Lecture 7
Lagrange multipliers and KKT conditions

Instructor: Prof. Gabriele Farina (✉ gfarina@mit.edu)*

With separation in our toolbox, in Lecture 5 we have been able to prove the characterization of the normal cone to the intersection of linear constraints. As a reminder, this was the result that we were able to prove.[1]

Theorem L7.1. Let $\Omega \subseteq \mathbb{R}^n$ be defined as the intersection of m linear inequalities

$$\Omega := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^\top x = b_i & \forall i = 1, ..., r \\ c_j^\top x \le d_j & \forall j = 1, ..., s \end{array} \right\}$$

Given a point $x \in \Omega$, define the index set of the "active" inequality constraints

$$I(x) := \left\{ j \in \{1, ..., s\} : c_j^\top x = d_j \right\}.$$

Then, the normal cone at any $x \in \Omega$ is given by

$$\begin{aligned} \mathcal{N}_\Omega(x) &= \left\{ \sum_{i=1}^r \mu_i a_i + \sum_{j \in I(x)} \lambda_j c_j : \mu_i \in \mathbb{R}, \lambda_j \in \mathbb{R}_{\ge 0} \right\} \\ &= \left\{ \sum_{i=1}^r \mu_i a_i + \sum_{j=1}^s \lambda_j c_j : \mu_i \in \mathbb{R}, \lambda_j \in \mathbb{R}_{\ge 0},\ \lambda_j (d_j - c_j^\top x) = 0 \quad \forall j = 1, ..., s \right\}, \end{aligned}$$

where the second equality simply rewrites the condition $j \in I(x)$ via *complementary slackness* (see Lecture 3).

L7.1 Karush-Kuhn-Tucker (KKT) conditions

The result of Theorem L7.1 gives a complete characterization of the normal cone for sets defined as intersections of linear constraints. We now turn our attention to more general constraint sets, defined as the intersection of *differentiable* functional constraints

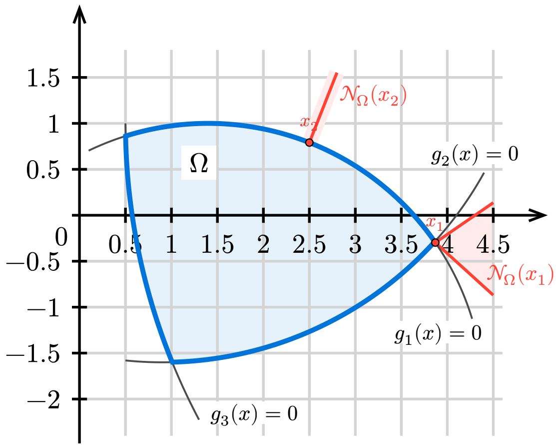
*These notes are class material that has not undergone formal peer review. The TAs and I are grateful for any reports of typos.

[1]For reasons of convenience, Theorem L7.1 makes a distinction between equality and inequality constraints. Of course, equality constraints can be rewritten as two inequality constraints (see also section L3.2.1 "A remark on equality constraints" from Lecture 3), but the notation will be more convenient this way for what is about to come in the rest of the lecture.

$$
\begin{aligned}
\min_{x} \quad & f(x) \\
\text{s.t.} \quad & h_i(x) = 0 \quad i \in \{1, ..., r\} \\
& g_j(x) \le 0 \quad j \in \{1, ..., s\}.
\end{aligned}
\tag{1}
$$

L7.1.1 The general idea

Consider any point $x^\star$ on the boundary of the feasible set Ω depicted on the side, which is the intersection of three inequality constraints $g_i(x) \le 0$ for $i \in \{1, 2, 3\}$ (in particular, in the figure the $g_i(x)$ are quadratic constraints).



> *The set of directions that form an obtuse angle with all directions that from $x^\star$ remain inside of the set coincides with the normal cone at the* linearization *of the constraints that are binding (holding with equality) at $x^\star$.*

The above observation suggests that for a nonlinear optimization problem with functional constraints, $-\nabla f(x)$ should belong to the normal cone to the linearization of the binding constraints at x. This condition goes under the name of *Karush-Kuhn-Tucker (KKT) optimality condition.*[2]

Since the binding linearized constraints are of the form

$$
\begin{aligned}
h_i(x^\star) + \langle \nabla h_i(x^\star), x - x^\star \rangle = 0 \quad &\Longrightarrow \quad \langle \nabla h_i(x^\star), x \rangle = \langle \nabla h_i(x^\star), x^\star \rangle - h_i(x^\star) \\
g_i(x^\star) + \langle \nabla g_i(x^\star), x - x^\star \rangle \le 0 \quad &\Longrightarrow \quad \langle \nabla g_i(x^\star), x \rangle \le \langle \nabla g_i(x^\star), x^\star \rangle - g_i(x^\star),
\end{aligned}
$$

from Theorem L7.1 we know that the normal cone of the linearization $\tilde{\Omega}$ of Ω around $x^\star$ is

$$
\mathcal{N}_{\tilde{\Omega}}(x^\star) = \left\{ \sum_{i=1}^{r} \mu_i \nabla h_i(x^\star) + \sum_{j \in I(x^\star)} \lambda_j \nabla g_j(x^\star) : \mu_i \in \mathbb{R}, \lambda_j \in \mathbb{R}_{\ge 0} \right\},
$$

where the set of binding inequality constraints $I(x^\star)$ is

$$
I(x^\star) := \{ j \in \{1, ..., s\} : g_j(x^\star) = 0 \}.
$$

(All equality constraints are always binding, and so we can directly sum over all $i = 1, ..., r$.)

By using the complementary slackness reformulation of $I(x)$, the first-order optimality conditions induced by the linearization of the feasible set are typically rewritten as follows.

[2]The KKT conditions used to be called "Kuhn-Tucker conditions" and were first published by Kuhn, H. W., & Tucker, A. W. [KT51]. It was later discovered that the same conditions had appeared more than 10 years earlier in the unpublished master's thesis of Karush, W. [Kar39]. An analysis of the history of the KKT conditions is given by Kjeldsen, T. H. [Kje00].

Definition L7.1 (KKT conditions). Consider a nonlinear optimization problem with differentiable objective function and functional constraints, in the form given in (1), and let x be a point in the feasible set ("Primal Feasibility"). The *KKT conditions* at x are given by

$$-\nabla f(x) = \sum_{i=1}^{r} \mu_i \nabla h_i(x) + \sum_{j=1}^{s} \lambda_j \nabla g_j(x) \qquad \text{("Stationarity")}$$

$$\mu_i \in \mathbb{R}, \quad \lambda_j \geq 0 \qquad \forall i = 1, ..., r, \quad j = 1, ..., s \qquad \text{("Dual feasibility")}$$

$$\lambda_j \cdot g_j(x) = 0 \qquad \forall j = 1, ..., s. \qquad \text{("Complementary slackness")}$$

In the definition above, we have noted in quotes the typical names for each of the different conditions. However, please do not get distracted by these names:

- What the KKT conditions are really saying is that $-\nabla f(x)$ must be in the normal cone to the linearization of the constraint set.
- The complementary slackness condition is just a fancier way of writing "if $j \notin I(x)$, then $\lambda_j = 0$".

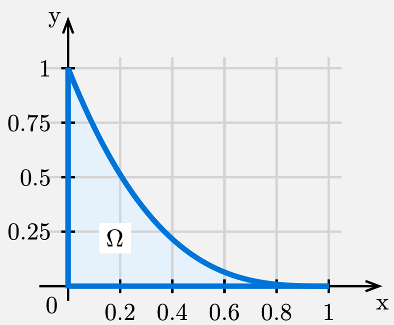
The KKT conditions are *often* necessary conditions for optimality (for example, in the picture above), but *not always*.

L7.1.2 Failure of the KKT conditions

It is important to keep in mind that in some cases *KKT conditions might fail* to hold at optimality. This typically happens when the linearization of the constraints *collapses*. To fix ideas, consider the following simple example.

Example L7.1 (Failure of KKT). Consider the problem

$$\begin{aligned} \min_{(x,y)} \quad & -x \\ \text{s.t.} \quad & y - (1-x)^3 \leq 0 \\ & x \geq 0 \\ & y \geq 0. \end{aligned}$$



Let's denote the objective and functional constraints as

- $f(x, y) := -x$,
- $g_1(x, y) := y - (1-x)^3 \leq 0$,
- $g_2(x, y) := -x \leq 0$,
- $g_3(x, y) := -y \leq 0$.

The feasible set Ω for this problem is shown on the right. At the optimal point $(x^\star, y^\star) := (1, 0)$, the gradients of the objective and the binding constraints (g_1 and g_3) are

$$\nabla f(x^\star, y^\star) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}; \qquad \nabla g_1(x^\star, y^\star) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \qquad \nabla g_3(x^\star, y^\star) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

It is then clear that there exist no Lagrange multipliers λ_1, λ_3 such that

$$-\nabla f(x^\star, y^\star) = \lambda_1 \nabla g_1(x^\star, y^\star) + \lambda_3 \nabla g_3(x^\star, y^\star).$$

The KKT conditions fail in this case.

The reason why the KKT conditions failed at the optimal point $(x^\star, y^\star) = (1, 0)$ in Example L7.1 is due to the fact that the linearization of constraint $g_1(x, y) \le 0$ around the optimal point $(x^\star, y^\star) = (1, 0)$ is $y \le 0$. This is parallel to the existing constraint $y \ge 0$, and fails to capture the fact that $x \le 1$ on the feasible set Ω.

L7.1.3 Constraint qualification

Conditions that prevent the degenerate behavior illustrated in Example L7.1 above go under the name of *constraint qualification*. Several constraint qualification conditions are known in the literature.

■ Concave and affine constraints. We already know that when the feasible set Ω is defined via linear constraints (that is, all h_i and g_j in (1) are affine functions), then no further constraint qualifications hold, and the necessity of the KKT conditions is implied directly by Theorem L7.1.

With only very little work, we can show that the same remains true if the g_j are allowed to be *concave* functions (that is, the $-g_j$ are convex functions).

Theorem L7.2 (Concave and linear constraints)**.** Let $x \in \Omega \subseteq \mathbb{R}^n$ be a minimizer of (1). If
- the binding inequality constraints $\{g_j\}_{j \in I(x)}$ are *concave* differentiable functions in a convex neighborhood of x; and
- the equality constraints $\{h_i\}_{i=1}^r$ are affine functions on $\mathbb{R}^n$,

then the KKT conditions hold at x.

(In fact, this might very well be in Pset 3…)

■ Linear independence of gradients. In Example L7.1, the linearization of two constraints coincided, causing problems. When all linearized constraints are linearly independent, the issue is avoided.

Theorem L7.3 (Linear independence of gradients)**.** Let $x \in \Omega \subseteq \mathbb{R}^n$ be a minimizer of (1). If all functions h_i, g_j are continuously differentiable and the multiset of gradients at x of all active constraints

$$\{\nabla h_i(x) : i = 1, ..., r\} \cup \left\{\nabla g_j(x) : j \in I(x)\right\}$$

is linearly independent,[3] then the KKT conditions hold at x.

[3] Remark: the union above is meant with repetition. Indeed, consider what would happen if in Example L7.1 we had made the third constraint $y = 0$ instead of $y \ge 0$.

In fact, the condition above is a special case of a much more general condition called *Mangasarian-Fromowitz constraint qualification* [MF67].

Example L7.2. One might wonder why the condition requires *continuous* differentiability rather than just differentiability (after all the KKT conditions only use the gradients).

To appreciate the need, consider the following example, due to Fernández, L. A. [Fer97]:

$$\begin{aligned} \min_{x,y} \quad & x \\ \text{s.t.} \quad & h(x,y) = 0 \\ & x, y \in \mathbb{R}, \end{aligned} \qquad \text{where} \quad h(x,y) := \begin{cases} y & \text{if } x \geq 0 \\ y - x^2 & \text{if } x < 0 \text{ and } y \leq 0 \\ y + x^2 & \text{if } x < 0 \text{ and } y > 0. \end{cases}$$

It is clear that for $x < 0$, $h(x,y) \neq 0$, and for $x \geq 0$, $h(x,y) = 0$ is equivalent to $y = 0$. Hence, the constraint $h(x,y) = 0$ is equivalent to the constraint $y = 0, x \geq 0$. It is evident that the unique minimizer is the point $(0,0)$.

As it turns out, the function h is differentiable at $(0,0)$, with

$$\nabla h(0,0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

However, the function is *not* continuously differentiable (it is not differentiable on the ray $(x,0)$ for $x < 0$). The KKT conditions in this case would require that there exists $\mu \in \mathbb{R}$ such that

$$-\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mu \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

which is clearly impossible. Hence, the minimizer does *not* satisfy the KKT conditions.

■ **Slater's condition.** Finally, we consider a popular constraint qualification condition for problems with *convex* inequality constraints and affine equality constraints.

Theorem L7.4 (Slater's condition [Sla59])**.** Let $x \in \Omega \subseteq \mathbb{R}^n$ be a minimizer of (1). If

- the binding inequality constraints $\{g_j\}_{j \in I(x)}$ are *convex* differentiable functions; and
- the equality constraints $\{h_i\}_{i=1}^r$ are affine functions; and
- there exists a feasible point x_0 that is *strictly* feasible for the binding inequality constraints, that is,

$$g_j(x_0) < 0 \qquad \forall j \in I(x)$$

then the KKT conditions hold at x.

The *strict* feasibility requirement is key and cannot be relaxed, even if the g_j are convex and thus rather well-behaved. For example, consider what would happen in the problem $\min x$ subject to $x^2 \leq 0$...

Similarly, the convexity requirement cannot be relaxed even under strict feasibility, as Example L7.1 shows.

A nice feature of Slater's condition is that it is *sufficient* for optimality, and not just *necessary*, when the objective function f is convex.

Theorem L7.5. If f is convex and the constraints satisfy Slater's condition, then the KKT conditions are *both* necessary and sufficient for optimality.

Proof. Necessity is guaranteed from Theorem L7.4. So, we focus on sufficiency. Let the problem satisfying Slater's conditions be

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0 \quad i = 1, ..., r \\ & Ax = b, \end{aligned}$$

and $x^\star$ be a feasible point satisfying the KKT conditions. In particular, we know that there exist multipliers $\lambda_i^\star \ge 0$ and $\mu^\star$ such that

$$\nabla f(x^\star) + \sum_{i=1}^{r} \lambda_i^\star \nabla g_i(x^\star) + A^\top \mu^\star = 0, \tag{2}$$

and that the complementary slackness condition

$$\lambda_i^\star g_i(x^\star) = 0 \qquad \forall i = 1, ..., r.$$

holds. Since f and all the g_i's are convex, Equation (2) implies that $x^\star$ is a minimizer of the function

$$L(x) := f(x) + \sum_{i=1}^{r} \lambda_i^\star g_i(x) + \langle \mu^\star, Ax - b \rangle,$$

as first-order optimality conditions are sufficient for convex functions. Furthermore, note that $L(x) \le f(x)$ for all feasible x, since $\lambda_i^\star \ge 0$ and all feasible x satisfy $g_i(x) \le 0$ and $Ax = b$. Hence, for any feasible x we can write

$$f(x) \ge L(x) \ge L(x^\star) = f(x^\star),$$

where the last equality follows from the complementary slackness conditions. □

L7.2 Further readings

The following books all contain an excellent and approachable treatment of the KKT condititions:

[Gül10] Güler, O. (2010). *Foundations of Optimization*. Springer. `https://link.springer.com/book/10.1007/978-0-387-68407-9`

[Jah07] Jahn, J. (2007). *Introduction to the Theory of Nonlinear Optimization*. Springer. `https://link.springer.com/book/10.1007/978-3-540-49379-2`

[Ber16] Bertsekas, D. P. (2016). *Nonlinear Programming* (3rd edition). Athena Scientific.

For a more advanced treatment with connections to metric regularity and nonsmooth analysis, I recommend the following book by Borwein and Lewis.

[BL06] Borwein, J., & Lewis, A. (2006). *Convex Analysis and Nonlinear Optimization.* Springer. `https://link.springer.com/book/10.1007/978-0-387-31256-9`

Bibliography for this lecture

[KT51] Kuhn, H. W., & Tucker, A. W. (1951). Nonlinear Programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 2, pp. 481–493). University of California Press.

[Kar39] Karush, W. (1939). *Minima of functions of several variables with inequalities as side conditions.*

[Kje00] Kjeldsen, T. H. (2000). A Contextualized Historical Analysis of the Kuhn–Tucker Theorem in Nonlinear Programming: The Impact of World War II. *Historia Math.*, *27*(4), 331–361. `https://doi.org/10.1006/hmat.2000.2289`

[MF67] Mangasarian, O. L., & Fromovitz, S. (1967). The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *J. Math. Anal. Appl.*, *17*(1), 37–47. `https://doi.org/10.1016/0022-247X(67)90163-1`

[Fer97] Fernández, L. A. (1997). Classroom note: On the limits of the Lagrange multiplier rule. *SIAM Review*, *39*(2), 292–297.

[Gül10] Güler, O. (2010). *Foundations of Optimization.* Springer. `https://link.springer.com/book/10.1007/978-0-387-68407-9`

[Jah07] Jahn, J. (2007). *Introduction to the Theory of Nonlinear Optimization.* Springer. `https://link.springer.com/book/10.1007/978-3-540-49379-2`

[Ber16] Bertsekas, D. P. (2016). *Nonlinear Programming* (3rd edition). Athena Scientific.

[BL06] Borwein, J., & Lewis, A. (2006). *Convex Analysis and Nonlinear Optimization.* Springer. `https://link.springer.com/book/10.1007/978-0-387-31256-9`

Changelog

- Feb 27, 2025: Added Theorem L7.5.
- Feb 28, 2025: Fixed two typos (thanks Nicolas Gorlo!)
- Mar 1, 2025: Added footnote 3 (https://piazza.com/class/m6lg9aspoutda/post/56)
- Mar 2, 2025: Added equation number to (2)
- Mar 15, 2025: Fixed indexing in Theorem L7.1 (thanks Khizer Shahid!)