

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

**Lecture - 105**  
**The problem of Exploding and Vanishing Gradients**

And that takes us to the Problem of Vanishing and Exploding Gradients ok. So, you want to see what is a problem with this back propagation through time, which could lead to certain interesting situations?

(Refer Slide Time: 00:24)

- We will now focus on  $\frac{\partial s_t}{\partial s_k}$  and highlight an important problem in training RNN's using BPTT

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \dots \frac{\partial s_{k+1}}{\partial s_k}$$

$$= \prod_{j=k}^{t-1} \frac{\partial s_{j+1}}{\partial s_j}$$

- Let us look at one such term in the product (i.e.,  $\frac{\partial s_{j+1}}{\partial s_j}$ )

$\frac{\partial s_j}{\partial s_{j-1}}$

34/41

So, we will focus on this  $\frac{\partial s_t}{\partial s_k}$  and let me just go back. So, remember that this formula had this  $\frac{\partial s_t}{\partial s_k}$ , where  $s_t$  could be the last time step and  $s_k$  could also be the first time step because you are summing over all the time steps, right.

(Refer Slide Time: 00:27)

• Finally we have

$$\frac{\partial \mathcal{L}_4(\theta)}{\partial W} = \frac{\partial \mathcal{L}_4 \theta}{\partial s_4} \frac{\partial s_4}{\partial W}$$

$$\frac{\partial s_4}{\partial W} = \sum_{k=1}^4 \frac{\partial s_4}{\partial s_k} \frac{\partial s_k}{\partial W}$$

$$\frac{\partial \mathcal{L}_t(\theta)}{\partial W} = \frac{\partial \mathcal{L}_t(\theta)}{\partial s_t} \sum_{k=0}^{t-1} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial W}$$

• This algorithm is called backpropagation through time (BPTT) as we backpropagate over all previous time steps

NPTEL  
Mitesh M. Khapra CS7015 (Deep Learning) 32/41

So, you could have a term which is  $s_t^T$  which is the last time step the first time step and the derivative of the last time step with respect to the first time step, right. So, that is a situation that we are dealing with. So, we will consider one such generic element which is  $\frac{\partial s_t}{\partial s_k}$  and we will just try to expand it. So, remember I have done this short circuiting, so I am now just going to expand it again. So, this is going to be  $t$  by  $t-1$ ,  $t-1$  by  $t-2$  and so on up to  $k+1$  by  $s_k$  ok or I can write it as this generic formula. Everyone find with this, I have just replace this as a product and written it more compactly.

Now, let us look at one such term here  $\frac{\partial s_{j+1}}{\partial s_j}$ . Now, just to confuse you guys from next slide I will go over to  $\frac{\partial s_j}{\partial s_{j-1}}$  or not confuse you I just did not pay attention to this. So, instead of  $s_{j+1}$  and  $j$  and I am going to do  $j$  and  $j-1$ , right, it remains the same does not matter.

(Refer Slide Time: 01:33)

$$a_j = [a_{j1}, a_{j2}, a_{j3}, \dots, a_{jd}]$$

$$s_j = [\sigma(a_{j1}), \sigma(a_{j2}), \dots, \sigma(a_{jd})]$$

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \dots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \dots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \dots & \\ 0 & 0 & \dots & \sigma'(a_{jd}) \end{bmatrix}$$

$$= \text{diag}(\sigma'(a_j))$$

- We are interested in  $\frac{\partial s_j}{\partial s_{j-1}}$

$$a_j = W s_{j-1} + b$$

$$s_j = \sigma(a_j)$$

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j} \frac{\partial a_j}{\partial s_{j-1}} = \text{diag}(\sigma'(a_j))W$$

- We are interested in the magnitude of  $\frac{\partial s_j}{\partial s_{j-1}}$  ← if it is small (large)  $\frac{\partial s_k}{\partial s_l}$  and hence  $\frac{\partial \mathcal{L}}{\partial W}$  will vanish (explode)

⏪ ⏩ ⏴ ⏵ 🔍 🔄 35/41

So, we are interested in this particular quantity. So, let us see what this derivative is. And remember that in the final formula we have a product of these quantities. So, I am looking at one such term in my final product. So, just to jog a memory  $a_j$  is the pre-activation which is given by this and then  $s_j$  is the hidden representation after activation after the nonlinearity which is given by  $\sigma$ . So, let me just write it down as  $s_j$  by  $s_{j-1}$  can be written as this chain rule which is first compute  $s_j$  with respect to  $a_j$  and then  $a_j$  with respect to  $s_{j-1}$ . Everyone has found. So, far at this point please raise your hands if you find ok.

Now, let me just write down  $a_j$  and  $s_j$  explicitly. So, remember that  $a_j$  is this  $d$  dimensional vector which are the entries  $a_{j1}, a_{j2}, \dots, a_{jd}$  and  $s_j$  is the corresponding activation applied vector which has these entries  $\sigma(a_{j1}), \sigma(a_{j2})$  and so on ok. Now, first question what is this quantity? Scalar? Vector? Metric? Tensor? Numerator is a.

Student: (Refer Time: 02:44).

Denominator is a.

Student: (Refer Time: 25:45).

That is why it is a matrix ok. So, that is the matrix that I am interested in. If I can give you that matrix and we are kind of done so, it help me filling in this matrix.

Tell me what this matrix is going to look like even before we start filling it ok. You are right, but it does not matter because you will have  $U^x$  and then you are taking the derivative with respect to  $s_j - 1$ , right so, this does not matter ok. So, everyone gets that you will have a  $U^x_j$  here, right but that does not matter because you are taking a derivative with respect to  $s_j$  so, that is a constant.

So,  $\frac{d}{ds_j} \frac{d}{da_j}$  is what? What does this matrix look like? How many of you see a diagonal matrix? Ok good so, it is straightforward, right. What is the first entry it is going to be  $\frac{d}{ds_1} \frac{d}{da_1}$  by  $\frac{d}{ds_1} \frac{d}{da_1}$ , what is that going to be? It will be something. But let us look at the second entry  $\frac{d}{ds_2} \frac{d}{da_1}$  what is this going to be? What this going to be?

Student: 0.

0, because it does not depend on that, right. So, now, you can see how the full matrix will look like all the off-diagonal elements are going to be 0s and diagonal elements are going to be  $\sigma_j$  everyone fine with this ok. So, this matrix I am going to just call it as diagonal  $\sigma_j$  this is a diagonal matrix which I have. And what is  $\frac{d}{ds_j} \frac{d}{da_j}$  by  $\frac{d}{ds_j} \frac{d}{da_j}$ ? Scalar? Vector? Matrix? Scalar.

Student: (Refer Time: 04:17).

Matrix. Which matrix?

Student: W (Refer Time: 04:18).

W, right ok. So, now, for some reason I am interested in the magnitude of this. Why I am interested in the magnitude of this? For some reason I am interested, let us see why. We will become clear that for some reason I am interested in.

(Refer Slide Time: 04:40)

$$\left\| \frac{\partial s_j}{\partial s_{j-1}} \right\| = \left\| \text{diag}(\sigma'(a_j))W \right\|$$
$$\leq \left\| \text{diag}(\sigma'(a_j)) \right\| \|W\|$$

$\because \sigma(a_j)$  is a bounded function (sigmoid, tanh)  $\sigma'(a_j)$  is bounded

$$\sigma'(a_j) \leq \frac{1}{4} = \gamma \text{ [if } \sigma \text{ is logistic]}$$
$$\leq 1 = \gamma \text{ [if } \sigma \text{ is tanh]}$$
$$\left\| \frac{\partial s_j}{\partial s_{j-1}} \right\| \leq \gamma \|W\|$$
$$\leq \gamma \lambda$$
$$\left\| \frac{\partial s_t}{\partial s_k} \right\| = \left\| \prod_{j=k+1}^t \frac{\partial s_j}{\partial s_{j-1}} \right\|$$
$$\leq \prod_{j=k+1}^t \gamma \lambda$$
$$\leq (\gamma \lambda)^{t-k}$$

- If  $\gamma \lambda < 1$  the gradient will vanish
- If  $\gamma \lambda > 1$  the gradient could explode

Mitesh M. Khapra CS7015 (Deep Learning) 36/41

And here I will write how I will write the magnitude of this, right. So, this is the norm that I am interested in. So, I have already said that this is actually equal to whatever is inside this norm. So, I can just write it as this norm so, I have norm of c is equal to norm of a b which is less than equal to.

Student: Norm a, norm b.

Norm a norm b, this is fine ok. Now, let us look at the norm of this. Now, going to say that sigma a j is actually a bounded function because, we are using sigmoid or tan h or something so, it is a bounded function ok. So, that mean sigma dash a j is also going to be bounded actually, can you tell me what is the bound for the logistic function for sigma dash a j.

If sigma is logistic function what sigma dash what is the bound for sigma dash. If I say 1 by 4 how many of you will agree with that? How many of you have a problem with that? If you do not understand this you not understand anything after that ok, still do not have a problem. So, for the logistic function the bound is actually 1 by 4, the maximum derivative that you can get if you have this curve so, then that would be 1 by 4, ok.

What about the tan h function? And that actually happens at this point, right 0.5, so 0.5 into 0.5 is 1 by 4. What about the tan h function? The bound is 1, right. So, this is, this clearly an upper bond on these things the derivative is going to be an upper bounded

thing that means, this magnitude is actually going to be upper bounded by something and I will just call it as  $\lambda$  sorry as  $\gamma$ . So, this quantity is bounded and I am going to call that bound as  $\gamma$ .

What about our weight matrix? It is again bounded, right we have real weights we do not have like blowing we do not have very large weights it is all bounded. So, it is still going to be some upper bound on this and I will call this magnitude as  $\gamma$ , right. So, this quantity on the left hand side, I can say that it is less than equal to some  $\gamma$  into  $\lambda$ .

Now, let us look at the product. So, this is a quantity that I was interested in and this is actually a product of various such quantities. So, what is it going to be now? Can you go to the next step? It will be  $\gamma$  into  $\lambda$  raise to  $t$  minus  $t$  minus  $k$ , right,  $t$  minus. It basically as  $t$  minus this product as  $t$  minus  $k$  terms, right. So, it will be  $\gamma \lambda$  raise to  $t$  minus  $k$ . Now, if  $\gamma$  or  $\lambda$ , or rather  $\gamma$  into  $\lambda$  if it is greater than 1, what will happen? What will happen to the series? Explore. If it is less than 1?

Student: (Refer Time: 07:20).

It will vanish, right so, you get that. So, that is why you have this vanishing an exploding gradients problem ok. But why what if this vanishes what vanishes? Let us go back. So, I have shown you that this quantity could vanish right if this vanishes the entire gradient could vanish. And if the gradient vanishes what would happen?

(Refer Slide Time: 07:39)

$a_j = [a_{j1}, a_{j2}, a_{j3}, \dots, a_{jd}]$   
 $s_j = [\sigma(a_{j1}), \sigma(a_{j2}), \dots, \sigma(a_{jd})]$

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \dots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \dots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \dots & \\ 0 & 0 & \dots & \sigma'(a_{jd}) \end{bmatrix}$$

$= \text{diag}(\sigma'(a_j))$


- We are interested in  $\frac{\partial s_j}{\partial s_{j-1}}$

$$a_j = W s_{j-1} + b$$

$$s_j = \sigma(a_j)$$

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j} \frac{\partial a_j}{\partial s_{j-1}} = \text{diag}(\sigma'(a_j))W$$

- We are interested in the magnitude of  $\frac{\partial s_j}{\partial s_{j-1}}$  ← if it is small (large)  $\frac{\partial s_k}{\partial s_l}$  and hence  $\frac{\partial \mathcal{L}}{\partial W}$  will va



Mitesh M. Khapra CS7015 (Deep Learning)

Student: No updates.

No updates and you just stuck where you are. If the gradient explodes what happens? Think in terms of the WB plane, you suddenly have a very large gradient what will happen is just gone way far from where you are right. Now, because your update is  $W$  is equal to  $W$  minus eta into this gradient and this you have got a very large value.

Now, I just going to move somewhere very far from where you are and that is never go where your suddenly jump to a different universe ok. So, that is the problem in training recurring neural networks. You could have this problem of exploding or vanishing gradients, and we have done a mathematical derivation of why you have this problem, ok.

(Refer Slide Time: 08:16)

- One simple way of avoiding this is to use truncated backpropagation where we restrict the product to  $\tau(< t - k)$  terms

NPTEL  
Mitesh M. Khapra CS7015 (Deep Learning)

So, one trick to do that is to avoid this is remember these are  $t$  minus  $k$  terms and the problem appears when your  $t$  minus  $k$  is or rather you are  $t$  is close to capital  $T$ , angle  $k$  is closed to 1, right. In those cases you will have many terms in the product you will have as many as  $T$  terms in the product, so even if your product is even if this product is slightly less than 1, if you raise it to capital  $T$  it is going to vanish, right. So, can you think of solution for this?

And the last module in the title of this lecture was truncated back propagation. Can you think of a solution for this? So, you do not back propagate through all the time steps yes, use an approximation that if you are at time step  $n$ . We are just going to look at  $n$  minus  $k$  time steps and we are not going to look all the way back, right that is the common trick used to avoid exploding and vanishing gradients.

What is the other thing that you could do to avoid exploding gradients? So, remember that you have some gradient, right. To think in terms of vectors we have some gradient vector  $W$  whose magnitude is very large, what will you do to avoid exploding gradients? In gradient descent you are always interested in the direction so, what can I do?

Student: (Refer Time: 09:35).

Just normalize it, right. So, you can just do this so, typically what is done is that you can it is a normalizing it you can just say that you will clip the gradient so that it is



magnitude is less than a certain  $k$ , right. So, normalize it in such a way that if its magnitude is greater than  $k$ , its magnitude becomes  $k$ . So, this is something typical that you will see when you use tensor flow where you have something with `tf.nn.clip` the gradients to a certain magnitude. And there are different ways of doing this, so I just give you an intuition that this is what is used for magnitude but there are other things that you can use for magnitude. So, just go back and look at that ok.

So, that is a back propagation through time with exploding and vanishing gradients and then the solution for that or a part for that is truncated back propagation ok. We have not yet done with this problem, we will again look at other solutions for handling this which will lead us to LSTMs which is Long Short Term Memory cells and gated recurrent units, so that we will do in the next lecture.