1 Outline

In this lecture, we study

- Saddle point problem,
- Fenchel duality.

2 Saddle point problem

Consider the following inequality constrained problem.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m. \end{aligned} \tag{18.1}$$

Note that

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

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 $g_i(x) \leq 0$ for all $i \in [m]$, then $\max_{\lambda \geq 0} \mathcal{L}(x, \lambda)$ is attained at $\lambda = 0$, in which case, $\max_{\lambda \geq 0} \mathcal{L}(x, \lambda) = f(x)$. This observation implies that

$$\min_x \max_{\lambda \geq 0} \mathcal{L}(x, \lambda) = \min_x \left\{ f(x) : \ g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \right\}.$$

Remember that the Lagrangian dual problem is given by

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

Then the weak duality theorem states that

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

Moreover, if strong duality holds, then the equality holds as follows.

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

More generally, consider a function $\phi(x, y)$ that is convex in x and concave in y. Then

$$\min_{x \in X} \max_{y \in Y} \phi(x, y) \tag{18.2}$$

where sets X and Y are convex is called a saddle point problem. Under certain conditions on X and Y, the minimum and maximum can be swapped.

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

Such a result is called a minimax theorem, and the strong Lagrangian duality theorem is an example.

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

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

In fact, von Neumann's minimax theorem states that it does not matter who moves first, because

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

2.2 Saddle point optimality

In general, we have the following relationship.

Theorem 18.1. *Consider the saddle point problem (18.2). 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. □

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 18.2. *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

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

Moreover, this implies that

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

By Theorem 18.1, it follows that the inequalities must hold with equality. □

A saddle point problem combines two convex optimization problems into one.

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

For any $(\bar{x},\bar{y})\in X\times Y$, Theorem 18.1 implies that

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

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

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

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

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

2.3 Primal-dual algorithm for saddle point problems

Let us consider an algorithm for solving the saddle point problem, whose pseudo-code is given as in Algorithm 1. The algorithm is called the *primal-dual subgradient method*. Note that at each

Algorithm 1 Primal-dual subgradient method

```
Initialize x_1 ∈ X and y_1 ∈ Y.
for t = 1, ..., T − 1 do
    Obtain g_{x,t} ∈ ∂_x φ(x_t, y_t) and g_{y,t} ∈ ∂_y φ(x_t, y_t).
    Update x_{t+1} = proj_X(x_t − η_t g_{x,t}) and y_{t+1} = proj_Y(y_t + η_t g_{y,t}) for some step size η_t > 0.
end for
Return x_{T+1}.
```

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 18.3. *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.

3 Fenchel duality

The Fenchel conjugate of a function $f : \mathbb{R}^d \to \mathbb{R}$ is given by

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \left\{ y^\top x - f(x) \right\}.$$

As $y^\top x - f(x)$ is linear in y, the conjugate function is always convex, regardless of f.

Lemma 18.4 (Fenchel-Young inequality)**.** *For $x \in dom(f)$ and $y \in dom(f^*)$,*

$$f(x) + f^*(y) \geq y^\top x.$$

Proof. Note that $f^*(y) = \sup_{x \in \mathrm{dom}(f)} (y^\top x - f(x)) \geq y^\top x - f(x)$. ☐

We discussed Lagrangian duality, and in fact, we can derive the Lagrangian dual function based on the conjugate function. Consider

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \\ & Cx \leq d. \end{aligned} \tag{18.3}$$

Then the associated Lagrangian dual function is given by

$$\begin{aligned} q(\lambda, \mu) &= \min_x \left\{ f(x) + \lambda^\top (Cx - d) + \mu^\top (Ax - b) \right\} \\ &= -d^\top \lambda - b^\top \mu + \min_x \left\{ f(x) + (C^\top \lambda + A^\top \mu)^\top x \right\} \\ &= -d^\top \lambda - b^\top \mu - \sup_x \left\{ -f(x) - (C^\top \lambda + A^\top \mu)^\top x \right\} \\ &= -d^\top \lambda - b^\top \mu - f^*(-C^\top \lambda - A^\top \mu). \end{aligned}$$

Note that the domain of $q(\lambda, \mu)$ is

$$\mathrm{dom}(q) = \left\{(\lambda, \mu) : \ -C^\top \lambda - A^\top \mu \in \mathrm{dom}(f^*)\right\}.$$

Then the Lagrangian dual problem is given by

$$\begin{aligned} \text{maximize} \quad & -d^\top \lambda - b^\top \mu - f^*(-C^\top \lambda - A^\top \mu) \\ \text{subject to} \quad & \lambda \geq 0 \\ & -C^\top \lambda - A^\top \mu \in \mathrm{dom}(f^*). \end{aligned} \tag{18.4}$$

In particular, when there is no inequality constraint, the associated Lagrangian dual function is given by

$$q(\mu) = -b^\top \mu - f^*(-A^\top \mu),$$

and the Lagrangian dual problem is given by

$$\begin{aligned} \text{maximize} \quad & -b^\top \mu - f^*(-A^\top \mu) \\ \text{subject to} \quad & -A^\top \mu \in \mathrm{dom}(f^*). \end{aligned} \tag{18.5}$$

3.1 Fenchel conjugate examples

Example 18.5. When $f(x) = c^\top x + d$ over $x \in \mathbb{R}^d$,

$$f^*(y) = \sup_{x \in \mathbb{R}^d} (y^\top x - c^\top x - d) = \begin{cases} -d, & \text{if } y = c, \\ +\infty, & \text{otherwise.} \end{cases}$$

Example 18.6. When $f(x) = \log(1 + e^x)$ over $x \in \mathbb{R}$,

$$f^*(y) = \sup_{x \in \mathbb{R}} (yx - \log(1 + e^x)) = \begin{cases} y \log y + (1-y)\log(1-y), & \text{if } 0 < y < 1, \\ 0, & \text{if } y \in \{0, 1\}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Example 18.7. When $f(x) = (1/2)x^\top Q x + p^\top x$ over $x \in \mathbb{R}^d$ for some positive definite Q,

$$f^*(y) = \sup_{x \in \mathbb{R}} \left(y^\top x - \frac{1}{2} x^\top Q x - p^\top x \right).$$

Note that the maximum is attained at $x = Q^{-1}(y - p)$. Therefore,

$$f^*(y) = \frac{1}{2}(y-p)^\top Q^{-1}(y-p).$$

Here,

$$\nabla f^*(y) = Q^{-1}(y - p),$$

which implies that $\nabla f(\nabla f^*(y)) = y$ and

$$\nabla f^*(y) = (\nabla f)^{-1}(y).$$

Example 18.8. When $f(x) = \sum_{i=1}^{d} x_i \log x_i$ over $x \in \mathbb{R}_{++}^d$,

$$f^*(y) = \sup_{x \in \mathbb{R}_{++}^d} \left(y^\top x - \sum_{i=1}^{d} x_i \log x_i \right) = \sup_{x \in \mathbb{R}_{++}^d} \left(\sum_{i=1}^{d} x_i (y_i - \log x_i) \right) = \sum_{i=1}^{d} e^{y_i - 1}.$$

Example 18.9. When $f(X) = -\log \det X$ over $X \in \mathbb{S}_{++}^d$,

$$f^*(Y) = \sup_{X \in \mathbb{S}_{++}^d} \left(\mathrm{tr}(Y^\top X) + \log \det X \right).$$

It is known that $\nabla \log \det X = X^{-1}$. Then the supremum is attained at $X = -Y^{-1}$, and therefore,

$$f^*(Y) = -d - \log \det(-Y).$$