Appunti di Calcolo Numerico

Capitolo 0

Numeri in virgola mobile e precisione finita

Maurizio Ciampa

Dipartimento di Matematica - Università di Pisa

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/`).

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 quesi 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. In questo caso, ma è *quasi sempre* così, la richiesta di una scrittura posizionale 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 è *più accurata* – la limitazione sull'errore relativo è *più stringente* – della prima.

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 754-2008) che prescrive regole – ampiamente condivise – per eseguire calcoli in virgola mobile in modo che il risultato sia *indipendente* dal dispositivo di calcolo utilizzato.

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

In questo capitolo descriviamo il tipo *numero in virgola mobile e precisione finita* ed 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*. 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.

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 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:

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

0.2 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 $2^1 \leqslant \sqrt{5} < 2^2$:

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

0.3 Osservazione
La condizione $\beta^{-1} \leqslant g < 1$ è equivalente a:

$$\textit{la scrittura posizionale di } g \textit{ in base } \beta \textit{ ha la forma } 0.c_1c_2\cdots \textit{ con } c_1 \neq 0$$

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*.

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]

[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.

– 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.

0.4 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 la scrittura posizionale di g in base dieci ha *lunghezza uno*, la scrittura posizionale di g in base due ha *lunghezza infinita*.

0.5 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.*

0.6 Esempio
Si consideri $F(10, 1)$.

– Poiché $\frac{1}{100} = 10^{-1}\, 0.1$ allora $\frac{1}{100} \in F(10, 1)$. Invece: $\frac{11}{100} \notin F(10, 1)$ perchè $\frac{11}{100} = 10^0\, 0.11$ e la 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.

– I possibili valori della frazione $0.c_1$ 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.

[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.

0.7 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}.

0.8 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 ha frazione non compatibile con la 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 *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 $\lim \xi_k = +\infty$.

0.9 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]

0.10 Esempio

Si consideri $F(10,3)$.

- 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 $\pi(\xi) = -\sigma(-\xi)$.

0.11 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.

[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.

0.12 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 sottonsieme 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*.

Il sottonsieme 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 *ed elementi denormalizzati*.

0.13 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 $c_1, \ldots, c_m$ sono le cifre (in base β) della frazione di ξ. Se invece ξ ha esponente minore di $b_{\min}$ – ovvero ξ è un elemento denormalizzato – allora $c_1 = 0, c_2, \ldots, c_m$ *non sono le cifre 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 $c_1 = c_2 = \cdots = c_m = 0$.

0.14 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*, $\xi_{\max} = \max F(10, 4, -99, 99) = 10^{99}\, 0.9999$ e la funzione successore *non è definita* in $\xi_{\max}$.[5]

[5] Analogamente, la funzione predecessore *non è definita* in $\min F(10, 4, -99, 99) = -\xi_{\max}$.

– 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*.

0.15 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 $M = F(10, 12)$.

Esercizi

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

E2 Indicare quali dei seguenti numeri reali appartengono ad $F(2, 3)$: *uno*, *un terzo*, *meno un sedicesimo*, *tre sedicesimi*, *zero*, π.

E3 Determinare il numero di elementi dell'insieme:

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

E4 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.

E5 ★ 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 0.4) 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* $2^k > 10^n$; questi due asserti provano che per k sufficientemente grande si ha $2^k \notin F_{10}$, e quindi che è falso anche l'asserto $F_2 \subset F_{10}$.

[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.

E6 ★ La dimostrazione dell'asserto (1) dell'Osservazione 0.8 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 $F(10, m)$.

E7 Sia $x = 3.7$ (scrittura in base dieci). Decidere se $x \in F(2, 8)$.

E8 Mostrare che tutti gli elementi positivi di $F(2, 4)$ con esponente maggiore o uguale a 4 sono interi, e poi determinare:

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

E9 ★ Siano $\texttt{esp}, \texttt{fraz} : F(\beta, m) \setminus \{\,0\,\} \to \mathbb{R}$ le funzioni definite da:

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

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

E10 Posto $\xi = 2^{-3}\, 0.1101 \in F(2, 4)$, indicare per quali numeri interi n si ha $4^n\, \xi \in F(2, 4)$.

E11 Si consideri $F(2, 10)$. Determinare il numero di elementi positivi con esponente -6, ovvero il numero di elementi dell'insieme B_{-6}.

E12 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 $\pi(-2^{-1}\, 0.101) = -\sigma(2^{-1}\, 0.101)$. Dedurne che

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

Determinare infine:

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

E13 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 $\xi_{\min}$.

E14 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).

E15 ★ 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.

E16 ♠ 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 $\xi^*_{\min}$.

E17 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.

Sia M l'insieme dei numeri in virgola mobile e precisione finita scelto ed x un numero reale *non* in M. Se si è scelto 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.

0.16 Osservazione

Si consideri $M = F(\beta, m)$ e sia $x \notin M$ un numero reale positivo. Se, in base β, $x = \beta^b \cdot 0.c_1 c_2 \cdots$ allora, posto $\xi_- = \beta^b \cdot 0.c_1 \cdots c_m$ (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 $F(10, 3)$.

0.17 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 si sceglie come arrotondato di x quello tra i due elementi adiacenti ad x che ha cifra c_m *pari* — questa scelta, utilizzabile *solo* se la base β è *pari* ed M è un insieme con *esponente non limitato* oppure con *esponente limitato ed elementi denormalizzati*, sarà indicata con la sigla RTTE[9] 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.

0.18 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 0.4) si ha:

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

dunque $x \notin M$. Inoltre, posto $\xi_- = 2^{-3}\, 0.11$ (l'elemento di M ottenuto da x *troncando* la scrittura della frazione alla seconda cifra) e $\xi_+ = \sigma(\xi_-) = 2^{-2}\, 0.10$ si ha:

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

ovvero ξ_- e ξ_+ sono gli elementi di M adiacenti ad x. 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_-.

0.19 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*. Infatti, se $\xi = \mathrm{rd}(x) \neq 0$ 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 $\mathrm{rd}(x) = \mathrm{rd}(y)$; se $y > \sup I$ allora $\mathrm{rd}(x) < \mathrm{rd}(y)$.

[9] Le sigle RTTE e RTTA sono abbreviazioni, rispettivamente, dei termini *round ties to even* e *round ties to away* utilizzati nello standard IEEE Std 754-2008.

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

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

Esercizi

E18 Calcolare l'arrotondato di $\frac{1}{4}$ in $F(3, 2)$.

E19 ★ 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:

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

E20 Calcolare l'arrotondato di $2^2\, 0.1011$ in $F(10, 2)$.

E21 Calcolare l'arrotondato di $\frac{1}{2}\,\xi_{\min}$ in $F(2, 5, -9, 9)$.

E22 Calcolare l'arrotondato di $\frac{1}{2}\,\xi_{\min}$ in $F_d(2, 5, -9, 9)$.

E23 Sia rd la funzione arrotondamento in $F(10, 3)$ con RTTE. Determinare tutti gli $x \in \mathbb{R}$ tali che $\mathrm{rd}(x) = 642$.

E24 Sia rd la funzione arrotondamento in $F(10, 3)$ con RTTE. Determinare:

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

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.

0.20 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*.

0.21 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 $F(10, 3)$.

0.22 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 0.11. 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 $x < 0$.

0.23 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 0.22 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*.

0.24 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 0.22 si esprimono:

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

e, quindi:

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

0.25 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 precisone di macchina quanto più *stringente* è, in base al Teorema 0.22, 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.

0.26 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 $x \mapsto |\,\delta(x)\,|$.

0.27 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 0.22.)

Esercizi

E25 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 0.22 e le tesi del Teorema 0.27.

E26 ♠ Utilizzare la funzione `number_properties` per creare una variabile di nome `u` e verificare, utilizzando la funzione `frexp`, che la precisione di macchina in *Scilab* è 2^{-53}.

E27 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 $\mathrm{rd}(\beta^{12}x) = \beta^{12}\xi$.

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.

0.28 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*

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

  Sono le *restrizioni* ad $M \times M$ delle corrispondenti funzioni sui numeri reali.[10]

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."[11]

0.29 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) "lo 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

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

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

[11] Le definizioni date delle funzioni predefinite corrispondenti alle operazioni aritmetiche e quelle dei confronti rispecchiano fedelmente la realtà (lo standard IEEE Std 754–2008 le *impone*). Invece, le definizioni date di funzioni predefinite corrispondenti alle funzioni elementari possono essere *troppo stringenti* (lo standard le *raccomanda* – ma non *impone*): in casi concreti le funzioni predefinite corrispondenti alle funzioni elementari possono essere definite in modo leggermente diverso, quindi "peggiore."

(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'unità non è unica:" 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: $\xi = 10^0\, 0.49; \xi \otimes 10^0\, 0.99 = \xi$).

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

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

può essere vuoto o avere più di un elemento: "l'inverso può non esistere o non essere unico" (ad esempio: $\xi = 10^0\, 0.20, \xi \otimes 10^1\, 0.50 = 1$ e $\xi \otimes 10^1\, 0.51 = 1$; $\xi = 10^1\, 0.89, \xi \otimes 10^0\, 0.11 = 10^0\, 0.98 < 1$ e $\xi \otimes 10^0\, 0.12 = 10^1\, 0.11 > 1$ e quindi, per la monotonia di $\otimes$ — (M.3) —, ξ non ha 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 $\mathrm{rd}(\text{sen } \xi) = 0$ ovvero 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 $\texttt{SEN}(\alpha) \neq 0$).

0.30 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 0.22, 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 $\xi^*_{\min} \leqslant |\, x \,| \leqslant \xi_{\max}$.

0.31 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:

```
√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:

```
sqrt(2) = 2^1 0.10110101000001001111001100110011111111001110111100 1101
```

In questi casi la Definizione 0.28 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) = 2.99999999999999955591079014993738383054733276367 1875
```

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 0.28.

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 $F(2, 53)$:

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

Questo valore è *leggermente più grande* del massimo conseguente alla Definizione 0.28.

Esercizi

E28 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: $\xi \oplus \alpha = 0$.

E29 ★ 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

In questa sezione completiamo la descrizione del 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 *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*.

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*.

0.32 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 0.25. Analogamente, dopo l'esecuzione della procedura in *Scilab* si ottiene:

`y = 0.43301270189 22192 98294151030859211459755897521972656 25`

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

`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 =` $\sqrt{2}$

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 0.31:

`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 0.31:

`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 0.31:

`x = 2.99999999999999955591079014993738383054733276367 1875`

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(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: vedere l'Esempio 0.29, A.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` $= \pi(1)$.

Abbiamo detto che nel Calcolo Numerico si cercano procedure da eseguire utilizzando un calcolatore che determinano approssimazioni di numeri reali e *informazioni sull'errore* commesso utilizzando i valori restituiti dalle procedure per approssimare tali numeri reali. I prossimi due esempi mostrano, in casi elementari, come ottenere le informazioni richieste.

0.33 Esempio
Per ottenere un'approssimazione del numero reale $\alpha = 1/\sqrt{2}$, si considerano le procedure seguenti (ottenute in modo ovvio):

```
A = 1 / sqrt(2);     ,     B = sqrt(2) / 2;
```

Detti, rispettivamente, a e b i valori delle variabili `A` e `B` al termine dell'esecuzione delle procedure in *Scilab*, determiniamo una *limitazione superiore* al valore assoluto dell'errore relativo (si ricordi quanto detto nell'Osservazione 0.26) commesso utilizzando ciascuno dei valori per approssimare α.
Scelto a come approssimazione di α, cerchiamo la limitazione superiore in *due passi*.

- *Primo passo*:

  Sostituendo al tipo *numero reale* il tipo *numero in virgola mobile e precisone finita* con $M = F(2, 53)$, detta `SQRT` la funzione predefinita corrispondente alla funzione elementare radice quadrata e tenuto conto che sia 1 che 2 sono elementi di $F(2, 53)$, si deduce che *Scilab* assegna ad `A` il valore:

  $$a = 1 \oslash \texttt{SQRT(2)}$$

  Utilizzando ripetutamente il Teorema 0.27 ed indicando con u la precisione di macchina in $F(2, 53)$ si ha:

  – Poiché `SQRT(2)` $= \mathrm{rd}(\sqrt{2})$, esiste $e_1 \in \mathbb{R}$ tale che:

  $$a = 1 \oslash \big((1 + e_1)\sqrt{2}\big) \qquad \text{con} \qquad |\, e_1 \,| \leqslant u$$

  – Poiché per ogni $\xi_1, \xi_2 \in F(2, 53)$ si ha $\xi_1 \oslash \xi_2 = \mathrm{rd}(\xi_1/\xi_2)$, esiste $e_2 \in \mathbb{R}$ tale che:

  $$a = \frac{1 + e_2}{(1 + e_1)\sqrt{2}} \qquad \text{con} \qquad |\, e_2 \,| \leqslant u$$

  Posto $f(x_1, x_2) = x_1/x_2$ si riscrive infine:

  $$a = f\big((1 + e_2)1, (1 + e_1)\sqrt{2}\,\big) \qquad \text{e} \qquad \alpha = f(1, \sqrt{2})$$

  Siano giunti a *rileggere* a ed α come *valori di una stessa funzione in punti vicini*. (*Esercizio*: mostrare che la distanza tra i due punti non supera $\sqrt{3}\,u$.)

- *Secondo passo*:

  Affrontiamo il seguente problema riguardante la funzione f introdotta nel primo passo: *dati* $x_1, x_2 \in \mathbb{R}$ *tali che* $f(x_1, x_2) \neq 0$ *e* $\epsilon_1, \epsilon_2 \in \mathbb{R}$ *con* $|\,\epsilon_2\,| < 1$, *determinare* $\epsilon \in \mathbb{R}$ *tale che*:

$$f\big((1+\epsilon_1)\,x_1, (1+\epsilon_2)\,x_2\big) = (1+\epsilon)\,f(x_1, x_2)$$

  Si cerca ϵ tale che:

$$\frac{(1+\epsilon_1)\,x_1}{(1+\epsilon_2)\,x_2} = (1+\epsilon)\,\frac{x_1}{x_2}$$

  ovvero tale che:

$$\frac{1+\epsilon_1}{1+\epsilon_2} = 1+\epsilon$$

  Quest'ultima relazione determina *univocamente* ϵ:

$$\epsilon = \frac{\epsilon_1 - \epsilon_2}{1+\epsilon_2}$$

Utilizzando il risultato appena ottenuto si riscrive:

$$a = f\big((1+e_2)1, (1+e_1)\sqrt{2}\,\big) = \Big(1 + \frac{e_2 - e_1}{1+e_1}\Big)\,f(1, \sqrt{2}) = \Big(1 + \frac{e_2 - e_1}{1+e_1}\Big)\,\alpha$$

che equivale a:

$$\frac{a-\alpha}{\alpha} = \frac{e_2 - e_1}{1+e_1}$$

Considerando le limitazioni $|\,e_1\,| \leqslant u$ e $|\,e_2\,| \leqslant u$ si ottiene infine:

$$\Big|\,\frac{a-\alpha}{\alpha}\,\Big| = \Big|\,\frac{e_2 - e_1}{1+e_1}\,\Big| \leqslant \frac{|\,e_2\,| + |\,e_1\,|}{|\,1+e_1\,|} \leqslant \frac{2u}{1-u} \approx 2u$$

Dunque: *il valore assoluto dell'errore relativo commesso approssimando* α *con* a *non supera una quantità che vale circa* $2u$.

Scelto invece b come approssimazione di α e procedendo come nel *primo passo* precedente si ottiene:

– *Scilab* assegna a `B` il valore:

$$b = \texttt{SQRT(2)} \oslash 2$$

– Con lo stesso e_1 del caso precedente si riscrive:

$$b = \big((1+e_1)\,\sqrt{2}\,\big) \oslash 2 \qquad \text{con} \qquad |\,e_1\,| \leqslant u$$

– Poiché per ogni $\xi \in F(2, 53)$ si ha $\xi \oslash 2 = \xi/2$ ('la divisione per 2 è esatta'), si ottiene:

$$b = (1+e_1)\,\frac{\sqrt{2}}{2} = (1+e_1)\,\alpha$$

Quanto ottenuto equivale a:

$$\frac{b-\alpha}{\alpha} = e_1$$

Considerando la limitazione $|\,e_1\,| \leqslant u$ si ottiene infine:

$$\Big|\,\frac{b-\alpha}{\alpha}\,\Big| \leqslant u$$

Dunque: *il valore assoluto dell'errore relativo commesso approssimando* α *con* b *non supera* u.

Si osservi che l'analisi svolta *non consente* di sapere quale tra a e b sia migliore come approssimazione di α (non sappiamo in quale caso l'errore relativo commesso nell'approssimazione sia più piccolo) ma, in base alle limitazioni ottenute: il *massimo* valore assoluto dell'errore relativo che si *rischia* di commettere è più piccolo usando b (u) piuttosto che a ($\approx 2u$). Questo è il *contesto*

usuale nel Calcolo Numerico: si hanno informazioni sul *massimo errore che si rischia di commettere* utilizzando i valori restituiti dalle procedure e *non* sull'errore *effettivamente commesso*.

0.34 Osservazione

Siano x un numero reale non nullo, ξ l'elemento di M scelto per approssimare x e:

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

l'errore relativo commesso nell'approssimazione.

Non è ragionevole sperare di ottenere una limitazione superiore per il valore assoluto dell'errore relativo più stringente di:

$$|\,\epsilon\,| \leqslant u$$

infatti quest'ultima è la limitazione che si ottiene quando $\xi = \mathrm{rd}(x)$, ovvero quando si approssima x con l'elemento di M ad esso *più vicino*.

0.35 Esempio

Sia $x \in (0, \frac{\pi}{2})$ un numero reale. Per ottenere un'approssimazione del numero reale $y = \operatorname{sen} x$, si considera la procedura seguente (ottenuta in modo ovvio):

```
y = sin(x);
```

Determiniamo una *limitazione superiore* al valore assoluto dell'errore relativo commesso approssimando y con il valore di `y` al termine dell'esecuzione della procedura in *Scilab*. Anche in questo caso cerchiamo la limitazione superiore in *due passi*.

- *Primo passo*:

  Sostituendo al tipo *numero reale* il tipo *numero in virgola mobile e precisone finita* con $M = F(2, 53)$, detta `SEN` la funzione predefinita corrispondente alla funzione elementare sen e considerato che in generale $x \notin F(2, 53)$, si deduce che *Scilab* assegna ad `y` il valore:

  $$y^* = \texttt{SEN(}\,\mathrm{rd}(x)\,\texttt{)}$$

  Utilizzando ripetutamente il Teorema 0.27 ed indicando con u la precisione di macchina in $F(2, 53)$ si ha:

  - Esiste $e_1 \in \mathbb{R}$ tale che:

    $$y^* = \texttt{SEN(}\,(1 + e_1)x\,\texttt{)} \qquad \text{con} \qquad |\,e_1\,| \leqslant u$$

  - Poiché per ogni $\xi \in F(2, 53)$ si ha $\texttt{SEN(}\xi\texttt{)} = \mathrm{rd}(\operatorname{sen}\xi)$, esiste $e_2 \in \mathbb{R}$ tale che:

    $$y^* = (1 + e_2)\operatorname{sen}\big(\,(1 + e_1)x\,\big) \qquad \text{con} \qquad |\,e_2\,| \leqslant u$$

  Siano giunti a *rileggere* y^* come *piccola perturbazione del valore della funzione* sen *in un punto vicino ad* x.

- *Secondo passo*:

  Affrontiamo il seguente problema riguardante la funzione sen: *dato* α *numero reale tale che* $\operatorname{sen}\alpha \neq 0$ *e* $\epsilon_* \in \mathbb{R}$*, determinare* $\epsilon \in \mathbb{R}$ *tale che*:

  $$\operatorname{sen}\big((1 + \epsilon_*)\,\alpha\big) = (1 + \epsilon)\,\operatorname{sen}\alpha$$

  Ricordando che $\operatorname{sen}\alpha \neq 0$, questa relazione determina *univocamente* ϵ:

  $$\epsilon = \frac{\operatorname{sen}\big((1 + \epsilon_*)\,\alpha\big) - \operatorname{sen}\alpha}{\operatorname{sen}\alpha}$$

  Poiché la funzione sen è regolare, in base al *Teorema di Lagrange* esiste un numero reale θ compreso tra α e $(1 + \epsilon_*)\,\alpha$ che consente di riscrivere:

  $$\operatorname{sen}\big((1 + \epsilon_*)\,\alpha\big) - \operatorname{sen}\alpha = (\cos\theta)\,\alpha\epsilon_*$$

  e quindi:

  $$\operatorname{sen}\big((1 + \epsilon_*)\,\alpha\big) = \Big(1 + \frac{\cos\theta}{\operatorname{sen}\alpha}\,\alpha\,\epsilon_*\Big)\,\operatorname{sen}\alpha$$

Per il risultato appena ottenuto, esiste un numero reale z compreso tra x e $(1+e_1)\,x$ tale che:

$$\operatorname{sen}\big((1+e_1)x\big) = \Big(1+\frac{\cos z}{\operatorname{sen} x}\,x\,e_1\Big)\,\operatorname{sen} x$$

Allora, posto:

$$t = \frac{\cos z}{\operatorname{sen} x}\,x\,e_1$$

si riscrive:

$$y^* = (1+e_2)(1+t)\operatorname{sen} x = (1+e_2)(1+t)\,y$$

che equivale a:

$$\frac{y^*-y}{y} = e_2 + t + e_2 t$$

Considerando la limitazione $|\,e_1\,| \leqslant u$ si ha $z \approx x$ e quindi $\cos z \approx \cos x$. Allora:

$$t \approx \frac{\cos x}{\operatorname{sen} x}\,x\,e_1 = \frac{x}{\tan x}\,e_1$$

Si ottiene perciò:

$$\Big|\,\frac{y^*-y}{y}\,\Big| = |\,e_2+t+e_2 t\,| \leqslant |e_2| + |t| + |e_2|\,|t| \approx |e_2| + \Big|\,\frac{x}{\tan x}\,\Big|\,|\,e_1| + |e_2|\,\Big|\,\frac{x}{\tan x}\,\Big|\,|\,e_1|$$

Ma, per ogni $x \in (0, \frac{\pi}{2})$ si ha:

$$\Big|\frac{x}{\tan x}\Big| < 1$$

Allora, ricordando che anche $|\,e_2\,| \leqslant u$, si deduce:

$$\Big|\,\frac{y^*-y}{y}\,\Big| \approx |e_2| + \Big|\,\frac{x}{\tan x}\,\Big|\,|\,e_1| + |e_2|\,\Big|\,\frac{x}{\tan x}\,\Big|\,|\,e_1| \leqslant 2u + u^2 \approx 2u$$

Dunque: *il valore assoluto dell'errore relativo commesso approssimando y con y^* non supera una quantità che vale circa $2u$.*

0.36 Osservazione

Siano $f : \mathbb{R} \to \mathbb{R}$ una funzione elementare, $\mathtt{F} : M \to M$ la funzione predefinita corrispondente ad f e x un numero reale.

La migliore approssimazione di $f(x)$ in M è $\mathrm{rd}\big(\,f(x)\,\big)$ *ma*, salvo casi particolari, *non è ragionevole* sperare di ottenere, utilizzando il calcolatore, un'approssimazione migliore di:

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

Negli esempi precedenti si sono (a) considerate *procedure* che determinano approssimazioni ξ in M del valore di opportune funzioni f (di $n \geqslant 1$ variabili reali a valori in $\mathbb{R}$) per un dato valore x dell'argomento e poi (b) determinate *informazioni sull'errore* commesso approssimando $f(x)$ con ξ.

In ciascun caso si è operato come segue:

(1) Si sono individuate opportunamente un numero *finito* j di funzioni predefinite $\mathrm{fp}_1, \ldots, \mathrm{fp}_j$ tali che, posto:

$$\phi = \mathrm{fp}_j \circ \cdots \circ \mathrm{fp}_1 \circ \mathrm{rd} : \mathbb{R}^n \to M$$

si ha: $\xi = \phi(x)$.

(2) Si è utilizzato il Teorema 0.27 per rileggere $\phi(x)$ nella forma:

$$\phi(x) = f\big(\,(1+e_{a1})\,x_1, \ldots, (1+e_{an})\,x_n\,\big)$$

oppure:

$$\phi(x) = (1+e_v)\,f\big(\,(1+e_{a1})\,x_1, \ldots, (1+e_{an})\,x_n\,\big)$$

(3) Si è determinato un numero reale t tale che:

$$f\big(\,(1+e_{a1})\,x_1, \ldots, (1+e_{an})\,x_n\,\big) = (1+t)\,f(x)$$

(4) Si è posto:

$$e = e_v + t + e_v\, t$$

cosicché:

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

(5) Infine, si è determinata una limitazione superiore per $|\, t\,|$.

0.37 Definizione
La *funzione* ϕ introdotta nel punto (1) si chiama *algoritmo* (utilizzato per approssimare i valori di f).

– *Esercizio*

Per ciascuno dei tre casi trattati negli esempi 0.33 e 0.35, determinare le funzioni f e ϕ ed il valore x dell'argomento. Indicare infine la forma scelta per riscrivere $\phi(x)$ nel punto (2).

I passaggi (2)–(5) si formalizzano con i due asserti seguenti.

0.38 Definizione (qualitativa di algoritmo accurato, stabile e di calcolo ben condizionato)
Siano f una funzione da $\Omega \subset \mathbb{R}$ in $\mathbb{R}$, ϕ da Ω in M un algoritmo ed $x \in \Omega$ tale che $f(x) \neq 0$.[12]

- L'algoritmo ϕ è *accurato* quando utilizzato per approssimare f in x se, posto:

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

l'errore relativo e risulta "*piccolo*" (cioé: se $\phi(x)$ è una "*piccola*" perturbazione moltiplicativa di $f(x)$).

- L'algoritmo ϕ è *stabile* quando utilizzato per approssimare f in x se: esistono $e_v, e_a \in \mathbb{R}$ tali che:

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

e gli errori relativi e_v ed e_a risultano "*piccoli*" (ovvero, se $\phi(x)$ è una "*piccola*" perturbazione moltiplicativa del valore di f in un punto "*vicino*" ad x).

- Il calcolo di f in x è *ben condizionato* se: *per ogni* $e_a \in \mathbb{R}$ "*piccolo*," posto:

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

l'errore relativo e_v^* risulta "*piccolo*" (cioè: se in *ogni* punto "*vicino*" ad x il valore di f è una "*piccola*" perturbazione moltiplicativa di $f(x)$).

Le definizioni sono *qualitative* perché non si è dato un significato quantitativo al termine "piccolo" associato ai vari errori relativi. In ogni caso si richiede che se q_1 e q_2 sono quantità "piccole" allora posto $(1+q_1)(1+q_2) = (1+q_t)$, ovvero $q_t = q_1 + q_2 + q_1 q_2$, la quantità q_t risulti a sua volta "piccola."
Si osservi infine che:

(1) 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 $f(x)$.

(2) Se $x = 0$ la proprietà di *stabilità* coincide con quella di *accuratezza*. Per ottenere una nozione più utile la *stabilità* va riformulata introducendo una perturbazione *additiva* di x.

(3) 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 $f(x)$.

Se $x = 0$ (e $f(0) \neq 0$) risulta $e_v^* = 0$ ed il calcolo è ben condizionato quale che sia f. Anche in questo caso, per ottenere una nozione più utile, la definizione va riformulata introducendo una perturbazione *additiva* di x.

[12] Le definizioni ed il Teorema successivo sono date nel caso di f funzione di *una* variabile. Le modifiche da apportare nel caso generale sono ovvie.

(4) Se $f(x) = 0$ e $\phi(x) = 0$ la relazione:

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

è verificata per $e = 0$, cioè: $\phi(x)$ *è* una "piccola" perturbazione moltiplicativa di $f(x)$ e quindi ϕ *è un algoritmo accurato quando utilizzato per approssimare* f *in* x.

(5) Se $x = 0$, $f(0) = 0$ e $\phi(0) = 0$ la relazione:

$$\phi(0) = (1 + e_2)\, f\big((1 + e_1)0\big)$$

è verificata per $e_1 = e_2 = 0$, cioè: $\phi(0)$ *è* una "piccola" perturbazione moltiplicativa del valore di f in un punto "vicino" a 0 e quindi ϕ *è un algoritmo stabile quando utilizzato per approssimare* f *in* $x = 0$.

0.39 Teorema (stabilità + buon condizionamento ⇒ accuratezza)
Siano f una funzione da $\Omega \subset \mathbb{R}$ in $\mathbb{R}$ ed $x \in \Omega$ tale che $x \neq 0$ e $f(x) \neq 0$. Sia infine ϕ da Ω in M l'algoritmo utilizzato per approssimare f in x.

Se ϕ *è* stabile ed *il calcolo di* f *in* x *è* ben condizionato*, allora* ϕ *è* accurato

(Dim: Per la stabilità si ha: esistono $e_v, e_a \in \mathbb{R}$ "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:

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

l'errore relativo e_v^* risulta "*piccolo.*" Si ottiene perciò:

$$\phi(x) = (1 + e_v)(1 + e_v^*) f(x) = (1 + t) f(x)$$

con $t = e_v + e_v^* + e_v e_v^*$ che risulta "piccolo." Dunque l'algoritmo è accurato.)

0.40 Esempio
Discutere l'accuratezza dell'algoritmo ϕ quando utilizzato per approssimare i valori della funzione f nei seguenti casi:

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

- *Stabilità dell'algoritmo.*
  Per ogni numero reale x esistono numeri reali e_1, e_2 tali che:

  $$\phi(x) = (1 + e_2)\, \operatorname{sen}\big((1 + e_1)x\big) \qquad \text{con } |\, e_1 \,| \leqslant u \text{ ed } |\, e_2 \,| \leqslant u$$

  Considerando "piccola" la precisione di macchina, l'algoritmo verifica la definizione di stabilità (con $e_v = e_2$, $e_a = e_1$) *per ogni* $x \in \mathbb{R}$.
- *Condizionamento del calcolo.*
  Assegnati $f : \Omega \to \mathbb{R}$ ($\Omega \subset \mathbb{R}$), $x \in \Omega$ tale che $f(x) \neq 0$ ed $e \in \mathbb{R}$ tale che $(1 + e)x \in \Omega$, sia e_v^* il numero reale tale che:

  $$f\big((1 + e)x\big) = (1 + e_v^*)\, f(x) \qquad \text{ovvero} \qquad e_v^* = \frac{f\big((1 + e)x\big) - f(x)}{f(x)}$$

  Si chiama *funzione di condizionamento* del calcolo di f in x la funzione definita da:

  $$C(x, e) = \left| \frac{f\big((1 + e)x\big) - f(x)}{f(x)} \right|$$

  Studiare il condizionamento del calcolo di f in x significa determinare, per x fissato ed e "piccolo," una *limitazione superiore* per $|\, e_v^* \,|$, ovvero per $C(x, e)$, in termini di $|\, e \,|$.

Se $f : \mathbb{R} \to \mathbb{R}$ è una funzione *con derivata prima continua*, utilizzando il Teorema di Lagrange si ottiene: dati numeri reali x ed e *esiste* un numero reale θ compreso tra x ed $(1+e)x$ tale che:

$$f\big((1+e)x\big) = f(x) + f'(\theta)\, ex$$

dunque, se $f(x) \neq 0$, si ottiene:

$$C(x,e) = \left| f'(\theta)\, \frac{x}{f(x)}\, e \right|$$

Poiché per lo studio del condizionamento si considera e piccolo, è usuale ricorrere all'approssimazione (studio "locale" del condizionamento):

$$C(x,e) \approx \left| f'(x)\, \frac{x}{f(x)} \right| |\, e\,|$$

e ricondurre così lo studio del condizionamento a quello del "coefficiente di amplificazione" (dipendente *solo* da x)

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

detto *numero di condizionamento* (del calcolo di f in x).
Nel caso in esame la funzione di condizionamento è definita per ogni $x \in \mathbb{R}$ non multiplo intero di π ed ogni $e \in \mathbb{R}$. La funzione ha derivata prima continua e quindi:

$$C(x,e) = \left| \cos\theta\, \frac{x}{\operatorname{sen} x}\, e \right| \qquad \text{con } \theta \text{ tra } x \text{ e } (1+e)x$$

e il numero di condizionamento è:

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

Per ogni $x \in (0, \frac{\pi}{2})$ si ha:

$$k(x) \leqslant 1$$

ed il calcolo *è* ben condizionato e quindi, per il Teorema precedente, l'algoritmo ϕ risulta *accurato.*
Invece:

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

ed il calcolo, invece, *non è* ben condizionato. In questo caso il Teorema citato *non consente* di concludere alcunché riguardo all'accuratezza dell'algoritmo (vedere l'Esercizio *E34*).

- *Esercizio*
  Utilizzare *Scilab* per ottenere il (un'approssimazione del) grafico della funzione

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

  per $x \in (0,\pi) \cup (\pi, 2\pi)$ e dedurne che il calcolo di sen x risulta ragionevolmente ben condizionato (e quindi, per il Teorema precedente, l'algoritmo ϕ risulta *accurato*) per $x \in (0, \pi - h) \cup (\pi + h, 2\pi - h)$ con h non troppo piccolo.

(2) $f(x) = 1/\sqrt{x}$, $\phi(x) = 1 \oslash \texttt{SQRT}(\mathrm{rd}(x))$ (entrambe definite per $x > 0$).

- *Stabilità dell'algoritmo.*
  Per ogni numero reale $x > 0$ esistono numeri reali e_1, e_2 ed e_3 tali che:

  $$\phi(x) = \frac{1+e_3}{(1+e_2)\sqrt{(1+e_1)x}} \qquad \text{con } |\, e_1\,| \leqslant u, |\, e_2\,| \leqslant u \text{ ed } |\, e_3\,| \leqslant u$$

  e quindi:

  $$\phi(x) = \frac{1+e_3}{1+e_2}\, f\big((1+e_1)x\big) = (1+t) f\big((1+e_1)x\big) \quad \text{con } t = \frac{e_3 - e_2}{1+e_2}$$

Utilizzando le limitazioni su e_2 ed e_3 si ottiene:

$$|\,t\,| \leqslant \frac{2u}{1-u} \approx 2u$$

Considerando "piccola" la precisione di macchina, l'algoritmo verifica la definizione di stabilità (con $e_v = t$, $e_a = e_1$) *per ogni* $x > 0$.

– *Condizionamento del calcolo.*

La funzione ha derivata prima continua. Per ogni $x > 0$ ed ogni e tale che $(1+e)x > 0$ si ha allora:

$$C(x,e) = \left| -\tfrac{1}{2}\,\frac{x\sqrt{x}}{\theta\sqrt{\theta}}\,e \right| \qquad \text{con } \theta \text{ tra } x \text{ e } (1+e)x$$

e:

$$k(x) = \tfrac{1}{2}$$

Possiamo ritenere il calcolo di f *ben condizionato* per ogni $x > 0$.

In base a quanto ottenuto ed al Teorema precedente: l'algoritmo ϕ è *accurato* quando utilizzato per approssimare f, *per ogni* $x > 0$.

0.41 Osservazione (stabilità delle funzioni predefinite)

Siano $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}$ una funzione elementare e $\phi : \Omega \cap M \to M$ la funzione predefinita corrispondente ad f (ovvero tale che: per ogni $\xi \in \Omega \cap M, \phi(\xi) = \mathrm{rd}\big(f(\xi)\big)$).[13]

L'algoritmo $\Phi : \Omega \to M$ definito da $\Phi(x) = \phi(\mathrm{rd}(x))$ è *stabile* quando utilizzato per approssimare f per ogni $x \in \Omega$.

(Infatti: per ogni x esistono e_1 ed e_2 tali che:

$$\Phi(x) = (1+e_2)f\big((1+e_1)x\big) \qquad \text{e} \qquad |\,e_1\,| \leqslant u, |\,e_2\,| \leqslant u$$

Come già osservato, le limitazioni su e_1 ed e_2 consentono di ritenerli in ogni caso "piccoli." Dunque l'algoritmo è stabile per ogni x.)

Siano $*$ un'operazione aritmetica, $f : \Omega \to \mathbb{R}$ la funzione definita da $f(x_1, x_2) = x_1 * x_2$ e $\circledast$ la pseudo-operazione aritmetica corrispondente a $*$.

L'algoritmo $\Phi : \Omega \to M$ definito da $\Phi(x) = \mathrm{rd}(x_1) \circledast \mathrm{rd}(x_2)$ è *stabile* quando utilizzato per approssimare f per ogni $(x_1, x_2) \in \Omega$.

(Infatti: per ogni x_1, x_2 esistono 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: $\Phi(x)$ è una "*piccola*" perturbazione moltiplicativa del valore di f in un punto "*vicino*" ad 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 mostra invece che la composizione di algoritmi stabili *non sempre* genera algoritmi a loro volta stabili e chiarisce perchè ciò accade.

0.42 Osservazione (algoritmi *non* stabili)

Siano $f, g : \mathbb{R} \to \mathbb{R}$ due funzioni e ϕ, γ gli algoritmi, *stabili* quando utilizzati per approssimare, rispettivamente, i valori di f e g, definiti da:

$$\phi(x) = \mathrm{rd}\Big(f\big(\mathrm{rd}(x)\big)\Big) \qquad , \qquad \gamma(x) = \mathrm{rd}\Big(g\big(\mathrm{rd}(x)\big)\Big)$$

Si vuole studiare la stabilità dell'algoritmo $\Phi(x) = \phi\big(\gamma(x)\big)$ quando utilizzato per approssimare i valori di $F(x) = f\big(g(x)\big)$.

[13] L'insieme di definizione di ϕ è, più correttamente: $\{\, x \in \Omega \text{ t.c. } \mathrm{rd}(x) \in \Omega \,\}$

Tenuto conto della definizione di γ, per ogni x esistono $e_1, e_2 \in \mathbb{R}$ tali che:

$$\Phi(x) = \phi\Big((1+e_2)\, g\big((1+e_1)x\big)\Big) \quad \text{con} \quad |\, e_1 \,| \leqslant u, |\, e_2 \,| \leqslant u$$

Tenuto conto della definizione di ϕ e del fatto che $(1+e_2)\, g\big((1+e_1)x\big) = \gamma(x)$ è un elemento di M, per ogni x esiste $e_3 \in \mathbb{R}$ tale che:

$$\Phi(x) = (1+e_3) f\Big((1+e_2)\, g\big((1+e_1)x\big)\Big) \quad \text{con} \quad |\, e_3 \,| \leqslant u$$

Per leggere $\Phi(x)$ come perturbazione moltiplicativa del valore di F in un opportuno punto, determiniamo $t \in \mathbb{R}$ tale che:

$$f\Big((1+e_2)\, g\big((1+e_1)x\big)\Big) = (1+t) f\Big(g\big((1+e_1)x\big)\Big)$$

Ponendo poi $(1+e_3)(1+t) = 1+\tau$ si ottiene:

$$\Phi(x) = (1+e_3)(1+t) f\Big(g\big((1+e_1)x\big)\Big) = (1+\tau) F\big((1+e_1)x\big)$$

con:

$$\tau = e_3 + t + t\, e_3$$

Per giudicare la stabilità di Φ occorre decidere se τ, ovvero t, sia "piccolo." In altri termini occorre indagare il *condizionamento* del calcolo di f in $g\big((1+e_1)x\big)$:

- Se il calcolo di f in $g\big((1+e_1)x\big)$ è *ben condizionato* allora t risulta "piccolo." Dunque anche τ lo è e l'algoritmo Φ risulta *stabile*.
- Se il calcolo di f in $g\big((1+e_1)x\big)$ *non è* ben condizionato allora l'algoritmo Φ *può* risultare *non stabile*.

Prima di mostrare un esempio di algoritmo non stabile, affrontiamo lo studio del condizionamento delle operazioni aritmetiche.

0.43 Osservazione (condizionamento delle operazioni aritmetiche)
Sia $*$ un'operazione aritmetica. La funzione di condizionamento in questo caso è definita da:

$$C(x_1, x_2; e_1, e_2) = \left| \frac{\big((1+e_1)x_1 * (1+e_2)x_2\big) - (x_1 * x_2)}{(x_1 * x_2)} \right|$$

ovvero:

$$(1+e_1)x_1 * (1+e_2)x_2 = \big(1 + C(x_1, x_2; e_1, e_2)\big)\,(x_1 * x_2)$$

Con semplici passaggi si ottiene, per l'*addizione*:

$$C(x_1, x_2; e_1, e_2) = \left| \frac{x_1}{x_1+x_2}\, e_1 + \frac{x_2}{x_1+x_2}\, e_2 \right|$$

per la *moltiplicazione*:

$$C(x_1, x_2; e_1, e_2) = |\, e_1 + e_2 + e_1 e_2 \,|$$

e per la *divisione*:

$$C(x_1, x_2; e_1, e_2) = \left| \frac{e_1 - e_2}{1+e_2} \right|$$

Il calcolo della moltiplicazione e della divisione è *sempre ben condizionato*, infatti per le rispettive funzioni di condizionamento si ha, per e_1 ed e_2 piccoli:

$$C(x_1, x_2; e_1, e_2) \leqslant |\, e_1 \,| + |\, e_2 \,| + |\, e_1 \,|\, |\, e_2 \,| \approx |\, e_1 \,| + |\, e_2 \,|$$

e, rispettivamente:

$$C(x_1, x_2; e_1, e_2) \leqslant \frac{|\, e_1 \,| + |\, e_2 \,|}{1 - |\, e_2 \,|} \approx |\, e_1 \,| + |\, e_2 \,|$$

Per il calcolo dell'addizione, invece, il condizionamento *dipende dagli addendi*:

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

$$C(x_1, x_2; e_1, e_2) \leqslant \left| \frac{x_1}{x_1 + x_2} \right| |\, e_1 \,| + \left| \frac{x_2}{x_1 + x_2} \right| |\, e_2 \,| < |\, e_1 \,| + |\, e_2 \,|$$

– *Se gli addendi hanno segno opposto, il condizionamento del calcolo è tanto peggiore quanto 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:

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

0.44 Esempio

Si consideri la funzione f definita, per ogni $x \geqslant 0$ da: $f(x) = x - \sqrt{x}$ e l'algoritmo ϕ definito, anch'esso per ogni $x \geqslant 0$, da: $\phi(x) = \mathrm{rd}(x) \ominus \mathtt{SQRT}\big(\mathrm{rd}(x)\big)$. Si vuole studiare la stabilità di ϕ quando utilizzato per approssimare i valori di f per $x > 0$.

Per ogni x esistono e_1, e_2 ed e_3 tali che:

$$\phi(x) = (1 + e_3)\Big((1 + e_1)x - (1 + e_2)\sqrt{(1 + e_1)x}\Big)$$

con $|\, e_1 \,| \leqslant u, |\, e_2 \,| \leqslant u$ ed $|\, e_3 \,| \leqslant u$.

Posto $\xi = \mathrm{rd}(x) = (1 + e_1)x$, per ogni $\xi \neq 0, 1$ si ha:

$$\xi - (1 + e_2)\sqrt{\xi} = (1 + t)(\xi - \sqrt{\xi})$$

con:

$$t = -\frac{\sqrt{\xi}}{\xi - \sqrt{\xi}}\, e_2$$

e quindi:

$$\phi(x) = (1 + e_3)(1 + t)\big((1 + e_1)x - \sqrt{(1 + e_1)x}\,\big) = (1 + e_3)(1 + t) f\big((1 + e_1)x\big)$$

Per decidere la stabilità di ϕ occorre indagare la grandezza di t, ovvero studiare il *condizionamento* del calcolo di $x_1 + x_2$ per $x_1 = \xi$ e $x_2 = -\sqrt{\xi}$ nel caso di perturbazione sul solo secondo addendo. Per quanto detto nell'Osservazione precedente (addizione con addendi di segno opposto), il coefficiente:

$$\left| \frac{\sqrt{\xi}}{\xi - \sqrt{\xi}} \right|$$

assume valori tanto più grandi quanto più $-\sqrt{\xi}/\xi = -1/\sqrt{\xi}$ è vicino a -1, cioè tanto più ξ è vicino a 1.

0.45 Esempio

Si consideri la funzione f definita, per ogni $x \geqslant 0$ da: $f(x) = x + \sqrt{x}$ e l'algoritmo ϕ definito, anch'esso per ogni $x \geqslant 0$, da: $\phi(x) = \mathrm{rd}(x) \oplus \mathtt{SQRT}\big(\mathrm{rd}(x)\big)$. Si vuole studiare la stabilità di ϕ quando utilizzato per approssimare i valori di f per $x > 0$.

Per ogni x esistono e_1, e_2 ed e_3 tali che:

$$\phi(x) = (1 + e_3)\Big((1 + e_1)x + (1 + e_2)\sqrt{(1 + e_1)x}\Big)$$

con $|\, e_1 \,| \leqslant u, |\, e_2 \,| \leqslant u$ ed $|\, e_3 \,| \leqslant u$.

Posto $\xi = \mathrm{rd}(x) = (1 + e_1)x$, poichè $x > 0$ implica $\xi > 0$ si ha:

$$\xi + (1 + e_2)\sqrt{\xi} = (1 + t)(\xi + \sqrt{\xi})$$

con:

$$t = \frac{\sqrt{\xi}}{\xi + \sqrt{\xi}}\, e_2$$

e quindi:

$$\phi(x) = (1+e_3)(1+t)\big((1+e_1)x + \sqrt{(1+e_1)x}\,\big) = (1+e_3)(1+t) f\big((1+e_1)x\big)$$

Per decidere la stabilità di ϕ occorre indagare la grandezza di t, ovvero studiare il *condizionamento* del calcolo di $x_1 + x_2$ per $x_1 = \xi$ e $x_2 = \sqrt{\xi}$ nel caso di perturbazione sul solo secondo addendo. Per quanto detto nell'Osservazione precedente (addizione con addendi di uguale segno), si ha certamente $|\, t \,| \leqslant |\, e_2 \,|$. L'algoritmo risulta *stabile* per ogni $x > 0$.

0.46 Esempio

Si consideri la funzione f definita, per ogni $x \geqslant 0$ da: $f(x) = x - \sqrt{x}$. Per ogni $x > 0$ si ha:

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

Sia ϕ l'algoritmo, definito anch'esso per ogni $x \geqslant 0$, da:

$$\phi(0) = 0 \qquad , \qquad \phi(x) = \Big(\mathrm{rd}(x) \otimes (\mathrm{rd}(x) \ominus 1)\Big) \oslash \Big(\mathrm{rd}(x) \oplus \mathtt{SQRT}(\mathrm{rd}(x))\Big) \quad \text{per } x > 0$$

Si vuole studiare la stabilità di ϕ quando utilizzato per approssimare i valori di f per $x > 0$.

Per ogni $x > 0$ poniamo: $\xi = \mathrm{rd}(x)$, $f_1(x) = x(x-1)$, $\phi_1(x) = \xi \otimes (\xi \ominus 1)$, $f_2(x) = x + \sqrt{x}$ e $\phi_2(x) = \xi \oplus \mathtt{SQRT}(\xi)$. Si ha:

- Per ogni $x > 0$:

$$f(x) = \frac{f_1(x)}{f_2(x)}$$

- Esistono numeri reali e_1, e_2 ed e_3 tali che $|\, e_1 \,| \leqslant u$, $|\, e_2 \,| \leqslant u$, $|\, e_3 \,| \leqslant u$ e:

$$\phi_1(x) = (1+e_3)(1+e_2)\xi(\xi - 1) = (1+e_3)(1+e_2) f_1\big((1+e_1)x\big)$$

  Posto $(1+e_3)(1+e_2) = 1 + t_{23}$ si ha:

$$\phi_1(x) = (1+t_{23}) f_1\big((1+e_1)x\big)$$

  e, tenuto conto delle limitazioni su e_2 ed e_3:

$$|\, t_{23} \,| \leqslant 2u + u^2 \approx 2u$$

  Se ne deduce che ϕ_1 è un algoritmo *stabile* quando utilizzato per approssimare f_1 per $x > 0$.

- Per quanto mostrato nell'Esempio 0.45, ϕ_2 è un algoritmo *stabile* quando utilizzato per approssimare f_2: per ogni x esiste un numero reale e_4 "piccolo" tale che:

$$\phi_2(x) = (1+e_4) f_2\big((1+e_1)x\big)$$

- Esiste un numero reale e_5 tale che $|\, e_5 \,| \leqslant u$ e:

$$\phi(x) = \phi_1(x) \oslash \phi_2(x) = (1+e_5)\,\frac{\phi_1(x)}{\phi_2(x)} = \frac{(1+e_5)(1+t_{23})}{1+e_4}\,\frac{f_1\big((1+e_1)x\big)}{f_2\big((1+e_1)x\big)}$$

  Posto:

$$1 + \theta = \frac{(1+e_5)(1+t_{23})}{1+e_4}$$

  si ha:

$$\phi(x) = (1+\theta)\, f\big((1+e_1)x\big)$$

  e, tenuto conto delle limitazioni su e_4, e_5 e t_{23}:

$$\theta = \frac{t_{23} - e_4 + e_5 + e_5 t_{23}}{1+e_4} \qquad\qquad |\, \theta \,| \leqslant \frac{5u + 4u^2 + u^3}{1-u} \approx 5u$$

Se ne deduce che ϕ è un algoritmo *stabile* quando utilizzato per approssimare f per ogni $x > 0$.

Esercizi

E30 Tenuto conto che 2^{-53} è la precisione di macchina in $F(2, 53)$, spiegare i risultati del punto (4) dell'Esempio 0.32.

E31 ★ Realizzando la procedura del punto (4) dell'Esempio 0.32 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` $\neq$ `y`.

E32 Sia $f(x) = \sqrt{x}/x$ (definita per $x > 0$). Determinare l'insieme di definizione e discutere l'accuratezza dell'algoritmo:

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

quando utilizzato per approssimare i valori di f.

E33 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. Sia poi $C(x, e)$ la funzione di condizionamento del calcolo di f in x. Per ogni $x \neq \alpha$ ed $e \in \mathbb{R}$ esiste θ tale che:

$$C(x, e) = \left| f'(\theta)\, \frac{x}{f(x)}\, e \right|$$

Mostrare che per ogni x, e del dominio si ha:

$$C(x, e) > L \left| \frac{x}{f(x)} \right| |\, e\,|$$

e dedurne che per ogni $e \neq 0$:

$$\lim_{x \to \alpha} C(x, e) = +\infty$$

E34 Si consideri il punto (1) dell'Esempio 0.40. Tenuto conto che in *Scilab* si ha: `%pi` = $\mathrm{rd}(\pi) < \pi$ e $\phi(\pi) =$ `SEN(%pi)` > 0:

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

(2) Mostrare che, posto per ogni $x \in$ (`%pi`, π):

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

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.
