Analisi numerica:

***CAPITOLO*: Analisi dell'errore**

Definizione (Errore assoluto): $\varepsilon_a = \hat{x} - x$

Notazione:
Dato $\hat{x}$ rappresenta il valore approssimato di x.

Definizione (Errore relativo): $\varepsilon_r = \frac{\hat{x}-x}{x}$

Teorema di rappresentazione in base.

Per ogni numero reale $x \neq 0$ ∃! un intero p ed una successione $\{d_i\}_{i\geq 1}$ con le seguenti proprietà:

1. $0 \leq d_i \leq B-1$
2. $d_i \neq 0$ (Normalizzazione)
3. d_i non definitivamente $B-1$ (Unicità di scrittura)

$$x = \text{segno}(x) B^p \sum_{i=1}^{\infty} d_i \, B^{-i}$$

Notazione:
B è la base di rappresentazione, gli interi d_i sono le cifre di rappresentazione, p è l'esponente, $\sum_{i=1}^{\infty} d_i \, B^{-i}$ è la mantissa.

Definizione (Numeri di macchina o numeri floating point):

Dati $B \geq 2$ base, $t \geq 1$ il numero di cifre, $m, M > 0$ i limite dell'esponente.

$$F(t, B, m, M) = \{0\} \cup \left\{ \pm B^p \sum_{i=1}^{\infty} d_i \, B^{-i}, d_i \neq 0, 0 \leq d_i \leq B-1, -m \leq p \leq M \right\}$$

Notazione:

$\varepsilon_r = \frac{\hat{x}-x}{x}$ dato dal troncamento è detto **errore relativo di rappresentazione**.

Osservazione:
$\varepsilon_r < B^{1-t}$

Notazione:
$B^{1-t} = u$ è detto precisione di macchina.

Definizione (Aritmetica di macchina):

$$\hat{c} = a[\text{op}]b = tronc(a \text{ op } b) = c(1+\delta)$$

Notazione:

δ è detto errore locale generato dall'operazione floating point.

Errori nel calcolo di una funzione:

Errore inerente:

$\epsilon_{in} = \frac{f(\tilde{x})-f(x)}{f(x)}$ dove $\tilde{x}$ è il valore calcolato in aritmetica di macchina di x.

È un errore dato solo dalla rappresentazione del numero.

Analisi dell'errore inerente:

Sviluppando Taylor otteniamo (Su $\mathbb{R}$): $\epsilon_{in} \doteq \delta_x \frac{xf'(x)}{f(x)}$ dove $\frac{xf'(x)}{f(x)}$ è detto **coefficiente di amplificazion**e e δ_x è l'errore che è presente alla base e che verrà amplificato dalla funzione.

Osservazione:

Questo calcolo a priori ci permette di lavorare esclusivamente con numeri di macchina al primo passo.

Osservazione:

Su $\mathbb{R}^n$ vale $\epsilon_{in} \doteq \sum_{i=1}^{n} \delta_{x_i} C_i$ dove $C_x = \frac{x_i \frac{\partial f(x)}{\partial x_i}}{f(x)}$

Osservazione (Funzioni razionali):

Una funzione razionale è data da un quoziente di due polinomi, sono quelle su cui si può lavorare in quanto sono calcolabili in un numero finito di operazioni aritmetiche (Algoritmo di calcolo).

Errore locale:

È l'errore introdotto da ogni operazione aritmetica, è limitato superiormente in valore assoluto dalla precisione di macchina.

Il valore calcolato non sarà $f(\tilde{x})$ ma $\varphi(\tilde{x})$

Errore algoritmico:

Generato dall'accumularsi degli errori locali.

$\epsilon_{alg} = \frac{\varphi(\tilde{x})-f(\tilde{x})}{f(\tilde{x})}$

Analisi errore algoritmico:

Dati due numeri $\widetilde{x_1} = x_1(1+\varepsilon_1)$; $\widetilde{x_2} = x_2(1+\varepsilon_2)$

Attenzione:

Gli ε_i sono errori di varia natura, non necessariamente di approssimazione.

Quindi $s = x_1 \text{ op } x_2$ diventa $\tilde{s} = (\widetilde{x_1} \text{ op } \widetilde{x_2})(1+\delta)$.
Sviluppando ricaviamo: $\tilde{s} \doteq (x_1 \text{ op } x_2)(1 + \delta + \mathrm{C}_1\varepsilon_1 + \mathrm{C}_2\varepsilon_2)$

Quindi ci basta calcolare i coefficienti di amplificazione.

Attenzione:

δ è un errore di rappresentazione quindi ci basta maggiorarlo con la precisione di macchina.

Operazione	C_1	C_2
Moltiplicazione	1	1
Divisione	1	-1
Addizione	$\dfrac{x_1}{x_1+x_2}$	$\dfrac{x_2}{x_1+x_2}$
Sottrazione	$\dfrac{x_1}{x_1-x_2}$	$-\dfrac{x_2}{x_1-x_2}$

Errore totale:

Si può calcolare direttamente come: $\epsilon_{tot} = \frac{\varphi(\tilde{x})-f(x)}{f(x)}$

Oppure: $\epsilon_{tot} = \epsilon_{in} + \epsilon_{alg} + \epsilon_{in}\epsilon_{alg} \doteq \epsilon_{in} + \epsilon_{alg}$

Parentesi su funzioni non razionali:

Vale la definizione di errore inerente ma non di errore algoritmico, introduciamo l'errore analitico: $\epsilon_{an} = \frac{f(x)-g(x)}{g(x)}$ dove $g(x)$ è la funzione non razionale e $f(x)$ la razionale che la approssima.

Osservazione:

In questo caso vale: $\epsilon_{tot} \doteq \epsilon_{in} + \epsilon_{alg} + \epsilon_{an}(\tilde{x})$

Domanda:

Può essere posta la seguente domanda: dato un algoritmo stabile per il calcolo di una $g(x)$ e sia $\bar{\alpha} \mid fl\big(g(\tilde{\alpha})\big) = \tilde{\alpha}$, si dia una maggiorazione di $|\tilde{\alpha} - \alpha|$.

Esempio con $g(x) = ax + \frac{b}{x}$.

$$\tilde{\alpha} - \alpha = \hat{a}\tilde{\alpha} + \frac{\hat{b}}{\tilde{\alpha}} - g(\alpha) = a(1+\varepsilon_a)\tilde{\alpha} + \frac{b(1+\varepsilon_b)}{\tilde{\alpha}} - g(\alpha) = g(\tilde{\alpha}) - g(\alpha) + a(\varepsilon_a)\tilde{\alpha} + \frac{b(\varepsilon_b)}{\tilde{\alpha}}$$

$$= g'(\alpha)(\tilde{\alpha}-\alpha) + a(\varepsilon_a)\alpha + \frac{b(\varepsilon_b)}{\alpha} \rightarrow (\tilde{\alpha}-\alpha)\big(1-g'(\alpha)\big) = a(\varepsilon_a)\alpha + \frac{b(\varepsilon_b)}{\alpha}$$

$$\rightarrow (\text{Essendo } g'(\alpha) = 2a-1 \rightarrow |\tilde{\alpha}-\alpha| \leq \frac{2u\left(|a\alpha| + \left|\frac{b}{\alpha}\right|\right)}{|2a-2|}$$

Analisi all'indietro:

Idea:
Il risultato effettivamente calcolato lo posso vedere come operazione priva di errore su dei valori "perturbati", conoscendo la misura delle perturbazioni possiamo maggiorare l'errore algoritmico sfruttando i coefficienti di perturbazione.
In pratica stiamo cercando $\hat{x}_1, \dots, \hat{x}_n \mid \varphi(x_1, \dots, x_n) = f(\hat{x}_1, \dots, \hat{x}_n)$

Osservazione:
In un algoritmo complesso si scarica sui vari coefficienti così da non farli accumulare.

Maggiorare l'errore algoritmico:
Avendo scaricato sui coefficienti e lavorando a questo punto con una funzione esatta dal punto di vista formale si tratta di calcolare un errore inerente.

Quindi $\varepsilon_{alg} \doteq \delta_{x_1} C_{x_1} + \cdots + \delta_{x_n} C_{x_n}$
Con C_{x_i} ottenuto a partire da $f(x)$ sfruttando le derivate parziali.

Attenzione:
Può non essere possibile assegnare tutti gli errori, in quel caso si deve svolgere l'intero S_n senza eliminare alcuna delle variabili e POI impostare il sistema.

Osservazione:
Se sto lavorando in base due e effettuo un'operazione $2 \cdot x$ questa non mi genera errore in quanto aumento giusto di 1 l'esponente.

Domanda: determinare una limitazione all'errore di rappresentazione

Dato ad esempio e^x affetto da un errore assoluto δ allora in pratica se devo studiare l'errore di approssimazione di $g(x) = x + 1 - \frac{1}{a} e^x$ ricavo che $\tilde{g}(x) = g(x) + \frac{\delta}{a}$.

Per capire quale è l'errore di approssimazione devo in pratica capire quali sono le intersezioni di $\tilde{g}(x)$ con la bisettrice e proiettare sull'asse delle x. Questo mi dice quale è l'intervallo di incertezza sotto il quale ogni stima più accurata diventa inutile.

***CAPITOLO*: Teoremi di Gerschgorin**

Definizione (Cerchi di Gerschgorin):

$K_i = \{z \in \mathbb{C} \mid |z - a_{i,i}| \leq \sum_{j=1;j\neq i}^{n} |a_{i,j}|\}$

Osservazione:

Sono i cerchi sul Piano di Gauss di centro i valori della diagonale della matrice e di raggio la somma dei valori assoluti dei coefficienti della riga data.

PRIMO teorema di Gershgorin:

Gli autovalori di una matrice A appartengono all'insieme: $\bigcup_{i=1}^{n} K_i$

Osservazione:

Sia v un autovettore per A, allora l'autovalore $\lambda \in K_h$ dove $h \mid |v_h| = \max_i |v_i|$

Proposizione (Trasposta):

Siccome A e A^T hanno gli stessi autovalori possiamo affermare che gli autovalori della matrice A appartengono all'intersezione dei cerchi di A e di A^T.

Proposizione (Invertibile):

Se nessun cerchio di Gerschgorin contiene l'origine significa che 0 non può essere autovalore, quindi A è invertibile.

SECONDO teorema di Gershgorin:

Se l'unione dei cerchi di Gerschgorin è formata da due sottoinsiemi disgiunti M_1 e M_2 dati dall'unione di m_1, m_2 cerchi, allora M_1 conterrà m_1 autovalori mentre M_2 ne conterrà m_2.

Osservazione:

Se $\exists$ un cerchio isolato questo avrà un solo autovalore al suo interno.

Osservazione:

Se un singolo cerchio interseca l'asse dei numeri reali allora l'autovalore contenuto sarà reale non potendo esserci anche il coniugato.

Definizione (Matrice fortemente dominante diagonale):

$A = (a_{i,j}) \mid |a_{i,i}| > \sum_{j=1;j\neq i}^{n} |a_{i,j}|$

Osservazione:

Una matrice fortemente dominante diagonale è non singolare.

Definizione (Matrice riducibile):

A è riducibile se $\exists P$ matrice di permutazione | $PAP^T = \begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{pmatrix}$

Osservazione:
Una matrice irriducibile è una matrice non riducibile.

Definizione informale (Grafo diretto associato ad A):

Data una matrice A il grafo diretto $G[A]$ è formato da *n* nodi che rappresentiamo a coppie i, j come collegati se $a_{i,j} \neq 0$

Definizione (Grafo fortemente connesso):

Se $\forall$ coppia di nodi (i, j) $\exists$ un percorso che li colleghi.

Teorema:
Una matrice è irriducibile $\leftrightarrow$ il suo grafo associato è fortemente connesso.

TERZO teorema di Gerschgorin:

Supponiamo che $\exists \lambda$ | $\lambda \in K_i \rightarrow \lambda \in \delta K_i$ allora se A è irriducibile $\rightarrow \lambda$ appartiene a tutti i cerchi di Gerschgorin e quindi a tutte le loro frontiere.

Definizione (Matrici irriducibilmente dominante diagonale):

$A = (a_{i,j})$ |

1. $|a_{i,i}| \geq \sum_{j=1;j\neq i}^{n} |a_{i,j}|$
2. A è irriducibile
3. $\exists k$ | $|a_{k,k}| > \sum_{j=1;j\neq k}^{n} |a_{k,j}|$

Osservazione Similitudine:

Siccome gli autovalori sono invarianti per similitudine $\rightarrow$ posso applicare una matrice invertibile ad A $(M \in GL(V), M^{-1}AM)$ allora gli autovalori apparterranno all'INTERSEZIONE di questi cerchi di Gerschgorin con quelli calcolati in A.

Idea:
Ogni volta riduciamo quanto possibile il raggio di UN cerchio.

***CAPITOLO*: Norme di vettori e matrici**

Definizione (Norma):

È un'applicazione $\| \|: \mathbb{C}^n \to \mathbb{R}$ che rispetti le seguenti proprietà:

1- $\|x\| \geq 0 \, , \|x\| = 0 \leftrightarrow x = 0,$
2- $\|\alpha x\| = |\alpha| \|x\|$
3- $\|x + y\| \leq \|x\| + \|y\|$

Esempio (Norma di Holder):

$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ con $p \geq 1$

Osservazione:

$\{x \in \mathbb{C}^n \mid \|x\| \leq 1\}$ è un insieme convesso.

Definizione (Prodotto scalare Hermitiano):

È una funzione $\langle \rangle: \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ che rispetti le seguenti proprietà:

1- $\langle x, y \rangle = \langle y, x \rangle$
2- $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle \; \forall \alpha \in \mathbb{C}$
   - $\langle \alpha x, y \rangle = \bar{\alpha} \langle x, y \rangle \; \forall \alpha \in \mathbb{C}$
3- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
4- $\langle x, x \rangle \geq 0, \langle x, x \rangle = 0 \leftrightarrow x = 0$

Esempio: Prodotto scalare euclideo

$\langle x, y \rangle = x^T y \;\; \forall x, y \in \mathbb{R}^n$

Esempio: Prodotto scalare Hermitiano:

$\langle x, y \rangle = x^H y \;\; \forall x, y \in \mathbb{C}^n$

Osservazione (Prodotto scalare su $\mathbb{R}^n$):

È una funzione bilineare simmetrica definita positiva.

Disuguaglianza di Cauchy-Schwarz

$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$

Osservazione:

Ogni prodotto scalare induce una norma, ad esempio: $\|x\| = (\langle x, x \rangle)^{\frac{1}{2}}$

Teorema del parallelogramma:

Condizione affinché una norma sia indotta da un prodotto scalare è che:

$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$

Teorema:

Ogni norma è uniformemente continua, ossia:

$\forall \varepsilon > 0\ \exists \delta > 0 \mid \forall x, y \in \mathbb{C}^n\ |x_i - y_i| \leq \delta \rightarrow |\|x\| - \|y\|| < \varepsilon$

Teorema di equivalenza delle norme:

$\forall$ coppia di norme su $\mathbb{C}^n\ \exists \alpha, \beta \mid \forall x \in \mathbb{C}^n$ vale $\alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1$

Proprietà norme comuni:

$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$

$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$

$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$

La norma 2 è invariante per trasformazioni unitarie, sia $y = Ux$ con $U^H U = Id$ allora:

$$\|y\|_2^2 = y^H y = (Ux)^H (Ux) = x^H x = \|x\|_2^2$$

$\|\mathrm{Id}\| \geq 1$

$\forall\ M$ Hermitiana $\rho(M)$ è una norma.

Definizione (Norma di matrici):

È una funzione $\|\ \|: \mathbb{C}^{n\times n} \rightarrow \mathbb{R}$

1- $\|A\| \geq 0\ , \|A\| = 0 \leftrightarrow A = 0,$
2- $\|\alpha A\| = |\alpha|\|A\|$
3- $\|A + B\| \leq \|A\| + \|B\|$
4- $\|AB\| \leq \|A\|\|B\|$

Esempio (Norma di Frobenius):

$\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{i,j}|^2\right)^{\frac{1}{2}}$

Definizione (Norma operatore):

È una norma di matrici indotta da una norma vettoriale.

Ad esempio:

$\|A\| = \max_{\|x\|=1} \|Ax\|$ che rappresenta il massimo allungamento dato dalla trasformazione.

Osservazione:

$\|A\| \geq \rho(A)$

Osservazione:

$\|Ax\| \leq \|A\|\|x\|$

Proprietà norme operatore:

$\|Id\| = 1$

Osservazione:

Questo ci permette di dire quando una norma non è indotta, ad esempio la Norma di Frobenius.

Osservazione:

$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{i,j}|$

$\|A\|_2 = \left(\rho(A^H A)\right)^{\frac{1}{2}}$

$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{i,j}|$

Notazione (Raggio spettrale):

$\rho(A) = \max|\mu| \ \ \forall \mu$ autovalore.

Teorema:

$\forall A$ matrice, $\forall \varepsilon > 0 \ \exists \ \| \ \|$ norma indotta $| \ \rho(A) \leq \|A\| \leq \rho(A) + \varepsilon$

Inoltre:

$\|A\| = \rho(A) \leftrightarrow$ I blocchi relativi al massimo auto valore hanno dimensione uno nella F. di J. di A.

Proprietà:

$\forall$ norma di matrici $\| \ \|, \forall A$ vale $\lim_{k\to\infty} \left\|A^k\right\|^{\frac{1}{k}} = \rho(A)$

Definizione (Numero di condizionamento di una matrice A in un sistema lineare $Ax = b$):

$\mu(A) = \|A\|\|A^{-1}\|$

Idea:

Studiare quanto le perturbazioni modifichino il risultato.

Osservazione:

Potremmo calcolarlo con i coefficienti di amplificazione ma risulterebbe troppo pesante quindi lo stimiamo sfruttando le norme.

Osservazione:

È funzione della norma scelta. Può essere utile sfruttare la matrice norma 2.

Osservazione (Maggiorare il numero di condizionamento):

$\|A^{-1}\| = \|KB\| \leq \|K\| \cdot \|B\|$

Osservazione:

Per S simmetrica $\mu_2 = \|S\|_2 \|S^{-1}\|_2 = \frac{\lambda_{max}}{\lambda_{min}}$

***CAPITOLO*: Forma normale di Schur:**

Idea:
La forma di Jordan sebbene intuitiva risulta pesante da calcolare e numericamente instabile per perturbazioni.

Teorema (Forma normale di Schur):
$\forall A \in \mathbb{C}^{n\times n}\ \exists T$ triangolare superiore e $U \in \mathbb{C}^{n\times n}$ unitaria | $U^H A U = T$

Osservazione:
Non è unica.

Definizione (Matrice quasi triangolare):
Una matrice $T \in \mathbb{R}^{n\times n}$ si dice quasi triangolare se si può scrivere nella forma: $\begin{pmatrix} T_1 & - & - \\ & \dots & - \\ 0 & & T_m \end{pmatrix}$ dove T_i sono matrici 2×2 con auto valori complessi coniugati oppure sono numeri reali.

Osservazione:
Gli autovalori di T sono quelli di T_i

Teorema (Forma reale di Schur):
$\forall A \in \mathbb{R}^{n\times n}\ \exists T \in \mathbb{R}^{n\times n}$ quasi triangolare e $Q \in \mathbb{R}^{n\times n}$ ortogonale | $Q^T A Q = T$

Osservazione:
$A = A^H \rightarrow T = T^H \rightarrow T$ è diagonale e reale
$A = -A^H \rightarrow T$ è diagonale e immaginaria

Teorema:
$A \in \mathbb{C}^{n\times n}$ ha forma Normale di Schur diagonale $\leftrightarrow AA^H = A^H A$ (A normale)

Osservazione:
Unitaria → Normale

***CAPITOLO*: Fattorizzazione LU e QR:**

Idea:

Risolvere i sistemi lineare $Ax = b$

Caso facile da calcolare 1:

a) Avendo T triangolare inferiore si effettua una sostituzione in avanti:

$$\begin{cases} x_1 = \frac{b_1}{a_{1,1}} \\ x_i = \left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j\right)/a_{i,i} \end{cases}$$

b) Avendo T triangolare superiore si effettua una sostituzione in avanti:

$$\begin{cases} x_1 = \frac{b_n}{a_{n,n}} \\ x_i = \left(b_{n-i} - \sum_{j=i+1}^{n} a_{n-i,j} x_j\right)/a_{n-i,n-i} \end{cases}$$

Caso facile da calcolare 2:

Se Q è una matrice unitaria il sistema $Qx = b$ si risolve semplicemente come: $x = Q^H b$

Osservazione:

Se $A = BC$ allora $Ax = b \rightarrow BCx = b \rightarrow \begin{cases} By = b \\ Cx = y \end{cases}$

Tipi di fattorizzazione:

1- LU, fattorizzazione in una matrice triangolare superiore ed una triangolare inferiore
2- PLU, "" e P matrice di permutazione
3- $P_1 L U P_2$ "" e P_1, P_2 matrici di permutazione
4- QR una matrice unitaria e una triangolare superiore

Teorema:

La fattorizzazione LU ∃! se tutti i minori principali sono invertibili.

Osservazione:

Se A è invertibile allora il teorema precedente è un se solo se.

Osservazione:

PLU ∃ sempre.

Teorema:

La fattorizzazione QR esiste sempre ma non è unica.

A invertibile → QR è unica a meno di trasformazioni con matrici unitarie diagonali.

***CAPITOLO*: Matrici elementari**

Definizione (Matrice elementare):

È una matrice della forma $M = Id - \sigma uv^H$ con σ numero complesso e $u, v \in \mathbb{C}^n$

Osservazione 1:

Se $x \in \mathbb{C}^n, Mx = x - \sigma u(v^H x)$

Quindi se x è ortogonale a v allora x viene mandato in se stesso.

Osservazione 2:

Se $x = u, Mu = (1 - \sigma u^H v)u$ quindi $(1 - \sigma u^H v)$ è autovalore.

Idea:

L'inversa sarà dello stesso tipo.

Teorema:

$M = Id - \sigma uv^H$ è non singolare $\leftrightarrow \sigma uv^H \neq 1$ e $M^{-1} = Id - \tau uv^H$ con $\tau = \frac{-\sigma}{1-\sigma v^H u}$

Teorema (Costo operazioni):

Sia M una matrice elementare, la risoluzione del sistema $Mx = y$ richiede $3n + 1$ moltiplicazioni, $3n$ addizioni e una divisione.

Teorema (Matrici elementari che scambiano):

$\forall x, y \in \mathbb{C}^n \backslash \{0\}\ \exists M$ elementare | $Mx = y$

Definizione (Matrici elementari di Gauss):

Sono matrici elementari caratterizzate da: $v = e^1, \sigma = 1, u | u_1 = 0$ ossia:

$$M = Id - u(e^1)^H = \begin{pmatrix} 1 & & 0 \\ -u_2 & 1 & \\ \vdots & & \ddots \\ -u_n & 0 & 1 \end{pmatrix}$$

Osservazione (Inversa):

$M^{-1} = Id + u(e^1)^H$ ossia: $M^{-1} = \begin{pmatrix} 1 & & 0 \\ u_2 & 1 & \\ \vdots & & \ddots \\ u_n & 0 & 1 \end{pmatrix}$

Osservazione (<u>Numero di condizionamento</u>):

M e M^{-1} hanno $\| \|_\infty = 1 + \max_i \frac{|x_i|}{|x_1|}$ quindi il Numero di condizionamento:

$\mu(M) = \left(1 + \max_i \frac{|x_i|}{|x_1|}\right)^2$

<u>Metodo di Gauss (Eliminazione Gaussiana):</u>

Prevede di scrivere una matrice $M_k \dots M_1 A = U$ dove U è una matrice triangolare superiore e le M_i sono le matrici elementari di Gauss che ogni volta "azzerano l'i-esima colonna al di sotto della diagonale.

Siccome sono matrici facili da invertire e triangolari inferiori il metodo di Gauss mi restituisce una fattorizzazione LU

Osservazione:

È esattamente l'algoritmo di Gauss visto a GAAL.

Osservazione:

Per individuare M di Gauss | $Mx = x_1 e^1$ con $x = (x_i), x_1 \neq 0$ basta porre $u_i = \frac{x_i}{x_1}$ ossia:

$$M = \begin{pmatrix} 1 & & & 0 \\ -\frac{x_2}{x_1} & 1 & & \\ \vdots & 0 & 1 & \\ -\frac{x_n}{x_1} & 0 & 0 & 1 \end{pmatrix}$$

Attenzione:

La fattorizzazione non è sempre garantita nel caso in cui l'elemento della diagonale della matrice trasformata sia 0.

Osservazione:

Il costo computazionale del metodo di Gauss è dell'ordine di $\frac{2}{3}n^3$

Osservazione:

Il metodo di Gauss NON è stabile perché con numeri piccoli la frazione può assumere valori elevati.

<u>Strategia del massimo pivot parziale:</u>

Riduce l'errore senza sostanzialmente modificare il costo computazionale.

Consiste in una permutazione di righe prima di applicare la matrice di triangolazione a A_k.

1- Si sceglie un indice h per cui $|a_{h,k}| \geq |a_{i,k}| \; \forall i > k$
2- Si permutano la riga h con la riga k.
3- Si applica la matrice di triangolazione: $A_{k+1} = M_k P_{k,h} A_k$

Osservazione:
Nel caso si incontri un pivot nullo ed esiste un elemento da non nullo da scambiare non interrompe l'algoritmo, nel caso non lo incontri semplicemente procede al passo successivo.

Strategia del massimo pivot totale:
Riduce ulteriormente l'errore ma aumenta il costo computazionale.
Si sceglie il massimo valore possibile $|a_{i,j}|$ e si porta sulla diagonale al posto k.

Matrici speciali (Matrici a banda di ampiezza $2q+1$):
È una matrice $A = (a_{i,j})$ se $a_{i,j} = 0$ per $|i-j| > q$.

Esempio:
Una matrice tridiagonale.

Osservazione:
La fattorizzazione LU di una matrice a banda è data da due matrici a banda. Inoltre tutte le matrici A_k sono matrici a banda di ampiezza $2k+1$.

Definizione (matrici elementari di Householder):
Sono matrici elementari hermitiane e unitarie (Ossia $M^H = M$ e $MM^H = Id$) della forma:
$M = Id - \beta uu^H$ con $\beta = 0$ o $\beta = \frac{2}{(u^H u)}$ e $u \neq 0$

Osservazione (Inversa):
$M^{-1} = M$

Metodo di Householder:
Prevede di scrivere una matrice $M_k \dots M_1 A = R$ dove R è una matrice triangolare superiore e le M_i sono le matrici elementari di Householder che ogni volta "azzerano l'i-esima colonna al di sotto della diagonale.
Siccome sono matrici facili da invertire e unitarie il metodo di Householder mi restituisce una fattorizzazione QR

Osservazione:
Per individuare M di Householder | $Mx = \alpha e^1$ con $x = (x_i)$ basta porre $u = x - \alpha e^1$ e $\beta = \frac{2}{u^H u}$ e potremmo calcolare α e ricavare il resto.

Oppure osservare che va bene la scelta:
$u_1 = x_1 \left(1 \pm \frac{\|x\|_2}{|x_1|}\right)$ se $x_1 \neq 0$ oppure:
$u_1 = \|x\|_2$ se $x_1 = 0$
Gli altri $u_i = x_i$ e scegliamo il segno per ridurre i problemi legati alla cancellazione.

Osservazione:
Questa metodo sebbene più complicato come passaggi è sempre funzionante.

Osservazione:
Il costo computazionale del metodo di Householder è dell'ordine di $\frac{2}{3}n^3$

Osservazione:
Il metodo è stabile anche senza sfruttare strategie di massimo pivot.

Definizione (Complemento di Schur):

Data una matrice A e il partizionamento: $A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$ con $A_{1,1}$ di dimensione k ed invertibile.
Il complemento di Schur di $A_{2,2}$ in A è $S = A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{2,1}$

Osservazione:
Vale la fattorizzazione LU di $A = \begin{pmatrix} \mathrm{Id} & 0 \\ A_{2,1}A_{1,1}^{-1} & \mathrm{Id} \end{pmatrix}\begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & S \end{pmatrix}$
Inoltre S coincide con la sottomatrice ottenuta iterando il metodo fino al passo k-esimo.

Proprietà:
$\det A = \det A_{1,1} \det S$
Quindi A invertibile → S invertibile.
S^{-1} coincide con la sottomatrice principale di A^{-1} formata dagli indici $i, j > k$

Proprietà utili matrici:

Se una matrice è della forma $A = \mathrm{Id}_n - B$ e dato $\mathrm{Spettro}(B) = \{\lambda_1, \dots, \lambda_n\}$ allora:
$\mathrm{Spettro}(A) = \{1 - \lambda_1, \dots, 1 - \lambda_n\}$.

Osservazione:
Una matrice della forma $B = uv^T$ ha come autovalore $0 \mid \mu_a(0) = (n-1)$ e
$v^T u \mid \mu_a(v^T u) = 1$
Di conseguenza $A = \mathrm{Id}_n - B$ ha $\mathrm{Spettro}(A) = \{1, \dots, 1, 1 - v^T u\}$

***CAPITOLO*: Metodi iterativi per sistemi lineari:**

Idea:

Nella risoluzione di sistemi lineari eccessivamente pesanti è possibile individuare una successioni di vettori che sotto date condizioni convergono alla soluzione del sistema. Il problema si riduce dunque a dimostrare la convergenza della famiglia di vettori associati al particolare sistema lineare.

Metodi stazionari:

Dato il sistema $Ax = b$ scriviamo la matrice $A = M - N$, $\det M \neq 0$ il sistema può essere scritto come: $Mx = Nx + b \rightarrow x = M^{-1}Nx + M^{-1}b \coloneqq Px + q$.

A questo punto fissato un valore $x^{(0)} \in \mathbb{C}^n$ possiamo generare una successione di vettori secondo la regola: $x^{(k+1)} = Px^{(k)} + q$

(Metodo iterativo stazionario, in quanto P (Matrice di iterazione) e q non dipendono da k)

Proposizione:

Se la successione $x^{(i)}$ ammette limite questo è la soluzione del sistema lineare $Ax = b$

Definizione (Vettore dell'errore di approssimazione al passo *k*-esimo):

$e^{(k)} = x^{(k)} - x$

Si può calcolare vedendo quanto il sistema si discosti da 0.

Per proprietà delle norme, fissata una norma qualsiasi e la sua norma indotta vale:
$\left\|e^{(k)}\right\| \leq \|P\|^k \left\|e^{(0)}\right\|$

Osservazione:

Se il $\rho(P) = 0$ allora il metodo fornisce dopo un numero di passi finiti la soluzione priva di errore.

Il numero di passi da eseguire è il minimo $k \mid P^k = 0$

Teorema:

Se esiste una norma indotta $\mid \|P\| < 1$ allora $\forall x^{(0)}$ vettore iniziale, la successione converge al valore di x.

Teorema:

Il metodo iterativo è convergente $\leftrightarrow \rho(P) < 1$

Osservazione pratica utile:

Se ho una matrice a blocchi e devo cercarne gli autovalori rispetteranno:
$\begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix}\begin{pmatrix} x_i \\ y_j \end{pmatrix} = \begin{pmatrix} \lambda x_i \\ \lambda y_j \end{pmatrix} \rightarrow \begin{cases} Bx_i = \lambda y_j \\ Ay_j = \lambda x_i \end{cases}$ risolvendolo possiamo individuare gli autovalori in funzione di quelli di B e di A.

Esempio:

$\begin{pmatrix} 0 & -B \\ -uv^T & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{cases} -By = \lambda x \\ -uv^T x = \lambda y \end{cases} \rightarrow u(v^T By) = \lambda^2 u \rightarrow$ se scelgo $y = u$ allora:
$\lambda = \pm\sqrt{v^T Bu} \rightarrow \rho(P) = \sqrt{v^T Bu}$

Definizione (Riduzione asintotica media):

Data la media geometrica $\theta_k\left(e^{(0)}\right) = \left(\frac{\|e^{(1)}\|}{\|e^{(0)}\|} \cdot \frac{\|e^{(2)}\|}{\|e^{(1)}\|} \dots \frac{\|e^{(k)}\|}{\|e^{(k-1)}\|}\right)^{\frac{1}{k}} = \left(\frac{\|P^k e^{(0)}\|}{\|e^{(0)}\|}\right)^{\frac{1}{k}}$ è: $\lim_{k\to\infty} \theta_k\left(e^{(0)}\right)$

Osservazione (Necessaria a confrontare due metodi iterativi):

$\theta\left(e^{(0)}\right) \leq \rho(P)$

Osservazione:

Se $e^{(0)}$ è un autovettore relativo all'autovalore di modulo massimo $\rightarrow \theta\left(e^{(0)}\right) = \rho(P)$

Metodo di Jacobi:

Si decompone $A = D - B - C$ con D diagonale, B triangolare inferiore e C triangolare superiore.

$$M = D\ ; N = B + C$$

Matrice di iterazione: $J = D^{-1}(B + C)$

Scritta in componenti l'iterazione diventa:
$x_i^{(k+1)} = \frac{1}{A_{i,i}}\left(b_i - \sum_{j=1;j\neq i}^{n} A_{i,j} x_j^{(k)}\right)$

Idea:

Per evitare il caso in cui $D \notin GL(V)$ possiamo applicare delle trasformazioni (Permutazioni) che eliminino gli 0 dalla diagonale.

Attenzione:

Se applichiamo P di permutazione non scordarsi il vettore delle soluzioni del sistema: $PAx = Pb$

Osservazione:

La non convergenza può essere messa in evidenza da $\det \mathrm{It} > 1$

Metodo di Gauss-Seidel:

Si decompone $A = D - B - C$ con D diagonale, B triangolare inferiore e C triangolare superiore.

$$M = D - B \; ; N = C$$

Matrice di iterazione: $G = (D - B)^{-1}C$

Scritta in componenti l'iterazione diventa:

$x_i^{(k+1)} = \frac{1}{A_{i,i}}\left(b_i - \sum_{j=1}^{i-1} A_{i,j}x_j^{(k)} - \sum_{j=i+1}^{n} A_{i,j}x_j^{(k+1)}\right)$

Teorema comune:

Se vale una delle seguenti condizioni allora $\rho(J) < 1 \; ; \; \rho(G) < 1$:

1. A è fortemente dominante diagonale.
2. A^T è fortemente dominante diagonale.
3. A è irriducibilmente dominante diagonale
4. A^T è irriducibilmente dominante diagonale

Teorema di Stein-Rosenberg :

Se la matrice A ha elementi diagonali non nulli e J ha elementi non negativi allora vale una sola delle seguenti proprietà:

1. $\rho(J) = \rho(G) = 0$
2. $0 < \rho(G) < \rho(J) < 1$
3. $\rho(J) = \rho(G) = 1$
4. $1 < \rho(J) < \rho(G)$

Teorema:

Se A è una matrice tridiagonale con elementi diagonali non nulli allora per ogni autovalore γ di J $\exists \mu$ autovalore per G $|\mu = \gamma^2$.

Per ogni autovalore non nullo μ di G $\exists \gamma$ autovalore di J | $\mu = \gamma^2$. In particolare vale: $\rho(G) = \rho(J)^2$

Metodo di Richardson:

Dato $Ax = b$ si scrive il metodo iterativo $x_{k+1} = x_k = \alpha(b - Ax_k)$ con α scelto da noi.

Se abbiamo già dimostrato la convergenza cerchiamo α che massimizzi la velocità di convergenza.

Un caso semplice è nel caso in cui gli autovalori siano $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

Il raggio spettrale della matrice di iterazione $P = (\mathrm{Id} - \alpha A)$ è $\rho(P) = \max|1 - \alpha\lambda_i|$

Impongo che tutti siano minori di uno e cerco il massimo dato dalla più bassa intersezione fra le rette $y = |1 - \alpha\lambda_i|$ in funzione di α. Questo α ottimizza il metodo iterativo.

***CAPITOLO*: Zeri di funzioni**

Idea:

Ricavare gli 0 di funzioni o di sistemi non lineari, il calcolo esplicito infatti è raramente attuabile soprattutto per la pesantezza del calcolo e per la mancanza di formule risolutive esplicite (Es. i polinomi di grado maggiore al 5).

CASO: $f: [a,b] \to \mathbb{R} \mid f(a)f(b) \leq 0$
Sappiamo che esiste almeno uno 0 nell'intervallo per il T. degli 0, vogliamo generare una successione che tenda allo 0.

METODO DI BISEZIONE O DICOTOMICO:

$a_0 = a$, $b_0 = b$; si calcola $e_k = \frac{a_k + b_k}{2}$ se $f(e_k) = 0$ abbiamo trovato lo 0, altrimenti si sostituisce all'elemento dell'intervallo $[a_k, b_k]$ di segno concorde, ad esempio $[a_{k+1}, b_{k+1}] = [e_k, b_k]$.

La maggiorazione dell'errore converge a 0 in modo esponenziale.
Infatti $b_k - a_k = \frac{1}{2^k}(b-a)$

Attenzione:

Siccome in aritmetica floating point i valori calcolati di $f(x)$ sono affetti da errore, non cerchiamo una soluzione α ma un **intervallo di incertezza**.

Osservazione:
Se f è derivabile è possibile dare una stima dell'intervallo di incertezza: l'approssimazione di primo ordine sarà: $\left[\alpha - \frac{\delta}{f'(\alpha)}, \alpha + \frac{\delta}{f'(\alpha)}\right]$ dove δ è l'errore massimo che la macchina può commettere su $f(x)$. (Sviluppato secondo i Polinomi di Taylor traslati di δ)

METODI DEL PUNTO FISSO (O di iterazione funzionale):

Idea:

Invece di provare a ricavare direttamente lo 0 si costruisce una $g(x) \mid g(\alpha) = \alpha \leftrightarrow f(\alpha) = 0$ di cui cerchiamo i punti fissi.

Esempio di costruzione:
$g(x) = x - \frac{f(x)}{h(x)}$ con $h(x) \neq 0$
Dalla g possiamo generare una successione di punti che (se) convergono al punto fisso secondo la regola: $\begin{cases} x_{k+1} = g(x_k) \\ x_0 \in \mathbb{R} \end{cases}$

Osservazione:

Nei metodi iterativi, stando noi lavorando con sistemi lineari vale che qualsiasi scelta del punto iniziale mi assicura la convergenza del metodo, per metodi del punto fisso si distingue invece fra convergenza globale e convergenza locale (Intorno del punto fisso).

Idea:

Assegnato un valore iniziale calcolare la sua immagine e proiettarla sulla bisettrice.

Teorema del Punto Fisso (1) (Condizione convergenza):

Sia $I = [\alpha - p, \alpha + p]$ e $g(x) \in C^1$ dove $g(\alpha) = \alpha$; $p > 0$.
Sia $\lambda = \max_{|x-\alpha|<p} |g'(x)|$.

Se $\lambda < 1 \rightarrow \forall x_0 \in I$ dato $x_{k+1} = g(x_k)$ vale:
$|x_k - \alpha| < \lambda^k p$.

Inoltre $\lim x_k = \alpha$ e α è l'unico punto fisso.

Corollario:

$x_{k+1} - \alpha = g'(\varepsilon_k)(x_k - \alpha)$
Con $\varepsilon_k \in (\alpha, x_k)$ quindi $|g'(\varepsilon_k) \rightarrow |g'(\alpha)$
In altre parole la convergenza è sempre più rapida.

Successione monotona:

Se $0 < g'(x) < 1$ nell'intervallo I allora:
- Se $x_0 > \alpha$ allora $\alpha < x_{k+1} < x_k$
- Se $x_0 < \alpha$ allora $\alpha > x_{k+1} > x_k$

Se $-1 < g'(x) < 0$ nell'intervallo I allora:
- Se $x_k > \alpha$ allora $x_{k+1} < \alpha$ e se $x_k < \alpha$ allora $x_{k+1} > \alpha$
- Le sottosuccessioni $\{x_{2k}\}$; $\{x_{2k+1}\}$ convergono ad α una crescendo e l'altra decrescendo.

Osservazione generale:

Sia $I = [a, b]$ un intervallo contenente α nel quale $1 > g'(x) > 0$, allora $\forall x_0 \in I, \{x_k\}$ converge in modo monotono ad α.

Limitazione a posteriori dell'errore:

Consideriamo $|x_k - x_{k+1}|$ che ci permette di decidere a priori una condizione di arresto all'iterazione del metodo.

Possiamo calcolarlo a partire dall'errore di approssimazione: $x_k - x_{k+1} = \left(1 - g'(\varepsilon_k)\right)(x_k - \alpha)$

Quindi una limitazione all'errore sarà:

$|x_k - \alpha| \leq \frac{1}{|1-g'(\varepsilon_k)|}\epsilon$ dove ϵ è la condizione di arresto $|x_k - x_{k+1}| < \epsilon$

Teorema (Floating Point):

Hp precedenti, sia $\tilde{x}_k$ la successione generata | $\tilde{x}_{k+1} = g(\tilde{x}_k) + \delta_k$ con $|\delta_k| \leq \delta$ errore commesso nel calcolo della funzione.

Posto $\sigma = \frac{\delta}{1-\lambda}$ se $\sigma < p$ vale: $|\tilde{x}_k - \alpha|(p-\alpha)\lambda^k + \sigma$

Osservazione:

È l'equivalente del teorema del Punto Fisso in aritmetica Floating Point.

Definizione (Velocità di convergenza):

Sia $\{x_k\}$ | $\lim x_k = \alpha$. Se $\exists\ \gamma = \lim \left|\frac{x_{k+1}-\alpha}{x_k-\alpha}\right|$

La convergenza di $\{x_k\}$ è detta:

Se $0 < \gamma < 1$ **Lineare** o **Geometrica**

Se $\gamma = 1$ **Sublineare**

Se $\gamma = 0$ **Superlineare**

Se $\gamma = 0$ si dice che la convergenza ha **ordine $\boldsymbol{p}$** se $\exists \lim \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^p} = \sigma$ con $0 < \sigma < \infty$

Idea:

Per le successioni generate da un metodo di punto fisso possiamo sfruttare le derivate.
Altrimenti può essere utile sviluppare la serie di Taylor associata e valutarla in α.

Teorema (Convergenza Lineare):

Sia $g(x) \in C^1([a,b])$ e $\alpha \in [a,b]$ punto fisso.

Se $\exists\ x_0 \in [a,b]$ | $\{x_k\}$ converga linearmente ad α con fattore γ, allora $\left|g'(\alpha)\right| = \gamma$

Viceversa se $0 < \left|g'(\alpha)\right| < 1$ allora:
$\exists I \subseteq [a,b]$ | $\forall\ x_0 \in I$ $\{x_k\}$ converge in modo lineare con fattore $\gamma = \left|g'(\alpha)\right|$

Teorema (Convergenza Sublineare):

Sia $g(x) \in C^1([a,b])$ e $\alpha \in [a,b]$ punto fisso.

Se $\exists\ x_0 \in [a,b]$ | $\{x_k\}$ converga Sublinearmente ad α allora $\left|g'(\alpha)\right| = 1$

Viceversa se $\left|g'(\alpha)\right| = 1$ allora:
$\exists I \subseteq [a,b]$ | $\forall\ x_0 \in I$ $\{x_k\}$ converge in modo Sublineare.

Teorema (Convergenza Superlineare):

Sia $g(x) \in C^p([a,b])$ con $p > 1$ e $\alpha \in [a,b]$ punto fisso.

Se $\exists\ x_0 \in [a,b]$ | $\{x_k\}$ converga Superlinearmente ad α allora:
$|g'(\alpha)| = \cdots = |g^{(p-1)}(\alpha)| = 0$ e $|g^{(p)}(\alpha)| \neq 0$

Viceversa se $|g'(\alpha)| = \cdots = |g^{(p-1)}(\alpha)| = 0$ e $|g^{(p)}(\alpha)| \neq 0$ allora:
$\exists I \subseteq [a,b]$ | $\forall\ x_0 \in I$ $\{x_k\}$ converge in modo Superlineare con ordine di convergenza p.

Definizione (Ordine almeno):

Se non possiamo sfruttare le derivate vale comunque il risultato:
la successione $\{x_k\}$ converge ad α con ordine almeno p se $\exists\ \beta$ costante | $|x_{k+1} - \alpha| \leq \beta |x_k - \alpha|^p$

Confronto asintotico tra due metodi:

Dati due metodi che convergano nella stessa maniera (Convergenza lineare o superlineare) denotando con c_i le operazioni aritmetiche per passo, γ_i il fattore di convergenza allora le riduzioni dell'errore al k_i-esimo passo saranno uguali a:

Convergenza lineare:

$\beta_1 \gamma_1^{k_1} = \beta_2 \gamma_2^{k_2}$ usando i logaritmi otteniamo $k_1 \log \gamma_1 = k_2 \log \gamma_2 + \log \frac{\beta_2}{\beta_1}$ dove $\log \frac{\beta_2}{\beta_1}$ può essere trascurato.
Già con questo possiamo individuare quale metodo si avvicina prima alla soluzione.
Siccome il costo totale è dato da: $c_i k_i$ allora $c_1 k_1 < c_2 k_2$ se solo se $\frac{c_1}{c_2} = \frac{\log \gamma_1}{\log \gamma_2}$

Convergenza superlineare di ordine *p*:

Per la convergenza superlineare vale la stima $\varepsilon_k \leq \beta \varepsilon_{k-1}^p$ quindi $\varepsilon_k \leq \beta \beta^p \beta^{p^2} \dots \beta^{p^{k-1}} \varepsilon_0^{p^k} = n r^{p^k}$
con: $n = \beta^{-\frac{1}{p-1}}$ e $r = \varepsilon_0 \beta^{\frac{1}{p-1}}$ quindi con ε_0 piccolo vale $r < 1$.
Quindi stavolta: $n_1 r_1^{p_1^{k_1}} = n_2 r_2^{p_2^{k_2}}$ quindi $k_1 = k_2 \frac{\log p_2}{\log p_1}$
Il 1° è migliore se:
$\frac{c_1}{c_2} \leq \frac{\log p_1}{\log p_2}$

Maniera semplice pratica:

Individuo il numero di operazioni (Costo computazionale) dei due metodi e divido per il loro ordine di convergenza $\left(\frac{C.\ comp}{\log p}\right)$, il minore è il migliore.

METODO delle Tangenti di Newton:

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

Teorema 1 (M. di Newton):

Sia $f(x) \in C^2([a,b])$ e $\alpha \in (a,b) \mid f(\alpha) = 0$.

Se $f'(\alpha) \neq 0$ allora $\exists\, I \subseteq [a,b]$ intorno di α | ogni punto dell'intorno converge con il Metodo di Newton ad α.

Inoltre se $f''(\alpha) \neq 0$ il metodo ha velocità di convergenza superlineare di ordine 2, se $f''(\alpha) = 0$ almeno 2.

Attenzione:
Se vale $f'(x) \neq 0\,; f''(x) = 0\,; f'''(x) \neq 0$ allora l'ordine di convergenza è 3.

Teorema 2 (M. di Newton):

Sia $f(x) \in C^p([a,b])$ e $\alpha \in (a,b) \mid f(\alpha) = 0$.

Se $f'(\alpha) = \cdots = f^{(p-1)}(\alpha) = 0\,, f^{(p)}(\alpha) \neq 0$ allora $\exists\, I \subseteq [a,b]$ intorno di α | ogni punto dell'intorno converge con il Metodo di Newton ad α.

La convergenza è **lineare** con fattore di convergenza $1 - \frac{1}{p}$

Teorema 3 (M. di Newton):

Sia $f(x) \in C^2$ sull'intervallo $I = [\alpha, \alpha + \rho] \mid f'(x)f''(x) > 0$ allora $\forall\, x_0 \in I$ la successione converge decrescendo ad α.

Sia $f(x) \in C^2$ sull'intervallo $I = [\alpha - \rho, \alpha] \mid f'(x)f''(x) < 0$ allora $\forall\, x_0 \in I$ la successione converge crescendo ad α.

Idea:

Nel caso in cui i tre teoremi non siano sufficienti a stabilire la convergenza locale o globale del sistema si può costruire in maniera esplicita la funzione di iterazione del metodo di Newton e studiarla direttamente.

Esempio:

Se ho dei punti in cui non mi è possibile ricavare la derivata (Neppure come limite del rapporto incrementale) o essa è infinita, se passo alla $g(x) = x - \frac{f(x)}{f'(x)}$ può darsi che io riesca a studiare il punto.

***CAPITOLO*: Interpolazione**

Idea:

Data una $f: [a, b] \to \mathbb{R}$ e un numero di punti x_i di cui si conosce il valore di $f(x_i)$ ricavare un valore approssimato di $f(\vartheta)$ per un $\vartheta \in [a, b]$.

Esistono diversi tipi di interpolazione dipendenti da quale categorie di funzioni usiamo per approssimare f.

Interpolazione lineare:

Siano $\varphi_0(x) \dots \varphi_n(x)$ da $[a, b] \to \mathbb{R}$ e (x_i, y_i) delle coppie di punti (Con $x_i \neq x_j$), vogliamo determinare i valori dei coefficienti a_i | $g(x) = \sum a_i \varphi_i(x)$ rispetti $g(x_i) = y_i$

Notazione:

$g(x_i) = y_i$ si dice Condizioni di interpolazione

I punti x_i si dicono Nodi di interpolazione

Interpolazione polinomiale:

Scegliamo come funzioni con cui approssimare $\varphi_i(x) = x^i$

La condizione diviene: $\sum a_i {x_s}^i = y_s$

Rappresentata in forma matriciale (Matrice di Vandermonde) diventa:

$$\begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Teorema 1:

Il determinante di una matrice di Vandermonde è:

$\det V_n = \prod_{0 \leq j < i \leq n} (x_i - x_j)$

Osservazione:

Se i nodi sono distinti la matrice è invertibile, ossia esiste un polinomio che interpoli quei punti.

Interpolazione polinomiale nella base di Lagrange:

Consideriamo come base di polinomi:

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

Questa costruzione ha delle utili proprietà:

$L_i(x_i) = 1$ e $L_i(x_s) = 0$

La matrice associata è la matrice identica e il polinomio di interpolazione è:

$$p(x) = \prod_{i=0}^{n} y_i L_i(x)$$

Resto dell'interpolazione:

Se è possibile calcolarlo è: $r_n(x) = f(x) - p_n(x)$ dove $p_n(x)$ è il polinomio di interpolazione di $f(x)$ considerato nei punti $x_0 < x_1 < \cdots < x_n$

Teorema (Valutare il resto dell'interpolazione):

Sia $f(x) \in C^{n+1}[a,b]$ e $p_n(x)$ il suo polinomio di interpolazione nei nodi $x_0 < x_1 < \cdots < x_n$ allora:

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

Osservazione:

Se la derivata $(n+1)$-esima non cambia di segno allora il resto cambia di segno ogni volta che la variabile x supera un nodo. Cambiando base è possibile evitare questo problema.

Osservazione (Convergenza):

Se prendiamo una successione di punti x_i ed interpoliamo con un polinomio non è detto che questi convergano uniformemente o puntualmente alla funzione che approssimano.

Controesempio utile (Funzione di Runge)

$f(x) = \frac{1}{1+x^2}$ sull'intervallo $[-5,5]$ la successione di punti equispaziati non porta alla convergenza dei polinomi di interpolazione ad $f(x)$.

Teorema (Condizione convergenza):

Se la funzione $f(x)$ è continua su $[a,b]$ allora ∃ una scelta di nodi per cui $p_n(x)$ converge uniformemente ad $f(x)$.

***CAPITOLO*: Trasformata discreta di Fourier**

Idea:

Dato un vettore complesso rappresentarlo in una base particolare (Base di Fourier) dalle particolari proprietà.

Osservazione:

Questo cambio di base:

1. Ha un costo computazionale basso $\left(\frac{3}{2} n \log_2 n \text{ operazioni purché } n = 2^k\right)$
2. Ben condizionata
3. Esistono algoritmi numericamente stabili per il calcolo $(\mathrm{FFT_n})$

Interpolazione alle radici n-esime dell'unità:

Sappiamo che le radici del polinomio $x^n - 1$ sono della forma $\xi_n^j = \cos\frac{2\pi j}{n} + i \sin\frac{2\pi j}{n}$

Definizione (Radice primitiva):

Una radice viene detta primitiva se è una generatrice del gruppo moltiplicativo delle radici n-esime.

Se interpoliamo sul Piano di Gauss scegliendo come nodi di interpolazione le radici n-esime (Detti nodi di Fourier), ossia: $x_j = \xi_n^j$

Se abbiamo le n coppie (x_j, y_j) per costruire la funzione passante per quei punti la condizione diviene la matrice di Vandermonde (Detta **matrice di Fourier**):

$$\Omega_n = \left(\xi_n^{ij \bmod n}\right)$$

Con le seguenti proprietà:

$$\sum_{i=0}^{n-1} \xi_n^{ki} = \begin{cases} n \text{ se } k \equiv 0\ (n) \\ 0 \text{ se } k \neq 0\ (n) \end{cases}$$

$$\Omega_n = \Omega_n^T$$

$$\Omega_n^H \Omega_n = n \cdot \mathrm{Id}$$

Quindi:

$$\Omega_n^{-1} = \frac{1}{n} \Omega_n^H$$

Osservazione:

Sfruttando l'inversa vale la seguente proprietà:

$z = \frac{1}{n} \Omega_n^H y$ con y punto da interpolare e z vettore dei coefficienti.

Proprietà:

La matrice così costruita è ben condizionata.

Algoritmi:

$\mathrm{DFT}_n(y) \coloneqq z = \frac{1}{n}\Omega_n^H y$ **Trasformata discreta di Fourier** (Individuazione dei coefficienti di un polinomio)

$\mathrm{IDFT}_n(z) \coloneqq y = \Omega_n z$ **Trasformata discreta inversa di Fourier** (Valutazione del polinomio nelle radici)

Osservazione:
Le componenti di z sono i coefficienti di rappresentazione del vettore y nella base data dalle colonne di Ω_n (Detta Base di Fourier).

Quindi DFT_n può essere visto come un cambiamento di base da quella di Fourier alla canonica e IDFT_n come la sua inversa.

Algoritmi FFT:

Sono gli algoritmi utilizzati per ridurre il costo computazionale per il calcolo di DFT_n e IDFT_n. Sfruttando giusto la struttura della matrice abbiamo un costo computazionale di n^2 moltiplicazioni e $n^2 - n$

Idea:

Ogni volta dividere in due algoritmi che lavorino un numero dimezzato di nodi, particolarmente efficiente se $n = 2^q$.

Consideriamo IDFT_n, vogliamo ricavare:

$y_i = \sum_{j=0}^{n-1} \xi_n^{ij} z_j$ e sia $n = 2^q$

Se separiamo la sommatoria in termine di indice pari e termini di indice dispari otteniamo:

$y_i = \sum_{j=0}^{\frac{n}{2}-1} \xi_n^{2ij} z_{2j} + \sum_{j=0}^{\frac{n}{2}-1} \xi_n^{i(2j+1)} z_{2j+1}$ ed essendo radici n-esime vale:

$\xi_n^2 = \xi_{\frac{n}{2}}$; quindi $\xi_n^{2ij} = \xi_{\frac{n}{2}}^{ij}$; $\xi_n^{i(2j+1)} = \xi_n^i \xi_{\frac{n}{2}}^{ij}$

L'espressione diviene quindi:

$$y_i = \sum_{j=0}^{\frac{n}{2}-1} \xi_{\frac{n}{2}}^{ij} z_{2j} + \xi_n^i \sum_{j=0}^{\frac{n}{2}-1} \xi_{\frac{n}{2}}^{ij} z_{2j+1}$$

Quindi:

$$\begin{pmatrix} y_0 \\ \vdots \\ y_{\frac{n}{2}-1} \end{pmatrix} = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}}) + \mathrm{Diag}\left(1, \xi_n, \dots, \xi_n^{\frac{n}{2}-1}\right) \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{dispari}})$$

$$\begin{pmatrix} y_{\frac{n}{2}} \\ \vdots \\ y_{n-1} \end{pmatrix} = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}}) + \mathrm{Diag}\left(1, \xi_n, \dots, \xi_n^{\frac{n}{2}-1}\right) \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{dispari}})$$

Questo conclude un passo dell'algoritmo.