ARITMETICA, ERRORI

Letizia SCUDERI

Dipartimento di Scienze Matematiche, Politecnico di Torino letizia.scuderi@polito.it

A.A. 2022/2023

Nell'insegnamento di **Metodi Numerici**, verranno proposti e analizzati metodi che consentono di ottenere una **soluzione numerica** di alcuni problemi matematici, per i quali metodi di risoluzione analitica non esistono oppure sono eccessivamente onerosi.

Per ottenere una soluzione numerica si utilizzano **algoritmi**, eseguibili da un calcolatore e dedotti dai metodi stessi.

Per gli algoritmi, che implementano i suddetti metodi, in questo insegnamento verrà usato il linguaggio di programmazione $\underline{Matlab}^{\ @}.$

Rappresentazione dei numeri

Di seguito verranno richiamati i concetti fondamentali riguardanti l'aritmetica del calcolatore.

Definizione

Si definisce **rappresentazione floating-point** di un numero reale *a* la seguente espressione

$$a = (-1)^s pN^q$$
, $s \in \{0,1\}$, $p \ge 0$ reale, q intero,

ove *N* rappresenta la **base** del sistema di numerazione.

In questo insegnamento verrà utilizzato il sistema di numerazione decimale, ovvero il sistema con ${\it N}=10$.

Definizione

La rappresentazione floating-point $a = (-1)^s pN^q$ del numero reale a si dice **normalizzata** se p soddisfa la condizione

$$N^{-1} \le p < 1$$

In tal caso, p e q sono univocamente determinati e

- il numero reale non negativo p si definisce mantissa di a;
- l'intero q si definisce esponente oppure caratteristica di a.

Per la memorizzazione di $a=(-1)^spN^q$, con $N^{-1} \leq p < 1$, su di un calcolatore è sufficiente memorizzare s, p e q della sua rappresentazione floating-point normalizzata.

Poiché un calcolatore riserva per la memorizzazione di tali quantità uno **spazio finito di memoria**, esso può memorizzare solo mantisse con un numero finito di cifre, per esempio massimo t cifre, ed esponenti appartenenti a un certo intervallo, per esempio $L \le q \le U$ con L < 0 e U > 0 interi.

Definizione

Si definiscono **numeri di macchina** i numeri con mantissa ed esponente esattamente rappresentabili negli spazi a loro riservati dal calcolatore.

L'insieme dei numeri di macchina è costituito da un numero finito di elementi ed è così definito:

$$\mathcal{F} = \{0\} \cup \{(-1)^s 0.a_1 a_2 ... a_t \cdot N^q, 0 \le a_i < N, a_1 \ne 0, L \le q \le U\}$$

dove

- N è la base del sistema di numerazione;
- s vale 0 (segno positivo) oppure 1 (segno negativo);
- $a_1, a_2, ..., a_t$ sono interi minori o uguali a N-1 e rappresentano le cifre della mantissa¹;
- t è il massimo numero di cifre della mantissa rappresentabili;
- q è l'esponente;
- L < 0 e U > 0 sono interi.

¹Da qui in avanti, con il termine *cifre della mantissa* $p = 0.a_1a_2...a_t$ si intenderanno le cifre $a_1, a_2, ..., a_t$ che seguono la virgola.

Il più piccolo numero di macchina positivo (diverso dallo zero) appartenente a ${\mathcal F}$ è

$$m = 0. \underbrace{10...0}_{t \text{ cifre}} \cdot N^L$$

Il più grande numero di macchina positivo appartenente a ${\mathcal F}$ è

$$M = 0. \underbrace{N-1 N-1 ... N-1}_{t \text{ cifre}} \cdot N^U$$

Definizione

Si definisce **regione di underflow** l'insieme dei numeri reali diversi da zero e appartenenti a (-m, m).

Si definisce **regione di overflow** l'insieme dei numeri reali appartenenti a $(-\infty, -M) \cup (M, +\infty)$.

In un calcolatore come vengono "trattati" i numeri reali che non appartengono all'insieme $\mathcal F$ dei numeri di macchina, ovvero che non sono numeri di macchina?

- Quando un'operazione genera un numero appartenente alla regione di underflow, questo viene approssimato con lo zero, il processo di calcolo non si arresta e l'utente riceve una segnalazione dell'avvenuto fenomeno.
- Quando, invece, il risultato di un'operazione appartiene alla regione di overflow, esso viene approssimato con il simbolo *Inf*, il processo di calcolo si arresta e l'utente riceve una segnalazione dell'avvenuto fenomeno. Pertanto, in quest'ultimo caso, è necessario evitare la regione di overflow. Ciò è generalmente possibile manipolando opportunamente l'espressione matematica che genera il fenomeno di overflow.
- Quando il dato di un problema oppure il risultato di un'operazione è un numero la cui mantissa ha più di t cifre, allora per la sua memorizzazione il calcolatore ricorre a una tecnica di arrotondamento.

La tecnica di arrotondamento generalmente utilizzata è la seguente.

Definizione

Tecnica di arrotondamento "rounding to even": la mantissa p viene approssimata con la mantissa di macchina \bar{p} più vicina e se p è equidistante da due mantisse di macchina consecutive allora p viene approssimata con quella delle due che ha l'ultima cifra pari.

Esempi

Siano N = 10 e t = 3.

- Se $p = 0.158\underline{1}4$, allora $\bar{p} = 0.158$.
- Se p = 0.1585432, allora $\bar{p} = 0.159$.
- Se $p = 0.158\underline{8}12$, allora $\bar{p} = 0.159$.
- Se p = 0.1585, allora $\bar{p} = 0.158$.
- Se p = 0.1595, allora $\bar{p} = 0.160$.

In seguito con la notazione \bar{a} verrà indicato il numero di macchina corrispondente ad a. Se a è un numero di macchina, allora $a=\bar{a}$; altrimenti $a\approx\bar{a}$. In questo ultimo caso, vale la seguente definizione.

Definizione

Si definisce **errore di arrotondamento** l'errore che si commette quando si sostituisce il numero reale $a \neq 0$ con il corrispondente numero di macchina \bar{a} . Per esso vale la seguente stima

$$\frac{|a-\bar{a}|}{|a|} \leq \frac{1}{2} N^{1-t}$$

Si ricordano inoltre le seguenti definizioni.

Definizione

Si definisce epsilon di macchina la quantità

$$eps = N^{1-t}$$

Si definisce precisione di macchina la quantità

$$\varepsilon_m = \frac{1}{2} N^{1-t}$$

La precisione di macchina è una costante caratteristica di ogni aritmetica floating-point; essa rappresenta il massimo errore relativo che si commette quando si approssima il numero reale a con il corrispondente numero di macchina \bar{a} . Ponendo $\varepsilon=(\bar{a}-a)/a$, ove \bar{a} è il numero di macchina corrispondente al numero reale a, si deduce la seguente uguaglianza

$$\bar{a} = a(1+\varepsilon), \quad |\varepsilon| \le \varepsilon_m$$

che esprime una relazione tra \bar{a} e a.

Operazioni di macchina

Il risultato di un'operazione aritmetica tra due numeri di macchina generalmente non è un numero di macchina.

In un calcolatore **non è possibile eseguire esattamente le operazioni** aritmetiche +, -, \times e /, ma solo le cosiddette *operazioni di macchina*, che vengono rappresentate con i simboli \oplus , \ominus , \otimes e \oslash .

Definizione

L'operazione di macchina \odot associa a due numeri di macchina un terzo numero di macchina, ottenuto arrotondando l'esatto risultato dell'operazione in questione.

Siano \bar{a}_1 e \bar{a}_2 due numeri di macchina e sia \odot l'operazione di macchina corrispondente all'operazione \cdot in aritmetica esatta. Si ha allora

$$ar{a}_1\odotar{a}_2=\overline{ar{a}_1\cdotar{a}_2}=(ar{a}_1\cdotar{a}_2)(1+arepsilon_\odot)$$

con $|\varepsilon_{\odot}| \leq \varepsilon_m$.

Per le operazioni di macchina rimane valida la proprietà commutativa, ma non valgono in generale le proprietà associativa e distributiva. Per esempio, in generale si ha

$$(\bar{a}_1 \oplus \bar{a}_2) \oplus \bar{a}_3 \neq \bar{a}_1 \oplus (\bar{a}_2 \oplus \bar{a}_3)$$

Un'ulteriore relazione anomala è

$$\bar{a}_1 \oplus \bar{a}_2 = \bar{a}_1$$

quando $|\bar{a}_2| \ll |\bar{a}_1|$, in particolare quando $|\bar{a}_2| < |\bar{a}_1| eps$. Infatti, in tal caso, il valore di $|\bar{a}_2|$ è "troppo piccolo" rispetto a $|\bar{a}_1|$ e non fornisce alcun contributo nella somma.

Osservazione

MATLAB implementa le specifiche del sistema standard floating point IEEE 754 del 1985 (divenuto lo standard internazionale IEC 559 del 1989), che in doppia precisione prevede:

la base N=2, la tecnica di arrotondamento precedentemente descritta e 64 bit per la rappresentazione di un numero macchina, di cui

- 1 bit per rappresentare il segno della mantissa;
- 52 bit per rappresentare le cifre della mantissa p (in realtà i bit per la mantissa sono 53, in quanto non si rappresenta il primo bit, che è certamente 1);
- 11 bit per rappresentare l'esponente q.

... continua osservazione

In Matlab sono predefiniti i seguenti valori:

- realmin fornisce il valore del più piccolo numero di macchina positivo e non nullo $m \approx 2.2 \cdot 10^{-308}$;
- realmax fornisce il valore del più grande numero di macchina $M \approx 1.8 \cdot 10^{308}$;
- eps fornisce il valore dell'epsilon di macchina eps $\approx 2.2 \cdot 10^{-16}$ (la precisione di macchina è dunque $\varepsilon_m \approx 1.1 \cdot 10^{-16}$).

Si osservi che 53 cifre per la mantissa in base 2 corrispondono a circa 16 cifre per la mantissa in base 10.

Ricordiamo inoltre la seguente definizione.

Definizione

Due espressioni/quantità e_1 ed e_2 si definiscono **equivalenti nell'aritmetica del calcolatore** quando, valutate nel calcolatore stesso, forniscono risultati che differiscono per una tolleranza relativa dell'ordine della precisione di macchina ε_m o minore, cioè quando

$$\frac{|\bar{e}_1 - \bar{e}_2|}{|\bar{e}_1|} \quad \text{oppure} \quad \frac{|\bar{e}_1 - \bar{e}_2|}{|\bar{e}_2|}$$

è dell'ordine della precisione di macchina ε_m o minore.

Ne consegue che ε_m rappresenta la massima precisione (relativa) di calcolo raggiungibile: non ha senso cercare di determinare approssimazioni con precisione relativa inferiore a ε_m .

Si supponga che l'arrotondamento dei numeri reali venga effettuato con una tecnica di arrotondamento (per esempio, quella del rounding to even). Allora vale la seguente definizione.

Definizione

• Si dice che un'approssimazione $\tilde{a} = (-1)^s \tilde{p} N^q$ del numero reale $a = (-1)^s p N^q$ ha **k decimali corretti nella base N** se

$$\frac{1}{2}N^{-(k+1)} < |a - \tilde{a}| \le \frac{1}{2}N^{-k};$$

Si dice che ã ha k cifre significative nella base N se

$$\frac{1}{2}N^{-(k+1)} < |p - \tilde{p}| \le \frac{1}{2}N^{-k}$$

cioè se k è il numero di decimali corretti presenti nella mantissa \tilde{p} .

Esempi

• Siano a = 0.589231 e $\tilde{a} = 0.58941326$. Poiché

$$\frac{1}{2} 10^{-(3+1)} < |a - \tilde{a}| = |p - \tilde{p}| = 0.18226 \, \cdot 10^{-3} < \frac{1}{2} 10^{-3}$$

deduciamo che $\tilde{a}=0.\underline{589}41326$ ha 3 decimali corretti e 3 cifre significative.

• Siano $a = 0.589231 \cdot 10^{-3}$ e $\tilde{a} = 0.58941326 \cdot 10^{-3}$. Da

$$|a - \tilde{a}| = 0.18226 \cdot 10^{-6}$$

deduciamo che $\tilde{a}=0.\underline{000589}41326$ ha 6 decimali corretti. Inoltre, essendo

$$|p - \tilde{p}| = 0.18226 \cdot 10^{-3},$$

deduciamo che $\tilde{a} = 0.58941326 \cdot 10^{-3}$ ha 3 cifre significative.

... segue esempi

• Siano $a = 0.589231 \cdot 10^2$ e $\tilde{a} = 0.58941326 \cdot 10^2$.

$$|a - \tilde{a}| = 0.18226 \cdot 10^{-1}$$

deduciamo che $\tilde{a} = 58.941326$ ha 1 decimale corretto.

Da

$$|p - \tilde{p}| = 0.18226 \cdot 10^{-3}$$

segue che $\tilde{a} = 0.58941326 \cdot 10^{-3}$ ha 3 cifre significative.

• Siano a = 0.2 e $\tilde{a} = 0.199999$.

$$|a - \tilde{a}| = |p - \tilde{p}| = 0.1 \cdot 10^{-5}$$

segue che $\tilde{a} = 0.199999$ ha 5 decimali corretti e 5 cifre significative.

Cancellazione numerica

La cancellazione numerica rappresenta una delle conseguenze più gravi della rappresentazione con precisione finita dei numeri reali.

In generale, la cancellazione numerica può essere così definita.

Definizione

Siano

$$\bar{a}_1 = (-1)^{s_1} \bar{p}_1 N^{q_1}, \quad \bar{a}_2 = (-1)^{s_2} \bar{p}_2 N^{q_2}$$

le rappresentazioni di macchina associate rispettivamente ai numeri reali

$$a_1 = (-1)^{s_1} p_1 N^{q_1}, \quad a_2 = (-1)^{s_2} p_2 N^{q_2}$$

La cancellazione numerica consiste in una perdita di cifre della mantissa e si verifica quando si esegue l'operazione di sottrazione fra due rappresentazioni di macchina \bar{a}_1 e \bar{a}_2 dello stesso segno $(s_1=s_2)$, circa uguali $(q_1=q_2$ e $\bar{p}_1\approx\bar{p}_2)$ e almeno una delle quali sia affetta dall'errore di arrotondamento $(\bar{p}_1\neq p_1$ e/o $\bar{p}_2\neq p_2)$.

Esempio

Siano N=10 e t=5. Si ha allora $\varepsilon_m=0.5\cdot 10^{-4}$. Consideriamo i numeri

$$a_1 = 0.157824831$$
 e $a_2 = 0.157348212$

Osserviamo che le prime tre cifre delle mantisse di a_1 e a_2 coincidono; inoltre, $\bar{a}_1=0.15782\neq a_1$ e $\bar{a}_2=0.15735\neq a_2$. In tal caso, risulta

	in aritmetica esatta		in aritmetica finita
1	= 0.157824831	$ar{a}_1$	= 0.15782
2	= 0.157348212	\bar{a}_2	= 0.15735

$$a_1 - a_2 = 0.000476619$$
 $\bar{a}_1 \ominus \bar{a}_2 = 0.00047$
= $0.476619 \cdot 10^{-3}$ = $0.47000 \cdot 10^{-3}$

Osserviamo che la mantissa di $\bar{a}_1 - \bar{a}_2$ ha solo le prime 2 cifre decimali in comune con la mantissa di $a_1 - a_2$. Si è verificata una perdita di tre cifre!

*a*₁

... segue esempio

Consideriamo ora i numeri

$$a_1 = 0.15782$$
 e $a_2 = 0.15735$

Osserviamo che le prime tre cifre delle mantisse di a_1 e a_2 coincidono; inoltre, $a_1 = \bar{a}_1$ e $a_2 = \bar{a}_2$. Eseguiamo l'operazione di sottrazione in aritmetica esatta e nell'aritmetica fissata:

in aritmetica esatta	in aritmetica finita
$a_1 = 0.15782$	$\bar{a}_1 = 0.15782$
$a_2 = 0.15735$	$\bar{a}_2 = 0.15735$

$$a_1 - a_2 = 0.00047$$
 $\bar{a}_1 \ominus \bar{a}_2 = 0.00047$
= 0.47 \cdot 10^{-3} = 0.47 \cdot 10^{-3}

Le due operazioni hanno fornito lo stesso risultato: non è avvenuta cancellazione numerica.

In questo caso gli operandi \bar{a}_1 e \bar{a}_2 non sono affetti dall'errore di arrotondamento e il risultato dell'operazione di sottrazione non presenta alcuna perdita di precisione; nel primo caso, invece, entrambi gli operandi \bar{a}_1 e \bar{a}_2 sono affetti dall'errore di arrotondamento e tali errori vengono amplificati dall'operazione di sottrazione.

Infatti, si ha

$$\frac{|a_1 - \bar{a}_1|}{|a_1|} \approx 0.3 \cdot 10^{-4} < \varepsilon_m$$
 $\frac{|a_2 - \bar{a}_2|}{|a_2|} \approx 0.1 \cdot 10^{-4} < \varepsilon_m$

mentre

$$\frac{|(a_1 - a_2) - (\bar{a}_1 - \bar{a}_2)|}{|a_1 - a_2|} \approx 0.1 \cdot 10^{-1} > \varepsilon_m$$

Esempio

Implementiamo in MATLAB il calcolo della quantità

$$y = \frac{(1+x)-1}{x}$$

per i valori di x uguali a 10^{-k} , k=1,...,15. Tenendo conto che in precisione infinita di calcolo $y\equiv 1$ qualunque sia x, calcoliamo l'errore assoluto |1-y| associato a y (che in questo caso coincide con l'errore relativo), per ogni valore di x considerato.

```
>> format short e
>> k = 1:15;
>> x = 10.^-k;
>> y = ((1+x)-1)./x;
>> err = abs(1-y);
>> [x' err']
```

... segue esempio

```
ans =
1.0000e-01 8.8818e-16
1.0000e-02 8.8818e-16
1.0000e-03 1.1013e-13
1.0000e-04 1.1013e-13
1.0000e-05 6.5512e-12
1.0000e-06 8.2267e-11
1.0000e-07 5.8387e-10
1.0000e-08 6.0775e-09
1.0000e-09 8.2740e-08
1.0000e-10 8.2740e-08
1.0000e-11 8.2740e-08
1.0000e-12 8.8901e-05
1.0000e-13 7.9928e-04
1.0000e-14 7.9928e-04
1.0000e-15 1.1022e-01
```

Si osservi che, al diminuire dell'ordine di grandezza di x, la perdita di precisione aumenta e il fenomeno della cancellazione è sempre più eclatante!

Talvolta, manipolando opportunamente le espressioni matematiche che definiscono un problema, è possibile evitare il fenomeno della cancellazione numerica; quando ciò non è possibile si dice che la cancellazione è **insita nel problema**.

Esempio

Sia
$$y = \sqrt{x + \delta} - \sqrt{x} \operatorname{con} x, x + \delta > 0.$$

La cancellazione numerica per $|\delta| << x$ si elimina razionalizzando:

$$y = (\sqrt{x+\delta} - \sqrt{x}) \frac{\sqrt{x+\delta} + \sqrt{x}}{\sqrt{x+\delta} + \sqrt{x}} = \frac{\delta}{\sqrt{x+\delta} + \sqrt{x}}$$

La cancellazione numerica per $\delta \approx -x$ non è, invece, eliminabile in quanto è insita nel problema.

La cancellazione numerica talvolta si può eliminare utilizzando gli sviluppi di Taylor.

Esempio

Sia
$$y = \frac{e^x - 1}{x} \operatorname{con} x \neq 0.$$

La cancellazione numerica per $x \approx 0$ si elimina sostituendo e^x con il corrispondente sviluppo di Taylor, centrato in 0 e di ordine n:

$$y = \frac{1}{x} \left(\cancel{1} + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} + o(x^n) - \cancel{1} \right)$$
$$= 1 + \frac{x}{2!} + \frac{x^2}{3!} + \dots + \frac{x^{n-1}}{n!} + o(x^{n-1})$$

Successivamente per il calcolo numerico di *y* si sommano soltanto i primi termini, ovvero quelli che danno contributo alla somma; i restanti si trascurano perché, essendo piccoli rispetto al valore a cui vanno sommati, non modificano il valore della somma.

Condizionamento di un problema numerico

Nello studio della propagazione degli errori bisogna distinguere il ruolo del *problema* dal ruolo dell'*algoritmo* utilizzato per risolvere tale problema.

A tale scopo si introducono le definizioni di problema numerico e algoritmo, e i corrispondenti concetti di *condizionamento di un problema* e stabilità di un algoritmo.

Definizione

Si definisce **problema numerico** una relazione funzionale f, del tipo y = f(x) (**esplicita**) oppure f(x, y) = 0 (**implicita**), tra i dati x (**input**) e i risultati y (**output**).

I dati x e i risultati y devono essere rappresentabili da numeri, vettori o matrici di numeri di dimensione finita.

Sia y = f(x) oppure f(x, y) = 0 un generico problema numerico. Si denotino con

- \bar{x} una perturbazione dei dati x di input,
- \bar{y} i risultati ottenuti a partire dai dati \bar{x} in **precisione infinita di** calcolo.

Definizione

Un problema numerico si dice **ben condizionato** se accade che l'errore relativo associato a \bar{y} è dello stesso ordine di grandezza dell'errore relativo associato a \bar{x} o minore:

$$\frac{||y-\bar{y}||}{||y||} \approx \frac{||x-\bar{x}||}{||x||};$$

altrimenti, si dice mal condizionato.

Pertanto, un problema è ben condizionato quando le perturbazioni nei dati non influenzano eccessivamente i risultati. Per studiare il condizionamento di un problema si possono determinare stime del tipo:

$$\frac{||y - \bar{y}||}{||y||} \le K(f, x) \frac{||x - \bar{x}||}{||x||}$$

Definizione

Si definisce **numero di condizionamento** del problema il più piccolo valore di K(f,x) per cui vale la suddetta disuguaglianza.

Se K non è eccessivamente grande, il problema è ben condizionato.

Stabilità di un algoritmo

Definizione

Per algoritmo si intende una sequenza finita di operazioni (aritmetiche e non) che consente di ottenere l'output di un problema a partire dai dati di input.

Data un'aritmetica con precisione finita, si denotino con

- \bar{x} l'arrotondamento dei dati x di input,
- \bar{y} i risultati dell'algoritmo ottenuti a partire dai dati \bar{x} in **precisione** infinita di calcolo,
- \tilde{y} i risultati dell'algoritmo ottenuti a partire dai dati \bar{x} in **precisione** finita di calcolo.

Per giudicare la bontà di un algoritmo per la risoluzione di un problema, bisogna dunque confrontare la risposta \tilde{y} con \bar{y} .

Definizione

Un algoritmo si dice **numericamente stabile** se accade che l'errore relativo associato al risultato \tilde{y} ha lo stesso ordine di grandezza della precisione di macchina o minore

$$\frac{||\bar{y} - \tilde{y}||}{||\bar{y}||} \approx \varepsilon_{m}$$

altrimenti, si dice instabile.

Pertanto, un algoritmo è numericamente stabile quando la sequenza delle operazioni non amplifica eccessivamente gli errori di arrotondamento presenti nei dati.

Esempio

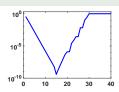
L'algoritmo per il calcolo di π definito dalla successione

$$x_1 = 2$$

 $x_n = 2^{n-1/2} \sqrt{1 - \sqrt{1 - 4^{1-n} x_{n-1}^2}} \qquad n \ge 2$

è instabile a causa del fenomeno della cancellazione numerica. Infatti, al crescere di n, la quantità $z=\sqrt{1-4^{1-n}x_{n-1}^2}$ si avvicina sempre più a 1 e genera una perdita di cifre nella successiva operazione 1-z.

Figura: Errore relativo $|\pi - x_n|/|\pi|$ al variare di n = 1, ..., 40



... segue esempio

Mediante la razionalizzazione è possibile eliminare la cancellazione numerica e l'algoritmo che così si ottiene

$$x_1 = 2$$

$$x_n = x_{n-1} \sqrt{\frac{2}{1 + \sqrt{1 - 4^{1-n}x_{n-1}^2}}} \qquad n \ge 2$$

è stabile.

Figura: Errore relativo $|\pi - x_n|/|\pi|$ al variare di n = 1, ..., 40

