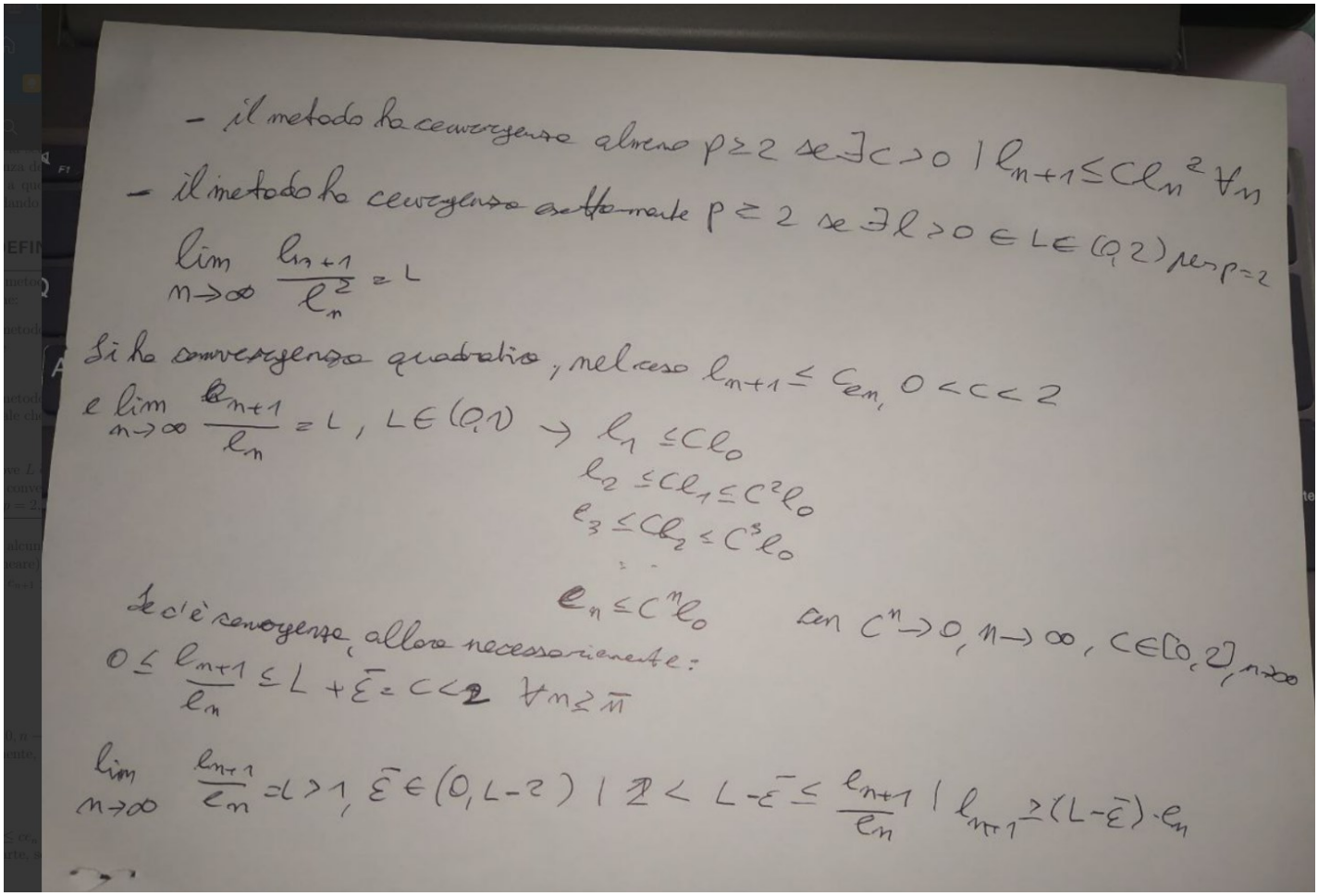
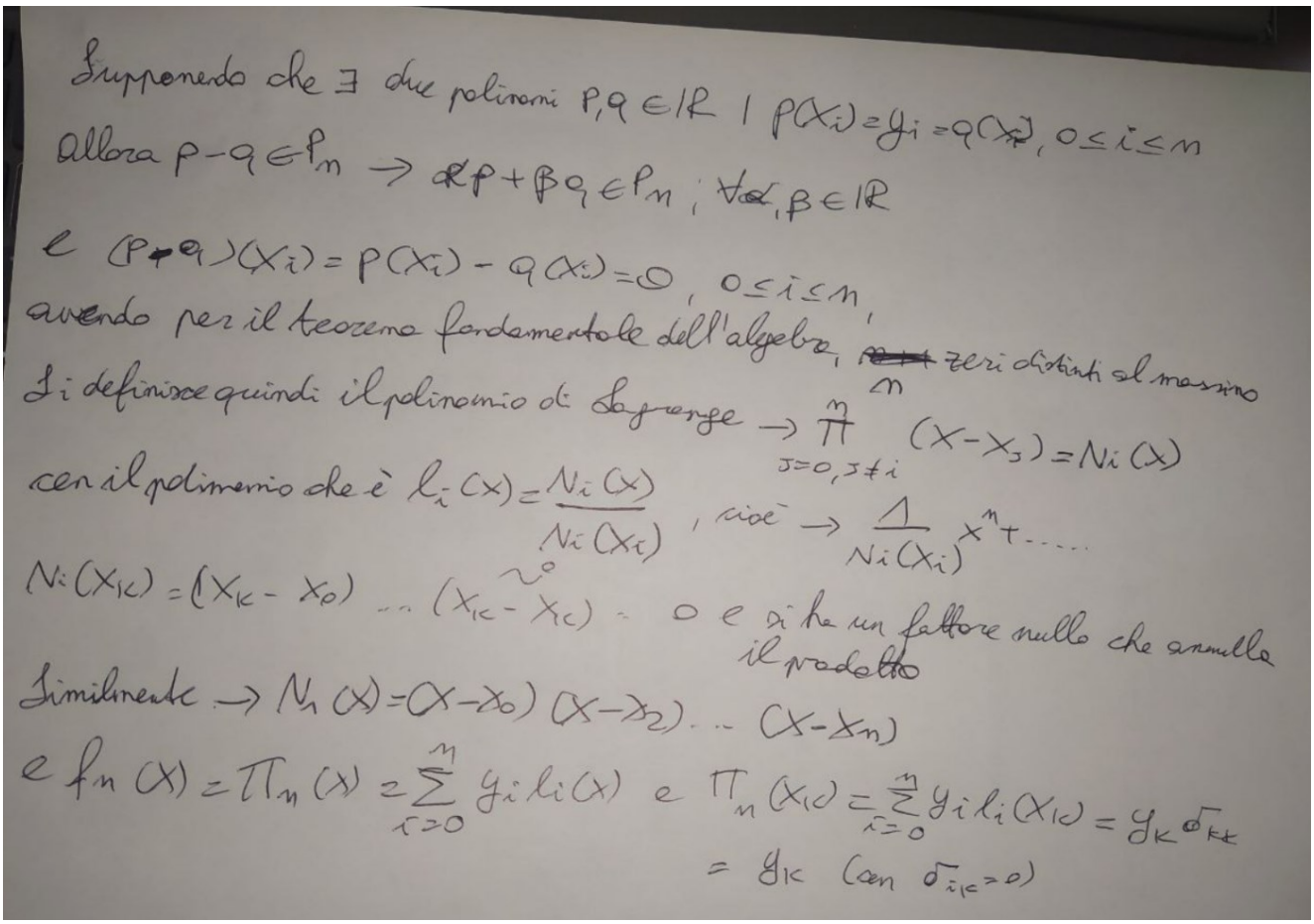
Possibili risposte a varie domande d'esame che ho trovato e messo qui su Mega; esse stanno in mezza facciata; al lettore come organizzarle e disporle. Si propone una trattazione completa dai soliti Appunti presenti sempre qui su Mega.

- **Perché il metodo di Newton per zeri semplici ha ordine di convergenza almeno 2? Quando ha ordine esattamente 2? (si dimostri la relazione fondamentale che lega e_{n+1} ed e_n)**

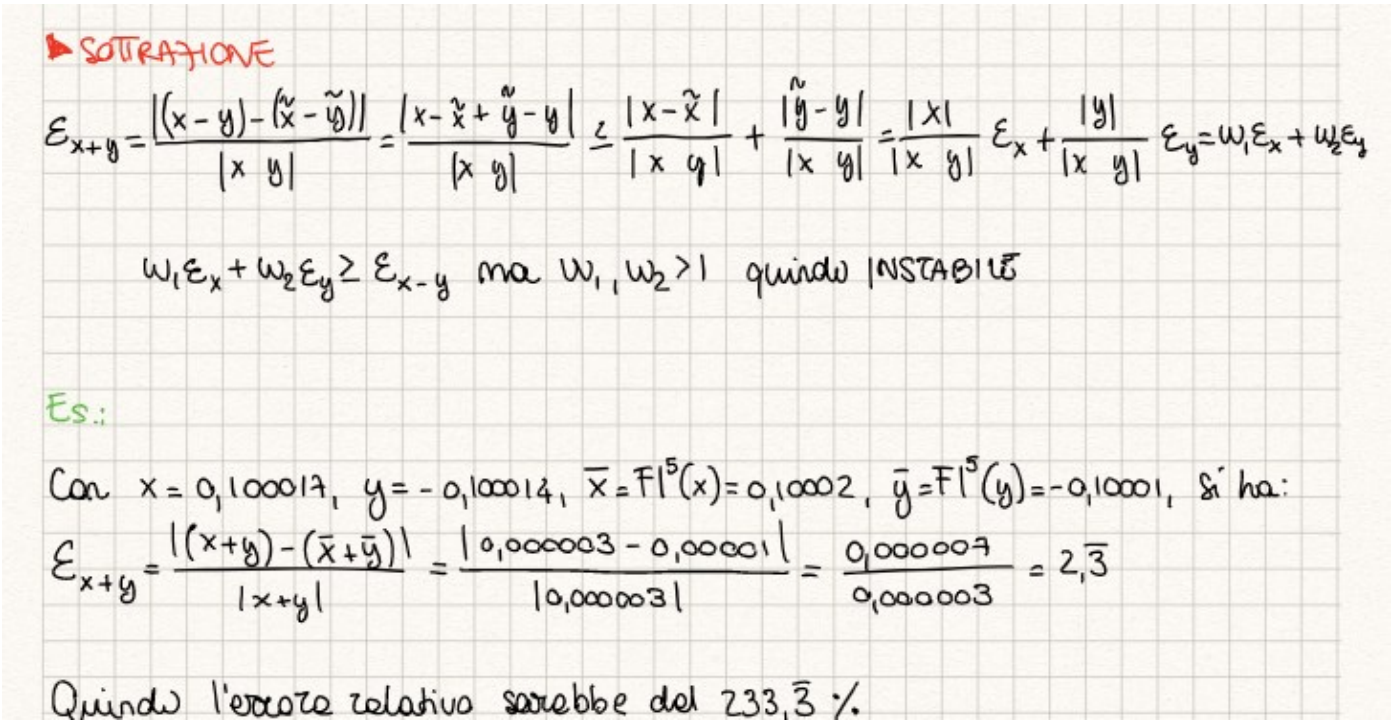


- **Perché il polinomio interpolatore di grado $\leq n$ su $n+1$ nodi distinti è unico? Si faccia un esempio in cui ha grado $< n$**

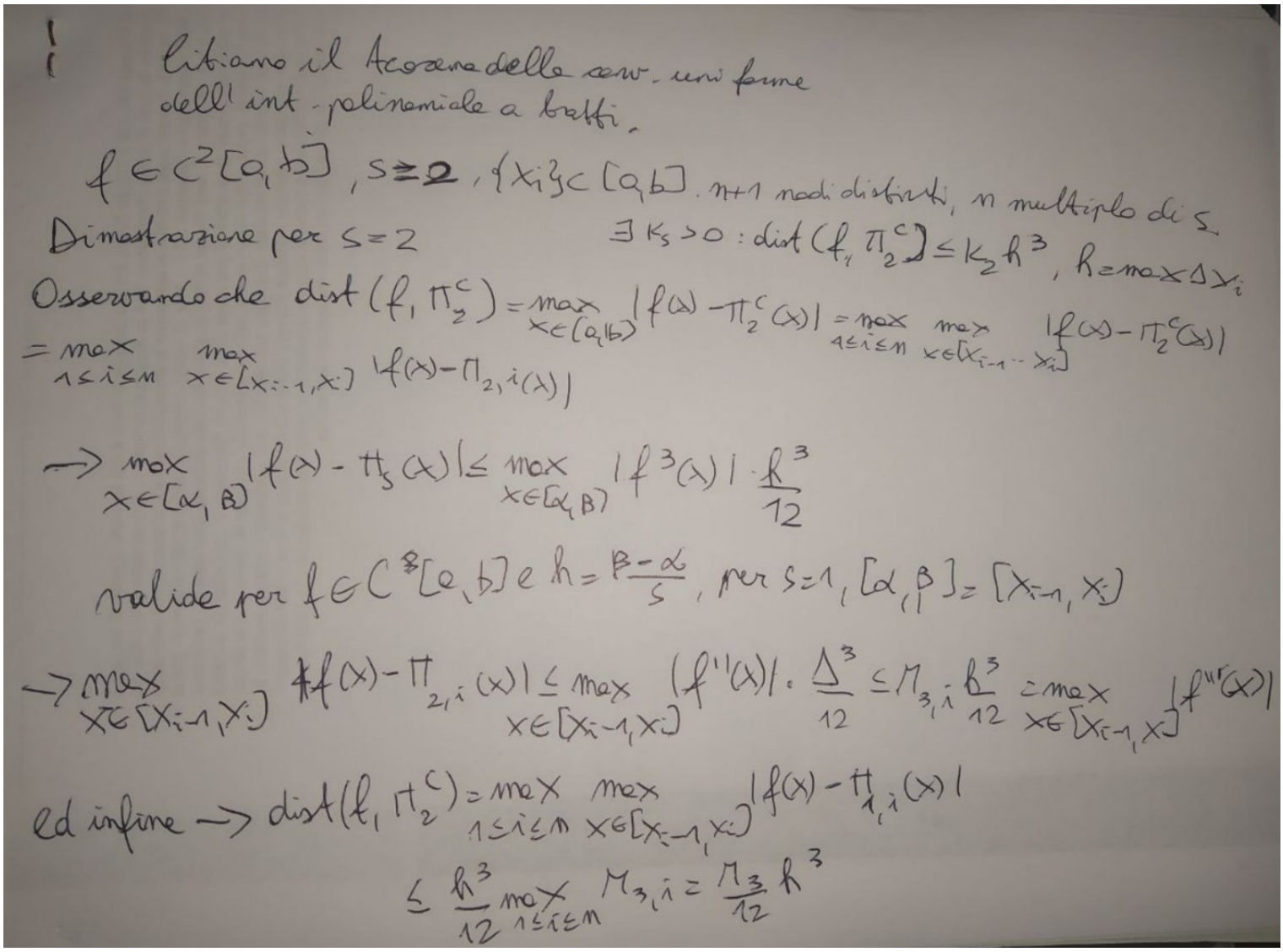


- **Perché la sottrazione tra numeri approssimati può essere instabile? (si ricavi la stima dell'errore e si faccia un esempio)**

La sottrazione può essere instabile per il fatto che solitamente |x| ≠ |y| e dunque può normalmente succedere che |x+y| < |x| o |x+y| < |y| e max {w1, w2} > 1.
Quando w_1, w_2 sono >> 1, la sottrazione è instabile.
Più nel dettaglio:



- **Perché l'interpolazione lineare a tratti a passo costante converge uniformemente con errore $O(h^2)$ se $f \in C^2[a,b]$? (si ricavi una stima dell'errore).**



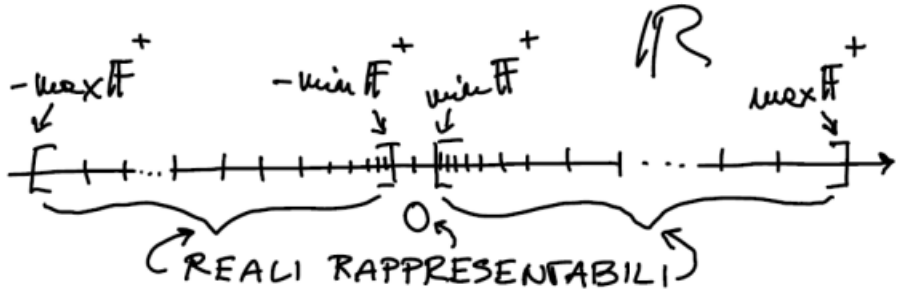
- **Che cos'è la precisione di macchina in un sistema floating-point $F(b,t,L,U)$ e come si calcola? (si ricavi il valore)**

Ricordiamo che i reali-macchina in base b a t cifre di mantissa e con range di esponenti $[L,U] \subset \mathbb{Z}$ sono definiti da

$$\mathbb{F}(b,t,L,U) = \{\mu \in \mathbb{Q} : \mu = sign(\mu)\cdot(0,\mu_1\mu_2\dots\mu_t)\cdot b^p\,,\ \mu_j \in \{0,1,\dots,b-1\}\,,\ \mu_1 \neq 0\,,\ p \in [L,U] \subset \mathbb{Z}\}$$

dove μ_j, $j = 1,\dots,t$ sono le cifre della mantissa e p l'esponente intero della potenza della base che sposta la virgola dove $L < 0$ e $U > 0$, quindi p appartiene all'intervallo di interi $L, L+1,\dots,-1,0,1,\dots,U-1,U$.

Rappresentato schematicamente si ha:



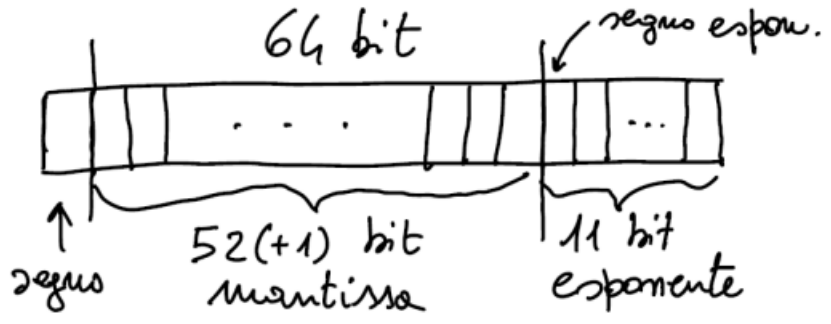
I reali macchina corrispondono all'insieme finito di tacche sull'asse reale e stanno nell'unione di due intervall simmetrici che rappresentano tutti i reali approssimabili:

$$\mathbb{F}^+ \subset [\min \mathbb{F}^+, \max \mathbb{F}^+]$$

$$\mathbb{F}^- \subset [-\max \mathbb{F}^+, -\min \mathbb{F}^+]$$

sapendo che se eccediamo a – max F+ si ha un overflow altrimenti se eccediamo max F+ si ha un overflow

Partendo da questa descrizione, si usa un sistema di rapprsentazione a 64 bit, con questo disegno:



In ciascuna delle caselline (bit) si può scrivere 0 oppure 1 (per il segno uno dei due rappresenta "+" e l'altro "-"). Per quanto riguarda la mantissa i bit sono 52 ma è come se fossero 53, perchè la prima cifra di mantissa deve essere non nulla e quindi per forza 1 (ovvero il processore tratta i numeri come se avessero mantissa $0,1d_2\dots d_{53}$ con $d_j \in \{0,1\}$, $2 \le j \le 53$) quindi la precisione di macchina è

$$\varepsilon_M = \frac{2^{1-53}}{2} = 2^{-53} \approx 10^{-16}$$

• Derivazione numerica col rapporto incrementale simmetrico

Innanzitutto definiamo il rapporto incrementale simmetrico: $\delta(h) = \dfrac{f(x+h) - f(x-h)}{2h}$

ottenuto scrivendo la formula di Taylor "da destra" e "da sinistra":

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

dove $\xi \in (x, x+h)$ e $\eta \in (x-h, x)$ da cui si ottiene, sottraendo membro a membro

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

usando il rapporto incrementale simmetrico "perturbato", usando la stima:

Ora

$$\begin{aligned}
|\delta(h) - \tilde{\delta}(h)| &= \frac{1}{2h}|f(x+h) - f(x-h)| - |\tilde{f}(x+h) - \tilde{f}(x-h)| \\
&= \frac{1}{2h}|(f(x+h) - \tilde{f}(x+h)) + (\tilde{f}(x-h) - f(x-h))| \\
&\le \frac{1}{2h}(|f(x+h) - \tilde{f}(x+h)| + |\tilde{f}(x-h) - f(x-h)|) \\
&\le \frac{1}{2h}(\varepsilon + \varepsilon) = \frac{2\varepsilon}{2h} = \frac{\varepsilon}{h}
\end{aligned}$$

Otteniamo quindi

$$|f'(x) - \tilde{\delta}(h)| \le dh^2 + \frac{\varepsilon}{h} = E(h)$$

Notando in ultimo che si può migliorare rispetto all'errore minimale qui a destra:

$$\begin{aligned}
E(h) &= dh^2 + \frac{\varepsilon}{h} \\
E'(h) &= \left(dh^2 + \frac{\varepsilon}{h}\right)' \\
&= 2dh - \frac{\varepsilon}{h^2} = 0 \Rightarrow h^3 = \frac{\varepsilon}{2d} \\
&\Rightarrow h^* = h^*(\varepsilon) = \left(\frac{\varepsilon}{2d}\right)^{\frac{1}{3}}
\end{aligned}$$

Inoltre $E''(h) = 2d + \frac{2\varepsilon}{h^3} > 0$ quindi $E(h)$ è convessa e h^* è di minimo. D'altra parte

$$\begin{aligned}
E(h^*) &= d(h^*)^2 + \frac{\varepsilon}{h^*} \\
&= d\left(\frac{\varepsilon}{2d}\right)^{\frac{2}{3}} + \varepsilon\left(\frac{2d}{\varepsilon}\right)^{\frac{1}{3}} \\
&= 2^{-2/3} \cdot d^{1/3} \cdot \varepsilon^{2/3} + (2d)^{1/3} \cdot \varepsilon^{2/3} \\
&= d^{1/3} \cdot (2^{-2/3} + 2^{1/3}) \cdot \varepsilon^{2/3}
\end{aligned}$$

cioè

$$h^* = O(\varepsilon^{1/3}) \quad \text{e} \quad E(h^*) = O(\varepsilon^{2/3})$$

• Perché moltiplicazione e divisione tra numeri approssimati sono operazioni stabili? Si ricavino le stime dell'errore

Per la moltiplicazione otteniamo:

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

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 $\tilde{x}y$

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

e usando la disuguaglianza triangolare:

Quindi da ($\bigstar$) otteniamo (ricordando che il modulo del prodotto è il prodotto dei moduli)

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

Questo perchè $\frac{|x-\tilde{x}|}{|x|} = \varepsilon_x$ e $\frac{|y-\tilde{y}|}{|y|} = \varepsilon_y$.
Ora, siccome $\tilde{x} \approx x$, possiamo dire almeno qualitativamente che $\frac{|\tilde{x}|}{|x|} \approx 1$ e quindi

$$\varepsilon_{xy} \leq \varepsilon_x + \frac{|\,\tilde{x}\,|}{|\, x\,|}\varepsilon_y \approx \varepsilon_x + \varepsilon_y$$

cioè che l'operazione di moltiplicazione è STABILE (l'errore relativo sul risultato è maggiorato da una quantità che è dell'ordine dell'errore sui dati).

Per la divisione:

Passiamo ora alla DIVISIONE: siccome la divisione $\frac{x}{y}$, $y \neq 0$ è la moltiplicazione per il reciproco, $\frac{x}{y} = x \cdot \frac{1}{y}$, ci basta analizzare la stabilità dell'operazione di reciproco

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

con l'ipotesi qualitativa che $|y| \approx |\tilde{y}|$ e quindi $\frac{|y|}{|\tilde{y}|} \approx 1$ ne deduciamo che anche che la DIVISIONE è un'operazione STABILE, perchè il reciproco è stabile e la moltiplicazione è stabile.
Anche in questo caso possiamo però quantificare, stimando meglio $\frac{|y|}{|\tilde{y}|}$.

sarà vero in tutte le situazioni "ragionevoli", visto che tipicamente l'errore sarà molto più piccolo di 1), allora

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

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

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

perché $\varepsilon_y^2 \ll \varepsilon_y < 1$ (per la prima volta usiamo qui il simbolo "$\ll$" molto minore)
Alla fine otteniamo

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

cioè abbiamo quantificato in modo più preciso la stima qualitativa

$$\frac{|\,y\,|}{|\,\tilde{y}\,|} \approx 1$$

- **Perché il residuo non pesato può non essere una buona stima dell'errore nel metodo di bisezione? (si ricavi la stima del residuo pesato in modo rigoroso).**

Volendo fare una stima del residuo: $f(x_n) \to f(\xi) = 0, \quad n \to \infty$

se e solo se:

$$\forall \{x_n\} : \lim_{n\to\infty} x_n = l \text{ si ha } \lim_{n\to\infty} f(x_n) = f(l)$$

che si esprime anche dicendo "f è continua se e solo se il limite si può trasportare 'dentro' la funzione".
Nel nostro caso $f(\xi) = 0$ quindi $f(x_n) \to 0$, $n \to \infty$ e anche $|f(x_n)| \to 0$, $n \to \infty$.
La quantità $|f(x_n)|$ si chiama "RESIDUO" perchè dice quanto "resta" ad f per annullarsi.
Viene allora spontanea questa domanda: siccome $f(x_n) \to 0$, $n \to \infty$, possiamo arrestare il processo di calcolo quando il residuo $|f(x_n)|$ è piccolo? In altre parole

$$|f(x_n)| \leq \varepsilon \overset{?}{\Longrightarrow} e_n \leq \varepsilon$$

La risposta è NO, in realtà

$$|f(x_n)| \leq \varepsilon \not\Rightarrow e_n \leq \varepsilon$$

e dunque usando il teorema del valor medio per "pesare" opportunamente la velocità di variazione nonché il teorema di permanenza del segno (sotto ipotesi che lo zero sia semplice, dunque che f'(ξ) ≠ 0.
Dunque volendo ricavare la stima pesata rigorosa:

DIMOSTRAZIONE della rappresentazione utilizzando il teorema del valor medio e supponendo che $x_n > \xi$ (l'altro caso è del tutto analogo), con $\alpha = \xi, \beta = x_n$

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

con $f(\xi) = 0$, cioè

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

che si può riscrivere come

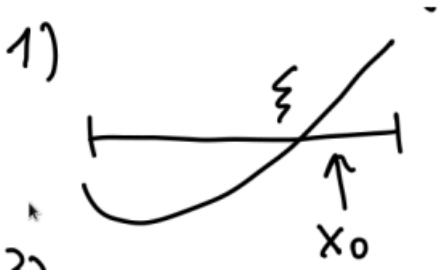
$$e_n = |x_n - \xi| = \frac{|f(x_n)|}{|f'(z_n)|}$$

■

che viene usato in varie stime pratiche.

- **Si dimostri la convergenza del metodo di Newton nel caso strettamente convesso o concavo**

Nel caso strettamente convesso, rappresentato graficamente da:



sapendo che f' può cambiare di segno e similmente f'' per il teorema degli zeri.
Si procede con una dimostrazione per induzione, partendo da:

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

Ma se $x_n \in (\xi, b]$ allora $f'(x_n) > 0$ e $f(x_n) > 0$, quindi x_{n+1} si ottiene da x_n sottraendo una quantità > 0, cioè $x_{n+1} < x_n$.
D'altra parte f è strettamente convessa, il che è equivalente a dire che la tangente sta "sotto al grafico" $\forall x \in [a, b]$.

E quindi:

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

Concludendo si passa al limite della formula che definisce il metodo:

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

e quindi:

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

• 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)

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

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

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

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

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

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

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

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

Quindi

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

e perciò

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

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:

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

• Sistema delle equazioni normali per l'approssimazione polinomiale ai minimi quadrati

Si parte da una premessa:

> Dati N punti $\{(x_i, y_i)\}$, $y_i = f(x_i)$, $1 \le i \le N$ 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 $V^tVa = V^ty$.
> Dove V^t è la trasposta di V.

osservando che poi il sistema ha dimensione (m + 1) x (m + 1):

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

e quindi:

$$V^tV \in \mathbb{R}^{(m+1)\times(m+1)} \text{ e } V^ty \in \mathbb{R}^{m+1}$$

Ora per la dimostrazione si deve dimostrare che *a* è di minimo assoluto:

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

ma

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

usando le proprietà del prodotto scalare, in particolare:

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

usata per scrivere: $\underset{\| \\ (Va-y,\,Vb)_N}{(Va - y,\, Vb)} = \underset{\| \\ (V^t(Va-y),\,b)_{m+1}}{(V^t(Va - y),\, b)}$

A questo punto:

- Dimostriamo per prima l'implicazione "⇐": assumendo che $V^tVa = V^ty$ abbiamo che:

$$V^tVa - V^ty = V^t(Va - y) = 0 \quad \text{e} \quad (V^t(Va - y),\, b) = \underset{\| \\ \text{vettore nullo in } \mathbb{R}^{m+1}}{(0, b)} = 0$$

da cui:

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

e poi:

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

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 $\varepsilon \to 0$ la disuguaglianza viene mantenuta, ottenendo:

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

Ma se vale $\forall$ versore, possiamo prendere $-v$ al posto di v e otteniamo:

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

avendo come unico vettore ortogonale il vettore nullo.

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

Ovvero a è soluzione del sistema delle equazioni normali.

- **Si mostri qual'e' l'idea del metodo di estrapolazione e si faccia un esempio di applicazione.**

Usando il calcolo approssimato della derivata con rapporti incrementali,

Chaimando $\phi(h)$ una di queste formule e $\tilde{\phi}(h)$ la formula in cui si usano i dati realmente misurati, cioè valori $\tilde{f}(t)$ con $|\tilde{f}(t) - f(t)| \leq \varepsilon$, abbiamo visto che

$$\left|\tilde{\phi}(h) - f'(x)\right| = O(h^p) + O\left(\frac{\varepsilon}{h}\right)$$

vedendo contemporaneamente una stima a posteriori dell'errore ed un aumento a costo computazionale basso dell'ordine di infinitesimo dell'errore.
Concentriamoci su uno dei due casi, dunque sull'aumento dell'ordine di infinitesimo dell'errore, partendo da una struttura asintotca che considera l'ordine $O(h^2)$ ottenuto dalle operazioni con la derivazione:

$$\delta_+(h) = f'(x) + \frac{f''(x)}{2}h + O(h^2)$$
$$\delta_+\left(\frac{h}{2}\right) = f'(x) + \frac{f''(x)}{2}\frac{h}{2} + O(h^2)$$

Si sfrutta nuovamente la struttura asintotica per eliminare la parte principale, arrivando ad una formula di ordine superiore come infinitesimo:

$$\delta_+(h) = f'(x) + \frac{f''(x)}{2}h + O(h^2)$$
$$2\delta_+\left(\frac{h}{2}\right) = 2f'(x) + \frac{f''(x)}{2}2\frac{h}{2} + O(h^2)$$

quindi

$$\begin{aligned} 2\delta_+\left(\frac{h}{2}\right) - \delta_+(h) &= 2f'(x) + \cancel{\frac{f''(x)}{2}}h + O(h^2) - \left(f'(x) + \cancel{\frac{f''(x)}{2}}h + O(h^2)\right) \\ &= 2f'(x) - f'(x) + O(h^2) - O(h^2) \\ &= f'(x) + O(h^2) \end{aligned}$$

(si noti che abbiamo usato il fatto che $2O(h^2) = O(h^2) : |u(h)| \leq \gamma h^2 \Rightarrow |k \cdot u(h)| \leq k \cdot \gamma h^2$ se k è una costante).
In definitiva:

$$\phi_1(h) = 2\delta_+\left(\frac{h}{2}\right) - \delta_+(h) = f'(x) + O(h^2)$$

cioè con una semplice speciale combinazione lineare delle formule con passo h e $\frac{h}{2}$ abbiamo ricavato una nuova formula con errore $O(h^2)$ invece di $O(h)$.

Quanto descritto è basato su una struttura asintotica precisa; questa è l'estrapolazione.

• Condizionamento di matrici e sistemi lineari

La domanda fa riferimento alle proposizioni 1/2 della Lezione 20:

Perturbazione del termine noto

Sia $A \in \mathbb{R}^{n\times n}$ una matrice non singolare, $x \in \mathbb{R}^n$ soluzione del sistema $Ax=b$, $b\neq 0$ e $\tilde{x} = x + \delta x$ soluzione del sistema perturbato $A\tilde{x} = \tilde{b} = b + \delta b$

Fissata una norma vettoriale $\|\cdot\|$ in $\mathbb{R}^n$ vale la seguente stima dell'errore relativo su x: $\frac{\|\delta x\|}{\|x\|} \leq k(A)\frac{\|\delta b\|}{\|b\|}$ dove $k(A) = \|A\|\cdot\|A^{-1}\|$.

Dim.

Data che $x = A^{-1}b \neq 0$ calcoliamo l'errore relativo in norma.

$\tilde{x} = x + \delta x = A^{-1}b + A^{-1}\delta b \Rightarrow \|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\cdot\|\delta b\|$.

Stimiamo quindi da sopra $\frac{1}{\|x\|}$, cioè da sotto $\|x\|$.

Essendo x la soluzione: $\|b\| = \|Ax\| \leq \|A\|\cdot\|x\|$ da cui $\|x\| \geq \frac{\|b\|}{\|A\|}$ e $\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$

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

Perturbazione sulla matrice

Sia $A \in \mathbb{R}^{n\times n}$ una matrice non singolare, $x \in \mathbb{R}^n$ soluzione del sistema $Ax=b$, $b\neq 0$ e $\tilde{x} = x + \delta x$ soluzione del sistema perturbato $\tilde{A}\tilde{x} = b$, $\tilde{A} = A + \delta A$

Fissata una norma vettoriale $\|\cdot\|$ in $\mathbb{R}^n$ vale la seguente stima dell'errore relativo su x: $\frac{\|\delta x\|}{\|\tilde{x}\|} \leq k(A)\cdot\frac{\|\delta A\|}{\|A\|}$ dove $k(A) = \|A\|\cdot\|A^{-1}\|$.

Dim.

Da $\tilde{A}\tilde{x} = (A + \delta A)(x + \delta x) = b$ otteniamo:

$Ax + A\delta x + \delta A\tilde{x} = b$ cioè:

$\delta x = A^{-1}(-\delta A\tilde{x}) = -A^{-1}\delta A\tilde{x}$ quindi:

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

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

Altre domande possibili (stando al 20/21):

1) Applicazione del MEG nel calcolo del determinante.

Innanzitutto diciamo che il MEG è basato su due proprietà fondamentali in merito a trasformazioni della matrice. Esse sono:

- la sostituzione alla riga k della somma della riga k con la riga i moltiplicata per uno scalare
- lo scambio di due righe porta il determinante a cambiare segno

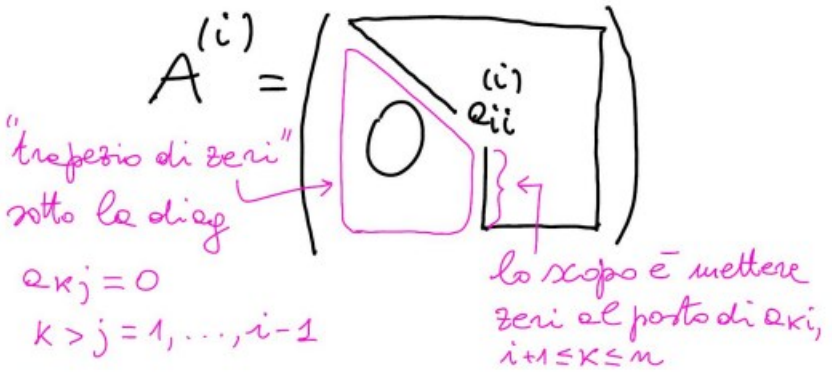
Esso si applica ripetutamente per mettere zeri in ogni colonna sotto la diagonale principale, in maniera tale da ottenere una matrice triangolare superiore.
Partendo ad esempio da una matrice A 3x3 (a destra) si arriva facilmente alla matrice U (a sinistra):

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 2 & 3 \\ -1 & -3 & 0 \end{pmatrix} \qquad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

Questo afferma che il MEG consiste in una sequenza di n-1 trasformazioni:

$$A^{(1)} = A \to A^{(2)} \to A^{(3)} \to \ldots \to A^{(n)} = U$$

ottenendo una matrice U finale che è triangolare superiore con la seguente struttura schematica:



Se $a_{ii}^{(i)} \neq 0$ (cioè se l'elemento diagonale di $A^{(i)}$ è non nullo) si può propagare a destra la struttura del trapezio di zeri con le trasformazioni

$$\mathcal{R}_k^{(i+1)} := \mathcal{R}_k^{(i)} + \left(-\frac{a_{ki}^{(i)}}{a_{ii}^{(i)}}\right) \cdot \mathcal{R}_i^{(i)}, \ i+1 \leq k \leq n$$

che mettono lo zero al posto k, i visto che

$$a_{ki}^{(i+1)} = a_{ki}^{(i)} + \left(-\frac{a_{ki}^{(i)}}{a_{ii}^{(i)}}\right) \cdot a_{ii}^{(i)} = 0$$

2) MEG e relazione con la fattorizzazione LU.

Per alcune classi di matrici, come quelle a diagonale strettamente dominante e le simmetriche definite positive, ciò già vale. Partendo dall'esempio di matrici 3 x 3, definiamo U (ottenuta alla fine del MEG):

$$U = A^{(3)} = T_{3,2}(-m_{3,2})\ T_{3,1}(-m_{3,1})\ T_{2,1}(-m_{2,1})\ A$$

Posto $\mathcal{L} = T_{3,2}(-m_{3,2})\ T_{3,1}(-m_{3,1})\ T_{2,1}(-m_{2,1})$ abbiamo

$$U = \mathcal{L}A \Rightarrow A = LU \text{ con } L = \mathcal{L}^{-1}$$

Avendo L che ha ordine invertito:

$$L = \mathcal{L}^{-1} = \overbrace{(T_{2,1}(-m_{2,1}))^{-1}\ (T_{3,1}(-m_{3,1}))^{-1}\ (T_{3,2}(-m_{3,2}))^{-1}}^{\text{ordine invertito}} = T_{2,1}(m_{2,1})\ T_{3,1}(m_{3,1})\ T_{3,2}(m_{3,2})$$

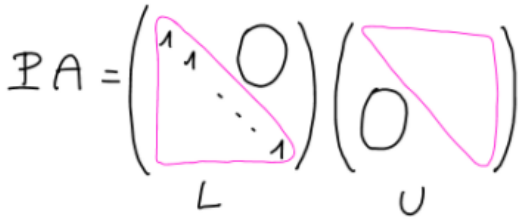
ricordando che $(T_{k,i}(\alpha))^{-1} = T_{k,i}(-\alpha)$

ed L è matrice triangolare inferiore (Lower Triangular): $L = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix}$

Otteniamo quindi una fattorizzazione PA = LU con:

- U matrice triangolare superiore
- P matrice di permutazione ottenuta come prodotto di matrici di scambio
- L matrice triangolare inferiore e con i moltiplicare rimescolati nel triangolo inferiore sotto la diagonale.

In sintesi grafica, si ha una fattorizzazione di costo dato dagli scambi e dai moltiplicatori usati, che sarà $2/3*n^3$:



3) Calcolo della matrice inversa.

Si parte da una proprietà del prodotto matrice-vettore, interpretabile come combinazione lineare delle colonne di B tramite i coefficienti di C:

$$z = Bc = c_1 \underbrace{C_1(B)}_{\text{colonna 1 di } B} + \cdots + c_n \underbrace{C_n(B)}_{\text{colonna } n \text{ di } B}$$

Applichiamo tale osservazione con c = e^i = (0 … 010 … 0)' ottenendo $Be^i = C_i(B)$.

Allora per $B = A^{-1}$

$$C_i(A^{-1}) = A^{-1}e^{(i)} \Leftrightarrow AC_i(A^{-1}) = e^{(i)}$$

cioè $C_i(A^{-1})$ è la soluzione di

$$Ax^{(i)} = e^{(i)},\ 1 \leq i \leq n$$

Il calcolo dell'inversa procede per colonne risolvendo M=n sistemi lineari sulla matrice A e con terminte noto n variabile. Usando il calcolo standard si avremmo un costo $2/3*n^4$ flops, invece usando MEG con la fattorizzazione LU e risolvendo le M=n coppie di sistemi triangolari si ha un costo $8/3*n^3$

4) Sistemi sovradeterminati.

Essi sono sistemi Ax=b su A in $R^{m \times n}$ e b in R^m con m > n (più equazioni che incognite):

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_n \end{pmatrix}$$

Essi hanno soluzione se e solo se b sta nel sottospazio di R^m generato dalle colonne di A:

$$b \in \langle C_1(A), \dots, C_n(A) \rangle \subset \mathbb{R}^m$$

Vogliamo quindi fare in modo che esista una soluzione x tale che $dist_2(b, Ax) = ||b - Ax||_2 = 0$
Nuovamente abbiamo un problema di minimo ϕ(x) se e solo se ϕ(x+z) >= ϕ(x).

Ora

$$\begin{aligned} \phi(x+z) &= \|b - A(x+z)\|_2^2 \\ &= (b - A(x+z), b - A(x+z)) \quad \longleftarrow \text{prod. scalare in } \mathbb{R}^m \\ &= (b - Ax - Az, b - Ax - Az) \\ &= (b - Ax, b - Ax) - 2(Az, b - Ax) + (Az, Az) \\ &= \phi(x) + 2(z, A^t(Ax - b)) + \|Az\|_2^2 \end{aligned}$$

- "⇑" Se $A^tAx = A^tb$ allora

$$\phi(x+z) = \phi(x) + \overbrace{\|Az\|_2^2}^{\geq 0} \geq \phi(x) \ \forall z$$

  cioè $\phi(x)$ è minimo

- "⇓" Se $\phi(x)$ è un minimo allora $\forall \varepsilon > 0$ e $\forall v \in \mathbb{R}^n$, $\|v\|_2 = 1$

$$\phi(x + \varepsilon v) = \phi(x) + 2(\varepsilon v, A^t(Ax - b)) + \|\varepsilon v\|_2^2 \geq \phi(x)$$

  cioè

$$2\varepsilon(v, A^t(Ax - b)) + \varepsilon^2 \geq 0$$

  e dividendo per ε

$$2(v, A^t(Ax - b)) + \varepsilon \geq 0 \ \forall \varepsilon, v$$

  da cui per $\varepsilon \to 0$

$$(v, A^t(Ax - b)) \geq 0 \ \forall v$$

  Ma allora prendendo $-v$

$$(-v, A^t(Ax - b)) \geq 0 \ \forall v$$

  cioè

$$(v, A^t(Ax - b)) \leq 0 \ \forall v$$

  e quindi

$$(v, A^t(Ax - b)) = 0 \ \forall v$$

  da cui $A^t(Ax - b) = 0$ perché l'unico vettore ortogonale a tutti i versori è il vettore nullo, cioè $A^tAx = A^tb$

In conclusione se A ha rango max = n la soluzione ai minimi quadrati del sistema sovradeterminato è unica.

6) Si ricavi una stima dell'errore relativo sulla soluzione di un sistema lineare non singolare in caso di vettore termine noto effetto da errori.

Sia $A \in \mathbb{R}^{n\times n}$ una matrice non singolare, $x \in \mathbb{R}^n$ soluzione del sistema $Ax = b$, $b \neq 0$ e $\tilde{x} = x + \delta x$ soluzione del sistema perturbato $A\tilde{x} = \tilde{b} = b + \delta b$.
Fissata una norma vettoriale $\|\cdot\|$ in $\mathbb{R}^n$, vale la seguente stima dell'errore "relativo" su x

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

Dimostrazione
Osserviamo che $x = A^{-1}b \neq 0$ quindi ha senso stimare l'errore relativo in norma (cioè l'errore assoluto $\|\delta x\|$ diviso per la "lunghezza" di x, cioè $\|x\|$).
Ora

$$\tilde{x} = x + \delta x = A^{-1}\tilde{b} = A^{-1}(b + \delta b) = \cancel{A^{-1}}b + A^{-1}\delta b$$

da cui otteniamo la stima dell'errore assoluto

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

Per stimare l'errore relativo dobbiamo stimare da sopra $\frac{1}{\|x\|}$, cioè da sotto $\|x\|$.
Siccome x è la soluzione

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

da cui

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

e

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

perciò

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

■