METODO DEL GRADIENTE[1]

Il metodo del gradiente (steepest descent "più ripida discesa") è il più antico, il più noto, uno dei metodi più usati, proposti per la minimizzazione non vincolata di funzioni differenziabili.

Il vettore gradiente $\nabla f(x_k)$ individua la direzione di più ripida salita della funzione f in x_k, cioè la direzione in cui la derivata direzionale è massima. L'idea alla base del metodo del gradiente è di cercare valori di minimo relativo iterativamente, muovendosi ad ogni passo nella direzione opposta a $\nabla f(x_k)$.

Sia x_k la soluzione ottenuta alla *k-esima* iterata, supponendo $\nabla f(x_k) \neq 0$, l'iterata successiva è definita da

$$x_{k+1} = x_k + \alpha_k d_k$$

dove

$$d_k = -\nabla f(x_k)$$

è la direzione opposta a quella del gradiente, o *antigradiente* di f in x_k. Mi muovo in maniera opposta (180°) a quella direzione indicata dal gradiente. Il procedimento termina non appena $\nabla f(x_k) = 0$.

L'algoritmo ora formulato non sembra rivestire particolare interesse poiché risulta molto costoso (ad ogni passo occorre calcolare n derivate direzionali e minimizzare la funzione lungo la direzione individuata). Il metodo diventa efficiente se applicato a problemi quadratici. L'algoritmo della massima discesa si caratterizza come metodo iterativo per la risoluzione anche di sistemi lineari.

---

[1] Metodo della più ripida discesa (steepest descent): è quella direzione che minimizza il prodotto scalare $d^T \nabla f(x_k)$.

Si può osservare che la direzione dell'antigradiente (normalizzata) è la direzione che minimizza la derivata direzionale in x_k tra tutte le direzioni che hanno norma euclidea unitaria (lunghezza 1), cioè è la soluzione del problema

$$-\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} \in \arg\min_{d} \; d^T \nabla f(x_k)$$

$$\|d\|_2 = 1$$ [2]

Dim.:

utilizziamo le condizioni di ottimalità di Karush-Kuhn-Tucker (i gradienti sono indipendenti)

$$\nabla f(x_k) + 2\lambda d = 0$$ [3]

$$d^T d = 1$$

segue

$$d = -\frac{\nabla f(x_k)}{2\lambda}$$

$$d^T d = 1 \Rightarrow -\frac{\nabla f(x_k)^T}{2\lambda} \cdot \left[-\frac{\nabla f(x_k)}{2\lambda} \right] = 1$$

$$\Rightarrow \nabla f(x_k)^T \cdot \nabla f(x_k) = 4\lambda^2$$

$$\lambda = \pm \frac{\|\nabla f(x_k)\|}{2}$$

scegliendo $\lambda > 0$ per un minimo ottengo

$$\nabla f(x_k) + 2 \cdot \frac{\|\nabla f(x_k)\|}{2} \cdot d = 0$$

$$d_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$$ [4]

c.v.d.

---

[2] La norma non è differenziabile, invece lo è la norma al quadrato, useremo $\|d\|_2^2 = 1$ al posto di $\|d\|_2 = 1$.

$\|d\|_2^2$ si può scrivere come $d^T \cdot d$.

[3] Gradiente della funzione obiettivo $+ \lambda$ per gradiente del vincolo.

[4] Sappiamo che $-\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} = \cos\alpha$, qual è la direzione che ha il coseno minimo (vicino a -1)?

L'ottimalità locale della direzione $-\nabla f(x_k)$ dipende dalla scelta della norma e che pertanto in un punto fissato x_k qualsiasi direzione di discesa può interpretarsi come una direzione di discesa più ripida in una norma opportuna. Se usiamo ad esempio la norma del sup (uniforme o di Chebyshev), $\|\cdot\|_\infty = 1$è un quadrato, la direzione di discesa più ripida sarà una direzione di discesa che dipende dalla norma utilizzata.

In genere, prendiamo come direzione di discesa più ripida l'antigradiente, perché la norma euclidea è quella più utilizzata.

Nel caso quadratico le curve di livello della funzione f sono ellissoidi concentriche, il gradiente è normale alla retta tangente in x_k alla curva di livello.

L'interesse della direzione $-\nabla f(x_k)$ risiede nel fatto che, se il gradiente è continuo essa costituisce una direzione di discesa *continua* rispetto a x, che si annulla se e solo se xè un punto stazionario. Questa proprietà assicura che con una scelta opportuna del passo α_k sia possibile stabilire un risultato di convergenza globale.

Algoritmo di Discesa del Gradiente

Dati: scegliamo punto iniziale $x_0 \in R^n$; $k = 0$

while una stopping rule non è soddisfatta ($\|\nabla f(x_k)\| \neq 0$)[5] **do**

$d_k := -\nabla f(x_k)$;

Calcola un passo $\alpha_k \in R^+$lungo d_k con una tecnica di ricerca unidimensionale

(ad esempio Armijo);

$x_{k+1} = x_k + \alpha_k d_k$;

$k = k + 1$

end

[5] Ci fermiamo quando il gradiente è vicino a zero.

Teorema (Convergenza globale del metodo del gradiente)

Sia $f : R^n \to R$ una funzione continuamente differenziabile su R^n.

Si supponga che L_0 sia compatto. Sia $\{x_k\}$ la successione prodotta dal metodo del gradiente. Si supponga inoltre che la ricerca unidimensionale[6] sia tale da soddisfare le condizioni:

1. $f(x_{k+1}) < f(x_k)$ (monotonicità)[7]
2. se $\nabla f(x_k) \neq 0 \quad \forall k$, si ha

$$\lim_{k\to\infty} \frac{d_k^T \nabla f(x_k)}{\|d_k\|} = 0^{8}.$$

Allora, o il metodo si arresta in un numero finito di iterazioni con $\nabla f(x_k) = 0$[9], oppure viene prodotta una successione infinita, tale che ogni punto di accumulazione $\bar{x}$ di $\{x_k\} \in L_0$ e soddisfa $\nabla f(\bar{x}) = 0$[10].

Dim.:

Siccome la direzione è l'antigradiente

$$d_k = -\nabla f(x_k)$$

si ha

$$\frac{d_k^T \nabla f(x_k)}{\|d_k\|} = -\|\nabla f(x_k)\|^{11}$$

Essendo soddisfatte le ipotesi del teorema di convergenza globale (condizione sufficiente), esiste una funzione di forzamento $\sigma(t) = t$[12] tale che sono soddisfatte le tesi di tale teorema, cioè:

o viene trovato un punto stazionario dopo un numero finito di passi, oppure viene generata una successione infinita di punti tale che[13]:

1. $\lim_{k\to\infty} \nabla f(x_k) = 0$, la successione dei gradienti converge a zero;

---

[6] Se usiamo la regola di Armijo (è una possibilità), abbiamo la convergenza globale.

[7] Sia garantita la discesa.

[8] È una ipotesi del teorema di convergenza globale. È una tesi del teorema di convergenza del metodo di Armijo.

[9] Tangente orizzontale.

[10] Non è garantita la convergenza della successione $\{x_k\}$. I punti limite $\bar{x}$ sono stazionari.

[11] Ho moltiplicato 1° e 2° membro per $\nabla f(x_k)$ e normalizzato d_k. Normalizzare un vettore v significa moltiplicare il vettore v per un opportuno scalare α scelto in modo tale che $\|\alpha \cdot v\| = 1$. Per far si che il vettore $\alpha \cdot v$ abbia norma 1, dobbiamo prendere $\alpha = \frac{1}{\|v\|}$, infatti con tale scelta di abbiamo $\left\|\frac{1}{\|v\|} \cdot v\right\| = \frac{1}{\|v\|} \cdot \|v\| = 1$.

[12] $\sigma(t) = t$ è la funzione identità. Se va a zero la funzione, va a zero anche l'argomento della funzione.

[13] O $\exists k$ con $\nabla f(x_k) = 0$, oppure la successione dei gradienti tende a zero $\{\nabla f(x_k)\} \to 0$.

2. ogni punto limite $\bar{x}$ di $\{x_k\}$ soddisfa $\nabla f(\bar{x}) = 0$, limite di ogni sottosuccessione è un punto stazionario.

c.v.d.

La convergenza globale può essere assicurata attraverso una tecnica di ricerca unidimensionale tipo-Armijo o Wolfe.

Il metodo del gradiente costituisce il prototipo di algoritmo globalmente convergente e può essere usato, in associazione ad altri metodi dotati di migliori caratteristiche di convergenza locale, per assicurare la convergenza globale.

Teorema (Metodo del gradiente con passo costante $\alpha_k = \eta > 0$)[14]

Sia $f : R^n \to R$ una funzione continuamente differenziabile su R^n.

Se $x_0 \in L_0$ [15] compatto e $\exists \mathrm{L} > 0$ (costante di Lipschitz) tale che

$$\|\nabla f(x) - \nabla f(y)\| \le \mathrm{L}\|x - y\| \qquad \forall\, x, y \in R^n \text{ [16]}$$

Supponiamo che scegliendo η valga la condizione

$$\varepsilon \le \eta \le \frac{2-\varepsilon}{\mathrm{L}} \qquad \varepsilon > 0$$

Allora, o viene trovato un punto stazionario dopo un numero finito di iterazioni, oppure la successione dei gradienti $\{x_k - \eta \nabla f(x_k)\}$ converge a zero[17].

Dim.:

Per ipotesi il gradiente di f, $\nabla f(x_k)$, gode di una proprietà più forte della continuità, è Lipschitz continuo.

Per la dimostrazione utilizziamo il seguente Lemma:

$\forall x \in R$, $\alpha > 0$, $d \in R^n$, utilizzando il teorema della media in forma integrale si può scrivere

$$f(x + \alpha \cdot d)$$

$$= f(x) + \alpha \int_0^1 d^T \nabla f(x + t \cdot \alpha \cdot d)\, dt \text{ [18]}$$

---

[14] η è il "learning rate", il passo è una costante, si usa nelle Reti Neurali. Rispetto al teorema precedente ha ipotesi in più. In questo momento ha grande importanza, è il metodo più usato nell'addestramento di reti neurali multistrato (metodo della backpropagation).
È un metodo molto lento, però ci assicura la convergenza.

[15] Rispetto al teorema precedente è stato aggiunto x_0.

[16] Le variazioni del gradiente non sono troppe elevate nei punti in cui lo valutiamo. È simile alla continuità, ma è più forte. C'è L che dice di quanto varia la funzione per spostarmi da x a y. La **lipschitzianità** risulta una condizione più forte della **continuità** e meno forte della **differenziabilità**. $|x|$ è Lipschitz continua ma non è differenziabile in $O(0;0)$.

[17] O $\exists k$ con $\nabla f(x_k) = 0$, oppure la successione dei gradienti tende a zero $\{x_k - \eta \nabla f(x_k)\} \to 0$.

$$= f(x) + \alpha \int_0^1 d^T \nabla f(x + t \cdot \alpha \cdot d) dt + \alpha \cdot d^T \nabla f(x) - \alpha \cdot d^T \nabla f(x)$$ [19]

$$= f(x) + \alpha \cdot d^T \nabla f(x) + \alpha \int_0^1 \left(d^T \nabla f(x + t \cdot \alpha \cdot d) - d^T \nabla f(x) \right) dt$$

$$\leq f(x) + \alpha \cdot d^T \nabla f(x) + \alpha \int_0^1 \|d\| \cdot \|\nabla f(x + t \cdot \alpha \cdot d) - \nabla f(x)\| dt$$ [20]

$$\leq f(x) + \alpha \cdot d^T \nabla f(x) + \alpha \int_0^1 \|d\| \cdot \mathrm{L} \|x + t \cdot \alpha \cdot d - x\| dt$$ [21]

$$\leq f(x) + \alpha \cdot d^T \nabla f(x) + \alpha \int_0^1 \|d\| \cdot \mathrm{L} \cdot t \cdot \alpha \cdot \|d\| dt$$

$$= f(x) + \alpha \cdot d^T \nabla f(x) + \frac{\alpha^2 \mathrm{L}}{2} \|d\|^2$$ [22]

Supponiamo che sia $x_k \in L_0$, utilizzando il lemma precedente si ha[23]

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) = f(x_k - \eta \nabla f(x_k))$$

$$\leq f(x_k) - \eta \nabla f(x_k)^T \nabla f(x_k) + \frac{\eta^2 \mathrm{L}}{2} \nabla f(x_k)^T \nabla f(x_k)$$

da qui segue

$$f(x_{k+1}) - f(x_k) \leq -\eta \|\nabla f(x_k)\|^2 + \frac{\eta^2 \mathrm{L}}{2} \|\nabla f(x_k)\|^2$$

$$f(x_{k+1}) - f(x_k) \leq \eta \left(-1 + \frac{\eta \mathrm{L}}{2} \right) \cdot \|\nabla f(x_k)\|^2$$ [24]

---

[18] $\int_0^1 d^T \nabla f(x + t \cdot \alpha \cdot d) dt$, per $t = 0$ ho $f(x)$, per $t = 1$ ho $f(x + \alpha \cdot d)$, descrive tutto il segmento che va da 0 a 1.

[19] Sommo e sottraggo $\alpha \cdot d^T \nabla f(x)$.

[20] $\left\| \int \right\| \leq \int \| \ \|$

[21] $\|\nabla f(x + t \cdot \alpha \cdot d) - \nabla f(x)\| \leq \mathrm{L} \|x + t \cdot \alpha \cdot d - x\| \leq \mathrm{L} \cdot t \cdot \alpha \|d\|$

[22] $\alpha \int_0^1 \|d\| \cdot \mathrm{L} \cdot t \cdot \alpha \cdot \|d\| dt = \alpha^2 \cdot \mathrm{L} \cdot \|d\|^2 \int_0^1 t dt = \alpha^2 \cdot \mathrm{L} \cdot \|d\|^2 \cdot \left[\frac{t^2}{2} \right]_0^1 = \frac{\alpha^2 \cdot \mathrm{L} \cdot \|d\|^2}{2}$.

[23] $\alpha = \eta$ e $d = -\nabla f(x_k)$. Inoltre $\|d\|^2 = |d^T| \cdot |d| = \nabla f(x_k)^T \cdot \nabla f(x_k)$.

[24] $f(x_{k+1}) - f(x_k)$ la vorremmo negativa per avere una successione monotona (decrescente), quindi dobbiamo imporre $-1 + \frac{\eta \mathrm{L}}{2} < 0$ dal momento che $-1 + \frac{\eta \mathrm{L}}{2}$ non sta diminuendo e $\|\nabla f(x_k)\|^2$ è una quantità positiva.

Scegliendo $\eta > 0$, se vogliamo che $-1+\frac{\eta \mathrm{L}}{2}$ diminuisca deve essere

$$-1+\frac{\eta \mathrm{L}}{2}<0$$

$$\frac{\eta \mathrm{L}}{2}<1$$

$$\eta<\frac{2}{\mathrm{L}}$$

oppure

$$\varepsilon \leq \eta \leq \frac{2-\varepsilon}{\mathrm{L}}$$ [25]

quindi la successione generata dal metodo è monotona (decrescente) perché

$$\begin{aligned}
&f\left(x_{k+1}\right)-f\left(x_k\right)<0 \\
&f\left(x_k\right)-f\left(x_{k+1}\right)>0 \\
&f\left(x_k\right)>f\left(x_{k+1}\right) \Rightarrow f \text{ è monotona decrescente}
\end{aligned}$$

Inoltre, se conosciamo L, allora possiamo trovare η.

Quindi tutta la successione $\{x_k\}$ rimane per ipotesi in L_0, chiuso e limitato; la successione $\{f(x_k)\}$ è monotona e limitata, siccome siamo in un compatto avrà minimo e quindi converge[26]. Inoltre

$$\begin{aligned}
&f\left(x_{k+1}\right)-f\left(x_k\right) \leq \eta\left(-1+\frac{\eta \mathrm{L}}{2}\right) \cdot\left\|\nabla f\left(x_k\right)\right\|^2 \\
&f\left(x_k\right)-f\left(x_{k+1}\right) \geq \eta\left(-1+\frac{\eta \mathrm{L}}{2}\right) \cdot\left\|\nabla f\left(x_k\right)\right\|^2 \\
&\left\|\nabla f\left(x_k\right)\right\|^2 \leq \frac{f\left(x_k\right)-f\left(x_{k+1}\right)}{\eta\left(-1+\frac{\eta \mathrm{L}}{2}\right)} \rightarrow 0 \\
&\left\|\nabla f\left(x_k\right)\right\|^2 \rightarrow 0
\end{aligned}$$

e questo implica

$$\nabla f(x_k) \rightarrow 0$$

quindi abbiamo la convergenza del gradiente.

c.v.d.

---

[25] I 2 ε potrebbero essere diversi, per semplicità abbiamo usato solo un ε.

[26] In questo caso non abbiamo bisogno del teorema di convergenza globale perché abbiamo tutto per dimostrare la convergenza del gradiente.

Rapidità di convergenza del metodo del gradiente

Per valutare la rapidità di convergenza del metodo del gradiente facciamo riferimento al comportamento del metodo nella minimizzazione di una funzione quadratica convessa[27]. Supponiamo che la funzione quadratica

$$f(x) = \frac{1}{2} x^T Q\, x \qquad Q \succ 0$$

abbia minimo (locale e globale) nell'origine, il metodo utilizzato sia quello del gradiente

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \quad {}^{28} \\ &= x_k - \alpha_k \cdot Q \cdot x_k \\ &= (I - \alpha_k Q) x_k \end{aligned}$$

Calcoliamo l'errore al passo $k+1$ [29]

$$\begin{aligned} \mathrm{e}_{k+1} &= \left\| x_{k+1} - x^* \right\| \\ &= \left\| x_{k+1} \right\| \end{aligned}$$

Quindi

$$\begin{aligned} \left\| x_{k+1} \right\| &= \left\| (I - \alpha_k Q) x_k \right\| \\ &= \sqrt{x_k^T \cdot (I - \alpha_k Q) \cdot (I - \alpha_k Q) \cdot x_k} \quad {}^{30} \\ &= \sqrt{x_k^T \cdot (I - \alpha_k Q)^2 x_k} \end{aligned}$$

---

[27] Usiamo la funzione quadratica, perché localmente rappresenta bene una funzione nell'intorno del punto di minimo.

$$f(x) = \frac{1}{2} x^T Q x + c^T x \qquad Q \succ 0$$

[28] $\nabla f(x_k) = Q \cdot x_k + c$

$\nabla^2 f(x_k) = Q$

[29] Lo studio della rapidità di convergenza viene effettuato supponendo che $\{x_k\}$ sia una successione convergente a $x^* \in R^n$ e consiste nello stimare la velocità con cui tende a zero l'errore e_k definito da una norma della differenza tra x_k e x^*, ossia $\mathrm{e}_k = \left\| x_k - x^* \right\|$. Nel nostro caso la distanza tra x_{k+1} e il punto di ottimo x^* che è zero.

[30] $(I - \alpha_k Q)$ è una matrice, I è la matrice identità. x_k è il vecchio iterato. $\sqrt{x_k^T \cdot (I - \alpha_k Q)^2 x_k}$ è molto complicata, vogliamo sostituire la matrice Q con autovalori eliminando così la parte matriciale.

Sia λ^* l'autovalore massimo della matrice $(I-\alpha_k Q)^2$, ricordando che

$$\lambda_{\min} x^T x \le x^T A x \le \lambda_{\max} x^T x \quad {}^{31}$$

abbiamo

$$x_k^T (I-\alpha_k Q)^2 x_k \le \lambda^* x_k^T x_k \quad {}^{32}$$

Immaginiamo, per semplicità, che λ_1, λ_n siano rispettivamente l'autovalore minimo e l'autovalore massimo di Q, allora vale la stima seguente

$$\frac{\left\| x_{k+1} - x^* \right\|}{\left\| x_k - x^* \right\|} \le \max\left\{ |1-\alpha_k \lambda_1|, |1-\alpha_k \lambda_n| \right\}$$

$$\left\| x_{k+1} \right\| \le \left\| x_k \right\| \max\left\{ |1-\alpha_k \lambda_1|, |1-\alpha_k \lambda_n| \right\}$$

Togliamo i moduli a $\max\left\{ |1-\alpha_k \lambda_1|, |1-\alpha_k \lambda_n| \right\}$, ricordando che $|a| = \max\{-a, +a\}$, abbiamo

$$\max\left\{ |1-\alpha_k \lambda_1|, |1-\alpha_k \lambda_n| \right\} =$$

$$\max\{1-\alpha_k \lambda_1, -1+\alpha_k \lambda_1, 1-\alpha_k \lambda_n, -1+\alpha_k \lambda_n\}$$

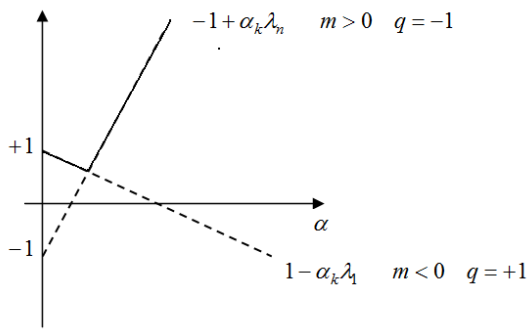
ma $\alpha_k \ge 0, \lambda_1 < \lambda_n$, quindi

$$1-\alpha_k \lambda_1 \ge 1-\alpha_k \lambda_n$$
$$-1+\alpha_k \lambda_n \ge -1+\alpha_k \lambda_1$$

in definitiva abbiamo

$$\left\| x_{k+1} \right\| \le \left\| x_k \right\| \max\{1-\alpha_k \lambda_1, -1+\alpha_k \lambda_n\}$$

che è una funzione di α



---

[31] $\lambda_{\min}$ è l'autovalore minimo, $\lambda_{\max}$ è l'autovalore massimo.

[32] Non conosciamo gli autovalori di $(I-\alpha_k Q)^2$.

Indichiamo con α^* (passo ideale) il migliore caso possibile

$$\alpha^* = 1 - \alpha_k \lambda_1 = -1 + \alpha_k \lambda_n$$ [33]

$$\alpha^* = \frac{2}{\lambda_1 + \lambda_n}$$

Al posto di α_k sostituisco $\frac{2}{\lambda_1 + \lambda_n}$ e ottengo

$$\begin{aligned}
\|x_{k+1}\| &\le \|x_k\| \cdot \left(1 - \frac{2}{\lambda_1 + \lambda_n} \cdot \lambda_1\right) \\
&= \|x_k\| \cdot \left(\frac{\lambda_1 + \lambda_n - 2\lambda_1}{\lambda_1 + \lambda_n}\right) \\
&= \|x_k\| \cdot \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right) \\
&= \|x_k\| \cdot \left(\frac{\frac{\lambda_n}{\lambda_1} - \frac{\lambda_1}{\lambda_1}}{\frac{\lambda_n}{\lambda_1} + \frac{\lambda_1}{\lambda_1}}\right) \\
&= \|x_k\| \cdot \left(\frac{\rho - 1}{\rho + 1}\right)
\end{aligned}$$

dove $\rho = \frac{\lambda_n}{\lambda_1}$ è il numero di condizionamento della matrice Q.

Osserviamo che, se $\rho \approx 1$ (autovalori max e min sono vicini), la convergenza è molto veloce, la convergenza viene ottenuta in un solo passo da ogni punto iniziale.

Se $\rho >> 1$ la convergenza è notevolmente lenta (autovalori max e min distanti tra di loro).

La rapidità di convergenza del metodo del gradiente peggiora quindi, in genere, al crescere di ρ, ossia all'aumentare del *mal condizionamento* della matrice Q.

---

[33] $\begin{cases} \alpha^* = 1 - \alpha_k \lambda_1 \\ \alpha^* = -1 + \alpha_k \lambda_n \end{cases} \Rightarrow 1 - \alpha_k \lambda_1 = -1 + \alpha_k \lambda_n \Rightarrow 2 - \alpha_k (\lambda_1 + \lambda_n) = 0 \Rightarrow \alpha_k = \frac{2}{\lambda_1 + \lambda_n}$.

zig-zagging

Vediamo un esempio pratico, sia f una funzione in due variabili

$$f(x,y)=\frac{1}{2}\left(x^2+\rho\cdot y^2\right)$$
$$\nabla f(x,y)=\begin{bmatrix} x \\ \rho\cdot y \end{bmatrix}$$
$$\nabla^2 f(x,y)=\begin{bmatrix} 1 & 0 \\ 0 & \rho \end{bmatrix}$$

dove $\rho>0$. Immaginiamo che $(\rho,1)$ sia il punto di partenza, sappiamo che il passo ottimo α_k [34] è

$$\alpha_k=-\frac{d_k^T\nabla f(x_k)}{d_k^T\cdot Q\cdot d_k}$$
$$=\frac{x^2+\rho^2\cdot y^2}{x^2+\rho^3\cdot y^2}$$

dopo una serie di calcoli elementari (noiosi), se partiamo da $x_0=\rho$ e $y_0=1$ otteniamo

$$x_{k+1}=\rho\left(\frac{\rho-1}{\rho+1}\right)^k$$
$$y_{k+1}=\left(-\frac{\rho-1}{\rho+1}\right)^k$$

Studiamo la convergenza:

- molto veloce se $\rho\approx 1$
- estremamente lenta se $\rho>>1$ o $\rho<<1$, in questo caso abbiamo il cosiddetto "*zig-zagging*".

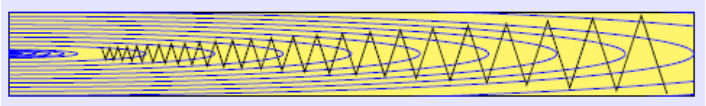
Nella seguente figura è illustrato il fenomeno dello "zig-zagging" relativo alla funzione $f(x,y)=\frac{1}{2}\left(x^2+\rho\cdot y^2\right)$ quando $\rho=50$

$f(x)=\frac{1}{2}x^TQx+c^Tx \qquad Q\succ 0$

[34] $\nabla f(x_k)=Q\cdot x_k+c$

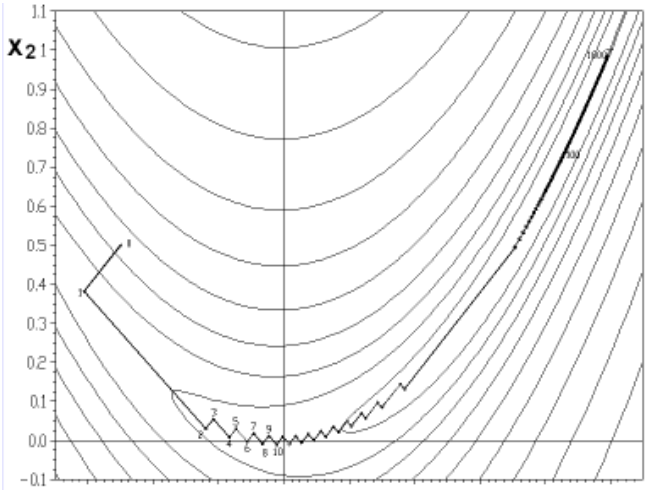
$\nabla^2 f(x_k)=Q$

$$\phi(\alpha)=f(x_k+\alpha d_k)=\frac{1}{2}\left(x_k+\alpha d_k\right)^T\cdot Q\cdot\left(x_k+\alpha d_k\right)+c^T\left(x_k+\alpha d_k\right)$$
$$\phi'(\alpha)=d_k^T\cdot Q\cdot\left(x_k+\alpha d_k\right)+c^Td_k$$
$$=d_k^T\cdot Q\cdot x_k+d_k^T\cdot Q\cdot\alpha\cdot d_k+c^Td_k \qquad :=0\Rightarrow$$
$$\alpha_k=-\frac{d_k^T\cdot\left(Q\cdot x_k+c\right)}{d_k^T\cdot Q\cdot d_k}=-\frac{d_k^T\nabla f(x_k)}{d_k^T\nabla^2 f(x_k)d_k}$$



come si può osservare, il metodo ripete la stessa direzione una volta si e una volta no, le direzioni sono quasi parallele, in questo caso il metodo è molto inefficiente. Una soluzione allo zig-zagging potrebbe essere quella di "scalare" gli assi in modo da far diventare le curve di livello da ellissoidali a sferiche.

Le considerazioni svolte nel caso quadratico, sono ovviamente indicative anche del comportamento locale del metodo nella minimizzazione di funzioni non quadratiche, come si può osservare nella seguente figura



dove è visibile il cosiddetto "effetto banana".

In generale, tanto più sono allungate le curve di livello (ellissi appiattite), tanto peggiore è la direzione dell'antigradiente. Inoltre possiamo dire che

- se $\frac{\lambda_n}{\lambda_1} \approx 1$, le ellissi diventano dei cerchi e la direzione del gradiente passa direttamente per il centro, l'antigradiente punta verso il centro (minimo); il metodo del gradiente convergerà in una sola iterazione;
- se $\frac{\lambda_n}{\lambda_1} >> 1$, le ellissi risultano fortemente eccentriche , sono allungate ("effetto banana"), il metodo convergerà lentamente seguendo una traiettoria a zig-zag; gli assi sono uno più corto l'altro più lungo, i semiassi hanno lunghezza inversamente proporzionale ai valori λ_n e λ_1.

Metodo Heavy Ball o "momentum"[35]

Il metodo del gradiente (senza memoria) utilizza solo le informazioni relative all'iterazione corrente, e quindi non tiene conto in alcun modo delle informazioni derivanti dalle iterazioni precedenti.

Al fine di accelerare la convergenza può essere conveniente introdurre nella regola di aggiornamento del punto corrente termini che riassumano le informazioni di una o più iterazioni precedenti.

Il metodo "Heavy Ball" (con memoria) è un metodo iterativo multistep a "due passi" (nel senso che utilizza informazioni di due iterazioni, quella corrente e quella precedente) in cui l'aggiornamento del punto corrente avviene con la seguente regola

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})$$

Ci limiteremo ad analizzare ed evidenziare i potenziali vantaggi del metodo (con memoria) Heavy Ball rispetto al metodo (senza memoria) del gradiente, al caso di funzione obiettivo quadratica convessa

$$f(x) = \frac{1}{2} x^T Q x \qquad Q \succ 0$$

e scegliendo in modo specifico i valori dei passi α e β abbiamo

$$\begin{aligned}
\|x_{k+1}\| &\le \frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}} \|x_k\| \\
&= \frac{\frac{\sqrt{\lambda_n}}{\sqrt{\lambda_1}} - \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1}}}{\frac{\sqrt{\lambda_n}}{\sqrt{\lambda_1}} + \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1}}} \|x_k\| \\
&= \frac{\sqrt{\rho} - 1}{\sqrt{\rho} + 1} \|x_k\|
\end{aligned}$$

da cui segue che in problemi mal condizionati, fissando un grado di precisione desiderato, il metodo Heavy Ball può richiedere, rispetto al metodo del gradiente, un numero di iterazioni molto più basso.

[35] Heavy Ball (palla pesante): se immaginiamo ogni iterazione come un punto della traiettoria di una palla che cade, i punti portano con sé un momento o una velocità. Metodo *momentum* nella letteratura sulle reti neurali

Teorema (Convergenza finita del metodo del gradiente nel caso quadratico con $\alpha_k = \frac{1}{\lambda_k}$)

Si consideri il problema di minimizzare la funzione quadratica strettamente convessa

$$\min \frac{1}{2} x^T \cdot Q \cdot x + c^T x \qquad Q \succ 0$$

Allora il metodo del gradiente definito dall'iterazione

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

con la scelta dei passi uguali agli inversi degli auto valori di Q

$$\alpha_k = \frac{1}{\lambda_k} \qquad k = 1,\dots,n$$

converge al massimo in n iterazioni al punto di minimo della funzione obiettivo.

Dim.:

ricordando l'espressione del gradiente della funzione quadratica, si può scrivere [36]

$$\begin{aligned}
\nabla f(x_{k+1}) &= Q \cdot x_{k+1} + c \\
&= Q \cdot \left(x_k - \frac{1}{\lambda_k} \nabla f(x_k) \right) + c \\
&= Q \cdot x_k - \frac{1}{\lambda_k} \cdot Q \cdot \nabla f(x_k) + c \\
&= \nabla f(x_k) - \frac{1}{\lambda_k} \cdot Q \cdot \nabla f(x_k) \\
&= \left(I - \frac{1}{\lambda_k} \cdot Q \right) \cdot \nabla f(x_k)
\end{aligned}$$

Applicando ripetutamente la formula precedente, si può porre

$$\nabla f(x_{k+1}) = \left(\prod_{i=1}^{k} \left(I - \frac{1}{\lambda_i} \cdot Q \right) \right) \cdot \nabla f(x_1)$$

Sia v_i un insieme di autovettori linearmente indipendenti di Q, associati agli autovalori λ_i, per cui

$$Q \cdot v_i = \lambda_i \cdot v_i$$

Possiamo esprimere

$$\nabla f(x_1) = \sum_{i=1}^{n} \beta_i \cdot v_i \qquad \beta_i \in R$$

da cui segue che

---

[36] $\nabla f(x_{k+1}) = Q \cdot x_{k+1} + c; \qquad \nabla f(x_k) = Q \cdot x_k + c$.

$$\nabla f\left(x_{k+1}\right)=\left(\prod_{i=1}^{k}\left(I-\frac{1}{\lambda_i}\cdot Q\right)\right)\cdot\sum_{j=1}^{n}\beta_j\cdot v_j$$

All'iterata $n+1$ abbiamo [37]

$$\begin{aligned}\nabla f\left(x_{n+1}\right)&=\left(\prod_{i=1}^{n}\left(I-\frac{1}{\lambda_i}\cdot Q\right)\right)\cdot\sum_{j=1}^{n}\beta_j\cdot v_j\\&=\sum_{j=1}^{n}\beta_j\cdot\left(\prod_{i=1}^{n}\left(I-\frac{1}{\lambda_i}\cdot Q\right)\cdot v_j\right)\\&=\sum_{j=1}^{n}\beta_j\cdot\left(\prod_{i=1}^{n}\left(v_j-\frac{1}{\lambda_i}\cdot Q\cdot v_j\right)\right)\\&=\sum_{j=1}^{n}\beta_j\cdot\left(\prod_{i=1}^{n}\left(v_j-\frac{1}{\lambda_i}\cdot\lambda_j\cdot v_j\right)\right)\qquad=0\quad(\text{per } j=i)\end{aligned}$$

e quindi possiamo concludere che il metodo del gradiente converge in (al più) n passi.

c.v.d.

[37] $Q\cdot v_i=\lambda_i\cdot v_i$, $Q\cdot v_j=\lambda_j\cdot v_j$.