Lecture 16: Geodesic Convexity & General Scaling Algorithms

Rafael Oliveira

University of Waterloo
Cheriton School of Computer Science

rafael.oliveira.teaching@gmail.com

March 10, 2021

Overview

- Lie Groups, Lie Algebras & Positive Definite Matrices
- Crash Course on Geodesic Convex Optimization
- Analysis of Scaling Problem for Conjugation Action
- Conclusion

A Lie-ttle bit of Lie Theory

- A Lie group is a "continuous" group think of $\mathbb{GL}(n)$

A Lie-ttle bit of Lie Theory

- A Lie group is a "continuous" group   think of $\mathbb{GL}(n)$
- Lie Algebra describes the "infinitesimal action" of the group near the identity

For $\mathbb{GL}(n)$ its Lie Algebra is $\mathrm{Mat}(n)$

For $\mathbb{SL}(n)$ its Lie Algebra is $\mathrm{Mat}(n)$ which are traceless

A Lie-ttle bit of Lie Theory

- A Lie group is a "continuous" group think of $\mathbb{GL}(n)$
- Lie Algebra describes the "infinitesimal action" of the group near the identity

For $\mathbb{GL}(n)$ its Lie Algebra is $\mathsf{Mat}(n)$

For $\mathbb{SL}(n)$ its Lie Algebra is $\mathsf{Mat}(n)$ which are traceless

- Exponential map gives us surjection (in our case)

$$\exp : \mathsf{Mat}(n) \to \mathbb{GL}(n)$$

- Matrix exponential:

$$\exp(A) = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots$$

Conjugation Action - Manifold of Positive Definite Matrices

- Let us consider the conjugation action
- $G = \mathbb{GL}(n)$, $V = \mathrm{Mat}(n)$

$$(h, A) \mapsto hAh^{-1}$$

Conjugation Action - Manifold of Positive Definite Matrices

- Let us consider the conjugation action
- $G = \mathbb{GL}(n)$, $V = \mathrm{Mat}(n)$

$$(h, A) \mapsto hAh^{-1}$$

- Our optimization problem:

$$\text{logcap}(A) = \inf_{h \in G} \log \|hAh^{-1}\|_F^2$$

A is in null cone $\iff$

logcap(A) $= -\infty$

A is not null cone iff

logcap(A) $> -\infty$

Conjugation Action - Manifold of Positive Definite Matrices

- Let us consider the conjugation action
- $G = \mathbb{GL}(n)$, $V = \mathrm{Mat}(n)$

$$(h, A) \mapsto hAh^{-1}$$

- Our optimization problem:

$$\inf_{h \in G} \log \|hAh^{-1}\|_F^2$$

- Our function:

$$f_A(h) = \log\left(\mathrm{tr}[hAh^{-1}h^{-T}A^T h^T]\right) = \log\left(\mathrm{tr}[XAX^{-1}A^T]\right) = f_A(X)$$

where $X = h^T h \in \mathbb{PD}(n)$

$PD(n) \cong GL(n)/U(n)$

$h \quad Uh$ (U unitary)

$f_A(h)$

$(Uh)^\dagger Uh = h^\dagger h$

$\inf_{X \in PD(n)} f_A(X)$

Conjugation Action - Manifold of Positive Definite Matrices

- Let us consider the conjugation action
- $G = \mathbb{GL}(n)$, $V = \mathrm{Mat}(n)$

$$(h, A) \mapsto hAh^{-1}$$

- Our optimization problem:

$$\inf_{h \in G} \log \|hAh^{-1}\|_F^2$$

- Our function:

$$f_A(h) = \log\left(\mathrm{tr}[hAh^{-1}h^{-T}A^T h^T]\right) = \log\left(\mathrm{tr}[XAX^{-1}A^T]\right) = f_A(X)$$

  where $X = h^T h \in \mathbb{PD}(n)$
- Naturally we obtain $f_A : \mathbb{PD}(n) \to \mathbb{R}$

Conjugation Action - Manifold of Positive Definite Matrices

- Let us consider the conjugation action
- $G = \mathbb{GL}(n)$, $V = \mathrm{Mat}(n)$

$$(h, A) \mapsto hAh^{-1}$$

- Our optimization problem:

$$\inf_{h \in G} \log \|hAh^{-1}\|_F^2$$

- Our function:

$$f_A(h) = \log\left(\mathrm{tr}[hAh^{-1}h^{-T}A^T h^T]\right) = \log\left(\mathrm{tr}[XAX^{-1}A^T]\right) = f_A(X)$$

  where $X = h^T h \in \mathbb{PD}(n)$
- Naturally we obtain $f_A : \mathbb{PD}(n) \to \mathbb{R}$
- What is so good about having a function from space of positive definite matrices?

Manifold of Positive Definite Matrices

- $\mathbb{PD}(n)$ is Riemannian submanifold of $\mathrm{Mat}(n)$
  1. real, smooth manifold
  2. tangent space given by *Hermitian matrices* $\mathrm{Her}(n)$
  3. positive definite inner product on tangent space

$$A, B \in \mathrm{Her}(n) \quad \langle A, B \rangle = \mathrm{tr}[AB]$$

Manifold of Positive Definite Matrices

- $\mathbb{PD}(n)$ is Riemannian submanifold of $\mathrm{Mat}(n)$
  1. real, smooth manifold
  2. tangent space given by *Hermitian matrices* $\mathrm{Her}(n)$
  3. positive definite inner product on tangent space

$$A, B \in \mathrm{Her}(n) \quad \langle A, B\rangle = \mathrm{tr}[AB]$$

- Geodesics given by

$$\exp_A(Z) = A^{1/2} \exp(Z) A^{1/2}$$

Hermitian

Manifold of Positive Definite Matrices

- $\mathbb{PD}(n)$ is Riemannian submanifold of $\mathrm{Mat}(n)$
  1. real, smooth manifold
  2. tangent space given by *Hermitian matrices* $\mathrm{Her}(n)$
  3. positive definite inner product on tangent space

$$A, B \in \mathrm{Her}(n) \quad \langle A, B \rangle = \mathrm{tr}[AB]$$

- Geodesics given by

$$\exp_A(Z) = A^{1/2} \exp(Z) A^{1/2}$$

- To go from A to B we use geodesic:

$$\gamma(t) = A^{1/2}(A^{-1/2} B A^{-1/2})^t A^{1/2}$$

$\gamma(0) = A \quad \gamma(1) = B$

Manifold of Positive Definite Matrices

- $\mathbb{PD}(n)$ is Riemannian submanifold of $\mathrm{Mat}(n)$
  1. real, smooth manifold
  2. tangent space given by *Hermitian matrices* $\mathrm{Her}(n)$
  3. positive definite inner product on tangent space

$$A, B \in \mathrm{Her}(n) \quad \langle A, B \rangle = \mathrm{tr}[AB]$$

- Geodesics given by

$$\exp_A(Z) = A^{1/2} \exp(Z) A^{1/2}$$

- To go from A to B we use geodesic:

$$\gamma(t) = A^{1/2}(A^{-1/2} B A^{-1/2})^t A^{1/2}$$

- Distance from A to B given by

$$\delta(A, B) = \| \log A - \log B \|_F$$

- Lie Groups, Lie Algebras & Positive Definite Matrices
- Crash Course on Geodesic Convex Optimization
- Analysis of Scaling Problem for Conjugation Action
- Conclusion

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\frac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$

$g_{a,b}(t)$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\frac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$
- Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the vector $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ such that:

$$\langle \nabla f(\mathbf{a}), \mathbf{b} \rangle = \partial_t f(\mathbf{a} + \mathbf{b} \cdot t)|_{t=0}$$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$
- Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the vector $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ such that:

$$\langle \nabla f(\mathbf{a}), \mathbf{b} \rangle = \partial_t f(\mathbf{a} + \mathbf{b} \cdot t)|_{t=0}$$

- Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the matrix $\nabla^2 f(\mathbf{a}) \in \mathbb{R}^{n \times n}$ such that:

$$\langle \mathbf{c}, \nabla^2 f(\mathbf{a}) \cdot \mathbf{b} \rangle = \partial_s \partial_t f(\mathbf{a} + \mathbf{b} \cdot t + \mathbf{c} \cdot s)|_{s,t=0}$$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\frac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$
- Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the vector $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ such that:

$$\langle \nabla f(\mathbf{a}), \mathbf{b} \rangle = \partial_t f(\mathbf{a} + \mathbf{b} \cdot t)|_{t=0}$$

- Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the matrix $\nabla^2 f(\mathbf{a}) \in \mathbb{R}^{n \times n}$ such that:

$$\langle \mathbf{c}, \nabla^2 f(\mathbf{a}) \cdot \mathbf{b} \rangle = \partial_s \partial_t f(\mathbf{a} + \mathbf{b} \cdot t + \mathbf{c} \cdot s)|_{s,t=0}$$

- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff $\nabla^2 f(\mathbf{a}) \succeq 0$ for all $\mathbf{a} \in \mathbb{R}^n$

Lipschitz, smooth and strong convex functions

- A map $F : \mathbb{R}^n \to \mathbb{R}^m$ is *L-Lipschitz* if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\|F(\mathbf{x}) - F(\mathbf{y})\|_2 \leq L \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

Lipschitz, smooth and strong convex functions

- A map $F : \mathbb{R}^n \to \mathbb{R}^m$ is *L-Lipschitz* if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\|F(\mathbf{x}) - F(\mathbf{y})\|_2 \leq L \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

- Function $f : \mathbb{R}^n \to \mathbb{R}$ *L-smooth* iff

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$

is L-Lipschitz.

Lipschitz, smooth and strong convex functions

- A map $F : \mathbb{R}^n \to \mathbb{R}^m$ is *L-Lipschitz* if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\|F(\mathbf{x}) - F(\mathbf{y})\|_2 \leq L \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

- Function $f : \mathbb{R}^n \to \mathbb{R}$ *L-smooth* iff

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$

  is L-Lipschitz.
- Equivalently, Hessian of $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla^2 f \preceq L \cdot I$$

Lipschitz, smooth and strong convex functions

- A map $F : \mathbb{R}^n \to \mathbb{R}^m$ is *L-Lipschitz* if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\|F(\mathbf{x}) - F(\mathbf{y})\|_2 \leq L \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

- Function $f : \mathbb{R}^n \to \mathbb{R}$ *L-smooth* iff

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$

  is L-Lipschitz.
- Equivalently, Hessian of $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla^2 f \preceq L \cdot I$$

- Function $f : \mathbb{R}^n \to \mathbb{R}$ *μ-strongly convex* iff

$$\nabla^2 f \succeq \mu \cdot I$$

Gradient descent - smooth functions

- **Input:** convex, L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{b} \in \mathbb{R}^n$ such that $\|\nabla f(\mathbf{b})\|_2 \leq \varepsilon$.

Gradient descent - smooth functions

- **Input:** convex, L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{b} \in \mathbb{R}^n$ such that $\|\nabla f(\mathbf{b})\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$ and $\eta < \dfrac{2}{L}$

↑ learning / descent parameter

Gradient descent - smooth functions

- **Input:** convex, L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{b} \in \mathbb{R}^n$ such that $\|\nabla f(\mathbf{b})\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$ and $\eta < \dfrac{2}{L}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \eta \cdot \nabla^{(k)}$

next step — current step — direction of gradient (steepest descent)

Gradient descent - smooth functions

- **Input:** convex, L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{b} \in \mathbb{R}^n$ such that $\|\nabla f(\mathbf{b})\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$ and $\eta < \frac{2}{L}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \eta \cdot \nabla^{(k)}$

1. f is L-smooth then Taylor + mean-value theorem, there is $\mathbf{z}$ in line from $\mathbf{x}$ to $\mathbf{y}$ such that:

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \cdot \langle \mathbf{y} - \mathbf{x}, \ \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \rangle \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2 \end{aligned}$$

$\preceq L \cdot I$

Gradient descent - smooth functions

1. f is L-smooth:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2$$

Gradient descent - smooth functions

1. f is L-smooth:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2$$

2. letting $\mathbf{y} = \mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \eta \nabla^{(k)}$ and $\mathbf{x} = \mathbf{x}^{(k)}$, we have:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - \eta \langle \nabla^{(k)}, \nabla^{(k)} \rangle + \frac{L}{2} \cdot \|\eta \nabla^{(k)}\|_2^2$$

$$= f(\mathbf{x}^{(k)}) - \eta \left(1 - \frac{\eta L}{2}\right) \cdot \|\nabla^{(k)}\|_2^2$$

$y - x = -\eta \nabla^{(k)}$

$\eta < \frac{2}{L} \Rightarrow 1 - \frac{\eta L}{2} > 0$

some constant

Gradient descent - smooth functions

1. f is L-smooth:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2$$

2. letting $\mathbf{y} = \mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \eta \nabla^{(k)}$ and $\mathbf{x} = \mathbf{x}^{(k)}$, we have:

$$\begin{aligned} f(\mathbf{x}^{(k+1)}) &\leq f(\mathbf{x}^{(k)}) - \eta \langle \nabla^{(k)}, \nabla^{(k)} \rangle + \frac{L}{2} \cdot \|\eta \nabla^{(k)}\|_2^2 \\ &= f(\mathbf{x}^{(k)}) - \eta \left(1 - \frac{\eta L}{2}\right) \cdot \|\nabla^{(k)}\|_2^2 \end{aligned}$$

3. Thus we have

$$\|\nabla^{(k)}\|_2^2 \leq \underbrace{\frac{1}{\eta \left(1 - \frac{\eta L}{2}\right)}}_{C_{\eta,L} := C} \cdot \left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^{(k+1)})\right)$$

Gradient descent - smooth functions

1. f is L-smooth:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2$$

2. letting $\mathbf{y} = \mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \eta \nabla^{(k)}$ and $\mathbf{x} = \mathbf{x}^{(k)}$, we have:

$$\begin{aligned} f(\mathbf{x}^{(k+1)}) &\leq f(\mathbf{x}^{(k)}) - \eta \langle \nabla^{(k)}, \nabla^{(k)} \rangle + \frac{L}{2} \cdot \|\eta \nabla^{(k)}\|_2^2 \\ &= f(\mathbf{x}^{(k)}) - \eta \left(1 - \frac{\eta L}{2}\right) \cdot \|\nabla^{(k)}\|_2^2 \end{aligned}$$

3. Thus we have

$$\|\nabla^{(k)}\|_2^2 \leq \frac{1}{\eta \left(1 - \frac{\eta L}{2}\right)} \cdot \left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^{(k+1)})\right)$$

4. Implies $\sum_{k=1}^{T} \|\nabla^{(k)}\|_2^2$ upper bounded. Thus $\|\nabla^{(k)}\|_2 \to 0$

$f(x^{(0)}) - f(x^{(T)}) \leq f(x^{(0)})$

Geodesic Convex Optimization Crash Course

- Recall the exponential map at $A \in \mathbb{PD}(n)$

geodesics through A → $$\exp_A(H) = A^{1/2} e^H A^{1/2}$$

↑ Hermitian

Geodesic Convex Optimization Crash Course

- Recall the exponential map at $A \in \mathbb{PD}(n)$

$$\exp_A(H) = A^{1/2} e^H A^{1/2}$$

- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ g-convex iff the univariate function

$$g_A(t) = f(\exp_A(t \cdot H))$$

is convex for every $\underbrace{A \in \mathbb{PD}(n)}_{\text{pt in manifold}}, \underbrace{H \in \mathrm{Her}(n)}_{\text{direction}}$

Geodesic Convex Optimization Crash Course

- Recall the exponential map at $A \in \mathbb{PD}(n)$

$$\exp_A(H) = A^{1/2} e^H A^{1/2}$$

- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ g-convex iff the univariate function

$$g_A(t) = f(\exp_A(t \cdot H))$$

is convex for every $A \in \mathbb{PD}(n), H \in \text{Her}(n)$

- Gradient of $f : \mathbb{PD}(n) \to \mathbb{R}$ at A is the vector $\nabla f(A) \in \text{Her}(n)$ such that:

$$\langle \nabla f(A), H \rangle = \partial_t f(\exp_A(t \cdot H))|_{t=0}$$

Geodesic Convex Optimization Crash Course

- Recall the exponential map at $A \in \mathbb{PD}(n)$

$$\exp_A(H) = A^{1/2} e^H A^{1/2}$$

- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ g-convex iff the univariate function

$$g_A(t) = f(\exp_A(t \cdot H))$$

  is convex for every $A \in \mathbb{PD}(n), H \in \mathrm{Her}(n)$
- Gradient of $f : \mathbb{PD}(n) \to \mathbb{R}$ at A is the vector $\nabla f(A) \in \mathrm{Her}(n)$ such that:

$$\langle \nabla f(A), H \rangle = \partial_t f(\exp_A(t \cdot H))|_{t=0}$$

- Hessian of $f : \mathbb{PD}(n) \to \mathbb{R}$ at A is the matrix $\nabla^2 f(A) \in \mathrm{Her}(n) \otimes \mathrm{Her}(n)$ such that:

$$\langle F, \nabla^2 f(A) \cdot H \rangle = \partial_s \partial_t f(\exp_A(t \cdot F + s \cdot H))|_{s,t=0}$$

Geodesic Convex Optimization Crash Course

- Recall the exponential map at $A \in \mathbb{PD}(n)$
$$\exp_A(H) = A^{1/2} e^H A^{1/2}$$
- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ g-convex iff the univariate function
$$g_A(t) = f(\exp_A(t \cdot H))$$
is convex for every $A \in \mathbb{PD}(n), H \in \mathrm{Her}(n)$
- Gradient of $f : \mathbb{PD}(n) \to \mathbb{R}$ at A is the vector $\nabla f(A) \in \mathrm{Her}(n)$ such that:
$$\langle \nabla f(A), H \rangle = \partial_t f(\exp_A(t \cdot H))|_{t=0}$$
- Hessian of $f : \mathbb{PD}(n) \to \mathbb{R}$ at A is the matrix $\nabla^2 f(A) \in \mathrm{Her}(n) \otimes \mathrm{Her}(n)$ such that:
$$\langle F, \nabla^2 f(A) \cdot H \rangle = \partial_s \partial_t f(\exp_A(t \cdot F + s \cdot H))|_{s,t=0}$$
- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ g-convex iff $\nabla^2 f(A) \succeq 0$ for all $A \in \mathbb{PD}(n)$

Smooth and strong g-convex functions

- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ *L-smooth* iff

$$\nabla^2 f \preceq L \cdot I$$

Smooth and strong g-convex functions

- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ *L-smooth* iff

$$\nabla^2 f \preceq L \cdot I$$

- Function $f : \mathbb{PD}(n) \to \mathbb{R}$ *μ-strongly convex* iff

$$\nabla^2 f \succeq \mu \cdot I$$

Gradient descent - geodesically smooth functions

- **Input:** g-convex, L-smooth function $f : \mathbb{PD}(n) \to \mathbb{R}$, $\varepsilon > 0$, initial point $A \in \mathbb{PD}(n)$
- **Output:** Find point $B \in \mathbb{PD}(n)$ such that $\|\nabla f(B)\| \leq \varepsilon$.

Gradient descent - geodesically smooth functions

- **Input:** g-convex, L-smooth function $f : \mathbb{PD}(n) \to \mathbb{R}$, $\varepsilon > 0$, initial point $A \in \mathbb{PD}(n)$
- **Output:** Find point $B \in \mathbb{PD}(n)$ such that $\|\nabla f(B)\|_2 \leq \varepsilon$.
- Start with your initial point $A^{(0)} = A$ and $\eta < \frac{2}{L}$

Gradient descent - geodesically smooth functions

- **Input:** g-convex, L-smooth function $f : \mathbb{PD}(n) \to \mathbb{R}$, $\varepsilon > 0$, initial point $A \in \mathbb{PD}(n)$
- **Output:** Find point $B \in \mathbb{PD}(n)$ such that $\|\nabla f(B)\|_2 \leq \varepsilon$.
- Start with your initial point $A^{(0)} = A$ and $\eta < \dfrac{2}{L}$
- Let $\nabla^{(k)} := \nabla f(A^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $A^{(k+1)} = \underbrace{\exp_{A^{(k)}}(-\eta \cdot \nabla^{(k)})}$

$\Leftrightarrow x^{(k)} - \eta \nabla^{(k)}$

but now geodesic through $A^{(k)}$ in direction $-\nabla^{(k)}$

Gradient descent - geodesically smooth functions

- **Input:** g-convex, L-smooth function $f : \mathbb{PD}(n) \to \mathbb{R}$, $\varepsilon > 0$, initial point $A \in \mathbb{PD}(n)$
- **Output:** Find point $B \in \mathbb{PD}(n)$ such that $\|\nabla f(B)\|_2 \leq \varepsilon$.
- Start with your initial point $A^{(0)} = A$ and $\eta < \dfrac{2}{L}$
- Let $\nabla^{(k)} := \nabla f(A^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $A^{(k+1)} = \exp_{A^{(k)}}(-\eta \cdot \nabla^{(k)})$
- Same analysis goes through.

- Lie Groups, Lie Algebras & Positive Definite Matrices
- Crash Course on Geodesic Convex Optimization
- Analysis of Scaling Problem for Conjugation Action
- Conclusion

Conjugation Action

- $G = \mathbb{GL}(n),\ V = \mathrm{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \mathrm{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*

- Conjugation action scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$
  - **Output:** Is A nilpotent?

$A \in \mathrm{Mat}(n)$, $\varepsilon > 0$ $f_A : PD(n) \to \mathbb{R}$

output $B \in PD(n)$ (if exists) s.t.

$\|\nabla f_A(B)\| \leq \varepsilon$

ε small enough $\Rightarrow$ solution to null-cone problem

Simultaneous conjugation action

$G = GL(n)$ $\qquad V = Mat(n)^m$

$(A_1, \ldots, A_m)$

$$X \circ (A_1, \ldots, A_m) = (XA_1X^{-1}, \ldots, XA_mX^{-1})$$

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*

- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?
- Our function

$$f_A(h) = \log \|hAh^{-1}\|_F = \text{tr}[XAX^{-1}A] = f_A(X)$$

can be thought of as $f_A : \mathbb{PD}(n) \to \mathbb{R}$

$X = h^\dagger h$

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*

- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?
- Our function

$$f_A(h) = \log \|hAh^{-1}\|_F = \text{tr}[XAX^{-1}A]$$

  can be thought of as $f_A : \mathbb{PD}(n) \to \mathbb{R}$

- $\inf_{X \in \mathbb{PD}(n)} f_A(X)$ exists iff A *not in nullcone*!

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*

- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?
- Our function

$$f_A(h) = \log \|hAh^{-1}\|_F = \text{tr}[XAX^{-1}A]$$

  can be thought of as $f_A : \mathbb{PD}(n) \to \mathbb{R}$

- $\inf_{X \in \mathbb{PD}(n)} f_A(X)$ exists iff A *not in nullcone*!
- Is this function convex? Smooth?

Is this function nice?

Equivariance & Gradient at each point

$$A \qquad (g, A) \qquad \boxed{Y = g \circ X}$$

$$f_A(X) = f_A(\qquad \boxed{f_A(g \circ X) = f_{g \circ A}(X)}$$

$$X = h h^+ \qquad \underbrace{\overbrace{g X g^+}^{Y}}_{g h h^+ g^+}$$

$$g X g^+ = Y$$

$$\|(gh)^{-1} A (gh)\| = tr\left[(gh)^{-1} A (gh)(gh)^+ A^+ (gh)^{-+}\right]$$

$$= tr\left[h (g^{-1} A g) h h^+ (g^{-1} A g)^+ h^{-+}\right]$$

$$= f_A(Y) \qquad = f_{g \circ A}(X)$$

Equivariance allows us to do:

understanding our function at I $\Rightarrow$ understand our function everywhere

Computing the Gradient

$$\nabla_A := \nabla f_A(I)$$

$$\langle \nabla_A, H \rangle = \partial_t f_A(e^{tH}) \Big|_{t=0}$$

$\exp_I(Ht)$

$$= \partial_t \log \operatorname{tr}\left[e^{tH} A e^{-tH} A^+\right] \Big|_{t=0}$$

$$= \frac{\operatorname{tr}\left[H e^{tH} A e^{-tH} A^+ - e^{tH} A H e^{-tH} A^+\right]}{\operatorname{tr}\left[e^{tH} A e^{-tH} A^+\right]} \Bigg|_{t=0}$$

$$\nabla_A = \frac{1}{\|A\|_F^2} \cdot (AA^+ - A^+A)$$

$$= \frac{\operatorname{tr}\left[AA^+H - A^+AH\right]}{\|A\|_F^2} = \frac{1}{\|A\|_F^2} \cdot \langle AA^+ - A^+A, H \rangle$$

Proving g-convexity

tangent space

- Given any direction $H \in \mathrm{Her}(n)$, need to show

$$\partial_t^2 f_A(e^{tH})|_{t=0} \geq 0$$

$$\partial_t f_A = \frac{\mathrm{tr}\left[H e^{tH} A e^{-tH} A^+ - e^{tH} A H e^{-tH} A^+\right]}{\mathrm{tr}\left[e^{tH} A e^{-tH} A^+\right]}$$

$$\partial_t^2 f_A \Big|_{t=0} = \frac{\|A\|^2 \cdot \mathrm{tr}\left[H^2 A A^+ - HAHA^+ - HAHA^+ + AH^2A^+\right]}{\|A\|_F^4} - \frac{\left(\mathrm{tr}\left[HAA^+ - AHA^+\right]\right)^2}{\|A\|_F^4}$$

$$\partial_t^2 f_A \Big|_{t=0} = \frac{\|A\|^2 \cdot \mathrm{tr}\left[H^2 A A^+ - \underline{HA}\,\underline{HA}^+ - HAHA^+ + AH^2A^+\right]}{\|A\|_F^4}$$

$$- \frac{\left(\mathrm{tr}\left[HAA^+ - AHA^+\right]\right)^2}{\|A\|_F^4}$$

$$= \frac{1}{\|A\|^2} \cdot \mathrm{tr}\Big[\underbrace{(HA - AH)}_{B}\underbrace{(A^+H - HA^+)}_{B^+}\Big]$$

$$- \frac{1}{\|A\|^4}\left(\mathrm{tr}\left[BB^+\right]\right) - \frac{1}{\|A\|^4}\left(\mathrm{tr}\left[HAA^+ - HA^+A\right]\right)^2$$

$$\gtrless 0$$

Conjugation Action - what is gradient descent doing?

$$\nabla_A = \frac{\ell}{\|A\|^2} \cdot \left(AA^{+} - A^{+}A\right)$$

orbit of $A^{(0)}$

$\downarrow$

$$M_k = \frac{A^{(n)}}{\|A^{(n)}\|_F}$$

$$\nabla_A \to 0$$

$$M_k M_k^{+} - M_k^{+} M_k \to 0$$

$$M_k M_k^{+} = M_k^{+} M_k \quad \leftarrow \text{ definition normal matrix}$$

$$M_k \in G \cdot A \qquad \nabla_{M_k} \to 0 \iff A \text{ similar to normal matrix}$$

Conjugation Action - Thoughts

need to show that

$\exists\ \varepsilon(n)$ s.t.

$\|\nabla_{A(n)}\| < \varepsilon(n)$

$\Rightarrow$ A not nilpotent.

also need to show that

$f_1(x)$ is L-smooth for some

L

Conclusion

- Today we learned the about scaling algorithms for non-commutative groups
- Geodesic Convexity
- Gradient descent algorithm for g-convex problems
- Still to see: how representation theory allows us to finish the analysis of the nullcone problem for the conjugation action (weight norms and weight margins)

what ε to choose? ← Which ε should we choose? robustness L-smooth