# UltimateCN

Università degli Studi di Padova

Dipartimento di Matematica

Corso di Laurea in Informatica

Anno accademico 2022 - 2023

*Autore:*
Nicola Baesso

# Premessa

Questa raccolta é stata ispirata dal giá ottimo documento "Calcolo Numerico Completo", e vuole estendere in maniera completa quanto presente.

La repository in cui segnalare problematiche o suggerire modifiche tramite PR si puo' trovare qui: https://github.com/nicolabaesso/UltimateCN

# Indice



# 1 Quiz parte teorica

Attenzione:
Il seguente file comprende:
1) tutte le risposte delle crocette
2) Tutte le dimostrazioni dell'ottimo file "syllabus.pdf" presente già su Mega (e correzioni/precisazioni dello stesso)
3) Ampliamenti alle risposte e completamenti dove mancanti alle dim. irrinunciabili e domande similari
4) Domande fuori dal syllabus chieste almeno una volta in esami
5) Domande fuori dal syllabus non chieste ma presenti a titolo di cautela

Credo sia uno strumento unico ed utile, forse il file definitivo per calcolo.

## 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:**

dove b è base, t è una serie di cifre di mantissa,
e l'esponente è compreso tra L ed U

- **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** $\leq 28$

**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 é:**

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 sostema 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 $A \in \mathbf{R^{n \times n}}$ 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**

# 2 Dimostrazioni irrinunciabili

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

**Risponde alla domanda:** Cos'é la precisione di macchina in un sistema floating-point F(b,t,L,U) e come si calcola? (si ricavi il valore). Anche chiesta come "Precisione di macchina".

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

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

(dove sign é la funzione segno, con valori negativi=-1, zero=0 e valori positivi=1)
il numero

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

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:

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

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

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

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

### 2.2.1 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 $\tilde{x}y$

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

$(*)$ Disugliaglianza 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 $\frac{|\, \tilde{x} \,|}{|\, x \,|}$

$$\frac{|\, \tilde{x} \,|}{|\, x \,|} = \underbrace{\frac{|\, \overbrace{x}^{=a} + \overbrace{\tilde{x} - x}^{=b} \,|}{|\, x \,|} \leq \frac{|\, x \,| + |\, \tilde{x} - x \,|}{|\, x \,|}}_{\text{Disuguaglianza Triangolare}} = 1 + \varepsilon_x$$

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.

### 2.2.2 Divisione

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

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

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

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

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

$$\left|\,1 + \frac{(\tilde{y} - y)}{y}\,\right| \geq \left|\,1 - \frac{|\,\tilde{y} - y\,|}{|\,y\,|}\,\right| = |\,1 - \varepsilon_y\,| = 1 - \varepsilon_y \qquad \Big(\text{perchè } \varepsilon_y < 1\Big)$$

da cui si ottiene

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

e quindi

$$\frac{|\,y\,|}{|\,\tilde{y}\,|} \leq \frac{|\,y\,|}{|\,y\,|(1 - \varepsilon_y)} = \frac{1 + \varepsilon_y}{(1 + \varepsilon_y)(1 - \varepsilon_y)} = \underbrace{\frac{1 + \varepsilon_y}{1 - \varepsilon_y^2} \approx 1 + \varepsilon_y}_{\text{Poiché } \varepsilon_y^2 \ll \varepsilon_y < 1}$$

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)

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

### 2.2.3 Somma Algebrica

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

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** $sign(x) = sign(y)$
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** $sign(x) \neq sign(y)$
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

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

**Esempio sottrazione (si inserisce il più semplice)**
Sia

$$F(10, 4, L, U)$$

con L e U di opportuna dimensione, i numeri

$$x = 0.10016$$

$$y = -0.10014$$

e le loro relative approssimazioni, ovvero

$$\tilde{x} = fl^4(x) = 0.1002$$

$$\tilde{y} = fl^4(y) = -0.1001$$

Si ha che

$$x \bigoplus y = fl^4(fl^4(x) + fl^4(y)) = fl^4(0.1002 - 0.1001) = 10^{-4}$$

quando $x + y = 4 * 10^{-5}$. Quindi l'errore relativo nel risultato è

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

Concludiamo calcolando i pesi:

$$w_1 = \frac{|x|}{|x+y|} \simeq \frac{10^{-1}}{4*10{-}5} = \frac{10^4}{4} = 2500$$

Analogamente, $w_2 \simeq 2500$.

## 2.3 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 $[a_n, b_n] \to [a_{n+1}, b_{n+1}]$ 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 $\xi \in (a,b)$

- $0 \leq |\,\xi - a_n\,|,\; |\,\xi - b_n\,| < \frac{b-a}{2^n} \underset{n\to\infty}{\longrightarrow} 0 \underset{\text{Teor. Carabinieri}}{\Longrightarrow} |\,\xi - a_n\,|,\; |\,\xi - b_n\,| \longrightarrow 0,\; n \to \infty$
- $0 \leq |\,\xi - x_n\,| < \frac{b-a}{2^{n+1}} \Longrightarrow |\,\xi - x_n\,| \longrightarrow 0,\; n \to \infty$

## 2.4 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)

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

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

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

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

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

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

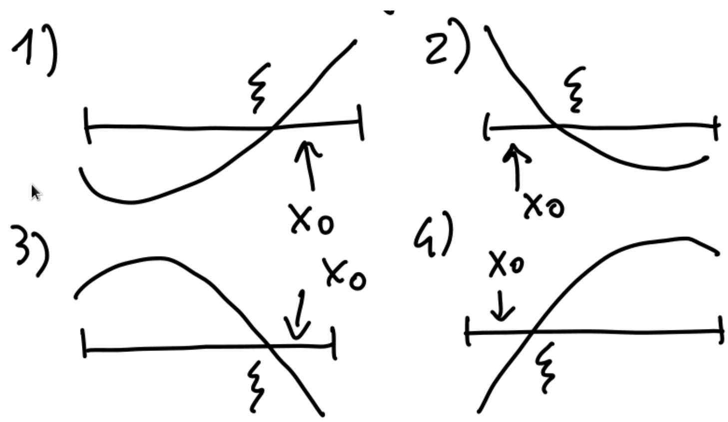
<u>Metodo di Newton</u>: Linearizzare iterativamente la funzione con la tangente nel punto

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

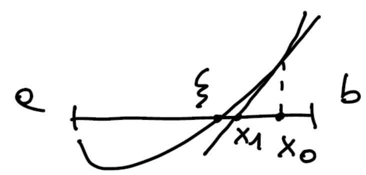
<u>Convergenza metodo di Newton</u>:

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

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




In questa dimostrazione di concentriamo sul caso 1)




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

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 $\xi$

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

$$x_{n+1} = x_n - \left.\tfrac{f(x_n)}{f'(x_n)}\right\} > 0$$

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

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

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

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

## 2.6 Velocità (ordine) di convergenza del metodo di Newton

Sia

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

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

**Dimostrazione**

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

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

$$\Downarrow$$

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

$$\Downarrow$$

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

$$\overset{aggiungendo\, i\, moduli}{\Downarrow}$$

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

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

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

## 2.7 Ordine di convergenza delle iterazioni di punto fisso

Sia $\xi$ 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:

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

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

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

$$\Downarrow$$

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

per 2) utilizziamo la formula di Taylor di grado $p-1$ centrata in $\xi$ 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
  $$\frac{e_{n+1}}{e_n^p} = \frac{|\phi^{(p)}(u_n)|}{p!} \underset{n\to\infty}{\longrightarrow} \frac{|\phi^{(p)}(\xi)|}{p!} \neq 0$$
  $e_n^p$ 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:
  $$\frac{e_{n+1}}{e_n^k} \underset{n\to\infty}{\rightarrow} \frac{|\phi^{(k)}(\xi)|}{k!} = L' \neq 0$$
  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 $p$.

## 2.8 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 $0 \leq i \leq n \quad \rightarrow \quad n+1$ modi di interpolare.
Poiché $\mathbb{P}_n$ è uno spazio vettoriale $\quad \Rightarrow \quad p - q \in \mathbb{R}_n$.
Allora:

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

$$\Downarrow$$

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

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

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

**Esistenza**

Definiamo il "polinomio di Lagrange":

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

dove

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

$l_i(x) \in \mathbb{P}_n$ 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":

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

Verifichiamo che interpola

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

## 2.9 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

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

**Dimostrazione** per $s = 1$.
Si ha che:

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

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

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

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

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

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

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

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

## 2.10 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 $\phi(a) = \sum\limits_{i=1}^{N}(y_i - \sum\limits_{j=0}^{m} a_j \cdot x_i^j)^2 \iff$ risolve il sistema $V^tVa = V^ty$

**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 $2 \times 2$ 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 $(u, v)_n = u^tv$ 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. $(u, Az)_n = (A^tu, z)_k \quad u \in \mathbb{R}^n,\ z \in \mathbb{R}^k,\ A \in \mathbb{R}^{n\times k}$

Dimostriamo "$\Leftarrow$":

Se $V^tVa = V^ty$ allora:

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

Ma allora

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

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 $\varepsilon \to 0$ 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

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

## 2.11 Stime di condizionamento per un sistema lineare

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

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

$\boxed{\textbf{Caso 1}}$ **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 $x$

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

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

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

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

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

da cui

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

e

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

perciò

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

$\boxed{\textbf{Caso 2}}$ **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**

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

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

$\boxed{\textbf{Caso 3}}$ **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:
$$\frac{\|\delta x\|}{\|x\|} \leq \frac{k(A)}{1 - k(A) \cdot \frac{\|\delta A\|}{\|A\|}} \cdot (\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|})$$

# 3 Dimostrazioni facoltative