8 Duality (continued) and KKT conditions

Linear inequality constraints We consider a convex optimization problem where the constraint set C is given by a finite number of linear inequalities

$$\min_{x \in C} f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is smooth, and

$$C = \{x \in \mathbb{R}^n : \langle a_i, x\rangle \le b_i \ (i = 1, \dots, m)\}.$$

Equivalently, we can write $C = \{x \in \mathbb{R}^n : Ax \le b\}$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We know that x^* is an optimal point for (1) if, and only if, $-\nabla f(x^*) \in N_C(x^*)$, where N_C indicates the normal cone.

It is not difficult to show that the normal cone for the polyhedron C at $x \in C$ is precisely the cone spanned[1] by $\{a_i : \langle a_i, x\rangle = b_i\}$ (active constraints). This can be written as:

$$N_C(x) = \left\{ \sum_{i=1}^m \lambda_i a_i : \lambda_i \ge 0 \text{ and } \lambda_i(b_i - \langle a_i, x\rangle) = 0 \ \forall i = 1, \dots, m \right\}.$$

Using matrix notations, we have $\sum_{i=1}^m \lambda_i a_i = A^T\lambda$, where $\{a_i\}$ are the columns of A^T. Thus, the necessary and sufficient conditions for optimality can be written as

$$x^* \in \arg\min_{Ax\le b} f(x) \iff \exists \lambda^* \in \mathbb{R}^m \text{ s.t. } \begin{cases} -\nabla f(x^*) = A^T\lambda^* \\ Ax^* \le b \\ \lambda^* \ge 0 \\ \lambda_i^*(b_i - \langle a_i, x^*\rangle) = 0 \ \forall i = 1, \dots, m. \end{cases} \tag{2}$$

The constraints on the RHS are known as *Karush-Kuhn-Tucker conditions of optimality*. The last condition $\lambda_i^*(b_i - \langle a_i, x^*\rangle) = 0$ is known as *complementary slackness*.

Just like with the case of linear equality constraints, the dual variable λ^* can be understood as the solution of some dual optimization problem. Define the Lagrangian by

$$L(x, \lambda) = f(x) + \langle \lambda, Ax - b\rangle,$$

and the dual function

$$g(\lambda) = \min_{x\in\mathbb{R}^n} L(x, \lambda).$$

Then one can immediately verify:

$$\text{Weak duality:} \quad g(\lambda) \le \min_{Ax\le b} f(x) \quad \forall \lambda \ge 0.$$

(Note the condition $\lambda \ge 0$.) To verify weak duality we note

$$g(\lambda) = \min_{x\in\mathbb{R}^n} L(x, \lambda) \le \min_{Ax\le b} L(x, \lambda) = \min_{Ax\le b} f(x) + \langle \lambda, Ax - b\rangle \overset{(*)}{\le} \min_{Ax\le b} f(x)$$

[1] The cone spanned by $\{a_i\}$ is $\{\sum_i \lambda_i a_i : \lambda_i \ge 0 \ \forall i\}$.

where in $(*)$ we used the fact that $\langle \lambda, Ax - b\rangle \le 0$ since $Ax - b \le 0$ and $\lambda \ge 0$. Now the dual problem is

$$\max_{\lambda \ge 0} g(\lambda) = \max_{\lambda \ge 0} \min_{x \in \mathbb{R}^n} L(x, \lambda).$$

The KKT conditions (2) assert that we have *strong duality* and that the optimal value of the dual (maximization) problem is equal to the optimal value of the primal (minimization) problem. Indeed, we already know from weak duality that the value of the dual problem is $\le$ the value of the primal problem. For the reverse, let x^* be a solution of the primal optimization problem, and let λ^* be the dual variable so that (2) hold. Then

$$g(\lambda^*) \overset{(a)}{=} L(x^*, \lambda^*) = f(x^*) + \langle \lambda^*, Ax^* - b\rangle \overset{(b)}{=} f(x^*)$$

where in (a) we used the fact that $\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + A^T \lambda^* = 0$ and in (b) we used the fact that $\langle \lambda^*, Ax^* - b\rangle = \sum_{i=1}^m \lambda_i^* (\langle a_i, x^*\rangle - b_i) = 0$ (complementary slackness).

General problems Consider a constrained convex optimization problem of the form

$$\min_{x \in \mathbb{R}^n} \{ f(x) \;\text{ s.t. }\; x \in C_1 \cap C_2 \cap \cdots \cap C_m \}.$$

A necessary and sufficient condition for x^* to be a minimizer is that

$$0 \in \partial(f + I_{C_1} + \cdots + I_{C_m})(x^*).$$

If[2] $\mathbf{int}(\mathbf{dom}\, f) \cap \mathbf{int}\, C_1 \cap \cdots \cap \mathbf{int}\, C_m \neq \emptyset$ (Slater's condition), then condition above is equivalent to

$$0 \in \partial f(x^*) + N_{C_1}(x^*) + \cdots + N_{C_m}(x^*).$$

Using expressions for the normal cones, one can arrive at explicit necessary and sufficient conditions of optimality as described above.

Duality Consider a generic convex optimization with convex nonlinear inequality constraints:

$$f^* = \min_{x \in \mathbb{R}^n} \{ f(x) \;\text{ s.t. }\; Ax = b,\; h_1(x) \le 0, \dots, h_m(x) \le 0 \}. \tag{3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h_1, \dots, h_m : \mathbb{R}^n \to \mathbb{R}$ are convex, and $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$. The Lagrangian in this case is

$$L(x, \nu, \lambda) = f(x) + \langle \nu, Ax - b\rangle + \sum_{i=1}^m \lambda_i h_i(x)$$

where $\nu \in \mathbb{R}^p, \lambda \in \mathbb{R}^m$. The dual function is

$$g(\nu, \lambda) = \min_{x \in \mathbb{R}^n} L(x, \nu, \lambda).$$

It can be easily verified that if $\lambda \ge 0$, then $g(\nu, \lambda) \le f^*$. The dual problem is

$$g^* = \max_{\lambda \ge 0, \nu \in \mathbb{R}^p} g(\nu, \lambda). \tag{4}$$

[2]If some of the C_i are polyhedral, then it is not necessary to take the interior. I.e., if we let I be the set of i such that C_i is polyhedral, then it is sufficient to assume $\mathbf{int}\,\mathbf{dom}\, f \cap (\cap_{i \in I} C_i) \cap (\cap_{i \notin I} \mathbf{int}\, C_i) \neq \emptyset$. If the function f is polyhedral (i.e., $\mathbf{epi}(f)$ is polyhedral), then $\mathbf{int}\,\mathbf{dom}\, f$ can be replaced by $\mathbf{dom}\, f$.

Assuming that there is $\bar{x}$ such that $A\bar{x} = b$ and $h_i(\bar{x}) < 0$ for all $i = 1, \ldots, m$ (*Slater's condition*), then the primal and dual problems have the same value, i.e., we have strong duality. In this case, and if we assume further that the primal and dual problems are attained at (x^*, ν^*, λ^*), then we must have

$$\begin{cases} f(x^*) = g(\nu^*, \lambda^*) \\ Ax^* = b, h_i(x^*) \leq 0 \ \forall i = 1, \ldots, m \\ \lambda^* \geq 0. \end{cases} \tag{5}$$

The condition $f(x^*) = g(\nu^*, \lambda^*)$ can be simplified further: indeed under the conditions above we have

$$f(x^*) = g(\nu^*, \lambda^*) = \min_{x \in \mathbb{R}^n} L(x, \nu^*, \lambda^*) \leq L(x^*, \nu^*, \lambda^*) = f(x^*) + \sum_{i=1}^{m} \lambda_i^* h_i(x^*) \leq f(x^*),$$

where we used the fact that $Ax^* = b, \lambda^* \geq 0$ and $h_i(x^*) \leq 0$. The above implies that all inequalities are actually equalities, i.e., we must have:

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} L(x, \nu^*, \lambda^*) \ \text{ and } \ \lambda_i^* h_i(x^*) = 0 \ \forall i = 1, \ldots, m.$$

The conditions (5) thus become the KKT conditions:

$$\begin{cases} x^* \in \operatorname{argmin}_{x \in \mathbb{R}^n} L(x, \nu^*, \lambda^*) & \\ \lambda_i^* h_i(x^*) = 0 \ \forall i = 1, \ldots, m & \text{(complementary slackness)} \\ Ax^* = b, h_i(x^*) \leq 0 \ \forall i = 1, \ldots, m & \text{(primal feasibility)} \\ \lambda^* \geq 0 & \text{(dual feasibility).} \end{cases} \tag{6}$$

EXERCISE: The goal of this exercise is to sketch a direct proof of strong duality for the primal-dual pair of problems (3)-(4), assuming Slater's condition holds. For simplicity we omit the linear equality constraints. Define the convex set $K = \{(t_0, t_1, \ldots, t_m) \in \mathbb{R}^{m+1} : \exists x \in \mathbb{R}^n \text{ with } f(x) \leq t_0, h_i(x) \leq t_i \ \forall i\}$, and note that $f^* = \inf\{t_0 : (t_0, 0, \ldots, 0) \in K\}$. (a) Use the supporting hyperplane theorem to show the existence of $(\lambda_i) \geq 0$ and $b \in \mathbb{R}$ such that $\lambda_0 f(x) + \sum_{i=1}^m \lambda_i h_i(x) \geq b$ for all $x \in \mathbb{R}^n$, and $\lambda_0 f^* = b$. (b) Use Slater's condition to argue that $\lambda_0 > 0$. Conclude.

Example: dual decomposition Duality can be a very useful tool algorithmically. Consider an optimization problem of the form

$$\min_{x \in \mathbb{R}^n} \quad f_1(x) + f_2(x).$$

We assume the functions f_1 and f_2 are held on two different computers/devices, e.g., the functions f_i involve some training data that cannot be shared. We "decouple" the problem by introducing a new variable y and enforcing the constraint $x = y$:

$$\min_{x, y \in \mathbb{R}^n} f_1(x) + f_2(y) \text{ s.t. } x = y.$$

The Lagrangian of this problem is $L(x, y, \lambda) = f_1(x) + f_2(y) + \langle \lambda, x - y \rangle = (f_1(x) + \langle \lambda, x \rangle) + (f_2(y) - \langle \lambda, y \rangle)$. The dual function is

$$g(\lambda) = \min_{x, y \in \mathbb{R}^n} L(x, y, \lambda) = \min_{x \in \mathbb{R}^n} (f_1(x) + \langle \lambda, x \rangle) + \min_{y \in \mathbb{R}^n} (f_2(y) - \langle \lambda, y \rangle)$$

and the dual problem is to maximize $g(\lambda)$. We have already seen that g is concave, as it is the pointwise minimum of linear functions. Furthermore a supgradient[3] for g is given by $x^*(\lambda) - y^*(\lambda)$ where $x^*(\lambda)$ and $y^*(\lambda)$ are minimizers in the definition of g. Thus a supgradient method to maximize g takes the form: for $k = 0, 1, \ldots$

$$\begin{cases} \text{Compute } x^*(\lambda_k) \in \operatorname{argmin}_{x \in \mathbb{R}^n} \{f_1(x) + \langle \lambda_k, x \rangle\} \\ \text{ and } y^*(\lambda_k) \in \operatorname{argmin}_{y \in \mathbb{R}^n} \{f_2(y) - \langle \lambda_k, y \rangle\} \\ \text{Update } \lambda_{k+1} = \lambda_k + t_k(x^*(\lambda_k) - y^*(\lambda_k)) \end{cases}$$

where t_k is a step size. The advantage of this method is that it only requires separate minimizations of f_1 and f_2, and so the computations of $x^*(\lambda_k)$ and of $y^*(\lambda_k)$ can be done separately (and in parallel) on the devices where each f_i is known. The optimal points $x^*(\lambda_k)$ and of $y^*(\lambda_k)$ are then communicated to the central server, who updates λ and sends it back to the devices holding f_1, f_2, etc.

[3] A supgradient for a concave function f at x is a vector q s.t. $f(y) \leq f(x) + \langle q, y - x \rangle$ for all y.