MIT 6.7220/15.084 — Nonlinear Optimization Tue, Feb 27th 2024

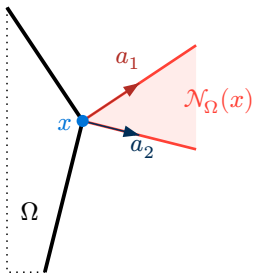
Lecture 5
Lagrange multipliers and KKT conditions

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

With separation in our toolbox, in this lecture we revisit normal cones, and extend our machinery to feasible sets defined by functional constraints.

1 A second look at the normal cone of linear constraints

In Lecture 2, we considered normal cones for a few classes of feasible sets that come up often: hyperplanes, affine subspaces, halfspaces, and intersection of halfspaces. In that latter case, we drew a picture and were convinced that the normal cone at a point at the intersection of halfspaces was given by the conic hull of the directions orthogonal to those halfspaces (see the picture on the right).



We now give the proof of this result. We will do that by invoking the machinery of separation seen in Lecture 4 to argue that a direction outside of the normal cone must form an acute angle with at least one direction that remains in the feasible set Ω.

Theorem 1.1. Let $\Omega \subseteq \mathbb{R}^n$ be defined as the intersection of m linear inequalities

$$\Omega := \{x \in \mathbb{R}^n : Ax \le b\}, \quad \text{where} \quad A = \begin{pmatrix} — & a_1^\top & — \\ & \vdots & \\ — & a_m^\top & — \end{pmatrix} \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m.$$

Given a point $x \in \Omega$, define the index set of the "binding" (also known as "active") constraints

$$I(x) := \Big\{j \in \{1, ..., m\} : a_j^\top x = b_j\Big\}.$$

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

$$\mathcal{N}_\Omega(x) = \left\{ \sum_{j \in I(x)} \lambda_j a_j : \lambda_j \ge 0 \right\} = \Big\{A^\top \lambda : \lambda^\top (b - Ax) = 0, \lambda \in \mathbb{R}^m_{\ge 0}\Big\},$$

where the second equality rewrites the condition $j \in I(x)$ via the *complementary slackness* condition (see Lecture 2).

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

Remark 1.1. As a reminder, the coefficients λ in the normal cone above are typically called "Lagrange multipliers".

Remark 1.2. The result above immediately implies the correctness of the normal cone to an affine subspace too. This is because we can rewrite the condition $Ax = b$ as the intersection of $Ax \leq b$ and $-Ax \leq -b$:

$$\Omega \coloneqq \{x \in \mathbb{R}^n : Ax = b\} = \left\{x \in \mathbb{R}^n : \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}\right\}.$$

Applying the characterization in Theorem 1.1 to any $x \in \Omega$ using the right-hand side formulation yields that

$$\mathcal{N}_\Omega(x) = \left\{ A^\top \lambda_1 - A^\top \lambda_2 : \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}^\top \begin{pmatrix} b - Ax \\ -b + Ax \end{pmatrix} = 0, \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \in \mathbb{R}^{2m}_{\geq 0} \right\}.$$

Since by hypothesis $x \in \Omega$, then $Ax = b$ and so the complementary slackness condition is vacuous, and we are left with

$$\mathcal{N}_\Omega(x) = \left\{ A^\top (\lambda_1 - \lambda_2) : \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \in \mathbb{R}^{2m}_{\geq 0} \right\}.$$

It is clear that for any $\lambda_1, \lambda_2 \in \mathbb{R}^m_{\geq 0}$, their difference is $\lambda_1 - \lambda_2 \in \mathbb{R}^m$. Similarly, for any value of $\lambda \in \mathbb{R}^m$, we can find $\lambda_1, \lambda_2 \in \mathbb{R}^m_{\geq 0}$ such that $\lambda = \lambda_1 - \lambda_2$. Thus,

$$\mathcal{N}_\Omega(x) = \{A^\top \lambda : \lambda \in \mathbb{R}^m\}$$

as we argued (albeit using a different technique) in Lecture 2.

1.1 Separating a point from a convex cone

Before we can prove Theorem 1.1, we will find it helpful to use the following corollary of separation for *convex cones*. A *cone* is a set with the property that the ray $\{\lambda \cdot x : \lambda \geq 0\}$ generated by any point x in the set is fully contained in the set.

Definition 1.1 (Cone). A set S is a *cone* if, for any $x \in S$ and $\lambda \in \mathbb{R}_{\geq 0}$, the point $\lambda \cdot x \in S$.

Convex cones are among the simplest convex sets, and they appear all the time in optimization theory.[1] In particular, in the next theorem we show that separation of a point from a nonempty closed convex cone can always be achieved using a hyperplane passing through the origin.

Theorem 1.2. Let $S \subseteq \mathbb{R}^n$ be a nonempty closed convex cone, and $y \notin S$ be a point in $\mathbb{R}^n$. Then, there exists a hyperplane *passing through the origin* that separates y from S; formally, there exists $u \in \mathbb{R}^n$ such that

$$\langle u, y \rangle < 0 \quad \text{and} \qquad \langle u, x \rangle \geq 0 \quad \forall x \in S.$$

[1] Hiriart-Urruty, J.-B., & Lemaréchal, C. [HL01], speaking of convex cones, say: *"they are important in convex analysis (the "unilateral" realm of inequalities), just as subspaces are important in linear analysis (the "bilateral" realm of equalities)".*

Proof. We already know from Lecture 4 that there exist $u \in \mathbb{R}^n, v \in \mathbb{R}$ such that

$$\langle u, y\rangle < v \quad \text{and} \qquad \langle u, x\rangle \geq v \quad \forall x \in S. \tag{1}$$

Consider any point $a \in S$. By definition of cone, $\lambda \cdot a \in S$ for all $\lambda \geq 0$. Thus, the separation condition on the right in (1) implies that $v \leq \lambda \cdot \langle u, a\rangle$ for all $\lambda \geq 0$. In particular, by plugging $\lambda = 0$, we find that $v \leq 0$, yielding $\langle u, y\rangle < 0$. Furthermore, dividing by λ we find that

$$\langle u, a\rangle \geq \frac{v}{\lambda} \qquad \forall \lambda \geq 0 \qquad \Longrightarrow \qquad \langle u, a\rangle \geq \sup_{\lambda \to \infty} \frac{v}{\lambda} = 0.$$

Since $a \in S$ was arbitrary, the statement follows. □

1.2 The proof of Theorem 1.1

Proof (of Theorem 1.1). Fix any $x \in \Omega$ and let

$$\mathcal{C}(x) \coloneqq \left\{ \sum_{j \in I(x)} \lambda_j a_j : \lambda_j \geq 0 \right\}.$$

We will show that $\mathcal{N}_\Omega(x) = \mathcal{C}(x)$ by proving the two directions of inclusion separately.

- We start by showing that any $d \in \mathcal{C}(x)$ belongs to $\mathcal{N}_\Omega(x)$, that is,

  $$\langle d, y - x\rangle \leq 0 \text{ for all } y \in \Omega.$$

  Let d be expressed as $\sum_{j \in I(x)} \lambda_j a_j$ with $\lambda_j \geq 0$. Then, for any $y \in \Omega$,

  $$\begin{aligned}
  \left\langle \sum_{j \in I(x)} \lambda_j a_j, y - x \right\rangle &= \sum_{j \in I(x)} \lambda_j \langle a_j, y - x\rangle \\
  &= \sum_{j \in I(x)} \lambda_j \big(\langle a_j, y\rangle - b_j\big) \qquad \text{(by definition of } I(x), \langle a_j, x\rangle = b_j) \\
  &\leq \sum_{j \in I(x)} \lambda_j \big(b_j - b_j\big) = 0. \qquad \text{(since } y \in \Omega \text{ and } \lambda_j \geq 0)
  \end{aligned}$$

  This shows that $d \in \mathcal{N}_\Omega(x)$ and concludes the proof of this direction of the inclusion.

- We now look at the other direction. Take any $d \notin \mathcal{C}(x)$. Since $\mathcal{C}$ is a nonempty closed convex cone [▷ you should verify this claim], by the conic separation result of Theorem 1.2, there must exist $u \in \mathbb{R}^n$ such that

  $$\langle u, d\rangle < 0, \quad \text{and} \qquad \langle u, a\rangle \geq 0 \quad \forall a \in \mathcal{C}(x). \tag{2}$$

  We argue that for $\delta > 0$ small enough, the point $y \coloneqq x - \delta \cdot u$ belongs to Ω. We do so by showing that it satisfies all the inequalities $a_j^\top x \leq b_j$ that define Ω:
  - if $j \in I(x)$, then $\langle a_j, x - \delta \cdot u\rangle = b_j - \delta \cdot \langle a_j, u\rangle \leq b_j$ since $\langle a_j, u\rangle \geq 0$ by (2).
  - if $j \notin I(x)$, then $b_j - \langle a_j, x\rangle > 0$. By continuity, small enough perturbations of x, in any direction, will not affect the strict inequality.

  Thus, the direction $\delta \cdot u$ remains inside of Ω starting from x. We now argue that it forms a strictly positive inner product with d. Indeed, note that from (2)

  $$\langle d, y - x\rangle = \langle d, -\delta \cdot u\rangle = -\delta \cdot \langle d, u\rangle > 0.$$

  This shows that $d \notin \mathcal{C}(x) \Longrightarrow d \notin \mathcal{N}_\Omega(x)$, completing the proof. □

2 Karush-Kuhn-Tucker (KKT) conditions

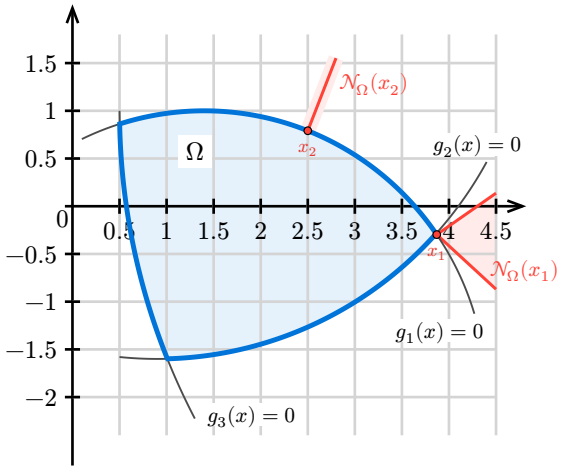
The result of Theorem 1.1 gives a complete characterization of the normal cone for sets defined as intersections of linear constraints. We now turn our attention to more general constraint sets, defined as the intersection of *differentiable*[2] functional constraints

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t.} \ & h_i(x) = 0 \quad i \in \{1, ..., r\} \\ & g_j(x) \leq 0 \quad j \in \{1, ..., s\}. \end{aligned} \tag{3}$$

2.1 The general idea

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

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



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.*[3]

Since the binding linearized constraints are of the form

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

from Theorem 1.1 we know that the normal cone of the linearization $\tilde{\Omega}$ of Ω around x^* is

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

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

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

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

---

[2] While in previous classes we could have gotten away with Gâteaux differentiability (*i.e.*, linearity of directional derivatives), in this lecture we assume *Fréchet* differentiability, that is, the fact that the functions can be locally approximated by their linearization at any point.

[3] 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].

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.

Definition 2.1 (KKT conditions). Consider a nonlinear optimization problem with differentiable objective function and functional constraints, in the form given in (3), and let x be a point in the feasible set ("Primal Feasibility"). The *KKT conditions* at x are given by

$$
\begin{aligned}
-\nabla f(x) &= \sum_{i=1}^{r} \lambda_i \nabla h_i(x) + \sum_{j=1}^{s} \mu_j \nabla g_j(x) && \text{("Stationarity")} \\
\lambda_i \in \mathbb{R}, \quad \mu_j \geq 0 \qquad & \forall i = 1, ..., r, \quad j = 1, ..., s && \text{("Dual feasibility")} \\
\mu_j \cdot g_j(x) &= 0 \qquad \forall j = 1, ..., s. && \text{("Complementary slackness")}
\end{aligned}
$$

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 $\mu_j = 0$".

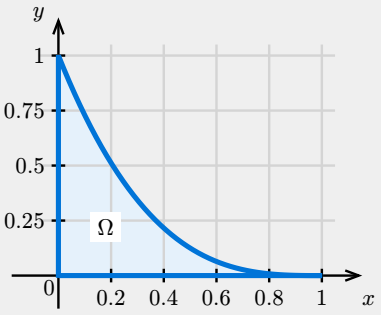
The KKT conditions are *often* necessary conditions for optimality (for example, in the picture above), but *not always*.

2.2 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 2.1 (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 $(x^*, y^*) := (1, 0)$, the gradients of the objective and the binding constraints (g_1 and g_3) are

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

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

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

The KKT conditions fail in this case.

The reason why the KKT conditions failed at the optimal point $(x^*, y^*) = (1, 0)$ in Example 2.1 is due to the fact that the linearization of constraint $g_1(x, y) \leq 0$ around the optimal point $(x^*, y^*) = (1, 0)$ is $y \leq 0$. This is parallel to the existing constraint $y \geq 0$, and fails the capture the fact that $x \leq 1$ on the feasible set Ω.

2.3 Constraint qualification

Conditions that prevent the degenerate behavior illustrated in Example 2.1 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 (3) are affine functions), then no further constraint qualifications hold, and the necessity of the KKT conditions is implied directly by Theorem 1.1.

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 2.1 (Concave and linear constraints). Let $x \in \Omega \subseteq \mathbb{R}^n$ be a minimizer of (3). 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.

■ **Linear independence of gradients.** In Example 2.1, the linearization of two constraints coincided, causing problems. When all linearized constraints are linearly independent, the issue is avoided.

Theorem 2.2 (Linear independence of gradients). Let $x \in \Omega \subseteq \mathbb{R}^n$ be a minimizer of (3). If all functions h_i are continuously differentiable and the set of gradients

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

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

For those interested, the conditions above is a special case of a much more general condition called *Mangasarian-Fromowitz constraint qualification* [MF67].

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

Theorem 2.3 (Slater's condition [Sla59]). Let $x \in \Omega \subseteq \mathbb{R}^n$ be a minimizer of (3). 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.

3 Further readings and bibliography

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

[HL01] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer, 2001. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-56468-0

[KT51] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2. University of California Press, Ewing, NJ, USA, pp. 481–493, Jan. 1951.

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

[Kje00] T. H. Kjeldsen, "A Contextualized Historical Analysis of the Kuhn–Tucker Theorem in Nonlinear Programming: The Impact of World War II," *Historia Math.*, vol. 27, no. 4, pp. 331–361, Nov. 2000, doi: 10.1006/hmat.2000.2289.

[MF67] O. L. Mangasarian and S. Fromovitz, "The Fritz John necessary optimality conditions in the presence of equality and inequality constraints," *J. Math. Anal. Appl.*, vol. 17, no. 1, pp. 37–47, Jan. 1967, doi: 10.1016/0022-247X(67)90163-1.

[Sla59] M. Slater, "Lagrange Multipliers Revisited," 1959.

[Gül10] O. Güler, *Foundations of Optimization*. New York, NY, USA: Springer, 2010. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-68407-9

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

[Ber16] D. P. Bertsekas, *Nonlinear Programming*, 3rd edition. Nashua, NH, USA: Athena Scientific, 2016.

[BL06] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*. New York, NY, USA: Springer, 2006. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-31256-9