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

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

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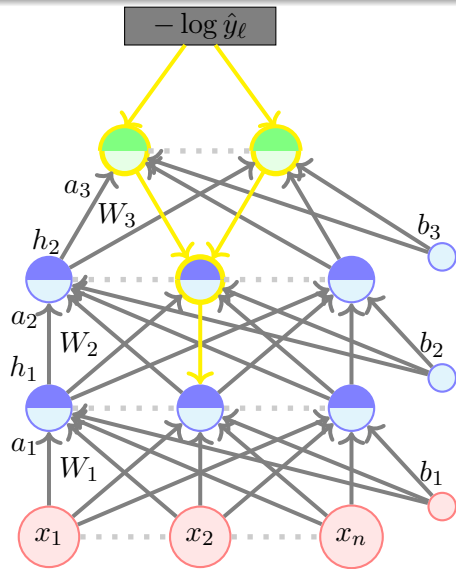
Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$
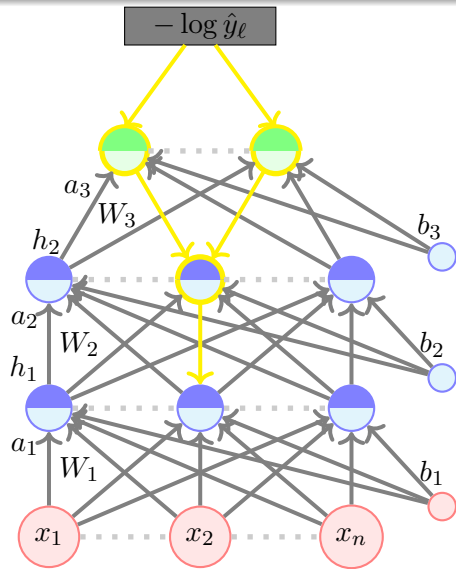
Recall that,

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

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

Recall that,

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

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

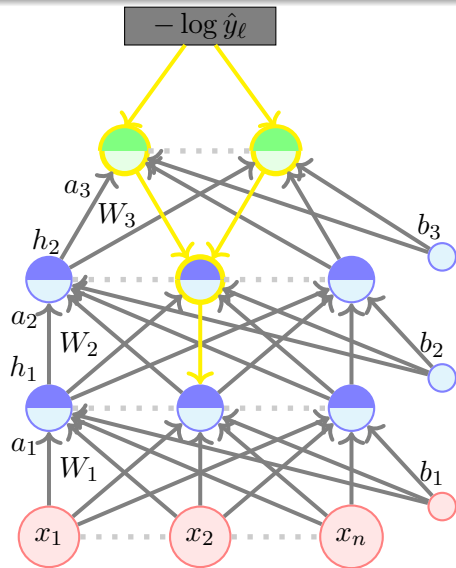$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}}$$

Recall that,

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

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

$$\frac{\partial \mathcal{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathcal{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$
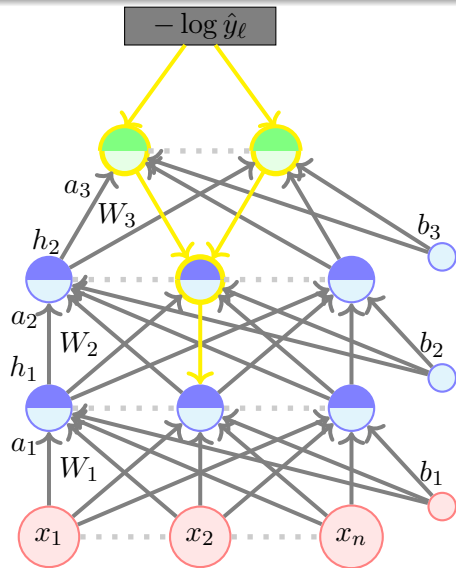
Recall that,

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

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

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

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

Recall that,

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

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

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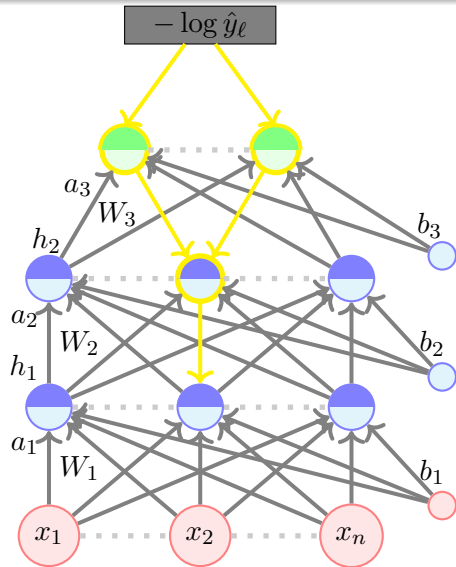
$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}$$
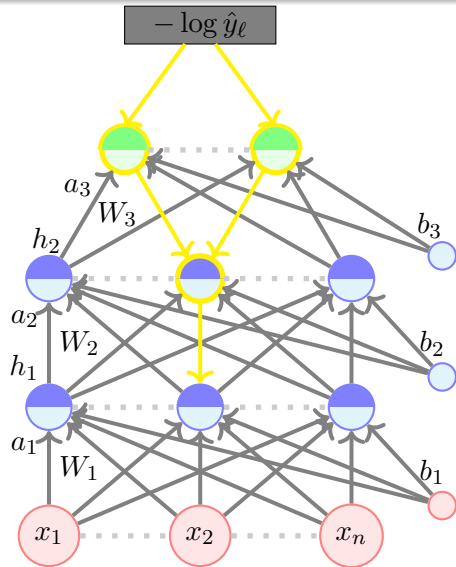
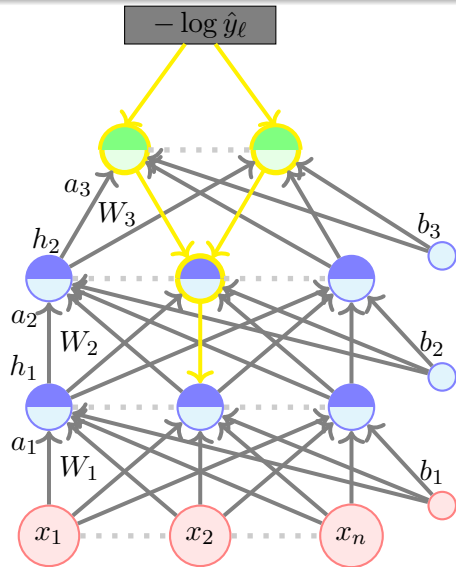$$\nabla_{W_k} \mathscr{L}(\theta) =$$

Recall that,

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

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

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

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

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \cdots & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k1n}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{knn}} \end{bmatrix}$$

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

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\[2mm] \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\[2mm] \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

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

$$\nabla_{W_k}\mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{k13}} \\ \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{k23}} \\ \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial \mathcal{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathcal{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathcal{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathcal{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \\ \frac{\partial \mathcal{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathcal{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathcal{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \\ \frac{\partial \mathcal{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathcal{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathcal{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

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

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\\\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\\\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\\\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\\\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

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

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial\mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k13}} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k23}} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial\mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial\mathscr{L}(\theta)}{\partial a_{ki}}\frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,3} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,3} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,3} \end{bmatrix} =$$

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

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial\mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k13}} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k23}} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial\mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial\mathscr{L}(\theta)}{\partial a_{ki}}\frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,3} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,3} \\[2mm] \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,3} \end{bmatrix} =$$

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

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\[2mm] \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\[2mm] \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\[2mm] \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\[2mm] \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

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

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial\mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k13}} \\\\ \frac{\partial\mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k23}} \\\\ \frac{\partial\mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial\mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial\mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial\mathscr{L}(\theta)}{\partial a_{ki}}\frac{\partial a_{ki}}{\partial W_{kij}}$$
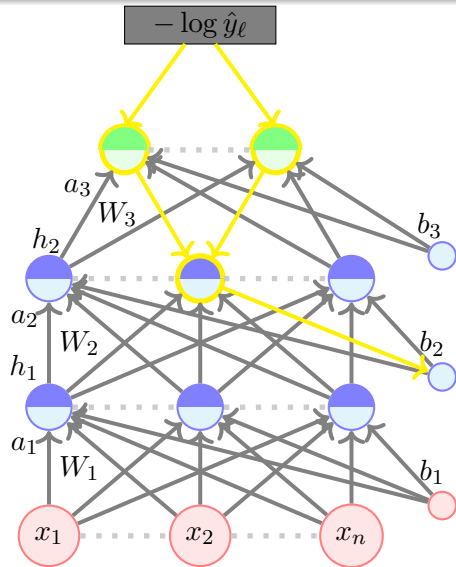
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k1}}h_{k-1,3} \\\\ \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k2}}h_{k-1,3} \\\\ \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,1} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,2} & \frac{\partial\mathscr{L}(\theta)}{\partial a_{k3}}h_{k-1,3} \end{bmatrix} = \nabla_{a_k}\mathscr{L}(\theta) \cdot \mathbf{h_{k-1}}^T$$

Finally, coming to the biases
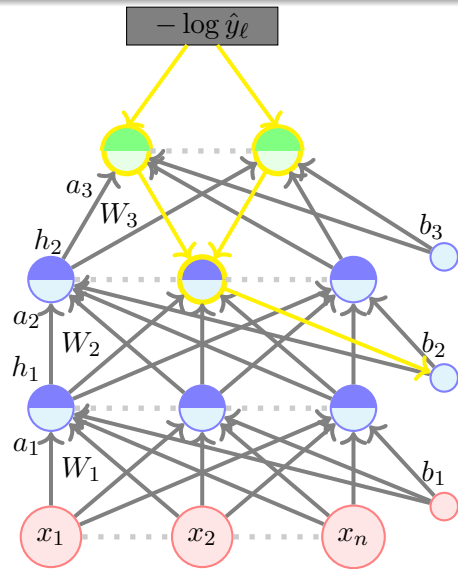
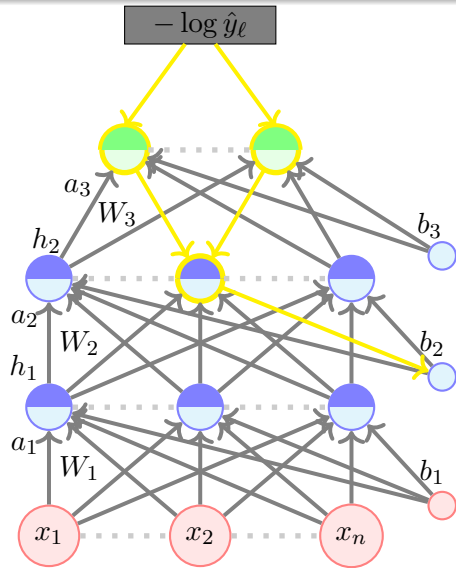Finally, coming to the biases

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

Finally, coming to the biases

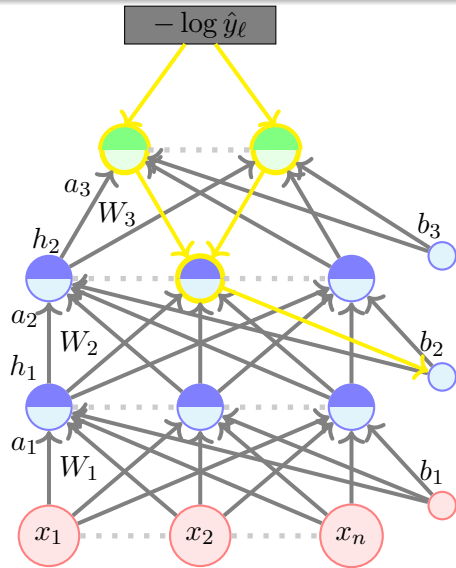$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$
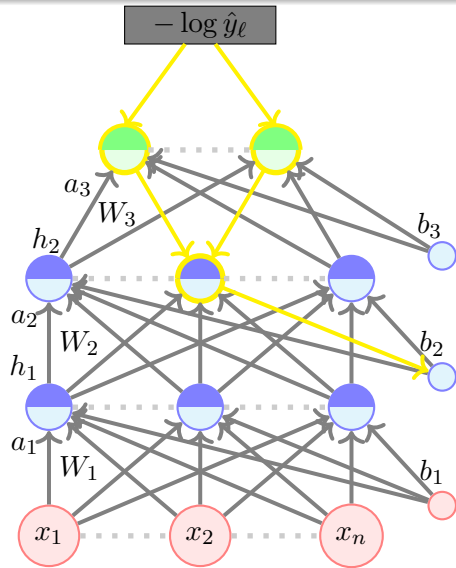
Finally, coming to the biases

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

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

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

Finally, coming to the biases

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

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

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

We can now write the gradient w.r.t. the vector b_k
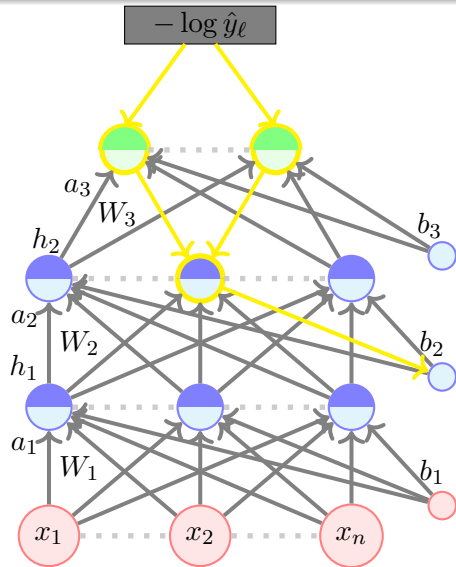
Finally, coming to the biases

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

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

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

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

$$\nabla_{\mathbf{b_k}} \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{a_{k1}} \\ \frac{\partial \mathcal{L}(\theta)}{a_{k2}} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta)}{a_{kn}} \end{bmatrix}$$

Finally, coming to the biases

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

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

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

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

$$\nabla_{\mathbf{b_k}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{a_{k1}} \\ \frac{\partial \mathscr{L}(\theta)}{a_{k2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{a_{kn}} \end{bmatrix} = \nabla_{\mathbf{a_k}} \mathscr{L}(\theta)$$