Backpropagation algorithm

$$\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4]$$

$$\mathbf{x} = [x_1, x_2, x_3, x_4]$$

Length and width of sepals

Length and width of petals

$$\mathbf{y} = [y_1, y_2, y_3]$$

Setosa

Virginica

Virginica

$$\mathbf{y} = \qquad\qquad \mathbf{x}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \qquad\qquad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} w^1_{11} & w^1_{12} & w^1_{13} & w^1_{14} \\ w^1_{21} & w^1_{22} & w^1_{23} & w^1_{24} \\ \vdots & \vdots & \vdots & \vdots \\ w^1_{n1} & w^1_{n2} & w^1_{n3} & w^1_{n4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b^1_1 \\ b^1_2 \\ \vdots \\ b^1_n \end{bmatrix}$$

$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\mathbf{y} \quad = \qquad \mathbf{W}^2\left(\mathrm{ReLU}\left(\mathbf{W}^1\mathbf{x}+\mathbf{b}^1\right)\right)+\mathbf{b}^2$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \qquad \mathrm{ReLU}\left(\begin{bmatrix} w_{11}^1 & w_{12}^1 & w_{13}^1 & w_{14}^1 \\ w_{21}^1 & w_{22}^1 & w_{23}^1 & w_{24}^1 \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1}^1 & w_{n2}^1 & w_{n3}^1 & w_{n4}^1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}+\begin{bmatrix} b_1^1 \\ b_2^1 \\ \vdots \\ b_n^1 \end{bmatrix}\right)$$

$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x}+\mathbf{b}^1$$
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}})$$

$$\mathbf{y} \;=\; \mathbf{W}^2\left(\mathrm{ReLU}\left(\mathbf{W}^1\mathbf{x}+\mathbf{b}^1\right)\right)+\mathbf{b}^2$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} w^1_{11} & w^1_{12} & \dots & w^1_{1n} \\ w^1_{21} & w^1_{22} & \dots & w^1_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w^1_{m1} & w^1_{m2} & \dots & w^1_{nn} \end{bmatrix} \mathrm{ReLU}\left(\begin{bmatrix} w^1_{11} & w^1_{12} & w^1_{13} & w^1_{14} \\ w^1_{21} & w^1_{22} & w^1_{23} & w^1_{24} \\ \vdots & \vdots & \vdots & \vdots \\ w^1_{n1} & w^1_{n2} & w^1_{n3} & w^1_{n4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b^1_1 \\ b^1_2 \\ \vdots \\ b^1_n \end{bmatrix}\right) + \begin{bmatrix} b^2_1 \\ b^2_2 \\ \vdots \\ b^2_m \end{bmatrix}$$

$$\begin{aligned} \mathbf{h}^a &= \mathbf{W}^1\mathbf{x}+\mathbf{b}^1 \\ \mathbf{h}^s &= \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}}) \\ \mathbf{o}^a &= \mathbf{W}^2\mathbf{h}^s+\mathbf{b}^2 \end{aligned}$$

$$p(\mathbf{y}) = \text{Softmax}\left[\mathbf{W}^2\left(\text{ReLU}\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right)\right) + \mathbf{b}^2\right]$$

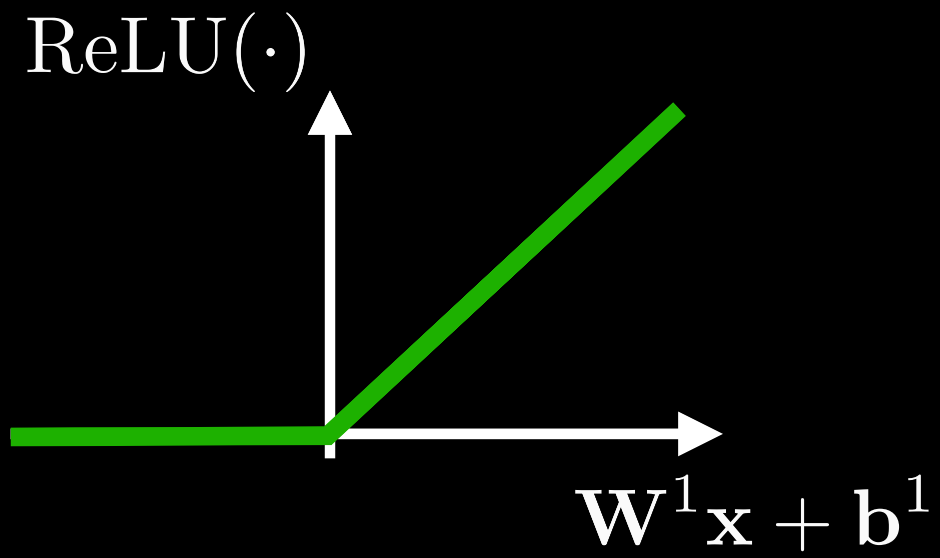
$$\begin{bmatrix} p(y_1) \\ p(y_2) \\ p(y_3) \end{bmatrix} = \text{S}\left(\begin{bmatrix} w_{11}^1 & w_{12}^1 & \dots & w_{1n}^1 \\ w_{21}^1 & w_{22}^1 & \dots & w_{2n}^1 \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1}^1 & w_{m2}^1 & \dots & w_{nn}^1 \end{bmatrix} \text{ReLU}\left(\begin{bmatrix} w_{11}^1 & w_{12}^1 & w_{13}^1 & w_{14}^1 \\ w_{21}^1 & w_{22}^1 & w_{23}^1 & w_{24}^1 \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1}^1 & w_{n2}^1 & w_{n3}^1 & w_{n4}^1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b_1^1 \\ b_2^1 \\ \vdots \\ b_n^1 \end{bmatrix}\right) + \begin{bmatrix} b_1^2 \\ b_2^2 \\ \vdots \\ b_m^2 \end{bmatrix}\right)$$

$$\begin{aligned} \mathbf{h}^a &= \mathbf{W}^1\mathbf{x} + \mathbf{b}^1 \\ \mathbf{h}^s &= \text{ReLU}(\mathbf{h}^{\text{a}}) \\ \mathbf{o}^a &= \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2 \\ \mathbf{o}^s &= \text{Softmax}(\mathbf{o}^{\text{a}}) \end{aligned}$$

$$p(\mathbf{y}) = \mathrm{Softmax}\left[\mathbf{W}^2\left(\mathrm{ReLU}\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right)\right) + \mathbf{b}^2\right]$$

$$p(\mathbf{y}) = \text{Softmax}\left[\mathbf{W}^2\left(\text{ReLU}\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right)\right) + \mathbf{b}^2\right]$$

$$\text{Softmax}_{\text{i}}(\mathbf{o}^{\text{a}}) = \frac{\exp(\text{o}_{\text{i}}^{\text{a}})}{\sum_{\text{k}}\exp(\text{o}_{\text{k}}^{\text{a}})}$$



$$L(\mathbf{x}, y) = \log p(y)$$

Log-likelihood

$$p(\mathbf{y}) = \mathrm{Softmax}\left[\mathbf{W}^2\left(\mathrm{ReLU}\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right)\right) + \mathbf{b}^2\right]$$

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

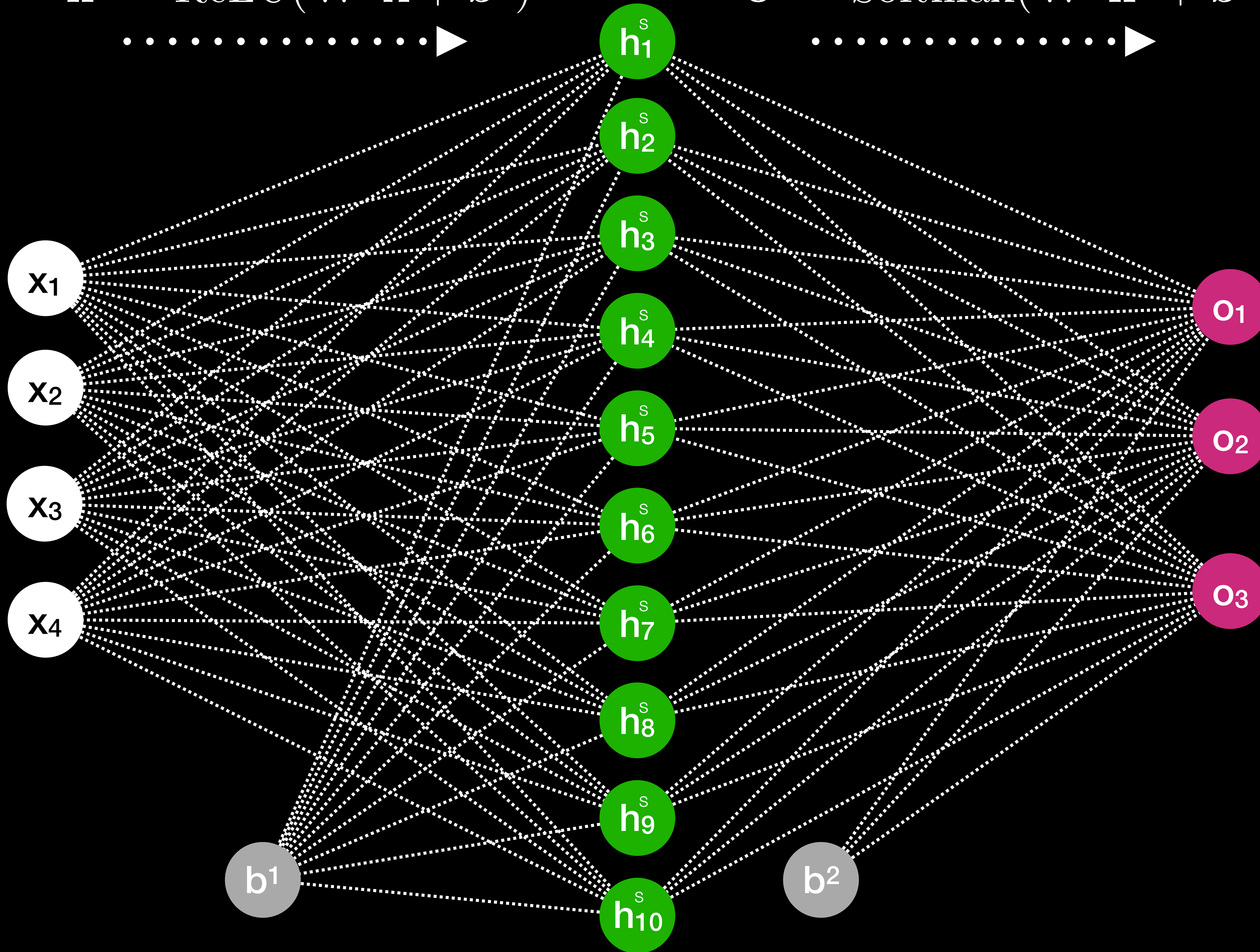
$$p(\mathbf{y}) = \mathrm{Softmax}\left[\mathbf{W}^2\left(\mathrm{ReLU}\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right)\right) + \mathbf{b}^2\right]$$

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^\mathrm{s} + \mathbf{b}^2)$$

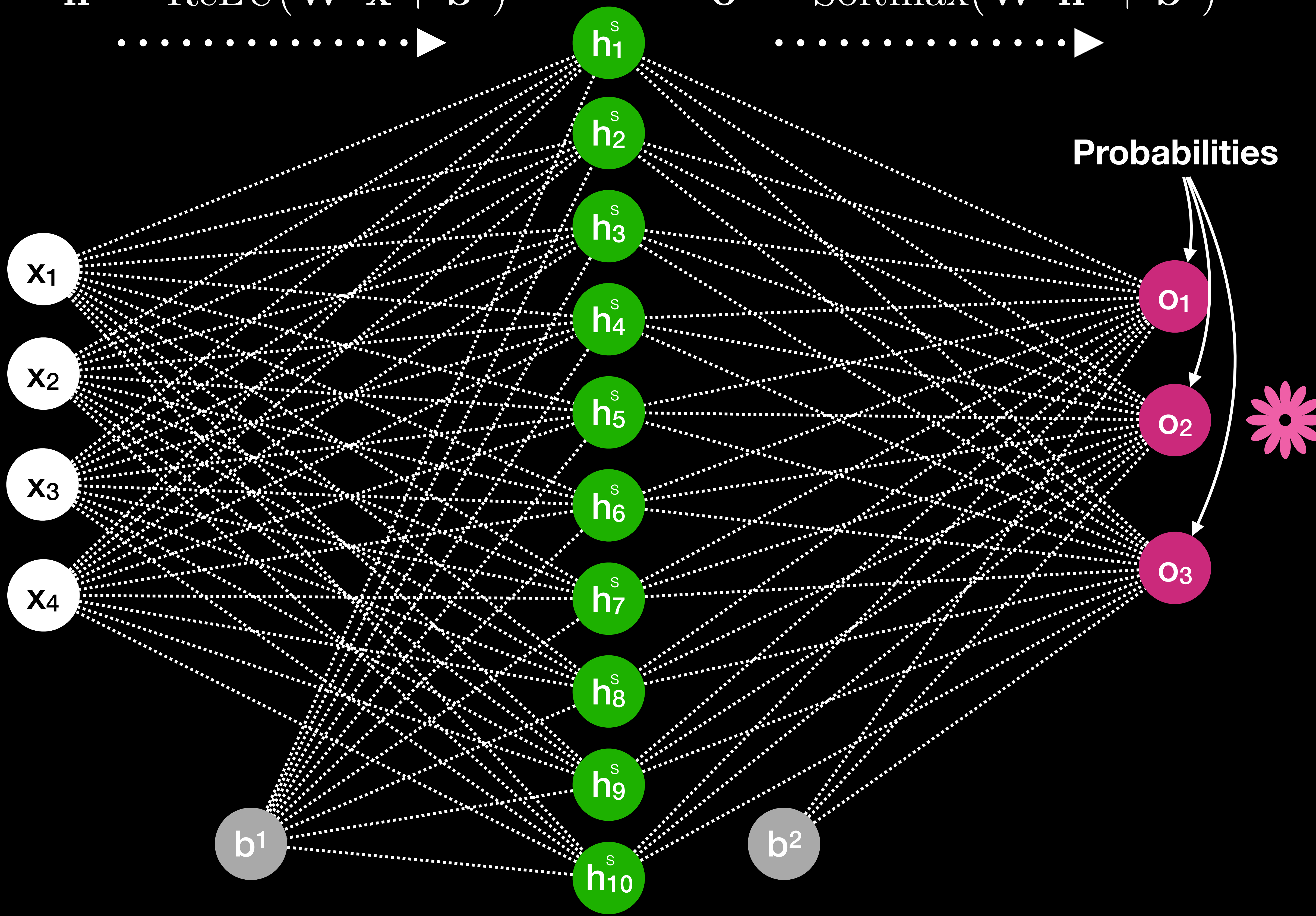
$$p(\mathbf{y}) = \mathrm{Softmax}\left[\mathbf{W}^2 \underbrace{\left(\mathrm{ReLU}\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right)\right)}^{\mathbf{h}^s} + \mathbf{b}^2\right]$$

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^{\mathrm{s}} + \mathbf{b}^2)$$

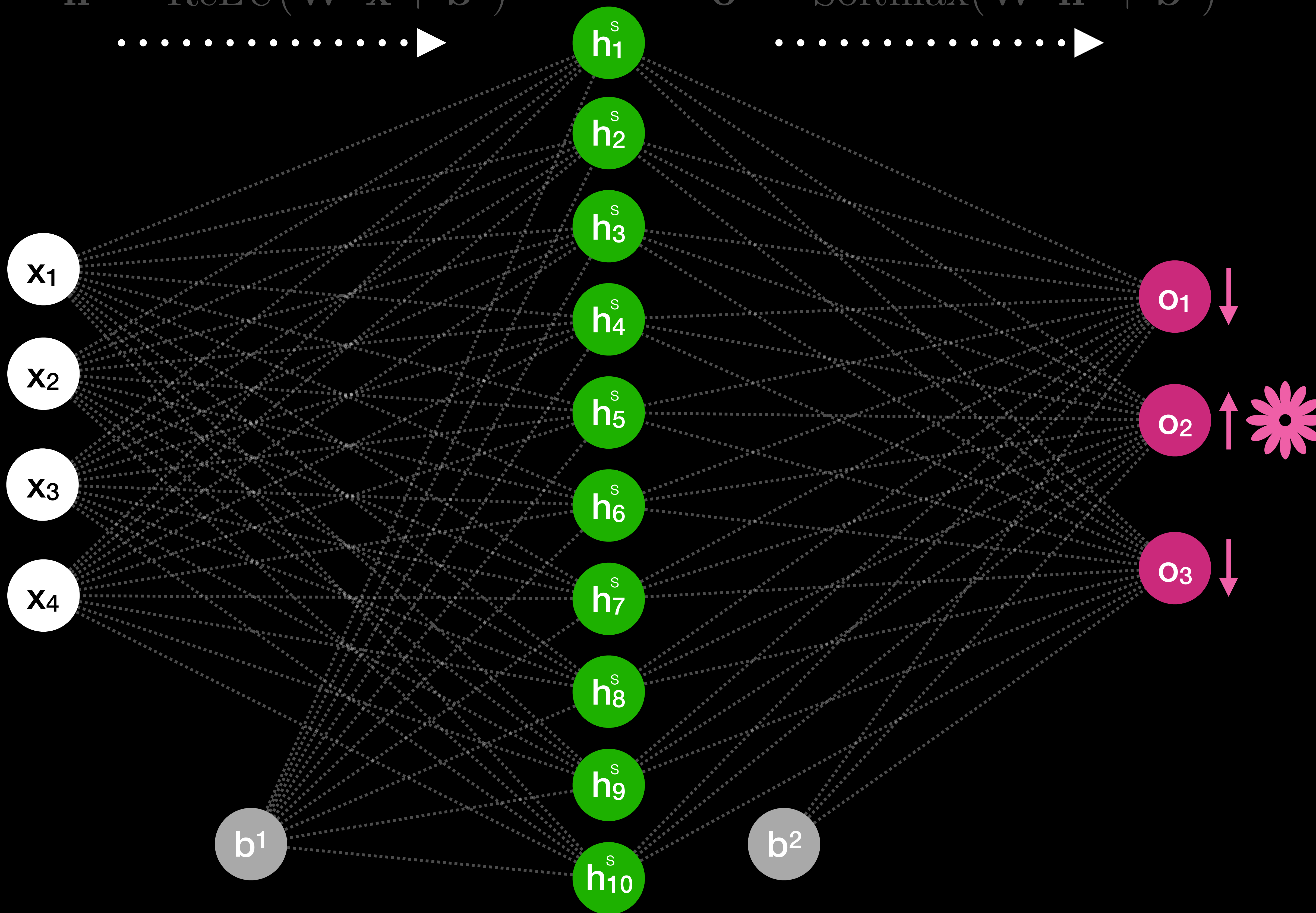
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$

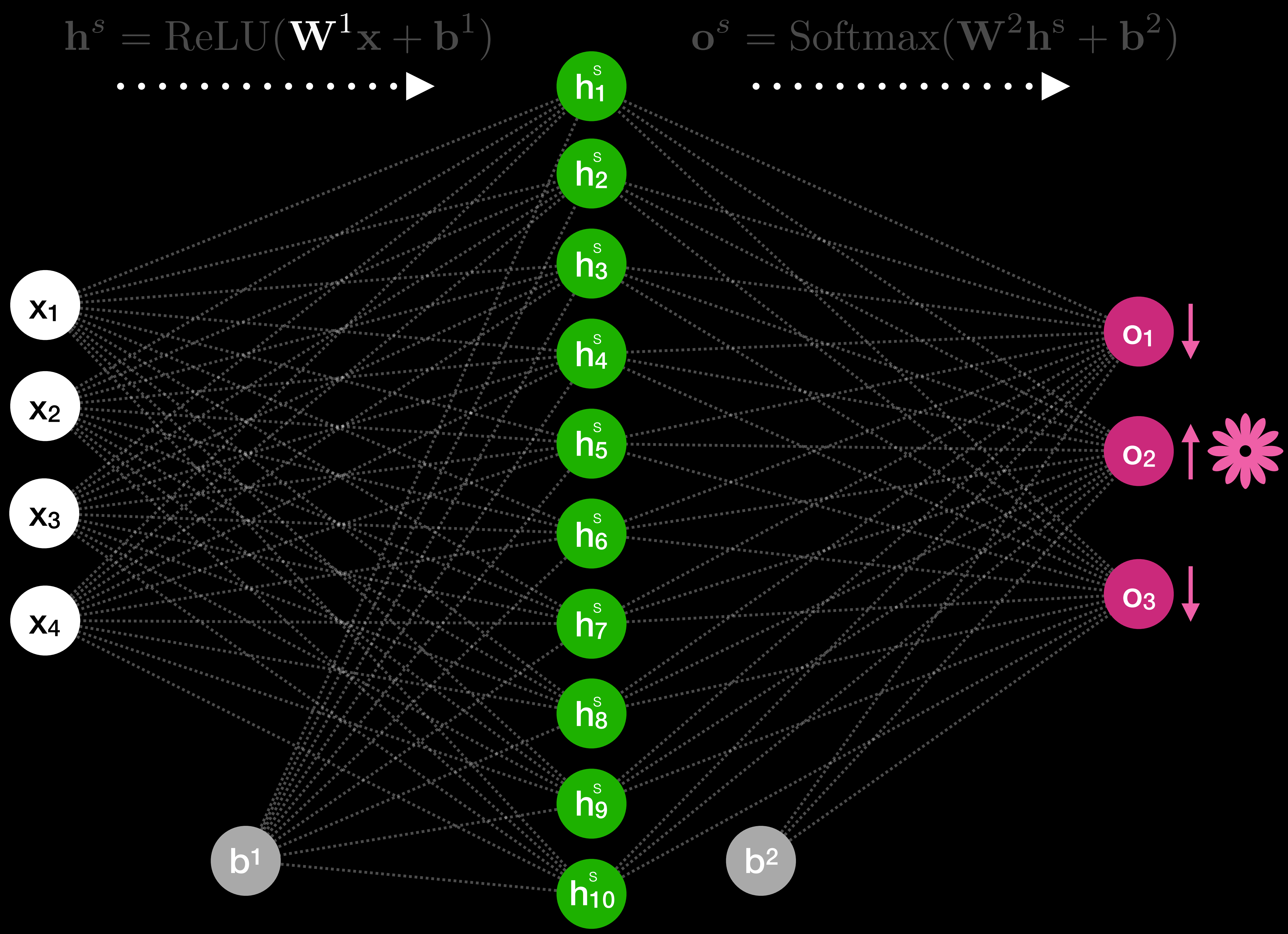


$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$



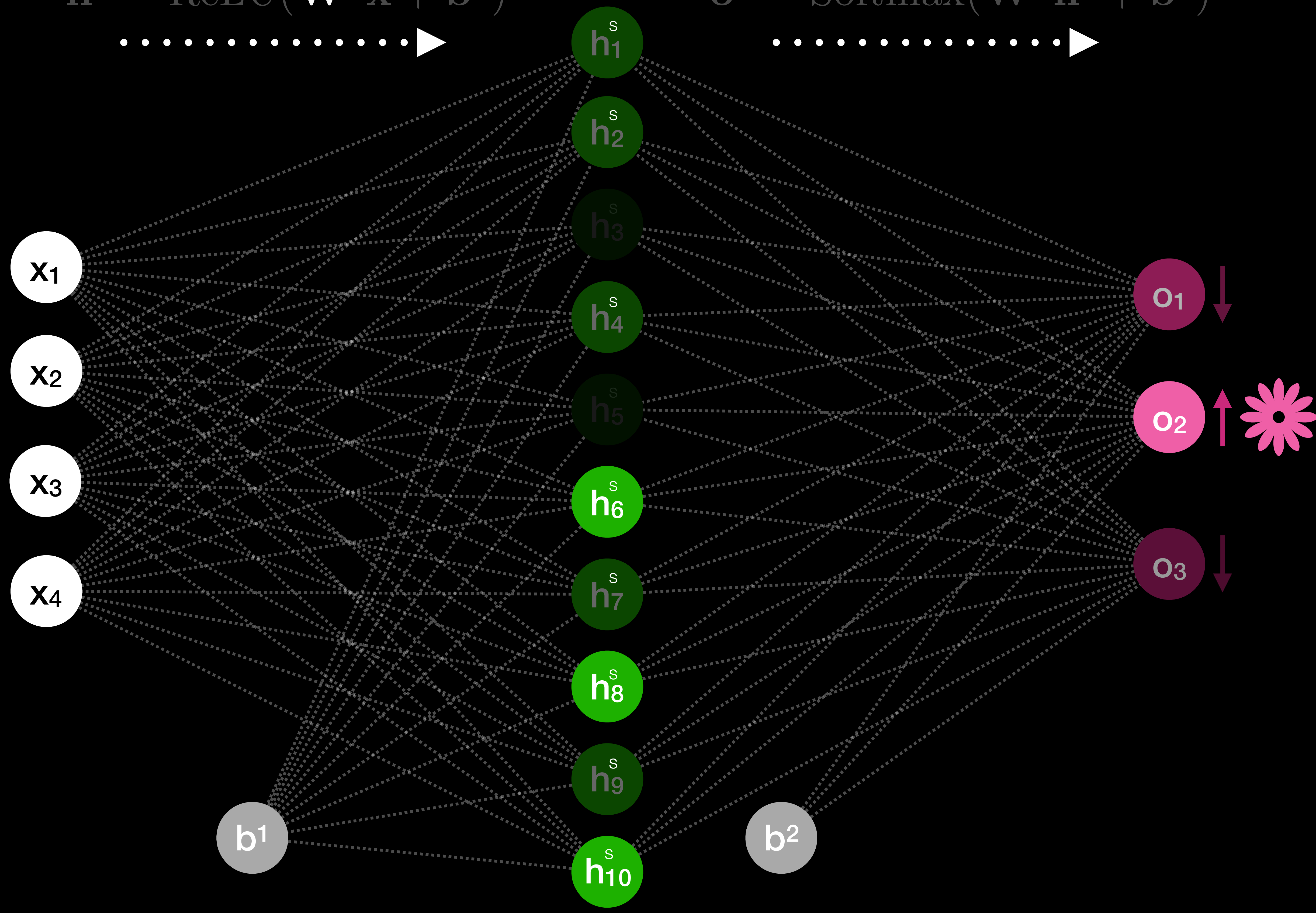
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$



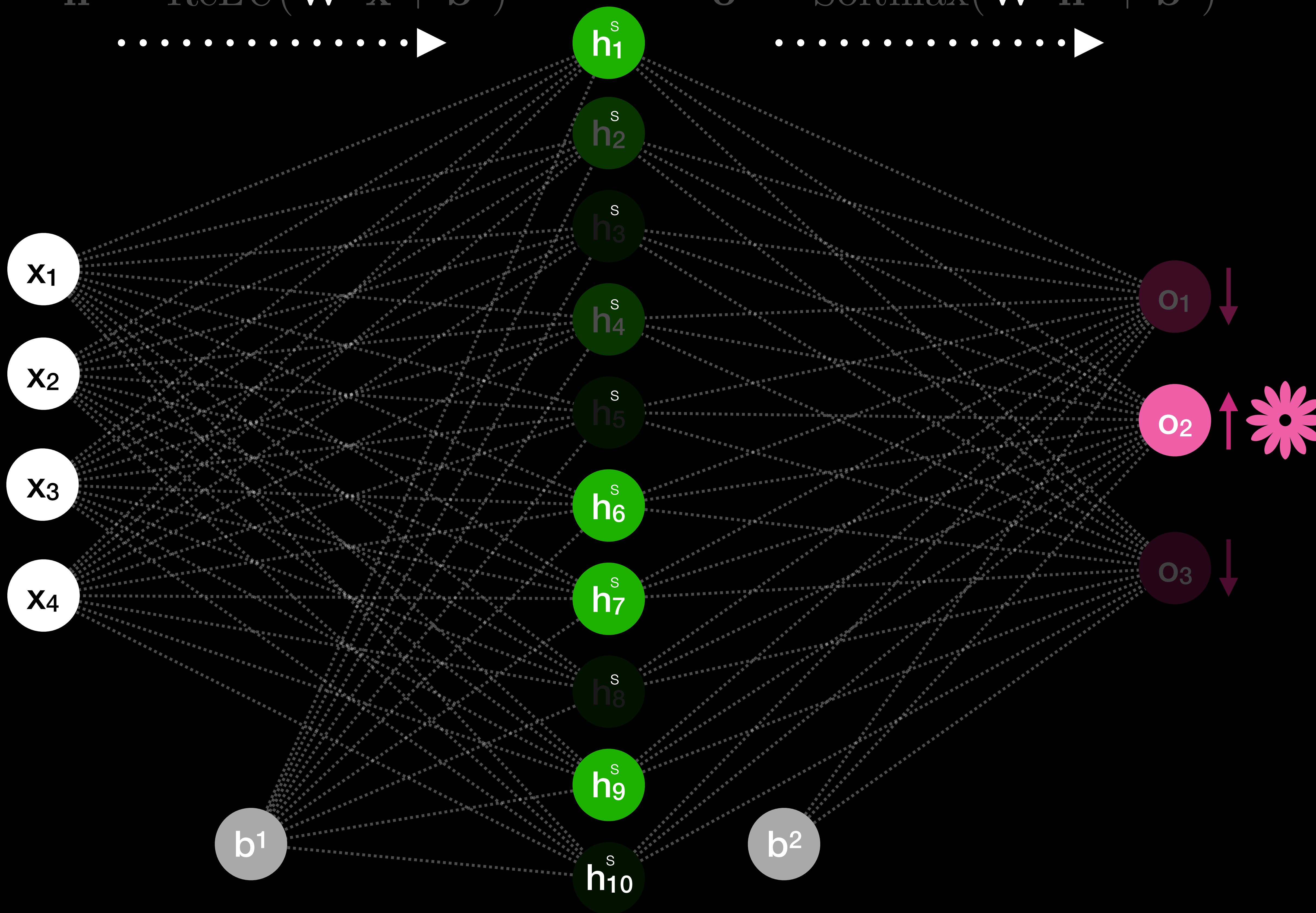


$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$
$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$
x_1
x_2
x_3
x_4
h^s_1
h^s_2
h^s_3
h^s_4
h^s_5
h^s_6
h^s_7
h^s_8
h^s_9
h^s_{10}
o_1
o_2
o_3
b^1
b^2

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$

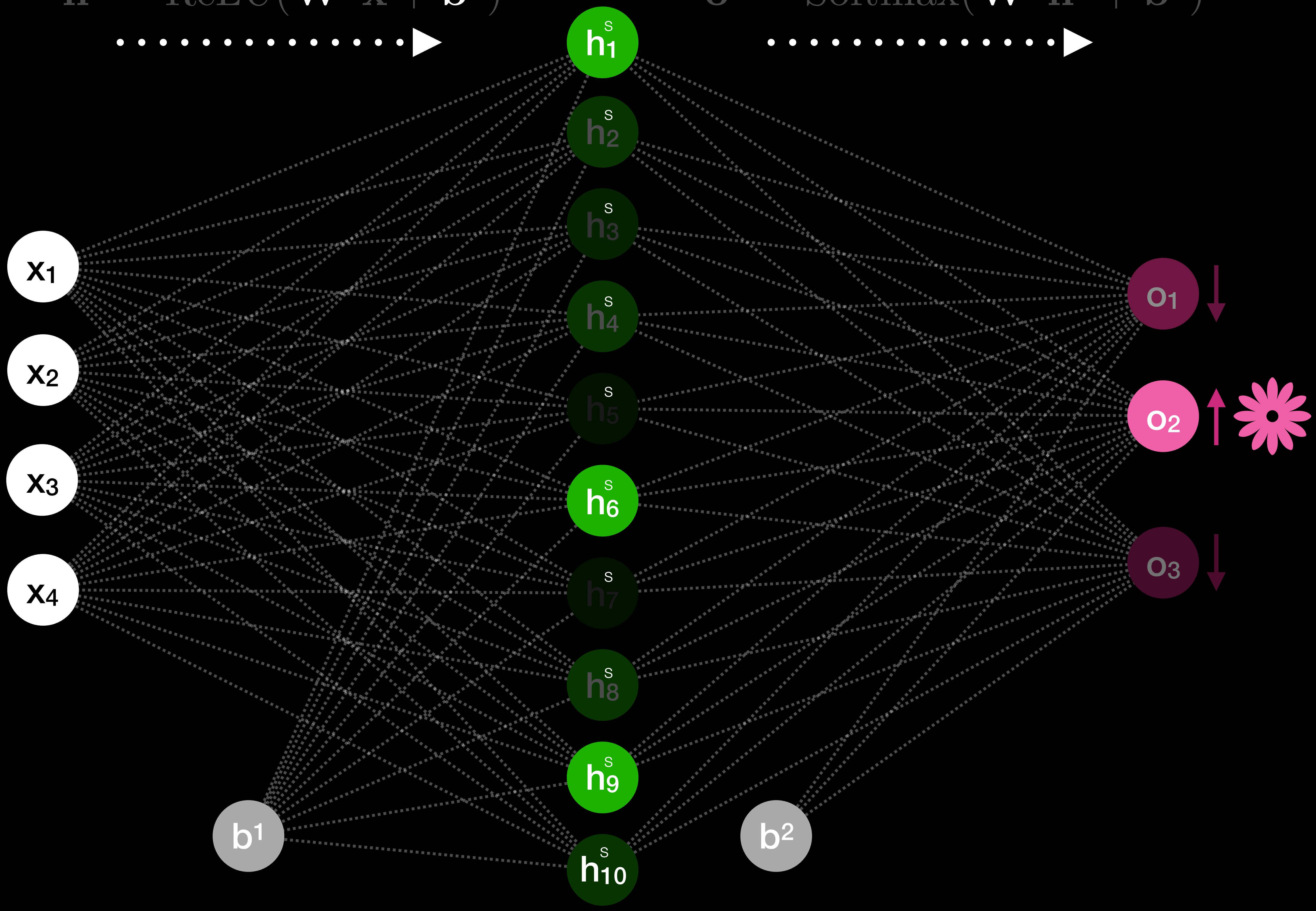


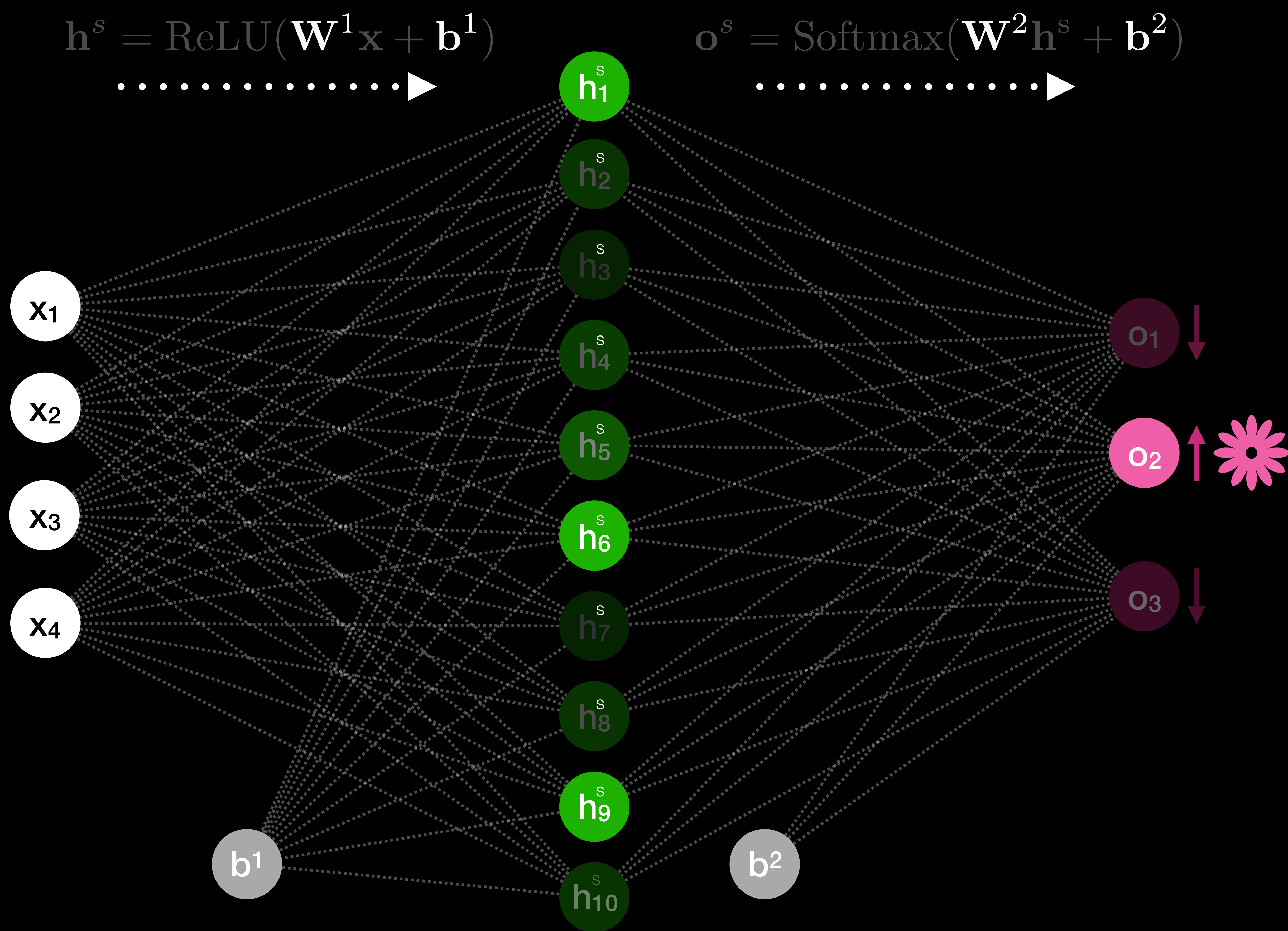
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \qquad \mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$



$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$





$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$
$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$
x_1
x_2
x_3
x_4
h_1^s
h_2^s
h_3^s
h_4^s
h_5^s
h_6^s
h_7^s
h_8^s
h_9^s
h_{10}^s
o_1
o_2
o_3
b^1
b^2

Optimal parameters?

Backpropagation

$$\mathcal{L} = -\log(\mathbf{o}^s)$$

$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{o}^\mathrm{a})$$

$$\mathbf{o}^a = \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2$$

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{h}^\mathrm{a})$$

$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} =$$

$$\mathcal{L} = -\log(\mathbf{o}^s)$$

$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{o}^{\mathrm{a}})$$

$$\mathbf{o}^a = \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2$$

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}})$$

$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^s}\frac{\delta\mathbf{o}^s}{\delta\mathbf{W}^2}$$

$$
\begin{aligned}
\mathcal{L} &= -\log(\mathbf{o}^s) \\
\mathbf{o}^s &= \mathrm{Softmax}(\mathbf{o}^{\mathrm{a}}) \\
\mathbf{o}^a &= \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2 \\
\mathbf{h}^s &= \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}}) \\
\mathbf{h}^a &= \mathbf{W}^1\mathbf{x} + \mathbf{b}^1
\end{aligned}
$$

$$
\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^s}\frac{\delta\mathbf{o}^s}{\delta\mathbf{o}^a}\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2}
$$

$$\mathcal{L} = -\log(\mathbf{o}^s)$$
$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{o}^{\mathrm{a}})$$
$$\mathbf{o}^a = \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2$$
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}})$$
$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} = \underbrace{\boxed{\frac{\delta\mathcal{L}}{\delta\mathbf{o}^s}\frac{\delta\mathbf{o}^s}{\delta\mathbf{o}^a}}}_{\mathbf{o}^s - \mathrm{onehot_m}(\mathrm{y})}\underbrace{\boxed{\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2}}}_{\mathbf{h}^s}$$

$$
\begin{aligned}
\mathcal{L} &= -\log(\mathbf{o}^s) \\
\mathbf{o}^s &= \mathrm{Softmax}(\mathbf{o}^{\mathrm{a}}) \\
\mathbf{o}^a &= \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2 \\
\mathbf{h}^s &= \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}}) \\
\mathbf{h}^a &= \mathbf{W}^1\mathbf{x} + \mathbf{b}^1
\end{aligned}
$$

$$
\begin{aligned}
\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a} &= \frac{\delta\mathcal{L}}{\delta\mathbf{o}^s}\frac{\delta\mathbf{o}^s}{\delta\mathbf{o}^a} = \mathbf{o}^s - \mathrm{onehot_m(y)} \\
\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} &= \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\right)^{\intercal} \cdot \mathbf{h}^s
\end{aligned}
$$

$$\mathcal{L} = -\log(\mathbf{o}^s)$$

$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{o}^{\mathrm{a}})$$

$$\mathbf{o}^a = \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2$$

$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}})$$

$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^s}\frac{\delta \mathbf{o}^s}{\delta \mathbf{o}^a} = \mathbf{o}^s - \mathrm{onehot_m(y)}$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}\frac{\delta \mathbf{o}^a}{\delta \mathbf{W}^2} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}\right)^{\intercal} \cdot \mathbf{h}^s$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}$$

$$\mathcal{L} = -\log(\mathbf{o}^s)$$
$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{o}^{\mathrm{a}})$$
$$\mathbf{o}^a = \mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2$$
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{h}^{\mathrm{a}})$$
$$\mathbf{h}^a = \mathbf{W}^1\mathbf{x} + \mathbf{b}^1$$

$$\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^s}\frac{\delta\mathbf{o}^s}{\delta\mathbf{o}^a} = \mathbf{o}^s - \mathrm{onehot}_{\mathrm{m}}(\mathrm{y})$$
$$\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\right)^\intercal \cdot \mathbf{h}^s$$
$$\frac{\delta\mathcal{L}}{\delta\mathbf{b}^2} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}$$

$$\frac{\delta\mathcal{L}}{\delta\mathbf{h}^s} = \mathbf{W}^2 \cdot \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}$$
$$\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a} \odot \mathbf{1}_{\mathbf{h}^{\mathbf{a}}(\mathbf{x})>\mathbf{0}}$$
$$\frac{\delta\mathcal{L}}{\delta\mathbf{W}^1} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a}\right)^\intercal \cdot \mathbf{x}$$
$$\frac{\delta\mathcal{L}}{\delta\mathbf{b}^1} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a}\right)$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \frac{\delta \mathbf{o}^a}{\delta \mathbf{W}^2} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}\right)^\intercal \cdot \mathbf{h}^s$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^1} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}\right)^\intercal \cdot \mathbf{x}$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^1} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}\right)$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \frac{\delta \mathbf{o}^a}{\delta \mathbf{W}^2} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}\right)^{\intercal} \cdot \mathbf{h}^s$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a}$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^1} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}\right)^{\intercal} \cdot \mathbf{x}$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^1} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}\right)$$

$$\Delta_{\mathbf{W}^2}\mathcal{L} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\right)^{\mathsf{T}} \cdot \mathbf{h}^s$$

$$\Delta_{\mathbf{b}^2}\mathcal{L} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}$$

$$\Delta_{\mathbf{W}^1}\mathcal{L} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a}\right)^{\mathsf{T}} \cdot \mathbf{x}$$

$$\Delta_{\mathbf{b}^1}\mathcal{L} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a}\right)$$

Gradients points along the greatest rate of increase

Objective: Minimize $\mathcal{L}$

$$\Delta_{\mathbf{W}^2}\mathcal{L} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\right)^{\intercal} \cdot \mathbf{h}^s \qquad \Delta_{\mathbf{W}^1}\mathcal{L} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a}\right)^{\intercal} \cdot \mathbf{x}$$

$$\Delta_{\mathbf{b}^2}\mathcal{L} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a} \qquad \Delta_{\mathbf{b}^1}\mathcal{L} = \left(\frac{\delta\mathcal{L}}{\delta\mathbf{h}^a}\right)$$

Gradients points along the greatest rate of increase

Objective: Minimize $\mathcal{L}$

$$\mathbf{W}^2 \leftarrow \mathbf{W}^2 - \eta\Delta_{\mathbf{W}^2}\mathcal{L}$$
$$\mathbf{b}^2 \leftarrow \mathbf{b}^2 - \eta\Delta_{\mathbf{b}^2}\mathcal{L}$$

$$\mathbf{W}^1 \leftarrow \mathbf{W}^1 - \eta\Delta_{\mathbf{W}^1}\mathcal{L}$$
$$\mathbf{b}^1 \leftarrow \mathbf{b}^1 - \eta\Delta_{\mathbf{b}^1}\mathcal{L}$$

$$\mathbf{W}^2 \leftarrow \mathbf{W}^2 - \eta\Delta_{\mathbf{W}^2}\mathcal{L}$$
$$\mathbf{b}^2 \leftarrow \mathbf{b}^2 - \eta\Delta_{\mathbf{b}^2}\mathcal{L}$$

$$\mathbf{W}^1 \leftarrow \mathbf{W}^1 - \eta\Delta_{\mathbf{W}^1}\mathcal{L}$$
$$\mathbf{b}^1 \leftarrow \mathbf{b}^1 - \eta\Delta_{\mathbf{b}^1}\mathcal{L}$$

$$\mathbf{W}^2 \leftarrow \mathbf{W}^2 - \eta\Delta_{\mathbf{W}^2}\mathcal{L}$$
$$\mathbf{b}^2 \leftarrow \mathbf{b}^2 - \eta\Delta_{\mathbf{b}^2}\mathcal{L}$$

$$\mathbf{W}^1 \leftarrow \mathbf{W}^1 - \eta\Delta_{\mathbf{W}^1}\mathcal{L}$$
$$\mathbf{b}^1 \leftarrow \mathbf{b}^1 - \eta\Delta_{\mathbf{b}^1}\mathcal{L}$$

$$\theta^{t+1} = \theta^t - \eta\Delta_\theta\mathcal{L}$$

$$\mathbf{W}^2 \leftarrow \mathbf{W}^2 - \eta \Delta_{\mathbf{W}^2} \mathcal{L}$$
$$\mathbf{b}^2 \leftarrow \mathbf{b}^2 - \eta \Delta_{\mathbf{b}^2} \mathcal{L}$$

$$\theta^{t+1} = \theta^t - \eta \Delta_\theta \mathcal{L}$$

$$\mathbf{W}^1 \leftarrow \mathbf{W}^1 - \eta \Delta_{\mathbf{W}^1} \mathcal{L}$$
$$\mathbf{b}^1 \leftarrow \mathbf{b}^1 - \eta \Delta_{\mathbf{b}^1} \mathcal{L}$$

$$\mathbf{W}^2 \leftarrow \mathbf{W}^2 - \eta \Delta_{\mathbf{W}^2} \mathcal{L}$$

$$\mathbf{b}^2 \leftarrow \mathbf{b}^2 - \eta \Delta_{\mathbf{b}^2} \mathcal{L}$$

$$\theta^{t+1} = \theta^t - \eta \Delta_\theta \mathcal{L}(f(\mathbf{x}^t, \theta), y^t)$$

$$\mathbf{W}^1 \leftarrow \mathbf{W}^1 - \eta \Delta_{\mathbf{W}^1} \mathcal{L}$$

$$\mathbf{b}^1 \leftarrow \mathbf{b}^1 - \eta \Delta_{\mathbf{b}^1} \mathcal{L}$$

Optimisation problem