Pre-Questions for Reading Group

Andre Wibisono

January 7, 2022

Instruction

Please solve **at least 4** of the following questions. There are 4 sections; choose at least 1 question from each section. Please choose the most interesting or challenging problems that you can solve. If you have seen the solution of a problem elsewhere, please choose a different problem.

Please do not discuss the questions or collaborate with anyone. You may consult textbooks or other references, but please cite the sources you use.

Write your solution using Latex with rigorous proofs. Submit the PDF **via Gradescope** on the course **OSG 1** with course number **6P3G4R** by **January 30, 2022**.

I – Optimization

1 Optimizing via Gradient Flow

Suppose we want to minimize an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$. Assume f is differentiable, and let $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$ be a global minimizer of f, so $f(x^*) = \min_{x\in\mathbb{R}^d} f(x)$.

Suppose we follow the **gradient flow** dynamics, which is the solution $\{X_t \in \mathbb{R}^d \colon t \geq 0\}$ to the following differential equation starting from an arbitrary $X_0 \in \mathbb{R}^d$:

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

where $\dot{X}_t = \frac{d}{dt} X_t$ is the time derivative and $\nabla f(X_t) \in \mathbb{R}^d$ is the gradient vector of f.

(a) Show that along gradient flow (1), the function value $f(X_t)$ is monotonically decreasing:

$$f(X_t) \leq f(X_s) \quad \text{for all} \quad 0 \leq s \leq t.$$

Does this imply $X_t \to x^*$ as $t \to \infty$?

(b) Now suppose that f is *gradient dominated*, which means it satisfies the following inequality with some constant $\alpha > 0$:

$$\|\nabla f(x)\|^2 \geq 2\alpha(f(x) - f(x^*)) \quad \text{for all} \quad x \in \mathbb{R}^d.$$

(Here and throughout, $\|\cdot\|$ denotes the ℓ_2-norm in $\mathbb{R}^d$.)

Show that along gradient flow (1), the function value $f(X_t)$ converges to the minimum $f(x^*)$ exponentially fast:

$$f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*)) \quad \text{for all} \quad t \geq 0.$$

(c) From part (b), what is the time complexity of gradient flow to reach an error of $\epsilon > 0$ in function value (i.e. how long should we wait until $f(X_t) - f(x^*) \leq \epsilon$)?

2 Optimizing via Gradient Descent

As in Problem 1, suppose we want to minimize an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$. Assume f is differentiable, and let $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$ be a global minimizer of f, so $f(x^*) = \min_{x\in\mathbb{R}^d} f(x)$.

Suppose we use the **gradient descent** algorithm, which is the following update starting from an arbitrary $x_0 \in \mathbb{R}^d$:

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

where $\eta > 0$ is a fixed step size. Note that as $\eta \to 0$, gradient descent (2) recovers the gradient flow dynamics (1).

Assume that f is L-*smooth*, which means the gradient map $\nabla f \colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz for some $0 < L < \infty$:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \quad \text{for all} \quad x, y \in \mathbb{R}^d.$$

(a) Show that along gradient descent (2) with small enough step size:

$$0 \le \eta \le \frac{2}{L}$$

the function value $f(x_k)$ is monotonically decreasing:

$$f(x_{k+1}) \le f(x_k) \quad \text{for all} \quad k \ge 0.$$

What happens when η is too large? Give a concrete example where the behavior above breaks.

(b) Assume further that f is *gradient dominated* with constant $\alpha > 0$:

$$\|\nabla f(x)\|^2 \ge 2\alpha(f(x) - f(x^*)) \quad \text{for all} \quad x \in \mathbb{R}^d.$$

Show that along gradient descent (2) with small step size:

$$0 \le \eta \le \frac{1}{L}$$

the function value $f(x_k)$ converges to the minimum $f(x^*)$ exponentially fast:

$$f(x_k) - f(x^*) \le (1 - \alpha\eta)^k \left(f(x_0) - f(x^*)\right) \quad \text{for all} \quad k \ge 0.$$

(c) From part (b), if we choose the largest step size $\eta = \frac{1}{L}$, what is the iteration complexity of gradient descent to reach error ϵ in function value (i.e., how long should we run gradient descent until $f(x_k) - f(x^*) \le \epsilon$)?

Compare your answer with the time complexity of gradient flow in Problem 1(c).

3 Optimizing via Proximal Gradient Method

As in Problem 1, suppose we want to minimize an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$. Assume f is differentiable, and let $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$ be a global minimizer of f, so $f(x^*) = \min_{x\in\mathbb{R}^d} f(x)$.

Suppose we use the **proximal gradient method**, which is the following update starting from an arbitrary $x_0 \in \mathbb{R}^d$:

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^d} \; f(x) + \frac{1}{2\eta}\|x - x_k\|^2 \tag{3}$$

where $\eta > 0$ is step size. Assume we can solve the regularized optimization problem in (3) in each iteration. If f is differentiable, then the update (3) is equivalent to the following implicit update rule:

$$x_{k+1} = x_k - \eta\nabla f(x_{k+1}). \tag{4}$$

Note that as $\eta \to 0$, proximal gradient method (4) recovers the gradient flow dynamics (1). Whereas gradient descent (2) is a forward discretization of gradient flow (1), proximal gradient (4) is a backward discretization.

(a) Show that along proximal gradient method (3) with any step size $\eta > 0$, the function value $f(x_k)$ is monotonically decreasing:

$$f(x_{k+1}) \le f(x_k) \quad \text{for all} \quad k \ge 0.$$

(b) Assume that f is *gradient dominated* with constant $\alpha > 0$:

$$\|\nabla f(x)\|^2 \ge 2\alpha(f(x) - f(x^*)) \quad \text{for all} \quad x \in \mathbb{R}^d.$$

Show that along proximal gradient method (3) with any step size $\eta > 0$, the function value $f(x_k)$ converges to the minimum $f(x^*)$ exponentially fast:

$$f(x_k) - f(x^*) \le \frac{1}{(1+\alpha\eta)^k}\,(f(x_0) - f(x^*)) \quad \text{for all} \quad k \ge 0.$$

What prevents us from choosing a large step size $\eta \to \infty$?

(c) Under the same setting as in part (b), show that in fact:

$$f(x_k) - f(x^*) \le \frac{1}{(1+\alpha\eta)^{2k}}\,(f(x_0) - f(x^*)) \quad \text{for all} \quad k \ge 0.$$

Show with a concrete example that this is the correct rate of convergence.

II – Sampling

4 Sampling via Ornstein-Uhlenbeck

Suppose we want to sample from the Gaussian distribution $\nu = \mathcal{N}(0, A^{-1})$ on $\mathbb{R}^d$ with mean $0 \in \mathbb{R}^d$ and covariance matrix $\Sigma = A^{-1}$ for some symmetric positive definite matrix $A \succ 0$, $A \in \mathbb{R}^{d \times d}$.

Suppose we follow the **Ornstein-Uhlenbeck (OU) process**, which is the stochastic process $(X_t)_{t \geq 0}$ in $\mathbb{R}^d$ that evolves following the stochastic differential equation (SDE):

$$dX_t = -AX_t \, dt + \sqrt{2} \, dW_t \tag{5}$$

where $(W_t)_{t \geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$ starting at $W_0 = 0$. Suppose we start at $X_0 \sim \rho_0$ for some arbitrary initial probability distribution ρ_0. Let ρ_t be the probability distribution of X_t for all $t \geq 0$.

(a) Solve the SDE (5) and find the explicit solution of ρ_t in terms of ρ_0 and t.

(b) Show that as $t \to \infty$, the distribution ρ_t converges to the target distribution $\nu = \mathcal{N}(0, A^{-1})$. Characterize the convergence speed in terms of the eigenvalues of A.

(c) Suppose we want to sample from a general target distribution ν with density $\nu(x) \propto e^{-f(x)}$. How should we modify the OU process (5) to make it converge to ν?

5 Sampling via Discretized Ornstein-Uhlenbeck

As in Problem 4, suppose we want to sample from the Gaussian distribution $\nu = \mathcal{N}(0, A^{-1})$ on $\mathbb{R}^d$ with mean $0 \in \mathbb{R}^d$ and covariance matrix $\Sigma = A^{-1}$ for some $A \succ 0$, $A \in \mathbb{R}^{d \times d}$.

In discrete time, we can discretize the OU process (5) to get the following algorithm:

$$x_{k+1} = x_k - \eta A x_k + \sqrt{2\eta} \, z_k \tag{6}$$

where $\eta > 0$ is step size and $z_k \sim \mathcal{N}(0, I)$ is an independent Gaussian random variable in $\mathbb{R}^d$. Suppose we start at $x_0 \sim \rho_0$, and let ρ_k be the probability distribution of x_k for all $k \geq 0$.

(a) Solve the recursion (6) and find the explicit solution of ρ_k in terms of ρ_0 and k.

(b) Show that for small $\eta > 0$, the distribution ρ_k converges to a limiting distribution $\nu_\eta = \lim_{k \to \infty} \rho_k$, and show that $\nu_\eta \neq \nu$. What is the range of η for which this happens?

(c) Explain why discretizing the OU process (5) for sampling results in a biased algorithm (6) (converges to the wrong limit $\nu_\eta \neq \nu$), while in optimization, discretizing gradient flow (1) results in a consistent algorithm (converging to the correct minimizer x^*) via e.g. gradient descent (2) or proximal gradient (3). How would you fix the problem?

6 Random Walk on a Graph

Let $G = (V, E)$ be a connected, undirected, unweighted graph on the vertex set $V = \{1, \dots, n\}$ with no self-loops. Let $N(i) \subset V$ be the set of neighbors of node i, and $d_i = |N(i)|$ the degree of node i. Let $A \in \{0,1\}^{n \times n}$ be the adjacency matrix of G, and let $D \in \mathbb{R}^{n \times n}$ be the diagonal degree matrix. Let $K = D^{-1}A$ be the *random walk matrix*, and let $L = K - I$ be the *Laplacian matrix*, where $I \in \mathbb{R}^{n \times n}$ is the identity matrix.

Consider the following discrete-time random walk on G: At each iteration $k \geq 0$ we are at a random vertex $x_k \in V$, and at the next iteration we jump to one of its neighbors $x_{k+1} \in N(x_k)$ uniformly at random. If x_k has probability distribution $\rho_k \in \mathbb{R}^n$ (represented as a row vector) and x_{k+1} has probability distribution $\rho_{k+1} \in \mathbb{R}^n$, then

$$\rho_{k+1} = \rho_k K. \tag{7}$$

(a) Describe what is the stationary probability distribution $\nu \in \mathbb{R}^n$ of the random walk (7). Show that along the random walk (7), ρ_k converges to ν exponentially fast in χ^2-divergence:

$$\chi^2_\nu(\rho_k) \leq (1 - \lambda)^k \chi^2_\nu(\rho_0)$$

for some $0 < \lambda \leq 1$. Describe what is λ in terms of the eigenvalues of the Laplacian L.

(Recall the χ^2-divergence between ρ and ν is $\chi^2_\nu(\rho) = \mathrm{Var}_\nu(\frac{\rho}{\nu}) = \sum_{i \in V} \nu(i)\big(\frac{\rho(i)}{\nu(i)} - 1\big)^2$.)

(b) Now suppose we follow a *continuous-time random walk* on G, where we jump following K (7) after every random time interval with an exponential distribution: At each time $t \geq 0$ we are at a random vertex $X_t \in V$, then we draw a random $\tau \sim \mathrm{Exp}(1)$ from the exponential distribution with mean 1, and we wait until time $t + \tau$ before jumping to one of its neighbors $X_{t+\tau} \in N(X_t)$ uniformly at random.

Let $\rho_t \in \mathbb{R}^n$ be the probability distribution of X_t at time $t \geq 0$ following the process above. Show that ρ_t evolves following the *heat equation*, which is the differential equation:

$$\dot{\rho}_t = \rho_t L \tag{8}$$

where $\dot{\rho}_t = \frac{d}{dt}\rho_t$. Show that along (8), ρ_t converges to ν exponentially fast in χ^2-divergence:

$$\chi^2_\nu(\rho_t) \leq e^{-\lambda t} \chi^2_\nu(\rho_0)$$

where λ is the same constant as in part (a).

III – Games

7 Min-Max Optimization via Simultaneous Gradient Descent

Consider the min-max optimization problem:

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^n} x^\top A y \tag{9}$$

where $A \in \mathbb{R}^{n\times n}$ is an arbitrary payoff matrix. Suppose we want to find the equilibrium point $(x^*, y^*) = (0, 0) \in \mathbb{R}^{2n}$. We define the distance from $(x, y) \in \mathbb{R}^{2n}$ to the equilibrium as

$$H(x, y) = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2.$$

(a) In continuous time, suppose we follow the *simultaneous gradient flow* dynamics:

$$\begin{aligned}\dot{X}_t &= -AY_t\\ \dot{Y}_t &= A^\top X_t \ .\end{aligned}$$

Show that (X_t, Y_t) remains bounded for all $t \geq 0$. Furthermore, show that the distance to equilibrium is preserved:

$$H(X_t, Y_t) = H(X_0, Y_0) \quad \text{for all } t \geq 0.$$

(b) In discrete time, suppose we follow the *simultaneous gradient descent* algorithm with step size $\eta > 0$:

$$\begin{aligned}x_{k+1} &= x_k - \eta A y_k\\ y_{k+1} &= y_k + \eta A^\top x_k \ .\end{aligned}$$

Assume A is invertible. Show that for any $\eta > 0$, (x_k, y_k) becomes unbounded as $k \to \infty$. Furthermore, show that the distance to equilibrium increases exponentially fast:

$$H(x_k, y_k) \geq (1 + \eta^2\alpha^2)^k H(x_0, y_0) \quad \text{for all } k \geq 0$$

where $\alpha > 0$ is the smallest singular value of A.

8 Min-Max Optimization via Optimistic Gradient Descent

Suppose we want to solve the min-max optimization problem (9) for some payoff matrix $A \in \mathbb{R}^{n\times n}$. Assume A is invertible.

(a) Suppose we follow the *simultaneous proximal gradient* method with step size $\eta > 0$:

$$\begin{aligned} x_{k+1} &= x_k - \eta A y_{k+1} \\ y_{k+1} &= y_k + \eta A^\top x_{k+1} \ . \end{aligned}$$

Show that for any $\eta > 0$, (x_k, y_k) converges to the equilibrium as $k \to \infty$. Furthermore, show that the distance to equilibrium decreases exponentially fast:

$$H(x_k, y_k) \leq \frac{1}{(1+\eta^2\alpha^2)^k} H(x_0, y_0) \quad \text{for all } k \geq 0$$

where $\alpha > 0$ is the smallest singular value of A.

(b) In general, the proximal method can be difficult to implement. An approximation is the *optimistic gradient method* with step size $\eta > 0$:

$$\begin{aligned} x_{k+1} &= x_k - 2\eta A y_k + \eta A y_{k-1} \\ y_{k+1} &= y_k + 2\eta A^\top x_k - \eta A^\top x_{k-1} \end{aligned}$$

starting from arbitrary $(x_{-1}, y_{-1}) = (x_0, y_0) \in \mathbb{R}^{2n}$.

Show that for small $\eta > 0$, the iterates (x_k, y_k) converge to the equilibrium exponentially fast:

$$H(x_k, y_k) \leq (1 - \eta^2\beta)^k H(x_0, y_0) \quad \text{for all } k \geq 0$$

for some $\beta > 0$. Describe what is the constant β in terms of the singular values of A. What is the range of η for which this result holds?

9 Min-Max Optimization via Alternating Gradient Descent

Suppose we want to solve the min-max optimization problem (9) for some arbitrary $A \in \mathbb{R}^{n \times n}$. We have seen from Problem 7 that simultaneous gradient flow conserves the distance to the equilibrium in continuous time.

Now suppose we follow the *alternating gradient descent* algorithm with step size $\eta > 0$:

$$\begin{aligned} x_{k+1} &= x_k - \eta A y_k \\ y_{k+1} &= y_k + \eta A^\top x_{k+1} \ . \end{aligned}$$

Show that for all $\eta > 0$, the iterates (x_k, y_k) remain bounded for all $k \geq 0$. Show that the distance to equilibrium $H(x_k, y_k)$ is *not* conserved. However, show that there is a conserved quantity, i.e. a function $H_\eta(x, y) \in \mathbb{R}$ such that

$$H_\eta(x_k, y_k) = H_\eta(x_0, y_0) \quad \text{for all } \ k \geq 0.$$

What is the relationship between H_η and H?

IV – Miscellaneous

10 Stein's identity

(a) Let $Z \sim \mathcal{N}(0, I)$ be a standard Gaussian random variable in $\mathbb{R}^d$. Show that for any differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}[\nabla f(Z)] = \mathbb{E}[Z\, f(Z)] \ .$$

(b) If $Z \sim \mathcal{N}(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$, what does the identity above become?

11 Recursion

Show that for any $x_1, \ldots, x_k \geq 0$ with $x_k > 0$,

$$\sum_{i=1}^{k} \frac{x_i}{\sqrt{x_i + x_{i+1} + \cdots + x_k}} \leq 2\sqrt{\sum_{i=1}^{k} x_i} \ .$$

12 Partition

A *partition* of a natural number $n \in \mathbb{N}$ is an (ordered) sequence $(k_1, \ldots, k_m)$ where each $k_i \geq 1$ and $k_1 + \cdots + k_m = n$. Let $P(n)$ denote the set of all partitions of n. Show that for all $n \geq 2$,

$$\sum_{(k_1,\ldots,k_m)\in P(n)} \frac{(-1)^m}{m!} \frac{1}{\prod_{i=1}^{m} k_i} = 0 \ .$$