RISOLUZIONE DI SISTEMI LINEARI: METODI ITERATIVI (Relazione tra residuo ed errore e principi dei metodi iterativi)

In questa lezione introdurremo il concetto di residuo per un sistema lineare e analizzeremo la relazione fondamentale tra residuo ed errore. Successivamente, esamineremo i principi che guidano la costruzione dei metodi iterativi e, infine, presenteremo alcuni esempi concreti di tali metodi.
Iniziamo osservando che esiste una relazione algebrica semplice tra l'errore commesso nella risoluzione di un sistema lineare e il residuo ad esso associato. Consideriamo un generico sistema lineare

$$A\boldsymbol{x} = \boldsymbol{b}.$$

Sia $\boldsymbol{x}^*$ la soluzione approssimata ottenuta mediante un generico metodo iterativo e sia $\boldsymbol{r}^*$ il corrispondente residuo, definito come il vettore

$$\boldsymbol{r}^* = \boldsymbol{b} - A\boldsymbol{x}^*.$$

Dunque, prendiamo il termine noto $\boldsymbol{b}$ e sottraiamo il vettore di $A\boldsymbol{x}^*$ (ovvero il prodotto matrice A per vettore $\boldsymbol{x}^*$, e quindi ancora un vettore). Definiamo la differenza $\boldsymbol{b} - A\boldsymbol{x}^*$ come il vettore residuo. Se $\boldsymbol{x}^*$ coincidesse con la soluzione esatta $\boldsymbol{x}^* = \boldsymbol{x}$, allora il residuo risulterebbe nullo

$$\boldsymbol{r}^* = \boldsymbol{b} - A\boldsymbol{x}^* = 0,$$

perché la soluzione esatta soddisfa il problema $A\boldsymbol{x} = \boldsymbol{b}$. Pertanto, sulla soluzione esatta il residuo è pari a zero. Tuttavia, quando la soluzione approssimata differisce da quella esatta, il residuo sarà generalmente diverso da zero. Il nostro obiettivo è quindi analizzare la relazione tra il residuo e l'errore. Ricordando che l'errore è definito come la differenza tra la soluzione esatta $\boldsymbol{x}$ e la soluzione calcolata (o approssimata) $\boldsymbol{x}^*$

$$\boldsymbol{e}^* = \boldsymbol{x} - \boldsymbol{x}^*,$$

possiamo ricavare che

$$\boldsymbol{r}^* = \boldsymbol{b} - A\boldsymbol{x}^* = A\boldsymbol{x} - A\boldsymbol{x}^* = A(\boldsymbol{x} - \boldsymbol{x}^*) = A\boldsymbol{e}^*,$$

dove abbiamo sostituito $\boldsymbol{b} = A\boldsymbol{x}$, da cui

$$\boldsymbol{r}^* = A\boldsymbol{e}^*,$$

Dunque, possiamo trovare il residuo, applicando la matrice A all'errore rispetto alla soluzione approssimata $e^* = \boldsymbol{x} - \boldsymbol{x}^*$.
Inoltre, invertendo la matrice A posso trovare l'errore applicando A^{-1} al residuo

$$\boldsymbol{e}^* = A^{-1}\boldsymbol{r}^*.$$

Abbiamo dunque trovato la relazione che intercorre fra residuo ed errore.
Vogliamo sfruttare questa relazione per ottenere stime dell'errore in funzione del residuo, che saranno poi utilizzate nell'analisi dei metodi iterativi. Considerando la norma della relazione tra errore e residuo $\boldsymbol{e}^* = A^{-1}\boldsymbol{r}^*$, otteniamo

$$\|\boldsymbol{e}^*\| = \|A^{-1}\boldsymbol{r}^*\|.$$

Per la proprietà delle norme abbiamo che (la norma del prodotto è minore uguale al prodotto delle norme)

$$\|\boldsymbol{e}^*\| = \|A^{-1}\boldsymbol{r}^*\| \leq \|A^{-1}\|\|\boldsymbol{r}^*\|.$$

Moltiplicando e dividendo per la norma di A otteniamo

$$\|\boldsymbol{e}^*\| = \|A^{-1}\boldsymbol{r}^*\| \leq \|A^{-1}\|\|A\| \frac{\|\boldsymbol{r}^*\|}{\|\boldsymbol{A}\|},$$

e ricordando che la norma di A per la norma di A^{-1} è il numero di condizionamento di A

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

troviamo che

$$\|\boldsymbol{e}^*\| = \|A^{-1}\boldsymbol{r}^*\| \leq \|A^{-1}\|\|A\| \frac{\|\boldsymbol{r}^*\|}{\|A\|} = K(A) \frac{\|\boldsymbol{r}^*\|}{\|A\|}.$$

Abbiamo quindi trovato che l'errore in norma è minore o uguale del numero di condizionamento di A per la norma del residuo, diviso per la norma di A

$$\|\boldsymbol{e}^*\| \leq K(A) \frac{\|\boldsymbol{r}^*\|}{\|A\|}.$$

Se volessimo tradurre questa relazione per l'errore relativo

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}\|} = \frac{\|\boldsymbol{e}^*\|}{\|\boldsymbol{x}\|},$$

basterebbe dividere la relazione precedentemente trovata $\|\boldsymbol{e}^*\| \leq K(A)\frac{\|\boldsymbol{r}^*\|}{\|A\|}$ per $\|\boldsymbol{x}\|$ (ovvero, la norma della soluzione esatta) a sinistra e a destra, ottenendo

$$\frac{\|\boldsymbol{e}^*\|}{\|\boldsymbol{x}\|} \leq K(A) \frac{\|\boldsymbol{r}^*\|}{\|A\|\|\boldsymbol{x}\|}.$$

Ricordando che $A\boldsymbol{x} = \boldsymbol{b}$ e quindi $\|A\boldsymbol{x}\| = \|\boldsymbol{b}\|$ da cui $\|\boldsymbol{b}\| = \|A\boldsymbol{x}\| \leq \|A\|\|\boldsymbol{b}\|$ e dato che questa quantità è a denominatore possiamo usarla come maggiorazione dell'ultima relazione trovata, ovvero

$$\frac{\|\boldsymbol{e}^*\|}{\|\boldsymbol{x}\|} \leq K(A) \frac{\|\boldsymbol{r}^*\|}{\|A\|\|\boldsymbol{x}\|} \leq K(A) \frac{\|\boldsymbol{r}^*\|}{\|\boldsymbol{b}\|}.$$

Quindi l'errore relativo è minore o uguale del numero di condizionamento per il residuo relativo $\frac{\|\boldsymbol{r}^*\|}{\|\boldsymbol{b}\|}$

$$\frac{\|\boldsymbol{e}^*\|}{\|\boldsymbol{x}\|} \leq K(A) \frac{\|\boldsymbol{r}^*\|}{\|\boldsymbol{b}\|}.$$

Infatti, possiamo associare alla norma di $\boldsymbol{b}$ una misura della dimensione del residuo e quindi questo possiamo interpretare $\frac{\|\boldsymbol{r}^*\|}{\|\boldsymbol{b}\|}$ come residuo relativo.

Il numero di condizionamento della matrice A svolge, anche in questo contesto, il ruolo di possibile amplificatore del residuo nella determinazione dell'errore. In conclusione, se $K(A)$ è relativamente piccolo e, quindi, se la matrice A è ben condizionata, a un residuo di piccola entità corrisponderà un errore ridotto. Questo risultato è particolarmente incoraggiante, poiché in molti casi avremo la possibilità di controllare quantitativamente i residui, mentre, in generale, l'errore sulla soluzione non è direttamente accessibile, dato che la soluzione esatta non è nota a priori.

Di conseguenza, se il residuo è noto e la matrice è ben condizionata, si può immediatamente concludere che anche l'errore sarà contenuto, a condizione che il residuo sia piccolo. Al contrario, se la matrice A è mal condizionata, ossia se $K(A)$ assume valori elevati, può accadere che, pur in presenza di un residuo di piccola entità, l'errore risulti comunque grande.

Vediamo a questo proposito un esempio su un semplice sistema 2×2

$$A\boldsymbol{x} = \boldsymbol{b},$$

dove

$$A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 0.8642 \\ 0.1440 \end{bmatrix}.$$

La soluzione di questo sistema è facilmente ricavabile e risulta

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}.$$

Supponendo di aver calcolato una soluzione $\boldsymbol{x}^*$ uguale a

$$\boldsymbol{x}^* = \begin{bmatrix} 0.991 \\ -0.487 \end{bmatrix},$$

l'errore assoluto che si commette è un errore significativo

$$\boldsymbol{e}^* = \begin{bmatrix} 1.01 \\ -1.52 \end{bmatrix},$$

e quindi l'errore relativo

$$\frac{\|\boldsymbol{e}^*\|_2}{\|\boldsymbol{x}\|_2} \approx 39\%,$$

quindi un errore grande. Invece, il residuo associato a $\boldsymbol{x}^*$ è estremamente piccolo

$$\boldsymbol{r}^* = \boldsymbol{b} - A\boldsymbol{x}^* = \begin{bmatrix} -10^{-8} \\ 10^{-8} \end{bmatrix}.$$

Dunque, in corrispondenza della soluzione $\boldsymbol{x}^*$, a un residuo estremamente piccolo può corrispondere un errore significativo. Questo rappresenta un caso in cui, anche in un sistema 2×2, si osserva come residui di piccola entità possano essere associati a errori decisamente elevati. Sulla base della teoria generale sviluppata in precedenza, ci si aspetta che la causa di questo comportamento sia il mal condizionamento della matrice A. In effetti, in questo caso A^{-1} si può calcolare facilmente a mano ed ha questa espressione

$$A^{-1} = \begin{pmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{pmatrix} 10^8.$$

Quindi, il numero di condizionamento di A, calcolato nella norma infinito, è

$$K_\infty(A) = \|\|A^{-1}\|_\infty \|A\|\|_\infty \approx 3.3 \cdot 10^8.$$

Abbiamo dunque individuato un esempio di matrice fortemente mal condizionata, per la quale è possibile determinare un opportuno termine noto $\boldsymbol{b}$ tale che la soluzione del sistema associato presenti un residuo estremamente piccolo, mentre l'errore commesso risulti invece significativamente elevato.

Esaminiamo ora i principi fondamentali che guidano la costruzione dei metodi iterativi. Ricordiamo che l'obiettivo dei metodi iterativi è quello di generare una successione di vettori $\boldsymbol{x}^{(k)}$ tali che il limite per k che tende all'infinito sia uguale a $\boldsymbol{x}$, essendo $\boldsymbol{x}$ la soluzione esatta del sistema $A\boldsymbol{x} = \boldsymbol{b}$

$$\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}.$$

Dal punto di vista pratico, non è desiderabile proseguire le iterazioni k indefinitamente, ma si preferisce arrestare il processo al minimo indice $k = n$, per cui l'errore risulti inferiore o uguale a una tolleranza ϵ fissata a priori

$$\left\|\boldsymbol{x}^{(n)} - \boldsymbol{x}\right\| \leq \epsilon.$$

Pertanto, possiamo decidere di fermarci non appena l'errore raggiunga un valore minore o uguale a una tolleranza predefinita, che potrebbe essere ad esempio 10^{-2}, 10^{-3}, 10^{-6}.

Vediamo ora come costruire metodi iterativi, ossia quale formula ci permette di generare vari tipi di metodi. La strategia chiave consiste nello scomporre la matrice A del sistema come differenza di due matrici P ed N

$$A = P - N.$$

È importante precisare che tale approccio non corrisponde a una fattorizzazione. Nella fattorizzazione, infatti, A viene rappresentata come prodotto di matrici, ad esempio nella fattorizzazione LU, mentre qui A viene espressa come differenza di due matrici.
L'unica ipotesi che facciamo è che P sia non singolare. Allora, da $A\boldsymbol{x} = \boldsymbol{b}$, e considerando la scomposizione $A = P - N$, segue che

$$A\boldsymbol{x} = (P - N)\boldsymbol{x} = P\boldsymbol{x} - N\boldsymbol{x} = \boldsymbol{b},$$

e dunque

$$P\boldsymbol{x} = N\boldsymbol{x} + \boldsymbol{b},$$

dove nell'ultimo passaggio abbiamo semplicemente portato $N\boldsymbol{x}$ a secondo membro. In questa relazione $P\boldsymbol{x} = N\boldsymbol{x} + \boldsymbol{b}$, troviamo la soluzione del sistema sia a sinistra che a destra dell'uguale. Pertanto, possiamo pensare di impostare un processo iterativo, in cui, dato un valore iniziale $\boldsymbol{x}^{(0)}$ scelto arbitrariamente, si possa generare una successione di vettori $\boldsymbol{x}^{(k)}$ così costruita

Assegnato $\boldsymbol{x}^{(0)}$, si generi una successione $\{\boldsymbol{x}^{(k)}\}$ risolvendo

$$P\boldsymbol{x}^{(\boldsymbol{k})} = N\boldsymbol{x}^{(\boldsymbol{k-1})} + \boldsymbol{b}, \qquad k \geq 1,$$

Quindi, supponendo noto $\boldsymbol{x}^{(0)}$, per $k = 1$, calcoliamo

$$P\boldsymbol{x}^{(\boldsymbol{1})} = N\boldsymbol{x}^{(\boldsymbol{0})} + \boldsymbol{b}, \qquad k = 1.$$

Noto $\boldsymbol{x}^{(1)}$, possiamo calcolare $\boldsymbol{x}^{(2)}$ (per $k = 2$), e poi $\boldsymbol{x}^{(3)}$, $\boldsymbol{x}^{(4)}$ e così via. In generale, per ogni $k \geq 1$, possiamo ottenere il vettore $\boldsymbol{x}^{(k)}$ in funzione del vettore precedente $\boldsymbol{x}^{(k-1)}$. In questo modo, abbiamo costruito una formula generale che ci consente di generare iterativamente una successione di vettori, seguendo la struttura del problema iniziale.
A questo punto, è opportuno esaminare se tale metodo iterativo così costruito sia in grado di convergere alla soluzione esatta del sistema. Poiché P e N sono scelte arbitrarie, con l'unica condizione che P sia non singolare, è possibile costruire una famiglia infinita di metodi iterativi. Tuttavia, non sempre è possibile garantire che questi metodi convergano, ossia che la successione $\{\boldsymbol{x}^{(k)}\}$ si avvicini alla soluzione esatta $\boldsymbol{x}$ del sistema lineare. Pertanto, è necessario esaminare i principi teorici alla base della convergenza dei metodi iterativi.
Ripartendo dalla formula

$$P\boldsymbol{x}^{(\boldsymbol{k})} = N\boldsymbol{x}^{(\boldsymbol{k-1})} + \boldsymbol{b}, \qquad k \geq 1,$$

introduciamo l'errore al passo k tra la soluzione calcolata $\boldsymbol{x}^{(k)}$ e la soluzione esatta $\boldsymbol{x}$

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}.$$

Sottraendo membro a membro le seguenti due equazioni

$$P\boldsymbol{x} = N\boldsymbol{x} + \boldsymbol{b}$$
$$P\boldsymbol{x}^{(\boldsymbol{k})} = N\boldsymbol{x}^{(\boldsymbol{k-1})} + \boldsymbol{b}\,'$$

otteniamo

$$P\boldsymbol{x}^{(\boldsymbol{k})} - P\boldsymbol{x} = N\boldsymbol{x}^{(\boldsymbol{k-1})} - N\boldsymbol{x},$$

da cui

$$P\left(\boldsymbol{x}^{(\boldsymbol{k})} - \boldsymbol{x}\right) = N\left(\boldsymbol{x}^{(\boldsymbol{k-1})} - \boldsymbol{x}\right).$$

Ricordando che $\boldsymbol{e^{(k)} = x^{(k)} - x}$, segue che

$$P\boldsymbol{e^{(k)}} = N\boldsymbol{e^{(k-1)}}.$$

Infine, moltiplicando a sinistra e destra per P^{-1}, otteniamo

$$\boldsymbol{e^{(k)}} = P^{-1}N\boldsymbol{e^{(k-1)}}.$$

Abbiamo ottenuto una relazione ricorsiva che lega l'errore al passo $k-1$ con l'errore al passo k. In particolare, tale relazione stabilisce che l'errore al passo k può essere espresso come il prodotto (matrice per vettore) di una matrice $B = P^{-1}N$ per il vettore errore al passo precedente $k-1$

$$\boldsymbol{e^{(k)}} = B\boldsymbol{e^{(k-1)}}.$$

La matrice B è quindi definita come

$$B = P^{-1}N.$$

Da questa definizione si evince il motivo per cui abbiamo richiesto che P forse non singolare. Infatti, è evidente che, se vogliamo risolvere

$$P\boldsymbol{x^{(k)}} = N\boldsymbol{x^{(k-1)}} + \boldsymbol{b},$$

per calcolare $\boldsymbol{x^{(k)}}$, è necessario essere certi che il sistema sia risolubile, ovvero che P sia non singolare.
La matrice $B = P^{-1}N$ è chiamata matrice di interazione associata alla scomposizione (o splitting) $A = P - N$.
È possibile verificare ricorsivamente (partendo dalla relazione $\boldsymbol{e^{(k)}} = B\boldsymbol{e^{(k-1)}}$) che esiste una relazione che lega l'errore al passo k con l'errore al passo inziale, ovvero

$$\boldsymbol{e^{(k)}} = B^k\boldsymbol{e^{(0)}}.$$

Abbiamo concluso questa prima parte di analisi affermando che l'errore al passo k del metodo iterativo può essere espresso come l'errore al passo iniziale ($k = 0$) moltiplicato per la matrice di iterazione B, e più precisamente come il prodotto della matrice B elevata alla potenza k-esima. Ricordiamo che la potenza k-esima di una matrice non è altro che il prodotto della matrice per sé stessa ripetuto k volte.

A questo punto, desideriamo comprendere quali conclusioni possiamo trarre riguardo alla convergenza del metodo iterativo. È infatti fondamentale che il metodo iterativo sia convergente per ogni scelta possibile del vettore iniziale $\boldsymbol{x^{(0)}}$, e quindi per ogni possibile errore iniziale $\boldsymbol{e^{(0)}}$. In altre parole, vogliamo garantire che la convergenza si verifichi indipendentemente dal vettore iniziale scelto, il che equivale a richiedere che il limite di $\boldsymbol{e^{(k)}}$ per k che tende all'infinito sia uguale zero

$$\text{Convergenza } \forall \boldsymbol{x^{(0)}} \quad \Leftrightarrow \quad \lim_{k\to\infty} \boldsymbol{e^{(k)}} = 0.$$

Ciò implica che l'errore, al progressivo aumentare dei passi iterativi, deve tendere a zero, garantendo così la convergenza della successione $\{\boldsymbol{x^{(k)}}\}$ verso la soluzione esatta del sistema.
Poiché abbiamo visto che $\boldsymbol{e^{(k)}} = B^k\boldsymbol{e^{(0)}}$, allora, per avere convergenza del metodo iterativo, dovremmo avere che la matrice B^k per k che tende all'infinito tende a 0

$$\lim_{k\to\infty} B^k = 0.$$

Questo implica che le potenze k-esime della matrice B all'infinito tendono alla matrice nulla

$$B^k \to 0 \quad \text{per} \quad k \to \infty.$$

Tale condizione è vera se e solo se la norma di B (per una qualsiasi norma) è minore di uno

$$\|B\| < 1.$$

Inoltre, ciò è equivalente a richiedere che il raggio spettrale della matrice B è minore di 1

$$\|B\| < 1 \qquad \Leftrightarrow \qquad \rho(B) < 1.$$

È infatti possibile verificare che il raggio spettrale $\rho(B)$ è sempre minore o uguale alla norma di B per ogni norma possibile

$$\rho(B) \leq \|B\|.$$

Pertanto, possiamo concludere una condizione sufficiente per la convergenza di un metodo iterativo per la risoluzione di sistema lineari è che il raggio spettrale della matrice B sia minore di 1

$$\rho(B) < 1.$$

Inoltre, la convergenza sarà tanto più rapida quanto più piccola è la norma della matrice di iterazione B (o, equivalentemente, quanto più piccolo è il raggio spettrale di B).

In sintesi, abbiamo introdotto una famiglia generale di metodi iterativi per la risoluzione di sistemi lineari attraverso lo splitting della matrice A, ossia $A = P - N$. Questo genera una successione di vettori $\boldsymbol{x}^{(k)}$ che soddisfano la relazione $P\boldsymbol{x}^{(\boldsymbol{k})} = N\boldsymbol{x}^{(\boldsymbol{k-1})} + \boldsymbol{b}$ $(k \geq 1)$**.** Per garantire la convergenza del metodo iterativo, è necessario che esista una norma della matrice di iterazione B tale che $\|B\| < 1$, o equivalentemente che il raggio spettrale di B sia minore di 1 $(\rho(B) < 1)$.