

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

Module – 5.5
Lecture – 05
Nesterov Accelerated Gradient Descent

So let us look at Nesterov Accelerated Gradient Descent.

(Refer Slide Time: 00:16)

Question

- Can we do something to reduce these oscillations ?
- Yes, let's look at Nesterov accelerated gradient

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

So, now we know that momentum based gradient descent is good at these gentle regions, it moves really fast, but we do not, still do not like it, because it has this problem of oscillations. It has this problem that it overshoots its objective right, it is good and then it has to take a lot of u turns. So, can we do something about reducing this oscillation? So, the answer is always yes. So, let us look at Nesterov accelerated gradient descent ok.

(Refer Slide Time: 00:41)

The slide features a purple header with the word "Intuition" in white. Below it, a light blue box contains two bullet points: "• Look before you leap" and "• Recall that $update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$ ". To the right of this box, the equation $w_{t+1} = w_t - \gamma \cdot update_{t-1} - \eta \nabla w_t$ is written in red. The bottom of the slide includes the NPTEL logo, the name "Mitesh M. Khapra", and the text "CS7015 (Deep Learning) : Lecture 5". A small video inset of the speaker is visible in the bottom right corner.

So, the idea here is very simple, look before you leap ok. Now remember that this was the update rule for momentum based gradient descent and I will write it down again w_{t+1} is equal to w_t minus γ into $update_{t-1}$ minus η into the gradient at the current point ok

So, you see that actually I am taking two steps; one is this step and then one more step and I could just this is one way of visualizing right, that I move according to the history and then I move a bit more according to the current gradient. So, everyone sees that there is a two step movement happening here.

Now, can you think what could have been done, look before you leap ok. So, we will see what we can do.

(Refer Slide Time: 01:34)

Intuition

- Look before you leap
- Recall that $update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$
- So we know that we are going to move by at least by $\gamma \cdot update_{t-1}$ and then a bit more by $\eta \nabla w_t$
- Why not calculate the gradient ($\nabla w_{look.ahead}$) at this partially updated value of w ($w_{look.ahead} = w_t - \gamma \cdot update_{t-1}$) instead of calculating it using the current value w_t

$w_{t+1} = [w_t - \text{---}] - \text{---}$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 5 46/87

So, we know that we are going to move at least by this one way, that is fixed we know that our history is telling us to move at least by this one and then we will move a bit more by the gradient ok

So, now can you think about it, I am at least going to move this much. What if I had some way of looking ahead and then do something at that point, this is what you are saying. Of course, I can verify it, but I am sure it will become clear once I show the equations, but I just want you to think about it a bit wait, it is very simple it will become absolutely clear once I show you the answer, but just think about it a bit ok.

So, here is the answer it, why not compute the gradients add this look ahead point right. So, you are again adding it in two steps; minus the history and then minus the current gradient. So, take this value, call it the look ahead point. I know that I am going to move by this much. So, let me not compute the gradients at the current point, let me move by this much, then compute the gradients and see what happens at that point. Does it make sense? Ok.

(Refer Slide Time: 02:52)

The slide is divided into two main sections. The top section, titled 'Intuition', contains a bulleted list of points. The bottom section, titled 'Update rule for NAG', contains three equations. Handwritten red annotations include circles around the equations and a handwritten ∇w_t next to the second equation. A small video inset of a speaker is visible in the bottom right corner of the slide area.

Intuition

- Look before you leap
- Recall that $update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$
- So we know that we are going to move by at least by $\gamma \cdot update_{t-1}$ and then a bit more by $\eta \nabla w_t$
- Why not calculate the gradient ($\nabla w_{look.ahead}$) at this partially updated value of w ($w_{look.ahead} = w_t - \gamma \cdot update_{t-1}$) instead of calculating it using the current value w_t

Update rule for NAG

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

We will have similar update rule for b_t

NPTEL | MITESH M. KHAPRA | CS7015 (Deep Learning) : Lecture 5

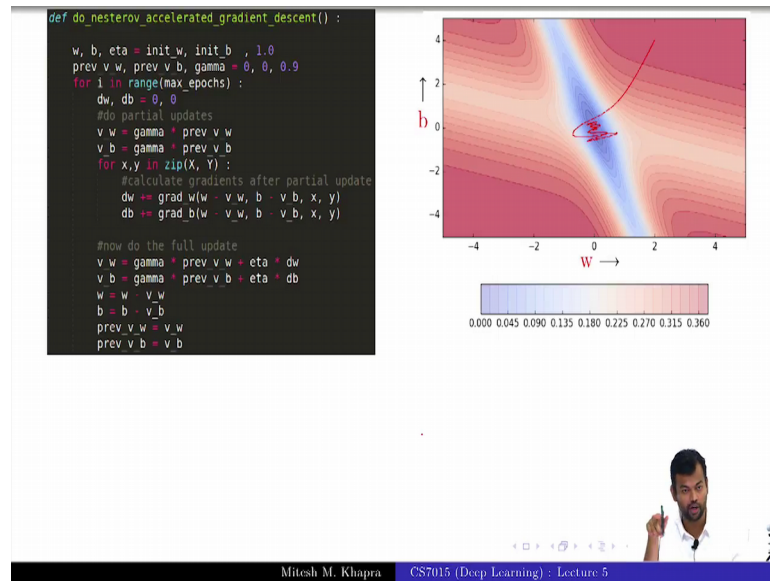
So, this is the equation right that first I move by that one step. I had to make a two step movement. So, I will move by that one step right, then I will compute the gradient at that position, not at my current position right, this was earlier gradient at point t , now I have already moved a bit. So, I can compute the gradient there and then move in the direction of that gradient. Does that make sense? How many of you are completely lost, you are completely lost good, may not good, but yeah good that you raised your hand ok. So, ok

So, you understood this that there is a two step movement right w_t minus history minus the current gradient, gradient computed at time step t ok. Now you know that you are already going to move by the history right. So, why not just move there and then compute a gradient at that point you are anyways made some movement you compute the gradient at that point and then decide which is the direction to move in right..

So, that is what this look ahead value is ok. I know it is still not clear to many of you and I am very confident it will become clear in the next 5 minutes; we will show you one more visualization for this, but this stay with, stay with me for a while. As long as you get the intuition I am fine I will move ahead and then I will explain it again in a different way ok, this is fine ah, that should become clear good that you asked that question ok. So, ask me again on the blank slide that I have and then I, it should be complete ok

So, for right now let me just show you what will happen with the code, and then I will again explain it with a different way ok.

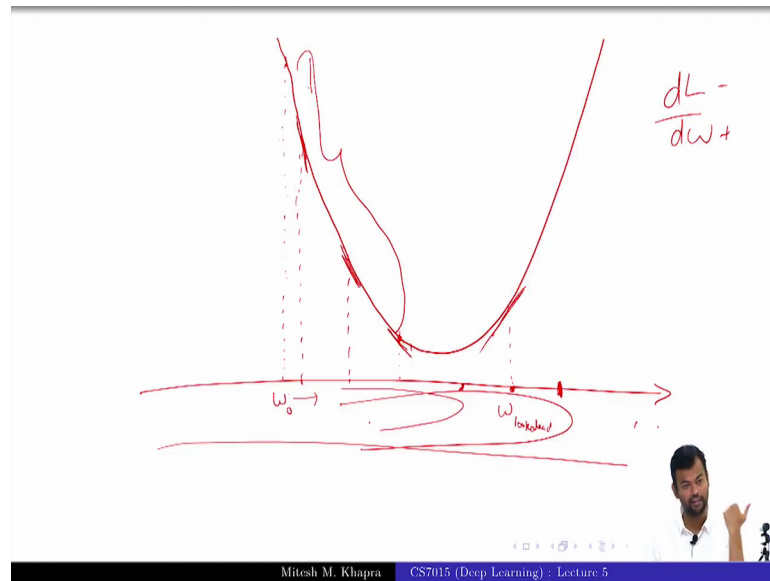
(Refer Slide Time: 04:35)



So, this is what momentum based gradient descent it ok. Now let us see what nested or accelerated gradient descent will do. Again the code is simple you can just read it up and I have started executing, you see this blue curve coming over there fine, and now I keep running this. Now what will happen? You see that all the u turns of the blue curve are inside the u turns of the red curve.

So, the objective is being achieved at least empirically I have showed you that right, its taking shorter u turns. What is probably not clear to all of you is, why is this happening. Is it clear to everyone, why is this happening, can everyone visualize that ok. So, let us see why this is happening, I will give you an alternate explanation for this ok.

(Refer Slide Time: 05:38)



So, suppose this is my error surface right on a two by and I have a single variable with respect to which I am trying to optimize. So, this is my w ok. I started off with some initial value w_0 ok.

Now, what is the gradient at this point; positive negative, negative right, because when I am going to increase w the function is actually going to decrease right. So, right. So, the slope is negative. So, where will I move, this is the number line right. So, this line is actually the number line, because it is a single variable. So, where will I move positive side of the number line or the negative side of the number; positive side. The derivative is negative; I am going to move in the direction opposite to the derivative. So, I am going to move in the positive direction right. So, I will end up somewhere here. Is that clear, fine with everyone ok.

So, now I am somewhere here, what is the derivative at this point ok? Now what is the derivative here; positive negative, negative right when I am increasing w my loss function is decreasing. So, my dL/dw is going to be negative, this is positive this is negative right ok. So, again I will move in this direction.

So, what is happening a lot of negative updates are getting? Sorry a lot of positive updates are getting accumulated right. And now because of my momentum I am not going to move only by this derivative I am also going to move by the history right. So, I will end up somewhere further ok. Is that fine?

So, now at this point what is the derivative, again negative, when I am increasing w the function is decreasing. So, what is my update positive or negative? Positive. So, now, you see that a lot of positive updates are getting accumulated right, my momentum is building up. So, now, what will happen? Now if I just move further, then again I will get a let me just put it here right. So, I am again moving largely in the positive direction, because this guy is also positive, all my history was also positive. So, I have moved in the positive direction

Now, what will happen at this point, what is the derivative here. No it is still its negative sorry sorry. So, again I am going to move in the positive side of the number line ok. Now at this point I want you break down the movement into two points; one is what my history was telling me, which was all these positive updates, but of course, I will not make such a large update, because I am waiting them exponentially right so, but its telling me to move in the positive direction ok, and I know that the gradient at this point is negative, but I want you to ignore that for now. I just want you to focus on the history. If I just move according to the history where will I end up? I will end up somewhere here right, because the history is very positive. So, I will keep moving in the positive direction and this is my w look ahead ok

Now, what will happen if I compute the gradient here?

Student: Positive.

The gradient is.

Student: Positive.

Positive. So, where will I move?

Student: Negative.

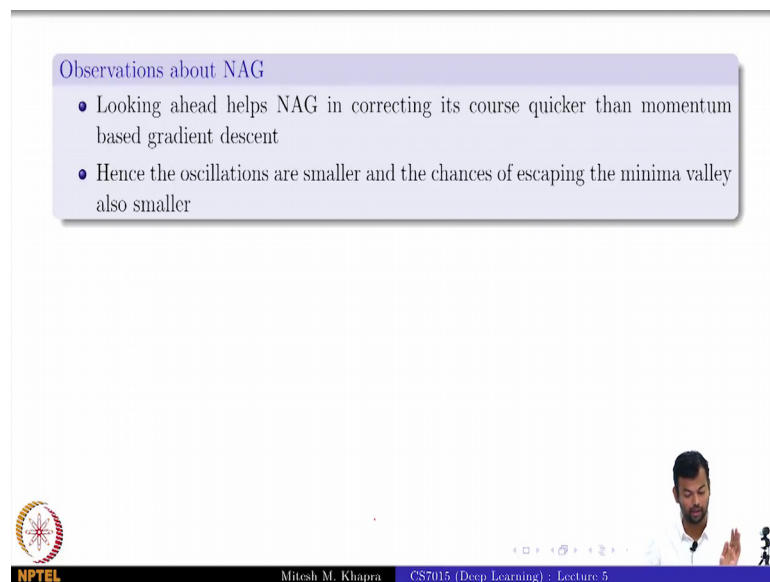
Negative. So, you see now why momentum works, because you are able to look ahead to this point. Instead of what should I have actually done is I should have looked at the gradient at this point, the history is positive; the gradient is also telling me to move positive. So, I would have moved a large positive and I would have ended somewhere here. Instead I just moved by the history, I checked where I end up I end up here.

Now, let me see whether what is the gradient at this point? Have I already overshoot my overshoot my objective, when would I overshoot my objective, it has the sign of the gradient changes right, it became from negative to positive and now since its positive because as I am increasing w , the loss is also increasing. So, now, where will I move? Negative.

So, now, what is the second step actually its again bringing me close to here. So, instead of taking this large u turn, I end up taking this small u turn. Is this clear to everyone now? How many if you still do not get it, how many if you get it now, good sure ok. So, this is what and now we can relate it to what was happening on the figure

So, let us go back right. So, you saw that I was making these smaller u turns, because when I was at this point right, I already moved by the history I knew I would land up somewhere here, where I would need to go back right. So, I already accounted for that and made a very small movement. Is this clear, everyone gets this how the Nesterov of accelerated gradient descent works, sure raise your hands.

(Refer Slide Time: 10:42).



Observations about NAG

- Looking ahead helps NAG in correcting its course quicker than momentum based gradient descent
- Hence the oscillations are smaller and the chances of escaping the minima valley also smaller

NPTEL Mitesh M. Khurana CS7015 (Deep Learning) : Lecture 5

So, looking ahead helps nag in correcting its course quicker than momentum based gradient descent right. So, it is already looking ahead where do I land up and already making a correction if required, if not required it will again move in the right direction right. So, the update is this guy plus the gradient and my update happens on the original value not on the look ahead value.

So, her confusion was perhaps that I am doing w look ahead minus update, where this update again has this quantity you know that is what your confusion was, but I am not doing w locate I am using w_t there; everyone gets this ok. So, that is where ah. Now it is clear that why the oscillations are smaller in the case of nag and it is able to correcting its course quicker ok. So, that is why we will end this module.