RISOLUZIONE DI SISTEMI LINEARI: METODI ITERATIVI (Il metodo di rilassamento SOR e la matrice di precondizionamento)

In questa lezione verrà presentata una generalizzazione del metodo di Gauss-Seidel (GS), illustrato nella lezione precedente, mediante l'introduzione del metodo di rilassamento denominato SOR (acronimo di "Successive Over Relaxation"). Successivamente, i metodi finora esaminati (Jacobi, GS e SOR) verranno reinterpretati in una nuova veste, che permetterà di generalizzare ulteriormente i metodi introdotti e di ottenere nuove famiglie di metodi iterativi. Tale generalizzazione si articola attraverso il concetto di matrice di precondizionamento.

Il metodo di rilassamento, anche detto SOR, rappresenta una generalizzazione del metodo di GS, in quanto consente l'introduzione di un parametro di accelerazione, opportunamente scelto per assicurare che la convergenza verso la soluzione del sistema lineare avvenga nel modo più rapido possibile. Questo costituisce il primo esempio di metodo iterativo dipendente da un parametro.

Introduciamo la struttura del metodo di rilassamento mediante un'interpretazione geometrica. Consideriamo l'asse reale, sul quale viene definito il parametro di rilassamento, che indichiamo con ω. Inoltre, consideriamo un ipotetico asse in $\mathbb{R}^n$, ossia lo spazio n-dimensionale in cui i punti rappresentano i vettori. Il vettore generato dal metodo di GS corrisponde alla scelta $\omega = 1$ del parametro di rilassamento. Di conseguenza, se il parametro di rilassamento viene scelto uguale a 1, si ottiene il metodo di Gauss-Seidel; mentre, per un valore di ω diverso da 1, si ricava il metodo di rilassamento SOR, che produce un nuovo vettore.

Esaminiamo ora nel dettaglio le formule che consentono di introdurre la successione iterativa del metodo SOR. Supponendo di disporre della soluzione al passo k, ovvero $\boldsymbol{x}^{(k)}$, si procede inizialmente con un'iterata del metodo di GS

$$y_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right).$$

La formula presentata corrisponde esattamente a quella del metodo di Gauss-Seidel, con la differenza che la soluzione al passo $k+1$ viene denominata $y_i^{(k+1)}$ anziché $x_i^{(k+1)}$, in quanto $y_i^{(k+1)}$ rappresenta una soluzione intermedia. Successivamente, $y_i^{(k+1)}$ viene modificata secondo il seguente algoritmo

$$x_i^{(k+1)} = \omega y_i^{(k+1)} + (1-\omega) x_i^{(k)}.$$

Definiamo quindi la nuova soluzione del metodo SOR come una combinazione lineare, attraverso il parametro ω, della soluzione di transito $y_i^{(k+1)}$ e della soluzione precedente $x_i^{(k+1)}$. Il parametro ω è il parametro di rilassamento. Questo algoritmo introduce un elemento di memoria, in quanto il nuovo valore dipende sia dalla soluzione di transito ottenuta tramite il metodo di GS sia dalla soluzione precedente. Si osservi, inoltre, che se $\omega = 1$ si ottiene

$$x_i^{(k+1)} = y_i^{(k+1)}$$

riottenendo in tal modo la formula del metodo di GS. Nel caso in cui $\omega \neq 1$, si ottiene una formula più generale che consente di modificare e accelerare la convergenza del metodo.

La matrice di iterazione del SOR dipenderà da ω ed è espressa dalla formula

$$B(\omega) = (I + \omega D^{-1}E)^{-1}[(1-\omega)I - \omega D^{-1}F].$$

Per garantire la convergenza del metodo SOR, è essenziale che il raggio spettrale della matrice di iterazione $B(\omega)$ sia minore di uno, ovvero $\rho(B(\omega)) < 1$. Di conseguenza, la scelta del parametro ω non può essere arbitraria, bensì deve soddisfare condizioni specifiche che assicurino tale proprietà, garantendo così la convergenza del metodo.

Andiamo dunque a caratterizzare il raggio spettrale dalla materia di iterazione in funzione di ω. Possiamo riscrivere $B(\omega)$ come il prodotto di due matrici B_1 e B_2

$$B(\omega) = (I + \omega D^{-1}E)^{-1}[(1-\omega)I - \omega D^{-1}F] = B_1 B_2,$$

dove

$$B_1 = (I + \omega D^{-1}E)^{-1}$$
$$B_2 = [(1-\omega)I - \omega D^{-1}F]'$$

Ricordando che, se una matrice B è il prodotto di due matrici B_1 e B_2, allora il determinante è il prodotto dei determinanti

$$\det\big(B(\omega)\big) = \det(B_1)\det(B_2).$$

Il determinante di queste due matrici B_1 e B_2 è facilmente determinabile, poiché, essendo matrici triangolari, basta esaminare la loro diagonale principale (in quanto il determinante di una matrice triangolare corrisponde al prodotto degli elementi sulla diagonale principale). In particolare, B_1 presenta sulla diagonale principale tutti elementi uguali a 1, per cui il suo determinante risulta pari a 1; B_2 invece presenta sulla diagonale principale tutti elementi uguali a $1-\omega$ e, di conseguenza, il suo determinante è $(1-\omega)^n$.

Quindi il determinante di $B(\omega)$ è

$$\det\big(B(\omega)\big) = \det(B_1)\det(B_2) = 1 \cdot (1-\omega)^n = (1-\omega)^n.$$

Inoltre, possiamo osservare che il prodotto degli autovalori di $B(\omega)$ in valore assoluto è uguale a

$$\prod_{i=1}^{n} |\lambda_i(B(\omega))| = |1-\omega|^n .$$

Questa relazione deriva direttamente dal fatto che il determinante di $B(\omega)$ è il prodotto degli autovalori della matrice. Estraendo dal prodotto il massimo dei moduli degli autovalori, ossia il raggio spettrale, ed elevandolo alla n-esima potenza, si ottiene la seguente disuguaglianza

$$\big[\rho\big(B(\omega)\big)\big]^n \geq \prod_{i=1}^{n} \big|\lambda_i\big(B(\omega)\big)\big| = |1-\omega|^n ,$$

da cui, prendendo la radice ennesima ad entrambi i membri, troviamo che il raggio spettrale è maggiore o uguale del valore assoluto di $1-\omega$

$$\rho\big(B(\omega)\big) \geq |1-\omega|.$$

Quindi la condizione necessaria per avere convergenza con il metodo di SOR è

$$0 < \omega < 2,$$

poiché il raggio spettrale $\rho\big(B(\omega)\big)$ deve sempre essere minore di uno e quindi $|1-\omega| < 1$.

Questa condizione necessaria per la convergenza del metodo SOR diventa anche sufficiente (ovvero che basti prendere ω compreso fra 0 e 2, per essere certi della convergenza) nei casi in cui

- A è simmetrica e definita positiva (SDP). Se A è SDP, allora la condizione $0 < \omega < 2$, non è solo necessaria, ma anche sufficiente per la convergenza del SOR. Quindi, prendendo un qualunque ω compreso fra 0 e 2, la convergenza del SOR è assicurata.

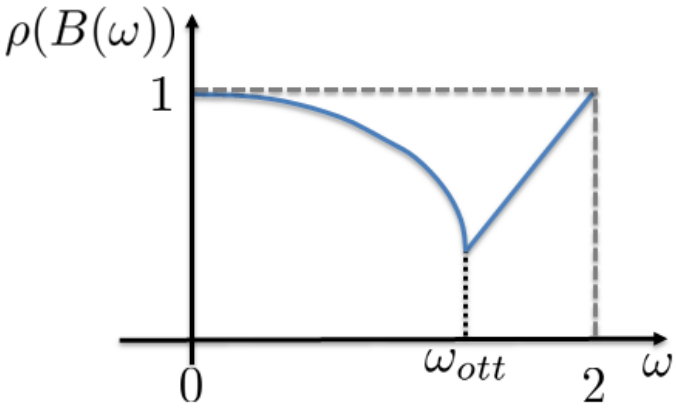

Figura 1: Andamento del raggio spettrale della matrice di iterazione $B(\omega)$ in funzione del parametro ω.

- A è non solo simmetrica definita positiva (SDP) ma anche tridiagonale, la scelta del parametro di rilassamento ω nell'intervallo $0 < \omega < 2$ garantisce la convergenza del metodo SOR. Inoltre, in questo caso, è possibile determinare un valore ottimale ω_{ott} che massimizza la velocità di convergenza. Questo valore è dato dalla formula:

$$\omega_{ott} = \frac{2}{1 + \sqrt{1 - \rho(B_{GS})}}.$$

Negli altri casi più generali non avremo la possibilità di utilizzare questa rappresentazione. E dunque siamo indotti a capire come scegliere omega ottimale. Per far questo cerchiamo di capire qual è il comportamento del raggio spettrale della matrice di iterazione in funzione di ω.
L'andamento tipico della funzione raggio spettrale è rappresentato in Figura 1.
Osservando la Figura 1, possiamo interpretare qualitativamente il comportamento della funzione $\rho(B(\omega))$ all'interno dell'intervallo $0 < \omega < 2$ osservando che essa assume un andamento con un minimo in corrispondenza del valore ottimale ω_{ott}. Inoltre, la funzione è sempre minore di 1 nell'intervallo di convergenza e assume il valore 1 agli estremi, ovvero per $\omega = 0$ e $\omega = 2$, dove la convergenza non è garantita poiché il raggio spettrale è esattamente uguale a 1

$$\rho(B(0)) = \rho(B(2)) = 1$$

Ciò implica che per questi valori il metodo SOR non converge. Inoltre, il valore del raggio spettrale dipende in modo sensibile dalla scelta di ω. Infatti, sperimentalmente si osserva che anche piccole variazioni di ω possono determinare variazioni significative nel numero di iterazioni necessarie per raggiungere la convergenza. Questo evidenzia l'importanza della scelta di ω, in quanto un valore non ottimale può rallentare notevolmente il processo iterativo.

Vogliamo ora fornire una nuova interpretazione ai metodi iterativi analizzati fino a questo punto (Jacobi, GS e SOR), introducendo un quadro più generale che consenta di estenderli e di ottenere nuove famiglie di metodi iterativi. Questa reinterpretazione si basa sul concetto di matrice di precondizionamento.

L'idea alla base di questa generalizzazione è quella di riscrivere i metodi iterativi in una forma più astratta, evidenziando il ruolo della decomposizione della matrice del sistema lineare.
Partiamo andando a riscrivere il residuo alla k-esima iterazione

$$\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}.$$

Si consideri un generico metodo iterativo finalizzato alla determinazione di una soluzione $x^{(k)}$ al passo k. Ricordando la formulazione generale del metodo iterativo

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

da cui possiamo (alla destra dell'uguale) sommare e sottrarre la matrice P alla matrice N, ottenendo

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

Ricordando ora che per lo splitting della matrice abbiamo $A = P - N$. Di conseguenza si ha

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

Inoltre, $\boldsymbol{b} - A\boldsymbol{x}^{(k)}$ è il residuo al passo k, $\boldsymbol{r}^{(k)}$, quindi

$$P\boldsymbol{x}^{(\boldsymbol{k+1})} = P\boldsymbol{x}^{(\boldsymbol{k})} + \boldsymbol{r}^{(\boldsymbol{k})},$$

da cui

$$P(\boldsymbol{x}^{(\boldsymbol{k+1})} - \boldsymbol{x}^{(\boldsymbol{k})}) = \boldsymbol{r}^{(\boldsymbol{k})}.$$

Di conseguenza, una singola iterazione di un generico metodo iterativo, corrispondente alla decomposizione nelle matrici P e N, può essere espressa nella seguente forma: l'operatore P applicato all'incremento $\boldsymbol{x}^{(\boldsymbol{k+1})} - \boldsymbol{x}^{(\boldsymbol{k})}$ risulta equivalente al residuo al passo k, ovvero $\boldsymbol{r}^{(\boldsymbol{k})}$.
Tale formulazione, più compatta, consente di evidenziare il meccanismo mediante il quale si passa dal residuo alla nuova iterata $\boldsymbol{x}^{(\boldsymbol{k+1})}$ attraverso la matrice P. La matrice P è chiamata matrice di precondizionamento.
Possiamo definire P, specificandola per i vari casi visti finora: Jacobi (P_J), GS (P_{GS}) e SOR (P_{SOR}):

- Jacobi: $P_J = D$;
- GS: $P_{GS} = D + E$;
- SOR: $P_{SOR} = \frac{1}{\omega}(D + \omega E)$.

Quindi possiamo riscrivere i metodi visti fino ad ora, alla luce della nuova struttura. L'aspetto interessante di questa riscrittura è che i metodi iterativi possono caratterizzarsi questa volta non più attraverso lo splitting della matrice A, ma tramite la matrice di precondizionamento P, grazie all'introduzione del concetto di residuo $\boldsymbol{r}^{(\boldsymbol{k})}$.
Modificando la matrice P, si modifica il metodo iterativo, determinando così una corrispondenza biunivoca tra la scelta di P e il metodo iterativo adottato. La matrice P è denominata matrice di precondizionamento, o precondizionatore di A, poiché esercita un effetto di scalatura sugli elementi della matrice A. In altri termini, come sarà approfondito successivamente, l'introduzione di P consente di sostituire la matrice A con la matrice P^{-1}A, la quale, a priori, risulta meglio condizionata rispetto ad A.
Vediamo quali sono i criteri che si possono seguire per determinare la matrice P. Per individuare una nuova matrice P abbiamo due esigenze da soddisfare:

1. Minimizzare il raggio spettrale della matrice di iterazione. La prima esigenza è quella di assicurare una convergenza, la più rapida possibile. Si può verificare che la matrice di iterazione associata alla scelta di P è

$$B = I - P^{-1}A,$$

ovvero, l'identità I meno $P^{-1}A$. Ne consegue che, se si desidera ottenere una matrice B con il raggio spettrale il più ridotto possibile, è opportuno garantire che $P^{-1}A$ sia, in senso euristico, il più vicino possibile all'identità. In tal modo, la matrice B risulterà prossima allo zero. In realtà, ciò che riveste maggiore importanza è che gli autovalori di $P^{-1}A$ si avvicinino, per quanto possibile, a quelli dell'identità I, i quali, come noto, sono tutti pari a 1. Pertanto, l'azione di P, attraverso la sua inversa applicata ad A, deve avere l'effetto di ridurre la dispersione degli autovalori, restringendoli a un insieme che tenda ad avvicinarsi a quello di I. Se tale condizione viene soddisfatta, la matrice B presenterà un raggio spettrale contenuto, determinando una più rapida convergenza del metodo iterativo.

2. Mantenere basso il costo computazionale. Un ulteriore aspetto cruciale è il contenimento del costo computazionale necessario per la risoluzione del sistema. In particolare, ad ogni passo dell'iterazione, è necessario risolvere un sistema della forma
$$P(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = \boldsymbol{r}^{(k)}.$$
La complessità della risoluzione di tale sistema dipende dalla struttura di P: quanto più semplice risulta la matrice, tanto più efficiente sarà il calcolo. Durante il corso, sono state analizzate le classi di matrici caratterizzate da una ridotta complessità computazionale nella risoluzione dei sistemi lineari. Il paradigma di riferimento è rappresentato dalle matrici diagonali, le quali permettono una risoluzione estremamente efficiente. Anche le matrici triangolari e tridiagonali presentano una bassa complessità computazionale. In particolare, le matrici tridiagonali richiedono un numero di operazioni dell'ordine di n, mentre per la risoluzione di sistemi con matrici triangolari è necessario un numero di operazioni dell'ordine di n^2.

Esaminiamo ora alcuni esempi di precondizionatori. In primo luogo, consideriamo il caso dei precondizionatori diagonali, che soddisfano pienamente la seconda esigenza. Con un precondizionatore diagonale, il sistema da risolvere risulta estremamente semplice. Un esempio comune è rappresentato dal precondizionatore di Jacobi, ossia la matrice diagonale che ha come elementi diagonali quelli corrispondenti agli elementi diagonali della matrice di partenza, cioè
$$P_J = \text{diag}(a_{11}, a_{22}, \dots, a_{nn}).$$

Questo precondizionatore prende il nome di Jacobi proprio perché il metodo di Jacobi si ottiene scegliendo $P = P_J$.
Un altro esempio di precondizionatore diagonale è quello definito in norma due. In questo caso, anziché utilizzare direttamente gli elementi diagonali della matrice A, si considerano i valori
$$P_2 = \text{diag}(c_1, c_2, \dots, c_n),$$

dove ogni c_i è definito come
$$c_i = \left(\sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}.$$

Quindi, c_i rappresenta la radice quadrata della somma dei quadrati degli elementi della i-esima riga della matrice A, il che equivale a una sorta di media quadratica degli elementi di ciascun elemento diagonale.

Infine, accenniamo ad alcune scelte meno ovvie, ma altrettanto efficaci: le cosiddette fattorizzazioni incomplete. È noto che una matrice A può essere fattorizzata come il prodotto di due matrici L e U. Nel processo di fattorizzazione, le matrici L e U possono contenere elementi non nulli anche nelle posizioni in cui la matrice di partenza aveva elementi nulli. Le fattorizzazioni incomplete si basano sull'idea delle fattorizzazioni LU (o della fattorizzazione di Cholesky), ma mantengono quella che viene chiamata struttura di "sparsità" della matrice originale. In altre parole, si conserva la posizione degli zeri nella matrice iniziale, mentre si determinano a priori gli elementi non nulli nelle matrici triangolari inferiori e superiori. Questo tipo di fattorizzazioni, denominate rispettivamente ILU (Incomplete LU) e IC (Incomplete Cholesky), sono varianti delle fattorizzazioni LU e di Cholesky, modificate per preservare la struttura sparsa (ovvero gli elementi nulli) della matrice iniziale.