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)

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

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

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

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

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

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

Perché moltiplicazione e addizione sono operazioni stabili? Si ricavino le stime dell'errore

Dati $x, y \in \mathbb{R}$ e $*$ operazione binaria fondamentale, si ha che in $F(b,t,L,U)$ l'operazione risulta: $x * y = fl^t(fl^t(x) * fl^t(y))$

Dati $x, y \neq 0$, gli errori relativi su tali quantità saranno rispettivamente $\varepsilon_x = \frac{|x - \tilde{x}|}{|x|}$ e $\varepsilon_y = \frac{|y - \tilde{y}|}{|y|}$ mentre quello sul risultato dell'operazione sarà: $\varepsilon_{x*y} = \frac{|x*y - \tilde{x}*\tilde{y}|}{|x*y|}$

Un'operazione è stabile se l'ordine di grandezza di ε_{x*y} è vicino a quello di ε_x e ε_y

- MOLTIPLICAZIONE

$$\varepsilon_{xy} = \frac{|xy - \tilde{x}\tilde{y}|}{|xy|} = \frac{|xy - \tilde{x}y + \tilde{x}y - \tilde{x}\tilde{y}|}{|xy|} = \frac{|(x-\tilde{x})y + (y-\tilde{y})\tilde{x}|}{|xy|} \leq \frac{|(x-\tilde{x})y| + |(y-\tilde{y})\tilde{x}|}{|xy|} =$$

$$= \frac{|x-\tilde{x}||y|}{|x||y|} + \frac{|y-\tilde{y}||\tilde{x}|}{|x||y|} = \varepsilon_x \frac{\cancel{|y|}}{\cancel{|y|}} + \varepsilon_y \frac{|\tilde{x}|}{|x|} \Rightarrow \varepsilon_{xy} \approx \varepsilon_x + \varepsilon_y \quad \text{STABILE}$$

▶ ADDIZIONE

$$\varepsilon_{x+y} = \frac{|(x+y)-(\tilde{x}+\tilde{y})|}{|x+y|} = \frac{|x-\tilde{x}+y-\tilde{y}|}{|x+y|} \le \frac{|x-\tilde{x}|}{|x+y|} + \frac{|y-\tilde{y}|}{|x+y|} = \frac{|x|}{|x+y|}\varepsilon_x + \frac{|y|}{|x+y|}\varepsilon_y = w_1\varepsilon_x + w_2\varepsilon_y$$

$w_1\varepsilon_x + w_2\varepsilon_y \ge \varepsilon_{x+y}$ ma $w_1, w_2 \le 1$ quindi STABILE

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

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

<u>Perché la sottrazione tra numeri approssimati può essere instabile? (si ricavi la stima dell'errore e si faccia un esempio)</u>

▶ SOTTRAZIONE

$$\varepsilon_{x+y} = \frac{|(x-y)-(\tilde{x}-\tilde{y})|}{|x\ \ y|} = \frac{|x-\tilde{x}+\tilde{y}-y|}{|x\ \ y|} \le \frac{|x-\tilde{x}|}{|x\ \ y|} + \frac{|\tilde{y}-y|}{|x\ \ y|} = \frac{|x|}{|x\ \ y|}\varepsilon_x + \frac{|y|}{|x\ \ y|}\varepsilon_y = w_1\varepsilon_x + w_2\varepsilon_y$$

$w_1\varepsilon_x + w_2\varepsilon_y \ge \varepsilon_{x-y}$ ma $w_1, w_2 > 1$ quindi INSTABILE

Es.:

Con $x = 0{,}100017$, $y = -0{,}100014$, $\bar{x} = Fl^5(x) = 0{,}10002$, $\bar{y} = Fl^5(y) = -0{,}10001$, si ha:

$$\varepsilon_{x+y} = \frac{|(x+y)-(\bar{x}+\bar{y})|}{|x+y|} = \frac{|0{,}000003 - 0{,}00001|}{|0{,}000003|} = \frac{0{,}000007}{0{,}000003} = 2{,}\overline{3}$$

Quindi l'errore relativo sarebbe del $233{,}\overline{3}$ %

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

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

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

Convergenza 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 $\xi \in (a,b)$

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

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

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

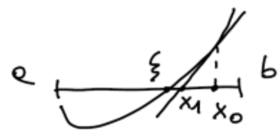
$$\begin{cases} y = 0 \\ y = f(x_n) + f'(x_n)(x - x_n) \end{cases} \Rightarrow \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Convergenza metodo di Newton:

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

e converge all'unico zero ξ di f in $[a,b]$

Si considera il caso:



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

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 ξ

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

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

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

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

Velocità (ordine) di convergenza del metodo di Newton

Applicando Taylor centrato in x_n e calcolata in ξ con il resto di Lagrange del secondo ordine:

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

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

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

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

ma c_n è succ. limitata:

$$c_n = \frac{\max|f''(x)|}{\min|f'(x)|}\cdot\frac{1}{2} = c \ \ \forall x \in [c,d] \Rightarrow e_{n+1} \le c e_n^2 \ \text{ e } \lim \frac{e_{n+1}}{e_n^2} = c \ \text{ quindi}$$

il metodo converge con velocità quadratica

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)

Usando Taylor:

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

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

quindi:

$$x_{n+1} - \xi = \frac{f''(z_n)}{2f'(x_n)}(\xi - x_n)^2 \longrightarrow |x_{n-1} - \xi| = \frac{|f''(z_n)|}{|2f'(x_n)|}|\xi - x_n|^2 \quad \text{che diventa:}$$

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

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.

Ordine di convergenza delle iterazioni di punto fisso

Sia ξ punto fisso di $\phi \in C^p(I)$, $p \geq 1$ con intervallo I di $\mathbb{R}$, supponiamo di essere in ipotesi che garantiscano la convergenza a ξ di $x_{n+1} = \phi(x_n)$, $n \geq 0$ con $x_n \in I$, allora $\{x_n\}$ ha:

① ordine esattamente $p = 1 \iff 0 < |\phi'(\xi)| < 1$

② ordine esattamente $p > 1 \iff \phi^{(j)}(\xi) = 0$, $1 \leq j \leq p-1$ e $\phi^{(p)}(\xi) \neq 0$

Dim ①: $e_{n+1} = |\phi'(z_n)| e_n$ con $z_n \in \text{int}(\xi, x_n)$ per il teorema del valor medio

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

Dim ②: si utilizza la formula di Taylor di grado $p-1$ centrata in ξ con resto p-esimo in forma di Lagrange:

$$x_{n+1} = \phi(x_n) = \phi(\xi) + \phi'(\xi)(x_n - \xi) + \frac{\phi''(\xi)}{2}(x_n - \xi)^2 + \dots + \frac{\phi^{(p-1)}(\xi)}{(p-1)!}(x_n - \xi)^{p-1} + \frac{\phi^{(p)}(u_n)}{p!}(x_n - \xi)$$

con $u_n \in \text{int}(\xi, x_n)$ e $u_n \xrightarrow[n\to\infty]{} \infty$

Per dimostrare la condizione sufficiente "$\Leftarrow$" si applicano le condizioni $\phi^{(j)}(\xi) = 0$ per $1 \leq j \leq p-1$ e $\phi^{(p)}(\xi) \neq 0$. Inoltre $\phi(\xi) = \xi$. Risulta, quindi:

$x_{n+1} - \xi = \frac{\phi^{(p)}(u_n)}{p!}(x_n - \xi)^p$ che passando ai moduli diventa:

$$\frac{e_{n+1}}{e_n^p} = \frac{|\phi^{(p)}(u_n)|}{p!} \xrightarrow[n\to\infty]{} \frac{|\phi^{(p)}(\xi)|}{p!} \neq 0 \quad \text{quindi } \{x_n\} \text{ ha ordine } p$$

Per dimostrare la condizione necessaria "$\Rightarrow$" si suppone per assurdo che $\exists\, j < p$ t.c. $\phi^{(j)}(\xi) \neq 0$ e $\{x_n\}$ ha ordine p. Si avrebbe quindi, tramite Taylor:

$$\frac{e_{n+1}}{e_n^k} \to \frac{|\phi^{(k)}(\xi)|}{k!} = L' \neq 0 \quad \text{con } k = \min\{j < p : \phi^{(j)}(\xi) \neq 0\}$$

ma $\frac{e_{n+1}}{e_n^p} = \frac{e_{n+1}}{e_n^k} \cdot e_n^{k-p}$ e so che $\frac{e_{n+1}}{e_n^p} \to L \neq 0$

Quindi:

$\frac{e_{n+1}}{e_n^k} \cdot e_n^{k-p} \to \infty$ perché $\frac{e_{n+1}}{e_n^k} \to L'$ e $e_n^{k-p} \to \infty$ perché $e_n \to 0$ e $k - p < 0$

cioè $\frac{e_{n+1}}{e_n^p} \to \infty$, $n \to \infty$ che contraddice l'ipotesi dell'ordine finito

Esistenza del polinomio interpolatore di grado <= n su n+1 nodi distinti

ESISTENZA

Dati n+1 nodi distinti $\{x_i\}_{0\le i\le n}$ considero per ogni nodo fissato il polinomio elementare di Lagrange:

$\ell_i(x) = \frac{N_i(x)}{N_i(x_i)}$ con $N_i(x_i) = \prod_{j=0,\, j\neq i}^{n} (x - x_j) \neq 0$ e $N_i(x)$ polinomio grado n

Quindi $\ell_i(x)$ ha grado n e $\ell_i(x_k) = \delta_{ik} = \begin{cases} 0 & i\neq k \\ 1 & i=k \end{cases}$ (delta di Kronecker)

Possiamo quindi definire il polinomio interpolatore di Lagrange:

$f_n(x) = \Pi_n(x) = \sum_{i=0}^{n} y_i \ell_i(x)$ con $\Pi_n(x) \in \mathbb{P}_n$

Verifica dell'interpolazione:

$\Pi_n(x_k) = \sum_{i=0}^{n} y_i \ell_i(x_k) = \sum_{i=0}^{n} y_i \delta_{ik} = y_k \delta_{kk} = y_k$, $0\le k\le n$ $(\delta_{ik}=0)$

Unicità del grado del polinomio interpolatore di grado <= n su n+1 nodi distinti (esempio con grado < n)

UNICITA`

Supponiamo $\exists\, p,q \in \mathbb{P}_n$ t.c. $p(x_i) = y_i = q(x_i)$ con $0\le i\le n$.

Allora il polinomio $p-q \in \mathbb{P}_n$ e si ha:

$(p-q)(x_i) = p(x_i) - q(x_i) = 0$, $0\le i\le n$ cioè $p-q$ avrebbe $n+1$ zeri dist.

Ma $p-q$ può avere al massimo n zeri distinti, a meno che non sia un p. nullo. Quindi $(p-q)(x) = 0\ \forall x \Rightarrow p(x) = q(x)\ \forall x$

Es.:

Un esempio di polinomio di grado $< n$ che interpola $n+1$ punti si ha con $f(x) = x^2$ dove, prendendo $n \ge 3$ punti, si avrà sempre il polinomio interpolatore uguale a $f(x)$ e, quindi, di grado 2 che è quindi $< n$.

Perché l'interpolazione 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)

Siano $f \in C^{s+1}[a,b]$, $s \ge 0$ e $\{x_i\} \subset [a,b]$ $n+1$ nodi distinti, con n multiplo di s. Allora $\exists K_s > 0 : dist(f, \Pi_s^c) \le K_s h^{s+1}$, $h = \max \Delta x_i$

Dimostrazione per $s=1$ ($f \in C^2[a,b]$)

$dist(f, \Pi_1^c) = \max_{x\in[a,b]} |f(x) - \Pi_1^c(x)| = \max_{0\le i\le n-1} \max_{x\in[x_i,x_{i+1}]} |f(x) - \Pi_1^c(x)| =$

$= \max_{1\le i\le n} \max_{x\in[x_{i-1},x_i]} |f(x) - \Pi_{1,i}(x)|$

Si ricava quindi la stima dell'errore: $\max_{x\in[x_{i-1},x_i]} |f''(x)| \cdot \frac{h^2}{8} = M_{2,i} \frac{h^2}{8}$ da cui

$dist(f, \Pi_1^c) = \max_{1\le i\le n} \max_{i\in[x_{i-1},x_i]} |f(x) - \Pi_{1,i}(x)| \le \frac{h^2}{8} \max_{1\le i\le n} M_{2,i} = \frac{M_2}{8} h^2$

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

Sistema delle equazioni normali per approssimazione polinomiale/ Sistema delle equazioni normali per approssimazione ai minimi quadrati

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 $V^tVa = V^ty$

Dire che $a \in \mathbb{R}^{m+1}$ e' di minimo assoluto per $\phi(a)$ equivale a dire: $\phi(a+b) \ge \phi(a) \;\forall\, b \in \mathbb{R}^{m+1}$ ma $\phi(a+b) = \phi(a) + 2(V^t(Va-y), b) + (Vb, Vb)$

Dim "$\Leftarrow$": assumendo che $V^tVa = V^ty$ abbiamo:
$V^tVa - V^ty = V^t(Va - y) = 0$ e $(V^t(Vb-y), b) = (0, b)$ vettore nullo in $\mathbb{R}^{m+1} = 0$

Dim "$\Rightarrow$": assumendo che $\phi(a+b) \ge \phi(a) \;\forall\, b \in \mathbb{R}^{m+1}$ allora:
$\phi(a+b) = \phi(a) + 2(V^t(Vb-y), b) + (Vb, Vb) \ge \phi(a) \;\forall\, b$ cioè:
$2(V^t(Vb-y), b) + (Vb, Vb) \ge 0 \;\forall\, b$.
Ponendo $b = \varepsilon v$, con v vettore versore $((v,v)=1)$. Dividendo per $\varepsilon > 0$ e considerando $\varepsilon \to$ otteniamo:
$(V^t(Va-y), v) \ge 0 \;\forall v$
Ma essendo la disuguaglianza valida per ogni v, allora possiamo sostituire $-v$ a v e otteniamo:
$(V^t(Va-y), -v) = -(V^t(Va-y), v) \ge 0 \;\forall v$ cioè

$0 \le (V^t(Va-y), v) \le 0 \Rightarrow (V^t(Va-y), v) = 0 \;\forall v$

Ma essendo il vettore nullo l'unico vettore ortogonale a tutti i vettori:
$V^t(Va-y) = 0 \iff V^tVa = V^ty$ ovvero "a" soluzione del sistema

Derivazione numerica rapporto incrementale simmetrico

quindi è soluzione del sistema delle equazioni normali

DERIVAZIONE SIMMETRICA CON RAPPORTO INCREMENTALE SIMMETRICO

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

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

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

da cui si ottiene, sottraendo membro a membro

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

$$= f(x) + O(h^2)$$

con $|f'(x) - \delta(h)| = \frac{1}{12} |f'''(\xi) + f'''(\eta)| h^2 \le \frac{1}{12} (|f'''(\xi)| + |f'''(\eta)|) h^2 \le d h^2$

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

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

$$\delta(h) = \frac{f'(x+h) - f(x-h)}{2h}$$ è $O(h^2)$ per $f \in C^3(I_x)$

Rapporto incrementale standard

derivazione numerica standard col rapporto incrementale

Supponendo $f \in C^2(I_t)$, con I intervallo di derivazione, f derivabile, dato il rapporto incrementale:

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

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

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

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

Se $\tilde{\delta}_+$ approssima δ_+, si ha la seguente stima dell'errore:

$$|f'(x) - \tilde{\delta}_+(h)| \le |f'(x) - \delta_+(h)| + |\delta_+(h) - \tilde{\delta}(h)|$$

da cui

$$|\delta_+(h) - \tilde{\delta}_+(h)| = \frac{f(x+h) - f(x) - \tilde{f}(x+h) + \tilde{f}(x)}{h}$$

$$\le \frac{1}{h}\left((f(x+h) - \tilde{f}(x+h)) + (\tilde{f}(x) - f(x)) \right)$$

$$\le \frac{1}{h}\left(|f(x+h) - \tilde{f}(x+h)| + |\tilde{f}(x) - f(x)| \right)$$

$$\le \frac{2}{h} \|f - \tilde{f}\|_\infty$$

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

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

se $h \to 0$, $E(h) \to \infty$ 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 $f \in C^2$

Formule di quadratura composte

Per le formule di quadratura composte ci riconduciamo a due casi, facendo riferimento nel caso di calcolo di nodi equispaziati:

- Per s = 1 alla formula dei trapezi, in cui l'integrale viene approssimato alla somma delle aree dei trapezi lineari corrispondenti all'interpolazione lineare a tratti.

$$\begin{aligned} I_n^{trap}(f) &= I(\textstyle\prod_1^c) \\ &= \int_a^b \textstyle\prod_1^c(x)dx \\ &= \sum(\text{aree trapezi}) \\ &= \overbrace{\frac{h}{2}\cdot(f(x_0)+f(x_1))}^{\text{area trapezio 1}} + \overbrace{\frac{h}{2}\cdot(f(x_1)+f(x_2))}^{\text{area trapezio 2}} + \ldots + \\ &+ \underbrace{\frac{h}{2}\cdot(f(x_{n-2})+f(x_{n-1}))}_{\text{area trapezio (n-1)-esimo}} + \underbrace{\frac{h}{2}\cdot(f(x_{n-1})+f(x_n))}_{\text{area trapezio n-esimo}} \\ &= \frac{h}{2}(f(x_0)+f(x_n)) + \sum_{i=1}^{n-1} h\cdot f(x_i) \end{aligned}$$

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

- Per s = 2, integrando la funzione quadratica a tratti, si ottiene la formula delle parabole:

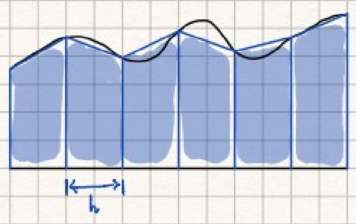
$$I_n^{parab}(f) = I\left(\textstyle\prod_2^c\right) = \sum(\text{aree trapezi parabolici}) = \sum_{i=0}^{n} w_i \cdot f(x_i), \text{ con } w_i = \begin{cases} h/3, & i = 0, n \text{ pari} \\ 4h/3, & i \text{ dispari} \\ 2h/3, & i \text{ pari } 2 \leq i \leq n-2 \end{cases}$$

$$\int_\alpha^\beta \textstyle\prod_2(x)dx = \frac{h}{3}\cdot f(\alpha) + \frac{4}{3}\cdot h \cdot f\left(\frac{\alpha+\beta}{2}\right) + \frac{h}{3}\cdot f(\beta)$$

Errore formula trapezi

La formula dei trapezi utilizza l'interpolazione lineare a tratti, imponendo $s=1$ l'integrale viene approssimato con la somma di aree di trapezi lineari. L'i-esimo trapezio ha altezza $h=\frac{b-a}{n}$ e basi $f(x_i-1)$ e $f(x_i)$ con $1 \le i \le n$, 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:

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



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

Per ricavare una stima dell'errore possiamo usare la stima $|I(f)-I_n(f)| = |I(f)-I(f_n)| \le |I(f-f_n)| \le (b-a)\,dist(f,f_n)$. Se $dist \to 0$ allora ci sarà convergenza, altrimenti potrebbero presentarsi problemi di divergenza.

Per quanto riguarda le formule di quadrature composte ottenute come $I_n(f)=I(\Pi_s^c)$, con n multiplo di s: $|I(f)-I_n(f)| \le (b-a)\,dist(f,\Pi_s^c) \le$ $\le (b-a)K_s \cdot h^{s+1}$ se $f \in C^{s+1}[a,b]$ con $h=\max \Delta x_i$. Quindi per qualsiasi distribuzione dei nodi per cui $h \to 0$ 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 $O(h^2)$.

Condizionamento di sistemi lineari

La prima risposta delle due risponde anche ad un'altra domanda:
Si ricavi una stima dell'errore relativo sulla soluzione di un sistema lineare non singolare in caso di vettore termine noto effetto da errore:

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.

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

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

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 $\hat{A}\tilde{x} = b$, $\hat{A} = A + \delta A$

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

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

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

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.

Sistema delle equazioni normali per approssimazione lineare

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

$$V = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ \vdots \\ 1 & x_{m+1} & x_{m+1}^2 & \dots & x_{m+1}^m \\ \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^m \end{pmatrix}$$

La sottomatrice $V \in \mathbb{R}^{(m+1)(m+1)}$ è matrice di Vandermonde per l'interpolazione di grado $\le m$ 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 $m = 1$

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

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

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

Costo computazionale MEG/Metodo di Gauss (indicato come possibile domanda file 20/21)

Il costo computazione del meg è dato dall'analisi tra ciclo interno, composto da n moltiplicazioni ed n some, scritto come:

$$
\begin{aligned}
c_n^{meg} &\sim \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} 2n \\
&= 2n \sum_{i=1}^{n-1} (n-i) \\
&\underset{j=n-i}{=} 2n \sum_{j=1}^{n-1} j \\
&= 2n \cdot \frac{n(n-1)}{2} \\
&= n^3 - n^2 \sim n^3, \quad n \to \infty
\end{aligned}
$$

Vedendo però che le operazioni vettoriali non ha senso farle sui vettori riga, le facciamo solo sul segmento di vettori con indici da i + 1 ad n, verificando che otteniamo:

$$
\begin{aligned}
c_n^{meg} &\sim \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} 2(n-i) \\
&= 2 \sum_{i=1}^{n-1} (n-i)^2 \\
&\underset{j=n-i}{=} 2 \sum_{j=1}^{n-1} j^2 \sim \frac{2}{3} n^3
\end{aligned}
$$

Ottenendo infine:

$$
\frac{(n-1)^3}{3} < \sum_{j=1}^{n-1} j^2 < \frac{n^3}{3} - 1
$$

Fattorizzazione QR per soluzioni di sistemi sovradeterminati (indicato come possibile domanda file 20/21)

Sia $A \in \mathbb{R}^{m \times n}$, $m \geq n$, $rango(A) = n$. Fattorizzando $A = QR$, si ha che

$$
A^t A = (QR)^t QR = R^t Q^t QR = R^t IR = R^t R
$$

e

$$
A^t b = R^t Q^t b
$$

quindi il sistema $A^t Ax = A^t b$ diventa

$$
R^t Rx = R^t Q^t b
$$

ma essendo R (e quindi R^t) invertibile

$$
(R^t)^{-1} R^t Rx = Rx = (R^t)^{-1} R^t Q^t b = Q^t b
$$

cioè il sistema $A^t Ax = A^t b$ equivale al sistema triang. sup.

$$
Rx = d = Q^t b
$$

che si può facilmente risolvere con la sostituzione all'indietro.

Nella pratica andiamo a risolvere un sistema perturbato del tipo ~R~x=~d, con R molto meglio condizionata di A^tA.

Applicazione del MEG nel calcolo del determinante (indicato come possibile domanda file 20/21)

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

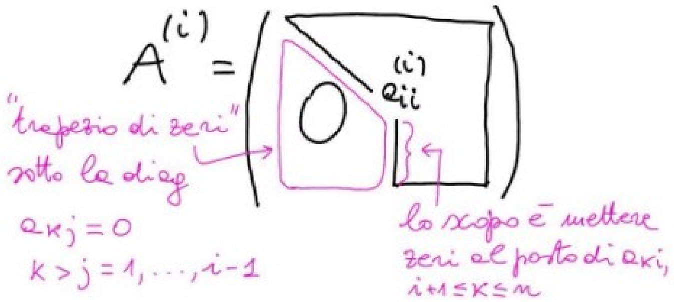
Esso si applica ripetutamente per mettere zeri in ogni colonna sotto la diagonale principale, in maniera tale da ottenere una matrice triangolare superiore.
Partendo ad esempio da una matrice A 3x3 (a destra) si arriva facilmente alla matrice U (a sinistra):

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

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

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

Meg e relazione con la fattorizzazione LU (indicato come possibile domanda file 20/21)

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

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

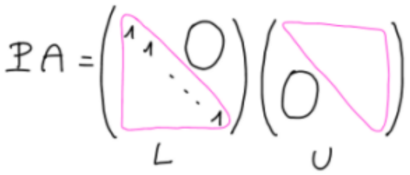
Avendo L che ha ordine invertito:

$$\begin{aligned} L = \mathcal{L}^{-1} &= \overbrace{(T_{2,1}(-m_{2,1}))^{-1}\ (T_{3,1}(-m_{3,1}))^{-1}\ (T_{3,2}(-m_{3,2}))^{-1}}^{\text{ordine invertito}} \qquad \text{ricordando che } (T_{k,i}(\alpha))^{-1} = T_{k,i}(-\alpha) \\ &= T_{2,1}(m_{2,1})\ T_{3,1}(m_{3,1})\ T_{3,2}(m_{3,2}) \end{aligned}$$

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 $2/3 \cdot n^3$:



Calcolo della matrice inversa (indicato come possibile domanda file 20/21)

Un modo semplice per calcolare l'inversa si basa su una proprietà del prodotto matrice-vettore che abbiamo usato spesso, cioè che z = Bc si può interpretare come combinazione lineare delle colonne di B tramite i coeff. di c, cioè:

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

Applicando l'osservazione con:

(cioè con $c_i = 1$, $c_j = 0$ e $j \neq i$)

$$c = e^{(i)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

cioè $c_i = 1$ e $c_j = 0,\ j \neq i$, otteniamo

$$Be^{(i)} = C_i(B)$$

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

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

Possiamo quindi calcolare l'inversa "per colonne" risolvendo M = n sistemi lineari, tutti con matrice A, in cui il termine noto varia tra gli n vettori coordinati della base canonica.

Se usassimo n volte il metodo standard avremmo un costo $2/3n^4$.

Invece calcolando una volta col meg la fattorizzazione LU e risolvendo le M = n coppie di sistemi triangolari come segue, si ha il calcolo dell'inversa a costo cubico:

$$\begin{cases} Ly^{(i)} = \underline{P}e^{(i)} \\ Ux^{(i)} = y^{(i)} \end{cases}$$

si ha un costo

$$c_n^{(2)} \sim \frac{2}{3}n^3 + 2n^2 n = \frac{8}{3}n^3\ flops$$