CAPITOLO 3

Metodi iterativi stazionari e non stazionari per la soluzione di sistemi lineari

Un metodo iterativo per risolvere un sistema lineare $A\mathbf{x} = \mathbf{b}$ è un algoritmo che prende il via da una qualsiasi approssimazione iniziale $\mathbf{x}^{(0)}$ della soluzione $\mathbf{x}$ e la modifica successivamente - generando $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... - nel tentativo di migliorare l'effettiva approssimazione (secondo un certo criterio) di $\mathbf{x}$.

Premettiamo alla trattazione dei metodi iterativi per la soluzione di sistemi lineari alcune proprietà della norma dell'energia. Sia $A \in \mathbb{C}^{n\times n}$ una matrice hermitiana definita positiva. Denotiamo con $A^{\frac{1}{2}}$ l'unica matrice hermitiana definita positiva soluzione dell'equazione matriciale $X^2 = A$. Ricordiamo che la norma vettoriale in $\mathbb{C}^n$ data da

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^H A\mathbf{x}} = \|A^{\frac{1}{2}}\mathbf{x}\|_2,$$

viene detta *norma dell'energia* (o A-norma) del vettore $\mathbf{x}$ ed è la norma indotta dal prodotto scalare dato da $(\mathbf{x},\mathbf{y})_A = (A\mathbf{x},\mathbf{y}) = (\mathbf{x},A\mathbf{y}) = \mathbf{y}^H A\mathbf{x}$, detto A-prodotto scalare. Per la norma operatore indotta dalla norma dell'energia vale il seguente risultato.

PROPOSIZIONE 3.1. *Sia $M \in \mathbb{C}^{n\times n}$, si ha*

$$\|M\|_A = \|A^{\frac{1}{2}} M A^{-\frac{1}{2}}\|_2. \tag{3.1}$$

DIMOSTRAZIONE. Vale

$$\|M\|_A = \max_{\|\mathbf{x}\|_A=1} \|M\mathbf{x}\|_A = \max_{\|A^{\frac{1}{2}}\mathbf{x}\|_2=1} \|A^{\frac{1}{2}} M\mathbf{x}\|_2.$$

Posto $\mathbf{y} = A^{\frac{1}{2}}\mathbf{x}$, si ottiene la tesi, avendo

$$\|M\|_A = \max_{\|\mathbf{y}\|_2=1} \|A^{\frac{1}{2}} M A^{-\frac{1}{2}}\mathbf{y}\|_2 = \|A^{\frac{1}{2}} M A^{-\frac{1}{2}}\|_2.$$

□

Ovviamente la norma spettrale e la norma dell'energia di A coincidono, e sono date, per la Proposizione 2.10, dal suo autovalore più grande. Illustriamo alcune proprietà della norma dell'energia.

PROPOSIZIONE 3.2. *Sia $M \in \mathbb{C}^{n\times n}$ tale che AM è hermitiana, si ha*

$$\|M\|_A = \rho(M).$$

Inoltre, se M è normale, allora $\|M\|_A = \|M\|_2$.

DIMOSTRAZIONE. Da (3.1) si ha $\|M\|_A = \|A^{\frac{1}{2}} M A^{-\frac{1}{2}}\|_2 = \|A^{-\frac{1}{2}} AM A^{-\frac{1}{2}}\|_2$. Dato che per ipotesi AM è hermitiana, lo è anche $A^{-\frac{1}{2}} AM A^{-\frac{1}{2}}$, quindi, per la Proposizione 2.10, $\|A^{-\frac{1}{2}} AM A^{-\frac{1}{2}}\|_2 = \rho(A^{-\frac{1}{2}} AM A^{-\frac{1}{2}}) = \rho(A^{\frac{1}{2}} M A^{-\frac{1}{2}})$.
Inoltre, dato che matrici simili hanno medesimo spettro, si ha $\rho(A^{\frac{1}{2}} M A^{-\frac{1}{2}}) = \rho(M)$.
Per quanto riguarda la seconda parte dell'asserto, se inoltre la matrice M è normale, ancora per la Proposizione 2.10 vale $\|M\|_2 = \rho(M)$, quindi $\|M\|_A = \|M\|_2$. □

COROLLARIO 3.3. *Se $M \in \mathbb{C}^{n\times n}$ ha autovalori reali e positivi, che assumiamo ordinati in modo che $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ e se AM è hermitiana, si ha*

$$\|M\|_A = \lambda_1.$$

Inoltre e AM^{-1} è hermitiana, si ha $\kappa_A(M) = \frac{\lambda_1}{\lambda_n}$.

DIMOSTRAZIONE. La tesi si ottiene direttamente dalla prima parte della Proposizione 3.2 e dall'osservazione che $\rho(M) = |\lambda_1| = \lambda_1$ e $\rho(M^{-1}) = \frac{1}{|\lambda_n|} = \frac{1}{\lambda_n}$. □

OSSERVAZIONE 3.1. *Per $M \in \mathbb{C}^{n\times n}$ invertibile tale che $AM = M^H A$ e $AM^{-1} = M^{-H} A$ vale, per la Proposizione 3.2, $\|M\|_A \leq \|M\|_2$ e $\|M^{-1}\|_A \leq \|M^{-1}\|_2$ (dato che, come è noto, nessuna norma operatore di una matrice può essere minore del suo raggio spettrale). Dunque si ha $\|M\|_A \|M^{-1}\|_A = \kappa_A(M) \leq \kappa_2(M) = \|M\|_2 \|M^{-1}\|_2$.*

3.1. Metodo di Richardson (stazionario)

Dato il sistema lineare $A\mathbf{x} = \mathbf{b}$ e una matrice non singolare P, il metodo iterativo

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha P^{-1} \mathbf{r}^{(k)},$$

con $\alpha \in \mathbb{R} \neq 0$, per la soluzione del suddetto sistema, è detto *metodo di Richardson*, o metodo di Richardson stazionario. Ricordiamo che $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$ è il residuo al passo k-esimo e notiamo che la matrice di iterazione B_α associata, tale cioè che $\mathbf{x}^{(k+1)} = B_\alpha \mathbf{x}^{(k)} + \alpha P^{-1}\mathbf{b}$, è

$$B_\alpha = I_n - \alpha P^{-1} A.$$

È utile definire il residuo precondizionato $\mathbf{z}^{(k)} = P^{-1}\mathbf{r}^{(k)}$ e osservare esplicitamente che $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha \mathbf{z}^{(k)}$ e che $\mathbf{r}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k+1)} = \mathbf{r}^{(k)} - \alpha A\mathbf{z}^{(k)}$.

Il metodo può essere schematizzato come segue

per $k = 0, 1, \ldots$
  si risolve il sistema lineare $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$
  si aggiorna la soluzione $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha \mathbf{z}^{(k)}$
  si aggiorna il residuo $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha A\mathbf{z}^{(k)}$
fine

I seguenti teoremi riguardano l'analisi della convergenza del metodo di Richardson.

TEOREMA 3.4. *Sia* $\det(P) \neq 0$*, il metodo di Richardson converge se e solo se*

$$\frac{2\Re(\lambda_i)}{\alpha|\lambda_i|^2} > 1, \quad \forall i = 1,\dots,n,$$

dove $\lambda_i \in \sigma(P^{-1}A)$*, con* $i = 1,\dots,n$*. (Notiamo che la tesi implica che la parte reale di tutti gli autovalori di* $P^{-1}A$ *deve avere lo stesso segno e che tale segno deve essere necessariamente lo stesso di* α*).*

DIMOSTRAZIONE. La ben nota condizione necessaria e sufficiente di convergenza di un metodo iterativo (per ogni scelta del vettore iniziale), ovvero $\rho(B_\alpha) < 1$, è equivalente a $|1-\alpha\lambda_i|^2 < 1$, ovvero $\Re(1-\alpha\lambda_i)^2 + \Im(1-\alpha\lambda_i)^2 = (1-\alpha\Re(\lambda_i))^2 + \alpha^2\Im(\lambda_i)^2 = 1 + \alpha^2\Re(\lambda_i)^2 - 2\alpha\Re(\lambda_i) + \alpha^2\Im(\lambda_i)^2 < 1$, da cui segue $2\alpha\Re(\lambda_i) > \alpha^2|\lambda_i|^2$, per ogni $i = 1,\dots,n$. Dividendo per $\alpha^2|\lambda_i|^2$ si ha la tesi. □

TEOREMA 3.5. *Sia* $\det(P) \neq 0$*. Se* $P^{-1}A$ *ha autovalori reali positivi, che possiamo assumere ordinati in modo che* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$*, allora si ha che*

i) il metodo di Richardson converge se e solo se $0 < \alpha < \dfrac{2}{\lambda_1}$; (3.2)

ii) $\min\limits_{\alpha\in\left(0,\frac{2}{\lambda_1}\right)} \rho(B_\alpha) = \rho(B_{\alpha^*}) = \dfrac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}, \qquad con\ \alpha^* = \dfrac{2}{\lambda_1+\lambda_n}.$ (3.3)

DIMOSTRAZIONE. Per quanto riguarda il primo punto, l'ipotesi sugli autovalori di $P^{-1}A$ implica che α sia positivo e che la condizione di convergenza del teorema precedente sia equivalente a $\alpha < \frac{2}{\lambda_i}$ per $i = 1:n$, e quindi alla (3.2). Per la seconda parte dell'asserto, si può osservare che che per la funzione $\rho(B_\alpha) = \max\{|1-\alpha\lambda_1|,\dots,|1-\alpha\lambda_n|\}$, nell'intervallo $0 < \alpha < \frac{2}{\lambda_1}$ vale

$$\rho(B_\alpha) = \begin{cases} 1-\alpha\lambda_n & \text{se } 0 < \alpha \leq \frac{2}{\lambda_1+\lambda_n} \\ \alpha\lambda_1 - 1 & \text{se } \frac{2}{\lambda_1+\lambda_n} \leq \alpha < \frac{2}{\lambda_1} \end{cases}$$

e che il valore minimo di $\rho(B_\alpha)$ viene assunto quando $\alpha\lambda_1 - 1 = 1 - \alpha\lambda_n$, ovvero quando $\alpha = \frac{2}{\lambda_1+\lambda_n}$, ovvero il valore ottimale α^* in (3.3).
Si ha inoltre

$$\rho(B_{\alpha^*}) = \alpha^*\lambda_1 - 1 = 1 - \alpha^*\lambda_n = 1 - \frac{2}{\lambda_1+\lambda_n}\lambda_n = \frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}.$$

□

Assumiamo ora che la matrice dei coefficienti del sistema lineare sia hermitiana definita positiva e scegliamo come P una matrice hermitiana definita positiva. La seguente osservazione ci permette di affermare che gli autovalori di $P^{-1}A$ sono reali e positivi.

OSSERVAZIONE 3.2. *Nell'ipotesi che sia A che P sono matrici hermitiane definite positive, allora si ha che anche $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ è hermitiana definita positiva. Osserviamo infatti che vale $(P^{-\frac{1}{2}}AP^{-\frac{1}{2}})^H = P^{-\frac{H}{2}}A^HP^{-\frac{H}{2}} = P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ e che per ogni $\mathbf{x}\neq 0$, $\mathbf{x}^HP^{-\frac{1}{2}}AP^{-\frac{1}{2}}\mathbf{x} = (P^{-\frac{1}{2}}\mathbf{x})^HA(P^{-\frac{1}{2}}\mathbf{x}) > 0$.*
Osserviamo inoltre che, dato che $P^{\frac{1}{2}}(P^{-1}A)P^{-\frac{1}{2}} = P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$, allora $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ e $P^{-1}A$ sono simili, quindi hanno gli stessi autovalori (reali e positivi in quanto autovalori di una matrice hermitiana definita positiva).

I seguenti risultati relativi alla monotonia della convergenza del metodo di Richardson coinvolgono sia la A-norma che il condizionamento in norma spettrale.

TEOREMA 3.6. *Siano A e P hermitiane definite positive. Se $0 < \alpha < \frac{2}{\lambda_1}$, dove λ_1 è l'autovalore più grande di $P^{-1}A$, il metodo di Richardson converge monotonamente rispetto alla A-norma, avendo*

$$\|\mathbf{e}^{(k+1)}\|_A \leq \rho(B_\alpha)\|\mathbf{e}^{(k)}\|_A, \tag{3.4}$$

con $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$.

DIMOSTRAZIONE. Grazie all'osservazione precedente, la convergenza è data dal Teorema 3.5. Inoltre, dato che P è una matrice hermitiana, la matrice $AB_\alpha = A - \alpha AP^{-1}A$ è anch'essa hermitiana e, per la Proposizione 3.2, si ha $\|B_\alpha\|_A = \rho(B_\alpha)$. Inoltre, per la consistenza del metodo, si ha

$$\|\mathbf{e}^{(k+1)}\|_A = \|\mathbf{x} - \mathbf{x}^{(k+1)}\|_A = \|B_\alpha\mathbf{x} + P^{-1}\mathbf{b} - (B_\alpha\mathbf{x}^{(k)} + P^{-1}\mathbf{b})\|_A = \|B_\alpha\mathbf{e}^{(k)}\|_A$$

e vale dunque

$$\|\mathbf{e}^{(k+1)}\|_A = \|B_\alpha\mathbf{e}^{(k)}\|_A \leq \|B_\alpha\|_A\|\mathbf{e}^{(k)}\|_A = \rho(B_\alpha)\|\mathbf{e}^{(k)}\|_A.$$

□

TEOREMA 3.7. *Siano A, P e $P^{-1}A$ hermitiane definite positive. L'iterazione di Richardson con parametro ottimale*

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \frac{2}{\lambda_1 + \lambda_n}P^{-1}\mathbf{r}^{(k)}, \tag{3.5}$$

dove $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ sono gli autovalori di $P^{-1}A$, converge monotonamente rispetto alla norma $\|\cdot\|_A$, e vale

$$\|\mathbf{e}^{(k+1)}\|_A \leq \frac{\kappa_2(P^{-1}A) - 1}{\kappa_2(P^{-1}A) + 1}\|\mathbf{e}^{(k)}\|_A. \tag{3.6}$$

DIMOSTRAZIONE. Il Teorema 3.5 assicura che per il raggio spettrale della matrice di iterazione ottimale vale

$$\rho(B_{\frac{2}{\lambda_1+\lambda_n}}) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{\frac{\lambda_1}{\lambda_n} - 1}{\frac{\lambda_1}{\lambda_n} + 1} = \frac{\kappa_2(P^{-1}A) - 1}{\kappa_2(P^{-1}A) + 1}, \tag{3.7}$$

dove solo nell'ultimo passaggio abbiamo sfruttato che $P^{-1}A$ hermitiana è definita positiva. Infatti, per il Corollario 2.11 e per l'ordinamento degli autovalori di $P^{-1}A$, vale $\kappa_2(P^{-1}A) = \frac{\lambda_1}{\lambda_n}$. La tesi segue dal Teorema 3.6. □

In effetti, l'ipotesi che la matrice $P^{-1}A$ sia hermitiana definita positiva può essere eliminata. Vale infatti il seguente teorema.

TEOREMA 3.8. *Siano A e P hermitiane definite positive. Denotati $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ gli autovalori reali e positivi di $P^{-1}A$, per l'errore dell'iterazione di Richardson con parametro ottimale (3.5) vale la (3.6).*

DIMOSTRAZIONE. Per l'Osservazione 3.2 la matrice $P^{-1}A$ ha autovalori reali e positivi. Per il Corollario 3.3, dato che la matrice $P^{-1}A$ è tale che $AP^{-1}A$ è hermitiana e la matrice $(P^{-1}A)^{-1} = A^{-1}P$ è tale che $AA^{-1}P = P$ è hermitiana, vale $\kappa_A(P^{-1}A) = \frac{\lambda_1}{\lambda_n}$. Quindi si ha

$$\rho(B_{\frac{2}{\lambda_1+\lambda_n}}) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{\frac{\lambda_1}{\lambda_n} - 1}{\frac{\lambda_1}{\lambda_n} + 1} = \frac{\kappa_A(P^{-1}A) - 1}{\kappa_A(P^{-1}A) + 1},$$

Per ottenere la disuguaglianza in (3.6) - che ha maggior senso perché riguarda il numero di condizionamento in norma spettrale ovvero, come abbiamo visto nel capitolo precedente, senza dubbio il numero di condizionamento computazionalmente più ragionevole - basta notare che, per l'Osservazione 3.1 e dato che $(x-1)/(x+1)$ è una funzione crescente di x, si ha $\frac{\kappa_A(P^{-1}A)-1}{\kappa_A(P^{-1}A)+1} \leq \frac{\kappa_2(P^{-1}A)-1}{\kappa_2(P^{-1}A)+1}$.

□

Dalla presenza del numero di condizionamento di $P^{-1}A$ nella formula dell'errore in (3.6), si comprende quanto sia importante la scelta della matrice P, generalmente definita *matrice di precondizonamento* o *precondizionatore*, che dovrà essere sia di agevole invertibilità che tale da garantire una maggiore velocità di convergenza del metodo. Di fatto, se si scegliesse di non precondizionare il sistema, i Teoremi 3.6 e 3.8 condurrebbero immediatamente al seguente risultato.

COROLLARIO 3.9. *Sia A hermitiana definita positiva. Se $0 < \alpha < \frac{2}{\lambda_1}$, dove λ_1 è l'autovalore più grande di A, si ha convergenza monotona dell'iterazione di Richardson e*

$$\|\mathbf{e}^{(k+1)}\|_A \leq \rho(I_n - \alpha A)\|\mathbf{e}^{(k)}\|_A.$$

Inoltre, se $\alpha = \frac{2}{\lambda_1+\lambda_n}$, dove λ_n è l'autovalore più piccolo di A, per l'errore dell'iterazione di Richardson con parametro ottimale

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \frac{2}{\lambda_1 + \lambda_n}\mathbf{r}^{(k)}, \tag{3.8}$$

si ha

$$\|\mathbf{e}^{(k+1)}\|_A \leq \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\|\mathbf{e}^{(k)}\|_A. \tag{3.9}$$

La maggiorazione in (3.9), per quanto osservato alla fine della dimostrazione del Teorema 3.8, è meno stringente della maggiorazione in (3.6) ove il precondizionatore P sia scelto in modo tale che $\kappa_2(P^{-1}A) < \kappa_2(A)$.

3.2. Tecniche di precondizionamento e fattorizzazioni incomplete

Se P_S è una matrice non singolare, il sistema $A\mathbf{x} = \mathbf{b}$ può essere equivalentemente riformulato come $P_S^{-1}A\mathbf{x} = P_S^{-1}\mathbf{b}$. Questo precondizionamento a sinistra è perfettamente analogo a quello operato nel metodo di Richardson. Si può alternativamente precondizionare il sistema $A\mathbf{x} = \mathbf{b}$ a destra, con una matrice P_D non singolare, ottenendo il sistema $AP_D^{-1}\mathbf{y} = \mathbf{b}$, con medesima soluzione $\mathbf{x} = P_D^{-1}\mathbf{y}$. Naturalmente si può quindi pensare di precondizionare da ambo le parti e riscrivere il sistema come $P_S^{-1}AP_D^{-1}\mathbf{y} = P_S^{-1}\mathbf{b}$ con medesima soluzione $\mathbf{x} = P_D^{-1}\mathbf{y}$.

Assumiamo che la matrice dei coefficienti del sistema lineare sia hermitiana definita positiva e che il metodo iterativo che si vuole usare richieda necessariamente questà proprietà. Il sistema precondizionato dovrà quindi avere matrice dei coefficienti hermitiana definita positiva.
Anche se la matrice di precondizionamento P_S [o P_D] è scelta hermitiana definita positiva, sappiamo che $P_S^{-1}A$ [o AP_D^{-1}] in generale non è neanche hermitiana. Al contrario, se - scelta la matrice hermitiana definita positiva P - si definiscono le matrici di precondizionamento $P_S = P_D = P^{\frac{1}{2}}$, ovvero si segue la terza strategia menzionata, si ha che $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ è ancora una matrice hermitiana definita positiva (si veda l'Osservazione 3.2).
Come si deve scegliere la matrice hermitiana definita positiva P?
Euristicamente, un precondizionatore funziona meglio quando lo spettro della matrice $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ è *clusterizzato*. In tal caso, denotati $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ gli autovalori di $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$, per il Corollario 2.11 si avrebbe $\kappa_2(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) = \frac{\lambda_1}{\lambda_n} \approx 1$, e quindi un condizionamento favorevole della nuova matrice dei coefficienti $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$. D'altro canto, sempre dall'Osservazione 3.2, si ha che $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ e $P^{-1}A$ sono simili, quindi si cerca P hermitiana definita positiva tale che $P^{-1}A$ abbia spettro clusterizzato in modo da avere, per quanto osservato nella dimostrazione del Teorema 3.8, $\kappa_A(P^{-1}A) = \frac{\lambda_1}{\lambda_n} \approx 1$.
Alternativamente, con minore costo computazionale, se della matrice hermitiana definita positiva P è nota la decomposizione di Cholesky $P = LL^H$, si possono scegliere le matrici di precondizionamento $P_S = L$ e $P_D = L^H$, in modo che la matrice $L^{-1}AL^{-H}$ è hermitiana definita positiva. Vale infatti $(L^{-1}AL^{-H})^H = L^{-1}A^HL^{-H} = L^{-1}AL^{-H}$ e, per ogni $\mathbf{x} \neq 0$, $\mathbf{x}^HL^{-1}AL^{-H}\mathbf{x} = (L^{-H}\mathbf{x})^HA(L^{-H}\mathbf{x}) > 0$. Anche in questo caso gli autovalori $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ di $L^{-1}AL^{-H}$ (che coincidono con gli autovalori di $P^{-1}A$ dato che $L^HP^{-1}AL^{-H} = L^H(LL^H)^{-1}AL^{-H} = L^HL^{-H}L^{-1}AL^{-H} = L^{-1}AL^{-H}$) dovranno essere contenuti in un intervallo sufficientemente piccolo dell'asse reale positivo, in modo da avere un efficiente precondizionamento, ovvero $\kappa_A(P^{-1}A) = \frac{\lambda_1}{\lambda_n} \approx 1$.
Infine, si può alternativamente partire da una matrice triangolare inferiore L e

definire il precondizionatore $P = LL^H$, matrice hermitiana definita positiva per costruzione, in modo che gli autovalori di $P^{-1}A$ (che coincidono con quelli della matrice hermitiana definita positiva $L^{-1}AL^{-H}$ per quanto appena osservato) siano il più possibile clusterizzati. Come vedremo, questa è la strategia seguita nel precondizionamento tramite fattorizzazione di Cholesky incompleta (Paragrafo 3.2.2), tecnica molto efficiente per clusterizzare il più possibile (ad 1) lo spettro e ottenere un condizionamento migliore del nuovo sistema lineare: in effetti, la matrice $L = L_{\text{inc}}$ dà luogo ad una matrice di precondizionamento $P = L_{\text{inc}}L_{\text{inc}}^H$ la cui inversa risulta *quasi*-inversa della matrice A (sarebbe una vera e propria inversione se A avesse tutti elementi non nulli).

Per una generica matrice dei coefficienti A non singolare - e quindi per un metodo iterativo che non richieda ulteriori proprietà - si usa di preferenza la prima [o la seconda] tecnica di precondizionamento e si sceglie P_S [o P_D] in modo che $P_S^{-1}A$ [o AP_D^{-1}] sia *quasi*-normale e abbia autovalori contenuti in una regione sufficientemente piccola del piano complesso. Infatti, il Corollario 2.11 assicura che se $P_S^{-1}A$ [o AP_D^{-1}] fosse normale, allora il numero di condizionamento $\kappa_2(P_S^{-1}A)$ [o $\kappa_2(AP_D^{-1})$] - dato quindi dal rapporto tra il modulo dell'autovalore dominante e il modulo dell'autovalore dominato - risulterebbe prossimo ad 1. In caso di quasi-normalità, si spera quindi in un condizionamento quasi analogamente favorevole, in caso di clusterizzazione dello spettro. È in questo contesto che si fa uso della la fattorizzazione LU incompleta (Paragrafo 3.2.1), la cui inversa risulta *quasi*-inversa della matrice A e che in genere produce una discreta clusterizzazione (ad 1) dello spettro.

In ogni caso una caratteristica essenziale di un precondizionatore P è senza dubbio che sia contenuto il costo, anche in termini di memoria, necessario per risolvere il sistema $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ (ovvero per avere il residuo precondizionato $P^{-1}\mathbf{r}^{(k)}$).

3.2.1. Fattorizzazione LU incompleta. L'idea di base delle fattorizzazioni incomplete è quella di preservare sparsità e struttura (*pattern*) della matrice dei coefficienti $A \in \mathbb{C}^{n\times n}$ del sistema lineare assegnato. Assumiamo che per tale matrice esista unica la fattorizzazione LU - senza necessità cioè di ricorrere alla strategia del pivoting parziale che ne potrebbe alterare la struttura (per esempio tridiagonale, pentadiagonale, ecc.). Come è noto, il processo di fattorizzazione LU potrebbe generare un gran numero di elementi non nulli in corrispondenza ad elementi nulli della matrice di partenza; tale fenomeno (*fill in*) sarebbe gravoso dal punto di vista computazionale anche perché non consentirebbe la memorizzazione delle matrici di fattorizzazione L e U in una area di memoria analoga a quella occupata da A (nel caso che, come accade generalmente in MATLAB, venga sfruttata la sparsità nel processo di memorizzazione). Ricordiamo inoltre che la nostra finalità non è ora quella di fattorizzare la matrice e risolvere il sistema - che, ricordiamo, ha matrice dei coefficienti sparsa e di grandi dimensioni - con un metodo diretto, bensì quella di precondizionare il sistema assegnato per

poi risolverlo con un metodo iterativo; è dunque in questo contesto che vogliamo costruire delle approssimazioni dei fattori L e U della fattorizzazione LU di A.

L'algoritmo di fattorizzazione incompleta è essenzialmente quello della fattorizzazione LU con la differenza che ad ogni passo k viene calcolato

- il moltiplicatore m_{ik}, $i = k+1, ..., n$, solo se $a_{ik}^{(0)} = a_{ik} \neq 0$;
- il nuovo elemento $a_{ij}^{(k+1)}$, $j = k+1, ..., n$, solo se $a_{ij}^{(0)} = a_{ij} \neq 0$.

Ciò ci permette di sostenere un costo computazionale ben inferiore a quello della fattorizzazione LU e di preservare la sparsità di A. L'algoritmo è dunque

per $k = 1, \ldots, n-1$
  per $i = k+1, \ldots, n$
    se $a_{ik}^{(0)} \neq 0$
      si calcola $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$
      per $j = k+1, \ldots, n$
        se $a_{ij}^{(0)} \neq 0$
          si calcola $a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}$
        fine
      fine
    fine
  fine
fine

Costruiamo la matrice triangolare inferiore L_{inc}, che ha sulla diagonale tutti 1 e nella parte strettamente inferiore ha elementi nulli, laddove A aveva elementi nulli, e i moltiplicatori m_{ij} calcolati dall'algoritmo, altrimenti. Anche la matrice triangolare superiore $U_{\text{inc}} = A^{(n-1)}$ ha elementi nulli laddove A aveva elementi nulli. Notiamo che la matrice $L_{\text{inc}}U_{\text{inc}}$ coincide con la matrice A in ogni suo elemento diverso da zero; ovvero, denotando con R la matrice residuo $A - L_{\text{inc}}U_{\text{inc}}$, si ha

$$R_{ij} = 0, \qquad \text{per gli indici } (i,j) \text{ tali che } a_{ij} \neq 0. \tag{3.10}$$

3.2.2. Fattorizzazione di Cholesky incompleta. Assumiamo ora hermitiana definita positiva la matrice dei coefficienti - sparsa e di grandi dimensioni - del sistema lineare assegnato. Costruiamo una versione incompleta della fattorizzazione di Cholesky. In questo caso vogliamo dunque approssimare la matrice L triangolare inferiore - con elementi reali positivi sulla diagonale - che dà luogo alla fattorizzazione di Cholesky di A, con una matrice triangolare inferiore L_{inc} che abbia la medesima struttura di sparsità della parte triangolare inferiore di A. L'algoritmo di fattorizzazione incompleta è essenzialmente quello della fattorizzazione di Cholesky con la differenza che viene calcolato l'elemento ℓ_{ij}, $i > j$, solo se $a_{ij} \neq 0$. Si ha

per $j = 1, \ldots, n-1$

si calcola $\ell_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} |\ell_{jk}|^2}$
per $i = j+1, \ldots, n$
se $a_{ij} \neq 0$
si calcola $\ell_{ij} = (a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\bar{\ell}_{jk})/\ell_{jj}$
fine
fine
fine

L_{inc} è dunque formata da elementi nulli, laddove A aveva elementi nulli, oppure dagli elementi ℓ_{ij} calcolati dall'algoritmo. Notiamo che la matrice di precondizionamento $L_{\text{inc}}L_{\text{inc}}^H$ coincide con la matrice A in ogni suo elemento diverso da zero; ovvero, posto $R = A - L_{\text{inc}}L_{\text{inc}}^H$, vale la (3.10).

3.3. Metodo di Richardson non stazionario

Sia $\alpha_k \in \mathbb{R} \neq 0$ un paramentro dipendente dal passo k, il metodo di Richardson non stazionario può essere schematizzato come segue

per $k = 0, 1, \ldots$
si risolve il sistema lineare $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$
si calcola α_k
si aggiorna la soluzione $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}$
si aggiorna il residuo $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{z}^{(k)}$
fine

Ricordiamo che il calcolo di α^* in (3.3) per il metodo di Richardson stazionario richiede la conoscenza degli autovalori estremali di $P^{-1}A$, se sono soddisfatte le ipotesi del Teorema 3.5, quindi risulta di scarsa utilità pratica. Nel ben noto metodo del gradiente *steepest descent*, che illustreremo nel Paragrafo 3.5 e che è un metodo di Richardson non stazionario (non precondizionato), se la matrice dei coefficienti è hermitiana definita positiva, la valutazione del parametro dinamico α_k - pur se ripetuta ad ogni passo k - risulta comunque molto meno onerosa.

3.4. Metodi del gradiente

Per semplificare la trattazione, assumiamo d'ora in avanti (anche nel prossimo capitolo) che il sistema $A\mathbf{x} = \mathbf{b}$ abbia matrice dei coefficienti A, e termine noto $\mathbf{b}$ ad elementi in $\mathbb{R}$ (e quindi vale $\mathbf{x} \in \mathbb{R}^n$). Questa restrizione riguarda dunque la trattazione sia dei metodi del gradiente, precondizionati e non (in questo capitolo), che dei metodi basati su iterazioni in sottospazi di Krylov (nel prossimo capitolo).

Consideriamo il funzionale $\Phi : \mathbb{R}^n \to \mathbb{R}$,

$$\Phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}, \tag{3.11}$$

dove $A \in \mathbb{R}^{n\times n}$ è una matrice simmetrica definita positiva.

Il gradiente $\nabla\Phi(\mathbf{x})$ soddisfa

$$\nabla\Phi(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = -r(\mathbf{x}), \tag{3.12}$$

dove $r(\mathbf{x})$ è il residuo in $\mathbf{x}$ del sistema assegnato. Osserviamo infatti che

- $\nabla(\mathbf{x}^T\mathbf{b}) = \nabla(\sum_{i=1}^n \mathbf{x}_i\mathbf{b}_i) = \mathbf{b}$, dato che $\frac{\partial(\sum_{i=1}^n \mathbf{x}_i\mathbf{b}_i)}{\partial\mathbf{x}_h} = \mathbf{b}_h$, per $h = 1,\dots,n$;
- $\nabla(\mathbf{x}^T A\mathbf{x}) = \nabla(\sum_{i=1}^n\sum_{j=1}^n a_{ij}\mathbf{x}_i\mathbf{x}_j) = A\mathbf{x} + A^T\mathbf{x}$, dato che $\frac{\partial(\sum_{i=1}^n\sum_{j=1}^n a_{ij}\mathbf{x}_i\mathbf{x}_j)}{\partial\mathbf{x}_h} = \sum_{j=1}^n a_{hj}\mathbf{x}_j + \sum_{i=1}^n a_{ih}\mathbf{x}_i = (A\mathbf{x})_h + (A^T\mathbf{x})_h$, per $h = 1,\dots,n$.

Osserviamo che la matrice dei coefficienti del sistema coincide con la matrice hessiana associata alla forma quadratica Φ in (3.11), quindi l'ipotesi che A sia simmetrica definita positiva comporta che il vettore $\mathbf{x}^* \in \mathbb{R}^n$ tale che $\nabla\Phi(\mathbf{x}^*) = 0$, ovvero tale che $r(\mathbf{x}^*) = 0$, ovvero l'unica soluzione del sistema assegnato, risulta anche essere il punto di minimo globale di Φ.

Nei metodi del gradiente il problema è ricondotto dunque a determinare il punto di minimo di Φ partendo da un certo punto $\mathbf{x}^{(0)} \in \mathbb{R}^n$ e scegliendo ad ogni passo k una opportuna direzione $\mathbf{p}^{(k)}$ lungo la quale muoversi (per avvicinarsi il più rapidamente possibile alla soluzione $\mathbf{x}^*$) con un opportuno passo α_k. Si ha

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{p}^{(k)}. \tag{3.13}$$

La seguente proposizione illustra come calcolare il parametro α_k, dati i vettori $\mathbf{x}^{(k)}$ e $\mathbf{p}^{(k)}$, in modo da individuare il punto di minimo locale $\mathbf{x}^{(k+1)}$ per il funzionale Φ lungo la direzione $\mathbf{p}^{(k)}$ a partire dal punto $\mathbf{x}^{(k)}$.

PROPOSIZIONE 3.10. *Al k-esimo passo di un metodo del gradiente, data una direzione $\mathbf{p}^{(k)}$, il passo α_k tale che*

$$\alpha_k = \arg\min_{\alpha\in\mathbb{R}} \Phi(\mathbf{x}^{(k)} + \alpha\mathbf{p}^{(k)})$$

è dato da

$$\alpha_k = \frac{\mathbf{r}^{(k)T}\mathbf{p}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}}, \tag{3.14}$$

avendo denotato con $\mathbf{r}^{(k)}$ il residuo in $\mathbf{x}^{(k)}$. Inoltre, se la direzione $\mathbf{p}^{(k)}$ è una direzione di decrescita, ovvero se $\mathbf{p}^{(k)T}\nabla\Phi(\mathbf{x}^{(k)}) < 0$, allora $\alpha_k > 0$.

DIMOSTRAZIONE. Si cerca α che annulli la derivata rispetto ad α della funzione composta $\Phi(\mathbf{y}(\alpha))$, dove $\mathbf{y} : \mathbb{R} \to \mathbb{R}^n$ è tale che $\mathbf{y}(\alpha) = \mathbf{x}^{(k)} + \alpha\mathbf{p}^{(k)}$. Si ha

$$\frac{\partial\Phi(\mathbf{y}(\alpha))}{\partial\alpha} = \sum_{i=1}^n \frac{\partial\Phi}{\partial\mathbf{y}_i}\frac{d\mathbf{y}_i}{d\alpha} = \nabla\Phi(\mathbf{y}(\alpha))^T\mathbf{y}'(\alpha) = (A\mathbf{y}(\alpha) - \mathbf{b})^T\mathbf{p}^{(k)} =$$

$$\mathbf{y}(\alpha)^T A^T\mathbf{p}^{(k)} - \mathbf{b}^T\mathbf{p}^{(k)} = (\mathbf{x}^{(k)} + \alpha\mathbf{p}^{(k)})^T A\mathbf{p}^{(k)} - \mathbf{b}^T\mathbf{p}^{(k)},$$

quindi

$$\frac{\partial \Phi(\mathbf{y}(\alpha))}{\partial \alpha} = 0 \quad \text{se} \quad \alpha = \frac{(\mathbf{b} - A\mathbf{x}^{(k)})^T \mathbf{p}^{(k)}}{\mathbf{p}^{(k)T} A \mathbf{p}^{(k)}},$$

ovvero se α soddisfa la (3.14). Inoltre se la direzione $\mathbf{p}^{(k)}$ soddisfa $\mathbf{p}^{(k)T} \nabla \Phi(\mathbf{x}^{(k)}) < 0$, ovvero $\mathbf{p}^{(k)T} \mathbf{r}^{(k)} > 0$, allora, dato che A è simmetrica definita positiva, dalla (3.14) si ha $\alpha_k > 0$. □

Vale inoltre la seguente caratterizzazione.

PROPOSIZIONE 3.11. *Individuare il punto di minimo locale* $\mathbf{x}^{(k+1)}$ *per il funzionale* Φ *lungo una data direzione di decrescita* $\mathbf{p}^{(k)}$ *a partire dal punto* $\mathbf{x}^{(k)}$ *equivale a minimizzare lungo tale direzione la norma dell'energia dell'errore* $\mathbf{e}^{(k+1)} = \mathbf{x} - \mathbf{x}^{(k+1)}$*, dove* $\mathbf{x}$ *è la soluzione del sistema.*

DIMOSTRAZIONE. Consideriamo la funzione composta $\Phi(\mathbf{y}(\alpha))$, con $\mathbf{y}(\alpha) = \mathbf{x}^{(k)} + \alpha \mathbf{p}^{(k)}$. Si ha $\Phi(\mathbf{y}(\alpha)) = \frac{1}{2}\mathbf{y}(\alpha)^T A \mathbf{y}(\alpha) - \mathbf{y}(\alpha)^T \mathbf{A}\mathbf{x}$. D'altro canto si ha $\|\mathbf{x} - \mathbf{y}(\alpha)\|_A^2 = (\mathbf{x} - \mathbf{y}(\alpha))^T A(\mathbf{x} - \mathbf{y}(\alpha)) = \mathbf{x}^T A\mathbf{x} - 2\mathbf{y}(\alpha)^T A\mathbf{x} + \mathbf{y}(\alpha)^T A \mathbf{y}(\alpha)$. Ovvero $\|\mathbf{x} - \mathbf{y}(\alpha)\|_A^2 = \mathbf{x}^T A\mathbf{x} + 2\Phi(\mathbf{y}(\alpha))$. Quindi minimizzare $\Phi(\mathbf{y}(\alpha))$ equivale a minimizzare $\|\mathbf{x} - \mathbf{y}(\alpha)\|_A^2$, dato che la quantità $\mathbf{x}^T A\mathbf{x}$ non dipende da α. □

OSSERVAZIONE 3.3. *Dato che* $\|\mathbf{b} - A\mathbf{y}(\alpha)\|_{A^{-1}} = \|\mathbf{x} - \mathbf{y}(\alpha)\|_A$*, minimizzare lungo la direzione* $\mathbf{p}^{(k)}$ *la A-norma di* $\mathbf{e}^{(k+1)}$ *equivale a minimizzare lungo la direzione* $\mathbf{p}^{(k)}$ *la* A^{-1}*-norma di* $\mathbf{r}^{(k+1)}$*.*

Vale il seguente risultato di ortogonalità.

LEMMA 3.12. *Il residuo al passo* $(k+1)$*-esimo di un metodo del gradiente è ortogonale alla direzione* $\mathbf{p}^{(k)}$*.*

DIMOSTRAZIONE. Si ha che $\mathbf{r}^{(k+1)T} \mathbf{p}^{(k)} = (\mathbf{b} - A\mathbf{x}^{(k+1)})^T \mathbf{p}^{(k)} = (\mathbf{b} - A\mathbf{x}^{(k)} - \alpha_k A\mathbf{p}^{(k)})^T \mathbf{p}^{(k)} = \mathbf{r}^{(k)T} \mathbf{p}^{(k)} - \alpha_k \mathbf{p}^{(k)T} A \mathbf{p}^{(k)}$. Da (3.14), si ha la tesi $\mathbf{r}^{(k+1)T} \mathbf{p}^{(k)} = 0$. □

3.5. Metodo del gradiente steepest descent (SD)

La scelta della direzione nel metodo steepest descent (Cauchy, 1847) è quella più naturale: si pone

$$\mathbf{p}^{(k)} = \mathbf{r}^{(k)} \tag{3.15}$$

e, da (3.14), si ha

$$\alpha_k = \frac{\mathbf{r}^{(k)T} \mathbf{r}^{(k)}}{\mathbf{r}^{(k)T} A \mathbf{r}^{(k)}}. \tag{3.16}$$

Si tratta dunque di un metodo di Richardson non stazionario non precondizionato. La scelta (3.15) comporta la seguente proprietà.

PROPOSIZIONE 3.13. *Nel metodo steepest descent direzioni di decrescita successive, ovvero residui successivi, sono tra loro ortogonali.*

DIMOSTRAZIONE. Applicando il Lemma 3.12 alla direzione $\mathbf{p}^{(k)}$ in (3.15) si ha $\mathbf{r}^{(k+1)T}\mathbf{p}^{(k)} = \mathbf{r}^{(k+1)T}\mathbf{r}^{(k)} = \mathbf{p}^{(k+1)T}\mathbf{p}^{(k)} = 0$. □

L'algoritmo SD è il seguente:

per $k = 0, 1, \ldots$

si calcola la lunghezza del passo $\alpha_k = \dfrac{\mathbf{r}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{r}^{(k)T}A\mathbf{r}^{(k)}}$

si calcola la soluzione approssimata $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{r}^{(k)}$

si aggiorna il residuo $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{r}^{(k)}$

fine

OSSERVAZIONE 3.4. *Nell'implementazione dell'algoritmo sarà doveroso tenere presente che $A\mathbf{r}^{(k)}$ compare sia nel calcolo di α_k che nel calcolo del nuovo residuo. Il prodotto $A\mathbf{r}^{(k)}$ deve quindi essere eseguito una volta sola e memorizzato.*

Notiamo che nel metodo di Richardson non stazionario con passo (3.16), il passo è funzione del residuo nell'iterazione precedente, mentre nel metodo di Richardson stazionario con parametro ottimale è funzione degli autovalori estremali di A; il costo computazionale complessivo per la valutazione dei passi risulterà molto minore del costo per il calcolo dello spettro di A. Nonostante ciò, il seguente risultato mostra che l'errore del metodo SD soddisfa la (3.9).

PROPOSIZIONE 3.14. *Nel metodo SD vale, per $k = 0, 1, \ldots$*

$$\|\mathbf{e}^{(k+1)}\|_A \leq \frac{\kappa_2(A)-1}{\kappa_2(A)+1}\|\mathbf{e}^{(k)}\|_A \leq \left(\frac{\kappa_2(A)-1}{\kappa_2(A)+1}\right)^{k+1}\|\mathbf{e}^{(0)}\|_A. \tag{3.17}$$

DIMOSTRAZIONE. Se si fa un passo del metodo di Richardson stazionario non precondizionato con il passo ottimale α^* in (3.3) a partire da $\mathbf{x}^{(k)}$, si ha $\mathbf{x}_R^{(k+1)} := \mathbf{x}^{(k)} + \alpha^*\mathbf{r}^{(k)}$ e vale $\|\mathbf{e}_R^{(k+1)}\|_A \leq \frac{\kappa_2(A)-1}{\kappa_2(A)+1}\|\mathbf{e}^{(k)}\|_A$ (si veda Corollario 3.9), avendo denotato con $\mathbf{e}_R^{(k+1)}$ l'errore $\mathbf{x} - (\mathbf{x}^{(k)} + \alpha^*\mathbf{r}^{(k)})$. Grazie alla ottimalità del parametro dinamico α_k nei metodi del gradiente (si veda Proposizione 3.11), se si pone $\mathbf{p}^{(k)} = \mathbf{r}^{(k)}$ si ha $\alpha_k = \arg\min_{\alpha\in\mathbb{R}}\|\mathbf{x} - (\mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)})\|_A$, quindi $\|\mathbf{e}^{(k+1)}\|_A \leq \|\mathbf{e}_R^{(k+1)}\|_A$, avendo denotato con $\mathbf{e}^{(k+1)}$ l'errore $\mathbf{x} - (\mathbf{x}^{(k)} + \alpha_k\mathbf{r}^{(k)})$. □

La convergenza del metodo del gradiente steepest descent può essere molto lenta se è grande il numero di condizionamento $\kappa_2(A) = \lambda_1/\lambda_n$, avendo denotato con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ gli autovalori di A.
Per illustrare meglio la problematica, consideriamo una matrice diagonale Λ di ordine 2 che sia simmetrica definita positiva, ovvero con elementi diagonali $\lambda_1 \geq \lambda_2 > 0$. Le linee di livello della forma quadratica (3.11), $\Phi(x_1, x_2) = c$, sono circonferenze se $\lambda_1 = \lambda_2$, altrimenti sono ellissi concentriche tanto più eccentriche (ovvero tanto più "allungate") quanto più grande è il rapporto λ_1/λ_2. Con il metodo SD, da $\mathbf{x}^{(k)}$ ci si muove lungo la direzione del gradiente $\nabla(\Phi(\mathbf{x}^{(k)}))$ quindi ortogonalmente alle curve di livello. Se $\lambda_1 = \lambda_2$, in un solo passo da $\mathbf{x}^{(0)}$ si arriva

alla soluzione. Se $\lambda_1 >> \lambda_2$, il percorso è lungo (e a zig-zag dato che direzioni successive sono ortogonali).

Applicando invece il metodo del gradiente coniugato, descritto nel Paragrafo 3.6, ad un sistema lineare con matrice dei coefficienti Λ, vedremo che - qualsiasi sia il rapporto λ_1/λ_2 tra gli elementi diagonali di Λ - si giungerà a convergenza al più in 2 passi.

Queste considerazioni si possono estendere da $\mathbb{R}^2$ a $\mathbb{R}^n$: le superfici di livello sono iperelissoidi e l'interpretazione geometrica è la medesima. In questo caso, il metodo del gradiente coniugato converge al più in n passi.

3.6. Metodo del gradiente coniugato (CG)

Il metodo del gradiente coniugato (Hestenes-Stiefel, 1950) costruisce la nuova direzione di discesa come combinazione lineare della direzione precedente e del residuo. L'algoritmo è il seguente:

si definisce la direzione di ricerca $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$

per $k = 0, 1, \ldots$

si calcola la lunghezza del passo $\alpha_k = \dfrac{\mathbf{r}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}}$

si aggiorna la soluzione $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$

si aggiorna il residuo $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{p}^{(k)}$

si calcola $\beta_{k+1} = \dfrac{\mathbf{r}^{(k+1)T}\mathbf{r}^{(k+1)}}{\mathbf{r}^{(k)T}\mathbf{r}^{(k)}}$

si aggiorna la direzione di ricerca $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_{k+1}\mathbf{p}^{(k)}$

fine

OSSERVAZIONE 3.5. *Nell'implementazione dell'algoritmo bisogna calcolare una sola volta e memorizzare sia* $A\mathbf{p}^{(k)}$ *che* $\mathbf{r}^{(k+1)T}\mathbf{r}^{(k+1)}$*, tenendo inoltre conto che il prodotto scalare sarà usato anche nel passo successivo.*

Il passo (dinamico) che compare nell'algoritmo si ottiene da (3.14):

$$\alpha_0 = \frac{\mathbf{r}^{(0)T}\mathbf{p}^{(0)}}{\mathbf{p}^{(0)T}A\mathbf{p}^{(0)}} = \frac{\mathbf{r}^{(0)T}\mathbf{r}^{(0)}}{\mathbf{p}^{(0)T}A\mathbf{p}^{(0)}};$$

$$\alpha_k = \frac{\mathbf{r}^{(k)T}\mathbf{p}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}} = \frac{\mathbf{r}^{(k)T}(\mathbf{r}^{(k)} + \beta_k\mathbf{p}^{(k-1)})}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}} = \frac{\mathbf{r}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}}, \quad \text{per } k \geq 1,$$

dove nell'ultimo passaggio si è tenuto conto del fatto che per il Lemma 3.12 si ha $\mathbf{r}^{(k)T}\mathbf{p}^{(k-1)} = 0$.

Per ricavare la proprietà fondamentale del metodo CG, ovvero il Teorema 3.17, facciamo uso dei seguenti risultati.

LEMMA 3.15. *Se si ha, per $k \geq 1$, $\mathbf{r}^{(k)T}\mathbf{r}^{(k-1)} = 0$ e, per $k > 1$, $\mathbf{p}^{(k-2)T}A\mathbf{p}^{(k-1)} = 0$, allora vale*

$$\beta_k = -\frac{\mathbf{r}^{(k)T}A\mathbf{p}^{(k-1)}}{\mathbf{p}^{(k-1)T}A\mathbf{p}^{(k-1)}}.$$

(In effetti, spesso viene data nell'Algoritmo CG questa formulazione alternativa di β_k, che però risulta computazionalmente meno conveniente.)

DIMOSTRAZIONE. È immediato verificare che nelle ipotesi del lemma valgono le sequenti identità:
per $k \geq 1$, dato che $\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \alpha_{k-1}A\mathbf{p}^{(k-1)}$, vale

$$\mathbf{r}^{(k)T}A\mathbf{p}^{(k-1)} = -\mathbf{r}^{(k)T}\frac{\mathbf{r}^{(k)} - \mathbf{r}^{(k-1)}}{\alpha_{k-1}} = -\frac{\mathbf{r}^{(k)T}\mathbf{r}^{(k)}}{\alpha_{k-1}};$$

per $k > 1$, dato che $\mathbf{p}^{(k-1)} = \mathbf{r}^{(k-1)} + \beta_{k-1}\mathbf{p}^{(k-2)}$, vale

$$\mathbf{p}^{(k-1)T}A\mathbf{p}^{(k-1)} = -\mathbf{r}^{(k-1)T}\frac{\mathbf{r}^{(k)} - \mathbf{r}^{(k-1)}}{\alpha_{k-1}} + \beta_{k-1}\mathbf{p}^{(k-2)T}A\mathbf{p}^{(k-1)} = \frac{\mathbf{r}^{(k-1)T}\mathbf{r}^{(k-1)}}{\alpha_{k-1}};$$

per $k = 1$, dato che $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$, vale

$$\mathbf{p}^{(0)T}A\mathbf{p}^{(0)} = -\mathbf{r}^{(0)T}\frac{\mathbf{r}^{(1)} - \mathbf{r}^{(0)}}{\alpha_0} = \frac{\mathbf{r}^{(0)T}\mathbf{r}^{(0)}}{\alpha_0}.$$

□

TEOREMA 3.16. *Vale la seguente identità tra sottospazi*

$$\text{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(k-1)}\} = \text{span}\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(k-1)}\}. \tag{3.18}$$

Inoltre, i residui sono mutuamente ortogonali,

$$\mathbf{r}^{(k)T}\mathbf{r}^{(j)} = 0, \quad \textit{per } j < k, \tag{3.19}$$

e le direzioni di discesa sono $A-$coniugate,

$$\mathbf{p}^{(k)T}A\mathbf{p}^{(j)} = 0, \quad \textit{per } j < k. \tag{3.20}$$

DIMOSTRAZIONE. È immediato verificare la (3.18) osservando che $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$ e, per $k \geq 0$, $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_{k+1}\mathbf{p}^{(k)}$.
Dimostriamo ora contemporaneamente le identità in (3.19) e (3.20), procedendo per induzione su k.
Per prima cosa verifichiamo che valgono $\mathbf{r}^{(1)T}\mathbf{r}^{(0)} = \mathbf{r}^{(1)T}\mathbf{p}^{(0)} = 0$ e $\mathbf{p}^{(1)T}A\mathbf{p}^{(0)} = (\mathbf{r}^{(1)} + \beta_1\mathbf{p}^{(0)})^T A\mathbf{p}^{(0)} = 0$, rispettivamente per il Lemma 3.12 e per il Lemma 3.15 relativo al caso $k = 1$. A questo punto assumiamo entrambi i risultati veri al passo $(k-1)$-esimo e

i) mostriamo la (3.19), ovvero che

$$\mathbf{r}^{(k)T}\mathbf{r}^{(j)} = \mathbf{r}^{(k-1)T}\mathbf{r}^{(j)} - \alpha_{k-1}\mathbf{p}^{(k-1)T}A\mathbf{r}^{(j)} = 0, \quad \text{per } j < k.$$

In effetti, se $j < k-1$, i due addendi sono entrambi uguali a zero per le ipotesi induttive e per la (3.18); inoltre, se $j = k-1$, dalla definizione di α_{k-1} si ha

$$\mathbf{r}^{(k)T}\mathbf{r}^{(k-1)} = \mathbf{r}^{(k-1)T}\mathbf{r}^{(k-1)} - \frac{\mathbf{r}^{(k-1)T}\mathbf{r}^{(k-1)}}{\mathbf{p}^{(k-1)T}A\mathbf{p}^{(k-1)}}\mathbf{p}^{(k-1)T}A\mathbf{r}^{(k-1)}$$

e, da $\mathbf{p}^{(k-1)} = \mathbf{r}^{(k-1)} + \beta_{k-1}\mathbf{p}^{(k-2)}$, il denominatore si può riscrivere come

$$\mathbf{p}^{(k-1)T}A(\mathbf{r}^{(k-1)} + \beta_{k-1}\mathbf{p}^{(k-2)}) = \mathbf{p}^{(k-1)T}A\mathbf{r}^{(k-1)},$$

dato che $\beta_{k-1}\mathbf{p}^{(k-1)T}A\mathbf{p}^{(k-2)} = 0$ per l'ipotesi induttiva sulle direzioni di decrescita A-coniugate, quindi semplificando si ottiene la (3.19);

ii) mostriamo la (3.20), ovvero che

$$\mathbf{p}^{(k)T}A\mathbf{p}^{(j)} = \mathbf{r}^{(k)T}A\mathbf{p}^{(j)} + \beta_k\mathbf{p}^{(k-1)T}A\mathbf{p}^{(j)} = 0 \quad \text{per } j < k.$$

In effetti, se $j < k-1$, il primo addendo è nullo per la (3.19) dato che

$$A\mathbf{p}^{(j)} = -\frac{\mathbf{r}^{(j+1)} - \mathbf{r}^{(j)}}{\alpha_j} \in \text{span}\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(k-1)}\},$$

mentre il secondo addendo è nullo per l'ipotesi induttiva relativa alle direzioni A-coniugate; infine, se $j = k-1$, la (3.19) e l'ipotesi induttiva sulle direzioni A-coniugate consentono di applicare il Lemma 3.15, quindi semplificando si ottiene la (3.20). □

Il seguente risultato è di fondamentale importanza e implica che, al netto dagli errori di calcolo, il metodo CG può essere considerato un metodo diretto!

TEOREMA 3.17. *Sia $A \in \mathbb{R}^{n\times n}$ simmetrica definita positiva, applicando il metodo CG al sistema $A\mathbf{x} = \mathbf{b}$, a partire da una qualsiasi approssimazione iniziale $\mathbf{x}^{(0)}$, si ha la convergenza al più in n passi alla soluzione esatta $\mathbf{x}$.*

DIMOSTRAZIONE. Dalla (3.19), $\mathbf{r}^{(n)} = \mathbf{0}$ perché si ha $\text{span}\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(n-1)}\} \equiv \mathbb{R}^n$ e $\mathbf{r}^{(n)T}\mathbf{r}^{(j)} = 0$ per $j < n$, quindi $\mathbf{x}^{(n)} = \mathbf{x}$, ovvero la tesi. □

Rispetto al metodo SD si può dimostrare un fattore di abbattimento dell'errore più favorevole:

$$\|\mathbf{e}^{(k+1)}\|_A \leq 2\left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}\right)^{k+1} \|\mathbf{e}^{(0)}\|_A. \tag{3.21}$$

OSSERVAZIONE 3.6. *Per il Teorema 3.17, in aritmetica esatta si ha $\mathbf{e}^{(m)} = \mathbf{0}$ ad un certo passo $m \leq n$. In aritmetica finita, in caso di riscontrata perdita di ortogonalità dei residui (e quindi di A-ortogonalità delle direzioni di discesa) ad un certo passo $k \leq n$ per l'accumulo di errori di arrotondamento, si può provare a far ripartire il metodo con $\mathbf{x}^{(0)} := \mathbf{x}^{(k)}$ e dunque con $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(k)}$ (strategia del restart).*

3.7. Metodi del gradiente precondizionati

Al fine di migliorare il condizionamento del sistema assegnato $A\mathbf{x} = \mathbf{b}$, con matrice dei coefficienti A simmetrica definita positiva, si usano le tecniche di precondizionamento illustrate nel Paragrafo 3.2. Abbiamo osservato che se si precondiziona il sistema a sinistra con una matrice P simmetrica definita positiva, si ottiene il sistema $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$ con matrice dei coefficienti $P^{-1}A$ non più simmetrica definita positiva. Dunque, per non perdere la proprietà della matrice dei coefficienti di essere simmetrica definita positiva (condizione base per definire un metodo del gradiente), si dovrebbe applicare il metodo del gradiente per esempio a $L_{\text{inc}}^{-1}AL_{\text{inc}}^{-T}\mathbf{y} = L_{\text{inc}}^{-1}\mathbf{b}$, con medesima soluzione $\mathbf{x} = L_{\text{inc}}^{-T}\mathbf{y}$, in modo da clusterizzare lo spettro di $P^{-1}A$, dove $P = L_{\text{inc}}L_{\text{inc}}^{T}$, e migliorarne il condizionamento (dato che la matrice precondizionata ha i medesimi autovalori di $P^{-1}A$). Ovviamente, come si è detto, il discorso rimane valido per una qualsiasi matrice di precondizionamento simmetrica definita positiva P, usando i fattori della sua decomposizione di Cholesky, oppure, ad esempio, $P_S = P_D = P^{\frac{1}{2}}$.

Tuttavia, di fatto, per definire le formule dell'algoritmo precondizionato del gradiente, per esempio dell'algoritmo del gradiente applicato al sistema precondizionato $L_{\text{inc}}^{-1}AL_{\text{inc}}^{-T}\mathbf{y} = L_{\text{inc}}^{-1}\mathbf{b}$, si applica soltanto in maniera implicita il metodo del gradiente a tale sistema, senza mai formarlo esplicitamente. Questa strategia nel definire gli schemi precondizionati del gradiente si basa sulle seguenti considerazioni.

OSSERVAZIONE 3.7. *Se A è simmetrica definita positiva per il prodotto scalare euclideo, $P^{-1}A$ risulta simmetrica definita positiva per il P-prodotto scalare. Infatti, si ha*

$$(P^{-1}A\mathbf{x},\mathbf{y})_P = (A\mathbf{x},\mathbf{y}) = (\mathbf{x},A\mathbf{y}) = (\mathbf{x},P^{-1}A\mathbf{y})_P,$$
$$(P^{-1}A\mathbf{x},\mathbf{x})_P = (A\mathbf{x},\mathbf{x}) > 0, \quad \forall \mathbf{x} \neq \mathbf{0}.$$

Di fatto si sta minimizzando il medesimo funzionale, dato che

$$\Phi_P(\mathbf{x}) = \frac{1}{2}(P^{-1}A\mathbf{x},\mathbf{x})_P - (P^{-1}\mathbf{b},\mathbf{x})_P = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} = \Phi(\mathbf{x}),$$

e quindi annullando il medesimo gradiente $\nabla\Phi_P(\mathbf{x}) = \nabla\Phi(\mathbf{x}) = -r(\mathbf{x})$.

Seguendo questa strategia, data la direzione $\mathbf{p}^{(k)}$, il passo ottimale α_k è il medesimo in (3.14) :

$$\alpha_k = \frac{(\mathbf{p}^{(k)},\mathbf{z}^{(k)})_P}{(P^{-1}A\mathbf{p}^{(k)},\mathbf{p}^{(k)})_P} = \frac{\mathbf{z}^{(k)T}P\mathbf{p}^{(k)}}{\mathbf{p}^{(k)T}PP^{-1}A\mathbf{p}^{(k)}} = \frac{\mathbf{r}^{(k)T}\mathbf{p}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}},$$

avendo denotato con $\mathbf{z}^{(k)}$ il residuo precondizionato in $\mathbf{x}^{(k)}$. Inoltre, se la direzione $\mathbf{p}^{(k)}$ è una direzione di decrescita, ovvero se $(\nabla\Phi(\mathbf{x})^{(k)},\mathbf{p}^{(k)})_P < 0$, allora $\alpha_k > 0$.

Rimangono invariate le considerazioni fatte sia nella Proposizione 3.11 che nella Osservazione 3.3. Infine, per quanto riguarda il Lemma 3.12, si dimostra analogamente che $(\mathbf{p}^{(k)},\mathbf{z}^{(k+1)})_P = \mathbf{r}^{(k+1)T}\mathbf{p}^{(k)} = 0$. Dunque anche il risultato nel Lemma 3.12 rimane invariato.

OSSERVAZIONE 3.8. *Come criterio di arresto nei metodi del gradiente (anche se precondizionati) si usa generalmente controllare il rapporto tra la norma euclidea del residuo (non la norma euclidea del residuo precondizionato) e la norma euclidea del termine noto. Il calcolo della norma dell'energia del residuo aumenterebbe nettamente il costo computazionale ad ogni passo del metodo.*

3.8. Metodo SD precondizionato (PSD)

Si sceglie la direzione

$$\mathbf{p}^{(k)} = \mathbf{z}^{(k)} = P^{-1}\mathbf{r}^{(k)}, \tag{3.22}$$

con P simmetrica definita positiva. In linea con (3.14), si ha

$$\alpha_k = \frac{\mathbf{z}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{z}^{(k)T}A\mathbf{z}^{(k)}} > 0. \tag{3.23}$$

L'algoritmo PSD è dunque il seguente:

per $k = 0, 1, \ldots$
si risolve il sistema lineare $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$
si calcola la lunghezza del passo $\alpha_k = \dfrac{\mathbf{z}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{z}^{(k)T}A\mathbf{z}^{(k)}}$
si aggiorna la soluzione $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{z}^{(k)}$
si aggiorna il residuo $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{z}^{(k)}$
fine

Quindi l'algoritmo PSD è un metodo di Richardson non stazionario. Analogamente ai risultati di convergenza del metodo di Richardson stazionario ottimale, compare il numero di condizionamento di $P^{-1}A$ nella maggiorazione della norma dell'errore $\mathbf{e}^{(k+1)}$:

PROPOSIZIONE 3.18. *Nel metodo steepest descent precondizionato con P simmetrica definita positiva vale, per* $k = 0, 1, \ldots$

$$\|\mathbf{e}^{(k+1)}\|_A \leq \frac{\kappa_2(P^{-1}A) - 1}{\kappa_2(P^{-1}A) + 1}\|\mathbf{e}^{(k)}\|_A \leq \left(\frac{\kappa_2(P^{-1}A) - 1}{\kappa_2(P^{-1}A) + 1}\right)^{k+1}\|\mathbf{e}^{(0)}\|_A. \tag{3.24}$$

DIMOSTRAZIONE. Si applica la Proposizione 3.11 al metodo precondizionato (con direzione di decrescita definita in (3.22) e passo α_k in (3.23)) e si procede come nella dimostrazione del risultato in (3.17) - relativo al metodo non precondizionato - facendo prima un passo del metodo di Richardson stazionario precondizionato con parametro ottimale, $\mathbf{x}^{(k)} + \alpha^*\mathbf{z}^{(k)}$, in modo da ottenere - per il Teorema 3.8 - la maggiorazione in (3.6), e sfruttando poi l'ottimalità di α_k. □

3.9. Metodo CG precondizionato (PCG)

Posto $\mathbf{p}^{(0)} = \mathbf{z}^{(0)}$, la direzione di ricerca al passo $(k+1)$-esimo, $k \geq 0$ è data da $\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} + \beta_{k+1}\mathbf{p}^{(k)}$, dove

$$\beta_{k+1} = \frac{(\mathbf{z}^{(k+1)}, \mathbf{z}^{(k+1)})_P}{(\mathbf{z}^{(k)}, \mathbf{z}^{(k)})_P} = \frac{\mathbf{z}^{(k+1)T}\mathbf{r}^{(k+1)}}{\mathbf{z}^{(k)T}\mathbf{r}^{(k)}}.$$

Il passo, per la (3.14) e per il Lemma 3.12, è dato da

$$\alpha_k = \frac{\mathbf{z}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}} > 0.$$

Si perviene dunque al seguente algoritmo:

si risolve il sistema lineare $P\mathbf{z}^{(0)} = \mathbf{r}^{(0)}$
si definisce la direzione di ricerca $\mathbf{p}^{(0)} = \mathbf{z}^{(0)}$
per $k = 0, 1, \ldots$
  si calcola la lunghezza del passo $\alpha_k = \dfrac{\mathbf{z}^{(k)T}\mathbf{r}^{(k)}}{\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}}$
  si aggiorna la soluzione $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{p}^{(k)}$
  si aggiorna il residuo $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{p}^{(k)}$
  si risolve il sistema lineare $P\mathbf{z}^{(k+1)} = \mathbf{r}^{(k+1)}$
  si calcola $\beta_{k+1} = \dfrac{\mathbf{z}^{(k+1)T}\mathbf{r}^{(k+1)}}{\mathbf{z}^{(k)T}\mathbf{r}^{(k)}}$
  si aggiorna la direzione di ricerca $\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} + \beta_{k+1}\mathbf{p}^{(k)}$
fine

Analogamente al Lemma 3.15, si ha la seguente formulazione alternativa per β_k:

$$\beta_k = -\frac{(P^{-1}A\mathbf{p}^{(k-1)}, \mathbf{z}^{(k)})_P}{(P^{-1}A\mathbf{p}^{(k-1)}, \mathbf{p}^{(k-1)})_P} = -\frac{\mathbf{z}^{(k)T}A\mathbf{p}^{(k-1)}}{\mathbf{p}^{(k-1)T}A\mathbf{p}^{(k-1)}}.$$

Si ha inoltre $\text{span}\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(k-1)}\} = \text{span}\{\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(k-1)}\}$, quindi - analogamente al Teorema 3.16 - si dimostra che sia i residui che i residui precondizionati sono mutuamente ortogonali,

$$\mathbf{r}^{(k)T}\mathbf{r}^{(j)} = \mathbf{z}^{(k)T}\mathbf{z}^{(j)} = 0, \quad \text{per } j < k,$$

e che le direzioni di discesa sono $A-$coniugate,

$$\mathbf{p}^{(k)T}A\mathbf{p}^{(j)} = 0, \quad \text{per } j < k.$$

La convergenza in n passi del metodo PCG si dimostra analogamente al Teorema 3.17. Infine, vale

$$\|\mathbf{e}^{(k+1)}\|_A \leq 2\left(\frac{\sqrt{\kappa_2(P^{-1}A)} - 1}{\sqrt{\kappa_2(P^{-1}A)} + 1}\right)^{k+1} \|\mathbf{e}^{(0)}\|_A.$$