Optimization Oct 30, Nov 1, Nov 4, 2024

Lectures 10-12: Nonlinear and Convex Optimization Problems

Dan A. Iancu *Please report any typos to* dan.iancu@stanford.edu

1 Constructing the Dual

Paralleling our developments for linear optimization,[1] the starting point will be the following (primal) convex optimization problem:

Primal Problem

$$(\mathscr{P}) \quad \min_x \quad f_0(x) \\ f_i(x) \leq 0, \quad i = 1, \ldots, m \\ x \in X. \tag{1}$$

We henceforth call this the **primal problem** and concisely refer to as problem $(\mathscr{P})$.

Recall that in a convex optimization problem, the relevant domain $X \subseteq \mathbb{R}^n$ on which functions are defined must be a **convex set** and the functions $f_0, f_1, \ldots, f_m$ are real-valued convex functions on X. Note that our formulation does not include any equality constraints. In convex optimization, we can allow **equality constraints only involving linear functions**, which without loss of generality, can be included in the definition of the convex domain X. So if there are equality constraints, say $Ax = b$, you can think of the set X as the affine subspace of $\mathbb{R}^n$ corresponding to these equality constraints, $X = \{x \in \mathbb{R}^n : Ax = b\}$.

Some of the results that we are going to state will make reference to the **interior** of X. A point x **is in the interior of** X if $B(x, r) := \{y : \|y - x\| \leq r\} \subset X$ for some $r > 0$. However, in many cases it is helpful to work with sets X that are not full-dimensional (for instance, in the example above where X includes equality constraints). For that purpose, we need to define the following definition.

Definition 1 (Relative Interior)**.** *The relative interior of a set X is:*

$$\text{rel int}(X) := \big\{x \in X \, : \, \exists\, r > 0 \text{ so that } B(x, r) \cap \text{aff}(X) \subseteq X\big\}. \tag{2}$$

Recall that $\text{aff}(X)$ is the **affine hull** of X, i.e., the set of all affine combinations of points in X, $\text{aff}(X) := \{\theta_1 x_1 + \cdots + \theta_k x_k \, : \, x_i \in X, \, \sum_{i=1}^k \theta_i = 1\big\}$. In words, the relative interior of X is the interior defined relative to the affine hull of X. This gives a proper notion of the interior even for sets that are not full-dimensional.

What is the relative interior of the following sets?

[1] Our discussion here is inspired by the lecture notes Ben-Tal and Nemirovski (2023) and by the treatment in the books "Convex Optimization" (Boyd and Vandenberghe, 2004) and "Convex Optimization Theory" (Bertsekas, 2009).

- $\{(x, y) \in \mathbb{R}^2 \mid (x, y) \in [0, 1]^2\}$
- $\{(x, y) \in \mathbb{R}^2 \mid x + y = 1, x \geq 0, y \geq 0\}$
- $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$

Throughout, we assume that the relative interior of X is non-empty. We also assume that an optimal solution to $(\mathscr{P})$ exists and denote it by $x^\star$, and we let $p^\star = f_0(x^\star)$ denote the optimal value in $(\mathscr{P})$.

Mirroring our developments in linear optimization, we want to examine questions like:

1. For x feasible for $(\mathscr{P})$, how to quantify the optimality gap $f_0(x) - p^\star$?
2. How to certify that a specific $x^\star$ is **optimal** in $(\mathscr{P})$?

Duality theory will yet again help us answer such questions. We will formulate the dual problem as a lower bound on the primal, discuss briefly weak duality, and derive sufficient conditions under which strong duality holds.

To construct **lower bounds** on the optimal value $p^\star$ of $(\mathscr{P})$, let us define the **Lagrangean function**. For any $\lambda \geq 0$ (we will be using λ to denote dual variables, to avoid confusions with p), consider:

$$\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x). \tag{3}$$

By construction, it can be immediately seen that

$$\mathcal{L}(x, \lambda) \leq f_0(x), \text{ for any } x \text{ feasible in } (\mathscr{P}), \tag{4}$$

so $\mathcal{L}(x, \lambda)$ is a lower bound for $f(x)$. To derive a lower bound on the optimal value $p^\star$ in problem $(\mathscr{P})$, we can minimize $\mathcal{L}(x, \lambda)$ over $x \in X$. Therefore, let us define:

$$g(\lambda) := \inf_{x \in X} \mathcal{L}(x, \lambda). \tag{5}$$

We can immediately infer that $g(\lambda)$ is a valid lower bound, $g(\lambda) \leq p^\star$, and therefore it is natural to consider the problem of finding the best lower bound:

Dual Problem

$$(\mathscr{D}) \quad \sup_{\lambda \geq 0} g(\lambda). \tag{6}$$

Just like in linear optimization, this problem is called **the dual of the primal problem** $(\mathscr{P})$ and for conciseness, we also refer to it as problem $(\mathscr{D})$. Note that the dual problem is a convex optimization problem because the function $g(\lambda)$ is concave. In fact, this would be true even if the primal problem were non-convex!

The following **weak duality** result is immediate.

Theorem 1 (Weak Duality)**.** *If x is feasible for $(\mathscr{P})$ and $\lambda \geq 0$, then:*

$$g(\lambda) \leq f(x).$$

In particular, $d^\star \leq p^\star$.

We would obviously like to develop a **strong duality** result that would tell us that $p^\star = d^\star$. However, the situation with (nonlinear) convex optimization is unfortunately more subtle than with linear optimization: even when the primal $(\mathscr{P})$ has a finite optimal solution, there may be a non-zero duality gap. The following example shows how this can arise.

Example 1 (Non-zero duality gap)**.** *Consider the convex problem*

$$\begin{aligned} \underset{(x,y)\in X}{\text{minimize}}\ & e^{-x} \\ & x^2/y \leq 0 \end{aligned}$$

with variables x, y and domain $X = \{(x, y) \mid y \geq 1\}$. We have $p^\star = 1$. The Lagrangian is $L(x, y, \lambda) = e^{-x} + \lambda x^2/y$, and the dual function is

$$g(\lambda) = \inf_{x,y\geq 1} \left(e^{-x} + \lambda \frac{x^2}{y} \right) = \begin{cases} 0 & \lambda \geq 0, \\ -\infty & \lambda < 0, \end{cases}$$

so we can write the dual problem as

$$d^\star = \max_{\lambda \geq 0} 0$$

with optimal value $d^\star = 0$. The optimal duality gap is $p^\star - d^\star = 1$.

Moreover, even when the primal problem admits a (finite) optimal solution, the dual may not necessarily admit an optimal solution, as in the following example.

Example 2 (No dual optimal solution)**.** *Consider the optimization problem:*

$$\begin{aligned} \underset{x\in\mathbb{R}}{\text{minimize}}\ & x \\ & x^2 \geq 0 \end{aligned}$$

The optimal solution is trivially $x^\star = 0$, so $p^\star = 0$. The dual function is:

$$g(\lambda) = \inf_{x\in\mathbb{R}} \left\{ x + \lambda x^2 \right\} = \begin{cases} -\frac{1}{4\lambda} & \text{if } \lambda > 0, \\ -\infty & \text{if } \lambda \leq 0. \end{cases}$$

Thus $d^\star = 0$ (with $\lambda \to \infty$) and $p^\star = d^\star$, but the dual does not admit an optimal solution.

Thus, in the subsequent developments, our main goal is to provide sufficient conditions under which **strong duality** holds. These conditions are sometimes called **constraint qualifications** and several variations exist. Perhaps the most prominent and useful of these is **Slater's condition**, which we define next.

Definition 2 (Slater Condition)**.** *Let $X \subseteq \mathbb{R}^n$ and $f_1, \ldots, f_m$ be real-valued functions on X. We say that f_i satisfy the Slater condition on X if there exists $x \in \operatorname{rel\,int}(X)$ such that*

$$f_j(x) < 0, \quad j = 1, \ldots, m.$$

This condition essentially asks that there exists a point x that is **strictly feasible** because the inequality constraints hold strictly. The condition can actually be further refined if some of the functions f_i are affine: for instance, if $f_1, \ldots, f_r$ are affine, then we only require $f_i(x) \leq 0$ for $i = 1, \ldots, r$ and $f_i(x) < 0$ for $i = r+1, \ldots, m$, i.e., the strict inequality is only required for non-linear functions.

As we will see, Slater's condition implies that strong duality holds and also that the dual optimal value is attained, i.e., that there exists a feasible $x^\star$ in the primal and a $\lambda^\star \geq 0$ such that $f(x^\star) = g(\lambda^\star)$.

1.1 A Geometric View of Duality

Before proving our main strong duality result, we provide a natural geometric interpretation of the dual construction that will also make the proof more clear. The construction is depicted in Figure 1. To introduce it, assume that there is only one inequality constraint in $(\mathscr{P})$ (i.e., $m = 1$), and let

$$\mathcal{G} := \{(u, t) \in \mathbb{R}^2 : \exists x \in \mathbb{R}^n,\ t = f_0(x),\ u = f_1(x)\}$$

denote the set of values taken by the objective and constraint over the set $x \in X$. The optimal value of the primal is then expressed as:

$$p^\star = \inf\big\{t \ :\ (u, t) \in \mathcal{G},\, u \leq 0\big\},$$

and we can see that $(\mathscr{P})$ is feasible if and only if $\mathcal{G}$ intersects the left-half plane. Note that when evaluating the dual function, we are minimizing the affine function $\lambda \cdot u + 1 \cdot t$ over $(u, t) \in \mathcal{G}$, so we can write:

$$g(\lambda) = \min_{u,t} \big\{(\lambda, 1)^\intercal (u, t) \ :\ (u, t) \in \mathcal{A}\big\},$$

If the minimum is finite, then the inequality $(\lambda, 1)^\intercal (u, t) \geq g(\lambda)$ defines a supporting hyperplane for the set $\mathcal{G}$ and the intersection between this hyperplane and the vertical axis $u = 0$ gives the value of the dual, $g(\lambda)$.

Nothing would change if we replaced $\mathcal{G}$ by its "upper extension" $\mathcal{A} = \mathcal{G} + \mathbb{R}^2_+ \equiv \big\{(u, t) : \exists\, x \in X,\, t \geq f_0(x), u \geq f_1(x)\big\}$ because $\mathcal{A}$ includes all the points in $\mathcal{G}$ and points that are strictly "worse" for the optimization that defines the dual value. In this case, because $(0, p^\star) \in \operatorname{bd}(\mathcal{A})$, we have $p^\star \geq g(\lambda)$ for any $\lambda \geq 0$, so weak duality always holds.

If the problem is convex, then the set $\mathcal{A}$ will also be convex. And if Slater's condition holds, then the interior of $\mathcal{A}$ will intersect the left half-plane, and strong duality will hold.

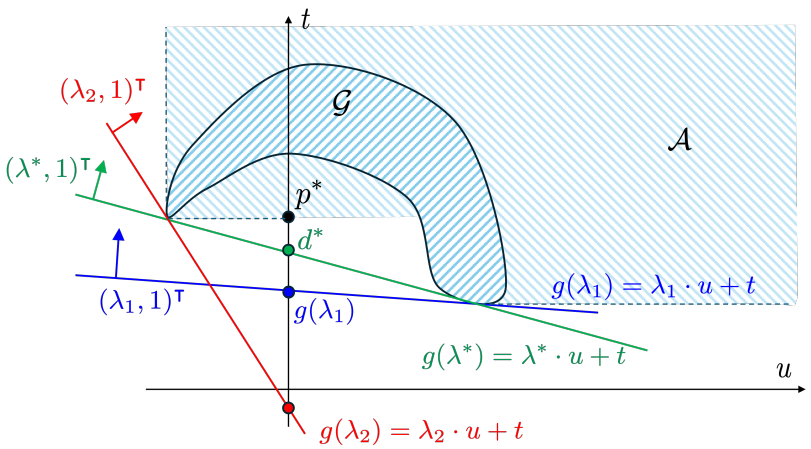

Figure 1: Geometric interpretation of the dual function and lower bound $g(\lambda) \leq p^\star$ for problem $(\mathscr{P})$ with one inequality constraint. Given $\lambda \geq 0$, to find $g(\lambda)$ we must minimize $t + \lambda \cdot u$ over $(u, t) \in \mathcal{G}$. This yields a supporting hyperplane for $\mathcal{G}$ who intersection with the vertical axis $u = 0$ yields $g(\lambda)$. Here, strong duality does not hold because the optimal dual solution $\lambda^\star$ yields a lower bound $d^\star$ so that $d^\star < p^\star$. Note that nothing would change if we replaced $\mathcal{G}$ by its upper extension $\mathcal{A} = \mathcal{G} + \mathbb{R}^2_+ \equiv \big\{(u, t) : \exists\, x \in X,\, t \geq f_0(x), u \geq f_1(x)\big\}$.

2 Strong Duality

The following result formalizes the intuition above.

Theorem 2 (Strong Duality in Convex Optimization)**.** *Let $X \subset \mathbb{R}^n$ be convex, let $f_0, f_1, \ldots, f_m$ be real-valued convex functions on X, and let $f_1, \ldots, f_m$ satisfy the Slater condition on X. Then, $p^\star = d^\star$ and the dual problem attains its optimal value.*

Proof. We adopt a proof that leverages the geometric intuition developed above. For problem $(\mathscr{P})$, let use define the upper-extension $\mathcal{A}$ as:

$$\mathcal{A} = \big\{(u, t) \in \mathbb{R}^m \times \mathbb{R} : \exists\, x \in X,\, t \geq f_0(x), u_i \geq f_i(x),\, i = 1, \ldots, m\big\}. \tag{7}$$

The set $\mathcal{A}$ is convex because it is the projection of the convex set $\{(x, u, t) : x \in X, t \geq f_0(x), u_i \geq f_i(x),\, i = 1, \ldots, m\}$ onto the (u, t) coordinates.

We next define a second convex set $\mathcal{B}$ as

$$\mathcal{B} = \{(0, s) \in \mathbb{R}^m \times \mathbb{R} \mid s < p^\star\}.$$

These sets are depicted in Figure 2. We claim that $A \cap B = \emptyset$. To see this, suppose $(u, t) \in A \cap B$. Because $(u, t) \in B$, we have $u = 0$ and $t < p^\star$. But $(u, t) \in A$, which implies that there exists an $x \in X$ with $f_i(x) \leq u_i = 0$, $i = 1, \ldots, m$ and with $f_0(x) \leq t < p^\star$, which contradicts $p^\star$ being the optimal value of $(\mathscr{P})$.

By the separating hyperplane theorem, there exists $(\lambda, \mu) \in \mathbb{R}^{m+1} \neq 0$ and b with:

$$\forall\, (u, t) \in A, \quad \lambda^\intercal u + \mu t \geq b, \tag{8a}$$

$$\forall\, (u, t) \in B, \quad \lambda^\intercal u + \mu t \leq b. \tag{8b}$$

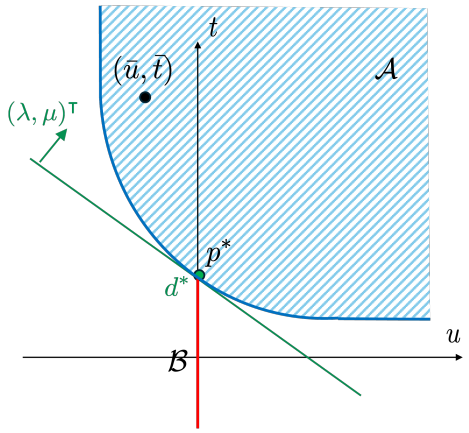

Figure 2: Illustration for the strong duality proof. The set $\mathcal{A}$ is shaded and the set $\mathcal{B}$ is the (red) vertical segment, not including the point $(0, p^\star)$. The sets are convex and do not intersect, so a separating hyperplane must exist. The Slater condition guarantees that the separating hyperplane is non-vertical, because it must separate a point $(\bar{u}, \bar{t}) = (f_1(\bar{x}), f_0(\bar{x}))$ corresponding to a Slater point $\bar{x}$.

We claim that (8a) implies that $\lambda \geq 0$ and $\mu \geq 0$. Otherwise, we would have

$$\inf_{(u,t)\in\mathcal{A}} (\lambda^\intercal u + \mu t) = -\infty$$

because the recession cone of $\mathcal{A}$ contains the rays e_i (for $i = 1, \ldots, m+1$), which would contradict (8a).

Moreover, condition (8b) simplifies to $\mu t \leq b$ for all $t < p^\star$, and hence, $\mu p^\star \leq b$. Together with (8a), we conclude that we identified a $\lambda \geq 0$ such that for any $x \in X$,

$$\mathcal{L}(x, \lambda) := \sum_{i=1}^{m} \lambda_i f_i(x) + \mu f_0(x) \geq b \geq \mu p^\star. \tag{9}$$

Case 1. Assume that $\mu > 0$. Then, we can divide inequality (9) by μ to obtain

$$\mathcal{L}(x, \lambda/\mu) \geq p^\star, \, \forall \, x \in X,$$

from which it follows that $g(\lambda/\mu) \geq p^\star$. By weak duality, $g(\lambda/\mu) \leq p^\star$, so in fact $g(\lambda/\mu) = p^\star$, which implies that strong duality holds and the dual optimum is attained.

Case 2. Now suppose $\mu = 0$. From (9), we conclude that:

$$\sum_{i=1}^{m} \lambda_i f_i(x) \geq 0, \, \forall \, x \in X.$$

Applying this to the point $\bar{x}$ that satisfies the Slater condition, we have

$$\sum_{i=1}^{m} \lambda_i f_i(\bar{x}) \geq 0.$$

Because $\bar{x}$ satisfies the Slater condition, we have $f_i(\bar{x}) < 0$ for $i = 1, \ldots, m$, which together with $\lambda \geq 0$ implies that we must have $\lambda = 0$. But this contradicts the separating hyperplane assumption that $(\lambda, \mu) \neq 0$. □

It should be noted that other versions of strong duality results are possible depending on the constraint qualification conditions. Moreover, if we are solely interested in guaranteeing strong duality and that the primal problem achieves is optimal value without insisting that the dual should achieve its optimum, other conditions are possible.

Proposition 1 (Convex Programming Duality - Existence of Primal Optimal Solutions)**.** *Assume that* $(\mathscr{P})$ *is feasible, that the convex functions* $f_i, i = 0, \ldots, m$ *are closed, and that the function*

$$F(x, 0) = \begin{cases} f_0(x) & \text{if } f_i(x) \leq 0, i = 1, \ldots, m,\ x \in X, \\ +\infty, \text{otherwise}, \end{cases}$$

has compact level sets. Then $p^\star = d^\star$ *and the set of optimal solutions of* $(\mathscr{P})$ *is nonempty and compact.*

For a proof, the interested reader can refer to Proposition 5.3.7 of Bertsekas (2009). The compactness requirement of this proposition is reasonable if either X is compact or if X is closed and f_0 has compact level sets. For instance, the latter happens if f_0 is a continuous and **coercive** function, i.e., $\lim_{\|x\| \to \infty} f_0(x) = +\infty$. However, this proposition does not guarantee the existence of a dual optimal solution; Example 2 exactly illustrates this point.

2.1 Explicit Equality Constraints

The developments above assumed that any equality constraints are included in the definition of the set X. In applications, it is often useful to write out the equality constraints explicitly, so we now briefly extend the theory above to accommodate this. Consider a more general convex optimization problem:

$$\begin{aligned} \text{minimize } & f_0(x) \\ \text{subject to } & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & Ax = b \\ & x \in X. \end{aligned} \tag{10}$$

where $f_i, i = 0, \ldots, m$ are convex and without loss of generality, we assume the matrix $A \in \mathbb{R}^{p \times n}$ has rank p. We define the Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with problem (10) as

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^\intercal (Ax - b),$$

where we use $\nu \in \mathbb{R}^p$ to denote the Lagrange multipliers associated with the linear constraints $Ax = b$. The dual objective can be written as:

$$g(\lambda, \nu) := \inf_{x \in X} \mathcal{L}(x, \lambda, \nu),$$

and the dual problem becomes:

$$\begin{aligned} &\text{maximize } g(\lambda, \nu) \\ &\text{subject to } \lambda \geq 0. \end{aligned} \tag{11}$$

The main thing to note is that the Lagrange multipliers ν associated with equality constraints are not sign-constrained, so ν can be both positive and negative values in the dual problem. This is completely consistent with the developments in linear optimization and should not come as a surprise.

2.2 Nonlinear Farkas Lemma

The previous developments also highlight that non-linear version of the Farkas Lemma is readily available. Specifically, we have the following result.

Proposition 2 (Nonlinear Farkas Lemma)**.** *Let $X \subset \mathbb{R}^n$ be convex, let $f_0, f_1, \ldots, f_m$ be real-valued convex functions on X, and assume $f_1, \ldots, f_m$ satisfy the Slater condition on X. Then, the following system of inequalities has a solution*

$$\exists\, x \;:\; f_0(x) < z, \quad f_j(x) \leq 0,\ j = 1, \ldots, m, \quad x \in X, \tag{12}$$

if and only if the following system has no solution:

$$\exists\, \lambda \;:\; \inf_{x \in X} \left[f(x) + \sum_{j=1}^{m} \lambda_j f_j(x) \right] \geq z, \quad \lambda_j \geq 0,\ j = 1, \ldots, m.$$

The proof essentially mirrors the arguments used in the strong duality proof. As in the linear case, the Farkas Lemma provides a very powerful certificate of feasibility and its role is essentially equivalent to strong duality.

3 Applications of Convex Duality

3.1 Minimum Euclidean Distance Problem

Consider the problem of finding the minimum Euclidean distance from a given point y to an affine set $\{z : Az = \tilde{b}\}$. This problem can be written as:

$$\min_z \ \|z - y\|_2^2 : Az = \tilde{b},$$

where $A \in \mathbb{R}^{p \times n}$, $\tilde{b} \in \mathbb{R}^p$ and we assume that A has rank p. With a change of variables $x := z - y$ and by letting $b := \tilde{b} - Ay$, this can be reformulated as:

$$\min_x \ x^\intercal x : Ax = b.$$

The Lagrangian for this problem is

$$L(x, \nu) = x^\intercal x + \nu^\intercal (Ax - b),$$

where $\nu \in \mathbb{R}^p$ is the dual variable associated with the constraint $Ax = b$. The dual function is given by $g(\nu) = \inf_x L(x, \nu)$. Since $L(x, \nu)$ is a convex quadratic function of x, we can find the minimizing x from the optimality condition

$$\nabla_x L(x, \nu) = 2x + A^\intercal \nu = 0 \quad \Leftrightarrow \quad x = -\frac{1}{2} A^\intercal \nu. \tag{13}$$

Therefore, the dual function is

$$g(\nu) = L\left(-\frac{1}{2} A^\intercal \nu, \nu\right) = -\frac{1}{4} \nu^\intercal A A^\intercal \nu - b^\intercal \nu,$$

which is a concave quadratic function with domain $\mathbb{R}^p$.

Note that the primal problem trivially satisfies the Slater condition (provided that it is feasible). Therefore, $p^\star = d^\star$. Moreover, to find the optimal value of the dual, which is an unconstrained convex optimization problem, we can simply set the gradient equal to zero, which leads to:

$$-\frac{1}{2} A A^\intercal \nu = b.$$

Because A is assumed to have rank p, the matrix $AA^\intercal$ is an invertible $p \times p$ matrix, so we obtain the optimal dual solution $\nu^\star = -2(AA^\intercal)^{-1} b$. Moreover, the optimal value of the dual (and, by strong duality, the primal) is:

$$p^\star = d^\star = g(\nu^\star) = b^\intercal (AA^T)^{-1} b.$$

Moreover, this also implies from (13) that the optimal primal solution is $x^\star = -\frac{1}{2} A^\intercal \nu^\star = A^\intercal (AA^\intercal)^{-1} b$! (Note that this point is feasible in the primal and it achieves the optimal value $p^\star$!)

3.2 Quadratic Programs

The example above is a special instance of the problem of minimizing a quadratic function subject to linear constraints – a problem that is generically called a **quadratic program (QP)**. When there are only equality constraints, this problem can be reformulated as:

$$\min_x \; x^\intercal Q x : Ax = b,$$

where $Q \succ 0$ is a positive definite matrix. We can follow an identical line of reasoning as above to form the Lagrangian and derive the dual and the optimal solution.

We now consider the case with **inequality** constraints, which we write more generally:

$$\underset{x}{\text{minimize}} \; \frac{1}{2} x^\intercal Q x + c^\intercal x \tag{14}$$

$$Ax \leq b \tag{15}$$

where $Q \in \mathbb{R}^{n \times n}$ and $Q \succ 0$. The Langragian function is:

$$\mathcal{L}(x, \lambda) = \frac{1}{2} x^\intercal Q x + c^\intercal x + \lambda^\intercal (Ax - b)$$

and the dual function is:

$$g(\lambda) = -\lambda^\intercal b + \inf_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} x^\intercal Q x + c^\intercal x + \lambda^\intercal A x \right\}.$$

By taking the gradient, we can see that the infimum is achieved at $x = -Q^{-1}(c + A^\intercal \lambda)$. Therefore, the fual function becomes:

$$g(\lambda) = -\frac{1}{2} \lambda^\intercal A Q^{-1} A^\intercal \lambda - \lambda^\intercal (b + A Q^{-1} c) - \frac{1}{2} c^\intercal Q^{-1} c.$$

Assuming the primal is feasible (i.e., $Ax \leq b$ is feasible), strong duality always holds for this problem so we can solve either the primal or the dual. The dual problem entails maximizing a concave quadratic function subject to the constraints $\lambda \geq 0$ and in this case, it may be simpler to solve than the primal because it is very easy to project onto the feasible set, so we can apply a gradient descent algorithm with a correction step. (More on that later in the course!)

3.3 Quadratically Constrained Quadratic Programs

Here, we consider a **quadratically constrained quadratic program (QCQP)**:

$$\begin{aligned} &\text{minimize } \frac{1}{2} x^\intercal P_0 x + q_0^\intercal x + r_0 \\ &\text{subject to } \frac{1}{2} x^\intercal P_i x + q_i^\intercal x + r_i \leq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{16}$$

where $P_0 \succ 0$ is an $n \times n$ positive definite matrix and $P_i \succeq 0$ are $n \times n$ positive semidefinite matrices, $i = 1, \ldots, m$. The Lagrangian is:

$$\mathcal{L}(x, \lambda) = \frac{1}{2} x^\intercal P(\lambda) x + q(\lambda)^\intercal x + r(\lambda),$$

where

$$P(\lambda) = P_0 + \sum_{i=1}^m \lambda_i P_i, \quad q(\lambda) = q_0 + \sum_{i=1}^m \lambda_i q_i, \quad r(\lambda) = r_0 + \sum_{i=1}^m \lambda_i r_i.$$

It is possible to derive an expression for $g(\lambda)$ for general λ, but it is rather complicated. However, because $\lambda \geq 0$ in our case, we have $P(\lambda) \succ 0$ and therefore:

$$g(\lambda) = \inf_x L(x, \lambda) = -\frac{1}{2} q(\lambda)^\intercal P(\lambda)^{-1} q(\lambda) + r(\lambda).$$

We can therefore express the dual problem as

$$\begin{aligned} &\text{maximize } -\frac{1}{2} q(\lambda)^\intercal P(\lambda)^{-1} q(\lambda) + r(\lambda) \\ &\text{subject to } \lambda \geq 0. \end{aligned} \tag{17}$$

The Slater condition states that strong duality between (16) and (17) holds if the quadratic inequality constraints are strictly feasible, i.e., there exists an x with

$$\frac{1}{2} x^\intercal P_i x + q_i^\intercal x + r_i < 0, \quad i = 1, \ldots, m.$$

3.3.1 A nonconvex quadratic problem with strong duality

A special QCQP instance also provides one of the rare examples where strong duality holds for a nonconvex problem. Specifically, consider the problem of minimizing a nonconvex quadratic function over the unit ball,

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} A x + 2b^\mathsf{T} x \\ \text{subject to} & x^\mathsf{T} x \leq 1, \end{array} \tag{18}$$

where $A \in S^n$ is a symmetric matrix but $A \not\succeq 0$ (i.e., A is not positive semidefinite) and $b \in \mathbb{R}^n$. Because $A \not\succeq 0$, this is not a convex problem. This problem is sometimes called the trust region problem and arises from minimizing a second-order approximation of a (non-convex) function over the unit ball, which is the region in which the approximation is assumed to be approximately valid.

The Lagrangian is

$$\mathcal{L}(x, \lambda) = x^\mathsf{T} A x + 2b^\mathsf{T} x + \lambda(x^\mathsf{T} x - 1) = x^\mathsf{T}(A + \lambda I)x + 2b^\mathsf{T} x - \lambda,$$

so the dual function is given by

$$g(\lambda) = \begin{cases} -b^\mathsf{T}(A + \lambda I)^\dagger b - \lambda & A + \lambda I \succeq 0,\ b \in \mathcal{R}(A + \lambda I), \\ -\infty & \text{otherwise}, \end{cases}$$

where $M^\dagger$ is the (Moore-Penrose) pseudo-inverse of M, i.e., $(M^\mathsf{T} M)^{-1} M^\mathsf{T}$ for a full-rank matrix M. The Lagrange dual problem is thus

$$\begin{array}{ll} \text{maximize} & -b^\mathsf{T}(A + \lambda I)^\dagger b - \lambda \\ \text{subject to} & A + \lambda I \succeq 0,\ b \in \mathcal{R}(A + \lambda I), \end{array} \tag{19}$$

with variable $\lambda \in \mathbb{R}$. Although it is not readily obvious from the expression above, this is a convex optimization problem. In fact, it is readily solved since it can be expressed as

$$\begin{array}{ll} \text{maximize} & -\displaystyle\sum_{i=1}^{n} \frac{(q_i^\mathsf{T} b)^2}{\lambda_i + \lambda} - \lambda \\ \text{subject to} & \lambda \geq -\lambda_{\min}(A), \end{array}$$

where λ_i and q_i are the eigenvalues and corresponding (orthonormal) eigenvectors of A, and we interpret $(q_i^\mathsf{T} b)^2/0$ as 0 if $q_i^\mathsf{T} b = 0$ and as ∞ otherwise.

The Slater condition is trivially satisfied in problem (18) and we actually have zero optimal duality gap: the optimal values of (18) and (19) are always the same. In fact, a more general result holds: strong duality holds for any optimization problem with quadratic objective and a single quadratic inequality constraint, provided Slater's condition holds. And extensions are also possible for two-sided quadratic constraints, i.e., constraints of the form $\ell \leq x^\mathsf{T} P x \leq u$, provided that the matrices P and A are simultaneously diagonalizable (see Ben-Tal and Teboulle (1996) for details).

3.4 Entropy Maximization

Consider the problem of maximizing the entropy in a distribution. The entropy is given by $-\sum_{i=1}^n x_i \log x_i$, so this problem can be written as the following minimization:

$$\begin{aligned} \text{minimize } & f_0(x) = \sum_{i=1}^n x_i \log x_i \\ \text{subject to } & Ax \leq b, \\ & 1^\intercal x = 1. \end{aligned} \tag{20}$$

Here, x is typically a distribution and the constraints $x \geq 0$ are assumed to be embedded in the constraints $Ax \leq b$ (so the matrix $A \in \mathbb{R}^{m \times n}$ has rank $n < m$.) The problem is a convex optimization problem because the functions $x \log x$ are convex on the domain $x > 0$ (take the second derivative for a proof!) With $\lambda \in \mathbb{R}^m$ denoting the dual variables for the inequality constraints and $\nu \in \mathbb{R}$ denoting the dual variable for the equality constraint, the Lagrangian can be written as:

$$\mathcal{L}(x, \lambda, \nu) = \sum_{i=1}^n x_i \log x_i + \lambda^\intercal (Ax - b) + \nu(1^\intercal x - 1). \tag{21}$$

The gradient with respect to the primal variable x_i is:

$$\frac{\partial \mathcal{L}}{\partial x_i} = \log x_i + 1 + \lambda^\intercal A_i + \nu, \tag{22}$$

where A_i is the i-th column of the matrix A. The first-order optimality condition yields:

$$x_i = \exp\bigl(-1 - \lambda^\intercal A_i - \nu\bigr), \tag{23}$$

and the dual function is given by:

$$g(\lambda, \nu) = -b^\intercal \lambda - \nu - \sum_{i=1}^n \exp\bigl(-A_i^\intercal \lambda - \nu - 1\bigr) = -b^\intercal \lambda - \nu - e^{-\nu - 1} \sum_{i=1}^n e^{-A_i^\intercal \lambda}.$$

3.5 Regularized Support Vector Machines

Consider a binary classification problem as shown in Figure 3. Given m data points $x_i \in \mathbb{R}^n$, each of which is associated with a label $y_i \in \{-1, 1\}$, the problem is to find a hyperplane that separates, as much as possible, the two classes.

The two classes are separable by a hyperplane $H(w, b) = \{x : w^\intercal x + b = 0\}$, where $w \in \mathbb{R}^n$, $w \neq 0$, and $b \in \mathbb{R}$, if and only if $w^\intercal x_i + b \geq 0$ for $y_i = +1$ and $w^\intercal x_i + b \leq 0$ if $y_i = -1$. Thus, the following conditions on (w, b):

$$y_i(w^\intercal x_i + b) \geq 0, \quad i = 1, \ldots, m \tag{24}$$

would ensure that the data set is separable by a linear classifier. In this case, the parameters w and b allow us to predict the label associated with a new point x, via $y = \text{sign}(w^\intercal x + b)$.

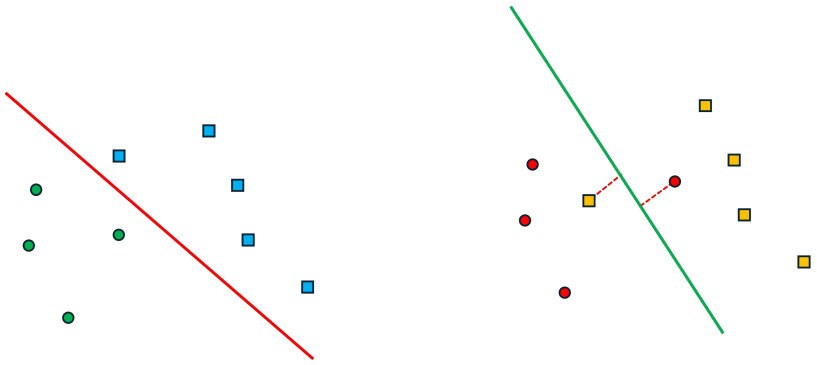

Figure 3: Two problem instances for the binary classification problem. The instance on the left is separable: we can find a hyperplane that separates the blue squares from the green circles. In contrast, the instance on the right is non-separable, so one typically seeks a hyperplane that minimizes the total errors committed. With the hinge loss, the errors correspond to the sum of the distances from the points to the hyperplane.

The feasibility problem – finding (w, b) that satisfy the above separability constraints – is an LP. If the data set is strictly separable (i.e., every inequality in (24) holds strictly), then we can rescale the constraints and transform them into

$$y_i(w^\intercal x_i + b) \geq 1, \quad i = 1, \ldots, m.$$

However, in practice the two classes may not be linearly separable. In this case, we would like find a hyperplane that minimizes the total number of classification errors. Strictly speaking, the objective function corresponding to the number of mistakes has the form:

$$\sum_{i=1}^{m} \psi(y_i(w^\intercal x_i + b)),$$

where $\psi(t) = 1$ if $t < 0$, and 0 otherwise. Unfortunately, this is non-convex and rather hard to minimize (it would require solving an IP!) As an alternative, we can replace the objective with an upper bound formed by using **the hinge function**, $h(t) = (1-t)_+ = \max(0, 1-t)$. Our problem becomes one of minimizing a piecewise linear "loss" function:

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(w^\intercal x_i + b))_+.$$

At optimality, the value of the loss function can be read from Figure 3: it equals the sum of the lengths of the dotted lines from data points that are wrongly classified to the hyperplane.

In practice, we often want to control the robustness of the resulting classifier and also to guarantee that an optimal classifier is unique. It turns out that these objectives can be achieved by solving the following regularized problem:

$$\min_{w,b} \; C \cdot \sum_{i=1}^{m} (1 - y_i(w^\intercal x_i + b))_+ + \frac{1}{2}\|w\|_2^2,$$

where $C > 0$ is a parameter that controls the trade-off between robustness and performance on the training set (a greater C encourages performance at the expense of robustness). This problem can be written as a QP, by introducing slack variables:

$$\min_{w,b,v} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^m v_i \quad : \quad v \geq 0, \ y_i(w^\intercal x_i + b) \geq 1 - v_i, \ i = 1, \ldots, m,$$

or, more compactly:

$$\min_{w,b,v} \ \frac{1}{2}\|w\|_2^2 + C1^\intercal v \quad : \quad v \geq 0, \ v + Z^\intercal w + by \geq 1,$$

where $Z^\intercal \in \mathbb{R}^{m\times n}$ is the matrix with rows given by $y_i \cdot x_i^\intercal$.

The corresponding Lagrangian is

$$\mathcal{L}(w, b, \lambda, \mu) = \frac{1}{2}\|w\|_2^2 + Cv^\intercal 1 + \lambda^\intercal(1 - v - Z^\intercal w - by) - \mu^\intercal v,$$

where $\mu \in \mathbb{R}^m$ corresponds to the sign constraints on v. The dual function is given by

$$g(\lambda, \mu) = \min_{w,b} \mathcal{L}(w, b, \lambda, \mu).$$

We can readily solve for w by taking derivatives, which leads to $w(\lambda, \mu) = Z\lambda$. Taking derivatives with respect to v yields the constraint $C \cdot 1 = \lambda + \mu$, while taking derivatives with respect to b leads to the dual constraint $\lambda^\intercal y = 0$. We obtain

$$g(\lambda, \mu) = \begin{cases} \lambda^\intercal 1 - \frac{1}{2}\|Z\lambda\|_2^2 & \text{if } \lambda^\intercal y = 0, \ \lambda + \mu = C \cdot 1, \\ +\infty & \text{otherwise.} \end{cases}$$

We obtain the dual problem

$$d^\star = \max_{\lambda\geq 0, \mu \geq 0} g(\lambda, \mu) = \max_{\lambda} \lambda^\intercal 1 - \frac{1}{2}\lambda^\intercal Z^\intercal Z\lambda \quad : \quad 0 \leq \lambda \leq C \cdot 1, \ \lambda^\intercal y = 0.$$

Strong duality holds, because the primal problem is a QP (note that we can always produce an interior point with b sufficiently large). Importantly, the dual objective depends only on the so-called kernel matrix $K = Z^\intercal Z \in S_+^m$, and the dual problem involves only m variables and $m + 1$ constraints. Hence, the only dependence on the number of dimensions (features) n is via the required computation of the kernel matrix, that is, on scalar products $x_i^\intercal x_j$, $1 \leq i \leq j \leq m$. Thus, duality allows a great reduction in the computational effort, compared to solving the original QP in n variables and m constraints. This is known as the "kernel trick."

Duality also shows that the optimal value of the problem is a convex function of the kernel matrix, which allows us to understand how the results depend on the data matrix (and consider robust objectives related to that, as we will discuss later in the course).

4 Saddle-point Theory

Our previous discussion may have made it seem like the primal and dual have slightly different roles. In this section we give a different interpretation of Lagrange duality that will appear more symmetric. To simplify the discussion, we consider again only the case with inequality constraints, as in (1) (equality constraints can be readily accommodated). First note that

$$\sup_{\lambda \geq 0} L(x, \lambda) = \sup_{\lambda \geq 0} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right) = \begin{cases} f_0(x) & \text{if } f_i(x) \leq 0,\ i = 1, \ldots, m, \\ \infty & \text{otherwise.} \end{cases}$$

Indeed, if x is not feasible and $f_i(x) > 0$ for some i, then $\sup_{\lambda \geq 0} L(x, \lambda) = \infty$ by taking $\lambda_i \to \infty$. And if $f_i(x) \leq 0$, $i = 1, \ldots, m$, then the optimal choice of λ is $\lambda = 0$, and $\sup_{\lambda \geq 0} L(x, \lambda) = f_0(x)$. This means that we can express **the optimal value of the primal problem** as

$$p^\star = \inf_{x \in X} \sup_{\lambda \geq 0} L(x, \lambda).$$

By the definition of the dual function, we also have

$$d^\star = \sup_{\lambda \geq 0} \inf_{x \in X} L(x, \lambda).$$

Thus, weak duality can be expressed as the inequality:

$$\sup_{\lambda \geq 0} \inf_{x \in X} L(x, \lambda) \leq \inf_{x \in X} \sup_{\lambda \geq 0} L(x, \lambda) \tag{25}$$

whereas strong duality is equivalent to the equality:

$$\sup_{\lambda \geq 0} \inf_{x \in X} L(x, \lambda) = \inf_{x \in X} \sup_{\lambda \geq 0} L(x, \lambda). \tag{26}$$

It is worth putting these results into the context of the more general comparison of the following two optimization problems:

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) \quad \textit{versus} \quad \inf_{w \in W} \sup_{z \in Z} f(w, z). \tag{27}$$

In this context, a weak duality relation (25) holds irrespective of the properties of f and the feasible sets in question, so we have

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) \leq \inf_{w \in W} \sup_{z \in Z} f(w, z)$$

for any $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and any $W \subseteq \mathbb{R}^n$ and $Z \subseteq \mathbb{R}^m$. This general inequality is called **the max-min inequality**. Strong duality means that **the order of the minimization over x and the maximization over $\lambda \geq 0$** can be switched without affecting the result. When equality holds, i.e.,

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) = \inf_{w \in W} \sup_{z \in Z} f(w, z) \tag{28}$$

we say that f (and W and Z) satisfy the **strong max-min** property or **the saddle-point property**. We refer to a pair $w^\star \in W$, $z^\star \in Z$ as a saddle-point for f (and W and Z) if

$$f(w^\star, z) \leq f(w^\star, z^\star) \leq f(w, z^\star)$$

for all $w \in W$ and $z \in Z$. In other words, $w^\star$ minimizes $f(w, z^\star)$ (over $w \in W$) and $z^\star$ maximizes $f(w^\star, z)$ (over $z \in Z$):

$$f(w^\star, z^\star) = \inf_{w \in W} f(w, z^\star), \quad f(w^\star, z^\star) = \sup_{z \in Z} f(w^\star, z).$$

Returning to our discussion of Lagrange duality, we see that if $x^\star$ and $\lambda^\star$ are primal and dual optimal points for a problem in which strong duality obtains, they form a saddle-point for the Lagrangian. The converse is also true: If (x, λ) is a saddle-point of the Lagrangian, then x is primal optimal, λ is dual optimal, and the optimal duality gap is zero. The following result actually formalizes and proves this.

Theorem 3 (Saddle Point Optimality Condition in Convex Programming)**.** *Let $(\mathscr{P})$ be an optimization program, $\mathcal{L}(x, \lambda)$ be its Lagrangian function, and let $x^\star \in X$. Then:*

*(i) A **sufficient condition** for $x^\star$ to be an optimal solution to $(\mathscr{P})$ is the existence of the vector of Lagrange multipliers $\lambda^\star \geq 0$ such that $(x^\star, \lambda^\star)$ is a saddle point of the Lagrange function $\mathcal{L}(x, \lambda)$, i.e., satisfies:*

$$\mathcal{L}(x, \lambda^\star) \geq \mathcal{L}(x^\star, \lambda^\star) \geq \mathcal{L}(x^\star, \lambda) \quad \forall x \in X,\ \lambda \geq 0. \tag{29}$$

*(ii) If $(\mathscr{P})$ is a convex optimization problem and satisfies the Slater condition, then the above condition is also **necessary** for the optimality of $x^\star$: if $x^\star$ is optimal for $(\mathscr{P})$, then there exists $\lambda^\star \geq 0$ such that $(x^\star, \lambda^\star)$ is a saddle point of the Lagrangian function.*

Proof. (i): Assume that for a given $x^\star \in X$ there exists $\lambda^\star \geq 0$ such that (29) is satisfied. We prove that $x^\star$ is optimal for $(\mathscr{P})$. First, $x^\star$ is feasible: indeed, if $f_j(x^\star) > 0$ for some j, then $\sup_{\lambda \geq 0} L(x^\star, \lambda) = +\infty$, which is forbidden by the second inequality in (29). Because $x^\star$ is feasible, $\sup_{\lambda \geq 0} L(x^\star, \lambda) = f_0(x^\star)$, and we conclude from the second inequality in (29) that $\mathcal{L}(x^\star, \lambda^\star) = f_0(x^\star)$. Now, the first inequality in (29) reads

$$f_0(x) + \sum_{j=1}^{m} \lambda_j^\star f_j(x) \geq f_0(x^\star) \quad \forall x \in X.$$

This inequality implies that $x^\star$ is optimal: indeed, if x is feasible for $(\mathscr{P})$, then the left side of the inequality is $\leq f_0(x)$ because $\lambda^\star \geq 0$ and $f_j(x) \geq 0$, so $f(x) \geq f(x^\star)$.

(ii): Assume that $(\mathscr{P})$ is a convex program, $x^\star$ is its optimal solution, and the problem satisfies the Slater condition. Then, we prove there exists $\lambda^\star \geq 0$ such that $(x^\star, \lambda^\star)$ is a saddle point of the Lagrange function. From the Convex Programming Duality Theorem,

the dual problem $(\mathscr{D})$ has a solution $\lambda^\star \geq 0$, and the optimal value of the dual problem equals $f(x^\star)$:

$$f_0(x^\star) = g(\lambda^\star) \equiv \inf_{x \in X} \Big[f_0(x) + \sum_{j=1}^{m} \lambda_j^\star f_j(x) \Big].$$

In particular, this implies that

$$f_0(x^\star) \leq \mathcal{L}(x^\star, \lambda^\star) = f_0(x^\star) + \sum_{j=1}^{m} \lambda_j^\star f_j(x^\star).$$

But all the terms in the sum $\sum_{j=1}^{m} \lambda_j^\star f_j(x^\star)$ are negative (because $x^\star$ and $\lambda^\star$ are feasible for the primal and dual, respectively), so the inequality above implies that each term must actually be zero. So $\lambda_j^\star \cdot f_j(x^\star) = 0$ and we have $f(x^\star) = \mathcal{L}(x^\star, \lambda^\star)$. Therefore:

$$\mathcal{L}(x^\star, \lambda^\star) = f(x^\star) = \inf_{x \in X} \mathcal{L}(x, \lambda^\star).$$

Because $x^\star$ is feasible for $(\mathscr{P})$, we have $\mathcal{L}(x^\star, \lambda) \leq f(x^\star)$ for $\lambda \geq 0$, implying

$$L(x^\star, \lambda) \leq L(x^\star, \lambda^\star) \leq L(x, \lambda^\star)$$

for all $x \in X$ and $\lambda \geq 0$. □

4.1 Game interpretation

The saddle-point properties developed above also bare a natural interpretation in terms of a continuous zero-sum game between a decision maker and an adversary. If the first player chooses $w \in W$, and the second player selects $z \in Z$, then player 1 pays an amount $f(w, z)$ to player 2. Player 1 therefore wants to minimize f, while player 2 wants to maximize f.

The critical comparison between

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) \quad \textit{versus} \quad \inf_{w \in W} \sup_{z \in Z} f(w, z). \tag{30}$$

then boils down to the order of play. Suppose that player 1 makes their choice first, and then player 2, after learning the choice of player 1, makes their selection. This corresponds to the second game above. Player 2's will seek to maximize the payoff $f(w, z)$ and so will choose $z \in Z$ to maximize $f(w, z)$. Critically, Player 2's choice z is allowed to depend on the choice w made by Player 1 so the resulting payoff, which is $\sup_{z \in Z} f(w, z)$, will also depends on w, the choice of the first player. Player 1 knows (or assumes) that player 2 will follow this strategy, and so will choose $w \in W$ to make this worst-case payoff to player 2 as small as possible. Thus player 1 chooses

$$\operatorname{argmin}_{w \in W} \sup_{z \in Z} f(w, z),$$

which results in the payoff

$$\inf_{w \in W} \sup_{z \in Z} f(w, z)$$

from player 1 to player 2. In this game, Player 2 has an informational advantage over player 1 because she makes her choice after observing the choice of Player 1.

Now suppose the order of play is reversed: player 2 must choose $z \in Z$ first, and then player 1 chooses $w \in W$ (with knowledge of z). Following a similar argument, if the players follow the optimal strategy, player 2 should choose $z \in Z$ to maximize $\inf_{w \in W} f(w, z)$, which results in the payoff

$$\sup_{z \in Z} \inf_{w \in W} f(w, z)$$

from player 1 to player 2.

The max-min inequality states the (intuitively obvious) fact that it is better for a player to go second, or more precisely, for a player to know their opponent's choice before choosing. In other words, the payoff to player 2 will be larger if player 1 must choose first. The optimal duality gap for the problem is exactly equal to the advantage afforded to the player who goes second. If strong duality holds – or equivalently, the saddle-point property holds – there is no advantage to playing second. If $(w^\star, z^\star)$ is a saddle-point for f (and W and Z), then it is called a solution of the game.

4.2 Sion Mini-max Result

One of the most celebrated results in optimization is the Sion-Kakutani Theorem that allows interchanging the order of minimization and maximization in a minimax problem.

Theorem 4 (Sion-Kakutani). *Let $X \subseteq \mathbb{R}^n$ and $\Lambda \subseteq \mathbb{R}^m$ be convex and compact subsets and let $f : X \times \Lambda \to \mathbb{R}$ be a continuous function that is convex in $x \in X$ for any fixed $\lambda \in \Lambda$ and that is concave in $\lambda \in \Lambda$ for any fixed $x \in X$. Then,*

$$\min_{x \in X} \max_{\lambda \in \Lambda} f(x, \lambda) = \max_{\lambda \in \Lambda} \min_{x \in X} f(x, \lambda).$$

We note that slight generalizations of this result are also possible. (Λ only needs to be convex – so no need for compactness – and f only needs to be lower semicontinuous and quasi-convex on X and upper semicontinuous and quasi-concave on Λ. We omit further details here.) A proof is slightly outside the scope of these notes, but we direct the interested reader to Ben-Tal and Nemirovski (2023) and Bertsekas (2009) for more details.

5 Optimality Conditions

We next discuss optimality conditions for optimization problems. We will be concerned with the following primal optimization problem:

$$\begin{array}{lll} (\mathscr{P}) \ \min_x & f_0(x) & \\ & f_i(x) \leq 0, & i = 1, \ldots, m \\ & h_i(x) = 0, & i = 1, \ldots, s \\ & x \in X. & \end{array} \tag{31}$$

The question we are interested in is the following: *"Assume that we are given a feasible solution $x^\star$ to $(\mathscr{P})$. What are the conditions (necessary, sufficient, necessary and sufficient)*

for $x^\star$ to be optimal?" We intend to answer this question under the following assumptions on the problem primitives:

- $x^\star$ is an interior point of the domain of the problem X;
- The functions f, $g_1, \ldots, g_m$ and $h_i, \ldots, h_s$ are smooth at $x^\star$: at least once continuously differentiable in a neighborhood of the point (for second-order conditions, we would need to require more smoothness!)

Importantly, we stress that **we are not going to impose structural convexity assumptions**, unless explicitly stated otherwise.

Before stating the conditions, we note that the only kinds of conditions that we should hope for are **necessary** conditions for the optimality of $x^\star$ and **sufficient** conditions for the **local optimality** of $x^\star$. In particular, we cannot possibly hope for global optimality conditions without imposing some other global requirements (such as convexity).

Letting λ denote the dual variables for the inequality constraints $f_j(x) \leq 0$ and ν denote the dual variables for the equality constraints $h_j(x) = 0$, recall from the developments in the previous section that if we have an optimal solution $x^\star$ for the primal $(\mathscr{P})$ and an optimal solution $\lambda^\star, \nu^\star$ for its dual so that strong duality holds, this implies:

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\ &= \inf_{x \in X} \Big[f(x) + \sum_{j=1}^{m} \lambda_j^\star f_j(x) + \sum_{j=1}^{s} \nu_j^\star h_j(x) \Big] \\ &\leq f_0(x^\star) + \sum_{j=1}^{m} \lambda_j^\star f_j(x^\star) \\ &\leq f_0(x^\star), \end{aligned}$$

The first inequality follows because $x^\star$ is feasible in $(\mathscr{P})$ so $f_j(x^\star) \leq 0$ (we omit writing the term $+\sum_{j=1}^{s} \nu_j^\star h_j(x^\star)$, which is zero anyway) and the last inequality also uses that $\lambda^\star$ is feasible in $(\mathscr{D})$, so $\lambda^\star \geq 0$. But this implies that:

$$\lambda_i^\star \cdot f_i(x^\star) = 0, \; i = 1, \ldots, m. \tag{32}$$

This condition, which we already encountered in linear optimization, is called **complementary slackness**, and it can be expressed equivalently as

$$\lambda_i^\star > 0 \Rightarrow f_i(x^\star) = 0 \quad \Leftrightarrow \quad f_i(x^\star) < 0 \Rightarrow \lambda_i^\star = 0. \tag{33}$$

These conditions will be very important as they will allow us to establish necessary (and sufficient) optimality conditions for optimization problems.

5.1 Karush-Kuhn-Tucker (KKT) Optimality Conditions

Let $x^\star \in X$ be a point that in the domain for the primal $(\mathscr{P})$ and let $\lambda^\star \in \mathbb{R}^m$ be dual variables corresponding to the inequality constraints and $\nu^\star$ be dual variables for the equality

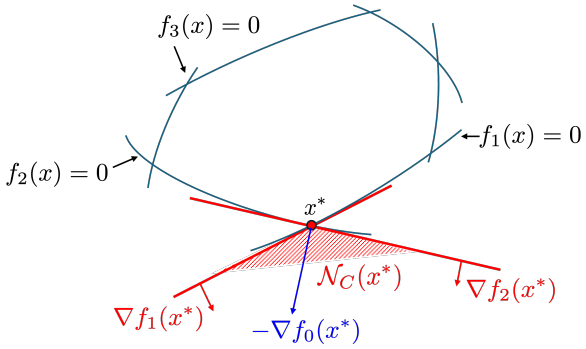

Figure 4: Illustration of KKT conditions. Here, the feasible set is the intersection of several inequality constraints $f_i(x) \leq 0$. At the optimal point $x^\star$, only $f_1(x)$ and $f_2(x)$ are active constraints. The Stationarity Condition requires that the negative of gradient of the objective, $-\nabla f_0(x^\star)$, can be expressed as a conic combination of the the gradients of all active constraints, i.e., $\nabla f_1(x^\star)$ and $-\nabla f_2(x^\star)$ here. (The set of all conic combinations of these gradients is denoted by $\mathcal{N}_C(x^\star)$ and is called the **normal cone** at $x^\star$.)

constraints. The **Karush-Kuhn-Tucker (KKT) conditions** at $x^\star \in X$ are given by:

$$
\begin{aligned}
&0 = \nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \cdot \nabla f_i(x^\star) + \sum_{i=1}^{s} \nu_i^\star \cdot \nabla h_i(x^\star), && \text{(“Stationarity”)} \\
&f_i(x^\star) \leq 0, \ i = 1, \ldots, m && \text{(“Primal Feasibility 1”)} \\
&h_i(x^\star) = 0, \ i = 1, \ldots, s, && \text{(“Primal Feasibility 2”)} \\
&\lambda^\star \geq 0 && \text{(“Dual Feasibility”)} \\
&\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m && \text{(“Complementary Slackness”).}
\end{aligned}
$$

In this definition, we noted common names for each condition in quotes. The rationale for the conditions should be clear from our previous developments involving the primal-dual.

To visualize these conditions, consider a case without equality constraints ($s = 0$). Note that the Stationarity Condition (which also corresponds to the derivative of the Lagrangian vanishing) together with the Complementarity Slackness condition (which states that $\lambda_i^\star = 0$ for any inequality constraints $f_i(x) \leq 0$ that are not active) yield:

$$
-\nabla f_0(x^\star) = \sum_{i : f_i(x^\star) = 0} \lambda_i^\star \cdot \nabla f_i(x^\star).
$$

This means that at optimality, $-\nabla f_0(x^\star)$ can be written as a conic combination of the gradients of all the constraints that are active at $x^\star$. The cone of all such directions is known as the **normal cone** at $x^\star$ and it denoted by $\mathcal{N}_C(x^\star)$. Note that $\mathcal{N}_C(x^\star)$ contains all the directions $d \in \mathbb{R}^n$ that “point away” from the feasible set, i.e., $\mathcal{N}_C(x^\star) := \left\{ d \in \mathbb{R}^n : d^\intercal (y - x^\star) \geq 0 \right\}$.[2] The geometric intuition of these conditions is depicted in Figure 4, and should be reminiscent of the optimality conditions we saw in linear optimization.

[2]Equivalently, the directions in $-\mathcal{N}_C(x^\star)$ allow moving from $x^\star$ while remaining inside the feasible set.

In some cases, the KKT conditions **may fail to hold at optimality**. Typically that happens when the linearization of the constraints collapses. Consider the following example.

Example 3 (Failure of KKT Conditions.)**.** *Consider the optimization problem*

$$\begin{aligned} \min_{x \in \mathbb{R}} \ & x \\ & x^3 \geq 0. \end{aligned}$$

In this example, $f_0(x) = x$ and $f_1(x) = -x^3$. The feasible set is $(-\infty, 0]$ and the optimal solution is $x^\star = 0$. The KKT condition fails because $\nabla f_0(x^\star) = 1$ while $\nabla f_1(x^\star) = 0$, so there is no $\lambda \geq 0$ such that $-\nabla f_0(x^\star) = \lambda \nabla f_1(x^\star)$. Note that in this case, we are **not** dealing with a convex optimization problem!

Here is a more subtle example of the KKT condition failing, in which the constraint gradients do not vanish.

Example 4 (Failure of KKT Conditions.)**.** *Consider the optimization problem*

$$\begin{aligned} \min_{x,y \in \mathbb{R}} \ & -x \\ & y - (1-x)^3 \leq 0 \\ & x, y \geq 0 \end{aligned}$$

Here, $f_0(x, y) := -x$, $f_1(x, y) := y - (1-x)^3$, $f_2(x, y) := -x$ and $f_3(x, y) := -y$. The feasible set is illustrated in Figure 5. At the optimal point $(x^\star, y^\star) := (1, 0)$, the gradients

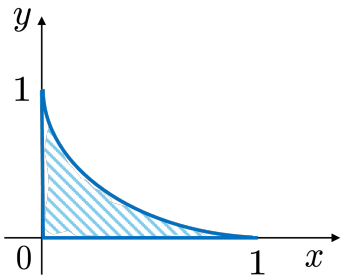

Figure 5: KKT Conditions Failing. (Figure not drawn to scale.)

of the objective and binding constraints f_1 and f_3 are

$$\nabla f_0(x^\star, y^\star) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \nabla f_1(x^\star, y^\star) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \nabla f_3(x^\star, y^\star) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

It is clear that no Lagrange multipliers λ_1, λ_3 satisfy

$$-\nabla f_0(x^\star, y^\star) = \lambda_1 \nabla f_1(x^\star, y^\star) + \lambda_3 \nabla f_3(x^\star, y^\star),$$

so the KKT conditions fail in this case.

The reason why the KKT conditions failed in the last example is because the linearization of constraint $f_1 \leq 0$ around the optimal point $(1, 0)$ is $y \leq 0$, which is parallel to the existing constraint $y \geq 0$ and fails the capture the fact that $x \leq 1$ on the feasible set.

Luckily, several **constraint qualification** conditions exist to prevent such pathological behavior; we highlight some examples below. In all of these conditions, $x^\star$ is the candidate point for which we want to check local optimality and we let $I(x^\star) := \{i \in \{1, \ldots, m\} : f_i(x^\star) = 0\}$ denote the set of indices of all **active inequality constraints**. We restrict attention to cases where the functions $\{f_i\}) =_{i=1,\ldots,m}$ and $\{h_j\}_{j=1,\ldots,s}$ are differentiable. If any of these constraint qualification conditions hold, then the KKT conditions are **necessary** for $x^\star$ to be locally optimal.

1. Affine constraints. If the feasible set is defined by linear constraints (i.e., all h_i and f_j are affine functions), then no further constraint qualifications are necessary and the KKT conditions are necessary at $x^\star$.

2. Slater's condition. This is the condition we are already familiar with, which we can relax slightly by only making reference to **active** constraints. Specifically, the relaxed Slater's condition holds if the functions f_i appearing in **active** inequality constraints $\{f_i : i \in I(x^\star)\}$ are convex and there exists a **feasible point** $\bar{x}$ in the relative interior of the domain $x \in \text{rel int}(X)$ that is strictly feasible for these, i.e.,

$$f_j(\bar{x}) < 0 \quad \forall j \in I(x^\star),$$

and if all the functions $\{h_j\}_{j=1,\ldots,s}$ appearing in equality constraints are affine.

3. Linearly independent gradients for active constraints. Suppose that the gradients of all active constraints at $x^\star$ are linearly independent, i.e., the vectors:

$$\{\nabla f_j(x^\star) : j \in I(x^\star)\} \cup \{\nabla h_j(x^\star) : j = 1, \ldots, s\}$$

has linearly independent vectors. Then, the KKT conditions are necessary at $x^\star$.

A point $x^\star$ where the gradients of active constraints are linearly independent is also referred to as a **regular** point. You may recall regular points from multivariate calculus, where regularity is a necessary condition for the implicit function theorem to hold.

4. Mangasarian-Fromovitz. Suppose the gradients of all equality constraints

$$\{\nabla h_j(x^\star) : j = 1, \ldots, r\}$$

are linearly independent and there exists a vector $d \in \mathbb{R}^n$ such that

$$d^\intercal \nabla f_j(x^\star) < 0,\ i \in I(x^\star), \qquad d^\intercal \nabla h_j(x^\star) = 0,\ j = 1, \ldots, s,$$

then the KKT conditions are necessary at $x^\star$.

As it turns out, these constraint qualification conditions satisfy a specific "pecking order," in the sense that some conditions are stronger and imply others. For instance, one can

show that condition (3) requiring linearly independent gradients implies the Mangasarian-Fromovitz condition (4). In principle, even more relaxed conditions are possible; we refer the interested reader to the lecture notes Burke (2012) and the article Peterson (1973) for a more thorough overview. However, the most practical conditions to check are the Slater condition (when dealing with a convex optimization problem) or the Mangasarian-Fromovitz condition for a more general (smooth) non-linear optimization problem.

5.2 Second Order Optimality Conditions

Under additional smoothness assumptions on the objective and constraints, we can also state a set of second-order optimality conditions that make use of Hessian information.

Second Order Necessary Optimality Conditions

Theorem 5 (Necessary Conditions)**.** *Consider problem* $(\mathscr{P})$ *stated in*(31) *and assume that* $x^\star$ *is a feasible solution and* $f_0, f_1, \ldots, f_m, h_1, \ldots, h_s$ *are twice continuously differentiable in a neighbourhood of* $x^\star$*. Let* $I(x^\star) := \{i \in \{1, \ldots, m\} : f_i(x^\star) = 0\}$ *denote the indices of all* ***active*** *inequality constraints at* $x^\star$ *and assume that* $x^\star$ *is* ***regular****, i.e., the gradients*

$$\{\nabla f_j(x^\star) : j \in I(x^\star)\} \cup \{\nabla h_j(x^\star) : j = 1, \ldots, s\}$$

of all active constraints at $x^\star$ *are linearly independent. Then, if* $x^\star$ *is locally optimal, there exist unique Lagrange multipliers* $\lambda_i^\star \geq 0$ *and* $\nu_j^\star$ *such that*

(i) $(\lambda^\star, \nu^\star)$ *certify that* $x^\star$ *is a KKT point of (P):*

$$\nabla_x \mathcal{L}(x^\star; \lambda^\star, \nu^\star) = \nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{j=1}^{s} \nu_j^\star \nabla h_j(x^\star) = 0 \tag{34a}$$

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m \tag{34b}$$

(ii) The Hessian $\nabla_x^2 \mathcal{L}(x^\star; \lambda^\star, \mu^\star)$ *of* $\mathcal{L}$ *in* x *is positive semidefinite on the orthogonal complement* $M^\star$ *to the set of gradients of active constraints at* $x^\star$*:*

$$d^T \nabla_x^2 \mathcal{L}(x^\star; \lambda^\star, \mu^\star)\, d \geq 0 \text{ for any } d \in M^\star$$
$$\text{where } M^\star := \{d \mid d^T \nabla f_i(x^\star) = 0, \forall\, i \in I(x^\star),\ d^T \nabla h_j(x^\star) = 0,\ j = 1, \ldots, s\}.$$

The Second Order Necessary Optimality Conditions actually state some intuitive facts. To see it, it first helps to develop an intuition for the subspace $M^\star$ involved in the necessary condition. This is the subspace obtained by linearizing all the constraints that are active at $x^\star$, so the affine space $x^\star + M^\star$ is exactly a tangent plane to the surface $\mathcal{S}$ where all constraints active at $x^\star$ are still active. Thus, directions $d \in M^\star$ are tangent to $\mathcal{S}$ at $x^\star$. When $x^\star$ is regular, moving forward or backwards along any such direction $d \in M^\star$ allows us to stay "very close" to $\mathcal{S}$. So when x^* is locally optimal, it must be that no direction

from M^* leads to a desired decrease of the objective. Indeed, if such a direction d existed that, we could improve on x^* by implementing a small step along this tangent direction, which would improve the objective only by an infinitesimal shift of second order.

In a similar fashion, we can also state a set of **sufficient** second-order conditions that would guarantee that a point $x^\star$ is a local optimum.

Second Order **Sufficient** Optimality Conditions

Theorem 6 (Sufficient Conditions)**.** *Under the same premises as stated in Theorem 5, assume that there exist Lagrange multipliers $\lambda_i^\star \geq 0$ and $\nu_j^\star$ such that:*

(i) $(\lambda^\star, \nu^\star)$ certify that $x^\star$ is a KKT point of $(\mathscr{P})$, i.e., (34a) and (34b) hold.

*(ii) The Hessian $\nabla_x^2 \mathcal{L}(x^\star; \lambda^\star, \mu^\star)$ of $\mathcal{L}$ in x is **positive definite** on the orthogonal complement $M^{\star\star}$ to the set of gradients of equality constraints and active inequality constraints **associated with positive Lagrange multipliers** $\lambda_i^\star$:*

$$d^\intercal \nabla_x^2 \mathcal{L}(x^\star; \lambda^\star, \mu^\star) d > 0 \text{ for any } d \in M^{\star\star}$$

where

$$M^{\star\star} := \{d \mid d^\intercal \nabla f_i(x^\star) = 0, \forall\, i \in I(x^\star) : \lambda_i^\star > 0 \text{ and } d^\intercal \nabla h_j(x^\star) = 0,\ j = 1, \ldots, s\}.$$

Then, $x^\star$ is locally optimal for $(\mathscr{P})$.

Note that the sufficient condition involves a stronger requirement on the Hessian: it should be positive **definite** in the subspace $M^{\star\star}$.

We omit proofs for these reasons due to space limitations. The interested reader can refer to Ben-Tal and Nemirovski (2023) or the book Borwein and Lewis (2006).

As we stated at the onset, the conditions above provide necessary and sufficient conditions for **local** optimality. When the optimization problem exhibits other (global) properties – for instance, when we deal with **convex optimization problems** like the ones we discussed in the previous sections – these conditions actually become **necessary and sufficient** for global optimality.

5.3 Examples

5.3.1 A Consumer's Constrained Consumption Problem

Consider a consumer trying to maximize his utility function $u(x)$ by choosing which bundle of goods $x \in \mathbb{R}_n^+$ to purchase. The goods have prices $p > 0$ and the consumer has a budget

$B > 0$. The consumer's problem can be stated as:

$$\begin{aligned} &\text{maximize } u(x) \\ &\text{such that } p^\intercal x \leq B \\ &\qquad\quad x \geq 0, \end{aligned}$$

where $u(x)$ is a concave utility function.

Let us express the KKT conditions when the utility function $u(x)$ is differentiable. We first convert this into the following equivalent problem:

$$\begin{aligned} \text{minimize } & -u(x) \\ (\lambda \rightarrow) \ & p^\intercal x \leq B \\ (\mu \rightarrow) \ & -x \leq 0, \end{aligned}$$

With $\lambda \in \mathbb{R}_+, \mu \in \mathbb{R}^n_+$ denoting the Lagrange multipliers, the Lagrangian becomes:

$$\mathcal{L}(x, \lambda, \mu) = -u(x) + \lambda(p^\intercal x - B) - x^\intercal \mu.$$

This is a convex optimization problem and the Slater condition is trivially satisfied (with a sufficiently small choice $x > 0$). The KKT conditions are therefore necessary and sufficient for optimality. These conditions at a primal point x and dual point λ, μ can be written as:

$0 = -\dfrac{\partial u}{\partial x_i} + \lambda p_i - \mu_i,\ i = 1, \ldots, n$	("Stationarity")
$p^\intercal x \leq B, \quad x \geq 0$	("Primal Feasibility")
$\lambda \geq 0,\ \mu \geq 0$	("Dual Feasibility")
$\lambda \cdot (p^\intercal x - B) = 0$	("Complementary Slackness" 1)
$\mu_i \cdot x_i = 0$	("Complementary Slackness" 2).

We distinguish two cases, depending on whether $p^\intercal x < B$ holds.

Case 1. If the consumer's budget constraint is not binding, $p^\intercal x < B$, then $\lambda = 0$ from the complementary slackness condition, and we have

$$\frac{\partial u}{\partial x_i} = -\mu_i.$$

Because for any $x_i > 0$, we must have $\mu_i = 0$, this implies that the optimal consumption bundle satisfies:

$$\frac{\partial u}{\partial x_i} = 0 \ \text{ for any } x_i > 0.$$

In words, the consumer purchases the unconstrained optimal amount of each good i.

Case 2. If $p^\intercal x = B$, then it is possible to have $\lambda = 0$ or $\lambda > 0$. The former case would lead to the same qualitative insights as Case 1. If $\lambda > 0$, then we have:

$$\frac{\partial u}{\partial x_i} = \lambda p_i \ \text{ for any } x_i > 0,$$

or equivalently,

$$\frac{\frac{\partial u}{\partial x_i}}{p_i} = \lambda \quad \text{for any } x_i > 0.$$

The quantity on the left has a very nice interpretation as the "bang-for-the-buck" for good i: it is the marginal increase in utility if the consumer increased the consumption of the good by a small amount δ (beyond the optimal consumption) , $\frac{\partial u}{\partial x_i} \cdot \delta$, divided by the price of the extra purchase, δp_i. So the condition states a very well-known economic fact that when the consumer is budget-constrained, the bang-for-the-buck for all the goods that are consumed ($x_i > 0$) must be equal at optimality.

Moreover, note that the stationarity condition also implies that for any i with $x_i > 0$ and j with $x_j = 0$, we have:

$$\frac{\frac{\partial u}{\partial x_i}}{x_i} = \lambda > \frac{\frac{\partial u}{\partial x_j}}{x_j} = \lambda - \mu_j,$$

so the bang-for-the-buck for goods that are consumed must be (weakly) larger than for goods that are not consumed.

6 Fenchel Duality

In this section we briefly sketch out the elegant and concise theory of Fenchel Duality, which can be used to gain a deeper understanding of optimality conditions stated earlier as well as to appreciate some important constructions in optimization.

We start by defining a central concept in convex optimization and convex optimization: the **conjugate** of a function f.

Conjugate of a function

Let $f : \mathbb{R}^n \to \mathbb{R}$. The function $f^* : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f^*(y) = \sup_{x \in \text{dom}(f)} \left\{ y^\intercal x - f(x) \right\} \tag{35}$$

is called the **conjugate** of f.

The construction is depicted in Figure 6. The rationale behind the definition is to be able to describe f in terms of the affine functions that are majorized by f, i.e., supporting hyperplanes to epi(f). When f is a closed convex function that is also proper (i.e., does not take value $-\infty$ anywhere), this description is actually accurate and the transformation is symmetric, i.e., f can be recovered by taking the conjugate of its conjugate f^*. The conjugacy transformation thus provides an alternative view of a convex function, which often reveals interesting properties and is useful for analysis and computation.

Note that regardless of the structure of f, the conjugate function f^* is convex, because it is the pointwise supremum of affine functions of y:

$$x^\intercal y - f(x) \, \forall \, x \in \text{dom}(f).$$

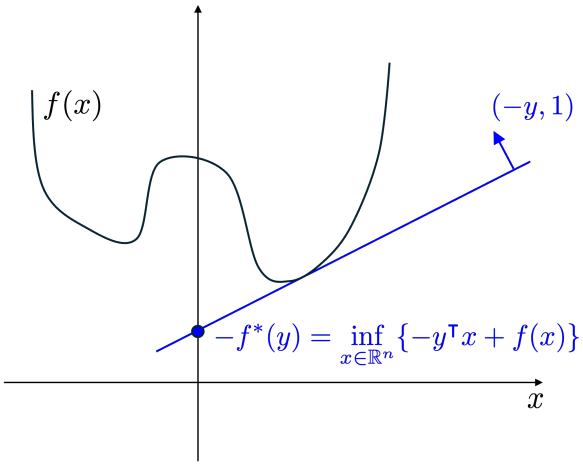

Figure 6: Visualization of the conjugate function $f^*(y) = \sup_{x \in \mathrm{dom}(f)} \{y^\intercal x - f(x)\}$ of a function f. The crossing point of the vertical axis with the hyperplane with normal $(-y, 1)$ that supports the epigraph of f is exactly $-f^*(y)$.

6.1 Basic Examples

We present a few examples of conjugate functions.

The zero function.

Example 5. *For $f(x) = 0$, the conjugate will depend on the relevant domain:*

- *If $f : \mathbb{R} \to \mathbb{R}$, then $f^* : \{0\} \to \mathbb{R}$ and $f^*(y) = 0$.*
- *If $f : \mathbb{R}_+ \to \mathbb{R}$, then yx is unbounded for $y > 0$ and for $y < 0$, it achieves its maximum for $x = 0$. We then have $f^* : (-\infty, 0] \to \mathbb{R}$ and $f^*(y) = 0$.*
- *If $f : [-1, 1] \to \mathbb{R}$, then yx achieves its maximum for $x = \mathrm{sign}(y)$ and we have $f^* : \mathbb{R} \to \mathbb{R}$ and $f^*(y) = |y|$.*
- *If $f : [0, 1] \to \mathbb{R}$, then for $y < 0$, the function yx achieves its maximum value of 0 at $x = 0$, and for $y \geq 0$ it achieves its maximum of y at $x = 1$. So we have $f^* : \mathbb{R} \to \mathbb{R}$ and $f^*(y) = y^+$.*

Affine functions.

Example 6. *Consider $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = a^\intercal x + b$. Note that $y^\intercal x - a^\intercal x - b$ is finite if and only if $y = a$, in which case it equals $-b$. Therefore $f^* : \{a\} \to \mathbb{R}$ and $f^*(a) = -b$.*

Absolute value.

Example 7. *Consider $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = |x|$. Note that $y^\intercal x - |x|$ has a finite supremum 0 if and only if $y \in [-1, 1]$. Therefore, $f^* : [-1, 1] \to \mathbb{R}$ and $f^*(y) = 0$.*

Negative logarithm.

Example 8. *Consider $f : (0, \infty) \to \mathbb{R}$ with $f(x) = -\log x$. The function $yx + \log x$ is unbounded above if $y \geq 0$ and reaches its maximum at $x = -1/y$ otherwise. Therefore, $f^* : (-\infty, 0) \to \mathbb{R}$ and $f^*(y) = -\log(-y) - 1$ for $y < 0$.*

Exponential.

Example 9. *Consider $f : \mathbb{R} \to \mathbb{R}, f(x) = e^x$. Then, $yx - e^x$ is unbounded if $y < 0$. For $y > 0$, $yx - e^x$ reaches its maximum at $x = \log y$, so we have $f^*(y) = y \log y - y$. For $y = 0$,*

$$f^*(y) = \sup_x -e^x = 0.$$

In summary, $f^* : \mathbb{R}_+ \to \mathbb{R}$ and

$$f^*(y) = \begin{cases} y \log y - y & y > 0 \\ 0 & y = 0. \end{cases} \tag{36}$$

Negative entropy.

Example 10. *Consider $f : [0, \infty) \to \mathbb{R}, f(x) = x \log x$ (with the convention $\lim_{x \to 0} f(x) = 0$). The function $yx - x \log x$ is bounded above on $[0, \infty)$ for all y and attains its maximum at $x = e^{y-1}$. Hence $f^* : \mathbb{R} \to \mathbb{R}$ and $f^*(y) = e^{y-1}$.*

Inverse.

Example 11. *Consider $f(x) = 1/x$ defined on $0, \infty$. For $y > 0$, $yx - 1/x$ is unbounded above. For $y = 0$ this function has supremum 0; for $y < 0$ the supremum is attained at $x = (-y)^{-1/2}$. Therefore $f^* : [0, \infty) \to \mathbb{R}$ and $f^*(y) = -2(-y)^{1/2}$.*

Strictly Convex Quadratic Function

Example 12. *Consider $f : \mathbb{R}^n \to \mathbb{R}, f(x) = \frac{1}{2} x^T Q x$, where $Q \succ 0$. The function $y^T x - \frac{1}{2} x^T Q x$ attains its maximum at $x = Q^{-1} y$ for any y, so $f^* : \mathbb{R}^n \to \mathbb{R}$ and $f^*(y) = \frac{1}{2} y^T Q^{-1} y$.*

Indicator Function

Example 13. *Let I_S be the indicator function of a (not necessarily convex) set $S \subset \mathbb{R}^n$, i.e., $I_S(x) = 0$ on $\operatorname{dom} I_S = S$ and $I_S(x) = +\infty$ otherwise. Its conjugate is*

$$I_S^*(y) = \sup_{x \in S} y^T x,$$

which is the support function of the set S.

6.2 Conjugate of Conjugate and Convex Envelope

For a function $f : \mathbb{R}^n \to \mathbb{R}$, consider the conjugate of the conjugate function f^* (or the **double conjugate**) denoted by f^{**} and given by

$$f^{**}(x) = \sup_{y \in \mathbb{R}^n} \{y^T x - f^*(y)\}, \quad x \in \mathbb{R}^n.$$

The next proposition shows that f^{**} is the **convex closure** or **convex envelope** of f, i.e., the function that has as epigraph the closure of the convex hull of $\operatorname{epi}(f)$. In particular, the last part of the result shows that under a few mild technical conditions, $f^{**} = f$ for a convex function f.

Conjugacy Theorem

Theorem 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that $\operatorname{epi}(f)$ is a closed set and let f^{**} be the double-conjugate. Then,*

a) We have $f(x) \geq f^{**}(x)$, $forall\, x \in \mathbb{R}^n$.

b) If f is convex, then $f(x) = f^{**}(x)$, $\forall\, x \in \mathbb{R}^n$.

c) $f^{**}(x)$ equals the convex envelope of f, i.e., the largest convex function $g(x)$ satisfying $g(x) \leq f(x)$ for any $x \in \mathbb{R}$.

For a proof, we refer the interested reader can refer to Bertsekas (2009).

This result has important implications, albeit more for theory than practice! Specifically, it can be shown that the optimal value in the problem of minimizing an **arbitrary** (i.e., potentially non-convex) closed function f – if finite – is the same as the optimal value when minimizing the convex envelope of f. Therefore, **IF** we had access to the convex function f^{**}, we could solve a convex optimization problem to determine the optimal value of any function f. Obviously, the challenge here lies gaining access to f^{**}: in general, that function is extremely difficult to compute or even approximate for arbitrary functions f!

6.3 Important Inequalities

An immediate consequence of the definition of the conjugate is the following inequality, which is called **Fenchel** (or **Fenchel-Young**) inequality:

Fenchel-Young Inequality

$$f^*(y) \geq y^\intercal x - f(x).$$

More importantly, the conjugates of functions allow us to restate the strong duality result in a very concise and illuminating form. To appreciate this, consider the following optimization problem:

$$\begin{aligned} &\text{minimize } f_1(x) + f_2(x) \\ &\text{subject to } x \in X_1 \cap X_2 \end{aligned}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ and $X_i \subseteq \mathbb{R}^n$ for $i = 1, 2$. Let's assume that the optimal value is finite and equal to $p^\star$. Then, the problem can be converted into:

$$\begin{aligned} &\text{minimize } f_1(y) + f_2(z) \\ &\text{subject to } z = y,\ z \in X_1,\ y \in \cap X_2. \end{aligned}$$

Moreover, we can dualize the constraint $z = y$ and construct a dual lower bound. Specifically, for any $\lambda \in \mathbb{R}^n$, define the following functions:

$$\begin{aligned} g(\lambda) &= \inf_{y \in X_1, z \in X_2} \{f_1(y) + f_2(z) + (z - y)^\intercal \lambda\} \\ &= -\sup_{y \in X_1} \{y^\intercal \lambda - f_1(y)\} + \inf_{z \in X_2} \{z^\intercal \lambda + f_2(z)\} \\ &= -\sup_{y \in X_1} \{y^\intercal \lambda - f_1(y)\} - \sup_{z \in X_2} \{-z^\intercal \lambda - f_2(z)\} \\ &:= -g_1(\lambda) - g_2(-\lambda), \end{aligned}$$

where $g_1(\lambda)$ is the conjugate of f_1 and $g_2(\lambda)$ is the conjugate of f_2.

Clearly, for any λ, $g(\lambda)$ is a lower bound on $p^\star$, and we can form the following dual problem:

$$\max_{\lambda \in \mathbb{R}^n} \{-g_1(\lambda) - g_2(-\lambda)\},$$

which is actually equivalent to the problem

$$\min_{\lambda \in \mathbb{R}^n} \{g_1(\lambda) + g_2(-\lambda)\},$$

which has a very similar form to the primal problem.[3]

Then, the following main result holds.

[3] This could be made to look even more symmetric by considering instead a function $f_1(x) - f_2(x)$ and defining the convex and concave conjugates. We preferred to not introduce additional notation and instead obtain the slight asymmetry in the definitions of the primal and dual.

Fenchel Duality

Assume that f_1 and f_2 are convex and either (i) rel int$(\text{dom}(f_1)) \cap$ rel int$(\text{dom}(f_2) \neq \emptyset$ or (ii) $\text{dom}(f_i)$ are polyhedral and f_i can be extended to a real-valued convex function over $\mathbb{R}^n$ for $i = 1, 2$. Then, there exists $\lambda^\star \in \mathbb{R}^n$ such that

$$p^\star = g(\lambda^\star)$$

and strong duality holds.

For a proof, see Bertsekas (2009) or Borwein and Lewis (2006).

This result is essentially a restatement of the strong duality result for convex optimization. It is worth noting that condition (i) is simply a restatement of the Slater condition in this new framework (the Slater condition has been replaced with the existence of a point x in the relative interior of the domains of f_1 and f_2), while condition (ii) is primarily concerned with the polyhedral case. So the theorem is simply reinterpreting – rather than extending or generalizing – the previous results.

References

Aharon Ben-Tal and Arkadi Nemirovski. Lecture notes on optimization, convex analysis, nonlinear programming theory, and nonlinear programming algorithms. Technical report, Georgia Tech and Technion, 2023. URL https://www2.isye.gatech.edu/~nemirovs/OPTIIILN2023Spring.pdf.

Aharon Ben-Tal and Marc Teboulle. Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Mathematical Programming*, 72:51 – 64, 1996.

Dimitri Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

J. Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006. URL https://link.springer.com/book/10.1007/978-0-387-31256-9.

Steven Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

J.V. Burke. Numerical optimization. course notes. Technical report, University of Washington, 2012. URL https://sites.math.washington.edu/~burke/crs/516/notes/cq_lec.pdf.

D.W. Peterson. A review of constraint qualifications in finite-dimensional spaces. *SIAM Review*, 1973.