LAGRANGE MULTIPLIERS

MATH 195, SECTION 59 (VIPUL NAIK)

Corresponding material in the book: Section 14.8

What students should definitely get: The Lagrange multiplier condition (one constraint, two constraints and in principle more than two constraints), the application to finding absolute extreme values.

What students should hopefully get: Situations where Lagrange multipliers fail, the underlying logic behind Lagrange multipliers, how to use Lagrange multipliers for piecewise smooth situations.

Executive summary

Words ...

(1) Two of the reasons why the derivative of a function may be zero: the function is constant around the point, or the function has a local extreme value at the point.

Version for many variables: two of the reasons why the gradient vector of a function of many variables may be zero: the function is constant around the point, or the function has a local extreme value at the point.

Version for function restricted to a subset smooth around a point: two of the reasons why the gradient vector may be *orthogonal* to the subset at the point: the function is constant on the subset around the point, or the function has a local extreme value (relative to the subset) at the point.

(2) For a function f defined on a subset smoooth around a point (i.e., with a well defined tangent and normal space), if f has a local extreme value at the point when restricted to the subset, then ∇f lives in the normal direction to the subset (this includes the possibility of it being zero).

(3) For a codimension one subset of $\mathbb{R}^n$ defined by a condition $g(x_1, x_2, \ldots, x_n) = k$, if a point $(a_1, a_2, \ldots, a_n)$ gives a local extreme value for a function f of n variables, and if ∇g is well defined and nonzero at the point, then there exists a real number λ such that $\nabla f(a_1, a_2, \ldots, a_n) = \lambda \nabla g(a_1, a_2, \ldots, a_n)$. Note that λ may be zero.

(4) Suppose a codimension r subset of $\mathbb{R}^n$ is given by r independent constraints $g_1(x_1, x_2, \ldots, x_n) = k_1$, $g_2(x_1, x_2, \ldots, x_n) = k_2$, and so on till $g_r(x_1, x_2, \ldots, x_n) = k_r$. Suppose ∇g_i are nonzero for all i at a point $(a_1, a_2, \ldots, a_n)$ of local extreme value for a function f relative to this subset. Suppose further that all the ∇g_i are linearly independent. Then $\nabla f(a_1, a_2, \ldots, a_n)$ is a linear combination of the vectors $\nabla g_1(a_1, a_2, \ldots, a_n)$, $\nabla g_2(a_1, a_2, \ldots, a_n)$, $\ldots$, $\nabla g_r(a_1, a_2, \ldots, a_n)$. In other words, there exist real numbers λ_1, λ_2, $\ldots$, λ_r such that:

$$\nabla f(a_1, a_2, \ldots, a_n) = \lambda_1 \nabla g_1(a_1, a_2, \ldots, a_n) + \lambda_2 \nabla g_2(a_1, a_2, \ldots, a_n) + \cdots + \lambda_r \nabla g_r(a_1, a_2, \ldots, a_n)$$

(5) The Lagrange condition may be violated at points of local extremum where ∇g is zero, or more generally, where the ∇g_i fail to be linearly independent. This may occur either because the tangent and normal space are not well defined or because the functions fail to capture it well.

Actions ...

(1) Suppose we want to maximize and minimize f on the set $g(x_1, x_2, \ldots, x_n) = k$. Assume $\nabla g(x_1, x_2, \ldots, x_n)$ is defined everywhere on the set and never zero. Suppose ∇f is also defined. Then, all local maxima and local minima are attained at points where $\nabla f = \lambda \nabla g$ for some real number λ. To find these, we solve the system of $n + 1$ equations in the $n + 1$ variables x_1, x_2, $\ldots$, x_n, namely the n scalar equations from the Lagrange condition and the equation $g(x_1, x_2, \ldots, x_n) = k$.

To find the actual extreme values, once we've collected all candidate points from the above procedure, we evaluate the function at all these and find the largest and smallest value to find the absolute maximum and minimum.

(2) If there are points in the domain where ∇g takes the value 0, these may also be candidates for local extreme values, and the function should additionally be evaluated at these as well to find the absolute maximum and minimum.
(3) A similar procedure works for a subset given by r constraints. In this case, we have the equation:

$$\nabla f(a_1, a_2, \ldots, a_n) = \lambda_1 \nabla g_1(a_1, a_2, \ldots, a_n) + \lambda_2 \nabla g_2(a_1, a_2, \ldots, a_n) + \cdots + \lambda_r \nabla g_r(a_1, a_2, \ldots, a_n)$$

as well as the r equations $g_1(x_1, x_2, \ldots, x_n) = k_1$, $g_2(x_1, x_2, \ldots, x_n) = k_2$, and so on. In total, we have $n + r$ equations in $n + r$ variables: the x_1, x_2, ..., x_n and the λ_1, λ_2, ..., λ_r.

1. Lagrange multipliers: basic formulation with one constraint

1.1. **The two key ideas.** We summarize two key ideas behind Lagrange multipliers:

- If a scalar function is constant on a subset of $\mathbb{R}^n$, its directional derivative along any direction tangent to the subset at a point on the subset is zero. Thus, the gradient of the function at any point in the subset (if nonzero) is orthogonal to the subset.
- Consider a scalar function and a subset of $\mathbb{R}^n$. At any point in the subset where the function attains a local extreme value relative to the subset, the directional derivative along any direction tangent to the subset at the point is zero. Thus, the gradient of the function at the point (if nonzero) is orthogonal to the subset.

Roughly speaking, we're saying that there are two reasons (among many!) why the directional derivative along all tangents at a point should be zero: one, the function is constant around the point, and the other, the function attains a local extreme value at the point.

The key insight behind Lagrange multipliers is to combine these insights and ask: *on a subset defined by one function being constant, how do we find the local extreme values of another function*? The idea is to use the fact that both for the function that's constant and the function that is attaining an extreme value, the gradient is normal (orthogonal) to the subset. If the subset is given by a single constraint, then it has codimension one, so the normal space is one-dimensional, and this forces the gradient vectors for the two functions to be scalar multiples of each other (with suitable assumptions of being nonzero).

1.2. **Getting started.** We know that for a differentiable function f defined in an open domain in $\mathbb{R}^n$, if the function has a local extreme value at a point $(a_1, a_2, \ldots, a_n)$, then the directional derivative of f along *every* direction in $\mathbb{R}^n$ is zero. More specifically, if f has a local extreme value *along* a particular line (both directions on that line) then the directional derivative along that particular direction is zero.

Now, suppose we want to maximize f, not on an open domain in $\mathbb{R}^n$, but on a subset of smaller dimension that has a well defined tangent space at the point. Then, the key idea is that a *necessary* condition for a point to have a local extreme value is that the directional derivative along all tangent directions to that subset are zero. *However, the directional derivative along non-tangential directions may well be nonzero.*

In other words, the gradient of the function f does not have any component tangential to the subset. Thus, the gradient of f, ∇f, is *either zero* or is a vector *perpendicular* to the tangent space at the point, i.e., a vecor *in* the normal space at the point.

We now turn to some specific cases for $n = 2$ and $n = 3$.

1.3. **In two dimensions: setup.** Suppose we have a function f in two variables, and a smooth curve γ. We want to find the extreme values of f along the curve γ. Suppose f attains a local maximum at (x_0, y_0) in the curve γ. This just means that f has a local maximum *relative* to the curve γ, i.e., if we take points in the curve γ close to the point (x_0, y_0), then the f-value at those points in γ is less than or equal to the value $f(x_0, y_0)$.

This implies that if we move slightly *along* γ, or tangential to γ, then the directional derivative of f is zero. The reason: the directional derivative along one direction on γ is less than or equal to zero, because the function is smaller if we move a bit in that direction. Similarly, the directioal derivative along the reverse direction is less than or equal to zero. Since these directional derivatives are negatives of each other, this forces both of them to be zero.

Thus, the directional derivative of f along the tangent direction to γ is zero. This means that the dot product of the gradient of f and the unit tangent vector to γ is zero, so ∇f is either equal to zero or points in a direction perpendicular to the tangent direction to γ.

1.4. **γ as the level curve of g.** We now continue with the same setup as above, now setting up γ as a level curve of another differentiable function g, i.e., γ is defined as the set $g(x,y) = k$ for some constant k. We want to find the tangent and normal vectors at a point (x_0, y_0) to this curve.

Since g is not changing along γ, ∇g has a component of zero along γ *at every point of* γ. In particular, if ∇g is nonzero, it is along the normal direction to γ.

Thus, the upshot is that if (x_0, y_0) is a point of extreme value for f on the curve γ defined as $g(x,y) = k$, then $\nabla g(x_0, y_0)$, if nonzero, is normal to the curve at the point, and $\nabla f(x_0, y_0)$, if nonzero, is normal to the curve at the point. Thus, $\nabla f(x_0, y_0)$ is a scalar multiple of $\nabla g(x_0, y_0)$, i.e., there is a constant λ such that:

$$\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0)$$

This constant λ is termed a *Lagrange multiplier*.

1.5. **Case of $n = 3$.** We quickly state the similar result for 3 variables. We want to find extreme values for a function $f(x,y,z)$ on the *surface* (codimension one subset) $g(x,y,z) = k$. Then, if ∇g is nonzero, it is in the normal direction to the surface, and thus, at a point (x_0, y_0, z_0) where f has a local extreme, ∇f is a scalar multiple of ∇g, i.e., we have:

$$\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)$$

1.6. **General statement.** Suppose we have two functions f and g, both of n variables. In other words, $f(x_1, x_2, \ldots, x_n)$ is a function of n variables and $g(x_1, x_2, \ldots, x_n)$ is also a function of n variables. Suppose further that k is a real number, and suppose that ∇g is nonzero everywhere on the codimension one subset $g(x_1, x_2, \ldots, x_n) = k$. Suppose $(a_1, a_2, \ldots, a_n)$ is a point satisfying $g(a_1, a_2, \ldots, a_n) = k$, and such that f has a local extreme value at $(a_1, a_2, \ldots, a_n)$ *when restricted to the subset* $g(x_1, x_2, \ldots, x_n) = k$. Then, there exists a scalar λ such that:

$$(\nabla f)(a_1, a_2, \ldots, a_n) = \lambda \nabla g(a_1, a_2, \ldots, a_n)$$

In other words, the directional derivative of f is a scalar multiple of the directional derivative of g at the point. Another way of thinking of this is that the directional derivative of f has no component tangential to the subset $g(x_1, x_2, \ldots, x_n) = k$.

Note that if f has a local extreme value at the point $(a_1, a_2, \ldots, a_n)$ with respect to the whole space (and not just the codimension one subset $g(x_1, x_2, \ldots, x_n) = k$) then in fact $\nabla f(a_1, a_2, \ldots, a_n)$ is the zero vector, so $\lambda = 0$ in this case. This is a much stronger condition.

2. Lagrange multipliers: multiple constraints

The Lagrange multiplier applications we have seen so far concentrate on codimension one subsets, i.e., subsets that are given as solutions to $g(x,y) = k$ for a single constraint g. However, the ideas generalize a little further to multiple constraints. The key difference is that the normal space is more than one-dimensional.

2.1. **Arithmetic of dimension and codimension.** Here's a quick recall of the arithmetic of dimension and codimension. If a subset of $\mathbb{R}^n$ is specified in a top-down fashion by r independent scalar equality constraints, then the subset has *codimension* r and *dimension* $n - r$. The way to think of this is that we start with the whole n-dimensional space and each new constraint reduces the dimension by 1, provided it is independent of all the previous constraints.

Let's recall what this means for tangent and normal spaces. If the subset is sufficiently smooth around a point in the subset, then we can define the tangent space to the subset about the point. The tangent space is a flat (linear) space at the point, and it has the same dimension as the subset, which in our case is $n - r$. The *normal space* is a space of dimension r orthogonal to the tangent space at the point, i.e., every vector in the normal space is orthogonal to every vector in the tangent space.

We now turn to a new aspect: *actually describing the normal space*. We know that for a function given by $g(x_1, x_2, \ldots, x_n) = k$, then the normal vector (unique up to scaling) at a point $(a_1, a_2, \ldots, a_n)$ is given by $\nabla g(a_1, a_2, \ldots, a_n)$, if that vector is nonzero.

Suppose we consider the subset of $\mathbb{R}^n$ satisfying this collection of r constraints, where all the functions g_i are differentiable functions:

$$\begin{aligned} g_1(x_1, x_2, \ldots, x_n) &= k_1 \\ g_2(x_1, x_2, \ldots, x_n) &= k_2 \\ \cdot &= \cdot \\ g_r(x_1, x_2, \ldots, x_n) &= k_r \end{aligned}$$

Then, at a point $(a_1, a_2, \ldots, a_n)$ in the subset, each of the gradient vectors $\nabla g_1(a_1, a_2, \ldots, a_n)$, $\nabla g_2(a_1, a_2, \ldots, a_n)$, $\ldots$, $\nabla g_r(a_1, a_2, \ldots, a_n)$, if nonzero, is orthogonal to the subset at the point $(a_1, a_2, \ldots, a_n)$. If the constraints are all independent at the point, then we get a bunch of *linearly independent* (whatever that means) vectors that *span* (whatever that means) the normal space.

For a function f on $\mathbb{R}^n$, if $(a_1, a_2, \ldots, a_n)$ is a point in the $(n-r)$-dimensional subset where f attains an extreme value, then the directional derivative of f along any direction tangent to the subset is zero. Thus, the gradient vector ∇f is in the normal space. So, we can find constants λ_1, λ_2, $\ldots$, λ_r such that:

$$(\nabla f)(a_1, a_2, \ldots, a_n) = \lambda_1 \nabla g_1(a_1, a_2, \ldots, a_n) + \lambda_2 \nabla g_2(a_1, a_2, \ldots, a_n) + \cdots + \lambda_r \nabla g_r(a_1, a_2, \ldots, a_n)$$

Unfortunately, a deeper understanding of the ideas here requires a rudimentary understanding of linear algebra, which very few of you have had.

2.2. **Curves in $\mathbb{R}^3$.** Consider the case $n = 3$ and $r = 2$, i.e., we have a curve in $\mathbb{R}^3$ given by a pair of independent scalar equality constraints. Let's say the constraints are as follows:

$$\begin{aligned} g_1(x, y, z) &= k_1 \\ g_2(x, y, z) &= k_2 \end{aligned}$$

Suppose further that ∇g_1 is not the zero vector anywhere on the curve and ∇g_2 is also not the zero vector anywhere on the curve. Suppose further that ∇g_1 and ∇g_2 are linearly independent everywhere on the curve, i.e., it is never the case that the vectors are scalar multiples of each other.[1]

Then, at a point (x_0, y_0, z_0) on the curve, the normal space is spanned by the vectors $\nabla g_1(x_0, y_0, z_0)$ and $\nabla g_2(x_0, y_0, z_0)$. If (x_0, y_0, z_0) is a point of local extreme for a function f relative to the curve, then the theory of Lagrange multipliers tells us that:

$$\nabla f(x_0, y_0, z_0) = \lambda_1 \nabla g_1(x_0, y_0, z_0) + \lambda_2 \nabla g_2(x_0, y_0, z_0)$$

3. Using Lagrange multipliers

3.1. **Finding critical points: equation setup in codimension one.** We consider optimization for a codimension one subset in $\mathbb{R}^n$, of a function $f(x_1, x_2, \ldots, x_n)$ on the subset of $\mathbb{R}^n$ given by the equation $g(x_1, x_2, \ldots, x_n) = k$, with ∇g not a zero vector anywhere on the subset. Then, we need to solve the equation:

$$\begin{aligned} \nabla f(x_1, x_2, \ldots, x_n) &= \lambda \nabla g(x_1, x_2, \ldots, x_n) \\ g(x_1, x_2, \ldots, x_n) &= k \end{aligned}$$

[1]For two vectors, linear independence just means that neither is a scalar multiple of the other. The situation is considerably more complicated for more than two vectors.

The first of these is a vector equation. In particular, when we consider it coordinate-wise, we get n scalar equations. Counting the second equation as well, we get a total of $n+1$ scalar equations. There are $n+1$ variables: x_1, x_2, ..., x_n and λ. We thus have a system of $n+1$ equations in $n+1$ variables. The solution space is thus expected to be zero-dimensional, i.e., we *expect* that the set of solutions is a discrete collection of isolated points. These are the *critical points*.

3.2. **Finding critical points: equation setup in codimension r.** In the codimension r setup in $\mathbb{R}^n$ discussed earlier, we get the following equations:

$$\begin{aligned}
\nabla f(x_1, x_2, \ldots, x_n) &= \lambda_1 \nabla g_1(x_1, x_2, \ldots, x_n) + \lambda_2 \nabla g_2(x_1, x_2, \ldots, x_n) + \cdots + \lambda_r \nabla g_r(x_1, x_2, \ldots, x_n) \\
g_1(x_1, x_2, \ldots, x_n) &= k_1 \\
g_2(x_1, x_2, \ldots, x_n) &= k_2 \\
\cdot &= \cdot \\
g_r(x_1, x_2, \ldots, x_n) &= k_r
\end{aligned}$$

The first equation is n scalar equations, so we have a total of $n+r$ scalar equations. The number of variables is also $n+r$: the n variables x_1, x_2, ..., x_n, and the r variables λ_1, λ_2, ..., λ_r. The number of equations equals the number of variables, so we expect the solution set to be a bunch of isolated points.

3.3. **Absolute maxima and minima.** After we have computed all the critical points, we need to figure out which of them give rise to local maxima, which of them give rise to local minima, which of them give neither, and what the absolute maximum and minimum are. For local maxima and minima, we need an analogue of the second derivative test, which is too hard to develop and conceptually justify here. So, we simply avoid that question and only concentrate on finding the absolute maxima/minima.

As was the case with our earlier discussion of maxima/minima, we simplify matters and only consider the case where the space $g(x_1, x_2, \ldots, x_n) = k$ is a closed bounded set. Fortunately, since g is assumed to be continuous, the set $g(x_1, x_2, \ldots, x_n) = k$ is automatically a closed subset. Boundedness is something we can check for separately. Once we have established this, we can use the extreme value theorem, and conclude that the absolute maximum/minimum are attained. The strategy for finding them is as follows in a closed bounded subset without any boundary:

> Set up equations using Lagrange multipliers (as discussed above) and solve to find all critical points, that are candidates for the absolute maximum/minimum. Then, compare the function values at all these points. The smallest among these gives the absolute minimum, and the largest among these gives the absolute maximum.

Note that the situation becomes a little more complicated for a closed bounded subset that has a boundary, because the boundary is a smaller dimension subset. In this case, we need to separately find critical points relative to the boundary. Fortunately, this does not happen for typical subsets defined by conditions of the form $g(x_1, x_2, \ldots, x_n) = k$.

3.4. **Piecewise smooth curves, curves where the gradient vector becomes zero.** Recall that a necessary condition for the Lagrange condition for codimension one to hold at a point of local extremum is that ∇g be nonzero, i.e., the gradient of g have a well defined direction. If $\nabla g = 0$, the Lagrange condition may be violated at a local extreme value.

Similarly, for the Lagrange condition to hold in higher codimension, what we need is that ∇g_i be nonzero for each i, and further, that all the ∇g_i be linearly independent (whatever that means). Since the higher codimension case requires some knowledge of linear algebra, we'll skip it for now and stick to the codimension one case.

In the codimension one case, the condition $\nabla g = 0$ could arise for either of two reasons. First, there is no well defined normal direction to the codimension one subset at the point. This may happen because of a sharp cusp-like appearance or sudden direction change near the point, like the vertex of a cone. Second, the normal direction may be well defined but the function g may simply have been chosen poorly. For instance, consider $g(x, y) = (x - y)^3$ in two dimensions. The level curves for this are lines parallel to $y = x$. For each

such line, there is a well defined normal direction. However, for the line $y = x$ itself, although a normal direction does exist, ∇g takes the value zero.

To handle these kinds of situations we add in the following caveat to the Lagrange method:

> In addition to testing all points where the Lagrange condition holds, *also* test all points where $\nabla g = 0$ (if such points exist), when trying to find absolute maximum and minimum.

The technique does not work if we end up with infinitely many points satisfying $\nabla g = 0$. In this case, an alternative approach might work.

> Find a new function h and a constant l such that the set $g(x_1, x_2, \ldots, x_n) = k$ is the same as the set $h(x_1, x_2, \ldots, x_n) = l$, but such that ∇h is never zero on the set $h(x_1, x_2, \ldots, x_n) = l$.

This technique allows us to deal with piecewise smooth curves and the analogous surfaces. In the examples we gave of $g(x, y) = (x - y)^3$, the set $g(x, y) = 0$ can also be described as the set $h(x, y) = 0$ where $h(x, y) = x - y$. The function h has the advantage over g that its gradient is never zero, so it always provides a nonzero vector in the normal direction.

3.5. **An application to a polygon.** Suppose we want to find the maximum and minimum of a differentiable function f of two variables on a triangle with three vertices (here, triangle refers only to the boundary, not the interior region). The triangle is piecewise linear, and *on the interior of any side, all normal vectors point in the same direction.* The points we need to test are as follows:

- The three vertices of the triangles
- The solutions to the Lagrange condition on each side: Note that on any fixed side, all normal vectors can be taken to be the same, so we just need to solve $\nabla f = \lambda$ times a constant vector, along with the equation constraining the point to be on the line for that side of the triangle and the inequality constraining it to be between the vertices.

After finding a hopefully finite list of points, we evaluate the function at each of them and determine the local extreme values.