Module 4.7: Backpropagation: Computing Gradients w.r.t. Parameters

Quantities of interest (roadmap for the remaining part):

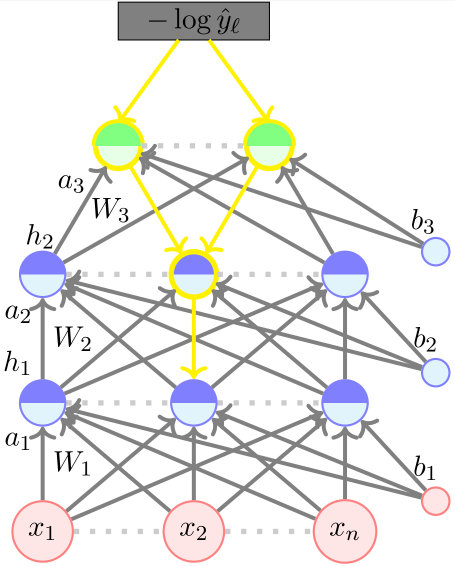
- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{11}}}_{\text{Talk to the weight directly}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_1}{\partial W_{11}}}_{\text{and now talk to the weights}}$$

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

Recall that,

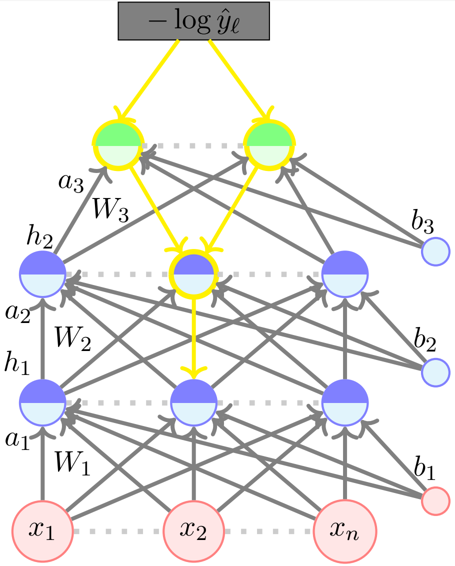
$$a_k = b_k + W_k h_{k-1}$$



Recall that,

$$a_k = b_k + W_k h_{k-1}$$

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

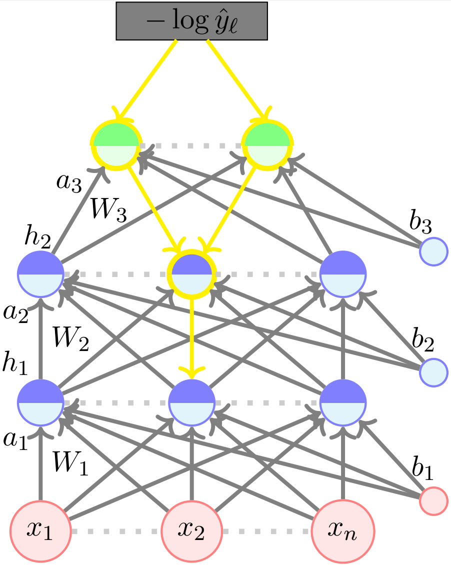


Recall that,

$$a_k = b_k + W_k h_{k-1}$$

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$



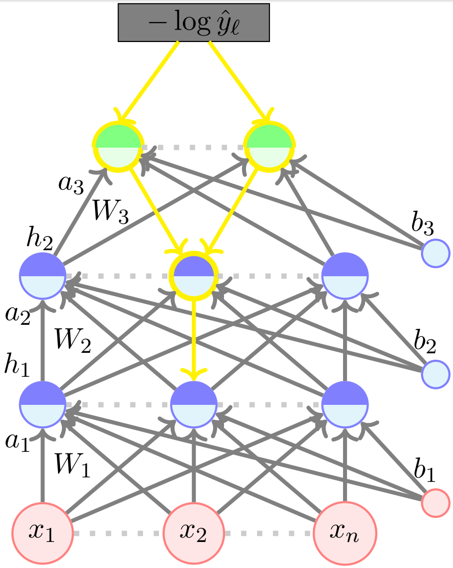
Recall that,

$$a_k = b_k + W_k h_{k-1}$$

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}}$$



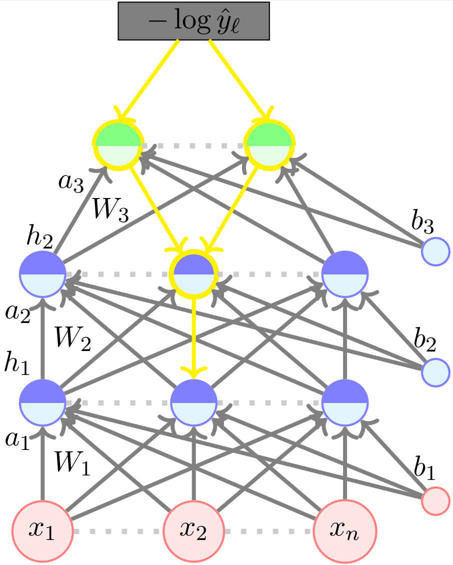
Recall that,

$$a_k = b_k + W_k h_{k-1}$$

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

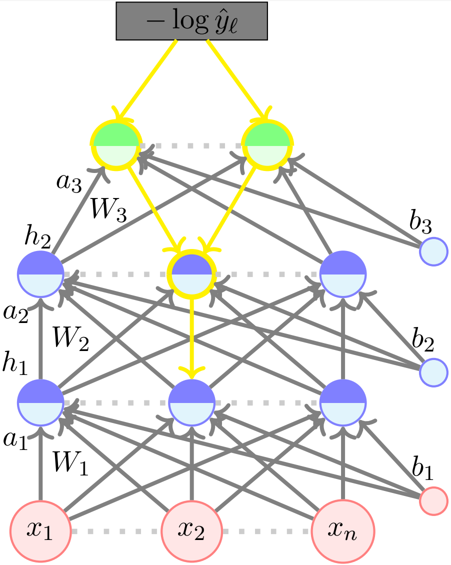
$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$



Recall that,

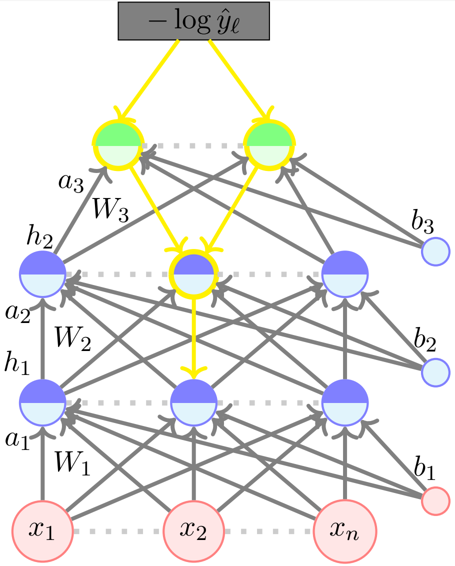
$$
\begin{aligned}
a_k &= b_k + W_k h_{k-1} \\
a_{ki} &= b_{ki} + W_{kij} h_{k-1,j} \\
\frac{\partial a_{ki}}{\partial W_{kij}} &= h_{k-1,j} \\
\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} &= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}} \\
&= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}
\end{aligned}
$$



Recall that,

$$
\begin{aligned}
a_k &= b_k + W_k h_{k-1} \\
a_{ki} &= b_{ki} + W_{kij} h_{k-1,j} \\
\frac{\partial a_{ki}}{\partial W_{kij}} &= h_{k-1,j} \\
\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} &= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}} \\
&= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}
\end{aligned}
$$

$$\nabla_{W_K} \mathscr{L}(\theta) =$$



Recall that,

$$a_k = b_k + W_k h_{k-1}$$

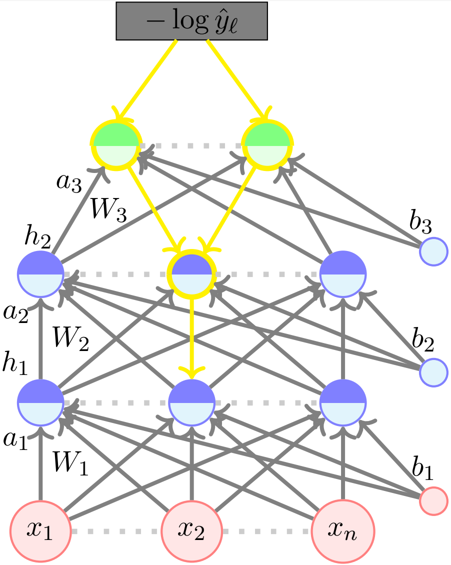
$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}$$

$$\nabla_{W_K} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \cdots & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k0n-1}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k,n-1,n-1}} \end{bmatrix}$$



Intentionally left blank

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times3}$ and see what each entry looks like

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} {\color{blue}h_{k-1,0}} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} {\color{red}h_{k-1,1}} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} {\color{blue}h_{k-1,0}} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} {\color{red}h_{k-1,1}} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} {\color{blue}h_{k-1,0}} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} {\color{red}h_{k-1,1}} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

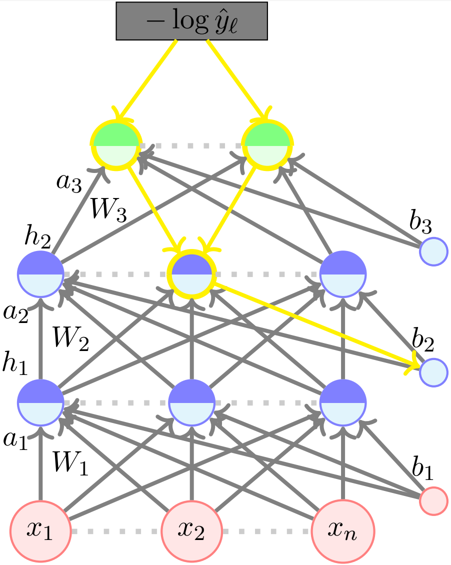
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

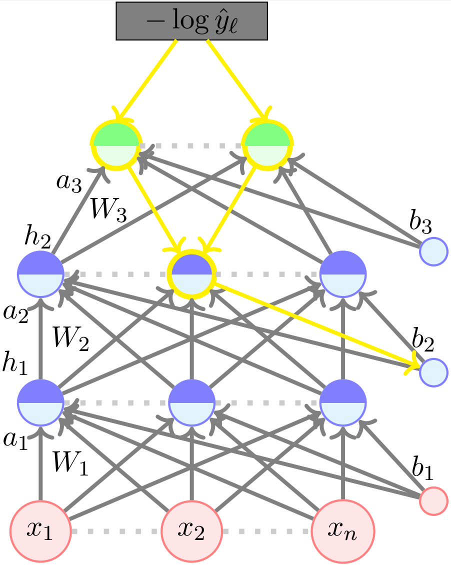
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} = \nabla_{a_k}\mathscr{L}(\theta) \cdot h_{k-1}^T$$

Finally, coming to the biases



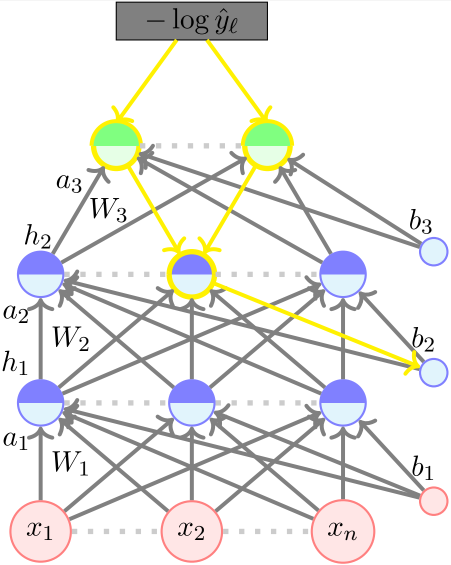
Finally, coming to the biases

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$



Finally, coming to the biases

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$
$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

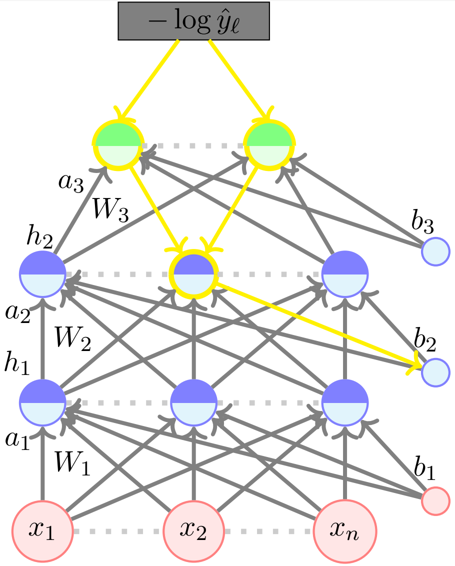


Finally, coming to the biases

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

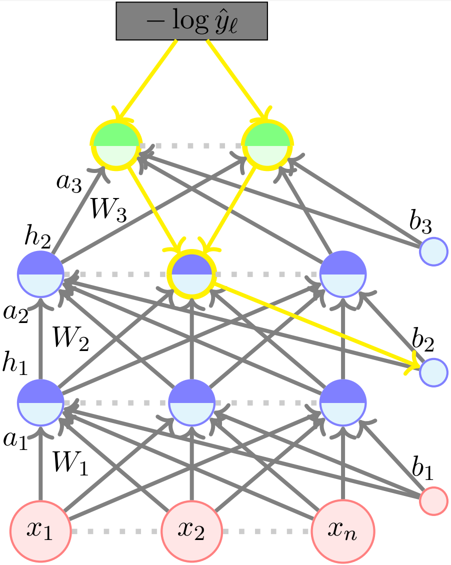
$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}$$



Finally, coming to the biases

$$
\begin{aligned}
a_{ki} &= b_{ki} + W_{kij} h_{k-1,j} \\
\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} &= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}} \\
&= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}
\end{aligned}
$$

We can now write the gradient w.r.t. the vector b_k



Finally, coming to the biases

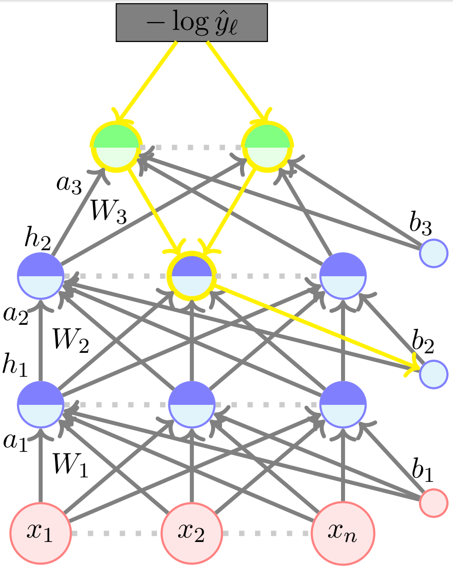
$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

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

We can now write the gradient w.r.t. the vector b_k

$$\nabla_{b_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{a_{k0}} \\ \frac{\partial \mathscr{L}(\theta)}{a_{k1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{a_{kn}} \end{bmatrix}$$



Finally, coming to the biases

$$a_{ki} = b_{ki} + W_{kij} h_{k-1,j}$$
$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$
$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}$$

We can now write the gradient w.r.t. the vector b_k

$$\nabla_{b_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{a_{k_0}} \\ \frac{\partial \mathscr{L}(\theta)}{a_{k_1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{a_{k_n}} \end{bmatrix} = \nabla_{a_\mathbf{k}} \mathscr{L}(\theta)$$