

Module 4.5: Backpropagation: Computing Gradients w.r.t. the Output Units

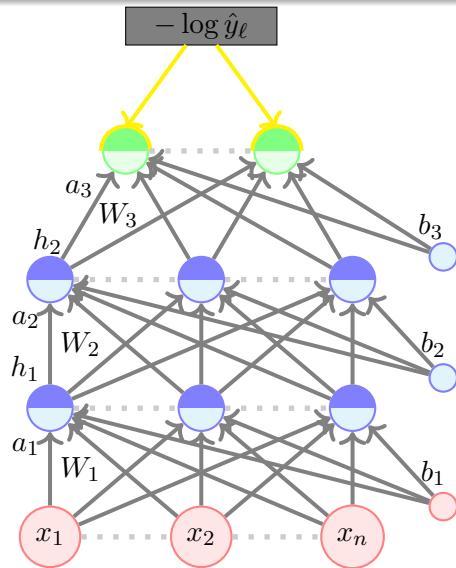
Quantities of interest (roadmap for the remaining part):

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

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

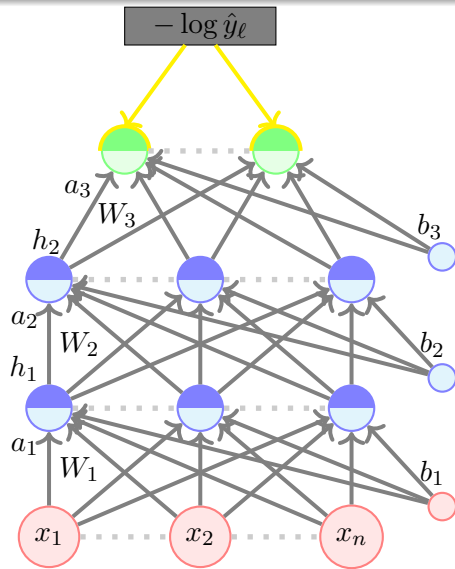
- Our focus is on *Cross entropy loss* and *Softmax* output.

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



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

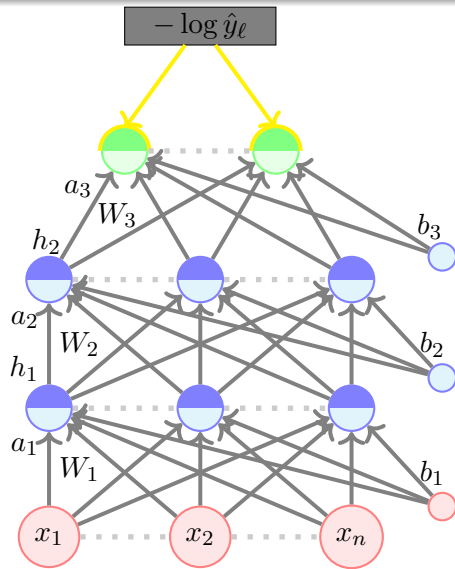
$$\mathcal{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$



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

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

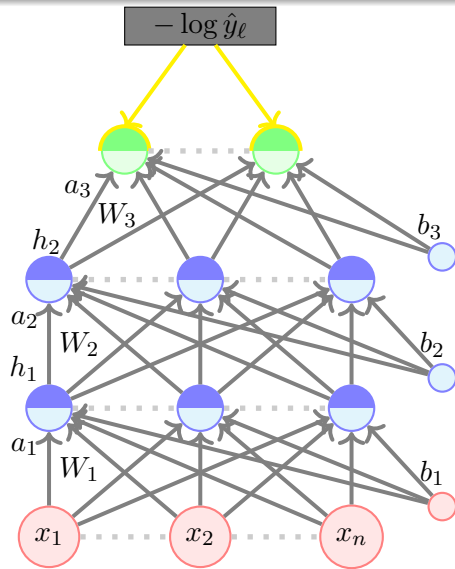
$$\frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) =$$



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

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

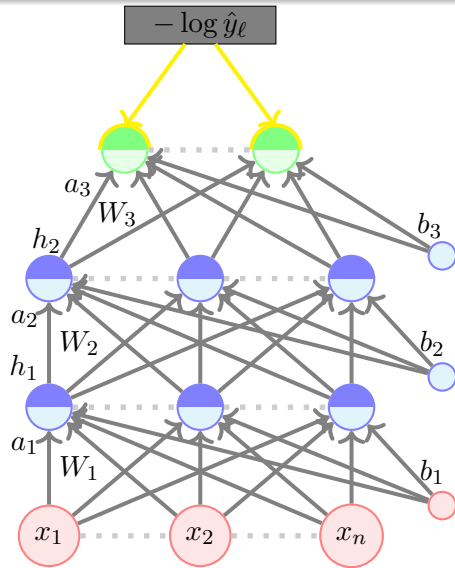
$$\frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) = \frac{\partial}{\partial \hat{y}_i} (-\log \hat{y}_\ell)$$



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

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

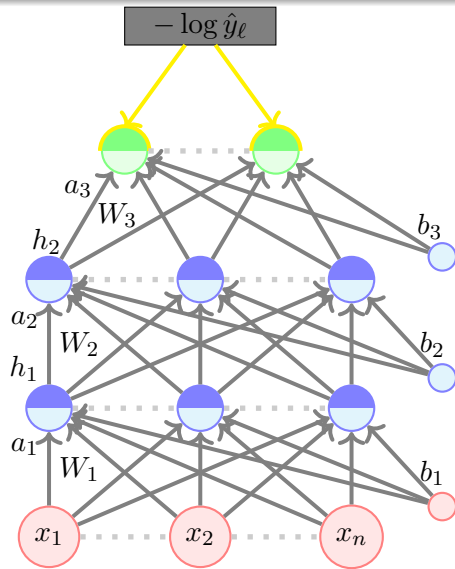
$$\begin{aligned} \frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) &= \frac{\partial}{\partial \hat{y}_i} (-\log \hat{y}_\ell) \\ &= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell \end{aligned}$$



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

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

$$\begin{aligned} \frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) &= \frac{\partial}{\partial \hat{y}_i} (-\log \hat{y}_\ell) \\ &= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell \\ &= 0 \quad \text{otherwise} \end{aligned}$$

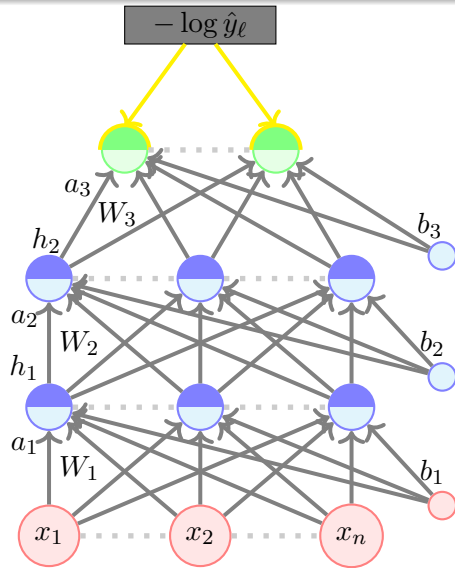


Let us first consider the partial derivative
w.r.t. i -th output

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

$$\begin{aligned} \frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) &= \frac{\partial}{\partial \hat{y}_i} (-\log \hat{y}_\ell) \\ &= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell \\ &= 0 \quad \text{otherwise} \end{aligned}$$

More compactly,



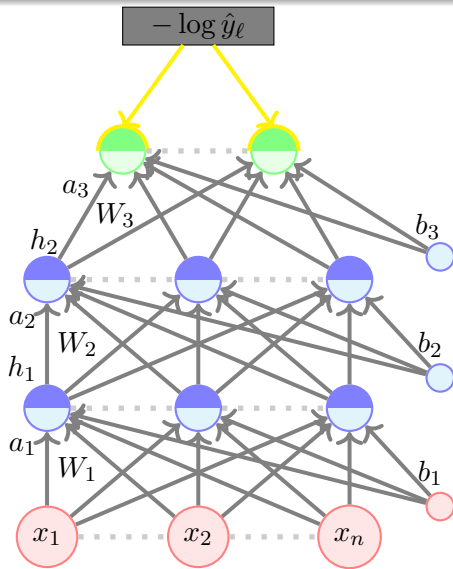
Let us first consider the partial derivative
w.r.t. i -th output

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

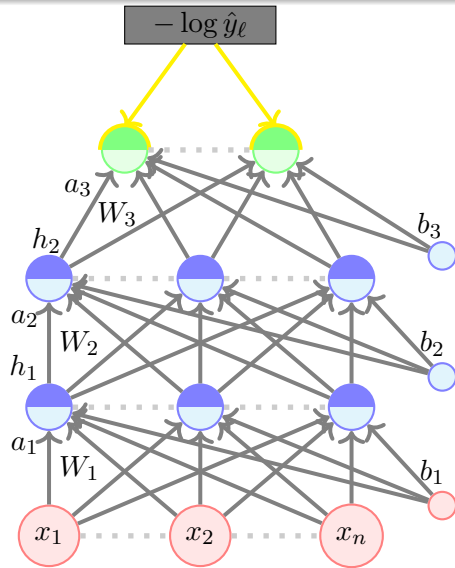
$$\begin{aligned} \frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) &= \frac{\partial}{\partial \hat{y}_i} (-\log \hat{y}_\ell) \\ &= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell \\ &= 0 \quad \text{otherwise} \end{aligned}$$

More compactly,

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

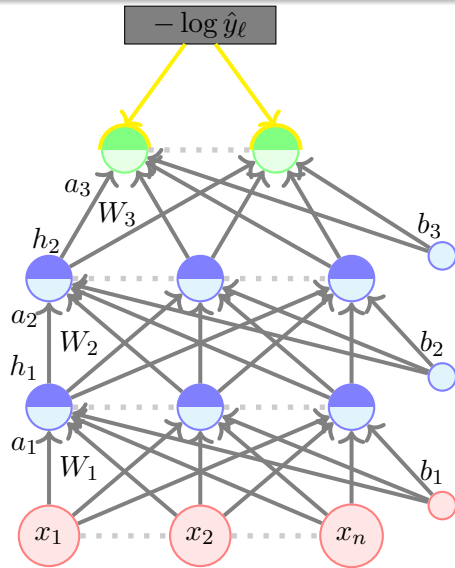


$$\frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$



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

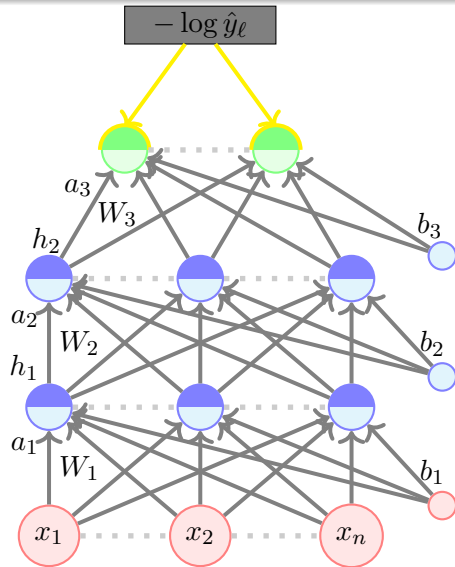
We can now talk about the gradient
w.r.t. the vector \hat{y}



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

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

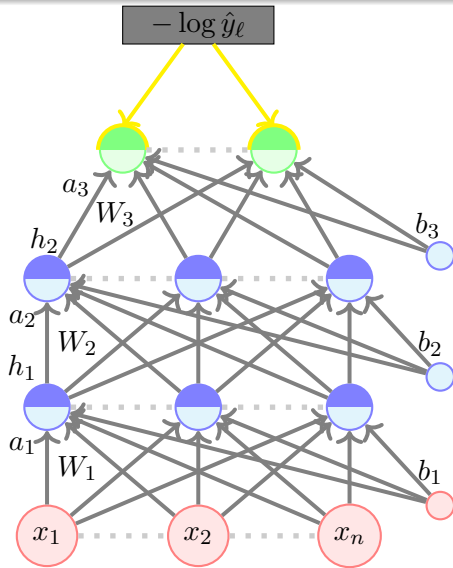
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \\ \\ \end{bmatrix}$$



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

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

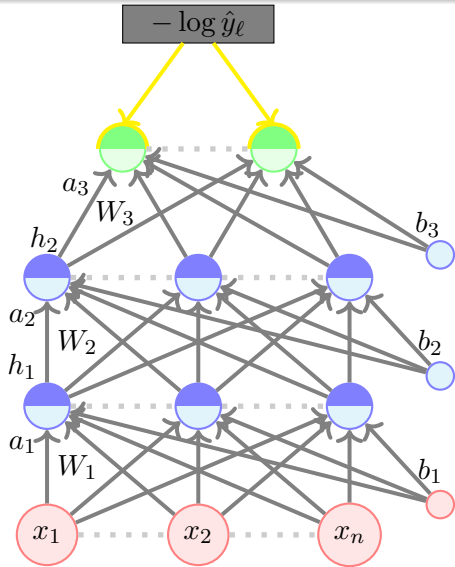
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \end{bmatrix}$$



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

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

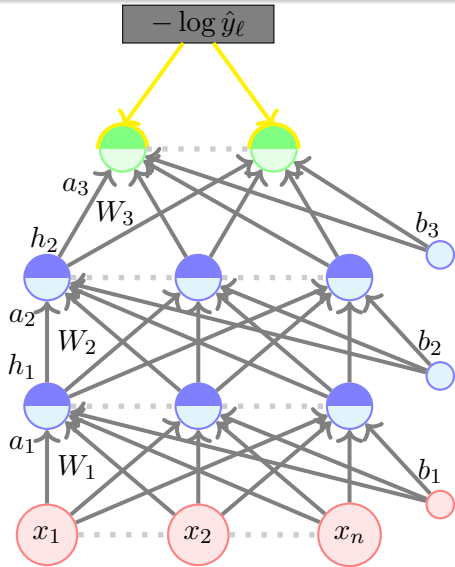
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \end{bmatrix}$$



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

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

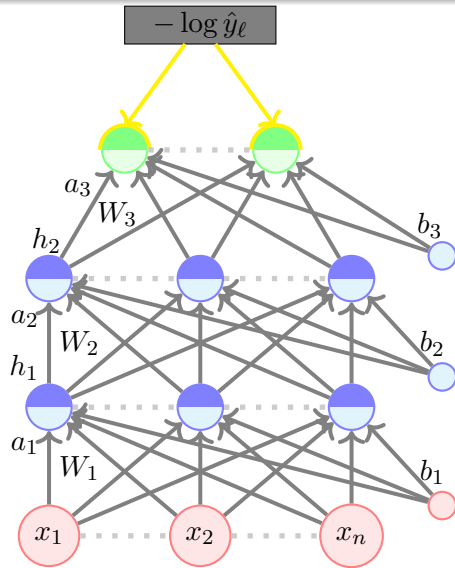
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix}$$



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

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

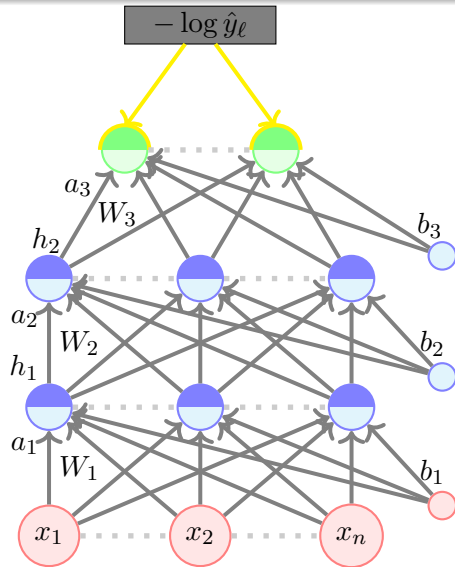
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}$$



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

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

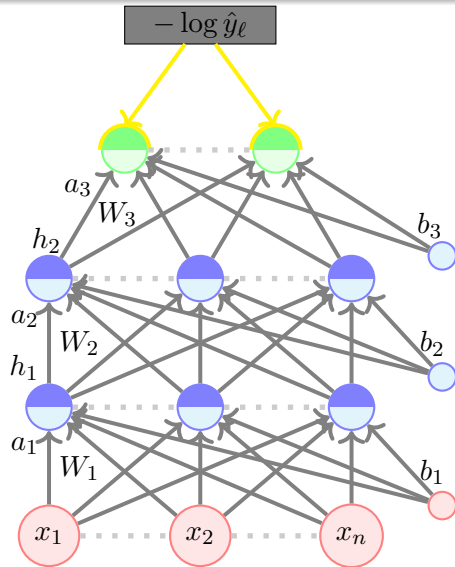
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \\ \\ \end{bmatrix}$$



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

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

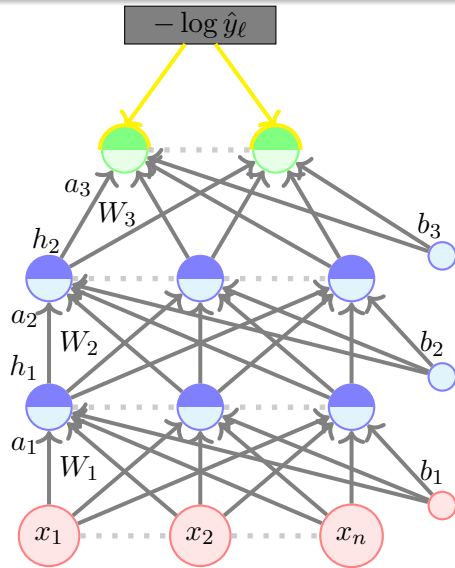
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \vdots \\ \mathbb{1}_{\ell=k} \end{bmatrix}$$



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

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

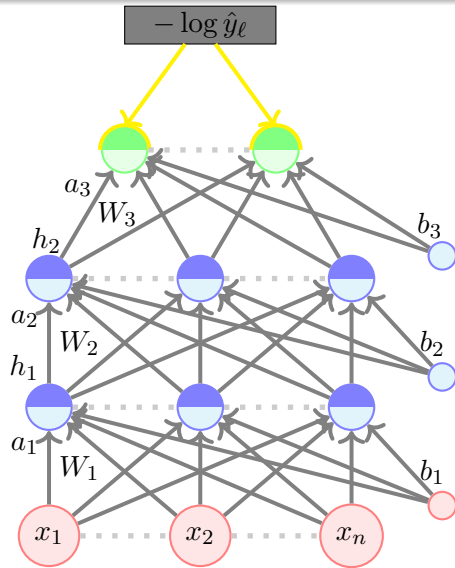
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \end{bmatrix}$$



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

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

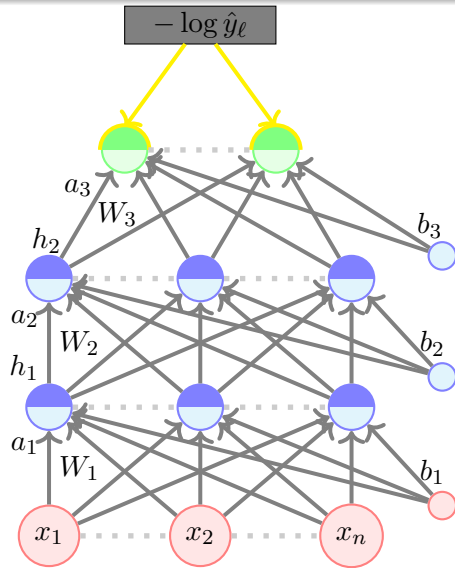
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \end{bmatrix}$$



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

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

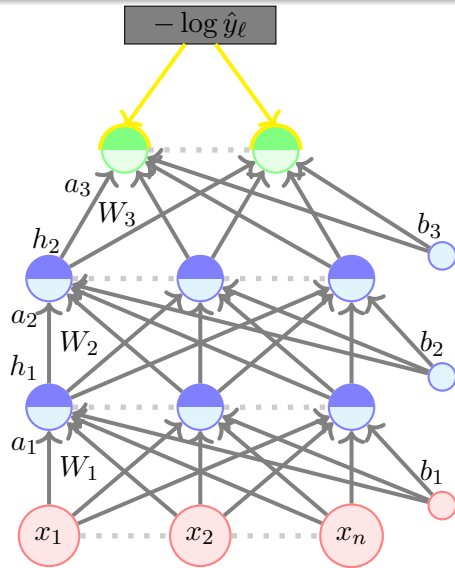
$$\nabla_{\hat{y}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \\ \mathbb{1}_{\ell=k}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



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

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

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

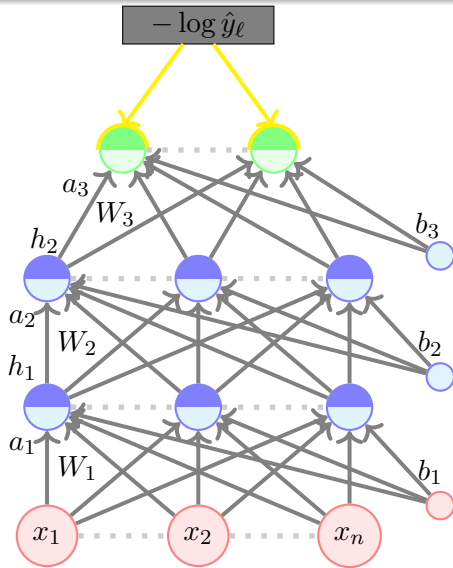


$$\frac{\partial}{\partial \hat{y}_i} (\mathcal{L}(\theta)) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

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

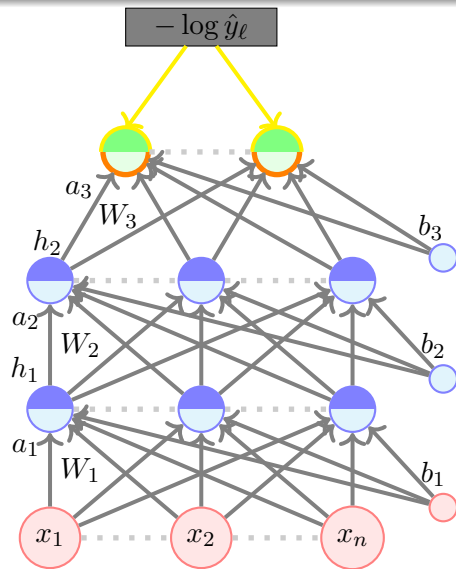
$$\begin{aligned} \nabla_{\hat{y}} \mathcal{L}(\theta) &= \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\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

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



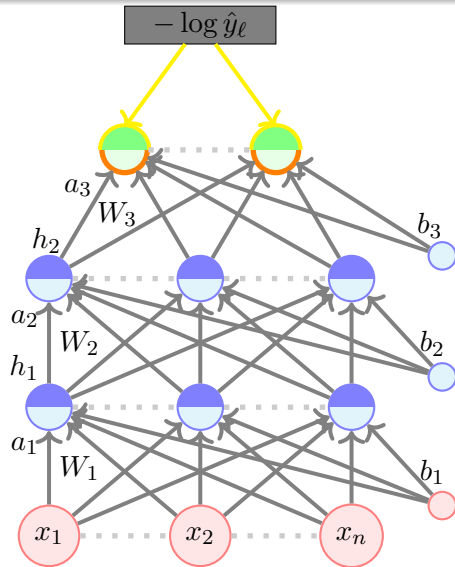
What we are actually interested in is

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



What we are actually interested in is

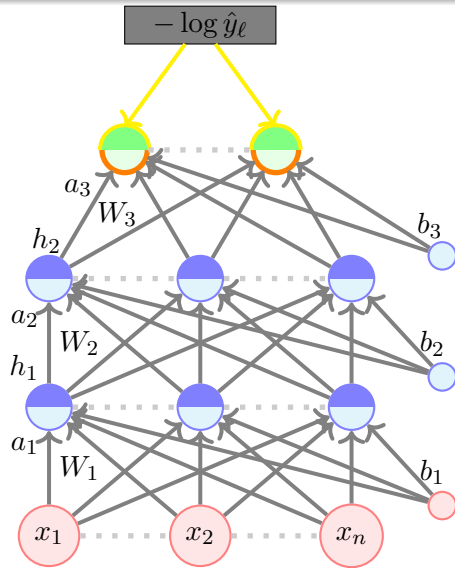
$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial a_{Li}} &= \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}} \\ &= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}\end{aligned}$$



What we are actually interested in is

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial a_{Li}} &= \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}} \\ &= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}\end{aligned}$$

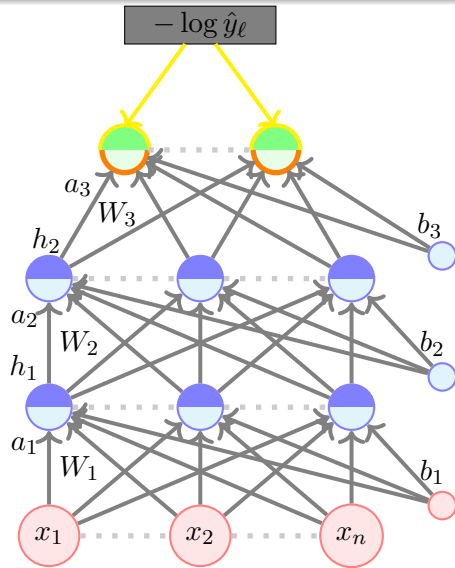
Does \hat{y}_ℓ depend on a_{Li} ?



What we are actually interested in is

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial a_{Li}} &= \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}} \\ &= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}\end{aligned}$$

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

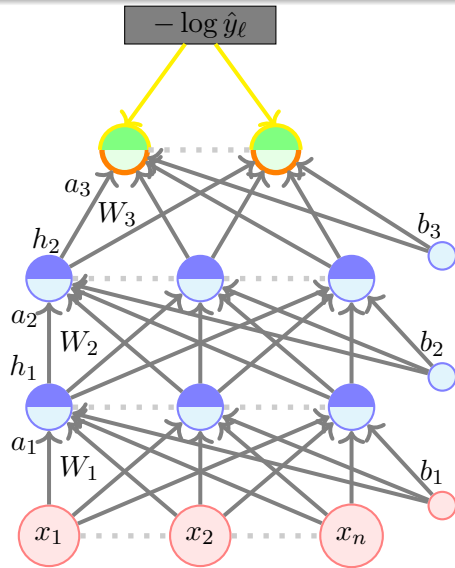


What we are actually interested in is

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial a_{Li}} &= \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}} \\ &= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}\end{aligned}$$

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

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



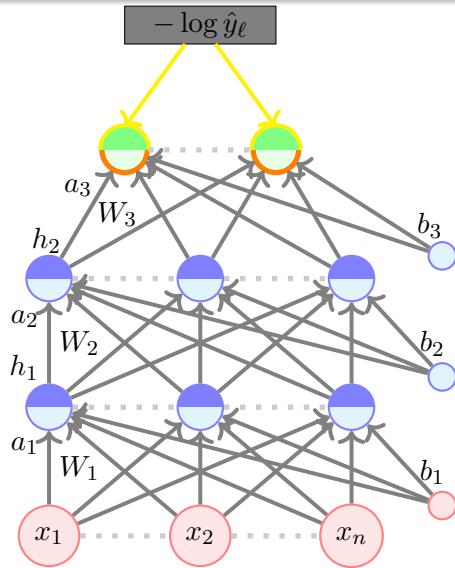
What we are actually interested in is

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial a_{Li}} &= \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}} \\ &= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}\end{aligned}$$

Does \hat{y}_ℓ 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



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

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

$$\begin{aligned}\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\ &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_\ell}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_\ell}
\end{aligned}$$

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

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

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

$$\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} \exp(\mathbf{a}_L)_{i'})^2} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}}
\end{aligned}$$

$$\begin{aligned}
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} (\exp(\mathbf{a}_L)_{i'})^2)} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} \text{softmax}(\mathbf{a}_L)_\ell - \text{softmax}(\mathbf{a}_L)_\ell \text{softmax}(\mathbf{a}_L)_i \right)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} \exp(\mathbf{a}_L)_{i'})^2} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} \text{softmax}(\mathbf{a}_L)_\ell - \text{softmax}(\mathbf{a}_L)_\ell \text{softmax}(\mathbf{a}_L)_i \right) \\
&= \frac{-1}{\hat{y}_\ell} (\mathbb{1}_{(\ell=i)} \hat{y}_\ell - \hat{y}_\ell \hat{y}_i)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \text{softmax}(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}}
\end{aligned}$$

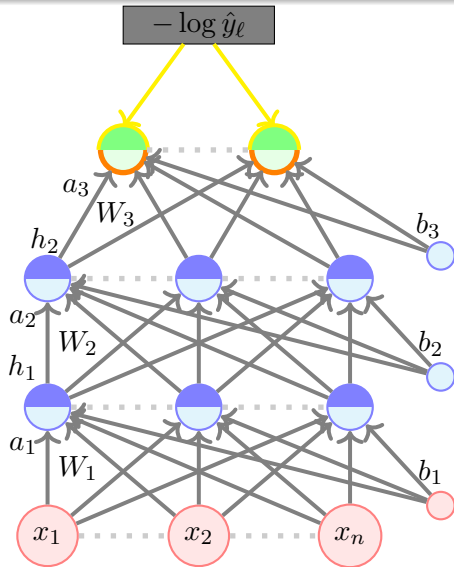
$$\begin{aligned}
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} (\exp(\mathbf{a}_L)_{i'})^2)} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} \text{softmax}(\mathbf{a}_L)_\ell - \text{softmax}(\mathbf{a}_L)_\ell \text{softmax}(\mathbf{a}_L)_i \right) \\
&= \frac{-1}{\hat{y}_\ell} (\mathbb{1}_{(\ell=i)} \hat{y}_\ell - \hat{y}_\ell \hat{y}_i) \\
&= -(\mathbb{1}_{(\ell=i)} - \hat{y}_i)
\end{aligned}$$

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

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

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

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

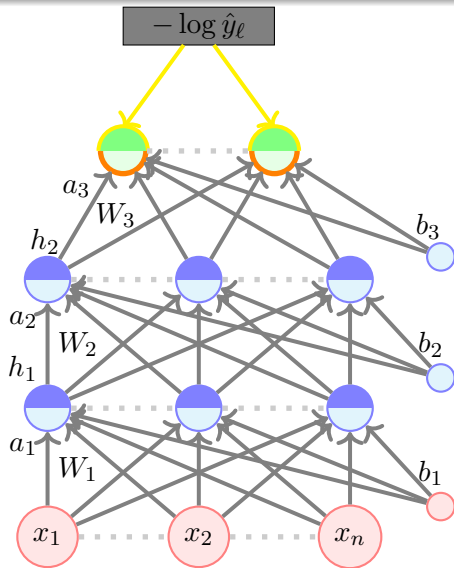


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

$$\nabla_{\mathbf{a}_L} \mathcal{L}(\theta)$$

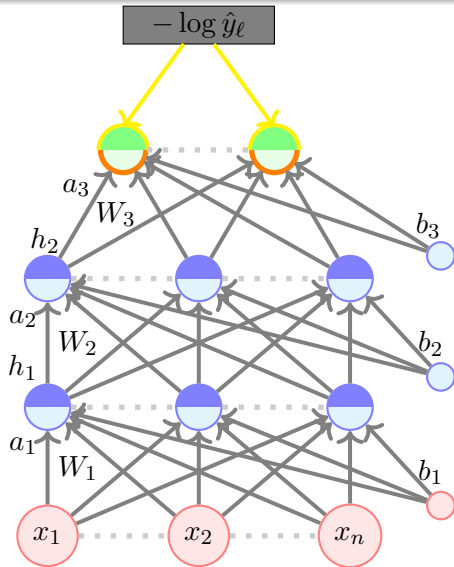


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

$$\nabla_{\mathbf{a}_L} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial a_{L1}} \\ \frac{\partial \mathcal{L}(\theta)}{\partial a_{L2}} \\ \frac{\partial \mathcal{L}(\theta)}{\partial a_{L3}} \end{bmatrix}$$

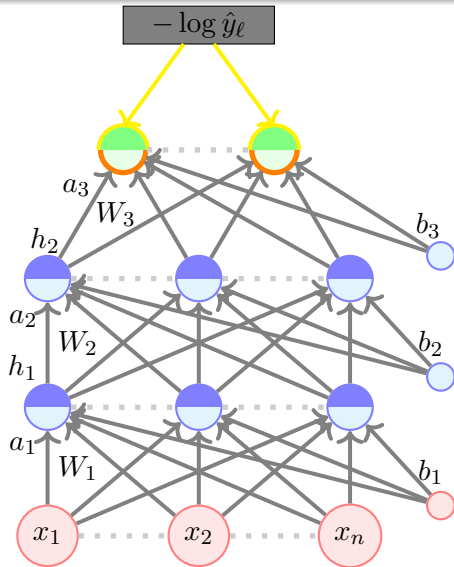


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

$$\nabla_{\mathbf{a}_L} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial a_{L1}} \\ \vdots \end{bmatrix}$$

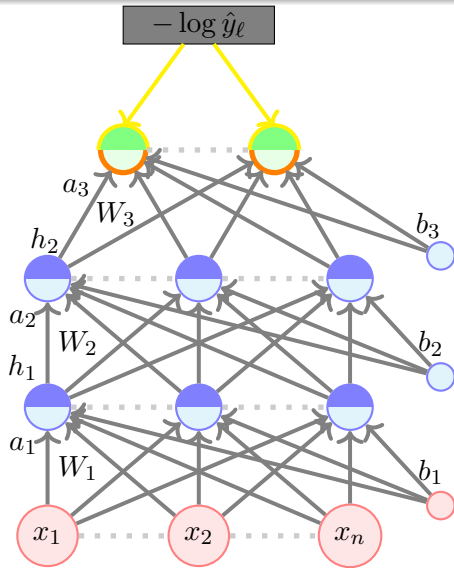


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

$$\nabla_{\mathbf{a}_L} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial a_{Lk}} \end{bmatrix}$$

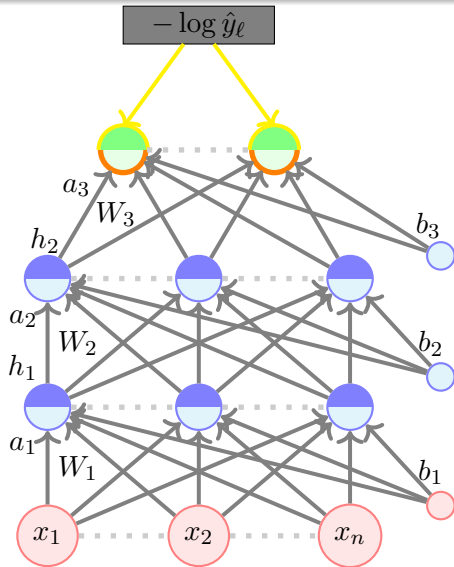


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

$$\nabla_{\mathbf{a}_L} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} \\ \\ \end{bmatrix}$$

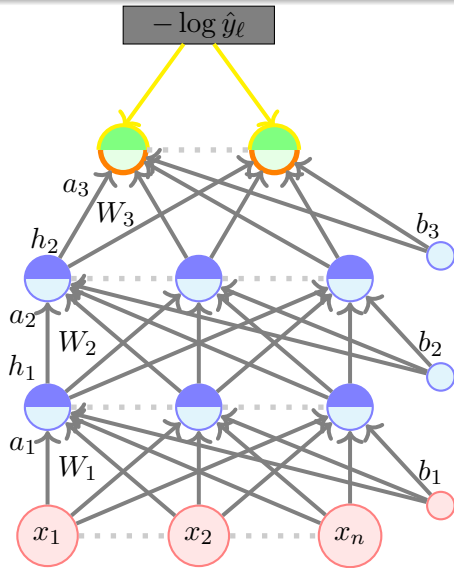


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

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

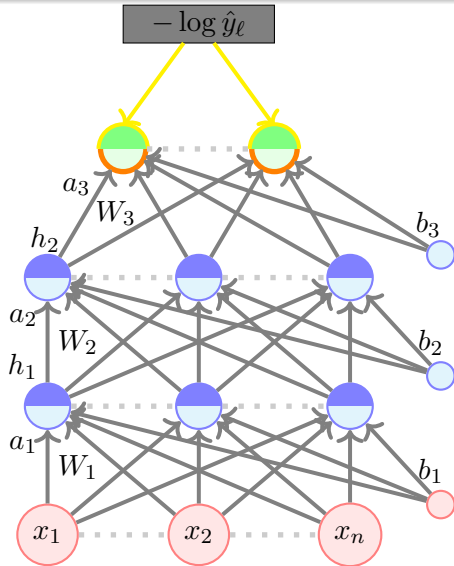


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

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

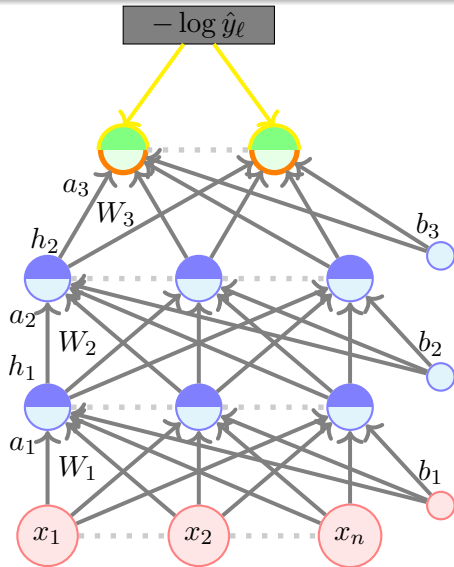


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

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

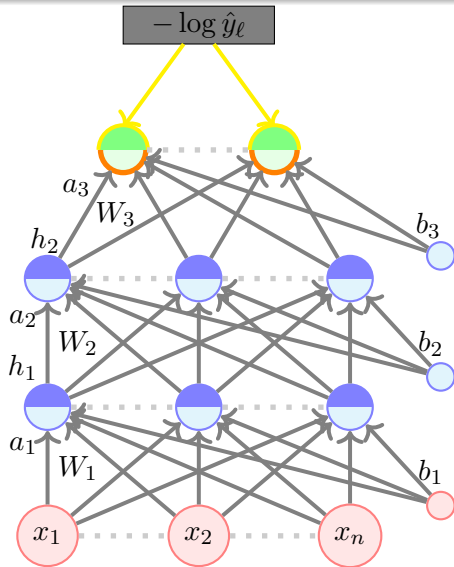


So far we have derived the partial derivative w.r.t. the i -th element of \mathbf{a}_L

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

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

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



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

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

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

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

