Deep Learning

Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

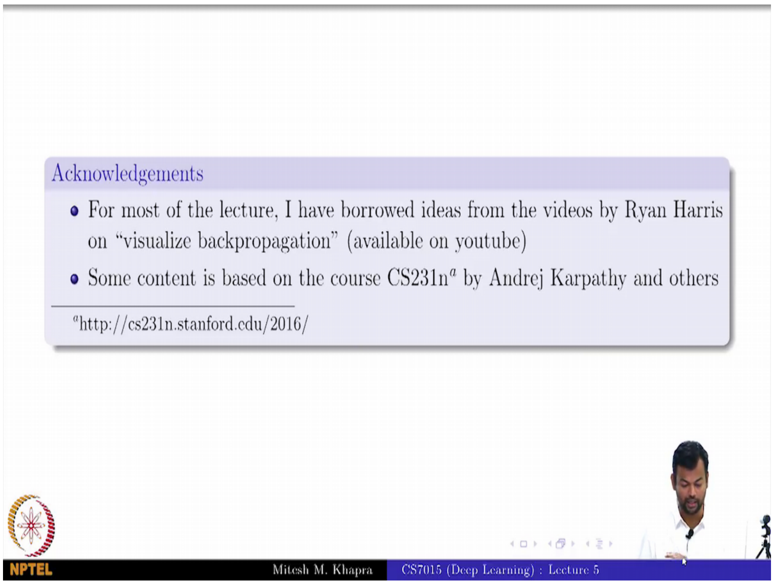
Lecture - 05
Gradient Descent (GD), Momentum Based GD, Nesterov Accelerated GD, Stochastic GD, AdaGrad, RMSProp, Adam

Welcome to lecture 5 of the course on Deep Learning. So, today we look at some variants of gradient descent. So, we will just quickly do a recap of gradient descent and then look at some variants of it, or some ways of improving it, which is momentum based gradient descent, Nesterov of accelerated gradient descent, stochastic gradient descent, AdaGrad RMSProp and Adam.

So, just to set the context. So, we started with this gradient descent algorithm for a single sigmoid neuron, and then we saw how to extend to network of neurons with back propagation. So, we realized that all we need is the gradients or the partial derivatives, with respect to all the weights and biases. Once we compute that we can just use the gradient descent update rule.

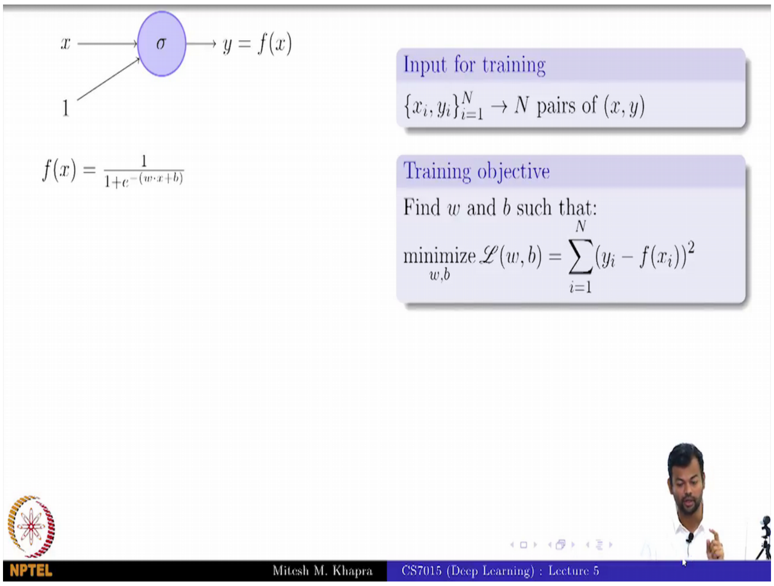
Now, today what we are going to see is, are there better update rules which lead to faster conversion or better performance in various ways. So, that is why we are going to look at all these different variants or methods of improving on gradient descent ok. So, that is the context.

(Refer Slide Time: 01:18)



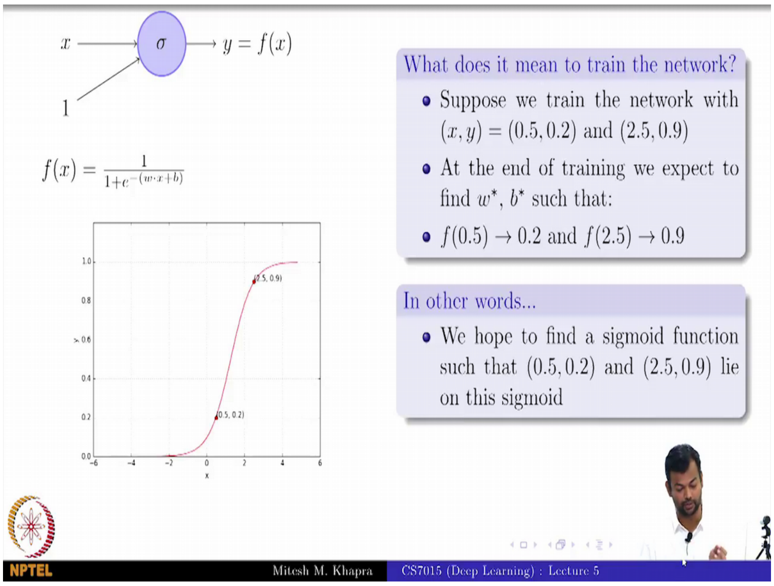
I will just quickly rush through. So, for most of the lecture, I have borrowed ideas from the videos by Ryan Harris on visualize back propagation and some content is based on this course by Andrej Karpathy and others, when I talk about some tips for learning rate and so on. So, you can just look at those also. So, we will just quickly rush through the first two modules which we have already done.

(Refer Slide Time: 01:46)

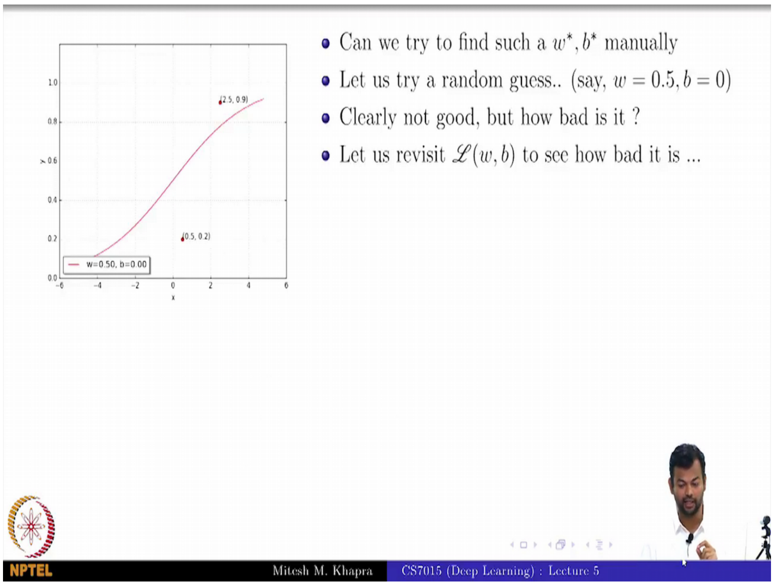


Which was, we were interested in learning the weights and biases for this very toy network, with just 1 input and 1 output, and we started by doing something known as guesswork where we were just trying to adjust these weights and biases by hand.

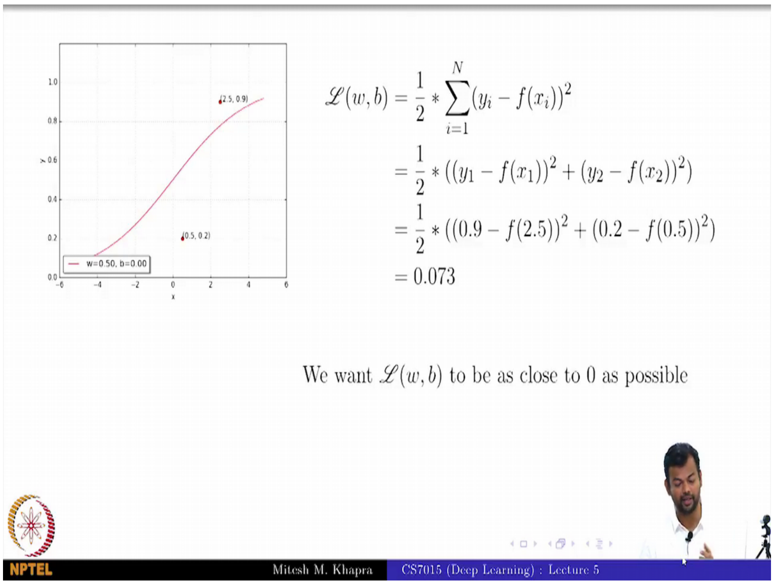
(Refer Slide Time: 01:52)



(Refer Slide Time: 01:56)

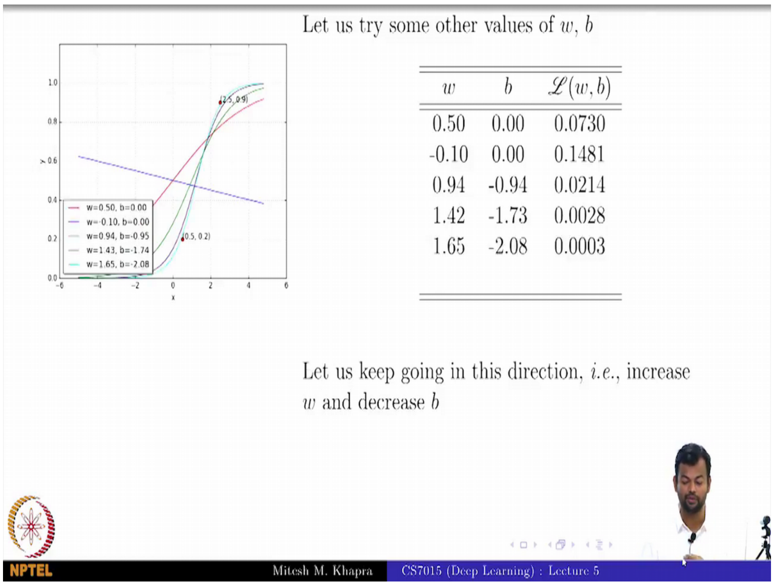


(Refer Slide Time: 01:58)

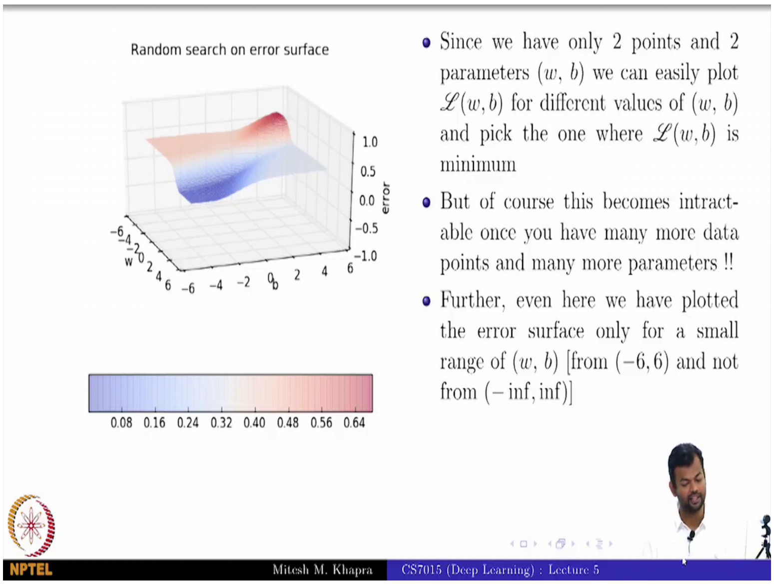


And we realized that its clearly not good and, but we still try to do a very smart guess work, where we were driven by this loss function, which was telling us whether this guess, the current guess is better than the previous guess or not. And we just kept following our guess work and try to reach to some solution, and for this toy network it was very easy to do that.

(Refer Slide Time: 02:06)

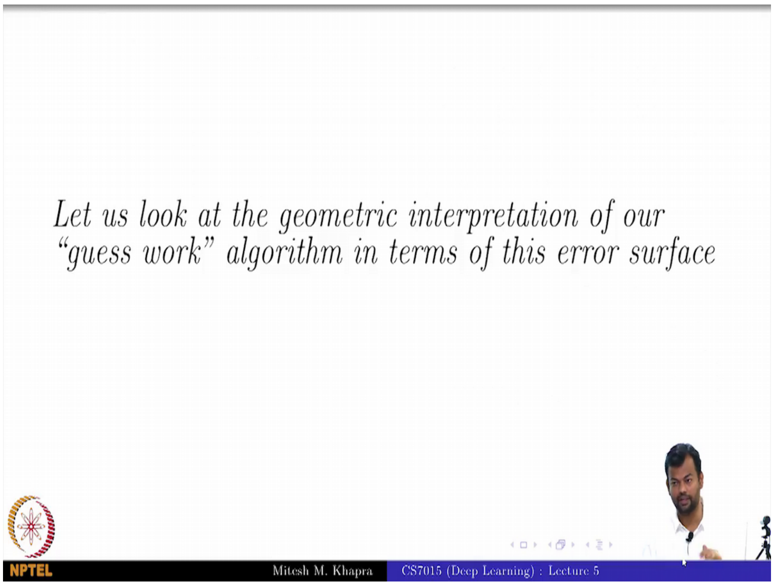


(Refer Slide Time: 02:17)



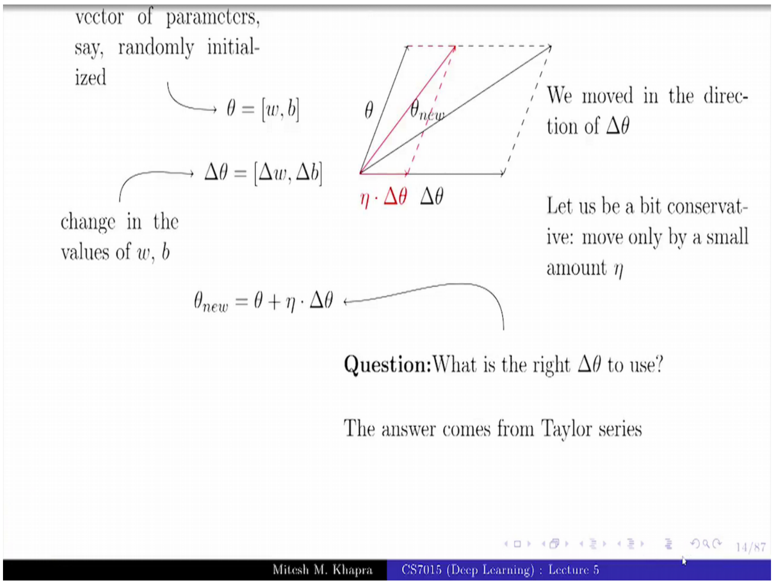
And what we were actually doing is, there is this error surface which exists, which can be plotted for all possible values of w comma b. And what we were trying to do with this guesswork is, trying to find path over the cellar surface, so that we enter into the better regions. So, red is bad, blue is good; the darker the shade of blue the better. And this of course, becomes intractable when you have many parameters and so on.

(Refer Slide Time: 02:38)



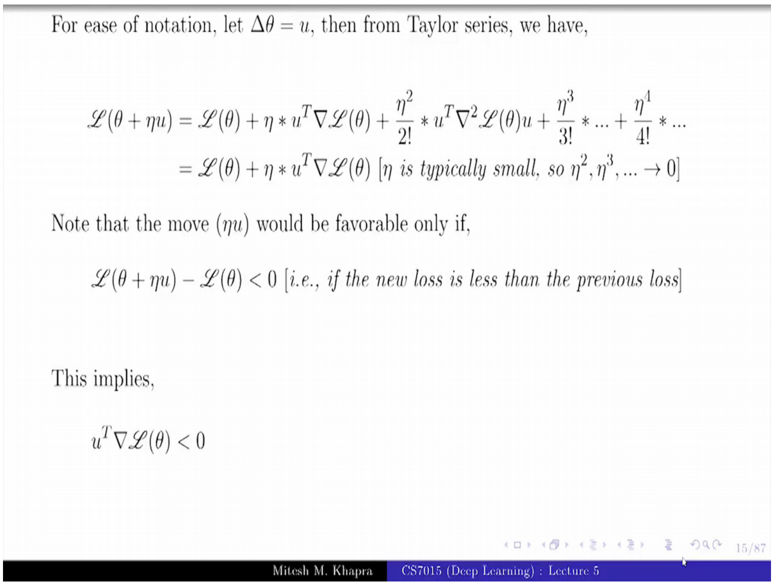
So, we wanted to have a better way of navigating the error surface. So, this is exactly what we were doing with the guesswork algorithm.

(Refer Slide Time: 02:48)



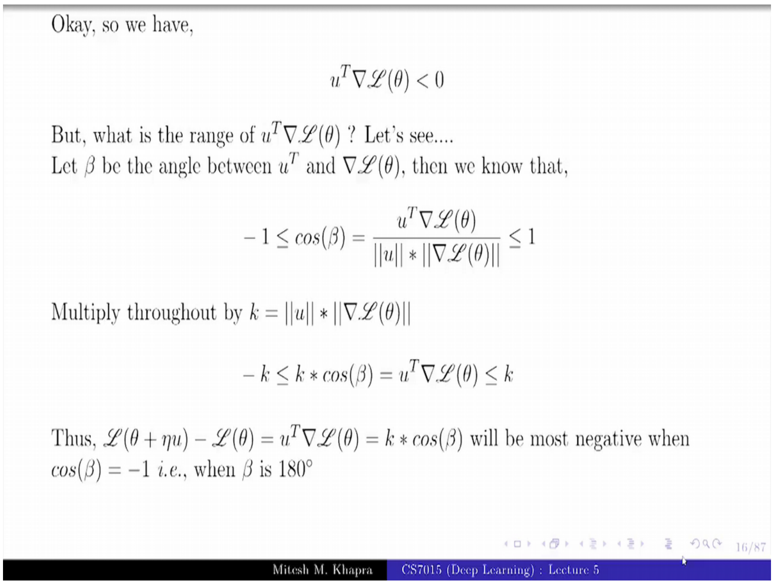
So, then this better way actually we realized that we could arrive at it from a very principled solution from, starting from Taylor series.

(Refer Slide Time: 02:55)

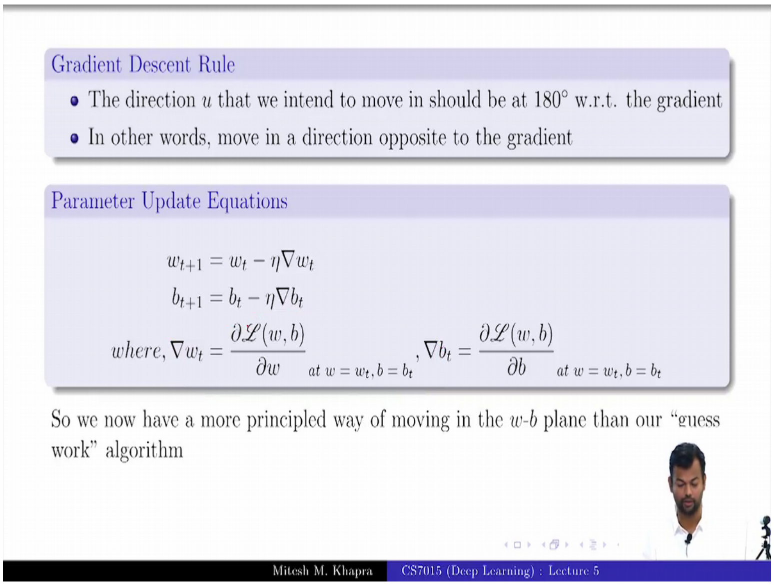


And we went to this derivative, where we finally came up with this rule that move in the direction opposite to the gradient.

(Refer Slide Time: 02:58)

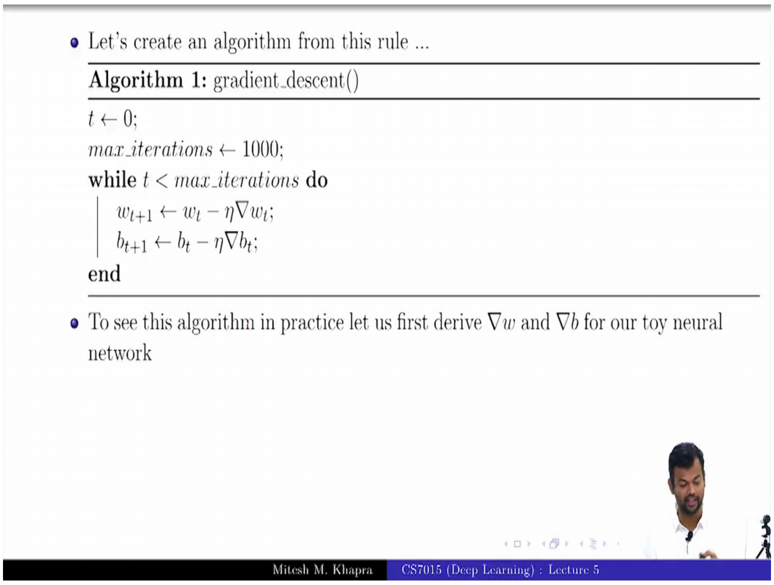


(Refer Slide Time: 03:03)



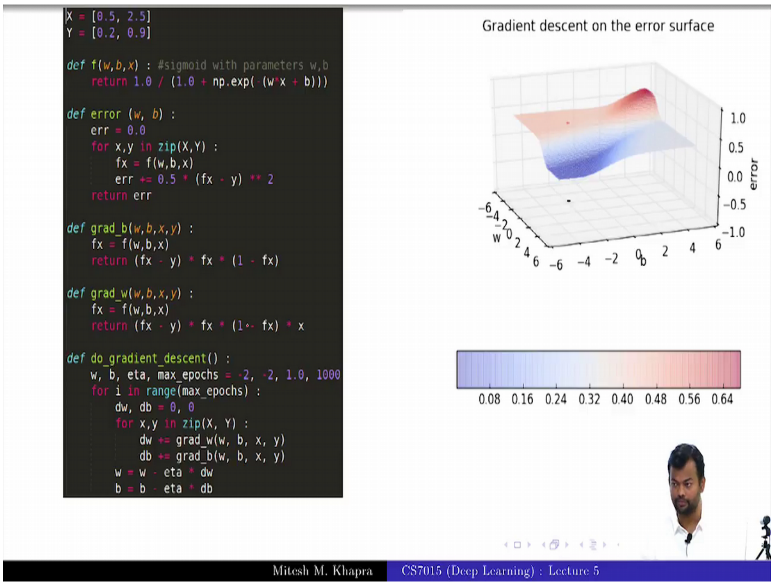
So, that is the rule that we have been sticking to since then. And we also along the way realize some of these things which we defined carefully which was, what is, what exactly this quantity means, which is the partial derivative with respect to w evaluated at a particular weight comma bias configuration. And because this is an iterative process, you are at a certain value of weight and bias and you need to change it from there.

(Refer Slide Time: 03:30)



And we then created an algorithm out of this and when we ran this, we actually derived the full derivative and so on.

(Refer Slide Time: 03:38)

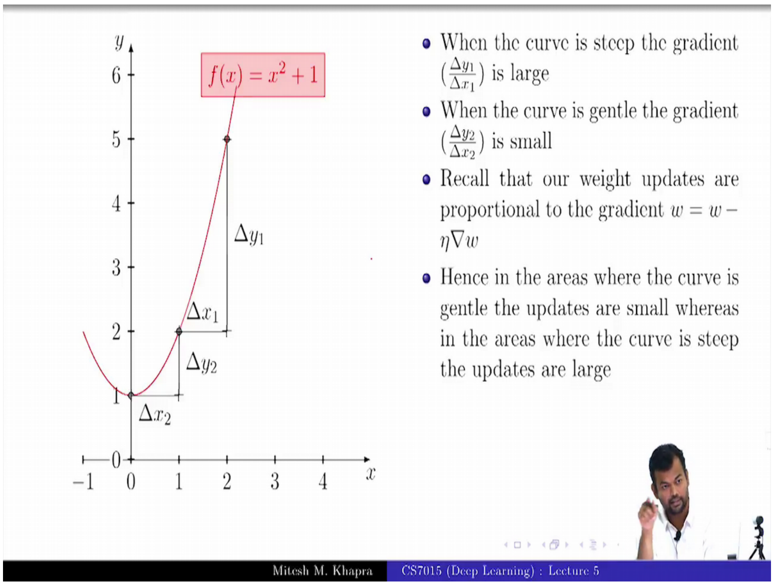


And then when we finally, ran this algorithm. So, this is where, now I will slow down. So, when we ran this algorithm. So, let us see what was happening here right. So, I will just start the algorithm from the beginning.

So, we are now going to run this code and you tell me something that you observe ok. So, I am just clicking. So, there is no change in the phase at which I am clicking this right. So, every click of this is one time step and I am just continuously clicking this I will start now, do you observe something [FL] ok. Do you observe something?

It was initially slow then suddenly picked up and then it again became slow. Why did this happen? The slope is small why ok. How many of you completely understand why this slow and fast moment was there, please raise your hands good. So, that is what we will focus on now right. So, we will try to see this.

(Refer Slide Time: 04:30)



So, we will, I hope this has been fixed ok. So, let us take a simple function which is f of x equal to x square plus 1 right, this is how it will look like. Now in these portions of the curve, the curve is actually very steep right and in these portions the curve is a bit gentle and of course, it becomes very gentle over here right. All of you can see the pen marks properly.

So, now let us see what this means; this steep and fast and small. So, let us look at a region which is steep ok. Now what I am going to do is, I am going to change my x by 1, I move my x from 1 to 2. How much did my y change. All you need to do is just substitute in this formula right for 2 it evaluates to 5, for 1 it evaluates to 2. So, when you move from 1 to 2, your function changed from 2 to 5. So, there is a large change in the function for 1 unit change in your value of x, everyone sees that.

Now, let me do the same at a gentle portion of the curve, I will do it here. Now when I changed the x by 1 unit, again 1 unit right, it is the same change which I did earlier. I changed from zero to 1, how much did my y change.

Student: 1.

1. Now actually what is this quantity; delta y 1 by delta x 1.

Student: Slope.

It is the slope, it is the derivative at that point. So, what are you inferring from this. What happens to the derivative when you are at steep slopes.

Student: It is high.

Derivative is high, because the change in y is much faster than the change in x. What happens to the derivative when you are at the gentle slopes.

Student: Smaller.

Smaller, because the change in y is small or relatively smaller as compared to the change in x or it could also missing, but just these two are relatively different, is what I am trying to impress upon right. And so; that means, the derivatives at the steep slopes are larger in magnitude, whereas, for the gentle slopes they are smaller in magnitude.

Now, can you relate it to the observation that you had on the previous slide. When we were at the plateau it was a very dash slope, gentle slope what would the derivatives be

Student: Small.

Small now what are our updates, you have w is equal to w minus the derivative. Now the derivative is small what will happen to the updates.

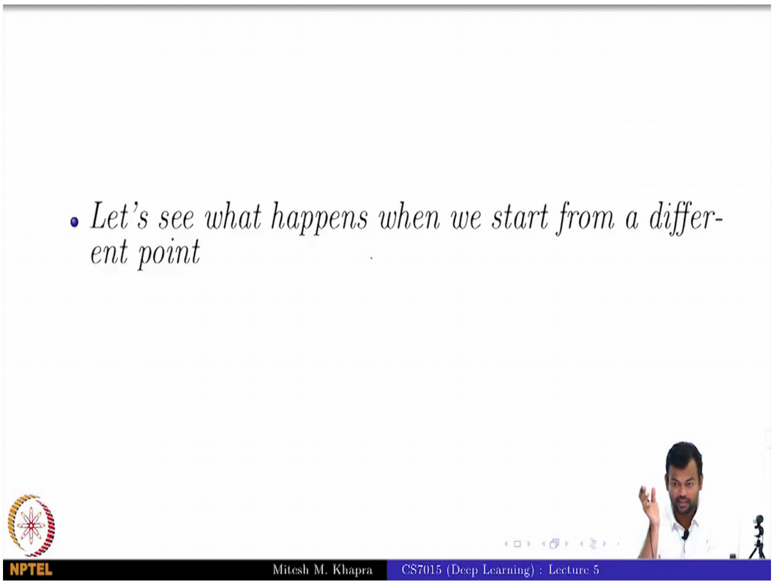
Student: Small.

They will be small. What would happen if the derivative is large.

Student: The updates would be large.

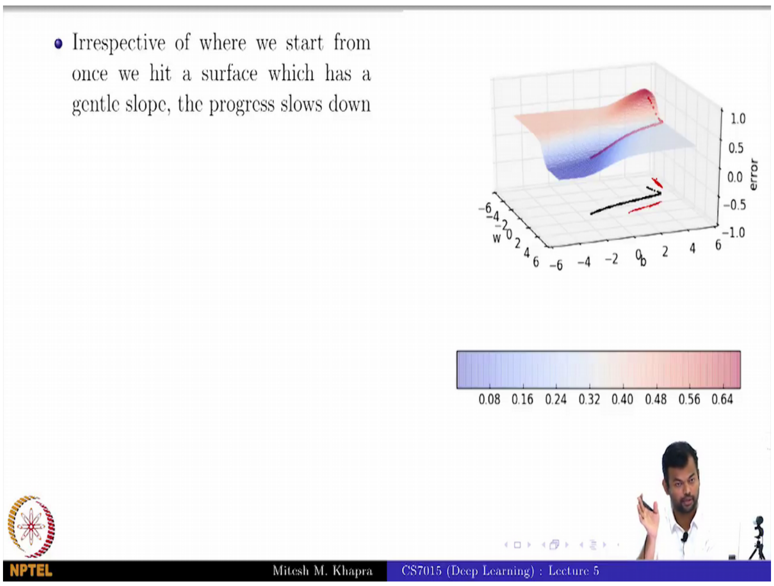
The updates would be large. Therefore, in the gentle areas you are moving slowly and in the steep areas you are moving fast ok. You get this picture very clearly. Now this is going to be the basis of a lot of things that we do today. So, it is very essential to that you understand this perfectly ok. All of you get this properly; good

(Refer Slide Time: 07:32)



Now, now you might say that this was only that special point again and I always get those questions. So, let us see what happens, if you start from a different point.

(Refer Slide Time: 07:40)



So, now again the same gradient descent algorithm I am going to run, but instead of starting at this point which was my random initialization, I just happened to choose a very different random initialization which is here ok, everyone sees that ok

Now, let us see what happens, what do you expect initially fast movement, because the steep, the slope is a bit steep. Now what would happen? It will become slow because you have entered a gentle slope region and then again fast right. So, and then again it will become slow

So, see in this gentle region right, the changes in w are so small that all your black points are actually indistinguishable from each other, it is almost like a snakes body whereas, in these steep slopes, you can see a large change in the w. You can see gaps between the values of w. So, this is irrespective of where you start from. Gentle means slow movement, steep means fast movement that is the basis.