UNIVERSITÀ DEGLI STUDI DI ROMA TOR VERGATA
MACROAREA DI SCIENZE MATEMATICHE, FISICHE E NATURALI

LAUREA IN MATEMATICA

TITOLO

METODO DEL SUB-DIFFERENZIALE E APPLICAZIONI

Relatore:
Dott. Cristian Mendico

Laureanda:
Michela Rossi

Matricola:
0292560

Anno Accademico 2023/2024

A chi non respira più con me,
ma continua a vivere al mio fianco,
i miei nonni.

Ai miei genitori,
coloro che mi hanno dato le ali,
insegnato a volare
e lasciato che volassi da sola.

In particolare a mia madre,
grazie per aver combattuto con me.

Al professor Gaetano Corvo,
grande matematico e talentuoso sassofonista,
la cui gentilezza ha lasciato
un'impronta indelebile nella mia vita.
Ovunque lei sia,
so che mi ha seguito fin dal primo traguardo.

Indice

Introduzione

Sempre più diffuso è l'uso del Machine Learning, un "settore" dell'intelligenza artificiale, il cui scopo è quello di creare sistemi che apprendono e migliorano le proprie performance in base ai dati che si forniscono.

La presente tesi si propone di condurre un'analisi sulla velocità convergenza del Metodo del Subdifferenziale e di esplorare le sue applicazioni nel contesto del Machine Learning.

Il Metodo del Subdifferenziale consente di individuare, sotto determinate condizioni, il minimo di una funzione, denominata "funzione obiettivo" o $loss\ function$.

In particolare, si analizza il comportamento del metodo su funzioni non lisce e fortemente convesse.

Per affrontare questa analisi, è necessario richiamare diversi concetti dell'Analisi Matematica, tra cui la teoria della dualità, la convessità, i massimi e minimi in più variabili e le principali proprietà delle funzioni convesse.

È importante osservare che, sebbene il Metodo del Subdifferenziale non sia tra i più efficienti in termini di velocità di convergenza, è rinomato per la sua semplicità che lo rende relativamente agevole da implementare.

Nello specifico, il **Primo Capitolo** si focalizza sull'analisi e la definizione di concetti fondamentali; tra questi, troviamo la definizione di problema primale standard, soluzioni ammissibili e soluzioni di base a esso associate. Inoltre, si fornisce una panoramica approfondita riguardante il concetto di convessità e le funzioni convesse.

Successivamente, si procede con una sezione di richiami sullo studio delle funzioni in più variabili

e, in particolare, sulla ricerca di massimi e minimi.

Nel **Secondo Capitolo**, si introduce il concetto di subdifferenziale, che riveste un ruolo centrale nell'elaborato.

Il **Terzo Capitolo** presenta il Metodo del Subdifferenziale e si presterà particolare attenzione alle sue applicazioni su funzioni obiettivo convesse e non lisce.

Nell'ambito dell'apprendimento automatico si lavora con campioni, cioè dati, che si assume siano stati estratti in modo indipendente e identicamente distribuiti (i.i.d.), vale a dire che la probabilità di ottenere un determinato campione non dipende dai campioni estratti in precedenza e provengono tutti dalla stessa fonte di probabilità con la stessa distribuzione.

Formalmente, si suppone di avere n campioni rappresentati da una coppia di variabili casuali $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, dove $\mathcal{X}$ è lo spazio degli input e $\mathcal{Y}$ è lo spazio degli output.

Il nostro obiettivo è trovare una *predictor function* $f : \mathcal{X} \to \mathbb{R}$, che abbia un *rischio (atteso) ridotto*, cioè un margine di errore (perdita) sufficientemente "piccolo" su dati non osservati.

Il rischio atteso associato alla funzione f è definito come segue:

$$\mathcal{R}(f) := \mathbb{E}[\ell(y, f(x))]$$

dove $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ è detta *loss function.*

Solitamente, la loss function è convessa rispetto alla seconda variabile; questa caratteristica, comune alla gran parte delle loss functions, è un'assunzione importante che semplifica notevolmente lo studio e la risoluzione dei problemi di ottimizzazione del modello.

Si definisce inoltre un *rischio empirico* associato alla funzione f dato da

$$\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

Si intuisce facilmente che il rischio empirico rappresenta una stima dell'errore del modello basata sul dataset di cui si dispone, cioè i dati osservati.

In generale, risolvere un problema di ottimizzazione con alta precisione è computazionalmente costoso.

Ci si potrebbe chiedere quale sia una buona stima per la soluzione. Di seguito cerchiamo una risposta a questa domanda.

Si consideri un insieme di predictor functions $\{f_\theta\}_{\theta\in\mathbb{R}^d}$ parametrizzate dalla variabile θ, si denoti con $\hat{\theta}$ l'output del modello costruito e con θ^* il valore di θ tale che

$$\mathcal{R}(f_{\theta^*}) \in \left\{ \mathcal{R}(f_\theta) \mid \min_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta) \right\}$$

cioè un possibile valore di θ la cui f_θ associata minimizzi il rischio atteso.

La quantità $\mathcal{R}(f_{\theta^*}) - \inf_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta)$ può essere riscritta come segue:

$$\mathcal{R}(f_{\theta^*}) - \inf_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta) = \mathcal{R}(f_{\theta^*}) - \hat{\mathcal{R}}(f_{\hat{\theta}}) + \hat{\mathcal{R}}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\theta^*}) + \hat{\mathcal{R}}(f_{\theta^*}) - \mathcal{R}(f_\theta)$$

dove

- $\mathcal{R}(f_{\theta^*}) - \hat{\mathcal{R}}(f_{\hat{\theta}})$, $\hat{\mathcal{R}}(f_{\theta^*}) - \mathcal{R}(f_\theta)$ sono detti errori di stima
- $\hat{\mathcal{R}}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\theta^*})$ è detto errore di ottimizzazione

La risposta alla domanda iniziale è che è sufficiente raggiungere una precisione dell'ordine dell'errore di stima, che tipicamente è $O(\frac{1}{\sqrt{n}})$ oppure $O(\frac{1}{n})$; si può dimostrare infatti che questa stima è tra le migliori per evitare problemi di overfitting, ossia problemi in cui il modello si adatta "troppo bene" al dataset che ha a disposizione, comportando una scarsa capacità di descrivere i dati non osservati; questo accade perché il modello memorizza i dati di addestramento anziché comprendere come descriverli.

Nel resto del Capitolo vengono presentati tre risultati principali, a cui si arriva gradualmente. Il primo risultato è lo studio del Metodo del Subgradiente su funzioni L-regolari, ossia funzioni che sono differenziabili quasi ovunque senza richiedere ipotesi di differenziabilità e fortemente convesse.

Questo rappresenta il primo risultato importante, ottenuto sotto ipotesi molto forti: l'obiettivo successivo è quello di indebolire tali ipotesi.

Pertanto, il secondo risultato consiste nello studio del metodo su funzioni L-regolari, ma non più fortemente convesse.

Infine, si indebolisce l'ipotesi di L-regolarità, osservando che l'ipotesi di convessità è invece necessaria e non può essere indebolita. In questo caso, si stima il minimo della distanza della funzione considerata dalla soluzione ottima.

Seguono poi alcuni esempi di funzioni di perdita comuni e diffuse in questo ambito, con alcune applicazioni pratiche.

Infine, si analizzano degli esempi di funzioni non lisce che hanno un punto di minimo e si studia la velocità di convergenza del metodo su queste funzioni.

Richiami

1.1 Cenni di Programmazione Lineare

Il contenuto di questo capitolo è stato tratto e rielaborato da [3], ad eccezione della sezione 1.3 che è stata tratta e rielaborata da [4].

1.1.1 Problema primale standard

Definizione 1.1.1. Si definisce **problema primale standard** (P) un problema di minimo della forma

$$\begin{cases} \min\ c \cdot x \\ \text{s.t. } Ax = b, x \geq 0 \end{cases}$$

dove A $\in \mathbb{M}^{m \times n}$, b $\in \mathbb{R}^n$.

Definizione 1.1.2. Si definisce **soluzione ammissibile** per (P) un vettore $x \in \mathbb{R}^n$ che verifichi le condizioni della definizione 1.1.1.

Denotando con $\left\{a^j \,|j = 1, .., n\right\}$ l'insieme delle colonne di A, si definisce **soluzione di base ammissibile** per (P) un vettore $x \in \mathbb{R}^n$ che sia soluzione ammissbile e tale che l'insieme

$$\left\{a^j \mid x_j \neq 0,\ j = 1, .., n\right\}$$

sia costituito da vettori linearmente indipendenti.

Nota 1.1.3. Se x è soluzione di base ammissibile per (P) allora x ha al più m entrate diverse da zero.

Definizione 1.1.4. Si definisce **soluzione ottima** per (P) una soluzione $x \in \mathbb{R}^n$ ammissibile che realizza anche il minimo di $c \cdot x$.
Si definisce **soluzione di base ottima** per (P) una soluzione $x \in \mathbb{R}^n$ ottima e di base.

Teorema 1.1.5. Se esiste una soluzione ammissibile per (P), allora esiste una soluzione di base ammissibile per (P).
Se esiste una soluzione ottima per (P), allora esiste una soluzione di base ottima per (P).

1.2 Convessità

1.2.1 Insiemi convessi

Definizione 1.2.1. Dato un insieme $C \subseteq \mathbb{R}^n$, si dice che C è **convesso** se per ogni coppia di punti $x, y \in C$, si ha che il segmento congiungente x ed y è contenuto in C, ossia

$$\{tx + (1-t)y \mid t \in [0,1]\} \subset C$$

Definizione 1.2.2. Dato $\tilde{A} = \left\{a^1, ..., a^p\right\} \subset \mathbb{R}^n$, si definisce **combinazione convessa** di $a^1, .., a^p$ l'insieme

$$\left\{\sum_{i=1}^{p} t_i a^i \mid t_i \geq 0 \ , \ \sum_{i=1}^{p} t_i = 1\right\}$$

Definizione 1.2.3. Dato $\tilde{A} = \left\{a^1, ..., a^p\right\} \subset \mathbb{R}^n$, si definisce **politopo convesso** $\mathcal{P}$ generato da $\left\{a^1, ..., a^p\right\}$ l'insieme

$$\left\langle a^1, ..., a^p \right\rangle := \left\{\sum_{i=1}^{p} t_i a^i \mid t_i \geq 0, \ \sum_{i=1}^{p} t_i = 1\right\}$$

Teorema 1.2.4 (**di Carathéodory**)**.** Dato $\tilde{A} = \{a^1, ..., a^p\} \subset \mathbb{R}^n$, detto $\mathcal{P}$ il politopo convesso generato da $\tilde{A}$, ogni punto $\mathbf{b} \in \mathcal{P}$ si scrive come combinazione convessa di al più $n+1$ elementi di $\tilde{A}$, ossia

$$b = \sum_{i=1}^{n+1} t_i a_{j_i}$$

dove $1 \leq j_1 < j_2 < ... < j_{n+1} \leq p$, $t_i \geq 0$, $\sum_{i=1}^{n+1} t_i = 1$.

Dimostrazione. Poniamo $A = \begin{bmatrix} a^1 & a^2 & ... & a^p \\ 1 & 1 & ... & 1 \end{bmatrix} \in \mathbb{M}^{(n+1)\times p}$.

Consideriamo il sistema lineare $Ax = \begin{bmatrix} b \\ 1 \end{bmatrix}, x \geq 0$.

Per il Teorema 1.1.5, esiste una soluzione ammissibile $x \in \mathbb{R}^p$ e quindi una soluzione di base ammissibile. Ne segue che x ha al più $n+1$ entrate $x_{i_1}, ...x_{i_{n+1}}$ non nulle, corrispondenti alle colonne di A linearmente indipendenti.

Per ogni $j = 1, ..., n+1$, poniamo $t_j := x_{i_j}$.

Siccome x è soluzione di $Ax = b$, si ottiene la tesi.

□

Definizione 1.2.5. Dato $C \subseteq \mathbb{R}^n$ convesso, si definisce **estremo** di C un elemento $e \in C$ tale che non esistono $x, y \in C$ per cui vale $e = tx + (1-t)y \quad \text{con } t \in (0,1)$.

Equivalentemente se per ogni $x, y \in \mathbb{R}^n$, $e = x$ oppure $e = y$.

Teorema 1.2.6. In un insieme chiuso, limitato e convesso ogni elemento è combinazione convessa degli estremi.

Definizione 1.2.7. Dato $C \subseteq \mathbb{R}^n$ un insieme chiuso, convesso e $0 \in C$, si definisce **duale polare** di C l'insieme

$$C^0 = \{y \in \mathbb{R}^n \mid x \cdot y \leq 1, \ \ x \in \mathrm{C}\}$$

Definizione 1.2.8. Dato a $\in \mathbb{R}^n$, $b \in \mathbb{R}$, un'espressione della forma

$$a \cdot x + b = 0$$

definisce un **iperpiano** in $\mathbb{R}^n$.

Definizione 1.2.9. Siano $\mathbb{H}_1, \mathbb{H}_2$ due sottoinsiemi di $\mathbb{R}^n$.

L'iperpiano $a \cdot x + b = 0$ **separa** $\mathbb{H}_1$ e $\mathbb{H}_2$ se

- $a \cdot x + b \geq 0$ per ogni $x \in \mathbb{H}_1$
- $a \cdot x + b \leq 0$ per ogni $x \in \mathbb{H}_2$

L'iperpiano $a \cdot x + b = 0$ **separa strettamente** $\mathbb{H}_1$ e $\mathbb{H}_2$ se

- $a \cdot x + b > 0$ per ogni $x \in \mathbb{H}_1$
- $a \cdot x + b < 0$ per ogni $x \in \mathbb{H}_2$

Teorema 1.2.10. Sia $C \subseteq \mathbb{R}^n$ un insieme non vuoto, chiuso e convesso.

Supponendo che $y \notin C$, esiste un iperpiano $a \cdot x + b = 0$ che separa strettamente C ed y.

Teorema 1.2.11. Dato $C \subseteq \mathbb{R}^n$ un insieme chiuso, convesso e $0 \in C$ e C^0 il suo duale polare si ha che:

(i) C^0 è chiuso, convesso e $0 \in C^0$

(ii) $(C^0)^0 = C$

Dimostrazione. $0 \in C^0$ perché $x \cdot 0 = 0 < 1$ per ogni $x \in C$.

Mostriamo che C^0 é chiuso.

Fissato x e posto $f(y) = x \cdot y$, si può affermare che $C_x^0 := f^{-1}((-\infty; 1])$ è un chiuso poiché preimmagine di un chiuso mediante funzione continua. Anche C^0 è un chiuso poiché intersezione arbitraria di chiusi, cioè

$$C^0 = \bigcap_{x \in C} C_x^0$$

Resta da verificare la continuità di f.

Fissato $x \in C$, siano $y_1,\ y_2 \in C_x^0$.

$|f(y_2) - f(y_1)| = |x \cdot y_2 - x \cdot y_1| = |x| \cdot |y_2 - y_1|$.

In realtà f è più che continua, infatti é lipschitziana di costante $|x|$.

Mostriamo che C^0 é convesso.

Siano $y_1,\ y_2 \in C^0$ e $t \in [0,1]$.

Mostriamo che la loro combinazione convessa é contenuta in C^0.

$x \cdot (ty_1 + (1-t)y_2) = tx \cdot y_1 + x \cdot y_1 - tx \cdot y_1 \leq t + 1 - t = 1$

(ii) Si osserva che

$$(C^0)^0 = \left\{ z \in \mathbb{R}^n \mid y \cdot z \leq 1,\ \ y \in C^0 \right\}$$

Mostriamo che $C \subseteq (C^0)^0$.

Sia $x \in C$; per ogni $y \in C^0$ si ha che $y \cdot x \leq 1$, cioé $x \in (C^0)^0$. Di conseguenza, $C \subseteq (C^0)^0$.

Mostriamo che $(C^0)^0 \subseteq C$. Per assurdo supponiamo che $(C^0)^0 \setminus C \neq \emptyset$.

Sia $z \in (C^0)^0 \setminus C$. Essendo C chiuso, convesso e non vuoto, per il Teorema 1.2.10, esiste un iperpiano che separa C e z, cioè

$$a \cdot x + b > 0 \qquad \text{per ogni } x \in C : \tag{1.2.1}$$

$$a \cdot z + b < 0. \tag{1.2.2}$$

Per ipotesi, $0 \in C$ e dalla 1.2.1 si ottiene che $b > 0$. Ponendo $y = -\frac{a}{b}$, la 1.2.1 si riscrive come

$y \cdot x < 1$ per ogni $x \in C$, che equivale a $y \in C^0$. Siccome $z \in (C^0)^0$ allora $y \cdot z \leq 1$, cioè $a \cdot z + b \geq 0$, assurdo per la 1.2.2.

Quindi $(C^0)^0 \setminus C = \emptyset$, da cui segue che $(C^0)^0 \subseteq C$.

□

1.2.2 Funzioni Convesse

Definizione 1.2.12. Una funzione F: $\mathbb{R}^n \to \mathbb{R}$ si dice **convessa** se per ogni coppia di punti $x, y \in \mathbb{R}^n$ e per ogni t $\in [0,1]$, vale la seguente disuguaglianza

$$F(tx + (1-t)y) \leq tF(x) + (1-t)F(y).$$

Se la disuguaglianza vale in senso stretto, F si dice fortemente convessa.

Nota 1.2.13. Si denota con $\mathscr{E} = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \mid y \geq F(x),\ x \in \mathbb{R}^n \right\} \subset \mathbb{R}^{n+1}$ l'epigrafico di F.

Osservazione 1.2.14. F è convessa se e soltanto se il suo epigrafico è un insieme convesso.

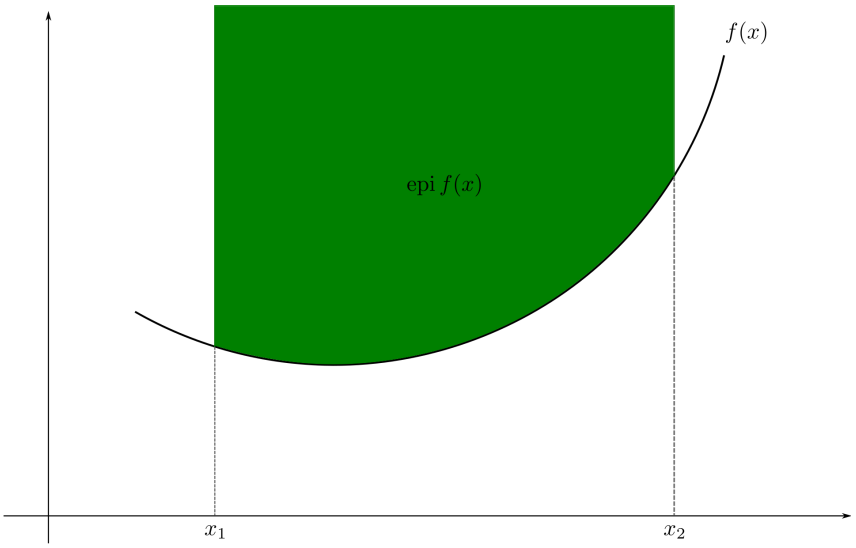

Figura 1.1: Grafico di una funzione convessa

Definizione 1.2.15. Si definisce matrice hessiana di $F(x)$ e si denota con $H(F(x))$ la matrice quadrata delle derivate seconde di F; si scriverà $H(F(x)) \geq 0$ se H è semidefinita positiva.

Teorema 1.2.16 (Richiamo unidimensionale)**.**

(i) Sia f: $\mathbb{R} \to \mathbb{R}$ continua e differenziabile, allora f è convessa se e solo se

$$f(x) + f'(x)(\hat{x} - x) \ \leq f(\hat{x})$$

per ogni x, $\hat{x} \in \mathbb{R}$.

(ii) Se f: $\mathbb{R} \to \mathbb{R}$ è derivabile due volte, allora f è convessa se e solo se $f''(x) \geq 0$ per ogni $x \in \mathbb{R}$.

(iii) Se f: $\mathbb{R} \to \mathbb{R}$ è convessa allora è continua.

Dimostrazione. $(i)(\Rightarrow)$ f è convessa, quindi vale la 1.2.12 che equivale a

$$\frac{f(x + t(\hat{x} - x)) - f(x)}{t} \leq f(\hat{x}) - f(x) \tag{1.2.3}$$

Mandando $t \to 0$, si ottiene

$$f'(x)(\hat{x} - x) \leq f(\hat{x}) - f(x)$$

$(\Leftarrow)$ Si ponga $z = tx + (1-t)\hat{x}$. Per ipotesi

$$f(\hat{x}) \geq f(z) + f'(z)(\hat{x} - z) \tag{1.2.4}$$

$$f(x) \geq f(z) + f'(z)(x - z) \tag{1.2.5}$$

Moltiplicando la 1.2.5 per t e la 1.2.4 per $(1-t)$ si ottiene

$tf(x) + (1-t)f(\hat{x}) \geq tf(z) + tf'(z)(x-z) + (1-t)f(z) + (1-t)f'(z)(\hat{x} - z) =$

$= f(z) + f'(z)[t(x-z) + (1-t)(\hat{x} - z)]$

Tuttavia,

$t(x-z) + (1-t)(\hat{x} - z) = t(x - tx - (1-t)\hat{x}) + (1-t)(\hat{x} - tx - (1-t)\hat{x}) = tx - t^2x - t\hat{x} +$

$t^2\hat{x} + \hat{x} - tx - (1-t)\hat{x} - t\hat{x} + t^2x + t\hat{x} - t^2\hat{x} = 0$

Quindi si ottiene,

$$tf(x) + (1-t)f(\hat{x}) \geq f(tx + (1-t)\hat{x})$$

cioè la tesi.

$(iii, cenni)$

Passo 1. Si mostra che f è limitata su ogni intervallo $[a, b]$ di misura finita. Si scriva ogni elemento $x \in [a, b]$ come combinazione convessa di a e b, cioè $x = ta + (1-t)b$ con $t \in [0.1]$.

Dalla convessità di f, si ottiene che $f(x) \leq 2\max\{|f(a)|, |f(b)|\}$.

Se $x \in [\frac{a+b}{2}, b]$, facendo alcuni calcoli e sfruttando la convessità di f si ottiene che

$$f(x) \geq -2\left(|f\left(\frac{a+b}{2}\right)| + |f(a)|\right)$$

Allo stesso modo, se $x \in [a, \frac{a+b}{2}]$, si ottiene che

$$f(x) \geq -2\left(|f\left(\frac{a+b}{2}\right)| + |f(b)|\right)$$

Quindi, poiché

$$\sup_{x\in[a,b]} |f| \leq 4\left(|f(a)|, |f\left(\frac{a+b}{2}\right)|, |f(b)|\right)$$

f è limitata su $[a, b]$.

Passo 2. Per semplicità supponiamo $[a, b] = [-1, 1]$. Siano $-1 \leq x < y \leq 1$.

Si scriva $y = tx + (1-t)2$, con $t = \frac{2-y}{2-x}$.

Dalla convessità di f si ottiene

$f(y) \leq tf(x) + (1-t)f(2) = tf(x) + f(2) - tf(2) + f(x) - f(x) = (1-t)(f(2) - f(x)) + f(x)$

cioè

$$f(y) - f(x) \leq (1-t)(f(2) - f(x)) \leq \cdots \leq 2|y-x| \sup_{x\in[-2,2]} |f|$$

Con gli stessi procedimenti si ottiene una stima per x e si ottiene che

$$f(x) - f(y) \leq 2|y - x| \sup_{x \in [-2,2]} |f|$$

cioè

$$|f(x) - f(y)| \leq 2|y - x| \sup_{x \in [-2,2]} |f|$$

Siccome $\sup_{x \in [-2,2]} |f|$ è finito, allora f è continua.

Per generalizzare a un generico intervallo $[a, b]$ si trasforma $[-1, 1]$ in $[a, b]$

□

Teorema 1.2.17.

(i) Sia F: $\mathbb{R}^n \to \mathbb{R}$ continua e differenziabile, allora F è convessa se e solo se

$$F(x) + \nabla F(x) \cdot (\hat{x} - x) \ \leq F(\hat{x})$$

per ogni x, $\hat{x} \in \mathbb{R}^n$.

(ii) Se F: $\mathbb{R}^n \to \mathbb{R}$ è derivabile due volte, allora F è convessa se e solo se $H(F(x)) \geq 0$ per ogni $x \in \mathbb{R}^n$.

(iii) Se F: $\mathbb{R}^n \to \mathbb{R}$ è convessa allora F è continua.

Dimostrazione. Fissati x, $y \in \mathbb{R}^d$, si definisca

$$\kappa : \mathbb{R} \longrightarrow \mathbb{R}$$

$$t \mapsto \kappa(t) := F(x + ty) \tag{1.2.6}$$

(i) La dimostrazione consisterà in due passi.

Passo 1. Dimostrare che F convessa $\iff$ κ convessa.

($\Rightarrow$) Siano x, $y \in \mathbb{R}^n$, $s \in [0, 1]$ e t_1, $t_2 \in \mathbb{R}$. Si ponga

$$z_1 = x + t_1 y, \quad z_2 = x + t_2 y$$

Dalla convessità di F si ha che

$$F(sz_1 + (1-s)z_2) \leq sF(z_1) + (1-s)F(z_2) = s\kappa(t_1) + (1-s)\kappa(t_2) \tag{1.2.7}$$

Inoltre,

$$F(sz_1 + (1-s)z_2) = F(s(x + t_1 y) + (1-s)(x + t_2 y)) = F(st_1 y + x + t_2 y - st_2 y) =$$

$$= F(x + y(st_1 + t_2(1-s))) = \kappa(st_1 + t_2(1-s)) \tag{1.2.8}$$

Da 1.2.7 e 1.2.8 si ottiene

$$\kappa(st_1 + t_2(1-s)) \leq s\kappa(t_1) + (1-s)\kappa(t_2)$$

cioè la convessità di κ.

($\Leftarrow$) Sia $s \in [0,1]$, $\tilde{x}$, $\tilde{y} \in \mathbb{R}^n$.

Si ponga in 1.2.6 $x = \tilde{y}$, $y = \tilde{x} - \tilde{y}$ Siccome κ è convessa, si ha che

$$F(s\tilde{x} + (1-s)\tilde{y}) = F(\tilde{y} + s(\tilde{x} - \tilde{y})) = \kappa(s) = \kappa(s \cdot 1 + (1-s) \cdot 0) \leq$$

$$\leq s\kappa(1) + (1-s)\kappa(0) = sF(\tilde{x}) + (1-s)F(\tilde{y})$$

cioè la convessità di F.

Passo 2. Nel caso unidimensionale, si è visto nel Teorema 1.2.16 che per ogni $t_1, t_2 \in \mathbb{R}$

$$\kappa(t_1) + \kappa'(t_1)(t_2 - t_1) \leq \kappa(t_2) \tag{1.2.9}$$

Dalla definizione di κ, si ottiene che $\kappa'(t) = \nabla F(x + ty) \cdot y$.

Quindi, scegliendo $t_1 = 0$ e $t_2 = 1$ in 1.2.9, si ottiene

$$\kappa(0) + \kappa'(0) = F(x) + \nabla F(x) \cdot y \leq \kappa(1) = F(x + y) \tag{1.2.10}$$

Si ponga $\hat{x} = x + y$, la 1.2.10 diventa

$$F(x) + \nabla F(x) \cdot (\hat{x} - x) \leq F(\hat{x}) \tag{1.2.11}$$

da cui la tesi.

(ii) Nel caso unidimensionale, dal punto (ii) del Teorema 1.2.16, se κ è differenziabile due volte allora κ è convessa se e soltanto se $\kappa''(t) \geq 0$ per ogni $t \in \mathbb{R}$.

$$\kappa'(t) = \nabla F(x + ty) \cdot y$$

$$\kappa''(t) = y^T \cdot \nabla^2 F(x + ty) \cdot y \geq 0 \tag{1.2.12}$$

Ponendo $t = 0$ nella 1.2.12, si ha che

$$y^T \cdot \nabla^2 F(x) \cdot y \geq 0$$

da cui la tesi.

$(iii, cenni)$

Si considera $Q_0^{k,n} := [-k, k] \times \cdots \times [-k, k]$, l'ipercubo di dimensione n di centro l'origine e lati di lunghezza $2k$ paralleli agli assi coordinati.

Passo 1. Per induzione si dimostra che $F\big|_{Q_0^{k,n}}$ è limitata.

Il caso $n = 1$ è quello del Teorema 1.2.16.

Per $n > 1$, si considera $Q_{0,i}^{k,n-1} := [-k, k] \times \cdots \times [-k, k] \times \{x_n = i\}$.

$Q_{0,i}^{k,n-1}$ è un ipercubo di dimensione $n - 1$ quindi per ipotesi induttiva, F è limitata su $Q_{0,i}^{k,n-1}$, quindi lo sarà per $i = -k$, $i = 0$ e $i = k$.

Fissato $x' \in Q_0^{k,n-1}$, si definisce la funzione $g(x_n) := f(x', x_n)$ e usando la convessità di F, si ottiene una minorazione per il $\sup_{x \in Q_0^{k,n}} |F|$ con una quantità finita.

Passo 2. Si fissino $x, y \in \mathbb{R}^n$ tali che $0 < |x-y| \leq 1$ e sia $z = \frac{y-x}{|y-x|}$. Si definisca $\kappa(t) := F(x+tz)$; dal Teorema 1.2.16, si ottiene

$$\frac{\kappa(t) - \kappa(0)}{t} \leq 2 \max_{|s| \leq 2} |\kappa(s)|$$

Si conclude scegliendo $t = |y - x|$.

□

Interpretazione geometrica 1. Ricordando che l'espressione $\nabla F(x) \cdot (\hat{x} - x)$ rappresenta l'iperpiano tangente ad F in x, la condizione (i) del Teorema 1.2.17 indica che il grafico delle funzioni convesse giace sopra ognuno dei suoi iperpiani tangenti.

Esempio 1.2.18. In $\mathbb{R}^2$ gli iperpiani sono le rette, ed effettivamente quanto asserito nell' interpretazione geometrica 1 corrisponde con quanto realmente accade. Consideriamo infatti l'insieme

$$\left\{(x, y) \in \mathbb{R}^2 \mid y = e^x\right\}$$

cioè il grafico della funzione $F(x) = e^x$, strettamente convessa.

Nella figura 1.2 sono stati riportate alcune rette tangenti al grafico e si può osservare che il grafico giace sopra ognuna di esse.

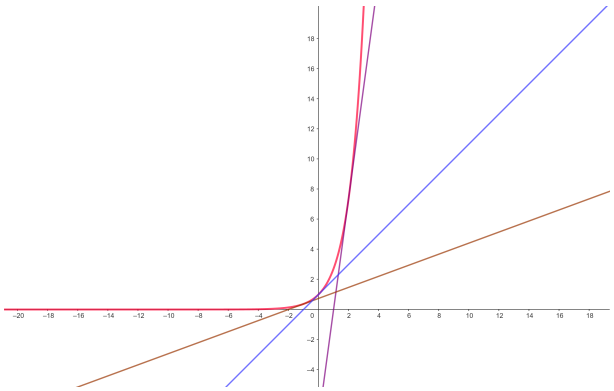

Figura 1.2: alcune rette tangenti al grafico $f(x) = e^x$

1.3 Massimi, minimi, punti critici

Definizione 1.3.1. Sia F: $\mathbb{R}^n \to \mathbb{R}$, $x_0 \in \mathbb{R}^n$ è

- **punto di minimo relativo** per F se esiste $R > 0$ tale che per ogni $x \in B_R(x_0)$ si ha che $F(x_0) \leq F(x)$.
- **punto di massimo relativo** per F se esiste $R > 0$ tale che per ogni $x \in B_R(x_0)$ si ha che $F(x_0) \geq F(x)$.

Se le disuguaglianze valgono in modo stretto, x_0 si dice rispettivamente di minimo/massimo relativo stretto

Definizione 1.3.2. Data F: $\mathbb{R}^n \to \mathbb{R}$ derivabile e $x_0 \in \mathbb{R}^n$ tale che $\nabla F(x_0) = 0$. Allora x_0 è un punto critico per F.
Se x_0 è un punto critico, ma non è né di massimo né di minimo si dice che x_0 è un punto di sella.

Lemma 1.3.3. Una matrice $A = (a_{i,j})_{i,j=1}^n$ è definita positiva se e soltanto se esiste una costante $m > 0$ tale che

$$\lambda^T \cdot A \cdot \lambda \geq m|\lambda|^2 \tag{1.3.1}$$

per ogni $\lambda \in \mathbb{R}^n$.
Una matrice $A = (a_{i,j})_{i,j=1}^n$ è definita negativa se e soltanto se esiste una costante $m < 0$ tale che

$$\lambda^T \cdot A \cdot \lambda \leq m|\lambda|^2$$

per ogni $\lambda \in \mathbb{R}^n$.

Dimostrazione. ($\Rightarrow$) Si consideri la forma quadratica Q associata alla matrice A.

$$Q(\lambda) = \lambda^T \cdot A \cdot \lambda = \sum_{i,j=1}^{n} a_{i,j}\lambda_i\lambda_j$$

e si faccia variare λ nel compatto

$$K = \{\lambda \in \mathbb{R}^n \ : \ |\lambda| = 1\}$$

Essendo Q continua su K compatto, per il Teorema di Weierstrass, esiste $\lambda_0 \in K$ tale che

$$\lambda_0 = \min_{\lambda \in K} Q(\lambda)$$

Risulta quindi che

$$Q(\lambda) = \sum_{i,j=1}^{n} a_{i,j}\lambda_i\lambda_j \geq Q(\lambda_0) \tag{1.3.2}$$

per ogni $\lambda \in K$.

Siccome A è definita positiva e $\lambda_0 \neq 0$ la disuguaglianza in 1.3.2 in realtà è una disuguaglianza stretta.

Ponendo $m = Q(\lambda_0)$, si ottiene $\sum_{i,j=1}^{n} a_{i,j}\lambda_i\lambda_j \geq m$ per ogni $\lambda \in K$.

Se invece $\lambda \in \mathbb{R}^n \setminus \{0\}$, basta considerare $\mu := \frac{\lambda}{|\lambda|}$: in questo modo, $\mu \in K$ e dalla 1.3.2,si ottiene

$$m \leq \sum_{i,j=1}^{n} a_{i,j}\mu_i\mu_j \leq \sum_{i,j=1}^{n} a_{i,j}\frac{\lambda_i}{|\lambda_i|} \cdot \frac{\lambda_j}{|\lambda_j|} = \frac{1}{|\lambda|^2}\sum_{i,j=1}^{n} a_{i,j}\lambda_i\lambda_j$$

cioè la tesi.

Per $\lambda = 0$ la tesi è ovvia.

($\Leftarrow$) Se vale la 1.3.1, essendo $m > 0$ la matrice è chiaramente definita positiva.

□

Lemma 1.3.4. Sia $f : \mathbb{R} \to \mathbb{R}$, $f \in C^2(\mathbb{R})$, $x_0 \in \mathbb{R}$

- se x_0 è un punto di minimo relativo allora

  (i) $f'(x_0) = 0$

  (ii) $f''(x_0) \geq 0$

- x_0 è un punto di massimo relativo allora

  (i) $f'(x_0) = 0$

  (ii) $f''(x_0) \leq 0$

Dimostrazione. Dalla continuità di f, esiste un intorno U_{x_0} tale che $f(x_0) \leq f(x)$ per ogni $x \in U_{x_0}$.

Se $x \in U_{x_0} \setminus \{x_0\}$ e $x > x_0$ allora $\frac{f(x)-f(x_0)}{x-x_0} \geq 0$.

Se $x \in U_{x_0} \setminus \{x_0\}$ e $x < x_0$ allora $\frac{f(x)-f(x_0)}{x-x_0} \leq 0$.

quindi

$$f'_+(x_0) = \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0} \geq 0 \geq \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0} = f'_-(x_0)$$

Essendo f derivabile in x_0, si deve avere che $f'_+(x_0) = f'_-(x_0)$, cioè $f'(x_0) = 0$.

La (i) è quindi dimostrata.

Si supponga, per assurdo, che $f''(x_0) < 0$; per il teorema di permanenza del segno risulta

$$f''(x_0) = \frac{f'(x) - f'(x_0)}{x - x_0} = \frac{f'(x)}{x - x_0} < 0$$

per ogni $x \in U_{x_0} \setminus \{x_0\}$. Dunque

$$f'(x) = \begin{cases} < 0 & \text{se } x \in U_{x_0} \cap (x_0, +\infty) \\ > 0 & \text{se } x \in U_{x_0} \cap (-\infty, x_0) \end{cases}$$

Quindi f decresce su $U_{x_0} \cap (x_0, +\infty)$ e cresce su $U_{x_0} \cap (-\infty, x_0)$, quindi x_0 non può essere un punto di minimo relativo per f. □

Teorema 1.3.5 (Condizioni necessarie per l'esistenza di massimi e minimi relativi)**.**

Data F: $\mathbb{R}^n \to \mathbb{R}$, $F \in C^2(\mathbb{R}^n)$, $x_0 \in \mathbb{R}^n$

- se x_0 è un punto di minimo relativo allora

  (i) $\nabla F(x_0) = 0$

  (ii) $H(F(x_0)) \geq 0$

- x_0 è un punto di massimo relativo allora

  (i) $\nabla F(x_0) = 0$

  (ii) $H(F(x_0)) \leq 0$

Dimostrazione. Sia $x \in \mathbb{R}^n$. Si definisca

$$\begin{aligned} \phi : \mathbb{R} &\longrightarrow \mathbb{R} \\ t &\mapsto F(x_0 + tx) \end{aligned}$$

Si osservi che se F ha un punto di minimo relativo in x_0, allora ϕ ha un punto di minimo relativo per $t = 0$.
Di conseguenza, per il Lemma 1.3.4

- $\phi'(0) = 0$
- $\phi''(0) \geq 0$

cioè

(A) $\phi'(0) = \nabla^T F(x_0) \cdot x = 0$

(B) $\phi''(0) = x^T \cdot H(F(x_0)) \cdot x \geq 0$

Per arbitrarietà di x, si ottiene che $H(F(x_0)) \geq 0$.
Inoltre, scegliendo $x = \nabla F(x_0)$ in (A), si ottiene $||\nabla F(x_0)||_2^2 = 0$, cioè $\nabla F(x_0) = 0$.

□

Teorema 1.3.6 (Condizioni sufficienti per l'esistenza di massimi e minimi relativi)**.**

Data F: $\mathbb{R}^n \to \mathbb{R}$, $F \in C^2(\mathbb{R}^n)$, $x_0 \in \mathbb{R}^d$ punto critico

- se $H(F(x_0)) > 0$ allora x_0 è un punto di minimo
- se $H(F(x_0)) < 0$ allora x_0 è un punto di massimo
- se $H(F(x_0))$ è semidefinita allora x_0 è un punto di sella

Dimostrazione. $H(F(x_0))$ è definita positiva, quindi per il Lemma 1.3.3 esiste una costante $m > 0$ tale che $\lambda^T \cdot H(F(x_0)) \cdot \lambda \geq m|\lambda|_2^2$, per ogni $\lambda \in \mathbb{R}^n$.
Scrivendo lo sviluppo di Taylor di F al secondo ordine, si ottiene

$$F(x+\lambda) = F(x_0) + \frac{1}{2}\lambda^T \cdot H(F(x_0)) \cdot \lambda + o(|\lambda|^2)$$

cioè

$$F(x+\lambda) - F(x_0) = \frac{1}{2}\lambda^T \cdot H(F(x_0)) \cdot \lambda + o(|\lambda|^2) \geq \frac{m}{2}|\lambda|^2 + o\left(|\lambda|^2\right) = |\lambda|^2\left(\frac{m}{2} + \frac{o\left(|\lambda|^2\right)}{|\lambda|^2}\right)$$

Sapendo che

$$\lim_{\lambda \to 0} \frac{o(|\lambda|^2)}{|\lambda|^2} = 0$$

usando la definizione di limite, cioè

$$\forall\, \epsilon > 0\ \exists\, \delta > 0 \text{ tale che se } |\lambda| < \delta \text{ allora } \frac{o(|\lambda|^2)}{|\lambda|^2} < \epsilon$$

e scegliendo $\epsilon = \frac{m}{4}$, si ottiene

$$\frac{o(|\lambda|^2)}{|\lambda|^2} < \frac{m}{4} \qquad \text{per ogni } \lambda \in \mathbb{R}^n \setminus \{0\} \text{ e } |\lambda| < \delta$$

Pertanto,

$$F(x_0+\lambda)-F(x_0) \geq |\lambda|^2 \left\{ \frac{m}{2} + \frac{o(|\lambda|^2}{|\lambda|^2} \right\} \geq |\lambda|^2 \left\{ \frac{m}{2} - \frac{m}{4} \right\} = \frac{m}{4}|\lambda|^2$$

per ogni $\lambda \in \mathbb{R}^n$ e $|\lambda| < \delta$

Quindi x_0 è un punto di minimo per F in un intorno $I_\delta(x_0)$, cioè x_0 è punto di minimo locale per F.

□

Teorema 1.3.7. Sia F: $\mathbb{R}^n \to \mathbb{R}$ convessa, differenziabile e x_0 è un punto critico per F, allora x_0 è punto di minimo globale.

Se inoltre F è fortemente convessa, x_0 è di minimo globale stretto ed è unico.

Dimostrazione. Per ipotesi, F è convessa e differenziabile in x_0, quindi per il Teorema 1.2.17 si ha che

$$F(x) \geq F(x_0) + \nabla F(x_0) \cdot (x - x_0) \tag{1.3.3}$$

per ogni $x \in \mathbb{R}^n$.

Essendo x_0 un punto critico, si ha che

$$F(x) \geq F(x_0)$$

per ogni $x \in \mathbb{R}^n$, cioè x_0 è un punto di minimo globale.

Se F è strettamente convessa, la 1.3.3 vale in senso stretto, quindi si ha

$$F(x) > F(x_0)$$

per ogni $x \in \mathbb{R}^n$, cioè x_0 è un punto di minimo globale stretto ed è unico.

□

Subdifferenziale

I riferimenti bibliografici del presente capitolo sono stati tratti e rielaborati da [3].

Il concetto di subdifferenziale entra in gioco quando si ha a che fare con funzioni non differenziabili.

Come si può definire un concetto analogo a quello di gradiente per funzioni non differenziabili?

Definizione 2.0.1. Data F: $\mathbb{R}^n \to \mathbb{R}$ convessa, fissato $x \in \mathbb{R}^n$, si definisce **subdifferenziale** di F in x l'insieme

$$\partial F(x) = \{r \in \mathbb{R}^n \mid r \cdot (x - \hat{x}) \geq F(x) - F(\hat{x}),\ \textit{per ogni}\ \hat{x} \in \mathbb{R}^n\}\,.$$

Interpretazione geometrica. Si definisce **iperpiano di supporto al grafico** di F in x la funzione $G(y) = a \cdot y + b$, dove $G(y) \leq F(y)$ per $y \in \mathbb{R}^n$ e $G(x) = F(x)$.

Il subdifferenziale $\partial F(x)$ indica quindi la pendenza di tutti gli iperpiani di supporto al grafico in x; inoltre, se F è differenziabile in x, l'unico iperpiano di supporto è l'iperpiano tangente e $\partial F(x) = \{\nabla F(x)\}$.

In generale $\partial F(x)$ può contenere più di un elemento poiché il grafico di F potrebbe avere, per esempio, un punto angoloso; in questo caso esisterà più di un iperpiano di supporto al grafico di F nel punto considerato.

Teorema 2.0.2 (Proprietà del subdifferenziale)**.** Data F: $\mathbb{R}^n \to \mathbb{R}$ convessa, si ha che per ogni $x \in \mathbb{R}^n$, $\partial F(x)$ è non vuoto, convesso e chiuso.

Dimostrazione. Sia $y \in \mathbb{R}^n$. Mostriamo che $\partial F(y)$ è non vuoto. Siccome F è convessa, quindi continua, si ha che il suo epigrafico $\mathscr{E}$ è un insieme convesso e chiuso. Dato $k > 0$, si consideri

$$e^k = \begin{bmatrix} y \\ F(y) - \frac{1}{k} \end{bmatrix} \notin \mathscr{E}$$

Per il Teorema 1.2.10, esiste un iperpiano $a_k \cdot x + b_k = 0 \subset \mathbb{R}^{n+1}$ che separa e^k ed $\mathscr{E}$, cioè

$$a_k \cdot x + b_k > 0 \qquad \text{per ogni } x \in E \tag{2.0.1}$$

$$a_k \cdot e^k + b_k < 0 \tag{2.0.2}$$

Posto $a_k = \begin{bmatrix} c_k \\ d_k \end{bmatrix}$, con $c_k \in \mathbb{R}^n$, $d_k \in \mathbb{R}$ dalla 2.0.2 si ha che

$$c_k \cdot y + d_k \left(f(y) - \frac{1}{k} \right) - b_k < 0$$

Equivalentemente,

$$-c_k \cdot y - d_k \left(F(y) - \frac{1}{k} \right) + b_k > 0 \tag{2.0.3}$$

F è definita su $\mathbb{R}^n$, quindi per ogni $\hat{x} \in \mathbb{R}^n$ si ha che

$$c_k \cdot \hat{x} + d_k F(\hat{x}) + b_k > 0 \tag{2.0.4}$$

Da 2.0.3 ed 2.0.4, si ottiene

$$c^k \cdot (y - \hat{x}) + d_k \left(F(\hat{x}) - F(y) + \frac{1}{k} \right) > 0 \tag{2.0.5}$$

con $\hat{x} \in \mathbb{R}^n$. Scegliendo $\hat{x} = y$, si ha $\frac{d_k}{k} > 0$. Ponendo $r_k = \frac{c_k}{d_k}$, dalla 2.0.5 segue che

$$F(\hat{x}) + \frac{1}{k} \geq r_k \cdot (\hat{x} - y) + F(y) \tag{2.0.6}$$

Mostriamo che $\{r_k\}_{k=1}^{\infty} \subset \mathbb{R}^n$ è limitata. In questo modo, per il teorema di Bolzano-Weierstrass, $\{r_k\}$ ammette un'estratta $\{r_{k_j}\}$ convergente a un elemento $r \in \mathbb{R}^n$.

Si supponga $r_k \neq 0$, sostituendo in 2.0.6 $\hat{x} = y + \frac{r_k}{|r_k|}$. Ne segue:

$$F\left(y + \frac{r_k}{|r_k|}\right) + \frac{1}{k} \geq |r_k| + F(y)$$

da cui si ottiene

$$|r_k| \leq (1 + \max_{x \in B(x,1)} |F(x)| + |F(y)|) := M$$

per ogni $k \in \mathbb{N}^+$. Quindi, per il teorema di Bolzano-Weierstrass, si ha che

$$\lim_{j \to +\infty} r_{k_j} = r \ \ esiste$$

dove $\{r_{k_j}\}$ è un'estratta di $\{r_k\}$. Ponendo $k = k_j \to +\infty$ in 2.0.6 si ottiene:

$$F(\hat{x}) \geq r \cdot (\hat{x} - y) + F(y)$$

con $\hat{x} \in \mathbb{R}^n$. Da cui $r \in \partial F(y)$.

Mostriamo che $\partial F(y)$ è convesso, cioè che la combinazione convessa di due elementi in $\partial F(y)$ è ancora contenuta in $\partial F(y)$.

Siano r_1, $r_2 \in \partial F(y)$. Bisogna verificare che preso t$\in [0,1]$, si ha che $tr_1 + (1-t)r_2 \in \partial F(y)$. Si osserva che $tr_1 \leq r_1$ e $(1-t)r_2 \leq r_1$, quindi se $r_1 \cdot (\hat{x} - y) \leq F(\hat{x}) - F(y)$, allora $tr_1 \cdot (\hat{x} - y) \leq t(F(\hat{x}) - F(y))$; analogamente, si ottiene che $(1-t)r_2 \cdot (\hat{x} - y) \leq (1-t)(F(\hat{x}) - F(y))$. In definitiva, considerando:

$$\begin{cases} tr_1 \cdot (\hat{x} - y) \leq t(F(\hat{x}) - F(y)) \\ (1-t)r_2 \cdot (\hat{x} - y) \leq (1-t)(F(\hat{x}) - F(y)) \end{cases}$$

sommando membro a membro si ottiene

$$[tr_1 + (1-t)r_2] \cdot (\hat{x} - x) \leq F(\hat{x}) - F(y)$$

cioè $tr_1 + (1-t)r_2 \in \partial F(y)$.

Si osserva che in $\mathbb{R}^n$ un insieme è chiuso se e soltanto se è chiuso per successioni.

Mostriamo che $\partial F(x)$ è chiuso per successioni, ossia considerata $\{x_n\} \subset \partial F(x)$ successione convergente e posto

$$\tilde{x} = \lim_{x \to \infty} x_n$$

si ha che $\tilde{x} \in \partial F(x)$.

Sia $\{r_k\}_{k=1}^{\infty} \subseteq \partial F(x)$ come sopra e sia $\tilde{r}$ il suo limite. Per ogni $k \in \mathbb{N}^+$ e per ogni $\hat{x} \in \mathbb{R}^n$, vale

$$F(x) + r_k \cdot (\hat{x} - x) \leq F(\hat{x})$$

Mandando $k \to +\infty$ si ottiene

$$F(x) + \tilde{r} \cdot (\hat{x} - x) \leq F(\hat{x})$$

per ogni $\hat{x} \in \mathbb{R}^n$, e quindi $\tilde{r} \in \partial F(x)$. Di conseguenza $\partial F(x)$ è chiuso per successioni, cioè chiuso. □

Ottimizzazione in Machine Learning

I riferimenti bibliografici di questo capitolo sono stati tratti e rielaborati da [2], ad eccezione della sezione 3.3 che è stata studiata e rielaborata da [1].

Sia $F : \mathbb{R}^d \to \mathbb{R}$. Si supponga di voler trovare $\hat{x} \in \mathbb{R}^d$ tale che

$$\hat{x} = \min_{x \in \mathbb{R}^d} F(x)$$

Di seguito un metodo iterativo utilizzato per la ricerca di tale minimo.

3.1 Il Metodo del Subgradiente

Passo 1. Si sceglie $x_0 \in \mathbb{R}^d$.

Per ogni $i \geq 1$ si ponga

$$x_i = x_{i-1} - \gamma_i z_{i-1} \tag{3.1.1}$$

dove $z_{i-1} \in \partial F(x_{i-1})$.

Chiaramente, se F è differenziabile, z_{i-1} coincide con $\nabla F(x_{i-1})$ e il metodo prende il nome di Metodo del Gradiente.

Passo 2. Come scegliere γ_i? Si possono usare diversi modi:

(a) costante: si può scegliere una sequenza del passo costante in cui si assegna lo stesso valore $\gamma := \gamma_i$ a ogni passo dell'algoritmo.

Il vantaggio di questa scelta potrebbe essere la facilità di implementazione, tuttavia non è sempre detto sia il più efficiente in termini di velocità di convergenza. In alcuni casi, una sequenza "dinamica" del passo γ_i potrebbe avere una convergenza più rapida. Tuttavia, se si ha una approfondita conoscenza del problema e si osserva che un certo valore γ_0 funziona "sufficientemente bene" per la buona riuscita dell'algoritmo, si può optare per selezionare $\gamma_i = \gamma_0$ per ogni i.

(b) decrescente: si può scegliere una sequenza $\{\gamma_i\}$ decrescente, per esempio, scelto γ_0, si può porre $\gamma_i = \frac{\gamma_0}{i}$. Questa scelta non ha nulla di interessante e la convergenza del metodo dipende ovviamente dalle caratteristiche del metodo stesso.

(c) ricerca di linea: questo metodo consiste nel cercare una direzione di discesa lungo la quale la funzione obiettivo f verrà ridotta, per poi calcolare la dimensione del passo γ_i: è qui che entra il gioco il metodo del gradiente, il quale insieme al metodo quasi-Newton costituisce uno degli algoritmi più utilizzati nella ricerca di linea.

Quando si arresta l'algoritmo?

Anche l'arresto dell'algoritmo può variare seconda del problema specifico. Alcune delle strategie più comuni per determinare quando arrestare l'algoritmo sono:

- numero massimo di iterazioni: si fissa un numero massimo di iterazioni oltre il quale l'algoritmo si arresta.

- convergenza della funzione obiettivo: si fissa una soglia $\epsilon > 0$ e si controlla la variazione della funzione obiettivo: se la differenza tra i valori della funzione obiettivo tra due iterazioni consecutive è inferiore alla soglia predefinita ϵ, l'algoritmo si arresta.

- convergenza del gradiente: si fissa una soglia $\epsilon > 0$ e si controlla se la norma del gradiente della funzione obiettivo è inferiore a ϵ, l'algoritmo si arresta.

Definizione 3.1.1. Data $F : \mathbb{R}^d \to \mathbb{R}$ differenziabile, si dice μ-**fortemente convessa**, con $\mu > 0$ se per ogni x, $y \in \mathbb{R}^d$ si verifica

$$F(x) \geq F(y) + \nabla F(y)^T \cdot (x - y) + \frac{\mu}{2}\|x - y\|_2^2$$

Osservazione 3.1.2. Una funzione μ-fortemente convessa è chiaramente fortemente convessa: dal Teorema 1.2.17, supponendo l'ipotesi di μ-forte convessità si ha subito la disuguaglianza stretta in (i).
(Nella Definizione 3.1.1, basta prendere $x = \hat{x}$, $y = x$ come nel Teorema 1.2.17; essendo $\mu > 0$, il termine $\frac{\mu}{2}\|x - y\|_2^2$ consente di avere la disuguaglianza stretta.)

Definizione 3.1.3. Data $F : \mathbb{R}^d \to \mathbb{R}$ differenziabile, si dice $L-$**regolare** se per ogni x, $y \in \mathbb{R}^d$ si verifica

$$|F(x) - F(y) - \nabla F(y)^T \cdot (x - y)| \leq \frac{L}{2}\|x - y\|_2^2$$

Nota 3.1.4. La condizone di L-regolarità implica che la funzione non può variare "troppo rapidamente" in un intorno di qualsiasi punto del dominio, ed L rappresenta un limite superiore su quanto rapidamente può cambiare la funzione.

Nota 3.1.5. Quanto asserito nella Definizione 3.1.3 equivale a dire che il gradiente di F è L-Lipschitz-continuo, cioè per ogni x, $y \in \mathbb{R}^d$ si ha che $\|\nabla F(x) - \nabla F(y)\|_2^2 \leq L^2\|x - y\|_2^2$.
Se si hanno ipotesi più forti sulla regolarità di F, supponiamo per esempio che F sia differenziabile due volte, si può dire che $-LI \preceq H(F(x)) \preceq LI$: ciò significa che la matrice hessiana di F in x ossia la curvatura della funzione in x, è limitata inferiormente da LI. Analogamente, la curvatura è superiormente limitata, il che significa che non ci sono rischi di "instabilità" intorno al punto considerato.

3.1.1 Il metodo applicato a funzioni L-regolari e μ-fortemente convesse

Lemma 3.1.6 (**Disuguaglianza di Lojasiewicz**)**.** Data $F : \mathbb{R}^d \to \mathbb{R}$ differenziabile, μ-fortemente convessa avente un unico elemento $\tilde{x}$ tale che $F(\tilde{x}) := \min_{x \in \mathbb{R}^d} F(x)$, allora per ogni $x \in \mathbb{R}^d$ vale

$$\|\nabla F(x)\|_2^2 \geq 2\mu(F(x) - F(\tilde{x})) \tag{3.1.2}$$

Dimostrazione. Per ipotesi, F è μ-fortemente convessa quindi per ogni $x, y \in \mathbb{R}^d$ si verifica la 3.1.1. Fissato $y \in \mathbb{R}^d$, poniamo $K(x) := F(y) + \nabla F(y)^T \cdot (x - y) + \frac{\mu}{2}\|x - y\|_2^2$: si osserva che il valore $x^* = y - \frac{1}{\mu}\nabla F(x)$ costituisce un punto di minimo; infatti calcolando $\nabla K(x)$, si ottiene

$$\nabla K(x) = \nabla F(y)^T + \mu(x - y) \tag{3.1.3}$$

Ponendo la 3.1.3 $= 0$, si trova che

$$x^* = -\frac{1}{\mu}\nabla F(y) + y \tag{3.1.4}$$

Inoltre K è convessa, (si verifica in modo immediato usando la definizione); dal Teorema 1.3.7, si ottiene che x^* è un punto di minimo globale per K.
Sapendo che F è μ-fortemente convessa, sostituendo $x = \tilde{x}$ nella definizione 3.1.1 si ottiene:

$$F(\tilde{x}) \geq K(\tilde{x}) \geq K(x^*) = F(y) - \frac{1}{\mu}\|\nabla F(y)\|_2^2 + \frac{1}{2\mu}\|\nabla F(y)\|_2^2 = F(y) - \frac{1}{2\mu}\|\nabla F(y)\|_2^2$$

cioè

$$F(\tilde{x}) \geq F(y) - \frac{1}{2\mu}\|\nabla F(y)\|_2^2$$

che equivale a

$$\|\nabla F(y)\|_2^2 \geq 2\mu(F(y) - F(\tilde{x}))$$

□

Teorema 3.1.7. Data $F : \mathbb{R}^d \to \mathbb{R}$ differenziabile, L-regolare e μ-fortemente convessa, si ponga in 3.1.1 $\gamma_i := \frac{1}{L}$. Per ogni $i \geq 0$, si verifica che

$$F(x_i) - F(x^*) \leq \left(1 - \frac{1}{k}\right)^i (F(x_0) - F(x^*)) \leq \exp\left(-\frac{i}{k}\right)(F(x_0) - F(x^*)) \tag{3.1.5}$$

dove si è indicato con:

- $x^* := \min_{x \in \mathbb{R}^d} F(x)$
- $k = \frac{L}{\mu}$

Dimostrazione. Da 3.1.1

$$x_i = x_{i-1} - \gamma_i \nabla F(x_{i-1})$$

allora,

$$F(x_i) = F(x_{i-1} - \gamma_i \nabla F(x_{i-1}))$$

F è $L-$regolare, quindi vale la disuguaglianza della definizione 3.1.3. Ponendo

(a) $y = x_{i-1}$

(b) $x = x_i$

si ottiene

$F(x_i) \leq F(x_{i-1}) + \frac{L}{2}\|\frac{1}{L}\nabla F(x_{i-1})\|_2^2 - \nabla F(x_{i-1})^T \cdot \frac{1}{L}\nabla F(x_{i-1}) = F(x_{i-1}) - \frac{1}{2L}\|\nabla F(x_{i-1})\|_2^2$

cioè

$$F(x_i) - F(x^*) \leq F(x_{i-1}) - F(x^*) - \frac{1}{2L}\|\nabla F(x_{i-1})\|_2^2$$

Usando il Lemma 3.1.6, si ottiene

$$F(x_i) - F(x^*) \leq \left(1 - \frac{\mu}{L}\right)(F(x_{i-1}) - F(x^*)) \leq \exp\left(-\frac{\mu}{L}\right)(F(x_{i-1}) - F(x^*))$$

Iterando si ottiene la tesi. □

Osservazione 3.1.8. Se $F : \mathbb{R}^d \to \mathbb{R}$ differenziabile, L-regolare e μ-fortemente convessa, la disuguaglianza

$$1 - \frac{\mu}{L} \leq \exp\left(-\frac{\mu}{L}\right)$$

è sempre verificata.

Dimostrazione. Osserviamo che $\frac{\mu}{L} < 1$, infatti, supponendo regolarità fino al secondo ordine e scrivendo lo sviluppo di Taylor di F centrato in x^*, si ottiene

$$F(x) - F(x^*) = \frac{1}{2}(x - x^*)^T \cdot H(F(x^*)) \cdot (x - x^*)$$

Dalla ipotesi di μ-convessità, si ottiene

$$\frac{1}{2}(x - x^*)^T \cdot H(F(x^*)) \cdot (x - x^*) \geq \frac{\mu}{2}||x - x^*||_2^2 = \frac{\mu}{2}(x - x^*)^T \cdot I_d \cdot (x - x^*) \tag{3.1.6}$$

dove con I_d si è indicata la matrice quadrata identità in $\mathbb{R}^d$.
Sottraendo ambo i membri la quantità $\frac{\mu}{2}(x - x^*)^T \cdot I_d \cdot (x - x^*)$, si ottiene

$$(x - x^*)^T \cdot [H(F(x)) - \mu I_d] \cdot (x - x^*) \geq 0$$

cioè la matrice $H(F(x)) - \mu I_d$ è semidefinita positiva, quindi se $\{\lambda_i\}_{i=1}^d \subset \mathbb{C}$ è l'insieme degli autovalori associati ad $H(F(x))$, allora $\lambda_i - \mu \geq 0$ per ogni i.
Analogamente, utilizzando l'ipotesi di L-regolarità, si ottiene

$$\frac{1}{2}(x - x^*)^T \cdot H(F(x)) \cdot (x - x^*) \leq \frac{L}{2}||x - x^*|| = \frac{L}{2}(x - x^*)^T \cdot I_d \cdot (x - x^*)$$

cioè

$$(x - x^*)^T \cdot [H(F(x)) - L I_d] \cdot (x - x^*) \leq 0$$

ovvero, $L \geq \lambda_i$ per ogni i.

Quindi,

$$\frac{\mu}{L} \leq 1$$

cioè

$$1 - \frac{\mu}{L} < 0 \tag{3.1.7}$$

Si noti che la disuguaglianza non è stretta ma se $\mu = L$, tutti gli autovalori sarebbero uguali tra loro e il caso sarebbe poco interessante.

Sfruttando il fatto che per ogni $x \in \mathbb{R} \cap (-1, +\infty)$ si ha la seguente disuguaglianza

$$log(1 + x) \leq x$$

applichiamo il log alla quantità $1 - \frac{\mu}{L}$ ottenendo $log(1 - \frac{\mu}{L}) \leq -\frac{\mu}{L}$.

Essendo la funzione $g(x) - e^x$ crescente, segue immediatamente che

$$1 - \frac{\mu}{L} = g(log(1 - \frac{\mu}{L})) \leq g(-\frac{\mu}{L}) = \exp\left(-\frac{\mu}{L}\right)$$

□

3.1.2 Cosa accade se si indebolisce l'ipotesi di μ-convessità?

In questa sezione si esaminerà cosa accade indebolendo l'ipotesi di μ-convessità con la più semplice ipotesi di convessità; sarà però necessario aggiungere l'ipotesi di *"co-coercività"*: in questo modo si potrà avere una convergenza dell'algoritmo con ordine $\frac{1}{i}$. Formalizziamo il tutto.

Lemma 3.1.9 (Co-coercività)**.** Sia $F : \mathbb{R}^d \to \mathbb{R}$ una funzione convessa e L-regolare, allora per ogni x, $y \in \mathbb{R}^d$ si ha che

$$\frac{1}{L}\|\nabla F(y) - \nabla F(x)\|_2^2 \leq (\nabla F(y) - \nabla F(x))^T \cdot (y - x) \tag{3.1.8}$$

In più si può affermare che

$$F(x) \geq F(y) + \nabla F(y)^T \cdot (x - y) + \frac{1}{2L}\|\nabla F(x) - \nabla F(y)\|_2^2 \tag{3.1.9}$$

Nota 3.1.10. Se F è come nella 3.1.8, si dice che F è **co-coerciva**.

Dimostrazione. Dimostrando la 3.1.9, la 3.1.8 seguirà immediatamente applicandola rispettivamente a x e y e sommando le due espressioni. Infatti si avrà:

$$\frac{1}{2L}||\nabla F(x) - \nabla F(y)||_2^2 \leq F(x) - F(y) - \nabla F(y)^T \cdot (x - y)$$

$$\frac{1}{2L}||\nabla F(x) - \nabla F(y)||_2^2 \leq F(y) - F(x) + \nabla F(x)^T \cdot (x - y)$$

Sommando entrambi i membri si ottiene che:

$$\frac{1}{L}||\nabla F(x) - \nabla F(y)||_2^2 \leq (\nabla F(x) - \nabla F(y))^T \cdot (x - y)$$

cioè

$$\frac{1}{L}||\nabla F(x) - \nabla F(y)||_2^2 = \frac{1}{L}||\nabla F(y) - \nabla F(x)||_2^2 \leq (\nabla F(y) - \nabla F(x))^T \cdot (y - x)$$

Dimostriamo la 3.1.9. Fissato $y \in \mathbb{R}^d$, sia $G : \mathbb{R}^d \to \mathbb{R}$ definita da

$$G(x) = F(x) - x^T \nabla F(y)$$

Per $x = y$ si ha un punto di minimo per G.

Infatti $\nabla G(x) = \nabla F(x) - \nabla F(y) = 0 \Rightarrow x = y$.

Inoltre, G è convessa; dal Teorema 1.2.17 si ha che y è un punto di minimo per G.

Si osserva inoltre che G è ancora $L-$regolare. Prendendo

$$x = x - \frac{1}{L}\nabla G(x)$$

nella Definizione 3.1.3 si ha che

$$G\left(x-\frac{1}{L}\nabla G(x)\right) \leq G(x)+\nabla G(x)^T\cdot\left(-\frac{1}{L}\nabla G(x)\right)+\frac{L}{2}\left\|-\frac{1}{L}\nabla G(x)\right\|_2^2 = G(x)-\frac{1}{2L}\left\|\nabla G(x)\right\| \tag{3.1.10}$$

Siccome in y si è in corrispondenza di un punto di minimo, ricordando come è definita $G(x)$, osservando che $\nabla G(x)=\nabla F(x)-\nabla F(y)$ e usando la disuguaglianza 3.1.10, si ottiene

$$G(y)=F(y)-y^T\nabla F(y)\leq G(x-\frac{1}{L}\nabla G(x))\leq G(x)-\frac{1}{2L}\|\nabla G(x)\|=$$

$$=F(x)-x^T\cdot\nabla F(y)-\frac{1}{2L}\|\nabla F(x)-\nabla F(y)\|$$

cioè

$$F(x)-x^T\cdot\nabla F(y)-\frac{1}{2L}\|\nabla F(x)-\nabla F(y)\|\geq F(y)-y^T\nabla F(y)$$

che equivale a

$$F(x)\geq F(y)+(x^T-y^T)\cdot\nabla F(y)+\frac{1}{2L}\|\nabla F(x)-\nabla F(y)\|$$

□

Teorema 3.1.11. Data $F:\mathbb{R}^d\to\mathbb{R}$ L-regolare e convessa, si ponga in 3.1.1 $\gamma_i:=\frac{1}{L}$. Per ogni $i\geq 0$, si verifica che

$$F(x_i)-F(x^*)\leq\frac{L}{2i}||x_0-x^*|| \tag{3.1.11}$$

dove x^* è un punto di minimo globale.

Dimostrazione. Consideriamo la funzione

$$V(x_i):=i[F(x_i)-F(x^*)]+\frac{L}{2}||x_i-x^*||_2^2$$

Obiettivo: dimostrare che la funzione V decresce all'aumentare delle iterazioni.

Si consideri

$$V(x_i) - V(x_{i-1}) = i[F(x_i) - F(x^*)] + \frac{L}{2}||x_i - x^*||_2^2 - (i-1)[F(x_{i-1}) - F(x^*)] - \frac{L}{2}||x_{i-1} - x^*||_2^2$$

$$= iF(x_i) + \frac{L}{2}||x_i - x^*||_2^2 - iF(x_{i-1}) + F(x_{i-1}) - F(x^*) - \frac{L}{2}||x_{i-1} - x^*||_2^2$$

cioè

$$V(x_i) - V(x_{i-1}) = i[F(x_i) - F(x_{i-1})] + F(x_{i-1}) - F(x^*) + \frac{L}{2}||x_i - x^*||_2^2 - \frac{L}{2}||x_{i-1} - x^*||_2^2$$

Si osserva che:

(1) $F(x_i) - F(x_{i-1}) \leq -\frac{1}{2L}||\nabla F(x_{i-1})||_2^2$ (segue dalla Definizione 3.1.3 e da 3.1.1 sostituendo il valore di γ_i)

(2) $F(x_{i-1}) - F(x^*) \leq \nabla F(x_{i-1})^T \cdot (x_{i-1} - x^*)$ (segue dal Teorema 1.2.17)

(3) $\frac{L}{2}||x_i - x^*||_2^2 - \frac{L}{2}||x_{i-1} - x^*||_2^2 = \frac{L}{2}||x_{i-1} - \gamma\nabla F(x_{i-1}) - x^*||_2^2 - \frac{L}{2}||x_{i-1} - x^*||_2^2 =$
$= \frac{L}{2}||x_{i-1} - x^*||_2^2 + \frac{L}{2}\gamma^2||\nabla F(x_{i-1})||_2^2 - L\gamma(x_{i-1} - x^*)^T \cdot \nabla F(x_{i-1}) - \frac{L}{2}||x_{i-1} - x^*||_2^2 =$
$= -L\gamma(x_{i-1} - x^*)^T \cdot \nabla F(x_{i-1}) + \frac{L}{2}\gamma^2||\nabla F(x_{i-1}||_2^2$

Ne segue che

$$\begin{aligned} V(x_i) - V(x_{i-1}) \leq & -\frac{1}{2L}\|\nabla F(x_{i-1})\|_2^2 \\ & + \nabla F(x_{i-1})^T(x_{i-1} - x^*) - L\gamma(x_{i-1} - x^*)^T\nabla F(x_{i-1}) \\ & + \frac{L\gamma^2}{2}\|\nabla F(x_{i-1})\| = -\frac{i-1}{2L}||\nabla F(x_{i-1})||_2^2 \leq 0 \end{aligned}$$

quindi, per ogni i si ha che

$$V(x_i) \leq V(x_{i-1})$$

cioè V è non crescente rispetto alle iterate. Ricordando la definizione di V, si ha che

$$i[F(x_i) - F(x^*)] \leq V(x_i) \leq V(x_0) = \frac{L}{2}||x_0 - x^*||_2^2$$

da cui

$$F(x_i) - F(x^*) \leq \frac{L}{2i}||x_0 - x^*||_2^2$$

□

Nota 3.1.12. Nel caso **non convesso**, non ci sono garanzie circa la convergenza a un minimo globale come invece è assicurata nel caso convesso, ma si può affermare che almeno una delle iterazioni ha un gradiente "sufficientemente piccolo". Tuttavia questo non dà risposte sulla convergenza al minimo globale, perciò l'ipotesi di convessità è necessaria per avere speranza di convergenza.

Si potrebbe però provare a rimuovere l'ipotesi di regolarità. Di seguito si esaminerà cosa accade.

3.2 Analisi del comportamento del Metodo del Gradiente senza ipotesi di regolarità

Definizione 3.2.1. Data $F : \mathbb{R}^d \to \mathbb{R}$ se per ogni x, $y \in \mathbb{R}^d$ vale che

$$|F(x) - F(y)| \leq B||x - y||_2$$

si dice che F è **B-Liptschitz-continua**.

Definizione 3.2.2. Sia $\mathcal{P}$ una proprietà e sia E un insieme su cui è definita una misura.
Si dice che $\mathcal{P}$ vale in E quasi ovunque (q.o.) se vale per ogni $x \in E$ tranne al più un sottoinsieme di E avente misura nulla.

Osservazione 3.2.3. A priori le funzioni come in Definizione 3.2.1 non sono differenziabili. Si può dimostrare che lo sono quasi ovunque. (crf. Teorema di Rademacher)
Quello che si andrà a studiare nel prossimo teorema è la convergenza del Metodo del Subgradiente su questo particolare insieme di funzioni.

Lemma 3.2.4. Sia $F : \mathbb{R}^d \to \mathbb{R}$ B-Liptschitz-continua e convessa. Fissato $x \in \mathbb{R}^d$ si ha che $\|z\|_2 \leq B$ per ogni $z \in \partial F(x)$.

Dimostrazione. Sia $z \in \partial F(x)$.
Per ogni $y \in \mathbb{R}^n$, $F(y) - F(x) \geq z \cdot (y - x)$.
Si scelga $y = x + tz$, $t \geq 0$.

(i) Dalla definizione di subdifferenziale si ha $F(x+tz)-F(x) \geq z\cdot(x+tz-x) = z\cdot tz = t\|z\|_2^2$.

(ii) Dall'ipotesi di B-Liptschitz-continuità si ha $F(x + tz) - F(x) \leq B\|tz\|_2 = Bt\|z\|_2$.

cioè $t\|z\|_2^2 \leq F(x + tz) - F(x) \leq Bt\|z\|_2$
Da cui $t\|z\|_2^2 \leq Bt\|z\|_2$, cioè la tesi.

□

Teorema 3.2.5. Data $F : \mathbb{R}^d \to \mathbb{R}$ B-Liptschitz-continua e convessa; sia x^* un punto di minimo per F tale che $||x_0 - x^*||_2 \leq D$, si ponga in 3.1.1 $\gamma_i := \frac{D}{B\sqrt{i}}$. Per ogni $i \geq 1$, si verifica che

$$\min_{0\leq k\leq i-1} F(x_k) - F(x^*) \leq DB\frac{2 + log(i)}{2\sqrt{i}} \tag{3.2.1}$$

Dimostrazione. Si consideri

$$||x_i - x^*||_2^2$$

Ricordiamo che nella 3.1.1, z_{i-1} è un qualsiasi elemento di $\partial F(x_{i-1})$; in realtà, una buona scelta, ricordando lo scopo finale dell'algoritmo, cioè quello di minimizzare la quantità $||x_i - x^*||_2^2$, sarebbe quella di considerare l'elemento di norma minima.

In ogni caso, ai fini della dimostrazione, la stima ottenuta sarà la seguente:

$$||x_i - x^*||_2^2 = ||x_{i-1} - \gamma_i z_{i-1} - x^*||_2^2 = ||x_{i-1} - x^*||_2^2 + \gamma_i^2 ||z_{i-1}||_2^2 - 2\gamma_i z_{i-1}^T \cdot (x_i - x^*) \quad (3.2.2)$$

Dalla convessità di F, (ricordando che F è q.o. differenziabile, quindi $z_{i-1} = \nabla F(x_{i-1})$ per quasi ogni i) sfruttando il Teorema 1.2.17 e tenendo a mente l'ipotesi di B-Liptschitz continuità, si ottiene

$$||z_{i-1}||_2 \leq \frac{|F(x) - F(x_{i-1})|}{||x - x_{i-1}||_2} \leq B$$

Sostituendo in 3.2.2, si ottiene

$$||x_i - x^*||_2^2 \leq ||x_{i-1} - x^*||_2^2 + \gamma_i^2 B^2 - 2\gamma_i [F(x_{i-1}) - F(x^*)] \quad (3.2.3)$$

Si osserva che stavolta la funzione $x \mapsto ||x - x^*||_2^2$ non è decrescente a causa del termine $\gamma_i^2 B^2$, quindi non si possono replicare i passaggi del Teorema 3.1.11.

Si riscriva la 3.2.3 in questo modo

$$\gamma_i [F(x_{i-1}) - F(x^*)] \leq \frac{1}{2} \left(||x_{i-1} - x^*||_2^2 - ||x_i - x^*||_2^2 \right) + \frac{1}{2} \gamma_i^2 B^2 \quad (3.2.4)$$

La 3.2.4 vale per ogni i, quindi si considera la somma finita

$$\sum_{k=1}^{i} \gamma_k [F(x_{k-1}) - F(x^*)] \leq \frac{1}{2} \sum_{k=1}^{i} \left(||x_{k-1} - x^*||_2^2 - ||x_k - x^*||_2^2 \right) + \frac{1}{2} B^2 \sum_{k=1}^{i} \gamma_k^2 \quad (3.2.5)$$

Si osserva che la prima somma finita del termine a destra equivale a

$$\sum_{k=1}^{i} \left(||x_{k-1} - x^*||_2^2 - ||x_k - x^*||_2^2 \right) = ||x_0 - x^*||_2^2 - ||x_i - x^*||_2^2 \leq ||x_0 - x^*||_2^2 \quad (3.2.6)$$

cioè

$$\sum_{k=1}^{i} \gamma_k [F(x_{k-1}) - F(x^*)] \leq ||x_0 - x^*||_2^2 + \frac{1}{2} B^2 \sum_{k=1}^{i} \gamma_k^2 \tag{3.2.7}$$

Dividendo ambo i membri per $\sum_{k=1}^{i} \gamma_k$ si ottiene

$$\frac{1}{\sum_{k=1}^{i} \gamma_k} \sum_{k=1}^{i} \gamma_k [F(x_{k-1}) - F(x^*)] \leq \frac{||x_0 - x^*||_2^2}{2 \sum_{k=1}^{i} \gamma_k} + B^2 \frac{\sum_{k=1}^{i} \gamma_k^2}{2 \sum_{k=1}^{i} \gamma_k} \tag{3.2.8}$$

Il termine di sinistra nella 3.2.8 è una media pesata quindi si avrà che

$$\min_{1 \leq k \leq i-1} F(x_k) - F(x^*) \leq \frac{1}{\sum_{k=1}^{i} \gamma_k} \sum_{k=1}^{i} \gamma_k [F(x_{k-1}) - F(x^*)] \tag{3.2.9}$$

Si ponga $\gamma_k = \frac{D}{B\sqrt{k}}$ in 3.2.9 e si osservi che:
(1) $\sum_{k=1}^{i} \gamma_k = \frac{D}{B} \sum_{k=1}^{i} \frac{1}{\sqrt{k}} \geq \frac{D}{B} \sum_{k=1}^{i} \frac{1}{\sqrt{i}} = \frac{D}{B}\sqrt{i} \iff \frac{1}{\sum_{k=1}^{i} \gamma_k} \leq \frac{B}{D\sqrt{i}}$
(2) $\sum_{k=1}^{i} \gamma_k^2 = \frac{D^2}{B^2} \sum_{k=1}^{i} \frac{1}{k} = \frac{D^2}{B^2} \left[1 + \sum_{k=2}^{i} \frac{1}{k}\right] \leq \frac{D^2}{B^2} \left[1 + \int_1^i \frac{1}{k} dk\right] = \frac{D^2}{B^2} [1 + log(i)]$
Maggiorando come in 3.2.8, si ottiene

$$\min_{1 \leq k \leq i} F(x_{k-1}) - F(x^*) \leq \frac{1}{2 \sum_{k=1}^{i} \gamma_k} \left[D^2 + B^2 \sum_{k=1}^{i} \gamma_k^2\right] \leq \frac{B}{2D\sqrt{i}} \left[D^2 + B^2 \frac{D^2}{B^2} (1 + log(i))\right]$$

$$= \frac{BD}{2\sqrt{i}} [2 + log(i)]$$

□

3.3 Alcune loss-function in Machine Learning

In questa sezione si andranno a studiare alcune tra le più comuni loss function utilizzate in Machine Learning.

In questo ambito, esistono due tipi di problemi: di classificazione o di regressione.

Il primo consiste, appunto, nel "classificare" i dati in input: supponendo di avere n input, corrispondenti ai nomi di alcuni pazienti che devono ritirare le risposte delle analisi del sangue, in un problema di classificazione ci si preoccupa di classificare i pazienti in modo corretto, per esempio come diabetico o non diabetico.

Il secondo, invece, non consiste nel "classificare" ma nello "stimare" alcuni valori reali. Si chiama regressione perché quello che effettivamente si fa é "regredire" una funzione a partire da alcuni input.

Notazione:

n = numero di campioni

y_i = valore effettivo per l'i-esimo campione

$\hat{y}_i$ = valore previsto per l'i-esimo campione

Esempio 3.3.1 (**Mean Absolute Error** o **L1 Loss**, poco funzionale)**.**

$M.A.E. = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \sum_{i=1}^{n}|e_i|$

La $M.A.E.$ rappresenta la media della somma degli scarti tra i valori effettivi e quelli calcolati dal metodo. Questa funzione è perlopiù utilizzata per problemi di regressione, ma se l'errore medio è molto piccolo, potrebbe essere più efficiente utilizzare l'*errore quadratico medio* che introdurremo successivamente.

Aspetto negativo: Quando si utilizza il $M.A.E.$ in metodi di ottimizzazione come quello del gradiente, bisogna considerare la possibilità che i gradienti siano "grandi"; poiché ciò si verifica anche quando la perdita è bassa, cioè quando si è abbastanza vicini alla soluzione, questo non è affatto ottimale per l'apprendimento: c'è infatti il rischio di oscillare attorno al minimo e, addirittura, di allontanarsi.

In questo caso si considera infatti un'altra loss function, la Huber-loss.

Si osserva subito che la $M.A.E.$ non è differenziabile ovunque, infatti per $e_i = 0$ non ammette derivata prima. Questo è un altro motivo per cui è conveniente usare altre loss, come la $M.S.E.$.

Esempio 3.3.2 (**Mean Squared Error**)**.**

$M.S.E. = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y_i}|^2 = \sum_{i=1}^{n} |e_i|^2$

La $M.S.E$ è simile alla $M.A.E.$, differisce per il termine della sommatoria che, stavolta, è quadratico; in questo modo, purtroppo, si ha un errore maggiore quando le previsioni sono "lontane" dalla soluzione. D'altro canto, per errori piccoli si ha una convergenza al minimo più rapida, inoltre la $M.S.E.$ è differenziabile ovunque.

In ogni caso, la scelta tra $M.A.E.$ e $M.S.E.$ dipende fortemente dai dati in input con cui si sta lavorando.

Sulla base di questi esempi, si può introdurre un'ulteriore loss, ancora più "funzionale", in quanto restituisce una percentuale di "correttezza" e, quindi, un vero e proprio confronto con la soluzione effettiva.

Esempio 3.3.3 (**Mean Absolute Percentage Error**)**.**

$M.A.P.E. = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y_i}}{y_i}\right|$

Come già accennato, il vantaggio nell'utilizzare la $M.A.P.E.$, è quello di avere un'idea di quanto il modello funzioni "bene" o "male".

Inoltre, potendo confrontare le percentuali, consente di confrontare le prestazioni dei modelli di regressione su diversi set di dati in modo "veloce": si supponga di voler addestrare un modello di regressione sul fondo di investimento aperto da X e su quello aperto da Y e si supponga di avere quantità molto diverse in valore assoluto.

L'utilizzo di $M.A.E.$ non sarà molto funzionale per confrontare le prestazioni del modello, in quanto si conosce l'errore ma quest'ultimo non è effettivamente rapportato a nulla. D'altra parte, con $M.A.P.E.$ si considera l'errore in termini di percentuale e si considerano due quantità effettivamente confrontabili senza bisogno di manipolazioni.

Esempio 3.3.4 (***Huber-loss***)**.**

$$Huber - Loss = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & se \ \ |y_i - \hat{y}_i| \leq \delta \\ \delta|y_i - \hat{y}_i| - \frac{\delta^2}{2} & se \ \ |y_i - \hat{y}_i| > \delta \end{cases}$$

Figura 3.1: Huber-Loss per alcuni valori di δ

Legenda:

- la funzione g corrisponde alla $Huber - loss$ per $\delta = 3$
- la funzione f corrisponde alla $Huber - loss$ per $\delta = 1$
- la funzione h corrisponde alla $Huber - loss$ per $\delta = 0.2$
- la funzione l corrisponde alla $Huber - loss$ per $\delta = 0.1$

Per delta relativamente piccoli, come si vede anche nella Figura 3.1, la funzione di perdita diventa relativamente "piatta" e, per questo, l'aumento della perdita richiede molto tempo.

Per delta più grandi, invece, la "ripidità" della funzione aumenta.

Si osservi che:

- per δ piccoli, la perdita diventa relativamente insensibile a errori più grandi e ciò potrebbe essere positivo in caso di errori grandi, ma negativo per errori piccoli.
- per δ grande, la perdita diventa sempre più sensibile a errori più grandi.

Effettivamente il comportamento della $Huber - Loss$ sembra simile a qualcosa già visto, infatti riguardo il $M.A.E.$ (in cui si aveva insensibilità rispetto errori grandi) e il $M.S.E.$ (più sensibile a questi ultimi) sono state fatte precedentemente trattazioni simili.

Si può infatti affermare che con la $Huber-loss$ si ha un comportamento simile a $M.A.E.$ quando $\delta \to 0$ e a $M.S.E.$ quando $\delta \to \infty$.

La presenza del termine δ ha quindi un ruolo ben preciso, consente infatti di "controllare" la ripidità della funzione. Si sta quindi parlando di un "approccio dinamico", infatti quando si verificano errori di grandi dimensioni, è possibile riprovare con un δ inferiore; se invece gli errori sono troppo piccoli per essere rilevati dalla $Huber - loss$, si può aumentare il δ.

Quali sono gli svantaggi della *Huber-loss*?

Lo svantaggio sta proprio nell' "approccio dinamico", infatti, poiché il valore δ va configurato manualmente, ciò che richiede più tempo è, appunto, la ricerca di un valore δ adatto al set di dati considerato. É un problema che, in alcuni casi, potrebbe essere poco pratico e costoso.

Esempio 3.3.5 (***Hinge-loss***)**.**

La $Hinge - Loss$ è una loss function usata in problemi di classificazione.

$Hinge - loss = \max\{0, 1 - ty\}$

Dove t prende il nome di target mentre y è l'output della funzione decisionale del classificatore. Per funzione decisionale si intende una funzione che prende in input il set di dati che si sta considerando, cioè gli x^i e produce un valore numerico che rappresenta la confidenza o la distanza del modello rispetto all'obiettivo effettivo.

Solitamente, tale funzione è definita come segue:

$$f(x) = \omega \cdot x + b$$

dove ω è il vettore dei pesi, scelto in modo da massimizzare la capacità del modello e renderlo il più esatto possibile.
Il target, invece, è l'etichetta di classe corretta associata a un set di dati in input in un problema di classificazione. In altre parole, è la "risposta corretta" che il modello di classificazione dovrebbe imparare a predire.
Quello che si vuole fare è trovare il vettore dei pesi ω in modo da minimizzare, in questo caso, la $Hinge - Loss$ così da produrre predizioni il più possibile vicine alle etichette vere; per farlo si utilizzano algoritmi di ottimizzazione come l'Algoritmo del Gradiente.
In problemi di classificazione binaria, il target t, generalmente, appartiene all'insieme $\{\pm 1\}$.
Di seguito un esempio pratico per comprendere meglio il significato dei termini.

> *Si supponga di avere un set di dati che rappresentano le email nella posta elettronica del signor X e di voler classificare le email come "spam" o "non spam".*

Innanzitutto, si comprende bene che le email spam verranno etichettate con $t = 1$ e, ovviamente, quelle non spam con l'etichetta $t = -1$.
Ecco come potrebbe apparire il set di dati:

E-mail	**Target**
1- Vuoi vincere un nuovo smartphone? Premi sul tasto qui in basso!	+1
2- AMAZON-Conferma ordine n°1749012	-1
3- Vinci un coupon da 500 €da spendere dove vuoi	+1
4- Conferma il tuo appuntamento in banca per il giorno 28/10/2024	-1

Tuttavia, il modello non sa riconoscere a priori quali email possono essere classificate come spam e quali no. Per questo motivo, è necessario fornire un dataset adeguato, che descriva al meglio le caratteristiche di una possibile spam-email.

Di seguito un semplice esempio.

n°Email	**Numero di parole in maiuscolo**	**Frequenza della parola "offerta"**	**Frequenza di punti esclamativi e interrogativi**	**Target**
1	10	3	7	+1
2	1	0	1	-1
3	2	2	5	+1

In questo caso, il dataset è costituito da un vettore $x \in \mathbb{R}^4$, dove

- x_1=numero dell'email in ordine di arrivo
- x_2=numero di parole in maiuscolo presenti nella email
- x_3=numero di volte che compare la parola "offerta" nella email
- x_4=numero di volte che compaiono punti esclamativi e interrogativi

Nel primo caso, l'input sarà $x^1 = [1, 10, 3, 7]$ e si impone il target a 1, cioè spam.
Durante il processo di addestramento, il modello apprende le caratteristiche in input in modo da poter fare previsioni accurate sulle nuove email non viste durante l'addestramento.
Tornando alla $Hinge - loss$, studiamo l'utilizzo che ha in machine learning.

- se $ty > 1$, la perdita è zero. Questo significa che il modello ha fatto una previsione corretta.
- se $ty \leq 1$ la perdita è $1 - ty$. In questo caso, la perdita aumenta linearmente man mano che il modello si allontana dalla classificazione corretta.

Un aspetto importante della $Hinge - Loss$ è che è sensibile solo alle classificazioni errate, mentre è insensibile alle istanze che sono classificate correttamente.
Consideriamo la $Hinge - Loss$ con target 1.
Possono verificarsi tre possibilità:

- la previsione è corretta e si verifica quando $y \geq 1$.
- la previsione è molto errata, che si verifica quando $y < 0$.
- la previsione non è corretta, ma ci si sta avvicinando $0 \leq y < 1$.

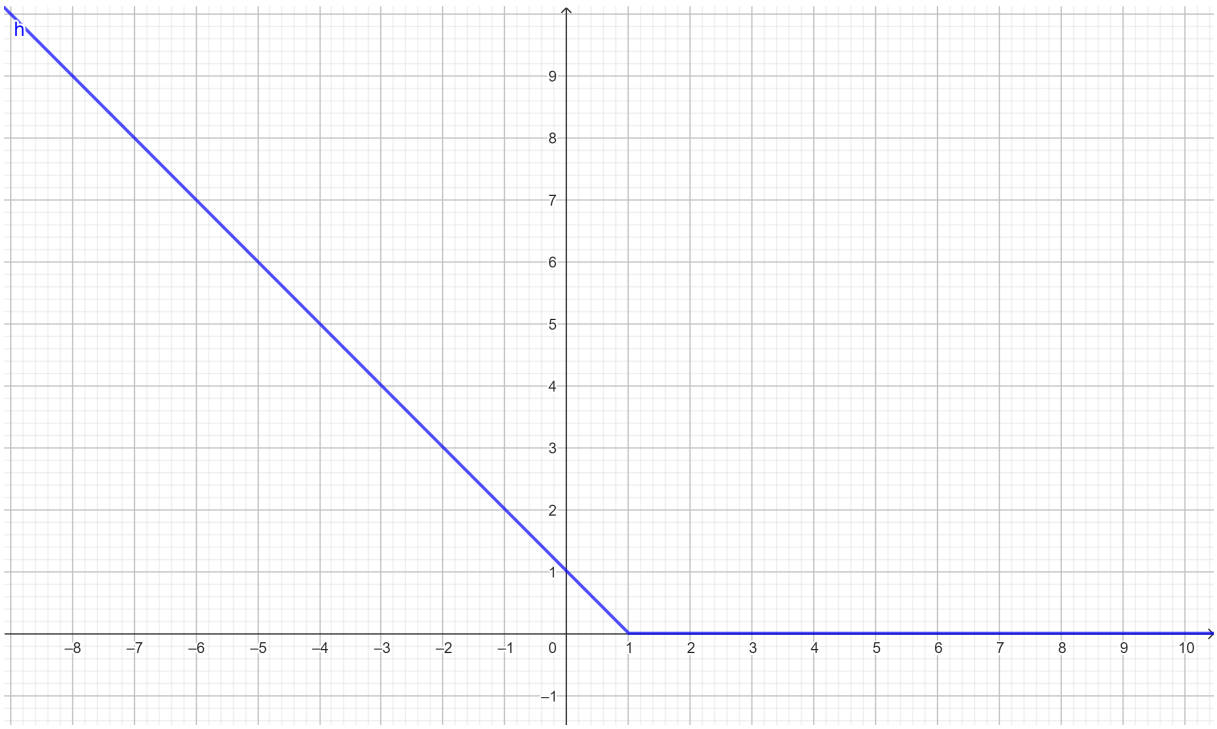

Figura 3.2: Hinge-Loss per $t = 1$

Nel primo caso, per esempio quando $y = 1.2$, la $Hinge - loss$ sarà $\max(0, -0.2) = 0$. Quindi, la perdita è zero.

Nel secondo caso, per esempio quando $y = -0.5$, l'output della loss sarà 1.5.

Nel terzo caso, per esempio quando $y = 0.9$, la funzione di perdita in uscita sarà 0.1, cioè ci si sta avvicinando e questo è indicato anche dalla perdita piccola ma non nulla.

Ciò che essenzialmente accade con la $Hinge - Loss$ è che quando la previsione è corretta o molto corretta, non ha molta importanza, ma quando non lo è, si cerca di correggerla. Il processo di correzione continua finché la previsione non è completamente corretta, o quando si decide di interrompere il processo "dall'esterno".

3.4 Aumentare la velocità di convergenza del metodo del gradiente: l'algoritmo di Nesterov

L'algoritmo del gradiente è tra i più semplici metodi di ottimizzazione, ma non è il più efficace; per questo motivo, nel tempo sono state studiate e introdotte diverse varianti al fine di accelerarne la convergenza.

Una delle modifiche più significative consiste nel metodo di Nesterov, proposto dal matematico russo Yurii Nesterov.

Questo metodo apporta una semplice modifica all'algoritmo del gradiente, migliorando i tassi di convergenza, in particolare per le funzioni fortemente convesse.

L'algoritmo di Nesterov si basa su un simultaneo aggiornamento di due successioni, x_i e y_i, dove, mantenendo la notazione, x_i rappresenta l'i-esima iterata nel metodo del gradiente.

Si può dimostrare che una scelta ottimale del passo è $\gamma_i = \frac{1}{L}$ per ogni i.

Per funzioni L-regolari e μ-fortemente convesse, le iterate dell'algoritmo con γ_i come sopra sono definite come segue:

$$x_i = y_{i-1} - \frac{1}{L}\nabla F(y_{i-1})$$
$$y_i = x_i + \frac{1 - \sqrt{\frac{\mu}{L}}}{1 + \sqrt{\frac{\mu}{L}}}(x_i - x_{i-1})$$

Un aspetto negativo potrebbe essere che il metodo di Nesterov su funzioni L-regolari e μ-fortemente convesse richiede la conoscenza del valore μ, cosa che per il metodo del gradiente classico non è necessaria.

Per funzioni convesse, invece, si aggiunge una semplice modifica che dipende da i, cioè

$$x_i = y_{i-1} - \frac{1}{L}\nabla F(y_{i-1})$$
$$y_i = x_i + \frac{i-1}{i+2}(x_i - x_{i-1})$$

così facendo, nel primo caso si ottiene la seguente stima tra la distanza dall'ottimo e la k-esima iterata.

$$F(x_i) - F(x^*) \leq L||x_0 - x^*||_2^2 \left(1 - \sqrt{\frac{\mu}{L}}\right)^i$$

mentre nel secondo caso si ottiene

$$F(x_i) - F(x^*) \leq \frac{2L||x_0 - x^*||_2^2}{(i+1)^2}$$

.

Implementazione dell'algoritmo

In questo capitolo si esamineranno i dati raccolti da vari test fatti su alcune funzioni convesse, quasi ovunque differenziabili e aventi minimo utilizzando l'Algoritmo del Gradiente e il Metodo di Nesterov implementati in Matlab. Di seguito gli pseudocodici di entrambi.

Notazione:

- f = funzione da minimizzare
- x = variabile dipendente
- x_0 = dato iniziale
- gamma = parametro del metodo
- eps = tolleranza in input
- x_opt = valore di x ottimo
- f_opt = valore di f valutata in x_opt
- iter = numero di iterazioni
- f_val = valore di f valutato in qualche iterata

Algoritmo 1: Metodo del Gradiente

Data: f, x, x_0, $gamma$, eps
Result: x_opt, f_opt, iter

$x\_j \leftarrow x\_0$;
$f\_val \leftarrow \text{subs}(f, x, x\_j)$;
$f\_opt \leftarrow f\_val$;
iter $\leftarrow$ 1;
gradf $\leftarrow$ diff(f);
gradf1 $\leftarrow$ subs(gradf, x, x_j);
$x\_i \leftarrow x\_j$-gamma*gradf1;

while $|x\_i - x\_j| > \text{eps}$ **do**
    $f\_val \leftarrow \text{subs}(f, x, x\_j)$;
    $x\_j \leftarrow x\_i$;
    iter $\leftarrow$ iter + 1;
    gradf1 $\leftarrow$ subs(gradf, x, x_j);
    $x\_i \leftarrow x\_j$-gamma* gradf1;
end
$f\_opt \leftarrow f\_val$;
$x\_opt \leftarrow x\_j$

Algoritmo 2: Metodo di Nesterov

Data: f, x, x_0, $gamma$, eps
Result: x_opt, f_opt, iter

$x\_j \leftarrow x\_0$;
$y\_j \leftarrow x\_0$;
$f\_val \leftarrow \text{subs}(f, x, x\_j)$;
$f\_opt \leftarrow f\_val$;
iter $\leftarrow$ 1;
gradf $\leftarrow$ diff(f);
gradf1 $\leftarrow$ subs(gradf, x, y_j);
$x\_i \leftarrow y\_j$-gamma*gradf1;

while $|x\_i - x\_j| > \text{eps}$ **do**
    $f\_val \leftarrow \text{subs}(f, x, x\_i)$;
    $x\_j \leftarrow x\_i$;
    iter $\leftarrow$ iter + 1;
    gradf1 $\leftarrow$ subs(gradf, x, y_j);
    $x\_i \leftarrow x\_j$-gamma* gradf1;
    $y\_j \leftarrow x\_i + \frac{(iter-1)}{(iter+2} * (x\_i - x\_j)$;
end
$f\_opt \leftarrow f\_val$;
$x\_opt \leftarrow x\_j$

Di seguito i risultati raccolti dal test sulla funzione $f(x) = |x|^3$

	METODO DEL GRADIENTE	**METODO DI NESTEROV**
Dato iniziale x_0	2	2
Tolleranza scelta	0.1	0.1
Minimo calcolato	0.49710	-0.25300
Minimo reale	0	0
Valutazione di f nel minimo calcolato	0.12284	0.01619
Valutazione di f nel minimo reale	0	0
Numero di iterazioni	4	4
Tempo impiegato	0.04239 s	0.04947 s

Provando a ridurre la tolleranza, si ottengono i dati raccolti nella tabella che segue.

	METODO DEL GRADIENTE	**METODO DI NESTEROV**
Dato iniziale x_0	2	2
Tolleranza scelta	0.01	0.01
Minimo calcolato	0.17738	-0.18787
Minimo reale	0	0
Valutazione di f nel minimo calcolato	0.00558	0.00663
Valutazione di f nel minimo reale	0	0
Numero di iterazioni	15	9
Tempo impiegato	0.30894 s	0.08479 s

Nelle tabelle sottostanti si esaminerà cosa accade cambiando dato iniziale, in particolare nella prima si avrà una tolleranza di 0.1 e nella seconda di 0.01.

	METODO DEL GRADIENTE	**METODO DI NESTEROV**
Dato iniziale x_0	3	3
Tolleranza scelta	0.1	0.1
Minimo calcolato	0.29999	-0.16760
Minimo reale	0	0
Valutazione di f nel minimo calcolato	0.02699	0.00470
Valutazione di f nel minimo reale	0	0
Numero di iterazioni	2	6
Tempo impiegato	0.02593 s	0.04915 s

	METODO DEL GRADIENTE	**METODO DI NESTEROV**
Dato iniziale x_0	2	2
Tolleranza scelta	0.01	0.01
Minimo calcolato	0.29999	-0.16760
Minimo reale	0	0
Valutazione di f nel minimo calcolato	0.02699	0.00470
Valutazione di f nel minimo reale	0	0
Numero di iterazioni	2	6
Tempo impiegato	0.02593 s	0.04915 s

Di seguito un esempio significativo, in cui si è scelta una tolleranza eps = 0.0055.

	METODO DEL GRADIENTE	**METODO DI NESTEROV**
Dato iniziale x_0	2	2
Tolleranza scelta	0.0055	0.0055
Minimo calcolato	0.13286	-0.13589
Minimo reale	0	0
Valutazione di f nel minimo calcolato	0.130915	0.00251
Valutazione di f nel minimo reale	0	0
Numero di iterazioni	21	16
Tempo impiegato	1791.3954 s (28 min circa)	2.86960 s

I dati dimostrano come il Metodo di Nesterov sia effettivamente più veloce e i tempi di output lo dimostrano.

La scelta del valore eps è cruciale: tale valore rappresenta infatti il punto medio tra 0.01, tolleranza i cui dati raccolti sono riportati sopra e 0.001, valore che, dopo oltre circa 15 ore di compilazione, non restituiva alcun output.

Di seguito invece verranno raccolti i dati riguardanti la funzione $f(x) = -cos(|x|)$, sulla quale, scegliendo in modo adatto il dato iniziale, vale a dire, ovviamente, dove la funzione è convessa e senza scegliere un punto estremale, il Metodo del Gradiente è particolarmente efficiente.

METODO DEL GRADIENTE				
Dato iniziale x_0	1.5	1.5	2	3
Tolleranza scelta	0.1	0.01	0.1	0.1
Minimo calcolato	0.0209	0.000001	0.00141	0.00236
Minimo reale	0	0	0	0
Valutazione di f nel minimo calcolato	-0.99978	0.99999	-0.99999	-0.99999
Valutazione di f nel minimo reale	-1	-1	-1	-1
Numero di iterazioni	3	4	4	7
Tempo impiegato	0.08131 s	0.07169 s	0.09189 s	1.9783 s

Infine, testando il metodo sulla funzione $f(x) = x^2 - 3|x|$, si è riusciti ad arrivare a una tolleranza di ordine 10^{-9}, scegliendo ovviamente un dato iniziale $x_0 = 5$.
Utilizzando il Metodo di Nesterov, le iterazioni compiute, con la tolleranza eps = 0.000000001, sono state ben 113; il tempo di convergenza calcolato è di 3.9445 s, il valore minimo calcolato, con la precisione di 5 cifre dopo la virgola è 1.5, coincidente col minimo effettivo; di conseguenza anche la funzione valutata nel minimo risulta avere una esattezza a 5 cifre dopo la virgola.
Con lo stesso dato iniziale e la medesima tolleranza, il Metodo del Gradiente risulta avere un'ottima convergenza: sono state contate 93 iterazioni e un tempo di convergenza di 4.0278 s. Gli output, alla quinta cifra dopo la virgola, sono gli stessi di cui sopra.

Bibliografia

[1] Loss functions in machine learning explained, https://www.datacamp.com/tutorial/loss-function-in-machine-learning.

[2] Francis Bach. Learning theory from first principles. Draft of a book, version of Sept, 6:2021, 2021.

[3] Lawrence C Evans. Mathematics 170 mathematical methods for optimization finite dimensional optimization.

[4] Nicola Fusco, Paolo Marcellini, and Carlo Sbordone. Lezioni di analisi matematica due. Zanichelli, 2020.

Ringraziamenti

Al termine di questo mio elaborato, vorrei ringraziare chi ha contribuito alla sua realizzazione. Ringrazio con sincera gratitudine il Dr. Mendico, incontrato per la prima volta in uno dei corsi che mi ha insegnato molto e mi ha permesso di capire ciò che fosse più adatto ai miei interessi. Lo ringrazio per la sua inestimabile guida, il suo costante impegno e l'interesse mostrato durante il processo di redazione di questo elaborato. La sua disponibilità e competenza hanno arricchito ogni fase di questo lavoro, trasformandolo in un percorso di apprendimento significativo e appassionante. La sua dedizione ha sicuramente contribuito in modo determinante al successo di questo progetto e per questo gli sono profondamente grata.

"La scelta di un giovane dipende dalla sua inclinazione,

ma anche dalla fortuna di incontrare un grande maestro."