On the Method of Lagrange Multipliers

Reza Nasiri Mahalati

November 6, 2016

Most of what is in this note is taken from the *Convex Optimization* book by Stephen Boyd and Lieven Vandenberghe. This should hopefully demystify the method of lagrange multipliers to some extent, and help you understand why and when this method works.

The Lagrange dual function

Generally, an optimization problem in the standard form is given by:

\[
\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad h_i(x) = 0, \quad i = 1, \ldots, p,
\end{align*}
\]

(1)

with variable \( x \in \mathbb{R}^n \). We assume its domain \( \mathcal{D} = \cap_{i=0}^m \text{dom } f_i \cap \cap_{i=1}^p \text{dom } h_i \) is nonempty, and denote the optimal value of (1) by \( p^* \). We use \( \text{dom} \) to denote the domain of a function. As an example, the general norm minimization with equality constraints that we discussed in class is a special case of (1) where:

\[
\begin{align*}
f_0(x) &= (1/2)\|Ax - b\|^2 \\
h_i(x) &= \tilde{c}^T_i x - d_i, \quad i = 1, \ldots, p,
\end{align*}
\]

(2)

and there are no inequality constraints (i.e. there are no \( f_i(x) \quad i = 1, \ldots, m \)). We simply write the \( p \) equality constraints in the matrix form as \( Cx - d = 0 \).

The basic idea in Lagrangian duality is to take the constraints in (1) into account by augmenting the objective function with a weighted sum of the constraint functions. We define the *Lagrangian* \( L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \) associated with the problem (1) as

\[
L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x),
\]

with \( \text{dom } L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \). We refer to \( \lambda_i \) as the *Lagrange multiplier* associated with the \( i \)th inequality constraint \( f_i(x) \leq 0 \); similarly we refer to \( \nu_i \) as the Lagrange multiplier associated with the \( i \)th equality constraint \( h_i(x) = 0 \). The vectors \( \lambda \) and \( \nu \) are called the *dual variables* or *Lagrange multiplier vectors* associated with the problem (1).
We define the Lagrange dual function (or just dual function) \( g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R} \) as the minimum value of the Lagrangian over \( x \), for \( \lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p \),

\[
g(\lambda, \nu) = \min_{x \in D} L(x, \lambda, \nu) = \min_{x \in D} \left( f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right).
\]

When the Lagrangian is unbounded below in \( x \), the dual function takes on the value \(-\infty\).

Since the dual function is the pointwise minimum of a family of affine functions of \((\lambda, \nu)\), it is always concave and hence, we can always find its maximum.

### Lower bounds on optimal value

The dual function yields lower bounds on the optimal value \( p^* \) of the problem (1). For any \( \lambda \succeq 0 \) and any \( \nu \) we have

\[
g(\lambda, \nu) \leq p^* \tag{3}
\]

This important property is easily verified. Suppose \( \tilde{x} \) is a feasible point for the problem (1), i.e., \( f_i(\tilde{x}) \leq 0 \) and \( h_i(\tilde{x}) = 0 \), and \( \lambda \succeq 0 \). Then we have

\[
\sum_{i=1}^m \lambda_i f_i(\tilde{x}) + \sum_{i=1}^p \nu_i h_i(\tilde{x}) \leq 0,
\]

since each term in the first sum is nonpositive, and each term in the second sum is zero, and therefore

\[
L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^m \lambda_i f_i(\tilde{x}) + \sum_{i=1}^p \nu_i h_i(\tilde{x}) \leq f_0(\tilde{x}).
\]

Hence

\[
g(\lambda, \nu) = \min_{x \in D} L(x, \lambda, \nu) \leq L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x}).
\]

Since \( g(\lambda, \nu) \leq f_0(\tilde{x}) \) holds for every feasible point \( \tilde{x} \), the inequality (3) follows. The lower bound (3) is illustrated in figure 1, for a simple problem with \( x \in \mathbb{R} \) and one inequality constraint. The inequality (3) holds, but is vacuous, when \( g(\lambda, \nu) = -\infty \). The dual function gives a nontrivial lower bound on \( p^* \) only when \( \lambda \succeq 0 \) and \( (\lambda, \nu) \in \text{dom} g \), i.e., \( g(\lambda, \nu) > -\infty \). We refer to a pair \((\lambda, \nu)\) with \( \lambda \succeq 0 \) and \( (\lambda, \nu) \in \text{dom} g \) as dual feasible.

### Linear approximation interpretation

The Lagrangian and lower bound property can be given a simple interpretation, based on a linear approximation of the indicator functions of the sets \( \{0\} \) and \( -\mathbb{R}_+ \). We first rewrite the original problem (1) as an unconstrained problem,

\[
\text{minimize} \quad f_0(x) + \sum_{i=1}^m I_-(f_i(x)) + \sum_{i=1}^p I_0(h_i(x)), \tag{4}
\]

where \( I_- : \mathbb{R} \to \mathbb{R} \) is the indicator function for the nonpositive reals,

\[
I_-(u) = \begin{cases} 
0 & u \leq 0 \\
\infty & u > 0,
\end{cases}
\]
Figure 1: Lower bound from a dual feasible point. The solid curve shows the objective function $f_0$, and the dashed curve shows the constraint function $f_1$. The feasible set is the interval $[-0.46, 0.46]$, which is indicated by the two dotted vertical lines. The optimal point and value are $x^* = -0.46$, $p^* = 1.54$ (shown as a circle). The dotted curves show $L(x, \lambda)$ for $\lambda = 0.1, 0.2, \ldots, 1.0$. Each of these has a minimum value smaller than $p^*$, since on the feasible set (and for $\lambda \geq 0$) we have $L(x, \lambda) \leq f_0(x)$. 
and similarly, \( I_0 \) is the indicator function of \( \{0\} \). In the formulation (4), the function \( I_-(u) \) can be interpreted as expressing our irritation or displeasure associated with a constraint function value \( u = f_i(x) \): It is zero if \( f_i(x) \leq 0 \), and infinite if \( f_i(x) > 0 \). In a similar way, \( I_0(u) \) gives our displeasure for an equality constraint value \( u = h_i(x) \). We can think of \( I_- \) as a “brick wall” or “infinitely hard” displeasure function; our displeasure rises from zero to infinite as \( f_i(x) \) transitions from nonpositive to positive.

Now suppose in the formulation (4) we replace the function \( I_-(u) \) with the linear function \( \lambda_i u \), where \( \lambda_i \geq 0 \), and the function \( I_0(u) \) with \( \nu_i u \). The objective becomes the Lagrangian function \( L(x, \lambda, \nu) \), and the dual function value \( g(\lambda, \nu) \) is the optimal value of the problem

\[
\minimize \quad L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x).
\]

In this formulation, we use a linear or “soft” displeasure function in place of \( I_- \) and \( I_0 \). For an inequality constraint, our displeasure is zero when \( f_i(x) = 0 \), and is positive when \( f_i(x) > 0 \) (assuming \( \lambda_i > 0 \)); our displeasure grows as the constraint becomes “more violated”. Unlike the original formulation, in which any nonpositive value of \( f_i(x) \) is acceptable, in the soft formulation we actually derive pleasure from constraints that have margin, i.e., from \( f_i(x) < 0 \).

Clearly the approximation of the indicator function \( I_-(u) \) with a linear function \( \lambda_i u \) is rather poor. But the linear function is at least an underestimator of the indicator function. Since \( \lambda_i u \leq I_-(u) \) and \( \nu_i u \leq I_0(u) \) for all \( u \), we see immediately that the dual function yields a lower bound on the optimal value of the original problem.

**The Lagrange dual problem**

For each pair \( (\lambda, \nu) \) with \( \lambda \geq 0 \), the Lagrange dual function gives us a lower bound on the optimal value \( p^* \) of the optimization problem (1). Thus we have a lower bound that depends on some parameters \( \lambda, \nu \). A natural question is: What is the best lower bound that can be obtained from the Lagrange dual function?

This leads to the optimization problem

\[
\maximize \quad g(\lambda, \nu) \\
\text{subject to} \quad \lambda \geq 0.
\]

This problem is called the Lagrange dual problem associated with the problem (1). In this context the original problem (1) is sometimes called the primal problem. The term dual feasible, to describe a pair \( (\lambda, \nu) \) with \( \lambda \geq 0 \) and \( g(\lambda, \nu) > -\infty \), now makes sense. It means, as the name implies, that \( (\lambda, \nu) \) is feasible for the dual problem (6). We refer to \( (\lambda^*, \nu^*) \) as dual optimal or optimal Lagrange multipliers if they are optimal for the problem (6).

The Lagrange dual problem (6) is always a convex optimization problem, since the objective to be maximized is concave and the constraint is convex. Therefore, we can always solve this problem.
Weak duality

The optimal value of the Lagrange dual problem, which we denote $d^\star$, is, by definition, the best lower bound on $p^\star$ that can be obtained from the Lagrange dual function. In particular, we have the simple but important inequality

$$d^\star \leq p^\star,$$

which holds for any general optimization problem. This property is called weak duality.

The weak duality inequality (7) holds when $d^\star$ and $p^\star$ are infinite. For example, if the primal problem is unbounded below, so that $p^\star = -\infty$, we must have $d^\star = -\infty$, i.e., the Lagrange dual problem is infeasible. Conversely, if the dual problem is unbounded above, so that $d^\star = \infty$, we must have $p^\star = \infty$, i.e., the primal problem is infeasible.

We refer to the difference $p^\star - d^\star$ as the optimal duality gap of the original problem, since it gives the gap between the optimal value of the primal problem and the best (i.e., greatest) lower bound on it that can be obtained from the Lagrange dual function. The optimal duality gap is always nonnegative.

The bound (7) can sometimes be used to find a lower bound on the optimal value of a problem that is difficult to solve, since the dual problem is always convex and can be solved efficiently, to find $d^\star$.

Strong duality

If the equality

$$d^\star = p^\star$$

holds, i.e., the optimal duality gap is zero, then we say that strong duality holds. This means that the best bound that can be obtained from the Lagrange dual function is tight.

Strong duality does not, in general, hold. But if the primal problem (1) is convex, i.e., of the form

$$\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m, \\
& \quad Cx = d,
\end{align*}$$

with $f_0, \ldots, f_m$ convex functions, we usually (but not always) have strong duality. There are many results that establish conditions on the problem, beyond convexity, under which strong duality holds. These conditions are called constraint qualifications.

The optimization problems we will be solving in EE263 are always convex, and since we don’t work with inequality constraints in this course, we need not worry about constraint qualifications. In other words, strong duality always holds in EE263, except for the case where the constraint $Cx = d$ cannot be satisfied for any $x \in \mathcal{D}$, which means the problem is infeasible and cannot be solved.

Now suppose that the primal and dual optimal values are attained and equal (so, in particular, strong duality holds). Let $x^\star$ be a primal optimal and $(\lambda^\star, \nu^\star)$ be a dual optimal
point. This means that

\[
    f_0(x^*) = g(\lambda^*, \nu^*) \\
    = \inf_x \left( f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right) \\
    \leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\
    \leq f_0(x^*).
\]

The first line states that the optimal duality gap is zero, and the second line is the definition of the dual function. The third line follows since the infimum of the Lagrangian over \( x \) is less than or equal to its value at \( x = x^* \). The last inequality follows from \( \lambda_i^* \geq 0, f_i(x^*) \leq 0, \) \( i = 1, \ldots, m, \) and \( h_i(x^*) = 0, \) \( i = 1, \ldots, p. \) We conclude that the two inequalities in this chain hold with equality.

We can draw several interesting conclusions from this. For example, since the inequality in the third line is an equality, we conclude that \( x^* \) minimizes \( L(x, \lambda^*, \nu^*) \) over \( x \). (The Lagrangian \( L(x, \lambda^*, \nu^*) \) can have other minimizers; \( x^* \) is simply a minimizer.)

Another important conclusion is that

\[
    \sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0.
\]

Since each term in this sum is nonpositive, we conclude that

\[
    \lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m. \tag{10}
\]

This condition is known as *complementary slackness*; it holds for any primal optimal \( x^* \) and any dual optimal \( (\lambda^*, \nu^*) \) (when strong duality holds). We can express the complementary slackness condition as

\[
    \lambda_i^* > 0 \implies f_i(x^*) = 0,
\]

or, equivalently,

\[
    f_i(x^*) < 0 \implies \lambda_i^* = 0.
\]

Roughly speaking, this means the \( i \)th optimal Lagrange multiplier is zero unless the \( i \)th constraint is active at the optimum.

**KKT conditions**

We now assume that the functions \( f_0, \ldots, f_m, h_1, \ldots, h_p \) are differentiable (and therefore have open domains), but we make no assumptions yet about convexity. As above, let \( x^* \) and \( (\lambda^*, \nu^*) \) be any primal and dual optimal points with zero duality gap. Since \( x^* \) minimizes \( L(x, \lambda^*, \nu^*) \) over \( x \), it follows that its gradient must vanish at \( x^* \), i.e.,

\[
    \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0.
\]
Thus we have
\begin{align*}
f_i(x^*) &\leq 0, \quad i = 1, \ldots, m \\
h_i(x^*) &= 0, \quad i = 1, \ldots, p \\
\lambda_i^* &\geq 0, \quad i = 1, \ldots, m \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) &= 0, \quad (11)
\end{align*}

which are called the Karush-Kuhn-Tucker (KKT) conditions.

To summarize, for any optimization problem with differentiable objective and constraint functions for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions (11).

When the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal. In other words, if \( f_i \) are convex and \( h_i \) are affine, and \( \tilde{x}, \tilde{\lambda}, \tilde{\nu} \) are any points that satisfy the KKT conditions
\begin{align*}
f_i(\tilde{x}) &\leq 0, \quad i = 1, \ldots, m \\
h_i(\tilde{x}) &= 0, \quad i = 1, \ldots, p \\
\lambda_i &\geq 0, \quad i = 1, \ldots, m \\
\lambda_i f_i(\tilde{x}) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^p \tilde{\nu}_i \nabla h_i(\tilde{x}) &= 0,
\end{align*}
then \( \tilde{x} \) and \( (\tilde{\lambda}, \tilde{\nu}) \) are primal and dual optimal, with zero duality gap.

What we did with the method of Lagrange multipliers in class, was precisely to form and solve the KKT system. To see this, note that the KKT conditions for our general norm minimization problem would be:
\begin{align*}
C \tilde{x} - d &= \nabla \nu L(\tilde{x}, \tilde{\nu}) = 0, \\
\nabla f_0(\tilde{x}) + \sum_{i=1}^p \tilde{\nu}_i \nabla h_i(\tilde{x}) &= \nabla x L(\tilde{x}, \tilde{\nu}) = 0,
\end{align*}
which is exactly the system of equations we got from the method of Lagrange multipliers.