Math 484: Nonlinear Programming[1] Mikhail Lavrov

Chapter 5, Lecture 7: KKT Duality

March 27, 2019 *University of Illinois at Urbana-Champaign*

1 The dual program

Recall the saddle point version of the KKT theorem: given the optimization problem

$$(P) \qquad \begin{cases} \underset{\mathbf{x}\in S}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{g}(\mathbf{x}) \le \mathbf{0} \end{cases}$$

a pair $\mathbf{x}^* \in S$ and $\boldsymbol{\lambda}^* \ge \mathbf{0}$ are an optimal solution and a sensitivity vector, respectively, when they satisfy the three conditions

1. $L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \le L(\mathbf{x}, \boldsymbol{\lambda}^*)$ for all $\mathbf{x} \in S$
2. $L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \ge L(\mathbf{x}^*, \boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \ge \mathbf{0}$.
3. For $i = 1, 2, \dots, m$, either $\lambda_i^* = 0$ or $g_i(\mathbf{x}^*) = 0$.

The first condition tells us that $\mathbf{x}^*$ is a global minimizer of something; ultimately, that it's the global minimizer of $f(\mathbf{x})$ over the feasible region of P.

The second condition tells us that $\boldsymbol{\lambda}^*$ is a global maximizer of something. The obvious thing that it's the global maximizer of is the function $L(\mathbf{x}^*, \boldsymbol{\lambda})$ of $\boldsymbol{\lambda}$. But this is not a very useful statement before we know $\mathbf{x}^*$.

Instead, define

$$h(\boldsymbol{\lambda}) = \inf\{L(\mathbf{x}, \boldsymbol{\lambda}) : \mathbf{x} \in S\}.$$

The idea is that for each $\boldsymbol{\lambda}$, we separately pick the value of $\mathbf{x}$ that minimizes $L(\mathbf{x}, \boldsymbol{\lambda})$, and set that to be the value of $h(\boldsymbol{\lambda})$. This has the usual caveats: if $L(\mathbf{x}, \boldsymbol{\lambda})$ is not bounded below on S, then $h(\boldsymbol{\lambda}) = -\infty$, and it's possible that we can get arbitrarily close to the lower bound $h(\boldsymbol{\lambda})$ but never achieve it.[2]

How does this $h(\boldsymbol{\lambda})$ relate to values of the Lagrangian in general?

Theorem 1.1 (KKT duality)**.** *Suppose that P has an optimal solution $\mathbf{x}^* \in S$ and a sensitivity vector $\boldsymbol{\lambda}^* \ge \mathbf{0}$. Then for all $\boldsymbol{\lambda} \ge 0$,*

$$h(\boldsymbol{\lambda}) \le h(\boldsymbol{\lambda}^*) = f(\mathbf{x}^*).$$

Proof. The value of $h(\boldsymbol{\lambda}^*)$ is known from condition 1 of the KKT theorem. Since $L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \le L(\mathbf{x}, \boldsymbol{\lambda}^*)$ for all $\mathbf{x} \in S$, we must have $h(\boldsymbol{\lambda}^*) = \inf\{L(\mathbf{x}, \boldsymbol{\lambda}^*) : \mathbf{x} \in S\} = L(\mathbf{x}^*, \boldsymbol{\lambda}^*)$.

[1] This document comes from the Math 484 course webpage: https://faculty.math.illinois.edu/~mlavrov/courses/484-spring-2019.html

[2] There's also the third caveat that $h(\boldsymbol{\lambda}) = +\infty$ if S is the empty set, but optimization problems where S is the empty set aren't very interesting.

Remember that

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

and by complementary slackness, each term of the sum $\sum_{i=1}^m$ simplifies to 0 when we evaluate $L(\mathbf{x}^*, \boldsymbol{\lambda}^*)$. So

$$h(\boldsymbol{\lambda}^*) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = f(\mathbf{x}^*).$$

What about $h(\boldsymbol{\lambda})$ for other values of $\boldsymbol{\lambda}$?

We have $h(\boldsymbol{\lambda}) \le L(\mathbf{x}^*, \boldsymbol{\lambda})$ just by the general idea of inf. That is, $h(\boldsymbol{\lambda})$ is supposed to be the smallest we can make $L(\mathbf{x}, \boldsymbol{\lambda})$ by varying $\mathbf{x}$; one possibility is to plug in $\mathbf{x}^*$, so the value of $h(\boldsymbol{\lambda})$ has to be as small as that or smaller.

Moreover, by condition 2 of the KKT theorem, $L(\mathbf{x}^*, \boldsymbol{\lambda}) \le L(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, which we've already seen is $h(\boldsymbol{\lambda}^*)$. Chaining these two together, we get $h(\boldsymbol{\lambda}) \le h(\boldsymbol{\lambda}^*)$. □

This motivates us to define the dual program D as:

$$(D) \qquad \begin{cases} \underset{\boldsymbol{\lambda} \in \mathbb{R}^m}{\text{maximize}} & h(\boldsymbol{\lambda}) = \inf\{L(\mathbf{x}, \boldsymbol{\lambda}) : \mathbf{x} \in S\} \\ \text{subject to} & \boldsymbol{\lambda} \ge \mathbf{0}. \end{cases}$$

By the theorem we just proved, in the nice case of the KKT theorem (when both $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ exist) the maximum value of D is the same as the minimum value of P, and the optimal dual solution is the same as the sensitivity vector.

1.1 An example

The dual program is a bit weird: it doesn't have any constraints on it, apart from $\boldsymbol{\lambda} \ge \mathbf{0}$, and its objective function is defined in terms of another optimization problem. To help make sense of it, here is a simple example:

$$(P) \qquad \begin{cases} \underset{(x,y) \in \mathbb{R}^2}{\text{minimize}} & x^2 + y^2 \\ \text{subject to} & x + 2y \ge 4. \end{cases}$$

Really, the constraint should be rewritten as $4 - x - 2y \le 0$, at which point P becomes a convex program in standard form.

First, we must find a better formula for the dual objective function h. Here, there is only one constraint on P, so λ is one-dimensional, and the Lagrangian $L(x, y, \lambda)$ is given by

$$\begin{aligned} L(x, y, \lambda) &= x^2 + y^2 + \lambda(4 - x - 2y) \\ &= (x^2 - \lambda x) + (y^2 - 2\lambda y) + 4\lambda. \end{aligned}$$

We define $h(\lambda) = \inf\{L(x, y, \lambda) : x, y, \in \mathbb{R}\}$, so let's minimize $L(x, y, \lambda)$ with respect to x and y.

Conveniently, x and y don't interact, so we can minimize them separately. We minimize $x^2 - \lambda x$ by setting x to be the vertex of the parabola. The vertex of a parabola $ax^2 + bx + c$ is at $x = -\frac{b}{2a}$, so $x = \frac{\lambda}{2}$. We minimize $y^2 - 2\lambda y$ by setting y to be the vertex of *this* parabola: $y = \lambda$. We get

$$h(\lambda) = L(\tfrac{\lambda}{2}, \lambda, \lambda) = \frac{\lambda^2}{4} + \lambda^2 + \lambda\left(4 - \frac{\lambda}{2} - 2\lambda\right) = 4\lambda - \frac{5}{4}\lambda^2.$$

Now that we've done the hard part and figured out what h is, the dual problem is given by

$$(D) \qquad \begin{cases} \underset{\lambda \in \mathbb{R}}{\text{maximize}} & h(\lambda) = 4\lambda - \frac{5}{4}\lambda^2 \\ \text{subject to} & \lambda \geq 0. \end{cases}$$

This is another parabola maximization problem. We set λ equal to the vertex of the parabola:

$$\lambda^* = -\frac{4}{2 \cdot \frac{5}{4}} = \frac{8}{5}.$$

At this point, we have already figured out that the optimal x and y to take are $x = \frac{\lambda}{2}$ and $y = \lambda$, so we get the optimal solution $(x^*, y^*) = (\frac{4}{5}, \frac{8}{5})$.

2 The duality gap

So far, we were in the nice case: when $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ are both guaranteed to exist. What about in the not-so-nice case: when P might not have an optimal solution, or a sensitivity vector?

There is still nothing stopping us from making the definition

$$(D) \qquad \begin{cases} \underset{\boldsymbol{\lambda} \in \mathbb{R}^m}{\text{maximize}} & h(\boldsymbol{\lambda}) = \inf\{L(\mathbf{x}, \boldsymbol{\lambda}) : \mathbf{x} \in S\} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}. \end{cases}$$

It even has some nice properties: for any $\mathbf{x} \in S$ satisfying $g(\mathbf{x}) \leq \mathbf{0}$, and for any $\boldsymbol{\lambda} \geq \mathbf{0}$, we have

$$h(\boldsymbol{\lambda}) \leq L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \leq f(\mathbf{x}).$$

(The first inequality, again, holds because $h(\boldsymbol{\lambda})$ is the inf of $L(\mathbf{x}, \boldsymbol{\lambda})$ over all $\mathbf{x}$; the second inequality holds because all terms in the sum are nonpositive when $g(\mathbf{x}) \leq \mathbf{0}$ and $\boldsymbol{\lambda} \geq \mathbf{0}$.)

The relationship $h(\boldsymbol{\lambda}) \leq f(\mathbf{x})$, whenever $\mathbf{x}$ and $\boldsymbol{\lambda}$ are feasible for their respective programs, is called *weak duality.*

But there's nothing that guarantees that the optimal value of D coincides with the optimal value of P. There might be a *duality gap*: though $h(\boldsymbol{\lambda}) \leq f(\mathbf{x})$ always holds for feasible $\mathbf{x}$ and $\boldsymbol{\lambda}$, there may be a gap between the largest possible value of h and the smallest possible value of f.

Here is an example (from the textbook) of this happening. Here, we have a convex program, but it doesn't satisfy the Slater condition, and it will turn out not to have a sensitivity vector.

$$(P) \qquad \begin{cases} \underset{(x,y) \in \mathbb{R}^2}{\text{minimize}} & f(x, y) = e^{-y} \\ \text{subject to} & \sqrt{x^2 + y^2} - x \leq 0. \end{cases}$$

The key idea here is that $\sqrt{x^2 + y^2}$ is the total distance from (x, y) to the origin, and x is the horizontal distance (provided x is positive). These only match when $x \geq 0$ and $y = 0$; in all other cases, the reverse inequality $\sqrt{x^2 + y^2} - x > 0$ will hold. So any point $(x, 0)$ with $x \geq 0$ is the optimal solution to P, with $f(x, 0) = e^0 = 1$.

Now let's look at the dual program. Here, we pick some $\lambda \ge 0$; again, because there is only one constraint, λ is one-dimensional. Then

$$h(\lambda) = \inf\{L(x,y,\lambda) : x,y \in \mathbb{R}\} = \inf\{e^{-y} + \lambda(\sqrt{x^2+y^2} - x) : x,y \in \mathbb{R}\}.$$

The problem is that we can make y arbitrarily large and still make the penalty $\lambda(\sqrt{x^2+y^2} - x)$ arbitrarily small. This would be obnoxious to specify with concrete numbers. But pick any $\lambda \ge 0$. Pick a y large enough that e^{-y} is as small as you like. No matter what λ and y are, we still have

$$\begin{aligned} \lim_{x\to\infty} \lambda(\sqrt{x^2+y^2} - x) &= \lim_{x\to\infty} \lambda \cdot \frac{(\sqrt{x^2+y^2} - x)(\sqrt{x^2+y^2} + x)}{\sqrt{x^2+y^2} + x} \\ &= \lim_{x\to\infty} \frac{\lambda y^2}{\sqrt{x^2+y^2} + x} = 0. \end{aligned}$$

So we can make the $\lambda(\sqrt{x^2+y^2} - x)$ term as small as we like, too. This means that for any $\lambda \ge 0$, $h(\lambda) = 0$.

So the dual program here is really stupid:

$$(D) \qquad \begin{cases} \underset{\lambda\in\mathbb{R}}{\text{maximize}} & h(\lambda) = 0 \\ \text{subject to} & \lambda \ge 0. \end{cases}$$

The dual objective value of 0 doesn't match the primal objective value of 1.

Despite examples like this one, we often hope with our fingers crossed that our program has a sensitivity vector, even when there is no condition guaranteeing it. In this case, strong duality holds, and a lot of the time, solving the dual program will give us a matching optimal primal solution along the way. (We saw this in the first example.)

If we find a primal feasible $\mathbf{x}$ and a dual feasible $\boldsymbol{\lambda}$ with $f(\mathbf{x}) = h(\boldsymbol{\lambda})$, this guarantees that they are optimal. So if we're lucky, even though we didn't have a condition telling us the things would work out in advance, we know we get the right answer when we're done.

The downside is that we have to solve *two* optimization problems to get there: one is in the definition of $h(\lambda)$, and one is the dual program itself. But we can hope that each one of them will individually be easier than the original problem.

Even when strong duality doesn't apply, solving the dual program gives us a lower bound on the primal objective value, which is partial progress of sorts.