Math 484: Nonlinear Programming[1] Mikhail Lavrov

Chapter 6, Lecture 4: The Penalty Method and KKT Duality

April 10, 2019 *University of Illinois at Urbana-Champaign*

Recal how KKT duality works. Given the optimization problem

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

we define a function

$$h(\boldsymbol{\lambda}) = \inf\{L(\mathbf{x}, \boldsymbol{\lambda}) : \mathbf{x} \in S\} = \inf\left\{ f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) : \mathbf{x} \in S \right\}$$

and solve the dual problem

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

1 Review of the duality gap

Whenever the saddle point form of the KKT theorem applies, we know that P has an optimal solution $\mathbf{x}^*$ and D has an optimal solution $\boldsymbol{\lambda}^*$ with $f(\mathbf{x}^*) = h(\lambda^*)$.

But to guarantee that this happens, we need some strong conditions on P. It is not even enough to assume that P is convex. We need P to be convex, superconsistent (that is, to have an $\mathbf{x}^{(0)} \in S$ with $\mathbf{g}(\mathbf{x}) < \mathbf{0}$) and to actually have an optimal solution $\mathbf{x}^*$.

In general, we can only guarantee that whenever $\mathbf{x}$ is feasible for P and $\boldsymbol{\lambda}$ is feasible for D, we have $f(\mathbf{x}) \geq h(\boldsymbol{\lambda})$. Define

$$MP = \inf\{f(\mathbf{x}) : \mathbf{x} \in S, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\},$$
$$MD = \sup\{h(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \geq \mathbf{0}\}.$$

That is, MP is (essentially) the minimum value of P; it may not be exactly achievable, say if we are minimizing a function like e^x, but we can get arbitrarily close. In exactly the same way, MD is the maximum value of MD. Because $f(\mathbf{x}) \geq h(\boldsymbol{\lambda})$ always holds for feasible $\mathbf{x}$ and $\boldsymbol{\lambda}$, we always have the relationship

$$MP \geq MD.$$

If $MP \neq MD$, we say that there is a duality gap.

Today, we will use the penalty method to get some conditions under which $MP = MD$ and there is no duality gap.

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

This is a bit weaker than having a pair $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfying the KKT conditions. It might be that one or both of MP or MD are "ideal" lower or upper bounds that can be approached arbitrarily close but never reached.

One interesting thing we can do under this hypothesis is solve the dual problem, finding MD, and declare that we now know MP, even if for some reason it is impossible to find a primal optimal solution $\mathbf{x}^*$. We can't do this unless we know that $MP = MD$. So, before, we could only do this when faced with a superconsistent convex program. Now, we will have more freedom.

2 Applying the penalty method

Theorem 2.1 (Theorem 6.3.1 in the textbook)**.** *Suppose that the problem*

$$(P) \qquad \begin{cases} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \end{cases}$$

(note that the domain is all of $\mathbb{R}^n$) is a convex program, that there is a point $\mathbf{x}$ with $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ (but we're not asking for $\mathbf{g}(\mathbf{x}) < \mathbf{0}$), that f and $g_1, g_2, \ldots, g_m$ have continuous gradients, and that f is coercive.

Then $MP = MD$.

Proof. Under these hypotheses, the penalty method is guaranteed to work. When we define

$$F_k(\mathbf{x}) = f(\mathbf{x}) + k\sum_{j=1}^{m}[g_j^+(\mathbf{x})]^2 = f(\mathbf{x}) + k\sum_{j=1}^{m}[\max\{0, g_j(\mathbf{x})\}]^2$$

we know that for all $k > 0$, $F_k(\mathbf{x})$ has an global minimizer $\mathbf{x}^*(k)$, and that there is an unbounded sequence $k_1 < k_2 < k_3$ so that as $k_i \to \infty$, $\mathbf{x}^*(k_i) \to \mathbf{x}^*$ for some $\mathbf{x}^*$.

(That $\mathbf{x}^*$ turns out to be the optimal solution of P.)

We're imagining a case where we've decided that solving P directly is too hard for us, so we're not sure what $\mathbf{x}^*$ is, and we're not sure what the $\mathbf{x}^*(k)$ are, either. But we know they exist.

In particular, if F_k has a global minimizer, it has a critical point. At $\mathbf{x}^*(k)$, we have $\nabla F_k(\mathbf{x}^*(k)) = 0$. That is,

$$\nabla F_k(\mathbf{x}^*(k)) = \nabla f(\mathbf{x}^*(k)) + \sum_{j=1}^{m} 2kg_j^+(\mathbf{x}^*(k))\nabla g_j(\mathbf{x}^*(k)) = \mathbf{0}.$$

Relating this back to the KKT theorem: let

$$\boldsymbol{\lambda}^*(k) = (2kg_1^+(\mathbf{x}^*(k)), 2kg_2^+(\mathbf{x}^*(k)), \ldots, 2kg_m^+(\mathbf{x}^*(k))).$$

This is the vector chosen so that we have

$$\nabla F_k(\mathbf{x}^*(k)) = \nabla f(\mathbf{x}^*(k)) + \sum_{j=1}^{m} \boldsymbol{\lambda}^*(k)\nabla g_j(\mathbf{x}^*(k)) = \nabla_{\mathbf{x}} L(\mathbf{x}^*(k), \boldsymbol{\lambda}^*(k))$$

where L is the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda} \cdot \mathbf{g}(\mathbf{x})$.

Because we assumed that P is a convex program, the Lagrangian is a convex function of $\mathbf{x}$ (when $\boldsymbol{\lambda}$ is fixed). The equation we've just written down tells us that $\mathbf{x}^*(k)$ is a critical point of $L(\mathbf{x}, \boldsymbol{\lambda}^*(k))$, and therefore that it is a global minimizer of $L(\mathbf{x}, \boldsymbol{\lambda}^*(k))$.

In other words, $h(\boldsymbol{\lambda}^*(k)) = L(\mathbf{x}^*(k), \boldsymbol{\lambda}^*(k))$, where h is the dual objective function mentioned earlier.

This is the "KKT side of things". From the "penalty method side of things", we know that we have this unbounded sequence of penalty factors k_1, k_2, k_3 with $\mathbf{x}^*(k_i)$ converging to the optimal solution $\mathbf{x}^*$ as $k_i \to \infty$. For each k_i in the sequence, we have

$$\begin{aligned} f(\mathbf{x}^*(k_i)) &\leq f(\mathbf{x}^*(k_i)) + 2k_i \sum_{j=1}^{m} [g_j^+(\mathbf{x}^*(k_i))]^2 \\ &= f(\mathbf{x}^*(k_i)) + \sum_{j=1}^{m} 2k_i g_i^+(\mathbf{x}^*(k_i)) g_j(\mathbf{x}^*(k_i)) \\ &= f(\mathbf{x}^*(k_i) + \sum_{j=1}^{m} \boldsymbol{\lambda}^*(k_i) g_j(\mathbf{x}^*(k_i)) \\ &= L(\mathbf{x}^*(k_i), \boldsymbol{\lambda}^*(k_i)) = h(\boldsymbol{\lambda}^*(k_i)). \end{aligned}$$

Now we have conflicting information about this value. On the one hand, from the KKT side of things, we have $h(\boldsymbol{\lambda}^*(k_i)) \leq MD$, because MD is an upper bound on h and $h(\boldsymbol{\lambda}^*(k_i))$ is just some value of h. On the other hand, from the penalty method side of things, we know that $f(\mathbf{x}^*(k_i))$ converges to $f(\mathbf{x}^*) = MP$ as $k_i \to \infty$.

So in the limit, we get the inequality $MP \leq MD$. Since we *always* have the reverse inequality $MP \geq MD$, we must have $MP = MD$. □

3 What if f is not coercive?

Finally, we can apply this method even when f is not coercive, by "coercing" f into being coercive first.

Pick some $\epsilon > 0$, and define $f^\epsilon(\mathbf{x}) = f(\mathbf{x}) + \epsilon \|\mathbf{x}\|^2$.

Lemma 3.1. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then f^ϵ remains convex, but is also coercive, for all $\epsilon > 0$.*

Proof. Since f is convex, it has a subgradient $\mathbf{d}$ at $\mathbf{0}$:

$$f(\mathbf{x}) \geq f(\mathbf{0}) + \mathbf{d} \cdot \mathbf{x}.$$

So

$$f^\epsilon(\mathbf{x}) \geq f(\mathbf{0}) + \mathbf{d} \cdot \mathbf{x} + \epsilon \|\mathbf{x}\|^2.$$

The quadratic term $\epsilon \|\mathbf{x}\|^2$ dominates the constant and linear terms, so $f(\mathbf{0}) + \mathbf{d} \cdot \mathbf{x} + \epsilon \|\mathbf{x}\|^2$ is coercive, and therefore $f^\epsilon(\mathbf{x})$ is coercive as well. □

Now, given a convex program

$$(P) \qquad \begin{cases} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \end{cases}$$

which satisfies all the hypotheses of the previous theorem, except that f is not coercive, define

$$(P^\epsilon) \qquad \begin{cases} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} & f(\mathbf{x}) + \epsilon\|\mathbf{x}\|^2 \\ \text{subject to} & \mathbf{g}(\mathbf{x}) \leq \mathbf{0}. \end{cases}$$

We now know that for any $\epsilon > 0$, P^ϵ has no duality gap: $MP^\epsilon = MD^\epsilon$. Solving the dual problem D^ϵ, then, tells us MP^ϵ even if optimal solutions to P^ϵ are hard to find?

Does this tell us anything about P? Yes: it turns out that $MP^\epsilon \to MP$ as $\epsilon \to 0$.

To see this, first imagine that P has an optimal solution $\mathbf{x}^*$. Then on one hand, $f(\mathbf{x}^*) = MP \leq MP^\epsilon$ for all $\epsilon > 0$, since adding the $\epsilon\|\mathbf{x}\|^2$ term can only make the optimal value smaller. On the other hand, $MP^\epsilon \leq f(\mathbf{x}^*) + \epsilon\|\mathbf{x}^*\|^2$, since $\mathbf{x}^*$ is feasible for MP^ϵ (even if it is not optimal for it).

This means that MP^ϵ is sandwiched between $f(\mathbf{x}^*)$ and $f(\mathbf{x}^*)+\epsilon\|\mathbf{x}^*\|^2$, and by the squeeze theorem $MP^\epsilon \to f(\mathbf{x}^*)$ as $\epsilon \to 0$.

This was a small lie; there is not necessarily an optimal solution $\mathbf{x}^*$ that achieves the value MP. But we can repeat the same argument for a point $\mathbf{x}$ with $f(\mathbf{x})$ arbitrarily close to MP, getting an upper bound on MP^ϵ that's arbitrarily close to MP as well. Either way, $MP^\epsilon \to MP$ as $\epsilon \to 0$.

In summary, if we can solve the dual problem D^ϵ for any $\epsilon > 0$, we can take the limit as $\epsilon \to 0$ of MD^ϵ, and find MP (but not, necessarily, the optimal solution that goes with it, if there even is one).