

Calcolo Numerico

Federico Matteoni

A.A. 2019/20

Indice

1	Aritmetica di macchina	5
1.1	Rappresentazione in macchina	5
1.1.1	Aritmetica di Macchina	6
1.1.2	Operazioni di Macchina	7
1.1.3	Errore Inerente	9
1.1.4	Errore Algoritmico	9
1.1.5	Errore Totale	10
1.2	Tecniche per l'Analisi degli Errori	10
1.2.1	Coefficiente di Amplificazione	10
2	Problemi dell'Algebra Lineare Numerica	13
2.1	Norme Matriciali e Vettoriali	13
2.1.1	Norma Vettoriale	13
2.1.2	Norma Matriciale	15
2.1.3	Teoremi di Caratterizzazione	15
2.1.4	Condizionamento della risoluzione di un sistema lineare	16
2.2	Autovalori e Autovettori	17
2.2.1	Polinomio caratteristico	17
2.2.2	Diagonalizzazione	17
2.2.3	Teorema di Gerschgorin	18
2.3	Fattorizzazione LU	19
2.3.1	Fattorizzazione LU	20
2.3.2	Calcolo della fattorizzazione LU	22
2.3.3	Predominanza Diagonale	25
2.3.4	Eliminazione Gaussiana con pivoting	26
2.4	Metodi Iterativi	27
2.4.1	Come nascono	28
2.4.2	Implementazione di un metodo iterativo	28
2.4.3	Convergenza	29
2.4.4	Metodi	30
2.4.5	Teorema	31
2.4.6	Esempi	32
3	Equazioni Non Lineari	37
3.1	Metodo di Bisezione	37
3.2	Metodi di Iterazione Funzionale	39
3.2.1	Metodo delle Tangenti	40
4	Esercitazioni	43
5	Matlab	53
5.1	Esponenziale	53

Introduzione

Prof.: Luca Gemignani

Calcolo Numerico Metodi numerici per risolvere problemi matematici con il calcolatore. In questo corso ce ne occuperemo dal punto di vista numerico, perché metodi di risoluzione diversi performano in maniera diversa sulla macchina. Cerchiamo di capire quali sono i metodi di interesse e cosa aspettarci dalla loro implementazione.

Il **metodi numerici approssimano la soluzione di problemi matematici**.

Inoltre, il computer **impatta** sul calcolo perché lavora con approssimazioni dei numeri, che su moli elevate di dati e di elaborazioni finiscono per perturbare il risultato ottenuto.

Tipici problemi

$$Ax = b$$

$$Ax = \lambda x$$

$$f(x) = 0$$

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

Matlab Matrix Laboratory, strumento di implementazione per verificare e constatare i risultati teorici.

Informazioni d'esame Compitini, che se complessivamente passati rendono l'orale facoltativo. In alternativa, appelli scritti + orale.

Capitolo 1

Aritmetica di macchina

Modello per capire cosa aspettarci dal punto di vista degli errori dell'esecuzione.

Esempio Per calcolare il limite

$$\lim_{x \rightarrow \infty} \sqrt{x+1} - \sqrt{x}$$

ottengo un caso indeterminato ($\infty - \infty$). Posso semplificare l'espressione ad esempio razionalizzando, e con pochi passaggi ottengo la seguente uguaglianza

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

Quindi dal punto di vista matematico, le due espressioni sono equivalenti. Ciò **non è sempre vero per la rappresentazione ed il calcolo in macchina**: le due espressioni forniranno risultati completamente diversi. Si rende **necessario**, quindi, **capire quale metodo si comporta meglio** rispetto agli altri.

1.1 Rappresentazione in macchina

Rappresentare i numeri Siamo comunemente abituati a rappresentare un numero in diverse forme.

Ad esempio, $0.1 = \frac{1}{10} = 10 \cdot 0.01$. In generale, per ogni numero ho diversi metodi di rappresentazione \Rightarrow In macchina dobbiamo poter **rappresentare i numeri in maniera univoca**.

Base di numerazione $B \in \mathbb{N}, B > 1$, poiché in base 1 non si può contare.

Una base B ha cifre nell'insieme $\{0, 1, \dots, B-1\}$

Teorema Dato $x \in \mathbb{R}$, con $x \neq 0$

$\exists! (\{d_i\}_{i \geq 1} \text{ con } d_1 \neq 0 \text{ e } d_i \text{ non definitivamente uguali a } B-1) \wedge (p \in \mathbb{Z}) \mid x = \text{segno}(x) \cdot B^p \cdot \sum_{i=1}^{\infty} d_i \cdot B^{-i}$

Considerazioni

Se $x \in \mathbb{C}$ allora viene rappresentato come coppia di numeri reali, quindi il problema si riconduce sempre alla loro rappresentazione

Lo 0 viene rappresentato in modo speciale, poiché non esiste una sua rappresentazione normalizzata

$d_{i \geq 1}$ è una **successione** di cifre

La rappresentazione è **normalizzata** se $d_1 \neq 0$, cioè se la prima cifra è diversa da 0

Non può avere tutte cifre uguali a $B-1$ da un certo punto in poi, la rappresentazione "collassa" al numero successivo

p è detto **esponente**

$\sum_{i=1}^{\infty} d_i \cdot B^{-i}$ è detta **mantissa**

Questa rappresentazione si chiama **in virgola mobile** o **floating point**

Segno	Esponente	Mantissa
-------	-----------	----------

Esempi Ponendo $B = 10$

$$x = 0.01 \Rightarrow x = +10^{-1} \cdot (0.1)$$

$$x = 1.35 \Rightarrow x = +10^1 \cdot (0.135)$$

$$x = 0.0023 \Rightarrow x = +10^{-2} \cdot (0.23)$$

Numeri di Macchina Nei computer ho **registri di lunghezza finita**, quindi essi vengono partizionati: una parte viene riservata alla rappresentazione dell'**esponente** e il resto alla rappresentazione della **mantissa**. L'**insieme dei numeri di macchina** F è quindi così definito

$$F(B, t, m, M) = \left\{ \pm B^p \cdot \sum_{i=1}^t d_i \cdot b^{-i} \mid d_1 \neq 0 \wedge -m \leq p \leq M \right\} \cup \{0\}$$

dove

t sono le **cifre della mantissa**

L'esponente p è compreso tra i valori $-m$ e M

Lo 0 è incluso ma rappresentato a parte

Esempio Ipotizzando di usare registri da 32 bit, posso stanziare 8 bit per l'esponente p (quindi 1 bit per il segno e 7 bit per il valore) e i restanti 24 bit per la mantissa (1 per il segno, 23 per il valore). Quanti numeri posso rappresentare?

p di 7 bit $\Rightarrow 2^7 - 1 = 127 \Rightarrow -127 \leq p \leq 127$ ma lo 0 è rappresentato due volte

$$x = \pm 2^p \sum_{i=1}^2 d_i \cdot 2^{-i}, \text{ con } d_i \in \{0, 1\}, d_1 \neq 0 \Rightarrow d_1 = 1 \text{ sempre}$$

Vedremo che con una serie di accorgimenti è possibile aumentare i numeri esattamente rappresentabili.

Dato quindi $F(B, t, m, M)$, osservo che:

$$F(B, t, m, M) \text{ ha cardinalità finita } N = 2B^{t-1}(B-1)(M+m+1) + 1$$

$$\text{Se } x \in F(B, t, m, M) \wedge x \neq 0 \Rightarrow \omega = B^{-m-1} \leq |x| \leq B^M(1 - B^{-t}) = \Omega$$

Quindi non è possibile rappresentare esattamente numeri non nulli minori di ω . Si può introdurre una rappresentazione **denormalizzata** quando $p = -m$: la condizione $d_1 \neq 0$ può essere abbandonata e posso così rappresentare numeri positivi e negativi compresi in modulo tra B^{-m-t} e $B^{-m}(B^{-1} - B^{-t})$

Analogamente se $p = M$ si introducono rappresentazioni speciali per i simboli $\pm\infty$ e NaN (**not a number**).

1.1.1 Aritmetica di Macchina

La rappresentazione di un numero $x \in \mathbb{R}, x \neq 0$ in macchina significa **approssimare** x con un numero $\bar{x} \in F(B, t, m, M)$ commettendo un **errore relativo** di rappresentazione

$$\epsilon_x = \frac{\bar{x} - x}{x} = \frac{\eta_x}{x}, x \neq 0$$

quanto più piccolo possibile in valore assoluto. La quantità $\eta_x = \bar{x} - x$ è detta **errore assoluto** della rappresentazione. L'errore relativo è importante per la **valutazione qualitativa** dell'errore: nelle misurazioni astronomiche un errore assoluto di 1 cm è *qualitativamente diverso* da un errore assoluto di 1 cm nella misurazione di un tavolo.

Dato $x \in \mathbb{R}, x \neq 0$, distinguiamo due casi:

1. $|x| < \omega$ (**underflow**) oppure $|x| > \Omega$ (**overflow**)
2. $\omega \leq |x| \leq \Omega$

Nel secondo caso abbiamo quattro tecniche di approssimazione:

1. **Arrotondamento:** x approssimato con il numero rappresentabile \bar{x} più vicino
2. **Troncamento:** x approssimato con il più grande numero rappresentabile \bar{x} tale che $|\bar{x}| \leq |x|$
3. *Round toward $+\infty$:* x approssimato al più piccolo numero rappresentabile maggiore del dato
4. *Round toward $-\infty$:* x approssimato al più grande numero rappresentabile minore del dato

Per semplicità considereremo una macchina che opera per troncamento sull'insieme $F(B, t, m, M)$.

Indicheremo con $trn(x) = \bar{x}$ il risultato dell'approssimazione di x con troncamento e più in generale $fl(x)$ l'**approssimazione in macchina del dato** x nel sistema floating point considerato.

Teorema Sia $x \in \mathbb{R}$ con $\omega \leq |x| \leq \Omega$. Si ha

$$|\epsilon_x| = \left| \frac{trn(x) - x}{x} \right| \leq u = B^{1-t}$$

Si osserva che:

$u = B^{1-t}$ è detta **precisione di macchina** ed è **indipendente dalla grandezza del numero**, ma è caratteristica dell'aritmetica floating point, dell'insieme dei numeri rappresentabili e dalla tecnica di approssimazione. Se ad esempio si opera con arrotondamento, u si dimezza.

Per valutare la precisione di macchina possiamo determinare il più piccolo numero di macchina maggiore di 1. Dato x tale numero, abbiamo $x - 1 = |x - 1| = B^{1-t}$ essendo $1 = B^1 \cdot B^{-1}$ rappresentato con esponente $p = 1$. Il seguente script MatLab fornisce il valore richiesto:

```

eps = 0.5;
eps1 = eps + 1;
while(eps > 1)
    eps = 0.5 * eps;
    eps1 = eps + 1;
end
eps = 2 * eps;

```

Dal teorema si ricava che dato $x \in \mathbb{R}$, in assenza di overflow/underflow, vale $fl(x) = x(1 + \epsilon_x)$ con $|\epsilon_x| \leq u$.

Questa relazione esprime il modo in cui viene descritto generalmente il legame tra numero reale e sua rappresentazione in macchina.

1.1.2 Operazioni di Macchina

Per le **operazioni aritmetiche** in un sistema floating point si pone un analogo problema di approssimazione, in quanto **il risultato di un'operazione eseguita tra due numeri di macchina in generale non sarà un numero di macchina**.

Operazioni Indichiamo con $\oplus, \ominus, \otimes, \oslash$ le **operazioni aritmetiche di macchina** corrispondenti relativamente all'addizione, sottrazione, prodotto e divisione. Si richiede che le operazioni siano interne all'insieme dei numeri di macchina. Una ragionevole definizione, derivata dal teorema precedente e in assenza di overflow/underflow, risulta:

$$\forall a, b \in F(B, t, m, M), a \oplus b = fl(a + b) = (a + b)(1 + \epsilon_1), |\epsilon_1| \leq u$$

con ϵ_1 detto **errore locale dell'operazione**. Sempre in assenza di overflow/underflow, se $a, b \in \mathbb{R}$ si ha

$$fl(a + b) = fl(a) \oplus fl(b) = (a(1 + \epsilon_a) + b(1 + \epsilon_b))(1 + \epsilon_1) \doteq (a + b) + a\epsilon_a + b\epsilon_b + (a + b)\epsilon_1$$

dove con \doteq si indica che l'eguaglianza vale **considerando le sole componenti lineari negli errori** e trascurando le componenti di ordine superiore (**analisi al primo ordine dell'errore**), in virtù del fatto che gli ϵ sono quantità piccole $0 < \epsilon < 1$, quindi trascurabili negli ordini superiori al primo.

Si ottiene che, in assenza di overflow/underflow, se $a, b \in \mathbb{R}, a + b \neq 0$, allora

$$\frac{fl(a + b) - (a + b)}{a + b} \doteq \frac{a}{a + b}\epsilon_a + \frac{b}{a + b}\epsilon_b + \epsilon_1$$

che esprime la dipendenza dell'errore totale commesso nel calcolo della somma tra due numeri reali rispetto agli errori generati dall'approssimazione dei dati iniziali (**errore inerente**) e agli errori generati dall'algoritmo di calcolo (**errore algoritmico**) visto come sequenza di operazioni aritmetiche.

Esempio Analizziamo cosa succede in macchina quando proviamo a calcolare $f(x) = x^2 - 1 = (x - 1)(x + 1)$. La **prima situazione di errore** sia ha sulla rappresentazione di x che, **in generale, non è un numero di macchina**.

$$x \rightarrow \tilde{x} = x(1 + \epsilon_x)$$

con $|\epsilon_x| \leq u$. Inoltre, sempre in generale, **le operazioni aritmetiche non sono operazioni di macchina**

$$f(x) = (\tilde{x} \otimes \tilde{x}) \ominus 1 = (\tilde{x} \ominus 1) \otimes (\tilde{x} \oplus 1)$$

Poniamo $g_1(x) = (\tilde{x} \otimes \tilde{x}) \ominus 1$ e $g_2(x) = (\tilde{x} \ominus 1) \otimes (\tilde{x} \oplus 1)$. La formula per l'**errore totale** dell'operazione è

$$\epsilon_{tot1} = \frac{g_1(x) - f(x)}{f(x)}$$

$$\epsilon_{tot2} = \frac{g_2(x) - f(x)}{f(x)}$$

Sviluppiamo $g_1(x)$

$$\begin{aligned} g_1(x) &= ((x(1 + \epsilon_x) \cdot x(1 + \epsilon_x))(1 + \epsilon_1) - 1)(1 + \epsilon_2) \doteq \\ &\doteq (x^2(1 + 2\epsilon_x)(1 + \epsilon_1) - 1)(1 + \epsilon_2) \doteq \\ &\doteq (x^2((1 + 2\epsilon_x + \epsilon_1) - 1)(1 + \epsilon_2) \doteq \\ &\doteq x^2(1 + 2\epsilon_x + \epsilon_1 + \epsilon_2) - (1 + \epsilon_2) = \\ &= (x^2 - 1) + 2x^2\epsilon_x + x^2\epsilon_1 + (x^2 - 1)\epsilon_2 = g_1(x) \\ &|\epsilon_i| \leq u \end{aligned}$$

Quindi, portando al primo membro dell'uguaglianza $\epsilon_{tot1} = \frac{g_1(x) - f(x)}{f(x)}$ tutti i fattori $(x^2 - 1) = f(x)$ si ottiene

$$\epsilon_{tot1} = \frac{2x^2}{x^2 - 1}\epsilon_x + \frac{x^2}{x^2 - 1}\epsilon_1 + \epsilon_2$$

Si evince che **l'errore totale è la somma di due componenti**:

Errore inerente o inevitabile: **l'errore di rappresentazione di x** , che vale 0 se x è numero di macchina ed è **proprietà della funzione**. Nell'esempio, $\epsilon_{in} = \frac{2x^2}{x^2 - 1}\epsilon_x$.
Se l'errore inerente è piccolo si dice che **la funzione è numericamente stabile**, viceversa è **numericamente instabile**.

Errore algoritmico, locale all'operazione e **proprietà dell'algoritmo**. Nell'esempio, $\epsilon_{alg} = \frac{x^2}{x^2 - 1}\epsilon_1 + \epsilon_2$.
Se è piccolo, si dice che **l'espressione è ben condizionata e poco sensibile rispetto alla perturbazione**.
Viceversa, è **mal condizionata**.

Sviluppiamo ora $g_2(x)$ (gli errori δ_i sono analoghi agli ϵ_i , si usa una notazione differente per evidenziare che hanno valori diversi, esseno propri del calcolo e dei valori usati in esso)

$$\begin{aligned} g_2(x) &= ((x(1 + \epsilon_x) - 1)(1 + \delta_1))((x(1 + \epsilon_x) + 1)(1 + \delta_2))(1 + \delta_3) \doteq \\ &\doteq (x^2(1 + \epsilon_x)^2 - 1)(1 + \delta_1 + \delta_2 + \delta_3) \doteq \\ &\doteq (x^2(1 + 2\epsilon_x) - 1)(1 + \delta_1 + \delta_2 + \delta_3) = \\ &= x^2(1 + 2\epsilon_x + \delta_1 + \delta_2 + \delta_3) - (1 + \delta_1 + \delta_2 + \delta_3) = \\ &= (x^2 - 1) + 2x^2\epsilon_x + (x^2 - 1)(\delta_1 + \delta_2 + \delta_3) = g_2(x) \\ &|\epsilon_x| \leq u, |\delta_i| \leq u \end{aligned}$$

Come prima, porto gli $f(x)$ al primo membro dell'uguaglianza $\epsilon_{tot2} = \frac{g_2(x) - f(x)}{f(x)}$ e ottengo

$$\epsilon_{tot2} = \frac{2x^2}{x^2 - 1}\epsilon_x + \delta_1 + \delta_2 + \delta_3$$

Notiamo come

$\epsilon_{in} = \frac{2x^2}{x^2-1}\epsilon_x$, come il calcolo precedente. Infatti, ripetiamo, l'errore inerente è una proprietà della funzione e non di come essa viene calcolata

$\epsilon_{alg} = \delta_1 + \delta_2 + \delta_3$, diverso poiché abbiamo seguito un calcolo differente

Per poter comparare i due algoritmi e scegliere il migliore, analizziamo gli errori algoritmici. L'analisi **va eseguita in valore assoluto**, poiché non si ha alcuna informazione sul segno degli errori, seguendo quindi le regole di comparazione dei valori assoluti:

$$|a + b| \leq |a| + |b|$$

$$|a \cdot b| = |a| \cdot |b|$$

$$\begin{aligned}\epsilon_{alg1} &= \left| \frac{x^2}{x^2-1}\epsilon_1 + \epsilon_2 \right| \leq \left| \frac{x^2}{x^2-1}\epsilon_1 \right| + |\epsilon_2| = \left| \frac{x^2}{x^2-1} \right| \cdot |\epsilon_1| + |\epsilon_2| \leq \frac{x^2}{|x^2-1|}u + u = \left(\frac{x^2}{|x^2-1|} + 1 \right)u \\ \epsilon_{alg2} &= |\delta_1 + \delta_2 + \delta_3| \leq |\delta_1| + |\delta_2| + |\delta_3| \leq 3u\end{aligned}$$

In ϵ_{alg1} , quindi, l'errore algoritmo è minore o uguale a $\left(\frac{x^2}{|x^2-1|} + 1 \right)u$, che però diventa arbitrariamente grande quando x si avvicina ad 1. Il secondo algoritmo è dunque **preferibile**, poiché l'**errore algoritmico è limitato**.

Esempio Calcoliamo ora l'espressione $f(x) = x^2 + 1$. Poniamo $g(x) = (\tilde{x} \otimes \tilde{x}) \oplus 1$ e sviluppiamo, ottenendo

$$g(x) \doteq x^2(1 + 2\epsilon_x + \epsilon_1 + \epsilon_2) + (1 + \epsilon_2)$$

L'errore totale è quindi

$$\epsilon_{tot} = \frac{g(x) - f(x)}{f(x)} = \frac{2x^2}{x^2+1}\epsilon_x + \frac{x^2}{x^2+1}\epsilon_1 + \epsilon_2$$

Studiamo le componenti

$$|\epsilon_{in}| = \left| \frac{2x^2}{x^2+1}\epsilon_x \right| = \left| \frac{2x^2}{x^2+1} \right| \cdot |\epsilon_x| \leq \frac{2x^2}{x^2+1}u \leq 2u$$

perché $\frac{x^2}{x^2+1} \leq 1$. La funzione quindi **non è suscettibile rispetto alle perturbazioni dei dati in ingresso** ed è **ben condizionata**.

$$\epsilon_{alg} = \left| \frac{x^2}{x^2+1}\epsilon_1 + \epsilon_2 \right| \leq \left| \frac{x^2}{x^2+1}\epsilon_1 \right| + |\epsilon_2| = \frac{x^2}{x^2+1}|\epsilon_1| + |\epsilon_2| \leq |\epsilon_1| + |\epsilon_2| \leq 2u$$

Quindi l'algoritmo è **numericamente stabile** perché la componente dell'errore algoritmo è piccola ($\leq 2u$). Questo è il caso migliore che può capitare: **funzione ben condizionata** e **algoritmo numericamente stabile**.

Succede perché non vi è la sottrazione al numeratore. La sottrazione, in macchina, lavora usando molte delle cifre "sporche" della mantissa, cioè quelle che fanno parte del rumore e dell'errore di rappresentazione. Si chiama **errore di cancellazione**.

1.1.3 Errore Inerente

Definizione Si dice **errore inerente** o **inevitabile** generato nel calcolo di $f(x) \neq 0$ la quantità

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

Osserviamo che l'errore inerente è una **proprietà della funzione**, ovvero del problema matematico, e misura la **sensibilità rispetto alla perturbazione del dato iniziale**. Quindi, è indipendente dall'algoritmo di calcolo.

Se l'errore inerente è qualitativamente elevato in valore assoluto, diciamo che **il problema matematico è mal condizionato**, altrimenti si dice che è **ben condizionato**.

1.1.4 Errore Algoritmico

Definizione Si dice **errore algoritmico** generato nel calcolo di $f(x) \neq 0$ la quantità

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

Osserviamo che l'errore algoritmico è **dipendente dall'algoritmo utilizzato**: la funzione g dipende dall'algoritmo usato per calcolare $f(x)$. In generale, **differenti algoritmi conducono a differenti errori algoritmici**.

Se l'errore algoritmico è qualitativamente elevato in valore assoluto si dice che **l'algoritmo è numericamente instabile**, altrimenti si dice che è **numericamente stabile**.

1.1.5 Errore Totale

Definizione Si dice **errore totale** generato nel calcolo di $f(x) \neq 0$ mediante l'algoritmo specificato da g la quantità

$$\epsilon_{tot} = \frac{g(\tilde{x}) - f(x)}{f(x)}$$

L'errore totale misura la **differenza relativa tra l'output atteso e l'output effettivamente calcolato**.

Teorema In un'analisi al primo ordine, vale

$$\epsilon_{tot} = \epsilon_{in} + \epsilon_{alg}$$

Dim

$$\epsilon_{tot} = \frac{g(\tilde{x}) - f(x)}{f(x)} = \frac{f(\tilde{x}) - f(x)}{f(\tilde{x})} \cdot \frac{f(\tilde{x})}{f(x)} + \frac{f(\tilde{x}) - f(x)}{f(x)} = \epsilon_{alg}(1 + \epsilon_{in}) + \epsilon_{in} \doteq \epsilon_{alg} + \epsilon_{in}$$

Viene espresso il fatto che nel calcolo di una funzione razionale, in un'analisi al primo ordine, le due fonti di generazione d'errore forniscono contributi separati che possono essere analizzati indipendentemente.

L'obiettivo dell'analisi numerica è **trovare algoritmi numericamente stabili per problemi ben condizionati**.

1.2 Tecniche per l'Analisi degli Errori

1.2.1 Coefficiente di Amplificazione

Dalla seguente relazione

$$\epsilon_{in} = \frac{f(\tilde{x}) - f(x)}{f(x)} = \frac{f(\tilde{x}) - f(x)}{\tilde{x} - x} \cdot \frac{x}{f(x)} \cdot \frac{\tilde{x} - x}{x}$$

si ricava che la **differenziabilità di $f(x)$ è essenziale** per il controllo dell'errore inerente. Se assumiamo che $f(x)$ è derivabile due volte e continua in (a, b) , allora vale lo sviluppo di Taylor

$$f(\tilde{x}) = f(x) + f'(x)(\tilde{x} - x) + \frac{f''(\xi)}{2}(\tilde{x} - x)^2, \quad |\xi - x| \leq |\tilde{x} - x|$$

da cui si ottiene

$$\epsilon_{in} = \frac{f(\tilde{x}) - f(x)}{f(x)} \doteq \frac{f'(x)}{f(x)} x \epsilon_x = c_x \epsilon_x, \quad c_x = \frac{f'(x)}{f(x)} x$$

La quantità $c_x = c_x(f) = \frac{f'(x)}{f(x)} x$ è detta **coefficiente di amplificazione** e fornisce una **misura del condizionamento del problema**.

In generale, se $f: \Omega \rightarrow \mathbb{R}$ è definita su un insieme aperto di \mathbb{R}^n , differenziabile due volte su Ω ed il segmento di estremi $\tilde{\mathbf{x}}$ e \mathbf{x} è contenuto in Ω allora vale

$$\epsilon_{in} = \frac{f(\tilde{\mathbf{x}}) - f(\mathbf{x})}{f(\mathbf{x})} \doteq \frac{1}{f(\mathbf{x})} \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{x}) x_i \epsilon_{x_i} = \sum_{i=1}^n c_{x_i}(f) \epsilon_{x_i}$$

con

$$c_{x_i}(f) = \frac{1}{f(\mathbf{x})} \frac{\partial f}{\partial x_i}(\mathbf{x}) x_i, \quad 1 \leq i \leq n$$

detti **coefficienti di amplificazione della funzione f rispetto alla variabile x_i** .

Esempio per $f(x) = (x^2 + 1)/x$ si ha

$$c_x = \frac{2 - (x^2 + 1)}{x^2} \cdot \frac{x}{x^2 + 1} \cdot x = \frac{x^2 - 1}{x^2 + 1}$$

e poiché $|c_x| \leq 1$ il problema risulta ben condizionato.

Coefficienti

Per le operazioni aritmetiche si ottiene

$$f(x, y) = x + y \quad \epsilon_{in} \doteq c_x \epsilon_x + c_y \epsilon_y \quad c_x = \frac{x}{x+y} \quad c_y = \frac{y}{x+y}$$

$$f(x, y) = x - y \quad \epsilon_{in} \doteq c_x \epsilon_x + c_y \epsilon_y \quad c_x = \frac{x}{x-y} \quad c_y = -\frac{y}{x-y}$$

$$f(x, y) = x \cdot y \quad \epsilon_{in} \doteq c_x \epsilon_x + c_y \epsilon_y \quad c_x = 1 \quad c_y = 1$$

$$f(x, y) = \frac{x}{y} \quad \epsilon_{in} \doteq c_x \epsilon_x + c_y \epsilon_y \quad c_x = 1 \quad c_y = -1$$

Ne segue, come visto in precedenza, che la sottrazione di due numeri vicini tra loro è potenziale causa di elevata amplificazione degli errori relativi (**cancellazione numerica**). Si nota anche come le operazioni moltiplicative siano ben condizionate.

Esempio Siano $x = 0.2178 \cdot 10^2$ e $y = 0.218 \cdot 10^2$, supponendo di operare con troncamento in base $B = 10$ e $t = 3$ cifre di rappresentazione ($u = 10^{1-t} = 10^{-2}$). Si ha $\tilde{x} = 0.217 \cdot 10^2$ e $\tilde{y} = y$.

Pertanto $\tilde{x} \oplus \tilde{y} = -0.001 \cdot 10^2 = -0.1$ mentre $x - y = -0.0002 \cdot 10^2 = -0.2 \cdot 10^{-1}$ e quindi $|\epsilon_{in}| = \frac{0.8}{0.2} = 0.4$.

Capitolo 2

Problemi dell'Algebra Lineare Numerica

2.1 Norme Matriciali e Vettoriali

I principali problemi dell'algebra lineare concernono la **risoluzione di sistemi lineari** ed il **calcolo di autovalori/autovettori di matrici**. Per studiarne il condizionamento è fondamentale avere degli strumenti per valutare la distanza tra vettori e tra matrici.

La risoluzione di sistemi lineari può essere eseguita in aritmetica reale, ma gli autovalori/autovettori possono essere complessi. Servono quindi **strumenti che operino su spazi vettoriali** F^n e $F^{n \times n}$, con $n \geq 1$ e $F \in \{\mathbb{R}, \mathbb{C}\}$

Condizionamento su problemi reali $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $G_m = \frac{f(\tilde{x}) - f(x)}{f(x)}$

Problema: calcolo della soluzione di un sistema lineare

Input $A \in \mathbb{R}^{2 \times 2}$, $b \in \mathbb{R}^2$

Output $x \in \mathbb{R}^2 \mid Ax = b$

Condizionamento: $A, B \rightarrow$ devono essere approssimati a valori di macchina

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \hat{A} = \begin{bmatrix} \hat{a} & \hat{b} \\ \hat{c} & \hat{d} \end{bmatrix} \text{ con } \hat{a} = a(1 + \epsilon_a) \dots$$

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \rightarrow \hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \text{ con } \hat{b}_i = b_i(1 + \epsilon_{b_i}) \dots$$

$$Ax = b \rightarrow \hat{A}\hat{x} = \hat{b}$$

Quando \hat{x} è vicino a x ? Per valutare la distanza fra vettori, ho bisogno delle **norme vettoriali**

2.1.1 Norma Vettoriale

Definizione $f : \mathbb{R}^n \rightarrow \mathbb{R}$ si dice **Norma Vettoriale** se soddisfa le seguenti proprietà

1. $f(x) \geq 0 \wedge (f(x) = 0 \Leftrightarrow x = 0)$
2. $f(\alpha x) = |\alpha|f(x) \quad \forall x \in \mathbb{R}^n \quad \forall \alpha \in \mathbb{R}$
3. $f(x + y) \leq f(x) + f(y) \quad \forall x, y \in \mathbb{R}^n$

Norma Euclidea

$$f(x) = \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

con $x = [x_1, x_2, \dots, x_n]$.

Norma Infinito

$$f(x) = \|x\|_\infty = \max_{i=1}^n |x_i|$$

Dimostrazione che è una norma vettoriale

1. $f(x) = \max |x_i|$ è sempre ≥ 0
 $f(x) = \max |x_i| = 0 \Leftrightarrow |x_i| = 0 \quad \forall i \Leftrightarrow x = 0$
2. $f(\alpha x) = \max |\alpha x_i| = \max |\alpha| |x_i| = |\alpha| \cdot \max |x_i| = |\alpha| f(x)$
3. $f(x + y) = \max |x_i + y_i| \leq \max (|x_i| + |y_i|) \leq \max |x_i| + \max |y_i| = f(x) + f(y)$

Norma Uno

$$f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$$

In generale le norme di uno stesso vettore hanno valori differenti

Esempio $x = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

$$\|x\|_1 = |2| + |3| = 5$$

$$\|x\|_2 = \sqrt{2^2 + 3^2} = \sqrt{4 + 9} = \sqrt{13}$$

$$\|x\|_\infty = 3$$

Distanza Data $f : \mathbb{R}^n \rightarrow \mathbb{R}$ **norma**, allora possiamo definire la **distanza su \mathbb{R}^n** come

$$dist(x, y) = \|x - y\| = f(x - y)$$

1. $f(x) \geq 0$ e $f(y) \geq 0 \Rightarrow dist \geq 0$ e ($dist = 0 \Leftrightarrow x = y$)
2. $f(\alpha x) = |\alpha| f(x)$
3. $f(x + y) \leq f(x) + f(y)$

Anche la distanza assume valori differenti su norme differenti

Esempio $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

$$dist(x, y) = \|x - y\|_1 = \left\| \begin{bmatrix} -1 \\ -3 \end{bmatrix} \right\|_1 = 4$$

$$dist(x, y) = \|x - y\|_2 = \left\| \begin{bmatrix} -1 \\ -3 \end{bmatrix} \right\|_2 = \sqrt{10}$$

$$dist(x, y) = \|x - y\|_\infty = \left\| \begin{bmatrix} -1 \\ -3 \end{bmatrix} \right\|_\infty = 3$$

Equivalenza topologica Può succedere che x e y siano vicini in norma uno e lontani in norma 2? **No** per l'**equivalenza topologica delle norme**: quantitativamente differenti, ma qualitativamente danno gli stessi risultati.

$$\|x - y\|_1 \text{ è piccolo} \Rightarrow \|x - y\|_2 \text{ è piccolo}$$

$$\|x - y\|_\infty \text{ è grande} \Rightarrow \|x - y\|_1 \text{ è grande}$$

2.1.2 Norma Matriciale

Definizione $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ si dice **norma matriciale** se soddisfa le seguenti proprietà

1. $f(A) \geq 0 \wedge (f(A) = 0 \Leftrightarrow A = 0 \Leftrightarrow a_{ij} = 0 \ \forall i, j)$
2. $f(\alpha A) = |\alpha| f(A)$
3. $f(A + B) \leq f(A) + f(B)$
4. $f(A \cdot B) \leq f(A) \cdot f(B)$

Difficile definire le norme metriciali, quindi **costruiamo le norme matriciali a partire dalle norme vettoriali**.

Definizione Data $f_V : \mathbb{R}^n \rightarrow \mathbb{R}$ norma vettoriale, si dice **norma matriciale indotta (compatibile)** la norma $f_M : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ definita da

$$f_M(A) = \max_{\{x \in \mathbb{R}^n \mid f_V(x)=1\}} f_V(Ax)$$

1. Prendo i vettori $x \in \mathbb{R}^n \mid f_V(x) = 1$, in generale sono infiniti
2. Calcolare Ax , in generale infiniti vettori
3. Prendo $f_V(Ax)$, infiniti **valori**
4. Calcolo il **massimo** di questi valori

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1$$

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$$

Esempio $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \rightarrow \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$.

$\|x\|_2 = 1 \Leftrightarrow \sqrt{x_1^2 + x_2^2} = 1 \Leftrightarrow x_1^2 + x_2^2 = 1$ che è l'equazione per la circonferenza di centro 0 e raggio 1. Quindi tutti i vettori per cui $\|x\|_2 = 1$ sono tutti quelli la cui punta giace sulla circonferenza, quindi tanti quanti i punti della circonferenza **quindi infiniti**.

2.1.3 Teoremi di Caratterizzazione

La definizione quindi **non è pratica** (calcolabile) in generale, perché dovrei fare un massimo su infiniti termini. Quindi si introducono dei **teoremi di caratterizzazione**.

Teorema 1 $\|A\|_2 = \sqrt{\rho(A^T A)}$

Dove $\rho(B) = \max |\lambda_i|$ e $\lambda_i \in \text{spettro}(B)$

$\rho(B)$ è il **raggio spettrale** di B , cioè il **modulo dell'autovalore di modulo massimo** ($\text{spettro}(B)$ è l'insieme degli autovalori).

Teorema 2 $\|A\|_\infty = \max_{i=1}^n \sum_{j=1}^n |a_{ij}|$ **somme per riga poi prendo il massimo**

Teorema 2 $\|A\|_1 = \max_{j=1}^n \sum_{i=1}^n |a_{ij}|$ **somme per colonna poi prendo il massimo**

Da questi teoremi, se ne deriva che **calcolare la norma uno e la norma infinito è più semplice che calcolare la norma euclidea**.

Vantaggi delle norme matriciali \rightarrow **ulteriorie proprietà** delle norme matriciali indotte

Teorema Sia $\|\bullet\|$ una norma vettoriale e $\|\cdot\|$ la norma matriciale indotta. Allora $\forall A \in \mathbb{R}^{n \times n}$, $\forall v \in \mathbb{R}^n$ si ha $\|Av\| \leq \|A\| \cdot \|v\|$

Dimostrazione Due casi

$$v = 0 \quad \|Av\| = \|0\| = 0 \\ \|A\| \cdot \|v\| = \|A\| \cdot 0 = 0$$

$$v \neq 0 \quad \|Av\| = \|A \frac{v}{\|v\|} \cdot \|v\|\| \text{ e chiamo il vettore } A \frac{v}{\|v\|} = z. \text{ Quindi, per la seconda propriet\`a avr\`o} \\ = \|z\| \cdot \|v\| = \|A \frac{v}{\|v\|}\| \cdot \|v\| \leq \|A\| \cdot \|v\|$$

2.1.4 Condizionamento della risoluzione di un sistema lineare

$$Ax = b \longrightarrow \widehat{A}\widehat{x} = \widehat{b}$$

Per semplicit\`a di analisi, $\widehat{A} = A$ e considero la perturbazione solo su b .

$$Ax = b \longrightarrow A\widehat{x} = \widehat{b}$$

Supponiamo anche A invertibile.

Misurare il condizionamento significa misurare quanto x e \widehat{x} sono vicini, attraverso

$$\epsilon_{in} = \frac{\|x - \widehat{x}\|}{\|x\|}$$

Studiamo la quantit\`a

$$\|x - \widehat{x}\| = \|A^{-1}b - A^{-1}\widehat{b}\| = \\ \widehat{b} = b + f \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \rightarrow \widehat{b} = \begin{bmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{bmatrix} = \begin{bmatrix} b_1(1 + \epsilon_1) \\ b_2(1 + \epsilon_2) \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} b_1\epsilon_1 \\ b_2\epsilon_2 \end{bmatrix} = b + f \\ = \|A^{-1}b - A^{-1}(b + f)\| = \|-A^{-1}f\| = \|A^{-1}f\| \leq \|A^{-1}\| \cdot \|f\|$$

$$\text{Quindi } \|x - \widehat{x}\| \leq \|A^{-1}\| \cdot \|f\|$$

Siccome $\|x\|$ sta al denominatore, maggiorazione a denominatore usa una quantit\`a pi\`u piccola.

$$Ax = b \Rightarrow \|Ax\| = \|b\| \Leftrightarrow \|b\| = \|Ax\| \leq \|A\| \cdot \|x\|$$

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

Da questo ottengo

$$\frac{\|x - \widehat{x}\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|f\|}{\frac{\|b\|}{\|A\|}} = \|A\| \cdot \|A^{-1}\| \cdot \frac{\|f\|}{\|b\|}$$

Quindi

$$\epsilon_{in} = \frac{\|x - \widehat{x}\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\widehat{b} - b\|}{\|b\|}$$

$\frac{\|\widehat{b} - b\|}{\|b\|}$ **perturbazione relativa sul termine noto**
 $\|A\| \cdot \|A^{-1}\|$ **coefficiente di amplificazione**

Quindi se $\|A\| \cdot \|A^{-1}\|$ \`e grande, allora il sistem \`e mal condizionato.

Se $\|A\| \cdot \|A^{-1}\|$ \`e piccolo, allora \`e ben condizionato. Piccolo quanto?

Osservazione $\|A \cdot A^{-1}\| = \|I\| \leq \|A\| \cdot \|A^{-1}\|$, ma $\|I\| = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1$
 Quindi $\|A\| \cdot \|A^{-1}\| \geq 1$ che quindi \`e un coefficiente di amplificazione effettivo, con valore minimo pari a 1.

2.2 Autovalori e Autovettori

Autovalore $\lambda \in \mathbb{C}$ è **autovalore** di $A \in \mathbb{C}^{n \times n}$ se $\exists x \neq 0 \mid A \cdot x = \lambda \cdot x$
 Attenzione a $x \neq 0$.

Autovettore x è detto **autovettore destro** relativo all'autovalore λ

Si osserva che

$$Ax = \lambda x \Leftrightarrow Ax - \lambda x = 0 \Leftrightarrow (A - \lambda I)x = 0$$

Quindi possiamo alternativamente dire

Definizione $\lambda \in \mathbb{C}$ è **autovalore** di $A \in \mathbb{C}^{n \times n} \Leftrightarrow \det(A - \lambda I) = 0$

2.2.1 Polinomio caratteristico

$p(\lambda) = \det(A - \lambda I)$ è **polinomio caratteristico** di A

Definizione $\lambda \in \mathbb{C}$ è **autovalore** di $A \in \mathbb{C}^{n \times n} \Leftrightarrow p(\lambda) = 0$ con $p(z) = \det(A - zI)$ polinomio caratteristico di A
 Quindi, per trovare gli autovalori un metodo è trovare il polinomio caratteristico e trovarne gli zeri.
 Gli **autovalori sono gli zeri del polinomio caratteristico**.

Osservazione L'autovettore non è univocamente determinato, ma è determinato a meno di una costante moltiplicativa, quindi in generale sono infiniti autovettori per ogni autovalore.

$p(z) = \det(A - zI)$ con $A \in \mathbb{C}^{n \times n}$ è un **polinomio di grado n** . Possiamo esser anche più precisi: siccome è un polinomio il cui termine di grado più alto si ottiene moltiplicando gli elementi della diagonale principale, può essere scritto come

$$p(z) = (-1)^n z^n + p_{n-1} z^{n-1} + \dots + p_1 z + p_0$$

Teorema Fondamentale dell'Algebra Un polinomio di grado n , su \mathbb{C} , ha n zeri eventualmente contati con la loro molteplicità algebrica. Quindi $p(z)$ su \mathbb{C} ha n zeri $\lambda_1 \dots \lambda_n$ eventualmente ripetuti.
 Quindi possiamo scriverlo come

$$p(z) = (-1)^n \cdot \prod_{i=1}^n (z - \lambda_i)$$

che si può ulteriormente riscrivere mettendo insieme tutti gli autovalori uguali

$$p(z) = (-1)^n \cdot \prod_{i=1}^k (z - \lambda_i)^{\sigma_i}$$

con k numero di zeri distinti, cioè di autovalori distinti ($\lambda_1 \neq \lambda_2 \neq \dots \neq \lambda_k$) e σ_i è detta **molteplicità algebrica dell'autovalore λ_i** .

Molteplicità Algebrica La **molteplicità algebrica** non è altro che il numero di **volte che l'autovalore si ripete** come zero del polinomio caratteristico.

Molteplicità Geometrica Se λ è autovalore di A , allora $(A - \lambda I)$ è una **matrice singolare** \Rightarrow
 $\tau = \dim(\text{Ker}(A - \lambda I)) \geq 1$ e τ_i è la **molteplicità geometrica** dell'autovalore λ_i

Teorema La **molteplicità algebrica** è maggiore o uguale della **molteplicità geometrica**, cioè $\sigma_i \geq \tau_i$

2.2.2 Diagonalizzazione

Definizione $A \in \mathbb{C}^{n \times n}$ si dice **diagonalizzabile** se $\exists S \in \mathbb{C}^{n \times n}$ invertibile $\mid S^{-1} \cdot A \cdot S = D$ **diagonale**
 Gli **autovalori di A sono gli autovalori di D** , questo perché

$$\det(D - \lambda I) = \det(S^{-1}AS - \lambda I) = \det(S^{-1}AS - \lambda S^{-1}S) = \det(S^{-1}(A - \lambda I)S) = \det(S^{-1}) \cdot \det(A - \lambda I) \cdot \det(S) = \det(A - \lambda I)$$

poiché $\det(S) \cdot \det(S^{-1}) = \det(SS^{-1}) = \det(I) = 1$. Quindi D e A hanno lo stesso polinomio caratteristico, quindi stessi autovalori.

Teorema $A \in \mathbb{C}^{n \times n}$ **diagonalizzabile** $\Leftrightarrow \sigma_i = \tau_i$ per $i = 1 \dots k$, cioè le molteplicità algebriche e geometriche sono coincidenti.

Esempi di matrici diagonalizzabili

$k = n$, cioè la matrice ha tutti autovalori distinti. Quindi $\sigma_i = \tau_i = 1$ per $i = 1 \dots n$

(**Teorema Spettrale**) A diagonalizzabile se $A = A^T$, cioè se A è **simmetrica** \Rightarrow **autovalori** $\in \mathbb{R}$

Diagonalizzare $S^{-1}AS = D \Leftrightarrow AS = SD \Leftrightarrow AS \cdot e_j = SD \cdot e_j$ per $j = 1 \dots n$

Sappiamo che $S \cdot e_j$ corrisponde alla j -esima colonna di S , che chiamiamo S_j . Inoltre $D \cdot e_j$ è semplicemente $d_j \cdot e_j$ perché D è diagonale e ha solo l'elemento d_j sulla j -esima colonna.

Quindi $\Leftrightarrow AS_j = Sd_j e_j = d_j S e_j = d_j S_j$

Questo ci dice che **le colonne di S sono gli autovettori di A .**

Calcolare gli autovalori di una matrice è, in generale, un problema difficile, perché gli autovalori non sono funzioni razionali dei coefficienti della matrice.

Quindi **non esiste algoritmo che esegue un numero finito di operazioni razionali per calcolare gli autovalori.**

In generale si utilizzano **metodi iterativi** che **costruiscono successioni convergenti agli autovalori**. Per la convergenza di queste successioni, è **importante avere teoremi di localizzazione che mi dicono dove stanno gli autovalori nel piano complesso.**

2.2.3 Teorema di Gerschgorin

Sia $A \in \mathbb{C}^{n \times n}$ e siano K_i gli **insiemi** definiti da

$$K_i = \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}| \right\} \quad \text{per } i = 1 \dots n$$

Allora se $\lambda \in \mathbb{C}$ è autovalore di A

$$\Rightarrow \lambda \in \bigcup_{i=1}^n K_i$$

Gli autovalori di A quindi appartengono ad almeno uno dei K_i , i cosiddetti **cerchi di Gerschgorin**, dove a_{ii} è l'elemento della diagonale principale, $|z - a_{ii}|$ è il **centro** e $\sum_{j=1, j \neq i}^n |a_{ij}|$ è il **raggio**, formato dalla somma degli elementi della stessa riga escluso l'elemento della diagonale principale.

Dimostrazione λ autovalore di A se $\exists x \neq 0 \mid Ax = \lambda x$. Riscrivendo per componenti, otteniamo

$$Ax = \lambda x \Leftrightarrow \sum_{j=1}^n a_{ij} x_j = \lambda x_i \quad \text{per } i = 1 \dots n$$

Porto al primo membro

$$\Leftrightarrow \sum_{j=1, j \neq i}^n a_{ij} x_j = (\lambda - a_{ii}) x_i \quad \text{per } i = 1 \dots n$$

$x \neq 0$, quindi esiste almeno una componente $\neq 0$. Prendiamo quella di modulo massimo x_p , quindi $|x_p| = \|x\|_\infty$

$$Ax = \lambda x \Rightarrow \sum_{j=1, j \neq p}^n a_{pj} x_j = (\lambda - a_{pp}) x_p \Rightarrow \left| \sum_{j=1, j \neq p}^n a_{pj} x_j \right| = |(\lambda - a_{pp}) x_p|$$

Per le proprietà del valore assoluto

$$\Rightarrow |\lambda - a_{pp}| \cdot |x_p| \leq \sum_{j=1, j \neq p}^n |a_{pj}| \cdot |x_j|$$

$x_p \in \mathbb{C}$ in generale, quindi posso dividere per $|x_p| \in \mathbb{R}$, sicuramente > 0 e sicuramente positivo quindi la disequazione non cambia verso

$$\Rightarrow |\lambda - a_{pp}| \leq \sum_{j=1, j \neq p}^n |a_{pj}| \cdot \frac{|x_j|}{|x_p|}$$

Siccome x_p è la componente di modulo massimo, allora tutte le $\frac{|x_j|}{|x_p|} \leq 1$ posso concludere che

$$\Rightarrow |\lambda - a_{pp}| \leq \sum_{j=1, j \neq p}^n |a_{pj}|$$

Per la definizione di K_p , questa relazione $\Rightarrow \lambda \in K_p$. Dato che non conosco p in generale,

$$\Rightarrow \lambda \in \bigcup_{i=1}^n K_i$$

Abbiamo considerato i cerchi **per riga**. Ma posso anche considerare i cerchi per colonna, perché **gli autovalori di A sono gli autovalori di A^T** in generale.

Questo perché $\det(A - \lambda I) = \det((A - \lambda I)^T) = \det(A^T - \lambda I)$

Esempio Data $A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix}$

$$K_1 = \{z \in \mathbb{C} \mid |z - 3| \leq 1\}$$

$$K_2 = \{z \in \mathbb{C} \mid |z - 3| \leq 2\}$$

$$K_3 = \{z \in \mathbb{C} \mid |z - 3| \leq 2\} = K_1$$

2.3 Fattorizzazione LU

Problema Dato $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, determinare $x \in \mathbb{R}^n \mid Ax = b$. Assunzione: A invertibile.

Quindi, convenzionalmente, A è la **matrice dei coefficienti**, b è il **termine noto** e x il **vettore delle incognite**.

x è **univocamente determinato** $x = A^{-1}b$ (rappresentazione teorica)

La **difficoltà è determinata dalla struttura di A** . Ci sono **alcuni casi più semplici**:

A matrice **diagonale**: $A = a_{ij}$ con $a_{ij} = 0$ se $i \neq j$

Quindi $A = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \\ 0 & & \ddots & 0 \\ 0 & & & d_n \end{bmatrix}$ A invertibile $\Leftrightarrow d_j \neq 0$ per $j = 1 \dots n$, questo perché $\det(A) = \prod_{i=1}^n d_i$

$$\begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \Leftrightarrow \begin{cases} d_1 x_1 = b_1 \\ \vdots \\ d_n x_n = b_n \end{cases} \Leftrightarrow \begin{cases} x_1 = \frac{b_1}{d_1} \\ \vdots \\ x_n = \frac{b_n}{d_n} \end{cases}$$

Calcolabile in $O(n)$ operazioni

```
for k = 1:n
    x(k) = b(k)/a(k, k);
end
```

A **triangolare superiore** = $\begin{bmatrix} n & n & n \\ 0 & n & n \\ 0 & 0 & n \end{bmatrix}$ o **triangolare inferiore** = $\begin{bmatrix} n & 0 & 0 \\ n & n & 0 \\ n & n & n \end{bmatrix}$

Triangolare superiore se $a_{ij} = 0$ per $i > j$

Triangolare inferiore se $a_{ij} = 0$ per $i < j$

A invertibile \Leftrightarrow gli elementi sulla diagonale sono $\neq 0$

$\det(A) = \prod_{i=1}^n a_{ii}$, quindi $\det(A) \neq 0 \Leftrightarrow a_{ii} \neq 0$ per $i = 1 \dots n$

Backward Substitution Usato per risolvere un sistema triangolare superiore

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ 0 & & a_{nn} \end{bmatrix} \text{ con } a_{ij} = 0 \text{ se } i > j \rightarrow Ax = b \Leftrightarrow \begin{cases} \sum_{j=1}^n a_{1j}x_j = b_1 \\ \sum_{j=2}^n a_{2j}x_j = b_2 \\ \vdots \\ a_{nn}x_n = b_n \end{cases}$$

$$\sum_{j=k}^n a_{kj}x_j = b_k$$

$x_{k+1} \dots x_n \rightarrow x_k$, quindi $a_{kk}x_k = b_k - \sum_{j=k+1}^n a_{kj}x_j$, da cui ricavo

$$x_k = \frac{1}{a_{kk}}(b_k - \sum_{j=k+1}^n a_{kj}x_j)$$

Di conseguenza k lo scorriamo "all'indietro": $k = n:-1:1$

```
x(n) = b(n)/a(n, n)
for k = n - 1:-1:1
    s = 0
    for j = k + 1:n
        s = s + a(k, j)*x(j)
    end
    x(k) = (b(k) - s)/a(k, k)
end
```

Costo computazionale

$$\sum_{k=1}^{n-1} (n-k) = (n-1) + (n-2) + \dots + 1 = \frac{n(n-1)}{2} = \frac{n^2}{2} + O(n)$$

Quindi ha un costo di $O(n^2)$ operazioni

Forward Substitution Usato per risolvere un sistema triangolare inferiore

$$A = \begin{bmatrix} a_{11} & & 0 \\ \vdots & \ddots & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \text{ con } a_{ij} = 0 \text{ se } i < j \rightarrow Ax = b \Leftrightarrow \begin{cases} a_{11}x_1 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j = b_n \end{cases}$$

Matrice generica Per una matrice generica $A \in \mathbb{R}^{n \times n}$, l'idea è di cercare di ridurre A in una forma più semplice, mantenendo l'equivalenza dei sistemi lineari.

$$Ax = b \rightarrow Fx = g$$

con sistemi equivalenti e F più semplice di A : ad esempio F triangolare.

Cominciamo domandandoci se A può essere riscritta come **prodotto di matrici triangolari**.

2.3.1 Fattorizzazione LU

$A = L \cdot U$ con L, U triangolari

$$Ax = b \Leftrightarrow L \cdot Ux = b \Leftrightarrow \begin{cases} Ly = b \\ Ux = y \end{cases}$$

Definizione $A \in \mathbb{R}^{n \times n}$ si dice **fattorizzabile LU** se $\exists L$ matrice triangolare inferiore con elementi = 1 sulla diagonale principale e U triangolare superiore tale che $A = L \cdot U$

$$\mathbf{Esempio} \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_2 \\ 0 & u_3 \end{bmatrix} = \begin{cases} 0 = u_1 \\ 1 = u_3 \\ 1 = l \cdot u_1 \\ 0 = l \cdot u_3 + u_2 \end{cases} \quad \dots \text{non ci sono soluzioni}$$

$\Rightarrow A$ non ammette fattorizzazione LU

Esempio $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_3 \\ 0 & u_2 \end{bmatrix} = \begin{cases} 0 = u_1 \\ 1 = u_3 \\ 0 = l \cdot u_1 \\ 0 = l \cdot u_3 + u_2 \end{cases} = \begin{cases} u_1 = 0 \\ u_3 = 1 \\ 0 = 0 \\ l + u_2 = 0 \\ u_2 = -l \end{cases} = \begin{bmatrix} 1 & 0 \\ l & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & -l \end{bmatrix}$

\Rightarrow infinite fattorizzazioni LU , perché $l \in \mathbb{R}$

Esempio $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_3 \\ 0 & u_2 \end{bmatrix} = \begin{cases} 1 = u_1 \\ 1 = u_3 \\ 1 = l \cdot u_1 \Rightarrow l = 1 \\ 1 = l \cdot u_3 + u_2 \Rightarrow u_2 = 0 \end{cases}$

\Rightarrow fattorizzazione LU **unica**

Teorema Esistenza ed unicità della fattorizzazione

Sia $A \in \mathbb{R}^{n \times n}$ e siano $A_k = A(1:k, 1:k)$ con $k = 1 \dots n$ le **sottomatrici principali di testa**.

Se A_1, \dots, A_{n-1} sono invertibili $\Rightarrow \exists!$ LU di A

Dim Per induzione sulla dimensione della matrice

$n = 1$, $A = [a] \Rightarrow a = 1 \cdot a$

$$* = \left[\begin{array}{c|c} A_{n-1} & z \\ \hline v^T & \alpha \end{array} \right] = \left[\begin{array}{c|c} L_{n-1} & 0 \\ \hline x^T & 1 \end{array} \right] \left[\begin{array}{c|c} U_{n-1} & y \\ \hline 0^T & \beta \end{array} \right] \Leftrightarrow \begin{cases} A_{n-1} = L_{n-1}U_{n-1} \\ z = L_{n-1}y \\ v^T = x^T U_{n-1} \\ \alpha = x^T y + \beta \end{cases} \Leftrightarrow \begin{cases} y = L_{n-1}^{-1}z \\ x^T = v^T U_{n-1}^{-1} \end{cases}$$

Con A_{n-1} matrice di ordine $n-1$

Le sue sottomatrici principali di testa di ordine $1 \dots n-2$ sono quelle di A e quindi, per ipotesi, invertibili

\Rightarrow per ipotesi $\exists!$ fattorizzazione LU di A_{n-1}

Per ipotesi, A_{n-1} invertibile, quindi

$$0 \neq \det(A_{n-1}) = \det(L_{n-1}U_{n-1}) =$$

Per Binet

$$= \det(L_{n-1})\det(U_{n-1}) = \det(U_{n-1})$$

Fine

Matrice Freccia (\searrow) Anche detta **arrow matrix** o **arrowhead matrix** $A = \begin{bmatrix} 1 & & & x_1 \\ & \ddots & & \vdots \\ & & 1 & \vdots \\ x_1 & \dots & \dots & x_n \end{bmatrix}$

2.3.2 Calcolo della fattorizzazione LU

Processo di eliminazione gaussiana

Strumento Per il calcolo della fattorizzazione LU utilizzeremo le **matrici elementari di Gauss**.

Matrice elementare di Gauss Una matrice $E \in \mathbb{R}^{n \times n}$ si dice **matrice elementare di Gauss** se $\exists k$ con $1 \leq k \leq n$ e $v \in \mathbb{R}^n$ con $v_1 = v_2 = \dots = v_k = 0$ tale che

$$E = I - v \cdot e_k^T$$

con e_k k -esimo vettore della base canonica, quindi lungo n e con tutti gli elementi $= 0$ tranne un 1 in k -esima posizione. Abbiamo quindi l'identità meno un vettore colonna per un vettore riga, quindi $v \cdot e_k^T$ **non è un prodotto scalare** ma è un **prodotto tra una matrice $n \times 1$ per una matrice $1 \times n$** , quindi la si esegue riga per colonna.

Esempi Con $n = 4$

$$k = 1 \quad E = \begin{bmatrix} 1 & & & \\ -v_2 & 1 & & \\ -v_3 & & 1 & \\ -v_4 & & & 1 \end{bmatrix} = I - \begin{bmatrix} 0 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

$$k = 2 \quad E = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -v_3 & 1 & \\ & -v_4 & & 1 \end{bmatrix} = I - \begin{bmatrix} 0 \\ 0 \\ v_3 \\ v_4 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$$

$$k = 3 \quad E = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -v_4 & 1 \end{bmatrix} = I - \begin{bmatrix} 0 \\ 0 \\ 0 \\ v_4 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$$

$$k = 4 \quad E = I$$

Quindi le **matrici elementari di Gauss** sono matrici **triangolari inferiori**, con **elementi diagonali = 1** ed altri elementi non nulli soltanto in una colonna sotto l'elemento diagonale.

Proprietà

1. $E = I - v \cdot e_k^T$ è **invertibile** e $E^{-1} = I + v \cdot e_k^T$ ed è **ancora elementare di Gauss**

Dim: $(I - v \cdot e_k^T) \cdot (I + v \cdot e_k^T) = I + \cancel{v e_k^T} - \cancel{v e_k^T} - v(e_k^T v) e_k^T = I$

2. Dato $x \in \mathbb{R}^n$ con $x_k \neq 0 \Rightarrow \exists E \in \mathbb{R}^{n \times n}$ elementare di Gauss tale che $E \cdot x = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

Dim: cerchiamo $E = I - v \cdot e_k^T$, quindi dobbiamo determinare v .

$$(I - v \cdot e_k^T)x = x - v(e_k^T \cdot x) = x - vx_k, \text{ quindi voglio che } x - vx_k = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\text{Questo significa, scambiando le componenti, che } \Leftrightarrow vx_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix}$$

$$\Leftrightarrow v_1 = \dots = v_k = 0 \wedge v_j x_k = x_j \text{ per } j = k+1..n$$

Questo significa che $v_j = \frac{x_j}{x_k}$ per $j = k+1..n$

3. Il **prodotto** Ex con E matrice elementare di Gauss **costa $O(n)$ operazioni**.

Questo perché $E = I - v \cdot e_k^T$, quindi $Ex = (I - v \cdot e_k^T)x = x - vx_k$

$$= \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} - x_k v_{k+1} \\ \vdots \\ x_n - x_k v_n \end{bmatrix} \quad \text{Quindi calcolo solamente le ultime } n - k \text{ componenti, ognuna richiedente un prodotto ed}$$

una sottrazione. Di conseguenza $n - k$ operazioni moltiplicative e $n - k$ operazioni additive $\Rightarrow O(n)$ operazioni aritmetiche.

Vediamo come si esegue il calcolo.

Calcolo Supponiamo la matrice A , che al tempo $t = 0$ (corrispondente alla matrice originale senza passaggi di calcolo effettuati) è

$$A^{(0)} = \begin{bmatrix} a_{11}^{(0)} & \dots & a_{1n}^{(0)} \\ \vdots & & \\ a_{n1}^{(0)} & \dots & a_{nn}^{(0)} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Primo passo, partiamo da una supposizione: $a_{11}^{(0)} \neq 0 \Rightarrow E_1 = I - v \cdot e_1^T$ tale che $E_1 \cdot \begin{bmatrix} a_{11}^{(0)} \\ \vdots \\ a_{n1}^{(0)} \end{bmatrix} = \begin{bmatrix} a_{11}^{(0)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

$v_j = \frac{a_{j1}^{(0)}}{a_{11}^{(0)}}$ per $j = 2..n$ ($v_1 = 0$ per costruzione), chiamati **moltiplicatori di Gauss**.

A questo punto posso calcolare $A^{(1)} = E_1 A^{(0)}$

$$A^{(1)} = \begin{bmatrix} 1 & & & \\ -v_2 & \ddots & & \\ \vdots & & \ddots & \\ -v_n & & & 1 \end{bmatrix} \cdot \begin{bmatrix} a_{11}^{(0)} & \dots & a_{1n}^{(0)} \\ \vdots & & \\ a_{n1}^{(0)} & \dots & a_{nn}^{(0)} \end{bmatrix} = \begin{bmatrix} a_{11}^{(1)} & \dots & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}$$

Gli elementi sulla prima riga non cambiano, $a_{1i}^{(0)} = a_{1i}^{(1)}$

$$a_{ij}^{(1)} = i\text{-esima riga di } E_1 \times j\text{-esima colonna di } A^{(0)} = \begin{bmatrix} -v_1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} a_{1j}^{(0)} \\ \vdots \\ a_{nj}^{(0)} \end{bmatrix}$$

$$= a_{ij}^{(0)} - v_1 a_{1j}^{(0)} = a_{ij}^{(0)} - \left(\frac{a_{i1}^{(0)}}{a_{11}^{(0)}} \right) a_{1j}^{(0)}$$

Quindi

$$a_{ij}^{(1)} = a_{ij}^{(0)} - \left(\frac{a_{i1}^{(0)}}{a_{11}^{(0)}} \right) \cdot a_{1j}^{(0)} \text{ per } i = 2..n, j = 2..n$$

Costo computazionale: $O(n)$ operazioni per calcolare i moltiplicatori, poi aggiornare $A^{(1)}$ con $O(n^2)$ operazioni (precisamente $(n-1)^2$ moltiplicazioni e $(n-1)^2$ addizioni)

Secondo passo, andiamo a mettere elementi nulli nella seconda colonna.

$$A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & \dots & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix} \text{ e suppongo } a_{22}^{(1)} \neq 0. \text{ Gli elementi } a_{kk}^{(k-1)} \text{ sono chiamati } \mathbf{pivot} \text{ perché sono alla}$$

basa della costruzione dei moltiplicatori.

$$\text{Posso determinare } E_2 \text{ tale che } E_2 \cdot \begin{bmatrix} a_{22}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{bmatrix} = \begin{bmatrix} a_{12}^{(1)} \\ a_{22}^{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \text{ quindi}$$

$$E_2 = I - v e_2^T = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -v_3 & \ddots & \\ & \vdots & & \ddots \\ & -v_n & & & 1 \end{bmatrix} \text{ e } v_j = \frac{a_{j2}^{(1)}}{a_{22}^{(1)}} \text{ per } j = 3..n \text{ e la costruzione si ripete.}$$

$$A^{(2)} = E_2 \cdot A^{(1)} = \begin{bmatrix} a_{11}^{(2)} & & \cdots & & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & & & \\ \vdots & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}$$

Proseguirò analogamente, supponendo $a_{33}^{(2)} \neq 0$, mettendo 0 nella terza colonna e così via.

In conclusione Se $a_{kk}^{(k-1)} \neq 0$ per $k = 1..n-1$, allora posso determinare $E_1 \dots E_{n-1}$ matrici elementari di Gauss tali che

$$E_{n-1} \dots E_1 A^{(0)} = A^{(n-1)} = U$$

Quindi U è matrice diagonale superiore per via della costruzione (E_i ha messo zeri nella i -esima colonna).

Quindi posso scrivere $E_{n-1} \dots E_1 A = U$, ma le matrici elementari di Gauss sono invertibili, quindi

$A = (E_{n-1}^{-1} \dots E_1^{-1}) \cdot U$ e se chiamo $(E_{n-1}^{-1} \dots E_1^{-1}) = L$ ottengo

$$A = L \cdot U$$

Questo strumento funziona con l'ipotesi di lavoro di avere gli elementi pivot non nulli, cioè $a_{kk}^{(k-1)} \neq 0$

Abbiamo anche l'ipotesi di esistenza ed unicità, cioè $\det A_k \neq 0$ per $k = 1..n-1$

Concludiamo che $A_k \neq 0$ per $k = 1..n-1 \Rightarrow a_{kk}^{(k-1)} \neq 0$

2.3.3 Predominanza Diagonale

Predominanza Diagonale $A \in \mathbb{R}^{n \times n}$ è **predominante diagonale per righe** se l'elemento della diagonale *pesa di più* della somma degli elementi della stessa riga in modulo, cioè

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$$

Ci riferiamo a questa quando diciamo "predominante diagonale" senza specificare.

Ad esempio $\begin{bmatrix} -3 & 1 \\ 1 & -2 \end{bmatrix}$ lo è, mentre $\begin{bmatrix} -3 & 1 \\ 1 & -1 \end{bmatrix}$ non lo è ($i = 2$ non soddisfa)

Predominanza Diagonale per colonne $A \in \mathbb{R}^{n \times n}$ è **predominante diagonale per colonne** se l'elemento della diagonale *pesa di più* della somma degli elementi della stessa colonna in modulo, cioè

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ji}|$$

La **predominanza diagonale** è importante per il seguente teorema.

Teorema Se $A \in \mathbb{R}^{n \times n}$ è **predominante diagonale** $\Rightarrow A$ è **invertibile**

Dim Segue dal teorema di Gerschgorin, che dice λ autovalore di $A \Rightarrow \lambda \in \bigcup_{i=1}^n K_i$.

Mostriamo che $0 \notin \bigcup_{i=1}^n K_i \Rightarrow 0$ non è autovalore $\Rightarrow A$ è invertibile.

$K_i = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|\}$, $0 \in ? K_i$

$|0 - a_{ii}| = |a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$ per predominanza diagonale, $\Rightarrow 0 \notin K_i$

Vale $\forall i \Rightarrow 0 \notin \bigcup_{i=1}^n K_i$ quindi A è invertibile.

La dimostrazione per colonne è analoga, utilizzando il teorema di Gerschgorin per colonne.

Corollario $A \in \mathbb{R}^{n \times n}$ predominante diagonale $\Rightarrow \exists!$ la fattorizzazione LU di A

Dim Se A è predominante diagonale $\Rightarrow A_k$ è predominante diagonale per $k = 1..n-1$

$\Rightarrow A_k$ è invertibile \Rightarrow per il teorema di esistenza ed unicità, A ammette unica la fattorizzazione LU

La **predominanza diagonale** è una **condizione sufficiente** per: **invertibilità ed esistenza ed unicità di LU**
In generale, il metodo di eliminazione Gaussiana **non preserva** la sparsità: si riempiono le matrici.

2.3.4 Eliminazione Gaussiana con pivoting

Vogliamo migliorare la **stabilità** dell'eliminazione gaussiana e **trovare approcci differenti** applicabili a matrici sparse.

I due problemi del **calcolare la fattorizzazione LU** e **risolvere il sistema lineare**, entrambi richiedenti il **processo di eliminazione gaussiana**, sono quindi profondamente **collegati**.

Esempio Partiamo da un esempio: $A \in \mathbb{R}^{n \times n}$ invertibile.

$$A = A^{(0)} = \begin{bmatrix} a_{11}^{(0)} & \dots & a_{1n}^{(0)} \\ \vdots & & \\ a_{n1}^{(0)} & \dots & a_{nn}^{(0)} \end{bmatrix}$$

La matrice è invertibile, quindi se guardo la prima colonna posso dire che esiste un elemento $\neq 0$ perché se la prima colonna avesse tutti elementi nulli allora la matrice sarebbe singolare. In generale ce ne possono ovviamente essere tanti di elementi $\neq 0$, per motivi di **stabilità numerica** scelgo l'**elemento di modulo massimo della prima colonna**. Quindi suppongo che l'elemento di modulo massimo sia sulla riga j (in caso di più elementi di modulo massimo solitamente si sceglie il primo j trovato)

$$\exists a_{j1}^{(0)} \neq 0 \text{ con } j : |a_{j1}^{(0)}| = \max_{(k=1:n)} |a_{k1}^{(0)}|$$

A questo punto, **scambio la j -esima riga con la prima**

$$A^{(0)} \longrightarrow A^{(\frac{1}{2})} = \begin{bmatrix} a_{j1}^{(0)} & \dots & a_{jn}^{(0)} \\ \vdots & & \\ a_{11}^{(0)} & \dots & a_{1n}^{(0)} \\ \vdots & & \end{bmatrix} \quad \begin{array}{l} I \text{ riga} \\ \\ j\text{-esima riga} \end{array}$$

Dal punto di vista matriciale, scambiare le righe corrisponde a moltiplicare la matrice per una matrice P_1 composta sostanzialmente dalla matrice identità con la prima e la j -esima colonna scambiate.

$$A^{(\frac{1}{2})} = \begin{bmatrix} & & & & 1 & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ 1 & & & & & & \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{bmatrix} \cdot A^{(0)}$$

P_1 si chiama **matrice di permutazione**. Hanno la proprietà che $P^{-1} = P^T$, cioè che l'inversa coincide con la trasposta.

Una volta eseguito lo scambio di riga, possiamo procedere con un passo di eliminazione su $A^{(\frac{1}{2})} = p_1 \cdot A^{(0)}$ ed ottenere $A^{(1)} = E_1 \cdot P_1 \cdot A^{(0)}$ utilizzando $a_{j1}^{(0)}$ portato in posizione 1,1 come **pivot**. Ottengo

$$A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \dots \\ \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}$$

Per continuare il processo di eliminazione dovrò lavorare su $\begin{bmatrix} a_{22}^{(1)} & \dots \\ \vdots & \vdots \\ a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}$ e quindi guardare la seconda colonna

di $A^{(1)}$ sulla quale posso dire che $\exists j \geq 2 : a_{j2}^{(1)} \neq 0$ quindi che c'è almeno un elemento $\neq 0$ sulla seconda colonna. Posso dirlo perché se tutti fossero $= 0$, allora la prima e la seconda colonna sarebbero **linearmente dipendenti**, quindi $A^{(1)}$ sarebbe **singolare**. Ma $A^{(1)}$ singolare $\Rightarrow A^{(0)}$ singolare $\Leftrightarrow A$ singolare, ma siamo partiti da una matrice invertibile, quindi è **assurdo**. Di conseguenza, la seconda colonna non può avere elementi tutti nulli.

Quindi determino

$$|a_{j2}^{(1)}| = \max_{k=2:n} |a_{k2}^{(1)}|$$

Dopodiché scambio la seconda e la j -esima riga...

Il processo **termina** perché a **nessun punto del processo** potrà avere tutti elementi nulli nella parte attiva della colonna considerata.

In termini matriciali, il processo diventa

$$E_{n-1}P_{n-1}E_{n-2}\dots E_1P_1A^{(0)} = U$$

La differenza sta che in ogni passo multiplico per una coppia **matrice di eliminazione** e **matrice di permutazione**. Porto tutte le matrici da sinistra di $A^{(0)}$ a destra e ottengo

$$A^{(0)} = \underline{P_1^T E_1^{-1} \dots P_{n-1}^T E_{n-1}^{-1}} U$$

Con $P_1^T E_1^{-1} \dots P_{n-1}^T E_{n-1}^{-1} = L$, quindi

$$A^{(0)} = L \cdot U$$

Ma in generale **L non è triangolare**. MATLAB ci dice che è *psicologicamente triangolare*, cioè che si crede essere triangolare inferiore ma non lo è, perché è **infestata dalle matrici di permutazione** P_i . Quindi è una **matrice psicologicamente triangolare perché è permutazione di una matrice triangolare**.

MATLAB Il comando `lu` di MATLAB quindi restituisce questa L e questa U

$$[L, U] = \text{lu}(A)$$

Con U triangolare superiore, ed L permutazione di una matrice triangolare inferiore.

Pivoting La nostra tecnica di **pivoting** (o **partial pivoting** perché il pivot è cercato solamente in una colonna) quindi ha le seguenti **proprietà**:

1. Rende il **processo applicabile ad ogni matrice invertibile**.
Quindi ogni matrice invertibile ha una fattorizzazione *tipo*-LU
2. **Migliora la stabilità dell'algoritmo di eliminazione gaussiana**, perché scambiare le righe e rendere il pivot più grande migliora la stabilità dell'algoritmo
3. **Non altera il costo computazionale**
4. **Può aumentare il fill-in**

2.4 Metodi Iterativi

Idea Ricerca di un metodo alternativo per **trattare matrici sparse**.

Quindi il problema è risolvere $Ax = b$ con A invertibile e sparsa. A sparsa = molti elementi sono nulli.

Vogliamo **sfruttare questa proprietà**, non sfruttata dal processo di eliminazione. Un esempio di matrice sparsa è

$$A = \begin{bmatrix} 1 & x & \dots & x \\ x & \ddots & & \\ \vdots & & \ddots & \\ x & & & 1 \end{bmatrix}$$

Sparsità Quantitativamente, la sparsità si potrebbe definire nel seguente modo $\text{nnz}(A) \ll n^2$, dove con $\text{nnz}(A)$ in MATLAB si intende il numero di elementi $\neq 0$ della matrice A .

Quindi si accettano valori come $\text{nnz}(A) = \begin{cases} O(n) \\ O(n \cdot \log p_n) \end{cases}$ ma principalmente $\text{nnz}(A) = O(n)$

Il metodo di eliminazione gaussiana è indicato come **metodo diretto**, perché **in un numero finito di passi determina la soluzione del sistema**.

I **metodi iterativi** adottano un **approccio differente**, cioè di **costruire una successione $\{x^{(k)}\}$ di vettori tali che $x^{(k)} \rightarrow x$ ($x^{(k)}$ converge a x) con x soluzione del sistema lineare**.

Siccome la successione è potenzialmente infinita, ho bisogno di un **criterio di arresto**.

Def La successione di vettori $\{x^{(k)}\}_{k \in \mathbb{N}}$ con $x^k \in \mathbb{R}^n$ **converge a x** (cioè $\lim_{k \rightarrow +\infty} x^{(k)} = x$) solamente se

$$\lim_{k \rightarrow +\infty} x^{(k)} = x \Leftrightarrow \lim_{k \rightarrow +\infty} \|x^{(k)} - x\| = 0$$

Per l'equivalenza topologica delle norme, vale per qualsiasi norma. Posso scegliere la norma infinito

$$\|x^{(k)} - x\|_{\infty} = \max_{j=1:n} |x_j^{(k)} - x_j|$$

e se $\max_{j=1:n} |x_j^{(k)} - x_j| \rightarrow 0$ per $k \rightarrow +\infty$ allora $|x_j^{(k)} - x_j| \rightarrow 0$ per $k \rightarrow +\infty$ per $j = 1 \dots n$

Quindi una successione di vettori tende ad un vettore se **componente per componente le successioni delle componenti tendono alle componenti del vettore obiettivo**.

2.4.1 Come nascono

Un'idea è partire dalla relazione $Ax = b$ e provo a scrivere A come differenza di due matrici $A = M - N$ con M invertibile. Quindi potrei dire

$$Ax = b \Leftrightarrow (M - N)x = b \Leftrightarrow Mx = Nx + b \Leftrightarrow x = M^{-1}Nx + M^{-1}b$$

e quindi $x = Px + q$ con $P = M^{-1}N$ e $q = M^{-1}b$.

Dunque

$$Ax = b \Leftrightarrow \begin{cases} x = Px + q \\ P = M^{-1}N \\ q = M^{-1}b \\ A = M - N \end{cases}$$

Perché Questo perché per risolvere $x = Px + q$ posso usare un metodo composto in questo modo $\begin{cases} x^{(0)} \in \mathbb{R}^n \\ x^{(k+1)} = Px^{(k)} + q \end{cases}$

Teorema Se la successione generata in questo modo ha $\lim_{k \rightarrow +\infty} x^{(k)} = x \in \mathbb{R}^n$ allora $x = Px + q$ e quindi è soluzione del sistema lineare

Dim Sappiamo per ipotesi che $x^{(k)} \rightarrow x$ e sappiamo sempre per ipotesi che $x^{(k+1)} = Px^{(k)} + q$. Da queste ipotesi ottengo

$$x = \lim_{k \rightarrow +\infty} x^{(k+1)} = \lim_{k \rightarrow +\infty} Px^{(k)} + q = Px + q$$

2.4.2 Implementazione di un metodo iterativo

Un metodo iterativo, quindi, parte da un **valore iniziale** (vettore o matrice), costruiamo una **sequenza di vettori/matrici** tendenzialmente infinita a cui, quindi, forniamo un criterio di arresto. Vediamo l'implementazione di un metodo iterativo particolare.

$A^{(0)} = B$ matrice **simmetrica** invertibile e **generiamo una sequenza di matrici**.

$$A^{(k)} = \frac{A^{(k-1)} + (A^{(k-1)})^{-1}}{2} \text{ con } k \geq 1$$

Criterio di arresto Ho due criteri per vedere se sono sufficientemente vicino ad una soluzione:

$$\frac{\|A^{(k)} - A^{(k-1)}\|_{\infty}}{\|A^{(k)}\|_{\infty}} \leq \text{tolleranza}$$

Può non andare bene perché non ho garanzie, in generale, che il metodo converga e che quindi la differenza diventi, ad un certo punto, inferiore alla tolleranza indicata.

$$k \geq \text{MAX_ITER}$$

2.4.3 Convergenza

Vediamo nel dettaglio cosa intendiamo per convergenza dei metodi iterativi. In generale, un metodo iterativo converge quando genera una sequenza convergente di matrici. Il nostro problema computazionale è calcolare $Ax = b$, lo riformuliamo con $A = M - N$ con M invertibile. Quindi riscriviamo come $(M - N)x = b \Leftrightarrow Mx = Nx + b \Leftrightarrow x = M^{-1}Nx + M^{-1}b$ e indico $M^{-1}N = P$ e $M^{-1}b = q$.

Ho ricondotto il problema $Ax = b$ al problema di trovare un vettore $x \mid x = Px + q$. Possiamo scegliere un vettore iniziale $x^{(0)} \in \mathbb{R}^n$ e costruiamo una successione $x^{(k+1)} = Px^{(k)} + q$ cioè

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ x^{(k+1)} = Px^{(k)} + q \end{cases} \quad \text{che solitamente lo si implementa come} \quad \begin{cases} x^{(0)} \in \mathbb{R}^n \\ Mx^{(k+1)} = Nx^{(k)} + b \end{cases}$$

Fissata A , la **convergenza dipende in generale da come seleziono M , N e dalla scelta del vettore iniziale**. Vogliamo **eliminare questa dipendenza dalla scelta del vettore iniziale**. Questo perché non possiamo considerare metodi che convergono per determinati vettori iniziali e divergono per altri, perché a quel punto la scelta del vettore iniziale diventa un problema di pari complessità a quello della risoluzione di $Ax = b$

Convergente Un metodo iterativo $x^{(k+1)} = Px^{(k)} + q$ con $k \geq 0$ per risolvere $Ax = b$ con $P = M^{-1}N$ e $q = M^{-1}b$, e $A = M - N$, si dice **convergente** se $\forall x^{(0)} \in \mathbb{R}^n \Rightarrow \lim x^{(k)} = x$ con x soluzione del sistema, cioè la **successione generata converge alla soluzione del sistema lineare**.

Partiamo col cercare delle condizioni di convergenza. Sappiamo che il metodo iterativo genera delle successioni con $x^{(k+1)} = Px^{(k)} + q$ e sappiamo che la soluzione del sistema lineare soddisfa $x = Px + q$, di conseguenza

$$x^{(k+1)} - x = P(x^{(k)} - x)$$

Chiamiamo $e^{(k)} = x^{(k)} - x$, $e^{(k)}$ misura una sorta d'**errore** cioè quanto dista la k -esima iterata dalla soluzione del sistema. La relazione precedente diventa

$$e^{(k+1)} = Pe^{(k)} \quad k \geq 0$$

Teorema Il metodo $x^{(k+1)} = Px^{(k)} + q$ è **convergente** se \exists una norma matriciale indotta dalla norma vettoriale $\|\cdot\|$ tale che $\|P\| < 1$

Dim $e^{(k+1)} = Pe^{(k)} \Leftrightarrow \|e^{(k+1)}\| = \|Pe^{(k)}\| \leq \|P\| \cdot \|e^{(k)}\| \leq \|P\|^2 \cdot \|e^{(k-1)}\| \leq \dots \leq \|P\|^{k+1} \cdot \|e^{(0)}\|$
 $0 \leq \|e^{(k+1)}\| \leq \|P\|^{k+1} \cdot \|e^{(0)}\|$, che per $k \rightarrow \infty$

$$0 \rightarrow 0$$

$$\|P\|^{k+1} \cdot \|e^{(0)}\| \rightarrow 0 \text{ perché } \|P\| < 1 \text{ per ipotesi} \Rightarrow \|P\|^{k+1} \rightarrow 0 \text{ per } k \rightarrow \infty$$

$$\Rightarrow \|e^{(k+1)}\| \rightarrow 0 \text{ per il teorema del confronto.}$$

Quindi se trovo una norma come sopra, posso dire che il metodo è convergente: il teorema fornisce una **condizione sufficiente**.

Adesso cerchiamo una condizione necessaria.

Teorema Se il metodo $x^{(k+1)} = Px^{(k)} + q$ è convergente, allora $\rho(P) < 1$
 $\rho(P)$ è il **raggio spettrale** di P , il **modulo dell'autovalore di modulo massimo** di P .
 Quindi il metodo **non converge** se $\rho(P) \geq 1$, mentre **può convergere** se $\rho(P) < 1$.

Dim Se è convergente, allora la successione converge $\forall x^{(0)}$ e quindi $\forall e^{(0)}$ ho $e^{(k)} \rightarrow 0$
 Prendo $e^{(0)} = v \mid Pv = \lambda v$ con $|\lambda| = \rho(P)$ quindi un autovettore relativo all'autovalore di modulo massimo.

$$e^{(k+1)} = Pe^{(k)} = \dots = P^{k+1}e^{(0)} = P^{k+1}v$$

$$P^{k+1}v = \lambda^{k+1}v = \lambda^{k+1}e^{(0)}$$

$$\Rightarrow \|e^{(k+1)}\| = |\lambda|^{k+1} \cdot \|e^{(0)}\| \neq 0$$

$$\|e^{(k+1)}\| \rightarrow 0 \Leftrightarrow |\lambda| < 1$$

Si può dimostrare che è una condizione anche sufficiente, quindi **necessaria e sufficiente**.

Teorema Il metodo $x^{(k+1)} = Px^{(k)} + q$ è convergente $\Leftrightarrow \rho(P) < 1$.

Esempio

$$x^{(k+1)} = Px^{(k)} + q$$

$$P = \begin{bmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix}$$

Convergente?

$$\|P\|_1 = \max \left\{ \frac{1}{2} + \frac{1}{3}, \frac{2}{3} + \frac{1}{4} \right\} = \max \left\{ \frac{5}{6}, \frac{11}{12} \right\} = \frac{11}{12} < 1$$

$$\|P\|_\infty = \frac{1}{2} + \frac{2}{3} = \frac{3+4}{6} = \frac{7}{6} > 1$$

$$\|P\|_\infty > 1, \text{ ma } \|P\|_1 < 1$$

Quindi il metodo è convergente

2.4.4 Metodi

Nella pratica, studiare le proprietà di P matrice d'iterazione può essere complicato, quindi necessitiamo di risultati che ci dicano se un metodo converge o meno in base alle proprietà soddisfatte da P .

Problema Il nostro problema è sempre la risoluzione di un sistema lineare $Ax = b$, sempre con l'idea classica dei metodi iterativi cioè di partizionare la matrice $A = M - N$ con M invertibile.

Suddivisione

In entrambi i metodi scriviamo A come **somma di 3 matrici** $A = L + U + D$ con

$L = \text{tril}(A, -1)$ cioè la **parte strettamente triangolare inferiore** di A , completata con 0 sulla diagonale principale e nella parte superiore.

$U = \text{tril}(A, 1)$ cioè la **parte strettamente triangolare superiore** di A , completata con 0 sulla diagonale principale e nella parte inferiore.

$D = \text{diag}(A)$ cioè la **diagonale** di A

Un esempio

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \rightarrow L = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{bmatrix} \quad U = \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix} \Rightarrow A = L + U + D$$

Nel metodo Jacobi Vengono scelte $M = D$
 $N = -L - U$

Applicabilità (un metodo è applicabile se posso costruire la successione).

Siccome in Jacobi M è diagonale, Jacobi è applicabile $\Leftrightarrow M$ è invertibile $\Leftrightarrow a_{ii} \neq 0$ con $i = 1 \dots n$ cioè che **tutti gli elementi sulla diagonale principale siano $\neq 0$**

Se il metodo è applicabile, posso generare la sequenza.

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ Mx^{(k+1)} = Nx^{(k)} + b \end{cases}$$

$$\begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{bmatrix} = Nx^{(k)} + b$$

$$a_{jj}x_j^{(k+1)} = b_j - \sum_{l=1, l \neq j}^n (a_{jl} \cdot x_l^{(k)}) \text{ per } j = 1 \dots n$$

Iterazione di Jacobi

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{l=1, l \neq j}^n (a_{jl} \cdot x_l^{(k)}) \right)$$

Per $j = 1 \dots n$ e $k \geq 0$

Molto semplice: richiede due vettori per l'implementazione $x^{(k)}$ e $x^{(k+1)}$

Il numero di operazioni moltiplicative è nell'ordine del numero di operazioni additive, quindi considero le moltiplicative: sono fatte all'interno della sommatoria tutte le volte che incontro un elemento di A che sia $\neq 0$. Quindi **un'iterazione di Jacobi costa $\text{nnz}(A)$** , dove $\text{nnz}(A)$ è il numero degli elementi di A che sono $\neq 0$.

Quindi, nel caso di **matrici sparse, un passo del metodo di Jacobi costa poco**.

Metodo di Gauss-Seidel

Il metodo di Gauss-Seidel si sviluppa a partire dall'iterazione di Jacobi

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{l=1, l \neq j}^n (a_{jl} \cdot x_l^{(k)}) \right)$$

La prima osservazione che possiamo fare è che possiamo spezzare la sommatoria, sommando da 1 a $j-1$ e poi da $j+1$ a n , sempre saltando j .

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{l=1}^{j-1} (a_{jl} \cdot x_l^{(k)}) - \sum_{l=j+1}^n (a_{jl} \cdot x_l^{(k)}) \right)$$

Ma quando calcolo $x_j^{(k+1)}$ ho già calcolato $x_1^{(k+1)} \dots x_{j-1}^{(k+1)}$, quindi posso sostituire $x_l^{(k)}$ con $x_l^{(k+1)}$

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{l=1}^{j-1} (a_{jl} \cdot x_l^{(k+1)}) - \sum_{l=j+1}^n (a_{jl} \cdot x_l^{(k)}) \right)$$

Sostanzialmente il costo computazionale non è cambiato. Ma dal punto di vista pratico, utilizzare l'informazione più "fresca" disponibile, cioè la x al tempo $k+1$, ha un impatto sulla convergenza: **Gauss-Seidel converge più rapidamente**, cioè in un numero minore di passi.

Però, però, la componente parallela essendo un **metodo intrinsecamente sequenziale**. Gauss-Seidel, però, permette di mantenere un solo vettore in memoria poiché posso "sovrascrivere" le componenti calcolate passo dopo passo.

Confronto

In **Gauss-Seidel** il partizionamento è $\begin{matrix} M = L + D \\ N = -U \end{matrix}$, che quindi prende la parte triangolare inferiore di A , mentre

Jacobi partiziona $\begin{matrix} M = D \\ N = -L - U \end{matrix}$ quindi prendendo la parte diagonale di A .

Applicabilità La stessa, cioè $\Leftrightarrow a_{ii} \neq 0$ per $i = 1 \dots n$

Calcolo parallelo Meglio Jacobi

Convergenza In generale nella pratica è meglio Gauss-Seidel

Implementazione Gauss-Seidel può richiedere un solo vettore

Costo computazionale Medesimo e dipende dalla sparsità della matrice: $O(\text{nnz}(A))$

2.4.5 Teorema

Sia $A \in \mathbb{R}^{n \times n}$ **predominante diagonale** $\Rightarrow A$ è invertibile \wedge Jacobi e Gauss-Seidel sono applicabili e convergenti.

Dim A predominante diagonale $\Rightarrow A$ invertibile **OK**

A predominante diagonale $\Leftrightarrow |a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$ per $i = 1 \dots n$

$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}| \geq 0$ per $i = 1 \dots n \Rightarrow |a_{ii}| > 0 \Rightarrow a_{ii} \neq 0$ per $i = 1 \dots n$ quindi i metodi sono applicabili **OK**

La convergenza dipende dalla matrice d'iterazione P , più precisamente dagli autovalori di P . Vogliamo dimostrare che gli autovalori di P hanno tutti modulo < 1 , perché quindi $\rho(P) < 1 \Rightarrow$ convergente.

Se cerchiamo per quali valori il determinante è 0 allora il verso del calcolo non cambia perché cambierebbe solo il segno, perché $\det(A) = -\det(-A)$

$$\det(\lambda I - P) = \det(\lambda I - M^{-1}N) = \det(\lambda M^{-1}M - M^{-1}N) = \det(\lambda M^{-1}(\lambda M - N)) = \det(M^{-1}) \cdot \det(\lambda M - N)$$

E dato che M è invertibile, $\det(M^{-1}) \neq 0$, quindi scopriamo che

$$\det(\lambda I - P) = 0 \Leftrightarrow \det(\lambda M - N) = 0$$

Quindi λ autovalore di $P \Leftrightarrow \det(\lambda M - N) = 0$ cioè se $\lambda M - N$ è singolare

Per quali $\lambda \in \mathbb{C}$ $\lambda M - N$ è una matrice singolare? Se $|\lambda| \geq 1$, mostriamo che $\lambda M - N$ è predominante diagonale e quindi invertibile.

Consideriamo il metodo di Gauss-Seidel $M = L + D$, $N = -U$ (per Jacobi c'è una dimostrazione analoga)

$$\lambda M - N = \lambda(L + D) + U$$

Quindi l'elemento sulla diagonale principale è λa_{ii} , quindi la condizione di predominanza diagonale (ricordiamo che se non è specificato si sottintende predominanza diagonale per righe) è la seguente.

$$|\lambda a_{ii}| > \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^n |\lambda a_{ij}|$$

Sappiamo che la matrice è predominante diagonale, cioè che

$$|a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^n |a_{ij}|$$

In \mathbb{C} non c'è un ordinamento, quindi non ci sono le disequazioni, e $\lambda \in \mathbb{C}$ ma $|\lambda| \in \mathbb{R}$ quindi posso moltiplicare ambo i membri per $|\lambda| > 0$

$$\begin{aligned} |\lambda| |a_{ii}| &> |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + |\lambda| \sum_{j=i+1}^n |a_{ij}| \\ |\lambda a_{ii}| &> \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^n |\lambda a_{ij}| \end{aligned}$$

Nella seconda sommatoria lo abbiamo lasciato fuori perché, siccome $|\lambda| \geq 1$ allora vale la seguente disequazione

$$\sum_{j=1}^{i-1} |\lambda a_{ij}| + |\lambda| \sum_{j=i+1}^n |a_{ij}| \geq \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^n |\lambda a_{ij}|$$

Dimostrando che

$$|\lambda a_{ii}| > \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^n |\lambda a_{ij}|$$

2.4.6 Esempi

$$A = \begin{bmatrix} 1 & & & \alpha \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}, n \geq 2$$

1. Per quali valori di α , A è predominante diagonale?
2. Per quali valori di α il metodo di G-S è convergente?
3. Per quali valori di α il metodo di Jacobi è convergente?

1. Ricordiamo la condizione di predominanza diagonale

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}| \text{ per } i = 1 \dots n$$

Vediamo per righe

$$i = 1 \rightarrow 1 > |\alpha| \Leftrightarrow |\alpha| < 1$$

$$i = 2 \dots n \rightarrow 1 > |-1| = 1 \text{ impossibile}$$

Quindi A **non è mai predominante diagonale**. Quindi la risposta è **per nessun valore di α** .

$x_n \neq 0$ perché altrimenti tutte le componenti sarebbero $= 0$

Dividendo per x_n , ottengo che $\lambda^n = -\alpha$ quindi gli autovalori soddisfano questa relazione.

Questo significa $|\lambda^n| = |-\alpha| = |\alpha|$, quindi $|\lambda|^n = |\alpha| \Leftrightarrow |\lambda| = \sqrt[n]{|\alpha|}$, quindi **tutti gli autovalori della matrice J hanno lo stesso modulo**.

Quindi la condizione per la convergenza è $\sqrt[n]{|\alpha|} < 1 \Leftrightarrow |\alpha| < 1$, che in questo caso è la stessa di G-S.

Notare che $\rho(J) = \sqrt[n]{|\alpha|}$, quindi $\neq \rho(G) = |\alpha|$. Quando il metodo è convergente su questa matrice $0 \leq \rho(G) < \rho(J) < 1$, quindi su questa matrice **G-S converge più rapidamente di Jacobi**.

Esempio

$$A = \begin{bmatrix} 1 & & & \theta \\ & \ddots & & \vdots \\ & & \ddots & \theta \\ \theta & \dots & \theta & 1 \end{bmatrix}$$

$$A = M - N$$

$$M = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ \theta & \dots & \theta & 1 \end{bmatrix}$$

$$N = \begin{bmatrix} & & & -\theta \\ & & & \vdots \\ & 0 & & -\theta \\ 0 & \dots & \dots & 0 \end{bmatrix}$$

Convergente? $P = M^{-1}N = M^{-1} [0 \dots |0|v] = [0 \dots |0|M^{-1}v]$ con $M^{-1}v = x \Leftrightarrow Mx = v$

$$\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ \theta & \dots & \theta & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} -\theta \\ \vdots \\ -\theta \\ 0 \end{bmatrix} \Leftrightarrow \begin{cases} x_1 = x_2 = \dots = x_{n-1} = -\theta \\ \theta x_1 + \theta x_2 + \dots + \theta x_{n-1} + x_n = 0 \\ \Leftrightarrow x_n = \theta^2 + \theta^2 + \dots + \theta^2 = (n-1)\theta^2 \end{cases}$$

$$P = \begin{bmatrix} & -\theta & \\ & \vdots & \\ & -\theta & \\ (n-1)\theta^2 & & \end{bmatrix}$$

$\rho(P) = ?$ P è triangolare superiore, quindi i suoi autovalori sono sulla diagonale principale:

$$\lambda = 0$$

$$\lambda = (n-1)\theta^2$$

$\Rightarrow \rho(P) = (n-1)\theta^2$, di conseguenza è **convergente** $\Leftrightarrow (n-1)\theta^2 < 1 \Leftrightarrow \theta^2 < \frac{1}{n-1}$ quindi se

$$-\frac{1}{\sqrt{n-1}} < \theta < \frac{1}{\sqrt{n-1}}$$

$$\|P\|_\infty = \max\{|\theta|, (n-1)\theta^2\} < 1$$

Esempio

$$A = \begin{bmatrix} n & -1 & \dots & -1 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \dots & -1 & n \end{bmatrix}$$

Studiare la convergenza di Jacobi e G-S.

La somma sugli elementi di ogni riga in valore assoluto, esclusi gli elementi della diagonale principale è $= n-1$. Quindi

$$n > n-1 \Rightarrow A \text{ è predominante diagonale}$$

$\Rightarrow J$ e G-S convergono.

Esempio

$$A = \begin{bmatrix} 1 & & & \theta \\ & \ddots & & \vdots \\ & & \ddots & \theta \\ \theta & \dots & \theta & 1 \end{bmatrix} \quad M = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \quad N = \begin{bmatrix} & & & -\theta \\ & 0 & & \vdots \\ -\theta & \dots & -\theta & 0 \end{bmatrix}$$

$P = M^{-1}N = N$, quindi condizione sufficiente su θ affinché il metodo sia convergente è

$$\|P\|_{\infty} = \|P\|_1 < 1 \Leftrightarrow (n-1)|\theta| < 1 \Leftrightarrow |\theta| < \frac{1}{n-1}$$

$$-\frac{1}{n-1} < \theta < \frac{1}{n-1}$$

Condizione necessaria e sufficiente è $\rho(P) < 1$, $\rho(P) = ?$

$$\det \left(\begin{bmatrix} \lambda & & & \theta \\ & \ddots & & \vdots \\ & & \ddots & \theta \\ \theta & \dots & \theta & \lambda \end{bmatrix} \right) = ?, \text{ notiamo che } \forall \lambda \neq 0 \text{ la matrice ammette } LU$$

$$\begin{bmatrix} \lambda & & & \theta \\ & \ddots & & \vdots \\ & & \ddots & \theta \\ \theta & \dots & \theta & \lambda \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ \frac{\theta}{\lambda} & \dots & \frac{\theta}{\lambda} & 1 \end{bmatrix} \cdot \begin{bmatrix} \lambda & & & \theta \\ & \ddots & & \vdots \\ & & \lambda & \theta \\ & & & x \end{bmatrix} \quad \text{con} \quad \begin{aligned} x + (n-1)\frac{\theta^2}{\lambda} &= \lambda \Leftrightarrow \\ \Leftrightarrow x &= \lambda - (n-1)\frac{\theta^2}{\lambda} \end{aligned}$$

Di conseguenza $\det(\lambda I - P) = \det(U) \cdot \det(L)$ ma $\det(L) = 1$, quindi

$$\det(\lambda I - P) = \lambda^{n-1} \cdot \left[\lambda - \frac{(n-1)\theta^2}{\lambda} \right] = \lambda^n - (n-1)\theta^2\lambda^{n-2} = \lambda^{n-2} \cdot (\lambda^2 - (n-1)\theta^2) = \begin{aligned} &\lambda = 0 \\ &\lambda = \pm\sqrt{n-1} \cdot \theta \end{aligned}$$

Di conseguenza $\rho(P) = \sqrt{n-1} \cdot |\theta| < 1 \Leftrightarrow |\theta| < \frac{1}{\sqrt{n-1}}$

Capitolo 3

Equazioni Non Lineari

Vogliamo adesso analizzare i metodi per la risoluzione di equazioni non lineari.

Una prima importante **osservazione** è la seguente: **tutti i metodi** che si trovano in letteratura **per la risoluzione di equazioni non lineari sono metodi iterativi. Non abbiamo metodi diretti.**

Metodi iterativi significa che generano delle successioni che, sotto opportune ipotesi, convergono alla soluzione α dell'equazione non lineare.

Quali metodi Possiamo distinguerli in un metodo (**metodo di bisezione**) ed una classe di metodi (**metodi di iterazione funzionale**). Solitamente vengono usati in combinazione tra loro.

Tra i loro **elementi distintivi** ci sono la **velocità di convergenza**, cioè quante iterate vanno calcolate per avere una buona approssimazione della soluzione, e la **regolarità della funzione** (continua, derivabile...).

Sono elementi distintivi perché in generale **i metodi di iterazione funzionale** (es: metodo delle tangenti, o di Newton) **al prezzo di una richiesta di maggiore regolarità della funzione, presenteranno migliori proprietà di convergenza rispetto al metodo di bisezione.**

Il **metodo di bisezione** richiede poca regolarità della funzione ma fornisce una **convergenza generalmente lenta**. I metodi di iterazione funzionale richiedono maggiore regolarità ma le successioni che producono convergono più rapidamente.

3.1 Metodo di Bisezione

Metodo antico, usato sin dagli antichi greci, ma anche il più semplice. Trova la sua natura d'essere nel teorema di esistenza degli zeri: data $f : [a, b] \rightarrow \mathbb{R}$, $f \in C^0([a, b])$ allora $f(a)f(b) < 0 \Rightarrow \exists \alpha \in [a, b] \mid f(\alpha) = 0$

Esiste un punto, ovviamente non è detto che α sia univocamente determinato. Notazione

$$f \text{ continua su } [a, b] \Leftrightarrow f \in C^0([a, b])$$

$$f \text{ continua e derivata prima continua su } [a, b] \Leftrightarrow f \in C^1([a, b])$$

$$f \text{ continua e derivata prima e seconda continua su } [a, b] \Leftrightarrow f \in C^2([a, b])$$

Il metodo di bisezione nasce dal seguente problema: dati $a < b \in \mathbb{R} \mid f(a)f(b) < 0$ con $f \in C^0([a, b])$ determinare $\alpha \mid f(\alpha) = 0$. L'idea del metodo è cercare di ridurre l'ampiezza dell'intervallo costruendo una successione di intervalli sempre più piccoli che si porti dietro almeno una delle radici.

```
a[1] = a; b[1] = b;
for k=1:inf
    c(k) = (a[k] + b[k])/2; % punto di mezzo di [ak, bk]
    if (f(c[k])f(a[k]) <= 0)
        a[k+1] = a[k]; b[k+1] = c[k];
    else
        a[k+1] = c[k]; b[k+1] = b[k];
    end
end
```

Intervallo di separazione per f Intervallo che contiene un'unica soluzione per $f(x) = 0$

Teorema Se $[a_1, b_1] = [a, b]$ è un **intervallo di separazione** per $f \in C^0([a, b])$ con $f(a)f(b) < 0$ allora le successioni $a_k, b_k, c_k \rightarrow \alpha$

Dim Assumendo che l'intervallo iniziale sia un intervallo di separazione, è un'ipotesi semplificata. Quindi nell'intervallo $[a, b]$ non solo c'è una radice ma è unica, e la chiamiamo α . Per costruzione $\alpha \in [a_k, b_k]$. La distanza tra α e a_k la possiamo sempre migliorare con l'ampiezza dell'intervallo.

$$0 \leq |\alpha - a_k| \leq |b_k - a_k| = b_k - a_k = \frac{b_{k-1} - a_{k-1}}{2} = \dots = \frac{b_1 - a_1}{2^{k-1}} \Rightarrow \lim_{k \rightarrow +\infty} |\alpha - a_k| = 0$$

Analogamente per

$$b_k \quad 0 \leq |\alpha - b_k| \leq \frac{b_1 - a_1}{2^{k-1}}$$

$$c_k \quad 0 \leq |\alpha - c_k| \leq \frac{b_1 - a_1}{2^k} \quad (\text{notare il } 2^k, \text{ leggero vantaggio, perché la distanza tra } c_k \text{ e } \alpha \text{ al primo passo si può migliorare con la semiampiezza dell'intervallo}).$$

Abbiamo mostrato che gli a_k, b_k, c_k convergono alla soluzione.

Supponiamo di voler determinare un'approssimazione della radice α , denotandola con $\bar{\alpha}$, tale per cui $|\bar{\alpha} - \alpha| < \epsilon$ allora se applico il metodo di bisezione ottengo una successione c_k tale per cui $|c_k - \alpha| \leq \frac{b_1 - a_1}{2^k}$

Chiedo che $\frac{b_1 - a_1}{2^k} \leq \epsilon$, così posso prendere $\bar{\alpha} = \epsilon$. Quindi chiedo un k sufficientemente grande da permettermi

$$\frac{b_1 - a_1}{2^k} \leq \epsilon \Leftrightarrow 2^k \geq \frac{b_1 - a_1}{\epsilon} \Leftrightarrow k \geq \log_2 \left(\frac{1}{\epsilon} \right) + \log_2 (b_1 - a_1)$$

da cui si evince che il numero di passi dipende da $\log_2 \left(\frac{1}{\epsilon} \right)$.

Quindi se volessi una precisione di 10^{-14} , dovrò fare un numero di passi dell'ordine di $\log_2(10^{14}) = 14 \log_2(10)$. Fornisce un criterio **a priori** sul numero di passi per ottenere una certa precisione.

Vantaggi Costo relativamente basso per ogni passo (solamente una funzione di valutazione per passo) e l'essere applicabile a funzioni con poca regolarità (richiede solo la continuità).

Difetti Si generalizza male a funzioni in più variabili e a valori complessi (necessita di $f(a)f(b) < 0$).

Ma soprattutto la convergenza è relativamente lenta. In termini di efficienza complessiva del metodo viene soppiantato da altri metodi. Parliamo dei concetti di convergenza interessati quando si parla di convergenza lenta.

Supponiamo di avere una successione $\{x_k\}_{k \in \mathbb{N}}$ con $x_k \rightarrow \alpha$ e $\forall k \ x_k \neq \alpha$ (cioè non diventa costante in un numero finito di passi)

$\{x_k\}_{k \in \mathbb{N}}$ **converge linearmente** se

$$\lim_{k \rightarrow +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|} = l < 1$$

Se $l = \frac{1}{2}$ significa che dopo un passo ho ridotto di $\frac{1}{2}$, dopo due passi di $\frac{1}{4}$... e per arrivare ad una precisione, per esempio, di 10^{-12} ci metto un po'.

$\{x_k\}_{k \in \mathbb{N}}$ **converge almeno quadraticamente** se

$$\lim_{k \rightarrow +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = c \in \mathbb{R}$$

Con c costante reale arbitraria.

Diremo **quadraticamente** se $c \neq 0$, **più che quadraticamente** se $c = 0$ (la convergenza potrebbe essere anche più rapida).

Quindi in una convergenza lineare, l'errore al passo $k+1$ $|x_{k+1} - \alpha| \simeq l|x_k - \alpha|$

In una convergenza almeno quadratica, l'errore al passo $k+1$ $|x_{k+1} - \alpha| \simeq c|x_k - \alpha|^2$, con quindi una decrescita nettamente differente e **molto più rapida**.

Esempio $l = \frac{1}{2}$, $c = 1$, $|x_0 - \alpha| = \frac{1}{2}$

Con convergenza lineare ho $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \dots, \frac{1}{2^{k-1}}$

Con convergenza quadratica ho $\frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^4}, \frac{1}{2^8}, \dots, \frac{1}{2^{2^{k-1}}}$

Il problema è che il **metodo di bisezione presenta tipicamente una convergenza lineare**.

3.2 Metodi di Iterazione Funzionale

Seguono un po' la stessa idea che ha portato alla nascita dei metodi iterativi per risolvere sistemi lineari. Posso riformulare

$$f(x) = 0 \Leftrightarrow x = g(x)$$

Vari esempi:

$$f(x) = 0 \Leftrightarrow x = x - f(x), \text{ con } g(x) = x - f(x)$$

$$f(x) = 0 \Leftrightarrow x = x - \frac{f(x)}{h(x)}, \text{ con } g(x) = x - \frac{f(x)}{h(x)}$$

$$f(x) = 0 \Leftrightarrow x = g(x) \Rightarrow \begin{cases} x_0 \in \mathbb{R} \\ x_{k+1} = g(x_k) \end{cases}$$

Oltre ad una radice α per cui $f(\alpha) = 0$ posso trovare un **punto fisso** $x = g(x)$ ottenendo che $f(\alpha) = 0 \Leftrightarrow \alpha = g(\alpha)$. Quindi il **calcolo degli zeri di f è ricondotto al calcolo dei punti fissi di g** . Bisogna capire quando la successione ha speranze di convergere. Molto difficile trovare successioni convergenti per ogni punto iniziale, quasi impossibile.

Teorema del punto fisso Data $g : [a, b] \rightarrow \mathbb{R}$, con $g \in C^1([a, b])$, $\alpha \in (a, b) \mid g(\alpha) = \alpha$

Se $\exists \rho > 0 \mid |g'(x)| < 1 \forall x \in I_\alpha = [\alpha - \rho, \alpha + \rho] \subseteq [a, b]$ allora il metodo $\begin{cases} x_0 \in I_\alpha \\ x_{k+1} = g(x_k) \end{cases}$ genera successioni tali che:

1. $x_k \in I_\alpha \forall k \geq 0$
2. $\lim_{k \rightarrow +\infty} x_k = \alpha$

Dim g è una funzione di classe $C^1 \Rightarrow g'$ continua $\Rightarrow |g'|$ continua quindi, per Wierstrass (una funzione continua in un intervallo chiuso e limitato ammette massimo e minimo) $\Rightarrow \max_{x \in I_\alpha} |g'(x)| = \lambda < 1$

Dimostro per induzione che la successione generata dal metodo $\begin{cases} x_0 \in I_\alpha \\ x_{k+1} = g(x_k) \end{cases}$ soddisfa la relazione

$$|x_k - \alpha| \leq \lambda^k \rho \quad \forall k \geq 0$$

Passo base $k = 0$

$$|x_0 - \alpha| \leq \lambda^0 \rho = \rho \text{ ok per l'ipotesi } x_0 \in I_\alpha$$

Passo induttivo $k \Rightarrow k+1$

$$|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| = \text{per } \mathbf{Lagrange} |g'(\xi_k)(x_k - \alpha)| \text{ con } \xi_k \text{ punto tra } x_k \text{ e } \alpha$$

Lagrange dice che $f(a) - f(b) = f'(\xi)(a - b)$ con ξ tra a e b

Per ipotesi induttiva $|\xi_k - \alpha| \leq |x_k - \alpha| \leq \lambda^k \rho < \rho$ perché $\lambda < 1$, quindi $\xi_k \in I_\alpha$

$$\text{Quindi } |x_{k+1} - \alpha| = |g'(\xi_k)| \cdot |x_k - \alpha| \leq \lambda \cdot \lambda^k \rho = \lambda^{k+1} \rho$$

Traiamo che tutti i punti della successione appartengono a I_α , ma concludiamo anche che

$$0 \longleftarrow 0 \leq |x_k - \alpha| \leq \lambda^k \rho \longrightarrow 0$$

Quindi $|x_k - \alpha| \rightarrow 0 \Leftrightarrow x_k \rightarrow \alpha$

Corollario Data $g : [a, b] \rightarrow \mathbb{R}$, con $g \in C^1([a, b])$, $\alpha \in (a, b) \mid g(\alpha) = \alpha$, se riesco a trovare un intervallo **centrato in α** tale per cui $|g'(x)| < 1$ allora per ogni punto iniziale di quell'intervallo la successione converge ad α . Cioè

se $|g'(x)| < 1 \Rightarrow \exists \rho \geq 0 \mid$ posto $I_\alpha = [\alpha - \rho, \alpha + \rho]$ il metodo $\begin{cases} x_0 \in I_\alpha \\ x_{k+1} = g(x_k) \end{cases}$ genera successioni che soddisfano le precedenti proprietà 1 e 2

Dim $h(x) = |g'(x)| - 1$ continua in $[a, b]$

Allora $h(\alpha) = |g'(\alpha)| - 1 < 0 \Rightarrow$ per il teorema della permanenza del segno \exists un intorno di α dove $h(x) < 0 \Leftrightarrow |g'(x)| < 1$ in cui vale la dimostrazione precedente.

Il corollario ci dice che *esiste un intorno*, ma non ci dà informazioni su quanto ampio sia o quanto ci si possa spostare da α . Ci dà un risultato sulla **convergenza locale**.

Un metodo è localmente convergente se possiamo trovare un intorno tale per cui per ogni punto iniziale preso dall'intorno tutti i punti stanno lì dentro e la successione converge.

Esempio Dire se ci sono e, in caso, quante e dove sono le soluzioni dell'equazione $f(x) = x - \cos x = 0$

$x - \cos x = 0 \Leftrightarrow x = \cos x$, possiamo vedere l'intersezione dei grafi e notare che tra $-\frac{\pi}{2}$ e $\frac{\pi}{2}$ c'è una e una sola soluzione.

Questo perché $f_1(x) = x \Leftrightarrow f_1\left(\frac{\pi}{2}\right) = \frac{\pi}{2} > 1$ quindi non possono esserci soluzioni a destra di $\frac{\pi}{2}$ perché f_1 è strettamente crescente e \cos è compreso tra -1 e 1.

Analogamente $f_1\left(-\frac{\pi}{2}\right) = -\frac{\pi}{2} < -1$ quindi non ci possono essere soluzioni a sinistra di $-\frac{\pi}{2}$.

Quindi $f(x)$ ammette una ed una sola soluzione ed è compresa tra $-\frac{\pi}{2}$ e $\frac{\pi}{2}$.

Analiticamente noto che $\lim_{x \rightarrow \pm\infty} x - \cos x = \lim_{x \rightarrow +\infty} x \left(1 - \frac{\cos x}{x}\right) = \pm\infty$ e $f \in C^\infty(\mathbb{R}) \Rightarrow \exists$ almeno una $x \mid f(x) = 0$

La soluzione è unica per via della monotonia non decrescente di $f(x) = x - \cos x$

Quindi ho una soluzione α tra 0 e $\frac{\pi}{2}$, per cui posso applicare il metodo di bisezione ed avere subito la soluzione.

Alternativamente posso dire $x - \cos x = 0 \Leftrightarrow x = \cos x$ e immaginare un metodo iterativo del tipo $\begin{cases} x_0 \in \mathbb{R} \\ x_{k+1} = \cos x_k \end{cases}$

Quindi avrei $g(x) = \cos x$ e α **punto fisso** di g .

$$|g'(\alpha)| = |-\sin \alpha| = |\sin \alpha| < 1$$

perché $-1 \leq \sin x \leq 1$ ed α è un punto interno.

Quindi posso dire che il **metodo è localmente convergente**, quindi \exists un intorno circolare di $\alpha \mid x_0 \in I_\alpha \Rightarrow x_k \rightarrow \alpha$ cioè se parto da un punto sufficientemente vicino ad α allora la successione converge ad α . Ma questa è un'informazione debole, perché non è facile trovare l'intorno di α . Possiamo usare il teorema del punto fisso, che ci dice di vedere dov'è soddisfatta $|g'(x)| < 1$

$$|g'(x)| < 1 \Leftrightarrow |\sin x| < 1 \text{ soddisfatta se } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

Quindi qualunque intorno centrato in α che sta in $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, preso un punto in quell'intorno la successione converge ad α .

Possiamo vedere se $\alpha < \frac{\pi}{4}$ o $\alpha > \frac{\pi}{4}$, così da avere un'indicazione su come scegliere il punto di partenza.

3.2.1 Metodo delle Tangenti

Metodo di Newton Il più efficiente metodo di iterazione funzionale, grazie anche alla semplicità delle condizioni di convergenza espresse in termini di $f(x)$ e non di $g(x)$. Genera successioni che convergono velocemente e con solo due valutazioni di funzioni per passo.

Efficiente inteso in termine generale: una valutazione complessiva del costo computazionale e della velocità di convergenza.

Definizione Vogliamo trovare lo zero α di una equazione non lineare. Supponiamo un punto iniziale x_0 , da cui vogliamo generare una successione che sperabilmente converga ad α .

L'idea è la seguente: prendiamo la retta tangente a $f(x)$ nel punto $(x_0, f(x_0))$. Questa retta intersecherà l'asse delle ascisse in un nuovo punto, che chiamiamo x_1 e **ripetiamo la costruzione**.

$$\begin{cases} y - f(x_0) = f'(x_0) \cdot (x - x_0) \\ y = 0 \end{cases} \Rightarrow x_1 - x_0 = -\frac{f(x_0)}{f'(x_0)} \Leftrightarrow x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$\text{Metodo delle Tangenti} \quad \begin{cases} x_0 \in \mathbb{R} \\ x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \end{cases}$$

Questo è un metodo di iterazione funzionale, perché $x_{k+1} = g(x_k)$, con $g(x) = x - \frac{f(x)}{f'(x)}$. Quindi dobbiamo avere $g \in C^1 \Leftrightarrow f' \in C^1 \Leftrightarrow f \in C^2$

Convergenza Locale

Teorema Sia $f \in C^2([a, b])$ e $f(\alpha) = 0$, $f'(\alpha) \neq 0$, $\alpha \in (a, b)$

Allora il **metodo delle tangenti** è **localmente convergente** in α è la **convergenza** è **almeno quadratica**.

Cioè $\exists I_\alpha = [\alpha - \rho, \alpha + \rho] \mid \forall x_0 \in I_\alpha$ valgono

1. $x_k \in I_\alpha$
2. $\lim_{k \rightarrow +\infty} x_k = \alpha$
Fino a qui per la convergenza locale
3. Se $\forall k \ x_k \neq \alpha \Rightarrow \lim_{k \rightarrow +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = c \in \mathbb{R}$

Dim $f'(\alpha) \neq 0 \Rightarrow \exists \overline{I}_\alpha \mid \forall x \in \overline{I}_\alpha \ f'(x) \neq 0$

Su \overline{I}_α posso definire $g : \overline{I}_\alpha \rightarrow \mathbb{R}$ con $g(x) = x - \frac{f(x)}{f'(x)}$

$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$ continua perché composizione di funzioni continue, quindi $g \in C^1$.

$g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{f'(\alpha)^2} = 0 \Rightarrow |g'(\alpha)| < 1$

\Rightarrow per il corollario del teorema del punto fisso $\exists I_\alpha \subseteq \overline{I}_\alpha \mid$ si ha convergenza locale.

Prendiamo un punto iniziale $x_0 \in I_\alpha$ e supponiamo che la successione generata soddisfi la proprietà $x_k \neq \alpha \ \forall k$

Ricordiamo l'espressione dell'errore inerente, con la **formula di Taylor arrestata al secondo ordine** di centro x_k e raggio $\alpha - x_k$

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + \frac{f''(\xi_k)(\alpha - x_k)^2}{2} \quad \text{con } \alpha < \xi_k < x_k$$

Posso portare $f(x_k)$ al primo membro, dividere per $f'(x_k)$ e portare x_k al primo membro, ottenendo

$$x_k - \frac{f(x_k)}{f'(x_k)} = \alpha + \frac{f''(\xi_k)}{f'(x_k)} \cdot \frac{(\alpha - x_k)^2}{2}$$

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

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

Posso dividere per $(\alpha - x_k)^2$ e spostare α

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

Ora posso fare il limite. Sapendo che $x_k \rightarrow \alpha$

$$\lim_{k \rightarrow +\infty} \frac{|x_{k+1} - \alpha|}{|\alpha - x_k|^2} = \frac{1}{2} \cdot \left| \frac{f''(\alpha)}{f'(\alpha)} \right| \in \mathbb{R}$$

Convergenza in largo

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

Se $\exists \delta > 0$ tale che $f'(x) \neq 0 \wedge f(x)f''(x) > 0 \ \forall x \in (\alpha, \alpha + \delta] \subseteq [a, b]$ allora $\forall x_0 \in (\alpha, \alpha + \delta]$ la successione generata dal metodo delle tangenti a partire da x_0 converge ad α . Questo vale anche per gli intervalli sinistri di α

Questo ha delle conseguenze:

$f \in C^2(\mathbb{R}^+)$ e $f'(\alpha) \neq 0 \Rightarrow$ il metodo è localmente convergente in α

$\forall x_0 \geq \alpha$ il metodo delle tangenti genera successioni convergenti in α

Esempio $f(x) = e^x + x = 0$

$f \in C^\infty(\mathbb{R})$, $\lim_{x \rightarrow +\infty} f(x) = +\infty$, $\lim_{x \rightarrow -\infty} f(x) = -\infty$

Quindi $f(x)$ va da $-\infty$ a $+\infty$, di conseguenza ha almeno una soluzione

$f'(x) = e^x + 1 > 0$, questo $\forall x \in \mathbb{R}$

$f''(x) = e^x > 0$, questo $\forall x \in \mathbb{R}$

$$\begin{cases} f(0) = e^0 = 1 > 0 \\ f(-1) = e^{-1} - 1 < 0 \end{cases} \Rightarrow -1 < \alpha < 0$$

Convergenza locale? Sì, perché $f'(\alpha) \neq 0$

$\forall x_0 \in \mathbb{R}$ ho $x_k \rightarrow \alpha$, convergenza globale

Esempio $f(x) = \frac{1}{x} - a = 0$

Può sembrare banale e venire da dire $x = \frac{1}{a}$ ma non è banale. Vediamo il metodo delle tangenti associato:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{\left(\frac{1}{x_k} - a\right)}{\frac{1}{x_k^2}} = x_k + x_k^2 \cdot \left(\frac{1}{x_k} - a\right) = x_k + x_k - ax_k^2 = 2x_k - ax_k^2 = x_k(2 - ax_k)$$

Non è banale perché implementando il metodo delle tangenti si nota che non c'è bisogno di alcuna divisione. Ho due prodotti e un'addizione.

Questo è importante perché nei primi elaboratori elettronici l'aritmetica di macchina implementava solamente tre operazioni: addizione, sottrazione e moltiplicazione. La divisione era implementata calcolando il reciproco del numero con il metodo delle tangenti e moltiplicando. Quindi b/a era implementata calcolando $1/a$ con Newton e calcolando $b \cdot (1/a)$

Capitolo 4

Esercitazioni

Studiare il **condizionamento** del problema, la **stabilità** degli algoritmi e il **costo computazionale** (stimare il numero di operazioni) di

$$f(x) = x^n = e^{n \log(x)}$$

con $x > 0$

Condizionamento Dato che non è dipendente dal modo in cui lo si calcola, calcoliamo il condizionamento per $f(x) = x^n$ con la formula $\epsilon_{in} = \frac{f'(x)}{f(x)} \cdot x \cdot \epsilon_x$

$$\epsilon_{in} = \frac{n \cdot x^{n-1} \cdot x}{x^n} \epsilon_x = n \epsilon_x$$

quindi

$$|\epsilon_{in}| = |n \cdot \epsilon_x| \leq n \cdot u$$

Dato che $C_x = n$ possiamo dire che il problema è **ben condizionato**

Stabilità Dobbiamo considerare i due algoritmi, perché algoritmi differenti producono errore algoritmico diverso.

$$f(x) = x^n = (((x \cdot x) \cdot x) \cdot x) \dots \cdot x$$

Non possiamo calcolare $x \cdot x$ ma calcoleremo $x \otimes x = x^2(1 + \epsilon_1)$, assumendo $x \in F(B, t, m, M)$ per il calcolo dell'errore algoritmico.

Successivamente calcoliamo $x \cdot x \cdot x$, cioè $(x \otimes x) \otimes x = x^3 \cdot (1 + \epsilon_1) \cdot (1 + \epsilon_2)$.

Quindi $x \cdot \dots \cdot x$ n volte diventerà $x^n \prod_{i=1}^{n-1} (1 + \epsilon_i)$. Quindi

$$g(x) = x^n \prod_{i=1}^{n-1} (1 + \epsilon_i) = x^n (1 + \sum_{i=1}^{n-1} \epsilon_i)$$

$$\epsilon_{alg} = \frac{g(x) - f(x)}{f(x)} = \frac{x^n (1 + \sum_{i=1}^{n-1} \epsilon_i) - x^n}{x^n} = \frac{x^n \sum_{i=1}^{n-1} \epsilon_i}{x^n} = \sum_{i=1}^{n-1} \epsilon_i$$

Quindi

$$|\epsilon_{alg}| = \left| \sum_{i=1}^{n-1} \epsilon_i \right| \leq \sum_{i=1}^{n-1} |\epsilon_i| \leq (n-1)u$$

Quindi l'algoritmo è **numericamente stabile**. In questo caso, ϵ_{in} è dello stesso ordine di ϵ_{alg}

<grafo dell'errore>

L'errore accumulato è $\delta_k = \epsilon_k + \delta_{k-1}$, si somma all'errore locale della moltiplicazione (ϵ_k).

$$|\delta_k| = |\epsilon_k + \delta_{k-1}| \leq |\epsilon_k| + |\delta_{k-1}| \leq |\delta_{k-1}| + u \leq |\delta_{k-2}| + 2u \leq |\delta_1| + (k-1)u \leq |\delta_0| + ku = ku$$

Quindi possiamo dire $|\delta_n| \leq n \cdot u$. Notiamo che gli errori non si amplificano, perché è un'operazione di moltiplicazione, ma vengono semplicemente sommati. La relazione trovata lo dimostra.

Per quanto riguarda il secondo algoritmo

$$f(x) = e^{n \cdot \log(x)}$$

Si osserva che **la funzione è razionale** ($f(x) = x^n$) ma **viene calcolata in macchina usando le funzioni non razionali** $h(x) = e^x$ e $g(x) = \log(x)$. In generale, quindi, non posso eseguire l'analisi dell'errore algoritmico con gli strumenti precedentemente usati. Questo **a meno che non si sappia com'è calcolata la funzione non razionale**. Aggiriamo il problema dicendo che

$$\text{Exp}(x) = e^x \cdot (1 + \epsilon_1)$$

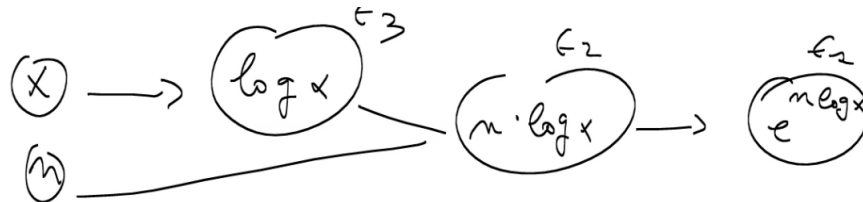
con $|\epsilon_1| \leq u$ e

$$\text{Log}(x) = \log(x) \cdot (1 + \epsilon_2)$$

con $|\epsilon_2| \leq u$. Queste sono approssimazioni, in generale $|\epsilon_1|, |\epsilon_2| \leq k \cdot u$ con k piccola che adesso trascuriamo perché l'analisi qualitativa è analoga.

In questo caso, trattiamo le funzioni non razionali come razionali.

Analisi del grafo



$$|\epsilon_1| \leq u, |\epsilon_2| \leq u, |\epsilon_3| \leq u$$

Per $x \rightarrow \log(x)$ allora $C_x = \frac{1}{\log(x)} = \frac{1}{\log(x)}$ ma non amplifica errore su x , perché per il calcolo dell'errore algoritmi assumiamo numeri di macchina.

Per $x \rightarrow e^x$ allora $C_x = \frac{e^x}{x} = x$. Quindi ϵ_2 ha come coefficiente di amplificazione pari a $n \cdot \log(x)$.

Di conseguenza calcolo l'errore algoritmico partendo dalla cima dell'albero e verso le radici (destra verso sinistra)

$$\epsilon_{alg} \doteq \epsilon_1 + (n \cdot \log(x))(\epsilon_2 + \epsilon_3)$$

Concludiamo che $|\epsilon_{alg}| = |\epsilon_1 + n \log(x)(\epsilon_2 + \epsilon_3)| \leq |\epsilon_1| + |n \log(x)(\epsilon_2 + \epsilon_3)| \leq u + |n \log(x)| \cdot |\epsilon_2 + \epsilon_3| \leq u + n |\log(x)| (|\epsilon_2| + |\epsilon_3|) \leq u + 2un |\log(x)| = u \cdot (1 + 2 \cdot n \cdot |\log(x)|)$ cioè

$$|\epsilon_{alg}| \leq u \cdot (1 + 2 \cdot n \cdot |\log(x)|)$$

Concludiamo che l'algoritmo è **numericamente instabile per x piccolo o x è grande**, ma è stabile in un intorno di 1.

L'analisi della stabilità ci dice che il primo algoritmo è preferibile.

Costo computazionale Il **primo algoritmo** $\Rightarrow (n - 1)$ operazioni moltiplicative, quindi **O(n)**.

Il **secondo algoritmo**, più complicato perché ci sono due operazioni non razionali oltre alla moltiplicazione.

Esercizio Supposto $n = 2^k$ e $f(x) = x^n = x^{2^k}$, determinare un algoritmo che calcola $f(x)$ con $k = \log_2(n)$ operazioni moltiplicative. Eseguire anche l'analisi dell'errore algoritmico per l'algoritmo trovato.

Esercizio 1 $A = I + \alpha ee^T$ con $e^T = [1 \dots]^T$ vettore colonna, quindi $n \times 1$. $e \cdot e^T$ è quindi una matrice $n \times n$.

Quindi $A = I + \alpha \begin{bmatrix} 1 & \dots & 1 \\ & \dots & \\ 1 & \dots & 1 \end{bmatrix} = I + \begin{bmatrix} \alpha & \dots & \alpha \\ & \dots & \\ \alpha & \dots & \alpha \end{bmatrix}$

1. Per quali valori di α , A è invertibile?
2. Mostrare che l'inversa è $B = I + \beta ee^T$
3. Analizzare il condizionamento di A

1. Matrice è singolare $\Leftrightarrow Ax = 0$ ha una soluzione non nulla $\Leftrightarrow \begin{cases} x_1 + \alpha \sum x_i = 0 \\ \vdots \\ x_n + \alpha \sum x_i = 0 \end{cases} \Rightarrow x_1 = x_2 = \dots = x_n = x$

$$x + \alpha \sum_{i=1}^n x_i = x + n\alpha x = x(1 + n\alpha)$$

$$\text{Se } 1 + n\alpha \neq 0, x = 0 \Rightarrow x_1 = \dots = x_n = 0$$

$$\text{Se } 1 + n\alpha = 0, \alpha = -\frac{1}{n}$$

$$A \text{ singolare} \Leftrightarrow \alpha = -\frac{1}{n}$$

2. $I = (I + \alpha ee^T)(I + \beta ee^T) \Leftrightarrow I = I + \beta ee^T + \alpha ee^T + \alpha \beta ee^T ee^T$ con α, β scalari quindi commutano

$$I = I + \beta ee^T + \alpha ee^T + \alpha \beta e(e^T e)e^T \text{ con } e^T e \text{ scalare} = n \Leftrightarrow 0 = (\beta + \alpha + \alpha \beta n)ee^T \text{ con } ee^T \neq 0$$

$$\Leftrightarrow \alpha + \beta + \alpha \beta n = 0$$

$$\Leftrightarrow \beta = \frac{-\alpha}{1 + \alpha n} \text{ con } \alpha \neq -\frac{1}{n}$$

$$\text{Quindi l'inversa è } I + \beta ee^T \Leftrightarrow \beta = \frac{-\alpha}{1 + \alpha n}$$

3. $A = \begin{bmatrix} \alpha + 1 & \alpha & \dots & \alpha \\ \alpha & \alpha + 1 & \alpha & \dots & \alpha \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \alpha & \dots & \alpha + 1 \end{bmatrix} \|A\|_\infty \dots$

$$K_\infty(A) = (|\alpha + 1| + (n-1)|\alpha|)(|\beta + 1| + (n-1)|\beta|)$$

$$\text{Mal condizionato per } |\alpha| \rightarrow +\infty$$

Esercizio 2 $A = I + \theta e_1 e_n^T + \theta e_n e_1^T$ $A = (a_{ij})$ con $a_{ij} = \min(i, j)$

1. Dire se A ammette fattorizzazione LU
2. In caso affermativo, determinare la fattorizzazione
3. Dire se la fattorizzazione è unica
4. Mostrare che A è invertibile
5. Scrivere un programma Matlab per la risoluzione di $Ax = b$ con costo lineare

$$A \in \mathbb{R}^{n \times n} \text{ con } a_{ij} = \min(i, j)$$

Ad esempio $n = 4$, abbiamo $A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$ Difficile applicare il teorema, perché valutare la complessità delle

sottomatrici principali di testa ha la stessa complessità: sono dello stesso tipo della matrice A .

$$\text{Analizziamo i casi minori, } n = 2 \text{ quindi } A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_2 \\ 0 & u_3 \end{bmatrix} \text{ e avremo } u_1 = 1, u_2 = 1, u_3 = 1 \text{ e } l = 1$$

$$\text{Quindi } A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$n = 3 \text{ quindi } A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \text{ e usando il teorema, procediamo "upgradando" la fattorizzazione di ordine 2}$$

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ Ottenendo di nuovo una triangolare}$$

superiore con tutti 1 e una triangolare inferiore con tutti 1. Ma non posso dire che *chiaramente* anche la matrice di ordine n generico si fattorizza così, perché fattorizzare per n non dà alcuna informazione sulla fattorizzazione di

$m > n$.

Posso *dimostrare* che A si fattorizza LU con due matrici una triangolare inferiore e una triangolare superiore con tutti 1

$B = LU = (b_{ij})$ con $b_{ij} = (i\text{-esima riga di } L) \times (j\text{-esima colonna di } U) = 1 + 1 + \dots + 1$ per $\min(i, j)$ volte quindi $= \min(i, j)$ e quindi $B = A$

Quindi ho dimostrato che A si fattorizza LU con due matrici una triangolare inferiore con tutti 1 per una triangolare superiore per tutti 1.

Invertibile? $\det A = \det(LU) = \det L \cdot \det U = 1 \cdot 1 = 1 \neq 0$ quindi A è invertibile.

Osservo che $A = L \cdot U$ e il termine della sottomatrice principale di ordine k di A , cioè $A = \left[\begin{array}{c|c} A_k & \\ \hline & \end{array} \right] = \left[\begin{array}{c|c} L_k & \\ \hline & \end{array} \right] \left[\begin{array}{c|c} U_k & \\ \hline & \end{array} \right]$

Quindi $A_k = L_k \cdot U_k$ quindi la fattorizzazione LU di A implica una fattorizzazione LU di tutte le radici principali di testa.

Risoluzione di $Ax = b$, usando la fattorizzazione LU perché $LUx = b$ e $Ux = y$ e $Ly = b$.

```

y(1) = b(1)/l(1, 1);
for k = 2:n
    s = 0;
    for j = 1:k-1
        s = s + l(k, j)*y(j);
    end
    y(k) = (b(k) - s)/l(k, k);
end

```

In generale questo algoritmo ha un costo quadratico. Ma gli elementi di l sono = 1

```

y(1) = b(1);
for k = 2:n
    s = 0;
    for j = 1:k-1
        s = s + y(j);
    end
    y(k) = b(k) - s;
end

```

Non faccio più operazioni moltiplicative, ma il costo rimane quadratico. Ma posso fare un'osservazione: $s = s + y_j$ è una somma crescente

```

y(1) = b(1);
s = 0
for k = 2:n
    s = s + y(k - 1);
end

```

Diventando così a costo lineare $O(n)$

Esercizio 2 $A = \begin{bmatrix} 1 & -\gamma_1 & & \\ & 1 & \ddots & \\ & & 1 & -\gamma_{n-1} \\ \gamma_0 & & & 1 \end{bmatrix}$ Proviamo che se $|\gamma_i| < 1$ allora A è invertibile.

Possiamo farlo con **Gerschgorin**:

$$K_i = \{z \in \mathbb{C} \mid |z - 1| \leq |\gamma_i|\} \text{ per } i = 1..n-1$$

$$K_n = \{z \in \mathbb{C} \mid |z - 1| \leq |\gamma_0|\}$$

$$1 - |\gamma_i| > 0 \Leftrightarrow 1 > |\gamma_i| \Leftrightarrow |\gamma_i| < 1$$

$$|\gamma_i| < 1 \Rightarrow 1 - |\gamma_i| > 0 \Rightarrow 0 \notin K_i \Rightarrow 0 \notin \bigcup_{i=1}^n K_i \Rightarrow A \text{ è invertibile}$$

Fattorizzazione LU $A_k = \begin{bmatrix} 1 & -\gamma_1 & & \\ & 1 & \ddots & \\ & & 1 & -\gamma_{k-1} \end{bmatrix}$

Triangolare superiore con elementi diagonali $\neq 0 \Rightarrow A_k$ invertibile per $k = 1..n-1 \Rightarrow \exists!$ LU di A

$$A = \left[\begin{array}{cccc|c} 1 & -\gamma_1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & & \\ \hline \gamma_0 & & & & 1 \end{array} \right] = \left[\begin{array}{ccc|c} & & & \\ & I_{n-1} & & 0 \\ \hline x_1 & \dots & x_{n-1} & 1 \end{array} \right] \cdot \left[\begin{array}{cccc|c} 1 & -\gamma_1 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & & 0 \\ & & & -\gamma_{n-2} & \\ \hline & & & 1 & -\gamma_{n-1} \end{array} \right]$$

$$[x_1 \dots x_n] \cdot \left[\begin{array}{cccc|c} 1 & -\gamma_1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & & \\ & & & -\gamma_{n-2} & \\ \hline & & & 1 & \end{array} \right] = [\gamma_0 \ 0 \dots 0] \rightarrow \begin{cases} x_1 = \gamma_0 \\ -x_1\gamma_1 + x_2 = 0 \Leftrightarrow x_2 = \gamma_0\gamma_1 \\ -x_2\gamma_2 + x_3 = 0 \Leftrightarrow x_3 = x_2\gamma_2 = \gamma_0\gamma_1\gamma_2 \\ \vdots \\ x_{n-1} = \gamma_0\gamma_1 \dots \gamma_{n-1} \end{cases}$$

$$[x_1 \dots x_{n-1}] \cdot \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ -\gamma_{n-1} \end{array} \right] + \beta = 1 \Leftrightarrow \beta = 1 + x_{n-1}\gamma_{n-1} = 1 + \gamma_0\gamma_1 \dots \gamma_{n-2}\gamma_{n-1} \Leftrightarrow \beta = 1 + \prod_{i=0}^{n-1} \gamma_i$$

Determiniamo per quali valori di γ_i A è invertibile.

$$\text{Sappiamo che } A \text{ invertibile} \Leftrightarrow \det(U) \neq 0 \Leftrightarrow \det(U) = \beta \neq 0 \Leftrightarrow \prod_{i=0}^{n-1} \gamma_i \neq -1$$

$$\text{Se } |\gamma_i| < 1 \Rightarrow \left| \prod_{i=0}^{n-1} \gamma_i \right| = \prod_{i=0}^{n-1} |\gamma_i| < 1 \quad A = \begin{bmatrix} a & & 1 \\ & \ddots & \\ 1 & & a \end{bmatrix} \in \mathbb{R}^{n \times n} \text{ con } a \in \mathbb{R}$$

Per quali a , A ammette LU? Abbiamo $A_k = \begin{bmatrix} a & & \\ & \ddots & \\ & & a \end{bmatrix} \in \mathbb{R}^{k \times k}$ per $k = 1 \dots n-1$

$$\det(A_k) = a^k \neq 0 \Leftrightarrow a \neq 0, \text{ quindi } A_k \text{ per } k = 1 \dots n-1 \text{ è invertibile} \Leftrightarrow a \neq 0$$

Quindi, per il **teorema di esistenza ed unicità**, $a \neq 0 \Rightarrow \exists!$ fattorizzazione LU di A

$$A_{n-1} = \begin{bmatrix} a & & \\ & \ddots & \\ & & a \end{bmatrix} = I_{n-1} \cdot \begin{bmatrix} a & & \\ & \ddots & \\ & & a \end{bmatrix}$$

$$A = \left[\begin{array}{ccc|c} & & & \\ & I_{n-1} & & 0 \\ \hline \frac{1}{a} & 0 & \dots & 0 \\ & & & 1 \end{array} \right] \left[\begin{array}{c|c} a & 1 \\ \ddots & 0 \\ & \vdots \\ & 0 \\ \hline 0^T & a - \frac{1}{a} \end{array} \right]$$

$$I_{n-1} \cdot z = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Leftrightarrow z = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$x^T \cdot \begin{bmatrix} a & & & \\ & \ddots & & \\ & & a & \end{bmatrix} = [1 \ 0 \ \dots \ 0] \Rightarrow \begin{cases} x_1 a = 1 \\ x_2 a = 0 \\ \vdots \\ x_{n-1} a = 0 \end{cases} \Leftrightarrow \begin{cases} x_1 = \frac{1}{a} \\ x_2 = x_3 = \dots = x_{n-1} = 0 \end{cases}$$

$$\frac{1}{a} + \beta = a \Leftrightarrow \beta = a - \frac{1}{a}$$

Dimostrare che A è invertibile per $|a| > 1$ Usiamo **Gerschgorin** Per $a > 0$

$$K_1 = K_n = \{z \in C \mid |z - a| \leq 1\}$$

$$K_2 = \dots = K_{n-1} = \{z \in C \mid |z - a| = 0\}$$

Per $a > 1 \Leftrightarrow a - 1 > 0$ ottengo $0 \notin \bigcup K_i$

$\Rightarrow A$ è invertibile per $a > 1$

Per $a < 0 \Rightarrow a < -1 \Leftrightarrow a + 1 < 0 \Rightarrow 0 \notin \bigcup K_i$

$\Rightarrow A$ è invertibile per $a < -1$

$\Rightarrow A$ **invertibile** per $|a| > 1$

Determinante di A Usiamo **Laplace**

$A = LU \Rightarrow \det(A) = \det(L) \cdot \det(U)$ (per **Binet**)

$$\det(U) = \prod_{k=1}^n u_{kk} = a^{n-1} \left(a - \frac{1}{a}\right) = a^n - a^{n-2} = a^{n-2}(a^2 - 1)$$

A è singolare $\Leftrightarrow U$ è singolare $\Leftrightarrow a = 0 \vee a = \pm 1$

$$f(a) = a^n - a^{n-2}$$

$$C_a = \frac{f'(a)}{f(a)} \cdot a$$

$$|\epsilon_{in}| = |C_a| \cdot |\epsilon_a| \leq |C_a| a$$

$$C_a = \frac{n \cdot a^{n-1} - (n-2) \cdot a^{n-3}}{a^n - a^{n-2}} \cdot a$$

$$C_a = \frac{\dot{a}^{n-2}(n \cdot a^2 - (n-2))}{\dot{a}^{n-2}(a^2 - 1)}$$

$$C_a = \frac{n \cdot a^2 - (n-2)}{a^2 - 1}$$

Mal condizionato per $a \rightarrow \pm 1$

Cosa posso dire quando $|a| \rightarrow +\infty$? $a^2 \rightarrow +\infty$

$$\lim_{|a| \rightarrow +\infty} C_a = \lim_{|a| \rightarrow +\infty} \frac{n \cdot a^2 - (n-2)}{a^2 - 1} = \lim_{|a| \rightarrow +\infty} \frac{\dot{a}^2(n - \frac{n-2}{a^2})}{\dot{a}^2(1 - \frac{1}{a^2})} = n$$

Quindi è **ben condizionato** per $|a| \rightarrow +\infty$ $A = \begin{bmatrix} 1 & \theta & \dots & \theta \\ -\theta & \ddots & & \\ \vdots & & \ddots & \\ -\theta & & & 1 \end{bmatrix}$ con $\theta \in \mathbb{R}$

Determinare condizioni sufficienti su θ affinché A abbia fattorizzazione LU Predominanza diagonale

$$i = 1 \rightarrow |1| = 1 > (n-1)|\theta| \Leftrightarrow |\theta| < \frac{1}{n-1}$$

$$\Leftrightarrow -\frac{1}{n-1} < \theta < \frac{1}{n-1}$$

$$i = 2 \dots n \rightarrow |1| = 1 > |-\theta| = |\theta| \Leftrightarrow |\theta| < 1 \\ \Leftrightarrow -1 < \theta < 1$$

Quindi A è predominante diagonale $\Leftrightarrow -\frac{1}{n-1} < \theta < \frac{1}{n-1} \Rightarrow \exists!$ la fattorizzazione LU

Determinare tutti i valori di θ per i quali A ammette fattorizzazione LU Determinante?

$$\det \begin{pmatrix} 1 & \theta & \dots & \theta \\ -\theta & \ddots & & \\ \vdots & & \ddots & \\ -\theta & & & 1 \end{pmatrix} = \det \begin{pmatrix} 1 & & & -\theta \\ & \ddots & & \vdots \\ & & \ddots & -\theta \\ \theta & \dots & \theta & 1 \end{pmatrix}$$

Questo perché passo da $\begin{bmatrix} 1 & \theta & \dots & \theta \\ -\theta & \ddots & & \\ \vdots & & \ddots & \\ -\theta & & & 1 \end{bmatrix}$ a $\begin{bmatrix} 1 & & & -\theta \\ & \ddots & & \vdots \\ & & \ddots & -\theta \\ \theta & \dots & \theta & 1 \end{bmatrix}$ semplicemente scambiando righe e colonne.

Questa osservazione è utile perché di $\begin{bmatrix} 1 & \theta & \dots & \theta \\ -\theta & \ddots & & \\ \vdots & & \ddots & \\ -\theta & & & 1 \end{bmatrix}$ a $\begin{bmatrix} 1 & & & -\theta \\ & \ddots & & \vdots \\ & & \ddots & -\theta \\ \theta & \dots & \theta & 1 \end{bmatrix}$ so calcolare la fattorizzazione LU .

$$\left[\begin{array}{ccc|c} I_{n-1} & & & \\ \hline \theta & \dots & \theta & 1 \end{array} \right] \cdot \left[\begin{array}{ccc|c} & & & -\theta \\ & & & \vdots \\ & & & -\theta \\ \hline 0 & & & 1 + (n-1)\theta^2 \end{array} \right]$$

Questo mi dice che $\det(A) = 1 + (n-1)\theta^2 \neq 0 \forall \theta$

Quindi le sottomatrici principali di testa sono invertibili $\forall \theta$, quindi **ammette unica fattorizzazione $LU \forall \theta$**

Determinare E_1 matrice elementare di Gauss tale che $E_1 \cdot A$ abbia 0 nella prima colonna esclusa la prima posizione $E_1 = ?$, ricordando la costruzione della matrice di Gauss ottengo

$$E_1 = \begin{bmatrix} 1 & & & & \\ +\theta & \ddots & & & \\ \vdots & & \ddots & & \\ +\theta & & & \ddots & 1 \end{bmatrix}$$

$$E_1 A = \begin{bmatrix} 1 & & & & \\ +\theta & \ddots & & & \\ \vdots & & \ddots & & \\ +\theta & & & \ddots & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & \theta & \dots & \theta \\ -\theta & \ddots & & \\ \vdots & & \ddots & \\ -\theta & & & 1 \end{bmatrix} = \begin{bmatrix} 1 & \theta & \dots & \dots & \dots & \theta \\ 0 & 1 + \theta^2 & \theta^2 & \dots & \dots & \theta^2 \\ \vdots & \theta^2 & 1 + \theta^2 & \theta^2 & \dots & \theta^2 \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \theta^2 \\ 0 & \theta^2 & \dots & \dots & \theta^2 & 1 + \theta^2 \end{bmatrix}$$

Con $1 + \theta^2$ sulla diagonale principale e θ^2 fuori da essa.

Notiamo che A è una matrice sparsa, cioè *tanti elementi di A sono 0*, che è una proprietà che vogliamo sfruttare. Ma in generale **il metodo di eliminazione gaussiana non preserva la sparsità della matrice**. Si riempiono sia $A^{(k)}$ sia L e U : questo si chiama **fill-in**.

Esercizio

$$A = \begin{bmatrix} 2+h & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2+h \end{bmatrix}$$

1. Determinare condizioni sufficienti su h tali da garantire esistenza ed unicità della fattorizzazione LU

2. Sotto tali condizioni, scrivere un programma per il calcolo dei fattori L e U

1. Proviamo attraverso la predominanza diagonale

$$i = 1, n \rightarrow |2+h| > 1$$

$$i = 2 \dots n-1 \rightarrow |2+h| > 2$$

Quindi la condizione sarà $|2+h| > 2 \Leftrightarrow \begin{cases} 2+h > 2 \\ 2+h \geq 0 \end{cases} \cup \begin{cases} -(2+h) > 2 \\ 2+h \leq 0 \end{cases} \Leftrightarrow \begin{cases} h > 0 \\ h \geq -2 \end{cases} \cup \begin{cases} h < -4 \\ h \leq -2 \end{cases}$
 \Rightarrow la matrice è predominante diagonale quando $h < -4 \vee h > 0$

2. Abbiamo le condizioni, vediamo come si comportano L ed U

Abbiamo una matrice del tipo $\begin{bmatrix} a & b & & \\ b & \ddots & \ddots & \\ & \ddots & \ddots & b \\ & & b & a \end{bmatrix}$ con $a = 2+h$ e $b = -1$. Il primo passo della fattorizzazione è

mettere 0 nella prima colonna. L'unico elemento da annullare è quello in posizione $(2,1)$, basta quindi prendere $-\frac{b}{a}$ come elemento corrispondente in E_1

$$E_1 = \begin{bmatrix} 1 & & & \\ -\frac{b}{a} & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

$$E_1 A = \begin{bmatrix} a & b & & \\ 0 & \bar{a} & b & \\ \vdots & b & a & \ddots \\ 0 & & \ddots & \ddots \end{bmatrix}$$

La matrice su cui dovrò lavorare, cioè $E_1 A$ privata della prima riga e della prima colonna, è ancora tridiagonale. In questo caso, e in generale nelle matrici cosiddette "a banda", il metodo di eliminazione gaussiana preserva la sparsità.

E_2 sarà analogamente con un elemento $-\frac{b}{a} \dots$

La U sarà una bidiagonale superiore, mentre la L sarà bidiagonale inferiore.

Esempio

$$A = \begin{bmatrix} \alpha & 1 & & 0 \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & 1 \\ \alpha & \dots & \dots & \alpha \end{bmatrix} \in \mathbb{R}^{n \times n} \text{ con } n \geq 2$$

Predominante diagonale?

$$i = 1 \rightarrow |\alpha| > 1$$

$$i = 2 \rightarrow \begin{cases} n = 2 & |\alpha| > |a| \\ n > 2 & |\alpha| > |a| + 1 \end{cases}$$

Quindi non è predominante diagonale per nessun valore α

Per quali α , G-S converge?

$$M = \begin{bmatrix} \alpha & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ \alpha & \dots & \dots & \alpha \end{bmatrix} \quad N = \begin{bmatrix} & -1 & & 0 \\ & & \ddots & \\ 0 & & & -1 \end{bmatrix}$$

Devo studiare $\rho(G) = \max_{i=1}^n |\lambda_i|$

$$G = \begin{bmatrix} \alpha & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ \alpha & \dots & \dots & \alpha \end{bmatrix}^{-1} \cdot \begin{bmatrix} -1 & & 0 \\ & \ddots & \\ 0 & & -1 \end{bmatrix}$$

$$\left(\alpha \cdot \begin{bmatrix} 1 & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ 1 & \dots & \dots & 1 \end{bmatrix} \right)^{-1} = \frac{1}{\alpha} \cdot \begin{bmatrix} 1 & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ 1 & \dots & \dots & 1 \end{bmatrix}^{-1}$$

Sapendo che $\begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ 0 & -1 & 1 & \\ 0 & 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$, allora ipotizzo che

$$\begin{bmatrix} 1 & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ 1 & \dots & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & & & \\ -1 & \ddots & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \ddots & & 0 \\ & & \ddots & \\ & 0 & & 1 \end{bmatrix}$$

$$M^{-1} = \frac{1}{\alpha} \begin{bmatrix} 1 & & & \\ -1 & \ddots & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix}$$

$$G = M^{-1}N = \frac{1}{\alpha} \begin{bmatrix} 1 & & & \\ -1 & \ddots & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -1 & & 0 \\ & \ddots & \\ 0 & & -1 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} 0 & -1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & \dots & \dots \\ & & & \ddots & \ddots & \\ & & & & \ddots & \ddots \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{\alpha} & & & \\ 0 & \frac{1}{\alpha} & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & -\frac{1}{\alpha} \end{bmatrix}$$

G triangolare superiore \Rightarrow gli autovalori stanno sulla diagonale principale e sono $\lambda = 0$
 $\lambda = \frac{1}{\alpha} \Rightarrow \rho(G) = \frac{1}{|\alpha|}$

Quindi $\frac{1}{|\alpha|} < 1 \Leftrightarrow 1 < |\alpha| \Leftrightarrow |\alpha| > 1$

$\det(\lambda I - M^{-1}N) = \det(\lambda M^{-1}M - M^{-1}N) = \det(M^{-1} \cdot (\lambda M - N)) = (\text{per Binet}) \det(M^{-1}) \cdot \det(\lambda M - N)$

Quindi $\det(\lambda I - G) = 0 \Leftrightarrow \det(\lambda M - N) = 0$

$$\det \left(\alpha \lambda \begin{bmatrix} 1 & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ 1 & \dots & \dots & 1 \end{bmatrix} + \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix} \right) = \det \left(\begin{bmatrix} \alpha \lambda & 1 & & \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & 1 \\ \alpha \lambda & \dots & \dots & \alpha \lambda \end{bmatrix} \right)$$

$$M = \begin{bmatrix} \alpha & & & \\ \vdots & \ddots & & \\ \alpha & \dots & \alpha \end{bmatrix} \Rightarrow \det(M^{-1}N) = \det(M^{-1}) \cdot \det(N) = \frac{\det(N)}{\det(M)} = \frac{\det(N)}{\alpha^n} = \frac{0}{\alpha^n} = 0$$

Capitolo 5

Matlab

11 bit esponente rappresentato in traslazione, 52 bit mantissa (53 cifre rappresentabili per il bit nascosto) $\Rightarrow u = B^{1-t} = 2^{-52}$

5.1 Esponenziale

$f(x) = e^x$ approssimato con Taylor

$$e^x \simeq 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!}$$

Implementiamo un algoritmo che dati x , n in input restituisce $y = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!}$ in output. Potrebbe essere utile un programma del tipo

```
y = 1;
for k = 1 .. n
    y = y + x^(k)/factorial(k)
end
```

Dal punto di vista del costo computazionale, però, ci sono dei problemi. Al passo k faccio k prodotti per x^k e k prodotti per $k!$, quindi **circa $2k$ prodotti**.

Il costo totale è del tipo $C = 2 + 2 \cdot 2 + 2 \cdot 3 + \dots + 2 \cdot n = 2(1 + 2 + 3 + \dots + n)$.

Ricordiamo $1 + 2 + \dots + n = \frac{n(n+1)}{2}$, quindi $C = 2 \cdot \frac{n(n+1)}{2} = O(n^2)$ quindi costo quadratico.

Una variazione più efficiente

```
y = 1; p = 1; m = 1;
for k = 1 .. n
    p = p * x;      % accumula le potenze
    m = m * k;      % accumula il fattoriale
    y = y + p/m     % accumula la somma
end
```

Al passo k ho 4 operazioni aritmetiche (o 3 operazioni moltiplicative, perché alcune volte si afferma che $t_* \gg t_+$ in termini di operazioni tra bit (**costo booleano**))

Il costo totale $C = 4n$ operazioni (o $3n$).

I due algoritmi possono produrre numeri molto grandi. Una versione che argina questo problema è la seguente.

```
y = 1; z = 1;
for k = 1 .. n
    z = z * x/k;    % controllare l'esplosione dei numeri
    y = y + z;
end
```

Codice MatLab (file: myexp.m)

```
function [y] = myexp(x,n)
    % y = 1 + x + (x^2)/2 + ... + (x^n)/(n!)
    y = 1; z = 1;
    for k = 1:n
        z = z * x/k;
        y = y + z;
    end
end
```