9 Lagrangian Duality

In the previous lectures, we saw that LP duality has remarkable implications in the theory of LP and ILP. Today, we are going to see that the theory of duality can be extended to other optimizations. The very fundamental idea for the duality theory is to provide lower (upper) bounds for the minimization (maximization) optimization.

We are going to talk about a method which is widely attributed to Lagrange, but of course he did not know at his time about the general duality theory. In fact, what we call the Lagrangian duality theory is formalized in the twentieth century, along with the LP duality. The Lagrangian duality establishes a weak duality result (a loose bound), but for convex functions it can be a strong duality theory (exact optimal value). For the differentiable convex functions, there is an equivalent for the complementary slackness conditions, known as the Karush-Kuhn-Tucker (KKT) conditions. These are very popular concepts in the modern optimization theory.

Let us start from the most general form of the Lagrange duality theory. I will focus without loss of generality on minimization problems. Take a general optimization problem as in (2):

$$\begin{array}{c} f^* = \min\limits_{\mathbf{x}\in\Psi} f(\mathbf{x}) \\ \text{subject to} \\ \begin{array}{ll} g_1(\mathbf{x}) \leq 0 & h_1(\mathbf{x}) = 0 \\ g_2(\mathbf{x}) \leq 0 & h_2(\mathbf{x}) = 0 \\ \vdots \quad , & \vdots \\ g_m(\mathbf{x}) \leq 0 & h_p(\mathbf{x}) = 0 \end{array} \end{array} \tag{75}$$

Now, we are going to perform similar steps as in the LP duality theory. Define a dual variable $u_i \leq 0$ for each constraint $g(\mathbf{x}) \leq 0$. If a constraint is in the form $g(\mathbf{x}) \leq 0$ we will take $u_i \geq 0$, but this is not the case in our example. Take also free dual variables $v_i \in \mathbb{R}$ for the constraints $h_i(\mathbf{x}) = 0$. Remember that in the linear LP case, we made the linear combination of the constraints:

$$\Phi(\mathbf{x}, \mathbf{u}, \mathbf{v}) = u_1 g_1(\mathbf{x}) + u_2 g_2(\mathbf{x}) + \ldots + u_m g_m(\mathbf{x}) + v_1 h_1(\mathbf{x}) + v_2 h_2(\mathbf{x}) + \ldots + v_p h_p(\mathbf{x}) \geq 0 \tag{76}$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_m)^T$ and $\mathbf{v} = (v_1, v_2, \ldots, v_m)^T$. In the LP theory at this point, we would define constraints over the dual variables, in order to enforce $\Phi(\mathbf{x}, \mathbf{u}, \mathbf{v})$ to be similar to $f(\mathbf{x})$. Then we could simply get a good inequality for $f(\mathbf{x})$. However, for general functions, we cannot have any kind of similarity between $\Phi(\mathbf{x}, \mathbf{u}, \mathbf{v})$ and f. A brilliant observation by Lagrange helps us to proceed with this issue. Define the **Lagrangian form**:

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) - \Phi(\mathbf{x}, \mathbf{u}, \mathbf{v}) \tag{77}$$

and consider the minimization

$$\Gamma(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x}\in\mathbf{\Psi}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \tag{78}$$

Notice that the optimization in (78) is unconstrained. The function $\Gamma(\mathbf{u}, \mathbf{v})$ is called the **Lagrangian dual** function. I am going to show that for any $\mathbf{u} \leq 0$ and $\mathbf{v}$, we have that $f^* \geq \Gamma(\mathbf{u}, \mathbf{v})$:

Denote the optimal solution of (75) by $\mathbf{x}^*$. Then, $\Gamma(\mathbf{u}, \mathbf{v}) \leq L(\mathbf{x}^*, \mathbf{u}, \mathbf{v})$ simply because $\Gamma(\mathbf{u}, \mathbf{v})$ is the minimum value. Further, we know that $\mathbf{x}^*$ is feasible in (75). Hence, according to (76), we have that $\Phi(\mathbf{x}^*, \mathbf{u}, \mathbf{v}) \geq 0$ and $L(\mathbf{x}^*, \mathbf{u}, \mathbf{v}) = f^* - \Phi(\mathbf{x}^*, \mathbf{u}, \mathbf{v}) \leq f^*$. This gives the result as

$$\Gamma(\mathbf{u}, \mathbf{v}) \leq L(\mathbf{x}^*, \mathbf{u}, \mathbf{v}) \leq f^*. \tag{79}$$

Finally, notice that $f^* \geq \Gamma(\mathbf{u}, \mathbf{v})$ is true for any $\mathbf{u} \leq 0$ and $\mathbf{v}$. Hence, the best lower bound is obtained by maximizing $\Gamma(\mathbf{u}, \mathbf{v})$ over all feasible dual variables:

$$\begin{array}{c} \Gamma^* = \max\limits_{\mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^p} \Gamma(\mathbf{u}, \mathbf{v}) \\ \text{subject to} \\ \mathbf{u} \geq \mathbf{0} \end{array} \tag{80}$$

The optimization in (80) is called the **Lagrangian dual optimization**. The **Lagrangian weak duality** states that $\Gamma^* \leq f^*$.

Example 23. Support Vector Machine: Stroke (brain attack) is the second most important cause of death in the world. In order to decrease the death chance and also its long-lasting effects, it is important to diagnose the stroke in a very short time. The main difficulty is that there exist two different types of stroke, ischemic (clots) and hemorrhagic (bleeding), which require different types of treatment. Computer-assisted techniques are designed to enhance diagnosing the stroke type by rapidly collecting a number of clinical parameters, such as blood pressure, a number of electroencephalogram (EEG) parameters and etc.

Currently, there does not exists a good model to link the clinical parameters and the type of stroke. Our purpose is to learn a simple model from a number of clinical trials, which enhances predicting the stroke type. Let us assume that there are m diagnosed patients and denote the clinical parameters of the k^{th} patient by

$$\mathbf{x}_k = \begin{bmatrix} x_{1,k} \\ x_{2,k} \\ \dots \\ x_{n,k} \end{bmatrix}, \tag{81}$$

where m is the number of parameters for each patient. Let $t_k = 1$ if the k^{th} patient is diagnosed by the ischemic stroke and $t_k = 0$ otherwise. We are to learn a simple rule to detect the type of the stroke. Our detector consists of a weighting vector $\mathbf{w}$ and a threshold b. For a given data vector $\mathbf{x}$, it will set $t = 0$ if $\mathbf{w}^T\mathbf{x} + b < 0$ and $t = 1$ otherwise. The SVM method suggests to solve the following optimization to obtain the best predictor:

$$\begin{gathered} \min_{\mathbf{w}\in\mathbb{R}^n, b\in\mathbb{R}} \|\mathbf{w}\|_2^2 \\ \forall i \in [m], \quad \begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq 1 & t_i = 1 \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1 & t_i = 0 \end{cases} \end{gathered} \tag{82}$$

Let us write the dual optimization. First, define the dual variable u_i for the i^{th} constraint. Notice that $u_i \geq 0$ for $t_i = 1$ and $u_i \leq 0$ for $t_i = 0$. Now, let us write the Lagrangian form

$$L(\mathbf{w}, b, \mathbf{u}) = \|\mathbf{w}\|_2^2 - \sum_{i|t_i=1} (\mathbf{w}^T\mathbf{x}_i + b - 1)u_i - \sum_{i|t_i=0} (\mathbf{w}^T\mathbf{x}_i + b + 1)u_i \tag{83}$$

which can also be written as

$$L(\mathbf{w}, b, \mathbf{u}) = \|\mathbf{w}\|_2^2 - \mathbf{w}^T\mathbf{c} - b\beta + \alpha \tag{84}$$

where

$$\mathbf{c} = \sum_{i=1}^m u_i\mathbf{x}_i, \quad \beta = \sum_{i=1}^m u_i, \quad \alpha = \sum_{i|t_i=1} u_i - \sum_{i|t_i=0} u_i \tag{85}$$

Now, let us calculate the Lagrange dual function by minimizing the Lagrangian form:

$$\begin{aligned} &\min_{\mathbf{w}=(w_1,w_2,\dots,w_m),b} \|\mathbf{w}\|_2^2 - \mathbf{w}^T\mathbf{c} - \beta b + \alpha \\ &= \min_{w_1\in\mathbb{R}}(w_1^2 - c_1w_1) \\ &+ \min_{w_2\in\mathbb{R}}(w_2^2 - c_2w_3) \\ &+ \dots \\ &+ \min_{w_n\in\mathbb{R}}(w_n^2 - c_nw_n) \\ &+ \min_{b\in\mathbb{R}}(-\beta b) \\ &+ \alpha \end{aligned} \tag{86}$$

where we used the fact that the Lagrangian form is **separable** i.e., it can be written as the sum of $n+1$ terms and the i^{th} term only depends on the i^{th} variable. Notice that

$$\min_{w\in\mathbb{R}}(w^2 - cw) = -\frac{c^2}{4} \tag{87}$$

and

$$\min_{b\in\mathbb{R}}(-\beta b) = \begin{cases} 0 & \beta = 0 \\ -\infty & \beta \neq 0 \end{cases} \tag{88}$$

Hence, we obtain that

$$\Gamma(\mathbf{u}) = \begin{cases} -\frac{\|\sum_{i=1}^{m} u_i \mathbf{x}_i\|_2^2}{4} + \sum_{i|t_i=1} u_i - \sum_{i|t_i=0} u_i & \sum_{i=1}^{m} u_i = 0 \\ -\infty & \sum_{i=1}^{m} u_i \neq 0 \end{cases} \tag{89}$$

Since the maximal solution of the dual never happens at $-\infty$, the dual optimization is given by

$$\begin{gathered} \max_{\mathbf{u}\in\mathbb{R}^m} -\frac{\|\sum_{i=1}^{m} u_i \mathbf{x}_i\|_2^2}{4} + \sum_{i|t_i=1} u_i - \sum_{i|t_i=0} u_i \\ \text{subject to} \\ \sum_{i=1}^{n} u_i = 0 \\ \forall i \in [m] \begin{cases} u_i \geq 0 & t_i = 1 \\ u_i \leq 0 & t_i = 0 \end{cases} \end{gathered} \tag{90}$$

In some cases, strong Lagrangian duality holds i.e., we have that $\Gamma^* = f^*$. Here, is one example:

Theorem 2. Suppose that the functions $f(\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ and $g_i(\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ are convex and the functions $h_j(\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ are affine. Furthermore, suppose that there is a feasible point $\mathbf{x}_1$, such that $g_i(\mathbf{x}_1) < 0$ for all i. Then, strong duality holds.

The second part of the above conditions (the existence of $\mathbf{x}_1$) is called the Slater's condition. It can be replaced by other conditions, which are generally known as **constraint qualifications**.

Example 24. The conditions of strong duality hold for the SVM example if it is feasible. Hence, strong duality holds for the SVM example.

9.1 Karush-Kuhn-Tucker Conditions

There are also different generalizations of the complementary slackness conditions. The most popular one is the so-called Karush-Kuhn-Tucker (KKT) condition. Unlike the complementary slackness conditions, the KKT theorem is not in general "if and only if". The general statement of the theorem is as follows

Theorem 3. Consider the primal optimization in (75). Suppose that a point $\mathbf{x}_0$ is a primal optimal solution and the set of constraints $I = \{i_1, i_2, \ldots, i_l\}$ are active. Suppose that at $\mathbf{x}_0$, g_i are continuous for $i \notin I$ and differentiable for $i \in I$. Further, suppose that $\nabla g_i(\mathbf{x}_0)$ for $i \notin I$ are linearly independent and f is also differentiable. Then, there is a dual feasible point $\mathbf{u}$ and $\mathbf{v}$ such that.

1. For each i, if the i^{th} inequality constraint is inactive, then $u_i = 0$.
2. We have that

$$\nabla f(\mathbf{x}_0) - \sum_{i\in I} u_i \nabla g_i(\mathbf{x}_0) - \sum_{i=1}^{p} v_i \nabla h_i(\mathbf{x}_0) = \mathbf{0} \tag{91}$$

Notice that there might be other points, which satisfy the conditions above, including Items 1 and 2, but are not global optimal points (and not even local ones!). Any point that satisfies all the conditions in Theorem 3 is called **a KKT point**. There is actually a more restricted form of the KKT theorem, which is in the form of "if and only if". It is often called the **sufficient KKT condition**:

Theorem 4. Suppose that the functions $f(\mathbf{x})$ and $g_i(\mathbf{x})$ are convex and $h_j(\mathbf{x})$ are affine. Furthermore, the point $\mathbf{x}_0$ is a KKT point. Then, $\mathbf{x}_0$ is a (global) optimal solution for the primal optimization.