lezione 29/9/2016 - Calcolo Numerico

prova orale con elaborato, ricevimento su appuntamento (cecilia.magherini@unipi.it)

definisce e studia metodi per determinare un'approssimazione della soluzione di problemi matematici

principali proprietà che si richiedono ad un metodo numerico:

- fornire un'approssimazione accurata
- avere un basso costo computazionale
- essere facilmente implementabile su un calcolatore

Misure dell'errore

soluzione esatta $x \in \mathbb{R}$

soluzione approssimata $\tilde{x} \in \mathbb{R}$

$\Delta x = \tilde{x} - x$ errore assoluto

Supponiamo che $10^{-4} \leq \Delta x \leq 10^{-3}$: grande o piccolo? $x = 10^{8} \rightarrow$ piccolo, $x = 10^{-8} \rightarrow$ grande

errore relativo $\frac{\Delta x}{x} = \frac{\tilde{x} - x}{x}$ con $x \neq 0$ $\quad \frac{\tilde{x}}{x} = 1 + \varepsilon_x$

ε_x deve essere confrontato con 1 per conoscere la sua entità

$\varepsilon_x = 10^{-3}$ $(0{,}01\,\%)$ $\rightarrow$ piccolo

$\varepsilon_x = 0{,}9$ $(90\,\%)$ $\rightarrow$ grande

Tipologie di errore:

- errori di discretizzazione o troncamento
- errori di convergenza
- errori di round-off

errori legati alla definizione del metodo numerico

problema esatto, dominio continuo $\longrightarrow$ problema discreto, dominio discreto

esempio: $f: \mathbb{R} \rightarrow \mathbb{R}$ $\quad x_0 \in \mathbb{R}$ $\quad$ voglio approssimare $f'(x_0)$

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h}$$

($\lim_{h \to 0}$: su un dominio continuo)

Scelgo un $\bar{h} \neq 0$ di modulo piccolo $\quad f'(x_0) \simeq \frac{f(x_0+\bar{h}) - f(x_0)}{\bar{h}} = \varphi_{\bar{h}}(x_0)$

$f(x_0+\bar{h}) = f(x_0) + f'(x_0)(x_0+\bar{h}-x_0) + \frac{1}{2}f''(\xi)\bar{h}^2$ ξ compreso tra x_0 e $x_0+\bar{h}$

$\varphi_{\bar{h}}(x_0) - f'(x_0)$ [errore assoluto] $= \frac{1}{2}f''(\xi)\bar{h} = O(\bar{h})$ $\left[\frac{g(\bar{h})}{\bar{h}} \leq C\right]$

errori di convergenza

(errori causati dall'utilizzo di un metodo iterativo)

$\{x_n\}_{n\in\mathbb{N}}$ $\lim_{n\to+\infty} x_n = x^*$ soluzione esatta del problema

numero finito di iterazioni N $x_N - x^*$ errore assoluto di convergenza

$\begin{cases} x_0 = 2 \\ x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right) \end{cases}$ $\lim_{n\to+\infty} x_n = \sqrt{2}$

devo stabilire un criterio di arresto : $n = 4$ $x_4 - \sqrt{2} \simeq 10^{-12}$

errori di round-off

(errori causati dall'utilizzo di un calcolatore)

- errori di rappresentazione → numeri interi, numeri reali
- errori introdotti durante il calcolo

numeri naturali

sistema posizionale in base $b = 10$ (diverso valore in base alla posizione)

N cifre $\alpha_1\alpha_2\dots\alpha_N$ $\alpha_i \in \{0,1,2,\dots,9\}$ $i = 1,\dots,N$

$n = \alpha_1\cdot 10^{N-1} + \alpha_2\cdot 10^{N-2} + \dots + \alpha_N\cdot 10^0 \longleftrightarrow \alpha_1\alpha_2\dots\alpha_N$

N cifre => rappresentazione esatta di tutti i naturali dell'insieme $\{0,1,\dots,10^N-1\}$

sistema posizionale in base $b \in \mathbb{N}$ con $b \geq 2$

$0,1,\dots,b-1$ corrispondono ai primi b naturali esempio $b = 2$ insieme $\{0,1\}$

$\alpha_1\alpha_2\dots\alpha_N \longleftrightarrow n = \alpha_1\cdot b^{N-1} + \alpha_2 b^{N-2} + \dots + \alpha_N b^0$

gli interi rappresentabili appartengono all'insieme $\{0,1,\dots,b^N-1\}$

rappresentazione "tipo" complemento a due

$\alpha_0\alpha_1\dots\alpha_N$ $\alpha_i \in \{0,1,\dots,b-1\}$ $i = 1\dots N$ $\alpha_0 \in \{+,-\}$

$\updownarrow$

$n = \alpha_0(\alpha_1 b^{N-1} + \alpha_2 b^{N-2} + \dots + \alpha_N b^0)$ in questo caso verrebbe rappresentato 0 con segno

master COPY COPISTERIA 050/831125 380/9837745

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

quindi $0 \le n \le b^N - 1$ se $a_0 = +$

$-b^N \le n \le -1$ se $a_0 = -$

memorizzazione estesa a $\{-b^N, \dots, -1, 0, 1, \dots, b^N - 1\}$ senza errori

---

rappresentazione con shift

fisso $r \in \mathbb{N}$ → parametro detto shift

$a_1 a_2 \dots a_N$ $a_i \in \{0, 1, \dots, b-1\}$ $i = 1, \dots, N$

$a_1 \dots a_N \longleftrightarrow n = (a_1 \cdot b^{N-1} + \dots + a_N b^0) - r$

$\{-r, -r+1, \dots, b^N - 1 - r\}$

Numeri reali

$b = 10$ $x \in \mathbb{R}$, $x \neq 0$ $x = \xi \cdot 10^{\eta}$ (η: esponente) $|\xi| \in [1, 10)$ $\eta \in \mathbb{Z}$

(ξ: mantissa)

esempio $x = 78{,}54 = 7{,}854 \cdot 10^1$ $\xi = 7{,}854$ $\eta = 1$

questo metodo è noto come notazione scientifica o in virgola mobile (floating point) normalizzata (poiché ho imposto il vincolo $|\xi| \in [1, 10)$)

$7{,}854 = 7 \cdot 10^0 + 8 \cdot 10^{-1} + 5 \cdot 10^{-2} + 4 \cdot 10^{-3}$

base $b \in \mathbb{N}$ con $b \ge 2$

se $x \neq 0$ $x = \xi \cdot b^{\eta}$ $|\xi| \in [1, b)$ $\eta \in \mathbb{Z}$

$a_0 \dots a_m$, $\beta_1 \dots \beta_s$ $a_0 \in \{+, -\}$ $a_i, \beta_j \in \{0, 1 \dots, b-1\}$ $a_1 \neq 0$ $i = 1 \dots m$, $j = 1 \dots s$

$\xi = a_0 (a_1 \cdot b^0 + a_2 \cdot b^{-1} + \dots + a_m \cdot b^{1-m})$

$\eta = (\beta_1 b^{s-1} + \beta_2 b^{s-2} + \dots + \beta_s \cdot b^0) - r$ con r shift → η funzione di $\beta_1, \beta_2 \dots \beta_s$

master cozy COPISTERIA 070/8312126 388/9837745

definiamo quindi l'insieme dei numeri macchina

$$M = \{0\} \cup \{x = f \cdot b^{n} \text{ con } f = f(\alpha_0, \alpha_1, \dots, \alpha_m)\ n = n(\beta_1, \beta_2, \dots \beta_s)\}$$

proprietà

* la cardinalità di M è finita
* $1 \le |f| \le b(1 - b^{-m}) < b \quad m \ge 1$
* $n \in [-r, b^s - r - 1] \cap \mathbb{Z}$
* Se $x \in M \setminus \{0\}$ allora $\underbrace{b^{-r}}_{R_{min}} \le |x| \le \underbrace{(1 - b^{-m}) b^{\varphi}}_{R_{max}} \quad \varphi = b^s - r$

$$|f| = \alpha_1 b^0 + \alpha_2 b^{-1} + \dots + \alpha_m b^{1-m} \ge \alpha_1 \ge 1 \quad \text{(vincolo di normalizzazione)}$$

$m = 3$

$$|f| = \alpha_1 b^0 + \alpha_2 b^{-1} + \alpha_3 b^{-2} \le (b-1)b^0 + (b-1) \cdot b^{-1} + (b-1)b^{-2}$$

$$b - 1 + 1 - b^{-1} + b^{-1} - b^{-2} = b(1 - b^{-3})$$

$$\Downarrow$$

$$b(1 - b^{-m}) \text{ con } m \ge 1$$

$|f| = 1 \quad n = -r$

$R_{min} = b^{-r}$

$R_{max} = b(1 - b^{-m}) \cdot b^{b^s - 1 - r}$

$|f| = b(1 - b^{-m}) \quad n = b^s - 1 - r$

lo shift è scelto in modo tale che $R_{min}^{-1} \approx R_{max}$

$$(b^{-r})^{-1} \approx (1 - b^{-m}) b^{b^s - r}$$

$$b^r \approx b^{b^s - r} \longrightarrow r \approx b^s - r \longrightarrow r \approx \frac{b^s}{2}$$

criterio per valutare la possibilità di rappresentare il reciproco di un numero

$$M \subset I = [-R_{max}, -R_{min}] \cup \{0\} \cup [R_{min}, R_{max}]$$

$\subset$ senso stretto ($\text{card} M < +\infty$)



$x \in I$ potrebbe non essere rappresentabile in maniera esatta

$fl : I \to M \quad fl(x) - x$ errore di rappresentazione di x dovuto all'uso di un calcolatore

master COPY COPISERIA 050/8312126 388/9857745

lezione 3/10/2016 - Calcolo Numerico

$fl : I \to M$

condizioni

$fl(0) = 0$

$fl(x) = -fl(-x) \qquad x \in [-\Omega_{max}, -\Omega_{min}]$

$x = \underbrace{(\alpha_1, \alpha_2, \dots, \alpha_m \dots)_b}_{\alpha_1 b^0 + \alpha_2 b^{-1} + \alpha_3 b^{-2} + \dots} \cdot b^q$ $\qquad q \in [-\gamma, b^s - 1 - \gamma] \cap \mathbb{Z}$

rappresentabile in maniera esatta

$fl(x) = \hat{\rho} \cdot b^q \qquad \hat{\rho} \approx \rho$

- rappresentazione con troncamento $\hat{\rho} = (\alpha_1 \dots \alpha_m)_b$ [m cifre selezionate]
- rappresentazione con arrotondamento $\hat{\rho} = (\alpha_1 \dots \hat{\alpha}_m)_b \qquad \hat{\alpha}_m = \begin{cases} \alpha_m \text{ se } \alpha_{m+1} < b/2 \\ 1 + \alpha_m \text{ se } \alpha_{m+1} > b/2 \end{cases}$

esempio : $b = 10$ (base 10) $m = 3$ (cifre utilizzabili nella rappresentazione della mantissa) $s >> 1$

$x = 7{,}854 \cdot 10^1 \longrightarrow fl(x) = 7{,}85 \cdot 10^1$ troncamento
arrotondamento $(\alpha_4 < b/2 = 5)$

$x = 7{,}856 \cdot 10^1 \longrightarrow fl(x) = 7{,}85 \cdot 10^1$ troncamento
$fl(x) = 7{,}86 \cdot 10^1$ arrotondamento $(\alpha_4 > b/2 = 5)$

Teorema

Se $x \in I \setminus \{0\}$ allora $fl(x) = x(1 + \varepsilon_x)$ con $|\varepsilon_x| \le u = \begin{cases} b^{1-m} & \text{troncamento} \\ \frac{1}{2} b^{1-m} & \text{arrotondamento con } b \text{ pari} \end{cases}$

$|\varepsilon_x|$: errore relativo di rappresentazione; u: precisione di macchina

Dimostrazione troncamento

$x = \rho \cdot b^q \qquad fl(x) = \hat{\rho} \cdot b^q \qquad \varepsilon_x = \dfrac{fl(x) - x}{x} = \dfrac{\hat{\rho} \cdot b^q - \rho \cdot b^q}{\rho \cdot b^q} = \dfrac{|\hat{\rho} - \rho|}{|\rho|} \le |\hat{\rho} - \rho|$

$\frac{1}{|\rho|} \le 1$

$|\hat{\rho} - \rho| = |\rho| - |\hat{\rho}| = (\alpha_1 \dots \alpha_m \alpha_{m+1} \dots)_b - (\alpha_1 \dots \alpha_m)_b = (0,0,0 \dots \alpha_{m+1} \alpha_{m+2} \dots)_b$

troncamento

$= (\alpha_{m+1} b^{-m} + \alpha_{m+2} b^{-m-1} \dots)$

$= b^{-m} \underbrace{[\alpha_{m+1} + \alpha_{m+2} b^{-1} \dots]}_{\le b} \le b^{1-m} = u$