10-725/36-725: Convex Optimization **Spring 2015**

Lecture 13: February 25

Lecturer: Ryan Tibshirani *Scribes: Chen Kong, Chao Liu, Shanghang Zhang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

13.1 Dual Norm

Definition 13.1 *Let $\|x\|$ be a norm, then its dual norm $\|x\|_*$ is*

$$\|x\|_* = \max_{\|z\| \leq 1} z^T x.$$

Theorem 13.2 *If $\|x\|$ is a norm and $\|x\|_*$ is the dual norm of it, $\|z^T x| \leq \|z\| \|x\|_*$ holds.*

Let's use some examples to have a look at dual norms. The dual norm of l_p nor is l_q norm, i.e. $(\|x\|_p)_* = \|x\|_q$, where $1/p + 1/q = 1$. The dual norm of trace norm is operator norm i.e. $(\|X\|_{tr})_* = \|X\|_{op} = \sigma_1(X)$.

Theorem 13.3 *Dual norm of dual norm is the primal norm i.e. $\|x\|_{**} = \|x\|$.*

13.2 Conjugate Function

Definition: Given a function $f : \mathbb{R}^n \to \mathbb{R}$, its conjugate $f^* : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f^*(y) = \max_x y^T x - f(x)$$

Because $f^*(y)$ is the pointwise maximum of affine functions in y "indexed" by x, $f^*(y)$ is always convex in y.

Intuition in 1D: The intuition of the conjugate function of a 1D function $f(x)$ is as follows: For the function $f(x)$, given the slope y, we search along the x-axis to find out the x value that maximizes the difference between the line $g(x) = yx$ and the function $f(x)$. Once we have found the optimal x^*, we define a function with slope y and passing through $(x^*, f(x^*))$, the intercept of the function with the y-axis is $-f^*(y)$.

Note that the Conjugate functions does not give us anything new itself. But it enables us to derive the dual problems more quickly.

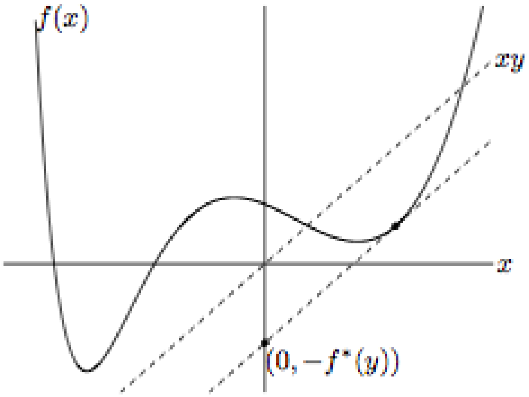

Figure 13.1: Conjugate function of $f(x)$

13.2.1 Properties of Conjugate function

- Fenchel's inequality: for any x and y, we have:

$$f(x) + f^*(y) \le x^T y$$

- For any function $f(x)$, the conjugate of conjugate function is no greater than the original function:

$$f^{**}(x) \le f(x)$$

- If $f(x)$ is closed and convex, then $f^{**}(x) = f(x)$
- If $f(x)$ is closed and convex, then

$$\begin{aligned} x \in \partial f^*(y) \quad &\Leftrightarrow \quad y \in \partial f(x) \\ &\Leftrightarrow \quad f(x) + f^*(y) = x^T y \end{aligned}$$

- If $f(u, v) = f_1(u) + f_2(v)$, with $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$, then

$$f^*(w, z) = f_1^*(w) + f_2^*(z)$$

13.2.2 Conjugate of some functions

Simple quadratic: $f(x) = \frac{1}{2}x^T Qx$ with $Q \succ 0$.

By taking the partial derivative of $y^T x - f(x)$ *w.r.t.* y and set it to zero, we have $y = Q^{-1}x$. Plugging that into $f^*(y) = y^T x - f(x)$, we have

$$f^*(y) = \frac{1}{2}y^T Q^{-1} y$$

The Fenchel's inequality gives the following inequality:

$$\frac{1}{2}x^T Qx + \frac{1}{2}y^T Q^{-1} y \ge x^T y$$

for $Q \succ 0$.

Indicator function: $f(x) = I_C(x)$.

Its conjugate is given by $f^*(y) = I^*_C(y) = \max\limits_{x \in C} y^T x$. This can be easily seen since $I_C(x) = -\infty$ for $x \notin C$.

$f^*(y) = \max\limits_{x \in C} y^T x$ is called the *support function* of C.

Norm function: $f(x) = \|f(x)\|$

Because the dual norm of dual norm is the original norm, *i.e.* $\|x\|_{**} = \|x\|$, we have:

$$f(x) = \max_{\|z\|_* \leq 1} z^T x$$

, from the definition of the dual norm.

We can see that $f(x)$ is the support function of set $\{z|\|z\|_* \leq 1\}$. We see from the last example that the conjugate of an indicator function is a support function, and the indicator function of a convex set is convex. So the conjugate of a support function is the indictor function. More specifically, we have:

$$f^*(y) = I_{\|z\|_* \leq 1}(y)$$

13.3 Lasso Dual

We now introduce the Lasso Dual problem to see why they are useful and how they come up. We will derive the lasso dual in this section. We are going to see a couple of things in this example. The first thing is the dual trick, and the second thing is the use of conjugate. The lasso problem is:

$$min_\beta \frac{1}{2}\|y - x\beta\|_2^2 + \lambda\|\beta\|_1 \tag{13.1}$$

How many variables does the dual function take? What is the dimension? It's a little tricky for there is no constraints for the function. It's equal to the minimum of the lasso function over all x. The second term of the constraints is 0 so there is no constraints. We are going to introduce some axiliant variables here to create fake constraints, which will give us dual variables to derive the dual. This is our first dual trick. You can do this in more than one way. Let's rewrite it in the following way.

$$\Leftrightarrow min_{z,\beta} \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \qquad s.t. \quad X\beta = z \tag{13.2}$$

Now the Lagrangian is as follows. Our primal variables are z and β. The dual variable is u.

$$L(z, \beta, u) = \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta) \tag{13.3}$$

Let's minimize that over all z and β to get the dual function to be the function of u. So we want to minimize the Lagrangian over all z and β by breaking it up as follows:

$$min_{z,\beta}\, L(z, \beta, u) = min_z\{\frac{1}{2}\|y - z\|_2^2 + u^T z\} + min_\beta\{\lambda\|\beta\|_1 - (X^T u)^T\beta\} \tag{13.4}$$

Minimize the first part over all z and take the gradient and let it equal to 0. We will get the minimizer here to be (y-u), and we plug this in the first part. The second part may be tricky, because it involves the L1 norm which is not differentiable. We rewrite it into the conjugate function:

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 - \lambda max_\beta\{\frac{(X^T u)^T}{\lambda}\beta - \|\beta\|_1\} \tag{13.5}$$

So we get the dual function as follows:

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 - I\{\|X^T u\|_\infty \leq \lambda\} \tag{13.6}$$

The lasso dual is to maximize (6) over all u, which is the same as (7)

$$min_u \frac{1}{2}\|y-u\|_2^2 \qquad s.t. \quad \|X^T u\|_\infty \leq \lambda \tag{13.7}$$

(7) is the lasso dual. It is easy because we use the conjugate trick during derivation. So after introducing the constrain, we write down the lasso's Lagrangian. Minimize it to get the dual function, and end up with the following problem:

$$max_u \frac{1}{2}(\|y\|_2^2 - \|y-u\|_2^2) \qquad s.t. \quad \|X^T u\|_\infty \leq \lambda \tag{13.8}$$

Let's think about the first point we made, which is that we can also get primal solutions from dual solutions. We can see the strong duality holds here. It does because if we go back to the primal problem, we have to look at (2). Because it is the problem whose dual we took with the fake constrain z. The strong duality holds between this problem and its dual, because it's equivalent to that problem. All we have to do it to find z and β such that $z = X\beta$. Slaters condition holds, and hence so does strong duality. The lasso's dual attains the same objective value as the lasso's primal does. But we should be careful here. If we maximize (7) over all u, we'll get $g(u^*) = f(x^8)$. It does not mean that if I minimize (6) over all u, the criteria value will equal to the f(*). (6) is just the transform version of the dual. (6) and (7) do not have the same criteria value. the optimal value of (6) is not the optimal lasso objective value.

The thing we care more about is how do we get the lasso solution from the dual solution. We go back to Lagrangian, which is

$$L(z, \beta, u) = \frac{1}{2}\|y-z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta) \tag{13.9}$$

If we were to plug in the optimal dual solution here, then both β and z are characterized by minimizing this function. Let's look at the minimizer over z first. The minimizer over z is $z = y - u$, which is $X\beta = y - u$. So given the dual solution u, any lasso solution β satisfies $X\beta = y - u$. This is from KKT stationarity condition for z. So the lasso fit is just the dual residual.

The dual is not only used for solving a problem, but also for understanding some problems' properties. Let's take a look at the dual problem here:

$$min_u \|y-u\|_2^2 \qquad s.t. \quad u \in C \tag{13.10}$$

What's this problem ask for? As illustrated in Figure 13.2, we have subset C and the dual solution given by projecting y onto C. Furthermore I claim C is actually polyhedron. You can explain it in the following ways:

$$C = \{u : \|X^T u\|_\infty \leq \lambda\} = (X^T)^{-1}\{V : \|V\|_\infty \leq \lambda\} \tag{13.11}$$

From our calculation of Lagrangian, the primal is actually the residual from this projection. So it's the vector y-u.We can use pretty simple geometry to make statements of lasso problem. This method has been used in a few papers. What's one thing we get immediately out of this is the non-trivial fact that the solution $x\beta(y)$ is Lipschitz continuous in y with Lipschitz constant L equals to one. Here is a simple fact of projection on convex set. If we were to project y on a convex set, and then I were to project some other vector y' to the convex set, then the distance between their projections is no bigger than the distance between y' and y. It's the property of convex set, ie. the projection is not expensive. So we have:

$$\|X\beta(y) - X\beta(y')\|_2 \leq \|y - y'\|_2 \quad y, y' \tag{13.12}$$

13.4 Conjugate and dual problems

Conjugates appear frequently in derivation of dual problems, via

$$f^*(u) = \min_x f(x) - u^T x,$$

in minimization of the Lagrangian.
Consider

$$\min_x f(x) + g(x)$$

$$\Leftrightarrow \min_{x,z} f(x) + g(z) \text{ subject to } x = z$$

Then Lagrange dual function is

$$g(u) = \min_{x,z} f(x) + g(z) + u^T(z - x) = -f^*(u) - g^*(-u)$$

Hence dual problem is

$$\max_u -f^*(u) - g^*(-u).$$

Example: Indicator function: dual of

$$\min_x f(x) + I_C(x)$$

is

$$\max_u -f^*(u) - I_C^*(-u),$$

where I_C^* is the support function of C.

Example: Norms: dual of

$$\min_x f(x) + \|x\|$$

is

$$\max_u -f^*(u) \text{ subject to } \|u\|_* \leq 1$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

13.5 Dual subtleties

Often, we will transform the dual into an equivalent problem and still call this the dual. Under strong duality, we can use solutions of the (transformed) dual problem to characterize or compute primal solutions. But the optimal value of this transformed dual problem is not necessarily the optimal primal value.

A common trick in deriving duals for unconstrained problems is to first transform the primal by adding a dummy variable and an equality constraint. Usually there is ambiguity in how to do this, and different choices lead to different dual problems!

Consider

$$\min \quad f(x) \quad \text{subject to} \quad h_i(x) \leq 0, i = 1, \dots, m \quad l_j(x) = 0, j = 1, \dots, r$$

If f and $h_1, \dots, h_m$ are closed and convex, and $l_1, \dots, l_r$ are affine, then the dual of the dual is the primal.

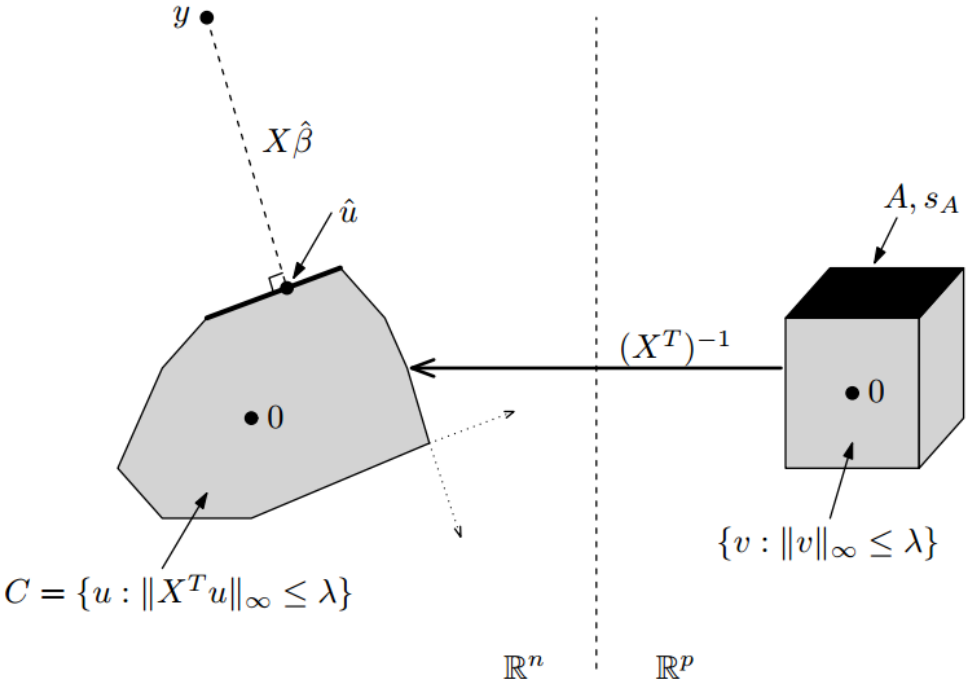

Figure 13.2: Lasso dual problem