

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or to view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

Last time we saw things about gradients and directional derivatives. Before that we studied how to look for minima and maxima of functions of several variables.

And today we are going to look again at min/max problems but in a different setting, namely, one for variables that are not independent. And so what we will see is you may have heard of Lagrange multipliers.

And this is the one point in the term when I can shine with my French accent and say Lagrange's name properly.

OK. What are Lagrange multipliers about? Well, the goal is to minimize or maximize a function of several variables.

Let's say, for example, f of x, y, z , but where these variables are no longer independent.

They are not independent. That means that there is a relation between them. The relation is maybe some equation of the form g of x, y, z equals some constant.

You take the relation between x, y, z , you call that g and that gives you the constraint. And your goal is to minimize f only of those values of x, y, z that satisfy the constraint. What is one way to do that?

Well, one to do that, if the constraint is very simple, we can maybe solve for one of the variables.

Maybe we can solve this equation for one of the variables, plug it back into f , and then we have a usual min/max problem that we have seen how to do.

The problem is sometimes you cannot actually solve for x, y, z in here because this condition is too complicated and then we need a new method. That is what we are going to do.

Why would we care about that? Well, one example is actually in physics. Maybe you have seen in thermodynamics that you study quantities about gases, and those quantities that involve pressure, volume and temperature. And pressure, volume and temperature are not independent of each other.

I mean you know probably the equation $PV = NRT$.

And, of course, there you could actually solve to express things in terms of one or the other.

But sometimes it is more convenient to keep all three variables but treat them as constrained.

It is just an example of a situation where you might want to do this. Anyway, we will look mostly at particular examples, but just to point out that this is useful when you study guesses in physics.

The first observation is we cannot use our usual method of looking for critical points of f .

Because critical points of f typically will not satisfy this condition and so won't be good solutions.

We need something else. Let's look at an example, and we will see how that leads us to the method.

For example, let's say that I want to find the point closest to the origin -- -- on the hyperbola xy equals 3 in the plane. That means I have this hyperbola, and I am asking myself what is the point on it that is the closest to the origin?

I mean we can solve this by elementary geometry, we don't need actually Lagrange multipliers, but we are going to do it with Lagrange multipliers because it is a pretty good example. What does it mean?

Well, it means that we want to minimize distance to the origin.

What is the distance to the origin?

If I have a point, at coordinates (x, y) and then the distance to the origin is square root of x squared plus y squared. Well, do we really want to minimize that or can we minimize something easier?

Yeah. Maybe we can minimize the square of a distance. Let's forget this guy and instead -- Actually, we will minimize f of x, y equals x squared plus y squared, that looks better, subject to the constraint $xy = 3$. And so we will call this thing g of x, y to illustrate the general method.

Let's look at a picture. Here you can see in yellow the hyperbola xy equals three. And we are going to look for the points that are the closest to the origin.

What can we do? Well, for example, we can plot the function x squared plus y squared, function f . That is the contour plot of f with a hyperbola on top of it. Now let's see what we can do with that. Well, let's ask ourselves, for example, if I look at points where f equals 20 now. I think I am at 20 but you cannot really see it. That is a circle with a point whose distant square is 20. Well, can I find a solution if I am on the hyperbola? Yes, there are four points at this distance. Can I do better?

Well, let's decrease for distance.

Yes, we can still find points on the hyperbola and so on.

Except if we go too low then there are no points on this circle anymore in the hyperbola. If we decrease the value of f that we want to look at that will somehow limit value beyond which we cannot go, and that is the minimum of f .

We are trying to look for the smallest value of f that will actually be realized on the hyperbola.

When does that happen? Well, I have to backtrack a little bit. It seems like the limiting case is basically here. It is when the circle is tangent to the hyperbola. That is the smallest circle that will hit the hyperbola. If I take a larger value of f , I will have solutions. If I take a smaller value of f , I will not have any solutions anymore.

So, that is the situation that we want to solve for.

How do we find that minimum? Well, a key observation that is valid on this picture, and that actually remain true in the completely general case, is that when we have a minimum the level curve of f is actually tangent to our hyperbola.

It is tangent to the set of points where x, y equals three, to the hyperbola.

Let's write that down. We observe that at the minimum the level curve of f is tangent to the hyperbola.

Remember, the hyperbola is given by the equal g equals three, so it is a level curve of g .

We have a level curve of f and a level curve of g that are tangent to each other. And I claim that is going to be the general situation that we are interested in.

How do we try to solve for points where this happens?

How do we find x, y where the level curves of f and g are tangent to each other? Let's think for a second.

If the two level curves are tangent to each other that means they have the same tangent line. That means that the normal vectors should be parallel. Let me maybe draw a picture here. This is the level curve maybe f equals something. And this is the level curve g equals constant. Here my constant is three.

Well, if I look for gradient vectors, the gradient of f will be perpendicular to the level curve of f .

The gradient of g will be perpendicular to the level curve of g . They don't have any reason to be of the same size, but they have to be parallel to each other. Of course, they could also be parallel pointing in opposite directions.

But the key point is that when this happens the gradient of f is parallel to the gradient of g .

Well, let's check that. Here is a point.

And I can plot the gradient of f in blue.

The gradient of g in yellow. And you see, in most of these places, somehow the two gradients are not really parallel. Actually, I should not be looking at random points. I should be looking only on the hyperbola. I want points on the hyperbola where the two gradients are parallel.

Well, when does that happen? Well, it looks like it will happen here. When I am at a minimum, the two gradient vectors are parallel.

It is not really proof. It is an example that seems to be convincing. So far things work pretty well.

How do we decide if two vectors are parallel?

Well, they are parallel when they are proportional to each other. You can write one of them as a constant times the other one, and that constant usually one uses the Greek letter lambda. I don't know if you have seen it before. It is the Greek letter for L .

And probably, I am sure, it is somebody's idea of paying tribute to Lagrange by putting an L in there. Lambda is just a constant.

And we are looking for a scalar lambda and points x and y where this holds. In fact, what we are doing is replacing min/max problems in two variables with a constraint between them by a set of equations involving, you will see, three variables.

We had min/max with two variables x , y , but no independent. We had a constraint g of x , y equals constant. And that becomes something new.

That becomes a system of equations where we have to solve, well, let's write down what it means for gradient f to be proportional to gradient g . That means that $f_{\text{sub } x}$ should be lambda times $g_{\text{sub } x}$, and $f_{\text{sub } y}$ should be lambda times $g_{\text{sub } y}$. Because the gradient vectors here are $f_{\text{sub } x}$, $f_{\text{sub } y}$ and $g_{\text{sub } x}$, $g_{\text{sub } y}$. If you have a third variable z then you have also an equation $f_{\text{sub } z}$ equals lambda $g_{\text{sub } z}$.

Now, let's see. How many unknowns do we have in these equations? Well, there is x , there is y and there is lambda. We have three unknowns and have only two equations. Something is missing.

Well, I mean x and y are not actually independent.

They are related by the equation g of x , y equals c , so we need to add the constraint g equals c .

And now we have three equations involving three variables.

Let's see how that works. Here remember we have f equals x squared y squared and $g = xy$. What is f sub x ?

It is going to be $2x$ equals λ times, what is g sub x, y .

Maybe I should write here f sub x equals λ g sub x just to remind you. Then we have f sub y equals λ g sub y . F sub y is $2y$ equals λ times g sub y is x . And then our third equation g equals c becomes xy equals three.

So, that is what you would have to solve.

Any questions at this point? No.

Yes? How do I know the direction of a gradient? Do you mean how do I know that it is perpendicular to a level curve?

Oh, how do I know if it points in that direction on the opposite one? Well, that depends.

I mean we'd seen in last time, but the gradient is perpendicular to the level and points towards higher values of a function. So it could be -- Wait.

What did I have? It could be that my gradient vectors up there actually point in opposite directions.

It doesn't matter to me because it will still look the same in terms of the equation, just λ will be positive or negative, depending on the case. I can handle both situations.

It's not a problem. I can allow λ to be positive or negative. Well, in this example, it looks like λ will be positive.

If you look at the picture on the plot.

Yes? Well, because actually they are not equal to each other. If you look at this point where the hyperbola and the circle touch each other, first of all, I don't know which circle I am going to look at. I am trying to solve, actually, for the radius of the circle.

I am trying to find what the minimum value of f is.

And, second, at that point, the value of f and the value of g are not equal.

g is equal to three because I want the hyperbola x equals three. The value of f will be the square of a distance, whatever that is.

I think it will end up being 6, but we will see.

So, you cannot really set them equal because you don't know what f is equal to in advance. Yes?

Not quite. Actually, here I am just using this idea of finding a point closest to the origin to illustrate an example of a min/max problem.

The general problem we are trying to solve is minimize f subject to g equals constant. And what we are going to do for that is we are really going to say instead let's look at places where gradient f and gradient g are parallel to each other and solve for equations of that. I think we completely lose the notion of closest point if we just look at these equations.

We don't really say anything about closest points anymore.

Of course, that is what they mean in the end.

But, in the general setting, there is no closest point involved anymore. OK.

Yes? Yes.

It is always going to be the case that, at the minimum, or at the maximum of a function subject to a constraint, the level curves of f and the level curves of g will be tangent to each other.

That is the basis for this method.

I am going to justify that soon. It could be minimum or maximum.

In three-dimensions it could even be a saddle point.

And, in fact, I should say in advance, this method will not tell us whether it is a minimum or a maximum. We do not have any way of knowing, except for testing values.

We cannot use second derivative tests or anything like that.

I will get back to that. Yes?

Yes. Here you can set y equals to favor x . Then you can minimize x squared plus nine over x squared. In general, if I am trying to solve a more complicated problem, I might not be able to solve. I am doing an example where, indeed, here you could solve and remove one variable, but you cannot always do that. And this method will still work.

The other one won't. OK.

I don't see any other questions. Are there any other questions?

No. OK.

I see a lot of students stretching and so on, so it is very confusing for me. How do we solve these equations?

Well, the answer is in general we might be in deep trouble.

There is no general method for solving the equations that you get from this method. You just have to think about them. Sometimes it will be very easy.

Sometimes it will be so hard that you cannot actually do it without the computer. Sometimes it will be just hard enough to be on Part B of this week's problem set.

I claim in this case we can actually do it without so much trouble, because actually we can think of this as a two by two linear system in x and y . Well, let me do something.

Let me rewrite the first two equations as $2x - \lambda y = 0$.

And $\lambda x - 2y = 0$. And $xy = 3$.

That is what we want to solve. Well, I can put this into matrix form. Two minus lambda, lambda minus two times x , y equals $0,0$.

Now, how do I solve a linear system matrix times x, y equals zero? Well, I always have an obvious solution. x and y both equal to zero.

Is that a good solution? No, because zero times zero is not three. We want another solution, the trivial solution. $0,0$ does not solve the constraint equation xy equals three, so we want another solution. When do we have another solution? Well, when the determinant of a matrix is zero. We have other solutions that exist only if determinant of a matrix is zero.

M is this guy. Let's compute the determinant.

Well, that seems to be negative four plus lambda squared.

That is zero exactly when lambda squared equals four, which is lambda is plus or minus two.

Already you see here it is a the level of difficulty that is a little bit much for an exam but perfectly fine for a problem set or for a beautiful lecture like this one.

How do we deal with -- Well, we have two cases to look at.

Lambda equals two or lambda equals minus two.

Let's start with lambda equals two.

If I set lambda equals two, what does this equation become?

Well, it becomes $x = y$. This one becomes $y = x$.

Well, they seem to be the same. $x = y$.

And then the equation $xy = 3$ becomes, well, $x^2 = 3$. I have two solutions.

One is $x = \sqrt{3}$ and, therefore, $y = \sqrt{3}$ as well, or negative root three and negative root three.

Let's look at the other case. If I set lambda equal to negative two then I get $2x = -2y$.

That means $x = -y$. The second one, $2y = -2x$. That is $y = -x$.

Well, that is the same thing. And $xy = 3$ becomes negative $x^2 = 3$. Can we solve that?

No. There are no solutions here.

Now we have two candidate points which are these two points, root three, root three or negative root three, negative root three. OK.

Let's actually look at what we have here.

Maybe you cannot read the coordinates, but the point that I have here is indeed root three, root three.

How do we see that lambda equals two?

Well, if you look at this picture, the gradient of f , that is the blue vector, is indeed twice the yellow vector, gradient g . That is where you read the value of lambda. And we have the other solution which is somewhere here.

Negative root three, negative root there. And there, again, lambda equals two. The two vectors are proportional by a factor of two. Yes?

No, solutions are not quite guaranteed to be absolute minima or maxima. They are guaranteed to be somehow critical points end of a constraint.

That means if you were able to solve and eliminate the variable that would be a critical point. When you have the same problem, as we have critical points, are they maxima or minima?

And the answer is, well, we won't know until we check. More questions?

No. Yes?

What is a Lagrange multiplier? Well, it is this number λ that is called the multiplier here.

It is a multiplier because it is what you have to multiply gradient of g by to get gradient of f .

It multiplies.

Let's try to see why is this method valid?

Because so far I have shown you pictures and have said see they are tangent. But why is it that they have to be tangent in general? Let's think about it.

Let's say that we are at constrained min or max.

What that means is that if I move on the level g equals constant then the value of f should only increase or only decrease. But it means, in particular, to first order it will not change. At an unconstrained min or max, partial derivatives are zero. In this case, derivatives are zero only in the allowed directions.

And the allowed directions are those that stay on the levels of this g equals constant. In any direction along the level set $g = c$ the rate of change of f must be zero.

That is what happens at minima or maxima.

Except here, of course, we look only at the allowed directions. Let's say the same thing in terms of directional derivatives.

That means for any direction that is tangent to the constraint level g equal c , we must have df over ds in the direction of u equals zero. I will draw a picture.

Let's say now I am in three variables just to give you different examples. Here I have a level surface g equals c . I am at my point.

And if I move in any direction that is on the level surface, so I move in the direction u tangent to the level surface, then the rate of change of f in that direction should be zero.

Now, remember what the formula is for this guy.

Well, we have seen that this guy is actually $\nabla f \cdot u$.

That means any such vector u must be perpendicular to the gradient of f . That means that the gradient of f should be perpendicular to anything that is tangent to this level. That means the gradient of f should be perpendicular to the level set.

That is what we have shown.

But we know another vector that is also perpendicular to the level set of g . That is the gradient of g .

We conclude that the gradient of f must be parallel to the gradient of g because both are perpendicular to the level set of g . I see confused faces, so let me try to tell you again where that comes from.

We said if we had a constrained minimum or maximum, if we move in the level set of g , f doesn't change.

Well, it doesn't change to first order.

It is the same idea as when you are looking for a minimum you set the derivative equal to zero.

So the derivative in any direction, tangent to g equals c , should be the directional derivative of f , in any such direction, should be zero.

That is what we mean by critical point of f .

And so that means that any vector u , any unit vector tangent to the level set of g is going to be perpendicular to the gradient of f . That means that the gradient of f is perpendicular to the level set of g .

If you want, that means the level sets of f and g are tangent to each other. That is justifying what we have observed in the picture that the two level sets have to be tangent to each other at the prime minimum or maximum.

Does that make a little bit of sense?

Kind of. I see at least a few faces nodding so I take that to be a positive answer.

Since I have been asked by several of you, how do I know if it is a maximum or a minimum?

Well, warning, the method doesn't tell whether a solution is a minimum or a maximum.

How do we do it? Well, more bad news.

We cannot use the second derivative test.

And the reason for that is that we care actually only about these specific directions that are tangent to variable of g .

And we don't want to bother to try to define directional second derivatives. Not to mention that actually it wouldn't work. There is a criterion but it is much more complicated than that. Basically, the answer for us is that we don't have a second derivative test in this situation. What are we left with?

Well, we are just left with comparing values.

Say that in this problem you found a point where f equals three, a point where f equals nine, a point where f equals 15.

Well, then probably the minimum is the point where f equals three and the maximum is 15. Actually, in this case, where we found minima, these two points are tied for minimum. What about the maximum?

What is the maximum of f on the hyperbola?

Well, it is infinity because the point can go as far as you want from the origin. But the general idea is if we have a good reason to believe that there should be a minimum, and it's not like at infinity or something weird like that, then the minimum will be a solution of the Lagrange multiplier equations. We just look for all the solutions and then we choose the one that gives us the lowest value. Is that good enough?

Let me actually write that down.

To find the minimum or the maximum, we compare values of f at the various solutions -- -- to Lagrange multiplier equations.

I should say also that sometimes you can just conclude by thinking geometrically. In this case, when it is asking you which point is closest to the origin you can just see that your answer is the correct one.

Let's do an advanced example. Advanced means that -- Well, this one I didn't actually dare to put on top of the other problem sets. Instead, I am going to do it.

What is this going to be about? We are going to look for a surface minimizing pyramid. Let's say that we want to

build a pyramid with a given triangular base -- -- and a given volume. Say that I have maybe in the x, y plane I am giving you some triangle.

And I am going to try to build a pyramid.

Of course, I can choose where to put the top of a pyramid.

This guy will end up being behind now.

And the constraint and the goal is to minimize the total surface area. The first time I taught this class, it was a few years ago, was just before they built the Stata Center. And then I used to motivate this problem by saying Frank Gehry has gone crazy and has been given a triangular plot of land he wants to put a pyramid.

There needs to be the right amount of volume so that you can put all the offices in there. And he wants it to be, actually, covered in solid gold.

And because that is expensive, the administration wants him to cut the costs a bit. And so you have to minimize the total size so that it doesn't cost too much.

We will see if MIT comes up with a triangular pyramid building. Hopefully not.

It could be our next dorm, you never know.

Anyway, it is a fine geometry problem.

Let's try to think about how we can do this.

The natural way to think about it would be -- Well, what do we have to look for first?

We have to look for the position of that top point.

Remember we know that the volume of a pyramid is one-third the area of base times height. In fact, fixing the volume, knowing that we have fixed the area of a base, means that we are fixing the height of the pyramid.

The height is completely fixed. What we have to choose just is where do we put that top point? Do we put it smack in the middle of a triangle or to a side or even anywhere we want?

Its z coordinate is fixed. Let's call h the height.

What we could do is something like this.

We say we have three points of a base.

Let's call them p_1 at $(x_1, y_1, 0)$; p_2 at $(x_2, y_2, 0)$; p_3 at $(x_3, y_3, 0)$.

This point p is the unknown point at (x, y, h) . We know the height.

And then we want to minimize the sum of the areas of these three triangles. One here, one here and one at the back. And areas of triangles we know how to express by using length of cross-product.

It becomes a function of x and y .

And you can try to minimize it. Actually, it doesn't quite work.

The formulas are just too complicated.

You will never get there. What happens is actually maybe we need better coordinates. Why do we need better coordinates? That is because the geometry is kind of difficult to do if you use x, y coordinates.

I mean formula for cross-product is fine, but then the length of the vector will be annoying and just doesn't look good. Instead, let's think about it differently.

I claim if we do it this way and we express the area as a function of x, y , well, actually we can't solve for a minimum. Here is another way to do it.

Well, what has worked pretty well for us so far is this geometric idea of base times height.

So let's think in terms of the heights of side triangles.

I am going to use the height of these things.

And I am going to say that the area will be the sum of three terms, which are three bases times three heights.

Let's give names to these quantities.

Actually, for that it is going to be good to have the point in the xy plane that lives directly below p .

Let's call it q . P is the point that coordinates x, y, h . And let's call q the point that is just below it and so its coordinates are $x, y, 0$. Let's see.

Let me draw a map of this thing. p_1, p_2, p_3 and I have my point q in the middle. Let's see.

To know these areas, I need to know the base.

Well, the base I can decide that I know it because it is part of my given data. I know the sides of this triangle. Let me call the lengths a_1 , a_2 , a_3 . I also need to know the height, so I need to know these lengths.

How do I know these lengths? Well, its distance in space, but it is a little bit annoying.

But maybe I can reduce it to a distance in the plane by looking instead at this distance here. Let me give names to the distances from q to the sides. Let's call u_1 , u_2 , u_3 the distances from q to the sides.

Well, now I can claim I can find, actually, sorry. I need to draw one more thing.

I claim I have a nice formula for the area, because this is vertical and this is horizontal so this length here is u_3 , this length here is h .

So what is this length here? It is the square root of u_3 squared plus h squared. And similarly for these other guys. They are square roots of a u squared plus h squared. The heights of the faces are square root of u_1 squared times h squared.

And similarly with u_2 and u_3 . So the total side area is going to be the area of the first faces, one-half of base times height, plus one-half of a base times a height plus one-half of the third one.

It doesn't look so much better. But, trust me, it will get better. Now, that is a function of three variables, u_1 , u_2 , u_3 .

And how do we relate u_1 , u_2 , u_3 to each other?

They are probably not independent.

Well, let's cut this triangle here into three pieces like that. Then each piece has side -- Well, let's look at it the piece of the bottom.

It has base a_3 , height u_3 . Cutting base into three tells you that the area of a base is one-half of a_1 , u_1 plus one-half of a_2 , u_2 plus one-half of a_3 , u_3 . And that is our constraint.

My three variables, u_1 , u_2 , u_3 , are constrained in this way. The sum of this figure must be the area of a base. And I want to minimize that guy.

So that is my g and that guy here is my f .

Now we try to apply our Lagrange multiplier equations.

Well, partial f of a partial u_1 is -- Well, if you do the calculation, you will see it is one-half a_1 , u_1 over square root of

$u_1^2 + h^2 = \lambda$, what is $\partial g / \partial a_1$?

That one you can do, I am sure. It is $\frac{1}{2} a_1$.

Oh, these guys simplify. If you do the same with the second one -- -- things simplify again.

And the same with the third one. Well, you will get, after simplifying, u_3 over square root of $u_3^2 + h^2 = \lambda$.

Now, that means this guy equals this guy equals this guy.

They are all equal to λ . And, if you think about it, that means that $u_1 = u_2 = u_3$. See, it looked like scary equations but the solution is very simple.

What does it mean? It means that our point q should be equidistant from all three sides.

That is called the incenter. Q should be in the incenter.

The next time you have to build a golden pyramid and don't want to go broke, well, you know where to put the top.

If that was a bit fast, sorry. Anyway, it is not completely crucial. But go over it and you will see it works. Have a nice weekend.