1 Lagrange multipliers for a single-constraint problem

We will consider the problem

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad g(x) = 0
\end{align*}
\]

where \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g : \mathbb{R}^n \to \mathbb{R} \) are functions with continuous first derivatives. As an example problem, consider

\[
\begin{align*}
\text{minimize} & \quad x + y \\
\text{subject to} & \quad x^2 + 2y^2 - 4 = 0.
\end{align*}
\]

Here is a geometric representation of this problem:

The black curve is the set of feasible solutions: points \((x, y)\) such that \(x^2 + 2y^2 - 4 = 0\). The red lines are lines of the form \(x + y = c\): points where the objective function achieves a certain value.

We want to choose a red line that’s as far left as possible while still intersecting the black curve. (And then pick a point on the intersection.) Intuitively speaking, this means choosing a red line that’s tangent to the black curve.

In general, a curve \(g(x) = 0\) has a tangent hyperplane at a point \(x^*\) given by

\[
\nabla g(x^*) \cdot (x - x^*) = 0.
\]
In this two-dimensional case, the gradient of $g(x, y) = x^2 + 2y^2 - 4$ is $(2x, 4y)$, so if the optimal solution is a point $(x^*, y^*)$, the tangent line has the form

$$\begin{bmatrix} 2x^* \\ 4y^* \end{bmatrix} \cdot \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} = 0 \iff 2x^*(x) + 4y^*(y) = 2(x^*)^2 + 4(y^*)^2.$$ 

We want this to be a line of the form $x + y = c$, so the coefficients of $x$ and $y$ must be equal: we want $2x^* = 4y^*$.

Now, solving $2x = 4y$ together with $x^2 + 2y^2 - 4 = 0$ gives us the two solutions; $(2\sqrt{2}/3, \sqrt{2}/3)$ and $(-2\sqrt{2}/3, -\sqrt{2}/3)$. Of these, the positive solution is a maximizer and the negative solution is the minimizer we want.

This approach generalizes.

**Theorem 1.1.** If $x^*$ is an optimal solution to the problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to $g(x) = 0$

then either $\nabla f(x^*) = \lambda \nabla g(x^*)$ for some $\lambda \in \mathbb{R}$, or else $\nabla g(x^*) = \mathbf{0}$ (assuming that $\nabla f, \nabla g$ exist and are continuous).

**Proof.** This will only be an informal sketch of a proof; the formal proof is very similar, but involves lots more epsilons.

Let $x$ be any point satisfying $g(x) = 0$. To see how good $x$ is at minimizing $f$, we compare $x$ to another point $y$ satisfying $g(y) = 0$, very close to $x$.

To a first approximation, we can get such points $y$ by taking $y = x + tu$, where $t > 0$ is some small scalar, and $u$ is a direction satisfying $\nabla g(x) \cdot u = 0$. This choice of $u$ means that $g$ is not changing in that direction, so $g(y)$ will stay 0 because $g(x)$ was 0.

(Of course, this is only an approximation, since $g$ is not necessarily linear and so $\nabla g(x)$ may change as we move away from $x$. In practice, we still know that $\nabla g(x) \cdot u$ can be made arbitrarily close to 0 provided that $y$ is sufficiently close to $x$.)

Moving from $x$ to $x + tu$ changes the objective value $f$ by approximately $\nabla f(x) \cdot tu = t(\nabla f(x) \cdot u)$. If it is possible to choose $u$ such that $\nabla f(x) \cdot u \neq 0$, then we can choose a small positive and small negative $t$ to move to a point better than $x$.

So if $x^*$ is an optimal solution, it should not be possible to choose such an $u$: whenever $\nabla g(x^*) \cdot u = 0$, we should have $\nabla f(x^*) \cdot u = 0$. This happens when $\nabla f(x^*)$ and $\nabla g(x^*)$ are parallel: $\nabla f(x^*) = \lambda \nabla g(x^*)$ for some $\lambda \in \mathbb{R}$.

The exceptional case when this reasoning does not work is when $\nabla g(x^*) = \mathbf{0}$. In this case, we cannot approximate $g(x)$ near $x^*$ by using the gradient $\nabla g(x^*)$: the dominant term will be the second-order term coming from $Hg(x^*)$, if it exists, or else complete chaos occurs if the second derivative does not exist. So all such points must be checked separately. \qed
We call the constant $\lambda$ the “Lagrange multiplier” of $g$.

You should think of the $\nabla g(x^*) = 0$ case of this theorem as a “the problem is ill-behaved” case. For a “nice” constraint, there should not be any points at all where $g(x) = 0$ and also $\nabla g(x) = 0$. But sometimes, there are such points, and we have to check them separately, because we can’t make predictions at such points.

Sometimes, everything is horrible. For example, if we change the constraint $x^2 + 2y^2 - 4 = 0$ in our example to $(x^2 + 2y^2 - 4)^2 = 0$, this is an equivalent constraint that is ill-behaved everywhere it is satisfied. About such a constraint, no conclusions can be made.

The theorem is not an if-and-only-if statement: an optimal solution $x^*$ is guaranteed to satisfy the conditions, but so can other points. Assuming that an optimal solution exists, it will be one of the points satisfying the conditions.

Whether an optimal solution exists or not is a question that we need other methods to answer. For example, we know that an optimal solution exists if the set $\{x \in \mathbb{R}^n : g(x) = 0\}$ is bounded, because it must be closed. We also know that an optimal solution exists if $f(x)$ is coercive.

## 2 Generalizations

### 2.1 Multiple constraints

Now let’s generalize and consider the problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to

$$g_1(x) = 0,$$

$$g_2(x) = 0,$$

$$\vdots$$

$$g_m(x) = 0.$$  

(We often write these constraints as $g(x) = 0$, thinking of $g_1, g_2, \ldots, g_m$ as components of a vector-valued function $g : \mathbb{R}^n \to \mathbb{R}^m$.)

**Theorem 2.1.** Assume that $\nabla f, \nabla g_1, \nabla g_2, \ldots, \nabla g_m$ exist and are continuous.

If $x^*$ is an optimal solution to the problem above, then either

$$\nabla f(x^*) + \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*) + \cdots + \lambda_m \nabla g_m(x^*) = 0$$

or else the gradients $\nabla g_1, \nabla g_2, \ldots, \nabla g_m$ are linearly dependent at $x^*$.

**Proof.** As an even less formal sketch of a proof, let’s consider how our previous argument changes.

Now, if we are moving from a point $x$ such that $g_i(x) = 0$ for $i = 1, \ldots, m$ to another point $y$ very close to $x$, then we should be moving in a direction $u$ such that $\nabla g_i(x) \cdot u = 0$ for all $i$. 

3
If there is such a direction such that, additionally, $\nabla f(x) \cdot u \neq 0$, then going a short distance in the direction $u$ or $-u$ will increase the value of $f$. So at an optimal point $x^*$, there cannot be any such direction: having $\nabla g_i(x^*) \cdot u = 0$ for all $i$ must imply $\nabla f(x^*) \cdot u = 0$.

This can only happen if $\nabla f(x^*)$ can be written as a linear combination of $\nabla g_1(x^*), \nabla g_2(x^*), \ldots, \nabla g_m(x^*)$.

It’s a harder to see why the problem becomes ill-behaved when $\nabla g_1, \nabla g_2, \ldots, \nabla g_m$ are linearly dependent at a point. Intuitively, this represents a situation where some of the hypersurfaces $\{x: g_i(x) = 0\}$ act like they’re tangent to each other, where the first-order approximation is again no longer sufficient to understand the nearby feasible solutions.

2.2 Inequalities

It’s somewhat standard to use Lagrange multipliers to solve a problem like

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$

subject to $g(x) \leq 0$.

Here, we consider the cases $g(x) = 0$ and $g(x) < 0$ separately. In the first case, we use the Lagrange multiplier method. In the second case, we are in an open set with no boundaries, so an optimal solution must be a critical point of $f$. So we find all cases where $\nabla f(x) = 0$, and check the sign of $g(x)$ at those points.

Things are trickier when we have multiple inequalities. For example, consider a problem of the form

$$\min_{(x,y) \in \mathbb{R}^2} \quad f(x,y)$$

subject to $x^2 + 2y^2 - 4 \leq 0,$

$$2x^2 + y^2 - 4 \leq 0.$$  

Some functions, such as $f(x,y) = x + y$, will be minimized at a point where both constraints are tight (satisfied with equality). Other functions, such as $f(x,y) = x$, are minimized at a point where only one constraint is tight. Still other functions, such as $f(x,y) = x^2 + y^2$, will be minimized at a point in the interior of both constraints.

The number of cases to check increases exponentially with the number of constraints. Our goal in the next few weeks will be to:

- Extend the method of Lagrange multipliers to inequalities in a way that considers all of these cases at once.
- Understand a different set of conditions for when a problem is sufficiently well-behaved to apply the method.