Il metodo del Gradiente Coniugato

Gerardo Toraldo

Università di Napoli Federico II

A.A. 2014-2015

Minimizzazione di una funzione quadratica convessa

$$\min f(x) = \frac{1}{2} x^T A x - b^T x \quad \text{s.t.} \quad x \in \Re^n \tag{1}$$

$$\text{con} \quad A \in \Re^{n \times n} \quad \text{simmetrica definita positiva} \quad e \quad b \in \Re^n$$

Risolvere il problema (1) è equivalente a risolvere uno dei problemi seguenti:

$$Ax = b,$$

$$\min_{x \in \Re} \|Ax - b\|.$$

Metodo Steepest Descent (SD)

$x_0 \in \Re^n$; $k \leftarrow 0$;

$g_0 \leftarrow Ax_0 - b$;

while `not_stopcondition`

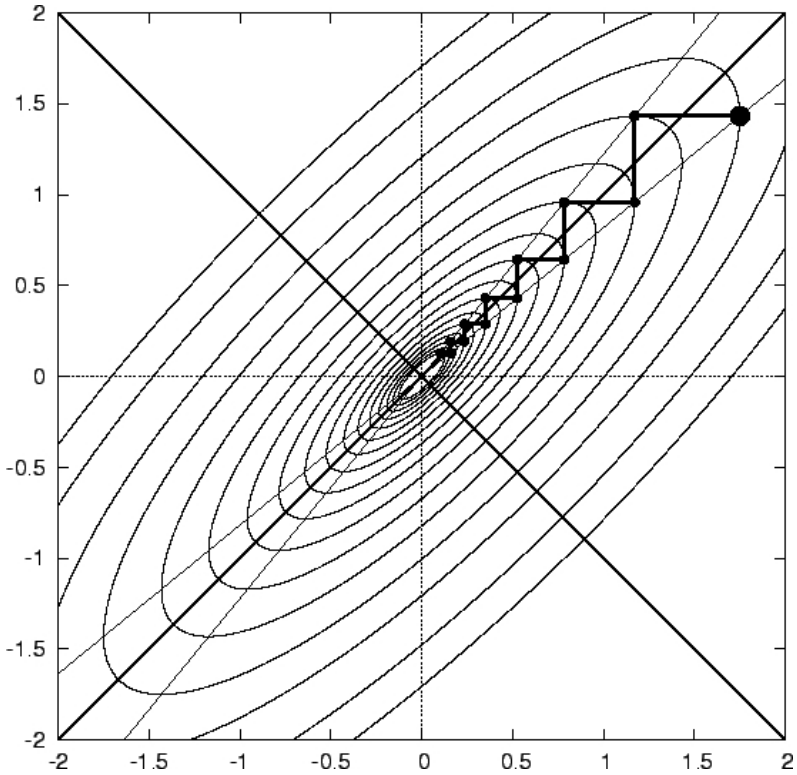
$\alpha_k = \frac{g_k^T g_k}{g_k^T A g_k}$;

$x_{k+1} \leftarrow x_k - \alpha_k g_k$;

$g_{k+1} \leftarrow g_k - \alpha_k A g_k$;

$k \leftarrow k + 1$;

endwhile



Notiamo che $x_{k+1} = x_k + v^*$ dove

$$v^* = \text{argmin}\, f(x_k + v) \;\; \text{con } v \in \text{Span}\{g_k\}$$

Algoritmo di discesa con "passo ottimo"

$x_0 \in \Re^n$; $g_0 \leftarrow Ax_0 - b$;; $k \leftarrow 0$;

$g_0 \leftarrow Ax_0 - b$;

while *not_stopcondition*

scegli v_k;

$\alpha_k = -\frac{v_k^T g_k}{v_k^T A v_k}$;

$x_{k+1} \leftarrow x_k + \alpha_k v_k$;

$g_{k+1} \leftarrow g_k + \alpha_k A v_k$;

$k \leftarrow k + 1$;

endwhile

Notiamo che

$$x_{k+1} = x_k + v^*, \text{ con}$$
$$v^* = \operatorname{argmin} f(x_k + v)$$
$$\text{e } v \in \operatorname{Span}\{v_k\}$$

Ad ogni passo la funzione obiettivo $f(x)$ **viene minimizzata in uno spazio affine unidimensionale (o, equivalentemente,** $f(x_0 + v)$ **viene minimizzata in uno spazio vettoriale unidimensionale)**

E' possibile minimizzare $f(x_0 + v)$ in **sottospazi di dimensione via via crescente?**

$$x_1 = \operatorname{argmin} f(x) \text{ con } x = x_0 + v,\ v \in \operatorname{Span}\{v_0\}$$
$$x_2 = \operatorname{argmin} f(x) \text{ con } x = x_0 + v,\ v \in \operatorname{Span}\{v_0, v_1\}$$
$$\ldots\ldots\ldots\ldots$$
$$x_{k+1} = \operatorname{argmin} f(x) \text{ con } x = x_0 + v,\ v \in \operatorname{Span}\{v_0, v_1, ..., v_k\}$$

Algoritmo di discesa con "passo ottimo"

$x_0 \in \Re^n$; $g_0 \leftarrow Ax_0 - b$;; $k \leftarrow 0$;

$g_0 \leftarrow Ax_0 - b$;

while *not_stopcondition*

scegli v_k;

$\alpha_k = -\frac{v_k^T g_k}{v_k^T A v_k}$;

$x_{k+1} \leftarrow x_k + \alpha_k v_k$;

$g_{k+1} \leftarrow g_k + \alpha_k A v_k$;

$k \leftarrow k + 1$;

endwhile

Notiamo che

$$x_{k+1} = x_k + v^*, \text{ con}$$
$$v^* = \operatorname{argmin} f(x_k + v)$$
$$\text{e } v \in \operatorname{Span}\{v_k\}$$

Ad ogni passo la funzione obiettivo $f(x)$ **viene minimizzata in uno spazio affine unidimensionale (o, equivalentemente,** $f(x_0 + v)$ **viene minimizzata in uno spazio vettoriale unidimensionale)**

E' possibile minimizzare $f(x_0 + v)$ in **sottospazi di dimensione via via crescente**?

$$x_1 = \operatorname{argmin} f(x) \text{ con } x = x_0 + v,\ v \in \operatorname{Span}\{v_0\}$$
$$x_2 = \operatorname{argmin} f(x) \text{ con } x = x_0 + v,\ v \in \operatorname{Span}\{v_0, v_1\}$$
$$\ldots\ldots\ldots\ldots$$
$$x_{k+1} = \operatorname{argmin} f(x) \text{ con } x = x_0 + v,\ v \in \operatorname{Span}\{v_0, v_1, ..., v_k\}$$

Minimizzazione in sottospazi

Il problema

$$\min f(x) = \frac{1}{2}x^T Ax - b^T x \ \text{ s.t. } \ x = x_0 + v \ \text{ con } v \in S = \text{Span}\{v_0, v_1, ..., v_k\}$$

può essere formulato come segue:

$$\min f(x_0 + \alpha_0 v_0 + \alpha_1 v_1 + ... + \alpha_k v_k) \quad \Leftrightarrow$$
$$\min f(x_0 + V\alpha) \ \text{ con } \ \alpha \in \Re^{k+1}, \ \ V = [v_0, v_1, ..., v_k] \in \Re^{n \times k+1}$$

Quindi

$$\overline{\alpha} = \text{argmin } f(x_0 + V\alpha) \quad \Leftrightarrow \quad \nabla_\alpha f(x_0 + V\overline{\alpha}) = 0$$
$$\Leftrightarrow \quad V^T \nabla f(x_0 + V\overline{\alpha}) = 0$$

Condizioni di ortogonalità

$$v_i^T \nabla f(x_0 + V\overline{\alpha}) = 0, \ \ i = 0, 1, ..., k \quad \Leftrightarrow \quad \nabla f(x_0 + V\overline{\alpha}) \in S^\perp = \text{Span}\{v_0, v_1, ..., v_k\}^\perp$$

Minimizzazione in sottospazi

Il problema

$$\min f(x) = \frac{1}{2}x^T Ax - b^T x \ \text{ s.t. } \ x = x_0 + v \ \text{ con } v \in S = \text{Span}\{v_0, v_1, ..., v_k\}$$

può essere formulato come segue:

$$\begin{gathered} \min f(x_0 + \alpha_0 v_0 + \alpha_1 v_1 + ... + \alpha_k v_k) \quad \Leftrightarrow \\ \min f(x_0 + V\alpha) \ \text{ con } \ \alpha \in \Re^{k+1}, \ \ V = [v_0, v_1, ..., v_k] \in \Re^{n \times k+1} \end{gathered}$$

Quindi

$$\begin{aligned} \overline{\alpha} = \text{argmin } f(x_0 + V\alpha) \quad &\Leftrightarrow \quad \nabla_\alpha f(x_0 + V\overline{\alpha}) = 0 \\ &\Leftrightarrow \quad V^T \nabla f(x_0 + V\overline{\alpha}) = 0 \end{aligned}$$

Condizioni di ortogonalità

$$v_i^T \nabla f(x_0 + V\overline{\alpha}) = 0, \ \ i = 0, 1, ..., k \quad \Leftrightarrow \quad \nabla f(x_0 + V\overline{\alpha}) \in S^\perp = \text{Span}\{v_0, v_1, ..., v_k\}^\perp$$

Minimizzazione in un sottospazio bidimensionale: perché scegliere direzioni coniugate?

Supponiamo di aver eseguito le prime due iterazioni:

$$x_1 = x_0 + \alpha_0 v_0;\ x_2 = x_1 + \alpha_1 v_1.$$

La scelta del passo ottimo garantisce che

$$\nabla f(x_1)^T v_0 = 0 \ \text{ e } \ \nabla f(x_2)^T v_1 = 0.$$

Richiedere che x_2 **minimizzi la funzione in** $\mathbf{Span}\{v_0, v_1\}$ significa richiedere che valga anche la condizione di ortogonalità

$$\nabla f(x_2)^T v_0 = 0.$$

Ricordando che $\nabla f(x_2) = \nabla f(x_1) + \alpha_1 A v_1$, abbiamo

$$\nabla f(x_2)^T v_0 = \nabla f(x_1)^T v_0 + \alpha_1 (A v_1)^T v_0 = \alpha_1 v_0^T A v_1$$

e, dunque, x_2 minimizza la funzione nello spazio affine

$$A = \{x : x = x_0 + \alpha_0 v_0 + \alpha_1 v_1, \ \ \alpha_0, \alpha_1 \in \Re\}$$

se e solo se **le direzioni di ricerca** v_0 **e** v_1 **sono** A**-coniugate, ovvero** $v_0^T A v_1 = 0$.

Minimizzazione in un sottospazio bidimensionale: perché scegliere direzioni coniugate?

Supponiamo di aver eseguito le prime due iterazioni:

$$x_1 = x_0 + \alpha_0 v_0;\ x_2 = x_1 + \alpha_1 v_1.$$

La scelta del passo ottimo garantisce che

$$\nabla f(x_1)^T v_0 = 0 \ \text{ e } \ \nabla f(x_2)^T v_1 = 0.$$

Richiedere che x_2 **minimizzi la funzione in** $\mathbf{Span}\{v_0, v_1\}$ significa richiedere che valga anche la condizione di ortogonalità

$$\nabla f(x_2)^T v_0 = 0.$$

Ricordando che $\nabla f(x_2) = \nabla f(x_1) + \alpha_1 A v_1$, abbiamo

$$\nabla f(x_2)^T v_0 = \nabla f(x_1)^T v_0 + \alpha_1 (A v_1)^T v_0 = \alpha_1 v_0^T A v_1$$

e, dunque, x_2 minimizza la funzione nello spazio affine

$$A = \{x : x = x_0 + \alpha_0 v_0 + \alpha_1 v_1, \ \ \alpha_0, \alpha_1 \in \Re\}$$

se e solo se **le direzioni di ricerca** v_0 **e** v_1 **sono** A**-coniugate, ovvero** $v_0^T A v_1 = 0$.

Il metodo delle direzioni coniugate

Proposizione [Coniugatezza e indipendenza lineare]

Se i vettori $v_1, v_2, ..., v_k$ sono **A-coniugati**, allora essi sono anche **linearmente indipendenti**

Metodo delle direzioni coniugate: metodo di discesa con "passo ottimo" in cui le direzioni di ricerca sono mutuamente coniugate.

:-(Necessità di determinare delle direzioni coniugate

:-) Il metodo termina dopo un numero finito di passi (al piú n)

Metodo delle direzioni coniugate

$v_0, v_1, ..., v_{n-1}$ vettori non nulli A-coniugati
$x_0 \in \Re^n$; $g_0 \leftarrow Ax_0 - b$; $k \leftarrow 0$;
while $k < n$ ***and*** $\|g_k\| > tol$
$\alpha_k = -\frac{v_k^T g_k}{v_k^T A v_k}$;
$x_{k+1} \leftarrow x_k + \alpha_k v_k$;
$g_{k+1} \leftarrow g_k + \alpha_k A v_k$;
$k \leftarrow k + 1$;
endwhile

Come scegliere le direzioni v_k?

Metodo del Gradiente Coniugato

$x_0 \in \Re^n$; $g_0 \leftarrow Ax_0 - b$; $v_0 = -g_0$; $k \leftarrow 0$;
while $k < n$ ***and*** $\|g_k\| > tol$
$\alpha_k = -\frac{v_k^T g_k}{v_k^T A v_k}$;
$x_{k+1} \leftarrow x_k + \alpha_k v_k$;
$g_{k+1} \leftarrow g_k + \alpha_k A v_k$;
$\beta_k \leftarrow ?$;
$v_{k+1} = -g_{k+1} + \beta_k v_k$;
$k \leftarrow k + 1$;
endwhile

La direzione di ricerca all'iterazione k è costruita come combinazione lineare dell'antigradiente in tale iterazione e della direzione di ricerca all'iterazione precedente.

Il Metodo del gradiente coniugato: proprietà

Siano $v_0, v_1, ..., v_k$ le direzioni A-coniugate generate dall'algoritmo CG e siano

$$x_i = x_0 + v^*, \text{ con } v^* = \operatorname{argmin} f(x_0 + v),\ v \in \operatorname{Span}\{v_0, v_1, ..., v_{i-1}\},\ \ i \leq k+1,$$

le corrispondenti approssimazioni della soluzione. Quale deve essere il valore di β_k affinché la direzione

$$v_{k+1} = -g_{k+1} + \beta_k v_k$$

sia A-coniugata alle precedenti?

Essendo $g_i = \beta_{i-1} v_{i-1} - v_i$, si ha

$$\operatorname{Span}\{g_0, g_1, ..., g_i\} = \operatorname{Span}\{v_0, v_1, ..., v_i\}\ \ \forall i \leq k$$

e, ricordando la formula ricorsiva che lega i gradienti, si ha anche

$$\operatorname{Span}\{v_0, v_1, ..., v_i\} = \operatorname{Span}\{g_0, g_1, ..., g_i\} = \operatorname{Span}\{g_0, Ag_0, ..., A^i g_0\}\ \ \forall i \leq k.$$

Infine, in virtú della coniugatezza delle direzioni e della ottimalità del passo, risulta

$$g_i \perp v_j,\ \ j = 1, ..., i-1 \ \text{ e quindi } \ g_i \perp g_j,\ \ j = 0, ..., i-1.$$

Il Metodo del gradiente coniugato: proprietà

Siano $v_0, v_1, ..., v_k$ le direzioni A-coniugate generate dall'algoritmo CG e siano

$$x_i = x_0 + v^*, \text{ con } v^* = \operatorname{argmin} f(x_0 + v),\ v \in \operatorname{Span}\{v_0, v_1, ..., v_{i-1}\},\ \ i \leq k+1,$$

le corrispondenti approssimazioni della soluzione. Quale deve essere il valore di β_k affinché la direzione

$$v_{k+1} = -g_{k+1} + \beta_k v_k$$

sia A-coniugata alle precedenti?

Essendo $g_i = \beta_{i-1} v_{i-1} - v_i$, si ha

$$\operatorname{Span}\{g_0, g_1, ..., g_i\} = \operatorname{Span}\{v_0, v_1, ..., v_i\}\ \ \forall i \leq k$$

e, ricordando la formula ricorsiva che lega i gradienti, si ha anche

$$\operatorname{Span}\{v_0, v_1, ..., v_i\} = \operatorname{Span}\{g_0, g_1, ..., g_i\} = \operatorname{Span}\{g_0, Ag_0, ..., A^i g_0\}\ \ \forall i \leq k.$$

Infine, in virtú della coniugatezza delle direzioni e della ottimalità del passo, risulta

$$g_i \perp v_j,\ \ j = 1, ..., i-1 \ \text{ e quindi } \ g_i \perp g_j,\ \ j = 0, ..., i-1.$$

Il metodo del Gradiente Coniugato: proprietà

Da $v_{k+1} = -g_{k+1} + \beta_k v_k$ discende

$$v_{k+1}^T A v_i = -g_{k+1}^T A v_i + \beta_k v_k A v_i, \quad i = 1, ..., k.$$

Inoltre, per $i \leq k$,

$$0 = g_{k+1}^T g_i = g_{k+1}^T (g_{i-1} + \alpha_{i-1} A v_{i-1}) \quad \Rightarrow \quad g_{k+1}^T A v_{i-1} = 0$$

Sfruttando la coniugatezza di $v_0, v_1, ..., v_k$ si ha quindi

$$v_{k+1}^T A v_i = 0, \quad i = 0, 1, , ..., k-1, \tag{2}$$

$$v_{k+1}^T A v_k = -g_{k+1}^T A v_k + \beta_k v_k A v_k. \tag{3}$$

Da (2) segue che, per qualsiasi scelta di β_k, la direzione v_{k+1} è coniugata a $v_0, v_1, ..., v_{k-1}$, mentre da (3) segue che affinché v_{k+1} sia coniugata anche a v_k deve essere

$$\beta_k = \frac{g_{k+1}^T A v_k}{v_k A v_k}.$$

figuracg.jpg

Il metodo del Gradiente Coniugato: proprietà (riepilogo)

Abbiamo dimostrato che il metodo del Gradiente Coniugato

- genera una sequenza di direzioni coniugate

$$v_k = -g_k + \beta_{k-1} v_{k-1},$$

  con $\beta_k = (g_{k+1}^T A v_k)/(v_k^T A v_k)$;

- costruisce una sequenza di punti

$$x_k = x_0 + v^*, \text{ con } v^* = \operatorname{argmin} f(x_0 + v),\ v \in \operatorname{Span}\{v_0, v_1, ..., v_{k-1}\};$$

- converge in al piú n passi;

- è tale che

$$\operatorname{Span}\{v_0, v_1, ..., v_k\} = \operatorname{Span}\{g_0, g_1, ..., g_k\} = \operatorname{Span}\{g_0, Ag_0, ..., A^k g_0\}$$

Il metodo del Gradiente Coniugato: espressioni di α_k e β_k

- Da $v_k = -g_k + \beta_{k-1} v_{k-1}$ e $v_{k-1}^T g_k = 0$ discende

$$\alpha_k = -\frac{v_k^T g_k}{v_k^T A v_k} = \frac{g_k^T g_k}{v_k^T A v_k}.$$

- Da $Av_k = (g_{k+1} - g_k)/\alpha_k$ discende

$$\beta_k = \frac{g_{k+1}^T A v_k}{v_k^T A v_k} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}.$$

Metodo del Gradiente Coniugato

$x_0 \in \Re^n$; $g_0 \leftarrow Ax_0 - b$; $v_0 = -g_0$; $k \leftarrow 0$;
while $k < n$ ***and*** $\|g_k\| > tol$

$\alpha_k \leftarrow \frac{g_k^T g_k}{v_k^T A v_k}$ $\boxed{-\frac{v_k^T g_k}{v_k^T A v_k}}$;

$x_{k+1} \leftarrow x_k + \alpha_k v_k$;

$g_{k+1} \leftarrow g_k + \alpha_k A v_k$;

$\beta_k \leftarrow \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$ $\boxed{\frac{g_{k+1}^T A v_k}{v_k^T A v_k}}$;

$v_{k+1} \leftarrow -g_{k+1} + \beta_k v_k$;

$k \leftarrow k + 1$;

endwhile

Complessità computazionale per iterazione

- Un solo prodotto matrice vettore per iterazione (Av_k);
- Il prodotto Ax_k non viene mai eseguito;
- memoria di lavoro: 4 vettori (x, v, Av, e r)

Il metodo del Gradiente Coniugato: convergenza

Convergenza e autovalori distinti

Se m è il numero di autovalori distinti di A, allora il metodo CG converge alla soluzione in un numero di passi $k \leq m$.

Dimostrazione. Siano $\lambda_1, \lambda_2, ..., \lambda_m$ gli autovalori distinti di $A = UDU^T$. Supponiamo per assurdo che i vettori $g_0, g_1, ..., g_m$ siano tutti non nulli. Osserviamo inoltre che

$$\prod_{i=1}^{m}(A - \lambda_i I) = U\left(\prod_{i=1}^{m}(D - \lambda_i I)\right)U^T = 0.$$

Allora le matrici

$$I = A^0, A, A^2, ..., A^m$$

sono linearmente dipendenti, e tali sono i vettori

$$g_0, Ag_0, A^2 g_0, ..., A^m g_0.$$

Ma

$$\mathrm{Span}\{g_0, Ag_0, A^2 g_0, ..., A^m g_0\} = \mathrm{Span}\{g_0, g_1, g_2, ..., g_m\},$$

e ciò è in contrasto l'ipotesi di assurdo.

Il metodo del Gradiente Coniugato: convergenza

Convergenza e autovalori distinti

Se m è il numero di autovalori distinti di A, allora il metodo CG converge alla soluzione in un numero di passi $k \leq m$.

Dimostrazione. Siano $\lambda_1, \lambda_2, ..., \lambda_m$ gli autovalori distinti di $A = UDU^T$. Supponiamo per assurdo che i vettori $g_0, g_1, ..., g_m$ siano tutti non nulli. Osserviamo inoltre che

$$\prod_{i=1}^{m}(A - \lambda_i I) = U \left(\prod_{i=1}^{m}(D - \lambda_i I) \right) U^T = 0.$$

Allora le matrici

$$I = A^0, A, A^2, ..., A^m$$

sono linearmente dipendenti, e tali sono i vettori

$$g_0, Ag_0, A^2 g_0, ..., A^m g_0.$$

Ma

$$\text{Span}\{g_0, Ag_0, A^2 g_0, ..., A^m g_0\} = \text{Span}\{g_0, g_1, g_2, ..., g_m\},$$

e ciò è in contrasto l'ipotesi di assurdo.

il Metodo del gradiente coniugato: errore assoluto

Definiamo errore assoluto al passo k la quantità

$$e_k = x_k - x^*.$$

Essendo $Ax^* = b$, si ha

$$e_k^T A v_i = (Ax_k - Ax^*)^T v_i = g_k^T v_i = 0, \quad i = 0, 1, ..., k-1$$

(e_k è A-ortogonale (coniugato) alle prime k direzioni di ricerca)

Inoltre, essendo $Ae_0 = g_0$, si ha

$$\text{Span}\{g_0, g_1, ..., g_k\} = \text{Span}\{v_0, v_1, ..., v_k\} = \text{Span}\{Ae_0, A^2 e_0, ..., A^{k+1} e_0\}$$

Il metodo del gradiente coniugato: errore assoluto

Si osservi che $x_{k+1} \in X_k$, con

$$X_k = \{x : x = x_0 + v\}, \quad v \in \text{Span}\{v_0, v_1, ..., v_k\},$$

e quindi $e_{k+1} \in E_k$, con

$$E_k = \{e : e = e_0 + v\}, \quad v \in \text{Span}\{v_0, v_1, ..., v_k\} = \text{Span}\{Ae_0, A^2e_0, ..., A^{k+1}e_0\}.$$

Esistono quindi dei coefficienti reali $\gamma_1, \gamma_2, ..., \gamma_{k+1}$ tali che

$$\begin{aligned} e_{k+1} &= e_0 + \sum_{i=1}^{k+1} \gamma_i A^i e_0 \\ &= \left(1 + \sum_{i=1}^{k+1} \gamma_i A^i\right) e_0 = p_{k+1}(A)e_0, \end{aligned}$$

dove $p_{k+1} \in \Pi^1_{k+1}$ e Π^1_{k+1} è l'insieme dei polinomi di grado $k+1$ che valgono 1 nell'origine.

Il metodo del gradiente coniugato: errore assoluto

Caratterizzazione dell'errore assoluto

L'errore assoluto e_{k+1} risolve il problema

$$\begin{aligned} &\min\ u^T A u \\ &s.t.\ \ u \in U = \{u\ :\ u = e_0 + s,\ s \in S\} \\ &\text{con } S = \text{Span}\{v_0, v_1, ..., v_k\} = \text{Span}\{Ae_0, A^2 e_0, ..., A^{k+1} e_0\} \end{aligned} \tag{4}$$

Dimostrazione. Il problema (4) può essere riformulato come

$$\begin{aligned} &\min\ (e_0 + V\alpha)^T A (e_0 + V\alpha) \\ &\text{con } V = [v_0, v_1, ..., v_k] \text{ e } \alpha \in \Re^{k+1} \end{aligned}$$

o, equivalentemente,

$$\min\ \alpha^T V^T A V \alpha + 2 e_0^T A V \alpha$$

(minimizzazione di una funzione convessa).

Il metodo del gradiente coniugato: errore assoluto

Caratterizzazione dell'errore assoluto

L'errore assoluto e_{k+1} risolve il problema

$$\begin{aligned} &\min\ u^T A u \\ &s.t.\ \ u \in U = \{u\ :\ u = e_0 + s,\ s \in S\} \\ &\text{con } S = \text{Span}\{v_0, v_1, ..., v_k\} = \text{Span}\{Ae_0, A^2 e_0, ..., A^{k+1} e_0\} \end{aligned} \tag{4}$$

Dimostrazione. Il problema (4) può essere riformulato come

$$\begin{aligned} &\min\ (e_0 + V\alpha)^T A (e_0 + V\alpha) \\ &\text{con}\ \ V = [v_0, v_1, ..., v_k]\ \ \text{e}\ \ \alpha \in \Re^{k+1} \end{aligned}$$

o, equivalentemente,

$$\min\ \alpha^T V^T A V \alpha + 2 e_0^T A V \alpha$$

(minimizzazione di una funzione convessa).

Il metodo del gradiente coniugato: errore assoluto

Scrivendo le condizioni di ottimo si ha:

$$\begin{aligned} V^T AV\alpha + V^T Ae_0 = 0 \quad &\Leftrightarrow \quad V^T A(V\alpha + e_0) = 0 \quad \Leftrightarrow \\ V^T Au = 0 \quad &\Leftrightarrow \quad v_i^T Au = 0, \;\; i = 0, 1, ..., k \end{aligned}$$

L'ultima relazione è proprio la condizione di A-ortogonalità tra e_{k+1} e le direzioni di ricerca $v_0, ..., v_k$ e quindi vale la tesi.

La proposizione precedente può essere anche formulata come segue:

Caratterizzazione dell'errore assoluto

$$\begin{aligned} \|e_{k+1}\|_A &= \min_{u \in U} \|u\|_A = \min_{\gamma \in R^{k+1}} \|e_0 + \sum_1^{k+1} \gamma_i A^i e_0\| \\ &= \min \|p(A)e_0\|_A, \;\; p \in \Pi_{k+1}^1. \end{aligned}$$

Il metodo del gradiente coniugato: errore e spettro di A

Si ha ($u_1, u_2, ..., u_n$ base ortonormale di autovettori di A)

$$\|p(A)(x_0 - x^*)\|_A^2 = (x_0 - x^*)^T A \left(p(A)\right)^2 (x_0 - x^*) =$$

$$\left(\sum_1^n \gamma_i u_i\right)^T A \left(p(A)\right)^2 \left(\sum_1^n \gamma_i u_i\right) = \left(\sum_1^n \gamma_i u_i\right)^T \sum_1^n \gamma_i A \left(p(A)\right)^2 u_i =$$

$$\left(\sum_1^n \gamma_i u_i\right)^T \sum_1^n \gamma_i A \left(p(\lambda_i)\right)^2 u_i = \left(\sum_1^n \gamma_i u_i\right)^T \sum_1^n \gamma_i \lambda_i \left(p(\lambda_i)\right)^2 u_i =$$

$$\sum_1^n \gamma_i^2 \lambda_i \left(p(\lambda_i)\right)^2 \leq \max_i \left(p(\lambda_i)\right)^2 \sum_1^n \gamma_i^2 \lambda_i = \max_i \left(p(\lambda_i)\right)^2 \|x_0 - x^*\|_A^2$$

Stima dell'errore relativo

$$\frac{\|e_{k+1}\|_A^2}{\|e_0\|_A^2} \leq \min_{p \in \Pi_{k+1}^1} \max_{1 \leq i \leq n} \left(p(\lambda_i)\right)^2$$

Il metodo del gradiente coniugato: errore e spettro di A

Si ha ($u_1, u_2, ..., u_n$ base ortonormale di autovettori di A)

$$\|p(A)(x_0 - x^*)\|_A^2 = (x_0 - x^*)^T A \left(p(A)\right)^2 (x_0 - x^*) =$$

$$\left(\sum_1^n \gamma_i u_i\right)^T A \left(p(A)\right)^2 \left(\sum_1^n \gamma_i u_i\right) = \left(\sum_1^n \gamma_i u_i\right)^T \sum_1^n \gamma_i A \left(p(A)\right)^2 u_i =$$

$$\left(\sum_1^n \gamma_i u_i\right)^T \sum_1^n \gamma_i A \left(p(\lambda_i)\right)^2 u_i = \left(\sum_1^n \gamma_i u_i\right)^T \sum_1^n \gamma_i \lambda_i \left(p(\lambda_i)\right)^2 u_i =$$

$$\sum_1^n \gamma_i^2 \lambda_i \left(p(\lambda_i)\right)^2 \leq \max_i \left(p(\lambda_i)\right)^2 \sum_1^n \gamma_i^2 \lambda_i = \max_i \left(p(\lambda_i)\right)^2 \|x_0 - x^*\|_A^2$$

Stima dell'errore relativo

$$\frac{\|e_{k+1}\|_A^2}{\|e_0\|_A^2} \leq \min_{p \in \Pi_{k+1}^1} \max_{1 \leq i \leq n} \left(p(\lambda_i)\right)^2$$

I polinomi di Chebyshev

$$\cos[(k+1)\theta] = 2\cos\theta\cos k\theta - \cos[(k-1)\theta].$$
$$\cosh[(k+1)\theta] = 2\cosh\theta\cosh k\theta - \cosh[(k-1)\theta],$$
$$\cosh x = \cos(ix) = \frac{e^x + e^{-x}}{2}, \quad \cosh^{-1} x = \ln(x + \sqrt{x^2-1}), \quad i = \sqrt{-1}.$$

Il polinomio di Chebyshev di prima specie di grado k si può definire come segue:

$$T_k(\cosh\theta) = \cosh(k\theta), \quad \theta = \cosh^{-1} x$$
$$\Leftrightarrow \quad T_k(x) = \cosh(k\cosh^{-1} x) = \cosh(k\ln(x + \sqrt{x^2-1}))$$
$$\Leftrightarrow \quad T_k(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\right).$$

$$T_0(x) = 1,\ T_1(x) = x \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Teorema

Sia

$$Q(x) = T_k\left(\frac{2x-(b+a)}{b-a}\right) / T_k\left(\frac{-(b+a)}{b-a}\right).$$

Allora

$$\min_{p\in\Pi_k^1} \max_{x\in[a,b]} |p(x)| = \max_{x\in[a,b]} |Q(x)| = \left|T_k\left(\frac{-(b+a)}{b-a}\right)\right|^{-1}.$$

I polinomi di Chebyshev

$$\cos[(k+1)\theta] = 2\cos\theta\cos k\theta - \cos[(k-1)\theta].$$
$$\cosh[(k+1)\theta] = 2\cosh\theta\cosh k\theta - \cosh[(k-1)\theta],$$
$$\cosh x = \cos(ix) = \frac{e^x + e^{-x}}{2}, \quad \cosh^{-1} x = \ln(x + \sqrt{x^2-1}), \quad i = \sqrt{-1}.$$

Il polinomio di Chebyshev di prima specie di grado k si può definire come segue:

$$T_k(\cosh\theta) = \cosh(k\theta), \quad \theta = \cosh^{-1} x$$

$$\Leftrightarrow \quad T_k(x) = \cosh(k\cosh^{-1} x) = \cosh(k\ln(x + \sqrt{x^2-1}))$$

$$\Leftrightarrow \quad T_k(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\right).$$

$$T_0(x) = 1,\ T_1(x) = x \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Teorema

Sia

$$Q(x) = T_k\left(\frac{2x-(b+a)}{b-a}\right) / T_k\left(\frac{-(b+a)}{b-a}\right).$$

Allora

$$\min_{p\in\Pi_k^1} \max_{x\in[a,b]} |p(x)| = \max_{x\in[a,b]} |Q(x)| = \left|T_k\left(\frac{-(b+a)}{b-a}\right)\right|^{-1}.$$

I polinomi di Chebyshev

$$\cos[(k+1)\theta] = 2\cos\theta\cos k\theta - \cos[(k-1)\theta].$$
$$\cosh[(k+1)\theta] = 2\cosh\theta\cosh k\theta - \cosh[(k-1)\theta],$$
$$\cosh x = \cos(ix) = \frac{e^x + e^{-x}}{2}, \quad \cosh^{-1} x = \ln(x + \sqrt{x^2-1}), \quad i = \sqrt{-1}.$$

Il polinomio di Chebyshev di prima specie di grado k si può definire come segue:

$$T_k(\cosh\theta) = \cosh(k\theta), \quad \theta = \cosh^{-1} x$$
$$\Leftrightarrow \quad T_k(x) = \cosh(k\cosh^{-1} x) = \cosh(k\ln(x + \sqrt{x^2-1}))$$
$$\Leftrightarrow \quad T_k(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\right).$$

$$T_0(x) = 1,\ T_1(x) = x \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Teorema

Sia

$$Q(x) = T_k\left(\frac{2x-(b+a)}{b-a}\right) / T_k\left(\frac{-(b+a)}{b-a}\right).$$

Allora

$$\min_{p\in\Pi_k^1} \max_{x\in[a,b]} |p(x)| = \max_{x\in[a,b]} |Q(x)| = \left|T_k\left(\frac{-(b+a)}{b-a}\right)\right|^{-1}.$$

I polinomi di Chebyshev

$$\cos[(k+1)\theta] = 2\cos\theta\cos k\theta - \cos[(k-1)\theta].$$
$$\cosh[(k+1)\theta] = 2\cosh\theta\cosh k\theta - \cosh[(k-1)\theta],$$
$$\cosh x = \cos(ix) = \frac{e^x + e^{-x}}{2}, \quad \cosh^{-1} x = \ln(x + \sqrt{x^2-1}), \quad i = \sqrt{-1}.$$

Il polinomio di Chebyshev di prima specie di grado k si può definire come segue:

$$T_k(\cosh\theta) = \cosh(k\theta), \quad \theta = \cosh^{-1} x$$
$$\Leftrightarrow \quad T_k(x) = \cosh(k\cosh^{-1} x) = \cosh(k\ln(x + \sqrt{x^2-1}))$$
$$\Leftrightarrow \quad T_k(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\right).$$

$$T_0(x) = 1,\ T_1(x) = x \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Teorema

Sia

$$Q(x) = T_k\left(\frac{2x-(b+a)}{b-a}\right)/T_k\left(\frac{-(b+a)}{b-a}\right).$$

Allora

$$\min_{p\in\Pi_k^1}\max_{x\in[a,b]}|p(x)| = \max_{x\in[a,b]}|Q(x)| = \left|T_k\left(\frac{-(b+a)}{b-a}\right)\right|^{-1}.$$

Il metodo del gradiente coniugato: errore relativo e condizionamento

Consideriamo $T_k(s(\tau))$, con

$$s(\tau) = \frac{2\tau - (\lambda_1 + \lambda_n)}{\lambda_n - \lambda_1}$$

Osserviamo che $-1 \leq s(\tau) \leq 1 \Leftrightarrow \lambda_1 \leq \tau \leq \lambda_n$. Inoltre, considerato

$$s_0 = s(0) = -\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} = -\frac{\kappa + 1}{\kappa - 1}, \quad \text{con } \kappa = \frac{\lambda_n}{\lambda_1},$$

risulta

$$T_k(s_0) = \frac{(-1)^k}{2}\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right).$$

Il metodo del gradiente coniugato: errore relativo e condizionamento (cont.)

Teorema

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \frac{2}{\left(\frac{a-1}{a+1}\right)^k + \left(\frac{a+1}{a-1}\right)^k} \leq 2\left(\frac{a-1}{a+1}\right)^k, \quad \text{con} \quad a = \sqrt{\kappa(A)}$$

Dimostrazione. Consideriamo il polinomio

$$Q(\tau) = \frac{T_k(s(\tau))}{T_k(s_0)}.$$

Si ha:

$$\frac{\|e_k\|_A^2}{\|e_0\|_A^2} \leq \max_{1\leq i\leq n} |Q(\lambda_i)| \leq \max_{\lambda_1\leq\tau\leq\lambda_n} |Q(\tau)| = \frac{1}{|T_k(s_0)|},$$

da cui segue la prima disuguaglianza del teorema. Per dimostrare la seconda disuguaglianza basta osservare che

$$\frac{2}{\left(\frac{a-1}{a+1}\right)^k + \left(\frac{a+1}{a-1}\right)^k} = \frac{2\left(\frac{a-1}{a+1}\right)^k}{\left(\frac{a-1}{a+1}\right)^{2k} + 1}.$$

Il metodo del gradiente coniugato: cluster di autovalori

Il risultato del teorema precedente può essere raffinato, per tenere conto, in particolare, del caso in cui gli autovalori siano concentrati intorno a λ_1.

Teorema

Sia m un intero minore di n. Per il metodo CG vale la seguente maggiorazione per gli errori[a]:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 2\left(\frac{a-1}{a+1}\right)^{k-m}, \quad \text{con} \quad a = \sqrt{\kappa_m(A)} \text{ dove } \kappa_m(A) = \sqrt{\frac{\lambda_{n-m}}{\lambda_1}}$$

[a] *Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations*, T. Gergelits, · Z. Strakos

Esempio:
$\Lambda = \{1,\ 1.1,\ 1.2,\ 1.3,\ 1.4,\ 1.5,\ 1.6,\ 1.7,\ 1.8,\ 1000\}$, se consideriamo $m = 1$ avremo $\kappa_m(A) = \lambda_9/\lambda_1 = 1.8$ e quindi $\frac{a-1}{a+1} \simeq 0,146$

$\Lambda = \{1,\ 999.1,\ 999.2,\ 999.3,\ 999.4,\ 999.5,\ 999.6,\ 999.7,\ 999.8,\ 1000\}$, se consideriamo $m = 1$ avremo $\kappa_m(A) = \lambda_9/\lambda_1 = 999.8$ e quindi $\frac{a-1}{a+1} \simeq 0,939$

Si osservi che in entrambi gli esempi è $\kappa(A) = 1000$

Metodi di Krylov

$$Ax = b \ \text{(A non necessariamente sdp)} \tag{5}$$

Se A è non singolare la soluzione di (5) è $x^* = A^{-1}b$. A partire da un punto iniziale x_0 un **metodo iterativo di Krylov** calcola la generica iterazione $h-$ma minimizzando **un certo errore** nello spazio affine $h-$dimensionale:

$$x_0 + K_h, \ \text{ dove } \ K_h = \text{Span}\{r_0, Ar_0, ..., A^{h-1}r_0\}$$

dove $r_0 = Ax_0 - b$ (residuo iniziale)

- $K_h = K_h(A, r_0)$ $k-$mo spazio di Krylov;
- $r(x) = Ax - b$ residuo in x ($r_h = b - Ax_h$);
- Se A è non singolare, allora $K_n = \Re^n$, e quindi un metodo di Krylov termina dopo (al piú) n passi.

CG

Il metodo CG é un metodo di Krylov. L'iterazione x_k é soluzione del problema

$$\min_{x \in x_0 + K_k} \|x^* - x\|_A, \text{ dove } K_k = \text{Span}\{r_0, Ar_0, , ..., A^{k-1} r_0\} \tag{6}$$

Minimizzazione dell'errore

L'iterazione $k-$ma determina in $x_0 + K_k$ il punto che ha distanza minima dalla soluzione x^* secondo la norma indotta da A

GMRES

Nel metodo GMRES, l'iterazione x_k è soluzione del problema

$$\min_{x \in x_0 + K_k} \|b - Ax\|, \ dove \ K_k = \text{Span}\{r_0, Ar_0, ..., A^{k-1}r_0\}$$

Minimizzazione del residuo

Notiamo che, se $x \in x_0 + K_k$, allora

$$r(x) = -Ax + b = r_0 - \sum_0^k \gamma_i A^i r_0 = p(A)r_0, \quad \text{con} \quad p \in \pi_k^1$$

Teorema (dimostrazione: kelley)

Sia A non singolare, e sia x_k la $k-$ma iterazione generata da GMRES. Allora

$$\|r_k\| = \min_{p \in \pi_k^1} \|p(A)r_0\|$$

$$\|r_k\| = \min_{p \in \pi_k^1} \|p(A)r_0\| \Rightarrow \|r_k\| \leq \|r_0\| \min_{p \in \pi_k^1} \|p(A)\|$$

e quindi vale la seguente

Maggiorazione del $k-$mo residuo

$$\frac{\|r_k\|}{\|r_0\|} \leq \|p(A)\| \text{ per ogni } p \in \pi_k^1 \tag{7}$$

Necessitá di stime dell'errore piú "operative" di (7)(ad es., collegate allo spettro o al condizionamento di A). Occorre fare ulteriori ipotesi su A

Maggiorazione del $k-$mo residuo

$$\text{Se } \|I - A\| \leq \rho < 1, \text{ allora } \|g_k\| \leq \rho^k \|r_0\|$$

Dimostrazione: in (7) si prenda $p(A) = (I - A)^k$.

Matrici Diagonalizzabili

Una matrice A si dice **diagonalizzabile** se esiste una matrice non singolare $V \in \mathcal{C}^{n\times n}$ tale che

$$A = V^{-1}\Lambda V \Leftrightarrow VAV^{-1} = \Lambda$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 \\ 0 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 0 & \lambda_1 \end{pmatrix}$$

con V **matrice diagonale** costituita dagli autovalori di A.

1. $VA = \Lambda V$ (le colonne $k-$di V é autovettore destro di A di autovalore λ_k. Analogamente le righe di V^{-1} sono autovettori sinistri di A).
2. Dalla invertibilitá di V segue che esiste una base di autovettori d/s (**condizione necessaria e sufficiente per la diagonalizzazione**)
3. Se V è **ortogonale**, allora V ha come colonne gli autovettori di A, $V^{-1} = V^H$, con V^H complessa coniugata di V.
4. Se A è Hermitiana (simmetrica), gli autovettori possono essere scelti in maniera tale da formare una base ortonormale di $\mathcal{C}^n$ ($\Re^n$)

Matrici Diagonalizzabili: stima del residuo

Se $A = V^{-1}\Lambda V$, allora, per ogni polinomio $p(\cdot)$ si ha che

$$p(A) = V^{-1}p(\Lambda)V. \tag{8}$$

Da (7-8) segue che

$$\begin{aligned} \|r_k\| &= \min_{p\in\pi_k^1} \|p(A)r_0\| \leq \|V\|\|p(\Lambda)\|\|V^{-1}\|\|r_0\| \ \ \forall p \in \pi_k^1 &(9)\\ &\leq \kappa(V) \max_{i\leq n} |p(\lambda_i)|\|r_0\| \ \ \forall p \in \pi_k^1 &(10)\end{aligned}$$

Teorema (terminazione finita per GMRES (A diagonalizzable)

Se A possiede solo k autovalori distinti $\lambda_1, \lambda_2, ..., \lambda_k$, allora GMRES termina dopo (al piú) k passi.

Dimostrazione Il polinomio

$$p(x) = \prod_{i=1}^{k} \left(\frac{\lambda_i - x}{\lambda_i}\right)$$

è tale che $p(0) = 1$, $p(\Lambda) = 0$, e quindi, dalla (9) segue la tesi

GMRES (Generalized Minimal RESidual method)

Nel metodo GMRES, l'iterazione x_k é soluzione del problema

$$\min_{x \in x_0 + K_k} \|b - Ax\|, \ dove \ K_k = \mathrm{Span}\{r_0, Ar_0, ..., A^{k-1} r_0\} \tag{11}$$

Supponendo di conoscere una base ortonormale $\{u_1, u_2, ..u_k\}$ di K_k, allora il problema (11) potrebbe essere formulato come

$$\min \|b - A(x_0 + y_1 u_1 + y_2 u_2 + .. + y_k u_k)\| \tag{12}$$

o, equivalentemente,

$$\min \|b - A(x_0 + U_k y)\|, \ y \in \Re^k \Longleftrightarrow \min \|r_0 - A(U_k y)\|, \ y \in \Re^k \tag{13}$$

dove $U = [u_1, u_2, ..., u_k]$

GMRES: Costruzione di U_{k+1}

Costruzione iterativa dei vettori u_i.
Alla prima iterazione:

$$r_0 = Ax_0 - b; u_1 = r_0/\|r_0\| \tag{14}$$

Supponiamo di voler costruire u_2 tale che $\{u_1, u_2\}$ sia una base ortonormale di $\mathrm{Span}\{r_0, Ar_0\}$. Possiamo scegliere: $u_2 = Au_1 + \beta u_1$. Imponendo la condizione di ortonormalitá:

$$u_2^T u_1 = 0 \Leftrightarrow (Au_1 + \beta u_1)^T u_1 = 0 \Leftrightarrow \beta = -h^{11} \text{ dove } h^{11} = (Au_1)^T u_1 = u_1^T A u_1;$$

$$\text{e quindi } u_2 = \frac{Au_1 - h^{11}u_1}{\|Au_1 - h^{11}u_1\|}$$

Analogamente, volendo costruire u_3 tale che $\{u_1, u_2, u_3, \}$ sia una base ortonormale di $\mathrm{Span}\{g_0, Ag_0, A^2 g_0\}$:

$$u_3 = Au_2 + \alpha u_1 + \beta u_2;$$

$$u_1^T u_3 = 0 \Leftrightarrow \alpha = -u_1^T A u_2 = -h^{12};$$

$$u_2^T u_3 = 0 \Leftrightarrow \beta = -u_2^T A u_2 = -h^{22}; \text{e quindi } u_3 = \frac{Au_2 - h^{12}u_1 - h^{22}u_2}{\|Au_2 - h^{12}u_1 - h^{22}u_2\|}$$

GMRES: Algoritmo di Arnoldi

Algoritmo Arnoldi

1. $g_0 = Ax_0 - b; u_1 = g_0/\|g_0\|$
2. **for i=1:k**

$$u_{i+1} = \frac{Au_i - \sum_{j=1}^{i} h^{ji} u_j}{\|Au_i - \sum_{j=1}^{i} h^{ji} u_j\|} \quad \text{dove } h^{ji} = (Au_j)^T u_i \tag{15}$$

È immediato osservare che u_{i+1} è linearmente indipendente dai vettori $u_1, \ldots u_i$, appartiene a K_{i+1}; inoltre, per $k \leq i$

$$u_k^T \left(Au_i - \sum_{j=1}^{i} h^{ji} u_j \right) = u_k^T Au_i - \sum_{j=1}^{i} h^{ji} u_k^T u_j = u_k^T Au_i - h^{ki} u_k^T u_k = 0$$

e quindi i vettori $u_1, \ldots u_i, u_{i+1}$, rappresentano una base ortonormale di K_{i+1}

Algoritmo di Arnoldi: *The happy breakdown theorem*

Sia A non singolare, e supponiamo che

$$Au_i - \sum_{j=1}^{i} h^{ji} u_j = 0$$

Allora

$$x^* = A^{-1}b \in x_0 + K_i$$

Dimostrazione: Kelley (lemma 3.4.1)

Posto $\gamma_i = \|Au_i - \sum_{j=1}^{i} h^{ji} u_j\|$, considerata la matrice

$$P = \begin{bmatrix} h^{11} & \gamma_1 & & & & \\ h^{21} & h^{22} & \gamma_2 & & & \\ .. & .. & .. & .. & .. & .. \\ h^{n-1,1} & h^{n-1,2} & .. & .. & h^{n-1,n-1} & \gamma_{n-1} \end{bmatrix}, \tag{16}$$

se $H = P^T$, e H_k è la sottomatrice di H ottenuta prendendo le prime k colonne e $k+1$ righe,

$$H_k = \begin{bmatrix} h^{11} & h^{21} & .. & .. & h^{k1} \\ \gamma_1 & h^{22} & & & h^{k2} \\ & \gamma_2 & .. & .. & .. \\ & & .. & .. & h^{kk} \\ & & .. & & \gamma_k \end{bmatrix}, \tag{17}$$

allora, per le direzioni di ricerca generate dal metodo di Arnoldi vale la formula

$$\boxed{AU_k = U_{k+1} H_k}$$

Osserviamo che la matrice H_k é **Hessemberg superiore**.

GMRES: Algoritmo di Arnoldi

Supponiamo che l'algoritmo proceda senza *breakdown*. Ricordiamo che alla generica iterazione occorre risolvere il problema

$$\min \|g_0 - A(U_k)y\|$$

Essendo $U_{k+1}H_k = AU_k$ e osservando che $g_0/\|g_0\| = U_{k+1}e_1$, tale problema puó essere scritto come

$$\min \|g_0 - (AU_k)y\| \Leftrightarrow \min \|g_0 - (U_{k+1}H_k)y\| \Leftrightarrow \min \|U_{k+1}\left(\|g_0\|e_1 - H_k y\right)\|$$

Quindi, la generica iterazione $k-$ma puó essere calcolata come $x_k = x_0 + H_k y_k$, con

$$y_k = \operatorname{argmin} \|\|g_0\|e_1 - H_k y\|$$

GMRES: Algoritmo di Arnoldi

$$y_k = \operatorname{argmin} \|\|g_0\|e_1 - H_k y\|^2 = \operatorname{argmin} \frac{1}{2} y^T A_k y - b_k^T y$$

dove $A_k = H_k^T H_k$ e $b_k = H_k^T \|g_0\| e_1$, e quindi ad ogni passo del metodo di Arnoldi occorerá risolvere un sistema della forma $y = (H_k^T H_k)^{-1} H_k^T \|g_0\| e_1$

GMRES (Arnoldi)

$x_0 \in \Re^n$; $k \leftarrow 0$;

$r_0 \leftarrow b - Ax_0$, $u_1 = r_0/\|r_0\|$,

while `not_stopcondition`

$k = k + 1$;

for $j = 1...k$

$h_{jk} \leftarrow (Au_k)^T uj$

endfor

$u_{k+1} \leftarrow Au_k - \sum_{j=1}^{k} h_{jk} u_j$, $h_{k+1,k} \leftarrow \|u_{k+1}\|$

$u_{k+1} \leftarrow u_{k+1}/\|u_{k+1}\|$

$y_k \leftarrow (H_k^T H_k)^{-1} H_k^T \|g_0\| e_1$

endwhile

$x_{sol} = x_0 + U_k y_k$

GMRES: Algoritmo di Arnoldi

Complessità di tempo: all'iterazione $k-$ma

- Calcolo dei coefficienti h_{jk}: un **prodotto matvet** $(Av_k) + k\mathbf{vetvet}$;
- Risoluzione di un sistema $k \times k$ (**la matrice è fattorizzata come prodotto di matrici di Hessemberg**)
- altri due prodotti $\mathbf{vetvet}$

Al termine un **prodotto matvet** per ricostruire la soluzione

Complessità di spazio:
Occorre memorizzare la matrice H e tutte le direzioni di ricerca u_i

Inconvenienti numerici:

- Il fenomeno del *fill-in* rende il metodo poco adatto ai problemi sparsi; in particolare, dover conservare tutte le direzioni di ricerca pone seri problemi di occupazione di memoria (possibile rimedio: restart)
- L'accumulo dell'errore di roundoff puó produrre la perdita di ortogonalità fra le direzioni di ricerca (e il conseguente deterioramento delle proprietà di convergenza del metodo)(possibile rimedio: ortogonalizzazione di Gram-Schmidt modificata)

per i rimedi: vedi Kelley.

Precondizionamento - CG

Obiettivo: risolvere un sistema equivalente a quello di partenza, con un indice di condizionamento piú favorevole.
Strategia Determinare una matrice non singolare M, che sia una "approssimazione di A^{-1}", in modo tale che lo spettro di MA "si stringa" intorno a 1, riconducendo la risoluzione di $Ax = b$ alla risoluzione di

$$MAx = Mb \tag{18}$$

Nota: se A è spd, non è detto che AM lo sia (quindi (18) non è applicabile a CG). Si considera quindi il piú generale schema di precondizionamento

$$MANy = Mb \ \text{ con } \ y = N^{-1}x \tag{19}$$

con M e N precondizionatori sinistro e destro rispettivamente.
Nota: Se A è sdp e $N = M^T$, allora la matrice del sistema (19) sarà sdp

Precondizionamento CG - Jacobi

Metodo di Jacobi per $Ax = b$:

$$A = L + D + U$$
$$Ax = b \Leftrightarrow (L + D + U)x = b;\ Dx = b - (L + U)x$$
$$x_{k+1} = D^{-1}b - D^{-1}(L + U)x_k$$

Precondizionatore sinistro

$$M = (m_{ij}) : m_{ij} = \begin{cases} a_{ii} \text{ se } i = j \\ 0 \text{ altrimenti} \end{cases}$$

Precondizionamento CG - Cholesky incompleto

Si consideri una matrice B spd e sia $B = U^T U$ la sua fattorizzazione di Cholesky. Nel sistema precondizionato $MANy = Mb$ in (19)possiamo sceglire $M = U^{-T}, \;\; N = U^{-1}$.

Si osservi che, se $B = A$,

$$U^{-T} A U^{-1} y = U^{-T} b \Leftrightarrow U^{-T} U^T U U^{-1} y = U^{-T} b \Leftrightarrow$$
$$U^{-T} U^T U U^{-1} y = U^{-T} b \Leftrightarrow Iy = U^{-T} b \qquad (20)$$

La matrice M dovrà essere scelta

1. in modo da "approssimare bene" A ed avere condizionamento piú favorevole $(\kappa(U^{-T} A U^{-1}) \simeq 1$;
2. tale che il costo computazionale per calcolare U sia accettabile;
3. U sia facilmente "invertibile"

Teorema

Sia

$$Q(x) = T_k\left(\frac{2x-(b+a)}{b-a}\right)/T_k\left(\frac{-(b+a)}{b-a}\right)$$

Allora $\max|Q(x)|_{x\in[a,b]} = \min_{p\in\pi_k^1}\max|p(x)|_{x\in[a,b]}$

Metodi di Newton inesatti

Il metodo di Newton per il sistema di equazioni non lineare $F(x) = 0$

$$x_{k+1} = x_k + s_k \text{ dove}$$

$$F'(x_k)s_k = -F(x_k) \Leftrightarrow r_k = F'(x_k)s_k + F(x_k) = 0$$

Metodi Newton-inesatti: (44) é risolto in maniera approssimata con un **residuo**

$$r_k = F'(x_k)s_k + F(x_k)$$

(possibilmente) **non nullo** (e passo $s_k = F'(x_k)^{-1}r_k - F'(x_k)^{-1}F(x_k)$). Che accuratezza richiedere su r_k? Si considera, come "misura di accuratezza" il rapporto

$$\eta_k = \frac{\|r_k\|}{\|F(x_k)\|} \simeq \frac{\|F(x_{k+1})\|}{\|F(x_k)\|}$$

Metodo di **Newton**

$x_0 \in \Re^n$; $k \leftarrow 0$;
while *not_stopcondition*
  solve$F'(x_k)s = F(x_k)$;
  $x_{k+1} \leftarrow x_k + s$; $k \leftarrow k + 1$;
endwhile

Metodo di **Newton Inesatto**

$x_0 \in \Re^n$; $k \leftarrow 0$;
while *not_stopcondition*
  approxsolve$F'(x_k)s = F(x_k)$;
  $x_{k+1} \leftarrow x_k + s$; $k \leftarrow k + 1$;
endwhile

Teorema: Convergenza metodi NI

Supponiamo che $\{\eta_k\}$ sia tale che $\eta_k < t < 1$, allora esiste $\varepsilon > 0$ tale che, se $\|x_0 - x^*\| < \varepsilon$ la successione $\{x_k\}$ converge a x^* **q-linearmente** e, in particolare,

$$\|x_{k+1} - x^*\|_* < t\|x_k - x^*\|_* \text{ dove } \|y\|_* = \|F'(x^*)y\| \tag{21}$$

Osserviamo che, poiché $F'(x_*)$ è non singolare, per ogni $\gamma > 0$ esiste $\rho > 0$ tale che, se $\|x - x^*\| < \rho$ allora

$$\|F'(x) - F'(x^*)\| < \gamma$$
$$\|F'(x)^{-1} - F'(x^*)^{-1}\| < \gamma$$
$$\|F(x) - F'(x^*)(x - x_*)\| < \gamma\|x - x_*\|$$

e, di conseguenza

$$\|F(x_k)\| = \|\,[F(x_k) - F'(x^*)e_k] + F'(x^*)e_k\| \le 2\gamma\|e_k\|$$

Inoltre,

$$\frac{1}{\mu}\|y\| \le \|y\|_* \le \mu\|y\| \text{ dove } \mu = \max\{\|F'(x^*)\|, \|F'(x^*)^{-1}\|\}$$

$$F'(x^*)e_{k+1} = F'(x^*)(e_k + s_k) = F'(x^*)\left[e_k + F'(x_k)^{-1}r_k - F'(x_k)^{-1}F(x_k)\right]$$
$$= F'(x^*)F'(x_k)^{-1}\left[F'(x_k)e_k + r_k - F(x_k)\right] =$$
$$F'(x^*)F'(x_k)^{-1}\left[r_k + \underbrace{(F'(x_k) - F'(x^*))e_k} + \underbrace{(-F(x_k) + F'(x^*)e_k)}\right];$$

$$\begin{aligned}\|F'(x^*)F'(x_k)^{-1}\| &= \|F'(x^*)(F'(x_k)^{-1} - F'(x^*)^{-1}) + I\| \\ &\leq 1 + \|F'(x^*)\| \cdot \|F'(x_k)^{-1} - F'(x^*)^{-1}\| \leq 1 + \gamma\mu\end{aligned}$$

$$\begin{aligned}\|e_{k+1}\|_* &\leq (1+\gamma\mu)(\|r_k\| + \|F'(x_k) - F'(x^*)\|\|e_k\| + \|F(x_k) - F'(x^*)e_k\|) \\ &\leq (1+\gamma\mu)(\eta_k\|F(x_k)\| + \gamma\|e_k\| + \gamma\|e_k\|) \leq (1+\gamma\mu)(2\gamma\eta_k + 2\gamma)\|e_k\| \\ &\leq (1+\gamma\mu)(2\gamma\eta_k + 2\gamma)\mu\|e_k\|_*\end{aligned} \tag{22}$$

Lemma: (maggiorazione di $\|F\|$).

Se

$$\alpha = \max\{\|F'(x_*)\| + \frac{1}{2\|F(x_*)^{-1}\|}, 2\|F(x_*)^{-1}\|\}$$

allora

$$\frac{1}{\alpha}\|x - x^*\| \leq \|F(x)\| \leq \alpha\|x - x^*\|$$

Teorema: Velocità di Convergenza metodi NI

Supponiamo che la successione $\{x_k\}$ generata da un metodo NI converga a x^*. La convergenza è superlineare se e solo se

$$\|r_k\| = o(\|F(x_k)\|) \text{ per } k \to \infty$$

Dimostrazione: supponiamo $\|e_{k+1}\| = o(\|e_k\|)$

$$r_k = \underbrace{F'(x_k)s_k} + F(x_k) = \underbrace{F'(x_k)e_{k+1} - F'(x_k)e_k} + F(x_k) - F'(x^*)e_k + F'(x^*)e_k =$$

$$\left[F(x_k) - F'(x^*)e_k\right] + \left[(F'(x_k) - F'(x^*))e_k\right] + \left[(F'(x^*) + F'(x_k) - F'(x^*))e_{k+1}\right]$$

e quindi

$$\|r_k\| \leq o(\|e_k\|) + o(1)\|e_k\| + \left[\|F'(x^*)\| + o(1)\right] o(\|e_k\|)$$

e quindi $\|r_k\| = o(\|e_k\|) = o(\|F(x_k)\|) \;\; k \to \infty$

Se $\|r_k\| = o(\|F(x_k)\|)$, ricordando (22) avremo

$$\begin{aligned}\|e_{k+1}\|_* &\leq (1+\gamma\mu)(\|r_k\| + \|F'(x_k) - F'(x^*)\|\|e_k\| + \|F(x_k) - F'(x^*)e_k\|) \\ &= o(\|F(x_k)\|) + o(1)(\|e_k\|) + o(\|e_k\|) = o(\|e_k\|)\end{aligned}$$

Teorema: Velocità di Convergenza metodi NI

Supponiamo che la successione $\{x_k\}$ generata da un metodo NI converga a x^*. Se $\eta_k \to 0$, allora la convergenza è **q-superlineare**.

Il generico metodo di Newton inesatto per il sistema di equazioni non lineare $F(x) = 0$ puo' essere scritto nella forma

$$x_{k+1} = x_k + s_k \quad \text{dove} \tag{23}$$

$$H_k s_k = -F(x_k) \tag{24}$$

In questo caso

$$r_k = F'(x_k)\left(-H_k F(x_k)\right) + F(x_k) \quad \text{e quindi}$$

$$\|r_k\| = \|F'(x_k)\left(-H_k F(x_k)\right) + F(x_k)\| \leq \|I - F'(x_k)H_k\| \cdot \|F(x_k)\|$$

In particolare, la successione η_k è maggiorata dalla successione ρ_k dove $\rho_k = \|I - F'(x_k)H_k\|$

Teorema: Velocità di Convergenza metodi NI

Supponiamo che la successione $\{x_k\}$ generata da un metodo NI converga a x^*. Se $F'(x)$, è Lipschitz continua in x^* allora la convergenza è **q-quadratica**, cioè esiste $C > 0$ tale che

$$\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^2.$$

Metodo di **Newton Inesatto globalizzato**

$x_0 \in \Re^n$; $k \leftarrow 0$; $t \in (0,1)$

while $\|F(x_k)\| > tol$

trova $\eta_k \in [0,1)$ e s_k tali che

$\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\|$ (C1) &&

$\|F(x_k + s_k)\| \leq [1 - t(1-\eta_k)]\|F(x_k)\|$ (C2)

$k \leftarrow k+1$; $x_{k+1} = x_k + s_k$

endwhile

Notiamo che, posto

$$\text{pred}_k = \|F(x_k)\| - \|F(x_k) + F'(x_k)s_k\|, \quad \text{ared}_k = \|F(x_k)\| - \|F(x_k + s_k)\|$$

$$(\text{C1}) \Rightarrow \text{pred}_k \geq (1-\eta_k)\|F(x_k)\| \tag{25}$$

$$(\text{C2}) \Rightarrow \text{ared}_k \geq t(1-\eta_k)\|F(x_k)\| \tag{26}$$

e quindi la condizione su s_k implica

$$\text{ared}_k \geq t(\text{pred}_k)$$

che puó essere interpretata come una condizione di fedeltá del modello lineare di F ripetto ad F.

Metodo di **Newton Inesatto con *Backtracking***

$x_0 \in \Re^n$; $k \leftarrow 0$; $\eta_{max} \in [0, 1)$, $0 < \theta_{min} < \theta_{max} < 1$

while $\|F(x_k)\| > tol$

  $\eta_k \in [0, \eta_{max}]$ calcola s_k tale che

    $\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\|$

  while $\|F(x_k + s_k)\| > [1 - t(1 - \eta_k)]\|F(x_k)\|$

    scegli $\Theta \in [\Theta_{min}, \Theta_{max}]$

    $s_k = \Theta s_k$, $\eta_k = 1 - \Theta(1 - \eta_k)$

  endwhile

  $k \leftarrow k + 1$; $x_{k+1} = x_k + s_k$

endwhile

Teorema: Convergenza NIB

Supponiamo che la successione $\{x_k\}$ generata da un metodo NIB abbia un punto di accumulazione x^*, con $F'(x_*)$ invertibile, allora $F(x_*) = 0$ e $x_k \to x_*$. Inoltre, per k sufficientemente grande, i valori iniziali di s_k ed η_k saranno accettabili.

Se la successione è limitata allora (in alternativa):

- Converge ad una soluzione x_*
- Ammette punti di accumulazione con Jacobiano singolare.

Implementazione di NIB: scelta dei parametri

$$\eta_{max} = 0.9; \quad \theta_{min} = 0.1; \; \theta_{max} = 0.5; \quad t = 1e-4$$

Termini forzanti e convergenza:

- $\eta_k = \eta < 1 \;\Rightarrow\;$ convergenza **lineare**
- $\eta_k \to 0 \;\Rightarrow\;$ convergenza **superlineare**
- $\eta_k = O(\|F(x_k)\|) \;\Rightarrow\;$ convergenza **quadratica**

Una possibile scelta per i termini forzanti:

$\eta_k = \min\{\eta_{max}, \widetilde{\eta}_k\}$ dove

$$\widetilde{\eta}_k = \frac{|\|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})s_{k-1}\||}{\|F(x_{k-1})\|}$$

Metodo delle secanti

Il metodo di Newton per il sistema di equazioni non lineare $F(x) = 0$

$$x_{k+1} = x_k - (F'(x_k))^{-1} F(x_k) \tag{27}$$

è una immediata generalizzazione del metodo delle tangenti (Newton-Raphson) a funzioni di piú variabili (formalmente ottenuta sostituendo la derivata della funzione con lo Jacobiano del sistema). Nel caso di funzioni di una variabile, le varianti del metodo di Newton che non utilizzano le derivate (secanti, falsa posizione, ...)

$$x_{k+1} = x_k - \frac{b-a}{F(b) - F(a)} F(x_k) \tag{28}$$

sono formalmente ottenute da (27) sostituendo la derivata con una approssimazione alle differenze finite $\frac{F(b)-F(a)}{b-a}$ (con a, b opportunamente scelti).
Un tale procedimento non è generalizzabile a funzioni di piú variabili. Occorre stabilire delle opportune condizioni da imporre alla approssimazione B_k della matrice Jacobiana $F'(x_k)$

Condizione secante

$$B_k s_k = y_k, \quad \text{dove} \quad s_k = x_k - x_{k-1}, \; y_k = F(x_k) - F(x_{k-1}) \tag{29}$$

La condizione (29) ammette infinite soluzioni (che costituiscono uno spazio affine di dimensione $n^2 - n$)

Condizione secante: Broyden (good)

$$B_{k+1} = \text{argmin}_{X \in \Re^{n\times n}} \|X - B\|_F \quad \text{s.t.} Xs_k = y_k \ \text{ con } B = B_k \tag{30}$$

Il problema (30) nelle n^2 variabili x_{ij} puó essere formulato come [1]

$$\min \Phi(X) = XX - 2BX \ \ s.t. \ \ Xs = y, \tag{31}$$

o, equivalentemente,

$$\min \Phi(X) = 0.5(XX - 2BX) \ \ s.t. \ \ \Theta_i(X) = X^i s = y_i, \ i = 1..n, \tag{32}$$

dove X^i è la $i-$ma riga di X e gli indici di s_k, y_k sono omessi per semplificare le notazioni. Le condizioni di Lagrange per (32) sono

$$\nabla\Phi(X) = X - B = \sum_i^n \lambda_i \nabla\Theta_i(X) \Leftrightarrow X^i - B^i = \lambda_i s \Leftrightarrow X - B = \lambda s^T \tag{33}$$

Per ricavare il vettore dei parametri di Lagrange in (33), moltiplicando ambo i membri per s e ricordando il vincolo $Xs = y$ si ottiene

$$Xs - Bs = \lambda s^T s \Leftrightarrow y - Bs = \lambda s^T s \ \text{ da cui} \tag{34}$$

$$\lambda = \frac{y - Bs}{s^T s} \ \text{ e quindi } \ X = B + \lambda s^T = B + \frac{ys^T - (Bs)s^T}{s^T s} \tag{35}$$

[1] il prodotto matrice matrice va inteso come prodotto componente per componente: $C = AB, \ c_{ij} = a_{ij} b_{ij}$

Condizione secante: Broyden (good)

$$B_k = B_{k-1} + \frac{y_k s_k^T - (B_{k-1}s_k)s_k^T}{s_k^T s_k} = B_{k-1} + p_k s_k^T \quad \text{dove} \quad p_k = \frac{y_k - B_{k-1}s_k}{s_k^T s_k} \tag{36}$$

- Con la (36) all'iterazione k viene calcolata una approssimazione della matrice Jacobiana $F'(x_k)$ a partire da quella calcolata all'iterazione $k-1$ mediante un *update* di rango 1.
- La risoluzione di un sistema lineare è l'operazione piú onerosa per il calcolo del passo ad ogni iterazione;
- In generale la successione $\{B_k\}$ non converge a $F'(x^*)$;
- La formula (36) non preserva simmetria e pd.
- Convergenza superlineare: $\lim_{k\to\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = 0$

Metodo di **Broyden (diretto)** ***Good***

$x_0 \in \Re^n$; $B_0 \in \Re^{n\times n}$; $k \leftarrow 1$;
$F_0 = F(x_0)$; $s_1 \leftarrow -B_0^{-1}F_0$;
$x_1 \leftarrow x_0 + s_1$, $F_1 \leftarrow F(x_1)$;
$y_1 = F_1 - F_0$
while $\|F(x_k)\| > tol$
$B_k \leftarrow B_{k-1} + \frac{y_k s_k^T - (B_{k-1}s_k)s_k^T}{s_k^T s_k}$
$s_{k+1} \leftarrow -B_k^{-1}F_k$;
$x_{k+1} \leftarrow x_k + s_{k+1}$; $F_{k+1} = F(x_{k+1})$
$k \leftarrow k+1$; $y_k = F_k - F_{k-1}$
endwhile

Condizione secante: Broyden (bad)

Idea: fare una update di una **matrice che approssimi l'inversa della matrice Jacobiana**.

Formula di Sherman Morrison Woodbury

$$(A + vw^T)^{-1} = A^{-1} - \frac{A^{-1}vw^TA^{-1}}{1 + w^TA^{-1}v}$$

$$\left(A^{-1} - \frac{A^{-1}vw^TA^{-1}}{1 + w^TA^{-1}v}\right)(A + vw^T) = I + A^{-1}vw^T - \frac{A^{-1}vw^T + A^{-1}vw^TA^{-1}vw^T}{1 + w^TA^{-1}v} =$$

$$I + \frac{A^{-1}vw^T + A^{-1}vw^T\overbrace{w^TA^{-1}v} - A^{-1}vw^T - A^{-1}v\overbrace{w^TA^{-1}v}w^T}{1 + w^TA^{-1}v} = I$$

Utilizzando SMW con $B_{k-1} \to A,\ p_k \to v,\ s_k \to w,\ H = (B)^{-1}$ si ha

$$H_k = (B_k)^{-1} = (B_{k-1} + p_ks_k^T)^{-1} = H_{k-1} - \frac{H_{k-1}p_ks_k^TH_{k-1}}{1 + s_k^TH_{k-1}p_k} = \tag{37}$$

$$H_{k-1} - \frac{(H_{k-1}y_k - s_k)s_k^TH_{k-1}/s_k^Ts_k}{1 + \frac{s_kH_{k-1}y_k - s_k^Ts_k}{s_k^Ts_k}} = H_{k-1} + \frac{(s_k - H_{k-1}y_k)s_k^TH_{k-1}}{s_k^TH_{k-1}y_k} \tag{38}$$

Condizione secante: Broyden (bad)

$$H_k = H_{k-1} + \frac{(s_k - H_{k-1}y_k)s_k^T H_{k-1}}{s_k^T H_{k-1} y_k} \tag{39}$$

- Con la (39) all'iterazione k viene calcolata una approssimazione dell'inversa della matrice Jacobiana a partire da quella calcolata all'iterazione $k-1$ mediante un *update* di rango 1.
- Un prodotto matrice vettore è il calcolo piú oneroso richiesto per il calcolo del passo ad ogni iterazione;
- In generale la successione $\{H_k\}$ non converge a $F'(x^*)^{-1}$;
- La formula (39) non preserva simmetria e pd.
- Convergenza superlineare: $\lim_{k\to\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = 0$

Matrici di Broyden: invertibilitá

Matrix Determinant Lemma (MDL)

Siano $A \in \Re^{n\times n}$, $u, v \in \Re^n$, con A non singolare. Allora

$$\det\left(A + uv^T\right) = \left(1 + v^T A^{-1} u\right) \det(A)$$

Matrici di Broyden: invertibilitá

$$B_{k-1} \text{ invertibile} \quad \Rightarrow \quad B_k \text{ invertibile sse } s_k^T B_{k-1}^{-1} y_k \neq 0$$
$$H_{k-1} \text{ invertibile} \quad \Rightarrow \quad H_k \text{ invertibile sse } s_k^T H_{k-1} y_k \neq 0$$

Dimostrazione:

$$\det(B_k) = \det\left(B_{k-1} + \frac{y_k s_k^T - (B_{k-1}s_k)s_k^T}{s_k^T s_k}\right) = \det\left(B_{k-1} + p_k s_k^T\right) \quad \left[p_k = \frac{y_k - B_{k-1}s_k}{s_k^T s_k}\right]$$

$$\text{Applicando MDL si ha } \det(B_k) = \det(B_{k-1})\left(1 + s_k^T B_{k-1}^{-1} \frac{y_k - B_{k-1}s_k}{s_k^T s_k}\right)$$

$$= \det(B_{k-1})\left(1 + \frac{s_k^T B_{k-1}^{-1} y_k}{s_k^T s_k} - 1\right) = \det(B_{k-1}) \frac{s_k^T B_{k-1}^{-1} y_k}{s_k^T s_k}$$

Broyden per $F(x) = Ax - b$

$$B_k = B_{k-1} + \frac{y_k s_k^T - (B_{k-1}s_k)s_k^T}{s_k^T s_k} = B_{k-1} + p_k s_k^T \text{ dove } p_k = \frac{y_k - B_{k-1}s_k}{s_k^T s_k}$$

Se $F(x) = Ax - b$, allora $y_k = As_k$ e, posto $E_{k-1} = B_{k-1} - A$, si ha

$$p_k = \frac{(A - B_{k-1})s_k}{s_k^T s_k} = -\frac{E_{k-1}s_k}{s_k^T s_k}, \text{ e quindi } B_k = B_{k-1} - E_{k-1}\frac{s_k s_k^T}{s_k^T s_k}$$

$$E_k = E_{k-1}\left(I - \frac{s_k s_k^T}{s_k^T s_k}\right) \Rightarrow \|E_k\|_2 \leq \|E_{k-1}\|_2 \text{ essendo } \left\|I - \frac{s_k s_k^T}{s_k^T s_k}\right\|_2 = 1$$

Al crescere delle iterazioni, le matrici di Broyden "si avvicinano" ad A.

Convergenza dei Metodi di Broyden

Broyden: convergenza finita per $Ax = b$ (Gay 1979)

Nel caso in cui sia $F(x) = Ax - b$, il metodo di Broyden *Good* [Bad] **converge in al piú** $2n$ **iterazioni**, purché la matrice B_0 $[H_0]$ sia non singolare e durante le iterazioni sia soddisfatta la condizione $s_k^T B_{k-1}^{-1} y_k \neq 0$ $[s_k^T H_{k-1} y_k \neq 0]$

Broyden: convergenza per $F \in C^1$

Nel caso in cui $F(x)$ sia di classe C^1, x_* sia tale che $F(x_*) = 0$, $\det(F'(x_*)) \neq 0$ con $F'(x)$ Lipschitz continua in un intorno di x_*, e inoltre si verifichi $s_k^T B_{k-1}^{-1} y_k \neq 0$ $[s_k^T H_{k-1} y_k \neq 0]$
il metodo di Broyden *Good* [Bad] è localmente convergente intorno a x_*, con velocitá di convergenza superlineare.