Lagrange duality

In the previous lecture we looked at three examples of optimization problems in which we aimed to minimize a convex function under convex inequality constraints and/or affine equality constraints. In each case we derived a set of necessary and sufficient conditions for having found a minimizer $\boldsymbol{x}^\star$ that involved the introduction of some mysterious additional variables $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$. Here we will provide an alternative perspective on these problems and provide a bit more intuition as to how to interpret these additional variables.

The Lagrangian

In this lecture we will consider an optimization program of the form

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; & f(\boldsymbol{x}) \\ \text{subject to} \;\; & g_m(\boldsymbol{x}) \leq 0, \quad m = 1, \ldots, M_I \\ & h_m(\boldsymbol{x}) = 0, \quad m = 1, \ldots, M_E. \end{aligned} \tag{1}$$

We will focus on the case where the objective function f and the inequality constraints g_m are convex, and the equality constraints h_m are affine (note that for equality constraints, convexity is equivalent to being affine). However, in general much of what we have to say applies to arbitrary (nonconvex) problems as well so we will be clear when we are or are not assuming convexity. We will take the domain of all of the f_m and h_m to be all of $\mathbb{R}^N$ below; this just simplifies the exposition, we can easily replace this with the intersections of the $\operatorname{dom} f_m$ and $\operatorname{dom} h_m$. We will also assume that the feasible set

$$\mathcal{C} = \{\boldsymbol{x} \;:\; f_m(\boldsymbol{x}) \leq 0,\; h_m(\boldsymbol{x}) = 0,\; \text{for all } m\}$$

is non-empty and a subset $\mathbb{R}^N$.

The **Lagrangian** takes the constraints in the program above and integrates them into the objective function. Specifically, the Lagrangian $\mathcal{L} : \mathbb{R}^N \times \mathbb{R}^{M_I} \times \mathbb{R}^{M_E} \to \mathbb{R}$ associated with (1) is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\boldsymbol{x}) + \sum_{m=1}^{M_I} \lambda_m g_m(\boldsymbol{x}) + \sum_{m=1}^{M_E} \nu_m h_m(\boldsymbol{x})$$

For reasons that will become clearer below, the $\boldsymbol{x}$ above are referred to as **primal variables**, and the $\boldsymbol{\lambda}, \boldsymbol{\nu}$ as either **dual variables** or **Lagrange multipliers**.

The Lagrangian allows us to transform the *constrained* optimization problem in (1) into an *unconstrained* one. Specifically, suppose for the moment that we are interested in a problem of the form in (1) but without equality constraints. Consider the problem given by

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m g_m(\boldsymbol{x}). \tag{2}$$

To get some intuition, suppose that we set the $\lambda_1, \ldots, \lambda_M$ to be very large (positive) numbers. In this case, violating any of the constraints (allowing $g_m(\boldsymbol{x}) > 0$) will result in a very large penalty being added to the objective function, so that by setting the corresponding λ_m to be large we will eventually guarantee that the resulting solution will satisfy the desired constraints.

The problem here is that large values of λ_m not only avoid the setting where $g_m(\boldsymbol{x}) > 0$, but actually encourages $g_m(\boldsymbol{x}) \ll 0$ (since we can potentially benefit by not just satisfying the constraints but by exceeding them by a large margin).

This raises a natural question: can we set $\boldsymbol{\lambda}$ so that the solution to the unconstrained problem (2) is the same as the constrained problem (1)? Here we will provide an answer in the case where the objective function f and the constraints $g_1, \ldots, g_M$ are both convex and differentiable.

Suppose that $\boldsymbol{x}^\star$ is a solution to the constrained problem (1). If we want $\boldsymbol{x}^\star$ to be a solution to (2), then a necessary and sufficient condition is

$$\nabla \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) = \nabla f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m \nabla g_m(\boldsymbol{x}^\star) = \mathbf{0}. \tag{3}$$

At this point you might want to compare (3) with conditions 4 in the second two examples from the previous lecture. (Hint: they are the same!)

If we knew $\boldsymbol{x}^\star$ already, finding a $\boldsymbol{\lambda}$ that would make the unconstrained and constrained problems equivalent (meaning that they both have the same solution $\boldsymbol{x}^\star$) would just amount to finding a $\boldsymbol{\lambda}$ such that (3) holds. Unfortunately, this might not seem to be particularly useful since $\boldsymbol{x}^\star$ is what we are trying to find to begin with.

To see how we might compute a $\boldsymbol{\lambda}$ that makes the unconstrained and constrained problems equivalent, we will need to begin our first exploration of one of the deepest and most important ideas of optimization: **duality**.

The Lagrange dual function

We can think of the unconstrained optimization problem (2) as actually representing a family of different optimization problems (depending on $\boldsymbol{\lambda}$). For any fixed $\boldsymbol{\lambda}$, imagine solving (2) and computing the minimal value of the objective function – we can think of this as actually defining a function that maps $\boldsymbol{\lambda} \in \mathbb{R}^M$ to $\mathbb{R}$. Specifically, returning to the case where we have both inequality and equality constraints, the **Lagrange dual function** $d(\boldsymbol{\lambda}, \nu) : \mathbb{R}^{M_I} \times \mathbb{R}^{M_E} \to \mathbb{R}$ is the minimum[1] of the Lagrangian over all $\boldsymbol{x} \in \mathbb{R}^N$:

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^N} \left(f(\boldsymbol{x}) + \sum_{m=1}^{M_I} \lambda_m g_m(\boldsymbol{x}) + \sum_{m=1}^{M_E} \nu_m h_m(\boldsymbol{x}) \right).$$

Note that since the dual is the pointwise infimum of a family of affine functions in $\boldsymbol{\lambda}, \boldsymbol{\nu}$, the Lagrange dual function is **always concave**, regardless of whether or not f, g_m, and h_m are convex. While we will not stress this much here, this is a remarkable fact and can be very useful when dealing with nonconvex problems.

A key fact about the dual function is that it can provide a lower bound on the optimal value of the original program. In the discussion below, we assume throughout that $\boldsymbol{\nu}$ and $\boldsymbol{\lambda} \geq 0$ are arbitrary. Our main claim is that if $p^\star = f(\boldsymbol{x}^\star)$ is the optimal value for (1),[2] then we have

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star.$$

This is very easy to show. Specifically, for any feasible point $\boldsymbol{x}'$, we must have $g_m(\boldsymbol{x}') \leq 0$ for all m and also $h_m(\boldsymbol{x}') = 0$ for all m, and

[1] We are writing inf instead of min here since we in general cannot be sure that the minimum exists. It very well may be that $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is $-\infty$.

[2] We use $p^\star$ instead of $f^\star$ to indicate the optimal value of the *primal* problem, which we will soon be opposing to the optimal value of the *dual* problem.

hence

$$\sum_{m=1}^{M_I} \lambda_m g_m(\boldsymbol{x}') + \sum_{m=1}^{M_E} \nu_m h_m(\boldsymbol{x}') \leq 0.$$

From this we have that

$$\mathcal{L}(\boldsymbol{x}', \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\boldsymbol{x}'),$$

meaning that

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\boldsymbol{x} \in \mathbb{R}^N} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathcal{L}(\boldsymbol{x}', \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\boldsymbol{x}').$$

Since this holds for all feasible $\boldsymbol{x}'$, including the minimizer of (1), we have $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$.

The (Lagrange) dual problem

Given that $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ provides a lower bound on $p^\star$, if you wanted to get an idea of what $p^\star$ looks like (for example, to see if you are close to convergence), it is natural to see how large you can make this lower bound. This gives rise to what we call the **(Lagrange) dual problem** of (1):

$$\underset{\boldsymbol{\lambda} \in \mathbb{R}^{M_I}, \boldsymbol{\nu} \in \mathbb{R}^{M_E}}{\text{maximize}} \quad d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \quad \text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}. \tag{4}$$

The dual optimal value $d^\star$ is

$$d^\star = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} d(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \inf_{\boldsymbol{x} \in \mathbb{R}^N} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

Since $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$, we know that

$$d^\star \leq p^\star.$$

The quantity $p^\star - d^\star$ is called the **duality gap**. If $p^\star = d^\star$, then we say that (1) and (4) exhibit **strong duality**.

We will soon discuss when strong duality holds, but first, why is it important? Suppose that $\boldsymbol{x}^\star$ is a solution to the original constrained problem (1) – which we will call the **primal problem** to distinguish it from the dual problem – and suppose that $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ is a solution to the dual problem (4). It turns out that if we have strong duality, then $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ is exactly what we need to make $\boldsymbol{x}^\star$ the solution to the unconstrained problem (2).

To see why, note that if we have strong duality then

$$\begin{aligned} f(\boldsymbol{x}^\star) &= d(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \\ &= \inf_{\boldsymbol{x} \in \mathbb{R}^N} \left(f(\boldsymbol{x}) + \sum_{m=1}^{M_I} \lambda_m^\star g_m(\boldsymbol{x}) + \sum_{m=1}^{M_E} \nu_m^\star h_m(\boldsymbol{x}) \right) \\ &\leq f(\boldsymbol{x}^\star) + \sum_{m=1}^{M_I} \lambda_m^\star g_m(\boldsymbol{x}^\star) + \sum_{m=1}^{M_E} \nu_m^\star h_m(\boldsymbol{x}^\star) \\ &\leq f(\boldsymbol{x}^\star). \end{aligned} \tag{5}$$

where the last inequality follows from the facts that we must have $\lambda_m^\star \geq 0$ and $g_m(\boldsymbol{x}^\star) \leq 0$ and that $h(\boldsymbol{x}^\star) = 0$. Looking at this entire chain of inequalities, where the first and last term are both $f(\boldsymbol{x}^\star)$, means that

$$f(\boldsymbol{x}^\star) = \min_{\boldsymbol{x} \in \mathbb{R}^N} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star).$$

In words, a solution to the primal problem $\boldsymbol{x}^\star$ is also a minimizer of $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$.

Strong duality and Slater's condition

As we have just seen, when we have strong duality there is a very close connection between the solutions of the primal and dual problems. So when can we expect strong duality to hold? For nonconvex problems, we rarely have strong duality, but for convex problems we usually (but not always) do.

Convexity is not quite enough to ensure strong duality, but there are additional conditions that we can require that will ensure that strong duality holds. Perhaps the most commonly encountered such condition is called **Slater's condition**. Informally, Slater's condition simply says that the feasible set has a non-empty interior. More formally, Slater's condition can be expressed as:

> **Slater's condition**: There exists at least one $\bar{\boldsymbol{x}}$ such that for each inequality constraint g_m, either g_m is affine or
>
> $$g_m(\bar{\boldsymbol{x}}) < 0.$$
>
> That is, there is an $\bar{\boldsymbol{x}}$ that is *strictly* feasible for all non-affine constraints.

Nearly all of the optimization problems that we will encounter in this course will satisfy this condition. There are, however, convex problems that do not. As a simple example, let $\boldsymbol{p}_1 = [1, 0]^{\mathrm{T}}$ and $\boldsymbol{p}_2 = [-1, 0]^{\mathrm{T}}$ and consider the constraints

$$g_1(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{p}_1\|_2^2 - 1 \leq 0$$
$$g_2(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{p}_2\|_2^2 - 1 \leq 0.$$

Note that the only $\boldsymbol{x}$ satisfying both constraints is $\boldsymbol{x} = 0$ and there are no *strictly* feasible points.

Certificates of (sub)optimality

One potential application of the above facts is to serve as a way of measuring how far away we are from finding an optimal solution to our optimization problem. To see this recall that any dual feasible[3] $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ gives us a lower bound on $p^\star$, since $d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$. Thus, if we have a primal feasible $\boldsymbol{x}$, then we know that

$$f(\boldsymbol{x}) - p^\star \;\leq\; f(\boldsymbol{x}) - d(\boldsymbol{\lambda}, \boldsymbol{\nu}).$$

We will refer to $f(\boldsymbol{x}) - d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as the duality gap for the primal/dual (feasible) variables $\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}$. We know that

$$p^\star \in [d(\boldsymbol{\lambda}, \boldsymbol{\nu}), f(\boldsymbol{x})], \quad \text{and likewise} \quad d^\star \in [d(\boldsymbol{\lambda}, \boldsymbol{\nu}), f(\boldsymbol{x})].$$

If we are ever able to reduce this gap to zero, then we know that $\boldsymbol{x}$ is primal optimal, and $\boldsymbol{\lambda}, \boldsymbol{\nu}$ are dual optimal.

There are certain kinds of "primal-dual" algorithms that produce a series of (feasible) points $\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k$ at every iteration. We can then use

$$f(\boldsymbol{x}_k) - d(\boldsymbol{\lambda}_k, \boldsymbol{\nu}_k) \leq \epsilon,$$

as a stopping criteria, and know that our answer would yield an objective value no further than ϵ from optimal.

[3] We simply need $\boldsymbol{\lambda} \geq \mathbf{0}$ for $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ to be dual feasible.

Examples

1. **Inequality LP**. Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \langle \boldsymbol{x}, \boldsymbol{c}\rangle \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.$$

Answer: The Lagrangian is

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= \langle \boldsymbol{x}, \boldsymbol{c}\rangle + \sum_{m=1}^{M} \lambda_m \left(\langle \boldsymbol{x}, \boldsymbol{a}_m\rangle - b_m\right) \\ &= \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}.\end{aligned}$$

This is a linear functional in $\boldsymbol{x}$ — it is unbounded below unless

$$\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0}.$$

Thus

$$\begin{aligned}d(\boldsymbol{\lambda}) &= \inf_{\boldsymbol{x}} \left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right) \\ &= \begin{cases} -\langle \boldsymbol{\lambda}, \boldsymbol{b}\rangle, & \boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0} \\ -\infty, & \text{otherwise.} \end{cases}\end{aligned}$$

So the Lagrange dual program is

$$\begin{aligned}\underset{\boldsymbol{\lambda}\in\mathbb{R}^M}{\text{maximize}} \ -\langle \boldsymbol{\lambda}, \boldsymbol{b}\rangle \quad \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = -\boldsymbol{c} \\ & \boldsymbol{\lambda} \geq \boldsymbol{0}.\end{aligned}$$

2. **Least-squares**. Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

Check that the duality gap is zero.

Answer: The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}.$$

This is quadratic in $\boldsymbol{x}$ and will attain its minimum for

$$\boldsymbol{x} = -\frac{1}{2}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}.$$

Thus

$$\begin{aligned} d(\boldsymbol{\nu}) &= \frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu} - \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} \\ &= -\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu}, \end{aligned}$$

and the Lagrange dual problem is

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^M}{\text{maximize}} -\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu}.$$

Note that this will be maximized when $-\frac{1}{2}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \boldsymbol{b}$, which, when substituted into the dual problem yields

$$-\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} + \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \left\| -\frac{1}{2}\boldsymbol{A}^{T}\boldsymbol{\nu} \right\|_2^2,$$

which shows that strong duality holds.