Lecture 8. Support Vector Machine II

Lecturer: Jie Wang Date: May 7, 2021

Duality plays an important role in analyzing SVM. Besides interesting theoretical results, duality also motives many efficient algorithms for solving SVM. In this section, we introduce ***elements of Lagrangian duality***. There are several different approaches to Lagrangian duality. We follow the approach introduced in [1, 2], which are based on geometric observations.

1 The Primal Problem

We consider the problem—that is, the ***primal problem***—as follows.

$$\begin{aligned} \min_{\mathbf{x}} \, & f(\mathbf{x}) \\ \text{s.t.} \, & g_i(\mathbf{x}) \leq 0, \, i = 1, \dots, m, \\ & h_i(\mathbf{x}) = 0, \, i = 1, \dots, p, \\ & \mathbf{x} \in X, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [m]$, and $h_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [p]$, are all continuously differentiable, and $X \subseteq \mathbb{R}^n$. To simplify notations, let $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector function whose i^{th} component is g_i, and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ be a vector function whose i^{th} component is h_i. Then, we can write the problem in (1) in a more compact form as follows.

$$\begin{aligned} \min_{\mathbf{x}} \, & f(\mathbf{x}) \\ \text{s.t.} \, & \mathbf{g}(\mathbf{x}) \leq 0, \\ & \mathbf{h}(\mathbf{x}) = 0, \\ & \mathbf{x} \in X. \end{aligned} \tag{2}$$

We denote the ***feasible set*** of (2) by

$$D_0 = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X\}. \tag{3}$$

Each element in D_0 is called a ***feasible solution***. The ***optimal function value*** is

$$f^* = \inf_{\mathbf{x} \in D_0} f(\mathbf{x}). \tag{4}$$

Assumption 1. Feasibility and Boundedness *The feasible set is nonempty and the objective function is bounded from below, that is,*

$$-\infty < f^* = \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) < \infty.$$

Remark 1. Notice that, Assumption 1 does not assume the existence of the optimum of the problem in (2).

2 Geometric Observations

We used to analyze and/or solve optimization problems by focusing on the problem domain D_0, as the variable $\mathbf{x}$ lies in D_0, and so does the optimum we are looking for (if it exists). Surprisingly, taking the perspective of the ***constraint-cost pairs*** as $\mathbf{x}$ goes over X, that is, the subset of $\mathbb{R}^{m+p+1}$

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) : \mathbf{x} \in X\}, \tag{5}$$

brings us fresh insights. Figure 1 shows a simple example of S for problems with only one inequality constraint. Indeed, the key idea to Lagrangian duality in [1, 2] is to *interpret the primal problem (2) by the geometric properties of the set S via hyperplanes*.

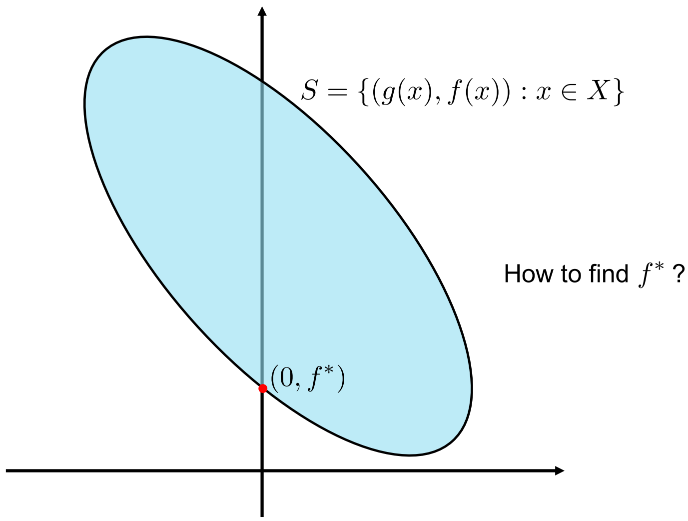

Figure 1: Illustration of the set S of the constraint-cost pairs for a simple problem with only one inequality constraint.

2.1 The Lagrangian

To illustrate the idea of Lagrangian duality from a geometric perspective, we first consider a simple problem with only one inequality constraint. We show the set S of constraint-cost pairs in Figure 1. We can see that, the optimal function value f^* of the primal problem is indeed the second component of the red dot, that is, the point with the smallest value of the second component among the points whose first components are non-positive.

Thus, a natural question arises, *instead of solving the primal problem (2), can we find the optimal function value f^* by analyzing the set S?* The answer is yes. The working horse is the (simple) hyperplanes. The linear function that specifies the hyperplanes is the so-called ***Lagrangian***.

Definition 1. Associated with the primal problem, we define the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}).$$

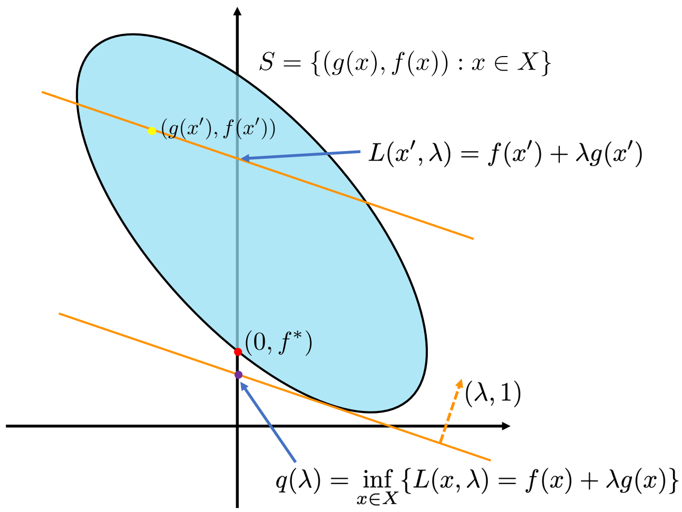

Figure 2: Illustration of the hyperplanes specified by the Lagrangian.

2.2 Hyperplanes Defined by the Lagrangian

Hyperplanes can be specified by level sets of linear functions. Given a constant c, the Lagrangian defines a hyperplane in $\mathbb{R}^{m+p+1}$—where the set S of constraint-cost pairs lies in—by

$$H_L(c) = \{(\mathbf{y}, \mathbf{w}, z) : z + \langle \lambda, \mathbf{y} \rangle + \langle \mu, \mathbf{w} \rangle = c, z \in \mathbb{R}, \mathbf{y} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^p\}.$$

The normal of $H_L(c)$ is $(\lambda, \mu, 1)$, which implies that the hyperplane is ***nonvertical*** (why?).

Figure 2 shows the hyperplanes defined by the Lagrangian for a simple problem with only inequality constraints. The two hyperplanes share the same normal vector $(\lambda, 1)$. As the function value of the Lagrangian at the yellow point $(g(x'), f(x'))$ is clearly given by

$$L(x', \lambda) = f(x') + \lambda g(x'),$$

the hyperplane which go through the point $(g(x'), f(x'))$ is

$$H_L(L(x', \lambda)) = \{(y, z) : z + \lambda y = L(x', \mu) = f(x') + \lambda g(x')\}.$$

Moreover, we can see that, *the hyperplane $H_L(L(x', \lambda))$ intercepts the vertical axis $\{(0, z) : z \in \mathbb{R}\}$ at the level $L(x', \lambda)$*.

2.3 The Lagrangian Dual Function

The geometric properties we observe in Section 2.2 lead us to the fact that, for a nonvertical hyperplane, the level of interception of the vertical axis is indeed the (linear) function value that defines the hyperplane. Thus, given a vector $(\lambda, \mu, 1) \in \mathbb{R}^{m+p+1}$, if we define

$$q(\lambda, \mu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu), \tag{6}$$

the hyperplane $H_L(q(\lambda, \mu))$ intercepts the vertical axis at the level $q(\lambda, \mu)$ if it exists. Figure 2 shows a simple example where a hyperplane intercepts the vertical axis at the level $q(\lambda)$.

In general, What is the relationship between $q(\lambda, \mu)$ and f^*? In view of Figure 2, a reasonable guess would be

$$q(\lambda, \mu) \leq f^*, \forall \lambda \geq 0,$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ and $\lambda \geq 0$ is an abbreviation for $\lambda_i \geq 0$, $i \in [m]$. Indeed, the above guess is true, and we have the result as follows.

Lemma 1. *For any $\lambda \geq 0$, the following result holds:*

$$q(\lambda, \mu) \leq f^*.$$

Proof. By definition, we have

$$\begin{aligned} q(\lambda, \mu) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu) \\ &= \inf_{\mathbf{x} \in X} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}) \\ &\leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}). \end{aligned}$$

The definition of D_0 implies that, for any $\mathbf{x} \in D_0$, we have

$$g_i(\mathbf{x}) \leq 0, i \in [m], \text{ and } h_i(\mathbf{x}) = 0, i \in [p].$$

Thus, the above inequality becomes

$$q(\lambda, \mu) \leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}) \leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) = f^*,$$

which completes the proof. □

The function $q(\lambda, \mu)$ is the so-called ***Lagrangian dual function***. The domain of q is the set for which $q(\lambda, \mu)$ is finite:

$$\mathbf{dom}\, q = \{(\lambda, \mu) : q(\lambda, \mu) > -\infty\}.$$

We can similarly define the ***dual feasible set*** by

$$D_1 = \{(\lambda, \mu) : \lambda \geq 0\} \cap \mathbf{dom}\, q = \{(\lambda, \mu) : \lambda \geq 0, q(\lambda, \mu) > -\infty\}.$$

Remark *2.* We do not require that $\lambda \geq 0$ for the points in $\mathbf{dom}\,(q)$.

A surprising result is that, the Lagrangian dual function q is concave, no matter the primal problem is convex or not.

Theorem 1. *The domain of q is convex and q is concave over* $\mathbf{dom}\,(q)$.

Proof. We first show that $\mathbf{dom}\,(q)$ is convex.

Suppose that $q(\lambda_1, \mu_1)$ and $q(\lambda_2, \mu_2)$ are finite and $(\lambda_1, \mu_1) \neq (\lambda_2, \mu_2)$. Let $\theta \in [0, 1]$.

$$\begin{aligned} q(\theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\mu_2) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\mu_2) \\ &= \inf_{\mathbf{x} \in X} \theta L(\mathbf{x}, \lambda_1, \mu_1) + (1-\theta) L(\mathbf{x}, \lambda_2, \mu_2) \\ &\geq \inf_{\mathbf{x} \in X} \theta L(\mathbf{x}, \lambda_1, \mu_1) + \inf_{\mathbf{x} \in X} (1-\theta) L(\mathbf{x}, \lambda_2, \mu_2) \\ &= \theta q(\lambda_1, \mu_1) + (1-\theta) q(\lambda_2, \mu_2) \\ &> -\infty. \end{aligned}$$

Thus, we have $\mathbf{dom}\,(q)$ is convex.

The concavity of q can easily be seen by noting that q is the infimum of a set of linear functions of (λ, μ). □

References

[1] M. Bazaraa, H. Sherali, and C. Shetty. *Nonlinear Programming.* Wiley-Interscience, 2006.

[2] D. P. Bertsekas. *Nonlinear Programming, 3ed.* Athena Scientific, 2016.