Lagrange Multipliers without Permanent Scarring

Dan Klein

1 Introduction

This tutorial assumes that you want to know what Lagrange multipliers are, but are more interested in getting the intuitions and central ideas. It contains nothing which would qualify as a formal proof, but the key ideas need to read or reconstruct the relevant formal results are provided. If you don't understand Lagrange multipliers, that's fine. If you don't understand vector calculus at all, in particular gradients of functions and surface normal vectors, the majority of the tutorial is likely to be somewhat unpleasant. Understanding about vector spaces, spanned subspaces, and linear combinations is a bonus (a few sections will be somewhat mysterious if these concepts are unclear).

Lagrange multipliers are a mathematical tool for constrained optimization of differentiable functions. In the basic, unconstrained version, we have some (differentiable) function $f(x_1, \ldots x_n) : R^n \rightarrow R$ that we want to maximize (or minimize). We can do this by first find extreme points of f, which are points where the gradient ∇f is zero, or, equivlantly, each of the partial derivatives is zero. If we're lucky, points like this that we find will turn out to be (local) maxima, but they can also be minima or saddle points. We can tell the different cases apart by a variety of means, including checking properties of the second derivatives or simple inspecting the function values. Hopefully this is all familiar from calculus, though maybe it's more concretely clear when dealing with functions of just one variable.

All kinds of practical problems can crop up in unconstrained optimization, which we won't worry about here. One is that f and its derivative can be expensive to compute, causing people to worry about how many evaluations are needed to find a maximum. A second problem is that there can be (infinitely) many local maxima which are not global maxima, causing people to despair. We're going to ignore these issues, which are as big or bigger problems for the constrained case.

In constrained optimization, we have the same function f to maximize as before. However, we also have some restrictions on which points in R^n we are interested in. The points which satisfy our constraints are refered to as the *feasible* region. A simple constraint on the feasible region is to add boundaries, such as insisting that each x_i be positive. Boundaries complicate matters because extreme points on the boundaries will not, in general, meet the zero-derivative criterion, and so must be searched for in other ways. You probably had to deal with boundaries in calculus class. Boundaries correspond to inequality constraints, which we will say relatively little about in this tutorial.

Lagrange multipliers can help deal with both equality constraints and inequality constraints. For the majority of the tutorial, we will be concerned only with equality constraints, which restrict the feasible region to points lying on some surface inside R^n. Each constraint will be given by a function $g(x_1, \ldots x_n)$, and we will only be interested in points x where $g(x) = 0$.[1]

[1] If you want a $g(x) = c$ constraint, you can just move the c to the left: $g(x) - c = 0$.

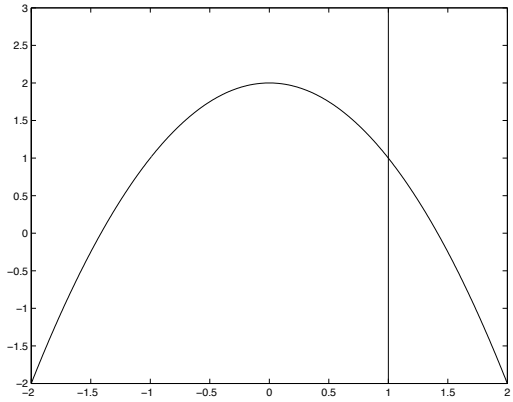

Figure 1: A one-dimensional domain... with a constraint. Maximize the value of $2 - x^2$ while satisfying $x - 1 = 0$.

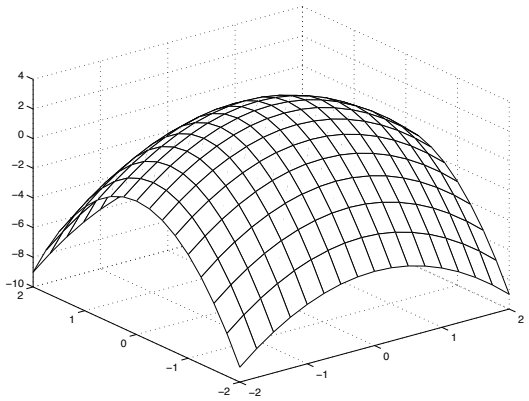

Figure 2: The paraboloid $2 - x^2 - 2y^2$.

2 Trial by Example

Let's do some example maximizations. First, we'll have an example of not using Lagrange multipliers.

2.1 A No-Brainer

Let's say you want to know the maximum value of $f(x) = 2 - x^2$ subject to the constraint $x - 1 = 0$ (see figure 1). Here we can just substitute our value for x (1) into f, and get our maximum value of $2 - 1^2 = 1$. It isn't the most challenging example, but we'll come back to it once the Lagrange multipliers show up. However, it highlights a basic way that we might go about dealing with constraints: substitution.

2.2 Substitution

Let $f(x, y) = 2 - x^2 - 2y^2$. This is the downward cupping paraboloid shown in figure 5. The unconstrained maximum is clearly at $x = y = 0$, while the unconstrained minimum is not even defined (you can find points with f as low as you like). Now let's say we constrain x and y to lie on the unit circle. To do this, we add the constraint $g(x, y) = x^2 + y^2 - 1 = 0$. Then, we maximize (or minimize) by first solving for one of the variables explicitly:

$$x^2 + y^2 - 1 = 0 \quad (1)$$

$$x^2 = 1 - y^2 \quad (2)$$

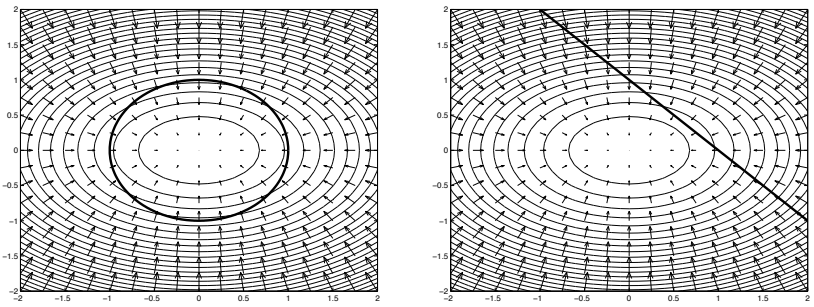

Figure 3: The paraboloid $2-x^2-2y^2$ along with two different constraints. Left is the unit circle $x^2+y^2=1$, right is the line $x+y=1$.

(3)

and substitute into f

$$\begin{aligned} f(x,y) &= 2-x^2+2y^2 && (4)\\ &= 2-(1-y^2)-2y^2 && (5)\\ &= 1-y^2 && (6) \end{aligned}$$

Then, we're back to a one-dimensional unconstrained problem, which has a maximum at $y=0$, where $x=\pm 1$ and $f=1$. This shouldn't be too surprising; we're stuck on a circle which trades x^2 for y^2 linearly, while y^2 costs twice as much from f.

Finding the constrained minimum here is slightly more complex, and highlights one weakness of this approach; the one-dimensional problem is still actually somewhat constrained in that y must be in $[-1,1]$. The minimum f value occurs at both these boundary points, where $x=0$ and $f=0$.

2.3 Inflating Balloons

The main problem with substitution is that, despite our stunning success in the last section, it's usually very hard to do. Rather than inventing a new problem and discovering this the hard way, let's stick with the f from the last section and consider how the Lagrange multiplier method would work. Figure 3(left) shows a contour plot of f. The contours, or level curves, are ellipses, which are wide in the x dimension, and which represent points which have the same value of f. The dark circle in the middle is the feasible region satisfying the constraint $g=0$. The arrows point in the directions of greatest increase of f. Note that the direction of greatest increase is always perpendicular to the level curves.

Imagine the ellipses as snapshots of an inflating and balloon. As the ballon expands, the value of f along the ellipse decreases. The size-zero ellipse has the highest value of f. Consider what happens as the ellipse expands. At first, the values of f are high, but the ellipse does not intersect the feasible circle anywhere. When the long axis of the ellipse finally touches the circle at $(\pm 1,0)$, $f=1$ as in figure 4(left). This is the maximum constrained value for f – any larger, and no point on the level curve will be in the feasible circle. The key thing is that, at $f=1$, the ellipse is tangent to the circle.[2]

The ellipse then continues to grow, f dropping, intersecting the circle at four points, until the ellipse surrounds the circle and only the short axis endpoints are still touching. This is the minimum ($f=0$, $x=0$, $y=\pm 1$). Again, the two curves are tangent. Beyond this value, the level curves do not intersect the circle.

The curves being tangent at the minimum and maximum should make intuitive sense. If the two curves were not tangent, imagine a point (call it p) where they touch. Since the curves aren't tangent, then the curves will cross, meeting at p, as in figreffig:crossing(right). Since the f contour (light curve) is a level curve, the points to one side of the contour have greater f value, while the points on the other side have lower f value. Since we may move anywhere along g and still satisfy the constraint, we can nudge p along g to either side of the contour to either increase or decrease f. So p cannot be an extreme point.

[2]Differentiable curves which touch but do not cross are tangent, but feel free to verify it by checking derivatives!

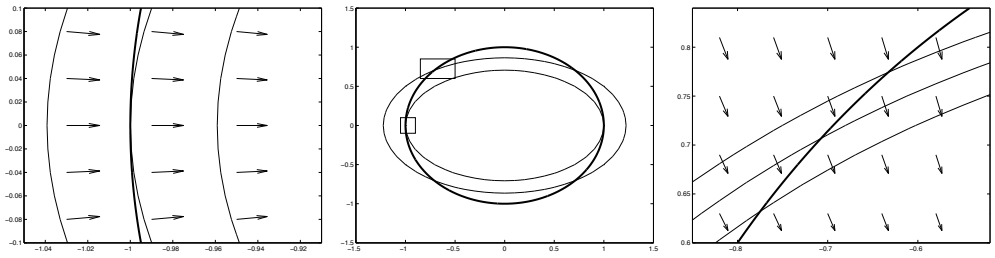

Figure 4: Level curves of the paraboloid, intersecting the constraint circle.

This intuition is very important; the entire enterprise of Lagrange multipliers (which are coming soon, really!) rests on it. So here's another, equivalent, way of looking at the tangent requirement, which generalizes better. Consider again the zooms in figure 4. Now think about the normal vectors of the contour and constraint curves. The two curves being tangent at a point is equivalent to the normal vectors being parallel at that point. The contour is a level curve, and so the gradient of f, ∇f, is normal to it. But that means that at an extreme point p, the gradient of f will be perpendicular to g as well. This should also make sense – the gradient is the direction of steepest ascent. At a solution p, we must be on g, and, while it is fine for f to have a non-zero gradient, the direction of steepest ascent had better be perpendicular to g. Otherwise, we can project ∇f onto g, get a non-zero direction along g, and nudge p along that direction, increasing f but staying on g. If the direction of steepest increase and decrease take you off perpendicularly off of g, then, even if you are not at an unconstrained maximum of f, there is no local move you can make to increase f which does not take you out of the feasible region g.

Formally, we can write our claim that the normal vectors are parallel at an extreme point p as:

$$\nabla f(p) = \lambda \nabla g(p) \tag{7}$$

So, our method for finding extreme points[3] which satisfy the constraints is to look for point where the following equations hold true:

$$\nabla f(x) = \lambda \nabla g(x) \tag{8}$$
$$g(x) = 0 \tag{9}$$

We can compactly represent both equations at once by writing the *Lagrangian*:

$$\Lambda(x, \lambda) = f(x) - \lambda g(x) \tag{10}$$

and asking for points where

$$\nabla \Lambda(x, \lambda) = 0 \tag{11}$$

The partial derivatives with respect to x recover the parallel-normals equations, while the partial derivative with respect to λ recovers the constraint $g(x) = 0$. The λ is our first Lagrange multiplier.

Let's re-solve the circle-paraboloid problem from above using this method. It was so easy to solve with substition that the Lagrange multiplier method isn't any easier (if fact it's harder), but at least it illustrates the method. The Lagrangian is:

$$\Lambda(x, \lambda) = f(x) - \lambda g(x) \tag{12}$$
$$= 2 - x_1^2 - 2x_2^- \lambda(x_1^2 + x_2^2 - 1) \tag{13}$$

and we want

$$\nabla \Lambda(x, \lambda) = \nabla f(x) - \lambda \nabla g(x) = 0 \tag{14}$$
$$\tag{15}$$

[3] We can sort out after we find them which are minima, maxima, or neither.

which gives the equations

$$\frac{\partial}{\partial x_1}\Lambda(x,\lambda) = -2x_1 - 2\lambda x_1 = 0 \quad (16)$$

$$\frac{\partial}{\partial x_2}\Lambda(x,\lambda) = -4x_2 - 2\lambda x_2 = 0 \quad (17)$$

$$\frac{\partial}{\partial \lambda}\Lambda(x,\lambda) = x_1^2 + x_2^2 - 1 = 0 \quad (18)$$

$$(19)$$

From the first two equations, we must have either $\lambda = -1$ or $\lambda = -2$. If $\lambda = -1$, then $x_2 = 0, x_1 = \pm 1$, and $f = 1$. If $\lambda = -2$, then $x_2 = \pm 1, x_1 = 0$, and $f = 0$. These are the minimum and maximum, respectively.

Let's say we instead want the constraint that x and y sum to 1 ($x + y - 1 = 0$). Then, we have the situation in figure **??**(right). Before we do anything numeric, convince yourself from the picture that the maximum is going to occur in the (+,+) quadrant, at a point where the line is tangent to a level curve of f. Also convince yourself that the minimum will not be defined; that f values get arbitrarily low in both directions along the line away from the maximum. Formally, we have

$$\Lambda(x,\lambda) = f(x) - \lambda g(x) \quad (20)$$

$$= 2 - x_1^2 - 2x_2^- \lambda(x_1 + x_2 - 1) \quad (21)$$

and we want

$$\nabla\Lambda(x,\lambda) = \nabla f(x) - \lambda \nabla g(x) = 0 \quad (22)$$

$$(23)$$

which gives

$$\frac{\partial}{\partial x_1}\Lambda(x,\lambda) = -2x_1 - \lambda = 0 \quad (24)$$

$$\frac{\partial}{\partial x_2}\Lambda(x,\lambda) = -4x_2 - \lambda = 0 \quad (25)$$

$$\frac{\partial}{\partial \lambda}\Lambda(x,\lambda) = x_1 + x_2 - 1 = 0 \quad (26)$$

$$(27)$$

We can see from the first two equations that $x_1 = 2x_2$, which, with, since they sum to one, means $x_1 = 2/3$, $x_2 = 1/3$. At those values, $f = 4/3$ and $\lambda = -4/3$.

So what do we have so far? Given a function and a constraint, we can write the Lagrangian, differentiate, and solve for zero. Actually solving that system of equations can be hard, but note that the Lagrangian is a function of n+1 variables (n x_i plus λ) and so we do have the right number of equations to hope for unique, existing solutions: n from the x_i partial derivatives, plus one from the λ partial derivative.

2.4 More Dimensions

If we want to have mutliple constraints, this method still works perfectly well, though it get harder to draw the pictures to illustrate it. To generalize, let's think of the parallel-normal idea in a slightly different way. In unconstrained optimization (no constraints), we knew we were at a local extreme because the gradient of f was zero – there was no local direction of motion which increased f. Along came the constraint g and dashed all hopes of the gradient being completely zero at a constrained extreme p, because we were confined to g. However, we still wanted that there be no direction of increase *inside* the feasible region. This occured whenever the gradient at p, while probably not zero, had no components which were perpendicular to the normal of g at p. To recap: in the presence of a constraint, $\nabla f(p)$ does not have to be zero at a solution p, it just has to be entirely contained in the (one-dimensional) subspace spanned by $\nabla g(p)$.

The last statement generalizes to multiple constraints. With multiple constraints $g_i(x) = 0$, we will insist that a solution p satisfy each $g_i(p) = 0$. We will also want the gradient $\nabla f(p)$ to be non-zero along the

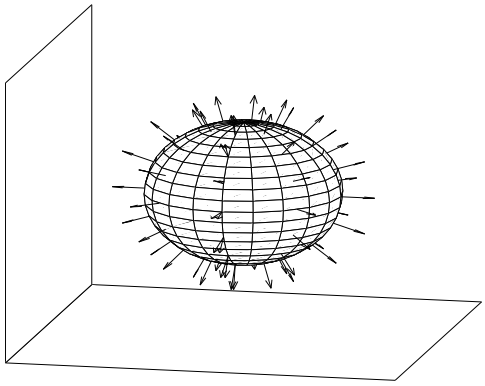

Figure 5: A spherical level curve of the function $f(x) = |x|$ with two constraint planes, $y = -1$ and $z = -1$.

directions that p is free to vary. However, given the constraints, p cannot make any local movement along vectors which have any component perpendicular to any constraint. Therefore, our condition should again be that $\nabla f(p)$, while not necessarily zero, is entirely contained in the subspace spanned by the $\nabla g_i(p)$ normals. We can express this by the equation

$$\nabla f(x) \quad = \quad \sum_i \lambda_i \nabla g_i(x) \tag{28}$$

Which asserts that $\nabla f(p)$ be a linear combination of the normals, with weights λ_i.

It turns out that tossing all the constraints into a single Lagrangian accomplishes this:

$$\Lambda(x, \lambda) \quad = \quad f(x) - \sum_i \lambda_i g_i(x) \tag{29}$$

It should be clear that differentiating $\Lambda(x, \lambda)$ with respect to λ_i and setting equal to zero recovers the ith constraint, $g_i(x) = 0$, while differentiating with respect to the x_i recovers the assertion that the gradient of f have no components which aren't spanned by the constraints normals.

As an example of multiple constraints, consider figure **??**. Imagine that f is the distance from the origin. Thus, the level surfaces of f are concentric spheres with the gradient pointing straight out of the spheres. Let's say we want the minimum of f subject to the constraints that $y = -1$ and $z = -1$, shown as planes in the figure. Again imagine the spheres as expanding from the center, until it makes contact with the planes. The unconstrained minimum is, of course, at the origin, where ∇f is zero. The sphere grows, and f increases. When the sphere's radius reaches one, the sphere touches both planes individually. At the points of contact, the gradient of f is perpendicular to the touching plane. Those points would be solutions if that plane were the only constraint. When the sphere reaches a radius of $\sqrt{2}$, it is touching both planes along their line of intersection. Note that the gradient is *not* zero at that point, *nor* is it perpendicular to either surface. However, it is parallel to an (equal) combination of the two planes' normal vectors, or, equivalently, it lies inside the plane spanned by those vectors (the plane $x = 0$, [not shown due to my lacking matlab skills]).

A good way to think about the effect of adding constraints is as follows. Before there are any constraints, there are n dimensions for x to vary along when maximizing, and we want to find points where all n dimensions have zero gradient. Every time we add a constraint, we restrict one dimension, so we have less freedom in maximizing. However, that constraint also removes a dimension along which the gradient must be zero. So, in the "nice" case, we should be able to add as many or few constraints (up to n) as we wish, and everything should work out.[4]

[4]In the "not-nice" cases, all sorts of things can go wrong. Constraints may be unsatisfiable (e.g. $x = 0$ and $x = 1$, or subtler situations can prevent the Lagrange multipliers from existing [more].

3 The Lagrangian

The Lagrangian $\Lambda(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x)$ is a function of $n + m$ variables (remember that $x \in R^n$, plus one for each of the m $\lambda_i \in \lambda$). Differentiating gives the corresponding $n + m$ equations, each set to zero, to solve. The n equations from differentiating with respect to each x_i recovers our gradient conditions. The m equations from differentiating with respect the λ_i recover the constraints g_i. So the numbers give us some confidence that we have the right number of equations to hope for point solutions.

It's helpful to have an idea of what the Lagrangian actually means. There are two intuitions, described below.

3.1 The Lagrangian as an Encoding

First, we can look at the Lagrangian as an encoding of the problem. This view is easy to understand (but doesn't really get us anywhere). Whenever the constraints are satisfied, the g_i are zero, and so at these point, regarless of the value of the λ_i multipliers, $\Lambda(x, \lambda) = f(x)$. This is a good fact to keep in mind.

You could imagine using the Lagrangian to do constrained maximization in the following way. You move x around R^n looking for a maximum value of f. However, you have no control over λ, which gets set in the worst way possible for you. Therefore, when you choose x, $lambda$ is chosen to minimize Λ. Formally, the problem is to find the x which gives

$$f^* = \max_x(\min_\lambda \Lambda(x, \lambda)) \tag{30}$$

Now remember that if your x happens to satisfy the constraints, $\Lambda(x, \lambda) = f(x)$, regardless of what λ is. However, if x does not satisfy the constraints, some $g_i(x) \neq 0$. But then, λ_i can be fiddled to make $\Lambda(x, \lambda))$ as small as desired, and $\min_\lambda \Lambda(x, \lambda)$ will be $-\infty$. So f^* will be the maximum value of f subject to the constraints.

3.2 Reversing the Scope

The problem with the above view of the Lagrangian is that it really doesn't accomplish anything beyond encoding the constraints and handing us back the same problem we started with: find the maximum value of f, ignoring the values of x which are not in the feasible region. More usefully, we can switch the $\min$ and $\max$ from the previous section, and the result still holds:

$$f^* = \min_\lambda(\max_x \Lambda(x, \lambda)) \tag{31}$$

This is part of the full Kuhn-Tucker theorem (cite), which we aren't going to prove rigorously. However, the intuition behind why it's true is important. Before we examine why this reversal should work, let's see what it accomplishes if it's true.

We originally had a constrained optimization problem. We would very much like for it to become an unconstrained optimization problem. Once we fix the values of the λ_i multipliers, $\Lambda(x, \lambda)$ becomes a function of x alone. We might be able to maximize that function (it's unconstrained!) relatively easily. If so, we would get a solution for each λ, call it $x^*(\lambda)$. But then we can do an unconstrained minimization of $x^*(\lambda)$ over the space of λ. We would then have our solution.

It might not be clear why that's any different that fixing x and finding a minimizing value $\lambda^*(x)$ for each x. It's different in two ways. First, unlike $x^*(\lambda)$, $\lambda^*(x)$ would not be continuous. (Remember that it's negative infinity almost everywhere and jumps to $f(x)$ for x which satisfy the constraints.) Second, it is often the case that we can find a closed-form solution to $x^*(\lambda)$ while we have nothing useful to say about $\lambda^*(x)$. This is also a general instance of switching to a dual problem when a primal problem is unpleasant in some way. [cites]

3.3 Duality

Let's say we're convinced that it would be a good thing if

$$\max_x(\min_\lambda \Lambda(x, \lambda)) = \min_\lambda(\max_x \Lambda(x, \lambda)) \tag{32}$$

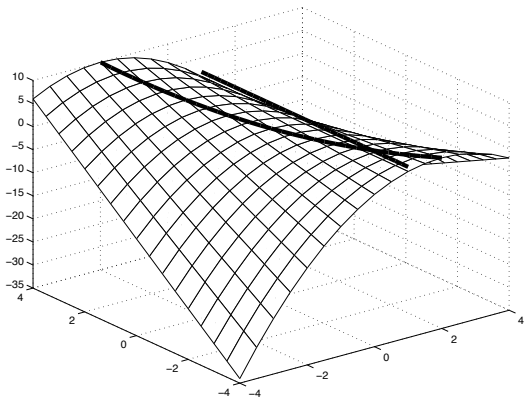

Figure 6: The Lagrangian of the paraboloid $2 - x^2$ with the constraint $x - 1 = 0$.

Now, we'll argue why this is true, examples first, intuition second, formal proof elsewhere. Recall the no-brainer problem: maximize $f(x) = 2 - x^2$ subject to $x - 1 = 0$. Let's form the Lagrangian.

$$\Lambda(x, \lambda) \quad = \quad 2 - x^2 - \lambda(x - 1) \tag{33}$$

This surface is plotted in figure **??**. The straight dark line is the value of Λ at $x = 1$. At that value, the constraint is satisfied and so, as promised, λ has no effect on Λ and $Lambda(x, \lambda) = f(x) = 1$. At each x, $\lambda^*(x)$ be $-\infty$, except for $x = 1$ where $\lambda^*(x) = 1$. The curving dark line is the value of Λ at $x = x^*(\lambda)$ for all λ. The minimum Λ value along this $x^*(\lambda)$ line is at $x = 1, \lambda = 2$, where $f = 1$, which is the maximum (and only) value of $f(x)$ among the point(s) satisfying the constraint.

3.4 A sort-of proof of sort-of the Kuhn-Tucker thereom

The Lagrangian is hard to plot when $n > 1$. However, let's consider what happens in the environment of a point p which satisfies the constraints, and which is a local maximum among the points satisfying the constraints. Since each $g_i(p) = 0$, the derivatives of $\Lambda(x, \lambda)$ with respect to each λ_i are zero. $\nabla f(p)$ may not be zero. But if $\nabla f(p)$ has any component which is not in the space spanned by the constraint normals $\nabla g_i(p)$, then we can nudge p in a direction inside the allowed region, increasing $f(p)$. Since p is a local minimum inside that region, that isn't possible. So $\nabla f(p)$ is in the space spanned by the constraint normals $\nabla g_i(p)$, and can therefore be written as a (unique) linear combination of these normals. Let $\nabla f(p) = \sum_i \lambda_i \nabla g_i(p)$ be that combination. Then clearly $\nabla f(p) - \sum_i \lambda_i \nabla g_i(p) = 0$.

Now consider a vector λ' near λ. $\nabla f(p) - \sum_i \lambda'_i \nabla g_i(p)$ cannot still be zero, because the linear combination weights λ are unique. But $\nabla \Lambda(p, \lambda') = \nabla f(p) - \sum_i \lambda'_i \nabla g_i(p)$ is non-zero. Thus, fixing λ' and allowing p to vary, there is some direction (either $\nabla \Lambda(p, \lambda')$ or the reverse direction where we could nudge p to increase Λ. Therefore, at $\lambda(p)$, $x^*(\lambda)$ is at a local minimum.

Another way to remember this intuitively is that λ is probably not zero, and, if we set it to zero (a huge nudge), $\Lambda(x, 0) = f(x)$, and so the maximum of Λ is the unconstrained maximum of f, which can only be larger than $f(p)$.

Let's look another more example. Recall the paraboloid (figure 5) with the constraint that x and y sum to one. The maximum value occured at $(x, y) = (2/3, 1/3)$, where $f = 4/3$. The λ value was $-4/3$. Figure 7 shows what happens when we nudge λ up and down slightly. At $\lambda = 0$, the Lagrangian Λ is just the original surface f. Its maximum value (2) is at the origin (which obviously doesn't satisfy the constraint). At $\lambda = -4/3$, the maximum value of the Lagrangian is at $p = (2/3, 1/3)$, (which does satisfy the constraints). The gradient of f is not zero, but it is perpendicular to the constraint line, so p is a local maximum along that line. Another way of thinking of this is that the gradient of f (the top arrow field) is balanced at that point by the scaled gradient of the constraint (the second arrow field down). We can see the effect by adding these two fields, which forms the gradient of the Lagrangian (third arrow field). This gradient is zero at p with the right λ. If we nudge λ up to $-4/3 + 0.1$, then suddenly the gradient of f is no longer completely cancelled out by $\lambda \nabla g$, and so we can

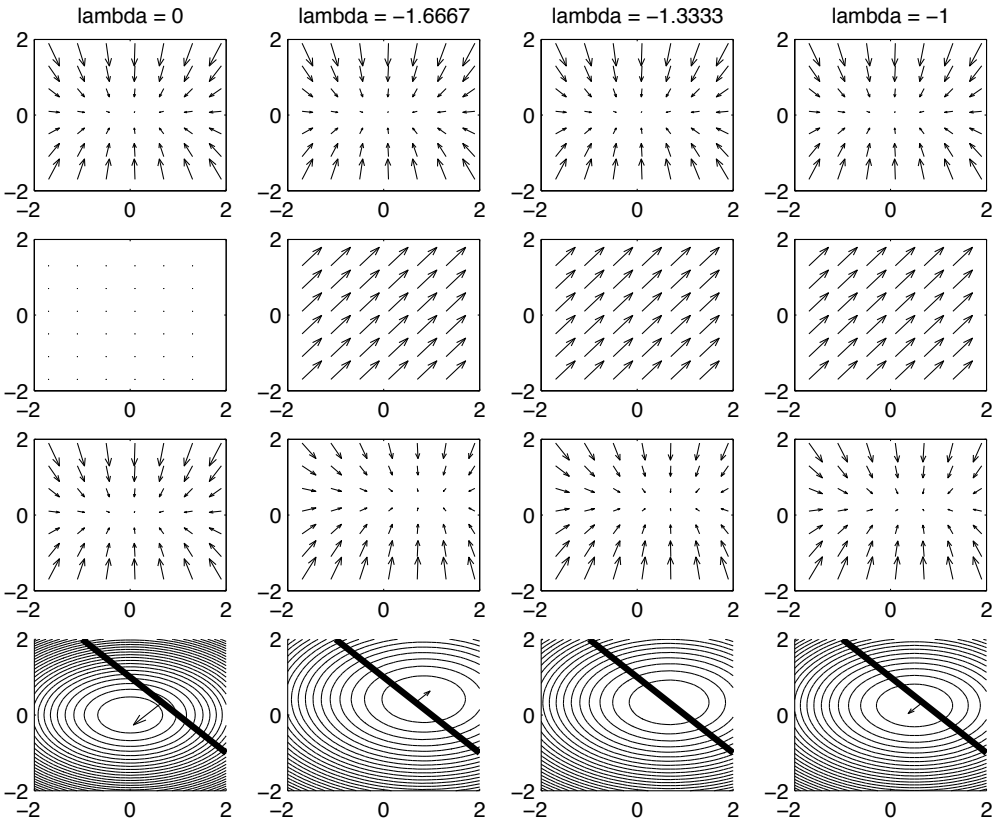

Figure 7: Lagrangian surfaces for the paraboloid $2 - x^2 - 2y^2$ with the constraint $x + y - 1 = 0$.

increase the lagrangian by nudging p toward the origin. Similarly, if we nudge λ down to $-4/3 - 0.1$, then the gradient of f is over-cancelled and we can increase the Lagrangian by nudging p away from the origin.

3.5 What do the multipliers mean?

A useful aspect of the Lagrange multiplier method is that the values of the multipliers at solution points often has some significance. Mathematically, a multiplier λ_i is the value of the partial derivative of Λ with respect to the constraint g_i. So it is the rate at which we could increase the Lagrangian if we were to raise the target of that constraint (from zero). But remember that at solution points p, $\Lambda(p, \lambda) = f(p)$. Therefore, the rate of increase of the Lagrangian with respect to that constraint is also the rate of increase of the maximum constrained value of f with respect to that constraint.

In economics, when f is a profit function and the g_i are constraints on resource amounts, λ_i would be the amount (possibly negative!) by which profit would rise if one were allowed one more unit of resource i. This rate is called the *shadow price* of i, which is interpreted as the amount it would be worth to relax that constraint upwards (by R&D, mining, bribery, or whatever means).

[Physics example?]

4 A bigger example than you probably wanted

This section contains a big example of using the Lagrange multiplier method in practice, as well as another case where the multipliers have an interesting interpretation.

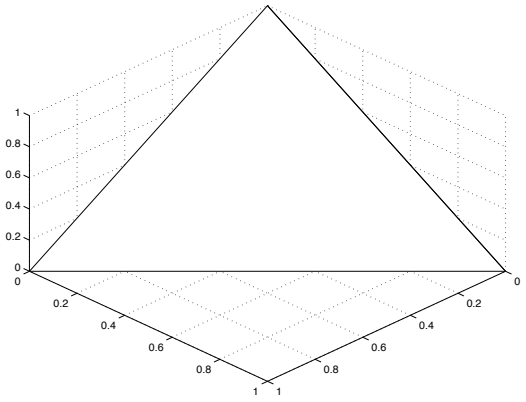

Figure 8: The simplex $x + y + z = 1$.

4.1 Maximum Entropy Models

5 Extensions

5.1 Inequality Constraints

The Lagrange multiplier method also covers the case of inequality constraints. Constraints of this form are written $h(x) \geq 0$. The key observation about inequality constraints work is that, at any given x, a $h(x)$ has either $h(x) = 0$ or $h(x) > 0$, which are qualitatively very different. The two possibilities are shown in figure **??**. If $h(x) = 0$ then h is said to be *active* at x, otherwise it is *inactive*. If h is active at x, then h is a lot like an equality constraint; it allows x to be maximum if the gradient of f, $\nabla f(x)$, is either zero or pointing towards negative values of h (which violate the constraint). However, if the gradient is pointing towards positive values of h, then there is no reason that we cannot move in that direction. Recall that we used to write

$$\nabla f(x) = \lambda \nabla g(x) \tag{34}$$

for a (single) equality constraint. The interpretation was that, if x is a solution, $\nabla f(x)$ must be entirely in the direction of the normal to $g(x)$, $\nabla g(x)$. For inequality constraints, we write

$$\nabla f(x) = \mu \nabla h(x) \tag{35}$$

but, if x is a maximum, then if $\nabla f(x)$ is non-zero, it not only has to be parallel $\nabla g(x)$, but it must actually point in the opposite sense along that direction (i.e., out of the feasible side and towards the forbidden side). We can actually enforce this very simply, by restricting the multiplier to be negative (or zero). Positive mutlipliers mean that the direction of increasing f is in the same direction as increasing $h(x)$ – but points in that situation certainly aren't solutions, as we want to increase f and we are allowed to increase h.

If h is inactive at x ($h(x) > 0$), then we want to be even stricter about what values of μ are acceptable from a solution. In fact, in this case, μ must be zero at x. (Intuitively, if h is inactive, then nothing should change at x if we drop h). [better explanation]

In summary, for inequality constraints, we add them to the Lagrangian just as if they were equality constraints, except that we require that $\mu \leq 0$ and that, if $h(x)$ is not zero, then μ is. The situation that one or the other can be non-zero, but not both, is referred to as *complementary slackness*. This situation can be compactly written as $\mu h(x) = 0$. Bundling it all up, complete with multiple constraints, we get the general Lagrangian:

$$\Lambda(x, \lambda, \mu) = f(x) - \sum_i \lambda_i g_i(x) - \sum_j \mu_j h_j(x) \tag{36}$$

The Kuhn-Tucker theorem (or our intuitive arguments) tell us that if a point x is a maximum of f subject to the constraints g_i and h_i, then:

$$\nabla \Lambda(x, \lambda, \mu) \quad = \quad \nabla f(x) - \sum_i \lambda_i \nabla g_i(x) - \sum_j \mu_j \nabla h_j(x) = 0 \tag{37}$$

$$\forall i, \mu_i \leq 0 \tag{38}$$

$$\sum_j \mu_j h_j(x) = 0 \tag{39}$$

The second condition takes care of the restriction on active inequalities. The third condition is a somewhat cryptic way of insisting that for each i, either μ_i is zero or $h_i(x)$ is zero.

Now is probably a good time to point out that there is more to the Kuhn-Tucker theorem than the above statement. The above conditions are called the *first-order* conditions. All (local) maxima will satisfy them. The theorem also gives *second order* conditions on the second derivative (Hessian) matrices which distinguish local maxima from other situations which can trigger "false alarms" with the first-order conditions. However, in many situations, one knows in advance that the solution will be a maximum (such as in the maximum entropy example).

Caveat about globals?

6 Conclusion

This tutorial only introduces the basic concepts of the Langrange multiplier methods. If you are interested, there are many detailed texts on the subject [cites]. The goal of this tutorial was to supply some intuition behind the central ideas so that other, more comprehensive and formal sources become more accessible.

Feedback requested!