VELOCITA' DI CONVERGENZA

Impareremo a costruire e valutare algoritmi di ottimizzazione per il caso non vincolato. Immaginiamo di avere un algoritmo che genera una successione di punti $x_0, x_1, ..., x_k$, chiamo $\{x_k\}$ la successione. Ci accontenteremo di punti stazionari $(\nabla = 0)$, supponiamo che la successione $\{x_k\}$ converga a x^* (ottimo del mio problema).

- Se $\exists \sigma \in [0,1)$ tale che

$$\left\| x_{k+1} - x^* \right\| \le \sigma \left\| x_k - x^* \right\| \text{ [1]}$$

per k abbastanza grande, allora la successione **converge linearmente** (almeno).[2]

- Dividiamo 1° e 2° membro per $\left\| x_k - x^* \right\|$ $\qquad x_k \ne x^*$

Se al limite il rapporto

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

è zero (il moto è accelerato), allora la successione **converge superlinearmente** (è accelerata).

- Se

$$\left\| x_{k+1} - x^* \right\| \le \sigma \left\| x_k - x^* \right\|^p \text{ [3]}$$

allora la successione **converge con ordine p**. Se $p = 2$: **quadraticamente**.

Esempi in R:

- $\frac{1}{k}$ converge linearmente a 0 $\quad 1; \frac{1}{2}; \frac{1}{3}; \frac{1}{4}; \frac{1}{5}; ...; \frac{1}{k}$ $\quad$ velocità lineare
- $\frac{1}{k^2}$ converge linearmente a 0 $\quad 1; \frac{1}{4}; \frac{1}{9}; \frac{1}{16}; \frac{1}{25}; ...; \frac{1}{k^2}$ $\quad$ velocità lineare
- $\frac{1}{2^k}$ converge linearmente a 0 $\quad 1; \frac{1}{2}; \frac{1}{4}; \frac{1}{8}; \frac{1}{16}; \frac{1}{32}; \frac{1}{64}; ...; \frac{1}{2^k}$ $\quad$ velocità
- $\frac{1}{k^k}$ converge superlinearmente a 0 $\quad 1; \frac{1}{4}; \frac{1}{27}; \frac{1}{256}; ...; \frac{1}{k^k}$ $\quad$ molto rapida

---

[1] $\left\| x_{k+1} - x^* \right\|$ l'errore al passo $k+1$, cioè quello che ci manca per arrivare a convergenza, è limitato da una costante $\sigma < 1$, $\left\| x_k - x^* \right\|$ è l'errore al passo k.

[2] è legato alla rapidità, alla velocità con cui si avvicina a x^* (punto ottimo, punto limite).

[3] $\left\| x_k - x^* \right\|^p$ la distanza fra x_k e l'ottimo è elevato a p. Più grande è p, più veloce è l'algoritmo. Raramente si trova $p \ge 2$, Newton ha $p = 2$.

- $\frac{1}{2^{k^k}}$ converge quadraticamente a 0 $\quad \frac{1}{2};\frac{1}{16};\frac{1}{512};...;\frac{1}{2^{k^k}}$ $\qquad$ Metodo di Newton

Funzione Coerciva

Il teorema di Weierstrass[4] (esistenza di un punto di minimo globale di un problema di ottimizzazione) si applica in modo diretto solamente a problemi *vincolati* in cui l'insieme ammissibile $X \subset R^n$ è compatto ovvero è chiuso e limitato. Il teorema dà una condizione *sufficiente* di esistenza (posso avere un minimo globale anche se l'insieme non è compatto), ma non *necessaria*: ad esempio, il problema

$$\min_{\substack{x \in R \\ x \geq 0}} x^2$$

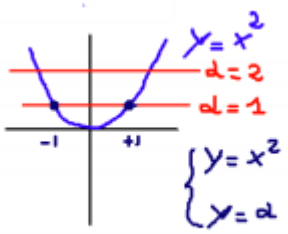
ha la soluzione globale $x^* = 0$, anche se l'insieme ammissibile, costituito da tutto il semiasse $x \geq 0$ è chiuso ma non limitato. Analogamente, il problema

$$\min_{\substack{x \in R \\ -1 < x < +1}} x^2$$

ha la soluzione globale $x^* = 0$, anche se l'insieme ammissibile è costituito dall'intervallo $(-1;+1)$, limitato ma non chiuso.

Nel caso non vincolato, in cui $X = R^n$ non è possibile applicare direttamente il teorema di Weierstrass. Allo scopo di definire condizioni sufficienti di esistenza, si definisce l'*insieme di livello* L_α della funzione obiettivo $f(x)$ relativo a un punto α come l'insieme dei punti in cui la funzione ha valore minore o uguale a quello di α:

$$L_\alpha = \{x \in R^n : f(x) \leq \alpha\}$$



Una funzione $f(x)$ si dice ***radialmente illimitata***, o, anche, ***coerciva***, se gode della proprietà di tendere a $+\infty$ quando ci si allontana dall'origine in R^n, cioè se risulta:

$$\lim_{\|x\| \to \infty} f(x) = +\infty$$

---

[4] Sia $f : R^n \to R$ una funzione continua e sia $X \subset R^n$ un insieme chiuso e limitato (compatto). Allora esiste un punto di minimo globale di f in X.

Condizione di Coercività (C.N.S.)

Sia $f : R^n \to R$ una funzione continua. C.N.S. perché tutti gli insiemi di livello di livello di f siano compatti è che f sia coerciva, ossia che, per ogni successione $\{x_k\}$ tale che

$$\lim_{k\to\infty} \|x_k\| = \infty ,$$

deve essere

$$\lim_{k\to\infty} f(x_k) = \infty$$

Esempi:

- $f(x) = \|x\|^q \qquad q > 0$ è coerciva.
- $f(x) = \frac{1}{2} x^T Q x + c^T x$ è coerciva.

Convergenza di metodi non vincolati con ricerche unidimensionali

Gli algoritmi che studieremo sono procedimenti iterativi per la determinazione di *punti stazionari* di f, ossia punti in cui $\nabla f(x) = 0$, possono essere descritti dal seguente schema nella maggior parte degli algoritmi di ottimizzazione (non tutti):

Data: $x_0 \in R^n$ (punto di "innesco"iniziale, soluzione iniziale)[5]. Poni $k = 0$.

While una *stopping rule* [6]non è soddisfata esegui (finché il gradiente non ha la norma vicino a zero)

Calcola una *direzione di ricerca* $d_k \in R^n$ (direzione lungo ci si sposta).

Determina un *passo* $\alpha_k \in R$ lungo d_k (lungo la direzione di quanto spostarci).

Determina il nuovo punto $x_{k+1} = x_k + \alpha_k d_k$ (prossima iterata)[7].

Poni $k = k + 1$.

End While

Esistenza di punti limite (sequenze monotone)

L'analisi delle proprietà di convergenza consiste nello studio del comportamento delle successioni $\{x_k\}, \{f(x_k)\}, \{\nabla f(x_k)\}$. L'esistenza di punti limite di $\{x_k\}$ può essere assicurata imponendo che tutti i punti della successione siano in un insieme compatto. Supponendo che l'insieme di livello L_0 sia

---

[5] Metodo basato sulla ricerca lineare (line search).

[6] Criterio di arresto. Fissando, ad esempio, un $\varepsilon > 0$ sufficientemente piccolo e arrestando l'algoritmo quando $\|f(x_k)\| \le \varepsilon$.

[7] Nei metodi Trust Region, passo e direzione vengono effettuati in un unico iterato.

compatto, una condizione *sufficiente* perché i punti x_k restino in un insieme compatto è che la successione dei valori della funzione obiettivo $\{f(x_k)\}$ sia *monotona non crescente*.[8]

Teorema (esistenza di punti limite)

Sia f continua.

Sia $\{x_k\}$ la successione prodotta dall'algoritmo.

Si supponga che $f(x_{k+1}) \leq f(x)$, (abbiamo un algoritmo di discesa, la funzione non cresce).

Inoltre (ipotesi forte, difficile da ottenere), supponiamo che $L_0 = \{x : f(x) \leq f(x_0)\}$ sia compatto (conosco x_0, il punto di partenza, che sta su un insieme di livello compatto, cioè chiuso e limitato)[9].

Allora:

1. $x_k \in L_0$, tutti i punti generati x_k stanno sulle curve di livello;
2. $\exists$ almeno un punto limite, la successione $\{x_k\}$ ha almeno un punto limite (non che converge);
3. ogni punto limite di $\{x_k\}$ appartiene a L_0;
4. la successione $\{f(x_k)\}$ converge.

Dim.:

1. questo è banale, siccome abbiamo ipotizzato che la funzione non cresca, tutti i punti dopo il primo avranno il valore della funzione $\leq$ della funzione nel punto iniziale, resterò nella curva di livello iniziale;
2. sicuramente esiste una sottosuccessione convergente, questo deriva dalla compatezza, siccome tutti i punti stanno su un insieme compatto posso sempre trovare una sottosuccessione convergente;[10]
3. 2 e 3 vanno insieme, L_0 compatto è difficile da garantire;
4. se è vero che l'algoritmo genera una successione di punti monotona, allora la successione dei valori della funzione converge. Questo perché la funzione è continua e la possiamo restringere a L_0, ammette sempre minimo, è limitata dal di sotto, e dall'Analisi segue che una successione di numeri reali monotona e limitata dal di sotto deve avere un limite e quindi converge.

---

[8] Ci sono algoritmi (non monotoni) dove non vale questa proprietà.

[9] Se la funzione è coerciva ogni curva di livello è compatta.

[10] Non è facile, in generale, dimostrare che la successione generata da un algoritmo di ottimizzazione ammette un limite. Tuttavia se esiste una successione $\{x_k\}$ tale che $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$, e esiste un punto limite $\bar{x}$ di $\{x_k\}$, allora, per ogni fissato $N > 0$ si ha che tutte le sottosuccessioni $\{x_{k+j}\}$, per $j = 1,\ldots,N$ convergono a $\bar{x}$.

Funzione di forzamento

Per esprimere varie condizioni di convergenza, introduciamo il concetto di *funzione di forzamento*[11] .

Una funzione $\sigma : R^+ \to R^+$ è una **funzione di forzamento** se, per ogni successione $\{t_k\} \geq 0$, si ha che

$$\lim_{k\to\infty} \sigma(t_k) = 0 \Rightarrow \lim_{k\to\infty} t_k = 0$$

Cioè tutte le volte che $f \to 0$, il suo argomento $\to 0$, è una funzione che ***forza*** il suo argomento ad andare a 0 quando la funzione stessa $\to 0$.

Esempi:

$$t^2 \to 0 \Rightarrow t \to 0$$
$$t^3 \to 0 \Rightarrow t \to 0$$

Utilizziamo la funzione di forzamento nel Teorema di convergenza globale (da non confondere con l'ottimo globale di un problema). Si chiama convergenza globale perché garantisce la convergenza da qualunque punto io parta. Con ipotesi deboli (tranne la compattezza) ho la convergenza globale, cioè l'algoritmo ci conduce a un punto stazionario.[12]

Teorema (condizione sufficiente di convergenza globale)

Sia $f : R^n \to R$ una funzione continuamente differenziabile su R^n, sia $x_0 \in R^n$ il punto di partenza (innesco), supponiamo che l'insieme di livello $L_0 = \{x : f(x) \leq f(x_0)\}$ sia compatto (ipotesi forte). Sia $\{x_k\}$ la successione generata dall'algoritmo e si supponga che $d_k \neq 0$ ogni volta che $\nabla f(x_k) \neq 0$ (l'algoritmo mi dirà quale direzione scegliere e di quanto spostarmi). Supponiamo inoltre che valgono le condizioni:

1. $f(x_{k+1}) \leq f(x_k) \quad \forall k$, cioè l'algoritmo è monotono, la scelta del nuovo punto non peggiora la funzione obiettivo;
2. se $\nabla f(x_k) \neq 0 \quad \forall k$, si ha (se $\nabla = 0$ l'algoritmo si ferma, non succede quasi mai)

$$\lim_{k\to\infty} \frac{d_k^T}{\|d_k\|} \nabla(f(x_k)) = 0$$ [13]

---

[11] Nel passato c'erano tantissime proprietà di convergenza, alla fine si è visto che si potevano racchiudere utilizzando una generalizzazione di *funzione di forzamento*.

[12] Il metodo di Newton non gode di questa proprietà.

[13] 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 che il vettore $\alpha \cdot v$ abbia norma 1, dobbiamo prendere $\alpha = \frac{1}{\|v\|}$, infatti con tale scelta di α

3. esiste una funzione di forzamento $\sigma : R^+ \to R^+$ tale che, per $d_k \neq 0$ si ha

$$\frac{\left| d_k^T \nabla f(x_k) \right|}{\| d_k \|} \geq \sigma\left(\| \nabla f(x_k) \| \right)$$ [14]

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

1. esistono punti limite; ogni punto limite $\{x_k\}$ sta nell'insieme di livello iniziale L_0 e $\{f(x_k)\}$ ha limite (è il Teorema precedente sull'esistenza dei punti limite, le ipotesi sono uguali);
2. $\lim_{k \to \infty} \nabla f(x_k) = 0$, la successione dei gradienti converge a zero;
3. ogni punto limite $\bar{x}$ di $\{x_k\}$ soddisfa $\nabla f(\bar{x}) = 0$, limite di ogni sottosuccessione è un punto stazionario.

Dim.:

La 1ª proprietà si dimostra attraverso il Teorema sull'esistenza dei punti limite. Siccome abbiamo ipotizzato che la funzione non cresca, tutti i punti dopo il primo avranno il valore della funzione $\leq$ della funzione nel punto iniziale, resterò nella curva di livello iniziale.

Se è vero che l'algoritmo genera una successione di punti monotona, allora la successione dei valori della funzione converge. Questo perché la funzione è continua e la possiamo restringere a L_0, ammette sempre minimo, è limitata dal di sotto, e dall'Analisi segue che una successione di numeri reali monotona e limitata dal di sotto deve avere un limite e quindi converge.

Dimostriamo la tesi n° 2, per ipotesi

$$\begin{aligned} 0 &= \lim_{k \to \infty} \frac{d_k^T}{\| d_k \|} \nabla\left(f(x_k) \right) \\ &= \lim_{k \to \infty} \frac{\left| d_k^T \nabla\left(f(x_k) \right) \right|}{\| d_k \|} \\ &\geq \lim_{k \to \infty} \sigma\left(\| \nabla f(x_k) \| \right) = 0 \quad \Rightarrow \{ \nabla f(x_k) \} \to 0 \end{aligned}$$ [15]

---

abbiamo $\left\| \frac{1}{\|v\|} \cdot v \right\| = \frac{1}{\|v\|} \cdot \|v\| = 1$. Se per esempio abbiamo il vettore $v(1,1)$; il vettore normalizzato è $\frac{v}{\|v\|} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$. Infatti $\left\| \frac{1}{\sqrt{2}} \cdot v \right\| = \frac{1}{\sqrt{2}} \cdot \|v\| = \frac{1}{\sqrt{2}} \cdot \sqrt{2} = 1$. $d_k^T \nabla(f(x_k))$ è legato alla derivata direzionale, lungo d stiamo trovando un ottimo locale.

[14] $\frac{\left| d_k^T \nabla f(x_k) \right|}{\| d_k \|} \to 0$ ma resta $\geq \sigma\left(\| \nabla f(x_k) \| \right)$. Questo è un modo per dire che non ci piace la perpendicolarità.

Dimostriamo la tesi n° 3. Data la sottosuccessione $\{x_k\}_{k\in K} \to \bar{x}$, per la continuità di $\nabla f(x_k)$, $\lim_{k\in K, k\to\infty} \nabla f(x_k) = \nabla f(\bar{x}) = 0$, il gradiente del punto limite è zero.

c.v.d.

Considerazioni sulle ipotesi del Teorema (condizione sufficiente di convergenza globale)

Le ipotesi enunciate nel teorema precedente possono essere ricondotte a ipotesi sulla direzione di ricerca d_k e sulla scelta del passo α_k.

1ª ipotesi (condizione di monotonicità $f(x_{k+1}) \le f(x_k) \quad \forall k$): nella maggior parte degli algoritmi la direzione d_k viene scelta come *direzione di discesa*, questo vuol dire che, per valori abbastanza piccoli di α risulta $f(x_k + \alpha d_k) < f(x_k)$.

Una condizione sufficiente perché d_k sia di discesa è che risulti

$$d_k^T \nabla f(x_k) < 0 .$$

Se d_k è una direzione di discesa, la condizione $f(x_{k+1}) \le f(x_k) \quad \forall k$ può essere soddisfatta con una scelta opportuna del passo α_k (un buon algoritmo è tale se ci dice di quanto vogliamo spostarci lungo una direzione). In particolare, per valori sufficientemente piccoli di a_k si può assicurare che risulti

$$f(x_{k+1}) < f(x_k) .$$

2ª ipotesi: se $\nabla f(x_k) \ne 0 \quad \forall k$, si ha $\lim_{k\to\infty} \frac{d_k^T}{\|d_k\|} \nabla(f(x_k)) = 0$, ha un significato più tecnico; essa impone che la derivata direzionale lungo la direzione normalizzata $\frac{d_k^T}{\|d_k\|}$ converga a zero e viene utilizzata per assicurare che tenda a zero il primo membro di $\frac{\left|d_k^T \nabla f(x_k)\right|}{\|d_k\|} \ge \sigma\left(\|\nabla f(x_k)\|\right)$. Il soddisfacimento di questa ipotesi può essere ottenuto attraverso la ricerca unidimensionale per il calcolo di α_k (Metodo di Armijo).

---

[15] $0 = \lim_{k\to\infty} \frac{\left|d_k^T \nabla\left(f(x_k)\right)\right|}{\|d_k\|}$, è 0 anche il limite del valore assoluto. $0 \ge \lim_{k\to\infty} \sigma\left(\|\nabla f(x_k)\|\right) = 0$, se $\sigma \to 0$ il suo argomento $\to 0$ (definizione di funzione di forzamento).

3ª ipotesi: esiste una funzione di forzamento $\sigma : R^+ \to R^+$ tale che, per $d_k \neq 0$ si ha

$$\frac{\left|d_k^T \nabla f(x_k)\right|}{\|d_k\|} \geq \sigma\left(\|\nabla f(x_k)\|\right),$$

questa ipotesi pone delle limitazioni sulla scelta di d_k. In particolare, se si assume come funzione di forzamento la funzione lineare $\sigma(t) = ct$, con $c > 0$ (se $ct \to 0$ allora $t \to 0$), e si suppone $d_k^T \nabla f(x_k) < 0$, possiamo scrivere

$$\frac{\left|d_k^T \nabla f(x_k)\right|}{\|d_k\|} \geq c\|\nabla f(x_k)\|$$

$$-\frac{d_k^T \nabla f(x_k)}{\|d_k\|} \geq c\|\nabla f(x_k)\|$$ [16].

Divido per $\|\nabla f(x_k)\|$ (immaginiamo che non sono arrivato a convergenza):

$$-\frac{d_k^T \nabla f(x_k)}{\|d_k\|\|\nabla f(x_k)\|} \geq c$$

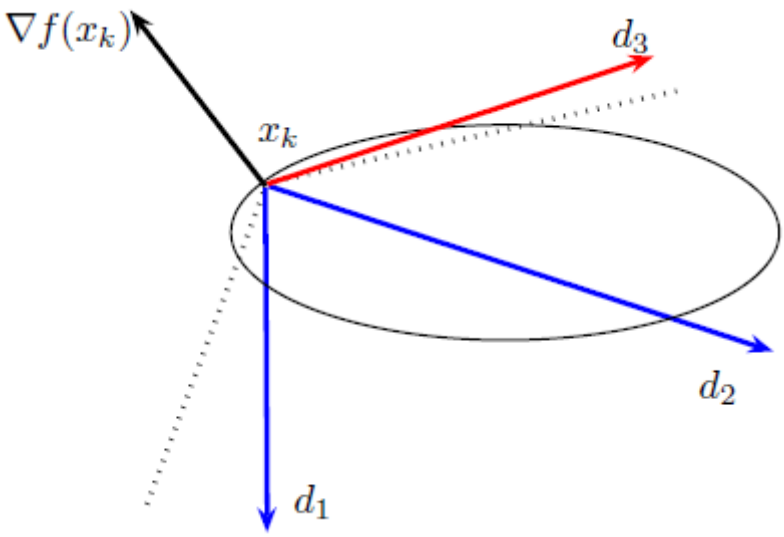
Da un punto di vista geometrico, $-\frac{d_k^T \nabla f(x_k)}{\|d_k\|\|\nabla f(x_k)\|} \geq c$ [17] equivale a richiedere che il coseno dell'angolo tra la direzione d_k e la direzione dell'antigradiente $-\nabla f(x_k)$ si mantenga sempre maggiore di una quantità indipendente da k. Ciò evita che, al crescere di k, la direzione d_k possa tendere a divenire ortogonale al gradiente. Questo significa che il metodo che stiamo considerando genera direzioni che sono discoste dalla ortogonalità con il gradiente, cioè non si avvicinano mai a direzioni normali con il gradiente. In particolare, per $d_k = -\nabla f(x_k)$, la disuguaglianza precedente è soddisfatta per $c = 1$.

---

[16] Se d è di discesa $\left|d_k^T \nabla f(x_k)\right| = -d_k^T \nabla f(x_k) = d_k^T \cdot \left(-\nabla f(x_k)\right)$. $-\nabla f(x_k)$ è l'antigradiente.

[17] In generale $\frac{v^T \cdot w}{\|v\| \cdot \|w\|} = \cos\alpha$.

Il significato geometrico di $-\frac{d_k^T \nabla f(x_k)}{\|d_k\| \|\nabla f(x_k)\|} \geq c$, nota come *condizione d'angolo* la possiamo vedere nella seguente figura



Nella figura si suppone che le direzioni che soddisfano la condizione d'angolo, applicate in x_k, devono rimanere entro il cono delimitato dalle linee tratteggiate (che non variano con k), come le direzioni d_1 e d_2. La direzione d_3 non soddisfa la condizione d'angolo, nonostante sia una direzione di discesa.

Ipotesi: compattezza dell'insieme di livello iniziale (difficile da verificare perché non la possiamo tracciare): per ipotesi l'insieme di livello iniziale L_0 è compatto, tuttavia in molti casi questa ipotesi può non essere soddisfatta, oppure può non essere possibile stabilire a priori se L_0 è compatto (non è limitato) in corrispondenza a una funzione obiettivo assegnata. Con le stesse ipotesi del Teorema precedente (condizione sufficiente di convergenza globale), omettendo l'ipotesi di compattezza, abbiamo il seguente teorema:

Teorema (condizione sufficiente di convergenza globale bis)

Sia $f: R^n \to R$ una funzione continuamente differenziabile su R^n, sia $x_0 \in R^n$ il punto di partenza (innesco). Sia $\{x_k\}$ la successione generata dall'algoritmo e si supponga che $d_k \neq 0$ ogni volta che $\nabla f(x_k) \neq 0$. Supponiamo inoltre che valgono le condizioni:

1. $f(x_{k+1}) \leq f(x_k) \quad \forall k$, cioè l'algoritmo è monotono;
2. se $\nabla f(x_k) \neq 0 \quad \forall k$, si ha

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

3. esiste una funzione di forzamento $\sigma : R^+ \to R^+$ tale che, per $d_k \neq 0$ si ha

$$\frac{\left| d_k^T \nabla f(x_k) \right|}{\| d_k \|} \geq \sigma\left(\| \nabla f(x_k) \| \right)$$

Allora

- viene trovato un punto stazionario dopo un numero finito di passi
- o viene generata una successione infinita di punti tale che:

$$\lim_{k \to \infty} f(x_k) = -\infty$$ [18]

- o valgono le stesse proprietà del teorema precedente (condizione sufficiente di convergenza globale).

[18] Decresce a $-\infty$, può darsi che la funzione però non abbia minimo, quindi l'algoritmo diverge. Possiamo nel codice inserire controlli, se vediamo che l'algoritmo diverge lo fermiamo e partiamo da un altro x_0.