Appunti di Metodi di Ottimizzazione

Ernesto Palchetti

30 giugno 2023

Indice

Capitolo 1

Introduzione

1.1 Problemi di Ottimizzazione

L'ottimizzazione prende un problema reale, definisce un modello matematico che descriva il problema e realizza un algoritmo che ottenga una soluzione al problema.

Il generico problema di ottimizzazione è della forma

$$\min_{x \in S} f(x)$$

dove x è la variabile di decisione (quella su cui abbiamo controllo), S è l'insieme ammissibile, definito da vincoli, e f è la funzione obiettivo, ossia la quantità che vogliamo, senza perdere di generalità, minimizzare[1].

Esempio 1.1.1 (Selezione del Portafoglio)**.** Si hanno n asset su cui è possibile investire. $x_1, ..., x_n$ è la frazione del budget che investiamo in ciascun asset. x_i è la frazione di budget investita nell'asset i. I vincoli sono $x_i \geq 0 \ \forall i$ e $\sum_{i=1}^{n} x_i = 1$. Se definiamo $e = (1, ..., 1)^T$, la seconda diventa

$$e^T x = 1.$$

Definiamo μ_i il ritorno atteso dell'investimento i-esimo e σ_{ij} la covarianza degli investimenti i e j. Se $i = j$ si ha la varianza (ossia la variabilità rispetto a μ_i).

La teoria ci dice che la funzione obiettivo è

$$\delta(x) = x^T \Sigma x - \mu^T x.$$

Il termine $-\mu^T x$ premia i ritorni attesi alti (ricordiamo che $\delta(x)$ è da minimizzare). Per quanto riguarda il termine

$$x^T \Sigma x = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j,$$

se la quantità $\sigma_{ij} x_i x_j$ è alta significa che abbiamo investito tanto su una coppia di investimenti con alta variabilità, ossia abbiamo rischiato tanto.

$$\min_{\substack{e^T x = 1 \\ x \geq 0}} x^T \Sigma x - \mu^T x.$$

[1] Se volessimo massimizzare, infatti, si tratterebbe di minimizzare $-f$.

I problemi di ottimizzazione si dividono in

- ottimizzazione continua: $x \in \mathbb{R}^n$;
- ottimizzazione intera: $x \in \mathbb{Z}^n$. Un caso particolare è $x \in \mathbb{Z}_2^n$, ossia quella binaria;
- ottimizzazione mista-intera: $x \in \mathbb{R}^n$, $z \in \mathbb{Z}^n$.

L'ottimizzazione continua si divide in lineare, quando la funzione obiettivo e i vincoli sono espressi da funzioni lineari ($f(x) = a^T x$ e $Ax \geq b$) e non lineare altrimenti.

Possiamo dividere, inoltre, in ottimizzazione differenziabile e non differenziabile a seconda che sia disponibile o meno il gradiente della funzione obiettivo. Infine, si parla di ottimizzazione vincolata se $S \subset \mathbb{R}^n$, ossia l'insieme ammissibile è sottoinsieme proprio di $\mathbb{R}^n$, oppure non vincolata se $S = \mathbb{R}^n$.

Ci occuperemo principalmente di ottimizzazione continua, non lineare e differenziabile, sia vincolata che non.

1.2 Problemi di Stima di Modelli Matematici

Studiamo un sistema fisico che, preso in input x, restituisce y tramite una funzione incognita f. Scegliamo dei parametri μ e cerchiamo di definire un modello matematico che, tramite una funzione $\tilde{f}(\cdot;\mu)$, trovano un outpu approssimato $\tilde{f}(x;\mu) = \tilde{y}$.

Il sistema fisico ci permette di effettuare delle misurazioni per ottenere delle coppie (x_i, y_i). La scelta dei parametri è atta a minimizzare la funzione di errore $e\big((\tilde{f}(x;\mu)), y\big)$

$$\sum_{\mu} \sum_{i=1}^{N} e\big((\tilde{f}(x_i;\mu)), y_i\big).$$

Di consueto consideriamo l'errore quadratico

$$e\big((\tilde{f}(x;\mu)), y\big) = (y_i - \tilde{f}(x_i;\mu))^2,$$

oppure assoluto

$$e\big((\tilde{f}(x;\mu)), y\big) = |y_i - \tilde{f}(x_i;\mu)|.$$

In questi casi il problema diventa

$$\min_{\mu} ||\tilde{f}(x;\mu) - y||_2^2, \qquad \min_{\mu} ||\tilde{f}(x;\mu) - y||_1.$$

Questo è un esempio di ottimizzazione non vincolata.

Definizione 1.2.1 (Norma)**.** Una norma è una funzione $|| \cdot || : \mathbb{R}^n \to \mathbb{R}^+$ con le seguenti proprietà:

(i) $||x|| \geq 0 \ \forall x \in \mathbb{R}^n$;

(ii) $||x|| = 0 \ \Leftrightarrow \ x = 0$;

(iii) $||\alpha x|| = |\alpha| ||x|| \ \forall x \in \mathbb{R}^n, \ \forall \alpha \in \mathbb{R}$ (omogeneità);

(iv) $||x + y|| \leq ||x|| + ||y||$ (disuguaglianza triangolare.)

Esempio 1.2.1. Esempi di norme sono

- norma euclidea
$$||x||_2 = \sqrt{\sum x_i^2};$$
- norma uno
$$||x||_1 = \sum |x_i|;$$
- norma infinito
$$||x||_\infty = \max_i |x_i|.$$

1.3 Apprendimento Supervisionato

$$x \in \mathbb{R}^n \xrightarrow{f} y$$

y può appartenere a $\mathbb{R}$ e si parla di *regressione* oppure $y \in \mathbb{Z}_2$ e si parla di *classificazione*.

La forma di f è spesso ignota, dobbiamo quindi scegliere una classe di funzioni $\mathcal{H}$ da cui pescare $\tilde{f}$. Una volta scelta, $\tilde{f} \in \mathcal{H}$, essa dipende da dei parametri w detti *pesi*. Come vengono scelti i pesi?

Definiamo una misura di errore (*Loss*)

$$\mathcal{L}\big(f(\cdot; w), y\big).$$

I pesi vengono scelti per risolvere

$$\min_w \mathbb{E}_{p(x,y)}\big[\mathcal{L}(f(x; w), y)\big].$$

Questa, detta *funzione di rischio*, dipende dalla distribuzione di probabilità dei dati, spesso ignota. Occorre quindi trovare un surrogato. Nell'ambito dell'apprendimento supervisionato eseguiamo delle misurazioni

$$D = \{(x_i, y_i), i = 1, ..., N\},\ |D| = N.$$

Possiamo quindi pensare di minimizzare la norma degli errori commessi sui dati. Tale funzione $\mathcal{L}(w)$ è detta funzione di rischio empirico.

Ciò che rischiamo minimizzando questa è che il modello si comporti perfettamente sui dati già ottenuti ma non sia flessibile su nuovi. Pertanto si tende a minimizzare

$$\mathcal{L}(w) + \Omega(w),$$

dove Ω è detta *regolarizzatore* e quantifica la complessità del modello. I regolarizzatori più comuni sono $||w||_2$, $||w||_1$ e $||w||_0$. Quest'ultima, usata di rado, conta il numero di valori non nulli di w.

1.4 Richiami di Spazi Metrici

Definizione 1.4.1 (Punto di Minimo Globale)**.** Un punto x^* si dice *punto di minimo globale* per il problema $\mathcal{P}$ se

$$f(x^*) \leq f(x),\ \forall x \in S.$$

$f(x^*)$ è detto minimo globale di $\mathcal{P}$.

Spesso non siamo interessati a $f(x^*)$ ma a x^* e quindi cerchiamo

$$\arg\min_{x \in S} f(x).$$

Esempio 1.4.1. Se $S = \emptyset$ non si hanno soluzioni ottimali poiché non vi sono soluzioni ammissibili. Se f è illimitata inferiormente su S non avrà minimo. Casi come $f = e^{-x}$ non ammettono minimo.

Definizione 1.4.2. L'insieme $S \subseteq \mathbb{R}^n$ si dice aperto se $\forall \bar{x} \in S$, $\exists \varepsilon > o$ tale che

$$B_\varepsilon(\bar{x}) \subset S,$$

dove

$$B_\varepsilon(\bar{x}) = \big\{x \in \mathbb{R}^n \mid ||\bar{x} - x|| \leq \varepsilon\big\}.$$

Definizione 1.4.3. S si dice chiuso se il suo complementare è un aperto.

Definizione 1.4.4. S si dice limitato se $\exists M > 0$ tale che $||x|| \leq M$ $\forall x \in S$.

Definizione 1.4.5. S si dice compatto se soddisfa le seguenti proprietà:

(i) $\forall \{x_k\} \subseteq S$, $\exists X \subseteq \{x_k\}$ sotto-successione convergente;

(ii) $\forall \{x_k\} \subseteq S$ tale che $\lim_{k\to\infty} x_k = \bar{x}$, $\bar{x} \in S$.

Lemma 1.4.1. *Sia $S \subseteq \mathbb{R}^n$, allora*

- *S è limitato se e solo se vale la proprietà (i) della definizione precedente;*
- *S è chiuso se e solo se vale la (ii);*
- *S è compatto se e solo se è chiuso e limitato.*

Teorema 1.4.1 (Weierstrass)**.** *Sia $S \subset \mathbb{R}^n$ compatto e sia $f : \mathbb{R}^n \to \mathbb{R}$ continua, allora f ammette minimo su S.*

Dimostrazione.
Sia $Y = f(S)$ l'immagine di S tramite f.

$$Y = \{y \in \mathbb{R} \mid \exists x \in S \ : \ f(x) = y\}.$$

$Y \subseteq \mathbb{R}$. Sia

$$l = \inf f(S), \qquad \exists \{y_k\} \subseteq f(S) \text{ tc } \lim_{k\to\infty} y_k = l.$$

Definiamo $\{x_k\}$ tale che $f(x_k) = y_k$ $\forall k$ (x_k esiste poiché $y_k \in f(S)$). $\{x_k\} \subseteq S$ compatto (quindi limitato), allora esiste una sotto-successione $\{x_{k_j}\} \subseteq \{x_k\}$ tale che

$$\lim_{j\to\infty} x_{k_j} = \bar{x}.$$

Dalla chiusura di S, $\bar{x} \in S$. Dato che f è continua, se $x_{k_j} \to \bar{x}$, $f(x_{k_j}) \to f(\bar{x})$ e $\{f(x_{k_j})\} \subseteq \{y_k\}$, allora

$$f(\bar{x}) = l \in Y$$

ed è un minimo. Infine, $\bar{x}$ è punto di minimo globale. □

Non è detto che un insieme ammissibile sia compatto. Se $S = \mathbb{R}^n$, il Teorema di Weierstrass non vale.

Definizione 1.4.6 (Insieme di Livello)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$, definiamo l'insieme di livello α

$$\mathcal{L}_f(\alpha) = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}.$$

Proposizione 1.4.1. *Sia $f : \mathbb{R}^n \to \mathbb{R}$, se f ha un insieme di livello compatto, allora ammette minimo su $\mathbb{R}^n$.*

Dimostrazione.
Supponiamo $\mathcal{L}_f(\alpha)$ compatto, per Weierstrass f ammette minimo su $\mathcal{L}_f(\alpha)$. Sia

$$x^* = \min_{x \in \mathcal{L}_f(\alpha)} f(x).$$

Sia $z \in \mathbb{R}^n$, possiamo avere due situazioni:

- Se $z \in \mathcal{L}_f(\alpha)$, allora $f(x^*) \leq f(z)$ per definizione di minimo;
- Se $z \notin \mathcal{L}_f(\alpha)$, allora
$$f(z) > \alpha \geq f(x^*).$$

Allora $f(x^*) \leq f(z)\ \forall z \in \mathbb{R}^n$ ed è dunque un minimo globale. □

Definizione 1.4.7 (Funzione Coerciva)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$, f si dice *coerciva* se, $\forall \{x_k\} \subset S$ tale che $||x_k|| \to \infty$,

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

Proposizione 1.4.2. *f è coerciva se e solo se f ha tutti insiemi di livello compatti.*

Corollario 1.4.1. *$f : \mathbb{R}^n \to \mathbb{R}$ coerciva, allora ammette minimo in $\mathbb{R}^n$.*

Osservazione 1.4.1. La condizione di coercività è necessaria ma non sufficiente.

Esempio 1.4.2 (Funzioni Coercive)**.** Le norme sono funzioni coercive, dalla definizione stessa di coercività.

Osservazione 1.4.2. Se abbiamo un problema di ottimizzazione del tipo

$$\min_{z \in \mathbb{R}^n} f(x),$$

con f limitata inferiormente, possiamo aggiungere un termine di penalità e risolvere

$$\min_{x \in \mathbb{R}^n} f(x) + \tau ||x||^2,$$

problema in cui la funzione obiettivo è coerciva, pertanto ammette soluzione.

1.4.1 Funzioni Quadratiche

Definizione 1.4.8 (Matrice semidefinita positiva)**.** Una matrice quadrata $Q \in \mathbb{R}^{n \times n}$ è semidefinita positiva se

$$x^T Q x \geq 0 \qquad \forall x \in \mathbb{R}^n.$$

Definizione 1.4.9 (Matrice definita positiva)**.** Una matrice quadrata $Q \in \mathbb{R}^{n \times n}$ è definita positiva se

$$x^T Q x > 0 \qquad \forall x \neq 0.$$

Proposizione 1.4.3. *Se Q è simmetrica, allora tutti i suoi autovalori sono reali.*

$$\lambda_1, ..., \lambda_n \in \mathbb{R}.$$

Se Q è simmetrica e semidefinita positiva, gli autovalori sono tutti non negativi. Se è simmetrica e definita positiva, gli autovalori sono tutti positivi.

Teorema 1.4.2 (Min-Max Theorem)**.** *Se Q è una matrice simmetrica,*

$$x^T Q x \geq \lambda_{min} ||x||^2,$$

dove λ_{min} è il minimo autovalore di Q. Inoltre

$$|c^T x| \leq ||c|| ||x||, \quad c, x \in \mathbb{R}^n. \tag{1.1}$$

Osservazione 1.4.3. Se sviluppiamo l'equazione (1.1), otteniamo

$$c^T x \leq ||c|| ||x||, \qquad c^T x \geq -||c|| ||x||. \tag{1.2}$$

Definizione 1.4.10. Una funzione $f : \mathbb{R}^n \to \mathbb{R}$ si dice quadratica se è della forma

$$f(x) = \frac{1}{2} x^T Q x + c^T x, \tag{1.3}$$

con Q matrice simmetrica.

Osservazione 1.4.4. Possiamo sviluppare l'Equazione (1.3) ottenendo

$$f(x) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j q_{ij} + \sum_{i=1}^{n} c_i x_i.$$

Proposizione 1.4.4. *Una funzione quadratica f è coerciva se e solo se Q è definita positiva.*

Dimostrazione.

$\Leftarrow$ Se Q è definita positiva, dall'Equazione (1.2) e dal Teorema 1.4.2,

$$f(x) = \frac{1}{2} x^T Q x + c^T x \geq \frac{1}{2} \lambda_{min} ||x||^2 - ||c|| ||x|| = ||x||^2 \left(\frac{1}{2} \lambda_{min} - \frac{||c||}{||x||} \right). \tag{1.4}$$

Dato che $\lambda_{min} > 0$ perché Q è definita positiva, se $||x|| \to +\infty$, la quantità in (1.4) tende a $+\infty$ e con essa anche f.

$\Rightarrow$ Supponiamo per assurdo che Q non sia definita positiva.

$$\exists y \in \mathbb{R}^n, \ y \neq 0 \text{ tc } y^T Q y \leq 0.$$

Senza perdere di generalità, supponiamo che $c^T y < 0$. Costruiamo una sequenza $\{x_k\}$ definita da

$$x_k = ky \ \forall k.$$

$$f(x_k) = \frac{1}{2} k^2 y^T Q y + k c^T y = k \left(\frac{1}{2} k y^T Q y + c^T y \right). \tag{1.5}$$

$y^T Q y \leq 0$ per ipotesi, $c^T y < 0$ [2], per $k \to +\infty$, la quantità in (1.5) non può tendere a $+\infty$ ma $||x_k|| = |k| ||y|| > 0$ e

$$\lim_{k \to +\infty} ||x_k|| = +\infty.$$

La funzione f è dunque non coerciva.

[2] Se fosse positiva possiamo comunque trovare un valore di k per cui $2k y^T Q y \leq -c^T y$.

□

Corollario 1.4.2. *Se la funzione obiettivo f di un problema di ottimizzazione è quadratica, allora il problema ammette una soluzione se Q è definita positiva.*

Osservazione 1.4.5. Spesso il problema non risiede nel trovare una soluzione ma nel verificare che essa lo sia effettivamente. Quello che facciamo talvolta è accontentarci di una soluzione più debole.

Definizione 1.4.11 (Punto di Minimo Locale)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$ e $S \subseteq \mathbb{R}^n$, si dice che $\bar{x}$ è un *punto di minimo locale* se

$$f(\bar{x}) \leq f(x), \quad \forall x \in S \cap B(\bar{x}, \varepsilon), \ \exists \varepsilon > 0. \tag{1.6}$$

Corollario 1.4.3. *Se x^* è minimo globale, è anche minimo locale.*

Dimostrazione.
Deriva direttamente dalle definizioni di minimo locale e globale. □

Osservazione 1.4.6. Esistono situazioni in cui il minimo globale non è difficile da trovare.

1.4.2 Convessità

Definizione 1.4.12. L'insieme S si dice convesso se $\forall x, y \in S$ e $\forall \lambda \in [0, 1]$ vale che

$$\lambda x + (1 - \lambda)y \in S.$$

Definizione 1.4.13. Sia $S \subseteq \mathbb{R}^n$ convesso e f continua, si dice che f è convessa su S se $\forall x, y \in S$ e $\forall \lambda \in [0, 1]$,

$$f\big((1 - \lambda)x + \lambda y\big) \leq (1 - \lambda)f(x) + \lambda f(y).$$

Esempio 1.4.3. $B(\bar{x}, \varepsilon)$ è convessa. Infatti, per $\lambda \in [0, 1]$, consideriamo la distanza tra il punto $\lambda y + (1 - \lambda)z$ e $\bar{x}$

$$||\lambda y + (1 - \lambda)z - \bar{x}|| = ||\lambda y + (1 - \lambda)z - \lambda\bar{x} - (1 - \lambda)\bar{x}|| =$$

$$= ||\lambda(y - \bar{x}) + (1 - \lambda)(z - \bar{x})|| \leq ||\lambda(y - \bar{x})|| + ||(1 - \lambda)(z - \bar{x})|| =$$

$$= \lambda||y - \bar{x}|| + (1 - \lambda)||z - \bar{x}|| \leq \lambda\varepsilon + (1 - \lambda)\varepsilon = \varepsilon.$$

Allora $\big(\lambda y + (1 - \lambda)z\big) \in S$.
$||x|| : \mathbb{R}^n \to \mathbb{R}$ è una funzione convessa.

$$||\lambda x + (1 - \lambda)y|| \leq ||\lambda x|| + ||(1 - \lambda)y|| = \lambda||x|| + (1 - \lambda)||y||.$$

Questa è proprio la definizione di convessità.

Proposizione 1.4.5. *Sia $S \subset \mathbb{R}^n$ convesso e $f : \mathbb{R}^n \to \mathbb{R}$ continua e convessa, allora se $\bar{x}$ è punto di minimo locale per f su S, $\bar{x}$ è punto di minimo globale per f su S.*

Dimostrazione.
Sia $\bar{x}$ minimo locale per f su S e sia $y \in S$. Per la convessità di f,

$$f\big((1 - \lambda)\bar{x} + \lambda y\big) \leq (1 - \lambda)f(\bar{x}) + \lambda f(y), \quad \forall \lambda \in [0, 1].$$

Detto $z := (1-\lambda)\bar{x} + \lambda y$, per la convessità di S, $z \in S$. Se prendiamo $\lambda > 0$ ma sufficientemente piccolo, $\exists \varepsilon$ tale che

$$z \in B(\bar{x}, \varepsilon),$$

e in essa vale la definizione di minimo locale, per cui

$$f(\bar{x}) \leq f(z) \leq (1-\lambda)f(\bar{x}) + \lambda f(y),$$

$$f(\bar{x}) \leq f(\bar{x}) - \lambda f(\bar{x}) + \lambda f(y)$$

$$\lambda f(\bar{x}) \leq \lambda f(y) \;\Leftrightarrow\; f(\bar{x}) \leq f(y), \quad \forall y \in S.$$

□

Definizione 1.4.14 (Funzione Strettamente Convessa)**.** Sia $S \subseteq \mathbb{R}^n$ convesso, $f : \mathbb{R}^n \to \mathbb{R}$ continua è detta *strettamente convessa* su S se $\forall x, y \in S$ e $\forall \lambda \in (0,1)$,

$$f\big((1-\lambda)x + \lambda y\big) < (1-\lambda)f(x) + \lambda f(y).$$

Esempio 1.4.4. Una parabola rivolta verso l'alto è una funzione strettamente convessa.

Proposizione 1.4.6. *Sia $S \subseteq \mathbb{R}^n$ convesso e $f : \mathbb{R}^n \to \mathbb{R}$ continua e strettamente convessa, se x^* è punto di minimo globale, allora è unico.*

Dimostrazione.
Supponiamo per assurdo che $\bar{x}$ e $\bar{y}$ siano due punti di minimo globali distinti,

$$f(\bar{x}) = f(\bar{y}) \leq f(x), \quad \forall x \in S.$$

$\forall \lambda \in (0,1)$,

$$f\big((1-\lambda)\bar{x} + \lambda\bar{y}\big) < (1-\lambda)f(\bar{x}) + \lambda f(\bar{y}) = (1-\lambda)f(\bar{x}) + \lambda f(\bar{x}) = f(\bar{x}).$$

Chiamiamo $z : (1-\lambda)\bar{x} + \lambda\bar{y} \in S$ e $f(z) < f(\bar{x})$ contraddicendo la definizione di minimo globale. □

Come si verifica quindi che un punto sia effettivamente di minimo senza controllare ogni punto? Ogni definizione, infatti, è caratterizzata da una proprietà universale e richiederebbe il calcolo di $f(x)$ $\forall x$, e questo vale anche per la definizione di convessità.

1.5 Richiami di Analisi

Definizione 1.5.1 (Derivata Direzionale)**.** Si dice derivata direzionale di f in x lungo la direzione d, se finita, la quantità

$$D(x,d) = \lim_{t \to \infty} \frac{f(x+td) - f(x)}{t}. \tag{1.7}$$

Esempio 1.5.1. Se prendiamo $d = e_i$ vettore della base canonica di $\mathbb{R}^n$

$$e_i = \begin{cases} 1 & j = i \\ 0 & \text{altrimenti} \end{cases}$$

allora

$$D(x, e_i) = \frac{\partial f(x)}{\partial x_i}.$$

Definizione 1.5.2 (Gradiente)**.** Si dice *gradiente* di f in $x \in \mathbb{R}^n$ il vettore

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}. \tag{1.8}$$

Proposizione 1.5.1. *Se* $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ *esiste ed è continuo su* $\mathbb{R}^n$ $\forall x \in \mathbb{R}^n$, $\forall d \in \mathbb{R}^n$

$$D(x, d) = \nabla f(x)^T d.$$

Definizione 1.5.3 (Matrice Hessiana)**.** Si dice *matrice hessiana* di f in $x \in \mathbb{R}^n$ $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}$

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{(\partial x_1)^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{(\partial x_n)^2} \end{pmatrix}$$

Se $\nabla^2 f$ è continua in $\mathbb{R}^n$ si dice che f è due volte continuamente differenziabile.

Esempio 1.5.2.

- $f(x) = c^T x$, $c \in \mathbb{R}^n$, $\nabla f(x) = c$ e $\nabla^2 f(x) = 0 \in \mathbb{R}^{n \times n}$.
- $f(x) = \frac{1}{2} x^T Q x$ con $Q \in \mathbb{R}^{n \times n}$ simmetrica, $\nabla f(x) = Qx$ e $\nabla^2 f(x) = Q$. Infatti,

  $$\frac{\partial}{\partial x_i} \frac{1}{2} \sum_h \sum_j q_{hj} x_h x_j = \frac{1}{2} \frac{\partial}{\partial x_i} \Bigg(\sum_{\substack{j \neq h \neq i \neq j \\ j \neq i}} x_h x_j q_{hj} + \sum_{\substack{j=i \\ h \neq i}} x_h x_i q_{hi} + \sum_{\substack{j \neq i \\ h = i}} x_i x_j q_{ij} + \sum_i x_i^2 q_{ii} \Bigg) =$$

  dato che la prima non dipende da i e la seconda e la terza coincidono per la simmetria di Q,

  $$\frac{1}{2} \Bigg(0 + 2 \frac{\partial}{\partial x_i} \sum_{\substack{j=i \\ h \neq i}} x_h x_j q_{hj} + \frac{\partial}{\partial x_i} \sum_i x_i^2 q_{ii} \Bigg) = \frac{1}{2} \Bigg(2 \sum_{h \neq i} q_{hi} x_i + 2 \sum_i x_i q_{ii} \Bigg) = \sum_h q_{hi} x_h = q_i^T x.$$

  Dato che vale $\forall i$ $\nabla f(x) = Qx$.
- $f = ||x||_2^2 = x^T x = x^T I x$, $\nabla f(x) = 2x$, $\nabla^2 f(x) = 2I$ (è un caso particolare del precedente).
- $f(x) = \frac{1}{2} ||Ax - b||_2^2 = \frac{1}{2} (Ax - b)^T (Ax - b) = \frac{1}{2} x^T A^T A x - x^T A^T b + \frac{1}{2} b^t b$. $\nabla f(x) = A^T A x - A^T b$ e $\nabla^2 f(x) = A^T A$.

Definizione 1.5.4 (Jacobiano)**.** Data la funzione $F : \mathbb{R}^n \to \mathbb{R}^n$, $F(x) = (f_1(x), f_2(x), ..., f_n(x))^T$, definiamo lo Jacobiano di F la matrice

$$JF(x) = \begin{pmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_n(x)^T \end{pmatrix}$$

1.5.1 Convessità e Differenziabilità

Sia $f : \mathbb{R}^n \to \mathbb{R}$, è continuamente differenziabile due volte, allora

(i) f è convessa se e solo se $\forall \bar{x} \in \mathbb{R}^n$

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \quad \forall x \in \mathbb{R}^n.$$

Ossia, l'iperpiano tangente al grafico sottostima la funzione.

(ii) f è convessa se e solo se $\nabla^2 f(x)$ è semi-definita positiva $\forall x \in \mathbb{R}^n$.

(iii) Se $\nabla^2 f(x)$ è definita positiva $\forall x \in S$, allora f è strettamente convessa.

Esempio 1.5.3.

$$f(x) = \frac{1}{2} Qx, \qquad \nabla^2 f(x) = Q.$$

f quadratica è convessa se e solo se Q è semi-definita positiva. Se è definita positiva allora è strettamente convessa.

Esempio 1.5.4.

$$f(x) = ||Ax - b||^2, \qquad \nabla^2 f(x) = A^T A.$$

$\forall x \in \mathbb{R}$,

$$x^T A^T A x = (Ax)^T (Ax) = ||Ax||^2 \geq 0.$$

$A^T A$ è semi-definita positiva, quindi f è convessa. Se poi A è di rango massimo, $A^T A$ è definita positiva e f è strettamente convessa.

Definizione 1.5.5. Sia $f : \mathbb{R}^n \to \mathbb{R}$, diciamo che una direzione $d \in \mathbb{R}^n$ è *di discesa* in $\bar{x} \in \mathbb{R}^n$ se $\exists \bar{t} > 0$ tale che

$$f(\bar{x} + td) < f(\bar{x}), \qquad \forall t \in (0, \bar{t}].$$

Proposizione 1.5.2. *Se $\bar{x} \in \mathbb{R}^n$ è punto di minimo locale per f, allora non esistono direzioni di discesa in $\bar{x}$.*

Osservazione 1.5.1. Le condizioni date fino ad ora sono di tipo universale.

Osservazione 1.5.2. La condizione data non è sufficiente.

Esempio 1.5.5.

$$f(x) = \begin{cases} x \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}.$$

f è continua in $\mathbb{R}$ ma $x = 0$ non ha direzioni di discesa. Comunque prenda $\bar{t}$ ci sono delle oscillazioni. Tuttavia $\bar{x}$ non è un ottimo locale.

Proposizione 1.5.3. *Sia $f : \mathbb{R}^n \to \mathbb{R}$ continuamente differenziabile, sia $d \in \mathbb{R}^n$ e $\bar{x} \in \mathbb{R}^n$, se vale che $\nabla f(\bar{x})^T d < 0$ allora d è di discesa in $\bar{x}$. Se invece $\nabla f(\bar{x})^T d > 0$ è di salita.*

Osservazione 1.5.3. Se $\nabla f(\bar{x})^T d < 0$

$$\lim_{t \to o^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t} < 0,$$

per t sufficientemente piccolo,

$$f(\bar{x} + td) < f(\bar{x}).$$

Osservazione 1.5.4. Se $\nabla f(\bar{x})^T d = 0$ non si può dire nulla su d con solo queste ipotesi.

Osservazione 1.5.5. *Il gradiente è la direzione di massima salita.* Infatti,

$$\max_{||d||=1} \nabla f(x)^T d = \max_{||d||=1} ||\nabla f(x)||||d|| \cos\theta =$$

con θ angolo formato da d e ∇f,

$$= ||\nabla f(x)|| \max_{||d||=1} \cos\theta,$$

poiché $\nabla f(x)$ non dipende da d. Il massimo si ha quindi quando $d//\nabla f(x)$.

Allora $-\nabla f(x)$ (l'*antigradiente*) è la direzione di massima discesa.

Proposizione 1.5.4. *Se $\bar{x}$ è punto di minimo locale per $f : \mathbb{R}^n \to \mathbb{R}$ continuamente differenziabile, allora $\nabla f(\bar{x}) = 0$.*

Dimostrazione.
Assumiamo $\bar{x}$ minimo locale, allora non esiste una direzione di discesa in $\bar{x}$. Assumiamo per assurdo che $\nabla f(\bar{x}) \neq 0$. Preso $\bar{d} = -\nabla f(x)$,

$$\nabla f(\bar{x})^T \bar{d} = -||\nabla f(\bar{x})|| < 0.$$

$\bar{d}$ risulta essere una direzione di discesa per $\bar{x}$ e questo genera un assurdo. □

Definizione 1.5.6. Un punto $\bar{x}$ per cui vale

$$\nabla f(\bar{x}) = 0$$

si dice *punto stazionario.*

Proposizione 1.5.5. *Sia $f : \mathbb{R}^n \to \mathbb{R}$ convessa e differenziabile, allora d è di discesa in $\bar{x}$ se e solo se*

$$\nabla f(\bar{x})^T d < 0.$$

Dimostrazione.
Dimostriamo che se d è di discesa e f è convessa, allora $\nabla f(\bar{x})^T d < 0$. Per la definizione di convessità, con $x = \bar{x} + td$,

$$f(\bar{x} + td) \geq f(\bar{x}) + \nabla f(\bar{x})^T d.$$

d è di discesa, quindi, per $t > 0$ sufficientemente piccolo, vale che $f(\bar{x}) > f(\bar{x} + td)$,

$$f(\bar{x}) > f(\bar{x} + td) \geq f(\bar{x}) + t\nabla f(\bar{x})^T d.$$

Allora

$$0 > t\nabla f(\bar{x})^T d \;\Rightarrow\; \nabla f(\bar{x})^T d < 0.$$

□

Proposizione 1.5.6. *Sia $f : \mathbb{R}^n \to \mathbb{R}$ convessa e differenziabile, allora $x^* \in \mathbb{R}^n$ è punto di minimo globale per f se e solo se $\nabla f(x^*) = 0$.*

Dimostrazione.
Una direzione è ovvia.

Dato che f è convessa,

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \quad \forall x \in \mathbb{R}^n,$$

$$f(x) \geq f(\bar{x}), \quad \forall x \in \mathbb{R}^n.$$

Questa non è altro che la definizione di minimo globale. □

1.5.2 Richiami di Analisi

Teorema 1.5.1 (Della Media)**.** *Sia $f : \mathbb{R}^n \to \mathbb{R}$ continuamente differenziabile due volte,*

$$f(x+d) = f(x) + \nabla f(\xi)^T d \qquad \exists \xi = x + td,\ t \in (0,1). \tag{1.9}$$

Teorema 1.5.2 (Taylor)**.** *Sia $f : \mathbb{R}^n \to \mathbb{R}$ continuamente differenziabile due volte,*

$$f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(\xi) d, \qquad \xi = x + td,\ t \in (0,1). \tag{1.10}$$

Teorema 1.5.3.

$$f(x+d) = f(x) + \nabla f(x)^T d + \beta(x,d), \qquad \frac{\beta(x,d)}{||d||} \overset{||d||\to 0}{\longrightarrow} 0. \tag{1.11}$$

Teorema 1.5.4.

$$f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + \beta(x,d), \qquad \frac{\beta(x,d)}{||d||^2} \overset{||d||\to 0}{\longrightarrow} 0. \tag{1.12}$$

Definizione 1.5.7. Sia $f : \mathbb{R}^n \to \mathbb{R}$ due volte continuamente differenziabile, sia $\bar{x} \in \mathbb{R}^n$ e $d \in \mathbb{R}^n$, diciamo che d è una *direzione a curvatura negativa* se vale

$$d^T \nabla^2 f(x) d < 0.$$

Proposizione 1.5.7. *Sia $f : \mathbb{R}^n \to \mathbb{R}$ differenziabile e siano $\bar{x} \in \mathbb{R}^n$, $d \in \mathbb{R}^n$. Se $\nabla f(\bar{x})^T d = 0$ e $d^T \nabla^2 f(\bar{x}) d < 0$[3] allora d è di discesa in $\bar{x}$.*

Dimostrazione.
Consideriamo

$$f(\bar{x} + td) = f(\bar{x}) + t\nabla f(\bar{x})^T d + \frac{1}{2} t^2 d^T \nabla^2 f(\bar{x}) d + \beta(\bar{x}, d),$$

$$f(\bar{x} + td) - f(\bar{x}) = \frac{1}{2} t^2 d^T \nabla^2 f(\bar{x}) d + \beta(\bar{x}, d),$$

$$\frac{f(\bar{x} + td) - f(\bar{x})}{t^2} = \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + \frac{\beta(\bar{x}, d)}{t^2},$$

la quale tende a qualcosa minore di 0 poiché il primo fattore a destra è negativo per ipotesi e il secondo tende a 0 per $t \to 0$. Allora necessariamente

$$f(\bar{x} + td) - f(\bar{x}) < 0,$$

e d è di discesa. □

Proposizione 1.5.8 (Condizione Necessaria di Ottimalità del Secondo Ordine)**.** *Sia $\bar{x} \in \mathbb{R}^n$ punto di minimo locale per $f : \mathbb{R}^n \to \mathbb{R}$ due volte differenziabile con continuità, allora*

$$\nabla f(\bar{x}) = 0,$$

e $\nabla^2 f(\bar{x})$ è semi-definita positiva.

[3] Ossia è a curvatura negativa.

Dimostrazione.
La prima è data dalla condizione necessaria di ottimalità del primo ordine. Per quanto riguarda il secondo, assumiamo per assurdo che $\nabla^2 f(\bar{x})$ non sia semi-definita positiva, allora $\exists y \in \mathbb{R}^n$ tale che

$$y^T \nabla^2 f(\bar{x}) y < 0,$$

e y è una direzione a curvatura negativa in $\bar{x}$. Inoltre, per ipotesi, $\nabla f(\bar{x}) = 0$, quindi, per il Teorema precedente, y è di discesa in $\bar{x}$ e questo è assurdo poiché $\bar{x}$ è minimo locale. □

Proposizione 1.5.9 (Condizione Sufficiente per Ottimalità del Secondo Ordine)**.** *Sia $f : \mathbb{R}^n \to \mathbb{R}$ due volte differenziabile con continuità con $\nabla f(\bar{x}) = 0$ e $\nabla^2 f(\bar{x})$ definita positiva, allora $\bar{x}$ è un minimo locale. Se $\nabla f(\bar{x}) = 0$ e $\nabla^2 f(\bar{x})$ è semi-definita positiva $\forall x \in B(\bar{x}, \varepsilon)$, allora $\bar{x}$ è minimo locale.*

Proposizione 1.5.10 (Funzioni Quadratiche)**.** *Siano $Q \in \mathbb{R}^{n \times n}$ simmetrica e $b \in \mathbb{R}^n$*

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

allora

(i) f ammette minimo se e solo se $\exists \bar{x}$ tale che $Q\bar{x} = b$ e Q è semi-definita positiva;

(ii) Ogni $\bar{x}$ tale che $Q\bar{x} = b$ con Q semi-definita positiva è un minimo globale;

(iii) Il minimo globale x^* è unico se e solo se Q è definita positiva.

Dimostrazione.

(i) $\nabla f(x) = Qx - b$, $Q\bar{x} = b$ significa che $\nabla f(\bar{x}) = 0$ e $\nabla^2 f(x) \equiv Q$. Sia dunque $\bar{x}$ minimo globale, l'implicazione $\Rightarrow$ segue dalla condizione di ottimalità del secondo ordine. Per quanto riguarda l'implicazione $\Leftarrow$, essa si ha poiché, se $\exists \bar{x}$ tale che $Q\bar{x} = b$ con Q semi-definita positiva, allora f è convessa (quadratica con Q semi-definita positiva) e il punto $\bar{x}$ soddisfa la condizione necassaria e sufficiente di ottimalità nel caso convesso.

(ii) Segue dalla convessità di f e dalla condizione necessaria e sufficiente di ottimalità del primo ordine per funzioni convesse.

(iii) L'implicazione $\Leftarrow$ si ha poiché Q è definita positiva, allora f è strettamente convessa e i minimi globali, se esistono, sono unici. f è anche coerciva e tale minimo globale esiste.

Per la $\Rightarrow$, x^* esiste unico, supponiamo per assurdo che Q non sia definita positiva, poiché vale la (i) e Q è semi-definita positiva e non invertibile (ha un autovalore nullo). Sappiamo che $Q\bar{x} = b$ e quindi $Qx = b$ ammette almeno una soluzione. Poiché Q è singolare , ne esistono infinite ed esistono infiniti punti tali che $Qx = b$ ($\nabla f(x) = 0$) e questo è assurdo per l'unicità di x^* e per il (ii).

□

Capitolo 2

Algoritmi di Ottimizzazione Non Vincolata

2.1 Caratterizzazione Generale

Gli algoritmi di ottimizzazione non vincolata presentano uno schema generale. A partire da un punto $x^0 \in \mathbb{R}^n$, fino a che x^k non soddisfa la condizione $\nabla f(x^k) = 0$ si calcola uno spostamento s_k e si aggiorna il valore di x^k.

Algorithm 1 Schema Generale Non Vincolata

$x^0 \in \mathbb{R}^n$, $k = 0$
while $\nabla f(x^k) \neq 0$ **do**
  Calcolo uno spostamento s_k
  $x^{k+1} = x^k + s_k$
end while
return x^k

Osservazione 2.1.1. Si tratta di un algoritmo iterativo.

L'algoritmo produce una sequenza di soluzioni $\{x^k\} \subset \mathbb{R}^n$ a cui sono associate le sequenze $\{f(x^k)\} \subset \mathbb{R}$ e $\{\nabla f(x^k)\} \subset \mathbb{R}^n$. Possiamo avere due situazioni

- $\exists \bar{k}$ tale che $\nabla f(x^{\bar{k}}) = 0$ (*convergenza finita*);
- $\{x^k\}$ è una sequenza infinita (la situazione più frequente).

Idealmente vorremmo che la sequenza convergesse (anche rapidamente) verso un ottimo del problema. Più formalmente richiediamo

(i) l'esistenza di *punti di accumulazione* per $\{x^k\}$, ossia che $\exists k_1, k_2, ...$ tale che

$$\lim_{j \to \infty} x^{k_j} = \bar{x}.$$

In altre parole, che ammetta una sotto-successione convergente;

(ii) che $\{x^k\}$ convergesse asintoticamente verso la stazionarietà;

(iii) che convergesse velocemente.

2.1.1 Esistenza di Punti di Accumulazione

Proposizione 2.1.1. *Se f è continua, $x^0 \in \mathbb{R}^n$ tale che l'insieme*

$$L_f(x^0) = \{x \mid f(x) \leq f(x^0)\}^1$$

compatto e la sequenza $\{x^k\}$ tale che

$$f(x^{k+1}) \leq f(x^k)\ \forall k.$$

Allora

(i) la sequenza $\{x^k\}$ ammette punti di accumulazione;

(ii) la sequenza $\{f(x^k)\}$ converge ad un valore f^* finito.

Osservazione 2.1.2. L'ipotesi della compattezza di $L_f(x^0)$ sarebbe soddisfatta nel caso di funzioni obiettivo coercive. Inoltre è richiesto che f sia non crescente sulla successione x^k.

Dimostrazione.

(i) Per $k = 0$, $x^0 \in L_f(x^0)$ per definizione. Assumiamo come ipotesi induttiva che $x^k \in L_f(x^0)$, allora, per definizione di $L_f(x^0)$,

$$f(x^k) \leq f(x^0).$$

Per ipotesi

$$f(x^{k+1}) \leq f(x^k) \leq f(x^0),$$

pertanto $x^{k+1} \in L_f(x^0)$ e $\{x^k\} \subset L_f(x^0)$ compatto. Dunque la sequenza $\{x^k\}$ ammette una sotto-successione convergente e quindi un punto di accumulazione: $\exists\{x^{k_j}\} \subset \{x^k\}$ tale che

$$\lim_{j\to\infty} x^{k_j} = \bar{x}.$$

(ii) Consideriamo $\{f(x^k)\}$ successione monotona non crescente per ipotesi.

$$f(x^k) \to f^*.$$

Per la compattezza di $L_f(x^0)$ e poiché $\{x^k\} \subseteq L_f(x^0)$ con f continua, il limite f^* è finito per il Teorema di Weierstrass.

□

2.1.2 Stazionarietà

Ci sono tre tipi di convergenza a stazionarietà.

- $$\lim_{k\to\infty} x^k = \bar{x}, \qquad \nabla f(\bar{x}) = 0.$$

  La successione converge ad un punto stazionario.

[1] La notazione richiederebbe $L_f(f(x^0))$.

- $$\lim_{k\to\infty} ||\nabla f(x^k)|| = 0.$$

  Per la continuità del gradiente, tutti i punti di accumulazione sono stazionari. Se esiste una sotto-successione convergente $x^{k_j} \to \bar{x}$,

  $$\nabla f(x^{k_j}) \overset{j\to\infty}{\to} \nabla f(\bar{x}),$$

  ma $||\nabla f(x^k)|| \overset{k\to\infty}{\to} 0$ allora

  $$\lim_{j\to\infty} ||\nabla f(x^{k_j})|| = ||\nabla f(\bar{x})|| = \lim_{k\to\infty} ||\nabla f(x^k)|| = 0,$$

  allora ogni punto di accumulazione è stazionario.

- $$\liminf_{k\to\infty} ||\nabla f(x^k)|| = 0.$$

  Ossia esiste un punto di accumulazione stazionario.

  Esempio 2.1.1.

  $$\nabla f(x) = (x-1)(x-2).$$

  $\{x^k\} = 1, 1, 1, 1, ...,\ x^k \to 1$ e converge ad un punto che è stazionario.

- $\{x^k\} = 1, 2, 1, 2, 1, 2..., ||\nabla f(x^k)|| = 0\ \forall k.$
- $\{x^k\} = 1, -1, 1, -1..., \{\nabla f(x^k)\} = 0, 6, 0, 6, ...$ e

  $$\liminf_{k\to\infty} ||\nabla f(x^k)|| = 0.$$

  Il punto di accumulazione $x = 1$ è stazionario.

Osservazione 2.1.3. Delle condizioni precedenti, la prima implica la seconda che implica la terza. La prima è la più forte e la terza è la più debole.

Osservazione 2.1.4. Per la continuità, dalla terza condizione, $\forall \varepsilon > 0\ \exists \bar{k} > c$ tale che

$$||\nabla f(x^{\bar{k}})|| \leq \varepsilon.$$

Esempio 2.1.2. Se f è strettamente convessa e $f(x^{k+1}) < f(x^k)$, le condizioni non garantiscono la stazionarietà. Infatti

$$\min_{x\in\mathbb{R}} \frac{1}{2}x^2,$$

ammette un ottimo globale $x^* = 0$ con $f(x^*) = 0$. Definiamo x^0 tale che $||x^0|| > 1$ e

$$x^{k+1} = x^k - \alpha_k f'(x^k) = x^k - \alpha_k x^k = x^k(1-\alpha_k), \qquad \alpha_k = 2 - \frac{\varepsilon_k}{||x^k||}.$$

Allora

$$x^{k+1} = x^k\big(1 - 2 + \frac{\varepsilon_k}{||x^k||}\big) = x^k\left(-1 + \frac{\varepsilon}{||x^k||}\right) = -x^k + \frac{x^k}{||x^k||}\varepsilon_k =$$

$$= -x^k + sign(x^k)\varepsilon_k = sign(x^k)\big(-\frac{x^k}{sign(x^k)} + \varepsilon_k\big) = sign(x^k)(-|x^k| + \varepsilon_k).$$

Ricapitolando

$$x^{k+1} = sign(x^k)(-|x^k| + \varepsilon_k). \tag{2.1}$$

Supponiamo $|x_k| > 1$,

$$|x^{k+1}| = |sign(x^k)(-|x^k| + \varepsilon_k)| = \big|-|x^k| + \varepsilon_k\big|.$$

Se $0 < \varepsilon_k < |x^k| - 1$, allora il membro in valore assoluto è negativo, pertanto

$$\left|-|x^k| + \varepsilon_k\right| = |x^k| - \varepsilon_k > |x^k| - |x^k| + 1 = 1.$$

allora $|x^{k+1}| > 1$. Pertanto $|x^k| > 1\ \forall k$ e

$$f(x^k) = \frac{1}{2}(x^k)^2, \qquad f(x^{k+1}) = \frac{1}{2}\big(-|x^k| + \varepsilon_k\big)^2 = \frac{1}{2}\big(|x^k| - \varepsilon_k\big)^2 < \frac{1}{2}(x^k)^2 = f(x^k).$$

$f(x^k)$ è strettamente decrescente. Qualsiasi punto limite $\bar{x}$ è però tale che $|\bar{x}| \geq 1$ poiché $|x^k| > 1\ \forall k$. $f(x^k)$ non converge a 0 e la sequenza $\nabla f(x^k)$ non converge a 0.

2.1.3 Velocità di Convergenza

Definizione 2.1.1. Supponiamo che $x^k \to \bar{x}$, diciamo che x^k ha un *tasso di convergenza*

- *sublineare* se
$$\lim_{k\to\infty} \frac{||x^{k+1} - \bar{x}||}{||x^k - \bar{x}||} = 1.$$
- *lineare* se
$$\lim_{k\to\infty} \frac{||x^{k+1} - \bar{x}||}{||x^k - \bar{x}||} = c \in (0, 1).$$
- *superlineare* se
$$\lim_{k\to\infty} \frac{||x^{k+1} - \bar{x}||}{||x^k - \bar{x}||} = 0.$$
- *quadratico* se
$$\lim_{k\to\infty} \frac{||x^{k+1} - \bar{x}||}{||x^k - \bar{x}||^2} = c \in \mathbb{R}.$$

2.2 Classificazione di Algoritmi di Ottimizzazione

La prima divisione riguarda la disponibilità o meno di $f, \nabla f$ e $\nabla^2 f$.

- Algoritmi di primo ordine: se in ogni momento è possibile calcolare $(f, \nabla f)$ in x;
- Algoritmi di secondo ordine: se in ogni momento è possibile calcolare $(f, \nabla f, \nabla^2 f)$ in x;
- Algoritmi di ordine zero: se si ha a disposizione solo f. Sono anche detti derivative free o ottimizzazione black box.[2]

Gli algoritmi convergenti si dividono in

[2] Non sono proprio lo stesso concetto.

- Algoritmi *globalmente* convergenti: $\forall x^0 \in \mathbb{R}$, $\{x^k\}$ converge a $\bar{x}$ con $\nabla f(\bar{x}) = 0$;
- Algoritmi *localmente* convergenti: $x\{^k\}$ converge a $\bar{x}$ con $\nabla f(\bar{x}) = 0$ $\forall x^0$ scelto in un intorno ooportuno di $\bar{x}$.

Inoltre abbiamo tre tipologie di algorimto:

- Metodi di tipo Linesearch;
- Metodi di tipo Trust Region (regione di confidenza);
- Metodi di tipo Direct Search.

Tutti questi tipi sono detti algoritmi di discesa.

Esempio 2.2.1 (Regressione Regolarizzata o di Ridge)**.** Si ha un dataset

$$D = \left\{ (x^i, y^i) \mid x^i \in \mathbb{R}^p, y^i \in \mathbb{R},\ i = 1, ..., N \right\}.$$

Cerchiamo una funzione lineare (un iperpiano) che minimizzi l'errore quadratico medio (residuo quadratico medio)

$$\min_w \frac{1}{2} \sum_{i=1}^{N} (y_i - w^T x^i)^2 = \min_w \frac{1}{2} ||Xw - Y||^2, \tag{2.2}$$

dove $r^i = y_i - w^T x^i$ è detto residuo. L'Equazione (2.2) definisce la regressione lineare ai minimi quadrati. Definita $\mathcal{L}(w, X, Y) := \frac{1}{2}||Xw - Y||^2$ la Loss function, aggiungiamo un termine di penalizzazione

$$\min_w \mathcal{L}(w, X, Y) + \lambda ||w||^2 = \min_w \frac{1}{2} ||Xw - Y||^2 + \lambda ||w||^2. \tag{2.3}$$

L'Equazione (2.3) definisce la Regressione di Ridge o Regressione regolarizzata.

$$\min_{w \in \mathbb{R}^p} \frac{1}{2} ||Xw - y||^2 + \frac{\lambda}{2} ||w||^2 = \frac{1}{2} w^T x^T x w - \frac{2}{2} y^T x w + \frac{1}{2} y^T y + \frac{\lambda}{2} w^T w.$$

Osservazione 2.2.1. $f(w) \geq \frac{\lambda}{2}||w||^2$ che tende a ∞ per $k \to \infty$.

Indaghiamo esistenza e unicità della soluzione.

$$\nabla f(w) = x^T x w - y^T x + \lambda w.$$

$$\nabla^2 f(w) = x^T x - \lambda I.$$

Dato che $\nabla^2 f(w)$ è definita positiva, la funzione è strettamente convessa.

Sia $u \in \mathbb{R}^p$,

$$u^T (x^T x + \lambda I) u = u^T x^T x u + \lambda u^T I u = ||xw||^2 + \lambda ||u||^2 \geq \lambda ||u||^2 > 0.$$

Ora, sia w^* tale che

$$X^T X w^* - y^T x + \lambda w^* = 0,$$
$$(x^T x + \lambda I) w^* = y^T x. \tag{2.4}$$

L'Equazione (2.4) è detto sistema delle Equazioni Normali. Abbiamo due modi per risolverle, l'inversione della matrice $x^T x + \lambda I$ (non singolare) o un algortimo iterativo.

Algorithm 2 Metodo Linesearch

$x^0 \in \mathbb{R}^n$ $k = 0$
while $\nabla f(x^k) \neq 0$ **do**
  scelgo una direzione $d_k \in \mathbb{R}^n$
  scelgo un passo $\alpha_k > 0$
  $x^{k+1} = x^k + \alpha_k d_k$
  $k = k + 1$
end while

2.3 Metodi di Tipo Linesearch

Lo schema generale di un metodo linesearch è quello dell'Algoritmo **??**

Definizione 2.3.1. Sia $\sigma : \mathbb{R}^+ \to \mathbb{R}^+$, diciamo che σ è una *funzione di forzamento* se $\forall \{t_k\} \subseteq \mathbb{R}^+$ tale che

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

allora vale

$$\lim_{k\to\infty} t_k = 0,$$

Esempio 2.3.1. $\sigma(t) = ct$ con $c > 0$ è tale che, se

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

allora $t_k \to 0$ per $k \to +\infty$.

Proposizione 2.3.1. *Sia f continuamente differenziabile, $\mathcal{L}_f(x^0)$ compatto, $\{x^k\}$ una sequenza tale che $\nabla f(x^k) \neq 0$ $\forall k$ e assumiamo che*

- $$f(x^{k+1}) \leq f(x^k) \tag{2.5}$$
- $$\lim_{k\to\infty} \frac{\nabla f(x^k)^T d_k}{||d_k||} = 0 \tag{2.6}$$
- $$\frac{\nabla f(x^k)^T d_k}{||d_k||} \geq \sigma(||\nabla f(x^k)||), \tag{2.7}$$

  con σ di forzamento.

Allora

(i) $\{x^k\} \subseteq \mathcal{L}_f(x^k)$;

(ii) $\{x^k\}$ *ha punti di accumulazione, ognuno dei quali in* $\mathcal{L}_f(x^0)$;

(iii) $\{f(x^k)\}$ *converge;*

(iv)

$$\lim_{k\to\infty} ||\nabla f(x^k)|| = 0;$$

(v) ogni punto di accumulazione di $\{x^k\}$ è stazionario.

Osservazione 2.3.1. La (2.6) forza la scelta del passo α_k, la (2.7) è detta *condizione d'angolo* e dipende dalla scelta di d_k.

Dimostrazione.
I punti (i), (ii) e (iii) seguono direttamente dalla continuità di f, la compattezza di $\{\mathcal{L}_f(x^0)\}$ e dalla monotonia di $\{f(x^k)\}$.

(iv) Dalle ipotesi, sappiamo che

$$\frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq \sigma(||\nabla f(x^k)||) \geq 0.$$

Se passiamo al limite per $k \to \infty$ si ha che il membro di sinistra va a 0. Allora

$$\lim_{k\to\infty} \sigma(||\nabla f(x^k)||) = 0,$$

allora

$$\lim_{k\to\infty} ||\nabla f(x^k)|| = 0.$$

(v) Sia $K \subseteq \{0, 1, ...\}$ sotto-sequenza tale che

$$\lim_{\substack{k\to\infty \\ k\in K}} x^k = \bar{x}.$$

allora

$$\lim_{\substack{k\to\infty \\ k\in K}} ||\nabla f(x^k)|| = ||\nabla f(\bar{x})||$$

per la continuità del gradiente e della norma. Per il punto (iv) il membro di sinistra va a 0, allora $\bar{x}$ è stazionario.

□

Osservazione 2.3.2. La condizione (ii) può essere riscritta come

$$\nabla f(x^k)^T \frac{d_k}{||d_k||} \to 0,$$

e la normalizzazione rende tale limite indipendente dal modulo di d_k. Pertanto deve essere la derivata direzionale ad amdare a 0 e non la lunghezza della direzione.

Inoltre, non deve verificarsi che asintoticamente d_k sia perpendicolare al gradiente.

Sia $\alpha(t) = ct$, $c > 0$. Dalla (2.7),

$$\frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq c||\nabla f(x^k)||.$$

Se $\nabla f(x^k)^T d_k < 0$, allora

$$\frac{\nabla f(x^k)^T d_k}{||d_k|| \cdot ||\nabla f(x^k)||} \leq -c < 0,$$

e, come abbiamo visto in precedenza,

$$\frac{\nabla f(x^k)^T d_k}{||d_k|| \cdot ||\nabla f(x^k)||} = \cos\big(\theta(d_k, \nabla f(x^k))\big).$$

Se quindi prendessimo come direzione l'antigradiente?, Scelto $d_k = -\nabla f(x^k)$,

$$\frac{\nabla f(x^k)^T(-\nabla f(x^k))}{||\nabla f(x^k)|| \cdot || - \nabla f(x^k)||} = -1.$$

Allora scegliere come direzione l'antigradiente soddisfa la condizione di angolo (2.7).

Se H_k simmetrica definita positiva con $\lambda_{min}(H_k) \geq m \ \forall k$ e $\lambda_{max}(H_k) \leq M$, possiamo prendere come direzione di discesa $d_k = -H_k \nabla f(x^k)$. Questo perché

$$\nabla f(x^k)^T d_k = -\nabla f(x^k)^T H_k \nabla f(x^k) < 0,$$

se $\nabla f(x^k) \neq 0$.

$$|\nabla f(x^k)^T d_k| = | - \nabla f(x^k)^T H_k \nabla f(x^k)| \geq m||\nabla f(x^k)||^2,$$
$$||d_k|| = ||H_k \nabla f(x^k)|| \leq M||\nabla f(x^k)||,$$

allora

$$\frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq \frac{m}{M}||\nabla f(x^k)||.$$

Scegliendo $c = \frac{m}{M}$, la scelta di d_k soddisfa la condizione d'angolo.

2.4 Line Search

Scelto α_k opportuno, considerando $x^{k+1} = x^k + \alpha_k d_k$, cerchiamo di minimizzare

$$f(x^k + \alpha_k d_k).$$

Possiamo avere due situazioni

$$\alpha_k = \arg\min_{\alpha} f(x^k + \alpha d_k)$$

e si parla di ricerche di linea *esatte*, oppure

$$\alpha_k \approx \arg\min_{\alpha} f(x^k + \alpha d_k)$$

e si tratta di ricerche di linea *inesatte*.

2.4.1 Ricerche di Linea Esatte

Consideriamo

$$\varphi(\alpha) = f(x^k + \alpha d_k) \qquad \varphi'(\alpha) = \nabla f(x^k + \alpha d_k)^T d_k$$
$$\varphi(0) = f(x^k) \qquad \varphi'(0) = \nabla f(x^k)^T d_k.$$

Come già anticipato, la ricerca di linea esatta consiste nel cercare

$$\alpha_k = \arg\min_{\alpha} \varphi(\alpha),$$

ossia cercare tra gli α per cui $\varphi'(\alpha) = 0$.

Esempio 2.4.1.

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

Usiamo Taylor,

$$\varphi(\alpha) = f(x^k + \alpha d_k = f(x^k) + \alpha \nabla f(x^k)^T d_k + \frac{1}{2}\alpha^2 d_k^T \nabla^2 f(x^k) d_k.$$

Risolvendo

$$\varphi'(\alpha) = \nabla f(x^k)^T d_k + \alpha d_k^T Q d_k = 0,$$

otteniamo

$$\alpha^* = -\frac{\nabla f(x^k)^T d_k}{d_k^T Q d_k}.$$

Per una f quadratica è possibile eseguire una ricerca di linea esatta perché siamo in grado di calcolare il passo ottimo.

2.4.2 Ricerche di Linea di Tipo Inesatto

Quello che vorremmo è ancora

$$\alpha_k \approx \arg\min_{\alpha} f(x^k + \alpha d_k),$$

ma nella pratica cerchiamo di scegliere α_k per imporre un sufficiente decremento di f.

2.4.3 Condizione di Armijo

Dati d_k tale che $\nabla f(x^k)^T d_k < 0$ e $\gamma \in (0,1)$,

$$f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k. \tag{2.8}$$

2.4.4 Metodo di Armijo

Siano $\gamma \in (0,1)$, $\delta \in (0,1)$, $\alpha_0 > 0$

$$\alpha_k = \min_{j>0}\{\delta^j \alpha_0 \mid f(x^k + \alpha_0 \delta^j d_k) \leq f(x^k) + \gamma \alpha_0 \delta^j \nabla f(x^k)^T d_k\}.$$

Algorithm 3 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 = α δ
end while
return α
```

Teorema 2.4.1 (Terminazione Finita del Metodo di Armijo)**.** *Sia d_k tale che $\nabla f(x^k)^T d_k < 0$, allora il metodo di Armijo termina in un numero finito di passi restituendo un valore α_k tale che*

$$f(x^{k+1}) = f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^t d_k.$$

inoltre, vale una delle seguenti condizioni

(i) $\alpha_k = \alpha_0$;

(ii) $\alpha_k \leq \delta\alpha_0$ *e*

$$f\left(x^k + \frac{\alpha_k}{\delta}d_k\right) > f(x^k) + \gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k.$$

Dimostrazione.
Assumiamo per assurdo che l'algoritmo non termini, allora $\forall j = 0, 1, ...$

$$f(x^k + \delta^j\alpha_0 d_k) > f(x^k) + \gamma\alpha_0\delta^j\nabla f(x^k)^T d_k \qquad \forall j$$

$$\frac{f(x^k + \delta^j\alpha_0 d_k) - f(x^k)}{\alpha_0\delta^j} > +\gamma\nabla f(x^k)^T d_k \qquad \forall j$$

Ora, il membro di sinistra, per $j \to \infty$, tende a $\nabla f(x^k)^T d_k$, allora

$$\nabla f(x^k)^T d_k \geq \gamma\nabla f(x^k)^T d_k,$$

$$(1-\gamma)\nabla f(x^k)^T d_k \geq 0,$$

e questo genera un assurdo. □

Per come è costruito l'algoritmo, possiamo avere due scelte,

(i) $\alpha_k = \alpha_0$;

(ii) $\alpha_k < \alpha_0$, poiché abbiamo ridotto il passo almeno una volta. In particolare, $\alpha_k = \alpha_0\delta$. Inoltre, alla penultima iterazione (l'ultimo passo in cui può avvenire la riduzione), la condizione di Armijo non deve essere soddisfatta. Allora abbiamo

$$f\left(x^k + \frac{\alpha_k}{\delta}\right) > f(x^k) + \gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k.$$

Proposizione 2.4.1 (Proprietà di Convergenza dei Metodi Linesearch con ricerca di Armijo)**.** *Sia $f \in C^1(\mathbb{R}^n)$, $x^0 \in \mathbb{R}^n$ tale che $\mathcal{L}_f(x^0)$ è compatto e sia $\{x^k\}$ la sequenza prodotta dall'algoritmo, tale che*

$$\nabla f(x^k)^T d_k < 0 \ \forall k. \tag{2.9}$$

Se

$$\alpha_0(k) \geq \frac{1}{||d_k||}\sigma\left(\frac{|\nabla f(x^k)^T d_k|}{||d_k||}\right),$$

per qualche σ funzione di forzamento. Allora

(i)

$$f(x^{k+1}) < f(x^k).$$

(ii)

$$\lim_{k\to\infty}\frac{|\nabla f(x^k)^T d_k|}{||d_k||} = 0. \tag{2.10}$$

Osservazione 2.4.1. La (2.9) definisce d_k come direzione di discesa e sottintende una convergenza infinita (non si realizza mai $\nabla f(x^k)^T d_k = 0$).

Dimostrazione.

(i) Per Armijo,

$$f(x^{k+1}) \leq f(x^k) + \gamma\alpha_k \nabla f(x^k)^T d_k,$$

con $\gamma > 0$, $\alpha_k > 0$ e $\nabla f(x^k)^T d_k < 0$, dunque

$$f(x^{k+1}) \leq f(x^k) + \gamma\alpha_k \nabla f(x^k)^T d_k < f(x^k).$$

(ii) Consideriamo

$$f(x^k) - f(x^{k+1}) \geq -\gamma\alpha_k \nabla f(x^k)^T d_k,$$

moltiplico e divido a destra per $||d_k||$,

$$f(x^k) - f(x^{k+1}) \geq -\gamma||d_k||\alpha_k \frac{\nabla f(x^k)^T d_k}{||d_k||} = \gamma\alpha_k||d_k|| \frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq 0.$$

dato che $-\nabla f(x^k)^T d_k = |\nabla f(x^k)^T d_k|$. Per il punto (i) sappiamo che $f(x^k)$ è monotona decrescente e $\{x^k\} \subseteq \mathcal{L}_f(x^0)$ compatto. Allora $\{f(x^k)\}$ ammette limite ed esso è finito[3].

$$f(x^k) \to \bar{f} \in \mathbb{R}.$$

$$f(x^k) - f(x^{k+1}) \geq \gamma\alpha_k||d_k|| \frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq 0,$$

passando al limite per $k \to \infty$

$$0 = \bar{f} - \bar{f} \geq \lim_{k\to\infty} \gamma\alpha_k||d_k|| \frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq 0. \tag{2.11}$$

Per il Teorema dei Carabinieri, il limite nella (2.11) fa 0.

Consideriamo ora la sequenza

$$\left\{ \frac{|\nabla f(x^k)^T d_k|}{||d_k||} \right\}. \tag{2.12}$$

$\{x^k\} \subseteq \mathcal{L}_f(x^0)$ compatto, dunque è limitata. $\nabla f(x)$ è continuo, allora $\{\nabla f(x^k)\}$ è limitata. $\frac{d_k}{||d_k||}$ è un vettore di norma unitaria $\forall k$ e $\{\frac{d_k}{||d_k||}\}$ è limitata. Allora la successione in (2.12) è limitata e ammette punti di accumulazione.

Assumiamo che la (ii) della Proposizione 2.4.1. Allora esiste una sotto-successione K tale che

$$\lim_{\substack{k\to\infty \\ k\in K}} \frac{\nabla f(x^k)^T d_k}{||d_k||} = -\mu < 0. \tag{2.13}$$

Infatti, nel caso ci fossero punti di accumulazione diversi da 0, essi dovrebbero essere negativi.

Ora, $\{x^k\}$ è limitata e $\{\frac{d_k}{||d_k||}\}$ è limitata, quindi esiste una sottosuccessione $K_1 \subseteq K$ tale che

$$\lim_{\substack{k\to\infty \\ k\in K_1}} x^k = \bar{x}, \qquad \lim_{\substack{k\to\infty \\ k\in K_1}} \frac{d_k}{||d_k||} = \bar{d}.$$

Dunque

$$\lim_{\substack{k\to\infty \\ k\in K_1}} \frac{\nabla f(x^k)^T d_k}{||d_k||} = \nabla f(\bar{x})^T \bar{d} = -\mu < 0.$$

[3] f è continua, $L_f(x^0)$ è compatto e possiamo applicare Weierstrass.

Dato che il limite della (2.11) tende a 0, anche la sotto-successione deve tendere a 0 e, di conseguenza,

$$\lim_{k\to\infty} \alpha_k ||d_k|| = 0.$$

Possiamo avere quindi due situazioni

(a) $\exists \bar{k}$ tale che $\forall k > \bar{k}$ $\alpha_k = \alpha_0(k)$.

(b) $\exists K_2 \subseteq K_1$ tale che $\alpha_k \leq \delta\alpha_0(k)$ e $k \in K_1$

$$f(x^k + \alpha_k/\delta d_k) > f(x^k) + \gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k.$$

(a) $\alpha_k/\delta = \alpha_0(k)$. Abbiamo richiesto che

$$\alpha_0(k) \geq \frac{1}{||\alpha_k||}\sigma\left(\frac{|\nabla f(x^k)^T d_k|}{||d_k||}\right),$$

$$\alpha_k||d_k|| \geq \sigma\left(\frac{|\nabla f(x^k)^T d_k|}{||d_k||}\right) \geq 0. \tag{2.14}$$

Nella (2.14) il membro di sinistra va a 0, pertanto

$$\sigma\left(\frac{|\nabla f(x^k)^T d_k|}{||d_k||}\right) \overset{k\to 0}{\longrightarrow} 0,$$

e, per le proprietà di σ funzione forzamento, deve tendere a 0 l'argomento. Allora

$$\lim_{\substack{k\to\infty \\ k\in K_1}} \frac{|\nabla f(x^k)^T d_k|}{||d_k||} = 0.$$

Questo contraddice l'ipotesi per cui tale limite fa $-\mu < 0$. Questo caso non si può avere.

(b)

$$f(x^k + \alpha_k/\delta d_k) - f(x^k) > \gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k. \tag{2.15}$$

Sfruttando il Teorema della Media,

$$f(x^k + \alpha_k/\delta d_k) - f(x^k) = \frac{\alpha_k}{\delta}\nabla f(\xi^k)^T d_k, \qquad \xi^k = x^k + t^*\frac{\alpha_k}{\delta}d_k, \quad t^* \in (0,1).$$

$$\frac{\alpha_k}{\delta}\nabla f(\xi^k)^T d_k > \gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k,$$

$$\frac{\alpha_k}{\delta}\nabla f(\xi^k)^T \frac{d_k}{||d_k||} > \gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k \frac{1}{||d_k||}.$$

Ora, $x^k \to \bar{x}$ in K_1, dunque anche in K_2. $\xi^k = x^k + \alpha_k/\delta t^* d_k$. $x^k \to \bar{x}$ e $t^*\alpha_k/\delta d_k \to 0$,[4] allora $\xi^k \to \bar{x}$ per $k \in K_2$ e tendente a infinito.

$$\lim_{\substack{k\to\infty \\ k\in K_2}} \nabla f(\xi^k)^T \frac{d_k}{||d_k||} \geq \lim_{\substack{k\to\infty \\ k\in K_2}} \gamma\nabla f(x^k)^T \frac{d_k}{||d_k||},$$

$$\nabla f(\bar{x})^T \bar{d} \geq \gamma\nabla f(\bar{x})^T \bar{d}.$$

$$(1-\gamma)\nabla f(\bar{x})^T \bar{d} \geq 0.$$

$1 - \gamma > 0$ allora $\nabla f(\bar{x})^T \bar{d} \geq 0$ ma $\bar{d}$ deve essere di discesa in $\bar{x}$. Abbiamo raggiunto un assurdo anche in questo caso.

[4] Rivedi.

La tesi risulta dunque dimostrata.

□

Proposizione 2.4.2 (Caso Quadratico)**.** *Sia* $f(x) = \frac{1}{2}x^TQx + c^Tx$ *con* Q *definita positiva. Sia* $x^k \in \mathbb{R}^n$ *e* d_k *tale che* $\nabla f(x^k)^T d_k < 0$*. Sia* α^* *il passo ottimale definito come segue,*

$$\alpha^* = \frac{-\nabla f(x^k)^T d_k}{d_k^T Q d_k},$$

esso soddisfa la condizione di Armijo

$$f(x^k + \alpha^* d_k) \leq f(x^*) + \gamma \alpha^* \nabla f(x^k)^T d_k,$$

se e solo se $\gamma \in (0, 1/2)$*.*

Dimostrazione.
Usiamo Taylor

$$f(x^k + \alpha^* d_k) = f(x^k) + \alpha^* \nabla f(x^k)^T d_k + \frac{1}{2}\alpha^{*2} d_k^T Q d_k = f(x^k) + \alpha^*(\nabla f(x^k)^T d_k + \frac{1}{2}\alpha^* d_k Q d_k) =$$

$$= f(x^k) + \alpha^* \left(\nabla f(x^k)^T d_k + \frac{1}{2} \frac{-\nabla f(x^k)^T d_k d_k^T Q d_k}{d_k^T Q d_k} \right) =$$

$$= f(x^k) + \frac{1}{2}\alpha^*(\nabla f(x^k)^T d_k).$$

Ora,

$$f(x^k) + \gamma \alpha^* \nabla f(x^k)^T d_k \geq f(x^k + \alpha^* d_k) = f(x^k) + \frac{1}{2}\alpha^* \nabla f(x^k)^T d_k.$$

$$\frac{1}{2}\alpha^* \nabla f(x^k)^T d_k \leq \gamma \alpha^* \nabla f(x^k)^T d_k.$$

Necessariamente $\gamma \leq \frac{1}{2}$ poiché $\nabla f(x^k)^T d_k < 0$.

□

2.5 Metodo del Gradiente

Algorithm 4 Metodo del Gradiente

Dato $x^0 \in \mathbb{R}^n$ e $k = 0$
while $\nabla f(x^k) \neq 0$ **do**
  scelgo $d_k = -\nabla f(x^k)$
  calcolo α_k con Armijo
  $x^{k+1} = x^k + \alpha_k d_k$
  $k = k + 1$
end while

2.5.1 Convergenza del Metodo del Gradiente

Proposizione 2.5.1 (Convergenza del Metodo del Gradiente)**.** *Sia x^0 tale che $\mathcal{L}_f(x^0)$ sia compatto, $f \in C^1(\mathbb{R}^n)$, allora la sequenza prodotta dall'algoritmo ammette punti di accumulazione e ogni punto di accumulazione è stazionario.*

Dimostrazione.
Valgono tutte le condizioni generali per la convergenza di metodi linesearch.

(i) $f(x^{k+1}) \leq f(x^k)$;

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

(iii)
$$\frac{|\nabla f(x^k)^T d_k|}{||d_k||} \geq \sigma\bigg(||\nabla f(x^k)||\bigg).$$

Le prime due condizioni derivano da Armijo. La terza dalla scelta dell'antigradiente, infatti

$$\frac{||\nabla f(x^k)||^2}{||\nabla f(x^k)||} = ||\nabla f(x^k)|| \geq c \cdot ||\nabla f(x^k)||,$$

con $c < 1$ e questa è di forzamento[5]. □

Proposizione 2.5.2. *Il tasso di convergenza del metodo del gradiente è sublineare.*

Definizione 2.5.1 (Fortemente Convesssa)**.** Una funzione $f : \mathbb{R}^n \to \mathbb{R}$ si dice *fortemente convessa* se esiste $\mu > 0$ tale che $f(x) - \mu||x||^2$ è una funzione convessa. Equivalentemente,

$$f(x) = g(x) + \mu||x||^2, \tag{2.16}$$

con g convessa.

Proposizione 2.5.3. *Se f è fortemente convessa, allora il metodo del gradiente converge in modo lineare.*

Tornando alla forma di x^{k+1} nel metodo del gradiente,

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k).$$

Cosa ottengo se scelgo $\alpha_k = \alpha \forall k$ (costante)? In generale non funziona il metodo poiché tende a mancare la soluzione quando ci si avvicina all'ottimo. Facendo però ulteriori assunzioni vediamo che le cose migliorano.
Assumiamo che

- $f \in C^2(\mathbb{R}^n)$;
- ∇f sia lipshitz-continuo con costante di Lipshitz L;

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \quad \forall x, y \in \mathbb{R}^n. \tag{2.17}$$

[5] Come visto nell'Esempio 2.3.1

Proposizione 2.5.4. *$f \in C^2(\mathbb{R}^n)$ con ∇f lipshitz-continuo se e solo se*

$$|u^T \nabla^2 f(y) u| \leq L||u||^2, \quad \forall u, y \in \mathbb{R}^n. \tag{2.18}$$

Sotto le ipotesi fatte,

$$f(x+td) \overset{Taylor}{=} f(x) + t\nabla f(x)^T d + \frac{1}{2}t^2 d^T \nabla^2 f(y) d, \quad \exists y \in \mathbb{R}^n$$

$$\leq f(x) + t\nabla f(x)^T d + \frac{1}{2}t^2 |d^T \nabla^2 f(x) d| \leq f(x) + t\nabla f(x)^T d + \frac{1}{2}L||d||^2.$$

Pertanto

$$f(x+td) \leq f(x) + t\nabla f(x)^T d + \frac{1}{2}L||d||^2. \tag{2.19}$$

Proposizione 2.5.5. *Sia $f \in C^2(\mathbb{R}^n)$, ∇f Lipshitz-continuo di costante L. Sia $x^0 \in \mathbb{R}^n$ tale che $\mathcal{L}_f(x^0)$ è compatto e $\{x^k\}$ ottenuta con la formula di aggiornamento del metodo del gradiente a passo costante $\alpha_k = 1/L$,*

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k). \tag{2.20}$$

Allora

(i) $\{f(x^k)\}$ è decrescente e vale che

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}||\nabla f(x^k)||^2;$$

(ii) $\{x^k\}$ ha punti di accumulazione, ciascuno dei quali è stazionario.

Dimostrazione.
$\alpha = \frac{1}{L}$ e $d_k = -\nabla f(x^k)$. Applichiamo la (2.19),

$$f(x^{k+1}) = f(x^k + \alpha d_k) \leq f(x^k) + \alpha \nabla f(x^k)^T d_k + \frac{1}{2}\alpha^2 L||d_k||^2 =$$

$$= f(x^k) - \frac{1}{L}||\nabla f(x^k)||^2 + \frac{1}{2L}||\nabla f(x^k)||^2 = f(x^k) - \frac{1}{2L}||\nabla f(x^k)||^2.$$

Questo dimostra il punto (i). Ora, $\{f(x^k)\}$ è decrescente, quindi $\{x^k\} \subseteq \mathcal{L}_f(x^o)$ compatto e $\{x^k\}$ ammette punti di accumulazione[6].

$$\lim_{k\to\infty} f(x^k) = f^* \in \mathbb{R}.$$

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L}||\nabla f(x^k)||^2,$$

Passando al limite

$$\lim_{k\to\infty} f(x^{k+1}) - f(x^k) \leq \lim_{k\to\infty} -\frac{1}{2L}||\nabla f(x^k)||^2,$$

$$0 = f^* - f^* \leq \lim_{k\to\infty} -\frac{1}{2L}||\nabla f(x^k)||^2 \leq 0.$$

Allora

$$\lim_{k\to\infty} -\frac{1}{2L}||\nabla f(x^k)||^2 = 0 \;\Rightarrow\; \lim_{k\to\infty} ||\nabla f(x^k)||^2 = 0.$$

pertanto i punti di accumulazione di $\{x^k\}$ sono stazionari e questo completa la dimostrazione. □

Osservazione 2.5.1. Conoscere L non è scontato, a volte è impossibile, a volte è costoso. Il risultato precedente non è utilizzabile direttamente nella pratica ma fornisce una giustificazione alla scelta di un passo costante se scelto opportunamente.

[6] Ancora poiché f continua, $\mathcal{L}_f(x^0)$ compatto e f decrescente, possiamo applicare Weierstrass.

2.5.2 Complessità Computazionale degli Algoritmi di Ottimizzazione

Risulta evidente che la definizione di tasso di convergenza può essere riformulata su f invece che su x^k, ottenendo che, a seconda del risultato del limite

$$\lim_{k\to\infty} \frac{f(x^{k+1}) - f^*}{f(x^k) - f^*},$$

si ha che, se il limite fa

- 1 il tasso è sub-lineare;
- $\rho \in (0,1)$ il tasso è lineare;
- 0 il tasso è super-lineare.

Osservazione 2.5.2. Nella pratica gli algoritmi non girano all'infinito ma tipicamente terminano quando

$$f(x^k) - f^* < \varepsilon.$$

- Quante iterazioni occorrono per ottenere un'accuratezza ε?
- Che accuratezza ottengo con k iterazioni?

Definizione 2.5.2. Se f è convessa,

- Diciamo che un algoritmo ha *iteration error* $O(h(k))$ se $f(x^k) - f^* = O(h(k))$.
- Diciamo che un algoritmo ha *iteration complexity* $O(\hat{h}(\varepsilon))$ se

$$\min\{k \mid f(x^k) - f^* < \varepsilon\} = O(h(\varepsilon)).$$

Qual è il legame tra loro?

- Se ho $f(x^k) - f^* \approx \frac{1}{k}$, allora

$$f(x^k) - f^* < \varepsilon \;\Rightarrow\; k > \frac{1}{\varepsilon} \;\Rightarrow\; k = O\left(\frac{1}{\varepsilon}\right).$$

- Se ho $f(x^k) - f^* \approx \frac{1}{k^2}$, allora

$$f(x^k) - f^* < \varepsilon \;\Rightarrow\; k > \frac{1}{\sqrt{\varepsilon}} \;\Rightarrow\; k = O\left(\frac{1}{\sqrt{\varepsilon}}\right).$$

Osservazione 2.5.3. Se cerco un'accuratezza ε, il numero di cifre significative che ottengo è $\log(1/\varepsilon)$. La seguente tabella mostra il numero di iterazioni necessarie per ottenere un certo numero di cifre significative

Cifre $\log(1/\varepsilon)$	ε	Costo $1/\varepsilon$
1	0.1	10
2	0.01	100
3	0.001	1000.

Il costo di ogni cifra è esponenziale nel numero di iterazioni.

Ora, se l'algoritmo ha $\varepsilon = O(1/k)$,

$$\lim_{k\to\infty} \frac{f(x^{k+1} - f^+)}{f(x^k) - f^*} \approx \lim_{k\to\infty} O\left(\frac{k}{k+1}\right) = 1.$$

Si ottiene che è sub-lineare. Analogamente, con $\rho \in (0,1)$,

iteration error	iteration complexity	tasso
$O\left(\frac{1}{k}\right)$	$O\left(\frac{1}{\varepsilon}\right)$	sub-lineare
$O(\rho^k)$	$O\left(\log\left(\frac{1}{k}\right)\right)$	lineare
$O(\rho^{2^k})$	$O\left(\log\left(\log\left(\frac{1}{k}\right)\right)\right)$	super-lineare

Nel caso della terza riga della tabella, trovare una cifra significativa aggiuntiva costa meno di quelle precedenti.

Proposizione 2.5.6. *Sia $f \in C^2(\mathbb{R}^n)$ convessa, ∇f lipshitz-continuo di costante L e $x^0 \in \mathbb{R}^n$ tale che $\mathcal{L}_f(x^0)$ compatto,*

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k),$$

$x^k \to \bar{x}$, $f(x^k) \to f^*$, allora

$$f(x^k) - f^* \leq L\frac{||x^0 - \bar{x}||^2}{k},$$

ossia la complessità del metodo del gradiente a passo costante è $O(1/k)$.

Dimostrazione.

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L}||\nabla f(x^k)||^2$$

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}||\nabla f(x^k)||^2 \tag{2.21}$$

Per la convessità

$$f(\bar{x}) \geq f(x^k) + \nabla f(x^k)^T(\bar{x} - x^k).$$

$$f(x^k) \leq f(\bar{x}) - \nabla f(x^k)^T(\bar{x} - x^k) = f(\bar{x})\nabla f(x^k)^T(x^k - \bar{x}). \tag{2.22}$$

Dalla (2.21) e dalla (2.22),

$$f(x^{k+1}) \leq f(\bar{x}) + \nabla f(x^k)^T(x^k - \bar{x}) - \frac{1}{2L}||\nabla f(x^k)||^2.$$

$$f(x^{k+1}) - f(\bar{x}) \leq \nabla f(x^k)^T(x^k - \bar{x}) - \frac{1}{2L}||\nabla f(x^k)||^2.$$

Raccogliamo $L/2$

$$f(x^{k+1}) - f(\bar{x}) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x^k)^T(x^k - \bar{x}) - \frac{1}{L^2}||\nabla f(x^k)||^2\right).$$

$$f(x^{k+1}) - f(\bar{x}) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x^k)^T(x^k - \bar{x}) - \frac{1}{L^2}||\nabla f(x^k)||^2 + ||x^k - \bar{x}||^2 - ||x^k - \bar{x}||^2\right). \tag{2.23}$$

Ora, posto $a := x^k - \bar{x}$ e $b := \frac{1}{L}\nabla f(x^k)$,

$$||a - b||^2 = ||a||^2 + ||b||^2 - 2a^T b,$$

$$-||a-b||^2 = -||a||^2 - ||b||^2 + 2a^T b.$$

Allora, la (2.23) diventa

$$= \frac{L}{2}(2a^T b - ||b||^2 + ||x^k - \bar{x}|| - ||a||^2) =$$

$$= \frac{L}{2}(-||a+b||^2 + ||x^k - \bar{x}||^2),$$

$$f(x^{k+1}) - f(\bar{x}) \leq \frac{L}{2}(-||x^k - \bar{x} - \frac{1}{L}\nabla f(x^k)||^2 + ||x^k - \bar{x}||^2).$$

Ricordando che $x^k - \frac{1}{L}\nabla f(x^k) = x^{k+1}$,

$$f(x^{k+1}) - f(\bar{x}) \leq \frac{L}{2}(+||x^k - \bar{x}||^2 - ||x^{k+1} - \bar{x}||^2).$$

Passando alle sommatorie da entrambi i membri,

$$\sum_{t=0}^{k}\big[f(x^{t+1} - f(\bar{x}))\big] \leq \frac{L}{2}\sum_{t=0}^{k}(||x^t - \bar{x}||^2 - ||x^{t+1} - \bar{x}||^2).$$

la somma a destra è telescopica, pertanto

$$\sum_{t=0}^{k}\big[f(x^{t+1} - f(\bar{x}))\big] \leq \frac{L}{2}(||x^0 - \bar{x}||^2 - ||x^{k+1} - \bar{x}||^2) \leq \frac{L}{2}||x^0 - \bar{x}||^2,$$

poiché $-||x^{k+1} - \bar{x}||^2 \leq 0$. Ora, poiché $f(x^k)$ è decrescente, $f(x^{t+1}) \geq f(x^{k+1})\ \forall t < k$. Allora

$$\sum_{t=0}^{k}\big[f(x^{t+1} - f(\bar{x}))\big] \geq \sum_{t=0}^{k}\big[f(x^{k+1}) - f(\bar{x})\big] = (k+1)[f(x^{k+1}) - f(\bar{x})].$$

Riassumendo

$$(k+1)[f(x^{k+1}) - f(\bar{x})] \leq \sum_{t=0}^{k}\big[f(x^{t+1} - f(\bar{x}))\big] \leq \frac{L}{2}||x^0 - \bar{x}||^2.$$

Ossia,

$$f(x^{k+1}) - f^* \leq \frac{L||x^0 - \bar{x}||^2}{2(k+1)}.$$

□

Proposizione 2.5.7. *Se f è fortemente convessa, allora l'algoritmo ha complessità $O(\log(1/\varepsilon))$ e quindi tasso di convergenza lineare.*

Osservazione 2.5.4. Abbiamo visto che spesso, in problemi di machine learning, si ha $f(x) = L(x) + \lambda||x||^2$. Dal punto di vista dell'ottimizzazione, il termine $\lambda||x||^2$ rende f coerciva e garantisce dunque l'esistenza della soluzione. Inoltre, se la Loss function è convessa, f è fortemente convessa e il metodo del gradiente diventa a tasso lineare.

2.6 Varianti del Metodo del Gradiente

2.6.1 metodo Heavy-Ball o Momentum

La formula di aggiornamento diventa

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1}).$$

Il metodo si muove di una combinazione tra l'antigradiente e la direzione dell'ultimo spostamento effettuato. Possiamo riformularla come segue

$$\begin{cases} y^k = x^k - \alpha_k \nabla f(x^k) \\ x^{k+1} = y^k + \beta_k (x^k - x^{k-1}) \end{cases}, \qquad \begin{cases} v^k = x^k - x^{k-1} \\ x^{k+1} = x^k + v^k - \alpha_k \nabla f(x^k) \end{cases}.$$

Il metodo del gradiente dipende fortemente dalla scelta del punto iniziale. Il termine $\beta_k(x^k - x^{k-1})$ tende ad evitare le oscillazioni dovute ad un cattivo punto di partenza.

Pros Con f fortemente convessa, la complessità è migliore. Nel caso di f quadratica, α_k e β_k possono essere scelti in modo ottimale in forma chiusa.

Cons Nel caso generale non sappiamo come scegliere β_k. Esistono esempi in cui il metodo non converge anche sotto forti ipotesi di regolarità.

2.6.2 Metodo del Gradiente Accelerato o di Nesterov

La formula di aggiornamento diventa

$$x^{k+1} = x^k + \beta_k (x^k - x^{k-1}) - \alpha_k \nabla f(x^k + \beta_k (x^k - x^{k-1})).$$

$$\begin{cases} v^k = \beta_k (x^k - x^{k-1}) \\ x^{k+1} = x^k + v^k - \alpha_k \nabla f(x^k + v^k) \end{cases} \qquad \begin{cases} y^k = x^k - \alpha_k \nabla f(x^k + \beta_k (x^k - x^{k-1})) \\ x^{k+1} = y^k + \beta_k (x^k - x^{k-1}). \end{cases}$$

Il senso goemetrico di tale scelta resta abbastanza oscuro. Di contro l'algoritmo, con β_k scelto seguendo uno schema opportuno e prefissato, ha complessità $O(1/k^2)$, ossia $O(\sqrt{\varepsilon})$ nel caso di f convessa. Questo è ottimale per metodi del primo ordine, ossia non si può costruire un metodo con complessità migliore con solo il gradiente a disposizione.

2.7 Metodo delle Direzioni Coniugate

Consideriamo il caso dell'ottimizzazione di funzioni quadratiche

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x - c^T x,$$

con $Q \in \mathbb{R}^{n \times n}$ simmetrica e definita positiva.

Definizione 2.7.1. Data Q simmetrica e definita positiva, diciamo che $d_0, d_1, ..., d_{m-1}$ m, si dicono direzioni *mutuamente coniugate* rispetto a Q se

$$d_i^T Q d_j = 0, \qquad \forall i, j = 0, ..., m-1, i \neq j.$$

Proposizione 2.7.1. *Se Q è simmetrica e definita positiva e $d_0, d_1, ..., d_{n-1}$ sono direzioni mutuamente coniugate rispetto a Q, allora sono anche linearmente indipendenti.*

Dimostrazione.
Consideriamo

$$\sum_{i=0}^{n-1} \alpha_i d_i = 0.$$

Moltiplichiamo per Q

$$\sum_{i=0}^{n-1} \alpha_i Q d_i = 0,$$

e ora per d_j^T $j \in \{0, ..., n-1\}$,

$$d_j^T \sum_{i=0}^{n-1} \alpha_i Q d_i = 0,$$

$$\sum_{i=0}^{n-1} \alpha_i d_j^T Q d_i = \alpha_j d_j^T Q d_j.$$

Dato che $d_j^T Q d_j > 0$ poiché Q è definita positiva, $\alpha_j = 0$ e questo può essere ripetuto $\forall j = 0, ..., n-1$. □

Ora, in un problema di ottimizzazione, raggiunto l'ottimo, si ha

$$0 = \nabla f(x^*) = Qx^* - c.$$

Osservazione 2.7.1. Se $d_0, ..., d_{n-1}$ sono direzioni mutuamente coniugate, sono anche una base, allora

$$x^* = \sum_{i=0}^{n-1} \alpha_i d_i,$$

$$Qx^* = \sum_{i=0}^{n-1} \alpha_i Q d_i,$$

$$d_j^T Q x^* = \sum_{i=0}^{n-1} \alpha_i d_j^T Q d_i = \alpha_j d_j^T Q d_j,$$

$$\alpha_j = \frac{d_j^T Q x^*}{d_j^T Q d_j} = \frac{d_j^T c}{d_j^T Q d_j}.$$

$$x^* = \sum_{i=0}^{n-1} \frac{d_i^T c}{d_i^T Q d_i} d_i.$$

Possiamo trovare x^* senza passare dall'inversione della matrice Q.

Ora, sia $x^0 \in \mathbb{R}^n$

$$x^* - x^0 = \sum_{i=0}^{n-1} \alpha_i d_i.$$

Allora, visto come processo iterativo,

$$x^* = x^0 + \alpha_0 d_0 + \alpha_1 d_1 + ... + \alpha_{n-1} d_{n-1},$$

il sui passo k-esimo è

$$x^k = x^0 + \alpha_0 d_0 + ... + \alpha_{k-1} d_{k-1}.$$

Facendo come in precedenza,

$$Q(x^* - x^0) = \sum_{i=0}^{n-1} \alpha_i Q d_i,$$

e

$$Q(x^k - x^0) = \sum_{i=0}^{k-1} \alpha_i Q d_i,$$

Poi

$$d_k^T Q(x^* - x^0) = \sum_{i=0}^{n-1} \alpha_i d_k^T Q d_i = \alpha_k d_k^T Q d_k, \tag{2.24}$$

e

$$d_k^T Q(x^k - x^0) = \sum_{i=0}^{k-1} \alpha_i d_k^T Q d_i = 0. \tag{2.25}$$

Allora, dalla (2.24),

$$\alpha_k = \frac{d_k^T Q(x^* - x^0)}{d_k^T Q d_k} = \frac{d_k^T Q(x^* - x^k + x^k - x^0)}{d_k^T Q d_k} =$$

$$\frac{d_k^T Q(x^* - x^k)}{d_k^T Q d_k} + \frac{d_k^T Q(x^k - x^0)}{d_k^T Q d_k} =$$

il secondo fattore è nullo per la (2.25),

$$\frac{d_k^T Q(x^* - x^k)}{d_k^T Q d_k} = \frac{d_k^T (Qx^* - Qx^k)}{d_k^T Q d_k} = \frac{d_k^T (c - Qx^k)}{d_k^T Q d_k} = \frac{-\nabla f(x^k)^T d_k}{d_k^T Q d_k} = \alpha_k.$$

Questa è la forma del passo α_k ottimo nella ricerca di linea con funzioni quadratiche.

Proposizione 2.7.2. *Sia Q simmetrica e definita positiva, sia $x^0 \in \mathbb{R}^n$, $d_0, ..., d_{n-1}$ direzioni mutuamente coniugate,*

$$x^{k+1} = x^k + \alpha_k d_k,$$
$$\alpha_k = \frac{-\nabla f(x^k)^T d_k}{d_k^T Q d_k},$$

Allora

(i) se chiamiamo $g_k := \nabla f(x^k)$,

$$g_k^T d_i = 0, \quad \forall i = 0, ..., k-1.$$

(ii) $\exists \bar{k} \leq n$ tale che

$$x^{\bar{k}} = x^*.$$

Osservazione 2.7.2. La (ii) dice che si raggiunge l'ottimo in un numero finito di passi, si ha quindi la proprietà di convergenza finita.

Dimostrazione.

(i) Sia $i \leq k$,

$$x^k = x^i + \sum_{j=i}^{k-1} \alpha_j d_j,$$

$$Qx^k = Qx^i + \sum_{j=i}^{k-1} \alpha_j Q d_j,$$

$$Qx^k - c = Qx^i - c + \sum_{j=i}^{k-1} \alpha_j Q d_j.$$

Ora, il membro di sinistra è g_k. Vediamo $g_k^T d_i$

$$g_k^T d_i = d_i^T (Qx^i - c) + \sum_{j=i}^{k-1} \alpha_j d_i^T Q d_j = d_i^T g_i + \alpha_i d_i^T Q d_i.$$

Dunque,

$$g_k^T d_i = d_i^T g_i + \frac{-g_i^T d_i}{d_i^T Q d_i} d_i^T Q d_i = d_i^T g_i - g_i^T d_i = 0,$$

dove abbiamo usato più volte che $d_i^T g_i \in \mathbb{R}$ ed è dunque uguale al suo trasposto.

(ii) Dato che, per la (i), $g_n^T d_j = 0 \ \forall j < n$, allora $g_n \perp d_i \ \forall i = 0, ..., n-1$, le quali costituiscono una base. Pertanto $g_n = 0$.

$$Qx^n - c = 0 \ \Rightarrow \ x^n = x^*.$$

□

Come ottengo le direzioni mutuamente coniugate.
Esiste un metodo iterativo che permette di trovarle coi passi

$$\alpha_k = \frac{-g_k^T d_k}{d_k^T Q d_k}, \qquad \beta_{k+1} := \frac{||g_{k+1}||^2}{||g_k||^2}.$$

Ora,

$$d_{k+1} = -g_{k+1} + \beta_{k+1} d_k. \tag{2.26}$$

Con questo si ottiene il *Metodo del Gradiente Coniugato*. Si può dimostrare che, se $d_0 = -\nabla f(x^0)$, $d_0, ..., d_k$ sono mutuamente coniugate.

2.8 Metodo del Gradiente Coniugato

Osservazione 2.8.1.

$$-g_k^T d_k = -g_k^T(-g_k + \beta_k d_{k-1}) = ||g_k||^2 - \beta_k g_k^T d_{k-1} = ||g_k||^2, \tag{2.27}$$

poiché $g_k^T d_{k-1} = 0$ per la Proposizione precedente.

La g_{k+1} si ha poiché

$$g_{k+1} = Qx^{k+1} - c = Q(x^k + \alpha_k d_k) - c = Qx^k - c + \alpha_k Q d_k = g_k + \alpha_k Q d_k. \tag{2.28}$$

Se n è grande, può essere un processo costoso ma spesso ci possiamo accontentare di qualcosa di meno.

Algorithm 5 Metodo del Gradiente Coniugato

$x^0 \in \mathbb{R}^n$ $d_0 = -g_0$ $k = 0$
while $||g_k|| \neq 0$ **do**
  calcola $\alpha_k = \frac{||g_k||^2}{d_k^T Q d_k}$
  $x^{k+1} = x^k + \alpha_k d_k$
  $g_{k+1} = g_k + \alpha_k Q d_k$
  $\beta_{k+1} = \frac{||g_{k+1}||^2}{||g_k||^2}$
  $d_{k+1} = -g_{k+1} + \beta_{k+1} d_k$
  $k = k + 1$
end while

Osservazione 2.8.2. Si tratta del metodo preferito in caso di funzioni quadratiche. Si usa spesso per la risuoluzione di sistemi lineari con Q simmetrica e definita positiva.

Cosa succede se applichiamo il metodo nel caso generale? Possiamo estenderlo ma occorrono alcune modifiche. La condizione del *while* va modificata perché non arriviamo ad una convergenza finita

$$||g_k|| \leq \varepsilon.$$

Il gradiente va calcolato in maniera diversa. α_k è scelto con una linesearch. Per quanto riguarda β_k, le formule non sono tutte equivalenti.

$$\beta_{k+1} = \frac{||\nabla f(x^{k+1})||^2}{||\nabla f(x^k)||^2}.$$

Resta da scegliere la linesearch per la ricerca di α_k.

Fino ad ora abbiamo usato

$$d_{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} d_k = -\nabla f(x^k + \alpha_k d_k) + \frac{||\nabla f(x^{k+1})||^2}{||\nabla f(x^k)||^2}.$$

Quello che vogliamo è ottenere

$$0 > d_{k+1}^t g_{k+1} = -||g_{k+1}||^2 + \beta_k d_k^T g_{k+1}.$$

In questo caso la condizione di Armijo non basta.

2.8.1 Ricerca di Linea di Tipo Wolfe

- Condizione di Wolfe debole

$$f(x^{k+1}) = f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k. \tag{2.29}$$
$$\nabla f(x^k + \alpha_k d_k)^T d_k \geq \sigma \nabla f(x^k)^T d_k, \qquad \gamma \in (0, 1/2),\ \sigma \in (\gamma, 1).$$

- Condizione di Wolfe forte

$$f(x^{k+1}) = f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k. \tag{2.30}$$
$$|\nabla f(x^k + \alpha_k d_k)^T d_k| \geq \sigma |\nabla f(x^k)^T d_k|, \qquad \gamma \in (0, 1/2),\ \sigma \in (\gamma, 1).$$

Osservazione 2.8.3. Mentre è relativamente facile verificare la condizione di Armijo, per verificare le condizioni di Wolfe è necessario calcolare ripetutamente il gradiente.

2.9 Metodo di Newton

I metodi visti fino ad ora sono metodi del primo ordine. Vediamo ora uno del secondo, ossia in cui, dato x^*, sono disponibili $f(x^k)$, $\nabla f(x^k)$ e $\nabla^2 f(x^k)$.

Partiamo col costruire un'approssimante con lo sviluppo di Taylor

$$m_k(x) = f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k).$$

Se $\nabla^2 f(x^k)$ è definita positiva, allora $m_k(x)$, in quanto funzione quafratica con Q definita positiva, è strettamente comnvessa. Questa ha quindi minimo nel punto $\hat{x}$ in cui $\nabla m_k(\hat{x}) = 0$.

$$\nabla m_k(\hat{x}) = \nabla f(x^k) + \nabla^2 f(x^k)(\hat{x} - x^k) = 0,$$

$$\nabla^2 f(x^k)(\hat{x} - x^k) = -\nabla f(x^k).$$

$$\hat{x} - x^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k),$$

$$\hat{x} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k). \tag{2.31}$$

dalla (2.31) si ottiene la formula di aggiornamento del metodo di Newton

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k), \tag{2.32}$$

la quale può essere vista come

$$x^{k+1} = x^k + \alpha_k d_k, \qquad \alpha_k = 1, \quad d_k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$

Esempio 2.9.1.

$$f(x) = \sqrt{1 + x^2}$$

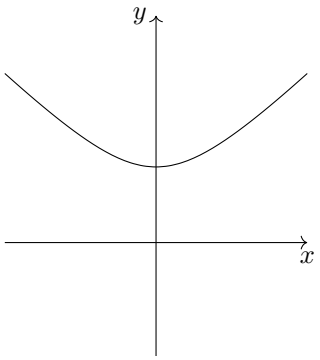

Figura 2.1: Figura dell'Esempio 2.9.1.

$$\nabla f(x) = f'(x) = \frac{x}{\sqrt{1 + x^2}}, \qquad \nabla^2 f(x) = f''(x) = \frac{1}{(1 + x^2)^{3/2}}.$$

Nel caso di funzioni ad una variabile, il metodo diventa

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}$$

$$x^{k+1} = x^k - \frac{x^k}{\sqrt{1+(x^k)^2}}(1+\big(x^k)^2\big)^{\frac{3}{2}} = x^k - x^k\big(1+(x^k)^2\big) = x^k\big(-(x^k)^2\big) = -(x^k)^3.$$

Se $x^0 = 1$, $x^1 = -1$, $x^2 = 1$, e cosìì via. Se $x^0 = -1$ la situazione è simile e se partiamo con $x^0 = 2$, $x^1 = 8$, $x^2 = 512$ mentre è evidente che il minimo della funzione sia in $x = 0$.

Come si può osservare dall'Esempio 2.9.1, il Metodo di Newton non ha la proprietà di convergenza globale.

Osservazione 2.9.1. Notiamo che la funzione obiettivo f non viene mai controllata. Occorre solo il caloclo di gradiente e hessiano.

Osservazione 2.9.2.

$$d_k = -(\nabla^2 f\big(x^k)\big)^{-1}\nabla f(x^k)$$

Nella pratica non si ottiene invertendo l'hessiano ma risolvendo il sistema

$$\nabla^2 f(x^k)d_k = \nabla f(x^k),$$

per questioni di efficienza e di robustezza.

Proposizione 2.9.1. *Sia $f \in C^2(\mathbb{R}^n)$, $x^* \in \mathbb{R}^n$ tale che $\nabla f(x^*) = 0$ e sia $\nabla^2 f(x^*)$ non-singolare. Allora $\exists \varepsilon > 0$ tale che $\forall x^0 \in B(x^*, \varepsilon)$, la sequenza $\{x^k\}$ ottenuta con*

$$x^{k+1} = x^k - (\nabla^2 f\big(x^k)\big)^{-1}\nabla f(x^k)$$

ha le seguenti proprietà

(i) $\{x^k\} \subset B(x^*, \varepsilon)$;

(ii) $x^k \to x^*$;

(iii) la convergenza è super-lineare

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||} = 0.$$

Inoltre, se $\nabla^2 f(x)$ è Lipshitz-continua, cioè

$$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L||x - y||, \quad \forall x, y \in \mathbb{R}^n$$

allora

(iv) la convergenza è quadratica

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} = \rho > 0.$$

Dimostrazione.
Richiamiamo il Teorema della media integrale: sia $F : \mathbb{R}^n \to \mathbb{R}^n$ con $JF : \mathbb{R}^n \to \mathbb{R}^{n\times n}$,

$$F(x) = F(y) + \int_0^1 JF\big(y + t(x-y)\big)(x-y)dt. \tag{2.33}$$

Ora, poiché $\nabla^2 f(x^*)$ è non-singolare e $\nabla^2 f(x)$ è continua, $\exists \varepsilon_1 > 0$ tale che

$$||(\nabla^2 f\big(x^k\big))^{-1}|| \leq \mu, \quad \forall x \in B(x^*, \varepsilon_1).$$

Sempre per la continuità di $\nabla^2 f(x)$, $\exists \varepsilon > 0,\ \varepsilon \leq \varepsilon_1$ tale che $\exists \sigma \in (0,1)$ per cui

$$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq \frac{\sigma}{\mu}, \quad \forall x, y \in B(x^*, \varepsilon). \tag{2.34}$$

Supponiamo che $x^k \in B(x^*, \varepsilon)$, vediamo $x^{k+1} - x^*$,[7]

$$x^{k+1} - x^* \overset{(2.32)}{=} -(\nabla^2 f\big(x^k\big))^{-1}\nabla f(x^k) + (x^k - x^*),$$

Moltiplichiamo per $(\nabla^2 f\big(x^k\big))^{-1}\nabla^2 f(x^k)$ l'ultimo fattore,[8]

$$x^{k+1} - x^* = -(\nabla^2 f\big(x^k\big))^{-1}\nabla f(x^k) + (\nabla^2 f\big(x^k\big))^{-1}\nabla^2 f(x^k)(x^k - x^*).$$

$$x^{k+1} - x^* = (\nabla^2 f\big(x^k\big))^{-1}\big[\nabla f(x^k) - \nabla^2 f(x^k)(x^k - x^*)\big],$$

dato che $\nabla f(x^*) = 0$ per ipotesi, possiamo aggiungere un fattore $-\nabla f(x^*)$ dentro le quadre,

$$x^{k+1} - x^* = -(\nabla^2 f\big(x^k\big))^{-1}\big[\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\big].$$

Passo alle norme

$$||x^{k+1} - x^*|| = \big|\big|(\nabla^2 f\big(x^k\big))^{-1}\big[\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\big]\big|\big| \leq,$$

$$\leq \big|\big|(\nabla^2 f\big(x^k\big))^{-1}\big|\big| \cdot \big|\big|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\big|\big|,$$

dato che $x^k \in B(x^*, \varepsilon) \subset B(x^*, \varepsilon_1)$,

$$x^{k+1} - x^* \leq \mu\big|\big|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\big|\big|.$$

Applicando la (2.33) a $\nabla f(x)$,

$$\nabla f(x^k) - \nabla f(x^*) = \int_0^1 \nabla^2 f\big(x^* + t(x^k - x^*)\big)(x^k - x^*)dt,$$

pertanto

$$x^{k+1} - x^* \leq \mu\left|\left|\int_0^1 \nabla^2 f\big(x^* + t(x^k - x^*)\big)(x^k - x^*)dt - \nabla^2 f(x^k)(x^k - x^*)\right|\right| \leq$$

$$\leq \mu\left|\left|\int_0^1 \Big(\big[\nabla^2 f\big(x^* + t(x^k - x^*)\big) - \nabla^2 f(x^k)\big](x^k - x^*)\Big)dt\right|\right| \leq$$

[7] Se tale distanza fosse minore di ε possiamo usare l'induzione sapendo che $x^0 \in B(x^*, \varepsilon)$.
[8] $\nabla^2 f(x^k)$ è ben definita poiché $x^k \in B(x^*, \varepsilon)$ e $\nabla^2 f(x^*)$ è non-singolare.

$$\leq \mu \int_0^1 \Big|\Big| \Big(\big[\nabla^2 f\big(x^* + t(x^k - x^*)\big) - \nabla^2 f(x^k)\big](x^k - x^*) \Big) \Big|\Big| dt \leq$$

$$\leq \mu \int_0^1 \Big|\Big| \nabla^2 f\big(x^* + t(x^k - x^*)\big) - \nabla^2 f(x^k) \Big|\Big| \cdot \big|\big|(x^k - x^*)\big|\big| dt.$$

Dato che $x^*, x^k \in B(x^*, \varepsilon)$ convesso, $x^* + t(x^k - x^*) \in B(x^*, \varepsilon)\ \forall t \in (0,1)$, allora, per la (2.34),

$$||x^{k+1} - x^*|| \leq \mu \int_0^1 \frac{\sigma}{\mu} ||x^k - x^*|| dt = \sigma ||x^k - x^*||.$$

Dato che $\sigma \in (0,1)$, se $x^k \in B(x^*, \varepsilon)$, anche $x^{k+1} \in B(x^*, \varepsilon)$ poiché la distanza di x^{k+1} da x^* è minore di ε. Se dunque $x^0 \in B(x^*, \varepsilon)$, $\{x^k\} \subset B(x^*, \varepsilon)$. Si ha dunque la (i).

Ripartendo da

$$||x^{k+1} - x^*|| \leq \sigma ||x^k - x^*|| \leq \sigma^2 ||x^{k-1} - x^*|| \leq ... \leq \sigma^k ||x^0 - x^*||,$$

allora, dato che $\sigma^k \to 0$ per $k \to \infty$ e $||x^0 - x^*||$ è una costante,

$$\lim_{k \to \infty} ||x^{k+1} - x^*|| \leq \lim_{k \to \infty} \sigma^k ||x^0 - x^*|| = 0,$$

e si ha la (ii). Ora, da

$$||x^{k+1} - x^*|| \leq \mu \int_0^1 \Big|\Big| \nabla^2 f\big(x^* + t(x^k - x^*)\big) - \nabla^2 f(x^k) \Big|\Big| \cdot \big|\big|(x^k - x^*)\big|\big| dt.$$

$$0 \leq \frac{||x^{k+1} - x^*||}{||x^k - x^*||} \leq \mu \int_0^1 \Big|\Big| \nabla^2 f\big(x^* + t(x^k - x^*)\big) - \nabla^2 f(x^k) \Big|\Big| dt$$

Passando al limite per $k \to \infty$, si ha $x^k \to x^*$ e

$$0 \leq \lim_{k \to \infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||} \leq \mu \int_0^1 \Big|\Big| \nabla^2 f(x^*) - \nabla^2 f(x^*) \Big|\Big| dt = 0.$$

Allora x^k converge super-linearmente.

Se poi $\nabla^2 f(x)$ è Lipshitz-continua,

$$||x^{k+1} - x^*|| \leq \mu \int_0^1 \Big|\Big| \nabla^2 f\big(x^* + t(x^k - x^*)\big) - \nabla^2 f(x^k) \Big|\Big| \cdot \big|\big|(x^k - x^*)\big|\big| dt \leq$$

$$\leq \mu \int_0^1 L \big|\big|\big(x^* + t(x^k - x^+) - x^k\big)\big|\big| \cdot \big|\big|(x^k - x^*)\big|\big| dt = \mu \int_0^1 L \big|\big|\big(1-t)(x^* - x^k\big)\big|\big| \cdot \big|\big|(x^k - x^*)\big|\big| dt =$$

$$= L\mu \int_0^1 \big|\big|(x^k - x^*)\big|\big|^2 (1-t) dt = L\mu \big|\big|(x^k - x^*)\big|\big|^2 \int_0^1 (1-t) dt = \frac{1}{2} \mu L \big|\big|(x^k - x^*)\big|\big|^2.$$

Pertanto

$$\frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} \leq \frac{\mu L}{2},$$

$$\lim_{k \to \infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} \leq \frac{\mu L}{2}.$$

x^k converge in modo quadratico.

□

Riassumiamo in una tabella le complessità e i tassi di convergenza di alcuni metodi con f convessa o fortemente convessa.

Metodo	f convessa		f fortemente convessa	
	Tasso	Complessità	Tasso	Complessità
Gradiente	sub-lineare	$O(1/\varepsilon)$	lineare	$O(\log(1/\varepsilon))$
Nesterov	sub-lineare	$O(1/\sqrt{\varepsilon})$	lineare	$O(\log(1/\varepsilon))$
Newton	super-lineare	$O(1/\sqrt{\varepsilon})$ fino a $O(1/\sqrt[3]{\varepsilon})$	sub-lineare	$O(\log\log(1/\varepsilon))$

Osservazione 2.9.3. Nel caso di f convessa, il tasso è solo sub-lineare e non superlineare poiché la Proposizione 2.9.1 richiede $\nabla^2 f$ invertibile, ossia definita positiva. Siamo dunque nel caso di f fortemente convessa. La convessità garantisce solo $\nabla^2 f$ semi-definita positiva.

Osservazione 2.9.4. Guardando solo il tasso, il metodo di Newton è dominante ma va considerato il costo di ogni iterazione. Mentre eseguirne una per il metodo del gradiente costa circa $O(n)$, una stessa per il metodo di Newton può costare anche $O(n^3)$. Questo può rendere preferibile un metodo più lento ma computazionalmente più conveniente, soprattutto se la dimensione del problema cresce.

Che problemi possono presentarsi col metodo di Newton?

- $\nabla^2 f(x^k)$ non è invertibile. L'update non è applicabile a questa iterazione e, solitamente, si utilizza un passo del metodo del gradiente.
- $\nabla^2 f(x^k)$ non è definita positiva. Questo fa sì che $-(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$ non sia di discesa. Si ovvia come sopra, utilizzando un'iterazione del metodo del gradiente.
- $\alpha_k = 1$ non è "buono". In questa situazione si esegue una ricerca di linea.

Definizione 2.9.1. Un algoritmo si dice *Metodo di Newton Globalmente Convergente* se produce una sequenza $\{x^k\}$ tale che

(i) x^k ha punti di accumulazione;

(ii) ogni punto di accumulazione è stazionario e non è un punto di massimo locale;[9]

(iii) Se x^k appartiene ad un intorno di x^* che soddisfi le ipotesi della Proposizione 2.9.1, allora $\exists \bar{k}$ tale che

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1}\nabla f(x^k), \quad \forall k \geq \bar{k}.$$

Questo significa che definitivamente, il metodo di Newton diventa convergente.[10]

Le varianti viste in precedenza per ovviare ai problemi del metodo sono alcuni dei modi conosciuti per costruire metodi come sopra.

Caso Generale

Sia il metodo del gradiente che il metodo di Newton possono essere riassunti con la formula

$$x^{k+1} = x^k - \alpha_k H_k \nabla f(x^k),$$

dove, scegliendo $H_k = I$ si ha il metodo del gradiente e scegliendo $\alpha_k = 1$ e H_k l'inversa della matrice Hessiana di f si ha il metodo di Newton.

[9] Nel metodo del gradiente si ha che $\{f(x^k)\}$ è monotona decrescente, mentre nel metodo di Newton non si controlla mai la funzione obiettivo, pertanto, è possibile incappare in un massimo locale se si entra in un suo intorno opportuno.

[10] Da un valore $\bar{k}$ in poi, il passo è sempre quello del metodo di Newton.

2.10 Metodi Quasi Newton

Prendiamo come esempio il caso quadratico

$$\min_{xin\mathbb{R}^n} \frac{1}{2}x^T Qx - c^T x, \qquad \nabla f(x) = Qx - c, \qquad \nabla^2 f(x) = Q.$$

Se calcoliamo $\nabla f(x) - \nabla f(y)$ si ha

$$Qx - c - (Qy - c) = Q(x - y) = \nabla^2 f(x)(x - y).$$

Quindi, in uno schema iterativo, dat x^k e x^{k+1},

$$\nabla f(x^{k+1}) - \nabla f(x^k) = Q(x^{k+1} - x^k). \tag{2.35}$$

L'Equazione (2.35) è detta Equazione Quasi-Newton.

Caso Generico Non Lineare

Nel caso generico non lineare, la formula di aggiornamento è del tipo

$$x^{k+1} = x^k - \alpha_k B_k^{-1} \nabla f(x^k).$$

$$x^{k+1} = x^k - \alpha_k H_k \nabla f(x^k).$$

dove

$$B_k \approx \nabla^2 f(x^k), \qquad H_k \approx (\nabla^2 f(x^k))^{-1}.$$

Chiamiamo

$$s_k := x^{k+1} - x^k, \qquad y_k := \nabla f(x^{k+1}) - \nabla f(x^k).$$

L'idea dei metodi quasi-Newton è quella di usare lo schema le approssimazioni precedenti forzando

$$B_{k+1} s_k = y_k \qquad [y_k = Q s_k],$$

$$s_k = H_{k+1} y_k, \qquad [s_k = Q^{-1} y_k].$$

dove abbiamo riportato tra parentesi quadre il caso quadratico.

Per costruire le matrici possiamo notare che

$$B_{k+1} = B_k + sB_k, \quad H_{k+1} = H_k + sH_k$$

che è utilizzata per trovare $d_k = -B_k \nabla f(x^k)$ (in realtà tramite la risoluzione del sistema $B_k d_k = -\nabla f(x^k)$) che poi, a sua volta, permette il calcolo di $x^{k+1} = x^k + \alpha_k d_k$. Lo spsotamento $s_k = \alpha_k d_k$ e

$$y_k = \nabla f(x^k + \alpha_k d_k) - \nabla f(x^k).$$

Si procede come nell'Algoritmo 6

La formula di aggiornamento diretta è data da

$$B_{k+1} = B_k + s_k, \qquad B_{k+1} s_k = y_k,$$

La formula di aggiornamento inversa è

$$H_{k+1} = H_k + s_k, \qquad H_{k+1} y_k = s_k.$$

Algorithm 6 Algoritmo dei metodi quasi-Newton

Dati H_0, B_0
Calcolo d_k
Calcolo s_k, y_k
Aggiorno $B_{k+1} = B_k + \Delta B_k$ tramite $B_{k+1}s_k = y_k$[11]
$k = k + 1$

La formula di aggiornamento di rango 1, in versione diretta

$$B_{k+1} = B_k + \rho_k u_k v_k^T, \qquad \rho_k \in \mathbb{R}, u_k, v_k \in \mathbb{R}^n,$$

mentre l'inversa è

$$H_{k+1} = H_k + \rho_k u_k v_k^T.$$

La formula di rango uno considerata migliore è la Formula di Broyden. La formula di aggiornamento di rango 2 è

$$B_{k+1} = B_k + alpha_k u_k u_k^T + b_k v_k v_k^T, \qquad H_{k+1} = H_k + \alpha_k u_k u_k^T + b_k v_k v_K^T.$$

Esiste un gruppo di formule considerate migliori.

2.10.1 Formule BFGS

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k}. \tag{D}$$

Mostriamo che è soddisfatta l'equazione di quasi Newton,

$$B_{k+1}s_k = B_k s_k + \frac{y_k y_k^T}{y_k^T s_k} s_k - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k} s_k =$$

$$= B_k s_k + y_k \frac{y_k^T s_k}{y_k^T s_k} - B_k s_k \frac{s_k^T B_k s_k}{s_k^T B_k s_k} = y_k,$$

pertanto $B_{k+1}s_k = y_k$. Analogamente si potrebbe mostrare che

$$H_{k+1} = H_k + \left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k y_k} - \frac{s_k y_k^T H_k + H_k y_k s_k^T}{s_k^T y_k}.$$

Le formule BFGS conservano dunque la validità dell'equazione dei metodi quasi-Newton ma le matrici devono essere definite positive. La H_k così ottenuta lo è?

Proposizione 2.10.1. *Sia H_k definita positiva e H_{k+1} ottenuta con una formula di aggiornamento BFGS, allora H_{k+1} è definita positiva se e solo se $s_k^T y_k > 0$.*

Se H^0 è definita positiva e $s_k^T Y_k > 0$, $\forall k$, allora H_k è definita positiva $\forall k$.

Osservazione 2.10.1. Se H_0 è definita positiva e $s_k^T y_k > 0$ $\forall k$, H_k è definita positiva $\forall k$.

Come facciamo a garantire $s_k^T y_k > 0$?

$$(x^{k+1} - x^k)^T = (\nabla f(x^{k+1}) - \nabla f(x^k)) > 0$$

Dato che $x^{k+1} = x^k + \alpha_k d_k$,

$$\alpha_k d_k^T (\nabla f(x^{k+1}) - \nabla f(x^k)) > 0.$$

Ora, $\alpha_k > 0$,

$$d_k^T(\nabla f(x^{k+1}) > d_k^T \nabla f(x^k))$$
$$d_k^T(\nabla f(x^k + \alpha_k d_k) > d_k^T \nabla f(x^k))$$

Per la condizione di Wolfe, $\nabla f(x^k + \alpha_k d_k)^T d_k \geq \sigma \nabla f(x^k)^T d_k$ con $\sigma \in (0,1)$. Allora tale condizione garantisce che

$$\nabla f(x^k + \alpha_k d_k)^T d_k \geq \nabla f(x^k)^T d_k > \nabla f(x^k)^T d_k.$$

Proposizione 2.10.2 (Proprietà di Convergenza delle BFGS)**.**

- *Se f è convessa si ha convergenza globale;*
- *Se f è fortemente convessa si ha convergenza superlineare.*

2.10.2 Problemi a Lunga Scala

Siamo interessati al solito problema

$$\min_{x \in \mathbb{R}^n} f(x), \qquad n > 1000.$$

L'applicazione dei metodi QN in questo caso richiederebbe la costruzione di matrici B e $H \in \mathbb{R}^{n \times n}$, ossia enormi. Oltre a dun problema di memoria, si ha anche un gigantesco costo di aggiornamento. Il calcolo di $d_k = -H_k \nabla f(x^k)$ diventa proibitivo.

Per ovviare a ciò e poter continuare ad utilizzare il metodo possiamo utilizzare

$$H_{k+1} = V_k^T H_k V^k + \rho_k s_k s_k^T, \qquad \rho_k = \frac{1}{y_k^T s_k}, \qquad V_k = I - \rho_k y_k s_k^T.$$

Osservazione 2.10.2. Si tratta di una formula ricorsiva che dipende da $H_k,\ y_k,\ s_k$ e estendendo la relazione ricorsiva si può calcolare H_{k+1} a partire da H_0 e dalla coppia (s_z, y_z) $z = 0, ..., k$.

Se fermiamo la ricorsione dopo m passi

$$H_{k+1} = f(H_0, \delta_0, y_0, ..., \delta_k, y_k) = f(H_{k-m}, \delta_{k-m}, y_{k-m}, ..., \delta_k, y_k).$$

Approssimiamo $H_{k-m} \approx \gamma I$.

Le considerazioni fatte nell'Osservazione precedente rendono il metodo olto conveniente dal punto di vista della memoria. Si è visto sperimentalmente che, per $m \in \{2, 3, ..., 10\}$ funziona molto bene. Si deve comunque affronatre in calcolo di

$$d_k = -H_k \nabla f(x^k).$$

Per ovviare a questo si usa l'algoritmo ricorsivo HG. Esso calcola d_k direttamente dalle m coppie (δ_z, y_z) $z = k - m + 1, ..., k$ effettuando $4mn$ moltiplicazioni. Mettendo insieme tutto,

- Calcoliamo col metodo troncato H_{k+1};
- Memorizziamo $\{(\delta_{k-m}, y_{k-m}), ..., (\delta_k, y_k)\}$;
- Calcoliamo d_k con HG.

Questo metodo, detto L-BFGS (L sta per limited memory) è il metodo migliore e quello di default per la ricerca nell'ottimizzazione non vincolata.

Osservazione 2.10.3. La convergenza del metodo non è globale.

Capitolo 3

Metodi Trust Region e Derivative Free

3.1 Apprendimento Automatico

Nell'ambito dell'apprendimento supervisionato, abbiamo problemi della forma

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(w; X, y) + \Omega(w).$$

Quando la Loss function è convessa, risulta conveniente risolvere ul problema di ottimizzazione con i metodi trattati fino ad ora.

Problemi di questo tipo sono, per esempio, la regressione logistica, la regressione multinomiale o i modellli ARMA per l'analisi di serie storiche.

3.1.1 Regressione Logistica

Nella regressione logistica la Loss function è fata da

$$\mathcal{L}(w; X, y) = \sum_{i=1}^{N} \log\left(1 + e^{-y^{(i)} w^T x^{(i)}}\right),$$

dove consideriamo $y \in \{-1, 1\}^n$ invece che in $\{0, 1\}^n$ [1].

Se chiamiamo $z = Xw$, $z_i = w^T x^{(i)}$,

$$\mathcal{L}(w; X, y) = \phi(z, y) = \sum_{i=1}^{n} \log\left(1 + e^{-y^{(i)} z_i}\right).$$

$$\nabla_w \mathcal{L}(w; X, y)^T = \nabla_w \phi(Xw, y) = \nabla_z \phi(Xw, y)^T \frac{\partial Xw}{\partial w}.$$

$$\frac{\partial Xw}{\partial w} = X.$$

$$(\nabla_z \phi(z; y))_i = \frac{\partial}{\partial z_i} \phi(z; y) = \frac{\partial}{\partial z_i} \sum_{j=1}^{N} \log\left(1 + e^{-y^{(j)} zj}\right) =$$

[1] Altrimenti andrebbe modificata la forma.

solo un termine della sommatoria dipende da z_i e quindi sopravvive solo quello

$$= \frac{\partial}{\partial z_i} \log\left(1 + e^{-y^{(i)} z_i}\right) = \frac{-y^{(i)} e^{-y^{(i)} z_i}}{1 + e^{-y^{(i)} z_i}} = -y^{(i)} \frac{e^{-y^{(i)} z_i}}{1 + e^{-y^{(i)} z_i}} =$$

$$-y^{(i)} \sigma\big(-y^{(i)} z_i\big),$$

dove σ è la funzione sigmoide

$$\sigma(t) = \frac{e^{tx}}{1 + e^{tx}} = \frac{1}{1 + e^{-tx}}.$$

Ora,

$$(\nabla_z \phi(z, y))_i = -y^{(i)} \sigma(-y^{(i)} z_i),$$
$$(\nabla_z \phi(Xw, y))_i = -y^{(i)} \sigma(-y^{(i)} w^T x^{(i)}),$$

Detto r tale gradiente

$$\nabla_z \phi(Xw, y) := r,$$
$$\nabla_w \mathcal{L}(w; X, y) = r^T X = (X^T r)^T.$$

Senza eseguire i calcoli,

$$\nabla_w^2 \mathcal{L}(w; X, y) = X^T D X,$$

dove D è una matrice diagolane in cui

$$D_{ii} = \sigma(-y^{(i)} w^T x^{(i)}) \sigma(y^{(i)} w^T x^{(i)}).$$

3.2 Metodi Trust Region

Nei metodi linesearch, si definisce una direzione e si cerca di ottimizzare la funzione lungo tale direzione. Nei metodi *Trust Region*, invece, dal punto x^k, si esplora un intorno del punto. Lo spostamento s^k è scelto in una regione detta *di confidenza* (trust region) Δ_k. Ad esempio, la regione può essere costituita da $B(x^k, \rho_k)$. Invece di utilizzare la funzione obiettivo, nell'intorno si utilizza una sua approssimazione, un modello $m_k(x)$ tale che

$$m_k(x) \approx f(x), \quad \forall x \in \Delta_k.$$

Se prendiamo come approssimazione lo sviluppo di Taylor del secondo ordine, otteniamo

$$m_k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k).$$

In generale, nei modelli quadratici,

$$m_k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T B_k (x - x^k),$$

con B_k matrice simmetrica.

Osservazione 3.2.1. B_k non ha bisogno di ipotesi aggiuntive oltre alla simmetria. Dato che minimizziamo in un intorno compatto, non occorre che la funzione sia convessa (ossia B_k semi-definita positiva) e tantomeno che sia strettamente convessa (B_k definita positiva).

Il sottoproblema Trust Region diventa

$$\arg \min_{||x-x^k|| \leq \rho_k} m_k(x),$$

in cui appunto $x \in B(x^k, \rho_k)$. Il problema è dunque vincolato. Per esso possiamo dunque usare uno qualsiasi dei metodi per l'ottimizzazione vincolata, ma solitamente i metodi che funzionano meglio sono quelli che presentano un metodo di risoluzione del sottoproblema specializzato.

Detta $\hat{x}$ la soluzione del sottoproblema, possiamo essere certi di avere qualcosa di meglio di x^k? Dato che non abbiamo minimizzato f ma una sua approssimante, questo non è garantito, ciò che è certo è che $m_k(\hat{x}) \leq m_k(x^k)$. Occorre una misura di bontà della soluzione trovata,

$$\eta_k = \frac{f(x^k) - f(\hat{x})}{m_k(x^k) - m_k(\hat{x})}.$$

Il denominatore è la differenza ottenuta su m_k tra $\hat{x}$ e x^k, ossia il *decremento atteso* secondo il modello. Come già osservato,

$$m_k(x^k) - m_k(\hat{x}) \geq 0.$$

Osservazione 3.2.2. Per i modelli quadratici $m_k(x^k) = f(x^k)$ perché nel centro di approssimazione, tale approssimazione è esatta.

Il numeratore è la variazione osservata della funzione obiettivo.

Dunque, η_k quantifica l'accordo tra modello e funzione obiettivo per quanto riguarda il decremento. Si possono avere tre situazioni

- Se $\eta_k < 0$ significa che la funzione è cresciuta mentre il modello è decrementato. $\hat{x}$ non è una buiona soluzione e m_k non è affidabile su Δ_k.
- Se $\eta_k > 0$ ma piccolo, $\hat{x}$ è buono (anche f decresce) ma m_k non è molto affidabile su Δ_k.
- Se $\eta_k >> 0$, $\hat{x}$ è buona e m_k è un modello affidabile su Δ_k.

Se $\hat{x}$ è buona, $x^{k+1} = \hat{x}$, altrimenti $x^{k+1} = x^k$ ma, se m_k è buono $\rho_{k+1} \geq \rho_k$, se invece esso non lo è, $\rho_{k+1} < \rho_k$.

Lo schema formale è quindi l'Algoritmo 7.

L'efficienza del metodo dipende dalla risoluzione del sottoproblema.

3.2.1 Sufficiente Decremento del Metodo Quadratico

Siamo interessati a

$$\min_{||s|| \leq \rho_k} m_k(x^k + s) = \min_{||s|| \leq \rho_k} f(x^k) + \nabla f(x^k)^T s + \frac{1}{2} s^T B_k s.$$

Se $s = 0$, $\nabla m_k(x^k + s) = \nabla f(x^k)$. Supponiamo ora di utilizzare uno spostamento lungo l'antigradiente, ossia

$$s = -\tau \nabla f(x^k),$$

il problema diventa

$$\min_\tau f(x^k) - \tau \nabla f(x^k) \nabla f(x^k) + \frac{1}{2} \tau^2 \nabla f(x^k)^T B_k \nabla f(x^k) =$$

$$= \min_\tau f(x^k) - \tau ||\nabla f(x^k)||^2 + \frac{1}{2} \tau^2 \nabla f(x^k)^T B_k \nabla f(x^k)$$

Algorithm 7 Algoritmo Generale Trust Region.

Dati $x^0 \in \mathbb{R}^n$, $\rho_0 > 0$, $0 < c_2 < c_1 < 1$, $0 < \gamma_2 < \gamma_1 < 1$, $k = 0$
while $\nabla f(x^k) \neq 0$ **do**
$m_k(x) = f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T B_k(x - x^k)$
Risolvo (In modo esatto o approssimato) il sottoproblema

$$\arg \min_{||x - x^k|| \leq \rho_k} m_k(x) = \hat{x}$$

Calcolo

$$\eta_k = \frac{f(x^k) - f(\hat{x})}{m_k(x^k) - m_k(\hat{x})}$$

if $\eta_k \geq c_2$ **then**
$x^{k+1} = \hat{x}$
else
$x^{k+1} = x^k$
end if
if $\eta_k \geq c_1$ **then**
$\rho_{k+1} \in [\rho_k, +\infty)$
end if
if $c_2 \leq \eta_k < c_1$ **then**
$\rho_{k+1} \in [\gamma_1 \rho_k, \rho_k)$
end if
if $\eta_k < c_2$ **then**
$\rho_{k+1} \in [\gamma_2 \rho_k, \gamma_1 \rho_k)$
end if
$k = k + 1$
end while

La condizione $||s|| \leq \rho_k$ si trasforma in

$$||\tau \nabla f(x^k)|| \leq \rho_k \quad \Leftrightarrow \quad \tau \leq \frac{\rho_k}{||\nabla f(x^k)||}.$$

Osservazione 3.2.3. Notiamo che tale quantità è positiva.

Si può notare che la funzione obiettivo è una parabola. La concavirtà dipende dal segno di $\nabla f(x^k)^T B_k \nabla f(x^k)$, allora

- Se $\nabla f(x^k)^T B_k \nabla f(x^k) < 0$, la parobola è rivolta verso il basso e ha un massimo in

$$\frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^T B_k \nabla f(x^k)}.$$

Questa quantità è negativa in queste ipotesi, pertanto il massimo precede il vincolo, come

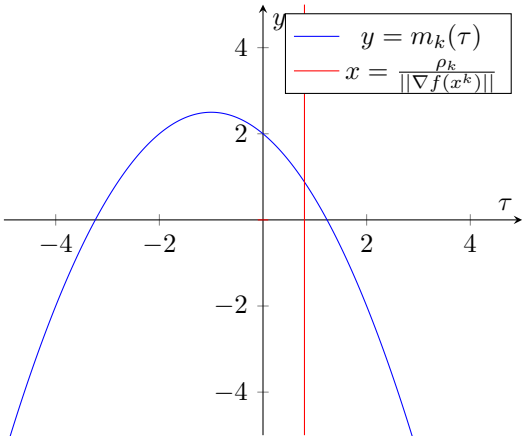

Figura 3.1: $m_k(\tau)$ nel caso $\nabla f(x^k)^T B_k \nabla f(x^k) < 0$.

osservato nell'Osservaxzione 3.2.3. Allora il punto di minimo si ha in

$$\tau^* = \frac{\rho_k}{||\nabla f(x^k)||}.$$

- Se $\nabla f(x^k)^T B_k \nabla f(x^k) > 0$ la parabola è rivolta verso l'alto e si ha un minimo in

$$\frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^T B_k \nabla f(x^k)} > 0.$$

Allora, se il vincolo precede il minimo, il vincolo è la soluzione del problema, se il vicnolo si trova dopo il minimo, è il minimo stesso (poiché è ammissibile) la soluzione, pertanto

$$\tau^* = \min\left\{ \frac{||\nabla f(x^k)||^2}{\nabla f(x^k)^T B_k \nabla f(x^k)}, \frac{\rho_k}{||\nabla f(x^k)||} \right\}.$$

τ^* è detto passo di Cauchy e $x^k - \tau^* \nabla f(x^k)$ è detto punto di Cauchy.

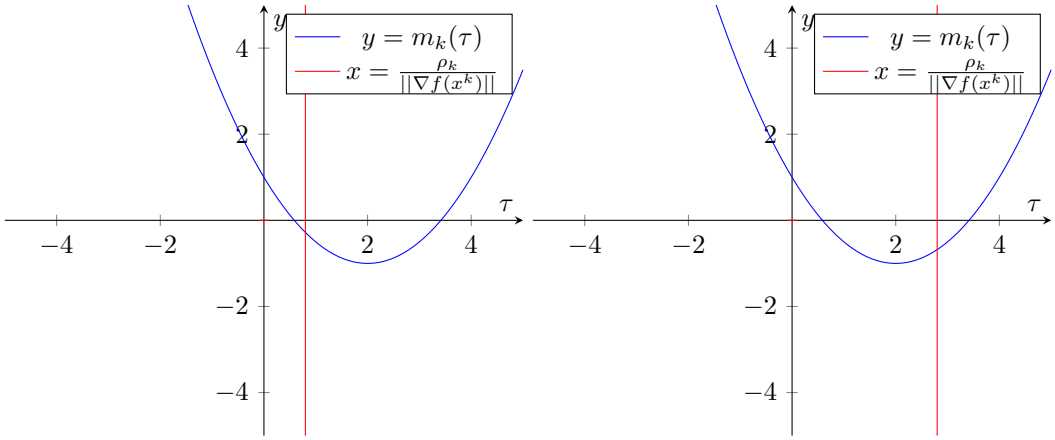

Figura 3.2: $m_k(\tau)$ nel caso $\nabla f(x^k)^T B_k \nabla f(x^k) > 0$.

Definizione 3.2.1. In un metodo Trust Region si dice che abbiamo sufficiente decremento del modello quadratico se la soluzione $\hat{x}$ del sottoproblema è tale che

$$m_k(\hat{x}) \leq m_k(x^k - \tau^* \nabla f(x^k)).$$

Si possono dimostrare proprietà di convergenza globale per metodi Trust Region che soddisfano questa proprietà per ogni k.

3.3 Metodi Derivative Free

Ancora una volta siamo interessati a

$$\min_{x \in \mathbb{R}^n} f(x),$$

ma in questo caso f è di tipo black box. Può quindi succedere che f esista ma sia troppo costoso da calcolare, oppure che non sia disponibile (talvolta f sono valori misurati e non esiste un'espressione esplicita).

Osservazione 3.3.1. Non è il caso in cui f non è differenziabile o non è smooth.

La prima iodea che potremmo avere è di utilizzare un metodo come le *differenze finite* per approssimare la derivata.

3.3.1 Approssimazione alle Differenze Finite

$$\frac{\partial}{\partial x_i} f(x) \approx \frac{f(x + \varepsilon e_i) - f(x)}{\varepsilon}, \quad \varepsilon > 0,$$

con e_i i-esimo vettore della base canonica di $\mathbb{R}^n$. L'errore commesso con tale approssimazione è

$$\left| \frac{\partial}{\partial x_i} f(x) - \frac{f(x + \varepsilon e_i) - f(x)}{\varepsilon} \right| \leq c_f \varepsilon.$$

Mantenendo ε piccolo, il gradiente sembra essere una buona approssimazione. Tuttavia, dato che lavoriamo in aritmetica finita, detta μ la precisione di macchina,

$$\left|\frac{\partial}{\partial x_i}f(x) - \frac{f(x+\varepsilon e_i)-f(x)}{\varepsilon}\right| \leq c_f\varepsilon + \frac{c_2(f,\mu)}{\varepsilon},$$

e i due termini entrano in competizione: mandando a 0 il primo si fa esplodere ils econdo. Inoltre, se la f è misurata con del rumore, non disponiamo di f ma di $\hat{f}$,

$$\hat{f}(x) = f(x) + r(x).$$

Allora l'errore diventa

$$\left|\frac{\partial}{\partial x_i}f(x) - \frac{f(x+\varepsilon e_i)-f(x)}{\varepsilon}\right| \leq c_f\varepsilon + \left|\frac{r(x+\varepsilon e_i)-r(x)}{\varepsilon}\right|.$$

Per mandare a 0 il primo termine si fa esplodere il secondo. Nella pratica questi schemi non sono applicabili quasi mai, soprattutto se serve un risultato con una determinata precisione.

3.3.2 Direct Search e Metodo delle Coordinate

Se consideriamo l'insieme delle direzioni coordinate e anticoordinate, una di queste deve necessariamente essere una direzione di discesa, ossia deve formare con $-\nabla f(x^k)$ un angolo acuto. infatti, preso

$$D = \{e_1, e_2, ..., e_n, -e_1, -e_2, ..., -e_n\},$$

e considerato

$$\nabla f(x^k) = \sum_{i=1}^{n} a_i e_i,$$

$$-\nabla f(x^k)^T \nabla f(x^k) < 0$$

$$0 > -\sum_{i=1}^{n}(a_i e_i)^T \nabla f(x^k) = -\sum_{i=1}^{n} a_i e_i^T \nabla f(x^k).$$

Almeno uno dei prodotti nella sommatoria deve essere non nullo, allora $\exists i \in \{1, ..., n\}$ tale che

$$e_i^T \nabla f(x^k) \neq 0.$$

Se $e_i^T \nabla f(x^k) < 0$ allora $e_i \in D$ è di discesa, se, invece, $e_i^T \nabla f(x^k) > 0$, $-e_i^T \nabla f(x^k) < 0$ e $-e_i \in D$ è di discesa.

Se considero

$$x^k \pm \alpha e_i,$$

con α sufficientemente piccolo, almeno uno di essi garantisce un decremento di f.

Osservazione 3.3.2. La condizione di uscita del *while* non può essere posta sul gradiente perché non è disponibile.

Aggiungiamo alcune ipotesi, $f \in C^1(\mathbb{R}^n)$ e $\mathcal{L}_f(x^0)$ compatto.

Lemma 3.3.1. *Sotto le ipotesi precedenti,*

$$\lim_{k\to\infty} \alpha_k = 0. \tag{3.1}$$

Algorithm 8 Metodo delle Coordinate

Dati $\alpha_0 > 0$, $\theta \in (0,1)$ $x^0 \in \mathbb{R}^n$ $k = 0$ e D come sopra
while $\alpha_k > \varepsilon$ **do**
  if $\exists d \in D$ tale che $f(x^k + \alpha_k d) < f(x^k)$ **then**
    $x^{k+1} = x^k + \alpha_k d$
    $\alpha_{k+1} = \alpha_k$
  else
    $x^{k+1} = x^k$
    $\alpha_k = \theta \alpha_k$
  end if
  $k = k + 1$
end while

Proposizione 3.3.1. *La sequenza $\{x^k\}$ prodotta dal metodo delle coordinate ha punti di accumulazione e almeno uno di essi è stazionario.*

Dimostrazione.
Per costruzione dell'algoritmo, $\alpha_{k+1} \leq \alpha_k$ e per la (3.1), $\alpha_k \to 0$, pertanto $\exists K \subseteq \{0, 1, ...\}$ sottosequenza tale che

$$\alpha_{k+1} < \alpha_k, \quad k \in K.$$

Sempre per costruzione, $f(x^{k+1}) \leq f(x^k)$ $\forall k$, pertanto $\{x^k\} \subseteq \mathcal{L}_f(x^0)$ compatto e

$$\{x^k\}_{k \in K} \subseteq \mathcal{L}_f(x^0).$$

Allora esiste una sottosequenza $K_1 \subseteq K$ convergente,

$$\lim_{\substack{k \to \infty \\ k \in K_1}} x^k = \bar{x}.$$

$\forall k \in K_1$ vale che $f(x^k + \alpha_k) \geq f(x^k)$ poiché, se α_{k+1} decrementa rispetto a α_k significa che abbiamo un passo di insuccesso. Allora

$$f(x^k \pm \alpha_k e_i) \geq f(x^k), \quad \forall i = 1, ..., n.$$

Per il Teorema della Media

$$f(x^k + \alpha_k e_i) = f(x^k) + \alpha_k e_i^T \nabla f(u_k), \quad u_k = x^k + t_i^+ \alpha_k e_i, \quad t_i^+ \in (0,1).$$

$$f(x^k - \alpha_k e_i) = f(x^k) - \alpha_k e_i^T \nabla f(v_k), \quad v_k = x^k + t_i^- \alpha_k e_i, \quad t_i^- \in (0,1).$$

Per ogni i,

$$\alpha_k e_i^T \nabla f(u_k) \geq 0, \qquad -\alpha_k e_i^T \nabla f(v_k) \geq 0,$$

allora $\forall i$ e $\forall k \in K_1$,

$$e_i^T \nabla f(u_k) \geq 0, \qquad e_i^T \nabla f(v_k) \leq 0.$$

Ora,

$$\lim_{\substack{k \to \infty \\ k \in K_1}} u_k = \lim_{\substack{k \to \infty \\ k \in K_1}} x^k + t_i^+ \alpha_k e_i = \bar{x},$$

$$\lim_{\substack{k \to \infty \\ k \in K_1}} v_k = \lim_{\substack{k \to \infty \\ k \in K_1}} x^k + t_i^- \alpha_k e_i = \bar{x},$$

passando al limite

$$e_i^T \nabla f(\bar{x}) \geq 0, \quad e_i^T \nabla f(\bar{x}) \leq 0,$$

$$\frac{\partial}{\partial x_i} f(\bar{x}) = e_i^T \nabla f(\bar{x}) = 0, \quad \forall i = 1, ..., n.$$

pertanto $\bar{x}$ è stazionario. □

Capitolo 4

Ottimizzazione Multi-Obiettivo

Introduzione

Si cerca di risolvere il problema

$$\min_{x\in\mathbb{R}^n} F(x),$$

nel caso non vincolato, con

$$F(x) = [f_1(x), ..., f_m(x)]^T, \qquad m > 1.$$

$F : \mathbb{R}^n \to \mathbb{R}^m$, $f_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, ..., m$.

Ora, mentre nel caso singolo è facile definire una soluzione x migliore di un'altra y basandosi sul valore della funzione obiettivo f, nel caso multiplo, occorre definire un *ordine parziale* su $\mathbb{R}^n$. Se $x, y \in \mathbb{R}^n$,

$$F(x) \leq F(y) \;\Leftrightarrow\; f_j(x) \leq f_j(y), \quad j = 1, ..., m.$$
$$F(x) < F(y) \;\Leftrightarrow\; f_j(x) < f_j(y), \quad j = 1, ..., m.$$

Osservazione 4.0.1. Mentre nel caso singolo $f(x) \nleq f(y)$ vuol dire direttamente $f(x) > f(y)$, nel caso multiplo, trattandosi di ordine parziale, $F(x) \nleq F(y)$ non implica $F(x) > F(y)$ ma vuol dire che $\exists \bar{j} \in \{1, ..., m\}$ tale che $f_{\bar{j}}(x) \geq f_{\bar{j}}(y)$.

Definizione 4.0.1 (Dominanza)**.** Siano $x, y \in \mathbb{R}^n$, diciamo che x domina y ($F(x) \lneqq F(y)$) se

$$F(x) \leq F(y) \;\wedge\; \exists \bar{j} \text{ tc } f_{\bar{j}}(x) < f_{\bar{j}}(y).$$

Osservazione 4.0.2. In altre parole x domina y se almeno per una funzione x risulta migliore di y e per le altre al massimo la pareggia.

Definizione 4.0.2 (Ottimo di Pareto)**.** $x^* \in \mathbb{R}^n$ è detto *ottimo di PAreto* per il problema (4) se $\nexists y \in \mathbb{R}^n$ tale che

$$F(y) \lneqq F(x^*).$$

$x^* \in \mathbb{R}^n$ è detto *ottimo di PAreto locale* per il problema (4) se $\nexists y \in B(x^*, \varepsilon)$ con $\varepsilon > 0$ tale che

$$F(y) \lneqq F(x^*).$$

L'insieme degli ottimi di Pareto è detto *Set di Pareto*, l'immagine di tale insieme è detta *Fronte di Pareto* e la sua geometria è oggetto di studio. Una delle geometrie più semplici, che però può risultare anche da funzioni molto complesse, è la forma iperbolica.

L'ottimalità di Pareto è un concetto molto forte, allora spesso si ricerca qualcosa di più debole.

Definizione 4.0.3. $x^* \in \mathbb{R}^n$ è ottimo di Pareto debole [locale] se $\nexists y \in \mathbb{R}^n$ [$\nexists yinB(x^*, \varepsilon)$ con $\varepsilon > 0$] tale che

$$F(y) < F(x^+).$$

Esempio 4.0.1. Vediamo un caso unidimensionale ()$x \in \mathbb{R}$).

$$\min_{x\in[0,1]} [-x, ||x||_0]^T.$$

Gli ottimi di Pareto sono $x = 1$ e $x = 0$ ma gli ottimi di Pareto deboli sono $x \in [0, 1]$.

Molti dei risultati ottenuti per il caso singolo si possono ottenere anche per il caso multiplo. Ad esempio, per il problema convesso, sappiamo che un ottimo locale è anche ottimo globale.

Definizione 4.0.4 (Problema Convesso)**.** Si dice problema convesso un problema in cui l'insieme ammissibile è convesso e f_j è convessa $\forall j = 1, ..., m$ (ossia F è convessa per componenti).

Proposizione 4.0.1. *Dato un problema convesso, se $x^* \in \mathbb{R}^n$ è ottimo di Pareto locale, allora x^* è un ottimo globale.*

Dimostrazione.
Supponiamo per ssurdo che x^* non sia ottimo globale. Allora $\exists \hat{x} \in \mathbb{R}^n$ tale che $F(\hat{x}) \leq F(x^*)$ e $\exists \bar{j}$ tale che

$$f_{\bar{j}}(\hat{x}) \leq f_{\bar{j}}(x^*).$$

Consideriamo i punti

$$(1 - \lambda)x^* + \lambda\hat{x}, \quad \lambda \in [0, 1].$$

$\exists \bar{\lambda} \in (0, 1)$ tale che $z := (1 - \bar{\lambda})x^* + \bar{\lambda}\hat{x} \in B(X^*, \varepsilon)$. Per la convessità di F per componenti

$$F(z) \leq (1 - \bar{\lambda})F(x^*) + \bar{\lambda}F(\hat{x}) \leq (1 - \bar{\lambda})F(x^*) + \bar{\lambda}F(x^*) = F(x^*),$$

dove abbiamo usato che $F(\hat{x}) \leq F(x^*)$.

Esiste però $\bar{j}$ tale che $f_{\bar{j}}(z) < f_{\bar{j}}(x^*)$? No, altrimenti, dato che $z \in B(x^*, \varepsilon)$, x^* non sarebbe nemmeno ottimo locale. Necessariamente $F(z) = F(x^*)$. Se così fosse, però

$$F(x^*) = F\big((1 - \bar{\lambda})x^* + \bar{\lambda}\hat{x}\big) \leq (1 - \bar{\lambda})F(x^*) + \bar{\lambda}F(\hat{x}),$$

ossia

$$\bar{\lambda}F(x^*) \leq \bar{\lambda}F(\hat{x})$$
$$F(x^*) \leq F(\hat{x})$$

contraddicendo le ipotesi. Si ha dunque la tesi. □

Proposizione 4.0.2 (Condizione necessaria del 1° ordine)**.** *$x^* \in \mathbb{R}^n$ ottimo di Pareto, allora $\nexists d \in \mathbb{R}^n$ tale che*

$$\nabla f_j(x^*)^T d < 0, \quad j = 1, ..., m.$$

Nel caso multi-obiettivo non è possibile calcolare tutti i gradienti e controllare che facciano 0, questo perché nella pratica non succede a tutti contemporaneamente. Si ricorre dunque alla stazionarietà di Pareto.

Definizione 4.0.5. x^* si dice stazionario di Pareto se

$$\forall d \in \mathbb{R}^n \ \exists \bar{j} \in \{1, ..., m\} \text{ tc } \nabla f_{\bar{j}}(x^*)^T d \geq 0.$$

Più compatta

$$\forall d \in \mathbb{R}^n \ \max_{j=1,...,m} \nabla f_j(x^*)^T d \geq 0,$$

o ancora

$$\min_{d \in \mathbb{R}^n} \max_{j=1,...,m} \nabla f_j(x^*)^T d = 0.$$

Proposizione 4.0.3 (Condizione sufficiente per il problema convesso)**.**

- *Se x^* è Pareto stazionario, x^* è un ottimo di PAreto debole;*
- *Se $\nabla^2 f_j(x) > 0 \ \forall x \in \mathbb{R}^n$ e $\forall j = 1, ..., m$, se x^* è Pareto stazionario, allora è un ottimo di Pareto forte.*

4.1 Metodi per Ottimizzazione Multi-Obiettivo

4.1.1 Metodo Scalarizzato o Pesato

Quello che facciamo è considerare m pesi $w_1, ..., w_m$ tali che $\sum_{j=1}^m w_j = 1$ e risolvere

$$\min_{x \in \mathbb{R}^n} \sum_{j=1}^m w_j f_j(x).$$

Questo metodo, sebbene sia la prima e più immediata strategia per trattare i problemi multipli, presenta tre problemi:

(i) la scelta dei pesi può essere decisiva e costosa;

(ii) Spesso si incorre in problemi illimitati;

(iii) Si passa dal caso vettoriale al caso scalare.

4.1.2 Metodo del Gradiente

Nel caso singolo si aveva

$x_0 \in \mathbb{R}^n$, $k = 0$
while $||\nabla f(x_k)|| \neq 0$ **do**
  $d_k = -\nabla f(x_k)$
  Cerco $\alpha_k \in \mathbb{R}^n$ (con Armijo)
  $x_{k+1} = x_k + \alpha_k d_k$
  $k = k + 1$
end while

Cosa occorre cambiare nel caso multiplo? Condizione di uscita del ciclo, direzione e passo. La prima si risolve cambiando la stazionarietà con la Pareto stazionarietà. Da essa si ottiene come calcolare la direzione. Infatti, se x^* non è Pareto stazionario, $\exists d \in \mathbb{R}^n$ di discesa per tutte le funzioni.

$$d_k \in \arg\min_{d \in \mathbb{R}^n} \max_{j=1,...,m} \nabla f(x^*)^T d.$$

Il problema così posto potrebbe però essere illimitato, allora restringiamo la ricerca alle sole direzioni con norma minore o uguale a 1.

$$d_k \in \arg\min_{||d||\leq 1} \max_{j=1,\dots,m} \nabla f(x^*)^T d.$$

Altrimenti, possiamo aggiungere una penalità su d ottenendo la forma seguente (in cui si ha l'uguale perché la soluzione è unica dato che la funzione obiettivo è coerciva)

$$d_k = \arg\min_{d\in\mathbb{R}^n} \max_{j=1,\dots,m} \nabla f(x^*)^T d + \frac{1}{2}||d||^2.$$

Infine, il passo è trovato come con Armijo, dimezzando un passo iniziale $\alpha = 1$ fino a che non risolve

$$\forall j = 1, ..., m, \quad f_j(x_k + \alpha_k d_k) \leq f_j(x_k) + \gamma\alpha\nabla f_j(x_k)^T d_k.$$

Tale condizione è realizzata dato che d_k è di discesa per tutte le f_j.

L'algoritmo diventa

$x_0 \in \mathbb{R}^n$, $k = 0$
while x^+ non è Pareto stazionario **do**
  $d_k = \arg\min_{d\in\mathbb{R}^n} \max_{j=1,\dots,m} \nabla f(x^*)^T d + \frac{1}{2}||d||^2$
  Cerco $\alpha_k \in \mathbb{R}^n$ (con Armijo)
  $x_{k+1} = x_k + \alpha_k d_k$
  $k = k + 1$
end while

Osservazione 4.1.1. Le prime due modifiche trasformano quelli che nel primo algoritmo erano solo calcoli, in veri e propri problemi di ottimizzazione a loro volta. Mentre nel caso bidimensionale esistono ancora formule chiese per d_k, per i casi multidimensionali no. Se però $n >> m$, esiste una formulazione duale del problema della ricerca di d_k con solo m variabili e il problema non esplode.

Come si può dare in pasto ad un solver il problema della ricerca di d_k? Si tradforma nel seguente sistema con l'aggiunta di una variabile e l'imposizione di m vincoli,

$$\min_{\substack{t\in\mathbb{R}\\ d\in\mathbb{R}^n}} t + \frac{1}{2}||d||^2, \qquad \nabla f_j(x_k)^T d_k \leq t \ \forall j = 1, ..., m.$$

Capitolo 5

Ottimizzazione Vincolata

5.1 Introduzione

Si cerca la soluzione al problema

$$\min_{x \in S \subset \mathbb{R}^n} f(x). \tag{5.1}$$

A differenza del caso non vincolato, il problema non richiede solo la valutazione della funzione obiettivo ma anche la verifica che un'eventuale soluzione sia ammissibile. Occorre inoltre modificare il concetto di direzione di discesa.

Definizione 5.1.1. Sia $x \in S$, $d \in \mathbb{R}^n$ è una *direzione ammissibile* se $\exists \bar{t} > 0$ tale che

$$x + td \in S, \quad \forall \bar{t} \in [0, \bar{t}].$$

Proposizione 5.1.1. *Se $\bar{x} \in S$ è un punto di minimo per (5.1), allora non esistono direzioni ammissibili e di discesa in $\bar{x}$.*

Proposizione 5.1.2. *Sia $\bar{x} \in S$ di minimo e $f \in C^1(\mathbb{R}^n)$, allora non esiste una direzione $d \in \mathbb{R}^n$ ammissibile e tale che*

$$\nabla f(\bar{x})^T d < 0.$$

Proposizione 5.1.3. *Sia $\bar{x} \in S$ di minimo e $f \in C^2(\mathbb{R}^n)$, allora non esiste una direzione $d \in \mathbb{R}^n$ ammissibile e tale che*

$$\nabla f(\bar{x})^T d = 0, \qquad d^T \nabla^2 f(\bar{x}) d < 0,$$

ossia a curvatura negativa.

Definizione 5.1.2. $\bar{x} \in \mathbb{R}^n$ si dice *punto stazionario* se $\nabla f(\bar{x})^T d \geq 0$ per ogni d ammissibile.

Osservazione 5.1.1. Se ogni direzione è ammissibile, l'unica opzione è che $\nabla f(\bar{x}) = 0$.

Da ora in poi assumiamo che S sia convesso.

Proposizione 5.1.4. *Sia $\bar{x} \in \mathbb{R}^n$, $\forall x \in S$ con $x \neq \bar{x}$, $d = x - \bar{x}$ è ammissibile in $\bar{x}$.*

Dimostrazione.
$\forall \lambda \in [0,1]$, $\lambda x + (1-\lambda)\bar{x} \in S$, $\forall x, \bar{x} \in S$.

$$\bar{x} + \lambda(x - \bar{x}) \in S \quad \forall \lambda \in [0,1].$$

Questo vuol dire che $x - \bar{x}$ è ammissibile con $\bar{t} = 1$. □

Proposizione 5.1.5. *Se $\bar{x} \in S$ è punto di minimo allora $\nabla f(\bar{x})^T(x - \bar{x}) \geq 0 \ \forall x \in S$.*

Dimostrazione.
Se per assurdo $\hat{x} \in S$ è tale che $\nabla f(\bar{x})^T(\hat{x} - \bar{x}) < 0$, allora $d = \hat{x} - \bar{x}$ è ammissibile (poiché S è convesso) e di discesa. □

Proposizione 5.1.6. *Se f è convessa allora $\bar{x}$ è di minimo globale se e solo se $\nabla f(\bar{x})^T(x-\bar{x}) \geq 0$ $\forall x \in S$.*

Dimostrazione.

- $\Rightarrow$ Già vista.
- $\Leftarrow$ Supponiamo che $\nabla f(\bar{x})^T(x - \bar{x}) \geq 0 \ \forall x \in S$, per la convessità di f, vale che

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}), \ \forall x \in S.$$

Ma, dato che $\nabla f(\bar{x}^T(x - \bar{x})) \geq 0$,

$$f(x) \geq f(\bar{x}), \ \forall x \in S.$$

□

5.2 Vincoli Poliedrali

Definizione 5.2.1. Un poliedro è un insieme

$$P = \{x \mid Ax \leq b, \ A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m\}.$$

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix}, \qquad \begin{cases} a_1^T x \leq b_1 \\ \vdots \\ a_m^T x \leq b_m \end{cases}$$

I punti dell'insieme soddisfano un sistema di disuguaglianze lineari.

Le direzioni ammissibili sono definite dove i vincoli valgono con l'uguaglianza.

$$I(x) = \left\{i \in \{1, ..., m\} \mid a_i^T x = b_i\right\}$$

è l'insieme degli indici dei vincoli attivi. Se $\bar{x}$ è un punto interno, $I(\bar{x}) = \emptyset$.

Proposizione 5.2.1. *Se $\bar{x} \in P$, $d \in \mathbb{R}^n$ è ammissibile in $\bar{x}$ se e solo se $a_i^T d \leq 0$, $\forall i \in I(\bar{x})$.*

Dimostrazione.

- $\Rightarrow$ d è ammissibile in $\bar{x}$ e per assurdo $j \in I(\bar{x})$ tale che $a_j^T d > 0$.
  $\forall t > 0$, $(\bar{x} + td)^T a_j = \bar{x}^T a_j + td^T a_j > b_J$ poiché $\bar{x}^T a_j = b_j$ e $d^T a_j > 0$, e questo è assurdo poiché d è ammissibile.
- $\Leftarrow$ $a_i^T d \leq 0 \ \forall i \in I(\bar{x})$. Vogliamo mostrare che $\exists \bar{t}$ tale che $\bar{x} + td \in P \ \forall t \in [0, \bar{t}]$.

(i) $i \in I(\bar{x})$,
$$(\bar{x} + td)^T a_i = \bar{x}^T a_i + t a_i^T d = b_i + t a_i^T d \leq b_i., \qquad \forall t.$$

(ii) $i \notin I(\bar{x})$ e $a_i^T d \leq 0$
$$(\bar{x} + td)^T a_i = \bar{x}^T a_i + t a_i^T d \leq b_i + t a_i^T d \leq b_i, \qquad \forall t.$$

(iii) $i \notin I(\bar{x})$ e $a_i^T d > 0$
$$(\bar{x} + td)^T a_i = \bar{x}^T a_i + t a_i^T d,$$
risulta $\leq b_i$ se
$$t \leq \frac{b_i - \bar{x}^T a_i}{a_i^T d},$$
quantità positiva in queste ipotesi. Pertanto
$$(\bar{x} + td)^T a_i \leq b_i, \qquad t \in \left[0, \frac{b_i - \bar{x}^T a_i}{a_i^T d}\right].$$
Basterà scegliere
$$\bar{t} = \min_{\substack{i \notin I(\bar{x}) \\ a_i^T d > 0}} \frac{b_i - \bar{x}^T a_i}{a_i^T d}.$$

□

Proposizione 5.2.2. *Sia* $\bar{x} \in \{x \mid x^T a_i \leq b_i \ \forall i = 1, ..., m \text{ e } \mu_j^T x = b_j \forall j = 1, ..., p\}$. d *è ammissibile se e solo se* $a_i^T d \leq 0 \ \forall i \in I(\bar{x})$ *e* $\mu_j^T d = 0 \ \forall j = 1, ..., p$.

Dimostrazione.
$$\mu_j^T x = b_j \ \Leftrightarrow \ \begin{cases} \mu_j^T x \leq b_j \\ -\mu_j^T x \leq b_j \end{cases}$$
Se $\bar{x} \in S$, $\mu_j^T \bar{x} = b_j$ e i due vincoli sono attivi, allora
$$\mu_j^T d \geq 0, \qquad \mu_j^T d \leq 0,$$
allora $\mu_j^T d = 0$. □

5.3 Vincoli di Box

Il vincolo è del tipo
$$S = \{x \mid l_i \leq x_i \leq u_i, \ i = 1, ..., n\}.$$
Analizzando i casi possibili, se $\bar{x}_i = l_i$, e_i è ammissibile[1].

Se $\bar{x}_i = u_i$, $-e_i$ è ammissibile. Se $l_i < x_i < u_i$, sia e_i che $-e_i$ sono ammissibili.

Una soluzione ottimale $\bar{x}$ deve avere $\nabla f(\bar{x})^T d \geq 0$ per ogni d ammissibile, pertanto si hanno tre casi:

[1]Ricordiamo che con e_i si intende l'i-esimo vettore della base canonica di $\mathbb{R}^n$.

- Se $\bar{x}_i = l_i$
$$0 \leq \nabla f(\bar{x}) e_i = \frac{\partial f}{\partial x_i}(\bar{x}), \tag{5.2}$$
ossia $\frac{\partial f}{\partial x_i}(\bar{x}) \geq 0$.
- Se $\bar{x}_i = u_i$, $-\nabla f(\bar{x})^T e_i \geq 0$,
$$0 \geq \nabla f(\bar{x}) e_i = \frac{\partial f}{\partial x_i}(\bar{x}), \tag{5.3}$$
ossia $\frac{\partial f}{\partial x_i}(\bar{x}) \leq 0$.
- Se $l_i < \bar{x}_i < u_i$, valgono contemporaneamente la (5.2) e la (5.3),
$$\nabla f(\bar{x}) e_i = \frac{\partial f}{\partial x_i}(\bar{x}) = 0. \tag{5.4}$$

Ricapitolando, se $\bar{x}$ è ottimale

$$\frac{\partial f}{\partial x_i}(\bar{x}) \begin{cases} = 0 & l_i < \bar{x}_i < u_i \\ \geq 0 & \bar{x}_i = l_i \\ \leq 0 & \bar{x}_i = u_i \end{cases}$$

5.4 Vincoli di Simplesso

Il vincolo è del tipo
$$S = \{x \mid x \geq 0 \text{ e } e^T x = 1\}.$$

Osservazione 5.4.1. Il vincolo di non negatività si può trasformare nella forma dei vincoli già visti con
$$e_h^T x = x_h \geq 0 \quad h = 1, ..., n \quad \Rightarrow \quad -e_h^T x \leq 0 \; h = 1, ..., n. \tag{5.5}$$

$d = e_i - e_j$ è banalmente ammissibile rispetto al vincolo $e^T x = 1$, infatti,
$$e^T d = e^T (e_i - e_j) = e^T e_i - e^T e_j = 1 - 1 = 0.$$

Consideriamo $\bar{x}_j > 0$, allora $e_i - e_j$ è ammissibile, infatti, se $h \neq i, j$,
$$-e_h^T d = -e_h^T (e_i - e_j) = -e_h^T e_i + e_h^T e_j = 0 \leq 0.$$

Se $h = i$,
$$-e_h^T d = -e_i^T (e_i - e_j) = -e_i^T e_i + e_h^T e_j = -1 \leq 0.$$

Infine, se $h = j$ vuol dire che $j \notin I(\bar{x})$ poiché $\bar{x}$ non soddisfa il vincolo con l'uguaglianza, allora
$$e_h^T d \leq 0 \quad \forall h \in I(\bar{x}).$$

Come si comporta $\bar{x}$ soluzione ottimale con tali vincoli? $\nabla f(\bar{x})^T d \geq 0$, ossia
$$\nabla f(\bar{x})^T (e_i - e_j) \geq 0,$$
$$\frac{\partial f}{\partial x_i}(\bar{x}) - \frac{\partial f}{\partial x_j}(\bar{x}) \geq 0,$$
$$\frac{\partial f}{\partial x_i}(\bar{x}) \geq \frac{\partial f}{\partial x_j}(\bar{x}),$$
$\forall j$ tale che $\bar{x}_j \neq 0$ $\forall i$.

Osservazione 5.4.2. Da tale relazione si vede come il valore della derivata sia uguale lungo le componenti realive a vincoli non attivi. Pertanto, se $\bar{x}_j, \bar{x}_l \neq 0$,

$$\frac{\partial f}{\partial x_j}(\bar{x}) = \frac{\partial f}{\partial x_l}(\bar{x}).$$

5.5 Proiezione su Insiemi Convessi

Siamo inteerssati al problrema

$$\min_{y \in S} \frac{1}{2}||x-y||^2, \qquad x \in \mathbb{R}^n, \tag{5.6}$$

con S chiuso e convesso non vuoto. Il problema ammette soluzione poiché la funzione obiettivo è coerciva. La soluzione è unica perché f è strettamente convessa. Detta $p(x)$ la soluzione, essa è detta *proiezione* euclidea di x su S.

$$p(x) = \arg\min_{y \in S} \frac{1}{2}||x-y||^2.$$

Proposizione 5.5.1. *Sia $p : \mathbb{R}^n \to S$ e $p(x)$ proiezione di x su S, valgono le seguenti proprietà*

(i) $p(x)$ è la soluzione del problema (5.6) *se e solo se*

$$(x-p(x))^T(y-p(x)) \leq 0, \quad \forall x, y \in S;$$

(ii) $p(x)$ è un'operazione non espansiva, ossia

$$||p(x)-p(y)|| \leq ||x-y||,$$

inoltre è una funzione continua.

Dimostrazione.

(i) Dato che $p(x)$ è l'unico ottimo del problema, valgono le condizioni sufficienti di ottimalità: se d è ammissibile,

$$\nabla f(p(x))^T d \geq 0 \quad \Leftrightarrow \quad -(x-p(x))^T d \geq 0.$$

Dato che S è convesso, $d = (y-p(x))$ è ammissibile in $p(x)$ $\forall y \in S$.

$$-(x-p(x))^T(y-p(x)) \geq 0 \quad \forall y \in S,$$

$$(x-p(x))^T(y-p(x)) \leq 0 \quad \forall y \in S.$$

(ii) Siano $x, z \in \mathbb{R}^n$ e $p(x), p(z)$ le relative proiezioni, dal punto (i) si ha che

$$(z-p(x))^T(p(z)-p(x)) \leq 0 \qquad (x-p(z))^T(p(x)-p(z)) \leq 0,$$

ossia

$$(z-p(x))^T(p(z)-p(x)) \leq 0 \qquad (p(z)-x)^T(p(z)-p(x)) \leq 0,$$

sommando si ha

$$(z-p(x))^T(p(z)-p(x)) + (p(z)-x)^T(p(z)-p(x)) \leq 0,$$

$$(z-p(x)+p(z)-x)^T(p(z)-p(x)) \leq 0 \;\Rightarrow\; (z-x)^T(p(z)-p(x))+(p(z)-p(x))^T(p(z)-p(x)) \leq 0,$$
$$(z-x)^T(p(z)-p(x)) \leq -||(p(z)-p(x))||^2,$$
$$||p(z)-p(x)||^2 \leq (z-x)^T(p(z)-p(x)) \leq ||z-x|| \cdot ||p(z)-p(x)||,$$
$$||p(z)-p(x)|| \leq ||z-x||.$$

□

La proiezione diventa utile nell'analisi di ottimalità per problemi di ottimizzazione.

Proposizione 5.5.2. *$\bar{x}$ è stazionario se e solo se $\bar{x} = p(\bar{x} - \nabla f(\bar{x}))$.*

Osservazione 5.5.1. Ogni condizione di ottimalità data fino ad ora presupponeva la verifica di una proprietà universale, questa, invece, necessita di un signolo controllo da eseguire iterativamente.

Dimostrazione.

$$\bar{x} = p(\bar{x} - \nabla f(\bar{x})) \Leftrightarrow (\bar{x} - \nabla f(\bar{x}) - \bar{x})^T(x - \bar{x}) \leq 0 \;\; \forall x \in S \Leftrightarrow$$
$$\Leftrightarrow -\nabla f(\bar{x})^T(x - \bar{x}) \leq 0 \;\; \forall x \in S \Leftrightarrow \nabla f(\bar{x})^T(x - \bar{x}) \geq 0 \;\; \forall x \in S.$$

□

Proposizione 5.5.3. *Se f è covessa, allora $\bar{x}$ è ottimo globale se e solo se $\bar{x} = p(\bar{x} - \nabla f(\bar{x}))$.*

In generale la proiezione non è facilmente calcolabile, ma ci sono situazioni in cui questo è fattibile.

Esempio 5.5.1 (Proiezione su insiemi definiti da vincoli di box)**.**

$$\min_{l_i \leq y_i \leq u_i} \frac{1}{2}||x-y||^2 = \min_{l_i \leq y_i \leq u_i} \frac{1}{2}\sum_{i=1}^{n}(x_i - y_i)^2.$$

Possiamo scomporre il problema lungo le i componenti,

$$y_i^* = \arg \min_{l_i \leq y_i \leq u_i} (x_i - y_i)^2.$$

Se $l_i \leq x_i \leq u_i$, allora $y_i = x_i$. Se $x_i < l_i$, allora $y_i = l_i$ e se $x_i > u_i$, $y_i = u_i$.

Con vincoli di Simplesso, la proiezione è calcolabile con un algoritmo che termina in un numero finito di passi (noto a priori).

5.6 Algoritmi di Tipo Linesearch per Problemi Vincolati

La forma del problema è la stessa del caso non vincolato,

$$x^{k+1} = x^k + \alpha_k d_k.$$

Come individuiamo la direzione e il passo? Vogliamo che d_k sia ammissibile e di discesa, il passo lo troviamo col metodo di Armijo con le dovute precauzioni imposte dalle limitazioni dell'insieme ammissibile.

Esempio 5.6.1. Sia $d_k = (z^k - x^k)$ ammissibile, se S è convesso.

$$x^k + \alpha d_k.$$

$x^{k+1} \in S$ se $\alpha \leq 1$.

Osservazione 5.6.1. Dall'Esempio precedente si intuisce come sia una buona scelta partire col metodo di Armijo con $\alpha_0 \leq 1$.

Proposizione 5.6.1. *Sia $\{x^k\}$ la sequenza prodotta da un algoritmo Line-Search con vincolo con d_k ammissibile e di discesa e α_k trovato con Armijo. Allora*

(i) $f(x^{k+1}) < f(x^k)\ \forall k$;

(ii)

$$\lim_{k\to\infty} \nabla f(x^k)^T d_k = 0.$$

La scelta di d_k non ricade automaticamente sull'antigradiente poiché può non essere ammissibile, scegliamo quindi

$$d_k = \hat{x}^k - x^k.$$

Possiamo scegliere $\hat{x}^k$ principalmente in due modi.

5.6.1 Metodo del Gradiente Proiettato

In questo caso

$$\hat{x}^k = p(x^k - \nabla f(x^k)).$$

Se $d_k = \hat{x}^k - x^k$. Vediamo che questa è ammissibile e di discesa: x^k e $\hat{x}^k \in S$ convesso, pertanto è ammissibile in x^k.

$$(x^k - \nabla f(x^k) - \hat{x}^k)^T (x^k - \hat{x}^k) \leq 0,$$
$$(x^k - \hat{x}^k)^T (x^k - \hat{x}^k) - \nabla f(x^k)^T (x^k - \hat{x}^k) \leq 0,$$
$$||x^k - \hat{x}^k||^2 \leq \nabla f(x^k)^T (x^k - \hat{x}^k),$$
$$\nabla f(x^k)^T (\hat{x}^k - x^k) \leq -||x^k - \hat{x}^k||^2.$$

Si possono avere due casi:

- $x^k \neq \hat{x}^k$, $x^k - \hat{x}^k$ è di discesa;
- $x^k = \hat{x}^k$, $x^k = p(x^k - \nabla f(x^k))$ e quindi x^k è stazionario.

Algorithm 9 Metodo del Gradiente Proiettato.

Dati $x^0 \in S$ e $k = 0$
while $x^k - p(x^k - \nabla f(x^k)) > \varepsilon$ **do**
  $d_k = \hat{x}^k - x^k$
  Cacolo α_k lungo d_k con Armijo
  $x^{k+1} = x^k + \alpha_k d_k$
  $k = k + 1$
end while

Osservazione 5.6.2. Nel caso non vincolato, se applico l'algoritmo precedente ritrovo l'altigradiente come direzione di ricerca. Si tratta quindi di una generalizzazione dello stesso metodo.

Proposizione 5.6.2. *Sia S compatto e sia $\{x^k\}$ la sequenza prodotta dal metodo del gradiente proiettato, allora $\{x^k\}$ ha punti di accumulazione ognuno dei quali è stazionario.*

Dimostrazione.
L'esistenza dei punti di accumulazione si ha per costruzione. $x^k \in S\ \forall k$ e S è compatto. Per definizione di compatto, $\{x^k\}$ ammette punti di accumulazione.

Sia ora $K \subset \{0, 1, 2, ...\}$ una tale che

$$\lim_{\substack{k\to\infty \\ k\in K}} x^k = \bar{x} \in S.$$

Consideriamo $\hat{x}^k = p(x^k - \nabla f(x^k))$, per la continuità di p e di ∇f,

$$\lim_{\substack{k\to\infty \\ k\in K}} \hat{x}^k = \lim_{\substack{k\to\infty \\ k\in K}} p(x^k - \nabla f(x^k)) = p(\bar{x} - \nabla f(\bar{x})) = \hat{x}.$$

Visto che abbiamo utilizzato la ricerca di Armijo,

$$d_k = \hat{x}^k - x^k,$$

$$0 = \lim_{\substack{k\to\infty \\ k\in K}} \nabla f(x^k)^T d_k = \lim_{\substack{k\to\infty \\ k\in K}} \nabla f(x^k)^T(\hat{x}^k - x^k) = \nabla f(\bar{x})^T(\hat{x} - \bar{x}).$$

$$(\bar{x} - \nabla f(\bar{x}) - \hat{x})^T(\bar{x} - \hat{x}) \leq 0,$$

$$\nabla f(\bar{x})^T(\hat{x} - \bar{x}) \leq -||\bar{x} - \hat{x}||^2.$$

il membro di sinistra va a 0, pertanto

$$||\bar{x} - \hat{x}||^2 \leq 0,$$

ossia $\bar{x} = \hat{x} = p(\bar{x} - \nabla f(\bar{x}))$. $\bar{x}$ è dunque stazionario. □

5.6.2 Metodo Frank-Wolfe

Sia S compatto e

$$z_k := \min_{x\in S} \nabla f(x^k)^T(x - x^k).$$

Posto

$$\hat{x}^k := \arg\min_{x\in S} \nabla f(x^k)^T(x - x^k), \tag{5.7}$$

$$z^k = 0 \ \Rightarrow \nabla f(x^k)^T(x - x^k) \geq 0\ \forall x \in S,$$

ossia x^k è stazionario. Se infatti,

$$z^k < 0 \ \Rightarrow \nabla f(x^k)^T(\hat{x}^k - x^k) < 0$$

e $d_k = (\hat{x}^k - x^k)$ è di discesa.

Proposizione 5.6.3. *Sia $\{x^k\}$ la sequenza prodotta dal metodo FW, allora $\{x^k\}$ ha punti di accumulazione, ognuno dei quali è stazionario.*

Algorithm 10 Metodo di Frank-Wolfe

Dati $x^0 \in S$ $K = 0$
while TRUE **do**
  Calcolo $\hat{x}^k$ con l'Equazione (5.7)
  if $\nabla f(x^k)^T(\hat{x}^k - x^k) > -\varepsilon$ **then**
    BREAK
  end if
end while

Dimostrazione.
Per costruzione, $\{x^k\} \subset S$ compatto. Quindi ammette punti di accumulazione. Sia $K \subseteq \{0, 1, ...\}$ tale che

$$\lim_{\substack{k\to\infty \\ k\in K}} x^k = \bar{x},$$

$$d_k = \hat{x}^k - x^k,\ \hat{x}^k \in S,$$

$$||d_k|| = ||\hat{x}^k - x^k|| \leq ||\hat{x}^k|| + ||x^k|| \leq M,$$

$\exists M$ data la compattezza (e quindi limitatezza) di S. Allora esiste $K_1 \subseteq K$ tale che

$$\lim_{\substack{k\to\infty \\ k\in K_1}} d_k = \bar{d}.$$

Per le proprietà del metodo di Armijo,

$$0 = \lim_{\substack{k\to\infty \\ k\in K_1}} \nabla f(x^k)^T d_k = \nabla f(\bar{x})^T \bar{d}.$$

Dato $d_k = \hat{x}^k - x^k$, sia $z \in S$ arbitrario, per definizione di $\hat{x}^k$,

$$\nabla f(x^k)^T(\hat{x}^k - x^k) \leq \nabla f(x^k)^T(z - x^k)$$

$$\lim_{\substack{k\to\infty \\ k\in K_1}} \nabla f(x^k)^T d_k \leq \lim_{\substack{k\to\infty \\ k\in K_1}} \nabla f(x^k)^T(z - x^k),$$

$$0 \leq \nabla f(\bar{x})^T(z - x^k),$$

ossia

$$\nabla f(x^k)^T(z - x^k) \geq 0 \forall z \in S,$$

ovvero $\bar{x}$ è stazionario. □

5.7 Problemi con Vincoli in Forma Analitica

$$\min_{\substack{g_i(x)\leq 0\ \ i=1,\dots,m \\ h_i(x)=0\ \ i=1,\dots,p}} f(x). \tag{5.8}$$

Con $f, g_1, ..., g_m, h_1, ..., h_p \in C^1(\mathbb{R}^n)$. Ossia abbiamo m vincoli di disuguaglianza e p vincoli di uguaglianza.

Proposizione 5.7.1 (Condizioni di Fritz-John)**.** *Se x^* è un punto di minimo locale per* (5.8)*, allora esistono dei moltiplicatori $\mu^* \in \mathbb{R}^m$, $\lambda_0 \in \mathbb{R}$, $\lambda^* \in \mathbb{R}^p$ tali che*

(i) $g_i(x^*) \leq 0 \ \forall i = 1, ..., m;$

(ii) $h_i(x^*) = 0 \ \forall i = 1, ..., p;$

(iii) $g_i(x^*)\mu_i^* = 0 \ \forall i = 1, ..., m;$

(iv) $\lambda_0^*, \mu_i^* \geq 0 \ \forall i = 1, ..., m;$

(v) $(\lambda_0^*, \lambda^*, \mu^*) \neq \underline{0};$

(vi)

$$\lambda_0^* \nabla f(x^*) + \sum_{i=1}^{n} \mu_i^* \nabla g_i(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) = 0.$$

Non dimostriamo la Proposizione ma osserviamo che le prime due condizioni sono dette di ammissibilità di x^*, la (iii) è detta di compatibilità e dice che, ove il vincolo non è attivo, il relativo moltiplicatore deve essere nullo, la (iv) e la (v) sono dette di ammissibilità dei moltiplicatori, e ci concentriamo sul significato della (vi).

Osservazione 5.7.1. La funzione lagrangiana applicata al problema è

$$\mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{i=1}^{p} \lambda_i h_i(x) = \lambda_0 f(x) + \mu^T g(x) + \lambda^T h(x).$$

Questa funzione penalizza i vincoli non rispettati. Il suo gradiente rispetto a x è

$$\nabla_x \mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 \nabla f(x) + \sum_{i=1}^{m} \mu_i \nabla g_i(x) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x).$$

Quello che chiediamo nella (vi) è che $\nabla_x \mathcal{L}(x^*, \lambda_0^*, \lambda^*, \mu^*) = 0$.

Osservazione 5.7.2. Tali condizioni tendono a non essere usate, per il motivo che evidenziamo nell'esempio successivo.

Esempio 5.7.1. Cerchiamo la soluzione al problema

$$\min_{y^2=0} f(x, y).$$

Per la Proposizione precedente, se (x^*, y^*) è ottimale, esistono $\lambda_0, \lambda_1 \in \mathbb{R}$ tali che

$$\begin{cases} y^{*2} = 0 \\ \lambda_0 \geq 0 \\ (\lambda_0, \lambda_1) \neq (0, 0) \\ \lambda_0 \nabla f(x^*, y^*) + \lambda_1 \begin{pmatrix} 0 \\ 2y^* \end{pmatrix} = 0 \end{cases}$$

dato che $\nabla h(x) = (0, 2y)^T$.
Se scegliamo $\lambda_0 = 0$ e $\lambda_1 = 1$, le condizioni di Fritz-John sono verificate per ogni punto $(x, 0)$ $\forall x$. Si nota quindi come si sia persa la dipendenza dalla funzione obiettivo f e tutto dipenda solo da S (dai vincoli).

Per fare in modo che resti la dipendenza da f, occorre che λ_0 non sia nullo.

Proposizione 5.7.2 (Condizioni di Karush, Kuhn- Tucker o KKT)**.** *Se x^* è un punto di minimo locale e viene soddisfatta una condizione di regolarità dei vincoli (constrain qualifications) in x^*, allora valgono le condizioni di F-J con $\lambda_0^* \neq 0$, cioè $\exists \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^m$ tali che*

(i) $g_i(x^*) \leq 0 \ \forall i = 1, ..., m;$

(ii) $h_i(x^*) = 0 \ \forall i = 1, ..., p;$

(iii) $g_i(x^*)\mu_i = 0 \ \forall i = 1, ..., m;$

(iv) $\mu_i \geq 0 \ \forall i = 1, ..., m;$

(v)

$$\nabla f(x^*) + \sum_{i=1}^{n} \mu_i \nabla g_i(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) = 0.$$

Osservazione 5.7.3. Esistono varie condizionio di regolarità, alcune più forti di altre.

Osservazione 5.7.4. La (v) può essere riscritta come

$$-\nabla f(x^*) = \sum_{i=1}^{n} \mu_i \nabla g_i(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*),$$

dunque l'antigradiente si può scrivere come combinazione lineare dei gradienti dei vincoli con $\mu_i \geq 0$ ($\mu_i = 0$ se il vincolo g_i non è attivo).

5.7.1 Interpretazione geometrica

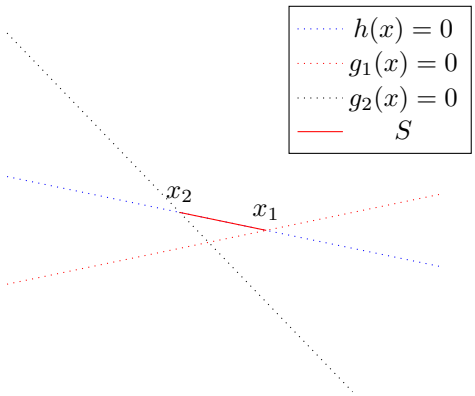

Figura 5.1: Esempio di insieme definito da due vincoli di disuguaglianza e uno di uguaglianza.

$$-\nabla f(x^*) = \sum_{i=1}^{2} \mu_i \nabla g_i(x^*) + \lambda \nabla h(x^*)$$

L'antigradiente è combinazione lineare dei gradienti dei vincoli non attivi (se sono di disuguaglianza). Pertanto l'antigradiente è ortogonale al vincolo di uguaglianza . Nella situazione in

Figura 5.1, se non siamo in x_1 o x_2, l'antigradiente può solo essere ortogonale a $h(x)$. In x_2 $\nabla f(x_1)$ è combinazione lineare di $\nabla h(x_1)$ e $\nabla g_1(x_1)$ con coefficiente $\mu_1 \geq 0$, quindi può stare solo nel semipiano inferiore definito dal vincolo g_1. Analogamente per x_1.

5.7.2 Condizioni di Regolarità dei Vincoli

(i) Tutti i vincoli sono lineari (LCQ);

(ii) I gradienti dei vincoli lineari ($\nabla h_i(x^*)$) e i gradienti $\nabla g_j(x^*)$ $j \in I(x^*)$ (solo quelli attivi) sono linearmente indipendenti (LICQ);

(iii) Condizione di Slatev (SCQ): $f, g_1, ..., g_m, h_1, ..., h_p$ sono tutte convesse ed esiste un punto $\bar{x}$ tale che
$$h_i(\bar{x}) = 0 \ i = 1, ..., m \quad g_i(\bar{x}) < 0 \ i = 1, ..., p.$$

Dimostriamo la seconda

Dimostrazione.
Sia x^* punto di minimo locale, allora sicuramente valgono le FJ. Ossia $\exists \lambda_0, \lambda, \mu$ *non tutti nulli* tali che $\mu_i g_i(x^*) = 0$ $i = 1, ..., m$ e $\mu_i = 0$ $\forall i \notin I(x^*)$ e

$$\lambda_o \nabla f(x^*) + \lambda^T Jh(x^*) + \mu^T Jg(x^*) = 0.$$

Supponiamo per assurdo che $\lambda_0 = 0$,

$$\sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0.$$

Sappiamo che $\mu_i = 0$ se $i \notin I(x^*)$, allora

$$\sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i \in I(x^*)} \mu_i \nabla g_i(x^*) = 0.$$

Per la condizione LICQ, i gradienti in gioco sono però linearmente indipendenti, pertanto i moltiplicatori dovrebbero essere tutti nulli e questo è assurdo. □

5.7.3 Casi Particolari

(i)
$$\min_{\substack{Ax=b \\ Cx \leq d}} f(x).$$

Le KKT sono condizioni necessarie di ottimalità. Se f è convessa sono anche condizioni sufficienti.

(ii) Vincoli di Box
$$l_i \leq x_i \leq u_i \quad i = 1, ..., N.$$

Le portiamo nella forma del problema (5.8)

$$\min_{\substack{l_i - x_i \leq 0 \ \ i=1,...,N \\ x_i - u_i \leq 0 \ \ i=1,...,N}} f(x).$$

Associamo λ_i^+ ai vincoli di upper bound e λ_i^- a quelli di lower bound.
x^* è ottimale, allora

$$\lambda_i^+, \lambda_i^- \geq 0, \quad \lambda_i^+(x_i^* - u_i) = 0 \ \forall i = 1, ..., N, \quad \lambda_i^-(l_i - x_i^*) = 0 \ \forall i = 1, ..., N.$$

$$\mathcal{L}(x, \lambda^+, \lambda^-) = f(x) + \sum_{i=1}^{N} \lambda_i^+(x_i - u_i) + \sum_{i=1}^{N} \lambda_i^-(l_i - x_i),$$

$$\frac{\partial \mathcal{L}}{\partial x_j}(x^*, \lambda^+, \lambda^-) = \frac{\partial f}{\partial x_j}(x^*) + \lambda_j^+ - \lambda_j^-.$$

Si possono verificare tre casi

- $l_i < x_i^* < u_i$, λ_j^+ e λ_j^- devono essere nulli, pertanto $\frac{\partial f}{\partial x_j}(x^*) = 0$.
- $x_i^* = l_j$, $\lambda_j^+ = 0$ e
$$\frac{\partial f}{\partial x_j}(x^*) = \lambda_j^- \geq 0.$$
- $x_j^* = u_j$, allora $\lambda_j^- = 0$ e
$$\frac{\partial f}{\partial x_j}(x^*) = -\lambda_j^+ \leq 0.$$

Allora

$$\frac{\partial f}{\partial x_j} \begin{cases} = 0 & l_j < x_j^* < u_i \\ \leq 0 & x_j^* = u_i \\ \geq 0 & x_j^* = l_i \end{cases}$$

(iii) Vincoli di Simplesso

$$\min_{\substack{e^T x = 1 \\ x_i \geq 0 \ i=1,...,N}} f(x)$$

oppure

$$\min_{\substack{e^T x - 1 = 0 \\ -x_i \leq 0 \ i=1,...,N}} f(x)$$

x^* ottimale, allora esistono λ associato al vincolo di uguaglianza e μ al vincolo di disuguaglianza tali che

$$\mu_i \geq 0, \quad \mu_i x_i = 0 \ i = 1, ..., N,$$

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{N} \mu_i(-x_i) + \lambda(e^T x - 1).$$

$$0 = \frac{\partial \mathcal{L}}{\partial x_j}(x^*, \lambda, \mu) = \frac{\partial f}{\partial x_j}(x^*) - \mu_j + \lambda.$$

Ora, se $x_j^* > 0$ ($x_j^* \neq 0$) $\mu_j^* = 0$ e

$$\frac{\partial f}{\partial x_j}(x^*) = -\lambda \leq \mu_i - \lambda = \frac{\partial f}{\partial x_i}(x^*) \ i = 1, ..., N.$$

Le derivate relative ai vincoli non attivi sono minori o uguali alle altre e uguali a quelle relative ad altri vincoli non attivi.

Capitolo 6

Applicazione all'Apprendimento Automatico

6.1 Regressione Lineare

Si risolve il problema

$$\min_w \frac{1}{2}||Xw - y||^2 + \lambda||w||^2. \tag{6.1}$$

In caso $\lambda \neq 0$ si ha che la soluzione è unica poiché f è coerciva e strettamente convessa[1]. Se $\lambda = 0$ l'esistenza della soluzione è meno ovvia e l'unicità è persa.

Il problema diventa

$$\min_{\substack{z \\ x^T z = 0}} \frac{1}{2}||z - y||^2. \tag{6.2}$$

Questo problema ha la funzione obiettivo strettamente convessa e coerciva. L'insieme S è chiuso, quindi ammette una soluzione z^*. Si tratta di un problema convesso con vincoli lineari, allora le KKT sono condizioni necessarie e sufficienti di ottimalità.

$$\mathcal{L}(z, \lambda) = f(z) + \lambda^T(X^T z) = \frac{1}{2}||z - y||^2 + \lambda^T(X^T z).$$

$$\frac{\partial \mathcal{L}}{\partial z}(z^*) = (z^* - y) + X\lambda = 0, \qquad y = z^* + X\lambda.$$

Per l'ammissibilità di z^*,

$$z^* : \ X^T z^* = 0$$

$$X^T y = X^T(z^* + X\lambda) = X^T z^* + X^T X\lambda = X^T X\lambda,$$

dato che $X^T z^* = 0$. Allora otteniamo le equazioni normali

$$X^T y = X^T X\lambda,$$

di cui l'esistenza della soluzione è garantita dalle KKT. Si ha dunque soluzione al problema di ottimizzazione con

$$f = w^T X^T X w - w^T X^T y, \quad \nabla_w f = X^T X w - X^T y.$$

Con w^* tale che

$$X^T X w^* = X^T y.$$

[1] Come si vede nella 2.3.

6.2 Support Vector Machines

Si tratta di una classe di classificatori lineari su un dataset

$$D = \{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathbb{R}^n \; y^{(i)} \in \{-1, 1\} \; i = 1, ..., N\},$$

6.2.1 Problemi di classificazione

Il dataset è definito da una funzione ignota $\varphi : \mathbb{R}^p \to \{-1, 1\}$ tale che $\varphi(x^{(i)}) = y^{(i)}$. Cerchiamo una funzione $\tilde{h}$ tale che

$$\text{segno}(\tilde{h}(x, \bar{w})) = \varphi(x) \; \forall x.$$

$$\bar{w} \in \arg\min \mathcal{L}(w, D) + \lambda\Omega(w),$$

dove $\mathcal{L}$ è la loss e $\lambda\Omega$ un regolarizzatore. Nel caso dei classificatori lineari cerchiamo un iperpiano che divida in due il dataset e separi in due parti differenti gli esempi con label opposte. La loss è

$$\mathcal{L}(w, D) = \frac{1}{N} \sum_{i=1}^{N} l(w^T x^{(i)} + b, y^{(i)}).$$

La l può essere scelta in vari modi

- La 1-loss è delal forma
$$\mathcal{X}(y^{(i)}(w^T x^{(i)} + b) \leq 0).$$
Questa valuta se si è classificato correttamente l'esempio $x^{(i)}$.
- La *log-loss* è della forma
$$\log\left(1 + e^{-y^{(i)}(w^T x^{(i)} + b)}\right)$$
Questa aggiunge anche una valutazione su quanto sia l'errore.
- La *hinge loss*
$$\max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}.$$
Se la classificazione è corretta questa fa 0. Altriemnti si paga in modo lineare in $y^{(i)}(w^T x^{(i)} + b)$.

Osservazione 6.2.1. $w^T x^{(i)} + b$ è concorde con $y^{(i)}$ se la classificazione è corretta.

Osservazione 6.2.2. La hinge loss è continua ma non differenziabile.

La hinge loss è utilizzata nelle SVM.

6.2.2 SVM

Cerchiamo la funzione $C(x)$ tale che $C(x^{(i)}) = y^{(i)} \; \forall i$ che sia del tipo $C(x) = w^T x + b$.

w^* è la soluzione di un problema di minimo (regolarizzato)

$$(w^*, b^*) \in \arg\min_{w,b} \sum_{i=1}^{N} l(w, b, x^{(i)}, y^{(i)}) + \frac{\lambda}{2}||w||^2.$$

Consideriamo $l(w, b, x^{(i)}, y^{(i)})$ come la funzione *hinge loss*

$$H(w, b, x^{(i)}, y^{(i)}) = \max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}.$$

Osservazione 6.2.3. $y^{(i)}(w^T x^{(i)} + b)$ è una misura di bontà della classificazione. Il prodotto $y^{(i)}(w^T x^{(i)} + b)$ è negativo se la vera label e la classificazione non coincidono ed è invece positivo se coincidono.

Il problema diventa

$$\min_{w,b} \frac{1}{2}||w||^2 + \sum_{i=0}^{N} \max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}. \tag{6.3}$$

Lo riscriviamo come

$$\min_{\substack{w,b,\xi_i \\ \xi_i \geq 1 - y^{(i)}(w^T x^{(i)} + b) \ \forall i \\ \xi_i \geq 0}} \frac{1}{2}||w||^2 + C\sum_{i=0}^{N} \xi_i \tag{6.4}$$

Si vede che, se w^*, b^* sono ottimali per il rpoblema (6.3), allora (w^+, b^*, ξ^*) è soluzione ottimale per il problema (6.4). Infatti,

$$\xi_i^* \geq \max\{0, 1 - y^{(i)}(w^T x^{(i)}) + b\}.$$

Se però si avesse ξ_i^* maggiore di tale massimo, potremmo trovare $\bar{\xi}$ esattamente uguale a tale massimo e questo risulterebbe essere ammissibile e con la funzione obiettivo minore di quanto raggiunto con ξ_i^*, dato che la parte dipendenmte da w^* e b^* resterebbe inalterata. Questo è assurdo.

Con tale formulazione, la funzione è quadratica, convessa e con vincoli lineari.

Solitamente consideriamo il caso $C = +\infty$, allora $\xi_i = 0 \ \forall i$ e chiedere $1 - y^{(i)}(w^T x^{(i)} + b) \leq 0$, pertanto

$$y^{(i)}(w^T x^{(i)} + b) \geq 1,$$

significa chiedere che sia classificato perfettamente il dataset. Così facendo corriamo il rischio di fare overfitting. Il problema, inoltre, può non ammettere soluzione se le due classi non sono linearmente separabili.

Proposizione 6.2.1. *Se consideriamo il problema*

$$\min_{\substack{x \\ g(x) \leq 0}} f(x),$$

con $f, g \in C^1(\mathbb{R}^n)$ e convesse. x^* soluzione e assumiamo che μ^* sia un vettore di moltiplicatori tali che (x^*, μ^*) soddisfi le KKT, allora (x^*, μ^*) è soluzione ottimale del problema duale di Wolfe

$$\max_{\substack{x,\mu \\ \mu \geq 0 \\ \nabla_x \mathcal{L}(x,\mu) = 0}} \mathcal{L}(x, \mu) = \max_{\substack{x,\mu \\ \mu \geq 0 \\ \nabla_x \mathcal{L}(x,\mu) = 0}} f(x) + \mu^T g(x).$$

Dimostrazione.
Le KKT consistono in

$$\begin{aligned} \mu^* &\geq 0 \\ g(x^*) &\leq 0 \\ \mu_i^* g_i(x^*) &= 0 \\ \nabla_x \mathcal{L}(x^*, \mu^*) &= 0 \end{aligned}$$

$$\mathcal{L}(x^* \mu^*) = f(x^*) + \sum_i \mu_i^* g_i(x^*) = f(x^*),$$

dato che $\mu_i g_i(x^*) = 0$ per ogni i. Sia (x, μ) una generica soluzione ammissibile per il duale, dato che $g_i(x^*) \leq 0$ e $\mu_i \geq 0$,

$$\mathcal{L}(x^*, \mu^*) = f(x^*) \geq f(x^*) + \sum_i \mu_i g_i(x^*) = \mathcal{L}(x^*, \mu).$$

$\mathcal{L}(x, \mu)$ è convessa rispetto a x poiché lo è f e la somma di funzioni convesse con coefficienti non negativi è convessa. Allora

$$\mathcal{L}(x^*, \mu) \geq \mathcal{L}(x, \mu) + \nabla_x \mathcal{L}(x, \mu)(x^* - x),$$

ma $\nabla_x \mathcal{L}(x, \mu) = 0$ poiché è ottimale per il duale, allora

$$\mathcal{L}(x^*), \mu^* \geq \mathcal{L}(x^*, \mu) \geq \mathcal{L}(x, \mu).$$

La tesi segue dall'arbitrarietà di (x, μ). □

Il problema (6.4) è quindi convesso con vincoli lineari e le KKT sono condizioni necessarie e sufficienti di ottimalità. Se (w^*, b^*, ξ^*) è ottimale, allora eisstono $\alpha^* \in \mathbb{R}^n$ e $\mu^* \in \mathbb{R}^n$ tali che $(w^+, b^*, \xi^*, \alpha^*, \mu^*)$ soddisfano le KKT. Come è fatto il problema duale per tale problema

$$\begin{array}{c} \max \\ {\scriptstyle \alpha \geq 0} \\ {\scriptstyle \mu \geq 0} \\ {\scriptstyle \nabla_\alpha \mathcal{L}(w,b,\xi,\alpha,\mu)=0} \\ {\scriptstyle \nabla_b \mathcal{L}(w,b,\xi,\alpha,\mu)=0} \\ {\scriptstyle \nabla_\xi \mathcal{L}(w,b,\xi,\alpha,\mu)=0} \end{array} \quad \mathcal{L}(w, b, \xi, \alpha, \mu),$$

e

$$\mathcal{L} = \frac{1}{2}||w||^2 + C\sum_i \xi_i + \sum_i \mu_i(-\xi_i) + \sum_i \alpha_i(1 - y^{(i)}(w^T x^{(i)} + b) - \xi_i)$$

dove il vincolo su ξ è stato posto in forma standard ($-\xi_i \leq 0$).

Approfondiamo i vincoli

$$0 = \nabla_w \mathcal{L}(w, b, \xi, \alpha, \mu) = w - \sum_i \alpha_i y^{(i)x^{(i)}},$$

$$w = \sum_i \alpha_i y^{(i)x^{(i)}}.$$

$$0 = \nabla_b \mathcal{L}(w, b, \xi, \alpha, \mu) = -\sum_i \alpha_i y^{(i)}, , \qquad \alpha^T y = 0.$$

$$0 = \nabla_\xi \mathcal{L}(w, b, \xi, \alpha, \mu) = \sum_i C e_i - \sum_i \mu_i e_i - \sum_i \alpha_i e_i, \qquad C - \mu_i - \alpha_i = 0 \ \ \forall i \qquad \alpha_i = C - \mu_i \ \ \forall i.$$

Ora la lagrangiana

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2}||w||^2 + C\sum_i \xi + \sum_i -\mu_i \xi + \sum_i \alpha_i - \sum_i \alpha_i y^{(i)} w^T x^{(i)} - b\sum_i \alpha_i y^{(i)} - \sum_i \alpha_i \xi_i$$

Il termine $\sum_i \alpha_i y^{(i)} = 0$. Dall'ultimo vincolo

$$\sum_i \xi(C - \mu_i - \alpha_i) = 0,$$

allora la lagrangiana diventa

$$\mathcal{L}(w,\alpha,\mu) = \frac{1}{2}||w||^2 + \sum_i \alpha_i - \sum_i \alpha_i y^{(i)} w^T x^{(i)}.$$

Possiamo scrivere w in funzione di α,

$$\mathcal{L}(w,\alpha) = \frac{1}{2}w^t w + \sum_i \alpha_i - \sum_i \alpha_i y^{(i)} w^T x^{(i)} = \frac{1}{2}w^t w + \sum_i \alpha_i - w^T \sum_i \alpha_i y^{(i)} x^{(i)} =$$

$$= \sum_i \alpha_i + w^T \big(\frac{1}{2}w - \sum_i \alpha_i y^{(i)} x^{(i)},\big)$$

$$w = \sum_i \alpha_i y^{(i)},$$

allora

$$\mathcal{L}(w,\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} = e^T\alpha - \frac{1}{2}Q^T Q\alpha$$

con $Q_{ij} = y^{(i)} y^{(j)} x^{(i)} x^{(j)}$. Il problema diventa

$$\max_{\substack{\alpha\geq 0 \\ \alpha^T y=0 \\ \alpha_i = C-\mu_i \\ \mu_i \geq 0}} \; e^T\alpha - \frac{1}{2}\alpha Q^T Q\alpha$$

ossia

$$\max_{\substack{\alpha\geq 0 \\ \alpha^T y=0 \\ \alpha_i \leq C}} e^T\alpha - \frac{1}{2}\alpha Q^T Q\alpha \qquad \max_{\substack{\alpha^T y=0 \\ 0\leq\alpha_i\leq C}} e^T\alpha - \frac{1}{2}\alpha Q^T Q\alpha.$$

Il duale è solo funzione dei moltiplicatori α. Possiamo calcolare α^+ come soluzione di

$$\max_{\substack{\alpha \\ 0\leq\alpha\leq C \\ \alpha^T y=0}} -\frac{1}{2}\alpha^T Q\alpha + e^T\alpha \qquad \min_{\substack{\alpha \\ 0\leq\alpha\leq C \\ \alpha^T y=0}} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha.$$

Osservazione 6.2.4. Il problema è quadrato, strettamente convesso e con vincoli lineari.

Una volta trovato α^* , possiamo superare le altre quantità con

$$w^* = \sum_i \alpha_i^* y^{(i)} x^{(i)}$$

$$\xi_i^* = \max\big\{0, 1 - y^{(i)}({w^*}^T x^{(i)} + b^*)\big\}$$

$$\mu_i^* = C - \alpha_i^*$$

$$b^* = \frac{y^{(i)}}{{w^*}^T x^{(i)}} \;\forall i \text{ tc } 0 < \alpha_i^* < C.$$

Osservazione 6.2.5.

$$\alpha_i^* \big(1 - y^{(i)({w^*}^T x^{(i)} + b^*)} - \xi_i^*\big) = 0,$$

$$\mu_i^* \xi_i^* = (C - \alpha_i^*)\xi_i^* = 0,$$

allora se $\alpha_i^* \in (0, C) \to \xi_i^* = 0$, $y^{(i)}({w^*}^T x^{(i)} + b^*) = 1$. Ossia, i punti stanno esattamente sulla superficie di separazione.

Se $\alpha_i^* < C$, allora $\xi_i^* = 0$.

Il problema duale è comodo da ottimizzare poiché i vincoli sono più semplici da trattare. L'elemento Q_{ij} è un prodotto scalare, la matrice Q è quindi una matrice di similarità. $u^T v$ si può infatti considerare una misura di similarità.

Possiamo però usare misure differenti. Le misure sono dette Kermel,

$$k: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}.$$

$$Q_{ij} = y_i y_j k(x_i, x_j) \qquad h(x) = \sum_{i=1}^{n} \alpha_i y^{(i)} k(x^{(i)}, x) + b.$$

Quali funzioni Kermel sono utilizzabili? Un Kermel si dice valido se $\forall D$ dataset, la matrice

$$Q_{ij} = y^{(i)} y^{(j)} k(x_i, x_j),$$

è semi-definita positiva, ossia se e solo se k rappresenta un prodotto scalare in uno spazio di dimensione maggiore.

k è valido, dati $u, v \in \mathbb{R}^n$ $\exists \phi : \mathbb{R}^n \to \mathbb{R}^m$ con $n \geq m$ tali che

$$u, v \in \mathbb{R}^n \xrightarrow{\phi} \phi(u), \phi(v) \in \mathbb{R}^n, \qquad \phi(u)^T \phi(v) = k(u, v).$$

Osservazione 6.2.6. Risulta importante sapere che k è valido ma non è necessario sapere come è fatto.

Utilizzare un kernel permette di costruire classificatori non lineari poiché quello che è lineare in $k(\mathbb{R}^n, \mathbb{R}^n)$ non lo è necessariamente in $\mathbb{R}^m \times \mathbb{R}^m$.

Esempio 6.2.1. $k(u, v) = e^{-n||u \cdot v||^2}$, è detto RBF.

6.3 Metodi di Decomposizione

$$\min_{x \in X \subseteq \mathbb{R}^n} f(x),$$

può essere visto come

$$\min_{\substack{x_1 \in X_1 \subseteq \mathbb{R}^{n_1} \\ \vdots \\ x_m \in X_m \subseteq \mathbb{R}^{n_m}}} f(x_1, ..., x_m).$$

Abbiamo cioè diviso le variabili in m gruppi. Utilizziamo quindi algoritmi che si occupino di risolvere dei sottoproblemi della forma

$$\min_{x_i \in X^{n_i}} f(\bar{x}_1, ..., x_i, ..., \bar{x}_m).$$

Osservazione 6.3.1. Questo processo risulta utile se il numero delle variabili è grande e se c'è possibilità di parallelizzare il calcolo. Inoltre, se il problema complesto è difficile, talvolta i sottoproblemi sono facili.

6.3.1 Metodi di Decomposizione Sequenziali

Il più semplice dei metodi di questo tipo è Gauss-Seidel.

$$x_i^{k+1} = \arg\min_{\xi \in X_i} f(x_1^{k+1}, ..., x_{i-1}^{k+1}, \xi_i, x_{i+1}^k, ..., x_m^k).$$

Questo metodo ha proprietà di convergenza a punti stazionari se f èconvessa o se f è strettamente convessa per componenti rispetto ad ogni blocco. Un caso particolarmente buono è $m = 2$.

Osservazione 6.3.2. Trovare l'argmin di un sottoproblema può non essere banale.

6.3.2 Metodi di Decomposizione Paralleli

L'algoritmo di riferimento è quello di Jacobi.

$$\tilde{x}_i^{k+1} = \arg\min_{\xi_i \in X_i} f(x_1^k, ..., x_{i-1}^k, \xi_i, x_{i+1}^k, x_m^k),$$

$\hat{x}_i^{k+1} \in \mathbb{R}^n$ è dato da

$$\hat{x}_i^{k+1} = (x_1^k, ..., x_{i-1}^k, \tilde{x}_i^{k+1}, x_{i+1}^k, ..., x_m^k).$$

$$x^{k+1} = \arg\min_{x=\hat{x}_i^{k+1}} \min_{i=1,...,m} f(x).$$

6.3.3 Schemi con Blocchi Sovrapposti

$$\tilde{x}_w^{k+1} = \arg\min_{X_{w_k}} f(x_{w_k}, x_{\bar{w}_k}^k).$$

$$x^{k+1} = \begin{cases} \tilde{x}^{k+1} & i \in w_k \\ x_i^k \end{cases}$$

La convergenza dello schema dipende dalla regola di selezione di w_k. Due regole che funzionano sono la cilcica

$$\exists M > 0 \text{ tc } \forall k, \forall i \ \exists l(k) leq M \text{ tc } i \in w_{k+l(k)}.$$

Ogni Mk iterazioni ho preso tutte le variabili.

Un'altra regola è Gauss-Southwell,

$$\forall k \ \exists i(k) \in w_k \text{ tc } \left| \frac{\partial f(x^k)}{\partial x_{i(k)}} \right| \geq \frac{\partial f(x^k)}{\partial x_h} \qquad h = 1, ..., n.$$

Ossia si inserisce in w_k la variabile a massima discesa.

6.4 Algoritmo di Decomposizione per SVM

$$\min_{\substack{\alpha \\ \alpha^T y = 0 \\ 0 \leq \alpha \leq C}} \frac{1}{2} \alpha^T Q \alpha.$$

La complessità del problema risiede nei seguenti punti

- il vincolo lineare di uguaglianza non è separabile.
- Q è tipicamente densa ($Q_{ij} = 0$ se $x_i^T x_j = 0$ o quasi) ed ha dimensione $N \times N$ con N numero di esempi.

Usando un working set w e il suo complementare $\bar{w}$, il sottoproblema all'iterazione k è

$$\min_{\substack{\alpha_w \\ 0 \leq \alpha_w \leq C \\ \alpha_w^T y_w = -\alpha_{\bar{w}}^{k\,T} y_{\bar{w}}}} \frac{1}{2} \alpha_w Q_{ww} \alpha_w$$

Il secondo vincolo deriva dal fatto che, se $\alpha^T y = 0$,

$$\alpha^T y = 0 \Rightarrow \alpha_w^T y_w + \alpha_{\bar{w}}^T y_{\bar{w}} = 0 \Rightarrow \alpha_w^T y_w = -\alpha_{\bar{w}}^T y_{\bar{w}}$$

Quante variabili devo scegliere nel working set? Se scelgo una sola varibaile $w = \{i\}$,

$$\alpha_i y_i = -\sum_{j \neq i} \alpha_j^k y_j = \bar{b}, \;\; \Rightarrow \alpha_i = \frac{\bar{b}}{y_i},$$

e questa si pione uguale a α_i^k. Questa scelta dà sempre questo risultato, pertanto perdiamo libertà. Se si scelgono due (o più) variabili tornaiamo ad avere problemi sensati. Nel caso di due variabili, i sottoproblemi sono risolvibili in forma analitica. Con più di due variabili occorre un solver per i sottoproblemi.

$$\min_{\substack{\alpha_i,\alpha_j \\ 0\leq\alpha_i,\alpha_j\leq C \\ \alpha_i y^i+\alpha_j y^j=-\sum_{h\neq i,j}\alpha_h^k y^h}} \begin{pmatrix}\alpha_i & \alpha_j\end{pmatrix} \begin{pmatrix}Q_{ii} & Q_{ij} \\ Q_{ji} & Q_{jj}\end{pmatrix} \begin{pmatrix}\alpha_i \\ \alpha_j\end{pmatrix} - \alpha_i - \alpha_j + p^T \begin{pmatrix}\alpha_i \\ \alpha_j\end{pmatrix}$$

Risolvendo il secondo vincolo e sostituendo si ha un problema di minimo in una variabile su una parabola ristretto ad un intervallo. I metodi così costruti sono detti *sequential-minimal-optimization* (SMO). Resta da definire come scegliere le due variabili del working set.

Siamo, come al solito, interessati ad una direzione d di discesa, ammissibile e con solo due componenti diverse da 0.

$$d^{ij} = \begin{pmatrix}0 & \cdots & 0 & d_i & 0 & \cdots & 0 & d_j & 0 & \cdots 0\end{pmatrix}.$$

Il set delle direzioni ammissibili in $\bar{\alpha}$ è

$$D(\bar{\alpha}) = \{d \in \mathbb{R}^n \mid d^T y = 0 \text{ e } d_i \geq 0 \text{ se } i \in l(\bar{\alpha}) \text{ e } d_i \leq 0 \text{ se } i \in u(\bar{\alpha})\},$$

dove abbiamo i9ndicato con $l(\bar{\alpha})$ l'insieme degli indici j per cui $\bar{\alpha}_j = 0$ e con $u(\bar{\alpha})$ quelli per cui $\bar{\alpha}_j = C$.

Ora

$$d^{ij\,T} y = 0 \;\Rightarrow\; y_i d_i + y_j d_j = 0,$$
$$d_i = \frac{1}{y_i}, \qquad d_j = \frac{1}{y_j}$$

può costituire una buona scelta. Questa non è però sempre una direzione ammissibile. Questo vale se $\bar{\alpha}_i \in (0, C)$. Se $i \in l(\bar{\alpha})$, la scelta può funzionare se e solo se $y_i = 1$ e se $i \in u(\bar{\alpha})$ se e solo se $y_i = -1$.

Un ragionamento analogo può essere fatto su j invertendo i segni di y_j. Da ora in avanti consideriamo

$$l(\bar{\alpha}) = l^+(\bar{\alpha}) \cup l^\cdot(\bar{\alpha}) = \{i \mid \bar{\alpha}_i = 0,\; y^{(i)} = 1\} \cup \{i \mid \bar{\alpha}_i = 0,\; y^{(i)} = -1\}.$$

$$u(\bar{\alpha}) = u^+(\bar{\alpha}) \cup u^\cdot(\bar{\alpha}) = \{i \mid \bar{\alpha}_i = C,\; y^{(i)} = 1\} \cup \{i \mid \bar{\alpha}_i = C,\; y^{(i)} = -1\}.$$

Proposizione 6.4.1. *Sia d^{ij} fatta come sopra, essa è ammissibile se e solo se*

$$i \in R(\bar{\alpha} :) = l^+(\bar{\alpha}) \cup u^-(\bar{\alpha}) \cup \{i \mid \bar{\alpha}_i \in (0, C)\},$$

$$j \in S(\bar{\alpha} :) = l^-(\bar{\alpha}) \cup u^+(\bar{\alpha}) \cup \{i \mid \bar{\alpha}_i \in (0, C)\}.$$

Proposizione 6.4.2. *d^{ij} come sopra è di discesa in $\bar{\alpha}$ se e solo se*

$$\frac{\nabla_i f(\bar{\alpha})}{y_i} < \frac{\nabla_j f(\bar{\alpha})}{y_j}. \tag{6.5}$$

Dimostrazione.
f è convessa e differenziabile (è quadratica). $d \in \mathbb{R}^n$ è di discesa in $\bar{\alpha}$ se e solo se $\nabla f(\bar{\alpha})^T d < 0$, ossia

$$d_i \nabla_i f(\bar{\alpha}) + d_j \nabla_j f(\bar{\alpha}) < 0,$$
$$\frac{\nabla_i f(\bar{\alpha})}{y_i} - \frac{\nabla_j f(\bar{\alpha})}{y_j} < 0$$
$$\frac{\nabla_i f(\bar{\alpha})}{y_i} < \frac{\nabla_j f(\bar{\alpha})}{y_j}.$$

□

Algorithm 11 SMO

Dati Q, y
$k = 0, \alpha^0 = 0$
$\nabla f(\alpha^0) = Q\alpha^0 - e = -e$
while Criterio di arresto non soddisfatto **do**
scegliamo $i \in R(\alpha^k)$ e $j \in S(\alpha^k)$ tali che vaòga la (6.5)
$w_k = \{i, j\}$
$\hat{\alpha}_{w_k} = \arg\ min_{w_k} f(\alpha_{w_k}, \alpha^k_{\bar{w}_k})$
$\alpha_h^{k+1} = \hat{\alpha}_i$ se $h = i$, $\alpha_h^{k+1} = \hat{\alpha}_j$ se $h = j$, $\alpha_h^{k+1} = \alpha_h^k$ altrimenti
$\nabla f(\alpha^{k+1}) = Q\alpha^{k+1} - e = Q(\alpha^{k+1} - \alpha^k) + Q\alpha^k - e =$
$= Q(\alpha^{k+1} - \alpha^k) + \nabla f(\alpha^k) = Q_i(\alpha_i^{k+1} - \alpha_i^k) + Q_j(\alpha_i^{k+1} - \alpha_i^k) + \nabla f(\alpha^k)$
end while

Osservazione 6.4.1. Come si nota nell'Algoritmo 11, l'aggiornamento del gradiente richiede solo Q_i e Q_j

Proposizione 6.4.3. *Sia α^+ ottimo globale per il problema duale di SVM, allora*

$$\max_{h \in R(\alpha^*)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\} \leq \min_{h \in S(\alpha^*)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\}. \tag{6.6}$$

Se α^* con è ottimale, $\exists i \in R(\alpha^*)$ e $j \in S(\alpha^*)$ tali che

$$\frac{-\nabla_i f(\alpha^*)}{y_i} > \frac{-\nabla_j f(\alpha^+)}{y_j},$$

ossia esiste una direzione di discesa. Dettii i^* e j^* che realizzano rispettivamente il massimo e il minimo della (6.6), (i^*, j^*) è detta *most violating point.*

Proposizione 6.4.4. *Sia α^+ prodotto da SMO e w_k la most violating point, allora*

$$\lim_{k \to \infty} \alpha^k = \alpha^*,$$

e $\forall \varepsilon > 0$, in un numero finito di iterazioni,

$$\max_{h \in R(\alpha^*)} \left\{ \frac{-\nabla_h f(\alpha^k)}{y_h} \right\} \leq \min_{h \in S(\alpha^*)} \left\{ \frac{-\nabla_h f(\alpha^k)}{y_h} \right\} - \varepsilon.$$

6.5 Metodi Stocastici per Problemi di Somme Finite

Il problema è

$$\min_{x\in\mathbb{R}^N} f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x).$$

f è una media di N termini. Questo problema è detto di tipo *Finite Sum*. I metodi *Stocastici* sono algoritmi che effettuano iterazioni del tipo

$$x^{k+1} = x^k + \alpha d_k, \qquad d_k = -\frac{1}{|B|}\sum_{i\in B}\nabla f_i(x), \qquad B \subseteq \{1, ..., N\}.$$

Per la linearità del gradiente

$$\nabla f = \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x),$$

ma quello che calcoliamo noi è fatto su un sottoinsieme. Il passo α viene scelto tipicamente costante o segue una sequenza predefinita $\{\alpha_k\}$.

Osservazione 6.5.1. Il vantaggio di usare l'approssimazione

$$\nabla f(x) \approx \frac{1}{|B|}\sum_{i\in B}\nabla f_i(x),$$

è che ci permette di non usare la loss totale ma solo di alcuni elementi. B è chiamato *mini batch*. L'insieme intero è detto *full batch*. α è detto *learning rate* in ambito machine learning.

6.6 Metodo del Gradiente Stocastico

Si sceglie $i_k \in \{1, ..., N\}$ con probabilità uniforme e si setta

$$x^{k+1} = x^k - \alpha\nabla f_{i_k}(x^k).$$

$$E[\nabla f_{i_k}(x^k)] = \sum_{i=1}^{N}\nabla f_i(x^k)\cdot p(i_k) = \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x^k) = \nabla f(x^k).$$

In media quindi sappiamo che otteniamo una direzione di discesa.

Lavoriamo sotto l'ipotesi che esista $G > 0$ tale che

$$||\nabla f_i|| \leq G, \quad \forall i, x,$$

che f sia limitata inferiormente e $C^2(\mathbb{R}^n)$ con ∇f Lipshitz-continuo e che f^* sia un valore finito e ottimale, vale la seguente

Proposizione 6.6.1. *Sia $\{x^k\}$ la sequenza prodotta da SGD con passi $\{\alpha_k\}$ che soddisfano*

$$\sum_{k=0}^{+\infty}\alpha_k = \infty, \qquad \sum_{k=0}^{+\infty}(\alpha_k)^2 < +\infty.$$

Assumiamo inoltre che, ad ogni iterazione k, l'algoritmo restituisca in output una soluzione $z^k = x^\tau$ con probabilità

$$\mathbb{P}(\tau = t) = \frac{\alpha_t}{\sum_{i=0}^{k-1}\alpha_i}, \quad \forall t = 0, ..., k-1.$$

Allora

$$\lim_{k\to\infty} E\big[||\nabla f(z^k)||^2\big] = 0.$$

Dimostrazione.
Per il Teorema di Taylor $\exists y_k$ tale che

$$f(x^{k+1}) = f\big(x^k - \alpha_k \nabla f_{i_k}(x^k)\big) = f(x^k) - \alpha_k \nabla f_{i_k}(x^k)^T \nabla f(x^k) + \frac{\alpha_k^2}{2} \nabla f(x^k)^T \nabla^2 f(y_k) \nabla f_{i_k}(x^k) \leq$$

$$\leq f(x^k) - \alpha_k \nabla f_{i_k}(x^k)^T \nabla f(x^k) + \frac{\alpha_k^2}{2} L ||\nabla f_{i_k}(x^k)||^2 \leq$$

$$\leq f(x^k) - \alpha - k \nabla f_{i_k}(x^k)^T \nabla f(x^k) + \frac{\alpha_k^2}{2} L G^2.$$

Vediamone il valore atteso

$$E\big[f(x^{k+1})\big] \leq E\big[f(x^k)\big] - E\big[\alpha_k \nabla f_{i_k}(x^k)^T \nabla f(x^k)\big] + E\big[\frac{\alpha_k^2 G^2 L}{2}\big] =$$

$$= E\big[f(x^k)\big] - \alpha_k E\big[\nabla f_{i_k}(x^k)^T \nabla f(x^k)\big] + \frac{\alpha_k^2 G^2 L}{2} = (*)$$

Ora,

$$E\big[\nabla f_{i_k}(x^k) | x^k\big] = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x^k) = \nabla f(x^k),$$

allora

$$(*) = E\big[f(x^k)\big] - \alpha_k E\big[||\nabla f(x^k)||^2\big] + \frac{\alpha_k^2 G^2 L}{2}.$$

Se applichiamo questa maggioraziobne ricorsivamente, otteniamo

$$E\big[f(x^{k+1})\big] - E\big[f(x^0)\big] \leq -\sum_{t=0}^{k} \alpha_t E\big[||\nabla f(x^t)||^2\big] + \frac{G^2 L}{2} \sum_{t=0}^{k} \alpha_t^2.$$

$$E\big[f(x^{k+1})\big] \leq f(x^0) - \sum_{t=0}^{k} \alpha_t E\big[||\nabla f(x^t)||^2\big] + \frac{G^2 L}{2} \sum_{t=0}^{k} \alpha_t^2.$$

$$\sum_{t=0}^{k} \alpha_t E\big[||\nabla f(x^t)||^2\big] \leq f(x^0) - E\big[f(x^{k+1})\big] + \frac{G^2 L}{2} \sum_{t=0}^{k} \alpha_t^2 \leq$$

$$\leq f(x^0) + \frac{G^2 L}{2} \sum_{t=0}^{k} \alpha_t^2 - f^*,$$

poiché $E\big[f(x^{k+1})\big] \geq f^*$.

$$E\big[||\nabla f(z^{k+1})||^2\big] = \sum_{t=0}^{k} E\big[||\nabla f(x^t)||^2\big] \mathbb{P}(z^{k+1} = x^t) =$$

$$\sum_{t=0}^{k} E\big[||\nabla f(x^t)||^2\big] \frac{\alpha_t}{\sum_{i=0}^{k} \alpha_i} = \frac{1}{\sum_{i=0}^{k} \alpha_i} \sum_{t=0}^{k} \alpha_t E\big[||\nabla f(x^k)||^2\big].$$

Allora,

$$E\big[||\nabla f(z^{k+1})||^2\big] \leq \frac{1}{\sum_{i=0}^{k} \alpha_i} \left(f(x^0) + \frac{G^2 L}{2} \sum_{t=0}^{k} \alpha_t^2 - f^* \right).$$

Ma $\sum_i \alpha_i \to +\infty$, $f(x^0), f^*$ finiti e $\sum_i \alpha_t^2 < \infty$. Se passiamo al limite,

$$\lim_{k\to\infty} E\big[||\nabla f(z^{k+1})||^2\big] \leq 0.$$

□

Osservazione 6.6.1. Il passo deve andare a 0 ma non troppo velcoemente, ad esempio $\left\{\frac{\alpha_0}{k+1}\right\}$ soddisfa le condizioni.

6.6.1 Complessità

Metodo	Convex	Strong Convex
GD	$\theta(\varepsilon)$	$O\big(N\log(1/\varepsilon)\big)$
SGD	$\theta(1/\varepsilon^2)$	$\theta(1/\varepsilon)$

Osservazione 6.6.2. Se N è grande, il vantaggio di GD si ha in corrispondenza di ε molto piccoli.

6.6.2 Addestramento di Reti Neurali

Il problema che affrontiamo è

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(w) = \sum_{i=1}^{N} l_i(w).$$

$\mathcal{L}$ è "fortissimamente non convessa" ed è un obiettivo surrogato, ossia la minimizzazione avviene sui dati a disposizione ma vorremmo che valesse su tutti.
par Spesso non siamo interessati ad un ottimo globale ma a un buon ottimo locale. Si è infatti osservato che le soluzioni buone sono quelle ampie di modo che che, in caso di perturbazioni, esse non si ripercuotano troppo sulla rete. Una particolarità delle reti è che il costo di $\nabla\mathcal{L}(w)$

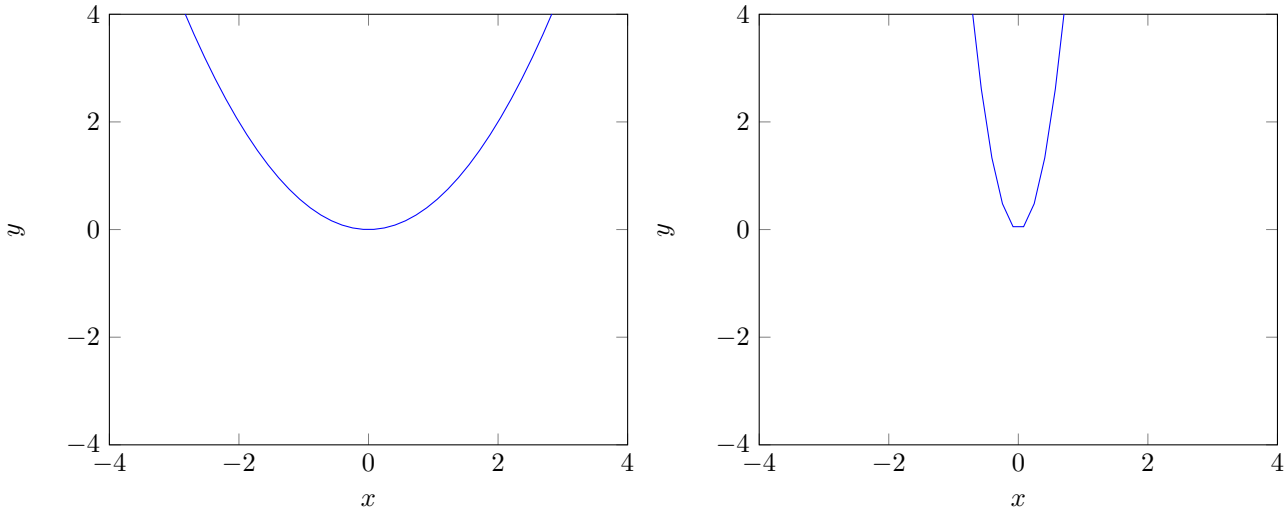

Figura 6.1: Minimo locale ampio e minimo globale stretto.

è circa il doppio di $\mathcal{L}(w)$ invece che circa le N volte della maggior parte dei casi. Questo è

dovuto all'algoritmo di *Back Propagation* per il caloclo del gradiente. Tuttavia il calcolo di $\mathcal{L}$ è particolarmente costoso.

I vantaggi di utilizzare SGD sono

- I dati di training presentano spesso ridondanza, pertanto non occorrono tutti gli esempi per migliorare la classificazione.
- Esperienza computazionale: se si sceglie α in modo oppportuno, si vede cge si arriva presto ad avere una buona precisione.
- La complessitò è $O(1/\varepsilon)$ indipendentemente da N.
- Per la natura di SGD, è impossibile triovare minimi *sharp* (stretti, legati a fenomeni di overfitting).

I vantaggi di GD sono

- Se si usa tutto il gradiente possiamo creare algoritmi davvero buoni.
- Si tratta di un algoritmo parallelizzabile.
- Consente di raggiungere un'ottima precisione.

La soluzione comunemente adottata dai solver è detta *Mini Batch SGD*, ossia

$$1 < |B| << N.$$

Algorithm 12 Mini Batch SGD

Dati w^0, $\{\alpha^k\}$ predefinita, $K = 0$
while Condizione di arresto **do**
  $w_0^k = w^k$
  Dividiamo $\{1, ..., N\}$ in N/M blocchi $B_0, ..., B_{N/M-1}$ di dimensione M
  for $t = 0, ..., N/M$ **do**
    $w_{t+1}^k = w_t^k - \alpha_k \frac{1}{M} \sum_{i \in B_t} \nabla f_i(w_t^k)$
  end for
  $w^{k+1} = w_{N/M}^k$
  $k = k + 1$
end while

Una singola iterazione del *while* nell'Algoritmo 12 è detta *epoca*.
Esistono molte varianti dell'algoritmo.

- Accelerazione, possiamo aggiungere un termine di tipo Momentum/Nesterov,

$$w^{k+1} = w^k - \alpha_k \nabla f(w^k) + \beta_k (w^k - w^{k-1}),$$

$$w^{k+1} = w^k - \alpha_k \nabla f\big(w^k + \beta_k (w^k - w^{k-1})\big) + \beta_k (w^k - w^{k-1}).$$

  Questo può risultare utile poiché il termine di memoria riduce lo "zig zag" dovuto alla stocasticità. Inoltre, l'informazione dei termini di tipo Momentum è a basso costo.

- Tecniche di *adaptive-learning rates*

$$x_i^{k+1} = x_i^k - \alpha_k \nabla_i f(x^k),$$

dove α_k^i dipende dall'iterazione ma anche dalla componente. Lo step-size viene aggiornato in base all'andamento del processo di ottimizzazione. Si hanno quindi n step-size associati alle n variabili e aggiornati indipendentemente.

$$x^{k+1} = x^k - \begin{pmatrix} \nabla_1 f(x^k)\alpha_k^1 \\ \vdots \\ \nabla_n f(x^k)\alpha_k^n \end{pmatrix}$$

Questo è utile poiché i gradienti potrebbero essere diversi componente per componente. α non può quindi essere uniformato. Esistono vari metodi per gestire questi passi.

Ado Gand

$$S_i^{k+1} = S_i^k + (\nabla_i f(x^k))^2,$$

La quantità S accumula la somma dei quadrati delle derivate parziali,

$$S_i^{k+1} = \sum_{t=0}^{k} (\nabla_i f(x^t))^2.$$

$$x_i^{k+1} = x_i^k - \frac{\alpha_0}{\sqrt{s_i^k + \varepsilon}} \nabla_i f(x^k).$$

In questo caso, se i gradienti sono grandi, il passo diminuisce sempre di più. I passi però tendono ad essere troppo grandi.

6.6.3 RMSprop

$$S^{k+1} = \rho S_i^k + (1-\rho)\nabla_i f(x^k), \qquad \rho \in (0,1),$$
$$= (1-\rho)\nabla_i f(x^k)^2 + \rho(1-\rho)(\nabla_i f(x^k))^2 + \rho^2(1-\rho)(\nabla_i f(x^k))^2.$$

Osservazione 6.6.3. I termini più vecchi incidono sempre meno.

6.6.4 Ado Delta

S_i^{k+1} si trova come sopra, ma

$$M^{k+1} = \rho_2 m^k + (1-\rho_2)(x^k - x^{k-1})^2,$$

$$x^{k+1} = x^k + \frac{\sqrt{m^k + \varepsilon}}{\sqrt{S_i^k + \varepsilon}} \nabla_i f(x^k).$$

Qui, oltre a penalizzare i gradienti grandi, valorizziamo le componenti che hanno dato grandi spostamenti. Non ci sono qui valori da settare come iperparametri.

6.6.5 Adam

Questo è l'algoritmo preferito in generale (*adaptive moment estimation*)

$$v_i^{k+1} = \rho v_i^k + (1-\rho)(\nabla_i f(x^k))^2, \qquad \rho \in (0,1),$$
$$m_i^{k+1} = \beta m_i^k + (1-\beta)\nabla_i f(x^k).$$

Queste due formule approssimano media e varianza della direzione di ottimizzazione. In generale $m_i^0, v_i^0 = 0$ ma questo introduce un bias in queste stime,

$$m_i^k = (1-\beta)\sum_{t=0}^{k} \beta^{k-t}\nabla_i f(x^{k-t}),$$

$$E\big[m_i^k\big] = (1-\beta)E\big[\nabla_i f(x^k)\big]\sum_{t=0}^{k}\beta^{k-t} = (1-\beta)E\big[\nabla_i f(x^k)\big]\frac{1-\beta^{k+1}}{1-\beta} =$$
$$= E\big[\nabla_i f(x^k)\big](1-\beta^{k+1}).$$

Il valore atteso non è quindi lo stesso di $\nabla_i f(x^k)$. Si definiscono allora

$$\hat{m}_i^k = \frac{m_i^k}{1-\beta^k}, \qquad \hat{v}_i^k = \frac{v_i^k}{1-\beta^k}.$$

Se $k = 0$,

$$\hat{m}_i^k = \frac{m_i^1}{beta} = \frac{1}{1-\beta}\big(\beta m_i^0 + (1-\beta)\nabla_i f(x^0)\big) = \nabla_i f(x^0).$$

Alla prima iterazione m è la derivata parziale stessa.

$$x_i^{k+1} = x_i^k - \frac{\alpha_0}{\sqrt{\hat{v}_i^k + \varepsilon}}\hat{m}_i^k.$$

In generale funziona ma ci sono casi in cui non si comporta troppo bene.

6.7 Calcolo del Gradiente nel Deep Learning, Algoritmo di Back Propagation

Nelle reti neurali, le opzioni per il calcolo del gradiente sono

- Ricavare ∇f in forma analitica. Questa non è un'opzione.
- Usare le differenze finite. Questo non funziona a causa degli errori che si propagano lungo la rete.
- Usare strumenti di calcolo simbolico. Questa via è ancora non gestibile per quanto riguarda i costi.
- Differenziazione automatica AD.

Quest'ultima è un insieme di tecniche per le derivate che sfrutta la regola di derivazione a catena. Se $f = f(g(x))$,

$$\frac{\partial f}{\partial x_j} = \sum_i \frac{\partial f}{\partial g_i}\frac{\partial g_i}{\partial x_j}.$$

L'algoritmo di Back Propagation, o AD *reverse mode* fa proprio questo. Ne vediamo il funzionamento su un esempio.

Esempio 6.7.1.

$$f(x_1, x_2) = \log(x_1) + x_1 x_2 - \sin(x_2).$$

Chiamiamo

$v_{-1} = x_1, \qquad v_0 = x_2, \qquad v_1 = \log(v_{-1}), \qquad v_2 = v_{-1}v_0, \qquad v_3 = \sin v_0, \qquad v_4 = v_1 + v_2, \qquad v_5 = v_4 - v_3.$

Infine $y = v_5$. Poniamo tale scomposizione in un grafo di computazione. Se cerco la derivata in

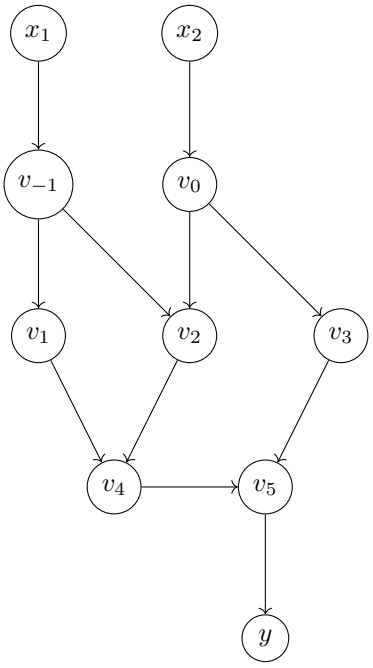

Figura 6.2: Grafo dell'Esempio 6.7.1.

$(2, 5)$,

$v_{-1} = 2, \qquad v_0 = 5, \qquad v_1 = \log 2, \qquad v_2 = 10, \qquad v_3 = \sin 5, \qquad v_4 = \log 2 + 10, \qquad v_5 = \log 2 + 10 - \sin 5.$

$$y \approx 11,652.$$

Tornando indietro calcoliamo il gradiente usnado la notazione

$$\bar{v}_t = \frac{\partial y}{\partial v_t}.$$

$$\bar{v}_5 = \frac{\partial y}{\partial y} = 1.$$

$$\bar{v}_4 = \bar{v}_5 \frac{\partial v_5}{\partial v_4} = \bar{v}_5 \cdot 1 = \bar{v}_5 = 1.$$

$$\bar{v}_3 = \bar{v}_5 \frac{\partial v_5}{\partial v_3} = \bar{v}_5 \cdot (-1) = -1.$$

$$\bar{v}_2 = \bar{v}_4 \frac{\partial v_4}{\partial v_2} = \bar{v}_4 \cdot 1 = 1.$$

$$\bar{v}_1 = \bar{v}_4 \frac{\partial v_4}{\partial v_1} = \bar{v}_4 \cdot 1 = 1.$$

$$\bar{v}_0 = \bar{v}_2 \frac{\partial v_2}{\partial v_0} + \bar{v}_3 \frac{\partial v_3}{\partial v_0} = 1 \cdot v_{-1} + (-1) \cos v_0.$$

$$\bar{v}_{-1} = \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}} = \bar{v}_2 v_0 + \bar{v}_1 \frac{1}{v_{-1}}.$$

Indice analitico