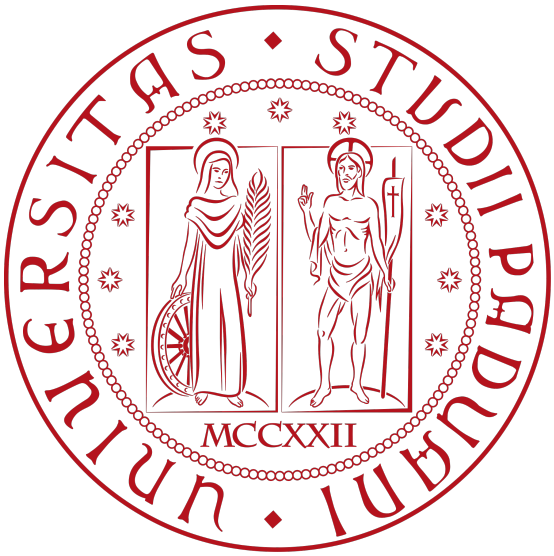


Riassunto di Calcolo Numerico

Indice

1 Introduzione

Questo è un breve riassunto che ho scritto dopo aver fallito per innumerevoli volte la prova scritta di calcolo numerico. Per ogni domanda non riporterò la dimostrazione esatta che bisogna scrivere all'esame, ma concetti che aiutano a capire il senso (e quindi a memorizzare gli argomenti).
Prima di leggere questo documento consiglio di rivedere la teoria di analisi, in particolare gli argomenti che vengono trattati anche nel corso di calcolo.
Consiglio di leggere quanto scritto nelle prossime pagine con occhio critico: sicuramente saranno presenti molti errori, spero però che questi appunti siano utili per comprendere come studiare le varie dimostrazioni.
In seguito riporto alcune cose utili, che saranno ricorrenti nelle varie dimostrazioni:

- La lettera greca ξ il prof la utilizza (solitamente quando parla dei metodi iterativi per risolvere le equazioni) per indicare la vera soluzione dell'equazione. Invece la successione x_n indica la soluzione che ho trovato all'n-esimo passaggio
- Con stima si intende dare un valore ad una quantità che non conosco, confrontandola con delle quantità note. Quando si parla di stima si avranno delle disuguaglianze, e qui i teoremi che vengono utilizzati di solito sono il teorema dei carabinieri, disuguaglianza triangolare e altre disuguaglianze fondamentali (queste compaiono solo nella domanda sulle stime di condizionamento)
- stimare "da sotto" vuol dire (letteralmente), in una frazione, considerare solo il denominatore. Quando, in seguito, si riuniscono numeratore e denominatore, è importante cambiare il verso della disequazione (vedi ad es. la domanda sulla precisione di macchina).

2 Domande del Syllabus

2.1 Precisione di macchina

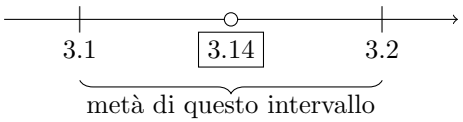
Si parte da un numero reale, scritto in notazione floating point:

$$x = sign(x)(0, d_1 d_2 ....dt...) \cdot b^p$$

È importante tenere a mente che la 1° cifra dopo la virgola (d_1) è diversa da zero.
Quindi si scrive il numero arrotondato a t cifre di mantissa, e si scrive la definizione formale di arrotondamento.
Poi diciamo che la precisione di macchina è l'errore relativo di arrotondamento, e scriviamo l'errore di arrotondamento:

$$\frac{|x - fl^t(x)|}{|x|}, |x| \neq 0$$

Quindi stimiamo questo rapporto in due parti: prima il numeratore.
Le cifre dei due numeri sono uguali fino alla $t-1$, cambia dalla t in poi (per l'arrotondamento), e con un esempio di può vedere che questa quantità è $\leq \frac{b^{-t}}{2}$
ESEMPIO: se $x = 3.14$ e voglio arrotondare ai decimi, l'errore che compio è $\leq$ mezzo decimo (in questo caso di 0.04):



Ora stimiamo il denominatore. Sappiamo che è una quantità positiva (c'è il valore assoluto), è $\neq 0$, e la prima cifra decimale (d_1) deve essere diversa da zero. Questo serve ad evitare che ci siano più rappresentazioni dello stesso numero.
Quindi $|x|$ è almeno ($\geq$) $0.1 \cdot b^p$ (non ci serve uno specifico p), ovvero $b^{-1} \cdot b^p => b^{p-1}$.
Siccome nella definizione di errore relativo $|x|$ "sta sotto", dobbiamo scrivere il reciproco:

$$\frac{1}{|x|} \leq \frac{1}{b^{p-1}}$$

Da notare il fatto che è cambiato il verso della disequazione.
Quindi uniamo i due pezzi, e otteniamo che:

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

Che è la nostra precisione di macchina.

2.2 Analisi di stabilità delle operazioni

Questa è la più lunga di tutte le dimostrazioni, ma:

- all'esame il prof ne chiede sempre metà (solitamente moltiplicazione e divisione oppure somma algebrica)
- Non serve impararsi tutto a memoria: gran parte del testo sono operazioni algebriche. Si utilizza la diseguaglianza triangolare e si moltiplica/divide per una certa quantità (questo nel caso della somma algebrica, nella moltiplicazione si somma/sottrae).
- Si inizia dalla definizione, ovvero l'errore relativo che ho su un'operazione con numeri approssimati:

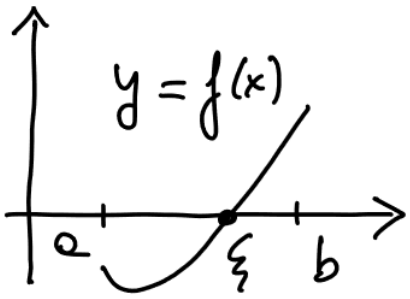
$$\varepsilon_{x \star y} = \frac{|x \star y - \widetilde{x} \star \widetilde{y}|}{|x \star y|}$$

  Dove al posto di $\star$ si mette una delle operazioni, e $\widetilde{x}$, $\widetilde{y}$ (con la tilde) sono i numeri approssimati.

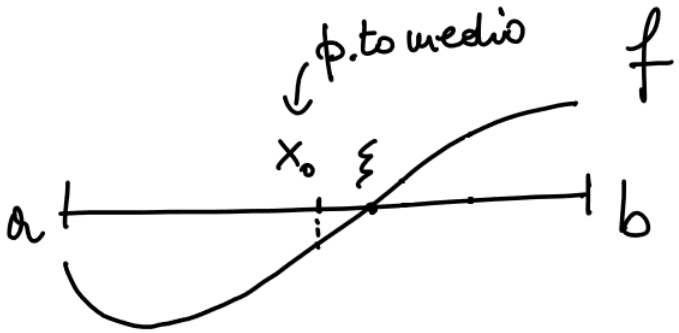
- La sottrazione (ovvero la somma algebrica nel caso in cui il segno dei due numeri è diverso) è l'unica operazione instabile (quando x e y sono vicini in termini relativi)

2.3 Convergenza del metodo di bisezione

Questa è la prima domanda in cui il prof inizia a utilizzare la lettera ξ (csi), che semplicemente indica la soluzione che stiamo cercando. Invece con x (oppure x_n) indica la "soluzione" che abbiamo trovato a una certa iterazione. In generale, più iterazioni abbiamo e più la nostra x si avvicina a ξ.
Con convergenza di un metodo (in questo caso la bisezione) si intende il fatto che, facendo iterazioni, ci avviciniamo sempre di più alla soluzione ξ, lo "zero" della funzione.
Il metodo di bisezione funziona applicando iterativamente il teorema degli zeri (o teorema di Bolzano).



Se ho una funzione continua in un intervallo $[a, b]$ e $f(a) \cdot f(b) \leq 0$ (ovvero hanno segno opposto), allora esiste un punto ξ, con $f(\xi) = 0$, che appartiene all'intervallo (a, b).
Il metodo consiste nell'individuare il punto medio, e continuare il procedimento sulla metà dell'intervallo in cui vale la condizione che i due estremi hanno segno opposto.
Esempio:



In questa funzione abbiamo come estremi a, b. Individuiamo il punto medio x_0, e la successiva iterazione la facciamo sul semintervallo $[a, x]$, perché $f(a)f(x_0) \leq 0$.
Il punto medio è $x_n = \frac{a_n+b_n}{2}$.
Poi si individuano tre successioni: a_n, b_n, x_n, tali che:

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

$\frac{b-a}{2^n}$ è la lunghezza dell'intervallo $b_n - a_n$, mentre $\frac{b-a}{2^{n+1}}$ è la distanza tra il punto medio e ξ, e non può superare metà dell'intervallo $b_n - a_n$.
Quindi si dimostra che le tre successioni convergono ad uno zero, utilizzando il teorema dei carabinieri:

$$0 \leq |\xi - a_n|, |\xi - b_n| < \frac{b-a}{2^n} \to 0, n \to \infty \overset{teor.carabinieri}{\Longrightarrow} |\xi - a_n|, |\xi - b_n| \to 0, n \to \infty$$

Qui semplicemente diciamo che $|\xi - a_n|, |\xi - b_n|$ sono due quantità positive (per via dei valori assoluti), e da quello che abbiamo detto prima sono $\leq$ di $\frac{b-a}{2^n}$.

Calcolando il $\lim_{n\to\infty} \frac{b-a}{2^n}$ otteniamo che è zero. Quindi possiamo utilizzare il teorema dei carabinieri per dire che le due successioni tendono a zero.
Lo stesso ragionamento lo utilizziamo per la terza successione:

$$0 \leq |\xi - x_n| < \frac{b-a}{2^{n+1}} \implies |\xi - x_n| \to 0, n \to \infty$$

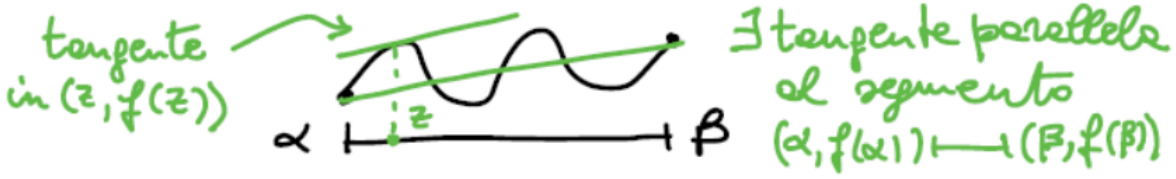
2.4 Stima dell'errore con residuo pesato (bisezione)

Prima di tutto è bene tenere a mente perché il residuo non pesato (detto anche stima a priori) non è una buona stima dell'errore.
Poi, partiamo dalle ipotesi: sono le stesse del metodo di bisezione, con l'aggiunta che la funzione, oltre a essere continua in $[a, b]$, deve anche essere derivabile, e questa derivata deve essere $\neq 0$ in tutto l'intervallo.
Quindi abbiamo che l'errore

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

Dove z_n è un punto che appartiene all'intervallo (x_n, ξ), ma non sappiamo esattamente quanto vale.
Si dimostra utilizzando il teorema del valor medio (detto anche teorema di Lagrange), che dice che se f è continua in $[a, b]$ e derivabile in (a, b), allora esiste un punto z in (a, b) : la retta che passa per $f(a)$ e $f(b)$ è parallela alla tangente in z:

$$\frac{f(b) - f(a)}{b - a} = f'(z)$$



Siccome z_n è nell'intervallo (x_n, ξ), e x_n tende a ξ, allora la derivata di z_n ha un valore molto vicino a quello della derivata di ξ.
Con la notazione $int(x_n, \xi)$ indichiamo un intervallo che ha come estremi x_n e ξ, ma non sappiamo quale dei due è più grande.
La dimostrazione si fa con un'applicazione del teorema del valor medio. Supponiamo di essere nel caso in cui $\xi < xn$ (l'altro caso è analogo), e quindi $a = \xi$, $b = x_n$, e "sostituiamo" questi valori nella formula del valor medio:

$$f(x_n) - f(\xi) = f'(z_n)(x_n - \xi) \to \text{il denominatore lo portiamo sopra}$$

Che si può riscrivere come

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

2.5 Convergenza globale del metodo di Newton

Il metodo di Newton è un altro metodo iterativo per trovare gli zeri di una funzione. Rispetto alla bisezione converge molto più rapidamente, ma ha bisogno di ipotesi più forti[1].
Questo è il metodo di Newton:

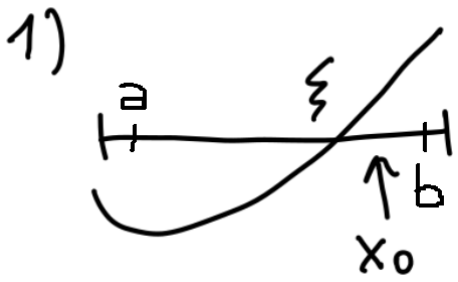
$$\begin{cases} y = 0 \\ y = f(x_n) + f'(x_n)(x - x_n) \end{cases} \quad \Rightarrow \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

[1] vedi lezione 9 - confronto tra bisezione e metodo di Newton

Queste sono le ipotesi:

$$\begin{cases} f \in C^2[a,b] & \text{la funzione ha derivata seconda e } f''(x) \text{ è continua} \\ f(a)f(b) < 0 & f(a) \text{ e } f(b) \text{ hanno segno opposto} \\ f''(x) > 0 \forall x \in [a,b] & \text{dimostriamo solo il caso di concavità stretta} \\ x_0 : f(x_0)f''(x_0) > 0 & \text{punto iniziale "scelto bene"} \end{cases}$$

Riguardo la scelta del punto iniziale, è una condizione importante, altrimenti il metodo non converge alla soluzione. Nel nostro caso, dove abbiamo concavità stretta ($f''(x) > 0$), dobbiamo per forza scegliere un punto $f(x) > 0$.
Nella dimostrazione vengono illustrati graficamente quattro casi differenti, in base al segno di $f''(x)$ e se $f(a) > f(b)$ (oppure viceversa). Basta farne solo uno (gli altri sono uguali), per comodità scegliamo il primo, che è quello che tratta anche il prof. a lezione:
CASO (1):



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

Dobbiamo dimostrare due cose: innanzitutto, per induzione, se l'n-esimo termine appartiene all'intervallo $(\xi, b]$, allora lo stesso vale per il termine $n+1$-esimo. La dim. si fa a parole.
Poi bisogna dimostrare che x_n è decrescente. Qui si riprende la definizione di metodo di Newton:

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

Guardando il grafico del caso (1) si capisce che, nell'intervallo $(\xi, b]$, $f(x)$ è sempre > 0.
Stessa cosa per $f'(x_n)$, che potrebbe essere negativo solo se $f''(x)$ cambiasse segno in $(\xi, b]$, cosa che escludiamo (vedi le ipotesi iniziali).
Quindi x_{n+1} è dato da x_n a cui viene sottratta una quantità positiva. Questo è sufficiente per dire che la successione x_n è decrescente.
Infine diciamo che, visto che la successione è decrescente, per il teorema della monotonia diciamo che il limite (per $n \to \infty$) della successione è il suo estremo inferiore, che il prof indica con la lettera η.
Quindi passiamo al limite della formula, e diciamo che:

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

Dopo un po' di proprietà di limiti e continuità otteniamo

$$\eta = x_n - \frac{f(\eta))}{f'(\eta)}$$

Infine diciamo che

$$\frac{f(\eta)}{f'(\eta)} = 0 \Rightarrow f(\eta) = 0 \Rightarrow \eta = \xi$$

2.6 Ordine di convergenza del metodo di Newton

Per prima cosa è importante riguardare la formula di Taylor, servirà per la dimostrazione.
Le ipotesi di partenza sono le stesse della domanda precedente, con l'aggiunta che la derivata prima deve essere $\neq 0$ per ogni x contenuto in un sottointervallo di $[a, b]$. Questo ci servirà più avanti, per evitare la divisione per zero.
Dalle ipotesi, facciamo la seguente affermazione:

$$e_{n+1} \leq ce_n^2, \quad n \geq 0, \quad c = \frac{1}{2} \cdot \frac{M_2}{m_1}$$

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

In pratica stiamo affermando che l'errore cala od ogni iterazione di un fattore quadratico moltiplicato per una costante c. La costante c (non l'ho capita molto bene) così definita misura la velocità di variazione della derivata seconda (una cosa simile compare nella stima dell'errore con residuo pesato).
Per dimostrare che l'affermazione appena fatta è vera, calcoliamo lo zero della funzione, $f(\xi)$ con la formula di Taylor, [2], che ci consente di avere una buona approssimazione del risultato:

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

Non so perché serve mettere z_n, però viene messo nella derivata seconda.
Sappiamo che $f(\xi)$ è zero, quindi lo togliamo. Portiamo a sinistra $f(x_n)$, che cambierà segno, e dividiamo entrambi i menbri per $f'(x_n)$, e otteniamo questo:

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

Adesso a sinistra abbiamo letteralmente la definizione del metodo di Newton, quindi possiamo sostituire il rapporto con $x_{n+1} - x_n$, semplifichiamo il termine $-x_n$, e otteniamo questo:

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

Da qui ritorniamo subito all'affermazione iniziale.
Intanto l'errore al passo $n+1$ è la differenza in modulo tra il valore calcolato al passo n+1 e la soluzione vera: $e_{n+1} = |x_{n+1} - \xi|$. Aggiungiamo il modulo perché l'errore è una quantità positiva (vogliamo sapere quanto è grande l'errore).
$(\xi - x_n)^2$ è l'errore al passo n, al quadrato, ovvero e_n^2 (qui non abbiamo bisogno del modulo perché è tutto elevato al quadrato).
Quello che resta è la costante moltiplicativa c (dove, per qualche motivo, bisogna aggiungere i moduli).

2.7 Ordine di convergenza delle iterazioni di punto fisso

Punto fisso:

Il teorema delle contrazioni afferma, spiegato in termini poco formali, che una funzione f che "contrae" le distanze ha un unico punto fisso, ...
Link utile: https://www1.mat.uniroma1.it/people/orsina/Frattali.pdf

2.8 Esistenza e unicità dell'interpolazione polinomiale

UNICITÀ:
Per dimostrare che il polinomio interpolatore è unico si procede per assurdo, supponendo di avere due

[2] https://it.wikipedia.org/wiki/Serie_di_Taylor

polinomi, p,q, diversi, ed entrambi che interpolano in tutti i punti x_i (importante dire che se i punti campionati sono $n+1$, il grado massimo dei polinomi è n).
Siccome stiamo affermando che entrambi i polinomi sono interpolatori, allora i loro valori nei punti campionati sono uguali, quindi:

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

Da questo possiamo dire che $p-q$ ha $n+1$ zeri distinti.
Quindi, per il teorema fondamentale dell'algebra, $p-q$ può avere al massimo n zeri distinti, a meno che non sia il polinomio nullo. Quindi, si vede che, se $p-q=0$, vale che $p=q$, che va contro l'ipotesi iniziale (e quindi si dimostra l'unicità).
ESISTENZA:
Per dimostrare l'esistenza, utilizziamo il metodo del polinomio di Lagrange $l_i(x)$. Prima affermiamo il metodo, e poi dimostriamo che trova un polinomio interpolatore.

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

Il prof scrive il polinomio con una scrittura diversa (distingue numeratore e denominatore) ma equivalente.
$l_i(x)$ è un polinomio perché il numeratore è un polinomio, e il denominatore è un numero $\neq 0$.
Osserviamo che questo polinomio si annulla in tutti i valori tranne uno, dove si semplificano tutti i termini e vale 1, questo il prof lo indica come delta di Kronecker, δ_{ik}.
Adesso definiamo il polinomio interpolatore di Lagrange, indicato con $f_n(x)$, che è la somma di tutti i polinomi l_i moltiplicato per i corrispondenti valori campionati y_i:

$$f_n(x) = \sum_{i=0}^{n} y_i l_i(x)$$

Infine, per verificare che il polinomio interpola, si sostituisce a $l_i(x_k)$ il delta di Kronecker, che poi si sewmplifica perché se $i=k$ vale 1, altrimenti vale 0.
Non è semplicissima da capire questa dimostrazione, però consiglio di guardare i primi minuti di questo video, dove viene mostrato un esempio applicativo:
https://www.youtube.com/watch?v=GITFkl8I3qI

2.9 Convergenza uniforme dell'interpolazione lineare a tratti

2.10 Stime di condizionamento: perturbazione termine noto

Partiamo innanzitutto da alcune definizioni, utili ad aggiungere un po' di contesto:
NORMA VETTORIALE: La norma indica il modulo (lunghezza) del vettore, si utilizza quando bisogna confrontare la grandezza di diversi vettori o matrici.[3]
SISTEMA LINEARE: è un sistema di equazioni lineari, ossia un sistema costituito da equazioni in più incognite ove ogni incognita compare con esponente 1. Nel nostro caso avremo un sistema di n equazioni in n incognite (questo ci serve perché la matrice dei coefficienti A deve essere invertibile).
La perturbazione è un errore generico (non sappiamo se è dovuto ad approssimazioni o ha altre cause), e va a modificare il risultato. Nel nostro caso, abbiamo che la perturbazione è sul termine noto, che è b, quindi il termine noto con l'errore lo indichiamo con $\tilde{b}$, definito nel seguente modo:
$\tilde{b} = b + \delta b$ ovvero b perturbato è determinato dal valore originale (b) a cui aggiungiamo l'errore (δb), che può essere una quantità positiva o negativa.
Quindi, avendo un errore sul termine noto del sistema, questo avrà conseguenze anche sulle soluzioni dello stesso, e quindi la perturbazione sarà presente anche sul vettore delle soluzioni x.
Se il sistema non affetto da errori è $Ax = b$, il sistema perturbato sarà $A\tilde{x} = \tilde{b}$
Il condizionamento è la risposta del sistema agli errori sui dati:se il sistema è mal condizionato, un piccolo errore sui dati comporta ad una grande differenza sulla soluzione.

[3] https://www.youmath.it/lezioni/algebra-lineare/matrici-e-vettori/882-norma-e-prodotto-scalare.html

Nella dimostrazione[4] non è obbligatorio scrivere entrambe le disuguaglianze fondamentali (visto che utilizziamo solo la prima):
$\|Ax\| \leq \|A\| \cdot \|x\|$ 1° disuguaglianza fondamentale
Ora scriviamo le ipotesi:

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

Fissata una norma $\|\cdot\| \in \mathbb{R}$, affermiamo che vale la seguente stima dell'errore relativo su x (le norme vanno messe in tutte le disuguaglianze della dimostrazione):

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

La dimostrazione inizia osservando che il sistema di partenza può essere scritto come $x = A^{-1}b$, perché A^{-1} è l'inversa di A, ovvero $A^{-1} = \frac{1}{A}$.

$$\begin{cases} \tilde{x} = x + \delta x \\ \tilde{x} = A^{-1}\tilde{b} = A^{-1}(b + \delta b) = A^{-1}b + A^{-1}\delta b \end{cases} \Rightarrow \quad \|\delta x\| = \|A^{-1}\delta b\| \leq_{*} \|A^{-1}\| \cdot \|\delta b\|$$

* 1° disusuaglianza fondamentale
$\|\delta x\| = \|A^{-1}\delta b\|$ la otteniamo dalle ipotesi iniziali e da un pezzo della dimostrazione. Infatti $\delta x = x + \tilde{x}$. Poco fa abbiamo visto che $\tilde{x} = A^{-1}\tilde{b}$, che è uguale a $x + A^{-1}\delta b$. A questo punto otteniamo $x + \delta x = x + A^{-1}\delta b$. A questo punto semplifichiamo x e mettiamo le norme.
Finora abbiamo stimato $\|\delta x\|$, adesso stimiamo il denominatore, $\|\frac{1}{x}\|$, da sotto $\|x\|$:

$$\|b\| = \|Ax\| \leq_{*} \|A\| \cdot \|x\|$$

* 1° disusuaglianza fondamentale
Da cui

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

e

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

Perciò (si mettono insieme le parti colorate per ritornare alla stima iniziale):

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

[4] Per approfondire: http://dm.unife.it/~tinti/Didattica/Labcn/lucidi9.pdf