METODO DI NEWTON

Data $f(x)$ continuamente differenziabile, x_k punto di partenza, ci spostiamo di un passo s, se s non è troppo grande l'approssimazione quadratica è buona.

Uso lo sviluppo in serie Taylor:

$$f(x_k + s) = f(x_k) + s^T \nabla f(x_k) + \frac{1}{2} s^T \nabla^2 f(x_k) s + o\left(\|s\|^2\right)$$

Modello quadratico locale:

per valori sufficientemente piccoli di s si può allora pensare di approssimare $f(x_k + s)$ con la funzione quadratica che ha nel punto x_k lo stesso valore della funzione (stesso gradiente, stessa tg, stessa curvatura, Hessiana)

$$q_k(s) = f(x_k) + s^T \nabla f(x_k) + \frac{1}{2} s^T \nabla^2 f(x_k) s$$

$$\nabla q_k(s) = \nabla f(x_k) + \nabla^2 f(x_k) s$$

della matrice Hessiana non sappiamo niente

Se $\nabla^2 f(x_k) \succ 0$ **il punto di minimo di** $q_k(s)$ **è**

$$\nabla q_k(s) = 0 \Rightarrow \nabla f(x_k) + \nabla^2 f(x_k) s$$

$$s_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k) \qquad \text{passo di Newton}$$

Usiamo il minimo di questo modello come predittore di dove andare nella prossima iterazione

Il **metodo di Newton** è definito dall'iterazione:

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

Da ricordare che $\nabla^2 f(x_k) s_k = -\nabla f(x_k)$ è una direzione, quindi $\alpha_k = 1 \quad \forall k$

Nota: Poiché la matrice Hessiana potrebbe NON essere definita positiva la direzione $s_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ potrebbe non essere di discesa.

PROBLEMI DEL METODO DI NEWTON

- Se l'Hessiana è non singolare il metodo potrebbe essere indefinito.
- Se l'Hessiana non è definita positiva potrebbe generare direzioni che non sono di discesa.
- La convergenza globale non è garantita.
- Invertire l'Hessiana ha un costo computazionale elevato e numericamente instabile
- La FO non interviene direttamente.

Teorema di convergenza locale del metodo di Newton

Sia $g(x) = \nabla f(x) \ \wedge \ H(x) = \nabla^2 f(x)$

Se $f \in \mathscr{C}^2$

- $\exists\, x^* : g(x^*) = 0$ esiste un punto stazionario (ottimalità 1° ordine)
- $H(x^*)$ è non singolare cioè invertibile (ipotesi forte)

allora se l'algoritmo parte vicino al punto stazionario, esiste un intorno $B(x^*, \varepsilon) > 0$ tale che se $x_0 \in B$ la successione $\{x_k\}$ generata dal metodo di Newton

$$x_{k+1} = x_k - H(x_k)^{-1} g(x_k)$$

resterà in B e convergerà a x^* con rapidità di convergenza superlineare (se partiamo dalla sferetta NON usciremo più dalla sferetta).

Inoltre, se $\exists\, L > 0$ tale che:

$$\|H(x) - H(y)\| \le L\|x - y\| \quad \forall\, x, y \quad \text{(Hessiana non cambia troppo velocemente)}$$

allora la convergenza è quadratica (è il massimo che si può avere in pratica)

Dim.:

La continuità e la non singolarità di $H(\cdot)$ in x^* implica la non singolarità in $B(x^*, \varepsilon)$ per ε abbastanza piccolo. Inoltre, scegliendo ε abbastanza piccolo, esiste $\mu > 0$ tale che

$$\left\| H(x)^{-1} \right\| \le \mu \quad [1]$$

e, per continuità, esiste $\sigma \in (0,1)$ tale che

$$\|H(x) - H(y)\| \le \frac{\sigma}{\mu} \qquad x, y \in B \quad [2]$$

Premessa: ho trovato una sferetta dove:

1. la matrice H è invertibile dappertutto;
2. la norma dell'inversa non è troppo grande $\le \mu$;
3. la variazione dell'Hessiana non è troppo grande $\le \sigma / \mu$

---

[1] Norme matriciali sono indotte da norme vettori. E' una matrice i cui elementi sono funzioni di x. Se f è continua $\Rightarrow$ sono continui gli auto valori, il determinante…Vuol dire la norma dell'inversa non può essere troppo grande.

[2] La norma è continua, la differenza non è troppo grande.

Richiamo: $\|A\|_2 = \max\left\{ \|Ax\|_2 : \|x\|_2 = 1, x \text{ nel bordo sfera unitaria} \right\} = \sqrt{\lambda_{\max}(A^T A)}$

Devo dimostrare che se siamo nella sfera all'iterazione x_k sarò nella sfera anche all'iterazione x_{k+1}.

Il **metodo di Newton** è definito dall'iterazione $x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ cioè $x_{k+1} = x_k - H(x_k)^{-1} g(x_k)$, tolgo a dx e sx x^* (ottimo)

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - H(x_k)^{-1} g(x_k) \\ &= -H(x_k)^{-1} \cdot \left[-H(x_k)(x_k - x^*) + g(x_k) - g(x^*) \right] \qquad g(x^*) = 0 \text{ per ip. elem. neutro} \end{aligned}$$

$$\left\| -H(x_k)^{-1} \cdot [-H(x_k) \cdot (x_k - x^*) + g(x_k) - g(x^*)] \right\| \le \left\| H(x_k)^{-1} \right\| \cdot \left\| -H(x_k) \cdot (x_k - x^*) + g(x_k) - g(x^*) \right\|$$ [3]

Considero la distanza di x_{k+1} da x_k, che non dovrà essere più grande di ε :

$$\begin{aligned} \left\| x_{k+1} - x^* \right\| &\le \left\| H(x_k)^{-1} \right\| \cdot \left\| -H(x_k) \cdot (x_k - x^*) + g(x_k) - g(x^*) \right\| \\ &\le \mu \left\| -H(x_k) \cdot (x_k - x^*) + g(x_k) - g(x^*) \right\| \end{aligned}$$

Poiché $g()$ è differenziabile, la differenza dei gradienti $g(x_k) - g(x^*)$ si può porre a

$$g(x_k) - g(x^*) = \int_0^1 H[x^* + \lambda(x_k - x^*)] \cdot (x_k - x^*) d\lambda$$ [4]

infatti dal teorema fondamentale del calcolo integrale (Torricelli-Barrow):[5]

---

[3] $\|x \cdot y\| \le \|x\| \cdot \|y\|$

[4] è lo stesso λ visto per le funzioni convesse. Si consideri il segmento $x_k \overline{\qquad} x^*$, $H(x^* + \lambda(x_k - x^*))$, $\lambda = 0 \Rightarrow H(x^*)$, $\lambda = 1 \Rightarrow H(x_k)$, funzione vettoriale, la guardo rispetto a λ che è uno scalare monodimensionale.

[5] Se la funzione $f(x)$ è continua nell'intervallo $[a,b]$, la sua funzione integrale $F(x) = \int_a^x f(t)dt$ è derivabile e la sua derivata in ogni punto x di $[a,b]$ è uguale alla funzione f nello stesso punto, cioè $F'(x) = f(x) \quad \forall x \in [a,b]$. La funzione integrale F è una primitiva della funzione f . Poiché tutte le primitive di una funzione differiscono per una costante, se $\varphi(x)$ è un'altra primitiva di $f(x)$, dovrà essere $\varphi(x) = F(x) + k$ ovvero $\varphi(x) = \int_a^x f(t)dt + k$. In particolare se $x = a \quad \varphi(a) = \int_a^a f(t)dt + k$ cioè $\varphi(a) = k$. Allora la funzione $\varphi(x)$ può essere scritta come $\varphi(x) = \int_a^x f(t)dt + \varphi(a)$, da cui ricaviamo che $\int_a^x f(t)dt = \varphi(x) - \varphi(a)$. In particolare, se $x = b$, si ha la formula di Newton-Leibniz $\int_a^b f(t)dt = \varphi(b) - \varphi(a)$ che afferma che **l'integrale definito fra a e b di una funzione continua $f(x)$ è la differenza fra i valori assunti da una generica primitiva di f** nei punti a **e** b. Esempio: $\int_0^3 (x^3 + 1)dx = \left[\frac{1}{4}x^4 + x + c \right]_0^3 = \left(\frac{81}{4} + 3 + c \right) - (c) = \frac{93}{4}$

$$\int_0^1 f'(\lambda)d\lambda = f(1) - f(0)$$

$$f(1) = f(0) + \int_0^1 f'(\lambda)d\lambda$$

Ad ogni λ corrisponde un vettore in R^n, quindi ho un vettore di derivate, è una funzione di funzione (funzione composta). Dato λ passo in R^n con $x^* + \lambda(x_k - x^*)$ e poi passo a g.

$$\begin{aligned} f(\lambda) &= g(\lambda x_k + (1-\lambda)x^*) \\ &= g(x^* + \lambda(x_k - x^*)) \\ f(1) &= g(x_k) \\ f(0) &= g(x^*) \end{aligned}$$

$$g(x_k) - g(x^*) = \int_0^1 H(x^* + \lambda(x_k - x^*)) \cdot (x_k - x^*)d\lambda$$ [6]

Quindi

$$\begin{aligned} & -H(x_k)\cdot(x_k - x^*) + g(x_k) - g(x^*) = \\ & -H(x_k)\cdot(x_k - x^*) + \int_0^1 H(x^* + \lambda(x_k - x^*))\cdot(x_k - x^*)d\lambda = \\ & \int_0^1 (H(x^* + \lambda(x_k - x^*))\cdot(x_k - x^*) - H(x_k)\cdot(x_k - x^*))d\lambda = \\ & \int_0^1 (H(x^* + \lambda(x_k - x^*)) - H(x_k))\cdot(x_k - x^*)d\lambda \end{aligned}$$

In definitiva:

$$\begin{aligned} \left\| x_{k+1} - x^* \right\| &\le \mu \left\| \int_0^1 (H(x^* + \lambda(x_k - x^*)) - H(x_k))\cdot(x_k - x^*)d\lambda \right\| \\ &\le \mu \int_0^1 \left\| (H(x^* + \lambda(x_k - x^*)) - H(x_k)) \right\| \cdot \left\| (x_k - x^*) \right\| d\lambda \end{aligned}$$ [7]

---

[6] moltiplicato per $(x_k - x^*)$ perché è la derivata di $\lambda(x_k - x^*)$ rispetto a λ

[7] $\left\| \int \quad \right\| \le \int \| \quad \|$

siccome i 2 punti $x^* + \lambda(x_k - x^*)$ e x_k stanno nella sferetta e per ipotesi sappiamo che $\|H(x) - H(y)\| \le \frac{\sigma}{\mu}$ $x, y \in B$, in più posso portare fuori l'integrale $\|x_k - x^*\|$ perché non dipende da λ, segue che

$$\begin{aligned}\|x_{k+1} - x^*\| &\le \mu \|x_k - x^*\| \int_0^1 \frac{\sigma}{\mu} d\lambda \\ &\le \mu \|x_k - x^*\| \cdot \frac{\sigma}{\mu} \cdot [\lambda]_0^1 \\ &\le \sigma \|x_k - x^*\|\end{aligned}$$

Ciò implica $x_{k+1} \in B(x^*, \varepsilon)$, perché $\sigma \in (0,1)$ e $\|x_k - x^*\|$ è limitata, quindi tutti gli x_{k+1} non "scappano" mai fuori dalla sferetta.

Per induzione, se $x_0 \in B(x^*, \varepsilon)$ si ha $x_k \in (x^*, \varepsilon)$ $\forall k$. Possiamo allora applicare iterativamente $\|x_{k+1} - x^*\| \le \sigma \|x_k - x^*\|$ a partire da x_0, ottenendo

$$\|x_k - x^*\| \le \sigma^k \|x_0 - x^*\|$$ [8],

per cui, essendo $\sigma < 1$, si ha che la successione dei punti $\{x_k\}$ converge a x^* (addirittura non usiamo sottosuccessioni).[9] Il risultato che ottengo è la convergenza dell'algoritmo.

La convergenza è superlineare:

inoltre dalla limitazione

$$\|x_{k+1} - x^*\| \le \mu \int_0^1 \|(H(x^* + \lambda(x_k - x^*)) - H(x_k))\| \cdot \|(x_k - x^*)\| d\lambda,$$

supponendo $x_k \ne x^*$ $\forall k$, dividendo per $\|x_k - x^*\|$ otteniamo:

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \le \mu \int_0^1 \|(H(x^* + \lambda(x_k - x^*)) - H(x_k))\| d\lambda$$ [10]

Siccome in precedenza abbiamo dimostrato che il metodo converge, cioè abbiamo dimostrato che x_k è sempre più vicino all'ottimo x^*, andando a limite per $k \to \infty$, la differenza tra

---

[8] $\|x_k - x^*\|$, la norma dell'errore, converge a 0, cioè $\sigma^k \to 0$ quando $k \to \infty$

[9] Ad ogni passo k riduco di σ $\frac{1}{2}; \frac{1}{4}; \frac{1}{8}; \frac{1}{16}$.....

[10] Rapporto tra distanze x_{k+1} e x_k dall'ottimo

$H(x^* + \lambda(x_k - x^*))$ e $H(x_k)$ tende a zero, cioè $H(x^* + \lambda(x_k - x^*))$ e $H(x_k)$ si avvicinano sempre più. Segue allora, per la continuità di H :

$$\lim_{k\to\infty} \frac{\left\|x_{k+1} - x^*\right\|}{\left\|x_k - x^*\right\|} = 0$$

il che prova che la rapidità di convergenza è Q-superlineare.

La convergenza è quadratica:

infine, se vale l'ipotesi che $\exists L > 0$ tale che:

$$\left\|H(x) - H(y)\right\| \le L\left\|x - y\right\| \quad \forall\, x, y \quad \text{(Hessiana non cambia troppo velocemente)}$$

dalla

$$\begin{aligned}
\left\|x_{k+1} - x^*\right\| &\le \mu \int_0^1 \left\|(H(x^* + \lambda(x_k - x^*)) - H(x_k))\right\| \cdot \left\|(x_k - x^*)\right\| d\lambda \\
&\le \int_0^1 L\left\|x^* + \lambda(x_k - x^*) - x_k\right\| \cdot \left\|x_k - x^*\right\| d\lambda = L\int_0^1 \left\|x^* + \lambda x_k - \lambda x^* - x_k\right\| \cdot \left\|x_k - x^*\right\| d\lambda \\
&= L\int_0^1 \left\|- x_k(1-\lambda) + x^*(1-\lambda)\right\| \cdot \left\|x_k - x^*\right\| d\lambda \\
&= L\int_0^1 \left\|(1-\lambda)\cdot(x^* - x_k)\right\| \cdot \left\|x_k - x^*\right\| d\lambda \\
&= L\left\|x_k - x^*\right\|^2 \left[\lambda - \frac{\lambda^2}{2}\right]_0^1 \\
&= \frac{L}{2}\left\|x_k - x^*\right\|^2
\end{aligned}$$

In definitiva si ottiene:

$$\left\|x_{k+1} - x^*\right\| \le \frac{L}{2}\left\|x_k - x^*\right\|^2,$$

per cui la rapidità di convergenza è quadratica.

Quindi abbiamo un algoritmo con convergenza locale. Se partiamo vicino all'ottimo l'algoritmo è estremamente veloce.

c.v.d.

OSSERVAZIONI

Il metodo di Newton nasce per risolvere equazioni, $\nabla f(x) = 0$ (gradiente = 0). La funzione obiettivo non gioca un ruolo importante nel metodo di Newton.

Il metodo di Newton quando è usato come un metodo di soluzione di un sistema di equazioni non lineari $F(x) = 0$ diventa:

$$x_{k+1} = x_k - J(x_k)^{-1} F(x_k)$$

dove $J(x)$ è la matrice Jacobiana di $F(x)$, cioè la matrice delle derivate parziali.

Applicazione: calcolare la radice quadrata

$$F(x) = x^2 - a \qquad J(x) = 2x \qquad J(x)^{-1} = \frac{1}{2x}$$

$$\begin{aligned} x_{k+1} &= x_k - \frac{(x_k^2 - a)}{2x_k} \\ &= \frac{2x_k^2 - x_k^2 + a}{2x_k} \\ &= \frac{x_k^2 + a}{2x_k} \\ &= \frac{1}{2}\left(x_k + \frac{a}{x_k} \right) \end{aligned}$$ [11]

$\sqrt{2} = 33$

$$x_1 = \frac{1}{2}\left(33 + \frac{2}{33} \right) = 16{,}53$$

$$x_2 = \frac{1}{2}\left(16{,}53 + \frac{2}{16{,}53} \right) = 8{,}3$$

$$x_3 = \frac{1}{2}\left(8{,}3 + \frac{2}{8{,}3} \right) = 4{,}2$$

$$x_4 = \frac{1}{2}\left(4{,}2 + \frac{2}{4{,}2} \right) = 2{,}3$$

$$x_5 = \frac{1}{2}\left(2{,}3 + \frac{2}{2{,}3} \right) = 1{,}5$$

$$x_6 = \frac{1}{2}\left(1{,}5 + \frac{2}{1{,}5} \right) = 1{,}41$$

$$x_7 = \frac{1}{2}\left(1{,}41 + \frac{2}{1{,}41} \right) = 1{,}4142$$ [12]

---

[11] $x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k} \right)$ è l'iterazione presente nelle calcolatrici

[12] da insegnare alle elementari

Globalizzazione del metodo di Newton

Esistono varianti convergenti del metodo di Newton. Si può rimediare al difetto principale del metodo di Newton. Qui diamo solo un cenno e si chiama modifica globalmente convergente del metodo di Newton.

Def.: sia $f : R^n \to R$ due volte continuamente differenziabile e supponiamo che l'insieme di livello $L_0 = \{x \in R^n : f(x) \le f(x_0)\}$ sia limitato. Diremo che l'algoritmo caratterizzato dall'iterazione

$$x_{k+1} = x_k + s_k$$

è una modifica globalmente convergente del metodo di Newton se valgono le seguenti proprietà:

- ogni punto limite di $\{x_k\}$ è un punto stazionario di f in L_0;
- nessun punto limite di $\{x_k\}$ è un punto di massimo locale per f;
- se $\{x_k\}$ converge a un punto di minimo locale x^* di f e $\nabla^2 f(x^*)$ soddisfa le ipotesi del teorema di convergenza di Newton, esiste un k^* tale che, $\forall k \ge k^*$ la direzione s_k coincide con la direzione di Newton, ossia risulta

$$s_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$ [13]

Le modifiche globalmente convergenti possono essere ottenute attraverso:

- metodi di ricerca unidimensionale modificati in cui la direzione di Newton è "deviata"[14];
- metodi di tipo trust region (regione di confidenza), dove viene usata la minimizzazione di una approssimazione quadratica di f in un intorno sferico di centro x_k.

I metodi basati su tecniche di ricerca unidimensionale sono classificati in:

- ***metodi di tipo Newton ibrido***, basati sulla combinazione del metodo di Newton con un metodo globalmente convergente, come il metodo del gradiente (ogni tanto, 10 iterazioni, faccio una iterazione col gradiente);
- ***metodi basati su modifiche della matrice Hessiana*** basati sulla modifica dell'Hessiana per ottenere una matrice definita positiva (invece di $-(H(x_k))^{-1}$ uso una matrice vicina a H ma $\succ 0$);
- ***metodi di ricerca curvilinea*** (invece della ricerca unidimensionale).

---

[13] nel metodo di Newton non c'è un passo, passo e direzione si scelgono in un colpo solo.

[14] lavora sul passo, per esempio Armijo. Tutti i metodi basati sul gradiente, posso moltiplicare il gradiente per una matrice definita positiva $\succ 0$ per avere una direzione di discesa. Posso fare anche $\nabla + cI$ dove c è una costante e I è la matrice identità, la costante ha l'effetto di spostare gli auto valori della matrice.

Lunghezza del passo nei metodi tipo Newton convergenti

In tutti i casi, perché siano soddisfatte le condizioni della modifica globalmente convergente del metodo di Newton, un requisito fondamentale è che la ricerca unidimensionale fornisca il passo unitario (come nell'iterazione di Newton), quando la direzione di ricerca venga scelta coincidente con la direzione di Newton, al posto di

$$-\left(H\left(x_k\right)\right)^{-1}\nabla f\left(x_k\right)$$

potrei mettere

$$-\alpha_k\left(H\left(x_k\right)\right)^{-1}\nabla f\left(x_k\right)$$

quindi

$$x_{k+1}=x_k-\alpha_k\left(H\left(x_k\right)\right)^{-1}g\left(x_k\right)$$

si pone $\alpha_k \approx 1$