<u>Sistemi lineari</u>

1. Generalità sui sistemi lineari

Consideriamo un sistema lineare

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3 \\ \\ a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + \ldots + a_{mn}x_n = b_m \end{cases}$$

Noti i coefficienti a_{ij}, *i=1,..m, j=1,n,* e le componenti del termine noto b_i, i=1,..*m*, si vuole individuare il vettore x incognito di n componenti che soddisfa contemporaneamente le *m relazioni lineari.*

In formato matriciale, il sistema po' essere espresso come:

$$Ax=b$$

con $A \in M(m \times n), x \in R^n, b \in R^m$

Matrice dei coefficienti

Vettore termini noti

$$\underbrace{\underbrace{\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{\mathbf{b}}}_{\mathbf{Ax=b}}$$

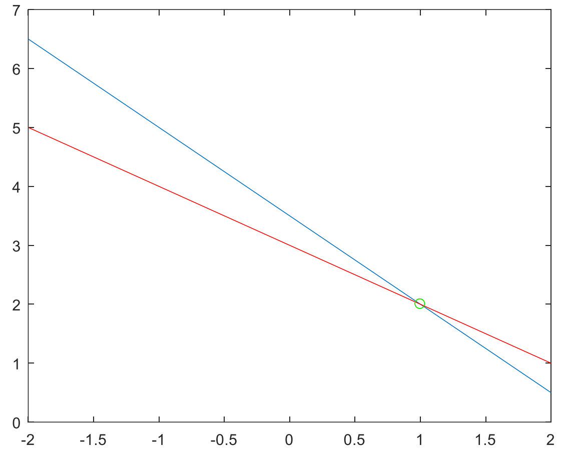
Vettore incognite

Un sistema lineare si dice **compatibile** se ammette almeno una soluzione, si dice **incompatibile** nel caso non ammetta alcuna soluzione

Interpretazione geometrica di sistemi lineari in R^2

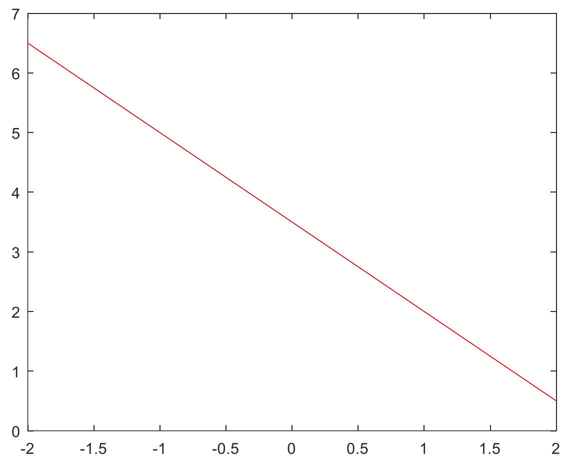
Punto di intersezione tra due rette

$$\begin{cases} 3x+2y=7 \\ x+y=3 \end{cases}$$



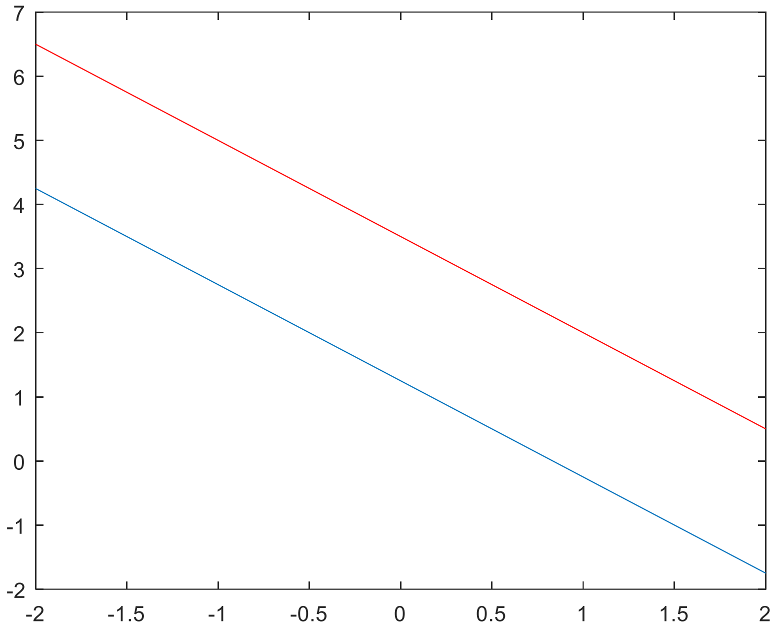
$$\begin{cases} 3x+2y=7 \\ 6x+4y=14 \end{cases}$$

Le due rette coincidono, infinite soluzioni



$$\begin{cases} 3x+2y=7 \\ 6x+4y=5 \end{cases}$$

Le due rette sono parallele, nessuna soluzione.



Teorema di Rouchè- Capelli

Il sistema lineare Ax=b ammette soluzioni se e solo se la matrice dei coefficienti A e la matrice completa [A b] hanno lo stesso rango:

$$\mathrm{Rank}(A)=r=\mathrm{Rank}([A\ b])$$

Altrimenti se (Rank(A)≠r≠Rank([A b])) il sistema non ammette soluzioni

Si possono verificare i seguenti casi:

Caso 1) $m<n$, cioè abbiamo **più incognite che relazioni lineari tra di esse**. Il sistema è del tipo

$$\begin{bmatrix} \\ & A & \\ & m<n & \\ \\ \end{bmatrix} \begin{bmatrix} \\ \\ x \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ b \\ \\ \end{bmatrix}$$

e si dice **indeterminato.** Se indichiamo con k il rango della matrice A e $k<n$, allora il sistema ammette ∞^{n-k} soluzioni.

2) Caso $m>n$, cioè abbiamo meno incognite che relazioni lineari tra di esse. Il sistema è del tipo:

$$\begin{bmatrix} \\ \\ & A & \\ & m>n & \\ \\ \\ \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ b \\ \\ \\ \\ \end{bmatrix}$$

e si dice sistema **sovradeterminato.** In questo caso il sistema non ammette una soluzione esatta, ma una soluzione approssimata.

3) Caso $m=n$, cioè il numero di incognite è uguale al numero di relazioni lineari tra di esse. Il sistema è della forma

$$\begin{bmatrix} & & \\ & A & \\ & m=n & \\ & & \\ & & \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ b \\ \\ \\ \\ \end{bmatrix}$$

e si dice sistema normale e sotto opportune ipotesi può ammettere una ed una sola soluzione.

Per il momento ci interesseremo di sistemi normali, con $n=m$.

Definizione:

A (nxn) è detta NON singolare se soddisfa una delle seguenti condizioni equivalenti:

1) $\det(A) \neq 0$

2) Se esiste la matrice inversa A^{-1} di A

3) rank(A)=n

Altrimenti A è singolare.

Teorema: *Condizione necessaria e sufficiente affinché il sistema lineare $Ax=b$, $A \in M(n \times n)$, $x, b \in R^n$ ammetta una ed una sola soluzione, comunque si scelga b, è che la matrice A sia a rango massimo (cioè che la matrice A sia invertibile); si ha perciò:*

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

Se il sistema è omogeneo (**b=0**), ed A è non singolare allora esiste solo la soluzione nulla **x=0**.

Un metodo ben noto per la soluzione di sistemi lineari è basato sulla regola di Cramer, che calcola la componente j-esima della soluzione facendo uso del calcolo del determinante (realizzato mediante la formula di Laplace). Questo metodo comporta una complessità computazionale di circa $(n+1)!$ operazioni.

Questo significa che per risolver un sistema con n=10 occorrerebbe fare 39916800 operazioni. Per n=20 , $(n+1)!=10^{21}$. Quindi se sono necessari 10^{-9} secondi per fare un prodotto, per risolvere un sistema di n=20 equazioni in n=20 incognite sono necessari 10^{12} secondi, che equivalgono a circa 1/3 di 10^{5} anni.

Nella pratica quindi vengono utilizzati altri metodi per la soluzione di un sistema lineare.

2. Metodi numerici per la soluzione di un sistema lineare.

I metodi per la soluzione di sistemi lineari vengono usualmente divisi in due raggruppamenti:

1) **Metodi diretti** $\Rightarrow$ Questi metodi, in assenza di errori di arrotondamento, conducono alla soluzione esatta in un numero finito di passi. Essi sono adatti per la soluzione di sistemi con **matrice dei coefficienti densa e di moderate dimensioni**.

2) **Metodi iterativi** $\Rightarrow$ Questi metodi generano una successione di soluzioni, che, sotto opportune ipotesi, convergono alla soluzione del sistema. La matrice dei coefficienti non viene modificata durante il calcolo e quindi è più agevole sfuttarne la sparsità. Sono adatti, quindi, per la soluzione di sistemi con **matrice dei coefficienti di grandi dimensioni e sparsa**. In assenza di errori di arrotondamento conducono alla soluzione esatta in un numero infinito di passi.

Durante il nostro corso ci interesseremo in particolare dei metodi iterativi per la soluzione di sistemi lineari. Di seguito solo un breve accenno al funzionamento dei metodi diretti, che si basano sulla fattorizzazione della matrice dei coefficienti.

3. Metodi di fattorizzazione per la soluzione di sistemi lineari normali.

Una classe di metodi di fattorizzazione consiste nel fattorizzare la matrice A nel prodotto di due matrici **triangolari,** una **inferiore L** ed una **superiore U**, tali che L ed U siano a **rango massimo**. Infatti, se

$$A=LU$$

il problema originale

$$Ax=b \quad \text{diventa} \quad LUx=b$$

che si può spezzare in due problemi di più facile soluzione

$$\begin{cases} Ly = b \\ Ux = y \end{cases}.$$

Fattorizzazione LU (di Gauss)

Fra i vari tipi esistenti di algoritmi di fattorizzazione riveste un ruolo importante la fattorizzazione LU di Gauss, che, nella sua versione base, esiste sotto l'ipotesi che la matrice A e tutti i minori principali di ordine k, k=1,..,n sia a rango massimo . Essa si basa sull'algoritmo di eliminazione di Gauss, che consiste nel sottoporre il sistema lineare che dobbiamo risolvere ad opportune trasformazioni in modo tale da eliminare successivamente le incognite dalle varie equazioni fino a ridursi ad un sistema triangolare superiore con matrice dei coefficienti U, e termine noto trasformato y, e quindi costruire la matrice L, triangolare inferiore, memorizzando i moltiplicatori utilizzati per trasformare la matrice del sistema A in triangolare superiore.

Costo computazionale della fattorizzazione $\frac{1}{3}\ O(n^3)$

$$L^{-1}[A \quad | \quad b] = [U|y]$$

$$L^{-1}[A \quad | \quad b] = [L^{-1}A \quad | \quad L^{-1}b] = [U|y]$$

$$L^{-1}A = U \Rightarrow A = LU$$
$$L^{-1}b = y \Rightarrow Ly = b$$

Soluzione di sistemi con matrici triangolari

Sia $L = \begin{bmatrix} l_{11} & 0 & \dots & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ l_{31} & l_{32} & l_{33} & & \\ \dots & \dots & \dots & \dots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{bmatrix}$ una matrice *triangolare inferiore* ed

$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix}$ una matrice *triangolare superiore*

La soluzione di sistemi lineari con queste matrici dei coefficienti si ottiene facilmente mediante sostituzione.

Soluzione del sistema Lx=b, nel caso in cui la matrice L sia triangolare inferiore

$$\begin{cases} l_{11}x_1 & = b_1 \\ l_{21}x_1 + l_{22}x_2 & = b_2 \\ \dots & \\ l_{i1}x_1 + l_{i2}x_2 + \dots + l_{ii}x_i & = b_i \\ \dots & \\ l_{n1}x_1 + l_{n2}x_2 + \quad \dots + l_{nn}x_n & = b_n \end{cases}$$

Sostituzione in avanti (Forward Substitution)

$$\begin{cases} x_1 = \dfrac{b_1}{l_{11}} \\ x_2 = (b_2 - l_{21}x_1)/l_{22} \\ \ldots \\ x_i = (b_i - (l_{i1}x_1 + l_{i2}x_2 + \ldots l_{i,i-1}x_{i-1}))/l_{i,i} \qquad i = 2,\ldots,n \end{cases}$$

Cioè, in forma algoritmica:

```
for  i=1,2,..,n
     x_i=b_i
          for j=1,2,…,i-1
              x_i=x_i-l_ij·x_j
          end for j
     x_i=x_i/l_ii
end for i
```

Soluzione del sistema Ux=b, nel caso in cui la matrice U sia triangolare superiore

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1,i}x_i + \cdots \qquad u_{1,n-1}\, x_{n-1} + u_{1n}x_n = b_1 \\ u_{i,i}x_i + u_{i,i+1}x_{i+1} + \cdots + u_{i,n}\, x_n = b_i \\ u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n = b_{n-1} \\ u_{nn}x_n = b_n \end{cases}$$

Sostituzione all'indietro (Backward Substitution)

$$\begin{cases} x_n = \dfrac{b_n}{u_{n,n}} \\ x_{n-1} = (b_{n-1} - u_{n-1,n}x_n)/u_{n-1,n-1} \\ \ldots \\ x_i = (b_i - (u_{i,i+1}x_{i+1} + u_{i,i+2}x_{i+2} + \ldots u_{i,n}x_n))/u_{i,i} \qquad i = n-1, n-2, \ldots, 1 \end{cases}$$

cioè, in forma algoritmica,

```
for  i=n,n-1,...,1
     x_i=b_i
          for j=i+1, ...,n
             x_i=x_i-u_ij·x_j
          end for j
x_i=x_i/u_ii
end for i
```

Complessità computazionale:

Calcoliamo la complessità computazionale dell'algoritmo di sostituzione in avanti.

Per calcolare la prima componente della soluzione è necessaria 1 divisione. Per calcolare la seconda componente della soluzione sono necessarie 1 moltiplicazione ed 1 divisione, quindi 2 operazioni moltiplicative

Per calcolare la componente i-esima della soluzione sono necessarie *i*-1 moltiplicazioni ed *1* divisione, quindi *i* operazioni moltiplicative.

La complessità computazionale in termini di operazioni moltiplicative è :

$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, quindi dell'ordine di $O\left(\frac{n^2}{2}\right)$.

Lo stesso risultato vale per la complessità computazionale dell'algoritmo di sostituzione all'indietro.

4. Metodi iterativi per la soluzione di sistemi lineari

Per la risoluzione di un sistema lineare $Ax = b$, oltre ai metodi diretti, è possibile utilizzare anche i metodi iterativi che raggiungono la soluzione esatta come limite di un procedimento iterativo.

Mentre i metodi diretti si basano su una fattorizzazione della matrice A e hanno una complessità di $O(n^3)$ i metodi iterativi si basano su una decomposizione della matrice A e presentano una complessità di $O(kn^2)$, dove k è il numero di iterazioni.

Pertanto tali metodi risultano particolarmente convenienti quando la matrice A del sistema è di grandi dimensioni e sparsa. Infatti, quando la matrice A è sparsa, cioè il numero degli elementi non nulli è di molto inferiore al numero degli elementi nulli, applicando i metodi diretti può accadere che vengano generati elementi non nulli in corrispondenza degli elementi nulli della matrice di partenza (fenomeno del **fill-in**).

Questo non avviene applicando i metodi iterativi in quanto essi si limitano ad utilizzare gli elementi non nulli della matrice senza toccare gli elementi nulli.

4.1 Metodi iterativi basati sulla decomposizione di A

Sia A matrice di ordine n e sia dato il sistema

$$Ax = b, \qquad \text{con } \det A \neq 0. \tag{1}$$

Una famiglia di metodi iterativi si ottiene utilizzando una **decomposizione della matrice *A*** nella forma

$$A = M - N \qquad \text{con } \det M \neq 0.$$

In tal modo il sistema (1) diventa

$$(M-N)\,x = b,$$

cioè

$$Mx = Nx + b$$

e quindi si ottiene

$$x = M^{-1}Nx + M^{-1}b \tag{1.1}$$

La soluzione di (1) risolve anche (1.1)

Questa formulazione suggerisce di considerare metodi le cui iterazioni siano fornite dalle iterazioni successive

$$x^{(k)} = M^{-1}Nx^{(k-1)} + M^{-1}b \qquad \text{per } k=1,2,\ldots.$$

L'idea dei metodi iterativi è quindi quella di partire da un vettore $x^{(0)}$ iniziale arbitrario, stima iniziale della soluzione del sistema (1) e di costruire una successione di iterati $x^{(k)}$ mediante il seguente procedimento iterativo

$$\boxed{x^{(k)} = Tx^{(k-1)} + q} \qquad \text{per } k=1,2,\ldots. \quad (2)$$

dove $T= M^{-1}N$, è detta la **matrice di iterazione** del metodo iterativo e $q= M^{-1}b$.

I metodi iterativi della forma (2) sono detti usualmente iterativi stazionari, perché T e q non dipendono dall'indice di iterazione k.

4.2 Esempi di metodi iterativi basati sulla decomposizione di *A*

1) **Il metodo di Jacobi** (o degli spostamenti simultanei)

2) **Il metodo di Gauss-Seidel** (o degli spostamenti successivi).

In entrambi i metodi si considera la matrice A del sistema decomposta come somma di 3 matrici, cioè

$$A = D + E + F$$

con

$$D=\begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix};\quad E=\begin{bmatrix} 0 & & & 0 \\ a_{21} & 0 & & \\ a_{31} & & \ddots & \\ \vdots & & & 0 \\ a_{n1} & & a_{nn-1} & 0 \end{bmatrix};\quad F=\begin{bmatrix} 0 & a_{12} & & a_{1n} \\ & 0 & & \\ & & \ddots & \\ & & & a_{n-1n} \\ & & & 0 \end{bmatrix}.$$

e si suppone che $a_{ii} \neq 0$ i=1,2,..n.

<u>Metodo di Jacobi</u>

Nel metodo di Jacobi la decomposizione di A nella forma A=M-N si ottiene scegliendo M=D e N=- (E+F).

Pertanto il procedimento iterativo (2) diviene

$$x^{(k)} = -D^{-1} (E+F)\, x^{(k-1)} + D^{-1}b \qquad \text{per } k=1,2,.. \quad (3)$$

In termini di componenti la (3), equivale a calcolare la i-esima componente dell'iterato k-esimo come

$$x_i^{(k)} = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} -a_{ij}x_j^{(k-1)} + b_i}{a_{ii}} \qquad i = 1,2..,n, \qquad (4)$$

in quanto la matrice D^{-1} è una matrice diagonale con elementi uguali ai reciproci degli elementi di D.

La matrice di iterazione del metodo di Jacobi è quindi data da

$$\boldsymbol{T_J = M^{-1}N = -D^{-1}(E+F)}$$

Osservazioni:

1) L'algoritmo di Jacobi è definito se gli elementi diagonali di *A* sono diversi da 0, cioè $a_{ii} \neq 0$. In caso contrario, sempre sotto l'ipotesi che *A* sia non singolare,

si possono riordinare le equazioni, e le incognite del sistema, in modo da rendere il metodo definito.

2) Nel metodo di Jacobi ogni elemento dell'iterato (k) è indipendente dagli altri, pertanto il metodo è programmabile in forma parallela.

Esempio:

Si consideri il seguente sistema lineare:

$$\begin{aligned} 10x_1 - x_2 + 2x_3 \qquad &= 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 &= 25 \\ 2x_1 - x_2 + 10x_3 - x_4 \quad &= -11 \\ 3x_2 - x_3 + 8x_4 \quad &= 15 \end{aligned}$$

Il metodo di Jacobi consiste nello scegliere un vettore iniziale $x^0 = (x_1^0, \ldots, x_4^0)$ e nel ricavare

$$\begin{aligned} x_1^{(1)} &= (x_2^{(0)} - 2x_3^{(0)} + 6)/10 \\ x_2^{(1)} &= (x_1^{(0)} + x_3^{(0)} - 3x_4^{(0)} + 25)/11 \\ x_3^{(1)} &= (-2x_1^{(0)} + x_2^{(0)} + x_4^{(0)} - 11)/10 \\ x_4^{(1)} &= (-3x_2^{(0)} + x_3^{(0)} + 15)/8 \end{aligned}$$

Si va poi avanti in questo modo finché non si ritiene di essere sufficientemente vicini alla soluzione. Occorre pertanto definire un criterio d'arresto.

<u>Metodo di Gauss-Seidel</u>

Nel metodo di Gauss-Seidel la decomposizione di A nella forma $A=M-N$ si ottiene scegliendo $M=E+D$ e $N=-F$. In questo caso la matrice di iterazione del metodo di Gauss Seidel è data da $\boldsymbol{T_G = M^{-1}N = -(E+D)^{-1}F}$ ***e la soluzione al passo k si ottiene come***

$$x^{(k)} = -(E+D)^{-1}F\, x^{(k-1)} + (E+D)^{-1}b \qquad \text{per } k=1,2,.. \qquad (5)$$

Per esprimere la soluzione in termini di componenti, partiamo da

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

Sostituendo ad M la matrice E+D e ad N la matrice –F, otteniamo

$$(E+D)\, x^{(k)} = -F\, x^{(k-1)} + b \qquad \text{per } k=1,2,..$$

$$D\, x^{(k)} = -E\, x^{(k)} \quad -F\, x^{(k-1)} + b \qquad \text{per } k=1,2,..$$

da cui

$$x^{(k)} = -\, D^{-1}\, (E\, x^{(k)} + F\, x^{(k-1)} - b\,) \qquad \text{per } k=1,2,.. \qquad (6)$$

e, in termini di componenti,

$$x_i^{(k)} = \frac{-\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + b_i}{a_{ii}} \qquad i = 1,2..,n \qquad (7)$$

Questo metodo ha la caratteristica di utilizzare, per calcolare la nuova componente i-esima di un iterato, le nuove componenti già calcolate dell'iterato stesso.

Osservazione:

Il metodo di Gauss-Seidel non si presta ad essere parallelizzato in quanto ogni nuova componente dell'iterato (k) dipende da tutte le nuove componenti dello stesso iterato che sono state appena calcolate.

Esempio:

Si consideri il sistema lineare dell'esempio precedente; partendo dall'iterato iniziale $x^0 = (x_1^0, ..., x_4^0)$ si ottiene:

$$x_1^{(1)} = (x_2^{(0)} - 2x_3^{(0)} + 6)\Big/10$$
$$x_2^{(1)} = (x_1^{(1)} + x_3^{(0)} - 3x_4^{(0)} + 25)\Big/11$$
$$x_3^{(1)} = (-2x_1^{(1)} + x_2^{(1)} + x_4^{(0)} - 11)\Big/10$$
$$x_4^{(1)} = (-3x_2^{(1)} + x_3^{(1)} + 15)\Big/8$$

e si continua così fino a convergenza.

4.3 Convergenza

Le prime domande a cui dobbiamo dare risposta sono: la successione di iterati $x^{(k)}$ è una successione convergente? In caso affermativo, il suo limite è proprio la soluzione x del sistema (1)?

<u>Definizione:</u> Convergenza

Il procedimento iterativo

$$x^{(k)} = Tx^{(k-1)} + q$$

*si dice "**convergente**" se, per ogni vettore iniziale $x^{(0)}$, la successione $\{x^{(k)}\}$ converge ad un vettore limite y;*

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

cioè se, per ogni $\varepsilon>0$, esiste un indice ν tale che per ogni $k>\nu$ si ha

$$\| x^{(k)} - y \| \leq \varepsilon.$$

Il seguente teorema dà una risposta affermativa alla seconda delle domande che ci eravamo posti.

Teorema:

Se il sistema (1) ammette un' unica soluzione x e se il processo iterativo (2) è convergente, allora il vettore limite y coincide con la soluzione x, cioè

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

Dimostrazione:

Si parte da

$$Mx^{(k)} = Nx^{(k-1)} + b$$

e si considera il limite per $k \to \infty$ di entrambi i membri

$$\lim_{k\to\infty} M\, x^{(k)} = \lim_{k\to\infty} \left(N\, x^{(k-1)} + b\right)$$

Poichè $x^{(k)}$ è per ipotesi una successione convergente ad y segue che

$$My \quad = Ny \quad + b$$

da cui si ottiene $(M\text{-}N)y = b$. Essendo $M\text{-}N=A$, si ha $Ay=b$ e quindi $y=x$, poiché per ipotesi AX=b ha un'unica soluzione x.

4.4 Convergenza dei Metodi Iterativi

Per effettuare lo studio della convergenza di un metodo iterativo si considera la matrice di iterazione *T*, tipica del metodo stesso.

Inoltre si definisce l' **errore commesso al passo *k*** come il vettore

$$e^{(k)} = x^{(k)} - x, \quad k = 0,1,\dots.$$

e il **vettore residuo al passo *k*** dato da

$$r^{(k)} = Ax^{(k)} - b, \quad k = 0,1,\dots.$$

Si osservi che queste due quantità sono legate dalla seguente relazione:

$$r^{(k)} = Ax^{(k)} - b = Ax^{(k)} - Ax = A(x^{(k)} - x) = Ae^{(k)}.$$

Consideriamo ora la relazione relativa al valore esatto x

$$Mx = Nx + b$$

e la relazione analoga, al generico passo k,

$$M\, x^{(k)} = Nx^{(k-1)} + b.$$

Sottraendo la prima dalla seconda, otteniamo $Me^{(k)} = Ne^{(k-1)}$ da cui

$$e^{(k)} = M^{-1}Ne^{(k-1)} = Te^{(k-1)} = T^2e^{(k-2)} = \ldots = T^k e^{(0)}$$

Affinché il procedimento sia convergente si deve avere che, comunque si sceglie $x^{(0)}$, ciascuna componente del vettore $e^{(k)}$ tenda a 0 per k che tende a ∞.

Questo equivale a cercare delle condizioni per cui

$$\lim_{k\to\infty} T^k e^{(0)} = 0$$

che equivale a

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

Teorema: **Condizione necessaria e sufficiente alla Convergenza**

Sia A=M-N una matrice di ordine n, con det(A)≠0, e $T=M^{-1}N$ *la matrice di iterazione del procedimento iterativo (2). Condizione necessaria e sufficiente per la convergenza del procedimento iterativo, comunque si scelga il vettore iniziale* $x^{(0)}$*, al vettore soluzione x del sistema Ax=b , è che*

$$\rho(T) < 1$$

ovvero che il raggio spettrale (che corrisponde all'autovalore di modulo massimo) della matrice di iterazione T sia minore di 1.

4.5 Condizioni Sufficienti per la Convergenza:

Il calcolo del raggio spettrale della matrice di iterazione T è, anche utilizzando algoritmi idonei, piuttosto oneroso. Si preferisce quindi o considerare condizioni anche più restrittive, ma facilmente verificabili, o individuare classi di matrici per cui la convergenza è garantita da risultati teorici. I seguenti teoremi servono a questi scopi.

Il teorema che segue ci garantisce una condizione sufficiente alla convergenza.

Teorema:

Se, per una qualche norma, risulta $||T||<1$*, allora il processo iterativo*

$$x^{(k)} = Tx^{(k-1)} + M^{-1}b \qquad per\ k=1,2,\ldots.$$

è convergente per ogni $x^{(0)}$*.*

Dim:

Dalla definizione di autovalore di una matrice si ha $Tx=\lambda x$, con $x\neq 0$, da cui si ottiene

$$|\lambda|\ ||x|| = ||Tx|| \leq ||T||\ ||x||$$

e quindi

$$|\lambda| \leq ||T||.$$

Poiché, per ipotesi, si ha $||T||<1$ allora anche il raggio spettrale di T risulta minore di 1 e, per il teorema precedente, il procedimento iterativo è convergente.

Seguono ora due teoremi che garantiscono la convergenza per classi particolari di matrici.

Teorema:

Se la matrice A è a ***diagonale strettamente dominante****, cioè*

$$|a_{ii}| > \sum_{\substack{k=1 \\ k\neq i}}^{n} |a_{ik}| \quad i = 1,2,\ldots,n$$

Allora sia il metodo di Jacobi che quello di Gauss-Seidel convergono e si ha

$$||T_G|| \leq ||T_J|| < 1$$

Teorema:

Se la matrice A è simmetrica e definita positiva, il metodo di Gauss-Seidel è convergente.

4.6 . Velocità di convergenza e condizionamento di una matrice

Un procedimento iterativo, convergente se l'autovalore di modulo massimo è minore di 1, è tanto più velocemente convergente quanto più piccolo è l'autovalore di modulo massimo della matrice di iterazione T.

Un problema strettamente legato alla convergenza del procedimento iterativo è quello del mal condizionamento della matrice A. L'influenza del condizionamento di A è infatti responsabile del rallentamento o addirittura della perdita della convergenza nei metodi iterativi.

Esempio:

$$A = \begin{bmatrix} .96326 & .81321 \\ .81321 & .68685 \end{bmatrix} \qquad\qquad b = \begin{bmatrix} .88824 \\ .74988 \end{bmatrix}$$

A è una matrice simmetrica definita positiva.

Il metodo di Gauss-Seidel, partendo dai valori iniziali $x^{(0)} = \begin{bmatrix} .33116 \\ .70000 \end{bmatrix}$ non converge alla soluzione

$$x^* = \begin{bmatrix} .39473 \\ .62470 \end{bmatrix}.$$

Questo comportamento si verifica per qualsiasi scelta di $x_2^{(0)}$ compresa tra .4 e .8.

Infatti l'indice di condizionamento di A è K(A)= 8936.

4.7 .Accelerazione di un metodo iterativo

Ora ci si chiede se è possibile accelerare la convergenza di un metodo iterativo e addirittura se è possibile rendere convergente un metodo che non lo è. Come risposta a questo quesito è nata una famiglia di metodi nota come **metodi di rilassamento**.

L'idea alla base di questi metodi è la seguente: poiché la velocità di convergenza di un metodo iterativo dipende dal raggio spettrale della matrice di iterazione associata al metodo, un modo per cercare di accelerare la convergenza è quello di far dipendere la matrice di iterazione da un parametro, detto **parametro di rilassamento,** e di scegliere tale parametro in modo tale che la matrice abbia minimo raggio spettrale.

Vediamo come viene generato un metodo di rilassamento a partire dal metodo di Gauss-Seidel.

Il metodo di Gauss-Seidel

$$x^{(k)} = -D^{-1}\,(E\,x^{(k)} + F\,x^{(k-1)} - b\,) \qquad \text{per } k=1,2,..$$

può essere riscritto nella forma

$$x^{(k)} = x^{(k-1)} + r^{(k)}$$

dove

$$r^{(k)} = x^{(k)} - x^{(k-1)} = -D^{-1}[Ex^{(k)} + Fx^{(k-1)} - b] - x^{(k-1)}. \quad (8)$$

Modificando il metodo di Gauss-Seidel come

$$x^{(k)} = x^{(k-1)} + \omega r^{(k)} \quad (9)$$

e scegliendo opportunamente il parametro ω>0 si può accelerare la convergenza in modo significativo. A seconda di come viene scelto tale parametro si distinguono i metodi di rilassamento in

a) ***under-relaxation methods*** con 0< ω<1

b) ***over-relaxation methods*** ω>1.

I metodi di ***under-relaxation*** possono essere usati per ottenere convergenza su certi sistemi per cui non si ha convergenza con il metodo di Gauss-Seidel, mentre i metodi di ***over-relaxation*** possono essere usati per accelerare la convergenza in sistemi in cui il metodo di Gauss-Seidel converge ma lentamente.

Questi ultimi metodi vengono chiamati metodi SOR (Successive Over-Relaxation) e vengono spesso impiegati per risolvere sistemi lineari che si incontrano nella soluzione numerica di alcune equazioni alle derivate parziali.

Il metodo SOR si ottiene sostituendo la (8) nella (9), ovvero:

$$x^{(k)} = (1-\omega)x^{(k-1)} - \omega D^{-1}[Ex^{(k)} + Fx^{(k-1)} - b], \quad (10)$$

da cui si ricava

$$x^{(k)} = (1-\omega)x^{(k-1)} + \omega[-D^{-1}[Ex^{(k)} + Fx^{(k-1)} - b]], \quad (11)$$

In termini di componenti si ha che la componente i-esima della soluzione al passo k, $x_i^{(k)}$, diviene

$$x_i^{(k)} = (1-\omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right] \quad i = 1,2..,n.$$

Questa variante del metodo di Gauss Seidel, dipendente dal parametro ω, è detta anche metodo di Gauss-Seidel estrapolato e può essere vista come

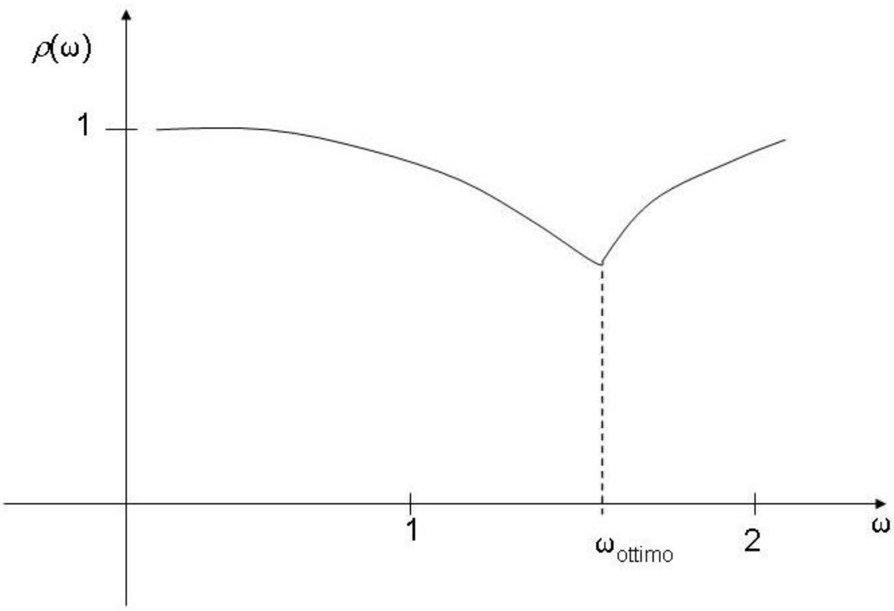
$$\begin{cases} \tilde{x}_i^{(k)} = \dfrac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}\right] \\ x_i^{(k)} = (1-\omega) x_i^{(k-1)} + \omega \tilde{x}_i^{(k)} \end{cases}$$

La matrice di iterazione del metodo è

$$T_\omega = (D+\omega E)^{-1}[(1-\omega)D - \omega F]. \tag{12}$$

Nell'applicazione di un metodo di rilassamento dipendente dal parametro nasce il problema della scelta ottimale del parametro ω che, oltre ad assicurare la convergenza, renda minimo il raggio spettrale della matrice di iterazione T_ω in modo da ottenere la massima velocità di convergenza.

E' noto infatti che, per matrici simmetriche definite positive e per $0 < \omega < 2$, il valore di $\rho(T_\omega)$ si mantiene <1, ma c'è un punto in cui è minimo. Il valore corrispondente al minimo del raggio spettrale è detto ω_{ottimo}.



In genere il calcolo di ω_{ottimo} è un problema molto laborioso, per cui si suole andare per tentativi.

4.8 Criterio d'arresto

Poiché con un metodo iterativo non è ovviamente possibile calcolare in generale la soluzione in un numero finito di iterazioni, occorre individuare dei criteri per l'arresto del procedimento. I criteri più comunemente usati, che consistono nel fissare una tolleranza ε che tiene conto della precisione utilizzata nei calcoli, sono i seguenti:

$$||x^{(k)} - x^{(k-1)}|| \leq \varepsilon \tag{13}$$

oppure se $x^{(k)} \neq 0$

$$\frac{||x^{(k)} - x^{(k-1)}||}{||x^{(k)}||} \leq \varepsilon. \tag{14}$$

La scelta della tolleranza ε nel criterio d'arresto viene fatta considerando la percentuale d'errore da cui sono affetti i dati iniziali.

La scelta del tipo di norma, in genere, dipende dallo specifico problema in esame; comunque norme comunemente usate sono la norma ∞ e la norma 2