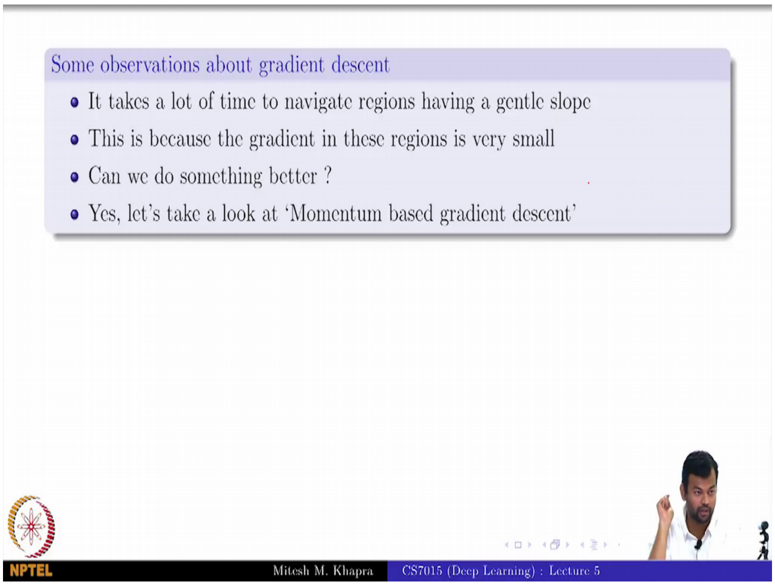
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 5.4
Lecture – 05
Momentum Based Gradient Descent

Ok. So, now, we will do a Momentum Based Gradient Descent ok. In this module we will look at momentum based gradient descent.

(Refer Slide Time: 00:19)



Some observations about gradient descent

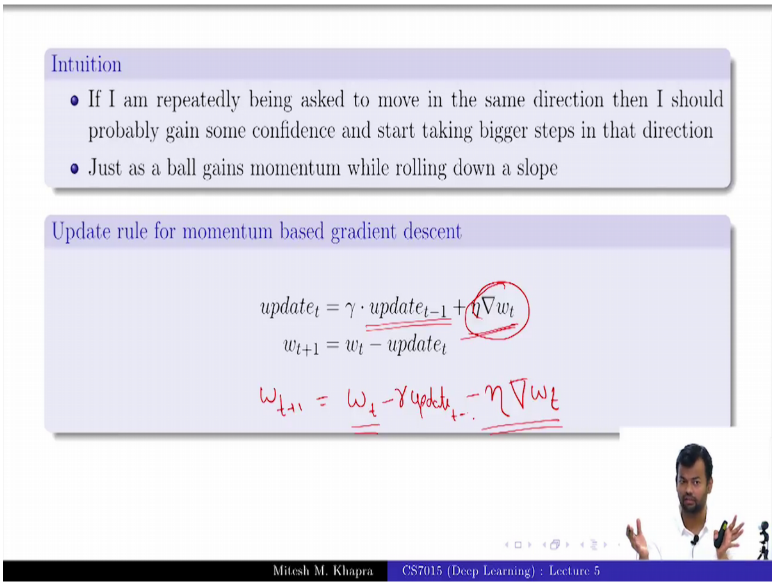
- It takes a lot of time to navigate regions having a gentle slope
- This is because the gradient in these regions is very small
- Can we do something better ?
- Yes, let's take a look at 'Momentum based gradient descent'

So, what were the observations about gradient descent that, it takes a lot of time to navigate regions having a gentle slope. So, what is the practical implication of this? in practice why it what does this need to, what does this mean right, it takes more time. So, remember we had said this math iteration equal to 1000.

Now if you are initialization happens to be such that you are stuck in this large flat region, then those 1000 iterations just keep moving around that flat region right, you will not enter into one of the valleys and valleys is what you are interested in right, because values is where you will have some minima for your function right

So, if you have a very very gentle slope, then for 1000 iterations you will keep moving around that gentle slope right, that is why this has a practical implication. Now this was because the gradient in these regions were small, can we do something better that is the question right. So, yes we can and we will take a look at momentum based gradient descent ok

(Refer Slide Time: 01:14)



Intuition

- If I am repeatedly being asked to move in the same direction then I should probably gain some confidence and start taking bigger steps in that direction
- Just as a ball gains momentum while rolling down a slope

Update rule for momentum based gradient descent

$$update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - update_t$$

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

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, here is the analogy which I give, my ts have heard this at least 10 times. So, I will just repeat it the 11 time for them. So, I hope that is the one which I want to use here yeah ok. So, now, suppose you are standing at the velachery gate and you want to go to phoenix market city, something that all if you can relate to today. So, you want to go to phoenix market city and you ask the security guy at the gate that where do I go right.

So, he will say take a left, no take a right. So, I am slightly dyslexic, actually I have a left right dyslexia. So, take, take a right ok. So, you will say he has told me to move right, but you would still be a bit cautious right, we will just keep moving slowly in that direction right, that is how we find ask for directions, you keep moving slowly in that direction right

Now, 100 steps later or 100 meters later you find another guy and you ask him or her where is phoenix market city. He again points to in the same direction, keep moving left right ok. So, now, you will, what will happen, you will increase your space and then you ask again someone, when you read the signal where it is and he again points in that direction what will happen, move even fast right.

So, what is happening here, if a lot of people are pointing you in the same direction, you better start taking larger and larger steps in that direction? Does that make sense, that is how we find directions and move around ok. So, just like a ball gains momentum as it goes down a slope right, it is constantly moving in that direction, so it starts moving faster. So, now, can you tell me a way of incorporating this? I have been moving in a certain direction these directions are nothing, but the gradients right. And now at this point someone asked me again to move in the same direction, what should I do?

Student: Take a bigger step.

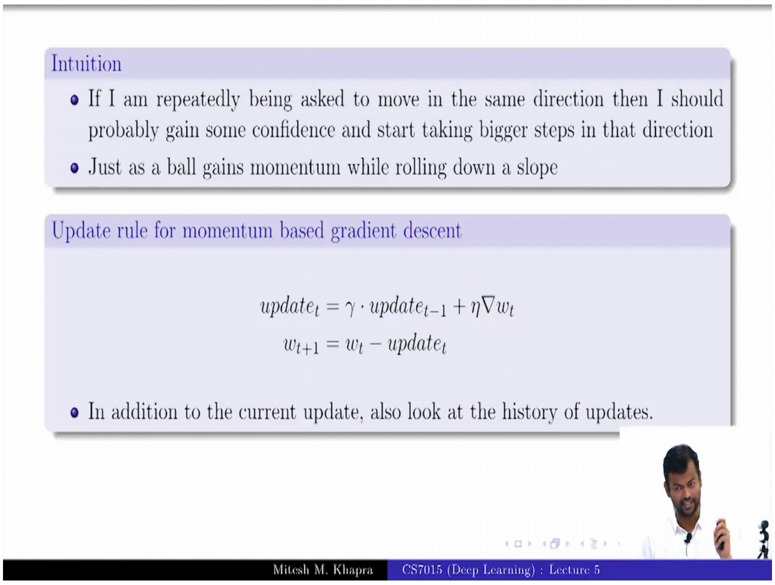
Take a bigger step. So, can you think or try to imagine, how would you do this mathematically?

Student: (Refer Time: 03:10)

Ok. So, it is probably there are a few ways to do it. So, let us see. So, what I am doing here is, this is my current gradient right. So, I asked that guy at the signal, he asked me to move in that direction. So, that is this direction and this is all my history, whatever I did till step t minus 1 ok. So, now, what I will do is, I will. So, earlier I was moving like this. This is what my update rule was wt plus 1 is equal to wt minus in the direction of the gradient right, I will moving in the direction opposite to the gradient

Now, what I have is in addition to that I have this gamma update t minus 1. So; that means, whatever I had done up till step t minus 1 I will also take that into account. So, I will end up taking a larger step. Is that clear? If it is not clear it will become clear on the next slide, but is it clear? Ok.

(Refer Slide Time: 04:18)



Intuition

- If I am repeatedly being asked to move in the same direction then I should probably gain some confidence and start taking bigger steps in that direction
- Just as a ball gains momentum while rolling down a slope

Update rule for momentum based gradient descent

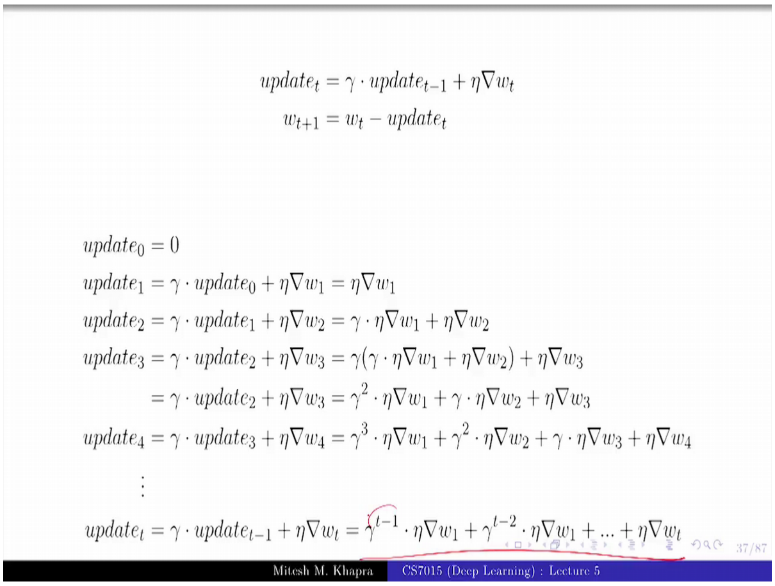
$$update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - update_t$$

- In addition to the current update, also look at the history of updates.

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, let us see what this means right. So, it basically means that in addition to the current step also look at the history. There are three guys who earlier pointed you in the same direction. So, maybe this direction makes sense right. So, start accumulating that and move faster ok.

(Refer Slide Time: 04:31)



$$update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - update_t$$

$$\begin{aligned}
update_0 &= 0\\
update_1 &= \gamma \cdot update_0 + \eta \nabla w_1 = \eta \nabla w_1\\
update_2 &= \gamma \cdot update_1 + \eta \nabla w_2 = \gamma \cdot \eta \nabla w_1 + \eta \nabla w_2\\
update_3 &= \gamma \cdot update_2 + \eta \nabla w_3 = \gamma(\gamma \cdot \eta \nabla w_1 + \eta \nabla w_2) + \eta \nabla w_3\\
&= \gamma \cdot update_2 + \eta \nabla w_3 = \gamma^2 \cdot \eta \nabla w_1 + \gamma \cdot \eta \nabla w_2 + \eta \nabla w_3\\
update_4 &= \gamma \cdot update_3 + \eta \nabla w_4 = \gamma^3 \cdot \eta \nabla w_1 + \gamma^2 \cdot \eta \nabla w_2 + \gamma \cdot \eta \nabla w_3 + \eta \nabla w_4\\
&\vdots\\
update_t &= \gamma \cdot update_{t-1} + \eta \nabla w_t = \gamma^{t-1} \cdot \eta \nabla w_1 + \gamma^{t-2} \cdot \eta \nabla w_1 + ... + \eta \nabla w_t
\end{aligned}$$

37/87

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, let us just break this down and see right. So, this is what the update rule is. Sorry this is all my updates and this is the update rule. So, at time step zero my update is zero, because not started yet. At time step 1 this is what it will look like right. And this is nothing, but just move in the direction of the opposite to the gradient, because this minus sign will come later on right in the next equation. Is it clear so far? Ok.

Now, what will happen update two? So, its gamma times update 1 plus the gradient at the current step. So, remember here everything is positive, I am adding the gradients, because my final negative sign is going to come in the next equation ok. So, do not get confused with that ok. Eventually I am going to move in the direction opposite that opposite will come from this negative sign ok

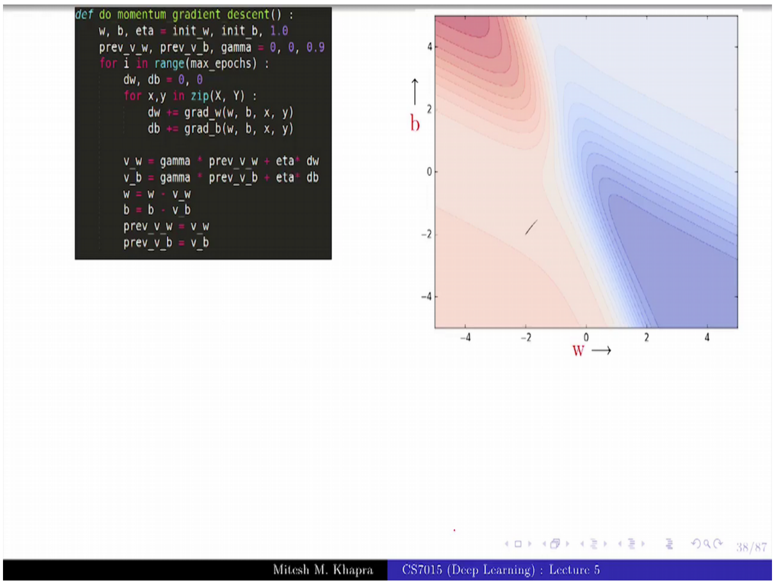
So, what is happening I am moving in the current direction plus a fraction of the direction which was pointed earlier right ok. Then does this make sense. So, can you tell me in general what is happening here at the th time step what is happening, what kind of average am I taking, weighted average, but it is a dash weighted average. This is an exponentially weighted average ok. So, let us look at this right

So, when I am at step 4, I have most faith in the current gradient right and this gamma is always I will just set it to less something less than 1 right. So, I have a fractional trust in the previous gradient, even smaller trust in the previous guy and even smaller trusts in the previous guys. So, I am taking an average of all my gradients, but it is an exponentially weighted average. Does that make sense? My maximum faith lies in the current guy and then decaying faith in the previous guys ok.

And as I move further and further away from the last guy that I checked right I will give lesser and lesser weightage to that. So, everyone understands what is happening here. Anyone who has a problem is, just raise your hands if you understand this good

So, in general this is going to be the formula and you see that as, as I form problem here no. As t is larger this fraction is going to become smaller and smaller right. So, you are first the first step that you take, will have lesser and lesser weightage as t increases. Everyone gets this fine.

(Refer Slide Time: 06:57)



```
def do_momentum_gradient_descent() :
    w, b, eta = init_w, init_b, 1.0
    prev_v_w, prev_v_b, gamma = 0, 0, 0.9
    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)

        v_w = gamma * prev_v_w + eta* dw
        v_b = gamma * prev_v_b + eta* db
        w = w - v_w
        b = b - v_b
        prev_v_w = v_w
        prev_v_b = v_b
```

38/87

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

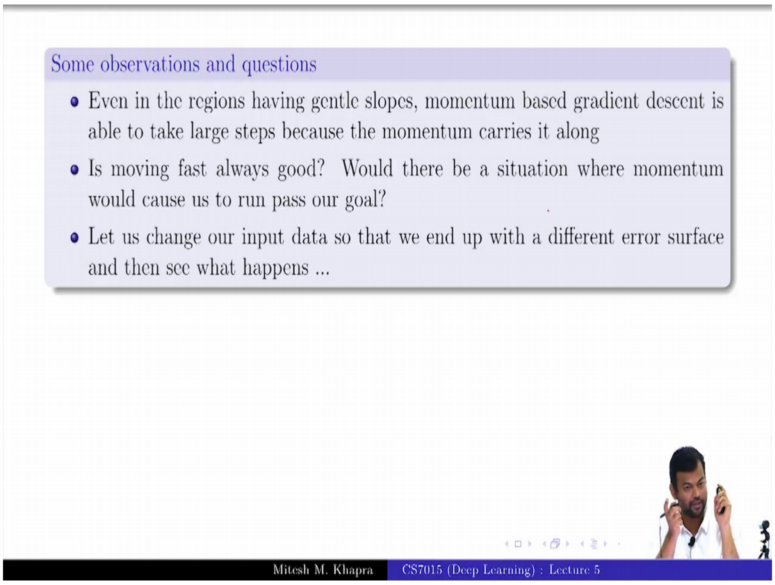
So, now this is the code for momentum based gradient descent . I will just give you a minute to stare at the code and see if it makes sense. So, this much part is ok, you are just computing the gradients with respect to all the points right. And now we are keeping this running sum ok, which is the previous gradients and the current gradient right and then you are just subtracting that running sum. Is that clear, everyone gets the code ok

Now, this sorry looking black curve that you see here; that is gradient this, this guy ok, this black curve that you see here, that is gradient descent when I have run it for around 100 iterations ok. Now I am going to run momentum base gradient descent and each click is going to be one step and I want you to observe what happens ok. So, slowly a red curve will start appearing on the figure.

Initially it will not be visible, so do not worry there is nothing wrong with your eyesight; one, how many if you already see the red part, I see it. 2, 3, 4, 5, 6, no now you can see it as is nothing great about7 8 9. I want you to observe something here 11, 12, 13, 14 came back right. So, gradient descent I ran it 400 iterations, it was just stuck here right this was a pointand I ran this for less than like around 15 or 20 is what we counted right and so, already entered into the valley right.

So, momentum base gradient descent is good, you see that wicked smile on my face and you know it is a trick question ok. So, we are moving fast right.

(Refer Slide Time: 09:02)



Some observations and questions

- Even in the regions having gentle slopes, momentum based gradient descent is able to take large steps because the momentum carries it along
- Is moving fast always good? Would there be a situation where momentum would cause us to run pass our goal?
- Let us change our input data so that we end up with a different error surface and then see what happens ...

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

Even in the regions where the slope was gentle right that is the beginning of the, beginning of our trajectory right, this was the gentle region, even that I was very quickly able to navigate right, within 5 to 6 steps I was away from that part right. So, even in the regions where the slope was gentle I was able to move fast, but is moving fast always good, philosophical questionright

So, would there will be a situation where momentum would cause us to run fast ago. Same thing now instead of walking you are in a car, you ask the person at the security whether I should go there, he says yes go in the right direction you keep moving there, someone else you keep accelerating, what will happen eventually? You will go fast phoenix market city then what will you do

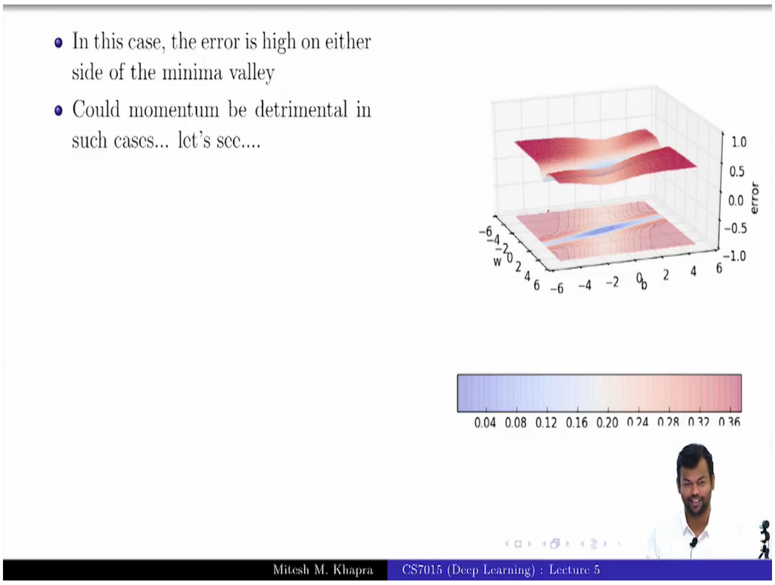
Student: Take a.

Take a u turn come back. Again while taking a u turn what will you do?

Student: (Refer Time: 09:57).

Overshoot and come to the signal and then go back again right. So, you see this you will end up taking a lot of u turns. So, let us change the input data a bit and see what happens to momentum based gradient descent ok.

(Refer Slide Time: 10:11)



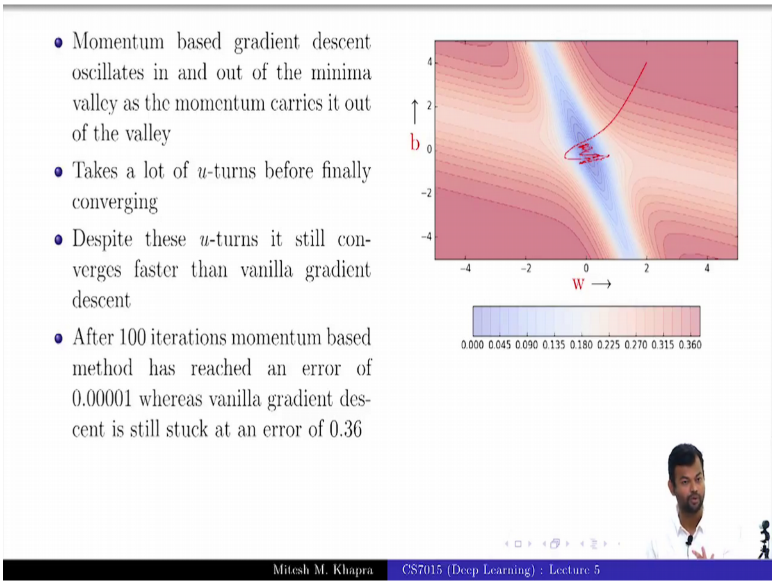
- In this case, the error is high on either side of the minima valley
- Could momentum be detrimental in such cases... let's see....

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, this is what my data looks like now . So, this is not what my data looks like, this is what my error surface looks like. So, earlier we had this error surface something like a flying carpet. Now I have a very peculiar error surface. This is again for the two parameter problem right w comma b; that means, I want to learn a sigmoid function, where I have these two plateaus at the top, the dark red regions that you see and then a very sharp valley fine. Can you tell me how I would have come up with this kind of an error surface, what are the points that I would have chosen. Just hold on to that partok.

So, I have this kind of an error surface fine, the error is high on either side of the valley right. Now could momentum be detrimental in this case, yes, no maybe I do not care, I do not care fine.

(Refer Slide Time: 11:07)



- Momentum based gradient descent oscillates in and out of the minima valley as the momentum carries it out of the valley
- Takes a lot of u-turns before finally converging
- Despite these u-turns it still converges faster than vanilla gradient descent
- After 100 iterations momentum based method has reached an error of 0.00001 whereas vanilla gradient descent is still stuck at an error of 0.36

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, let us see this is the, is this the 2 d equivalent of that 3 d surface, everyone gets it. I can perfectly verify that you get it ok, everyone gets it I will assume right. So, these are the very high plateaus where the error is very high, very sharp and narrow valley where the error is low ok

So, now again this sorry looking black curve is what I have done with gradient descent after some 100 iterations or something. Now I am going to run momentum based gradient descent and you have to help me understanding what is going to happen ok. Again you will soon start seeing that red curve appear 1 2 3 4 5 6 ok, what will happen now? It is already fast that is known, it was that black curve was after 100 iterations or. So, it is fast now tell me what will happen

Student: (Refer Time: 12:00).

He will go out, is actually almost come out of the valley right, it is almost at the top of the valley. Now what will do, take a u turn. Now what will I do. Again take a u turn. Now I will keep doing this, I will take now smaller and smaller u turns and it will converge right. So, what happens here is, because of this speedy movement and which is very analogous to that car movement which I described.

This overshoot your goal you will have to take the u turn come back if you are again careless you will have to keep taking these u turns, but you will finally, end up at the location that you want right . It takes a lot of u turns before converging. Despite these u turns it still converges faster than gradient descent right, because gradient descent can just not move at those gentle slopes right it just cannot move from there, because the gradient is almost 0, because the slope is flat right and it just cannot move, but even with this lot of u turn and lot of rework, after 100 iterations momentum base gradient descent has reached an error of almost0 whereas, gradient descent is still stuck at the plateau at an error of 0.36 ye. So, see you have reached the minima now.

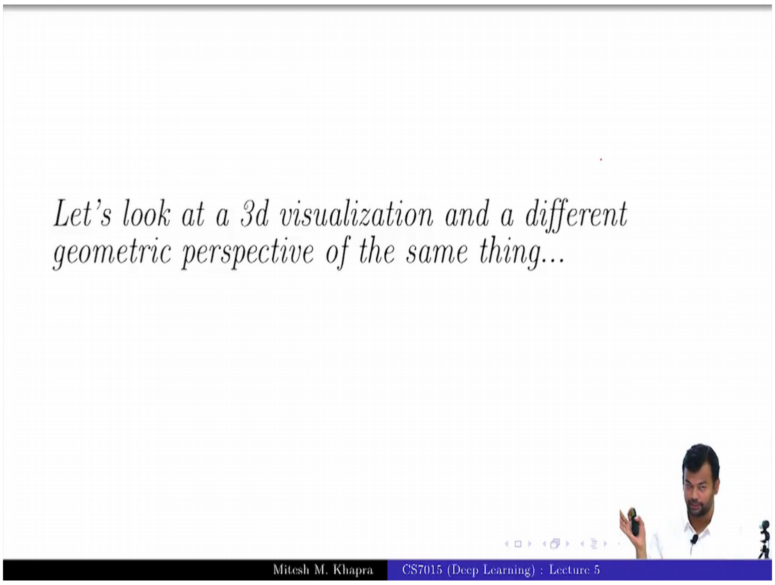
Student: Ye.

Right and now you will be navigating there right, but you know that now your loss is very slow low. So, you could end that right, you know that your loss is very close to 0. So, you could have a condition that once you have reached something very close to 0 you could end that, even if you are making these very small movements now you could just stop there

Student: But in the plateau regions is also 0.

But the loss is high right. So, if the loss is high and you are not moving you cannot stop, but if the loss is low and you are not making movements you can just stop there right. So, you can just end, you can define that as your convergence condition right fine.

(Refer Slide Time: 13:44)

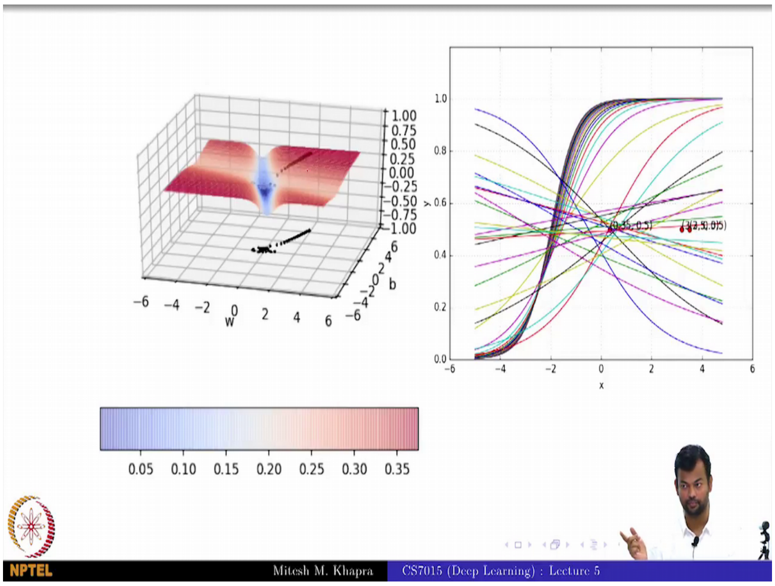


Let's look at a 3d visualization and a different geometric perspective of the same thing...

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, let us look at, we will come back to 3 d now. We look at a 3 d visualization and a very different interpretation of what is happening. I really want you to understand what exactly is happening in this example which I had picked up right.

(Refer Slide Time: 13:59)



Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5

So, this is what the 3 d surface looks like view from a different angle, you have these two plateaus and the very sharp valley. Now, this is the corresponding sigmoid function where I started with. So, what I am trying to tell you is that, this is a sigmoid function corresponding to w equal to 6, oh no sorry w equal to 2 and b equal to 6.

This is the sigmoid function that I got once I plug that value. So, sigmoid is 1 over 1 plus e raised to minus w x plus b and I have plugged in the values of w and b and plotted it for all the values of x and this is the sigmoid that I got ok. So, that is my starting point. Is this good, how do you define good or bad

Student: (Refer Time: 14:43).

What do you expect at the end of training? It should pass through all your training points and these are my training points ok. Is it passing through them, no its way off right? Ok. So, now, let us start this momentum based gradient descent and what just see how my sigmoid function changes ok. So, right now I am on the gentle slope, even that momentum base gradient descent it is going to be fast, but not dramatically fast, because still building up the momentum right .

So, it is you see that these sigmoid that I am drawing here; they are almost indistinguishable from each other. I have already drawn three sigmoids here. So, I will just go back. So, there was this initial guy then I draw drew a red one then one more and then one more, but they are all very close to each other

Now, keep viewing both these sides in parallel, what happens here on this figure and what happens to this sigmoid and I will ask you questions. So, still I am moving a bit slowly, because I am still building the momentum right, it takes time to build that moment. Now I have slowly started building the momentum my sigmoids have started moving towards where they should be. Everyone gets this what is happening here ok

Now, tell me what will happen. As I enter the valley, I am almost entering the valley what will happen. I have gained this momentum now. So, my w comma b values are going to change much faster now. So, what will happen to these sigmoids, they no longer stick to each other; we will start seeing a difference they are already moving away from each other ok. So, that is what is happening to the function ok. Now you see even faster changes ok, now what will happen. I have entered the valley; this is how my sigmoid looks at this point. Now tell me what will happen?

Student: (Refer Time: 16:34).

It will go fast, what will happen to your sigmoid, how many of you know what will happen to the sigmoid ok. I will tell you what happens and then it will be obvious right. So, now, I am entering the valley, all of us know that I am going to come out of the value of the other side right. So, let us see what happens when I come out of the valley from the other side, the sigmoid changes that is why you have this situation that your error is high on both sides right, because on this side you have these kind of sigmoids, on the other side you have the other sigmoids and somewhere in between lies the solution. Where does the solution lie, add a very flat sigmoid right

So, now I start, this is where the oscillations will happen. So, notice what will happen to the sigmoids, they will toggle between these two orientations ok, just see what happens to the sigmoids, you see it again moves keeps moving, keeps moving, it keeps oscillating around the solution and then finally, you reach the solution. So, you see that, should I repeat this ok.

So, when I am on one side of this valley, I have one kind of sigmoids right. Now when I move to the other side of the valley I have this others kind of sigma and take a u turn. So, when I u turn take a u turn I again overshoot and go to the other side, and this keeps happening and I keep toggling till I reach my final solution right

So, these are all the oscillations that you are seeing. So, can you visualize this, what is happening? Do you understand all these, relates to the actual function that you are trying to learn ok, this is fine ok. So, that is why we will end this module, this was on momentum base gradient descent. Now we will see a nesterov accelerated gradient descent.