10725/36725 Optimization
Recitation , Oct 16

Adona Iosif
aiosif@cs.cmu.edu

I spent this recitation reviewing duality theory, and trying to provide a coherent view of what we did and why we did it.

1 The primal and dual problems

The core idea of duality is that for any general optimization problem, called a *primal problem*, we can construct a second optimization problem, called the *dual problem*, which is in some sense equivalent to the original problem.

Primal problem:

$$\begin{aligned} f^* = \min_{x \in R^n} & f(x) \\ \text{subject to } & h_i(x) \leq 0, i \in 1, ..., m \\ & l_j(x) \leq 0, j \in 1, ..., p \end{aligned}$$

Dual problem:

$$\begin{aligned} g^* = \max_{u \in R^m, v \in R^p} & g(u, v) \\ \text{subject to } & u \geq 0 \end{aligned}$$

You can think about duality as trying a 'different perspective' on a question, and it has exactly the advantages you expect a different perspective to bring:

- The dual problem might be *simpler* to solve than the primal problem. e.g. the dual often has simpler constraints than the primal
- The dual problem might be *faster* to solve than the primal problem. e.g. the dual might have fewer variables than the primal, or be better conditioned, leading to optimization algorithms converging faster.

- The dual formulation might *bring new insights* into the primal solution (might help 'characterize' the primal solution). etc.

Weak and Strong Duality

But what do we mean by 'equivalent problems in some sense'? Well, under the most general assumptions (=we don't know anything about f, f could even be non-convex), we are guaranteed by construction of the dual (we'll see below) that $g^* \leq f^*$, in other words that solving the dual gives us a lower bound on the primal. This is not perfect, but can be useful in certain applications.

In many cases though, we can say something even stronger. For example, if f, h_i are convex and l_j are affine ('Slater's conditions'), we can say under very weak conditions that *strong duality* holds: $f^* = g^*$. **This is incredibly powerful**, and it's what makes us be able to solve the dual to get the primal solution. Let's now get into the details.

2 Deriving the dual from the primal

You can think of the process of deriving the dual as trying to construct simple lower bounds on the primal objective, and then choosing as tight a bound as we can in order to get a good approximation.

Practically, to derive the dual we simply follow the following steps:

1. **Write the Lagrangian** The Lagrangian is defined as a linear combination of the original objective f and the constraints h_i, l_j:

$$L(x, u, v) = f(x) + \sum_i u_i h_i(x) + \sum_j v_j l_j(x) \tag{1}$$

   where $u_i \geq 0$. By construction, if C is the constraint set of the primal problem $C = \{x : h_i(x) \leq 0, l_j(x) = 0\}$, we have:

$$f(x) \geq L(x, u, v), \forall x \in C$$

   We can see this easily: on C, hs are negative (so $u \cdot h$s are negative as well - that's why we required $u \geq 0$), ls are 0, and thus we're adding two non-positive quantities to f, so we must get an underestimate.

2. **Minimize the Lagrangian to get the dual function** From above, we have that:

$$f^* = \min_{x \in C} f(x) \geq \min_{x \in C} L(x, u, v), \forall u \geq 0, v$$

   But if minimizing L over C provides a lower bound, minimizing L over all of R^n, an even greater space, will also provide a (less tight) lower bound. This is a place where

we can potentially lose a lot of tightness in our bound, but it's worth it, because it makes our optimization problem unconstrained, and thus much simpler. In conclusion, in this step we compute what's called the *dual function*:

$$g(u, v) = \min_{x \in R^n} L(x, u, v) \tag{2}$$

To recap, we know that: $f^* = \min_{x \in C} f(x) \geq g(u, v), \forall u \geq 0, v$.

3. **Write out the new optimization problem** Finally, as we said: if all $g(u, v)$ provide lower bounds, might as well take the best lower bound. We thus write out the dual problem:

$$\begin{aligned} g^* = \max_{u \in R^m, v \in R^p} & \; g(u, v) \\ \text{subject to } & \; u \geq 0 \end{aligned} \tag{3}$$

Equations (1), (2) and (3) are all that's necessary to derive the dual of a problem. The tricky part is usually in equation (2): it might be difficult to do the minimization in (2) in closed form. However, we will see tools which help us do that minimization later (hint: conjugate functions). For now, let's see an example.

Example: Entropy Maximization This is an example taken from Boyd [TODO: page]. We wish to minimize the following function (negative entropy):

$$\begin{aligned} \min_{x \in R^n} & \sum_{i=1}^{n} x_i \log x_i \\ \text{subject to } & Ax \leq b \\ & 1^T x = 1 \end{aligned}$$

Before we even think about duality, take a moment to think about algorithms we've learned that could tackle this problem. Really, the only way we know how to approach constrained problems is using some sort of projected e.g. gradient descent. The issue though is that the constraints are very hard to project onto: we don't know how to project either on a general polyhedra ($Ax \leq b$) or a probability simplex ($1^T x = 1$). So in the primal we're stuck.

How about the dual? Well, first, notice that since entropy is concave (physics, anyone?), negative entropy will be convex. Further, all the constraints are affine, so we know by Slater's conditions that the primal and dual optima will be equal: $f^* = g^*$.

Now let's compute the dual:

1. **Write the Lagrangian** $L(x, u, v) = \sum_{i=1}^{n} x_i \log x_i + u^T(Ax - b) + v(1^T x - 1), u \geq 0$
2. **Minimize the Lagrangian to get the dual function** We minimize the Lagrangian,

as always, by taking the derivative and setting it to zero:

$$0 = \frac{d}{dx_i} L(x, u, v) = \log x_i + x_i \frac{1}{x_i} + u^T A_i + v_i = \log x_i + u^T A_i + v_i + 1$$
$$x_i = e^{-u^T A_i - v_i - 1}$$

Once we got the optimal x, we plug it back into $L(x, u, v)$ to get $g(u, v)$:

$$\begin{aligned} g(u, v) &= \sum_{i=1}^{n} [x_i(-u^T A_i - v_i - 1) + u^T(A_i x_i - b_i) + v x_i] - v \\ &= -u^T b - v - \sum_{i=1}^{n} x_i \\ &= -u^T b - v - e^{-v_i - 1} \sum_{i=1}^{n} e^{-u^T A_i} \end{aligned}$$

3. **Write out the new optimization problem** We can now write the dual problem:

$$\begin{aligned} g^* = &\max_{u \in R^m, v \in R^p} -u^T b - v - e^{-v_i - 1} \sum_{i=1}^{n} e^{-u^T A_i} \\ &\text{subject to } u \geq 0 \end{aligned}$$

Ok, so the primal was hard, how about the dual? Well, the dual is trivial! It's a convex optimization problem with smooth objective and a trivial constraint ($u \geq 0$). So we can easily apply projected gradient descent, solve the dual, and recover the primal objective value!

One little observation though: we now know how to compute f^*, but we still don't know how to compute x^*. We'll see that coming up next though, through the KKT conditions!

3 Tools to make deriving the dual easier: conjugate functions

The conjugate function definition might seem slightly arbitrary in the beginning, but it arrises very naturally from duality. To see that, take a very general optimization problem:

$$\begin{aligned} f^* = &\min_{x \in R^n} f(x) \\ &\text{subject to } Ax = b \end{aligned}$$

and try to compute its dual. As usual, we first compute the Lagrangian:

$$L(x, u) = f(x) + u^T(b - Ax)$$

and then minimize it to get the dual:

$$\begin{aligned} g(u) = \min_x L(x,u) &= u^T b + \min_x (f(x) - (A^T u)^T x) \\ &= u^T b - \max_x ((A^T u)^T x - f(x)) \end{aligned}$$

The structure $\max_y (y^T x - f(x))$ is what we call the *conjugate function* $f^*(y)$. As we can see, it appears regularly in duality (in this case y = $A^T u$), so it's clearly worth understanding better, and deriving rules about how it behaves. What people do is that they:

- Derive the conjugates of important functions (e.g. the conjugate of a norm $||x||$ is always $I_{\{z:||z||_* \leq 1\}}(y)$)
- Derive rules for how conjugation works with basic operations (e.g. $[f(ax)]^* = f^*(\frac{x}{a})$, +wikipedia has a whole table of them)

Once they have those two building blocks, they can easily take more complex looking problems ($[||ax||_1 + ||bx||_\infty]^*$?), identify the components (some norms, and some basic operations), and modularly solve the problem. Let's see how this works in a few examples!

Example: L-1 norm We want to derive the dual problem of the following primal:

$$\begin{aligned} &\min_x ||x||_1 \\ &\text{subject to } Ax = b \end{aligned}$$

This might have looked scary once, but see how easy it is now! From above:

$$\begin{aligned} g(u) &= u^T b - (||\cdot||_1)^*(A^T u) \\ &= u^T b - I_{\{z:||z||_\infty \leq 1\}}(A^T u) \end{aligned}$$

So the dual is:

$$\begin{aligned} &\max_u u^T b \\ &\text{subject to } ||A^T u||_\infty \leq 1 \end{aligned}$$

Example: Trace norm Again, we derive the dual problem of the following primal:

$$\begin{aligned} &\min_X ||X||_* \\ &\text{subject to } Tr(A^T X) = b \end{aligned}$$

Notice here that we used $Tr(A^T X) =< A, X >$, the inner product between two matrices, to express a linear equation in the matrix variable X. Again, if we just look at this problem, it seems very hard to solve with no extra tools. Let's take the dual though! As before:

$$\begin{aligned} g(u) &= ub - (||\cdot||_*)^*(uA) \\ &= ub - I_{\{z:||z||_{op} \leq 1\}}(uA) \end{aligned}$$

So the dual problem is:

$$\max_u ub$$
$$\text{subject to } ||uA||_{op} \leq 1$$

Well, this is trivial to solve:

$$|u| \leq \frac{1}{||A||_{op}} = \frac{1}{\sigma_1(A)}$$

so $u^* = \frac{sign(b)}{\sigma_1(A)}$, $g^* = \frac{|b|}{\sigma_1(A)}$.

To recap, we took a primal problem we didn't know how to solve, we did one line of calculations based on knowing properties of the conjugate functions, and we got a dual problem which **we knew how to solve in closed form**! Neat, aye? :)

There is only one snag. Yes, we know $f^* = g^*$, but we still don't know X^*. How do we get X^*? Well, here is where the KKT conditions help!

4 KKT Conditions

You can get an intuitive grasp of where the KKT conditions come from by looking at how we derived the dual problem. As we saw in Section 2, we derived the dual problem by constructing a string of lower bounds on f^*:

$$f^* = \min_{x \in C} f(x) \geq \min_{x \in C} L(x, u, v) \geq \min_{x \in R^n} L(x, u, v) = g(u, v), \forall u \geq 0, v$$

In particular:

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

We can trivially see then that we can have strong duality $f^* = g^*$ if and only if all those inequalities are actually equalities. In other words, iff the following hold:

- [**stationarity**] x^* minimizes the Lagrangian at u^*, v^* (the 2nd inequality):

$$0 \in \partial f(x^*) + \sum_i u_i^* h_i(x^*) + \sum_j v_j^* l_j(x^*)$$

- [**complementary slackness**] Lagrangian is equal to f on C (the 1st inequality), so:

$$u_i^* h_i(x^*) = 0, \forall i$$

- [**primal and dual feasibility**] $h_i(x^*) \leq 0$, $l_j(x^*) = 0$ for all i, j, and $u \geq 0$

As with duality, there are many ways the KKT conditions come in useful. We can:

- *Directly solve the KKT conditions to get the primal and dual solutions* (x^*, u^*, v^*). We saw this in class when we wrote the KKT conditions for: $\min_x \frac{1}{2}x^T Qx$, subject to $Ax = 0$. (Lecture 13, Slide 13, yayks!)
- Solve the dual, get u^*, v^*, plug them back in the Lagrangian, and now *minimize the Lagrangian to get* x^*. This addresses exactly our problem above: until now, we only knew how to use the dual problem to get f^*, not x^*. We now have a recipe for how to recover x^* from the solutions of the dual problem. We'll see an example below!
- Use the KKT conditions to characterize x^*.

Example: Trace norm Recall the following trace norm problem:

$$\min_X ||X||_*$$
$$\text{subject to } Tr(A^T X) = b$$

We derived the dual above as:

$$\max_u u^T b$$
$$\text{subject to } ||uA||_{op} \leq 1$$

And determined that the optimal solution is: $u^* = \frac{sign(b)}{\sigma_1(A)}$, $g^* = \frac{|b|}{\sigma_1(A)}$. We can now plug u^* in the stationarity condition to say that X^* minimizes:

$$L(X, u^*) = ||X||_* + u^*(b - Tr(A^T X)) = ||X||_* + \frac{sign(b)}{\sigma_1(A)}(b - Tr(A^T X))$$

This is now an unconstrained optimization problem so we can apply any algorithm (e.g. sub-gradient descent) to solve it, and recover X^*!

To recap all that was said:

- Duality provides a new perspective on an optimization problem.
- We can use the dual problem to recover f^*, or to get extra insights.
- We can use the KKT conditions to recover x^*, or to get extra insights.
- You should feel at home with: Lagrangians, dual function, dual problem, conjugate functions, and the KKT conditions. They come up everywhere!

The recitation also covered a geometric explanation of duality, but it's easier to follow with a whiteboard than written, so please check the video of the recitation and/or Boyd, pages 232-233. Enjoy!