▷ Lezione IV

Metodi iterativi per sistemi lineari

Libro Capitolo 5, 5.9, 5.10, 5.11.

Metodi iterativi per sistemi lineari

Dati una matrice $A \in \mathbb{R}^{n\times n}$ e un vettore $\boldsymbol{b} \in \mathbb{R}^n$, vogliamo determinare $\boldsymbol{x} \in \mathbb{R}^n$ tale che risolva il sistema lineare.

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

Un metodo iterativo costruisce una successione $\{\boldsymbol{x}^k\}$, dove ogni $\boldsymbol{x}^k \in \mathbb{R}^n$, che converge alla soluzione $\boldsymbol{x}$ per $k \to \infty$. Abbiamo quindi

$$\boldsymbol{x}^0 \xrightarrow{\text{un passo}} \boldsymbol{x}^1 \xrightarrow{\text{un passo}} \boldsymbol{x}^2 \xrightarrow{\text{un passo}} \dots \xrightarrow{\text{un passo}} \boldsymbol{x}^k \xrightarrow{k\to\infty} \boldsymbol{x}.$$

Più formalmente possiamo introdurre la seguente definizione.

Definizione 4.1 - Metodo iterativo convergente

Dato un vettore iniziale $\boldsymbol{x}^0$, un metodo iterativo si dice convergente se, detta x la soluzione esatta del sistema lineare,

$$\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x} \quad \iff \quad \lim_{k\to\infty} \boldsymbol{e}^k = \boldsymbol{0},$$

dove $\boldsymbol{e}^k = \boldsymbol{x} - \boldsymbol{x}^k$ è l'errore commesso al passo k.

Una forma generale per i metodi iterativi è data dall'espressione seguente

$$\boldsymbol{x}^{k+1} = B\boldsymbol{x}^k + \boldsymbol{g},$$

dove $B \in \mathbb{R}^{n\times n}$ è detta matrice di iterazione, che generalmente può dipendere da A, e $\boldsymbol{g}$ un vettore che viene costruito partendo da A e $\boldsymbol{b}$. La matrice B e il vettore $\boldsymbol{g}$ non possono essere presi in maniera arbitraria, infatti devono essere scelti in modo che il metodo sia consistente.

Definizione 4.2 - Metodo iterativo consistente

Un metodo della forma

$$\boldsymbol{x}^{k+1} = B\boldsymbol{x}^k + \boldsymbol{g},$$

è detto consistente se B e $\boldsymbol{g}$ sono tali che

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{g},$$

ossia se la soluzione esatta soddisfa esattamente il metodo numerico.

Notiamo che se il metodo è consistente, se per un certo k abbiamo che $\boldsymbol{x}^k = \boldsymbol{x}$ allora anche $\boldsymbol{x}^{k+1} = B\boldsymbol{x} + \boldsymbol{g} = \boldsymbol{x}$, ossia il metodo si "ferma" sulla soluzione esatta.

Teorema 4.1

Denotiamo con $\boldsymbol{e}^k = \boldsymbol{x} - \boldsymbol{x}^k$ è l'errore commesso al passo k, ogni metodo consistente soddisfa

$$\boldsymbol{e}^{k+1} = B\boldsymbol{e}^k$$

Proof. A partire dalla definizione di $\boldsymbol{e}^{k+1}$ sfruttiamo la consistenza del metodo per ottenere

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

□

La matrice di iterazione B applicata all'errore al passo k, $\boldsymbol{e}^k$, restituisce l'errore al passo successivo $\boldsymbol{e}^{k+1}$. Prima di introdurre il seguente teorema introduciamo cosa intendiamo per raggio spettrale della matrice, è un numero positivo che indichiamo con $\rho(B)$ ed è dato da

$$\rho(B) = \max_{i=1,\dots,n} |\lambda_i(B)|,$$

con $\lambda_i(B)$ autovalore i-esimo di B. Il raggio spettrale soddisfa la seguente proprietà

$$\lim_{k\to\infty} B^k = 0 \quad \iff \quad \rho(B) < 1.$$

ed inoltre per ogni norma di matrice $\|\cdot\|$ otteniamo che

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

Teorema 4.2 - Condizione necessaria e sufficiente di convergenza

Un metodo iterativo consistente della forma

$$\boldsymbol{x}^{k+1} = B\boldsymbol{x}^k + \boldsymbol{g},$$

è convergente se e solo se

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

Proof. Mostriamo l'implicazione che se il raggio spettrale soddisfa $\rho(B) < 1$ allora il metodo è convergente. Essendo per ipotesi un metodo consistente, allora vale

$$\boldsymbol{e}^{k+1} = B\boldsymbol{e}^k,$$

inoltre, ripetendo il ragionamento per $k-1$, $k-2$ ... si ottiene che

$$\boldsymbol{e}^{k+1} = B^{k+1}\boldsymbol{e}^0.$$

Dato che $\rho(B) < 1$ sappiamo che $\lim_{k\to\infty} B^k = 0$, quindi otteniamo che

$$\lim_{k\to\infty} \boldsymbol{e}^{k+1} = \lim_{k\to\infty} B^{k+1}\mathbf{e}^0 = \mathbf{0}.$$

Mostriamo ora l'implicazione inversa, ovvero vogliamo mostrare che se il raggio spettrale è maggiore di 1 allora il metodo non è convergente. Supponiamo che $\rho(B) \geq 1$, cioè esiste almeno un autovalore λ di B tale che $|\lambda| \geq 1$. Scelgo $\boldsymbol{x}^0$ tale che l'errore iniziale, $\boldsymbol{e}^0 = \boldsymbol{x} - \boldsymbol{x}^0$, sia uguale all'autovettore associato a λ. Abbiamo quindi

$$B\boldsymbol{e}^0 = \lambda \boldsymbol{e}^0$$

e otteniamo che l'errore al passo k si esprime come

$$\boldsymbol{e}^k = B\boldsymbol{e}^{k-1} = B^2\boldsymbol{e}^{k-2} = \ldots = B^k\boldsymbol{e}^0. = \lambda^k \boldsymbol{e}^0$$

Quest'ultima espressione non può convergere a zero per $k \to \infty$ dato che $|\lambda| > 1$. Quindi il raggio spettrale della matrice di iterazione per un metodo convergente deve necessariamente essere minore di 1. □

Esempio 4.1

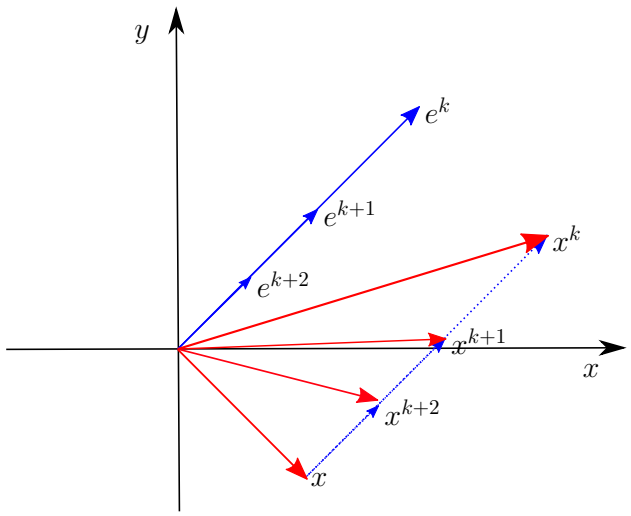
Sia data la matrice di iterazione B seguente

$$B = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

chiaramente gli autovalori associati a B sono $\lambda_1 = \lambda_2 = \frac{1}{2}$, che implica che il raggio spettrale è dato da $\rho(B) = \frac{1}{2}$. Supponendo che l'errore al passo k sia dato dal vettore $\boldsymbol{e}^k = [5, 7]^\top$ possiamo calcolare l'errore al passo $k+1$ come

$$\boldsymbol{e}^{k+1} = B\boldsymbol{e}^k = \frac{1}{2}\begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

Graficamente, proseguendo con le iterazioni, otteniamo il seguente andamento



in cui si osserva che, nel caso in cui $\rho(B) < 1$, applicare la matrice B ad un vettore lo fa accorciare. Di conseguenza, se l'errore diminuisce, la soluzione numerica si avvicina a quella esatta.

Consideriamo ora la matrice A scomposta tramite il seguente splitting additivo, ovvero data $P \in$

$\mathbb{R}^{n\times n}$ abbiamo che

$$A = P - (P - A), \tag{1.1}$$

dove la matrice P viene detta matrice precondizionatore. Deriviamo ora la matrice di iterazione B e il vettore $\boldsymbol{g}$ data la scomposizione proposta, risulta

$$A\boldsymbol{x} = \boldsymbol{b} \quad \to \quad [P - (P - A)]\boldsymbol{x} = \boldsymbol{b} \quad \to \quad P\boldsymbol{x} = (P - A)\boldsymbol{x} + \boldsymbol{b} \quad \to \quad \boldsymbol{x} = P^{-1}(P - A)\boldsymbol{x} + P^{-1}\boldsymbol{b}$$

Se definiamo $B = P^{-1}(P-A)$ e $\boldsymbol{g} = P^{-1}\boldsymbol{b}$, allora possiamo scrivere il metodo iterativo, che risulta consistente per costruzione, come

$$\boldsymbol{x}^{k+1} = B\boldsymbol{x}^k + \boldsymbol{g}.$$

Definiamo il residuo del problema al passo k come

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

chiaramente il residuo è nullo nel caso in cui $\boldsymbol{x}^k$ è uguale alla soluzione esatta del problema lineare $\boldsymbol{x}$. Abbiamo che

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

Ricordiamo che la formulazione di metodi diretti e iterativi per sistemi lineari ha lo scopo di evitare di dover invertire la matrice A per poi applicarla al sistema $A^{-1}A\boldsymbol{x} = A^{-1}\boldsymbol{b}$ per ottenere $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Questa operazione avrebbe infatti un costo computazionale proibitivo. Tuttavia, notiamo che considerando la scomposizione (1.1), ad ogni passo k dobbiamo comunque risolvere un sistema lineare dato da

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

quindi questo approccio ha senso solo se la risoluzione del sistema lineare in P è estremamente meno costosa che la risoluzione del sistema originale in A. Devo quindi scegliere P in modo che sia invertibile ma allo stesso tempo che il sistema lineare associato sia facile da risolvere. Diversi metodi iterativi sono caratterizzati da diverse scelte di P, sempre richiedendo che P sia invertibile e che il sistema lineare associato sia rapido da risolvere. Tipiche scelte di P sono matrice diagonale e matrice triangolare.

Metodo di Jacobi

Il metodo di Jacobi definisce la matrice di precondizionamento P come la matrice diagonale D, che ha per diagonale la stessa della matrice A. Ovvero $P = D = \text{diag}(A)$ dove

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & & & & \\ a_{n1} & & \dots & a_{n,n-1} & a_{nn} \end{bmatrix} \xrightarrow{D=\text{diag}(A)} D = \begin{bmatrix} a_{11} & 0 & \dots & & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \dots & & & & \\ 0 & & \dots & 0 & a_{nn} \end{bmatrix}$$

Abbiamo quindi che il metodo iterativo di Jacobi risulta scritto come

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

Abbiamo sfruttato il fatto che D^{-1} è una matrice diagonale con elementi $\frac{1}{a_{ii}}$ e inoltre, premoltiplicare per una matrice diagonale equivale a effettuare uno scaling delle righe. Possiamo quindi definire la matrice di iterazione B_J e il vettore $\boldsymbol{g}$ associati al metodo di Jacobi abbiamo

$$\boldsymbol{g} = D^{-1}\boldsymbol{b}. \quad \text{e} \quad B_J = D^{-1}(D - A) = I - D^{-1}A = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \dots & -\frac{a_{2n}}{a_{22}} \\ \dots & & & & \\ -\frac{a_{n1}}{a_{nn}} & & \dots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{bmatrix}$$

Data l'espressione di B_J posso scrivere la soluzione del problema componente per componente, ovvero per il passo k come

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

notiamo che l'elemento i-esimo del vettore $\boldsymbol{x}^{k+1}$ è calcolabile indipendentemente dalle altri componenti, quindi l'algoritmo è facilmente parallelizzabile ottenendo la soluzione del problema ancora più velocemente.

Essendo un metodo iterativo della forma $\boldsymbol{x}^{k+1} = B_J \boldsymbol{x}^k + \boldsymbol{g}$, allora il metodo di Jacobi converge se e solo se $\rho(B_J) < 1$. Esistono condizioni solo sufficienti per la convergenza che sono molto più facili da verificare, in particolare abbiamo che, se A è una matrice a dominanza diagonale stretta (per righe o per colonne), allora $|\lambda_i| < 1$ per ogni i e quindi il metodo converge.

Metodo di Gauss-Seidel

Scomponiamo la matrice A come $A = D - E - F$ dove

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & & 0 \\ 0 & a_{22} & 0 & \ldots & 0 \\ \ldots & & & & \\ 0 & & \ldots & 0 & a_{nn} \end{bmatrix} \quad E = \begin{bmatrix} 0 & 0 & \ldots & & 0 \\ -a_{21} & 0 & 0 & \ldots & 0 \\ \ldots & & & & \\ -a_{n1} & & \ldots & -a_{n,n-1} & 0 \end{bmatrix}$$

$$F = \begin{bmatrix} 0 & -a_{12} & \ldots & & -a_{1n} \\ 0 & 0 & -a_{23} & \ldots & -a_{2n} \\ \ldots & & & & \\ 0 & & \ldots & 0 & 0 \end{bmatrix}.$$

Il metodo di Gauss-Seidel pone la matrice di precondizionamento $P = D - E$, ossia la parte triangolare inferiore di A. Infatti, notiamo che $D - E = T$ dove $t_{ij} = a_{ij}$ se $j \leq i$, $t_{ij} = 0$ se $j > i$, e $T - A = F$. ottenere

$$P\boldsymbol{x}^{k+1} = (P - A)\boldsymbol{x}^k + \boldsymbol{b} \quad \rightarrow \quad (D - E)\boldsymbol{x}^{k+1} = F\boldsymbol{x}^k + \boldsymbol{b}.$$

La forma componente per componente del metodo è data da

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

Anche in questo caso, grazie alla struttura di P, ad ogni iterazione abbiamo un sistema semplice da risolvere, utilizzando una sostituzione in avanti. Dato che il calcolo di x_i^{k+1} richiede di aver già calcolato i valori di x_j^{k+1} per $j = 1, \ldots, i-1$ allora l'algoritmo di Gauss-Seidel è più difficile da parallelizzare rispetto al metodo di Jacobi. Tuttavia, il metodo di Gauss-Seidel spesso converge più velocemente rispetto al metodo di Jacobi, ossia raggiunge una certa soglia di tolleranza sull'errore con un numero minore di iterazioni.

La matrice di iterazione di Gauss-Seidel è data da

$$B_{GS} = (D - E)^{-1} F$$

(in alternativa, $B_{GS} = T^{-1}(T - A)$) e sappiamo che se $\rho(B_{GS}) < 1$ allora il metodo risulta convergente. Nel caso particolare in cui A è una matrice tridiagonale, ovvero se $a_{ij} = 0$ per $|i| > j + 1$, allora abbiamo che $\rho(B_{GS}) = [\rho(B_J)]^2$. Quindi il metodo di Jacobi converge se e solo se il metodo di Gauss-Seidel converge, inoltre se convergono allora quest'ultimo converge più velocemente. Come negli altri metodi studiati, verificare che $\rho(B_{GS}) < 1$ è solitamente molto costoso e quindi da evitare, ci affidiamo quindi a condizioni solo sufficienti per sapere se a priori il metodo di Gauss-Seidel converge o meno. Queste condizioni sono le seguenti:

- se la matrice A è a dominanza diagonale stretta per righe, o per colonne
- oppure se la matrice A è simmetrica e definita positiva.

Metodo di Richardson

Consideriamo la scomposizione $A = P - (P - A)$ usata precedentemente. Possiamo esprimere un'iterazione del generico metodo iterativo come

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

Il metodo di Richardson stazionario è definito generalizzando la precedente con un valore α costante come

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha P^{-1}\boldsymbol{r}^k,$$

mentre il metodo di Richarson dinamico o non-stazionario è definito dato un valore α^k variabile per ogni passo k come

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

Per i metodi di Richardson la matrice di iterazione risulta quindi

$$B = I - \alpha_k P^{-1}A.$$

Teorema 4.3

Supponiamo che P sia non singolare e che $P^{-1}A$ sia definita positiva, quindi con $\lambda_i > 0$. Ordiniamo gli autovalori di $P^{-1}A$ nel seguente modo

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0.$$

Allora il metodo di Richardson stazionario converge se e solo se

$$0 < \alpha < \frac{2}{\lambda_1}$$

inoltre esiste un valore ottimale di α detto α_{opt} di valore

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n}$$

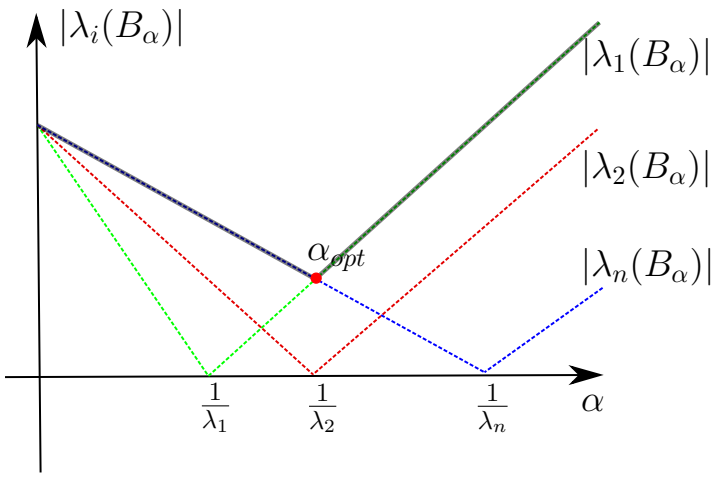
tale che il raggio spettrale della matrice di iterazione B è minimo, pari a

$$\rho_{opt} = \min_{\alpha} \rho(B_\alpha) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}.$$

Proof. Data l'espressione di B abbiamo che i suoi autovalori si possono esprimere come $\lambda_i(B_\alpha) = 1 - \alpha\lambda_i$. Sappiamo che un metodo iterativo è convergente quando $|\lambda_i(B_\alpha)| < 1$ per tutti gli $i = 1, \ldots, n$. Nel nostro caso abbiamo

$$|1 - \alpha\lambda_i| < 2 \quad \Rightarrow \quad 0 < \alpha < \frac{2}{\lambda_1}$$

Verifichiamo ora il valore di α_{opt}: come detto gli autovalori della matrice B_α sono dati da $\lambda_i(B_\alpha) = 1 - \alpha\lambda_i$ per $i = 1, \ldots, n$. Ricordando che $\rho(B_\alpha) = \max_{i=1,\ldots,n} |\lambda_i(B_\alpha)|$, allora consideriamo il seguente grafico che rappresenta ogni $|\lambda_i(B_\alpha)|$ al variare di α



Essendo il raggio spettrale il massimo tra tutti gli autovalori, allora per ogni α i suoi valori sono dati dall'inviluppo evidenziato in grigio nel grafico. Il valore ottimale di α, ovvero quello che rende minimo il raggio spettrale di B_α, è identificato dall'intersezione tra $1-\alpha\lambda_n = \alpha\lambda_1 - 1$. Otteniamo quindi il valore ottimale ricercato e il corrispondente raggio spettrale. □

Metodo del gradiente

Consideriamo una matrice $A \in \mathbb{R}^{n\times n}$ simmetrica e definita positiva, si vuole risolvere il problema di determinare $\boldsymbol{x} \in \mathbb{R}^n$ tale che

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

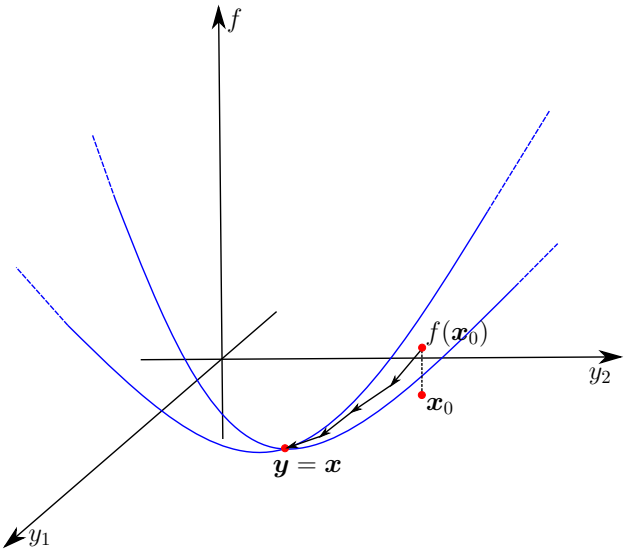
Notiamo che se A è simmetrica e definita positiva allora anche A^{-1} è simmetrica e definita positiva, infatti $\lambda(A) = 1/\lambda(A^{-1})$ ovvero

$$A\boldsymbol{w} = \lambda\boldsymbol{w} \quad \rightarrow \quad A^{-1}A\boldsymbol{w} = \lambda A^{-1}\boldsymbol{w} \quad \rightarrow \quad \lambda^{-1}\boldsymbol{w} = A^{-1}\boldsymbol{w}.$$

Costruiamo ora la funzione $f : \mathbb{R}^n \rightarrow \mathbb{R}$ quadratica associata alla risoluzione del sistema lineare data da

$$f(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^\top A\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{b},$$

dove $\boldsymbol{y} \in \mathbb{R}^n$ è un vettore. Graficamente essa può essere rappresentata come un paraboloide se $\boldsymbol{y} \in \mathbb{R}^2$



Vogliamo mostrare che la soluzione del sistema lineare equivale alla ricerca del punto di minimo della funzione f. Volendo calcolare il minimo di f dobbiamo risolvere il problema di determinare $\boldsymbol{y}$ tale che

$$\mathbf{0} = \nabla f(\boldsymbol{y}) = A\boldsymbol{y} - \boldsymbol{b}.$$

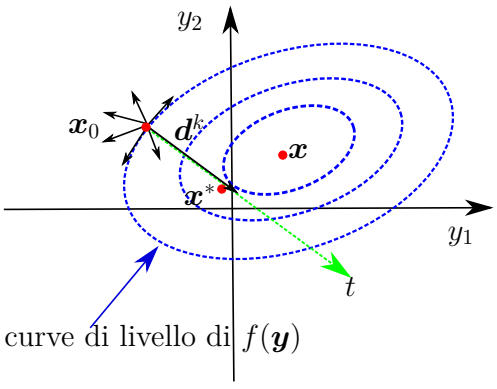
ovvero calcolare il minimo $\boldsymbol{y}$ di f è equivalente a determinare $\boldsymbol{x}$ che risolvere il sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$. Abbiamo quindi l'equivalenza $\boldsymbol{y} = \boldsymbol{x}$. Il metodo del gradiente costruisce una sequenza di soluzioni che converge verso il minimo di f.

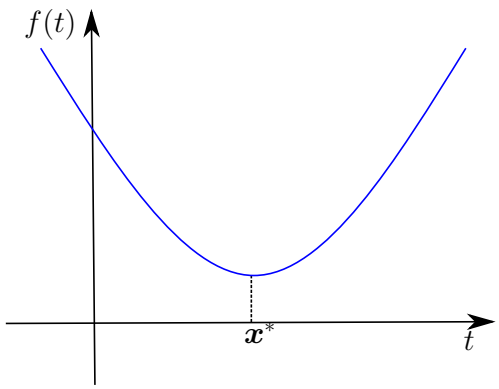
La funzione f può essere definita in modi diversi, ottenendo lo stesso punto di minimo ma un valore minimo differente (in altre parole, si può utilizzare un paraboloide traslato).

Il metodo del gradiente fa parte dei metodi di discesa, tali per cui partendo da un $\boldsymbol{x}^0$ ad ogni $k \geq 1$ prima si ricerca una direzione di discesa $\boldsymbol{d}^k$, successivamente viene scelto un passo α_k con cui avanzare, ed infine si aggiorna il valore $\boldsymbol{x}^{k+1}$

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k \boldsymbol{d}^k \quad \text{tale per cui} \quad f(\boldsymbol{x}^{k+1}) < f(\boldsymbol{x}^k).$$

Un metodo di discesa è caratterizzato principalmente dalla scelta della direzione $\boldsymbol{d}^k$ e dal passo α_k.

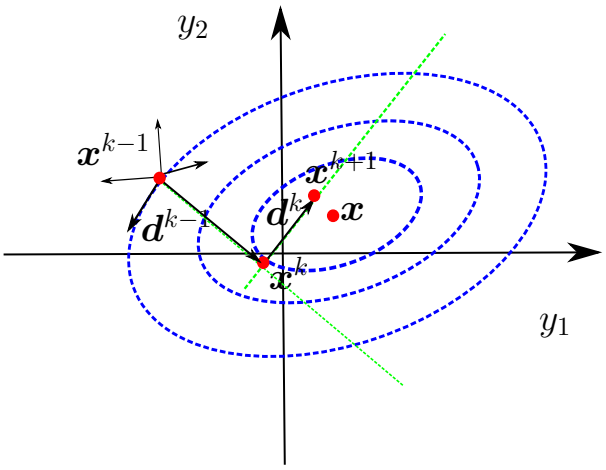




Data una direzione di discesa $\boldsymbol{d}^k$ allora possiamo calcolare il valore del passo ottimale, richiedendo che la funzione f sia minima lungo la direzione data da $\boldsymbol{d}^k$. Tale funzione rimane quadratica ma dipendente solo dall'ascissa curvilinea t, essa è definita come $f(\boldsymbol{x}^k + t\boldsymbol{d}^k)$. Il valore minimo si

realizza per un certo t pari al passo α_k che soddisfa

$$\alpha_k = \frac{(\boldsymbol{d}^k)^\top \boldsymbol{r}^k}{(\boldsymbol{d}^k)^\top A \boldsymbol{d}^k}.$$



Approfondimento 4.1

Dimostriamo che il valore di α_k corrisponde al minimo di $f(\boldsymbol{x}^k + t\boldsymbol{d}^k)$:

$$f(\boldsymbol{x}^k + t\boldsymbol{d}^k) = f(t) = \frac{1}{2}(\boldsymbol{x}^k + t\boldsymbol{d}^k)^\top A(\boldsymbol{x}^k + t\boldsymbol{d}^k) - (\boldsymbol{x}^k + t\boldsymbol{d}^k)^\top \boldsymbol{b}$$

$$\frac{df(t)}{dt} = (\boldsymbol{d}^k)^\top A(\boldsymbol{x}^k + t\boldsymbol{d}^k) - (\boldsymbol{d}^k)^\top \boldsymbol{b} = 0$$

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

Per quanto riguarda $\boldsymbol{d}^k$ una scelta che sembrerebbe ottimale è quella di massima discesa partendo dal punto $\boldsymbol{x}^k$, ovvero data da

$$\boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k) = -(A\boldsymbol{x}^k - \boldsymbol{b}) = \boldsymbol{b} - A\boldsymbol{x}^k = \boldsymbol{r}^k,$$

quindi se scelgo $\boldsymbol{d}^k = \boldsymbol{r}^k$, il metodo di discesa è detto metodo del gradiente. Inoltre possiamo calcolare il nuovo valore del residuo che è dato da

$$\boldsymbol{r}^{k+1} = \boldsymbol{b} - A\boldsymbol{x}^{k+1} = \boldsymbol{b} - A(\boldsymbol{x}^k + \alpha_k \boldsymbol{r}^k) = \boldsymbol{r}^k - \alpha_k A\boldsymbol{r}^k,$$

dove il termine $A\boldsymbol{r}^k$, che computazionalmente costa $\mathcal{O}(n^2)$, è già stato calcolato per determinare il valore di α_k. L'algoritmo è il seguente.

Algoritmo 4.1 - Metodo del gradiente

Data: Dato x^0, calcolo $\boldsymbol{r}^0 = \boldsymbol{b} - A\boldsymbol{x}^0$;
while *(criterio di convergenza)* **do**
  $\alpha_k \leftarrow \dfrac{(\boldsymbol{r}^k)^\top \boldsymbol{r}^k}{(\boldsymbol{r}^k)^\top A\boldsymbol{r}^k}$;
  $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha_k \boldsymbol{r}^k$;
  $\boldsymbol{r}^{k+1} \leftarrow \boldsymbol{r}^k - \alpha_k A\boldsymbol{r}^k$;
  $k \leftarrow k + 1$;
end

È possibile dimostrare che per il metodo del gradiente otteniamo che le direzione generate sono, a coppie, ortogonali ovvero

$$\boldsymbol{d}^k \cdot \boldsymbol{d}^{k+1} = 0 \quad \Longleftrightarrow \quad \boldsymbol{r}^k \cdot \boldsymbol{r}^{k+1} = 0 \quad k = 0, \dots,$$

tuttavia non è garantito che la direzione $k+1$-esima risulti ortogonale a tutte le direzioni da $k-1$ a 0. Infatti sostituendo l'espressione prima di $\boldsymbol{r}^{k+1}$ e successivamente di α_k otteniamo

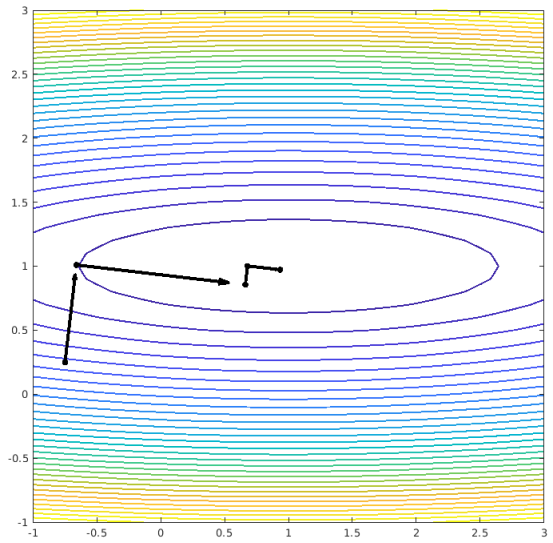
$$\boldsymbol{r}^k \cdot \boldsymbol{r}^{k+1} = \boldsymbol{r}^k \cdot (\boldsymbol{r}^k - \alpha_k A\boldsymbol{r}^k) = \boldsymbol{r}^k \cdot \boldsymbol{r}^k - \frac{(\boldsymbol{r}^k)^\top \boldsymbol{r}^k}{(\boldsymbol{r}^k)^\top A\boldsymbol{r}^k}(\boldsymbol{r}^k)^\top A\boldsymbol{r}^k = 0.$$

Esempio 4.2

Consideriamo il sistema $A\boldsymbol{x} = \boldsymbol{b}$ con

$$A = \begin{bmatrix} \frac{1}{20} & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} \frac{1}{20} \\ 1 \end{bmatrix}.$$

Ovviamente la soluzione esatta del sistema è $\boldsymbol{x} = [1, 1]^\top$. Supponiamo di scegliere come *guess* iniziale $\boldsymbol{x}^0 = [-\frac{3}{4}, \frac{1}{4}]^\top$ e osserviamo in figura le iterazioni del metodo del gradiente sovrapposte alle linee di livello della funzione f. Osserviamo che le direzioni di discesa sono perpendicolari alle linee di livello e, a due a due fra loro.



Teorema 4.4 - Convergenza del metodo del gradiente

Se $A \in \mathbb{R}^{n\times n}$ è simmetrica e definita positiva, allora il metodo del gradiente converge alla soluzione del sistema $A\boldsymbol{x} = \boldsymbol{b}$ per ogni dato iniziale $\boldsymbol{x}^0$. Inoltre, vale la seguente stima dell'errore $\boldsymbol{e}^k = \boldsymbol{x} - \boldsymbol{x}^k$ data da

$$\left\|\boldsymbol{e}^k\right\|_A \leq C^k \left\|\boldsymbol{e}^0\right\|_A \quad \text{con} \quad C = \frac{\text{cond}(A) - 1}{\text{cond}(A) + 1}$$

dove $\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^\top A\boldsymbol{x}}$ (detta norma dell'energia o $A-$norma) e C è detto fattore di abbattimento dell'errore.

Ricordiamo che, sotto le ipotesi del teorema, il numero di condizionamento della matrice A è dato da $\text{cond}(A) = \lambda_{\max}/\lambda_{\min}$. Inoltre, dato che il fattore di abbattimento dell'errore è minore strettamente di 1 allora abbiamo che

$$\lim_{k\to\infty} \left\|\boldsymbol{e}^k\right\|_A = 0.$$

Notiamo inoltre che C è costante e dipendente solamente dalla matrice A, risulta tanto più piccolo quanto il numero di condizionamento di A è vicino a 1, mentre C è tanto più grande e vicino a 1 quanto più $\text{cond}(A) \gg 1$. Possiamo quindi concludere che se una matrice ha condizionamento $\text{cond}(A) \approx 1$, la matrice A è detta ben condizionata e il metodo del gradiente converge velocemente. Al contrario se $\text{cond}(A) \gg 1$ allora A è detta matrice mal condizionata e il metodo del gradiente converge molto lentamente dato che $C \approx 1$. Infatti se consideriamo in cui $A \in \mathbb{R}^{2\times 2}$ allora l'eccentricità degli ellissi, curve di livello di f, è legata al condizionamento di A: più $\text{cond}(A)$ è alto e più l'ellisse risulta schiacciato e quindi le direzioni generate dal gradiente, essendo in questo specifico caso tutte ortogonali tra di loro, convergono molto più lentamente al minimo di f.

Metodo del gradiente coniugato

Il metodo del gradiente genera le direzioni $\boldsymbol{d}^k$ partendo dal gradiente della funzione, ci chiediamo se tale direzioni siano ottimali rispetto al problema considerato. Osserviamo che

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k \boldsymbol{d}^k = \boldsymbol{x}^{k-1}\alpha_{k-1}\boldsymbol{d}^{k-1} + \alpha_k \boldsymbol{d}^k = \ldots = \boldsymbol{x}^0 + \sum_{j=0}^{k} \alpha_j \boldsymbol{d}^j$$

e quindi il termine $\boldsymbol{x}^{k+1} - \boldsymbol{x}^0$ è una combinazione lineare delle k direzioni di discesa considerate $\boldsymbol{d}^j = \boldsymbol{r}^j$. Per $k = n$ tali vettori non sono, in generale, linearmente indipendenti ma anche se lo fossero non è detto che i coefficienti α_j trovati corrispondano allo sviluppo di $\boldsymbol{x} - \boldsymbol{x}^0$ su tale base. Quindi potrebbero servire molte più iterazioni di n per poter ottenere tale richiesta.

Il metodo del gradiente coniugato garantisce la scelta ottimale delle direzioni di discesa $\boldsymbol{d}^k$, tali da essere A-ortogonali o A-coniugate tra loro. Tale condizione permette di avere le direzioni di discesa tutte linearmente indipendenti tra loro essendo A non singolare.

Definizione 4.3 - A-ortogonalità

Data una matrice $A \in \mathbb{R}^{n\times n}$ simmetrica, due vettori $\boldsymbol{x}$ e $\boldsymbol{y} \in \mathbb{R}^n$ sono detti A-ortogonali se

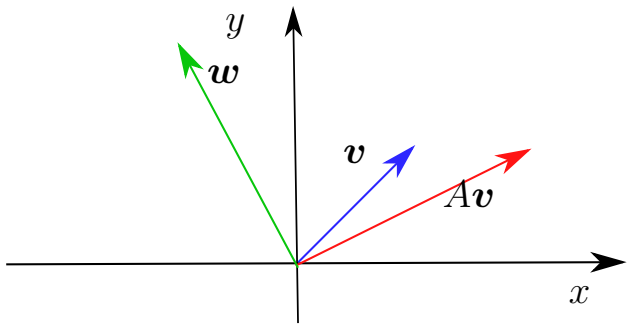
$$(A\boldsymbol{y})^\top \boldsymbol{x} = 0 \quad \Longleftrightarrow \quad \boldsymbol{y}^\top A\boldsymbol{x} = 0.$$

Esempio 4.3

Consideriamo la matrice

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix},$$

simmetrica definita positiva, con autovalori $\lambda_1 = 2$, $\lambda_2 = 1$ e autovettori paralleli a $\mathbf{i}$ e $\mathbf{j}$. Consideriamo il vettore $\mathbf{v} = [1, 1]^\top$. Abbiamo che $A\mathbf{v} = [2, 1]^\top$. Un vettore $\mathbf{y}$ A-ortogonale a $\mathbf{v}$ è tale che $2y_1 + y_2 = 0$, per esempio $\mathbf{w} = [-1, 2]^\top$. Notiamo che $\mathbf{v} \cdot \mathbf{w} \neq 0$. Osserviamo inoltre che se consideriamo un vettore $\mathbf{v} = [2, 0]^\top$, parallelo agli autovettori di A, allora $A\mathbf{v}$ risulta parallelo a $\mathbf{v}$ e in questo caso il vettore A-ortogonale è anche semplicemente ortogonale.



Abbiamo che il metodo del gradiente coniugato, in aritmetica esatta, converge alla soluzione esatta in al più n iterazioni. Quindi, seppure il metodo del gradiente coniugato è stato derivato come metodo iterativo esso convergerà in un numero finito di passi. Nel caso in cui n sia grande, lo schema viene comunque bloccato prima di fare n passi ottenendo quindi una soluzione approssimata.

Algoritmo 4.2 - Metodo del gradiente coniugato

Data: Dato x^0, calcolo $\boldsymbol{r}^0 = \boldsymbol{b} - A\boldsymbol{x}^0$ e impongo $\boldsymbol{d}^0 = \boldsymbol{r}^0$;
while *(criterio di convergenza)* **do**
$\alpha_k \leftarrow \dfrac{(\boldsymbol{d}^k)^\top \boldsymbol{r}^k}{(\boldsymbol{d}^k)^\top A\boldsymbol{d}^k}$;
$\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha_k \boldsymbol{d}^k$;
$\boldsymbol{r}^{k+1} \leftarrow \boldsymbol{r}^k - \alpha_k A\boldsymbol{d}^k$;
$\beta_k \leftarrow \dfrac{(A\boldsymbol{d}^k)^\top \boldsymbol{r}^{k+1}}{(A\boldsymbol{d}^k)^\top \boldsymbol{d}^k}$;
$\boldsymbol{d}^{k+1} \leftarrow \boldsymbol{r}^{k+1} - \beta_k \boldsymbol{d}^k$;
$k \leftarrow k + 1$;
end

Notiamo che la direzione di discesa coincide con il residuo solo alla prima iterazione, successivamente viene "corretta" sottraendo le direzioni precedentemente utilizzate. Le direzioni $\{\boldsymbol{d}^{k+1}\}$ calcolate sono quindi A-ortogonali tra loro.

Teorema 4.5 - Convergenza del metodo del gradiente coniugato

Sia $A \in \mathbb{R}^{n\times n}$ una matrice simmetrica e definita positiva, allora il metodo del gradiente coniugato converge alla soluzione del sistema $A\boldsymbol{x} = \boldsymbol{b}$, per ogni $\boldsymbol{x}^0 \in \mathbb{R}^n$ in al più n iterazioni. Inoltre, vale la stima

$$\left\|\boldsymbol{e}^k\right\|_A \leq \frac{2c^k}{1+c^{2k}} \left\|\boldsymbol{e}^0\right\|_A \quad \text{dove} \quad c = \frac{\sqrt{\text{cond}(A)}-1}{\sqrt{\text{cond}(A)}+1}.$$

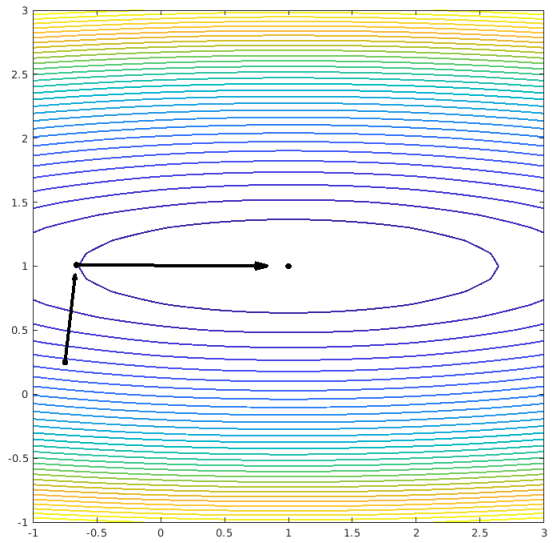
Abbiamo anche in questo caso che il fattore di abbattimento dell'errore converge a zero e lo fa più velocemente che nel caso del metodo del gradiente.

Esempio 4.4

Consideriamo il sistema dell'esempio precedente, $A\boldsymbol{x} = \boldsymbol{b}$ con

$$A = \begin{bmatrix} \frac{1}{20} & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} \frac{1}{20} \\ 1 \end{bmatrix}.$$

Ovviamente la soluzione esatta del sistema è $\boldsymbol{x} = [1,1]^\top$. Supponiamo di scegliere come *guess* iniziale $\boldsymbol{x}^0 = [-\frac{3}{4}, \frac{1}{4}]^\top$ e osserviamo in figura le iterazioni del metodo del gradiente coniugato sovrapposte alle linee di livello della funzione f. Osserviamo che le direzioni di discesa sono tali per cui la soluzione viene ottenuta in sole 2 iterazioni.



Sia il metodo del gradiente che il metodo del gradiente coniugato possono essere migliorati usando un precondizionatore, cioè trovando una matrice P invertibile, detto precondizionatore, tale che

$$\text{cond}(P^{-1}A) \ll \text{cond}(A).$$

Approfondimento 4.4

Esistono diverse scelte possibili per la matrice P. Lo scopo è trovare P tale che il condizionamento di $P^{-1}A$ sia il più piccolo possibile, e in questo senso la scelta più efficace sarebbe $P = A$: infatti $A^{-1}A = I$ ha condizionamento unitario. Tuttavia, la matrice A non è in generale facile da invertire, e, se lo fosse, non avremmo bisogno di ricorrere a un metodo iterativo per risolvere il sistema. Un buon precondizionatore P sarà quindi i) facile da invertire ii) "simile" alla matrice A. Consideriamo un esempio:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 100 \end{bmatrix}$$

Il condizionamento di A è $\text{cond}(A) = 50.26$ dato da $\frac{\lambda_{max}}{\lambda_{min}} = \frac{100.0102}{1.9898}$. Proviamo a utilizzare come P la diagonale di A: una matrice diagonale è infatti facile da invertire. Otteniamo

$$P^{-1}A = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{100} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 100 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ \frac{1}{100} & 1 \end{bmatrix}$$

e il condizionamento di $P^{-1}A$ è ridotto a $\text{cond}(P^{-1}A) = 2.63$.

Vogliamo quindi risolvere il seguente problema precondizionato utilizzando il metodo del gradiente o del gradiente coniugato seguente

$$A\boldsymbol{x} = \boldsymbol{b} \quad \xrightarrow{\text{moltiplico per } P^{-1}} \quad P^{-1}A\boldsymbol{x} = P^{-1}\boldsymbol{b}.$$

Avendo quindi il fattore di abbattimento dell'errore molto minore del sistema di partenza, possiamo ottenere la soluzione con meno passi e quindi in meno tempo.

I metodi di discesa generano una successione $\{\boldsymbol{x}^k\}$ tale che converge alla soluzione esatta $\boldsymbol{x}$ per $k \to \infty$. Chiaramente dobbiamo terminare l'algoritmo utilizzando un opportuno criterio d'arresto. Una possibilià è un criterio basato sul residuo $\boldsymbol{r}^k$, data una tolleranza ϵ abbiamo

$$\frac{\left\|\boldsymbol{r}^k\right\|}{\|\boldsymbol{b}\|} < \epsilon.$$

Tale criterio è efficace se il numero di condizionamento di A non è troppo elevato, infatti abbiamo che

$$\frac{\left\|\boldsymbol{e}^k\right\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\frac{\left\|\boldsymbol{r}^k\right\|}{\|b\|}.$$

Un'altra possibilità è basare il criterio d'arresto sull'incremento della soluzione calcolata, ovvero

$$\frac{\left\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\right\|}{\|\boldsymbol{b}\|} < \epsilon.$$