Convexity bootcamp part 2: duality

Administrative:

- New room? Siebel 1214?
- Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Strict convexity means $\nabla^2 f \succ 0$ everywhere, which allows the possibility that $\inf_x(\nabla^2 f)(x) = 0$, whereas strong convexity means $\nabla^2 f \succeq \lambda I$ everywhere for some $\lambda > 0$. For example, $g(x) = e^x$ has $g'' > 0$ thus g is strictly convex, whereas $\inf_x g''(x) = 0$.
- Another way to view local-to-global: we'll show (time permitting) that gradient descent on a strongly convex and smooth function takes $\mathcal{O}(\ln(1/\epsilon))$ iterations for accuracy ϵ, which roughly means we get one bit of the optimum per iteration (i.e., one bit of the global optimum using nothing but local information).
- Someone put a cat sticker on my mailbox ?!

Constrained optimization without duality

Earlier we discussed first order conditions for $\inf_x f(x)$ with $f : \mathbb{R}^d \to \mathbb{R}$, meaning there were no constraints. The task now is to add constraints, or more generally consider minimization of the form

$$\inf_x f(x) + g(Ax),$$

where f and g are both convex and A is a matrix; for instance, regular constrained optimization of f arises by setting $A = I$ and $g = \iota_S$ for some convex set S.

Duality will be our deepest look into the structure of these optimization problems, but sometimes we can get somewhere just by hitting the above optimization problem with first order conditions.

Theorem (Pshenichnyi-Rockafellar)**.** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and S is convex with $\text{int}(S) \neq \emptyset$ (note: this last bit is another "constraint qualification", to be explained soon). Then $\inf_{x \in S} f(x)$ has minimum point $\bar{x} \in S$ iff $\partial f(\bar{x}) \cap -N_S(\bar{x}) \neq \emptyset$, where $N_S(x)$ is the **normal cone** of S at x, defined as

$$N_S(x) := \left\{ s \in \mathbb{R}^d : \forall x' \in S \centerdot \left\langle s, x - x' \right\rangle \geq 0 \right\}.$$

(In class, a picture was drawn: normal cone are those directions which move "at least orthogonally" from the set. The idea is that if your (negative) gradient has any positive projection onto your set, then you should be able to move slightly in that direction and decrease your objective value.)

To prove this, we'll use a quick lemma.

Lemma. $\partial \iota_S(x) = N_S(x)$.

Proof. For any $x \in \text{dom}(\iota_S) = S$,

$$\begin{aligned} \partial \iota_S(x) &= \left\{ s \in \mathbb{R}^d : \forall x' \centerdot \iota_S(x') \geq \iota_S(x) + \left\langle s, x' - x \right\rangle \right\} \\ &= \left\{ s \in \mathbb{R}^d : \forall x' \in S \centerdot 0 \geq \left\langle s, x' - x \right\rangle \right\}, \end{aligned}$$

the second equality since the inequality is always true for $x' \notin S$, so we can restrict the quantification over S.

Proof (of theorem)**.** Applying the first order conditions to $f + \iota_S$,

$$f(\bar{x}) = \inf_{x \in S} f(x) \iff 0 \in \partial(f + \iota_S)(\bar{x})$$
$$\overset{(\star)}{\iff} 0 \in \partial f(\bar{x}) + \partial \iota_S(\bar{x}) \iff 0 \in \partial f(\bar{x}) + N_S(\bar{x}) \iff \partial f(\bar{x}) \cap -N_S(\bar{x}) \neq \emptyset,$$

where the step $(\star)$ used the constraint qualification (see the subgradient rule from the last lecture).

Example. Consider (the constrained form of) the *LASSO*

$$\inf \left\{ \frac{1}{2}\|Xw - y\|_2^2 : w \in \mathbb{R}^d, \|w\|_1 \leq 1/\lambda \right\}.$$

(Picture drawn in class with level sets around the unconstrained optimum, and then normal cones to the l_1 ball. Different projection properties of the l_2 ball drawn and discussed.)

The Fenchel Conjugate

Suppose, for sake of easier discussion, that we're given a differentiable convex function. In this case, first-order optimality conditions tell us that $\bar{x}$ minimizes f iff $\nabla f(\bar{x}) = 0$. With this in mind, wouldn't it be great if we could simply *invert* ∇f, and get a handle on an optimal set just by writing $(\nabla f)^{-1}(0)$?

This is exactly one of the things that the Fenchel conjugate will give us. Later we will see that the dual space is a sort of gradient space, and gradients and the Fenchel conjugate are our best way of moving between the dual and primal.

Speaking more concretely, we can heuristically derive the Fenchel conjugate as follows, playing fast and loose and ignoring all questions of invertibility, uniqueness, etc.:

$$x = (\nabla f)^{-1}(s) \quad \text{“} \iff \text{”} \quad s = \nabla f(x) \quad \text{“} \iff \text{”} \quad 0 = s - \nabla f(x) \quad \text{“} \iff \text{”} \quad x = \operatorname{argmax}_{x'} \langle s, x' \rangle - f(x').$$

Despite all the assumptions and heuristics, this derivation lands at a well-defined object: define the **Fenchel conjugate** f^* of $f : \mathbb{R}^d \to \mathbb{R} \cup \infty$ as

$$f^*(s) := \sup\{\langle s, x \rangle - f(x) : x \in \mathbb{R}^d\}.$$

This function has many interesting properties, but first let's see an example.

Proposition. The *unique* $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $f = f^*$ is $f(x) = \|x\|_2^2/2$.

Remark. Arguably this is the best reason for including the "1/2" on quadratics.

Proof. Let's first check that $g(x) = \|x\|_2^2/2$ satisfies $g = g^*$. In order to highlight the new concepts, the proof will step through things a little more slowly than usual. Taking the gradient of the expression within the supremum inside $g^*(s)$ and setting to zero gives $s - x = 0$; by first-order conditions, this means the sup is attained at $x = s$ for all s, and therefore (by plugging back in)

$$g^*(s) := \langle s, s \rangle - \|s\|_2^2/2 = g(s).$$

(Note that we both spelled out a derivation *and* a proof. To *derive* the Fenchel conjugate, we check some derivatives as the easiest way to compute the sup. But then we were left with an expression which we *prove* is the optimum of the sup due to first-order conditions.)

To prove the uniqueness needs some tricks. Let $f : \mathbb{R}^d \to \mathbb{R}$ be given with $f = f^*$; it's our goal to show $f = g$. Let's try to exhibit the expression $\|s\|_2^2$ somehow from the definition of f^*. We can do this by lower bounding the sup by its value at s:

$$f(s) = f^*(s) = \sup_x \langle s, x \rangle - f(x) \geq \langle s, s \rangle - f(s) = \|s\|_2^2 - f(s).$$

Rearranging this gives the inequality

$$f^*(s) = f(s) \geq \|s\|_2^2/2.$$

The proof is complete by invoking the following fact (which is proved in a moment): if $f \geq h$, then $f^* \leq h^*$. This completes the proof since it gives the chain of inequalities

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

starts and ends the same, and is thus a chain of equalities.

In turn, the fact is a consequence of the definition of conjugate: $f \geq h$ implies, for any $s \in \mathbb{R}^d$, that

$$f^*(s) = \sup_x \langle s, x\rangle - f(x) \leq \sup_x \langle s, x\rangle - h(x) = h^*(s).$$

□

The conjugate satisfies many convenient rules.

- f^* is convex, even if f is not (this follows from the supremum rule for convex functions: each function $s \mapsto \langle s, x\rangle - f(x)$ is an affine and thus convex function in s (it doesn't matter that they are not necessarily convex in x)).
- If f is convex, then f^* is closed convex.
- If $f \geq h$, then $f^* \leq h^*$ (as above).
- If $f(x) = g(cx)$, then $f^*(s) = cg^*(x/c)$.
- $f = f^{**}$ iff f is closed convex.
- **(Fenchel-Young inequality.)** $f(x) + f^*(s) \geq \langle x, s\rangle$. **Note:** this inequality, together with it's equality case in the following bullet point, is perhaps the single most important low-level property given today.
- If f is closed convex,

$$f(x) + f^*(s) = \langle s, x\rangle \iff s \in \partial f(x) \iff x \in \partial f^*(s).$$

  This last bit captures the "inverse subgradient" property we wanted.

Example. Conjugates show up in many places. Consider the standard "Chernoff method" for proving concentration inequalities: by Markov's inequality,

$$\Pr[X \geq a] \leq \inf_{t\geq 0} \Pr[\exp(tX) \geq \exp(ta)] \leq \inf_{t\geq 0} \frac{\mathbb{E}(\exp(tX))}{\exp(ta)}.$$

Since $\ln(\cdot)$ is monotone increasing along $\mathbb{R}_{++}$,

$$\inf_{t\geq 0} \frac{\mathbb{E}(\exp(tX))}{\exp(ta)} = \exp\left(\ln\left(\inf_{t\geq 0} \frac{\mathbb{E}(\exp(tX))}{\exp(ta)}\right)\right) = \exp\left(\inf_{t\geq 0} \ln \mathbb{E}(\exp(tX)) - ta\right) = \exp\left(-\Psi^*(a)\right),$$

defining $\Psi(t) := \ln \mathbb{E} \exp(tX)$ (the log of the moment generating function) when $t \geq 0$ and $\Psi(t) = \infty$ otherwise. Thus convexity arises in concentration.

Fenchel Duality

As stated above, sometimes we need a more serious look into the behavior of an optimization problem. The **dual problem** is another convex problem which lower bounds the **primal problem** $\inf_x f(x) + g(Ax)$, an object we saw above. As discussed, the dual space is a sort of gradient space.

Theorem (Fenchel Duality)**.** Let convex $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, and matrix $A : \mathbb{R}^d \to \mathbb{R}^n$ be given. Assume $\inf_x f(x) + g(Ax) > -\infty$, and the **constraint qualification**

$$0 \in \operatorname{int}(\operatorname{dom}(g) - A\operatorname{dom}(f)).$$

Then

$$\inf \left\{ f(x) + g(Ax) : x \in \mathbb{R}^d \right\} = \max \left\{ -f^*(A^\top s) - g^*(-s) : s \in \mathbb{R}^n \right\}.$$

A pair $(\bar{x}, \bar{s})$ are optimal iff $A^\top \bar{s} \in \partial f(\bar{x})$ and $-\bar{s} \in \partial g(A\bar{x})$.

Remarks. ("There's a lot to say here".)

- **(Essential proof technique; constraint qualifications.)** The key way I know to prove such theorems is the *perturbation technique*, due to Rockafellar. The idea is to look at the function

  $$h(u) := \inf \left\{ f(x) + g(Ax + u) : x \in \mathbb{R}^d \right\}.$$

  $h(0)$ is simply the primal optimal value, but by studying h in the vicinity of 0, we can reason that the optimization problem is well-behaved around its optimum, which lets us prove equality to the dual and the optimality conditions.

  For the purpose of the proof, it will suffice for us to exhibit a subgradient of h at 0. Note that $\operatorname{dom}(h) = \operatorname{dom}(g) - A\operatorname{dom}(f)$, thus the constraint qualification is nothing more than $0 \in \operatorname{int}(\operatorname{dom}(h))$, meaning a *sufficient* condition for the $\partial h(0)$ to be nonempty. (Superficially, the constraint qualification is as above, a statement that the optimization problem is well-behaved at the optimum.)

  Constraint qualifications for certain problems can be weaker, the weakest being $0 \in \operatorname{dom}(g) - A\operatorname{dom}(f)$; this is the case for *polyhedral programs* (all epigraphs are polyhedra), see for instance (Rockafellar 1970, Theorem 31.1).

  Constraint qualifications become a real mess in infinite dimensions; see for instance (Zălinescu 2002, Theorem 2.8.4).

- **(Duality gap.)** Let $x \in \mathbb{R}^d$ be given, and set $s := \partial g(Ax)$. By the theorem, the quantity

  $$f(x) + g(Ax) + f^*(A^\top s) + g^*(-s) \geq 0$$

  (where the inequality follows from the Fenchel-Young inequality), the **duality gap**, is the gold standard for determining convergence of numerical algorithms (or variants which exhibit a better dual point than s). By Fenchel-Young, if x is optimal, then the duality gap is zero.

- **The dual optimum is attained.** This is an interesting piece of asymmetry.

We won't prove the theorem fully in these notes; a full proof using the perturbation technique can be found in (Borwein and Lewis 2000, Theorem 3.3.5). *(note to future matus: consider adapting that one geometric duality proof in Rockafellar (1970) to the general setting.)*

Example. Let's close with an example on empirical risk minimization.

Take $\ell : \mathbb{R} \to \mathbb{R}_+$ to be a univariate, nonnegative loss function. Let data $((x_i, y_i))_{i=1}^n$ be given, and collect it into a matrix $A \in \mathbb{R}^{n \times d}$ as $A_{ij} := -(x_i)_j y_i$. Define

$$\begin{aligned} p(w) \quad & \text{penalty function / regularizer} \\ \ell_i(z) := \ell(z_i) \quad & \text{coordinate-wise loss.} \end{aligned}$$

With this notation, the **regularized empirical risk minimization** problem can be written

$$\inf_{w \in \mathbb{R}^d} \sum_{i=1}^n \ell(-y_i x_i^\top w) + p(w) = \inf_{w \in \mathbb{R}^d} \sum_{i=1}^n \ell_i(Aw) + p(w).$$

In order to form the dual, we need another convenient property of Fenchel conjugates.

Lemma. Let convex $f : \mathbb{R} \to \mathbb{R}$ be given, and define $g : \mathbb{R}^d \to \mathbb{R}$ as $g(x) := \sum_i f(x_i)$. Then $g^*(s) = \sum_i f^*(s_i)$.

This leads to the following primal-dual equation.

$$\inf \left\{ \sum_i \ell_i(Aw) + p(w) : w \in \mathbb{R}^d \right\} = \max \left\{ -\sum_i \ell_i^*(-s) - p^*(A^\top s) : s \in \mathbb{R}^n \right\}.$$

The dual vector space has the same dimension as the number of examples, and can be viewed as a search over weightings on examples, which is very suggestive in the case of SVMs and boosting (as will be discussed when they are reached).

References

Borwein, Jonathan, and Adrian Lewis. 2000. *Convex Analysis and Nonlinear Optimization.* Springer Publishing Company, Incorporated.

Rockafellar, R. Tyrrell. 1970. *Convex Analysis.* Princeton University Press.

Zălinescu, Constantin. 2002. *Convex Analysis in General Vector Spaces.* World scientific.