Essentials of Convex Optimization

Max Welling
Department of Computer Science
University of Toronto
10 King's College Road
Toronto, M5S 3G5 Canada
welling@cs.toronto.edu

Abstract

This is a note to explain duality and convex optimization. It is based on Stephen Boyd's book, chapter 5 (available online).

1 Lagrangians and all that

Most kernel-based algorithms fall into two classes, either they use spectral techniques to solve the problem, or they use convex optimization techniques to solve the problem. Here we will discuss convex optimization.

A constrained optimization problem can be expressed as follows,

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 0 \; \forall i \\ & h_j(\mathbf{x}) = 0 \; \forall j \end{aligned} \tag{1}$$

That is we have inequality constraints and equality constraints. We now write the primal Lagrangian of this problem, which will be helpful in the following development,

$$\mathcal{L}_P(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x}) + \sum_j \nu_j h_j(\mathbf{x}) \tag{2}$$

where we will assume in the following that $\lambda_i \geq 0 \;\; \forall i$. From here we can define the dual problem by,

$$\mathcal{L}_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathcal{L}_P(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{3}$$

This objective can actually become $-\infty$ for certain values of its arguments. We will call parameters $\boldsymbol{\lambda} \geq 0, \boldsymbol{\nu}$ for which $\mathcal{L}_D > -\infty$ dual feasible.

It is important to notice that the dual Lagrangian is a concave function of $\boldsymbol{\lambda}, \boldsymbol{\nu}$ because it is a pointwise infimum of a function. Hence, even if the primal is not convex, the dual is certainly concave!

It is not hard to show that

$$\mathcal{L}_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^* \tag{4}$$

where p^* is the primal optimal point. This simply follows because $\sum_i \lambda_i f_i(\mathbf{x}) + \sum_j \nu_j h_j(\mathbf{x}) \leq 0$ for a primal feasible point x^*.

Thus, the dual problem always provides lower bound to the primal problem. The optimal lower bound can be found by solving the dual problem,

$$\begin{aligned} \text{maximize}_{\boldsymbol{\lambda},\boldsymbol{\nu}} \quad & \mathcal{L}_D(\boldsymbol{\lambda},\boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0 \; \forall i \end{aligned} \tag{5}$$

which is therefore a convex optimization problem. If we call d^* the dual optimal point we always have: $d^* \leq p^*$, which is called weak duality. $p^* - d^*$ is called the duality gap. Strong duality holds when $p^* = d^*$. Strong duality is very nice, in particular if we can express the primal solution $\mathbf{x}^*$ in terms of the dual solution $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$, because then we can simply solve the dual problem and convert to the answer to the primal domain since we know that solution must then be optimal. Often the dual problem is easier to solve.

So when does strong duality holds. Up to some mathematical details the answer is: *if the primal problem is convex and the equality constraints are linear*. This means that $f_0(\mathbf{x})$ and $\{f_i(\mathbf{x})\}$ are convex functions and $h_j(\mathbf{x}) = A\mathbf{x} - b$.

The primal problem can be written as follows,

$$p^* = \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \mathcal{L}_P(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) \tag{6}$$

This can be seen as follows by noting that $\sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \mathcal{L}_P(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) = f_0(\mathbf{x})$ when $\mathbf{x}$ is feasible but ∞ otherwise. To see this first check that by violating one of the constraints you can find a choice of $\boldsymbol{\lambda},\boldsymbol{\nu}$ that makes the Lagrangian infinity. Also, when all the constraints are satisfied, the best we can do is maximize the additional terms to be zero, which is always possible.

The dual problem by definition is given by,

$$d^* = \sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \inf_{\mathbf{x}} \mathcal{L}_P(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) \tag{7}$$

Hence, the "sup" and "inf" can be interchanged if strong duality holds, hence the optimal solution is a saddle-point. It is important to realize that the order of maximization and minimization matters for arbitrary functions (but not for convex functions). Try to imagine a "V" shapes valley which runs diagonally across the coordinate system. If we first maximize over one direction, keeping the other direction fixed, and then minimize the result we end up with the lowest point on the rim. If we reverse the order we end up with the highest point in the valley.

There are a number of important necessary conditions that hold for problems with zero duality gap. These Karush-Kuhn-Tucker conditions turn out to be sufficient for convex optimization problems. They are given by,

$$\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \tag{8}$$

$$f_i(\mathbf{x}^*) \leq 0 \tag{9}$$

$$h_j(\mathbf{x}^*) = 0 \tag{10}$$

$$\lambda_i^* \geq 0 \tag{11}$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0 \tag{12}$$

The first equation is easily derived because we already saw that $p^* = \inf_{\mathbf{x}} \mathcal{L}_P(\mathbf{x},\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$ and hence all the derivatives must vanish. This condition has a nice interpretation as a "balancing of forces". Imagine a ball rolling down a surface defined by $f_0(\mathbf{x})$ (i.e. you are doing gradient descent to find the minimum). The ball gets blocked by a wall, which is the constraint. If the surface and constraint is convex then if the ball doesn't move we have reached the optimal solution. At that point, the forces on the ball must balance. The

first term represent the force of the ball against the wall due to gravity (the ball is still on a slope). The second term represents the reaction force of the wall in the opposite direction. The λ represents the magnitude of the reaction force, which needs to be higher if the surface slopes more. We say that this constraint is "active". Other constraints which do not exert a force are "inactive" and have $\lambda = 0$. The latter statement can be read of from the last KKT condition which we call "complementary slackness". It says that either $f_i(\mathbf{x}) = 0$ (the constraint is saturated and hence active) in which case λ is free to take on a non-zero value. However, if the constraint is inactive: $f_i(\mathbf{x}) \leq 0$, then λ must vanish. As we will see soon, the active constraints will correspond to the support vectors in SVMs!

Complementary slackness is easily derived by,

$$
\begin{aligned}
f_0(\mathbf{x}^*) = \mathcal{L}_D(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) &= \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_i \lambda_i^* f_i(\mathbf{x}) + \sum_j \nu_j h_j(\mathbf{x}) \right) \\
&\leq f_0(\mathbf{x}^*) + \sum_i \lambda_i^* f_i(\mathbf{x}^*) + \sum_j \nu_j h_j(\mathbf{x}^*) \qquad (13) \\
&\leq f_0(\mathbf{x}^*) \qquad (14)
\end{aligned}
$$

where the first line follows from the definition the second because the inf is always smaller than any $\mathbf{x}^*$ and the last because $f_i(\mathbf{x}^*) \leq 0$, $\lambda_i^* \geq 0$ and $h_j(\mathbf{x}^*) = 0$. Hence all inequalities are equalities and each term is negative so each term must vanish separately.