Lecture 15

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

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

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

15.1 A first-order optimality condition

So far we have seen two examples of first order optimality conditions: for unconstraint optimization ($\nabla f(\boldsymbol{x}) = 0$) and for linear programming. We now generalize these to the setting of constrained convex optimization.

Theorem 15.1. *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a convex, differentiable function, and*

$$\mathcal{F} = \{\boldsymbol{x} : f_i(\boldsymbol{x}) \leq 0,\ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$$

a feasible set, with f_i convex. Then $\boldsymbol{x}^*$ is an optimal point of the optimization problem

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

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

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle \geq 0. \tag{15.1}$$

Proof. Suppose $\boldsymbol{x}^*$ is such that (1) holds. Then, since f is a convex function, for all $\boldsymbol{y} \in \mathcal{F}$ we have, by Theorem 2.10.1,

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

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

$$\frac{df}{d\lambda} f(\boldsymbol{z}(\lambda))|_{\lambda=0} = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle < 0.$$

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

Example 15.2. In the absence of constraints, $\mathcal{F} = \mathbb{R}^n$, and the statement says that

$$\forall \boldsymbol{y} \in \mathbb{R}^n \colon \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle \geq 0.$$

If there was a $\boldsymbol{y}$ such that $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle > 0$, then replacing $\boldsymbol{y}$ by $2\boldsymbol{x} - \boldsymbol{y}$ we also have the converse inequality, and therefore the optimality condition is equivalent to saying that $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. We therefore recover the well-known first order optimality condition from Lecture 2.

Geometrically, the first order optimality condition means that the set

$$\{\boldsymbol{x} : \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} \rangle = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^* \rangle\}$$

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

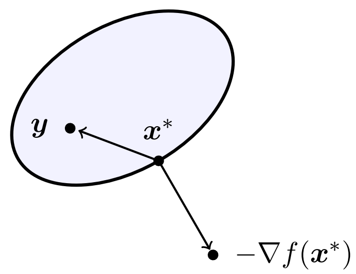

Figure 15.1: Optimality condition

15.2 Lagrangian duality

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

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

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

$$\nabla f(x^*, y^*) = \lambda \nabla h(x^*, y^*). \tag{15.1}$$

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

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

then (15.1) says that $\nabla \mathcal{L}(x^*, y^*, \lambda) = 0$ for some λ. The intuition is as follows. The set

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

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 15.3. 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 $(\pm\sqrt{2}, \pm 1)$, $(0, \pm\sqrt{2})$. 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

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{f}(\boldsymbol{x}) \leq \boldsymbol{0} \\ & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \end{aligned} \tag{15.2}$$

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

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

Assume that $\mathcal{D}$ is not empty and let p^* be the optimal value of (15.2).

The *Lagrangian* of the system is defined as

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

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 15.4. The *Lagrange dual* of the problem (15.2) is the function

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

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 15.5. *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 $\boldsymbol{x}^*$ be a feasible point for (15.2), that is,

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

Then for $\boldsymbol{\lambda} \geq \mathbf{0}$ and μ,

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

In particular,

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

Since this holds for *all* feasible $\boldsymbol{x}^*$, it holds in particular for the $\boldsymbol{x}^*$ that minimizes (15.2), for which $f(\boldsymbol{x}^*) = p^*$. $\square$

A point $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ with $\boldsymbol{\lambda} \geq \mathbf{0}$ and $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathrm{dom}(g)$ is called a *feasible point* of the dual problem.

Example 15.6. Consider a linear programming problem of the form

$$\begin{aligned} \text{minimize} \quad & \langle \boldsymbol{c}, \boldsymbol{x} \rangle \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \mathbf{0}. \end{aligned}$$

The inequality constraints are $-x_i \leq 0$, while the equality constraints are $\boldsymbol{a}_i^\top \boldsymbol{x} - b_i$. The Lagrangian has the form

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \langle \boldsymbol{c}, \boldsymbol{x} \rangle - \sum_{i=1}^n \lambda_i x_i + \sum_{j=1}^m \mu_j (\boldsymbol{a}_j^\top \boldsymbol{x} - b_j) \\ &= (\boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu})^\top \boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{\mu}.\end{aligned}$$

The infimum over $\boldsymbol{x}$ of this function is $-\infty$ unless $\boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0}$. The Lagrange dual is therefore

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} -\boldsymbol{\mu}^\top \boldsymbol{b} & \text{if } \boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0} \\ -\infty & \text{else.} \end{cases}$$

From Lemma 15.5 we conclude that

$$\max\{-\boldsymbol{b}^\top \boldsymbol{\mu} : \boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0},\ \boldsymbol{\lambda} \geq \boldsymbol{0}\} \leq \min\{\boldsymbol{c}^\top \boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0},\ \boldsymbol{x} \geq \boldsymbol{0}\}.$$

Note that if we write $\boldsymbol{y} = -\boldsymbol{\mu}$ and $\boldsymbol{s} = \boldsymbol{\lambda}$, then we get the dual version of the linear programming problem we started out with, and in this case we know that

$$\max_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = p^*.$$

The *Lagrange dual* of the optimization problem (15.2) is the problem

$$\text{maximize} \quad g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{subject to} \quad \boldsymbol{\lambda} \geq \boldsymbol{0}. \tag{15.3}$$

We have seen that if q^* is the optimal value of (15.3), then $q^* \leq p^*$, and the example above implies that 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.