CAPITOLO II

SISTEMI LINEARI

In questo capitolo ci occuperemo delle risoluzione numerica dei sistemi lineari del tipo

$$Ax=b$$

con $A \in \mathcal{L}(C^n, C^n)$, $x,b \in C^n$, ed $|A| \neq 0$.

I sistemi lineari rappresentano dei modelli per una vasta gamma di problemi attinenti alle scienze applicate ed all'ingegneria quali il calcolo delle reti elettriche, la dinamica discreta di popolazioni, il calcolo delle strutture statiche ecc. Spesso i sistemi derivanti da tali problemi sono di dimensioni moderate, e non presentano difficoltà di risoluzione dal punto di vista numerico anche se le matrici che intervengono sono generalmente matrici piene e non strutturate. Per tali sistemi saranno generalmente sufficienti i metodi diretti che verranno esposti nel paragrafo 2.

Le più grosse difficoltà sorgono invece nella risoluzione dei sistemi lineari che si ottengono dalla discretizzazione di certi sistemi di equazioni differenziali, ordinarie o a derivate parziali. Se la discretizzazione è abbastanza fine, allora essa può dar luogo ad un sistema lineare di dimensione molto grande. D'altra parte la particolare natura delle equazioni differenziali insieme a particolari scelte delle discretizzazioni, danno luogo generalmente a sistemi lineari di forma particolare che consentono l'uso di certe tecniche iterative che sono presentate e studiate nel prossimo paragrafo di questo capitolo.

Consideriamo quindi, a titolo di esempio, due tipici problemi differenziali del secondo ordine e una loro possibile discretizzazione.

Osserviamo preliminarmente che la derivata seconda u"(t) di una funzione può essere approssimata con la seguente *differenza centrale*

$$u''(t) \approx \frac{u(t-h) - 2u(t) + u(t+h)}{h^2}$$

Sommando gli sviluppi in serie di Taylor fino al quarto ordine di u(t-h) e u(t+h) si ottiene infatti:

$$u''(t) = \frac{u(t-h) - 2u(t) + u(t+h)}{h^2} + O(h^2)$$

Problema dei due punti.

Data una funzione $g(t) \geq 0$ in [0,1] , si consideri , per una assegnata f(t), l'equazione differenziale

$$u''(t) - g(t)u(t) = f(t) \qquad t \in [0,1]$$

con le condizioni agli estremi:

$$u(0)= a, \qquad u(1)= b.$$

Si può dimostrare che tale problema ammette una ed una sola soluzione per ogni termine noto f(t). Consideriamo ora una discretizzazione dell'intervallo [0,1] di passo h=1/N, con N intero,

$$t_i=ih \quad i=0,1,... N$$

e indichiamo con u_i le approssimazioni di $u(t_i)$, e con g_i ed f_i i valori $g(t_i)$ e $f(t_i)$ rispettivamente. Approssimando $u''(t_i)$, per i=1,...N-1, con la differenza centrale $\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2}$ si ottiene,cambiando segno all'equazione e tenendo conto che u_0=a ed u_N=b, il seguente sistema lineare:

$$(2 + h^2g_1)u_1 - u_2 = -h^2f_1 + a$$

.....

$$-u_{i-1} +(2 + h^2g_i)u_i - u_{i+1} = -h^2f_i$$

.....

$$-u_{N-2} +(2 + h^2g_{N-1})u_{N-1}= -h^2f_{N-1}+ b$$

la cui matrice :

$$A = \begin{Vmatrix} 2 + h^2g_1 & -1 & & & \\ -1 & 2 + h^2g_2 & -1 & & \\ & . & . & . & \\ & & -1 & 2 + h^2g_i & -1 \\ & & & . & . & . \\ & & & & -1 & 2 + h^2g_{N-2} & -1 \\ & & & & & -1 & 2 + h^2g_{N-1} \end{Vmatrix}$$

risulta simmetrica e tridiagonale.
Dimostriamo ora che la matrice A risulta anche definita positiva. A tale scopo osserviamo che

$$x^H A x = x^H T x + x^H G x$$

dove :

$$T=\begin{Vmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & . & . & . & & \\ & & . & . & . & \\ & & & . & . & . \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{Vmatrix} \quad , G=diag(h^2g_1,...h^2g_i...,h^2g_{N-1})$$

Poichè la matrice G è evidentemente semidefinita positiva a seguito della condizione $g(t)\geq 0$, A risulterà definita positiva se tale è T. Per il corollario 1.6 si ha, per ogni $x\neq 0$, $\lambda_1 x^H x \leq x^H T x \leq \lambda_n x^H x$.
D'altra parte, per il teorema di Gerschgorin, tutti gli autovalori di T sono ≥ 0. Se dimostro che T è non singolare, allora posso dire che tutti gli autovalori sono positivi e quindi, per la relazione precedente, anche $0 < x^H T x$.
Non resta dunque che dimostrare la non singolarità di T , per ogni dimensione della matrice. A tale scopo si osserva immediatamente che, detta T_k la matrice tridiagonale T di dimensione k, si ha

$$\det(T_{k+1})=2\det(T_k) - \det(T_{k-1})$$

dalla quale, per induzione, si ricava

$$\det(T_k) = k+1.$$

La matrice T è dunque non singolare e, di conseguenza, definita positiva. Ciò assicura, come abbiamo osservato prima, che anche A è definita positiva.

Problema di Dirichelet.

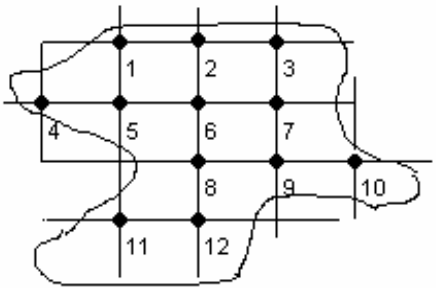
Sia u(x,y) una funzione definita in un dominio D del piano R^2, il cui bordo sarà indicato con Γ_D , che soddisfa l'equazione a derivate parziali

$$\frac{\partial^2 u}{\partial^2 x}+\frac{\partial^2 u}{\partial^2 y} = f(x,y) \qquad\qquad x,y \in D$$

con la *condizione al contorno*,

$$u(x,y)=0 \qquad x,y \in \Gamma_D.$$

Effettuiamo una reticolazione del dominio D con rette parallele agli assi coordinati, distanti tra loro una quantità positiva h, ed enumeriamo con un indice progressivo i nodi P_i che risultano *interni* al dominio.



Usando le differenze centrali per approssimare le derivate parziali seconde, si ottiene:

$$\frac{\partial^2 u(P_i)}{\partial^2 x} = \frac{u(P_{i-1})- 2u(P_i)+ u(P_{i+1})}{h^2}$$

e

$$\frac{\partial^2 u(P_i)}{\partial^2 y} = \frac{u(P_{i-r_i})- 2u(P_i)+ u(P_{i+s_i})}{h^2}$$

dove i punti P_{i-r_i} e P_{i+s_i} sono sovrastanti e sottostanti il punto P_i nel reticolo.

Per ogni punto interno del reticolo si avrà dunque

$$\frac{\partial^2 u(P_i)}{\partial^2 x} + \frac{\partial^2 u(P_i)}{\partial^2 y} = \frac{u(P_{i-r_i})+u(P_{i-1})- 4u(P_i)+ u(P_{i+1})+u(P_{i+s_i})}{h^2}$$

intendendo che se un punto P_i, prossimo al bordo, è tale che il nodo precedente, o seguente o sovrastante o sottostante del reticolo esce dal domino, in tale nodo si assegna il valore nullo.

In riferimento alla figura, il sistema lineare nelle incognite $u(P_i)$ i=1,...,12 ha una matrice *tridiagonale a blocchi* della forma:

$$A = \left\|\begin{array}{cccccccccccc}
4 & -1 & & -1 & & & & & & & & \\
-1 & 4 & -1 & & -1 & & & & & & & \\
 & -1 & 4 & & & -1 & & & & & & \\
 & & & 4 & -1 & & & & & & & \\
-1 & & & -1 & 4 & -1 & & & & & & \\
 & -1 & & & -1 & 4 & -1 & & & & & \\
 & & -1 & & & -1 & 4 & & -1 & & & \\
 & & & -1 & & & & 4 & -1 & & -1 & \\
 & & & & -1 & & & -1 & 4 & -1 & & \\
 & & & & & -1 & & & -1 & 4 & & \\
 & & & & & & -1 & & & & 4 & -1 \\
 & & & & & & & -1 & & & -1 & 4
\end{array}\right\| \qquad (2.1)$$

E' facile rendersi conto che la matrice risulta simmetrica. Inoltre, con ragionamenti simili all'esempio precedente, si dimostra anche che è definita positiva. E' anche facile capire che se l'equazione fosse stata di dimensione 3, cioè:

$$\frac{\partial^2 u}{\partial^2 x} + \frac{\partial^2 u}{\partial^2 y} + \frac{\partial^2 u}{\partial^2 z} = f(x,y,z) \qquad x,y,z \in D$$

il sistema sarebbe risultato *pentadiagonale a blocchi* e la sua dimensione sarebbe cresciuta considerevolmente. Se il dominio D fosse, per esempio, un cubo con lo spigolo unitario, un reticolo con 10 punti interni su ogni coordinata darebbe luogo ad un sistema di dimensione 1000.

1.METODI ITERATIVI.

I metodi iterativi per la risoluzione dei sistemi lineari si fondano sulla trasformazione dell'equazione $Ax=b$ in un problema equivalente di punto fisso. Ciò si ottiene attraverso uno *spezzamento* (*splitting*) della matrice A in

$$A = N - P \quad , \quad \text{con N non singolare e } P=N-A,$$

che dà luogo a

$$Nx = Px + b$$
$$x= N^{-1}Px + N^{-1}b$$

Il problema $Ax=b$ è cosi trasformato nel problema di punto fisso

$$x = Mx + a \qquad (2.2)$$

dove $a=N^{-1}b$ ed $M = N^{-1}P = I-N^{-1}A$. La matrice M è detta *matrice di iterazione* ed è definita univocamente dallo splitting.

Il problema di punto fisso viene quindi affrontato con l'iterazione

$$x^{k+1}=Mx^k +a, \qquad (2.3)$$

o, più precisamente, con l'iterazione

$$Nx^{k+1} = Px^k + b$$

a partire da un vettore iniziale x^0 assegnato. Affinchè questo approccio possa essere vantaggioso occorre che il sistema

$$Ny = Px^k + b$$

sia risolvibile per y in maniera "diretta", cioè con un costo trascurabile rispetto al costo richiesto per la risoluzione del problema originale $Ax=b$.

Sottraendo (2.2) da (2.3) si ottiene, per gli errori $e^k := x^k - x$,

$$e^{k+1} = Me^k = M^2e^{k-1} =... = M^{k+1}e^0$$

Se la matrice M è convergente allora l'errore tende a zero, qualunque sia il vettore iniziale x^0, mentre può accadere che l'errore tenda a zero senza che la matrice sia

infinitesima. Ciò accade quando M^k tende ad una matrice il cui nucleo includa e^0. Se invece si vuole che l'errore tenda a zero per ogni vettore iniziale, allora M deve essere convergente. Abbiamo dunque dimostrato il seguente teorema.

TEOREMA 2.1. *Condizione necessaria e sufficiente affinchè l'iterazione* (2.3) *converga per ogni vettore iniziale* x^0 *è che la matrice* M *sia infinitesima o, equivalentemente, che* $\rho(M)<1$.

Dunque nel proporre un metodo iterativo, lo splitting deve essere scelto con i seguenti criteri:

1) N deve essere non singolare.
2) N deve essere invertibile a costo trascurabile.
3) La matrice di iterazione M che ne deriva deve essere convergente.
4) Il raggio spettrale di M deve essere più piccolo possibile.

Quest'ultima affermazione discende dalla seguente analisi asintotica dell'errore.

Analisi asintotica dell'errore.

Nel valutare la bontà di un metodo iterativo attraverso l'analisi della successione e^k degli errori , si deve osservare che questa dipende dal punto iniziale x^0 col quale si innesca l'iterazione e dalla particolare norma con la quale viene valutata l'ampiezza degli errori. Una valutazione corretta della velocità di convergenza del metodo deve prescindere da entrambi questi fattori. Partendo quindi dalla relazione ricorsiva sugli errori

$$e^n = Me^{n-1} = M^n e^0$$

si ottiene, per una norma arbitraria,:

$$||e^n|| \leq ||M^n|| ||e^0||$$

e quindi,

$$\frac{||e^n||}{||e^0||} \leq ||M^n||$$

Dunque in n iterazioni il fattore di smorzamento dell'errore $\frac{||e^n||}{||e^0||}$ è maggiorato da $||M^n||$, e quindi la quantità $\frac{||e^n||}{||e^0||}$, che rappresenta, dopo n passi, il fattore medio di smorzamento ad ogni passo, è a sua volta maggiorato da

$$\sqrt[n]{\frac{||e^n||}{||e^0||}} \leq \sqrt[n]{||M^n||} \ .$$

Il termine a destra della disuguaglianza ammette limite, per $n \to \infty$, pari a $\rho(M)$. Di conseguenza il termine a sinistra, che in generale non ammette limite, possiede certamente il massimo limite e per esso si avrà:

$$\max \lim_{n\to\infty} \sqrt[n]{\frac{||e^n||}{||e^0||}} \leq \rho(M).$$

Osserviamo infine che scegliendo x^0 in modo tale che e^0 sia autosoluzione di M associata all'autovalore μ di modulo massimo, si ha

$$e_n=M^n e_0=\mu^n e_0$$

Da ciò si ricava

$$||e^n|| =\rho^n(M)||e^0||$$

e quindi

$$\sqrt[n]{\frac{||e^n||}{||e^0||}}=\rho(M).$$

Di conseguenza rispetto a tutti i possibili vettori iniziali ed a tutte le norme si ha:

$$\max_{e^0}\left(\max \lim_{n\to\infty} \sqrt[n]{\frac{||e^n||}{||e^0||}}\right)= \rho(M)$$

Il termine a sinistra si chiama **fattore asintotico di convergenza** e rappresenta il peggiore fattore medio di cui viene smorzato asintoticamente l'errore iniziale, indipendentemente dal vettore iniziale e dalla norma con cui si misura l'errore.

Il fatto che il fattore asintotico di convergenza coincide col raggio spettrale della matrice di iterazione del metodo, giustifica l'asserzione che un metodo iterativo è, *in*

generale, tanto più veloce quanto più piccolo è il raggio spettrale. Ciò non esclude che per certi vettori iniziali un metodo risulti più veloce di un altro di raggio spettrale inferiore.

Per dare una stima a priori del numero di iterazioni necessarie per avere un errore al di sotto di una assegnata tolleranza, è utile il concetto di **ordine asintotico di convergenza**. Esso rappresenta il numero di iterazioni necessarie per smorzare asintoticamente l'errore di un fattore 10^{-1} e verrà indicato con R.

Sulla base dei risultati precedenti si può dire che asintoticamente, cioè dopo un numero sufficientemente grande di iterazioni, sarà sufficiente che R soddisfi la disuguaglianza:

$$\rho^R \leq 10^{-1} \qquad \text{e quindi , poiche } \rho<1, \quad R \geq -1/\log_{10}\rho$$

Come prima, questa è una stima pessimistica inquanto il fattore asintotico di smorzamento per l'iterazione che si sta calcolando potrebbe essere, in realtà, più piccolo. Si osservi che se un altro metodo ha il raggio spettrale che è il quadrato di ρ, il suo ordine asintotico di convergenza è la metà. Si noti infine che disponendo di una maggiorazione del raggio spettrale, purchè inferiore ad 1, si dispone di una maggiorazione di R.

Metodo di Jacobi (metodo delle sostituzioni simultanee).

Il metodo di Jacobi consiste nel decomporre la matrice A in

$$N=D:=\mathrm{diag}(a_{11},....,a_{nn}) \qquad e \qquad P=N-A.$$

Poichè N deve essere invertibile, si deve premoltiplicare A per una matrice di permutazione in modo che la matrice permutata abbia sulla diagonale elementi tutti non nulli. Si vede che ciò è sempre possibile a condizione che A stessa sia non singolare.

, Risulta $N^{-1}=\mathrm{diag}(\frac{1}{a_{11}},....,\frac{1}{a_{nn}})$ e per la matrice di iterazione $M=(m_{ij})$ si ha:

$$m_{ij}= -\frac{a_{ij}}{a_{ii}} \quad , \quad \text{per } i\neq j \, , \quad \text{ed } m_{ii}=0.$$

L'iterazione di Jacobi si scrive dunque:

$$x_i^{k+1}= -\sum_{j\neq i} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}} \qquad i=1,...,n$$

Una matrice A si dice **a predominanza diagonale stretta**, se per ogni riga si ha:

$$\sum_{j\neq i} |a_{ij}|<|a_{ii}| \qquad (i=1,...,n) .$$

Poichè la predominanza diagonale stretta implica $\sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|}<1$ per ogni i, allora si dimostra immediatamente il seguente teorema

Teorema 2.2. *Se la matrice A è a predominanza diagonale stretta allora il metodo di Jacobi è convergente*.

Dim: $\| M \|_\infty = \max_i \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$

Un'altra condizione sufficiente per la convergenza è data da $\| M \|_1 <1$, cioè da

$$\max_j \sum_{i\neq j} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Si badi bene che la precedente relazione non significa la **predominanza diagonale per colonne**, che si esprime invece con $\sum_{i\neq j} |a_{ij}|<|a_{jj}|$ (j=1,...,n), ma una relazione più complessa che può comunque essere utile in qualche caso. D'altra parte la predominanza diagonale per colonne assicura anch'essa la convergenza del matodo di Jacobi.

Teorema 2.3. *Se la matrice A è a predominanza diagonale stretta per colonne, il metodo di Jacobi converge*.

Dim. Data la matrice B, indichiamo con B^{-T} la trasposta dell'inversa di B. La predominanza diagonale per colonne di A implica la predominanza diagonale per righe di A^T, la cui matrice di iterazione di Jacobi è

$$M'=D^{-T}P^T=D^{-1}P^T.$$

Per tale matrice si ha, in base al teorema precedente, $\rho(D^{-1}P^T)<1$ e quindi anche $\rho((D^{-1}P^T)^T)=\rho(PD^{-1})<1$. Poichè PD^{-1} è simile a $D^{-1}P$ anche $\rho(D^{-1}P)<1$.

Metodo di Gauss-Seidel (metodo delle sostituzioni successive).

Un'altra decomposizione di A in N-P, con N facilmente invertibile, è data da N=L+D e P=-U dove D è la diagonale ed L e U sono la parte triangolare inferiore e superiore di A. L'iterazione $Nx^{k+1} = Px^k + b$ assume la forma

$$
\begin{aligned}
&a_{11}\, x_1^{k+1} && = -a_{12}\, x_2^{k} - a_{13}\, x_3^{k} - \ldots\ldots - a_{1n}\, x_n^{k} + b_1 \\
&a_{21}\, x_1^{k+1} + a_{22}\, x_2^{k+1} && = \quad - a_{23}\, x_3^{k} - \ldots\ldots - a_{2n}\, x_n^{k} + b_2 \qquad \ldots\ldots \\
&\ldots\ldots \\
&a_{n1}\, x_1^{k+1} + a_{n2}\, x_2^{k+1} + \ldots\ldots + a_{nn}\, x_n^{k+1} && = \qquad\qquad + b_n
\end{aligned}
$$

Tale sistema si risolve facilmente ed ogni componente x_i^{k+1} è data da:

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{k} + \frac{b_i}{a_{ii}} \qquad i=1,\ldots,n$$

Come si può osservare, nel metodo delle sostituzione successive il valore aggiornato di ogni singola componente x_j^{k+1} viene utilizzato immediatamente per il calcolo delle componenti successive relative alla stessa iterazione, mentre nel metodo delle sostituzioni simultanee i valori aggiornati di tutte le componenti vengono sostituiti simultaneamente ai valori x_j^k alla fine dell'iterazione. A differenza del metodo di Jacobi, ora non è immediato valutare la matrice di iterazione $M = N^{-1}P = (L+D)^{-1}(-U)$ e quindi una sua norma, di conseguenza la nostra analisi della convergenza sarà effettuata analizzando l'errore componente per componente. A tale scopo, osservato che la soluzione del sistema Nx=Px + b si può scrivere

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}} \qquad i=1,\ldots,n$$

si ottiene, per gli errori $e_j^k := x_j^k - x_j$

$$e_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^{k}$$

$$|e_i^{k+1}| \le \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{k+1}| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{k}| \qquad i=1,...,n$$

$$\left|e_i^{k+1}\right| \le \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| \|e^{k+1}\|_\infty + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \|e^{k}\|_\infty \qquad i=1,...,n.$$

Detto M l'indice per il quale il secondo termine della disuguaglianza è massimo, e chiamato per brevità $R_M = \sum_{j=1}^{M-1} \left|\frac{a_{Mj}}{a_{MM}}\right|$ ed $S_M = \sum_{j=M+1}^{n} \left|\frac{a_{Mj}}{a_{MM}}\right|$, si ha

$$\left|e_i^{k+1}\right| \le R_M \|e^{k+1}\|_\infty + S_M \|e^{k}\|_\infty \quad \text{per ogni } i$$

e quindi

$$\|e^{k+1}\|_\infty \le R_M \|e^{k+1}\|_\infty + S_M \|e^{k}\|_\infty$$

$$(1-R_M)\|e^{k+1}\|_\infty \le S_M \|e^{k}\|_\infty$$

Se supponiamo A a predominanza diagonale stretta, allora $R_M+S_M<1$, da cui segue che $R_M<1$ ed $S_M<1$ e quindi

$$\|e^{k+1}\|_\infty \le \frac{S_M}{1-R_M} \|e^{k}\|_\infty \le \left(\frac{S_M}{1-R_M}\right)^{k+1} \|e^{0}\|_\infty$$

dove, per la predominanza diagonale, è ancora $\frac{S_M}{1-R_M}<1$. Abbiamo dimostrato dunque il seguente teorema.

Teorema 2.4. *Se la matrice A è a predominanza diagonale stretta allora il metodo di Gauss-Seidel è convergente*.

Sebbene la predominanza diagonale sia sufficiente per la convergenza di entrambi i metodi, esistono matrici A per le quali un metodo converge e l'altro no. In certi casi si può tuttavia fare un confronto tra il metodo di Jacobi e quello di Gauss-Seidel. Vale infatti il seguente teorema che viene enunciato senza dimostrazione.

Teorema 2.5. *Se la matrice* A *è tridiagonale , allora per le matrici di iterazione di Jacobi e di Gauss-Seidel vale la seguente relazione* $\rho(M_{GS})=\rho^2(M_J)$.

Il teorema rivela che i due metodi convergono o divergono entrambi e, quando convergono, il metodo di Gauss-Seidel è più veloce.

Nel caso di matrici A hermitiane e definite positive, si può dare il seguente criterio per la convergenza di un metodo iterativo generato da uno splitting.

Lemma 2.6: *Sia* A *hermitiana e definita positiva e sia* N *una matrice non singolare tale che* $Q:=N + N^H - A$ *sia ancora definita positiva. Allora la matrice di iterazione* $M=I-N^{-1}A$ *è convergente*.

Dim Sia λ un autovalore di M ed u una corrispondente autosoluzione. Allora:

$Mu=\lambda u$
$NMu=\lambda Nu$
$N(I-N^{-1}A)u=\lambda Nu$
$(N-A)u=\lambda Nu$
$(1-\lambda)Nu=Au$

Poichè A è non singolare, $Au\neq 0$ e quindi $(1-\lambda)\neq 0$, per cui:

$$Nu=\frac{Au}{1-\lambda}$$
$$u^HNu = \frac{u^HAu}{1-\lambda}$$

e, passando ai coniugati,

$$u^HN^Hu = \frac{u^HAu}{1-\bar{\lambda}} .$$

Sommando infine le ultime due relazioni:

$$u^H(N^H+N)u = u^HAu\left(2\,\mathcal{R}_e\left(\frac{1}{1-\lambda}\right)\right)$$

$$\left(2\,\mathcal{R}_e\left(\frac{1}{1-\lambda}\right)\right) = \frac{u^H(N^H+N)u}{u^HAu} = \frac{u^H(Q+A)u}{u^HAu} = \frac{u^HQu}{Au} + 1>1.$$

Sia $\lambda=\alpha+i\beta$, allora

$$\frac{1}{1-\lambda} = \frac{1-\alpha+i\beta}{(1-\alpha)^2+\beta^2}$$

da cui:

$$2\,\mathcal{R}e\left(\frac{1}{1-\lambda}\right) = \frac{2(1-\alpha)}{(1-\alpha)^2+\beta^2}$$

e la condizione $2\,\mathcal{R}e\left(\frac{1}{1-\lambda}\right) > 1$ diventa $\alpha^2+\beta^2<1$, cioè $|\lambda|<1$.

Con questo criterio, che non appare utile per la convergenza del metodo di Jacobi, si puo dare il seguente teorema per la convergenza del metodo di Gauss-Seidel.

Teorema 2.7. *Se la matrice A è hermitiana e definita positiva, il metodo di Gauss-Seidel è convergente*.

Dim. Per il metodo di Gauss-Seidel si ha $N=L+D$ e, poichè A è hermitiana, $A=L+D+L^H$. Quindi $Q=N + N^H - A = L+D+L^H+D-(L+D+L^H) = D$ che risulta definita positiva avendo tutti gli elementi positivi.

Osservazione. Il teorema precedente, pur avendo delle ipotesi abbastanza restrittive, può essere molto utile in casi più generali appena si osservi che il sistema $Ax=b$,con A non singolare, è equivalente al sistema $A^HAx=A^Hb$ la cui matrice A^HA è hermitiana e definita positiva. Nell'adottare questa strategia si deve, però, tener conto del rischio dovuto al fatto che il condizionamento della matrice A^HA è peggiore di quello di A (si veda il capitolo sul condizionamento).

Sulla base dei teoremi appena esposti, possiamo dire che il metodo di Gauss-Seidel è convergente per entrambi i sistemi generati dai problemi presentati all'inizio del capitolo, le cui matrici A sono simmetriche e definite positive. In particolare per il problema dei due punti, poichè la matrice A è anche tridiagonale, il metodo di Jacobi è ancora convergente ma con ordine di convergenza doppia.

Si dimostra che il raggio spettrale della matrice di iterazione di Jacobi relativa al sistema tridiagonale $Tx=b$ è $\rho(M_J)=\cos(\pi/(N+1))$ dove N è la dimensione di T. Osserviamo che per $N=10$ si ha $\rho(M_J)\approx 0.9595$ e quindi $\rho(M_{GS})\approx 0.9206$ che rivela una velocità di

convergenza molto lenta anche per il metodo di Gauss-Seidel. E' necessario quindi cercare degli ulteriori metodi, o delle modifiche ai metodi visti, che siano più veloci.

Il metodo SOR (successive over-relaxation method).

Il metodo SOR è una modifica del metodo di Gauss-Seidel e consiste nell'assumere come valore aggiornato della componente i-esima non il valore fornito dal metodo di Gauss-Seidel, che ora indichiamo con:

$$\bar{x}_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{k} + \frac{b_i}{a_{ii}}$$

bensì il valore

$$x_i^{k+1} = x_i^{k} + \omega\,(\bar{x}_i^{k+1} - x_i^{k}).$$

In altre parole si incrementa o si riduce il salto che il metodo di Gauss-Seidel provocherebbe alla componente i-esima con un opportuno **parametro di rilassamento** ω. Siccome per $\omega=1$ si ritrova il metodo di Gauss-Seidel, è ragionevole sperare che per valori diversi da 1 si abbia un metodo più veloce.

Fatte le opportune sostituzioni si trova

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

da cui si ricava:

$$a_{ii} x_i^{k+1} + \omega \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} = a_{ii}(1-\omega) x_i^{k} - \omega \sum_{j=i+1}^{n} a_{ij} x_j^{k} + \omega\, b_i$$

che rappresenta la riga i-esima del seguente sistema:

$$Dx^{k+1} + \omega L x^{k+1} = D(1-\omega)x^{k} - \omega U x^{k} + \omega b$$

$$(D + \omega L)x^{k+1} = \big((1-\omega)D - \omega U\big)x^{k} + \omega b.$$

Dividendo per ω , si ottiene lo splitting

$$\left(\frac{D}{\omega} + L\right)x^{k+1} = \left(\frac{1-\omega}{\omega} D - U\right) x^{k} + b$$

ed il metodo iterativo si esprime con

$$x^{k+1} = H_\omega x^k + \omega(D + \omega L)^{-1} b$$

dove la matrice di iterazione è:

$$H_\omega = (D + \omega L)^{-1}\big((1- \omega)D - \omega U\big).$$

Si noti che mentre nel metodo di Gauss-Seidel tutta la diagonale D appartiene alla parte N dello splitting, in SOR la parte $\frac{D}{\omega}$ sta in N mentre $\frac{1- \omega}{\omega} D$ sta in P.

Si tratta ora di vedere per quali valori del parametro di rilassamento si ha $\rho(H_\omega)<1$ ma, soprattutto, per quali valori di ω si ha un metodo più veloce di Gauss-Seidel ed in particolare quale è il valore ottimale di ω che minimizza $\rho(H_\omega)$.

Teorema 2.8 (di Kahan). *Per ogni matrice* A *tale che* $|D|\neq 0$, *si ha:*

$$\det(H_\omega)=(1 - \omega)^n$$

$$\rho(H_\omega)\geq |1 - \omega|.$$

Dim. Poichè il determinante di una matrice triangolare, o della sua inversa, dipende solo dagli elementi diagonali, si ha:

$$\det(H_\omega)=\det (D + \omega L)^{-1} \det\big((1- \omega)D - \omega U\big)=\det D^{-1} \det\big((1- \omega)D\big)=$$
$$=\det\big((1- \omega)I\big)=(1 - \omega)^n$$

$\rho(H_\omega)\geq|1-\omega|$ discende immediatamente dal fatto che il determinante di una matrice è il prodotto dei suoi autovalori.

Corollario 2.9. *Condizione necessaria affinchè* $\rho(H_\omega)<1$ *è* $0<\omega<2$.

Dim. $\rho(H_\omega)<1 \Rightarrow |1-\omega|<1 \Rightarrow 0<\omega<2$.

Nel caso di matrici hermitiane e definite positive, la condizione del corollario 2.9 è anche sufficiente per la convergenza:

Teorema 2.10 (di Reich-Ostrowski). *Se la matrice A è hermitina e definita positiva, condizione necessaria e sufficiente affinchè* $\rho(H_\omega)<1$ *è* $0<\omega<2$.

Dim. La matrice di splitting del metodo SOR è $N=(\frac{D}{\omega}+L)$per cui la matrice Q del lemma 2.6 è:

$$Q = \frac{D}{\omega} + L + \frac{D}{\omega} + L^H - (L + D + L^H) = D\left(\frac{2}{\omega} - 1\right)$$

che risulta definita positiva per $0<\omega<2$.

Metodi iterativi a blocchi.

Quando la matrice del sistema da risolvere presenta una struttura a blocchi, quale quella generata dalla discretizzazione del problema di Dirichlet, i metodi iterativi ora visti possono essere implementati a blocchi. Per fissare le idee, riferiamoci proprio all'esempio citato la cui matrice rappresenteremo ora con la seguente struttura *tridiagonale a blocchi:*

$$A = \begin{pmatrix} D_1 & A_1 & & \\ C_2 & D_2 & A_2 & \\ & C_3 & D_3 & A_3 \end{pmatrix}$$

dove i blocchi diagonali D_1,D_2,D_3,D_4 sono quadrati ed hanno dimensione, rispettivamente, 3,4,3,2, mentre i blocchi sopra e sottodiagonali A_i e C_i sono rettangolari ed inoltre $A_i=C_{i+1}^T$ per i=1,2,3. Ripartendo anche il vettore delle incognite $x=u(P_i)$ e dei termini noti $f=f(P_i)$ i=1,...,12 in 4 sottovettori $x=(x_1,x_2,x_3,x_4)$ ed $f=(f_1,f_2,f_3,f_4)$ di dimensioni uguali alle dimensioni dei blocchi diagonali, il sistema assume la seguente forma "tridiagonale a blocchi"

$$\begin{aligned} D_1x_1+A_1x_2 &= f_1 \\ C_2x_1+D_2x_2+A_2x_3 &= f_2 \\ C_3x_2+D_3x_3+A_3x_3 &= f_3 \\ C_4x_3+D_4x_4 &= f_4 \end{aligned}$$

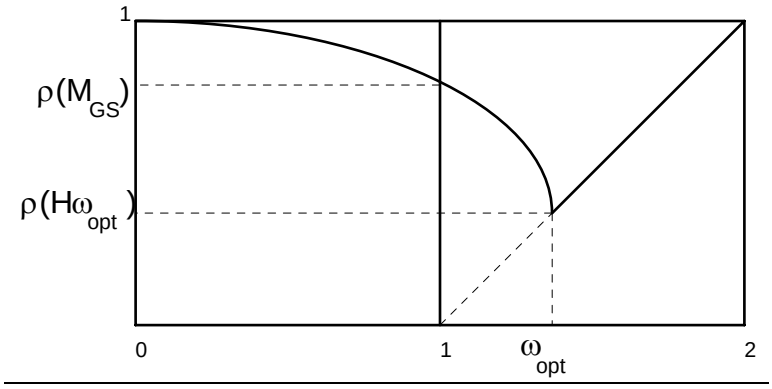
A tale sistema a blocchi, rappresentato formalmente come un sistema usuale, si possono applicare i metodi iterativi visti che si chiameranno metodo di "Jacobi, Gauss-Seidel ed SOR a blocchi". Ciascuno di loro richiede, al blocco i-esimo, la risoluzione di un sistema di matrice D_i.

Per le matrici tridiagonali a blocchi vale il seguente teorema analogo al teorema 2.5 che si riferiva ai metodi iterativi che ora chiameremo *puntuali* per distinguerli da quelli a blocchi.

Teorema 2.11. *Se la matrice* A *è tridiagonale a blocchi, allora per le matrici di iterazione di Jacobi e di Gauss-Seidel a blocchi vale la seguente relazione* $\rho(M_{GS})=\rho^2(M_J)$.

Concludiamo il paragrafo enunciando un teorema relativo alle matrici A hermitiane, definite positive e tridiagonali a blocchi che riassume in parte i teoremi già enunciati e che fornisce esplicitamente il valore del parametro ottimale di rilassamento in funzione del raggio spettrale della matrice di iterazione di Gauss-Seidel.

Teorema 2.12. *Se* A *è una matrice hermitiana, definita positiva e tridiagonale a blocchi, allora, per i metodi puntuali ed a blocchi, si ha* $\rho(M_{GS})=\rho^2(M_J)<1$. *Inoltre i metodi* SOR *puntuali ed* SOR *a blocchi, convergono per ogni* $0<\omega<2$ *ed il valore ottimale del parametro è* $\omega_{opt}=\frac{2}{1+\sqrt{1-\rho(M_{GS})}}>1$ *al quale corrisponde un raggio spettrale* $\rho(H_{\omega_{opt}}) = \omega_{opt} -1$. *Il grafico di* $\rho(H_\omega)$ *è dato, qualitativamente, dalla seguente figura*:



Abbiamo già osservato che il sistema relativo al problema di Dirichelet, presentato all'inizio, può essere risolto col metodo di Gauss-Seidel puntuale. Il teorema precedente ci assicura anche la convergenza del metodo di Gauss-Seidel a blocchi e di Jacobi a blocchi. In particolare quello di Jacobi è più lento.

Criteri d'arresto.

Per una effettiva implementazione dei metodi iterativi appena visti, è necessario un criterio d'arresto cioè un meccanismo automatico che interrompa il processo iterativo in base ad una stima dell'errore. Le stime dell'errore relative alla iterazione k-esima possono essere *stime a priori* oppure *stime a posteriori*. Le prime sono fondate solo sui dati del problema e sul punto iniziale dell'iterazione e danno una stima, a priori, del numero di iterazioni necessarie per approssimare la soluzione con un errore inferiore ad una tolleranza assegnata. Le seconde invece fanno uso anche dei valori forniti da ciascuna iterazione e sono quindi, presumibilmente, migliori.

Per una stima a posteriori, si osservi che

$$||x_{k+1}-x|| \le ||M||\,||x_k-x|| = ||M||\,||x_k-x_{k+1}+x_{k+1}-x|| < ||M||(||x_k-x_{k+1}||+||x_{k+1}-x||)$$

$$(1-||M||)||x_{k+1}-x|| \le ||M||\,||x_k-x_{k+1}||$$

$$||x_{k+1}-x|| \le \frac{||M||}{1-||M||}\,||x_k-x_{k+1}||.$$

Disponendo di una valutazione di ||M|| (<1) e memorizzando ad ogni passo gli ultimi due valori della traiettoria si ottiene, ad ogni iterazione, una stima *a posteriori* dell'errore.

Una stima *a priori* ricavata, in funzione del primo passo della traiettoria x_1-x_0, si può ottenere dalla precedente nel seguente modo.
Si osservi che

$$x_k-x_{k+1} = M(x_{k-1}-x_k) = \ldots = M^k(x_0-x_1)$$

e che

$$||x_k-x_{k+1}|| \le ||M||^k||x_1-x_0||.$$

Sostituendo quest'ultima nella precedente stima a posteriori si ottine:

$$||x-x_{k+1}|| \le \frac{1}{1-||M||}||M||^{k+1}||x_1-x_0||.$$

Da una stima di ||M|| (<1) e dalla conoscienza di x_1 e x_0 si può stimare a priori quante iterazioni occorrono per avere $||x-x_k|| < TOL$.

Un'altro criterio di arresto è fondato sul residuo $r_k:=Ax_k-b$ per il quale

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

da cui

$$||x_k-x||\leq||A^{-1}||\ ||r_k||.$$

Per una valutazione corretta dell'errore bisognerebbe disporre della quantità $||A^{-1}||$. In generale si può dire che, siccome $Ax=b$, si ha $||b|| \leq ||A||\ ||x||$ e $1/||x||\leq||A||/||b||$ e quindi

$$\frac{||x_k-x||}{||x||} \leq ||A^{-1}||\ ||A||\ \frac{||r_k||}{||b||}\ .$$

Ciò indica che un piccolo valore relativo del residuo assicura un piccolo valore dell'errore relativo solo per sistemi ben condizionati.

Esempio:

Consideriamo il sistema

$$\begin{cases} x+y=2 \\ x+1.01y=2.01 \end{cases}$$

e supponiamo di aver calcolato una soluzione approssimata

$$\bar{z}=(\bar{x},\bar{y})=(10,-8)$$

che fornisce un residuo r (0; -0.09), $||r||_\infty$= 0.09.
Per il termine noto b si ha $||b||_\infty$= 2.01 e quindi $\dfrac{||r||_\infty}{\|b\|_\infty} \approx 0.045$.

La soluzione vera invece è: z=(1,1). L'errore relativo è quindi $\dfrac{||z-\bar{z}||_\infty}{\|z\|_\infty} = 9$ che risulta circa 200 volte maggiore di $\dfrac{||r||_\infty}{\|b\|_\infty}$