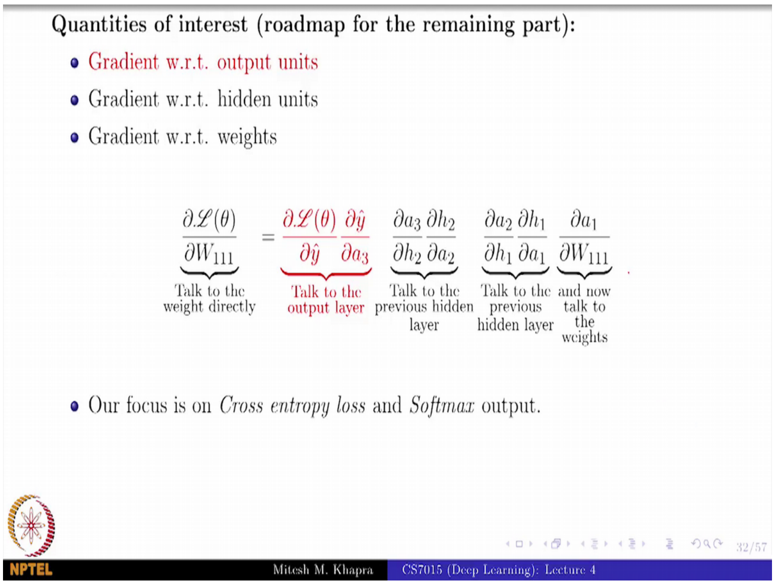
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 4.5
Lecture - 04
Backpropagation: Computing Gradients w. r. t. the Output Units

Now, we go to the next module where we will first see how to Compute the Gradient with respect to the Output Units, well that was the first guy in our chain right that is the first person that we need to talk to ok.

(Refer Slide Time: 00:23)



Quantities of interest (roadmap for the remaining part):

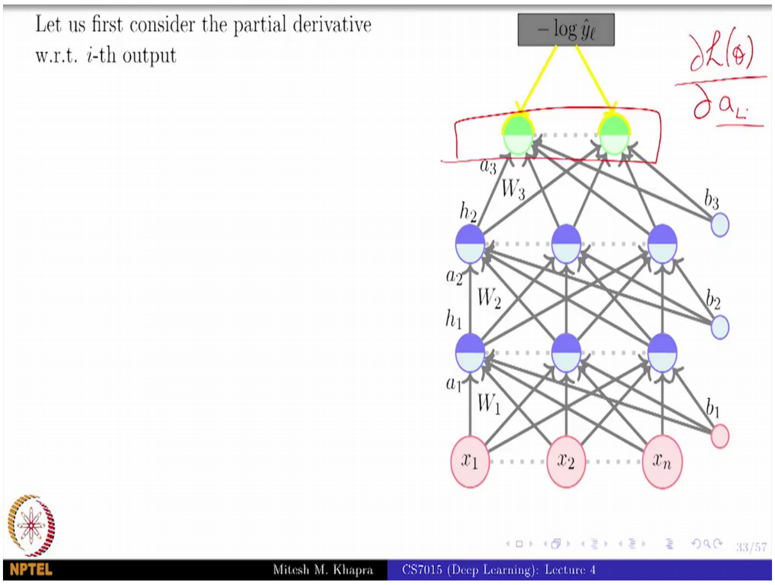
- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\text{Talk to the weight directly}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\text{and now talk to the weights}}$$

- Our focus is on *Cross entropy loss* and *Softmax* output.

So, that is the part that we are going to focus on.

(Refer Slide Time: 00:26)



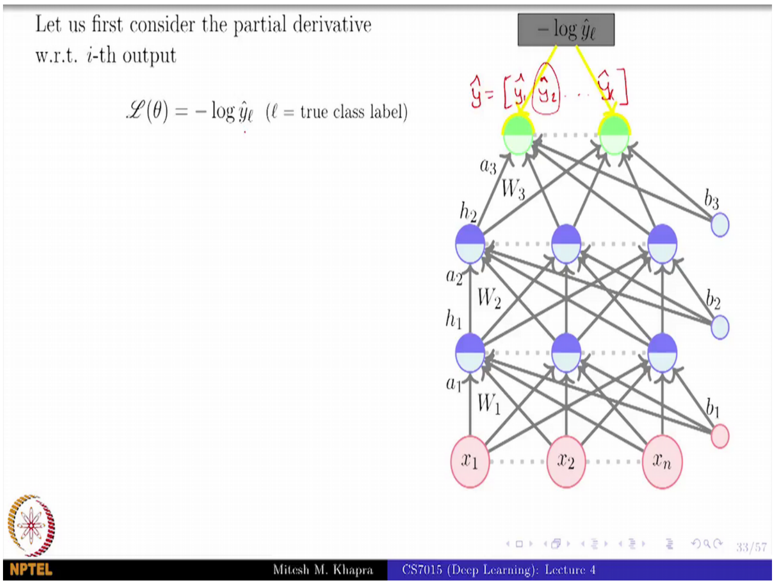
Let us first consider the partial derivative w.r.t. i-th output

$-\log \hat{y}_\ell$

$\frac{\partial \mathscr{L}(\theta)}{\partial a_{L_i}}$

So, this is the output and when I say I want to compute the gradient with respect to output unit, what do you actually mean; what is the quantity that I am looking for? I will help you out, actually what I meant by output unit is this entire thing right. So, I actually meant aLs ok, but it is it is a fair answer and even y hat is a fair answer ok. In fact, am going to start with y hat and then go to here. So, I will have to start with this guy and then come to this guy right.

(Refer Slide Time: 00:54)



Let us first consider the partial derivative w.r.t. i-th output

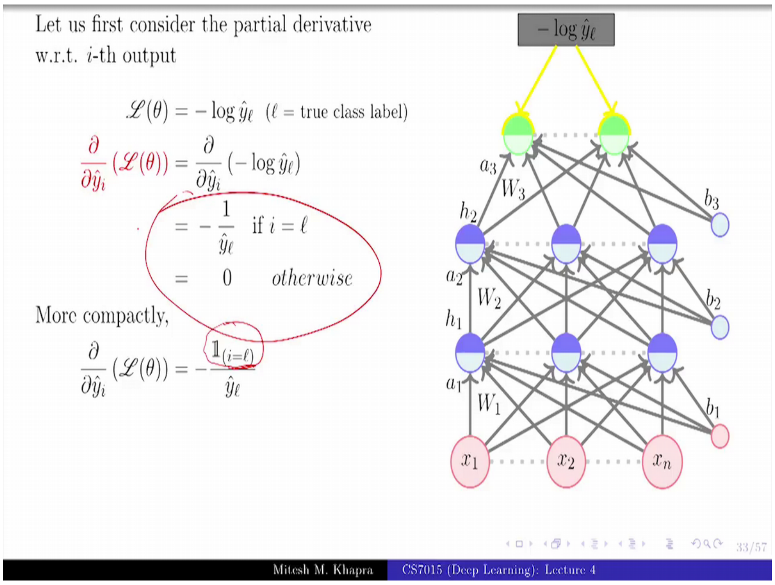
$\mathscr{L}(\theta) = -\log \hat{y}_\ell$ (ℓ = true class label)

$-\log \hat{y}_\ell$

$\hat{y} = [\hat{y}_1 \hat{y}_2 \cdots \hat{y}_k]$

So, this is the loss, this is y hat which is equal to y 1 hat, y 2 hat up to yk hat. So, these are the k values that we have here and we are looking at cross entropy; that means, we are looking at the classification problem, right? So, we have got a distribution over the k classes; that is what y hat looks like. And we know that one of these guys is the right class maybe say y 2. So, the loss function is minus log of y hat 2, because 2 is the correct class in this toy example that I am considering ok. So, the loss function I am just repeating the definition right that is how the loss function is ok.

(Refer Slide Time: 01:30)



Let us first consider the partial derivative w.r.t. i-th output

$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}(\mathscr{L}(\theta)) = \frac{\partial}{\partial \hat{y}_i}(-\log \hat{y}_\ell)$$

$$= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell$$

$$= \quad 0 \quad \textit{otherwise}$$

More compactly,

$$\frac{\partial}{\partial \hat{y}_i}(\mathscr{L}(\theta)) = -\frac{\mathbb{1}_{(i=\ell)}}{\hat{y}_\ell}$$

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

Now, oh god so again this is what our y hat looks like ok. Now I want to compute the gradient with respect to any of the output units right. So, it could be y 1, y 2, y 3, y 4 up to yk right. So, this i actually can take values from 1 to k, in this case 1 to 2 right ok. Now can you tell me what is this loss? Ok this much is fine; can you tell me what is this derivative? Minus 1 by minus 1 by y hat L if y is equal to L.

Student: (Refer Time: 02:08).

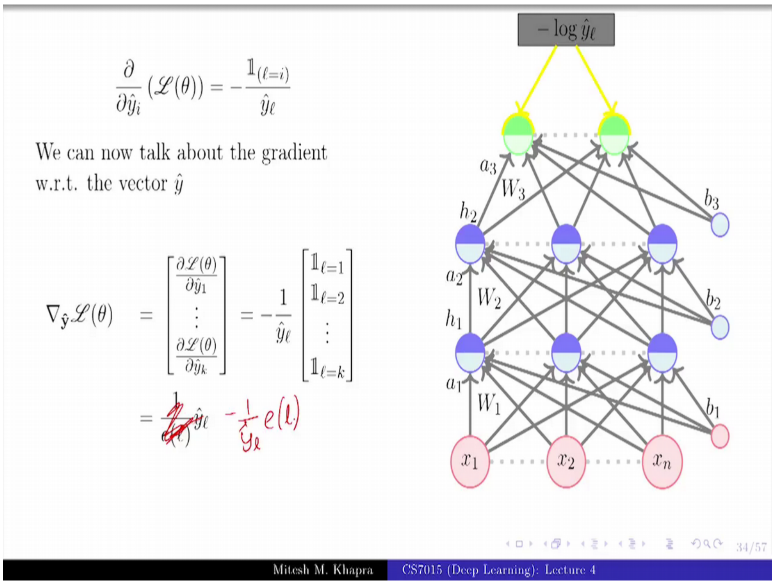
And 0 otherwise, how many of you get that? Cool ok, so it is a very simple thing that you can think of this as z and this is y, only if z is equal to y then the derivative would exist otherwise it is going to be 0 right ok. So, how do I write this fe part using?

Student: (Refer Time: 02:27).

How many of you have seen indicator variables before good? So, this is what you are telling me right, it is going to be minus 1 by y hat l, if i is equal to l ok. And if i is not equal to l, then these 2 things are not related, it this is a function of something else and you are taking a derivative with respect to a different quantity.

So, it is a constant with respect to that constant e and the answer would be 0 ok. Now I am going to write this as this right. So, this is the same as saying so this variable actually this is known as the indicator variable, it takes on the value 1. If the condition in the bracket holds, otherwise it takes on the value 0. So, this is exactly I am writing exactly this, but in a more compact manner ok, is that clear to everyone? Ok.

(Refer Slide Time: 03:15)



$$\frac{\partial}{\partial \hat{y}_i}(\mathscr{L}(\theta)) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(0)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}\begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \\ \mathbb{1}_{\ell=k} \end{bmatrix}$$

$$= \frac{1}{\cancel{e}(\ell)}\hat{y}_\ell \quad -\frac{1}{y_\ell}e(\ell)$$

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

So, this is what the quantity this is the quantity, that we have computed with respect to one of the output units ok. So, this is what; derivative, partial derivative gradient, how many of you say derivative? No one likes derivative; partial derivative? That is always the safest choice partially [FL] right and gradient oh; there is one brave soul who say is gradient do not worry well fix that ok.

So, this is the partial derivative y because my y hat is actually a vector, and I am taking the derivative with respect to one of those guys ok. Now if I want the gradient with respect to y hat, what would that look like? A vector which is a collection of?

Student: (Refer Time: 04:01).

Partial derivatives ok. So, let us see this is the quantity that I am interested in am interested in the gradient of the loss function with respect to the vector y hat. So, remember the vector y hat is y 1 hat y 2 hat up to yk hat, right? So, this gradient is going to be a collection of the partial derivatives with respect to y 1 hat y 2 hat and so on right ok.

Now, what is each of these quantities, how many of you are fine with this? How many of you not fine in this? I did not see as many hands that has going to play ok, how many were fine with it again please raise your hands? Up up up fine ah, is this right? So, it is simple right, so this quantity the derivative is either going to be 0 or is it going to it is going to be 1 by y 1 hat, right? If l is equal to 1 right? And that is exactly what I have done. So now, how many elements here are actually going to be nonzero? At a time how many of these going to be nonzero? 1, which one?

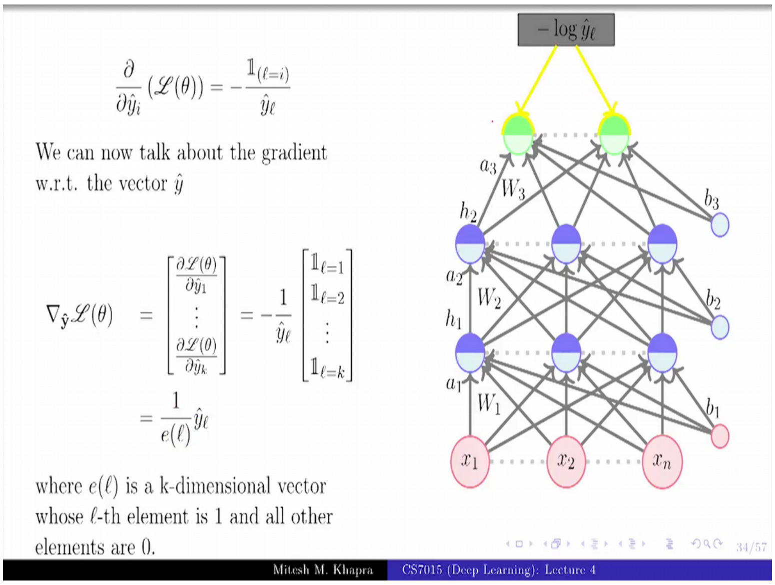
Student: (Refer Time: 05:04).

The 1 corresponding to l, right? Everything else is going to be 0. So, this is a dash vector y not vector ok. So now, am going to write 1 hot vector like this, what have we done? Ok where el is what 1 hot vector, such that; it is l th entry is 1 ok, that is what am going that is how am going to define e l, is that fine with everyone ok?

And so, you see the story how did how we went about computing this. We started with a partial derivative with respect to 1 of t guys right we found a formula for y i; we saw that this formula is generic enough. And So now, we can compute the gradient which is a collective of all these yis where i ranges from 1 to k, right? And then we just put that in a gradient vector.

So, this story is going to repeat throughout the lecture where we try to compute the gradient with respect to 1 guy and then generalize oh sorry, we compute the partial derivative with respect to 1 guy and then generalize and try to find the gradient fine ok.

(Refer Slide Time: 06:18)



$$\frac{\partial}{\partial \hat{y}_i}(\mathscr{L}(\theta)) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}\begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \\ \mathbb{1}_{\ell=k} \end{bmatrix}$$

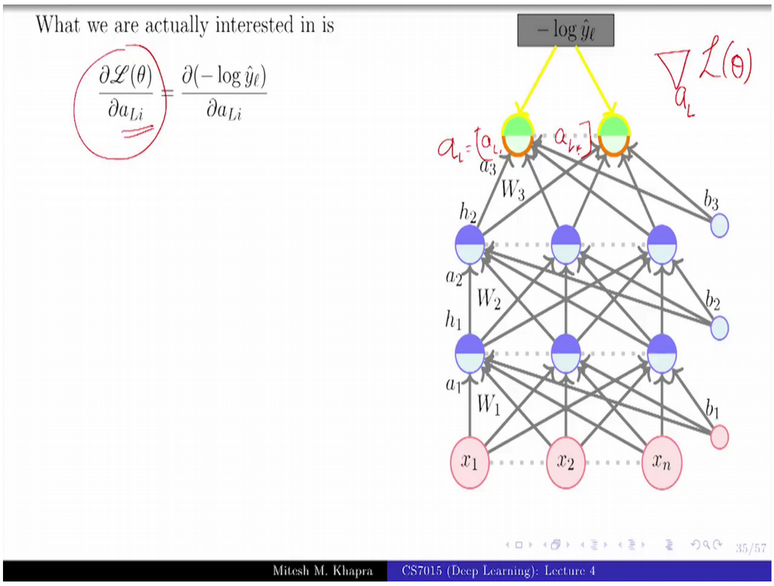
$$= \frac{1}{e(\ell)}\hat{y}_\ell$$

where $e(\ell)$ is a k-dimensional vector whose ℓ-th element is 1 and all other elements are 0.

34/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

(Refer Slide Time: 06:19)



What we are actually interested in is

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$

$\nabla_{a_L} \mathcal{L}(\theta)$

$a_L = [a_{L1} \quad a_{Lk}]$

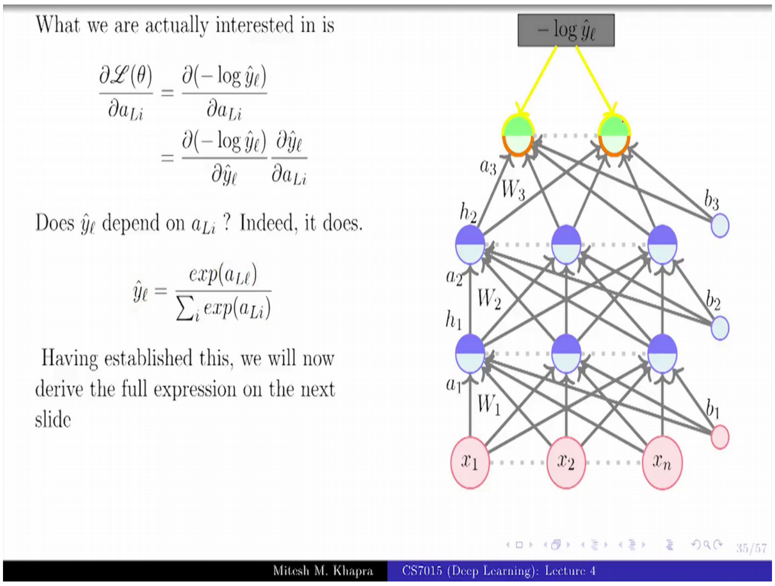
35/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

.

So, what if I what do I have so far? I have this quantity. What does till which part of the diagram am I currently? The dash green part dark green part I am till, here I need to go till the light green party that is collectively the output unit ok. Although I have divided into 2 halves, but when I say output unit I mean that output neuron right complete neuron ok. So, what I am actually interested in is these quantities all more specifically ok. This is what I am interested in, what is this? One of those guys right this aL is actually a l 1 up to aL k right. So, this is one of those guys; so, this is going to be the gradient or this is going to be the derivative, a partial derivative sorry ok, now what do how do we proceed from here ok?

(Refer Slide Time: 07:26)



What we are actually interested in is

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$
$$= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}$$

Does $\hat{y}_\ell$ depend on a_{Li} ? Indeed, it does.

$$\hat{y}_\ell = \frac{exp(a_{L\ell})}{\sum_i exp(a_{Li})}$$

Having established this, we will now derive the full expression on the next slide

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

35/57

Now, I will again have to compute this, you already know that good, but before that I want you to answer one question right. So, y hat l ok, what is y hat l? It is the output corresponding to the correct class, does it depend on an arbitrary aL i? So, in the previous thing we saw that only when i is equal to l there is a connection. In this case is there a connection always or only when i is equal to l.

Student: (Refer Time: 08:02).

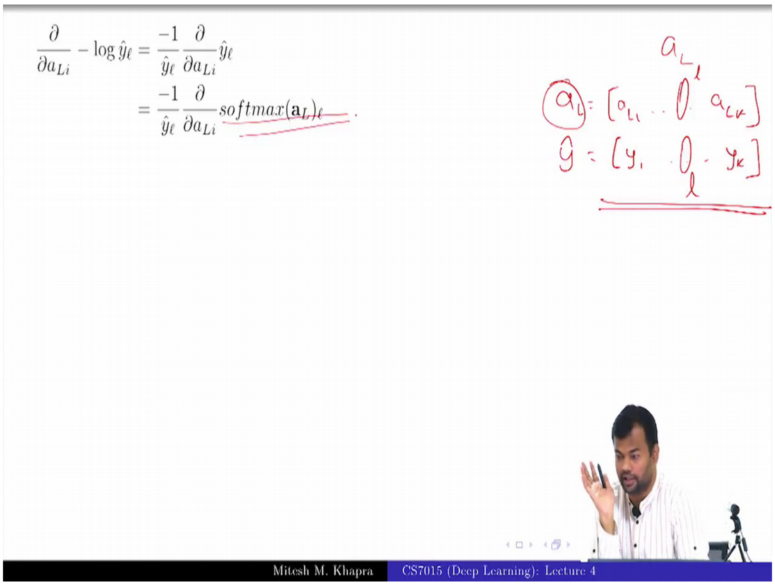
Always why? Softmax so?

Student: (Refer Time: 08:04).

Denominator has all the aLis right. So, this is there it is y hat l in the numerator of course, it only has this unit which corresponds to the l th probably did not choose my variables very well. So, l th component of a capital L right? And but in the denominator you have the entire sum which means; that every output guy here; each of these dark green guys depends on each of the dash green guys light green guys good.

So, that is at least settled that we always the we can always compute this partial derivative, we do not need an if else here there is no thing like l is equal to i then what will happen it will always have this partial derivative is that clear to everyone ok?

(Refer Slide Time: 08:53)



$$\frac{\partial}{\partial a_{Li}} -\log \hat{y}_\ell = \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell$$

$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \underline{\underline{softmax(\mathbf{a}_L)_\ell}}.$$

a_{L_l}

$\boxed{a_L} = [a_{L_1} \dots 0 \; a_{L_k}]$

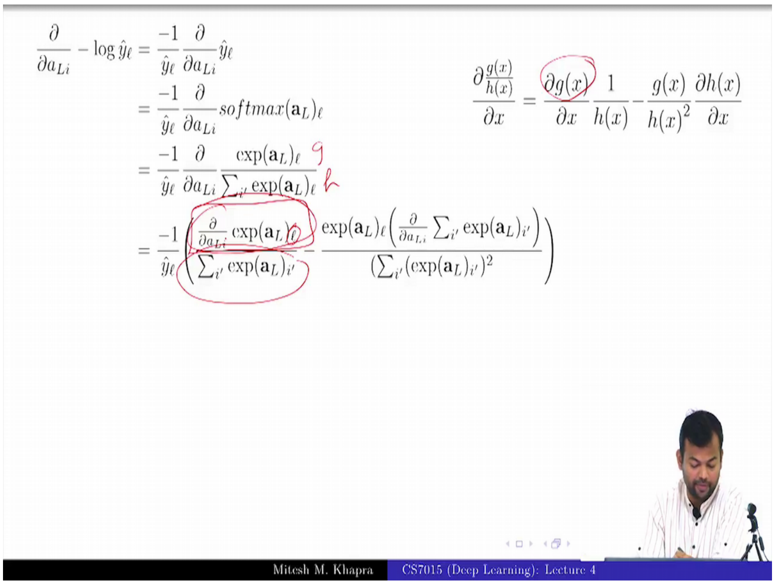
$\underline{\underline{y = [y_1 \; \underset{l}{0} \cdot y_k]}}$

So, we will now derive the full expression for this. So, this is what we are interested in is this fine; so this is a function of the form; so you are taking how do I say this. So, this is log of a function. So, first you will take the derivative with respect to log and then push the partial derivative inside right. So, that would be minus 1 by y hat l and then the derivative with respect to y hat l ok. Now what is y hat l? The softmax function right.

So, it is the l th entry of the softmax function applied to that output vector what is the output vector? aL right. So, it is the l th entry of the softmax or l th entry of the function applied to the output vector is that fine? Everyone gets this? I do not see a lot of thoughtful nodding right.

So, this was our aL what is our output; right? So now, one of these guys here is the l th guy and one of these guys here is the l th guy right. So, what you do is; you take this you apply the softmax function to it which again gives you a vector and now you are interested in the l th component of that vector that is what this quantity means, it should be clear.

(Refer Slide Time: 10:25)



$$\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell = \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell$$

$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} softmax(\mathbf{a}_L)_\ell$$

$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell \quad g}{\sum_{i'} \exp(\mathbf{a}_L)_\ell \quad h}$$

$$= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} (\exp(\mathbf{a}_L)_{i'})^2} \right)$$

$$\frac{\partial \frac{g(x)}{h(x)}}{\partial x} = \frac{\partial g(x)}{\partial x} \frac{1}{h(x)} - \frac{g(x)}{h(x)^2} \frac{\partial h(x)}{\partial x}$$

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

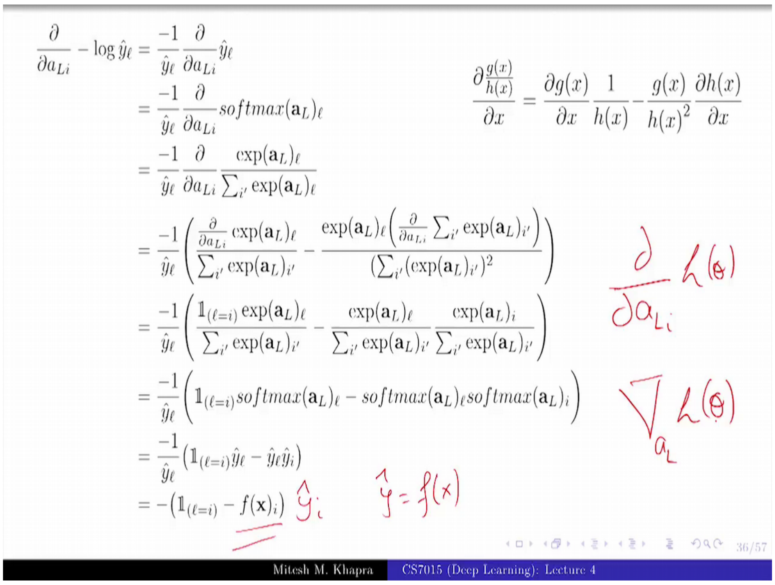
Now, now I will just do some simple math stuff here and we should be able to derive this, is it fine am just replaced by the actual softmax formula? This is a derivative of the form u by v right, what is the formula for that? Yeah it perfectly right yeah. So, this is what it would be right, I mean it is you all know this I am not going to spend time on this right.

So now am just going to substitute the values here, yeah it is getting a bit nasty, but it is not very difficult right. So, so this so this is our g of x so, am taking the derivative of that then this is this 1 over h of x, you can just figure it out right anyway it everyone just read this for a few seconds and let me know if this is not clear. This is g this is h in this formula right have just substituted the gs and hs in the, how many of you get this? Ok, how many if you are still struggling? Ok not ok, if this is clear then the rest of it should be fine. Now what is this quantity going to be? It is derivative of the form e raise to x right so it is e raise to x always.

Student: (Refer Time: 11:33).

If i is equal to l right; so now we have this dependence because we are looking at a numerator, but the numerator only depends on the l th entry right. So now, you are trying to take the derivative of the l th entry with respect to some arbitrary i th entry. So, only if l is equal to i you will get the derivative right.

(Refer Slide Time: 11:52)



$$\frac{\partial}{\partial a_{Li}} -\log \hat{y}_\ell = \frac{-1}{\hat{y}_\ell}\frac{\partial}{\partial a_{Li}}\hat{y}_\ell$$

$$= \frac{-1}{\hat{y}_\ell}\frac{\partial}{\partial a_{Li}} softmax(\mathbf{a}_L)_\ell$$

$$= \frac{-1}{\hat{y}_\ell}\frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'}\exp(\mathbf{a}_L)_{i'}}$$

$$= \frac{-1}{\hat{y}_\ell}\left(\frac{\frac{\partial}{\partial a_{Li}}\exp(\mathbf{a}_L)_\ell}{\sum_{i'}\exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell\left(\frac{\partial}{\partial a_{Li}}\sum_{i'}\exp(\mathbf{a}_L)_{i'}\right)}{(\sum_{i'}(\exp(\mathbf{a}_L)_{i'})^2}\right)$$

$$= \frac{-1}{\hat{y}_\ell}\left(\frac{\mathbb{1}_{(\ell=i)}\exp(\mathbf{a}_L)_\ell}{\sum_{i'}\exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'}\exp(\mathbf{a}_L)_{i'}}\frac{\exp(\mathbf{a}_L)_i}{\sum_{i'}\exp(\mathbf{a}_L)_{i'}}\right)$$

$$= \frac{-1}{\hat{y}_\ell}\left(\mathbb{1}_{(\ell=i)} softmax(\mathbf{a}_L)_\ell - softmax(\mathbf{a}_L)_\ell softmax(\mathbf{a}_L)_i\right)$$

$$= \frac{-1}{\hat{y}_\ell}(\mathbb{1}_{(\ell=i)}\hat{y}_\ell - \hat{y}_\ell\hat{y}_i)$$

$$= -(\mathbb{1}_{(\ell=i)} - \underline{f(\mathbf{x})_i}) \quad \hat{y}_i \qquad \hat{y} = f(x)$$

$$\frac{\partial\frac{g(x)}{h(x)}}{\partial x} = \frac{\partial g(x)}{\partial x}\frac{1}{h(x)} - \frac{g(x)}{h(x)^2}\frac{\partial h(x)}{\partial x}$$

$$\frac{\partial}{\partial a_{Li}} \mathcal{L}(\theta)$$

$$\nabla_{a_L} \mathcal{L}(\theta)$$

36/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

So, that is this correct, ok?

Now, what about this; how many terms in the summation would remain?

Student: (Refer Time: 12:02).

1, which one?

Student: (Refer Time: 12:04).

Where i dash is equal to i right. So, the i th guy would remain the rest of it is straightforward right. This square I have just divided into 2 parts ah, now let us see; can you simplify this? Because I cannot ok, can you simplify this? What is this?

Student: Softmax.

Softmax and which entry of the softmax?

Student: (Refer time: 12:36).

L th entry, i th entry, l th entry with the saw with the indicator variable ok, but what is this? This is our input hidden layer output ok. So, now let us see, what is the next step? Ok this is should have been y hat i, but y hat is equal to f of x right. So, we can fix this unit. So, fine so we have actually what do we have now? We have the derivative of the loss function with respect to the i th unit of the output layer, right? And which part of the output layer? The pre activation pattern ok, now what am I going to do? I have a formula which tells me how to compute this, what was I actually interested in? So now, how am I going to go from here to there? I just put all the partial derivatives into a.

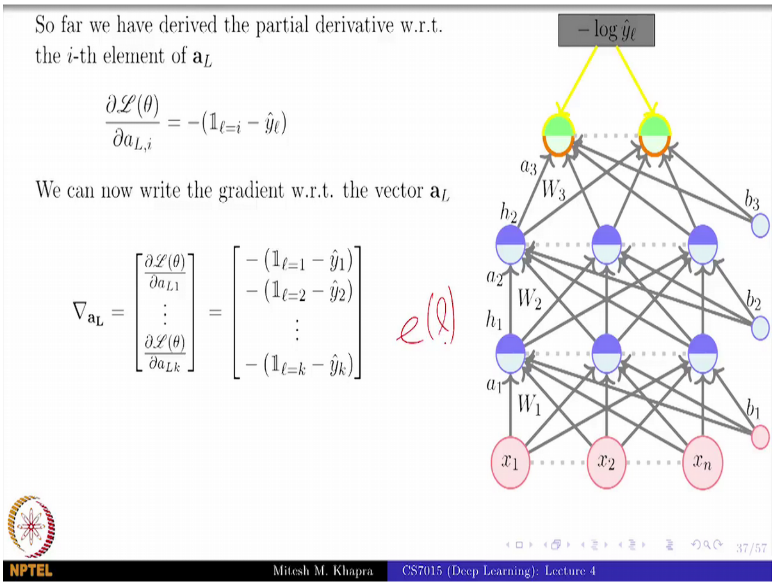
Student: Vector.

Vector, and that vector is the?

Student: (Refer Time: 13:58).

Gradient good.

(Refer Slide Time: 14:00)



So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{L,i}} = -(\mathbb{1}_{\ell=i} - \hat{y}_\ell)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \vdots \\ -(\mathbb{1}_{\ell=k} - \hat{y}_k) \end{bmatrix} \quad e(\ell)$$

NPTEL Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4 37/57

So, we have this one formula it is ok, if some of you did not get this derivation right, it is very, very straightforward. If you go back and look at it I am pretty sure you will get it is nothing in this is, very simple elementary stuff right, except for some degree here and there. So, now what would this look like?

We should add actually l theta here, this would look like a collection of all the partial derivatives. We have a generic formula, what will we do now? What is the first entry? Minus in indicator l equal to 1 minus y hat 1. Which is the variable that we are indexing over i right not l, oh god oh we are indexing or ok, have I goofed up oh that is wrong, is it oh yeah that is wrong fine. Ok then this is fine we are indexing over I and then we can do this.

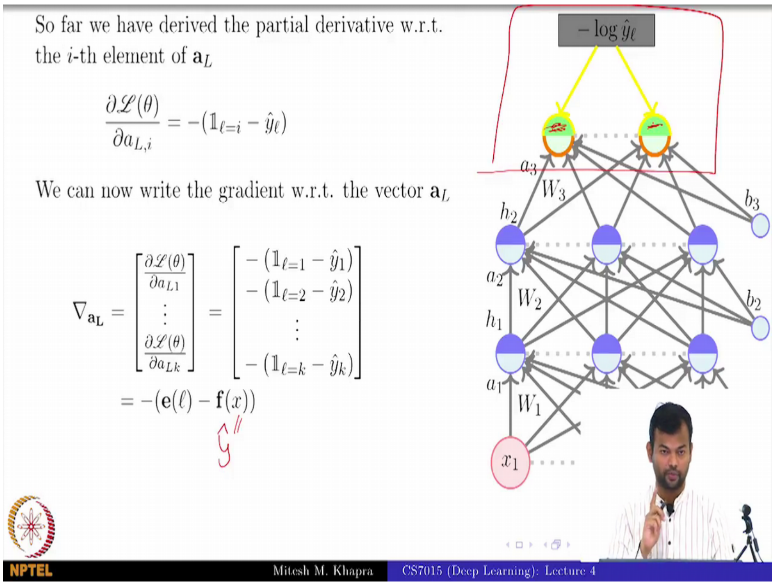
Now, can you simplify this? I am looking for this is the element wise difference of 2.

Student: (Refer Time: 15:38).

Of the indicator vector and.

Student: y hat.

(Refer Slide Time: 15:44)



So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{L,i}} = -(\mathbb{1}_{\ell=i} - \hat{y}_\ell)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \vdots \\ -(\mathbb{1}_{\ell=k} - \hat{y}_k) \end{bmatrix}$$

$$= -(\mathbf{e}(\ell) - \underset{\hat{y}}{\mathbf{f}(x)})$$

Y hat, oh hey we should change all this, y hat is equal to f of x right, but I just want it to be consistent as y hat. So, is this fine this is a simplification fine, right? So, we have come a long way right you have finish this part ok, we have got the gradients with respect to the output units ok. This much part is a clear to everyone moduler bit of the math which you can go back and look at it this entire derivation is fine, but you get the concept right that we start with one unit from there grow the gradient then keep going applying the chain rule right.

So, we started with the dark green guys and then went to the light green guys ok. Now we have the derivative with respect to the entire light green vector ok. And that is what we had started off with that we wanted the gradient with respect to the output units ok. So, that is where will end that module.