza A in (si scrive A come) prodotto di fattori $F_1, \ldots, F_j$ *semplici*, ovvero ciascuno appartenente ad una delle categorie diagonale, triangolare, ortogonale, di permutazione, e si verifica l'invertibilità di A controllando (facilmente) l'invertibilità di *ciascuno* dei fattori.

– *Passo 2*:

Se qualcuno dei fattori $F_1, \ldots, F_j$ risulta non invertibile (e quindi A risulta non invertibile) si rinuncia a calcolare la soluzione del sistema, altrimenti si calcolano (facilmente):

(1) la soluzione c_1 del sistema $F_1 x = b$;
(2) la soluzione c_2 del sistema $F_2 x = c_1$;
$\vdots$
(j) la soluzione c_j del sistema $F_j x = c_{j-1}$.

Poiché:

$$Ac_j = F_1 \cdots F_{j-1}(F_j c_j) \overset{(j)}{=} F_1 \cdots F_{j-2}(F_{j-1}c_{j-1}) \overset{(j-1)}{=} \cdots \overset{(2)}{=} F_1 c_1 \overset{(1)}{=} b$$

dall'unicità della soluzione del sistema $Ax = b$ si ottiene $c_j = x^*$. Dunque, per determinare la soluzione del sistema $Ax = b$ si risolvono tanti sistemi *semplici* quanti sono i fattori di A ottenuti nel Passo 1.

Resta da descrivere come determinare una fattorizzazione di A in prodotto di fattori semplici. Ci limiteremo a discutere le due fattorizzazioni più comunemente usate nel contesto della soluzione dei sistemi di equazioni lineari: la *fattorizzazione LR con pivoting* e la *fattorizzazione QR*, definite nell'asserto seguente.

2.2.1 Definizione (fattorizzazioni LR, LR con pivoting e QR di una matrice quadrata)
Sia $A \in \mathbb{R}^{n\times n}$.

Una *fattorizzazione LR* di A è *una coppia* di matrici $S, D \in \mathbb{R}^{n\times n}$ tali che:

∗ $A = SD$;

∗ Il fattore sinistro S è *triangolare inferiore con* $s_{kk} = 1, k = 1, \ldots, n$;

∗ Il fattore destro D è *triangolare superiore*.

Una *fattorizzazione LR con pivoting* di A è *una terna* di matrici $S, D, P \in \mathbb{R}^{n\times n}$ tali che:

∗ La matrice P è *di permutazione*;

∗ La coppia S, D è una fattorizzazione LR di PA.

In particolare sussiste la fattorizzazione:

$$A = P^{\mathsf{T}} SD$$

Una *fattorizzazione QR* di A è *una coppia* di matrici $U, T \in \mathbb{R}^{n\times n}$ tali che:

∗ $A = UT$;

∗ Il fattore sinistro U è *ortogonale*;

∗ Il fattore destro T è *triangolare superiore*.

Si osservi che le tre fattorizzazioni riducono l'invertibilità di A a quella del solo *fattore destro* (D per le fattorizzazioni LR e LR con pivoting, T per la fattorizzazione QR).

Nella prossima Sezione si mostra come una fattorizzazione LR con pivoting possa essere determinata rileggendo opportunamente la procedura di *eliminazione di Gauss*. Analogamente, mostreremo più avanti come una fattorizzazione QR possa essere determinata rileggendo opportunamente la procedura di *ortonormalizzazione di Gram-Schmidt*.

2.3 Fattorizzazione LR con pivoting: la procedura EGP

Assegnata una matrice $A \in \mathbb{R}^{n\times n}$, la procedura EGP (*Eliminazione di Gauss con Pivoting*), di intestazione:

$$(S, D, P) = \mathrm{EGP}(A)$$

determina una fattorizzazione LR con pivoting di A.

La fattorizzazione determinata consente (a) di verificare l'invertibilità di A constatando se $d_{11} \neq 0, \ldots, d_{nn} \neq 0$ ed eventualmente (b) di determinare la soluzione del sistema $Ax = b$ calcolando $c = \mathrm{SA}(S, Pb)$ e poi $x^* = \mathrm{SI}(D, c)$.

Prima di dare una descrizione della procedura, introduciamo la nozione di *matrice elementare di Gauss*.

2.3.1 Definizione (matrice elementare di Gauss)

Una matrice $n \times n$ ad elementi reali si chiama *matrice elementare di Gauss* se è ottenuta dalla matrice identità I scegliendo un indice j in $1, \ldots, n-1$, numeri reali $\lambda_{j+1,j}, \ldots, \lambda_{nj}$ ed operando in I la sostituzione:

$$e_j = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \longleftarrow \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \lambda_{j+1,j} \\ \vdots \\ \lambda_{nj} \end{bmatrix}$$

Si osservi che una matrice elementare di Gauss è dunque una particolare matrice *triangolare inferiore con elementi uguali a uno sulla diagonale*, dunque *invertibile*.

2.3.2 Esempio

Siano λ_{21} e λ_{31} numeri reali. La matrice:

$$H = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 1 & 0 \\ \lambda_{31} & 0 & 1 \end{bmatrix}$$

è elementare di Gauss (ottenuta dalla matrice identità sostituendo la prima colonna con...). Sia poi:

$$A = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} \in \mathbb{R}^{3\times 3}$$

Costruendo il prodotto *per righe* si constata che:

$$HA = \begin{bmatrix} r_1 \\ \lambda_{21} r_1 + r_2 \\ \lambda_{31} r_1 + r_3 \end{bmatrix}$$

Inoltre si verifica che:

$$H^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -\lambda_{21} & 1 & 0 \\ -\lambda_{31} & 0 & 1 \end{bmatrix}$$

In generale: Se H è una matrice elementare di Gauss, l'inversa H^{-1} si ottiene ha H *cambiando segno agli elementi al di sotto della diagonale*. Anche la matrice H^{-1} è elementare di Gauss, in particolare è triangolare inferiore con elementi uguali a uno sulla diagonale.

La procedura EGP opera come segue:

- pone $A_1 = A$;
- per $k = 1, \ldots, n-1$ determina *opportunamente* $P_k \in \mathbb{R}^{n\times n}$ di permutazione e $H_k \in \mathbb{R}^{n\times n}$ elementare di Gauss e pone:

$$A_{k+1} = H_k P_k A_k$$

$*$ pone $D = A_n$, $P = P_{n-1} \cdots P_1$ e $S = P(P_1^{-1} H_1^{-1} \cdots P_{n-1}^{-1} H_{n-1}^{-1})$.

Le matrici di permutazione P_k ed elementari di Gauss H_k sono determinate in modo che la matrice D risulti *triangolare superiore* e la matrice S risulti *triangolare inferiore con con elementi uguali a uno sulla diagonale.*

Si osservi che al termine della procedura si ha:

$$D = H_{n-1} P_{n-1} \cdots H_1 P_1 \, A$$

da cui, essendo ciascuno dei fattori P_k e H_k *invertibile*:

$$A = P_1^{-1} H_1^{-1} \cdots P_{n-1}^{-1} H_{n-1}^{-1} \, D$$

La matrice:

$$\Sigma = P_1^{-1} H_1^{-1} \cdots P_{n-1}^{-1} H_{n-1}^{-1}$$

non è, in generale, triangolare inferiore (la coppia Σ, D è una fattorizzazione di A ma *non* di tipo LR) *ma* la matrice:

$$S = P\Sigma$$

è triangolare inferiore con uno sulla diagonale e $SD = P(\Sigma D) = PA$. Dunque la *terna* di matrici S, D, P è una fattorizzazione LR con pivoting della matrice A.

Restano da discutere due punti: (i) come la procedura determina le matrici di permutazione P_k ed elementari di Gauss H_k e (ii) come mai Σ non è triangolare inferiore e $P\Sigma$ è triangolare inferiore con elementi uguali a uno sulla diagonale. Illustreremo questi punti descrivendo dettagliatamente il comportamento della procedura in due esempi.

2.3.3 Esempio

Sia:

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

La procedura opera così:

$*$ Pone $A_1 = A$;

$*$ Pone $k = 1$.

- Constata che $A_1(1,1) \neq 0$ e pone di conseguenza:

$$P_1 = I \quad \text{e} \quad T_1 = P_1 A_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 2 & 1 & 0 \\ -2 & 0 & 0 & -1 \\ -1 & 1 & 2 & -1 \end{bmatrix}$$

  Così facendo si ha $T_1(1,1) \neq 0$.

- Considera la matrice elementare di Gauss:

$$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 1 & 0 & 0 \\ \lambda_{31} & 0 & 1 & 0 \\ \lambda_{41} & 0 & 0 & 1 \end{bmatrix}$$

  e cerca valori di $\lambda_{21}, \lambda_{31}$ e λ_{41} tali che gli elementi di posto $(2,1)$, $(3,1)$ e $(4,1)$ della matrice $H_1 T_1$ siano *zero*. Le tre condizioni equivalgono alle equazioni:

$$\lambda_{j1} \, T_1(1,1) + T_1(j,1) = 0 \quad \text{per } j = 2,3,4$$

  Poiché $T_1(1,1) \neq 0$ le equazioni determinano, ciascuna, *un solo valore* di λ_{j1}:

$$\lambda_{21} = -\frac{T_1(2,1)}{T_1(1,1)} = -2 \quad , \quad \lambda_{31} = -\frac{T_1(3,1)}{T_1(1,1)} = 2 \quad , \quad \lambda_{41} = -\frac{T_1(4,1)}{T_1(1,1)} = 1$$

- Con i valori trovati costruisce:

$$A_2 = H_1 T_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 2 & 2 & -1 \end{bmatrix}$$

  Si osservi che la prima riga di A_2 è copia della prima riga di T_1.

$*$ Pone $k = 2$.

- Constata che $A_2(2,2) = 0$ e cerca $j > 2$ tale che $A_2(j,2) \neq 0$. Constatato che $A_2(3,2) \neq 0$, indicata con P_{23} la matrice di permutazione che *scambia le righe* 2 *e* 3, pone di conseguenza:

$$P_2 = P_{23} \quad \text{e} \quad T_2 = P_2 A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 2 & -1 \end{bmatrix}$$

  Così facendo si mantengono gli zeri ottenuti al passo precedente (in magenta) e si ha $T_2(2,2) \neq 0$.

- Considera la matrice elementare di Gauss:

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{32} & 1 & 0 \\ 0 & \lambda_{42} & 0 & 1 \end{bmatrix}$$

  e cerca valori di λ_{32} e λ_{42} tali che gli elementi di posto $(3,2)$ e $(4,2)$ della matrice $H_2 T_2$ siano *zero*. Le due condizioni equivalgono alle equazioni:

$$\lambda_{j2}\, T_2(2,2) + T_2(j,2) = 0 \quad \text{per } j = 3, 4$$

  Poiché $T_2(2,2) \neq 0$ le equazioni determinano, ciascuna, *un solo valore* di λ_{j2}:

$$\lambda_{32} = -\frac{T_2(3,2)}{T_2(2,2)} = 0 \quad , \quad \lambda_{42} = -\frac{T_2(4,2)}{T_2(2,2)} = -1$$

- Con i valori trovati costruisce:

$$A_3 = H_2 T_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

  Si noti che la scelta di H_2 *mantiene* i tre zeri ottenuti al passo precedente (in blu) e le prime *due* righe di A_3 sono copia delle prime due righe di T_2.

$*$ Pone $k = 3$.

- Constata che $A_3(3,3) \neq 0$ e pone di conseguenza:

$$P_3 = I \quad \text{e} \quad T_3 = P_3 A_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

  Così facendo si mantengono gli zeri ottenuti al passo precedente (in magenta) e si ha $T_3(3,3) \neq 0$.

- Considera la matrice elementare di Gauss:

$$H_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{43} & 1 \end{bmatrix}$$

  e cerca un valore di λ_{43} tale che l'elemento di posto $(4,3)$ della matrice H_3T_3 sia *zero*. La condizione equivale all'equazione:

$$\lambda_{43}\, T_3(3,3) + T_3(4,3) = 0$$

  Poiché $T_3(3,3) \neq 0$ l'equazione determina *un solo valore* di λ_{43}:

$$\lambda_{43} = -\frac{T_3(4,3)}{T_3(3,3)} = 0$$

- Con il valore trovato costruisce:

$$A_4 = H_3T_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = D$$

  Si noti che la scelta di H_3 *mantiene* i gli zeri ottenuti ai passi precedenti (in blu) e le prime *tre* righe di A_4 sono copia delle prime tre righe di T_3.

I valori $T_1(1,1), T_2(2,2)$ e $T_3(3,3)$ che la procedura utilizza come *divisori* per determinare i vari elementi λ_{ij}, e che *ritroviamo sulla diagonale della matrice finale* D, si chiamano *pivot*. La tecnica utilizzata per determinare le matrici P_k (e quindi i pivot) si chiama *pivoting*.

Come preannunciato, la matrice $\Sigma = H_1^{-1}P_2^{-1}H_2^{-1}H_3^{-1}$ *non è* triangolare inferiore:

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

ma, posto $P = P_3P_2P_1 = P_2$ si ha invece:

$$S = P\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -1 & 1 & 2 & 1 \end{bmatrix}$$

che *è* triangolare inferiore con elementi uguali a uno sulla diagonale. Per capire come ciò accada si osservi che:

$$P\Sigma = P_2H_1^{-1}P_2^{-1}H_2^{-1}H_3^{-1}$$

e che:

$$P_2H_1^{-1}P_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \equiv H_1^{-1}(2)$$

è triangolare inferiore con elementi uguali a uno sulla diagonale. La matrice $H_1^{-1}(2)$ è *il risultato dell'azione della permutazione* P_2 *sulle righe e colonne di* H_1^{-1}.

Più in generale, se:

$$P = P_3P_2P_1 \quad \text{e} \quad \Sigma = P_1^{-1}H_1^{-1}P_2^{-1}H_2^{-1}P_3^{-1}H_3^{-1}$$

allora:

$$P\Sigma = P_3P_2P_1P_1^{-1}H_1^{-1}P_2^{-1}H_2^{-1}P_3^{-1}H_3^{-1} = P_3\big(P_2H_1^{-1}P_2^{-1}\big)H_2^{-1}P_3^{-1}H_3^{-1}$$

e, con la notazione introdotta sopra:

$$P\Sigma = P_3 H_1^{-1}(2) H_2^{-1} P_3^{-1} H_3^{-1}$$

Adesso, ricordando che $P_3^{-1}P_3 = I$, si riscrive:

$$P\Sigma = \left(P_3 H_1^{-1}(2) P_3^{-1}\right)\left(P_3 H_2^{-1} P_3^{-1}\right) H_3^{-1} = H_1^{-1}(2,3) H_2^{-1}(3) H_3^{-1}$$

Le matrici $H_1^{-1}(2,3)$, $H_2^{-1}(3)$ e H_3^{-1} sono triangolari inferiori con uno sulla diagonale e tale è il loro prodotto.

2.3.4 Esempio

Sia:

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

La procedura opera così:

* Pone $A_1 = A$;
* Pone $k = 1$.
  - Constata che $A_1(1,1) \neq 0$ e pone di conseguenza:

    $$P_1 = I \quad \text{e} \quad T_1 = P_1 A_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 2 & 1 & 0 \\ -2 & -2 & 0 & -1 \\ -1 & -1 & 2 & -1 \end{bmatrix}$$

    Così facendo si ha $T_1(1,1) \neq 0$.
  - Considera la matrice elementare di Gauss:

    $$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 1 & 0 & 0 \\ \lambda_{31} & 0 & 1 & 0 \\ \lambda_{41} & 0 & 0 & 1 \end{bmatrix}$$

    e cerca valori di $\lambda_{21}, \lambda_{31}$ e λ_{41} tali che gli elementi di posto $(2,1)$, $(3,1)$ e $(4,1)$ della matrice $H_1 T_1$ siano *zero*. Le tre condizioni equivalgono alle equazioni:

    $$\lambda_{j1}\, T_1(1,1) + T_1(j,1) = 0 \quad \text{per } j = 2,3,4$$

    Poiché $T_1(1,1) \neq 0$ le equazioni determinano, ciascuna, *un solo valore* di λ_{j1}:

    $$\lambda_{21} = -\frac{T_1(2,1)}{T_1(1,1)} = -2 \quad , \quad \lambda_{31} = -\frac{T_1(3,1)}{T_1(1,1)} = 2 \quad , \quad \lambda_{41} = -\frac{T_1(4,1)}{T_1(1,1)} = 1$$

  - Con i valori trovati costruisce:

    $$A_2 = H_1 T_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 2 & -1 \end{bmatrix}$$

    Si osservi che la prima riga di A_2 è copia della prima riga di T_1.
* Pone $k = 2$.
  - Constata che $A_2(2,2) = A_2(3,2) = A_2(4,2) = 0$ (non esiste $j > 2$ tale che $A_2(j,2) \neq 0$). Pone di conseguenza: $P_2 = I$ e $H_2 = I$ da cui:

    $$A_3 = P_2 H_2 A_2 = A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 2 & -1 \end{bmatrix}$$

    Si noti che la scelta di H_2 *mantiene* i tre zeri ottenuti al passo precedente (in blu) e le prime *due* righe di A_3 sono copia delle prime due righe di T_2.

∗ Pone $k = 3$.

- Constata che $A_3(3,3) = 0$ e cerca $j > 3$ tale che $A_3(j,3) \neq 0$. Constata che $A_3(4,3) \neq 0$ e pone di conseguenza:

$$P_3 = P_{34} \quad \text{e} \quad T_3 = P_3 A_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

  Così facendo si mantengono gli zeri ottenuti al passo precedente (in magenta) e si ha $T_3(3,3) \neq 0$.

- Considera la matrice elementare di Gauss:

$$H_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{43} & 1 \end{bmatrix}$$

  e cerca un valore di λ_{43} tale che l'elemento di posto $(4,3)$ della matrice $H_3 T_3$ sia *zero*. La condizione equivale all'equazione:

$$\lambda_{43}\, T_3(3,3) + T_3(4,3) = 0$$

  Poiché $T_3(3,3) \neq 0$ l'equazione determina *un solo valore* di λ_{43}:

$$\lambda_{43} = -\frac{T_3(4,3)}{T_3(3,3)} = 0$$

- Con il valore trovato costruisce:

$$A_4 = H_3 T_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix} = D$$

  Si noti che la scelta di H_3 *mantiene* i gli zeri ottenuti ai passi precedenti (in blu) e le prime *tre* righe di A_4 sono copia delle prime tre righe di T_3.

I pivot, in questo caso, sono i valori $T_1(1,1)$ e $T_3(3,3)$ che *ritroviamo sulla diagonale della matrice finale* D.

Anche in questo caso la matrice $\Sigma = H_1^{-1} P_3^{-1}$ *non è* triangolare inferiore:

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

ma, posto $P = P_3 P_2 P_1 = P_3$ si ha invece:

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

che *è* triangolare inferiore con elementi uguali a uno sulla diagonale. Procedendo come nell'esempio precedente si osserva che:

$$S = P\Sigma = P_3 H_1^{-1} P_3^{-1} \equiv H_1^{-1}(3)$$

è triangolare inferiore con uno sulla diagonale ed è *il risultato dell'azione della permutazione* P_3 *sulle righe e colonne di* H_1^{-1}.

2.3.5 Esempio (uso della procedura EGP)

Siano: $\mathrm{EGP}(A) = (S, D, P)$ con:

$$S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad , \quad D = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \quad , \quad P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

e:

$$b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

In questo esempio si mostra come utilizzare la fattorizzazione determinata da EGP per calcolare $\det A$, risolvere il sistema $Ax = b$ e calcolare A^{-1}.

Si ha:

$$\det A = \det(P^{-1}SD) = \det P^{\mathsf{T}} \det S \det D = (-1) \cdot 1 \cdot (-2) = 2$$

La matrice A è quindi invertibile e *la* soluzione del sistema si determina come segue: (i) Si calcola la soluzione del sistema $Sx = Pb$ con la procedura SA:

$$c = \mathrm{SA}(S, Pb) = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

e poi (ii) Si calcola la soluzione x^* del sistema $Ax = b$ risolvendo con la procedura SI il sistema $Dx = c$:

$$x^* = \mathrm{SI}(D, c) = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Infine, siano $e_1, \ldots, e_n$ le colonne della matrice identità. Per definizione, la k-esima colonna di $A^{-1} = (y_1, \ldots, y_n) \in \mathbb{R}^{n \times n}$ verifica la relazione:

$$Ay_k = e_k$$

ovvero è *la* soluzione del sistema $Ax = e_k$ e si calcola come mostrato nel punto precedente:

$$c_k = \mathrm{SA}(S, Pe_k) \quad , \quad y_k = \mathrm{SI}(D, c_k)$$

Risolvendo $2n$ sistemi con matrice triangolare si ottiene:

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

Esercizi

E5 Siano:

$$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 1 & 0 & 0 \\ \lambda_{31} & 0 & 1 & 0 \\ \lambda_{41} & 0 & 0 & 1 \end{bmatrix} \quad \text{e} \quad H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{32} & 1 & 0 \\ 0 & \lambda_{42} & 0 & 1 \end{bmatrix}$$

Calcolare H_1H_2 *per colonne* e verificare che il numero di operazioni aritmetiche richiesto per *costruire* H_1H_2 è *zero*.

E6 Siano $A, B \in \mathbb{R}^{4\times4}$ matrici triangolari inferiori con uno sulla diagonale. Costruire il prodotto AB *per righe* e verificare che a sua volta è una matrice triangolare inferiore con uno sulla diagonale.

E7 Siano $A, B \in \mathbb{R}^{4\times 4}$ matrici triangolari superiori. Costruire il prodotto AB *per colonne* e verificare che a sua volta è una matrice triangolare superiore.

E8 Siano:

$$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 1 & 0 & 0 \\ \lambda_{31} & 0 & 1 & 0 \\ \lambda_{41} & 0 & 0 & 1 \end{bmatrix}$$

e $P_2 = P_{24}, P_3 = P_{34}$. Calcolare $H_1^{-1}(2,3)$ e constatare che è triangolare inferiore con uno sulla diagonale.

E9 Siano:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 2 & 1 & 0 \\ -2 & 0 & 0 & -1 \\ -1 & 1 & 2 & 0 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Applicare la procedura EGP ed utilizzare la fattorizzazione ottenuta per lo studio del sistema di equazioni lineari $Ax = b$.

E10 Siano S, D, P come nell'Esempio 2.3.5 ed $A = P^{\mathsf{T}}SD$. Al primo passo della procedura EGP applicata ad A è necessario scegliere se scambiare la prima riga con la seconda o con la terza. Constatare che le due scelte portano a valori *diversi* delle matrici prodotte da EGP.

2.4 Norme di vettori e matrici

La procedura EGP consente di ricercare la soluzione x^* del sistema $Ax = b$ con il seguente procedimento descritto in un linguaggio che consente l'uso del tipo *numero reale*:

$(S, D, P) = \mathrm{EGP}(A)$;
`se` esiste k tale che $d_{kk} = 0$ `allora` arresta il procedimento e dichiara A non invertibile;
  `altrimenti`
    $c = \mathrm{SA}(S, Pb)$;
    $x^* = \mathrm{SI}(D, c)$

La discussione dell'uso del calcolatore per eseguire il procedimento richiede di sostituire al tipo *numero reale* il tipo *numero in virgola mobile e precisione finita*. Questa sostituzione consiste di *due* passaggi nel primo dei quali si sostituiscono le costanti a valore in $\mathbb{R}$ con gli arrotondati in M. Tra le costanti presenti nel procedimento vi sono *i dati* che individuano il sistema da studiare: la matrice A e la colonna b. La sostituzione *cambia* la matrice A di elementi a_{ij} nella matrice A' di elementi $\mathrm{rd}(a_{ij})$ e la colonna b di elementi b_i nella colonna b' di elementi $\mathrm{rd}(b_i)$. Dunque la sostituzione di tipo *cambia il sistema in esame*: il calcolatore decide dell'invertibilità di A' ed eventualmente determina un'approssimazione ξ della soluzione $\hat{x}$ del sistema $A'x = b'$. Supponendo che ξ sia un'approssimazione accurata di $\hat{x}$, occorre chiedersi se essa risulti *anche* un'approssimazione accurata di x^*. Quest'ultima condizione è l'oggetto dello studio del *condizionamento del calcolo di* x^* che consiste appunto nel determinare quanto *lontano* può essere $\hat{x}$ da x^* rispetto alla *distanza* di A' da A e di b' da b.

Allo studio del condizionamento premettiamo alcune nozioni riguardanti la *norma* di vettori e matrici.

2.4.1 Definizione (norma, spazio normato)
Sia V uno spazio vettoriale su $\mathbb{R}$. Una funzione $N : V \to \mathbb{R}$ si dice *norma* in V se ha le tre proprietà seguenti:

(1) Per ogni $v \in V$: $N(v) \geqslant 0$ e $N(v) = 0 \Rightarrow v = 0$

(2) Per ogni $v \in V$ e $\alpha \in \mathbb{R}$: $N(\alpha v) = |\,\alpha\,|\, N(v)$

(3) Per ogni $v, w \in V$: $N(v + w) \leqslant N(v) + N(w)$ (disuguaglianza triangolare)

Il numero reale $N(v)$ si chiama *norma di* v e la coppia (V, N) si chiama *spazio normato*.

2.4.2 Esempio

(1) Sia V lo spazio vettoriale su $\mathbb{R}$ dei vettori del piano. La funzione che a v associa *la lunghezza del segmento orientato che rappresenta* v verifica le proprietà richieste dalla definizione di norma.

(2) Sia $V = \mathbb{R}^n$. Le funzioni $N_1, N_2, N_\infty : V \to \mathbb{R}$ definite da:

$$N_1(v) = |\, v_1 \,| + \cdots + |\, v_n \,| \equiv \| v \|_1$$

$$N_2(v) = \sqrt{v_1^2 + \cdots + v_n^2} \equiv \| v \|_2$$

$$N_\infty(v) = \max\{\, |\, v_1 \,|, \ldots, |\, v_n \,|\,\} \equiv \| v \|_\infty$$

verificano la definizione di norma (N_2 è la usuale *norma euclidea*).

2.4.3 Definizione (distanza tra vettori)

Con la nozione di norma è possibile introdurre quella di *distanza*: se (V, N) è uno spazio normato, per ogni $a, b \in V$ il numero reale $N(v - w)$ si chiama *distanza* tra a e b.

2.4.4 Esercizio

Si considerino i seguenti elementi di $\mathbb{R}^2$:

$$a = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad , \quad c = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Decidere quale tra b e c è più lontano da a utilizzando, per misurare la distanza tra elementi di $\mathbb{R}^2$, prima la norma N_1 poi la norma N_∞.

2.4.5 Definizione (intorno sferico)

Siano $v \in \mathbb{R}^n$ e r un numero reale non negativo. Si chiama *intorno sferico* (chiuso) di *centro* v e *raggio* r l'insieme:

$$I(v, r) = \{\, x \in \mathbb{R}^n : N(x - v) \leqslant r \,\}$$

ovvero l'insieme degli elementi di $\mathbb{R}^n$ che distano da v non più di r.

2.4.6 Esempio

Nella Figura 2 è rappresentato il bordo degli intorni sferici di centro $v = 0 \in \mathbb{R}^2$ e raggio $r = 1$ nei casi $N = N_2, N_1$ e N_∞.

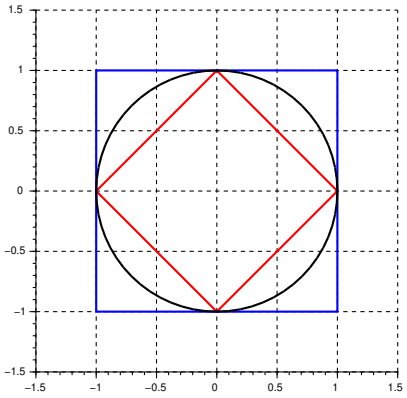


Figura 2: $I(0, 1)$ con N_2 (in nero), N_1 (in rosso) e N_∞ (in blu).

2.4.7 Definizione (norma di matrice)

Siano N una norma in $\mathbb{R}^n$ e $A \in \mathbb{R}^{n \times n}$. La *norma di* A *indotta da* N è:

$$\| A \|_N = \sup \left\{ \frac{N(Av)}{N(v)}, v \neq 0 \right\}$$

2.4.8 Esempio

In $\mathbb{R}^n$ con norma N si ha: $\| I \|_N = 1, \| 0_{n\times n} \| = 0$.

2.4.9 Osservazione (sulla definizione di norma di matrice)

(1) Per ogni $v \neq 0$ si ha:

$$\frac{N(Av)}{N(v)} = N\left(\frac{Av}{N(v)}\right) = N\left(A\,\frac{v}{N(v)}\right)$$

e quindi:

$$\left\{ \frac{N(Av)}{N(v)}, v \neq 0 \right\} = \{\, N(Av), N(v) = 1 \,\}$$

L'insieme B dei vettori v tali che $N(v) = 1$ è *chiuso e limitato* e la funzione $F : B \to \mathbb{R}$ definita da $F(v) = N(Av)$ è *continua*. Allora, per il Teorema di Weierstrass, F ha *massimo* e *minimo*, ovvero: esistono v^* e v_* tali che:

$$N(Av^*) = \max\{\, N(Av), N(v) = 1 \,\} \quad \text{e} \quad N(Av_*) = \min\{\, N(Av), N(v) = 1 \,\}$$

Allora:

$$\| A \|_N = \max\{\, N(Av), N(v) = 1 \,\}$$

(2) Se A è invertibile si ha:

$$\| A^{-1} \|_N = (\min\{\, N(Av), N(v) = 1 \,\})^{-1}$$

Infatti:

$$\| A^{-1} \| = \sup \left\{ \frac{N(A^{-1}v)}{N(v)}, v \neq 0 \right\}$$

ovvero, posto $w = A^{-1}v$ e osservato che essendo A^{-1} invertibile si ha $v \neq 0 \Leftrightarrow w \neq 0$:

$$\| A^{-1} \| = \sup \left\{ \frac{N(w)}{N(Aw)}, w \neq 0 \right\}$$

Adesso si osservi che se $\Omega \subset \mathbb{R}$ si ha:

$$\sup \Omega = (\, \inf\{\, 1/x, x \in \Omega \,\}\,)^{-1}$$

dunque:

$$\| A^{-1} \| = \left(\inf \left\{ \frac{N(Aw)}{N(w)}, w \neq 0 \right\} \right)^{-1}$$

(3) Dai risultati precedenti si deduce che, posto:

$$C = \{\, Av, N(v) = 1 \,\}$$

si ha: *la norma di A è il minimo valore di r tale che $C \subset I(0, r)$* e *la norma di A^{-1} è il massimo valore di r tale che*[32] $C \subset \mathbb{R}^n \setminus I^\circ(0, r)$.

2.4.10 Osservazione (formule di calcolo)

Il *calcolo* di $\| A \|$ per N generica è proibitivo. Nei *casi particolari* $N = N_1, N_2$ e N_∞ si ha, dette $a_1, \ldots, a_n$ le colonne di A:

$$\| A \|_1 = \max\{\, N_1(a_1), \ldots, N_1(a_n) \,\}$$

$$\| A \|_2 = \sqrt{\max\{\, \text{autovalori di } A^{\mathsf{T}} A \,\}}$$

$$\| A \|_\infty = \| A^{\mathsf{T}} \|_1$$

2.4.11 Osservazione (Proprietà della norma indotta)

Siano N una norma in $\mathbb{R}^n$ ed $A \in \mathbb{R}^{n\times n}$. Allora:

[32] Con $I^\circ(v, r)$ si indica l'intorno sferico *aperto* di centro $v \in \mathbb{R}^n$ e raggio r, ovvero: $\{\, x \in \mathbb{R}^n : N(x - v) < r \,\}$.

(i) Per ogni elemento v di $\mathbb{R}^n$ si ha: $N(Av) \leqslant \| A \|_N N(v)$;

(ii) Esiste un elemento non nullo w di $\mathbb{R}^n$ tale che: $N(Aw) = \| A \|_N N(w)$. In particolare esiste $w \in \mathbb{R}^n$ con $N(w) = 1$ tale che $N(Aw) = \| A \|_N$.

(L'asserto (i) segue dalla definizione di norma di matrice, quello (ii) da quanto osservato nel punto (1) dell'Osservazione 2.4.9.)

Sia poi $B \in \mathbb{R}^{n\times n}$. Allora $AB \in \mathbb{R}^{n\times n}$ e:

(iii) $\| AB \|_N \leqslant \| A \|_N \| B \|_N$.

(Infatti: Sia $v^* \in \mathbb{R}^n$ con $N(v^*) = 1$ tale che $N(ABv^*) = \| AB \|_N$. Utilizzando due volte la proprietà (i) si ottiene: $\| AB \|_N = N(ABv^*) \leqslant \| A \|_N N(Bv^*) \leqslant \| A \|_N \| B \|_N$.)

2.4.12 Osservazione

L'insieme $\mathbb{R}^{n\times n}$ con le usuali definizioni di somma e multiplo è uno spazio vettoriale su $\mathbb{R}$. Si ha:

(1) Se N è una norma in $\mathbb{R}^n$ allora la funzione $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ definita da $f(A) = \| A \|_N$ è una norma in $\mathbb{R}^{n\times n}$ e per ogni $A, B \in \mathbb{R}^{n\times n}$ il numero reale $\| A - B \|_N$ si chiama *distanza* tra A e B.

(2) Lo spazio vettoriale $\mathbb{R}^{n\times n}$ è *isomorfo* allo spazio vettoriale $\mathbb{R}^{n^2}$. La corrispondenza che realizza l'isomorfismo è quella che alla matrice A di colonne $a_1, \ldots, a_n$ associa, prendendo a prestito la notazione da *Scilab*, il vettore $a = [a_1; \ldots; a_n]$. Ciascuna delle funzioni $f_1, f_2, f_\infty : \mathbb{R}^{n\times n} \to \mathbb{R}$ definite da:

* $f_1(A) = N_1(a) = \Sigma_{i,j=1}^n | a_{ij} |$
* $f_2(A) = N_2(a) = \sqrt{\Sigma_{i,j=1}^n | a_{ij} |^2}$ (detta anche *norma di Frobenius* di A)
* $f_\infty(A) = N_\infty(a) = \max\{ | a_{ij} | \text{ con: } i, j = 1, \ldots, n \}$

è una *norma* in $\mathbb{R}^{n\times n}$.

(3) Siano $A \in \mathbb{R}^{n\times n}$ e $v \in \mathbb{R}^n$. Si ha:

$$\| Av \|_2 \leqslant f_2(A) \| v \|_2$$

(Infatti, dette $r_1, \ldots, r_n$ le righe di A ed omettendo il pedice alla norma due:

$$\| Av \| = \sqrt{| r_1 v |^2 + \cdots + | r_n v |^2}$$

Utilizzando la *disuguaglianza di Schwarz* si ha:

$$\sqrt{| r_1 v |^2 + \cdots + | r_n v |^2} \leqslant \sqrt{\| r_1 \|^2 \| v \|^2 + \cdots + \| r_n \|^2 \| v \|^2}$$

ed infine:

$$\sqrt{\| r_1 \|^2 \| v \|^2 + \cdots + \| r_n \|^2 \| v \|^2} = \sqrt{(\| r_1 \|^2 + \cdots + \| r_n \|^2) \| v \|^2}$$

da cui l'asserto.)

Esercizi

E11 In Figura 3, ottenuta utilizzando *Google Maps*, sono riportate due porzioni della cartina stradale di Manhattan. Quale funzione tra N_1, N_2 e N_∞ è utilizzata per misurare le distanze nei due casi?

E12 Verificare che le funzioni N_1 ed N_∞ sono norme in $\mathbb{R}^n$ secondo la Definizione 2.4.1.

E13 Siano N una norma in $\mathbb{R}^n$ e $\alpha \in \mathbb{R}$. Utilizzare la Definizione 2.4.7 per calcolare $\| \alpha I \|_N$.

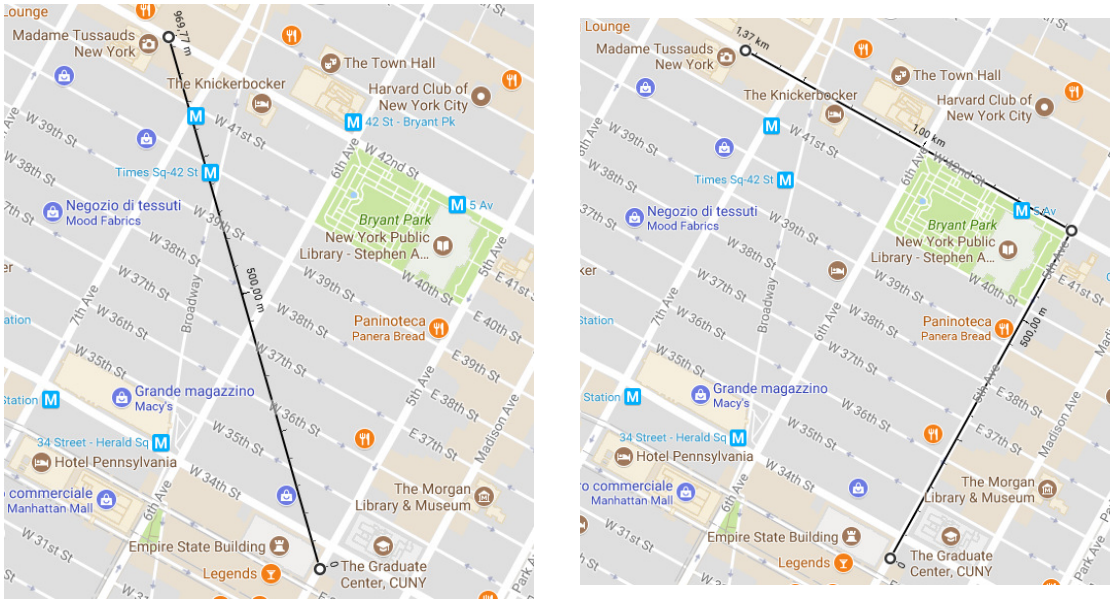


Figura 3: Distanze tra l'*Empire State Building* ed il museo delle cere *Madame Tussauds*.

E14 ★ Sia N una norma in $\mathbb{R}^n$. Dimostrare che: (i) *se* $A \in \mathbb{R}^{n\times n}$ e $\| A \|_N = 0$ *allora* $A = 0_{n\times n}$. (ii) per ogni $A, B \in \mathbb{R}^{n\times n}$ si ha: $\| A + B \|_N \leqslant \| A \|_N + \| B \|_N$.

(Suggerimento per il punto (ii): si consideri $w \in \mathbb{R}^n$ tale che $N(w) = 1$ e $\| A + B \|_N = N\big((A + B)w\big)$.)

E15 Si considerino le funzioni f_1 e f_2 definite nell'Osservazione 2.4.12. Calcolare $f_1(I)$ e $f_2(I)$ e dedurre dal risultato che f_1 e f_2 sono norme *non indotte*.

E16 Si consideri la funzione f_∞ definita nell'Osservazione 2.4.12 e siano:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{e} \quad B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Calcolare $f_\infty(A)$, $f_\infty(B)$ e $f_\infty(AB)$. Dedurre dal risultato che f_∞ è una norma *non indotta*.

E17 ★ Sia:

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

Determinare $\| A \|_\infty$ e $v^* \in \mathbb{R}^3$ tale che $N_\infty(v^*) = 1$ e $N_\infty(Av^*) = \| A \|_\infty$.

2.5 Condizionamento del calcolo della soluzione di un sistema

Siano $A \in \mathbb{R}^{n\times n}$ una matrice *invertibile*, $b \in \mathbb{R}^n$ una colonna *non nulla* e si consideri il sistema di equazioni lineari $Ax = b$. Il sistema ha una sola soluzione: l'unico vettore $x^* \in \mathbb{R}^n$ (che risulta non nullo) tale che $Ax^* = b$. Siano poi $A' \in \mathbb{R}^{n\times n}$ una matrice *invertibile*, $b' \in \mathbb{R}^n$ e si consideri il sistema di equazioni lineari $A'x = b'$. Anche questo sistema ha una sola soluzione: l'unico vettore $\hat{x} \in \mathbb{R}^n$ tale che $A'\hat{x} = b'$.

2.5.1 Definizione (perturbazioni, scostamento e loro misure relative)

La matrice $\delta A = A' - A$ si chiama *perturbazione* del dato A, la colonna $\delta b = b' - b$ si chiama *perturbazione* del dato b e la colonna $\delta x = \hat{x} - x^*$ si chiama *scostamento* della soluzione x^*.

La matrice A' e la colonna b' si chiamano *dati perturbati* ed il sistema $A'x = b'$ si chiama *sistema perturbato*.

Scelta una norma in $\mathbb{R}^n$, le misure *relative* di δA, δb e δx sono, rispettivamente:

$$\epsilon_A = \frac{\| \delta A \|}{\| A \|} \quad , \quad \epsilon_b = \frac{\| \delta b \|}{\| b \|} \qquad \epsilon_x = \frac{\| \delta x \|}{\| x^* \|}$$

Si osservi che tutte le quantità sono ben definite perché in ciascuna il denominatore è certamente diverso da zero.

Lo studio del condizionamento del calcolo della soluzione del sistema $Ax = b$ consiste nel considerare delle *piccole perturbazioni* dei dati A, b e discutere quanto grande *può essere* ϵ_x rispetto a quanto grandi *sono* ϵ_A e ϵ_b.

2.5.2 Osservazione (sull'ipotesi di invertibilità di A')

Perché il sistema perturbato abbia una sola soluzione si è posta l'ipotesi che la matrice perturbata A' sia invertibile. A tal proposito si ha: Se ϵ_A è *sufficientemente piccolo* allora la matrice perturbata $A + \delta A$ è certamente invertibile (asserto (3) dell'Osservazione 2.5.6).

L'ipotesi fatta è quindi *ragionevole* purché si considerino perturbazioni *piccole*, come usuale nel contesto dello studio del condizionamento.

Si analizzano prima due casi particolari poi il caso generale.

- $\delta A = 0$, $\delta b \neq 0$

  Per lo scostamento δx si ha:

  $$\delta x = \hat{x} - x^* = A^{-1}(b + \delta b) - A^{-1}b = A^{-1}\delta b$$

  da cui, per il punto (i) dell'Osservazione 2.4.11:

  $$\| \delta x \| = \| A^{-1}\delta b \| \leqslant \| A^{-1} \| \, \| \delta b \|$$

  In termini di misura relativa si ottiene:

  $$\epsilon_x = \frac{\| \delta x \|}{\| x^* \|} \leqslant \frac{\| A^{-1} \| \, \| \delta b \|}{\| x^* \|}$$

  Ricordando che x^* è la soluzione del sistema $Ax = b$ si ha:

  $$\| b \| = \| Ax^* \| \leqslant \| A \| \, \| x^* \| \qquad \text{ovvero} \qquad \frac{1}{\| x^* \|} \leqslant \frac{\| A \|}{\| b \|}$$

  Infine, sostituendo:

  $$\epsilon_x \leqslant \| A^{-1} \| \, \| A \| \, \frac{\| \delta b \|}{\| b \|} = \| A^{-1} \| \, \| A \| \, \epsilon_b$$

  Si osservi che per il punto (ii) dell'Osservazione 2.4.11 esistono una colonna $w \neq 0$ tale che:

  $$\| A^{-1}w \| = \| A^{-1} \| \, \| w \|$$

  ed una colonna $y \neq 0$ tale che:

  $$\| Ay \| = \| A \| \, \| y \|$$

  Allora, per $\delta b = w$ e $x^* = y$ (ovvero $b = Ay \neq 0$) si ha:

  $$\epsilon_x = \frac{\| A^{-1}\delta b \|}{\| x^* \|} = \frac{\| A^{-1}w \|}{\| y \|} = \frac{\| A^{-1} \| \, \| w \|}{\| y \|} = \| A^{-1} \| \, \| A \| \, \frac{\| w \|}{\| b \|} = \| A^{-1} \| \, \| A \| \, \epsilon_b$$

  Introdotto il *numero di condizionamento* di A:

  $$c(A) = \| A \| \, \| A^{-1} \|$$

  il risultato ottenuto si riscrive nella forma:

2.5.3 Teorema (di condizionamento per $\delta A = 0$)

Sia $A \in \mathbb{R}^{n \times n}$ invertibile. Allora:

(i) Per ogni vettore non nullo $b \in \mathbb{R}^n$ ed ogni $\delta b \in \mathbb{R}^n$ si ha:

$$\epsilon_x \leqslant c(A)\,\epsilon_b$$

(ii) Esistono un vettore δb ed un vettore non nullo b tali che:

$$\epsilon_x = c(A)\,\epsilon_b$$

- $\delta A \neq 0$, $\delta b = 0$

  Si ricordi che si considerano solo perturbazioni δA tali che $A + \delta A$ invertibile. Si ha:

$$(A + \delta A)\hat{x} = b = Ax^*$$

da cui:

$$A\,\delta x = A(\hat{x} - x^*) = -\delta A\,\hat{x}$$

Per lo scostamento δx si ottiene allora:

$$\delta x = -A^{-1}\delta A\,\hat{x}$$

e quindi per il punto (i) dell'Osservazione 2.4.11:

$$\| \delta x \| = \| A^{-1}\delta A\,\hat{x}\| \leqslant \| A^{-1}\delta A \|\,\| \hat{x} \|$$

Per punto (iii) dell'Osservazione 2.4.11 si ha poi:

$$\| A^{-1}\delta A \| \leqslant \| A^{-1} \|\,\| \delta A \|$$

Introducendo come misura relativa dello scostamento la quantità (si osservi che $\| \hat{x} \| \neq 0$ perché $b \neq 0$):

$$\hat{\epsilon}_x = \frac{\| \delta x \|}{\| \hat{x} \|}$$

si ottiene:

$$\hat{\epsilon}_x \leqslant \| A^{-1} \|\,\| \delta A \| = \| A^{-1} \|\,\| A \|\,\frac{\| \delta A \|}{\| A \|} = \| A^{-1} \|\,\| A \|\,\epsilon_A$$

Si osservi che per il punto (ii) dell'Osservazione 2.4.11 per ogni δA esiste una colonna $w \neq 0$ tale che:

$$\| A^{-1}\delta A\,w \| = \| A^{-1}\delta A \|\,\| w \|$$

Inoltre, per qualche matrice Z tale che $A + Z$ invertibile (ad esempio per $Z = \alpha\,I$ con $\alpha \in \mathbb{R}$ sufficientemente piccolo) si ha:

$$\| A^{-1}Z \| = \| A^{-1} \|\,\| Z \|$$

Allora, per $\delta A = Z$ e $\hat{x} = w$ (e quindi per $b = (A + Z)w$) si ha:

$$\| \delta x \| = \| A^{-1}\delta A\,\hat{x}\| = \| A^{-1}Z\,w\| = \| A^{-1}Z \|\,\| w\| = \| A^{-1} \|\,\| Z \|\,\| w\| = \| A^{-1} \|\,\| \delta A \|\| \hat{x} \|$$

e quindi:

$$\| \hat{\epsilon}_x \| = \| A^{-1} \|\,\| \delta A \| = \| A^{-1} \|\,\| A \|\,\epsilon_A$$

Con le notazioni introdotte si riscrive il risultato ottenuto nella forma:

2.5.4 Teorema (di condizionamento per $\delta b = 0$)

Sia $A \in \mathbb{R}^{n\times n}$ invertibile. Allora:

(i) Per ogni vettore non nullo $b \in \mathbb{R}^n$ ed ogni $\delta A \in \mathbb{R}^{n\times n}$ tale che $A + \delta A$ invertibile si ha:

$$\hat{\epsilon}_x \leqslant c(A)\,\epsilon_A$$

(ii) Esistono una matrice δA tale che $A + \delta A$ invertibile ed un vettore non nullo b tali che:

$$\hat{\epsilon}_x = c(A)\,\epsilon_A$$

- $\delta A \neq 0$ e $\delta b \neq 0$

  Dati $\delta b \in \mathbb{R}^n$ tale che $b + \delta b$ è non nullo e $\delta A \in \mathbb{R}^{n\times n}$ tale che $A + \delta A$ è invertibile, siano: $\hat{x}$ la soluzione del sistema perturbato $(A + \delta A)x = b + \delta b$, $\hat{x}_b$ la soluzione del sistema perturbato $Ax = b + \delta b$ e x^* la soluzione del sistema $Ax = b$. Allora, posto:

$$\epsilon_A = \frac{\| \delta A \|}{\| A \|} \quad , \quad \epsilon_b = \frac{\| \delta b \|}{\| b \|} \quad \text{e} \quad \epsilon_x = \frac{\| \hat{x} - x^* \|}{\| x^* \|}$$

  si ha:

  (1) Per quanto mostrato nei casi particolari:

$$\frac{\| \hat{x} - \hat{x}_b \|}{\| \hat{x} \|} \leqslant c(A)\, \epsilon_A \qquad \text{e} \qquad \frac{\| \hat{x}_b - x^* \|}{\| x^* \|} \leqslant c(A)\, \epsilon_b$$

  (2) $\epsilon_x = \dfrac{\| \hat{x} - x^* \|}{\| x^* \|} \leqslant \dfrac{\| \hat{x} - \hat{x}_b \|}{\| x^* \|} + \dfrac{\| \hat{x}_b - x^* \|}{\| x^* \|} = \dfrac{\| \hat{x} - \hat{x}_b \|}{\| \hat{x} \|} \dfrac{\| \hat{x} \|}{\| x^* \|} + \dfrac{\| \hat{x}_b - x^* \|}{\| x^* \|}$

  (3) $\dfrac{\| \hat{x} \|}{\| x^* \|} \leqslant \dfrac{\| \hat{x} - x^* \|}{\| x^* \|} + 1 = \epsilon_x + 1$

  Quindi:

$$\epsilon_x \leqslant c(A)\, \epsilon_A\, (\epsilon_x + 1) + c(A)\, \epsilon_b$$

  ovvero:

$$\big(1 - c(A)\, \epsilon_A\big)\, \epsilon_x \leqslant c(A)\, (\epsilon_A + \epsilon_b)$$

  Si ottiene infine:

2.5.5 Teorema (di condizionamento)

Sia $A \in \mathbb{R}^{n\times n}$ invertibile. Allora: per ogni vettore non nullo $b \in \mathbb{R}^n$, ogni $\delta b \in \mathbb{R}^n$ tale che $b + \delta b$ è non nullo e ogni $\delta A \in \mathbb{R}^{n\times n}$ tale che $A + \delta A$ è invertibile e $c(A)\, \epsilon_A < 1$ si ha:

$$\epsilon_x \leqslant \frac{c(A)}{1 - c(A)\, \epsilon_A}\, (\epsilon_A + \epsilon_b)$$

2.5.6 Osservazione

(1) Ponendo $\delta b = 0$ nell'asserto del Teorema di condizionamento, si ottiene la seguente *versione alternativa* del Teorema di condizionamento per $\delta b = 0$:

Per ogni vettore non nullo $b \in \mathbb{R}^n$ e ogni $\delta A \in \mathbb{R}^{n\times n}$ tale che $A + \delta A$ invertibile e $c(A)\, \epsilon_A < 1$ si ha:

$$\epsilon_x \leqslant \frac{c(A)\, \epsilon_A}{1 - c(A)\, \epsilon_A}$$

(2) Sia N una norma in $\mathbb{R}^n$ ed $A \in \mathbb{R}^{n\times n}$ invertibile. Allora: $c(A) \geqslant 1$.
(Infatti: $I = A^{-1}A$ e quindi $1 = \| I \|_N = \| A^{-1}A \|_N \leqslant \| A^{-1} \|_N \| A \|_N = c(A)$.)

(3) Siano $A \in \mathbb{R}^{n\times n}$ invertibile e $\delta A \in \mathbb{R}^{n\times n}$. Se $c(A)\, \epsilon_A < 1$ allora $A + \delta A$ invertibile.
(Infatti: $c(A)\, \epsilon_A = \| A^{-1} \| \| \delta A \|$ e quindi per l'ipotesi:

$$\| A^{-1}\delta A \| \leqslant \| A^{-1} \| \| \delta A \| < 1$$

Inoltre: $A + \delta A = A\,(I + A^{-1}\delta A)$, dunque:

$$A + \delta A \text{ invertibile} \quad \Leftrightarrow \quad I + A^{-1}\delta A \text{ invertibile}$$

Infine: Se per qualche $v \neq 0$ si ha $(I + A^{-1}\delta A)v = 0$ allora si ha anche: $v = -A^{-1}\delta A\, v$ e quindi $N(v) = N(A^{-1}\delta A\, v) \leqslant \| A^{-1}\delta A \|\, N(v)$, ovvero $\| A^{-1}\delta A \| \geqslant 1$.)

Dunque:

$$\epsilon_A < \frac{1}{c(A)} \quad \Rightarrow \quad A + \delta A \text{ invertibile}$$

(4) Sia $x^* \in \mathbb{R}^n$ un vettore non nullo e $\delta x \in \mathbb{R}^n$. Si consideri, per ogni k tale che $x_k^* \neq 0$, la misura relativa della *componente k-esima* dello scostamento:

$$\frac{|\,\delta x_k\,|}{|\,x_k^*\,|}$$

Poiché per ogni vettore $y \in \mathbb{R}^n$ ed ogni k si ha $|\,y_k\,| \leqslant \|\,y\,\|$, allora:

$$\frac{|\,\delta x_k\,|}{|\,x_k^*\,|} \leqslant \frac{\|\,\delta x\,\|}{|\,x_k^*\,|} = \frac{\|\,\delta x\,\|}{\|\,x^*\,\|}\,\frac{\|\,x^*\,\|}{|\,x_k^*\,|} = \epsilon_x\,\frac{\|\,x^*\,\|}{|\,x_k^*\,|}$$

con:

$$\frac{\|\,x^*\,\|}{|\,x_k^*\,|} \geqslant 1$$

Dunque: *se la componente k-esima del vettore x^* è molto vicina a zero, la misura relativa della componente k-esima dello scostamento può essere molto maggiore della misura relativa del vettore scostamento* (vedere l'Esercizio *E19*).

(5) Scelti $M = F(\beta, m)$ e la funzione arrotondamento rd, siano: $A \in \mathbb{R}^{n\times n}$ di elementi a_{ij} una matrice invertibile, $b \in \mathbb{R}^n$ di elementi b_i una colonna non nulla, A' la matrice di elementi $\mathrm{rd}(a_{ij})$, b' la colonna di elementi $\mathrm{rd}(b_i)$, $\delta A = A' - A$ e $\delta b = b' - b$. Scelta una norma in $\mathbb{R}^n$ tra N_1, N_2 e N_∞ e detta u la precisione di macchina in M si ha:

$$\epsilon_b \leqslant u \qquad \text{e} \qquad \epsilon_A \leqslant u$$

Se:

$$c(A)\,u < 1$$

allora:

(5.a) Per quanto mostrato nel punto (2), la matrice A' è invertibile;

(5.b) Dette x^* e $\hat{x}$ rispettivamente la soluzione del sistema $Ax = b$ e la soluzione del sistema $A'x = b'$ per il Teorema di condizionamento si ha:

$$\epsilon_x \leqslant 2\,\frac{c(A)\,u}{1 - c(A)\,u}$$

2.5.7 Osservazione (applicazione del Teorema di condizionamento)

Dati $A \in \mathbb{R}^{n\times n}$ invertibile, b elemento non nullo di $\mathbb{R}^n$ e $\hat{x} \in \mathbb{R}^n$, si utilizza $\hat{x}$ per approssimare la soluzione x^* del sistema $Ax = b$. Per ottenere informazioni sull'accuratezza dell'approssimazione, si introduce il vettore:

$$r = A\hat{x} - b$$

detto *residuo* di $Ax = b$ associato ad $\hat{x}$.

(1) Si consideri la seguente *interpretazione* di $\hat{x}$:

$$\hat{x} \text{ è la soluzione del } \textit{sistema perturbato } Ax = b + r$$

Per il Teorema di condizionamento con $\delta A = 0$: posto $\delta b = r$ si ha:

$$\frac{\|\,\hat{x} - x^*\,\|}{\|\,x^*\,\|} \leqslant c(A)\,\frac{\|\,r\,\|}{\|\,b\,\|}$$

ovvero si ottiene *una limitazione dell'errore relativo commesso approssimando x^* con $\hat{x}$.*

(2) Siano $\hat{x} \neq 0$ e $M \in \mathbb{R}^{n\times n}$ tale che $M\hat{x} = -r$.

(La condizione $\hat{x} \neq 0$ è sufficiente a garantire l'esistenza di matrici M tali che $M\hat{x} = -r$. Ad esempio:

$$M = -\frac{r\hat{x}^{\mathsf{T}}}{\hat{x}^{\mathsf{T}}\hat{x}}$$

Se $\hat{x} = 0$, invece, esistono matrici con la proprietà richiesta se e solo se anche $r = 0$.)

Se $A + M$ invertibile, si consideri la seguente *interpretazione* di $\hat{x}$:

$$\hat{x} \text{ è la soluzione del } \textit{sistema perturbato } (A + M)x = b$$

Posto $\delta A = M$ e:

$$\alpha = c(A)\,\frac{\| M \|}{\| A \|}$$

la versione alternativa del Teorema di condizionamento con $\delta b = 0$ (punto (1) dell'Osservazione 2.5.6) consente di dedurre che se $\alpha < 1$ allora:

$$\frac{\| \hat{x} - x^* \|}{\| x^* \|} \leqslant \frac{\alpha}{1 - \alpha}$$

ovvero si ottiene *una limitazione dell'errore relativo commesso approssimando x^* con $\hat{x}$.*

2.5.8 Esempio

Si consideri $\mathbb{R}^2$ con norma N_1 e siano:

$$A = \begin{bmatrix} 20 & 1 \\ 0 & 20 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 10 \\ 10 \end{bmatrix} \quad , \quad \hat{x} = \tfrac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Si approssima la soluzione x^* del sistema $Ax = b$ con $\hat{x}$. Per l'accuratezza dell'approssimazione si ha:

(1) Dopo aver calcolato A^{-1} si ottiene:

$$c(A) = \| A^{-1} \| \| A \| = \tfrac{21}{400}\, 21 = \tfrac{441}{400} \approx 1$$

Inoltre il *residuo* di $Ax = b$ associato ad $\hat{x}$ vale:

$$r = A\hat{x} - b = \tfrac{1}{2} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(2) Si interpreta $\hat{x}$ come soluzione del sistema $Ax = b + r$. La misura relativa della perturbazione è:

$$\epsilon_b = \frac{\| r \|}{\| b \|} = \tfrac{1}{40}$$

e, utilizzando il Teorema di condizionamento con $\delta A = 0$:

$$\epsilon_x \leqslant c(A)\,\epsilon_b = \tfrac{441}{400}\,\tfrac{1}{40} \equiv \alpha_1 \approx 2.76 \cdot 10^{-2}$$

(3) Posto:

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

si osserva che $A + \delta A$ è invertibile e si interpreta $\hat{x}$ come soluzione del sistema $(A + \delta A)x = b$. La misura relativa della perturbazione è:

$$\epsilon_A = \frac{\| \delta A \|}{\| A \|} = \tfrac{1}{21}$$

e risulta:

$$\alpha_2 \equiv c(A)\,\epsilon_A = \tfrac{441}{400}\,\tfrac{1}{21} = \tfrac{21}{400} < 1$$

Allora, per la versione alternativa del Teorema di condizionamento con $\delta b = 0$:

$$\epsilon_x \leqslant \frac{\alpha_2}{1 - \alpha_2} \approx 5.54 \cdot 10^{-2}$$

La limitazione ottenuta nel secondo caso è *peggiore* di quella ottenuta nel primo (infatti: $5.54 \cdot 10^{-2} > 2.76 \cdot 10^{-2}$). Se ne conclude che, utilizzando la norma N_1, l'errore relativo commesso approssimando x^* con $\hat{x}$ non supera $\alpha_1 \approx 2.76 \cdot 10^{-2}$.

E18 Si consideri $\mathbb{R}^2$ con norma N_1 e sia:

$$b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Disegnare su un piano cartesiano l'insieme di tutti i vettori b' ottenuti sommando a b le perturbazioni δb tali che $\epsilon_b \leqslant \frac{1}{10}$.

E19 Si consideri $\mathbb{R}^3$ con norma N_∞ e siano:

$$x^* = \begin{bmatrix} 1 \\ 10^{-2} \\ 0 \end{bmatrix} \quad , \quad \delta x = 10^{-4} \begin{bmatrix} 1 \\ 10 \\ -1 \end{bmatrix}$$

Determinare la misura relativa del vettore scostamento ϵ_x e, per ogni k tale che $x^*_k \neq 0$, la misura relativa della componente k-esima dello scostamento.

E20 Siano $M = F(2, 53)$ e A, b, A', b' come nel punto (5) dell'Osservazione 2.5.6. Utilizzare la limitazione mostrata nel punto (5.b) per ottenere una condizione sufficiente su $c(A)$ in modo che sia $\epsilon_x < 10^{-6}$.

E21 ★ Siano $M = F(2, 53)$ e $b, b', \delta b$ come nel punto (5) dell'Osservazione 2.5.6. Utilizzare il Teorema 0.27 del Capitolo 0 per dimostrare che, indicando con u la precisione di macchina, per $N = N_1, N_2$ e N_∞ si ha: $N(\delta b) \leqslant u\, N(b)$, ovvero $\epsilon_b \leqslant u$.

E22 ★ Dimostrare che:

$$\epsilon_x < 1 \quad \Rightarrow \quad \hat{\epsilon}_x \leqslant \frac{\epsilon_x}{1 - \epsilon_x}$$

E23 Nell'esempio finale si è ottenuto:

$$\epsilon_x \leqslant \alpha_1 \approx 2.76 \cdot 10^{-2}$$

(1) Utilizzare il Teorema di condizionamento con $\delta b = 0$ per ottenere:

$$\hat{\epsilon}_x \leqslant \alpha_2 = 5.25 \cdot 10^{-2}$$

(2) Dedurre dalla limitazione su $\hat{\epsilon}_x$ che:

$$\| \delta x \|_1 \leqslant \alpha_2$$

(3) Utilizzare il risultato dell'Esercizio precedente per ottenere, dalla limitazione su ϵ_x, una limitazione su $\hat{\epsilon}_x$ e dedurne una nuova limitazione su $\| \delta x \|_1$.

(4) Rappresentare su un piano cartesiano i vettori δx che verificano le limitazioni trovate in (2) e (3) e dedurne un insieme che certamente contiene l'effettivo vettore δx.

E24 Determinare la soluzione del sistema dell'Esempio 2.5.8 e controllare che le limitazioni trovate sono soddisfatte.

E25 Siano $\hat{x}$ e r come nell'Esempio 2.5.8. Determinare:

$$M = -\frac{r\hat{x}^\mathsf{T}}{\hat{x}^\mathsf{T}\hat{x}}$$

e verificare che $M\hat{x} = -r$. Posto poi:

$$\epsilon_A = \frac{\| M \|}{\| A \|}$$

verificare che:

$$\alpha_3 = c(A)\, \epsilon_A < 1$$

Dunque $A + M$ è invertibile e $\hat{x}$ è l'unica soluzione del sistema perturbato $(A + M)x = b$. Utilizzare il Teorema di condizionamento per ottenere una limitazione dell'errore relativo commesso approssimando x^* con $\hat{x}$ e confrontare la limitazione ottenuta con quelle già ricavate.

2.6 Uso del tipo *numero in virgola mobile*: la procedura EGPP

Si ricordi che, assegnata una matrice $A \in \mathbb{R}^{n\times n}$ ed una colonna $b \in \mathbb{R}^n$, la procedura EGP permette la ricerca della soluzione della soluzione del sistema $Ax = b$ con il seguente procedimento descritto con un linguaggio che consente l'uso del tipo *numero reale*:

$(S, D, P) = \mathrm{EGP}(A)$;
`se` esiste k tale che $d_{kk} = 0$ `allora` arresta il procedimento e dichiara A non invertibile;
  `altrimenti`
    $c = \mathrm{SA}(S, Pb)$;
    $x^* = \mathrm{SI}(D, c)$

Sostituendo al tipo *numero reale* il tipo *numero in virgola mobile e precisione finita* il procedimento si modifica come segue:

$(\hat{S}, \hat{D}, \hat{P}) = \mathtt{EGP}(A')$;
`se` esiste k tale che $\hat{d}_{kk} = 0$ `allora` arresta il procedimento e dichiara A' non invertibile;
  `altrimenti`
    $\hat{c} = \mathtt{SA}(\hat{S}, \hat{P}b')$;
    $\xi = \mathtt{SI}(\hat{D}, \hat{c})$

dove: A' e b' indicano, rispettivamente, la matrice e la colonna ottenute arrotondando in M ciascuna componente di A e b ed `EGP`, `SA` ed `SI` indicano le procedure ottenute sostituendo il tipo *numero reale* rispettivamente nelle procedure EGP, SA ed SI.

Il vettore finale ξ è utilizzato per approssimare x^*. Le osservazioni seguenti mostrano che l'approssimazione è *potenzialmente non accurata* ma una piccola modifica della procedura EGP la rende invece *quasi sempre* accurata quanto consentito dal condizionamento di A.

2.6.1 Osservazione (interpretazione di `SI`)
Siano $M = F(\beta, m)$, $T \in \mathbb{R}^{2\times 2}$ una matrice triangolare superiore invertibile e $c \in \mathbb{R}^2$. La procedura SI determina la soluzione del sistema:

$$\begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

calcolando:

(1) $x_2 = c_2/t_{22}$;

(2) $s_1 = c_1 - t_{12}x_2$ e poi $x_1 = s_1/t_{11}$ ovvero: $x_1 = (c_1 - t_{12}x_2)/t_{11}$.

La procedura `SI`, *supponendo che gli elementi di T e c siano elementi di M*, calcola:

(1) $\xi_2 = c_2 \oslash t_{22}$;

(2) $\sigma_1 = c_1 \ominus (t_{12} \otimes \xi_2)$ e poi $\xi_1 = \sigma_1 \oslash t_{11}$.

Ricordando la definizione di pseudo-operazioni aritmetiche, per il Teorema 0.2.13 si ha: esistono numeri reali $e_1, \ldots, e_4$ ciascuno di valore assoluto non superiore alla precisione di macchina u tali che:

$$\xi_2 = (1 + e_1)\, c_2/t_{22} \quad , \quad \sigma_1 = (1 + e_3)\left(c_1 - (1 + e_2)\, t_{12}\xi_2\right) \quad \text{e} \quad \xi_1 = (1 + e_4)\, \sigma_1/t_{11}$$

Posto poi:

$$t'_{22} = t_{22}/(1 + e_1) \neq 0 \quad , \quad t'_{12} = (1 + e_2)\, t_{12} \quad \text{e} \quad t'_{11} = t_{11}/\big((1 + e_3)(1 + e_4)\big) \neq 0$$

si ottiene:

$$\xi_2 = c_2/t'_{22} \quad \text{e} \quad \xi_1 = (c_1 - t'_{12}\xi_2)/t'_{11}$$

ovvero $\mathtt{SI}(T, c)$ è la soluzione del sistema:

$$\begin{bmatrix} t'_{11} & t'_{12} \\ 0 & t'_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

con:

$$t'_{22} \approx t_{22} \quad , \quad t'_{12} \approx t_{12} \quad \text{e} \quad t'_{11} \approx t_{11}$$

In generale si dimostra la seguente *interpretazione*:

$$\texttt{SI}(T, c) = \text{SI}(T', c) \quad \text{con} \quad T' \text{ triangolare superiore invertibile tale che } t'_{ij} \approx t_{ij} \text{ per ogni } i, j$$

L'interpretazione mostra anche che l'algoritmo `SI` è *stabile* (Definizione 0.4.5) quando utilizzato per approssimare la funzione SI. Più precisamente, l'algoritmo è *stabile all'indietro*: per ogni valore d dell'argomento, `SI` fornisce *il valore* della funzione SI in un punto vicino a d. Si ricordi che la stabilità richiede, per ogni valore d dell'argomento, che l'algoritmo di fornisca *un'approssimazione accurata* del valore della funzione in un punto vicino a d.

2.6.2 Osservazione (inadeguatezza della procedura EGP)

L'osservazione precedente mostra che $\xi = \texttt{SI}(\hat{D}, \hat{c}) = \text{SI}(\hat{D}', \hat{c})$. Per giudicare l'accuratezza di ξ come approssimazione di $x^* = \text{SI}(D, c)$ occorre dunque studiare il *condizionamento* del calcolo della soluzione del sistema $Dx = c$, ovvero indagare il *numero di condizionamento* della matrice D.

– *Esempio*

Sia $\alpha \in (0, \frac{1}{2})$ e:

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

La procedura EGP applicata ad A produce:

$$S = \begin{bmatrix} 1 & 0 \\ 1/\alpha & 1 \end{bmatrix} \quad , \quad D = \begin{bmatrix} \alpha & 1 \\ 0 & -1/\alpha \end{bmatrix} \quad , \quad P = I$$

Allora:

$$D^{-1} = \begin{bmatrix} 1/\alpha & 1 \\ 0 & -\alpha \end{bmatrix}$$

e il numero di condizionamento di D, utilizzando ad esempio la norma infinito in $\mathbb{R}^2$ è:

$$c(D) = \| D^{-1} \| \, \| D \| = \frac{\alpha + 1}{\alpha^2}$$

Si osservi che:

$$\lim_{\alpha \to 0} c(D) = +\infty$$

e che, essendo:

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

risulta:

$$c(A) = (1 + \alpha)^2 < 3$$

Dunque: per α sufficientemente piccolo, *il fattore destro D prodotto dalla procedura* EGP *ha numero di condizionamento arbitrariamente più alto di quello di A.* Ovvero il procedimento basato su EGP ha trasformato il sistema $Ax = b$, con *buone* proprietà di condizionamento, nel sistema *equivalente* $Dx = c$ che ha però proprietà di condizionamento, per α piccolo, *pessime.*

L'esempio mostra quindi che il procedimento che usa la procedura EGP *può* generare un'approssimazione ξ non accurata.

2.6.3 Osservazione (procedura EGPP)

Per ovviare al problema evidenziato nell'esempio dell'osservazione precedente, si modifica la ricerca delle matrici di permutazione nella procedura EGP. Precisamente, per $k = 1, \ldots, n-1$: *se* $A_k(k, k) = \cdots = A_k(n, k) = 0$ si pone $P_k = I$ *altrimenti* si determina un indice $j \geqslant k$ tale che:

$$\max\{ \, | \, A_k(k, k) \, | , \ldots , | \, A_k(n, k) \, | \, \} = | \, A_k(j, k) \, |$$

e si pone:

$$P_k = \begin{cases} I & \text{se } j = k \\ P_{kj} & \text{se } j > k \end{cases}$$

La procedura così ottenuta si chiama EGPP (*E*liminazione di *G*auss con *P*ivoting *P*arziale). Applicata alla matrice A dell'esempio fornisce:

$$\mathrm{EGPP}(A) = \Big(\begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix}, I, P_{12} \Big)$$

e quindi $c(D) = 1$.

In generale si ha: *Per ogni numero intero positivo n esiste un numero reale k_n (che dipende dalla norma scelta in $\mathbb{R}^n$) tale che: per ogni matrice $A \in \mathbb{R}^{n\times n}$ invertibile, la procedura EGPP applicata ad A produce un fattore destro D con $c(D) \leqslant k_n\, c(A)$.*

– *Dimostrazione.*

Sia $\mathrm{EGPP}(A) = (S, D, P)$. Allora:

$$D = S^{-1}PA \quad \text{e} \quad D^{-1} = A^{-1}P^{-1}S$$

Scelta in $\mathbb{R}^n$ una tra le norme N_1, N_2 e N_∞, per ogni matrice di permutazione M si ha: $\| M \| = 1$ e quindi:

$$c(D) \leqslant c(S)\, c(A)$$

La tecnica del *pivoting parziale* garantisce che per ogni $k = 1, \ldots, n-1$ *il valore assoluto degli elementi non nulli della matrice H_k non supera uno.* Allora lo stesso vale per gli elementi della matrice S. Quindi:

$$\| S \|_1 \leqslant n \quad \text{e} \quad \| S \|_\infty \leqslant n$$

Per ogni $M \in \mathbb{R}^{n\times n}$ si ha: $\| M \|_2 \leqslant \sqrt{\| M \|_1 \| M \|_\infty}$. Allora si ha anche:

$$\| S \|_2 \leqslant n$$

Inoltre si ha:

$$S = PP_1^{-1}H_1^{-1} \cdots P_{n-1}^{-1}H_{n-1}^{-1} \quad \Rightarrow \quad S^{-1} = H_{n-1}P_{n-1} \cdots H_1P_1P^{-1}$$

da cui:

$$\| S^{-1} \|_\infty \leqslant \| H_{n-1} \|_\infty \cdots \| H_1 \|_\infty \leqslant 2 \cdots 2 = 2^{n-1}$$

Infine, per ogni $M \in \mathbb{R}^{n\times n}$ si ha: $\| M \|_1 \leqslant \sqrt{n}\, \| M \|_2$ e $\| M \|_2 \leqslant \sqrt{n}\, \| M \|_\infty$. Dunque:

$$\| S^{-1} \|_1 \leqslant n\, 2^{n-1} \quad \text{e} \quad \| S^{-1} \|_2 \leqslant \sqrt{n}\, 2^{n-1}$$

Dalle disuguaglianze ottenute si ricava:

$$c_1(S) \leqslant n^2\, 2^{n-1} \quad , \quad c_2(S) \leqslant n^{3/2}\, 2^{n-1} \quad , \quad c_\infty(S) = n\, 2^{n-1}$$

da cui l'asserto.

Esercizi

E26 Siano $T \in \mathbb{R}^{2\times 2}$ una matrice triangolare inferiore invertibile e $c \in \mathbb{R}^2$. Mostrare che, *supponendo che gli elementi di T e c siano elementi di M*, sussiste la seguente *interpretazione*, simile a quella data per `SI`:

$$\texttt{SA}(T, c) = \mathrm{SA}(T', c) \quad \text{con} \quad T' \text{ triangolare inferiore tale che } t'_{ij} \approx t_{ij} \text{ per ogni } i, j$$

E27 Sia:

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

Determinare $\mathrm{EGPP}(A)$.

Soluzione:

$$S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \quad , \quad D = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad , \quad P = P_{23}P_{13} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

2.7 Fattorizzazione QR: la procedura GS

Assegnata una matrice $A \in \mathbb{R}^{n\times n}$, la procedura GS (*procedimento di Gram-Schmidt*) di intestazione:

$$(U, T) = \mathrm{GS}(A)$$

cerca una *fattorizzazione QR* di A.

Se la procedura trova una fattorizzazione, la soluzione del sistema $Ax = b$ si determina calcolando $x^* = \mathrm{SI}(T, U^\mathsf{T} b)$.

L'esempio seguente mostra come utilizzare il procedimento di ortonormalizzazione di Gram-Schmidt per cercare una fattorizzazione QR di una matrice A.

2.7.1 Esempio

Si consideri $\mathbb{R}^3$ con prodotto scalare canonico (per ogni $a, b \in \mathbb{R}^3$: $a \cdot b = b^\mathsf{T} a$) e sia $A \in \mathbb{R}^{3\times 3}$ di colonne a_1, a_2, a_3.

– *Primo passo*

Si cercano $\Omega = (\omega_1, \omega_2, \omega_3) \in \mathbb{R}^{3\times 3}$ a colonne ortogonali non nulle e $\Theta \in \mathbb{R}^{3\times 3}$ triangolare superiore con uno sulla diagonale tali che $\Omega\Theta = A$, ovvero tali che:

$$\omega_1 = a_1 \quad , \quad \omega_1\theta_{12} + \omega_2 = a_2 \quad , \quad \omega_1\theta_{13} + \omega_2\theta_{13} + \omega_3 = a_3$$

Se esistono matrici siffatte, allora *necessariamente*:

$$\omega_1 = a_1 \quad , \quad \omega_2 = a_2 - \omega_1\theta_{12} \quad , \quad \omega_3 = a_3 - \omega_1\theta_{13} - \omega_2\theta_{23}$$

e, dalla seconda uguaglianza:

$$\omega_2 \cdot \omega_1 = 0 \quad \text{se e solo se} \quad (\omega_1 \cdot \omega_1)\,\theta_{12} = a_2 \cdot \omega_1$$

dalla terza:

$$\omega_3 \cdot \omega_1 = 0 \quad \text{se e solo se} \quad (\omega_1 \cdot \omega_1)\,\theta_{13} + (\omega_2 \cdot \omega_1)\,\theta_{23} = a_3 \cdot \omega_1$$
$$\omega_3 \cdot \omega_2 = 0 \quad \text{se e solo se} \quad (\omega_1 \cdot \omega_2)\,\theta_{13} + (\omega_2 \cdot \omega_2)\,\theta_{23} = a_3 \cdot \omega_2$$

La procedura seguente determina Ω e Θ con le proprietà richieste se e solo se *le colonne di A sono linearmente indipendenti*:

$\omega_1 = a_1$;
`se` $\omega_1 = 0$ `allora`: interrompi la costruzione;
`altrimenti`: $\theta_{12} = \dfrac{a_2 \cdot \omega_1}{\omega_1 \cdot \omega_1}$; $\theta_{13} = \dfrac{a_3 \cdot \omega_1}{\omega_1 \cdot \omega_1}$;
$\omega_2 = a_2 - \omega_1\theta_{12}$;
`se` $\omega_2 = 0$ `allora`: interrompi la costruzione;
`altrimenti`: $\theta_{23} = \dfrac{a_3 \cdot \omega_2}{\omega_2 \cdot \omega_2}$;
$\omega_3 = a_3 - \omega_1\theta_{13} - \omega_2\theta_{23}$;
`se` $\omega_3 = 0$ `allora`: interrompi la costruzione;
`altrimenti`: $\Omega = (\omega_1, \omega_2, \omega_3) \quad , \quad \Theta = \begin{bmatrix} 1 & \theta_{12} & \theta_{13} \\ 0 & 1 & \theta_{23} \\ 0 & 0 & 1 \end{bmatrix}$

– *Secondo passo*

Siano Ω e Θ le matrici determinate dal *Primo passo* e:

$$\Delta = \mathrm{diag}(\| \omega_1 \|, \| \omega_2 \|, \| \omega_3 \|) \in \mathbb{R}^{3\times 3}$$

Si ricordi che $\omega_1 \neq 0, \omega_2 \neq 0, \omega_3 \neq 0$ e quindi Δ è invertibile. La coppia:

$$U = \Omega\,\Delta^{-1} \quad , \quad T = \Delta\,\Theta$$

è una fattorizzazione QR di A.

La procedura seguente, descritta in un linguaggio che consente l'uso del tipo *numero reale*, formalizza il procedimento descritto nell'esempio:

$[U,T] = \mathrm{GS}(A)$

// A matrice $n \times n$ ad elementi reali.
//
// ** Primo passo: cerca Ω matrice $n \times n$ a colonne ortogonali non nulle e Θ
// matrice $n \times n$ triangolare superiore con uno sulla diagonale tali che $\Omega\Theta = A$
//

$\omega_1 = a_1$;

`per` $k = 1, \ldots, n-1$ `ripeti`:
 `se` $\omega_k = 0$ `allora`: interrompi la costruzione e segnala A non invertibile;
 `altrimenti`:
 $d_k = \omega_k \cdot \omega_k$
 `per` $j = k+1, \ldots, n$ `ripeti`: $\theta_{kj} = \dfrac{a_j \cdot \omega_k}{d_k}$;
 $\omega_{k+1} = a_{k+1} - (\omega_1\theta_{1,k+1} + \cdots + \omega_k\theta_{k,k+1})$;

`se` $\omega_n = 0$ `allora`: interrompi la costruzione e segnala A non invertibile;
 //
 // ** Secondo passo: costruisce la fattorizzazione QR normalizzando le colonne di Ω
 //
 `altrimenti`:
 $d_n = \omega_n \cdot \omega_n$;
 $\Delta = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$;

$$U = (\omega_1, \ldots, \omega_n)\,\Delta^{-1}; \qquad T = \Delta \begin{bmatrix} 1 & & & \\ & \ddots & \theta_{ij} & \\ & 0 & \ddots & \\ & & & 1 \end{bmatrix}$$

La procedura GS termina correttamente (ovvero: determina una fattorizzazione QR di A) *se e solo se* la matrice A è *invertibile*.

2.7.2 Osservazione (non unicità della fattorizzazione QR, la funzione predefinita `qr`)

(1) Se U, T è una fattorizzazione QR di A ed $E \in \mathbb{R}^{n\times n}$ è una matrice diagonale tale che $|\,e_{11}\,| = 1, \ldots, |\,e_{nn}\,| = 1$ allora la coppia:

$$U' = UE \quad , \quad T' = ET$$

è a sua volta una fattorizzazione QR di A. Dunque: *la fattorizzazione QR non è unica.*

(2) La procedura GS mostra che *qualunque matrice invertibile ammette fattorizzazione QR.* Esistono altre procedure per la ricerca di una fattorizzazione QR di una matrice, più generali di GS e preferibili ad essa da un punto di vista numerico. Queste procedure terminano correttamente *in ogni caso* e dunque mostrano che *qualunque matrice $n \times n$ ammette fattorizzazione QR. Scilab* ha una funzione predefinita per il calcolo di una fattorizzazione QR di una matrice che utilizza una di queste altre procedure, il *metodo di Householder*:[33]

* `qr`

 Questa *funzione predefinita* restituisce un'approssimazione di una fattorizzazione QR di un'assegnata matrice A. Precisamente, se `A` è una matrice $n \times n$:

 ```
 [U,T] = qr(A)
 ```

 restituisce la coppia `U`,`T` di matrici $n \times n$, `T` triangolare superiore, che *approssima* una fattorizzazione QR di `A`. Come già osservato `A` *può* non essere invertibile (vedere l'Esercizio *E32*).

[33] Si veda, ad esempio: `https://en.wikipedia.org/wiki/QR_decomposition#Using_Householder_reflections`.

2.7.3 Osservazione (uso del tipo *numero in virgola mobile*)

Sia qr una procedura che per ogni matrice $A \in \mathbb{R}^{n\times n}$ determina una fattorizzazione QR di A. Si ricordi che, assegnata una matrice $A \in \mathbb{R}^{n\times n}$ ed una colonna $b \in \mathbb{R}^n$, la procedura qr permette la ricerca della soluzione sistema $Ax = b$ con il seguente procedimento descritto con un linguaggio che consente l'uso del tipo *numero reale*:

$(U, T) = \mathrm{qr}(A)$;
`se` esiste k tale che $t_{kk} = 0$ `allora` arresta il procedimento e dichiara A non invertibile;
  `altrimenti`
    $x^* = \mathrm{SI}(T, U^\mathsf{T} b)$

Sostituendo al tipo *numero reale* il tipo *numero in virgola mobile e precisione finita* il procedimento si modifica come segue:

$(\hat{U}, \hat{T}) = \mathtt{qr}(A')$;
`se` esiste k tale che $\hat{t}_{kk} = 0$ `allora` arresta il procedimento e dichiara A' non invertibile;
  `altrimenti`
    $\xi = \mathtt{SI}(\hat{T}, \hat{U}^\mathsf{T} * \hat{b})$

dove: A' e b' indicano, rispettivamente, la matrice e la colonna ottenute arrotondando in M ciascuna componente di A e b, `qr` ed `SI` indicano le procedure ottenute sostituendo il tipo *numero reale* rispettivamente nelle procedure qr ed SI e $\hat{U}^\mathsf{T} * \hat{b}$ indica la matrice ottenuta sostituendo le pseudo-operazioni aritmetiche alle operazioni aritmetiche nel prodotto riga per colonna $\hat{U}^\mathsf{T}\hat{b}$.

Il vettore finale ξ è utilizzato per approssimare x^* e l'approssimazione è *sempre* accurata quanto consentito dal condizionamento di A. Infatti, anche in questo caso per giudicare l'accuratezza occorre studiare il *condizionamento* del calcolo della soluzione del sistema $Tx = U^\mathsf{T} b$, ovvero indagare il *numero di condizionamento* della matrice T.

Scelta in $\mathbb{R}^n$ la norma euclidea N_2, si studia:

$$c_2(T) = \| T^{-1} \|_2 \, \| T \|_2$$

Si ha:

$$A = UT \quad \Rightarrow \quad T = U^\mathsf{T} A \qquad \text{e} \qquad A^{-1} = T^{-1} U^\mathsf{T} \quad \Rightarrow \quad T^{-1} = A^{-1} U$$

Dunque:

$$\| T \|_2 \leqslant \| U^\mathsf{T} \|_2 \, \| A \|_2 \quad \text{e} \quad \| T^{-1} \|_2 \leqslant \| A^{-1} \|_2 \, \| U \|_2$$

Ma: per ogni matrice M ortogonale si ha $\| M \|_2 = 1$. Perciò:

$$\| T \|_2 \leqslant \| A \|_2 \quad \text{e} \quad \| T^{-1} \|_2 \leqslant \| A^{-1} \|_2 \quad \Rightarrow \quad c_2(T) \leqslant c_2(A)$$

In questo caso *la procedura di fattorizzazione QR produce un fattore destro con proprietà di condizionamento non peggiori di quelle della matrice iniziale.*

Esercizi

E28 Sia:

$$A = \begin{bmatrix} 0 & 1 & 6 \\ 1 & -1 & 1 \\ 2 & 3 & 2 \end{bmatrix}$$

Determinare $\mathrm{GS}(A)$.

E29 ★ Verificare che, se A è una matrice invertibile, la coppia U, T definita da $\mathrm{GS}(A)$ è una fattorizzazione QR di A.

E30 ★ Sia $A \in \mathbb{R}^{3\times 3}$. Dimostrare che se $\mathrm{GS}(A)$ termina prematuramente allora A non è invertibile.

E31 Sia:

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

Determinare $\mathrm{GS}(A)$ e dedurre dal risultato due diverse fattorizzazioni QR di A.

E32 ♠ Sia $A \in \mathbb{R}^{n\times n}$ la matrice nulla. Per $n = 5$ verificare che la procedura GS applicata ad A termina prematuramente mentre $\mathtt{qr}(A)$ determina una fattorizzazione QR *esatta*.

E33 ★ Dimostrare, utilizzando la definizione di norma indotta, che: *se M è una matrice ortogonale allora* $\| M \|_2 = 1$.

2.8 Costo

La nozione di *costo*, già implicitamente adottata in precedenza, è quella "aritmetica:"

2.8.1 Definizione (costo aritmetico)
Sia ϕ un algoritmo. Si chiama *costo aritmetico* di ϕ *il numero* $C(\phi)$ *di pseudo-operazioni aritmetiche eseguite per calcolare* $\phi(x)$.[34]

Si osservi che il costo di un algoritmo deve essere una quantità almeno approssimativamente proporzionale al *tempo* necessario al calcolatore per portare a termine il calcolo. La definizione adottata riesce nell'intento se:

(a) La maggior parte del tempo impiegato dal calcolatore per calcolare ϕ è speso nell'esecuzione di pseudo-operazioni aritmetiche.

(b) Il tempo necessario per eseguire ciascuna pseudo-operazione aritmetica è lo stesso, in particolare *non dipende dagli operandi*.

Il sussistere della condizione (a) *dipende da* ϕ. Ad esempio, se ϕ è un algoritmo per il calcolo della norma infinito di un vettore il costo aritmetico è *zero* – per questo algoritmo *nessuna* delle funzioni predefinite calcolate è una pseudo-operazione aritmetica – ma non è vero che il calcolatore impiega tempo zero a calcolare $\phi(x)$. Nel seguito la definizione di costo sarà applicata solo ad algoritmi che calcolano *quasi esclusivamente* pseudo-operazioni aritmetiche.

Il sussistere della condizione (b), invece, *non dipende da* ϕ ma *dipende dalla scelta di* M. Si consideri, ad esempio, il calcolo in $F(2, 53)$ dello pseudo-prodotto $2^{b_1} \otimes 2^{b_2} = 2^{b_1+b_2}$. *Non è ragionevole* supporre che il tempo necessario per il calcolo sia indipendente da quali numeri interi b_1 e b_2 occorre sommare (calcolare la somma si due numeri interi *non può* richiedere un tempo *indipendente* dal numero di cifre necessario per rappresentare gli addendi – si pensi al solo tempo necessario per *leggere gli addendi e scrivere la somma*). L'ipotesi (b) è verificata, invece, qualora M sia un insieme di numeri in virgola mobile con *esponente limitato*.

Vediamo alcuni esempi di determinazione del costo di un algoritmo e poi confrontiamo il costo degli algoritmi dedotti dai due procedimenti proposti per la ricerca della soluzione di un sistema di equazioni lineari.

2.8.2 Esempio

∗ *Prodotto riga per colonna*

Sia $\mathtt{prc}_n$ l'algoritmo ottenuto dal procedimento di calcolo del *prodotto riga per colonna* $a^{\mathsf{T}}b$, con $a, b \in \mathbb{R}^n$, sostituendo ciascuna operazione aritmetica con la corrispondente pseudo-operazione e specificando l'ordine di composizione delle pseudo-operazioni. La corrispondenza biunivoca tra operazioni aritmetiche e pseudo-operazioni così stabilita rende possibile determinare il costo di $\mathtt{prc}_n(a, b)$ calcolando il numero di operazioni aritmetiche in $a^{\mathsf{T}}b$. Dette a_i, b_i le componenti di a, b si ha:

$$a^{\mathsf{T}}b = a_1b_1 + \cdots + a_nb_n$$

Il calcolo di $a^{\mathsf{T}}b$ richiede n prodotti e $n - 1$ somme. Allora:

$$C(\mathtt{prc}_n) = 2n - 1$$

[34] Il costo va valutato *nel caso peggiore*: il numero di pseudo-operazioni aritmetiche eseguite per calcolare $\phi(x)$ potrebbe dipendere da x.

∗ *Prodotto matrice per colonna*

Sia $\texttt{pmc}_n$ l'algoritmo ottenuto dal procedimento di calcolo del *prodotto matrice per colonna* Ab, con $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. Si ragiona come nel caso del prodotto riga per colonna, ovvero si calcola il numero di operazioni aritmetiche in Ab. Dette $r_1, \ldots, r_n$ le righe di A si ha:

$$Ab = \begin{bmatrix} r_1 b \\ \vdots \\ r_n b \end{bmatrix}$$

Ciascuna delle n componenti del vettore Ab è un prodotto riga per colonna. Il calcolo di Ab richiede quindi n^2 prodotti e $n(n-1)$ somme. Allora:

$$C(\texttt{pmc}_n) = n\, C(\texttt{prc}_n) = 2n^2 - n$$

∗ *Prodotto matrice triangolare per colonna*

Sia $\texttt{pmtc}_n$ l'algoritmo ottenuto dal procedimento di calcolo del *prodotto matrice triangolare per colonna* Tc, con $T \in \mathbb{R}^{n\times n}$ matrice triangolare, $c \in \mathbb{R}^n$. Si ragiona come nel caso del prodotto matrice per colonna. Dette $r_1, \ldots, r_n$ le righe di T si ha:

$$Tc = \begin{bmatrix} r_1 c \\ \vdots \\ r_n c \end{bmatrix}$$

Ciascuna delle n componenti del vettore Tc è un prodotto riga per colonna. Questa volta, però, ciascuna componente ha un costo diverso. Supponendo, ad esempio, T triangolare superiore ed evitando di calcolare operazioni con risultato noto a priori (per ogni $x \in \mathbb{R}$ si ha $0\,x = 0$ e $0 + x = x$) si ottiene:

$$C(\texttt{pmtc}_n) = C(\texttt{prc}_n) + C(\texttt{prc}_{n-1}) + \cdots + C(\texttt{prc}_1) = 2(1 + 2 + \cdots + n) - n = n^2$$

Il costo di $\texttt{pmtc}_n$ è circa *metà* del costo di $\texttt{pmc}_n$.

∗ *Procedura* SI

Come già sappiamo dal Paragrafo 2.1 il calcolo di $\mathrm{SI}(T, c)$ per $T \in \mathbb{R}^{n\times n}$ e $c \in \mathbb{R}^n$ richiede n divisioni e $\frac{1}{2}\, n(n-1)$ prodotti e somme. Allora:

$$C(\texttt{SI}) = n^2$$

Il costo è uguale a quello di $\texttt{pmtc}_n$. Gli stessi risultati si ottengono per `SA`.

∗ *Procedura* EGPP

Il calcolo di $\texttt{EGPP}(A)$ per $A \in \mathbb{R}^{n\times n}$ risulta richiedere:

$$\sum_{k=1}^{n-1} k = \tfrac{1}{2}\, n(n-1) \text{ divisioni e } \sum_{k=1}^{n-1} k^2 = \tfrac{1}{6}\,(n-1)n(2n-1) \text{ prodotti e somme}$$

In totale:

$$C(\texttt{EGPP}) = \tfrac{2}{3}\, n^3 - \tfrac{1}{2}\, n^2 - \tfrac{1}{6}\, n$$

Si osservi che il calcolo di $\texttt{EGPP}(A)$ richiede anche *confronti*, in numero trascurabile rispetto alle pseudo-operazioni aritmetiche.

Il procedimento per per la ricerca della soluzione di un sistema di equazioni lineari che utilizza la fattorizzazione LR ha dunque costo:

$$C(\texttt{EGPP}) + C(\texttt{SA}) + C(\texttt{SI}) = \tfrac{2}{3}\, n^3 + \text{ termini di grado inferiore in } n$$

Un calcolo analogo per il procedimento per per la ricerca della soluzione di un sistema di equazioni lineari che utilizza la fattorizzazione QR porta ad un costo:[35]

$$C(\texttt{qr}) + C(\texttt{pmc}_n) + C(\texttt{SI}) = \tfrac{4}{3}\, n^3 + \text{ termini di grado inferiore in } n$$

[35] Si veda la pagina di Wikipedia citata in precedenza. Si osservi che il calcolo della fattorizzazione QR richiede anche $n-1$ radici quadrate.

che è circa *il doppio* del precedente. Si osservi che si è scelto di esprimere il costo dei procedimenti mostrando esplicitamente solo *il termine dominante* al crescere di n. In entrambi i casi *il termine dominante è generato dalla procedura di ricerca della fattorizzazione.*

2.8.3 Osservazione (calcolo della matrice inversa)

La funzione predefinita `inv` di *Scilab* cerca un'approssimazione della matrice inversa di una data matrice usando la procedura `EGPP`. Sia $A \in \mathbb{R}^{n\times n}$ una matrice invertibile. Dette $e_1, \ldots, e_n$ le colonne della matrice identica, il calcolo di $Y = A^{-1}$ avviene in questo modo:

$[S, D, P] = \mathrm{EGPP}(A)$

`se` $d_{kk} = 0$ per qualche k `allora` arresta la procedura e dichiara A non invertibile;
`altrimenti`
`per` $k = 1, \ldots, n$ `ripeti`:
$c = \mathrm{SA}(S, Pe_k)$;
$y_k = \mathrm{SI}(D, c)$;

$Y = (y_1 \ldots, y_n)$.

Dunque si calcolano le n colonne della matrice inversa come soluzione degli n sistemi lineari $Ay = e_1, \ldots, Ay = e_n$. Tutti i sistemi hanno *la stessa matrice* e per la soluzione degli n sistemi è sufficiente calcolare *una sola volta* la fattorizzazione di A. Questo fa sì che il termine dominante del costo del procedimento sia ancora un multiplo di n^3.

Esercizi

E34 ★ Verificare che il calcolo della matrice H_k nel passo k-esimo di EGPP(A) richiede $n - k$ divisioni e che il calcolo del prodotto $H_k P_k A^{(k)}$ richiede $(n - k)^2$ prodotti e somme.

E35 Verificare che il calcolo di EGPP(A) richiede non più di $\frac{1}{2}\, n(n-1)$ confronti. Determinare l'errore relativo commesso approssimando il numero di pseudo-operazioni aritmetiche e confronti richiesto dal calcolo di EGPP(A) con il numero delle sole pseudo-operazioni aritmetiche.

E36 Sia $\mathtt{ata}_n$ l'algoritmo ottenuto dal procedimento di calcolo del prodotto $A^{\mathsf{T}}A$ con $A \in \mathbb{R}^{n\times n}$. Determinare $C(\mathtt{ata}_n)$. Si tenga conto che la matrice $A^{\mathsf{T}}A$ è *simmetrica.*

---

3 Interpolazione

Se $\Omega \subset \mathbb{R}$ è un intervallo, si indica con $C(\Omega)$ lo *spazio vettoriale su* $\mathbb{R}$ delle funzioni continue da Ω in $\mathbb{R}$.

3.1 Interpolazione polinomiale

Siano k un numero intero non negativo, $P_k(\mathbb{R})$ lo spazio vettoriale su $\mathbb{R}$ dei *polinomi a coefficienti reali di grado al più* k (che consideriamo come *sottospazio* di $C(\mathbb{R})$) $x_0, \ldots, x_k$ numeri reali *distinti*[36] e $y_0, \ldots, y_k$ numeri reali. Il *problema dell'interpolazione polinomiale* consiste nel *determinare* gli elementi $p \in P_k(\mathbb{R})$ che *interpolano i dati*:

$$(x_0, y_0), \ldots, (x_k, y_k)$$

ovvero tali che:

$$p(x_0) = y_0, \ldots, p(x_k) = y_k$$

3.1.1 Osservazione

Si consideri il problema di intepolazione polinomiale di determinare gli elementi di $P_k(\mathbb{R})$ che interpolano i dati: $(x_0, y_0), \ldots, (x_k, y_k)$. Si ha:

(a) *Interpretazione geometrica*

Considerati i dati $(x_0, y_0), \ldots, (x_k, y_k)$ come *coordinate di* $k+1$ *punti* in un piano cartesiano, il problema dell'interpolazione polinomiale consiste nel determinare *gli elementi di* $P_k(\mathbb{R})$ *il cui grafico contiene tutti i punti* assegnati.

(b) *Riformulazione*

Ricordato che lo spazio vettoriale $P_k(\mathbb{R})$ ha dimensione $k+1$, sia $q_0(x), \ldots, q_k(x)$ una sua *base*. Allora: $p(x) = a_0 q_0(x) + \cdots + a_k q_k(x)$ *è soluzione del problema di interpolazione polinomiale* se e solo se:

$$\begin{aligned} p(x_0) = a_0 q_0(x_0) + \cdots + a_k q_k(x_0) = y_0 \\ \vdots \\ p(x_k) = a_0 q_0(x_k) + \cdots + a_k q_k(x_k) = y_k \end{aligned}$$

ovvero *se e solo se la colonna dei coefficienti*:

$$\begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix} \in \mathbb{R}^{k+1}$$

è soluzione del sistema di equazioni lineari:

$$\begin{bmatrix} q_0(x_0) & q_1(x_0) & \cdots & q_k(x_0) \\ \vdots & \vdots & & \vdots \\ q_0(x_k) & q_1(x_k) & \cdots & q_k(x_k) \end{bmatrix} z = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}$$

Si osservi che si ottengono *tante equazioni quante sono le condizioni* richieste a $p(x)$ e *il numero di incognite è pari alla dimensione* di $P_k(\mathbb{R})$. Inoltre, poiché $q_0(x), \ldots, q_k(x)$ è una base di $P_k(\mathbb{R})$, l'insieme delle soluzioni del problema di interpolazione polinomiale e quello delle soluzioni del sistema di equazioni lineari sono in *corrispondenza biunivoca*.

3.1.2 Teorema (di esistenza ed unicità della soluzione)

Per ogni k numero intero non negativo, $x_0, \ldots, x_k$ numeri reali distinti e $y_0, \ldots, y_k$ numeri reali, *esiste un solo elemento in* $P_k(\mathbb{R})$ *che interpola i dati* $(x_0, y_0), \ldots, (x_k, y_k)$.

Dimostrazione. Siano $\ell_0(x), \ldots, \ell_k(x)$ gli elementi di $P_k(\mathbb{R})$ definiti da:

$$\ell_j(x) = \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_k)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_k)} \quad , \quad j = 0, \ldots, k$$

Si osservi che per ogni j si ha:

[36] Ovvero: $i \neq j \Rightarrow x_i \neq x_j$

(a) $\ell_j(x_j) = 1$

(b) se $i \neq j$ allora $\ell_j(x_i) = 0$

I $k+1$ polinomi $\ell_0(x), \ldots, \ell_k(x)$ sono *linearmente indipendenti*. Infatti, se $a_0, \ldots, a_k$ sono coefficienti tali che per ogni $x \in \mathbb{R}$ si ha:

$$a_0\ell_0(x) + \cdots + a_k\ell_k(x) = 0$$

allora per $j = 0, \ldots, k$ si ha:

$$0 = a_0\ell_0(x_j) + \cdots + a_k\ell_k(x_j) = a_j$$

Dunque $\ell_0(x), \ldots, \ell_k(x)$ sono una *base* di $P_k(\mathbb{R})$ detta *base di Lagrange* relativa ai punti $x_0, \ldots, x_k$. Inoltre:

$$\begin{bmatrix} \ell_0(x_0) & \ell_1(x_0) & \cdots & \ell_k(x_0) \\ \vdots & \vdots & & \vdots \\ \ell_0(x_k) & \ell_1(x_k) & \cdots & \ell_k(x_k) \end{bmatrix} = I$$

Per quanto detto nel punto (b) dell'Osservazione 3.1.1, il problema di interpolazione polinomiale ha *una ed una sola soluzione*:

$$p(x) = y_0\ell_0(x) + \cdots + y_k\ell_k(x)$$

L'espressione trovata prende il nome di *forma di Lagrange del polinomio interpolante*.

3.1.3 Esercizio

Determinare l'elemento di $P_2(\mathbb{R})$ che interpola i dati: $(-1, 0), (0, 1), (2, -2)$.

Soluzione.

Numerati i dati nell'ordine delle ascisse crescenti, la *base di Lagrange* di $P_2(\mathbb{R})$ relativa ai punti x_0, x_1, x_2 è:

$$\ell_0(x) = \frac{(x-0)(x-2)}{(-1-0)(-1-2)} = \tfrac{1}{3}(x^2 - 2x) \qquad , \qquad \ell_1(x) = \frac{(x+1)(x-2)}{(0+1)(0-2)} = -\tfrac{1}{2}(x^2 - x - 2)$$

e:

$$\ell_2(x) = \frac{(x+1)(x-0)}{(2+1)(2-0)} = \tfrac{1}{6}(x^2 + x)$$

La *forma di Lagrange* dell'elemento cercato è allora:

$$p(x) = 0 \cdot \ell_0(x) + 1 \cdot \ell_1(x) - 2 \cdot \ell_2(x)$$

Lo stesso elemento può essere individuato utilizzando la più usuale *base di Vandermonde* di $P_2(\mathbb{R})$:

$$1, x, x^2$$

In questo caso il sistema di equazioni a cui si riducono le condizioni di interpolazione è:

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

Risolvendo il sistema si ottiene l'elemento cercato in *forma di Vandermonde*:

$$p(x) = 1 \cdot 1 + \tfrac{1}{6} \cdot x - \tfrac{5}{6} \cdot x^2$$

3.1.4 Osservazione (forma di Newton del polinomio interpolante)

La scelta della base di $P_k(\mathbb{R})$ *non* fa cambiare la soluzione del problema di interpolazione in esame *ma* può agevolarne o meno la determinazione. La *base di Lagrange*, di non immediata manipolazione, genera il sistema di equazioni lineari *più semplice* da risolvere per la determinazione dei coefficienti perché la matrice del sistema è I. La più usuale *base di Vandermonde*, invece, genera un sistema di equazioni *non semplice* da risolvere perché la matrice del sistema è la *matrice di Vandermonde*

(vedere l'Esercizio *E2*) e la soluzione del sistema richiede la sua fattorizzazione. Una terza scelta è la *base di Newton* relativa ai punti $x_0, \ldots, x_{k-1}$:

$$1, x - x_0, (x - x_0)(x - x_1), \ldots, (x - x_0) \cdots (x - x_{k-1})$$

Gli elementi di questa base sono polinomi *di grado crescente* e le condizioni di interpolazione si traducono in un sistema di equazioni lineari con *matrice triangolare inferiore*, dunque semplice da risolvere.

3.1.5 Esercizio

Assegnati i dati $(0, 1), (-1, 2), (3, 10), (1, 10)$:

(a) Determinare la *forma di Newton* del polinomio interpolante utilizzando i dati nell'ordine in cui sono stati assegnati;

(b) Determinare la forma di Newton del polinomio interpolante ordinando i dati secondo ascisse crescenti;

(c) Calcolare il *valore* del polinomio interpolante in $-1, 0, 1, 2, 3, 4$.

Soluzione.

(a) Sono stati assegnati quattro dati, dunque $k = 3$. Utilizzando i dati nell'ordine in cui sono stati assegnati la *base di Newton* di $P_3(\mathbb{R})$ risulta:

$$1, x, x(x + 1), x(x + 1)(x - 3)$$

Le condizioni di interpolazione si traducono nel sistema:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 3 & 12 & 0 \\ 1 & 1 & 2 & -4 \end{bmatrix} x = \begin{bmatrix} 1 \\ 2 \\ 10 \\ 10 \end{bmatrix}$$

La soluzione b del sistema si ottiene con la procedura SA:

$$b = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -2 \end{bmatrix} \quad \Rightarrow \quad p(x) = 1 \cdot 1 - 1 \cdot x + 1 \cdot x(x + 1) - 2 \cdot x(x + 1)(x - 3)$$

(b) *Lasciata al lettore.*

(c) Un procedimento per calcolare il valore di p in $a \in \mathbb{R}$ è, dette $x_0, \ldots, x_k$ le ascisse dei dati nell'ordine considerato e b la colonna di componenti i coefficienti $b_0, \ldots, b_k$:

(1) $r_0 = 1; r_1 = a - x_0$; `per` $j = 2, \ldots, k$ `ripeti` $r_j = r_{j-1}(a - x_{j-1})$;

(2) $r = (r_0, \ldots, r_k)$; $p(a) = r\, b$

In (1) si calcolano i valori degli elementi della base di Newton in a, in (2) si calcola il valore in a della combinazione lineare che realizza il polinomio interpolante. La procedura richiede $2k$ somme e $2k - 1$ prodotti. Per il calcolo in m punti sono richieste $2mk$ somme e $2m(k-1)$ prodotti e quindi il costo del calcolo è: $(4k - 1)m$. Nel caso in esame si ottiene:

$$\begin{bmatrix} p(-1) \\ p(0) \\ p(1) \\ p(2) \\ p(3) \\ p(4) \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 2 & -4 \\ 1 & 2 & 6 & -6 \\ 1 & 3 & 12 & 0 \\ 1 & 4 & 20 & 20 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 10 \\ 17 \\ 10 \\ -23 \end{bmatrix}$$

Esercizi

E1 Verificare che le due forme, di *Lagrange* e *Vandermonde*, dell'elemento di $P_2(\mathbb{R})$ che interpola i dati dell'Esercizio 3.1.3 individuano lo stesso polinomio.

E2 Siano k un numero intero non negativo e $1, \ldots, x^k$ la *base di Vandermonde* di $P_k(\mathbb{R})$. Assegnati $x_0, \ldots, x_k$ numeri reali *distinti* e $y_0, \ldots, y_k$ numeri reali, verificare che il sistema di equazioni lineari a cui si riducono le condizioni di interpolazione è:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^k \\ 1 & x_1 & x_1^2 & \cdots & x_1^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^k \end{bmatrix} x = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix}$$

In base al Teorema di esistenza ed unicità della soluzione, la matrice del sistema, nota come *matrice di Vandermonde* relativa ai punti $x_0, \ldots, x_k$, è *invertibile*.

E3 ♠ Realizzare in *Scilab* una procedura che *dati* un vettore a di componenti $a_1, \ldots, a_m$ e due vettori x e b di componenti $x_0, \ldots, x_k$ e, rispettivamente, $b_0, \ldots, b_k$, *restituisce* il vettore p di componente j-esima:

$$p_j = b_0 + b_1(a_j - x_0) + \cdots + b_k(a_j - x_0)\cdots(a_j - x_{k-1}) \quad , \quad j = 1, \ldots, m$$

Utilizzare poi la procedura per disegnare, su uno stesso piano cartesiano, in rosso *il grafico* del polinomio del punto (a) dell'Esercizio 3.1.5 e con crocette *i dati* interpolati.

E4 Dopo aver rappresentato i dati $(0,0), (1,1), (3,3), (4,4)$ su un piano cartesiano, determinare la forma di Vandermonde e la forma di Newton *del* polinomio interpolante.

E5 ★ Per ogni $n \in \mathbb{N}$ si ha:

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = Q(n)$$

con $Q \in P_3(\mathbb{R})$. Determinare la forma di Newton di Q e dedurne la forma di Vandermonde.

3.2 Problema lineare di interpolazione

Siano F un sottospazio vettoriale di $C(\mathbb{R})$ di dimensione *finita* d, $L_0, \ldots, L_k$ applicazioni lineari da F in $\mathbb{R}$ e $y_0, \ldots, y_k$ numeri reali. Il *problema lineare di intepolazione* consiste nel *determinare* gli elementi $f \in F$ che verificano le condizioni:

$$L_0(f) = y_0, \ldots, L_k(f) = y_k$$

3.2.1 Esempio

(1) Il problema di interpolazione polinomiale definito da $k, x_0, \ldots, x_k$ e $y_0, \ldots, y_k$ è il problema lineare di interpolazione definito da: $F = P_k(\mathbb{R})$, $L_0(f) = f(x_0), \ldots, L_k(f) = f(x_k)$ e $y_0, \ldots, y_k$.

Infatti: $P_k(\mathbb{R})$ è un sottspazio di $C(\mathbb{R})$ di dimensione finita e per ogni $a \in \mathbb{R}$ l'applicazione $L : P_k(\mathbb{R}) \to \mathbb{R}$ definita da $L(p) = p(a)$ è tale che:

$$p, q \in P_k(\mathbb{R}) \quad \Rightarrow \quad L(p+q) = (p+q)(a) = p(a) + q(a) = L(p) + L(q)$$

e:

$$p \in P_k(\mathbb{R}), \alpha \in \mathbb{R} \quad \Rightarrow \quad L(\alpha\, p) = (\alpha\, p)(a) = \alpha\, p(a) = \alpha\, L(p)$$

ovvero *è lineare*.

(2) Il problema lineare di interpolazione definito da $F = P_3(\mathbb{R})$, $L_0(p) = p(0)$, $L_1(p) = p(2)$ e $y_0 = 2, y_1 = -6$ *non è un problema di interpolazione polinomiale*.

Infatti: Si cercano in $P_3(\mathbb{R})$, spazio vettoriale di dimensione *quattro*, elementi che verificano *solo due* condizioni di interpolazione. In particolare, al problema in esame *non si applica* il Teorema 3.1.2 di esistenza ed unicità.

(3) Il problema lineare di interpolazione definito da $F = P_2(\mathbb{R})$, $L_0(p) = p(0)$,

$$L_1(p) = \int_0^1 p(\theta)d\theta$$

$L_2(p) = p'(0)$ e $y_0 = 2, y_1 = -6, y_2 = 4$ *non è un problema di interpolazione polinomiale.*

Infatti: Si cercano in $P_2(\mathbb{R})$, spazio vettoriale di dimensione tre, elementi che verificano tre condizioni *ma non tutte del tipo richiesto dal problema di interpolazione polinomiale.*

Si osservi che assegnati numeri reali a, b tali che $a < b$, l'applicazione $L : P_3(\mathbb{R}) \to \mathbb{R}$ definita da:

$$L(p) = \int_a^b p(\theta)d\theta$$

è lineare, come pure, per ogni intero positivo j, quella definita da:

$$L(p) = p^{(j)}(a)$$

(4) Il problema lineare di interpolazione definito da $F = \text{span}\{\, 1, \operatorname{sen} x, \cos x \,\}$, $L_0(f) = f(\pi)$,

$$L_1(f) = \int_0^{2\pi} f(\theta)d\theta$$

e $y_0 = 2, y_1 = -6$ *non è un problema di interpolazione polinomiale.*

Si osservi che le funzioni $1, \operatorname{sen} x$ e $\cos x$, sono funzioni da $\mathbb{R}$ in $\mathbb{R}$ *linearmente indipendenti*: *se* $a_1, a_2, a_3 \in \mathbb{R}$ sono tali che per ogni $x \in \mathbb{R}$ si ha $a_1 + a_2 \operatorname{sen} x + a_3 \cos x = 0$ *allora* $a_1 = a_2 = a_3 = 0$.

3.2.2 Osservazione (Riformulazione di un problema lineare di interpolazione)

Ricordato che lo spazio vettoriale F ha dimensione d, sia $b_1(x), \ldots, b_d(x)$ una sua *base*. Allora: $g(x) = a_1 b_1(x) + \cdots + a_d b_d(x)$ *è soluzione del problema lineare di interpolazione* se e solo se:

$$\begin{aligned} L_0(g) &= a_1 L_0(b_1) + \cdots + a_d L_0(b_d) = y_0 \\ &\vdots \\ L_k(g) &= a_1 L_k(b_1) + \cdots + a_d L_k(b_k) = y_k \end{aligned}$$

ovvero *se e solo se la colonna dei coefficienti*:

$$\begin{bmatrix} a_1 \\ \vdots \\ a_d \end{bmatrix} \in \mathbb{R}^d$$

è soluzione del sistema di equazioni lineari:

$$\begin{bmatrix} L_0(b_1) & L_0(b_2) & \cdots & L_0(b_d) \\ \vdots & \vdots & & \vdots \\ L_k(b_1) & L_k(b_2) & \cdots & L_k(b_d) \end{bmatrix} z = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}$$

Si osservi che si ottengono $k+1$ equazioni (*tante quante sono le condizioni* richieste a g) e il numero di incognite è d (*pari alla dimensione di* F), ovvero la matrice del sistema è $(k+1) \times d$. Inoltre, poiché $b_1(x), \ldots, b_d(x)$ è una base di F, l'insieme delle soluzioni del problema lineare di interpolazione e quello delle soluzioni del sistema di equazioni lineari sono in *corrispondenza biunivoca.*

3.2.3 Esempio

Determinare gli elementi di $P_2(\mathbb{R})$ che verificano le condizioni:

$$p(1) = 2 \quad , \quad \int_0^6 p(t)\, dt = 0$$

Soluzione

Si verifica che il problema posto è lineare di interpolazione, dunque risolubile studiando un sistema di equazioni lineari. Il sistema risulta di *due equazioni* in *tre incognite* quindi il problema può avere *zero* soluzioni oppure *infinite*.

Si consideri la base di Newton di $P_2(\mathbb{R})$:

$$1\ ,\ x-1\ ,\ (x-1)x$$

Le condizioni si traducono nel sistema di *due* equazioni in *tre* incognite:

$$\begin{bmatrix} 1 & 0 & 0 \\ 6 & 12 & 54 \end{bmatrix} z = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Le soluzioni del sistema sono:

$$\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 9 \\ -2 \end{bmatrix} \quad , \quad t \in \mathbb{R}$$

dunque le (infinite) soluzioni del problema lineare di interpolazione sono:

$$p_t(x) = 2 \cdot 1 - (1 + 2\,t) \cdot (x-1) - 2\,t \cdot (x-1)x \quad , \quad t \in \mathbb{R}$$

Esercizi

E6 Studiare i problemi lineari di interpolazione proposti nell'Esempio 3.2.1, punti (2), (3) e (4).

E7 Determinare gli elementi $p \in P_2(\mathbb{R})$ che verificano le condizioni:

$$p(0) = 0 \quad , \quad p'(0) = 0 \quad , \quad p(1) = 1$$

Sia q uno di essi. Dimostrare che la funzione:

$$f(x) = \begin{cases} 0 & \text{per } x < 0 \\ q(x) & \text{per } x \geqslant 0 \end{cases}$$

è derivabile per ogni $x \in \mathbb{R}$ e la funzione derivata prima è continua.

E8 ★ Sia T un numero reale positivo. Per ogni funzione continua $f : \mathbb{R} \to \mathbb{R}$ ed ogni numero reale x sia:

$$\bar{f}(x) = \frac{1}{T} \int_x^{x+T} f(t)\,dt$$

il *valor medio* di f sull'intervallo $[x, x+T]$.

(a) Verificare che per ogni $x \in \mathbb{R}$ l'applicazione $L : C(\mathbb{R}) \to \mathbb{R}$ definita da $L(f) = \bar{f}(x)$ è *lineare* su $\mathbb{R}$.

(b) Assegnati numeri reali y_0 e y_1, determinare gli elementi $p \in P_1(\mathbb{R})$ che verificano le condizioni:

$$\bar{p}(0) = y_0 \qquad , \qquad \bar{p}(1) = y_1$$

3.3 Campionamento e ricostruzione

Alcuni interessanti problemi lineari di interpolazione hanno origine dal problema della *ricostruzione* di una funzione continua a partire da un insieme assegnato di suoi valori, detti *campioni*.[37]

Siano k un numero intero non negativo, $I \subset \mathbb{R}$ un intervallo non degenere e $t_0, \ldots, t_k$ numeri reali *distinti* in I. La funzione $c : C(I) \to \mathbb{R}^{k+1}$ definita da:

$$c(f) = \begin{bmatrix} f(t_0) \\ \vdots \\ f(t_k) \end{bmatrix}$$

si chiama *funzione di campionamento agli istanti* (di campionamento) $t_0, \ldots, t_k$. L'applicazione c risulta *lineare* e *non invertibile*.

Un'applicazione *lineare* $r : \mathbb{R}^{k+1} \to C(I)$ tale che:

$$\text{per ogni } y \in \mathbb{R}^{k+1} \text{ si ha: } c\big(r(y)\big) = y$$

ovvero tale che $r(y)$ *interpola i dati* $(t_0, y_0), \ldots, (t_k, y_k)$, si chiama *funzione di ricostruzione* (relativa a c).

3.3.1 Esempio (ricostruzione mediante interpolazione polinomiale)

Sia c la funzione di campionamento agli istanti $t_0, \ldots, t_k$. Dette $y_0, \ldots, y_k$ le componenti di $y \in \mathbb{R}^{k+1}$, la funzione $\rho : \mathbb{R}^{k+1} \to C(\mathbb{R})$ definita da:

$$\rho(y) = \text{l'elemento di } P_k(\mathbb{R}) \text{ che interpola i dati } (t_0, y_0), \ldots, (t_k, y_k)$$

è una funzione di ricostruzione relativa a c, dunque *esistono* funzioni di ricostruzione relative a c.

Infatti: Utilizzando la *forma di Lagrange* del polinomio interpolante si constata che ρ *è lineare*; inoltre, per definizione, $\rho(y)$ interpola i dati $(t_0, y_0), \ldots, (t_k, y_k)$.

Si è osservato che la funzione di campionamento c *non è invertibile* dunque, scelta comunque una funzione di ricostruzione r, esistono funzioni continue f tali che: $r\big(c(f)\big) \neq f$. Per *misurare* quanto diverse sono f e $r\big(c(f)\big)$ si adotta la definizione seguente:

3.3.2 Definizione (errore di ricostruzione)

Siano $I \subset \mathbb{R}$ un intervallo *limitato* non degenere, c la funzione di campionamento agli istanti $t_0, \ldots, t_k \in I$, r una funzione di ricostruzione relativa a c e $f \in C(I)$. Il numero reale non negativo:

$$e(f) = \max_{t \in I} \big| f(t) - r\big(c(f)\big)(t) \big|$$

si chiama *errore di ricostruzione* di f. Si osservi che $e(f) = 0$ se e solo se $f = r\big(c(f)\big)$.

Il *problema del campionamento e ricostruzione* consiste nel *determinare condizioni sufficienti a garantire un errore di ricostruzione soddisfacentemente piccolo.*

Esercizi

E9 Siano $t_0 = 0, t_1 = 1$ e $t_2 = 2$ istanti di campionamento in $I = [0, 2]$ e c la funzione di campionamento a tali istanti. Studiare il seguente problema lineare di interpolazione: determinare gli elementi $p \in P_3(\mathbb{R})$ tali che $c(p) = 0$.

E10 Si considerino la funzione c di campionamento agli istanti t_0, t_1, t_2 e, dette y_0, y_1, y_2 le componenti di $y \in \mathbb{R}^3$, la funzione $\rho : \mathbb{R}^3 \to C(\mathbb{R})$ definita da:

$$\rho(y) = \text{l'elemento di } P_2(\mathbb{R}) \text{ che interpola i dati } (t_0, y_0), (t_1, y_1), (t_2, y_2)$$

Dimostrare che assegnati $z, w \in \mathbb{R}^3$ si ha:

$$\rho(3z + 7w) = 3\,\rho(z) + 7\,\rho(w)$$

[37] Si pensi, ad esempio, alla registrazione e riproduzione di un brano musicale con tecniche *digitali*.

E11 Siano $t_0 = 0, t_1 = 1$ e $t_2 = 2$ istanti di campionamento in $I = [0, 2]$, c la funzione di campionamento a tali istanti e e_1, e_2, e_3 gli elementi della base canonica di $\mathbb{R}^3$. Per $k = 1, 2, 3$ determinare l'elemento $p_k \in P_2(\mathbb{R})$ tale che $c(p) = e_k$.

E12 ★ Siano $t_0 = 0, t_1 = 1$ e $t_2 = 2$ istanti di campionamento in $I = [0, 2]$ e $c : C(I) \to \mathbb{R}^3$ la funzione di campionamento a tali istanti. Detta c_* la *restrizione* di c a $P_2(\mathbb{R})$, dimostrare che la funzione di ricostruzione mediante interpolazione polinomiale definita nell'Esempio 3.3.1 è la *funzione inversa* di c_*.

A Ricostruzione con interpolazione polinomiale

Si consideri la funzione di ricostruzione mediante interpolazione polinomiale definita nell'Esempio 3.3.1. Si ha:

3.3.3 Teorema (errore di ricostruzione nell'interpolazione polinomiale)
Siano k un numero intero non negativo, $I \subset \mathbb{R}$ un intervallo non degenere, $t_0, \dots, t_k$ istanti di campionamento *distinti* in I. *Se* $f : I \to \mathbb{R}$ è una funzione *con derivata* $(k+1)$*-esima continua* e p_k è il polinomio che interpola i campioni di f, ovvero i dati $(t_0, f(t_0)), \dots, (t_k, f(t_k))$, *allora*:

$$\textit{Per ogni } t \in I \textit{ esiste } \theta \in I \textit{ tale che:} \qquad f(t) - p_k(t) = \frac{f^{(k+1)}(\theta)}{(k+1)!}\,(t - t_0)\cdots(t - t_k)$$

Se l'intervallo I è anche chiuso e limitato, posto $M_{k+1} = \max_{x \in I} | f^{(k+1)}(x) |$, *allora* per l'errore di ricostruzione relativo ad f si ha:

$$e(f) = \max_{t \in I} | f(t) - p_k(t) | = \max_{t \in I} \left| \frac{f^{(k+1)}(\theta)}{(k+1)!}\,(t - t_0)\cdots(t - t_k) \right| \leqslant \frac{M_{k+1}}{(k+1)!}\,(\operatorname{mis} I)^{k+1}$$

Dimostrazione: Omessa.[38] Si osservi che l'espressione della differenza $f(t) - p_k(t)$ ricorda quella della *forma di Lagrange del resto* per la formula di Taylor. Si osservi anche che θ *dipende da* t (vedere l'Esercizio *E13*).

3.3.4 Esempio
(1) Siano $I = [0, 1]$ e $f(t) = e^{-t}$. La funzione f ha derivate di ordine comunque elevato e per ogni intero non negativo j si ha $M_j = \max_{x \in I} | f^{(j)}(x) | = 1$. Inoltre $\operatorname{mis} I = 1$, dunque dal Teorema precedente, scelti *comunque* $k + 1$ istanti di campionamento distinti:

$$e(f) \leqslant \frac{1}{(k+1)!}$$

Si ha inoltre:

$$\lim_{k \to \infty} \frac{1}{(k+1)!} = 0$$

e quindi $\lim_{k \to \infty} e(f) = 0$: *per ottenere un errore di ricostruzione piccolo quanto si vuole è sufficiente utilizzare un numero opportunamente elevato di istanti di campionamento.*

(2) Siano $I = [0, 2\pi]$, ω un numero reale positivo e $f(t) = \operatorname{sen} \omega t$. La funzione f ha derivate di ordine comunque elevato e per ogni intero non negativo j si ha $M_j = \omega^j$. Inoltre $\operatorname{mis} I = 2\pi$, dunque dal Teorema 3.3.3, scelti *comunque* $k + 1$ istanti di campionamento distinti:

$$e(f) \leqslant \frac{(2\pi\omega)^{k+1}}{(k+1)!}$$

Anche in questo caso si ha (vedere l'Esercizio *E14*):

$$\lim_{k \to \infty} \frac{(2\pi\omega)^{k+1}}{(k+1)!} = 0$$

e quindi $\lim_{k \to \infty} e(f) = 0$: *per ottenere un errore di ricostruzione piccolo quanto si vuole è sufficiente utilizzare un numero opportunamente elevato di istanti di campionamento.*

[38] Vedere: M. Ciampa, "Calcolo Numerico, a.a. 2011/2012," Teorema 4.12, p.109-110. Il testo è reperibile sulla pagina web del corso.

I due esempi illustrano l'asserto seguente: *Se* f ha derivate di ordine comunque elevato e la successione M_j è tale che:

$$\lim_{k\to\infty} \frac{M_{k+1}}{(k+1)!}\,(\operatorname{mis} I)^{k+1} = 0$$

allora $\lim_{k\to\infty} e(f) = 0$ e *l'errore di ricostruzione può essere reso arbitrariamente piccolo scegliendo sufficientemente grande il numero degli istanti di campionamento, con l'unico vincolo che siano distinti.*

3.3.5 Esempio

Siano $I = [0, 1]$, f la funzione continua definita da:

$$f(t) = \begin{cases} t \operatorname{sen} \frac{\pi}{t} & \text{per } t \neq 0 \\ 0 & \text{per } t = 0 \end{cases}$$

e, per ogni numero intero k non negativo:

$$t_j = \frac{1}{j+1} \qquad j = 0, 1, 2, \ldots, k$$

gli istanti di campionamento. In questo caso il Teorema 3.3.3 non è utilizzabile (f non è derivabile per $t = 0$) ma: per ogni k e $j = 0, 1, 2, \ldots, k$ si ha $f(t_j) = 0$ e quindi per ogni numero intero k non negativo l'elemento $p_k \in P_k(\mathbb{R})$ che interpola i campioni di f, ovvero i dati $(t_0, 0), \ldots, (t_k, 0)$, è il *polinomio nullo.*[39] Dunque *per ogni* k si ha:

$$e(f) = \max_{t\in I} |\, f(t) - p_k(t) \,| = M_0 > 0$$

Per la funzione assegnata non è sufficiente aumentare il numero k *di istanti di campionamento per ottenere un errore di ricostruzione piccolo quanto si vuole.*

Questo esempio mostra che vi sono funzioni continue per le quali *non è sufficiente* utilizzare un numero di istanti di campionamento opportunamente elevato per ottenere un errore di ricostruzione piccolo quanto si vuole. In generale non solo il numero ma *anche il valore* degli istanti di campionamento ha un ruolo essenziale per rendere piccolo l'errore di ricostruzione.

3.3.6 Definizione (criterio di scelta degli istanti di campionamento)

Sia I un intervallo non degenere. Un *criterio di scelta degli istanti di campionamento in* I è *una funzione* che per ogni numero intero non negativo k restituisce un insieme di $k + 1$ istanti di campionamento *distinti* in I.

3.3.7 Esempio

(a) Il criterio di scelta degli istanti di campionamento in $I = [0, 1]$ utilizzato nell'Esempio 3.3.5 è definito per ogni numero intero non negativo k dall'insieme degli istanti:

$$t_j = \frac{1}{j+1} \qquad j = 0, 1, 2, \ldots, k$$

(b) Nel *campionamento uniforme* il criterio di scelta degli istanti di campionamento in $I = [a, b]$ è definito per ogni numero intero non negativo k dall'insieme degli istanti:

$$t_j = a + j\,\frac{b-a}{k} \qquad j = 0, 1, 2, \ldots, k$$

Sussistono gli asserti seguenti:

- *Per ogni* funzione continua $f : I \to \mathbb{R}$, *esiste* un criterio di scelta degli istanti di campionamento in I tale che $\lim_{k\to\infty} e(f) = 0$;
- *Fissato* un criterio di scelta degli istanti di campionamento in I, *esiste* una funzione continua $f : I \to \mathbb{R}$ tale che $\lim_{k\to\infty} e(f) \neq 0$.

[39] Si osservi che questo accade perchè la funzione di ricostruzione è *lineare.*

Il primo asserto afferma che *in teoria* è possibile campionare e ricostruire con errore arbitrariamente piccolo *qualunque* funzione continua. Il secondo asserto afferma però che *non esiste* un criterio di scelta degli istanti di campionamento che garantisce un errore di ricostruzione arbitrariamente piccolo campionando *ogni* funzione continua.

3.3.8 Osservazione (Condizionamento della ricostruzione con interpolazione polinomiale)
Nel campionamento e ricostruzione di una funzione continua f, i campioni sono spesso ottenuti mediante un procedimento di *misura* dei valori di f agli istanti di campionamento $t_0, \dots, t_k$. Per tenere conto dell'effetto sulla ricostruzione di inevitabili *errori* nel procedimento di acquisizione dei campioni, si studia il *condizionamento* del problema della ricostruzione, ovvero la grandezza della variazione della funzione ricostruita in termini della grandezza degli errori sui campioni.

Siano I un intervallo chiuso e limitato non degenere, k un numero intero non negativo, $t_0, \dots, t_k$ istanti di campionamento *distinti* in I ed $r : \mathbb{R}^{k+1} \to C(I)$ una funzione di ricostruzione. Dato $y \in \mathbb{R}^{k+1}$ sia $\delta \in \mathbb{R}^{k+1}$ la *perturbazione* di componenti $\delta_0, \dots, \delta_k$ e si considerino le funzioni continue $r(y)$ e $r(y+\delta)$. Scelto di misurare la variazione della funzione ricostruita con:

$$\max_{t \in I} | r(y+\delta) - r(y) |$$

per la linearità della funzione di ricostruzione si ha:

$$\max_{t \in I} | r(y+\delta) - r(y) | = \max_{t \in I} | r(\delta) |$$

Nel caso di ricostruzione mediante interpolazione polinomiale, detti $\ell_0(t), \dots, \ell_k(t)$ gli elementi della base di Lagrange di $P_k(\mathbb{R})$ relativi agli istanti $t_0, \dots, t_k$, si ottiene:

$$| r(\delta) | = | \delta_0 \, \ell_0(t) + \cdots + \delta_k \, \ell_k(t) | \leqslant | \delta_0 | \, | \ell_0(t) | + \cdots + | \delta_k | \, | \ell_k(t) |$$

Introdotta la misura della perturbazione $\| \delta \|_\infty$ si deduce:

$$| \delta_0 | \, | \ell_0(t) | + \cdots + | \delta_k | \, | \ell_k(t) | \leqslant \| \delta \|_\infty \big(| \ell_0(t) | + \cdots + | \ell_k(t) | \big)$$

da cui, posto:

$$\lambda(t_0, \dots, t_k) = \max_{t \in I} \big(| \ell_0(t) | + \cdots + | \ell_k(t) | \big)$$

si ha:

$$\max_{t \in I} | r(y+\delta) - r(y) | \leqslant \lambda(t_0, \dots, t_k) \, \| \delta \|_\infty$$

La disuguaglianza è *la migliore possibile* nel senso che: *esiste* $\delta \in \mathbb{R}^{k+1}$ *tale che*:

$$\max_{t \in I} | r(y+\delta) - r(y) | = \lambda(t_0, \dots, t_k) \, \| \delta \|_\infty$$

Infatti:

- Sia t^* tale che:

$$| \ell_0(t^*) | + \cdots + | \ell_k(t^*) | = \max_{t \in I} \big(| \ell_0(t) | + \cdots + | \ell_k(t) | \big) = \lambda(t_0, \dots, t_k)$$

- Scelto $\Delta > 0$, siano $\delta_0, \dots, \delta_k$ tali che:

$$| \delta_0 | = \cdots = | \delta_k | = \Delta \quad \text{e, per } j = 0, \dots, k: \quad \delta_j \, \ell_j(t^*) \geqslant 0$$

Allora, per ogni j, essendo $\delta_j \, \ell_j(t^*) \geqslant 0$ si ha: $\delta_j \, \ell_j(t^*) = | \delta_j \, \ell_j(t^*) | = \Delta \, | \ell_j(t^*) |$. Se ne deduce che:

$$\begin{aligned} | r(\delta)(t^*) | &= | \delta_0 \, \ell_0(t^*) + \cdots + \delta_k \, \ell_k(t^*) | = \delta_0 \, \ell_0(t^*) + \cdots + \delta_k \, \ell_k(t^*) = \\ &= | \delta_0 \, \ell_0(t^*) | + \cdots + | \delta_k \, \ell_k(t^*) | = \Delta \big(| \ell_0(t^*) | + \cdots + | \ell_k(t^*) | \big) = \\ &= \| \delta \|_\infty \, \lambda(t_0, \dots, t_k) \end{aligned}$$

Dunque *il condizionamento del problema della ricostruzione mediante interpolazione polinomiale dipende dal valore del coefficiente* $\lambda(t_0, \dots, t_k)$. Sussiste il seguente asserto: *per ogni criterio di scelta degli istanti di campionamento si ha*:

$$\lambda(t_0, \dots, t_k) > \frac{\log(k+1)}{8\sqrt{\pi}}$$

Dunque: *il condizionamento peggiora all'aumentare del numero degli istanti di campionamento.*

Ci si trova, in pratica, di fronte a due *esigenze contrastanti*: scegliere k *abbastanza elevato* da garantire un basso errore di ricostruzione e scegliere k *non troppo elevato* per non rendere troppo cattive le proprietà di condizionamento della ricostruzione.

Esercizi

E13 Si consideri il Teorema 3.3.3. Dimostrare che: *se θ non dipende da t allora f è un polinomio di grado al più $k+1$.*

E14 Sia a un numero reale positivo e j la *parte intera superiore* di $2a$. Dimostrare che, posto:

$$N = \frac{a^j}{j!}$$

per ogni numero intero $k > j$ si ha:

$$\frac{a^k}{k!} = N\,\frac{a}{j+1}\cdots\frac{a}{k} < N\,(\tfrac{1}{2})^{k-j}$$

Ne segue:

$$\lim_{k\to\infty}\frac{a^k}{k!} = 0$$

E15 Si considerino i dati dell'Esempio 3.3.4 parte (1). Determinare k in modo che $e(f) < 10^{-3}$.

E16 Siano $I = [a,b]$ un intervallo non degenere e $\ell_0(t), \ell_1(t)$ la base di Lagrange di $P_1(\mathbb{R})$ relativa agli istanti di campionamento a, b. Determinare analiticamente:

$$\lambda(a,b) = \max_{t\in I}\,\big(\,|\,\ell_0(t)\,| + |\,\ell_1(t)\,|\,\big)$$

Siano poi $I = [0,1]$ e $\ell_0(t), \ldots, \ell_3(t)$ la base di Lagrange $P_3(\mathbb{R})$ relativa agli istanti di campionamento $0, \frac{1}{3}, \frac{2}{3}, 1$. Utilizzare *Scilab* per ottenere, graficamente, una stima di:

$$\lambda(0,\tfrac{1}{3},\tfrac{2}{3},1) = \max_{t\in I}\,\big(\,|\,\ell_0(t)\,| + \cdots + |\,\ell_3(t)\,|\,\big)$$

B Ricostruzione con funzioni continue lineari a tratti

Siano $I = [a,b]$ un intervallo non degenere, $a = t_0 < t_1 < \cdots < t_k = b$ istanti di campionamento e, per $j = 1, \ldots, k$, $I_j = (t_{j-1}, t_j)$. Indichiamo con τ l'insieme aperto unione dei k intervalli $I_1, \ldots, I_k$.

Una funzione $f : [a,b] \to \mathbb{R}$ è *lineare a tratti su τ* se per ogni $j = 1, \ldots, k$ esiste $p_j \in P_1(\mathbb{R})$ tale che $f = p_j$ su I_j. Il termine "lineare a tratti" fa riferimento al grafico di f su τ che, appunto, è unione di segmenti.

3.3.9 Definizione (spazio vettoriale delle funzioni continue e lineari a tratti)

Si indica con $S(\tau)$ l'insieme di *tutte* le funzioni $f : I \to \mathbb{R}$ *continue* e *lineari a tratti su τ*. Si verifica facilmente che $S(\tau)$ è uno spazio vettoriale su $\mathbb{R}$.

3.3.10 Osservazione

(a) Dati $y_0, \ldots, y_k \in \mathbb{R}$ esiste un solo elemento di $S(\tau)$ che interpola i dati $(t_0, y_0), \ldots, (t_k, y_k)$.

Infatti: Per $j = 1, \ldots, k$ sia p_j l'unico elemento di $P_1(\mathbb{R})$ che interpola i dati (t_{j-1}, y_{j-1}), (t_j, y_j) e sia poi f la funzione continua tale che $f = p_1$ su $I_1, \ldots, f = p_k$ su I_k. Allora $f \in S(\tau)$ e $f(t_0) = y_0, \ldots, f(t_k) = y_k$. Sia inoltre g un altro elemento di $S(\tau)$ che interpola gli stessi dati. Allora $f - g \in S(\tau)$. Se fosse $f(t) - g(t) \neq 0$ per $t \in I_j$, detto q_j l'elemento di $P_1(\mathbb{R})$ che coincide con $f - g$ su I_j, si avrebbe: (a) $q_j(t) \neq 0$ e quindi $q_j \neq 0$, e (b) q_j è l'unico elemento di $P_1(\mathbb{R})$ che interpola i dati $(t_{j-1}, 0), (t_j, 0)$, ovvero $q_j = 0$: assurdo.

(b) Lo spazio $S(\tau)$ ha dimensione $k+1$.

Infatti: Per $i = 0, \ldots, k$, sia s_i l'elemento di $S(\tau)$ che vale *uno* in t_i e *zero* in tutti gli altri istanti di campionamento. Questi elementi sono univocamente determinati per quanto mostrato nel punto (a). Allora si ha:

– Se $a_0, \ldots, a_k \in \mathbb{R}$ sono coefficienti tali che $\phi = a_0 s_0 + \cdots + a_k s_k = 0$, allora $0 = \phi(t_0) = a_0$, $\ldots$, $0 = \phi(t_k) = a_k$: gli elementi $s_0, \ldots, s_k$ sono *linearmente indipendenti*.

– Sia $\sigma \in S(\tau)$. Si verifica che l'elemento $\sigma(t_0)\, s_0 + \cdots + \sigma(t_k)\, s_k \in S(\tau)$ interpola i dati $(t_0, \sigma(t_0)), \ldots, (t_k, \sigma(t_k))$. Ma anche σ interpola gli stessi dati. Per l'unicità stabilita nel punto (a) si ha:

$$\sigma = \sigma(t_0)\, s_0 + \cdots + \sigma(t_k)\, s_k$$

ovvero σ è una combinazione lineare di $s_0, \ldots, s_k$: gli elementi $s_0, \ldots, s_k$ sono *generatori di* $S(\tau)$.

Dunque: $s_0, \ldots, s_k$ sono una *base* di $S(\tau)$, che chiameremo *base canonica*.

(c) Il problema a cui si riferisce il punto (a) è *lineare di interpolazione*.

3.3.11 Osservazione (Ricostruzione con funzioni continue e lineari a tratti)
Sia c la funzione di campionamento agli istanti $t_0, \ldots, t_k$. Dette $y_0, \ldots, y_k$ le componenti di $y \in \mathbb{R}^{k+1}$, la funzione $\rho : \mathbb{R}^{k+1} \to C(I)$ definita da:

$$\rho(y) = \text{l'elemento di } S(\tau) \text{ che interpola i dati } (t_0, y_0), \ldots, (t_k, y_k)$$

è una funzione di ricostruzione relativa a c.

Infatti: Utilizzando la base canonica di $S(\tau)$ si ha:

$$\rho(y) = y_0 s_0 + \cdots + y_k s_k$$

Allora si constata facilmente che ρ *è lineare*; inoltre, per definizione, $\rho(y)$ interpola i dati $(t_0, y_0), \ldots, (t_k, y_k)$.

3.3.12 Teorema (errore di ricostruzione con funzioni continue lineari a tratti)
Se $f : I \to \mathbb{R}$ è una funzione *con derivata seconda continua* e σ è l'elemento di $S(\tau)$ che interpola i campioni di f, ovvero i dati $(t_0, f(t_0)), \ldots, (t_k, f(t_k))$, posto $M_2 = \max_{x \in I} |\, f''(x)\,|$ e $h = \max\{\, \operatorname{mis} I_1, \ldots, \operatorname{mis} I_k \,\}$, *allora* per l'errore di ricostruzione relativo ad f si ha:

$$e(f) = \max_{t \in I} |\, f(t) - \sigma(t)\,| \leqslant \frac{M_2}{8}\, h^2$$

Dimostrazione: Per ogni $j = 1, \ldots, k$ esiste $p_j \in P_1(\mathbb{R})$ tale che $\sigma = p_j$ su I_j. Allora, dal Teorema 3.3.3 sull'errore di ricostruzione con interpolazione polinomiale, per $j = 1, \ldots, k$ si ha:[40]

$$\textit{Per ogni } t \in \bar{I}_j \textit{ esiste } \theta_j \in \bar{I}_j \textit{ tale che: } f(t) - \sigma(t) = f(t) - p_j(t) = \frac{f''(\theta_j)}{2}\,(t - t_{j-1})(t - t_j)$$

Per ogni $t \in \bar{I}_j$ si ha poi:

$$\left|\, \frac{f''(\theta_j)}{2}\,(t - t_{j-1})(t - t_j) \,\right| \leqslant \frac{M_2}{2} \max_{t \in \bar{I}_j} \left|\, (t - t_{j-1})(t - t_j) \,\right|$$

e:

$$\max_{t \in \bar{I}_j} \left|\, (t - t_{j-1})(t - t_j) \,\right| = \left(\frac{t_j - t_{j-1}}{2} \right)^2 = \frac{(\operatorname{mis} I_j)^2}{4}$$

ovvero:

$$\left|\, \frac{f''(\theta_j)}{2}\,(t - t_{j-1})(t - t_j) \,\right| \leqslant \frac{M_2}{8}\,(\operatorname{mis} I_j)^2$$

Dunque:

$$\max_{t \in \bar{I}_j} |\, f(t) - \sigma(t)\,| \leqslant \frac{M_2}{8}\,(\operatorname{mis} I_j)^2$$

Infine:

$$e(f) = \max_{t \in I} |\, f(t) - \sigma(t)\,| = \max_j \max_{t \in \bar{I}_j} |\, f(t) - \sigma(t)\,| \leqslant \max_j \frac{M_2}{8}\,(\operatorname{mis} I_j)^2 = \frac{M_2}{8}\, h^2$$

[40] Se $J = (a, b)$ è un intervallo aperto, si indica con $\bar{J}$ la *chiusura* di J ovvero l'intervallo chiuso $[a, b]$.

Il Teorema 3.3.12 mostra che per la ricostruzione con funzioni continue e lineari a tratti si ha: *Se* f ha derivata seconda continua *allora* $\lim_{h\to 0} e(f) = 0$, ovvero: *per ottenere un errore di ricostruzione arbitrariamente piccolo è sufficiente utilizzare un insieme di istanti di campionamento con* h *opportunamente piccolo.*

3.3.13 Esempio

(a) Sia $[a, b]$ un intervallo non degenere. Per il criterio di scelta degli istanti di campionamento del *campionamento uniforme* (Esempio 3.3.7 punto (b)) si ha:

$$h = \frac{b-a}{k}$$

Dunque è possibile ottenere h piccolo quanto si vuole scegliendo il numero di istanti di campionamento $k+1$ opportunamente grande.

(b) Sia $[a, b] = [0, 1]$ e si consideri il criterio di scelta degli istanti di campionamento definito per ogni numero intero k da:

$$t_j = \frac{j}{j+1} \quad \text{per } j = 0, \ldots, k-1 \quad \text{e} \quad t_k = 1$$

Allora:

$$h = \begin{cases} 1 \text{ per } k = 1 \\ \frac{1}{2} \text{ per } k > 1 \end{cases}$$

Questo criterio di scelta degli istanti di campionamento *non consente* di ottenere $h < \frac{1}{2}$.

3.3.14 Esempio

Siano $f(t) = \operatorname{sen} t$ e $I = [0, 2\pi]$. Poiché $M_2 = 1$, utilizzando il criterio di scelta degli istanti di campionamento in I del *campionamento uniforme* si ha:

$$e(f) \leqslant \frac{1}{8}\left(\frac{2\pi}{k}\right)^2$$

Per ottenere un errore di ricostruzione non superiore a 10^{-n}, n numero intero positivo, è sufficiente scegliere k tale che:

$$\frac{1}{8}\left(\frac{2\pi}{k}\right)^2 \leqslant 10^{-n}$$

ovvero:

$$k^2 \geqslant \frac{1}{8}\,\frac{4\pi^2}{10^{-n}}$$

e quindi:

$$k \geqslant \frac{2\pi}{\sqrt{8}}\sqrt{10^n}$$

3.3.15 Osservazione (Condizionamento della ricostruzione con funzioni continue e lineari a tratti)

Procedendo come nell'analoga Osservazione 3.3.8 e con le medesime notazioni, nel caso di ricostruzione con funzioni continue e lineari a tratti, utilizzando la base canonica di $S(\tau)$ si ottiene:

$$|\, r(\delta)\,| = |\, \delta_0 s_0(t) + \cdots + \delta_k s_k(t)\,| \leqslant |\, \delta_0\,|\,|\, s_0(t)\,| + \cdots + |\, \delta_k\,|\,|\, s_k(t)\,|$$

Introdotta la misura della perturbazione $\|\, \delta\,\|_\infty$ si deduce:

$$|\, \delta_0\,|\,|\, s_0(t)\,| + \cdots + |\, \delta_k\,|\,|\, s_k(t)\,| \leqslant \|\, \delta\,\|_\infty \big(|\, s_0(t)\,| + \cdots + |\, s_k(t)\,|\big)$$

Ma per ogni $t \in I$ e $j = 0, \ldots, k$ si ha: $s_j(t) \geqslant 0$, dunque:

$$|\, s_0(t)\,| + \cdots + |\, s_k(t)\,| = s_0(t) + \cdots + s_k(t)$$

e si constata inoltre che:

$$s_0(t) + \cdots + s_k(t) = 1$$

Allora:

$$\max_{t\in I} |\, r(y+\delta) - r(y)\,| = \max_{t\in I} |\, r(\delta)\,| \leqslant \|\,\delta\,\|_\infty$$

Questa disuguaglianza mostra che *il condizionamento del problema della ricostruzione con funzioni continue e lineari a tratti è sempre buono.*

Esercizi

E17 Sia $I = [0,2]$, $\tau = (0,1) \cup (1,2)$ e $f : I \to \mathbb{R}$ la funzione continua e lineare a tratti definita da:

$$f(t) = \begin{cases} 1+t & \text{per } t \in (0,1) \\ 3-t & \text{per } t \in (1,2) \end{cases}$$

Determinare $f(0), f(1)$ e $f(2)$.

E18 Siano $I = [0,1]$ e $t_0 = 0, t_1 = \frac{1}{2}, t_2 = 1$ gli istanti di campionamento. Determinare gli elementi $\sigma \in S(\tau)$ che verificano le condizioni:

$$\int_0^{\frac{1}{2}} \sigma(x)\,dx = 0 \quad , \quad \sigma(0) = 1 \quad , \quad \int_{\frac{1}{2}}^{1} \sigma(x)\,dx = -1$$

E19 Siano $I = [0,4]$ e $\tau = (0,1) \cup (1,2) \cup (2,3) \cup (3,4)$. Detta $s_0, \ldots, s_4$ la base canonica di $S(\tau)$, disegnare il grafico di $\sigma = 4s_0 - s_1 + 2s_2 + s_3 - 2s_4$.

E20 ★ Dimostrare che *la disuguaglianza finale dell'Osservazione 3.3.15 è la migliore possibile* nel senso che: *esiste* $\delta \in \mathbb{R}^{k+1}$ *tale che*:

$$\max_{t\in I} |\, r(y+\delta) - r(y)\,| = \|\,\delta\,\|_\infty$$

E21 Siano $I = [0,1]$ e $f(t) = e^{-t}$. Scelto di utilizzare il campionamento uniforme e la ricostruzione con funzioni continue e lineari a tratti, determinare il numero di istanti di campionamento in modo che $e(f) < 10^{-3}$. Confrontare la risposta con quella dell'Esercizio *E15*. Discutere il risultato del confronto.

---

4 Approssimazione nel senso dei minimi quadrati

In questo Capitolo si affrontano due problemi che richiedono di determinare un vettore che rende *minimo* il valore di una *somma di quadrati*: il problema di determinare la soluzione di un sistema di equazioni lineari *nel senso dei minimi quadrati* ed il problema di determinare la funzione che meglio approssima dati assegnati *nel senso dei minimi quadrati*.

4.0.1 Definizione (Soluzione di un sistema nel senso dei minimi quadrati)
Siano $A \in \mathbb{R}^{n\times k}$ e $b \in \mathbb{R}^n$. Un vettore $x^* \in \mathbb{R}^k$ è una *soluzione del sistema* $Ax = b$ *nel senso dei minimi quadrati* se verifica una delle tre proprietà *equivalenti*:

(A) Per ogni $y \in \mathbb{R}^k$ si ha: $\| Ax^* - b \|_2 \leqslant \| Ay - b \|_2$;

(B) Per ogni $y \in \mathbb{R}^k$ si ha: $\| Ax^* - b \|_2^2 \leqslant \| Ay - b \|_2^2$;

(C) Il vettore x^* è un *minimo assoluto* della funzione $n : \mathbb{R}^k \to \mathbb{R}$, *norma del residuo*, definita da:

$$n(x) = \| Ax - b \|_2$$

Si osservi che *se* x^* *è soluzione di* $Ax = b$ *allora è anche soluzione di* $Ax = b$ *nel senso dei minimi quadrati*. Infatti: $n(x^*) = 0$ e per ogni $x \in \mathbb{R}^k$ si ha $n(x) \geqslant 0$.

4.0.2 Definizione (Funzione che meglio approssima dati assegnati nel senso dei minimi quadrati)
Siano I un intervallo non degenere, F un sottospazio vettoriale *di dimensione finita* dello spazio delle funzioni continue da I in $\mathbb{R}$, $x_0, \ldots, x_k$ numeri reali in I e $y_0, \ldots, y_k$ numeri reali. Un elemento f^* di F è una *funzione che meglio approssima i dati* $(x_0, y_0), \ldots, (x_k, y_k)$ *nel senso dei minimi quadrati* se: *Per ogni* $f \in F$ *si ha*:

$$\big(f^*(x_0) - y_0\big)^2 + \cdots + \big(f^*(x_k) - y_k\big)^2 \leqslant \big(f(x_0) - y_0\big)^2 + \cdots + \big(f(x_k) - y_k\big)^2$$

ovvero se f^* è un *minimo assoluto* della funzione $\mathrm{SQ} : F \to \mathbb{R}$, *scarto quadratico*, definita da:

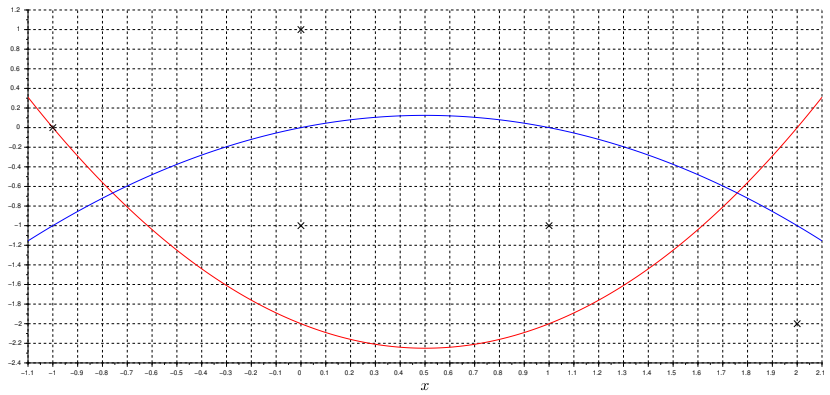
$$\mathrm{SQ}(f) = \big(f(x_0) - y_0\big)^2 + \cdots + \big(f(x_k) - y_k\big)^2$$

Interpretando i dati come coordinate di punti in un piano cartesiano, lo scarto quadratico ha un semplice *significato geometrico*. Il valore $\mathrm{SQ}(f)$ è somma di $k+1$ addendi. Il j-esimo addendo è il quadrato della lunghezza del segmento individuato, sulla retta verticale di ascissa x_j, dal valore $f(x_j)$ e dall'ordinata del j-esimo dato y_j. Il valore $\mathrm{SQ}(f)$ *è una misura dello scostamento del grafico di* $f(x)$ *dai dati*.

Si osservi che, contrariamente a quanto richiesto nei problemi di interpolazione, i numeri reali $x_0, \ldots, x_k$ *non* devono necessariamente essere distinti.

Esercizi

E1 Nella figura seguente sono rappresentati i dati (×), il grafico della funzione f (in rosso) e quello della funzione g (in blu). Determinare $\mathrm{SQ}(f)$, $\mathrm{SQ}(g)$ e decidere quale funzione tra f e g meglio approssima i dati nel senso dei minimi quadrati.



E2 ★ Dimostrare che $\mathrm{SQ}(f) = 0$ *se e solo se* f interpola i dati.

4.1 Migliore approssimazione in spazi con prodotto scalare

Alla ricerca delle soluzioni di un sistema nel senso dei minimi quadrati e delle funzioni che meglio approssimano i dati nel senso dei minimi quadrati premettiamo la nozione di migliore approssimazione in uno spazio con prodotto scalare e la sua ricerca.

Sia V uno spazio vettoriale su $\mathbb{R}$ con *prodotto scalare*. Indicato con $a \cdot b$ il prodotto scalare di a e b, si indica con $\| a \|$ la norma di a indotta dal prodotto scalare:

$$\| a \| = \sqrt{a \cdot a}$$

Siano poi W un sottospazio vettoriale di V di *dimensione finita* e v un elemento di V.

4.1.1 Definizione (Migliore approssimazione in uno spazio con prodotto scalare)
Un elemento w^* di W è una *migliore approssimazione di v in W* se verifica una delle due proprietà *equivalenti*:

(A) Per ogni $w \in W$ si ha: $\| v - w^* \| \leqslant \| v - w \|$;

(B) Il vettore w^* è un *minimo assoluto* della funzione $d : W \to \mathbb{R}$, *distanza da v*, definita da:

$$d(w) = \| v - w \|$$

Come vedremo, la nozione di migliore approssimazione in uno spazio con prodotto scalare è strettamente connessa a quella di *proiezione ortogonale*:

4.1.2 Definizione (Proiezione ortogonale)
Un elemento w^* di W è *proiezione ortogonale* di v su W se il vettore $v - w^*$ è ortogonale a W, ovvero se:

$$\textit{Per ogni } w \in W \textit{ si ha: } (v - w^*) \cdot w = 0 \qquad (*)$$

Sia $w_1, \ldots, w_j$ un insieme di *generatori* di W. La condizione $(*)$ è *equivalente* a:

$$(v - w^*) \cdot w_i = 0 \qquad i = 1, \ldots, j$$

Allora una combinazione lineare $a_1 w_1 + \cdots + a_j w_j$ è proiezione ortogonale di v su W *se e solo se* per $i = 1, \ldots, j$ si ha: $\big(v - (a_1 w_1 + \cdots + a_j w_j)\big) \cdot w_i = 0$, ovvero *se e solo se*:

$$\begin{array}{ccccccc} v \cdot w_1 & = & (a_1 w_1 + \cdots + a_j w_j) \cdot w_1 & = & a_1 (w_1 \cdot w_1) + \cdots + a_j (w_j \cdot w_1) & & \\ & \vdots & & \vdots & & & (**) \\ v \cdot w_j & = & (a_1 w_1 + \cdots + a_j w_j) \cdot w_j & = & a_1 (w_1 \cdot w_j) + \cdots + a_j (w_j \cdot w_j) & & \end{array}$$

Posto:

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_j \end{bmatrix} \quad , \quad F = \begin{bmatrix} w_1 \cdot w_1 & \cdots & w_j \cdot w_1 \\ \vdots & & \vdots \\ w_1 \cdot w_j & \cdots & w_j \cdot w_j \end{bmatrix} \quad , \quad c = \begin{bmatrix} v \cdot w_1 \\ \vdots \\ v \cdot w_j \end{bmatrix}$$

la condizione $(**)$ si riformula: *una combinazione lineare $a_1 w_1 + \cdots + a_j w_j$ è proiezione ortogonale di v su W se e solo se la colonna a è soluzione del sistema $Fx = c$*, detto *sistema delle equazioni normali*.

4.1.3 Teorema (di esistenza ed unicità della proiezione ortogonale)
Esiste un solo elemento di W proiezione ortogonale di v su W.

Dimostrazione. Sia $w_1, \ldots, w_r$ una *base ortonormale* di W (sicuramente esistente...). Allora la matrice F del sistema delle equazioni normali è la matrice identica ed il sistema ha una sola soluzione:

$$a = \begin{bmatrix} v \cdot w_1 \\ \vdots \\ v \cdot w_r \end{bmatrix}$$

L'unico elemento di W proiezione ortogonale di v su W è:

$$w^* = v \cdot w_1 \, w_1 + \cdots + v \cdot w_r \, w_r$$

Si osservi che il Teorema appena dimostrato prova che *le equazioni normali $Fx = c$ hanno sempre almeno una soluzione.* Infatti: *le componenti di ciascuna soluzione delle equazioni normali sono coordinate, rispetto ai generatori di W, della proiezione ortogonale.* L'esistenza della proiezione ortogonale w^* garantisce l'esistenza di almeno una combinazione lineare dei generatori di W tale che $a_1 w_1 + \cdots + a_j w_j = w^*$. Inoltre, se $w_1, \ldots, w_j$ sono generatori *linearmente dipendenti* (e quindi *non* una base) di W allora esistono *infinite* combinazioni lineari che generano lo stesso elemento w^*. In tal caso *il sistema $Fx = c$ ha infinite soluzioni* (in particolare: la matrice F *non* è invertibile). Se, invece, gli elementi sono generatori *linearmente indipendenti* (e quindi *una base*) di W allora esiste *una sola* combinazione lineare che genera l'elemento w^*. In tal caso *il sistema $Fx = c$ ha una sola soluzione* (in particolare: la matrice F *è* invertibile).

4.1.4 Teorema (di esistenza ed unicità della migliore approssimazione)
La proiezione ortogonale w^* di v su W è l'*unica* migliore approssimazione di v in W, ovvero: w^* è *l'elemento di W più vicino a v.*

Dimostrazione. Per ogni $w \in W$ si ha:

$$\| v - w \|^2 = \| (v - w^*) + (w^* - w) \|^2$$

Il primo addendo, per definizione, è ortogonale a W ed il secondo è un elemento di W. Allora, per il Teorema di Pitagora:

$$\| (v - w^*) + (w^* - w) \|^2 = \| v - w^* \|^2 + \| w^* - w \|^2$$

e quindi:

$$\| v - w \|^2 = \| v - w^* \|^2 + \| w^* - w \|^2$$

Poiché $\| w^* - w \|^2 \geqslant 0$ e si ha $\| w^* - w \|^2 = 0$ *se e solo se* $w = w^*$, allora:

(a) $\| v - w \|^2 \geqslant \| v - w^* \|^2$ ovvero $\| v - w \| \geqslant \| v - w^* \|$: w^* è *una* migliore approssimazione;

(b) $\| v - w \|^2 = \| v - w^* \|^2$, ovvero $\| v - w \| = \| v - w^* \|$, *se e solo se* $w = w^*$: w^* è *l'unica* migliore approssimazione.

4.1.5 Esercizio
(1) Siano $V = \mathbb{R}^4$ con prodotto scalare canonico,

$$W = \text{span}\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\} \qquad \text{e} \qquad v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

Determinare la migliore approssimazione di v in W.

Soluzione: Occorre determinare la proiezione ortogonale w^* di v su W. Poiché si ha *un solo* generatore di W, le equazioni normali si riducono ad *una* equazione in *una* incognita. Si ha:

$$F = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = 2 \qquad , \qquad c = v \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = 1$$

e le equazioni normali sono: $2x = 1$. Si ottiene l'unica soluzione $a = \frac{1}{2}$ da cui:

$$w^* = \tfrac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Per il Teorema di esistenza ed unicità della migliore approssimazione in uno spazio con prodotto scalare, la migliore approssimazione di v in W è w^*.

(2) Siano $V = \mathbb{R}^4$ con prodotto scalare canonico,

$$W = \operatorname{span}\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} \} \qquad \text{e} \qquad v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

Determinare la migliore approssimazione di v in W.

Soluzione: Come nel caso precedente, occorre determinare la proiezione ortogonale w^* di v su W. Poiché si hanno *due* generatori di W, le equazioni normali sono un sistema di *due* equazioni in *due* incognite. Poiché i generatori sono *dipendenti*, il sistema ha *infinite* soluzioni. Si ha:

$$F = \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix} \qquad , \qquad c = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

e le equazioni normali sono: $Fx = c$. L'insieme delle soluzioni è:

$$S = \left\{ \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \lambda \in \mathbb{R} \right\}$$

Tutti gli elementi di S individuano lo *stesso* elemento di W:

$$(\tfrac{1}{2} + 2\lambda) \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} - \lambda \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} = \tfrac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = w^*$$

Si ottiene, come giusto, *lo stesso elemento di* W ottenuto nel punto precedente. Per il Teorema di esistenza ed unicità della migliore approssimazione in uno spazio con prodotto scalare, la migliore approssimazione di v in W è w^*.

(3) Siano $I = [0, 2\pi]$ e $V = C(I)$ con prodotto scalare definito da:

$$f \cdot g = \frac{1}{\pi} \int_0^{2\pi} f(t) g(t) \, \mathrm{d}t$$

Siano infine:

$$W = \operatorname{span}\{ 1, \cos t, \operatorname{sen} t \} \quad \text{e} \quad v = t^2$$

Determinare la migliore approssimazione di v in W.

Soluzione: Come nel caso precedente, occorre determinare la proiezione ortogonale w^* di v su W. Si osservi che in questo esempio lo spazio vettoriale V ha *dimensione infinita*. Poiché si hanno *tre* generatori di W, le equazioni normali sono un sistema di *tre* equazioni in *tre* incognite. Poiché i generatori sono *indipendenti*, il sistema ha *una* soluzione. Si ha:

$$\begin{array}{llll} 1 \cdot 1 = \frac{1}{\pi} \int_0^{2\pi} \mathrm{d}t = 2 & & & v \cdot 1 = \frac{8}{3}\pi^2 \\ 1 \cdot \cos t = 0 & \cos t \cdot \cos t = 1 & & v \cdot \cos t = 4 \\ 1 \cdot \operatorname{sen} t = 0 & \cos t \cdot \operatorname{sen} t = 0 & \operatorname{sen} t \cdot \operatorname{sen} t = 1 & v \cdot \operatorname{sen} t = -4\pi \end{array}$$

e le equazioni normali sono:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x = \begin{bmatrix} \frac{8}{3}\pi^2 \\ 4 \\ -4\pi \end{bmatrix}$$

L'unica soluzione è:

$$a = \begin{bmatrix} \frac{4}{3}\,\pi^2 \\ 4 \\ -4\,\pi \end{bmatrix}$$

da cui:

$$w^* = \tfrac{4}{3}\,\pi^2 + 4\cos t - 4\pi \operatorname{sen} t$$

Per il Teorema di esistenza ed unicità della migliore approssimazione in uno spazio con prodotto scalare, la migliore approssimazione di v in W è w^*.

L'elemento determinato è il *minimo assoluto su* W della funzione $d : C(I) \to \mathbb{R}$ definita da:

$$d(f) = \| v - f \| = \left[\frac{1}{\pi} \int_0^{2\pi} \big(t^2 - f(t)\big)^2 \,\mathrm{d}t \right]^{\frac{1}{2}}$$

– Osservazione: *Serie di Fourier*

Sia $g \in C(I)$. Si ricordi che, introdotto in $C(I)$ il prodotto scalare definito nel punto (3) dell'Esercizio, e posto:

$$a_0 = g \cdot 1 \quad , \quad a_k = g \cdot \cos kt \quad , \quad b_k = g \cdot \operatorname{sen} kt \qquad k = 1, 2, \ldots$$

si chiama *serie di Fourier generata dalla funzione* g la *serie*:

$$\tfrac{1}{2}\,a_0 + \sum_{k=1}^{\infty} \big(a_k \cos kt + b_k \operatorname{sen} kt \big)$$

ovvero, introdotte le *somme parziali*:

$$s_0(t) = \tfrac{1}{2}\,a_0 \quad , \quad s_j(t) = \tfrac{1}{2}\,a_0 + \sum_{k=1}^{j} \big(a_k \cos kt + b_k \operatorname{sen} kt \big) \qquad j = 1, 2, \ldots$$

la *successione*: $s_0(t), s_1(t), \ldots$

Posto:

$$W_k = \operatorname{span} \{ 1, \cos t, \operatorname{sen} t, \ldots, \cos kt, \operatorname{sen} kt \} \qquad k = 0, 1, 2, \ldots$$

la migliore approssimazione w_k^* di g in W_k, ovvero il minimo assoluto *su* W_k della funzione $d : C(I) \to \mathbb{R}$ definita da:

$$d(f) = \| g - f \| = \left[\frac{1}{\pi} \int_0^{2\pi} \big(g(t) - f(t)\big)^2 \,\mathrm{d}t \right]^{\frac{1}{2}}$$

è s_k. Poiché $W_0 \subset W_1 \subset W_2 \subset \cdots$, la successione $d(s_0), d(s_1), d(s_2), \ldots$ è *non crescente*, dunque *convergente*. Un risultato classico dell'Analisi Matematica mostra che $\lim d(s_k) = 0$, risultato che si esprime anche con l'asserto:

$$g(t) = \tfrac{1}{2}\,a_0 + \sum_{k=1}^{\infty} \big(a_k \cos kt + b_k \operatorname{sen} kt \big) \qquad \textit{nel senso della convergenza in media}$$

Si osservi che la sola continuità di $g(t)$ *non assicura* la validità *puntuale* dell'uguaglianza in tutti i punti dell'intervallo I. Si consideri ad esempio il caso $I = [0, 2\pi]$ e $g(t) = t^2$. Per ogni j si ha $s_j(0) = s_j(2\pi)$ perciò se $\lim s_j(0) = g(0)$ allora $\lim s_j(2\pi) = g(0)$ Ma $g(0) \neq g(2\pi)$, dunque la convergenza puntuale della serie *non può sussistere* sull'intero intervallo I.

Lo studio della convergenza della successione $s_0(t), s_1(t), \ldots$ — ovvero del *significato* dell'uguaglianza precedente — è argomento vasto e non elementare.[41]

[41] Si veda, ad esempio: J.E. Marsden: *Elementary Classical Analysis*, Capitolo 10.

E3 ★ Sia $W = \text{span}\{ w_1, \ldots, w_j \}$ un sottospazio vettoriale di V, spazio vettoriale su $\mathbb{R}$ con prodotto scalare. Dimostrare che $v \in V$ è ortogonale a W se e solo se:

$$v \cdot w_i = 0 \qquad i = 1, \ldots, j$$

E4 ★ Sia V uno spazio vettoriale su $\mathbb{R}$ con prodotto scalare. Dimostrare il Teorema di Pitagora: *Siano a e b due elementi di V. Allora*:

$$a \cdot b = 0 \qquad \Rightarrow \qquad \| a + b \|^2 = \| a \|^2 + \| b \|^2$$

E5 Si consideri un sistema di riferimento cartesiano nello spazio. Siano poi π il piano di equazione $3x_1 - x_2 = 0$ e P il punto di coordinate $(4, 1, 0)$.

(a) Verificare che $P \notin \pi$.

Posto $V = \mathbb{R}^3$ con prodotto scalare canonico, $W = \{ x \in \mathbb{R}^3 : 3x_1 - x_2 = 0 \}$ e:

$$v = \begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix}$$

(b) Determinare una base di W;

(c) Determinare la migliore approssimazione w^* di v in W;

(d) Il punto P^* di coordinate le componenti di w^* è il punto di π più vicino a P. Determinare la distanza di P da π.

4.2 Calcolo delle soluzioni di un sistema nel senso dei minimi quadrati

Siano $A \in \mathbb{R}^{n \times k}$ di colonne $a_1, \ldots, a_k \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ e si consideri $\mathbb{R}^n$ con prodotto scalare canonico. Allora:

– Le soluzioni di $Ax = b$ nel senso dei minimi quadrati sono le *coordinate* rispetto ad $a_1, \ldots, a_k$ della migliore approssimazione di b in $\text{span}\{ a_1, \ldots, a_k \}$, ovvero della proiezione ortogonale di b su $\text{span}\{ a_1, \ldots, a_k \}$.

Infatti: La migliore approssimazione di b in $\text{span}\{ a_1, \ldots, a_k \}$ è il minimo assoluto y^* della funzione $d : \text{span}\{ a_1, \ldots, a_k \} \to \mathbb{R}$ definita da:

$$d(y) = \| b - y \| = \| y - b \|$$

Allora, posto $y^* = Ax^*$, per ogni $x \in \mathbb{R}^k$ si ha $Ax \in \text{span}\{ a_1, \ldots, a_k \}$ e quindi:

$$\| Ax - b \|_2 = \| Ax - b \| \geqslant \| y^* - b \| = \| Ax^* - b \|_2$$

Dunque, per definizione, x^* è una soluzione di $Ax = b$ nel senso dei minimi quadrati. Le componenti di x^* sono coordinate rispetto ad $a_1, \ldots, a_k$ della migliore approssimazione y^*.

– Le coordinate rispetto ad $a_1, \ldots, a_k$ della migliore approssimazione di b in $\text{span}\{ a_1, \ldots, a_k \}$, ovvero della proiezione ortogonale di b su $\text{span}\{ a_1, \ldots, a_k \}$, sono *le soluzioni* del sistema delle equazioni normali definite dai generatori $a_1, \ldots, a_k$. Ricordando che per ogni $a, b \in \mathbb{R}^n$ si ha $a \cdot b = b^\mathsf{T} a$, il sistema delle equazioni normali definite dai generatori $a_1, \ldots, a_k$ si scrive:

$$A^\mathsf{T} A x = A^\mathsf{T} b$$

ovvero:

4.2.1 Teorema

Le soluzioni di $Ax = b$ nel senso dei minimi quadrati sono le soluzioni del sistema delle equazioni normali:

$$A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$$

4.2.2 Osservazione

Si consideri il sistema delle equazioni normali relative ad $Ax = b$. Come già osservato, il sistema ha sempre *almeno una* soluzione. Si ha inoltre:

(i) La matrice $A^{\mathsf{T}}A \in \mathbb{R}^{k\times k}$ delle equazioni normali è *simmetrica* e *semidefinita positiva*.[42] La matrice è poi *definita positiva*, in particolare *invertibile*, se e solo se le colonne di A sono *linearmente indipendenti*.

Infatti: Per ogni $x \in \mathbb{R}^k$ si ha:

$$A^{\mathsf{T}}Ax \cdot x = x^{\mathsf{T}}A^{\mathsf{T}}Ax = (Ax)^{\mathsf{T}}Ax = Ax \cdot Ax = \| Ax \|^2 \geqslant 0$$

dunque la matrice $A^{\mathsf{T}}A$ è semidefinita positiva. Inoltre, si ha:

$$A^{\mathsf{T}}Ax \cdot x = \| Ax \|^2 = 0 \qquad \text{se e solo se} \qquad Ax = 0$$

e quindi $A^{\mathsf{T}}A$ è definita positiva se e solo se la condizione $Ax = 0$ è equivalente a $x = 0$, ovvero se e solo se le colonne di A sono linearmente indipendenti.

(ii) Sia $S(A, b) \subset \mathbb{R}^k$ l'insieme delle soluzioni di $Ax = b$ nel senso dei minimi quadrati. Sussistono i risultati seguenti:[43]

- Esiste *un solo* elemento di minima norma in $S(A, b)$.
- La funzione da $\mathbb{R}^n$ in $\mathbb{R}^k$ che associa a b l'elemento di minima norma in $S(A, b)$ è un'*applicazione lineare*. La matrice $k \times n$ che la definisce si chiama *pseudoinversa di* A e si indica con A^+.

Se $n \geqslant k$ e le colonne di A sono linearmente indipendenti allora $S(A, b)$ ha un solo elemento:

$$x^* = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}b$$

Dunque x^* è l'elemento di minima norma in $S(A, b)$ e la matrice pseudoinversa di A è:

$$A^+ = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$$

Se, inoltre, $n = k$ allora $A^+ = A^{-1}$.

La ricerca delle soluzioni del sistema $Ax = b$ nel senso dei minimi quadrati è dunque ricondotta alla costruzione e soluzione delle equazioni normali. Un procedimento *numericamente* preferibile alla determinazione della soluzione delle equazioni normali si ottiene estendendo la nozione di fattorizzazione QR al caso di matrici *non quadrate*.

4.2.3 Definizione (fattorizzazione QR, caso non quadrato)

Sia $A \in \mathbb{R}^{n\times k}$ con $n \geqslant k$. La coppia U, T è una *fattorizzazione QR* di A se:

- $A = UT$
- il fattore sinistro U è una matrice $n \times k$ ad elementi reali con *colonne ortonormali* rispetto al prodotto scalare canonico in $\mathbb{R}^n$;
- il fattore destro T è una matrice $k \times k$ ad elementi reali *triangolare superiore*.

[42] Si ricordi che una matrice simmetrica $M \in \mathbb{R}^{n\times n}$ è *semidefinita positiva* se per ogni $x \in \mathbb{R}^n$ si ha $Mx \cdot x \geqslant 0$. Se, inoltre, $Mx \cdot x > 0$ per *tutti* gli $x \neq 0$, la matrice è *definita positiva*. Se $x \neq 0$ e $Mx = 0$ allora $Mx \cdot x = 0$ ed M *non è* definita positiva, ovvero: Se M è definita positiva allora $Mx = 0$ se e solo se $x = 0$, ovvero M è invertibile.

[43] Si veda, ad esempio: L. Aceto e M. Ciampa: *Complementi di Algebra e Fondamenti di Geometria, Capitolo 4, Decomposizione ai valori singolari* (`http://pagine.dm.unipi.it/~a008363/x-appunti.php`).

La ricerca di una fattorizzazione QR può essere effettuata, se le colonne di A sono linearmente indipendenti, con la procedura GS definita nel Paragrafo 2.7 del Capitolo 2. Esistono procedure per la ricerca di una fattorizzazione QR *più generali* di GS e ad essa preferibili da un punto di vista numerico. La funzione predefinita `qr` di *Scilab* realizza una di queste ultime.

– `qr`

Questa *funzione predefinita* restituisce una coppia di matrici che approssima una fattorizzazione QR di una matrice assegnata. Precisamente, se `A` è una matrice $n \times k$ con $n \geqslant k$ e:

```
[U,T] = qr(A,'e')
```

allora la coppia `U,T` *approssima* una fattorizzazione QR di `A`. Come già osservato le colonne di `A` *possono* essere linearmente dipendenti.

4.2.4 Esempio

Sia:

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

Una fattorizzazione QR si determina utilizzando opportunamente la procedura GS. Procedendo come nell'Esempio 2.30:

– *Primo passo*

Si cercano $\Omega = (\omega_1, \omega_2) \in \mathbb{R}^{3\times 2}$ a colonne ortogonali e $\Theta \in \mathbb{R}^{2\times 2}$ triangolare superiore con uno sulla diagonale tali che $\Omega\Theta = A$, ovvero tali che, dette a_1, a_2 le colonne di A:

$$\omega_1 = a_1 \qquad , \qquad \omega_1\theta_{12} + \omega_2 = a_2$$

Se esistono matrici siffatte allora, *necessariamente*:

$$\omega_1 = a_1 \qquad , \qquad \theta_{12} = \frac{\omega_1 \cdot a_2}{\omega_1 \cdot \omega_1} = 1$$

e quindi:

$$\omega_2 = a_2 - \omega_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Dunque:

$$\Omega = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad , \qquad \Theta = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

– *Secondo passo*

Si *normalizzano* le colonne di Ω lasciando inalterato il risultato del prodotto. Posto:

$$\Delta = \mathrm{diag}(\,\|\,\omega_1\,\|, \|\,\omega_2\,\|\,) = \mathrm{diag}(\,\sqrt{2}, 1\,)$$

si pone:

$$U = \Omega\Delta^{-1} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} \qquad , \qquad T = \Delta\Theta = \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \end{bmatrix}$$

Se A è una matrice a colonne linearmente indipendenti e (U, T) è una fattorizzazione QR di A allora:

$$A^+ = (A^\mathsf{T}A)^{-1}A^\mathsf{T} = (T^\mathsf{T}T)^{-1}T^\mathsf{T}U^\mathsf{T} = T^{-1}(T^\mathsf{T})^{-1}T^\mathsf{T}U^\mathsf{T} = T^{-1}U^\mathsf{T}$$

Per la matrice in esame si ha allora:

$$A^+ = \begin{bmatrix} \frac{1}{\sqrt{2}} & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

Siano A *a colonne linearmente indipendenti* e U, T una fattorizzazione QR di A. Si ha:

Il sistema delle equazioni normali $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ *è* equivalente *al sistema* $Tx = U^{\mathsf{T}}b$

Infatti, tenuto conto che $U^{\mathsf{T}}U = I$, si ricava:

$$A^{\mathsf{T}}A = T^{\mathsf{T}}T \quad \text{e} \quad A^{\mathsf{T}}b = T^{\mathsf{T}}U^{\mathsf{T}}b$$

dunque il sistema delle equazioni normali si riscrive:

$$T^{\mathsf{T}}Tx = T^{\mathsf{T}}U^{\mathsf{T}}b$$

L'asserto si ottiene considerando che se le colonne di A sono linearmente indipendenti allora *la matrice* T, e quindi T^{T}, *è invertibile*. Infatti, ragionando per assurdo: Se $Ty = 0$ per qualche $y \neq 0$ allora $Ay = UTy = 0$ per qualche $y \neq 0$, ovvero: le colonne di A sono linearmente dipendenti.

Si ha, inoltre:[44]

$$c_2(A^{\mathsf{T}}A) = \big(\, c_2(T) \,\big)^2$$

ovvero:

Le proprietà di condizionamento di T *sono* (*quasi sempre*) migliori *di quelle di* $A^{\mathsf{T}}A$

Dunque: un procedimento per la ricerca delle soluzioni del sistema $Ax = b$ nel senso dei minimi quadrati *numericamente preferibile* alla costruzione e soluzione delle equazioni normali $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ è quello di *calcolare una coppia* U, T *fattorizzazione QR di* A *e poi risolvere il sistema* $Tx = U^{\mathsf{T}}b$.

4.2.5 Osservazione (`backslash`)

Scilab ha una funzione predefinita per la ricerca delle soluzioni di un sistema di equazioni lineari: `backslash`.

– `backslash`

Questa *funzione predefinita* restituisce un vettore che approssima una soluzione o una soluzione nel senso dei minimi quadrati del sistema di equazioni lineari descritto dai dati di ingresso. Precisamente, detta u la precisione di macchina, dati `A` matrice $n \times k$ e `b` colonna ad n componenti, `backslash(A,b)` o, più usualmente, `A\b`, restituisce la colonna a k componenti così determinata:

`Se` $n = k$ `allora`:

· $[\mathtt{S}, \mathtt{D}, \mathtt{P}] = \mathtt{EGPP}(\mathtt{A})$;

· `Se` $\det \mathtt{D} = 0$ `allora: rcond` $= 0$; `altrimenti: rcond` = una stima di $c_1(\mathtt{A})^{-1}$;

· `Se rcond` $> 20u$ `allora:` $\mathtt{A\backslash b} = \mathtt{SI}\big(\mathtt{D}, \mathtt{SA}(\mathtt{S}, \mathtt{Pb})\big)$;

`Se` $n \neq k$ `oppure rcond` $\leqslant 20\,u$ `allora`:

· `A\b` = una colonna che approssima una soluzione di $\mathtt{A}x = \mathtt{b}$ nel senso dei minimi quadrati determinata utilizzando opportunamente una fattorizzazione QR di A.

4.2.6 Esempio

Siano:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

La matrice A non è invertibile ma b è uguale alla prima colonna di A e il sistema ha infinite soluzioni:

$$S(A, b) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad , \quad \lambda \in \mathbb{R}$$

Le soluzioni di $Ax = b$ nel senso dei minimi quadrati coincidono con le soluzioni.

In *Scilab* si ha:

[44] Si veda, ad esempio: M. Ciampa: *Calcolo Numerico* (`http://pagine.dm.unipi.it/~a008363/x-appunti.php`), Osservazione 5.21 punto (c).

```
-->A = [1,1;1,1]; b = [1,1]';

-->x = A \ b
Warning :
matrix is close to singular or badly scaled. rcond =    0.0000D+00
computing least squares solution. (see lsq).

 x  =

    1.
    0.
```

Dopo aver avvisato l'utilizzatore che la stima di $c_1(A)^{-1}$ è inferiore a $20u \approx 2 \cdot 10^{-15}$ (e quindi $c_1(A)$ è maggiore di $(20u)^{-1} \approx 4 \cdot 10^{14}$), *Scilab* assegna ad `x` un valore che *approssima* una delle soluzioni di $Ax = b$ nel senso dei minimi quadrati:

```
-->x == [1,0]'
 ans  =

  F
  T

-->format(25)

-->x
 x  =

    0.99999999999999889777
    0.
```

Si osservi che l'elemento di $S(A, b)$ approssimato da *Scilab non è* quello di minima norma. Tale elemento, per quanto detto al punto (ii) dell'Osservazione 4.2.2, è A^+b. La funzione predefinita `pinv` di *Scilab* restituisce un'approssimazione della matrice pseudoinversa. Si ha:

```
-->y = pinv(A)*b
 y  =

    0.5
    0.5
```

4.3 Calcolo delle funzioni che meglio approssimano dati assegnati nel senso dei minimi quadrati

Siano I un intervallo non degenere, F un sottospazio vettoriale *di dimensione finita* dello spazio delle funzioni continue da I in $\mathbb{R}$, $x_0, \ldots, x_k$ numeri reali in I e $y_0, \ldots, y_k$ numeri reali. Ricordiamo che un elemento f^* di F è una *funzione che meglio approssima i dati* $(x_0, y_0), \ldots, (x_k, y_k)$ *nel senso dei minimi quadrati* se: *Per ogni* $f \in F$ *si ha*:

$$\big(f^*(x_0) - y_0\big)^2 + \cdots + \big(f^*(x_k) - y_k\big)^2 \leqslant \big(f(x_0) - y_0\big)^2 + \cdots + \big(f(x_k) - y_k\big)^2$$

ovvero se f^* è un *minimo assoluto* della funzione *scarto quadratico* $\mathrm{SQ} : F \to \mathbb{R}$ definita da:

$$\mathrm{SQ}(f) = \big(f(x_0) - y_0\big)^2 + \cdots + \big(f(x_k) - y_k\big)^2$$

Sia $f_1, \ldots, f_j$ una *base* di F. Il problema si traduce allora nella ricerca di $a_1, \ldots, a_j \in \mathbb{R}$ tali che $a_1 f_1(x) + \cdots + a_j f_j(x)$ sia un minimo assoluto della funzione SQ. Poiché per ogni $f \in F$ si ha:

$$\mathrm{SQ}(f) = \big(f(x_0) - y_0\big)^2 + \cdots + \big(f(x_k) - y_k\big)^2 = \left\| \begin{bmatrix} f(x_0) - y_0 \\ \vdots \\ f(x_k) - y_k \end{bmatrix} \right\|_2^2$$

allora:

$$\mathrm{SQ}\big(a_1 f_1(x) + \cdots + a_j f_j(x)\big) = \left\| \begin{bmatrix} a_1 f_1(x_0) + \cdots + a_j f_j(x_0) - y_0 \\ \vdots \\ a_1 f_1(x_k) + \cdots + a_j f_j(x_k) - y_k \end{bmatrix} \right\|_2^2$$

Posto:

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_j \end{bmatrix} \quad , \quad A = \begin{bmatrix} f_1(x_0) & \cdots & f_j(x_0) \\ \vdots & & \vdots \\ f_1(x_k) & \cdots & f_j(x_k) \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}$$

si riscrive:

$$\mathrm{SQ}\big(a_1 f_1(x) + \cdots + a_j f_j(x)\big) = \| Aa - b \|_2^2$$

Osservato che A e b sono la matrice e colonna del sistema di equazioni che traduce le condizioni di *interpolazione* dei dati con elementi di F e ricordata la definizione di soluzione del sistema $Ax = b$ nel senso dei minimi quadrati si deduce che: *Le coordinate delle funzioni che meglio approssimano i dati nel senso dei minimi quadrati sono le componenti delle soluzioni nel senso dei minimi quadrati del sistema che traduce le condizioni di interpolazione dei dati.*

4.3.1 Esercizio

Determinare gli elementi di $P_1(\mathbb{R})$ che meglio approssimano i dati:

$$(0,1) \quad , \quad (0,2) \quad , \quad (1,1) \quad , \quad (2,0)$$

nel senso dei minimi quadrati.

Soluzione. Sia $1, x$ una base di $P_1(\mathbb{R})$. Il sistema che traduce le condizioni di interpolazione dei dati con un elemento di $P_1(\mathbb{R})$ è $Ax = b$ con:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix}$$

I coefficienti che individuano gli elementi di $P_1(\mathbb{R})$ che meglio approssimano i dati sono le soluzioni del sistema nel senso dei minimi quadrati. Poiché le colonne di A sono linearmente indipendenti il sistema ha *una sola* soluzione nel senso dei minimi quadrati. Il sistema delle equazioni normali è:

$$\begin{bmatrix} 4 & 3 \\ 3 & 5 \end{bmatrix} x = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

da cui:

$$x^* = \begin{bmatrix} \frac{17}{11} \\ -\frac{8}{11} \end{bmatrix}$$

e l'elemento cercato è:

$$p^*(x) = \tfrac{17}{11} - \tfrac{8}{11}\, x$$

Si osservi che, detta b^* la proiezione ortogonale di b sul sottospazio di $\mathbb{R}^4$ generato dalle colonne di A si ha: $b^* = Ax^*$ e quindi:

$$\| Ax^* - b^* \| = 0$$

Dunque: *l'elemento p^* migliore approssimazione dei dati $(x_i, b_{i+1}), i = 0, \ldots, 3$ interpola i dati $(x_i, b^*_{i+1}), i = 0, \ldots, 3$.*

Le nozioni di soluzione di un sistema di equazioni lineari nel senso dei minimi quadrati e di funzione che meglio approssima i dati nel senso dei minimi quadrati possono essere estese modificando le funzioni n e SQ con l'introduzione di un coefficiente positivo, detto *peso*, per ciascun addendo. L'esempio seguente illustra queste estensioni.

4.3.2 Esercizio

Sia F un sottospazio vettoriale di dimensione due dello spazio delle funzioni continue da I in $\mathbb{R}$ e p_0, p_1, p_2 numeri reali positivi. Determinare gli elementi di F minimi assoluti della funzione $\mathrm{SQ} : F \to \mathbb{R}$ definita da:

$$\mathrm{SQ}(f) = p_0 \left(f(x_0) - y_0\right)^2 + p_1 \left(f(x_1) - y_1\right)^2 + p_2 \left(f(x_2) - y_2\right)^2$$

Soluzione. Per ogni $f \in F$ si ha:

$$\mathrm{SQ}(f) = p_0\,(f(x_0)-y_0)^2 + p_1\,(f(x_1)-y_1)^2 + p_2\,(f(x_2)-y_2)^2 = \left\| \begin{bmatrix} \sqrt{p_0}\,\big(f(x_0)-y_0\big) \\ \sqrt{p_1}\,\big(f(x_1)-y_1\big) \\ \sqrt{p_2}\,\big(f(x_2)-y_2\big) \end{bmatrix} \right\|_2^2$$

Procedendo come sopra si ottiene che, detta f_1, f_2 una base di F e posto:

$$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad , \quad \Delta = \mathrm{diag}(\sqrt{p_0}, \sqrt{p_1}, \sqrt{p_2}) \quad , \quad A = \begin{bmatrix} f_1(x_0) & f_2(x_0) \\ f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

si riscrive:

$$\mathrm{SQ}\big(a_1 f_1(x) + a_2 f_2(x)\big) = \| \Delta A a - \Delta b \|_2^2$$

Si deduce che: *Le coordinate delle funzioni che meglio approssimano i dati nel senso dei minimi quadrati*, con pesi p_0, p_1, p_2, *sono le componenti delle soluzioni nel senso dei minimi quadrati del sistema* $\Delta A x = \Delta b$. Queste ultime sono le soluzioni del sistema delle equazioni normali:

$$A^{\mathsf{T}} \Delta^2 A x = A^{\mathsf{T}} \Delta^2 b$$

Esercizi

E6 Siano:

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

Determinare una fattorizzazione QR di A utilizzando la procedura GS ed utilizzarla per calcolare le soluzioni del sistema $Ax = b$ nel senso dei minimi quadrati.

E7 ♠ Verificare che in *Scilab* dopo gli assegnamenti:

```
-->A = [1,1,0;1,1,1]';
```

```
-->[U,T] = qr(A,'e');
```

la coppia (`U`, `T`) è un'approssimazione della fattorizzazione QR di `A`:

$$U = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & 0 \\ 0 & -1 \end{bmatrix} \quad , \quad T = \begin{bmatrix} -\sqrt{2} & -\sqrt{2} \\ 0 & -1 \end{bmatrix}$$

Confrontare la fattorizzazione con quella ottenuta nell'Esempio 4.2.4.

E8 Siano A e A^+ come nell'Esempio 4.2.4. Determinare A^+A e AA^+.

Il primo risultato giustifica il termine "pseudoinversa" utilizzato per la matrice A^+. Dimostrare che per ogni $B \in \mathbb{R}^{n\times k}$ a colonne linearmente indipendenti si ha:

$$B^+ B = I$$

E9 ♠ Utilizzare la funzione `backslash` per determinare gli elementi di $P_1(\mathbb{R})$ che meglio approssimano i dati:

$$(0,1) \quad , \quad (0,2) \quad , \quad (1,1) \quad , \quad (2,0)$$

nel senso dei minimi quadrati e disegnare, su uno stesso piano cartesiano, i dati ed il grafico dell'elemento ottenuto.

E10 Sia $F : \mathbb{R}^3 \to \mathbb{R}$ definita da:

$$F(x) = x_1^2 + 2x_2^2 + 3x_3^2 + (x_1 - x_2 + x_3 - 1)^2 + (-x_1 - 4x_2 + 2)^2$$

Determinare una matrice $A \in \mathbb{R}^{n\times 3}$ ed una colonna $b \in \mathbb{R}^n$ tali che per ogni $x \in \mathbb{R}^3$ si abbia:

$$F(x) = \| Ax - b \|_2^2$$

Determinare poi il minimo assoluto di F.

E11 Assegnato un sistema di riferimento cartesiano in un piano π, siano $c_1, \ldots, c_j \in \mathbb{R}^2$ i vettori delle coordinate di j punti distinti di π. Si consideri la funzione $\lambda : \mathbb{R}^2 \to \mathbb{R}$ definita da:

$$\lambda(x) = \| x - c_1 \|^2 + \cdots + \| x - c_j \|^2$$

- Dare un'interpretazione geometrica della funzione λ.
- Determinare $A \in \mathbb{R}^{2j\times 2}$ e $b \in \mathbb{R}^{2j}$ tali che per ogni $x \in \mathbb{R}^2$ si abbia:

$$\lambda(x) = \| Ax - b \|_2^2$$

- Siano:

$$c_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad , \quad c_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad , \quad c_3 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

  Determinare il minimo assoluto di λ.

E12 Assegnato un sistema di riferimento cartesiano in un piano π, siano $c_1, \ldots, c_j \in \mathbb{R}^2$ i vettori delle coordinate di j punti distinti di π. Assegnati $k_1, \ldots, k_j$ numeri reali *positivi*, si consideri la funzione $\mathrm{EP} : \mathbb{R}^2 \to \mathbb{R}$ definita da:

$$\mathrm{EP}(x) = \tfrac{1}{2}\, k_1 \| x - c_1 \|^2 + \cdots + \tfrac{1}{2}\, k_j \| x - c_j \|^2$$

- Dare un'interpretazione meccanica della funzione EP.
- Determinare $A \in \mathbb{R}^{2j\times 2}$ e $b \in \mathbb{R}^{2j}$ tali che per ogni $x \in \mathbb{R}^2$ si abbia:

$$\mathrm{EP}(x) = \tfrac{1}{2} \| Ax - b \|_2^2$$

- Siano:

$$c_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad , \quad c_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad , \quad c_3 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

  e $k_1 = k_2 = 1, k_3 = 4$. Determinare il minimo assoluto di EP.
