MITOCW | ocw-18_02-f07-lec09_220k

 Today we are going to see how to use what we saw last time about partial derivatives to handle minimization or maximization problems involving functions of several variables. Remember last time we said that when we have a function, say, of two variables, x and y, then we have actually two different derivatives, partial f, partial x, also called f sub x, the derivative with respect to x keeping y constant. And we have partial f, partial y, also called f sub y, where we vary y and we keep x as a constant. And now, one thing I didn't have time to tell you about but hopefully you thought about in recitation yesterday, is the approximation formula that tells you what happens if you vary both x and y. f sub x tells us what happens if we change x a little bit, by some small amount delta x. f sub y tells us how f changes, if you change y by a small amount delta y. If we do both at the same time then the two effects will add up with each other, because you can imagine that first you will change x and then you will change y. Or the other way around. It doesn't really matter. If we change x by a certain amount delta x, and if we change y by the amount delta y, and let's say that we have z= f(x, y) then that changes by an amount which is approximately f sub x times delta x plus f sub y times delta y. And that is one of the most important formulas about partial derivatives. The intuition for this, again, is just the two effects of if I change x by a small amount and then I change y. Well, first changing x will modify f, how much does it modify f? The answer is the rate change is f sub x. And if I change y then the rate of change of f when I change y is f sub y. So all together I get this change as a value of f. And, of course, that is only an approximation formula. Actually, there would be higher order terms involving second and third derivatives and so on. One way to justify this -- Sorry. I was distracted by the microphone. OK. How do we justify this formula? Well, one way to think about it is in terms of tangent plane approximation. Let's think about the tangent plane with regard to a function f. We have some pictures to show you. It will be easier if I show you pictures. Remember, partial f, partial x was obtained by looking at the situation where y is held constant. That means I am slicing the graph of f by a plane that is parallel to the x, z plane. And when I change x, z changes, and the slope of that is going to be the derivative with respect to x. Now, if I do the same in the other direction then I will have similarly the slope in a slice now parallel to the y, z plane that will be partial f, partial y. In fact, in each case, I have a line. And that line is tangent to the surface. Now, if I have two lines tangent to the surface, well, then together they determine for me the tangent plane to the surface. Let's try to see how that works. We know that f sub x and f sub y are the slopes of two tangent lines to this plane, two tangent lines to the graph. And let's write down the equations of these lines. I am not going to write parametric equations. I am going to write them in terms of x, y, z coordinates. Let's say that partial f of a partial x at the given point is equal to a. That means that we have a line given by the following conditions. I am going to keep y constant equal to y0. And I am going to change x. And, as I change x, z will change at the rate that is equal to a. That would be z = 0 a(x - x0). That is how you would describe a line that, I guess, the one that is plotted in green here, been dissected with the slice parallel to the x, z plane. I hold y constant equal to y0. And z is a function of x that varies with a rate of a. And now if I look similarly at the other slice, let's say that the partial with respect to y is equal to b, then I get another line which is obtained by the fact that z now will depend on y. And the rate of change with respect to y will be b. While x is held constant equal to x0. These two lines are both going to be in the tangent plane to the surface. They are both tangent to the graph of f and together they determine the plane. And that plane is just given by the formula z = z0 a(x - x0) b (y - y0). If you look at what happens -- This is the equation of a plane. z equals constant times x plus constant times y plus constant. And if you look at what happens if I hold y constant and vary x, I will get the first line. If I hold x constant and vary y, I get the second line. Another way to do it, of course, would provide actually parametric equations of these lines, get vectors along them and then take the cross-product to get the normal vector to the plane. And then get this equation for the plane using the normal vector. That also works and it gives you the same formula. If you are curious of the exercise, do it again using parametrics and using cross-product to get the plane equation. That is how we get the tangent plane. And now what this approximation formula here says is that, in fact, the graph of a function is close to the tangent plane. If we were moving on the tangent plane, this would be an actual equality. Delta z would be a linear function of delta x and delta y. And the graph of a function is near the tangent plane, but is not quite the same, so it is only an approximation for small delta x and small delta y. The approximation formula says the graph of f is close to its tangent plane. And we can use that formula over here now to estimate how the value of f changes if I change x and y at the same time. Questions about that? Now that we have caught up with what we were supposed to see on Tuesday, I can tell you now about max and min problems. That is going to be an application of partial derivatives to look at optimization problems. Maybe ten years from now, when you have a real job, your job might be to actually minimize the cost of something or maximize the profit of something or whatever. But typically the function that you will have to strive to minimize or maximize will depend on several variables. If you have a function of one variable, you know that to find its minimum or its maximum you look at the derivative and set that equal to zero. And you try to then look at what happens to the function. Here it is going to be kind of similar, except, of course, we have several derivatives. For today we will think about a function of two variables, but it works exactly the same if you have three variables, ten variables, a million variables. The first observation is that if we have a local minimum or a local maximum then both partial derivatives, so partial f partial x and partial f partial y, are both zero at the same time. Why is that? Well, let's say that f of x is zero. That means when I vary x to first order the function doesn't change. Maybe that is because it is going through... If I look only at the slice parallel to the x-axis then maybe I am going through the minimum. But if partial f, partial y is not 0 then actually, by changing y, I could still make a value larger or smaller. That wouldn't be an actual maximum or minimum. It would only be a maximum or minimum if I stay in the slice. But if I allow myself to change y that doesn't work. I need actually to know that if I change y the value will not

But if I allow myself to change y that doesn't work. I need actually to know that if I change y the value will not change either to first order. That is why you also need partial f, partial y to be zero. Now, let's say that they are both zero. Well, why is that enough? It is essentially enough because of this formula telling me that if both of these guys are zero then to first order the function doesn't change. Then, of course, there will be maybe quadratic terms that will actually turn that, you know, this won't really say that your function is actually constant. It will just tell you that maybe it will actually be quadratic or higher order in delta x and delta y. That is what you expect to have at a maximum or a minimum. The condition is the same thing as saying that the tangent plane to the graph is actually going to be horizontal. And that is what you want to have. Say you have a minimum, well, the tangent plane at this point, at the bottom of the graph is going to be horizontal. And you can see that on this equation of a tangent plane, when both these coefficients are 0 that is when the equation becomes z equals constant: the horizontal plane. Does that make sense? We will have a name for this kind of point because, actually, what we will see very soon is that these conditions are necessary but are not sufficient. There are actually other kinds of points where the partial derivatives are zero. Let's give a name to this. We say the definition is (x0, y0) is a critical point of f -- -- if the partial derivative, with respect to x, and partial derivative with respect to y are both zero. Generally, you would want all the partial derivatives, no matter how many variables you have, to be zero at the same time. Let's see an example. Let's say I give you the function f(x;y)= x^2 - 2xy 3y^2 2x - 2y. And let's try to figure out whether we can minimize or maximize this. What we would start doing immediately is taking the partial derivatives. What is f sub x? It starts with 2x - 2y 0 2. Remember that y is a constant so this differentiates to zero. Now, if we do f sub y, that is going to be 0-2x 6y-2. And what we want to do is set these things to zero. And we want to solve these two equations at the same time. An important thing to remember, and maybe I should have told you a couple of weeks ago already, if you have two equations to solve, well, it is very good to try to simplify them by adding them together or whatever, but you must keep two equations. If you have two equations, you shouldn't end up with just one equation out of nowhere. For example here, we can certainly simplify things by summing them together. If we add them together, well, the x's cancel and the constants cancel. In fact, we are just left with 4y for zero. That is pretty good. That tells us y should be zero. But then we should, of course, go back to these and see what else we know. Well, now it tells us, if you put y = 0 it tells you 2x 2 = 0. That tells you x = - 1. We have one critical point that is (x, y) = (- 1; 0). Any questions so far? No. Well, you should have a question. The question should be how do we know if it is a maximum or a minimum? Yeah. If we had a function of one variable, we would decide things based on the second derivative. And, in fact, we will see tomorrow how to do things based on the second derivative. But that is kind of tricky because there are a lot of second derivatives. I mean we already have two first derivatives. You can imagine that if you keep taking partials you may end up with more and more, so we will have to figure out carefully what the condition should be. We will do that tomorrow. For now, let's just try to look a bit at how do we understand these things by hand? In fact, let me point out to you immediately that there is more than maxima and minima. Remember, we saw the example of x^2 y^2. That has a critical point. That critical point is obviously a minimum. And, of course, it could be a local minimum because it could be that if you have a more complicated function there is indeed a minimum here, but then elsewhere the function drops to a lower value. We call that just a local minimum to say that it is a minimum if you stick two values that are close enough to that point. Of course, you also have local maximum, which I didn't plot, but it is easy to plot. That is a local maximum. But there is a third example of critical point, and that is a saddle point. The saddle point, it is a new phenomena that you don't really see in single variable calculus. It is a critical point that is neither a minimum nor a maximum because, depending on which direction you look in, it's either one or the other. See the point in the middle, at the origin, is a saddle point. If you look at the tangent plane to this graph, you will see that it is actually horizontal at the origin. You have this mountain pass where the ground is horizontal. But, depending on which direction you go, you go up or down. So, we say that a point is a saddle point if it is neither a minimum or a maximum. Possibilities could be a local min, a local max or a saddle. Tomorrow we will see how to decide which one it is, in general, using second derivatives. For this time, let's just try to do it by hand. I just want to observe, in fact, I can try to, you know, these examples that I have here, they are x^2 y^2, y^2 - x^2, they are sums or differences of squares. And, if we know that we can put things as sum of squares for example, we will be done. Let's try to express this maybe as the square of something. The main problem is this 2xy. Observe we know something that starts with x^2 - 2xy but is actually a square of something else. It would be x^2 - 2xy y^2, not plus 3y2. Let's try that. So, we are going to complete the square. I am going to say it is x minus y squared, so it gives me the first two terms and also the y2. Well, I still need to add two more y^2, and I also need to add, of course, the 2x and - 2y. It is still not simple enough for my taste. I can actually do better. These guys look like a sum of squares, but here I have this extra stuff, 2x - 2y. Well, that is 2 (x - y). It looks like maybe we can modify this and make this into another square. So, in fact, I can simplify this further to (x - y 1)^2. That would be (x - y)^2 2(x - y), and then there is a plus one. Well, we don't have a plus one so let's remove it by subtracting one. And I still have my 2y^2. Do you see why this is the same function? Yeah. Again, if I expand x minus y plus one squared, I get (x - y)^2 2 (x - y) 1. But I will have minus one that will cancel out and then I have a plus 2y^2. Now, what I know is a sum of two squared minus one. And this critical point, (x,y) = (-1;0), that is actually when this is zero and that is zero, so that is the smallest value. This is always greater or equal to zero, the same with that one, so that is always at least minus one. And minus one happens to be the value at the critical point. So, it is a minimum. Now, of course here I was very lucky. I mean, generally, I couldn't expect things to simplify that much. In fact, I cheated. I started from that, I expanded, and then that is how I got my example. The general method will be a bit different, but you will see it will actually also involve completing squares. Just there is more to it than what we have seen. We will come back to this tomorrow. Sorry? How do I know that this equals -- How do I know that the whole function is greater or equal to negative one? Well, I wrote f of x, y as something squared plus 2y^2 - 1. This squared is always a positive number and not a negative. It is a square. The square of something is always non-negative. Similarly, y^2 is also always non-negative. So if you add something that is at least zero plus something that is at least zero and you subtract one, you get always at least minus one. And, in fact, the only way you can get minus one is if both of these guys are zero at the same time.

minus one. And, in fact, the only way you can get minus one is if both of these guys are zero at the same time. That is how I get my minimum. More about this tomorrow. In fact, what I would like to tell you about now instead is a nice application of min, max problems that maybe you don't think of as a min, max problem that you will see. I mean you will think of it that way because probably your calculator can do it for you or, if not, your computer can do it for you. But it is actually something where the theory is based on minimization in two variables. Very often in experimental sciences you have to do something called least-squares intercalation. And what is that about? Well, it is the idea that maybe you do some experiments and you record some data. You have some data x and some data y. And, I don't know, maybe, for example, x is -- Maybe your measuring frogs and you're trying to measure how bit the frog leg is compared to the eyes of the frog, or you're trying to measure something. And if you are doing chemistry then it could be how much you put of some reactant and how much of the output product that you wanted to synthesize generated. All sorts of things. Make up your own example. You measure basically, for various values of x, what the value of y ends up being. And then you like to claim these points are kind of aligned. And, of course, to a mathematician they are not aligned. But, to an experimental scientist, that is evidence that there is a relation between the two. And so you want to claim -- And in your paper you will actually draw a nice little line like that. The functions depend linearly on each of them. The question is how do we come up with that nice line that passes smack in the middle of the points? The question is, given experimental data xi, yi -- Maybe I should actually be more precise. You are given some experimental data. You have data points x1, y1, x2, y2 and so on, xn, yn, the question would be find the "best fit" line of a form y equals ax b that somehow approximates very well this data. You can also use that right away to predict various things. For example, if you look at your new homework, actually the first problem asks you to predict how many iPods will be on this planet in ten years looking at past sales and how they behave. One thing, right away, before you lose all the money that you don't have yet, you cannot use that to predict the stock market. So, don't try to use that to make money. It doesn't work. One tricky thing here that I want to draw your attention to is what are the unknowns here? The natural answer would be to say that the unknowns are x and y. That is not actually the case. We are not going to solve for some x and y. I mean we have some values given to us. And, when we are looking for that line, we don't really care about the perfect value of x. What we care about is actually these coefficients a and b that will tell us what the relation is between x and y. In fact, we are trying to solve for a and b that will give us the nicest possible line for these points. The unknowns, in our equations, will have to be a and b, not x and y. The question really is find the "best" a and b. And, of course, we have to decide what we mean by best. Best will mean that we minimize some function of a and b that measures the total errors that we are making when we are choosing this line compared to the experimental data. Maybe, roughly speaking, it should measure how far these points are from the line. But now there are various ways to do it. And a lot of them are valid they give you different answers. You have to decide what it is that you prefer. For example, you could measure the distance to the line by projecting perpendicularly. Or you could measure instead, for a given value of x, the difference between the experimental value of y and the predicted one. And that is often more relevant because these guys actually may be expressed in different units. They are not the same type of quantity. You cannot actually combine them arbitrarily. Anyway, the convention is usually we measure distance in this way. Next, you could try to minimize the largest distance. Say we look at who has the largest error and we make that the smallest possible. The drawback of doing that is experimentally very often you have one data point that is not good because maybe you fell asleep in front of the experiment. And so you didn't measure the right thing. You tend to want to not give too much importance to some data point that is far away from the others. Maybe instead you want to measure the average distance or maybe you want to actually give more weight to things that are further away. And then you don't want to do the distance with a square of the distance. There are various possible answers, but one of them gives us actually a particularly nice formula for a and b. And so that is why it is the universally used one. Here it says list squares. That's because we will measure, actually, the sum of the squares of the errors. And why do we do that? Well, part of it is because it looks good. When you see this plot in scientific papers they really look like the line is indeed the ideal line. And the second reason is because actually the minimization problem that we will get is particularly simple, well-posed and easy to solve. So we will have a nice formula for the best a and the best b. If you have a method that is simple and gives you a good answer then that is probably good. We have to define best. Here it is in the sense of minimizing the total square error. Or maybe I should say total square deviation instead. What do I mean by this? The deviation for each data point is the difference between what you have measured and what you are predicting by your model. That is the difference between y1 and axi plus b. Now, what we will do is try to minimize the function capital D, which is just the sum for all the data points of the square of a deviation. Let me go over this again. This is a function of a and b. Of course there are a lot of letters in here, but xi and yi in real life there will be numbers given to you. There will be numbers that you have measured. You have measured all of this data. They are just going to be numbers. You put them in there and you get a function of a and b. Any questions? How do we minimize this function of a and b? Well, let's use your knowledge. Let's actually look for a critical point. We want to solve for partial d over partial a= 0, partial d over partial b = 0. That is how we look for critical points. Let's take the derivative of this with respect to a. Well, the derivative of a sum is sum of the derivatives. And now we have to take the derivative of this quantity squared. Remember, we take the derivative of the square. We take twice this quantity times the derivative of what we are squaring. We will get 2(yi - axi) b times the derivative of this with respect to a. What is the derivative of this with respect to a? Negative xi, exactly. And so we will want this to be 0. And partial d over partial b, we do the same thing, but different shading with respect to b instead of with respect to a. Again, the sum of squares twice yi minus axi equals b times the derivative of this with respect to b is, I think, negative one. Those are the equations we have to solve. Well, let's reorganize this a little bit. The first equation. See, there are a's and there are b's in these equations. I am going to just look at the coefficients of a and b. If you have good eyes, you can see probably that these are actually linear equations in a and b. There is a lot of clutter with all these x's and y's all over the place. Let's actually try to expand things and make that more apparent. The first thing I will do is actually get rid of these factors of two. They are just not very important. I can simplify things.

first thing I will do is actually get rid of these factors of two. They are just not very important. I can simplify things. Next, I am going to look at the coefficient of a. I will get basically a times xi squared. Let me just do it and should be clear. I claim when we simplify this we get xi squared times a plus xi times b minus xiyi. And we set this equal to zero. Do you agree that this is what we get when we expand that product? Yeah. Kind of? OK. Let's do the other one. We just multiply by minus one, so we take the opposite of that which would be axi plus b. I will write that as xia plus b minus yi. Sorry. I forgot the n here. And let me just reorganize that by actually putting all the a's together. That means I will have sum of all the xi2 times a plus sum of xib minus sum of xiyi equal to zero. If I rewrite this, it becomes sum of xi2 times a plus sum of the xi's time b, and let me move the other guys to the other side, equals sum of xiyi. And that one becomes sum of xi times a. Plus how many b's do I get on this one? I get one for each data point. When I sum them together, I will get n. Very good. N times b equals sum of yi. Now, this quantities look scary, but they are actually just numbers. For example, this one, you look at all your data points. For each of them you take the value of x and you just sum all these numbers together. What you get, actually, is a linear system in a and b, a two by two linear system. And so now we can solve this for a and b. In practice, of course, first you plug in the numbers for xi and yi and then you solve the system that you get. And we know how to solve two by two linear systems, I hope. That's how we find the best fit line. Now, why is that going to be the best one instead of the worst one? We just solved for a critical point. That could actually be a maximum of this error function D. We will have the answer to that next time, but trust me. If you really want to go over the second derivative test that we will see tomorrow and apply it in this case, it is quite hard to check, but you can see it is actually a minimum. I will just say -- -- we can show that it is a minimum. Now, the event with the linear case is the one that we are the most familiar with. Least-squares interpolation actually works in much more general settings. Because instead of fitting for the best line, if you think it has a different kind of relation then maybe you can fit in using a different kind of formula. Let me actually illustrate that with an example. I don't know if you are familiar with Moore's law. It is something that is supposed to tell you how quickly basically computer chips become smarter faster and faster all the time. It's a law that says things about the number of transistors that you can fit onto a computer chip. Here I have some data about -- Here is data about the number of transistors on a standard PC processor as a function of time. And if you try to do a best-line fit, well, it doesn't seem to follow a linear trend. On the other hand, if you plug the diagram in the log scale, the log of a number of transitions as a function of time, then you get a much better line. And so, in fact, that means that you had an exponential relation between the number of transistors and time. And so, actually that's what Moore's law says. It says that the number of transistors in the chip doubles every 18 months or every two years. They keep changing the statement. How do we find the best exponential fit? Well, an exponential fit would be something of a form y equals a constant times exponential of a times x. That is what we want to look at. Well, we could try to minimize a square error like we did before. That doesn't work well at all. The equations that you get are very complicated. You cannot solve them. But remember what I showed you on this log plot. If you plot the log of y as a function of x then suddenly it becomes a linear relation. Observe, this is the same as ln of y equals ln of c plus ax. And that is the linear best fit. What you do is you just look for the best straight line fit for the log of y. That is something we already know. But you can also do, for example, let's say that we have something more complicated. Let's say that we have actually a quadratic law. For example, y is of the form ax^2 bx c. And, of course, you are trying to find somehow the best. That would mean here fitting the best parabola for your data points. Well, to do that, you would need to find a, b and c. And now you will have actually a function of a, b and c, which would be the sum of the old data points of the square deviation. And, if you try to solve for critical points, now you will have three equations involving a, b and c, in fact, you will find a three by three linear system. And it works the same way. Just you have a little bit more data. Basically, you see that this best fit problems are an example of a minimization problem that maybe you didn't expect to see minimization problems come in. But that is really the way to handle these questions. Tomorrow we will go back to the question of how do we decide whether it is a minimum or a maximum. And we will continue exploring in terms of several variables.