Optimization Theory and Algorithm **Lecture 6 - 09/29/2021**

Lecture 6

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Junbo Hao

1 KKT Optimality Condition

1.1 Benefit of Strong Duality

Example 1.1. (Dual norm of dual norm is the primal norm.) Define dual norm of $\mathbf{x} \in \mathbb{R}^n$ w.r.t. $\|\cdot\|$ is $\|\mathbf{x}\|_* = \sup_{\|\mathbf{z}\| \leqslant 1} \mathbf{z}^\top \mathbf{x}$. Prove that

$$\|\mathbf{x}\|_{**} = \|\mathbf{x}\|.$$

Proof. Consider a trivial problem (given $\mathbf{x}$),

$$\min_{\mathbf{y}} \|\mathbf{y}\|,$$
$$s.t.\ \mathbf{y} = \mathbf{x}$$

where the optimal value $p^* = \|\mathbf{x}\|$. Let

$$L(\mathbf{u}, \boldsymbol{\nu}) = \|\mathbf{y}\| + \boldsymbol{\nu}^\top (\mathbf{x} - \mathbf{y}) = \|\mathbf{y}\| - \mathbf{y}^\top \boldsymbol{\nu} + \mathbf{x}^\top \boldsymbol{\nu}.$$

Thus, Lagrange dual function is

$$g(\boldsymbol{\nu}) = \inf_{\mathbf{y}} L(\mathbf{y}, \boldsymbol{\nu}) = \begin{cases} \mathbf{x}^\top \boldsymbol{\nu}, \ \|\boldsymbol{\nu}\|_* \leqslant 1, \\ -\infty, \ otherwise. \end{cases}$$

Then the dual problem is

$$\max_{\boldsymbol{\nu}} \mathbf{x}^\top \boldsymbol{\nu},$$
$$s.t.\ \|\boldsymbol{\nu}\|_* \leqslant 1.$$

According to the definition of dual norm and strong duality, then $\|\mathbf{x}\|_{**} = \|\mathbf{x}\|$. ■

Example 1.2. (Dual Gradient Ascent) Consider

$$\min_{\mathbf{x}} f(\mathbf{x}),$$
$$s.t.\ c(\mathbf{x}) = 0.$$

Lagrangian: $L(\mathbf{x}, \boldsymbol{\nu}) = f(\mathbf{x}) + \boldsymbol{\nu}^\top c(\mathbf{x})$. Thus,

$$g(\boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\nu}) = L(\mathbf{x}^*(\boldsymbol{\nu}), \boldsymbol{\nu}).$$

The dual problem is

$$\max_{\boldsymbol{\nu}} g(\boldsymbol{\nu}).$$

Because we have

$$\nabla g(\boldsymbol{\nu}) = \frac{\partial L}{\partial \mathbf{x}^*} \frac{\partial \mathbf{x}^*}{\partial \boldsymbol{\nu}} + \frac{\partial L}{\partial \boldsymbol{\nu}} = c(\mathbf{x}),$$

where $\frac{\partial L}{\partial \mathbf{x}^*} = 0$. Based on that, the dual gradient assent algorithm is

$$\textbf{Step 1: } \mathbf{x}^t = \arg\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\nu}^t), \tag{1}$$

$$\textbf{Step 2: } \boldsymbol{\nu}^{t+1} = \boldsymbol{\nu}^t + s_t c(\mathbf{x}^t)). \tag{2}$$

1.2 Karush-Kuhn-Tucker Conditions

Theorem 1.3. *(KKT Optimality Conditions) Let* $\mathbf{x}^*$ *and* $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ *be the primal and dual optimal points of optimization problem of* (**??**) *with zero dual gap, then the following KKT conditions hold:*

$$\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \text{ (stationary point)}, \tag{3}$$

$$f_i(\mathbf{x}^*) \leqslant 0, \text{ (primal feasible)} \tag{4}$$

$$h_j(\mathbf{x}^*) = 0, \text{ (primal feasible)} \tag{5}$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \text{ (complementary slackness)} \tag{6}$$

$$\lambda_i \geqslant 0, \text{ (dual feasible)} \tag{7}$$

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

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

Theorem 1.4. *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. KKT conditions tell us that $\mathbf{x}^*$ is primally feasible, namely $f_i(\mathbf{x}^*) \leqslant 0$ 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 1.5. (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 1.6. *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*

$$\mathbf{a}^\top \mathbf{x} - b \leqslant 0 \text{ for any } \mathbf{x} \in C, \text{ and } \mathbf{a}^\top \mathbf{x} - b \geqslant 0 \text{ for any } \mathbf{x} \in D. \tag{8}$$

Proof. 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

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

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

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

- 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:

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

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 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).

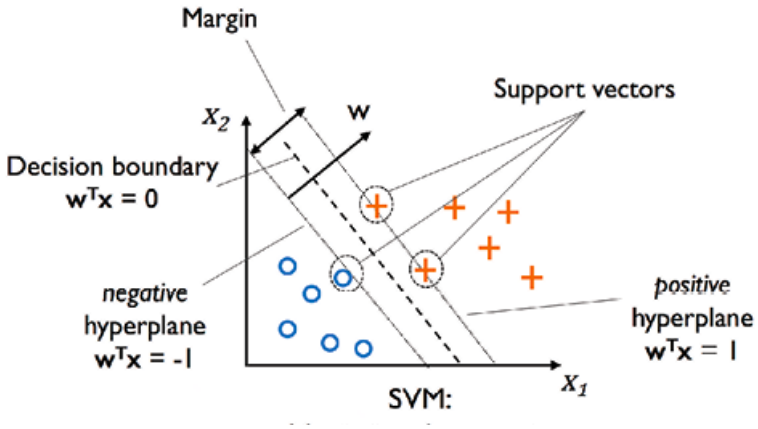

Figure 1: Support Vector Machine

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

$$\max_{\mathbf{w},b} \mathbf{d} = \frac{2}{\|\mathbf{w}\|}, \tag{9}$$
$$s.t.\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geqslant 1, i = 1, \ldots, n. \tag{10}$$

It is equivalent to

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \tag{11}$$
$$s.t.\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geqslant 1, i = 1, \ldots, n. \tag{12}$$

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:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0, \tag{13}$$

$$\nabla_b L(\mathbf{w}, b, \alpha) = -\sum_i \alpha_i y_i = 0, \tag{14}$$

$$\alpha_i \geqslant 0, \tag{15}$$

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geqslant 1, \tag{16}$$

$$\alpha_i[y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] = 0. \tag{17}$$

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:

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

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