

Backpropagation algorithm

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

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

Length and width of
sepals

Length and width of
petals

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

Setosa

Virginica

Virginica



$$\mathbf{y} \; = \;$$

$$\mathbf{x}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} =$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

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

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

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

$$\mathbf{y} = \mathbf{W}^2 (\text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)) + \mathbf{b}^2$$

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

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

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

$$\mathbf{y} = \mathbf{W}^2 (\text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)) + \mathbf{b}^2$$

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

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

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

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

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

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

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

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

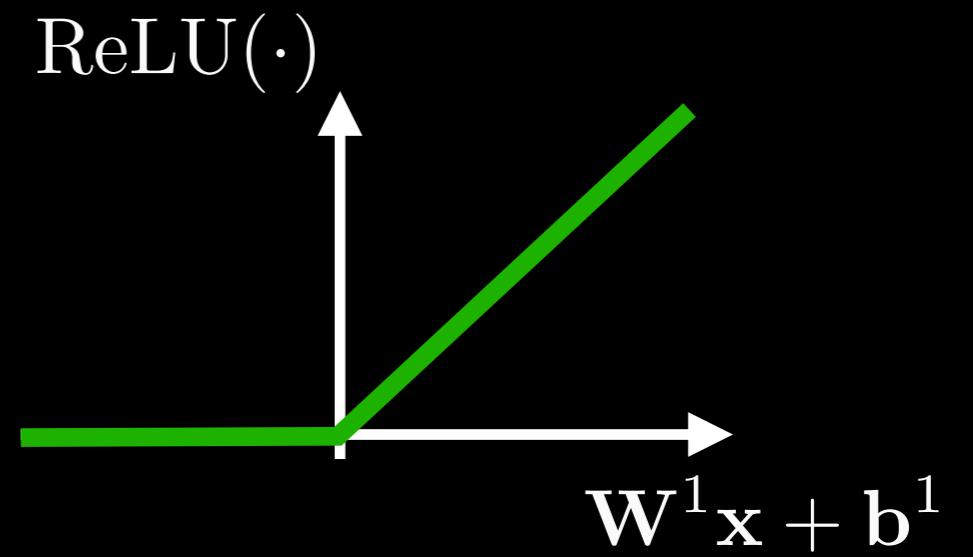
$$\mathbf{o}^a = \mathbf{W}^2 \mathbf{h}^s + \mathbf{b}^2$$

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

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

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

$$\text{Softmax}_i(\mathbf{o}^a) = \frac{\exp(o_i^a)}{\sum_k \exp(o_k^a)}$$



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

Log-likelihood

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

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

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

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

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

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

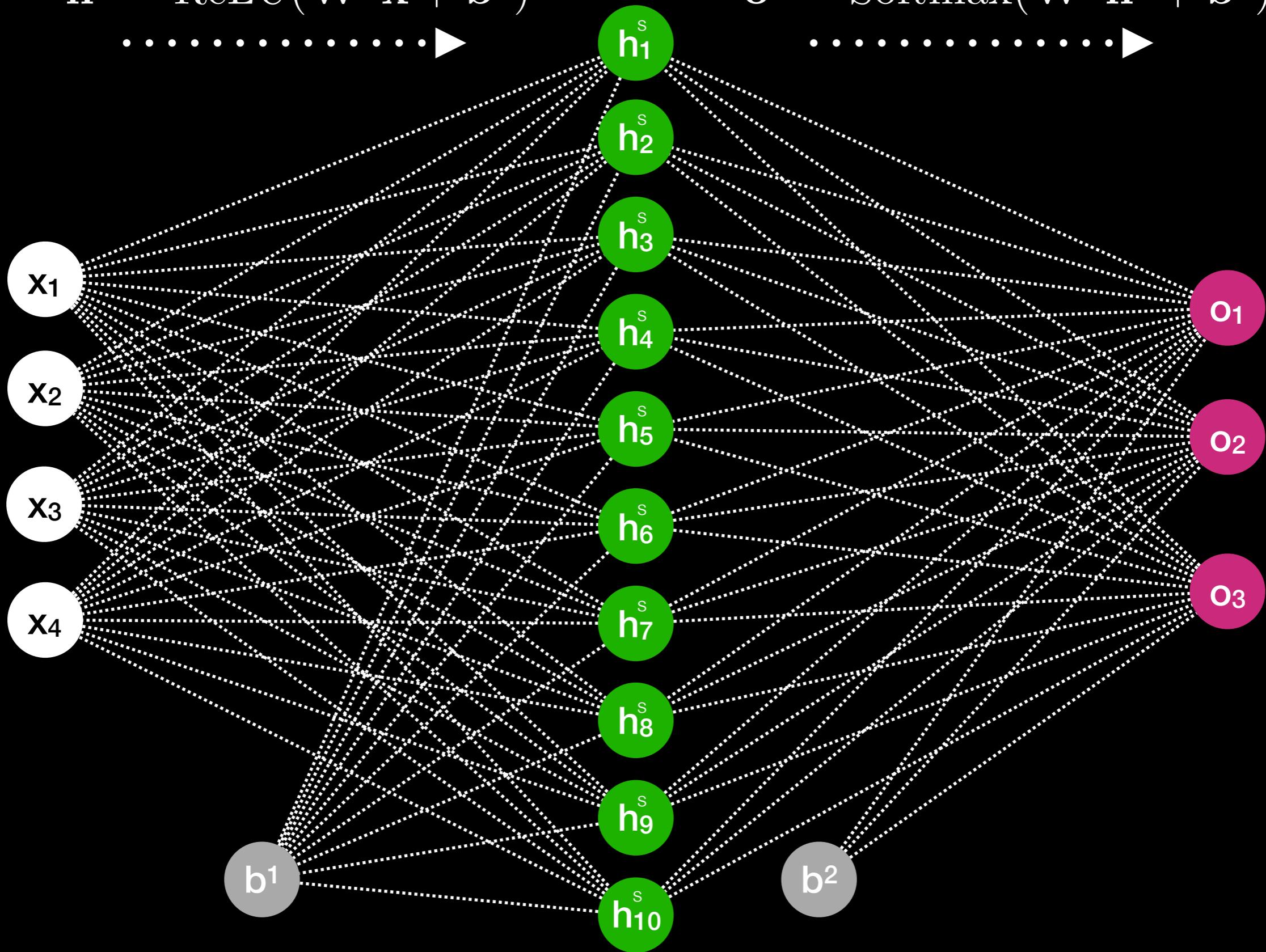
\mathbf{h}^s

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

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

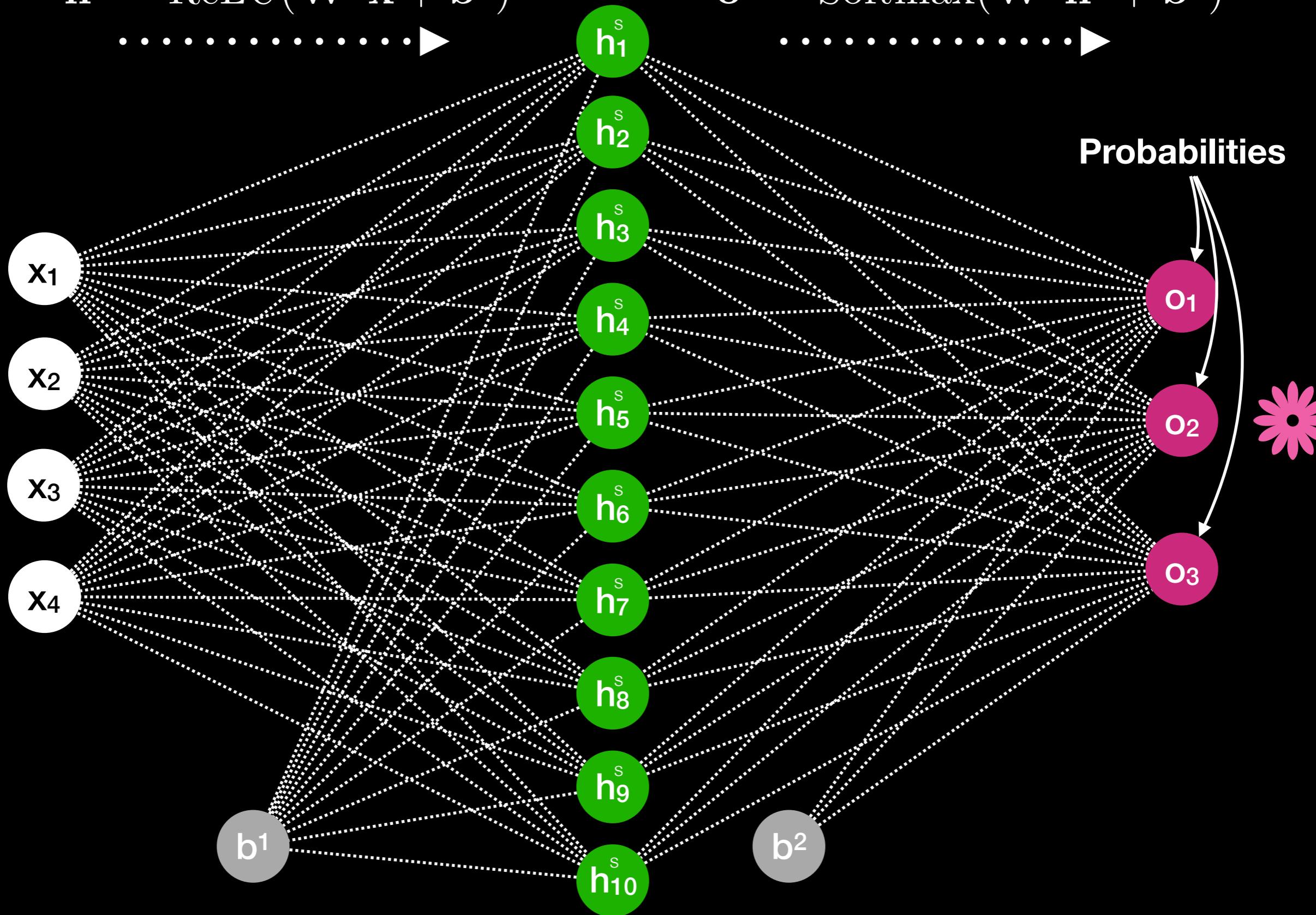
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



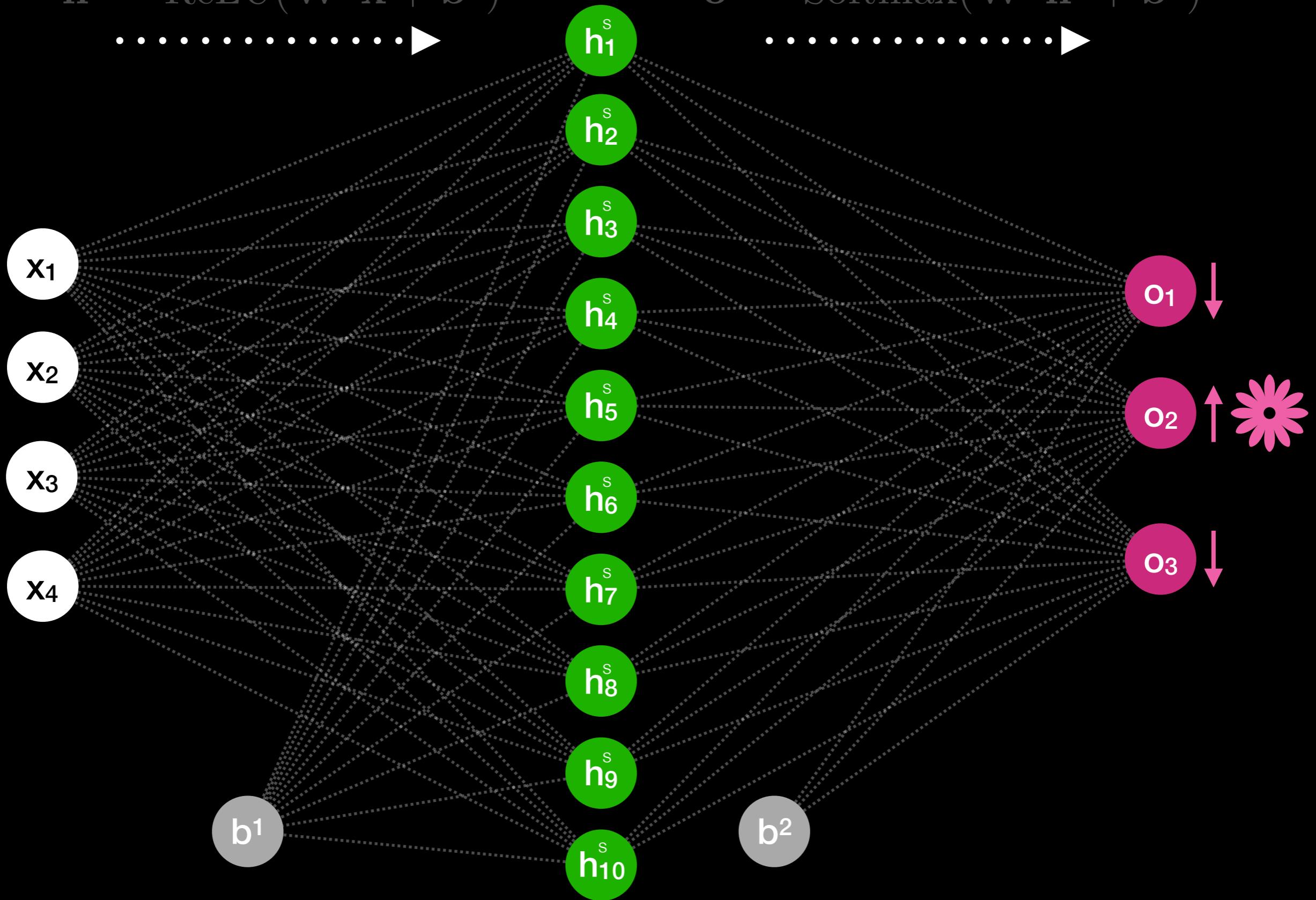
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



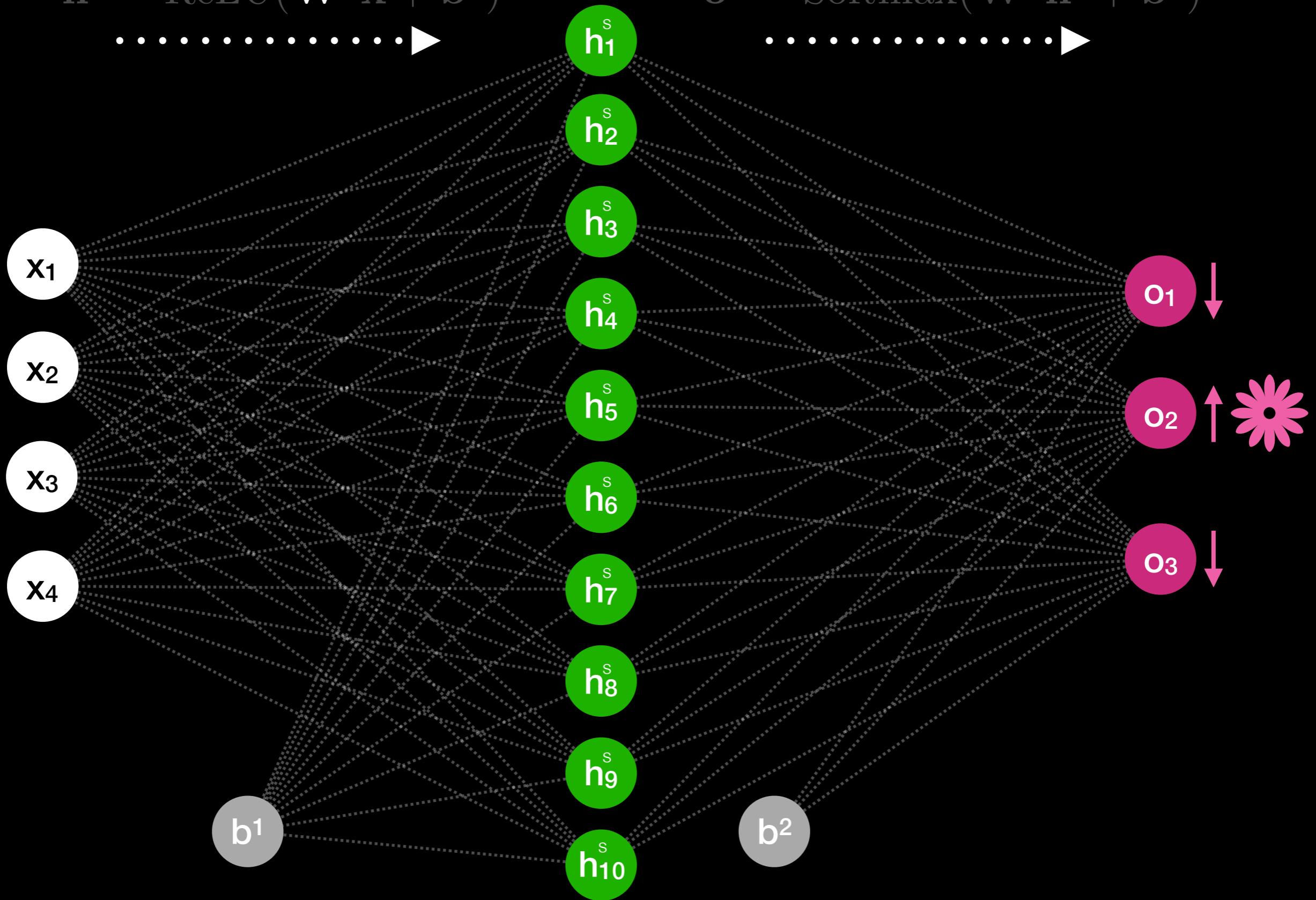
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



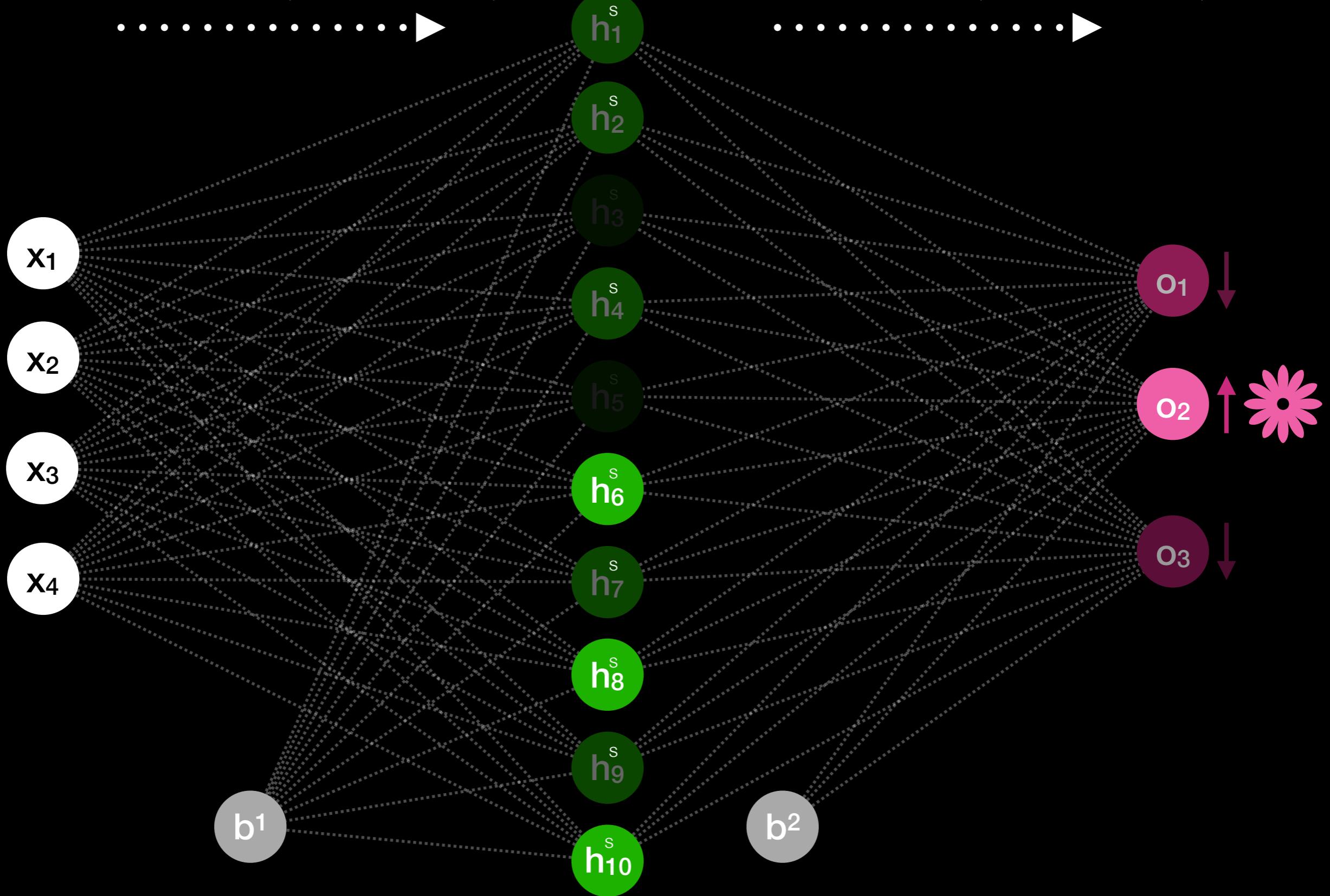
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



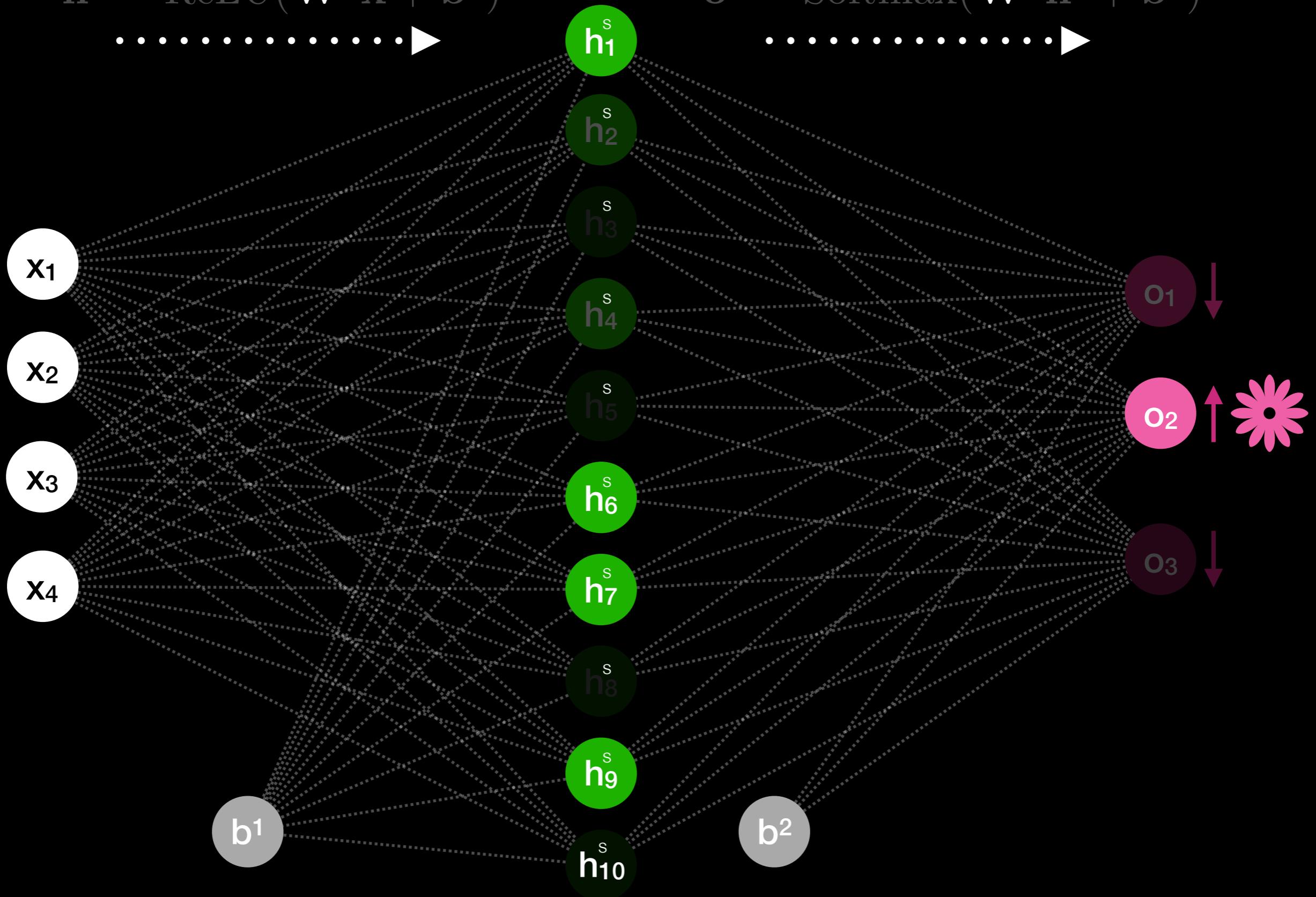
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



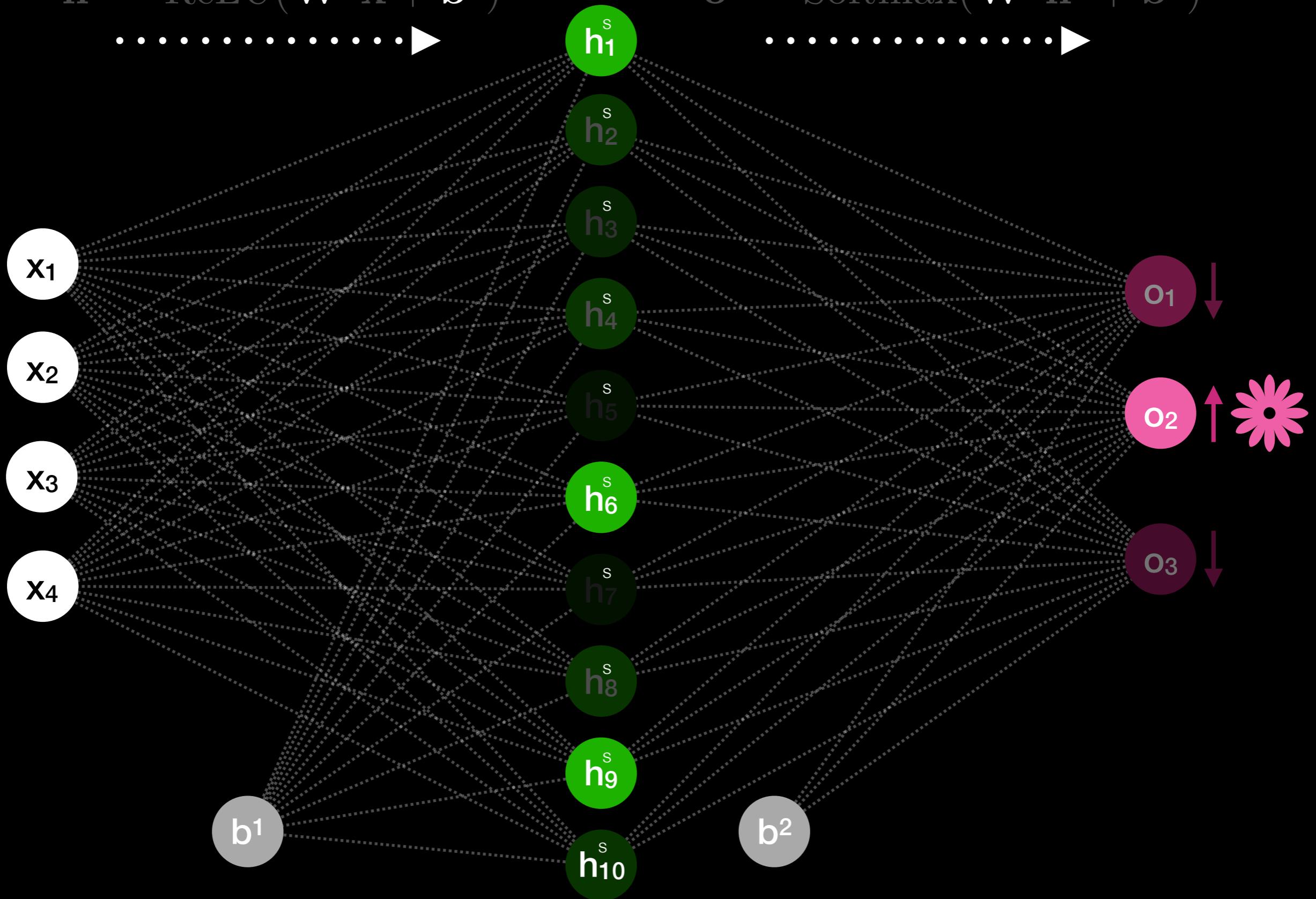
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



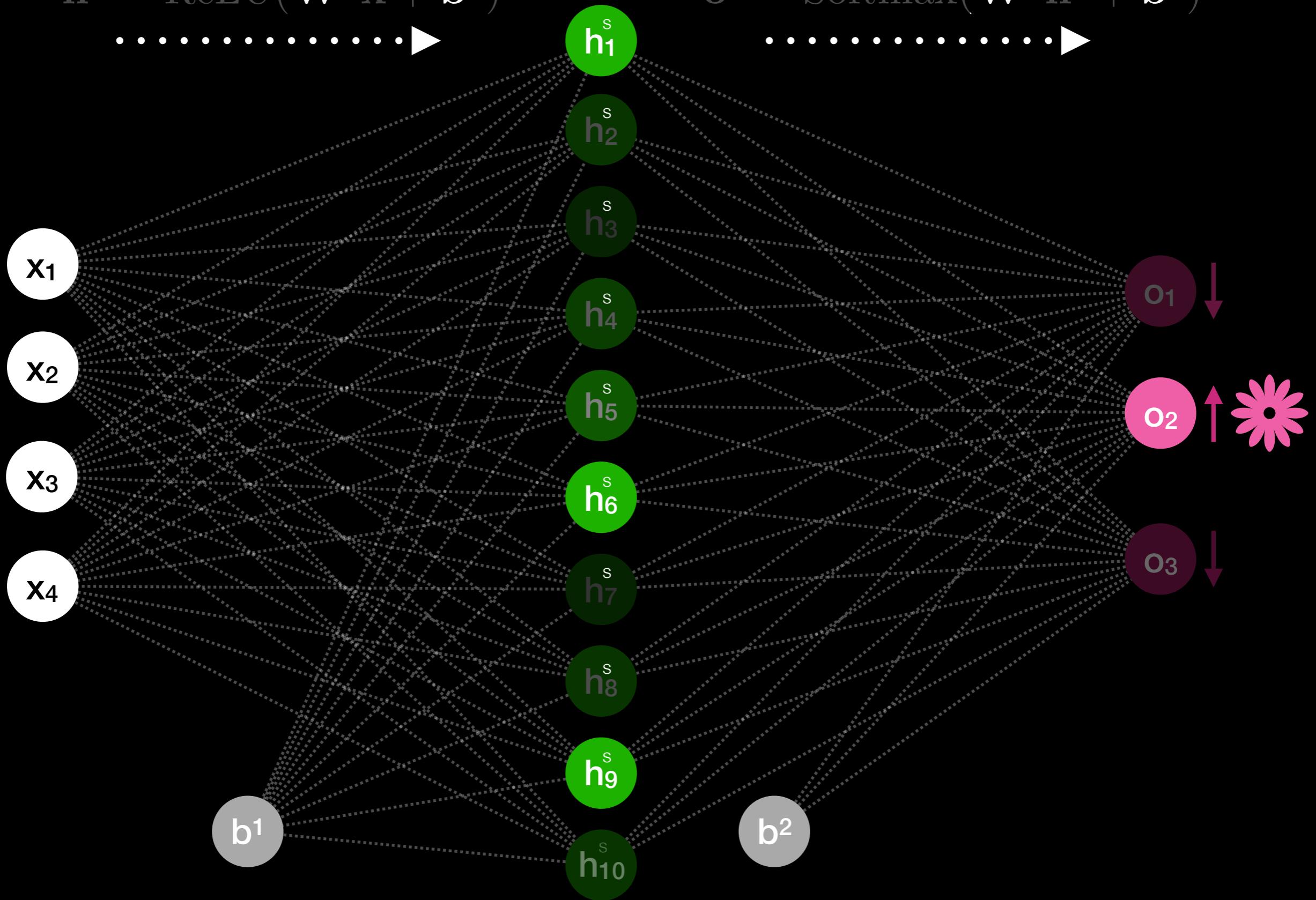
$$\mathbf{h}^s = \text{ReLU}(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1)$$

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



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

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



Optimal parameters?

Backpropagation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^s} \frac{\delta \mathbf{o}^s}{\delta \mathbf{o}^a} = \mathbf{o}^s - \text{onehot}_m(y) \quad (\text{matrix})$$

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

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

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

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^s} \frac{\delta \mathbf{o}^s}{\delta \mathbf{o}^a} = \mathbf{o}^s - \text{onehot}_m(y) \quad (\text{matrix})$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \frac{\delta \mathbf{o}^a}{\delta \mathbf{W}^2} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \right)^T \cdot \mathbf{h}^s \quad (\text{matrix product})$$

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

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

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^s} \frac{\delta \mathbf{o}^s}{\delta \mathbf{o}^a} = \mathbf{o}^s - \text{onehot}_m(y) \quad (\text{matrix})$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{W}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \frac{\delta \mathbf{o}^a}{\delta \mathbf{W}^2} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \right)^T \cdot \mathbf{h}^s \quad (\text{matrix product})$$

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \quad (\text{matrix})$$

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

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

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^s} \frac{\delta \mathbf{o}^s}{\delta \mathbf{o}^a} = \mathbf{o}^s - \text{onehot}_m(y)$$

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a} = \frac{\delta \mathcal{L}}{\delta \mathbf{h}^s} \odot \mathbf{1}_{\mathbf{h}^a(\mathbf{x}) > 0}$$

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^1} = \frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}$$

Basic matrix operations

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^1} = \frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}$$

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

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

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^1} = \frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}$$

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

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

Gradients points along the greatest rate of increase

Objective: Minimize \mathcal{L}

$$\begin{aligned}\nabla_{\mathbf{W}^2} \mathcal{L} &= \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \frac{\delta \mathbf{o}^a}{\delta \mathbf{W}^2} = \left(\frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \right)^\top \cdot \mathbf{h}^s & \nabla_{\mathbf{W}^1} \mathcal{L} &= \left(\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a} \right)^\top \cdot \mathbf{x} \\ \nabla_{\mathbf{b}^2} \mathcal{L} &= \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} & \nabla_{\mathbf{b}^1} \mathcal{L} &= \left(\frac{\delta \mathcal{L}}{\delta \mathbf{h}^a} \right)\end{aligned}$$

Gradients points along the greatest rate of increase

Objective: Minimize \mathcal{L}

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

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

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

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

$$\begin{aligned}\mathbf{W}^2 &\leftarrow \mathbf{W}^2 - \eta \nabla_{\mathbf{W}^2} \mathcal{L} \\ \mathbf{b}^2 &\leftarrow \mathbf{b}^2 - \eta \nabla_{\mathbf{b}^2} \mathcal{L}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^1 &\leftarrow \mathbf{W}^1 - \eta \nabla_{\mathbf{W}^1} \mathcal{L} \\ \mathbf{b}^1 &\leftarrow \mathbf{b}^1 - \eta \nabla_{\mathbf{b}^1} \mathcal{L}\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Optimisation problem