1 Outline

In this lecture, we cover

- minimax optimization and applications,
- saddle point,
- minimax theorems,
- gradient descent ascent (GDA).

2 Introduction to Mininax Optimization

Let $X \subseteq \mathbb{R}^d$, and let $Y \subseteq \mathbb{R}^p$. Given a function $\phi : X \times Y \to \mathbb{R}$, we consider the following optimization problem.

$$\min_{x\in X} \max_{y\in Y} \phi(x, y).$$

The problem is referred to as **min-max optimization** and **minimax optimization**. The problem can be interpreted as a game between two players, making decisions on x and y, respectively. The x-player chooses a solution x from the domain X. Given x chosen by the x-player, the y-player chooses a solution y from the domain Y to maximize the value of $\phi(x, y)$. The problem has many applications in machine learning such as

- zero-sum game,
- constrained optimization,
- nonsmooth optimization,
- distributionally robust optimization,
- generative adversarial networks,
- sharpness-aware minimization.

In this lecture, we briefly discuss the first three applications. We will study the other two applications in depth later.

2.1 Zero-Sum Game

Suppose that we have two adversarial players. Player 1 chooses from d actions $i \in [d]$ while player 2 chooses from m actions $j \in [m]$. If player 1 chooses $i \in [d]$ and player 2 chooses $j \in [m]$, then player 1 loses a_{ij} while player gains a_{ij}. This is called a zero-sum game.

Both players can *randomize* their strategies, meaning that player 1 chooses $x \in \Delta_d = \{x \in [0, 1]^d : \mathbf{1}^\top x = 1\}$ and player 2 chooses $y \in \Delta_m = \{y \in [0, 1]^m : \mathbf{1}^\top y = 1\}$. Then $x^\top Ay$ is the expected loss for player 1 and also the expected gain for player 2.

Suppose that player 1 knows player 2's strategy, given by a vector $y \in \Delta_m$. Then player 1 will choose a strategy $x \in \Delta_d$ so that the expected loss can be minimized and incurs a loss of

$$\min_{x\in\Delta_d} x^\top Ay.$$

Given that player 2 knows player 1 will do this for any y, player 2 should choose y to maximize the expected gain so that player 2 obtains a gain of

$$\max_{y\in\Delta_m}\min_{x\in\Delta_d} x^\top Ay.$$

2.2 Constrained Optimization

Consider the following inequality constrained problem.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0 \quad \text{for } i = 1,\dots,m. \end{aligned} \tag{13.1}$$

Note that

$$\max_{\lambda\ge 0} \quad \mathcal{L}(x,\lambda) \quad = \quad \max_{\lambda\ge 0}\left\{ f(x) + \sum_{i=1}^m \lambda_i g_i(x)\right\}.$$

If $g_i(x) > 0$ for some $i \in [m]$, then we can send λ_i to $+\infty$, making $\mathcal{L}(x,\lambda)$ arbitrarily large. On the other hand, if $g_i(x) \le 0$ for all $i \in [m]$, then $\max_{\lambda\ge 0}\mathcal{L}(x,\lambda)$ is attained at $\lambda = 0$, in which case, $\max_{\lambda\ge 0}\mathcal{L}(x,\lambda) = f(x)$. This observation implies that

$$\min_x \max_{\lambda\ge 0} \quad \mathcal{L}(x,\lambda) \quad = \quad \min_x \{f(x) : \; g_i(x) \le 0 \text{ for } i = 1,\dots,m\}.$$

2.3 Nonsmooth Optimization

Let us consider

$$\min_{x\in\mathbb{R}^d} \quad f(x) + g(Ax)$$

where

- f is smooth and convex,
- g is strongly convex but nonsmooth,
- A is a $p \times d$ matrix.

We may reformulate the problem by a minimax optimization problem based on **Fenchel duality**. Here, we may rewrite $g(Ax)$ as

$$g(Ax) = \max_{y\in\mathbb{R}^p}\left\{ y^\top Ax - g^*(y)\right\}$$

where $g^*(y)$ is the **Fenchel conjugate** of g. Then the original problem with the composite objective can be written as

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^p} \quad f(x) + y^\top Ax - g^*(y).$$

Here, $y^\top Ax$ is smooth in x and y. Moreover, it is known that the Fenchel conjugate of a strongly convex function is smooth. Hence, the second formulation is a minimax optimization problem with a smooth objective.

2.4 Distributionally Robust Optimization

Stochastic optimization problems have the following form:

$$\min_{x} \quad \mathbb{E}_{\xi \sim \mathcal{P}} \left[\ell(x, \xi) \right]$$

where

- $\ell(x, \xi)$ is the loss under decision x and data ξ,
- $\mathcal{P}$ is the (unknown) distribution of data ξ.

A common practice is to estimate the distribution $\mathcal{P}$ based on data samples. Given n data points $\xi^1, \ldots, \xi^n$, we take

$$\mathcal{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi^i}$$

where δ_ξ is the **Dirac distribution** of ξ where the probability mass of 1 is given to ξ. Here, we call $\mathcal{P}_n$ an empirical distribution. **Empirical Risk Minimization (ERM)** approximates the stochastic optimization problem by replacing the true distribution $\mathcal{P}$ with the empirical distribution:

$$\min_{x} \quad \mathbb{E}_{\xi \sim \mathcal{P}_n} \left[\ell(x, \xi) \right] \quad = \quad \min_{x} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(x, \xi^i).$$

Although ERM works well in practice in general, there exist some cases in which ERM suffers with poor generalization performances. Under such scenarios, the empirical distribution $\mathcal{P}_n$ perhaps does not approximate the true distribution well.

Inspired by the issue, **Distributionally Robust Optimization (DRO)** considers ambiguity in inferring the true distribution based on the empirical distribution and aims to make a robust decision even under such distributional ambiguity. The way it works is to consider a "family" of distributions around $\mathcal{P}_n$ not just $\mathcal{P}_n$ itself. To be more specific, we consider

$$\mathcal{F}_n = \{ \mathcal{Q} : \ d(\mathcal{Q}, \mathcal{P}_n) \leq \rho \}$$

where

- $\mathcal{Q}$ denotes a probability distribution,
- $d(\mathcal{Q}, \mathcal{P}_n)$ is a discrepancy function to meausre the discrepancy between $\mathcal{Q}$ and $\mathcal{P}_n$,
- ρ is the radius on how much a distribution in the family $\mathcal{F}_n$ can differ from the empirical distribution $\mathcal{P}_n$.

Here, typical choices for the discrepancy function include

- the Kullback-Leibler (KL) divergence,
- the Wasserstein distance.

For these choices of the discrepancy function, we have statistical guarantees that the true distribution $\mathcal{P}$ belongs to the family $\mathcal{F}_n$ with high probability under a proper choice of the radius ρ. Then we consider

$$\min_{x} \max_{\mathcal{Q} \in \mathcal{P}_n} \quad \mathbb{E}_{\xi \sim \mathcal{Q}} \left[\ell(x, \xi) \right].$$

Here, the inner maximization captures the expected loss under a worst-case distribution from the family $\mathcal{F}_n$.

3 Saddle Point and Minimax Theorems

We start by proving the following result.

Theorem 13.1. *Consider the minimax optimization problem. Then the following statement holds.*

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \geq \max_{y\in Y}\min_{x\in X}\phi(x,y).$$

Proof. Note that for any $(x,y)\in X\times Y$, we have $\phi(x,y)\geq \min_{x\in X}\phi(x,y)$. Taking the maximum of each side over $y\in Y$, we obtain $\max_{y\in Y}\phi(x,y)\geq \max_{y\in Y}\min_{x\in X}\phi(x,y)$. As this inequality holds for every $x\in X$, taking the minimum of the left-hand side over $x\in X$ preserves the inequality. If done so, we deduce that $\min_{x\in X}\max_{y\in Y}\phi(x,y)\geq \max_{y\in Y}\min_{x\in X}\phi(x,y)$, as required. $\square$

Let us get back to the constrained optimization problem

$$\begin{aligned} \text{minimize}\quad & f(x)\\ \text{subject to}\quad & g_i(x)\leq 0 \quad \text{for } i=1,\ldots,m. \end{aligned} \tag{13.2}$$

Recall that its Lagrangian is given by $\mathcal{L}(x,\lambda)$. Moreover, the Lagrangian dual function is given by

$$q(\lambda)=\min_x \mathcal{L}(x,\lambda).$$

Then the Lagrangian dual problem is given by

$$\max_{\lambda\geq 0}\quad q(\lambda)\quad = \quad \max_{\lambda\geq 0}\min_x \quad \mathcal{L}(x,\lambda).$$

Then Theorem 13.1 states that

$$\min_x\max_{\lambda\geq 0}\mathcal{L}(x,\lambda)\geq \max_{\lambda\geq 0}\min_x\mathcal{L}(x,\lambda).$$

In fact, we know that if strong duality holds, then the equality holds as follows.

$$\min_x\max_{\lambda\geq 0}\mathcal{L}(x,\lambda)= \max_{\lambda\geq 0}\min_x\mathcal{L}(x,\lambda).$$

In general, when does such equality hold for a minimax optimization problem? In this section, we provide other sufficient conditions under which the equality holds.

3.1 Saddle Point

We say that a solution $(x^*,y^*)\in X\times Y$ is a **saddle point** to the problem $\min_{x\in X}\max_{y\in Y}\phi(x,y)$ if

$$\phi(x^*,y)\leq\phi(x^*,y^*)\leq\phi(x,y^*)$$

for all $(x,y)\in X\times Y$. If (x^*,y^*) is a saddle point, then

$$\phi(x^*,y^*)=\max_{y\in Y}\phi(x^*,y)=\min_{x\in X}\phi(x,y^*).$$

Theorem 13.2. *If (x^*,y^*) is a saddle point, then*

$$\min_{x\in X}\max_{y\in Y}\phi(x,y)=\phi(x^*,y^*)=\max_{y\in Y}\min_{x\in X}\phi(x,y).$$

Proof. By definition, we obtain

$$\max_{y\in Y}\phi(x^*,y)\leq\phi(x^*,y^*)\leq\min_{x\in X}\phi(x,y^*).$$

Moreover, this implies that

$$\min_{x\in X}\max_{y\in Y}\phi(x^*,y)\leq\phi(x^*,y^*)\leq\max_{x\in X}\min_{x\in X}\phi(x,y^*).$$

By Theorem 13.1, it follows that the inequalities must hold with equality. $\square$

3.2 Minimax Theorems

We provide two more sufficient conditions. The following theorem is due to John von Neumann.

Theorem 13.3. *Assume that the following conditions are satisfied.*

- *X and Y are closed convex sets, and one of them is bounded.*
- *$\phi(x, y)$ is convex in x for any fixed y.*
- *$\phi(x, y)$ is concave in y for any fixed x.*

Then $\min_{x\in X} \max_{y\in Y} \phi(x, y) = \max_{y\in Y} \min_{x\in X} \phi(x, y)$.

For the zero-sum game, we know that Δ_m and Δ_d are both bounded.

$$\min_{x\in\Delta_d} x^\top \max_{y\in\Delta_m} Ay = \max_{y\in\Delta_m} \min_{x\in\Delta_d} x^\top Ay.$$

We also have the following result.

Theorem 13.4. *Assume that the following conditions are satisfied.*

- *X and Y are closed convex sets,*
- *$\phi(x, y)$ is strongly convex in x for any fixed y.*
- *$\phi(x, y)$ is strongly concave in y for any fixed x.*

Then $\min_{x\in X} \max_{y\in Y} \phi(x, y) = \max_{y\in Y} \min_{x\in X} \phi(x, y)$.

4 Gradient Descent Ascent Algorithm

From the minimax optimization problem $\min_{x\in X} \max_{y\in Y} \phi(x, y)$, we may consider

$$\begin{aligned} \text{Primal :} \quad & \min_{x\in X} \left\{ \overline{\phi}(x) := \max_{y\in Y} \phi(x, y) \right\} \\ \text{Dual :} \quad & \max_{y\in Y} \left\{ \underline{\phi}(y) := \min_{x\in X} \phi(x, y) \right\}. \end{aligned}$$

For any $(\bar{x}, \bar{y}) \in X \times Y$, Theorem 13.1 implies that

$$\overline{\phi}(\bar{x}) = \max_{y\in Y} \phi(\bar{x}, y) \geq \min_{x\in X} \phi(x, \bar{y}) = \underline{\phi}(\bar{y}).$$

We say that a point $(\bar{x}, \bar{y}) \in X \times Y$ is an ϵ-saddle point if

$$0 \leq \overline{\phi}(\bar{x}) - \underline{\phi}(\bar{y}) = \max_{y\in Y} \phi(\bar{x}, y) - \min_{x\in X} \phi(x, \bar{y}) \leq \epsilon.$$

Note that if $(\bar{x}, \bar{y}) \in X \times Y$ is an ϵ-saddle point, then

$$\begin{aligned} \overline{\phi}(\bar{x}) - \min_{x\in X} \overline{\phi}(x) &\leq \epsilon, \\ \max_{y\in Y} \underline{\phi}(y) - \underline{\phi}(\bar{y}) &\leq \epsilon. \end{aligned}$$

Algorithm 1 Gradient Descent Ascent

Initialize $x_1 \in X$ and $y_1 \in Y$.
for $t = 1, \ldots, T-1$ **do**
  Obtain $g_{x,t} \in \partial_x \phi(x_t, y_t)$ and $g_{y,t} \in \partial_y \phi(x_t, y_t)$.
  Update $x_{t+1} = \mathrm{proj}_X(x_t - \eta_t g_{x,t})$ and $y_{t+1} = \mathrm{proj}_Y(y_t + \eta_t g_{y,t})$ for some step size $\eta_t > 0$.
end for
Return x_{T+1}.

Let us consider an algorithm for solving the minimax optimization problem, whose pseudo-code is given as in Algorithm 1. The algorithm is called the **gradient descent ascent** method. Note that at each iteration, we simultaneously update both the primal variables x and the dual variables y. We assumed that $\phi(x, y)$ is convex in x and concave in y. $\partial_x \phi(x, y)$ is the subdifferential of $\phi(x, y)$ for a fixed y, and $\partial_y \phi(x, y)$ is the superdifferential of $\phi(x, y)$ for a fixed x.

Theorem 13.5. *Let $\bar{x}_T$ and $\bar{y}_T$ be defined as*

$$\bar{x}_T = \left(\sum_{t=1}^T \eta_t\right)^{-1} \sum_{t=1}^T \eta_t x_t, \quad \bar{y}_T = \left(\sum_{t=1}^T \eta_t\right)^{-1} \sum_{t=1}^T \eta_t y_t.$$

Then for any $(x, y) \in X \times Y$,

$$\phi(\bar{x}_T, y) - \phi(x, \bar{y}_T) \leq \frac{1}{2\sum_{t=1}^T \eta_t} \left(\|(x_1, y_1) - (x, y)\|_2^2 + \sum_{t=1}^T \eta_t^2 \|(g_{x,t}, g_{y,t})\|_2^2 \right).$$

Assuming that $\|(g_x, g_y)\|_2^2 \leq L^2$ for any $g_x \in \partial_x \phi(x, y)$ and $g_y \in \partial_y \phi(x, y)$ and that $\|(x_1, y_1) - (x, y)\|_2^2 \leq R^2$, we can set $\eta_t = R/(L\sqrt{T})$. Then for any $(x, y) \in X \times Y$,

$$\phi(\bar{x}_T, y) - \phi(x, \bar{y}_T) \leq \frac{LR}{\sqrt{T}}.$$

In particular,

$$\max_{y \in Y} \phi(\bar{x}_T, y) - \min_{x \in X} \phi(x, \bar{y}_T) \leq \frac{LR}{\sqrt{T}}.$$

Then setting $T = O(1/\epsilon^2)$, we know that $(\bar{x}_T, \bar{y}_T)$ is an ϵ-saddle point.