Glossario di Calcolo Numerico

Riccardo Graziani

Anno Accademico 2024/2025

Indice

1 Sistema binario e floating-point IEEE754

(Def.) Se $x \in \mathbb{R}$ e $N \in \mathbb{N}$ allora:

$$x = \pm x_{\mathrm{n}} N^{\mathrm{n}} + x_{\mathrm{n}-1} N^{\mathrm{n}-1} + ... + x_0 + x_{-1} N^{-1} + ... + x_{-\mathrm{r}} N^{-\mathrm{r}} \Rightarrow (x)_{\mathrm{N}}$$

in cui: $n \in \mathbb{N}$, $r \in \mathbb{N} \cup \infty$, $x_{\mathrm{j}} \in \{0, 1, ... \text{ N - } 1\}$, $\forall j = \mathrm{n}, \mathrm{n} - 1, ... -\mathrm{r}$

(Def.) Usando la notazione binaria, un numero $x \neq 0$ é scritto come:

$$(x)_2 = (-1)^{\mathrm{s}} \cdot (2)^{\mathrm{e}-\mathrm{b}} \cdot 1.f$$

in cui: s é il segno, e - b é l'esponente con **bias** che serve per avere e ≥ 0 per non doverne memorizzare il **segno** e *1.f* é la mantissa.

1.1 Massimo rappresentabile in Float32

Per rappresentare il nostro M_{MAX} possiamo imporre:

- f = 111...1 (in questo caso f = 32)
- e = 11111111 ovvero $\sum_{k=0}^{7}(2)^{\mathrm{k}} = \frac{1-2^8}{1-2} = 2^8$ - 1 = 255

Con questi valori possiamo esprimere *1.f* come:

$$1.f = 1.1...1 = \sum_{k=0}^{-23} 2^{\mathrm{k}} = \sum_{k=0}^{23}(\frac{1}{2})^{\mathrm{k}} = \frac{1-(\frac{1}{2})^{24}}{1-\frac{1}{2}} = 2 - 2^{-23}$$

e calcolare e - b = 255 - 127 = 128.
Quindi possiamo esprimere M_{MAX} come:

$$M_{\mathrm{MAX}} = (2 - 2^{-23})2^{128}$$

Analogamente M_{MIN} é espresso come:

$$M_{\mathrm{MIN}} = 2^{-127}$$

avendo *1.f* = 1.00..., $e = 0$, e - b = -127.

1.2 Spaziatura dei numeri in Float32

Dato x ∈ Float32, $|x| \neq \mathrm{M_{MAX}}$, x ≠ 0 allora sono ben definiti il **precedente** prec(x) e **successivo** succ(x).

(Def.) Definisco la funzione **distanza da F** come:

$$F \to \mathbb{R}$$
$$x \to \text{Distanza da F}$$

in cui d(x) = $2^{\mathrm{e\ -\ b\ -\ 23}}$. Con alcuni esempi di x si ricava che la spaziatura varia in modo proporzionale a $|x|$.

(Def.) Dato x ∈ $]-M_{\mathrm{MAX}}, M_{\mathrm{MAX}}[$ definisco due metodi float:

- fl^{TR}: floating point per **troncamento**
- fl^{AR}: floating point per **arrotondamento**

espressi in maniera funzionale come:

- $fl^{TR} = \begin{cases} \text{-M}_{\text{MAX}} & \forall x \leq -M_{\text{MAX}} \\ \text{Rappresentazione bin. troncata} & \forall x =]\text{-M}_{\text{MAX}}, \text{M}_{\text{MAX}}[\\ \text{M}_{\text{MAX}} & \forall x \geq M_{\text{MAX}} \end{cases}$

- $fl^{AR} = \begin{cases} \text{-M}_{\text{MAX}} & \forall x \leq -M_{\text{MAX}} \\ \text{Rappresentazione bin. arrotondata} & \forall x =]\text{-M}_{\text{MAX}}, \text{M}_{\text{MAX}}[\\ \text{M}_{\text{MAX}} & \forall x \geq M_{\text{MAX}} \end{cases}$

2 Errore relativo ed errore assoluto

(Def.) Sia $\tilde{x} \in \mathbb{R}$ **approssimazione** di $x \in \mathbb{R}$ allora definisco:

- $\text{ERR}_{\text{ASS}}(\tilde{x}) = |x - \tilde{x}|$
- $\text{ERR}_{\text{REL}}(\tilde{x}) = \frac{|x-\tilde{x}|}{|x|}$

Con questo possiamo definire gli errori di approssimazione relativi ed assoluti come sopra ponendo $\tilde{x} = fl^{TR}$, fl^{AR}:

- $ERR_{\text{ASS}}(fl^{\text{TR/AR}}) = \left|x - fl^{\text{TR/AR}}(x)\right|$
- $ERR_{\text{REL}}(fl^{\text{TR/AR}}) = \frac{\left|x-fl^{\text{TR/AR}}(x)\right|}{|x|}$

Volendo stimare l'errore di rappresentazione in troncamento chiediamo che $|fl(x) - x| \leq ?$; supponendo $x>0$:

$$|fl(x) - x| \leq |fl(x) - succ(x)| = d(fl(x)) = 2^{\text{e(x)}-\text{b}} \cdot 2^{-\text{nf}}$$

in cui diventa:

- $ERR_{\text{ASS}}(fl^{\text{TR}}(x)) \leq 2^{-\text{nf}} \cdot 2^{\text{e(x)}-\text{b}}$
- $ERR_{\text{REL}}(fl^{\text{TR}}(x)) \leq \frac{2^{-\text{nf}} \cdot 2^{\text{e(x)}-\text{b}}}{|x|} \leq 2^{-\text{nf}}$

Invece per l'arrotondamento chiediamo che $|x - fl(x)| \leq ?$; supponendo $x>0$:

$$|x - fl(x)| \leq \frac{|fl(x) - prec(fl(x))|}{2}$$

in cui diventa:

- $ERR_{\text{REL}}(fl^{\text{AR}}(x)) \leq \frac{2^{-\text{nf}}}{2}$

dove nf é il numero di bit per f.

3 Precisione macchina, operazioni macchina e loro proprietá

(Def.) Definisco come **precisione macchina** il piú piccolo numero rappresentabile > 0 in floating point per cui vale che:

$$fl(1 + \epsilon_{\text{MACH}}) \neq 1$$

noi assumiamo che $\epsilon_{\text{MACH}} = 2^{-\text{nf}}$.

(Def.) Per ogni operazione reale $*$ definiamo un'operazione macchina $\circledast$ definita come:

$$x \circledast y = fl(fl(x) * fl(y))$$

Ció porta alla rottura dell'algebra tradizionale, ad esempio nella moltiplicazione macchina:

- non vale la proprietá commutativa;
- non vale la proprietá associativa;
- non vale la proprietá distributiva rispetto all'addizione;
- gli elementi neutri non sono unici;
- non vale la proprietá di cancellazione;

4 Stabilitá di un algoritmo, stabilitá delle operazioni macchina e cancellazione numerica

(Def.) Un algoritmo si dice **stabile** se:

$$ERR_{\text{REL}}(OUTPUT) \leq C_{\text{STAB}} ERR_{\text{REL}}(INPUT)$$

ovvero non amplifica in maniera incontrollata gli errori presenti sui dati.
Ad esempio l'errore di stabilitá della somma é:

$$ERR_{\text{REL}}(x \oplus y) \leq \epsilon_{\text{MACH}} + \epsilon_{\text{MACH}}(\epsilon_{\text{MACH}} + 1)\frac{|x| + |y|}{|x + y|}$$

Si presenta un problema quando $x \approx -y$, infatti se $y = -x + \delta$ allora:

$$\frac{|x| + |y|}{|x + y|} = \frac{|x| + |x - \delta|}{|\delta|}$$

in cui per $\delta \to 0^+$ la quantitá sopra va a $+\infty$.

Se $x = 1, \delta = \epsilon_{\text{MACH}}$:

$$\frac{|x| + |x - \delta|}{|\delta|} = \frac{1 + 1 + \epsilon_{\text{MACH}}}{\epsilon_{\text{MACH}}} > \frac{1}{\epsilon_{\text{MACH}}}$$

Dal quale si ricava che $ERR_{\text{REL}}(x \oplus y) \leq 1 + \epsilon_{\text{MACH}}$ ovvero un errore di piú del 100%. Quando ció accade si parla di **cancellazione numerica**, ovvero un fenomeno che si verifica quando un'operazione matematica provoca l'eliminazione di cifre significative, comportando nel risultato di una perdita netta in termini di precisione rispetto al valore originale.

5 Problemi matematici e buona posizione

(Def.) Le caratteristiche principali dei problemi **ben posti** sono:

- definizione del **dominio** (F) e dei **dati ammissibili** (D);
- $\forall d \in \text{D } \exists!\ x \in \text{X} : \text{F}(x, d) = 0$;
- $\text{d} \to x(d)$ ovvero la soluzione del problema é **continua**;

6 Condizionamento numerico assoluto e relativo di un problema ben posto

(Def.) Un problema matematico si dice **ben condizionato** se a piccole variazioni di dati corrispondono piccole variazioni della soluzione. Il condizionamento locale é definito come:

$$limsup_{\tilde{\text{d}} \to \text{d}} = \frac{\left|x(\tilde{d})\right| - |x(d)|}{\left|(\tilde{d} - d)\right|^{\alpha}} \leq k_{\alpha}(d), 0 < \alpha \leq 1$$

in cui $k_\alpha(d)$ é il **numero di condizionamento** e limsup indica il massimo valore raggiungibile da $\frac{\left|x(\tilde{d})\right|-|x(d)|}{\left|(\tilde{d}-d)\right|^\alpha}$ e si definisce locale perché misura la fluttuazione del problema rispetto alle variazioni dell'input in un punto d.
Il condizionamento globale é definito come:

$$\sup_{\mathrm{d}\in\mathrm{D}} k_\alpha(d) \leq k_\alpha < +\infty$$

e misura la fluttuazione del problema rispetto alle variazioni dell'input.
Dalla definizione si ha che se $k_\alpha(d)$ é noto allora $\exists$I intorno di d tale che:

$$\left|x(\tilde{d}) - x(d)\right| \leq 2k_\alpha(d) \left|(\tilde{d} - d)\right|^\alpha$$

7 Numero di condizionamento di una matrice e stima dell'errore relativo della soluzione di un sistema lineare con termine noto affetto da errore

Vogliamo trovare $x \in X : F(x,d) = 0, d \in D$, possiamo usare il metodo numerico, per cui $x_\mathrm{n} \in X_\mathrm{n} : F_\mathrm{n}(x_\mathrm{n}, d) = 0, d \in D, n \in \mathbb{N}$ e $X_\mathrm{n}, F_\mathrm{n}$ approssimano X, F in senso opportuno.

Si vorrebbe che:

- $x_\mathrm{n}(d) \to x(d)$ se $n \to +\infty \ \forall d \in \mathrm{D}$
- prende il nome di **convergenza del metodo**
- c'é convergenza uniforme se si chiede che $\sup_{\mathrm{d}\in\mathrm{D}} |x_\mathrm{n}(d) - x(d)| \to 0$

(Def.) Il metodo $x \in X : F(x,d) = 0, d \in D$ si dice **consistente** se:

$$\lim_{\mathrm{n}\to+\infty} F_\mathrm{n}(x(d), d) = 0$$

(Th. Lax-Richmeyer) Se un metodo é consistente allora:

$$\textbf{convergente} \Leftrightarrow \textbf{stabile}$$

7.1 Analisi del problema Ax = b

Il problema $Ax = b$ é ben posto se:

- $A \in M_{\text{nxn}}$ ed é **invertibile**
- $b \in \mathbb{R}^{\text{n}}$

La soluzione del sistema risulta essere $x = A^{-1}b$ ma potrebbe essere » 1.

(Def.) Ogni volta che scelgo una norma su $\mathbb{R}^{\text{n}}$ rimane definita una **norma indotta** su $M_{\text{nxn}}(\mathbb{R})$. Sia $\|\cdot\| : M_{\text{nxn}}(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ allora:

$$A \to \sup_{\text{x}\neq 0} \frac{\|Ax\|}{\|x\|}$$

ovvero A é associata al $\sup_{\text{x}\neq 0} \frac{\|Ax\|}{\|x\|}$ e ció permete di calcolare il **condizionamento**.

7.1.1 Condizionamento relativo di Ax = b

Considerando b dato, $D = \mathbb{R}^{\text{n}}$ ed A rappresentabile in modo esatto abbiamo:

- $\tilde{b} \approx b$
- $\tilde{b} = b + \delta b$
- $\tilde{x} = x + \delta x$
- $A\tilde{x} = \tilde{b}$

Vogliamo stimare l'$ERR_{\text{REL}} = \frac{\|\delta x\|}{\|x\|}$, prendendo le due formule:

- $A(x + \delta x) = b + \delta b$
- $Ax = b$

unendole si ottiene che: $A\delta x = \delta b$. Siccome sappiamo che A é **invertibile** allora:

$$\delta x = A^{-1}\delta b \Rightarrow \|\delta x\| = \|A^{-1}\delta b\|$$

Secondo le proprietá delle norme indotte sulle matrici sappiamo che:

$$A \to \sup_{\text{x}\neq 0} \frac{\|Ax\|}{\|x\|} \geq \frac{\|A\bar{x}\|}{\|\bar{x}\|} \Rightarrow \|A\bar{x}\| \leq \|A\|_*\|\bar{x}\|_*$$

in cui $\bar{x}$ é un certo x fissato. Otteniamo dunque che:

$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|_* \|\delta b\|$$

che é la stima del **condizionamento assoluto**. Per ottenere peró l'errore relativo dobbiamo dividere per $\|x\|$, sapendo che:

$$\|b\| = \|Ax\| \leq \|A\|_* \|x\| \Rightarrow \|x\| = \frac{\|b\|}{\|A\|_*}$$

quindi ora possiamo dire che:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|\delta x\|}{\|b\|} \cdot \|A\|_* \leq \|A^{-1}\|_* \|A\|_* \cdot \frac{\|\delta b\|}{\|b\|}$$

da cui si ottiene che:

$$ERR_{\text{REL}} \frac{\|\delta x\|}{\|x\|} \leq ERR_{\text{REL}} \frac{\|\delta b\|}{\|b\|}$$

ovvero che il problema é stabile (per la definizione).
Ricaviamo anche il condizionamento relativo globale del problema:

$$\frac{ERR_{\text{REL}}(x)}{ERR_{\text{REL}}(b)} \leq \|A^{-1}\|_* \|A\|_*$$

che viene denominato: **numero condizionamento su A in norma** $\|\cdot\|$ e viene indicato come $cond(A, \|\cdot\|)$.

In sostanza ricaviamo due conclusioni:

- potremo anche perdere molta precisione su $cond(A, \|\cdot\|)$ come 10^8 l'ERR_{REL} su x potrebbe creare 10^{-8} anche su uno ϵ_{MACH};
- la stima che abbiamo ottenuto é una **worst case scenario**

7.1.2 Caso generale con A e b da rappresentare con errori

Dall'equazione $Ax = b$ otteniamo che $(\mathbb{1} + A^{-1}\delta A)\delta x = A^{-1}\delta b - A^{-1}\delta A x$ (lo svolgimento si trova negli appunti del prof. NoteL05). Da questa formula vogliamo sapere se $(\mathbb{1} + A^{-1}\delta A)$ é invertibile.

(Def.) Definiamo la **serie di Neumann** nella formulazione:

$$(\mathbb{1} + B)^{-1} = \sum_{k=0}^{+\infty} (-1)^{\mathrm{k}} B^{\mathrm{k}}$$

e sappiamo che B é **quadrata** e **invertibile**, supponiamo poi sia **diagonalizzabile**:

$$B^{\mathrm{k}} = P^{-1}\Lambda P P^{-1}\Lambda P ... P^{-1}\Lambda P$$

sapendo che $P^{-1}P = \mathbb{1}$ otteniamo che:

$$B^{\mathrm{k}} = P^{-1}\Lambda^{\mathrm{k}} P$$

e sostituendo a $(\mathbb{1} + B)$ otteniamo:

$$(\mathbb{1} + B) = P^{-1}(\mathbb{1} + \Lambda)P$$

Sostituendo ció che abbiamo ricavato alla serie di Neumann scopriamo che per B quadrata e diagonalizzabile con $|\lambda_{\mathrm{i}}| < 1$ si ha che $(\mathbb{1} + B)$ é invertibile e che $(\mathbb{1} + B)^{-1} = \sum_{k=0}^{+\infty} (-B)^{\mathrm{k}}$.
Con questo si ricava che l'errore di $\tilde{A}\tilde{x} = \tilde{b}$ é proporzionale al quadrato della perturbazione.

8 Metodi per la soluzione di Ax = b

Abbiamo due strade possibili:

- Metodi Diretti: costruiscono le soluzioni numeriche e si possono tradurre in fattorizzazioni di matrice $A = B \cdot C$
- Metodi Iterativi: costruiscono una successione di vettori di approssimazioni delle soluzioni

8.1 Metodo di sostituzione indietro

Supponendo che A sia triangolare superiore, posso definire la soluzione ricorsiva di x come:

$$x_{\mathrm{n}-1} = \frac{b_{\mathrm{n}-1} - A_{\mathrm{n}-1,\mathrm{n}} \cdot x_{\mathrm{n}}}{A_{\mathrm{n}-1,\mathrm{n}-1}}$$

8.2 Metodo di sostituzione avanti

Supponendo che A sia triangolare inferiore, posso definire la soluzione ricorsiva di x come:

$$x_{\mathrm{n}-1} = \frac{b_{\mathrm{n}-1} - A_{\mathrm{n},\mathrm{n}-1} \cdot x_{\mathrm{n}}}{A_{\mathrm{n}-1,\mathrm{n}-1}}$$

9 Fattorizzazione LU

9.1 Fattorizzazione LU senza pivoting

Per risolvere $Ax = b$ con A invertibile vogliamo ricondurci alla soluzione di sistemi triangolari, per i quali possiamo usare gli algoritmi di sostituzione. In particolare vogliamo fattorizzare A come $A = LU$ con L matrice triangolare inferiore con valori diagonali pari a 1 e U triangolare superiore. Grazie a ció la fattorizzazione LU puó costrutire la soluzione numerica componente per componente, interpretando le norme come stabilizzazioni di matrici, in particolare per $A = LU$.

Fatto ció il sistema $Ax = b$ diventa $LUx = b$ e possiamo calcolare la soluzione y di $Ly = b$ ponendo $Ux = y$ e risolvendolo per sostituzione. Osserviamo che la moltiplicazione a sx di una M_{nxn} equivale al passo elementare dell'eliminazione Gaussiana.
Definiamo per ogni vettore $m^{(\mathrm{k})} = (m_1^{(\mathrm{k})}, ..., m_{\mathrm{n}-\mathrm{k}}^{(\mathrm{k})}) \in \mathbb{R}^{\mathrm{n}-\mathrm{k}}$:

$$\mathbb{L}^{(\mathrm{k})}(m^{(\mathrm{k})}) = \mathbb{I}_{\mathrm{n}} - u^{(\mathrm{k})} e_{\mathrm{k}}^{\mathrm{t}}$$

dove $u_{\mathrm{i}}^{(\mathrm{k})} = 0$ se $i \leq k$ oppure $u_{\mathrm{i}}^{(\mathrm{k})} = m_{\mathrm{i}}^{(\mathrm{k})}$ se $i = k{+}1, ..., n$ dove $e_{\mathrm{k}}^{\mathrm{t}}$ rappresenta il k-esimo vettore della base canonica.
Graficamente otteniamo che:

$$\mathbb{L}^{(\mathrm{k})}(m^{(\mathrm{k})}) = \begin{bmatrix} \mathbb{I}_{\mathrm{k}} & \mathbb{O}_{\mathrm{k},\mathrm{n}-\mathrm{k}} \\ M^{(\mathrm{k})} & \mathbb{I}_{\mathrm{n}-\mathrm{k}} \end{bmatrix}$$

Calcolando $\mathbb{L}^{(\mathrm{k})}(m^{(\mathrm{k})})A$ scopriamo che abbiamo lasciato invariate le prime k righe di A e ottenuto ciascuna delle righe $k + 1, ..., n$ di $\mathbb{L}^{(\mathrm{k})}(m^{(\mathrm{k})})$ come

combinazione lineare della i-esima e k-esima riga di A.
Se scegliamo:

$$m_{\mathrm{i-k}}^{(\mathrm{k})} = \frac{a_{\mathrm{i,k}}}{a_{\mathrm{k,k}}} i = k+1, ..., n$$

otteniamo che $\mathbb{L}^{(\mathrm{k})}(m^{(\mathrm{k})})A$ rappresenta il k-esimo passo dell'eliminazione Gaussiana. Possiamo calcolare la matrice U come $A^{(\mathrm{k})} = \mathbb{L}^{(\mathrm{k})}(m^{(\mathrm{k})})A^{(\mathrm{k}-1)}$ e $U = A^{(\mathrm{n}-1)}$, che rende U triangolare superiore.

9.1.1 Schema fattorizzazione LU senza pivoting

Data $A \in M_{\mathrm{nxn}}(\mathbb{R})$ invertibile e $L^0 = \mathbb{I}_{\mathrm{n}}$, allora per $k = 1, 2, ..., n-1$:

- calcolo $m_{\mathrm{i-k}}^{(\mathrm{k})} = \frac{a_{\mathrm{i,k}}}{a_{\mathrm{k,k}}} i = k+1, ..., n$
- $u_{\mathrm{i}}^{(\mathrm{k})} = 0$ se $i \leq k$ oppure $u_{\mathrm{i}}^{(\mathrm{k})} = m_{\mathrm{i}}^{(\mathrm{k})}$ se $i = k+1, ..., n$
- $\widetilde{L}^{(\mathrm{k})} = \mathbb{I}_{\mathrm{n}} - u^{(\mathrm{k})} e_{\mathrm{k}}^{\mathrm{t}}$
- $L^{\mathrm{k}} = L^{(\mathrm{k}-1)} + u^{(\mathrm{k})} e_{\mathrm{k}}^{\mathrm{t}}$
- $A^{(\mathrm{k})} = \widetilde{L}^{(\mathrm{k})} A^{(\mathrm{k}-1)}$
- $U = A^{(\mathrm{n}-1)}, L = L^{(\mathrm{n}-1)}$

I problemi principali di questo schema sono:

- secondo il lemma dell'invertibilitá delle matrici $\tilde{L}^{(\mathrm{K})}$ in caso di pivot nullo la fattorizzazione LU puó non essere applicabile ed inoltre puó diventare instabile a causa di una possibile **cancellazione numerica** per pivot $\to 0$;

Un esempio dell'algoritmo si trova nelle note del prof. NoteL07.

9.2 Fattorizzazione LU con pivoting parziale per righe

La fattorizzazione LU con pivoting parziale per righe di A che risolve alcuni problemi della precedente metodologia, in quanto riesce a evitare il caso del pivot nullo cercando delle **permutazioni** delle righe della matrice A tali per cui il pivot é **non nullo**.

(Def.) Se σ é una permutazione dell'insieme $(1, ..., n)$, essa puó essere associata alla matrice:

$$P = (p_{\mathrm{i,j}})_{\mathrm{i,j=1...n}}; \quad p_{\mathrm{i,j}} = 1 \text{ se } \sigma(i) = j \text{ altrimenti } 0$$

Le proprietá delle matrici di permutazione sono:

- se P, Q sono matrici di permutazione $\Rightarrow PQ, QP$ sono di permutazione
- $PP^{\mathrm{t}} = P^{\mathrm{t}}P = \mathbb{I}_{\mathrm{n}}$ cioé sono ortogonali
- $\det P = \pm 1$
- se σ é lo scambio di due elementi, allora P é detta di scambio e $P = P^{\mathrm{t}} = P^{-1}$

Se P é una matrice di permutazione che rappresenta σ ed A ha la stessa dimensione allora:

- PA significa permutare le righe di A
- AP significa permutare le colonne di A

Se nella fattorizzazione LU incontriamo una riga con pivot nullo, possiamo trovare una permutazione di A per ottenere un pivot non nullo.

(Def.) Per effettuare il pivoting parziale per righe possiamo, ad ogni passo della fattorizzazione LU, prima di agire su $A^{(\mathrm{k}-1)}$ per calcolare $A^{(\mathrm{k})}$ scambio la k-esima riga di $A^{(\mathrm{k}-1)}$ con una riga $s^{(\mathrm{k})}$ che soddisfi:

$$\left|a^{(\mathrm{k}-1)}_{\mathrm{s^{(k)},k}}\right| = \max \left|a^{(\mathrm{k}-1)}_{\mathrm{i,k}}\right| : i = k, ..., n$$

ovvero trovare il max elemento della k-esima colonna e scambiare la riga.

9.2.1 Schema fattorizzazione LU con pivoting parziale per righe

Definendo $L^{(\mathrm{k})} = [\hat{L}^{(\mathrm{k})}]^{-1}$ ottengo che:

- $L = L^{(1)}L^{(2)}...L^{(\mathrm{n}-1)}$
- $P = P^{(\mathrm{n}-1)}...P^{(1)}$

e si ottiene che $LU = PA$ in cui:

- L é triangolare inferiore con elementi diagonali pari ad 1
- U matrice diagonale superiore
- P matrice di permutazione

Il pivoting parziale inoltre ha la proprietá di rendere l'algoritmo di fattorizzazione **stabile**.

9.3 Punti fissi e lemma delle contrazioni

(Def.) Sia $F : \mathbb{R}^{\mathrm{n}} \to \mathbb{R}^{\mathrm{n}}, x^* \in \mathbb{R}^{\mathrm{n}}$, x^* si dice punto fisso di F se $F(x^*) = x^*$.

(Lemma: Delle contrazioni) Sia $F : \mathbb{R}^{\mathrm{n}} \to \mathbb{R}^{\mathrm{n}}$ e $L<1 : \forall x, y \in \mathbb{R}^{\mathrm{n}}$ allora:

$$\|F(x) - F(y)\| \leq L\|x - y\|$$

Ovvero per ogni scelta di $x^0 \in \mathbb{R}^{\mathrm{n}}$ le successioni $x^{\mathrm{k+1}} = F(x^{\mathrm{k}})$ converge ad un $x^* \in \mathbb{R}^{\mathrm{n}}$. Inoltre x^* é un punto fisso e vale che:

$$e_{\mathrm{k+1}} = \|x^{\mathrm{k+1}} - x^*\| \leq L\|x^{\mathrm{k}} - x^*\| = Le_{\mathrm{k}}$$

(Dim. Lemma delle contrazioni) Sia $g : \mathbb{R}^n \to \mathbb{R}^n$ una funzione continua su un intervallo $I \subset \mathbb{R}^n$. Supponiamo che g soddisfi la **condizione di contratto**, cioè esiste una costante $L \in [0, 1)$ tale che:

$$\|g(x) - g(y)\| \leq L\|x - y\| \quad \text{per ogni} \quad x, y \in I.$$

Dimostriamo che g ha un **unico punto fisso** in I, ossia esiste un unico $x^* \in I$ tale che $g(x^*) = x^*$. Inoltre, mostreremo che la successione definita da:

$$x_{n+1} = g(x_n),$$

con un punto iniziale $x_0 \in I$, converge a x^*. Consideriamo la successione $\{x_n\}$ definita da $x_{n+1} = g(x_n)$, con $x_0 \in I$. Per ogni coppia di iterazioni successive, possiamo scrivere:

$$\|x_{n+1} - x_n\| = \|g(x_n) - g(x_{n-1})\| \leq L\|x_n - x_{n-1}\|.$$

Iterando questa disuguaglianza, otteniamo:

$$\|x_{n+1} - x_n\| \leq L\|x_n - x_{n-1}\| \leq L^2\|x_{n-1} - x_{n-2}\| \leq \cdots \leq L^n\|x_1 - x_0\|.$$

Poiché $L \in [0,1)$, la quantità $L^n \to 0$ per $n \to \infty$. Questo implica che:

$$\|x_{n+1} - x_n\| \to 0 \quad \text{quando} \quad n \to \infty.$$

Inoltre, possiamo dimostrare che la successione $\{x_n\}$ è di tipo **Cauchy**. Per ogni $m > n$, abbiamo:

$$\|x_m - x_n\| \leq \|x_m - x_{m-1}\| +_{m-1} -x_{m-2}\| + \cdots +_{n+1} -x_n\|.$$

Usando la contrazione su ogni termine:

$$\|x_m - x_n\| \leq \left(L^n + L^{n+1} + \cdots + L^{m-1}\right) \|x_1 - x_0\|.$$

La somma geometrica si semplifica come:

$$\|x_m - x_n\| \leq \frac{L^n(1 - L^{m-n})}{1 - L} \|x_1 - x_0\|.$$

Poiché $L^n \to 0$ quando $n \to \infty$, segue che $|x_m - x_n| \to 0$. Dunque, $\{x_n\}$ è una successione di Cauchy, e poiché $\mathbb{R}$ è completo, converge a un limite x^*:

$$\lim_{n \to \infty} x_n = x^*.$$

Supponiamo per assurdo che esistano due punti fissi distinti x_1^* e x_2^* tali che:

$$g(x_1^*) = x_1^* \quad \text{e} \quad g(x_2^*) = x_2^*.$$

Applicando la condizione di contratto:

$$\|g(x_1^*) - g(x_2^*)\| \leq L \|x_1^* - x_2^*\|.$$

Ma poiché $g(x_1^*) = x_1^*$ e $g(x_2^*) = x_2^*$, abbiamo:

$$\|x_1^* - x_2^*\| \leq L \|x_1^* - x_2^*\|.$$

Poiché $L < 1$, segue che:

$$\|x_1^* - x_2^*\| = 0 \quad \implies \quad x_1^* = x_2^*.$$

Quindi il punto fisso è unico. Poiché la successione $\{x_n\}$ converge a x^*, e g è continua, possiamo passare al limite nell'iterazione $x_{n+1} = g(x_n)$:

$$x^* = \lim_{n \to \infty} x_{n+1} = \lim_{n \to \infty} g(x_n) = g\left(\lim_{n \to \infty} x_n\right) = g(x^*).$$

Quindi x^* è il punto fisso di g. Abbiamo dimostrato che, se g soddisfa la condizione di contratto su un intervallo I con $L < 1$, allora esiste un unico punto fisso $x^* \in I$, e la successione generata da $x_{n+1} = g(x_n)$ converge a x^*.

Dalla dimostrazione del lemma concludiamo che $x^{\mathrm{k}+1} = F(x_{\mathrm{k}})$ é una successione di Cauchy e quindi ha limite in $\mathbb{R}^{\mathrm{n}}$ e ricaviamo le condizioni necessarie perché il punto fisso di F sia soluzione di $Ax = b$:

- $x^* = Ex^* + q \Rightarrow (\mathbb{1} - E)x^* = q$
- $Ax^* = b \Rightarrow x^* = A^{-1}b$

Che di dá la formula: $(\mathbb{1} - E)A^{-1}b = q$.

10 Metodi iterativi lineari stazionari per la soluzione di Ax = b

Vogliamo cercare una successione $x^{(\mathrm{k})} \in \mathbb{R}^{\mathrm{n}}$, $\forall n \in \mathbb{N}$ che approssimi x soluzione. Il motivo é presto detto; supposto di avere:

$$A \in M_{\mathrm{nxn}}, n \approx 10^5$$

e di lavorare con una CPU che esegue 4 cicli per clock con un clock di 2.5Ghz, ció significa poter effettuare: 10^{10} operazioni al secondo e di dover effettuare $n^3 \approx 10^{15}$ operazioni di algoritmo LU, ció si tradurrebbe in:

$$\frac{10^{15}}{10^{10}} = 10^5 \text{ sec } \approx \text{ 30 ore}$$

Inoltre un metodo diretto non riesce a calcolare la soluzione esatta a causa degli errori di rappresentazione e della loro propagazione.
L'idea dei metodi iterativi stazionari lineari é espressa come una successione:

- x^0 valore dato
- $x^{\mathrm{k}+1} = Ex^{\mathrm{k}} + q$ con $E \in M_{\mathrm{nxn}}(\mathbb{R})$, $q \in \mathbb{R}^{\mathrm{n}}$

in cui vogliamo che $x^{\mathrm{k}} \to x^*$ con $x^* : Ax^* = b$. Il nostro metodo iterativo si appoggia al concetto dei punti fissi appena spegati, infatti possiamo riscrivere il metodo iterativo come:

$$x^{\mathrm{k}+1} = F(x^{\mathrm{k}})$$

Notiamo che se $x^{\mathrm{k}} \to \bar{x}$ allora se F é continua anche $F(x^{\mathrm{k}}) \to F(\bar{x})$ e $x^{\mathrm{k+1}} \to \bar{x}$ quindi il metodo converge automaticamente poiché:

$$\bar{x} = \lim_{\mathrm{n}\to+\infty} x^{\mathrm{k}}$$

é punto fisso. Quindi vogliamo creare F in modo che se x^* é punto fisso allora é soluzione di $Ax = b$:

- con x^* punto fisso $x^* = Ex^* + q$
- con x^* soluzione di $Ax = b$: $x^* = A^{-1}b$

dunque $(\mathbb{1} - E)A^{-1}b = q$ é condizione necessaria alla consistenza.

10.1 Metodo di Richardson

Partendo dalla condizione necessaria per la consistenza il metodo di Richardson é definito come:

- x^0 dato
- $x^{\mathrm{k+1}} = (\mathbb{1} - A)x^{\mathrm{k}} + b$

Le ipotesi del lemma delle contrazioni sono rispettate (vedi NoteL08 per dimostrazione) a patto che $\|E\|_*<1$ il che dipende da quale norma uso, dall'analisi delle proprietá spettrali del metodo (vedi sempre NoteL08) ricavo che se $|1 - \lambda_{\mathrm{i}}| <1\ \forall i$ e A é diagonalizzabile allora Richardson converge.
Si puó inoltre dimostrare che se $\rho(E)<1$ allora $x^{\mathrm{k+1}} = Ex^{\mathrm{k}} + q$ é una successione convergente in una norma:

$$\rho(E) = \max|\lambda| : \lambda \text{ autovalore di E}$$

che é definito come **raggio spettrale**.

Possiamo definire un'ulteriore forma al metodo di Richardson detto **Metodo di Richardson precondizionato** il quale é definito come:

- $x^0 \in \mathbb{R}^{\mathrm{n}}$
- $P^{-1}b + (\mathbb{1} - P^{-1}A)x^{\mathrm{k}} = x^{\mathrm{k+1}}$

in cui $P^{-1}b = q$ e $(\mathbb{1} - P^{-1}A) = E$ e in cui dobbiamo scegliere opportunamente P.
Possiamo scomporre $A = L + D + U$ dove L é triangolare inferiore, D é diagonale e U é triangolare superiore, in quanto é difficile che $\rho(E)<1$ per ogni iterazione ottenuta in Richardson.

10.2 Metodo di Jacobi

Parto dal metodo di Richardson precondizionato e pongo $P = D$ da cui ottengo:

- $Ax = b$ diventa $D^{-1}Ax = D^{-1}b$
- il punto fisso $F(x) \Rightarrow D^{-1}b - D^{-1}Ax + x = x$

Dall'esempio svolto in NoteL09 otteniamo che Jacobi converge $\forall x^0$, il motivo é spiegato dal seguente lemma.

(Lemma: Dei cerchi di Gershgorin) Se $A = (A_{\mathrm{i,j}})_{\mathrm{i,j=1...n}}$ e λ é autovalore di A allora:

$$\lambda \in \cup_{\mathrm{i=0}}^{\mathrm{n}} B(A_{\mathrm{i,i}}, \sum_{j \neq i} |A_{\mathrm{i,j}}|)$$

ovvero gli autovalori di A sono contenuti nell'unione di tutti i cerchi con: centro i valori della diagonale di A e raggio la somma dei restanti valori nella stessa riga del centro.

(Prop.) Se $A \in M_{\mathrm{nxn}}$ é **strettamente diagonalmente dominante**, cioé $\forall i = 1...n \, |A_{\mathrm{i,i}}| > \sum_{j \neq i} |A_{\mathrm{i,j}}|$ allora il metodo di Jacobi converge $\forall x^0 \in \mathbb{R}^{\mathrm{n}}$.

(Dim. Prop.) Sappiamo che $E = \mathbb{1} - D^{-1}A$ dunque $D^{-1}A = 1$ se $i = j$ altrimenti $D^{-1}A = a_{\mathrm{i,j}}$ se $i \neq j$. Facendo $\mathbb{1} - D^{-1}A$ annullo la diagonale e cambio il segno ai restanti termini della matrice.
Se $\lambda \in$ autovalori di E, dal lemma di Gershgorin ho che:

$$\lambda \in \cup_{\mathrm{i=1}}^{\mathrm{n}} B(0, r_{\mathrm{i}})$$

$$r_{\mathrm{i}} = \sum_{j \neq i} \frac{|a_{\mathrm{i,j}}|}{|a_{\mathrm{i,i}}|} = \frac{\sum_{j \neq i} |a_{\mathrm{i,j}}|}{|a_{\mathrm{i,i}}|} < 1$$

poiché A é strettamente diagonalmente dominante ne risulta che $|\lambda| < 1$ ■

10.3 Metodo di Gauss-Seidel

A differenza di Jacobi questo metodo prevede di usare $P = L + D$ e $A = P + U$ da cui si ottiene la formula (con dimostrazione NoteL09):

$$P^{-1}b - P^{-1}Ux^{\mathrm{k}} = x^{\mathrm{k+1}}$$

che é equivalente alla formula:

$$Px^{\mathrm{k}+1} = b - Ux^{\mathrm{k}}$$

che viene risolta per sostituzione in avanti.
Tipicamente Gauss-Seidel ha prestazioni migliori di Jacobi ma é piú complesso trovare le condizioni per la convergenza.

10.4 Criterio di arresto per metodi lineari

Un possibile criterio é quello del **residuo relativo** secondo il quale dovrei fermare il metodo quando:

$$res_{\mathrm{rel}}^{(\mathrm{k})} = \frac{\|Ax^{(\mathrm{k})} - b\|}{\|b\|} < toll$$

Possiamo stimare l'errore relativo con il residuo relativo:

$$err_{\mathrm{rel}}^{(\mathrm{k})} = \frac{\|x^* - x^{(\mathrm{k})}\|}{\|x^*\|} \leq \|A\|_* \|A^{-1}\|_* res_{\mathrm{rel}}^{(\mathrm{k})} = Cond(A, \|\cdot\|_*) res_{\mathrm{rel}}^{(\mathrm{k})}$$

Ovvero usando nella pratica questo criterio su sistemi malcondizionati rischiamo di perdere precisione nel calcolo della soluzione proporzionalmente al numero di condizionamento della matrice.

Un secondo criterio é quello dello **step** secondo il quale dovrei fermarmi quando:

$$\|s^{(\mathrm{k})}\| = \|x^{(\mathrm{k}+1)} - x^{(\mathrm{k})}\| < toll$$

Dalla dimostrazione (NoteL09) ne risulta che:

$$err_{\mathrm{rel}}^{(\mathrm{k})} \leq \frac{\|A\|_* \|s^{(\mathrm{k})}\|}{1 - \|E\|_* \|b\|}$$

Ovvero che se troviamo una matrice con un raggio spettrale distante da 1 allora lo step é un'ottimo criterio anche su problemi malcondizionati. Se siamo interessati all'errore dobbiamo considerare anche la dimensione di $\|A\|_*$ e $\|b\|$.

11 Sistemi sovradeterminati e tecnica dei minimi quadrati con approssimazione della soluzione

Data $A \in M_{\text{mxn}}$, $b \in \mathbb{R}^{\text{n}}$, $m>n$ voglio trovare la soluzione di $Ax = b$.
Ipotizzo che le colonne di A siano **linearmente indipendenti**, ovvero:

$$\sum_j A_{\text{i,j}} c_{\text{j}} = 0 \quad \forall i \Rightarrow c = 0$$

allora $Ax = b$ ha soluzione se e solo se b appartiene allo spazio lineare generato:

$$b \in spaz : A(:,1), A(:,2), ..., A(:,n) = Im(A)$$

Se y é una soluzione, allora $A(x-y) = 0 \Rightarrow (x-y) = 0$.
Voglio trovare una generalizzazione del concetto di soluzione, e lo posso fare studiando i minimi di:

$$f(x) = \|Ax - b\|_2^2$$

in cui se x^* é soluzione allora $f(x^*) = \|Ax^* - b\| = 0$ e $f(x^*) \geq 0 \quad \forall x$.
Abbiamo due casi da studiare, ponendo che A sia di rango pieno:

- se $b \in Im(A)$ allora $\exists!\ x^*$ soluzione;
- se $b \notin Im(A)$ allora $\nexists\ x^*$ soluzione;

Considerando la generalizzazione di F:

- $\mathbb{R}^{\text{n}} \to \mathbb{R}$
- $x \to \|Ax - b\|_2^2$

Dobbiamo trovare $x^* : F(x^*) \leq F(x) \quad \forall x \in \mathbb{R}^{\text{n}}$. Osserviamo che se $b \in Im(A)$ allora $\exists!\ x^*$ e x^* soddisfa la disequazione.

11.1 Soluzione ai minimi quadrati

L'espressione $x^* : F(x^*) \leq F(x) \quad \forall x \in \mathbb{R}^{\text{n}}$ viene chiamata **minimi quadrati** e la sua soluzione é detta **soluzione ai minimi quadrati**.

(Th. Caso particolare delle proiezioni ortogonali) Sia $A \in M_{\text{mxn}}(\mathbb{R})$, $m>n$,

$rk(A) = n$, $b \in \mathbb{R}^{\mathrm{n}}$ e denotiamo con $F : \mathbb{R}^{\mathrm{n}} \to \mathbb{R}$ la formula $F(x) = \|Ax-b\|_2^2$. Allora $\exists!\ x^* \in \mathbb{R}^{\mathrm{n}} : F(x^*) = \min_{\mathrm{x}^* \in \mathbb{R}^{\mathrm{n}}} F(x)$, inoltre x^* é caratterizzato dalle **equazioni normali**:

$$A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}b$$

(Dim. Th. Caso particolare delle proiezioni ortogonali) Definisco $f_{\mathrm{v,x}}(t) = F(x + tv)\ \forall v \neq 0, v \in \mathbb{R}^{\mathrm{n}}$. Fissati x, v allora $f_{\mathrm{v,x}}(\cdot)$ é:

$$f_{\mathrm{v,x}}(t) = \|A(x + tv) - b\|_2^2 = [A(x + tv) - b]^{\mathrm{T}}[A(x + tv) - b] =$$

$$(x + tv)^{\mathrm{T}}A^{\mathrm{T}}A(x + tv) - (x + tv)^{\mathrm{T}}A^{\mathrm{T}}b - b^{\mathrm{T}}A(x + tv) + \|b\|_2^2 =$$

$x^{\mathrm{T}}A^{\mathrm{T}}A(x-tv)+tv^{\mathrm{T}}A^{\mathrm{T}}A(x+tv)-x^{\mathrm{T}}A^{\mathrm{T}}b-tv^{\mathrm{T}}A^{\mathrm{T}}b-b^{\mathrm{T}}Ax-tb^{\mathrm{T}}Av+\|b\|_2^2 =$

$x^{\mathrm{T}}A^{\mathrm{T}}Ax+tx^{\mathrm{T}}A^{\mathrm{T}}Av+tv^{\mathrm{T}}A^{\mathrm{T}}Ax+t^2v^{\mathrm{T}}A^{\mathrm{T}}Av-x^{\mathrm{T}}A^{\mathrm{T}}b-tv^{\mathrm{T}}A^{\mathrm{T}}b-b^{\mathrm{T}}Ax-tb^{\mathrm{T}}Av+\|b\|_2^2 =$

$t^2v^{\mathrm{T}}A^{\mathrm{T}}Av+t[x^{\mathrm{T}}A^{\mathrm{T}}Av+v^{\mathrm{T}}A^{\mathrm{T}}Ax-v^{\mathrm{T}}A^{\mathrm{T}}b-b^{\mathrm{T}}Av]+x^{\mathrm{T}}A^{\mathrm{T}}Ax-x^{\mathrm{T}}A^{\mathrm{T}}b-b^{\mathrm{T}}Ax+\|b\|_2^2 =$

é un polinomio di secondo grado in t, per trovare il minimo valore annullo la derivata:

$$f'_{\mathrm{v,x}}(t) = 2tv^{\mathrm{T}}A^{\mathrm{T}}Av + [x^{\mathrm{T}}A^{\mathrm{T}}Av + v^{\mathrm{T}}A^{\mathrm{T}}Ax - v^{\mathrm{T}}A^{\mathrm{T}}b - b^{\mathrm{T}}Av]$$

notiamo che $x^{\mathrm{T}}A^{\mathrm{T}}Av = (v^{\mathrm{T}}A^{\mathrm{T}}Ax)^{\mathrm{T}}$, ma essendo numeri: $x^{\mathrm{T}}A^{\mathrm{T}}Av = v^{\mathrm{T}}A^{\mathrm{T}}Ax$

$$f'_{\mathrm{v,x}}(t) = 2tv^{\mathrm{T}}A^{\mathrm{T}}Av + 2v^{\mathrm{T}}A^{\mathrm{T}}Ax - 2v^{\mathrm{T}}A^{\mathrm{T}}b$$

esiste un t^* che minimizza $f_{\mathrm{v,x}}(\cdot)$ e varia con x, v. Mi accorgo che se prendo $x = x^*$ sol. delle equazioni normali ho $f'_{\mathrm{v,x}^*}(t) = 2tv^{\mathrm{T}}A^{\mathrm{T}}Av + 2v^{\mathrm{T}}(A^{\mathrm{T}}Ax^* - A^{\mathrm{T}}b)$ e ho un minimo per $f_{\mathrm{v,x}^*}$ in $t = 0$. Questo vale $\forall v \in \mathbb{R}^{\mathrm{n}} - [0]$ cioé:

$$f_{\mathrm{v,x}^*}(0) < f_{\mathrm{v,x}^*}(t)\ \forall t \neq 0 \in \mathbb{R}, \forall v \in \mathbb{R}^{\mathrm{n}} - [0]$$

$$F(x^* + 0v) < F(x^* + tv)\ \forall t \in \mathbb{R}, \forall v \in \mathbb{R}^{\mathrm{n}} - [0]$$

Dunque se x^* risolve le eq. normali minimizza F su $\mathbb{R}^{\mathrm{n}}$. Ma $\exists$ una soluzione di $A^{\mathrm{T}}Ax = A^{\mathrm{T}}b$? Si e dipende ancora dalle ipotesi:

- $A^{\mathrm{T}}A$ é simmetrica $\Rightarrow A^{\mathrm{T}}A = U^{\mathrm{T}}\Lambda U$ dove U é ortogonale e Λ é diagonale con autovettori sulla diagonale
- se $\lambda = 0$ fosse autovalore avrei $A^{\mathrm{T}}Ax = 0$ con x autovettore, quindi $x^{\mathrm{T}}A^{\mathrm{T}}Ax = 0$

- peró $x^{\mathrm{T}}A^{\mathrm{T}}Ax = \|Ax\|_2^2 \Rightarrow Ax = 0$ e l'ipotesi di rango pieno implica $x = 0$, quindi x **non é vettore**, cioé $\lambda \neq 0$ e $A^{\mathrm{T}}A$ ha solo **autovettori** $\neq 0$

Quindi deduco da questo che:

- se x^* risolve le equazioni normali allora minimizza F su $\mathbb{R}^{\mathrm{n}}$;
- dato che $A^{\mathrm{T}}A$ é invertibile ne risulta che $A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}b$ ha soluzione ed essa é unica;

che conclude la dimostrazione ■

Il metodo per calcolare $x^* : A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}b$ é il seguente:

- costruisco $A^{\mathrm{T}}A$ e $A^{\mathrm{T}}b$ e risolvo il sistema lineare con uno dei metodi giá visti;
- se $A \in M_{\mathrm{mxn}}$ con $m>n$ e $rk(A) = n$ allora esiste la sua fattorizzazione $QR(qr(A))$ in cui: $QR = A$ con $Q \in M_{\mathrm{mxm}}$ ortogonale e $R \in M_{\mathrm{nxn}}$ triangolare superiore
- costruisco $A^{\mathrm{T}}A = R^{\mathrm{T}}Q^{\mathrm{T}}QR = R_0^{\mathrm{T}}Q_0^{\mathrm{T}}Q_0R_0 = R_0^{\mathrm{T}}R_0$ poiché $Q_0^{\mathrm{T}}Q_0 = \mathbb{1}_{\mathrm{n}}$
- costruisco $A^{\mathrm{T}}b = R^{\mathrm{T}}Q^{\mathrm{T}}b = R_0^{\mathrm{T}}Q_0^{\mathrm{T}}b$
- si ottiene che $A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}b \Rightarrow R_0^{\mathrm{T}}R_0x^* = R_0^{\mathrm{T}}Q_0^{\mathrm{T}}b$, dato che R_0 é invertibile il sistema si semplifica a $R_0x^* = Q_0^{\mathrm{T}}b$ e dato che R_0 é triangolare superiore il sistema si puó risolvere con sostituzione all'indietro

Il metodo QR é potenzialmente piú pesante in quanto $m>>n$ ma é anche molto piú accurato. Il metodo QR ci permette di ridurre il $Cond$, infatti senza QR avremo $Cond(A^{\mathrm{T}}A, \|\cdot\|) \approx 10^{12}$ mentre con QR avremo $Cond(R_0, \|\cdot\|) \approx 10^6$, ovvero $Cond(A^{\mathrm{T}}A, \|\cdot\|)$ risulta essere il quadrato di $Cond(R_0, \|\cdot\|)$ (dimostrato in NoteL10).

12 Ricerca degli zeri di funzione

Il problema della ricerca degli zeri ci chiede di trovare $x^* \in \mathbb{R} : f(x^*) = 0$ data $f : [a, b] \to \mathbb{R}$ continua.

(Th. Degli zeri) Se $f(a)f(b)<0$ allora $\exists x^* \in [a,b]$ con $f(x^*) = 0$ Il th. degli zeri fornisce una condizione sufficiente ma non necessaria, ad esempio in $f : [-1,1], f = x^2$ il th. degli zeri non funziona.

12.1 Metodo di bisezione

Il teorema degli zeri ci fornisce un metodo numerico:

- $a_{\mathrm{k}+1} = a_{\mathrm{k}}$ se $f(a_{\mathrm{k}})f(\frac{a_{\mathrm{k}}+b_{\mathrm{k}}}{2}) \leq 0$ altrimenti $\frac{a_{\mathrm{k}}+b_{\mathrm{k}}}{2}$
- $b_{\mathrm{k}+1} = \frac{a_{\mathrm{k}}+b_{\mathrm{k}}}{2}$ se $f(a_{\mathrm{k}})f(\frac{a_{\mathrm{k}}+b_{\mathrm{k}}}{2}) \leq 0$ altrimenti b_{k}

a_{k} é una successione crescente limitata da sopra $a_{\mathrm{k}}<b \quad \forall k$, mentre b_{k} é una successione decrescente limitata da sotto $b_{\mathrm{k}}>a \quad \forall k$. Studiando $\left|a^{\mathrm{k}+1} - b^{\mathrm{k}+1}\right|$ noto che:

$$\left|a^{\mathrm{k}+1} - b^{\mathrm{k}+1}\right| = \frac{1}{2}\left|a^{\mathrm{k}} - b^{\mathrm{k}}\right| = \frac{1}{2} \cdot \frac{1}{2}\left|a^{\mathrm{k}-1} - b^{\mathrm{k}-1}\right| = 2^{-\mathrm{k}}\left|a-b\right| \to 0$$

Quindi so che $x_1 = x_2 = x^*$:

- $f(x^*) = \lim_{\mathrm{k}\to+\infty} f(a_{\mathrm{k}})$
- $f(x^*) = \lim_{\mathrm{k}\to+\infty} f(b_{\mathrm{k}})$
- $0 \leq f(x^*)^2 = \lim_{\mathrm{k}\to+\infty} f(a_{\mathrm{k}})f(b_{\mathrm{k}}) \leq 0$

ovvero ne risulta che $f(x^*) = 0$ ■

12.2 Condizionamento nella ricerca degli zeri

In questo problema il "dato" é la funzione f continua in $[a,b] \in \mathbb{R}$, che appartiene a $\mathbb{C}^0([a,b])$ ovvero l'insieme delle funzioni continue in $[a,b] \in \mathbb{R}$. Notiamo che $\mathbb{C}^0([a,b])$ é uno spazio vettoriale, sul quale possiamo definire la **norma uniforme**:

$$\|f\|_{\mathrm{u}} = \sup_{\mathrm{x}\in[\mathrm{a},\mathrm{b}]} |f(x)| = \max_{\mathrm{x}\in[\mathrm{a},\mathrm{b}]} |f(x)|$$

la quale rispetta le caratteristiche principali di una norma.

12.3 Condizionamento delle radici di f

Supponiamo che $\tilde{f} : [a,b] \longrightarrow \mathbb{R}$ e $\|\tilde{f} - f\|_{\mathrm{u}} < \epsilon$ e consideriamo $x^* \in [a,b]$: $f(x^*) = 0$ e $\tilde{x} \in [a,b] : f(\tilde{x}) = 0$, allora:

$$|f(\tilde{x})| = \left|f(\tilde{x}) - \tilde{f}(\tilde{x})\right| \leq \sup_{\mathrm{x}\in[\mathrm{a,b}]} \left|f(x) - \tilde{f}(x)\right| = \|f - \tilde{f}\|_{\mathrm{u}} < \epsilon$$

Supponiamo che $f \in \mathbb{C}^1([a,b])$, allora:

$$f(\tilde{x}) = f(x^*) + f^{'}(\xi)(\tilde{x} - x^*)$$

dove $\xi = \tilde{x}x^*$ ovvero il segmento aperto che congiunge $\tilde{x}$ e x^*.
Unendo le ultime due formule ottengo:

$$|f(\tilde{x})| \leq \|f - \tilde{f}\|_{\mathrm{u}} = \left|f^{'}(\xi)\right| |\tilde{x} - x^*|$$

Se $f^{'}(\xi) \neq 0$ allora posso dividere e diventa:

$$|\tilde{x} - x^*| \leq \|f - \tilde{f}\|_{\mathrm{u}} \frac{1}{|f^{'}(\xi)|}$$

dove ξ é dentro il segmento aperto. Se $f^{'}(x^*) \neq 0$ allora $f^{'}(\xi) \neq 0$ in un intorno I di x^*, dunque per $\tilde{x}$ sufficientemente vicino ad x^* abbiamo:

$$err_{\mathrm{ass}}(\tilde{x}) \leq C \cdot err_{\mathrm{ass}}(\tilde{f})$$

dove $C = \frac{1}{\min_{\mathrm{I}}\left|f^{'}\right|}$. (Si veda esempio chiarificatore nelle NoteL11)

(Def. Molteplicitá di una radice) Sia $f : [a,b] \longrightarrow \mathbb{R}, f \in \mathbb{C}^{\mathrm{k}}([a,b]), k \in \mathbb{N}$ e $x^* \in [a,b]$:

$$f(x^*) = f^{'}(x^*) = ...f^{(\mathrm{k}-1)}(x^*) = 0, f^{\mathrm{k}}(x^*) \neq 0$$

allora x^* ha molteplicitá k, e se x^* ha molteplicitá 1 si dice **radice semplice**. Da questa definizione ricaviamo che:

- se x^* é una radice semplice allora se $\left|f^{'}(x^*)\right| >> 0$ é ben condizionata mentre se $\left|f^{'}(x^*)\right| \approx 0$ é malcondizionata
- se x^* ha molteplicitá >1 allora é malcondizionata

Riassumendo, abbiamo che:

- La soluzione approssimata $|f(\tilde{x})|$ é minore dell'errore nel calcolare f ovvero $\|f - \tilde{f}\|$
- supponendo che x^* abbia molteplicitá k>1 allora posso dire che $f(\tilde{x}) = \frac{f^{(\mathrm{k})}(\xi)}{k!}(\tilde{x} - x^*)^{\mathrm{k}}$ (sviluppo di Taylor per $f(\tilde{x})$)
- mettendo assieme questi due punti otteniamo che:

$$|\tilde{x} - x^*| \leq \sqrt[k]{\frac{k!}{f^{(\mathrm{k})}(\xi)}} \|f - \tilde{f}\|_{\mathrm{u}}^{\frac{1}{\mathrm{k}}}$$

Il metodo di bisezione presenta peró diverse limitazioni forti:

- se f **non cambia segno** non si puó neanche applicare il metodo
- se si puó applicare converge ma **non é molto veloce** (ha una velocitá lineare, cioé dimezza l'errore ad ogni passo)

12.4 Metodo di Newton per la ricerca degli zeri di funzione

Il metodo di Newton é descritto come:

- x_0 valore dato
- $x_{\mathrm{k}+1} = x_{\mathrm{k}} - \frac{f(x_{\mathrm{k}})}{f'(x_{\mathrm{k}})}$

(Th. Convergenza locale per radici semplici) Sia $f \in \mathbb{C}^2([a,b]), x^*$ radice semplice di $f \in [a,b]$, allora esiste un intorno I di x^* tale che il metodo di Newton inizializzato con $x_0 \in I$ converge ad x^*. Inoltre si ha che:

$$\lim_{\mathrm{k}\to+\infty} \frac{|e_{\mathrm{k}+1}|}{|e_{\mathrm{k}}|^2} = \left|\frac{f''(x^*)}{2f'(x^*)}\right|$$

ció permette al metodo di Newton di essere estremamente piú veloce del metodo di bisezione. (Dim. importante del Th. nelle NoteL12)

(Dim. Metodo di Newton) Considero l'espansione di Taylor di $f(x)$ intorno a un punto x_{n}:

$$f(x) = f(x_{\mathrm{n}}) + f'(x_{\mathrm{n}})(x - x_{\mathrm{n}})$$

supponendo che x_{n+1} una migliore approssimazione della radice di $f(x)$ imposto $f(x_{n+1}) \approx 0$:

$$f(x_n) + f^{'}(x_n)(x_{n+1} - x_n) = 0$$

risolvendo per x_{n+1}:

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

che dimostra la formula del metodo di Newton ■

(Dim. Th. di convergenza locale per radici semplici) Considero la formula iterativa del metodo di newton:

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

sostituisco $f(x_n)$ utilizzando la serie di Taylor di $f(x)$ intorno a x^*:

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

dove c é un punto intermedio tra x_n e x^*. Siccome $f(x^*) = 0$ poiché x^* é una radice semplice (ossia $f(x^*) = 0$ e $f^{'}(x^*) \neq 0$) otteniamo:

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

sostituisco questa espressione alla formula del metodo:

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

siccome $f^{'}(x_n) \to f^{'}(x^*)$ vicino alla radice, possiamo dire che $f^{'}(x_n) \approx f^{'}(x^*)$:

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

semplifichiamo il primo termine:

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

l'errore $e_{n+1} = x_{n+1} - x^*$ diventa:

$$e_{n+1} = -\frac{\frac{f''(\xi)}{2}(x_n - x^*)^2}{f'(x^*)}$$

e poiché $e_n = x_n - x^*$ otteniamo:

$$e_{n+1} = -\frac{f''(\xi)}{2f'(x^*)}e_n^2$$

dividendo per e_n^2 e applicando il limite per $n \to +\infty$:

$$\lim_{n\to+\infty} \frac{|e_{n+1}|}{|e_n|^2} = \left|\frac{f''(\xi)}{2f'(x^*)}\right|$$

ovvero l'errore e_{n+1} é proporzionale al quadrato di e_n, e ció implica una velocitá di convergenza maggiore rispetto al metodo di bisezione ■
Possiamo applicare il criterio dello step al metodo di Newton, da cui otteniamo che:

$$|s_k| = |x_{k+1} - x_k| = \left|x_k - \frac{f(x_k)}{f'(x_k)} - x_k\right| = \left|\frac{f(x_k)}{f'(x_k)}\right| < toll$$

se pensiamo che $|f'(x_k)| \approx |f'(x^*)|$ possiamo pensare a questo criterio come un criterio basato sul residuo ma **pesato** dall'inversa della derivata, ovvero:

$$f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) + \frac{f''(\xi_k)}{2}(x_{k+1} - x_k)^2 =$$

$$f'(x_k)(\frac{f(x_k)}{f'(x_k)} - x_k + x_{k+1}) + \frac{f''(\xi_k)}{2}(x_{k+1} - x_k)^2 =$$

$$\frac{f''(\xi_k)}{2}s_k^2 \quad \xi_k \in x_k x^*$$

Imponendo $\mu_k \in x_{k+1}x^*$ su $f(x_{k+1})$:

$$f(x_{k+1}) = f(x_k) + f'(\mu_k)(x_{k+1} - x^*) =$$

$$f'(\mu_k)(x_{k+1} - x^*) =$$

$$f'(\mu_k)e_{k+1}$$

e uguagliando i due termini si ottiene che:

$$e_{k+1} = \frac{f''(\xi_k)}{2f'(\mu_k)} s_k^2$$

Se k $\to +\infty$, sia ξ_k che μ_k tendono a x^* visto che $f \in \mathbb{C}^2$ si ha che:

$$\lim_{k\to+\infty} \frac{\left|f''(\xi_k)\right|}{2\left|f'(\mu_k)\right|} = \frac{\left|f''(x^*)\right|}{2\left|f'(x^*)\right|} = C$$

da cui si ricava che $|e_{k+1}| = C|s_k^2|$. Se mi fermo quando $|s_k| < toll$ ho che $|e_{k+1}| = C \cdot toll^2$ ovvero che Newton si implementa con un ciclo **while**. Osserviamo che dato che Newton é uno schema iterativo e che x^* é un punto fisso.

(Lemma delle contrazioni in $\mathbb{R}$) Sia $g : [a, b] \to \mathbb{R}$ tale che:

- $g([a, b]) \subseteq [a, b]$
- $\exists L < 1 : \forall x, y \in [a, b] \quad |g(x) - g(y)| < L|x - y|$

allora si ha che:

- $\exists! x^* \in [a, b] : g(x^*) = x^*$
- $\forall x_0 \in [a, b]$ la successione generata da $x_{k+1} = g(x_k)$ converge a x^*
- se $g'(x^*) \neq 0$ e $g \in \mathbb{C}^1$ allora $\lim_{k\to+\infty} \frac{|x_{k+1}-x^*|}{|x_k-x^*|} = |g'(x^*)| < 1$
- se $g'(x^*) = g''(x^*) = \cdots = g^{\mathrm{m}-1}(x^*) = 0$ e $g^{\mathrm{m}}(x^*) \neq 0$ allora $\lim_{k\to+\infty} \frac{|x_{k+1}-x^*|}{|x_k-x^*|} = \frac{|g^{\mathrm{m}}(x^*)|}{m!}$

(Dim. Lemma delle contrazioni in $\mathbb{R}$) Siano $g'(x^*) = g''(x^*) = \cdots = g^{\mathrm{m}-1}(x^*) = 0$ e $g^{\mathrm{m}}(x^*) \neq 0$ espando $g(x)$ con la sua approssimazione di Taylor di grado m:

$$g(x) = g(x^*) + g'(x^*)(x - x^*) + \frac{g''(x^*)}{2}(x - x^*)^2 + \cdots + \frac{g^m(\xi)}{m!}(x - x^*)^m$$

siccome la derivazione di $g'(x^*) \ldots g^{m-1}(x^*) = 0$ ottengo che:

$$g(x) - g(x^*) = \frac{g^m(\xi)}{m!}(x - x^*)^m$$

prendo $x = x_k$:

$$|g(x_k) - g(x^*)| = \frac{|g^m(\xi)|}{m!}|x_k - x^*|^m \ = \frac{|g^m(\xi)|}{m!}|e_k|^m$$

siccome $|g(x_k) - g(x^*)| = |x_{k+1} - x^*| = |e_{k+1}|$:

$$|e_{k+1}| = \frac{|g^m(\xi)|}{m!}|e_k|^m$$

divido per $|e_k|^m$ e prendo il limite:

$$\lim_{k\to+\infty} \frac{|e_{k+1}|}{|e_k|^m} = \lim_{k\to+\infty} \frac{|g^m(\xi)|}{m!} = \frac{|g^m(\xi)|}{m!}$$

se x^* é radice semplice di f abbiamo che:

$$g^{'}(x) = (x - \frac{f(x)}{f^{'}(x)}) = 1 - \frac{f^{'}(x)^2 - f(x)f^{''}(x)}{f^{'}(x)^2} = \frac{f(x) \cdot f^{''}(x)}{f^{'}(x)^2}$$

$$g^{'}(x^*) = \frac{f(x^*) \cdot f^{''}(x^*)}{f^{'}(x^*)^2} = 0 \quad m \geq 2$$

$$g^{''}(x) = \frac{[f^{'}(x)f^{''}(x) + f(x)f^{'''}(x)]f(x)^2 - f(x)f^{''}(x) \cdot 2f^{'}(x)f^{''}(x)}{f^{'}(x)^4}$$

$$C = \left|\frac{g^{''}(x^*)}{2}\right| = \frac{|f^{'}(x^*)f^{''}(x^*)f^{'}(x^*)^2|}{2|f^{'}(x^*)|^4} = \frac{|f^{''}(x^*)|}{2|f^{'}(x^*)|}$$

cosa succede se x^* non é semplice?

$$g(x) = \begin{cases} x - \frac{f(x)}{f^{'}(x)} & x \neq x^* \\ x^* & x = x^* \end{cases}$$

Supponiamo che $f \in \mathbb{C}^2([a,b]), \quad f^{'}(x) \neq 0 \quad \forall x \in [a,b] - (x^*)$, se $x \neq x^*$ non ci sono problemi:

$$g^{'}(x) = 1 - \frac{f^{'}(x)^2 - f(x)f^{''}(x)}{f^{'}(x)^2} = \frac{f(x)f^{''}(x)}{f^{'}(x)^2}$$

$f(x) = f(x^*) + f(x^*)(x - x^*) + \frac{f^{''}(\xi)}{2}(x - x^*)^2 = \frac{f^{''}(\xi)}{2}(x - x^*)^2$ per Newton

$$f^{'}(x) = f^{'}(x^*) + f^{''}(\mu)(x - x^*)$$
$$f^{''}(x) = f^{''}(x^*) + f^{'''}(\theta)(x - x^*)$$
$$g^{'}(x) = \frac{\frac{f^{''}(\xi)}{2}(x - x^*)^2 \cdot f^{''}(x^*) + \frac{f^{''}(\xi)}{2} f^{'''}(\theta)(x - x^*)^3}{f^{''}(\mu)^2 (x - x^*)^2} =$$
$$g^{'}(x) = \frac{f^{''}(\xi) f^{''}(x^*)}{2 f^{''}(\mu)^2} + \frac{f^{''}(\xi) f^{'''}(\theta)}{2 f^{''}(\mu)^2}(x - x^*) \quad \xi, \mu = x^* \text{ per } x \to x^*$$
$$\lim_{x \to x^*} g^{'}(x) = \frac{1}{2} + 0$$

quindi si ottiene che:

$$g^{'}(x) = \begin{cases} \frac{f^{''}(x) f(x)}{2 f^{'}(x)^2} & x \neq x^* \\ \frac{1}{2} & x = x^* \end{cases} \quad \textbf{prolungamento per continuitá}$$

$|g^{'}(x) = \frac{1}{2}|$, se $g \in \mathbb{C}^2$ in x^* allora esiste un intorno $[x^* - \delta, x^* + \delta]$ per cui $|g^{'}(x) < 1| \quad \forall x \in I_\delta$ per δ sufficientemente piccolo $g(I_\delta) \subseteq I_\delta$ e dunque per il lemma delle contrazioni se $x_0 \in I_\delta$ allora $x_{k+1} = f(x_k)$ converge a x^* ■

Ne ricaviamo che per radici di molteplicitá > 1 il metodo di Newton ha ancora carattere di convergenza locale, ma perde la **convergenza quadratica** e avremo:

$$\lim_{k \to +\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \frac{1}{2}$$

A volte si puó conoscere la molteplicitá di una radice dalla specifica applicazione pratica, in questo caso possiamo usare il **Metodo di Newton modificato**:

$$x_{k+1} = x_k - m \frac{f(x_k)}{f^{'}(x_k)} \quad \text{m = molteplicitá}$$

12.5 Metodo della secante e metodi Newton-like

Se non sappiamo scrivere $f^{'}$ oppure é troppo costoso farlo, possiamo usare dei metodi strutturati come:

$$x_{k+1} = x_k - \frac{f(x_k)}{P_k} \quad P_k \approx f^{'}(x_k)$$

I due metodi di cui parliamo sono:

- $P_k = f^{'}(x_0) \quad \forall k \in \mathbb{N}$ ovvero il **Metodo della tangente fissa**
- $P_k = \frac{f(x_k)-f(x_{k-1})}{x_K - x_{k-1}}$ ovvero il **Metodo delle secanti variabili** con ordine di convergenza per **radici semplici**: $\frac{1+\sqrt{5}}{2}$

13 Interpolazione di funzioni

Dato $x \longrightarrow f(x)$ e disponendo di misurazioni $(x_i, f(x_i) = y_i) : i = 0 \dots n$ voglio costruire un modello del fenomeno

$$x \longrightarrow g(x) \qquad \text{(g simile ad f)}$$

con le seguenti caratteristiche:

- facile da **costruire**;
- g semplice in **forma** e nelle **operazioni**;

(Def.) Si dice che g **interpola** i dati (x_i, y_i) se vale che $g(x_i) = y_i \quad \forall i$. Potrebbe darsi che $f : [a, b] \longrightarrow \mathbb{R}$ sia una funzione **nota o calcolabile** ma molto complessa (come l'output di un algoritmo). In tal caso si puó voler produrre una funzione:

$$g : [a, b] \longrightarrow \mathbb{R} : g \approx f$$

in un senso opportuno (dipendente dall'**applicazione specifica**).
Tale funzione g viene detta **modello surrogato** e puó essere costruito per interpolazione, ossia:

$$g(x_i) = f(x_i) \quad i = 0 \dots n$$

Considero $\phi_0, \phi_1, \dots, \phi_n \in \mathbb{C}^0([a, b])$ dove:

$$\Phi = span.(\phi_0, \dots, \phi_n) \quad \text{Spazio lineare generato}$$

Se é spazio vettoriale allora ho:

$$X = (x_0, x_1, \dots, x_m) \quad x_i \in [a, b]$$

e cerchiamo $\phi \in \Phi : \phi_i(x_i) = y_i \quad \forall i = 0 \dots m$. Se un tale ϕ esiste allora é un **interpolante** dei dati (x_i, y_i) e si ha che:

$$\phi(x_i) = \sum_{j=0}^{n} c_j \phi_j(x_i) = y_i \quad \forall i = 0 \dots n$$

dove abbiamo che $\phi(x_i)$ é la valutazione di $\phi \in \Phi$ su x_i.

13.1 Matrice di Vandermonde

(Def. Matrice di Vandermonde) Impostando il sistema $Ac = y$ con $A_{i,j} = \phi_j(x_i) \quad i = 0 \dots m \quad j = 0 \dots n, \quad \vec{y} = [y_0, \dots, y_m]$:

$$Ac = \begin{bmatrix} \sum_{j=0}^{n} A_{0,j}c_j \\ \sum_{j=0}^{n} A_{1,j}c_j \\ \dots \\ \sum_{j=0}^{n} A_{m,j}c_j \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{n} \phi_j(x_0)c_j \\ \sum_{j=0}^{n} \phi_j(x_1)c_j \\ \dots \\ \sum_{j=0}^{n} \phi_j(x_m)c_j \end{bmatrix} \tag{1}$$

La matrice $A = \phi_j(x_i) \quad i = 0...m \quad j = 0...n$ é detta **matrice di Vandermonde** o **matrice di interpolazione**.
Da notare come se A sia **quadrata** e **invertibile** allora Ay^{-1} é soluzione **unica** del **sistema di Vandermonde** $Ac = y$.
Se $m > n$ e le colonne di A sono linearmente indipendenti allora esiste soluzione solo se:

$$y \in span(A:0, A:1, ..., A:n)$$

e in tal caso é unica.
(Th.) Nel caso **algebrico** ($\phi_j(x) = x^j, \quad j = 0...n$) della matrice di Vandermonde V con $m = n$, si ha che:

$$\exists V^{-1} \Leftrightarrow x_i \neq x_j \quad \forall i \neq j$$

(Dim.) Sia $A \in \mathbb{M}_{nxn}$ con $m = n$:

$$A = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix}$$

Tolgo ad ogni colonna tranne la prima le precedenti moltiplicate per x_0. Questa operazione corrisponde alla moltiplicazione per matrice triangolare superiore con 1 sulla diagonale, che non modifica il determinante:

$$det = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & (x_1 - x_0) & (x_1^2 - x_1x_0) & \dots & (x_1^n - x_1^{n-1}x_0) \\ \vdots & & & & \\ 1 & (x_n - x_0) & (x_n^2 - x_nx_0) & \dots & (x_n^n - x_n^{n-1}x_0) \end{bmatrix}$$

raccogliendo il termine $(x_n - x_0)$ su ogni riga diventa:

$$det \begin{vmatrix} (x_1 - x_0) \cdot 1 & (x_1 - x_0)x_1 & \dots & (x_1 - x_0)x_1^{n-1} \\ (x_1 - x_0) \cdot 1 & (x_2 - x_0)x_2 & \dots & (x_2 - x_0)x_2^{n-1} \\ \vdots & & & \\ (x_1 - x_0) \cdot 1 & (x_n - x_0)x_n & \dots & (x_n - x_0)x_n^{n-1} \end{vmatrix}$$

che equivale al prodotto di:

$$det \begin{pmatrix} (x_1 - x_0) & 0 & \dots & \\ 0 & (x_2 - x_0) & 0 & \dots \\ \vdots & & & \\ 0 & \dots & 0 & (x_n - x_0) \end{pmatrix} \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{pmatrix}$$

che raccogliendo diventa:

$$\prod_{i=1}^{n}(x_i - x_0) det \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{pmatrix}$$

Si nota che la matrice a destra ha la stessa forma di A ma ha una colonna e una riga **in meno**. Quindi:

$$det.A = \prod_{i_1=1}^{n}(x_{i_1} - x_0) \prod_{i_2=2}^{n}(x_{i_2} - x_1) det \begin{pmatrix} 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ 1 & x_3 & x_3^2 & \dots & x_3^{n-1} \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{pmatrix}$$

Iterando questo passaggio si arriva ad avere:

$$det.A = \prod_{0 \le i < j \le n} (x_j - x_i)$$

da cui osserviamo che:

- se $x_i \neq x_j \quad \forall i \neq j$ allora $det.A$ é il prodotto di termini non nulli $\Rightarrow det.A \neq 0$ cioé A é **invertibile**;
- se $det.A = 0 \Rightarrow$ almeno uno dei fattori deve essere nullo e dunque $x_i = x_j$ per almeno una coppia di indici $(i, j) \quad i \neq j$; ■

Se prendiamo punti distinti pari al grado massimo aumentato di uno la matrice di Vandermonde é **invertibile** e:

$$\exists! \quad p \in P^n \text{ polinomio interpolante di grado} \le \text{n} \ : p(x_i) = y_i$$

13.2 Polinomi di Lagrange

Supponiamo di avere a disposizione $l_0(x), l_1(x), \ldots, l_n(x)$ tali che:

$$l_j(x_i) = \delta_{i,j}$$

dove $\delta_{i,j}$ é il **delta di kronecker** definito come:

- 1 se $i = j$
- 0 se $i \neq j$

con $l_j \in \Phi$. Allora $\phi \in \Phi : \phi(x_i) = y_i$ si scrive:

$$\phi(x) = \sum_{i=0}^{n} y_i l_i(x); \quad \phi(x_k) = \sum_{i=0}^{n} y_i l_i(x_k) = y_k l_k(x_k) = y_k \quad \forall k = 0...n$$

Se dispongo dei l_i il problema dell'interpolazione é risolto, ma devo verificare che esistano $l_i \quad \forall i = 0...n$.

13.3 Base di Lagrange

Se A é invertibile allora la base di Lagrange esiste ed é unica:

$$l_k(x) = \sum_j c_{jk}\phi_j(x) = \begin{bmatrix} l_k(x_0) = \sum_j c_{jk}\phi_j(x_0) \\ l_k(x_1) = \sum_j c_{jk}\phi_j(x_1) \\ \cdots \\ l_k(x_k) = \sum_j c_{jk}\phi_j(x_k) \\ l_k(x_n) = \sum_j c_{jk}\phi_j(x_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 1 \\ 0 \end{bmatrix}$$

Costruisco il vettore $c_{:,k}$ soluzione di $Ac_{:,k} = e_k$ dove e_k é costruito sopra come:

$$c_{:,k} = \begin{bmatrix} c_{0,k} \\ c_{1,k} \\ \cdots \\ c_{n,k} \end{bmatrix}$$

Ma allora si ha che:

$$A[c_{:,0}, c_{:,1}, \ldots, c_{:,n}] \Rightarrow A^{-1} = [e_0, e_1, \ldots, e_n] \Rightarrow \mathbb{1}$$

Quindi $c_{:,k}$ é la k-esima colonna di A^{-1}:

$$l_k(x) = \sum_{j=0}^{n} A^{-1}_{jk} \phi_j(x)$$

Se consideriamo il caso **algebrico** la base di Lagrange é detta **polinomio di Lagrange** l_j, la cui formula é:

$$l_j(x) = \prod_{j \neq i} \frac{x - x_i}{x_j - x_i}$$

13.4 Condizionamento dell'interpolazione

Sia $f : \mathbb{C}^0([a,b])$ e $\tilde{f} \in \mathbb{C}^0([a,b])$ sua approssimazione in **norma uniforme**:

$$\|f - \tilde{f}\|_u = \max_{x \in [a,b]} |f(x) - \tilde{f}(x)|$$

e sia p il polinomio che interpola f su $x_0, \ldots, x_n$:

$$p(x_i) = f(x_i) \quad \forall i = 0, \ldots, n$$

e $\tilde{p}$ il polinomio che interpola $\tilde{f}$ sugli stessi punti:

$$\tilde{p}(x_i) = \tilde{f}(x_i) \quad \forall i = 0, \ldots, n$$

Vogliamo una stima di $\|p - \tilde{p}\|_u$:

$$\|p - \tilde{p}\|_u = \max_{x \in [a,b]} |p(x) - \tilde{p}(x)| =$$

$$\max_{x \in [a,b]} |\sum_{i=0}^{n} (f(x_i) - \tilde{f}(x_i)) l_i(x)| \leq \max_{x \in [a,b]} \sum_{i=0}^{n} |f(x_i) - \tilde{f}(x_i)||l_i(x)|$$

poiché $x_i \in [a,b]$ allora $|f(x_i) - \tilde{f}(x_i)| \leq \|f - \tilde{f}\|_u$ per definizione di norma uniforme, dunque abbiamo che:

$$\max_{x \in [a,b]} \sum_{i=0}^{n} |f(x_i) - \tilde{f}(x_i)||l_i(x)| \leq \|f - \tilde{f}\|_u \max_{x \in [a,b]} \sum_{i=0}^{n} |l_i(x)|$$

quindi ne ricaviamo che:

$$\|p - \tilde{p}\|_u \leq \|f - \tilde{f}\|_u \max_{x \in [a,b]} \sum_{i=0}^{n} |l_i(x)|$$

in cui $\max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|$ é la **massima amplificazione dell'errore**.

(Def. Costante di Lebesgue) Dalla precedente disuguaglianza definiamo la **costante di Lebesgue** come:

$$\Lambda(x_0, x_1, \ldots, x_n, [a,b]) = \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|$$

che ci permette di misurare il **condizionamento assoluto** del problema dell'interpolazione. Alcune osservazioni da fare su questa costante:

- costanti di Lebesgue **molto grandi** distruggono la qualitá dell'interpolante;
- necessariamente $\Lambda \to +\infty$ se $n \to +\infty$ indipendentemente da come scelgo i nodi;
- la velocitá con cui $\Lambda \to +\infty$ se $n \to +\infty$ dipende da come scelgo i nodi;

Sia $f \in \mathbb{C}^0([a,b])$ e considero $x_0^{(n)}, \ldots, x_n^{(n)}$, ponendo $n \to +\infty$ cosa succede a $\|f - p_n\|_u$ dove p_n é l'interpolante di f su $x_0^{(n)}, \ldots, x_n^{(n)}$? Vorremo poter concludere che $p_n \to f$ ma in realtá si misura che $\|f - p_n\|_u \to 0$ se $n \to +\infty$.

Fissati i nodi di interpolazione per ogni grado posso introdurre l'**operatore di interpolazione** I_n:

$$\mathbb{C}^0([a,b]) \longrightarrow P^n$$

$$f \longrightarrow \sum_{i=0}^{n} l_i(x) f(x_i^{(n)})$$

Alcune osservazioni utili su I_n:

- I_n é **lineare** ossia $I_n(af + bg) = aI_n(f) + bI_n(g) \quad \forall a, b \in \mathbb{R} \quad \forall f, g \in \mathbb{C}^0([a,b])$;
- I_n é un **operatore di proiezione** su P^n ossia $I_n(p) = p$ se $p \in P^n$;

$I_n(p)$ é l'unico polinomio di grado $\leq n$ che vale $p(x_i^{(n)})$ in $x_i^{(n)}$ per $i = 0, \ldots, n$ ma anche p soddisfa la stessa proprietá dunque $I_n(p) = p$.

13.5 Teoremi importanti

13.5.1 Teorema di approssimazione di Weiestrass

(Th.) Se $f \in \mathbb{C}^0([a,b])$ allora $\quad \forall \epsilon > 0 \quad \exists p_\epsilon$ polinomio tale che $\|f - p_\epsilon\|_u < \epsilon$.
Da notare come questo teorema non ci dia informazioni sul grado di p_ϵ.

13.5.2 Teorema di Jackson

(Th.) Se $f \in \mathbb{C}^k([a,b])$ e $k \geq 1$ allora $\exists c$:

$$\inf_{p \in P^n} \|f - p\| \leq c \|f^{(k)}\|_u n^{-k}$$

dove:

- $\inf_{p \in P^n} \|f - p\|$ é quanto bene riesco ad approssimare f con un polinomio di grado al piú n;
- $\|f^{(k)}\|_u$ misura di quanto f é regolare;

13.6 Stima di Lebesgue dell'errore di interpolazione

Partiamo dall'uguaglianza:

$$\|f - I_n(f)\|_u = \|f - p_n + p_n - I_n(f)\|_u$$

in cui p_n é il polinomio di grado $< n$ che realizza:

$$\|f - p\|_u = \min_{q \in P^n} \|f - q\|_u$$

per cui il nostro q esiste davvero e non é unico abbiamo che:

$$\|f - p_n + p_n - I_n(f)\|_u \leq \|f - p_n\|_u + \|p_n - I_n(f)\|_u =$$

$$\|f - p_n\|_u + \|I_n(p_n) - I_n(f)\|_u =$$

$\|f - p_n\|_u + \|I_n(p_n - f)\|_u$ poiché I_n é operatore di proiezione

Ora espandiamo il secondo termine:

$$(I_n(p_n - f))(x) = \sum_{i=0}^{n} (p_n(x_i^{(n)}) - f(x_i^{(n)})) l_i(x) \Rightarrow$$

$$\|I_n(p_n - f)\|_u = \max_{x\in[a,b]} |\sum_{i=0}^{n}(p_n(x_i^{(n)}) - f(x_i^{(n)}))l_i(x)| \leq$$

$$\max_{x\in[a,b]} \sum_{i=0}^{n} |p_n(x_i^{(n)}) - f(x_i^{(n)})||l_i(x)|$$

osservo poi che:

$$|p_n(x_i^{(n)}) - f(x_i^{(n)})| \leq \|p_n - f\|_u \quad \forall i = 0, \ldots, n \text{ perché } x_i^{(n)} \in [a, b]$$

quindi abbiamo che:

$$\max_{x\in[a,b]} \sum_{i=0}^{n} |p_n(x_i^{(n)}) - f(x_i^{(n)})||l_i(x)| \leq \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|\|p_n - f\|_u$$

Unendo tutto quello che abbiamo scritto otteniamo che:

$$\|f - I_n(f)\|_u \leq \|f - p_n\|_u + \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|\|p_n - f\|_u \leq$$

$$(1 + \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|)\|p_n - f\|_u =$$

$$(1 + \Lambda([a, b], x_0^{(n)}, \ldots, x_n^{(n)})) \min_{p\in P^n} \|f - p\|_u$$

Facendo ció abbiamo spezzato la stima in due fattori:

- $1+\Lambda$ dipende solo da $[a, b], x_0, \ldots, x_n$ dunque $1+\Lambda \to +\infty$ se $n \to +\infty$;
- $\min_{p\in P^n} \|f - p\|_u$ dipende solo da f, ad esempio Jackson mi dice che questo fattore $\to 0$ se f é almeno $\mathbb{C}^1$

Ci eravamo chiesti se $\|f - I_n(f)\|_u \to 0$ se $n \to +\infty$: ció dipende da quali nodi scelgo a da quanto "liscia" é f.

13.6.1 Nodi cattivi

Un classico esempio di nodo cattivo é:

$$[a, b] = [-1, 1] \quad x_i^n \text{ nodi equspaziati}$$

poiché ottengo:

$$\Lambda_n = \Lambda([-1, 1], x_0^n, \ldots, x_n^n) \; e^{n^L}$$

13.6.2 Nodi buoni

Un esempio di nodi buoni sono i nodi di **Chebyshev** in $[-1, 1]$:

$$\cos \frac{2\pi}{2n} i + \frac{\pi}{2n} \quad i = 0, \ldots, 2n-1 \text{ non contengono gli estremi}$$

da cui si ottiene: $\Lambda_n \; 1 + log(n)$.
Un altro esempio sono i nodi di **Chebyshev-Lobatto**:

$$\cos \frac{\pi}{n} i \quad i = 0, \ldots, n$$

da cui si ottiene $\Lambda_n \; a + log(n)$ con $a \approx 2$.

13.7 Matrice di Vandermonde rettangolare

Quello che interessa a noi nella pratica é il valore di p su punti di valutazione.
Pensiamo di fissare:

$$x_0^{eval}, x_1^{eval}, \ldots, x_N^{eval} \quad N >> n+1$$

allora la matrice dei polinomi interpolatori p diventa:

$$\begin{bmatrix} p(x_0^{eval}) \\ p(x_1^{eval}) \\ \ldots \\ p(x_N^{eval}) \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^n l_j(x_0^{eval}) y_j \\ \sum_{j=0}^n l_j(x_1^{eval}) y_j \\ \ldots \\ \sum_{j=0}^n l_j(x_N^{eval}) y_j \end{bmatrix} = \begin{bmatrix} l_0(x_0^{eval}) \ldots l_n(x_0^{eval}) \\ l_0(x_1^{eval}) \ldots l_n(x_1^{eval}) \\ \ldots \\ l_0(x_N^{eval}) \ldots l_n(x_N^{eval}) \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \ldots \\ y_n \end{bmatrix}$$

Notiamo che la matrice é nella forma:

$$[l_j(x_i^{eval})] \quad i = 0 \ldots N \quad j = 0 \ldots n$$

é una matrice di Vandermonde **rettangolare** rispetto alla base $l_0(x), \ldots, l_n(x)$ di P^n e ai punti $x_0^{eval}, \ldots, x_N^{eval}$ (non vale $l_j(x_i^{eval}) = \delta_{ij}$).
Posso calcolare questa Vandermonde in due modi:

- usando la formula: $l_j(x) = \prod_{j \neq i} \frac{x - x_i}{x_j - x_i}$, ma ció ha un costo computazionale **molto alto** oltre a possedere possibili **instabilitá**;
- uso la proprietá dei polinomi di Lagrange per cui $l_j(x) = \sum_{k=0}^n V_{k,j}^{-1} x^k$, questo se V é matrice di Vandermonde nella base $x^0, \ldots, x^n$ e relative ai punti $x_0, \ldots, x_n$

Dunque abbiamo che:

$$\begin{bmatrix} l_j(x_0^{eval}) \\ l_j(x_1^{eval}) \\ \dots \\ l_j(x_N^{eval}) \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^{n} V_{k,j}^{-1}(x_0^{eval})^k \\ \sum_{k=0}^{n} V_{k,j}^{-1}(x_1^{eval})^k \\ \dots \\ \sum_{k=0}^{n} V_{k,j}^{-1}(x_N^{eval})^k \end{bmatrix} = [(x_i^{eval})^k]V_{:,j}^{-1} \text{ j-esima colonna di } V^{-1}$$

Quindi in conclusione:

$$[l_j(x_i^{eval})] = V^{eval} \cdot V^{-1} \text{ dove } [V^{eval}]_{i,k} = (x_i^{eval})^k$$

Notiamo che nella formula per $A = [l_j(x_i^{eval})]$ c'é V^{-1}:

$$A = V^{eval} \cdot V^{-1} \Leftrightarrow A^T = V^{-T} \cdot (V^{eval})^T \Leftrightarrow V^T A^T = (V^{eval})^T$$

$$(V^T A^T)_{:,j} = V^T (A^T)_{:,j}$$

Posso calcolare le colonne di A^T una alla volta come soluzioni di $V^T x = (V^{eval})_{:,j}^T$ con le seguenti strategie:

- LU: $PV^T = LU$ con un costo di $O(n^3)$ diventa:

  $$V^T x = (V^{eval})_{:,j}^T \Leftrightarrow PV^T x = P(V^{eval})_{:,j}^T \Leftrightarrow LUx = P(V^{eval})_{:,j}^T$$

  la sostituzione costa $O(n^2)$:

  $$y = Ux \quad Ly = P(V^{eval})_{:,j}^T \quad Ux = y$$

  i due passi hanno un costo simile ma il secondo (la sostituzione) puó essere **parallelizzata**;

- QR: $V = QR$ é piú **stabile** di LU e si puó innestare se $Cond(V) >> 1$ allora $V \approx QR$ e uso la R come cambio di base. Moltiplicare a dx per una matrice la V vuol dire sostituire ogni colonna con una **combinazione lineare** delle altre:

  $$V^T x = y \Leftrightarrow R^{-T} V^T x = R^{-T} y$$

  e se V fosse davvero $V = QR$ allora avrei:

  $$R^{-T} V^T = R^{-T} R^T Q^T \Rightarrow R^{-T} V^T = Q^T$$

  matrice ortogonale **molto ben condizionata**. L'idea é di calcolare una volta QR, usare R^{-T} come precondizionamento e poi risolvere $R^{-T} V^T x = y$ con LU o con QR

Se uso QR la prima volta per il precondizionamento e la seconda volta per risolvere ho una soluzione esatta a precisione di macchina se:

$$Cond(V) < \frac{1}{\epsilon_{MACH}}$$

Si possono incontrare affermazioni del tipo:
"La matrice di Vandermonde é tipicamente **malcondizionata**, si sconsiglia il suo utilizzo per risolvere problemi di interpolazione, soprattutto con $n >> 1$".
Ció dipende fortemente da:

- i **nodi di interpolazione**;
- le **basi scelte** per i polinomi;

Se la base é adatta ai punti ho condizionamento piccolo, in generale **puó** essere enorme.

13.8 Rappresentazione dell'errore di interpolazione

Sia $f \in \mathbb{C}^{n+1}([a,b])$ con $x_0, \ldots, x_n$ nodi distinti. Sia $E_n f(x) = f(x) - p(x)$, dove $p \in P^n$ che interpola f su $x_0, \ldots, x_n$, allora:

$$\forall x \in [a,b]\ \exists \xi_x \in [a,b] : E_n f(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$$

(Dim.) Partendo da:

$$E_n f(x) = f(x) - p(x))$$

poniamo $G(z) : [a,b] \longrightarrow \mathbb{R}$ come:

$$G(z) = E_n f(z) - \frac{\prod(z - x_i) \cdot E_n f(x)}{\prod_{i=0}^{n}(x - x_i)} \quad x \in [a,b] \text{ fissato}$$

scegliendo un punto x_j punto di interpolazione otteniamo che:

$$G(x_j) = E_n f(x_j) - \frac{\prod_{i=0}^{n}(x_j - x_i) \cdot E_n f(x)}{\prod_{i=0}^{n}(x - x_i)} = 0 - 0 = 0$$

poiché x_j é di interpolazione e poiché nel prodotto c'é anche $i = j$. Si ottiene che:

$$G(x) = E_n f(x) - \frac{\prod_{i=0}^{n}(x - x_i) \cdot E_n f(x)}{\prod_{i=0}^{n}(x - x_i)} = 0$$

Si ha che G ha $n+2$ zeri in $[a,b]$. Per il Th. di Rolle $G^{'}$ ha $n+1$ zeri in $[a,b]$ (poiché $f \in \mathbb{C}^1([a,b]) \quad f(x_0) = f(x_1) \Rightarrow \exists \xi : f^{'}(\xi) = 0$), quindi per Rolle $G^{''}$ ha n zeri. In generale diciamo che $G^{(n+1)}$ ha 1 zero in $[a,b]$ chiamato ξ_x:

$$0 = G^{(n+1)}(\xi_x) = [(\frac{d}{dz})^{(n+1)}(f(z)-p(z))-(\frac{d}{dz})^{(n+1)}(\prod_{i=0}^{n}(z-x_i))\cdot\frac{E_nf(x)}{\prod_{i=0}^{n}(x-x_i)}]$$

dove $z = \xi_x$ otteniamo che:

$$= (\frac{d}{dz})^{(n+1)}f(\xi_x) - (\frac{d}{dz})^{(n+1)}p(\xi_x) - (\frac{d}{dz})^{(n+1)}(\prod_{i=0}^{n}(z-x_i)) \cdot \frac{E_nf(x)}{\prod_{i=0}^{n}(x-x_i)}$$

Analizzando ogni singolo componente:

$$\prod_{i=0}^{n}(z-x_i) = z^{n+1} + c_nz^{n+1} + \dots$$

$$(\frac{d}{dz})^{(n+1)}(\prod_{i=0}^{n}(z-x_i)) = (\frac{d}{dz})^{(n+1)}z^{n+1} + (\frac{d}{dz})^{(n+1)}(\dots) = 0$$

$$z^{n+1} = (n+1)z^n = (n+1)nz^{n-1} = \dots = (n+1)!$$

otteniamo alla fine che:

$$f^{(n+1)}(\xi_x) - 0 - (n+1)! \cdot \frac{E_nf(x)}{\prod_{i=0}^{n}(x-x_i)}$$

$$0 = f^{(n+1)}(\xi_x) - (n+1)! \cdot \frac{E_nf(x)}{\prod_{i=0}^{n}(x-x_i)} \Rightarrow f^{(n+1)}(\xi_x) = (n+1)! \cdot \frac{E_nf(x)}{\prod_{i=0}^{n}(x-x_i)}$$

e si ha che:

$$E_nf(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \cdot \prod_{i=0}^{n}(x-x_i) \quad \blacksquare$$

13.9 Rappresentazione dell'errore di approssimazione

In molte applicazioni capita di avere molti dati (componenti di una funzione):

$$x_i \quad y_i = f(x_i) \quad i = 0, \dots, N >> 1$$

potremo non avere controllo su **dove sono** gli x_i e potremo avere $N >> 1$: vogliamo cercare $p \in P^n : p(x_i) = y_i$.
Un possibile approccio é il seguente: fissato $n < N$ cerco $p \in P^n$ tale che:

$$\sum_{i=0}^{N} |p(x_i) - y_i|^2 = \min_{q \in P^n} (\sum_{i=0}^{N} |q(x_i) - y_i|^2)$$

che rappresenta la **somma degli scarti quadratici**. Il polinomio p che risolve tale formula é detto **soluzione ai minimi quadrati lineari** ma ancora non sappiamo se esiste e se é unico.
Fissata una base $P^n : \phi_0, \phi_1, \ldots, \phi_n \quad \phi_k(x) = x^k$:

$$\sum_{i=0}^{N} |q(x_i) - y_i|^2 = \sum_{i=0}^{N} |\sum_{i=0}^{n} (c_j \phi_j(x_i)) - y_i|^2 =$$

$$V = [\phi_j(x_i)] \quad j = 0, \ldots, n \quad i = 0, \ldots, N$$

$$(Vc)_i = \sum_{j=0}^{n} c_j V_{i,j} = \sum_{j=0}^{n} \phi_j(x_i) c_j$$

$$= \sum_{i=0}^{N} ((Vc)_i - y_i)^2 = \|Vc - y\|_2^2$$

quindi stiamo minimizzando $\|Vc-y\|_2^2$ su $c \in \mathbb{R}^{n+1}$ cioé vogliamo risolvere ai minimi quadrati il sistema lineare **sovradeterminato** $Vc = y$. Osserviamo che se il rango di V $rk.V = n+1$ cioé **massimo** allora il Th. delle proiezioni ortogonali implica che:

- $\sum_{i=0}^{N} |p(x_i) - y_i|^2 = \min_{q \in P^n} (\sum_{i=0}^{N} |q(x_i) - y_i|^2)$ ha soluzione unica $p(x) = \sum_{j=0}^{n} \phi_j(x) c_j^*$
- c^* é l'unica soluzione delle equazioni normali $V^T V c^* = V^T y$

13.9.1 Generalizzazione dei minimi quadrati pesati

Data:

$$\sum_{i=0}^{N} |p(x_i) - y_i|^2 = \min_{q \in P^n} (\sum_{i=0}^{N} |q(x_i) - y_i|^2)$$

costruiamo $W = diag(\sqrt{w_0}, \sqrt{w_1}, \ldots, \sqrt{w_N})$ con $w_i \geq 0$ e scopriamo che:

$$\sum_{i=0}^{N} |p(x_i) - y_i|^2 = \min_{q \in P^n} (\sum_{i=0}^{N} |q(x_i) - y_i|^2) \Rightarrow \min_{c \in \mathbb{R}^{n+1}} \|WVc - Wy\|_2^2$$

e dunque abbiamo una soluzione unica se WV ha rango $n+1$.
Cosa possiamo dire di $rk.V$ e $rk.WV$? Se $\exists\, i_0, i_1, \ldots, i_n$ tali che:

$$\begin{bmatrix} V_{i_0,:} \\ V_{i_1,:} \\ \ldots \\ V_{i_n,:} \end{bmatrix} n+1 \text{ é invertibile}$$

allora la V di potenza ha $rk. = n+1$. La matrice di partenza é una Vandermonde quando seleziono $i_0, i_1, \ldots, i_n$ come voglio ma diversi tra loro:

$$\begin{bmatrix} V_{i_0,:} \\ V_{i_1,:} \\ \ldots \\ V_{i_n,:} \end{bmatrix} = \begin{bmatrix} \phi_0(x_{i_0}) \ldots \phi_n(x_{i_0}) \\ \phi_0(x_{i_1}) \ldots \phi_n(x_{i_1}) \\ \vdots \\ \phi_0(x_{i_n}) \ldots \phi_n(x_{i_n}) \end{bmatrix}$$

Osserviamo che $x_{i_k} \neq x_{i_l} \forall k \neq l$ dunque Vandermonde **invertibile** per scelta di indici diversi tra loro ossia V ha rango massimo. Nel caso in cui avessi WV con $W = diag(\sqrt{w_0}, \sqrt{w_1}, \ldots, \sqrt{w_N})$:

- se $(i : w_i > 0) \geq n+1$ ho la stessa situazione
- se $(i : w_i > 0) < n+1$ il ragionamento non funziona

13.10 Prodotti scalari e matrici simmetriche definite

Sia $(\cdot,\cdot) : V \mathrm{x} V \longrightarrow \mathbb{R}$ simmetrica, bilineare e definita positiva. Se fissiamo una base $\phi_0, \phi_1, \ldots, \phi_n$ di V allora possiamo rappresentare il prodotto $(\cdot,\cdot)$ con una matrice **simmetrica definita positiva**:

$$(\sum_{i=0}^{n} c_i\phi_i, \sum_j b_j\phi_j) = \sum_{i=0}^{n} c_i(\phi_i, \sum_{i=0}^{n} b_j\phi_j) =$$

$$= \sum_{i=0}^{n}\sum_{j=0}^{n} c_i b_j(\phi_i, \phi_j) = c^T G b \quad G = [(\phi_i, \phi_j)]_{i,j=0\ldots n}$$

gramiano dal prodotto $(\cdot,\cdot)$ rispetto alla base $(\phi_0, \ldots, \phi_n)$.
Se $(\cdot,\cdot)$ é prodotto scalare su V allora induce una norma su V, basta porre $\|v\| = \sqrt{(v,v)}$. Se $v = \sum_{i=0}^{n} c_i\phi_i$ allora $\|v\| = \sqrt{c^T G c}$:

Al contrario se $(\cdot,\cdot)$ non é definita positivamente ma solo semi definita positivamente $(v,v) \geq 0$ allora $(\cdot,\cdot)$ **non induce** una norma ma solo una **seminorma**:

$$\|\cdot\| : V \longrightarrow [0, +\infty[$$

con le proprietá delle norme tranne $\|v\| = 0 \Rightarrow v = 0$.

13.10.1 Teorema di Pitagora

Sia V uno spazio vettoriale $(\cdot,\cdot)$ prodotto scalare su V e denotiamo con $\|\cdot\|$ la norma indotta da $(\cdot,\cdot)$ allora:

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2(u,v)$$

Il fatto che $\|\cdot\|$ derivi da $(\cdot,\cdot)$ é **fondamentale**.
(Dim.) Espandendo la formula:

$$\|u+v\|^2 = (u+v, u+v) = (u, u+v) + (v, u+v) =$$

$$(u,u) + (u,v) + (v,u) + (v,v) =$$

$$(u,u) + 2(u,v) + (v,v) = \|u\|^2 + 2(u,v) + \|v\|^2$$

13.10.2 Ortogonalitá e ortonormalitá

Se $(u,v) = 0$ diciamo che u e v sono **ortogonali**. Dato V e $(\cdot,\cdot)$ prodotto scalare su V esistono basi **ortogonali** e **ortonormali**.
Se $\psi_0, \psi_1, \ldots, \psi_n$ é base di V allora $(\psi_0, \ldots, \psi_n)$ é detta **base ortogonale** se:

$$(\psi_i, \psi_j) = \delta_{ij} c_i \quad c_i \text{ se } i = j \quad 0 \text{ se } i \neq j \quad c_i > 0$$

Si dice **ortonormale** se:

$$(\psi_i, \psi_j) = \delta_{ij} \quad 1 \text{ se } i = j \quad 0 \text{ se } i \neq j$$

Se $\phi_0, \phi_1, \ldots, \phi_n$ é base di V e G é il **gramiano** di $(\cdot,\cdot)$ in questa base allora per il Th. spettrale $U^T \Lambda U$:

$$\psi_k = \sum_{j=0}^{n} U_{k,j} \phi_j$$

$$(\psi_h, \psi_k) = (\sum_{i=0}^{n} U_{h,i}\phi_i, \sum_{j=0}^{n} U_{k,j}\phi_j) =$$

$$\sum_{i=0}^{n} U_{h,i} \sum_{j=0}^{n} U_{k,j}(\phi_i, \phi_j) \Rightarrow G_{ij} =$$

$$U_{h,i} G(U_{k,i}^T) = U_{h,i} U^T \Lambda U U_{k,i}^T$$

siccome U é ortogonale allora $UU^T = \mathbb{1}$:

$$(\mathbb{1})_{i,k} = (UU^T)_{i,k} = U(U^T)_{i,k} = U(U_{k,i})^T \Rightarrow e_k$$

$$(\psi_h, \psi_k) = U_{h,i} U^T \Lambda U U_{k,i}^T = e_h^T \Lambda e_k = \lambda_h \delta_{h,k}$$

Dunque le $(\psi_h)_{h=0,\dots,n}$ sono una **base ortogonale** e $\|\psi_h\|^2 = \lambda_h$ dunque se poniamo:

$$\tilde{\psi_h} = \frac{\psi_h}{\|\psi_h\|} = \frac{\psi_h}{\sqrt{\lambda_h}}$$

otteniamo una base **ortonormale**.

13.10.3 Identitá di Parseval

Sia $V \in V$ e $\tilde{\psi}_0, \dots, \tilde{\psi}_n$ **ortonormale**, allora:

$$v = \sum_{h=0}^{n} (v, \tilde{\psi_h}) \tilde{\psi_h} \Rightarrow \|v\|^2 = \sum_{h=0}^{n} \|(v, \tilde{\psi_h})\|^2$$

13.11 Teorema delle proiezioni ortogonali versione generale

Sia P sottospazio di C spazio vettoriale, sia $(\cdot, \cdot)$ applicazione lineare su C simmetrica e semi definita positivamente, che é definita positivamente se ristretta a P.
Allora $\forall\, f \in C\ \exists!$ l'elemento p di P tale che:

$$\|f - p\|^2 = \min_{q \in P} \|f - p\|^2$$

Tale p é caratterizzato dalle equazioni normali $(f - p, q) = 0\ \forall\, q \in P$, ovvero che l'errore $f - p$ é **ortogonale** a P.
Posso definire una proiezione Π proiezione ortogonale di f su P:

$$C \longrightarrow P$$

$$f \longrightarrow p \in P : \|f - p\|^2 = \min_{q \in P} \|f - p\|^2$$

Supponiamo di disporre di una base **ortonormale** di P $\psi_0, \ldots, \psi_n$ e scrivo le equazioni normali $(f-p,q)=0 \,\forall\, q \in P$ mi accorgo che posso farlo anche solo con q elemento di base:

$$(f-p,\psi_h)=0 \,\forall\, h$$

$$(f,\psi_h)=(p,\psi_h)=(\sum_{k=0}^{n} c_k\psi_k,\psi_k)=\sum_{k=0}^{n} c_k(\psi_k,\psi_h) \Rightarrow \sum_{k=0}^{n} c_k\delta_{h,k}=c_h$$

$$p=\sum_{k=0}^{n} c_k\psi_k=\sum_{k=0}^{n}(f,\psi_k)\psi_k$$

13.12 Nucleo di riproduzione

Sia $\psi_0, \ldots, \psi_n$ base ortonormale di P e supponiamo che P sia uno spazio di funzioni (es. $P=$ polinomi di grado $\leq n$), allora definisco:

$$K(x,y)=\sum_{k=0}^{n}\psi_k(x)\psi_k(y) \quad \text{ben definito poiché } \psi_k(x)\in\mathbb{R}$$

in cui K é detto **nucleo di riproduzione** ed ha le seguenti proprietá:

$$p(x)=(K(x,\cdot),p(\cdot))$$

$$K(x,\cdot)=\sum \psi_k(x)\psi_k(\cdot)$$

Se $f \in \mathbb{C} \geq P$:

$$\Pi f(x)=(K(x,\cdot),f(\cdot))$$

Usando l'identitá di Parseval:

$$p(\cdot)=\sum_{h=0}^{n}(p,\psi_h)\psi_j(\cdot)$$

$$(K(x,\cdot),p(\cdot))=(\sum_{k=0}^{n}\psi_k(x)\psi_k(\cdot),\sum_{h=0}^{n}(p,\psi_h)\psi_h(\cdot))$$

$$=\sum_{k=0}^{n}\psi_k(x)(\psi_k(\cdot),\sum_{h=0}^{n}(p,\psi_h)\psi_h(\cdot))$$

$$=\sum_{k=0}^{n}\psi_k(x)\sum_{h=0}^{n}(p,\psi_h)\delta_{h,k} \quad \text{poiché base ortonormale}$$

$$=\sum_{k=0}^{n}\psi_k(x)(p,\psi_h)=p(x)$$

13.13 Stima di Lebesgue dell'errore di approssimazione

Sia $P =$ polinomi di grado $\leq n$, $C =$ funzioni continue e $(p,q) = \sum_{i=0}^{N} w_i p(x_i) q(x_i)$ con $N >> n$ punti distinti e $w_i > 0$.
Per il Th. delle proiezioni ortogonali $\exists!\, p$ tale che:

$$\|f-p\|^2 = \min_{q\in P} \|f-p\|^2$$

$$p(x) = (K(x,\cdot), f(\cdot))$$

$$\|f-p\|_u = \|f-p^*+p^*-p\|_u \quad p^* \text{ migliore approx. uniforme di } f \text{ su } P$$

$$\leq \|f-p^*\|_u + \|p^*-p\|_u$$

$$\|p^*-p\|_u \leq ?$$

$$(\Pi f) - p^* = \Pi(f-p^*) \quad \text{perché proiezione}$$

$$\|p^*-p\|_u = \|\Pi(f-p^*)\|_u$$

$$= \max_{x\in[a,b]} |\Pi(f-p^*)(x)| = \max_{x\in[a,b]} |(K(x,\cdot),(f-p^*)(x))|$$

$$= \max_{x\in[a,b]} |\sum_{i=0}^{n} K(x,x_i)(f(x_i)-p^*(x_i))|$$

$$\text{ma dato che } |f(x_i)-p^*(x_i)| \leq \|f-p^*\|_u$$

$$\leq \max_{x\in[a,b]} \sum_{i=0}^{n} |K(x,x_i)| \cdot \|f-p^*\|_u$$

$$\text{siccome } \leq \|f-p^*\|_u + \|p^*-p\|_u$$

$\|f-p\|_u \leq \|f-p^*\|_u(1+\max_{x\in[a,b]} \sum_{i=0}^{n} |K(x,x_i)|)$ equivale alla costante di Lebesgue

14 Quadratura numerica

Sia $[a,b] \in \mathbb{R}$ e $f \in \mathbb{C}^0([a,b])$:

$$\int_a^b f(x)dx =? \approx \sum_{i=1}^{M} f(x_i)w_i \quad \text{formula di quadratura}$$

in cui $x_i \in [a,b]$, $w_i \in \mathbb{R}$. Definiamo le seguenti operazioni lineari su f:

- $I(f,[a,b]) = \int_a^b f(x)dx$
- $Q_{X,W}(f) = \sum_{i=1}^{M} f(x_i)w_i$

14.1 Formule di interpolazione

Consideriamo:

$$\int_a^b f(x)dx \approx \int_a^b p(x)dx$$

dove p interpola f su $x_0, \ldots, x_n \in [a, b]$. L'idea é quella di far diventare questa la formula di quadratura:

$$I(p, [a, b]) = I(\sum_{j=0}^{n} c_j \phi_j(c), [a, b]) \quad \text{dove } \phi_0, \ldots, \phi_n \text{ base di } P$$

$$= \sum_{j=0}^{n} c_j I(\phi_j(x), [a, b])$$

in cui c_j sono i coefficienti del polinomio interpolante e ϕ_j sono una base. Se uso la base di Lagrange, allora ho $c_j = f(x_j)$:

$$I(p, [a, b]) = \sum_{j=0}^{n} f(x_j) I(l_j(x), [a, b])$$

$$\int_a^b f(x)dx \approx \int_a^b p(x)dx = \sum_{j=0}^{n} f(x_j) \int_a^b l_j(x)dx$$

in cui pongo $w_j = \int_a^b l_j(x)dx$.

(Def.) Sia (X, W) una formula di quadratura su $[a, b]$ diciamo che (X, W) ha grado di esattezza polinomiale n se $\forall\, p \in P_n$ vale che:

$$I(p, [a, b]) = Q_{X,W}(p)$$

Nel caso di una formula interpolatoria si ha $X = (x_0, \ldots, x_n)$. Allora si scrive:

$$I(\sum c_j \phi_j, [a, b]) = Q_{X,W}(\sum_j c_j \phi_j) \quad \forall\, c \in \mathbb{R}^{n+1}$$

$$\sum_{j=0}^{n} c_j I(\phi_j, [a, b]) = \sum_{j=0}^{n} c_j Q_{X,W}(\phi_j)$$

Notiamo che se l'ultima equazione vale $\forall$ funzione di base allora vale anche nella forma in cui é scritta:

$$m = I(\phi_j, [a, b]) = Q_{X,W}(\phi_j)$$

$$Q_{X,W}(\phi_j) = \sum_{i=0}^{M} \phi_j(x_i) w_i = V^T w$$

in cui $\phi_j(x_i)$ é matrice di Vandermonde ma sto sommando sull'indice relativo al punto non alla base. Le formule con esattezza n soddisfano:

$$V^T w = m \quad \text{equazioni dei momenti}$$

in cui abbiamo:

- V matrice di Vandermonde sui nodi di quadratura
- w incognita pesi di quadratura
- m vettore dei momenti della base $\int_a^b \phi_j dx = m_j$

14.1.1 Formula del punto medio

La formula del punto medio é:

$$\int_a^b f(x)dx \approx f(\frac{a+b}{2})(b-a)$$

ed é esatta solo sui polinomi di grado 0 e 1.

14.1.2 Formula del trapezio

Dato $n = 1$ cerco una formula interpolativa su $[a, b]$ ad esempio con $x_0 = a, x_1 = b$:

$$\frac{b-a}{2} \cdot p(a) + \frac{b-a}{2} \cdot p(b) \Rightarrow w_0 \cdot p(x_0) + w_1 \cdot p(x_1)$$

in cui ho esattezza uguale a 1 e non a 2.
Una sostituzione utile per calcolare i pesi é la seguente:

$$x = a + (b-a)t \quad dx = (b-a)dt$$

$$\int_a^b f(x)dx = \int_0^1 f(a + (b-a)t)(b-a)dt$$

$$\tilde{f(t)} = f(a + (b-a)t) \quad \tilde{f} \in \mathbb{C}^0([0,1])$$

Dalla formula:

$$I(f,[a,b]) = (b-a)I(\tilde{f},[0,1]) \approx (b-a)\int_0^1 \tilde{p}(t)dt \quad (*)$$

dove $\tilde{p}$ é interpolante di $\tilde{f}$:

$$\tilde{p}(t) = \sum_{j=0}^{n} \tilde{l}_j(t)\tilde{f}(t_j) \quad t_0,\ldots,t_n \in [0,1]$$

$$(*) = (b-a)\sum_{j=0}^{n} \tilde{f}(t_j)\int_0^1 \tilde{l}_j(t)dt = \sum_{j=0}^{n} f(x_j)(b-a)\int_0^1 \tilde{l}_j(t)dt$$

$$\text{dove} \quad x_j = a + t_j(b-a) \quad w_j = (b-a)\int_0^1 \tilde{l}_j(t)dt$$

14.1.3 Formula della parabola

É un particolare esempio di **formule di Newton-Cotes** ossia interpolatorie con nodi **equispaziati**:

$$\int_a^b f(x)dx \approx (b-a)\sum_{j=0}^{2} f(x_j)\int_0^1 \tilde{l}_j(t)dt$$

14.2 Errore nelle formule di quadratura

Cosa possiamo dire sull'errore:

$$|\int_a^b f dx - Q_{X,W}(f)| \leq ?$$

Se $Q_{X,W}$ é interpolatoria possiamo usare la formula di rappresentazione dell'errore:

$$E_n f(x) = f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_n(x)$$

dove p interpola f su $x_0,\ldots,x_n$ e $\omega_n(x) = \prod_{j=0}^{n}(x-x_j)$. Prendiamo l'esempio in cui $n=1$ nella formula del trapezio:

$$\int_a^b f(x)dx - Q_{X,W}(f) = \int_a^b f(x)dx - (b-a)\int_0^1 \tilde{p}(t)dt$$

$$= (b-a)\int_0^1 (\tilde{f}(t) - \tilde{p}(t))dt$$

$$= (b-a)\int_0^1 E_n\tilde{f}(t)dt$$

$$= (b-a)\int_0^1 \frac{f^{(n+1)}(\xi_t)}{(n+1)!}\omega_n(t)dt \quad \text{se } f \in \mathbb{C}^{n+1}([a,b]) \quad (**)$$

Se $n = 1$ allora:

$$\omega_1(t) = (t-0)(t-1) = t^2 - t$$

ossia una parabola che vale 0 in $[0,1]$ cioé che non cambia di segno.

14.2.1 Teorema della media integrale

Siano $f, g \in \mathbb{C}^0([a,b])$, g di segno costante su $[a,b]$ allora $\exists\, c \in [a,b]$ tale che:

$$\int_a^b f(x)g(x)dx = f(c)\int_a^b g(x)dx$$

Considerando $(**)$:

$$\int_a^b f(x)dx - Q_{X,W}(f) = (b-a)\int_0^1 \frac{\tilde{f}''(\xi_t)}{2}\omega_1(t)dt$$

applicando il Th. media integrale:

$$= (b-a)\frac{\tilde{f}''(c)}{2}\int_0^1 \omega_1(t)dt = (b-a)\frac{\tilde{f}''(c)}{2}\int_0^1 t^2 - tdt$$

$$= (b-a)\frac{\tilde{f}''(c)}{2}[\frac{t^3}{3} - \frac{t^2}{2}]_0^1 = \frac{(b-a)\tilde{f}''(c)}{2}\cdot\frac{-1}{6} = -\frac{(b-a)}{12}\tilde{f}''(c)$$

Sapendo che $\tilde{f}''(t) = f''(a + t(b-a))(b-a)^2$:

$$\int_a^b f(x)dx - Q_{X,W}(f) = -\frac{(b-a)^3}{12}f''](c)$$

Da notare come se $f \in P_1$ allora $f''(c) = 0\ \forall\, c \in [a,b]$.

14.2.2 Errore nella formula della parabola con n = 2

$$\int_a^b f(x)dx - Q_{X,W}(f) = (b-a)\int_0^1 \frac{\tilde{f}'''(\xi_t)}{3!}\omega_2(t)dt$$

$$\omega_2(t) = \prod_{j=0}^{2}(t-t_j) = t(t-\frac{1}{2})(t-1) = (t-\frac{1}{2})\cdot\omega_1(t) \quad \text{cambia segno}$$

$$(b-a)\int_0^1 \frac{\tilde{f}'''(\xi_t)}{3!}\omega_2(t)dt = (b-a)[\int_0^{\frac{1}{2}} \frac{\tilde{f}'''(\xi_t)}{3!}\omega_2(t)dt + \int_{\frac{1}{2}}^1 \frac{\tilde{f}'''(\xi_t)}{3!}\omega_2(t)dt]$$

Poniamo $t = 1-s$ e $dt = -ds$ in quanto $\omega_2(1-s) = -\omega_2(s)$:

$$= (b-a)[\int_0^{\frac{1}{2}} \frac{\tilde{f}'''(\xi_t)}{3!}\omega_2(t)dt + \int_{\frac{1}{2}}^0 \frac{\tilde{f}'''(\xi_{1-s})}{3!}(-\omega_2(s))(-1)ds]$$

$$= (b-a)[\int_0^{\frac{1}{2}} \frac{\tilde{f}'''(\xi_t)}{3!}\omega_2(t) - \int_{\frac{1}{2}}^0 \frac{\tilde{f}'''(\xi_{1-s})}{3!}\omega_2(s)ds]$$

Nel secondo integrale effettuo il cambio $s \to t$:

$$= (b-a)[\int_0^{\frac{1}{2}} \frac{(\tilde{f}'''(\xi_t) - \tilde{f}'''(\xi_{1-t}))}{3!}\omega_2(t)dt]$$

Applico il Th. della media integrale poiché ω_2 non cambia di segno in $[0, \frac{1}{2}]$:

$$(b-a)\frac{(\tilde{f}'''(\xi_t) - \tilde{f}'''(\xi_{1-t}))}{3!}\cdot\int_0^{\frac{1}{2}}\omega_2(t)dt$$

in cui usiamo il valore medio (ossia la differenza degli $\tilde{f}$) e calcoliamo l'integrale di ω_2. Stimiamo la differenza della valutazione di f (poniamo per convenienza $\xi_{1-t} \to \eta$:

$$\frac{(b-a)}{3!}\frac{(\tilde{f}'''(\xi) - \tilde{f}'''(\eta))}{\xi-\eta}(\xi-\eta)\cdot\int_0^{\frac{1}{2}}\omega_2(t)dt$$

$$= \frac{(b-a)}{3!}\tilde{f}^{(IV)}(c)(\xi-\eta)\cdot\int_0^{\frac{1}{2}}\omega_2(t)dt \quad (***)\ \tilde{f}^{(IV)} \text{ per Th. di Lagrange}$$

Ricordando che $\omega_2(t) = t(t-\frac{1}{2})(t-1) = t^3 - \frac{3}{2}t + \frac{t}{2}$:

$$\int_0^{\frac{1}{2}}\omega_2(t)dt = [\frac{t^4}{4} - \frac{t^3}{2} + \frac{t^2}{4}]_0^{\frac{1}{2}} = 2^{-6}$$

$$(***) = \frac{(b-a)}{3!2^6}\tilde{f}^{(IV)}(c)(\xi-\eta)$$

Sapendo che $\tilde{f}^{(IV)}(t) = (b-a)^4 f^{(IV)}(a+t(b-a))$:

$$\frac{(b-a)^5}{3!2^6}f^{(IV)}(z)(\xi-\eta) = (\frac{(b-a)}{2})^5 \cdot \frac{1}{3!2}(\xi-\eta)f^{(IV)}(z) \quad z \in [a,b]$$

14.3 Stabilitá della quadratura

Cosa succede se consideriamo $f_\epsilon \approx f$ ad esempio supponiamo $\|f - f_\epsilon\|_u \leq \epsilon$ e cerchiamo di stimare:

$$|Q_{X,W}(f) - Q_{X,W}(f_\epsilon)| = |\sum_{i=1}^{M} f(x_i)w_i - f_\epsilon(x_i)w_i|$$

$$\leq \sum_{i=1}^{M} |(f(x_i) - f_\epsilon(x_i))w_i| = \sum_{i=1}^{M} |f(x_i) - f_\epsilon(x_i)||w_i|$$

Notando che $|f(x_i) - f_\epsilon(x_i)| \leq \epsilon$ poiché $x_i \in [a,b]$ e $\|f - f_\epsilon\|_u \leq \epsilon$:

$$\leq \|f - f_\epsilon\|_u \sum_{i=1}^{M} |w_i| \quad (*)$$

Supponiamo che la formula $Q_{X,W}$ abbia grado di esattezza almeno 0, allora:

$$\sum_{i=1}^{M} w_i = \sum_{i=1}^{M}(1 - w_i) = Q_{X,W}(1) = \int_a^b 1dx = (b-a)$$

$$(*) = |Q_{X,W}(f) - Q_{X,W}(f_\epsilon)| \leq \|f - f_\epsilon\|_u \cdot \frac{\sum_{i=1}^{M}|w_i|}{|\sum_{i=1}^{M} w_i|} \cdot |\sum_{i=1}^{M} w_i|$$

$$= \|f - f\epsilon\|_u (b-a)\frac{\sum_{i=1}^{M}|w_i|}{|\sum_{i=1}^{M} w_i|}$$

in cui $\frac{\sum_{i=1}^{M}|w_i|}{|\sum_{i=1}^{M} w_i|}$ é detto **fattore di stabilitá**.

Cosa succede se $w_i > 0 \, \forall \, i$?

$$\sum_{i=1}^{M} |w_i| = \sum_{i=1}^{M} w_i \quad \text{fattore di stabilitá a 1}$$

Significa che la quadratura a pesi **positivi** é **sempre stabile**, quindi noi vorremo avere sempre delle formule a pesi positivi.

14.3.1 Formule di Newton-Cotes

Sono formule interpolatorie a punti equidistanti, ma per $n \geq 7$ ha pesi che cambiano di segno ossia per $n >> 1$ tende ad essere instabile.

14.4 Formule composte

Sappiamo che:

$$\int_a^b f dx = \sum_{i=1}^{N} \int_{a_i}^{b_i} f(x) dx$$

$$\text{dove: } a_i = a + (i-1) \cdot \frac{b-a}{N} \quad b_i = a + i \cdot \frac{b-a}{N}$$

Possiamo dire che:

$$\int_a^b f dx = \sum_{i=1}^{N} \int_{a_i}^{b_i} f(x) dx \approx \sum_{i=1}^{N} Q_{X_i,W_i}(f)$$

detta **formula composta**.

14.4.1 Formula composta del trapezio

Scegliamo come Q_{X_i,W_i} la formula del trapezio su $[a_i, b_i]$:

- Q_{X_1,W_1} ha come nodi a_1, b_1 e come pesi $(b_i - a_i)\frac{1}{2}, (b_i - a_i)\frac{1}{2}$
- Q_{X_2,W_2} ha come nodi a_2, b_2 e come pesi $(b_2 - a_2)\frac{1}{2}, (b_2 - a_2)\frac{1}{2}$

Da cui possiamo costruire:

$$\sum_{i=1}^{N} Q_{X_i,W_i}(f) = \sum_{i=1}^{N} (f(a + (i-1)\frac{(b-a)}{N}) \cdot (b_i - a_i) + f(a + i\frac{(b-a)}{N})\frac{b_i - a_i}{2})$$

$$b_i - a_i = (a + i\frac{(b-a)}{N}) - (a + (i-1)\frac{(b-a)}{N}) = \frac{b-a}{N}$$

$b_i - a_i$ definito come passo di integrazione h.

$$\sum_{i=1}^{N} \frac{(b-a)}{N} (\frac{f(a + (i-1)\frac{b-a}{N})}{2} + \frac{f(a + i\frac{(b-a)}{N})}{2})$$

Nelle somme competono tutti i nodi interni 2 volte e i due esterni 1 volta.

$$Q_{X_W}(f) = \frac{b-a}{N} (\frac{f(a)}{2} + f(a_2) + f(a_3) + \cdots + f(a_N) + \frac{f(b)}{2})$$

14.4.2 Formula composta della parabola o di Simpson

Possiamo fare le stesse cose con la formula della parabola da cui otteniamo le formule composte di **Simpson**:

- nodi: $a, \frac{a+b}{2}, b$
- pesi: $(b-a)\frac{1}{6}.(b-a)\frac{4}{6}, (b-a)\frac{1}{6}$

Possiamo riscrivere i pesi mettendo in evidenza il passo h: $h(\frac{1}{3}, \frac{4}{3}, \frac{1}{3})$.

14.5 Errore delle formule composte

Sia $Q_{X,W}$ formula composta ottenuta con le formule semplici Q_{X_i,W_i}.

$$\int_a^b f(x)dx - Q_{X,W}(f) = \sum_{i=1}^{N} \int_a^b f(x)dx - \sum_{i=1}^{N} Q_{X_i,W_i}(f)$$

$$= \sum_{i=1}^{N} (\int_a^b f(x)dx - Q_{X_i,W_i}(f))$$

Allora nel caso del trapezio avevamo che:

$$\int_a^b f(x)dx - Q_{X,W}^{TRAP}(f) = -(b-a)^3 \frac{f''(\xi)}{12} \quad (**) \quad \xi \in [a,b]$$

Il trucco delle formule composte é scegliere $-(b-a) = h$ ossia usare sottointervalli $b_i - a_i << 1$ otteniamo un errore che va a 0.
Usiamo $(**)$ su ogni $[a_i, b_i]$:

$$\int_a^b f(x)dx - Q_{X,W}^{TRAP}(f) = h^3 \frac{f''(\xi_i)}{12} \quad \xi_i \in [a_i, b_i]$$

$$= -\frac{h^2}{12} \frac{(b-a)}{N} \sum_{i=1}^{N} f''(\xi_i) \quad (***) \quad \text{media di valutazioni di } f''$$

quindi abbiamo che:

$$\frac{\sum_{i=1}^{N} f''(\xi_i)}{N} \in [\min f''(x), \max f''(x)]$$

dunque per il Th. di Rolle $\exists\, c \in [a,b]$ tale che:

$$f''(c) = \frac{\sum_{i=1}^{N} f''(\xi_i)}{N}$$

$$(***) = \int_a^b f(x)dx - Q_{X,W}(f) = -\frac{(b-a)}{12} f''(c) h^2$$

Se $f \in \mathbb{C}^2([a,b])$ i trapezi composti hanno un errore di ordine $h^2 \Rightarrow$ lento. Possiamo provare ad usare su ogni sottointervallo la parabola:

$$\int_a^b f(x)dx - Q_{X,W}^{PAR}(f) = \sum_{i=1}^{N} h^5 \cdot \frac{1}{90} f^{(IV)}(\xi_i) \quad \xi_i \in [a_i, b_i]$$

$$\frac{h^4}{90} \frac{b-a}{2N} \sum_{i=1}^{N} f^{(IV)}(\xi_i) = \frac{h^4}{180}(b-a) \frac{\sum_{i=1}^{N} f^{(IV)}(\xi_i)}{N}$$

Usando il ragionamento sopra sulle medie otteniamo:

$$= \frac{b-a}{180} f^{(IV)}(c) \cdot h^4$$