

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

Lecture - 05
Explanation for why we need bias correction in Adam

(Refer Slide Time: 00:12)




So, in this video we will try to look at an explanation for why we need Bias Correction in Adam ok. Or in other words I want to explain why do I do this particular step why did I take m_t and v_t as it is, but why did I do this particular step which I called as the bias correction step ok.

(Refer Slide Time: 00:21)

Update equations for Adam

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * \nabla w_t$$
$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * (\nabla w_t)^2$$
$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} * \hat{m}_t$$

- Note that we are taking a running average of the gradients as m_t
- The reason we are doing this is that we don't want to rely too much on the current gradient and instead rely on the overall behavior of the gradients over many timesteps
- One way of looking at this is that we are interested in the expected value of the gradients and not on a single point estimate computed at time t
- However, instead of computing $E[\nabla w_t]$ we are computing m_t as the exponentially moving average
- Ideally we would want $E[m_t]$ to be equal to $E[\nabla w_t]$
- Let us see if that is the case



◀ ▶ ⏪ ⏩ ⏴ ⏵

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

So note that in the case of Adam if you look at this equation for m_t we are actually taking a running average of the gradients and storing it as m_t right. So this is the gradient and we are taking a running average or exponential running average of these gradients exponentially decaying running average right.

Ah So the reason we are doing that is that we do not want to rely on a single estimate, so we do not want to rely only on gradient of w_t we want to look at the overall behaviour of the gradients over multiple time steps and then take a decision. So that means, in one particular gradient at time t is actually pushing us in some direction we do not want to be very hasty and start moving there we want to accumulate the history and appropriately weigh everything in the history, that is the idea behind taking this running average of radiance ok.

And the other way of looking at is that we are interested in the expected value of the gradients and not the point estimate at time w_t right. At time t rather so gradient of w_t which is this quantity which is the point estimate at time t , we are not interested in that were interested in the expected value and our behaviour should be according to the expected value that is what we desire.

So however, instead of computing the expected value of this quantity which should have been ideal, we are computing empty as the exponentially moving average. So in the ideal case we would want that these two quantities are the same that the expected value of

empty the way I am computing it and the expected value of the gradient of w_t should be the same. If that is the same then I am fine because then; that means, I am just taking the expected value or the, of the gradient instead of relying on the point estimate ok. So, let us see if that is indeed the case.

(Refer Slide Time: 02:14)

• For convenience we will denote ∇w_t as g_t

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g_t$$

$$m_0 = 0$$

$$m_1 = \beta m_0 + (1 - \beta)g_1$$

$$= (1 - \beta)g_1$$

$$m_2 = \beta m_1 + (1 - \beta)g_2$$

$$= \beta(1 - \beta)g_1 + (1 - \beta)g_2$$

$$m_3 = \beta m_2 + (1 - \beta)g_3$$

$$= \beta(\beta(1 - \beta)g_1 + (1 - \beta)g_2) + (1 - \beta)g_3$$

$$= \beta^2(1 - \beta)g_1 + \beta(1 - \beta)g_2 + (1 - \beta)g_3$$

$$= (1 - \beta) \left(\sum_{i=1}^3 \beta^{3-i} g_i \right) (\beta^2 g_1 + \beta^1 g_2 + 1 g_3)$$

So, for convenience we are going to just denote this gradient w_t as g_t because it is cumbersome to write this grad symbol and we will just not make it so readable the derivation that we are going to do. So I am just going to replace that as g_t so what I have written is g_t here instead of grad w_t right. So from now on I will just use g_t for grad w_t is that fine ok, so we have this expression for m_t .

So, now let us just try to expand it and see what happens right so m_0 it is going to be 0 because that is my starting points I have no history nothings, so I will just going to keep it as 0, m_1 is my first time step at which it is going to be beta into m_0 so I am just substituted t minus 1 and t here. And in the original expression I have just substituted appropriate quantities for m of t minus 1 and g of t , so m of t minus 1 is 0 m_0 and g of t is g_1 and of course, $b_0 m_0$ itself was 0, so what will be left it is $1 - \beta g_1$.

Now, let us look at what happens is m_2 , m_2 is going to be beta m_1 plus $1 - \beta g_2$, but I already have an expression for m_1 , so I am just going to substitute that here and this is what I get ok. Now let us look at m_3 , m_3 is again going to be beta times m_2 plus

1 minus beta times g 3 and I have an expression for m 2 so I am going to substitute that here and see if that leads to something interesting right.

So, I have just substituted the value of m 2 here right and I already had the m 3 part here the, this term here as it is ok. And now let us see so this already starts looking something interesting you see some pattern here, in particular we could take these 1 minus beta terms outside they can be taken common and then you will be left with beta square g 1 plus beta square g 1 plus beta g 2 plus g 3. So let us try to write this more compactly right, so I have taken one minus beta common and then I have written the remaining terms as this particular summation and you can verify right.

So, when I is equal to 1 this is going to be beta 3 minus 1 which is beta square into g 1. When I is equal to 2 this is going to be beta 3 minus 2 which is going to be beta into g 2 and when I is going to be 3 this is going to be beta raise to 3 minus 3 which is beta raise to 0 which is just 1 into g 3 right. So we get back the same expression that we had here of course, there is a 1 minus beta outside, so this is a more compact way of writing it and this was for the 3th entry right this was for m 3, the third entry.


Now, what if we want to write it for the tth entry in general, what if we want to write the expression for m t right.

(Refer Slide Time: 05:05)

- For convenience we will denote ∇w_t as g_t

$$\begin{aligned}
 m_t &= \beta_1 * m_{t-1} + (1 - \beta_1) * g_t \\
 m_0 &= 0 \\
 m_1 &= \beta m_0 + (1 - \beta)g_1 \\
 &= (1 - \beta)g_1 \\
 m_2 &= \beta m_1 + (1 - \beta)g_2 \\
 &= \beta(1 - \beta)g_1 + (1 - \beta)g_2 \\
 m_3 &= \beta m_2 + (1 - \beta)g_3 \\
 &= \beta(\beta(1 - \beta)g_1 + (1 - \beta)g_2) + (1 - \beta)g_3 \\
 &= \beta^2(1 - \beta)g_1 + \beta(1 - \beta)g_2 + (1 - \beta)g_3 \\
 &= (1 - \beta) \sum_{i=1}^3 \beta^{t-i} g_i
 \end{aligned}$$

- In general,

$$m_t = (1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i$$


So, in general m_t we can write it as $(1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i$. So this 3 is here I have just replaced them by t s right you can just verify that this is from you can just generalize from the third entry to the t th entry fine.

(Refer Slide Time: 05:27)

• So we have, $m_t = (1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i$ $E[m_t]$

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

So now, let us see we have the following expression we have simplified the expression for m_t and written it more compactly, but what we were eventually interested in the expected value of m_t right, we wanted to show that certain things holds for the expected value of m_t .

(Refer Slide Time: 05:40)

• So we have, $m_t = (1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i$

• Taking Expectation on both sides

$$E[m_t] = E[(1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i]$$

$$E[m_t] = (1 - \beta) E[\sum_{i=1}^t \beta^{t-i} g_i]$$

$$E[m_t] = (1 - \beta) \sum_{i=1}^t E[\beta^{t-i} g_i]$$

$$= (1 - \beta) \sum_{i=1}^t \beta^{t-i} E[g_i]$$

• Assumption: All g_i 's come from the same distribution i.e. $E[g_i] = E[g] \forall i$

$E[g_1]$

$E[g_2]$

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

So, you just take expectation on both sides so this is what we will get ok. Now $1 - \beta$ is of course, a constant so I can move it outside the expectation, so then I get an expectation of a sum.

Now, the expectation of a sum is the same as the sum of expectations, so I can write it as a sum of expectations ok. Now again β is a constant so I can take it outside the expectation, so what I will be left with is β^i outside and expectation of g_i right. So this is actually expectation of g_1 when $i = 1$, then expectation of g_2 , expectation of g_3 and so on ok.

Now, we will make an assumption that all these g_i s; that means, the gradient at time step 1, the gradient at time step 2, the gradient at time step 3 and so on they all come from the same distribution ok. We are going to make that assumption so let us try to understand the implication of that right. So let us say this was a distribution from which g_1 came right suppose I am dealing with a scalar quantity and maybe this was the distribution from which g_1 came. Now g_2 could have come from a different distribution, g_3 could have come from a different distribution and if that was the case then expectation of g_1 would be different from the expectation of g_2 and so on.

So, what we have assumed to it will make things simple for us is that g_1, g_2, g_3 any g_i comes from the same distribution and hence you can say that the expectation of all these g_i s is going to be just the expectation of g , that is this one single distribution from these which these entries come this of course, a very strong assumption, but we are going to live with this assumption ok.

(Refer Slide Time: 07:16)

- So we have, $m_t = (1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i$
- Taking Expectation on both sides

$$E[m_t] = E[(1 - \beta) \sum_{i=1}^t \beta^{t-i} g_i]$$

$$E[m_t] = (1 - \beta) E[\sum_{i=1}^t \beta^{t-i} g_i]$$

$$E[m_t] = (1 - \beta) \sum_{i=1}^t E[\beta^{t-i} g_i]$$

$$= (1 - \beta) \sum_{i=1}^t (\beta^{t-i} E[g_i])$$

- Assumption: All g_i 's come from the same distribution i.e. $E[g_i] = E[g] \forall i$

$$E[m_t] = (1 - \beta) \sum_{i=1}^t (\beta^{t-i} E[g])$$

$$E[m_t] = (1 - \beta) E[g] \sum_{i=1}^t (\beta^{t-i})$$

$$= E[g] (1 - \beta) (\beta^{t-1} + \beta^{t-2} + \dots + \beta^0)$$

$$= E[g] (1 - \beta) \frac{1 - \beta^t}{1 - \beta}$$

the last fraction is the sum of a GP with common ratio = β

$$E[m_t] = E[g] (1 - \beta^t)$$

$$E[\frac{m_t}{1 - \beta^t}] = E[g]$$

$$E[\hat{m}_t] = E[g] \left(\frac{m_t}{1 - \beta^t} = \hat{m}_t \right)$$

Hence we apply the bias correction because the expected value of \hat{m}_t is the same as the expected value of g_t

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

So then this expectation of g_i just becomes expectation of g , so I have gotten rid of the index i ; that means, I can move it outside the summation right so this is what I will get now. These two have come out of the summation and inside I have this quantity, now let me just expand this quantity this is nothing but beta raise to t minus 1 plus beta raise to t minus 2 plus so on at last you will reach t minus t which is just going to be beta raise to 0.

So, this is nothing but a sum of a g p with common ratio beta and I can replace that sum by this formula, you know this is the formula for the sum of a g p with common ratio beta. So I have just replaced that and now what happens is this 1 minus beta and 1 minus beta cancel out, so I get this particular expression that the expected value of m_t is equal to the expected value of g into 1 minus beta t .

So, I will just take 1 minus beta t on the other side and I can move it inside the expectation because it is a constant it does not matter. So I will get as oh actually yeah I can just move it inside so I will get it as expectation of m_t over 1 minus beta is equal to expectation of g_t right and this quantity the one which I have circled is nothing but \hat{m}_t right this was exactly the bias correction that I was applying. If I go back to the previous slide or the slide before that, so this was exactly the bias correction that I was applying right.

So, what I have inside is this, so what I have shown is that if I apply the bias correction then the expected value of the bias corrected m_t is equal to the expected value of the gradient and that is actually what I wanted, I wanted that whatever m_t . I am computing if I look at its expected value it should be the same as the expected value of my gradients and that is what I have arrived at right.

Hence this bias correction makes sense and hence we apply this bias correction for Adam. So this we have shown for m_t , we had a similar expression for v_t right, so for m_t we had this bias correction as \hat{m}_t and similarly for v_t also we had this bias correction as \hat{v}_t so you can derive the same kind of derivation for v_t also and show that that bias correction makes sense right. So this is an explanation for why you do bias correction in the case of Adam ok, so we will end this lecture here ok.

Thank you.