13

Lagrangian Duality

In this chapter we study optimality conditions for convex problems of the form

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & f(\boldsymbol{w}) \\ \text{subject to} \quad & \boldsymbol{f}(\boldsymbol{w}) \leq \mathbf{0} \\ & \boldsymbol{h}(\boldsymbol{w}) = \mathbf{0}, \end{aligned} \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^n$, $\boldsymbol{f} = (f_1, \dots, f_m)^\top$, $\boldsymbol{h} = (h_1, \dots, h_p)^\top$, and the inequalities are componentwise. We assume that f and the f_i are convex, and the h_j are linear. It is also customary to write the conditions $\boldsymbol{h}(\boldsymbol{w}) = \mathbf{0}$ as $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$, with $h_j(\boldsymbol{w}) = \boldsymbol{a}_j^\top \boldsymbol{w} - b_j$, $\boldsymbol{a}_j^\top$ being the j-th row of $\boldsymbol{A}$. The **feasible set** of (1) is the set

$$\mathcal{F} = \{\boldsymbol{w} \mid f_i(\boldsymbol{w}) \leq 0 \ \text{ for } i \in [m], \ \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}\}$$

It is easy to see that $\mathcal{F}$ is convex if the f_i are convex. If the objective f and the f_i are also linear, then (1) is called a **linear programming** problem, and $\mathcal{F}$ is a **polyhedron**. Such problems have been studied extensively and can be solved with efficient algorithms such as the simplex method.

A first-order optimality condition

We first generalize the standard first-order optimality conditions for differentiable functions to the setting of constrained convex optimization.

Theorem 13.1. *Let $f \in C^1(\mathbb{R}^d)$ be a convex, differentiable function, and consider a convex optimization problem of the form (1). Then $\boldsymbol{w}^*$ is a minimizer of the optimization problem*

$$\text{minimize}\, f(\boldsymbol{w}) \quad \textit{subject to} \quad \boldsymbol{w} \in \mathcal{F}$$

if and only if for all $\boldsymbol{w} \in \mathcal{F}$,

$$\nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) \geq 0, \tag{13.1}$$

where $\mathcal{F}$ is the feasible set of the problem.

Proof. Suppose $\boldsymbol{w}^*$ is such that (13.1) holds. Then, since f is a convex function, for all $\boldsymbol{w} \in \mathcal{F}$ we have,

$$f(\boldsymbol{w}) \geq f(\boldsymbol{w}^*) + \nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) \geq f(\boldsymbol{w}^*),$$

which shows that $\boldsymbol{w}^*$ is a minimizer in $\mathcal{F}$. To show the opposite direction, assume that $\boldsymbol{w}^*$ is a minimizer but that (13.1) does not hold. This means that there exists a $\boldsymbol{w} \in \mathcal{F}$ such that $\nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) < 0$. Since both $\boldsymbol{w}^*$ and $\boldsymbol{w}$ are in $\mathcal{F}$ and $\mathcal{F}$ is convex, any point $\boldsymbol{v}(\lambda) = (1-\lambda)\boldsymbol{w}^* + \lambda \boldsymbol{w}$ with $\lambda \in [0,1]$ is also in $\mathcal{F}$. At $\lambda = 0$ we have

$$\frac{df}{d\lambda} f(\boldsymbol{v}(\lambda))|_{\lambda=0} = \nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) < 0.$$

Since the derivative at $\lambda = 0$ is negative, the function $f(\boldsymbol{v}(\lambda))$ is decreasing at $\lambda = 0$, and therefore, for small $\lambda > 0$, $f(\boldsymbol{v}(\lambda)) < f(\boldsymbol{v}(0)) = f(\boldsymbol{w}^*)$, in contradiction to the assumption that $\boldsymbol{w}^*$ is a minimizer. $\square$

Example 13.2. In the absence of constraints, $\mathcal{F} = \mathbb{R}^d$, and the statement says that

$$\forall \boldsymbol{w} \in \mathbb{R}^n \colon \nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) \geq 0.$$

Given $\boldsymbol{w}$ such that $\nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) \geq 0$, then replacing $\boldsymbol{w}$ by $2\boldsymbol{w}^* - \boldsymbol{w}$ we also have the converse inequality, and therefore the optimality condition is equivalent to saying that $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$. We therefore recover the well-known first order optimality condition.

Geometrically, the first order optimality condition means that the set

$$\{\boldsymbol{w} \mid \nabla f(\boldsymbol{w}^*)^\top \boldsymbol{w} = \nabla f(\boldsymbol{w}^*)^\top \boldsymbol{w}^*\}$$

defines a supporting hyperplane to the set $\mathcal{F}$.

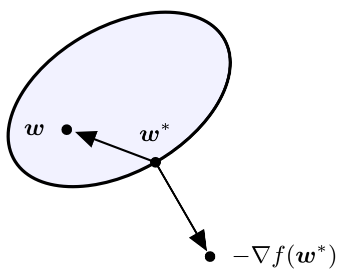

Figure 13.1: Optimality condition

Lagrangian duality

Recall the method of Lagrange multipliers. Given two functions $f(x, y)$ and $h(x, y)$, if the problem

$$\text{minimize}\, f(x, y) \qquad \text{subject to} \quad h(x, y) = 0$$

has a solution (x^*, y^*), then there exists a parameter λ, the *Lagrange multiplier*, such that

$$\nabla f(x^*, y^*) = \lambda \nabla h(x^*, y^*). \tag{13.2}$$

In other words, if we define the *Lagrangian* as

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda h(x, y),$$

then (13.2) says that $\nabla \mathcal{L}(x^*, y^*, \lambda) = 0$ for some λ. The intuition is as follows. The set

$$M = \{(x, y) \in \mathbb{R}^2 \mid h(x, y) = 0\}$$

is a curve in $\mathbb{R}^2$, and the gradient $\nabla h(x, y)$ is perpendicular to M at every point $(x, y) \in M$. For someone living inside M, a vector that is perpendicular to M is not visible, it is zero. Therefore the gradient $\nabla f(x, y)$ is zero as viewed from within M if it is perpendicular to M, or equivalently, a multiple of $\nabla h(x, y)$.

Alternatively, we can view the graph of $f(x, y)$ in three dimensions. A maximum or minimum of $f(x, y)$ along the curve defined by $h(x, y) = 0$ will be a point at which the direction of steepest ascent $\nabla f(x, y)$ is perpendicular to the curve $h(x, y) = 0$.

Example 13.3. Consider the function $f(x, y) = x^2 y$ with the constraint $h(x, y) = x^2 + y^2 - 3$ (a circle of radius $\sqrt{3}$). The Lagrangian is the function

$$\mathcal{L}(x, y, \lambda) = x^2 y - \lambda(x^2 + y^2 - 3).$$

Computing the partial derivatives gives the three equations

$$\begin{aligned} \frac{\partial}{\partial x}\mathcal{L} &= 2xy - 2\lambda x = 0 \\ \frac{\partial}{\partial y}\mathcal{L} &= x^2 - 2\lambda y = 0 \\ \frac{\partial}{\partial \lambda}\mathcal{L} &= x^2 + y^2 - 3 = 0. \end{aligned}$$

From the second equation we get $\lambda = \frac{x^2}{2y}$, and the first and third equations become

$$\begin{aligned} 2xy - \frac{x^3}{y} &= 0 \\ x^2 + y^2 - 3 &= 0. \end{aligned}$$

Solving this system, we get six critical point $(\pm\sqrt{2}, \pm 1)$, $(0, \pm\sqrt{3})$. To find out which one of these is the minimizers, we just evaluate the function f on each of these.

We now turn to convex problems of the more general form form

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & f(\boldsymbol{w}) \\ \text{subject to} \quad & \boldsymbol{f}(\boldsymbol{w}) \leq \mathbf{0} \\ & \boldsymbol{h}(\boldsymbol{w}) = \mathbf{0}, \end{aligned} \tag{13.3}$$

Denote by $\mathcal{D}$ the *domain* of all the functions f, f_i, h_j, i.e.,

$$\mathcal{D} = \operatorname{dom}(f) \cap \operatorname{dom}(f_1) \cap \cdots \cap \operatorname{dom}(f_m) \cap \operatorname{dom}(h_1) \cap \cdots \cap \operatorname{dom}(h_p).$$

Assume that $\mathcal{D}$ is not empty and let p^* be the optimal value of (13.3).

The **Lagrangian** of the system is defined as

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{w}) + \boldsymbol{\lambda}^\top \boldsymbol{f}(\boldsymbol{w}) + \boldsymbol{\mu}^\top h(\boldsymbol{w}) = f(\boldsymbol{w}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{p} \mu_i h_i(\boldsymbol{w}).$$

The vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are called the **dual variables** or **Lagrange multipliers** of the system. The domain of $\mathcal{L}$ is $\mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$.

Definition 13.4. The **Lagrange dual** of the problem (13.3) is the function

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{w} \in \mathcal{D}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

If the Lagrangian $\mathcal{L}$ is unbounded from below, then the value is $-\infty$.

The Lagrangian $\mathcal{L}$ is linear in the λ_i and μ_j variables. The infimum of a family of linear functions is concave, so that the Lagrange dual is a concave function. Therefore the negative $-g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is a convex function.

Lemma 13.5. *Let p^* be the optimal value of the problem 13.3. Then for any $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\lambda} \geq \mathbf{0}$ we have*

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq p^*.$$

Proof. Let $\boldsymbol{w}^*$ be a feasible point for (13.3), that is,

$$f_i(\boldsymbol{w}^*) \leq 0, \quad h_j(\boldsymbol{w}^*) = 0, \ 1 \leq i \leq m, \ 1 \leq j \leq p.$$

Then for $\boldsymbol{\lambda} \geq \mathbf{0}$ and any $\boldsymbol{\mu}$, since each $h_j(\boldsymbol{w}^*) = 0$ and $\lambda_j f_j(\boldsymbol{w}^*) \leq 0$,

$$\mathcal{L}(\boldsymbol{w}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{w}^*) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{w}^*) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{w}^*) \leq f(\boldsymbol{w}^*).$$

In particular,

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \mathcal{L}(\boldsymbol{w}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f(\boldsymbol{w}^*).$$

Since this holds for *all* feasible $\boldsymbol{w}^*$, it holds in particular for the $\boldsymbol{w}^*$ that minimizes (13.3), for which $f(\boldsymbol{w}^*) = p^*$. $\square$

The **Lagrange dual problem** of the optimization problem (13.3) is the problem

$$\operatorname{maximize} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}. \tag{13.4}$$

If q^* is the optimal value of (13.4), then $q^* \leq p^*$. In the special case of linear programming we actually have $q^* = p^*$. We will see that under certain conditions, we have $q^* = p^*$ for more general problems, but this is not always the case.

Notes

Duality is a central theme in optimization theory, and appears in many guises. The general idea is that an optimization problem can appear in two equivalent forms: a primal and a dual problem, where the solution of the dual problem provides a lower bound on the solution of the primal problem. The task is then to determine conditions under which this *duality gap* is zero. One of the features of duality is that the two complexity parameters of an optimization problem, namely the number of constraints and the number of variables, are exchanged under duality. This aspect can be used to simplify certain optimization problems. Another important aspect of duality is that one can often reformulate the requirement that certain primal and dual optimality conditions are satisfied as a system of equations. Duality in the setting of linear programming was first conjectured by John von Neumann, and subsequently developed by Tucker, Danzig, and others [2]. A treatment of convex optimization with a principled development of duality can be found in [1].

[1] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

[2] George Bernard Dantzig. *Linear programming and extensions*, volume 48. Princeton university press, 1998.