Metodi per l'ottimizzazione non vincolata

$$(P) \quad \min\{ f(x) : x \in \mathbb{R}^n \} \qquad (f : \mathbb{R}^n \to \mathbb{R})$$

METODI ITERATIVI

$$x^0, x^1, x^2, \dots, x^k, \dots\dots$$

Ricerca dei punti stazionari: $\nabla f(x^*) = 0$ (ipotesi minimale: f diff.)

Convergenza
- $\exists \bar{k}$ t.c. $\nabla f(x^{\bar{k}}) = 0$ (FINITEZZA)
- $\exists \lim x^k = x^*$ e $\nabla f(x^*) = 0$
- Tutti i punti di accumulazione di $\{x^k\}$ sono stazionari
- Almeno un punto di accumulazione di $\{x^k\}$ è stazionario

- DI DISCESA: $f(x^0) > f(x^1) > f(x^2) > \dots > f(x^k) > \dots\dots$

$$[\ f(x^{k+1}) < f(x^k)\ ]$$

("discesa non monotona": $\exists m \in \mathbb{N}$ t.c. $f(x^{k+m}) < f(x^k)$)

- DI RICERCA MONODIMENSIONALE ('line search'):

$$x^{k+1} = x^k + \underbrace{t_k}_{\text{passo di spostamento}} \overbrace{d^k}^{\text{direzione di spostamento}} \qquad (t_k \in \mathbb{R},\ d^k \in \mathbb{R}^n)$$

Come scegliere t_k e d^k in modo t.c. il metodo sia di discesa?

$\bar{x}$ punto non stazionario, cioè $\nabla f(\bar{x}) \neq 0$

$$f(\bar{x}+td) - f(\bar{x}) = t\, \nabla f(\bar{x})^T d + r(td) \qquad \lim_{t \to 0} \frac{f(\bar{x}+td) - f(\bar{x})}{t} = \nabla f(\bar{x})^T d$$

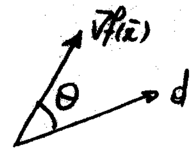
se $\nabla f(\bar{x})^T d < 0$, allora $f(\bar{x}+td) < f(\bar{x})$ se t è suff.nte piccolo

Per rendere (asintoticamente) massima la decrescita, conviene scegliere $d \in \mathbb{R}^n$ tale che $\nabla f(\bar{x})^T d$ sia minimo:

$$(P_d) \qquad \min \{ \nabla f(\bar{x})^T d \ : \ \|d\|_2 = 1 \}$$

Il vincolo $\|d\|_2 = 1$ è fondamentale, altrimenti $\nabla f(\bar{x})^T d < 0 \Rightarrow \nabla f(\bar{x})^T (td) \xrightarrow[t \to +\infty]{} -\infty$

$\nabla f(\bar{x})^T d = \|\nabla f(\bar{x})\|_2 \, \|d\|_2 \cos\theta$, dove θ è l'angolo compreso tra $\nabla f(\bar{x})$ e d (nota: per f e $\bar{x}$ fissati, $\theta = \theta(d)$)



$\nabla f(\bar{x})^T d = \|\nabla f(\bar{x})\|_2 \, \|d\|_2 \cos\theta = \|\nabla f(\bar{x})\|_2 \cos\theta \ \rightarrow$ (P_d) può essere riscritto nella forma $\|\nabla f(\bar{x})\|_2 \ \min \{ \cos\theta \ : \ \|d\|_2 = 1 \}$, il cui minimo si ottiene quando $\cos\theta = -1$, ovvero per $\theta = \pi \ \rightarrow$ da cui $\bar{d} = -\nabla f(\bar{x}) / \|\nabla f(\bar{x})\|$

["il gradiente è la direzione di massima crescita" / "l'opposto del gradiente è la direzione di massima decrescita"]

METODO DEL GRADIENTE

ipotesi: f differenziabile con continuità

$$d^k = -\nabla f(x^k) \ \rightsquigarrow \ x^{k+1} = x^k - t_k \nabla f(x^k)$$

Scelta (ideale) per il passo di spostamento

$$t_k \in \operatorname{argmin} \{ f(x^k - t \nabla f(x^k)) \ : \ t \geq 0 \} \qquad \text{RICERCA ESATTA}$$

$$\left(\text{ovvero } f(x^{k+1}) \leq f(x^k - t\nabla f(x^k)) \quad \forall t \geq 0 \right)$$

(in un generico algoritmo 'line search': $t_k \in \operatorname{argmin} \{ f(x^k + t d^k) : t \geq 0 \}$)

Metodo del Gradiente Esatto (più estesamente: "con ricerca esatta")

1) Scegliere $x^0 \in \mathbb{R}^n$; $k=0$
2) Se $\nabla f(x^k)=0$, allora STOP
3) Calcolare $t_k \in \operatorname{argmin}\{f(x^k - t\nabla f(x^k)) : t \geq 0\}$
4) $x^{k+1} = x^k - t_k \nabla f(x^k)$
5) $k=k+1$ e ritornare a 2)

Prop 1 Il metodo del gradiente esatto è un metodo di discesa

dim $\varphi(t) := f(x^k - t\nabla f(x^k))$

$f(x^k) = \varphi(0) \geq \varphi(t_k) = f(x^{k+1})$ [per def. di t_k]

Supponiamo $\varphi(0) = \varphi(t_k)$: $t=0$ punto di minimo di φ su $t \geq 0$

φ è derivabile con continuità

$\varphi'(t) = \nabla f(x^k - t\nabla f(x^k))^T(-\nabla f(x^k))$ [*]

$\varphi'(0) = -\|\nabla f(x^k)\|_2^2 < 0$ (a meno che x^k non sia stazionario)

quindi φ è localmente decrescente in 0 : $\varphi(t) < \varphi(0)$ $\forall t \in [0, \bar{t})$ contraddizione

Pertanto $\varphi(0) > \varphi(t_k)$ e quindi $f(x^k) > f(x^{k+1})$ ∎

Prop 2 Due direzioni successive del metodo del gradiente esatto sono ortogonali, ovvero $\nabla f(x^{k+1})^T \nabla f(x^k) = 0$

dim $0 = \varphi'(t_k) = \nabla f(x^{k+1})^T \nabla f(x^k)$ ∎

---

[*] $\phi : \mathbb{R} \to \mathbb{R}^n$ derivabile in $\bar{t} \in \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$ diff.bile in $\phi(\bar{t}) \in \mathbb{R}^n$ $\Rightarrow$ $g \circ \phi$ differenziabile in $\bar{t}$ e
$(g\circ\phi)'(\bar{t}) = \nabla g(\phi(\bar{t}))^T \dot{\phi}(\bar{t})$ (dove $\dot{\phi} = (\phi_1', \dots, \phi_n')$, $\phi = (\phi_1, \dots, \phi_n)$, $\phi_i : \mathbb{R} \to \mathbb{R}$)

Teorema (convergenza) Supponiamo che $\lim_{k\to+\infty} x^k = x^* \in \mathbb{R}^n$. Allora $\nabla f(x^*) = 0$

dim $0 = \nabla f(x^{k+1})^T \nabla f(x^k) \xrightarrow[x^k\to+\infty]{} \nabla f(x^*)^T \nabla f(x^*) = \|\nabla f(x)\|_2^2$

Quindi $\|\nabla f(x^*)\|_2 = 0$, da cui $\nabla f(x^*) = 0$ ∎

È possibile dimostrare che ogni punto di accumulazione di $\{x^k\}$ è stazionario.

La ricerca esatta nel caso quadratico

$f(x) = \frac{1}{2} x^T Q x + b^T x + c \quad , \quad Q = Q^T \in \mathbb{R}^{n\times n}, \ b \in \mathbb{R}^n, \ c \in \mathbb{R}$

$\nabla f(x) = Qx + b$.

Possiamo supporre che Q sia semidefinita positiva. Infatti, in caso contrario $\exists \bar{x} \in \mathbb{R}^n$ t.c. $\bar{x}^T Q \bar{x} < 0$ e $f(t\bar{x}) = \frac{1}{2} t^2 (\bar{x}^T Q \bar{x}) + t\, b^T \bar{x} + c \xrightarrow[t\to+\infty]{} -\infty$.

$\varphi(t) = f(x - t\nabla f(x))$, $\varphi'(t) = -\nabla f(x - t\nabla f(x))^T \nabla f(x) =$

(parabola se $\nabla f(x)^T Q \nabla f(x) > 0$)

$$
\begin{aligned}
&= -[Q(x - t\nabla f(x)) + b]^T \nabla f(x) = \\
&= -[Qx - tQ\nabla f(x) + b]^T \nabla f(x) = \\
&= -[\nabla f(x) - tQ\nabla f(x)]^T \nabla f(x) = \\
&= -\nabla f(x)^T \nabla f(x) + t\, \nabla f(x)^T Q \nabla f(x)
\end{aligned}
$$

$\varphi'(t) = 0 \iff t = \nabla f(x)^T \nabla f(x) / (\nabla f(x)^T Q \nabla f(x))$ se $\nabla f(x)^T Q \nabla f(x) > 0$
(sempre OK se Q definita positiva)

se invece $\nabla f(x)^T Q \nabla f(x) = 0$, allora

$\varphi'(t) \equiv -\nabla f(x)^T \nabla f(x) = -\|\nabla f(x)\|_2^2 < 0$ da cui

$f(x - t\nabla f(x)) = \varphi(t) = -\|\nabla f(x)\|_2^2\, t + \text{cost.} \xrightarrow[t\to+\infty]{} -\infty$: f inf.nte illimitata

Criticità del metodo del gradiente esatto

- Gli spostamenti sono ortogonali tra loro: crea "rigidità" negli spostamenti quando si è vicini ad un punto stazionario $\rightsquigarrow$ passi brevi e tante iterazioni

(Es: $f(x_1, x_2) = x_1^2 + 5x_2^2$)

La ricerca esatta nel caso quadratico (cont.)

Se Q è semidefinita positiva ma non definita positiva, non è detto che f sia inferiormente illimitata. Ad esempio se $b \equiv 0$, allora

$$\min \{ f(x) : x \in \mathbb{R}^n \} = c \quad \text{e} \quad \bar{x} = 0 \text{ è un punto di minimo.}$$

Più in generale, f ammette [un punto di] minimo (globale) se e solo se tutti gli autovettori relativi all'autovalore nullo di Q sono ortogonali al vettore b, ovvero $Qx = 0 \Rightarrow b^T x = 0$.

Esempio: $Q = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $c = 0$

Gli autovalori di Q sono $\lambda_{max} = 5$ e $\lambda_{min} = 0$ e i relativi autovettori sono $\bar{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ e $\hat{x} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Nota $b^T \hat{x} = 0$

$$x \in \mathbb{R}^2 \rightarrow \quad x = \alpha \bar{x} + \beta \hat{x} \quad \text{con } \alpha, \beta \in \mathbb{R}$$

$$f(x) = f(\alpha \bar{x} + \beta \hat{x}) = f(\alpha \bar{x}) = 25\alpha^2 + 5\alpha$$

$(Q\hat{x} = 0 \text{ e } b^T\hat{x} = 0)$

$$\min \{ f(x) : x \in \mathbb{R}^2 \} = \min \{ 25\alpha^2 + 5\alpha : \alpha \in \mathbb{R} \} = -1/4$$

$\bar{\alpha} = -1/10$ punto di minimo

$$x = \bar{\alpha} \bar{x} = \begin{bmatrix} -1/10 \\ -1/5 \end{bmatrix} \text{ punto di minimo di } f$$

Algoritmo 'finito' per il caso quadratico strettamente convesso

$f(x) = \frac{1}{2} x^T Q x + b^T x + c \qquad Q = Q^T \in \mathbb{R}^{n \times n}$ definita positiva, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$

Siano $\lambda_0, \lambda_1, \dots, \lambda_{n-1} > 0$ gli autovalori di Q

$$\boxed{x^{k+1} = x^k - \frac{1}{\lambda_k} \nabla f(x^k)} \qquad k = 0, 1, \dots, n-1$$

L'algoritmo individua un punto stazionario (e quindi di minimo) in al più n passi.

Prop Esiste $j \leq n$ tale che $\nabla f(x^j) = 0$

dim Supponiamo $\nabla f(x^j) \neq 0$ per ogni $j < n$.

$$\nabla f(x^n) = Q x^n + b = Q x^{n-1} - \frac{1}{\lambda_{n-1}} Q \nabla f(x^{n-1}) + b = \nabla f(x^{n-1}) - \frac{1}{\lambda_{n-1}} Q \nabla f(x^{n-1}) =$$

$$= \left(I - \frac{1}{\lambda_{n-1}} Q\right) \nabla f(x^{n-1}).$$

Iterando il procedimento $(n-1)$ volte si ottiene: $\nabla f(x^n) = \prod_{j=0}^{n-1} \left(I - \frac{1}{\lambda_j} Q\right) \nabla f(x^0)$

Siano $u_0, u_1, \dots, u_{n-1} \in \mathbb{R}^n$ autovettori relativi a $\lambda_0, \lambda_1, \dots, \lambda_{n-1}$ $(Q u_i = \lambda u_i)$ tali da costituire una base di $\mathbb{R}^n$. Allora $\nabla f(x^0) = \sum_{i=0}^{n-1} \alpha_i u_i$ per opportuni $\alpha_i \in \mathbb{R}$. Quindi:

$$\nabla f(x^n) = \left[\prod_{j=0}^{n-1} \left(I - \frac{1}{\lambda_j} Q\right)\right] \sum_{i=0}^{n-1} \alpha_i u_i = \sum_{i=0}^{n-1} \alpha_i \left(\prod_{j=0}^{n-1} \left(I - \frac{1}{\lambda_j} Q\right) u_i \right) =$$

$$= \sum_{i=0}^{n-1} \alpha_i \left[\prod_{j=0}^{n-1} \underbrace{\left(1 - \frac{1}{\lambda_j} \lambda_i\right)}_{=0 \ (j=i)} \right] u_i = 0$$

$\blacksquare$

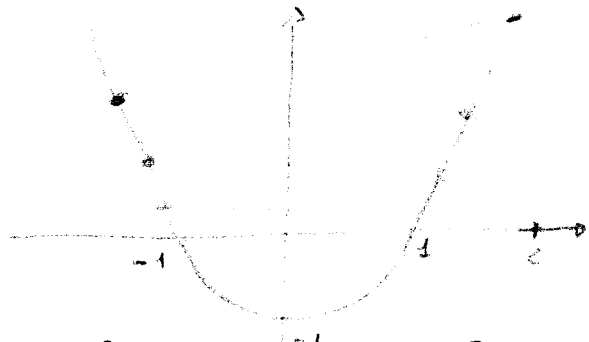
- La 'ricerca esatta' è computazionalmente onerosa: è un problema essa stessa di ottimizzazione ma monodimensionale $\leftarrow$ algoritmi richiedono criteri di STOP approx.
- La 'ricerca esatta' può condurre a passi molto brevi se la funzione ha curve di livello particolarmente "schiacciate" vicino ai punti stazionari (tipo la fz. "banana" di Rosenbrock $f(x_1,x_2) = 100(x_2 - x_1^2)^2 + (1-x_1)^2$)

Possibili Direzioni di Discesa

$$d^k = -D_k \nabla f(x^k) \qquad \text{con } D_k \in \mathbb{R}^{n\times n} \text{ definita positiva}$$

(d^k è una direzione di discesa: $\nabla f(x^k)^T d^k = -\nabla f(x^k)^T D_k \nabla f(x^k) < 0$)

Garantire che il metodo sia di discesa può non essere sufficiente:



$n = 1, \; f(x) = x^2 - 1$

$x^k = (-1)^k [2 - k/(k+1)], \; d^k = (-1)^{k+1}$

$f(x^k) = [2 - k/(k+1)]^2 - 1 \xrightarrow[k\to+\infty]{} 1$

$f(x^{k+1}) < f(x^k)$ ma $x^{2k+1} \to -1, \; x^{2k} \to 1$ e $f(-1) = f(1) = 0 > f(0) = -1$

Possibili Scelte per il Passo

1) Ricerca esatta: $t_k^{min} \in \operatorname{argmin} \{ f(x^k + t d^k) : t \geq 0 \}$
2) Minimizzazione limitata: $t_k \in \operatorname{argmin} \{ f(x^k + t d^k) : t \in [0, T] \}$, $T > 0$ fissato
3) Costante: $t_k = \bar{t}$ per qualche $\bar{t} > 0$
4) Passi decrescenti: $t_k \downarrow 0$ con $\sum_{k=0}^{\infty} t_k = +\infty$ (ad es. $t_k = 1/k$)

Versioni approssimate di 1) $\rightsquigarrow$ RICERCA INESATTA

Siano $x^k \in \mathbb{R}^n$ e $d^k \in \mathbb{R}^n$ t.c. $\nabla f(x^k)^T d^k < 0$

Sia $\varphi(t) = f(x^k + t d^k)$: $\varphi'(t) = \nabla f(x^k + t d^k)^T d^k$

Per forzare la diminuzione del valore di f a partire da x^k lungo d^k (ovvero del valore di $\varphi(t)$) ad essere di una certa consistenza, si può imporre la CONDIZIONE DI ARMIJO:

$$\text{(AJO)} \quad f(x^k + t d^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k \qquad c_1 \in (0,1) \text{ fissato}$$

- Se f è inf.nte limitata, (AJO) impedisce il caso $t \to +\infty$ (altrimenti si avrebbe $f(x^k + t d^k) \xrightarrow[t \to +\infty]{} -\infty$ in quanto $\nabla f(x^k)^T d^k < 0$).
- Poiché $\lim_{t \to 0} (f(x^k + t d^k) - f(x^k))/t = \varphi'(0) = \nabla f(x^k)^T d^k < 0$, dalla def. di limite segue che $\exists \bar{t}$ per cui (AJO) vale per ogni $t \in [0, \bar{t}]$ (si noti che $\nabla f(x^k)^T d^k < 0$ garantisce $\nabla f(x^k)^T d^k < c_1 \nabla f(x^k)^T d^k$)

Per evitare che il passo di spostamento sia eccessivamente piccolo (ciò è ammissibile per (AJO)), si può richiedere che $\varphi'(t)$ differisca sufficientemente da $\varphi'(0)$:

$$\text{(CUR)} \quad \nabla f(x^k + t d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k \qquad c_2 \in (c_1, 1) \text{ fissato}$$

Infatti $t \approx 0 \Rightarrow \nabla f(x^k + t d^k) \approx \nabla f(x^k)$ e quindi $\varphi'(t) = \nabla f(x^k + t d^k)^T d^k \simeq \varphi'(0) = \nabla f(x^k)^T d^k$

Questa coppia di condizioni è nota come CONDIZIONI DI WOLFE:

$$\text{(AJO)} \quad f(x^k + t_k d^k) \leq f(x^k) + c_1 t_k \nabla f(x^k)^T d^k \qquad c_1 \in (0,1) \text{ fissato}$$
$$\text{(CUR)} \quad \nabla f(x^k + t d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k \qquad c_2 \in (c_1, 1) \text{ fissato}$$

<u>Oss</u>: il passo fornito dalla ricerca esatta verifica le condizioni di Wolfe (CUR)

$\varphi'(t_k^{min}) = 0$ e $\varphi'(0) < 0 \Rightarrow$ (CUR)

Sia t suff.nte piccolo da soddisfare (AJO): $f(x^k + t_k^{min} d^k) = \varphi(t_k^{min}) \leq \varphi(t) =$

(t_k^{min} p.to di minimo)

$= f(x^k + t d^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k$ [redacted]

(t soddisfa (AJO)) ma ciò non garantisce (AJO) per t_k^{min} (accade nella "pratica")

Prop Supponiamo che f sia inf. nte limitata. Allora esiste un intervallo (a,b) per cui entrambe le condizioni di Wolfe valgono per ogni $t \in (a,b)$.

dim Sia $\tau_1 := \sup \{ \tau \mid \text{(AJO) vale per ogni } t \in [0,\tau] \}$: $\tau_1 > 0$

f inf. nte limitata $\Rightarrow \tau_1 < +\infty$. Per il teo (valor medio) $\exists\, \tau_2 \in (0, \tau_1)$

t.c. $f(x^k + \tau_1 d^k) - f(x^k) = \tau_1 \nabla f(x^k + \tau_2 d^k)^T d^k$

Inoltre $f(x^k + \tau_1 d^k) - f(x^k) = c_1 \tau_1 \nabla f(x^k)^T d^k$ (se fosse $<$, per la permanenza del segno τ_1 non sarebbe il max dell'insieme). Quindi:

$$\nabla f(x^k + \tau_2 d^k)^T d^k = c_1 \nabla f(x^k)^T d^k > c_2 \nabla f(x^k)^T d^k \quad (c_2 > c_1 \text{ e } \nabla f(x^k)^T d^k < 0)$$

Poiché ∇f è continua, la disuguaglianza vale in un intorno I_1 di τ_2

$\tau_2 \in [0, \tau_1) \Rightarrow$ (AJO) vale in un intorno I_2 di τ_2

$\rightarrow I_1 \cap I_2$ è l'intervallo cercato

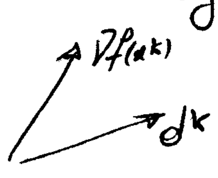
■

METODI DEL 'GRADIENTE' CON RICERCA INESATTA

1) Scegliere $x^0 \in \mathbb{R}^n$; $k = 0$
2) Se $\nabla f(x^k) = 0$, allora STOP
3) Scegliere $d^k \in \mathbb{R}^n$ tale che $\nabla f(x^k)^T d^k < 0$
4) Calcolare $t_k > 0$ che soddisfa (AJO) e (CUR) (condizioni di Wolfe)
5) $x^{k+1} = x^k + t_k d^k$
6) $k = k+1$ e ritornare a 2)

3) + (AJO) $\Downarrow$ metodo di discesa

Siano $\{x^k\}$, $\{t_k\}$ e $\{d^k\}$ le successioni generate dall'algoritmo e sia θ_k l'angolo compreso tra $\nabla f(x^k)$ e d^k:



$$\nabla f(x^k)^T d^k = \| \nabla f(x^k) \|_2 \, \| d^k \|_2 \cos \theta_k$$

Lemma Supponiamo che

- f sia inf.nte limitata ($\exists M \in \mathbb{R}$ t.c. $f(x) \geq M \ \forall x \in \mathbb{R}^n$)
- ∇f sia Lipschitziana, ovvero $\exists L > 0$ tale che

$$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2 \quad \forall x, y \in \mathbb{R}^n$$

Allora $\displaystyle\sum_{k=0}^{\infty} \cos^2 \theta_k \, \| \nabla f(x^k) \|_2^2 < +\infty$.

dim Sottraendo $-\nabla f(x^k)^T d^k$ ad entrambi i membri di (CvR) si ottiene la disuguaglianza $(\nabla f(x^k + t_k d^k) - \nabla f(x^k))^T d_k \geq (c_2 - 1) \nabla f(x^k)^T d^k$.

$$(c_2 - 1) \nabla f(x^k)^T d^k = (c_2 - 1) \| \nabla f(x^k) \|_2 \, \| d^k \|_2 \cos \theta_k$$

$$\wedge$$

$$(\nabla f(x^k + t_k d^k) - \nabla f(x^k))^T d^k \underset{\text{(dis. Schwartz)}}{\leq} \| \nabla f(x^k + t_k d^k) - \nabla f(x^k) \|_2 \, \| d^k \|_2 \underset{(\nabla f \text{ Lipschitziana})}{\leq} L \, t_k \| d^k \|_2^2$$

da cui $\displaystyle t_k \geq \frac{(c_2 - 1)}{L} \frac{\| \nabla f(x^k) \|_2}{\| d^k \|_2} \cos \theta_k > 0$ ($\cos \theta_k < 0$ e $c_2 < 1$)

Posti $f_{k+1} = f(x^k + t_k d^k)$ e $f_k = f(x^k)$, la condizione (AJO) si riscrive come

$$f_{k+1} \leq f_k + c_1 t_k \nabla f(x^k)^T d^k = f_k + c_1 t_k \underbrace{\| \nabla f(x^k) \|_2 \, \| d^k \|_2 \cos \theta_k}_{(<0 \text{ e })} \leq$$

$$\leq f_k - \underbrace{\boxed{c_1 \frac{(1 - c_2)}{L}}}_{\substack{\| \\ c > 0}} \| \nabla f(x^k) \|_2^2 \cos^2 \theta_k \longrightarrow f_{k+1} \leq f_k - c \| \nabla f(x^k) \|_2^2 \cos^2 \theta_k$$

Iterando a ritroso la disuguaglianza si ottiene: $\displaystyle f_{k+1} \leq f_0 - c \sum_{j=0}^{k} \| \nabla f(x^j) \|_2^2 \cos^2 \theta_j$

Quindi $\displaystyle\sum_{j=0}^{k} \| \nabla f(x^j) \|_2^2 \cos^2 \theta_j \leq (f_0 - f_{k+1})/c \leq \underbrace{(f_0 - M)/c}_{\text{costante}}$

Per $k \to +\infty$ si ottiene $\displaystyle\sum_{j=0}^{\infty} \| \nabla f(x^j) \|_2^2 \cos^2 \theta_j \leq (f_0 - M)/c < +\infty$

$$\sum_{k=0}^{\infty} \|\nabla f(x^k)\|_2^2 \cos^2\theta_k < +\infty \implies \|\nabla f(x^k)\|_2^2 \cos\theta_k \xrightarrow[k\to+\infty]{} 0$$

Teorema (convergenza) Supponiamo che

- f sia inf.nte limitata
- ∇f sia Lipschitziana
- $\exists\, \delta > 0$ tale che $\cos\theta_k \le -\delta \quad \forall k$

(equivalentemente: $\exists\, \sigma \in (0, \pi/2]$ t.c. $\theta_k \geq \pi/2 + \sigma$).

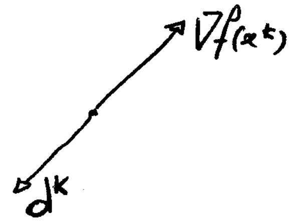
Allora ogni punto di accumulazione di $\{x^k\}$ è un punto stazionario.

dim Sia x^* un punto di accumulazione: $\exists \{x^{k_j}\}_j$ sottosuccessione t.c. $x^{k_j} \to x^*$

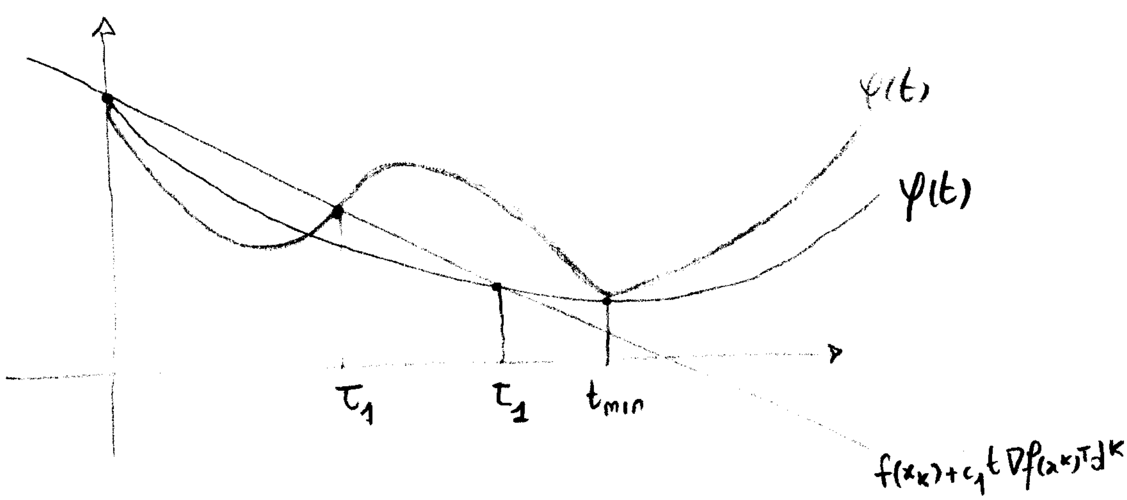
$$0 \xleftarrow[j\to+\infty]{} \cos^2\theta_{k_j} \|\nabla f(x^{k_j})\|_2^2 \geq \delta^2 \|\nabla f(x^{k_j})\|_2^2 \geq 0$$

da cui $\|\nabla f(x^*)\|_2 = \lim_{j\to+\infty} \|\nabla f(x^{k_j})\|_2 = 0$, ovvero $\nabla f(x^*) = 0$ ∎

Nota: $d^k = -\nabla f(x^k) \implies \cos\theta_k \equiv -1 \quad (\theta_k = \pi)$



Esempi in cui la ricerca esatta non soddisfa (AJO)



IL METODO DI NEWTON

Metodo di Newton-Raphson per risolvere il sistema di equazioni

$$F(x) = 0 \qquad \text{con} \qquad F : \mathbb{R}^n \to \mathbb{R}^n$$

$$x^{k+1} = x^k - [JF(x^k)]^{-1} F(x^k)$$

dove $JF(x^k)$ è la matrice jacobiana di F in x^k.

<u>Teorema</u> (NR) Supponiamo che F sia differenziabile 2 volte (con continuità) e sia $x^* \in \mathbb{R}^n$ tale che $F(x^*) = 0$. Se $JF(x^*)$ è invertibile, allora $\exists \delta > 0$ tale che per ogni $x^0 \in B(x^*, \delta)$, $x^k \to x^*$ ed inoltre esiste $M > 0$ t.c.

$$\|x^{k+1} - x^*\|_2 \leq M \|x^k - x^*\|_2^2 .$$

<u>Oss</u> Per la continuità di $JF : x \mapsto JF(x)$ e della f.z. determinante, $JF(x^*)$ invertibile ($\det JF(x^*) \neq 0$) $\Rightarrow$ $JF(x)$ è invertibile ($\det JF(x) \neq 0$) in un intorno di x^*.

(P) $\{ \min f(x) : x \in \mathbb{R}^n \}$ $\to$ applicare Newton-Raphson a $F = \nabla f$ per trovare punti stazionari di f: $\quad J(\nabla f)(x) = \nabla^2 f(x)$.

$d^k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$ è una direzione di discesa se $\nabla^2 f(x^k)$ è definita positiva ma non in generale: Newton non è necessariamente un metodo di discesa e può convergere a punti stazionari, che siano punti di massimo.

<u>Nota</u>: se $\nabla^2 f(x)$ è definita positiva per ogni x, allora f è strettamente convessa.

Per garantire la discesa, si possono rafforzare le ipotesi: invertibile $\rightsquigarrow$ definita positiva

$\nabla^2 f(x^*)$ definita positiva $\Rightarrow$ $\nabla^2 f(x)$ definita positiva per ogni x in un intorno di x^*.

(continuità di $\nabla^2 f$ richiesta)

Se $x^k \to x^*$, allora x^k appartiene definitivamente a tale intorno, e quindi d^k è una direzione di decrescita per k suff.nte grande (metodo def.nte di discesa)

__Teorema__ Supponiamo che f sia differenziabile 3 volte con continuità e sia $x^* \in \mathbb{R}$ un punto stazionario ($\nabla f(x^*) = 0$) per cui $\nabla^2 f(x^*)$ è definita positiva. Allora $\exists \delta > 0$ tale che per ogni $x^0 \in B(x^*, \delta)$, $x^k \to x^*$ ed inoltre $\exists M > 0$ tale che

$$\| x^{k+1} - x^* \|_2 \leq M \, \| x^k - x^* \|_2^2$$

$$\left(x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) \right)$$

__Oss__ le ipotesi su x^* garantiscono che sia un punto di minimo locale

__Nota 1__: il metodo è di natura locale: la convergenza è garantita, ma esclusivamente se il punto di partenza x^0 è sufficientemente vicino a x^*; altrimenti il metodo potrebbe non convergere o convergere ad un punto di max locale.

__Nota 2__: il passo $t_k \equiv 1$ (si possono comunque applicare anche 'ricerche monodim' (line search))

__Nota 3__: Ad ogni iterazione è richiesto il calcolo della matrice hessiana $\nabla^2 f(x^k)$ e della sua inversa $\leftarrow$ computazionalmente oneroso $\rightsquigarrow$ metodi quasi-Newton: non $\nabla^2 f(x^k)$ ma una sua approssimazione, aggiornata iterazione per iterazione tramite formule che utilizzano il risultato dell'iterazione collegando le derivate seconde alla variazione del gradiente.

__Altro punto di vista per la derivazione del metodo di Newton__

$$f(x+d) \approx f(x) + \nabla f(x)^T d \rightsquigarrow d = -\nabla f(x) \quad (\to \text{sviluppo I ordine} \rightsquigarrow \text{antigradiente})$$

Sviluppo II ordine: $f(x+d) \approx f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d = m_2(d)$

$\nabla^2 f(x)$ semidefinita positiva $\Rightarrow m_2(d)$ convessa

$\nabla m_2(d) = \nabla f(x) + \nabla^2 f(x) d$; $\nabla^2 f(x)$ definita positiva $\Rightarrow$ $\left(\nabla m_2(d) = 0 \Leftrightarrow d = -[\nabla^2 f(x)]^{-1} \nabla f(x) \right)$

(invertibile)

La direzione di Newton $-[\nabla^2 f(x)]^{-1} \nabla f(x)$ minimizza $m_2(d)$ se $\nabla^2 f(x)$ def. positiva.

Sviluppo I° ordine con resto esatto: $f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x + \tau d) d \quad \tau \in (0,1)$ opp.

$m_2(d)$ sostituisce $\nabla^2 f(x+\tau d)$ con $\nabla^2 f(x)$: l'approssimazione è "accurata" per $\|d\| \ll 1$.

Se f è una funzione quadratica, allora $f \equiv m_2$ ed il metodo di Newton termina in una unica iterazione.

Verifica ulteriore: $f(x) = \frac{1}{2} x^T Q x + b^T x$ con $Q = Q^T$ definita positiva.

$$f(x+d) = \frac{1}{2}(x+d)^T Q (x+d) + b^T(x+d) = \frac{1}{2} x^T Q x + b^T x + x^T Q d + b^T d + \frac{1}{2} d^T Q d$$

$$= \frac{1}{2} x^T Q x + b^T x + (Qx+b)^T d + \frac{1}{2} d^T Q d = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T Q d = m_2(d)$$

$\nabla f(x) = Qx + b$, $\nabla^2 f(x) = Q$ $\quad (\to \nabla f(x) = 0 \iff Qx = -b \iff x = -Q^{-1} b)$.

$x^1 = x^0 - Q^{-1}(Qx^0 + b) = x^0 - x^0 - Q^{-1} b = -Q^{-1} b$ quale che sia $x^0 \in \mathbb{R}^n$

<u>Nota 4</u>: il teorema fornisce anche un risultato sulla VELOCITÀ DI CONVERGENZA:

$\|x^{k+1} - x^*\|_2 \leq M \|x^k - x^*\|_2^2 \to$ il metodo di Newton ha convergenza superlineare di ordine 2

<u>Ricorda</u>: sia $x^k \to x^*$, con $x^k \neq x^*$ $\forall k$, e sia $p \geq 1$.

$$\lim_{k \to +\infty} (\sup) \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|_2^p} = \gamma$$

- $p = 1$, $\gamma \in (0,1)$ convergenza lineare
- $p = 1$, $\gamma = 1$ convergenza sublineare
- $p > 1$ convergenza superlineare di ordine p

$\|x^{k+1} - x^*\|_2 \leq \beta \|x^k - x^*\|_2^p$ con $\beta \in (0,1)$ $\to$ convergenza di ordine (almeno) p

<u>VELOCITÀ DI CONVERGENZA DEL METODO DEL GRADIENTE ESATTO</u>

Sia $\{x^k\}$ la successione generata dal metodo del gradiente esatto

<u>Teo 1</u> Supponiamo che f sia differenziabile 2 volte con continuità, e che $x^k \to x^*$ con $\nabla^2 f(x^*)$ definita positiva. Allora

$$f(x^{k+1}) - f(x^*) \leq [(\lambda_{max} - \lambda_{min})/(\lambda_{max} + \lambda_{min})]^2 (f(x^k) - f(x^*)),$$

dove $\lambda_{max} \geq \lambda_{min} > 0$ sono il più grande ed il più piccolo autovalore di $\nabla^2 f(x^*)$.

<u>Teo 2</u> Supponiamo che f sia differenziabile con continuità e ∇f sia Lipschitziana, che f sia <u>convessa</u> e che l'insieme dei punti di minimo sia non vuoto e limitato. Allora

$$f(x^k) - f^* = o(1/k)$$

(valore ottimo)

ovvero $\forall c > 0$ $\exists \bar{k}$ tale che $f(x^k) \leq f^* + c/k$ $\forall k \geq \bar{k}$

I METODI DEL GRADIENTE CONIUGATO

- CASO LINEARE

Risolvere $Ax=b$ con $b \in \mathbb{R}^n$, $A = A^T \in \mathbb{R}^{n \times n}$ definita positiva tramite la minimizzazione della funzione quadratica convessa $f(x) = \frac{1}{2} x^T A x - b^T x$ (in quanto risulta $\nabla f(x) = Ax - b$) modificando opportunamente la direzione di discesa del metodo del gradiente [esatto] ("coniugazione").

(HESTENES - STIEFEL 1952)

1) Scegliere $x^0 \in \mathbb{R}^n$; $k=0$
2) Se $r_k = b - Ax_k = 0$, allora STOP
3) $\beta_k = r_k^T r_k / r_{k-1}^T r_{k-1}$ se $k \geq 1$, $\beta_0 = 0$
4) $d_k = r_k + \beta_k d_{k-1}$ se $k \geq 1$, $d_0 = r_0$ ← coniugazione
5) $t_k = r_k^T r_k / d_k^T A d_k$
6) $x_{k+1} = x_k + t_k d_k$
7) $k = k+1$ e ritornare a 2)

Oss $\boxed{r_k = -\nabla f(x_k)}$

Prop (i) $r_k^T r_j = 0$ $j = 0, 1, \dots, k-1$ (ii) $d_k^T A d_{k-1} = 0$ (iii) $r_k^T d_j = 0$ $j = 0, 1, \dots, k-1$

Oss (i) garantisce che il metodo è diretto: la convergenza si ottiene in max n iterazioni

La formula al passo 5) fornisce il passo di ricerca esatta, infatti se $\varphi_k(t) = f(x_k + t d_k)$, risulta:

$$0 = \varphi_k'(t_k) = \nabla f(x_k + t_k d_k)^T d_k = (Ax_k + t_k A d_k - b)^T d_k =$$

$$= (t_k A d_k - r_k)^T d_k = t_k d_k^T A d_k - r_k^T r_k - \beta_k r_k^T d_{k-1} = t_k d_k^T A d_k - r_k^T r_k$$

da cui $0 = \varphi'(t_k) \iff t_k = r_k^T r_k / d_k^T A d_k$

d_K è una direzione di discesa:

$$\nabla f(x_K)^T d_K = -r_K^T d_K = -r_K^T r_K - \beta_K r_K^T d_{K-1} = -r_K^T r_K < 0 \quad (\text{se } r_K \neq 0).$$

- <u>CASO NONLINEARE</u>

Applicare la tecnica di coniugazione per la minimizzazione di $f: \mathbb{R}^n \to \mathbb{R}$ non necessariamente quadratica (Fletcher-Reeves 1964)

<u>Idea base</u>: sostituire il residuo r^k con $-\nabla f(x^k)$

<u>Difficoltà</u>: non esiste una formula esplicita per la ricerca esatta, e nel caso di ricerca inesatta bisogna garantire che il metodo sia di discesa.

$$d^k = -\nabla f(x^k) + \beta_k d^{k-1} : \quad \nabla f(x^k)^T d^k = -\|\nabla f(x^k)\|_2^2 + \beta_k \nabla f(x^k)^T d^{k-1}$$

Nel caso della ricerca esatta risulta $\nabla f(x^k)^T d^{k-1} = 0$ (infatti se $\varphi_{k-1}(t) = f(x^{k-1} + t d^{k-1})$, risulta $0 = \varphi'_{k-1}(t_{k-1}) = \nabla f(x^{k-1} + t_{k-1} d^{k-1})^T d^{k-1} = \nabla f(x^k)^T d^{k-1}$) e quindi il metodo è di discesa. Con la ricerca inesatta non c'è garanzia che $\nabla f(x^k)^T d^{k-1} \leq 0$. Ciò può essere ottenuto imponendo le <u>condizioni forti di Wolfe</u>:

$$(\text{AJO}) \quad f(x^k + t d^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k$$

$$(\text{CUR1}) \quad |\varphi'_k(t)| \leq c_2 |\varphi'_k(0)| \qquad (\varphi_k(t) = f(x^k + t d^k))$$

dove $0 < c_1 < c_2 < 1/2$. Nelle condizioni di Wolfe la condizione (AJO) è la medesima, mentre si richiede $c_2 < 1$ (invece di $c_2 < 1/2$) e la condizione di curvatura è

$$(\text{CUR}) \quad \varphi'_k(t) \geq c_2 \varphi'_k(0).$$

Se d^k è una direzione di discesa, e quindi $\varphi'_k(0) = \nabla f(x^k)^T d^k < 0$, allora (CUR1) può essere riscritta come $-|\varphi'_k(t)| \geq -c_2 |\varphi'_k(0)| = c_2 \varphi'_k(0)$ e pertanto implica (CUR) essendo $\varphi'_k(t) \geq -|\varphi'_k(t)|$.

METODO DEL GRADIENTE CONIUGATO DI FLETCHER-REEVES

1) Scegliere $x^0 \in \mathbb{R}^n$; $k=0$

2) Se $\nabla f(x^k)=0$, allora STOP

3) $\beta_k = \nabla f(x^k)^T \nabla f(x^k) / \nabla f(x^{k-1})^T \nabla f(x^{k-1})$ se $k \geq 1$, $\beta_0 = 0$

4) $d^k = -\nabla f(x^k) + \beta_k d^{k-1}$ se $k \geq 1$, $d^0 = -\nabla f(x^0)$

5) Calcolare $t_k > 0$ che soddisfa (AJO) e (CUR1)
(condizioni forti di Wolfe)

6) $x^{k+1} = x^k + t_k d^k$

7) $k = k+1$ e ritornare a 2)

Ipotesi: f inferiormente limitata

Lemma Se i passi t_k soddisfano (CUR 1), allora il metodo genera direzioni d^k tali che

$$-\frac{1}{1-c_2} \leq \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|_2^2} \leq \frac{2c_2-1}{1-c_2} .$$

Oss $c_2 \in (0, 1/2)$ garantisce che $(2c_2-1)/(1-c_2) < 0$ (infatti $\ell(c) = (2c-1)/(1-c)$ è monotona crescente e $\ell(0)=-1$, $\ell(1/2)=0$), quindi il metodo è di discesa e con dim analoga a quella di pagina 7 si può dimostrare che esiste un intervallo (a,b) in cui entrambe le condizioni (AJO) e (CUR1) valgono.

Supponendo che ∇f sia Lipschitziana, si può dimostrare $\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^k)\|_2^2 < +\infty$ dove θ_k è l'angolo compreso tra $\nabla f(x^k)$ e $-d^k$ (opportuna modifica della dim. di pagina 8).

Teo (convergenza) Supponiamo che ∇f sia Lipschitziana. Allora $\liminf_{k \to +\infty} \|\nabla f(x^k)\|_2 = 0$

È una forma molto debole di convergenza: garantisce soltanto l'esistenza di una sottosuccessione $\{x^{k_j}\}_j$ tale che $\|\nabla f(x^{k_j})\|_2 \xrightarrow[j\to+\infty]{} 0$, mentre la sottosuccessione $\{x^{k}\}$, potrebbe non convergere. Ovviamente se $x^k \to x^*$, allora è un punto stazionario di f, ma se $x^{k_s} \xrightarrow[s\to+\infty]{} x^*$ per qualche sottosuccessione $\{x^{k_s}\}_s$ ma $x^k \not\to x^*$ (cioè x^* è un punto di accumulazione della successione ma non è l'unico) allora la stazionarietà di x^* non è garantita.

Nel metodo del gradiente (con ricerca esatta) risulta $\cos\theta_k = 1$ per ogni k, mentre nel metodo del gradiente coniugato (nonlineare) non è garantito che esista un qualche $\delta > 0$ per cui $\cos\theta_k \geq \delta$ per ogni k.

Esaminiamo il caso $\cos\theta_k \approx 0$:

$$0 \approx \cos\theta_k = -\frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|_2 \|d^k\|_2} \overset{\text{lemma}}{\geq} \left(\frac{1-2c_2}{1-c_2}\right) \frac{\|\nabla f(x^k)\|_2}{\|d^k\|_2}$$

e quindi $\|\nabla f(x^k)\|_2 << \|d^k\|_2$. Inoltre $\cos\theta_k \approx 0$ rende molto probabile che si abbia $t_k \approx 0$ in quanto la direzione d^k è quasi ortogonale alla direzione di massima decrescita $-\nabla f(x^k)$: $t_k \approx 0 \Rightarrow x_{k+1} \approx x_k \Rightarrow \nabla f(x^{k+1}) \approx \nabla f(x^k)$

In questo caso abbiamo $\beta_{k+1} \approx 1$ e $\|\nabla f(x^{k+1})\|_2 \approx \|\nabla f(x^k)\|_2 << \|d^k\|_2$, e pertanto

$$d^{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} d^k \approx -\nabla f(x^{k+1}) + d^k \approx d^k$$

Quindi $\cos\theta_k \approx 0$ rende probabile che l'algoritmo produca una lunga sequenza di iterazioni poco utili

<u>Tecnica di restart</u>

Ogni $\bar{n}$ iterazioni porre $\beta_k = 0$, cioè effettuare un passo del metodo del gradiente con $d^k = -\nabla f(x^k)$. In questo modo si cancella la "memoria" delle direzioni precedenti (da cui "restart")

Oss Sia $\{x^{\bar{n}j}\}_j$ la sottosuccessione fornita da queste iterazioni. Poiché $\cos^2 \theta_{\bar{n}j} = 1$, risulta $\sum_{j=0}^{\infty} \| \nabla f(x^{\bar{n}j}) \|_2^2 < +\infty$ e quindi $\| \nabla f(x^{\bar{n}j}) \|_2 \xrightarrow[j \to +\infty]{} 0$

(analogamente nel caso in cui il "restart" non viene eseguito dopo un numero prefissato e costante di iterazioni --- in genere $\bar{n} \approx n$).

Variante di Polak-Ribiere (1969)

Nel caso lineare abbiamo

$$\beta_K = r_K^T r_K / r_{K-1}^T r_{K-1} \quad \text{ma anche} \quad \beta_K = r_K^T (r_K - r_{K-1}) / r_{K-1}^T r_{K-1}$$

poiché $r_K^T r_{K-1} = 0$. Questa variante utilizza, per analogia, questa seconda formula per calcolare il coefficiente β_K:

$$\beta_K^{PR} = \nabla f(x^k)^T (\nabla f(x^k) - \nabla f(x^{k-1})) / \nabla f(x^{k-1})^T \nabla f(x^{k-1})$$

Quindi, se $\nabla f(x^k) \approx \nabla f(x^{k-1})$, risulta $\beta_K^{PR} \approx 0$ e la tecnica di restart viene applicata (in maniera approssimata) automaticamente.

Non è possibile dimostrare alcun risultato di convergenza (se non in casi molto particolari [f fortemente convessa + ricerca esatta]) ed inoltre d^k potrebbe non essere una direzione di discesa.

Oss È possibile avere $\beta_K^{PR} < 0$, da cui la variante $\beta_K^{PR+} = \max\{\beta_K^{PR}, 0\}$ per cui con una opportuna variante delle condizioni forti di Wolfe è possibile dimostrare la convergenza del metodo nel caso generale.

METODI SENZA DERIVATE

(P) $\min\{f(x) : x \in \mathbb{R}^n\}$ ← metodi che richiedano soltanto il calcolo di valori della fz. obiettivo f.

Motivazioni:

- f non differenziabile

  oppure

- difficoltà nel calcolo delle derivate

Calcolo derivate

- Differenze finite (via sviluppi di Taylor)

$$\frac{\partial f}{\partial x_i}(x) = \frac{f(x+te_i)-f(x)}{t} + O(t) \qquad (r_x(t)/t \to 0)$$

$$\frac{\partial f}{\partial x_i}(x) = \frac{f(x+te_i)-f(x-te_i)}{2t} + O(t^2)$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{f(x+te_i+te_j)-f(x+te_i)-f(x+te_j)+f(x)}{t^2} + O(t)$$

- Differenziazione automatica

  "Scomposizione Fz. + regole differenziazione fz. composte"

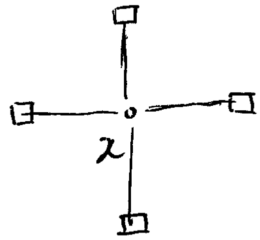
Consideriamo $\{e_1, \dots, e_n\} \subseteq \mathbb{R}^n$ la "base canonica" e $D = \{d_1, \dots, d_n, d_{n+1}, \dots, d_{2n}\}$

con $d_i = \begin{cases} e_i & 1 \le i \le n \\ -e_{i-n} & n+1 \le i \le 2n \end{cases}$

<u>Oss</u> Ogni $v \in \mathbb{R}^n$ è combinazione lineare a coefficienti non negativi dei vettori di D

$v = \sum_{i=1}^n \alpha_i e_i$ con $\alpha_i \in \mathbb{R}$ $\Rightarrow$ $v = \sum_{i=1}^n |\alpha_i| \hat{d}_i$ con $\hat{d}_i = \begin{cases} d_i (=e_i) & \text{se } \alpha_i > 0 \\ d_{i+n} & \text{se } \alpha_i < 0. \end{cases}$

__Idea__: valutare f nei punti $\{x + t d_i\}_{i=1..2n}$ per qualche fissato t; se si individua un valore migliore, iterare a partire dal nuovo punto, altrimenti diminuire t.



ALGORITMO DI RICERCA DIRETTA "A COMPASSO"

1) Scegliere $x^0 \in \mathbb{R}^n$, $t_0 > 0$; $k = 0$, $\ell = 0$

2) Se esiste $d \in D$ tale che $f(x^k + t_k d) < f(x^k)$,

allora $x^{k+1} = x^k + t_k d$, $t_{k+1} = t_k$

altrimenti $x^{k+1} = x^k$, $t_{k+1} = t_k/2$, $\ell = \ell + 1$, $z^\ell = x^k$, $t^\ell = t_k$

3) $k = k+1$ e ritornare a 2)

__Nota__: $\{z^\ell\}$ costituisce la sequenza delle iterazioni "fallimentari"

__CRITERIO DI ARRESTO__: $t_k \leq \delta$ per una tolleranza fissata $\delta > 0$.

Se la curva di sotto livello $L_f(x^0) = \{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$ è compatta, allora il passo di ricerca t_k viene aggiornato dopo un numero finito di iterazioni. Infatti la griglia $M_k = \{x_k + t_k \left(\sum_{i=1}^{2n} \mu_i d_i \right) : \mu_i \in \mathbb{Z}_+\}$ individua tutti i punti visitabili senza aggiornare il passo e $M_k \cap L_f(x^0)$ è un insieme finito poiché $L_f(x^0)$ è compatto.

__Teo__ (convergenza). Sia f differenziabile su $\mathbb{R}^n$. Supponiamo che

- $L_f(x^0)$ sia compatto
- ∇f sia Lipschitziana (ovvero $\exists L > 0$ t.c. $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\| \ \forall x, y$),

allora $\lim_{\ell \to +\infty} \|\nabla f(z^\ell)\|_2 = 0$

dim Sia $v \in \mathbb{R}^n$ con $\|v\|_2 = 1$: $\quad v = \sum_{i=1}^{n} \gamma_i \hat{d}_i \quad$ con $\gamma_i \geq 0$.

$$1 = v^T v = \Big(\sum_{i=1}^{n} \gamma_i \hat{d}_i\Big)^T \Big(\sum_{j=1}^{n} \gamma_j \hat{d}_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_i \gamma_j \hat{d}_i^T \hat{d}_j = \sum_{i=1}^{n} \gamma_i^2$$

quindi esiste k tale che $\gamma_k \geq 1/\sqrt{n}$. Inoltre $-(-\hat{d}_k)^T v = \hat{d}_k^T v = \gamma_k$.

Poiché $(-\hat{d}_k) \in D$, possiamo concludere: $\exists d \in D$ t.c. $-d^T v \geq 1/\sqrt{n}$.

Supponiamo $\nabla f(z^\ell) \neq 0$: $\quad v = \nabla f(z^\ell) / \|\nabla f(z^\ell)\|_2 \implies \exists d \in D$ t.c. $-\nabla f(z^\ell)^T d \geq \dfrac{\|\nabla f(z^\ell)\|_2}{\sqrt{n}}$

Poiché z^ℓ è ottenuto in una iterazione fallimentare, risulta $f(z^\ell + t_\ell d) \geq f(z^\ell)$ e dal teo (valor medio): $0 \leq f(z^\ell + t_\ell d) - f(z^\ell) = \nabla f(z^\ell + \gamma t_\ell d)^T d$ per $\gamma \in (0,1)$ opportuno.

Quindi:

$$0 \leq \frac{\|\nabla f(z^\ell)\|_2}{\sqrt{n}} \leq -\nabla f(z^\ell)^T d \leq [\nabla f(z^\ell + \gamma t_\ell d) - \nabla f(z^\ell)]^T d \leq \quad \text{(dis. Schwarz)}$$

$$\leq \|\nabla f(z^\ell + \gamma t_\ell d) - \nabla f(z^\ell)\|_2 \, \|d\|_2 \leq L \|\gamma t_\ell d\|_2 = L \gamma t_\ell \xrightarrow[\ell \to +\infty]{} 0$$

$(\|d\|_2 = 1,\ \nabla f$ Lipschitziana$)$ $\qquad (t_{\ell+1} = t_0 / 2^\ell)$ ∎

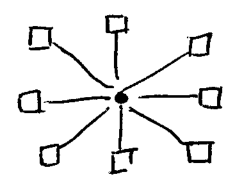
Cor Ogni punto di accumulazione di $\{z^\ell\}$ è un punto stazionario di f.

dim z^* punto di accumulazione $\implies \exists \{z^{\ell_j}\}$ sottosuccessione t.c. $z^{\ell_j} \to z^*$ e $\|\nabla f(z^{\ell_j})\|_2 \xrightarrow[j \to +\infty]{} \|\nabla f(z^*)\|_2$ per la continuità di ∇f e $\|\cdot\|_2$. Dal teo (convergenza) segue $\|\nabla f(z^*)\|_2 = 0$ e quindi $\nabla f(z^*) = 0$ per l'unicità del limite ∎

Nota: Poiché $\{z^\ell\} \subseteq L_f(x^0)$, la compattezza di $L_f(x^0)$ garantisce l'esistenza di un punto di accumulazione.

Nel caso non differenziabile $\{z^\ell\}$ potrebbe convergere a punti "non stazionari".

<u>Generalizzazioni dello schema algoritmico</u>

- $D \supseteq \{\pm b_1, \pm b_2, \dots, \pm b_n\}$ con $\{b_1, \dots, b_n\} \subseteq \mathbb{R}^n$ base $\longrightarrow$ 

- possibilità di considerare D_k differenti per ciascuna iterazione