01 - Numeri e aritemetica di macchina

Analisi Numerica

L'analisi numerica si occupa di dare una risposta a problemi che riguardano il mondo reale.

Esempio: voglio calcolare l'area della superficie terrestre

Quali sono gli step da seguire?

1. Conversione del problema del mondo reale a problema del mondo matematico
2. Trasformazione da problema matematico a problema numerico risolvibile su un calcolatore mediante un algoritmo

Nel fare questi passaggio tuttavia introduciamo delle approssimazioni ad ogni passo (raggio della terra approssimato, π approssimato, etc...)

Il secondo problema è capire se la soluzione trovata è affidabile e porta ad un risultato accettabile rispetto alla soluzione reale o meno.

Sorgenti di errori

1. **Errori del modello matematico** (prodotti dal passaggio da problema reale a problema matematico)
2. **Errori nel modello numerico-computazionale** (prodotti dal passaggio da problema matematico a problema numerico)
3. **Errori presenti nei dati** (dati sbagliati per misurazioni errate)
4. **Errori di arrotondamento nei dati e nei calcoli** (approssimazione numerica)

Classificazione dei problemi numerici/computazionali

- Problemi Diretti
- Problemi Inversi
- Problemi Di Identificazione

Siano **x** e **y** dati e risultati e ϕ la relazione tra dati e risultati.

Problemi Diretti

Conosco **x** e ϕ, voglio trovare **y**

Esempio: Calcolo integrale definiti

Problemi Inversi

Consoco **y** e ϕ, voglio trovate **x**

Esempio: Risoluzione sistema lineare

$$Ax = y$$

Conosco A e conosco **y**, voglio trovare la matrice colonna **x**

Problemi Di Identificazione

Conosco **x** e **y**, voglio trovare ϕ

Esempio: Approssimazione di dati

Tipi di problemi

Ben posto

Se la sua soluzione:

- esiste
- è unica
- dipende in modo continuo dai dati del problema

Mal posto

Se non è ben posto

Ben condizionato (Anticipazione)

Un problema dove a piccole perturbazioni sui dati corrispondano piccole perturbazioni sulle soluzioni.

Noi lavoreremo solo con problemi **ben posti** e **ben condizionati**.

Sistema di numerazione posizionale

Numeri reali in base β:

$$\pm(a_n ... a_0 . b_1 b_2 ...)_\beta$$

$$(\sum_{i=0}^{n} a_i \beta^i + \sum_{i=1}^{\infty} b_i \beta^{-i})$$

Tutte le cifre vanno da 0 a $\beta - 1$

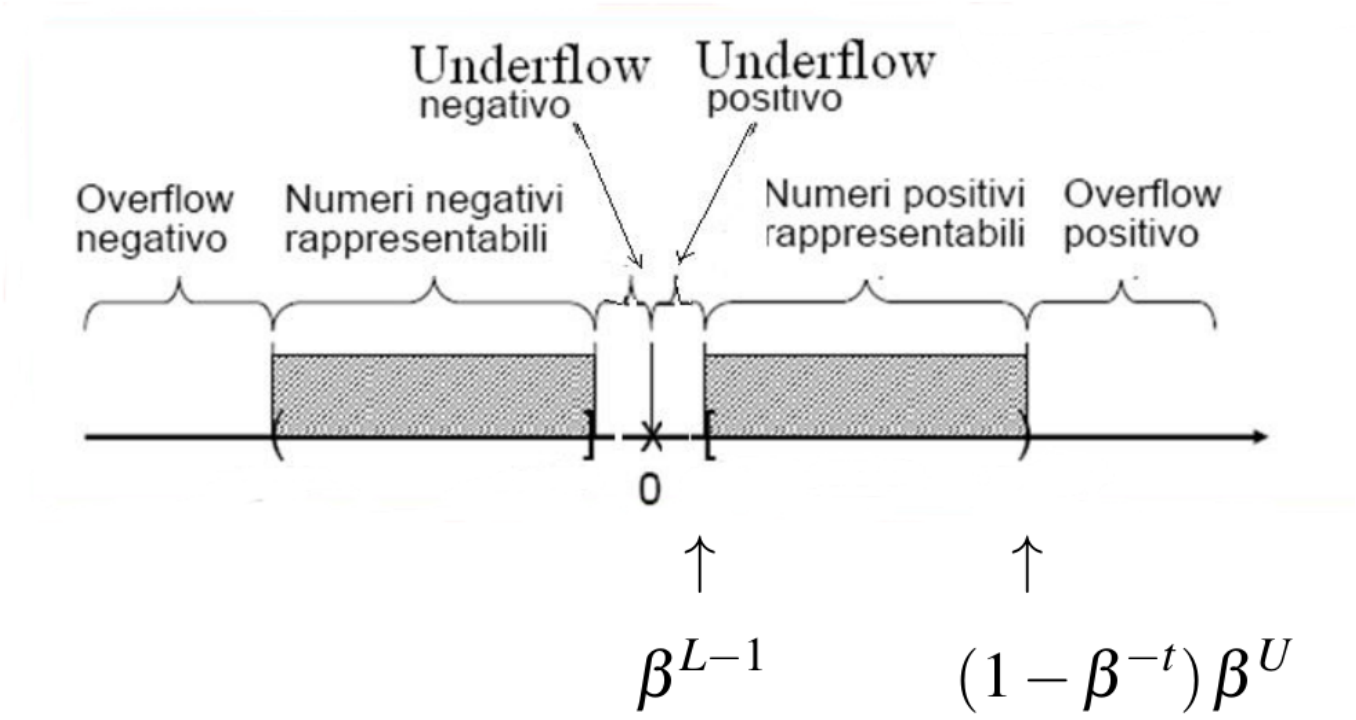
Rappresentzione normalizzata in base β

m = mantissa
p = esponente
β^{-1} è la più piccola mantissa e il limite superiore è 1

$$x = (\pm 0.d_1 d_2 d_3 ...)_\beta$$

$$\pm\beta^p \sum_{i>=1} (d_i)_\beta * \beta^{-i}$$

Numeri Di Macchina



Numeri usati nei calcolatori (sono finiti)
L'insieme dei numeri di macchina è definito come segue:

$$F(\beta, t, L, U)$$

- β = **base di rappresentazione**
- t = **numero di cifre della mantissima**
- L = **minimo valore dell'esponente** (negativo) [Lower Bound]
- U = **massimo valore dell'esponente** (positivo) [Upper Bound]

Quindi dato un qualsiasi numero $x = (\pm 0.d_1 d_2 d_3 ...)_\beta$ abbiamo che:

- $1 \leq d_i \leq \beta - 1 \forall i = 2, ..., t$
- $L \leq p \leq U$

Tipi di errori:

- **Overflow** numero troppo grande o troppo piccolo
- **Underflow** numero troppo vicino allo 0

Esempio:

Poichè in base 2 la prima cifra è per forza 1 allorà questo non verrà memorizzato.

$$F(2,3,-2,1) = \{\pm 0.1 d_1 d_2 \times 2^p\}$$

- $0 \leq d_1, d_2 \leq 1$
- $-2 \leq p \leq 1 => p = \{-2, -1, 0, 1\}$

p/m	**-2**	**-1**	**0**	**1**
100	$0.1 * 2^{-2}$	$0.1 * 2^{-1}$	$0.1 * 2^{0}$	$0.1 * 2^{1}$
101	$0.101 * 2^{-2}$	-	-	-
110	$0.110 * 2^{-2}$	-	-	$0.110 * 2^{1}$
111	$0.111 * 2^{-2}$	-	-	-

Qual'è la cardinalità dell'insieme F? **33** elementi

- **16** positivi
- **16** negativi
- lo **0**

Il più piccolo numero è nella prima entry della tabella, il più grande nell'ultima.

Il più piccolo numero positivo rappresentabile è β^{L-1}
Il più grande numero positivo rappresentabile è $(1-\beta^{-t})\beta^{U}$
Il più piccolo numero negativo rappresentabile è $-[\beta^{L-1}]$
Il più grande numero positivo rappresentabile è $-[(1-\beta^{-t})\beta^{U}]$

Osservazione : mentre β^{L} appartiene ad F, β^{U} non appartiene ad F

Conversione

$$0.100 \times 2^{-2} = (1 \times 2^{-1} + 0 \times 2^{-2} + 0 \times 2^{-3}) \times 2^{-2} = 1/8$$

Cardinalità di F

Nota: si osserva che i numeri positivi sono tanti quanti i numeri positivi.

Procedimento:

- Si osserva quanti segmenti di tipo $[\beta^p, \beta^{p+1}]$ sono contenuti in F
- Si contano il numero di elementi contenuti in ciascun segmento

Poichè in ciascun segmento definito tra due potenze della base β c'è sempre la stessa quantità di numeri di macchina

Domanda: Quanti segmenti contiene l'insieme F^+ (solo positivi)

- $U - L + 1$ = **numero di segmenti positivi**

Domanda: Quanti elementi ci sono in ciascuno di questi segmenti
Mi devo calcolare la differenza tra i due valori estremi del segmento, posso scegliere un qualsiasi

segmento quindi viene comodo lavorare sul primo segmento.
Il primo numero sarà dato da $(0.10000...0)\beta^{p+1}$
Il successivo sarà $(0.10000...1)\beta^{p+1}$

A questo punto definiamo lo **spacing (s)** è la differenza tra i due numeri che viene

- $(0.0000...1)\beta^{p+1}$
  Quindi poichè il numero tra parentesi è β^{-t} abbiamo che lo **spacing di** F è:

$$s = \beta^{p+1-t}$$

Quanti numeri di F sono presenti in $[\beta^p, \beta^{p+1}]$?

$$\frac{(\beta^{p+1} - \beta^p)}{s}\beta^{-p+1+t}$$

che è uguale a

$$(\beta - 1)\beta^{t-1}$$

La cardinalità di F^+ è quindi $(U - L + 1)(\beta - 1)\beta^{t-1}$

Questà andrà moltiplicata per due per aggiungere i numeri negativi e andrà sommato 1 per contare anche lo 0:

Cardinalità di F = 2 * $|F^+| + 1$

$$2 \times ((U - L + 1)(\beta - 1)\beta^{t-1}) + 1$$

Approssimazione (Floating)

Troncamento

Rimuovo tutte le cifre dopo la t-esima
$x = \pm 0.d_1 d_2 ... d_t d_{t+1} ... => \pm 0.d_1 d_2 ... d_t$

Esempio:
$F(10, 4, L, U)$
$x = 0.372145 => 0.3721$

Arrotondamento

Prendo il numero di macchina più vicino.
Si può usare solo per β pari (noi usiamo la base 10, il pc la base 2 quindi non cè problema)

Devo sommare $\beta/2$ alla $(t+1)$-esima cifra e poi troncare alla t-esima cifra.

B = 10, t = 4

EX1:
$x = 0.3798165$
Aggiungo $\beta/2$ => $x = 0.37986$ => $x = 0.3798$

EX2:
$x = 0.1265873$
Aggiungo $\beta/2$ => $x = 0.12663$ => $x = 0.1266$

Rounding to even

Nel caso il numero si trovi esattamente a metà tra due numeri di macchina viene scelto quello pari.

Normalizzazione

Prima di effettuare un arrotondamento/troncamento bisogna normalizzare il numero.
Ad esempio:
$x = 0.01236$ con $t = 3$ non va arrotondato subito a 0.012 ma prima va normalizzato in forma $x = 0.1236$ e poi si può arrotondare a 0.124

Precisione Di Macchina

La precisione di macchina **eps** è lo spacing relativo al segmento $[\beta^0, \beta^1] = [1, \beta]$:

$$eps = \beta^{1-t}$$

L' **unità di arrotondamento o Roundoff Unit** vale:

$$u = (1/2)eps = (1/2)\beta^{1-t}$$

Domanda: chi è il numero di macchina successivo ad 1?

- $1 + eps$

Errore assoluto

$$fl = floating$$
$$E_{ass} = |x - fl(x)|$$

Errore relativo

$$E_{rel} = |E_{ass}/x| = |(x - fl(x))/x| \ \ \forall x \neq 0$$

Errore approsimmazione

Nel caso di **approssimazione per troncamento** l'errore viene sempre $\leq eps$

$$E_{ass} \leq eps$$

Nel caso di **approssimazione per arrotondamento** l'errore viene sempre $\leq u$ ovvero $\leq (1/2)eps$

$$E_{ass} \leq u \leq (1/2)eps$$

Quindi l'approssimazione per arrotondamento porta sempre ad un errore minore

---

Errori Operazioni

Ricordiamo $fl_A(x)$ il **floating per arrotondamento** e

$$E_{rel}^{A} = |(fl_A(x) - x)/x| \leq u$$

e definiamo

$$\xi_x = (fl_A(x) - x)/x \ \leq u$$

$$x\xi_x = fl_A(x) - x$$

$$fl_A(x) = x\xi_x + x = x(1 + \xi_x)$$

Siano $x, y \in R - \{0\}$, quello che ci chiediamo e se appartengono anche ad F
Siano $\{\times, /, +, -\}$ l'insieme delle operazioni e $*$ una generica operazione

$$x * y$$

Risultato calcolato in F

- $x => fl_A(x) \in F$
- $y => fl_A(y) \in F$

L'operazione da fare ora è quindi:

$$fl_A(x) * fl_A(y)$$

Tuttavia non sappiamo se il risultato apparterrà ad F quindi l'operazione finale diventa:

$$fl_A(fl_A(x) * fl_A(y)) \in F$$

Calcolandoci l'errore relativo sostituendo i vari $fl_A(x)$ con

$$fl_A(x) = x\xi_x + x = x(1 + \xi_x)$$

troviamo che l'errore generico di una qualsiasi operazione è:

$$(\frac{x * y - (x(1 + \xi_x) * y(1 + \xi_y))(1 + \xi_r)}{|x * y|})$$

dove ξ_r è quello relativo all'operazione finale.

Moltiplicazione/Divisione

1. Si esegue il prodotto/divisione delle mantisse e si sommano/sottraggono gli esponenti
2. Si ricava il floating del risultato (si normalizza il numero troncando o arrotondando se necessario)

Errore relativo nel prodotto

$$E_{rel}^{prod} \leq |\xi_x| + |\xi_y| + |\xi_{prod}| \leq 3u$$

Il prodotto risulta quindi un operazione sempre stabile a prescindere dai numeri coinvolti.

Errore relativo nella divisione

$$E_{rel}^{div} \leq |\xi_x| + |\xi_y| + |\xi_{div}| \leq 3u$$

(Ricavato tramite espansione di Taylor al primo ordine).
Come la moltiplicazione **la divisione è un operazione sempre stabile** a prescindere dai numeri coinvolti

---

Somma algebrica

1. Si trasorma il numero con esponente minore in modo che i 2 numeri abbiano lo stesso esponente (uno dei due perde la forma in virgola mobile normalizzata)
2. Si sommando le mantisse (lasciando invariati gli esponenti)
3. Si ricava il floating del risultato (si normalizza il numero troncando o arrotondando se necessario)

Errore relativo nella somma

$$E_{rel}^{sum} = (|\frac{x}{x+y}| + |\frac{y}{x+y}| + 1)u$$

L'operazione non risulta stabile poichè dipende dai valori di x e y dato che se $x + y$ è vicino a 0 l'errore relativo cresce molto.
Il caso pericoloso lo si ha quando:

- x, y hanno segni discordi
- $|x|$ vicino a $|y|$

In questo caso otteniamo un fenomeno di **cancellazione numerica**
Questo fenomeno capita solo se almeno uno dei due non appartiene ad F.

In caso in cui invece x, y abbiano lo stesso segno:

- $|\frac{x}{x+y}| \leq 1$
- $|\frac{y}{x+y}| \leq 1$

Quindi

$$E_{rel}^{sum} = (|\frac{x}{x+y}| + |\frac{y}{x+y}| + 1)u \leq 3u$$

Study Case

Siano dati due numeri a, b e si vuole calcolare la loro media.
Due strade possibili:

1. $\frac{a+b}{2}$
2. $a + \frac{b-a}{2}$

Possiamo notare tramite i dovuti calcoli che in alcuni casi la prima strada porta ad un risultato maggiore che è oltre il range accettabile $[a, b]$ poichè fornisce come risultato un numero maggiore di b.

La seconda strada risulta più stabile poichè porta ad un risultato molto più vicino a quello effettivo.