RISOLUZIONE DI SISTEMI LINEARI: METODI DIRETTI (Metodi numerici diretti: sostituzioni in avanti e indietro)

In questa lezione iniziamo ad affrontare il problema della risoluzione dei sistemi lineari dal punto di vista numerico. Per cominciare, richiamiamo un risultato classico dell'algebra lineare: la cosiddetta regola di Cramer. Consideriamo un sistema lineare della forma

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

La regola di Cramer garantisce che, se la matrice A è non singolare, ovvero se il suo determinante è diverso da 0, $det(A) \neq 0$, allora il sistema ammette un'unica soluzione. Inoltre, la componente i-esima della soluzione $\boldsymbol{x}$, x_i, è data dal rapporto tra il determinante della matrice $A^{(i)}$ e il determinante di A

$$x_i = \frac{\det(A^{(i)})}{\det(A)} \qquad i = 1, \dots, n.$$

Abbiamo quindi n relazioni per le n componenti della soluzione. La matrice $A^{(i)}$ si ottiene dalla matrice A sostituendo la sua i-esima colonna con il vettore dei termini noti $\boldsymbol{b}$. In altre parole, al posto della i-esima colonna di A, si inseriscono i termini noti del sistema lineare.

Un'osservazione naturale da fare è la seguente: se disponiamo di una formula esplicita per calcolare la soluzione del sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$, perché è necessario sviluppare metodi numerici per la sua risoluzione? La risposta diventerà subito evidente: la regola di Cramer risulta impraticabile dal punto di vista numerico, poiché richiede un numero eccessivo di operazioni.

Osserviamo innanzitutto che la definizione della regola di Cramer richiede il calcolo di $n+1$ determinanti: il determinante della matrice A e i determinanti delle n matrici $A^{(i)}$. Si può dimostrare che il calcolo del determinante di una matrice quadrata di dimensione $n \times n$ richiede un numero di operazioni dell'ordine di $n!$ (ossia n fattoriale). Di conseguenza, il numero totale di operazioni necessarie per determinare le n componenti della soluzione del sistema lineare sarà approssimativamente pari a

$$(n+1)n! = (n+1)! \sim n! \qquad \text{per } n \text{ "grande".}$$

Questo rappresenta quindi il numero totale di operazioni necessarie per calcolare le soluzioni del sistema lineare utilizzando la formula esplicita della regola di Cramer.

Il fatto che il calcolo dei determinanti richieda circa $n!$ operazioni non è immediatamente evidente, ma lo diventa se si considera la formula di Laplace per calcolare il determinante in modo ricorsivo.

$$\det(A) = \begin{cases} a_{11} & \text{se } n = 1 \\ \sum_{j=1}^{n} (-1)^{(1+j)} a_{1j} \det(A_{1j}) & se\ n > 1' \end{cases}$$

Una verifica rigorosa del fatto che il calcolo dei determinanti necessario per risolvere il sistema lineare utilizzando la regola di Cramer richieda circa $n!$ operazioni è fornita nell'Approfondimento 1 di questa lezione.

Questo valore ($n!$) è estremamente elevato (se n è "grande"), tanto che, anche utilizzando un calcolatore molto potente, l'impiego della regola di Cramer risulta proibitivo dal punto di vista computazionale.

Supponendo di utilizzare un calcolatore moderno in grado di effettuare un miliardo di operazioni al secondo (10^9 opz/s, 1 Giga flops) i tempi di calcolo che servirebbero per risolvere un sistema $n \times n$ con la regola di Cramer sono

$$n = 10 \rightarrow t \sim 0.04\ s,$$

$$n = 15 \rightarrow t \sim 5.8\,h,$$
$$n = 20 \rightarrow t \sim 1620 \text{ anni.}$$

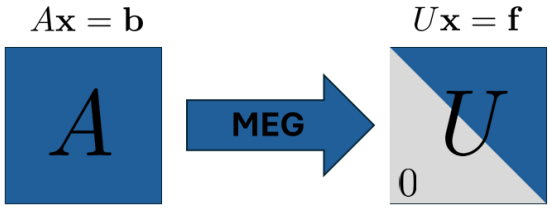
Per una matrice con $n = 10$ il tempo è circa qualche centesimo di secondo, ma già se abbiamo $n = 15$, troviamo quasi sei ore e con $n = 20$ abbiamo più di 1600 anni. La formula di Cramer risulta quindi inutilizzabile ogni volta che il numero di equazioni da risolvere è elevato. Nelle applicazioni ingegneristiche, accade frequentemente che il numero di righe di una matrice sia dell'ordine di centinaia, migliaia o addirittura milioni. È dunque evidente che l'uso della regola di Cramer non sia praticabile. Di conseguenza, è necessario adottare strategie e metodi alternativi.
Per i motivi menzionati poc'anzi, è quindi necessario ricorrere alla cosiddetta approssimazione numerica, o più precisamente alla risoluzione numerica dei sistemi lineari. I metodi di risoluzione numerica possono essere

- Metodi diretti,
- Metodi iterativi.

Un metodo diretto è un procedimento che permette di ottenere la soluzione del problema in un numero finito di passi (potenzialmente elevato, ma comunque finito), supponendo di operare in aritmetica esatta, ovvero senza errori di arrotondamento. I metodi diretti si basano sul Metodo di Eliminazione di Gauss o su sue generalizzazioni, applicabili a classi particolari di matrici, come le matrici a blocchi o a banda.
Un'altra categoria di metodi è quella dei cosiddetti metodi iterativi, i quali conducono alla soluzione $\boldsymbol{x}$ del sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$ come limite di una successione di vettori costruita progressivamente dal metodo stesso

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

L'obiettivo dei metodi iterativi è quindi la costruzione di una successione di vettori che, nel limite, convergano alla soluzione esatta $\boldsymbol{x}$ del sistema lineare. In questo caso, sarà necessario interrompere il processo iterativo non appena si raggiunge un'accuratezza ritenuta soddisfacente.
Concentriamoci per ora sui metodi diretti, in particolare sul Metodo di Eliminazione di Gauss che indicheremo sinteticamente con la sigla MEG. Il MEG viene utilizzato per risolvere il generico sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$, dove A è una matrice che, a priori, supporremo piena (ovvero, con tutti elementi a priori diversi da zero). Nel seguente riquadro schematico,



non abbiamo indicato esplicitamente gli elementi della matrice, ma abbiamo utilizzato un colore (blu) per rappresentare il fatto che A è, inizialmente, una matrice densa, ovvero con elementi non nulli distribuiti in tutta la sua struttura.
Il MEG trasforma il sistema iniziale $A\boldsymbol{x} = \boldsymbol{b}$ in un nuovo sistema lineare con una nuova matrice, che indicheremo con U. Questa matrice U è triangolare superiore (dove U sta per *upper*), il che significa che tutti gli elementi situati al di sotto della diagonale principale sono nulli.

Il metodo, dunque, opera trasformando la matrice generica A nella matrice triangolare superiore U. Il sistema $U\boldsymbol{x} = \boldsymbol{f}$ risultante è equivalente a quello originale $A\boldsymbol{x} = \boldsymbol{b}$, poiché la soluzione $\boldsymbol{x}$ rimane invariata. Tuttavia, sono cambiati sia la matrice (passando da A a U), sia il termine noto (da b a f). La matrice U dipende esclusivamente dalla matrice iniziale A, mentre il nuovo termine noto f dipende sia da A che da b.

$$f = f(A, b), \qquad U = U(A).$$

La trasformazione ottenuta semplifica notevolmente il problema, poiché la struttura triangolare superiore della matrice U rende il sistema molto più semplice da risolvere.

La caratterizzazione di U è quindi tale per cui

$$u_{km} = 0 \quad se\ k > m,$$

ovvero gli elementi della matrice che hanno un indice di riga maggiore dell'indice di colonna saranno tutti nulli.
L'obiettivo del MEG è dunque generare la matrice U, una matrice triangolare superiore, e ottenere il nuovo sistema triangolare superiore associato $U\boldsymbol{x} = \boldsymbol{f}$, che è equivalente al sistema di partenza $A\boldsymbol{x} = \boldsymbol{b}$.
Per eseguire questa operazione, procediamo attraverso una serie di passi successivi. In particolare, introduciamo una successione di trasformazioni applicate alla matrice iniziale A, che gradualmente la modificheranno fino a ottenere la matrice finale U.
Per costruire la matrice U, partiamo dalla matrice iniziale A e generiamo una serie di matrici intermedie, le quali avranno una struttura progressivamente modificata. In particolare, ad ogni passo k, la matrice ottenuta avrà gli elementi al di sotto della diagonale principale posti a zero fino alla riga k-esima. Il processo continua iterativamente fino a quando tutti gli elementi sotto la diagonale principale saranno annullati. La matrice finale, U, avrà quindi una forma triangolare superiore. Questo schema rappresenta il principio fondamentale del MEG. Al momento, lo stiamo analizzando dal punto di vista qualitativo, osservando la trasformazione della struttura delle matrici; in seguito, vedremo nel dettaglio il procedimento analitico che permette di applicare questa successione di trasformazioni.
Alla conclusione del MEG, il problema finale che dovremo affrontare sarà la risoluzione di un sistema triangolare superiore. È evidente che trasformare il sistema iniziale in un sistema con matrice triangolare superiore è vantaggioso solo se siamo in grado di risolvere quest'ultimo in modo semplice ed efficiente. Il nostro obiettivo, quindi, è comprendere come risolvere un sistema lineare con matrice triangolare superiore utilizzando metodi diretti. Esamineremo quindi il procedimento per risolvere un sistema triangolare, considerando sia i sistemi triangolari superiori sia quelli triangolari inferiori.
Nei sistemi triangolari superiori, la parte inferiore della matrice U dei coefficienti è composta interamente da zeri, mentre nei sistemi triangolari inferiori avviene il contrario, con la parte superiore della matrice L uguale a zero. Procediamo di seguito con la scrittura dettagliata, in forma algebrica, di tali sistemi. Consideriamo un sistema generico della forma

$$U\boldsymbol{x} = \boldsymbol{f} = \begin{cases} u_{11}x_1 + u_{12}x_2 + \cdots u_{1n}x_n = f_1 \\ \qquad\quad u_{22}x_2 + \cdots u_{2n}x_n = f_2 \\ \qquad\qquad\quad \vdots \\ \qquad\qquad\qquad\quad u_{nn}x_n = f_n \end{cases}.$$

In questo sistema, la prima equazione è completa, mentre nella seconda equazione manca la prima componente, nella terza mancano la prima e la seconda, e così via. Nell'ultima equazione, invece, risultano assenti tutte le componenti precedenti, fino alla $(n-1)$-esima. Questo schema conferisce al sistema una chiara struttura triangolare. È evidente che tale struttura suggerisce un metodo

particolarmente semplice per la risoluzione del sistema. In particolare, si può partire dall'ultima equazione, che presenta un unico termine incognito, e determinare direttamente la componente x_n. Successivamente, si procede risalendo all'equazione $(n-1)$-esima, poi alla $(n-2)$-esima, fino a raggiungere la prima equazione. Il metodo adottato per questa risoluzione è noto come algoritmo di sostituzione all'indietro, il quale sfrutta la struttura triangolare superiore per calcolare, in sequenza, tutte le componenti della soluzione.
Partendo dell'n-esima equazione, ricavando x_n, avremo

$$x_n = \frac{f_n}{u_{nn}} \qquad n-\text{esima equazione.}$$

Sostituendo questa espressione di x_n nell'equazione $(n-1)$-esima ed isolando la componente x_{n-1}, otteniamo

$$x_{n-1} = \frac{f_{n-1} - u_{n-1n}x_n}{u_{n-1n-1}} \qquad n-1-\text{esima equazione.}$$

In generale avremo, quindi, una formula compatta del tipo

$$x_n = \frac{f_n}{u_{nn}},$$
$$x_k = \frac{f_k - \sum_{m=k+1}^{n} u_{km}x_m}{u_{kk}} \quad \text{per } k = n-1, \dots, 1 \,.$$

Questa formula ci dice che la componente x_n è uguale a $\frac{f_n}{u_{nn}}$. Le altre componenti x_k (con $k = n-1, \dots, 1$), quindi le componenti dalla $(n-1)$-esima alla prima) sono date dal termine noto f_k a cui si sottrae la somma $\sum_{m=k+1}^{n} u_{km}x_m$, poiché stiamo lavorando su componenti della soluzione di indice maggiore di k, le quali sono già state calcolate ricorsivamente nei passi precedenti (dato che stiamo utilizzando la formula per $k = n-1, \dots, 1$). Il tutto viene poi diviso per u_{kk}. Questa divisione, come osserveremo in seguito, è lecita perché vedremo che $u_{kk} \neq 0 \ \forall k$. Chiameremo questo algoritmo sostituzioni all'indietro (in inglese, "*backward substitution*"), che consente di risolvere il sistema triangolare superiore.
In maniera speculare possiamo trattare il sistema triangolare inferiore. Infatti, se consideriamo ora il sistema della forma

$$L\boldsymbol{x} = \boldsymbol{f} = \begin{cases} l_{11}x_1 = f_1 \\ l_{21}x_1 + l_{22}x_2 = f_2 \\ \vdots \\ l_{n1}x_1 + l_{n2}x_2 + \cdots l_{nn}x_n = f_n \end{cases} .$$

Se ripetiamo il processo di prima in ordine inverso partendo dalla prima equazione, troviamo

$$x_1 = \frac{f_1}{l_{11}},$$

e sostituendo poi x_1 nella seconda equazione ed isolando x_2, otteniamo

$$x_2 = \frac{f_2 - l_{21}x_1}{l_{22}}.$$

Possiamo procedere in questo modo fino all'ultima equazione. La formula generale, detta algoritmo delle sostituzioni in avanti (o in inglese "*forward substituion*") è quindi

$$x_1 = \frac{f_1}{l_{11}},$$
$$x_k = \frac{f_k - \sum_{m=1}^{k-1} l_{km} x_m}{l_{kk}} \quad \text{per } k = 2, \dots, n$$

Osserviamo adesso che possiamo effettivamente dividere per gli elementi diagonali per ricavare le componenti della soluzione $\boldsymbol{x}$ (sia per le sostituzioni in avanti che all'indietro). Infatti, per le matrici triangolari, il determinante può essere facilmente calcolato come prodotto dei presenti sulla diagonale principale della matrice

$$\det(U) = \prod_{k=1}^{n} u_{kk}(U).$$

Se il sistema di partenza è non singolare (ipotesi verificata in quanto abbiamo supposto il sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$ non singolare), ovvero se il determinante di U è diverso da zero, $det(U) \neq 0$, allora tutti i fattori u_{kk} devono essere diversi da zero. Di conseguenza, tutti gli elementi diagonali (cioè, quelli situati sulla diagonale principale) della matrice U risultano diversi da zero

$$\det(U) \neq 0 \iff u_{kk} \neq 0 \ \forall k.$$

Per le stesse considerazioni vale anche che

$$\det(L) = \prod_{k=1}^{n} l_{kk}(L), \quad \det(L) \neq 0 \iff l_{kk} \neq 0.$$

Dunque, tutte le componenti x_k, ricavate utilizzando il metodo di sostituzioni in avanti o all'indietro, si possono effettivamente calcolare.
Abbiamo quindi esaminato due modi semplici per risolvere un sistema lineare: uno per un sistema triangolare superiore, utilizzando le sostituzioni all'indietro, e uno per un sistema triangolare inferiore, utilizzando le sostituzioni in avanti.
Si può verifica che l'algoritmo delle sostituzioni all'indietro (o in avanti) porta a un calcolo efficiente per la risoluzione del sistema lineare $U\boldsymbol{x} = \boldsymbol{f}$ (o $L\boldsymbol{x} = \boldsymbol{f}$), poiché richiederà un numero di operazioni dell'ordine di n^2 (dove n è la dimensione della matrice, ossia il numero di righe o colonne della matrice). Una verifica rigorosa del computo delle operazioni necessarie, ad esempio, nel caso delle sostituzioni in avanti (per i sistemi triangolari inferiori) è presentata nell' Approfondimento 2 di questa lezione.

Approfondimento 1:

Il numero totale di operazioni necessarie per determinare le n componenti della soluzione del sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$, utilizzando il metodo di Cramer, è approssimativamente pari a

$$(n+1)! \sim n! \qquad \text{per } n \text{ "grande".}$$

Questo rappresenta quindi il numero totale di operazioni necessarie per calcolare le soluzioni del sistema lineare utilizzando la formula esplicita della regola di Cramer.
Il fatto che il calcolo dei determinanti richieda circa $n!$ operazioni non è immediatamente evidente, ma lo diventa se si considera la formula di Laplace per calcolare il determinante in modo ricorsivo. Infatti, se A è una matrice di dimensione $n \times n$ il suo determinante può essere calcolato mediante la formula di Laplace

$$\det(A) = \begin{cases} a_{11} & \text{se } n = 1 \\ \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j}) & \text{se } n > 1 \end{cases},$$

ovvero, se n è uguale a uno la matrice è in realtà uno scalare e $det(A)$ coincide con a_{11}, mentre se n è maggiore di uno e sviluppando rispetto alla prima riga, $det(A)$ è la somma di $a_{1j}\ det(A_{1j})$ moltiplicato per un segno $(-1)^{1+j}$, dove A_{1j} è la matrice $n-1 \times n-1$ che si ottiene da A sopprimendo la prima riga e la j-esima colonna. Dalla formula si evince che il calcolo del determinante della matrice A può essere ridotto al calcolo di n determinanti di matrici di dimensione ridotta. È quindi evidente che, applicando ricorsivamente questo processo, possiamo ridurre progressivamente la dimensione delle matrici per le quali calcolare il determinante di A, fino a ottenere una matrice di dimensione 1, per la quale il calcolo del determinante diventa banale, poiché corrisponde semplicemente all'elemento che caratterizza la matrice stessa.

Vogliamo rappresentare questo processo in modo schematico. Definiamo con $D(n)$ il numero di operazioni necessarie per calcolare il determinante di una matrice A di dimensione $n \times n$. Se la matrice A è di dimensione 2×2 (ossia con due righe e due colonne), il determinante di A è

$$det(A) = a_{11}a_{22} - a_{12}a_{21}.$$

Dunque, abbiamo due moltiplicazioni e una addizione: tre operazioni. Quindi in questo caso molto semplice abbiamo

$$2 \text{ molt.} + 1 \text{ add.} = 3 \text{ operazioni} \Rightarrow D(2) = 3.$$

Dunque, il numero di operazioni per calcolare il determinante di una matrice di ordine due è uguale a tre. Se adesso consideriamo una matrice 3×3, il calcolo del suo determinante, secondo lo sviluppo rispetto alla prima riga, è

$$\det(A) = a_{11} det(A_{11}) - a_{12} \det(A_{12}) + a_{13} \det(A_{13}).$$

Dunque, abbiamo tre moltiplicazioni, due addizioni e tre determinanti di matrici 2×2. Poiché ogni determinante di una matrice 2×2 richiede 3 operazioni, troviamo

$$3 \text{ prod.}, 2 \text{ add.}, 3 \det 2 \times 2 = 14 \text{ operazioni} \Rightarrow D(3) = 14.$$

Possiamo applicare iterativamente questo processo a una matrice 4×4 e trovare una relazione tra $D(4)$ e $D(3)$. Più precisamente, in questo caso, si effettuano quattro moltiplicazioni, tre addizioni e il calcolo di quattro determinanti di matrici 3×3, per un totale di 63 operazioni

$$4 \text{ prod.}, 3 \text{ add.}, 4 \det 3 \times 3 = 63 \text{ operazioni} \Rightarrow D(4) = 63.$$

In generale si può ricavare che Il numero di operazioni per calcolare il determinante di una matrice $n \times n$ è

$$D(n) = n + (n-1) + nD(n-1),$$

ovvero n moltiplicazioni, n-1 addizioni e n volte il calcolo di determinanti di matrici $n-1 \times n-1$. Abbiamo quindi una formula ricorsiva che mette in relazione il costo computazionale per il calcolo dei determinanti. Possiamo osservare che $D(n)$ cresce come $n!$. Più precisamente, $\frac{D(n)}{n!}$ è maggiore di $\frac{1}{2}$ e minore di una costante C

$$\frac{1}{2} < \frac{D(n)}{n!} < C \Rightarrow D(n) \sim n!$$

Di conseguenza, il calcolo del determinante di una matrice di dimensione $n \times n$ richiede circa $n!$ operazioni. Poiché dobbiamo calcolare $n+1$ determinanti, abbiamo effettivamente verificato che l'applicazione della regola di Cramer comporterebbe un numero di operazioni dell'ordine di $n!$.

Approfondimento 2:

Abbiamo visto due modi semplici per risolvere un sistema lineare triangolare superiore con le sostituzioni all'indietro e un sistema triangolare inferiore con le sostituzioni in avanti. Vediamo adesso che questo calcolo è molto semplice perché richiederà un numero di operazioni dell'ordine di n^2 (dove n è la dimensione della matrice, ovvero il numero di righe o di colonne della matrice). Possiamo contare le operazioni da fare, per esempio, nel caso delle sostituzioni in avanti (nel caso quindi di sistemi triangolari inferiori). Le formule per le sostituzioni in avanti sono

$$x_1 = \frac{f_1}{l_{11}}, \qquad \text{1 div.}$$

$$x_k = \frac{f_k - \sum_{m=1}^{k-1} l_{km} x_m}{l_{kk}} \quad \text{per } k = 2, \dots, n \qquad \text{1 div.}, k-1 \text{ add.}, k-1 \text{ prod.}$$

Per calcolare x_1 dobbiamo fare solo una divisione. Per calcolare x_k dobbiamo fare, per ogni k, una divisione (perché dobbiamo dividere per l_{kk}), $k-1$ prodotti (dati da $l_{km}x_m$per $m = 1, \dots, k-1$), e poi $k-1$ addizioni (indicate dall'operazione di somma $\sum_{m=1}^{k-1}\cdot$). Dunque, in totale abbiamo

$$1 + \sum_{k=2}^{n} (2k-1) = n^2,$$

ovvero, una operazione di divisione 1, sommata a $(k-1)+(k-1)+1$ operazioni, ovvero $2k-1$ operazioni. Infine, abbiamo una somma con $k = 2, \dots, n$ perché nella formula delle sostituzioni in avanti troviamo x_k per $k = 2, \dots, n$. Facendo questa somma troviamo esattamente n^2 operazioni.

Abbiamo quindi verificato che la risoluzione di un sistema triangolare, sia esso superiore o inferiore, richiede n^2 operazioni. Questi algoritmi (delle sostituzioni in avanti e indietro) sono metodi diretti perché, come abbiamo visto, conducono alla risoluzione del sistema con un numero finito di passi (ovvero un numero finito di operazioni), e se non si commettessero errori di arrotondamento nelle operazioni, si arriverebbe alla soluzione esatta del sistema. Gli unici errori che si introducono, quindi, con questo approccio sono gli errori di arrotondamento. Vedremo in seguito come questi errori si riflettono nell'accuratezza della soluzione finale.