EE 227A: Convex Optimization and Applications **February 9, 2012**

Lecture 8: **Strong Duality**

Lecturer: Laurent El Ghaoui

Reading assignment: Sections 5.1 through 5.4 (included) of BV. Sections 4.1-4.2 of the WTB.

8.1 Strong duality for convex problems

8.1.1 Primal and dual problems

In this section, we consider a *convex* optimization problem

$$p^* := \min_x \ f_0(x) \ : \quad \begin{array}{ll} f_i(x) \leq 0, & i = 1, \cdots, m, \\ h_i(x) = 0, & i = 1, \cdots, p, \end{array} \tag{8.1}$$

where the functions $f_0, f_1, \ldots, f_m$ are convex, and $h_1, \ldots, h_p$ are affine. We denote by $\mathcal{D}$ the domain of the problem (which is the intersection of the domains of all the functions involved), and by $\mathcal{X} \subseteq \mathcal{D}$ its feasible set.

To the problem we associate the Lagrangian $\mathcal{L} : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with values

$$\mathcal{L}(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x).$$

The dual function is $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with values

$$g(\lambda, \nu) := \min_x \ \mathcal{L}(x, \lambda, \nu).$$

The associated dual problem is

$$d^* = \max_{\lambda \geq 0, \, \nu} \ g(\lambda, \nu).$$

8.1.2 Strong duality via Slater's condition

Duality gap and strong duality. We have seen how weak duality allows to form a convex optimization problem that provides a lower bound on the original (primal) problem, even when the latter is non-convex. The *duality gap* is the non-negative number $p^* - d^*$.

We say that *strong duality holds* for problem (8.1) if the duality gap is zero: $p^* = d^*$.

Slater's condition. We say that the problem satisfies *Slater's condition* if it is *strictly feasible,* that is:

$$\exists\, x_0 \in \mathcal{D} \;:\; f_i(x_0) < 0,\;\; i = 1, \ldots, m,\;\; h_i(x_0) = 0,\;\; i = 1, \ldots, p.$$

We can replace the above by a *weak form* of Slater's condition, where strict feasibility is not required whenever the function f_i is affine.

We then have the

Theorem 1 (Strong duality via Slater condition). *If the primal problem (8.1) is convex, and satisfies the weak Slater's condition, then strong duality holds, that is, $p^* = d^*$.*

Note that there are many other similar results that guarantee a zero duality gap. For example:

Theorem 2 (Quadratic convex optimization problems). *If f_0 is quadratic convex, and the functions $f_1, \ldots, f_m, h_1, \ldots, h_p$ are all affine, then the duality gap is always zero, provided one of the primal or dual problems is feasible. In particular, strong duality holds for any feasible linear optimization problem.*

A counterexample. Convexity alone is not enough to guarantee strong duality. Consider for example the convex problem

$$\min_{x,y>0}\; e^{-x} \;:\; x^2/y \leq 0,$$

with variables x and y, and domain $\mathcal{D} = \{(x, y) \mid y > 0\}$. We have $p^* = 1$. The Lagrangian is $L(x, y, \lambda) = e^{-x} + \lambda x^2/y$, and the dual function is

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

so we can write the dual problem as

$$d^* = \max_{\lambda}\; 0 \;:\; \lambda \geq 0$$

with optimal value $d^\star = 0$. The optimal duality gap is $p^\star - d^\star = 1$. In this problem, Slater's condition is not satisfied, since $x = 0$ for any feasible pair (x, y).

8.1.3 Geometric interpretation

Assume that there is only one inequality constraint in (8.1) ($m = 1$), and let

$$\mathcal{A} := \{(u, t) \;:\; \exists\, x \in \mathbf{R}^n,\;\; u \geq f_1(x),\;\; t \geq f_0(x)\}\,.$$

The problem is feasible if and only if $\mathcal{A}$ intersects the left-half plane. Furthermore, we have

$$p^* = \min_{u,t} \; t \;:\; (u,t) \in \mathcal{A}, \;\; u \leq 0.$$

and

$$g(\lambda) = \min_{u,t} \; (\lambda, 1)^T (u,t) \;:\; (u,t) \in \mathcal{A}.$$

If the minimum is finite, then the inequality $(\lambda, 1)^T (u,t) \geq g(\lambda)$ defines a supporting hyperplane, with slope $-\lambda$, of $\mathcal{A}$ at (u,t). (See Figs. 5.3 and 5.4 in [BV,p.233].)

If the problem is convex, then $\mathcal{A}$ is also convex. If Slater's condition holds, then the interior of $\mathcal{A}$ intersects the left-half plane, and strong duality holds. (See Fig. 5.6 in [BV,p.236].)

8.2 Examples

8.2.1 Minimum Euclidean distance problem

The minimum distance to an affine set mentioned in lecture 11 is

$$\min \; \frac{1}{2}\|x\|_2^2 \;:\; Ax = b, \tag{8.2}$$

where $A \in \mathbf{R}^{p \times n}$, $b \in \mathbf{R}^p$. The problem is convex, and satisfies Slater's condition (in fact, strong duality always holds for this convex quadratic problem). Hence, we know that $p^* = d^*$. This allows us to compute the optimal value of the problem analytically: $p^* = d^* = \frac{1}{2} b^T (AA^T)^{-1} b$.

We can also find a corresponding optimal point: for every ν, the point $x(\nu) = -A^T \nu$ achieves the minimum in the definition of the dual function $g(\nu)$. Let us set $x^* := x(\nu^*)$, where $\nu^* = -(AA^T)^{-1} b$ denotes the optimal dual variable. The point $x^* = A^T (AA^T)^{-1} b$ is optimal for the primal problem. Indeed, it is feasible, since $Ax^* = A^T A (AA^T)^{-1} b = b$, and its objective value equals to the optimal value $(1/2)\|x^*\|_2^2 = \frac{1}{2} b^T (AA^T)^{-1} b = d^* = p^*$. Hence, x^* is optimal, as claimed.

8.2.2 Linear optimization problem

Consider the LP in inequality form:

$$p^* = \min_{x} \; c^T x \;:\; Ax \leq b,$$

where $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$. Assume that the above problem is feasible, so that strong duality holds. Then the problem can be *equivalently* written in the dual form, as an LP:

$$p^* = d^* = \max_{\lambda} \; -b^T \lambda \;:\; \lambda \geq 0, \;\; A^T \lambda + c = 0.$$

The above LP is in standard form, with the number of constraints and variables exchanged.

Duality is another way to convert any LP in inequality form into a standard form, and vice-versa. (The other method, seen in lecture 5, is via the introduction of new variables.)

8.2.3 Support vector machine classification

Return to the example seen in lecture 5, which involved a binary classification problem. Given m data points $x_i \in \mathbf{R}^n$, 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. Let us denote $Z = [y_1 x_1, \ldots, y_m x_m] \in \mathbf{R}^{n \times m}$.

Ideally, we would like to minimize the number of errors on the training set $(x_i, y_i)_{i=1}^m$. This is hard as it involves a non-convex function. An upper bound on the number of errors is provided by the so-called *hinge loss* function

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

We'd also like to control robustness of the resulting linear classifier, and at the same time guarantee unicity. It turns out that these objectives can be achieved via the following problem:

$$\min_{w,b} \; C \cdot L(w, b) + \frac{1}{2}\|w\|_2^2.$$

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).

The above can be written as a QP, by introducing slack variables:

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

or, more compactly:

$$\min_{w,b,v} \; \frac{1}{2}\|w\|_2^2 + Cv^T\mathbf{1} \;\; : \;\; v \geq 0, \;\; v + Z^T w + by \geq \mathbf{1}.$$

The corresponding Lagrangian is

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

where $\mu \in \mathbf{R}^m$ 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 $C\mathbf{1} = \lambda + \mu$, while taking derivatives with respect to b leads to the dual constraint $\lambda^T y = 0$. We obtain

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

We obtain the dual problem

$$d^* = \max_{\lambda \geq 0,\, \mu \geq 0} g(\lambda, \mu) = \max_{\lambda} \; \lambda^T \mathbf{1} - \frac{1}{2} \lambda^T Z^T Z \lambda \;\; : \;\; 0 \leq \lambda \leq C\mathbf{1}, \;\; \lambda^T y = 0.$$

Strong duality holds, since the primal problem is a QP.

Note that the result depends only on the so-called *kernel matrix* $K = Z^T Z \in \mathcal{S}_+^m$, and the dual problem involves only m variables and m 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 $x_i^T x_j$, $1 \leq i \leq j \leq m$. 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".

Note also that duality allows to show that the optimal value of the problem is a convex function of the kernel matrix, which allows to optimize over it. We will elaborate on this later.

8.3 Minimax equality theorems

8.3.1 Minimax inequality

As seen in lecture 7, weak duality can be obtained as a consequence of the *minimax inequality*, valid for *any* function ϕ of two vector variables x, y, and any subsets $\mathcal{X}$, $\mathcal{Y}$:

$$d^* := \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y) \leq \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y) := p^*. \tag{8.3}$$

Minimax equality theorems identify cases for which the equality $p^* = d^*$ can be proven.

8.3.2 Saddle points

A point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called a *saddle point* if

$$\forall\, x \in \mathcal{X}, \;\; \forall\, y \in \mathcal{Y} \;\; : \;\; \phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*).$$

The existence of saddle points is related to the minimax equality, as follows:

Proposition 3. *(x^*, y^*) is a saddle point if and only if the minimax equality holds, and is attained, in the sense that*

$$x^* \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y), \;\; y^* \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y).$$

8.3.3 A minimax equality theorem

It can be shown that the gap $p^* - d^*$ is zero if:

- $\mathcal{X}$, $\mathcal{Y}$ are both convex, and one of them is compact.
- The function ϕ is convex-concave: $\phi(\cdot, y)$ is convex for every $y \in \mathcal{Y}$, and $\phi(x, \cdot)$ is concave for every $x \in \mathcal{X}$.
- The function ϕ is continuous.

8.3.4 Examples

The LASSO problem The following is a penalized least-norm problem, where the penalty is the l_1-norm (which is known empirically to encourage a sparse solution):

$$p^* := \min_w \ \|X^T w - y\|_2 + \lambda \|w\|_1.$$

Here $X = [x_1, \ldots, x_n]$ is a $p \times n$ matrix of data points, $y \in \mathbf{R}^n$ is a "response" vector, and $\lambda > 0$ is a parameter. The higher the λ, the more zeroes we find in the solution w. A zero in the j-th position means that the j-th feature (row of X) is not used by the model. Thus the above approach allows us to learn which few features are important.

The above can be put in standard SOCP format, and we can form the dual that way (see lecture 9). More directly, we can form a dual based on the minimax expression:

$$p^* := \min_w \max_{u,v} \left\{ u^T(y - X^T w) + v^T w \ : \ \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \right\}.$$

We can exchange the min and max due to the convex-concave property of the objective function, and to the fact that the dual feasible set is compact. We obtain

$$\begin{aligned} p^* &= \max_{u,v} \min_w \left\{ u^T(y - X^T w) + v^T w \ : \ \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} u^T y \ : \ v = Xu, \ \ \|v\|_\infty \leq \lambda, \ \ \|u\|_2 \leq 1 \\ &= \max_u u^T y \ : \ \|Xu\|_\infty \leq \lambda, \ \ \|u\|_2 \leq 1. \end{aligned}$$

Denote by a_j, $j = 1, \ldots, p$ the j-th feature vector (corresponding to the j-th *row* in X). If $\|Xu\|_\infty < \lambda$ for every u with $\|u\|_2 \leq 1$, that is, if $\lambda > \max_{1 \leq j \leq p} \|a_j\|_2$, then the constraint $\|Xu\|_\infty \leq \lambda$ is inactive in the dual and

$$p^* = \max_{u \,:\, \|u\|_2 \leq 1} u^T y = \|y\|_2.$$

This means that $w = 0$ is optimal for the primal problem, since the value above is attained for $w = 0$. Note that $\sigma_j := (1/\sqrt{n})\|a_j\|_2$ is nothing else than the empirical variance of feature j.

In a cross-validation setting, we often to solve the problem for a range of λ-values. Duality allowed us to prove that we can safely restrict our search to the interval $\lambda \in [0, \sqrt{n}\sigma]$, where $\sigma := (1/\sqrt{n})\sigma_j$ is the maximum variance among all the features. For normalized data, such that $\|a_j\|_2 = 1$ for every j, the interval is $[0, 1]$.

Exercises

1. *Duality via the conjugate function: examples from supervised learning.* Many supervised learning problems can be expressed in the form

$$\min_w \ f(X^T w)$$

where the matrix $X \in \mathbf{R}^{n\times m}$ contains problem data, and $w \in \mathbf{R}^n$ contains (regression or classification) coefficients. Here, n is the number of features, while m is the number of observations. A few examples, in which $y \in \mathbf{R}^m$ is given:

- Least-squares: $f(z) = \|z - y\|_2^2$.
- Support vector machine (hinge loss) classification: $f(z) = \sum_{i=1}^m \max(0, 1 - y_i z_i)$.
- Logistic regression: $f(z, y) = \sum_{i=1}^m \log(1 + e^{-y_i z_i})$.

In this exercise you will derive a dual based on the conjugate function of f, defined as

$$f^*(u) := \max_z \ z^T u - f(z).$$

(a) Find the conjugate functions for the three examples above.

(b) Express the original (unconstrained) problem as a constrained one, involving a new variable $z := X^T w$.

(c) Take the Lagrange dual for the constrained problem, and show that it can be written as

$$\max_u \ -f^*(-u) \ : \ Xu = 0.$$

(d) Express the dual for the three cases mentioned above.

(e) What becomes of the dual is we add a penalty term $p(w) = \|w\|_1$ to the original objective function?

2. *QP duality.* Consider the QP

$$\min_x \ \|x\|_2^2 \ : \ Ax \leq b.$$

We assume that $A \in \mathbf{R}^{m\times n}$, with $n \geq m$.

(a) Show that we can always assume that the solution lies in the range of A^T.

(b) Show that the problem's solution depends only on $K := AA^T$, and that the problem's complexity is linear in n, the number of variables.

3. *Scenarios with incomplete distribution information.* Consider the convex problem

$$\min_{x\in\mathcal{X}} c^T x,$$

where $\mathcal{X}$ is convex, and in which the n-vector c is actually random, with values in the set $\{c^1,\ldots,c^N\}$, where the "scenarios" $c^i \in \mathbf{R}^n$, $i=1,\ldots,N$, are given. One approach to handle uncertainty in the cost vector c is to minimize the "worst-case" objective:

$$\min_{x\in\mathcal{X}} \max_{1\le i\le N} (c^i)^T x.$$

Let us develop a less pessimistic solution, which relies on additional knowledge about the distribution of c.

We assume that the (discrete) distribution of c, which we refer to as a vector p ($p \ge 0$, $\mathbf{1}^T p = 1$, with $\mathbf{1}$ the n-vector of ones), is also only partially known. Specifically, we assume that its "Kullback-Leibler distance" to the uniform distribution is bounded, as follows:

$$h(p) := -\sum_{i=1}^n p_i \log p_i \ge \gamma,$$

where $\gamma > 0$ is a measure of how "far" the distribution is from the uniform one. The function h is often referred to as the "entropy" of p; it is non-negative, and defined everywhere on its domain $\mathcal{P} := \{p \ge 0 \ : \ \mathbf{1}^T p = 1\}$, with the convention $0 \log 0 = 0$.

(a) Show that h is concave. Plot it in the case $n = 1$. In the case $n = 3$, plot the set defined by $p \in \mathcal{P}$, $h(p) \ge \gamma$, for various values of γ. (You can plot the projection of the set on $\mathcal{P}$.)

(b) Show that the entropy constraint is strictly feasible if $\gamma < \gamma_{\max}$, where

$$\gamma_{\max} := \max_{p\in\mathcal{P}} h(p).$$

Compute $\gamma_{\max}$, and the corresponding optimal distribution.

(c) Assume from now on that $\gamma < \gamma_{\max}$. For a given $x \in \mathbf{R}^n$, solve the problem

$$\max_{p\in\mathcal{P}} \mathbf{E}_p(c^T x) \ : \ h(p) \ge \gamma,$$

where the symbol $\mathbf{E}_p$ denotes the expectation with respect to p. Interpret this approach.

(d) Show that the problem

$$\min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}} \mathbf{E}_p(c^T x) \ : \ h(p) \ge \gamma,$$

can be written as

$$\min_{x\in\mathcal{X},\, \lambda>0} \ -\gamma\lambda + \lambda \log \sum_{i=1}^N e^{(c^i)^T x/\lambda}.$$

Is the above formulation convex? *Hint:* search for the term "perspective function" in BV.

(e) What becomes of the problem when $\gamma = \gamma_{\max}$? Interpret.

(f) Can you generalize the approach to the case when the objective is not linear, but convex? Discuss.