Introduzione all'ottimizzazione non vincolata

A. Agnetis*

1 Introduzione

In questi appunti saranno presentati i concetti introduttivi relativi all'ottimizzazione di una funzione f di n variabili su un insieme X. La f prende il nome di *funzione obiettivo*, l'insieme X *insieme* o *regione ammissibile*. Ciascun punto $x = (x_1, x_2, \ldots, x_n)^T \in X$ costituisce una *soluzione ammissibile*. Il problema di *Programmazione Non Lineare* (PNL) consiste nel determinare un punto x^* appartenente all'insieme X tale da rendere minima la funzione f. Il problema può indicarsi in generale così:

$$\begin{aligned} &\min \ f(x) \\ &\quad x \in X \end{aligned} \tag{1}$$

Nel caso in cui X coincida con tutto R^n si parla di *ottimizzazione non vincolata*, altrimenti di *ottimizzazione vincolata*. Valgono le seguenti definizioni:

Definizione 1 *Un punto $x^* \in X$ è un* punto di minimo globale *di f su X se*

$$f(x^*) \leq f(x) \quad \textit{per ogni} \quad x \in X$$

□

Definizione 2 *Un punto $x^* \in X$ è un* punto di minimo locale *di f su X se esiste un intorno circolare $I(x^*, \epsilon)$ di x^*, avente raggio $\epsilon > 0$ tale che*

$$f(x^*) \leq f(x) \quad \textit{per ogni} \quad x \in X \cap I(x^*, \epsilon)$$

□

*Dipartimento di Ingegneria dell'Informazione - Università di Siena

Un punto x^* di minimo (globale o locale) si dice *stretto* se $f(x)$ è strettamente maggiore di $f(x^*)$ per ogni $x \neq x^*$ in X o in $I(x^*, \epsilon)$ rispettivamente. Ovviamente un punto di minimo globale è anche locale, mentre il viceversa può non essere vero.

Il problema di PNL (1) consiste nel determinare, se esiste, un punto di minimo globale, o, quando questo risultasse eccessivamente oneroso dal punto di vista risolutivo, almeno un punto di minimo locale. Come si vedrà, a seconda della forma della funzione obiettivo e della struttura dell'insieme ammissibile cambia sensibilmente la difficoltà del problema, e di conseguenza cambiano gli approcci risolutivi che vanno utilizzati.

Secondo una convenzione abbastanza comune, i vettori sono sempre pensati come vettori-*colonna*. Dunque, un vettore-riga sarà rappresentato come vettore-colonna trasposto.

Nel seguito si supporranno note le nozioni elementari di analisi (limite, derivata, integrale...) e di algebra (vettori, matrici, matrice inversa, norma, autovalori...). Altre nozioni saranno invece richiamate.

In questa dispensa tratteremo i problemi di ottimizzazione non vincolata, mentre i concetti di base di ottimizzazione vincolata saranno presentati in una dispensa successiva. Vedremo separatamente le *condizioni* di ottimalità dagli *algoritmi* risolutivi. Le prime costituiscono i risultati teorici in base ai quali è possibile caratterizzare le soluzioni ottime di un problema. I secondi sono invece lo strumento con cui ricercare una soluzione ottima.

2 Richiami di analisi e algebra

2.1 Derivate direzionali, gradiente, Hessiana

Come prima cosa, vogliamo richiamare alcuni concetti che generalizzano a R^n alcune nozioni relative alle funzioni di singola variabile. La prima di queste nozioni è quella di derivata. Mentre in R^1 la variabile indipendente può variare solo lungo la retta, in R^n si può considerare la variazione di x in una qualsiasi direzione.

DEFINIZIONE 3 *Si consideri una funzione $f : R^n \to R$, e un vettore $d \in R^n$. Sia $x \in R^n$ un punto in cui f è definita. Se esiste il limite*

$$\lim_{\lambda \to 0^+} \frac{f(x + \lambda d) - f(x)}{\lambda}$$

allora tale limite prende il nome di derivata direzionale di f nel punto x lungo la direzione d. □

Si noti che nel caso in cui $d = (0, 0, \ldots, 1, \ldots, 0)^T$, ossia se d è il vettore costituito da tutti zeri tranne un 1 in i-esima posizione, la derivata direzionale coincide con la nota

derivata parziale di f rispetto alla variabile x_i, indicata con

$$\frac{\partial f}{\partial x_i}.$$

Una direzione che riveste particolare importanza nell'ottimizzazione, è quella avente come componenti le n derivate parziali.

DEFINIZIONE 4 *Si consideri una funzione $f : R^n \to R$, e un punto $x \in R^n$. Se in x esistono le n derivate parziali $\frac{\partial f}{\partial x_i}$, $i = 1, \ldots, n$, definiamo* gradiente *di f in x il vettore $\nabla f(x) \in R^n$ avente come componenti le derivate parziali, ossia*

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1} \; \frac{\partial f}{\partial x_2} \; \ldots \; \frac{\partial f}{\partial x_n} \right]^T.$$

□

Ricordiamo che analogamente si può introdurre il concetto di derivata parziale seconda rispetto a due variabili, nell'ordine x_i e x_j, indicata con

$$\frac{\partial^2 f}{\partial x_i \partial x_j}.$$

In queste dispense, supporremo – tranne che quando esplicitamente indicato – che la funzione f sia almeno *due volte continuamente differenziabile*, dunque disponga di tutte le derivate parziali seconde continue, in tutto X. In questo caso possiamo dare la seguente definizione.

DEFINIZIONE 5 *Sia $f : R^n \to R$ due volte continuamente differenziabile in $x \in R^n$. Definiamo* matrice Hessiana *di f in x la matrice*

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}, \frac{\partial^2 f}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}, \frac{\partial^2 f}{\partial x_2^2}, \ldots, \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \ldots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}, \frac{\partial^2 f}{\partial x_n \partial x_2}, \ldots, \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

□

Si noti che nelle ipotesi di derivate parziali seconde continue, per il noto teorema di Schwarz, la matrice Hessiana risulta simmetrica.

A questo punto siamo in grado di enunciare (senza dimostrarlo) un risultato che rappresenta l'estensione a più variabili della formula di Taylor per funzioni di una sola variabile. Ricordiamo che, se una funzione di una sola variabile ha derivata continua in un

intorno di un punto x, e si considera il punto $x + h$, appartenente a tale intorno, allora è possibile esprimere l'incremento della funzione nel seguente modo (formula di Taylor arrestata ai termini del primo ordine):

$$f(x + h) = f(x) + f'(x)h + \beta_1(x, h)$$

dove $\beta_1(x, h)$ è un infinitesimo di ordine superiore rispetto ad h. Se poi possiede anche la derivata seconda continua, allora è possibile scrivere (formula di Taylor arrestata ai termini del secondo ordine):

$$f(x + h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \beta_2(x, h)$$

dove stavolta $\beta_2(x, h)$ è un infinitesimo di ordine superiore rispetto ad h^2. In altre parole, con la formula di Taylor è possibile *approssimare* il valore di una funzione in un punto "incrementato" $x+h$ utilizzando i valori delle derivate nel punto x, e tale approssimazione è tanto migliore quanto meno ci spostiamo dal punto iniziale, ossia quanto più piccolo è h.

In più dimensioni, il significato e la struttura della formula di Taylor sono molto simili. Stavolta però x, h e $\nabla f(x)$ sono *vettori* a n componenti, e inoltre l'Hessiana è una matrice $n \times n$, per cui le due formule diventano rispettivamente:

$$f(x + h) = f(x) + \nabla f(x)^T h + \beta_1(x, h) \tag{2}$$

$$f(x + h) = f(x) + \nabla f(x)^T h + \frac{1}{2}h^T \nabla^2 f(x)h + \beta_2(x, h) \tag{3}$$

ove $\beta_1(x, h)$ e $\beta_2(x, h)$ sono rispettivamente infinitesimi di ordine superiore rispetto alla *norma* dell'incremento h e al quadrato della norma di h.

Utilizzando la formula di Taylor (2), possiamo scoprire un legame tra alcuni dei concetti introdotti. Poniamo $h = \lambda d$, ove λ è uno scalare. La formula diventa:

$$f(x + \lambda d) = f(x) + \lambda \nabla f(x)^T d + \beta_1(x, \lambda, d)$$

dividendo tutto per λ, si ha

$$\frac{f(x + \lambda d) - f(x)}{\lambda} = \nabla f(x)^T d + \frac{\beta_1(x, \lambda, d)}{\lambda}$$

da cui, passando al limite per $\lambda \to 0$, si ha il seguente risultato:

TEOREMA 1 *La derivata direzionale di f nel punto x lungo la direzione d è data da $\nabla f(x)^T d$.* □

2.2 Convessità

Un concetto che riveste molta importanza in ottimizzazione è quello di *convessità* (che nel seguito tratteremo sempre con riferimento a insiemi in R^n).

DEFINIZIONE 6 *Dati due punti $x, y \in R^n$, e considerato uno scalare $\lambda \in [0,1]$, si dice* combinazione convessa *di x e y un qualunque punto ottenuto come*

$$\lambda x + (1-\lambda)y$$

inoltre, al variare di λ tra 0 e 1, si ottiene il segmento *(in R^n) che unisce x e y.* □

DEFINIZIONE 7 *Un insieme $X \subset R^n$ si dice* convesso *se, presi comunque due punti $x, y \in X$, il segmento che li unisce è interamente contenuto in X.* □

Consideriamo ora un insieme convesso X, e una funzione f definita su tale insieme.

DEFINIZIONE 8 *Una funzione f definita su un insieme convesso X si dice* convessa *se, presi comunque due punti $x, y \in X$, si ha che per ogni scalare $\lambda \in [0,1]$*

$$\lambda f(x) + (1-\lambda)f(y) \geq f(\lambda x + (1-\lambda)y) \tag{4}$$

(una funzione f tale che $-f$ è convessa, si dice concava*).* □

Il significato della (4) può essere facilmente compreso facendo riferimento a funzioni di una sola variabile. In tal caso, considerando il punto $\tilde{x} = \lambda x + (1-\lambda)y$ (che ovviamente appartiene all'intervallo $[x, y]$), il termine di sinistra rappresenta il valore dell'ordinata del punto che si trova sul segmento che unisce i due punti $(x, f(x))$ e $(y, f(y))$ in corrispondenza al valore $\tilde{x}$, mentre il termine di destra è il valore della funzione in corrispondenza dello stesso valore $\tilde{x}$. Dunque, se f è convessa, vuol dire che il grafico della funzione si trova sempre al di sotto di un segmento teso fra due suoi punti.

La Definizione 8 vale in generale, senza alcuna ipotesi sulle proprietà della funzione f. Se però aggiungiamo, come stiamo supponendo, che la f sia continuamente differenziabile, allora è possibile dire qualcosa in più.

TEOREMA 2 *Sia f una funzione continuamente differenziabile in R^n e convessa. Dati due punti $x, y \in R^n$,*

$$f(y) - f(x) \geq \nabla f(x)^T (y - x) \tag{5}$$

Dim.– Essendo la f convessa, vale la (4), che, ponendo $\epsilon = 1 - \lambda$, possiamo riscrivere come

$$f(x) + \epsilon(f(y) - f(x)) \geq f(x + \epsilon(y - x)) \tag{6}$$

in questa disuguaglianza, possiamo interpretare x come un punto e $x + \epsilon(y - x)$ come un "punto incrementato", ottenuto muovendosi nella direzione $y - x$ di una quantità pari a ϵ. Possiamo allora scrivere la formula di Taylor troncata al primo ordine (2), ossia

$$f(x + \epsilon(y - x)) = f(x) + \epsilon \nabla f(x)^T (y - x) + \beta(x, y, \epsilon) \tag{7}$$

dalle (6) e (7), si ha dunque che, se la f è convessa,

$$\epsilon(f(y) - f(x)) \geq \epsilon \nabla f(x)^T (y - x) + \beta(x, y, \epsilon)$$

da cui, dividendo per ϵ e passando al limite per $\epsilon \to 0$, si ha la tesi. □

È importante sapere che è vero anche il viceversa, anche se non lo dimostreremo, ossia:

Teorema 3 *Sia f una funzione continuamente differenziabile in X. Se, presi comunque due punti $x, y \in X$, risulta*

$$f(y) - f(x) \geq \nabla f(x)^T (y - x)$$

allora la f è convessa in X. □

Si noti che, nel caso monodimensionale, anche la (5) ha un'immediata interpretazione geometrica. Infatti, $f(x) + f'(x)(y - x)$ è l'ordinata del punto y sulla retta tangente alla curva in x. Se f è convessa, quindi, la curva della funzione si trova sempre al di sopra di una retta tangente in un suo punto.

Un altro concetto (come vedremo legato alla convessità) che è utile richiamare è quello di matrice *definita positiva*.

Definizione 9 *Data una matrice A quadrata di ordine n, e un insieme $Y \subseteq R^n$, essa si dice* definita positiva *se, per ogni $x \in R^n$, $x \neq 0$, si ha*

$$x^T A x > 0$$

Se invece, per qualsiasi $x \in R^n$, $x \neq 0$,si ha

$$x^T A x \geq 0$$

la matrice si dice semidefinita positiva su Y. *Una matrice A è* definita *o* semidefinita negativa *se $-A$ è definita o semidefinita positiva rispettivamente. In tutti gli altri casi, la matrice è* indefinita. [1] □

[1] In alcuni casi (soprattutto in problemi di ottimizzazione vincolata) risulta necessario dare una definizione più generale. Se $x^T A x > 0$ solo per $x \in Y$, con $Y \subset R^n$, allora si dice che A è *definita positiva in Y* (idem per matrici semidefinite).

Quella fornita dalla Definizione 9 non è l'unica definizione possibile per le matrici definite positive. Una caratterizzazione operativamente più utile, ma che vale solo per matrici simmetriche, è basata sul segno dei *minori principali*, che, ricordiamo, sono le n sottomatrici quadrate formate dall'intersezione delle prime i righe e delle prime i colonne, $i = 1, \ldots, n$. Si hanno allora i seguenti teoremi (di Sylvester):

TEOREMA 4 *Una matrice A simmetrica è definita positiva se e solo se i determinanti di tutti i minori principali di A sono positivi.* □

(Se A non è simmetrica, la condizione suddetta è necessaria ma non sufficiente).

TEOREMA 5 *Una matrice A simmetrica è semidefinita positiva se e solo se i determinanti di tutti i minori principali di A sono non negativi.* □

Una proprietà interessante delle matrici definite positive è espressa dal seguente teorema

TEOREMA 6 *Tutti gli autovalori di una matrice A simmetrica definita positiva sono positivi.* □

Si noti, tra l'altro, che ciò implica che una matrice definita positiva è non singolare, dal momento che altrimenti avrebbe tra i suoi autovalori anche lo 0. Inoltre vale quest'altra proprietà, che ci tornerà utile in seguito (§9).

TEOREMA 7 *Se A è una matrice definita positiva, e $\lambda_m(A)$ e $\lambda_M(A)$ indicano rispettivamente il minimo e il massimo autovalore di A, allora per ogni $x \in R^n$ si ha*

$$\lambda_m(A)\|x\|^2 \leq x^T A x \leq \lambda_M(A)\|x\|^2$$

dove la norma è quella euclidea. □

Consideriamo ora una funzione f con derivate seconde continue in tutto R^n. Dati due punti x, y, per la (3) possiamo sempre scrivere

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + \beta_2(x, y) \tag{8}$$

Supponiamo ora che l'Hessiana calcolata nel punto x sia definita positiva. Di conseguenza, di certo

$$(y - x)^T \nabla^2 f(x)(y - x) > 0$$

Essendo $\beta_2(x,y)$ infinitesimo di ordine superiore rispetto al termine di secondo grado, si ha che il segno di quest'ultimo prevale rispetto a quello di $\beta_2(x,y)$, *almeno in un intorno di* x. Dalla (8) segue allora

$$f(y) - f(x) \geq \nabla f(x)^T (y - x)$$

e dunque, invocando il Teorema 3, risulta dimostrato il seguente teorema:

TEOREMA 8 *Data una funzione f due volte continuamente differenziabile, sia x un punto in cui $\nabla^2 f(x)$ è definita positiva. Allora, almeno in un intorno di x, la f è convessa.* □

Un caso particolare di notevole interesse è quello relativo a funzioni la cui Hessiana è costante. Queste sono le funzioni del tipo $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, con $A \in R^{n\times n}, b \in R^n$ e $c \in R$, e sono dette funzioni *quadratiche*. È facile verificare che in questo caso (supponendo per semplicità A simmetrica), il gradiente nel punto $\bar{x}$ è dato da $A\bar{x} + b$, mentre l'Hessiana coincide con la matrice A, indipendentemente dal punto in cui essa è calcolata. In questo caso, dunque, se A è definita positiva, la funzione f è convessa (in tutto R^n).

3 Esistenza di un minimo

Esiste una serie di teoremi che danno condizioni necessarie e sufficienti per l'esistenza di punti di minimo. Tra questi riportiamo qui, senza dimostrazione, quelli più significativi per gli scopi di questa dispensa.

TEOREMA 9 *Se una funzione f è continua su un insieme X chiuso e limitato, allora f ammette un punto di minimo globale in X.* □

Concetti importanti sono quelli di *insieme* e *superficie di livello*.

DEFINIZIONE 10 *Data una funzione $f(x)$ definita in un insieme X, e un numero reale α, un* insieme di livello *di f su X è l'insieme di tutti i punti x in cui il valore della funzione non eccede α, ossia*

$$\mathcal{L}_f(X, \alpha) = \{x \in X : f(x) \leq \alpha\}$$

mentre una superficie di livello *è l'insieme dei punti x in cui f vale esattamente α.*

$$\mathcal{C}_f(X, \alpha) = \{x \in X : f(x) = \alpha\}$$

□

Nel caso di insiemi non limitati, ma chiusi, è possibile dare una condizione sufficiente di esistenza:

TEOREMA 10 *Se una funzione f è continua su un insieme X chiuso, e almeno uno dei suoi insiemi di livello è chiuso e limitato, allora f ammette un punto di minimo in X.* □

Si noti che quest'ultima condizione è sufficiente, ma non necessaria. Una situazione che spesso si presenta è quella che si ha quando la funzione f cresce indefinitamente allontanandosi dall'origine.

DEFINIZIONE 11 *Una funzione f continua in tutto R^n si dice* radialmente illimitata *o* coerciva *se*

$$\lim_{\|x\|\to\infty} f(x) = +\infty \tag{9}$$

□

(Nella (9) è da intendersi che la norma di x tende a infinito in qualsiasi direzione.) In questo caso è possibile dimostrare il seguente risultato:

TEOREMA 11 *Se una funzione f continua è radialmente illimitata, tutti i suoi insiemi di livello sono limitati e chiusi.* □

Combinando quest'ultimo teorema con il Teorema 10, risulta quindi che

TEOREMA 12 *Se una funzione f è radialmente illimitata, ammette un minimo globale.* □

4 Angolo tra due vettori e direzioni di discesa

Com'è noto, dati due punti $x, y \in R^n$, il loro *prodotto scalare* è dato da

$$x^T y = \sum_{i=1}^{n} x_i y_i$$

DEFINIZIONE 12 *Dati due vettori $x, y \in R^n$, l'*angolo *tra essi è quel numero θ (compreso fra 0 e π) tale che*

$$\cos\theta = \frac{x^T y}{\|x\|\ \|y\|}$$

□

Come è noto, se il prodotto scalare di due vettori è 0, i due vettori si dicono *ortogonali*, e in questo caso risulta quindi $\cos\theta = 0$. Se invece x e y hanno la stessa direzione, allora $y = \alpha x$ con $\alpha \in R$, e risulta (se $\alpha > 0$)

$$\cos\theta = \frac{\alpha x^T x}{|\alpha|\ ||x||^2} = 1.$$

Consideriamo un punto x, e una funzione $f(x)$. Da questo punto x possiamo pensare di muoverci seguendo una direzione $d \in R^n$. Muovendoci di una quantità $\lambda \in R$ lungo la direzione d, raggiungiamo il punto $x + \lambda d$. Siccome il nostro scopo ultimo è quello di trovare un punto di minimo della funzione f, risulta chiaramente di interesse stabilire se, muovendosi lungo la direzione d, la funzione cresce o decresce.

DEFINIZIONE 13 *Data una funzione f, una direzione d si dice* direzione di discesa *per f in x se esiste un $\bar{\lambda} > 0$ tale che*

$$f(x + \lambda d) < f(x)$$

per ogni $0 < \lambda < \bar{\lambda}$. □

In sostanza una direzione d è di discesa se, spostandosi da x lungo quella direzione, la funzione descresce – almeno, purché gli spostamenti siano sufficientemente piccoli. A questo punto, possiamo comprendere meglio il significato geometrico della derivata direzionale, come illustrato dal seguente teorema.

TEOREMA 13 *Si consideri una funzione f con gradiente continuo in tutto R^n. Dati $x, d \in R^n$, se la derivata direzionale di f in x rispetto alla direzione d è negativa, la direzione d è una direzione di discesa per f in x.*

Dim. – Dal Teorema 1 si ha

$$\lim_{\lambda \to 0} \frac{f(x + \lambda d) - f(x)}{\lambda} = \nabla f(x)^T d < 0$$

e quindi, per λ sufficientemente piccolo, si ha $f(x + \lambda d) - f(x) < 0$. □

Dunque, per trovare punti in cui la funzione ha un valore inferiore a quello che ha in x, si può seguire una direzione d la cui derivata direzionale è negativa, facendo attenzione a non compiere passi troppo "lunghi". Se scegliessimo invece una direzione per la quale $\nabla f(x)^T d > 0$, ribaltando la discussione avremmo che la f crescerebbe, e d sarebbe dunque una direzione di salita. Se invece $\nabla f(x)^T d = 0$, vuol dire che d è ortogonale al gradiente e non si può dire in generale se è una direzione di discesa o meno. Peraltro, osserviamo che se ci muoviamo lungo la superficie di livello $\mathcal{C}_f(X, f(x_0)) = \{x \in X : f(x) = f(x_0)\}$, in

ogni momento seguiamo una direzione d tangente alla superficie stessa. Ma non essendovi dunque in quella direzione variazione della f (perché rimaniamo sulla stessa superficie di livello), la derivata direzionale è nulla. Questo indica che la direzione del gradiente in un punto x è sempre ortogonale alla superficie di livello passante per quel punto.

Si noti che il segno della derivata direzionale dà informazioni sull'angolo tra la direzione d e il gradiente. Se tale segno è negativo, ciò vuol dire che l'angolo fra d e $\nabla f(x)$ è ottuso. In particolare, se d è la direzione *opposta* a quella del gradiente, allora $d = -\nabla f(x)$ e l'angolo è piatto, in quanto

$$\cos\theta = \frac{-\nabla f(x)^T \nabla f(x)}{\|\nabla f(x)\|^2} = -1.$$

e dunque l'antigradiente è sempre una direzione di discesa, mentre per lo stesso motivo il gradiente è sempre una direzione di salita. Questo fatto verrà ripreso più avanti.

5 Condizioni di ottimalità

Siamo ora in grado di introdurre alcune condizioni in base alle quali si possono caratterizzare i punti di minimo di una funzione.

TEOREMA 14 *Si consideri una funzione f con gradiente continuo in un punto $x^* \in R^n$. Condizione necessaria affinché x^* sia un punto di minimo locale per f è che*

$$\nabla f(x^*) = 0$$

Dim. – Se fosse $\nabla f(x^*) \neq 0$, allora $-\nabla f(x^*)$ sarebbe una direzione di discesa, e per il Teorema 13 esisterebbe un punto $x^* - \lambda \nabla f(x^*)$ tale che $f(x^* - \lambda \nabla f(x^*)) < f(x^*)$, contraddicendo cosi' l'ipotesi che x^* sia un minimo locale. □

Il Teorema 14 fornisce delle condizioni molto generali, dette *condizioni del* 1° *ordine*. Un punto che soddisfa tali condizioni si dice *punto stazionario*, e indichiamo con $\mathcal{PS}$ l'insieme dei punti stazionari per f, ossia

$$\mathcal{PS} = \{x : x \in R^n, \nabla f(x) = 0\}$$

Se la f è due volte continuamente differenziabile, si possono enunciare anche delle *condizioni del* 2° *ordine*.

TEOREMA 15 *Si consideri una funzione f con Hessiana continua in un punto $x^* \in R^n$. Condizioni necessarie affinché x^* sia un punto di minimo locale per f sono che:*

(a) $\nabla f(x^*) = 0$

(b) $\nabla^2 f(x^*)$ *è semidefinita positiva.*

Dim. – La (a) segue dal Teorema 14. Data una direzione $d \in R^n$, poiché f è due volte differenziabile, possiamo scrivere la formula di Taylor (3) con riferimento a un punto incrementato $x^* + \lambda d$, ove d è una qualsiasi direzione:

$$f(x^* + \lambda d) = f(x^*) + \lambda \nabla f(x^*)^T d + \frac{1}{2}\lambda^2 d^T \nabla^2 f(x^*) d + \beta_2(x^*, \lambda, d)$$

e, poiché in x^* il gradiente si annulla,

$$\frac{f(x^* + \lambda d) - f(x^*)}{\lambda^2} = \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{\beta_2(x^*, \lambda, d)}{\lambda^2}$$

dal momento che x^* è per ipotesi un minimo locale, per λ sufficientemente piccolo il termine di sinistra è sicuramente non negativo, e quindi risulta

$$\frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{\beta_2(x^*, \lambda, d)}{\lambda^2} \geq 0$$

da cui, passando al limite per $\lambda \to 0$, e osservando che d è una direzione qualsiasi, segue la tesi. □

Fin qui si sono viste condizioni necessarie di ottimalità. È possibile dare anche una condizione sufficiente:

Teorema 16 *Si consideri una funzione f con Hessiana continua in un intorno di un punto $x^* \in R^n$. Condizioni sufficienti affinché x^* sia un punto di minimo locale stretto per f sono che:*

(a) $\nabla f(x^*) = 0$

(b) $\nabla^2 f(x^*)$ *è definita positiva.*

Dim. – Basta riscrivere ancora la formula di Taylor, ove $x^* + \lambda d$ è un punto sufficientemente vicino a x^* tale che $\nabla^2 f(x)$ è continua. Sfruttando la (a), possiamo scrivere:

$$f(x^* + \lambda d) = f(x^*) + \frac{1}{2}\lambda^2 d^T \nabla^2 f(x^*) d + \beta_2(x^*, \lambda, d)$$

siccome $\nabla^2 f(x^*)$ è definita positiva, e poiché $\beta_2(x^*, \lambda, d)$ è un infinitesimo di ordine superiore, abbiamo che per qualunque d, e per λ sufficientemente piccolo,

$$\frac{1}{2}\lambda^2 d^T \nabla^2 f(x^*) d + \beta_2(x^*, \lambda, d) > 0$$

da cui la tesi. □

È facile rendersi conto che la discussione precedente può essere ripetuta in modo simmetrico relativamente ai punti di massimo. In particolare, l'annullamento del gradiente

è anche condizione necessaria (del 1° ordine) affinché un punto sia punto di massimo locale, mentre la corrispondente condizione necessaria del 2° ordine è che l'Hessiana sia semidefinita negativa; se oltre a soddisfare le condizioni del 1° ordine l'Hessiana è definita negativa, allora il punto in questione è di massimo locale stretto. Si noti che se in un punto x^* si annulla il gradiente, ma l'Hessiana è indefinita, possiamo *escludere* che x^* sia punto di minimo o di massimo. In tal caso, x^* è detto *punto di sella*.

5.1 Il caso convesso

Come ora vedremo, nel caso in cui la f è una funzione convessa, è possibile dimostrare alcune proprietà molto forti della soluzione ottima del problema. Anzitutto, vediamo che, sotto ipotesi molto generali, nel caso convesso la distinzione tra minimi locali e globali non sussiste.

TEOREMA 17 *Si consideri una funzione f convessa in R^n. Se x^* è un punto di minimo locale, è anche un punto di minimo globale.*

Dim.– Essendo x^* un punto di minimo locale, senz'altro $f(x) \geq f(x^*)$ per tutti i punti $x \in I(x^*, \epsilon)$. Supponiamo che x^* non sia un minimo globale. Deve esistere allora un punto z tale che $f(z) < f(x^*)$. Sia $\tilde{x}$ il generico punto del segmento che unisce z e x^*, ossia $\tilde{x} = \lambda z + (1-\lambda)x^*$. Per λ sufficientemente piccolo, $\tilde{x} \in I(x^*, \epsilon)$. D'altro canto, per la convessità avremo che

$$f(\tilde{x}) = f(\lambda z + (1-\lambda)x^*) \leq \lambda f(z) + (1-\lambda) f(x^*)$$

ma siccome stiamo supponendo $f(z) < f(x^*)$, da questa discende

$$f(\tilde{x}) < \lambda f(x^*) + (1-\lambda) f(x^*) = f(x^*)$$

il che contraddice il fatto che x^* è un minimo locale. □

Si noti che il Teorema 17 vale in ipotesi del tutto generali: non abbiamo nemmeno supposto la f differenziabile. Se lo è, vediamo ora che la convessità consente di dare una caratterizzazione dei punti di minimo più forte di quanto visto finora. Infatti, in generale, il soddisfacimento delle condizioni necessarie del primo e del second'ordine non basta a determinare la natura del punto in questione. Invece, nel caso particolare che la f sia convessa, le sole condizioni del 1° ordine divengono necessarie e sufficienti.

TEOREMA 18 *Si consideri una funzione f con gradiente continuo, e sia f convessa in R^n. Condizione necessaria e sufficiente affinché x^* sia un punto di minimo globale per f è che*

$$\nabla f(x^*) = 0$$

Dim. – La necessità deriva dal Teorema 14. Per quanto concerne la sufficienza, basta ricordare la (5), ove y è un qualunque punto di R^n:

$$f(y) - f(x^*) \geq \nabla f(x^*)^T (y - x^*)$$

per cui, se $\nabla f(x^*) = 0$, si ha che $f(y) \geq f(x^*)$. □

Dunque, nel caso convesso trovare un minimo locale equivale a trovare un minimo globale, e un punto è di minimo se e solo se soddisfa le condizioni del 1° ordine. C'è ancora una proprietà che contraddistingue il caso convesso: i punti di minimo formano essi stessi un insieme convesso.

TEOREMA 19 *Si consideri una funzione f, convessa in R^n. L'insieme dei punti di minimo della f è convesso.*

Dim.– Si considerino due punti di minimo distinti, x^* e x^{**}, e sia $\bar{f}$ il valore ottimo della f. Il generico punto sul segmento che li unisce è $\lambda x^* + (1-\lambda)x^{**}$. Per la convessità della f,

$$f(\lambda x^* + (1-\lambda)x^{**}) \leq \lambda f(x^*) + (1-\lambda) f(x^{**})$$

ma essendo sia x^* che x^{**} punti di minimo, $f(x^*) = f(x^{**}) = \bar{f}$ e quindi

$$f(\lambda x^* + (1-\lambda)x^{**}) = \bar{f}$$

e quindi anche tutti i punti del segmento sono punti di minimo. □

6 Schema generale degli algoritmi di minimizzazione non vincolata

Le condizioni di ottimalità che abbiamo visto possono talora essere utilizzate direttamente per calcolare la soluzione ottima di un problema di ottimizzazione non vincolata, ma spesso la forma della funzione obiettivo e/o il numero di variabili possono essere tali da rendere di fatto impossibili i calcoli in forma chiusa. In generale è allora necessario progettare e utilizzare un *algoritmo* iterativo. Va detto subito che, in generale, gli algoritmi di minimizzazione non vincolata consentono soltanto la determinazione di *punti stazionari* di f, senza in generale la garanzia che si tratti effettivamente di punti di minimo. Molto spesso tuttavia si è anche in grado di stabilire che i punti di $\mathcal{PS}$ raggiunti soddisfano le condizioni del 2° ordine.

La struttura di fondo degli algoritmi di minimizzazione non vincolata è molto semplice. Si considera un punto iniziale $x_0 \in R^n$, e si calcola il valore della funzione $f(x_0)$

```
Metodo_di_discesa
{
 Si fissa un punto iniziale x_0 ∈ R^n ; k := 0;
 while  ∇f(x_k) ≠ 0
        {
        si calcola una direzione di discesa d_k ∈ R^n
        si calcola un passo α_k ∈ R lungo d_k
        x_{k+1} := x_k + α_k d_k;
        k := k + 1;
        }
}
```

Figura 1: Schema generale degli algoritmi di minimizzazione non vincolata.

e del gradiente $\nabla f(x_0)$. Se quest'ultimo è il vettore nullo, si è già individuato un punto stazionario. Altrimenti, da x_0 ci si sposta in cerca di un punto x_1, possibilmente migliore di x_0. Per fare questo, appare logico scegliere una direzione di discesa, e lungo tale direzione muoversi di un opportuno *passo*. Trovato x_1, se esso appartiene a $\mathcal{PS}$ ci si ferma, altrimenti si cerca un nuovo punto x_2 e cosi' via. Al generico passo quindi, il nuovo punto x_{k+1} si ottiene come

$$x_{k+1} := x_k + \alpha_k d_k \tag{10}$$

Si ottiene cosi' una *successione* $x_0, x_1, x_2, \ldots, x_k, \ldots$ di punti. A tale schema generale di algoritmo ci si riferisce in genere con il termine di *metodo di discesa*, a indicare che a ogni passo risulta

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

Lo schema è dunque quello riportato in Figura 1.

Ciò che distingue un metodo di discesa da un altro sono i criteri da usare per effettuare le due scelte che caratterizzano il metodo: la direzione di discesa e la lunghezza del passo. Peraltro, da queste scelte dipendono le caratteristiche fondamentali dell'algoritmo, vale a dire la *convergenza* della successione $\{x_k\}$ a punti stazionari e la *rapidità* di convergenza. Nel seguito, supporremo sempre verificate le ipotesi matematiche in base alle quali la funzione f ammette almeno un punto di minimo (e dunque $\mathcal{PS}$ non è vuoto).

6.1 Convergenza

Prima di vedere un teorema che dà condizioni sotto le quali un algoritmo del tipo (10) converge a un punto stazionario, facciamo una ulteriore distinzione. Un importante criterio di classificazione degli algoritmi riguarda il punto iniziale della successione $\{x_k\}$. Infatti, in alcuni casi l'algoritmo può convergere o meno a seconda della scelta di x_0.

DEFINIZIONE 14 *Un algoritmo è* globalmente convergente *se esso è convergente per qualunque* $x_0 \in R^n$. *Un algoritmo è* localmente convergente *se è convergente solo per* $x_0 \in I(x^*)$, *ove* $I(x^*)$ *è un opportuno intorno di un punto* $x^* \in \mathcal{PS}$. □

Ovviamente la convergenza globale è preferibile a quella locale, anche perché tipicamente è molto difficile conoscere la struttura e l'estensione dell'intorno $I(x^*)$, ma se ne conosce soltanto l'esistenza. Tuttavia, la convergenza locale o globale non è l'unico criterio per giudicare la bontà di un algoritmo, come vedremo nel §6.2.

Per le nostre considerazioni è utile far riferimento alla funzione di una sola variabile $\phi(\alpha) = f(x_k + \alpha d_k)$, che indica il valore della f in funzione del passo α, allorché si sia fissata la direzione di discesa d_k. Sia $y = x_k + \alpha d_k$ il punto incrementato, e indichiamo con y^i la sua i-esima componente. Avendo supposto ∇f continuo, possiamo calcolare la derivata di ϕ come

$$\phi'(\alpha) = \frac{d\phi}{d\alpha} = \sum_{i=1}^{n} \frac{\partial f(y)}{\partial y^i} \frac{dy^i}{d\alpha}$$

considerando che la variazione della i-esima componente di y al variare di α è data dalla i-esima componente del vettore d_k, e ricordando la definizione di gradiente si ha quindi

$$\phi'(\alpha) = \nabla f(y)^T d_k = \nabla f(x_k + \alpha d_k)^T d_k \tag{11}$$

Si noti in particolare che l'inclinazione della retta tangente alla funzione ϕ per $\alpha = 0$ è proprio la derivata direzionale della f in x_k lungo la direzione d_k.

Anzitutto osserviamo che il fatto che, a ogni iterazione del metodo di discesa (algoritmo in Fig.1), si abbia una diminuzione della funzione obiettivo, non basta a garantire la convergenza dell'algoritmo a un punto stazionario. Ad esempio, si consideri la funzione di una sola variabile $f(x) = x^2 - 1$. Se alla k-esima ($k \geq 1$) iterazione dell'algoritmo in Figura 1 si produce il punto $x_k = (-1)^k(1 + 1/k)$, a ogni iterazione si ha $f(x_{k+1}) < f(x_k)$. La successione $\{x_k\}$ ha i due punti limite $x = 1$ e $x = -1$, e dunque $\{f(x_k)\}$ converge al valore 0. Invece, l'unico punto di minimo della f è $x = 0$, in cui $f(x) = -1$. Dunque, la successione dei valori $\{f(x_k)\}$ converge, ma al valore sbagliato. Il problema, in questo caso, è che è vero che la funzione f diminuisce a ogni iterazione, ma sempre di meno.

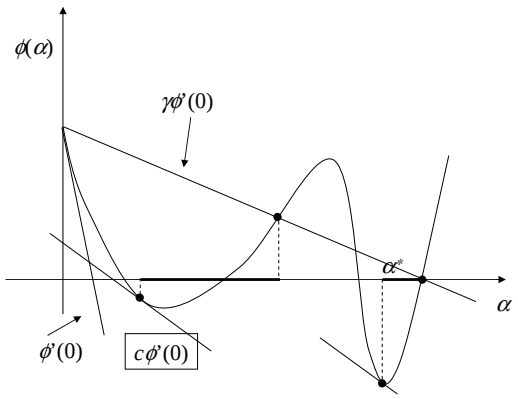

Figura 2: L'insieme di valori di α che soddisfano le (12) e (13).

Dunque, è desiderabile che, a ogni iterazione dell'algoritmo, il punto incrementato $x_k + \alpha d_k$ soddisfi una *condizione di sufficiente riduzione* della f, che possiamo esprimere in questi termini:

$$f(x_k + \alpha d_k) \leq f(x_k) + \gamma \alpha \nabla f(x_k)^T d_k \tag{12}$$

ove $0 < \gamma < 1$. Guardando il grafico della $\phi(\alpha)$ (Figura 2), dal momento che $\nabla f(x_k)^T d_k < 0$ è la derivata di ϕ in $\alpha = 0$, si vede che la condizione espressa dalla (12) significa che il nuovo punto deve trovarsi *sotto* la retta passante per il punto $(0, \phi(0))$ e avente pendenza $\gamma \nabla f(x_k)^T d_k$ (dunque negativa). Si noti che essendo $\gamma < 1$, tale retta è in valore assoluto meno pendente della tangente a ϕ in $\alpha = 0$. Scegliendo diversi valori per il parametro γ si può rendere la condizione più o meno restrittiva. In pratica, si usano in genere valori di γ abbastanza piccoli, dell'ordine di 10^{-4}.

La (12) da sola può non essere ancora sufficiente a garantire una buona efficienza dell'algoritmo. Questo perché, essendo comunque la condizione soddisfatta per valori sufficientemente piccoli di α, il rischio è che lo spostamento rispetto a x_k sia poco significativo. Ricordando che, per la (11), la pendenza della ϕ nel punto α è data da $\nabla f(x_k + \alpha d_k)^T d_k$, possiamo allora considerare la seguente condizione, chiamata *condizione di Wolfe*:

$$\nabla f(x_k + \alpha d_k)^T d_k \geq c \nabla f(x_k)^T d_k \tag{13}$$

ove $\gamma < c < 1$. (Questa condizione può anche scriversi $\phi'(\alpha) \geq c\phi'(0)$.) La (13) è senz'altro soddisfatta se in α la ϕ ha pendenza positiva (si ricordi che $\phi'(0) < 0$), mentre nel caso che la pendenza in α sia negativa, la condizione è soddisfatta solo se la pendenza è, in valore assoluto, inferiore a $c\phi'(0)$ (figura 2) . In sostanza, la (13) vincola il passo α

ad essere abbastanza lungo da percepire una significativa diminuzione (in valore assoluto) della derivata direzionale, fatto questo che indica un avvicinamento al minimo della f.

Si noti che, stavolta, valori troppo piccoli di α possono *non* soddisfare la condizione di Wolfe. Ci si può dunque chiedere se esistono effettivamente dei valori di α in cui sia la condizione di sufficiente riduzione che quella di Wolfe siano soddisfatte. Non è difficile dimostrare il seguente teorema.

TEOREMA 20 *Data una direzione di discesa d_k per una funzione f continuamente differenziabile, se la $\phi(\alpha)$ non è inferiormente illimitata per $\alpha > 0$, allora esistono $0 < \alpha_1 < \alpha_2$ tali che ogni $\alpha \in [\alpha_1, \alpha_2]$ soddisfa sia (12) che (13).*

Le due condizioni viste fanno riferimento al valore del passo α, mentre per quanto riguarda d_k richiedono solo che questa sia una direzione di discesa. Vediamo ora invece una condizione su d_k che richiede qualcosa in più.

DEFINIZIONE 15 *Un algoritmo di ottimizzazione del tipo (10)* soddisfa la condizione d'angolo *se esiste un $\epsilon > 0$ tale che, a ogni passo k*

$$\nabla f(x_k)^T d_k \leq -\epsilon \, \|\nabla f(x_k)\| \, \|d_k\| \tag{14}$$

con $\epsilon > 0$.

Come sappiamo, se $\nabla f(x_k)^T d_k < 0$, la direzione d_k è una direzione di discesa. La condizione d'angolo richiede qualcosa in più: dovendo essere soddisfatta a ogni iterazione k dell'algoritmo, dal momento che il termine di destra nella (14) è strettamente negativo, questa condizione implica che il coseno dell'angolo tra il gradiente e la direzione di ricerca si mantiene sempre strettamente inferiore a $-\epsilon$. Questo impedisce che, nel corso dell'algoritmo, d_k possa tendere a diventare ortogonale a $\nabla f(x_k)$. Si noti che non importa quanto piccolo sia ϵ, purché sia strettamente positivo.

Possiamo a questo punto enunciare, senza dimostrazione, un teorema che caratterizza la convergenza globale di un algoritmo del tipo (10).

TEOREMA 21 *Sia f una funzione con gradiente continuo in tutto R^n. Si consideri un algoritmo del tipo (10), in cui, a ogni iterazione k, la direzione d_k e il passo α_k sono tali da soddisfare le (12), (13) e (14), per opportune costanti $0 < \gamma < c < 1$, $\epsilon > 0$. Allora, per ogni scelta del punto iniziale x_0 tale che l'insieme di livello $\mathcal{L}_f(x_0)$ è limitato e chiuso, la successione di punti $\{x_k\}$ generata dalla (10) è tale che o (i) esiste un k per cui $\nabla f(x_k) = 0$ oppure (ii) la successione $\{\nabla f(x_k)\}$ tende a 0.*[2]

[2]Si noti che questo teorema non esclude l'esistenza di *più* punti limite, tutti caratterizzati dal fatto di avere gradiente nullo, e lo stesso valore di f. Se da un punto di vista matematico è facile costruire esempi in cui questo accade, dal punto di vista pratico è un fatto abbastanza raro.

Nei prossimi capitoli vedremo con maggior dettaglio i due aspetti caratterizzanti di un algoritmo del tipo (10). Precisamente, nel §7 ci occuperemo della determinazione di α_k, mentre nei §8 e §9 della scelta della direzione di discesa.

6.2 Rapidità di convergenza

In alcuni ambiti dell'ottimizzazione (es. programmazione lineare, ottimizzazione combinatoria) è possibile progettare algoritmi in grado di calcolare la soluzione ottima in un numero finito, ancorché eventualmente molto elevato, di iterazioni. In quei casi, è possibile basare una definizione di efficienza di un algoritmo sul numero di iterazioni che è necessario effettuare per giungere alla soluzione ottima. Nel caso della programmazione non lineare, come si è detto, tranne che in casi particolari (come il metodo di Newton per funzioni quadratiche, §9), gli algoritmi risolutivi producono una successione infinita di punti $\{x_k\}$. Supponendo che le condizioni di convergenza siano soddisfatte, e che un certo algoritmo converga a un punto stazionario x^*, occorre allora caratterizzare la rapidità con cui tale convergenza avviene. I metodi più utilizzati per misurare la convergenza fanno riferimento al rapporto tra gli scostamenti esistenti, a un'iterazione e alla successiva, tra la soluzione corrente x_k e il punto limite x^*, ossia

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|}$$

Le misure di rapidità di convergenza si basano sull'andamento di tale scostamento nel corso dell'algoritmo. Si noti che tale rapporto *non* fa riferimento al valore della f.

DEFINIZIONE 16 *Un algoritmo che costruisce una successione $\{x_k\}$ convergente a x^* ha* rapidità di convergenza lineare *se esistono un numero finito $C > 0$ e un valore $\tilde{k}$ tale che, per $k \geq \tilde{k}$, risulta*

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

DEFINIZIONE 17 *Un algoritmo che costruisce una successione $\{x_k\}$ convergente a x^* ha* rapidità di convergenza superlineare *se esistono una successione $\{\alpha_k\}$ tendente a 0 e un valore $\tilde{k}$ tale che, per $k \geq \tilde{k}$, risulta*

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq \alpha_k$$

DEFINIZIONE 18 *Un algoritmo che costruisce una successione $\{x_k\}$ convergente a x^* ha* rapidità di convergenza quadratica *se esistono un numero finito $C > 0$ e un valore $\tilde{k}$ tale che, per $k \geq \tilde{k}$, risulta*

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

Il significato di queste definizioni dovrebbe essere abbastanza chiaro, ricordando che, dal momento che l'algoritmo converge, almeno da una certa iterazione in poi i valori $\|x_k - x^*\|$ diventano sempre più piccoli. Evidentemente, gli algoritmi aventi convergenza lineare sono quelli meno efficienti, anche se, al loro interno, una differenza dal punto di vista pratico è senz'altro rappresentata dal valore di C (che prende il nome di *tasso di convergenza*). Tipicamente, una convergenza quadratica può definirsi "veloce", mentre quella lineare può risultare insoddisfacente se C è prossimo o addirittura superiore a 1 (in tal caso si parla talora di convergenza *sublineare*). In generale, si ha la seguente definizione:

DEFINIZIONE 19 *Un algoritmo che costruisce una successione $\{x_k\}$ convergente a x^* ha* rapidità di convergenza di ordine p *se esiste un numero $C > 0$ tale che, per tutti i k da un certo $\tilde{k}$ in poi, risulta*

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

Va detto tuttavia che algoritmi aventi $p > 2$ sono abbastanza rari. Gli algoritmi utilizzati nelle applicazioni reali in genere hanno convergenza superlineare o quadratica.

ESEMPIO 1 *Per esercizio, si possono applicare i concetti visti per caratterizzare la rapidità di convergenza delle seguenti successioni (convergenti a $x^* = 0$):*

$$\begin{aligned}
\{x_k\} &= \frac{1}{k} \quad \textit{(lineare)} \\
\{x_k\} &= \frac{1}{2^k} \quad \textit{(lineare)} \\
\{x_k\} &= \frac{1}{k!} \quad \textit{(superlineare)} \\
\{x_k\} &= \frac{1}{2^{2^k}} \quad \textit{(quadratica)}
\end{aligned}$$

□

7 Ricerca unidimensionale del passo α_k

In questo capitolo affrontiamo il problema della determinazione del passo α_k a ciascuna iterazione della (10) che dà il nuovo punto x_{k+1}, supponendo che d_k sia una direzione di discesa, ossia che $\nabla f(x_k)^T d_k < 0$. Questa ricerca di α_k prende il nome di *line search* (dal momento che avviene lungo una "linea", ossia la direzione d_k).

Gli algoritmi di line search consistono nel provare iterativamente diversi valori di α, fino a che certe *condizioni di arresto* sono verificate. Chiaramente, una possibilità è quella

di arrestarsi quando risultano soddisfatte sia la (12) che la (13). Questo si può fare (e anzi, si fa in alcuni approcci risolutivi), ma la descrizione dettagliata dei procedimenti che garantiscono di terminare con un valore di α che soddisfi ambedue le relazioni è abbastanza macchinosa ed è al di fuori dello scopo di queste dispense. Sappiamo d'altra parte (Teorema 13) che se si sceglie α_k sufficientemente piccolo, abbiamo la garanzia (in ipotesi di continuità della f, come stiamo supponendo) di riuscire a soddisfare (12). Se però il nuovo punto x_{k+1} risulta troppo vicino al precedente, la diminuzione di f può risultare molto piccola e quindi in definitiva la convergenza può risultare troppo lenta. Dunque, la scelta di α_k deve essere fatta in modo tale da consentire anche un significativo spostamento dal punto x_k, pur senza garantire, in generale, il soddisfacimento della condizione di Wolfe (§7.2).

7.1 Line search "esatta"

Dal momento che la funzione $\phi(\alpha) = f(x_k + \alpha d_k)$ esprime l'andamento della funzione f a partire dal punto x_k, muovendosi nella direzione d_k, può apparire logico cercare quel passo α^* che *minimizza* la $\phi(\alpha)$ (per $\alpha > 0$). Poiché d_k è una direzione di discesa, e l'insieme di livello

$$\{\alpha : \alpha \in R, \alpha \geq 0, \phi(\alpha) \leq \phi(0)\}$$

è chiuso e limitato, esiste sicuramente (Teorema 9) tale valore α^*, e dalle condizioni del 1° ordine si ha

$$\phi'(\alpha^*) = \nabla f(x_k + \alpha^* d_k)^T d_k = 0 \tag{15}$$

La (15) ha un'interessante interpretazione geometrica: scegliendo come passo il valore α^*, il gradiente nel punto incrementato $x_{k+1} = x_k + \alpha^* d_k$ risulta ortogonale alla direzione di ricerca d_k. Dunque, la superficie di livello corrispondente al valore $f(x_k + \alpha^* d_k)$ risulta tangente alla direzione d_k. In altre parole, per cercare il minimo della ϕ si può pensare di procedere nella direzione d_k (che, almeno inizialmente, di sicuro *non* è ortogonale al gradiente) e poi scegliere il punto in cui la ϕ è minima tra quelli in cui d_k risulta ortogonale al gradiente della f in quel punto.

Effettuare una line search esatta è comunque oneroso dal punto di vista pratico, soprattutto se la ϕ è non convessa. Occorrerebbero troppe valutazioni della funzione ϕ (o f, che è la stessa cosa) e del gradiente, anche solo per determinare un minimo locale della ϕ. Più concretamente, ci si può allora orientare a effettuare una minimizzazione non esatta della ϕ, ma ancora in grado di ottenere un'adeguata riduzione del valore della f a costi computazionali contenuti. Il valore $\hat{\alpha}$ prodotto dalla line search può allora vedersi come una *stima* del valore ideale α^*.

```
Metodo_di_Armijo
{
  α := α_0 ;
  while  f(x_k + α d_k) > f(x_k) + γ α ∇f(x_k)^T d_k
         do  α := σ α ;
  α_k := α
}
```

Figura 3: Schema del metodo di Armijo.

7.2 Backtracking e metodo di Armijo

Nel seguito, illustreremo un approccio iterativo che genera i valori di α in modo abbastanza accurato, e converge con accettabile rapidità, pur facendo a meno della verifica puntuale della condizione di Wolfe, e utilizzando esplicitamente solo la condizione di sufficiente riduzione.

L'approccio iterativo che stiamo per descrivere è di tipo *backtracking*. Nel seguito, indicheremo con $\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots$ i valori di α generati alle varie iterazioni, mentre $\hat{\alpha}$ indica il valore restituito dal metodo, e che verrà quindi utilizzato come passo.

L'approccio backtracking consiste nel considerare, inizialmente, un valore α_0 (che va scelto con una certa attenzione, come vedremo dopo). Se già α_0 soddisfa la condizione di riduzione, il procedimento termina e restituisce α_0. Altrimenti, si moltiplica α_0 per un *fattore di contrazione* $0 < \sigma \leq 1/2$ e si prova il valore così generato. Il procedimento prosegue in questo modo fino a trovare un valore $\hat{\alpha} = \sigma^i \alpha_0$ tale da soddisfare la (12). L'idea è che il valore α restituito dal metodo, oltre a soddisfare la (12), non sarà *troppo* piccolo, in quanto c'è da tener presente che il valore trovato all'iterazione precedente, ossia $\sigma^{i-1}\alpha_0$, non era stato ritenuto soddisfacente, ossia era ancora *troppo grande*.

In questa versione-base, l'approccio backtracking prevede di utilizzare a ogni iterazione sempre lo stesso valore del fattore di contrazione σ. In tal caso, il metodo prende il nome di *metodo di Armijo*, ed è riportato in Figura 3.

Come si vede, il metodo di Armijo è estremamente semplice, e anche per questo è uno dei metodi di ricerca unidimensionale più usati. È peraltro possibile dimostrare che scegliendo il passo secondo Armijo si ha effettivamente un metodo di discesa, e che ogni punto limite della successione $\{x_k\}$ è un punto stazionario. Per semplicità, vedremo qui solo la dimostrazione della prima condizione, che è abbastanza immediata.

TEOREMA 22 *Dato un punto x_k e una direzione di discesa d_k, il metodo di Armijo de-*

termina, in un numero finito di iterazioni, un valore di $\hat{\alpha}$ tale che

$$f(x_{k+1}) = f(x_k + \hat{\alpha} d_k) < f(x_k)$$

Dim.– Dimostriamo che il metodo di Armijo *termina*. Se infatti non terminasse, vorrebbe dire che a ogni iterazione risulterebbe sempre vera la condizione di ingresso nello while, ossia, alla j-esima iterazione, essendo $\alpha = \sigma^j \alpha_0$, sarebbe

$$\frac{f(x_k + \alpha_0 \sigma^j d_k) - f(x_k)}{\sigma^j \alpha_0} > \gamma \nabla f(x_k)^T d_k$$

Ora, al crescere dell'indice di iterazione j, si ha che $\sigma^j \to 0$, e dunque il termine di sinistra tende alla derivata direzionale di f in x_k nella direzione d_k, ma allora si ha

$$\nabla f(x_k)^T d_k > \gamma \nabla f(x_k)^T d_k$$

il che è un assurdo, essendo $0 < \gamma < 1$ e $\nabla f(x_k)^T d_k < 0$. Dunque, il metodo di Armijo termina. Di conseguenza, il valore prodotto α_k è tale che $f(x_k + \alpha_k d_k) \leq f(x_k) + \gamma \alpha_k \nabla f(x_k)^T d_k$. Dal fatto che $\gamma \alpha_k \nabla f(x_k)^T d_k < 0$, segue la tesi. □

Un aspetto delicato dei metodi di backtracking consiste nel fatto che il numero di iterazioni necessarie dipende in sostanza dalla scelta del valore iniziale α_0. Questo non è un problema nei metodi tipo-Newton, in quanto come vedremo (§9) in quei casi il metodo prescrive una stima iniziale fissa $\alpha_0 = 1$. In altri casi occorre però regolarsi diversamente (§8).

7.3 Interpolazione

In metodi di backtracking più generali del metodo di Armijo, il valore del fattore di contrazione può variare da un'iterazione all'altra. In effetti, quello che il metodo di Armijo fa a ogni iterazione non è altro che diminuire il valore corrente α_i in modo "controllato", ossia in un modo tale da salvaguardare la convergenza dell'algoritmo complessivo (10). Il fatto di moltiplicare α_i per un certo $\sigma < 1$ è un modo, ma non l'unico, di effettuare questa diminuzione. Vediamo un esempio di un metodo di backtracking in cui a ogni iterazione il nuovo valore di α viene calcolato per mezzo di una semplice *interpolazione*, che descriviamo nel seguito.

L'idea di base dell'interpolazione è quella di utilizzare una *rappresentazione approssimata* della ϕ nell'intervallo di valori di interesse, che sono quelli compresi tra $\alpha = 0$ e $\alpha = \alpha_i$ corrente. Tale rappresentazione è in genere un polinomio, che sceglieremo di secondo o terzo grado (che sono poi i casi di interesse pratico). Anziché moltiplicare il valore α_i per il fattore di contrazione, si cerca allora il punto di minimo $\bar{\alpha}$ di tale polinomio

approssimante nell'intervallo $(0, \alpha_i)$. Se la ϕ non è fortemente irregolare, ci sono buone speranze che l'approssimazione non si discosti troppo dalla realtà, e che quindi il valore $\bar{\alpha}$ sia non troppo distante da α^*.

Nel seguito, vediamo un metodo di interpolazione progettato per limitare il numero di volte che occorre valutare il gradiente della f. Dunque, riscrivendo la (12), come prima cosa si vuole verificare se la stima iniziale α_0 verifica o meno la condizione di riduzione, ossia

$$\phi(\alpha_0) \leq \phi(0) + \gamma\alpha_0\phi'(0) \tag{16}$$

se questa non è verificata, costruiamo un'*approssimazione quadratica* ϕ_q della ϕ, ossia una funzione del tipo

$$\phi_q(\alpha) = a\alpha^2 + b\alpha + c \tag{17}$$

Poiché abbiamo già disponibili tre quantità, vale a dire $\phi(0)$, $\phi'(0)$ e $\phi(\alpha_0)$, scegliamo i parametri della ϕ_q in modo tale che $\phi_q(0) = \phi(0)$, $\phi_q'(0) = \phi'(0)$ e $\phi_q(\alpha_0) = \phi(\alpha_0)$. È facile verificare con semplici passaggi che queste tre condizioni implicano che il polinomio approssimante sia

$$\phi_q(\alpha) = \frac{\phi(\alpha_0) - \phi(0) - \alpha_0\phi'(0)}{\alpha_0^2}\alpha^2 + \phi'(0)\alpha + \phi(0) \tag{18}$$

si osservi che $\phi_q(\alpha)$ risulta essere convessa (perché?). Derivando, si ottiene il punto di minimo, che verrà preso come nuovo valore α_1:

$$\alpha_1 = -\frac{\phi'(0)\alpha_0^2}{2[\phi(\alpha_0) - \phi(0) - \alpha_0\phi'(0)]} \tag{19}$$

A questo punto, si verifica la condizione di riduzione (12) per α_1. Se non è soddisfatta, si può procedere, tenendo conto che ora si dispone anche di un altro valore della ϕ, vale a dire $\phi(\alpha_1)$, e dunque si può tentare una stima più accurata utilizzando un'approssimazione *cubica* della ϕ, ossia

$$\phi_c(\alpha) = a\alpha^3 + b\alpha^2 + \alpha\phi'(0) + \phi(0) \tag{20}$$

(in cui si sono già imposti il valore $\phi(0)$ in 0 e il valore $\phi'(0)$ della derivata nello stesso punto). I valori di a e b possono essere ricavati imponendo il passaggio per i punti $(\alpha_0, \phi(\alpha_0))$ e $(\alpha_1, \phi(\alpha_1))$, ottenendo

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\alpha_0^2\alpha_1^2(\alpha_1 - \alpha_0)} \begin{bmatrix} \alpha_0^2 & -\alpha_1^2 \\ -\alpha_0^3 & \alpha_1^3 \end{bmatrix} \begin{bmatrix} \phi(\alpha_1) - \phi(0) - \phi'(0)\alpha_1 \\ \phi(\alpha_0) - \phi(0) - \phi'(0)\alpha_0 \end{bmatrix} \tag{21}$$

di nuovo, derivando si trova il punto di minimo per ϕ_c, che risulta essere interno all'intervallo $(0, \alpha_1)$, dato da

$$\alpha_2 = \frac{-b + \sqrt{b^2 - 3a\phi'(0)}}{3a} \tag{22}$$

Di nuovo si va a verificare la condizione di riduzione, e se questa non fosse soddisfatta si cercherebbe α_3 utilizzando una approssimazione cubica ancora della forma (20) in cui i valori di a e b sono calcolati per mezzo della (21), ma in cui α_0 e α_1 sono rimpiazzati rispettivamente da α_1 e α_2. Il metodo prosegue, utilizzando a ogni passo, per la determinazione di α_i, un polinomio di terzo grado. I parametri di questa espressione fanno uso della ϕ calcolata negli ultimi due punti, α_{i-2} e α_{i-1}. Benché sia possibile dimostrare che facendo cosi' si ottengono sempre punti interni all'intervallo $(0, \alpha_{i-1})$, può capitare che α_i risulti troppo vicino a α_{i-1} (e allora la line search rischia di essere troppo lenta) o al contrario troppo prossimo a 0 (e allora al solito la diminuzione di ϕ rischia di essere poco significativa). Per evitare questi inconvenienti, si può decidere che quando capitano casi di questo tipo, si pone semplicemente $\alpha_i = \alpha_{i-1}/2$.

8 Il metodo del gradiente

Il metodo del gradiente costituisce l'algoritmo di ottimizzazione più semplice, sia dal punto di vista concettuale che realizzativo. In molti metodi di discesa, la direzione di ricerca d_k viene determinata considerando un'opportuna approssimazione della funzione obiettivo. Nel *metodo del gradiente*, si fa riferimento a un'approssimazione lineare di $f(x_k + d)$, pensata come *funzione del vettore* d. Riprendendo la formula di Taylor arrestata al 1° ordine (2):

$$f(x_k + d) = f(x_k) + \nabla f(x_k)^T d + \beta_1(x_k, d)$$

l'idea del metodo del gradiente è quella di approssimare $f(x_k + d)$ con la funzione

$$\psi_k(d) = f(x_k) + \nabla f(x_k)^T d \tag{23}$$

e di scegliere come direzione di ricerca d_k quella direzione che minimizza $\psi_k(d)$ nella sfera di raggio unitario, ossia la soluzione del problema

$$\begin{aligned} &\min\ \psi_k(d) \\ &\quad \|d\| = 1 \end{aligned}$$

```
Metodo_del_gradiente
{
 Si fissa un punto iniziale x_0 ∈ R^n ; k := 0;
 while  ∇f(x_k) ≠ 0
        {
        si pone d_k := −∇f(x_k)
        si calcola il passo α_k lungo d_k (ad es.con Armijo)
        x_{k+1} := x_k + α_k d_k;
        k := k + 1;
        }
}
```

Figura 4: Il metodo del gradiente.

ovvero

$$\begin{aligned} \min\ & \nabla f(x_k)^T d \\ & \|d\| = 1 \end{aligned}$$

dove in questo caso la norma si suppone essere quella euclidea. Per la disuguaglianza di Hölder (il valore assoluto del prodotto scalare di due vettori non supera il prodotto delle loro norme euclidee), e poiché $\|d\| = 1$, si ha che $|\nabla f(x_k)^T d| \leq \|\nabla f(x_k)\|$, e dunque il minimo si ha scegliendo $d = -\nabla f(x_k)/\|\nabla f(x_k)\|$, che è la direzione dell'antigradiente (scegliendo $d = \nabla f(x_k)/\|\nabla f(x_k)\|$ troveremmo ovviamente il *massimo* di $\psi_k(d)$).

Dal momento che in definitiva $-\nabla f(x_k)$ minimizza la derivata direzionale della f, il metodo del gradiente è anche detto il metodo *della discesa più ripida*. La (10) diviene quindi

$$x_{k+1} := x_k - \alpha_k \nabla f(x_k) \tag{24}$$

Lo schema del metodo del gradiente è riportato in Figura 4.

Per quanto concerne la convergenza del metodo del gradiente, supponendo che le condizioni (12) e (13) siano soddisfatte utilizzando un opportuno algoritmo di line search, per poter applicare il Teorema 21 occorre verificare se è soddisfatta la condizione d'angolo (14):

$$\nabla f(x_k)^T d_k \leq -\epsilon \, \|\nabla f(x_k)\| \, \|d_k\|$$

evidentemente, essendo $d_k = -\nabla f(x_k)$, la condizione d'angolo è soddisfatta scegliendo $\epsilon = 1$. Dunque, il metodo converge. Inoltre, per il Teorema 21 la convergenza è *globale*, e questo costituisce ovviamente uno dei maggiori pregi del metodo del gradiente.

Dunque, qualunque sia il punto iniziale x_0, il metodo del gradiente converge a un punto stazionario. Tuttavia, al variare di x_0 può sia variare il punto stazionario a cui l'algoritmo converge, e sia la rapidità con cui tale convergenza avviene. Per avere un'indicazione quantitativa della rapidità di convergenza del metodo, si può fare riferimento a una funzione quadratica $f(x) = \frac{1}{2}x^TQx + b^Tx + c$. (È chiaro che per trovare il punto stazionario in questo caso non ci sarebbe bisogno di ricorrere a un metodo di discesa, dal momento che $\nabla f(x) = Qx + b$, e dunque basterebbe risolvere un sistema lineare. Tuttavia, qui ci interessa solo valutare la rapidità del metodo del gradiente.) Consideriamo allora una funzione quadratica $f(x) = x^TQx$, con Q matrice simmetrica definita positiva, il cui punto di minimo è evidentemente $x = (0, 0, \ldots, 0)^T$. È interessante osservare che le superfici di livello hanno in questo caso la forma di ellissoidi – ellissi in R^2 – il cui rapporto tra le lunghezze dei vari assi è dato dal rapporto tra i diversi autovalori della matrice (che essendo Q simmetrica, sono tutti reali). Indicando con λ_m e λ_M il più piccolo e il più grande autovalore, si può dimostrare che, utilizzando il metodo del gradiente, ed effettuando a ogni passo la line search esatta, tra i valori della f in due iterazioni successive intercorre la relazione

$$f(x_{k+1}) \leq (\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m})^2 f(x_k) \tag{25}$$

(Si può anche dimostrare che vi sono dei punti iniziali per i quali la (25) vale effettivamente con il segno di uguaglianza.) Ricordando il Teorema 7, si ha che

$$\lambda_m \|x\|^2 \leq x^TQx \leq \lambda_M \|x\|^2$$

da cui, per la (25), ed essendo l'origine il punto di minimo,

$$\lambda_m \|x_{k+1}\|^2 \leq x_{k+1}^T Q x_{k+1} = f(x_{k+1}) \leq (\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m})^2 x_k^T Q x_k \leq (\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m})^2 \lambda_M \|x_k\|^2$$

ossia, in definitiva

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leq (\frac{\lambda_M}{\lambda_m})^{1/2} (\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m}) \tag{26}$$

Dunque, in base alla Definizione 16, il metodo del gradiente ha rapidità di convergenza lineare. Si noti però che nel caso in cui gli autovalori sono tutti uguali (ossia, le superfici di livello sono sferiche), la (26) implica che si ha convergenza in un solo passo. Viceversa, se vi è una grande differenza tra il più grande e il più piccolo autovalore di Q, il valore di f può scendere anche molto lentamente, con convergenza dunque sublineare. Anche se questa analisi è fatta solo per funzioni quadratiche, le conclusioni sono abbastanza generali e infatti, dal punto di vista pratico, il metodo del gradiente viene ritenuto mediamente inefficiente, soprattutto in presenza di funzioni aventi superfici di livello a forte curvatura.

A proposito dell'utilizzo dei metodi di ricerca unidimensionale per la determinazione del passo α, va detto che per la scelta della stima iniziale α_0 non vi sono criteri di scelta specifici per il metodo del gradiente. In questo caso è importante calibrare tale scelta tenendo conto dell'informazione disponibile sulla funzione fino a quel momento. Un modo di procedere abbastanza diffuso e mediamente efficace è quello di considerare che, molto spesso, la variazione che possiamo attenderci per la funzione obiettivo una volta giunti all'iterazione k-sima è dello stesso ordine di grandezza di quella che si è avuta nell'ultima iterazione. Quest'ultima, considerando solo il termine del prim'ordine, è data da:

$$f(x_{k-1} + \alpha_{k-1}d_{k-1}) - f(x_{k-1}) \approx \alpha_{k-1}\nabla f(x_{k-1})^T d_{k-1}$$

(dove – attenzione – qui α_{k-1} indica il valore del passo all'iterazione precedente) e dunque questo criterio corrisponde a scegliere come valore iniziale α_0 del passo all'iterazione k:

$$\alpha_0 = \alpha_{k-1}\frac{\nabla f(x_{k-1})^T d_{k-1}}{\nabla f(x_k)^T d_k}$$

ossia, nel caso del metodo del gradiente:

$$\alpha_0 = \alpha_{k-1}\frac{\|\nabla f(x_{k-1})\|^2}{\|\nabla f(x_k)\|^2}$$

9 Metodo di Newton

I metodi di tipo Newton sono tra i metodi di maggiore importanza nell'ottimizzazione non vincolata. Noi vedremo dapprima il metodo di Newton nella sua forma più semplice, e quindi alcune modifiche che ne migliorano le caratteristiche di convergenza.

9.1 Il metodo di Newton "puro"

Analogamente al metodo del gradiente, anche il metodo di Newton si basa sul concetto di minimizzare un'approssimazione della f, ma stavolta quadratica. Sia f una funzione con Hessiana continua. Dalla formula di Taylor arrestata ai termini del 2° ordine (3), per valori sufficientemente piccoli della norma del vettore incremento h è possibile scrivere:

$$f(x_k + h) \approx f(x_k) + \nabla f(x_k)^T h + \frac{1}{2}h^T \nabla^2 f(x_k) h \tag{27}$$

Indichiamo con $q(h)$ il termine a destra della (27). Annullando il gradiente di $q(h)$ si ha

$$\nabla q(h) = \nabla f(x_k) + \nabla^2 f(x_k)h = 0$$

da cui, se $\nabla^2 f(x_k)$ è non singolare, possiamo ottenere:

$$h^* = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k) \tag{28}$$

```
Metodo_di_Newton
{
 Si fissa un punto iniziale x_0 ∈ R^n;  k := 0;
 while  ∇f(x_k) ≠ 0
        si pone  x_{k+1} := x_k − (∇²f(x_k))^{-1} ∇f(x_k)
        k := k + 1
}
```

Figura 5: Il metodo di Newton "puro".

Si noti che se l'Hessiana è definita positiva, h^* è proprio il vettore che minimizza $q(h)$. Il metodo di Newton consiste nell'utilizzare, come vettore incrementato, proprio h^* dato dalla (28), ossia si ha

$$x_{k+1} := x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \tag{29}$$

e l'algoritmo di ottimizzazione (10) assume la forma indicata in Figura 5.

In questa versione "pura" del metodo di Newton, anziché distinguere la scelta della direzione da quella del passo, viene specificato direttamente il vettore-incremento, o, in altre parole, viene fissata la direzione $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ (che prende il nome di *direzione di Newton*) e lungo questa direzione ci si muove di un passo pari a 1. Peraltro, si noti che il metodo in Figura 5 è applicabile anche se l'Hessiana non è definita positiva, purché risulti non singolare.

Poiché di fatto il metodo di Newton specifica direttamente il vettore incremento, non è possibile invocare direttamente il Teorema 21 per analizzare le proprietà di convergenza dell'algoritmo. In effetti, vale un risultato leggermente diverso, che enunciamo senza dimostrarlo. In questo teorema di fa uso della seguente definizione:

DEFINIZIONE 20 *Si consideri una matrice F di dimensioni $n \times n$, in cui ciascun elemento F_{ij} è una funzione di $x \in R^n$. La F si dice* lipschitziana *su un insieme aperto $\mathcal{D}$ se esiste una costante $L > 0$ tale che, per ogni $x, y \in \mathcal{D}$, si ha*

$$\|F(x) - F(y)\| \leq L\|x - y\|$$

□

In altre parole, una matrice di funzioni è lipschitziana se la sua norma non varia troppo rapidamente spostandosi da un punto a un altro; è dunque una condizione di regolarità abbastanza poco restrittiva. Vale, in definitiva, il seguente teorema

TEOREMA 23 *Data la funzione f, sia essa due volte continuamente differenziabile in R^n. Supponiamo inoltre che:*

(a) esiste un punto x^* tale che $\nabla f(x^*) = 0$ e che $\nabla^2 f(x^*)$ sia non singolare;

(b) $\nabla^2 f(x^*)$ è lipschitziana in R^n

allora esiste una sfera aperta $B(x^*, \epsilon)$ di centro in x^* e raggio ϵ tale che se $x_0 \in B(x^*, \epsilon)$, la successione $\{x_k\}$ generata dalla (29) rimane in $B(x^*, \epsilon)$ e converge a x^* con rapidità di convergenza quadratica. □

L'aspetto più importante del risultato espresso dal Teorema 23 è il fatto che il metodo di Newton converge con rapidità quadratica, il che lo rende significativamente più interessante rispetto al metodo del gradiente. Tuttavia, questo teorema esprime solo un risultato di convergenza *locale*.

9.2 Metodo di Newton modificato

Per ovviare al problema della convergenza locale, si può allora modificare il metodo, utilizzando la struttura dei metodi di discesa (Figura 1).

Precisamente, si può pensare di scegliere a ogni iterazione come direzione d_k la direzione di Newton $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$, ed effettuare poi una line search, ad esempio con Armijo, utilizzando come valore iniziale $\alpha_0 = 1$.

Andiamo allora a esaminare la convergenza globale del metodo di Newton così modificato, considerando dapprima il caso in cui l'Hessiana risulti definita positiva a ogni passo dell'algoritmo, ossia per ogni x_k. Per semplicità notazionale, nei passaggi che seguono indicheremo l'inversa dell'Hessiana calcolata in x_k con B_k, cosicché $d_k = -B_k \nabla f(x_k)$. Supponiamo, indicando con $\lambda_M(B_k)$ e $\lambda_m(B_k)$ rispettivamente il più grande e il più piccolo autovalore dell'inversa dell'Hessiana in x_k, che esistano due costanti M e m tali che

$$0 < m < \lambda_m(B_k) \leq \lambda_M(B_k) \leq M \tag{30}$$

ossia gli autovalori dell'inversa dell'Hessiana sono compresi tra m e M per qualunque k. (Si ricorda che che gli autovalori dell'inversa sono gli inversi degli autovalori.) Si noti che questa condizione è leggermente più restrittiva che non il semplice fatto che $\nabla^2 f(x_k)^{-1} > 0$, in quanto si richiede anche che, al crescere di k, gli autovalori della $\nabla^2 f(x_k)^{-1}$ non divengano né arbitrariamente grandi e né arbitrariamente prossimi a 0.

Si tratta dunque di vedere se la condizione d'angolo è soddisfatta dalla direzione di Newton $d_k = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$. Considerando la derivata direzionale nella direzione d_k nel punto x_k, si ha

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T B_k \nabla f(x_k)$$

e, applicando il risultato espresso nel Teorema 7 (ove al posto di x c'è $\nabla f(x_k)$), otteniamo

$$-\nabla f(x_k)^T B_k \nabla f(x_k) \leq -\lambda_m(B_k)\|\nabla f(x_k)\|^2 \tag{31}$$

Poiché per qualunque vettore d, $\|d\| = (d^T d)^{\frac{1}{2}}$, possiamo scrivere

$$\|d_k\| = \|B_k \nabla f(x_k)\| = (\nabla f(x_k)^T B_k^T B_k \nabla f(x_k))^{\frac{1}{2}}$$

ossia, essendo B_k simmetrica

$$\|d_k\| = \|B_k \nabla f(x_k)\| = (\nabla f(x_k)^T B_k^2 \nabla f(x_k))^{\frac{1}{2}} \tag{32}$$

D'altro canto, ancora ricordando il Teorema 7 (ove ancora, al posto di x poniamo $\nabla f(x_k)$)

$$(\nabla f(x_k)^T B_k^2 \nabla f(x_k))^{\frac{1}{2}} \leq (\lambda_M(B_k^2)\|\nabla f(x_k)\|^2)^{\frac{1}{2}} \tag{33}$$

e, ricordando che gli autovalori del quadrato di una matrice coincidono con i quadrati degli autovalori, si ha in definitiva, da (32) e (33)

$$\|d_k\| \leq \lambda_M(B_k)\|\nabla f(x_k)\| \tag{34}$$

a questo punto, sfruttando la (30) possiamo scrivere, dalla (31),

$$\nabla f(x_k)^T d_k < -m\|\nabla f(x_k)\|^2 \tag{35}$$

e dalla (34)

$$\|\nabla f(x_k)\| \geq \frac{\|d_k\|}{M} \tag{36}$$

Sempre facendo attenzione ai segni, si vede che maggiorando il secondo membro della (35) per mezzo della (36) si ottiene

$$\nabla f(x_k)^T d_k < -\frac{m}{M}\|\nabla f(x_k)\|\ \|d_k\|$$

e risulta dimostrato il seguente teorema:

TEOREMA 24 *La direzione di Newton $d_k = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$ soddisfa la condizione d'angolo con $\epsilon = m/M$.* □

Dunque, dal Teorema 21, se a ogni iterazione l'Hessiana è definita positiva, il metodo di discesa che utilizza la direzione di Newton converge globalmente a un punto stazionario. Peraltro, è possibile dimostrare che da un certo k in poi il passo $\alpha_k = 1$ soddisfa le condizioni del metodo di Armijo, e dunque si può di fatto usare il metodo di Newton "puro", e la convergenza quadratica è ancora assicurata.

Il Teorema 24 mostra che, poiché la direzione di Newton soddisfa la condizione d'angolo, è, tra l'altro, una direzione di discesa. Questo risultato è stato stabilito però sfruttando il fatto che l'Hessiana (e dunque la sua inversa) sia definita positiva. Se invece l'Hessiana non è sempre definita positiva, il Teorema 24 non vale più, e il metodo di discesa può risultare inapplicabile per due motivi: o perché in alcuni punti l'Hessiana può risultare singolare, oppure perché la direzione di Newton non è una direzione di discesa. Inoltre, può anche accadere che, pur risultando una direzione di discesa, la direzione di Newton e il gradiente sono talmente prossimi a essere ortogonali che non conviene seguire la direzione di Newton.

A questi inconvenienti è comunque possibile ovviare in modo molto semplice, utilizzando la direzione dell'antigradiente quando non sia possibile o conveniente seguire la direzione di Newton. Il metodo di Newton modificato in modo da tenere in conto tutte queste possibilità è raffigurato in Figura 6, in cui $\epsilon > 0$ è una quantità sufficientemente piccola. In sostanza, come si può vedere il metodo sceglie come direzione di discesa quella di Newton, la sua opposta o l'antigradiente, e successivamente effettua una line search con metodi standard, in modo da garantire la convergenza globale.

Si noti che l'algoritmo in Figura 6 soddisfa la condizione d'angolo, e dunque le proprietà di convergenza globale dell'algoritmo sono conservate.

10 Metodo del gradiente coniugato[2]

Studiando il metodo del gradiente, abbiamo osservato che la sua convergenza risulta essere notevolmente influenzata dagli autovalori della matrice Hessiana. In particolare, dalla (25) si vede che, quanto maggiore è il rapporto fra autovalore massimo e minimo, tanto peggiori sono le prestazioni del metodo. Il caso migliore si verifica quando gli autovalori dell'Hessiana sono tutti uguali: nel caso di una funzione quadratica, questo corrisponde al fatto che le curve di livello sono sfere (in R^n) e il metodo del gradiente riesce a trovare l'ottimo in un singolo passo (se la line search è esatta). A parte questo caso, in generale la direzione di massima pendenza risulta la scelta migliore solo localmente, ma utilizzata per uno spostamento "grande" non risulta essere sempre vantaggiosa.

[2]In collaborazione con Bernardetta Addis.

```
Metodo_di_Newton_Modificato
{
 Si fissa un punto iniziale x_0 ∈ R^n ; k := 0;
 while  ∇f(x_k) ≠ 0
        {
        calcola ∇²f(x_k)
         se ∇²f(x_k) e' singolare, poni d_k := −∇f(x_k)
        altrimenti
           {
            s := −(∇²f(x_k))^{-1} ∇f(x_k)
            se |∇f(x_k)^T s| < ε‖∇f(x_k)‖ ‖s‖, poni d_k := −∇f(x_k)
            altrimenti
               poni d_k := s oppure d_k := −s
               a seconda di quale delle due e' una direzione di discesa
           }
        si calcola il passo α_k lungo d_k (ad es.con Armijo)
        x_{k+1} := x_k + α_k d_k;
        k := k + 1;
        }
}
```

Figura 6: Il metodo di Newton modificato nella versione globalmente convergente.

Poiché non possiamo sempre supporre di conoscere la natura dell'Hessiana, è opportuno sviluppare metodi che riescano ad avere buone prestazioni nel caso generale. Molti di questi metodi possono pensarsi come modifiche al metodo del gradiente volte a correggere le sue debolezze. Ad esempio, il metodo di Newton può vedersi come un metodo che *modifica* la direzione dell'antigradiente, moltiplicandola per l'inversa dell'Hessiana. Questo può essere interpretato come un cambio di coordinate che permette di trasformare le curve di livello da ellissoidi a sfere. Infatti, nel caso di Hessiana uguale alla matrice identità, i due metodi (Newton e gradiente) sono equivalenti.

Nel seguito vediamo un metodo, detto *metodo dei gradienti coniugati*, che si inserisce in questa famiglia di metodi, ma al contrario del metodo di Newton non necessita del calcolo esplicito della inversa della matrice Hessiana, che per problemi di grandi dimensioni può essere molto oneroso. Inoltre, non incontra problemi qualora la Hessiana sia semidefinita.

Per comprendere l'idea generale su cui si basa il metodo, consideriamo dapprima il caso di funzioni quadratiche. In particolare, consideriamo anzitutto il caso in cui l'Hessiana sia *diagonale*. In questo caso, le curve di livello sono ellissoidi (in R^n) con assi paralleli agli assi coordinati, e possiamo pensare di ottimizzare la nostra funzione muovendoci *sequenzialmente* lungo gli assi coordinati. Infatti, la funzione obiettivo può essere considerata come somma di n funzioni quadratiche monodimensionali:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(q_{ii}x_i^2 + b_i x_i) + c \tag{37}$$

È evidente che in questo caso, minimizzando *in modo esatto* sequenzialmente rispetto a $x_1, x_2, \ldots, x_n$, ovvero rispetto alle direzioni rappresentate dagli assi coordinati, si ottiene l'ottimo desiderato, in esattamente n passi. Ci chiediamo allora se la stessa idea (minimizzare sequenzialmente rispetto alle diverse variabili) possa essere generalizzata al caso di funzione quadratica generica, pur mantenendo l'ipotesi che la funzione sia convessa. Introduciamo a questo proposito il concetto di *direzioni Q-coniugate*.

DEFINIZIONE 21 *Data una matrice Q simmetrica e definita positiva, un insieme di vettori $d_1, d_2, \ldots, d_n$ si dicono* Q-coniugati *se sono linearmente indipendenti e inoltre si ha che* $d_i^T Q d_j = 0 \quad \forall i \neq j$.

Osserviamo che le direzioni degli assi cartesiani sono I-coniugate (o, più in generale, Q-coniugate rispetto ad una generica matrice Q diagonale), e che dunque la proprietà di coniugazione ($d_i^T Q d_j = 0$) può essere interpretata come una generalizzazione della proprietà di ortogonalità ($d_i^T d_j = 0$, ovvero $d_i^T I d_j = 0$).

Vediamo ora cosa succede se, nella nostra funzione obiettivo quadratica, Q è definita positiva ma non diagonale, e abbiamo a disposizione n direzioni Q-coniugate. Essendo

le n direzioni $d_1, d_2, \ldots, d_n$ linearmente indipendenti, rappresentano una base in R^n, e dunque un vettore generico x può essere scritto come loro combinazione lineare:

$$x = \sum_{i=1}^{n} y_i d_i$$

Possiamo allora riscrivere la nostra funzione obiettivo

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

come

$$\begin{aligned} f(y) &= \frac{1}{2}(\sum_{i=1}^{n} y_i d_i^T) Q (\sum_{j=1}^{n} y_j d_j) + b^T (\sum_{i=1}^{n} y_i d_i) + c \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j d_i^T Q d_j + \sum_{i=1}^{n} y_i b^T d_i + c \end{aligned}$$

e quindi, ponendo $q'_{ii} = d_i^T Q d_i$ e $b'_i = b^T d_i$, e sfruttando la definizione di direzioni Q-coniugate, si ha

$$f(y) = \frac{1}{2} \sum_{i=1}^{n} (q'_{ii} y_i^2 + b'_i y_i) + c$$

In questo modo, la f ha la stessa forma della (37), e dunque, ottimizzando in successione rispetto alle variabili y_i (ossia rispetto alle direzioni Q-coniugate, in sequenza arbitraria), si giunge all'ottimo.

Il metodo del gradiente coniugato costruisce dunque una sequenza di direzioni $d_1, d_2, \ldots, d_n$, effettuando lungo ognuna di esse una ricerca lineare *esatta*. La differenza sostanziale rispetto ai metodi di discesa precedentemente visti è che le direzioni scelte hanno la caratteristica di essere coniugate rispetto alla matrice Hessiana.

Vediamo ora *come* ottenere le n direzioni coniugate. Al primo passo dell'algoritmo, la direzione scelta d_0 è semplicemente quella dell'antigradiente. A ciascun passo successivo si calcola una direzione in modo che risulti coniugata rispetto alla precedente. Come ora vedremo (senza dimostrazioni), ciò è possibile in modo semplice: il concetto è che, all'iterazione $(k+1)$-esima, si considera come direzione di discesa una direzione ottenuta sommando all'antigradiente $-\nabla f(x_{k+1})$ la direzione seguita al passo precedente d_k, moltiplicata per un coefficiente β_{k+1}:

$$d_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} d_k \tag{38}$$

Trasponendo e postmoltiplicando ambo i membri della (38) per Qd_k, otteniamo:

$$d_{k+1}^T Q d_k = -\nabla f(x_{k+1})^T Q d_k + \beta_{k+1} d_k^T Q d_k$$

```
Metodo_gradienti_coniugati
{
  Si fissa un punto iniziale x_0 ∈ R^n;  d_0 = −∇f(x_k);  k := 0;
  while  ∇f(x_k) ≠ 0
        {
        α_k := arg min_{α≥0} f(x_k + αd_k)
        x_{k+1} := x_k + α_k d_k;
        β_{k+1} := (∇f(x_{k+1})^T ∇f(x_{k+1})) / (∇f(x_k)^T ∇f(x_k));
        d_{k+1} := −∇f(x_{k+1}) + β_{k+1} d_k;
        k := k + 1;
        }
}
```

Figura 7: Algoritmo del gradiente coniugato.

Siccome noi vogliamo che d_{k+1} e d_k siano direzioni Q-coniugate, deve essere:

$$d_{k+1}^T Q d_k = 0$$

e si ottiene quindi

$$\beta_{k+1} = \frac{\nabla f(x_{k+1})^T Q d_k}{d_k^T Q d_k} \tag{39}$$

Si potrebbe dimostrare che tutte le direzioni ottenute in questo modo durante l'algoritmo, finché non si verifica la condizione di gradiente nullo, sono Q-coniugate. Peraltro, sfruttando alcune caratteristiche del metodo[3], si può dimostrare che per ogni k vale la seguente relazione:

$$\beta_{k+1} = \frac{\nabla f(x_{k+1})^T \nabla f(x_{k+1})}{\nabla f(x_k)^T \nabla f(x_k)} \tag{40}$$

La (40) risulta particolarmente utile nel voler applicare l'approccio anche a funzioni non quadratiche. Si noti che utilizzando la (40), possiamo infatti calcolare β_{k+1} a ogni iterazione *senza bisogno di utilizzare la matrice Hessiana.*

In effetti, mentre il metodo del gradiente coniugato (riassunto in Fig.7) converge in n passi al punto di minimo di una funzione quadratica convessa, qualunque sia il punto di partenza, sussistono interessanti proprietà di convergenza anche nel caso più generale

[3]È possibile dimostrare che tutti i gradienti nei punti via via generati sono tra loro ortogonali, ossia $\nabla f(x_k)^T \nabla f(x_h) = 0$ per $h \neq k$, e che inoltre tutti i gradienti sono pure ortogonali alle direzioni coniugate generate fino a quel momento, ossia $\nabla f(x_k)^T d_i = 0$ per $i = 1, \ldots, k-1$.

di funzioni convesse, o addirittura funzioni non convesse. L'importanza del metodo sta proprio nel fatto che, seppure in generale non è garantita la terminazione dopo n passi, si ha tipicamente una rapidità di convergenza molto buona, ed è possibile dimostrare alcuni risultati specifici. In particolare, si considera una tecnica che consiste nel far "ripartire" l'algoritmo ogni n iterazioni, ossia, ogni n iterazioni, prendere come direzione di discesa l'antigradiente anziché usare la (38) – o, equivalentemente, porre $\beta_k = 0$ quando k è un multiplo di n. In tal caso è possibile dimostrare che, sotto alcune ipotesi tecniche non particolarmente restrittive (tra cui il fatto che la f sia Lipschitziana), la sequenza dei punti $\{x_0, x_n, x_{2n}, \ldots\}$ ottenuti ogni n passi, e ponendo $\beta_k = 0$ se $k \bmod n = 0$, converge *globalmente* a un punto stazionario, *con rapidità di convergenza quadratica*.

Occorre sottolineare che le notevoli proprietà di convergenza del metodo del gradiente coniugato con "restart" sono garantite se la line search è effettuata in modo *esatto*. Vale la pena però accennare al fatto che l'algoritmo conserva molte delle sue caratteristiche di convergenza anche utilizzando altri metodi per la line search, ma occorre fare una certa attenzione: ad esempio, non è scontato che applicando ancora la (38), se x_{k+1} non è stato ottenuto tramite una line search esatta, la d_{k+1} sia ancora una direzione di discesa. In molti casi può essere necessario scegliere α in modo da soddisfare esplicitamente le condizioni di Wolfe. Per i dettagli si rimanda ai testi specialistici[3]. In ogni caso, il fatto che non sia necessario calcolare la Hessiana rende il metodo particolarmente indicato per problemi di grandi dimensioni.

La Tabella 1 confronta i tre metodi di ottimizzazione non vincolata visti, riassumendone le principali caratteristiche.

[3] J. Nocedal, S.J.Wright, Numerical Optimization, Springer-Verlag, 1999.

	metodo del gradiente	metodo di Newton	metodo del gradiente coniugato
Informazioni a ogni iterazione	$f(x_k)$ e $\nabla f(x_k)$	$f(x_k)$, $\nabla f(x_k)$ e $\nabla^2 f(x_k)$	$f(x_k)$ e $\nabla f(x_k)$
Calcoli richiesti a ogni iterazione	calcolo di $\nabla f(x_k)$	calcolo di $\nabla f(x_k)$, $\nabla^2 f(x_k)$ e soluzione del sistema lineare $\nabla^2 f(x_k)d = -\nabla f(x_k)$	calcolo di $\nabla f(x_k)$
Convergenza	globale	locale nella versione "pura", globale con modifiche	globale nella versione con "restart"
Comportamento con funzioni quadratiche	convergenza asintotica	convergenza in un solo passo	convergenza in al più n passi
Rapidità di convergenza	lineare	quadratica	quadratica nella versione con "restart"

Tabella 1: Confronto tra metodi di ottimizzazione non vincolata.