MIT 6.7220/15.084 — Nonlinear Optimization Thu, Feb 8th 2024

Lecture 2
First-order optimality conditions

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

First-order optimality conditions define conditions that optimal points need to satisfy. For this lecture, we will make the blanket assumption that we work with *differentiable* functions.

1 Unconstrained optimization

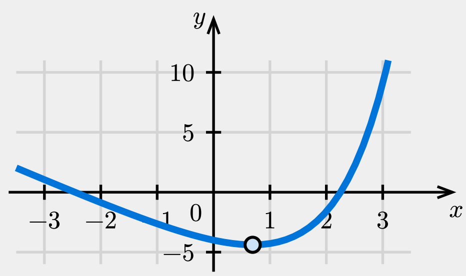
I'm pretty sure you have already encountered first-order optimality conditions for unconstrained optimization problems before. For example, consider the following optimization problem.

Example 1.1. Find a solution to the problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{R}, \end{aligned}$$

where the differentiable function $f : \mathbb{R} \to \mathbb{R}$, plotted on the right, is defined as

$$f(x) \coloneqq -2x + e^x - 5.$$



Solution. I expect that most students would have the same thought: *take the gradient of the function, set it to* 0*, and solve for* x! In this case, this leads to $-2 + e^x = 0$ which implies that the optimal point is $x^* = \log 2 \approx 0.693$. □

Now, in the above process we have been pretty informal. It is good to remember that when facing an optimization problem of the form $\min_{x \in \mathbb{R}^n} f(x)$, with $f(x)$ differentiable, solving $\nabla f(x) = 0$ has some limitations:

- It is only a *necessary* condition that all optimal points *need* to satisfy; but not all points that satisfy it are automatically optimal. [▷ For example, think about what happens with $f(x) = -x^2$? With $f(x) = x^3$? With $f(x) = x^3 + 3x^2 - 6x - 8$?]
- In other words, the solutions to $\nabla f(x) = 0$ form a list of *possible* minimizing points: solving $\nabla f(x) = 0$ allows us to *focus our attention on few promising candidate points* (some people call these "critical points"). It might give *false positives* but *never false negatives*: if a point fails the $\nabla f(x) = 0$ test, it cannot be optimal.

In practice, as you know from experience, *solving* $\nabla f(x) = 0$ *is a practical way of analytically solving unconstrained problems.* Today and next time, we will focus on the following two big questions:

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

- What is the correct generalization of the necessary condition $\nabla f(x) = 0$, when we are faced with a *constrained* optimization problem?
- Under what circumstances does $\nabla f(x) = 0$ also become sufficient for optimality?

2 Constrained optimization

In order to generalize the "$\nabla f(x) = 0$" condition to *constrained* optimization problems, it is important to make sure we are all on the same page as to why such a condition arises in the first place in unconstrained problems. From there, generalizing will be straightforward.

2.1 Where the zero gradient condition arises from in unconstrained optimization

The idea is very simple: if x is a minimizer of the function, then look at the values of the function $f : \mathbb{R}^n \to \mathbb{R}$ along the direction d. Clearly, $f(x + t \cdot d) \geq f(x)$ for all $t \geq 0$ (or x would not be a minimizer). Hence, the directional derivative $f'(x; d)$ of f at x along direction d,

$$f'(x; d) = \lim_{t \downarrow 0} \frac{f(x + t \cdot d) - f(x)}{t} \geq 0,$$

since the limit of a nonnegative sequence must be nonnegative.

By definition of gradient, we have $f'(x; d) = \langle \nabla f(x), d \rangle$, and so the previous inequality can be rewritten as

$$\langle \nabla f(x), d \rangle \geq 0 \qquad \forall d \in \mathbb{R}^n.$$

Because the above inequality must hold for all directions $d \in \mathbb{R}^n$, in particular it must hold for $d = -\nabla f(x)$, leading to

$$-\|\nabla f(x)\|^2 \geq 0 \qquad \iff \qquad \nabla f(x) = 0.$$

2.2 The constrained case

Now that we have a clearer picture of why the "$\nabla f(x) = 0$" condition arises in unconstrained problems, the extension to the constrained case is rather natural.

The main difference with the unconstrained case is that, in a constrained set, *we might be limited in the choices of available directions d along which we can approach x while remaining in the set.* Nonetheless, for any direction d such that $x + t \cdot d \in \Omega$ for all $t \geq 0$ sufficiently small, the above argument applies without changes, and we can still conclude that necessarily $\langle \nabla f(x), d \rangle \geq 0$.

So, the natural generalization of the "$\nabla f(x) = 0$" condition to constrained problems can be informally stated as follows: for the optimality of x it is *necessary* that

$$\langle \nabla f(x), d \rangle \geq 0 \qquad \text{for all } d \in \mathbb{R}^n \text{ that remain in } \Omega \text{ from } x. \tag{1}$$

In order to instantiate the above condition, two steps are required:

1. first, we need to determine what the set of "directions d that remain in Ω from x" is.
2. then, based on the directions above, see in what way they constrain $\nabla f(x)$. For example, we have seen before that when the set of all directions spans the entire space $\mathbb{R}^n$, then $\nabla f(x) = 0$.

Out of the two, usually the first point is the easiest. In all the cases that will be of our interest, we can determine the set of directions that remain in Ω from x by simply considering any other $y \in \Omega$ and considering the direction from x to y. This holds trivially if all line segments between x and any point in Ω are entirely contained in Ω, a condition known as *star-convexity at x*.

Definition 2.1 (Star-convexity at x). A set $\Omega \subseteq \mathbb{R}^n$ is said to be *star-convex* at a point $x \in \Omega$ if, for all $y \in \Omega$, the entire segment from x to y is contained in Ω. In symbols, if

$$x + t \cdot (y - x) \in \Omega \qquad \forall t \in [0, 1].$$

(Note that the condition is equivalent to "$t \cdot y + (1 - t) \cdot x \in \Omega$ for all $y \in \Omega$ and $t \in [0, 1]$", or also "$t \cdot x + (1 - t) \cdot y \in \Omega$ for all $y \in \Omega$ and $t \in [0, 1]$".)

In fact, for all our purposes today, we will only consider sets that are star-convex at all of their points. Such sets are simply called *convex*.

Definition 2.2 (Convex set). A set Ω is convex if it is star-convex at all of its points $x \in \Omega$. In other words, Ω is convex if all segments formed between any two points $x, y \in \Omega$ are entirely contained in Ω. In symbols, if

$$t \cdot x + (1 - t) \cdot y \in \Omega \qquad \forall x, y \in \Omega \text{ and } t \in [0, 1].$$

Under assumption of convexity, the condition (1) can be equivalently rewritten as follows.

Theorem 2.1 (First-order necessary optimality condition for a convex feasible set). Let $\Omega \subseteq \mathbb{R}^n$ be convex and $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. For a point $x \in \Omega$ to be a minimizer of f over Ω it is *necessary* that

$$\langle \nabla f(x), y - x \rangle \geq 0 \qquad \forall y \in \Omega.$$

2.3 Geometric intuition: normal cones

The condition established in Theorem 2.1 has the following geometric interpretation: the gradient of f at a solution $x \in \Omega$ must form an acute angle with all directions $y - x$, $y \in \Omega$. While this makes perfect sense, it is actually more customary, for mental visualization purposes, to flip signs and instead have the following useful mental picture: at any solution $x \in \Omega$, the opposite of the gradient $-\nabla f(x)$ must form an *obtuse* angle with all directions $y - x$, $y \in \Omega$. In other words, $-\nabla f(x)$ can only "look" in those directions in which the set is not in the 90° cone of vision.

Of course, depending on the shape of the set Ω and the particular point $x \in \Omega$, the set of directions that point away from the set might be extremely limited—for example we have seen earlier that when $\Omega = \mathbb{R}^n$, then no directions "point away" from Ω, and the only possible value for $-\nabla f(x)$ is therefore 0. This mental picture of "directions pointing away" from Ω is generally pretty useful, and we give it a name.

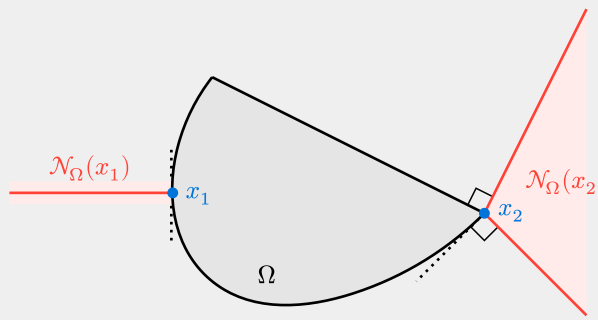
Definition 2.3 (Normal cone). Let $\Omega \subseteq \mathbb{R}^n$ be convex, and let $x \in \Omega$. The *normal cone to* Ω *at* x, denoted $\mathcal{N}_\Omega(x)$, is defined as the set

$$\mathcal{N}_\Omega(x) := \{d \in \mathbb{R}^n : \langle d, y - x \rangle \leq 0 \quad \forall y \in \Omega\}.$$

With this definition, the first-order necessary optimality condition for x, given in Theorem 2.1, can be equivalently written as

$$-\nabla f(x) \in \mathcal{N}_\Omega(x).$$

Example 2.1. As an example, here are a few normal cones computed for a convex set.



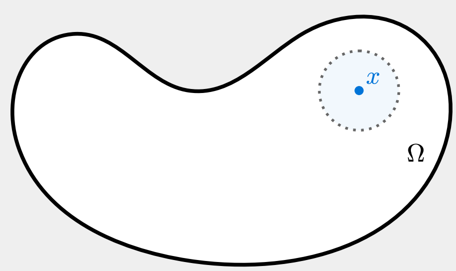
3 Normal cones to some notable sets, and applications

Let's build our intuition regarding normal cones by considering examples that are progressively harder. Along the way, we will see that first-order optimality conditions, in all their simplicity, imply some of the deepest results in optimization theory.

3.1 Point in the interior

Let's start from an easy example: the normal cone of a point in the interior of the feasible set.[1]

Example 3.1 (Normal cone at an interior point). What is the normal cone $\mathcal{N}_\Omega(x)$ of a point x in the *interior* of the feasible set Ω?



Solution. In this case, the normal cone contains *only the zero vector*, that is,

$$\mathcal{N}_\Omega(x) = \{0\}.$$

This is easy to prove: if any $d \neq 0$ were to belong to $\mathcal{N}_\Omega(x)$, then we could consider the point $x + \delta d$ for sufficiently small $\delta > 0$, and have

$$\langle d, x + \delta d - x \rangle = \delta \|d\|^2 > 0.$$

Hence, for a point x in the interior of Ω to be optimal, it is *necessary* that $\nabla f(x) = 0$. □

This fully recovers what we know for unconstrained domains, since there every point is in the interior of the feasible set.

[1] Remember that a point is in the *interior* of a set if you can draw a ball centered at the point, such that the ball is fully contained in the set.

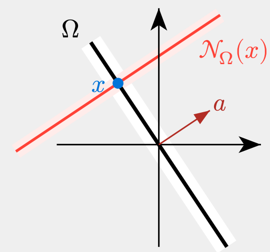
3.2 Point on a hyperplane / subspace

Next up, we consider the normal cone to a point on a hyperplane.

Example 3.2 (Normal cone to a hyperplane). Consider a hyperplane

$$\Omega := \{y \in \mathbb{R}^n : \langle a, y\rangle = 0\}, \quad \text{where } a \in \mathbb{R}^n, a \neq 0$$

and a point $x \in \Omega$.



It is pretty intuitive from the picture that

$$\mathcal{N}_\Omega(x) = \text{span}\{a\} = \{\lambda \cdot a : \lambda \in \mathbb{R}\}.$$

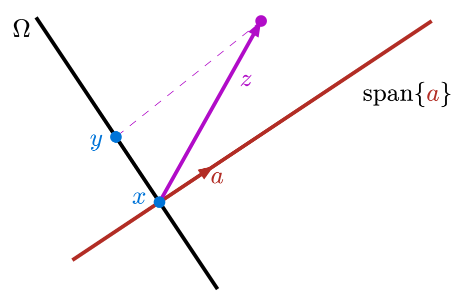
Solution. In order to convert this intuition into a formal proof, [▷ before continuing, try to think how you would go about proving this yourself!] we would need to show two things:

- all points in span$\{a\}$ *do* indeed belong to $\mathcal{N}_\Omega(x)$; by convexity, this means that we need to show that all points $z \in \text{span}\{a\}$ satisfy
$$\langle z, y - x\rangle \leq 0 \quad \forall y \in \Omega;$$
- none of the points outside of span$\{a\}$ belong to $\mathcal{N}_\Omega(x)$; that is, for any point $z \notin \text{span}\{a\}$, then there exists $y \in \Omega$ such that $\langle z, y - x\rangle > 0$.

The first point is straightforward: by definition of span, all points in span$\{a\}$ are of the form $\lambda \cdot a$ for some $\lambda \in \mathbb{R}$. But then, for all $y \in \Omega$,

$$\langle z, y - x\rangle = \langle \lambda \cdot a, y - x\rangle = \lambda \cdot \langle a, y\rangle - \lambda \cdot \langle a, x\rangle = 0 - 0 \leq 0,$$

where the last equality follows from the definition of Ω and the fact that both x and y belong to it. To prove the second point, we can let the geometric intuition guide us. Draw a vector $z \notin \text{span}\{a\}$ applied to x, and look at the picture:



We can project the point $x + z$ onto Ω, finding some $y \in \Omega$, and onto $x + \text{span}\{a\}$, finding some point $x + k \cdot a$:

$$z = (y - x) + k \cdot a.$$

We now show that z cannot be in $\mathcal{N}_\Omega(x)$, because it would have a positive inner product with $y - x$:

$$\begin{aligned}\langle z, y - x\rangle &= \langle (y - x) + k \cdot a, y - x\rangle \\ &= \|y - x\|^2 + k \cdot \langle a, y - x\rangle = \|y - x\|^2.\end{aligned}$$

Since z was not aligned with $\mathrm{span}\{a\}$ by hypothesis, then $y \neq x$, and therefore $\langle z, y - x\rangle > 0$ as we wanted to show. □

Remark 3.1. Because normal cones are insensitive to shifts in the set, the result above applies without changes to any *affine* plane $\Omega := \{y \in \mathbb{R}^n : \langle a, y\rangle = b\}$, with $a \in \mathbb{R}^n, b \in \mathbb{R}$. Again, $\mathcal{N}_\Omega(x) = \mathrm{span}\{a\} = \{\lambda \cdot a : \lambda \in \mathbb{R}\}$ at any $x \in \Omega$.

Remark 3.2. The same argument above, based on decomposing $x + z$ onto Ω and its orthogonal complement $\mathrm{span}\{a\}$ applies to lower-dimensional affine subspaces

$$\Omega := \{y \in \mathbb{R}^n : Ay = b\}.$$

In this case, we obtain that

$$\mathcal{N}_\Omega(x) = \mathrm{colspan}(A^\top).$$

(This immediately recovers Example 3.2 by considering $A = a^\top$)

In the case of Remark 3.2, the argument above with the projection goes through verbatim. In this case, one would need to project $x + z$ onto $\mathrm{colspan}(A^\top)$ and onto Ω.[2]

Remark 3.3 (Lagrange multipliers). The discussion we just had, shows that whenver we have a problem of the form

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & Ax = b \\ & x \in \mathbb{R}^n, \end{aligned}$$

at optimality it needs to hold that

$$-\nabla f(x) = A^\top \lambda, \qquad \text{for some } \lambda \in \mathbb{R}^d$$

where d is the number of rows of A. This necessity of being able to express—at optimality—the gradient of the objective as a combination of the constraints is very general. The entries of λ are an example of *Lagrange multipliers*.

In the next two subsections, we will see how the characterization of the normal cone to affine subspaces enables us to solve a couple of problems that arise in practice.

[2]The orthogonality of $\mathrm{colspan}(A^\top)$ and Ω is a reflection of the well-known linear algebra result that the orthogonal complement of the nullspace of a matrix is the span of the columns of the transpose matrix.

3.2.1 Application: projection onto an affine subspace

Example 3.3. Consider the nonempty set $\Omega := \{x \in \mathbb{R}^n : Ax = b\}$, where $A \in \mathbb{R}^{d\times n}$ is such that $AA^\top$ is invertible. Prove that the Euclidean projection x of a point z onto Ω, that is, the solution to[3]

$$\begin{aligned} \min_x \quad & \tfrac{1}{2}\|x - z\|_2^2 \\ \text{s.t.} \quad & x \in \Omega, \end{aligned}$$

is given by

$$x = z - A^\top (AA^\top)^{-1}(Az - b).$$

Solution. Since the gradient of the objective at any point x is $(x - z)$, from the first-order optimality conditions any solution x must satisfy

$$-(x - z) \in \mathcal{N}_\Omega(x).$$

From Remark 3.2, we know that at any $x \in \Omega$, $\mathcal{N}_\Omega(x) = \operatorname{colspan}(A^\top) = \{A^\top\lambda : \lambda \in \mathbb{R}^n\}$. So, at optimality there must exist $\lambda \in \mathbb{R}^d$ such that

$$-(x - z) = A^\top\lambda \quad \Longrightarrow \quad x = z - A^\top\lambda.$$

Furthermore, since $x \in \Omega$, we have $Ax = b$. Plugging the above expression for x we thus have

$$A(z - A^\top\lambda) = b \quad \Longrightarrow \quad (AA^\top)\lambda = Az - b.$$

Solving for λ and plugging back into $x = z - A^\top\lambda$ yields the result. □

3.2.2 Application: entropy-regularized linear optimization (softmax)

As a second example application, we will consider a real problem that comes up naturally in online learning and reinforcement learning: entropy-regularized best responses. We will have plenty of time in the future to talk about *where* such problem arises; for now, let's cut to the chase and consider what the optimization problem looks like.

Example 3.4. Consider the set of probability distributions over n actions $\{1, ..., n\}$ that have *full* support, that is, the set $\mathring{\Delta}^n := \{(x_1, ..., x_n) \in \mathbb{R}^n_{>0} : \quad x_1 + \cdots + x_n = 1\}$. Given an assignment of *values* v_i for each action $i = 1, ..., n$, the *entropy-regularized best response* given the values is the distribution that solves the following problem:

$$\begin{aligned} \min_x \quad & g(x) := -\sum_{i=1}^n v_i x_i + \sum_{i=1}^n x_i \log x_i \\ \text{s.t.} \quad & x \in \mathring{\Delta}^n, \end{aligned}$$

Show that the solution to this problem is the distribution that picks action i with probability proportional to the exponential of the value v_i of that action:

$$x_i = \frac{e^{v_i}}{\sum_{i=1}^n e^{v_i}}.$$

[3] We already know from Lecture 1 that the projection must exist since Ω is nonempty and closed.

Solution. We'll leave showing that the nonlinear optimization problem has a solution as exercise. Here, we show that the first-order optimality conditions imply that the solution necessarily has components proportional to e^{v_i}.

Pick any point $x \in \mathring{\Delta}^n$. The set of directions that remain inside $\mathring{\Delta}^n$ span the entire plane: the constraint $x_i > 0$ is *completely inconsequential* for the purposes of first-order optimality conditions. In other words, we are *exactly* in the same setting as Example 3.2, where in this case $a = 1 \in \mathbb{R}^n$. Hence, whatever the solution x to the problem might be, it is *necessary* that $-\nabla g(x)$ be in the normal cone $\mathcal{N}_{\mathring{\Delta}^n}(x) = \operatorname{span}\{1\} \subset \mathbb{R}^n$. So, there must exist $\lambda \in \mathbb{R}$ such that

$$\underbrace{\begin{pmatrix} -v_1 + 1 + \log x_1 \\ \vdots \\ -v_n + 1 + \log x_n \end{pmatrix}}_{-\nabla g(x)} \underbrace{= \lambda \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{\in \mathcal{N}_{\mathring{\Delta}^n}(x)} \qquad \iff \qquad \log x_i = \lambda - 1 + v_i \quad \forall i = 1, ..., n.$$

Exponentiating on both sides, we have

$$x_i = \exp(\lambda - 1 + v_i) = \alpha \cdot \exp(v_i), \qquad \text{where} \quad \alpha := \exp(\lambda - 1) \in \mathbb{R}.$$

This shows that at optimality there exists a proportionality constant α such that $x_i = \alpha \cdot e^{v_i}$ for all $i = 1, ..., n$. Since $\sum_{i=1}^n x_i = 1$, we find that

$$\alpha \sum_{i=1}^n e^{v_i} = 1 \qquad \Longrightarrow \qquad \alpha = \frac{1}{\sum_{i=1}^n e^{v_i}},$$

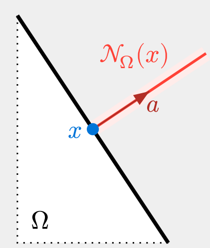
and the result follows. □

3.3 Point in a halfspace / at the intersection of halfspaces

Example 3.5 (Normal cone to a halfspace). Consider a halfspace

$$\Omega := \{y \in \mathbb{R}^n : \langle a, y \rangle \leq b\}, \quad \text{where } a \in \mathbb{R}^n, a \neq 0$$

and a point $x \in \Omega$.



We already know that if x is in the interior of Ω (that is, $\langle a, x \rangle < b$), then $\mathcal{N}_\Omega(x) = \{0\}$. On the other hand, if x is on the boundary of Ω, that is, $\langle a, x \rangle = b$, then it is pretty intuitive from the picture that

$$\mathcal{N}_\Omega(x) = \{\lambda \cdot a : \lambda \in \mathbb{R}_{\geq 0}\}.$$

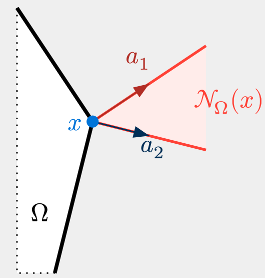
The crucial difference between the above result and that of Example 3.2, is that the normal cone to a *halfspace* only points in one direction (so $\lambda \geq 0$), as opposed to the case of the hyperplane where $\lambda \in \mathbb{R}$. The formal proof can be obtained by adapting the arguments we used in Example 3.2. [▷ You should try to work out the details!]

We now consider the case of the *intersection* of two halfspaces.

Example 3.6 (Normal cone to the intersection of two halfspaces). Consider the intersection Ω of two halfspaces:

$$\Omega := \left\{ y \in \mathbb{R}^n : \begin{array}{l} \langle a_1, y\rangle \le b_1 \\ \langle a_2, y\rangle \le b_2 \end{array} \right\}, \quad \text{where } a_1, a_2 \in \mathbb{R}^n, a_1, a_2 \neq 0,$$

and a point $x \in \Omega$.



If x is in the interior of Ω, or if x is on the boundary of one of the halfspaces but not both, then the normal cone follows from our prior results. So, consider x at the intersection of both halfspaces, that is, $\langle a_1, x\rangle = b_1, \langle a_2, x\rangle = b_2$.

It is pretty intuitive from the picture that the normal cone at x is the *conic hull* of a_1 and a_2, that is, the set obtained from summing the directions in all possible ways:

$$\mathcal{N}_\Omega(x) = \{\lambda_1 \cdot a_1 + \lambda_2 \cdot a_2 : \lambda_1, \lambda_2 \in \mathbb{R}_{\ge 0}\}.$$

Formally proving the above result requires a bit of simple machinery that we have not seen yet; it will be the topic of Lecture 5. For now, we take the result as a given. The generalization of the result to the case of m halfspaces is equally intuitive, and we present it next.

Theorem 3.1. Let $\Omega \subseteq \mathbb{R}^n$ be given as the intersection of m halfspaces $\langle a_j, x\rangle \le b_j$. Then, the normal cone at any point $x \in \Omega$ is obtained by taking combinations of all those a_j's for which $\langle a_j, x\rangle = b_j$. In symbols:

$$\mathcal{N}_\Omega(x) = \left\{ \sum_{j \in I(x)} \lambda_j \cdot a_j : \lambda_j \in \mathbb{R}_{\ge 0} \right\}, \qquad \text{where } I(x) = \{ j \in \{1, ..., m\} : \langle a_j, x\rangle = b_j \}.$$

The constraints j in $I(x)$ are often called the "active constraints" at $x \in \Omega$.

Remark 3.4. In the definition of $\mathcal{N}_\Omega(x)$ in Theorem 3.1, instead of summing over $j \in I(x)$, one could equivalently sum over all $j = 1, ..., m$, and then impose the constraint that $\lambda_j = 0$ for all $j \notin I(x)$:

$$\mathcal{N}_\Omega(x) = \left\{ \sum_{j=1}^{m} \lambda_j \cdot a_j : \lambda_j \ge 0, \lambda_j = 0 \text{ if } \langle a_j, x\rangle < b_j \right\}.$$

Because $\lambda_j \ge 0$, the condition that $\lambda_j = 0$ whenever $\langle a_j, x\rangle < b_j$ can be succintly rewritten as

$$\sum_{j=1}^{m} \lambda_j (b_j - \langle a_j, x\rangle) = 0.$$

The above condition is usually called "complementary slackness".

3.3.1 Application: derivation of linear programming duality

As a last illustration of the surprising power of first-order optimality conditions for nonlinear programming, we show how linear programming duality follows immediately as a corollary of the previous result on normal cones.

Consider the linear program

$$\begin{aligned} \max_{x} \quad & f(x) \coloneqq c^\top x \\ \text{s.t.} \quad & Ax \le b \\ & x \in \mathbb{R}^n, \end{aligned} \tag{P}$$

where the matrix A is assumed to have m rows. We will show that the first-order necessary optimality conditions imply that in order for x to be an optimal solution to (P), then the following “dual” problem also has a solution:

$$\begin{aligned} \min_{\lambda} \quad & g(\lambda) \coloneqq b^\top \lambda \\ \text{s.t.} \quad & A^\top \lambda = c \\ & \lambda \ge 0 \end{aligned} \tag{D}$$

From the first-order necessary optimality conditions for (P), we know that any optimal x^* must satisfy the condition (note that the problem is a maximization rather then minimization)

$$\nabla f(x^*) \in \mathcal{N}_\Omega(x), \qquad \text{where } \Omega \coloneqq \{x \in \mathbb{R}^n : Ax \le b\}. \tag{4}$$

From Theorem 3.1 and Remark 3.4, the normal cone $\mathcal{N}_\Omega(x)$ is a combination of the rows of A, that is,

$$\mathcal{N}_\Omega(x^*) = \{A^\top \lambda : \lambda^\top (b - Ax^*) = 0, \lambda \ge 0\}, \tag{5}$$

where the condition $\lambda^\top(b - Ax^*) = 0$ is the complementary slackness condition mentioned in Remark 3.4. Plugging (5) into (4), together with the fact that $\nabla f(x^*) = c$, we obtain that at optimality there exist $\lambda^* \in \mathbb{R}^n$ such that

$$c = A^\top \lambda^*, \quad (\lambda^*)^\top (b - Ax^*) = 0, \quad \lambda^* \ge 0.$$

Plugging in the fact that $A^\top \lambda^* = c$ in the middle equality, yields that

$$(\lambda^*)^\top b = c^\top x^*.$$

Hence, from the mere existence of an optimal point $x^* \in \Omega$ for (P), we deduced that there exists a point λ^* in the feasible set of (D) that achieves objective value $g(\lambda^*) = b^\top \lambda^* = c^\top x^* = f(x^*)$. To conclude that λ^* is optimal for (D), it just suffices to show that all other feasible points for (D) must achieve value $\ge c^\top x^*$. This is pretty straightforward. All feasible vectors λ for (D) satisfy $\lambda \ge 0$ and $A^\top \lambda = c$. Furthermore, $b \ge Ax^*$ since $x^* \in \Omega$. Hence, for any feasible λ of (D),

$$\begin{aligned} g(\lambda) &= b^\top \lambda \\ &\ge (Ax^*)^\top \lambda \\ &= (x^*)^\top A^\top \lambda \\ &= (x^*)^\top c = f(x^*). \end{aligned}$$

Since $g(\lambda) \ge f(x^*)$ for all feasible λ in (D), and we know that a feasible λ^* that achieves $g(\lambda^*) = f(x^*)$ exists. Then, such λ^* must clearly be optimal for (D).

Thus, with just a simple application of normal cones, we have shown the following result:

Theorem 3.2 (Strong linear programming duality). If (P) admits an optimal solution x^*, then (D) admits an optimal solution λ^*, such that:

- the values of the two problems coincide: $c^\top x^* = b^\top \lambda^*$; and
- the solution λ^* satisfies the complementary slackness condition $(\lambda^*)^\top (b - Ax^*) = 0$.