CALCOLO NUMERICO

Corso A

Autore
Giuseppe Acocella
2024/25

Ultima Compilazione - February 27, 2025

Contents

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 $\overline{x}$. Definiamo dunque i due errori ε_{IN} ed ε_x:

1. **Errore Inerente** (ε_{IN}): Assumendo una funzione f:

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

2. **Errore di Approssimazione** (ε_x): Assumendo una funzione f:

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

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 β di rappresentazione **esistono** e **sono unici**:

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

tali che:

$$x \;=\; segno(x)\, \beta^{\rho}\, (\sum_{i=1}^{\infty} d_i \beta^{-i})$$

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 $0.\overline{9}$ convergerebbe ad 1.

1.5 Insieme di Numeri di Macchina

Definiamo l'insieme Φ che permette la rappresentazione dei numeri di macchina:

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

1. β: **base**.
2. t cifre della **mantissa**.
3. $-n \leq \rho \leq M$, ossia i due **estremi** che contengono l'**esponente** ρ.
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** ρ.
4. $(\beta - 1)$ tutte le possibili **configurazioni** delle **cifre** rispetto alla base (escluso lo zero).
5. (β^{t-1}) avendo t **bit disponibili** e β 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 ω **più piccolo** e del numero Ω più grande.

1. Numero **più piccolo** rappresentabile ω:

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

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

2. Numero **più grande** rappresentabile Ω:

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

   Questo perchè vogliamo la **base** β 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 Φ:

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

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 ∞.

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

1. Se $\rho > M$ allora è **overflow**, e ad esempio in Matlab questo comportamento viene approssimato ad ∞.
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** (ϵ):
$$\epsilon = \tilde{x} - x$$
2. **Errore Relativo** (ϵ_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:

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

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

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

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 approcciamo 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 ϵ_{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:

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

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

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:

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

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

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

   Questa formula di ϵ_{TOT} si basa su una generica operazione

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

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

Calcolo dell'errore totale ϵ_{TOT}:

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

dove:

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

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

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

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

Dimostrazione

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

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

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

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

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:

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

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

$$\boxed{C_{amplificazione} = \frac{x}{f(x)}}$$

1. **Ben Condizionato**: Non ho punti nel dominio del coefficiente di amplificazione dove la funzione va a ∞.
2. **Mal Condizionato**: Ho dei punti in cui il coefficiente può andare a ∞. 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:

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

5. Considerando che:

   (a) Errore ϵ_x ed ϵ_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 ϵ_x ed approssimiamo al prim'ordine ignorando ϵ_x^2:

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

7. Dunque infine otteniamo la formula generica:

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

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 $EXP(x)$:

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

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 Condizionamento

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 $f : F^n \to \mathbb{R}$ tale che:

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

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***:
$$||x||_\infty = nax|x|$$

Equivalenza Topologica Date due norme $|| \, ||_{(1)}$ e $|| \, ||_{(2)}$ allora $\exists \, \alpha, \beta \in F$ 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 $f : F^{nxn} \to \mathbb{R}$ tale che:

1. $f(A) \geq 0$ e $f(A) = 0$ se e solo se $A = [0]_{nxn}$
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

****

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

3.2.2 Norma di Frobanius

****

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

$$||A||_F = traccia(A^H A)^1$$

[1]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**:

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

2. **Vettore Strettamente Positivo**:

   (a) Prendo vettore x e lo normalizzo:

   $$\frac{x}{||x||}$$

   (b) Prendo la norma del risultato e la pongo uguale alla norma di un ipotetico y:

   $$||y|| = ||\frac{x}{||x||}|| = \frac{1}{||x||} * ||x|| = 1$$

   (c) Seguendo la catena di uguaglianze precedenti ricaviamo:

   $$||\frac{x}{||x||}|| = 1$$

   (d) ******

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:

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

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

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

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

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

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

3.2.5 Matrici Simmetriche sui Reali

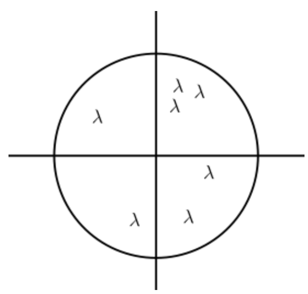
Elenchiamo proprietà caratterizzanti delle matrici simmetriche che verranno citate successivamente:

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 Hirch (con Dim.)

Definizione Se $||\,.\,||$ 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:

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

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

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

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

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

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

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

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 λ è un autovalore, allora questo deve essere all'interno dell'unione dei cerchi di Greshgorin della matrice.

Definizione $\boxed{\text{Se } \lambda \text{ è } \textbf{autovalore} \text{ di } A \Rightarrow \lambda \in \bigcup_{i=1}^{n} k_i}$

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

2. Assumo di aver effettuato il prodotto ad sx tra A e v:
$$\sum_{j=1}^{n} a_{ij} v_j = \lambda v_i \quad \forall\, i \in \{1\,, ...,\, n\}$$

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:

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

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 λ.

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. □

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

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) $rango(A) = 0$
   (c) $dim(ker(A)) = 0$
   (d) 0 non è un autovalore
   (e) $P(x) = det(A - xI) = (x - \lambda_1)(x - \lambda_2)....(x - \lambda_n)$
   (f) $P(0) = det(A) = (x - \lambda_1)(x - \lambda_2)....(x - \lambda_n) =$ prodotto degli autovalori λ_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} :

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

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

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

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

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* (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 $(\leq)$:

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

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:

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

4 Metodi Diretti per Risoluzione di Sistemi Lineari

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

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

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

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

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

   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:

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

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

   In questo caso le **soluzioni** si ottengono grazie al **metodo di sostituzione**. (Il metodo di sostituizione in avanti o in 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:

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

   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 $A \in \mathbb{R}^{nxn}$ 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 $A \in \mathbb{R}^{nxn}$, definiamo con A_k le sue sottomatrici quadrate di dimensione $k * k$. Questo darà contesto alla definizione formale del teorema.

Definizione Sia $A \in \mathbb{R}^{nxn}$ se $det(A_k) \neq 0 \, \forall k \in \{1, ..., n-1\}$ 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:

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

      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:

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

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:

$$E^{-1} = I + ve_k^T$$

Il metodo di Gauss può essere utilizzato senza scambio di righe se e solo se $a_{kk}^{k-1} \neq 0 \;\; \forall k = 1 \ldots n$.

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

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

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:

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