CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 15, 2021

Yale University

Last time

- Optimization on $\mathbb{R}^n$
- Continuous time: Gradient flow
- Discrete time: Gradient descent and proximal gradient
- Strong convexity $\Rightarrow$ Gradient dominated $\Rightarrow$ Sufficient growth
- Exponential convergence rate (with smoothness)

Today: Optimization on Manifold

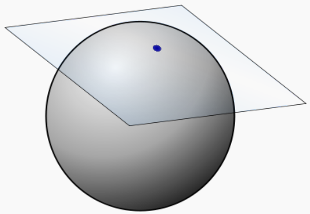
References

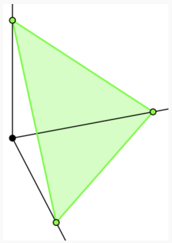
- Lee, *Introduction to Riemannian Manifolds*, 2nd ed, Springer, 2018
- Zhang & Sra, *First-order methods for geodesically convex optimization*, COLT, 2016
- ◦ Sra, *Some non-convex optimization problems through a geometric lens*, Harvard talk, 2019, `https://www.youtube.com/watch?v=ys2XPPijoDA`
- ◦ Vishnoi, *An Introduction to Geodesic Convexity*, IAS talk, 2018, `https://www.youtube.com/watch?v=hJdcd1SR_tA`
- Vishnoi, *Geodesic Convex Optimization: Differentiation on Manifolds, Geodesics, and Convexity*, `https://arxiv.org/pdf/1806.06373.pdf`
- Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018, Appendix C

Manifold

Manifold

A **manifold** $\mathcal{X}$ is a set which locally looks like Euclidean space

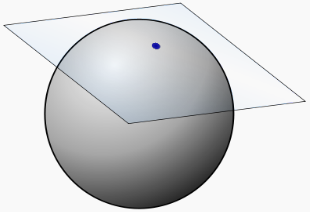


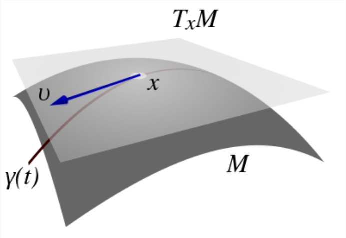


- $\mathcal{X} = \mathbb{R}^n$
- $\mathcal{X} = \mathbb{R}^n_+$ $= \{x \in \mathbb{R}^n \mid x_i > 0 \text{ for } i = 1, \dots, n\}$
- $\mathcal{X} = \Delta_{n-1}$ $= \{x \in \mathbb{R}^n \mid x_1 + \dots + x_n = 1, \; x_i \geq 0\}$
- $\mathcal{X} = \mathbb{S}_{n-1}$ $= \{x \in \mathbb{R}^n \mid x_1^2 + \dots + x_n^2 = 1\}$
- $\mathcal{X} =$ PSD matrices $= \{x \in \mathbb{R}^{n \times n} \mid x = x^T \succ 0\}$
- $\mathcal{X} =$ orthogonal matrices $= \{x \in \mathbb{R}^n \mid x \cdot x^T = I\}$

Tangent space

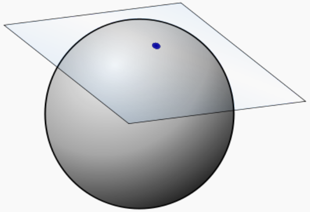
Every point $x \in \mathcal{X}$ has a *tangent space* $\mathrm{T}_x\mathcal{X}$ consisting of tangent vectors v (directions of motion such that $x + tv \in \mathcal{X}$ for small t)

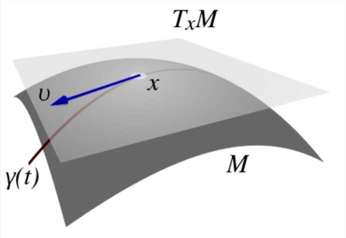




Tangent space

Every point $x \in \mathcal{X}$ has a *tangent space* $\mathrm{T}_x\mathcal{X}$ consisting of tangent vectors v (directions of motion such that $x + tv \in \mathcal{X}$ for small t)

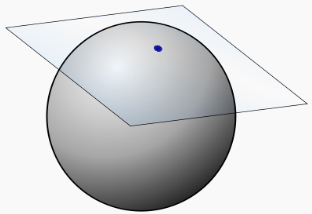


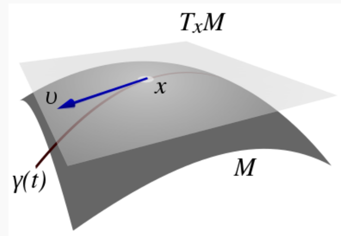


- $\mathcal{X} = \mathbb{R}^n$, $\mathrm{T}_x\mathcal{X} =$ $\mathbb{R}^n$
- $\mathcal{X} = \mathbb{R}^n_+$, $\mathrm{T}_x\mathcal{X} =$ $\mathbb{R}^n$
- $\mathcal{X} = \Delta_{n-1}$, $\mathrm{T}_x\mathcal{X} =$
- $\mathcal{X} = \mathbb{S}_{n-1}$, $\mathrm{T}_x\mathcal{X} =$
- $\mathcal{X} =$ PSD matrices, $\mathrm{T}_x\mathcal{X} =$
- $\mathcal{X} =$ orthogonal matrices, $\mathrm{T}_x\mathcal{X} =$

Tangent space

Every point $x \in \mathcal{X}$ has a *tangent space* $\mathrm{T}_x\mathcal{X}$ consisting of tangent vectors v (directions of motion such that $x + tv \in \mathcal{X}$ for small t)





- $\mathcal{X} = \mathbb{R}^n$, $\mathrm{T}_x\mathcal{X} = \mathbb{R}^n$
- $\mathcal{X} = \mathbb{R}^n_+$, $\mathrm{T}_x\mathcal{X} = \mathbb{R}^n$
- $\mathcal{X} = \Delta_{n-1}$, $\mathrm{T}_x\mathcal{X} = \{v \in \mathbb{R}^n \colon 1^\top v = 0\}$
- $\mathcal{X} = \mathbb{S}_{n-1}$, $\mathrm{T}_x\mathcal{X} = \{v \in \mathbb{R}^n \colon x^\top v = 0\}$
- $\mathcal{X} =$ PSD matrices, $\mathrm{T}_x\mathcal{X} =$ symmetric matrices
- $\mathcal{X} =$ orthogonal matrices, $\mathrm{T}_x\mathcal{X} =$ skew-symmetric matrices

$x \in S_{n-1}$, $x_1^2 + \dots + x_n^2 = 1$ (*)

$x + tv \in S_{n-1}$

$\Leftrightarrow (x_1+tv_1)^2 + \dots + (x_n+tv_n)^2 = 1$

$\Leftrightarrow \cancel{(x_1^2+\dots+x_n^2)}^{\,1} + 2t(x_1v_1+\dots+x_nv_n) + \cancel{t^2}^{\,0}(v_1^2+\dots+v_n^2) = \cancel{1}$

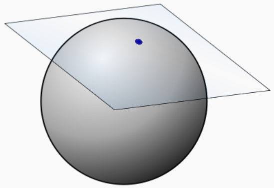
$\Leftrightarrow x_1v_1 + \dots + x_nv_n = 0$

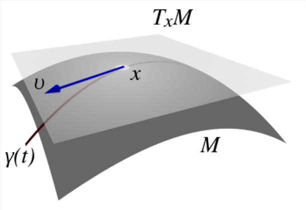
Metric on manifold

Every tangent space $\mathrm{T}_x\mathcal{X}$ has a Riemannian *metric*
$g_x : \mathrm{T}_x\mathcal{X} \times \mathrm{T}_x\mathcal{X} \to \mathbb{R}$ which is symmetric and positive definite

defines inner product: $\langle u, v\rangle_x = g_x(u, v)$

and square norm: $\|v\|_x^2 = g_x(v, v) \geq 0$





If $\mathcal{X} \subseteq \mathbb{R}^n$, g_x can be represented by a matrix $g(x) \in \mathbb{R}^{n\times n}$ $\succ 0$:

$$\|v\|_x^2 = v^\top g(x) v \qquad \forall\, v \in T_x\mathcal{X} \subseteq \mathbb{R}^n$$

- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric: $g(x) = I$

$$\langle u, v \rangle_x = \sum_{i=1}^{n} u_i v_i$$

- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric: $g(x) = \mathrm{diag}\left(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2}\right)$

$$\langle u, v \rangle_x = \sum_{i=1}^{n} \frac{u_i v_i}{x_i^2}$$

- $\mathcal{X} = \Delta_{n-1}$, Fisher metric: $g(x) = \mathrm{diag}\left(\frac{1}{x_1}, \ldots, \frac{1}{x_n}\right)$

$$\langle u, v \rangle_x = \sum_{i=1}^{n} \frac{u_i v_i}{x_i}$$

- $\mathcal{X} =$ PSD matrices, log-det metric:

$$\langle u, v \rangle_x = \mathrm{Tr}(x^{-1} u \, x^{-1} v)$$

Hessian manifold

A manifold $\mathcal{X} \subseteq \mathbb{R}^n$ is a **Hessian manifold** if the metric $g(x)$ is the Hessian of a convex function $\phi \colon \mathcal{X} \to \mathbb{R}$:

$$g(x) = \nabla^2 \phi(x)$$

Hessian manifold

A manifold $\mathcal{X} \subseteq \mathbb{R}^n$ is a **Hessian manifold** if the metric $g(x)$ is the Hessian of a convex function $\phi \colon \mathcal{X} \to \mathbb{R}$:

$$g(x) = \nabla^2 \phi(x)$$

- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric $g(x) = I$ is Hessian of $\phi(x) = \frac{1}{2}\|x\|^2$
- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric $g(x) = \mathrm{diag}(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2})$ is Hessian of

log-barrier $\quad \phi(x) = -\sum_{i=1}^{n} \log x_i$

- $\mathcal{X} = \Delta_{n-1}$, Fisher metric $g(x) = \mathrm{diag}(\frac{1}{x_1}, \ldots, \frac{1}{x_n})$ is Hessian of

(negative) entropy $\quad \phi(x) = \sum_{i=1}^{n} x_i \log x_i$

- $\mathcal{X} =$ PSD matrices, log-det metric is Hessian of

$$\phi(x) = -\log\det x = -\sum_{i=1}^{n} \log \lambda_i(x)$$

Hessian manifold

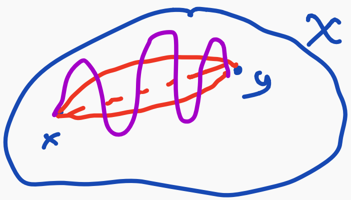
A manifold $\mathcal{X} \subseteq \mathbb{R}^n$ is a **Hessian manifold** if the metric $g(x)$ is the Hessian of a convex function $\phi\colon \mathcal{X} \to \mathbb{R}$:

$$g(x) = \nabla^2 \phi(x)$$

- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric $g(x) = I$ is Hessian of $\phi(x) = \frac{1}{2}\|x\|^2$
- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric $g(x) = \mathrm{diag}(\frac{1}{x_1^2}, \dots, \frac{1}{x_n^2})$ is Hessian of
$$\phi(x) = -\sum_{i=1}^n \log x_i$$
- $\mathcal{X} = \Delta_{n-1}$, Fisher metric $g(x) = \mathrm{diag}(\frac{1}{x_1}, \dots, \frac{1}{x_n})$ is Hessian of
$$\phi(x) = \sum_{i=1}^n x_i \log x_i$$
- $\mathcal{X} =$ PSD matrices, log-det metric is Hessian of
$$\phi(x) = -\log \det x$$

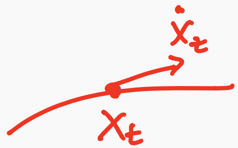
Distance on manifold

The distance between two points $x, y \in \mathcal{X}$ is



$$d(x, y) = \inf_X \int_0^1 \|\dot{X}_t\|_{X_t} \, dt$$

$$= \sqrt{\inf_X \int_0^1 \|\dot{X}_t\|_{X_t}^2 \, dt}$$



where infimum is over curves $X = (X_t)$ from $X_0 = x$ to $X_1 = y$

- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric $g(x) = I$

- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric $g(x) = \text{diag}(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2})$

- $\mathcal{X} =$ PSD matrices, log-det metric

- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric $g(x) = I$

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric $g(x) = \mathrm{diag}(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2})$

$$d(x, y) = \|\log x - \log y\|_2 = \sqrt{\sum_{i=1}^{n}(\log x_i - \log y_i)^2}$$

- $\mathcal{X} =$ PSD matrices, log-det metric

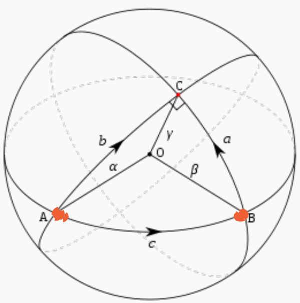
$$d(x, y) = \|\log(x^{-1}y)\|_{\mathrm{HS}} = \sqrt{\sum_{i=1}^{n}(\log \lambda_i(x^{-1}y))^2}$$

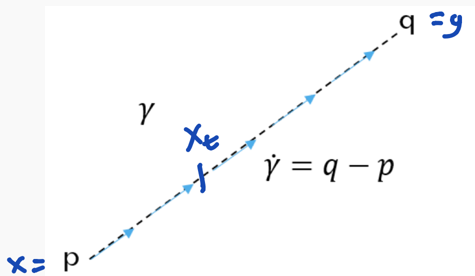
Geodesic on manifold

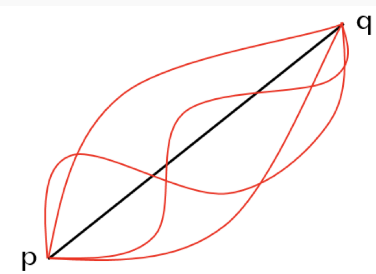
A **geodesic** is a (locally) shortest curve (X_t) between $x, y \in \mathcal{X}$

$$X_0 = x \quad \to \quad X_1 = y$$

$$d(x, X_t) = t\, d(x, y)$$







- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric $g(x) = I$

$$X_t = x + t(y - x) = (1 - t)x + ty$$

- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric $g(x) = \mathrm{diag}(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2})$

$$X_t = x \circ (y/x)^t = \left(x_1^{1-t} y_1^t, \ldots, x_n^{1-t} y_n^t\right)$$

- $\mathcal{X} = $ PSD matrices, log-det metric

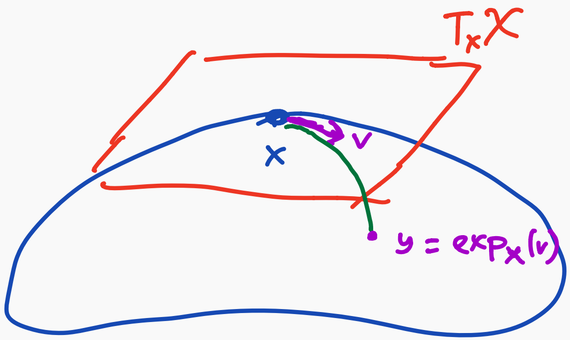
$$X_t = x^{1/2} \left(x^{-1/2} y x^{-1/2}\right)^t x^{1/2}$$

Exponential map

The **exponential map** at $x \in \mathcal{X}$ is

$$\exp_x : \mathrm{T}_x\mathcal{X} \to \mathcal{X}$$

that maps $v \in \mathrm{T}_x\mathcal{X}$ to $y = \exp_x(v) \in \mathcal{X}$ which is the position of the geodesic (X_t) at time $t = 1$ starting from $X_0 = x$, $\dot{X}_0 = v$



Exponential map

The **exponential map** at $x \in \mathcal{X}$ is

$$\exp_x \colon \mathrm{T}_x\mathcal{X} \to \mathcal{X}$$

that maps $v \in \mathrm{T}_x\mathcal{X}$ to $y = \exp_x(v) \in \mathcal{X}$ which is the position of the geodesic (X_t) at time $t = 1$ starting from $X_0 = x$, $\dot{X}_0 = v$

- The *logarithm map* is the inverse exponential map:

$$\log_x \colon \mathcal{X} \to \mathrm{T}_x\mathcal{X}$$

$$v = \log_x(y) \quad \Leftrightarrow \quad y = \exp_x(v)$$

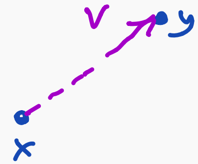
- Distance is $d(x, y) = \|\log_x(y)\|_x$

- $\mathcal{X} = \mathbb{R}^n$, Euclidean metric $g(x) = I$

$x = (\log p_1, \dots, \log p_n)$

$$\exp_x(v) = x + v$$

$$\log_x(y) = y - x$$



$p = (p_1, \dots, p_n)$

- $\mathcal{X} = \mathbb{R}^n_+$, log-barrier metric $g(x) = \mathrm{diag}(\frac{1}{x_1^2}, \dots, \frac{1}{x_n^2})$

$$\exp_x(v) = x \circ e^{v/x} = \left(x_1 e^{v_1/x_1}, \dots, x_n e^{v_n/x_n}\right)$$

$$\log_x(y) = x \circ \log(y/x) = \left(x_1 \log \frac{y_1}{x_1}, \dots, x_n \log \frac{y_n}{x_n}\right)$$

- $\mathcal{X} =$ PSD matrices, log-det metric

$$\exp_x(v) = x^{1/2} \exp(x^{-1/2} v\, x^{-1/2})\, x^{1/2}$$

$$\log_x(y) = x^{1/2} \log(x^{-1/2} y x^{-1/2})\, x^{1/2}$$

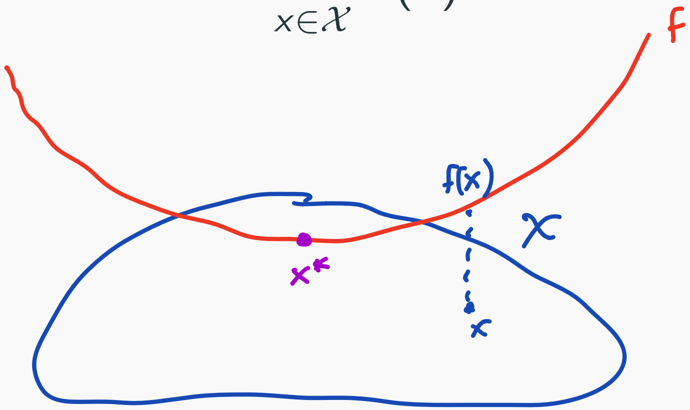
Optimization on Manifold

Optimization on manifold

Given a manifold $\mathcal{X}$ and an objective function $f : \mathcal{X} \to \mathbb{R}$

Want to solve

$$\min_{x \in \mathcal{X}} f(x)$$



Optimization on manifold

Given a manifold $\mathcal{X}$ and an objective function $f\colon \mathcal{X} \to \mathbb{R}$

Want to solve

$$\min_{x \in \mathcal{X}} f(x)$$

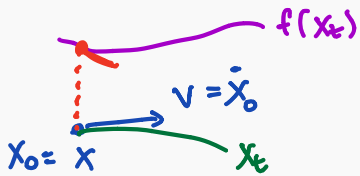
- Hosseini & Sra, *Matrix Manifold Optimization for Gaussian Mixtures*, NeurIPS 2015
- Sra, Vishnoi, & Yildiz, *On geodesically convex formulations for the Brascamp-Lieb constant*, APPROX/RANDOM 2018
- Allen-Zhu, Garg, Li, Oliveira, & Wigderson, *Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing*, STOC 2018

The **differential** of a function $f : \mathcal{X} \to \mathbb{R}$ at $x \in \mathcal{X}$ is a *cotangent vector* $df_x \in (\mathrm{T}_x\mathcal{X})^*$ which is a linear functional

$$df_x : \mathrm{T}_x\mathcal{X} \to \mathbb{R}$$

which gives directional derivative along any direction $v \in \mathrm{T}_x\mathcal{X}$

$$df_x(v) = \frac{d}{dt} f(x + tv)\Big|_{t=0} = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} v_i \quad = \langle \nabla f(x), v \rangle_{\ell_2}$$



where

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

Gradient on manifold

The **gradient** of $f : \mathcal{X} \to \mathbb{R}$ at $x \in \mathcal{X}$ is the tangent vector

$$\mathrm{grad}\, f(x) \in \mathsf{T}_x\mathcal{X}$$

corresponding to the differential df_x, so

$$df_x(v) = g_x(\mathrm{grad}\, f(x), v)$$

Gradient on manifold

The **gradient** of $f : \mathcal{X} \to \mathbb{R}$ at $x \in \mathcal{X}$ is the tangent vector

$$\operatorname{grad} f(x) \in \mathrm{T}_x\mathcal{X}$$

corresponding to the differential df_x, so

$$(\nabla f(x))^T v \quad = \quad df_x(v) = g_x(\operatorname{grad} f(x), v) = \underbrace{(\operatorname{grad} f(x))^T g(x)}_{\therefore\ = \nabla f(x)^T} v$$

- If $\mathcal{X} \subseteq \mathbb{R}^n$ and $g_x = g(x)$, then

$$\operatorname{grad} f(x) = g(x)^{-1} \nabla f(x)$$

where $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$

Gradient flow on manifold

The **gradient flow** of $f : \mathcal{X} \to \mathbb{R}$ is

$$\frac{d}{dt} X_t = \dot{X}_t = -\mathrm{grad}\, f(X_t) \quad \in T_{X_t}\mathcal{X}$$

- Descent flow:

$$\tfrac{d}{dt} f(X_t) = \langle \mathrm{grad}\, f(X_t), \dot{X}_t \rangle_{X_t} = -\|\mathrm{grad}\, f(X_t)\|_{X_t}^2 \leq 0$$

Gradient flow on manifold

The **gradient flow** of $f : \mathcal{X} \to \mathbb{R}$ is

$$\lim_{\eta \to 0} \frac{X_{t+\eta} - X_t}{\eta} = \frac{d}{dt} X_t \quad = \quad \dot{X}_t = -\mathrm{grad}\, f(X_t)$$

- Descent flow:

$$\tfrac{d}{dt} f(X_t) = \langle \mathrm{grad}\, f(X_t), \dot{X}_t \rangle_{X_t} = -\|\mathrm{grad}\, f(X_t)\|_{X_t}^2 \le 0$$

- If $\mathcal{X} \subseteq \mathbb{R}^n$ and $g_x = g(x)$, then gradient flow is

$$\dot{X}_t = -g(X_t)^{-1} \nabla f(X_t)$$

and $\frac{d}{dt} f(X_t) = -\nabla f(X_t)^\top g(X_t)^{-1} \nabla f(X_t)$

Natural gradient flow on Hessian manifold

If $\mathcal{X} \subseteq \mathbb{R}^n$ is a Hessian manifold with $g(x) = \nabla^2\phi(x)$, then **natural gradient flow**:

$$\dot{X}_t = -(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)$$

- When discretized in $\mathbb{R}^n$, this gives *natural gradient descent*:

$$x_{k+1} = x_k - \eta\,(\nabla^2\phi(x_k))^{-1}\nabla f(x_k)$$

- Hoffman, Blei, Wang, & Paisley, *Stochastic variational inference*, JMLR, 2013
- Amari, *Natural gradient works efficiently in learning*, Neural Computation, 1998

Mirror flow on Hessian manifold

If $\mathcal{X} \subseteq \mathbb{R}^n$ is a Hessian manifold with $g(x) = \nabla^2\phi(x)$, then

natural gradient flow $\dot{X}_t = -(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)$ is equivalent to

the **Mirror flow** in dual variable $Y_t = \nabla\phi(X_t) \in \mathbb{R}^n$:

$$\Leftrightarrow \quad X_t = \nabla\phi^*(Y_t)$$

$$\dot{Y}_t = \underbrace{\frac{d}{dt}\nabla\phi(X_t)}_{\nabla^2\phi(X_t)\cdot\dot{X}_t} = -\nabla f(X_t) = -\nabla f(\nabla\phi^*(Y_t))$$

$$\nabla^2\phi(X_t)\cdot \dot{X}_t = -\nabla f(X_t)$$

$$\Leftrightarrow \quad \dot{X}_t = -(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)$$

Mirror flow on Hessian manifold

If $\mathcal{X} \subseteq \mathbb{R}^n$ is a Hessian manifold with $g(x) = \nabla^2 \phi(x)$, then

natural gradient flow $\dot{X}_t = -(\nabla^2 \phi(X_t))^{-1} \nabla f(X_t)$ is equivalent to

the **Mirror flow** in dual variable $Y_t = \nabla \phi(X_t) \in \mathbb{R}^n$:

$$\dot{Y}_t = \tfrac{d}{dt} \nabla \phi(X_t) = -\nabla f(X_t)$$

- Pushforward of natural gradient flow under mirror map $y = \nabla \phi(x)$ with pushforward metric $\nabla^2 \phi^*(y) = (\nabla^2 \phi(x))^{-1}$

Mirror flow on Hessian manifold

Mirror flow in dual variable $Y_t = \nabla\phi(X_t) \in \mathbb{R}^n$:

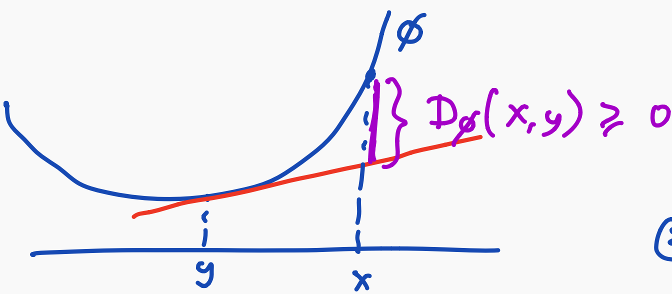
$$\dot{Y}_t = \frac{d}{dt}\nabla\phi(X_t) = -\nabla f(X_t)$$

- When discretized in $\mathbb{R}^n$, this gives *mirror descent*:

$$\nabla\phi(x_{k+1}) = \nabla\phi(x_k) - \eta\nabla f(x_k)$$

$$\Leftrightarrow \quad x_{k+1} = \arg\min_{x\in\mathbb{R}^n}\left\{\langle\nabla f(x_k), x - x_k\rangle + \frac{1}{\eta}D_\phi(x, x_k)\right\}$$

where $D_\phi(x,y) = \phi(x) - \phi(y) - \langle\nabla\phi(y), x - y\rangle$ is *Bregman divergence*



eg ① $\phi(x) = \frac{1}{2}\|x\|^2$

$D_\phi(x,y) = \frac{1}{2}\|x-y\|^2$

② $\phi(x) = \sum_{i=1}^{n} x_i \log x_i$

$D_\phi(x,y) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i}$

Mirror flow on Hessian manifold

Mirror flow in dual variable $Y_t = \nabla\phi(X_t) \in \mathbb{R}^n$:

$$\dot{Y}_t = \tfrac{d}{dt}\nabla\phi(X_t) = -\nabla f(X_t)$$

- When discretized in $\mathbb{R}^n$, this gives *mirror descent*:

$$\begin{aligned} &\nabla\phi(x_{k+1}) = \nabla\phi(x_k) - \eta\nabla f(x_k) \\ \Leftrightarrow \quad & x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D_\phi(x, x_k) \right\} \end{aligned}$$

  where $D_\phi(x,y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle$ is *Bregman divergence*

- *Multiplicative weight* on $\mathcal{X} = \Delta_{n-1}$ is mirror descent with $\phi =$ entropy

  [Arora, Hazan & Kale, *The Multiplicative Weights Update Method: A Meta-Algorithm and Applications*, Theory of Computing, 2012]

Gradient descent on manifold

Gradient descent on manifold

The **gradient descent** of $f \colon \mathcal{X} \to \mathbb{R}$ with step size $\eta > 0$ is

$$\begin{aligned} x_{k+1} &= \exp_{x_k}(-\eta \,\mathrm{grad}\, f(x_k)) \\ &= \arg\min_{x \in \mathcal{X}} \left\{ \langle \mathrm{grad}\, f(x_k), \log_{x_k}(x) \rangle_{x_k} + \frac{1}{2\eta} d(x, x_k)^2 \right\} \end{aligned}$$

- On $\mathcal{X} = \mathbb{R}^n$ with Euclidean metric, this is the usual gradient descent

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

- In general difficult to compute, can approximate via retraction

Proximal method on manifold

The **proximal method** of $f\colon \mathcal{X} \to \mathbb{R}$ with step size $\eta > 0$ is

$$\begin{aligned} x_{k+1} &= \exp_{x_k}(-\eta \operatorname{grad} f(x_{k+1})) \\ &= \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2\eta} d(x, x_k)^2 \right\} \end{aligned}$$

- On $\mathcal{X} = \mathbb{R}^n$ with Euclidean metric, this is the proximal method

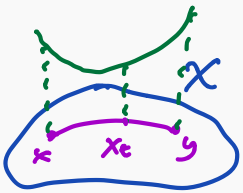
$$x_{k+1} = x_k - \eta \nabla f(x_{k+1})$$

- In general difficult to compute, need to solve implicit update, but more stable than gradient descent

Geodesic convexity

Geodesic convexity

A function $f\colon \mathcal{X} \to \mathbb{R}$ is **geodesically convex** if it is convex along any geodesic (X_t):



$\Leftrightarrow$ $\tilde{f}\colon [0,1] \to \mathbb{R}$, $\tilde{f}(t) = f(X_t)$ is convex (in the usual sense)

$\Leftrightarrow$ $(1-t)f(x) + tf(y) \geq f(X_t)$ for all $0 \leq t \leq 1$, $x, y \in \mathcal{X}$

$\Leftrightarrow$ $f(\exp_x(v)) \geq f(x) + \langle \operatorname{grad} f(x), v \rangle_x$ for all $x \in \mathcal{X}$, $v \in \mathrm{T}_x\mathcal{X}$

$\Leftrightarrow$ $\operatorname{Hess} f(x) \succeq 0$ for all $x \in \mathcal{X}$

where the *Hessian* $\operatorname{Hess} f(x)\colon \mathrm{T}_x\mathcal{X} \times \mathrm{T}_x\mathcal{X} \to \mathbb{R}$ is the quadratic form

$$\operatorname{Hess} f(x)[v,v] = \frac{d^2}{dt^2} f(\exp_x(tv))\Big|_{t=0}$$

- On $\mathcal{X} = \mathbb{R}^n$ with Euclidean metric, geodesic convexity is convexity
- On $\mathcal{X} = \mathbb{R}^n_+$ with log-barrier metric,

$$f(x) = -\sum_{i=1}^{n} \log x_i$$

  is geodesically *linear* (both convex and concave)
- On $\mathcal{X} =$ PSD matrices with log-det metric,

$$f(x) = -\log\det x$$

  is geodesically *linear* (both convex and concave)

Geodesic strong convexity

A function $f : \mathcal{X} \to \mathbb{R}$ is **geodesically α-strongly convex** if it is α-strongly convex along any geodesic (X_t):

$$\Leftrightarrow \quad (1-t)f(x) + tf(y) \geq f(X_t) + \frac{\alpha t(1-t)}{2} d(x,y)^2$$

$$\Leftrightarrow \quad f(\exp_x(v)) \geq f(x) + \langle \mathrm{grad}\, f(x), v \rangle_x + \frac{\alpha}{2} \|v\|_x^2$$

$$\Leftrightarrow \quad \mathrm{Hess}\, f(x) \succeq \alpha I$$

Geodesic smoothness

A function $f\colon \mathcal{X} \to \mathbb{R}$ is **geodesically L-smooth** if it is L-smooth along any geodesic:

$$\Leftrightarrow \quad f(\exp_x(v)) \leq f(x) + \langle \operatorname{grad} f(x), v \rangle_x + \frac{L}{2}\|v\|_x^2$$

$$\Leftrightarrow \quad \operatorname{Hess} f(x) \preceq LI$$

If f is α-strongly convex and L-smooth, define *condition number*

$$\kappa = \frac{L}{\alpha}$$

Gradient domination

A function $f\colon \mathcal{X} \to \mathbb{R}$ is α-**gradient dominated** if

$$\|\mathrm{grad}\, f(x)\|_x^2 \geq 2\alpha\,(f(x) - \min f)$$

A function $f\colon \mathcal{X} \to \mathbb{R}$ has α-**sufficient growth** if

$$f(x) - \min f \geq \frac{\alpha}{2} d(x, x^*)^2$$

Theorem: Geodesic strong convexity $\Rightarrow$ gradient domination $\Rightarrow$ sufficient growth (with the same constant α)

- Otto & Villani, *Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality*, Journal of Functional Analysis, 2000: Propositions 1' & 2'

Convergence rates

Convergence rate of gradient flow

Gradient flow

$$\dot{X}_t = -\mathrm{grad}\, f(X_t)$$

- If f is α-strongly convex, then

$$d(X_t, Y_t)^2 \leq e^{-2\alpha t} d(X_0, Y_0)^2$$

- If f is α-gradient dominated, then

$$f(X_t) - \min f \leq e^{-2\alpha t}(f(X_0) - \min f)$$

Convergence rate of gradient descent

Gradient descent

$$x_{k+1} = \exp_{x_k}(-\eta \operatorname{grad} f(x_k))$$

- If f is α-gradient dominated and L-smooth, with $\eta = \frac{1}{L}$,

$$f(x_k) - \min f \leq \left(1 - \frac{1}{\kappa}\right)^k (f(x_0) - \min f)$$

Convergence rate of proximal method

Proximal method

$$x_{k+1} = \exp_{x_k}(-\eta \,\mathrm{grad}\, f(x_k))$$

- If f is α-gradient dominated, for all $\eta > 0$,

$$f(x_k) - \min f \leq \frac{f(x_0) - \min f}{(1 + \alpha\eta)^k}$$

  - see [Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018], Appendix C

Extensions

- Zhang, Reddi & Sra, *Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds*, NeurIPS 2016
- Sundaramoorthi & Yezzi, *Variational PDEs for Acceleration on Manifolds and Application to Diffeomorphisms*, NeurIPS 2018
- Ahn & Sra, *From Nesterov's estimate sequence to Riemannian acceleration*, COLT 2020
- Criscitiello & Boumal, *An accelerated first-order method for non-convex optimization on manifolds*, arXiv:2008.02252, 2020
- Hamilton & Moitra, *No-go Theorem for Acceleration in the Hyperbolic Plane*, arXiv:2101.05657, 2021