10-725/36-725: Convex Optimization **Fall 2019**

Lecture 13: October 9

Lecturer: Ryan Tibshirani *Scribes: Montiel Abello, Cherie Ho, Sudharshan Suresh*

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

13.1 Review: KKT Conditions

Given a problem

$$\begin{aligned} \min_x & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & l_j(x) = 0, \quad j = 1, \ldots, r. \end{aligned}$$

The KKT conditions are:

- **Stationarity:** $0 \in \partial_x \Big(f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j l_j(x) \Big)$
- **Complementary Slackness:** $u_i \cdot h_i(x) = 0$ for all i
- **Primal Feasibility:** $h_i(x) \leq 0,\ l_j(x) = 0$ for all i, j
- **Dual Feasibility:** $u_i \geq 0$ for all i

The conditions are necessary for optimality under strong duality, and always sufficient.

13.2 Uses of Duality

1. **Duality gap can be used to check optimality and as a stopping criterion**:
   For x primal feasible, and u, v dual feasible,

   $$f(x) - f(x^*) \leq f(x) - g(u, v)$$

   RHS is the duality gap. Zero duality gap implies optimality.

2. **Duality can be used to characterize or compute primal solutions from dual solution**:
   Under strong duality, given dual optimal u^*, v^*, any primal solution x^* solves

   $$\min_x L(x, u^*, v^*)$$

13.2.1 When is dual easier?

Key facts about primal-dual relationship (some described below, some later):

- Dual has complementary **number of variables**: recall, number of primal constraints.
- Dual involves complementary **norms**: $\|\cdot\|$ becomes $\|\cdot\|_*$
- Dual has "identical" **smoothness**: L/m (Lipschitz constant of gradient by strong convexity parameter) is unchanged between f and its conjugate f^*
- Dual can "shift" **linear transformations** between terms. This leads to key idea: **dual decomposition**

13.2.2 Solving the primal via the dual

An important consequence of stationarity: under strong duality, given a dual solution u^*, v^*, any primal solution x^* solves

$$\min_x \; f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* l_j(x)$$

Often, solutions of this unconstrained problem can be expressed explicitly, giving an explicit **characterization** of primal solution from dual solutions.

Further, suppose the solution of this problem is unique; then it must be the primal solution x^*. This can be very helpful when the dual is easier to solve than the primal.

Example: Consider

$$\min_x \sum_{i=1}^n f_i(x_i) \;\; \text{subject to} \;\; a^T x = b$$

where each $f_i(x_i) = \frac{1}{2} c_i x_i^2$ (smooth and strictly convex). Its dual function is

$$\begin{aligned} g(v) &= \min_x \sum_{i=1}^n f(x_i) + v(b - a^T x) \\ &= bv + \sum_{i=1}^n \min x_i \{f_i(x_i) - a_i v x_i\} \\ &= bv - \sum_{i=1}^n f_i^*(a_i v), \end{aligned}$$

where each $f_i^*(y) = \frac{1}{2c_i} y^2$, called the conjugate of f_i.

Therefore the dual problem is

$$\max_v bv - \sum_{i=1}^n f_i^*(a_i v) \quad \Longleftrightarrow \quad \min_v \sum_{i=1}^n f_i^*(a_i v) - bv$$

This is a convex minimization problem with scalar variable - much easier to solve than primal.
Given v^*, the primal solution x^* solves

$$\min_x \sum_{i=1}^n \Big(f_i(x_i) - a_i v^* x_i \Big)$$

Strict convexity of each f_i implies that this has a unique solution, namely x^*, which we compute by solving $f_i'(x_i) = a_i v^*$ for each i. This gives $x_i^* = a_i v^*/c_i$.

13.3 Dual Norms

A general norm is defined as the form $\|x\|$, for example:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad \text{represents the } l_p \text{ norm for all } p \geq 1$$

$$\|X\|_{tr} = \sum_{i=1}^{r} \sigma_i(X) \qquad \text{represents the trace norm of a matrix, the sum of its singular values}$$

Definition 13.1 *The dual norm of x is expressed as:*

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

From this we get the inequality $|z^T x| \leq \|z\|\|x\|_*$, which is similar to the generalized Hölder's inequality. If we apply the definition to the l_p and trace norm:

$$(\|x\|_p)_* = \|x\|_q, \text{ where } \frac{1}{p} + \frac{1}{q} = 1$$
$$(\|X\|_{tr})_* = \|X\|_{op} = \sigma_1(X)$$

Theorem 13.2 *The dual of the dual norm, $\|x\|_{**} = \|x\|$*

Proof: We consider the problem:

$$\min_y \|y\| \quad \text{subject to } y = x$$

While this looks trivial, the result is actually not quite so. Its optimal value is $\|x\|$. Taking the Lagrangian:

$$L(y, u) = \|y\| + u^T(x - y) = \|y\| - y^T u + x^T u$$

By the definition of the dual norm $\|\cdot\|_*$:

$$\min_y\{\|y\| - y^T u\} = \begin{cases} -\infty & \|u\|_* > 1 \\ 0 & \|u\|_* \leq 1 \end{cases}$$

Thus, the Lagrangian dual is:

$$\max_u u^t x \quad \text{subject to } \|u\|_* \leq 1$$

This defines the dual of the dual norm, and by strong duality $f^* = g^* \implies \|x\| = \|x\|_{**}$ ■

13.4 Conjugate Functions

Definition 13.3 *For a function $f : \mathbb{R}^n \to \mathbb{R}$, a conjugate $f^* : \mathbb{R}^n \to \mathbb{R}$ is defined as:*

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

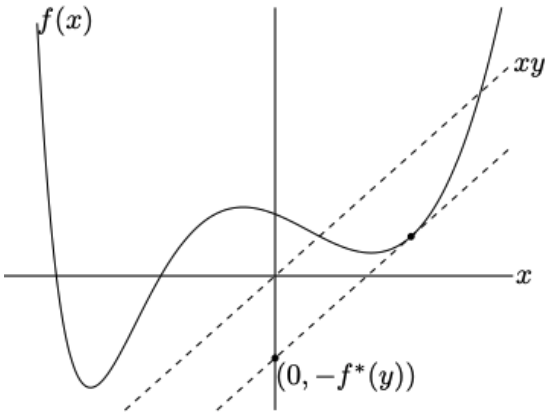

Figure 13.1: Illustration of the conjugate for a function $f(x)$ (from B&V Page 91)

The conjugate is always convex, even for non-convex f. This is because it is a pointwise maximum of convex (affine) functions in y. It can also be considered as the maximum difference between a linear function $y^T x$ and the function itself. If f is differentiable, the conjugate is called the Legendre transform.

13.4.1 Properties

- Fenchel's inequality for any x, y:
$$f(x) + f^*(y) \geq x^T y$$
- Taking the conjugate of a conjugate obeys $f^{**} \leq f$
- As a special case, if f is closed (sublevel sets are closed) and convex $f^{**} = f$
- If f is closed and convex, then for all x, y, Fenchel's inequality becomes a equality:
$$\begin{aligned} x \in \partial f^*(y) &\iff y \in \partial f(x) \\ &\iff f(x) + f^*(y) = x^T y \end{aligned}$$
- If $f(u, v) = f_1(u) + f_2(v)$, then we can split it into a term-by-term conjugation:
$$f^*(w, z) = f_1^*(w) + f_2^*(z)$$

13.4.2 Examples

- For a simple **quadratic example** $f(x) = \frac{1}{2} x^T Q x$, where $Q \succ 0$. Then:
$$f^*(y) = \max_x y^T x - \frac{1}{2} x^T Q x$$
$y^T x - \frac{1}{2} x^T Q x$ is strictly concave in y
$$= \frac{1}{2} y^T Q^{-1} y \qquad (x^* = Q^{-1} y)$$

- **Indicator function**: if $f(x) = I_c(x)$, then the conjugate is called the support function:

$$f^*(y) = I_c^*(y) = \max_{x \in C} \; y^T x$$

- **Norm**: if $f(x) = \|x\|$, then its conjugate is the indicator of the dual norm ball:

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

13.4.3 Example: Lasso Dual

This example shows how we can use conjugate functions to derive the dual problem, and a useful trick for "shifting" a linear transform from one part of the objective to another.

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$ the lasso problem is

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

It is not obvious at first what the dual problem will look like. The primal function has no constraints, so it's not clear how many dual variables we will have. We introduce auxiliary variables (here z) to create constraints, which will later give dual variables:

$$\min_{\beta, z} \quad \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \quad \text{subject to } z = X\beta$$

Now the Lagrangian is as below, with dual variable u and primal variables z, β:

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

We minimize the Lagrangian over z, β and use the conjugate of the L_1-norm derived previously.

$$\begin{aligned}
&\min_{\beta, z} \quad \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^\top (z - X\beta) \\
&= \min_z \left(\frac{1}{2}\|y - z\|_2^2 + u^\top z \right) + \min_\beta \left(\lambda\|\beta\|_1 + (X^\top u)^\top \beta \right) \\
&= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 + \min_\beta \lambda \left(\|\beta\|_1 - \frac{(X^\top u)^\top}{\lambda} \beta \right) \\
&= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 - \lambda I_{v:\|v\|_\infty \leq 1}(X^\top u / \lambda)
\end{aligned}$$

The indicator function is equivalent to the constraint $\|X^\top u\|_\infty \leq \lambda$, so the lasso dual problem is

$$\max_{u} \quad \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \quad \text{subject to } \|X^\top u\|_\infty \leq \lambda$$

Which is equivalent to

$$\min_u \quad \frac{1}{2}\|y-u\|_2^2 \quad \text{subject to } \|X^\top u\|_\infty \leq \lambda$$

The dual problem was derived by making the substitution $z = X\beta$, so the two are equivalent. Slater's condition holds, and therefore strong duality holds too. The dual attains the same objective value as the primal.

The KKT stationary condition for z gives us that at the dual solution, any lasso solution β satisfies $z = X\beta = y - u$

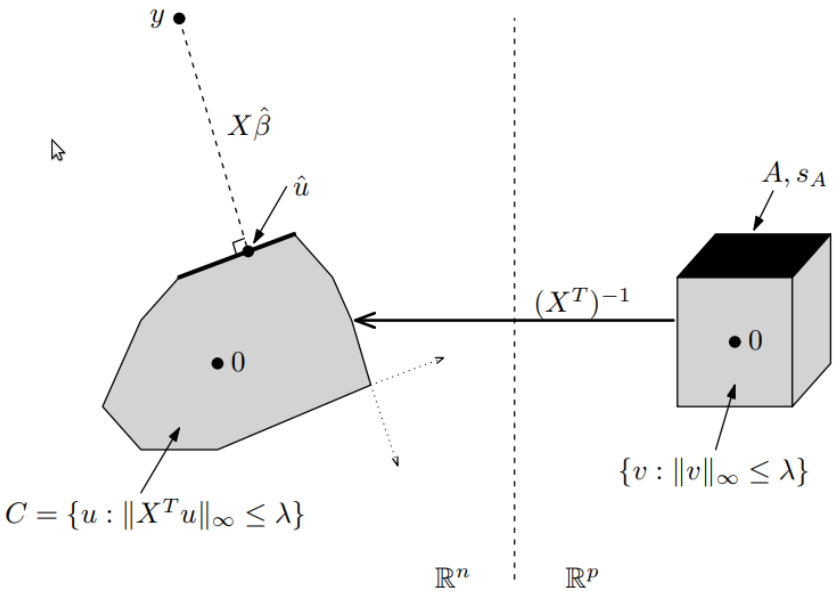

Figure 13.2: Visualization of Lasso Dual Problem, where its solution looks like projection of y onto set C, a polyhedron mapped from a hypercube.

The solution to the dual problem should look like projection of y onto the set $C = \{u : \|X^\top u\|_\infty \leq \lambda\} = (X^\top)^{-1}\{v : \|v\|_\infty \leq \lambda\}$. This concept is illustrated by Figure 13.2. The set C is a polyhedron which is mapped from a hypercube. This is the inverse image under linear map X. Every face of the polyhedron corresponds to an active set of signs that lasso would choose. When we project y onto a face, all y_i share the active set of signs of the lasso solution. Those active sets of signs are locally constant, and we see that lasso is stable as variable selector.

13.4.4 Relationship between conjugate and dual problems

The Lagrangian dual has a relationship to forming conjugates of functions. Conjugates appear frequently in the derivation of dual problems:

$$-f^*(u) = \min_x \quad f(x) - u^\top x$$

Examining a primal problem

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

As before, we make a substitution to introduce a constraint:

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

The Lagrange dual function is:

$$\begin{aligned} g(u) &= \min_x \quad f(x) + g(z) + u^\top (z - x) \\ &= -f^*(u) - g^*(u) \end{aligned}$$

So the dual problem is

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

Below are some of examples of the use of conjugates in deriving the dual problem. The indicator maps to a support function, and the norm (which is a special kind of support function) maps to the indicator:

- **Indicator function:**
  The dual of

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

  is

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

- **Norm:**
  The dual of

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

  is

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

13.4.5 Shifting linear transformation

We have seen in the previous section how a linear transformation can shift from one part of the objective to another in the dual formulation. For the problem

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

Which is equivalent to

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

The dual problem is

$$\max_u \quad -f^*(A^\top x) - g^*(-u)$$

This can often be a useful trick. If f is smooth but g is not, it is typically hard to solve the primal problem with methods such as projected gradient descent, proximal gradient descent, or the subgradient method. We can get around this with the dual. If f is smooth, f^* is also smooth. In the dual, though g is nonsmooth, projected gradient descent or proximal gradient descent are now possible.

13.5 Dual Subtleties

- Sometimes, like in the case of the Lasso Dual, we transform the dual to an equivalent problem but still call it the dual. With strong duality, we can use the solutions of this modified problem to compute primal solutions. **Note: The transformed dual's optimal value may not be the optimal value of the primal.**
- When presented with an unconstrained problem, a common trick is to add dummy variables with an equality constraint (e.g. Dual lasso). **This leads to an ambiguity in the process, we can theoretically get different duals for the same primal problem!** However, if the primal is convex, all the duals must give an equivalent solution, albeit with different solving difficulties.