Riassunto Ottimizzazione

Jay Senoner

April 24, 2024

1 Introduzione

Un **Problema di ottimizzazione** è un problema della forma

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

- **X**: Variabili, quantità di interesse che possono essere scelte
- **S**: Insieme ammissibile(dei vincoli sulle variabili **x**)
- **f(x)**: Funzione obiettivo $f : S \mapsto \mathbb{R}$, misura la bontà di una certa assegnazione delle variabili

Un problema di ottimizzazione generico può essere

- **Lineare** se $f(x) = c^T x, \quad S = \{x : Ax \leq b\}$
- **Non lineare**(la funzione obiettivo non è esprimibile come un prodotto scalare di 2 vettori)
- **Continui** se $S \subseteq \mathbb{R}^n$
- **Misti interi** se $S \subseteq \mathbb{R}^{n-m} \times \mathbb{Z}^m$
- **Interi** se $S \subseteq \mathbb{Z}^n$

Inoltre è possibile suddividere i problemi di ottimizzazione in **Vincolati** e **Non vincolati** rispettivamente quando l'insieme ammissibile è un sottoinsieme di $\mathbb{R}^n$ o $\mathbb{R}^n$ stesso.

Definizione 1.1. Una *norma* è un applicazione

$$||\cdot|| : \mathbb{R}^n \mapsto \mathbb{R}^+$$

che soddisfa le seguenti proprietà

- $||x|| = 0 \iff x = 0 \ \ \forall x \in \mathbb{R}^n$
- $||x + y|| \leq ||x|| + ||y|| \ \ \forall x, y \in \mathbb{R}^n$
- $||\alpha x|| = |\alpha|||x|| \ \ \forall x \in \mathbb{R}^n, \forall \alpha \in \mathbb{R}$

Vediamo adesso cosa significa risolvere un problema di ottimizzazione

Definizione 1.2. Un punto $x^* \in S$ si dice **Punto di ottimo globale** per f se:

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

Il valore $f(x^*)$ si dice **ottimo globale**

Se la funzione obiettivo è illimitata inferiormente su S, S è vuoto o la funzione obiettivo presenta asintoti orizzontali, il problema non ammette soluzione.

Definizione 1.3 (Insieme Aperto,Limitato,Compatto)**.** Un insieme S si dice

- **Aperto** se $\forall x \in S \ \ \exists \rho > 0$ tale che $\mathcal{B}_\rho(x) \subseteq S$
  (cioè per ogni punto in S esiste sempre un raggio positivo rho tale per cui la palla di centro nel punto e raggio rho è interamente contenuta in S)
- **Chiuso** se il complementare è aperto
- **Limitato** se $\exists M > 0 : ||x|| \leq M \ \ \forall x \in S$
- **Compatto** se $\forall \{x^k\} \in S \ \ \exists K \subseteq \{0, 1, ..\} : \lim_{k \in K} x^k = \bar{x} \in S$
  (Cioè per ogni possibile sequenza di valori in S esisterà sempre una sottosequenza della sequenza precedente che converge ad un valore limite e tale valore appartiene ad S (limitatezza + chiusura in $\mathbb{R}^n$))

Teorema 1.1 (Weierstrass)**.** *Sia $f : \mathbb{R}^n \to \mathbb{R}$ continua e $S \subseteq \mathbb{R}^n$ compatto.*
Allora f ammette minimo su S

Proof. Sia $Y = f(S)$ l'immagine di S secondo f, e sia $L = \inf Y \implies \exists \{y^k\} \subseteq Y$ tale per cui $\lim y^k = L$.
Definiamo la sequenza $\{x^k\} \subseteq S$ come la sequenza tale per cui $f(x^k) = y^k$.
Poichè S è compatto $\exists K \subseteq \{0, 1, ..\} : \lim x^k = \bar{x} \in S$. Quindi abbiamo

$$\lim f(x^k) = f(\lim x^k) = f(\bar{x}) \ \ e \ \ \lim f(x^k) = \lim y^k = L$$

Da cui si ha $f(\bar{x}) = L$. Poichè per ipotesi L è l'estremo inferiore dell'immagine di S secondo f, si ha la tesi. □

Definizione 1.4 (Insieme di livello)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$ e sia $\alpha \in \mathbb{R}$.
L'**Insieme di livello** $\mathcal{L}_f(\alpha)$ è definito come:

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

Ossia è l'insieme di tutti i valori del dominio per i quali la funzione assume valori minori di una costante reale alpha.

Proposizione 1.5 (Insiemi di livello compatti e esistenza soluzione)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$ continua, e supponiamo che la funzione abbia un insieme di livello $\mathcal{L}_f(\alpha)$ compatto. Allora f ammette minimo su $\mathbb{R}^n$.

Proof. Per Weierstrass, poichè f è continua sull'insieme di livello $\mathcal{L}_f(\alpha)$, allora f ammette minimo x^* in tale insieme . Consideriamo ora un generico $z \in \mathbb{R}^n$.

- Se $z \in \mathcal{L}_f(\alpha) \implies f(x^*) \leq f(z)$, poichè x^* minimizza f nell'insieme di livello.
- Se $z \notin \mathcal{L}_f(\alpha) \implies f(z) > \alpha \geq f(x^*)$ per definizione di insieme di livello.

$\forall z \in \mathbb{R}^n$ vale quindi che $f(x^*) \leq f(z)$, ossia la tesi. □

Questa proposizione in aggiunta al teorema di weierstrass fornisce delle condizioni di esistenza per il punto di minimo globale.

Definizione 1.6 (Funzione coerciva)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$. f si dice **Coerciva** se $\forall\{x^k\} \subseteq \mathbb{R}^n$ tale che $||x^k|| \to \infty$ si ha:

$$\lim f(x^k) = \infty$$

Le funzioni coercive godono di un importante proprietà: una funzione è continua e coerciva se e solo se ha tutti gli insiemi di livello compatti. Ciò implica per la proposizione 1.5 che se f è continua e coerciva allora ammette minimo su $\mathbb{R}^n$

Definizione 1.7 (Matrici definite/semidefinite positive)**.** Sia $Q \in \mathbb{R}^{n\times n}$. La matrice Q è detta **semidefinita positiva** se $\forall x \in \mathbb{R}^n$ vale

$$x^T Q x \geq 0$$

mentre è detta **definita positiva** se $\forall x \neq 0$ vale

$$x^T Q x > 0$$

Inoltre, se la matrice Q è simmetrica, si hanno le seguenti doppie implicazioni:

- Q è semidefinita positiva $\iff$ Ha tutti gi autovalori diversi da 0

- Q è definita positiva $\iff$ Ha tutti gli autovalori maggiori di 0
- (Minmax property): Se Q è semidefinita positiva $\forall x \in \mathbb{R}^n : x^T Q x \geq \lambda_{min}||x||^2$

Definizione 1.8 (Problema Quadratico)**.** Un **Problema Quadratico** è un problema della forma

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

con Q matrice simmetrica

Proposizione 1.9 (Coercività funzioni quadratiche)**.** Una funzione quadratica f è coerciva $\iff$ Q è definita positiva.

Proof. Dimostriamo innanzitutto che se Q è definita positiva $\implies$ f coerciva.

$$f(x) = \frac{1}{2} x^T Q x - c^T x \geq \frac{1}{2}\lambda_{min}||x||^2 - c^T x \geq \frac{1}{2}\lambda_{min}||x||^2 - ||c||||x||$$

Dove la prima disuguaglianza è ottenuta utilizzando la minmax property e la seconda è data dalla disuguaglianza di cauchy shwartz(prodotto scalare $\leq$ prodotto delle norme) Sia ora $\{x^k\} \subseteq \mathbb{R}^n$ tale che $||x^k|| \to \infty$.

$$\lim_{k\to\infty} f(x^k) \geq \lim_{k\to\infty} \frac{1}{2}\lambda_{min}||x||^2 - ||c||||x^k|| = \lim_{k\to\infty} ||x^k||(\frac{1}{2}\lambda_{min}||x^k|| - ||c||) = \infty$$

Il risultato del limite è dato dal fatto che λ_{min} è maggiore di zero poiche per ipotesi Q è definita positiva, e la norma di c è una quantità finita. Poichè abbiamo dimostrato che per una generica sequenza di valori del dominio tendente a infinito la funzione obiettivo tende a infinito, abbiamo mostrato che f è coerciva.
Ora dimostriamo che se f è coerciva $\implies$ Q definita positiva.
Supponiamo per assurdo che Q non sian definita positiva, ossia che

$$\exists y \in \mathbb{R}^n, y \neq 0 : \quad y^T Q y \leq 0$$

Senza perdita di generalità supponiamo inoltre che $c^T y \geq 0$. Costruiamo quindi una sequenza $\{x^k\}$ tale per cui $x^k = ky \ \ \forall k$. Consideriamo $f(x^k)$:

$$f(x^k) = \frac{1}{2} {x^k}^T Q x^k - c^T x^k = \frac{1}{2}(ky)^T Q (ky) - c^T (ky) = \frac{1}{2} k^2 y^T Q y - c^T k y = k(\frac{1}{2} k y^T Q y - c^T y)$$

Passando al limite la relazione si ottiene:

$$\lim_{k\to\infty} f(x^k) = \lim_{k\to\infty} k(\frac{1}{2} k y^T Q y - c^T y) < \infty$$

Il limite è sicuramente diverso da ∞, poichè le quantità $y^t Q y$ e $-c^T y$ sono ≤ 0 per ipotesi, da cui si deduce la non coercività della funzione f, ossia un assurdo. $\square$

Possiamo quindi derivare dalla proposizione precedente la seguente proprietà:

Proposizione 1.10 (Esistenza soluzione per problemi quadratici)**.** Se la matrice Q è definita positiva la funzione quadratica $f(x) = \frac{1}{2}x^TQx - c^Tx$ ammette minimo su $\mathbb{R}^n$.

Osserviamo che certificare l'ottimalità di una soluzione, ossia determinare, una volta trovato x^*, se $\forall x \in S \;\; f(x^*) \leq f(x)$, mediante un algoritmo, è un problema **NP-HARD**, ossia un problema intrattabile a livello di risorse computazionali necessarie per ottenere la certificazione di ottimalità globale. Ci concentriamo quindi su un altra classe di punti di minimo.

Definizione 1.11 (Punto di minimo locale)**.** Un punto $x^* \in S$ si dice **Punto di minimo locale** se per qualche $\rho > 0, \rho \in \mathbb{R}$:

$$f(x^*) \leq f(x) \;\; \forall x \in S \cap \mathcal{B}_\rho(x^*)$$

Da ora in poi considereremo quasi sempre soluzioni locali a problemi di ottimizzazione, e solo in rari casi cercheremo di determinare ottimi globali.

Definizione 1.12 (Insieme convesso)**.** Un insieme $S \subseteq \mathbb{R}^n$ si dice **Convesso** se $\forall x, y \in S$ e $\forall \lambda \in [0, 1]$ risulta:

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

Graficamente un insieme è convesso se presi due generici punti dell'insieme, il segmento che li congiunge è interamente contenuto nell'insieme S.

Proposizione 1.13. La palla $\mathcal{B}_\rho(\bar{x})$ è convessa.

Proof. Siano $x, y \in \mathcal{B}_\rho(\bar{x})$, e siano $\lambda \in [0, 1]$ e $z = \lambda x + (1 - \lambda)y$. Vogliamo verificare l'appartenenza di z alla palla $\mathcal{B}_\rho(\bar{x})$.

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

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

Dove la minoranza è data dalla disuguaglianza triangolare, e l'ultimo passaggio è dato dal fatto che per ipotesi sia x che y appartengono alla palla. □

Definizione 1.14 (Funzione convessa)**.** Sia $S \subseteq \mathbb{R}^n$ un insieme convesso e $f : S \to \mathbb{R}$ continua. Si dice che f è convessa se $\forall x, y \in S, \;\; \forall \lambda \in [0, 1]$

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

La condizione di convessità di una funzione ha un interpretazione geometrica di facile comprensione: una funzione si dice convessa nel momento in cui presi due generici punti nell'immagine secondo f di un sottoinsieme convesso del dominio, il segmento congiungente tali punti maggiora tutti gli altri punti dell'immagine di f calcolata sulla preimmagine del segmento secondo f.

Un esempio di funzione convessa è la norma, infatti se consideriamo $x, y \in \mathbb{R}^n$, $\lambda \in [0,1]$ e $z = \lambda x + (1-\lambda)y$ si ottiene:

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

Proposizione 1.15 (Sulle funzioni obiettivo convesse con vincoli convessi)**.** Sia $S \subseteq \mathbb{R}^n$ convesso e $f : S \to \mathbb{R}$ convessa. $\implies$ i punti di minimo locale per il problema

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

sono tutti e soli i punti di minimo globale(x minimo locale $\iff$ x minimo globale)

Proof. Dobbiamo solo dimostrare che se valgono le ipotesi della proposizione i punti di minimo locale sono anche globali, in quanto in ogni caso i punti di minimo globale saranno localmente minimi per f.
Consideriamo quindi $x^* \in S$ minimo locale, e sia $y \in S$ un punto di S. Definiamo il segmento $z = (1-\lambda)x^* + \lambda y$ con $\lambda \in [0,1]$. Ora:

- S è convesso $\implies$ $z \in S$.
- x^* è di minimo locale $\implies$ $\exists \rho > 0$ tale per cui $f(x^*) \leq f(x)$ $\forall x \in S \cap \mathcal{B}_\rho(x^*)$.
- Dobbiamo inoltre notare che per λ sufficientemente piccolo il punto sul segmento z che congiunge x^* e y appartiene alla palla di raggio rho e centro in x^*. Questo perchè per quanto il raggio rho possa essere piccolo, esisterà sempre un valore λ^* del parametro lambda che fa si che il punto $z(\lambda^*) \in \mathcal{B}_\rho(x^*)$.

Ricordando che per ipotesi x^* è minimo locale per f in S, si ha

$$f(x^*) \leq f(z) = f((1-\lambda)x^* + \lambda y) \leq (1-\lambda)f(x^*) + \lambda f(y)$$

Dove la disuguaglianza è data dall'ipotesi di convessità della funzione f. Si ha quindi

$$f(x^*) \leq f(x^*) - \lambda f(x^*) + \lambda f(y) \implies f(x^*) \leq f(y) \quad \forall y \in S$$

Cioè x^* è minimo globale per f. □

Definizione 1.16. Sia $f : S \to \mathbb{R}$, S convesso. f si dice **Strettamente convessa** se $\forall x, y \in S, \;\; x \neq y, \;\; \forall \lambda \in (0,1)$:

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

La definizione è analoga a quella di funzione convessa eccetto per il fatto che non sono strettamente convesse funzioni che assumono valori sul segmento congiungente i punti dell'immagine al di fuori dai valori assunti nei punti estremi. Intuitivamente, ad 1 variabile, tutte le funzioni costanti a tratti su un convesso potranno essere convesse ma non lo saranno mai strettamente, in quanto presi due punti in cui f è costante e pari allo stesso valore sicuramente il segmento che li congiunge sarà interamente contenuto nell'immagine della funzione sul convesso corrispondente.

Proposizione 1.17 (Unicità del punto di minimo per funzioni obiettivo strettamente convesse)**.** Sia $f : S \to \mathbb{R}$ strettamente convessa. Allora il punto di minimo, se esiste, è unico.

Proof. Per assurdo siano $\bar{x}$ e $\bar{y}$ due punti di minimo globale diversi tra loro. Sia poi z il segmento che li congiunge, $z = (1-\lambda)\bar{x} - \lambda\bar{y}$.

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

Dove la disuguaglianza è data dall'ipotesi di stretta convessità della funzione f, e le successive uguaglianze dal fatto che due punti di minimo globale assumeranno sempre lo stesso valore minimo della funzione sul domini Poichè $\bar{x}$ era punto di minimo globale per ipotesi, si ha un assurdo. □

Richiamiamo adesso alcune fondamentali nozioni di analisi matematica(2).

Definizione 1.18 (Derivata Direzionale)**.** Siano $f : \mathbb{R}^n \to \mathbb{R}$, $\bar{x} \in \mathbb{R}^n$ e $d \in \mathbb{R}^n$.
La **Derivata direzionale** della funzione f nel punto $\bar{x}$ lungo la direzione d è data da

$$D_f(\bar{x}, d) = \lim_{t \to 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

Definizione 1.19 (Derivata parziale)**.** Siano $f : \mathbb{R}^n \to \mathbb{R}$, $\bar{x} \in \mathbb{R}^n$ e $e_i \in \mathbb{R}^n$ una delle direzioni canoniche associata alla variabile x_i.
La **Derivata parziale** di f rispetto alla variabile x_i è data da.

$$\frac{\partial f(\bar{x})}{\partial x_i} = D_f(\bar{x}, e_i)$$

Definizione 1.20 (Gradiente)**.** Siano $f : \mathbb{R}^n \to \mathbb{R}$ e $\bar{x} \in \mathbb{R}^n$.
Il gradiente $\nabla f(\bar{x})$ è definito come:

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

Definizione 1.21 (Funzione continuamente differenziabile)**.** Se $\nabla f : \mathbb{R}^n \to \mathbb{R}$ esiste ed è continua su $\mathbb{R}^n$ allora la funzione f si dice **Continuamente differenziabile** e si indica con $f \in C^1(\mathbb{R}^n)$.

Proposizione 1.22 (Gradiente e derivata direzionale)**.** Se $f \in C^1(\mathbb{R}^n)$ allora $\forall x \in \mathbb{R}^n$ e $\forall d \in \mathbb{R}^n$

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

Cioè, se f è continuamente differenziabile, allora la derivata direzionale in un generico punto $x \in \mathbb{R}^n$ lungo una generica direzione $d \in \mathbb{R}^n$ è data dal prodotto scalare tra il gradiente calcolato nel suddetto punto e la direzione d di derivazione.

Proposizione 1.23 (Sul gradiente)**.** Il gradiente di una funzione $f : \mathbb{R}^n \to \mathbb{R}$ rappresenta la **direzione di massima salita**, ossia la direzione per la quale effettuando spostamenti si ottiene il massimo incremento della funzione obiettivo.

Proof. Consideriamo il problema

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

Notiamo che il massimo della funzione obiettivo coincide con il massimo della funzione $\cos(\theta(d, \nabla f(x))$, in quanto la norma della direzione è vincolata ad 1 e la norma del gradiente è una costante. La funzione coseno è massima quanto i vettori che formano l'angolo sono paralleli, per cui la soluzione è tale per cui $d \parallel \nabla f(x)$. Poichè i vincoli richiedono $||d|| = 1$, si ha

$$d^* = \frac{\nabla f(x)}{||\nabla f(x)||}$$

□

Definizione 1.24 (Hessiana)**.** Siano $f : \mathbb{R}^n \to \mathbb{R}$ e $\bar{x} \in \mathbb{R}^n$.

L'**Hessiana** di f nel punto $\bar{x}$ è la matrice $\nabla^2 f(\bar{x}) \in \mathbb{R}^{n\times n}$ definita da:

$$\begin{pmatrix} \frac{\partial^2 f(\bar{x})}{\partial {x_1}^2} & \frac{\partial^2 f(\bar{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\bar{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\bar{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\bar{x})}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f(\bar{x})}{\partial x_2 \partial x_n} \\ \vdots & & & \\ \frac{\partial^2 f(\bar{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\bar{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\bar{x})}{\partial {x_n}^2} \end{pmatrix}$$

Definizione 1.25 (2 volte continuamente differenziabile)**.** Se $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ esiste ed è continua su $\mathbb{R}^n$ allora la funzione f si dice **Due volte continuamente differenziabile** e si indica con $f \in C^2(\mathbb{R}^n)$. Se f è due volte continuamente differenziabile, ossia se tutte le possibili derivate parziali seconde soni continue, allora la matrice hessiana è una matrice simmetrica.

Definizione 1.26 (Jacobiano di funzioni vettoriali)**.** Siano $F : \mathbb{R}^n \to \mathbb{R}^m$ e $x \in \mathbb{R}^n$ tali per cui $F(\bar{x}) = (F_1(\bar{x}), F_2(\bar{x}), \cdots, F_m(\bar{x}))$. Si definisce lo **Jacobiano** di F come:

$$J_F(\bar{x}) = \begin{pmatrix} \nabla F_1(\bar{x})^T \\ \vdots \\ \nabla F_m(\bar{x})^T \end{pmatrix}$$

Osserviamo che si può interpretare l'hessiana di f come il jacobiano del gradiente ∇f.

Ad esempio, per una funzione lineare $f(x) = c^T x$ si ha:

$$\nabla f(x) = c, \quad \nabla^2 f(x) = 0$$

Mentre per una funzione quadratica $f(x) = \frac{1}{2} x^T Q x$ con Q matrice $n \times n$ a valori reali si ha

$$\nabla f(x) = Qx, \quad \nabla^2 f(x) = Q$$

Per la funzione $f(x) = \frac{1}{2}||x||^2 = \frac{1}{2} x^T x$ si ha

$$\nabla f(x) = x, \quad \nabla^2 f(x) = I$$

Infine per $f(x) = \frac{1}{2}||Ax - b||^2$ ci si può ricondurre al caso di forma quadratica dato che

$$||Ax - b|| = (Ax - b)^T (Ax - b) = x^T A^T A x - 2b^T A x + b^t b$$

per cui si ha

$$\nabla f(x) = A^T A x - b^T A, \quad \nabla^2 f(x) = A^T A$$

Per i calcoli che giustificano questi risultati, vedi quaderno.

Proposizione 1.27 (Funzioni convesse)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1(\mathbb{R}^n)$.
f è convessa $\iff$

$$\begin{cases} f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) & \forall x, \bar{x} \in \mathbb{R}^n \\ f(x + d) \geq f(x) + \nabla f(x)^T d & \forall x, d \in \mathbb{R}^n \end{cases}$$

Osserviamo che per $n = 1$ la prima condizione equivale ad affermare che la funzione obiettivo in un generico punto $x \in \mathbb{R}^n$ maggiora la retta tangente in un qualsiasi punto $\bar{x} \in \mathbb{R}^n$. Per $n \geq 1$ l'equazione della prima condizione rappresenta l'iperpiano tangente a f nel punto $\bar{x}$.

Proposizione 1.28 (Convessità e Hessiana)**.** Se $f \in C^2(\mathbb{R}^n)$ vale

- f è convessa $\iff$ $\nabla^2 f(x)$ è semidefinita positiva $\forall x \in \mathbb{R}^n$
- Se $\nabla^2 f(x)$ è definita positiva $\forall x \in \mathbb{R}^n \implies$ f è strettamente convessa.

Definizione 1.29 (Direzione di discesa)**.** Una direzione $d \in \mathbb{R}^n$ si dice **Direzione di discesa** per f nel punto $x \in \mathbb{R}^n$ se:

$$\exists \bar{t} > 0 : f(x + td) < f(x) \ \ \forall t \in [0, \bar{t})$$

Una direzione di discesa d per una funzione f in un punto x è tale per cui esiste un valore massimo $\bar{t}$ del passo tale per cui tutti i passi positivi minori di $\bar{t}$ lungo la direzione d causano uno stretto decremento della funzione obiettivo.

Proposizione 1.30 (Minimi locali e direzioni di discesa)**.** Se $\bar{x}$ è un punto di minimo locale per f allora non esistono direzioni di discesa per f in $\bar{x}$

Osserviamo che questa proposizione rappresenta una condizione necessaria di ottimalità locale, ma non sufficiente.

Proposizione 1.31 (Sulle direzioni di discesa)**.** Siano $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ e $f \in C^1(\mathbb{R}^n)$.
Se $D_f(x, d) = \nabla f(x)^T d < 0 \implies$ d è direzione di discesa

Proposizione 1.32 (Sulle direzioni di discesa)**.** Se $\nabla f(x) \neq 0$ allora la direzione $d = -\nabla f(x)$ è una direzione di discesa per f
La direzione $-\nabla f(x)$ si dice **Antigradiente** di f in x.

Proof. Basta considerare

$$\nabla f(x)^T d = -\nabla f(x)^T \nabla f(x) = -||\nabla f(x)||^2 < 0$$

per $\nabla f(x) \neq 0$. □

Proposizione 1.33. Se $\bar{x}$ è punto di minimo locale
$\implies \nabla f(\bar{x}) = 0$

I punti $x \in \mathbb{R}^n$ tali per cui $\nabla f(x) = 0$ si dicono **Punti stazionari**

Proposizione 1.34 (Convessità e direzioni di discesa)**.** Sia $f \in C^1(\mathbb{R}^n)$ convessa.
Una direzione $d \in \mathbb{R}^n$ è di discesa per f in un punto $x \in \mathbb{R}^n \iff \nabla f(\bar{x})^T d < 0$

Proof. Dimostriamo innanzitutto la proposizione: $\nabla f(\bar{x})^T d < 0 \implies d$ di discesa.(proposizione 1.31)

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

per t sufficientemente piccolo. In particolare quindi poichè il valore limite è negativo esisterà sempre un valore $\bar{t}$ del passo t tale per cui passi inferiori a $\bar{t}$ lungo d risultano in uno stretto decremento della funzione obiettivo, ossia d è una direzione di discesa.
Dimostriamo ora che se d è di discesa e $f \in C^1(\mathbb{R}^n)$ convessa $\implies \nabla f(\bar{x})^T d < 0$.
Se d è di discesa $\implies$ per t sufficientemente piccolo $f(\bar{x} + td) < f(\bar{x})$. Per ipotesi si ha poi che la funzione f è convessa e differenziabile, da cui si ha (proposizione 1.27)

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

Da cui si ha

$$f(\bar{x}) + t\nabla f(\bar{x})^T d \leq f(\bar{x} + td) < f(\bar{x})$$

per definizione di direzione di discesa per f in $\bar{x}$. Possiamo quindi ricavare

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

poichè $t > 0$ per definizione. □

Proposizione 1.35 (Funzioni convesse e punti stazionari)**.** Sia $f \in C^1(\mathbb{R}^n)$ convessa.
$\bar{x} \in \mathbb{R}^n$ è punto di minimo globale $\iff \nabla f(\bar{x}) = 0$

Proof. Dalla convessità di f si ha

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

Ma $\nabla f(\bar{x}) = 0 \implies f(x) \geq f(\bar{x}) \quad \forall x \in \mathbb{R}^n$ □

Richiami su teoremi di analisi 2
Sia $f \in C^1(\mathbb{R}^n)$.

Teorema 1.2 (Teorema della media)**.**

$$f(y) = f(x) + \nabla f(\xi)^T (y - x) \ \ \forall x, y \in \mathbb{R}^n$$

Per qualche $\xi = (1 - \lambda)x + \lambda y$ *con* $\lambda \in (0, 1)$*. Inoltre*

$$f(x + d) = f(x) + \nabla f(\xi)^T d$$

Per qualche $\xi = x + \theta d$ *con* $\theta \in (0, 1)$*.Infine*

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

Con $\lim_{||d|| \to 0} \frac{\alpha(x,d)}{||d||} = 0$

Teorema 1.3 (Formula di Taylor)**.**

$$f(x + d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(\xi) d$$

Per qualche $\xi = x + \theta d$ *con* $\theta \in (0, 1)$

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

Con $\lim_{||d|| \to 0} \frac{\beta(x,d)}{||d||^2} = 0$

Definizione 1.36 (Direzione a curvatura negativa)**.** Sia $f \in C^2(\mathbb{R}^n)$ e $x \in \mathbb{R}^n$. Una direzione $d \in \mathbb{R}^n$ si dice **Direzione a curvatura negativa** per f in x se vale:

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

Proposizione 1.37 (Caso $\nabla f(x)^T d = 0$)**.** Siano $x \in \mathbb{R}^n$ e $d \in \mathbb{R}^n$ tali che:

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

Allora d è direzione di discesa per f in x.

Proof. Applichiamo la formula di Taylor per $f(x + td)$:

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

Per ipotesi $\nabla f(x)^T d = 0$, per cui si ha:

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

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

Da Taylor si ha che per $t^2 \to 0$ vale $\beta(x,td) \to 0$ e per ipotesi della proposizione abbiamo che $d^T\nabla^2 f(x)d < 0$, da cui si ha che per t sufficientemente piccolo

$$\frac{1}{2}d^t\nabla^2 f(x)d + \frac{\beta(x,td)}{t^2} < 0$$

Da cui si ottiene che per t sufficientemente piccolo

$$\frac{f(x+td)-f(x)}{t^2} < 0 \implies f(x+td) < f(x)$$

□

Proposizione 1.38 (Condizioni necessarie ottimalità locale del secondo ordine)**.** Sia $\bar{x} \in \mathbb{R}^n$ un punto di minimo locale per f

1. $\nabla f(\bar{x}) = 0$
2. $\nabla^2 f(x)$ è semidefinita positiva

Proof. (1): Proposizione 1.33
(2): Per assurdo supponiamo che $\exists y \in \mathbb{R}^n : y^T\nabla^2 f(y)y < 0$. Si ha quindi che

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

Per la proposizione 1.37 la direzione y è di discesa per f in $\bar{x} \implies$ assurdo, poichè $\bar{x}$ è minimo locale, pertanto non esistono direzioni di discesa per f in $\bar{x}$. □

Proposizione 1.39 (Condizioni sufficienti ottimalità locale del secondo ordine)**.** Siano $f : \mathbb{R}^n \to \mathbb{R}$ e $\bar{x} \in \mathbb{R}^n$. Se valgono la condizione (1) e almeno una tra le condizioni (2) e (3):

1. $\nabla f(\bar{x}) = 0$
2. $\nabla^2 f(\bar{x})$ è definita positiva
3. $\nabla^2 f(\bar{x})$ è semidefinita in un intorno sferico del punto $\bar{x}$

Allora $\bar{x}$ è punto di minimo locale per f.
Poniamoci nel caso quadratico, dove la funzione obiettivo è data da $f(x) = \frac{1}{2}x^TQx - c^Tx$, in questo caso valgono le seguenti proprietà:

1. Il problema di minimo non vincolato ammette soluzione $\iff \exists \bar{x} \in \mathbb{R}^n : \quad Q\bar{x} - c = 0$ e Q è semidefinita positiva

2. Se Q è semidefinita positiva allora ogni punto $\bar{x} \in \mathbb{R}^n$ tale che $Q\bar{x} - c = 0$ è punto di minimo globale

3. Il punto di minimo locale esiste ed è unico $\iff$ Q è definita positiva

Proof. Dimostriamo le proprietà sopra esposte per i problemi quadratici, partendo da (1).
$\implies$
Supponiamo che esista un punto $\bar{x}$ di minimo globale. Per le condizioni necessarie di ottimalità del secondo ordine si ha:

- $\nabla f(\bar{x}) = 0 \iff Q\bar{x} - c = 0$
- $\nabla^2 f(\bar{x}) = Q$ è semidefinita positiva

□

Proof. (1) $\impliedby$ Per ipotesi abbiamo che $\exists \bar{x} : \nabla f(\bar{x}) = Q\bar{x} - c = 0$ e che $\nabla^2 f(\bar{x})$ è semidefinita positiva $\forall x$.
La tesi segue dalle condizioni sufficienti di ottimalità del secondo ordine e tenendo conto del fatto che f è convessa poichè la matrice Q è semidefinita positiva,quindi i punti di minimo locale sono minimi globali per il problema. □

Proof. La dimostrazione di (2) segue dalla convessità di f garantita dal fatto che Q è semidefinita positiva per ipotesi in (2) (proposizione 1.35 □

Proof. Dimostriamo (3)
$\impliedby$
Supponiamo che Q sia definita positiva: si ha che

- f è coerciva $\implies$ ammette minimno globale(esistenza minimo globale)
- $\nabla^2 f(x) = Q$ è definita positiva $\implies$ f è strettamente convessa $\implies$ il minimo globale è unico

□

Proof. (3) $\implies$
Supponiamo adesso che il punto di minimo globale x^* per il problema quadratico esista e sia unico. Supponiamo ora per assurdo che Q non sia definita positiva. Dal punto (1) sappiamo pero che se il problema ammette soluzione (cosa che abbiamo supposto)

allora la matrice Q deve essere semidefinita positiva. Pertanto la matrice Q deve necessariamente essere singolare, e quindi non invertibile. Consideriamo quindi il sistema

$$Qx = c$$

Poichè Q è singolare il sistema può non avere soluzioni o averne infinite. Entrambi i casi portano però ad un assurdo, in quanto è stato supposto che il minimo esista e sia unico. □

2 Regressione lineare ai minimi quadrati regolarizzata

Il problema di regressione lineare ai minimi quadrati regolarizzata è il problema

$$w^* = \arg\min_w \frac{1}{2}\sum_{i=1}^{N}(w^t x_i - y_i)^2 = \arg\min_w \frac{1}{2}||Xw - Y||^2$$

Aggiungiamo un regolarizzatore quadratico $\lambda||w||^2$:

$$w^* = \arg\min_w \frac{1}{2}(||Xw - Y||^2 + \lambda||w||^2)$$

Ci chiediamo quindi:

- w^* esiste sempre?
- w^*, quando esiste, è unico?

Proof. (1)

$$f(w) = \frac{1}{2}(w^T X^T X w - 2Y^T x w + Y^T Y) + \frac{\lambda}{2}w^T w$$

Sia $\{w^k\}$ tale che $||w^k|| \to \infty$

$$\lim_{k\to\infty} f(w^k) = \lim_{k\to\infty} \frac{1}{2}(||Xw^k - Y||^2 + \lambda||w^k||^2) = \infty$$

Si ha quindi che f è coerciva, inoltre f è continua, pertanto ammette minimo su $\mathbb{R}^n$ □

Proof. (2) $f(w)$ è della forma $f(x) = \frac{1}{2}||Ax - b||$, per cui conosciamo il gradiente $\nabla f(x) = A^T Ax - b^T A$ e l'hessiana $\nabla^2 f(x) = A^T A$:

$$\nabla f(w) = X^T X w - Y^T x + \lambda I w, \;\; \nabla^2 f(w) = X^T X + \lambda I$$

e l'hessiana $\nabla^2 f(w)$ è definita positiva. Per le proprietà viste sui problemi quadratici, la soluzione w^* esiste ed è unica: in particolare è l'unico punto tale per cui $\nabla f(w^*) = 0$:

$$\nabla f(w^*) = X^T X w^* - Y^T x + \lambda I w^* = 0$$

Cioè la soluzione del problema di regressione lineari ai minimi quadrati regolarizzata è data dalla soluzione del sistema lineare delle **Equazioni normali**

$$(X^T X + \lambda I)w = Y^T X$$

Osserviamo che calcolare la soluzione w^* risolvendo il sistema invertendo la matrice può risultare computazionalmente oneroso, poichè invertire una matrice è in generale un operazione $O(n^p)$ dove p è il numero di variabili del problema. Inoltre vi sono anche problemi di errori in artimentica finita nell'invertire una matrice. Si ricorre quindi a **metodi iterativi di ottimizzazione** □

3 Algoritmi Iterativi di Ottimizzazione

Algorithm 1 Schema generale algoritmo iterativo di ottimizzazione

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

Si possono verificare 2 casi

- **Convergenza finita**: $\exists \bar{k} > 0 : \quad \nabla f(x^{\bar{k}}) = 0$
- L'algoritmo produce delle sequenze infinite:

$$\{x^k\}, \quad f(x^k), \quad \{\nabla f(x^k)\}$$

Il caso di convergenza finita non si verifica quasi mai in contesti applicativi: è molto difficile che ad un passo k si abbia esattamente un punto x^k tale che $\nabla f(x^k) = 0$. Nella maggior parte dei casi si ottengono sequenze infinite per le quali vorremmo che

- $\{x^k\}$ abbia punti di accumulazione

- Arrivare in modo asintotico alla stazionarietà
- Convergenza "veloce"

Dobbiamo adesso formalizzare le proprietà sopra elencate per le sequenze prodotte da un algoritmo iterativo di ottimizzazione. Partiamo quindi dal definire formalmente il punto (1).

Proposizione 3.1 (Esistenza di punti di accumulazione)**.** Sia f continua e sia $x^0 \in \mathbb{R}^n$ tale che l'insieme di livello:

$$\mathcal{L}_0 = \{x \in \mathbb{R}^n : \quad f(x) \leq f(x^0)\}$$

è compatto. Supponiamo poi che l'algoritmo generi una sequenza $\{x^k\}$ del tipo $x^{k+1} = x^k + s_k$ tali che $f(x^{k+1}) \leq f(x^k) \implies$

1. $\{x^k\}$ ha punti di accumulazione $\bar{x} \in \mathcal{L}_0$
2. La sequenza $\{f(x^k)\}$ converge ad un valore $\bar{f} > -\infty$

Proof. (1): Per ipotesi $f(x^k) \leq f(x^{k+1})$. Quindi $f(x^{k+1}) \leq f(x^0) \implies x^{k+1} \in \mathcal{L}_0 \ \forall k$, da cui si ha che $\{x^k\} \in \mathcal{L}_0$. Ma per ipotesi $\mathcal{L}_0$ è compatto, pertanto la sequenza $\{x^k\}$ ha punti di accumulazione in $\mathcal{L}_0$.

□

Proof. (2): Essendo $f(x^{k+1}) \leq f(x^k)$ per ipotesi, si ha che la sequenza $\{f(x^k)\}$ è monotona non crescente. Consideriamo il suo valore limite $\bar{f} = \lim f(x^k)$. Inoltre consideriamo una sottosequenza $K \subseteq \{0, 1..\}$ tale che $\lim_{k \in K} x^k = \bar{x}$, che sicuramente esiste per il punto (1). Per la continuità di f vale che $\lim_{k \in K} f(x^k) = f(\bar{x}) \in \mathbb{R} \implies \bar{f} = f(\bar{x})$, poichè tutte le sottosequenze di una sequenza convergente convergono allo stesso valore limite. □

Per convergenza verso la stazionarietà possiamo intendere 3 situazioni differenti:

- $\lim x^k = \bar{x}$ e $\nabla f(\bar{x}) = 0$. In questo caso $\{x^k\}$ converge verso un punto stazionario
- $\lim ||\nabla f(x^k)|| = 0$. La norma del gradiente tende a 0: tutti gli eventuali punti di accumulazione sono stazionari
- $\lim \inf ||\nabla f(x^k)|| = 0$: Almeno un punto di accumulazione è stazionario.

Vediamo un esempio per chiarire meglio i concetti presentati sopra. Consideriamo una funzione di una variabile reale tale per cui $f'(x) = (x-1)(x-3)$ e vediamo 3 sequenze che convergono alla stazionarietà nei 3 modi visti sopra.

1. $\{x^k\} = \{0.9, 0.99, 0.999, 0.9999, 0.99999....\}$
2. $\{x^k\} = \{0.9, 2.99, 0.999, 2.9999, 0.99999....\}$
3. $\{x^k\} = \{1, 6, 1, 6, 1....\}$

Nel primo caso la sequenza converge al punto stazionario $\bar{x} = 1$. Nel secondo caso la sequenza è composta da due sottosequenze convergenti rispettivamente ai punti stazionari $\bar{x}_1 = 1$ e $\bar{x}_2 = 3$: ogni punto di accumulazione della sequenza è stazionario e la norma del gradiente della funzione obiettivo valutata su valori della sequenza tende a 0. Nel terzo caso la sequenza è nuovamente scomponibile in 2 sottosequenze, ma solo una di esse converge a un punto stazionario. Per tutti questi casi, poichè la convergenza finita è difficile da ottenere nella pratica, si stabilisce un **Criterio di arresto** alternativo dato da:

$$||\nabla f(x^k)|| < \epsilon$$

Ci chiediamo adesso se le condizioni di stretto decremento della funzione obiettivo su valori successivi della sequenza $\{x^k\}$ e l'ipotesi di stretta convessità della funzione obiettivo siano sufficienti per garantire la convergenza al minimo globale. Vediamo subito tramite un controesempio come queste condizioni non siano sufficienti a garantire la convergenza al minimo globale.
Consideriamo la funzione

$$\min f(x) = \frac{1}{2}x^2$$

E consideriamo un punto iniziale x^0 tale per cui $|x^0| > 1$ e $x^{k+1} = x^k - \alpha_k f'(x^k)$.
Artificio: Prendiamo il passo $\alpha_k = 2 - \frac{\epsilon_k}{|x^k|}$ con $0 < \epsilon_k < |x^k| - 1$:

$$x^{k+1} = x^k - \alpha_k x^k = x^k(1 - 2 + \frac{\epsilon_k}{|x^k|}) = x^k(\frac{\epsilon_k}{|x^k|} - 1) = -x^k + \epsilon_k \frac{x^k}{|x^k|} =$$

$$-x^k + \epsilon_k sgn(x^k) = sgn(x^k)(-\frac{x^k}{sgn(x^k)} + \epsilon_k) = sgn(x^k)(-|x^k| + \epsilon_k)$$

Adesso valutiamo $|x^{k+1}|$:

$$|x^{k+1}| = |sgn(x^k)(-|x^k| + \epsilon_k)| = |-|x^k| + \epsilon_k|$$

Notiamo che $-|x^k| + \epsilon_k < 0$ in quanto $0 < \epsilon_k < |x^k| - 1$, da cui si ha

$$= -(-|x^k| + \epsilon_k) = |x^k| - \epsilon_k$$

Per ipotesi $0 < \epsilon_k < |x^k| - 1$ da cui $|x^k| - \epsilon_k > 1 \implies |x^{k+1}| = |x^k| - \epsilon_k > 1$. Adesso

valutiamo $f(x^{k+1})$:

$$f(x^{k+1}) = \frac{1}{2}(x^{k+1})^2 = \frac{1}{2}(|x^k| - \epsilon_k)^2 < \frac{1}{2}(|x^k|)^2 = f(x^k)$$

Siamo partiti da una funzione obiettivo f strettamente convessa e abbiamo mostrato che $f(x^{k+1}) < f(x^k)$ sulla sequenza $\{x^k\}$ sopra definita. Tuttavia la sequenza $\{x^k\}$ è tale per cui $|x^k| > 1 \ \forall k \implies |\bar{x}| > 1 \ \forall \bar{x}$ punto di accumulazione della sequenza x^k. Ma l'unico punto stazionario della funzione obiettivo sopra definita è $x^* = 0$, pertanto la sequenza non raggiungerà mai la stazionarietà, poichè i suoi valori sono maggiori di 1 in modulo.
Abbiamo quindi mostrato che le ipotesi di stretta convessità della funzione obiettivo e di sequenza che garantisce uno stretto decremento della funzione obiettivo non danno alcuna garanzia di convergenza ad un ottimo.

Proposizione 3.2. Le condizioni

- f strettamente convessa
- $f(x^{k+1}) < f(x^k)$

non sono sufficienti a garantire la convergenza alla stazionarietà.

Chiariamo adesso che cosa si intende per convergenza "veloce" per una sequenza $\{x^k\}$

Definizione 3.3 (Tasso di convergenza)**.** Si dice che una sequenza $\{x^k\} \to x^*$ ha **Tasso di convergenza**

1. **Sublineare** se:
$$\lim \frac{||x^{k+1} - x^*||}{||x^k - x^*||} = 1$$
2. **Lineare** se:
$$\lim \frac{||x^{k+1} - x^*||}{||x^k - x^*||} = c \in (0, 1)$$
3. **Superlineare** se:
$$\lim \frac{||x^{k+1} - x^*||}{||x^k - x^*||} = 0$$
4. **Quadratico** se:
$$\lim \frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} = c < 1$$

Nota: La funzione
$$\frac{||x^{k+1} - x^*||}{||x^k - x^*||}$$

Misura la distanza tra il punto x^{k+1} e la soluzione, rapportata alla distanza tra il punto precedente x^k e la soluzione. Se il tasso di convergenza è sublineare, iterazioni successive dell'algoritmo comportano un aumento della distanza tra il valore corrente della sequenza e la soluzione. In particolare, aggiungere cifre significative alla soluzione è sempre più difficile, in quanto l'aggiunta di una nuova cifra ha un costo computazionale maggiore rispetto all'aggiunta della cifra precedente. Se il tasso di convergenza è lineare, la distanza è costante ad ogni passo ed aggiungere una cifra alla soluzione ha sempre un costo costante. Nel caso di tasso superlineare si ha una situazione inversa al caso lineare: l'aggiunta di una nuova cifra significativa è sempre meno dispendiosa rispetto all'aggiunta della cifra precedente. Il caso quadratico rappresenta il caso ideale per il tasso di convergenza(molto difficile raggiungere un tasso quadratico).

3.1 Classificazione algoritmi di ottimizzazione

Gli algoritmi di ottimizzazione si classificano in base alle informazioni a disposizione dell'algoritmo e al tipo di convergenza. In base alle informazioni a disposizione si hanno:

- **Algoritmi del primo ordine**: Sono note $f(x), \nabla f(x)$.
- **Algoritmi del secondo ordine**: Sono note $(f(x), \nabla f(x), \nabla^2 f(x))$
- **Algoritmi di ordine zero**: è nota solo la funzione obiettivo $f(x)$

In base al tipo di convergenza:

- **Convergenza globale**: La scelta di x^0 non influisce sulla convergenza alla soluzione:

$$\forall x^0 \in \mathbb{R}^n : \{x^k\}\text{ha punti di accumulazione stazionari}$$

- **Convergenza locale**: La scelta di x^0 può influire sulla convergenza alla soluzione: Sia $\bar{x} : \nabla f(\bar{x}) = 0$:

$$\exists \rho > 0 : \text{Se} x^0 \in \mathcal{B}_\rho(\bar{x}) \implies \{x^k\} \to \bar{x}$$

Le classi principali di algoritmi iterativi di ottimizzazione sono gli algoritmi **LineSearch**, **TrustRegion** e **DirectSearch**.

4 Algoritmi di tipo Line Search

Gli algoritmi di tipo Line Search sono algoritmi iterativi di ottimizzazione per i quali lo spostamento s_k che porta al valore successivo della sequenza $\{x^k\}$ è dato da uno

spostamento di un passo α_k lungo una direzione d_k, dove passo e direzione vengono ricalcolati ad ogni iterazione.

Algorithm 2 Schema generale algoritmo LineSearch

Require: $x^0 \in \mathbb{R}^n$
  $k = 0$
  while $\nabla f(x^k) \neq 0$ **do**
    Scelta di una direzione d_k
    Calcolo del passo $\alpha_k > 0$
    $x^{k+1} = x^k + \alpha_k d_k$
    $k++$
  end while

Definizione 4.1 (Funzione di forzamento)**.** Una funzione $\sigma : \mathbb{R}^+ \to \mathbb{R}^+$ si dice **Funzione di forzamento** se:

$$\forall \{x^k\} : \lim_{k\to\infty} \sigma(x^k) = 0 \implies \lim_{k\to\infty} x^k = 0$$

Proposizione 4.2 (Convergenza algoritmi LineSearch)**.** Sia $f \in C^1(\mathbb{R}^n)$ e $x^0 \in \mathbb{R}^n$: $\mathcal{L}_0$ è compatto. Sia poi $\{x^k\}$ tale che $x^{k+1} = x^k + \alpha_k d_k$. Se valgono le seguenti ipotesi:

1. $f(x^k) \geq f(x^{k+1})$
2. $\lim_{k\to\infty} \frac{\nabla f(x^k)^T d_k}{||d_k||} = 0$
3. Vale la **Condizione d'angolo**:

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

   Con σ funzione di forzamento

Allora valgono i seguenti risultati:

- (a): $\{x^k\} \subseteq \mathcal{L}_0$
- (b): $\{x^k\}$ ha punti di accumulazione
- (c): $f(x^k) \to \bar{f}$
- (d): $\lim_{k\to\infty} ||\nabla f(x^k)|| = 0$
- (e): Ogni punto di accumulazione della sequenza $\{x^k\}$ è stazionario

Proof. I punti (a),(b) e (c) sono già stati dimostrati per un caso più generale nella proposizione 3.1. Dimostriamo il punto (d).

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

per l'ipotesi (3) e per definizione di funzione di forzamento. Per l'ipotesi (2) risulta:

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

Per il teorema dei carabinieri

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

Poichè σ è funzione di forzamento:

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

□

Proof. Dimostriamo il punto (e).
Sia $K \subseteq \{0, 1, ...\}$ tale che

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

Risulta

$$\lim_{\substack{k\to\infty \\ k\in K}} ||\nabla f(x^k)|| = ||\nabla f(\bar{x})|| = \lim_{k\to\infty} ||\nabla f(x^k)|| = 0$$

Dove la prima uguaglianza è data dall'ipotesi $f \in C^1(\mathbb{R}^n)$ e la seconda è data dal punto (d) □

L'ipotesi (1) è ottenibile con scelte opportune di α_k e d_k, in particolare scegliendo α_k sufficientemente piccolo e d_k direzione di discesa. L'ipotesi (3) dipende dalla scelta di d_k, mentre l'ipotesi (2) dalla scelta di α_k. Analizziamo la condizione d'angolo (3): poniamo ad esempio $\sigma(t) = ct, \quad c > 0$, la condizione d'angolo diventa:

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

Scegliendo d_k come direzione di discesa si ha per definizione $\nabla f(x^k)^T d_k < 0$, da cui :

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

Ricordando che per definizione, se $u, v \in \mathbb{R}^n$:

$$\cos(\theta(u,v)) = \frac{u \cdot v}{||u||||v||}$$

La condizione d'angolo diventa

$$\cos(\theta(\nabla f(x^k), d_k)) \leq -c$$

che equivale a chiedere che la direzione d_k scelta al passo k formi un angolo maggiore di 90 gradi con il gradiente della funzione obiettivo nel punto x^k. Consideriamo ad esempio la direzione al passo k data dall'antigradiente della funzione obiettivo nel punto x^k: $d_k = -\nabla f(x^k)$. Risulta:

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

in quanto l'antigradiente in x^k forma un angolo di $\theta = \pi$ con il gradiente della funzione obiettivo nel punto $x^k \ \ \forall k$ e $\cos(\pi) = -1$.
Consideriamo adesso la direzione data da $d_k = -H_k \nabla f(x^k)$ dove $\forall k$ la matrice H_k è definita positiva e i suoi autovalori sono tali per cui $\lambda_{min} \geq m$ e $\lambda_{max} \leq M$ per qualche m, M. Risulta:

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

Poichè H_k è definita positiva. Applicando la minmax property e il bound sull'autovalore minimo di H_k si ha che

$$\nabla f(x^k)^T H_k \nabla f(x^k) \geq \lambda_{min}||\nabla f(x^k)||^2 \geq m||\nabla f(x^k)||^2 \implies \nabla f(x^k)^T d_k \leq -m||\nabla f(x^k)||^2$$

Calcoliamo $||d_k||$

$$||d_k|| = ||H_k \nabla f(x^k)|| \leq \lambda_{max}||\nabla f(x^k)|| \leq M||\nabla f(x^k)||$$

Da cui si ha

$$\frac{\nabla f(x^k)^T d_k}{||\nabla f(x^k)||||d_k||} \leq \frac{-m||\nabla f(x^k)||^2}{||\nabla f(x^k)||||d_k||} \leq \frac{-m||\nabla f(x^k)||^2}{M||\nabla f(x^k)||^2} = -\frac{m}{M}$$

Ponendo $c = \frac{m}{M}$ la direzione $d_k = -H_k \nabla f(x^k)$ rispetta la condizione d'angolo

4.1 Scelta del passo α_k: Ricerche di linea

Come si sceglie il passo α_k da prendere per spostarsi lungo la direzione di discesa d_k? Idealmente vorremmo prendere il passo α_k tale per cui si ottiene il massimo decremento possibile della funzione obiettivo lungo la direzione di discesa d_k:

$$\alpha_k = \arg\min_{\alpha>0} f(x^k + \alpha d_k)$$

Definiamo adesso la funzione $\phi(\alpha)$ come:

$$\phi(\alpha) = f(x^k + \alpha d_k) \implies \phi'(\alpha) = \nabla f(x^k + \alpha d_k)^T d_k$$

Da cui si ha che $\phi(0) = f(x^k)$ e $\phi'(0) = \nabla f(x^k)^T d_k$. Siamo quindi interessati a determinare il valore di α_k utilizzando un algoritmo linesearch. Gli algoritmi di tipo linesearch per la ricerca del passo α_k si dividono in

- **Linesearch Esatte**: $\alpha_k = \arg\min_{\alpha>0} \phi(\alpha)$
- **Linesearch Inesatte**: $\alpha_k \approx \arg\min_{\alpha>0} \phi(\alpha)$

In generale effettuare una linesearch esatta può essere un processo computazionalmente oneroso , a tale punto per cui può non valere la pena effettuare una linesearch esatta a fronte del decremento ottenuto al passo k dell'algoritmo. Questa osservazione ha valenza generale, ma esistono comunque casi in cui è possibile effettuare una linesearch esatta per α_k. Consideriamo il problema quadratico non vincolato e convesso(Q definita positiva):

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

Supponiamo di trovarci nel punto x^k e di voler effettuare uno spostamento lungo la direzione di discesa d_k. Consideriamo la funzione $\phi(\alpha)$ e applichiamo la formula di Taylor del secondo ordine(ricordando che $\nabla^2 f(x) = Q$:

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

Si nota che la funzione $\phi(\alpha$ rappresenta una parabola nella variabile α tale per cui il coefficiente del termine di grado massimo è positivo. Pertanto il minimo si realizza nel vertice:

$$\alpha^* = \frac{-b}{2a} = \frac{-\nabla f(x^k)^T d_k}{d_k^T Q d_k}$$

E' quindi possibile determinare il valore ottimo di α_k in forma chiusa $\implies$ si ha quindi linesearch esatta.

4.1.1 Linesearch Inesatte

In una linesearch inesatta si cerca un approssimazione del passo ottimo α_k che garantisce il massimo decremento della funzione obiettivo lungo una direzione di discesa d_k scelta al passo k. In particolare vogliamo garantire la proprietà di **Sufficiente decremento**:

$$f(x^k + \alpha_k d_k) \leq f(x^k) - \epsilon_k(\alpha_k); \quad \epsilon_k(\alpha_k) > 0$$

Un modo per garantire tale proprietà è imporre la **Condizione di Armijo** sul passo α_k. Siano $\gamma \in (0,1)$ e $d_k : \quad \nabla f(x^k)^T d_k < 0$:

$$f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k$$

In termini della funzione $\phi(\alpha)$ la condizione di Armijo si esprime come

$$\phi(\alpha_k) \leq \phi(0) + \gamma \alpha_k \phi'(0)$$

Per trovare un passo α_k che soddisfa la condizione di Armijo si utilizza l'algoritmo di ricerca di Armijo. Tale algoritmo consiste nel partire da un valore candidato $\alpha_0 > 0$ massimo per il passo α_k e di ridurre il valore α_0 ad ogni iterazione di un fattore $\delta \in (0,1)$ fino a quando il passo α_k è tale per cui la condizione di Armijo è rispettata.

Algorithm 3 Ricerca di Armijo

Require: $x^k, d_k \in \mathbb{R}^n : \nabla f(x^k)^T d_k < 0, \gamma \in (0,1), \delta \in (0,1), \alpha_0 > 0$

```
t = 0
while f(x^k + α_t d_k) > f(x^k) + γ α_t ∇f(x^k)^T d_k do
    α_{t+1} = δ α_t
    t++
end while
```

Proposizione 4.3 (Terminazione finita algoritmo di ricerca di Armijo)**.** Supponiamo $\nabla f(x^k)^T d_k < 0, \quad \alpha_0 > 0, \gamma, \delta \in (0,1)$. Si hanno i seguenti risultati:

- L'algoritmo di ricerca di Armijo produce in un numero finito di passi un valore α_k che soddisfa la condizione di Armijo
- Vale una delle seguenti proprietà:
  - $\alpha_k = \alpha_0$
  - $\alpha_k \leq \delta \alpha_0$ e $f(x^k + \frac{\alpha_k}{\delta} d_k) > f(x^k) + \gamma \frac{\alpha_k}{\delta} \nabla f(x^k)^T d_k$

Nota: La prima proprietà è equivalente ad affermare che il passo iniziale α_0 soddisfava già la condizione di Armijo, e quindi l'algoritmo termina restituendo α_0 senza aver effettuato alcuna iterazione. La seconda proprietà è equivalente ad affermare che è stata effettuata almeno 1 iterazione(in quanto $\delta\alpha_0 = \alpha_1$ e il passo $\alpha_{k-1} = \frac{\alpha_k}{\delta}$ trovato all'iterazione precedente non soddisfaceva la condizione di Armijo.

Proof. Supponiamo per assurdo che in un numero finito di passi l'algoritmo di ricerca di Armijo non riesca a produrre un valore del passo α_k che soddisfa la condizione di Armijo. Allora $\forall t = 0, 1, 2...$, ricordando che $\alpha_t = \delta^t\alpha_0$

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

Da cui si ottiene

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

Poichè $\delta \in (0, 1)$ si può osservare che per $t \to \infty$ il membro a sinistra tende alla derivata direzionale di f lungo d_k. Inoltre se $f \in C^1(\mathbb{R}^n)$ la derivata direzionale si può scrivere come prodotto scalare tra gradiente e direzione:

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

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

Si ha quindi un assurdo, poichè per ipotesi $\nabla f(x^k)^T d_k < 0$ ma $(1 - \gamma) > 0$ implica che $\nabla f(x^k)^T d_k \geq 0$ nella relazione sopra
Per il punto (2) sappiamo che per le istruzioni dell'algoritmo o accetto subito α_0 o il passo è stato ridotto almeno una volta, per cui $\alpha_k \leq \delta\alpha_0$. Inoltre alla penultima iterazione il controllo della condizione di Armijo non era stato superato, e il passo provato era α_k/δ, da cui:

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

□

Proposizione 4.4 (Convergenza metodi Line Search con ricerca di Armijo)**.** Sia $f \in C^1(\mathbb{R}^n)$, $x^0 \in \mathbb{R}^n : \mathcal{L}_0$ è compatto, e sia $\{x^k\}$ la sequenza prodotta da un algoritmo del tipo

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

Con $d_k : \nabla f(x^k)^T d_k < 0 \ \ \forall k$. Sia poi α_k il passo lungo d_k calcolato con una ricerca di

Armijo in cui il passo di prova iniziale α_0 soddisfa

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

con σ funzione di forzamento. Allora valgono

- (a) $f(x^{k+1}) < f(x^k)$
- (b) $\lim_{k\to\infty} \frac{|\nabla f(x^k)^T d_k|}{||d_k||} = 0$

Osservazione: Se usiamo una line search di tipo Armijo in un algoritmo di ottimizzazione line search abbiamo garanzia dei primi due punti della proposizione 4.2 sulla convergenza dei metodi di tipo line search.

Proof. Mostriamo che la condizione (a) di stretto decremento della funzione obiettivo è soddisfatta. Imponendo la condizione di Armijo su α_k si ha

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

Poichè $\gamma > 0, \quad \alpha_k > 0$ e $\nabla f(x^k)^T d_k < 0$ per ipotesi. □

Proof. Dimostriamo il punto (b)(**Warning: wall of text incoming!!)**
Per Armijo(scambiando i termini nella condizione):

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

Poiche $\nabla f(x^k)^T d_k < 0$ per ipotesi. Ora moltiplichiamo e dividiamo per $||d_k||$

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

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

Poichè per ipotesi $\gamma, \alpha_k > 0$(il resto dell'espressione è composto da norme e valori assoluti che sono ovviamente positivi).
Ora, $\{f(x^k)\}$ è decrescente per il punto (a) e l'insieme di livello $\mathcal{L}_0$ relativo al punto iniziale x^0 della sequenza $\{x^k\}$ è compatto, pertanto la sequenza $\{f(x^k)\}$ ha limite finito $\bar{f}$. Passiamo al limite per $k \to \infty$ la disuguaglianza (*), abbiamo che

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

In quanto per $k \to \infty f(x^k), f(x^{k+1}) \to \bar{f}$. Pertanto, dato che $\gamma\alpha_k||d_k|| > 0$, per il teorema dei carabinieri risulta

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

Consideriamo ora la sequenza

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

Poichè $\{x^k\} \subseteq \mathcal{L}_0$ è limitata e ∇f è continua, la sequenza $\{\nabla f(x^k)\}$ è limitata. Inoltre anche la sequenza $\{d_k/||d_k||\}$ è ovviamente limitata: ciò implica che la sequenza sopra definita ha punti di accumulazione.
Assumiamo ora che (b) sia falsa, ossia che:

$$\exists K \subseteq \{0, 1...\} : \lim_{\substack{k\to\infty \\ k\in K}} \frac{\nabla f(x^k)^T d_k}{||d_k||} = -\mu < 0$$

Inoltre poichè $\{x^k\}$ e $\{d_k/||d_k||\}$ sono limitate si ha che:

$$\exists K_1 \subseteq K : \lim_{\substack{k\to\infty \\ k\in K_1}} x^k = \bar{x}, \quad \lim_{\substack{k\to\infty \\ k\in K_1}} \frac{d_k}{||d_k||} = \bar{d}$$

Per la continuità del gradiente della funzione obiettivo si ha

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

Ricordando che tutte le sottosequenze di una sequenza convergente convergono allo stesso valore limite possiamo scrivere

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

Possiamo quindi dire che:

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

Possiamo adesso individuare due casi per α_k

1. $\exists \bar{k} \in K_1 : \quad \forall k \in K_1, \quad k \geq \bar{k} \implies \alpha_k = \alpha_0(k)$
2. $\exists K_2 \subseteq K_1 : \quad \alpha_k < \alpha_0(k) \quad \forall k \in K_2$

In pratica distinguiamo due possibili alternative: Nel primo caso possiamo individuare

un passo $\bar{k}$ dell'algoritmo line search per cui per tutti i passi successivi la ricerca di Armijo restituisce sempre il passo di prova iniziale(senza quindi effettuare iterazioni), mentre nel secondo caso ciò non accade, e quindi è possibile individuare una sottosequenza di valori k di tutti i passi dell'algoritmo in cui la ricerca di Armijo ha restituito un valore di α_k diverso dal valore iniziale α(e quindi sono state effettuate iterazioni della ricerca di Armijo).
Poniamoci adesso nel primo caso
Si ha che $\forall k > \bar{k}, \quad k \in K_1, \quad \alpha_k = \alpha_0(k)$. Applichiamo questa ipotesi:

$$\alpha_k ||d_k|| = \alpha_0 ||d_k|| \geq \frac{1}{||d_k||} \sigma \left(\frac{|\nabla f(x^k)^T d_k|}{||d_k||} \right) ||d_k|| \geq 0$$

per definizione di funzione di forzamento. Abbiamo quindi

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

Ricordando che $\lim_{\substack{k \to \infty \\ k \in K_1}} \alpha_k ||d_k|| = 0$ e applicando il teorema dei carabinieri si ha

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

Per definizione di funzione di forzamento si ha quindi

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

Ma avevamo assunto che (b) fosse falsa, e quindi che

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

Si ha quindi un assurdo(il limite non esiste)
Poniamoci adesso nel caso (2)
In questo caso si ha che $\alpha_k < \alpha_0(k) \ \ \forall k \in K_2$. Per la proposizione 4.3 sulla terminazione finita della ricerca di Armijo si ha

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

Per il teorema della media

$$f(x^k + \frac{\alpha_k}{\delta} d_k) = f(x^k) + \frac{\alpha_k}{\delta} \nabla f(\xi_k)^T d_k$$

con $\xi_k = x^k + t_k \frac{\alpha_k}{\delta} d_k, \quad t_k \in (0,1)$ Da cui possiamo scrivere

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

$$f(x^k + \frac{\alpha_k}{\delta} d_k) - f(x^k) = \frac{\alpha_k}{\delta} \nabla f(\xi_k)^T d_k$$

Componendo le due relazioni otteniamo

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

Dividendo per $||d_k||$:

$$\frac{\nabla f(\xi_k)^T d_k}{||d_k||} > \gamma \frac{\nabla f(x^k)^T d_k}{||d_k||}$$

Per $k \to \infty, \quad k \in K_2$

$$\frac{d_k}{||d_k||} \to \bar{d}, \quad x^k \to \bar{x}$$

Inoltre

$$\lim_{\substack{k \to \infty \\ k \in K_2}} \xi_k = \lim_{\substack{k \to \infty \\ k \in K_1}} x^k + t_k \frac{\alpha_k}{\delta} d_k = \bar{x}$$

Quindi per $k \to \infty, \quad k \in K_2$

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

Ma $(1-\gamma) > 0$ e sotto le condizioni di limite$\nabla f(\bar{x})^T \bar{d} \to -\mu < 0$, per cui si ha un assurdo □

5 Metodo del Gradiente

Algorithm 4 Metodo del Gradiente

Require: $x^0 \in \mathbb{R}^n$

$k = 0$

while $\nabla f(x^k) \neq 0$ **do**

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

Calcola il passo α_k lungo d_k mediante una ricerca di Armijo

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

$k++$

end while

Proposizione 5.1 (Convergenza verso la stazionarietà per il Metodo del Gradiente)**.** Sia $f \in C^1(\mathbb{R}^n), \quad x^0 \in \mathbb{R}^n$ tale che l'insieme di livello $\mathcal{L}_0$ è compatto.

$\implies$ Il metodo del Gradiente genera una sequenza $\{x^k\}$ che ammette punti di accumulazione, ognuno dei quali è stazionario.

Proof. Segue dalla proposizione 4.2 sulla convergenza dei metodi Line search. In particolare l'ipotesi (1) di decremento della funzione obiettivo sulla sequenza $\{x^k\}$ e l'ipotesi (2) della proposizione 4.2 sono soddisfatte per le proprietà degli algoritmi Line Search con ricerca di Armijo per la scelta del passo α_k ad ogni iterazione(vedi proposizione 4.4), e l'ipotesi (3), ovvero la condizione d'angolo, segue (come visto nell'esempio successivo alla proposizione 4.2) dalla scelta dell'antigradiente di f in x^k come direzione di discesa al passo k. □

Proposizione 5.2. Il metodo del gradiente ha un tasso di convergenza sublineare

Definizione 5.3 (Funzione fortemente convessa)**.** Una funzione $f : \mathbb{R}^n \to \mathbb{R}$ si dice **fortemente convessa** se $\exists g : \mathbb{R}^n \to \mathbb{R}$ convessa tale che:

$$f(x) = g(x) + \mu||x||^2 \quad \forall x \in \mathbb{R}^n, \mu > 0$$

Proposizione 5.4. Sia $f : \mathbb{R}^n \to \mathbb{R}$ fortemente convessa.
$\implies$ il metodo del gradiente ha tasso di convergenza lineare

Ci chiediamo adesso se il metodo del gradiente possa convergere anche con passi costanti. Vedremo che sotto alcune ipotesi la risposta è si.

Definizione 5.5 (Lipshitz-continuità)**.** Una funzione $f : \mathbb{R}^n \to \mathbb{R}^m$ si dice **Lipshitz-continua** su $\mathbb{R}^n$ se $\exists L > 0$ tale che:

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

Se $f \in C^2(\mathbb{R}^n)$ possiamo dare una definizione alternativa di lipshitz-continuità che coinvolge l'hessiana della funzione:

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

La più piccola costante L che realizza la disuguaglianza nella definizione di lipshitz-continuità si dice **costante di Lipshitz** della funzione f.

5.1 Metodo del Gradiente a passo costante

Nel metodo del gradiente a passo costante la formula di aggiornamento dei valori della sequenza $\{x^k\}$ è data da:

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

dove L è la costante di Lipshitz della funzione ∇f.

Proposizione 5.6. Sia $f \in C^2(\mathbb{R}^n)$ e ∇f Lipshitz-continuo. Allora $\forall x \in \mathbb{R}^n$, $\forall d \in \mathbb{R}^n$ e $\forall t \geq 0$ vale:

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

Proof. Per Taylor:

$$f(x+td) = f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(y) d \leq f(x) + t\nabla f(x)^T d + \left| \frac{t^2}{2} d^T \nabla^2 f(y) d \right|$$

Applicando la definizione di Lipshitz continuità per $f \in C^2(\mathbb{R}^n)$ si ha:

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

□

Proposizione 5.7 (Convergenza alla stazionarietà metodo del gradiente a passo costante)**.** Sia $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(\mathbb{R}^n)$, ∇f Lipshitz-continuo con costante L e $x^0 \in \mathbb{R}^n$ tale che l'insieme di livello $\mathcal{L}_0$ è compatto.
$\implies$ il metodo del gradiente con passo $\alpha_k = \frac{1}{L}$ genera una sequenza $\{x^k\}$ tale che:

1. $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}||\nabla f(x^k)||^2 \quad \forall k$
2. La sequenza $\{x^k\}$ ha punti di accumulazione, ognuno dei quali è stazionario

Proof. Dimostriamo il punto (1)
Per la proposizione 5.6 vale:

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

Ora nel metodo del gradiente a passo costante abbiamo $d_k = -\nabla f(x^k)$ e $\alpha_k = \frac{1}{L}$, per cui sostituendo si ha:

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

Da cui ricostruendo la catena di disuguaglianze possiamo ottenere

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

□

Proof. Dimostriamo il punto (2)
Per il punto (1) abbiamo

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

Ora osserviamo che $\{f(x^k)\}$ è monotona decrescente e che la sequenza $\{x^k\}$ ha punti di accumulazione in quanto per ipotesi $\{x^k\} \subseteq \mathcal{L}_0$ compatto. Quindi possiamo affermare che $f(x^k) \to f^*$ finito.
Sia $\bar{x}$ un qualsiasi punto di accumulazione della sequenza $\{x^k\}$:

$$\exists K \subseteq \{0, 1, ...\} : \lim_{\substack{k\to\infty \\ k\in K}} x^k = \bar{x}$$

Passiamo al limite per $k \to \infty, k \in K$ nella disuguaglianza (*):

$$\lim_{\substack{k\to\infty \\ k\in K}} f(x^{k+1}) - f(x^k) \leq \lim_{\substack{k\to\infty \\ k\in K}} -\frac{1}{2L}||\nabla f(x^k)||^2 \leq 0$$

Osserviamo poi che $\lim_{\substack{k\to\infty \\ k\in K}} f(x^{k+1}) - f(x^k) \to f^* - f^* = 0$ e che $\lim_{\substack{k\to\infty \\ k\in K}} -\frac{1}{2L}||\nabla f(x^k)||^2 \to -\frac{1}{2L}||\nabla f(\bar{x})||^2 \leq 0$. Per il teorema dei carabinieri si ha

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

Abbiamo mostrato che per un generico punto di accumulazione $\bar{x}$ della sequenza $\{x^k\}$ la norma del gradiente nel punto è 0, ossia il punto è stazionario. □

I problemi principali dell'algoritmo del gradiente a passo costante consistono nel non avere una funzione obiettivo due volte continuamente differenziabile, e nel non riuscire a calcolare la costante di Lipshitz L per determinate funzioni obiettivo. Se f è due volte continuamente differenziabile e non si riesce a calcolare L è possibile scegliere $\alpha \in \mathcal{B}_\rho(L)$ e mantenere le proprietà di convergenza viste.

6 Complessità degli algoritmi di ottimizzazione

In questa sezione ipotizziamo che:

- f sia limitata inferiormente da f^*
- f convessa
- ∇f Lipshitz continuo

Definizione 6.1 (Iteration error)**.** Si dice che un algoritmo ha **Iteration error** $\mathcal{O}(h(k))$ se vale

$$f(x^k) - f^* = \mathcal{O}(h(k))$$

Definizione 6.2 (Iteration complexity)**.** Si dice che un algoritmo ha **Iteration complexity** $\mathcal{O}(\hat{h}(\epsilon))$ se vale:

$$\min\{k : f(x^k) - f^* \leq \epsilon\} = \mathcal{O}(\hat{h}(\epsilon))$$

L'iteration error rappresenta l'errore che viene compiuto nell'approssimare il valore obiettivo f^* con il valore della funzione obiettivo calcolata nel k-esimo punto della sequenza $\{x^k\}$ in una data iterazione k.
L'iteration complexity rappresenta il numero minimo di iterazioni necessario per ottenere una precisione fissata(numero di cifre significative) per il valore obiettivo Osservazione:

- Iteration complexity $\mathcal{O}(\frac{1}{\epsilon}) \iff$ Iteration error $\mathcal{O}(\frac{1}{k})$
- Iteration complexity $\mathcal{O}(\frac{1}{\sqrt{\epsilon}}) \iff$ Iteration error $\mathcal{O}(\frac{1}{k^2})$

Mostriamo ora con un esempio numerico di come un algoritmo con iteration complexity $\mathcal{O}(\frac{1}{\epsilon})$ comporti un costo esponenziale per l'aggiunta di una nuova cifra significativa(non utilizzabile nella pratica)

ϵ	$\log(\frac{1}{\epsilon})$	costo per raggiungere un numero prefissato di cifre se complessità $\mathcal{O}(\frac{1}{\epsilon})$
0.1	1	10
0.01	2	100
0.001	3	1000
0.0001	4	10000

Come si può notare dalla tabella, il costo per raggiungere un numero fissato di cifre significative nel caso in cui l'iteration complexity dell'algoritmo sia $\mathcal{O}(\frac{1}{\epsilon})$ cresce esponenzialmente all'aumentare del numero di cifre richieste. Ciò implica che il costo necessario ad ottenere la cifra successiva sarà sempre maggiore della somma di tutti i costi necessari per ottenere le precedenti cifre.
Vediamo adesso come diversi tipi di iteration complexity comportano costi diversi per l'aggiunta di una nuova cifra significativa, e le implicazioni che questi costi hanno sul tasso di convergenza dell'algoritmo.

Iteration complexity	Iteration error	Costo per nuova cifra significativa
$\mathcal{O}(\frac{1}{\epsilon})$	$\mathcal{O}(\frac{1}{k})$	esponenziale $\implies$ tasso sublineare
$\mathcal{O}(\log(\frac{1}{\epsilon}))$	$\mathcal{O}(\rho^k), \rho < 1$	polinomiale $\implies$ tasso lineare
$\mathcal{O}(\log(\log(\frac{1}{\epsilon})))$	$\mathcal{O}(\rho^{2^k})$	costante $\implies$ tasso superlineare

L'iteration error è legato al tasso di convergenza di $\{f(x^k)\}$.Consideriamo il caso in cui l'iteration error è $\mathcal{O}(\frac{1}{k})$:

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

Si ha che il tasso di convergenza della sequenza $\{f(x^k)\}$ è sublineare. Consideriamo adesso il caso di iteration error $\mathcal{O}(\rho^k), \rho < 1$

$$\lim_{k\to\infty} \frac{f(x^{k+1}) - f^*}{f(x^k) - f^*} \approx \lim_{k\to\infty} \frac{\rho^{k+1}}{\rho^k} = \rho < 1$$

Si ha che il tasso di convergenza della sequenza $\{f(x^k)\}$ è lineare(come da tabella). Analogamente si può far vedere che ad un iteration error di $\mathcal{O}(\rho^{2^k})$ corrisponde un tasso di convergenza costante per la sequenza $\{f(x^k)\}$.

Proposizione 6.3 (Complessità del metodo del gradiente a passo costante)**.** Sia $f \in C^2(\mathbb{R}^n)$, ∇f Lipshitz continuo con costante L e $x^0 \in \mathbb{R}^n$ tale che l'insieme di livello $\mathcal{L}_0$ è compatto. Inoltre sia f convessa con valore minimo $f(x^*) = f^*$ e sia $\alpha_k = \frac{1}{L}$.
$\implies$ Il metodo del gradiente a passo costante genera una sequenza $\{x^k\}$ tale che:

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

Cioè ha iteration error(complexity)

$$\mathcal{O}\left(\frac{1}{k}\right) \iff \mathcal{O}\left(\frac{1}{\epsilon}\right)$$

Proof. Ricordiamo la proposizione 5.7 sul metodo del gradiente a passo costante:

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

Per ipotesi f è convessa, da cui si ha

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

(f è convessa in un punto se è maggiore della retta tangente ad f in quel punto).Possiamo

quindi ricavare

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

Otteniamo quindi la disuguaglianza

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

Da cui raccogliendo una costante $\frac{L}{2}$ possiamo ottenere:

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

Completiamo il binomio aggiungendo e sottraendo la quantità $||x^* - x^k||^2$:

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

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

Ricordando che

$$||a + b||^2 = a^Ta + b^Tb + 2a^Tb = ||a||^2 + ||b||^2 + 2a^Tb$$

e ponendo $(x^* - x^k) = a$ e $\frac{1}{L}\nabla f(x^k) = b$ si ha quindi

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

Sostituendo nella disuguaglianza otteniamo

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

$$= \frac{L}{2}\left(||x^* - x^k||^2 - ||x^* - (x^k - \frac{1}{L}\nabla f(x^k))||^2\right)$$

Osserviamo che per la formula di aggiornamento del metodo del gradiente a passo costante $x^k - \frac{1}{L}\nabla f(x^k) = x^{k+1}$. Sostituendo otteniamo

$$\frac{L}{2}\left(||x^* - x^k||^2 - ||x^* - x^{k+1}||^2\right)$$

Ricostruendo la catena di disuguaglianze si ottiene

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

Abbiamo determinato un upper bound per l'iteration error alla $k+1$-esima iterazione. Consideriamo ora la somma degli iteration error su k iterazioni complessive:

$$\sum_{t=0}^{k} f(x^{t+1}) - f(x^*) \leq \sum_{t=0}^{k} \frac{L}{2}\left(||x^* - x^t||^2 - ||x^* - x^{t+1}||^2\right)$$

A questo punto dobbiamo notare che la sommatoria al membro destro è la somma parziale di una serie telescopica, pertanto è esprimibile come differenza tra il primo e l'ultimo termine della sequenza.

$$\sum_{t=0}^{k} \frac{L}{2}\left(||x^* - x^t||^2 - ||x^* - x^{t+1}||^2\right) = \frac{L}{2}\left(||x^* - x^0||^2 - ||x^* - x^{k+1}||^2\right) \leq \frac{L}{2}(||x^* - x^0||^2)$$

Considerando le disuguaglianze ottenute si ha quindi

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

Inoltre $\{f(x^k)\}$ è monotona decrescente, per cui vale che $f(x^{t+1}) \geq f(x^{k+1}) \ \ \forall t \leq k \implies f(x^{t+1}) - f(x^*) \geq f(x^{k+1}) - f(x^*) \ \ \forall t \leq k$.Poichè nella somma parziale $t \in \{0..k\}$ possiamo scrivere:

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

Poichè la somma $\sum_{t=0}^{k} f(x^{k+1}) - f(x^*)$ non dipende da t possiamo scrivere:

$$\sum_{t=0}^{k} f(x^{k+1}) - f(x^*) = (k+1)(f(x^{k+1}) - f(x^*))$$

Sostituendo nell'ultima disuguaglianza ottenuta si ha (finalmente direi) la tesi

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

□

Proposizione 6.4. Se f è fortemente convessa il metodo del gradiente ha complessità $\mathcal{O}(\log(\frac{1}{\epsilon}))$

Questo fatto è particolarmente utile nel caso di problemi di apprendimento automatico:

$$\min_w \mathcal{L}(w) + \lambda\Omega(w)$$

Se la loss $\mathcal{L}(w)$ è convessa e si utilizza un regolarizzatore quadratico $\Omega(w) = ||w||^2$ la funzione obiettivo è fortemente convessa per definizione.

7 Metodi con termini di tipo momentum

Nei metodi di tipo momentum si aggiunge un termine dipendente dall'iterazione precedente per tentare di migliorare le proprietà di convergenza dell'algoritmo

7.1 Metodo Heavy-ball

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

Nel caso quadratico esiste una formula per la scelta di α_k e β_k ottimali, in tal caso si ha una convergenza lineare con costanti migliori di quelle del metodo del gradiente. Quest'ultimo risultato è generalizzabile al caso di funzione obiettivo due volte continuamente differenziabile, fortemente convessa e avente gradiente Lipshitz-continuo. L'aggiunta di un termine di tipo momentum aiuta ad alleviare oscillazioni e a diminuire il numero di passi in zone a bassa curvatura.(se la curvatura è bassa un metodo classico effettua molti più passi).

7.2 Metodo del gradiente accelerato di Nesterov

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

è la combinazione di un passo di gradiente puro e di un passo momentum puro(dove il gradiente non influisce). In nesterov il gradiente al passo corrente viene calcolato sul termine momentum. Il metodo del gradiente accelerato di Nesterov su funzioni convesse ha complessità

$$\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right) \iff \mathcal{O}\left(\frac{1}{k^2}\right)$$

Questa è la complessità ottimale per un metodo del primo ordine. L'algoritmo del gradiente accelerato di Nesterov è quindi teoricamente ottimale. Tuttavia dobbiamo notare che tale complessità comporta un tasso di convergenza della sequenza dei valori obiettivo $\{f(x^k)\}$ sublineare. Nel caso fortemente convesso si ha la stessa complessità

del metodo del gradiente ma con costanti migliori. Il metodo di Nesterov non ha riscontri pratici, e rilassando le ipotesi sulla funzione obiettivo il metodo potrebbe divergere.

8 Metodo del Gradiente Coniugato

Il metodo del gradiente coniugato è un algoritmo iterativo che risulta particolarmente efficiente nella risoluzione di problemi di ottimizzazione quadratici con funzioni obiettivo strettamente convesse e sistemi lineari di equazioni. In questa sezione verdremo il metodo del gradiente coniugato lineare, utilizzato nei casi descritti sopra, e il metodo del gradiente coniugato di Fletcher-Reeves, che utilizza i principi base del metodo del gradiente coniugato per minimizzare funzioni obiettivo non lineari generiche.

Definizione 8.1. Sia $Q \in \mathbb{R}^{n\times n}$ simmetrica e definita positiva. Un insieme di direzioni $\{d_0, ..., d_{m-1}\}, d_i \in \mathbb{R}^n, m \leq n$ si dice insieme di direzioni **mutuamente coniugate** rispetto a Q se $\forall i, j = 0, ..., m-1$ con $i \neq j$ si ha

$$d_i^T Q d_j = 0$$

Proposizione 8.2 (Direzioni mutuamente coniugate e indipendenza lineare)**.** Sia $\{d_0, ..., d_{m-1}\}$ un insieme di direzioni mutuamente coniugate rispetto ad una qualche matrice $Q \in \mathbb{R}^{n\times n}$. Le direzioni $\{d_0, ..., d_{m-1}\}$ sono linearmente indipendenti.

Proof. Scriviamo una generica combinazione lineare nulla delle direzioni mutuamente coniugate (ricordiamo che un insieme di vettori linearmente indipendenti è tale per cui ogni possibile combinazione lineare si annulla solo quando tutti gli scalari sono nulli).

$$\sum_{j=0}^{m-1} \alpha_j d_j$$

Sia $i \in \{0, 1..., m-1\}$, moltiplichiamo per $d_i^T Q$

$$\sum_{j=0}^{m-1} \alpha_j d_i^T Q d_j = 0$$

Ora per ipotesi di direzioni mutuamente coniugate rispetto a Q si ha $d_i^T Q d_j = 0 \ \ \forall i \neq j$. L'unico termine che rimane nella combinazione lineare è il termine per $j = i$:

$$\alpha_i d_i^T Q d_i = 0$$

Poichè Q è definita positiva $d_i^T Q d_i > 0$, da cui ricaviamo che necessariamente $\alpha_i = 0$. Per l'arbitrarietà dell'indice i si ha la tesi. □

Vediamo adesso come sfruttare le direzioni mutuamente coniugate per derivare un metodo in grado di risolvere problemi quadratici. Consideriamo il problema:

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

Ricordiamo innanzitutto che $\nabla f(x) = Qx - c$ e in particolare nel punto di minimo x^* vale $\nabla f(x^*) = Qx^* - c = 0 \implies Qx^* = c$.Consideriamo ora un insieme di n direzioni mutuamente coniugate $\{d_0, ..., d_{n-1}\}$, abbiamo visto che esse formano un insieme di vettori linearmente indipendenti, per cui possiamo esprimere il punto di minimo come combinazione lineare a coefficienti reali delle $\{d_0, ..., d_{n-1}\}$:

$$x^* = \sum_{s=0}^{n-1} \alpha_s^* d_s \implies Qx^* = \sum_{s=0}^{n-1} \alpha_s^* Q d_s$$

$$d_i^T Q x^* = \sum_{s=0}^{n-1} \alpha_s^* d_i^T Q d_s = \alpha_i^* d_i^T Q d_i$$

Poichè per $i \neq s \;\; d_i^T Q d_s = 0$. Ora dato che $\nabla f(x^*) = Qx^* - c = 0 \implies Qx^* = c$ possiamo scrivere:

$$d_i^T c = \alpha_i^* d_i^T Q d_i \implies \alpha_i^* = \frac{d_i^T c}{d_i^T Q d_i}$$

Riscrivendo l'espressione appena determinata per gli scalari nell'espressione di x^* come combinazione lineare di direzioni mutuamente coniugate si ha

$$x^* = \sum_{s=0}^{n-1} \frac{d_s^T c}{d_s^T Q d_s} d_s$$

Scriviamo $x^* - x^0$ come combinazione lineare delle direzioni mutuamente coniugate $\{d_0, ..., d_{n-1}\}$:

$$x^* - x^0 = \sum_{s=0}^{n-1} \alpha_s d_s \implies x^* = x^0 + \alpha_0 d_0 + \alpha_1 d_1 + ... + \alpha_{n-1} d_{n-1}$$

Si può interpretare la formula come un processo iterativo in cui

$$x^0 = x^0, \;\; x^1 = x^0 + \alpha_0 d_0, \;\; x^2 = x^0 + \alpha_0 d_0 + \alpha_1 d_1 ... \implies$$

$$\implies x^k = x^0 + \sum_{s=0}^{k-1} \alpha_s d_s$$

Consideriamo le quantità:

$$d_k^T Q(x^k - x^0) = d_k^T Q(\sum_{s=0}^{k-1} \alpha_s d_s) = \sum_{s=0}^{k-1} \alpha_s d_k^T Q d_s = 0$$

dato che $d_k^T Q d_s = 0 \ \ \forall s = 0...k-1$

$$d_k^T Q(x^* - x^0) = \sum_{s=0}^{n-1} \alpha_s d_k^T Q d_s = \alpha_k d_k^T Q d_k$$

Poichè l'unico termine diverso da zero nella somma è il termine per cui $s = k$. Date queste quantità possiamo riscrivere l'espressione del passo α_k nel processo iterativo:

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

Il secondo membro della somma è uguale a zero, poichè abbiamo mostrato che $d_k^T Q(x^k - x^0) = 0$.Possiamo quindi scrivere

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

dove abbiamo sfruttato il fatto che $c - Qx^k = -\nabla f(x^k)$ e che $Qx^* = c$ per x^* soluzione del problema quadratico. Notiamo infine che il risultato ottenuto per il passo α_k è lo stesso ottenuto effettuando una ricerca di linea esatta lungo la direzione d_k per un problema quadratico. Pertanto il passo α_k determinato rappresenta il passo che comporta il maggior decremento possibile della funzione obiettivo lungo la direzione d_k. Abbiamo quindi determinato un processo iterativo, detto **metodo delle direzioni coniugate**, per la soluzione di problemi quadratici:

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

Proposizione 8.3 (Convergenza finita metodo delle direzioni mutuamente coniugate su problemi quadratici)**.** Siano $\{d_0, .., d_{m-1}\}$ direzioni mutuamente coniugate e sia $x^0 \in \mathbb{R}^n$. Consideriamo il metodo iterativo $x^{k+1} = x^k + \alpha_k d_k$ dove $\alpha_k = -\frac{g_k^T d_k}{d_k^T Q d_k}$. Allora vale

- $\forall k$: si ha che $g_k^T d_i = 0 \ \ \forall i = 0...k-1$
- $\exists m \leq n$ tale che $x^m = x^*$

(il gradiente è ortogonale a ciascuna delle k-1 direzioni mutuamente coniugate calcolate alla k-esima iterazione, e in un numero di iterazioni minore o uguale al numero totale di direzioni mutuamente coniugate l'algoritmo converge esattamente all'ottimo.

Proof. Dimostriamo il punto (1):

$$x^k = x^i + \sum_{s=i}^{k-1} \alpha_s d_s$$

per $i \in \{0, .., k-1\}$. Consideriamo la quantità

$$g_k^T d_i = (Qx^k - c)^T d_i = (Qx^i + \sum_{s=i}^{k-1} \alpha_s Q d_s - c)^T d_i = (Qx^i - c)T d_i + (\sum_{s=i}^{k-1} \alpha_s d_i^T Q d_s)$$

Ora ricordando che $g_i^T = (Qx^i - c)^T$ e che $d_i^T Q d_s = 0 \ \ \forall i \neq s$ poichè le direzioni sono mutuamente coniugate si ha:

$$g_i^T d_i + \alpha_i d_i^T Q d_i$$

Ponendo $\alpha_i = \frac{-d_i^T g_i}{d_i^T Q d_i}$ si ottiene

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

Riconsiderando tutte le uguaglianze abbiamo ottenuto

$$g_k^T d_i = 0, \quad i \in \{0, ..., k-1\}$$

□

Proof. Dimostriamo il punto (2):
Supponiamo di aver effettuato n iterazioni ottenendo x^n: per il punto (1)

$$x^n : \quad g_n^T d_i = 0 \ \ \forall i = 0, .., n-1$$

Poichè $d_0, ..., d_{n-1}$ sono direzioni linearmente indipendenti si ha che $g_n = \nabla f(x^n) = 0$, ossia x^n è soluzione del problema quadratico. □

8.1 Metodo del gradiente coniugato

Fin'ora abbiamo supposto di avere un insieme di n direzioni mutuamente coniugate e abbiamo costruito un metodo iterativo avente convergenza finuta per il calcolo della soluzione di un problema quadratico, derivando l'espressione esatta del passo α_k che

causa il maggior decremento possibile lungo la direzione d_k. Scegliendo:($g_k = \nabla f(x^k)$)

$$d_{k+1} = -g_{k+1} + \beta_{k+1} d_k \text{con} \beta_{k+1} = \frac{||g_{k+1}||^2}{||g_k||^2}$$

si può dimostrare che la direzione al passo k+1 calcolata secondo questa formula è mutuamente coniugata con le direzioni $d_0, ..., d_k$ calcolate nei passi precedenti.

Algorithm 5 Metodo del gradiente coniugato

Require: $x^0 \in \mathbb{R}^n, Q \in \mathbb{R}^{n \times n}, c \in \mathbb{R}^n$

$k = 0$

$d_0 = -g_0$

while $||g_k|| \neq 0$ **do**

$\alpha_k = \frac{-g_k^T d_k}{d_k^T Q d_k}$

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

$\beta_{k+1} = \frac{||g_{k+1}||^2}{||g_k||^2}$

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

$k++$

end while

Osserviamo che è possibile risparmiare risorse di calcolo considerando

$$\alpha_k = \frac{-g_k^T d_k}{d_k^T Q d_k} = \frac{-g_k^T(-g_k + \beta_k d_{k-1})}{d_k^T Q d_k} = \frac{||g_k||^2 - \beta_k g_k^T d_{k-1}}{d_k^T Q d_k} = \frac{||g_k||^2}{d_k^T Q d_k}$$

Poichè $g_k^T d_{k-1} = 0$ per il punto (1) della proposizione 8.3. Inoltre:

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

dato che $g_k = Qx^k - c$.

Il metodo del gradiente coniugato può essere generalizzato al caso di una funzione non lineare generica. Il termine β_{k+1} ha diverse formulazioni che sono equivalenti nel caso quadratico, ma non lo sono nel caso di funzione obiettivo non lineare generica. In ogni caso utilizziamo la formula di **Fletcher-Reeves**: $\beta_{k+1} = \frac{||g_{k+1}||^2}{||g_k||^2}$.

Ponendoci nel caso in cui la funzione obiettivo è una funzione non lineare generica, dobbiamo garantire che la direzione d_k scelta al passo k sia una direzione di discesa: dobbiamo imporre quindi che $d_{k+1}^T g_{k+1} < 0$. Sostituendo a d_{k+1} la formula iterativa del metodo del gradiente coniugato otteniamo:

$$(-g_{k+1} + \beta_{k+1} d_k)^T g_{k+1} < 0 \implies -||g_{k+1}||^2 + \beta_{k+1} d_k^T g_{k+1} < 0$$

Sostituendo $g_{k+1} = \nabla f(x^{k+1}) = \nabla f(x^k + \alpha_k d_k)$ otteniamo la condizione

$$-||\nabla f(x^k + \alpha_k d_k)||^2 + \frac{||\nabla f(x^k + \alpha_k d_k)||^2}{||\nabla f(x^k)||^2} d_k^T \nabla f(x^k + \alpha_k d_k) < 0$$

La garanzia di valenza di questa proprietà dipende dalla scelta del passo α_k. Tuttavia in questo caso una ricerca di tipo Armijo non è sufficiente a garantire la proprietà voluta. Sono necessarie delle condizioni più forti:

8.2 Condizioni di Wolfe

8.2.1 Condizione di Wolfe debole

Siano $\gamma, \sigma \in \mathbb{R}$ tali che $\sigma, \gamma \in (0, 1)$. La **Condizione di Wolfe debole** è data dall'insieme di disuguaglianze

$$f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k d_k^T \nabla f(x^k) \quad (\textbf{Condizione di Armijo})$$

$$\nabla f(x^k + \alpha_k d_k)^T d_k \geq \sigma \nabla f(x^k)^T d_k \quad (\textbf{Condizione di curvatura})$$

Solitamente si sceglie γ abbastanza piccolo e σ più grande, ad esempio per il metodo del gradiente coniugato non lineare Nocedal e Wright hanno utlizzato $\gamma = 10^{-4}$ e $\sigma = 0.9$. Nella condizione di Wolfe debole, la condizione di Armijo garantisce la proprietà di sufficiente decremento della funzione obiettivo, mentre la condizione di curvatura assicura il sufficiente decremento del gradiente della funzione obiettivo. Le condizioni che compongono la condizione di Wolfe debole possono essere interpretate rispettivamente come un limite superiore ed inferiore al passo α_k. Notiamo che se $\phi(\alpha) = f(x^k + \alpha d_k)$ con x^k e α_k fissati per l'iterazione corrente, un algoritmo di ricerca di linea di tipo Wolfe debole può terminare con un valore di α_k che non è sufficientemente vicino ad un punto di minimo di $\phi(\alpha)$.

8.2.2 Condizione di Wolfe forte

La **Condizione di Wolfe forte**:

$$f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k d_k^T \nabla f(x^k)$$

$$|\nabla f(x^k + \alpha_k d_k)^T d_k| \leq \sigma |\nabla f(x^k)^T d_k|$$

forza un algorimo di ricerca di linea di tipo Wolfe forte a convergere ad un punto α_k che appartiene ad un intorno di un punto stazionario di $\phi(\alpha) = f(x^k + \alpha d_k)$

9 Metodi del secondo ordine

Nei metodi iterativi di ottimizzazione del secondo ordine si suppone di avere a disposizione informazioni del secondo ordine sulla funzione obiettivo, ossia di conoscere $f, \nabla f, \nabla^2 f$. I metodi del secondo ordine per l'ottimizzazione non vincolata richiedono quindi che la funzione obiettivo sia due volte continuamente differenziabile su tutto $\mathbb{R}^n$. Utilizziamo l'informazione del secondo ordine per costruire un modello quadratico che approssimi la funzione obiettivo in un punto $x \in \mathbb{R}^n$:

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

Se l'hessiana $\nabla^2 f(x^k)$ è definita positiva, la funzione che descrive il modello quadratico di approssimazione $m_k(x)$ è strettamente convessa, pertanto possiamo determinare il minimo imponendo $\nabla m_k(x) = 0$

$$\nabla m_k(x) = \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k)$$

Sia $\bar{x} \in \mathbb{R}^n$ il punto che soddisfa $\nabla m_k(\bar{x}) = 0$, si ha quindi:

$$\nabla^2 f(x^k)(\bar{x} - x^k) = -\nabla f(x^k) \implies \bar{x} - x^k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$$

Da cui ricaviamo un'espressione in forma chiusa per il punto di minimo del modello quadratico di approssimazione

$$\bar{x} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$$

Osserviamo che $\bar{x}$ è il punto di minimo dell'approssimazione quadratica $m_k(x)$ della funzione obiettivo nel punto x, che è in generale diverso dal punto di minimo di $f(x)$. Con un ragionamento simile a quello visto per il metodo delle direzioni mutuamente coniguate, possiamo interpretare l'espressione in forma chiusa come un processo iterativo:

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

Questo processo iterativo basato sull'approssimazione quadratica della funzione obiettivo è detto **Metodo di Newton**. Esso è fondamentalmente una ricerca di linea dove:

- $x^{k+1} = x^k + \alpha_k d_k$
- $\alpha_k = 1$
- $d_k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

E' facile mostrare che la direzione d_k del metodo di Newton è una direzione di discesa per f in x^k se l'hessiana della funzione obiettvo nel punto x^k è definita positiva:

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

Notiamo quindi che $\nabla f(x^k)^T [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) > 0 \iff [\nabla^2 f(x^k)]^{-1}$ è definita positiva $\iff \nabla^2 f(x^k)$ è definita positiva.
Inoltre è possibile mostrare mediante un controesempio(minimizzazione non vincolata della funzione $f(x) = \sqrt{1+x^2}$,vedi quaderno per calcoli) che il metodo di Newton può o meno convergere all'ottimo in base alla scelta del punto iniziale x^0 della sequenza $\{x^k\}$. Si ha quindi che il metodo di Newton non ha proprietà di convergenza globale

Proposizione 9.1 (Convergenza locale metodo di Newton)**.** Sia $f \in C^2(\mathbb{R}^n)$. Sia $x^* \in \mathbb{R}^n$ tale che:

- $\nabla f(x^*) = 0$
- $\nabla^2 f(x^*)$ è non singolare

Allora $\exists \epsilon > 0 : \ \forall x^0 \in \mathcal{B}_\epsilon(x^*)$ la sequenza $\{x^k\}$ prodotta dalla formula iterativa:

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

soddisfa:

1. $\{x^k\} \subseteq \mathcal{B}_\epsilon(x^*)$
2. $\lim_{k\to\infty} x^k = x^*$
3. $\lim_{k\to\infty} \frac{||x^{k+1}-x^*||}{||x^k-x^*||} = 0$ (Convergenza superlineare)

Inoltre se l'hessiana $\nabla^2 f(x)$ è Lipshitz-continua, il tasso di convergenza è quadratico.

Per dimostrare la proposizione sulla convergenza locale del metodo di Newton richiamiamo il seguente Teorema

Teorema 9.1 (Teorema della media in forma integrale)**.** *Sia $F : \mathbb{R}^n \to \mathbb{R}^m$, $\forall x, y \in \mathbb{R}^n$ si ha:*

$$F(x) = F(y) + \int_0^1 J_F(y + t(x-y))(x-y)dt$$

Dove J_F denota il Jacobiano della funzione F:

$$F(x) = (f_1(x), ..., f_m(x))^T \implies J_F(x) = \begin{pmatrix} \nabla f_1(x)^T \\ ... \\ \nabla f_m(x)^T \end{pmatrix}$$

Proof. Dimostriamo il punto (1). Dobbiamo mostrare che date le ipotesi della proposizione la sequenza $\{x^k\}$ prodotta dalla formula iterativa del metodo di Newton è interamente contenuta in un intorno dell'ottimo: $\{x^k\} \in \mathcal{B}_\epsilon(x^*)$.
Poichè l'hessiana $\nabla^2 f(x^*)$ è non singolare $\exists \epsilon_1 > 0$ e $\exists \mu > 0$ tali che:

$$||\nabla^2 f(x)^{-1}|| \leq \mu \ \ \forall x \in \mathcal{B}_{\epsilon_1}(x^*)$$

(Nota: Penso che derivi da qui) Inoltre, essendo $f \in C^2(\mathbb{R}^n)$ si ha che $\exists \epsilon \leq \epsilon_1$ e $\exists \sigma \in (0,1)$ tali che :

$$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq \frac{\sigma}{\mu} \ \ \forall x \in \mathcal{B}_\epsilon(x^*)$$

(Nota: questa proprio boh, è da trovare ma non so nemmeno come cercarla su internet D:) Osserviamo che per definizione $\mathcal{B}_\epsilon(x^*) \subseteq \mathcal{B}_{\epsilon_1}(x^*)$. Procediamo per induzione. Supponiamo che $x^0 \in \mathcal{B}_\epsilon(x^*)$, e supponiamo poi che $x^k \in \mathcal{B}_\epsilon(x^*)$. Per dimostrare che l'intera sequenza $\{x^k\} \subseteq \mathcal{B}_\epsilon(x^*)$ dobbiamo quindi mostrare che $x^{k+1} \in \mathcal{B}_\epsilon(x^*)$.

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

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

Moltiplichiamo il termine $(x^k - x^*)$ per la matrice identità $I = [\nabla^2 f(x^k)]^{-1} \nabla^2 f(x^k)$:

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

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

Poiche $\nabla f(x^*) = 0$ possiamo scrivere in modo equivalente:

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

Passando alla norma otteniamo:

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

L'obiettivo è mostrare che tale norma è minore o uguale al raggio ϵ della palla centrata in x^*. Utilizziamo la disuguaglianza di Cauchy-Shwartz per derivare:

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

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

Ora, per ipotesi induttiva abbiamo $x^k \in \mathcal{B}_\epsilon(x^*) \subseteq \mathcal{B}_{\epsilon_1}(x^*)$ da cui per l'ipotesi di non singolarità dell'hessiana della funzione obiettivo nel punto di ottimo si ha $||[\nabla^2 f(x^k)]^{-1}|| \leq \mu$, pertanto possiamo scrivere:

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

Applichiamo ora il teorema della media in forma integrale alla funzione $F = \nabla f$ nei punti x^k e x^*. E' necessario osservare che in generale vale $J(\nabla f)^T = \nabla^2(f)$:

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

Sostituendo l'espressione di $\nabla f(x^k)$ data dal teorema della media in forma integrale all'interno del limite superiore per $||x^{k+1} - x^*||$ si ottiene

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

Notiamo poi che il termine $\nabla f(x^k)(x^k - x^*)$ è costante rispetto alla variabile di integrazione t, e pertanto può essere portato sotto il segno di integrale senza dividere per nessuna costante(dato che gli estremi di integrazione sono 0-1). Possiamo quindi scrivere:

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

Raccogliendo il termine $x^k - x^*$:

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

Procediamo sfruttando la disuguaglianza $|| \int f|| \leq \int ||f||dt$:

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

Applichiamo nuovamente la disuguaglianza di Cauchy-Shwartz:

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

Ora poichè per ipotesi $x^k \in \mathcal{B}_\epsilon(x^*)$ si ha che:

$$x^* + t(x^k - x^*) = tx^k + (1-t)x^* \in \mathcal{B}_\epsilon(x^*) \ \ \forall t \in (0,1)$$

per la convessità di $\mathcal{B}_\epsilon(x^*)$. Ricapitolando abbiamo ottenuto la disuguaglianza:

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

Adesso poichè $x^k \in \mathcal{B}_\epsilon(x^*)$ per ipotesi e avendo ricavato $x^* + t(x^k - x^*) \in \mathcal{B}_\epsilon(x^*) \ \ \forall t \in (0,1)$ possiamo applicare la disuguaglianza

$$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq \frac{\sigma}{\mu} \ \ \forall x, y \in \mathcal{B}_\epsilon(x^*), \epsilon \leq \epsilon_1, \sigma \in (0,1)$$

per i valori $x = x^* + t(x^k - x^*)$ e $y = x^k$, ottenendo:

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

Concludiamo la dimostrazione considerando che $\sigma \in (0,1) \implies$

$$||x^{k+1} - x^*|| \leq \sigma ||x^k - x^*|| \leq ||x^k - x^*|| \leq \epsilon$$

Dato che per ipotesi induttiva $x^k \in \mathcal{B}_\epsilon(x^*) \implies ||x^k - x^*|| \leq \epsilon$. Si ha quindi

$$x^{k+1} \in \mathcal{B}_\epsilon(x^*)$$

□

Proof. Dimostriamo il punto (2): dobbiamo mostrare che la sequenza $\{x^k\}$ converge a x^*.Iteriamo k volte la disuguaglianza ottenuta nel punto precedente per la norma $||x^{k+1} - x^*||$

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

Passando al limite

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

poichè $\sigma \in (0,1)$. Per il teorema dei carabinieri si ha:

$$\lim_{k \to \infty} ||x^k - x^*|| = 0 \implies \{x^k\} \to x^*$$

□

Proof. Dimostriamo il punto (3): dobbiamo mostrare che il tasso di convergenza del metodo di Newton è superlineare. Partiamo dal considerare una delle disuguaglianze

ottenute al punto (1):

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

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

Osserviamo che per $k \to \infty$ si ha che $x^k \to x^*$ per il punto (2), da cui si ha che $\int_0^1 \left|\left|(\nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k))\right|\right| dt \to 0$. Per $k \to \infty$ si ha quindi:

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

Per il teorema dei carabinieri si ha quindi

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

□

Proof. Dimostriamo il punto (4): dobbiamo mostrare che nel caso di hessiana Lipshitz-conitnua il tasso di convergenza del metodo di Newton è quadratico. Supponiamo quindi che:

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

Si ha:

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

$$= L\mu \int_0^1 ||x^* - tx^* - x^k + tx^k|| \cdot ||x^k - x^*|| dt =$$

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

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

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

Abbiamo quindi ottenuto

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

□

Analizziamo i tassi di convergenza e le complessità degli algoritmi visti fino ad ora nel caso convesso e fortemente convesso:

Algoritmo	Caso convesso	Caso fortemente convesso
Gradiente	Tasso sublineare- $\mathcal{O}(\frac{1}{\epsilon})$	Tasso lineare - $\mathcal{O}(\log(\frac{1}{\epsilon}))$
Nesterov	Tasso sublineare- $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$	Tasso lineare - $\mathcal{O}(\log(\frac{1}{\epsilon}))$
Newton	Tasso sublineare- $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$	Tasso superlineare- $\mathcal{O}(\log(\log(\frac{1}{\epsilon})))$

Notiamo che nell'analisi non vengono considerati i costi delle singole iterazioni: un iterazione di un algoritmo del secondo ordine come Newton sarà generalmente più costosa, in quanto ad ogni passo è necessario calcoare gradiente, hessiana e inversa dell'hessiana: tipicamente un iterazione di Newton costa $\mathcal{O}(n^3)$, mentre un iterazione del metodo del gradiente ha un costo $\mathcal{O}(n)$.
Osserviamo inoltre che nel metodo di Newton il calcolo della direzione di discesa

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

al passo k richiede di invertire la matrice hessiana della funzione obiettivo nel punto x^k. Tuttavia si deve notare che l'operazione di inversione di una matrice è un operazione onerosa a livello di risorse computazionali impiegate, ed è inoltre un operazione che può portare a errori numerici in aritmetica finita. Per ottenere maggior robustezza ed efficienza nel calcolo della direzione di discesa del metodo di Newton si ricorre quindi alla soluzione del sistema lineare:

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

Osserviamo inoltre che nel caso in cui $\nabla^2 f(x^k)$ sia singolare o non definita positiva si può ottenere una direzione $d_k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$ che non è in generale una direzione di discesa. Per risolvere il problema quando una di queste due alternative si verifica si procede effettuando un passo dell'algoritmo del gradiente(si pone quindi $d_k = \nabla f(x^k)$).
Concludiamo la trattazione del metodo di Newton con un osservazione: il metodo di Newton non effettua alcun controllo sulla funzione obiettivo, ma utilizza esclusivamente informazioni del primo e del secondo ordine. Ciò può portare a iterazioni per le quali la funzione obiettivo non decresce, ossia vale $f(x^{k+1}) \geq f(x^k)$. Per ovviare a questi casi, si può rinunciare al passo costante $\alpha = 1$ ed effettuare una ricerca di linea di tipo Armijo, al fine di garantire un sufficiente decremento della funzione obiettivo.

Definizione 9.2 (Metodo di Newton globalmente convergente)**.** Un algoritmo iterativo si dice **metodo di Newton globalmente convergente** se produce sequenze $\{x^k\}$ con le seguenti proprietà:

- $\{x^k\}$ ha punti di accumulazione, ognuno dei quali è stazionario
- Nessuno dei punti di accumulazione della sequenza $\{x^k\}$ è un massimo locale
- Se $\{x^k\}$ converge ad un punto di minimo locale x^* allora $\exists \bar{k} : \forall k \geq \bar{k} \;\; x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

Ossia l'algoritmo non fa più uso di strategie di globalizzazione ed esegue un passo unitario nella direzione di Newton

10 Metodi Quasi-Newton

Consideriamo il problema quadratico:

$$\min_x \frac{1}{2} x^T Q x - c^T x \implies \nabla f(x) = Qx - c$$

Da cui si ha

$$\nabla f(x) - \nabla f(y) = Qx - c - (Qy - c) = Q(x - y) + c - c = Q(x - y)$$

Si ottiene quindi l'**Equazione Quasi-Newton**

$$\nabla f(x) - \nabla f(y) = Q(x - y)$$

Negli algoritmi iterativi si ha che per due iterate successive:

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

Negli algoritmi iterativi visti fin'ora avevamo le seguenti formule di aggiornamento:

- Gradiente: $x^{k+1} = x^k - \alpha_k I \nabla f(x^k)$
- Newton: $x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

Le formule di aggiornamento sopra riportate possono essere riassunte nel seguente modello generale:

$$x^{k+1} = x^k - \alpha_k B_k^{-1} \nabla f(x^k) \;\; B_k \text{ simmetrica e definita positiva}$$

L'idea generale dei metodi Quasi-Newton è quella di utilizzare lo schema di aggiornamento $x^{k+1} = x^k - \alpha_k B_k^{-1} \nabla f(x^k)$ prendendo $B_k \approx \nabla^2 f(x^k)$ e B_k che soddisfa

l'equazione Quasi-Newton.

Formule di aggiornamento dirette:

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

$$B_{k+1} = B_k + \Delta B_k$$

Nelle formule di aggiornamento dirette si costruisce una matrice B_k che approssima l'hessiana della funzione obiettivo nel punto x^k.

Formule di aggiornamento inverse:

$$x^{k+1} = x^k - \alpha_k H_k \nabla f(x^k) \quad H_k \approx [\nabla^2 f(x^k)]^{-1}$$

$$H_{k+1} = H_k + \Delta H_k$$

Se poniamo

$$s_k = x^k - x^{k-1} \quad y_k = \nabla f(x^k) - \nabla f(x^{k-1})$$

L'equazione Quasi newton può essere riscritta nelle forme:

$$B_k s_k = y_k \quad \text{per formule dirette}$$

$$s_k = H_k y_k \quad \text{per formule inverse}$$

Le perturbazioni devono essere quindi tale che:

$$B_{k+1} s_{k+1} = y_{k+1}$$

$$H_{k+1} y_{k+1} = s_{k+1}$$

Algorithm 6 Schema generale metodo Quasi-Newton

Require: $x^0 \in \mathbb{R}^n, B_0 \in \mathbb{R}^{n \times n}$ simmetrica e definita positiva

```
k = 0
while ∇f(x^k) ≠ 0 do
    d_k = −B_k^{−1} ∇f(x^k)
    x^{k+1} = x^k + α_k d_k
    Calcolo B_{k+1} : B_{k+1}(x^{k+1} − x^k) = ∇f(x^{k+1}) − ∇f(x^k)
    k++
end while
```

Per determinare la matrice B_{k+1} relativa al passo successivo si utilizzano diverse classi di formule di aggiornamento.

Aggiornamenti di rango 1:

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

Osserviamo che la matrice $\rho_k \mu_k v_k^T$ è di rango 1 poichè è data da 3 vettori riga generati come multipli dello stesso vettore v_k^T. L'efficacia dell'aggiornamento di rango 1 dipende dalla scelta di ρ_k, μ_k, v_k^T

Aggiornamenti di rango 2:

$$B_{k+1} = B_k + a_k \mu_k \mu_k^T + b_k v_k v_k^T \quad a_k, b_k \in \mathbb{R}, \mu_k, v_k \in \mathbb{R}^n$$

Si ricorre alla formula **BFGS**(Broyden, Fletcher, Goldfarb, Shanno):

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k s_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k} \quad \text{(Diretta)}$$

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

E' facile verificare che la formula BFGS soddisfa l'equazione Quasi-Newton. Scegliendo $d_k = -B_k^{-1} \nabla f(x^k)$ o $d_k = -H_k \nabla f(x^k)$ la direzione ottenuta è di discesa se la matrice $B_k(H_k)$ è definita positiva

Proposizione 10.1. Sia B_k definita positiva. Se $y_k^T s_k > 0$ allora B_{k+1} ottenuta tramite formula di aggiornamento BFGS è ancora definita postiva.

Come possiamo garantire che $s_k^T y_k > 0$?

$$s_k = x^{k+1} - x^k = x^k + \alpha_k d_k - x^k = \alpha_k d_k$$

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

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

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

Ricordando che la condizione di Wolfe debole è data da $d_k^T \nabla f(x^{k+1}) \geq \sigma d_k^T \nabla f(x^k) \quad \sigma \in (0, 1)$, con una ricerca di linea di tipo Wolfe possiamo garantire:

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

Possiamo quindi concludere che BFGS con una ricerca di linea di tipo Wolfe garantiscono che $d_k = -B_k^{-1} \nabla f(x^k)$ sia una direzione di discesa. Vediamo adesso le proprietà di convergenza del metodo BFGS:

- Se la funzione obiettivo è convessa, si può dimostrare che l'algoritmo converge globalmente all'ottimo
- Se la funzione obiettivo è fortemente convessa, il tasso di convergenza è superlineare.

11 Problemi a larga scala

Un **problema a larga scala** è un problema di ottimizzazione del tipo:

$$\min_{x\in\mathbb{R}^n} f(x) \quad n \approx 10^3 - 10^4$$

In questo contesto il metodo BFGS presenta dei problemi:

- Memorizzazione di $H_k \in \mathbb{R}^{n\times n}$
- Calcolo di $d_k = -H_k \nabla f(x^k)$

Possiamo esprimere la matrice H_{k+1} in modo ricorsivo

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

$$= ... = H(y_0, H_0, s_0, y_1, s_1, ..., y_k, s_k)$$

La matrice H_{k+1} è funzione di H_0 e delle sequenze $\{y_k\}$ e $\{s^k\}$. Ponendo $H_0 = \gamma I$ posso ricostruire H_{k+1} a partire da 1 scalare e $2k$ vettori in $\mathbb{R}^n$.
Nelle prime iterazioni la memoria richiesta è molto minore della memoria necessaria per gestire la matrice $n \times n$. Per semplificare ulteriormente il calcolo e diminuire la memoria impiegata, si può troncare la ricorsione:

$$H_{k+1} = H(y_0, H_0, s_0, y_1, s_1, ..., y_k, s_k) = H(H_{k-m}, y_{k-m}, s_{k-m}, .., y_k s_k)$$

Ad ogni iterazione si mantengono in memoria $2mn$ vettori invece di n^2. Si osserva sperimentalmente che con $m \approx 10$ si ottiene un ottima approssimazione dell'hessiana $\nabla^2 f(x^k)$. Nelle prime iterazioni si utilizza quindi BFGS con ricorsione, quando la memoria inizia a non bastare si attiva il troncamento della ricorsione.
Nota: esiste un algoritmo(metodo HG) che consente, con $4mn$ iterazioni, di svolgere in modo implicito il prodotto $H_k \nabla f(x^k)$ usando solo i valori di $\gamma, s_{k-m}, y_{k-m}, ..., s_k, y_k$.
Il metodo BFGS che utlizza questi meccanismi si chiama metodo **L-BFGS**(Limited memory BFGS). L-BFGS è considerato l'algoritmo di ottimizzazione non lineare non

vincolata più performante tra quelli disponibili. L-BFGS non ha proprietà di convergenza globale, neanche nel caso convesso! Nel caso fortemente convesso l'algoritmo ha un tasso di convergenza lineare

12 Ottimizzazione Vincolata

Un problema di ottimizzazione vincolata è un problema del tipo:

$$\min_{\substack{x \in S \\ S \subset \mathbb{R}^n}} f(x)$$

In un problema di ottimizzazione vincolata, quando si valuta una soluzione candidata x, non si considera più esclusivamente la qualità della soluzione (data dal valore obiettivo f(x)), ma se ne considera anche l'eventuale **ammissibilità**, ossia l'eventuale appartenenza all'insieme dei vincoli S, detto **Insieme ammissibile**.

Definizione 12.1 (Direzione ammissibile)**.** Una direzione $d \in \mathbb{R}^n$ si dice **direzione ammissibile** in $\bar{x} \in S$ se:

$$\exists \bar{t} > 0 : \quad \bar{x} + td \in S \ \ \forall t \in [0, \bar{t}]$$

Proposizione 12.2 (Condizioni necessarie di ottimalità del primo ordine)**.** Sia $f \in C^1(\mathbb{R}^n)$ e x^* minimo locale. Allora non esistono direzioni $d \in \mathbb{R}^n$ ammissibili in x^* tali che $\nabla f(x^*)^T d < 0$

Notiamo quindi che la condizione di stazionarietà nel caso vincolato è data dal fatto che se x^* è minimo locale allora $\nabla f(x^*)^T d \geq 0 \ \ \forall d$ammissibile. In particolare quindi non è più vero che se x è minimo locale allora $\nabla f(x) = 0$, in quanto si deve tener conto dell'ammissibilità del punto x. Notiamo inoltre che la condizione di stazionarietà nel caso vincolato implica la condizione di stazionarietà nel caso non vincolato: infatti se $S = \mathbb{R}^n$ la condizione x stazionario $\implies \nabla f(x)^T d \geq 0 \ \ \forall x \in \mathbb{R}^n \implies \nabla f(x) = 0$.

Proposizione 12.3 (Condizioni necessarie di ottimalità del secondo ordine)**.** Sia $f \in C^2(\mathbb{R}^n)$. Se $x^* \in \mathbb{R}^n$ è piunto di minimo locale allora non esistono direzioni $d \in \mathbb{R}^n$ ammissibili in x^* tali che:

1. $\nabla f(x^*)^T d = 0$
2. $d^T \nabla^2 f(x^*) d < 0$ (d a curvatura negativa)

Ricordiamo che se valgono le condizioni (1) e (2) della proposizione possiamo affermare che la direzione d è di discesa in x^*. Pertanto la condizione neessaria di ottimalità del

secondo ordine è sostanzialmente equivalente a quella del primo ordine, con la differenza che essendo la funzione obiettivo due volte continuamente differenziabile è possibile utilizzare l'hessiana per valutare se la direzione d è a curvatura negativa o meno.

Proposizione 12.4. Sia S convesso e $\bar{x} \in S$. La direzione $(x-\bar{x})$ è ammissibile $\forall x \in S$.

Proof. Se S è convesso $\implies \forall \lambda \in [0,1]$ si ha $\lambda\bar{x} + (1-\lambda)x \in S$, e quindi allo stesso modo $(1-\lambda)\bar{x} + \lambda x \in S$. Si ha che:

$$\bar{x} - \lambda\bar{x} + \lambda x \in S \implies \bar{x} + \lambda(x - \bar{x}) \in S \ \ \forall \lambda \in [0,1], \forall x \in S$$

Pertanto si ha che $\forall x \in S$ la direzione $d = (x - \bar{x})$ è ammissibile:

$$\bar{x} + \lambda(x - \bar{x}) \in S \ \ \forall \lambda \in [0,1] (\bar{t} = 1)$$

□

Proposizione 12.5 (Condizioni necessarie di ottimalità del primo ordine, Insieme ammissibile convesso)**.** Sia S convesso e x^* punto di minimo locale.

$$\nabla f(x^*)^T (x - x^*) \geq 0 \ \ \forall x \in S$$

Proof. Per assurdo sia $\hat{x} \in S: \ \ \nabla f(x^*)^T(\hat{x} - \bar{x}) < 0$. Poichè S è convesso la direzione $(\hat{x} - \bar{x})$ è ammissibile. Inoltre $(\hat{x} - x^*)$ è di discesa. Questo viola le CNO del primo ordine. □

Proposizione 12.6 (Condizioni necessarie e sufficienti ottimalità primo ordine,Insieme ammissibile e funzione obiettivo convessi)**.** Sia S convesso e $f \in C^1(\mathbb{R}^n)$ convessa. Allora x^* è punto di minimo globale $\iff$

$$\nabla f(x^*)^T (x - x^*) \geq 0 \ \ \forall x \in S$$

Proof. ($\implies$): Segue dalle condizioni necessarie di ottimalità del primo ordine con insieme ammissibile convesso.
($\impliedby$): Supponiamo che $\nabla f(x^*)^T (x - x^*) \geq 0 \ \ \forall x \in S$. f convessa

$$\implies f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*) \forall x \in \mathbb{R}^n$$

Ma allora $\nabla f(x^*)^T (x - x^*) \geq 0 \ \ \forall x \in S \implies \forall x \in S f(x) \geq f(x^*)$ Ovvero x^* è punto di minimo globale per il problema vincolato. □

Definizione 12.7 (Vincoli poliedrali)**.** Un insieme $P \subseteq \mathbb{R}^n$ definito da:

$$P = \{x \in \mathbb{R}^n : \quad Ax \leq b\}$$

con $A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$ è detto **Poliedro**

Definizione 12.8 (insieme dei vincoli attivi)**.** Si definisce l'**insieme dei vincoli attivi** in x: $I(x) \subseteq \{1, ..., m\}$ come:

$$I(x) = \{i \in \{1, ..., m\} : a_i^T x = b_i\}$$

Proposizione 12.9 (Direzioni di discesa e vincoli poliedrali)**.** Sia $\bar{x} \in P$, con P poliedro. Una direzione $d \in \mathbb{R}^n$ è ammissibile in $\bar{x} \iff a_i^T d \leq 0 \ \ \forall i \in I(x)$

Proof. Dimostriamo $\implies$:
Sia $d \in \mathbb{R}^n$ una direzione ammissibile in $\bar{x}$. Supponiamo per assurdo che $\exists i \in I(\bar{x})$: $a_i^T d > 0$.
Poichè d è ammissibile in $\bar{x}$ si ha che per $t > 0$ sufficientemente piccolo vale:

$$\bar{x} + td \in P \implies (\bar{x} + td)^T a_i \leq b_i$$

Esplicitando il prodotto scalare si ha

$$\bar{x}^T a_i + td^T a_i \leq b_i \implies td^T a_i \leq 0$$

Dove l'implicazione è data dal fatto che per $i \in I(\bar{x})$ vale $\bar{x}^T a_i = b_i$. Si ha quindi un assurdo, in quanto per ipotesi $t > 0$ e $d^T a_i > 0$. □

Proof. Dimostriamo $\impliedby$:
Supponiamo quindi che $a_i^T d \leq 0 \ \ \forall i \in I(\bar{x})$: vogliamo dimostrare che la direzione d è ammissibile, cioè che $\exists \bar{t} > 0 : \quad A(\bar{x} + td) \leq b \ \ \forall t \in [0, \bar{t}]$.
Sia $j \in \{1, ..., m\}$, identifichiamo 3 casi principali:

- $j \in I(\bar{x})$, in questo caso si ha:

  $$(\bar{x} + td)^T a_j = \bar{x}^T a_j + td^T a_j = b_j + td^T a_j$$

  Ora $t > 0, d^T a_j \leq 0$ per ipotesi $\implies b_j + td^T a_j \leq b_j$, ossia $(\bar{x} + td)$ soddisfa il vincolo $j \ \ \forall t > 0$

- $j \notin I(\bar{x})$ e $a_j^T d \leq 0$
  Poichè il vincolo j non fa parte dell'insieme dei vincoli attivi in $\bar{x}$ si ha $a_j^T \bar{x} < b_j$.

Calcolando $(\bar{x} + td)^T a_j$ si ha quindi:

$$(\bar{x} + td)^T a_j = \bar{x}^T a_j + td^T a_j = b_j + td^T a_j < b_j + td^T a_j \leq b_j$$

Cioè $\bar{x} + td$ soddisfa il j-esimo vincolo $\forall t > 0$

- $j \notin I(\bar{x})$ e $a_j^T d > 0$

$$(\bar{x} + td)^T a_j = \bar{x}^T a_j + td^T a_j \leq b_j \iff$$
$$\iff t \leq \frac{b_j - x^T a_j}{d^T a_j}$$

Pertanto, scegliendo opportunamente $\bar{t}$ posso comunque ottenere una direzione ammissibile:

$$t \leq \min_{\substack{j \notin I(\bar{x}) \\ d^T a_j > 0}} \frac{b_j - \bar{x}^T a_j}{d^T a_j} \implies \bar{t} = \frac{b_j - \bar{x}^T a_j}{d^T a_j}$$

Abbiamo visto che per tutti i casi possibili per il j-esimo vincolo si riesce sempre a determinare un opportuno $\bar{t} > 0$ tale per cui $\bar{x} + td$ è ammissibile $\forall t \in [0, \bar{t}]$

Interpretazione geometrica: I vettori $a_i : i \in \{0, ..., m\}$ sono i vettori ortogonali all'iperpiano definito dal vincolo i-esimo. Richiedere che $a_i^T d \leq 0$,ossia che il prodotto scalare tra il vettore a_i e il vettore direzione$d$ sia minore o uguale di zero, equivale a chiedere che l'angolo tra la direzione di discesa e il vettore ortogonale al vincolo i-esimo sia maggiore di 90 gradi.
Nel primo caso si ha che il j-esimo vincolo è attivo in $\bar{x}$, per cui valendo $a_j^T d \leq 0$ si ha che la direzione di discesa 'punta verso' la parte dello spazio ammissibile per il j-esimo vincolo, ossia per ogni possibile passo positivo nella direzione d il punto $\bar{x} + td$ sarà sempre ammissibile rispetto al vincolo j-esimo
Nel secondo caso il vincolo j-esimo non è attivo in $\bar{x}$ ma vale $a_j^T d \leq 0$. Poichè la direzione d forma un angolo maggiore di 90 gradi con il vettore ortogonale all'iperpiano che definisce il vincolo j-esimo, e il punto $\bar{x}$ non si trova su tale iperpiano,per ogni possibile passo positivo nella direzione d il punto $\bar{x} + td$ sarà sempre ammissibile rispetto al vincolo j-esimo
Nel terzo caso il vincolo j-esimo non è attivo in $\bar{x}$ e vale $a_j^T d > 0$. In questo caso l'angolo è minore di 90 gradi, e quindi esisterà un valore del passo t tale per cui uno spostamento lungo d a partire da $\bar{x}$ porta a violare il vincolo j-esimo. In questo caso pero è possibile trovare un limite superiore al valore del passo per il quale tutti i passi inferiori a tale limite lungo d a partire da $\bar{x}$ portano in un punto ammissibile.

□

Consideriamo ora il caso di insieme ammissibile dato da:

$$S = \{x \in \mathbb{R}^n : Ax \leq b, \mu_i^T x = c_i \ \ \forall i = 1, ..., p\}$$

La proposizione precedente può essere riscritta come:
d ammissibile in $\bar{x} \in S \iff a_i^T d \leq 0 \ \ \forall i \in I(\bar{x})$ e $\mu_i^T d = 0 \ \ \forall i = 1, .., p$

Proof.

$$\mu_i^T x = c_i \iff \begin{cases} \mu_i^T x \leq c_i \\ -\mu_i^T x \leq c_i \end{cases}$$

Poichè $\bar{x} \in S$ si ha che $\mu_i^T \bar{x} = c_i \ \ \forall i = 1, ..., p$, pertanto i vincoli $\mu_i^T x \leq c_i$ e $-\mu_i^T x \leq c_i$ sono entrambi vincoli attivi in $\bar{x}$. Per la proposizione precedente si deve quindi avere

$$\begin{cases} \mu_i^T d \leq 0 \\ \mu_i^T d \geq 0 \end{cases} \implies \mu_i^T d = 0$$

□

Definizione 12.10 (Vincoli di box)**.** Un insieme di **Vincoli di box** è un insieme del tipo:

$$S = \{x \in \mathbb{R}^n : l_i \leq x_i \leq u_i, \ \ i = 1, ..., n\}$$

Un insieme di n vincoli di box può essere visto come un insieme di 2n vincoli lineari di disuguaglianza:

- n vincoli del tipo: $x^T e_h \geq l_h$
- n vincoli del tipo: $x^T e_h \leq u_h$

Ricordiamo che se $\bar{x}$ è ottimale $\implies \nabla f(\bar{x})^T d \geq 0 \ \ \forall d$ ammissibile in $\bar{x}$.
Fissiamo l'i-esimo vincolo di box e consideriamo i possibili casi per $\bar{x}_i$:

- Se $\bar{x}_i = l_i$ il vincolo $-\bar{x}^T e_i \leq l_i$ è attivo.
  Consideriamo $d = e_i$: d è ammissibile rispetto al vincolo i-esimo $\bar{x}^T e_i = l_i$ in quanto $-e_i^T e_i = -1 \leq 0$, dove abbiamo utilizzato la proposizione 12.9 sui vincoli poliedrali attivi e il prodotto $a_i^T d$. Inoltre, se $h \neq i$ si ha che $-e_h^T e_i = 0$. Pertanto la direzione $d = e_i$ è ammissibile in $\bar{x}$ punto di ottimo. Per l'ottimalità di $\bar{x}$ si ha

$$\nabla f(\bar{x})^T d \geq 0 \implies \frac{\partial f(\bar{x})}{\partial x_i} \geq 0$$

- Analogamente al caso precedente, se $\bar{x}_i = u_i$ il vincolo $\bar{x}^T e_i \leq u_i$ è attivo.
  Si ha quindi che la direzione $d = -e_i$ è ammissibile in quanto $e_i^T(-e_i) = -1 \leq 0$ e $\forall h \neq i$ si ha $-e_h^T e_i = 0 \leq 0$. Per l'ottimalità di $\bar{x}$ si ha

$$\nabla f(\bar{x})^T d \geq 0 \implies -\frac{\partial f(\bar{x})}{\partial x_i} \geq 0 \implies \frac{\partial f(\bar{x})}{\partial x_i} \leq 0$$

- Infine se $l_i < \bar{x}_i < u_i$ le direzioni $d = e_i$ e $d = -e_i$ sono entrambe ammissibili. Dai risultati precedenti si deve quindi avere

$$\begin{cases} \frac{\partial f(\bar{x})}{\partial x_i} \geq 0 \\ \frac{\partial f(\bar{x})}{\partial x_i} \leq 0 \end{cases} \implies \frac{\partial f(\bar{x})}{\partial x_i} = 0$$

Definizione 12.11 (Vincoli di simplesso)**.** Un insieme di **vincoli di simplesso** è un insieme del tipo:

$$S = \{x \in \mathbb{R}^n : x \geq 0, e^T x = 1\}$$

13 Proiezioni e ottimizzazione vincolata

Sia $\bar{x} \in \mathbb{R}^n$, consideriamo il problema:

$$\min_{y \in S} \frac{1}{2}||y - \bar{x}||^2 = r(y)$$

Risolvere questo problema equivale a determinare il punto ammissibile più vicino al punto $\bar{x}$ dato. Osserviamo che

- $r(y)$ è coerciva $\implies$ Il problema ammette soluzione
- $r(y)$ è strettamente convessa $\implies$ La soluzione è unica

La soluzione del problema è data da

$$P(\bar{x}) = \arg\min_{y \in S} \frac{1}{2}||y - \bar{x}||^2$$

$P(\bar{x}) : \mathbb{R}^n \to S$ rappresenta la **Proiezione ortogonale** di $\bar{x}$ su S.

Proposizione 13.1. Sia $\bar{x} \in \mathbb{R}^n$. $P(\bar{x})$ è la proiezione di $\bar{x}$ su S $\iff$:

$$(\bar{x} - P(\bar{x}))^T (x - P(\bar{x})) \leq 0 \ \ \forall x \in S$$

Proof. $P(\bar{x})$ è soluzione del problema:

$$\min_{y \in S} \frac{1}{2}||y - \bar{x}||^2$$

Per le CNS di ottimalità(f convessa e continuamente differenziabile) ciò è equivalente a dire che se $y = P(\bar{x})$:

$$\nabla r(y)^T (x - y) \geq 0 \ \ \forall x \in S$$

Ora $\nabla r(y) = y - \bar{x}$, da cui si ha

$$(y - \bar{x})T(x - y) \geq 0 \implies (\bar{x} - y)^T (x - y) \leq 0$$

□

Proposizione 13.2. La proiezione $P : \mathbb{R}^n \to S$ è una mappa non espansiva, ossia vale:

$$||P(x) - P(y)|| \leq ||x - y|| \ \ \forall x, y \in \mathbb{R}^n$$

Significato: La proiezione è una mappa che non aumenta mai la distanza tra due punti distinti. Ciò implica che la proiezione è continua, in quanto si può notare che la non espansività dell'operazione di proiezione implica la sua Lipshitz-continuità con costante di Lipshitz $L = 1$.

Proposizione 13.3. Un punto ammissibile $\bar{x} \in S$ è stazionario $\iff$

$$\bar{x} = P(\bar{x} - \nabla f(\bar{x}))$$

Proof. Utilizzo la proposizione:

$$(y - P(y))^T (x - P(y)) \leq 0 \ \ \forall x \in S$$

ponendo $y = \bar{x} - \nabla f(\bar{x})$ ottengo $\forall x \in S$:

$$(\bar{x} - \nabla f(\bar{x}) - \bar{x})^T (x - \bar{x}) \leq 0 \implies -\nabla f(\bar{x})^T (x - \bar{x}) \leq 0$$

Si ottiene quindi

$$\nabla f(\bar{x})^T (x - \bar{x}) \geq 0 \ \ \forall x \in S$$

□

Proposizione 13.4 (Condizione necessaria di ottimalità)**.** Se $\bar{x}$ è minimo locale $\implies$ $P(\bar{x} - \nabla f(\bar{x})) = \bar{x}$.

Proposizione 13.5 (Condizione necessaria e sufficiente ottimalità(caso convesso))**.** Se f è convessa un punto $\bar{x}$ è di minimo locale $\iff P(\bar{x} - \nabla f(\bar{x})) = \bar{x}$

Per alcune classi di vincoli è facile calcolare la proiezione. Nel caso di vincoli di box:

$$S = \{x \in \mathbb{R}^n : x_i \in [l_i, u_i] \forall i = 1, ..., n\}$$

Se $y = P(x)$ vale:

$$y_i = \begin{cases} x_i & x_i \in (l_i, u_i) \\ l_i & x_i \leq l_i \\ u_i & x_i \geq u_i \end{cases}$$

Consideriamo ad esempio il problema:

$$\min_{\substack{l_1 \leq x \leq u_1 \\ l_2 \leq y \leq u_2}} \frac{1}{2} \left|\left| \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right|\right|^2 = \frac{1}{2}((x - \bar{x})^2 + (y - \bar{y})^2)$$

Posso variare indipendentemente x e y. Per $x, \bar{x} \in \mathbb{R}^n$ si ha:

$$\min_{l \leq x \leq u} \frac{1}{2}||x - \bar{x}||^2 = \frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x}_i)^2$$

Posso quindi minimizzare indipendentemente i termini della sommatoria:

$$\min_{l_i \leq x_i \leq u_i} (x_i - \bar{x}_i)^2$$

Nel caso di vincoli di simplesso:

$$S = \{x \in \mathbb{R}^n : x \geq 0, e^T x = 1\}$$

Esiste un algoritmo che calcola la proiezione in al massimo n passi.

14 Algoritmi iterativi di ottimizzazione vincolata

Supponiamo di dover risolvere il seguente problema:

$$\min_{x \in S} f(x) \text{ S convesso e compatto}$$

con un algoritmo iterativo di ottimizzazione di tipo line search:

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

Scegliamo quindi

- d_k: Direzione di discesa ammissibile $\implies$ ammissibilità di d_k garantita scegliendo $d_k = (x - x^k)$ con $x \in S$
- α_k: Passo lungo d_k tale per cui $x^k + \alpha_k d_k \in S$

Osserviamo che se scegliamo $d_k = \hat{x}^k - x^k$ con $\hat{x}^k \in S$, per $\alpha_k \leq 1$ il nuovo punto della sequenza $\{x^k\}$ è sicuramente un punto ammissibile. Infatti se:

$$d_k = (\tilde{x} - x^k), \tilde{x} \in S, \alpha_k = 1 \implies x^{k+1} = x^k + (\tilde{x} - x^k) = \tilde{x} \in S$$

Quindi limitando il passo $\alpha_k \leq 1$ lungo la direzione $d_k = (\tilde{x} - x^k), \tilde{x} \in S$ si nota che ponendo $\alpha_k = 1$(passo massimo) al più il nuovo punto della sequenza $\{x^k\}$ potrà finire sulla frontiera dell'insieme ammissibile S(se il punto $\tilde{x}$ utilizzato per costruire la direzione d_k era un punto sulla frontiera dell'insieme ammissibile)
Pertanto consideriamo

- $d_k = \hat{x}^k - x^k$ per qualche $\hat{x}^k \in S$ tale per cui la direzione d_k ottenuta risulta essere una direzione di discesa.
- α_k calcolato tramite una ricerca di linea di tipo Armijo con passo iniziale $\alpha_0 \leq 1$
  Proprietà di una ricerca di linea Armijo vincolata:
  - $x^{k+1} \in S$
  - $f(x^{k+1}) < f(x^k)$
  - $\lim_{k\to\infty} \nabla f(x^k)^T d_k = 0$

Abbiamo visto che scegliere $d_k = (\hat{x}^k - x^k), \hat{x}^k \in S$ garantisce l'ammissibilità di d_k. Dobbiamo adesso determinare quale punto $\hat{x}^k \in S$ garantisce che la direzione d_k sia una direzione di discesa per la funzione obiettivo in x^k.

15 Metodo del gradiente proiettato

Nel **Metodo del gradiente proiettato** si definisce:

$$d_k = \hat{x}^k - x^k, \quad \hat{x}^k = P(x^k - \nabla f(x^k))$$

Queste scelte garantiscono che:

- d_k sia una direzione ammissibile $\impliedby$ $\hat{x}^k \in S$ per definizione di mappa di proiezione

- d_k sia una direzione di discesa: Per la proposizione 13.1 si ha:

$$(x^k - \nabla f(x^k) - \hat{x}^k)^T(x^k - \hat{x}^k) \leq 0 \implies$$

$$(x^k - \hat{x}^k)^T(x^k - \hat{x}^k) - \nabla f(x^k)^T(x^k - \hat{x}^k) \leq 0 \implies$$

$$-\nabla f(x^k)^T(x^k - \hat{x}^k) \leq -||x^k - \hat{x}^k||^2$$

Pertanto si ha che

$$\nabla f(x^k)^T(\hat{x}^k - x^k) \leq -||x^k - \hat{x}^k|| \leq 0$$

Da cui possiamo distinguere 2 casi:

$$-||x^k - \hat{x}^k|| \implies \begin{cases} < 0 : x^k \neq \hat{x}^k \implies \nabla f(x^k)^T d_k < 0 \implies d_k \text{ è di discesa} \\ = 0 : x^k = \hat{x}^k \implies x^k = P(x^k - \nabla f(x^k)) \implies x^k \text{ stazionario} \end{cases}$$

Dove nel caso 2 abbiamo utilizzato la proposizione 13.3 sulla condizione di stazionarietà dei punti ammissibili($x \in S$ stazionario $\iff x = P(x - \nabla f(x))$)

Algorithm 7 Metodo del gradiente proiettato

Require: $x^0 \in S$

$k = 0$

while $||x^k - P(x^k - \nabla f(x^k))|| > \epsilon$ **do**

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

$d_k = \hat{x}^k - x^k$

Calcolo α_k lungo d_k con una ricerca di linea Armijo

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

$k++$

end while

Proposizione 15.1 (Proprietà di convergenza metodo del gradiente proiettato)**.** Sia $S \subseteq \mathbb{R}^n$ convesso e compatto, $f : \mathbb{R}^n \to \mathbb{R} : f \in C^1(\mathbb{R}^n)$, e sia $\{x^k\}$ la sequenza prodotta dal metodo del gradiente proiettato. La sequenza $\{x^k\}$ ha punti di accumulazione, ciascuno dei quali è stazionario

Proof. Innanzitutto notiamo che la sequenza $\{x^k\}$ prodotta dal metodo del gradiente proiettato è tale per cui $\{x^k\} \subset S$, poichè ad ogni passo dell'algoritmo il punto x^k è un punto ammissibile. Per la compattezza di S si ha quindi che la sequenza $\{x^k\}$ ammette punti di accumulazione. Dobbiamo mostrare che ciascun punto di accumulazione di

$\{x^k\}$ è un punto stazionario per la funzione obiettivo.
Sia $K \subseteq \{0, 1, ...\}$ tale che

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

Consideriamo $\hat{x}^k := P(x^k - \nabla f(x^k))$, per la continuità di ∇f e P si ha:

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

Per le proprietà della ricerca di linea di tipo Armijo vincolata si ha:

$$\lim_{\substack{k\to\infty \\ k\in K}} \nabla f(x^k)^T d_k = 0 \implies \lim_{\substack{k\to\infty \\ k\in K}} \nabla f(x^k)^T (\hat{x}^k - x^k) = \nabla f(\bar{x})^T (\hat{x} - \bar{x}) = 0$$

Per le proprietà della mappa di proiezione si ha:

$$(\bar{x} - \nabla f(\bar{x}) - \hat{x})^T (\bar{x} - \hat{x}) \leq 0 \implies$$

$$-\nabla f(\bar{x})^T (\bar{x} - \hat{x}) \leq -||\bar{x} - \hat{x}||^2 \implies$$

$$\nabla f(\bar{x})^T (\hat{x} - \bar{x}) \leq -||\bar{x} - \hat{x}||^2$$

Avendo ricavato precedentemente $\nabla f(\bar{x})^T (\hat{x} - \bar{x}) = 0$ possiamo scrivere

$$||\bar{x} - \hat{x}||^2 \leq 0 \implies ||\bar{x} - \hat{x}||^2 = 0 \implies \hat{x} = \bar{x}$$

Ma $\bar{x} = \hat{x} = P(\bar{x} - \nabla f(\bar{x})) \implies \bar{x}$ stazionario per la proposizione 13.3 □

16 Metodo di Frank-Wolfe

Nel **Metodo di Frank-Wolfe** scegliamo:

$$d_k = \hat{x}^k - x^k$$

Dato x^k consideriamo il problema:

$$\min_{x\in S} \nabla f(x^k)^T (x - x^k)$$

L'idea del metodo di Frank-Wolfe è quella di approssimare la funzione obiettivo del problema originale con una funzione lineare data dallo sviluppo di Taylor troncato al primo ordine, e confinata nell'insieme ammissibile S. Questa tecnica ha senso solo per problemi vincolati su un compatto S, per i quali le funzioni lineari ammettono minimo.

Se z^k è la soluzione del problema approssimato

$$z^k = \min_{x \in S} \nabla f(x^k)^T (x - x^k)$$

E scegliamo

$$\hat{x}^k \in \arg\min_{x \in S} \nabla f(x^k)^T (x - x^k)$$

Si possono verificare 2 casi:

- Se $z^k = 0 \implies \nabla f(x^k)^T (x - x^k) \geq 0 \ \ \forall x \in S$ in quanto z^k è il valore ottimo della funzione $\nabla f(x^k)^T (x - x^k)$. Per la convessità di f si ha quindi che il punto x^k è un punto stazionario del problema originale
- Se $z^k < 0$ si ha conseguentemente $\nabla f(x^k)^T (\hat{x}^k - x^k) < 0$, ossia la direzione

  $$d_k = \hat{x}^k - x^k$$

  è una direzione di discesa per la funzione obiettivo del problema originale nel punto x^k. L'ammissibilità della direzione d_k considerata è data dall'appartenenza del punto $\hat{x}^k$ all'insieme ammissibile S. Notiamo che il caso $z^k > 0$ non si può mai realizzare in quanto $\forall x^k \in S$ si ha che la funzione obiettivo in x^k vale 0.

Algorithm 8 Metodo di Frank-Wolfe

Require: $x^0 \in S$
  $k = 0$
  while $||\hat{x}^k - x^k|| \geq \epsilon$ **do**
    Calcola $\hat{x}^k \in \arg\min_{x \in S} \nabla f(x^k)^T (x - x^k)$
    $d_k = \hat{x}^k - x^k$
    Calcolo α_k lungo d_k con una ricerca di linea Armijo
    $x^{k+1} = x^k + \alpha_k d_k$
    $k++$
  end while

Proposizione 16.1 (Proprietà di convergenza metodo di Frank-Wolfe)**.** Sia S convesso e compatto, $f \in C^1(\mathbb{R}^n)$ e $\{x^k\}$ la sequenza prodotta dal metodo di Frank-Wolfe. La sequenza $\{x^k\}$ ammette punti di accumulazione, ciascuno dei quali è stazionario

Proof. Per costruzione l'algoritmo genera punti ammissibili ad ogni passo $\implies \{x^k\} \subset S$. Per la compattezza di S si ha quindi l'esistenza dei punti di accumulazione della sequenza $\{x^k\}$. Dobbiamo quindi dimostrare che tali punti sono tutti punti stazionari.

Sia $K \subset \{0, 1, ...\}$ tale che:

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

Sia inoltre $d_k = \hat{x}^k - x^k$.

$$||d_k|| = ||\hat{x}^k - x^k|| \leq ||\hat{x}^k|| + ||x^k||$$

Poichè $\hat{x}^k, x^k \in S \subset \mathbb{R}^n$ compatto(e quindi anche limitato per heine borel) si ha che per $M > 0$ vale $||\hat{x}^k|| \leq M, ||x^k|| \leq M$, da cui otteniamo

$$||d_k|| \leq 2M \ \ \forall k$$

abbiamo quindi che la sequenza $\{d_k\}$ è limitata. Per la limitatezza di $\{d_k\}$ si ha che $\exists K_1 \subset K$ tale che

$$\lim_{\substack{k\to\infty \\ k\in K_1}} d_k = \bar{d}$$

Per le proprietà della ricerca di Armijo vincolata si ha:

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

Sia $y \in S$. Per definizione di $\hat{x}^k$ deve valere:

$$\nabla f(x^k)^T(\hat{x}^k - x^k) \leq \nabla f(x^k)^T(y - x^k) \ \ \forall y \in S, \forall k$$

Passando al limite si ottiene:

$$0 = \nabla f(\bar{x})^T \bar{d} = \lim_{\substack{k\to\infty \\ k\in K_1}} \nabla f(x^k)^T d_k \leq \lim_{\substack{k\to\infty \\ k\in K_1}} \nabla f(x^k)^T(y - x^k) = \nabla f(\bar{x})^T(y - \bar{x})$$

Da cui si ha che

$$\nabla f(\bar{x})^T(y - \bar{x}) \geq 0 \ \ \forall y \in S$$

Per la condizione necessaria e sufficente di ottimalità nel caso convesso si ha che $\bar{x}$ è stazionario

□

17 Problemi con vincoli in forma analitica

Un problema con vincoli in forma analitica è un problema di ottimizzazione vincolata della forma:

$$\min_{\substack{x\in\mathbb{R}^n \\ h(x)=0 \\ g(x)\leq 0}} f(x)$$

dove le funzioni vettoriali

$$h : \mathbb{R}^n \to \mathbb{R}^p$$

$$g : \mathbb{R}^n \to \mathbb{R}^m$$

rappresentano rispettivamente p vincoli di uguaglianza e m vincoli di disuguaglianza:

$$h_i(x) = 0 \ \ \forall i = 1, .., p$$

$$g_i(x) \leq 0 \ \ \forall i = 1, ..., m$$

Proposizione 17.1 (Condizioni di Fritz-John)**.** Sia $x^* \in \mathbb{R}^n$ e f,g,h funzioni continuamente differenziabili. Se x^* è punto di minimo locale allora esistono $\lambda_0 \in \mathbb{R}, \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^m$ **moltiplicatori** tali che:

- Valgono le **condizioni di ammissibilità**:
  - $h(x^*) = 0$
  - $g(x^*) \leq 0$
- Valgono le **condizione di complementarietà**:
  - $\mu_i g_i(x^*) = 0 \ \ \forall i = 1, ..., m$
- Valgono le **Condizioni di ammissiblità duale**
  - $\lambda_0, \mu_i \geq 0 \ \ \forall i = 1, ..., m$
- $(\lambda_0, \lambda, \mu) \neq (0, 0, 0)$(moltiplicatori non tutti nulli)
- Si ha l'annullamento del gradiente della **funzione Lagrangiana** associata al problema:

$$\nabla \mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 \nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0$$

La funzione Lagrangiana associata al problema è data da:

$$\mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 f(x) + \sum_{i=1}^{p} \lambda_i h_i(x^*) + \sum_{i=1}^{m} \mu_i g_i(x^*)$$

Consideriamo ad esempio il problema:

$$\min_{y^2=0} f(x, y)$$

Se (x^*, y^*) è un punto di minimo locale allora valgono le condizioni di Fritz John:

$$\exists \lambda_0, \lambda_1 \in \mathbb{R} \implies \begin{cases} (y^*)^2 = 0 \\ \lambda_0 \geq 0 \\ (\lambda_0, \lambda_1) \neq (0, 0) \\ \lambda_0 \nabla f(x^*, y^*) + \lambda_1 \begin{pmatrix} 0 \\ 2y^* \end{pmatrix} = 0 \end{cases}$$

Dove $\mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 f(x, y) + \lambda_1 y^2 \implies \nabla \mathcal{L} = \lambda_0 \nabla f(x^*, y^*) + \lambda_1 \begin{pmatrix} 0 \\ 2y \end{pmatrix}$
Possiamo quindi scegliere $\lambda_0 = 0, \lambda_1 = 1$: tali moltiplicatori rispettano le condizioni di Fritz-John. Le condizioni di Fritz John con i valori dei moltiplicatori specificati implicano che un eventuale soluzione del problema è un punto della forma $(x, 0)$. Ma tali punti costituiscono l'insieme dei punti ammissibili per il problema! Pertanto in questo caso le condizioni di Fritz John non aggiungono alcuna informazione utile alla soluzione del problema. Questo esempio è utile per mostrare che le condizioni di Fritz John sono condizioni **deboli**. Inoltre, per $\lambda_0 = 0$, nelle condizioni di Fritz John non si considera più la funzione obiettivo $f(x)$. Dobbiamo quindi determinare un insieme di condizioni necessarie di ottimalità che siano più forti e che tengano conto del valore della funzione obiettivo a prescindere dai valori ottenuti per i moltiplicatori.

18 Condizioni KKT(Karush-Kuhn-Tucker)

Proposizione 18.1. Se x^* è un punto di minimo locale e valgono delle condizioni di regolarità dei vincoli(**constraint qualification**) allora $\exists \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^m$:

- $h_i(x^*) = 0 \ \ \forall i = 1, ..., p$
- $g_i(x^*) \leq 0 \ \ \forall i = 1, ..., m$
- $\mu_i g_i(x^*) = 0 \ \ \forall i = 1, ..., m$

- $\mu_1, ..., \mu_m \geq 0 \ \ \forall i = 1, ..., m$
- $$\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0$$
- Quest'utlima condizione è riscrivibile come:

$$-\nabla f(x^*) = \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*)$$

L'antigradiente è combinazione lineare dei gradienti dei vincoli. Per la condizione $\mu_i g_i(x^*) = 0$ si ha che $\mu_i = 0$ se il vincolo $g_i(x^*) \leq 0$ non è attivo

18.1 Constraint qualification

Le condizioni di regolarità dei vincoli, o **Constraint qualification** rappresentano un insieme di ipotesi necessarie sulle funzioni che realizzano i vincoli di uguaglianza e disuguaglianza all'applicabilità delle condizioni KKT come condizioni necessarie di ottimalità. In seguito sono presentati alcuni esempi di constraint qualification:

- **Linear constraint qualification(LCQ)**: Tutti i vincoli sono lineari
- **Linear Independence constraint (LICQ)**: I vettori:

$$\nabla h_i(x^*) \ \ \forall i = 1, ..., p$$

$$\nabla g_i(x^*) \ \ \forall i \in I(x^*)$$

  sono linearmente indipendenti

- **Mangasarian-Fromovitz constraint qualification(MFCQ)**: I vettori:

$$\nabla h_i(x^*) \ \ \forall i = 1, ..., p$$

  sono linearmente indipendenti ed esiste una direzione $d \in \mathbb{R}^n$ tale che:

$$\nabla h_i(x^*)^T d = 0 \ \ \forall i = 1, ..., p$$

$$\nabla g_i(x^*)^T d < 0 \ \ \forall i \in I(x^*)$$

- **Slater constraint qualification(SCQ)**: Se f,g,h sono convesse ed esiste un

punto $x^* \in \mathbb{R}^n : h_i(x^*) = 0 \ \ \forall i = 1, .., p$:

$$g_i(x^*) < 0 \ \ \forall i = 1, ..., m$$

Proposizione 18.2. Sia x^* un punto che soddisfa le condizioni di Fritz-John. Se x^* soddisfa LICQ per un problema di ottimizzazione con vincoli in forma analitica, allora x^* soddisfa le condizioni KKT per lo stesso problema.

Proof. Sia x^* un punto che soddisfa le condizioni di Fritz-John, si ha che $\exists \lambda_0 \in \mathbb{R}, \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^n$ tali che:

- $h(x^*) = 0$
- $g(x^*) \leq 0$
- $\lambda_0, \lambda, \mu \geq 0$
- $\mu_i g_i(x^*) = 0 \ \ \forall i$
- $(\lambda_0, \lambda, \mu) \neq (0, 0, 0)$
- $$\nabla \mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 \nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0$$

Per assurdo sia $\lambda_0 = 0$. La condizione sull'annullamento del gradiente della Lagrangiana diventa:

$$\sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0$$

Per i vincoli di complementarietà si ha che $\mu_i = 0 \ \ \forall i$ tale che $g_i(x^*) < 0$, ossia $\mu_i = 0 \ \ \forall i \notin I(x^*)$. Possiamo quindi riscrivere la condizione sul gradiente della Lagrangiana come:

$$\sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{i \in I(x^*)} \mu_i \nabla g_i(x^*) = 0$$

Per LICQ si ha che i vettori

$$\nabla h_i(x^*), i = 1, ..., p$$

$$\nabla g_i(x^*), i \in I(x^*)$$

sono linearmente indipendenti. Pertanto l'unico modo per ottenere una combinazione lineare nulla di tali vettori è avere tutti gli scalari nulli:

$$\lambda_i = 0 \ \ \forall i, \mu_i = 0 \forall i \in I(x^*)$$

Inoltre per i vincoli di complementarietà si ha che $\mu_i = 0 \ \ \forall i \notin I(x^*)$, da cui si ottiene che tutti i moltiplicatori devono essere nulli:

$$(\lambda_0, \lambda, \mu) = (0, 0, 0)$$

Ciò viola le condizioni di Fritz-John, pertanto si ha un assurdo. □

Proposizione 18.3. Le KKT sono condizioni necessarie di ottimalità, ma nel caso di funzione obiettivo convessa le KKT diventano condizioni necessarie e sufficienti di ottimalità globale.

Consideriamo il caso di vincoli di simplesso:

$$S = \{x \in \mathbb{R}^n : e^T x = 1, x \geq 0\}$$

Proposizione 18.4. Se $\bar{x}$ è ottimale vale:

$$\frac{\partial f(\bar{x})}{\partial x_i} \geq \frac{\partial f(\bar{x})}{\partial x_j} \ \ \forall i, j : \ \ \bar{x}_j > 0$$

L'insieme ammissibile può essere riscritto nella seguente forma:

$$\begin{cases} e^T x - 1 = 0 \\ -x_h \leq 0 \ \ \forall h = 1, ..., n \end{cases}$$

Scriviamo le condizioni KKT (LCQ soddisfatto):

$$\begin{cases} \mu_1, ..., \mu_h \geq 0 \\ \mu_i(-x_i) = 0 \ \ \forall i = 1, ..., n \end{cases}$$

Il Lagrangiano è dato da:

$$\mathcal{L} = f(x) + \lambda(e^T x - 1) + \sum_{i=1}^{n} \mu_i(-x_i)$$

Consideriamo l'i-esima componente del gradiente del Lagrangiano, ossia la derivata parziale del lagrangiano rispetto all'i-esima variabile:

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} + \lambda - \mu_i = 0 \ \ \forall i$$

Per le condizioni di complementarietà si ha che se $x_j > 0 \implies \mu_j = 0$ in quanto deve

valere $-x_j\mu_j = 0 \ \ \forall j$. Pertanto si ha:

$$\frac{\partial f(x)}{\partial x_i} + \lambda = 0 \implies \frac{\partial f(x)}{\partial x_i} = -\lambda$$

Possiamo inoltre considerare che :

$$-\lambda \leq -\lambda + \mu_i \ \ \forall i$$

dato che $\mu_i \geq 0$. Ma avevamo ricavato che:

$$-\lambda + \mu_i = \frac{\partial f(x)}{\partial x_i}$$

Da cui si ha che $\forall j : x_j > 0$ e $\forall i = 1, ..., n$:

$$\frac{\partial f(x)}{\partial x_j} \leq \frac{\partial f(x)}{\partial x_i}$$

19 Modelli lineari di classificazione binaria

19.1 Il problema dell'apprendimento supervisionato

Il **problema dell'apprendimento supervisionato** consiste nell'elaborare automaticamente previsioni sui valori di uscita di un sistema rispetto ad un input sulla base di una serie di esempi ideali, costituiti da coppie di input e di output, che gli vengono inizialmente forniti. Si cerca quindi di determinare una funzione:

$$h : \mathbb{R}^n \to Y$$

che associ ad un input $x \in \mathbb{R}^n$ il corrispettivo output in Y. Se $Y = \mathbb{R}$ il problema si dice di **regressione**, mentre se Y è un insieme finito e numerabile il problema si dice di **classificazione**. Un caso particolare è rappresentato dai **problemi di classificazione binaria**, nei quali si cerca di determinare a quale tra due classi appartiene l'input dato. In questo caso, per convenzione, le classi sono identificate dai valori $\{-1, 1\}$. In generale è difficile deterinare l'espressione esatta della funzione h, per cui dati i parametri w di un particolare modello di classificazione/regressione, si cerca:

$$\hat{h}(x, w) \approx h(x) \ \ \forall x \in \mathbb{R}^n$$

In generale non si ha la conoscenza di tutto lo spazio, ma si conosce solo un **Dataset**:

$$D = \{(x_i, y_i) : y_i = h(x_i), i = 1, ..., n\}$$

ossia un insieme di **esempi**, ovvero coppie input-predizione corretta. Il dataset può essere poi ulterioremente diviso in **training set**, ossia la parte del dataset utilizzata in fase di addestramento del modello(e quindi di inferenza dei suoi parametri) e **test set**, ossia la parte del dataset impiegata per la verifica della correttezza dei risultati restituiti dal modello. Si introduce una funzione di errore:

$$e : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$$

per misurare la differenza tra $y_i = h(x_i)$ e l'approssimazione $\hat{h}(w, x_i)$. L'obiettivo è quindi minimizzare l'errore atteso rispetto ai parametri del modello di classificazione/regressione:

$$\min_w \mathbb{E}[e(\hat{h}(x, w), y)]$$

Tuttavia non possiamo supporre di conoscere a priori la distribuzione che ha generato i dati, e quindi non sappiamo calcolare il valore atteso dell'errore che si compie nell'approssimare il reale valore di $h(x)$ con $\hat{h}(w, x)$. Non potendo calcolare il valore atteso, si minimizza il **rischio empirico**, ossia la media:

$$\min_w \sum_{i=1}^{n} e(h(x_i), w), y_i)$$

La funzione $e(h(x_i), w), y_i)$ prende il nome di funzione **Loss**. Si definisce quindi il problema di **Empirical Risk Minimization(ERM)** come il seguente problema di ottimizzazione:

$$\min_w \mathcal{L}(w, x, y) + \lambda\Omega(w)$$

$$\mathcal{L}(w, x, y) = \frac{1}{N} \sum_{i=1}^{N} l(w, x^{(i)}, y^{(i)})$$

19.2 Classificazione Binaria con modelli lineari

. Dato un input $x \in \mathbb{R}^n$, il problema di classificazione binaria consiste nel decidere a quale $y \in \{-1, 1\}^n$ appartiene l'input x. Risolvere un problema di classificazione binaria equivale a trovare una funzione $h(x)$ che definisca una superficie di separazione tra esempi di classe positiva e negativa. Nel fare ciò si può incorrere in problemi di:

- **Underfitting**: Non esiste un classificatore lineare che separa gli esempi nel

dataset con loss = 0 $\implies$ il dataset non è **linearmente separabile**

- **Overfitting**: Tutti gli esempi del train set sono stati classificati correttamente, ma gli esempi del test set non vengono classificati correttamente. In questo caso si impiega un regolarizzatore $\Omega(w)$.

Nel caso di classificatori binari lineari, la superficie di separazione è data da un iperpiano:

$$h(x) = w^T x + b$$

dove w L'iperpiano di separazione viene determinato risolvendo il problema di minimizzazione del rischio empirico:

$$\min_{w,b} \sum_{i=1}^{N} l(w^t x^{(i)} + b, y^{(i)}) + \Omega(w)$$

Solitamente per modelli lineari, al fine di evitare fenomeni di overfitting, si impiega un regolarizzatore quadratico $\Omega(w) = \frac{1}{2}||w||^2$. In un problema di classificazione binaria si possono impiegare diverse tipologie di funzione loss:

- **Log-Loss**:
$$l(x^{(i)}, y^{(i)}) = \log(1 + \exp -y^{(i)} h(w, x^{(i)}))$$

- **0-1 Loss**:
$$l(x^{(i)}, y^{(i)}) = \mathbb{1}\{y^{(i)} \neq h(w, x^{(i)})\}$$

  Problema: Loss non continua

- **Hinge-Loss**:
$$l(x^{(i)}, y^{(i)}) = \max\{0, 1 - y^{(i)} h(w, x^{(i)})\}$$

  Più il prodotto $y^{(i)} h(w, x^{(i)})$ è grande e più si ha confidenza sulla qualità della classificazione.

Il problema di classificazione binaria che utilizza una superficie di separazione lineare, un regolarizzatore quadratico e la hinge-loss come funzione di errore costituisce un modello di classificazione binaria detto **Support Vector Machines(SVM)**:

$$\min_{w,b} \frac{1}{2}||w||^2 + c \sum_{i=1}^{N} \max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}$$

La soluzione del problema **SVM** è l'**iperpiano a massimo margine**, ossia l'iperpiano che realizza la maggior separazione(detta **margine**) tra due esempi di classi distinte.

Il problema principale di questa rappresentazione della funzione obiettivo per un problema **SVM** consiste nella non differenziabilità della funzione obiettivo: ciò impedisce l'applicazione di qualsiasi metodo di ordine superiore al primo. Consideriamo quindi il problema di ottimizzazione vincolata:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i$$

$$\xi_i \geq 0 \ \ \forall i = 1, ..., N$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \ \ \forall i = 1, ..., N$$

E' possibile mostrare che l'ottimo del precedente problema di ottimizzazione vincolata corrisponde all'ottimo del problema **SVM** non vincolato. Infatti:

$$\begin{cases} \xi_i \geq 1 - y^{(i)}(w^T x^{(i)} + b) \\ \xi_i \geq 0 \ \ \forall i = 1, ..., N \end{cases} \implies \xi_i \geq \max\{0, 1 - y^{(i)}(w^T x^{(i)} + b)\}$$

In una soluzione ammissibile ξ. Pertanto minimizzando $\xi_i \ \ \forall i = 1, ..., N$ si ottiene l'ottimo del problema precedente. Abbiamo quindi mostrato l'equivalenza tra le soluzioni dei due problemi presentati.

19.3 Hard margin SVM

Consideriamo ora il caso in cui $c \to \infty$, in tal caso si ha $\xi_i = 0 \ \ \forall i$ e il problema diventa:

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 \ \ \forall i$$

Se il dataset è **linearmente separabile** si possono scegliere due iperpiani di separazione tra le due classi tali per cui la distanza tra di essi è la più grande possibile. La distanza tra di essi è comunemente chiamata margine, e l'iperpiano a massimo margine sarà dato dall'iperpiano parallelo ai due iperpiani di separazione che giace a metà distanza tra i due iperpiani. Per un dataset normalizzato le equazioni dei due iperpiani sono date da:

$$w^T x + b = 1, \ \ w^T x + b = -1$$

Il primo iperpiano definisce una frontiera di separazione per la quale ogni esempio che giace sull'iperipano o sopra di esso sarà classificato come un esempio di classe avente etichetta 1. Analogamente, il secondo iperpiano definisce una superficie di separazione

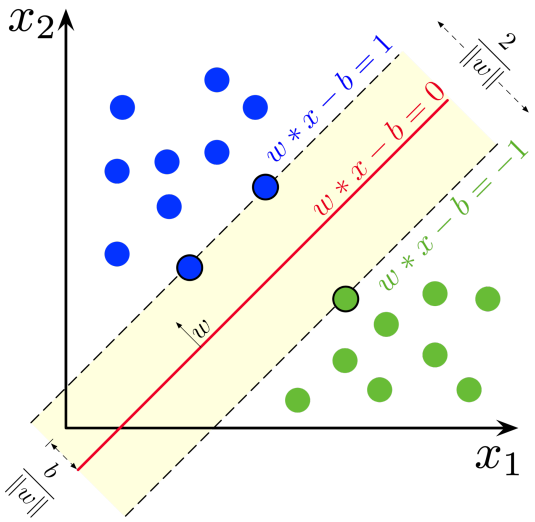


Figure 1: Hard-Margin SVM

per la quale ogni esempio che giace su di esso o sotto sarà classificato come un esempio avente classe -1. Geometricamente, la distanza tra i due iperpiani è $\frac{2}{||w||}$(vedi qui), pertanto per massimizzare la distanza tra i due iperpiani è necessario minimizzare $||w||$, in accordo con la definizione del problema. Inoltre si richiede anche che non vi siano esempi che giacciono nel margine definito dai due iperpiani, ossia si richiede che:

$$\begin{cases} w^T x^{(i)} + b \geq 1 & y^{(i)} = 1 \\ w^T x^{(i)} + b \leq -1 & y^{(i)} = -1 \end{cases} \implies y^{(i)}(w^T x^{(i)} + b) \geq 1 \quad \forall i$$

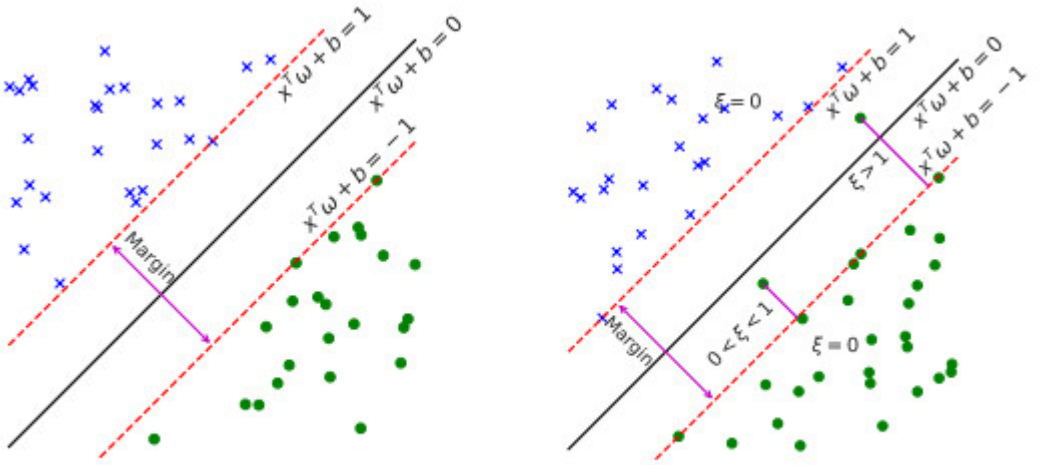
I vincoli sopra esplicitati specificano che ogni esempio deve trovarsi dalla parte corretta del margine. Il problema **SVM** che si ottiene per $c \to \infty$ si dice **Hard-Margin SVM**[1]: i valori dei parametri w, b ottenuti risolvendo tale problema determinano il classificatore binario:

$$x \to sgn(w^T x + b)$$

Un importante conseguenza dell'interpretazione geometrica del problema di SVM hard margin è il fatto che l'iperpiano di separazione a massimo margine è completamente determinato dagli esempi $(x^{(i)}, y^{(i)}) \subset D$ che si trovano a distanza minima dallo stesso. Tali esempi si dicono **Vettori di supporto**, e nel caso di dataset lineamente separabile tali esempi giacciono sugli iperpiani di separazione che definiscono il margine.

19.4 Soft margin SVM e Duale di Wolfe

Vi sono casi di Dataset per i quali il requisito di Hard margin è troppo stringente: tali casi riguardano dataset non linearmente separabili(dataset per i quali non si riesce a determinare un separatore lineare che definisca un classificatore a loss nulla), e dataset



(a) Hard margin. (b) Soft margin.

che presentano outliers. Rilassando il requisito di hard margin torniamo all'espressione precendentemente ricavata del problema SVM con decomposizione della hinge-loss. Tale espressione del problema SVM si dice problema **SVM Soft margin**.

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + c\sum_{i=1}^{N}\xi_i$$

$$\xi_i \geq 0 \quad \forall i = 1, ..., N$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \quad \forall i = 1, ..., N$$

Le variabili $\xi_i \;\; i = 1, ..., n$ si dicono variabili **Slack**, e misurano l'errore relativo ad esempi interni al margine o mal classificati. In particolare:

- $\xi_i = 0$ se l'esempio si trova dalla parte corretta rispetto al margine
- $0 < \xi_i < 1$ se l'esempio si trova dalla parte corretta rispetto all'iperpiano di classificazione, ma non rispetto al margine
- $\xi_i \geq 1$ se l'esempio è mal classificato, ovvero si trova dalla parte errata dell'iperpiano di classificazione

Proposizione 19.1 (Duale di Wolfe)**.** Siano $f : \mathbb{R}^n \to \mathbb{R}$ e $g : \mathbb{R}^n \to \mathbb{R}^m$ funzioni convesse e differenziabili. Sia poi $x^* \in \mathbb{R}^n$ una soluzione ottima del problema:

$$\min_{g(x)\leq 0} f(x)$$

Sia inoltre $\mu^* \in \mathbb{R}^m$ un vettore di moltiplicatori tale che la coppia (x^*, μ^*) soddisfa le condizioni KKT.

Allora la coppia (x^*, μ^*) è soluzione ottimale del problema **Duale di Wolfe**:

$$\max_{x,\mu} f(x) + \mu^T g(x) = \mathcal{L}(x, \mu)$$

$$\mu \geq 0$$

$$\nabla_x \mathcal{L}(x, \mu) = 0$$

dove $\mathcal{L}(x, \mu)$ denota il Lagrangiano del problema originale.

Proof. Per ipotesi la coppia (x^*, μ^*) soddisfa le condizioni KKT:

$$\begin{cases} \nabla_x \mathcal{L}(x^*, \mu^*) = 0 \\ \mu^* \geq 0 \\ g_i(x^*)\mu_i^* = 0 \ \ \forall i \\ g(x^*) \leq 0 \end{cases}$$

Pertanto si ha:

$$\mathcal{L}(x^*, \mu^*) = f(x^*) + \sum_{i=1}^{m} \mu_i^* g_i(x^*) = f(x^*)$$

Poiche per i vincoli di complementarietà si ha $\mu_i^* g_i(x^*) = 0 \ \ \forall i$. Sia ora $\bar{x}, \bar{\mu}$ una soluzione ammissibile del problema duale di Wolfe. Si ha che:

$$\mathcal{L}(x^*, \mu^*) = f(x^*) \geq f(x^*) + \sum_{i=1}^{m} \bar{\mu}_i g_i(x^*) = \mathcal{L}(x^*, \bar{\mu})$$

Poichè $g_i(x^*) \leq 0$ e $\bar{\mu}_i \geq 0$ per le condizioni KKT. Si può inoltre osservare che $\mathcal{L}(x, \mu)$ è convessa rispetto a x, poichè combinazione lineare di funzioni convesse a coefficienti positivi. Pertanto in ogni punto $\mathcal{L}(x, \mu)$ sarà maggiorata da un iperpiano tangente ad un punto dato(in questo caso $(\bar{x}, \bar{\mu})$):

$$\mathcal{L}(x^*, \bar{\mu}) \geq \mathcal{L}(\bar{x}, \bar{\mu}) + \nabla_x \mathcal{L}(x^*, \mu^*)(x^* - \bar{x})$$

Poichè valgono le KKT in (x^*, μ^*) si ha che $\nabla_x \mathcal{L}(x^*, \mu^*) = 0$ implica:

$$\mathcal{L}(x^*, \bar{\mu}) \geq \mathcal{L}(\bar{x}, \bar{\mu})$$

Componendo le due disuguaglianze ottenute per $\mathcal{L}(x^*, \mu^*)$ e $\mathcal{L}(x^*, \bar{\mu})$ si ha

$$\mathcal{L}(x^*, \mu^*) \geq \mathcal{L}(x^*, \bar{\mu}) \geq \mathcal{L}(\bar{x}, \bar{\mu})$$

Ma $(\bar{x}, \bar{\mu})$ è una generica soluzione ammissibile del duale, ed inoltre (x^*, μ^*) è soluzione ammissibile per il duale $\implies$ (x^*, μ^*) è soluzione ottima del duale per la disuguaglianza ricavata. □

Possiamo quindi applicare la proposizione appena dimostrata sul duale di Wolfe al problema SVM Soft- margin per determinare il problema SVM duale. Consideriamo quindi il problema SVM Soft margin riscritto con vincoli di disugaglianza in forma standard e associamo a tali vincoli un set di moltiplicatori:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i$$

$$\xi_i > 0 \ \ \forall i = 1, ..., N \implies -\xi_i \leq 0 \implies \mu_i$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \ \ \forall i = 1, ..., N \implies -\xi_i - y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0 \implies \alpha_i$$

Applichiamo la proposizione sul duale di Wolfe ottenendo il problema:

$$\max_{w,b,\xi,\alpha,\mu} \mathcal{L}(w, b, \xi, \alpha, \mu)$$

$$\mu \geq 0, \alpha \geq 0$$

$$\nabla_w \mathcal{L}(w, b, \xi, \alpha, \mu) = 0$$

$$\nabla_b \mathcal{L}(w, b, \xi, \alpha, \mu) = 0$$

$$\nabla_\xi \mathcal{L}(w, b, \xi, \alpha, \mu) = 0$$

Dove il Lagrangiano associato al problema originale ha la seguente espressione:

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N}(-\xi_i)\mu_i + \sum_{i=1}^{N} \alpha_i(-\xi_i - y^{(i)}(w^T x^{(i)} + b) + 1)$$

(da ora in poi $\mathcal{L}(w, b, \xi, \alpha, \mu) = \mathcal{L}$). Imponiamo che i gradienti del lagrangiano rispetto alle variabili del problema originale siano nulli, come da definizione del duale di Wolfe:

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} \implies w = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$$

$$\nabla_b \mathcal{L} = -\sum_{i=1}^{N} y^{(i)} \alpha_i = 0 \implies \sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = c - \mu_i - \alpha_i = 0 \implies \alpha_i = c - \mu_i$$

Pertanto il problema SVM duale avrà il seguente insieme ammissibile:

$$\begin{cases} w = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} & (1) \\ \sum_{i=1}^{N} \alpha_i y^{(i)} = 0 & (2) \\ \alpha_i = c - \mu_i \ \ \forall i & (3) \end{cases}$$

Possiamo inoltre trovare un altra espressione per la funzione obiettivo del problema duale. Consideriamo il Lagrangiano ottenuto separando i termini dell'ultima sommatoria:

$$\mathcal{L} = \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} -\mu_i \xi_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i + \sum_{i=1}^{N} -\alpha_i y^{(i)} (w^T x^{(i)} + b)$$

Possiamo riesprimere il Lagrangiano sfruttando i vincoli (1),(2) e (3) considerando che:

$$\sum_{i=1}^{N} -\alpha_i y^{(i)} (w^T x^{(i)} + b) = -\sum_{i=1}^{N} \alpha_i y^{(i)} (w^T x^{(i)}) - b \sum_{i=1}^{N} \alpha_i y^{(i)}$$

e che:

$$\sum_{i=1}^{N} c\xi_i - \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} \alpha_i \xi_i = \sum_{i=1}^{N} (c - \mu_i - \alpha_i) \xi_i$$

Da cui si ha che:

$$\mathcal{L} = \frac{1}{2} w^T w + \sum_{i=1}^{N} (c - \mu_i - \alpha_i) \xi_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y^{(i)} (w^T x^{(i)}) - b \sum_{i=1}^{N} \alpha_i y^{(i)}$$

Ora per il vincolo (3)($\alpha_i = c - \mu_i \ \ \forall i$) si ha che $\sum_{i=1}^{N} (c - \mu_i - \alpha_i) \xi_i = 0$, e il vincolo (2) implica $\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$, per cui nell'espressione del Lagrangiano si ha, raccogliendo per w^T:

$$\mathcal{L} = w^T (\frac{1}{2} w - \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}) + \sum_{i=1}^{N} \alpha_i$$

Ora per il vincolo (1) si ha che $w = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$, da cui possiamo scrivere:

$$\mathcal{L} = -\frac{1}{2} w^T w + \sum_{i=1}^{N} \alpha_i$$

Riapplicando nuovamente il vincolo (3) all'espressione ottenuta per il Lagrangiano della

funzione obiettivo si ha:

$$\mathcal{L} = -\frac{1}{2}[\sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}]^T [\sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}] + e^T\alpha = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} + e^T\alpha$$

Ricordando che $\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} = \alpha^T Q\alpha$, dove l'elemento di riga i e colonna j della matrice Q è dato da $Q_{ij} = y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)}$ si ha che il Lagrangiano associato al problema SVM Soft margin con i vincoli (1),(2) e (3) può essere espresso nella forma:

$$\mathcal{L} = -\frac{1}{2}\alpha^T Q\alpha + e^T\alpha$$

Possiamo scrivere l'espressione completa del duale di Wolfe del problema SVM, detto anche **problema SVM duale**(ricordando che il duale di Wolfe è un problema di massimo, pertanto per ottenre un problema di minimo si deve cambiare di segno la funzione obiettivo):

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$

$$\alpha \geq 0, \quad \alpha \leq c$$

$$\alpha^T y = 0$$

Dove il vincolo $\alpha \leq c$ deriva dal vincolo $\alpha_i = c - \mu_i, \mu_i \geq 0 \ \ \forall i$. Osserviamo che il problema SVM duale è un problema quadratico con vincoli lineari, pertanto è risolvibile con le tecniche viste per i problemi di ottimizzazione vincolata su problemi quadratici. Inoltre se Q è definita positiva, la funzione obiettivo è strettamente convessa. Risolvendo il duale SVM, e quindi determinando la soluzione ottima α^* del problema SVM duale, è possibile ricavare la soluzione del primale considerando che:

$$\begin{cases} w^* = \sum_{i=1}^{N} \alpha_i^* y^{(i)} x^{(i)} \\ \mu_i^* = c - \alpha_i^* \\ b^* = \frac{y^{(i)}}{(w^*)^T x^{(i)}} & i : \alpha_i^* \in (0, c) \\ \xi_i^* = \max\{0, 1 - y^{(i)}((w^*)^T x^{(i)} + b^*)\} \end{cases}$$

Sappiamo inoltre che i moltiplicatori α^*, μ^* devono soddisfare le KKT. Avendo associato α_i al vincolo $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$, portando il vincolo di disuguaglianza in forma standard si ha:

$$1 - \xi_i^* - y^{(i)}((w^*)^T x^{(i)} + b^*) \leq 0 \implies \alpha_i$$

Da cui otteniamo la seguente condizione per α_i^*:

$$\alpha_i^*(1 - \xi_i^* - y^{(i)}((w^*)^T x^{(i)} + b^*)) = 0$$

Analogamente, avendo associato μ_i al vincolo $-\xi_i \leq 0$ si ottiene la condizione per μ_i^*:

$$\xi_i^* \mu_i^* = 0 \implies \xi_i^*(c - \alpha_i^*) = 0$$

Adesso consideriamo i casi possibili per le componenti del moltiplicatore α^*, che ricordiamo essere limitate nell'intervallo $[0, c]$:

- Caso $\alpha_i^* \in (0, c)$: Per la condizione $\xi_i^*(c - \alpha_i^*) = 0$ si ha:

  $$\xi_i^* = 0$$

  Per la condizione $\alpha_i^*(1 - \xi_i^* - y^{(i)}((w^*)^T x^{(i)} + b^*)) = 0$, considerando che $\alpha_i^* > 0$ si ha:

  $$1 - y^{(i)}((w^*)^T x^{(i)} + b^*) = 0 \implies y^{(i)}((w^*)^T x^{(i)} + b^*) = 1$$

  I moltiplicatori $\alpha_i \in (0, c)$ sono associati ad esempi $(x^{(i)}, y^{(i)})$ per cui la variabile slack è nulla, ossia esempi che si trovano dalla parte corretta rispetto al margine, e quindi esempi per i quali non si paga alcuna penalità. Inoltre la seconda condizione equivale ad affermare che il punto $(x^{(i)}, y^{(i)})$ si trova esattamente sul bordo del margine corretto. Ciò implica che il punto $x^{(i)}$ è un **vettore di supporto**.

- Caso $\alpha_i^* = 0$: Per la condizione $\xi_i^*(c - \alpha_i^*) = 0$ si ha:

  $$\xi_i^* = 0$$

  Da cui ricordando il vincolo $y^{(i)}((w^*)^T x^{(i)} + b^*) \geq 1 - \xi_i^*$ del problema primale si ha:

  $$y^{(i)}((w^*)^T x^{(i)} + b^*) \geq 1$$

  Osserviamo quindi che gli esempi aventi moltiplicatori $\alpha_i^* = 0$ sono esempi aventi variabile slack nulla, e che si trovano dalla parte corretta rispetto al margine

- Caso $\alpha_i^* = c$. In questo caso la condizione $\xi_i^*(c - \alpha_i^*) = 0$ implica che l'esempio ξ potrebbe (non ne si ha la certezza) essere tale che:

  $$\xi_i^* > 0$$

  Se tale caso si verifica, per la condizione $\alpha_i^*(1 - \xi_i^* - y^{(i)}((w^*)^T x^{(i)} + b^*)) = 0$ si ha:

  $$y^{(i)}((w^*)^T x^{(i)} + b^*) = 1 - \xi_i^* \leq 1$$

  In questo caso l'esempio ξ potrebbe essere stato classificato male o trovarsi all'interno del margine.

Prendiamo come vettori di supporto tutti i vettori associati al moltiplicatori $\alpha_i^* > 0$: consideriamo quindi gli esempi che si trovano esattamente sull'iperpiano relativo alla zona corretta rispetto al margine (caso $\alpha_i^* \in (0, c)$ oppure gli esempi che hanno varaibile slack $\xi_i > 0$, ovvero esempi mal classificati/ interni al margine (case $\alpha_i^* = c$). L'iperpiano di separazione ottenuto è dato da:

$$w^* = \sum_{i=1}^{N} \alpha_i^* y^{(i)} x^{(i)}$$

Osserviamo quindi che solo gli esempi associati a moltiplicatori α_i^* postivi, e quindi solo i vettori di supporto, contribuiscono alla definizione dell'iperpiano di separazione: difatti è possibile affermare che w^* è combinazione lineare di vettori di supporto, pesata dai moltiplicatori α_i^*.

19.5 Kernel Trick

Consideriamo l'iperpiano passante per l'origine:

$$(w^*)^T x = \sum_{i=1}^{N} \alpha_i^* y^{(i)} x^T x^{(i)}$$

Possiamo vedere il prodotto scalare $x^T x^{(i)}$ come una misura di similarità tra vettori. Difatti, dati due vettori $x, y \in \mathbb{R}^n$, più il loro prodotto scalare $x \cdot y$ è vicino ad 1, maggiore sarà il grado di similitudine attribuibile ai due vettori(sfruttando il principio della cosine similarity). Nel problema SVM duale, l'elemento in riga i e colonna j della matrice Q è dato da:

$$Q_{ij} = y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)}$$

Vi sono casi di dataset non linearmente separabili per i quali se $\xi \in \mathbb{R}^p$ non si riesce a determinare un separatore lineare (p-1) dimensionale in grado di classificare i dati con una loss accettabile : in questi casi si ricorre al **kernel trick**. Una **funzione Kernel** è una funzione:

$$\mathcal{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

che sostituisce il prodotto scalare come misura di similarità all'interno dell'espressione dell'iperpiano di separazione ottimo. In particolare si ha:

$$h(x) = \sum_{i=1}^{N} \alpha_i^* y^{(i)} \mathcal{K}(x, x^{(i)})$$

Per utilizzare una funzione Kernel in SVM esso deve essere **valido**. Seguono due

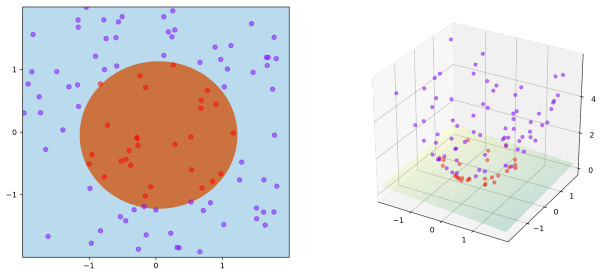


Figure 2: Kernel trick

possibili definizioni di validità di un kernel:

Definizione 19.2. Una funzione kernel $\mathcal{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ si dice **valida** se vale una delle seguenti affermazioni:

- La matrice Q : $Q_{ij} = y^{(i)}y^{(j)}x^{(i)^T}x^{(j)}$ è semidefinita positiva per ogni possibile dataset $\mathcal{D}$.
- Esiste una **feature map** che mappa la funzione Kernel nel prodotto interno(scalare) tra feature map di due input distinti in uno spazio di dimensione pari o superiore a quello dei dati. Formalizzando:

$$\exists \phi : \mathbb{R}^n \to \mathbb{R}^P, \quad p \geq n : \quad \mathcal{K}(u, v) = \phi(u)^T \phi(v) \quad \forall u, v \in \mathbb{R}^n$$

Osserviamo che in generale calcolare il prodotto scalare $\phi(u)^T\phi(v)$ può essere dispendioso in termini di risorse computazionali, dato che lo spazio delle feature (ossia lo spazio degli input $x^{(i)}$ nel dataset trasformati mediante la feature map ϕ) può essere in generale uno spazio ad alta dimensione(se non infinita). Tuttavi, utilizzando una SVM con kernel, non è necessario calcolare esplicitamente i prodotti scalari tra feature map, in quanto la funzione kernel consente di calcolare il prodotto scalare tra i vettori delle feature trasformate nello spazio di dimensioni superiore senza mappare esplicitamente i dati in tale spazio. Un esempio di funzione kernel è il kernel RBF(gaussiano)

$$\mathcal{K}(u, v) = e^{-\gamma||u-v||^2}$$

Il duale di Wolfe di un problema SVM che impiega una funzione Kernel valida permette di costruire classificatori non lineari, in quanto nonostante il separatore ottenuto nello spazio delle feature sia a tutti gli effetti un separatore lineare, rimappando i dati nello spazio di partenza si ottiene in generale un classificatore non lineare. Nei primi anni di sviluppo dei classificatori basati su SVM, date le scarse dimensioni dei dataset ($\approx 10^3$), si impiegava il metodo di Frank-Wolfe per la soluzione del duale SVM. Tuttavia, con

l'aumentare delle dimensioni dei dataset il metodo di Frank-Wolfe non è più ideale per la soluzione del duale di Wolfe del problema SVM. Si ricorre quindi a **Metodi di decomposizione**

20 Metodi di Decomposizione

I metodi di decomposizione risolvono problemi di ottimizzazione vincolata del tipo:

$$\min_{\substack{x \in S \subseteq \mathbb{R}^n \\ x=(x_1,...,x_m)}} f(x_1, ..., x_m)$$

Dove $x_1 \in \mathbb{R}^{n_1}, ..., x_m \in \mathbb{R}^{n_m}$. In pratica si sono divise le variabili in m gruppi, ciascuno con la sua dimensionalità $n_i \;\; i = 1, ..., m$. I metodi di decomposizione agiscono risolvendo sottoproblemi della forma:

$$\arg\min_{x_i \in S(\bar{x}_1,..,\bar{x}_{i-1},\bar{x}_{i+1},...,\bar{x}_m)} f(x_i|\bar{x}_1, .., \bar{x}_{i-1}, \bar{x}_{i+1}, ..., \bar{x}_m))$$

ossia(non so se è vero o no) fissati i valori di $m-1$ gruppi di variabili a valori ammissibili, si minimizza la funzione obiettivo rispetto all'i-esimo gruppo di variabili nel rispettivo insieme ammissibile ottenuto a seguito della scelta delle altre variabili nei restanti $m-1$ gruppi. I metodi di decomposizione sono utili quando:

- $n = n_1 + n_2 + ... + n_m$ è grande
- La funzione obiettivo f è separabile rispetto a blocchi di variabili
- Rispetto ai singoli blocchi di variabili la funzione obiettivo f è convessa ma non lo è complessivamente
- I vincoli sono separabili(e.g vincoli di box):

$$(x_1, ..., x_m) \in S_1 \subseteq \mathbb{R}^{n_1} \times S_2 \subseteq \mathbb{R}^{n_2} \times ... \times S_m \subseteq \mathbb{R}^{n_m}$$

I metodi di decomposizione si dividono in diverse classi in base al loro funzionamento.

20.1 Metodi Sequenziali

Il **Metodo di Gauss-Seidel** consiste nel risolvere l'i- esimo sottoproblema e conseguentemente nell'impiegare la soluzione trovata nei blocchi precedenti:

$$x_i^{k+1} = \arg\min_{x_i \in S_i} f(x_1^{k+1}, ..., x_{i-1}^{k+1}, x_i, x_i^k, ..., x_m^k)$$

La sequenza prodotta dal metodo di Gauss-Seidel converge a punti stazionari:

- Se f è convessa
- Se f è strettamente convessa rispetto a $x_i \ \ \forall i = 1, .., m$
- se $m = 2$

20.2 Metodi Paralleli

Il **Metodo di Jacobi** calcola in modo parallelo dei possibili aggiornamenti e prende quello che garantisce il maggior decremento:

$$\xi_i^{k+1} = \arg\min_{x_i \in S_i} f(x_i^k, ..., x_i, ..., x_m^k)$$

$$\begin{cases} x^{k+1} = \arg\min f(x) \\ x = (x_1^k, ..., \xi_i^{k+1}, ..., x_m), \quad i = 1, ..., M \end{cases}$$

20.3 Metodi con sovrapposizione di blocchi

Ad ogni iterazione si sceglie un **Working set**:

$$W_k \subseteq \{1, ..., n\}$$

Aggiorno quindi:

$$\tilde{x}_{W_k}^k = \arg\min_{X_{W_k}} f(x_{W_k}, x_{\bar{W}_k}^k)$$

$$x_i^{k+1} = \begin{cases} x_i^k & i \in \bar{W}_k \\ \tilde{x}_i^k & i \in W_k \end{cases}$$

Dove $\bar{W}_k = \{1, ..., n\}/W_k$. La convergenza di un metodo di questa classe dipende dalla scelta del working set W_k. Seguono due possibili regole per la scelta del working set:

- **Regola ciclica**:

$$\exists m > 0 : \quad \forall i = 1, ..., m \ \ \forall k \in \{0, 1, ..\} \ \ \exists \bar{t} \leq m : \quad i \in W_{k+\bar{t}}$$

- **Regola di Gauss-Southwell**

$$\forall k \ \ \exists i(k) \in \{1, ..., n\} : i(k) \in W_k \ \text{ e risulta:}$$

$$\left| \frac{\partial f(x^k)}{\partial x_{i(k)}} \right| \geq \left| \frac{\partial f(x^k)}{\partial x_j} \right| \ \ \forall j \in \{1, ..., n\}$$

Torniamo al problema duale SVM:

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$

$$\alpha \geq 0, \quad \alpha \leq c$$

$$\alpha^T y = 0$$

Generalmente, per via delle grandi dimensioni dei Dataset, la matrice Q nel problema SVM duale è densa e grande, pertanto la sua memorizzazione diventa proibitiva all'aumentare della dimensione del problema(dati di addestramento). Un ulteriore difficoltà alla soluzione del duale SVM è la presenza del vincolo $\alpha^T y = 0$ che lega tutte le variabili. Utilizziamo una strategia di decomposizione con sovrapposizione di blocchi, cosi da non dover memorizzare l'hessiana Q: scegliamo $W \subseteq \{1, ..., n\}$ e otteniamo sottoproblemi del tipo:

$$\min_{\alpha_w} \frac{1}{2}\alpha_w^T Q_{ww}\alpha_w + \alpha_w^T(Q_{w\bar{w}}\alpha_{\bar{w}}^k - e)$$

$$0 \leq \alpha_w \leq c$$

$$\alpha_w^T y_w = -(\alpha_{\bar{w}})^{k^T} y_{\bar{w}}$$

In questo schema di decomposizione, ad ogni iterazione k, il vettore delle variabili α^k è partizionato in due sottovettori $(\alpha_W^k, \alpha_{\bar{W}}^k)$, e a partire dal punto ammissibile corrente $\alpha_k = (\alpha_W^k, \alpha_{\bar{W}}^k)$ il sottovettore α_W^{k+1} è determinato risolvendo il sottoproblema sopra riportato. Il sottovettore relativo alle variabili fuori dal working set, $\alpha_{\bar{w}}^k$ non viene modificato a seguito di un iterazione, vale a dire che si ha $\alpha_{\bar{w}}^{k+1} = \alpha_{\bar{w}}^k$. La soluzione corrente viene aggiornata ponendo $\alpha^{k+1} = (\alpha_W^{k+1}, \alpha_{\bar{w}}^{k+1})$. Dobbiamo ora determinare come scegliere la cardinalità e gli indici del working set.

Proposizione 20.1 (Cardinalità del working set)**.** La cardinalità $|W|$ del working set deve essere necessariamente maggiore o uguale a 2 per comportare una variazione del punto corrente all'interno di un algoritmo iterativo di decomposizione a sovrapposizione di blocchi.

Proof. Mostriamo solo il caso $|W| = 1$ si ha $\alpha^{k+1} = \alpha^k$. Infatti se $W = \{i\}$ si ha:

$$\alpha_i^{k^T} y_i = -(\alpha_{\bar{w}})^{k^T} y_{\bar{w}}$$

e affinche il successivo punto α_i^{k+1} sia ammissibile dovra risultare:

$$\alpha_i^{k+1^T} y_i = -(\alpha_{\bar{w}})^{k^T} y_{\bar{w}}$$

Da cui si ha che necessariamente deve essere

$$\alpha_i^{k+1} = \alpha_i^k$$

□

21 Sequential Minimal Optimization

Gli algoritmi di decomposizione con sovrapposizione di blocchi sono classificati in base alla cardinalità $|W|$ del working set W. Per $|W| = 2$ il sottoproblema è risolvibile in forma chiusa e si parla di algoritmi di tipo **Sequential Minimal Optimization**. Poichè il sottoproblema su working set a cardinalità $|W| = 2$ è risolvibile analiticamente, gli algoritmi SMO non richiedono l'utilizzo di un solver per risolvere il sottoproblema. Se $W = \{i, j\}$ il sottoproblema diventa:

$$\min_{\alpha_i,\alpha_j} \frac{1}{2} \begin{pmatrix} \alpha_i \\ \alpha_j \end{pmatrix}^T \begin{pmatrix} Q_{ii} & Q_{ji} \\ Q_{ij} & Q_{jj} \end{pmatrix} \begin{pmatrix} \alpha_i \\ \alpha_j \end{pmatrix} - \alpha_i - \alpha_j + \begin{pmatrix} \alpha_i & \alpha_j \end{pmatrix} P_{\bar{w}}$$

Dove $P_{\bar{W}} = (Q_{w\bar{w}}\alpha_{\bar{w}}^k)$. Il sottoproblema ottenuto è un problema quadratico e convesso nelle variabli (α_i, α_j) che può essere risolto in forma chiusa. Adesso focalizziamo l'attenzione sulle regole che determinano la scelta delle variabili nel working set W. Vorremmo che:

- Il punto $\alpha^{k+1} = (\alpha_1^k, ..., \alpha_i^{k+1}, ..., \alpha_j^{k+1}, ..., \alpha_n^k)$ sia ammissibile
- Vi sia uno stretto decremento della funzione obiettivo : $f(\alpha^{k+1}) < f(\alpha^k)$

In altri termini vogliamo scegliere una direzione:

$$d^{ij} = (0, ..., 0, d_i, 0, ..., 0, d_j, 0, .., 0)^T$$

che sia ammisibile e di discesa.

Proposizione 21.1 (Direzioni ammissibili nel problema SVM duale)**.** Sia $\bar{\alpha} \in S$, con $S = \{0 \leq \alpha \leq c, \alpha^T y = 0\}$. L'insieme delle direzioni ammissibili in $\bar{\alpha}$ è dato da:

$$D(\bar{\alpha}) = \left\{ d \in \mathbb{R}^n : d^T y = 0, \;\; d_i \geq 0 \;\; \forall i \in L(\bar{\alpha}), \;\; d_i \leq 0 \;\; \forall i \in U(\bar{\alpha}) \right\}$$

Dove gli insiemi $L(\bar{\alpha})$ e $U(\bar{\alpha})$ sono dati da:

$$\begin{cases} L(\bar{\alpha}) = \{i : \bar{\alpha}_i = 0\} \\ U(\bar{\alpha}) = \{i : \bar{\alpha}_i = c\} \end{cases}$$

Proof. Supponiamo che $D(\bar{\alpha})$ sia l'insieme delle direzioni ammissibili in $\bar{\alpha}$. Per $d \in D(\bar{\alpha})$ si ha quindi che $\bar{\alpha} + td \in S \ \ \forall t \in [0, \bar{t}]$ per $\bar{t}$ sufficientemente piccolo. Pertanto deve valere:

$$y^T(\bar{\alpha} + td) = 0 \implies y^T\bar{\alpha} + y^T td = 0$$

Poiche $\bar{\alpha}$ è ammissibile vale $y^T\bar{\alpha} = 0$, pertanto l'ammissibilità del nuovo punto ottenuto spostandosi lungo la direzione d ammissibile implica:

$$y^T d = 0 \iff d^T y = 0$$

Inoltre poichè il nuovo punto $\bar{\alpha}+td$ deve essere ammissibile si deve avere $0 \leq \bar{\alpha}+td \leq c$, il che implica:

$$\begin{cases} d_i \geq 0 & \alpha_i = 0 \\ d_i \leq 0 & \alpha_i = c \end{cases}$$

Componendo le due condizioni ottenute si ha la tesi. □

Stiamo quindi cercando una direzione $d^{ij} \in D(\bar{\alpha})$ ammissibile e con due sole componenti non nulle in corrispondenza degli indici i e j. Le direzioni $d \in D(\bar{\alpha})$ siffatte sono tali per cui:

$$(d^{i,j})^T y = 0 \implies d_i y_i + d_j y_j = 0$$

Possiamo quindi scegliere:

$$d_i = \frac{1}{y_i}, \ \ d_j = -\frac{1}{y_j}$$

Ora supponiamo che $i \in L(\alpha^k)$, in tal caso si ha $\alpha_i^k = 0$ e affinche la direzione d^{ij} sia ammissibile deve valere $d_i \geq 0$, ossia si deve necessariamente avere $y_i = 1$.
Analogamente se $i \in U(\alpha_i^k)$ si ha $\alpha_i^k = c$, e affinche la direzione d^{ij} sia ammissibile deve valere $d_i \leq 0$ si deve necessariamente avere $y_i = -1$

Indice	Variabile α	Direzione ammissibile	Valore etichetta
$i \in L(\bar{\alpha})$	$\bar{\alpha}_i = 0$	$d_i \geq 0$	$y_i = 1$
$i \in U(\bar{\alpha})$	$\bar{\alpha}_i = c$	$d_i \leq 0$	$y_i = -1$
$j \in L(\bar{\alpha})$	$\bar{\alpha}_j = 0$	$d_i \geq 0$	$y_i = -1$
$j \in U(\bar{\alpha})$	$\bar{\alpha}_j = c$	$d_j \leq 0$	$y_i = 1$

Osserivamo inoltre che per $0 < \alpha_i < c$ non si hanno vincoli sull'etichetta y_i, e lo stesso vale per y_j. Consideriamo le seguenti partizioni degli insiemi L:

$$L(\alpha) = L^+(\alpha) \cup L^-(\alpha)$$

$$\begin{cases} L^+(\alpha) = \{i \in L(\alpha) : y_i = 1\} \\ L^-(\alpha) = \{i \in L(\alpha) : y_i = -1\} \end{cases}$$

e dell'insieme U:

$$U(\alpha) = U^+(\alpha) \cup U^-(\alpha)$$

$$\begin{cases} U^+(\alpha) = \{i \in U(\alpha) : y_i = 1\} \\ U^-(\alpha) = \{i \in U(\alpha) : y_i = -1\} \end{cases}$$

Proposizione 21.2 (Ammissibilità della direzione d^{ij})**.** Sia $d^{ij} = (0, ..., 0, \frac{1}{y_i}, 0, ..., 0, -\frac{1}{y_j}, 0, ..., 0)^T$. d^{ij} è ammissibile in $\alpha^k \iff$

- $i \in R(\alpha^k) = L^+(\alpha^k) \cup U^-(\alpha^k) \cup \{i : 0 < \alpha_i^k < c\}$
- $j \in S(\alpha^k) = L^-(\alpha^k) \cup U^+(\alpha^k) \cup \{j : 0 < \alpha_j^k < c\}$

Proof. Supponiamo per assurdo che d^{ij} sia ammissibile e che $j \notin S(\alpha^k)$.Ad esempio possiamo considerare $j \in U^-(\alpha^k)$(si può sempre mostrare un assurdo per ognuno degli insiemi appartenti al complementare di S). Allora si ha

$$d_j = -\frac{1}{y_j} = 1 > 0$$

Ma $\alpha_j^k = c$ poiche $j \in U^-(\alpha^k)$, pertanto deve essere

$$d_j \leq 0$$

Si ha quindi un assurdo. Viceversa supponiamo che $i \in R(\alpha^k)$ e $j \in S(\alpha^k)$, prendendo ad esempio $i \in L^+(\alpha_k)$ e $j \in U^+(\alpha^k)$. In tal caso si ha:

$$y^T d^{ij} = \frac{y_i}{y_i} + y_j \frac{-1}{y_j} = 1 - 1 = 0$$

La prima condizione di ammissibilità della direizone d^{ij} è rispettata. Ora, per $i \in L^+(\alpha_k)$ per avere d^{ij} ammissibile devo avere $d_i \geq 0$. In effetti:

$$d_i = \frac{1}{y_i} = 1 \geq 0$$

Mentre per $j \in U^+(\alpha^k)$ per avere d^{ij} ammissibile devo avere $d_i \leq 0$. In questo caso risulta:

$$d_j = -\frac{1}{y_j} = -1 < 0$$

Pertanto la direzione $d^{ij} = (0, ..., 0, \frac{1}{y_i}, 0, ..., 0, -\frac{1}{y_j}, 0, ..., 0)^T$ è sicuramente ammissibile scegliendo come working set $W = i, j$ con $i \in R(\alpha_k)$ e $j \in S(\alpha^k)$. □

Abbiamo quindi una condizione sulla scelta degli indici del working set che garantisce l'ammissibilità della direzione $d^{ij} = (0, ..., 0, \frac{1}{y_i}, 0, ..., 0, -\frac{1}{y_j}, 0, ..., 0)^T$. Dobbiamo considerare ora una regola che stabilisca in quali casi tale direzione è a tutti gli effetti una direzione di discesa.

Proposizione 21.3 (Come stabilire se d^{ij} è di discesa)**.**
La direzione $d^{ij} = (0, ..., 0, \frac{1}{y_i}, 0, ..., 0, -\frac{1}{y_j}, 0, ..., 0)^T$ è di discesa in $\alpha^k \iff$

$$\frac{1}{y_i}\frac{\partial f(\alpha^k)}{\partial \alpha_i} < \frac{1}{y_j}\frac{\partial f(\alpha^k)}{\partial \alpha_j}$$

Proof. La funzione obiettivo f è convessa, pertanto la direzione d^{ij} è di discesa se e solo se:

$$\nabla f(\alpha^k)^T d^{ij} < 0$$

Si ha:

$$\nabla f(\alpha^k)^T d^{ij} = \sum_h \frac{\partial f(\alpha^k)}{\partial \alpha_h} d_h$$

Notiamo che la derivata parziale della funzione obiettivo in α^k è nulla $\forall h \neq i, j$, da cui si ha:

$$\frac{\partial f(\alpha^k)}{\partial \alpha_i} d_i + \frac{\partial f(\alpha^k)}{\partial \alpha_j} d_j < 0 \implies \frac{1}{y_i}\frac{\partial f(\alpha^k)}{\partial \alpha_i} - \frac{1}{y_j}\frac{\partial f(\alpha^k)}{\partial y_j} < 0$$

□

Abbiamo quindi determinato che scegliendo ad ogni iterazione un working set $W_k = \{i, j\}$ con $i \in R(\alpha^k)$ e $j \in S(\alpha^k)$ tali per cui vale

$$\frac{1}{y_i}\frac{\partial f(\alpha^k)}{\partial \alpha_i} < \frac{1}{y_j}\frac{\partial f(\alpha^k)}{\partial y_j}$$

si ha la garanzia di ottenere dei punti aggiornati α^{k+1} che siano ammissibili e che garantiscano lo stretto decremento della funzione obiettivo. Nell'algoritmo SMO si parte con $\alpha^0 = 0$ poichè solitamente la soluzione ha molti zeri, e si conosce $\nabla f(\alpha^0)$. Osserviamo inoltre che queste condizioni non sono sufficienti a garantire che l'algoritmo

converga globalmente all'ottimo, ma in generale si dovrà imporre una condizione di sufficiente decremento.

Algorithm 9 Sequential Minimal Optimization

Require: $Q, \alpha^0 \in \mathbb{R}^n (= 0), \nabla f(\alpha^0) (= \nabla f(0) = -e)$

$k = 0$

while Criterio di arresto non soddisfatto **do**

Scelgo $W_k = \{i, j\}$ con $i \in R(\alpha^k), j \in S(\alpha^k)$ tali che $\frac{\nabla_i f(\alpha^k)}{y_i} < \frac{\nabla_j f(\alpha^k)}{y_j}$

Calcolo α_i^* e α_j^* risolvendo il sottoprobema in due variabili analiticamente.

Pongo $\alpha_h^{k+1} = \begin{cases} \alpha_i^* & h = i \\ \alpha_j^* & h = j \\ \alpha_h^k & \text{altrimenti} \end{cases}$

Calcolo $\nabla f(\alpha^{k+1})$

$k++$

end while

Notiamo che il calcolo del gradiente $\nabla f(\alpha^{k+1})$ può essere notevolmente semplificato:

$$\nabla f(\alpha^{k+1}) = Q\alpha^{k+1} - e$$

Aggiungendo e sottraendo la quantità $Q\alpha^k$ si ha:

$$= Q\alpha^{k+1} - Q\alpha^k + Q\alpha^k - e$$

Osservando che $\nabla f(\alpha^k) = Q\alpha^k - e$ si ha:

$$= \nabla f(\alpha^k) + Q(\alpha^{k+1} - \alpha^k) = \nabla f(\alpha^k) + Q \begin{pmatrix} \alpha_1^k - \alpha_1^k \\ \vdots \\ \alpha_i^{k+1} - \alpha_i^k \\ \vdots \\ \alpha_j^{k+1} - \alpha_j^k \\ \vdots \\ \alpha_n^k - \alpha_n^k \end{pmatrix}$$

Infine notiamo che nell'algoritmo solo le componenti di indice corrispondente ad uno degli indici del working set vengono aggiornate, pertanto nella differenza $\alpha^{k+1} - \alpha^k$ si avranno tutte le componenti nulle ad eccezione delle componenti di indice i, j. Il

gradiente $\nabla f(\alpha^{k+1})$ può quindi essere calcolato iterativamente con la seguente formula:

$$\nabla f(\alpha^{k+1}) = \nabla f(\alpha^k) + Q_i(\alpha_i^{k+1} - \alpha_i^k) + Q_j(\alpha_j^{k+1} - \alpha_j^k)$$

Dove Q_i e Q_j sono rispettivamente l'i-esima e la j-esima colonna della matrice Q. Possiamo quindi evitare di considerare la matrice Q nella sua interezza per aggiornare il gradiente ad ogni passo, risparmiando quindi risorse temporali e spaziali. Inoltre ponendo come punto iniziale $\alpha^0 = 0$ si ha $\nabla f(\alpha^0) = -e$, pertanto anche nel punto iniziale non è necessario sfruttare Q per calcolare il gradiente.

21.1 Proprietà di convergenza di un algoritmo SMO con regola di selezione del primo ordine

Vogliamo adesso studiare le proprietà di convergenza di un algoritmo SMO. La scelta del working set $W = \{i, j\}$ è cruciale per ottenere proprietà di convergenza per algoritmi SMO.

Proposizione 21.4 (Punto di ottimo globale del duale SVM)**.**
Un punto α^* è punto di minimo globale per il duale SVM $\iff$

$$\max_{i \in R(\alpha^*)} \left\{ -\frac{\nabla_i f(\alpha^*)}{y_i} \right\} \leq \min_{j \in S(\alpha^*)} \left\{ -\frac{\nabla_j f(\alpha^*)}{y_j} \right\}$$

Proof. Il problema è convesso con vincoli lineari $\implies$ le KKT sono condizioni necessarie e sufficienti di ottimalità globale. Consideriamo quindi il problema SVM duale e associamo dei moltiplicatori ai vincoli del problema:

$$\min_{\alpha} \frac{1}{2}\alpha^T Q \alpha - e^T \alpha = f(\alpha)$$

$$0 \leq \alpha \leq c \implies \begin{cases} -\alpha_i \leq 0 \implies \mu^- \in \mathbb{R}^n \\ \alpha_i - c \leq 0 \implies \mu^+ \in \mathbb{R}^n \end{cases}$$

$$\alpha^T y = 0 \implies \lambda \in \mathbb{R}$$

Le condizioni KKT sono date da:

$$\begin{cases} \mu_i^+, \mu_i^- \geq 0 \ \ \forall i \\ \mu_i^+(\alpha_i - c) = 0 \\ \mu_i^- \alpha_i = 0 \\ \nabla_i \mathcal{L}(\alpha, \lambda, \mu^+, \mu^-) = 0 \end{cases}$$

Calcoliamo $\mathcal{L}(\alpha, \lambda, \mu^+, \mu^-)$:

$$\mathcal{L}(\alpha, \lambda, \mu^+, \mu^-) = f(\alpha) + \lambda y^T \alpha + \sum_{i=1}^{n} \mu_i^-(-\alpha_i) + \sum_{i=1}^{n} \mu_i^+(\alpha_i - c)$$

Calcoliamo $\nabla_i \mathcal{L}$:

$$\nabla_i \mathcal{L} = \nabla_i f(\alpha) + \lambda y_i - \mu_i^- + \mu_i^+ = 0 \ \ \forall i$$

Consideriamo il valore di $\nabla_i f(\alpha) + \lambda y_i$, per le condizioni di complementarietà si ha:

$$\nabla_i f(\alpha) + \lambda y_i = \begin{cases} = 0 & \alpha_i \in (0, c) \\ = -\mu_i^+ \leq 0 & \alpha_i = c \\ = \mu_i^- \geq 0 & \alpha_i = 0 \end{cases}$$

Dividendo per l'etichetta y_i:

$$\frac{\nabla_i f(\alpha) + \lambda y_i}{y_i} = \begin{cases} = 0 & \alpha_i \in (0, c) \\ \leq 0 & i \in U^+(\alpha) \cup L^-(\alpha) \\ \geq 0 & i \in U^-(\alpha) \cup L^+(\alpha) \end{cases}$$

Per $\alpha_i \in (0, c)$ il numeratore si annulla, per cui l'espressione è 0. I seguenti casi sono determinati considerando che se $\alpha_i = c$ il numeratore è negativo, pertanto dividendo per $y_i : i \in U^+(\alpha)$ si ha, ricordando che $U^+(\alpha)$ è il sottoinsieme di $U(\alpha)$ degli indici associati ad etichette positive, complessivamente la frazione rimane negativa. Se invece $\alpha_i = 0$ il numeratore è positivo, ma dividendo per $y_i : i \in L^-(\alpha)$ stiamo dividendo per $y_i = -1$, rendendo l'espressione complessivamente negativa. Analogamente il numeratore è positivo quando $\alpha_i = 0$, quindi dividendo per $y_i : i \in L^+(\alpha)$, ossia per $y_i = 1$, l'espressione rimane complessivamente positiva. Infine, se risulta $\alpha_i = c$ il numeratore è negativo, e dividendo per $y_i : i \in U^-(\alpha)$, ossia per $y_i = -1$ stiamo rendendo positiva la frazione. Possiamo quindi ricavare che:

$$\frac{\nabla_i f(\alpha)}{y_i} = \begin{cases} = -\lambda & \alpha_i \in (0, c) \\ \leq -\lambda & i \in U^+(\alpha) \cup L^-(\alpha) \\ \geq -\lambda & i \in U^-(\alpha) \cup L^+(\alpha) \end{cases}$$

In conclusione si ha che per $i \in [U^+(\alpha) \cup L^-(\alpha) \cup \{i : 0 < \alpha_i < c\}] = S(\alpha)$ vale:

$$\frac{\nabla_i f(\alpha)}{y_i} \leq -\lambda$$

mentre per $i \in [U^-(\alpha) \cup L^+(\alpha) \cup \{i : 0 < \alpha_i < c\}] = R(\alpha)$ vale:

$$\frac{\nabla_i f(\alpha)}{y_i} \geq -\lambda$$

Abbiamo quindi:

$$\frac{\nabla_i f(\alpha)}{y_i} \geq -\lambda \geq \frac{\nabla_j f(\alpha)}{y_j} \quad \forall i \in R(\alpha), \forall j \in S(\alpha)$$

Invertendo i segni si ha:

$$-\frac{\nabla_i f(\alpha)}{y_i} \leq \lambda \leq \frac{\nabla_j f(\alpha)}{y_j} \quad \forall i \in R(\alpha), \forall j \in S(\alpha)$$

Da cui si ha la tesi:

$$\max_{i \in R(\alpha^*)} \left\{ -\frac{\nabla_i f(\alpha^*)}{y_i} \right\} \leq \min_{j \in S(\alpha^*)} \left\{ -\frac{\nabla_j f(\alpha^*)}{y_j} \right\}$$

□

Proposizione 21.5 (Corollario)**.** Sia α non ottimale. Allora:

$$\max_{h \in R(\alpha^*)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\} > \min_{h \in S(\alpha^*)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\}$$

Definizione 21.6 (Most violating pair(MVP))**.** Una coppia di indici (i^*, j^*) si dice **Most violating pair** se

$$i^* \in \arg\max_{h \in R(\alpha)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\} \quad j^* \in \arg\min_{h \in S(\alpha)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\}$$

Il corollario implica l'esistenza di una coppia di indici $(\bar{i}, \bar{j})$ con $\bar{i} \in R(\alpha)$ e $\bar{j} \in S(\alpha)$:

$$-\frac{\nabla_{\bar{i}} f(\alpha)}{y_{\bar{i}}} > -\frac{\nabla_{\bar{j}} f(\alpha)}{y_{\bar{j}}} \implies \frac{\nabla_{\bar{i}} f(\alpha)}{y_{\bar{i}}} < \frac{\nabla_{\bar{j}} f(\alpha)}{y_{\bar{j}}}$$

La condizione sopra riportata corrisponde esattamente alla condizione imposta sulla scelta di $i \in R(\alpha)$ e $j \in S(\alpha)$ che garantiva lo stretto decremento nell'algoritmo SMO. Per ogni possibile coppia di variabili è presente un gap tra le quantità $-\frac{\nabla_{\bar{i}} f(\alpha)}{y_{\bar{i}}}$ e $-\frac{\nabla_{\bar{j}} f(\alpha)}{y_{\bar{j}}}$ se α non è ottimo.
La Most violating pair produce il gap, e quindi la violazione delle condizioni KKT, più grande possibile tra tutte le altre coppie di indici $(i, j) : i \in R(\alpha), j \in S(\alpha)$. Scegliendo ad ogni iterazione nell'algoritmo SMO la most violating pair come working set si ottiene l'algoritmo SVMlight, algoritmo che garantisce proprietà di convergenza

ed è ampiamente impiegato nelle librerie per l'addestramento di SVM non lineari. Si può dimostrare la seguente proposizione:

Proposizione 21.7 (Convergenza algoritmo SVMlight)**.** Sia $\{\alpha^k\}$ la sequenza prodotta dall'algoritmo SVMlight. Tale sequenza ammette punti di accumulazione, ognuno dei quali è soluzione del duale SVM:

$$\lim_{k\to\infty} \alpha^k = \alpha^*, \quad \alpha^* \text{minimo globale}$$

21.2 Criterio di arresto

L'ultimo aspetto che resta da definire è il criterio di arresto. Siano

$$M(\alpha) = \max_{h\in R(\alpha)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\}$$

$$m(\alpha) = \min_{h\in S(\alpha)} \left\{ -\frac{\nabla_h f(\alpha^*)}{y_h} \right\}$$

Per α^* ottimale risulta:

$$M(\alpha^*) \leq m(\alpha^*)$$

Possiamo definire il seguente criterio di arresto:

$$M(\alpha^*) \leq m(\alpha^*) + \epsilon$$

Non è ovvio che tale criterio di arresto si attivi, poichè non si ha la continuità delle funzioni $m(\alpha^*)$ e $M(\alpha^*)$. Tuttavia vale:

Proposizione 21.8 (Criterio di arresto SVMlight)**.** $\forall \epsilon > 0$ l'algoritmo SVMlight produce una soluzione α^k che soddisfa:

$$M(\alpha^k) - m(\alpha^k) \leq \epsilon$$

22 Algoritmi per SVM lineari, SVM unbiased

Con dataset particolarmente grandi, i classificatori lineari sono spesso la soluzione preferibile, per due ragioni principali: in primo luogo, la dimensionalità dei dati può essere così elevata che la mappatura dei punti in spazi di dimensioni superiori tramite il kernel trick potrebbe non essere necessaria; in secondo luogo, se viene utilizzato il kernel lineare, è possibile sfruttare particolari caratteristiche del problema per rendere il processo di addestramento molto più veloce e consentire quindi di eseguire l'addestramento

con quantità di dati più grandi. Il problema SVM può essere scritto nella forma:

$$\min_w \frac{1}{2}||w||^2 + c\sum_{i=1}^{n} \max\{0, 1 - y^{(i)}w^t x^{(i)}\}$$

In questa formulazione del problema SVM il termine di bias b non viene introdotto esplicitamente nel modello (parliamo di formulazioni **unbiased**); ciò non rappresenta una restrizione dal punto di vista statistico: il bias può essere implicitamente incorporato nel modello aggiungendo una feature costante a tutti gli esempi nel dataset.

$$\begin{pmatrix} x_{11} & \cdots & x_{14} & 1 \\ \vdots & \cdots & \cdots & \vdots \\ x_{41} & \cdots & x_{44} & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_5 \end{pmatrix} \implies w_1x_{11} + w_2x_{12} + w_3x_{13} + w_4x_{14} + 1w_5$$

Essendo il bias interno a w, si regolarizza anche il bias. Il duale di wolfe di un problema SVM unbiased è dato da:

$$\min_\alpha \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$
$$0 \leq \alpha \leq c$$

L'algoritmo Dual Coordinate Descent permette di risolvere il problema duale di una SVM unbiased con kernel lineare. Esso è un metodo di decomposizione che aggiorna una sola variabile alla volta, e in cui l'ordine di scelta delle variabile può essere ciclico o casuale(dimostrato empiricamente che porta a prestazioni migliori). Nel caso di kernel lineare non è richiesto il calcolo della colonna Q_i dell'hessiana, che rappresenta il costo maggiore nell'algoritmo SMO. In generale i costi di iterazioni di SMO e DCD sono simili, ma SMO sceglie le variabili in modo più intelligente, e quindi risulta migliore in tutti i casi di kernel non lineare. Tuttavia se il kernel è lineare e la mole di dati di addestramento è significativa, conviene utilizzare DCD, in quanto si evita il costo di calcolo di due colonne dell'hessiana Q. Un altra strategia di semplificazione adottabile quando il dataset conta un numero diesempi superiore a 10^6 consiste nell'elevare al quadrato la hinge-loss, ottenendo il problema:

$$\min_w \frac{1}{2}||w||^2 + c\sum_{i=1}^{n} \max\{0, 1 - y^{(i)}w^t x^{(i)}\}^2$$

Si può dimostrare che questa espressione del duale SVM con hinge loss elevata al quadrato(e quindi differenziabile) ha la stessa espressione del duale SVM. L'algoritmo DCD è quindi adatto per la classificazione lineare su larga scala. Tuttavia, all'aumentare delle dimensioni dei problemi, questo approccio potrebbe diventare computazional-

mente insostenibile. Per tali casi, è stato proposto un metodo di tipo Newton specializzato per affrontare direttamente il problema primale continuamente differenziabile con hinge loss elevata al quadrato . Questi approcci sono stati implementati nella popolare libreria LIBLINEAR per la classificazione lineare su larga scala.

23 Problemi di somme finite

Un problema di ottimizzazione di somme finite è un problema del tipo:

$$\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} f_i(x) = f(x)$$

con N molto grande. Per questa classe di problemi si utilizzano principalmente **Metodi di tipo gradiente stocastico(SGD)** con **minibatch**. Un metodo di questo tipo è un metodo della forma:

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

Dove l'approssimazione dell'antigradiente della funzione obiettivo nel punto x^k è data da:

$$d_k = \frac{1}{|B|}\sum_{i\in B} -\nabla f_i(x^k)$$

Dove $B \subseteq \{1, ..., N\}, |B| << N$ è detto **minibatch**. Gli algoritmi che usano il gradiente esatto si dicono **batch optimizer**. Il passo α_k da effettuare al passo k lungo la direzione d_k che approssima l'antigradiente della funzione obiettivo nel punto x^k viene comunemente chiamato **learning rate**. Vi sono diverse possibili scelte per il learning rate α_k:

- Costante
- Sequenza decrescente predefinita: ad oggi è la soluzione comunemente impiegata
- Tramite linesearch: Algoritmi di ricerca di linea per il learning rate in un problema di somme finite sono ancora oggetto di ricerca, e rappresentano un importante problema aperto nel campo dell'AI e dell'ottimizzazione

Dobbiamo inoltre osservare che non si ha alcuna garanzia che la direzione $d_k = \frac{1}{|B|}\sum_{i\in B} -\nabla f_i(x^k)$ sia una direzione di discesa per la funzione obiettivo. Consideriamo il caso $|B| = 1$, e supponiamo di scegliere $B_k = \{i_k\}$ in modo uniforme, cosi che la probabilità di scegliere

un indice i come minibatch, P_i, sia uguale ad $\frac{1}{N}$. Calcoliamo il valore atteso della direzione d_k:

$$\mathbb{E}[d_k] = \mathbb{E}[-\nabla f_{i_k}(x^k)] = \sum_{i=1}^{N} P_i(-\nabla f_i(x^k)) = \frac{1}{N}\sum_{i=1}^{N} -\nabla f_i(x^k) = -\nabla f(x^k)$$

Il valore atteso del gradiente su un minibatch **standard**, ossia con $|B| = 1$, è il gradiente della funzione obiettivo $f(x)$. Si hanno quindi due versioni del metodo del gradiente stocastico

- SGD standard($|B| = 1$)
- SGD minibatch($|B| > 1$)

Proposizione 23.1 (Convergenza SGD standard)**.** Sia $||\nabla f_i(x)|| \leq G \ \ \forall i, \forall x \in \mathbb{R}^n$. Sia ∇f lipshitz continuo. Assumiamo poi che la sequenza dei learning rate $\{\alpha_k\}$ sia tale che:

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \ \ \sum_{t=0}^{\infty} \alpha_t^2 < \infty$$

e che ad ogni iterazione l'algortimo SGD standard restituisca in output una soluzione $z^{k+1} = x^\tau$ per qualche $\tau = 0, ..., k$ dove:

$$\mathbb{P}(\tau = t) = \frac{\alpha_t}{\sum_{i=0}^{k} \alpha_i}$$

Allora si ha

$$\lim_{k\to\infty} \mathbb{E}[||\nabla f(z^k)||^2] = 0$$

Un esempio di sequenza decrescente di learning rate che soddisfa le assunzioni della proposizione precedente è data da $\alpha_k = \frac{\alpha_0}{k}$

Algortimo	Caso convesso	Caso fortemente convesso
GD	$\mathcal{O}(\frac{1}{\epsilon})$	$\mathcal{O}(\log \frac{1}{\epsilon})$
SGD	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon})$

A livello asintotico il metodo del gradiente risulta superiore a SGD. Tuttavia per problemi di somme finite i fattori costanti inclusi nell'O-grande sono molto più elevati nel metodo del gradiente. Pertanto per problemi di somme finite SGD risulta superiore.

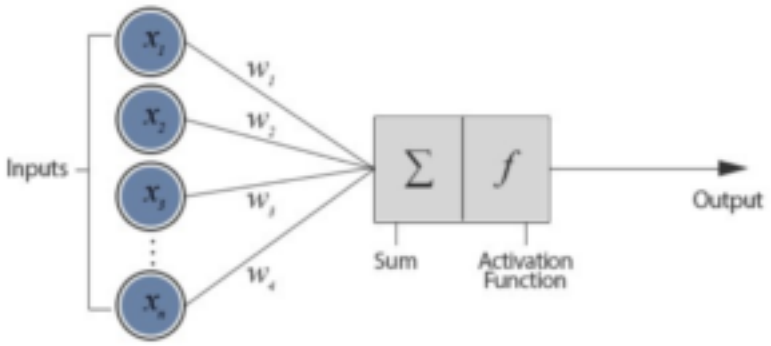


Figure 3: Neurone artificiale

Algorithm 10 Minibatch Stochastic Gradient Descent

Require: $x^0 \in \mathbb{R}^n, f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$

$k = 0$

while Criterio di arresto non soddisfatto **do**

Dividi $\{1, ..., N\}$ in M blocchi disgiunti di dimensione $\frac{N}{M} : B_1, ..., B_M$

for $i = 1, .., M$ **do**

$w_i = w_{i-1} - \alpha_k \frac{1}{|B_i|} \sum_{j \in B_i} \nabla f_j(w_{i-1})$

end for

$x^{k+1} = W_M$

$k++$

end while

Un iterazione su tutto il dataset, che in questo caso corrisponde ad un iterazione del ciclo esterno, prende il nome di **epoca**.

24 Reti neurali

Un **Neurone artificiale** è una funzione matematica vista come modellazione del neurone biologico, e rappresenta l'unità minima nelle reti neurali artificiali. Ad ogni input $(x_1, ..., x_n)$ associa una serie di pesi $(w_1, ..., w_n)$ che vengono passati ad una funzione di somma, restituendo $w^T x + b$. Successivamente tale risultato viene passato ad una funzione di attivazione σ: esempi di quest'ultima sono:

- Sigmoide: $\sigma(x) = \frac{1}{1+e^{-x}}$
- Tangente iperbolica
- ReLU : $\max\{0, x\}$

In generale quindi possiamo vedere un neurone artificiale come una funzione:

$$N(x) = \sigma(w^T x + b)$$

Componendo e collegando tra loro più **layer** di neuroni artificiali si possono costruire **reti neurali**. Addestrare un modello di rete neurale significa risolvere il seguente problema di ottimizzazione:

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(w) = \sum_{i=1}^{N} l_i(w)$$

Questo problema ha le seguenti caratteristiche:

- Funzione obiettivo non lineare
- Funzione obiettivo non convessa
- E' un problema di somme finite con N molto(ma molto) grande
- n è molto grande
- Calcolare la funzione obiettivo ha un costo significativo
- Calcolare il gradiente costa 2 volte tanto quanto il costo del calcolo della funzione obiettivo

Valutiamo adesso i pro e i contro dell'utilizzare gli algoritmi GD e SGD per l'addestramento di modelli di reti neurali. Pro SGD:

- Dati tipicamente ridondanti, non è necessario tutto il dataset per avere una direzione d_k di discesa
- Sperimentalmente funziona bene
- La costante N non appare nella valutazione della complessità

Pro GD:

- Complessità migliore $\mathcal{O}(\log \frac{1}{\epsilon})$
- Possibilità di utilizzare solver molto sofisticati
- Il calcolo di $\mathcal{L}$ è molto parallelizzabile.

Il grafico in figura 4 mostra come per precisioni più basse le prestazioni di SGD superino di gran lunga quelle di GD. In generale i valori di precisione utili per i problemi di apprendimento si assestano intorno a $(10^{-1}, 10^{-2})$, pertanto non importa risolvere il

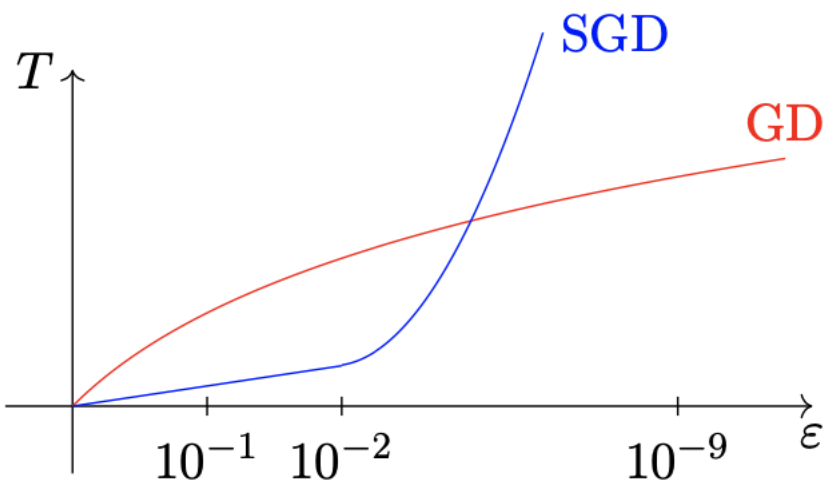


Figure 4: GD vs SGD

problema con alta precisione: utilizziamo comunque SGD anche se a livello asintotico è peggiore poichè permette di raggiungere valori utili impiegando un costo minore. La soluzione che viene tipicamete impiegata è SGD minibatch con $1 < |B| << N$. Un' ulteriore proprietà utile di SGD nell'addestramento di modelli di reti neurali è il suo effetto implicito di regolarizzazione, che implica una maggior difficoltà nel cadere in minimi "sharp", e una maggior facilità nel cadere in minimi più ampi. Vediamo adesso dei possibili miglioramenti a SGD per il caso particolare delle reti neurali:

- Direzione: Si utilizzano termini di tipo momentum/accelerazione:

$$w^{k+1} = w^k - \alpha_k \nabla f_{i_k}(w^k) + \beta(w^k - w^{k-1})$$

  Questa startegia aiuta a smorzare comportamenti irregolari dovuti alla randomizzazione nella scelta della direzione

- Learning rate adattivo:

$$w_i^{k+1} = w_i^k - \alpha_{i_k} \frac{\partial \mathcal{L}(w)}{\partial w_i}$$

  Si utilizza un learning rate differente per ogni peso. Si tenta euristicamente di stimare la dinamica dei gradienti (gradienti alti, pesi bassi, gradienti bassi, pesi alti). Il learning rate α_i^k viene aggiornato ad ogni iterazione secondo una qualche regola, tra le più famose troviamo:

  - RMSProp
  - AdaGrad

– AdaDelta

– Adam

25 Calcolo di $\nabla\mathcal{L}(w)$

Uno dei problemi chiave che sorgono nell'addestramento delle reti neurali riguarda il calcolo dei gradienti della funzione di perdita. Infatti, l'obiettivo del problema di ottimizzazione è la somma finita di funzioni (ignorando il termine di regolarizzazione) di (tipicamente) milioni di variabili, ognuna delle quali è la composizione concatenata di funzioni elementari. Pertanto, le tecniche di differenziazione numerica sono fuori discussione: il costo è troppo elevato (è richiesto un gran numero di valutazioni di funzioni per ottenere un'approssimazione del gradiente) e soffrono anche di significativi errori di approssimazione; quindi, l'approssimazione delle differenze finite è generalmente utilizzata solo per verificare la correttezza dell'implementazione dei gradienti, a bassa precisione. Mostriamo adesso come il metodo delle differenze finite sia fondamentalmente inapplicabile nel contesto dell'addestramento di modelli di reti neurali. Consideriamo la seguente approssimazione:

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}$$

Nel contesto dell'aritmetica finita possiamo derivare il seguente bound sull'errore di approssimazione:

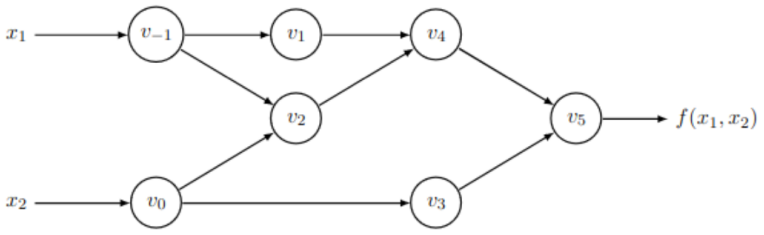
$$\left| \frac{\partial f(x)}{\partial x_i} - \frac{f(x + \epsilon e_i) - f(x)}{\epsilon} \right| \leq cf\epsilon + \frac{c(f, \mu)}{\epsilon}$$

Si ha un bound sull'errore nell'approssimare le derivate con le differenze finite che ha una componente proporzionale alla precisione e l'altra inversamente proporzionale alla precisione e dovuta all'aritmetica finita nei calcolatori. D'altra parte, derivare espressioni esplicite per i gradienti è troppo complesso; anche l'uso di strumenti di differenziazione simbolica, cioè software per manipolare le espressioni al fine di ottenere espressioni delle derivate, porta a formule molto lunghe e complesse e di conseguenza a calcoli massicci. Si ricorre quindi alle tecniche di **Differenziazione automatica**. Tali tecniche mirano a sfruttare in modo intelligente la chain rule:

$$f = f(g(x)) \implies \frac{\partial f}{\partial x_j} = \sum_i \frac{\partial f}{\partial g_i} \frac{\partial g_i}{\partial x_j}$$

Le reti neurali costituiscono fondamentalmente una concatenazione e composizione di

un gran numero di funzioni elementari, pertanto si cerca di decomporre le derivate in termini dati da funzioni elementari. Nelle reti neurali si utilizza la tecnica di **Back-propagation**, che consiste in una tecnica di differenziazione automatica applicata in "reverse mode". In figura 5 viene mostrato un esempio di algoritmo di backpropagation per il calcolo del gradiente della funzione $f(x_1, x_2) = \log(x_1) + x_1x_2 - \sin(x_2)$. Innanzitutto, tutte le operazioni elementari vengono mappate in un grafo di computazione, che è una struttura che consente di collegare le quantità che dipendono l'una dall'altra. In questo modo, le quantità calcolate che saranno necessarie per calcolare altri termini possono essere memorizzate, evitando calcoli duplicati. La computazione effettiva procede prima dando gli input alla funzione e calcolando i prodotti intermedi fino all'output della funzione; successivamente, il grafo viene attraversato all'indietro per calcolare tutte le derivate parziali. Va notato che, al fine di calcolare i gradienti, otteniamo come prodotto laterale intermedio il valore della funzione; quindi, in un'iterazione di discesa del gradiente, possiamo ottenere il valore della funzione nel punto corrente, eseguendo un passaggio in avanti attraverso il grafo di calcolo, e successivamente i gradienti, dopo un passaggio all'indietro attraverso il grafo. Come già detto, alcuni termini compaiono più volte durante il calcolo; per evitare calcoli duplicati, possiamo memorizzare questi valori per riutilizzarli quando necessario; naturalmente, non sarà difficile immaginare che in reti profonde e complesse i termini si ripetano in modo molto più massiccio rispetto all'esempio semplice in questione.



(a) Computation Graph

Forward Primal Trace			Reverse Adjoint (Derivative) Trace		
$v_{-1} = x_1$		$= 2$	$\bar{\boldsymbol{x}}_1 = \bar{\boldsymbol{v}}_{-1}$		$= \mathbf{5.5}$
$v_0 = x_2$		$= 5$	$\bar{\boldsymbol{x}}_2 = \bar{\boldsymbol{v}}_0$		$= \mathbf{1.716}$
$v_1 = \ln v_{-1}$		$= \ln 2$	$\bar{v}_{-1} = \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}}$	$= \bar{v}_{-1} + \bar{v}_1 / v_{-1}$	$= 5.5$
$v_2 = v_{-1} \times v_0$		$= 2 \times 5$	$\bar{v}_0 = \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0}$	$= \bar{v}_0 + \bar{v}_2 \times v_{-1}$	$= 1.716$
			$\bar{v}_{-1} = \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}}$	$= \bar{v}_2 \times v_0$	$= 5$
$v_3 = \sin v_0$		$= \sin 5$	$\bar{v}_0 = \bar{v}_3 \frac{\partial v_3}{\partial v_0}$	$= \bar{v}_3 \times \cos v_0$	$= -0.284$
$v_4 = v_1 + v_2$		$= 0.693 + 10$	$\bar{v}_2 = \bar{v}_4 \frac{\partial v_4}{\partial v_2}$	$= \bar{v}_4 \times 1$	$= 1$
			$\bar{v}_1 = \bar{v}_4 \frac{\partial v_4}{\partial v_1}$	$= \bar{v}_4 \times 1$	$= 1$
$v_5 = v_4 - v_3$		$= 10.693 + 0.959$	$\bar{v}_3 = \bar{v}_5 \frac{\partial v_5}{\partial v_3}$	$= \bar{v}_5 \times (-1)$	$= -1$
			$\bar{v}_4 = \bar{v}_5 \frac{\partial v_5}{\partial v_4}$	$= \bar{v}_5 \times 1$	$= 1$
$y = v_5$		$= 11.652$	$\bar{v}_5 = \bar{y}$	$= 1$	

Figure 5: Backpropagation

26 Project Work

- **Descrizione dell'argomento del project work:**
  E' ben noto che per garantire che il metodo di Newton abbia proprietà di convergenza globale è necessaria l'introduzione di una procedura di ricerca di linea per selezionare un passo adeguato lungo la direzione di discesa definita dal metodo di Newton. Poichè il calcolo dell'hessiana domina il costo computazionale di ogni iterazione, si può pensare di effettuare una ricerca di linea (quasi) esatta lungo la direzione definita dal metodo di Newton invece di effettuare una procedura di backtracking inesatta come la ricerca di Armijo.

- **Obiettivi del progetto:**

  – Implementare(in linguaggio Python, sfruttando la libreria numpy) classi e funzioni che permettano di caricare dati per costruire un istanza di un problema di regressione logistica con regolarizzatore $\updownarrow_2$, calcolare la loss e il gradiente della loss.

- Implementare i seguenti algoritmi
  - Metodo del gradiente con ricerca di Armijo
  - Metodo del gradiente con ricerca esatta(Greedy) [1]
  - Metodo di Newton standard[$\alpha = 1$]
  - Metodo di Newton con ricerca di Armijo
  - Metodo di Newton con ricerca esatta [1]
  - Metodo di Newton ibrido [1]
- **Esperimenti**:
  Effettuare gli esperimenti riportati in [1](Sezione 3, figura 1) con gli algoritmi sopra riportati. Devono inoltre essere forniti risultati che mostrano come l'errore evolve rispetto al tempo.

27 Regressione logistica

In questa sezione consideriamo il problema dell'adattamento di un modello di **regressione logistica**. Il problema di ottimizzazione relativo all'addestramento di modelli di regressione logistica è un problema della forma:

$$\min_{w \in \mathbb{R}^m} \mathcal{L}(w) + \lambda \Omega(w)$$

Dove $\mathcal{L}(w)$ è la funzione di verosimiglianza negativa del modello logistico (nota come **NLL**), e $\Omega(w)$ è un regolarizzatore convesso. Si può mostrare che la **NLL** del modello logistico è anch'essa una funzione convessa, e che essa può essere espressa in forma compatta(assumendo $Y = \{-1, 1\}$) come [2]:

$$\mathcal{L}(w; X, y) = \sum_{i=i}^{n} \log(1 + \exp(-y^{(i)} w^T x^{(i)}))$$

27.1 Calcolo del gradiente e dell'hessiana

Se poniamo $z = Xw$, ossia $z_i = w^T x^{(i)} \ \ \forall i \in \{1, ..., n\}$ si ha:

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

Siamo interessati a calcolare $\nabla_w \mathcal{L}(w; X, y)$. Per la regola della catena si ha:

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

Consideriamo adesso l'i-esima componente del gradiente $\nabla_z \phi(z,y)^T$:

$$\frac{\partial \phi(z,y)}{\partial z_i} = \frac{\partial \log(1+\exp(-y^{(i)} z_i))}{\partial z_i} = \frac{1}{1+\exp(-y^{(i)} z_i)} \exp(-y^{(i)} z_i)(-y^{(i)})$$

$$= -y^{(i)} \frac{1}{1+\exp(-y^{(i)} z_i)} \exp(-y^{(i)} z_i) = -y^{(i)} \frac{1}{1+\exp(y^{(i)} z_i)}$$

Ricordando l'espressione della funzione sigmoide $\sigma : \mathbb{R} \to [0,1]$:

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$

Possiamo esprimere la componente i-esima del gradiente della funzione $\phi(z,y)$ come:

$$\frac{\partial \phi(z,y)}{\partial z_i} = -y^{(i)} \sigma(-y^{(i)} z_i)$$

Pertanto si ha:

$$\nabla_z \phi(z,y) = (-y^{(1)} \sigma(-y^{(1)} z_1), ..., -y^{(n)} \sigma(-y^{(n)}) z_n)^T$$

D'altra parte si ha:

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

Da cui si ha che il gradiente della loss $\mathcal{L}(w)$ è esprimibile come:

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

Dove $r \in \mathbb{R}^n$ è tale che $r_i = -y^{(i)} \sigma(-y^{(i)} w^T x^{(i)}) \ \ \forall i = 1, ..., n$. (Nota: Se si impiega un regolarizzatore L2 del tipo $\Omega(w) = \lambda ||w||^2$, si deve aggiungere al gradiente il termine $2\lambda w$.) Con un procedimento analogo è possibile ricavare l'hessiana:

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

dove $D \in \mathbb{R}^{n \times n}$ è una matrice diagonale tale per cui $d_{ii} = \sigma(y^{(i)} w^T x^{(i)}) \sigma(-y^{(i)} w^T x^{(i)})$. (Nota: se si impiega un regolarizzatore quadratico si deve aggiungere all'hessiana il termine $2\lambda I$.

28 Greedy Newton e Hybrid Newton

Per minimizzare una funzione due volte continuamente differenziabile $f : \mathbb{R}^n \to \mathbb{R}$ il metodo di Newton standard genera iterate del tipo:

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

Nell'ottimizzazione il metodo di Newton è un potente strumento per minimizzare funzioni obiettivo non lineari. Ciò è principalmente dovuto alle sue propretà di convergenza superlineare nell'intorno di un punto di minimo locale 9.1. Tuttavia il metodo di Newton presenta una serie di debolezze note. Ad esempio, la convergenza globale del metodo non è garantita in generale, neanche per funzioni obiettivo strettamente decrescenti. Una delle tecniche principali per ovviare alla non convergenza del metodo è quella di rinunciare al passo costante $\alpha_k = 1$ e considerare:

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

Dove tipicamente il valore di α_k è determinato mediante una ricerca di linea di tipo Armijo con passo iniziale $\alpha_0 = 1$. Se si assume che il metodo raggiunga un iterata che si trovi in un intorno sufficientemente piccolo del punto di minimo, l'introduzione della ricerca di linea di tipo Armijo mantiene la convergenza superlineare del metodo di Newton standard. Nonostante il passo $\alpha_k = 1$ diventi asintoticamente ottimale in un intorno di un punto di minimo locale, tale passo potrebbe comunque non essere ottimo in un intorno del punto di un punto di minimo locale. Inoltre, utilizzare il passo $\alpha_k = 1$ o un passo minore ottenuto tramite ricerca di linea Armijo quando si è lontani da un punto di minimo locale può portare l'algoritmo a convergere lentamente. Per questi motivi in [1] viene proposto il metodo **Greedy Newton**:

$$\alpha_k \in \arg\min_{\alpha} f(x^k - \alpha \nabla^2 f(x^k)^{-1} \nabla f(x^k))$$

Il metodo Greedy Newton viene ricavato considerando un iterazione standard del metodo di Newton, nella quale il passo α_k non viene posto costante uguale ad 1 ma si effettua una ricerca di linea esatta lungo la direzione di discesa data dal metodo di Newton. Si ha che [1]:

- Tipicamente il metodo Greedy Newton non incrementa il costo del Metodo di Newton standard. Questo perchè tipicamente in un iterazione di Newton il costo computazionale è dominato dal calcolo dell'hessiana, pertanto si può effettuare una ricerca di linea esatta, e quindi valutare la funzione obiettivo $f(x^k - \alpha \nabla^2 f(x^k)^{-1} \nabla f(x^k))$ e le sue derivate(anche più volte) non incrementa significa-

tivamente il costo della singola iterazione. Ad esempio, in un problema di regressione logistica con n esempi, feature dense $x \in \mathbb{R}^p$ e etichette binarie $y \in \{-1, 1\}$:

$$f(w) = \sum_{i=1}^{n} \log(1 + \exp(-y^{(i)} w^T x^{(i)}))$$

il costo per calcolare l'hessiana è $\mathcal{O}(pn^2)$, ma il costo per valutare la funzione o una sua derivata direzionale è $\mathcal{O}(pn)$. Mediante bisezione su dominio limitato è possibile risolvere il problema di ricerca di linea esatta nel metodo greedy Newton con accuratezza ϵ in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ iterazioni , per cui il costo complessivo risulta essere $\mathcal{O}(pn + n\log(\frac{1}{\epsilon}))$. Osserviamo che il costo della bisezione può essere ulteriormente ridotto se si considera la struttura di composizione lineare del problema di regressione logistica [1]. Inoltre per risolvere il problema di ricerca di linea esatta è possibile sfruttare solver unidimensionali più efficienti, come il metodo delle secanti [1]

- La ricerca di linea esatta può portare a valori significativamente più piccoli dell'obiettivo rispetto alla ricerca di Armijo. Questo perchè in molti problemi il valore del passo ottimo può risultare molto maggiore del valore massimo di 1 considerato nelle implementazioni standard del metodo di Newton. Inoltre, anche se si considerasse una ricerca di Armijo che parte da 1 ed incrementa il passo di un fattore costante ad ogni iterazione, la condizione di Armijo potrebbe naturalmente escludere il valore ottimo del passo per il problema in una data iterazione. Si può mostrare che per funzioni non quadratiche il massimo valore del passo ammesso dalla condizione di Armijo può essere arbitrariamente peggiore del valore del passo ottimo.

28.1 Metodi Hybrid Newton

Nei metodi di Newton ibridi si combina il metodo del Gradiente al metodo di Newton, con lo scopo di creare metodi che abbiano la proprietà di velocizzare il tasso di convergenza lineare globale, mantenendo le proprietà di convergenza locale superlineare date dal metodo di Newton. [1]. Uno dei metodi di Newton ibridi più semplici possibili è il seguente:

- Sia x_N^{k+1} il passo di Newton puro e sia x_G^{k+1} il passo del metodo del gradiente con ricerca di linea esatta per il valore del passo:

$$x_G^{k+1} = x^k - \alpha_k^G \nabla f(x^k), \quad \alpha_k^G \in \arg\min_{\alpha} \{f(x^k - \alpha \nabla f(x^k))$$

- Se $f(x_G^{k+1}) < f(x_N^{k+1})$, prendi il passo del metodo del gradiente con ricerca di linea

esatta, altrimenti prendi il passo del metodo di Newton puro.

Negli esperimenti effettuati in [1], il metodo Hybrid Newton è risultato peggiore rispetto al metodo Greedy Newton.

28.2 Rergressione logistica: ricerca di linea esatta

Effettuare una ricerca di linea esatta per determinare il passo ottimo ad una data iterazione di un algoritmo di ottimizzazione per la soluzione del problema di regressione logistica equivale a risolvere il seguente problema:

$$\arg\min_{\alpha} \mathcal{L}(w + \alpha d_k) = \sum_{i=1}^{N} \log(1 + \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)}))$$

Per risolvere tale problema può essere utile conoscere la derivata rispetto al passo α della funzione $\mathcal{L}(w + \alpha d_k)$:

$$\frac{\partial \mathcal{L}(w + \alpha d_k)}{\partial \alpha} = \sum_{i=1}^{N} \frac{\partial}{\partial \alpha} \log(1 + \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)}))$$

$$= \sum_{i=1}^{N} \frac{1}{1 + \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)})} \frac{\partial}{\partial \alpha}(1 + \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)}))$$

$$= \sum_{i=1}^{N} \frac{1}{1 + \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)})} \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)}) \frac{\partial}{\partial \alpha}(-y^{(i)}(w + \alpha d_k)^T x^{(i)})$$

$$= \sum_{i=1}^{N} \frac{\exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)})}{1 + \exp(-y^{(i)}(w + \alpha d_k)^T x^{(i)})}(-y^{(i)} d_k^T x^{(i)})$$

Ora ricordando che $\frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} = \sigma(x)$ possiamo scrivere:

$$\frac{\partial \mathcal{L}(w + \alpha d_k)}{\partial \alpha} = \sum_{i=1}^{N} -y^{(i)} \sigma(-y^{(i)}(w + \alpha d_k)^T x^{(i)}) d_k^T x^{(i)}$$

References

[1] Betty Shea and Mark Schmidt. Greedy newton: Newton's method with exact line search. 2024.

[2] Hung Nghiep Tran and Atsuhiro Takasu. Analyzing knowledge graph embedding methods from a multi-embedding interaction perspective. 2023.