Statistical Machine Learning Notes 7 "bonus"

Lagrange Duality

Instructor: Justin Domke

1 Theory

Lagrange duality is a fundamental tool in machine learning (among many, many other areas). Though a full understanding of Lagrange duality is beyond the scope of this course, the basic intuition is not hard to explain. Let's begin by considering a general optimization problem

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0 \\ & h_i(\mathbf{x}) = 0. \end{aligned}$$

Here, we assume that the problem is convex, meaning that f_0 and f_i are convex functions, and that h_i are affine. Let the optimum of this problem be p^*.

The **Lagrangian** is defined to be

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x}) + \sum_j \nu_j h_j(\mathbf{x}).$$

The **Lagrange dual function** is defined to be

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

The key idea of Lagrange duality is the following observation:

Claim #1:
For any $\boldsymbol{\nu}$ and for any $\boldsymbol{\lambda} \geq \mathbf{0}$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$.

This is to say, for any $\boldsymbol{\nu}$, and any nonnegative $\boldsymbol{\lambda}$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ lower-bounds the true optimum. It is not hard to see why this is true. First off, for any $\mathbf{x}$ that obeys the constraints,

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}).$$

That is to say, the Lagrangian can only "help" legal solutions. Thus, in particular, if $\mathbf{x}^*$ is the optimum of the original problem

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\nu}) = p^*$$

For illegal solutions, however, this is not true. If $\mathbf{x}'$ violates some constraints, then for certain settings of the Lagrange multipliers, $\mathcal{L}(\mathbf{x}', \boldsymbol{\lambda}, \boldsymbol{\nu}) > f_0(\mathbf{x}')$. This is in fact the key to Lagrange duality's magic.

Given that the Lagrange dual function gives valid lower bounds for any $\forall \boldsymbol{\nu}, \forall \boldsymbol{\lambda} \geq \mathbf{0}$, it is natural to try to get the best/tightest/highest lower bound. This is called the **Lagrange dual problem**.

$$\begin{array}{rl} \max & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{s.t.} & \lambda \geq \mathbf{0}. \end{array}$$

Call the optimal value of the Lagrange dual problem d^*. It is obvious that $d^* \leq p^*$, always. But wouldn't it be nice if $d^* = p^*$? This is called **strong duality**. When does this hold? The following theorem gives some sufficent conditions. (Note this is on top of our assumption of convexity.)

Claim #2 (Slater's Theorem):
Strong duality holds if there exists a strictly feasible point, i.e. some $\mathbf{x}$ such that the inequality constraints are strictly satisfied, with

$$\begin{array}{rcl} f_i(\mathbf{x}) & < & 0 \\ h_j(\mathbf{x}) & = & 0. \end{array}$$

This can be slightly strengthened. If some f_i are affine ($f_i(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} + b$), then they don't need to be strictly satisfied.

Claim #3 (Slater's Theorem++):
Strong duality holds if there exists a strictly feasible point, i.e. some $\mathbf{x}$ such that the inequality constraints are strictly satisfied, with

$$\begin{aligned} f_i(\mathbf{x}) &\leq 0, \ f_i \text{ affine} \\ f_i(\mathbf{x}) &< 0, \ f_i \text{ non-affine} \\ h_j(\mathbf{x}) &= 0. \end{aligned}$$

2 Discussion

Fundamentally, that is all you need to know about Lagrange duality for this course. However, some discussion is helpful to appreciate the significance of Lagrange duality, and when it is useful.

Another way of writing Lagrange duality, when strong duality holds, is:[1]

$$\min_{\substack{\mathbf{x}:f_i(\mathbf{x})\leq 0 \\ h_j(\mathbf{x})=0}} f_0(\mathbf{x}) = \max_{\boldsymbol{\nu},\boldsymbol{\lambda}\geq\mathbf{0}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}).$$

Thus, we take a constrained problem, and convert it into a "maximin" problem. There are two things that have happened

1. We have made the problem *easier* in the sense that the complex constraints f_i, h_j have been replaced with the trivial constraint $\boldsymbol{\lambda} \geq \mathbf{0}$. This is almost always good.
2. We have made the problem harder in the sense that we now have two "nested", and "competing" optimization problems. This is possibly bad.

Though the above theory is always *true*, the *usefulness* of Lagrange duality depends greatly upon how easy it is to solve $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu})$. There are a bunch of situations:

- **Closed form solution**. This is what we want, and most of what you will see in the literature. This means that $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu})$ can be solved "symbolically". Usually, this is done by calculating the gradient $d\mathcal{L}/d\mathbf{x}$, setting it to zero, and solving for $\mathbf{x}$ in terms of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. If we write our closed-form solution for $\mathbf{x}$ in terms of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ as $\mathbf{x}^*(\boldsymbol{\lambda},\boldsymbol{\nu})$, we have the optimization problem

$$\max_{\boldsymbol{\nu},\boldsymbol{\lambda}\geq\mathbf{0}} \mathcal{L}(\mathbf{x}^*(\boldsymbol{\lambda},\boldsymbol{\nu}),\boldsymbol{\lambda},\boldsymbol{\nu}).$$

[1] Here we sloppily replace inf with min. This is only correct, of course, if the infimum is bounded below. In practice it often happens that for some $\boldsymbol{\nu}$, $\boldsymbol{\lambda}$, $\inf_{\mathbf{x}} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) = -\infty$. This is usually dealt with by constraining $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ to be such that this doesn't occur.

When this is possible, Lagrange duality essentially turns a constrained problem into an almost unconstrained one ("Almost" meaning the simple constraint that $\boldsymbol{\lambda} \geq \mathbf{0}$).

- **Fast Algorithm**. A less ideal case where Lagrange duality can still be useful is when $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ doesn't have a simple symbolic solution, but does have a fast *algorithm*. Depending on technical details of this algorithm (beyond the scope of this course), it may still be possible to recover, for example $\frac{d}{d\boldsymbol{\nu}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, or $\frac{d}{d\boldsymbol{\lambda}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, meaning an algorithm like gradient ascent might be used to solve the Lagrange dual problem.
- **No easy solution**. If $f_0(\mathbf{x})$ is complicated enough, it is going to be very hard to solve the inner minimization, no matter the constraints. Lagrange duality isn't *usually* used in these, though there are no hard and fast rules. Often, complicated functions can be turned into simpler functions by creating new variables and new constraints. For example $\min\limits_{x} |x| = \min\limits_{x,z} z$, s.t. $z \geq x, z \geq -x$. In less trivial situations, tricks like this can make it possible to use Lagrange duality when originally the problem seems too hard.

In this course, we will only encounter the best case, when closed-form solutions exist. Finally, take an example of where we can find a closed-form solution. Take the problem

$$\begin{array}{ll} \min & \frac{1}{2}\mathbf{x} \cdot \mathbf{x} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b}. \end{array}$$

This has the Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\nu}) = \frac{1}{2}\mathbf{x} \cdot \mathbf{x} + \boldsymbol{\nu}^T (A\mathbf{x} - \mathbf{b}).$$

For fixed $\boldsymbol{\nu}$, it is easy to solve for the minimizing $\mathbf{x}$. At the maximum, the derivative with respect to $\mathbf{x}$ must be zero, and so

$$\frac{d\mathcal{L}}{d\mathbf{x}} = \mathbf{0} = \mathbf{x} + A^T \boldsymbol{\nu}$$

$$\mathbf{x}^* = -A^T \boldsymbol{\nu}$$

Substituting this, we have that

$$\begin{aligned} \max_{\boldsymbol{\nu}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\nu}) &= \max_{\boldsymbol{\nu}} \frac{1}{2} \boldsymbol{\nu}^T A A^T \boldsymbol{\nu} + \boldsymbol{\nu}^T(-AA^T \boldsymbol{\nu} - \mathbf{b}). \\ &= \max_{\boldsymbol{\nu}} -\frac{1}{2} \boldsymbol{\nu}^T A A^T \boldsymbol{\nu} - \boldsymbol{\nu}^T \mathbf{b}. \end{aligned}$$

So we have transformed into an unconstrained quadratic maximization problem.