10-725: Convex Optimization **Spring 2023**

Lecture 12: February 23rd

Lecturer: Siva Balakrishnan

Today's lecture will focus on the Fenchel conjugate function, and the important role it plays in duality (sometimes called Fenchel duality).

12.1 Fenchel Conjugate

Suppose we're given a function f, then the Fenchel conjugate is defined to be:

$$f^*(y) = \sup_{x \in \text{dom}(f)} (y^T x - f(x)).$$

Some intuition: One way to understand this definition is that for any given slope y we could ask for the "highest" line that minorizes (i.e. lives below) the function, i.e. we could ask for the line with largest possible intercept. This intercept is exactly $-f^*(y)$. To see this, suppose we require that,

$$f(x) \geq y^T x - b, \quad \text{for all } x \in \text{dom}(f).$$

Several scalars b could satisfy this condition, in which case, we would like the line with largest intercept, i.e. we would try to make b as small as possible. We see that,

$$b \geq y^T x - f(x), \quad \text{for all } x \in \text{dom}(f),$$

so the smallest value that works is to set,

$$b = \sup_{x \in \text{dom}(f)} [y^T x - f(x)] = f^*(y).$$

Alternatively, this is saying that for any given slope y, the intercept of the largest affine minorant of f with slope y is $-f^*(y)$. Here is a figure from BV which summarizes this discussion:

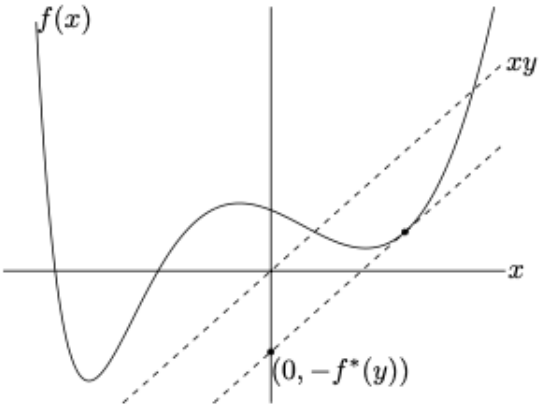

Figure 3.8 A function $f : \mathbf{R} \to \mathbf{R}$, and a value $y \in \mathbf{R}$. The conjugate function $f^*(y)$ is the maximum gap between the linear function yx and $f(x)$, as shown by the dashed line in the figure. If f is differentiable, this occurs at a point x where $f'(x) = y$.

An important property of the Fenchel conjugate is that irrespective of whether f is convex or not f^* is convex. This is because f^* is the pointwise maximum/supremum of affine functions.

12.1.1 Examples

Here are some nice examples to build some intuition.

Linear: Suppose that $f(x) = a^T x + b$. In this case, we would evaluate:

$$f^*(y) = \sup_x [x^T(y - a) - b].$$

So we conclude that,

$$f^*(y) = \begin{cases} -b, & \text{if } y = a \\ \infty, & \text{otherwise.} \end{cases}$$

Quadratic: Suppose that $f(x) = \frac{1}{2}x^T Q x$, with $Q \succ 0$. Then we can compute:

$$f^*(y) = \sup_x [x^T y - \frac{1}{2}x^T Q x] := \frac{1}{2}y^T Q^{-1} y,$$

just by taking the derivative and setting it to 0. Lets briefly notice a curious fact (that will turn out to be much more generally true).

$$\nabla f(x) = Qx$$
$$\nabla f^*(y) = Q^{-1}y,$$

so we have that $\nabla f^*(\nabla f(x)) = x$, and $\nabla f(\nabla f^*(y)) = y$, i.e. the gradient of the conjugate is the inverse of the gradient of the function.

Norm: Suppose that $f(x) = \|x\|$ for some norm. Then we observe that,

$$f^*(y) = \sup_x [x^T y - \|x\|] = \begin{cases} 0 & \text{if } \|y\|_* \leq 1 \\ \infty & \text{otherwise.} \end{cases}$$

In words, the conjugate is the convex indicator of the unit ball of the dual norm. To see this, observe that if $\|y\|_* \leq 1$, then by the Cauchy-Schwarz inequality, we know that $x^T y \leq \|x\|\|y\|_* \leq \|x\|$ so the term in the supremum is always ≤ 0, so the supremum corresponds to setting $x = 0$.

On the other hand, if $\|y\|_* > 1$, then we know that there must be some z, with $\|z\| \leq 1$ such that, $z^T y > 1$. Suppose in our optimization problem we take $x = tz$ then we get,

$$f^*(y) \geq t(z^T y - \|z\|),$$

and so letting $t \to \infty$, gives that $f^*(y) = \infty$.

Convex Indicator: Suppose that $f(x) = I_C(x)$, i.e. the function is a convex indicator, then:

$$f^*(y) = \sup_x [x^T y - f(x)] = \sup_{x \in C} x^T y.$$

This function is known as the support function of the set C. Notice that as a consequence of the definition,

$$x^T y - f^*(y) \leq 0, \quad \text{for any } x \in C,$$

so the set C lies on one side of this hyperplane. Further if the set is nice (closed) so that the sup is attained, we know that there is some $x_0 \in C$ such that, $x_0^T y - f^*(y) = 0$, i.e. the hyperplane $x^T y - f^*(y)$ is a supporting hyperplane of the convex set.

12.1.2 Fenchel's Inequality

The most basic property of the conjugate function that is immediate from its definition is that,

$$f^*(y) \geq x^T y - f(x).$$

This is known as Fenchel's inequality or Young's inequality or the Fenchel-Young inequality.

An immediate consequence is that for any function f, $f^{**} \leq f$. To see this (this is also a HW question), we notice that for any y by Fenchel's inequality,

$$f(x) \geq x^T y - f^*(y),$$

so that,

$$f(x) \geq \sup_y [x^T y - f^*(y)] = f^{**}(x).$$

12.1.3 Conjugate of Fenchel Conjugate

We've already seen via Fenchel's inequality that $f^{**} \leq f$. It is the case, that if the function f is closed (i.e. its epigraph is a closed set) and convex, then this holds with equality, i.e. $f^{**} = f$.

The Epigraph as an Intersection of Halfspaces: It's worth recalling/noting the following fact: any closed convex set is equal to the intersection of all the halfspaces which contain it. (This is very simple to prove as a consequence of the separating hyperplane theorem.)

This suggests that we could describe a closed convex function (a convex function whose epigraph is closed), by instead describing its epigraph. This would be a dual description of the convex function. Indeed, this dual description leads naturally to the Fenchel conjugate.

Concretely, suppose that $f^{**} = f$, then notice that,

$$\begin{aligned}
\text{epi}(f) &= \{(x, w) : f(x) \leq w\} \\
&= \{(x, w) : \sup_y [x^T y - f^*(y)] \leq w\} \\
&= \{(x, w) : x^T y - f^*(y) \leq w, y \in \mathbb{R}^d\} \\
&= \cap_{y \in \mathbb{R}^d} \{(x, w) : x^T y - w \leq f^*(y)\}.
\end{aligned}$$

So we have described the epigraph as a intersection of many halfspaces, and these halfspaces have intercepts which are defined in terms of the Fenchel conjugate.

When $f^{*}* \neq f$, then you can go through a similar chain of reasoning to show that the epigraph of f is contained in the intersection of a collection of halfspaces defined by f^* (but the containment is strict).

12.1.4 Inverse of Gradient

One of the other nice properties of the Fenchel conjugate is that it gives us an expression for the inverse of the gradient mapping. Recall, that we used this inverse in our description of the mirror descent algorithm (but it should not have been very obvious why the inverse exists, or when it does).

Here is the basic result: suppose that f is closed and convex. Then ∂f and ∂f^* are inverses

in the following sense:

$$y \in \partial f(x) \iff x \in \partial f^*(y).$$

When the function and its conjugate are in fact differentiable, this yields the fact that $\nabla f^*(\nabla f(x)) = x$.

We won't prove this rigorously (it's a part of your HW), but lets try to get some sense for why this is true. The first thing we'll need to know is the subgradient of a pointwise maximum (everything we're saying will generalize to a supremum), i.e. suppose:

$$f(x) = \max\{f_i(x) : i = 1, \ldots, m\},$$

then

$$\partial f(x) = \text{conv}\left[\bigcup_{i \in \mathcal{I}(x)} \partial f_i(x)\right],$$

where $\mathcal{I}(x) = \{i : f_i(x) = f(x)\}$ (i.e. the functions which attain the maximum). You can use this above fact on your HW (without proof).

Now, for the non-rigorous intuition (you'll need to be more careful using subgradient conditions in the HW). Suppose that all our functions are differentiable and we write

$$f^*(y) = \sup_x [x^T y - f(x)],$$

then if some x^* maximizes the RHS (non-rigorous), then by gradient optimality we must have that $y = \nabla f(x^*)$. On the other hand, we can write $f^*(y) = (x^*)^T y - f(x^*)$ so that $\nabla f^*(y) = x^*$. Putting these together gives us the claim that, $\nabla f(\nabla f^*(y)) = y$. Now, you can use the fact that $f^{**} = f$ to obtain the claim that, $\nabla f^*(\nabla f(x)) = x$.

12.1.5 Duality Between Smoothness and Strong Convexity

Another connection that you will explore a bit more in your HW is the duality between smoothness and strong convexity. Here is the result: if f is α-strongly convex, then f^* is $1/\alpha$-smooth. On the other hand, if f is convex, β-smooth, then f^* is $1/\beta$-strongly convex.

You can see this easily for the convex quadratics we discussed earlier, but it holds in general.

Remark: This connection is pretty interesting. We haven't yet discussed this yet, but if we could optimize a function f by optimizing its conjugate f^* instead, then we could transform problems with smoothness into ones with strong-convexity and vice versa. One reason why this is interesting is that accelerated methods achieve a faster rate for smooth problems – but acceleration is not possible for strongly convex problems which are not smooth. This suggests that at least if the conjugate is easy to compute (not always the case), then we might hope for an accelerated method for strongly convex problems which are not smooth by working in the dual.

12.2 Fenchel Duality

Part of the motivation for Fenchel duality is that our scheme for deriving duals (via Lagrange duality) seems somehow intricately tied to having a constrained optimization problem, and manipulating the constraints in some way.

However, for a problem like the LASSO (and many other problems) there aren't any constraints (but the objective has some structure), and it would be nice if we could rewrite these types of problems in order to gain some insights (or to derive new algorithms). Indeed, we have noticed before that every constrained program could be written as an unconstrained one (by using an indicator function for feasibility), and so it seems natural that these problems should have interesting duals.

12.2.1 The Simplest Case

Suppose I have an unconstrained problem,

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

You could imagine g is an indicator of a constraint set, or a regularizer. It's not clear how to use Lagrange duality here. However, notice that we could re-write this as the following constrained problem:

$$\begin{aligned} &\min_{x,z} \ f(x) + g(z) \\ &\text{subject to } x = z. \end{aligned}$$

Writing out the Lagrangian:

$$L(x, z, u) = f(x) + g(z) + u^T(x - z),$$

so the dual problem is:

$$\sup_u \inf_x [f(x) + u^T x] + \inf_z g(z) - u^T z = \sup_u -f^*(-u) - g^*(u).$$

This problem is sometimes called the Fenchel dual of the original problem. Fenchel's duality theorem gives conditions on f and g such that strong duality holds. Particularly, it says that if there is some $x_0 \in \text{dom}(f) \cap \text{cont}(g)$ then strong duality holds. $\text{cont}(g)$ is the set of points where the function g is continuous. In the case, when g is the indicator of a convex set, this condition will precisely recover Slater's conditions (i.e. we need a strictly feasible point).

12.2.2 A Slightly More Interesting Case

A slightly more interesting example, one that will include the LASSO as a special case, is problems of the form:

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

We could introduce the constraint $Ax = z$, to write,

$$\begin{aligned} &\min_{x,z} f(z) + g(x) \\ &\text{subject to } Ax = z, \end{aligned}$$

and then derive the Lagrange dual as above to see that the dual is:

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

As an application, suppose we took the LASSO program:

$$\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1 := \min_x f(Ax) + g(x),$$

then we would have the dual problem (needs a few steps of algebra to derive the conjugates):

$$\max_u \left[-y^T u - \frac{1}{2}\|u\|_2^2 - \mathbb{I}(\|A^T u\|_\infty \leq \lambda)\right] := \max_{\widetilde{u}} \left[y^T \widetilde{u} - \frac{1}{2}\|\widetilde{u}\|_2^2 - \mathbb{I}(\|A^T \widetilde{u}\|_\infty \leq \lambda)\right],$$

by replacing u by $-\widetilde{u}$ since it is unconstrained. This dual problem has the same optimizer as the following slightly more interpretable problem:

$$\begin{aligned} &\min_{\widetilde{u}} \ \frac{1}{2}\|y - \widetilde{u}\|_2^2, \\ &\text{subject to } \|A^T \widetilde{u}\|_\infty \leq \lambda. \end{aligned}$$

This dual is quite nice – it says that the dual to the LASSO is simply a projection problem, and it's a projection onto a polytope. In deriving this dual (in one of the steps of algebra to derive the conjugate) you would notice that $\widetilde{u} = y - Ax$, so given the dual optimal solution it's easy to derive the primal optimal solution (and strong duality holds). This dual can be used to show various properties of the LASSO fitted values (the fact that has some projection interpretation means that the fitted value inherit some type of Lipschitzness/non-expansiveness that we associate with projections).

More broadly, this idea of introducing constraints to derive a dual optimization problem, is at the heart of many dual optimization schemes – for instance, algorithms like dual ascent and ADMM (Alternating Direction Method of Multipliers) all will involve steps like this.