Math 484: Nonlinear Programming[1] Mikhail Lavrov

Chapter 5, Lecture 6: KKT Theorem, Gradient Form

March 25, 2019 *University of Illinois at Urbana-Champaign*

1 The Karush–Kuhn–Tucker theorem, gradient form

Last time, we proved:

Theorem 1.1 (Karush–Kuhn–Tucker theorem, saddle point form)**.** *Let P be any nonlinear program. Suppose that $\mathbf{x}^* \in S$ and $\boldsymbol{\lambda}^* \geq 0$. Then $\mathbf{x}^*$ is an optimal solution of P and $\boldsymbol{\lambda}^*$ is a sensitivity vector for P if and only if:*

1. *$L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \leq L(\mathbf{x}, \boldsymbol{\lambda}^*)$ for all $\mathbf{x} \in S$. (Minimality of $\mathbf{x}^*$)*
2. *$L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}^*, \boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \geq \mathbf{0}$. (Maximality of $\boldsymbol{\lambda}^*$)*
3. *$\lambda_i^* g_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$. (Complementary slackness)*

A much more practical form of the theorem, however, is the following:

Theorem 1.2 (Karush–Kuhn–Tucker theorem, gradient form)**.** *Let P be any nonlinear program where f and $g_1, \dots, g_m$ have continuous first partial derivatives. Suppose that $\mathbf{x}^* \in \operatorname{int}(S)$ is an optimal solution of P, and $\boldsymbol{\lambda}^* \geq \mathbf{0}$ is a sensitivity vector. Then*

1. *$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}$. That is, $\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}$.*
2. *For each i, $g_i(\mathbf{x}^*) \leq 0$.*
3. *For each i, either $\lambda_i^* = 0$ or $g_i(\mathbf{x}^*) = 0$.*

Furthermore, if P is a convex program, then the converse holds: if $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfy these conditions, then $\mathbf{x}^*$ is an optimal solution, and $\boldsymbol{\lambda}^*$ is a sensitivity vector.

Proof. Here, we just transform minimality of $\mathbf{x}^*$ and maximality of $\boldsymbol{\lambda}^*$ into nicer forms to deal with.

Awkwardly, if $\mathbf{x}^*$ is a boundary point of S, then we can't say much; maybe there are better points close to $\mathbf{x}^*$, but they're all outside S. (Maybe S consists only of the point $\mathbf{x}^*$, in which case nothing is guaranteed to be true.)

But if $\mathbf{x}^*$ is an interior point of S, and a global minimizer of $L(\mathbf{x}, \boldsymbol{\lambda}^*)$ (as a function of $\mathbf{x}$), then it's a critical point. Taking the gradient of $L(\mathbf{x}, \boldsymbol{\lambda}^*)$, we get

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) = \mathbf{0}.$$

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

As for maximality of $\boldsymbol{\lambda}^*$: since $\mathbf{x}^*$ satisfies $\mathbf{g}(\mathbf{x}^*) \le 0$, $L(\mathbf{x}^*, \boldsymbol{\lambda})$ is a linear function of $\boldsymbol{\lambda}$ where all coefficients $g_i(\mathbf{x}^*)$ are at most 0. So we want to set every component of $\boldsymbol{\lambda}^*$ to 0, except possibly the ones where $g_i(\mathbf{x}^*) = 0$, since those don't affect the value of L anyway.

Finally, if P is a convex program, then $L(\mathbf{x}, \boldsymbol{\lambda}^*)$ is a convex function of $\mathbf{x}$, so any critical point is a global minimizer. This lets us go the other way, and conclude the minimality of $\mathbf{x}^*$ from the conditions in this theorem. The maximality of $\boldsymbol{\lambda}^*$ always follows from conditions 2 and 3, because we understand linear functions.

So in the convex case, we can recover all three parts of the saddle point form of the KKT theorem, and then we use it to conclude that $\mathbf{x}^*$ is an optimal solution and $\boldsymbol{\lambda}^*$ is a sensitivity vector. □

2 Example problem

Consider the convex program

$$(P) \qquad \begin{cases} \underset{\mathbf{x} \in S}{\text{minimize}} & \dfrac{1}{x+y} \\ \text{subject to} & 2x + y^2 - 6 \le 0, \\ & 1 - x \le 0, \\ & 1 - y \le 0 \end{cases}$$

where $S = \{(x, y) : x, y > 0\}$. This choice of S guarantees that the objective function is convex on S, but also doesn't constrain us in any way: a point that satisfies $1 - x \le 0$ and $1 - y \le 0$ is guaranteed to be in the interior of S. The point $(1.5, 1.5)$ (for example) demonstrates that (P) is superconsistent, and therefore the KKT theorem is guaranteed to solve the problem for us.

Setting the gradient of the Lagrangian to $\mathbf{0}$ gives us the equation

$$\begin{bmatrix} -\frac{1}{(x+y)^2} \\ -\frac{1}{(x+y)^2} \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 2y \end{bmatrix} + \lambda_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

If you like terminology, this is sometimes called the stability condition.

We also have the complementary slackness equations

$$\lambda_1(2x + y^2 - 6) = \lambda_2(1 - x) = \lambda_3(1 - y) = 0.$$

In theory, at this point, we could check $2 \times 2 \times 2 = 8$ cases: either $\lambda_1 = 0$ or $2x + y^2 - 6 = 0$, either $\lambda_2 = 0$ or $1 - x = 0$, and either $\lambda_3 = 0$ or $1 - y = 0$. This is a fall-back option if you can't do anything clever, but fails to take advantage of the full power of the KKT theorem. (It is essentially applying the method of Lagrange multipliers, checking which inequalities are actually equations.)

In practice, we can often eliminate many cases at once. Sometimes it takes a clever strategy for which case to look at first, but the worst thing that happens if you don't guess the right case to start with is you end up doing a bit more work.

Here, let's first consider the case $\lambda_1 = 0$. Then the stability condition simplifies to

$$\begin{bmatrix} -\frac{1}{(x+y)^2} \\ -\frac{1}{(x+y)^2} \end{bmatrix} + \lambda_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which can only be satisfied by setting $\lambda_2 = \lambda_3 = -\frac{1}{(x+y)^2}$. But this guarantees that $\lambda_2, \lambda_3 < 0$, which is not allowed. So we can't have $\lambda_1 = 0$, and just like that we have rejected four of the eight cases.

Since $\lambda_1 \neq 0$, we must have $2x + y^2 - 6 = 0$. Now, almost all of the remaining cases are ones where we set either $x = 1$ or $y = 1$; in those cases, we can solve for both variables, so they're easy to check.

- Suppose we have $x = 1$. Then $y^2 - 4 = 0$, so $y = \pm 2$, and $y = 2$ is the only one that satisfies $1 - y \leq 0$. This gives us the point $(x, y) = (1, 2)$.

  Is this a valid solution? Going back to the stability condition, we have

  $$\begin{bmatrix} -\frac{1}{3^2} \\ -\frac{1}{3^2} \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 4 \end{bmatrix} + \lambda_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  Since $y \neq 1$, we must have $\lambda_3 = 0$. So we are left with two linear equations for λ_1 and λ_2:

  $$\begin{cases} -\frac{1}{9} + 2\lambda_1 - \lambda_2 = 0, \\ -\frac{1}{9} + 4\lambda_1 = 0. \end{cases}$$

  This gives us $\lambda_1 = \frac{1}{36}$ and $\lambda_2 = -\frac{1}{18}$. We reject this case, because $\lambda_2 < 0$, which isn't allowed.

- Suppose we have $y = 1$. Then $2x + 1^2 - 6 = 0$, so $x = 2.5$. This gives us the point $(x, y) = (2.5, 1)$.

  Is this a valid solution? Going back to the stability condition, we have

  $$\begin{bmatrix} -\frac{1}{3.5^2} \\ -\frac{1}{3.5^2} \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \lambda_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  Since $x \neq 1$, we must have $\lambda_2 = 0$. So we are left with two linear equations for λ_1 and λ_3:

  $$\begin{cases} -\frac{4}{49} + 2\lambda_1 = 0, \\ -\frac{4}{49} + 2\lambda_1 - \lambda_3 = 0. \end{cases}$$

  This gives us $\lambda_1 = \frac{2}{49}$ and $\lambda_3 = 0$, satisfying all the conditions.

At this point, we can stop, because we've already found a valid solution: $(x, y) = (2.5, 1)$ and $\boldsymbol{\lambda} = (\frac{2}{49}, 0, 0)$. Since P is a convex program, there is only one solution to the gradient KKT conditions, and it is the correct one.

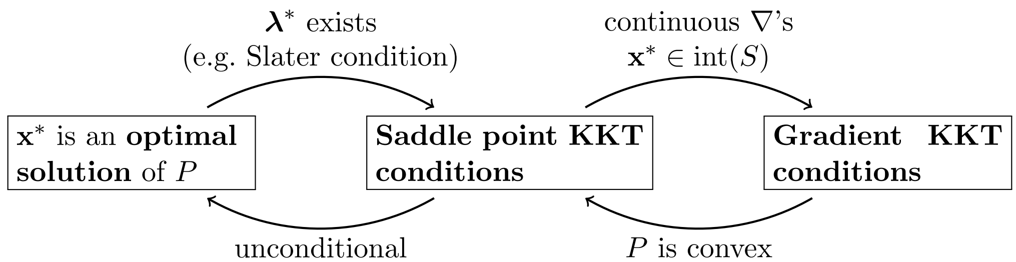
If we didn't stop, then the remaining case to consider would be the one where neither $x = 1$ nor $y = 1$, where we are forced to take $\lambda_2 = \lambda_3 = 0$ by complementary slackness. It turns out that having $\lambda_2 = \lambda_3 = 0$ is impossible when $x \neq 1$ and $y \neq 1$.

3 A summary of what we know

Ultimately, we want to find the optimal solution $\mathbf{x}^*$ of P. We also have this object we're calling a sensitivity vector $\boldsymbol{\lambda}^*$ which we may or may not care about.

We have two sets of conditions that $\mathbf{x}^* \in S$ and $\boldsymbol{\lambda}^* \geq \mathbf{0}$ can satisfy: the saddle point KKT conditions (that is, conditions 1–3 in the saddle point KKT theorem) and the gradient KKT conditions (that is, conditions 1–3 in the gradient KKT theorem).

The chart below summarizes the relationship between them. Each arrow is labeled with the additional conditions we need to go from one statement to another.



In more detail:

- If $\mathbf{x}^*$ is an optimal solution of P, then to conclude that $\mathbf{x}^*$ satisfies the saddle point KKT conditions (together with some $\boldsymbol{\lambda}^* \geq \mathbf{0}$) we need to know that a sensitivity vector exists.

  One condition that guarantees this is the Slater condition. The Slater condition holds if P is convex and superconsistent: that is, there is some feasible solution $\mathbf{x}$ for which the strict inequality $\mathbf{g}(\mathbf{x}) < \mathbf{0}$ holds.

- If $\mathbf{x}^* \in S$ and $\boldsymbol{\lambda}^* \geq \mathbf{0}$ satisfy the saddle point KKT conditions, then we don't need any further hypotheses to conclude that $\mathbf{x}^*$ is an optimal solution.

- If $\mathbf{x}^* \in S$ and $\boldsymbol{\lambda}^* \geq \mathbf{0}$ satisfy the saddle point KKT conditions, then to get from there to the gradient KKT conditions, we need two things.

  We need f and $g_1, g_2, \ldots, g_m$ to have continuous partial derivatives, so that a global minimizer of $L(\mathbf{x}, \boldsymbol{\lambda}^*)$ is a critical point.

  We also need $\mathbf{x}^* \in \text{int}(S)$, because a global minimizer on the boundary of S might still not be a critical point.

- If $\mathbf{x}^* \in S$ and $\boldsymbol{\lambda}^* \geq \mathbf{0}$ satisfy the gradient KKT conditions, and P is convex, then they also satisfy the saddle point KKT conditions.