Lecture 2

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Junbo Hao

1 Duality Theory

1.1 Motivation Examples

Example 1.1. Let us consider the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{x}} f(\mathbf{x}),\\ &s.t.\ c_i(\mathbf{x}) = 0, i = 1, \dots, m.\end{aligned}$$

If $c_i(\mathbf{x}) = \mathbf{a}_i^\top \mathbf{x} - b_i$, we have the optimality condition for constrains $A\mathbf{x} = \mathbf{b}$. If c_i is not a linear function, the optimality condition is $\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geqslant 0$, for all $\mathbf{y} \in \mathcal{X} = \{\mathbf{x} | c_i(\mathbf{x}) = 0, i = 1, \dots, m\}$. This means we have no an equation system to solve the optimal point compared with the equality constrains.

Example 1.2. LAD regression: $\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_1$.

- Sub-gradient descent: $\mathbf{x}^{t+1} = \mathbf{x}^t - s_t \partial_{\|\cdot\|_1}(A\mathbf{x}^t - b)$. The speed is $O(\frac{1}{\sqrt{T}})$.
- Proximal Gradient Descent: consider $\min_{\mathbf{x}} f(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_1$, where $f(\mathbf{x}) = 0$. Then the corresponding PGD algorithm is

$$\begin{cases}\mathbf{x}^{t+1} = prox_{\alpha\|A\mathbf{x}-\mathbf{b}\|_1}(\mathbf{x}^t),\\ prox_{\alpha\|A\mathbf{x}-\mathbf{b}\|_1}(\mathbf{x}^t) = \arg\min\{\frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^t\|^2 + \|A\mathbf{x} - \mathbf{b}\|_1\}.\end{cases}$$

Example 1.3. Fused LASSO [**?**]:

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda\|F\mathbf{x}\|_1, \tag{1}$$

where $F \in \mathbb{R}^{(n-1)\times n}$ and

$$F_{ij} = \begin{cases}1,\ j = i + 1,\\ -1,\ j = i,\\ 0,\ otherwise.\end{cases}$$

1.2 The Lagrange Dual Function

We consider that

$$\begin{aligned}\min_{\mathbf{x}} & f_0(\mathbf{x}),\\ s.t.\ & f_i(\mathbf{x}) \leqslant 0, i = 1, \ldots, m,\\ & h_j(\mathbf{x}) = 0, j = 1, \ldots, l.\end{aligned}$$

Definition 1.4. We define that *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ is

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{l} \nu_j h_j(\mathbf{x}), \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_l)^\top$ are denoted as *dual variables or Lagrange multipliers*.

Definition 1.5. Define the *Lagrange dual function* as

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}), \tag{3}$$

where $D = \{\cap_{i=0}^{m} \mathrm{dom}(f_i)\} \cap \{\cap_{j=1}^{l} \mathrm{dom}(h_j)\}$.

Theorem 1.6. *Let us define that* $p^* = \min_{\mathbf{x} \in \mathcal{X}} f_0(\mathbf{x})$, *then*

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leqslant p^*$$

for any $\boldsymbol{\lambda} \succeq 0$.

Proof. Suppose that $\bar{\mathbf{x}} \in \mathcal{X}$, then $\sum_{i=1}^{m} \lambda_i f_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \nu_j h_j(\bar{\mathbf{x}}) \leqslant 0$. Thus,

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leqslant L(\bar{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= f_0(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i f_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \nu_j h_j(\bar{\mathbf{x}}) \\ &\leqslant f_0(\bar{\mathbf{x}}), \end{aligned}$$

for all $\bar{\mathbf{x}} \in \mathcal{X}$. Therefore, $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leqslant f_0(\mathbf{x}^*) = p^*$. ■

Remark 1.7.
- *Theorem 1.6 shows the Lagrange dual function gives a nontrivial lower bound on* p^* *only when* $\boldsymbol{\lambda} \succeq 0$ *and* $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathrm{dom}(g)$. *We refer to a pair* $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathrm{dom}(g)$ *with* $\boldsymbol{\lambda} \succeq 0$ *as dual feasible variables.*
- $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ *is always concave.*

Definition 1.8. For each pair $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathrm{dom}(g)$ with $\boldsymbol{\lambda} \succeq 0$, the Lagrange dual function gives us a lower bound of p^*. A natural question is what is the best lower bound that can be obtained form the Lagrange dual function. This leads to the following optimization problem:

$$q^* = \max_{\boldsymbol{\lambda},\boldsymbol{\nu}} g(\boldsymbol{\lambda},\boldsymbol{\nu}), \tag{4}$$

$$s.t.\ \boldsymbol{\lambda} \succeq 0. \tag{5}$$

The previous problem is called *Lagrange dual problem* and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are the *dual optimal variables or optimal Lagrange multipliers*.

The Lagrange dual problem is a convex optimization since the objective to be maximized is concave and the constraint is convex, whether or not the primal problem is convex.

Definition 1.9. Weak Duality: $q^* \leqslant p^*$.

Strong Duality: $q^* = p^*$.

Remark 1.10.
- *Weak duality always holds. However, strong duality needs more well conditions.*
- *Let us discuss the following fact first:*

$$\sup_{\boldsymbol{\lambda}\succeq 0}\{f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x})\} = \begin{cases} f_0(\mathbf{x}), & f_i(\mathbf{x}) \leqslant 0, i = 1,\dots,m \\ \infty, & otherwise. \end{cases}$$

Thus, we have

$$p^* = \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda}\succeq 0} L(\mathbf{x},\boldsymbol{\lambda}),$$

$$q^* = \sup_{\boldsymbol{\lambda}\succeq 0} \inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda}).$$

Therefore, the weak duality implies that

$$\sup_{\boldsymbol{\lambda}\succeq 0} \inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda}) \leqslant \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda}\succeq 0} L(\mathbf{x},\boldsymbol{\lambda}).$$

Definition 1.11. We refer to a pair $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ as a *saddle-point* for f if

$$f(\bar{\mathbf{x}}, \mathbf{y}) \leqslant f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leqslant f(\mathbf{x}, \bar{\mathbf{y}}),$$

for all $(\mathbf{x}, \mathbf{y}) \in \mathrm{dom}(f)$. In other words, $\bar{\mathbf{x}}$ minimizes $f(\mathbf{x}, \bar{\mathbf{y}})$ and $\bar{\mathbf{y}}$ minimizes $f(\bar{\mathbf{x}}, \mathbf{y})$. Saddle-point problems play an important role in **Game Theory and Generative Adversarial Networks**.

Example 1.12.

$$\min \|\mathbf{x}\|^2,$$

$$s.t.\ A\mathbf{x} = \mathbf{b}.$$

- Lagrangian: $L(\mathbf{x},\boldsymbol{\nu}) = \|\mathbf{x}\|^2 + \boldsymbol{\nu}^\top (A\mathbf{x} - \mathbf{b})$.

- Lagrange Dual Function: $g(\boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\nu})$. We know that $\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\nu}) = 2\mathbf{x} + A^\top \boldsymbol{\nu} = 0$, thus $\mathbf{x}^* = -\frac{1}{2} A^\top \boldsymbol{\nu}$. Take $\mathbf{x}^*$ into Lagrangian, we obtain the Lagrange dual function
$$g(\boldsymbol{\nu}) = -\frac{1}{4} \boldsymbol{\nu}^\top A A^\top \boldsymbol{\nu} - \boldsymbol{\nu}^\top \mathbf{b}$$
.
- Dual problem: $\max -\frac{1}{4} \boldsymbol{\nu}^\top A A^\top \boldsymbol{\nu} - \boldsymbol{\nu}^\top \mathbf{b}$.
- Weak duality:
$$\sup_{\boldsymbol{\nu}} \{-\frac{1}{4} \boldsymbol{\nu}^\top A A^\top \boldsymbol{\nu} - \boldsymbol{\nu}^\top \mathbf{b}\} \leqslant \min_{\mathbf{x}} \{\|\mathbf{x}\|^2 | A\mathbf{x} = \mathbf{b}\}.$$

Example 1.13. (Linear Programming) Recall the example of transportation problem in OM.

$$\begin{aligned} &\min_{\mathbf{x}} \mathbf{c}^\top \mathbf{x}, \\ &s.t.\ A\mathbf{x} = \mathbf{b}, \\ &\mathbf{x} \succeq 0. \end{aligned}$$

- Lagrangian:
$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathbf{c}^\top \mathbf{x} - \boldsymbol{\lambda}^\top \mathbf{x} + \boldsymbol{\nu}^\top (A\mathbf{x} - \mathbf{b}) = (\mathbf{c} - \boldsymbol{\lambda} + A^\top \boldsymbol{\nu})^\top \mathbf{x} - \boldsymbol{\nu}^\top \mathbf{b}.$$
- Lagrange Dual Function:
$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\boldsymbol{\nu}^\top \mathbf{b},\ \mathbf{c} - \boldsymbol{\lambda} + A^\top \boldsymbol{\nu} = 0, \\ -\infty,\ otherwise. \end{cases}$$
- Dual problem:
$$\begin{aligned} &\max_{\boldsymbol{\lambda}, \boldsymbol{\nu}}\ -\boldsymbol{\nu}^\top \mathbf{b}, \\ &s.t.\ \mathbf{c} - \boldsymbol{\lambda} + A^\top \boldsymbol{\nu} = 0, \\ &\boldsymbol{\lambda} \succeq 0. \end{aligned}$$

  This is equivalent to
$$\begin{aligned} &\min_{\boldsymbol{\nu}} \boldsymbol{\nu}^\top \mathbf{b}, \\ &s.t.\ \mathbf{c} + A^\top \boldsymbol{\nu} \succeq 0. \end{aligned}$$

Example 1.14.

$$\begin{aligned} &\min_{\mathbf{x}} \|\mathbf{x}\|, \\ &s.t.\ A\mathbf{x} = \mathbf{b}. \end{aligned}$$

It seems that we cannot obtain the Lagrange dual function via the directly derivation. How to do? We will learn and adapt conjugate function to handle this problem.