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 $f(x_1, \ldots x_n) : R^n \to R$ 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: $g(x) - c = 0$.

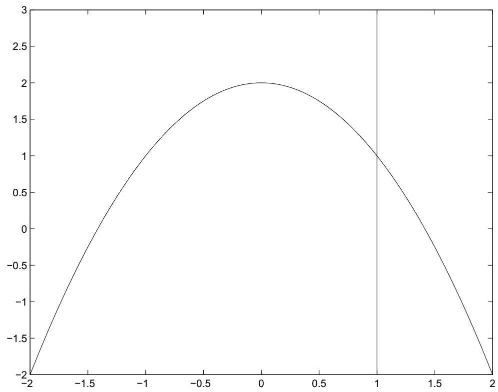


Figure 1: A one-dimensional domain... with a constraint. Maximize the value of $2 - x^2$ while satisfying $x - 1 = 0$.

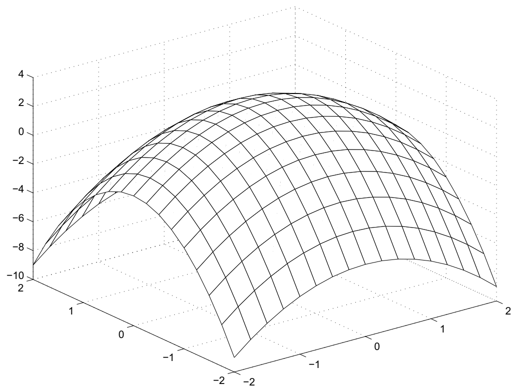


Figure 2: The paraboloid $2 - x^2 - 2y^2$.

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:

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

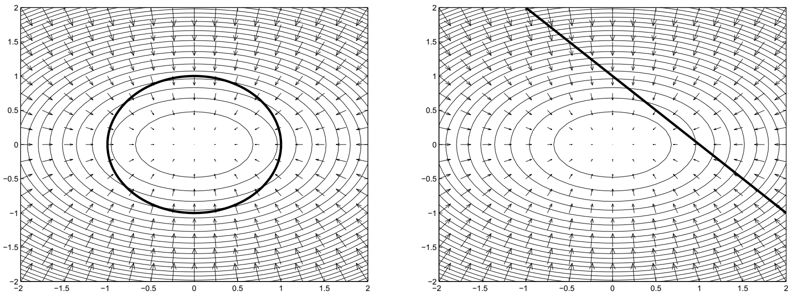


Figure 3: The paraboloid $2 - x^2 - 2y^2$ along with two different constraints. Left is the unit circle $x^2 + y^2 = 1$, right is the line $x + y = 1$.

$$\tag{3}$$

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!

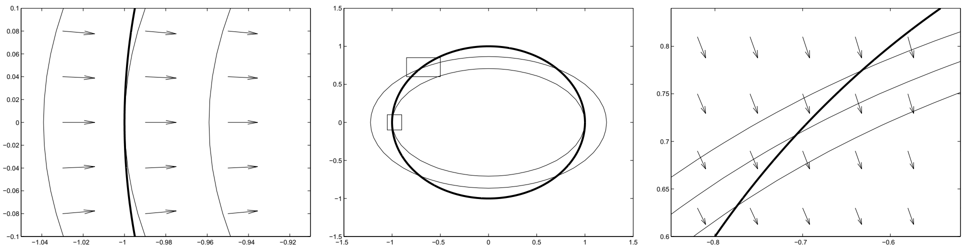


Figure 4: Level curves of the paraboloid, intersecting the constraint circle.

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:

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

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

$$\begin{aligned} \nabla f(x) &= \lambda \nabla g(x) && (8) \\ g(x) &= 0 && (9) \end{aligned}$$

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

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

and asking for points where

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

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:

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

and we want

$$\begin{aligned} \nabla \Lambda(x, \lambda) &= \nabla f(x) - \lambda \nabla g(x) = 0 && (14) \\ & && (15) \end{aligned}$$

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

which gives the equations

$$
\begin{aligned}
\frac{\partial}{\partial x_1}\Lambda(x,\lambda) &= -2x_1 - 2\lambda x_1 = 0 && (16)\\
\frac{\partial}{\partial x_2}\Lambda(x,\lambda) &= -4x_2 - 2\lambda x_2 = 0 && (17)\\
\frac{\partial}{\partial \lambda}\Lambda(x,\lambda) &= x_1^2 + x_2^2 - 1 = 0 && (18)\\
& && (19)
\end{aligned}
$$

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

$$
\begin{aligned}
\Lambda(x,\lambda) &= f(x) - \lambda g(x) && (20)\\
&= 2 - x_1^2 - 2x_2^{\bar{}}\lambda(x_1 + x_2 - 1) && (21)
\end{aligned}
$$

and we want

$$
\begin{aligned}
\nabla\Lambda(x,\lambda) &= \nabla f(x) - \lambda \nabla g(x) = 0 && (22)\\
& && (23)
\end{aligned}
$$

which gives

$$
\begin{aligned}
\frac{\partial}{\partial x_1}\Lambda(x,\lambda) &= -2x_1 - \lambda = 0 && (24)\\
\frac{\partial}{\partial x_2}\Lambda(x,\lambda) &= -4x_2 - \lambda = 0 && (25)\\
\frac{\partial}{\partial \lambda}\Lambda(x,\lambda) &= x_1 + x_2 - 1 = 0 && (26)\\
& && (27)
\end{aligned}
$$

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

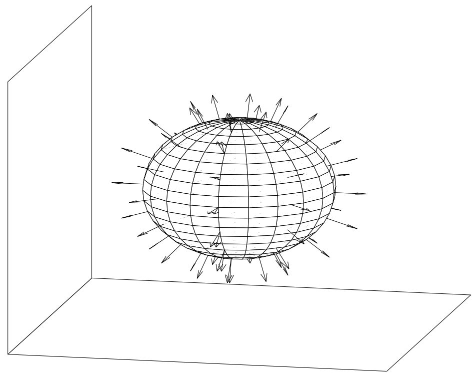


Figure 5: A spherical level curve of the function $f(x) = |x|$ with two constraint planes, $y = -1$ and $z = -1$.

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:

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

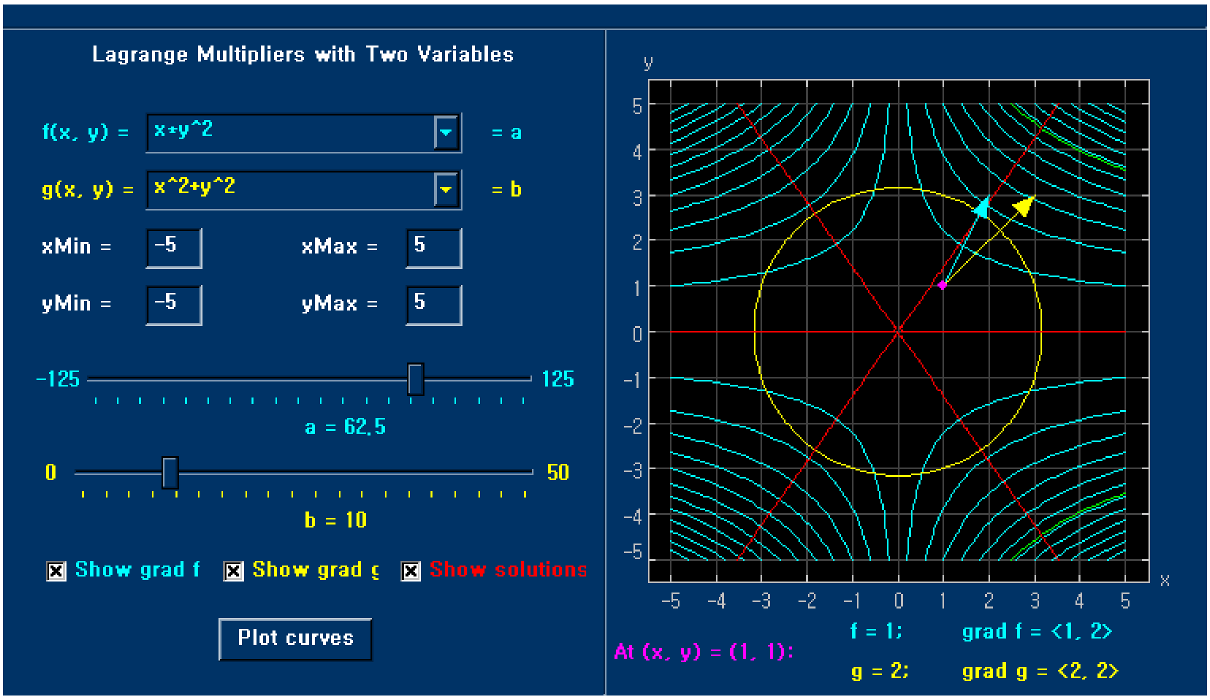
It should be clear that differentiating $\Lambda(x, \lambda)$ with respect to λ_i and setting equal to zero recovers the ith constraint, $g_i(x) = 0$, while differentiating with respect to the x_i recovers the assertion that the gradient of f have no components which aren't spanned by the constraints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

Lagrange Multipliers (Two Variables)



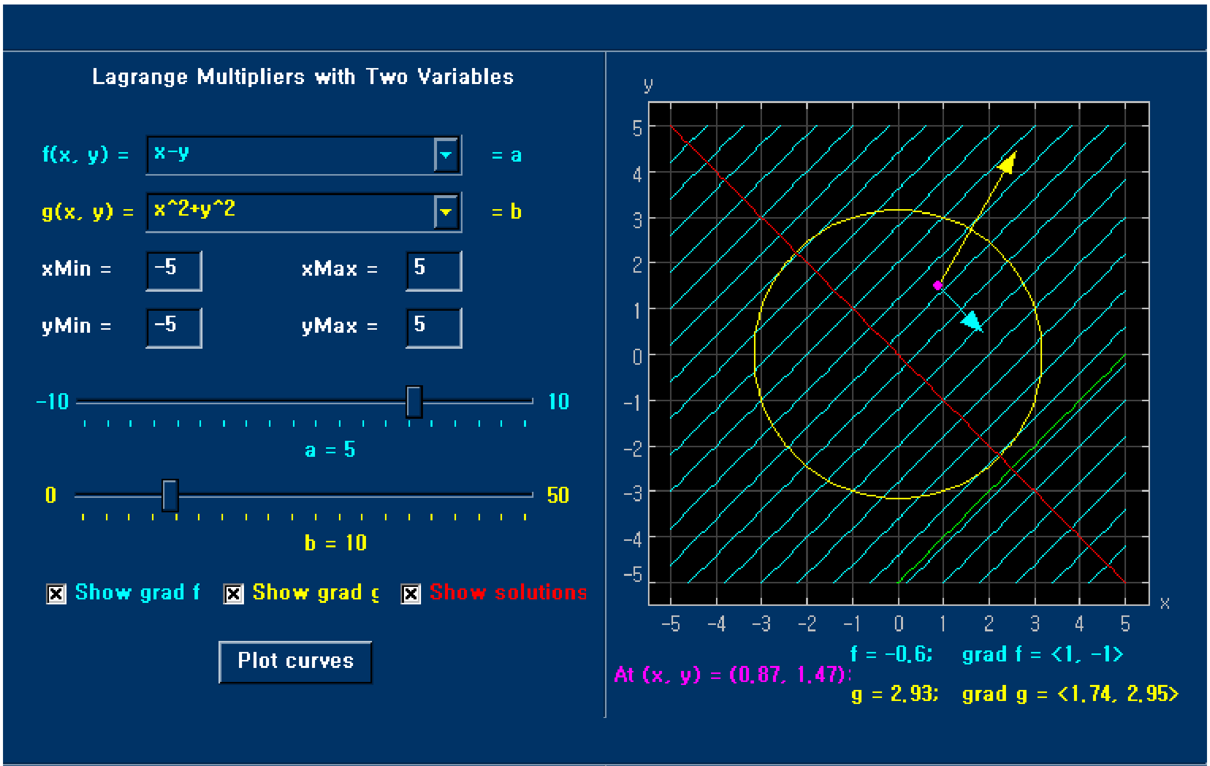
Directions:

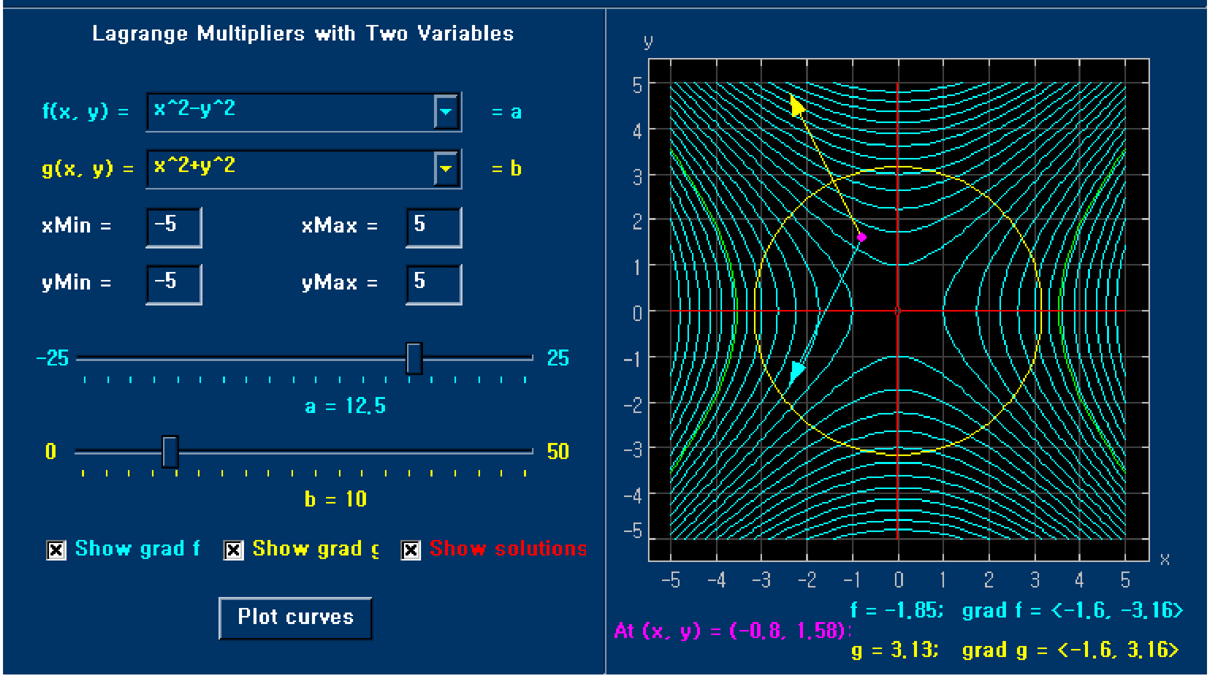
Setup. Enter the function to minimize / maximize, f(x,y), into the box in the upper-left corner. Enter the constraint, g(x,y), into the box immediately below. Click on the "Plot curves" button in the lower-left corner to update the display. Then, use the yellow slider control to set the value of b in the constraint equation g(x,y)=b.

The applet shows a contour plot of f (in blue), together with the level curve g(x,y)=b corresponding to the constraint equation (in yellow). You can use the blue slider control to move a highlighted level curve of f. The minima and maxima of f subject to the constraint correspond to the points where this level curve becomes tangent to the yellow curve g(x,y)=b.

Click in the contour plot to move the pink dot and display the gradient vectors of f and g at the given point. The components of grad(f) and grad(g) are displayed in the lower-right corner. As expected, the two gradient vectors are proportional to each other at a constrained minimum/maximum.

The red "Show solutions" button displays a red curve consisting of all points where grad(f) and grad(g) are proportional to each other. The Lagrange multiplier method tells us that constrained minima/maxima occur when this proportionality condition and the constraint equation are both satisfied: this corresponds to the points where the red and yellow curves intersect.





http://www-math.mit.edu/18.02/applets/LagrangeMultipliersTwoVariables.html