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

Lecture 11: Duality in general programs

Lecturer: Ryan Tibshirani *Scribes: Ben Cowley, Calvin Murdock, Dallas Card*

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.*

11.1 Background

As demonstrated last lecture, a linear program can be transformed into a dual problem by introducing a dual variable for each constraint, as summarized below:

Given $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $G \in \mathbb{R}^{r\times n}$, $h \in \mathbb{R}^r$:

$$\begin{array}{rl|rl}
\displaystyle\min_{x\in\mathbb{R}^n} & c^T x & \displaystyle\max_{u\in\mathbb{R}^m, v\in\mathbb{R}^r} & -b^T u - h^T v \\
\text{s.t.} & Ax = b & \text{s.t.} & -A^T u - G^T v = c \\
 & Gx \le h & & v \ge 0 \\
 & & & \\
\multicolumn{2}{c|}{\text{Primal LP}} & \multicolumn{2}{c}{\text{Dual LP}}
\end{array}$$

In these notes, we will see that we can derive a very similar dual problem for a general optimization problem using the Lagrangian. This allows us to define, for a general optimization problem (even a non-convex one), a dual problem which *is* convex, and the solution to which provides a lower-bound on the solution to the primal problem.

11.2 The Lagrangian

Consider the general constrained minimization problem

$$\begin{array}{ll}
\min & f(x) \\
\text{s.t.} & h_i(x) \le 0,\ i = 1, ...m \\
 & l_j(x) = 0,\ j = 1, ...r
\end{array}$$

Introduce new variables $u \in \mathbb{R}^m$, and $v \in \mathbb{R}^r$, with $u \ge 0$, and define the **Lagrangian** to be

$$L(x, u, v) := f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x)$$

(implicitly, $L(x, u, v) = -\infty$ if $u < 0$)

Observe that for **feasible** x and $u \ge 0$, $l_j(x) = 0$ and $u_i h_i(x) \le 0$; thus

$$L(x,u,v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x) \leq f(x)$$

This is illustrated in Figure 11.1 below.

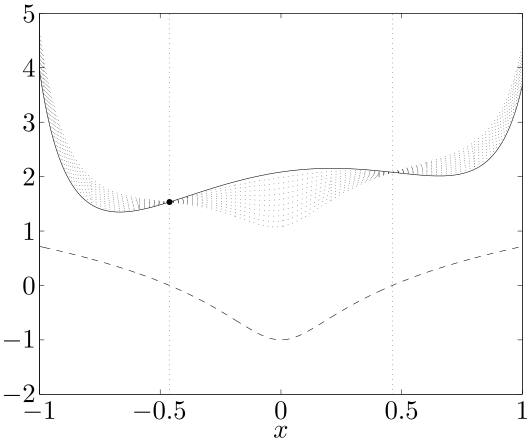

Figure 11.1: Lagrangian lower bound: solid line is $f(x)$; dashed line is a non-convex inequality constraint ($h(x) \leq 0$); the feasible region is $x \in [-0.46, 0.46]$, with an optimal value of $x = -0.46$. The dotted lines show $L(x, \lambda)$ for different values of the dual variable $\lambda \geq 0$. Note that $L(x, \lambda) \leq f(x)$ for all feasible x.

Now define the **Lagrange dual function** to be:

$$g(u,v) := \min_{x} \; L(x,u,v)$$

and observe that we have:

$$f^* \geq \min_{x \in C} \; L(x,u,v) \geq \min_{x} \; L(x,u,v) := g(u,v)$$

Thus, $g(u,v)$ gives a lower bound on f^* for any $u \geq 0$ and v. This is illustrated in Figure 11.2 below:

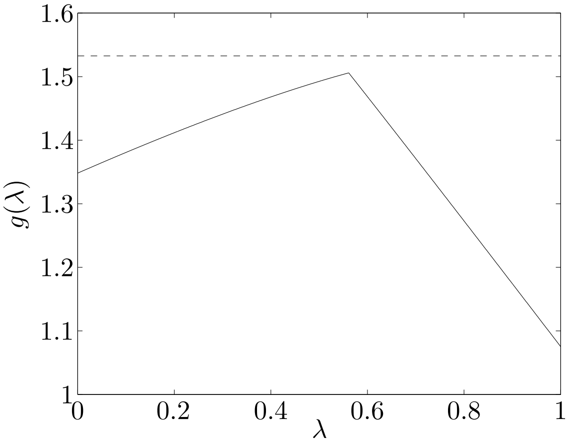

Figure 11.2: Lagrangian dual function: This figure shows the Lagrangian dual function, $g(\lambda)$, (solid line) for the problem shown in Figure 11.1, as a function of the dual variable λ. The optimal value of the primal objective function is shown by the dashed line. Note that neither the objective function nor the constraint is convex in the original problem, but the Lagrangian dual function is nevertheless concave.

Now that we have a lower bound on f^* in terms of the dual variables, u and v, we can obtain the best possible bound by maximizing the Lagrangian dual function:

$$\begin{array}{ll} \max\limits_{u,v} & g(u,v) \\ \text{s.t.} & u \geq 0 \end{array}$$

Based on the above, we are guaranteed to have **weak duality**, namely

$$f^* \geq g^*$$

where g^* is the optimal value of the dual maximization problem. Note that this holds even if the original problem is not convex, (as in Figure 11.1 and 11.2).

In addition, the resulting dual problem will always be a convex optimization problem (again, even if the original problem is not convex), as the dual objective can be written as a pointwise maximum of convex functions in (u,v), which is guaranteed to be convex:

$$g(u,v) = -\max_x \left[-L(x,u,v)\right] = -\max_x \left[-f(x) - \sum_{i=1}^m u_i h_i(x) - \sum_{j=1}^r v_j l_j(x)\right]$$

11.3 Example: Quadratic program

Consider the following quadratic program with $Q \succeq 0$:

$$\begin{array}{ll} \max\limits_{x\in\mathbb{R}^n} & \frac{1}{2}x^T Q x + c^T x \\ \text{s.t.} & Ax = b,\, x \geq 0 \end{array}$$

The Lagrange dual function is:

$$\begin{aligned} g(u,v) &= \min_x \ L(x,u,v) \\ &= \min_x \ \frac{1}{2}x^TQx + c^Tx + u^T(-x) + v^T(Ax-b) \end{aligned}$$

Differentiating gives $Qx + c - u + A^Tv = 0$. Thus, $x = Q^+(-c + u - A^Tv)$, where Q^+ is the generalized inverse of Q. Substituting this into the Lagrange dual function gives:

$$g(u,v) = -\frac{1}{2}(c - u + A^Tv)Q^+(c - u + A^Tv) - b^Tv$$

Note, however, that if $(c-u+A^Tv)$ is in the null space of Q (i.e. $Q(c-u+A^Tv) = 0$), then $g(u,v) = -b^Tv$, which, since v is unconstrained, could take on arbitrary values. Thus, we effectively have an additional constraint on the dual problem, and the Lagrangian dual function becomes:

$$g(u,v) = \begin{cases} -\frac{1}{2}(c - u + A^Tv)Q^+(c - u + A^Tv) - b^Tv & \text{if } c - u + A^Tv \perp \text{null}(Q) \\ -\infty & \text{otherwise} \end{cases}$$

If Q is positive-definite ($Q \succ 0$), then $x^TQx > 0$ and we can drop this condition.

11.4 Strong Duality and Slater's Condition

Recall that we always have weak duality ($f^* \leq g^*$).

However, when $f^* = g^*$, we call it **strong duality**.

This is motivated by Slater's condition:

If the primal problem is convex *and* there exists at least one $x \in \mathbb{R}^n$ that is strictly feasible (i.e., the convex inequality constraints are strictly negative: $h_i(x) < 0$, $i = 1, \ldots, m$), then strong duality holds.

For linear programs, the following hold:

1. The dual of the dual is the primal problem.
2. Strong duality holds if the primal problem is feasible.
3. Likewise, strong duality holds if the dual problem is feasible.
4. Thus, strong duality holds for linear programs, except when infeasible.

11.5 Example: Support Vector Machines

In this example, we derive the dual form of SVM. Note that the goal of SVM is to maximize the margin width $1/\beta$, and thus minimize β (and allowing for some errors, if the data is not linearly separable).

Consider the SVM problem:

Given $y_i \in \{-1,1\}, \mathbf{X} \in \mathbb{R}^{n\times p}$ (with rows $x_1,\ldots,x_n \in \mathbb{R}^{1\times p}$),

$$\begin{array}{ll} \min\limits_{\beta,\beta_0,\xi} & \frac{1}{2}\|\beta\|_2^2 + C\sum\limits_{i=1}^n \xi_i \\ \text{s.t.} & \xi_i \geq 0,\ i=1,\ldots,n \\ & y_i(x_i\beta+\beta_0) \geq 1-\xi_i,\ i=1,\ldots,n \end{array}$$

where $\|\beta\|_2^2$ minimizes the margin while $C\sum\limits_{i=1}^n \xi_i$ determines how many errors are allowed (i.e., how many support vectors are allowed within the margin).

We rewrite the primal problem as:

$$\begin{array}{ll} \min\limits_{\beta,\beta_0,\xi} & \frac{1}{2}\|\beta\|_2^2 + C\sum\limits_{i=1}^n \xi_i \\ \text{s.t.} & -\xi_i \leq 0,\ i=1,\ldots,n \\ & 1-\xi_i-y_i(x_i\beta+\beta_0) \leq 0,\ i=1,\ldots,n \end{array}$$

Now, we introduce dual variables v (where $v_i \geq 0$ corresponds to the $-\xi_i \leq 0$ constraints) and w (where $w_i \geq 0$ corresponds to the $1-\xi_i-y_i(x_i\beta+\beta_0) \leq 0$ constraints).

Recall to formulate the dual problem, we first compute the Lagrangian function, then minimize it with respect to the primal variables. The Lagrangian $\mathcal{L}$ is:

$$\mathcal{L}(\beta,\beta_0,\xi,v,w) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n v_i(-\xi_i) + \sum_{i=1}^n w_i(1-\xi_i-y_i(x_i\beta+\beta_0)$$

This can be reformulated as:

$$\mathcal{L}(\beta,\beta_0,\xi,v,w) = \frac{1}{2}\|\beta\|_2^2 - \sum_{i=1}^n w_iy_ix_i\beta - \sum_{i=1}^n w_iy_i\beta_0 + \sum_{i=1}^n (C-v_i-w_i)\xi_i + \sum_{i=1}^n w_i$$

We can minimize the Lagrangian with respect to β, β_0, and ξ separately. Also, for ease of notation, let $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$, where the ith row is y_ix_i.

Minimizing $\mathcal{L}$ w.r.t. β:
Note the only terms that $\mathcal{L}$ depends on β is

$$\min_\beta \frac{1}{2}\beta^T\beta - w^T\tilde{\mathbf{X}}\beta$$

Taking the gradient and setting it equal to zero yields the β minimizer:

$$\beta = w^T\tilde{\mathbf{X}}\beta$$

Minimizing $\mathcal{L}$ w.r.t. β_0:

$$\min_{\beta_0} -\sum_{i=1}^{n} w_i y_i \beta_0$$

Taking the gradient and setting it equal to zero yields a necessary constraint:

$$\sum_{i=1}^{n} w_i y_i = 0$$

Finally, minimizing $\mathcal{L}$ w.r.t. ξ:

$$\min_{\xi} \sum_{i=1}^{n} (C - v_i - w_i)\xi_i$$

Taking the gradient and setting it equal to zero yields another constraint:

$$C - v_i - w_i = 0, i = 1, \ldots, n \Rightarrow w_i = C - v_i, \text{ where } v_i \geq 0$$

Note that since w_i depends on v_i, we can remove v_i as a slack variable:

$$\Rightarrow w_i \leq C, i = 1, \ldots, n$$

Plugging in all these minimizers back into the Lagrangian (and thus minimizing the Lagrangian) yields:

$$g(w) = \frac{1}{2} w^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} w + w^T \mathbf{1}$$

Since $g(w)$ is putting a lower bound on the primal objective function, we would like to maximize it, with consideration of the constraints we used:

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2}\|\tilde{\mathbf{X}}^T w\|_2^2 + w^T\mathbf{1} \\ \text{s.t.} & 0 \leq w_i \leq C,\, i = 1, \ldots, n \\ & w^T y = 0 \end{array}$$

You can check that Slater's condition is satisfied, and that we have strong duality. We will also shortly learn that $\beta = w^T\tilde{\mathbf{X}}$ (which was the minimizer of the Lagrangian with respect to β). This gives us a mapping from the dual solution to the primal solution (i.e., we can solve the dual problem, which may be easier, and then easily map the dual optimal solution to the primal optimal solution). This correspondence comes from the KKT conditions.

11.6 Duality Gap

Given primal feasible x and dual feasible u, v:

$$f(x) - g(u, v) \text{ is called the duality gap}$$

Note that

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

This implies that if the duality gap $f(x) - g(u, v)$ is zero, then we've reached an optimal primal solution $(f(x) - f(x^*) = 0)$. Also, u, v are dual optimal.

The duality gap has a direct algorithmic use: if $f(x) - g(u, v) \leq \epsilon$, then we know $f(x) - f(x^*)$ (i.e., we know how close we are to the optimal solution). This is different from a criterion threshold method $(\frac{f(x^{k+1}) - f(x^k)}{x^k})$, which is a measure of the function value drop, not how close we are to the optimal value.

11.7 Dual Norms

Consider a norm $\|x\|$.

For example,

$$\ell_p \text{ norm: } \|x\|_p = \big(\sum_{i=1}^{n} |x_i|^p\big)^{1/p}, \text{ for } p \geq 1$$

$$\text{trace norm: } \|\mathbf{X}\|_{\text{tr}} = \sum_{i=1}^{r} \sigma_i(\mathbf{X})$$

We can define the dual norm $\|x\|_*$ as:

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

Dual norms can be useful in proofs, such as proving subgradients for particular norms. We can derive a Cauchy-Schwartz-like inequality from this definition:

$$\frac{|z^T x|}{\|z\|} \leq \|x\|_* \Rightarrow |z^T x| \leq \|z\| \|x\|_*$$

The example norms also have duals:

$$\ell_p \text{ norm dual: } (\|x\|_p)_* = \|x\|_q, \text{ where } \frac{1}{p} + \frac{1}{q} = 1$$

$$\text{Trace norm dual: } (\|\mathbf{X}\|_{\text{tr}})_* = \|\mathbf{X}\|_{\text{op}} = \sigma_{\max}(\mathbf{X})$$

Also note that the dual norm of the dual norm is the norm $(\|x\|_*)_* = \|x\|$.

References

[BV04] S. Boyd and L. Vandenberghe (2004), “Convex optimization”, Chapter 5.

[R70] R. T. Rockafellar (1970), “Convex analysis”, Chapters 28-30.