



**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:  
Prof. Bruno Iannazzo**

**Candidato:  
Luca Moroni**

**Matricola:  
311279**

---

**Anno Accademico 2020/2021**

# Contenuti

<b>1</b>	<b>Introduzione</b>	<b>3</b>
<b>2</b>	<b>Metodi di ottimizzazione su <math>R^n</math> e Varieta differenziabili</b>	<b>3</b>
2.1	Condizioni di ottimalità . . . . .	3
2.1.1	Ottimo locale ed ottimo globale . . . . .	3
2.1.2	Condizioni necessarie per l'ottimalità . . . . .	3
2.1.3	Il caso della convessità . . . . .	4
2.1.4	Condizioni Sufficienti per L'ottimalità . . . . .	4
2.2	Discesa del gradiente in $R^n$ . . . . .	5
2.2.1	Selezionare la direzione di discesa . . . . .	5
2.2.2	Selezione del passo . . . . .	7
2.3	Discesa del gradiente su varietà . . . . .	8
2.3.1	Retrazioni . . . . .	8
2.3.2	Metodi <i>line-search</i> . . . . .	9
2.3.3	Convergenza . . . . .	10
<b>3</b>	<b>Disco di poincare e Iperboloide</b>	<b>11</b>
3.1	Iperboloide . . . . .	11
3.1.1	Metrica . . . . .	11
3.1.2	Distanza . . . . .	11
3.1.3	Mappa Esponenziale . . . . .	11
3.1.4	Gradiente della funzione distanza . . . . .	11
3.2	Disco di Poincaré . . . . .	12
3.2.1	Metrica . . . . .	12
3.2.2	Distanza . . . . .	12
3.2.3	Mappa Esponenziale . . . . .	12
3.2.4	Gradiente della funzione distanza . . . . .	12
3.3	Iperboloide come modello conforme . . . . .	13
3.4	Media di Fréchet e problema del centroide . . . . .	13
<b>4</b>	<b>Metodi utilizzati con pseudocodice ed esperimenti</b>	<b>14</b>
4.1	Algoritmi e Pseudocodice . . . . .	14
4.1.1	Passo fisso . . . . .	14
4.1.2	Armijo . . . . .	14
4.1.3	Barzilai-Borwein . . . . .	14
4.2	Esperimenti . . . . .	14
<b>5</b>	<b>Appendice</b>	<b>14</b>
5.1	Varietà . . . . .	14
5.1.1	Vettori Tangenti . . . . .	15
5.1.2	Metriche distanze e e gradienti Riemanniani . . . . .	16
5.2	Topologia . . . . .	16

5.3	Definizioni Utili . . . . .	17
5.3.1	Teorema di Taylor . . . . .	17

# 1 Introduzione

Nella seguente trattazione verrà presentato il problema del minimo della funzione media di Frechet sulla varietà disco di Poincaré, verrà proposta una metodologia risolutiva che consiste nel trasportare il problema su una varietà conforme definita iperboloide, nella quale alcune formule per dimensione arbitraria sono computazionalmente meno onerose (e numericamente più stabili ?)

## 2 Metodi di ottimizzazione su $R^n$ e Varietà differenziabili

In tale sezione andremo ad esplicitare le metodologie generali per trovare un minimo non vincolato di una funzione costo  $f$  definita su una varietà differenziabile  $M$  (notiamo che  $R^n$  è un caso particolare di varietà differenziabile) con valori reali.

### 2.1 Condizioni di ottimalità

Sia  $f : R^n \rightarrow R$  non vincolata.

#### 2.1.1 Ottimo locale ed ottimo globale

Un vettore  $x^*$  è un minimo locale non vincolato per  $f$  se (informalmente) ha un valore in  $f$  non più grande di un suo intorno, più formalmente se esiste  $\epsilon > 0$  tale che

$$f(x^*) \leq f(x) \quad \forall x \quad \|x - x^*\| < \epsilon$$

Un vettore  $x^*$  è un minimo globale non vincolato rispetto ad  $f$  se ha un valore in  $f$  non più grande di ogni altro vettore, più formalmente è tale che

$$f(x^*) \leq f(x) \quad \forall x \in R^n$$

#### 2.1.2 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  $\Delta x$  è non negativa

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

perciò dovendo valere per  $\Delta x$  sia 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  $\Delta x$  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

$\nabla^2 f(x^*)$ : semidefinita positiva

**Proposizione 1 (Condizioni necessarie di ottimalità)** *Sia  $x^*$  un minimo locale non vincolato di  $f : R^n \rightarrow R$ , 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*

$\nabla^2 f(x^*)$ : semi-definita positiva

### 2.1.3 Il caso della convessità

Se la funzione  $f$  è convessa non ci sono distinzioni tra un minimo locale ed un minimo globale, tutti i punti di minimo locale sono punti di minimo globale. Perciò un fatto importante è che la condizione  $\nabla f(x^*) = 0$  è sufficiente per l'ottimalità, la dimostrazione è basata sulle proprietà di base della convessità della funzione  $f$ .

**Proposizione 2 (Funzioni Convesse)** *Sia  $f : R^n \rightarrow R$  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  $X$ .*

### 2.1.4 Condizioni Sufficienti per L'ottimalità

Non è difficile trovare esempi in cui le condizioni di ottimalità necessarie definite nella sezione precedente [ $\nabla f(x^*) = 0$  &  $\nabla^2 f(x^*) \geq 0$ ] portino ad identificare

punti non di minimo locale come ad esempio punti di sella o punti di massimo. Supponiamo di avere un vettore  $x^*$  che soddisfa le seguenti condizioni

- $\nabla f(x^*) = 0$
- $\nabla^2 f(x^*)$ : definita positiva

Allora abbiamo che per ogni  $\Delta x \neq 0$

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

Ciò implica che in  $x^*$  la variazione di  $f$  al secondo ordine data da un piccolo spostamento  $\Delta x$  è positiva, perciò  $f$  tende ad incrementare nell'intorno di  $x^*$  e ciò implica che 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  $f : R^n \rightarrow R$  doppiamente differenziabile con continuità in un insieme aperto  $S$ . Si supponga esistere  $x^* \in S$  che soddisfi le condizioni*

*$\nabla f(x^*) = 0$   $\nabla^2 f(x^*)$ : definita positiva*

*allora,  $x^*$  è un minimo locale non vincolato di  $f$ . In particolare esistono  $\gamma > 0$  e  $\epsilon > 0$  tali che*

$$f(x) \geq f(x^*) + \gamma/2 \|x - x^*\|^2, \quad \forall x \quad \text{con } \|x - x^*\| < \epsilon$$

## 2.2 Discesa del gradiente in $R^n$

Andiamo ora a definire nel dettaglio le principali metodologie computazionali di ottimizzazione non lineare su  $R^n$ . Tali metodologie saranno trattate con un occhio alle possibili applicazioni.

Consideriamo il problema di trovare il minimo non vincolato di una funzione  $f : R^n \rightarrow R$  la risoluzione analitica è infattibile, per problemi pratici, 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 si genera una sequenza  $x_1, x_2, \dots$  tali che  $f$  decresce ad ogni iterazione  $f(x^{k+1}) < f(x^k)$  e sperabilmente raggiungere un punto di minimo.

I principali algoritmi di discesa sono Metodi del Gradiente poichè basano le decisioni in base al gradiente della funzione  $f$ .

### 2.2.1 Selezionare la direzione di discesa

Sia  $v \in R^n - 0$ ,  $v^\perp$  è un iperpiano che divide lo spazio in due componenti connesse:

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

- $v'd = 0$  (iperpiano)

se  $v = \nabla f(x)$  ogni  $d$  tale che  $v'd > 0$  è una direzione di decrescita.

**Teorema 1** Sia  $f \in C^1(\Omega)$ ,  $\Omega$  aperto di  $R^n$  e sia  $x \in \Omega$  e  $\nabla f(x) \neq 0 \forall d \in R^n$  tale che  $\nabla f(x)'d < 0 \exists \alpha_0 > 0$  tale che  $f(x + \alpha d) < f(x)$  con  $\alpha \in (0, \alpha_0]$

Detto ciò possiamo esplitare una formula che definisce una classe di algoritmi

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

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

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

Ci sono varie possibilita nella scelta di  $d^k$  e di  $\alpha^k$ . 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 diretto che  $\nabla f(x^k)'d^k < 0$ .

A seconda della scelta di  $D^k$  abbiamo differenti metodologie applicabili.

#### Steepest Descent:

$$D^k = I, \quad k = 0, 1, \dots$$

#### Metodi di Newton:

$$D^k = (\nabla^2 f(x^k))^{-1}, \quad k = 0, 1, \dots$$

$\nabla^2 f(x^k)$  deve essere definita positiva, proprietà sempre verificata se  $f$  è convessa.

#### Steepest Descent Riscalata:

$D^k$  matrice diagonale.

#### Metodi di Schamsky:

$$D^k = (\nabla^2 f(x^0))^{-1}$$

#### Metodi di quasi Newton:

$$D^k = H(x^k) \approx \nabla^2 f(x^k)$$

I metodi di quasi newton sono utilizzati nella pratica dal momento in cui il calcolo della matrice hessiana è computazionalmente oneroso, in tal 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. 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 essiana, in particolare abbiamo che

$(\nabla f(x^{k+1}) - \nabla f(x^k)) \approx H(x^{k+1})(x^{k+1} - x^k)$  che sappiamo essere un uguaglianza nel caso di  $f$  quadratica. 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 metodo delle secanti (equazione delle secanti).

### 2.2.2 Selezione del passo

Ci sono numerose metodologie per la scelta del passo  $d^k$  nei metodi del gradiente, ne listiamo alcune.

#### Ricerca lineare esatta:

scegliamo  $\alpha^k$  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 più semplice da implementare in vari casi. definito un scalare  $s > 0$   $\alpha^k$  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  $\alpha^k$  in modo che garantisca una decrescita sufficiente a garantire la convergenza del metodo sotto opportune ipotesi. Il metodo di armijo è stato definito come segue. Siano fissi  $s, \beta$  e  $\sigma$  con  $0 < \beta < 1$ , e  $0 < \sigma < 1$  e sia  $\alpha^k = \beta^{m_k} s$ , dove  $m_k$  è il primo intero non negativo per cui vale

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

Solitamente si scelgono i valori come segue,  $\sigma \in [10^{-5}, 10^{-1}]$ , il fattore di riduzione  $\beta \in [1/10, 1/2]$ , possiamo scegliere  $s = 1$  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 sequenza di direzioni  $\{x^k\}_k$  è detta *limitata* se vale la seguente  $\limsup_{k \in K} \nabla f(x^k)' d^k < 0$

**Definizione 2** Una successione di direzioni  $\{x^k\}_k$ ,  $\{d^k\}_k$  è detta **gradient related** se per ogni sotto-successione di  $\{x^k\}_k$  che converge ad un punto non stazionario  $\{d^k\}_k$  è limitata. Perciò non sarà mai ortogonale al gradiente (sennò sarebbe stazionario)

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

#### Metodo a passo costante:

Si definisce un passo costante  $s > 0$  e si fissa  $\alpha^k = s$ ,  $k = 0, 1, \dots$



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  $\alpha^k \rightarrow 0$ . 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  $\sum_{n=1}^{\infty} \alpha^k = \infty$

## 2.3 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 ( $R^n$ ), possiamo approcciare le metodologie computazionali per trovare un minimo non vincolato di una funzione  $f : M \rightarrow R$  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.

### 2.3.1 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  $T_x M$  (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  $T_x M \quad \forall x \in M$  in  $M$  con le seguenti proprietà. Sia  $R_x$  la restrizione di  $R$  a  $T_x M$ .*

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

Ci riferiamo alla condizione  $DR_x(0_x) = id_{T_x M}$  come condizione di *rigidità locale*. Equivalentemente, per ogni vettore tangente  $\xi$  in  $T_x M$ , la curva  $\gamma_\xi : t \mapsto R_x(t\xi)$  soddisfa  $\gamma_\xi(0)' = \xi$ .

Oltre a trasformare elementi di  $T_x M$  in elementi di  $M$ , un secondo importante scopo della retrazione  $R_x$  è quello di trasformare una funzione costo ( $f : M \rightarrow R$ ) definita in un intorno di  $x \in M$  in una funzione costo definita sullo spazio vettoriale  $T_x M$ . 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  $T_x M$ . 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  $\text{grad } \hat{f}_x(0_x) = \text{grad } f(x)$ .

### 2.3.2 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  $\eta^k$  è in  $T_{x^k} M$  e  $t^k$  è uno scalare. Una volta scelta la retrazione  $R$  ci rimane da decidere la direzione  $\eta^k$  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  $\{\eta^k\}$ ,  $\eta^k \in T_{x^k} M$  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

$$\limsup_{k \rightarrow \infty, k \in K} \langle \text{grad } f(x^k), \eta^k \rangle < 0$$

**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  $\eta \in T_x M$ , 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  $f(x) - f(R_x(\beta^m \bar{\alpha} \eta)) \geq -\sigma \langle \text{grad } f(x), \beta^m \bar{\alpha} \eta \rangle_x$ . 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 però, nel contempo, delle restrizioni fondamentali che garantiscono la convergenza della sequenza di punti generati.

---

**Algorithm 1:** Accelerated Line Search (ALS)

---

**Input:** Manifold  $M$ ; funzione costo  $f$  differenziabile definita su  $M$ ;  
retraction  $R$ ; scalari  $\bar{\alpha} > 0$ ,  $c, \beta, \sigma \in (0, 1)$ ; initial iterate  $x^0 \in M$

**Output:** Sequenza  $\{x^k\}$

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

Prendere  $\eta^k \in T_x M$  tale per cui la sequenza  $\{\eta^i\}_{i=0,1,\dots}$  sia *gradient related*.

Selezionare  $x^{k+1}$  tale che

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

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 può essere visto come una generalizzazione della scelta del passo di Armijo descritto nella sotto-sezione 2.2, 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  $x^{k+1}$  garantendoci la convergenza ad un punto stazionario, come enunceremo formalmente, e ciò può portare alla definizione di algoritmi altamente efficienti.

### 2.3.3 Convergenza

Il concetto di convergenza può 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  $(U, \psi)$  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  $\{\psi(x^k)\}_{k=K,K+1,\dots}$  converge a  $\psi(x^*)$  (in tal caso applichiamo il concetto di convergenza di  $R^n$ , avendo  $\psi$  come codominio uno spazio euclideo). Il punto  $\psi^{-1}(\lim_{k \rightarrow \infty} \psi(x^k))$  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 2.2 rispetto al passo di Armijo in  $R^n$ , in tal caso però viene data una generalizzazione di quest'ultimo essendo, come già detto, *ALS* una generalizzazione del passo di Armijo e lavorando non più su  $R^n$  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

$\lim_{k \rightarrow \infty} \|\text{grad } f(x^k)\| = 0$ .

### 3 Disco di poincare e Iperboloide

Lo spazio iperbolico può essere definito in vari modi equivalenti, nessuno dei quali è prevalente in letteratura. Ciascuno di questi è un modello dello spazio iperbolico, noi descriveremo l'iperboloide ed il disco di poincare

#### 3.1 Iperboloide

Possiamo definire  $H^n$  come il luogo dei punti di norma -1 in  $R^{n:1}$ , dotato dell'usuale prodotto di minkowski ( $\langle u, v \rangle_{n:1} = \sum_{i=1}^n u_i v_i - u_{n+1} v_{n+1}$ ;  $u, v \in R^{n+1}$ ), questo luogo di punti ha in realta due componenti connesse, noi ne scegliamo una.

**Definizione 6 (Iperboloide)** *Il modello dell'iperboloide è definito nel modo seguente.*  
 $H^n = \{x \in R^{n+1} | \langle x, x \rangle_{n:1} = -1; \quad x_{n+1} > 0\}$

Lo spazio tangente in un punto  $p \in H^n$  è  $T_p H^n = \{x \in R^{n:1} | \langle p, x \rangle_{n:1} = 0\}$ .  
Possiamo asserire che l'iperboloide è una varietà riemanniana. Andiamo ora ad esplicitare gli strumenti necessari per effettuare ottimizzazione su questa varietà.

##### 3.1.1 Metrica

La metrica, definita nel *TangentBundle* dell'iperboloide, è la metrica indotta dal prodotto di Minkowski già esplicitata.

##### 3.1.2 Distanza

La distanza tra due punti dell'iperboloide rappresenta la lunghezza della geodedica che unisce i due punti, è definita come segue.

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

##### 3.1.3 Mappa Esponenziale

La mappa esponenziale definita su  $p \in H^n$  e  $v \in T_p H^n$ , che ha proprietà di retrazione, e che rappresenta la geodedica originaria in  $p$  con derivata in questo punto pari a  $v$ . E' definita come segue.

$$\operatorname{Exp}_p(v) = \cosh(\|v\|_{n:1})p + \sinh(\|v\|_{n:1})v / \|v\|_{n:1}$$

##### 3.1.4 Gradiente della funzione distanza

La deifinize dell'iperboloide e dello spazio di minkowski nel quale è costruito ci aiuta nel calcolo del gradiente della funzione distanza.

Data una funzione costo  $E$  qualsiasi possiamo calcolarne il gradiente in un punto  $p \in H^n$ , anch'esso arbitrario, passando dapprima per il gradiente definito nell'ambiente, in tal caso lo spazio di minkowski, ricordiamo che quest'ultimo è così definito.

$$\nabla_p^{R^{n:1}} E = (\frac{\partial E}{\partial x_1}|_p, \dots, \frac{\partial E}{\partial x_n}|_p, -\frac{\partial E}{\partial x_{n+1}}|_p)$$

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

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

Relativamente al problema della media di fréchet che andremo ad esplicitare definiremo la seguente formula rappresentante il gradiente su spazio di minkowski della funzione distanza.

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

### 3.2 Disco di Poincaré

Andiamo ora a definire l'altro modello dello spazio iperbolico, rappresentato come tutti i punti di norma strettamente minore di 1 in  $R^n$ .

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

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

Essendo  $D^n$  un sottoinsieme di  $R^n$  lo spazio tangente in ogni punto di  $D^n$  è  $R^n$ . Come fatto per l'iperboloide andiamo ad esplicitare i vari strumenti necessari per effettuare ottimizzazione.

#### 3.2.1 Metrica

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

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

#### 3.2.2 Distanza

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

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

#### 3.2.3 Mappa Esponenziale

La mappa esponenziale definita su  $p \in D^n$  e  $v \in T_p D^n$ , che ha proprietà di retrazione, e che rappresenta la geodetica originaria in  $p$  con derivata in questo punto pari a  $v$ . E' definita come segue.

$$\operatorname{Exp}_p(v) = \frac{\lambda_p(\cosh(\lambda_p\|v\|) + \langle p, \frac{v}{\|v\|} \rangle \sinh(\lambda_p\|v\|))p}{1 + (\lambda_p - 1)\cosh(\lambda_p\|v\|) + \lambda_p \langle p, \frac{v}{\|v\|} \rangle \sinh(\lambda_p\|v\|)} + \frac{\frac{1}{\|v\|} \sinh(\lambda_p\|v\|)v}{1 + (\lambda_p - 1)\cosh(\lambda_p\|v\|) + \lambda_p \langle p, \frac{v}{\|v\|} \rangle \sinh(\lambda_p\|v\|)}$$

#### 3.2.4 Gradiente della funzione distanza

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

$$\nabla_u^E d(u, v) = \frac{4}{b\sqrt{c^2-1}} \left( \frac{(\|v\|^2 - 2\langle u, v \rangle + 1)v}{a^2} - \frac{v}{a} \right) \text{ con } a = 1 - \|u\|^2, b = 1 - \|v\|^2, \\ c = 1 + \frac{2}{ab} \|u - v\|^2.$$

Sappiamo che per ricavare il gradiente riemanniano dal gradiente euclideo è sufficiente riscrivere  $\nabla^E$  per l'inverso della metrica riemanniana, perciò:  $\nabla_u^{D^n} d(u, v) = \frac{\nabla_u^E d(u, v)}{\lambda_u^2}$

### 3.3 Iperboloide come modello conforme

Come inizialmente accennato le due varietà trattate fin'ora sono due modelli conformi 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 proiezione stereografica  $\rho$  definita come segue.

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

A questo punto facciamo vedere che vale la seguente uguaglianza per  $n = 2$ , ci sarà utile nella sezione successiva.

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

Siano  $a, b \in D^n$

$$a = (a_1, a_2) \\ b = (b_1, b_2) \\ \rho^{-1}(a) = \frac{2}{1-r_a}(a_1, a_2, \frac{1+r_a}{2}) \\ \rho^{-1}(b) = \frac{2}{1-r_b}(b_1, b_2, \frac{1+r_b}{2}) \\ \text{con } r_a = (a_1^2 + a_2^2) \text{ e } r_b = (b_1^2 + b_2^2)$$

$$d_{H^n}(\rho^{-1}(a), \rho^{-1}(b)) = \operatorname{arccosh}(-\langle (\frac{2a_1}{1-r_a}, \frac{2a_2}{1-r_a}, \frac{1+r_a}{2}), (\frac{2b_1}{1-r_b}, \frac{2b_2}{1-r_b}, \frac{1+r_b}{2}) \rangle) = \operatorname{arccosh}(-(\frac{4a_1b_1}{(1-r_a)(1-r_b)} + \frac{4a_2b_2}{(1-r_a)(1-r_b)} - \frac{(1+r_a)(1+r_b)}{(1-r_a)(1-r_b)}) = \operatorname{arccosh}(\frac{(1+r_a)(1+r_b)-4a_1b_1-4a_2b_2}{(1-r_a)(1-r_b)}) = \operatorname{arccosh}(\frac{1+r_a+r_b+r_ar_b-4a_1b_1-4a_2b_2}{1-r_a-r_b+r_ar_b}) = \operatorname{arccosh}(\frac{1-r_a-r_b+2r_a+2r_b+r_ar_b-4a_1b_1-4a_2b_2}{1-r_a-r_b+r_ar_b}) = \operatorname{arccosh}(1 + \frac{2r_a+2r_b-4a_1b_1-4a_2b_2}{(1-r_a)(1-r_b)}) = \operatorname{arccosh}(1 + 2\frac{\|a-b\|^2}{(1-\|a\|^2)(1-\|b\|^2)}) = d_{D^n}(a, b)$$

### 3.4 Media di Fréchet e problema del centroide

**Definizione 8 (Media di Fréchet)** La funzione media di Fréchet di  $p$ -punti è così definita.

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

Il problema di trovare il centroide di  $p$ -punti consiste nel trovare il punto che min-

imizza la funzione media di Fréchet definita a partire dai punti in questione, il cui minimo esiste ed è unico. Tale problema può essere risolto direttamente in uno spazio euclideo il cui minimo è dato dalla media aritmetica dei p-punti. Mentre in uno spazio iperbolico, come nel nostro caso, la soluzione non è diretta e perciò si applicano le metodologie di ottimizzazione su varietà differenziabili viste precedentemente.

Diamo ora una congettura che possiamo affermare essere corretta in funzione dell'uguaglianza  $d_{D^n}(a, b) = d_{H^n}(\rho^{-1}(a), \rho^{-1}(b))$  che abbiamo verificato nella sezione precedente.

Dati  $a^{(i)}, \Theta \in D^n$ ; con  $i = 1 \dots p$  dall'uguaglianza precedente è facile notare che  $\frac{1}{p} \sum_{i=1}^p d_{D^n}^2(\Theta, a^{(i)}) = \frac{1}{p} \sum_{i=1}^p d_{H^n}^2(\rho^{-1}(\Theta), \rho^{-1}(a^{(i)}))$ , perciò essendo  $\rho$  un diffeomorfismo se  $\exists \Theta \in D^n$  per cui  $\frac{1}{p} \sum_{i=1}^p d_{D^n}^2(\Theta, a^{(i)})$  è minima allora esiste  $\exists \Psi \in H^n$  per cui è minima  $\frac{1}{p} \sum_{i=1}^p d_{H^n}^2(\Psi, \rho^{-1}(a^{(i)}))$ . Essendo inoltre  $f(x) = \hat{f} = f(\rho^{-1}(x))$  allora cercare il minimo di  $f$  equivale a cercare il minimo di  $\hat{f} \circ \rho^{-1}$ .

In funzione di tale congettura possiamo trasportare il problema definito su disco di Poincaré nella varietà Iperboloide e viceversa, notiamo che le formule necessarie ad effettuare ottimizzazione sulla varietà iperboloidi sono computazionalmente meno complesse rispetto a quelle definite per il disco di Poincaré.

## 4 Metodi utilizzati con pseudocodice ed esperimenti

In questa sezione andremo a definire lo pseudocodice degli algoritmi di ottimizzazione che abbiamo utilizzato per confrontare il calcolo del minimo della media di Fréchet sui modelli dello spazio iperbolico descritti nel capitolo precedente.

### 4.1 Algoritmi e Pseudocodice

#### 4.1.1 Passo fisso

#### 4.1.2 Armijo

#### 4.1.3 Barzilai-Borwein

### 4.2 Esperimenti

## 5 Appendice

### 5.1 Varietà

Sia  $M$  un insieme. Una bigiezione  $\phi$  di un sottoinsieme  $U$  di  $M$  in un sottoinsieme aperto di  $R^d$  è chiamata mappa d-dimensionale dell'insieme  $U$ . Data una carta  $(U, \psi)$  e  $x \in U$  gli elementi di  $\psi(x) \in R^d$  sono le coordinate di  $x$  nella carta  $(U, \psi)$ . Un *Atlas*  $(A)$  di  $M$  in  $R^d$  è un insieme di carte  $(U_a, \psi_a)$  dell'insieme  $M$  tali che:

- $\bigcup_a U_a = M$
- Per ogni coppia  $\alpha, \beta$  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  $\psi_\alpha \circ \psi_\beta^{-1} : R^d \rightarrow R^d$  è *smooth*.

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

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

Data una carta  $\psi$  su  $M$ , l'inverso  $\psi^{-1}$  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 un manifold, sostituendo  $t \mapsto (x + t\eta)$  con una curva su  $M$  attraverso  $x \in M$  (punto di derivazione) ( $\gamma(0) = x$ ) questo porta ad una derivata direzionale ben definita  $\frac{\partial f(\gamma(t))}{\partial t}|_{t=0}$ . Formalmente, sia  $M$  una varietà e  $\gamma$  una mappa differenziabile  $\gamma : R \rightarrow M | t \mapsto \gamma(t)$  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  $R$  in  $R$  con una ben definita derivata.

Sia  $x \in M$ ;  $\gamma(0) = x$  allora  $F_x(M)$  è l'insieme delle funzioni reali differenziabili definite in un intorno di  $x$ . Il mapping  $\dot{\gamma}(0)$  da  $F_x(M)$  in  $R$  è definito come:  $\dot{\gamma}(0)f = \frac{\partial f(\gamma(t))}{\partial t}|_{t=0}$  con  $f \in F_x(M)$ ; è chiamato vettore tangente alla curva  $\gamma$  in  $t = 0$ .

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

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

Tale curva  $\gamma$  è detta realizzare il vettore tangente  $\epsilon_x$

**Definizione 11 (Spazio Tangente)** Uno spazio tangente ad  $M$  in  $x$  denominato come  $T_x M$  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  $T_x M$  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  $T_x M$  mette a disposizione un'approssimazione locale di spazio vettoriale della varietà. Perciò tramite le Retrazioni possiamo effettuare ottimizzazione sullo spazio vettoriale  $T_x M$ .



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

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

Sia  $P : M \rightarrow N$  una mappa liscia tra due varietà  $M$  ed  $N$ . sia  $\epsilon_x$  un vettore tangente in un punto  $x$  di  $M$ , si può far vedere che la mappa  $DP(x)[\epsilon_x]$  da  $T_x M$  in  $T_{P(x)} N$  è definita da:

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

è un vettore tangente ad  $N$  in  $P(x)$ .

La mappa  $DP(X) : TM \rightarrow TN$  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  $T_x M$  con un prodotto interno  $\langle \cdot, \cdot \rangle_x$  bilineare, definito positivo e simmetrico. Il prodotto interno induce una norma  $\| \epsilon_x \|_x = \sqrt{\langle \epsilon_x, \epsilon_x \rangle_x}$  su  $T_x M$ . Una varietà il cui spazio tangente è munito di un prodotto interno prende il nome di varietà riemanniana  $(M, g)$  con  $g$  metrica riemanniana.

La lunghezza di una curva  $\gamma : [a, b] \rightarrow M$  in una varietà riemanniana  $(M, g)$  è definita da:  $L(\gamma) = \int_a^b \sqrt{g(\dot{\gamma}(t), \dot{\gamma}(t))} dt$

La distanza riemanniana in una varietà riemanniana connessa  $(M, g)$  è  $dist : M \times M \rightarrow \mathbb{R}$  definita da  $dist(x, y) = \inf_{\gamma \in \Gamma} L(\gamma)$ , dove  $\Gamma$  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 di  $f$  in  $x$  è definito come l'elemento di  $T_x M$  che soddisfa:  $\langle f(x), \epsilon \rangle_x = Df(x)[\epsilon]$ .

Sia  $M$  una sottovarietà immersa di una varietà riemanniana  $\hat{M}$  siccome ogni spazio tangente  $T_x M$  può essere trattato come sottospazio di  $T_x \hat{M}$  la metrica  $\hat{g}$  di  $\hat{M}$  induce una metrica  $g$  su  $M$ ,  $g_x(\epsilon, c) = \hat{g}_x(\epsilon, c)$  con  $\epsilon, c \in T_x M$ , 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 sottoinsieme di  $T$  è in  $T$
- L'intersezione di elementi di qualsiasi sotto collezione 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  $X$  hanno un intorno distinto.

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

### 5.3 Definizioni Utili

**Definizione 14 (Positività)** Sia  $A \in R^{n \times n}$  è definita positiva se è simmetrica e se  $v'Av > 0$  se  $v$  è un vettore  $\in R^n - 0$ , con  $v = 0$  è diretto che  $v'Av = 0$ .

**Definizione 15 (Convessità)** Una funzione  $f : \Omega \rightarrow R$  con  $\Omega \subseteq V$ ,  $V$  (spazio vettoriale) allora 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  $f : V \rightarrow R$ ,  $V$  (spazio vettoriale), allora una derivata parziale su  $x \in V$  è definita come:

$\frac{\partial f(x)}{\partial x_i} = \lim_{h \rightarrow 0} \frac{f(x + he_i) - f(x)}{h}$ ; 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  $x_0 \in R$  è un punto di accumulazione per un insieme  $E \subset R$  se comunque scelto un intorno  $B(x_0, \epsilon)$ , risulta che  $B(x_0, \epsilon)$  contiene almeno un punto di  $E$  diverso da  $x_0$ .

**Definizione 18 (Gradiente)** Consideriamo una funzione  $f$  definita su un insieme aperto  $A \subset R^n$  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:  $\nabla f = [\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n}]$ , inoltre per il teorema della formula del gradiente sappiamo che, dato un versore  $v$  esiste la derivata direzionale  $f_v$  in  $x$  e che  $f_v(x) = \nabla f(x) \cdot v$ .

#### 5.3.1 Teorema di Taylor

Il teorema di Taylor è un teorema che fornisce una sequenza di approssimazioni di una funzione differenziabile attorno ad un dato punto mediante i polinomi di Taylor, ovvero polinomi i cui coefficienti dipendono solo dalle derivate della funzione nel punto. Diamo un esempio pratico di approssimazione di Taylor del secondo ordine di una funzione a due variabili.

$$f(x_0+h, y_0+k) = f(x_0, y_0) + f_x(x_0, y_0)h + f_y(x_0, y_0)k + \frac{1}{2}[f_{xx}(x_0, y_0)h^2 + f_{xy}(x_0, y_0)hk + f_{yy}(x_0, y_0)k^2] + R(h, k); \text{ con } R(h, k) \in o(\|(h, k)\|^2).$$