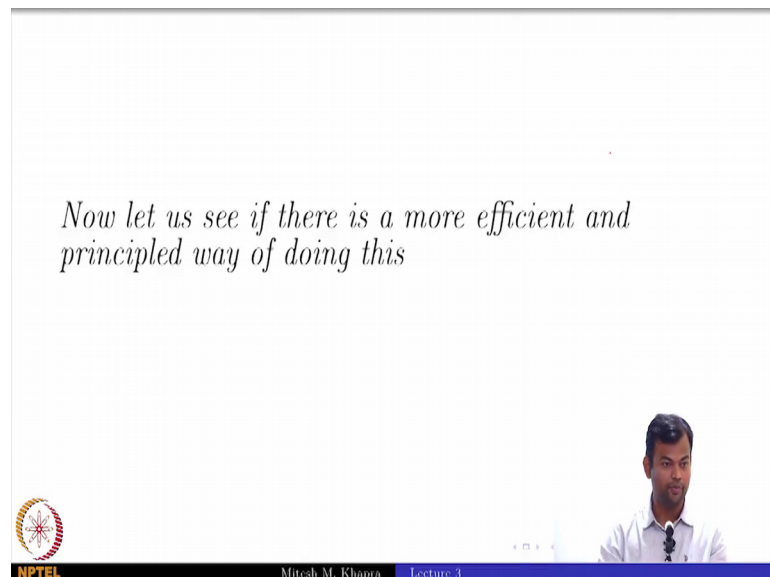


Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 3.4
Lecture - 03
Learning Parameter: Gradient Descent

And, we will move on to the next module where we will talk about Gradient Descent.
How many of you have done gradient descent before? Almost all of you ok.

(Refer Slide Time: 00:22)



So, what we want to do is find a more efficient and principled way of navigating the error surface. Is everyone with me in that?

(Refer Slide Time: 00:30)

Goal

Find a better way of traversing the error surface so that we can reach the minimum value quickly without resorting to brute force search!

The slide features a purple header with the word "Goal" and a light blue box containing the text. At the bottom, there is a video feed of the lecturer, the NPTEL logo, and the text "Mitesh M. Khapra Lecture 3".

And, the goal is to find a better way of doing this, fine.

(Refer Slide Time: 00:35)

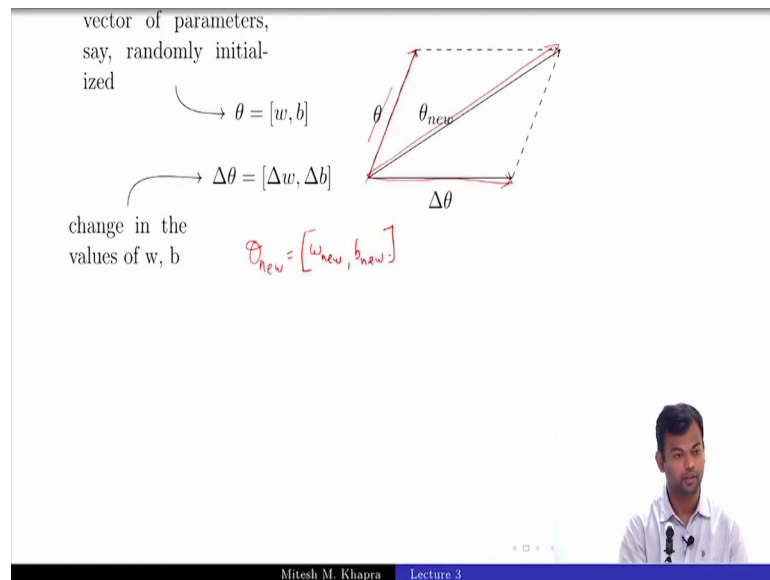
vector of parameters,
say, randomly initial-
ized

$\theta = [w, b]$
 $\in \mathbb{R}^2$

The slide shows handwritten text and a mathematical expression. A red arrow points from the text to the expression. At the bottom, there is a video feed of the lecturer, the NPTEL logo, and the text "Mitesh M. Khapra Lecture 3".

So, let us start by setting up things we will define some notations and some parameters and so on and from there on we will try to come to the algorithm, ok. So, my parameters in this case were w comma b , what I am going to do is I am going to put them into an array or a vector, right and call that vector as θ . So, θ is the vector of parameters and θ belongs to \mathbb{R}^2 ; \mathbb{R} what? \mathbb{R}^2 , right. There are two parameters here. So, it is a two dimensional vector, ok.

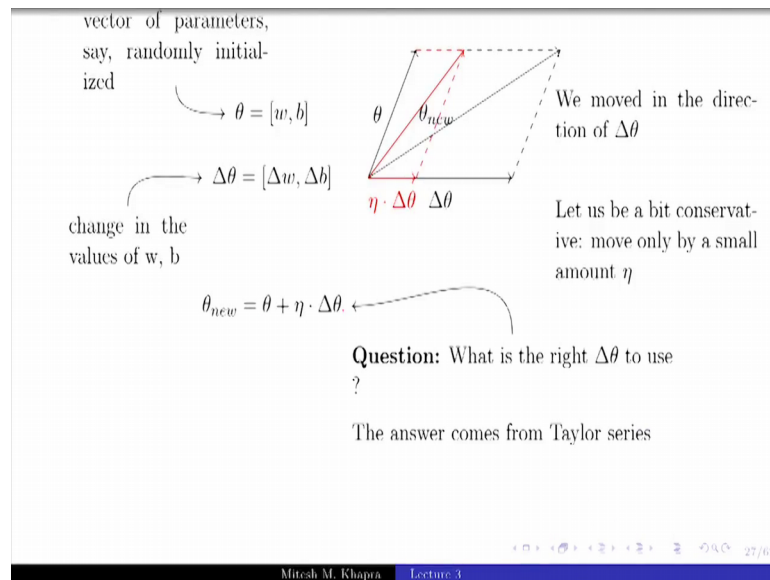
(Refer Slide Time: 01:08)



Now, what I want is again what I will do is, I do not know what the value of w comma b is. So, I started with a random guess. So, that is always going to be my starting point I will always start with a random guess and from there on move on to good values. Now, once I have started with a random guess I want you to tell me some changes that I can make to w and b. So, that I land up in better situations, right that means, I land up in situations where the error is less is that fine.

So, that change in w and b, I am going to call it as delta w and delta b and that again is a vector which is storing these two values, ok. So, this is the picture right I want to take theta and I want to add a small change to it. So, this is my theta, this vector is actually theta right this is the theta vector, I want to add a small change to it which is again a vector right this is delta w comma delta v such that I will get a new value for theta new right. So, theta new would be what actually theta new is equal to w new comma b new, right. Is that fine that is what theta new means, ok?

(Refer Slide Time: 02:17)



Now, what has happened is actually when I have added delta theta to theta I have moved in the direction of delta theta, right I have come from here to here, ok. Now, I am going to be a bit conservative and I am going to say that while I am in moving in the direction of delta theta I do not want to make a giant stride what I will do is I will just move by a small quantity in that direction, right.

So, this delta theta is this large magnitude. So, all I am saying is that I will not I move in that direction I am fine with that, ok, but I do not want to make a giant stride I will just take a small stride in that direction, right. So, eta is a scalar which actually scales down delta theta, ok. So, now, if I am going to take only a small step in that direction instead of this large change I will just get a smaller change theta new, right. So, red the red vector is actually going to be the movement which I make, that is the new value of theta. So, theta new is equal to the original theta plus a small step in the direction of delta theta.

So, everything is clear? You are done we are done with gradient descent? What is missing? What is delta theta, right? I am telling you I want to move in a certain direction, but, what is the right delta theta to use? How many of you know the answer to this? What is the answer? Move in the direction.

Student: (Refer Time: 03:39).

Opposite to the gradient, why; where does that answer come from? Not the ML class folks. How many of you know why we need to move in the direction opposite to the gradient? Why, ok. We will see, ok. So, that is the question that we need to answer. If I give you an answer to this question then what is it I am doing I am giving you a principled way such that you start from a random value of theta move in certain direction and you will ensure that your loss has decreased and then you have to keep doing this right. So, that is the set up and the answer to this comes from Taylor series.

(Refer Slide Time: 04:11)

For ease of notation, let $\Delta\theta = u$, then from Taylor series, we have,

$$\mathcal{L}(\theta + \eta u) = \mathcal{L}(\theta) + \eta * u^T \nabla \mathcal{L}(\theta) + \frac{\eta^2}{2!} * u^T \nabla^2 \mathcal{L}(\theta) u + \frac{\eta^3}{3!} * \dots + \frac{\eta^4}{4!} * \dots$$

$\mathcal{L}(w) = w^2$ $\mathcal{L}(w, b) = w^2 + b^2$
 $\frac{d\mathcal{L}(w)}{dw} = 2w$ $\frac{\partial \mathcal{L}(w, b)}{\partial w} = 2w$ $\begin{bmatrix} \frac{\partial \mathcal{L}(w, b)}{\partial w} \\ \frac{\partial \mathcal{L}(w, b)}{\partial b} \end{bmatrix}$ $\nabla_{\bar{w}, b} \mathcal{L}(\theta)$
 $\frac{\partial \mathcal{L}(w, b)}{\partial b} = 2b$

Mitesh M. Khapra Lecture 3

So, now what I am going to do is I am going to give you the right direction delta theta; fine and for ease of notation I am going to call it as u. So, remember what this delta theta is, what is it? Change in w comma change in b. So, it is a vector in R 2, remember that, I am just going to call it as u. Now, this is what Taylor series tells me what it tells me is that if I am at a certain value of theta and if I want to change that value a bit then what is going to be the new value of the loss function or any function for that point and this is the formula for that, ok. Now, what is let us see; what are some quantities here what is this quantity scalar, vector matrix?

Student: (Refer Time: 05:08).

Scalar, this?

Student: Vector.

Vector, we just did that right, it is a vector. What about this?

Student: (Refer Time: 05:16).

What is this quantity, actually?

Student: Gradient.

Gradient: what is the gradient? What is the gradient? No, you are telling me how to use the gradient, I am asking you what is the gradient you are giving me absolutely correct and absolutely useless definitions.

Student: (Refer Time: 05:36).

That is a very good answer, ok. So, now what I am going to do is I am going to digress a bit, and I am going to tell you something about derivatives partial, derivatives and gradients and then we will come back to this, ok. So, now, suppose you have a function L this is L in my handwriting this function of w , and say this function is w square, ok. Now, what is what is this called? A derivative of the function with respect to w , this is the derivative and you know this is $2w$, ok. Now, suppose I have a function b square, now what is this quantity?

Student: (Refer Time: 06:27).

Is a partial derivative of the function with respect to w ? Why partial?

Student: (Refer Time: 06:32).

Because, it is considering b as a constant and taking the derivative with respect to only one of the variables, right this happens to be and what is this quantity, oh sorry. So, is w comma b , right this is the partial derivative with respect to b , ok.

Now, can you tell me what is a gradient? The gradient is nothing, but it is just these two partial derivatives taken together and put into a vector, right. Now, suppose I had a function which depended on hundred variables, what would the gradient be size of the gradient?

Student: (Refer Time: 07:25).

R 100 it would lie, it would be a hundred dimensional case, ok. So, now can you tell me with this evidence in knowledge, but this primer can you tell me, what this is? This is a gradient vector?

Student: (Refer Time: 07:39).

Which is right there in front of you in a red ink.

Student: (Refer Time: 07:45).

This is what it is, right, fine. Everyone with that? So, actually the right way to write this and probably we need to correct in the slides would be theta. So, remember that theta is equal to w comma b. So, this is the derivative of L theta with respect to theta which is nothing, but the collection of the partial derivatives with respect to the components of theta, is it fine. So, everybody understands what is a derivative partial derivative and gradient, ok, fine. So, now, the gradient is a vector in this case, fine.

(Refer Slide Time: 08:25)

For ease of notation, let $\Delta\theta = u$, then from Taylor series, we have,

$$\mathcal{L}(\theta + \eta u) = \mathcal{L}(\theta) + \eta * u^T \nabla \mathcal{L}(\theta) + \frac{\eta^2}{2!} * u^T \nabla^2 \mathcal{L}(\theta) u + \frac{\eta^3}{3!} * \dots + \frac{\eta^4}{4!} * \dots$$

Handwritten notes on the slide:
 - The term $\mathcal{L}(\theta + \eta u)$ is circled in red and labeled "New".
 - The term $\mathcal{L}(\theta)$ is circled in red and labeled "old".
 - The term $\frac{\eta^2}{2!} * u^T \nabla^2 \mathcal{L}(\theta) u$ is circled in red.
 - A note below the equation says: $[\eta \text{ is typically small, so } \eta^2, \eta^3, \dots \rightarrow 0]$

Below the equation, two gradient vectors are shown:

$$\nabla_{\omega, b} \begin{bmatrix} \frac{\partial \mathcal{L}(\omega, b)}{\partial \omega} \\ \frac{\partial \mathcal{L}(\omega, b)}{\partial b} \end{bmatrix} \quad \begin{bmatrix} \frac{\partial^2 \mathcal{L}(\omega, b)}{\partial \omega^2} & \frac{\partial^2 \mathcal{L}(\omega, b)}{\partial b \partial \omega} \\ \frac{\partial^2 \mathcal{L}(\omega, b)}{\partial b \partial \omega} & \frac{\partial^2 \mathcal{L}(\omega, b)}{\partial b^2} \end{bmatrix}$$

Mitsh M. Khapra | Lecture 3

So, now what is this quantity? It is a?

Student: (Refer Time: 08:32).

No, it is what is this?

Student: The dot product.

The dot product between these two vectors, fine. Now, one last thing and many more things actually. So, what is this, square of the gradient?

Student: (Refer Time: 08:48).

This is not the square of the gradient. What is this? Hessian, fine, everyone knows the textbook what said can you tell me, what does it is a scalar, vector matrix?

Student: Matrix.

Matrix; what is the size of this matrix?

Student: 2 by 2.

2 by 2; what are the elements of this matrix?

Student: (Refer Time: 09:06).

Second order partial derivatives, right. So, it is the gradient of the gradient, right, is that fine. So, what does that mean you had this gradient, this is the gradient. Now, you want to take the gradient of this again with respect to w comma b , right that is what this means, it is a gradient of the gradient right. So, what that means, is we will take the gradient of the first quantity again with respect to w . So, that would be $\frac{\partial^2}{\partial w^2}$ by $\frac{\partial^2}{\partial w^2}$ what would this quantity be?

Student: (Refer Time: 09:58).

What would this be?

Student: (Refer Time: 10:06).

Is that fine and you can fill in this quantity, right? So, now, it is clear, what the hessian is? It is the derivative of the derivative and it would be a matrix, ok, is that clear to everyone. So, I have a habit of doing a lot of these basic stuff I know that the top 20 percent of the class gets really pissed off when I do this, but as a philosophy I teach for the bottom 30 percent of the class.

So, I do not mind that and the other thing is I use slides. So, I do not write a lot of math so, I can cover a lot of material despite doing all this basic stuff, right. So, I am going to

stick to that what I am trying to say is that write this in the feedback that you do not like this basic stuff, but it is just that I am going to ignore that feedback, I mean just being honest, right. So, I like doing this because it just takes me 10 minutes to do this and for the rest of the class I do not have to look at blank faces afterwards, right. So, it really helps me a lot fine. So, is that all clear all the quantities here are clear?

So, now, so, this is the gradient this is the Hessian and now eta, remember what did we say about eta?

Student: (Refer Time: 11:19).

It is a small quantity and what do we do with small quantities always in maths?

Student: (Refer Time: 11:25).

We ignore them. So, once we take their powers you are always ignore them. Whether it is correct or not who cares I mean someone has told it, it is good to ignore. So, we will ignore it, right. So, now, all these higher order terms we can ignore right; that means, I will only consider this, fine.

So, let us again look at what the setup is? The setup is that I have some value of theta I want to move away from that value such that what do you say about this loss compared to this loss? I will call this the new loss and I will call this the old loss. What is the relation between them?

Student: (Refer Time: 12:09).

The new law should be?

Student: (Refer Time: 12:13).

Less than; so if I or someone gives you a u , I am not getting ok, someone gives you this u then what does what when would you say it is a good u ?

(Refer Slide Time: 12:28)

For ease of notation, let $\Delta\theta = u$, then from Taylor series, we have,

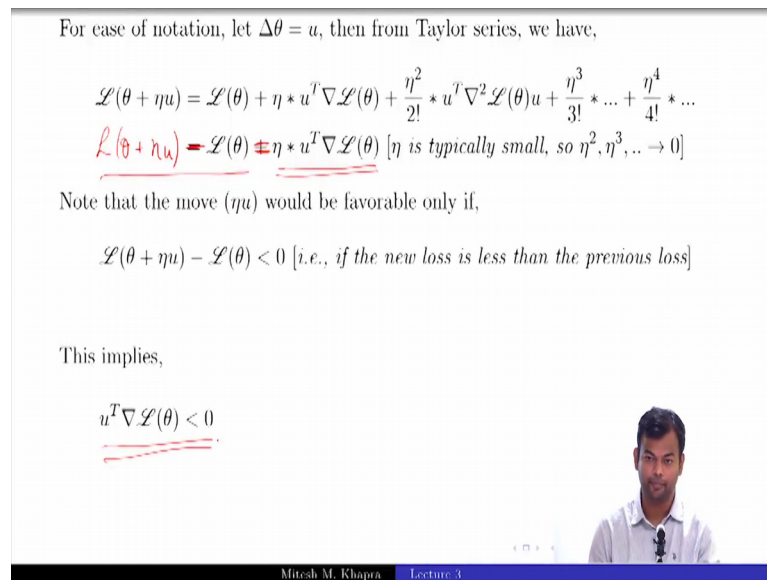
$$\mathcal{L}(\theta + \eta u) = \mathcal{L}(\theta) + \eta * u^T \nabla \mathcal{L}(\theta) + \frac{\eta^2}{2!} * u^T \nabla^2 \mathcal{L}(\theta) u + \frac{\eta^3}{3!} * \dots + \frac{\eta^4}{4!} * \dots$$

$\mathcal{L}(\theta + \eta u) \approx \mathcal{L}(\theta) + \eta * u^T \nabla \mathcal{L}(\theta)$ [η is typically small, so $\eta^2, \eta^3, \dots \rightarrow 0$]

Note that the move (ηu) would be favorable only if,

$$\mathcal{L}(\theta + \eta u) - \mathcal{L}(\theta) < 0 \text{ [i.e., if the new loss is less than the previous loss]}$$

This implies,

$$u^T \nabla \mathcal{L}(\theta) < 0$$


Mitsh M. Khapra Lecture 3

If this condition holds everyone agrees with that, right. So, I have found a good direction to move in if this condition holds. Now, this condition actually implies that this condition should hold right this is $\mathcal{L}(\theta + \eta u)$, right. So, if I just do minus here I get this, right. So, this quantity which should be less than equal to 0, implies that this quantity should be less than equal to 0 and remember η is a positive constant, ok. Why cannot it be negative?

Student: (Refer Time: 13:09).

Why? Because you wanted to take a small step in that direction, if we make it negative we will do what?

Student: (Refer Time: 13:16).

We will reverse the direction; we do not want that as of now, right. So, η is that for a positive quantity. So, that means, this quantity should be less than 0, is it fine with everyone?

(Refer Slide Time: 13:29)

Okay, so we have,

$$u^T \nabla \mathcal{L}(\theta) < 0$$

But, what is the range of $u^T \nabla \mathcal{L}(\theta)$? Let us see...

Let β be the angle between u^T and $\nabla \mathcal{L}(\theta)$, then we know that,

$$-1 \leq \cos(\beta) = \frac{u^T \nabla \mathcal{L}(\theta)}{\|u\| * \|\nabla \mathcal{L}(\theta)\|} \leq 1$$

multiply throughout by $k = \|u\| * \|\nabla \mathcal{L}(\theta)\|$

$$-k \leq k * \cos(\beta) = u^T \nabla \mathcal{L}(\theta) \leq k$$

Thus, $\mathcal{L}(\theta + \eta u) - \mathcal{L}(\theta) = u^T \nabla \mathcal{L}(\theta) = k * \cos(\beta)$ will be $\cos(\beta) = -1$ i.e., when β is 180°

Mitesh M. Khapra Lecture 3

Ok. So, so far after all this story what we are left with is this condition should hold for the u that I am trying to choose, so that I can be sure that I have chosen the correct u , right and the definition of correct u is that the loss at the previous step the loss of the new step should be less than the loss at the previous step, is that fine? Ok. So, that is what we have arrived at.

Now, what is the range of this quantity, that is why I asked you what is this? This is a?

Student: Dot product.

Dot product, I will leave it at that. So, now, you tell me what is the range of this, people from the ML class cannot answer. Did I cover this in the ML class? No? Ok, fine. What is the range of this? Not a very hard question.

Student: (Refer Time: 14:22).

Plus or minus?

Student: (Refer Time: 14:25) mod of u (Refer Time: 14:26).

Very good; how many of you understood that answer? He said plus or minus mod of u into mod of gradient the gradient vector, right why is it so, easy.

Let β be the angle between u^T and this between sorry it should not be u transpose between u and the gradient then we know that this condition holds, $\cos \beta$ is given by this quantity and we know that $\cos \beta$ lies between minus 1 and 1, ok. Now, if I just say that this quantity is equal to k then I can just get this condition, ok.

Now, let us see what are we trying to do? We are trying to find a u such that this quantity is negative, ok. Ok we are trying to find the use such that this quantity is negative, now I just stop at negative we would like to make it as negative as possible, right because the more than negative it is the more will be the decrease in my loss function right because this quantity tell me tells me how much my loss decreases. So, the more the negative it is the more the loss will decrease. So, let me make it as negative as possible.

Now, what is that value? When will that happen? When α is you know the answer you started with the answer.

Student: (Refer Time: 15:54).

No, what is that one phrase which you have marked up move in the direction ?

Student: (Refer Time: 16:07).

Now, think of that.

Student: (Refer Time: 16:09).

What would happen when this is the most negative it can be, what would the angle be?

Student: (Refer Time: 16:19).

180 degrees; how many of you get that because when this is the most negative; that means, the $\cos \beta$ is actually minus 1 and when is $\cos \beta$ minus 1 when the angle is 180 degrees; that means, u should be such that it is at 180 degrees to the gradient, hence repeat the phrase.

Student: (Refer Time: 16:42).

Move in a direction opposite to the gradient. Is that fine? Everyone gets it now? Why you need to move in the direction opposite to the gradient, ok, fine.

(Refer Slide Time: 16:53)

Gradient Descent Rule

- The direction u that we intend to move in should be at 180° w.r.t. the gradient
- In other words, move in a direction opposite to the gradient

Parameter Update Equations

$$\theta_{t+1} = \begin{cases} w_{t+1} \\ b_{t+1} \end{cases} = \begin{cases} w_t \\ b_t \end{cases} - \eta \begin{cases} \nabla w_t \\ \nabla b_t \end{cases}$$

where, $\nabla w_t = \frac{\partial \mathcal{L}(w, b)}{\partial w}$ at $w = w_t, b = b_t$, $\nabla b = \frac{\partial \mathcal{L}(w, b)}{\partial b}$ at $w = w_t, b = b_t$

$$\theta_{t+1} = \theta_t + \eta(-\nabla \mathcal{L}(\theta))$$

Mitesh M. Khapra Lecture 3 30/62

So, this is what the gradient descent rule is. You are at a particular value of theta you want to move to a new value of theta such that your new loss is less than the current loss, what gradient descent tells you is move in a direction opposite to the gradient, ok. So, are you fine with this? Now, with gradients I have come to scalars, but I will just explain what I have written here.

So, this quantity is nothing, but theta t plus 1, right, is equal to theta t, right and what is this, right. So, the new theta is equal to the current theta minus, why because we want to move in the direction opposite. So, it is basically theta t plus 1 is equal to theta t plus eta into a negative direction, right the direction negative to the gradient hence you get that minus 1 ok. So, is that clear?

Now, what are these quantities let me just take that carefully. So, this quantity is gradient of the loss function with respect to w, sorry the partial derivative of the loss function with respect to w evaluated at w is equal to w t and b equal to b t, what does that mean. So, remember when you are dealing with derivatives as always a formula and then a value add that at a particular value. So, what is the derivative of x square with respect to what does not matter 2x, ok.

So, derivative of x square with respect to x is 2x. What is the value of this derivative at x equal to 1? 2, right. So, you see the difference you have a formula which is 2x, now you substitute in a particular value and you get the value at that particular value, ok. So, that

is what this means because you are already at w_t comma b_t , now you cannot subtract a formula from here right you have to put subtract a value. So, you know what the formula is you plug in the values of w_t comma b_t get that value and subtract it from your current w_t , is that fine? So, everyone completely understands what is the gradient descent rule is, fine, good.

(Refer Slide Time: 19:14)

Gradient Descent Rule

- The direction u that we intend to move in should be at 180° w.r.t. the gradient
- In other words, move in a direction opposite to the gradient

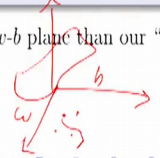
Parameter Update Equations

$$w_{t+1} = w_t - \eta \nabla w_t$$

$$b_{t+1} = b_t - \eta \nabla b_t$$

where, $\nabla w_t = \frac{\partial \mathcal{L}(w, b)}{\partial w}$ at $w = w_t, b = b_t$, $\nabla b = \frac{\partial \mathcal{L}(w, b)}{\partial b}$ at $w = w_t, b = b_t$

So we now have a more principled way of moving in the w - b plane than our "guess work" algorithm



30/62

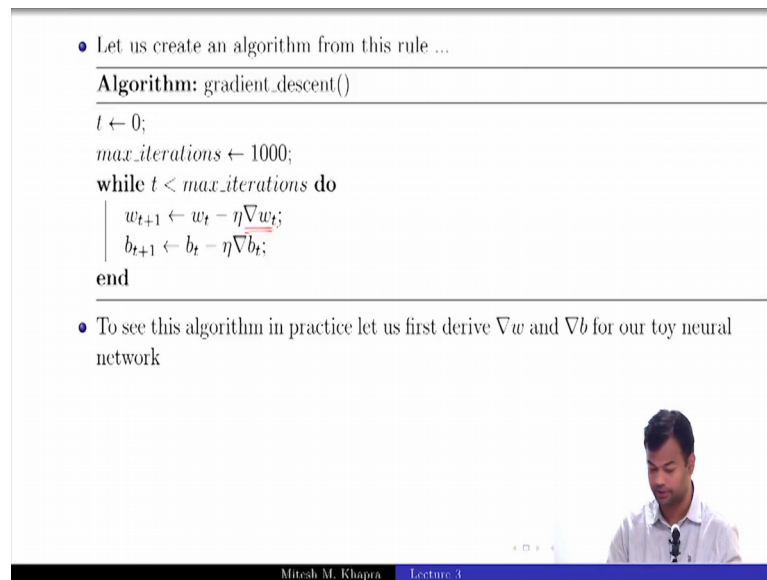
So, now we have a more principled way of moving in the w b plane. What do I mean by that? Remember this was our w b plane, this was our error, this is something what our error surface looked like, it was this flying carpet. I was randomly moving on the w b plane earlier, right and trying to guess what the errors or trying to compute the error and then settle for a particular value. Now, I have a more principled way of moving in the w b plane. I know what is the next step based on the current step I just need to move in the direction opposite to the gradient, ok.

So, let us try to. So, this is what it tells me for one step, but I need to keep doing this till, what is that golden word?

Student: (Refer Time: 19:58).

Convergence, right; I have to keep doing this till convergence, ok.

(Refer Slide Time: 20:02)



- Let us create an algorithm from this rule ...

Algorithm: gradient_descent()

```
t ← 0;
max_iterations ← 1000;
while t < max_iterations do
  | w_{t+1} ← w_t - η∇w_t;
  | b_{t+1} ← b_t - η∇b_t;
end
```

- To see this algorithm in practice let us first derive ∇w and ∇b for our toy neural network

Mitesh M. Khapra Lecture 3

So, let us create an algorithm out of this rule I will start a time step 0, I will do this for some max iterations; instead of saying till convergence I will do it for some iterations. At every iteration I will this is how I will update my weights. I will take the current weights, subtract the gradient from that and get the new state I mean not subtract the gradient subtract this quantity and get the new weights. So, now, is everything clear? Is the gradient descent algorithm done? Can you do it for the toy network which I had? Is there something still missing?

Student: (Refer Time: 20:38).

It has fine we will take a small value 0.01 or something. Actually, not told you what these are, right I means to write it you know these are derivatives, but what is this actually, ok. So, let us see that now. So, that is what we are going to see next, I am really going slow today, fine.

(Refer Slide Time: 21:00)

The slide contains the following elements:

- A diagram showing an input x and a bias 1 entering a blue circle labeled σ , which outputs $y = f(x)$.
- The sigmoid function formula: $f(x) = \frac{1}{1 + e^{-(wx + b)}}$
- A graph of the sigmoid function with two data points plotted: $(2.5, 0.9)$ and $(0.5, 0.2)$.
- Handwritten red notes: ∇_w , ∇_b , x_1, y_1 , x_2, y_2 , and x, y .
- Text: "Let's assume there is only 1 point to fit (x, y) ".
- A small video inset of the lecturer in the bottom right corner.
- A footer bar with the text "Mitesh M. Khapra Lecture 3".

Do you guys need a break? You are really in a hurry to wind up the lecture and go. Are you sure you do not need a break? Ok, I do not know whether to be happy about it or sad about it I will be happy, ok. I will assume that you are enjoying the lecture so much that you do not want a break, ok.

So, now we want to find out we are in the car quest is for this delta sorry the partial derivative with respect to w and partial derivative with respect to b , that is the thing which we had plugged in the formula, but we do not know what that is, right. So, we need to find that out. So, now, for simplicity let us assume there is only one point of it which is x comma y , right. So, earlier we had this x_1 y_1 and x_2 y_2 . Now, I am just assuming there is only one point which is x comma y .

(Refer Slide Time: 21:50)

Diagram: $x \rightarrow \sigma \rightarrow y = f(x)$

Equation: $f(x) = \frac{1}{1 + e^{-(wx+b)}}$

Graph: A plot of the sigmoid function with two data points: $(2, 0.9)$ and $(5, 0.2)$.

Text: Let's assume there is only 1 point to fit (x, y)

Equation: $\mathcal{L}(w, b) = \frac{1}{2} * (f(x) - y)^2$

Equation: $\nabla w = \frac{\partial \mathcal{L}(w, b)}{\partial w} = \frac{\partial}{\partial w} \left[\frac{1}{2} * (f(x) - y)^2 \right]$

Footer: Mitesh M. Khapra Lecture 3

So, now, what is a loss function earlier I had this summation over I equal to 1 to 2, but I have just one x comma y . So, I will just use that this is what my loss function and what are the quantities that I am interested in finding? One is the partial derivative of this loss function with respect to w , ok.

(Refer Slide Time: 22:11)

$\begin{aligned} \nabla w &= \frac{\partial}{\partial w} \left[\frac{1}{2} * (f(x) - y)^2 \right] \\ &= \frac{1}{2} * [2 * (f(x) - y) * \frac{\partial}{\partial w} (f(x) - y)] \\ &= (f(x) - y) * \frac{\partial}{\partial w} (f(x)) \\ &= (f(x) - y) * \frac{\partial}{\partial w} \left(\frac{1}{1 + e^{-(wx+b)}} \right) \\ &= (f(x) - y) * f(x) * (1 - f(x)) * x \end{aligned}$	$\begin{aligned} &\frac{\partial}{\partial w} \left(\frac{1}{1 + e^{-(wx+b)}} \right) \\ &= \frac{-1}{(1 + e^{-(wx+b)})^2} \frac{\partial}{\partial w} (e^{-(wx+b)}) \\ &= \frac{-1}{(1 + e^{-(wx+b)})^2} * (e^{-(wx+b)}) \frac{\partial}{\partial w} (-(wx+b)) \\ &= \frac{-1}{(1 + e^{-(wx+b)})} * \frac{e^{-(wx+b)}}{(1 + e^{-(wx+b)})} * (-x) \\ &= \frac{1}{(1 + e^{-(wx+b)})} * \frac{e^{-(wx+b)}}{(1 + e^{-(wx+b)})} * (x) \\ &= f(x) * (1 - f(x)) * x \end{aligned}$
--	---

Footer: Mitesh M. Khapra Lecture 3 33/62

So, let us do this lets actually derive this. So, this is what it looks like now you have to help me in deriving this what will I do first?

Student: (Refer Time: 22:18).

Tell me the next step.

Student: (Refer Time: 22:23).

2 into f of x minus y and push the gradient inside, of course the derivative is that fine? Anyone who has a problem with this? Next, y is a constant this is the true I remember. So, that is why this is a constant is not the predicted y , ok.

Now, this quantity, what is f of x actually?

Student: (Refer Time: 22:51).

Sigmoid function, right; so I will just write it. Now, this is the quantity that I need the derivative for. So, I will just write it here. What is the next step? This is of the form 1 over x . So, what will it be?

Student: (Refer Time: 23:04).

Minus 1 over x square and then you put the derivative and say it. Is that fine? Now, the quantity inside is of the form e raise to x . So, the derivative is e raise to x and you push it inside, is this fine? So, this on slide should come both these are coming together. So, is that fine? Ok. Now, what is this actually?

Student: f of x .

F of x ; what is this?

Student: (Refer Time: 23:32).

This is actually 1 minus f of x , just take my word for it for now, you can go home and work it out, right. So, this actually if you do 1 minus this and do some trickery you will get to this quantity, right. So, what you get is a very simple formula f of x into 1 minus f of x into x , I am going to substitute back here. So, now I exactly know what the partial derivative of w is, fine.

(Refer Slide Time: 23:57)

Diagram: $x \rightarrow \sigma \rightarrow y = f(x)$

Equation: $f(x) = \frac{1}{1+e^{-(w*x+b)}}$

Graph points: $(2.5, 0.9)$ and $(0.5, 0.2)$

Equation: $\frac{\partial}{\partial w} \sum_{i=1}^2 (f(x_i) - y_i)^2$

Text: So if there is only 1 point (x, y) , we have,

Equation: $\nabla w = (f(x) - y) * f(x) * (1 - f(x)) * x$

Footer: Mitesh M. Khapra Lecture 3

So, there is only one point, then this is what the partial derivative with respect to w is going to be of the loss function, right. If there were two points what would happen? If there were two points my loss function was this is a sum of two elements. And I am taking some derivative of a sum, I will get a sum of derivatives, right.

(Refer Slide Time: 24:23)

Diagram: $x \rightarrow \sigma \rightarrow y = f(x)$

Equation: $f(x) = \frac{1}{1+e^{-(w*x+b)}}$

Graph points: $(2.5, 0.9)$ and $(0.5, 0.2)$

Equation: $\frac{\partial}{\partial w} \sum_{i=1}^2 (f(x_i) - y_i)^2$

Text: So if there is only 1 point (x, y) , we have,

Equation: $\nabla w = (f(x) - y) * f(x) * (1 - f(x)) * x$

Text: For two points,

Equation: $\nabla w = \sum_{i=1}^2 (f(x_i) - y_i) * f(x_i) * (1 - f(x_i)) * x_i$

Equation: $\nabla b = \sum_{i=1}^2 (f(x_i) - y_i) * f(x_i) * (1 - f(x_i))$

Footer: Mitesh M. Khapra Lecture 3

So, how many of you will not cringe if I say this is the answer? Anyone who has a problem with this? You get this, how many if you do not get this? How many of you get

this? Good, fine. Now, can you do a similar thing for b? Can you tell me the answer without actually deriving it?

Student: (Refer Time: 24:41).

I can perfectly understand what you are saying.

Student: (Refer Time: 24:48).

x would not be there, right because this last x that you see here came because w into x was there, but b we are not multiplying x. So, what we will get is this. You can go home and check, ok.

(Refer Slide Time: 25:12)

The slide contains the following Python code on the left:

```
X = [0.5, 2.5]
Y = [0.2, 0.9]

def f(w,b,x) : #sigmoid with parameters w,b
    return 1.0 / (1.0 + np.exp(-(w*x + b)))

def error(w, b) :
    err = 0.0
    for x,y in zip(X,Y) :
        fx = f(w,b,x)
        err += 0.5 * (fx - y) ** 2
    return err

def grad_b(w,b,x,y) :
    fx = f(w,b,x)
    return (fx - y) * fx * (1 - fx)

def grad_w(w,b,x,y) :
    fx = f(w,b,x)
    return (fx - y) * fx * (1 - fx) * x

def do_gradient_descent() :
    w, b, eta, max_epochs = -2, -2, 1.0, 1000
    for i in range(max_epochs) :
        dw, db = 0, 0
        for x,y in zip(X, Y) :
            dw += grad_w(w, b, x, y)
            db += grad_b(w, b, x, y)
        w = w - eta * dw
        b = b - eta * db
```

On the right, there is a 3D plot titled "Random search on error surface" showing a surface of error as a function of weights w and bias b. The axes range from -6 to 6. Below the plot is a color bar for the weight w, ranging from 0.08 to 0.64. A red circle highlights the color bar with the handwritten equation $\Delta w = \sum_{i=1}^2$.

So, now we have everything that we need. Now, we actually have everything that we need, ok, no more trick questions. So, now, we will write code to do this, ok, we will actually implement the code and see what happens. So, these are the two data points that I had 0.5 comma 0.2 and 2.5 comma 0.9. The first thing which I need is something which can implement the sigmoid function. So, this is 1 over one plus e raise to minus w x plus b is that fine, ok.

Now, I need something which can compute the error. So, this is summation of half into f of x minus y the whole square, I go over all the data points summation of half into f of x minus y the whole square, is that fine? Now, what I will do is I will take this try out a lot

of values of w and b and plot the error surface ok, but this is only for illustration, in practice I will not do this. We just know that this error surface exist I just want to verify that whatever algorithm I come up with does not efficient navigation of this error surface, that is what I want to verify, that is why I am plotting this, ok.

Next time you need a function which can compute grad of b . We just saw this on the previous slide this is $f(x) = 1 - f(x)$, right, simple. Everyone is fine with this? Then I need a function which can compute the grad with respect to w same thing except that I have this x at the end. So, I have all the ingredients in place. Now, what would I do, what is the next thing that I will write? The main loop right, I will write the main loop now ok.

So, this is what the main loop look like looks like. I start with some random initialize for w and b . Remember that our initial theta which is composed of w and b is going to be some random guess. So, I started with the random guess which is $[-2, 2]$, I have chosen η to be 1; that means, I am not going to be conservative I am going to move in the direction of the gradient, ok. If I chosen at 0.01 and 0.001 I would have been conservative and I am going to run this till 1000 epochs which is my notion of conversions, ok.

Now, in each epoch what I am doing is for every data point, so, remember that this gradient with respect to w was a summation of i equal to 1 to 2 and that formula, right. So, for each data point I am computing the grad adding it, right. So, that is the summation part similar thing I am doing for b . Once I have computed the gradient which is the summation quantity I am just moving in the direction of the gradient, is that fine? Everyone understands the code, it is simple python code and it does exactly what I had shown in the pseudo code, ok.

Now, let us execute this code and see what happens, ok. So, I will start with my random point which was $[-2, 2]$ and now, I am going to actually run this code and keep plotting what happens on the figure, ok. So, just pay attention fine. So, now, here is how the code is running. See what is happening, what is happening actually? So, at every point I am changing my w , so that I am moving in the direction of the gradient I keep doing that as I keep doing that my error keeps decreasing, why? Because that is exactly what we got from Taylor series that if we do this the error is bound to decrease, right and

then we keep doing this and after a few iterations we will actually reach almost the value which is the zero error, right and this same thing would happen if you start from anywhere else it will keep moving in a principled way and reach the low error configuration, right.

Now, some of you would say that maybe this was the shortest path, right. It could have just rolled over from there, but that is not a principled way of doing that right we the principled way of doing it is to move in the direction of the gradient. You might take a longer route, but reach your destination, taking shortcuts is always risky, in life as well as here, right. So, so, do not please this is an advice for error assignments and so on, right. So, this is the more principled way and we will reach the solution, right. So, that is what is happening here. So, we have actually derived everything that we needed and this is all you need to write for gradient descent for this toy example that you had.

Now, answer this question. Now, suppose I had hundred such variables, instead of w comma b , I had hundred such variables. What would happen? You do not try to visualize it.

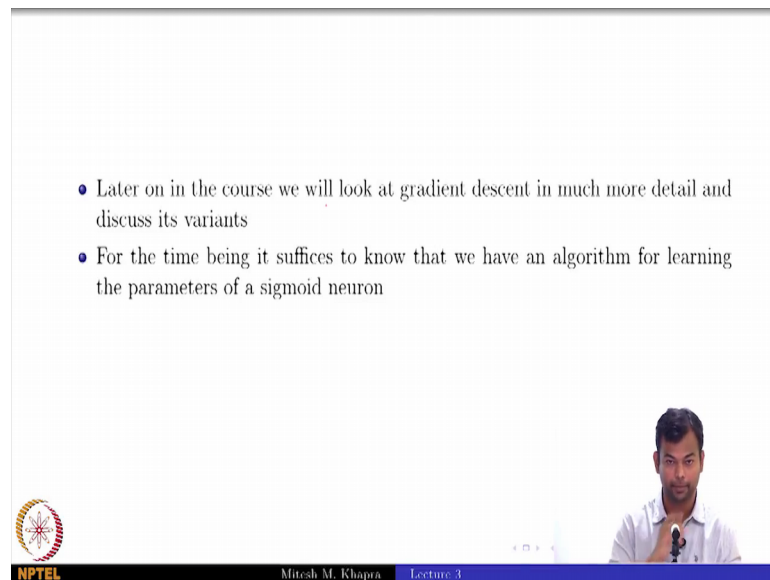
Student: (Refer Time: 29:48).

In terms of the code.

Student: (Refer Time: 29:52).

I will just need to have these functions for all of those, right. I will have to calculate it by hand, but still doable it is just a lot of tedious work of course, later on we will see a more refined way of doing this where we can do a lot of these computations at one go, right. So, we can directly start operating in vectors as opposed to scalars here, right. I am treating w and b separately here I could have actually had a function which tells me grad of theta directly, right. And later on we will see something like this ok, but for now the code is still running here.

(Refer Slide Time: 30:24)



The slide contains the following text:

- Later on in the course we will look at gradient descent in much more detail and discuss its variants
- For the time being it suffices to know that we have an algorithm for learning the parameters of a sigmoid neuron

The bottom of the slide features a video feed of a man in a white shirt speaking into a microphone. To the left of the video feed is the NPTEL logo. Below the video feed, the text 'NPTEL' is on the left, 'Mitesh M. Khapra' is in the center, and 'Lecture 3' is on the right.

Now, it suffices. So, later on we will see gradient descent in more detail in the course and we will also see a lot of variants of gradient descent, but for now it suffices that we have an algorithm which can learn the parameters of a sigmoid neuron, right. So, just as we had the perceptron learning algorithm, we have the gradient descent learning algorithm which can help us learn the parameters of the sigmoid neurons starting from random values. And it gives a principled approach for doing that.