R. Bevilacqua **O. Menchi**

ESERCIZI DI CALCOLO NUMERICO

Questa raccolta di esercizi si propone come integrazione degli

Appunti di Calcolo Numerico

(R. Bevilacqua, D. Bini, M. Capovani, O. Menchi, Servizio Editoriale Universitario di Pisa)

e contiene gli esercizi assegnati alle prove di esame di Calcolo Numerico dei corsi di studio in Informatica nel periodo 2001-08. Di una parte di essi è riportata la risoluzione.

Capitolo 1

Analisi dell'errore

1.1 Richiami di teoria

Prerequisiti: proprietà dei moduli, disuguaglianze e maggiorazioni di funzioni, nozioni elementari di calcolo differenziale, in particolare la formula di Taylor.

Il calcolatore opera su numeri rappresentati mediante sequenze finite di cifre, quindi già nella fase di immissione dei dati il troncamento delle cifre genera errori di rappresentazione. Inoltre l'uso di un'aritmetica finita introduce errori anche nell'esecuzione delle operazioni aritmetiche. Durante il calcolo gli errori si propagano ed è importante riuscire a valutarne la consistenza.

Rappresentazione in base

- Sia $\beta \geq 2$ la base scelta per la rappresentazione e sia x un numero reale non nullo; allora esistono e sono unici il numero intero p (detto **esponente** e la successione $\{d_i\}_{i=1,2,\ldots}$ di numeri interi (dette **cifre**), $0 \leq d_i < \beta$, $d_1 \neq 0$, non definitivamente uguali a $\beta - 1$, tali che

$$x = \mathrm{sgn}(x)\beta^p \sum_{i=1}^{\infty} d_i \beta^{-i},$$

dove $\mathrm{sgn}(x) = 1$ se $x > 0$, $\mathrm{sgn}(x) = -1$ se $x < 0$. Poiché $d_1 \neq 0$, la rappresentazione è **normalizzata**.

- Per la rappresentazione effettiva si usa la notazione posizionale

$$x = \pm\, (0.d_1 d_2 \ldots)_\beta \ \beta^p \quad \text{o} \quad x = \pm\, 0.d_1 d_2 \ldots \ \beta^p.$$

La base β, quando è chiara dal contesto, non viene indicata. la rappresentazione è finita se esiste un indice k tale che $d_k \neq 0$ e $d_i = 0$ per $i > k$.

- Per convertire un numero dalla base 10 alla base β, lo si scrive come somma della sua parte intera e della sua parte decimale, che si convertono separatamente. Per la parte intera si utilizza un procedimento di divisioni successive. Infatti se x è intero si ha

$$x = \beta^p \sum_{i=1}^{p} d_i \beta^{-i} = \beta\ q_1 + d_p,$$

dove q_1 è il quoziente della divisione di x per β e d_p è il resto. Si è così ricavata l'ultima cifra della rappresentazione in base β di x. Poiché

$$q_1 = \sum_{i=1}^{p-1} d_i \beta^{(p-1)-i},$$

riapplicando il procedimento a q_1 si ricava d_{p-1} e così via.

- Per la parte decimale si utilizza un procedimento di moltiplicazioni successive. Infatti, se $0 < x < 1$ come prima cosa si determina l'esponente p, che è minore o uguale a 0, mediante la successione

$$y_1 = \beta\, x, \quad y_2 = \beta^2\, x, \quad \ldots, \quad y_j = \beta^j\, x,$$

arrestandosi al primo indice j per cui $y_j \geq 1$. È

$$p = 1 - j \quad \text{e} \quad y_j = \beta \sum_{i=1}^{\infty} d_i \beta^{-i} = d_1 + y_{j+1}, \quad \text{dove} \quad y_{j+1} = \sum_{i=2}^{\infty} d_i \beta^{-(i-1)}.$$

Poiché $y_{j+1} < 1$, d_1 risulta essere la parte intera di y_j. Si è così trovata la prima cifra della rappresentazione in base β di x. Si riapplica il procedimento a y_{j+1}, ottenendo d_2 e così via.

Numeri di macchina

- Dati $t,\ m,\ M$, numeri interi tali che $t \geq 1$, $m,\ M > 0$, si definisce insieme dei **numeri di macchina** in base β con t **cifre significative**, l'insieme

$$\mathcal{F}_{(\beta,t,m,M)} = \{0\} \cup \{x \in \mathbf{R} \ :\ x = \mathrm{sgn}(x)\beta^p \sum_{i=1}^{t} d_i \beta^{-i}\},$$

dove $-m \leq p \leq M$ e $0 \leq d_i < \beta$ per $i = 1, 2, \ldots, t$, con $d_1 \neq 0$. Quindi tutti i numeri di macchina, eccetto lo zero, hanno una rappresentazione normalizzata. Si usa dire che i numeri di $\mathcal{F}_{(\beta,t,m,M)}$ sono rappresentati in virgola mobile (dall'inglese floating point).

- Il minimo numero positivo di $\mathcal{F}_{(\beta,t,m,M)}$ è

$$\omega = 0.1\ \beta^{-m} = \beta^{-m-1},$$

il massimo è

$$\Omega = \beta^M(\beta - 1)\sum_{i=1}^{t} \beta^{-i} = \beta^M(1 - \beta^{-t}).$$

L'insieme $\mathcal{F}_{(\beta,t,m,M)}$, oltre allo zero, contiene $(m+M+1)(\beta^t - \beta^{t-1})$ numeri positivi compresi fra ω e Ω e altrettanti numeri negativi compresi fra $-\Omega$ e $-\omega$.

- Se il numero reale non nullo $x = \pm 0.d_1 d_2 \ldots\ \beta^p$ è tale che $-m \leq p \leq M$, $d_1 \neq 0$ e $d_i = 0$, per $i > t$, allora $x \in \mathcal{F}_{(\beta,t,m,M)}$. Se x non appartiene a $\mathcal{F}_{(\beta,t,m,M)}$, si pone di rappresentare x con un numero di macchina $\widetilde{x}$.

• Se $p < -m$, si verifica la situazione di **underflow**, se $p > M$, si verifica la situazione di **overflow**. In tal caso il calcolo può essere interrotto o può continuare con l'assegnazione di valori predefiniti (per esempio ω e Ω).

• Se l'esponente p appartiene all'intervallo $[-m, M]$ ma x ha più di t cifre x viene rappresentato con uno dei due seguenti numeri di macchina:

$$\mathrm{trn}(x) = \beta^p \sum_{i=1}^{t} d_i \beta^{-i},$$

ottenuto per **troncamento** di x alla t-esima cifra, e

$$\mathrm{arr}(x) = \begin{cases} \mathrm{trn}(x) & \text{se } d_{t+1} < \beta/2, \\ \mathrm{trn}(x) + \beta^{p-t} & \text{se } d_{t+1} \geq \beta/2, \end{cases}$$

ottenuto per **arrotondamento** di x alla t-esima cifra.

• Per valutare l'errore commesso nel rappresentare un numero reale $x > 0$ con un numero di macchina $\widetilde{x}$, si considerano le seguenti quantità

$$\widetilde{x} - x \quad \text{errore assoluto}, \quad \epsilon_x = \frac{\widetilde{x} - x}{x} \quad \text{errore relativo}.$$

ϵ_x viene detto **errore di rappresentazione** di x.

• Se non si verificano situazioni di underflow o di overflow risulta

$$|\mathrm{trn}(x) - x| < \beta^{p-t} \quad \text{e} \quad |\mathrm{arr}(x) - x| \leq \frac{1}{2}\beta^{p-t},$$

dove il segno di uguaglianza vale se e solo se $d_{t+1} = \dfrac{\beta}{2}$ e $d_{t+i} = 0,\ i \geq 2$, e

$$\left|\frac{\widetilde{x} - x}{x}\right| < u, \quad \text{dove} \quad u = \begin{cases} \beta^{1-t} & \text{se } \widetilde{x} = \mathrm{trn}(x), \\ \dfrac{1}{2}\,\beta^{1-t} & \text{se } \widetilde{x} = \mathrm{arr}(x). \end{cases}$$

La quantità u è detta **precisione di macchina**. La relazione precedente può anche essere scritta nella forma

$$\widetilde{x} = x(1 + \epsilon_x), \quad \text{con} \quad |\epsilon_x| < u.$$

• In generale il risultato di un'operazione aritmetica fra numeri di macchina non è un numero di macchina. Occorre perciò definire un'aritmetica di macchina. Indicando con $\odot$ l'operazione di macchina che approssima l'operazione esatta $\cdot$, per tutti i numeri di macchina x e y per cui l'operazione non dia luogo a condizioni di underflow o di overflow, deve valere una relazione analoga a quella dell'approssimazione di un singolo numero, cioè

$$x \ \odot \ y = (x \ \cdot \ y)(1 + \epsilon), \quad |\epsilon| < u,$$

dove ϵ è detto **errore locale** dell'operazione.

- Non tutte le proprietà algebriche delle operazioni nel campo reale sono soddisfatte dalle operazioni di macchina. Ad esempio, non valgono la proprietà associativa dell'addizione e della moltiplicazione e quella distributiva della moltiplicazione rispetto all'addizione.

Errore nel calcolo di una funzione

- Sia $f(x_1, x_2, \ldots, x_n)$ una funzione razionale su $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Se $\widetilde{\boldsymbol{x}} = \{\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n\}$ è la rappresentazione di macchina di $\boldsymbol{x}$, e ψ è la funzione di macchina che approssima la f, il valore effettivamente calcolato è $\psi(\widetilde{\boldsymbol{x}})$. Per $f(\boldsymbol{x}) \neq 0$ e $f(\widetilde{\boldsymbol{x}}) \neq 0$, si definiscono i seguenti errori:

errore totale di $\psi(\widetilde{\boldsymbol{x}})$ rispetto a $f(\boldsymbol{x})$ $\quad \epsilon_{tot} = \dfrac{\psi(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x})}{f(\boldsymbol{x})}$,

errore inerente $\quad \epsilon_{in} = \dfrac{f(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x})}{f(\boldsymbol{x})}$,

errore algoritmico $\quad \epsilon_{alg} = \dfrac{\psi(\widetilde{\boldsymbol{x}}) - f(\widetilde{\boldsymbol{x}})}{f(\widetilde{\boldsymbol{x}})}$.

In un'analisi dell'errore al primo ordine vale

$$\epsilon_{tot} \doteq \epsilon_{alg} + \epsilon_{in}.$$

- Se la funzione f è differenziabile due volte in un intorno di $\boldsymbol{x}$, e $x_i \neq 0$ per $i = 1, \ldots, n$, in un'analisi dell'errore al primo ordine si ha

$$\epsilon_{in} \doteq \sum_{i=1}^{n} c_i \epsilon_i, \quad \text{dove} \quad c_i = \frac{x_i}{f(\boldsymbol{x})} \frac{\partial f(\boldsymbol{x})}{\partial x_i} \quad \text{e} \quad \epsilon_i = \frac{\widetilde{x}_i - x_i}{x_i}.$$

I coefficienti c_i sono detti **coefficienti di amplificazione**. Se almeno uno dei coefficienti di amplificazione ha modulo elevato, il problema del calcolo di $f(\boldsymbol{x})$ è **mal condizionato**.

- Se la funzione f è una delle quattro operazioni aritmetiche vale
  se $f(x_1, x_2) = x_1 \pm x_2$, è $\quad c_1 = \dfrac{x_1}{x_1 \pm x_2}$, $\quad c_2 = \dfrac{\pm x_2}{x_1 \pm x_2}$,
  se $f(x_1, x_2) = x_1 x_2$, è $\quad c_1 = 1, \quad c_2 = 1$,
  se $f(x_1, x_2) = x_1 / x_2$, è $\quad c_1 = 1, \quad c_2 = -1$.
- Nel caso dell'addizione di due numeri x_1 e x_2 di segno opposto (o della sottrazione di due numeri x_1 e x_2 dello stesso segno), non è in generale possibile dare una limitazione superiore del modulo dell'errore totale indipendente da x_1 e x_2. Tanto più piccolo è il modulo di $x_1 + x_2$, tanto più grandi sono i moduli dei coefficienti di amplificazione c_1 e c_2 (fenomeno di cancellazione numerica).
- Se la funzione f è composta da più operazioni aritmetiche, il valore $\psi(\widetilde{\boldsymbol{x}})$ è ottenuto sostituendo alle operazioni aritmetiche le corrispondenti operazioni di macchina, cioè è ottenuto implementando un algoritmo. In questo caso per determinare l'errore

algoritmico conviene utilizzare un grafo, che descrive la sequenza delle operazioni dell'algoritmo. Se l'errore algoritmico è superiormente limitabile in modulo si dice che l'algoritmo è **numericamente stabile**. In caso contrario si dice che l'algoritmo è **numericamente instabile** perché si può presentare una elevata propagazione dell'errore.

• Se la funzione f non è razionale, è necessario approssimarla con una funzione razionale g: tale approssimazione introduce un **errore analitico**

$$\epsilon_{an}(\boldsymbol{x}) = \frac{g(\boldsymbol{x}) - f(\boldsymbol{x})}{f(\boldsymbol{x})}.$$

Detta ancora ψ la funzione effettivamente calcolata al posto della g, se $g(\widetilde{\boldsymbol{x}}) \neq 0$, in un'analisi dell'errore al primo ordine si ha

$$\epsilon_{tot} \doteq \epsilon_{in} + \epsilon_{alg} + \epsilon_{an}.$$

• Il modo più semplice per ottenere un'approssimazione razionale di una funzione non razionale è quello di utilizzare la formula di Taylor troncata ad un termine opportuno, scelto in modo che l'errore analitico sia dello stesso ordine dell'errore algoritmico.

• Molte funzioni non razionali fra le più comuni possono essere calcolate utilizzando programmi inclusi in librerie di software, che implementano algoritmi per i quali gli errori algoritmici e analitici corrispondenti sono limitati superiormente in modulo da quantità dell'ordine della precisione di macchina. La valutazione dell'errore totale da cui è affetta una funzione non razionale può ancora essere fatta usando i grafi.

1.2 Esercizi svolti

1.2.1 Si considerino i due insiemi $\mathcal{F} = \mathcal{F}_{(10,t,m,M)}$, con $t = 3$, $m = 4$, $M = 5$ e $\mathcal{G} = \mathcal{G}_{(10,t,m,M)}$ definito come l'unione di $\mathcal{F}$ con l'insieme dei numeri non nulli della forma $x = \pm 10^{-m}\,(0.0d_2, \ldots, d_t)$, con $0 \leq d_i \leq 9$, per $i = 2, \ldots, t$ (quindi $\mathcal{G}$ contiene anche numeri piccoli non normalizzati).

a) Si calcolino i minimi positivi non nulli ω e i massimi Ω degli insiemi $\mathcal{F}$ e $\mathcal{G}$.

b) Si determinino le cardinalità degli insiemi $\mathcal{F}$ e $\mathcal{G}$.

c) Quale errore relativo di rappresentazione si commette volendo rappresentare $1.4\ 10^{-(t+m)}$ in $\mathcal{F}$ e in $\mathcal{G}$. (8/11/2007)

Soluzione

a) Per $\mathcal{F}$ si ha

$$\omega_{\mathcal{F}} = 10^{-4}\,(.1) = 10^{-5}, \qquad \Omega_{\mathcal{F}} = 10^{5}\,(.999) = 10^{5}\,(1 - 10^{-3}).$$

Per $\mathcal{G}$ si ha

$$\omega_{\mathcal{G}} = 10^{-4}\,(.001) = 10^{-7}, \qquad \Omega_{\mathcal{G}} = \Omega_{\mathcal{F}}.$$

b) La cardinalità di $\mathcal{F}$ è $1 + 2(M + m + 1) \cdot 9 \cdot 10^{t-1} = 18001$. I numeri di $\mathcal{G}$ non appartenenti ad $\mathcal{F}$ sono $2(10^2 - 1)$. Quindi la cardinalità di $\mathcal{G}$ è 18199.

c) Sia $x = 1.4\ 10^{-7} = 10^{-4}\,(.0014)$. In $\mathcal{G}$ è $\widetilde{x} = \mathrm{arr}(x) = 10^{-4}\,(.001) = 10^{-7}$. L'errore relativo è

$$\frac{\widetilde{x} - x}{x} = -\frac{4\ 10^{-8}}{10^{-7}} = -\frac{20}{7}\,10^{-1},$$

quindi assai superiore alla precisione di macchina che è $10^{-2}/2$. In $\mathcal{F}$ il numero x verrebbe rappresentato con $\omega_{\mathcal{F}}$ e il suo errore relativo sarebbe molto più alto.

1.2.2 Si consideri il numero reale x che ha in base 2 la seguente rappresentazione periodica

$$x = 0.\overline{10}_2$$

a) Si scrivano in $\mathcal{F}_{(2,6,m,M)}$ i numeri $x_t = \mathrm{trn}(x)$ e $x_a = \mathrm{arr}(x)$.

b) Si scriva x come frazione con numeratore e denominatore interi in base 10.

c) Si dica se la rappresentazione in base 10 di x è finita, oppure periodica, oppure infinita non periodica.

d) Si scrivano x_t e x_a in base 10 e si dica quanto valgono i loro errori relativi. (6/7/2004)

Soluzione

a) In $\mathcal{F}_{(2,6,m,M)}$ è $x_t = (0.101010)_2$ e $x_a = (0.101011)_2$.

b) Sapendo che la somma di una serie geometrica di ragione r, con $|r| < 1$, è data da

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r},$$

si ha

$$x = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \ldots = \frac{1}{2}\sum_{i=0}^{\infty}\frac{1}{4^i} = \frac{2}{3}.$$

c) La rapppresentazione di x in base 10 è periodica.

d)

$$x_t = (0.101010)_2 = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} = \frac{21}{32} \quad \text{e} \quad x_a = x_t + \frac{1}{64} = \frac{43}{64}.$$

$$\epsilon_t = \frac{x_t - x}{x} = -\frac{1}{64} \quad \text{e} \quad \epsilon_a = \frac{x_a - x}{x} = \frac{1}{128}.$$

1.2.3 Supponendo di operare con arrotondamento

a) si determinino i primi due interi positivi che hanno la stessa rappresentazione in $\mathcal{F}_{(2,3,5,4)}$.

b) Stessa domanda del punto a) per $\mathcal{F}_{(2,24,128,127)}$. (6/6/2002)

Soluzione

a) Con tre cifre significative gli interi da 1 a 8 vengono rappresentati esattamente, mentre $9_{10} = 2^4\,(.1001)_2$ viene arrotondato a $2^4\,(.101)_2 = 10_{10}$. Quindi 9 e 10 sono i due interi cercati.

b) Sulla base di quanto ottenuto al punto precedente si trova subito che i due interi cercati sono $2^{24}+1$ e $2^{24}+2$. Infatti tutti gli interi minori o uguali a 2^{24} si possono rappresentare esattamente con 24 cifre binarie, mentre $2^{24}+1$ deve essere arrotondato a $2^{24}+2$, che invece si può rappresentare esattamente.

1.2.4 Sia $\mathcal{F} = \mathcal{F}_{(2,4,m,M)}$ l'insieme dei numeri di macchina con arrotondamento. Sono dati i numeri

$$x = \frac{1}{10}, \qquad y = \frac{1}{3}, \qquad z = \frac{7}{9}.$$

a) Si calcolino i valori approssimati $\widetilde{x},\ \widetilde{y},\ \widetilde{z}$ in $\mathcal{F}$.

b) Per trovare $R = (xy)/z$ si calcolino

$$r_1 = (\widetilde{x} \otimes \widetilde{y}) \oslash \widetilde{z}, \qquad \text{e} \qquad r_2 = \widetilde{x} \otimes (\widetilde{y} \oslash \widetilde{z}).$$

Si dica se i due risultati sono uguali o diversi e se uno o entrambi sono uguali a $\mathrm{arr}(R)$. (4/11/2003)

Soluzione

a) In $\mathcal{F}_{(2,4,m,M)}$ è $x = 0.0\overline{0011}_2$, $y = 0.\overline{01}_2$, $z = 0.\overline{110001}_2$. Arrotondando alla quarta cifra significativa si ha

$$\widetilde{x} = 2^{-3}\ 0.1101_2, \quad \widetilde{y} = 2^{-1}\ 0.1011_2, \quad \widetilde{z} = 0.11_2.$$

b) Con il primo algoritmo, si ha

$$p_1 = \widetilde{x} \otimes \widetilde{y} = \mathrm{arr}(\widetilde{x} \times \widetilde{y}) = 2^{-4}\ \mathrm{arr}(\,0.1101_2 \times 0.1011_2\,).$$

Si calcola il prodotto dei due numeri in base 2 applicando la solita regola distributiva della moltiplicazione

$$\begin{array}{rcl} 0.1101 & \times & 0.1011 \\ \hline 1101 & & \\ 1101 & & \\ 1101 & & \\ \hline 0.10001111 & & \end{array}$$

e arrotondando si ottiene $p_1 = 2^{-4}\ 0.1001_2$. Poi si ha

$$r_1 = p_1 \oslash \widetilde{z} = \mathrm{arr}(\, p_1 \,/\, \widetilde{z}) = 2^{-4}\ \mathrm{arr}(\, 1001_2 \,/\, 1100_2\,).$$

Si calcola il quoziente dei due numeri in base 2 con la solita regola

$$\begin{array}{rr|l} 1\,0\,0\,1\,.\,0 & & 1\,1\,0\,0 \\ \cline{3-3} -\quad 1\,1\,0\,0 & & 0\,.\,1\,1 \\ \cline{1-1} \quad 1\,1\,0\,0 & & \\ -\quad 1\,1\,0\,0 & & \\ \cline{1-1} \quad 0 & & \end{array}$$

Il quoziente ha solo 2 cifre non nulle, quindi non c'è bisogno di arrotondare ed è $p_1 = 2^{-4}\ 0.11_2$.

Con il secondo algoritmo, si ha

$$d_2 = \widetilde{y} \oslash \widetilde{z} = \mathrm{arr}(\, \widetilde{y} \,/\, \widetilde{z}) = 2^{-1}\ \mathrm{arr}(\, 1011_2 \,/\, 1100_2\,).$$

Procedendo come sopra, si calcola il quoziente

$$1011_2 \,/\, 1100_2 = 0.11\overline{10}_2$$

e arrotondando si ottiene $d_2 = 2^{-1}\ 0.1111_2$. Poi si ha

$$r_2 = \widetilde{x} \otimes d_2 = \mathrm{arr}(\, \widetilde{x} \times d_2) = 2^{-4}\ \mathrm{arr}(\, 0.1101_2 \times 0.1111_2\,).$$

Si calcola il prodotto

$$0.1101_2 \times 0.1111_2 = 0.11000011_2,$$

e arrotondando si ottiene $r_2 = 2^{-4}\ 0.11_2$. In questo caso è risultato che $r_1 = r_2$. In generale però la proprietà associativa non vale per le operazioni di macchina. Poiché

$$\mathrm{arr}(R) = 2^{-4}\ 0.1011_2,$$

si ha $\mathrm{arr}(R) \neq r_1$.

È possibile evitare il calcolo delle operazioni in base 2 procedendo in questo modo:

si riconvertono $\widetilde{x}$, $\widetilde{y}$ e $\widetilde{z}$ in base 10, ottenendo

$$\widetilde{x} = \frac{13}{128}\,, \quad \widetilde{y} = \frac{11}{32}\,, \quad \widetilde{z} = \frac{3}{4}$$

si calcola p_1 moltiplicando le frazioni, convertendo il risultato in base 2 e arrotondando

$$p_1 = \mathrm{arr}\Big(\frac{13}{128} \times \frac{11}{32}\Big) = \mathrm{arr}\Big(\frac{143}{4076}\Big) = (0.00001001)_2$$

si riconverte p_1 in base 10 ottenedo

$$p_1 = \frac{9}{256}$$

si calcola r_1 dividendo le frazioni, convertendo il risultato in base 2 e arrotondando

$$r_1 = \mathrm{arr}\Big(\frac{9}{256} \,/\, \frac{3}{4}\Big) = \mathrm{arr}\Big(\frac{3}{64}\Big) = (0.000011)_2.$$

Per il secondo algoritmo si procede nello stesso modo.

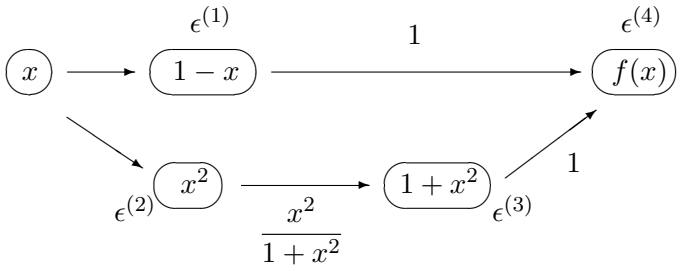
1.2.5 Quale dei due algoritmi ottenuti dall'identità

$$(1+x^2)(1-x) = 1 - x + x^2 - x^3$$

è più stabile? (20/1/2004)

Soluzione

Sia u la precisione di macchina. Dal grafo relativo al primo algoritmo, corrispondente alla funzione $f = (1+x^2)(1-x)$



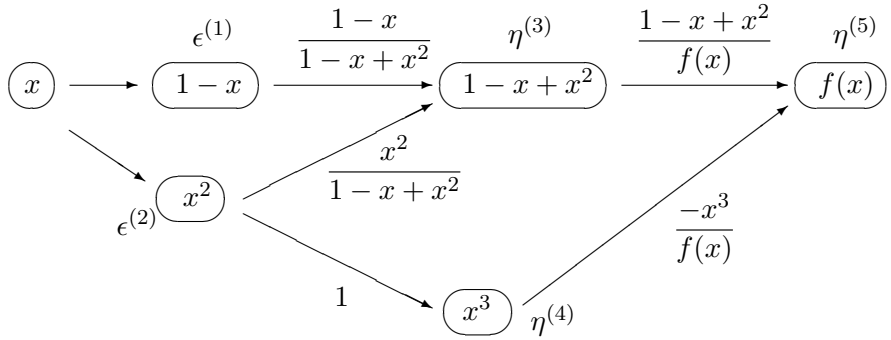
si ha che

$$\epsilon^{(1)}_{alg} = \epsilon^{(4)} + \epsilon^{(1)} + \epsilon^{(3)} + \frac{x^2}{1+x^2}\,\epsilon^{(2)},$$

e quindi

$$|\epsilon^{(1)}_{alg}| < u\Big(3 + \frac{x^2}{1+x^2}\Big) < 4u.$$

Dal grafo relativo al secondo algoritmo, corrispondente alla funzione $f = 1 - x + x^2 - x^3$



si ha che

$$\epsilon^{(2)}_{alg} = \eta^{(5)} + \frac{1-x+x^2}{f(x)}\Big(\eta^{(3)} + \frac{1-x}{1-x+x^2}\,\epsilon^{(1)} + \frac{x^2}{1-x+x^2}\,\epsilon^{(2)}\Big) - \frac{x^3}{f(x)}\,\Big(\eta^{(4)} + \epsilon^{(2)}\Big)$$

da cui

$$|\epsilon_{alg}^{(2)}| < u\Big(2 + \frac{|1 - x + x^2| + |x^3|}{|1 - x + x^2 - x^3|}\Big).$$

Quindi il primo algoritmo è stabile per ogni x, mentre il secondo non è stabile per x in un intorno di 1. Altrove anche il secondo algoritmo è stabile. Il primo appare comunque preferibile.

1.2.6 È data la funzione $g(x) = \mathrm{e}^{f(x)}$ per $x \in [0, 1]$.

a) Si dimostri che il coefficiente di amplificazione del calcolo di $g(x)$ è $c_g(x) = x\, f'(x)$.

b) Si dica se il calcolo di $g(x)$ è sempre ben condizionato nel caso in cui $f(x) = \sin x$.

c) Si dica se il calcolo di $g(x)$ è sempre ben condizionato nel caso in cui $f(x) = \sqrt{x}$.

d) Assumendo che il calcolo di $f(x)$ e della funzione esponenziale siano effettuabili con errore relativo limitabile dalla precisione di macchina u, si dimostri che per l'errore algoritmico di $g(x)$ risulta

$$|\epsilon_{alg}| < u(1 + |f(x)|). \qquad (5/2/2007)$$

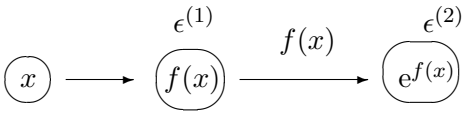
Soluzione

a) Risulta $\displaystyle c_g(x) = x\, \frac{f'(x)\, \mathrm{e}^{f(x)}}{\mathrm{e}^{f(x)}} = x\, f'(x)$.

b) In questo caso risulta $c_g(x) = x \cos x$. Per $x \in [0, 1]$ è $|c_g(x)| \leq 1$, quindi il calcolo di $g(x)$ è sempre ben condizionato.

c) In questo caso risulta $c_g(x) = \frac{1}{2}\sqrt{x}$. Per $x \in [0, 1]$ è $|c_g(x)| \leq \frac{1}{2}$, quindi il calcolo di $g(x)$ è sempre ben condizionato.

d) Dal grafo



risulta

$$\epsilon_{alg} < \epsilon_1\, f(x) + \epsilon_2,$$

quindi

$$|\epsilon_{alg}| < |\epsilon_1|\, |f(x)| + |\epsilon_2|,$$

da cui si ottiene la disuguaglianza richiesta, tenendo conto del fatto che $|\epsilon_1| < u$ e $|\epsilon_2| < u$.

1.2.7

a) Si dica per quali $x > 0$ è ben condizionato il problema del calcolo di

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

b) Si studi la stabilità dell'algoritmo nell'ipotesi che la radice quadrata di un numero di macchina venga calcolata da una funzione di libreria con un errore limitato in modulo dalla precisione di macchina.

c) Si dica con quale precisione va effettuato il calcolo affinchè l'errore algoritmico sia maggiorato in modulo da 10^{-6} per ogni $x \in [0, 1/4]$. (10/2/2004)

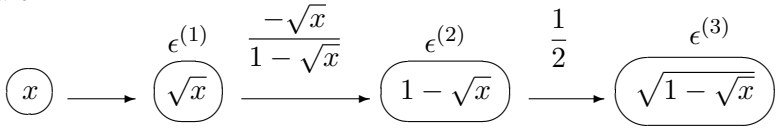
Soluzione

a) Deve essere $0 \le x \le 1$. L'errore inerente è

$$\epsilon_{in} = c_x\, \epsilon_x, \quad \text{dove} \quad c_x = \frac{x f'(x)}{f(x)} = -\frac{\sqrt{x}}{4(1 - \sqrt{x})}.$$

Quindi il problema risulta mal condizionato in un intorno sinistro di 1 e ben condizionato altrove.

b) Dal grafo



risulta

$$\epsilon_{alg} = \epsilon^{(3)} + \frac{1}{2}\,\epsilon^{(2)} - \frac{\sqrt{x}}{2(1 - \sqrt{x})}\,\epsilon^{(1)},$$

per cui

$$|\epsilon_{alg}| < u\Big(\frac{3}{2} + \frac{\sqrt{x}}{2(1 - \sqrt{x})}\Big),$$

dove u è la precisione di macchina. Quindi l'algoritmo risulta instabile in un intorno sinistro di 1 e stabile altrove.

c) La funzione $\dfrac{\sqrt{x}}{2(1 - \sqrt{x})}$ è crescente, quindi

$$\max_{x \in [0,1/4]} \frac{\sqrt{x}}{2(1 - \sqrt{x})} = \frac{\sqrt{1/4}}{2(1 - \sqrt{1/4})} = \frac{1}{2}.$$

Per $x \in [0, 1/4]$ risulta $|\epsilon_{alg}| < 2u$. Quindi l'errore algoritmico risulta maggiorato in modulo da 10^{-6} se $u = 10^{-6}/2$.

1.2.8 Si studi il condizionamento e la stabilità del calcolo di $f(x) = \sin(\alpha\, x)$ per $x \in \mathbf{R}$ ed α intero positivo. (17/1/2003)

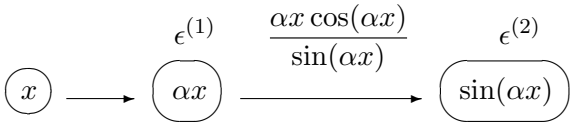
Soluzione

L'errore inerente è

$$\epsilon_{in} = c_x\, \epsilon_x, \quad \text{dove} \quad c_x = \frac{\alpha x\, \cos(\alpha x)}{\sin(\alpha x)}\,.$$

Il denominatore si annulla per $\alpha x = k\pi$ con k intero. Per $x \to 0$ abbiamo $\dfrac{\alpha x}{\sin(\alpha x)} \to 1$ e quindi $|\epsilon_{in}|$ è limitato ed il problema è ben condizionato. Il problema è invece mal condizionato per x negli intorni di $k\pi/\alpha$ con $k = 1,\ 2,\ \ldots$.

Dal grafo



risulta

$$\epsilon_{alg} = \epsilon^{(2)} + \frac{\alpha x\, \cos(\alpha x)}{\sin(\alpha x)}\, \epsilon^{(1)},$$

per cui

$$|\epsilon_{alg}| \leq u \left(1 + \left|\frac{\alpha x\, \cos(\alpha x)}{\sin(\alpha x)}\right|\right).$$

Quindi l'algoritmo risulta stabile dove il problema è ben condizionato.

1.2.9 È data la funzione

$$f(x) = \sqrt{\cos(2x)} = \sqrt{\cos^2 x - \sin^2 x}, \quad \text{per} \quad 0 \leq x \leq \pi/4.$$

Si studi il condizionamento del calcolo di $f(x)$ e si dica quale dei due algoritmi è più stabile. (18/6/2003)

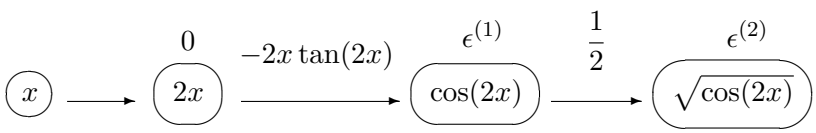
Soluzione

L'errore inerente è

$$\epsilon_{in} = c_x\, \epsilon_x, \quad \text{dove} \quad c_x = -\,x\, \tan(2x).$$

Quindi il problema del calcolo di $f(x)$ è malcondizionato nell'intorno sinistro di $\pi/4$.

Il grafo relativo al primo algoritmo, corrispondente a $f(x) = \sqrt{\cos(2x)}$, è



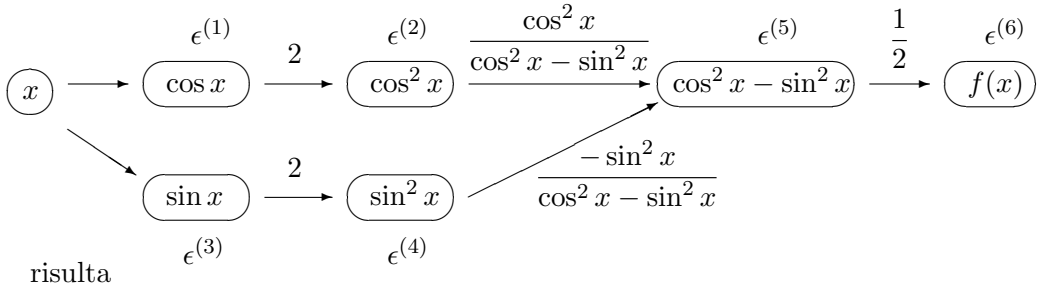
dove si è tenuto conto del fatto che l'errore locale della moltiplicazione di x per 2 è nullo. Quindi l'errore algoritmico risulta

$$\epsilon_{alg}^{(1)} = \epsilon^{(2)} + \frac{1}{2}\,\epsilon^{(1)},$$

da cui

$$|\epsilon_{alg}^{(1)}| < \frac{3}{2}\,u.$$

Il grafo relativo al secondo algoritmo, corrispondente a $f(x) = \sqrt{\cos^2 x - \sin^2 x}$ è



risulta

$$\epsilon_{alg}^{(2)} = \epsilon^{(6)} + \frac{1}{2}\left(\epsilon^{(5)} + \frac{\cos^2 x}{\cos^2 x - \sin^2 x}\left(\epsilon^{(2)} + 2\epsilon^{(1)}\right) - \frac{\sin^2 x}{\cos^2 x - \sin^2 x}\left(\epsilon^{(4)} + 2\epsilon^{(3)}\right)\right)$$

per cui

$$|\epsilon_{alg}^{(2)}| < u\left[1 + \frac{1}{2}\left(1 + \frac{3\sin^2 x + 3\cos^2 x}{|\cos^2 x - \sin^2 x|}\right)\right] = \frac{3}{2}\,u\,\left(1 + \frac{1}{|\cos 2x|}\right).$$

Quindi il primo algoritmo è sempre stabile, mentre il secondo algoritmo non è stabile nell'intorno sinistro di $\pi/4$. Il primo algoritmo è preferibile.

1.2.10 Si vuole calcolare il valore del polinomio

$$p(x) = x^n + 2x^{n-1} + 2^2x^{n-2} + \ldots + 2^{n-1}x + 2^n, \quad n > 1,$$

con il metodo di Ruffini-Horner per $0 < x < 1$.

a) Si studi l'errore algoritmico per il caso $n = 3$, in cui

$$p(x) = x(x(x + 2) + 4) + 8.$$

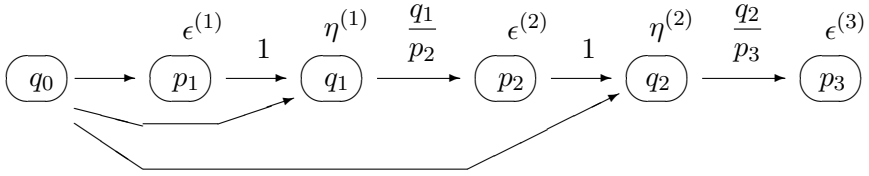
b) Si studi l'errore algoritmico per n generico. (24/7/2006)

Soluzione

a) Si pone

$$q_0 = x, \quad p_1 = q_0 + 2, \quad q_1 = x\,p_1, \quad p_2 = q_1 + 4, \quad q_2 = x\,p_2, \quad p_3 = q_2 + 8.$$

Quindi $p(x) = p_3$. Per l'errore algoritmico dal grafo



si ha

$$\epsilon_{alg} = \epsilon^{(3)} + \frac{q_2}{p_3}\Big(\eta^{(2)} + \epsilon^{(2)} + \frac{q_1}{p_2}\,(\eta^{(1)} + \epsilon^{(1)})\Big).$$

Maggiorando in modulo e tenendo conto che $x \in (0,1)$, si ha

$$|\epsilon_{alg}| < u\Big(1 + 2\,\frac{q_2}{p_3}\,(1 + \frac{q_1}{p_2})\Big).$$

Poichè le due frazioni q_1/p_2 e q_2/p_3 sono maggiorate dal valore che esse hanno in 1, si ha $q_1/p_2 < 3/7$ e $q_2/p_3 < 7/15$ e $|\epsilon_{alg}| < 7u/3$.

b) Per n generico si pone

$$q_0 = x, \quad p_i = q_{i-1} + 2^i, \quad q_i = x\,p_i, \quad i = 1, \ldots, n.$$

Il modulo dell'errore algoritmico risulta così maggiorato

$$|\epsilon_{alg}| < u\Big(1 + 2\,\frac{q_{n-1}}{p_n}\,\Big(1 + \frac{q_{n-2}}{p_{n-1}}\,\Big(1 + \frac{q_{n-3}}{p_{n-2}}\,\Big(1 + \ldots\,(1 + \frac{q_1}{p_2})\,\ldots\Big)\Big)\Big)\Big).$$

Tutte le frazioni sono maggiorate dal valore che esse hanno in 1. Poiché per $x = 1$ è $p_i = 2^{i+1} - 1$, si ha

$$\frac{q_{i-1}}{p_i} = 1 - \frac{2^i}{p_i} \le 1 - \frac{2^i}{2^{i+1} - 1} < \frac{1}{2} \quad \text{per ogni } i,$$

quindi

$$\begin{aligned} |\epsilon_{alg}| &< u\Big(1 + 2\,\frac{1}{2}\,\Big(1 + \frac{1}{2}\,\Big(1 + \frac{1}{2}\,\Big(1 + \ldots\,(1 + \frac{1}{2})\,\ldots\Big)\Big)\Big)\Big) \\ &= u\Big(1 + 2\,(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{2^{n-1}})\Big) < 3u. \end{aligned}$$

1.2.11 Si vuole approssimare la funzione $f(x) = 5e^x + 3e^{-x}$ con il polinomio $p(x) = 8 + 2x + 4x^2$ per $x \in [0,1]$. Si vuole valutare il polinomio operando in $\mathcal{F} = \mathcal{F}_{(2,20,128,127)}$ con arrotondamento.

a) Si dica qual è la precisione di macchina u di $\mathcal{F}$ e quanti elementi contiene l'insieme $\mathcal{G} = \mathcal{F} \cap [0,1]$.

b) Si individui una maggiorazione per l'errore analitico (si osservi che $p(x)$ è il polinomio di Taylor di secondo grado che approssima $f(x)$ nell'intorno di 0).

c) Si individui una maggiorazione per l'errore inerente assumendo $|\epsilon_x| < u$.

d) Si individui una maggiorazione per l'errore algoritmico assumendo che $p(x)$ sia calcolato con l'algoritmo suggerito dall'uguaglianza:

$$p(x) = 8 + x(2 + 4x).$$

e) Si individui una maggiorazione per l'errore totale.

f) Si suggerisca in che modo l'errore totale potrebbe venire ridotto. (7/11/2001)

Soluzione

a) È $u = 2^{-20}$. Si contano a parte 0 e 1. Restano da contare tutti i numeri del tipo $(0.1d_1 \dots d_{20})\, 2^p$ con $d_i \in \{0, 1\}$ e con p intero tale che $-128 \leq p \leq 0$. In totale $2 + 129 \cdot 2^{19}$ numeri.

b) Per definizione

$$\epsilon_{an} = \frac{f(x) - p(x)}{f(x)}.$$

Utilizzando la forma di Lagrange del resto otteniamo subito

$$\epsilon_{an} = \frac{x^3(5e^{\xi} - 3e^{-\xi})}{6\,(5e^x + 3e^{-x})}$$

con $\xi \in (0, x)$ e quindi

$$|\epsilon_{an}| \leq \frac{5e + 3}{6\,(5 + 1)} < \frac{18}{36} = \frac{1}{2}.$$

c) Il coefficiente di amplificazione risulta

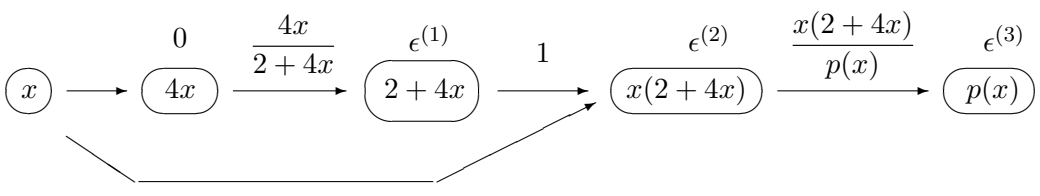
$$c_f = x\,\frac{5e^x - 3e^{-x}}{5e^x + 3e^{-x}},$$

quindi

$$|c_f| \leq \frac{5e + 3}{6} < 3,$$

e $|\epsilon_{in}| < 3u = 3 \cdot 2^{-20}$.

d) Dal grafo



per l'errore algoritmico otteniamo

$$\epsilon_{alg} = \epsilon^{(3)} + \frac{2x + 4x^2}{8 + 2x + 4x^2}\,(\epsilon^{(2)} + \epsilon^{(1)}).$$

Poichè $x \in [0, 1]$, risulta $\left|\dfrac{2x + 4x^2}{8 + 2x + 4x^2}\right| < 1$, quindi

$$|\epsilon_{alg}| < 3\,u = 3 \cdot 2^{-20}.$$

e) L'errore totale risulta

$$\epsilon_{tot} < \frac{1}{2} + 6 \cdot 2^{-20}.$$

f) Per ridurre l'errore totale occorre ridurre l'errore analitico. Per far questo si può per esempio utilizzare un polinomio di Taylor di grado maggiore.

1.2.12 Si vuole approssimare $f(x) = \sin x - \cos x$ con il polinomio di grado 2 ottenuto dalla formula di Taylor

$$p(x) = -1 + x + \frac{x^2}{2}$$

nell'intervallo $\mathcal{I} = [-\pi/8, \pi/8]$ (si osservi che $|\sin x - \cos x| \geq 1/2$ nell'intervallo $\mathcal{I}$).

a) Si dica se il problema del calcolo di $f(x)$ è ben condizionato per ogni $x \in \mathcal{I}$.

b) Si studi l'errore algoritmico commesso calcolando la funzione $p(x)$ in aritmetica finita.

c) Si studi l'errore analitico relativo commesso approssimando $f(x)$ con $p(x)$ per $x \in \mathcal{I}$. (6/11/2002)

Soluzione

a) L'errore inerente è

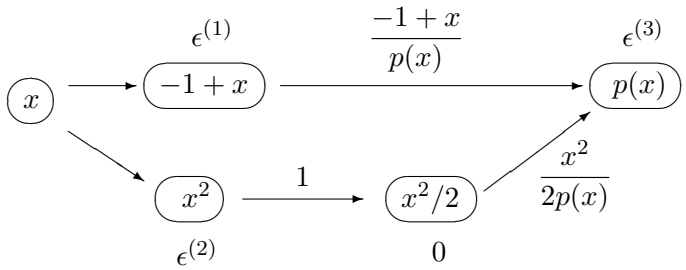
$$\epsilon = c_x\,\epsilon_x, \quad \text{dove} \quad c_x = x\,\frac{\cos x + \sin x}{\sin x - \cos x}.$$

Poiché $|\cos x + \sin x| \leq \cos(\pi/8) + \sin(\pi/8) \leq 3/2$, e $|\sin x - \cos x| \geq 1/2$ otteniamo

$$|c_x| \leq 3\,|x| \leq \frac{3\pi}{8}.$$

Quindi il problema è ben condizionato in $\mathcal{I}$.

b) Dal grafo



l'errore algoritmico risulta

$$\epsilon_{alg} = \epsilon^{(3)} + \frac{-1+x}{p(x)}\,\epsilon^{(1)} + \frac{x^2}{2\,p(x)}\,\epsilon^{(2)},$$

per cui

$$|\epsilon_{alg}| < u\left(1 + \frac{|x-1|}{|p(x)|} + \frac{x^2}{2\,|p(x)|}\right).$$

Il polinomio $p(x)$ al denominatore si annulla nei due punti $-1 \pm \sqrt{3}$ che non appartengono a $\mathcal{I}$. In $\mathcal{I}$ è $|p(x)| > 1/2$, $x^2 < 0.2$, $|x-1| < 1.4$, perciò

$$|\epsilon_{alg}| < 4u.$$

c) L'errore analitico può essere studiato ricorrendo al resto della formula di Taylor

$$f(x) = p(x) - \frac{x^3}{6}(\sin\xi + \cos\xi),$$

con ξ compreso tra 0 e x. Abbiamo allora

$$|\epsilon_{an}| = \frac{|x^3(\sin\xi + \cos\xi)|}{6\,|f(x)|}.$$

Poiché in $\mathcal{I}$ è $|f(x)| > 1/2$, $|\sin\xi| \le \sin(\pi/8) < 0.4$ e $\cos\xi \le 1$, risulta

$$|\epsilon_{an}| \le 0.03.$$

1.3 Esercizi proposti

1.3.1 Sia $\mathcal{F} = \mathcal{F}_{(\beta,t,m,M)}$ con $m = M$.

a) Si dica quali sono il minimo numero positivo ω ed il massimo numero positivo Ω di $\mathcal{F}$.

b) Si dica se i numeri $b = 1/\omega$ e $B = 1/\Omega$ appartengono a $\mathcal{F}$.

c) Si esamini in particolare il caso $\beta = 2$, $t = 8$ ed $m = M = 6$. (25/5/2004)

1.3.2 Si consideri l'insieme dei numeri di macchina $\mathcal{F} = \mathcal{F}_{(10,3,7,6)}$ in cui si assume di operare con troncamento.

a) Assegnato $x = 2006$, se ne calcoli la rappresentazione $\widetilde{x}$ in $\mathcal{F}$.

b) Si determinino tutti i numeri reali tali che la loro rappresentazione in $\mathcal{F}$ coincida con $\widetilde{x}$.

c) Si calcoli la precisione di macchina u e si determinino due numeri $z \in \mathcal{F}$ tali che
$$\frac{|z - x|}{|x|} \leq u. \qquad (7/11/2006)$$

1.3.3 Data l'equazione $x^2 - 3 = 0$, se ne approssima la soluzione positiva con il metodo delle tangenti
$$x_{i+1} = \frac{x_i^2 + 3}{2x_i}, \qquad \text{con} \quad x_0 = 4.$$
Si effettuino 4 passi del metodo operando in aritmetica finita in $\mathcal{F}_{(10,2,m,M)}$ con arrotondamentoe si confrontino i valori ottenuti con quelli che avremmo ottenuto operando in aritmetica esatta. (6/11/2002)

1.3.4 Si consideri l'insieme $\mathcal{F} = \mathcal{F}_{(2,2,1,2)}$ in cui si suppone di operare con arrotondamento.

a) Si calcoli la precisione di macchina u.

b) Si elenchino tutti i numeri positivi contenuti in $\mathcal{F}$.

c) Si calcoli la funzione $\widetilde{f}(x) = x \otimes x$ su ciascuno degli elementi di $\mathcal{F}$ individuati nel punto precedente, segnalando le situazioni di underflow ed overflow e calcolando, negli altri casi l'errore relativo $\dfrac{f(x) - \widetilde{f}(x)}{f(x)}$, dove $f(x) = x^2$. (13/9/2007)

1.3.5 È noto che nell'algebra dei numeri reali vale la proprietà di semplificazione, cioè $x(y/x) = y$. Si verifichi che questa proprietà non vale nell'algebra dei numeri di macchina, constatando che per i numeri $x = 1/9$ e $y = 1/5$ è
$$\widetilde{x} \otimes (\widetilde{y} \oslash \widetilde{x}) \neq \widetilde{y},$$
quando si operi in $\mathcal{F}_{(2,3,3,3)}$ con troncamento. (13/1/2005)

1.3.6 Si dica quale è il numero $y \in \mathcal{F} = \mathcal{F}_{(2,4,m,M)}$, tale che
$$|y - \sqrt{5}| = \min_{x \in \mathcal{F}} |x - \sqrt{5}|,$$
(è $\sqrt{5} \sim 2.23607$). Si verifichi che l'errore relativo di y è maggiorato in modulo dalla precisione di macchina. (9/7/2003)

1.3.7 Dati $x = 1/3$, $y = 1/5$, siano $\widetilde{x}$ e $\widetilde{y}$ i corrispondenti valori arrotondati in $\mathcal{F} = \mathcal{F}_{(2,4,m,M)}$.

a) Si calcoli $\widetilde{x} \oslash \widetilde{y}$, dove $\oslash$ rappresenta la divisione di macchina di $\mathcal{F}$ con arrotondamento del risultato, applicando il seguente algoritmo: si riconvertono $\widetilde{x}$ e $\widetilde{y}$ in frazioni decimali e si converte in $\mathcal{F}$ il risultato della divisione esatta.

b) Si confronti l'errore relativo del risultato con la sua limitazione teorica. (12/6/2006)

1.3.8 Si consideri l'insieme $\mathcal{F} = \mathcal{F}_{(2,3,1,2)}$.

a) Si dica qual è il più piccolo elemento positivo ω di $\mathcal{F}$ e qual è la distanza δ_1 tra ω e l'elemento di $\mathcal{F}$ immediatamente successivo.

b) Si dica qual è il più grande elemento positivo Ω di $\mathcal{F}$ e qual è la distanza δ_2 tra Ω e l'elemento di $\mathcal{F}$ immediatamente precedente.

c) Si dica qual è il numero di elementi positivi di $\mathcal{F}$ e quale sarebbe la distanza δ_3 fra gli elementi positivi di $\mathcal{F}$ se fossero equidistanti tra ω e Ω.

d) Quale caratteristica di $\mathcal{F}$ rendeva prevedibile che $\delta_1 < \delta_3 < \delta_2$? (15/1/2007)

1.3.9 Si consideri l'insieme $\mathcal{F}_{(2,t,m,M)}$ e si indichi rispettivamente con Ω, ω e u il massimo ed il minimo elemento positivo dell'insieme e la precisione di macchina nel caso in cui si operi con arrotondamento.

a) Si dica per quali valori di t si ha $u \leq 10^{-6}$.

b) Si dica per quali valori di M risulta $\Omega \geq 10^9$ (suggerimento: avendosi $t \geq 1$ risulta $1 - 2^{-t} \geq 1/2$).

c) Si dica per quali valori di m risulta $\omega \leq 10^{-9}$.

d) Scegliendo i più piccoli valori di t, M ed m compatibili con le tre richieste precedenti, si calcoli la cardinalità di $\mathcal{F}$. Quanti bit risultano necessari per rappresentare tutti gli elementi di $\mathcal{F}$? (7/6/2007)

1.3.10 Dati due numeri positivi a e b tali che $a^2 > b$, siano α_1 e α_2 le ascisse dei punti di intersezione della parabola $y = x^2 - 2ax + b$ con l'asse delle x. Si studi il condizionamento del calcolo di α_1 e α_2. (3/9/2008)

1.3.11 Sia $M = \begin{bmatrix} x & y \\ y & x \end{bmatrix}$, con $x, y \in \mathbf{R}$.

a) Si studi il condizionamento della funzione $f(x, y) = \det M$.

b) Si approssimi il valore del determinante quando $x = \pi, y = e$ operando in $\mathcal{F}_{(10,2,m,M)}$, assumendo $\widetilde{x} = 3.1$ e $\widetilde{y} = 2.7$, e si dia una maggiorazione dell'errore commesso. (4/6/2003)

1.3.12 Assegnata la matrice $A = \begin{bmatrix} a & b \\ c & 1 \end{bmatrix}$, si studi l'errore algoritmico del calcolo del suo determinante

a) quando il determinante viene calcolato con la formula di Laplace,

b) quando il determinante viene calcolato dopo aver posto A in forma triangolare mediante il metodo di Gauss (si assuma $a \neq 0$).

Si indichi con u la precisione di macchina dell'aritmetica utilizzata. (18/9/2002)

1.3.13 È data la funzione $d(x) = \det \begin{bmatrix} 1 & x \\ x & x \end{bmatrix}, \quad x \in \boldsymbol{R}$.

a) Si studi il condizionamento del calcolo di $d(x)$.

b) Si proponga un algoritmo che non presenti situazioni di instabilità per il calcolo di $d(x)$ e si dia una limitazione per l'errore algoritmico commesso. (12/9/2005)

1.3.14 Si consideri la funzione $f(x) = \dfrac{1}{3x+1} - \dfrac{1}{3x+2}$ per $x > 0$.

a) Si studi il condizionamento di $f(x)$.

b) Si verifichi che l'algoritmo è instabile per valori grandi di x.

c) Si trovi un altro algoritmo che calcoli $f(x)$ e che risulti stabile per ogni $x > 0$. (4/11/2003)

1.3.15 Si consideri in $\mathcal{F}_{(2,3,m,M)}$ la rappresentazione $\widetilde{x}$ del numero $x = 1/3$ e si calcoli $s = \displaystyle\sum_{i=1}^{6} \widetilde{x}$ secondo i due diversi algoritmi:

a) $z = \widetilde{x} + \widetilde{x} + \widetilde{x}, \qquad s = z + z,$

b) $t_1 = \widetilde{x}, \qquad t_i = t_{i-1} + \widetilde{x}$, per $i = 1, \ldots, 6, \qquad s = t_6,$

operando con troncamento dei risultati intermedi. Si confrontino i risultati ottenuti con il valore esatto e gli errori effettivi con le maggiorazioni degli errori ottenuti con i grafi. (16/9/2003)

1.3.16 Per ognuna delle seguenti funzioni si studi il condizionamento del calcolo della funzione e la stabilità degli algoritmi corrispondenti alle diverse espressioni

equivalenti proposte. Nel caso di funzioni non razionali, si supponga di usare funzioni di libreria con errori limitati in modulo dalla precisione di macchina.

a) $f(x) = (x-1)(x-1) = (x-2)x + 1$, (17/1/2008)

b) $f(x) = (x^2+2)(x+1) = ((x+1)x+2)x + 2$, (9/6/2008)

c) $f(x) = (x^2)^2 - 1 = (x^2+1)(x^2-1)$, (15/6/2004)

d) $f(x) = \dfrac{x-1}{x^2-1} = \dfrac{1}{x+1}$, per $x \neq \pm 1$, (9/6/2005)

e) $f(x) = x - 1 + \dfrac{1}{x} = \dfrac{x^2-x+1}{x}$, per $x \neq 0$, (3/7/2006)

f) $f(x) = \sqrt{x^3} = x\sqrt{x}$, per $x \geq 0$, (12/9/2006)

g) $f(x) = \sqrt{x(1-x)} = \sqrt{x - x^2}$, per $0 < x < 1$, (8/2/2006)

h) $f(x) = \sqrt{\dfrac{1-\sqrt{1-x^2}}{2}}$, per $0 < x < 1$, (7/2/2005)

i) $f(x) = \dfrac{2-\cos x}{2x}$, per $x \neq 0$, (1/7/2005)

j) $f(x) = 1 + \dfrac{1}{5+\sin x}$, per $x \in [0, 2\pi]$, (7/11/2006)

k) $f(x) = \tan\dfrac{x}{2} = \dfrac{\sin x}{1+\cos x}$, per $x \in (0, \pi)$, (17/11/2004)

l) $f(x) = \dfrac{\log(1+x^2)}{x}$, per $x \neq 0$, (18/1/2006)

m) $f(x) = \mathrm{e}^{\sin x} - \dfrac{1}{\mathrm{e}}$, per $x \in [0, \pi]$, (25/6/2008)

n) $f(x,y) = \dfrac{x}{x+y} = \dfrac{z}{1+z}$, dove $z = \dfrac{x}{y}$,
per $y \neq 0$ e $x + y \neq 0$, (19/7/2007)

o) $f(x,y) = \log x^2 - \log y^2 = 2(\log x - \log y) = \log\left(x^2/y^2\right)$,
per $x, y > 0$, (15/9/2004).

1.3.17 Per $x > 0$, si confrontino le stabilità del calcolo delle espressioni

a) $\sqrt{x^2}$ e $(\sqrt{x})^2$ *b)* $\sqrt{\sqrt{(x^2)^2}}$ e $((\sqrt{\sqrt{x}})^2)^2$

c) $\sqrt{\sqrt{\sqrt{((x^2)^2)^2}}}$ e $(((\sqrt{\sqrt{\sqrt{x}}})^2)^2)^2$.

Nel caso $x > 1$, cosa si può prevedere accada aumentando gli elevamenti a potenza e le estrazioni di radice nel procedimento più stabile? (28/6/2007)

1.3.18 Per calcolare la funzione $f(x) = x - \sin x$ in un punto $x \in (0, 1/10)$ si usa l'approssimazione

$$g(x) = \frac{x^3}{3!} - \frac{x^5}{5!}\,.$$

a) Si studi il condizionamento del calcolo di $f(x)$.

b) Si dia una maggiorazione del modulo dell'errore analitico relativo. (7/2/2008)

1.3.19 Per calcolare la funzione $f(x) = \log x$ quando $x \geq 1/2$ si può usare l'approssimazione

$$g(x) = \frac{x-1}{x}\,\left(1 + \frac{x-1}{2x}\right).$$

a) Si studi il condizionamento del calcolo di $f(x)$ per $1/2 \leq x \leq 3/2$.

b) Si studi la stabilità dell'algoritmo.

c) Si studi l'errore analitico relativo (si tenga conto del fatto che $g(x)$ è ottenuta dai primi due termini della formula di Taylor di $\log(1+y)$, con la sostituzione $y = (x-1)/x$). (21/7/2005)

1.3.20 Per $x > 0$ si vuole approssimare la funzione $f(x) = 3x + \dfrac{2}{x+1}$ con la funzione $\widetilde{f}(x) = 3x$. Si suppone che la precisione di macchina sia $u = 10^{-4}$.

a) Fissato $\gamma > 0$, si dimostri che se $x \geq \gamma$ per l'errore analitico relativo risulta $|\varepsilon_{an}(x)| \leq \dfrac{2}{3\gamma^2}$.

b) Si dimostri che per $x > 0$, assumendo che l'errore sul dato x in valore assoluto sia inferiore ad u, per l'errore inerente di $f(x)$ risulta $|\varepsilon_{in}(x)| \leq u$.

c) Si dia una maggiorazione per l'errore algoritmico del calcolo di $\widetilde{f}(x)$.

d) Si determini γ affinchè l'errore totale risulti in valore assoluto inferiore a $7u$.

e) Se in $\mathcal{F}_{(2,t,m,M)}$ si opera con arrotondamento, qual è il valore minimo di t affinchè la precisione di macchina u sia inferiore a quella assegnata. (8/11/2007)

1.3.21 Per $x \in [0, 1)$ si vuole approssimare la funzione

$$f(x) = \frac{1}{x^5 - 1} \quad \text{con la funzione} \quad g(x) = \frac{1}{x^4 - 1},$$

supponendo di effettuare i calcoli nel modo seguente: $x^4 = (x^2)^2$ e $x^5 = (x^2)^2\, x$. Ignorando l'errore inerente, si confrontino l'errore algoritmico $\epsilon_{alg}(f)$ di $f(x)$ con l'errore dell'approssimazione, dato dalla somma dell'errore algoritmico $\epsilon_{alg}(g)$ di $g(x)$ e dell'errore analitico relativo $\epsilon_{an} = (f(x) - g(x))/f(x)$. (11/7/2008)