Dipartimento di Matematica e Informatica
Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Metodi di ottimizzazione su varietà differenziabili per la media di Fréchet su disco di Poincaré

Relatore:
Bruno Iannazzo

Candidato:
Luca Moroni

Matricola:
311279

Anno Accademico 2020/2021

Alla mia famiglia ed alla loro estrema pazienza.

Contenuti

1 Introduzione

Nei giorni d'oggi sono molteplici le applicazioni pratiche nelle quali è necessario effettuare ottimizzazione di funzioni non lineari, ovvero trovarne un valore di minimoI; In particolare sono alla base del funzionamento di tecnologie molto utilizzate, tra cui le reti neurali il cui apprendimento pone le fondamenta sulla solida teoria matematica dei metodi numerici per l'ottimizzazione.
Negli ultimi anni si sta affermando una generalizzazione dell'ottimizzazione classica, dove la funzione da ottimizzare non ha come dominio un sottoinsieme dello spazio euclideo ma una varietà differenziabile.
Le applicazioni di queste nuove tecniche sono moltissime ad esempio possiamo pensare al calcolo di autospazi di una matrice, che rappresenta un problema molto comune e necessario da risolvere in molte applicazioni dell'algebra lineare; possiamo anche pensare alla decomposizione SVD la quale può essere anche modellizzata come un problema di ottimizzazione su varietà.
Un altro problema è l'analisi delle componenti indipendenti (Indipendent Component Analysis, ICA), conosciuto come separazione delle sorgenti e modellizzabile anch'esso come un problema di ottimizzazione su varietà, il quale ha ricevuto un forte interesse nella ricerca di questi ultimi anni dato dall'ampia applicazione nell'ambito biomedico.
L'applicazione su cui si basa la tesi è il calcolo del centroide di p-punti in una varietà riemanniana che può essere posto come un problema di ottimizzazione sulla varietà stessa, infatti è il minimo della funzione somma delle distanze al quadrato dai punti dati.
Questo problema ammette un unico minimo e questo accade in particolare per il disco di Poincaré che è una famosa varietà riemanniana ed è il più famoso modello di geometria non euclidea con curvatura negativa.
Verrà confrontata la risoluzione di tale problema con il corrispettivo problema trasferito nella varietà conforme iperboloide.
Confronteremo vari algoritmi di ottimizzazione su varietà per calcolare il centroide di p-punti e per questo faremo uso di metodi di discesa del gradiente.
É importante notare che tali metodi non possono essere implementati in modo naturale come avviene nel caso euclideo, è quindi necessaria una codifica del problema e sono necessari particolari strumenti, che verranno trattati ed esplicitati nella presente e questo poichè, contrariamente al caso euclideo, il gradiente non appartiene allo stesso spazio su cui è definita la funzione da minimizzare ma in uno spazio tangente alla varietà che è uno spazio ben diverso, perciò la normale somma del gradiente non garantirebbe il corretto funzionamento dei metodi che descriveremo, tra i quali sono presenti l'algoritmo a passo fisso, ricerca lineare ed altri più performanti.
Tale sperimentazione prende spunto dall'articolo [1], il nostro contributo consiste nel considerare una più ampia gamma di algoritmi tra cui il metodo di Barzilai Borwain riemanniano recentemente introdotto in [2] dove viene mostrata la sua

efficienza nel calcolo del centroide nella varietà delle matrici definite positive. I nostri esperimenti mostrano che, anche nel caso del disco di Poincaré e dell'iperboloide, questo algoritmo è estremamente più efficiente dei metodi classici del gradiente con ricerca lineare o passo fisso.
Capiremo inoltre se tali modelli conformi sono equivalenti, non solo da un punto di vista geometrico ma anche sotto un aspetto computazionale tale per cui effettuare ottimizzazione.
Inoltre tale trattazione vuole far comprendere la possibilità, in problemi di ottimizzazione su varietà, di interscambiare modelli conformi.
Gli esperimenti trattati vorrebbero infine dare l'incipit per continuare con altre sperimentazioni analoghe andando magari a trattare anche il modello del semispazio di Poincaré.

2 Metodi di ottimizzazione su $\mathbb{R}^n$ e varietà differenziabili

Nella presente trattazione i vettori verranno indicati tramite le ultime lettere minuscole dell'alfabeto x, v, w, che assumeremo essere vettori colonna.
Il trasposto di un vettore verrà identificato tramite il carattere apice ($'$), perciò per definire il prodotto scalare euclideo di due vettori u, v scriveremo $u'v$, mentre se si intende un prodotto scalare differente da quello euclideo, come è il prodotto scalare riemanniano, lo indicheremo con $\langle \cdot, \cdot \rangle$.
Ricordiamo che il prodotto scalare in una varietà riemanniana dipende in modo continuo da un punto nella varità.
In tale sezione andremo ad esplicitare le metodologie generali per trovare un minimo di una funzione f definita su una varieta differenziabile M a valori reali.
Notiamo che $\mathbb{R}^n$ è un caso particolare di varietà differenziabile.
Cercheremo inoltre di trovare quei punti che garantiscono l'ottimalità di una funzione, ovvero punti di massimo e di minimo, che descriveremo a breve.
Sia $f : \mathbb{R}^n \to \mathbb{R}$ non vincolata. Un vettore x^* è un minimo locale non vincolato per la funzione f se, informalmente, f assume un valore in x^* non più grande dei valori che assume in un intorno di x^* stesso. Più formalmente x^* è minimo locale per f se esiste $\varepsilon > 0$ tale che,

$$f(x^*) \leq f, \text{ per ogni } x \text{ con } \|x - x^*\| < \varepsilon,$$

$\|x\|$ rappresenta la norma euclidea del vettore x.
Un vettore x^* è un minimo globale non vincolato rispetto ad f se la funzione assume un valore in x^* non più grande dei valori che assume in ogni altro vettore nel proprio dominio, più formalmente x^* è tale che

$$f(x^*) \leq f, \text{ per ogni } x \in \mathbb{R}^n.$$

2.1 Condizioni necessarie per l'ottimalità

Se la funzione di costo f è differenziabile, possiamo utilizzare il gradiente e l'espansione in serie di Taylor per comparare il costo di un vettore con il costo di un suo intorno. Ci aspettiamo perciò che se x^* è un minimo non vincolato allora la variazione della funzione al primo ordine per una piccola variazione Δx è non negativa,

$$\nabla f(x^*)'\Delta x \geq 0.$$

Perciò valendo sia per Δx positivi che negativi abbiamo la seguente condizione necessaria,

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

Ci aspettiamo inoltre che anche la variazione della funzione al secondo ordine per una piccola variazione di Δx è non negativa,

$$\nabla f(x^*)'\Delta x + (1/2)\Delta x'\nabla^2 f(x^*)\Delta x \geq 0,$$

dato che la precedente osservazione ha imposto $\nabla f(x^*)'\Delta x = 0$ otteniamo che

$$\Delta x'\nabla^2 f(x^*)\Delta x \geq 0,$$

e che perciò

$$\nabla^2 f(x^*) \text{ è una matrice semidefinita positiva.}$$

Diamo il seguente risultato che sintetizza quanto detto fin'ora.

Proposizione 1 (Condizioni necessarie di ottimalità) *Sia x^* un minimo locale non vincolato di $f : \mathbb{R}^n \to R$, assumiamo f differenziabile con continuità in un insieme aperto S contenente x^*. Allora*

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

Se f è due volte differenziabile con continuità in S, allora

$$\nabla^2 f(x^*) \textit{ è una matrice semidefinita positiva.}$$

2.2 Il caso della convessità

Se la funzione f è convessa su un insieme convesso non ci sono distinzioni tra un minimo locale ed un minimo globale, tutti i punti di minimo locale sono anche punti di minimo globale.
Un fatto importante è che la condizione $\nabla f(x^*) = 0$ è sufficiente per l'ottimalità, da cui deriva il seguente risultato.

Proposizione 2 (Funzioni Convesse) *Sia $f : \mathbb{R}^n \to R$ una funzione convessa su un insieme convesso X.*

- *Un minimo locale di f su X è anche un minimo globale su X. Se inoltre f è strettamente convessa, allora esiste al più un minimo per f.*
- *Se f è convessa e l'insieme X è aperto, allora $\nabla f(x^*) = 0$ è una condizione necessaria e sufficiente per il vettore $x^* \in X$ per essere minimo globale di f su X.*

2.3 Condizioni sufficienti per l'ottimalità

Non è difficile trovare esempi in cui le condizione di ottimalita necessarie definite nelle sezione precedente [$\nabla f(x^*) = 0$ e $\nabla^2 f(x^*) \geq 0$] portino ad identificare punti che non sono di minimo locale, come ad esempio punti di sella o punti di massimo.
Supponiamo di avere un vettore x^* che soddisfa le seguenti condizioni

- $\nabla f(x^*) = 0$.
- $\nabla^2 f(x^*)$ è una matrice definita positiva.

Allora abbiamo che per ogni $\Delta x \neq 0$ con $\Delta x \in \mathbb{R}^n \setminus \{0\}$,

$$\Delta x' \nabla^2 f(x^*) \Delta x > 0.$$

Ciò implica che in x^* la variazione di f al secondo ordine data da un piccolo spostamento Δx è positiva, perciò f tende ad incrementare nell'intorno di x^* e di conseguenza le due condizioni di cui sopra sono necessarie e sufficienti per l'ottimalità locale di x^*.

Proposizione 3 (Condizioni di ottimalità sufficienti del secondo ordine) *Sia $f : S \subset \mathbb{R}^n \to R$ derivabile due volte con continuità. Si supponga che esista $x^* \in S$ che soddisfi le condizioni*

$$\nabla f(x^*) = 0 \quad e \quad \nabla^2 f(x^*) \text{è una matrice definita positiva.}$$

Allora, x^* è un minimo locale di f. In particolare esistono $\gamma > 0$ e $\varepsilon > 0$ tali che

$$f(x) \geq f(x^*) + \frac{\gamma}{2}\|x - x^*\|^2, \quad \textit{per ogni } x \quad \textit{con } \|x - x^*\| < \varepsilon$$

2.4 Discesa del gradiente in $\mathbb{R}^n$

Andiamo ora a definire nel dettaglio le principali metodologie computazionali di ottimizzazione non lineare su $\mathbb{R}^n$. Tali metodologie sarannno trattate con un occhio alle possibili applicazioni.
Consideriamo il problema di trovare il minimo non vincolato di una funzione $f : \mathbb{R}^n \to \mathbb{R}$. Nella maggior parte di casi non è possibile trovare una soluzione analitica, perciò l'approccio adottato è quello di applicare un algoritmo iterativo chiamato *iterative descent* che opera come segue: prendiamo un vettore iniziale x^0 e successivamente generiamo una sequenza $x^1, x^2, ...$ tale che f decresce ad ogni iterazione $f(x^{k+1}) < f(x^k)$ e sperabilmente converga un punto di minimo.
I principali algoritmi di discesa sono metodi del gradiente poichè sono basati sul gradiente della funzione f.

2.4.1 Selezionare la direzione di discesa

Sia $v \in \mathbb{R}^n \setminus 0$, $v^\perp$ è un iperpiano, contenente i punti tali che $v'd = 0$, divide lo spazio in due componenti connesse:

- $v'd > 0$ (semipiano).
- $v'd < 0$ (semipiano).

Se $v = \nabla f(x)$ ogni d tale che $v'd < 0$ è una direzione di decrescita, come mostrato nel seguente teorema.

Teorema 1 *Sia $f \in C^1(\Omega)$, con Ω aperto di $\mathbb{R}^n$ e sia $x \in \Omega$ e $\nabla f(x) \neq 0$ per ogni $d \in \mathbb{R}^n$ tale che $\nabla f(x)'d < 0$ ed esiste $\alpha^0 > 0$ tale che $f(x+\alpha d) < f(x)$ con $\alpha \in (0, \alpha^0]$*

Detto ciò possiamo dare una classe di algoritmi con $x^0 \in \Omega$ e dove $d_k \in \mathbb{R}^n$ è detta direzione ed α_k è detto passo, definiti secondo la seguente formula,

$$x^{k+1} = x^k + \alpha^k d^k \quad k = 0, 1, ...,$$

Dove se $\nabla f(x^k) \neq 0$ la direzione d^k è scelta in modo tale che

$$\nabla f(x^k)' d^k < 0.$$

Ci sono varie possibilita nella scelta di d^k e di α^k.

Molti metodi di gradiente sono definiti nella forma,

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

dove D^k è una matrice definita positiva e con $d^k = -D^k \nabla f(x^k)$ allora è immediato verificare che $\nabla f(x^k)' d^k < 0$.
A seconda della scelta di D^k abbiamo differenti metodi applicabili.

- **Steepest Descent**: $D^k = I, \quad k = 0, 1, ...$
- **Metodi di Newton**: $D^k = (\nabla^2 f(x^k))^{-1}, \quad k = 0, 1, ..., \nabla^2 f(x^k)$ deve essere definita positiva, proprietà sempre verificata se f è convessa.
- **Steepest Descent Riscalata**: D^k matrice diagonale.
- **Metodi di quasi Newton**: $D^k = H(x^k) \approx \nabla^2 f(x^k)$.

Nonostante i metodi di Newton convergono molto velocemente non sono utilizzati nella pratica, dal momento in cui il calcolo della matrice hessiana è computazionalmente oneroso.
Sono stati introdotti perciò i metodi di quasi Newton, nel qual caso la matrice D^k è scelta in modo tale che la direzione $d^k = -D^k \nabla f(x^k)$ tenda ad approssimare la direzione del metodo di Newton e D_k approssima la matrice hessiana. L'idea fondamentale riguardo questa metodologia è che tramite due iterate successive k, $k+1$ ed i valori x^k, x^{k+1}, $\nabla f(x^k)$, $\nabla f(x^{k+1})$ noi possiamo avere accesso a delle informazioni riguardo la matrice hessiana, in particolare abbiamo che

$$\nabla f(x^{k+1}) - \nabla f(x^k) \approx H(x^{k+1})(x^{k+1} - x^k).$$

Definiti $p^k = x^{k+1} - x^k$ e $q^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ allora D^{k+1} deve soddisfare l'uguaglianza $D^{k+1} p^k = q^k$, questo prende il nome di equazione delle secanti.

2.4.2 Selezione del passo

Ci sono numerose metodologie per la scelta del passo a^k nei metodi del gradiente, ne elenchiamo alcune.

Ricerca lineare esatta:
scegliamo α^k tale che minimizza la funzione costo f lungo la direzione d^k, perciò,

$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k).$$

Ricerca lineare esatta limitata:

Questa è una versione modificata della metodologia precedente, in vari casi più semplice da implementare.
Definito un scalare $s > 0$, α^k è scelto in modo tale che minimizza il costo di f nell'intervallo $[0, s]$,

$$f(x^k + \alpha^k d^k) = \min_{\alpha \in [0,s]} f(x^k + \alpha d^k).$$

Le due metodologie appena presentate sono solitamente implementate con l'aiuto delle tecniche risolutive dell'ottimizzazione in una variabile che è un problema certamente più facile da risolvere in modo analitico.

Metodo di Armijo:
Scegliere α^k in modo che garantisca una decrescita sufficiente necessaria per la convergenza del metodo sotto opportune ipotesi. Il metodo di armijo è stato definito come segue. Siano fissati λ, σ e γ con $0 < \sigma < 1$, $0 < \gamma < 1$ e $\lambda > 0$, sia $\alpha^k = \sigma^{m_k}\lambda$, dove m_k è il primo intero non negativo per cui vale

$$f(x^k) - f(x^k + \sigma^m \lambda d^k) \geq -\gamma \sigma^m \lambda \nabla f(x^k)' d^k.$$

Solitamente si scelgono i valori come segue, $\gamma \in [10^{-5}, 10^{-1}]$, il fattore di riduzione $\sigma \in [1/10, 1/2]$, possiamo scegliere $\lambda = 1$ e moltiplicare la direzione d^k per uno scalare.
Per questa metodologia diamo anche il seguente teorema che rappresenta un risultato importante.

Definizione 1 *Una successione di direzioni $\{d^k\}$ è detta limitata se vale la seguente,*

$$\|d^k\| < M \text{ per ogni } k \text{ con } M \in \mathbb{R}^+.$$

Definizione 2 *Una successione di direzioni $\{d^k\}$ è detta **gradient related** se per ogni sotto-successione di $\{x^k\}_{k \in K}$ che converge ad un punto non stazionario $\{d^k\}_{k \in K}$ è limitata ed inoltre vale,*

$$\limsup_{k \to \infty, k \in K} \nabla f(x)' d^k < 0.$$

Ora diamo il principale risultato che garantisce la convergenza globale per metodi di discesa che utilizzano la tecnica di Armijo.

Teorema 2 *Sia $\{x^k\}$ una sequenza infinita (non definitivamente uguale al suo limite) generata dal metodo $x^{k+1} = x^k + \alpha^k d^k$, dove α^k è generato con la regola di armijo e d^k è **gradient related**. Allora ogni punto limite è stazionario.*

Metodo a passo costante:
Si definisce un passo costante $s > 0$ e si fissa $\alpha^k = s, \quad k = 0, 1, ...$
Il passo costante è una metodologia veramente semplice da implementare ma si deve fare attenzione nella scelta del passo.
Se si sceglie un valore di s troppo grande allora il metodo divergerà, contrariamente se il passo s venisse scelto troppo piccolo l'ordine di convergenza risulterebbe troppo lento.

Metodo di diminuzione del passo:
Scegliamo il passo in modo tale che $\alpha^k \to 0$. Una problematica legata a questa metodologia è che il passo può diventare troppo piccolo per cui non può essere garantita la convergenza, per questa ragione viene richiesto che.

$$\sum_{n=1}^{\infty} \alpha^k = \infty.$$

2.5 Discesa del gradiente su varietà

Ora che abbiamo un contesto teorico e una consapevolezza di cosa voglia dire effettuare ottimizzazione su uno spazio euclideo n-dimensionale ($\mathbb{R}^n$).
Possiamo approcciare le metodologie computazionali per trovare un minimo non vincolato di una funzione $f : M \to \mathbb{R}$ analoghe a quella definite nella sezione precedente.
Prima di andare nel vivo del discorso dobbiamo però munirci degli strumenti necessari.
Richiamiamo la parte dedicata alle varietà differenziabili presente nell'appendice e aggiungiamo quanto segue.

2.5.1 Retrazioni

Su varietà differenziabili la nozione di muoversi nella direzione del vettore tangente rimanendo nella varietà stessa è generalizzato dal concetto di **Retrazione**.
Concettualmente una retrazione R su x, denominata come R_x, è una mappa che va da T_xM (spazio tangente in x rispetto a M) in M con una condizione di rigidità locale che preserva il gradiente in x.

Definizione 3 (Retrazione) *Una retrazione in una varietà M è una mappa liscia R che ha come dominio l'insieme dei T_xM per ogni $x \in M$ con le seguenti proprietà. Sia R_x la restrizione di R a T_xM.*

- *$R_x(0_x) = x$, dove 0_x denota l'elemento zero (origine) di T_xM.*
- *Con l'identificazione canonica $T_{0_x}T_xM \simeq T_xM$, R_x soddisfa $DR_x(0_x) = id_{T_xM}$,*
  dove id_{T_xM} denota la mappa identità su T_xM.

Ci riferiamo alla condizione $DR_x(0_x) = id_{T_xM}$ come condizione di *rigidità locale*. Equivalentemente per ogni vettore tangente ξ in T_xM, la curva $\gamma_\xi : t \mapsto R_x(t\xi)$ soddisfa $\gamma_\xi(0)' = \xi$.
Oltre a trasformare elementi di T_xM in elementi di M, un secondo importate scopo della retrazione R_x è quello di trasformare una funzione costo ($f : M \to R$) definita in un intorno di $x \in M$ in una funzione costo definita sullo spazio vettoriale T_xM. Nello specifico, data una funzione reale f su una varietà M su cui è definita una retrazione R, abbiamo che $\hat{f} = f \circ R$ denota il *pullback* di f attraverso R. Per $x \in M$, abbiamo che

$$\hat{f}_x = f \circ R_x$$

denota la restrizione di $\hat{f}$ su T_xM. Si noti che $\hat{f}_x$ è una funzione reale su uno spazio vettoriale. Dalla condizione di rigidità locale abbiamo che $D\hat{f}_x(0_x) = Df(x)$. Se M è dotato di una metrica riemanniana abbiamo

$$\nabla^{(R)} \hat{f}_x(0_x) = \nabla^{(R)} f(x).$$

Un caso particolare di retrazione è la **mappa esponenziale** definita su $p \in M$ e $v \in T_pM$ che ha proprietà di retrazione e rappresenta la curva geodedica calcolata nel punto 1 originaria in p con derivata in 0 pari a v.
In alcuni casi può essere necessario dover effettuare operazioni tra vettori appartenenti a spazi tangente diffenti come ad esempio nel caso dei metodi di ottimizzazzione quasi Newton.
Nell'ottimizzazione basata su retrazioni, lo strumento standard per trasportare vettori appartenenti ad uno spazio tangente in un altro spazio tangente, è il *trasporto vettoriale*.
Informalmente, dati due vettori $x, y \in M$ e due vettori tangenti ad x, chiamati $v_x, w_x \in T_xM$, un trasporto vettoriale $\mathcal{T}$ associato ad una retrazione R è una mappa differenziabile che genera un vettore appartenente ad $T_{R_x(w_x)}M$ e per semplicità utilizzeremo la notazione $\mathcal{T}_{x \to y}(v)$.
Un caso speciale di trasporto vettoriale è il *trasporto parallelo* che può essere interpretato come un trasporto vettoriale la cui retrazione associata è la mappa esponenziale.

2.5.2 Metodi *line-search*

I metodi di ricerca in linea (definiti *line-search*) su varietà si basano sulla formula di aggiornamento

$$x^{k+1} = R_{x^k}(t^k \eta^k),$$

dove η^k è in $T_{x^k}M$ e t^k è uno scalare. Una volta scelta la retrazione R ci rimane

da decidere la direzione η^k e la lunghezza del passo t^k. Per ottenere la convergenza del metodo delle restrizioni sulla scelta di questi due parametri devono essere fatte.

Definizione 4 (sequenza gradient related su varietà) *Data una funzione di costo f su una varietà riemanniana M, una sequenza $\{\eta^k\}$, $\eta^k \in T_xM$ è **gradient related** se per ogni sotto-sequenza $\{x^k\}_{k\in K}$ di $\{x^k\}$ che converge ad un punto non stazionario di f, la corrispondente sotto-sequenza $\{\eta^k\}_{k\in K}$ è limitata e soddisfa*

$$\limsup_{k\to\infty, k\in K} \langle \nabla^{(R)} f(x^k), \eta^k \rangle_{x^k} < 0.$$

Definizione 5 (Punto di Armijo su varietà) *Data una funzione f su una varietà riemanniana M dotata di una retrazione R, un punto $x \in M$, un vettore tangente $\eta \in T_xM$, e degli scalari $\bar{\alpha} > 0, \beta, \sigma \in (0,1)$, un **punto di Armijo su varietà** è $\eta^A = t^A\eta = \beta^m\bar{\alpha}\eta$, dove m è il più piccolo intero non negativo tale che*

$$f(x) - f(R_x(\beta^m\bar{\alpha}\eta)) \geq -\sigma\langle \nabla^{(R)} f(x), \beta^m\bar{\alpha}\eta\rangle_x.$$

*Il valore reale t^A è il **Passo di Armijo**.*

Andiamo ora a definire un algoritmo generico per la discesa di gradiente su varietà (**Accelerated Line Search**).
Tale algoritmo rappresenta una generalità di metodologie di ottimizzazione delineando nel contempo, delle restrizioni fondamentali che garantiscono la convergenza della sequenza di punti generati.

Algoritmo 1: Accelerated Line Search (ALS)

Input: Varietà M; funzione costo f differenziabile definita su M;
retrazione R; scalari $\bar{\alpha} > 0$, $c, \beta, \sigma \in (0,1)$; punto iniziale $x^0 \in M$
Output: Sequenza $\{x^k\}$
for $k = 0, 1, 2, \ldots$ **do**
Prendere $\eta^k \in T_xM$ tale per cui la sequenza $\{\eta^i\}_{i=0,1,\ldots}$ sia *gradient related.*
Selezionare x^{k+1} tale che

$$f(x^k) - f(x^{k+1}) \geq c(f(x^k) - f(R_{x^k}(t^{k^A}\eta^k))) \tag{1}$$

Dove t^{k^A} è il passo definito dalla regola di Armijo su varietà definita poco sopra.
end

Dallo pseudocodice appena esplicitato è chiaro che tale algoritmo puo essere visto come una generalizzazione della scelta del passo di Armijo descritto nella sezione 2.4.2, difatti scegliere $x^{k+1} = R_{x^k}(t^{k^A}\eta^k)$ soddisfa (1). Inoltre la condizione rilassata (1) ci permette un ampio spazio di manovra nella scelta di x^{k+1} garantendo la convergenza ad un punto stazionario (come enunceremo formalmente) e ciò puo portare alla definizione di algoritmi altamente efficienti.

2.5.3 Convergenza

Il concetto di convergenza puo essere ridefinito e generalizzato su varietà. Una sequenza infinita $\{x^k\}_{k=0,1,\dots}$ di punti su una varietà M è detta essere convergente se esiste una carta (U, ψ) di M, un punto $x^* \in M$ e $K > 0$ tali che x^k è in U per ogni $k \geq K$ e tale che la sequenza $\{\psi(x^k)\}_{x=K,K+1,\dots}$ converge a $\psi(x^*)$, in tal caso applichiamo il concetto di convergenza di $\mathbb{R}^n$, avendo ψ come codominio uno spazio euclideo.
Il punto $\psi^{-1}(\lim_{k\to\infty} \psi(x^k))$ è chiamato il *limite* della sequenza $\{x^k\}_{k=0,1,\dots}$.
Tutte le sequenze convergenti di una varietà di *Hausdorff* hanno uno ed un solo punto limite.
Diamo ora un importante risultato rispetto alla convergenza dell'algoritmo *ALS*.
Tale enunciato deriva direttamente dal risultato di convergenza definito nella sezione 2.4.2 rispetto al passo di Armijo in $\mathbb{R}^n$, in tal caso viene data una generalizzazione di quest'ultimo essendo, come già detto, ALS una generalizzazione del passo di Armijo e lavorando non più su $\mathbb{R}^n$ ma su varietà differenziabili.

Teorema 3 *Sia $\{x^k\}$ una sequenza infinita di iterate generate da ALS. Allora ogni punto di accumulazione di $\{x^k\}$ è un punto stazionario della funzione costo f*

Inoltre assumendo compattezza della varietà M possiamo enunciare il seguente.

Corollario 1 *Sia $\{x^k\}$ una sequenza finita di iterate generata da ALS. Assumiamo che l'insieme di livello $L = \{x \in M : f(x) \leq f(x^0)\}$ è compatto allora*

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

3 Disco di Poincaré e Iperboloide

Lo spazio iperbolico è un particolare spazio su cui è definita una geometria che soddisfa i primi quattro assiomi di euclide ma non il quinto, tale spazio può essere definito in vari modi equivalenti.
Noi descriveremo l'iperboloide ed il disco di Poincaré, ciascuno di questi è un modello dello spazio iperbolico, nessuno dei due è prevalente in letteratura.

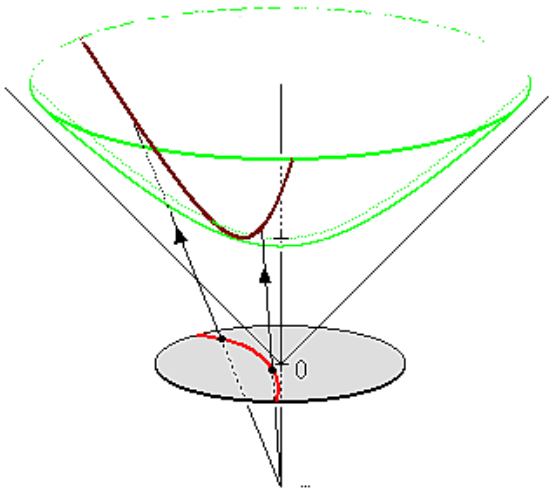

Figura 1: Rappresentazione della proiezione stereografica tra disco e iperboloide

3.1 Iperboloide

Possiamo definire $\mathbb{H}^n$ come il luogo dei punti tali che $\langle \cdot, \cdot \rangle_{n:1} = -1$. Sia $\mathbb{R}^{n:1}$ l'insieme $\mathbb{R}^{n+1}$ dotato dell'usuale prodotto di Minkowski ($\langle u, v \rangle_{n:1} = \sum_{i=1}^{n} u_i v_i - u_{n+1} v_{n+1}; \quad u, v \in \mathbb{R}^{n+1}$), questo luogo di punti ha in realta due componenti connesse, noi ne scegliamo una.
Diamo la definizione di tale modello dello spazio iperbolico.

Definizione 6 (Iperboloide) *Il modello dell'iperboloide è definito nel modo seguente.*

$$\mathbb{H}^n = \{x \in \mathbb{R}^{n+1} | \langle x, x \rangle_{n:1} = -1; \quad x^{n+1} > 0\}.$$

Si può dimostrare che lo spazio tangente in un punto $p \in \mathbb{H}^n$ è $T_p\mathbb{H}^n \approx \{x \in \mathbb{R}^{n:1} \mid \langle p, x \rangle_{n:1} = 0\}$. Il prodotto di Minkowski è definito positivo su $T_p\mathbb{H}^n$ e dunque induce una norma $\| \cdot \|_{n:1} = \sqrt{\langle \cdot, \cdot \rangle_{n:1}}$, perciò l'iperboloide è una **varietà riemanniana**.
Andiamo ora ad esplicitare gli strumenti necessari per effettuare ottimizzazione su questa varietà, le formule riportate di seguito, tra cui la derivata della funzione distanza, sono state riprese da [1]

3.1.1 Metrica

La metrica, definita nel *Fibrato Tangente* dell'iperboloide, è la metrica indotta dal prodotto di Minkowski già esplicitato.

3.1.2 Distanza

La distanza tra due punti dell'iperboloide rappresenta la lunghezza della geodedica che unisce i due punti, si può dimostrare che vale.

$$d_{\mathbb{H}^n}(u,v) = \operatorname{arccosh}(-\langle u,v\rangle_{n:1}); \quad u,v \in \mathbb{H}^n.$$

3.1.3 Mappa Esponenziale

La mappa esponenziale definita su $p \in \mathbb{H}^n$ applicata al vettore $v \in T_p\mathbb{H}^n$. É.

$$Exp_p(v) = cosh(\|v\|_{n:1})p + sinh(\|v\|_{n:1})\frac{v}{\|v\|_{n:1}}.$$

3.1.4 Gradiente della funzione distanza

La definizione dell'iperboloide e dello spazio di Minkowski nel quale è costruito ci aiuta nel calcolo del gradiente della funzione distanza.
Data una funzione costo $E : \mathbb{H}^n \to \mathbb{R}$ differenziabile, possiamo calcolarne il gradiente in un punto $p \in \mathbb{H}^n$, passando dapprima per il gradiente definito nell'ambiente, che nel nostro caso è lo spazio di Minkowski. Ricordiamo che quest'ultimo è cosi definito.

$$\nabla_p^{\mathbb{R}^{n:1}} E = \left(\frac{\partial E}{\partial x_1}|_{x=p}, ..., \frac{\partial E}{\partial x_n}|_{x=p}, -\frac{\partial E}{\partial x_{n+1}}|_{x=p} \right).$$

Una volta calcolato il gradiente nell'ambiente procediamo a proiettarlo nello spazio tangente a $\mathbb{H}^n$ nel punto $p \in \mathbb{H}^n$ come segue.

$$\nabla_p^{\mathbb{H}^n} E = \nabla_p^{\mathbb{R}^{n:1}} E + \langle p, \nabla_p^{\mathbb{R}^{n:1}} E\rangle_{n:1} p.$$

Relativamente al problema della media di Fréchet, che andremo ad esplicitare, si ha la seguente formula la quale rappresenta il gradiente su spazio di Minkowski della funzione distanza.

$$\nabla_u^{\mathbb{R}^{n:1}} d_{\mathbb{H}^n}(u,v) = -(\langle u,v\rangle_{n:1}^2 - 1)^{-1/2} v.$$

3.2 Disco di Poincaré

Andiamo ora a definire l'altro modello dello spazio iperbolico definito disco di Poincaré, definito come l'insieme dei punti di norma strettamente minore di 1 in $\mathbb{R}^n$.
Di seguito ne riportiamo una definizione formale.
Ricordiamo che la norma $\|\cdot\|$ senza pedice rappresenta la norma euclidea.

Definizione 7 (Disco di Poincaré) *Il modello del disco di Poincaré è definito nel modo seguente,*

$$\mathbb{D}^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}.$$

Essendo $\mathbb{D}^n$ un sottoinsieme di $\mathbb{R}^n$ lo spazio tangente in ogni punto di $\mathbb{D}^n$ è $\mathbb{R}^n$, è possibile definire il prodotto scalare $\langle \cdot, \cdot \rangle_x$, esplicitato come metrica nella sottosezione successiva, tale prodotto porta il disco di Poincaré ad essere una **varietà riemanniana**.
Come fatto per l'iperboloide andiamo ad esplicitare i vari strumenti necessari per effettuare ottimizzazione. Le formule riportate sono state riprese da [3], [4], mentre per il calcolo del gradiente della funzione distanza abbiamo fatto riferimento a [5].

3.2.1 Metrica

La metrica, definita nello spazio tangente del disco di poincaré $T_p\mathbb{D}^n$ rispetto ad un punto $p \in \mathbb{D}^n$, è la seguente.

$$g^{\mathbb{D}} = \lambda_p^2 g^E; \text{ con } \lambda_p = \frac{2}{1 - \|p\|^2} \text{ e con } g^E \text{ la metrica euclidea.}$$

3.2.2 Distanza

La distanza tra due punti del disco di Poincaré rappresenta la lunghezza della geodedica che unisce i due punti, è definita come segue.

$$d_{\mathbb{D}^n}(u, v) = \operatorname{arccosh}\left(1 + \frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)}\right)$$

3.2.3 Mappa Esponenziale

La mappa esponenziale definita su $p \in \mathbb{D}^n$ e $v \in T_p\mathbb{D}^n$. E' definita come segue.

$$\begin{aligned}\exp_p(v) = &\frac{\lambda_p(\cosh(\lambda_p\|v\|) + p'\frac{v}{\|v\|}\sinh(\lambda_p\|v\|))p}{1 + (\lambda_p - 1)\cosh(\lambda_p\|v\|) + \lambda_p p'\frac{v}{\|v\|}\sinh(\lambda_p\|v\|)} \\ &+ \frac{\frac{1}{\|v\|}\sinh(\lambda_p\|v\|)v}{1 + (\lambda_p - 1)\cosh(\lambda_p\|v\|) + \lambda_p p'\frac{v}{\|v\|}\sinh(\lambda_p\|v\|)}\end{aligned}$$

3.2.4 Gradiente della funzione distanza

Dati $u, v \in \mathbb{D}^n$ la formula del gradiente riemanniano della funzione distanza è dato da.

$$\nabla_u^E d_{\mathbb{D}^n}(u, v) = \frac{4}{b\sqrt{c^2 - 1}}\left(\frac{(\|v\|^2 - 2\langle u, v\rangle + 1)u}{a^2} - \frac{v}{a}\right)$$

$$\text{con } a = 1 - \|u\|^2, \quad b = 1 - \|v\|^2, \quad c = 1 + \frac{2}{ab}\|u - v\|^2$$

Sappiamo che per per ricavare il gradiente riemanniano dal gradiente euclideo è sufficiente riscalare ∇^E per l'inverso della metrica riemanniana, perciò,

$$\nabla_u^{\mathbb{D}^n} d_{\mathbb{D}^n}(u,v) = \frac{\nabla_u^E d_{\mathbb{D}^n}(u,v)}{\lambda_u^2}.$$

3.3 Iperboloide come modello conforme

Come inizialmente accennato le due varietà trattate finora sono due modelli conformi dello spazio iperbolico ovvero, esiste un diffeomorfismo invertibile conforme (mantiene gli angoli ma non le lunghezze) che porta da una varietà all'altra. Il diffeomorfismo in questione è la proieizone stereografica ρ definita come segue.

$$\rho : \mathbb{H}^n \to \mathbb{D}^n \,|\, x \to \frac{1}{x_{n+1}+1}(x_1, ..., x_n),$$
$$\rho^{-1} : \mathbb{D}^n \to \mathbb{H}^n \,|\, y \to \frac{2}{1-r}(y_1, ..., \frac{1+r}{2}) \text{ ; con } r = \|y\|^2,$$

A questo punto facciamo vedere che vale la seguente uguaglianza, che sarà utile nella sezione successiva.

$$d_{\mathbb{D}^n}(a,b) = d_{\mathbb{H}^n}(\rho^{-1}(a), \rho^{-1}(b))$$

Dimostreremo che l'uguaglianza è valida, dando dapprima una dimostrazione per $n = 2$, e poi la generalizzeremo per n qualsiasi.

Caso $n = 2$, siano $a, b \in \mathbb{D}^2$.

$$a = (a_1, a_2),$$
$$b = (b_1, b_2),$$

Applicando l'inversa della proiezione stereografica si ottiene,

$$\rho^{-1}(a) = \frac{2}{1-r_a}(a_1, a_2, \frac{1+r_a}{2})$$

$$\rho^{-1}(b) = \frac{2}{1-r_b}(b_1, b_2, \frac{1+r_b}{2})$$

con $r_a = (a_1^2 + a_2^2)$ e $r_b = (b_1^2 + b_2^2)$

A questo punto con semplici passaggi facciamo vedere che vale la seguente.

$$\begin{aligned}
d_{\mathbb{H}^2}(\rho^{-1}(a),\rho^{-1}(b)) &= \operatorname{arccosh}(-\langle(\frac{2a_1}{1-r_a},\frac{2a_1}{1-r_a},\frac{1+r_a}{1-r_a}),(\frac{2b_1}{1-r_b},\frac{2b_1}{1-r_b},\frac{1+r_b}{1-r_b})\rangle_{2:1}) \\
&= \operatorname{arccosh}(-(\frac{4a_1b_1}{(1-r_a)(1-r_b)}+\frac{4a_2b_2}{(1-r_a)(1-r_b)}-\frac{(1+r_a)(1+r_b)}{(1-r_a)(1-r_b)}) \\
&= \operatorname{arccosh}(\frac{(1+r_a)(1+r_b)-4a_1b_1-4a_2b_2}{(1-r_a)(1-r_b)}) \\
&= \operatorname{arccosh}(\frac{1+r_a+r_b+r_ar_b-4a_1b_1-4a_2b_2}{1-r_a-r_b+r_ar_b}) \\
&= \operatorname{arccosh}(\frac{1-r_a-r_b+2r_a+2r_b+r_ar_b-4a_1b_1-4a_2b_2}{1-r_a-r_b+r_ar_b}) \\
&= \operatorname{arccosh}(1+\frac{2r_a+2r_b-4a_1b_1-4a_2b_2}{(1-r_a)(1-r_b)}) \\
&= \operatorname{arccosh}(1+2\frac{r_a+r_b-2a_1b_1-2a_2b_2}{(1-r_a)(1-r_b)}) \\
&= \operatorname{arccosh}(1+2\frac{\|a-b\|^2}{(1-\|a\|^2)(1-\|b\|^2)}) \\
&= d_{\mathbb{D}^2}(a,b).
\end{aligned}$$

Facciamo vedere ora che l'uguaglianza vale anche nel caso in cui n è generico, siano $a,b\in\mathbb{D}^n$. Procediamo facendo vedere che i due termini dell'uguaglianza rappresentano la stessa quantità.

$$d_{\mathbb{D}^n}(a,b)=\operatorname{arccosh}(1+2\frac{\|a-b\|^2}{(1-\|a\|^2)(1-\|b\|^2)})$$

$$d_{\mathbb{H}^n}(\rho^{-1}(a),\rho^{-1}(b))=\operatorname{arccosh}(-\langle\rho^{-1}(a),\rho^{-1}(b)\rangle_{n:1})$$

Facciamo vedere ora, che gli argomenti della funzione $\operatorname{arccosh}$, nelle due equazioni riportate, rappresentano in realtà lo stesso valore.

$$\begin{aligned}
-\langle\rho^{-1}(a),\rho^{-1}(b)\rangle_{n:1} &= -\langle\frac{2}{1-\|a\|^2}\begin{bmatrix}a\\ \frac{1+\|a\|^2}{2}\end{bmatrix},\frac{2}{1-\|b\|^2}\begin{bmatrix}b\\ \frac{1+\|b\|^2}{2}\end{bmatrix}\rangle_{n:1} \\
&= \frac{4}{(1-\|a\|^2)(1-\|b\|^2)}\left(a'b-\frac{(1+\|a\|^2)(1+\|b\|^2)}{4}\right) \\
&= \frac{-4a'b+1+\|a\|^2+\|b\|^2+\|a\|^2\|b\|^2}{(1-\|a\|^2)(1-\|b\|^2)} \\
&= \frac{-4a'b+1+2\|a\|^2+2\|b\|^2-\|a\|^2-\|b\|^2+\|a\|^2\|b\|^2}{(1-\|a\|^2)(1-\|b\|^2)} \\
&= 1+2\frac{\|a-b\|^2}{(1-\|a\|^2)(1-\|b\|^2)}
\end{aligned}$$

3.4 Media di Fréchet e problema del centroide

In tale sezione formalizzeremo il concetto di **media di Fréchet** di p punti in una varietà differenziabile M, daremo inoltre un'importante risultato. Partiamo dapprima dalla seguente definizione.

Definizione 8 (Media di Fréchet) *Dati $x^{(1)}, ..., x^{(p)} \in M$, la media di Fréchet di $x^{(1)}, ..., x^{(p)}$ è l'argomento di minimo della funzione,*

$$f(\Theta) = \frac{1}{p} \sum_{i=1}^{p} d^2(\Theta, x^{(i)}).$$

Il problema di trovare la media di Fréchet (centroide) di p-punti consiste nel trovare il punto che minimizza la funzione **somma delle distanze al quadrato** definita a partire dai punti in questione, il cui minimo esiste ed è unico nella varietà disco di Poincaré. Tale problema può essere risolto direttamente in uno spazio euclideo, nel quale il minimo è dato dalla media aritmetica dei p-punti. Mentre nel disco di Poincaré non è nota una soluzione esplicita e perciò la soluzione viene approssimata applicando le metodologie di ottimizzazione su varietà differenziabili trattate nella sezione 2.5.
Per effettuare ottimizzazione abbiamo bisogno di calcolare il gradiente della funzione somma delle distanze al quadrato, avendo definito nella sezione precedente il gradiente della funzione distanza su disco e su iperboloide, possiamo calcolare il gradiente della funzione come segue,

$$\nabla f(\Theta) = \frac{2}{p} \sum_{i=1}^{p} d(\Theta, x^{(i)}) \cdot \nabla d(\Theta, x^{(i)}).$$

Diamo ora un importante risultato, la cui correttezza è dimostrata in funzione dell'uguaglianza $d_{\mathbb{D}^n}(a, b) = d_{\mathbb{H}^n}(\rho^{-1}(a), \rho^{-1}(b))$ che abbiamo verificato nella sezione precedente.
Dati $a^{(i)}, \Theta \in \mathbb{D}^n$; con $i = 1...p$ dall'uguaglianza precedente è facile notare che $\frac{1}{p}\sum_{i=1}^{p} d^2_{\mathbb{D}^n}(\Theta, a^{(i)}) = \frac{1}{p}\sum_{i=1}^{p} d^2_{\mathbb{H}^n}(\rho^{-1}(\Theta), \rho^{-1}(a^{(i)}))$, perciò essendo ρ un diffeomorfismo se esiste $\Theta \in \mathbb{D}^n$ per cui $\frac{1}{p}\sum_{i=1}^{p} d^2_{\mathbb{D}^n}(\Theta, a^{(i)})$ è minima allora esiste $\Psi \in \mathbb{H}^n$ per cui è minima $\frac{1}{p}\sum_{i=1}^{p} d^2_{\mathbb{H}^n}(\Psi, \rho^{-1}(a^{(i)}))$. Essendo inoltre $f(x) = \hat{f}(\rho^{-1}(x))$ allora cercare il minimo di f equivale a cercare il minimo di $\hat{f} \circ \rho^{-1}$.
In funzione di tale dimostrazione possiamo trasportare il problema definito su disco di Poincaré nella varieta iperboloide e viceversa.

4 Esperimenti numerici e implementazioni

In questa sezione andremo a definire lo pseudocodice degli algoritmi di ottimizzazione che abbiamo utilizzato nei nostri esperimenti.
Confronteremo tali strumenti per il calcolo della media di Fréchet sui modelli dello spazio iperbolico descritti nella sezione precedente, ovvero disco di Poincaré ed iperboloide, sfruttando il fatto che questi due modelli sono tra di loro conformi, ovvero che esiste una mappa conforme che li lega.
In tale sezione cambieremo la notazione con cui faremo riferimento ai vettori generati in una particolare iterazione.
Fin'ora abbiamo indicato con x^k il vettore x generato al passo k da uno specifico algoritmo, nello pseudocodice e nelle successive digressioni tale vettore verrà indicato con x_k.
Indicheremo inoltre, con $\|\cdot\|$ (senza pedice indicante un punto nella varietà) la norma euclidea di un vettore.

4.1 Algoritmi e Pseudocodice

Descriviamo tre algoritmi di ottimizzazione, di cui due sono algoritmi di primo ordine, discesa del gradiente con **passo fisso** e con la regola di **Armijo**, ovvero che utilizzano informazioni solamente sul gradiente nel punto per decidere il passo e la direzione dell'iterazione, mentre il terzo, **Barzilai-Borwein**, è un algoritmo di quasi Newton e perciò ha accesso ad informazioni di approssimazione al secondo ordine, sfruttando l'equazione delle secanti. Tutti gli algoritmi si basano sul calcolo del gradiente della funzione da minimizzare.

4.1.1 Metodo di discesa del gradiente a passo fisso

L'algoritmo a Passo fisso è un algoritmo del primo ordine molto elementare, la direzione scelta è l'opposta della direzione del gradiente, tale algoritmo, implementato su una varietà differenziabile, necessita solamente del gradiente riemanniano della funzione da minimizzare, di una retrazione e di un parametro α che denota il passo fisso applicato ad ogni iterazione.

Algoritmo 2: Algoritmo a passo fisso

Input: Varietà M; funzione costo f differenziabile definita su M; retrazione R; scalare $\alpha > 0$; punto iniziale $x_0 \in M$, condizione di arresto ε

Output: Sequenza $\{x_k\}$

for $k = 0, 1, 2, ...$ **do**

  $g_k \leftarrow \nabla^{(R)} f(x_k)$;

  if $\|g_k\| < \varepsilon$ **then**

    return

  end

  $d_k \leftarrow -\alpha g_k$;

  $x_{k+1} \leftarrow R_{x_k}(d_k)$;

end

4.1.2 Armijo

L'algoritmo di Armijo è anche esso un algoritmo del primo ordine, la direzione scelta è l'opposta della direzione del gradiente, per la selezione del passo si basa sul calcolo del punto di Armijo riemanniano, descritto nella sezione 2.5.2, per poter essere eseguito necessita, anche esso, del gradiente riemanniano della funzione da minimizzare, di una retrazione e di tre parametri $\gamma \in (0, 1)$, $\sigma \in (0, 1)$ e $\lambda > 0$. Solitamente il parametro λ deve essere calcolato ad ogni iterazione, necessario per effettuare il giusto rescaling di d_k, per semplicità di sperimentazione abbiamo deciso di assumere tale parametro fisso. L'implementazione riemanniana di tale algoritmo fa riferimento alle formule in [2].

Algoritmo 3: Algoritmo con metodo di Armijo

Input: Varietà M; funzione costo f differenziabile definita su M; retrazione R; scalari $\gamma \in (0, 1)$, $\sigma \in (0, 1)$, $\lambda > 0$; punto iniziale $x_0 \in M$, condizione di arresto ε

Output: Sequenza $\{x_k\}$

for $k = 0, 1, 2, ...$ **do**

  $g_k \leftarrow \nabla^{(R)} f(x_k)$;

  if $\|g_k\| < \varepsilon$ **then**

    return

  end

  Trovare il più piccolo $h = 0, 1, ...$, tale per cui

$$f(R_{x_k}(-\sigma^h \lambda g_k)) \leq f(x_k) - \gamma \sigma^h \lambda \langle g_k, g_k \rangle_{x_k};$$

  $x_{k+1} \leftarrow R_{x_k}(-\sigma^h \lambda g_k)$;

end

4.1.3 Metodo di Barzilai-Borwein

L'algoritmo di Barzilai-Borwein è un metodo quasi Newton e si basa sulla soluzione, per $k \geq 1$, di un problema dei minimi quadrati, che ci permette di avere ricavare informazioni sulla matrice hessiana, quindi del secondo ordine. Tale algoritmo è stato implementato in [2]. Andiamo ora a descrive l'algoritmo dapprima nella sua versione euclidea in modo tale da avere ben chiare le motivazioni teoriche, dopo di che descriveremo la versione riemanniana.
L'algoritmo si basa sulla risoluzione del seguente problema di minimo,

$$\min_t \|s_k t - y_k\|,$$

con $s_k = x_{k+1} - x_k$ ed $y_k = \nabla^{(R)} f(x_{k+1}) - \nabla^{(R)} f(x_k)$, il quale, assumendo che $x_{k+1} \neq x_k$ ha un'unica soluzione $t = \frac{s_k' y_k}{s_k' s_k}$. Perciò quando $s_k' y_k > 0$ il passo viene scelto come.

$$a_{k+1}^{BB} = \frac{s_k' s_k}{s_k' y_k}.$$

Per tradurre lalgoritmo dalla sua definizione su spazio euclideo in una versione compatibile con ottimizzazione su varietà differenziabili dobbiamo andare a ridefinire s_k ed y_k. Al passo $k+1$. L'hessiana è una mappa bilineare da $T_{x_{k+1}} M$ a $T_{x_{k+1}} M$, e siccome g_{k+1} e g_k (con $g_k = \nabla^{(R)} f(x_k)$) appartengono a due differenti spazi tangenti, considereremo il vettore $-\alpha_k g_k \in T_{x_k} M$ e lo trasporteremo in $T_{x_{k+1}} M$, quindi,

$$s_k = \mathcal{T}_{x_k \to x_{k+1}}(-\alpha_k g_k),$$

Da cui,

$$y_k = g_{k+1} - \mathcal{T}_{x_k \to x_{k+1}}(g_k).$$

Mentre per quanto riguarda il prodotto scalare utilizzeremo quello riemanniano, quindi

$$a_{k+1}^{BB} = \frac{\langle s_k, s_k \rangle_{x_{k+1}}}{\langle s_k, y_k \rangle_{x_{k+1}}}.$$

Per quanto riguarda l'operatore di trasporto vettoriale definito per le varietà disco di Poincaré ed iperboloide sono state riprese le implementazioni presenti in [6].

Algoritmo 4: Barzilai-Borwein Algorithm

Input: Varietà M; funzione costo f differenziabile definita su M; retrazione R trasporto vettoriale $\mathcal{T}$; scalari $\alpha_{\max} > \alpha_{\min} > 0$, $\alpha_0^{BB} \in [\alpha_{\min}, \alpha_{\max}]$; punto di partenza $x_0 \in M$

Output: Sequenza $\{x_k\}$

$g_0 \leftarrow \nabla^{(R)} f(x_0)$.

for $k = 0, 1, 2, \ldots$ **do**

  if $\|g_k\| < \varepsilon$ **then**

    return

  end

  $x_{k+1} \leftarrow R_{x_k}(-a_k^{BB} g_k)$;

  $f_{k+1} \leftarrow f(x_{k+1})$; $g_{k+1} \leftarrow \nabla^{(R)} f(x_{k+1})$;

  $s_k \leftarrow \mathcal{T}_{x_k \to x_{k+1}}(-\alpha_{k+1}^{BB} g_k)$;

  $y_k \leftarrow g_{k+1} - \mathcal{T}_{x_k \to x_{k+1}}(g_k)$;

  $\tau_{k+1} \leftarrow \frac{\langle s_k, s_k \rangle_{x_{k+1}}}{\langle s_k, y_k \rangle_{x_{k+1}}}$;

$$\begin{cases} \alpha_{k+1}^{BB} = \min\{\alpha_{\max}, \max\{\tau_{k+1}, \alpha_{\min}\}\} & se \langle s_k, y_k \rangle_{x_{k+1}} > 0; \\ \alpha_{k+1}^{BB} = \alpha_{\max} \quad altrimenti; \end{cases}$$

end

Nella nostra implementazione dell'algoritmo **Riemannian Barzilai-Borwein** abbiamo deciso di non applicare valori di riscalamento, tramite ricerca lineare esatta o non, al valore α_k^{BB}, per motivazioni sperimentali, sarebbe stato computazionalmente più oneroso, alcuni esperimenti avrebbero richiesto più tempo per essere eseguiti, considerando inoltre che l'implementazione riportata si comporta mediamente bene nella pratica.

4.2 Esperimenti

Gli esperimenti effettuati ci permetteranno di confrontare il problema di ottimizzazione della funzione **somma delle distanze al quadrato** implementato sulla varietà disco di Poincaré e dell'equivalente istanza trasportatta su iperboloide.
Tali esperimenti saranno utili per capire se queste due varietà sono equivalenti ed interscambiabili, non solo nella rappresentazione dello spazio iperbolico ma anche come varietà su cui effettuare ottimizzazione, in quanto la mappa conforme che lega tali varietà mantiene l'esistenza del minimo della funzione costo in questione, come dimostrato nella sezione 3.3.
Gli esperimenti effettuati si suddividono in due parti che in seguito descriveremo.
Per effettuare gli esperimenti ci siamo muniti di un dataset di circa duecento istanze del problema definito su disco di Poincaré, prendendo un numero fissato comune a tutte le istanze, di punti casuali sul disco (sampling effettuato sfruttando la funzione *rand* del file *PoincareFactory* della libreria manopt [6]), la dimensione

delle varietà da cui vengono presi questi punti casuali è comune a tutte le istanze del dataset.
Gli esperimenti che riporteremo fanno riferimento al sampling effettuato dal disco di dimensione $n = 2$.
Per ogni istanza abbiamo calcolato il punto x_0 di partenza come la media aritmetica dei punti dell'istanza.
Abbiamo calcolato inoltre il punto limite tramite l'algoritmo a passo fisso con un α molto piccolo e come criterio di arresto abbiamo scelto la condizione che la norma euclidea del gradiente g_k scendesse sotto un valore soglia (e.g. 10^{-9}).
Prima di poter effettivamente confrontare gli algoritmi implementati nelle due varietà abbiamo selezionato i parametri, uno per disco ed uno per iperboloide, per gli algoritmi che hanno necessità di dover modificare alcuni parametri, come α per il passo fisso e λ per l'algoritmo di Armijo.
Per effetturare tale selezione abbiamo preso un sotto insieme del dataset iniziale e, sia per l'algoritmo a passo fisso che per Armijo.
Abbiamo scelto un valore che rappresenta il numero di elementi in cui si suddivide lo spazio dei possibili valori del parametro di cui si deve trovare il valore ottimale (e.g 100, partiremo da 0.01 ed arriveremo a 0.99 con passo pari a 0.01). Per ogni record del sotto dataset si è applicato l'algoritmo con un parametro incrementale creando, per ogni istanza, una sequenza i cui valori rappresentano il numero di passi che l'algoritmo con il parametro in questione impiega per convergere al limite, o divergere (numero di passi prima di arrivare ad avere la norma della distanza dal limite di 10^{-4} , se diverge il numero sarà il massimo delle iterazioni possibili).
Una volta calcolate queste sequenze ne abbiamo fatta una media aritmetica e come parametro ottimo abbiamo scelto il valore che minimizza la sequenza media.
Per quanto riguarda la scelta di α ottimo per l'algoritmo a passo fisso abbiamo notato sperimentalmente che le sequenze di convergenza definite per l'implementazione su disco e su iperboloide, hanno un andamento quasi identico ed anche il medesimo punto di minimo che indica il parametro ottimo.

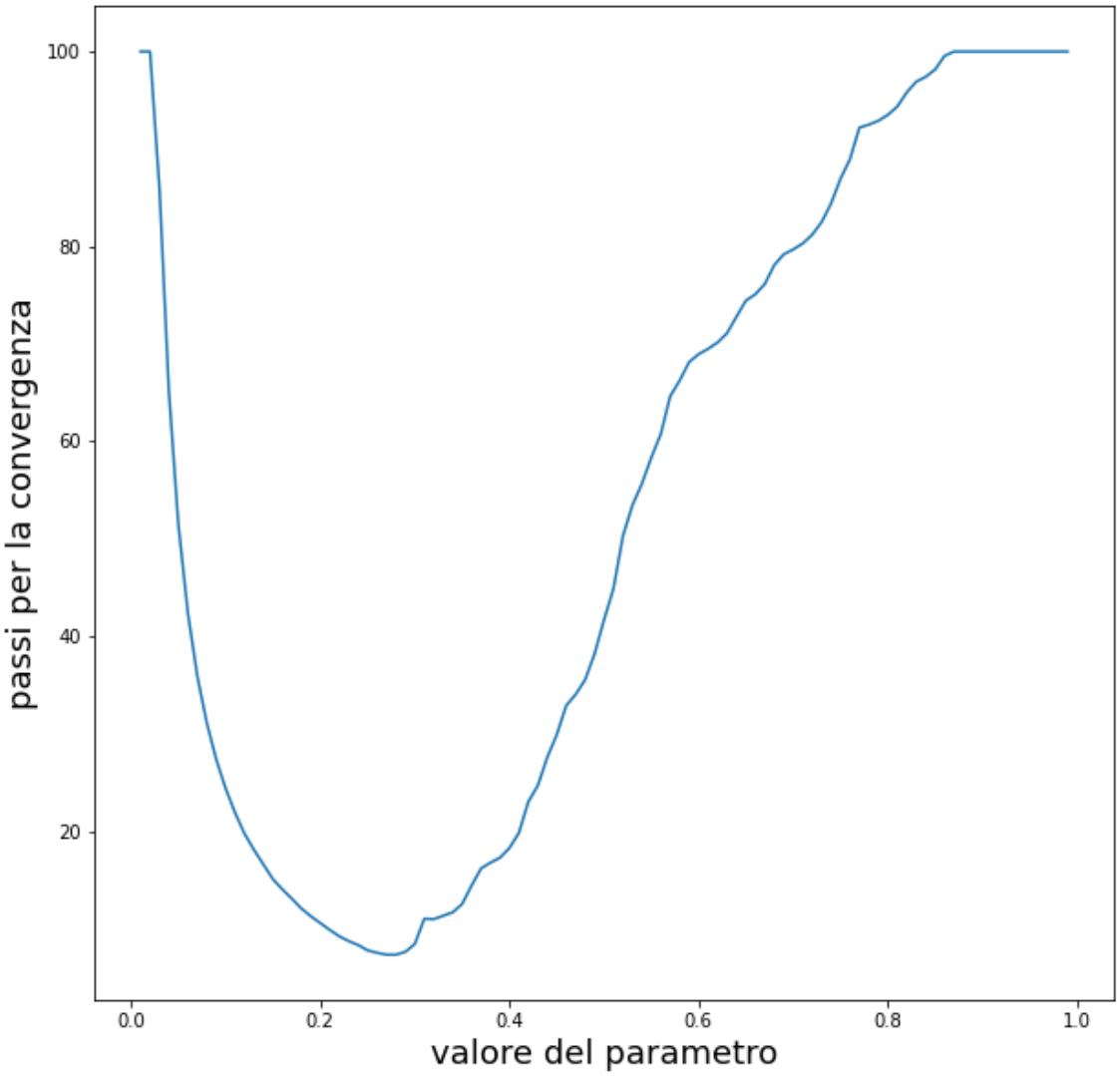

Figura 2: Sequenza relativa alla convergenza del'algoritmo a passo fisso implementato su disco. Punto di minimo 0.27, valore di convergeza minimo 7.35.

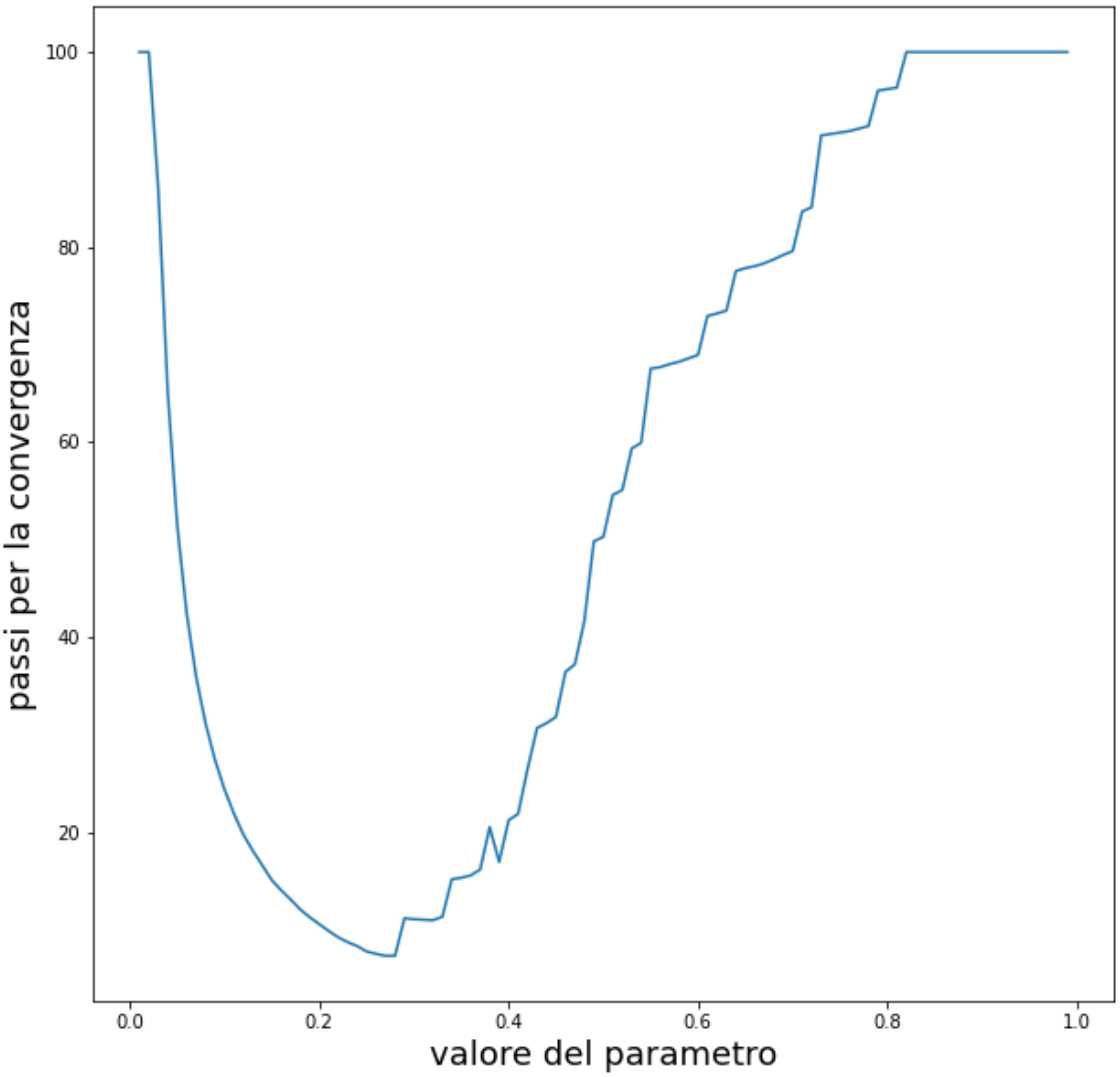

Figura 3: Sequenza relativa alla convergenza del'algoritmo a passo fisso implementato su iperboloide. Punto di minimo 0.27, valore di convergeza minimo 7.35.

Analogamente, per quanto riguarda le sequenze definite dalla scelta di λ per l'algoritmo di Armijo, si è notato un'andamento molto simile nelle due sequenze, anche in questo caso il punto di minimo coincide.

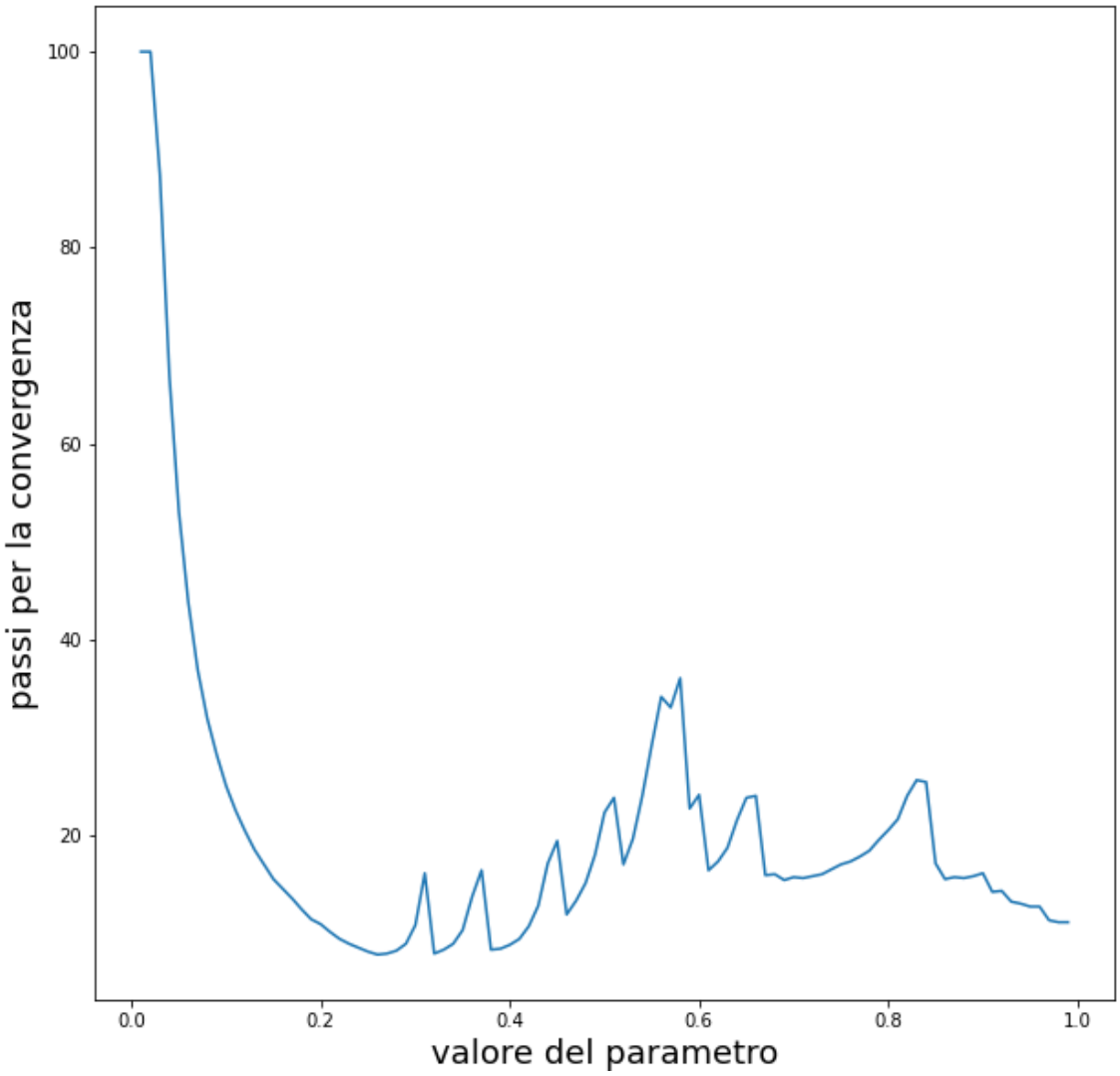

Figura 4: Sequenza relativa alla convergenza del'algoritmo di Armijo implementato su disco. Punto di minimo 0.26, valore di convergeza minimo 7.9.

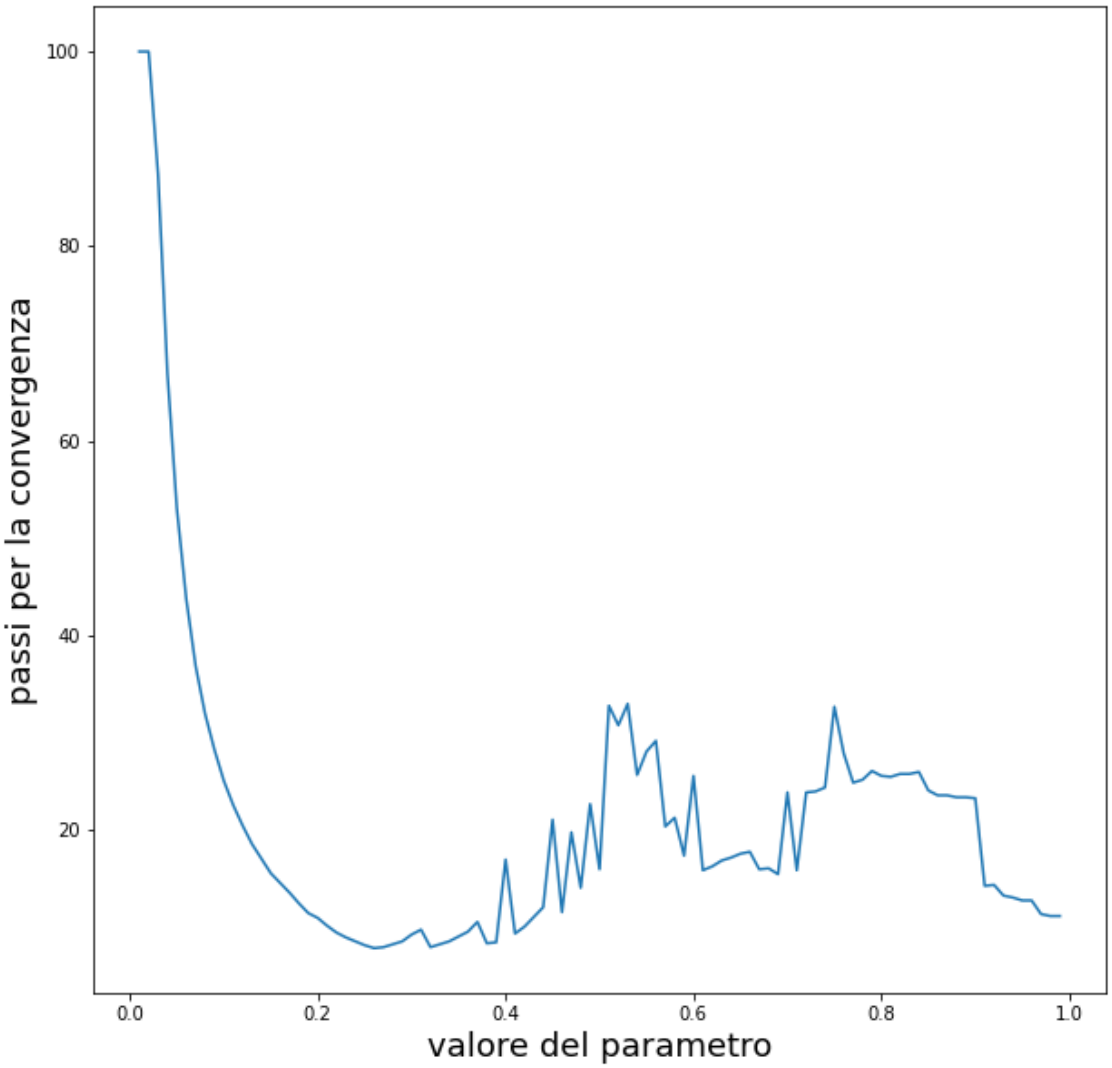

Figura 5: Sequenza relativa alla convergenza del'algoritmo di Armijo implementato su iperboloide. Punto di minimo 0.26, valore di convergeza minimo 7.9.

A questo punto, avendo scelto i parametri che permettono ai nostri algoritmi di performare al meglio, possiamo procedere con il confronto.
Riprendiamo ora il dataset totale.
Abbiamo eseguito tutti gli algoritmi su ogni istanza del dataset sia per l'implementazione su disco che per quella su iperboloide, creando un punto (a, b) in R^2 per ogni istanza, le cui coordinate sono a, il numero di passi per convergere al punto limite per l'implementazione su disco e b invece il numero di passi per convergere al punto limite per l'implementazione su iperboloide, creando cosi una nuvola di punti.
Tramite tale nuvola possiamo verificare se le metodologie sono effettivamente equivalenti.
In ogni grafico vengono costruite due rette di regressione su una versione ridotta

della nuvola di punti.
Toglieremo quei punti che hanno una molteplicità bassa (e.g. sotto $\frac{1}{40}$ della cardinalità totale del dataset), la prima retta di regressione viene calcolata tramite il metodo dei minimi quadrati, la seconda invece viene calcolata tramite un algoritmo meno sensibile ai valori anomali, la regressione di $Huber$ [1]. Noteremo quindi che se le rette di regressione tendono ad avere un coefficiente angolare "vicino" ad 1 i risultati ottenuti utilizzando le due varietà possono essere ritenuti equivalenti. Riportiamo di seguito i grafici dei nostri esperimenti, verrà riportato per ogni grafico un valore medio di convergenza per il disco e per l'iperboloide, tale valore rappresenterà un ulteriore indice di confronto.

[1] https://en.wikipedia.org/wiki/Huber_loss

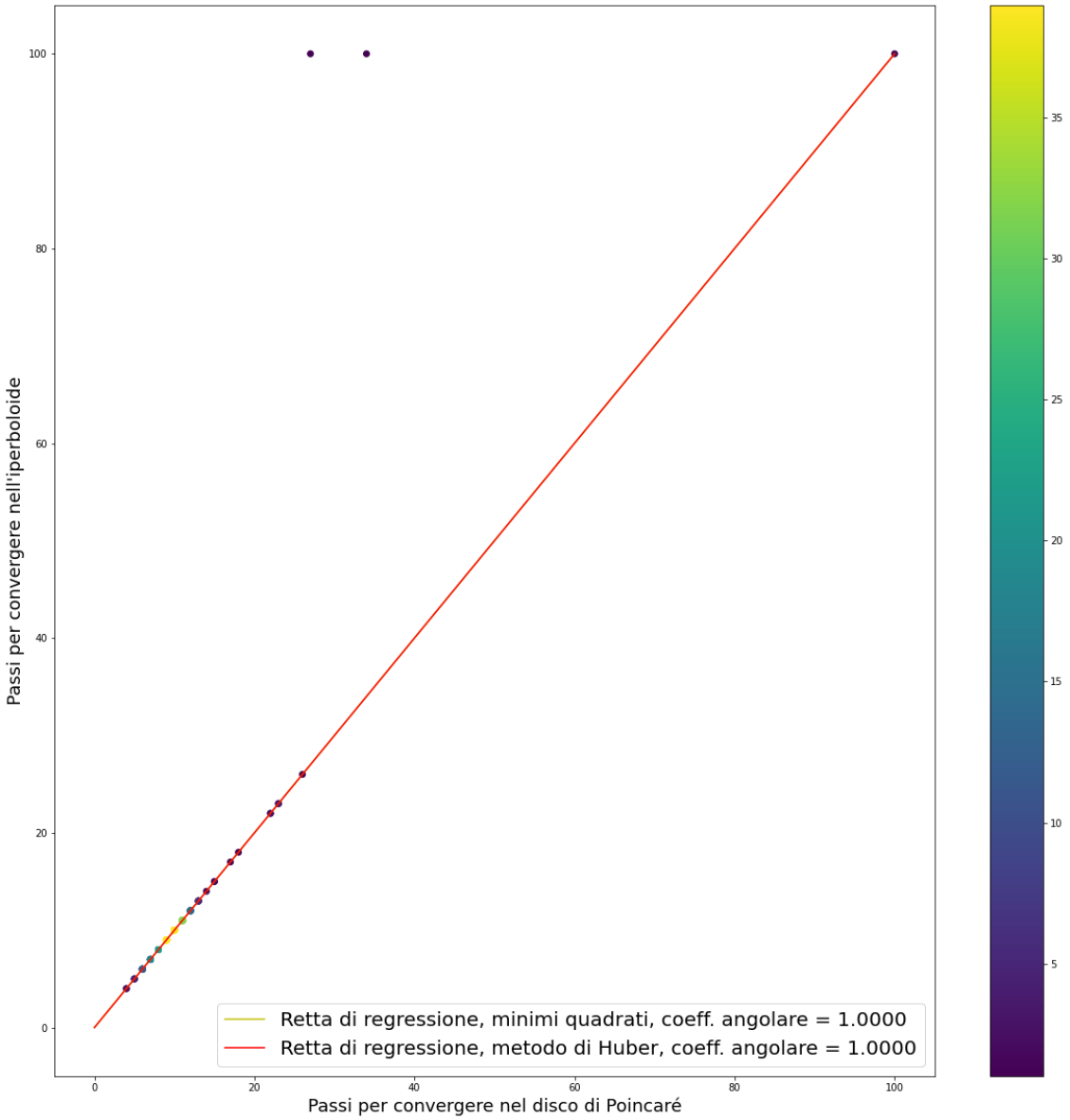

Figura 6: Nuvola di punti relativa all'algoritmo a passo fisso. Notiamo che i punti generati tendono a giacere sulla bisettrice e ciò indica un'equvalenza tra le due implementazioni. Valore medio di convergenza sul disco 10.5, valore medio di convergenza su iperboloide 11.2.

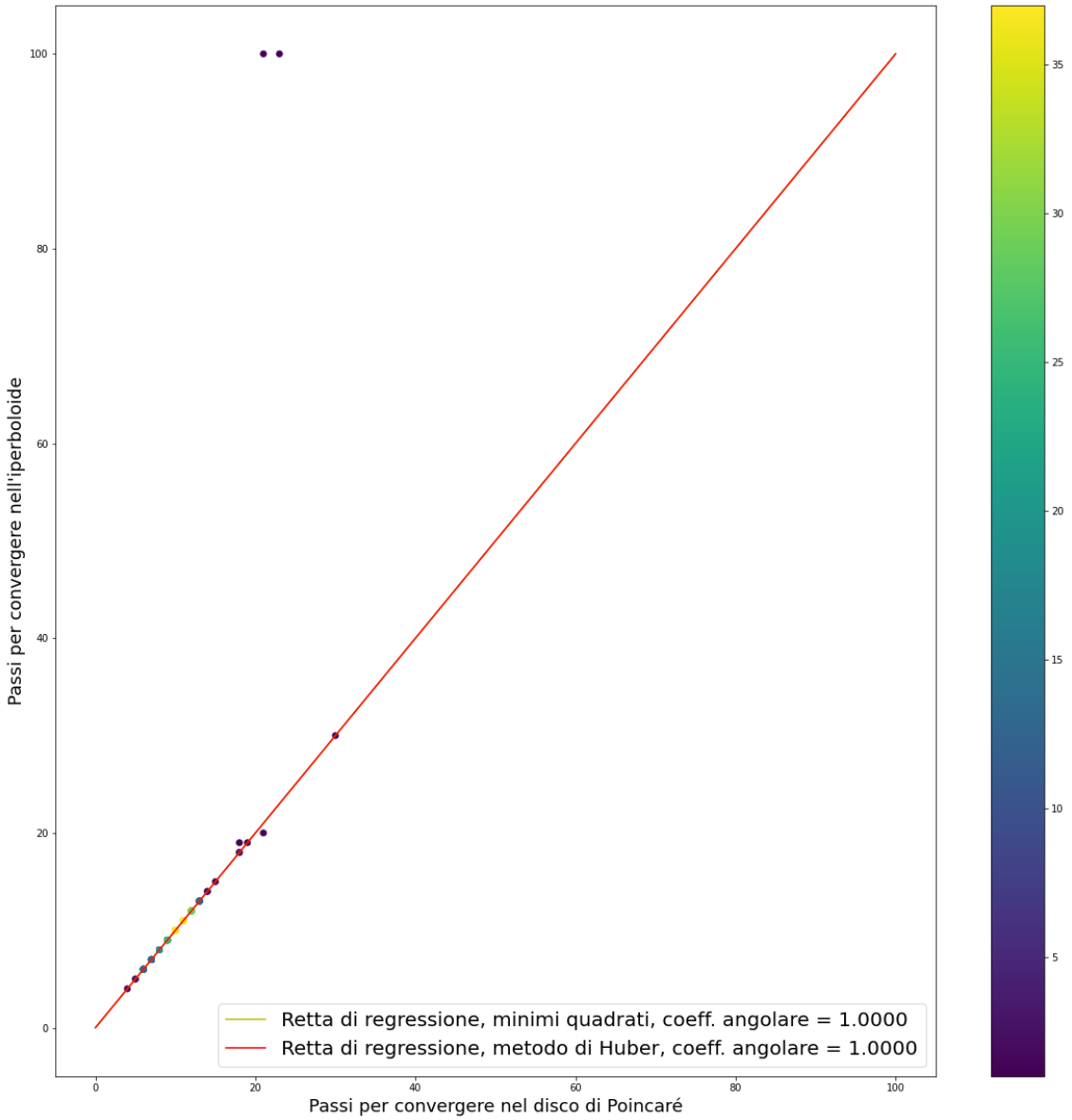

Figura 7: Nuvola di punti relativa all'algoritmo di Armijo. Notiamo che i punti generati tendono a giacere sulla bisettrice e ciò indica un equvalenza tra le due implementazioni. Valore medio di convergenza sul disco 10.4, valore medio di convergenza su iperboloide 11.2.

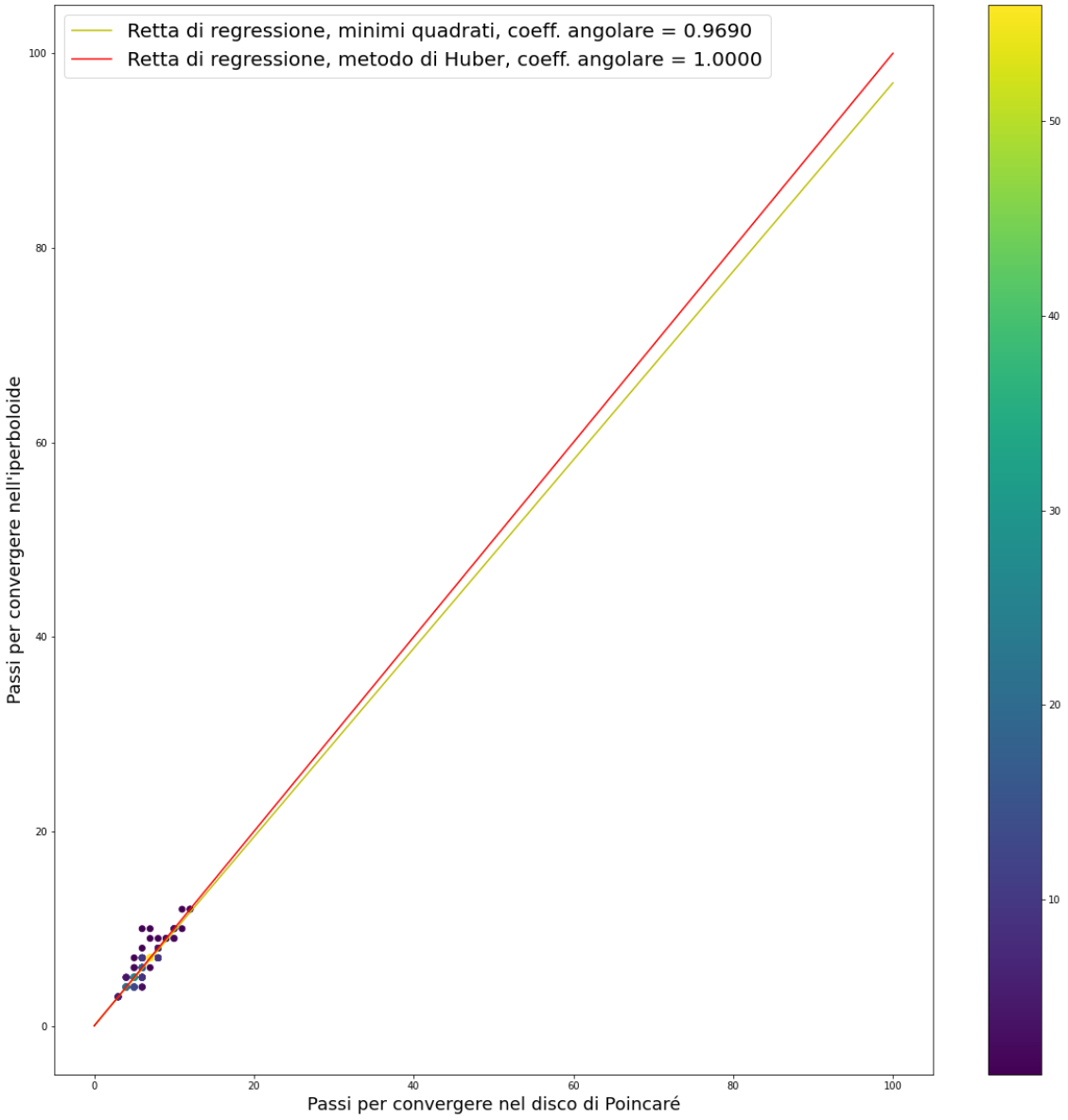

Figura 8: Nuvola di punti relativa all'algoritmo di Barzilai-Borwein. Notiamo un'equvalenza tra le due implementazioni. Valore medio di convergenza sul disco 6.13, valore medio di convergenza su iperboloide 6.8.

Durante le nostre sperimentazioni ci siamo resi conto che il grafico relativo all'algoritmo di Barzilai-Borwein tende ad avere i coefficienti delle rette di regressione "molto vicini" ad 1, ma non tutti i punti della nuvola tendono a giacere sulla bisettrice.
Igrafici relativi all'algoritmo di Armijo e l'algoritmo a passo fisso presentano una retta di regressione quasi equivalente alla bisettrice, e graficamente è facile vedere come tutti i punti, a meno di casi anomali, giacciono sulla bisettrice. Relativamente ai valori medi di convergenza, abbiamo valori abbastanza simili per tutti e tre gli algoritmi e possiamo notare una differenza di al più un passo e mezzo nella

convergenza media.
Possiamo asserire che le implementazioni su disco e su iperboloide, degli algoritmi ottimizzazzione per il calcolo della media di Fréchet sono pressoché equivalenti.
I risultati di tali sperimentazioni ci portano a congetturare che dato un problema di ottimizzazione su una varietà ,è possibile trasferire il problema da tale varietà ad una varietà conforme sempre tenendo presente che deve essere preservata l'esistenza di una soluzione ottima per la funzione costo.
Concludendo, un altro risultato delle nostre sperimentazioni è stato far vedere come l'algoritmo di Barzilai-Borwein si comporta meglio rispetto agli algoritmi a passo fisso e di ricerca lineare anche facendo riferimento al valore di convergenza medio delle implementazioni sulle due varietà riportate nelle descrizioni dei grafici. Riportiamo il link alla repository Github contenente i codici degli esperimenti: *https://github.com/Andrew-Wyn/Tesi_Ottimizzazione.*

5 Appendice

Richiamiamo qui alcuni concetti teorici necessari per la piena comprensione del presente elaborato.

5.1 Varietà

Sia M un insieme. Una bigezione ψ differenziabile di un sottoinsime U di M in un sottoinsieme aperto di $\mathbb{R}^d$ è chiamata mappa d-dimensionale dell'insieme U che definisce una *carta* (U, ψ). Data una carta (U, ψ) e $x \in U$ gli elementi di $\psi(x) \in \mathbb{R}^d$ sono le coordinate di x nella carta (U, ψ). Un *Atlante* (A) di M in R^d è un insieme di carte (U_a, ψ_a) dell'insieme M, dove a varia in un insieme di indici $\mathcal{A}$, tali che:

- $\bigcup_a U_a = M$, per $a \in \mathcal{A}$.
- Per ogni coppia $\alpha, \beta \in \mathcal{A}$ con $U_\alpha \cap U_\beta \neq \emptyset$ gli insiemi $\psi_\alpha(U_\alpha \cap U_\beta)$ e $\psi_\beta(U_\alpha \cap U_\beta)$ sono sottoinsiemi aperti di R^d ed il cambio di coordinate $\psi_\alpha \circ \psi_\beta^{-1} : \mathbb{R}^d \to \mathbb{R}^d$ è differenziabile.

Dato un'atlante A, sia A^+ l'insieme delle carte (U, ψ) tali che $A \bigcup \{(U, \psi)\}$ è ancora un'atlante. A^+ è un'atlante massimale generato da A. Un'atlante massimale dell'insieme M è chiamato struttura differenziabile su M.

Definizione 9 *Una varietà differenziabile d-dimensionale è una coppia (M, A^+), dove M è un insieme e A^+ è un atlante massimale su M in R^d, tale che la topologia indotta da A^+ è di Hausdorff.*

Data una carta ψ su M, l'inversa ψ^{-1} è chiamata parametrizzazione locale di M.

5.1.1 Vettori Tangenti

Riprendendo il concetto di derivata direzionale possiamo generalizzarla per un funzione f definita su una varietà, sustituendo $t \mapsto (x + t\eta)$ con una curva, $\gamma : U \to M$, con $0 \in U$ ed $\gamma(0) = x$, attraverso $x \in M$ (punto di derivazione) questo porta ad una derivata direzionale ben definita $\frac{\partial f(\gamma(t))}{\partial t}|_{t=0}$. Formalmente, sia M una varietà e γ una mappa differenziabile $\gamma : U \to M|t \mapsto \gamma(t)$ detta curva in M, data, una funzione f a valori reali definita su M, la funzione $f \circ \gamma : t \mapsto f(\gamma(t))$ è una funzione differenziabile da $\mathbb{R}$ in $\mathbb{R}$ con una ben definita derivata.
Sia $x \in M$; $\gamma(0) = x$ allora indichiamo con $\mathcal{F}_x(M)$ l'insieme delle funzioni reali differenziabili definite in un intorno di x. Il mapping $\dot{\gamma}(0)$ da $\mathcal{F}_x(M)$ in $\mathbb{R}$ è definito come: $\dot{\gamma}(0)f = \frac{\partial f(\gamma(t))}{\partial t}|_{t=0}$ con $f \in F_x(M)$; é chiamato vettore tangente alla curva γ in $t = 0$.

Definizione 10 (Vettore Tangente) *Un vettore tangente ε_x ad una varietà M in un punto $x \in M$ è una mappa da $F_x(M)$ in R tale che esiste una curva γ a valori in M con $\gamma(0) = x$ e che soddisfa,*

$$\varepsilon_x f = \dot{\gamma}(0)f = \frac{\partial f(\gamma(t))}{\partial t}|_{t=0} \quad f \in F_x(M).$$

Tale curva γ realizza il vettore tangente ε_x.

Definizione 11 (Spazio Tangente) *Lo spazio tangente ad M in x denominato come T_xM è l'insieme di tutti i vettori tangenti ad M in x. Questo insieme ammette una struttura di spazio vettoriale come segue: date $\dot{\gamma}_1(0)$ e $\dot{\gamma}_2(0)$ in T_xM e a, b in R allora,*

$$(a\dot{\gamma}_1(0) + b\dot{\gamma}_2(0))f = a(\dot{\gamma}_1(0)f) + b(\dot{\gamma}_2(0)f).$$

Lo spazio tangente T_xM mette a disposizione un'approssimazione locale della varietà tramite uno spazio vettoriale. Perciò tramite le retrazioni definite nella sezione 2.5.1 possiamo effettuare ottimizzazione sullo spazio vettoriale T_xM.

Definizione 12 (Fibrato tangente) *Data una varietà M, sia TM l'insieme di tutti i vettori tangenti ad M.*

$$TM = \bigcup_{x \in M} T_xM.$$

Sia $P : M \to N$ una mappa differenziabile tra due varietà M ed N. sia ε_x un vettore tangente in un punto x di M, si può far vedere che $DP(x)[\varepsilon_x]$ da $F_{P(x)}(N)$ in $\mathbb{R}$ è definita da:

$$(DP(x)[\varepsilon_x])f = \varepsilon_x(f \circ P),$$

è un vettore tangente ad N in $P(x)$.
La mappa $DP(X) : T_xM \to T_{P(x)}N | \varepsilon_x \mapsto DP(x)[\varepsilon_x]$ è una mappa lineare chiamata differenziale di P in x.

5.1.2 Metriche distanze e gradienti riemanniani

I vettori tangenti ad una varietà generalizzano la nozione di derivata direzionale. Questo può essere fatto munendo lo spazio tangente T_xM con un prodotto interno $\langle \cdot, \cdot \rangle_x$ bilineare, definito positivo e simmetrico. Il prodotto interno induce una norma $\|\varepsilon_x\|_x = \sqrt{\langle \varepsilon_x, \varepsilon_x \rangle_x}$ su T_xM. Una varietà il cui spazio tangente è munito di un prodotto interno, che dipende in modo differenziabile da x, prende il nome di varietà riemanniana (M, g) con g metrica riemanniana.
La lunghezza di una curva $\gamma : [a, b] \to M$ in una varietà riemanniana (M, g) è definita da,

$$L(\gamma) = \int_a^b \sqrt{g(\dot{\gamma}(t), g(\dot{\gamma}(t)))}.$$

La distanza riemanniana in una varietà riemanniana connessa (M, g) è $\mathrm{dist} : M \times M \to \mathbb{R} \,|\, \mathrm{dist}(x, y) \mapsto \inf_{\gamma \in \Gamma}(L(\gamma))$, dove Γ è l'insieme di tutte le curve in M che connettono x ed y.
Sia f la solita funzione liscia definita su (M, g), il gradiente riemanniano di f in x è definito come l'elemento di T_xM che soddisfa,

$$\langle \nabla^{(R)} f(x), \varepsilon \rangle_x = Df(x)[\varepsilon].$$

Sia M una sottovarietà immersa di una varietà riemanniana $\widehat{M}$ siccome ogni spazio tangente T_xM può essere trattato come sottospazio di $T_x\widehat{M}$ la metrica $\widehat{g}$ di $\widehat{M}$ induce una metrica g su M, $g_x(\xi, \eta) = \widehat{g}_x(\xi, \eta)$ con $\xi, \eta \in T_xM$, perciò M è una sotto varietà riemanniana.

5.2 Topologia

Una topologia su un insieme X è una collezione T di sottoinsiemi di X chiamati insiemi aperti, tali che:

- X e $\emptyset$ appartengono a T.
- L'unione di elementi di qualsiasi sotto collezione di T è in T.
- L'intersezione di elementi di qualsiasi sotto collezione finita di T è in T.

Uno spazio topologico è una coppia (X, T) dove X è un insieme e T è una topologia su X.

Definizione 13 (Hausdorff) *Un insieme X è T_2 o di Hausdorff se ogni coppia di punti distinti di X ammettono intorni distinti.*

Se X è di Hausdorff allora ogni sequenza di punti di X converge ad al più un punto di X.

5.3 Miscellanea

Definizione 14 (Positività) *Sia $A \in \mathbb{R}^{n\times n}$ è definita positiva se è simmetrica e se $v'Av > 0$ se v è un vettore $\in \mathbb{R}^n \setminus \{0\}$.*

Si osservi che se $v = 0$ allora che $v'Av = 0$.

Definizione 15 (Convessità) *Una funzione $f : \Omega \to \mathbb{R}$ con $\Omega \subset V$, V (spazio vettoriale) ed Ω convesso, è definita convessa se: $(f((\lambda)v + (1-\lambda)w) \leq \lambda f(v) + (1-\lambda)f(w)$, per $\lambda \in [0;1]$ e $v, w \in \Omega$.*

Definizione 16 (Derivate Parziali) *Data una funzione $f : V \to \mathbb{R}$, V (spazio vettoriale), allora una derivata parziale su $x \in V$ è definita come:*

$$\frac{\partial f(x)}{\partial x_i} = \lim_{h\to 0} \frac{f(x + he_i) - f(x)}{h};$$

con x_i componente i-esima di x ed e_i i-esimo vettore della base canonica.

Se una funzione è differenziabile in x allora tutte le derivate parziali esistono in x.

Definizione 17 (Punto di accumulazione) *Diciamo che $x \in \mathbb{R}$ è un punto di accumulazione per un insieme $E \subset \mathbb{R}$ se comunque scelto un intorno $B(x, \varepsilon)$, risulta che $B(x, \varepsilon)$ contiene almeno un punto di E diverso da x.*

Definizione 18 (Gradiente) *Consideriamo una funzione f definita su un insieme aperto $A \subset \mathbb{R}^n$ e sia $x \in A$ se esistono in x le derivate parziali rispetto ad $\{x_i\}_{i=1,\dots,n}$ allora è possibile costruire un vettore che ha per componenti le derivate parziali, perciò il gradiente è definito come.*

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n}\right].$$

Inoltre per la formula del gradiente sappiamo che, dato un versore v esiste la derivata direzionale f_v in x e che $f_v(x) = \nabla f(x)'v$.

Riferimenti bibliografici

[1] Benjamin Wilson and Matthias Leimeister. Gradient descent in hyperbolic space. *arXiv preprint arXiv:1805.08207*, 2018.

[2] Bruno Iannazzo and Margherita Porcelli. The Riemannian Barzilai-Borwein method with nonmonotone line search and the matrix geometric mean computation. *IMA J. Numer. Anal.*, 38(1):495–517, 2018.

[3] Geoopt, python library for manifold optimization.

[4] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1646–1655. PMLR, 10–15 Jul 2018.

[5] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30:6338–6347, 2017.

[6] Manopt, python library for manifold optimization.