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:

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

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

[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.

Gli esercizi contrassegnati dal simbolo ★ sono leggermente più astratti rispetto agli altri. 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

Si osservi che 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 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.

Esercizi

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

E4 Determinare il numero di elementi dell'insieme:

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

E5 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.

E6 ★ 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 $F_2 \subset F_{10}$.

E7 ★ 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 $F(10,m)$.

E8 Sia $x = 3.7$ (scrittura in base dieci). Decidere se $x \in F(2,8)$.

E9 Mostrare che tutti gli elementi positivi di $F(2,4)$ con esponente maggiore o uguale a 4 sono interi, e poi determinare:

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

E10 ★ 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.

E11 Posto $\xi = 2^{-3}\ 0.1101 \in F(2,4)$, indicare per quali numeri interi n si ha $4^n\,\xi \in F(2,4)$.

E12 Si consideri $F(2,10)$. Determinare il numero di elementi positivi con esponente -6, ovvero il numero di elementi dell'insieme B_{-6}.

E13 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:

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

E14 Assegnate una base β ed una precisione m, dimostrare che:

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

E15 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 $\xi_{\min}$.

E16 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).

E17 ★ 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.

E18 ♠ 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 $\xi^*_{\min}$.

E19 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.

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!
- 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

E20 Calcolare l'arrotondato di $\frac{1}{4}$ in $F(3,2)$.

E21 ★ 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:

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

E22 Calcolare l'arrotondato di $2^2\,0.1011$ in $F(10,2)$.

E23 Calcolare l'arrotondato di $\frac{1}{2}\,\xi_{\min}$ in $F(2,5,-9,9)$.

[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.

E24 Calcolare l'arrotondato di $\frac{1}{2}\,\xi_{\min}$ in $F_d(2,5,-9,9)$.

E25 Sia rd la funzione arrotondamento in $F(10,3)$ con RTTE. Determinare tutti gli $x \in \mathbb{R}$ tali che $\mathrm{rd}(x) = 642$.

E26 Sia rd la funzione arrotondamento in $F(10,3)$ con RTTE. Determinare:

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

E27 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.

E28 Sia $M = F(10,1)$. Determinare l'arrotondato in M di $x = \frac{19}{20}$ con RTTE.

E29 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.

E30 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.

E31 ♠ Utilizzare la funzione `number_properties` per ottenere, da *Scilab* la precisione di macchina e verificare, utilizzando la funzione `frexp`, che tale precisione di macchina è 2^{-53}.

E32 Dimostrare che, detta u la precisione di macchina in $M = F(\beta, m)$, si ha:

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

E33 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 $\mathrm{rd}(\beta^{12}x) = \beta^{12}\xi$.

E34 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

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

(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 $\mathtt{SEN}(\xi) = 0$ equivale a $\mathrm{rd}(\mathrm{sen}\ \xi) = 0$ ovvero $\mathrm{sen}\ \xi = 0$, 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: $1 \in M, 4 \in M$, $\mathtt{SEN}(1) > 0$ e $\mathtt{SEN}(4) < 0$ ma per ogni $\alpha \in M$ compreso tra 1 e 4 si ha $\mathtt{SEN}(\alpha) \neq 0$).

[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).

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 √ e si ottiene, ad esempio:

```
√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 1454746218587388284504413604736328125
```

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:

```
log10(1000) = 2.999999999999999555910790149937383830547332763671875
```

che *non* coincide con l'arrotondato di $\log_{10} 1000$ in $F(2, 53)$ – infatti: $\mathrm{rd}(\log_{10} 1000) = 3$. La definizione della funzione predefinita è quindi *diversa* da quella della Definizione 0.3.1.

Si ha inoltre:

$$\sigma\big(\texttt{log10(1000)}\big) = 3 = \mathrm{rd}(\log_{10} 1000)$$

e per l'errore relativo commesso approssimando $\log_{10} 1000$ con `log10(1000)`, detta u la precisione di macchina in $F(2, 53)$ si ha:

$$\left| \frac{\pi(3) - 3}{3} \right| = \frac{2^{-51}}{3} = \tfrac{4}{3}\, u$$

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

Esercizi

E35 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: $\xi \oplus \alpha = 0$.

E36 ★ 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*:

```
x = √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:

```
x = 1.41421356237309514547462185873882845044136047363281 25
```

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:

```
x = 1.41421356237
```

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

`x` = $\log_{10} 1000$

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

`x = LOG10(`rd(1000)`)`

Tenuto conto che $1000 \in F(2,53)$, il valore di `x` dopo l'esecuzione in *Scilab* è `log10(1000)` ovvero, si veda l'Esempio 0.3.4:

```
x = 2.99999999999999995559107901499373838305473327636718 75
```

Analogamente, tenuto conto che $1000 \in F(10,12)$, il valore di `x` dopo l'esecuzione con la calcolatrice tascabile *HP 49G* è `LOG(1000)` ovvero:

```
x = 3
```

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

`u` = 2^{-53}`;`
`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$\big($rd$(2)^{\text{rd}(-53)}\big)$`;`
`a = -u;`
`b = u;`
`x = (a` $\oplus$ `b)` $\oplus$ rd(1)`;`
`y = a` $\oplus$ `(b` $\oplus$ 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 $\mathtt{y} = f(x)$ quando utilizza il tipo *numero reale* e nell'assegnamento $\mathtt{y} = \phi(x)$ 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: 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:

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

Wuest'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 $x = 0$.

[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.

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: 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.*

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 piccolo tale che*:[16]

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

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:

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

si conclude:

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

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

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

Utilizzando i risultati ottenuti:

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

e, posto:

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

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 $f(x)$.

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.)

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

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

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

  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:

  $$c_f(x) = \left| \frac{x}{\tan x} \right|$$

  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:

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

  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*:

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

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

– *Esercizio*

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

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

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

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

  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 $\operatorname{sen} x$ 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:

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

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:

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

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:

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

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:

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

[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: $\texttt{COS}(\xi_n) = \mathrm{rd}(\cos \xi_n) = 1$.

E37 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.

E38 ★ Realizzando la procedura del punto (4) dell'Esempio 0.4.1 con la calcolatrice *HP 49G* si ottiene $\mathtt{x} = \mathtt{y} = 1$. Spiegare questi risultati e poi indicare come modificare l'assegnamento che definisce il valore di `u` in modo da ottenere anche in questo caso $\mathtt{x} \neq \mathtt{y}$.

E39 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.

E40 Per ogni $x > 0$ sia $f(x) = 1/\sqrt{x}$. Determinare l'insieme di definizione e discutere stabilità ed accuratezza dell'algoritmo:

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

quando utilizzato per approssimare i valori di f.

E41 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:

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

quando utilizzato per approssimare i valori di f.

E42 ★ 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:

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

E43 Si consideri l'Esempio 0.4.12. Tenuto conto che in *Scilab* si ha: $\mathtt{\%pi} = \mathrm{rd}(\pi) < \pi$ e $\phi(\pi) = \mathtt{SEN(\%pi)} > 0$:

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

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

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

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.

E44 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:

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

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

E45 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.
