Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt
Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz
Email: msimchow+ee227c@berkeley.edu

October 15, 2018

15 Fenchel duality and algorithms

In this section, we introduce the Fenchel conjugate. First, recall that for a real-valued convex function of a single variable $f(x)$, we call $f^*(p) := \sup_x px - f(x)$ its Legendre transform. This is illustrated in figure 1.

The generalization of the Legendre transform to (possibly nonconvex) functions of multiple variables is the Fenchel conjugate.

Definition 15.1 (Fenchel conjugate)**.** The Fenchel conjugate of $f : \mathbb{R}^n \to \mathbb{R}$ is

$$f^*(p) = \sup_x \langle p, x \rangle - f(x)$$

We now present several useful facts about the Fenchel conjugate. The proofs are left as an exercise.

Fact 15.2. *f^* is convex.*

Indeed, f^* is the supremum of affine functions and therefore convex. Thus, the Fenchel conjugate of f is also known as its convex conjugate.

Fact 15.3. *$f^*(f^*(x)) = f$ if f is convex.*

In other words, the Fenchel conjugate is its own inverse for convex functions. Now, we can also relate the subdifferential of a function to that of its Fenchel conjugate. Intuitively, observe that $0 \in \partial f^*(p) \iff 0 \in p - \partial f(x) \iff p \in \partial f(x)$. This is summarized more generally in the following fact.

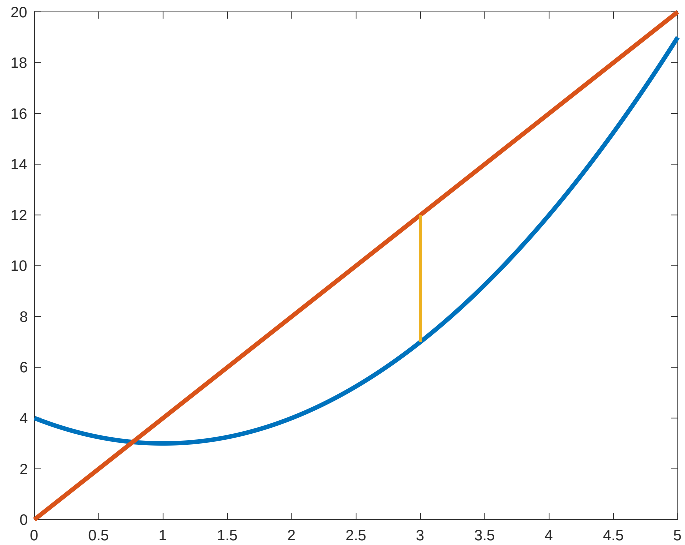

Figure 1: Legendre Transform. The function $f^*(p)$ is how much we need to vertically shift the epigraph of f so that the linear function px is tangent to f at x.

Fact 15.4. *The subdifferential of f^* at p is $\partial f^*(p) = \{x : p \in \partial f(x)\}$.*

Indeed, $\partial f^*(0)$ is the set of minimizers of f.

In the following theorem, we introduce Fenchel duality.

Theorem 15.5 (Fenchel duality)**.** *Suppose we have f proper convex, and g proper concave. Then*

$$\min_x f(x) - g(x) = \max_p g^*(p) - f^*(p)$$

where g^* is the concave conjugate of g, defined as $\inf_x \langle p, x \rangle - g(x)$.

In the one-dimensional case, we can illustrate Fenchel duality with Figure 2.

In the minimization problem, we want to find x such that the vertical distance between f and g at x is as small as possible. In the (dual) maximization problem, we draw tangents to the graphs of f and g such that the tangent lines have the same slope p, and we want to find p such that the vertical distance between the tangent lines is as large as possible. The duality theorem above states that strong duality holds, that is, the two problems have the same solution.

We can recover Fenchel duality from Lagrangian duality, which we have already studied. To do so, we need to introduce a constraint to our minimization problem in Theorem 15.5. A natural reformulation of the problem with a constraint is as follows.

$$\min_{x,z} f(x) - g(z) \text{ subject to } x = z \tag{1}$$

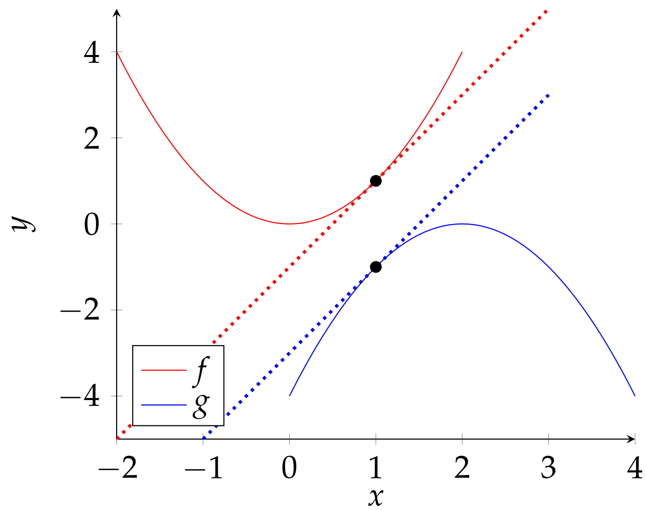

Figure 2: Fenchel duality in one dimension

15.1 Deriving the dual problem for empirical risk minimization

In empirical risk minimization, we often want to minimize a function of the following form:

$$P(w) = \sum_{i=1}^{m} \phi_i(\langle w, x_i \rangle) + R(w) \tag{2}$$

We can think of $w \in \mathbb{R}^n$ as the model parameter that we want to optimize over (in this case it corresponds to picking a hyperplane), and x_i as the features of the i-th example in the training set. $\phi_i(\cdot, x_i)$ is the loss function for the i-th training example and may depend on its label. $R(w)$ is the regularizer, and we typically choose it to be of the form $R(w) = \frac{\lambda}{2}\|w\|^2$.

The primal problem, $\min_{w \in \mathbb{R}^n} P(w)$, can be equivalently written as follows:

$$\min_{w,z} \sum_{i=1}^{m} \phi_i(z_i) + R(w) \text{ subject to } X^\top w = z \tag{3}$$

By Lagrangian duality, we know that the dual problem is the following:

$$\begin{aligned}
&\max_{\alpha \in \mathbb{R}^m} \min_{z,w} \sum_{i=1}^{m} \phi_i(z_i) + R(w) - \alpha^\top (X^\top w - z) \\
=& \max_{\alpha} \min_{w,z} \sum_{i=1}^{m} \phi_i(z_i) + \alpha_i z_i + R(w) - \alpha^\top X^\top w \\
=& \max_{\alpha} \left(- \min_{w,z} - \left(\sum_{i=1}^{m} \phi_i(z_i) + \alpha_i z_i \right) + (X\alpha)^\top w - R(w) \right) \\
=& \max_{\alpha} - \left(\sum_{i=1}^{m} \max_{z_i} \left(-\phi_i(z_i) - \alpha_i z_i \right) + \max_{w} (X\alpha)^\top w - R(w) \right) \\
=& \max_{\alpha} - \sum_{i=1}^{n} \phi_i^*(-\alpha_i) - R^*(X\alpha)
\end{aligned}$$

where ϕ_i^* and R^* are the Fenchel conjugates of ϕ_i and R^* respectively. Let us denote the dual objective as $D(\alpha) = \sum_{i=1}^{m} \phi_i^*(-\alpha_i) - R^*(X\alpha)$. By weak duality, $D(\alpha) \leqslant P(w)$.

For $R(w) = \frac{\lambda}{2}\|w\|^2$, $R^*(p) = \frac{\lambda}{2}\|\frac{1}{\lambda}p\|^2$. So R is its own convex conjugate (up to correction by a constant factor). In this case the dual becomes:

$$\max_{\alpha} \sum_{i=1}^{m} \phi_i^*(-\alpha_i) - \frac{\lambda}{2}\|\frac{1}{\lambda}\sum_{i=1}^{m} \alpha_i x_i\|^2$$

We can relate the primal and dual variables by the map $w(\alpha) = \frac{1}{\lambda}\sum_{i=1}^{m} \alpha_i x_i$. In particular, this shows that the optimal hyperplane is in the span of the data. Here are some examples of models we can use under this framework.

Example 15.6 (Linear SVM)**.** We would use the hinge loss for ϕ_i. This corresponds to

$$\phi_i(w) = \max(0, 1 - y_i x_i^\top w)\,, \quad -\phi_i^*(-\alpha_i) = \alpha_i y_i$$

Example 15.7 (Least-squares linear regression)**.** We would use the squared loss for ϕ_i. This corresponds to

$$\phi_i(w) = (w^\top x_i - y_i)^2\,, \quad -\phi_i^*(-\alpha_i) = \alpha_i y_i + \alpha^2/4$$

We end with a fact that relates the smoothness of ϕ_i to the strong convexity of ϕ_i^*.

Fact 15.8. *If ϕ_i is $\frac{1}{\gamma}$-smooth, then ϕ_i^* is γ-strongly convex.*

15.2 Stochastic dual coordinate ascent (SDCA)

In this section we discuss a particular algorithm for empirical risk minimization which makes use of Fenchel duality. Its main idea is picking an index $i \in [m]$ at random, and then solving the dual problem at coordinate i, while keeping the other coordinates fixed.

More precisely, the algorithm performs the following steps:

1. Start from $w^0 := w(\alpha^0)$

2. For $t = 1, \ldots, T$:

   (a) Randomly pick $i \in [m]$

   (b) Find $\Delta\alpha_i$ which maximizes

$$-\Phi_i\left(-(\alpha_i^{t-1} + \Delta\alpha_i)\right) - \frac{\lambda}{2m}\left\|w^{t-1} + \frac{1}{\lambda}\Delta\alpha_i x_i\right\|^2$$

3. Update the dual and primal solution

   (a) $\alpha^t = \alpha^{t-1} + \Delta\alpha_i$

   (b) $w^t = w^{t-1} + \frac{1}{\lambda}\Delta\alpha_i x_i$

For certain loss functions, the maximizer $\Delta\alpha_i$ is given in closed form. For example, for hinge loss it is given explicitly by:

$$\Delta\alpha_i = y_i \max\left(0, \min(1, \frac{1 - x_i^T w^{t-1} y_i}{\|x_i\|^2/\lambda m} + \alpha_i^{t-1} y_i)\right) - \alpha_i^{t-1},$$

and for squared loss it is given by:

$$\Delta\alpha_i = \frac{y_i - x_i^T w^{t-1} - 0.5\alpha_i^{t-1}}{0.5 + \|x\|^2/\lambda m}.$$

Note that these updates require both the primal and dual solutions to perform the update.

Now we state a lemma given in [SSZ13], which implies linear convergence of SDCA. In what follows, assume that $\|x_i\| \leqslant 1$, $\Phi_i(x) \geqslant 0$ for all x, and $\Phi_i(0) \leqslant 1$.

Lemma 15.9. *Assume Φ_i^* is γ-strongly convex, where $\gamma > 0$. Then:*

$$\mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})] \geqslant \frac{s}{m}\mathbb{E}[P(w^{t-1}) - D(\alpha^{t-1})],$$

where $s = \frac{\lambda m \gamma}{1 + \lambda m \gamma}$.

We leave out the proof of this result, however give a short argument that proves linear convergence of SDCA using this lemma. Denote $\epsilon_D^t := D(\alpha^*) - D(\alpha^t)$. Because the dual solution provides a lower bound on the primal solution, it follows that:

$$\epsilon_D^t \leqslant P(w^t) - D(\alpha^t).$$

Further, we can write:

$$D(\alpha^t) - D(\alpha^{t-1}) = \epsilon_D^{t-1} - \epsilon_D^t.$$

By taking expectations on both sides of this equality and applying Lemma 15.9, we obtain:

$$\begin{aligned}\mathbb{E}[\epsilon_D^{t-1} - \epsilon_D^t] &= \mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})] \\ &\geqslant \frac{s}{m}\mathbb{E}[P(w^{t-1}) - D(\alpha^{t-1})] \\ &\geqslant \frac{s}{m}\mathbb{E}[\epsilon_D^{t-1}].\end{aligned}$$

Rearranging terms and recursively applying the previous argument yields:

$$\mathbb{E}[\epsilon_D^t] \leqslant (1 - \frac{s}{m})\mathbb{E}[\epsilon_D^{t-1}] \leqslant (1 - \frac{s}{m})^t \epsilon_D^0.$$

From this inequality we can conclude that we need $O(m + \frac{1}{\lambda\gamma}\log(1/\epsilon))$ steps to achieve ϵ dual error.

Using Lemma 15.9, we can also bound the primal error. Again using the fact that the dual solution underestimates the primal solution, we provide the bound in the following way:

$$\begin{aligned}\mathbb{E}[P(w^t) - P(w^*)] &\leqslant \mathbb{E}[P(w^t) - D(\alpha^t)] \\ &\leqslant \frac{m}{s}\mathbb{E}[D(\alpha^{t+1}) - D(\alpha^t)] \\ &\leqslant \frac{m}{s}\mathbb{E}[\epsilon_D^t],\end{aligned}$$

where the last inequality ignores the negative term $-\mathbb{E}[\epsilon_D^{t-1}]$.

References

[SSZ13] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.