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

1 Introduzione

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

2 Metodi di ottimizzazione su R^n e Varieta differenziabili

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

2.1 Condizioni di ottimalità

Sia $f : R^n \to R$ 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

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

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

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

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 Δx è non negativa

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

perciò dovendo valere per Δx 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 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

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

Proposizione 1 (Condizioni necessarie di ottimalità) *Sia x^* un minimo locale non vincolato di $f : R^n \to R$, 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
$\nabla^2 f(x^*)$*: 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 $f : R^n \to R$ 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 convess 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 condizione di ottimalita necessarie definite nelle sezione precedente [$\nabla f(x^*) = 0 \quad \& \quad \nabla^2 f(x^*) \geq 0$] 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 Δx 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 $f : R^n \to R$ doppiamente differenziabile con continuità in un insieme aperto S. Si supponga esistere $x^* \in S$ che soddisfi le condizioni*
$\nabla f(x^*) \quad \nabla^2 f(x^*)$: definita positiva
allora, x^* è un minimo locale non vincolato di f. In particolare esistono $\gamma > 0$ e $\epsilon > 0$ tali che
$f(x) \geq f(x^*) + \gamma/2 \parallel x - x^* \parallel^2, \quad \forall x \quad con \parallel x - x^* \parallel < \epsilon$

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 sarannno trattate con un occhio alle possibili applicazioni.
Consideriamo il problema di trovare il minimo non vincolato di una funzione $f : R^n \to R$ 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 $x0$ e successivamente si genera una sequenza $x1, x2, ...$ 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 cio possiamo esplitare una formula che definisce una classe di algoritmi
$x^{k+1} = x^k + \alpha^k d^k \quad k = 0, 1, ...$
Dove se $\nabla f(x^k) \neq 0$ la direzione d^k è scelto 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 diretto che $\nabla f(x^k)' d^k < 0$.
A seconda della scelta di D^k abbiamo differenti metodologie applicabili.

Steepest Descent:
$D^k = I, \quad k = 0, 1, ...$

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

Metodi di quasi Newton:
$D^k = H(x^k) \approx \nabla^2 f(x^k)$

Nella pratica $h^k = (\nabla^2 f(x^k))^{-1} d^k$, per trovare h^k riscriviamo $(\nabla^2 f(x^k)) h^k = d^k$, risolvibile in tempo polinimiale tramite metodi di Krylov.

2.2.2 Selezione del passo

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

Ricerca lineare esatta:
scegliamo α^k tale che minimizza la funzione costo f lungo la direzione d^k, percio $f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k)$.

Ricerca lineare esatta limitata:

Questa è una versione modifica della metodologia precedente più semplice da implementare in vari casi. 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 a garantire la convergenza del metodo sotto opportune ipotesi. Il metodo di armijo è stato definito come segue. Siano fissi s, β e σ 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
$f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(^k)' d^k$.
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* $\limsup_{k \in K} \nabla f(x)' d^k < 0$

Definizione 2 *Una successione di direzioni $\{x^k\}_k, \quad \{d^k\}_k$ 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 α^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 $\alpha^k = s, \quad k = 0, 1, ...$
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 $\sum_{n=1}^{\infty} \alpha^k = \infty$

2.3 Discesa del gradiente su varietà

Ora che abbiamo un contesto teorico e una consapevolezza di cosa voglia dire effetutare ottimizzazione su uno spazio euclideo n-dimensionale (R^n), possiamo approcciare le metodologie computazionali per trovare un minimo non vincolato di una funzione $f : M \to R$ 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_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 $T_xM \quad \forall x \in M$ 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
grad $\hat{f}_x(0_x)$ =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 η^k è in $T_x M$ 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 $\{\eta^k\}$, $\eta^k \in T_x M$ 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*
$\limsup_{k\to\infty, k\in K} \langle \textit{grad}\, f(x^k), \eta^k \rangle < 0$

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*
$f(x) - f(R_x(\beta^m \bar{\alpha}\eta)) \geq -\sigma \langle \textit{grad}\, f(x), \beta^m \bar{\alpha}\eta \rangle_x$.
*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 $k = 0, 1, 2, \ldots$ **do**
  Prendere $\eta^k \in T_x M$ 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

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} grantendoci la convergenza ad un punto stazionario, come enunceremo formalmente, e ciò puo portare alla definizione di algoritmi altamente efficienti.

2.3.3 Convergenza

Il concetto di convergenza puo essere ridefinito e generalizzato su varietà. Una sequenza infinita $\{x^k\}_{k=0,1,...}$ 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 $\{\psi(x^k)\}_{x=K,K+1,...}$ converge a $\psi(x^*)$ (in tal caso applichiamo il concetto di convergenza di 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 $\{x^k\}_{k=0,1,...}$. Tutte le sequenze convergenti di una vairetà 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*
$\lim_{k\to\infty} \|grad\ f(x^k)\| = 0$.

3 Disco di poincare e Iperboloide

3.1 Iperboloide

3.2 Disco di Poincaré

3.3 Iperboloide come modello conforme

3.4 Media di Fréchet e problema del centroide

4 Metodi utilizzati con pseudocodice

4.1 subsection

5 Esperimenti (e Applicazioni)

5.1 subsection

6 Appendice

6.1 Varietà

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

- $\bigcup_a U_a = M$
- Per ogni coppia α, β con $U_\alpha \bigcap U_\beta \neq \emptyset$ gli insiemi $\psi_\alpha(U_\alpha \bigcap U_\beta)$ e $\psi_\beta(U_\alpha \bigcap U_\beta)$ sono sottoinsiemi aperti di R^d ed il cambio di coordinate $\psi_\alpha \circ \psi_\beta^{-1} : R^d \to R^d$ è $smooth$.

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

Definizione 6 *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 ψ su M, l'inverso ψ^{-1} è chiamata parametrizzazione locale di M.

6.1.1 Vettori Tangenti

6.1.2 Metriche distanze e e gradienti Riemanniani

6.2 Definizioni Utili

Definizione 7 (Positività)

Definizione 8 (Convessità)

Definizione 9 (Derivate Parziali)

Definizione 10 (Spazio di Hausdorff)

Definizione 11 (Punto di accumulazione)

Definizione 12 (Taylor a più variabili)

Definizione 13 (Gradiente)

Definizione 14 (Derivata di Fréchet)

Definizione 15 (Spazio topologico)

- positivita - convessita - insieme chiuso - derivate parziali - derivata di fréchet - teorema di rappresentazione di ritz - teorema di schwartz - teorema di weiestrass - gradiente - taylor - elementi base di topologia (?) - varieta differenziabili - spazio di hausdorff - punto di accumulazione