

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) Il polinomio interpolatore di $f(x) = x^2 + bx + c$ su 20 nodi distinti:

- A) ha grado 3
- B) ha grado 20
- C) ha grado 2
- D) ha grado 19

2) La somma algebrica di numeri approssimati:

- A) è sempre instabile
- B) è stabile quando i numeri hanno segno opposto
- C) è instabile quando i numeri hanno lo stesso segno
- D) può essere instabile quando i numeri hanno segno opposto

3) Il metodo di Newton (tangenti) quando converge:

- A) ha sempre convergenza quadratica
- B) ha sempre convergenza lineare
- C) può avere convergenza lineare
- D) può avere convergenza cubica

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) La moltiplicazione tra numeri approssimati:

- A) è sempre instabile
- B) è instabile quando i numeri hanno lo stesso segno
- C) è sempre stabile
- D) è instabile quando i numeri hanno segno opposto

2) L'interpolazione cubica a tratti a passo costante h :

- A) converge uniformemente con errore $O(h^4)$ per $f \in C^5[a, b]$
- B) non converge uniformemente se $f \in C^k[a, b]$ con $k < 5$
- C) converge uniformemente con errore $O(h^5)$ per $f \in C^3[a, b]$
- D) converge uniformemente con errore $O(h^4)$ per $f \in C^2[a, b]$

3) La precisione di macchina in un sistema floating-point $F(b, t, L, U)$ è:

- A) il più piccolo reale-macchina positivo
- B) $b^{L-t}/2$
- C) il massimo errore relativo di arrotondamento a t cifre di mantissa
- D) b^{L-U}

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) La formula di derivazione numerica col rapporto incrementale simmetrico $\delta(h)$ per $f \in C^5$ ha un errore teorico:

- A) $O(h^4)$
- B) $O(h^3)$
- C) $O(h^2)$
- D) $O(h^5)$

2) Il costo computazionale del Metodo di Eliminazione Gaussiana applicato a una matrice invertibile é:

- A) $\sim 5n^3/4$
- B) $O(n^3)$
- C) $O(n^2)$
- D) $\sim 2n^3/3$

3) Il metodo di Newton (tangenti) quando converge:

- A) può avere convergenza lineare
- B) ha sempre convergenza lineare
- C) può avere convergenza quadratica
- D) ha sempre convergenza quadratica

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) Il prodotto di numeri approssimati:

- A) è sempre stabile
- B) è instabile quando i numeri hanno segno opposto
- C) è sempre instabile
- D) è instabile quando i numeri hanno lo stesso segno

2) L'interpolazione lineare a tratti a passo costante

- A) converge uniformemente con errore $O(h^4)$ per $f \in C^5[a, b]$
- B) non converge uniformemente se $f \in C^k[a, b]$ con $k < 4$
- C) converge uniformemente con errore $O(h^2)$ per $f \in C^2[a, b]$
- D) converge uniformemente con errore $O(h^2)$ per $f \in C^3[a, b]$

3) Il polinomio interpolatore di $f(x) = x^3 + bx + c$ su 29 nodi distinti:

- A) ha grado 30
- B) ha grado 3
- C) ha grado ≤ 28
- D) ha grado 4

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) In un sistema floating-point $F(b, t, L, U)$ il più piccolo reale-macchina positivo è:

A) la precisione di macchina

B) b^{-U}

C) b^{L-1}

D) b^{L-U}

2) Il costo computazionale del Metodo di Eliminazione Gaussiana applicato a una matrice invertibile è:

A) $\sim 2n^4/3$

B) $\sim 2n^3/3$

C) $O(n^2)$

D) $\sim n^3$

3) L'interpolazione spline cubica a passo costante h per $f \in C^4[a, b]$ ha un errore:

A) $O(h^5)$ su f

B) $O(h^3)$ su f'

C) $O(h^3)$ su f''

D) $O(h^4)$ su f

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) La divisione tra numeri approssimati:

- A) è sempre stabile
- B) può essere instabile
- C) è instabile se i numeri hanno segno opposto
- D) è stabile se i numeri hanno lo stesso segno

2) L'interpolazione spline cubica a passo costante:

- A) converge uniformemente con errore $O(h^5)$ per $f \in C^5[a, b]$
- B) converge uniformemente con errore $O(h^4)$ per $f \in C^4[a, b]$
- C) converge uniformemente con errore $O(h^4)$ per $f \in C^6[a, b]$
- D) non converge mai uniformemente

3) In un sistema floating-point $F(b, t, L, U)$ il più piccolo reale-macchina positivo è:

- A) b^{L-U}
- B) $b^{1-t}/2$
- C) b^{-U}
- D) b^{L-1}

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) La divisione tra numeri approssimati:

- A) è sempre stabile
- B) è instabile quando i numeri hanno lo stesso segno
- C) può essere instabile quando i numeri hanno segno opposto
- D) è sempre instabile

2) L'interpolazione quadratica a tratti a passo costante

- A) converge uniformemente con errore $O(h^4)$ per $f \in C^5[a, b]$
- B) non converge uniformemente se $f \in C^k[a, b]$ con $k < 6$
- C) converge uniformemente con errore $O(h^3)$ per $f \in C^5[a, b]$
- D) converge uniformemente con errore $O(h^3)$ per $f \in C^3[a, b]$

3) La precisione di macchina in un sistema floating-point $F(b, t, L, U)$ è:

- A) il più piccolo reale-macchina positivo
- B) $b^{1-t}/2$
- C) il minimo reale-macchina positivo che sommato ad 1 dà un risultato > 1
- D) b^{L-1}

TEST A RISPOSTA MULTIPLA

Indicare TUTTE le affermazioni corrette

Risposte tipo: 1AD - 2A - 3BC
(ci possono essere più risposte corrette)

1) L'indice di condizionamento di una matrice invertibile $A \in \mathbb{R}^{n \times n}$ è:

- A) $\det(A)$
- B) l'autovalore di modulo massimo di A
- C) l'autovalore di modulo minimo di A
- D) $\|A\| \|A^{-1}\|$

2) Il costo computazionale del Metodo di Eliminazione Gaussiana applicato a una matrice invertibile é:

- A) $O(n^3)$
- B) $\sim n^3$
- C) $O(n^2)$
- D) $\sim n^4$

3) Le iterazioni di punto fisso per una contrazione:

- A) hanno sempre convergenza quadratica
- B) possono avere convergenza quadratica
- C) possono non convergere
- D) hanno sempre convergenza almeno lineare

Dalle dispense del
PROF. MARCO VIANELLO

Syllabus

Dimostrazioni Irrinunciabili

Università degli Studi di Padova

Dipartimento di Matematica

CORSO DI LAUREA IN INFORMATICA

Anno accademico 2020 - 2021

Autori:

MARKO

TOLDO DM

Premessa

Questa raccolta di appunti non intende essere un sostituto allo studio completo degli argomenti di calcolo numerico. Gli appunti sono stati scritti secondo quanto studiato e capito, di conseguenza potrebbe contenere errori/non essere esaustivo nella risposta agli argomenti del syllabus.

La repository contenente i sorgenti di questo documento si puo' trovare qui: <https://github.com/ToldoDM/SyllabusCalcolo>

Indice

1 Precisione di macchina come max errore relativo di arrotondamento nel sistema floating-point	3
2 Analisi di stabilità di moltiplicazione, divisione, addizione e sottrazione con numeri approssimati	4
2.1 Moltiplicazione	4
2.2 Divisione	4
2.3 Somma Algebrica	5
3 Convergenza del metodo di bisezione	6
4 Stima dell'errore con residuo pesato (metodo bisezione)	7
5 Convergenza globale del metodo di Newton ("delle tangenti") in ipotesi di convessità/concavità stretta	8
6 Velocità (ordine) di convergenza del metodo di Newton	10
7 Ordine di convergenza delle iterazioni di punto fisso	11
8 Esistenza e unicità dell'interpolazione polinomiale	12
9 Convergenza uniforme dell'interpolazione lineare a tratti	13
10 Stima delle equazioni normali per l'approssimazione polinomiale ai minimi quadrati	14
11 Stime di condizionamento per un sistema lineare	16

1 Precisione di macchina come max errore relativo di arrotondamento nel sistema floating-point

Definiamo arrotondamento a t cifre di un numero reale scritto in notazione floating-point

$$x = \text{sign}(x)(0, d_1 d_2 \dots d_t \dots) \cdot b^p$$

il numero

$$fl^t(x) = \text{sign}(x)(0, d_1 d_2 \dots \tilde{d}_t) \cdot b^p$$

dove la mantissa è stata arrotondata alla t -esima cifra

$$\tilde{d}_t = \begin{cases} d_t & \text{se } d_{(t+1)} < \frac{b}{2} \\ d_t + 1 & \text{se } d_{(t+1)} \geq \frac{b}{2} \end{cases}$$

Definiamo:

$$\text{Errore Relativo} \leftarrow \frac{\overbrace{|x - fl^t(x)|}^{\text{Errore Assoluto}}}{|x|} \quad \text{per } x \neq 0$$

Stimiamo il numeratore

$$\begin{aligned} |x - fl^t(x)| &= b^p \cdot \overbrace{|(0, d_1 d_2 \dots d_t \dots) - (0, d_1 d_2 \dots \tilde{d}_t)|}^{\text{Errore di arrotondamento a } t \text{ cifre dopo la virgola} \leq \frac{b^{-t}}{2}} \\ &\leq b^p \cdot \frac{b^{-t}}{2} = \frac{b^{p-t}}{2} \end{aligned}$$

Notiamo subito un aspetto: l'errore dipende da p , cioè dall'ordine di grandezza del numero (in base b).

Stimiamo da sopra $\frac{1}{|x|}$, ovvero da sotto $|x|$:

$$|x| = (0, d_1 d_2 \dots d_t \dots) \cdot b^p$$

Poiché $d_1 \neq 0$, p fissato, il minimo valore della mantissa è $0,100\dots = b^{-1}$. Quindi:

$$|x| \geq b^{-1} \cdot b^p = b^{p-1} \iff \frac{1}{|x|} \leq \frac{1}{b^{p-1}}$$

Otteniamo

$$\frac{|x - fl^t(x)|}{|x|} \leq \frac{\frac{b^{p-t}}{2}}{b^{p-1}} = \frac{b^{p-t+1-p}}{2} = \frac{b^{1-t}}{2} = \varepsilon_M$$

2 Analisi di stabilità di moltiplicazione, divisione, addizione e sottrazione con numeri approssimati

2.1 Moltiplicazione

$$\varepsilon_{xy} = \frac{|xy - \tilde{x}\tilde{y}|}{|xy|}, \quad x, y \neq 0$$

Usiamo la stessa tecnica che si usa per dimostrare che il limite del prodotto di due successioni o funzioni è il prodotto dei limiti, aggiungendo e togliendo a numeratore ad esempio $\tilde{x}y$

$$\begin{aligned} \varepsilon_{xy} &= \frac{|xy - \tilde{x}y + \tilde{x}y - \tilde{x}\tilde{y}|}{|y|} \\ &= \frac{\overbrace{|y(x - \tilde{x})|}^{=a} + \overbrace{|\tilde{x}(y - \tilde{y})|}^{=b}}{|xy|} \\ &\leq \frac{|y(x - \tilde{x})| + |\tilde{x}(y - \tilde{y})|}{|xy|} \quad (*) \end{aligned}$$

(*) Disuguaglianza triangolare: $||a| - |b|| \leq |a + b| \leq |a| + |b|$

Quindi otteniamo

$$\varepsilon_{xy} \leq \frac{|y||x - \tilde{x}|}{|xy|} + \frac{|\tilde{x}||y - \tilde{y}|}{|xy|} = \varepsilon_x + \frac{|\tilde{x}|}{|x|} \varepsilon_y$$

Questo perché $\frac{|x - \tilde{x}|}{|x|} = \varepsilon_x$ e $\frac{|y - \tilde{y}|}{|y|} = \varepsilon_y$.

Poiché $\tilde{x} \approx x \Rightarrow \frac{|\tilde{x}|}{|x|} \approx 1$ e possiamo quindi dire che la moltiplicazione è STABILE.

$$\varepsilon_{xy} \lesssim \varepsilon_x + \varepsilon_y$$

Però possiamo dare una stima più precisa di $\frac{|\tilde{x}|}{|x|}$

$$\frac{|\tilde{x}|}{|x|} = \frac{\overbrace{|x|}^{=a} + \overbrace{|\tilde{x} - x|}^{=b}}{|x|} \leq \frac{|x| + |\tilde{x} - x|}{|x|} = 1 + \varepsilon_x$$

Disuguaglianza Triangolare

e quindi

$$\varepsilon_{xy} \leq \varepsilon_x + (1 + \varepsilon_x) \varepsilon_y$$

Solitamente $\varepsilon_x \leq \varepsilon_M \approx 10^{-16} \Rightarrow 1 + \varepsilon_x$ è vicinissimo ad 1. Ma anche se $\varepsilon_x = 1$ (errore del 100%, molto grande) $\Rightarrow (1 + \varepsilon_x) = 2$ e la stabilità della moltiplicazione non cambia.

2.2 Divisione

La divisione è la moltiplicazione per il reciproco $\frac{x}{y} = x \cdot \frac{1}{y}$.
Analizzando quindi l'operazione di reciproco

$$\varepsilon_{\frac{1}{y}} = \frac{\left| \frac{1}{y} - \frac{1}{\tilde{y}} \right|}{\left| \frac{1}{y} \right|} = \frac{\frac{|\tilde{y} - y|}{|\tilde{y}y|}}{\frac{1}{|y|}} = \frac{|\tilde{y} - y|}{|y|} \cdot \frac{|y|}{|\tilde{y}|} \approx \varepsilon_y \quad \left(\text{questo perchè } \frac{|\tilde{y} - y|}{|y|} = \varepsilon_y \right)$$

Poiché $\frac{|y|}{|\tilde{y}|} \approx 1$ possiamo dedurre che il reciproco, e possiamo quindi la divisione, è STABILE.
 Però possiamo dare una stima più precisa di $\frac{|y|}{|\tilde{y}|}$

$$|\tilde{y}| = |y + \tilde{y} - y| = |y| \left| 1 + \frac{(\tilde{y} - y)}{y} \right|$$

usando la stima da sotto nella disuguaglianza triangolare

$$|a + b| \geq ||a| - |b||$$

$$a = 1 \text{ e } b = \frac{(\tilde{y} - y)}{y}$$

$$\left| 1 + \frac{(\tilde{y} - y)}{y} \right| \geq \left| 1 - \frac{|\tilde{y} - y|}{|y|} \right| = |1 - \varepsilon_y| = 1 - \varepsilon_y \quad \left(\text{perché } \varepsilon_y < 1 \right)$$

da cui si ottiene

$$|\tilde{y}| \geq |y|(1 - \varepsilon_y)$$

e quindi

$$\frac{|y|}{|\tilde{y}|} \leq \frac{|y|}{|y|(1 - \varepsilon_y)} = \frac{1 + \varepsilon_y}{(1 + \varepsilon_y)(1 - \varepsilon_y)} = \frac{1 + \varepsilon_y}{1 - \varepsilon_y^2} \approx 1 + \varepsilon_y$$

Poiché $\varepsilon_y^2 \ll \varepsilon_y < 1$

Quindi

$$\varepsilon_{\frac{1}{y}} = \varepsilon_y \frac{|y|}{|\tilde{y}|} \lesssim \varepsilon_y(1 + \varepsilon_y) \approx \varepsilon_y \Rightarrow \varepsilon_{\frac{1}{y}} \lesssim \varepsilon_y$$

Infine abbiamo che per la divisione vale (usando la stima della moltiplicazione)

$$\varepsilon_{\frac{x}{y}} \lesssim \varepsilon_x + \varepsilon_y$$

Es. instabilità sottrazione: pag. 33 appunti

2.3 Somma Algebrica

$$x + y = \begin{cases} \text{ADDIZIONE} & \text{se } \text{sign}(x) = \text{sign}(y) \\ \text{SOTTRAZIONE} & \text{se } \text{sign}(x) \neq \text{sign}(y) \end{cases}$$

Per la somma algebrica vale:

$$\begin{aligned} \varepsilon_{x+y} &= \frac{|(x + y) - (\tilde{x} + \tilde{y})|}{|x + y|}, \quad x + y \neq 0 \\ &= \frac{|x - \tilde{x} + y - \tilde{y}|}{|x + y|}, \quad a = x - \tilde{x} \text{ e } b = y - \tilde{y} \\ &\leq \frac{|x - \tilde{x}|}{|x + y|} + \frac{|y - \tilde{y}|}{|x + y|}, \quad \text{DISUGUAGLIANZA TRIANGOLARE} \\ &= \frac{|x|}{|x + y|} \cdot \frac{|x - \tilde{x}|}{|x|} + \frac{|y|}{|x + y|} \cdot \frac{|y - \tilde{y}|}{|y|} \\ &= w_1 \varepsilon_x + w_2 \varepsilon_y \quad \text{con } w_1 = \frac{|x|}{|x + y|}, w_2 = \frac{|y|}{|x + y|} \end{aligned}$$

Addizione $\text{sign}(x) = \text{sign}(y)$

In questo caso $|x + y| \geq |x|, |y| \Rightarrow w_1, w_2 \leq 1$. Quindi l'addizione è stabile $\varepsilon_{x+y} \lesssim \varepsilon_x + \varepsilon_y$

Sottrazione $\text{sign}(x) \neq \text{sign}(y)$

In questo caso $|x + y| \leq |x|$ e/o $|x + y| \leq |y| \Rightarrow \max\{w_1, w_2\} > 1$. Quindi la sottrazione è potenzialmente instabile (se w_1, w_2 troppo grandi).

Nel caso in cui $|x|, |y|$ siano molto vicini in termini relativi, si ha

$$|x + y| \ll |x|, |y| \Rightarrow w_1, w_2 \gg 1$$

3 Convergenza del metodo di bisezione

Il metodo di bisezione si basa sull'applicazione iterativa del Teorema degli zeri di funzioni continue:
Se $f(x) \in C[a, b]$ e $f(a)f(b) < 0$ (cioè f cambia segno) allora

$$\exists \xi : f(\xi) = 0, \xi \in (a, b)$$

Il procedimento consiste nel passare da $[a_n, b_n] \rightarrow [a_{n+1}, b_{n+1}]$ in cui uno degli estremi è diventato il punto medio

$$x_n = \frac{a_n + b_n}{2}$$

A meno che per qualche n non risulti $f(x_n) = 0$, si tratta di un processo infinito che ci permette di costruire tre successioni $\{a_n\}, \{b_n\}, \{x_n\}$ tali che:

- $|\xi - a_n|, |\xi - b_n| \leq b_n - a_n = \frac{b-a}{2^n}$
- $|\xi - x_n| < \frac{b_n - a_n}{2} = \frac{b-a}{2^{n+1}}$

È semplice dimostrare che tutte e tre le successioni convergono ad uno zero $\xi \in (a, b)$

- $0 \leq |\xi - a_n|, |\xi - b_n| < \frac{b-a}{2^n} \xrightarrow{n \rightarrow \infty} 0 \xRightarrow{\text{Teor. Carabinieri}} |\xi - a_n|, |\xi - b_n| \rightarrow 0, n \rightarrow \infty$
- $0 \leq |\xi - x_n| < \frac{b-a}{2^{n+1}} \implies |\xi - x_n| \rightarrow 0, n \rightarrow \infty$

4 Stima dell'errore con residuo pesato (metodo bisezione)

Vogliamo stimare l'errore di bisezione, applicato nelle seguenti ipotesi:

$$\left. \begin{array}{l} f \in C^1[a, b] \\ \{x_n\} \in [c, d] \subseteq [a, b] \\ f'(x) \neq 0, \forall x \in [c, d] \end{array} \right\} \Rightarrow e_n = |x_n - \xi| = \frac{|f(x_n)|}{|f'(z_n)|}, \quad n \geq n_0, \quad z_n \in \begin{cases} (x_n, \xi) \\ (\xi, x_n) \end{cases}$$

C^1 indica derivabile 1 volta con derivata continua.

Dimostriamolo utilizzando il teorema del valor medio

$$\text{Sia } f \in C[a, b] \text{ derivabile in } [a, b] \Rightarrow \exists z \in [a, b] : \frac{f(b) - f(a)}{b - a} = f'(z)$$

Consideriamo il caso $\xi < x_n$ (se $x_n < \xi$ la dimostrazione è analoga)

$$f(x_n) - f(\xi) = f'(z_n)(x_n - \xi), \quad z_n \in (\xi, x_n)$$

con $f(\xi) = 0$, cioè

$$|f(x_n)| = |f'(z_n)| |x_n - \xi|$$

che si può riscrivere come

$$e_n = |x_n - \xi| = \frac{|f(x_n)|}{|f'(z_n)|}$$

Osserviamo che:

- e_n è un "residuo pesato"
- $f'(x) \neq 0 \Rightarrow$ zero è semplice
- e_n è una stima a posteriori (serve aver calcolato x_n)

Siccome non conosciamo z_n , diamo delle stime pratiche dell'errore:

- Se è noto che $|f'(x)| \geq k > 0 \Rightarrow e_n = \frac{|f(x_n)|}{|f'(z_n)|} \leq \frac{|f(x_n)|}{k}$
- Se f' è nota, per n abbastanza grande si ha

$$\underbrace{f'(x_n) \approx f'(z_n)}_{\approx f'(\xi)} \Rightarrow e_n \approx \frac{|f(x_n)|}{|f'(z_n)|}$$

- Se f' non è nota, si può approssimare con

$$f'(z_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}, \quad \text{per } n \text{ abbastanza grande}$$

5 Convergenza globale del metodo di Newton ("delle tangenti") in ipotesi di convessità/concavità stretta

Metodo di Newton: Linearizzare iterativamente la funzione con la tangente nel punto

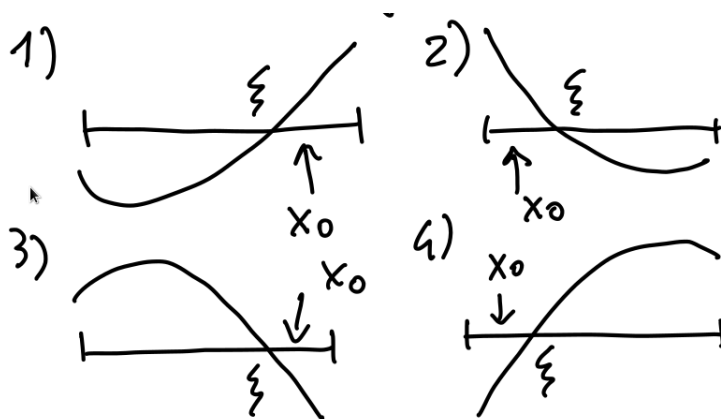
$$\begin{cases} y = 0 \\ y = f(x_n) + f'(x_n)(x - x_n) \end{cases} \Rightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Convergenza metodo di Newton:

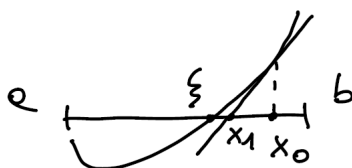
$$\begin{cases} f \in C^2[a, b] \\ f(a)f(b) < 0 \\ f''(x) > 0 \quad \forall x \in [a, b] \\ x_0 : f(x_0)f''(x_0) > 0 \end{cases} \Rightarrow \text{Il metodo di Newton è ben definito (cioè } f'(x_n) \neq 0) \\ \text{e converge all'unico zero } \xi \text{ di } f \text{ in } [a, b]$$

Dimostrazione

ci sono 4 casi possibili in base al segno di f'' ovvero



In questa dimostrazione di concentriamo sul caso 1)



- $f(a) < 0, f(b) > 0$
- $f''(x) > 0 \quad \forall x \in [a, b]$
- $x_0 \in [a, b]$

Dimostriamo come prima cosa: $x_n \in (\xi, b] \Rightarrow x_{n+1} \in (\xi, b]$

f è esattamente convessa \Rightarrow La tangente sta "sotto al grafico" $\forall x \in [a, b]$
 \Rightarrow La tangente in un punto $\in (\xi, b]$ interseca l'asse x "a destra" di ξ

Dimostriamo quindi: $x_{n+1} < x_n$ (cioè $\{x_n\}$ è decrescente)

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \Big\} > 0$$

Poiché per $x_n \in (\xi, b]$ si ha $f(x_n) > 0$. Inoltre $f'(x_n) > 0$ in $(\xi, b]$ altrimenti per avere uno zero f'' in $(\xi, b]$ dovrebbe cambiare segno.

Abbiamo quindi che $\{x_n\}$ è una successione decrescente, con $x_n > \xi \quad \forall n$.

Allora

$$\exists \lim_{n \rightarrow \infty} x_n = \inf\{x_n\} = \eta \quad \text{con} \quad \eta \geq \xi$$

Infine

$$\begin{aligned} \eta &= \lim x_{n+1} = \lim \left(x_n - \frac{f(x_n)}{f'(x_n)} \right) \\ &= \lim x_n - \lim \frac{f(x_n)}{f'(x_n)} \\ &= \lim x_n - \frac{\lim f(x_n)}{\lim f'(x_n)} \\ &= \lim x_n - \frac{f(\lim x_n)}{f'(\lim x_n)} \leftarrow \lim x_n = \eta \\ &= \eta - \frac{f(\eta)}{f'(\eta)} \end{aligned}$$

Quindi

$$\eta = \eta - \frac{f(\eta)}{f'(\eta)} \quad \text{con} \quad f'(\eta) \neq 0 \Rightarrow \frac{f(\eta)}{f'(\eta)} = 0 \Rightarrow f(\eta) = 0 \Rightarrow \eta = \xi$$

6 Velocità (ordine) di convergenza del metodo di Newton

Sia

$$\begin{cases} f \in C^2[a, b] \\ \xi \in [a, b] : f(\xi) = 0 \\ \{x_n\} \subset [c, d] \subseteq [a, b] \\ f'(x) \neq 0 \quad \forall x \in [c, d] \end{cases} \Rightarrow e_{n+1} \leq c e_n^2, \quad n \geq 0, \quad c = \frac{1}{2} \cdot \frac{M_2}{m_1}$$

$$\text{con } M_2 = \max_{x \in [c, d]} |f''(x)|, \quad m_1 = \min_{x \in [c, d]} |f'(x)| > 0$$

Dimostrazione

Applichiamo la formula di Taylor centrata in x_n e calcolata in ξ , con resto del II ordine in forma di Lagrange

$$\underbrace{f(\xi)}_{=0} = f(x_n) + f'(x_n)(\xi - x_n) + \frac{f''(z_n)}{2}(\xi - x_n)^2 \quad z_n \in \text{int}(x_n, \xi) \subset [c, d]$$

\Downarrow

$$\underbrace{-\frac{f(x_n)}{f'(x_n)}}_{=x_{n+1}-x_n} = \xi - x_n + \frac{f''(z_n)}{2f'(x_n)}(\xi - x_n)^2$$

\Downarrow

$$x_{n+1} = \xi + \frac{f''(z_n)}{2 \cdot f'(x_n)}(\xi - x_n)^2$$

aggiungendo i moduli

\Downarrow

$$e_{n+1} = |x_{n+1} - \xi| = c_n e_n^2 \quad \text{con} \quad c_n = \frac{1}{2} \frac{|f''(z_n)|}{|f'(x_n)|}$$

Applicando il teorema di Weierstrass (\exists di max, min assoluti in un intervallo chiuso e limitato)

$$|f''(z_n)| \leq \max_{x \in [c, d]} |f''(x)| = M_2, \quad |f'(x_n)| = \min_{x \in [c, d]} |f'(x)| > m_1$$

7 Ordine di convergenza delle iterazioni di punto fisso

Sia ξ punto fisso di $\phi \in C(I)$ e I è un intervallo chiuso (non necessariamente limitato) di \mathbb{R} . Supponiamo di essere nelle ipotesi in cui:

$$x_{n+1} = \phi(x_n) \quad \text{converge a } \xi \quad (\xi = \phi(\xi)) \quad \text{con } x_0 \in I$$

Allora:

- $\{x_n\}$ ha ordine esattamente $p = 1 \iff 0 < |\phi'(\xi)| < 1$
- $\{x_n\}$ ha ordine esattamente $p > 1 \iff \phi^{(j)}(\xi) = 0$ e $\phi^{(p)}(\xi) \neq 0$ con $1 \leq j \leq p-1$

Dimostrazione

1) si dimostra subito visto che

$$e_{n+1} = |\phi'(z_n)|e_n \quad \text{con } z_n \in (\xi, x_n)$$

\Downarrow

$$\lim_{n \rightarrow \infty} \frac{e_{n+1}}{e_n} = \left| \phi' \left(\lim_{n \rightarrow \infty} z_n \right) \right| = |\phi'(\xi)|$$

per 2) utilizziamo la formula di Taylor di grado $p-1$ centrata in ξ e calcolata in x_n , con il resto p -esimo in forma di Lagrange.

$$x_{n+1} = \phi(x_n) = \phi(\xi) + \phi'(\xi)(x_n - \xi) + \cdots + \frac{\phi^{(p-1)}(\xi)}{(p-1)!}(x_n - \xi)^{p-1} + \frac{\phi^{(p-1)}(\xi)}{(p-1)!} + \frac{\phi^{(p)}(u_n)}{p!}(x_n - \xi)^p$$

con $u_n \in (\xi, x_n)$

• Dimostriamo “ \Leftarrow ” (condizione sufficiente)

Da Taylor resta solo

$$x_{n+1} - \xi = \frac{\phi^{(p)}(u_n)}{p!}(x_n - \xi)^p$$

e passando ai moduli

$$\frac{e_{n+1}}{e_n^p} = \frac{|\phi^{(p)}(u_n)|}{p!} \xrightarrow{n \rightarrow \infty} \frac{|\phi^{(p)}(\xi)|}{p!} \neq 0$$

e_n^p ovvero per p , $\{x_n\}$ ha ordine esattamente p .

• Dimostriamo “ \Rightarrow ” (condizione necessaria)

Per ipotesi $\{x_n\}$ ha esattamente ordine $p > 1$.

Abbiamo per assurdo che $\exists j < p : \phi^{(j)}(\xi) \neq 0$, prendiamo $k = \min\{j < p : \phi^{(j)}(\xi) \neq 0\}$ e dal polinomio di Taylor iniziale si avrebbe:

$$\frac{e_{n+1}}{e_n^k} \xrightarrow{n \rightarrow \infty} \frac{|\phi^{(k)}(\xi)|}{k!} = L' \neq 0$$

ma allora

$$\frac{e_{n+1}}{e_n^p} = \frac{e_{n+1}}{e_n^k} \cdot e_n^{k-p}$$

$$\left(\frac{e_{n+1}}{e_n^k} \rightarrow L' \text{ ed } e_n^{k-p} \rightarrow \infty \text{ perchè } k-p < 0 \text{ ed } e_n \rightarrow 0 \right)$$

cioè

$$\frac{e_{n+1}}{e_n^p} \rightarrow \infty, \quad n \rightarrow \infty$$

contraddicendo l'ipotesi che $\{x_n\}$ abbia ordine esattamente p .

8 Esistenza e unicità dell'interpolazione polinomiale

Unicità

Supponiamo che \exists due polinomi $p, q \in \mathbb{P}_n$ (polinomi di grado $\leq n$), $p \neq q$, che interpolano $p(x_i) = y_i = q(x_i)$, con $0 \leq i \leq n \rightarrow n+1$ modi di interpolare.

Poiché \mathbb{P}_n è uno spazio vettoriale $\Rightarrow p - q \in \mathbb{R}_n$.

Allora:

$$(p - q)(x_i) = p(x_i) - q(x_i) = 0, \quad \forall 0 \leq i \leq n$$

$$\Downarrow \\ p - q \text{ ha } n + 1 \text{ zeri distinti}$$

Ma per il teorema fondamentale dell'algebra, $p - q$ può avere al massimo n zeri distinti, a meno che non sia il polinomio nullo

$$(p - q)(x) = 0 \quad \forall x \quad \Rightarrow \quad p(x) = q(x) \quad \forall x$$

Esistenza

Definiamo il "polinomio di Lagrange":

$$l_i(x) = \frac{N_i(x)}{N_i(x_i)}$$

dove

$$N_i(x) = \prod_{j=0, j \neq i}^n (x - x_j) = (x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)$$

$l_i(x) \in \mathbb{P}_n$ poiché $N_i(x) \in \mathbb{P}_n$ e $N_i(x_i)$ è un numero $\neq 0$.

Osserviamo che:

$$l_i(x_k) = \delta_{ik} = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}$$

Definiamo il "polinomio interpolatore di Lagrange":

$$f_n(x) = \prod_n(x) = \sum_{i=0}^n y_i l_i(x) \in \mathbb{P}_n$$

Verifichiamo che interpola

$$\begin{aligned} \prod_n(x_k) &= \sum_{i=0}^n y_i l_i(x_k) \\ &= \sum_{i=0}^n y_i \delta_{ik} \\ &= y_k \delta_{kk} \quad \leftarrow \quad \text{perchè } \delta_{ik} = 0, i \neq k \\ &= y_k, \quad 0 \leq k \leq n \end{aligned}$$

9 Convergenza uniforme dell'interpolazione lineare a tratti

Teorema

Convergenza uniforme dell'interpolazione polinomiale a tratti.

Siano $f \in C^{s+1}[a, b]$, $s \geq 0$ e $\{x_i\} \subset [a, b]$ $n + 1$ nodi distinti con n multiplo di s .

Allora

$$\exists k_s > 0 : \text{dist}(f, \prod_s^c) \leq k_s \cdot h^{s+1}, \quad h = \max \Delta x_i$$

Dimostrazione per $s = 1$.

Si ha che:

$$\begin{aligned} \text{dist}(f, \prod_1^c) &= \max_{x \in [a, b]} |f(x) - \prod_1^c(x)| \\ &= \max_{0 \leq i \leq n-1} \max_{x \in [x_i, x_{i+1}]} |f(x) - \prod_1^c(x)| \end{aligned}$$

Ricordiamo la stima dell'errore di interpolazione polinomiale a grado s :

$$\max_{x \in [\alpha, \beta]} |f(x) - \prod_s(x)| \leq \max_{x \in [\alpha, \beta]} |f^{(s+1)}(x)| \cdot \frac{h^{s+1}}{4(s+1)} \quad \text{con } h = \frac{\beta - \alpha}{s}$$

Applichiamo al nostro caso: $s = 1$, $[\alpha, \beta] = [x_{i-1}, x_i]$

$$\max_{x \in [x_{i-1}, x_i]} |f(x) - \prod_{1,i}(x)| \leq \max_{x \in [x_{i-1}, x_i]} |f''(x)| \cdot \frac{h^2}{8} = M_{2,i} \frac{h^2}{8}$$

con $M_2 = \max_{x \in [x_{i-1}, x_i]} |f''(x)|$ e $h = \Delta x_i$.

Da cui:

$$\text{dist}(f, \prod_1^c) \leq \frac{M_2}{8} h^2$$

con $M_2 = \max_{x \in [a, b]} |f''(x)|$

10 Stima delle equazioni normali per l'approssimazione polinomiale ai minimi quadrati

Dati N punti $\{(x_i, y_i)\} : y_i = f(x_i), 1 \leq i \leq N$ e $m < N$, il vettore $a \in \mathbb{R}^{m+1}$

minimizza $\phi(a) = \sum_{i=1}^N (y_i - \sum_{j=0}^m a_j \cdot x_i^j)^2 \iff$ risolve il sistema $V^t V a = V^t y$

Dimostrazione

Osserviamo le dimensioni degli elementi considerati

$$V \in \mathbb{R}^{N \times (m+1)}, \quad V^t \in \mathbb{R}^{(m+1) \times N}, \quad y \in \mathbb{R}^N, \quad a \in \mathbb{R}^{m+1}$$

Quindi per $m = 1$ non importa quanti dati N ci siano, il sistema sarà sempre 2×2 poiché ci saranno 2 coefficienti.

Dire che $a \in \mathbb{R}^{m+1}$ è di minimo (assoluto) per $\phi(a)$ significa:

$$\phi(a + b) \geq \phi(a) \quad \forall b \in \mathbb{R}^{m+1}$$

Osserviamo che

$$\begin{aligned} \phi(a + b) &= (y - V(a + b), y - V(a + b)) = (y - Va - Vb, y - Va - Vb) = \\ &= (y - Va, y - Va) + (y - Va, -Vb) + (-Vb, y - Va) + (-Vb, -Vb) = \\ &= \phi(a) + 2(Va - y, Vb) + (Vb, Vb) = \phi(a) + 2(V^t(Va - y), b) + (Vb, Vb) \end{aligned}$$

dove abbiamo usato le seguenti proprietà del prodotto scalare in \mathbb{R}^m (per chiarezza indicato con $(u, v)_n$; ricordiamo che $(u, v)_n = u^t v$ interpretando i vettori come vettori-colonna):

1. $(u, v)_n = (v, u)_n \quad u, v, w \in \mathbb{R}^n$
2. $(\alpha u, v)_n = \alpha(u, v)_n \quad \alpha \in \mathbb{R}$
3. $(u + v, w)_n = (u, w)_n + (v, w)_n$
4. $(u, Az)_n = (A^t u, z)_k \quad u \in \mathbb{R}^n, z \in \mathbb{R}^k, A \in \mathbb{R}^{n \times k}$

Dimostriamo “ \Leftarrow ”:

Se $V^t V a = V^t y$ allora:

$$V^t V a - V^t y = 0 \quad \iff \quad V^t (Va - y) = 0$$

Ma allora

$$\begin{aligned} \phi(a + b) &= \phi(a) + (Vb, Vb) \geq \phi(a) \quad b \in \mathbb{R}^{m+1} \\ &\quad \parallel \\ &\quad \sum_{i=1}^N (Vb)_i^2 \geq 0 \end{aligned}$$

Dimostriamo “ \Rightarrow ”:

Assumiamo che

$$\phi(a + b) \geq \phi(a) \quad \forall b \in \mathbb{R}^{m+1}$$

Allora:

$$\phi(a + b) = \phi(a) + 2(V^t(Va - y), b) + (Vb, Vb) \geq \phi(a) \quad \forall b$$

Cioè:

$$2(V^t(Va - y), b) + (Vb, Vb) \geq 0 \quad \forall b$$

Prendiamo $b = \varepsilon v$, con v versore (cioè vettore di lunghezza 1, $(v, v) = 1$). Si ha:

$$\begin{aligned} & 2(V^t(Va - y), \varepsilon v) + (V(\varepsilon v), V(\varepsilon v)) \\ &= 2\varepsilon(V^t(Va - y), v) + \varepsilon^2(Vv, Vv) \geq 0 \quad \forall \varepsilon \geq 0 \text{ e } \forall v \end{aligned}$$

Dividendo per $\varepsilon > 0$:

$$2(V^t(Va - y), v) + \varepsilon(Vv, Vv) \geq 0 \quad \forall \varepsilon \text{ e } \forall v$$

Per $\varepsilon \rightarrow 0$ si ha:

$$(V^t(Va - y), v) \geq 0 \quad \forall v$$

Ma se vale \forall versore, possiamo prendere $-v$:

$$(V^t(Va - y), -v) = -(V^t(Va - y), v) \geq 0 \quad \forall v$$

$$\Downarrow$$

$$(V^t(Va - y), v) \leq 0 \quad \forall v$$

Ma abbiamo che

$$0 \leq (V^t(Va - y), v) \leq 0 \iff (V^t(Va - y), v) = 0 \quad \forall v$$

L'unico vettore ortogonale a tutti i vettori è il vettore nullo. Quindi

$$V^t(Va - y) = 0 \iff V^tVa = V^ty$$

11 Stime di condizionamento per un sistema lineare

(i) $\|Ax\| \leq \|A\| \cdot \|x\|$ (1° disuguaglianza fondamentale)

(ii) $\|AB\| \leq \|A\| \cdot \|B\|$ (2° disuguaglianza fondamentale)

Caso 1 perturbazione termine noto

Sia

- $A \in \mathbb{R}^{n \times n}$ non singolare
- $x \in \mathbb{R}^n$ soluzione del sistema $Ax = b$ con $b \neq 0$
- $\tilde{x} = x + \delta x$ soluzione del sistema $A\tilde{x} = \tilde{b}$ con $\tilde{b} = b + \delta b$

Fissata una norma vettoriale $\|\cdot\|$ in \mathbb{R}^n , vale la seguente stima dell'errore relativo su x

$$\frac{\|\delta x\|}{\|x\|} \leq K(A) \frac{\|\delta b\|}{\|b\|} \quad \text{con} \quad k(A) \underset{\text{indice di condiz.}}{=} \|A\| \cdot \|A^{-1}\|$$

Dimostrazione

Osserviamo che $x = A^{-1}b \neq 0$ quindi ha senso studiare l'errore relativo (dividere per $\|x\|$).

Si ha

$$\begin{cases} \tilde{x} = x + \delta x \\ \tilde{x} = A^{-1}\tilde{b} = A^{-1}(b + \delta b) = \underbrace{A^{-1}b}_{=x} + A^{-1}\delta b \end{cases} \Rightarrow \|\delta x\| = \|A^{-1}\delta b\| \underset{1^\circ \text{ dis.fond.}}{\leq} \|A^{-1}\| \cdot \|\delta b\|$$

Per stimare $\frac{1}{\|x\|}$ da sopra, cioè da sotto $\|x\|$.

$$\|b\| = \|Ax\| \underset{1^\circ \text{ dis.fond.}}{\leq} \|A\| \cdot \|x\|$$

da cui

$$\|x\| \geq \frac{\|b\|}{\|A\|}$$

e

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

perciò

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\delta b\|}{\|x\|} \leq \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\delta b\|}{\|b\|} = k(A) \cdot \frac{\|\delta b\|}{\|b\|}$$

Caso 2 perturbazione matrice

Siano fatte le stesse ipotesi del caso 1, ma con $\tilde{A}\tilde{x} = b$, $\tilde{A} = A + \delta A$.

Vale la stima dell' "errore relativo" su x

$$\frac{\|\delta x\|}{\|\tilde{x}\|} \leq k(A) \cdot \frac{\|\delta A\|}{\|A\|}$$

Dimostrazione

$$\begin{cases} \tilde{A}\tilde{x} = (A + \delta A)(x + \delta x) \\ \quad = Ax + A\delta x + \delta A\tilde{x} \\ \quad = b + A\delta x + \delta A\tilde{x} \\ \tilde{A}\tilde{x} = b \end{cases} \Rightarrow A\delta x + \delta A\tilde{x} = 0 \iff \delta x = -A^{-1}(\delta A\tilde{x})$$

Quindi

$$\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta A \tilde{x}\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|\tilde{x}\|$$

e perciò

$$\frac{\|\delta x\|}{\|\tilde{x}\|} \leq \|A^{-1}\| \cdot \|\delta A\| = \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta A\|}{\|A\|} = k(A) \cdot \frac{\|\delta A\|}{\|A\|}$$

Caso 3 perturbazione termine noto e matrice

Stesse ipotesi degli altri casi ma con $\tilde{A}\tilde{x} = \tilde{b}$.

Si ha che se $k(A) \cdot \frac{\|\delta A\|}{\|A\|} < 1$ allora:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{k(A)}{1 - k(A) \cdot \frac{\|\delta A\|}{\|A\|}} \cdot \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

Le seguenti sono domande che sono state fatte in esami, ma non erano presenti nel syllabus.

Le raccolgo, per comodità, in un file unico.

Altre domande, presenti, in "Tutte le risposte esami" fanno parte di risposte syllabus, risposte relative ai pdf messi sui Moodle del prof e, per comodità, sempre lì presenti (si spera non servano).

Queste invece, è ben probabile servano.

Rapporto incrementale standard

derivazione numerica standard col rapporto incrementale

Supponiamo $f \in C^2(I; \mathbb{R})$, con I intervallo di derivazione, f derivabile, dato il rapporto incrementale:

$$S_+(h) = \frac{f(x+h) - f(x)}{h}, \quad h > 0$$

si avrà che $S_+(h) = f'(x) + O(h)$ ed espandendo la formula di Taylor si ottiene:

$$f(x+h) = f(x) + f'(x)h + \frac{f''(\xi)}{2}h^2, \quad \xi \in \text{int}(x, x+h)$$

$$\text{quindi } S_+(h) = f'(x) + \frac{f''(\xi)}{2}h$$

Se \tilde{S}_+ approssima S_+ , si ha la seguente stima dell'errore:

$$|f'(x) - \tilde{S}_+(h)| \leq |f'(x) - S_+(h)| + |S_+(h) - \tilde{S}_+(h)|$$

da cui

$$\begin{aligned} |S_+(h) - \tilde{S}_+(h)| &= \frac{f(x+h) - f(x) - \tilde{f}(x+h) + \tilde{f}(x)}{h} \\ &\leq \frac{1}{h} (|f(x+h) - \tilde{f}(x+h)| + |\tilde{f}(x) - f(x)|) \\ &\leq \frac{1}{h} (|f(x+h) - \tilde{f}(x+h)| + |\tilde{f}(x) - f(x)|) \\ &\leq \frac{2}{h} \|f - \tilde{f}\|_\infty \end{aligned}$$

ponendo $\varepsilon \geq \|f - \tilde{f}\|_\infty$ si ottiene che

$$|f'(x) - \tilde{S}_+(h)| \leq ch + \frac{2\varepsilon}{h} := E(h), \quad c = \frac{f''(\xi)}{2}$$

se $h \rightarrow 0$, $E(h) \rightarrow \infty$ e poiché $E'(h) = c - \frac{2\varepsilon}{h^2}$ il suo punto minimo è $\sqrt{\frac{2\varepsilon}{c}} = O(\sqrt{\varepsilon})$

Dunque l'errore minimo commesso dall'approssimazione della derivata utilizzando il rapporto incrementale $\frac{f(x+h) - f(x)}{h}$ è $O(\sqrt{\varepsilon})$ per $f \in C^2$

Rapporto incrementale simmetrico

DERIVAZIONE SIMMETRICA CON RAPPORTO INCREMENTALE SIMMETRICO

Assumendo che $f \in C^3(I)$, con I intervallo di derivazione, scriviamo la formula di Taylor "da destra" e "da sinistra" centrata in x

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) - \frac{h^3}{3} f'''(\xi)$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2} f''(x) - \frac{h^3}{3} f'''(\eta)$$

dove $\xi \in (x, x+h)$, $\eta \in (x-h, x)$

da cui si ottiene, sottraendo membro a membro

$$f(x+h) - f(x-h) = 2hf'(x) + O(h^3) \text{ e anche } S(h) = \frac{f(x+h) - f(x-h)}{2h}$$
$$= f'(x) + O(h^2)$$

con $|f'(x) - S(h)| = \frac{1}{12} |f'''(\xi) + f'''(\eta)| h^2 \leq \frac{1}{12} (|f'''(\xi)| + |f'''(\eta)|) h^2 \leq dh^2$

dove $d = \frac{1}{6} \max_{t \in I} |f'''(t)|$

Questo dimostra che l'errore commesso dall'approssimazione dello derivato in x con il rapporto incrementale simmetrico

$$S(h) = \frac{f(x+h) - f(x-h)}{2h} \text{ è } O(h^2) \text{ per } f \in C^3(I)$$

Nota: quanto segue fa parte della risposta: "Metodo di Newton, quando ha convergenza almeno/esattamente 2"; la risposta è completa al 90% con la dimostrazione "Velocità di convergenza del syllabus", ma le note finali di questo screen servono per rispondere al primo pezzo della domanda.

Usando Taylor:

$$f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{f''(z_n)}{2} (\xi - x_n)^2$$
$$-\frac{f(x_n)}{f'(x_n)} = \xi - x_n + \frac{1}{2f'(x_n)} f''(z_n) (\xi - x_n)^2 \text{ con } -\frac{f(x_n)}{f'(x_n)} = x_{n+1} - x_n \text{ per def.}$$

quindi:

$$x_{n+1} - \xi = \frac{f''(z_n)}{2f'(x_n)} (\xi - x_n)^2 \rightarrow |x_{n+1} - \xi| = \frac{|f''(z_n)|}{|2f'(x_n)|} |\xi - x_n|^2 \text{ che diventa:}$$
$$e_{n+1} = \frac{1}{2} \cdot \left| \frac{f''(z_n)}{f'(x_n)} \right| \cdot e_n^2 \text{ si ha quindi: } \lim \frac{e_{n+1}}{e_n^2} = \frac{1}{2} \left| \frac{f''(\xi)}{f'(\xi)} \right| \text{ da cui si deriva}$$

che se $f''(\xi) \neq 0$ e $f'(\xi) \neq 0$ l'ordine di convergenza è esattamente 2. Se $f''(\xi) = 0$, $f'(\xi) \neq 0$ e $\exists f'''(\xi)$ allora l'ordine è almeno 3.

Sistema delle equazioni normali per approssimazione lineare

Sapendo che dati N punti $\{(x_i, y_i)\}$, $y_i = f(x_i)$, $1 \leq i \leq N$ e $m < N$, se il vettore $a \in \mathbb{R}^{m+1}$ minimizza $\phi(a) = \sum_{i=1}^N (y_i - \sum_{j=0}^m a_j x_i^j)^2$ allora risolve il sistema

$V^T V a = V^T y$, si possono usare le proprietà di $V^T V$ per trovare il sistema relativo alla retta dei minimi quadrati. $V^T V$ è una matrice simmetrica e semidefinita positiva. Inoltre $(Vv, Vv) = 0 \iff Vv = 0$ e $(Vv, Vv) = (V^T V v, v)$

quindi $v = 0$ se V ha rango max cioè se ha almeno $m+1$ punti distinti tra i nodi di campionamento. Si ricava quindi una matrice V t.c.

$V = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m+1} & x_{m+1}^2 & \dots & x_{m+1}^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^m \end{pmatrix}$ La sottomatrice $V \in \mathbb{R}^{(m+1) \times (m+1)}$ è matrice di Vandermonde per l'interpolazione di grado $\leq m$ su $m+1$ nodi distinti, quindi è non singolare.

Questo evidenzia che, quindi, il rango della sottomatrice è $m+1$ e che le intere colonne $m+1$ di V sono linearmente indipendenti come vettori di \mathbb{R}^N . Quindi si possono calcolare gli elementi della matrice $V^T V$

e del vettore noto $V^T y$, con $m=1$

$$V^T V = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} = \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$$

$$V^T y = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \in \mathbb{R}^2$$

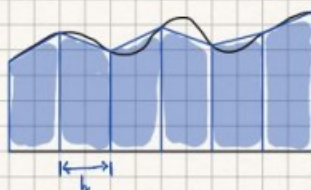
quindi il sistema è: $\begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$

Errore formula trapezi

La formula dei trapezi utilizza l'interpolazione lineare a tratti, imponendo $s=1$ l'integrale viene approssimato con la somma delle aree dei trapezi lineari. L' i -esimo trapezio ha altezza $h = \frac{b-a}{n}$ e basi $f(x_{i-1})$ e $f(x_i)$ con $1 \leq i \leq n$, si avrà quindi l'area $A = \frac{h}{2} (f(x_{i-1}) + f(x_i))$ quindi:

$\frac{h}{2} (f(x_{i-1}) + f(x_i)) + \sum_{i=1}^{n-1} h \cdot f(x_i)$, ottenendo così la formula dei trapezi:

$$I_n(f) = \sum_{i=0}^n w_i f(x_i) \quad \text{con} \quad w_i = \begin{cases} \frac{h}{2}, & i=0, n \\ h, & 1 \leq i \leq n-1 \end{cases}$$



$$I_n^{\text{trap}}(f) = I(\Pi_1^c) = \sum (\text{area trapezi lineari})$$

Per ricavare una stima dell'errore possiamo usare la stima $|I(f) - I_n(f)| = |I(f) - I(\Pi_n^c)| \leq |I(f - \Pi_n^c)| \leq (b-a) \text{dist}(f, \Pi_n^c)$. Se $\text{dist} \rightarrow 0$ allora ci sarà convergenza, altrimenti potrebbero presentarsi problemi di divergenza. Per quanto riguarda le formule di quadrature composte ottenute come $I_n(f) = I(\Pi_n^c)$, con n multiplo di s : $|I(f) - I_n(f)| \leq (b-a) \text{dist}(f, \Pi_n^c) \leq (b-a) K_s \cdot h^{s+1}$ se $f \in C^{s+1}[a, b]$ con $h = \max \Delta x_i$. Quindi per qualsiasi distribuzione dei nodi per cui $h \rightarrow 0$ se $f \in C^{s+1}[a, b]$ le formule sono sempre convergenti con un errore proporzionale a h^{s+1} , ma $s=1$ per i trapezi quindi per $f \in C^2$ sarà convergente con un errore $O(h^2)$.

1.5.1 Esempio 1

Consideriamo $\mathbb{F}(10, 4, L, U)$ (con L, U sufficienti per rappresentare i numeri che ci interessano) e

$$x = 0,10016$$

$$y = -0,10012$$

allora

$$\tilde{x} = fl^4(x) = 0,1002$$

$$\tilde{y} = fl^4(y) = -0,1001$$

eseguendo l'operazione-macchina di somma algebrica (che è una sottrazione visto che x e y hanno segno opposto) si ottiene

$$\begin{aligned}x \oplus y &= fl^4(fl^4(x) + fl^4(y)) \\&= fl^4(0,1002 - 0,1001) \\&= 10^{-4}\end{aligned}$$

scriveremo spesso i numeri in notazione standard per comodità)

Invece

$$x + y = 4 \cdot 10^{-5}$$

quindi l'errore relativo nel risultato è

$$\frac{|(x + y) - (x + y)|}{|x + y|} = \frac{|4 \cdot 10^{-5} - 10^{-4}|}{4 \cdot 10^{-5}} = \frac{6 \cdot 10^{-5}}{4 \cdot 10^{-5}} = \frac{3}{2} = 150\%$$