Optimization Theory and Algorithm II	**September 19, 2022**

Lecture 4

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 KKT Conditions

1.1 The Lagrange Dual Function

We consider that

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

Definition 1 *We define that Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ is*

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

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

Definition 2 *Define the Lagrange dual function as*

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}\in D} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}), \tag{2}$$

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

Theorem 1 *Let us define that $p^* = \min_{\mathbf{x}\in\mathcal{X}} f_0(\mathbf{x})$, then*

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

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

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

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

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

Remark 1
- *Theorem 1 shows the Lagrange dual function gives a nontrivial lower bound on p^* only when $\boldsymbol{\lambda} \succeq 0$ and $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in dom(g)$. We refer to a pair $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in dom(g)$ with $\boldsymbol{\lambda} \succeq 0$ as dual feasible variables.*

- $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ *is always concave.*

Definition 3 *For each pair* $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in dom(g)$ *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 from the Lagrange dual function. This leads to the following optimization problem:*

$$q^* = \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}), \tag{3}$$

$$s.t.\ \boldsymbol{\lambda} \succeq 0. \tag{4}$$

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 4 ***Weak Duality***: $q^* \leq p^*$.

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

Remark 2

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

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

  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

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

1.2 Benefit of Strong Duality

Theorem 2 *Suppose that* $\mathbf{x}^*$ *and* $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *are the primal and dual solution of optimization problem of* (**??**), *and strong duality holds. Then we have the following two facts:*

- $\sum_i \lambda_i^* f_i(\mathbf{x}^*) = 0$. *That is* $\lambda_i^* > 0, \Longrightarrow f_i(\mathbf{x}^*) = 0$ *or* $f_i(\mathbf{x}^*) < 0, \Longrightarrow \lambda_i^* = 0$. *This is also called "complementary slackness."*
- $\mathbf{x}^*$ *is the minimizer of* $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$, *that is*

$$\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0.$$

Proof 2 *Due to the strong duality, then*

$$\begin{aligned} p^* = f_0(\mathbf{x}^*) = q^* &= \inf_{\mathbf{x} \in D} \left\{ f_0(\mathbf{x}) + \sum_i \lambda_i^* f_i(\mathbf{x}) + \sum_j \nu_j^* h_j(\mathbf{x}) \right\} \\ &\leq f_0(\mathbf{x}^*) + \sum_i \lambda_i^* f_i(\mathbf{x}^*) + \sum_j \nu_j^* h_j(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*). \end{aligned}$$

This implies

$$\sum_i \lambda_i^* f_i(\mathbf{x}^*) = 0$$

and $\mathbf{x}^*$ *is the minimizer of* $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$. *In addition,*

$$\nabla L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = 0 \Longrightarrow \nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0.$$

Under strong duality, given a dual solution $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ any primal solution $\mathbf{x}^*$ solves

$$\min_{\mathbf{x}} f_0(\mathbf{x}) + \sum_i \lambda_i^* f_i(\mathbf{x}) + \sum_j \nu_j^* h_j(\mathbf{x}).$$

This means that we only need to solve an unconstrained problem we have familiar with them.

1.3 Karush-Kuhn-Tucker Conditions

- First appeared in publication by Kuhn and Tucker 1951.
- Later people found out that Karush had the condition in his unpublished master's thesis of 1939.
- Finally, it is called the Karush-Kuhn-Tucker conditions.

Theorem 3 *(KKT Optimality Conditions) Let* $\mathbf{x}^*$ *and* $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *be the primal and dual optimal points with zero dual gap, then the following KKT conditions hold:*

$$\begin{aligned}
\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \ \textit{(stationary point)}, \quad & (5)\\
f_i(\mathbf{x}^*) \le 0, \ \textit{(primal feasible)} \quad & (6)\\
h_j(\mathbf{x}^*) = 0, \ \textit{(primal feasible)} \quad & (7)\\
\lambda_i^* f_i(\mathbf{x}^*) = 0, \ \textit{(complementary slackness)} \quad & (8)\\
\lambda_i \ge 0, \ \textit{(dual feasible)} \quad & (9)
\end{aligned}$$

where $i = 1, \ldots, m$ *and* $j = 1, \ldots, l$.

Proof 3 *Combing the primal and dual feasible conditions and results of Theorem 2, we can justify the KKT optimality conditions.*

Next, let us show some insightful examples

Example 1 *For the unconstrained optimization, KKT optmality conditions say:* $\nabla f(\mathbf{x}^*) = 0$.

Example 2 *Let us consider the following general convex optimization with linear equality constrains.*

$$\min_{\mathbf{x}} \ f(\mathbf{x}), \quad (10)$$

$$s.t. \ A\mathbf{x} = \mathbf{b}. \quad (11)$$

Based on the KKT optimality conditions, we have

$$\begin{cases} A\mathbf{x}^* = \mathbf{b}, \\ \nabla f(\mathbf{x}^*) + A^\top \lambda^* = 0. \end{cases}$$

Recall that we have obtain these conditions by the general optimality conditions

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x} \rangle \geq 0$$

in the previous example.

Example 3

$$\begin{aligned} &\min_{\mathbf{x}} \ f_0(\mathbf{x}), \\ &s.t. \ \mathbf{x} \succeq 0. \end{aligned}$$

The Lagrangian: $L(\mathbf{x}, \boldsymbol{\lambda}) = f_0(\mathbf{x}) - \boldsymbol{\lambda}^\top \mathbf{x}$. *Then, the KKT conditions:*

$$\begin{aligned} \nabla f_0(\mathbf{x}^*) - \boldsymbol{\lambda}^* &= 0, \\ \mathbf{x}^* &\succeq 0, \\ \boldsymbol{\lambda}^* &\succeq 0, \\ \lambda_i^* x_i^* &= 0. \end{aligned}$$

Thus, $(\nabla f_0(\mathbf{x}^*))_i = \lambda_i^*$. *Finally, we have the optimality condition for* $\mathbf{x}^*$ *as*

$$\begin{aligned} (\nabla f_0(\mathbf{x}^*))_i x_i^* &= 0, \\ \nabla f_0(\mathbf{x}^*) &\succeq 0, \\ \mathbf{x}^* &\succeq 0. \end{aligned}$$

Theorem 3 shows the necessary condition of primal and dual optimal points which should satisfy. What about sufficient conditions?

Theorem 4 *Suppose that primal problem is convex,* $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *are any points that satisfies the KKT conditions, then* $\mathbf{x}^*$ *and* $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *are primal and dual optimal with zero dual gap.*

Proof 4 *KKT conditions tell us that* $\mathbf{x}^*$ *is primally feasible, namely* $f_i(\mathbf{x}^*) \leq 0$ *and* $h_j(\mathbf{x}^*) = 0$. *Since* $\boldsymbol{\lambda}^* \succeq 0$, *then* $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *is convex in* $\mathbf{x}$. *Thus, the condition* $\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0$ *indicates* $\mathbf{x}^*$ *minimizes* $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *over* $\mathbf{x}$. *Therefor,*

$$g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = f_0(\mathbf{x}^*) + \sum_i \lambda_i^* f_i(\mathbf{x}^*) + \sum_j \nu_j^* h_j(\mathbf{x}^*) = f_0(\mathbf{x}^*).$$

This means the zero dual gap. Obviously, $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *are primal and dual optimal points.*

Example 4

$$\begin{aligned} &\min_{\mathbf{x}} \ f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top P \mathbf{x} + q^\top \mathbf{x} + r, \\ &s.t. \ A\mathbf{x} = \mathbf{b}, \end{aligned}$$

where $P \succ 0$. *We know that this is a convex problem and its KKT conditions are*

$$\begin{cases} Ax^* = b, \\ P\mathbf{x}^* + q + A^\top \lambda^* = 0. \end{cases}$$

Based on Theorem 4, solving the so-called "KKT-system" can obtain the optimal solution.

Example 5 *(Support Vector Machine)*

Given a data set $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}, i = 1, \ldots, n\}$, *how to construct a linear classifier if the data set is separable?*

The basic idea is that we can use Separation Hyperplane Theorem to construct the classifier.

Recall that

Theorem 5 *Suppose that there are two convex sets* C *and* D *satisfies* $C \cap D = \emptyset$. *Then there exists* $\mathbf{a} \neq 0$ *and* b *such that*

$$\mathbf{a}^\top \mathbf{x} - b \leq 0 \text{ for any } \mathbf{x} \in C, \text{ and } \mathbf{a}^\top \mathbf{x} - b \geq 0 \text{ for any } \mathbf{x} \in D. \tag{12}$$

Proof 5 *Let* p, q *be the two pints which achieve*

$$\min_{\mathbf{x} \in C, \mathbf{y} \in D} \|\mathbf{x} - \mathbf{y}\| = \|p - q\|.$$

Then the hyperplan separates C *and* D *is*

$$\langle p - q, \mathbf{x} - \frac{p + q}{2} \rangle = 0,$$

that is

$$\langle p - q, \mathbf{x} \rangle - \frac{1}{2} \langle p - q, p + q \rangle = 0.$$

Thus, $\mathbf{a} = p - q$ *and* $b = \frac{1}{2} \langle p - q, p + q \rangle$.

Let us go back to the SVM example. According to the hyperplane separation theorem, we can construct the linear classifier by the following three steps:

- Step 1: Construct a positive and negtive convex hull

$$C_+ = \{\mathbf{x} | \mathbf{x} = \sum_{y_i = 1} \alpha_i \mathbf{x}_i, \sum_{y_i = 1} \alpha_i = 1, 0 \leq \alpha_i \leq 1\},$$
$$C_- = \{\mathbf{x} | \mathbf{x} = \sum_{y_i = -1} \alpha_i \mathbf{x}_i, \sum_{y_i = -1} \alpha_i = 1, 0 \leq \alpha_i \leq 1\}.$$

- Step 2: Find p and q for C_+ and C_-.
- Step 3: set $\mathbf{a} = p - q$ and $b = \frac{1}{2} \langle p - q, p + q \rangle$, we have the linear classifier $y = \mathbf{a}^\top \mathbf{x} + b$.

Q: How to find p and q? To this end, we need to find the optimal solution of the following optimization problem:

$$\begin{aligned} \min_{\alpha, \beta} \ & \frac{1}{2} \| \sum_{y_i = 1} \alpha_i \mathbf{x}_i - \sum_{y_i = -1} \beta_i \mathbf{x}_i \|^2, \\ s.t. \ & \sum_{y_i = 1} \alpha_i = 1, 0 \leq \alpha_i \leq 1, \\ & \sum_{y_i = -1} \beta_i = 1, 0 \leq \beta_i \leq 1. \end{aligned}$$

However, finding the optimal solution of the above optimization problem is relatively hard. Then in the machine learning community, another method called "maximal margin" approach that has been widely used

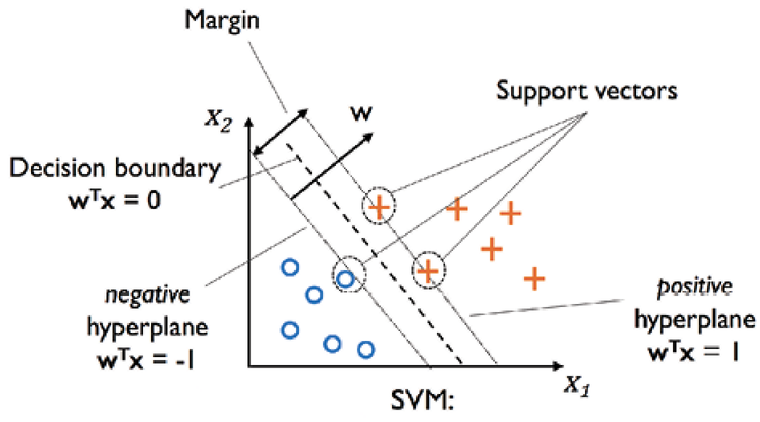

Figure 1: Support Vector Machine

to find the "optimal" linear classifier. The fundamental idea is to find two parallel hyperplanes (see Figure 1) which can separate the positive and negative point set with the maximal distance (margin).

With loss of generality, assume that the two parallel hyperplanes are $\langle \mathbf{w}, \mathbf{x} \rangle + b = 1$ and $\langle \mathbf{w}, \mathbf{x} \rangle + b = -1$. Then the maximal margin means

$$\max_{\mathbf{w},b} \mathbf{d} = \frac{2}{\|\mathbf{w}\|}, \tag{13}$$

$$s.t. \; y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, i = 1, \ldots, n. \tag{14}$$

It is equivalent to

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \tag{15}$$

$$s.t. \; y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, i = 1, \ldots, n. \tag{16}$$

Lagrangian:

$$L(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} - \sum_i \alpha_i [y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1].$$

KKT conditions:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0, \tag{17}$$

$$\nabla_b L(\mathbf{w}, b, \alpha) = -\sum_i \alpha_i y_i = 0, \tag{18}$$

$$\alpha_i \geq 0, \tag{19}$$

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \tag{20}$$

$$\alpha_i [y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] = 0. \tag{21}$$

So, it has $\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i$, then the linear classifier is $y = \langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = \sum_i \alpha_i^* y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b^*$. The point $\mathbf{x}_i$ is called the **support point** due to $\alpha_i \neq 0$. $\alpha_i \neq 0$ also indicates that point i lies on the support hyperplane.

Take $\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i$ into the Lagrangian, we have the Lagrange dual problem:

$$
\begin{aligned}
\max_{\alpha} \ & -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_i \alpha_i \\
s.t. \ & \alpha_i \geq 0, \\
& \sum_i \alpha_i y_i = 0.
\end{aligned}
$$

The primal and dual problems are convex, and the dual problem is quadratic.

References