

Corso di Calcolo Numerico

Dott.ssa M.C. De Bonis

Università degli Studi della Basilicata, Potenza

Facoltà di Ingegneria

Corso di Laurea in Ingegneria Meccanica

Sistemi di Numerazione

Sistema decimale

La base del sistema è **10**

Ogni numero viene rappresentato come

$$a = \pm a_m a_{m-1} \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-M}$$

con $0 \leq a_i \leq 9$

Il sistema è detto **posizionale**: il valore di ogni cifra varia in funzione della sua posizione nella rappresentazione decimale del numero

$$a = \pm \sum_{i=-M}^m a_i \times 10^i$$

Esempio

$$931.57 = 9 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 + 5 \times 10^{-1} + 7 \times 10^{-2}$$

La **rappresentazione decimale** di ogni numero reale è **unica**, eccetto quando la parte frazionaria contiene una sequenza di **9** consecutivi

Esempio

$$0.319999 \dots 9 \dots \quad \text{e} \quad 0.32$$

sono lo stesso numero

Qualunque intero $N > 1$ può essere scelto come base ed ogni numero reale **a** può essere scritto nella forma

$$a = \pm \sum_{i=-M}^m a_i \times N^i$$

ovvero

$$a = \pm a_m a_{m-1} \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-M}$$

con $0 \leq a_i \leq N - 1$

La **rappresentazione** di ogni numero reale in base N è **unica**, eccetto quando la parte frazionaria contiene una sequenza di cifre $a_k = N - 1$ consecutive.

Sistema binario

La base del sistema è **2**

Le cifre di questo sistema sono **0** e **1** e vengono dette **bit** (da **b**inary **d**igit): ogni numero reale **a** è rappresentato come una sequenza di **0** e **1**, ovvero

$$a = \pm \sum_{i=-M}^m a_i \times 2^i, \quad 0 \leq a_i \leq 1$$

Questo sistema è particolarmente interessante perché può essere realizzato con qualsiasi oggetto capace di assumere due stati diversi, uno per la cifra **0** e l'altro per la cifra **1**.

Esempi Stato di magnetizzazione e non di un nucleo di ferrite; conduttività e non di un diodo.

Rappresentazione dei numeri in un calcolatore

I numeri vengono rappresentati nel calcolatore secondo il **sistema binario** e, quindi, come una **sequenza di bit**.

Ad ogni numero reale viene riservato uno **spazio** finito di **memoria**, capace di contenere un numero finito di bit detto **parola**.

Detta l la lunghezza della parola, si possono rappresentare esattamente solo quei numeri la cui rappresentazione binaria consta di un numero di bit inferiore o uguale ad l .

Si parla di **numeri macchina**.

Tutte le operazioni fra i numeri macchina vengono effettuate utilizzando l'**aritmetica binaria**.

Numeri interi

Con parole di lunghezza l è possibile rappresentare tutti i numeri interi appartenenti all'intervallo

$$\left[-\frac{2^l}{2} + 1, \frac{2^l}{2} - 1 \right]$$

Esempio Se $l = 16$ sono rappresentabili tutti i numeri interi appartenenti all'intervallo

$$[-32767, 32767]$$

Numeri reali

Ogni numero reale **a** può essere scritto nella forma

$$a = pN^q$$

dove $p \in \mathbb{R}$, N è la base del sistema di numerazione e $q \in \mathbb{Z}$.

Questa rappresentazione, detta **in virgola mobile** (floating-point), non è unica, infatti

$$321.25 = 32.125 \times 10^1 = 0.32125 \times 10^3$$

La rappresentazione di **a** si dice **normalizzata** quando

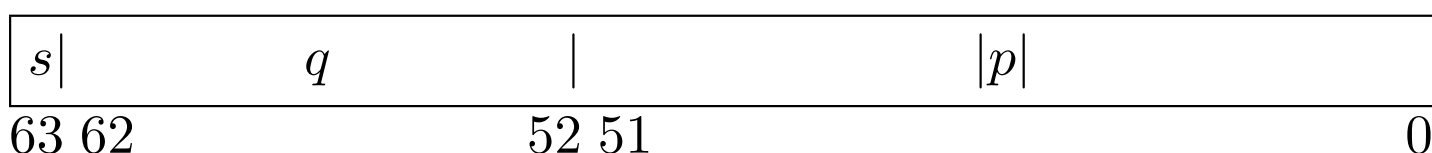
$$N^{-1} \leq |p| < 1.$$

Le cifre di p si dicono **cifre significative**

Esempi La rappresentazione normalizzata di $a = 92.25$ è
 $a = 0.9225 \times 10^2$;

la rappresentazione normalizzata di $a = 0.000718$ è
 $a = 0.718000 \times 10^{-3}$

Doppia Precisione



È possibile rappresentare tutti i numeri reali della seguente forma

$$\pm 0.d_1 d_2 \cdots d_{15} \times 10^n,$$

dove $0 < d_1 \leq 9, 0 \leq d_i \leq 9, i = 2, \dots, 15$, e $-308 \leq n \leq 308$

15 cifre significative

Conseguenze della rappresentazione dei numeri come parole di lunghezza fissa

1. L'insieme $\mathbb{R}$ dei numeri reali è infinito, ma, soltanto un sottoinsieme di $\mathbb{R}$, limitato inferiormente e superiormente, è rappresentabile in un calcolatore.

In seguito denoteremo con $\mathbb{F}$ il sottoinsieme dei numeri reali rappresentabili in un calcolatore.

2. Non sempre una operazione tra due o più numeri macchina produce un risultato.

Se tale operazione ha come risultato un numero più grande dell'estremo superiore di $\mathbb{F}$ si genera una situazione di **overflow**; se, invece, ha come risultato un numero più piccolo dell'estremo inferiore di $\mathbb{F}$ si genera una situazione di **underflow**.

3. L'insieme $\mathbb{R}$ dei numeri reali è denso, cioè tra due numeri reali r_1 e r_2 esiste sempre un numero reale r_3 tale che

$$r_1 < r_3 < r_2.$$

Invece, il sottoinsieme $\mathbb{F}$, oltre ad essere limitato inferiormente e superiormente, è anche “bucato”.

Esempio Utilizzando la rappresentazione in singola precisione, i numeri macchina

$$0.123456 \quad \text{e} \quad 0.123457$$

sono consecutivi e tra di essi non esiste alcun altro numero macchina.

Errori

Se $\mathbf{a}$ è un numero reale, un numero $\mathbf{A}$ che differisce di poco da a si dice **approssimazione di a** .

Se $A < a$, si parla di approssimazione per difetto.

Se $A > a$, si parla di approssimazione per eccesso.

La quantità

$$\Delta a = |a - A|$$

si chiama **errore assoluto**

La quantità

$$\delta a = \left| \frac{a - A}{a} \right|$$

si chiama **errore relativo**

Siano

$$a = pN^q \quad \text{e} \quad A = \bar{p}N^q$$

con

$$N^{-1} \leq |p|, |\bar{p}| < 1$$

Quando

$$\Delta_a \leq \frac{1}{2}N^{-k}, \quad k \geq 1,$$

si dice che A ha almeno k **cifre decimali corrette** nella base N .

Quando

$$\delta_a \leq \frac{1}{2}N^{-k}, \quad k \geq 1,$$

si ha anche

$$\left| \frac{p - \bar{p}}{p} \right| \leq \frac{1}{2}N^{-k} \Rightarrow |p - \bar{p}| \leq \frac{1}{2}N^{-k}|p| < \frac{1}{2}N^{-k}$$

e, dunque, A ha almeno k **cifre significative** nella base N .

Esempi

1. $a = 15.2000, A = 15.1997, N = 10$

$$\Delta_a = 0.3 \dots \times 10^{-3} < \frac{1}{2}10^{-3}$$

dunque A ha almeno 3 cifre decimali corrette (in base 10)

$$\delta_a = 0.197 \dots \times 10^{-4} < \frac{1}{2}10^{-4}$$

dunque A ha almeno 4 cifre significative

2. $a = 99.2000, A = 99.1997, N = 10$

$$\Delta_a = 0.3 \dots \times 10^{-3} < \frac{1}{2}10^{-3}$$

dunque A ha almeno 3 cifre decimali corrette

$$\delta_a = 0.3 \dots \times 10^{-5} < \frac{1}{2}10^{-5}$$

dunque A ha almeno 5 cifre significative

Cause principali di errori

Le sorgenti di errore possono essere, in generale, suddivise in cinque gruppi.

1. Errori di impostazione del problema

Non sempre le formule matematiche rappresentano modelli esatti dei fenomeni reali; generalmente questi modelli sono più o meno idealizzati. Infatti, nello studio dei vari fenomeni della natura, siamo costretti, al fine di rendere più semplice il problema, ad ammettere certe condizioni e ciò implica una serie di errori.

2. Errori di metodo

Talvolta succede che è difficile, se non impossibile, risolvere un problema impostato in termini rigorosi. In questo caso il problema in esame viene sostituito da un altro problema approssimato i cui risultati si distinguono di poco dal problema dato.

3. Errori di troncamento

Le funzioni che appaiono in formule matematiche sono spesso date in forma di successioni infinite o di serie e numerose equazioni matematiche non possono essere risolte che attraverso una descrizione di processi infiniti, i cui limiti sono appunto le soluzioni cercate.

Siccome un processo infinito non può, in generale, terminare in un numero finito di passi, siamo costretti a interrompere la successione infinita di passi computazionali, necessaria per ottenere un risultato esatto.

4. Errori di rappresentazione dei dati o di arrotondamento

Abbiamo visto che non tutti i numeri reali possono essere rappresentati in un calcolatore. Se il calcolatore utilizza una rappresentazione con t cifre per la mantissa, tutti i numeri reali compresi tra l'estremo inferiore e l'estremo superiore di $\mathbb{F}$ la cui mantissa ha un numero di cifre superiore a t dovranno in qualche modo essere “accorciati”, o, meglio, **arrotondati** a t cifre.

5. Errori nei calcoli

Effettuando calcoli tra numeri macchina approssimati, gli errori di rappresentazione dei dati iniziali si propagano in qualche misura sui risultati di tali calcoli. L'entità della propagazione dipende anche dall'algoritmo utilizzato.

Per un'analisi completa sarebbe necessario esaminare tutti gli errori, ma, in seguito, ci limiteremo essenzialmente al calcolo degli errori di rappresentazione dei dati, di calcolo e di metodo.

Errori di rappresentazione dei dati

Regola di arrotondamento

Per arrotondare un numero

$$a = 0.d_1 d_2 \dots d_k d_{k+1} d_{k+2} \dots \times N^q$$

fino a k cifre significative, si eliminano tutte le cifre $d_{k+1} d_{k+2} \dots$ a destra della k -esima e si rimpiazza d_k con la cifra c , dove

$$\begin{cases} c = d_k & \text{se } 0 \leq d_{k+1} \leq 4 \\ c = d_k + 1 & \text{se } 6 \leq d_{k+1} \leq 9 \\ c = d_k & \text{se } d_{k+1} = 5 \text{ e } d_k \text{ è pari} \\ c = d_k + 1 & \text{se } d_{k+1} = 5 \text{ e } d_k \text{ è dispari} \end{cases}$$

Esempi

1. Arrotondando il numero

0.31415926536

fino a 6, 5 e 4 cifre significative, si ottengono
rispettivamente in numeri approssimati

0.314159 0.31416 0.3142

2. Arrotondando il numero

0.7268165467112

fino a 6, 5 e 4 cifre significative, si ottengono
rispettivamente in numeri approssimati

0.726816 0.72682 0.7268

Arrotondando un numero reale $a = pN^q$ a k cifre significative si ottiene un numero macchina $A = \bar{p}N^q$ che è una approssimazione di a .

Si ha

$$\Delta_a = N^q |p - \bar{p}| \leq \frac{1}{2} N^{q-k}$$

e

$$\delta_a = \frac{\Delta_a}{|a|} \leq \frac{1}{2} \frac{N^{q-k}}{|p|N^q} \leq \frac{1}{2} \frac{N^{q-k}}{N^{q-1}} \leq \frac{1}{2} N^{1-k}$$

La quantità

$$eps = \frac{1}{2} N^{1-k}$$

è detta **precisione di macchina o epsilon-macchina**.

Essa è una costante propria di ciascun calcolatore e rappresenta la massima precisione di calcolo raggiungibile.

Pertanto, non ha senso cercare di determinare approssimazioni con errore relativo inferiore ad eps .

Denotiamo con fl l'operazione di arrotondamento. Dunque se A è il numero macchina che si ottiene arrotondando il numero reale a , scriviamo

$$A = fl(a)$$

Caratterizzazione La quantità eps è il più piccolo numero macchina per il quale si ha

$$fl(1 + eps) > 1.$$

Sia a un numero reale. Ponendo

$$\varepsilon = \frac{fl(a) - a}{a}, \quad |\varepsilon| = \delta_a \leq eps$$

si ottiene

$$fl(a) = a(1 + \varepsilon)$$

Errori nei calcoli

I risultati di operazioni aritmetiche tra numeri macchina generalmente non sono numeri macchina.

Si può verificare una situazione di underflow o di overflow, oppure la forma normalizzata del risultato ha un numero di cifre decimali superiore alla precisione con cui si sta lavorando.

Pertanto in un calcolatore è impossibile implementare esattamente le operazioni aritmetiche.

Dovremo accontentarci delle cosiddette **operazioni macchina**: a due numeri macchina viene associato un terzo numero macchina ottenuto arrotondando l'esatto risultato dell'operazione aritmetica.

Siano a e b due numeri reali e siano $A = fl(a)$ e $B = fl(b)$ i corrispondenti numeri macchina.

Denotando con $\oplus, \ominus, \otimes, \oslash$ le operazioni macchina corrispondenti rispettivamente alle operazioni aritmetiche $+, -, \times, /$, si ha

$$A \oplus B = fl(A + B) = (A + B)(1 + \varepsilon_1)$$

$$A \ominus B = fl(A - B) = (A - B)(1 + \varepsilon_2)$$

$$A \otimes B = fl(A \times B) = (A \times B)(1 + \varepsilon_3)$$

$$A \oslash B = fl(A/B) = (A/B)(1 + \varepsilon_4)$$

con

$$|\varepsilon_i| \leq eps, \quad i = 1, 2, 3, 4$$

Dunque, quando si effettua una operazione macchina si commette un errore dell'ordine della precisione di macchina eps .

Domanda Per le operazioni macchina valgono ancora le note proprietà (commutativa, associativa, distributiva, etc) delle operazioni aritmetiche?

La risposta non è sempre affermativa.

La proprietà commutativa permane:

$$A \oplus B = B \oplus A \quad \text{e} \quad A \otimes B = B \otimes A$$

mentre, in generale,

$$(A \oplus B) \oplus C \neq A \oplus (B \oplus C)$$

$$(A \otimes B) \otimes C \neq A \otimes (B \otimes C)$$

$$A \otimes (B \oplus C) \neq (A \otimes B) \oplus (A \otimes C)$$

Un'ulteriore relazione anomala è

$$A \oplus B = A, \quad \text{se } |B| \leq \frac{eps}{N} |A|$$

infatti

$$A \oplus B = A \times \left[1 \oplus \frac{eps}{N} \right] = A \times fl \left(1 + \frac{eps}{N} \right) = A$$

Conclusioni

- Le espressioni equivalenti in aritmetica esatta non lo sono generalmente in aritmetica con precisione finita.
- Due espressioni macchina sono equivalenti se differiscono per un ε dell'ordine della precisione macchina *eps*

Cancellazione Numerica

È un fenomeno che si verifica durante l'operazione di sottrazione tra due numeri “quasi uguali”.

Siano x_1 e x_2 due numeri reali. Se $x = x_1 - x_2$ è “molto piccolo”, l'errore relativo

$$\delta_x = \left| \frac{fl(x_1 - x_2) - x}{x} \right|$$

può essere “molto grande” e ciò produce una **perdita di cifre significative** nel calcolo di $fl(x_1 - x_2)$

Esempi $N = 10, k = 6$

1. $x_1 = 0.147554326 \times 10^2$

$$\Rightarrow X_1 = fl(x_1) = 0.147554 \times 10^2$$

$$x_2 = 0.147251742 \times 10^2$$

$$\Rightarrow X_2 = fl(x_2) = 0.147252 \times 10^2$$

$$fl(X_1 - X_2) = 0.000302 \times 10^2 = 0.302000 \times 10^{-1}$$

mentre

$$x_1 - x_2 = 0.302584 \times 10^{-1}$$

2.

$$x_1 = \sqrt{2.01} \Rightarrow X_1 = fl(x_1) = 0.141774 \times 10^1$$

$$x_2 = \sqrt{2} \Rightarrow X_2 = fl(x_2) = 0.141421 \times 10^1$$

$$fl(X_1 - X_2) = 0.353000 \times 10^{-2}$$

mentre

$$x_1 - x_2 = 0.35311255 \dots 10^{-2}$$

È sempre preferibile evitare la sottrazione tra numeri macchina “quasi uguali”.

Propagazione degli errori

Problema numerico: è una descrizione chiara e non ambigua di una connessione funzionale tra i dati e i risultati

Se x è il vettore dei dati detti **input** e y è il vettore dei risultati detti **output**, un problema numerico può essere espresso come

$$y = f(x)$$

Algoritmo: è una sequenza finita di operazioni logiche e non ambigue che opera sui dati x e produce come risultato il vettore y^* non necessariamente uguale a y

In generale, esistono diversi algoritmi per la risoluzione dello stesso problema numerico.

Condizionamento di un problema

Un problema numerico si dice **ben condizionato** se a “piccole” perturbazioni sui dati corrispondono “piccole” perturbazioni sui risultati; si dice **mal condizionato** se a “piccole” perturbazioni sui dati corrispondono “grandi” perturbazioni sui risultati.

Supponiamo che il dato x sia perturbato mediante un Δx e poniamo $\bar{x} = x + \Delta x$, in generale si ha

$$\left| \frac{f(x) - f(\bar{x})}{f(x)} \right| \leq k \left| \frac{\Delta x}{x} \right|$$

e k viene detto **indice o numero di condizionamento** del problema numerico f

Tanto più k è “piccolo” tanto più il problema è ben condizionato

Il condizionamento è una caratteristica propria del problema numerico e non ha alcun legame né con gli errori di arrotondamento delle operazioni macchina, né con gli algoritmi eventualmente scelti per risolverlo.

Condizionamento dell'operazione di somma

Definiamo $f(a, b) = a + b$ e perturbiamo a e b rispettivamente con Δa e Δb

$$\begin{aligned} \frac{|f(a, b) - f(a + \Delta a, b + \Delta b)|}{|f(a, b)|} &= \frac{|\Delta a + \Delta b|}{|a + b|} \leq \\ &\leq \frac{\max\{|a|, |b|\}}{|a + b|} \left(\frac{|\Delta a|}{|a|} + \frac{|\Delta b|}{|b|} \right) \end{aligned}$$

$a + b \rightarrow 0 \Rightarrow$ l'operazione di somma è
mal condizionata

$\Rightarrow$ **Cancellazione Numerica**

Stabilità di un Algoritmo

Poichè i dati di input x di un algoritmo sono sempre affetti dagli errori di arrotondamento (dell'ordine di eps) occorre confrontare i risultati y^* dell'algoritmo non con la soluzione $y = f(x)$ ma bensì con $\bar{y} = f(\bar{x})$ dove $\bar{x} = fl(x)$

Si dice che un algoritmo è **stabile** se la quantità

$$\frac{|f(\bar{x}) - y^*|}{|f(\bar{x})|}$$

è dell'ordine della precisione di macchina eps

$\Rightarrow$ un algoritmo si dice stabile se non amplifica gli errori di arrotondamento

Esempio

Consideriamo il problema numerico

$$f(x, y) = \sqrt{x} - \sqrt{y}$$

Abbiamo visto che se

$$y = x + \Delta x \quad \text{e} \quad |\Delta x| \quad \text{è piccolo}$$

il problema è mal condizionato (Cancellazione Numerica)

Osserviamo che

$$f(x, y) = \sqrt{x} - \sqrt{y} \tag{1}$$

$$= \frac{x - y}{\sqrt{x} + \sqrt{y}} \tag{2}$$

L'algoritmo che risolve il problema numerico f con la formula (1) è instabile, mentre l'algoritmo che risolve il problema numerico f con la formula (2) è sempre stabile.

Esempio

$$x = 2.01 \quad y = 2 \Rightarrow \sqrt{x} - \sqrt{y} = 0.35311255 \dots 10^{-2}$$

$$\bar{x} = fl(\sqrt{x}) = 0.141774 \times 10^1$$

$$\bar{y} = fl(\sqrt{y}) = 0.141421 \times 10^1$$

$$fl(\bar{x} - \bar{y}) = 0.353000 \times 10^{-2}$$

mentre

$$\frac{x - y}{fl(\bar{x} + \bar{y})} = 0.353114 \times 10^{-2}$$

Esercizi Calcolare

1. $\sqrt{1.0001} - 1$

2. $\sqrt{0.2} - \sqrt{0.1}$

3. $70 - \sqrt{4899}$

usando le formule (1) e (2).

Conclusioni

- La propagazione degli errori nei calcoli dipende sia dal condizionamento del problema numerico, sia dalla stabilità dell'algoritmo che si utilizza.
- Un problema ben condizionato, risolto con un algoritmo stabile conduce a risultati attendibili
- Si può pervenire a risultati non accurati se
 - un algoritmo stabile viene applicato ad un problema mal condizionato
 - un problema ben condizionato viene risolto con un algoritmo instabile
- A volte può accadere che un problema sia ben condizionato per certi dati iniziali e mal condizionato per altri (vedi l'operazione di somma)