CPSC 661: Sampling Algorithms in Machine Learning *Out: March 22, 2021*

Problem Set 2

Lecturer: Andre Wibisono *Due: April 12, 2021*

Instruction

Solve at least 3 of the following problems (feel free to solve as many as you'd like). Each problem has equal worth, so you can choose the ones that are most interesting to you. Collaboration is allowed and encouraged, but please write your own solution and acknowledge your collaborators. Submit the solutions as a single PDF file via Canvas. If there are questions, please post a discussion on Canvas or email `andre.wibisono@yale.edu`.

1 Strong convexity

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$. let $\alpha > 0$. Recall:

- f is α-strongly convex if $\nabla^2 f(x) \succeq \alpha I$ for all $x \in \mathbb{R}^n$.
- f is α-gradient dominated if $\|\nabla f(x)\|^2 \geq 2\alpha(f(x) - f(x^*))$ for all $x \in \mathbb{R}^n$.
- f has α-sufficient growth if $f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2$ for all $x \in \mathbb{R}^n$.

(a) Show that if f is α-strongly convex, then f is also α-gradient dominated.

(b) Show that if f is α-gradient dominated, then f also has α-sufficient growth.

2 Convex optimization

Assume $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and let $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$.

(a) Prove that the gradient flow dynamics $\dot{X}_t = -\nabla f(X_t)$ minimizes f at a polynomial rate:

$$f(X_t) - f(x^*) \leq \frac{\|X_0 - x^*\|^2}{2t}.$$

(b) Assume that f is L-smooth. Prove that the gradient descent algorithm $x_{k+1} = x_k - \eta \nabla f(x_k)$ with step size $0 < \eta \leq \frac{1}{L}$ minimizes f at a polynomial rate:

$$f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta k}.$$

3 Moreau-Yosida regularization

The **Moreau-Yosida regularization** of $f\colon \mathbb{R}^n \to \mathbb{R}$ is the function $f_\eta\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$f_\eta(x) = \inf_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\eta}\|x-y\|^2 \right\}$$

for some $\eta > 0$. Show that proximal method for f is equivalent to gradient descent for f_η:

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

That is, show that $\nabla f_\eta(x) = \nabla f(y)$ where y satisfies $y = x - \eta \nabla f(y)$.

4 Extremal trajectories for convex cost

Let $c\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function.

(a) Show that for any $T > 0$ and $x, y \in \mathbb{R}^n$,

$$\inf_{(X_t)} \int_0^T c(\dot{X}_t)\, dt = Tc\left(\frac{y-x}{T}\right)$$

where the infimum is over all curves (X_t) from $X_0 = x$ to $X_T = y$.

(b) Assume c is strictly convex. Show that the infimum is achieved uniquely by straight line:

$$X_t = x + \frac{t}{T}(y-x).$$

5 Rescaled gradient flow

Let $f\colon \mathbb{R} \to \mathbb{R}$ be the polynomial function

$$f(x) = \frac{1}{p}|x|^p$$

for some $p > 2$, with minimizer $x^* = 0$ and $f(x^*) = 0$. Let $X_0 > 0$ be a starting point.

(a) Show that the gradient flow dynamics $\dot{X}_t = -\nabla f(X_t)$ minimizes f at a polynomial rate:

$$f(X_t) = \Theta(t^{-\frac{p}{p-2}}).$$

(b) Show that the *rescaled gradient flow* of order p, which is the dynamics

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

minimizes f at an exponential rate:

$$f(X_t) = e^{-pt} f(X_0).$$

6 Rescaled gradient descent

Let $f\colon \mathbb{R} \to \mathbb{R}$ be the polynomial function

$$f(x) = \frac{1}{p}|x|^p$$

for some $p > 2$, with minimizer $x^* = 0$ and $f(x^*) = 0$. Let $x_0 > 0$ be a starting point.

(a) Show that the gradient descent algorithm $x_{k+1} = x_k - \eta \nabla f(x_k)$ with step size $0 < \eta \leq \frac{1}{2x_0^{p-2}}$ minimizes f at a polynomial rate:

$$f(x_k) = \Omega((\eta k)^{-\frac{p}{p-2}}).$$

(b) Show that the *rescaled gradient descent* of order p, which is the algorithm

$$x_{k+1} = x_k - \eta \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|^{\frac{p-2}{p-1}}}$$

with step size $0 < \eta < 1$, minimizes f at an exponential rate:

$$f(x_k) = (1-\eta)^{pk} f(x_0).$$

7 Fisher metric

Recall the Fisher metric on the simplex $\Delta_{n-1} = \{x \in \mathbb{R}^n \colon x_i \geq 0, \sum_{i=1}^n x_i = 1\}$ is given by the matrix $g(x) = \mathrm{diag}(\frac{1}{x_1}, \dots, \frac{1}{x_n})$.

Compute the geodesic and distance between two points $x, y \in \Delta_{n-1}$ under the Fisher metric.

8 Lower bound

Let ρ_1, ρ_2 be probability distributions on $\mathbb{R}^n$ with mean $\mu_1, \mu_2 \in \mathbb{R}^n$ and covariance $\Sigma_1, \Sigma_2 \succ 0$. Assume $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$. Show that the Wasserstein distance is lower bounded by:

$$W_2(\rho_1, \rho_2)^2 \geq \|\mu_1 - \mu_2\|_2^2 + \|\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}}\|_{\mathrm{HS}}^2.$$

9 Wasserstein projection

Fix $\mu \in \mathbb{R}^n$ and $\sigma^2 > 0$, and let $\mathcal{P}_{\mu,\sigma^2}$ denote the set of probability distributions ρ on $\mathbb{R}^n$ with mean $\mu = \mathbb{E}_\rho[X]$ and variance $\sigma^2 = \mathbb{E}_\rho[\|X - \mu\|^2]$. Let ν be a probability distribution on $\mathbb{R}^n$ with mean $\mu_0 = \mathbb{E}_\nu[X]$ and variance $\sigma_0^2 = \mathbb{E}_\nu[\|X - \mu_0\|^2]$. Consider the projection of ν to $\mathcal{P}_{\mu,\sigma^2}$:

$$\inf_{\rho \in \mathcal{P}_{\mu,\sigma^2}} W_2(\rho, \nu)^2$$

(a) Show the infimum is attained at $\tilde{\nu}$ given by

$$\tilde{\nu}(x) = a^n \, \nu(a(x-\mu) + \mu_0)$$

where $a = \sqrt{\sigma_0^2/\sigma^2}$. Describe the optimal transport map that sends ν to $\tilde{\nu}$.

(b) Show the minimum value is

$$W_2(\nu, \tilde{\nu})^2 = \|\mu_0 - \mu\|^2 + (\sqrt{\sigma_0^2} - \sqrt{\sigma^2})^2.$$

Hints

1. Consider the gradient flow dynamics from $X_0 = x$.
   (a) Show $\|\nabla f(X_t)\|^2$ converges to 0 exponentially fast and integrate.
   (b) Show $\mathcal{E}_t = \sqrt{\frac{2}{\alpha}(f(X_t) - f(x^*))} + \|X_t - x\|^2$ is decreasing.

2. (a) Show $\mathcal{E}_t = t(f(X_t) - f(x^*)) + \frac{1}{2}\|X_t - x^*\|^2$ is decreasing along gradient flow.
   (b) Find a similar functional that decreases along gradient descent.

4. Use Jensen's inequality.

7. Show that the simplex with the Fisher metric is isometric to the positive orthant of the sphere $\mathbb{S}^+_{n-1} = \{x \in \mathbb{R}^n \colon x_i \geq 0, \sum_{i=1}^n x_i^2 = 1\}$ with the Euclidean metric $g(x) = I$.

8. Write the optimization problem in terms of covariance and reduce to Gaussian case.