

---

# CALCOLO NUMERICO

---

**Corso A**

**Autore**

Giuseppe Acocella

2024/25

Ultima Compilazione - March 12, 2025

# Contents

<b>1</b>	<b>Introduzione</b>	<b>4</b>
1.1	Fasi dell'Analisi Numerica . . . . .	4
1.2	Errore Inerente ed Errore di Approssimazione . . . . .	4
1.3	Rappresentazione Virgola Fissa vs Virgola Mobile . . . . .	5
1.4	Teorema di Rappresentazione in Base . . . . .	5
1.4.1	Motivazioni e commenti . . . . .	5
1.5	Insieme di Numeri di Macchina . . . . .	5
1.5.1	Cardinalità dell'Insieme di Numeri di Macchina . . . . .	6
1.5.2	Numero più piccolo/più grande . . . . .	6
1.5.3	Standard IEEE . . . . .	6
<b>2</b>	<b>Studio dell'Errore</b>	<b>7</b>
2.1	Troncamento/Arrotondamento . . . . .	7
2.1.1	Teorema di Errore di Rappresentazione (con Dim.) . . . . .	7
2.2	Operazioni di Macchina . . . . .	9
2.2.1	Errore nella Somma e suo Relativo Ordine (con Dim.) . . . . .	9
2.2.2	Teorema di Errore di Calcolo Funzione Razionale (con Dim.) . . . . .	11
2.2.3	Condizionamento vs Stabilità . . . . .	12
2.2.4	Teorema Coefficiente di Amplificazione ed Errore Inerente (con Dim.) . . . . .	12
2.2.5	Errore di Calcolo Funzione Irrazionale . . . . .	13
<b>3</b>	<b>Algebra Lineare Numerica - Computazione e Condizionamento</b>	<b>14</b>
3.1	Norme Vettoriali . . . . .	14
3.1.1	Distanza, Norma 1, Norma 2, Norma Infinito . . . . .	14
3.2	Norma Matriciale . . . . .	15
3.2.1	Norma Matriciale indotta da Norma Vettoriale . . . . .	15
3.2.2	Norma di Frobanius . . . . .	15
3.2.3	Th. Compatibilità delle Norme (con Dim.) . . . . .	16
3.2.4	Metodi Iterativi su Norme . . . . .	16
3.2.5	Matrici Simmetriche sui Reali . . . . .	17
3.2.6	Th. di Hirsch (con Dim.) . . . . .	17
3.3	Utilities Greshgorin . . . . .	18
3.3.1	Cerchio i-esimo di Greshgorin . . . . .	18
3.3.2	Th. di Greshgorin (con Dim.) . . . . .	18
3.3.3	Invertibilità e Predominanza Diagonale . . . . .	19
3.3.4	II Th. di Gershgorin . . . . .	20
3.4	Condizionamento del Problema sulla Risoluzione di Sistemi Lineari . . . . .	20
3.4.1	Teorema sul Condizionamento di Norme Matriciali (con dim) . . . . .	20
<b>4</b>	<b>Metodi Diretti per Risoluzione di Sistemi Lineari</b>	<b>22</b>
4.1	Fattorizzazione LU . . . . .	23
4.2	Th. Condizioni Sufficienti per Esistenza ed Unicità Fattorizzazione LU (con Dim.) . . . . .	23
4.3	Matrici Elementari di Gauss . . . . .	24

4.4	Tecniche di Pivoting . . . . .	25
-----	--------------------------------	----

# 1 Introduzione

I temi principali trattati in questi appunti saranno riguardanti i processi matematici che ci permettono di analizzare la conversione da continuo a discreto, per poter fornire questi dati ad una macchina finita. Spesso questi approcci vengono utilizzati anche quando la complessità di un determinato algoritmo è troppo elevata e di conseguenza si preferisce analizzare delle approssimazioni discrete.

## 1.1 Fasi dell'Analisi Numerica

Elenchiamo le fasi dell'Analisi Numerica:

1. La prima fase è lo studio del **Mondo Reale** che osserviamo.
2. Grazie all'osservazione del **Mondo Reale** generiamo un **Modello Matematico Continuo** in una seconda fase.
3. La terza fase cerca di **discretizzare** il modello precedente in uno **discreto**. Questo genera un errore detto **errore analitico**.
4. Si cerca un **Metodo di Risoluzione** al **Modello Matematico Discreto** durante una quarta fase. Questo genera un errore detto **errore inerente**, dato dalla rappresentazione discreta di qualcosa di continuo.
5. L'ultima fase è quella della **Soluzione Approssimata** trovata dal **Metodo di Risoluzione** proposto. Questo produce un errore detto **errore algoritmico**.

## 1.2 Errore Inerente ed Errore di Approssimazione

Consideriamo una  $x$  continua, la sua rappresentazione su una macchina sarà  $\bar{x}$ . Definiamo dunque i due errori  $\varepsilon_{IN}$  ed  $\varepsilon_x$ :

1. **Errore Inerente** ( $\varepsilon_{IN}$ ): Assumendo una funzione  $f$ :

$$\varepsilon_{IN} = \frac{f(\bar{x}) - f(x)}{f(x)}$$

2. **Errore di Approssimazione** ( $\varepsilon_x$ ): Assumendo una funzione  $f$ :

$$\varepsilon_x = \frac{\bar{x} - x}{x}$$

### 1.3 Rappresentazione Virgola Fissa vs Virgola Mobile

Immaginiamo di avere una quantità  $k$  fissata di bit da poter utilizzare per rappresentare un numero su una macchina. Descriviamo due potenziali metodologie di rappresentazione:

1. **Numeri a virgola fissa:** Si compongono di un **segno**, una **parte intera** ed una **parte frazionaria**.
2. **Numeri a virgola mobile:** Si compongono di una **mantissa** ossia un numero compreso tra 0 ed 1 (estremi esclusi), un **segno** ed un **esponente**.

### 1.4 Teorema di Rappresentazione in Base

Sia  $x \in \mathbb{R}$ ,  $x \neq 0$ , allora scelta una base  $\beta$  di rappresentazione **esistono e sono unici**:

1. Un valore  $\rho \in \mathbb{Z}$  detto **esponente**.
2. Una successione  $\{d_i\}_{i=1,2,\dots}$  dette **cifre**.
3.  $d_i$  non tutte uguali a  $\beta - 1$  da un certo punto in poi.

tali che:

$$x = \text{segno}(x) \beta^\rho \left( \sum_{i=1}^{\infty} d_i \beta^{-i} \right)$$

#### 1.4.1 Motivazioni e commenti

1.  $d \neq 0$  altrimenti avrei **rappresentazioni diverse di stessi numeri**, di conseguenza cadrebbe l'**unicità** delle rappresentazioni.
2.  $d_i$  non tutte uguali a  $\beta - 1$  da un certo punto in poi altrimenti numeri come  $0.\bar{9}$  convergerebbe ad 1.

### 1.5 Insieme di Numeri di Macchina

Definiamo l'insieme  $\Phi$  che permette la rappresentazione dei numeri di macchina:

$$\Phi(\beta, t, m, M) = \{0\} \cup \{x \in \mathbb{R}, x = \text{segno}(x) \beta^\rho \left( \sum_{i=1}^t d_i \beta^{-i} \right)\}$$

1.  $\beta$ : **base**.
2.  $t$  cifre della **mantissa**.
3.  $-n \leq \rho \leq M$ , ossia i due **estremi** che contengono l'**esponente**  $\rho$ .
4. Sono necessarie delle ipotesi a supporto dell'**unicità** di questa formulazione:
  - (a)  $0 \leq d_i \leq \beta - 1$
  - (b)  $d_1 \neq 0$

### 1.5.1 Cardinalità dell'Insieme di Numeri di Macchina

$$\#\Phi(\beta, t, m, M) = 1 + 2(n + M + 1)(\beta - 1)(\beta^{t-1})$$

1. 1 è la cardinalità dello **zero**.
2. Il prodotto con 2 è dato dal **segno**.
3.  $(n + M + 1)$  tutte le possibili **configurazioni** dell'**esponente**  $\rho$ .
4.  $(\beta - 1)$  tutte le possibili **configurazioni** delle **cifre** rispetto alla base (escluso lo zero).
5.  $(\beta^{t-1})$  avendo  $t$  **bit disponibili** e  $\beta$  la **base**, allora ho tutte le possibili **combinazioni** (escluse tutte quelle che iniziano con lo zero).

### 1.5.2 Numero più piccolo/più grande

Analizziamo la rappresentazione del numero  $\omega$  **più piccolo** e del numero  $\Omega$  più grande.

1. Numero **più piccolo** rappresentabile  $\omega$ :

$$\omega = \beta^{-m}(0.10...0)_\beta = (\beta^{-m})(\beta^{-1}) = \beta^{-m-1}$$

Questo perchè vogliamo la nostra base  $\beta$  elevata al più piccolo estremo degli esponenti  $-m$  moltiplicata alla più piccola mantissa nella base corrente.

2. Numero **più grande** rappresentabile  $\Omega$ :

$$\Omega = \beta^M(0.[\beta - 1][\beta - 1][\beta - 1]) = \beta^M(1 - \beta^{-t})$$

Questo perchè vogliamo la **base**  $\beta$  elevata al più grande dei possibili esponenti  $M$ , ripetendo nella mantissa tutte le cifre più grandi permesse dalla base.

### 1.5.3 Standard IEEE

Lo **Standard IEEE** può essere rappresentato come istanza di  $\Phi$ :

$$Standard_{IEEE} = \Phi(2, 53, 1021, 1024)$$

Contando 1 bit per il segno e 11 bit per l'esponente, 52 bit vengono dedicati alla mantissa ed ad alcuni simboli speciali, come  $NaN$  oppure  $\infty$ .

**Underflow/Overflow** Una volta stabilito questo standard, se l'esponente  $\rho$  esce dall'intervallo  $[-m, M]$ , allora:

1. Se  $\rho > M$  allora è **overflow**, e ad esempio in Matlab questo comportamento viene approssimato ad  $\infty$ .
2. Se  $\rho < -m$  allora è **underflow**, e in Matlab questo comportamento viene approssimato a 0.

## 2 Studio dell'Errore

### 2.1 Troncamento/Arrotondamento

Se  $\rho \in [-m, M]$  allora possono succedere due cose:

1.  $x \in \mathbb{R}$  si rappresenta su  $t$  cifre della mantissa disponibili.
2.  $x \in \mathbb{R}$  ha bisogno di più cifre rispetto a quelle fornite per la mantissa. In questo caso posso operare in due modi:
  - (a) **Troncamento:**  $x$  viene rappresentato con il numero di macchina subito prima. Quindi  $x$  viene rappresentato con il numero di macchina  $\tilde{x}$  che sia più grande rappresentabile con  $|\tilde{x}| \leq |x|$ .
  - (b) **Arrotondamento:**  $x$  viene rappresentato con  $\tilde{x}$  numero di macchina più vicino.

**Errore Assoluto/Relativo** Definiamo due tipi di errore:

1. **Errore Assoluto** ( $\epsilon$ ):

$$\epsilon = \tilde{x} - x$$

2. **Errore Relativo** ( $\epsilon_x$ ):

$$\epsilon_x = \frac{\tilde{x} - x}{x}$$

#### 2.1.1 Teorema di Errore di Rappresentazione (con Dim.)

Sia  $x \in \mathbb{R}$ ,  $x \neq 0$  e  $\omega \leq |x| \leq \Omega$  allora:

1. Identificando con  $u$  la **precisione di macchina**:

$$|\epsilon_x| < u$$

ed oltre a questo:

- (a) Operando con **troncamento**:

$$u = \beta^{1-t}$$

- (b) Operando con **arrotondamento**:

$$u = \frac{1}{2}\beta^{1-t}$$

### Dimostrazione

1. Rappresentazione in base:

$$x = \beta^\rho \left( \sum_{i=1}^{\infty} d_i \beta^{-i} \right) \quad \text{con } \rho \in [-m, M]$$

2. Assumiamo di star considerando i numeri di macchina in troncamento, dunque cambia l'indice della sommatoria in  $t$  cifre:

$$x = \beta^\rho \left( \sum_{i=1}^t d_i \beta^{-i} \right) \quad \text{con } \rho \in [-m, M]$$

3. \*\*\*\*

$$|\tilde{x} - x| < |b - a|$$

4. \*\*\*

$$|x| \geq \beta^{\rho-1} = \beta^\rho (0.1)_\beta$$

5. Dunque alla fine possiamo ricavare  $\tilde{x}$  della nostra macchina:

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



## 2.2 Operazioni di Macchina

Assumendo  $\tilde{x}, \tilde{y}$  con  $\tilde{x}, \tilde{y} \in \Phi(10, t, 5, 5)$  ma con  $\tilde{x} + \tilde{y} \notin \Phi(10, t, 5, 5)$ , risulta necessario definire nuove operazioni, ossia delle **operazioni di macchina**:

1. **Operazione Somma**: Prendiamo la versione in floating point dell'operazione somma originale:

$$\tilde{x} \oplus \tilde{y} = fl(\tilde{x} + \tilde{y})$$

e l'**errore** generato sarà:

$$\epsilon = \frac{(\tilde{x} \oplus \tilde{y}) - (\tilde{x} + \tilde{y})}{(\tilde{x} + \tilde{y})}$$

Si associa quindi un'operazione reale ad una approssimativa di macchina, quindi anche che

$$|\epsilon| < u$$

2. **Operazione Differenza**: Allo stesso modo approssimiamo la differenza, anch'essa produrrà un errore:

$$\epsilon = \frac{(\tilde{x} \oplus \tilde{y}) - (x + y)}{(x + y)}$$

### 2.2.1 Errore nella Somma e suo Relativo Ordine (con Dim.)

Mostriamo per step questa dimostrazione:

1. Definizione di  $\tilde{x}$ :

$$\tilde{x} = x(1 + \epsilon_x) \quad , \quad \tilde{y} = y(1 + \epsilon_y)$$

2. Prendiamo in considerazione **solo** l'errore sull'**operazione somma**:

$$\tilde{x} \oplus \tilde{y} = (\tilde{x} + \tilde{y})(1 + \epsilon)$$

3. Sostituiamo le definizioni di (1.) in (2.):

$$\tilde{x} \oplus \tilde{y} = [x(1 + \epsilon_x) + y(1 + \epsilon_y)](1 + \epsilon)$$

4. Svolgiamo i prodotti:

$$\tilde{x} \oplus \tilde{y} = (x + y) + x\epsilon_x + y\epsilon_y + (x + y)\epsilon + x\epsilon_x\epsilon + y\epsilon_y\epsilon$$

5. Considerando il fatto che gli ultimi due operandi sono generati dal prodotto di due epsilon diversi, li ignoriamo effettuando un **approssimazione al prim'ordine**:

$$\tilde{x} \oplus \tilde{y} = (x + y) + x\epsilon_x + y\epsilon_y + (x + y)\epsilon$$

6. Tornando alla definizione di  $\epsilon_{TOT}$  dell'operazione somma:

$$\epsilon = \frac{(\tilde{x} \oplus \tilde{y}) - (x + y)}{(x + y)}$$

Effettuiamo sostituzione di  $\tilde{x} \oplus \tilde{y}$  approssimati al prim'ordine:

$$\begin{aligned} \epsilon_{TOT} &= \frac{[(x + y) + x\epsilon_x + y\epsilon_y + (x + y)\epsilon] - (x + y)}{(x + y)} = \\ &= \frac{x}{x + y}\epsilon_x + \frac{y}{x + y}\epsilon_y + \epsilon \end{aligned}$$

7. Otteniamo dunque i primi due operandi che rappresentano l'**errore inerente**, ossia quello che si propaga dalle precedenti operazioni, e l'**errore algoritmico**, causato dalla corrente operazione:

$$\epsilon_{TOT} = \frac{x}{x + y}\epsilon_x + \frac{y}{x + y}\epsilon_y + \epsilon$$

8. Rendendo generica la formula ottenuta definiamo dei **coefficienti di amplificazione**:

$$\epsilon_{TOT} = \epsilon_{OP} + C_1\epsilon_{TOT}^{(k)} + C_2\epsilon_{TOT}^{(s)}$$

Questa formula di  $\epsilon_{TOT}$  si basa su una generica operazione

$$z^{(i)} = z^{(k)} \text{ op } z^{(k)}$$

### 2.2.2 Teorema di Errore di Calcolo Funzione Razionale (con Dim.)

Calcolo dell'errore totale  $\epsilon_{TOT}$ :

$$\epsilon_{TOT} = \epsilon_{IN} + \epsilon_{ALG}$$

dove:

1. L'**Errore Inerente** dipende esclusivamente dal problema:

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

2. L'**Errore Algoritmico** dipende dalla scelta della funzione scelta:

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

#### Dimostrazione

1. Prendendo  $\epsilon_{TOT}$  (inteso come somma di  $\epsilon_{IN}$  e  $\epsilon_{ALG}$ ) sommiamo e sottraiamo  $f(\tilde{x})$ :

$$\begin{aligned} \epsilon_{TOT} &= \epsilon_{IN} + \epsilon_{ALG} = \\ &= \frac{g(\tilde{x}) - f(x) + f(\tilde{x}) - f(\tilde{x})}{f(x)} \end{aligned}$$

2. Distribuisco in modo tale da poter ricavare l'**errore inerente** (secondo operando):

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

3. Moltiplico e divido per  $f(\tilde{x})$  e scambio tra loro i due denominatori:

$$\epsilon_{TOT} = \frac{g(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})} * \frac{f(\tilde{x})}{f(x)} + \epsilon_{IN}$$

4. Notiamo che il primo fattore del primo operando corrisponde all'errore algoritmico:

$$\epsilon_{TOT} = \epsilon_{ALG} * \frac{f(\tilde{x})}{f(x)} + \epsilon_{IN}$$

5. Assumendo che  $\epsilon_{IN+1} = \frac{f(\tilde{x})}{f(x)}$  sostituiamo:

$$\epsilon_{TOT} = (\epsilon_{ALG})(\epsilon_{IN+1}) + \epsilon_{IN}$$

6. Approssimiamo  $\epsilon_{ALG} \doteq (\epsilon_{ALG})(\epsilon_{IN+1})$ :

$$\epsilon_{TOT} = \epsilon_{IN} + \epsilon_{ALG}$$

### 2.2.3 Condizionamento vs Stabilità

Descriviamo le differenze tra **condizionamento** e **stabilità**:

1. **Condizionamento**: Studio dell'errore, errore intrinseco.
2. **Stabilità**: Studio dell'errore algoritmico, rappresenta la stabilità numerica dell'algoritmo proposto.

### 2.2.4 Teorema Coefficiente di Amplificazione ed Errore Inerente (con Dim.)

Sia  $f(x) \in C^2$  (ossia derivabili due volte con entrambe le derivate continue), allora:

$$\epsilon_{IN} = \frac{x}{f(x)} f'(x) \epsilon_x$$

Grazie a questo possiamo determinare se un problema risulta mal condizionato o ben condizionato:

	somma	sottrazione	prodotto	divisione
c1	$\frac{x}{x+y}$	$\frac{x}{x-y}$	1	1
c2	$\frac{y}{x+y}$	$\frac{-y}{x-y}$	1	-1

1. **Ben Condizionato**: Non ho punti nel dominio del coefficiente di amplificazione dove la funzione va a  $\infty$ .
2. **Mal Condizionato**: Ho dei punti in cui il coefficiente può andare a  $\infty$ . Dunque un problema può essere mal condizionato in specifici intervalli.

### Dimostrazione

1. Riprendiamo lo sviluppo di Taylor fino al secondo ordine:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0) \frac{(x - x_0)^2}{2!}$$

2. Contestualizziamo ad  $\tilde{x}$ :

$$f(\tilde{x}) = f(x) + f'(x)(\tilde{x} - x) + f''(x) \frac{(\tilde{x} - x)^2}{2!}$$

3. Porto a sinistra  $f(x)$ :

$$f(\tilde{x}) - f(x) = f'(x)(\tilde{x} - x) + f''(x) \frac{(\tilde{x} - x)^2}{2!}$$

4. Moltiplico e divido per  $x$  il primo operando a sinistra e moltiplico e divido per  $x^2$  il secondo operando a sinistra:

$$f(\tilde{x}) - f(x) = x f'(x) \frac{(\tilde{x} - x)}{x} + x^2 \frac{f''(x) \frac{(\tilde{x} - x)^2}{2!}}{x^2}$$

5. Considerando che:

(a) Errore  $\epsilon_x$  ed  $\epsilon_x^2$ :

$$\boxed{\epsilon_x = \frac{(\tilde{x} - x)}{x}} \quad \boxed{\epsilon_x^2 = \frac{(\tilde{x} - x)^2}{x^2}}$$

6. Sostituiamo con  $\epsilon_x$  ed approssimiamo al prim'ordine ignorando  $\epsilon_x^2$ :

$$f(\tilde{x}) - f(x) = x f'(x) \epsilon_x$$

7. Dunque infine otteniamo la formula generica:

$$\boxed{\epsilon_{IN} = \sum_{i=1}^n \frac{x_i}{f(x_1, \dots, x_n)} \frac{\delta f}{\delta x_i} \epsilon_{x_i}}$$

### 2.2.5 Errore di Calcolo Funzione Irrazionale

Assumiamo di voler rappresentare in macchina  $e^x$ . E' necessario trovare una funzione che approssimi la funzione irrazionale  $e^x$ :

1. Definiamo  $e^x$  ed  $EXP(x)$ :

$$\boxed{e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}} \quad \boxed{EXP(x) = \sum_{i=0}^n \frac{x^i}{i!}}$$

2. Valutiamo quanto intercorre tra le due con l'errore di **Lagrange**:

$$e^x = EXP(x) + \frac{\epsilon_x^{(n+1)}}{(n+1)!}$$

In questo modo abbiamo stabilito che errore viene effettuato approssimando  $e^x$  con  $EXP(x)$ .

### 3 Algebra Lineare Numerica - Computazione e Con-dizionamento

Questo capitolo analizzerà strumenti base dell'algebra lineare che successivamente saranno richiesti per problemi su spazi vettoriali.

#### 3.1 Norme Vettoriali

La norma è una **funzione** che ci permette di ricavare un informazione quantitativa dato un oggetto di uno specifico spazio vettoriale.

**Definizione** Sia  $f : F^n \rightarrow \mathbb{R}$  tale che:

$$1. f(x) \geq 0 \text{ e } f(x) = 0 \text{ se e solo se } x = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

$$2. f(\alpha x) = |\alpha|f(x) \quad \forall \alpha \in F$$

$$3. f(x + y) \leq f(x) + f(y)$$

Allora  $f$  è una norma vettoriale su  $F$  e si definisce con  $\boxed{\| \cdot \|}$

##### 3.1.1 Distanza, Norma 1, Norma 2, Norma Infinito

Elenchiamo queste definizioni:

1. **Distanza:**

$$d(x, y) = \|x - y\|$$

2. **Norma 1:**

$$\|x\|_1 = \sqrt{\sum_{i=1}^n |x_i|}$$

3. **Norma 2:**

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

4. **Norma Inf\*:**

$$\|x\|_\infty = \max |x_i|$$

**Equivalenza Topologica** Date due norme  $\| \cdot \|_{(1)}$  e  $\| \cdot \|_{(2)}$  allora  $\exists \alpha, \beta \in \mathbb{R}$  tale che  $\forall v \in F^n$ :

$$\alpha \|v\|_{(2)} \leq \|v\| \leq \beta \|v\|_{(1)}$$

Da questo posso ottenere informazioni riguardo divergenza e convergenza se utilizzato "simil th. dei carabinieri".

## 3.2 Norma Matriciale

Contestualizziamo la norma alle matrici:

**Definizione** Sia  $f : F^{n \times n} \rightarrow \mathbb{R}$  tale che:

1.  $f(A) \geq 0$  e  $f(A) = 0$  se e solo se  $A = [0]_{n \times n}$
2.  $f(\alpha A) = |\alpha| f(A)$
3.  $f(A + B) = f(A) + f(B)$
4.  $f(AB) \leq f(A)f(B)$

### 3.2.1 Norma Matriciale indotta da Norma Vettoriale

\*\*\*\*\*

$$\|A\| = \max \|Av\| \quad \text{con} \quad \|v\| = 1$$

### 3.2.2 Norma di Frobanius

\*\*\*\*\*

$$\|A\|_F = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}$$

$$\|A\|_F = \text{traccia}(A^H A)^{1/2}$$

---

<sup>1</sup>La traccia in una matrice corrisponde alla somma degli elementi sulla diagonale principale. La matrice indicata con H è la trasposta coniugata, ossia matrice su cui abbiamo eseguito rispettivamente l'inversione tra righe e colonne e invertito i segni alle componenti immaginarie.

### 3.2.3 Th. Compatibilità delle Norme (con Dim.)

1. Il teorema afferma questo:

$$\boxed{\|Ax\| \leq \|A\| \|x\|}$$

**Dimostrazione** Dimostriamo il teorema:

1. **Vettore Nullo:**

$$x = 0 \Rightarrow vera, \quad 0 \leq \|A\| * 0$$

2. **Vettore Strettamente Positivo:**

- (a) Definizione di **norma matriciale indotta**:

$$\|A\| = \max_{\|z\|=1} \|Az\|$$

- (b) Portiamo fuori  $\frac{1}{\|v\|}$  grazie alla proprietà 2 delle norme matriciali e moltiplichiamo a sx e dx:

$$\|v\| \frac{1}{\|v\|} \|Av\| \leq \|A\| * \|v\|$$

- (c) Risolviamo i calcoli ed otteniamo:

$$\boxed{\|Av\| \leq \|A\| \|v\|}$$

### 3.2.4 Metodi Iterativi su Norme

Elenchiamo le caratteristiche del calcolo iterativo delle norme:

1. **Norma 1:** Somma delle **colonne**, ottengo un vettore, da questo prendo il massimo:

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}| \quad \text{con } a_{ij} \in A$$

2. **Norma Infinito:** Somma delle **righe**, ottengo un vettore, da questo prendo il massimo:

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}| \quad \text{con } a_{ij} \in A$$

3. **Norma 2:** Assumendo  $\varphi$  sia raggio spettrale della matrice in questione, dove:

$$\varphi(A) = \max_{i=1 \dots n} |\lambda_i|$$

$$\boxed{\|A\|_2 = \sqrt{\varphi(A^H A)}}$$



### 3.2.5 Matrici Simmetriche sui Reali

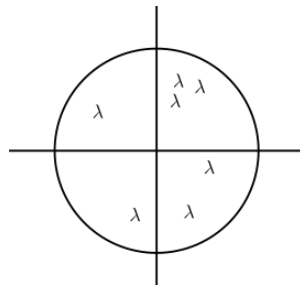
Elenchiamo proprietà caratterizzanti delle matrici simmetriche che verranno citate successivamente:

1.  $A = A^T$
2. Le matrici simmetriche sui reali:
  - (a) Sono **diagonalizzabili**
  - (b) Hanno **autovalori reali**
  - (c)  $(A^T A)^T = (A A^T)$
  - (d)  $\|A\|_2 = \varphi(A)$

### 3.2.6 Th. di Hirsch (con Dim.)

**Definizione** Se  $\|\cdot\|$  è una norma matriciale indotta, allora:

$$|\lambda^{(A)}| < \|A\|$$



Ossia ogni autovalore deve essere inferiore alla norma matriciale indotta. E' come se stessimo "localizzando" la posizione degli autovalori.

**Dimostrazione** Dimostriamo il teorema:

1. Dalla definizione di autovalore:

$$A v = \lambda v \quad \text{con} \quad \lambda \neq 0$$

2. Applico la norma a sx e dx ed inverto l'ordine:

$$\|\lambda v\| = \|A v\|$$

3. Proprietà (2.) e (4.) delle norme matriciali:

$$|\lambda| \|v\| = \|A v\| \leq \|A\| \|v\|$$

4. Considero dunque primo e terzo termine, dividendo a sx e dx per  $\|v\|$ :

$$\boxed{|\lambda| \leq \|A\|}$$

### 3.3 Utilities Greshgorin

Elenchiamo tutte gli oggetti e funzioni definiti sui cerchi di Gershgorin:

#### 3.3.1 Cerchio i-esimo di Greshgorin

Definiamo un cerchio come luogo geometrico dei punti, dato che successivamente sarà necessario alla localizzazione degli autovalori.

**Definizione** Sia  $K_i$  dove

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

1.  $a_{ii}$  corrisponde al **centro** del cerchio.
2.  $\sum_{j=1, j \neq i}^n |a_{ij}|$  corrisponde al **raggio** del cerchio.

#### 3.3.2 Th. di Greshgorin (con Dim.)

Il teorema di Greshgorin afferma che se un arbitrario  $\lambda$  è un autovalore, allora questo deve essere all'interno dell'unione dei cerchi di Greshgorin della matrice.

**Definizione** Se  $\lambda$  è **autovalore** di  $A \Rightarrow \lambda \in \bigcup_{i=1}^n K_i$

Spesso questo teorema viene utilizzato "al contrario", ossia se un valore **non** è all'interno dell'unione dei cerchi, allora sicuramente **non** è un autovalore della matrice.

**Dimostrazione** Dimostriamo il teorema:

1. Dalla definizione di autovalore:

$$A v = \lambda v$$
$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ v_n \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ v_n \end{bmatrix}$$

2. Assumo di aver effettuato il prodotto ad sx tra  $A$  e  $v$ :

$$\sum_{j=1}^n a_{ij} v_j = \lambda v_i \quad \forall i \in \{1, \dots, n\}$$

3. Tiro fuori dalla sommatoria il termine per  $i = j$ , lo porto a sinistra e metto in evidenza  $v_i$ :

$$\sum_{j=1, j \neq i}^n (a_{ij} v_j) = (\lambda - a_{ii}) v_i$$

4. Sapendo che:

$$\boxed{|\lambda - a_{ii}| |v_i| = |(\lambda - a_{ii})v_i|} \quad \text{e} \quad \boxed{\left| \sum_{j=1}^n a_{ij} v_j \right| \leq \sum_{j=1}^n |a_{ij}| |v_j|}$$

5. Prendo un  $p$  tale che:

$$0 \leq |v_p| = \|v\| = \max_{j=1..n} |v_j|$$

Ossia  $p$  deve fare in modo che la norma di  $v$  deve essere uguale al massimo delle componenti del vettore  $v$ . Assumiamo di non star lavorando sul vettore nullo grazie alla definizione di autovalore  $\lambda$ .

6. Istanzio il punto (3.) con la  $p$  appena definita in (5.):

$$|\lambda - a_{pp}| |v_p| \leq \sum_{j=1, j \neq i}^n |a_{pj}| |v_j|$$

7. Divido a sx e dx per  $|v_p|$ :

$$|\lambda - a_{pp}| \frac{|v_p|}{|v_p|} \leq \sum_{j=1, j \neq i}^n |a_{pj}| \frac{|v_j|}{|v_p|}$$

Sappiamo che  $\frac{|v_j|}{|v_p|} \leq 1$  perchè in un vettore possiamo avere più massimi, quindi componenti max con stessi valori.

8. Riesco dunque ad effettuare una maggiorazione grazie alle affermazioni precedenti:

$$\sum_{j=1, j \neq i}^n |a_{pj}| \frac{|v_j|}{|v_p|} \leq \sum_{j=1, j \neq i}^n |a_{pj}|$$

9. La sommatoria ottenuta nella maggiorazione del punto (8.) rispetta la definizione di *Cerchio  $i$ -esimo di Gershgorin*, di conseguenza abbiamo dimostrato il teorema.  $\square$

$$\boxed{\lambda \in K_p}$$

### 3.3.3 Invertibilità e Predominanza Diagonale

**Matrice Invertibile** Elenchiamo un paio di caratteristiche sull'invertibilità di matrici:

1. Se  $A$  è invertibile, allora:

(a)  $\det(A) \neq 0$

(b)  $\text{rango}(A) = n$

(c)  $\dim(\ker(A)) = 0$

(d) 0 non è un autovalore

(e)  $P(x) = \det(A - xI) = (x - \lambda_1)(x - \lambda_2) \dots (x - \lambda_n)$

(f)  $P(0) = \det(A) = (x - \lambda_1)(x - \lambda_2) \dots (x - \lambda_n) = \text{prodotto degli autovalori } \lambda_i$

**Predominanza Diagonale di Matrice per Riga** La predominanza vale quando in valore assoluto, l'elemento sulla diagonale principale è maggiore di tutti gli altri sulla riga, formalmente:

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

**Corollario** Se  $A$  è a **predominanza diagonale** allora  $A$  è **invertibile**. La dimostrazione di questo corollario si basa sul fatto che la definizione appena data si basa sul *Cerchio di Gershgorin* ma utilizzato al contrario, ossia stiamo affermando di **non** essere nel cerchio. Questo però ci porta ad essere esattamente al contrario rispetto \*\*\*\*.

### 3.3.4 II Th. di Gershgorin

**Definizione** Se l'unione  $M_1$  di  $k$  cerchi è **disgiunta** dall'unione  $M_2$  di  $(n - k)$  cerchi allora  $k$  autovalori appartengono ad  $M_1$  ed  $(n - k)$  ad  $M_2$ .

## 3.4 Condizionamento del Problema sulla Risoluzione di Sistemi Lineari

Assumiamo di avere una matrice  $A$  ed un vettore  $b$ ,  $Ax = b$ , sapendo che se  $A$  è invertibile allora  $x = A^{-1}b$ , altrimenti o non ha soluzioni o ne ha infinite. Elenchiamo come approssimeremo questi oggetti in oggetti discreti:

1. **Matrice**  $A$  ed i suoi **elementi**  $a_{ij}$  :

$$\tilde{A} = A + \Delta A \quad \tilde{a}_{ij} = a_{ij} + \epsilon f_{ij}$$

dove  $f_{ij} = (1 + \epsilon_{ij})$ , ossia l'errore su ogni componente, e  $\Delta A$  li contiene tutti.

2. **Vettore**  $b$  ed i suoi **elementi**  $b_i$  :

$$\tilde{b} = b + \delta b \quad \tilde{b}_i = b_i + (1 + f_i)$$

Il **condizionamento** verrà quindi calcolato su **norme**:

$$\frac{\|\tilde{x} - x\|}{\|x\|}$$

Vogliamo dunque esprimere la risoluzione in questa forma:

$$x = A^{-1}b$$

### 3.4.1 Teorema sul Condizionamento di Norme Matriciali (con dim)

**Definizione** Sia  $A$  invertibile e  $b \neq 0$ , allora:

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\tilde{b} - b\|}{\|b\|}$$

Dove  $\|A\| \|A^{-1}\|$  è detto **numero di condizionamento** di  $A$ .

**Dimostrazione** Dimostriamo questo teorema:

1. Partiamo dalla definizione di  $x$  ed  $\tilde{x}$ :

$$x = A^{-1}b \rightarrow Ax = b$$

$$\tilde{x} = A^{-1}\tilde{b} \rightarrow A\tilde{x} = \tilde{b}$$

2. Sostituiamo (1.) in  $\|\tilde{x} - x\|$ :

$$\|A^{-1}\tilde{b} - A^{-1}b\|$$

3. Raccolta di  $A^{-1}$  e compatibilità di norme a motivazione del  $(\leq)$ :

$$\|\tilde{x} - x\| = \|A^{-1}(\tilde{b} - b)\| \leq \|A^{-1}\| \|\tilde{b} - b\|$$

4. Scriviamo la forma standard di risoluzione di sistema lineare applicando le norme a  $sx$  e  $dx$ , e utilizziamo anche qui la compatibilità delle norme per  $(\leq)$  a  $dx$ :

$$\|Ax\| = \|b\| \rightarrow \|b\| = \|Ax\| \leq \|A\| \|x\|$$

5. Da questo possiamo dunque ricavare che:

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

6. Tornando dunque al punto (3.) possiamo dividere  $\|\tilde{x} - x\|$  per  $\|x\|$  ed a  $dx$  del  $\leq$  dividiamo per  $\frac{\|b\|}{\|A\|}$  rispettando quindi la maggiorazione (dato il punto precedente).

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\tilde{b} - b\|}{\frac{\|b\|}{\|A\|}}$$

7. Avendo concluso la dimostrazione definiamo  $\mu$ , il **numero di condizionamento** di  $A$ :

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

## 4 Metodi Diretti per Risoluzione di Sistemi Lineari

Assumiamo di avere  $Ax = b$ ,  $A \in \mathbb{R}^{n \times n}$ ,  $\det(A) \neq 0$ ,  $x = A^{-1}b$ , allora possiamo avere diversi casi:

1. **Matrice Diagonale:** La matrice  $A$  ha solo elementi sulla sua diagonale.

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & & \dots \\ \dots & & a_{33} & 0 \\ 0 & \dots & 0 & a_{nn} \end{bmatrix}$$

In questo caso possiamo ricavare le **soluzioni** in questo modo

$$x = A^{-1}b = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & & \dots \\ \dots & & \frac{1}{a_{33}} & 0 \\ 0 & \dots & 0 & \frac{1}{a_{nn}} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_n \end{bmatrix}$$

Ricordiamo che il costo di questa risoluzione risulta essere  $O(n)$ .

2. **Matrice Triangolare:** La matrice  $A$  è diagonale e di conseguenza gli *autovalori* sono gli elementi sulla diagonale. Dunque possiamo calcolare il determinante in questo modo:

$$\begin{bmatrix} a_{11} & \dots & \dots & a_{1n} \\ 0 & a_{22} & & \dots \\ \dots & & a_{33} & \dots \\ 0 & \dots & 0 & a_{nn} \end{bmatrix}$$

$$\det(A) = \prod_{i=1}^n a_{ii}$$

In questo caso le **soluzioni** si ottengono grazie al **metodo di sostituzione**. (Il metodo di sostituzione in avanti o indietro in base a se la matrice risulta triangolare superiore o inferiore). Il costo di questa risoluzione risulta essere  $O(n^2)$ .

3. **Matrice Piena:** La matrice  $A$  risulta piena:

$$\begin{bmatrix} a_{11} & \dots & \dots & a_{1n} \\ \dots & a_{22} & & \dots \\ \dots & & a_{33} & \dots \\ a_{n1} & \dots & \dots & a_{nn} \end{bmatrix}$$

Non si conosce un algoritmo che sia aderente al limite inferiore  $O(n^2)$  del problema. Di conseguenza l'algoritmo favorito in queste circostanze è quello di **Gauss**, caratterizzato da un costo in tempo asintotico  $O(n^3)$ , dato che bisogna, per ogni colonna, azzerare tutti gli elementi sotto la diagonale.

## 4.1 Fattorizzazione LU

Grazie al Th. di Gauss riusciamo ad ottenere anche una nuova **formulazione** di matrici piene sotto specifiche ipotesi.

**Definizione di Fattorizzabile** Una matrice  $A \in \mathbb{R}^{n \times n}$  è *fattorizzabile LU* se

1. Esiste  $L$  matrice triangolare inferiore con elementi diagonali uguali ad 1.
2. Esiste  $U$  matrice triangolare superiore.

Tale che

$$\boxed{A = LU}$$

## 4.2 Th. Condizioni Sufficienti per Esistenza ed Unicità Fattorizzazione LU (con Dim.)

Assumiamo di avere una matrice quadrata  $A \in \mathbb{R}^{n \times n}$ , definiamo con  $A_k$  le sue sottomatrici quadrate di dimensione  $k * k$ . Questo darà contesto alla definizione formale del teorema.

**Definizione** Sia  $A \in \mathbb{R}^{n \times n}$  se  $\det(A_k) \neq 0 \forall k \in \{1, \dots, n-1\}$  allora esiste la *fattorizzazione LU* di  $A$ .

**Dimostrazione** Procediamo a dimostrare per induzione questo teorema:

1. **Caso Base:** Prendiamo  $k = 1$ , quindi  $A = \begin{bmatrix} a_{11} \end{bmatrix}$  dunque:

$$L = \begin{bmatrix} 1 \end{bmatrix} \quad U = \begin{bmatrix} a_{11} \end{bmatrix} \quad \text{allora} \quad A = LU$$

2. **Caso Induttivo:** Assumiamo che la proprietà sia vera sulle matrici di dimensione  $n - 1$  (Ipotesi Induttiva).

(a) Vediamo le matrici  $A, L, U$  a blocchi rispettivamente in questo modo:

$$\left[ \begin{array}{c|c} A_{n-1} & x \\ \hline y^T & a_{nn} \end{array} \right] = \left[ \begin{array}{c|c} L_{n-1} & 0 \\ \hline w & 1 \end{array} \right] \left[ \begin{array}{c|c} U_{n-1} & z \\ \hline 0^T & \beta \end{array} \right]$$

- (b) Consideriamo ogni elemento della matrice  $A$  come prodotto dei blocchi delle matrici  $L$  ed  $U$ :

i. Prodotto prima riga di  $L$  prima colonna di  $U$ :

$$A_{n-1} = L_{n-1}U_{n-1} + 0 \cdot 0^T$$

Rimuovendo gli zeri, otteniamo questo primo blocco di  $A$  valido per ipotesi induttiva.

ii. Prodotto prima riga di  $L$  seconda colonna di  $U$ :

$$x = L_{(n-1)}z + 0\beta$$

La matrice  $L_{(n-1)}$  è valida per ipotesi induttiva, esisterà almeno una  $z$  per cui  $z = L_{n-1}^{-1}x$ .

iii. Prodotto seconda riga di  $L$  prima colonna di  $U$ :

$$y^T = w^T U_{n-1} + 1\ 0^T$$

Trasponendo otteniamo:

$$y = U_{n-1}^T w$$

iv. Prodotto seconda riga di  $L$  seconda colonna di  $U$ :

$$a_{nn} = w^T z + 1\beta$$

### 4.3 Matrici Elementari di Gauss

Definiamo la matrice  $E$ :

$$E = I - v e_k^T \quad \text{con} \quad e_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{e} \quad v_1 = v_2 = v_k$$

$E$  è quindi definita come **Matrice di Gauss**.

**Proprietà** Questo tipo di matrice gode di diverse proprietà caratteristiche:

1. Queste matrici sono **triangolari inferiori** con tutti 1 sulla diagonale e sono **invertibili**.
2. Vale che:

$$E^{-1} = I + v e_k^T$$

Questo statement si può dimostrare effettuando la moltiplicazione tra  $E$  ed  $E^{-1}$  ottenendo  $I$ .

3. Sia  $x \in \mathbb{R}$  con  $x_k \neq 0$ . Allora esiste una matrice elementare di Gauss tale che

$$Ex = \begin{bmatrix} x_1 & \dots & x_k & 0 & \dots & 0 \end{bmatrix}^T$$



4. Assumendo di avere due matrici elementari di Gauss:

$$\boxed{E = I - ve_k^T} \quad \boxed{\bar{E} = I - we_l^T}$$

allora

$$\boxed{E \bar{E} = I_n - ve_k^T - we_l^T}$$

che informalmente vuol dire che il prodotto tra le due matrici elementari di Gauss viene costruito posizionando semplicemente nella posizione corretta i due vettori  $v$  e  $w$  di fattori.

5. Il prodotto di  $Ey$ , dove  $E = I - ve_k^T$  è una matrice elementare di Gauss ed  $y$  un vettore, può essere calcolato in  $n - k$  operazioni moltiplicative, infatti ponendo  $Ey = z$  otteniamo che  $z_j = y_j$  per  $1 \leq j \leq k$  mentre  $z_j = y_j - v_j y_k$ .

Il metodo di Gauss può essere utilizzato senza scambio di righe se e solo se  $a_{kk}^{k-1} \neq 0 \quad \forall k = 1 \dots n$ , ossia se tutti i pivot risultano essere diversi da 0.

**Teorema** Dato un  $A \in \mathbb{R}$ :

$$A(1 : k, 1 : k) \neq 0 \Leftrightarrow p_{kk}^{k-1} \neq 0 \quad \forall k \in \{1, \dots, n-1\}$$

E quindi, considerando che  $Ly = b$ :

$$[A|B] \rightarrow_{E_1} \dots \rightarrow_{E_2} \dots \rightarrow_{E_{n-1}} = [U|y]$$

#### 4.4 Tecniche di Pivoting

Assumiamo di avere un pivot pari a 0, è necessario che si generi una **permutazione** della corrente matrice per fare in modo che il pivot in questione non sia nullo.

$$\begin{bmatrix} a_{11} & * & * & * \\ 0 & 0_{k+1} & & \dots \\ \dots & \dots & \dots & \dots \\ 0 & a_{nk} & 0 & * \end{bmatrix}$$

Questo mi causa però la **fattorizzazione LU** di una permutazione della matrice  $A$  e non di  $A$  stessa.

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

**Stabilità e Pivoting** Le tecniche di pivoting non sono utilizzate solo nel caso in cui un pivot risulti nullo, ma anche per questioni di stabilità degli algoritmi causati.

Si può dunque dimostrare che i fattori  $\tilde{L}$  e  $\tilde{U}$  calcolati sono tali per cui  $\tilde{L}\tilde{U} = A + E$ , dove  $E$  corrisponde all'errore in rappresentazione in numeri di macchina. Vale dunque che:

$$\frac{\|E\|}{\|\tilde{L}\| \|\tilde{U}\|} = O(u)$$