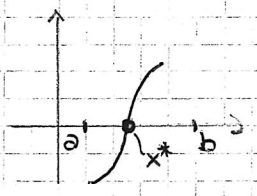
1. ERRORI ED ARITMETICA FINITA

TEOREMA DI ESISTENZA DEGLI ZERI

$f:[a,b] \to \mathbb{R}$, $f \in C^0([a,b])$, $f(a) \cdot f(b) < 0 \Rightarrow \exists\, x^* \in [a,b]$ t.c. $f(x^*) = 0$



→ mi dice che $f(x^*) = 0$, ma non come trovare x^*

PROPRIETÁ CHE SI RICHIEDONO AD UN BUON METODO

1. Approssimazione sufficientemente accurata della soluzione del problema
2. Basso costo computazionale (minor numero di conti possibili)
3. Facilmente implementabile su calcolatore elettronico

* Parlando di accuratezza si introduce il concetto di ERRORE che si distingue in 2 tipi:

ERRORE ASSOLUTO

x = soluzione esatta

$\hat{x}$ = soluzione approssimata fornita dal metodo

$$E.A. = \Delta x = \hat{x} - x$$

Supponiamo di avere $10^{-4} \leq |\Delta x| \leq 10^{3}$: Δx é grande o piccolo?

se $x = 10^{8} \Rightarrow \Delta x$ é piccolo

se $x = 10^{-8} \Rightarrow \Delta x$ é grande

ERRORE RELATIVO

$$E_x = \frac{\Delta x}{x} = \frac{\hat{x} - x}{x} = \frac{\hat{x}}{x} - 1 \qquad \hat{x} = x(E_x + 1) \Rightarrow \frac{\hat{x}}{x} = E_x + 1$$

(se non commetto errore $\frac{\hat{x}}{x} = 1$)

Quindi per stabilire se E_x é grande o piccolo va confrontato con 1:

$E_x = 0{,}001 = 0{,}1\%$ (piccolo)

$E_x = 0{,}9 = 90\%$ (Grande)



TIPOLOGIE DI ERRORE

1) ERRORE DI DISCRETIZZAZIONE O TRONCAMENTO
2) ERRORE DI CONVERGENZA
3) ERRORE DI ROUND-OFF

1) ERRORE DI DISCRETIZZAZIONE O TRONCAMENTO

Dovuto alla definizione del metodo stesso. Spesso il problema da risolvere é definito su un dominio continuo, ma il calcolatore elettronico rappresenta solo un sottoinsieme dei numeri naturali. Per procedere sostituisco al problema esatto un problema approssimato su un dominio discreto (= insieme finito di numeri). É in questo passaggio che si introducono gli errori di discretizzazione:

PROBLEMA ESATTO SU DOMINIO CONTINUO → PROBLEMA APPROSSIMATO SU DOMINIO DISCRETO

ESEMPIO

Vediamo l'approssimazione della derivata prima:

$f: \mathbb{R} \to \mathbb{R}$ derivabile in $\bar{x} \in \mathbb{R}$

Dalla definizione di derivata prima abbiamo che:

$$f'(\bar{x}) = \lim_{h \to 0} \frac{f(\bar{x}+h) - f(\bar{x})}{h}$$

↘ non calcolabile con calcolatore. Quindi scelgo un $\bar{h}$ piccolissimo e approssimo.

$$f'(\bar{x}) \cong \frac{f(\bar{x}+\bar{h}) - f(\bar{x})}{\bar{h}} = \varphi_{\bar{h}}(\bar{x}) = \text{approssimazione della derivata prima di } f \text{ in } \bar{x}$$

ERRORE ASSOLUTO DI DISCRETIZZAZIONE $= \varphi_{\bar{h}}(\bar{x}) - f'(\bar{x})$

Per stimare questo errore assumo che $f \in C^2(\mathbb{R})$. Quindi scrivo lo sviluppo di Taylor:

$$f(\bar{x}+\bar{h}) = f(\bar{x}) + f'(\bar{x})\bar{h} + \underbrace{\frac{1}{2} f''(\xi)\bar{h}^2}_{\text{resto di Lagrange con } \bar{x} \le \xi \le \bar{x}+\bar{h}}$$

$$|\xi - \bar{x}| \le \bar{h}$$

$$\underbrace{\frac{f(\bar{x}+\bar{h}) - f(\bar{x})}{\bar{h}}}_{\varphi_{\bar{h}}(\bar{x})} - f'(\bar{x}) = \frac{1}{2} f''(\xi)\bar{h}$$

$$\Rightarrow \varphi_{\bar{h}}(\bar{x}) - f'(\bar{x}) = \frac{1}{2} f''(\xi)\bar{h} = O(\bar{h})$$

o grande

ERRORE DI CONVERGENZA

Si parla quando si usano metodi iterativi: si parte da x_0;

$x_0 \to x_1 \to x_2 \to \dots \to x_n$ = SUCCESSIONE DELLE APPROSSIMAZIONI DELLA SOLUZIONE DEL PROBLEMA

Se il metodo è buono, la successione converge alla soluzione esatta.

$$\lim_{n \to \infty} x_n = x^* \quad \text{con } x^* = \text{soluzione esatta}$$

Mi serve un criterio di arresto per i metodi iterativi. Se mi fermo dopo aver calcolato x_n e lo prendo come approssimazione di x^*:

$$\Delta x = x_n - x^* = \text{ERRORE ASSOLUTO DI CONVERGENZA}$$

ESEMPIO

Approssimazione del calcolo di $\sqrt{2}$

$$x_0 = 2, \quad x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right) \quad n = 0, 1, 2 \dots$$

$x_0 = 2$

$x_1 = \frac{1}{2}\left(2 + \frac{2}{2}\right) = \frac{3}{2}$

$x_2 = \frac{1}{2}\left(\frac{3}{2} + \frac{2}{3}\cdot 2\right)$

$\lim_{n\to\infty} x_n = \sqrt{2}$

ma non posso andare avanti all'infinito; se mi fermo a x_4 ho $\Delta x = 10^{-2}$ (Piccolo)

3) ERRORI DI ROUND-OFF

Dovuti all'utilizzo del calcolatore elettronico. Si dividono in:

3.1) ERRORI DI RAPPRESENTAZIONE

3.2) ERRORI INTRODOTTI DURANTE IL CALCOLO

Prima di parlare degli errori di rappresentazione è utile capire la classificazione dei numeri:

a) Numeri naturali

b) Numeri interi

c) Numeri reali

a) NUMERI NATURALI ($\mathbb{N}_b$)

Supponiamo di lavorare con una base generica $b > 1$. Ho a disposizione b cifre che mi rappresentano i primi b numeri naturali. Ho N cifre ($= a_1, a_2, \dots a_N$) e $a_i \in \{0, 1, 2, \dots b-1\}$. Il numero naturale sarà: $n = a_1 b^{N-1} + a_2 b^{N-2} + \dots + a_N b^0$.

Con N cifre possiamo rappresentare senza errori i seguenti numeri naturali: $\{0, 1, 2, \dots b^N - 1\}$

I calcolatori usano $b = 2$.

ESEMPIO

$b = 10$, $N = 3$, $a_1 = 8$, $a_2 = 1$, $a_3 = 8$

$n = 8 \times 10^{(3-1)} + 1 \times 10^{(2-1)} + 8 \times 10^{(1-1)} = 800 + 10 + 8 = 818$

b) NUMERI INTERI

Posso avere 2 diverse rappresentazioni:

- RAPPRESENTAZIONE DI "TIPO" COMPLEMENTO A DUE

Date N cifre $(a_0, a_1, a_2, \dots a_N)$, con $a_0 \in \{+, -\}$ e $a_i \in \{0, 1, \dots b-1\}$

$(a_0 a_1 \dots a_N)$ $\quad n = \begin{cases} (a_1 b^{N-1} + a_2 b^{N-2} + \dots a_N b^0) & \text{se } a_0 = + \\ (a_1 b^{N-1} + a_2 b^{N-2} + \dots a_N b^0) - b^N & \text{se } a_0 = - \end{cases}$

Analizziamo il caso in cui $a_0 = -$: $\underbrace{(a_1 b^{N-1} + \dots + a_N b^0)}_{[0,\, b^N - 1]} - \underbrace{b^N}_{[-1,\, -b^N]}$

Posso rappresentare senza errori i seguenti interi:

$\{-b^N, \dots, -1, 0, \dots b^N - 1\}$

· RAPPRESENTAZIONE CON SHIFT

Fissiamo $b \geq 1$, $b \in \mathbb{N}$, consideriamo N cifre $(\alpha_1 \dots \alpha_N)$ con $\alpha_i \in \{0, 1, \dots b-1\}$ e ν = sRift, $\nu \in \mathbb{N}$:

$$n = (\alpha_1 b^{N-1} + \alpha_2 b^{N-2} + \dots + \alpha_N b^0) - \nu$$

Rappresentiamo senza errori i numeri: $\{-\nu, \dots, 0, 1, \dots b^N - 1 - \nu\}$

③ NUMERI REALI

Base qualunque $b \geq 1$, $b \in \mathbb{N}$ $\quad x \in \mathbb{R} \setminus \{0\}$

$x = \rho\, b^{\eta}$ $\quad \rho$ = MANTISSA $\quad 1 \leq |\rho| \leq b-1$

η = ESPONENTE $\quad \eta \in \mathbb{Z}$

Dati $\alpha_0, \dots \alpha_m$ $\quad \alpha_0 \in \{+, -\}$ $\quad \alpha_1, \beta_j \in \{0, 1, \dots b-1\}$

$\beta_0, \dots, \beta_s$

Uso α per la mantissa e β per l'esponente

$$\rho = \alpha_0 (\alpha_1 b^0 + \alpha_2 b^{-1} + \dots + \alpha_m b^{1-m})$$

$$\eta = (\beta_1 b^{s-1} + \dots + \beta_s b^0) - \nu$$

b = base

m = n° di α

s = n° di β

ν = sRift

INSIEME DEI NUMERI MACCHINA

$$M = \{0\} \cup \left\{ \rho \cdot b^{\eta} \;\middle|\; \begin{matrix} \rho = \rho(\alpha_0 \dots \alpha_m) \\ \eta = \eta(\beta_1 \dots \beta_s) \end{matrix} \right\}$$

· PROPRIETÀ DELL'INSIEME:

1) CARDINALITÀ $(M) < +\infty$

2) Se $x = \rho \cdot b^{\eta} \in M \Rightarrow$ $\quad 1 \leq |\rho| \leq b(1 - b^{-m}) < b$

$\eta \in [-\nu, b^s - \nu - 1] \cap \mathbb{Z}$

Per dimostrare la limitazione di ρ:

$\rho = \alpha_1 b^0 + \alpha_2 b^{-1} + \alpha_3 b^{-2} > \alpha_1 b^0 \geq 1 \cdot b^0 = 1$

$\Rightarrow \alpha_1 \geq 1 \Rightarrow |\rho| \geq 1$

$|\rho| \leq (b-1) b^0 + (b-1) b^{-1} + (b-1) b^{-2} = b^1 - b^0 + b^0 - b^{-1} + b^{-1} - b^{-2} =$

$= b(1 - b^{-3}) = b(1 - b^{-m}) < b$

3) Se $x \in M \setminus \{0\}$ $\quad r_1 \leq |x| \leq r_2$

$r_1 = b^{-\nu}$, $\quad r_2 = (1 - b^{-m}) b^{\varphi}$ $\quad \varphi = b^s - \nu$

Il valore più grande di $\rho = b(1 - b^{-m})$

Il valore più grande di $\eta = b^s - \nu - 1 = \varphi - 1$

$$x = \rho b^{\eta} \leq b(1 - b^{-m})\, b^{\varphi - 1} = b^{\varphi}(1 - b^{-m})$$

3.1) ERRORI DI RAPPRESENTAZIONE

$x \in [r_1, r_2]$, $\eta \in [-\nu, \varphi - 1]$ con $\varphi = b^s - \nu$

$x = \rho b^\eta = (\alpha_1, \alpha_2, ___ \alpha_m, \alpha_{m+1} \dots)_b \; b^\eta$

η lo posso rappresentare in modo esatto, ρ non è detto. Quindi il mio problema è la mantissa.

La mia approssimazione sarà per mezzo di : $fl(x) = \hat{\rho} b^\eta$

- RAPPRESENTAZIONE PER TRONCAMENTO : $\hat{\rho} = (\alpha_1, __, \alpha_m)$
- RAPPRESENTAZIONE PER ARROTONDAMENTO : $\hat{\rho} = (\alpha_1, __, \hat{\alpha}_m)$ con

$$\hat{\alpha}_m = \begin{cases} \alpha_m \text{ se } \alpha_{m+1} < \frac{b}{2} \\ \alpha_m + 1 \text{ se } \alpha_{m+1} \geq \frac{b}{2} \end{cases}$$



PROPOSIZIONE

Se $x \in I \setminus \{0\}$: $fl(x) = x(1+\varepsilon)$ con $|\varepsilon| \leq u$ → errore di rappresentazione

maggiorazione dell'errore relativo rispetto a fl = PRECISIONE DI MACCHINA

$$u = \begin{cases} b^{1-m} & \text{per troncamento} \\ \frac{1}{2} b^{1-m} & \text{per arrotondamento e } b \text{ pari} \end{cases}$$

DIMOSTRAZIONE (solo per troncamento)



$x = \rho b^\eta \qquad fl(x) = \hat{\rho} b^\eta$

$\rho = \alpha_0 \sum_{i=1}^{\infty} \alpha_i b^{1-i} \qquad \hat{\rho} = \alpha_0 \sum_{i=1}^{m} \alpha_i b^{1-i}$

$fl(x) = (1+\varepsilon)x \quad \text{con } |\varepsilon| \leq u$

$$|\varepsilon| = \left| \frac{fl(x) - x}{x} \right| = \left| \frac{\hat{\rho} b^\eta - \rho b^\eta}{\rho b^\eta} \right| = \left| \frac{\alpha_0 \hat{\rho} - \alpha_0 \rho}{\alpha_0 \rho} \right| = \left| \frac{|\hat{\rho}| - |\rho|}{|\rho|} \right| \leq \underbrace{||\hat{\rho}| - |\rho||}_{\text{è sicuramente negativo poiché } \rho \text{ ha più cifre di } \hat{\rho}} =$$

$$= |\rho| - |\hat{\rho}| = (\alpha_1 b^0 + \alpha_2 b^{-1} + _ + \alpha_m b^{1-m} + \alpha_{m+1} b^{-m} _) +$$

$$- (\alpha_1 b^0 + ___ + \alpha_m b^{1-m}) =$$

$$= \alpha_{m+1} b^{-m} + \alpha_{m+2} b^{-1-m} ___ \leq b^{-m} (\alpha_{m+1} + \alpha_{m+2} b^{-1} + _) \leq$$

$$\leq b \leq b^{1-m} = u$$



L'OVERFLOW e l'UNDERFLOW sono 2 problemi che generano grandi errori.

OVERFLOW : problema che si genera per $fl(x)$ t.c $|x| > r_2$.

a) si arresta l'esecuzione dell'errore

b) $fl(x) = \begin{cases} + \text{Inf } (+\infty) \\ - \text{Inf } (-\infty) \end{cases}$

UNDERFLOW : problema che si genera con $fl(x)$ t.c. $x \in (0, r_1)$.

a) round to zero $\Rightarrow fl(x) = 0$

b) gradual underflow : si rilassa il vincolo $\alpha_1 \neq 0$, $|\rho| < 1$