MIT 6.7220/15.084 — Nonlinear Optimization (Spring '25) Tue, Mar 4th 2025

Lecture 8
Lagrangian duality

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

We know since Lecture 3, when we introduced convexity, that the first-order optimality condition

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

is both *necessary* and *sufficient* for optimality in convex optimization problems—that is, problems in which the objective function f is convex, and the feasible set Ω is convex. With the machinery of Lecture 7, we now know that if Ω is expressed via functional constraints, and Slater's condition holds (or all the constraints are linear), the KKT conditions are both necessary and sufficient (and in fact, the normal cone to the linearization at each point matches the normal cone to the original feasible set).

In other words, consider a convex optimization problem of the form

$$(P) := \begin{cases} \min\limits_x \; f(x) & & f : \mathbb{R}^n \to \mathbb{R} \text{ convex} \\ \text{s.t. } g_j(x) \le 0 & j = 1, ..., r & g_j : \mathbb{R}^n \to \mathbb{R} \text{ convex} \\ \quad\; h_i(x) = 0 & i = 1, ..., s & h_i : \mathbb{R}^n \to \mathbb{R} \text{ affine,} \end{cases}$$

under the assumption that either all g_j are affine, or that Slater's condition holds at all points, that is, there exists a point x_0 such that $g_j(x_0) < 0$ for all $j = 1, ..., r$. Then, we know from Lecture 7 that the following two statements are equivalent:

A The point x^* is optimal for (P) $\iff$ **B** The point $x^* \in \Omega$ admits $\lambda^* \in \mathbb{R}^r_{\ge 0}, \mu^* \in \mathbb{R}^s$ such that

$$-\nabla f(x^*) = \sum_{j=1}^r \lambda_j^* \nabla g_j(x^*) + \sum_{i=1}^s \mu_i^* \nabla h_i(x^*),$$
$$\lambda_j^* g_j(x^*) = 0 \quad \forall j \in \{1, ..., r\}.$$

(As a reminder, the implication **A** $\implies$ **B** is necessity of KKT conditions and comes from constraint qualification. The reverse direction **B** $\implies$ **A** is sufficiency of the KKT conditions and requires convexity of f and g_j and the affinity of h_i.)

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

In this lecture, we ponder the depth of this statement by interpreting it from three different points of view. As usual, we will use the letter Ω to denote the feasible set of (P), that is, $\Omega := \{x \in \mathbb{R}^n : g_j(x) \leq 0, h_i(x) = 0 \text{ for all } i = 1, ..., s, j = 1, ..., r\}$.

L8.1 Point of view #1: Certification of optimality

In a sense, the multipliers λ^*, μ^* are nothing but a reflection of the expression for the normal cone to the linearization of the feasible set: it just so happens that the normal cone to linear constraints can be written as a linear combination of vectors, and λ^*, μ^* are simply these combination coefficients. However, an aspect worth paying attention to is the very different *conceptual flavor* of the two statements **A** and **B**:

- The statement **A** is asserting that x^* is better than *all* other points in the feasible set. It is a "*for all*" kind of statement.
- The statement **B** is asserting that x^* admits *one* set of multipliers λ^*, μ^* with certain properties that are easy to check. It is an "*exists*" kind of statement.

This equivalence for example implies that *verifying* whether a point is optimal is easy, at least assuming that all gradients of all functions involved only use rational numbers and a reasonable number of bits in their encoding. Indeed, if someone claims that x^* is optimal, we could ask them to show us the multipliers λ^*, μ^*. This ability to efficiently certify optimality is not at all trivial.

In fact, assuming that all gradients of all functions involved only use rational numbers and a reasonable number of bits in their encoding, we can go even further. Given a putative optimal solution x^*, we could compute $\nabla f(x^*)$ and the gradients $\nabla g_j(x^*), \nabla h_i(x^*)$ of the active constraints, and determine whether suitable $\lambda^* \in \mathbb{R}^r_{\geq 0}, \mu^* \in \mathbb{R}^s$ exist by using a linear program. This shows that we would not even need to ask for multipliers, but we could instead come up with them ourselves.

L8.2 Point of view #2: From constraints to penalizations

By rearranging the terms, the gradient equality in **B** can be written as

$$\nabla f(x^*) + \sum_{j=1}^{r} \lambda_j^* \nabla g_j(x^*) + \sum_{i=1}^{s} \mu_i^* \nabla h_i(x^*) = 0.$$

It is evident that the left-hand side has the flavor of the gradient of a function. In particular, it is the gradient with respect to x and evaluated at the point $(x^*; \lambda^*, \mu^*)$ of the function

$$\mathcal{L} : \mathbb{R}^n \times (\mathbb{R}^r_{\geq 0} \times \mathbb{R}^s) \to \mathbb{R}, \qquad \mathcal{L}(x; \lambda, \mu) := f(x) + \sum_{j=1}^{r} \lambda_j g_j(x) + \sum_{i=1}^{s} \mu_i h_i(x).$$

The function $\mathcal{L}$ is called the *Lagrangian function* associated with problem (P).

Given any $\mu \in \mathbb{R}^s$ and $\lambda \in \mathbb{R}^r_{\geq 0}$, the function $\mathcal{L}(x; \lambda, \mu)$ is a convex function of $x \in \mathbb{R}^n$. Since $\mathbb{R}^n$ is open and the Lagrangian is convex in x, we know from Lecture 3 that the condition

$\nabla_x \mathcal{L}(x; \lambda, \mu) = 0$ is both necessary and sufficient for (*i.e.*, it is equivalent to) optimality of x in the minimization of $\mathcal{L}(x; \lambda, \mu)$ over $\mathbb{R}^n$.

In light of the above observation, we can add a third equivalent characterization **C** to our block diagram with **A** and **B** above, as follows.

A The point x^* is optimal for (P) $\iff$

B The point $x^* \in \Omega$ admits $\lambda^* \in \mathbb{R}^r_{\geq 0}, \mu^* \in \mathbb{R}^s$ such that

$$-\nabla f(x^*) = \sum_{j=1}^{r} \lambda_j^* \nabla g_j(x^*) + \sum_{i=1}^{s} \mu_i^* \nabla h_i(x^*),$$
$$\lambda_j^* g_j(x^*) = 0 \quad \forall j \in \{1, ..., r\}.$$

$\Updownarrow$

C There exist penalization coefficients $\mu^* \in \mathbb{R}^s$, $\lambda^* \in \mathbb{R}^r_{\geq 0}$ such that

x^* is a minimizer of $\mathcal{L}(x; \lambda^*, \mu^*)$ over $\mathbb{R}^n$,
$x^* \in \Omega$, and $\lambda_j^* g_j(x^*) = 0 \quad \forall j \in \{1, ..., r\}$.

To appreciate why **C** is itself worthy of attention, we need to discuss the interpretation of the Lagrangian function.

L8.2.1 Interpretation of the Lagrangian function

The Lagrangian function is defined not on the feasible set Ω, but on the entire space $\mathbb{R}^n$. It can be thought of as a *relaxation* of the original problem, in which, instead of constraining x to satisfy all constraints $g_j(x) \leq 0$ and $h_i(x) = 0$, it simply penalizes those x that do not satisfy the constraints by adding in the objective a penalty term $\lambda_j g_j(x)$ for each constraint $g_j(x) \leq 0$ and a penalty term $\mu_j h_i(x)$ for the equality constraints. The coefficients $\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s$ control the magnitude of the penalization.

To build intuition, let's consider two extremes: zero penalization and infinite penalization. When the amount of penalization is zero, the Lagrangian function is simply the objective function f considered as a function on $\mathbb{R}^n$, that is, ignoring all constraints. In this case, it is clear that the infimum of $\mathcal{L}(x; 0, 0)$ is a lower bound on the value of $f(x)$, and in general we expect that if a minimum exists, it might very well lie outside of the feasible set Ω of (P). As we increase the penalization, the Lagrangian function becomes more and more sensitive to the constraints, and the minimum of $\mathcal{L}$ will move towards the feasible set Ω. A key observation is that, since the Slater's condition hold, when $\lambda \to +\infty$, the minimum of $\mathcal{L}$ will be attained in the interior of the inequality constraints.

In other words, as the penalization coefficients λ, μ are varied, the Lagrangian $\mathcal{L}(x; \lambda, \mu)$ interpolates between two different priorities: minimizing the original objective function f, at the cost of potentially violating the constrained set Ω; and finding any point in the (relative) interior of the feasible set Ω, at the cost of ignoring the objective function $f(x)$.

Example L8.1. As a concrete example, consider the optimization problem

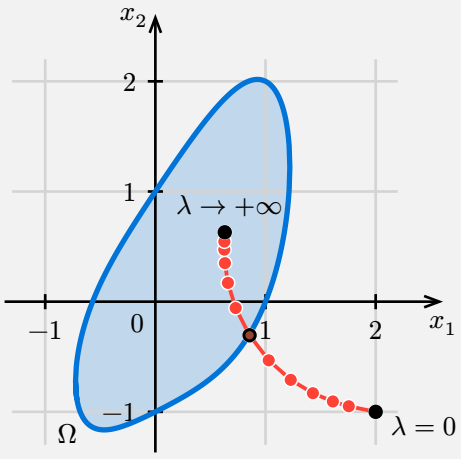
$$\begin{aligned} \min_x \quad & f(x) := (x_1 - 2)^2 + (x_2 + 1)^2 \\ \text{s.t.} \quad & g(x) := x_1^4 + (x_1 - x_2)^2 - (x_1 + 1) \leq 0 \\ & x \in \mathbb{R}^2, \end{aligned}$$

This problem is convex and satisfies Slater's conditions [▷ You should check this!].

The figure on the right shows the trajectory of the minimizers of the Lagrangian functions

$$\arg\min_{x \in \mathbb{R}^n} \{\mathcal{L}(x, \lambda) := f(x) + \lambda g(x)\}$$

as the penalization coefficient $\lambda \geq 0$ varies from 0 to $+\infty$.



L8.2.2 Existence of finite penalization coefficients

Given the tension between the two priorities encoded by the Lagrangian function and discussed above, it is natural to ask:

> *Is there an intermediate choice of penalization coefficients μ, λ that balances between the two priorities in a way that makes minimizing the Lagrangian equivalent to solving the original problem* (P)*?*

This is the content of the following theorem and the meaning of **C**.

Theorem L8.1. If the constrained problem (P) has a minimizer and attains optimal value value(P), there exist concrete penalization coefficients $(\lambda^*, \mu^*) \in \mathbb{R}^r_{\geq 0} \times \mathbb{R}^s$ such that:

1. any optimal solution x^* of the original problem is a minimizer of $\mathcal{L}(x; \lambda^*, \mu^*)$; and
2. the value of $\min_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda^*, \mu^*)$ is exactly equal to value(P).

Furthermore,

3. if $x^* \in \mathbb{R}^n$ is a minimizer of $\mathcal{L}(x; \lambda^*, \mu^*)$, and it satisfies $x^* \in \Omega, \lambda_j^* g_j(x^*) = 0$ for all $j = 1, ..., r$, then it is also an optimal solution of (P).

Proof.

1. Again, this is immediate from the implication **A** $\Longrightarrow$ **C**.
2. Let x^* be any minimizer of (P). From the implication **A** $\Longrightarrow$ **C**, we know that there exist coefficients $\lambda^* \in \mathbb{R}^r_{\geq 0}, \mu^* \in \mathbb{R}^s$ such that x^* is a minimizer of the penalized objective function $\mathbb{R}^n \ni x \mapsto \mathcal{L}(x; \lambda^*, \mu^*)$, and $\lambda_j^* g_j(x^*) = 0$ for all $j = 1, ..., r$. Hence,

$$\min_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda^*, \mu^*) = \mathcal{L}(x^*; \lambda^*, \mu^*) = f(x^*) + \sum_{j=1}^{r} \lambda_j^* g_j(x^*) + \sum_{i=1}^{s} \mu_i^* h_i(x^*) = f(x^*),$$

where in the last equality we used the fact that $h_i(x^*) = 0$ (since $x^* \in \Omega$), and the complementary slackness conditions $\lambda_j^* g_j(x^*) = 0$.

3. This is nothing but the implication **C** $\Longrightarrow$ **A**.

□

Remark L8.1. Theorem L8.1 asserts that if the Lagrangian is set up with suitable penalization coefficients λ^*, μ^*, the optimal *value* of the constrained problem (P) matches the optimal *value* of the Lagrangian $\mathbb{R}^n \ni x \mapsto \mathcal{L}(x; \lambda^*, \mu^*)$. Furthermore, all optimal points of the original problem (P) are minimizers for the Lagrangian.

However, when multiple optimal solutions exist for the Lagrangian, it is possible that *not all of them are optimal for the original problem.* As a simple example, consider the linear program

$$\left.\begin{array}{ll} \min\limits_x & 2x_2 \\ \text{s.t.} & -x_1 - x_2 \le 0 \\ & x_1 - x_2 \le 0 \end{array}\right\} \quad \longrightarrow \quad \mathcal{L}(x;\lambda) = 2x_2 + \lambda_1(-x_1 - x_2) + \lambda_2(x_1 - x_2).$$

The point $x^* = (0, 0)$ is optimal for the problem, and achieves the optimal value 0 of the objective. It admits the unique choice of Lagrange multipliers $\lambda^* = (1, 1)$. For that choice, however, the Lagrangian becomes

$$\mathcal{L}(x; \lambda^*) = 0.$$

Indeed, the value of the linear program is 0, but the Lagrangian has infinite minimizers and so we cannot extract the optimal solution x^* just by looking at the minimizers of $\mathcal{L}$.

L8.3 Point of view #3: A natural dual problem

Theorem L8.1 guarantees that if (P) admits an optimal solution, then there exist penalization coefficients λ^*, μ^* such that $\min_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda^*, \mu^*)$ is exactly equal to the value of the original problem. This shows that

$$\sup_{\lambda \in \mathbb{R}^r_{\ge 0}, \mu \in \mathbb{R}^s} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda, \mu) \ge \text{value}(P).$$

As it turns out, the reverse inequality is also true, and in fact it holds much more generally.

Theorem L8.2 (Weak duality)**.** For any choice of penalization coefficients $\lambda \in \mathbb{R}^r_{\ge 0}, \mu \in \mathbb{R}^s$,

$$\inf_x \mathcal{L}(x; \lambda, \mu) \le f(\tilde{x}) \qquad \forall \tilde{x} \in \Omega.$$

In fact, the inequality holds for any minimization problem with functional constraints —that is, even ignoring the requirement that g_j is convex, that h_i is affine, and any constraint qualification.

As a direct consequence, if (P) admits an optimal solution, then

$$\inf_x \mathcal{L}(x; \lambda, \mu) \le \text{value}(P).$$

Proof. Let $\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s$ be arbitrary, and $\tilde{x}$ be any point feasible for (P). Clearly,

$$\begin{aligned}\inf_x \mathcal{L}(x;\lambda,\mu) &\leq \mathcal{L}(\tilde{x};\lambda,\mu)\\ &= f(\tilde{x}) + \sum_{j=1}^{r} \lambda_j g_j(\tilde{x}) + \sum_{i=1}^{s} \mu_i h_i(\tilde{x})\\ &\leq f(\tilde{x}),\end{aligned}$$

where we used the fact that $h_i(\tilde{x}) = 0$ and $g_j(\tilde{x}) \leq 0$ for all $i \in \{1,...,s\}$ and $j \in \{1,...,r\}$ since $\tilde{x}$ is feasible for (P). □

The weak duality theorem Theorem L8.2 shows that minimizing the Lagrangian $\mathcal{L}$ with respect to x with any choice of penalization coefficients yields a *lower bound* on the value of (P); this fact is sometimes useful to come up with bounds for the optimal value of a problem.

When taken together, Theorem L8.1 and Theorem L8.2 imply the following corollary.

Corollary L8.1 (Strong duality)**.** If (P) admits a minimizer, the optimization problem

$$(D) := \begin{cases} \max\limits_{\mu,\lambda} & \inf_{x\in\mathbb{R}^n} \mathcal{L}(x;\lambda,\mu) \\ \text{s.t.} & \lambda \in \mathbb{R}^r_{\geq 0} \\ & \mu \in \mathbb{R}^s \end{cases}$$

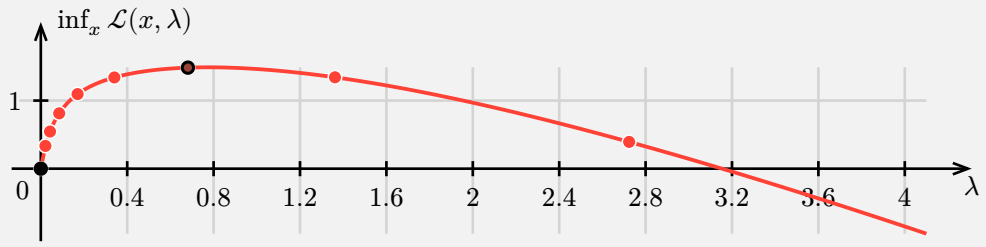
admits an optimal solution λ^*, μ^*, and matches the value of the original problem (P).

The problem (D) is called the *(Lagrangian) dual problem of* (P).

Example L8.2. The dual problem corresponding to the constrained optimization problem in Example L8.1 is by definition

$$\begin{cases} \max\limits_{\lambda} & \inf_{x\in\mathbb{R}^n} \mathcal{L}(x;\lambda) \\ \text{s.t.} & \lambda \geq 0. \end{cases}$$

The plot below shows the dual objective $\inf_{x\in\mathbb{R}^n} \mathcal{L}(x;\lambda)$.



As a final remark, we point out that whenever (P) has a solution, the value at optimality satisfies

$$\text{value}(P) = \min_{x\in\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^r_{\geq 0},\mu\in\mathbb{R}^s} \mathcal{L}(x;\lambda,\mu),$$

since it is immediate to check that given x,

$$\sup_{\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s} \mathcal{L}(x; \lambda, \mu) = \begin{cases} f(x) & \text{if } h_i(x) = 0, g_j(x) \leq 0 \text{ for all } i, j \\ +\infty & \text{otherwise.} \end{cases}$$

Hence, the strong duality theorem (Corollary L8.1) can be quite succinctly stated as follows: if (P) has an optimal solution, then

$$\underbrace{\max_{\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda, \mu)}_{\text{value}(D)} = \underbrace{\min_{x \in \mathbb{R}^n} \sup_{\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s} \mathcal{L}(x; \lambda, \mu)}_{\text{value}(P)}.$$

L8.4 Examples of dual problems

In general, the dual objective function $\inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda, \mu)$ might not be a nice, closed-form expression. However, in some cases we know that we can massage it into a nice, tractable form. We show two important examples below.

L8.4.1 Example #1: Dual of a linear program

As a sanity check, we can verify that the Lagrangian dual problem to a linear program recovers the usual notion of linear programming duality. As a reminder, a standalone proof of linear programming duality was given in Lecture 3 as a direct consequence of the characterization of the normal cone to the intersection of halfspaces.

Consider the primal linear program

$$\begin{aligned} \min_x \ & c^\top x \\ \text{s.t.} \ & Ax \leq b, \end{aligned}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. The Lagrangian function is given by

$$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m_{\geq 0}, \qquad \mathcal{L}(x; \lambda) = c^\top x + \lambda^\top (Ax - b).$$

The dual problem in this case is therefore

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda) &= \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \inf_{x \in \mathbb{R}^n} c^\top x + \lambda^\top (Ax - b) \\ &= \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \left\{ -\lambda^\top b + \inf_{x \in \mathbb{R}^n} (A^\top \lambda + c)^\top x \right\}. \end{aligned}$$

If a value of λ such that $A^\top \lambda + c \neq 0$ is picked, the inner infimum is $-\infty$. Hence, the only interesting choices of $\lambda \geq 0$ are those such that $A^\top \lambda + c = 0$, leading to the equivalent formulation of the dual problem as

$$\begin{aligned} \max_\lambda \ & -\lambda^\top b \\ \text{s.t.} \ & A^\top \lambda = -c \\ & \lambda \geq 0, \end{aligned}$$

as expected.

L8.4.2 Example #2: Dual of a convex quadratic program

As a second example, we can investigate the dual problem of a convex quadratic program

$$\begin{aligned} \min_x \ & \tfrac{1}{2} x^\top Q x + c^\top x \\ \text{s.t.} \ & Ax \leq b \end{aligned}$$

where $Q \in \mathbb{R}^{n \times n}$ is symmetric positive definite, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$.

The Lagrangian function is given by

$$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m_{\geq 0}, \qquad \mathcal{L}(x; \lambda) = \frac{1}{2} x^\top Q x + c^\top x + \lambda^\top (Ax - b).$$

The dual problem in this case is therefore

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda) &= \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \inf_{x \in \mathbb{R}^n} \frac{1}{2} x^\top Q x + c^\top x + \lambda^\top (Ax - b) \\ &= \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \left\{ -\lambda^\top b + \inf_{x \in \mathbb{R}^n} \frac{1}{2} x^\top Q x + (c + A^\top \lambda)^\top x \right\}. \end{aligned}$$

The inner infimum is bounded from below if and only if it has a minimizer, which is the case if and only if there exists a point $x \in \mathbb{R}^n$ that satisfies the first-order optimality conditions, *i.e.*, if there exists a point $x \in \mathbb{R}^n$ such that

$$0 = \nabla_x \left(\frac{1}{2} x^\top Q x + c^\top x + \lambda^\top (b - Ax) \right) = Qx + c + A^\top \lambda \qquad \iff \qquad x = -Q^{-1}(c + A^\top \lambda),$$

where we used the fact that a positive definite matrix Q is invertible. Substituting this back into the expression for the dual problem, we obtain

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m_{\geq 0}} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda) &= \max_{\lambda \in \mathbb{R}^m_{\geq 0}} -\lambda^\top b - \frac{1}{2} (c + A^\top \lambda)^\top Q^{-1} (c + A^\top \lambda) \\ &= \max_{\lambda \in \mathbb{R}^m_{\geq 0}} -\frac{1}{2} \lambda^\top (A Q^{-1} A^\top) \lambda - (b + A Q^{-1} c)^\top \lambda - \frac{1}{2} c^\top Q^{-1} c. \end{aligned}$$

In other words, the dual problem is

$$\begin{aligned} \max_\lambda \ & -\tfrac{1}{2} \lambda^\top (A Q^{-1} A^\top) \lambda - (b + A Q^{-1} c)^\top \lambda - \tfrac{1}{2} c^\top Q^{-1} c \\ \text{s.t.} \ & \lambda \geq 0, \end{aligned}$$

which is another quadratic problem.

L8.5 Failure of duality

Lagrangian duality is nothing but a reflection of the characterization of the normal cone to the intersection of the convex constraints, which is given by the KKT idea of linearizing the feasible set. As we discussed in Lecture 7, such a linearization usually gives an expression for the normal cone to the original feasible set, but sometimes it fails to do so. Unsurprisingly, when constraint qualifications fail, the Lagrangian dual problem and the original problem might not yield the same value.

When the KKT conditions are not necessary, the only result discussed today that keeps holding is weak duality (Theorem L8.2), which guarantees that

$$\sup_{\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \lambda, \mu) \leq \text{value}(P).$$

In particular, it is totally possible that $\sup_{\lambda \in \mathbb{R}^r_{\geq 0}, \mu \in \mathbb{R}^s} \inf_x \mathcal{L}(x; \lambda, \mu)$ is strictly lower than $f(x)$. When this happens, the Lagrangian dual problem is said to have a *duality gap*.

Changelog

- Mar 4, 2025: Fixed typo in domain of lambda, mu in Theorem L8.1.
- Mar 6, 2025: Fixed typo in dual of quadratic problem.