1 Outline

In this lecture, we study

- Karush-Kuhn-Tucker conditions for optimality,
- saddle point problems and primal-dual algorithms,
- Lagrangian duality based on the Fenchel conjugate,
- dual gradient method.

2 Karush-Kuhn-Tucker conditions

Remember that x^* is an optimal solution to

$$\min_{x \in C} \quad f(x)$$

where C is a convex set and f is differentiable if and only if

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall x \in C.$$

However, the structure of C may be arbitrary, which makes the condition difficult to verify. In this section, we present another way of verifying optimality. Namely, Karu-Kuhn-Tucker conditions, often referred to as KKT conditions.

2.1 Linear constraints

We consider problems of the following structure.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax \leq b \\ & Cx = d \end{aligned} \tag{16.1}$$

where

- $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$,
- $C \in \mathbb{R}^{\ell \times d}$ and $d \in \mathbb{R}^\ell$.

Theorem 16.1 (KKT conditions for linearly constrained problems)**.** *The linearly constrained problem as in (16.1) satisfies the following.*

1. *(Necessity) If x^* is a feasible solution to (16.1) and $f(x^*)$ is a local minimum, then there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ such that*

$$\nabla f(x^*)^\top + \lambda^{*\top} A + \mu^{*\top} C = 0 \quad \& \quad \lambda^{*\top}(Ax - b) = 0. \tag{$\star$}$$

2. *(Sufficiency) If f is convex, x^* is a feasible solution to (16.1), and there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ satisfying ($\star$), then x^* is an optimal solution to (16.1).*

2.2 General convex constraints

We consider problems of the following structure.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0 \quad \text{for } i = 1, \dots, m \\ & h_j(x) = 0 \quad \text{for } j = 1, \dots, \ell \end{aligned} \tag{16.2}$$

where

- f is convex,
- $g_1, \dots, g_m$ are convex,
- $h_1, \dots, h_\ell$ are affine.

Definition 16.2 (Slater's condition)**.** Suppose that $g_1, \dots, g_k$ are affine and $g_{k+1}, \dots, g_m$ are convex functions that are not affine. Then we say that the problem (16.2) satisfies Slater's condition if there exists a solution $\bar{x}$ such that

$$g_i(\bar{x}) \le 0 \text{ for } i = 1, \dots, k, \quad g_i(\bar{x}) < 0 \text{ for } i = k+1, \dots, m, \quad h_j(\bar{x}) = 0 \text{ for } j = 1, \dots, \ell.$$

Theorem 16.3 (KKT conditions for convex constrained problems)**.** *The convex programming problem as in (16.2) satisfies the following.*

1. *(Necessity) Assume that Slater's condition is satisfied. If x^* is a feasible optimal solution to (16.2), then there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ such that*

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^\ell \mu_j^* \nabla h_j(x^*) = 0 \quad \& \quad \lambda_i^* g_i(x^*) = 0 \text{ for all } i = 1, \dots, m. \tag{$\star\star$}$$

2. *(Sufficiency) If x^* is a feasible solution to (16.2) and there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ satisfying ($\star\star$), then x^* is an optimal solution to (16.2).*

3 Saddle point problem

Consider the following inequality constrained problem.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0 \quad \text{for } i = 1, \dots, m. \end{aligned} \tag{16.3}$$

Note that

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

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

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

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

$$\min_{x \in X} \max_{y \in Y} \phi(x, y) \tag{16.4}$$

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.

3.1 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 $x^\top A y$ 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

$$\min_{x \in \Delta_d} x^\top A y.$$

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

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

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

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

3.2 Saddle point optimality

In general, we have the following relationship.

Theorem 16.4. *Consider the saddle point problem (16.4). 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. $\square$

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 16.5. *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

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

By Theorem 16.4, it follows that the inequalities must hold with equality. $\square$

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 16.4 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

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

3.3 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 $x_1 \in X$ 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}.

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 16.6. *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 $(x,y) \in X \times Y$,

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

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

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

In particular,

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

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

4 Lagrangian duality based on the Fenchel conjugate

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

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \\ & Cx \le d. \end{aligned} \tag{16.5}$$

Then the associated Lagrangian dual function is given by

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

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

$$\text{maximize} \quad -b^\top \mu - f^*(-A^\top \mu) \tag{16.6}$$

Then the problem is equivalent to

$$(-1) \quad \times \quad \text{minimize} \quad f^*(-A^\top \mu) + b^\top \mu.$$

As f^* is convex, this dual forumulation is a convex minimization problem. Let us apply the subgradient method to the dual.

Given μ_t, let $g_t \in \partial\left(f^*(-A^\top \mu_t) + b^\top \mu_t\right)$. Then the subgradient method applies the following update rule.

$$\mu_{t+1} = \mu_t - \eta_t g_t.$$

Here, what is a subgradient g_t? Note that

$$\underbrace{\partial\left(f^*(-A^\top \mu_t) + b^\top \mu_t\right)}_{\text{subdifferential of } f^*(-A^\top\mu) + b^\top \mu \text{ at } \mu = \mu_t} = -A \underbrace{\partial f^*(-A^\top \mu_t)}_{\text{subdifferential of } f^*(\mu) \text{ at } \mu = -A^\top \mu_t} + b.$$

Hence, $g_t \in \partial\left(f^*(-A^\top \mu_t) + b^\top \mu_t\right)$ if and only if

$$g_t \in -A\partial f^*(-A^\top \mu_t) + b.$$

Therefore,

$$g_t = -Ax_t + b \quad \text{for some } x_t \in \partial f^*(-A^\top \mu_t).$$

Moreover, we have also observed that $x_t \in \partial f^*(-A^\top \mu_t)$ if and only if $-A^\top \mu_t \in \partial f(x_t)$. Here, $-A^\top \mu_t \in \partial f(x_t)$ holds if and only if $0 \in \partial f(x_t) + A^\top \mu_t$ which is equivalent to

$$x_t \in \operatorname*{argmin}_x f(x) + \mu_t^\top Ax.$$

Note that $\mu_t^\top b$ remains constant as x changes, so $x_t \in \operatorname{argmin}_x f(x) + \mu_t^\top Ax$ is equivalent to

$$x_t \in \operatorname*{argmin}_x f(x) + \mu_t^\top (Ax - b).$$

Therefore, the subgradient method applied to the dual problem proceeds with

$$
\begin{aligned}
x_t &\in \operatorname*{argmin}_x f(x) + \mu_t^\top (Ax - b), \\
\mu_{t+1} &= \mu_t + \eta_t (Ax_t - b).
\end{aligned}
$$

Here, $f(x) + \mu_t^\top (Ax - b)$ is the Lagrangian function $\mathcal{L}(x, \mu)$ at $\mu = \mu_t$. In words, the subgradient method applied to the dual problem works as follows. At each iteration t with a given dual multiplier μ_t, we find a minimizer of the Lagrangian function $\mathcal{L}(x, \mu_t)$. Then we use the corresponding dual subgradient $Ax_t - b$ to obtain a new multiplier μ_{t+1}.

Algorithm 2 Subgradient method for the dual problem

Initialize μ_1.
for $t = 1, \ldots, T-1$ **do**
  Obtain $x_t \in \operatorname{argmin}_x f(x) + \mu_t^\top (Ax - b)$,
  Update $\mu_{t+1} = \mu_t + \eta_t (Ax_t - b)$ for a step size $\eta_t > 0$.
end for

At each iteration, we find a minimizer of the Lagrangian function $\mathcal{L}(x, \mu_t)$, which gives rise to an unconstrained optimization problem. Hence, the dual approach is useful when there is a complex system of constraints.