

## 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  $f : \mathbb{R} \rightarrow \mathbb{R}$ , determinare un numero reale  $\alpha$  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  $(x_0, y_0), \dots, (x_k, y_k)$  e le funzioni  $f_0, \dots, f_k : \mathbb{R} \rightarrow \mathbb{R}$ , determinare numeri reali  $a_0, \dots, a_k$  tali che, posto  $f(x) = a_0 f_0(x) + \dots + a_k f_k(x)$  si abbia  $f(x_0) = y_0, \dots, f(x_k) = y_k$ .
- 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  $SQ : \mathbb{R}^k \rightarrow \mathbb{R}$  definita da  $SQ(x) = \|Ax - b\|^2$ .

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. 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:

$$\xi = 1 \quad , \quad \xi = 1.4142135623730951454746218587388284504413604736328125$$

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  $\xi$  è l'*errore relativo*:

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

Risposte soddisfacenti sono allora:

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

e:

$$\xi = 1.4142135623730951454746218587388284504413604736328125 \quad , \quad |\epsilon| < 2^{-53} \approx 10^{-16}$$

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*.<sup>1</sup> 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.

<sup>1</sup>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.

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  $\beta$  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:

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

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

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

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

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

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

#### 0.1.2 Esempio

Sia  $x = \sqrt{5}$ .

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

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

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

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

#### 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 0.09). In tal caso, delle due si considera quella *finita*. Con questa precisazione, la condizione  $\beta^{-1} \leq g < 1$  è equivalente a:

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

Le cifre  $c_1, c_2, \dots$  della scrittura posizionale di  $g$  in base  $\beta$  si possono ottenere, una alla volta, con la procedura seguente:<sup>2</sup>

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

- Passo 1:  $i = 1; t_i = g; (t_1 = 0.c_1c_2 \dots)$
- Passo 2:  $c_i = \lfloor \beta t_i \rfloor; (\beta t_i = c_i.c_{i+1}c_{i+2} \dots)$
- Passo 3:  $t_{i+1} = \beta t_i - \lfloor \beta t_i \rfloor; (t_{i+1} = 0.c_{i+1}c_{i+2} \dots)$
- 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  $10^{-1} \leq x < 10^0$ :

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

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

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

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 Definizione (numeri in virgola mobile, precisione)

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

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

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

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

#### 0.1.6 Esempio

Si consideri  $F(10, 1)$ .

- Poiché  $\frac{1}{100} = 10^{-2} 0.1$  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:

$$0.1, 0.2, \dots, 0.9$$

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

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

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:<sup>3</sup>

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

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

<sup>3</sup>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  $a$ .

### 0.1.7 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  $10^{b-1}$ .

### 0.1.8 Osservazione (Proprietà di $F(\beta, m)$ )

Si ha:

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

Infatti:  $\xi = (-1)^s \beta^b 0.c_1 \dots c_m = (-1)^s \beta^{b-m} c_1 \dots c_m \in \mathbb{Q}$  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  $\xi_k$  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  $\lim \xi_k = +\infty$ .

### 0.1.9 Definizione (Funzioni successore e predecessore)

Si consideri la rappresentazione degli elementi di  $F(\beta, m)$  sulla retta reale e sia  $\xi$  un elemento non nullo di  $F(\beta, m)$ . Il *successore* di  $\xi$ , che si indica con  $\sigma(\xi)$ , è "il primo elemento di  $F(\beta, m)$  a destra di  $\xi$ ." Il *predecessore* di  $\xi$ , che si indica con  $\pi(\xi)$ , è "il primo elemento di  $F(\beta, m)$  a sinistra di  $\xi$ ." Le funzioni  $\sigma$  e  $\pi$ , 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*.<sup>4</sup>

### 0.1.10 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  $\xi$  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  $\xi$  è il primo elemento di  $B_4$  e quindi il primo elemento a sinistra di  $\xi$  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  $\pi(\xi) = -\sigma(-\xi)$ .

### 0.1.11 Teorema (distribuzione degli elementi di $F(\beta, m)$ )

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

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

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

<sup>4</sup>Più formalmente: il primo elemento di  $F(\beta, m)$  a destra di  $\xi$  è il più piccolo elemento di  $F(\beta, m)$  maggiore di  $\xi$ ; il primo elemento di  $F(\beta, m)$  a sinistra di  $\xi$  è il più grande elemento di  $F(\beta, m)$  minore di  $\xi$ .

– *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.

**0.1.12 Definizione** (numeri in virgola mobile con esponente limitato ed elementi denormalizzati)

Siano  $\beta$  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*,  $b_{\min} \leq b \leq b_{\max}$ , si indica con:

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

e si chiama insieme dei numeri in virgola mobile (normalizzati) in base  $\beta$  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*,  $b_{\min} \leq b \leq b_{\max}$ , e da tutti i numeri reali  $x$  tali che:

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

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

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

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

**0.1.13 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  $\xi$  ha esponente maggiore o uguale a  $b_{\min}$  allora  $0.c_1 \cdots c_m$  è la scrittura posizionale (in base  $\beta$ ) della frazione di  $\xi$ . Se invece  $\xi$  ha esponente minore di  $b_{\min}$  – ovvero  $\xi$  è un elemento denormalizzato – allora  $0.0c_2 \cdots c_m$  *non* è la scrittura posizionale della frazione di  $\xi$ .

(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  $c_1 = c_2 = \cdots = c_m = 0$ .

**0.1.14 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  $\xi_{\max}$ .<sup>5</sup>
- 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*.

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

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  $\xi_{\max}$ .<sup>5</sup>
- 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 *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*.

### 0.1.15 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  $M = F_d(2, 53, -1021, 1024)$ <sup>6</sup> perché questi sono gli oggetti del tipo di dato numerico che *Scilab (Matlab, Octave)* consente di usare.<sup>7</sup> 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  $M = F_d(2, 24, -125, 128)$ .<sup>8</sup>
- 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,  $\pi$* .

*E3* Determinare il numero di elementi dell'insieme:

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

*E4* Dimostrare che  $F(2, 2) \subset F(2, 3)$  e che  $F(10, 1) \subset F(10, 2)$ . In generale:

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

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 0.1.4) e dedurre che sono falsi gli asserti  $F_2 \supset F_{10}$  e  $F_2 = F_{10}$ .

---

<sup>6</sup>Questo è il formato "binary64" dello IEEE Standard for Floating-Point Arithmetic.

<sup>7</sup>Nei linguaggi *Matlab* e *Octave* questo tipo di dato si chiama `double`.

<sup>8</sup>Questo è il formato "binary32" dello IEEE Standard for Floating-Point Arithmetic.



(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  $10^m = 10^{m+1} \cdot 0.1$  (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  $F_2 \subset F_{10}$ .

**E6** ★ La dimostrazione dell'asserto (1) dell'Osservazione 0.1.8 prova che: se  $\xi$  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  $\text{esp}, \text{fraz} : F(\beta, m) \setminus \{0\} \rightarrow \mathbb{R}$  le funzioni definite da:

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

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

**E10** Posto  $\xi = 2^{-3} \cdot 0.1101 \in F(2, 4)$ , 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:

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

Determinare poi:

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

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

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

**E13** Assegnate una base  $\beta$  ed una precisione  $m$ , dimostrare che:

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

**E14** Si consideri  $F(2, 3, -7, 7)$ . Determinare:

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

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

**E15** Si consideri  $F_d(2, 3, -7, 7)$ . Determinare:

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

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

**E16** ★ Sia  $\phi$  la funzione definita, per ogni elemento non nullo di  $F(\beta, m)$ , da  $\phi(\xi) = \sigma(\xi) - \xi$ . Mostrare che per ogni  $\xi$  si ha  $\phi(\xi) \in F(\beta, m)$ . Discutere la monotonia della funzione  $\phi$ .

E17 ♠ 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  $\xi_{\min}^*$ .

E18 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  $\beta$ ,  $x = \beta^b 0.c_1 c_2 \dots$  allora, posto  $\xi_- = \beta^b 0.c_1 \dots c_m$  (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  $\xi_-$  e  $\xi_+$  sono gli elementi di  $M$  adiacenti ad  $x$ .

*Esercizio:* Determinare gli elementi adiacenti ad  $x = \sqrt{2} = 1.4142\dots$  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  $\text{rd}(x)$ , 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  $\beta$  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  $c_m$  *pari*<sup>9</sup> — questa scelta, utilizzabile *solo* se la base  $\beta$  è *pari* ed  $M$  è un insieme con *esponente non limitato* oppure con *esponente limitato ed elementi denormalizzati*, sarà indicata con la sigla RTTE<sup>10</sup> 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  $\text{rd} : \mathbb{R} \rightarrow M$  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  $\text{rd}(x) = 2^{-3} 0.11 = \frac{3}{32}$ . 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  $\xi_-, \xi_+$  è  $2^{-3} 0.111 > x$  allora l'elemento di  $M$  più vicino ad  $x$  è ben definito:  $\xi_-$ .

### 0.2.5 Osservazione (Proprietà della funzione arrotondamento)

Siano  $M$  l'insieme dei numeri in virgola mobile e precisione finita ed  $\text{rd} : \mathbb{R} \rightarrow M$  la funzione arrotondamento in  $M$  scelti.

<sup>9</sup>Si ricordi che per la definizione di  $c_m$  si fa riferimento: nel caso di un elemento *normalizzato* alla scrittura  $(-1)^s \beta^b 0.c_1 \dots c_m$ , con  $c_1 \neq 0$ ; nel caso di un elemento *denormalizzato o nullo* alla scrittura  $(-1)^s \beta^{b_{\min}} 0.0c_2 \dots c_m$ .

<sup>10</sup>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.

- La funzione  $\text{rd}$  non è invertibile. Infatti, se  $\xi = \text{rd}(x) \neq 0$  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  $\text{rd}(y) = \xi$  include l'intervallo non vuoto  $(m_-, m_+)$ .
- La funzione  $\text{rd}$  è dispari:  $\text{rd}(-x) = -\text{rd}(x)$ . *Esercizio:* Verificare aiutandosi con un disegno!
- La funzione  $\text{rd}$  è non decrescente:  $x < y \Rightarrow \text{rd}(x) \leq \text{rd}(y)$ . Infatti, detto  $I$  l'insieme dei numeri reali  $t$  tali che  $\text{rd}(t) = \text{rd}(x)$  si ha: se  $y \in I$  allora  $\text{rd}(x) = \text{rd}(y)$ , altrimenti  $\text{rd}(x) < \text{rd}(y)$ .
- $\text{rd}(x) = x \Leftrightarrow x \in M$ .
- Se  $M = F(\beta, m)$  allora:  $\text{rd}(x) = 0 \Leftrightarrow x = 0$ .

### Esercizi

*E19* Calcolare l'arrotondato di  $\frac{1}{4}$  in  $F(3, 2)$ .

*E20* ★ Sia  $\xi = 3^b 0.c_1c_2c_3 \in F(3, 3)$ . Detto  $m$  il punto medio del segmento di estremi  $\xi$  e  $\sigma(\xi)$ , mostrare (aiutandosi con la rappresentazione grafica di tutti i numeri considerati) che:

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

e quindi che:

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

*E21* Calcolare l'arrotondato di  $2^2 0.1011$  in  $F(10, 2)$ .

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

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

*E24* Sia  $\text{rd}$  la funzione arrotondamento in  $F(10, 3)$  con RTTE. Determinare tutti gli  $x \in \mathbb{R}$  tali che  $\text{rd}(x) = 642$ .

*E25* Sia  $\text{rd}$  la funzione arrotondamento in  $F(10, 3)$  con RTTE. Determinare:

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

*E26* Sia  $M$  un insieme di numeri in virgola mobile con esponente limitato (ma senza elementi denormalizzati). Mostrare che la scelta RTTE non è utilizzabile per definire l'arrotondato in  $M$  di  $\frac{1}{2} \xi_{\min}$ .

*E27* Sia  $M = F(3, 2)$ . Mostrare che la scelta RTTE non è utilizzabile per definire l'arrotondato in  $M$  di  $x = \frac{11}{2}$ .

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  $\text{rd}$  la funzione arrotondamento in  $M$  scelti. La funzione  $\delta$  tale che:

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

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

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

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

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:

$$|\delta(x)| \leq \frac{1}{2} \beta^{b-m} \quad , \quad |\epsilon(x)| \leq \frac{1}{2} \beta^{1-m} \quad , \quad |\eta(x)| \leq \frac{1}{2} \beta^{1-m}$$

(*Infatti*:  $x$  è un numero reale positivo con esponente  $b$  dunque  $\beta^b \beta^{-1} \leq x < \beta^{b+1} \beta^{-1}$ ; la disuguaglianza relativa alla funzione  $\delta$  si ottiene immediatamente dal Teorema 0.1.11. Le altre disuguaglianze si ottengono utilizzando quella relativa a  $\delta$  e considerando che il valore minimo per  $x$  e per  $\text{rd}(x)$  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  $\beta$ , 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  $\text{rd}$  la funzione arrotondamento in  $F(\beta, m)$ , con  $\text{rd}_\ell$  quella in  $F(\beta, m, b_{\min}, b_{\max})$  e con  $\text{rd}_d$  quella in  $F_d(\beta, m, b_{\min}, b_{\max})$  si ottiene allora:

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

Dunque: se  $\xi_{\min}^* \leq x \leq \xi_{\max}$  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 illimitato* 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):

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

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

$$|\epsilon(x)| \leq u \quad , \quad |\eta(x)| \leq u$$

e, quindi:

$$|\delta(x)| \leq u |x| \quad \text{oppure} \quad |\delta(x)| \leq u |\text{rd}(x)|$$

### 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:

$$\text{in } F(2, 53): \text{rd}(\pi) = 3.141592653589793115997963468544185161590576171875$$

e:

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

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

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

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

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

$$\text{in } F(2, 53): \text{rd}(0.1) = 0.100000000000000055511151231257827021181583404541015625$$

e:

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

In questo caso:

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

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

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

– Esiste un numero reale  $d$  tale che:

$$\text{rd}(x) = x + d \quad \text{e} \quad |d| \leq u|x|$$

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

– Esiste un numero reale  $d$  tale che:

$$x = \text{rd}(x) + d \quad \text{e} \quad |d| \leq u|x|$$

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

– Esiste un numero reale  $e$  tale che:

$$\text{rd}(x) = (1 + e)x \quad \text{e} \quad |e| \leq u$$

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

– Esiste un numero reale  $t$  tale che:

$$x = (1 + t)\text{rd}(x) \quad \text{e} \quad |t| \leq u$$

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

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

---

### Esercizi

---

*E28* Siano  $x = \frac{5}{4}$  e  $\text{rd}$  la funzione arrotondamento in  $F(2, 2)$  con RTTE. Determinare  $\text{rd}(x)$  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.

*E29* ♠ 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  $2^{-53}$ .

*E30* Sia  $M = F(\beta, m)$ . Discutere ciascuno dei seguenti asserti:

- (1) l'errore relativo commesso approssimando  $x \in \mathbb{R}$  con  $\text{rd}(x)$  è minore o uguale ad  $u$ ;
  - (2) l'errore assoluto commesso approssimando  $x \in \mathbb{R}$  con  $\text{rd}(x)$  è minore o uguale ad 1;
  - (3) ★ se  $x \in \mathbb{R}$  e  $\xi \in M$  sono tali che  $\text{rd}(x) = \xi$  allora  $\text{rd}(\beta^{12}x) = \beta^{12}\xi$ .
- 

## 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  $\text{rd}$  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*

$$\oplus, \ominus, \otimes : M \times M \rightarrow M \quad \text{tali che} \quad \xi_1 \otimes \xi_2 = \text{rd}(\xi_1 * \xi_2)$$

e:

$$\oslash : M \times M \setminus \{0\} \rightarrow M \quad \text{tale che} \quad \xi_1 \oslash \xi_2 = \text{rd}(\xi_1 / \xi_2)$$

- *Funzioni predefinite corrispondenti alle funzioni elementari*

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

$$F(\xi) = \text{rd}(f(\xi))$$

- *Funzioni predefinite corrispondenti ai confronti*

$$<, \leq, =, \neq, \geq, > : M \times M \rightarrow \{V, F\}$$

Sono le *restrizioni* ad  $M \times M$  delle corrispondenti funzioni sui numeri reali.<sup>11</sup>

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.”<sup>12</sup>

### 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:

---

<sup>11</sup>I valori  $V$  e  $F$  sono codificati, rispettivamente, dagli elementi 1 e 0 di  $M$ . Dunque anche i confronti sono funzioni a valori in  $M$ .

<sup>12</sup>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 delle 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” (si veda l'Esempio 0.3.4).

(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  $\xi_1 > \xi_2 \Rightarrow \xi_1 \oplus \alpha \geq \xi_2 \oplus \alpha$ ).

(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  $\xi_1 > \xi_2 \Rightarrow \xi_1 \otimes \alpha \geq \xi_2 \otimes \alpha$ ).

(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  $\xi$

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

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  $\xi$  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) —,  $\xi$  non ha elemento inverso).

(F.1) La funzione predefinita **SEN**, corrispondente alla funzione elementare *sen*, ha *un solo zero*:  $\xi = 0$  (infatti: l’uguaglianza **SEN**( $\xi$ ) = 0 equivale a  $\text{rd}(\text{sen } \xi) = 0$  ovvero  $\text{sen } \xi = 0$ , 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  $\phi : M \rightarrow M$  è una funzione predefinita corrispondente ad una funzione elementare *continua*,  $\phi(\xi) < 0$  e  $\phi(\theta) > 0$ , non è detto che esista  $\alpha$  tale che  $\phi(\alpha) = 0$  (ad esempio:  $1 \in M, 4 \in M, \text{SEN}(1) > 0$  e  $\text{SEN}(4) < 0$  ma per ogni  $\alpha \in M$  compreso tra 1 e 4 si ha  $\text{SEN}(\alpha) \neq 0$ ).

### 0.3.3 Osservazione (errore relativo per le funzioni predefinite)

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

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

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  $\xi_{\min}^* \leq |x| \leq \xi_{\max}$ .

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

Si consideri la funzione elementare *radice quadrata*.

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

$$\sqrt{2} = 1.41421356237$$

che coincide ( $\sqrt{2} = 1.41421356237 3095 04880 \dots$ ) 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:

$$\sqrt{2} = 2^1 0.1011010100000100111100110011001111110011101111001100 1001 \dots$$

e:

```
sqrt(2) = 2^1 0.1011010100000100111100110011001111110011101111001101
```

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:  $\text{rd}(\log_{10} 1000) = 3$ . La definizione della funzione predefinita è quindi *diversa* da quella della Definizione 0.3.1.

Si ha inoltre:

$$\sigma(\log_{10}(1000)) = 3 = \text{rd}(\log_{10} 1000)$$

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

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

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

---

### Esercizi

---

*E31* Sia  $M = F(10, 2)$ . Dimostrare, utilizzando le proprietà della funzione `rd` che:

- (1) Per ogni  $\xi$  si ha:  $\xi \oplus (-\xi) = 0$ ;
- (2) Per ogni  $\xi$  esiste un solo  $\alpha$  tale che:  $\xi \oplus \alpha = 0$ .

*E32* ★ Sia  $M = F(\beta, m)$ . Discutere ciascuno dei seguenti asseriti:

- (1) Se  $\xi$  ed  $\alpha$  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  $\text{rd}$  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 = pi;  
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(pi);  
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  $\text{rd}$ , 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:

$$y = 0.43301270189\ 22192\ 9829415103085921145975589752197265625$$

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:

$$y = \sin \frac{\pi}{6} \cos \frac{\pi}{6} = \frac{\sqrt{3}}{4} = 0.4330127018922192\ 3\dots$$

- (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 0.3.4:

$$x = 1.4142135623730951454746218587388284504413604736328125$$

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*:

$$x = \text{LOG10}(\text{rd}(1000))$$

Tenuto conto che  $1000 \in F(2, 53)$ , il valore di  $x$  dopo l'esecuzione in *Scilab* è  $\log_{10}(1000)$  ovvero, si veda l'Esempio 0.3.4:

$$x = 2.9999999999999999555910790149937383830547332763671875$$

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

$$x = 3$$

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

$$\begin{aligned} u &= 2^{-53}; \\ a &= -u; \\ b &= u; \\ x &= a + b + 1; \\ y &= a + (b + 1); \end{aligned}$$

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

$$\begin{aligned} u &= \text{rd}(\text{rd}(2)^{\text{rd}(-53)}); \\ a &= -u; \\ b &= u; \\ x &= (a \oplus b) \oplus \text{rd}(1); \\ y &= a \oplus (b \oplus \text{rd}(1)); \end{aligned}$$

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:

$$x = 1 \quad , \quad y = 0.999999999999999988897769753748434595763683319091796875$$

ovvero  $x \neq y$ .

– *Esercizio*

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

## B - Studio dell'errore

In questa sezione consideriamo il caso elementare e frequente in cui la procedura consista nell'assegnamento  $y = f(x)$  quando utilizza il tipo *numero reale* e nell'assegnamento  $y = \phi(x)$  quando utilizza il tipo *numero in virgola mobile e precisione finita*, con  $f$  e  $\phi$  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  $\text{rd}$ , e date una funzione  $f$  da  $\Omega \subset \mathbb{R}^n$  in  $\mathbb{R}$  ed una funzione  $\phi$  da  $\Omega$  in  $M$  (detta *algoritmo*) tali che:

- esiste una sequenza *finita* di funzioni predefinite  $fp_1, \dots, fp_j$  tale che:

$$\phi = fp_j \circ \dots \circ fp_1 \circ rd$$

nel senso che per ogni  $x \in \Omega$ , il numero  $\phi(x)$  è ottenuto arrotondando le componenti  $x_1, \dots, x_n$  ed utilizzando poi opportunamente, nell'ordine, le funzioni predefinite  $fp_1, \dots, fp_j$

- detta  $f_1, \dots, f_j$  la sequenza di funzioni elementari o operazioni aritmetiche corrispondente alla sequenza  $fp_1, \dots, fp_j$  si ha:

$$f = f_j \circ \dots \circ f_1$$

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

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:*

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

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.

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  $\Omega$  tale che  $f(x) \neq 0$ .

L'algoritmo  $\phi$  è accurato quando utilizzato per approssimare  $f$  in  $x$  se, posto:

$$\phi(x) = (1 + e_t) f(x) \quad \text{ovvero} \quad e_t = \frac{\phi(x) - f(x)}{f(x)}$$

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  $\phi$  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  $rd(f(x))$  e che per il Teorema 0.2.13 si ha:

$$rd(f(x)) = (1 + e_t) f(x) \quad \text{con} \quad |e_t| \leq u$$

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:

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

Posto  $\text{rd}(r) = \hat{r}$  e  $\text{rd}(h) = \hat{h}$ , 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:

$$\text{rd}(\pi) = (1 + \theta)\pi \quad \text{con} \quad |\theta| \leq u$$

e:

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

Posto:

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

si ottiene:

$$\phi(r, h) = (1 + e_v) f(\hat{r}, \hat{h}) \quad \text{e} \quad |e_v| \leq 5u + \dots \approx 5u$$

In quest'ultima uguaglianza si interpreta  $\phi(r, h)$  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  $\Omega$  diverso da zero.<sup>13</sup>

L'algoritmo  $\phi$  è stabile quando utilizzato per approssimare  $f$  in  $x$  se esistono numeri reali piccoli  $e_v, e_a$  tali che:

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

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

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:

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

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  $\phi$  un algoritmo stabile quando utilizzato per approssimare  $f$  in  $x = 0$ .

#### 0.4.6 Osservazione

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

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

$$\text{rd}(f(\text{rd}(x)))$$

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  $x$ .

<sup>13</sup>La definizione è data nel caso di  $f$  funzione di una variabile. Le modifiche da apportare nel caso generale sono ovvie.

Tenuto conto che, utilizzando il Teorema 0.2.13:

$$\text{rd}\left(f(\text{rd}(x))\right) = (1 + e_v) f((1 + e_a) x) \quad \text{con} \quad |e_v| \leq u \text{ e } |e_a| \leq u$$

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:

$$\phi(r, h) = (1 + e_v) f(\hat{r}, \hat{h}) \quad \text{con} \quad |e_v| \leq 5u + \dots \approx 5u$$

dunque, tenuto conto che:

$$\hat{r} = \text{rd}(r) = (1 + \rho)r \quad \text{con} \quad |\rho| \leq u$$

e:

$$\hat{h} = \text{rd}(h) = (1 + \omega)h \quad \text{con} \quad |\omega| \leq u$$

l'algoritmo  $\phi$ , 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  $e_d$  piccolo tale che*:

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

Si ottiene:

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

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

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

ovvero:

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

si ottiene:

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

Tenuto conto delle limitazioni per  $\rho$  e  $\omega$  e che  $r$  ed  $h$  sono numeri positivi si ottiene poi:

$$|e_d^s| = \left| \frac{r}{r + h} \rho + \frac{h}{r + h} \omega \right| \leq \left| \frac{r}{r + h} \right| |\rho| + \left| \frac{h}{r + h} \right| |\omega| \leq |\rho| + |\omega| \leq 2u$$

e infine, posto:

$$1 + e_d = (1 + \rho)(1 + e_d^s) \quad \text{ovvero} \quad e_d = \rho + e_d^s + \rho e_d^s$$

si conclude:

$$f(\hat{r}, \hat{h}) = (1 + e_d) f(r, h) \quad \text{con} \quad |e_d| \leq 3u + 2u^2 \approx 3u$$

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

Utilizzando i risultati ottenuti:

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

e, posto:

$$1 + e_t = (1 + e_v)(1 + e_d) \quad \text{ovvero} \quad e_t = e_v + e_d + e_v e_d$$

risulta:

$$\phi(r, h) = (1 + e_t) f(r, h) \quad \text{con} \quad |e_t| \leq 8u + \dots \approx 8u$$

ovvero: l'algoritmo  $\phi$ , 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  $\Omega$  diverso da zero e  $f(x) \neq 0$ .<sup>14</sup>

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

$$f((1 + e_a)x) = (1 + e_d^f) f(x) \quad \text{ovvero} \quad e_d^f = \frac{f((1 + e_a)x) - f(x)}{f(x)}$$

l'errore relativo  $e_d^f$  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  $f((1 + e_a)x)$  come perturbazione moltiplicativa di  $f(x)$ . In questo caso la nozione di calcolo ben condizionato va definita interpretando  $f((1 + e_a)x)$  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 $\Rightarrow$ accuratezza)

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

Se  $\phi$  è un algoritmo stabile quando utilizzato per approssimare  $f$  in  $x$  e il calcolo di  $f$  in  $x$  è ben condizionato, allora  $\phi$  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:

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

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

$$e_d^f = \frac{f((1 + e_a)x) - f(x)}{f(x)} \quad \text{ovvero} \quad f((1 + e_a)x) = (1 + e_d^f) f(x)$$

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

$$1 + e_t = (1 + e_v)(1 + e_d^f) \quad \text{ovvero} \quad e_t = e_v + e_d^f + e_v e_d^f$$

si ottiene:

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

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

#### 0.4.11 Osservazione (condizionamento delle funzioni regolari)

Siano  $\Omega$  un intervallo di  $\mathbb{R}$ ,  $f : \Omega \rightarrow \mathbb{R}$  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:

$$f((1 + e_a)x) = (1 + e_d^f) f(x) \quad \text{ovvero} \quad e_d^f = \frac{f((1 + e_a)x) - f(x)}{f(x)}$$

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

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

dunque:

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

---

<sup>14</sup>La definizione è data nel caso di  $f$  funzione di una variabile. Le modifiche da apportare nel caso generale sono ovvie.

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

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

Introdotta il numero di condizionamento del calcolo di  $f$  in  $x$ :

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

si ottiene infine:

$$|e_d^f| \approx c_f(x) |e_a|$$

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:

$$f(x) = \text{sen } x \quad , \quad \phi(x) = \text{SEN}(\text{rd}(x))$$

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

- *Stabilità dell'algoritmo  $\phi$  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

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

L'algoritmo  $\phi$  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:

$$f((1 + e_a)x) = (1 + e_d^f) f(x) \quad \text{con} \quad |e_d^f| \approx c_f(x) |e_a|$$

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  $\phi$  quando utilizzato per approssimare  $f(x)$ :*

In base al Teorema 0.4.10, l'algoritmo  $\phi$  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:

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

e, utilizzando le limitazioni

$$|e_v| \leq u \quad , \quad |e_d^f| \approx c_f(x) |e_a| \leq u$$

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

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

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

– *Esercizio*

La funzione numero di condizionamento del calcolo di  $\sin x$ :

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

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

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

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 dedurre che il calcolo di  $\sin x$  risulta ragionevolmente ben condizionato (e quindi, per il Teorema 0.4.10, l'algoritmo  $\phi$  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:

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

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

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

per la *moltiplicazione*:

$$e_d^m = e_1 + e_2 + e_1 e_2$$

e per la *divisione*:

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

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:

$$|e_d^m| \leq |e_1| + |e_2| + |e_1||e_2| \approx |e_1| + |e_2|$$

e per la divisione:

$$|e_d^d| \leq \frac{|e_1| + |e_2|}{1 - |e_2|} \approx |e_1| + |e_2|$$

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:

$$|e_d^s| \leq \left| \frac{x_1}{x_1 + x_2} \right| |e_1| + \left| \frac{x_2}{x_1 + x_2} \right| |e_2| < |e_1| + |e_2|$$

– Se gli addendi hanno segno opposto, il condizionamento del calcolo 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:

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

e quindi:

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



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

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

e:

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

Infine:

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

e:

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

L'errore  $|e_d^s|$  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  $\text{fp}$  la funzione predefinita corrispondente ad  $f$ . Per la Definizione 0.3.1, per ogni  $\xi \in \Omega \cap M$  si ha:  $\text{fp}(\xi) = \text{rd}(f(\xi))$ .

Il procedimento utilizzato nell'Esempio 0.4.12 per mostrare la stabilità prova che l'algoritmo  $\phi$  definito da  $\phi(x) = \text{fp}(\text{rd}(x))$  — definito nell'insieme  $\Omega^* \subset \Omega$  dei punti  $x \in \Omega$  tali che  $\text{rd}(x) \in \Omega$  — 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  $\otimes$  la funzione predefinita (ovvero la pseudo-operazione aritmetica) corrispondente a  $*$ .

L'algoritmo  $\phi$  definito da  $\phi(x_1, x_2) = \text{rd}(x_1) \otimes \text{rd}(x_2)$  — definito nell'insieme  $\Omega^* \subset \Omega$  dei punti  $(x_1, x_2) \in \Omega$  tali che  $(\text{rd}(x_1), \text{rd}(x_2)) \in \Omega$  — 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:

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

e:

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

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  $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R}$  due funzioni e  $\phi_1, \phi_2$  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  $\phi_1$  quando utilizzato per approssimare  $f_1$  in  $x$ , esistono numeri reali  $e_{v1}, e_{a1}$  tali che:

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

Tenuto conto della stabilità di  $\phi_2$  quando utilizzato per approssimare  $f_2$  in  $\phi_1(x)$ , esistono numeri reali  $e_{v2}, e_{a2}$  tale che:

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

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

$$\gamma(x) = (1 + e_{v2}) f_2 \left( (1 + e_{a2})(1 + e_{v1}) f_1((1 + e_{a1})x) \right)$$

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

$$f_2 \left( (1 + e_{a2})(1 + e_{v1}) f_1((1 + e_{a1})x) \right) = (1 + e_d^{f_2}) f_2 \left( f_1((1 + e_{a1})x) \right)$$

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

$$\gamma(x) = (1 + e_{v2})(1 + e_d^{f_2}) f_2 \left( f_1((1 + e_{a1})x) \right) = (1 + e_{v2})(1 + e_d^{f_2}) g((1 + e_{a1})x)$$

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

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

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

- Se il calcolo di  $f_2$  in  $f_1((1 + e_{a1})x)$  è *ben condizionato* allora  $e_d^{f_2}$  risulta piccolo. Dunque anche  $e_v$  lo è e l'algoritmo  $\gamma$  è *stabile*.
- Se il calcolo di  $f_2$  in  $f_1((1 + e_{a1})x)$  *non* è ben condizionato allora l'algoritmo  $\gamma$  *può* risultare *non stabile*.

#### 0.4.16 Esempio

Siano:

$$f(x) = 1 - \cos x \quad , \quad \phi(x) = 1 \ominus \overset{2}{\ominus} \overset{1}{\text{COS}}(\text{rd}(x))$$

e si utilizzi  $\phi$  per approssimare il valore di  $f$  in  $x$ .

Per il Teorema 0.2.13, esistono numeri reali  $\theta, e_1$  ed  $e_2$  tali che:

$$\phi(x) = (1 + e_2) \left( 1 - (1 + e_1) \cos((1 + \theta)x) \right) \quad \text{e} \quad |\theta| \leq u, |e_1| \leq u, |e_2| \leq u$$

Per l'Osservazione 0.4.13, posto  $\xi = \text{rd}(x) = (1 + \theta)x$  e:

$$e_d^s = \frac{\cos \xi}{1 - \cos \xi} e_1$$

si ha:

$$\phi(x) = (1 + e_2)(1 - e_d^s) \left( 1 - \cos((1 + \theta)x) \right)$$

Poiché:

$$\lim_{y \rightarrow 0} \left| \frac{\cos y}{1 - \cos y} \right| = +\infty$$

il calcolo di  $1 - \cos \xi$  non è ben condizionato per  $\xi$  vicino a zero e quindi per l'osservazione precedente l'algoritmo  $\phi$  *può risultare* non stabile quando utilizzato per approssimare  $f(x)$  per  $x$  vicino a zero.

Per mostrare che effettivamente  $\phi$  non è stabile si osservi che, essendo:

$$c_f(x) = \left| \frac{f'(x)}{f(x)} x \right| = \left| \frac{\text{sen } x}{1 - \cos x} x \right| \quad \text{e} \quad \lim_{x \rightarrow 0} c_f(x) = 2$$

il calcolo di  $f(x)$  per  $x$  vicino a zero è ben condizionato. Se  $\phi$  fosse stabile quando utilizzato per approssimare il valore di  $f(x)$  allora, per il Teorema 0.4.10, sarebbe anche accurato: esisterebbe un numero reale  $e_t$  piccolo (ovvero che in valore assoluto non supera un multiplo *non troppo grande* di  $u$ ) tale che:

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

L'Esercizio E41 mostra, utilizzando *Scilab*, che esistono valori di  $x$  vicini a zero per i quali si ottiene  $|e_t| > 10^6 u$ .

---

Esercizi

---

*E34* 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.

*E35* ★ Realizzando la procedura del punto (4) dell'Esempio 0.4.1 con la calcolatrice *HP 49G* si ottiene  $x = y = 1$ . Spiegare questi risultati e poi indicare come modificare l'assegnamento che definisce il valore di  $u$  in modo da ottenere anche in questo caso  $x \neq y$ .

*E36* 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  $x$ .

*E37* Per ogni  $x > 0$  sia  $f(x) = 1/\sqrt{x}$ . Determinare l'insieme di definizione e discutere stabilità ed accuratezza dell'algoritmo:

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

quando utilizzato per approssimare i valori di  $f$ .

*E38* 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:

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

quando utilizzato per approssimare i valori di  $f$ .

*E39* ★ Siano  $f : \mathbb{R} \rightarrow \mathbb{R}$  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:

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

*E40* Si consideri l'Esempio 0.4.12. Tenuto conto che in *Scilab* si ha:  $\%pi = \text{rd}(\pi) < \pi$  e  $\phi(\pi) = \text{SEN}(\%pi) > 0$ :

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

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

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

si ha:

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

ovvero: per  $x$  vicino a  $\pi$  l'algoritmo  $\phi$  non è accurato quando utilizzato per approssimare  $\text{sen } x$ .

*E41* ♠ Si consideri l'Esempio 0.4.16.

(1) Per ogni  $x \in \mathbb{R}$  si ha:  $\cos x = \cos(\frac{1}{2}x + \frac{1}{2}x)$ . Dimostrare che:

$$f(x) = 2 (\text{sen}(x/2))^2$$

Siano  $M = F(2, 53)$ ,  $\text{SQR}$  la funzione predefinita corrispondente alla funzione quadrato e  $\psi : \mathbb{R} \rightarrow M$  l'algoritmo definito da:

$$\psi(x) = 2 \otimes \text{SQR}(\text{SEN}(\text{rd}(x) \oslash 2))$$

(2) Dimostrare che per ogni  $x \in \mathbb{R}$  esiste  $\epsilon \in \mathbb{R}$  tale che:

$$\psi(x) = (1 + \epsilon) f(x) \quad \text{e} \quad |\epsilon| \leq 5u + \dots \approx 5u$$

(3) Mostrare che per ogni  $x \in \mathbb{R}$ , posto:

$$E = \frac{\phi(x) - \psi(x)}{\psi(x)}$$

si ha:

$$\phi(x) = (1 + E)\psi(x) = (1 + E)(1 + \epsilon)f(x)$$

ovvero:

$$\phi(x) = (1 + e_t)f(x) \quad \text{con} \quad e_t = E + \epsilon + E\epsilon$$

Si osservi che, per la limitazione su  $|\epsilon|$ , se  $E = ku$  con  $k$  molto maggiore di 5 allora  $e_t \approx E = ku$ .

(4) Per ogni  $x \in \mathbb{R}$  tale che  $\psi(x) \neq 0$ , sia:

$$err(x) = \left| \frac{\phi(x) - \psi(x)}{\psi(x)} \right|$$

Detta  $u$  la precisione di macchina in  $M$  sia infine  $k(x)$  tale che:  $err(x) = k(x)u$ . Per disegnare il grafico di  $k(x)$  con  $x \in [10^{-10}, 10^{-2}]$  utilizzando *Scilab*, dopo aver opportunamente:

- creato una *function* di nome `err` da usare per approssimare i valori della funzione *err*
- creato una variabile di nome `u` ed averle assegnato valore  $u$
- assegnato un valore alla variabile `x`

si esegue il comando:

```
plot2d('nl', x, err(x) / u, style = 2)
```

Dedurne che esistono valori di  $x$  vicini a zero per i quali si ha  $\phi(x) = (1 + e_t)f(x)$  con  $|e_t| > 10^6 u$ .

---