Report

Diletta Goglia, Lisa Lavorati

A.A. 2019-20

1. Analisi teorica

1.1 Introduzione al problema e definizioni preliminari

(P) *is the problem of estimating the matrix norm* $||A||_2$ *for a (possibly rectangular) matrix* $A \in R^{m \times n}$*, using its definition as an (unconstrained) maximum problem.*

(A1) *is a standard gradient descent (steepest descent) approach.*

(A2) *is a a quasi-Newton method such as BFGS (one which does not require any approximations of the Hessian of the function)*

1.1.1 Norma euclidea di una matrice e di un vettore

Iniziamo dando la definizione di norma-2 (o **norma Euclidea**) di una matrice partendo dalla formula della **norma indotta**:

$||A||_2 := \max_{x \neq 0}(\frac{||Ax||_2}{||x||_2})$

con $A \in R^{m \times n}$ e $x \in R^n$.

La norma di A dunque sarà il massimo di tutti i rapporti possibili per ogni vettore $x \neq 0$.

La norma indotta, quindi, definisce la norma di una matrice in termini di norma di un vettore.

Definiamo la norma-2 del vettore x :

$||x||_2 := \sqrt{x^T x}$

Partendo da questa formula possiamo riscrivere la norma-2 del vettore Ax nel seguente modo:

$||Ax||_2 = \sqrt{A(x^T x)} = \sqrt{(Ax)^T(Ax)} = \sqrt{(x^T A^T)(Ax)} = \sqrt{x^T Q x}$

dove:

- A^T è la matrice **trasposta** di A
- $Q = A^T A$ è una matrice **quadrata**, **simmetrica** e **diagonalizzabile** (quindi anche **invertibile**).
  Inoltre, è possibile dimostrare che Q è **semidefinita positiva** e che dunque i suoi autovalori sono tutti non negativi. Infatti, Q si dice semidefinita positiva se la sua forma quadratica associata è semidefinita positiva, ovvero se per ogni vettore x: $x^T Q x \succeq 0$.
  Questo si dimostra semplicemente con:
  $x^T Q x = x^T A^T A x = ||Ax||_2^2 \geq 0 \Longleftrightarrow A^T A = Q \succeq 0$

La norma della matrice A diventa dunque:

$||A||_2 = \max \sqrt{\frac{x^T Q x}{x^T x}}$

1.1.2 Rappresentazione geometrica

Se guardiamo ad una matrice come un operatore lineare tra due spazi vettoriali, la norma indotta di una matrice ci dice qual è il massimo allungamento che l'operatore produce nel trasformare i vettori.

Geometricamente, quindi, la norma indotta rappresenta "la massima distorsione positiva della lunghezza che la trasformazione lineare $A : x \mapsto Ax$ produce".

1.1.3 Definizione della *objective function*

Dato che la radice quadrata è una **funzione monotona**, vale:

$$\max h(g(x)) = h(\max g(x))$$

Quindi possiamo scrivere:

$$\max \sqrt{\frac{x^T Qx}{x^T x}} = \sqrt{\max(\frac{x^T Qx}{x^T x})}$$

La funzione da massimizzare perciò diventa:

$$f(x) = \frac{x^T Qx}{x^T x}$$

Affronteremo tuttavia questo problema di ottimizzazione come un problema di **minimo**: dunque riscriviamo il problema come:

$$min\{f(x)\} = -\max\{-f(x)\}$$

La nostra *objective function* definitiva dunque è la seguente:

$$f(x) = -\frac{x^T Qx}{x^T x}$$

1.2 Verifica delle soluzioni

Il nostro obiettivo è quello di trovare una *optimal solution* (che indichiamo con la seguente notazione: x_*) tale che minimizzi il valore della funzione facendoci ottenere un *optimal value*, (indicato con $f_* = f(x_*)$).

Il risultato che ci aspettiamo, dunque, non è il valore della norma della matrice, ma il vettore x_* che minimizza il valore della funzione in oggetto (f_*).

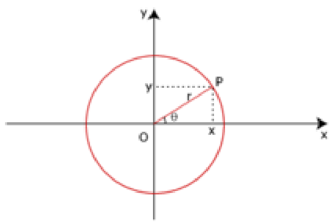
Ci aspettiamo inoltre che i vettori ottenuti (*feasible solution*s) candidati ad essere possibili soluzioni ottimali siano più di uno, a seconda del punto (vettore) di partenza nella ricerca dei una soluzione ottimale.

N.B. I punti di partenza corrispondono ai punti stazionari (vedi cap. "1.4.4. Punti stazionari").

1.2.1 Proporzionalità delle soluzioni

Possiamo eseguire un controllo sui vettori x_* ottenuti, verificando che siano effettivamente soluzioni semplicemente normalizzandoli, ovvero calcolando per ciascuno di essi il rapporto $\frac{x}{||x||}$ e verificando che il risultato sia uguale per ogni x_*.

Possiamo visualizzare ciò tramite un grafico in coordinate polari. Notiamo che tutti i vettori x_* ottenuti come soluzioni si trovano tutti lungo una stessa direzione, ovvero sono indipendenti dal raggio r.



Consideriamo un vettore x in $\mathbb{R}^2$.

In coordinate polari, otteniamo: $rcos\theta$ per la componente orizzontale, e $rsin\theta$ per quella verticale.

Da cui:

$$f(x) = -\frac{x^T Qx}{x^T x} = -\frac{r(cos\theta, sin\theta)Qr\binom{cos\theta}{sin\theta}}{r^2} = -(cos\theta, sin\theta)Q\binom{cos\theta}{sin\theta}$$

Questo significa che il valore della norma non dipende dal raggio della circonferenza, ovvero che tutte le soluzioni ottenute sono proporzionali. Verifichiamo ciò normalizzandoli, come suddetto.

Questa osservazione è confermata dai grafici delle curve di livello al capitolo 1.5.1 e può essere generalizzata per spazi multidimensionali $\mathbb{R}^n$ con $n > 2$, in cui θ sarà un angolo solido.

1.2.2 SVD

Come ulteriore controllo nella correttezza delle soluzioni, verifichiamo che il rapporto $\frac{x}{||x||}$ sia esattamente il vettore che costituisce la prima colonna della matrice V nella *Single Value Decomposition (SVD)* della matrice A di cui dobbiamo calcolare la norma.

In caso contrario, il vettore x_* ottenuto non può essere una soluzione.

Questo perché, data una matrice $A \in R^{m \times n}$, possiamo scriverne la *SVD*:

$$||A|| = ||U\Sigma V^T|| = ||U\Sigma|| = ||\Sigma||$$

in quanto la norma-2 è invariante nella moltiplicazione ortogonale sia a destra che a sinistra.

Nella *SVD*, U e V sono due matrici ortonormali, e Σ è una matrice diagonale i cui elementi, detti valori singolari della matrice di partenza A, hanno la proprietà di essere:

$$\sigma_i \geq 0 \; \forall i, \; \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_m$$

Inoltre:

$$||\Sigma|| = || \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & ... & \\ & & & \sigma_m \end{bmatrix} || = \sigma_1$$

Questo vale perché la norma di una matrice diagonale corrisponde al valore del suo elemento massimo, che in questo caso è σ_1.

Quindi, l'*optimal value* corrisponde a: $f_* = ||A|| = \sigma_1$.

Mentre l'*optimal solution* x_* che cerchiamo è esattamente la prima colonna della matrice Vnella *SVD* di A, ovvero il vettore v_1 che massimizza il rapporto $\frac{||Av_1||}{||v_1||}$.

$$\frac{||Av_1||}{||v_1||} = \sigma_1 = ||A|| = f_* \Rightarrow x_* = v_1$$

1.3 Gradiente

Il gradiente di $f(x)$ è stato calcolato a mano (vedere **Allegato 1**), riportiamo di seguito i passaggi principali per ottenerlo:

$$f(x) = -\frac{x^T A^T Ax}{x^T x} = -\frac{x^T Qx}{x^T x} = \frac{-\sum_{i=1}^n \sum_{j=1}^n x_j q_{ij} x_i}{\sum_{i=1}^n x_i^2}$$

$$\frac{\delta f}{\delta x_k} = \frac{-2x_k \sum_{i=1}^n \sum_{j=1}^n x_j q_{ij} x_i}{(\sum_{i=1}^n x_i^2)^2} - \frac{-\sum_{j=1}^n 2x_j q_{kj}}{\sum_{i=1}^n x_i^2} = \frac{2x_k(x^T Qx)}{(x^T x)^2} - \frac{\sum_{j=1}^n 2x_j q_{kj}}{x^T x}$$

quindi possiamo scrivere il gradiente come:

$$\nabla f = \frac{2x(x^T Qx)}{(x^T x)^2} - \frac{2Qx}{x^T x}$$

Allegato 1.

20200826_172739.jpg https://drive.google.com/file/d/1BwtT0Gs3b0E5jVxNNtv42P1iMwNmVnMT/view?usp=drivesdk

20200826_172717.jpg https://drive.google.com/file/d/1Bx5-XwCMyQfSTpDUybs8BB2InpvyLsKF/view?usp=drivesdk

20200826_172649.jpg https://drive.google.com/file/d/1C2yCDiHCtCBCtwpCNMtlqQU4Vd7s6vbh/view?usp=drivesdk

1.4 Proprietà interessanti della funzione

1.4.1 Dominio e codominio

La funzione è definita $\mathbb{R}^n_{\neq 0}$, ovvero è definita per ogni vettore x diverso dal vettore nullo, ed è **limitata inferiormente:**

$\exists m \in \mathbb{R}^+ : f(x) = -\frac{x^TQx}{x^Tx} \geq m \, \forall x \in \mathbb{R}^n_{\neq 0}$

Quindi:

$\exists m \in \mathbb{R}^+ : -f(x) = \frac{x^TQx}{x^Tx} \leq m \, \forall x \in \mathbb{R}^n_{\neq 0}$

$\iff ||-f(x)|| = ||\frac{x^TQx}{x^Tx}|| \leq \frac{||x|| \cdot ||Q|| \cdot ||x||}{||x|| \cdot ||x||} \leq ||Q|| \leq m$

1.4.2 Continuità e derivabilità

La funzione è continua e derivabile in ogni punto eccetto che in zero.

$\frac{\delta f}{\delta x_k} = \frac{2x_k(\sum_{i=1}^n \sum_{j=0}^n x_i x_j a_{ij})}{(\sum_{i=1}^n x_i^2)^2} - \frac{\sum_{j=1}^n 2x_j a_{kj}}{\sum_{i=1}^n x_i^2}$

Infatti possiamo ottenere le derivate parziali per un qualsiasi vettore $x \in R^n_{\neq 0}$.

Inoltre, la funzione è **Lipschitz continua.**

Per dimostrarlo possiamo usare la seguente relazione:

Se una funzione $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$ è lipschitziana e differenziabile, allora esiste una costante L tale per cui la jacobiana soddisfi:
$\|J_f(x)\| \leq L \quad \forall x \in U$

Per definizione, una funzione è Lipschitz continua in un intervallo I se $\exists L > 0$:

$|f(x) - f(y)| \leq L||x - y|| \ \ \forall x, y \in I$

Essendo la nostra funzione è definita in $\mathbb{R}^n_{\neq 0} \to \mathbb{R}^+$, continua e derivabile in ogni punto eccetto che in 0, allora questa sarà anche Lipschitz continua purché il suo gradiente sia limitato. Infatti:

Proposizione. Sia (a, b) un intervallo aperto in $\mathbb{R}$ eventualmente illimitato e sia

$f : (a, b) \to \mathbb{R}$ una funzione derivabile. Allora f è Lipschitz continua *se* la funzione

$\nabla f : (a, b) \to \mathbb{R}$ è limitata.

$\nabla f = \frac{2x(x^T Qx)}{(x^T x)^2} - \frac{2Qx}{x^T x}$

Il gradiente della funzione è limitato per ogni insieme I = { $x : ||x|| \geq \epsilon > 0$ } perché altrimenti si annullano i denominatori.

Dunque la funzione è Lipschitz continua su I.

Unit-sphere nello spazio vettoriale → { $x \in I$: $||x|| = 1$ } → in questa sfera la funzione è limitata da m (vedi sopra).

1.4.3 Non convessità

La funzione non è convessa in quanto discontinua all'interno del dominio (vedi sopra).

Questo significa che, se ci troviamo in un punto stazionario, non possiamo assicurare che questo sia anche un minimo globale.

In realtà, ai fini dell'ottimizzazione, la condizione di convessità non è necessaria per trovare una *optimal solution* (vedi anche cap. 2.1).

1.4.4 Punti stazionari

Riportiamo di seguito i passaggi principali del calcolo dei punti stazionari della funzione in oggetto (vedi **Allegato 2**):

$f(x) = -\frac{x^T Qx}{x^T x}$

$\nabla f = \frac{2x(x^T Qx)}{(x^T x)^2} - \frac{2Qx}{x^T x}$

$\nabla f(\bar{x}) = 0 \iff \frac{2\bar{x}(\bar{x}^T Q\bar{x})}{(\bar{x}^T \bar{x})^2} - \frac{2Q\bar{x}}{\bar{x}^T \bar{x}} = 0$

$\iff \frac{2\bar{x}(\bar{x}^T Q\bar{x}) - 2Q\bar{x}(\bar{x}^T \bar{x})}{(\bar{x}^T \bar{x})^2} = 0 \iff \bar{x}(\bar{x}^T Q\bar{x}) = Q\bar{x}(\bar{x}^T \bar{x})$

$\iff \frac{\bar{x}(\bar{x}^T Q\bar{x})}{(\bar{x}^T \bar{x})} = Q\bar{x}$

Dunque abbiamo osservato come i punti stazionari $\bar{x}$ della funzione siano tutti e soli i punti tali per cui $\bar{x}$ è un **autovalore** di Q, il cui **autovettore** corrispondente è $f(\bar{x})$.

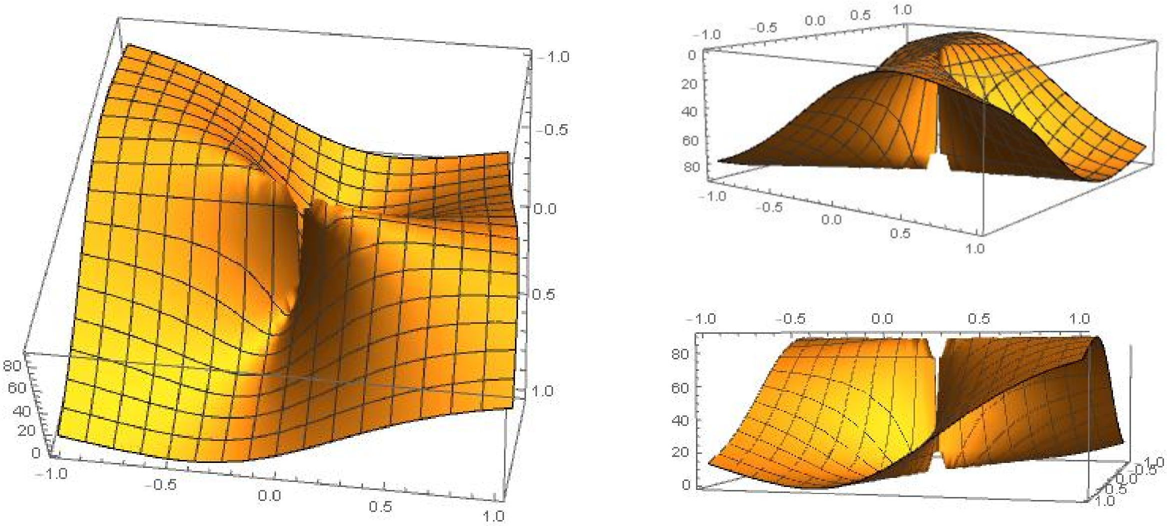
Allegato 2:

20200826_172623.jpg https://drive.google.com/file/d/1C81cj1Ly5Mr0n1AOIEpTg_1i3ek6hwUV/view?usp=drivesdk

I **punti stazionari,** ovvero **gli autovettori di Q,** sono i vettori da cui far partire gli algoritmi (infatti solo se sono in un punto stazionario ha senso controllare se è un minimo).

1.5 Grafico

Per visualizzare il comportamento della funzione in $\mathbb{R}^2$ ne abbiamo tracciato il grafico, per una matrice random $A \in \mathbb{R}^{3x2}$.

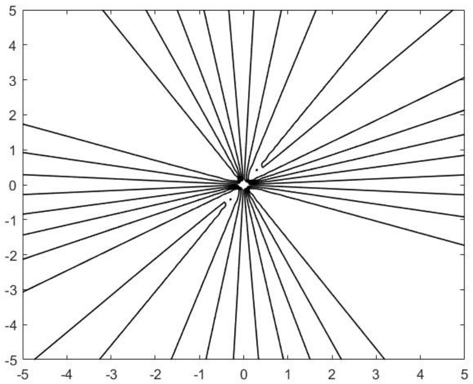


Possiamo vedere chiaramente la conferma di quanto dimostrato nell'analisi precedente. La funzione non è convessa, è discontinua (e quindi non differenziabile) in 0, ed è limitata inferiormente. Inoltre è Lipschitz continua, ovvero non varia "troppo" rapidamente (definito formalmente in 1.4.2).

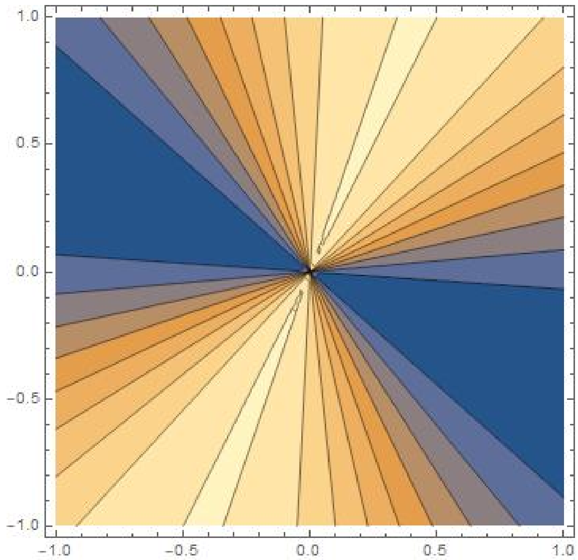
1.5.1 Level set

Le considerazioni fatte sull'indipendenza delle soluzioni rispetto al raggio sono confermate e ben visibili nel grafico delle curve di livello.

Lungo i raggi la funzione è costante, quindi le derivate parziali nelle direzioni radiali saranno 0.



Level set generated by MATLAB contour plot



Level set generated by Mathematica contour plot

2. Algoritmi

3. Definizione del dataset