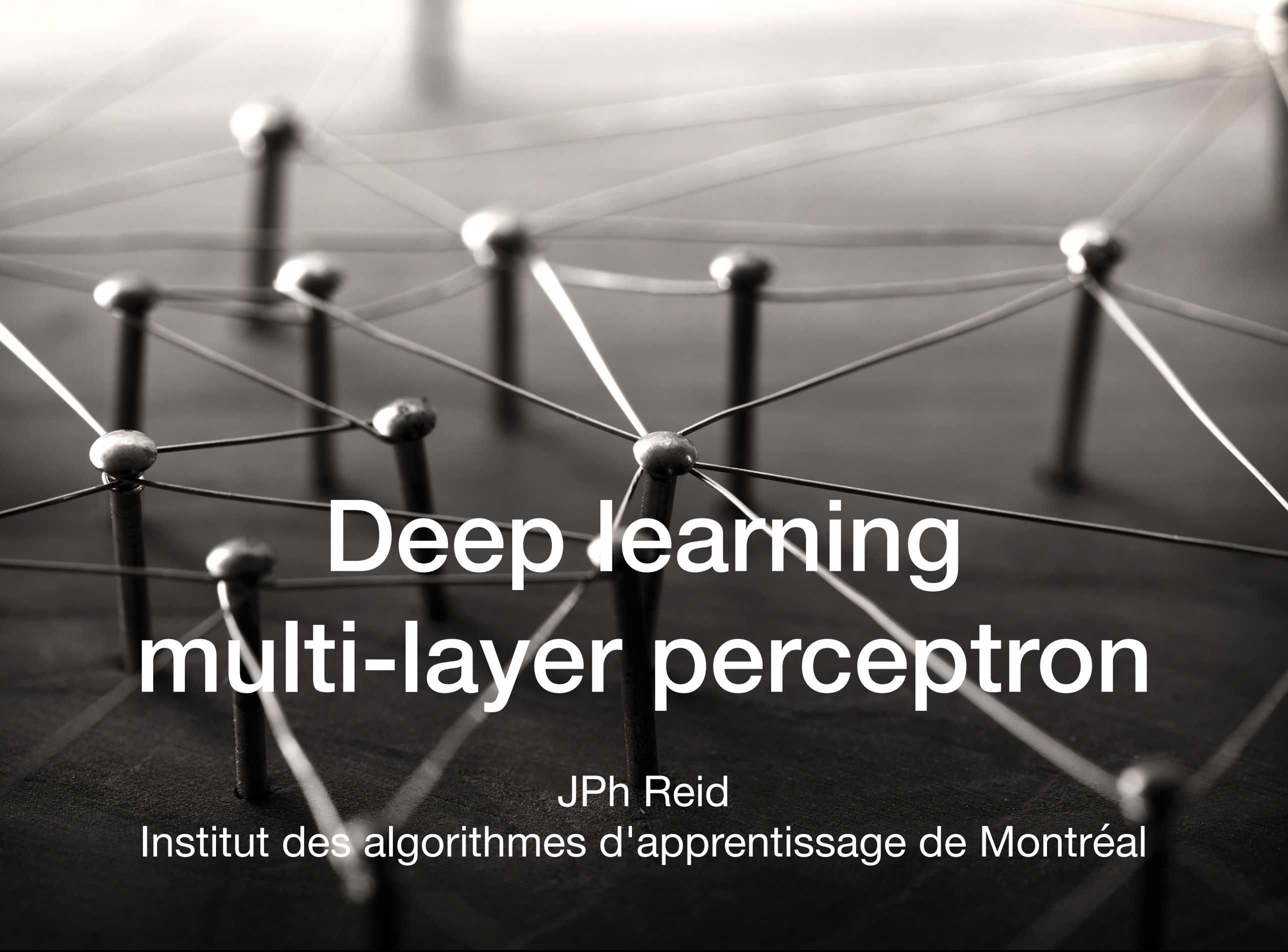


Deep learning multi-layer perceptron

JPh Reid
Institut des algorithmes d'apprentissage de Montréal

Machine learning

1913, A. Markov

1950, A. Turing

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

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

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

Length and width of sepals

Length and width of petals

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

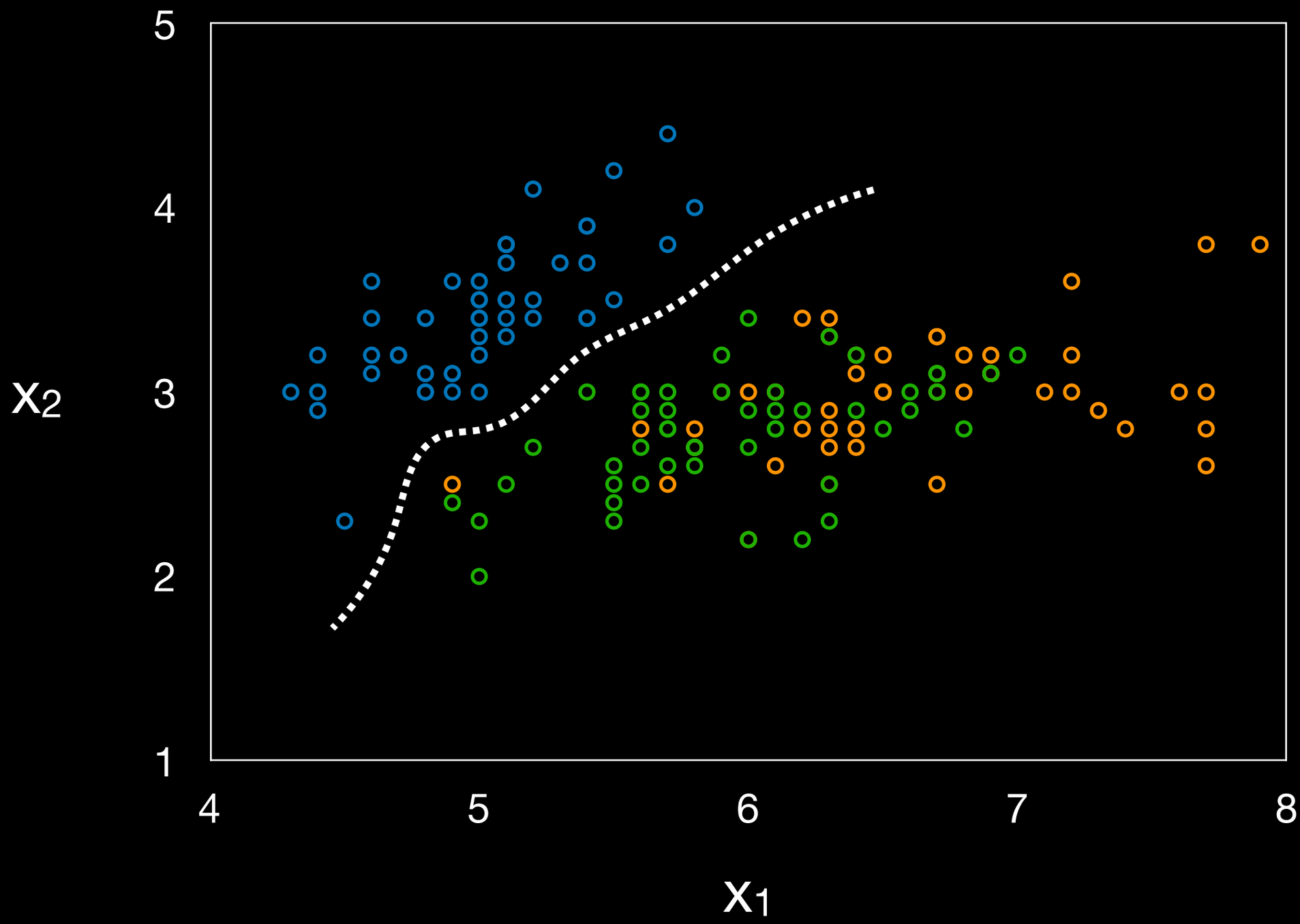
Setosa

Virginica

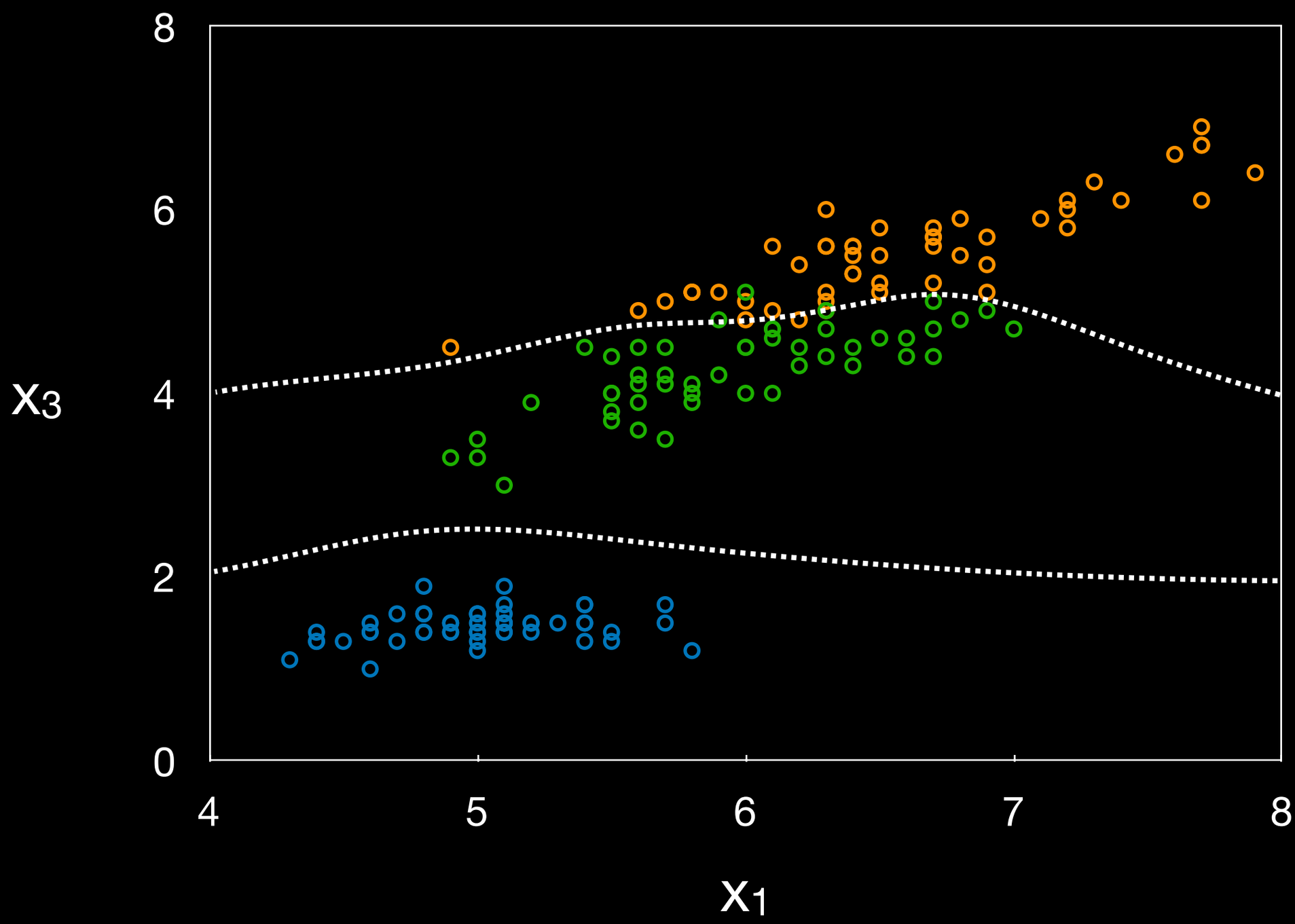
Virginica

$$\mathbf{x} = [\, x_1, x_2, x_3, x_4 \,] \qquad \mathbf{y} = [\, y_1, y_2, y_3 \,]$$

$$\mathbf{x} = [\, x_1, x_2, x_3, x_4 \,] \qquad \mathbf{y} = [\, y_1, y_2, y_3 \,]$$



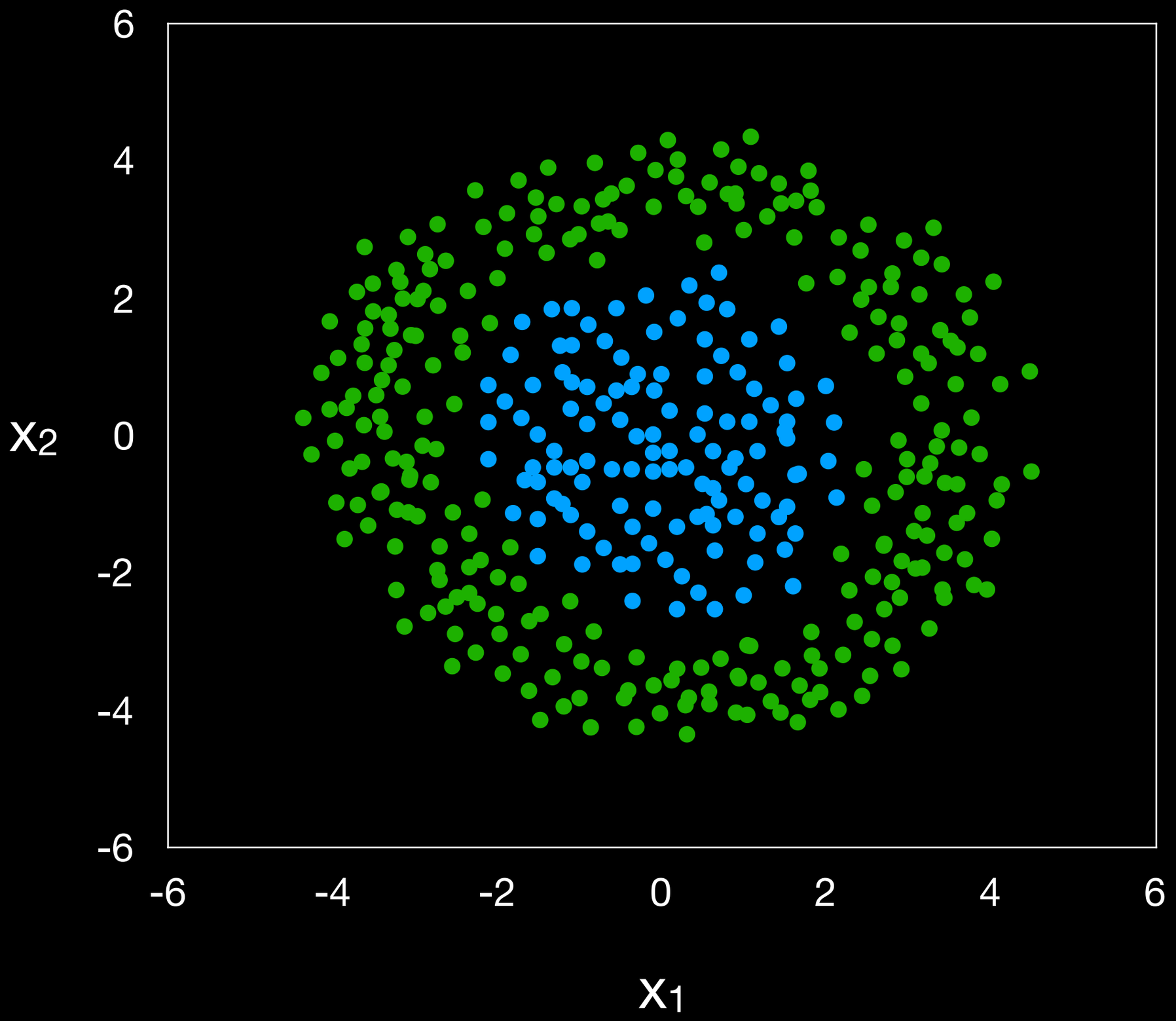
$$\mathbf{x} = [\, x_1, x_2, x_3, x_4 \,] \qquad \mathbf{y} = [\, y_1, y_2, y_3 \,]$$

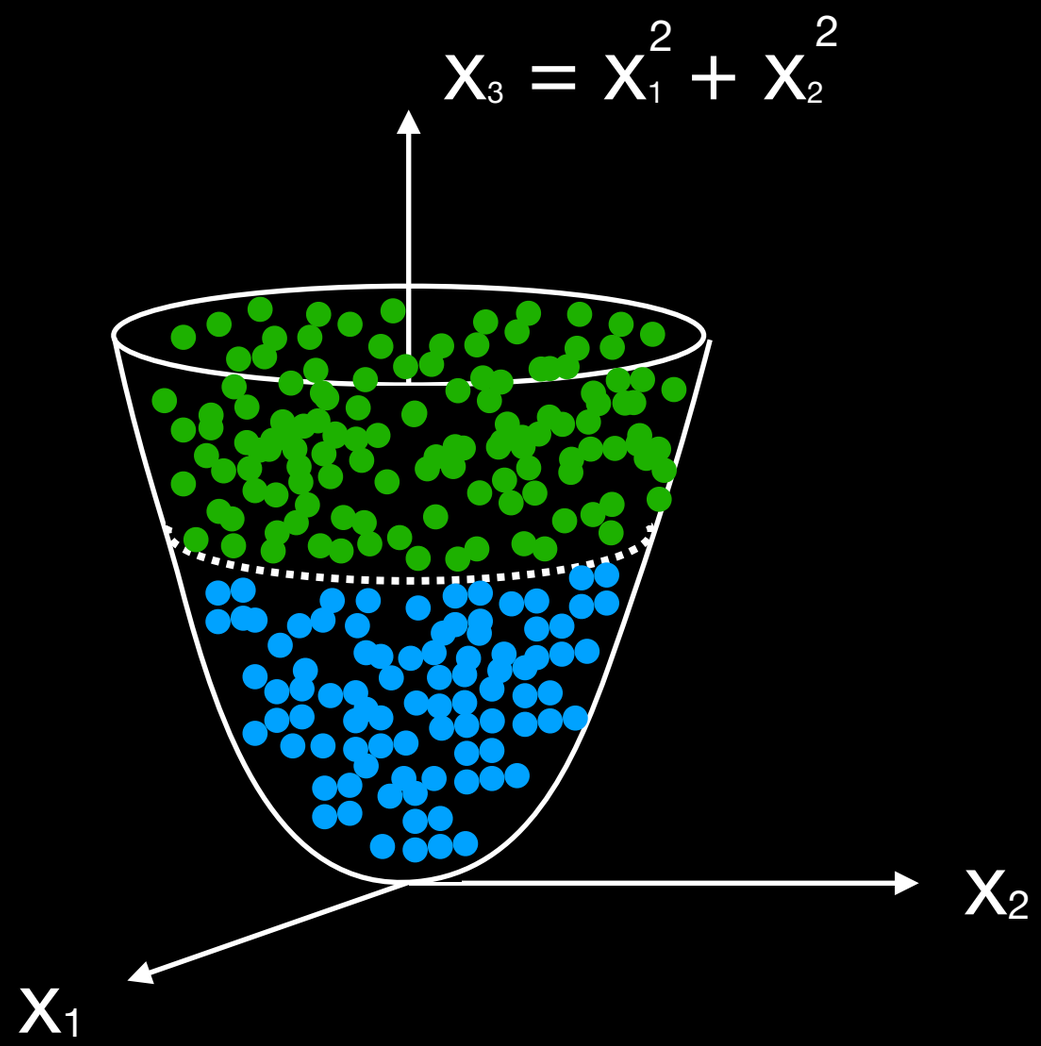


$$\mathbf{x} = [\, x_1, x_2, x_3, x_4 \,] \qquad \mathbf{y} = [\, y_1, y_2, y_3 \,]$$

$$\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \end{bmatrix}$$

$$\mathbf{y} = \mathbf{wx} + b$$


6
4
2
0
-2
-4
-6
x_2
-6
-4
-2
0
2
4
6
x_1


$x_3 = x_1^2 + x_2^2$
x_2
x_1

Deep learning

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

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

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

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

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

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

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

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

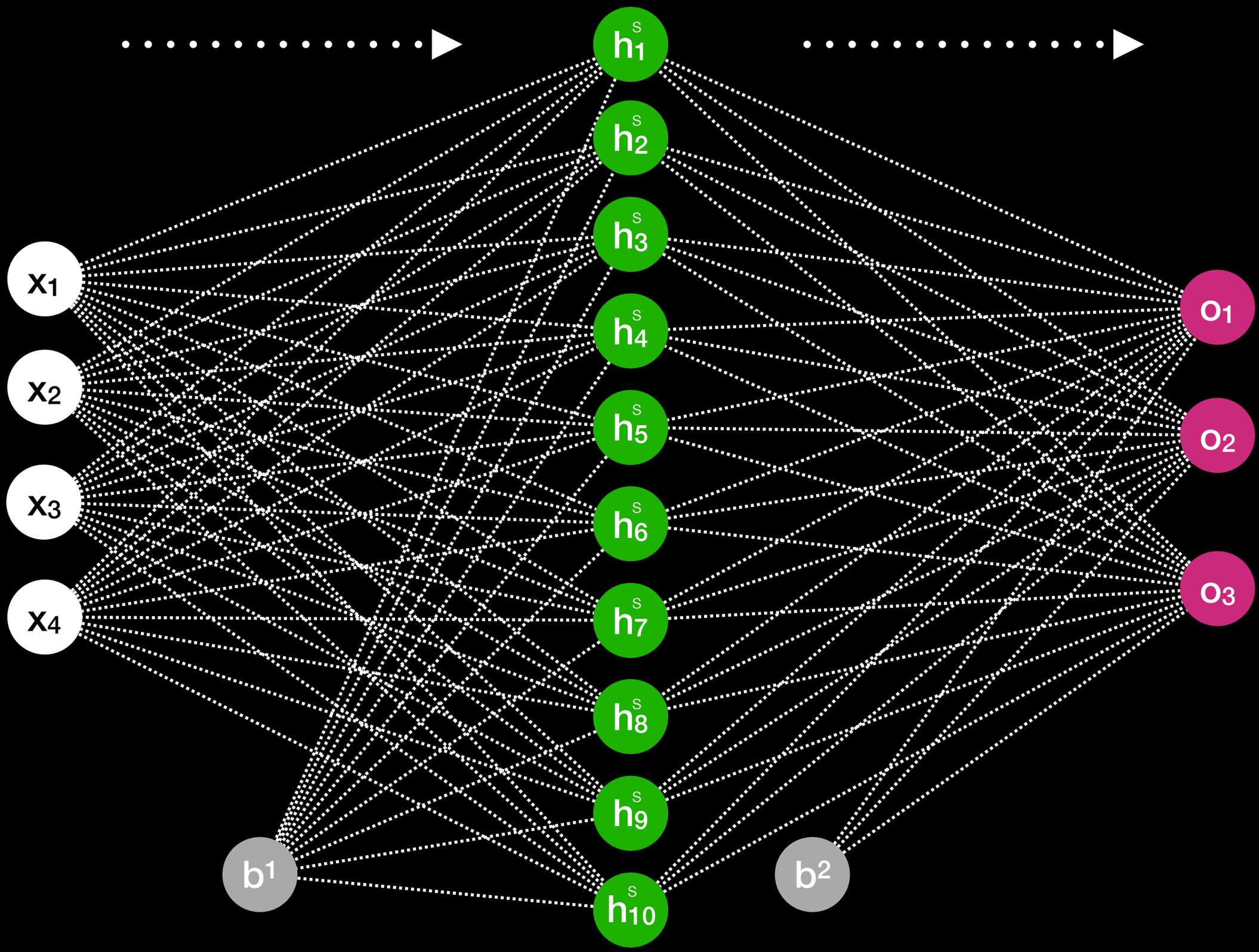
$$\mathbf{o} = \mathbf{W}^2\mathbf{h}^1 + \mathbf{b}^2$$

$$\mathbf{h}^1$$

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

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

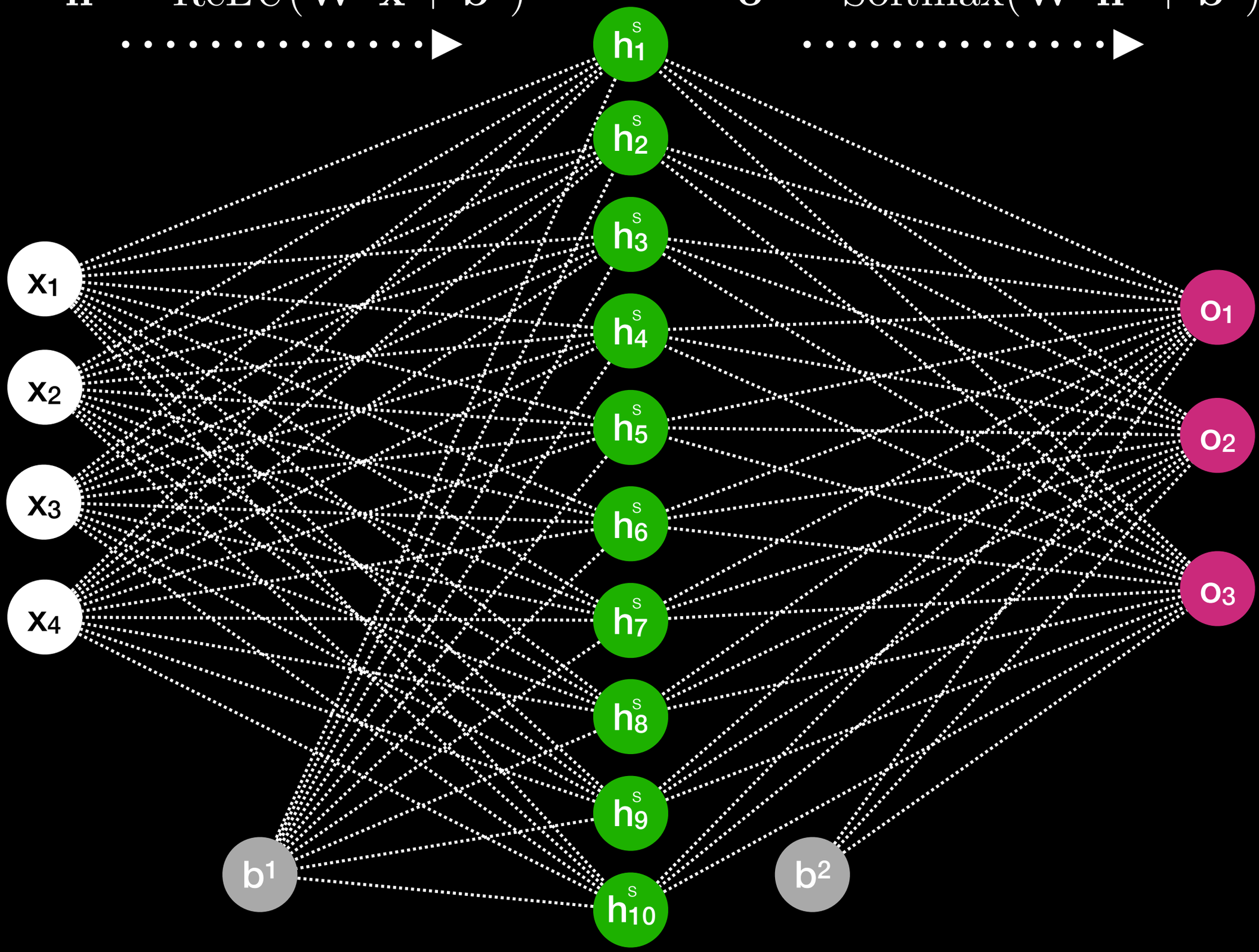
$$\mathbf{o} = \mathbf{W}^2\mathbf{h}^1 + \mathbf{b}^2$$



Example

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

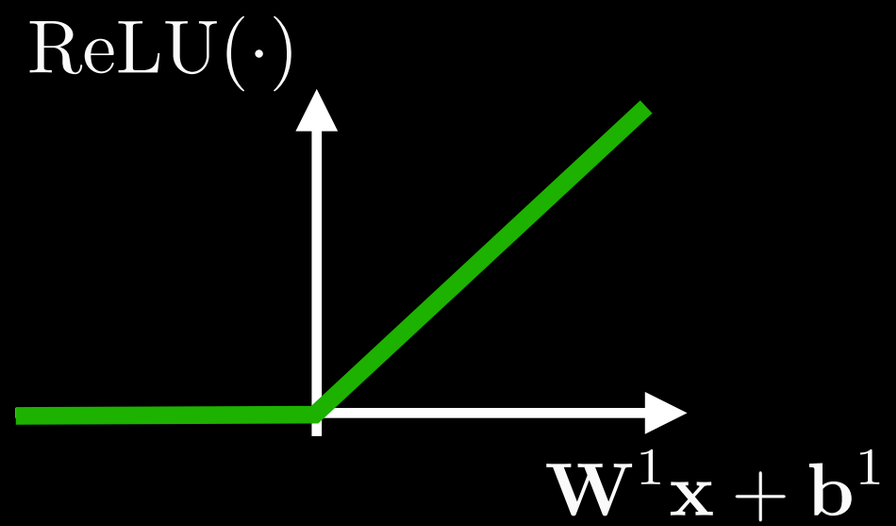
$$\mathbf{o}^s = \mathrm{Softmax}(\mathbf{W}^2\mathbf{h}^s + \mathbf{b}^2)$$



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

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

$$\mathrm{Softmax_i}(\mathbf{o}^a) = \frac{\exp(\mathrm{o_i^a})}{\sum_k \exp(\mathrm{o_k^a})}$$

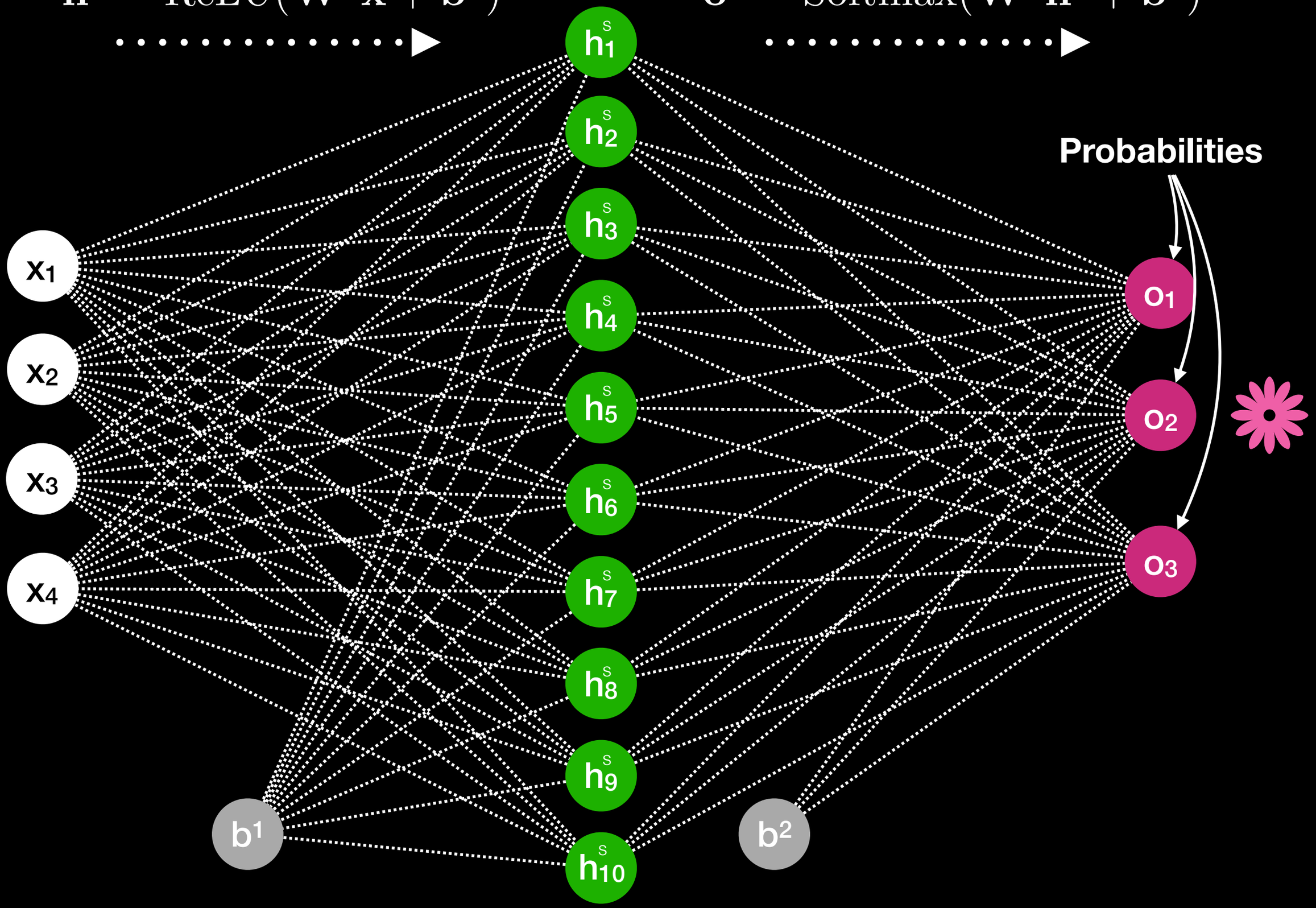


$$\mathcal{L} = -\log p(\mathbf{y})$$

Negative Log-likelihood

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

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



36 ans
canadien
célibataire
féministe
H2G 2K9
PhD
course
vélo
aucune allergie
40-60 heures
végétarien
droitier
aucun enfant
intègre
anxieux
patient
leader
persévérant
.....

DL

Application rejected

Training data

Interprétabilité

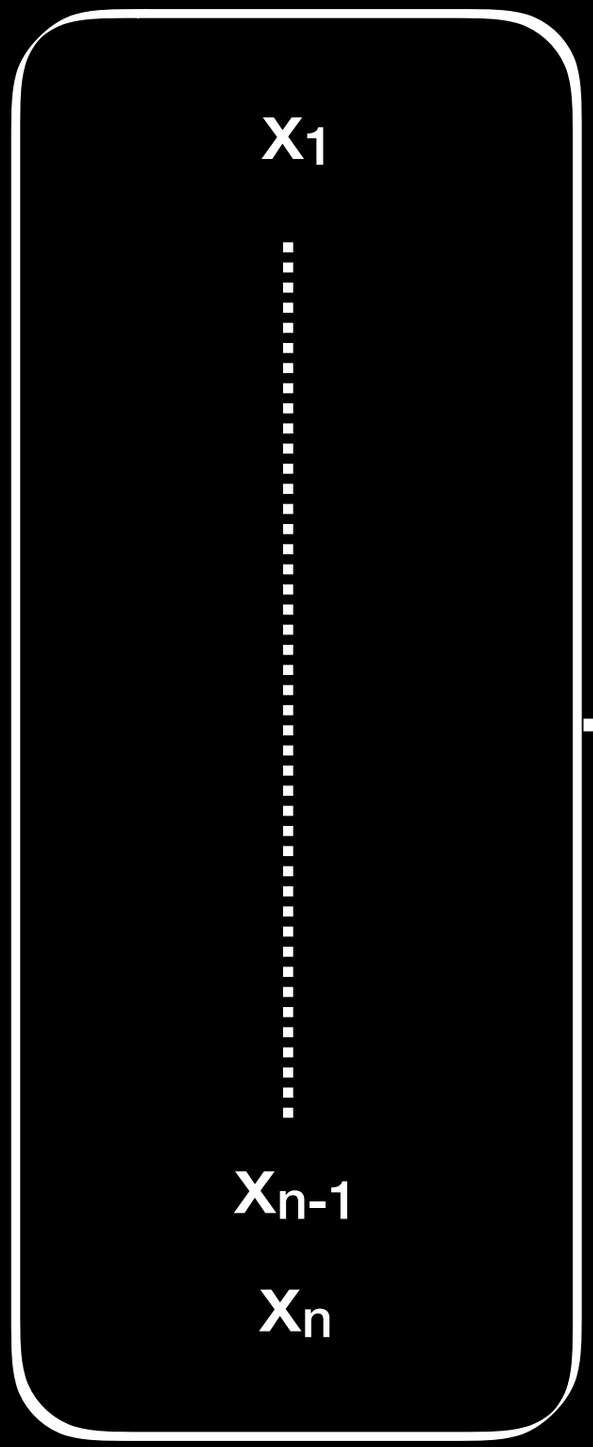
36 ans
canadien
célibataire
féministe
H2G 2K9
PhD
course
vélo
aucune allergie
40-60 heures
végétarien
droitier
aucun enfant
intègre
anxieux
patient
leader
persévérant
.....

DL

Application rejected

Olivier
Sven
Francis
Ramzy
Maude
Fazel
Adrien
Sophie
Arezoo
David
Samuel
Elena
Simon
Alexandre
Étienne
Gaël
Clement
Nicolas
Amirreza
JBay
Marie-Ève
Louis
Bastien
Patrick
Xigang
Shiyan
Hamideh

Interprétabilité



DL

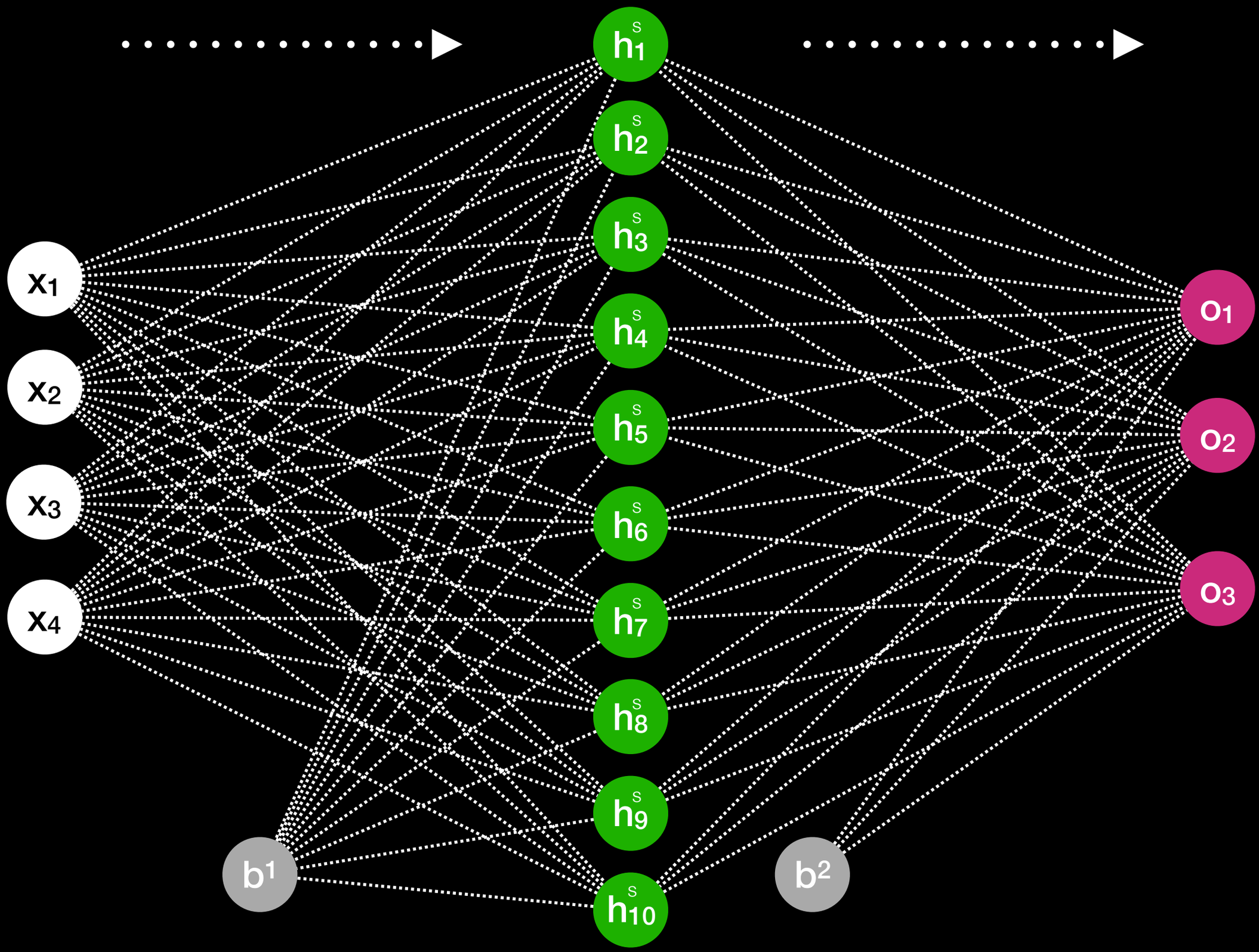
....

Interprétabilité

Olivier
Sven
Ramzy
Fazel
Adrien
Maude
Sophie
David
Samuel
Arezoo
Elena
Alexandre
Étienne
Simon
Clement
Gaël
Nicolas
Amirreza
JBay
Marie-Ève
Louis
Bastien
Patrick
Xigang
Shiyan
Hamideh

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

$$\mathbf{o} = \mathbf{W}^2\mathbf{h}^1 + \mathbf{b}^2$$



Backpropagation algorithm

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

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

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

Length and width of sepals

Length and width of petals

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

Setosa

Virginica

Virginica

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

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

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

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

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

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

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

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

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

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

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

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

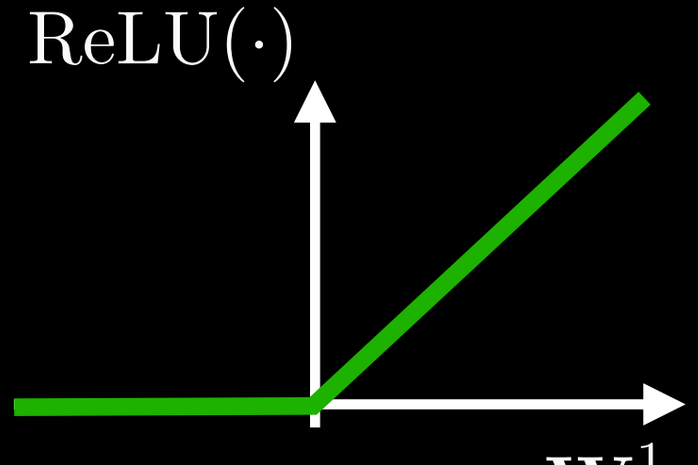
$$\begin{bmatrix} p(y_1) \\ p(y_2) \\ p(y_3) \end{bmatrix} = \text{S}\left(\begin{bmatrix} w^1_{11} & w^1_{12} & \cdots & w^1_{1n} \\ w^1_{21} & w^1_{22} & \cdots & w^1_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w^1_{m1} & w^1_{m2} & \cdots & w^1_{nn} \end{bmatrix} \text{ReLU}\left(\begin{bmatrix}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

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

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

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

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



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

Log-likelihood

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

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

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

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

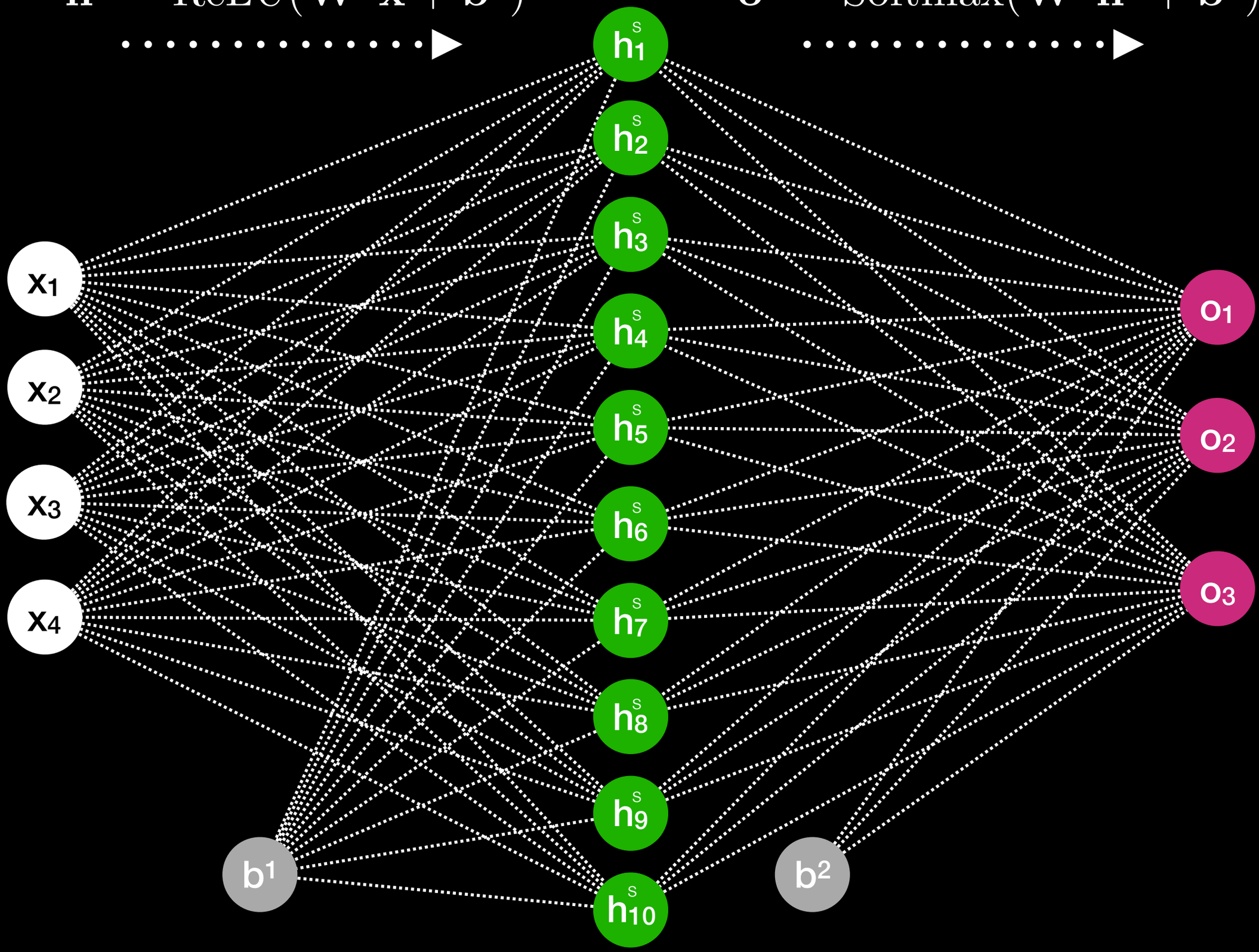
$\mathbf{h}^s$

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

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

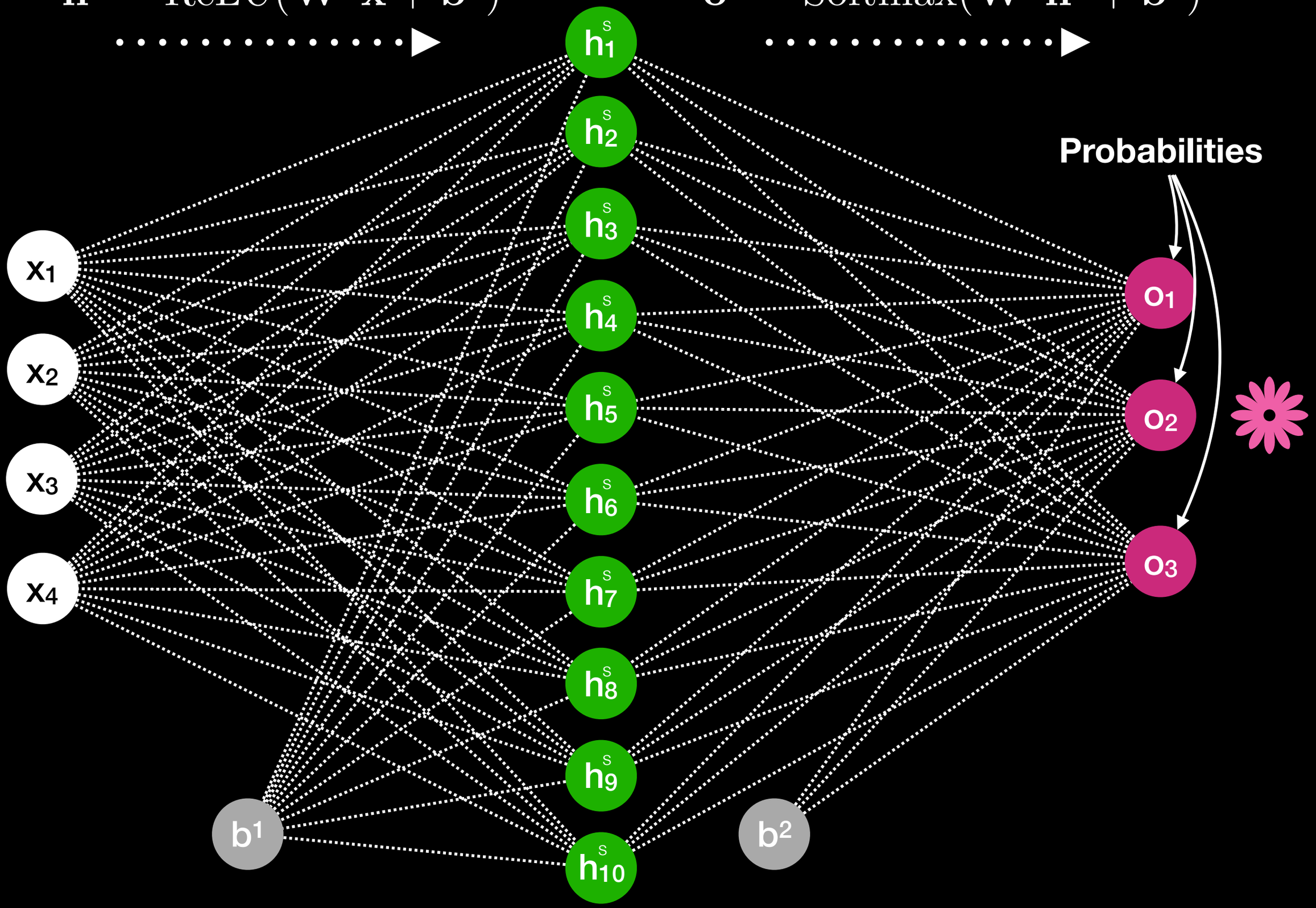
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



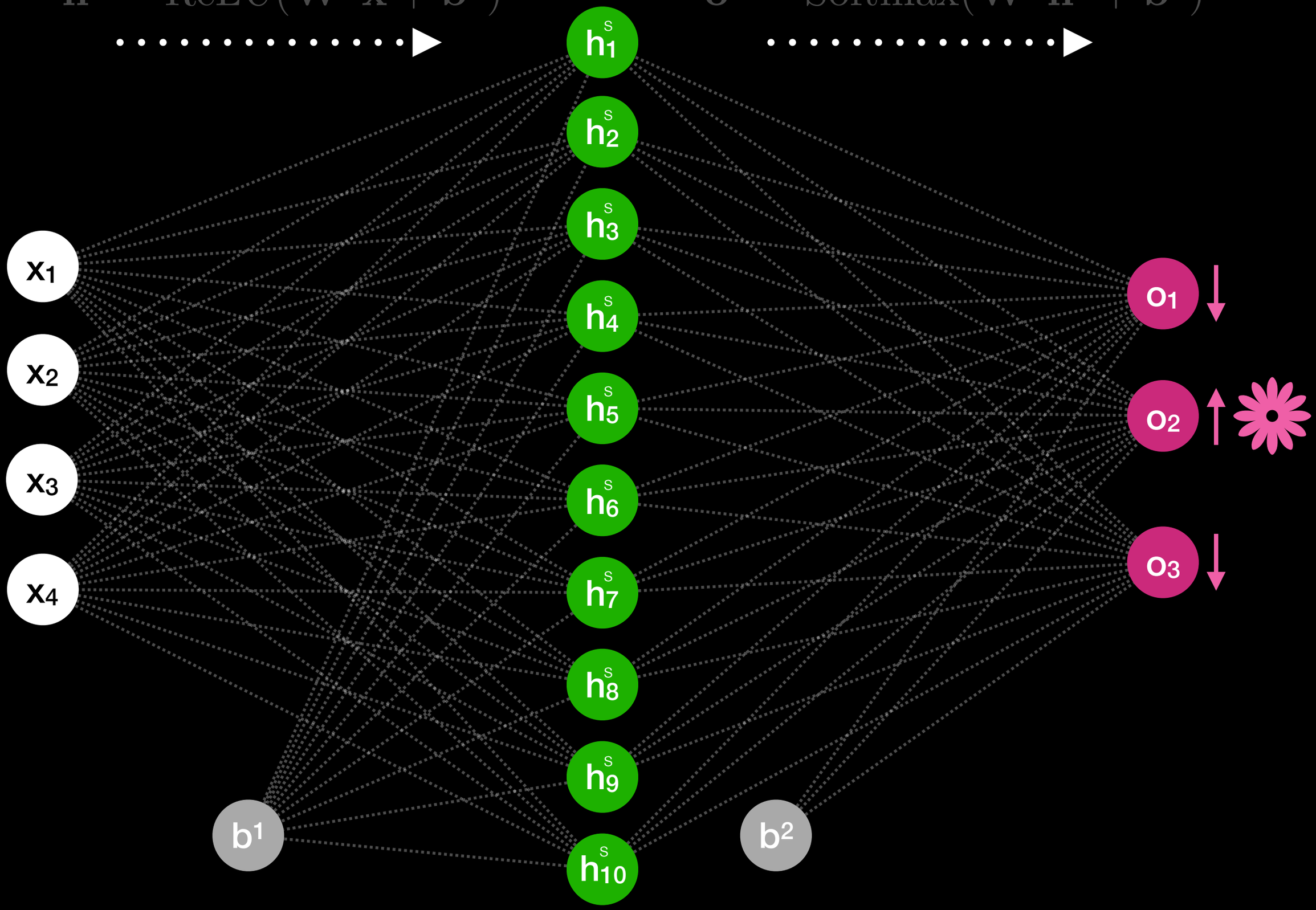
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



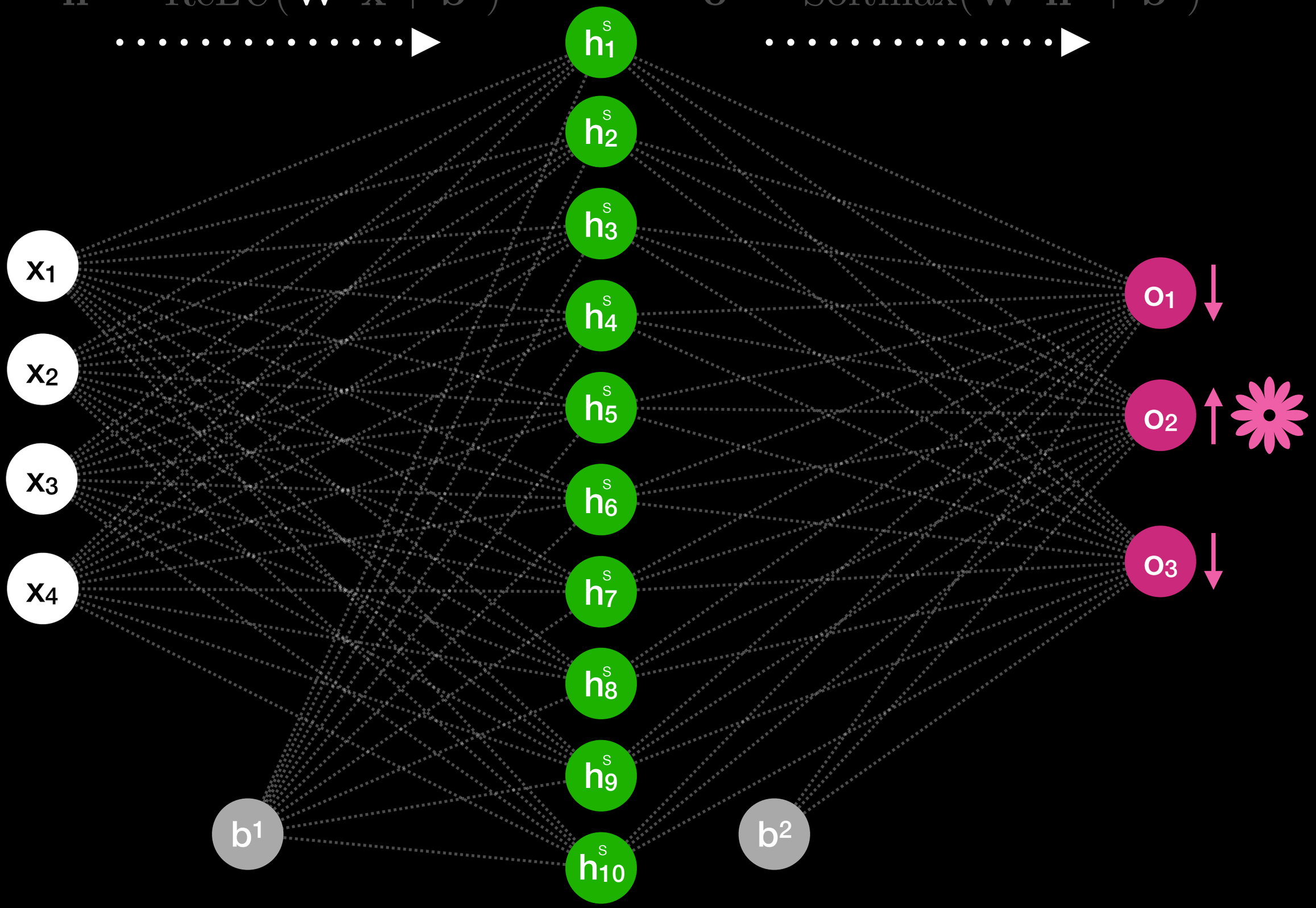
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



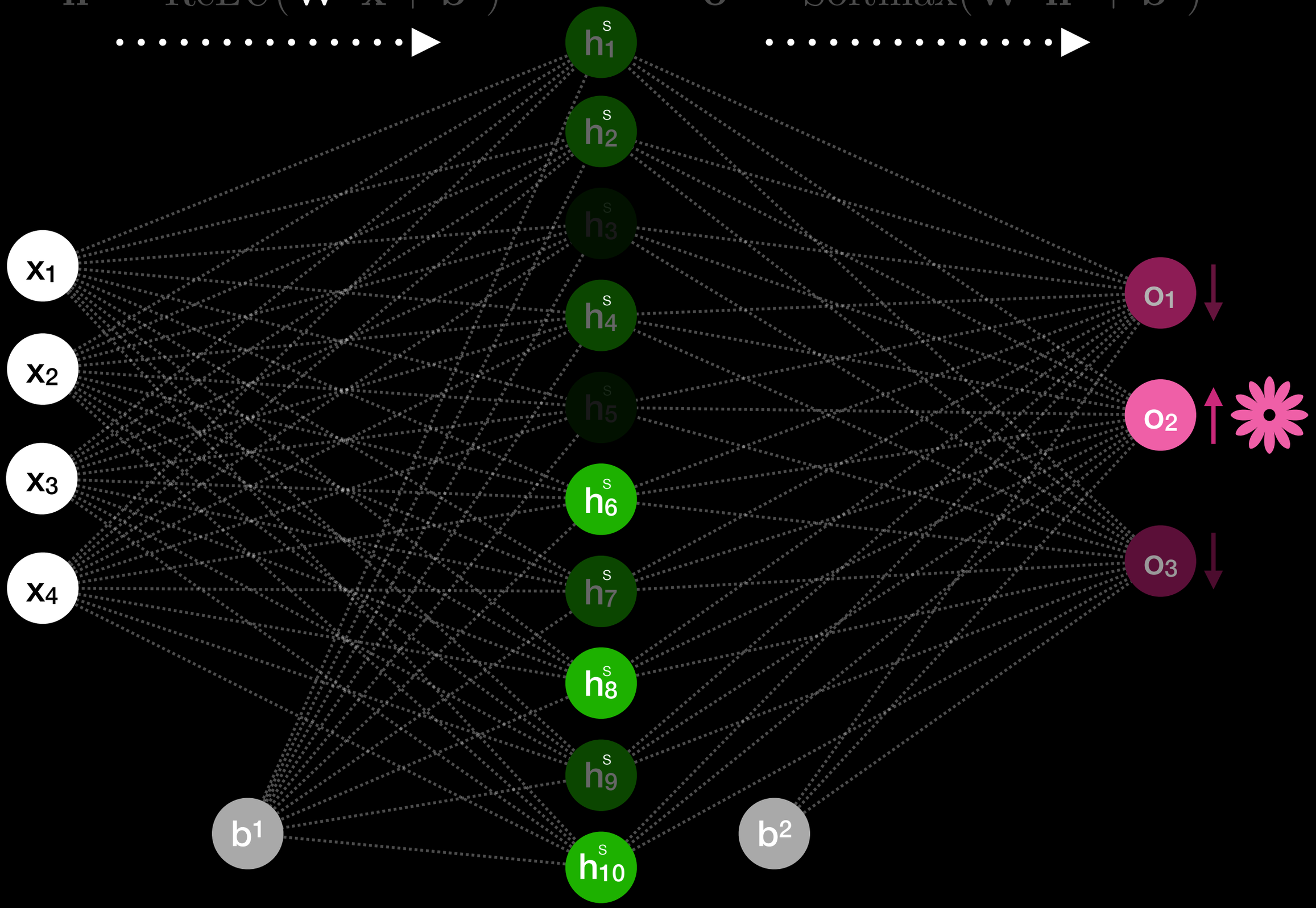
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



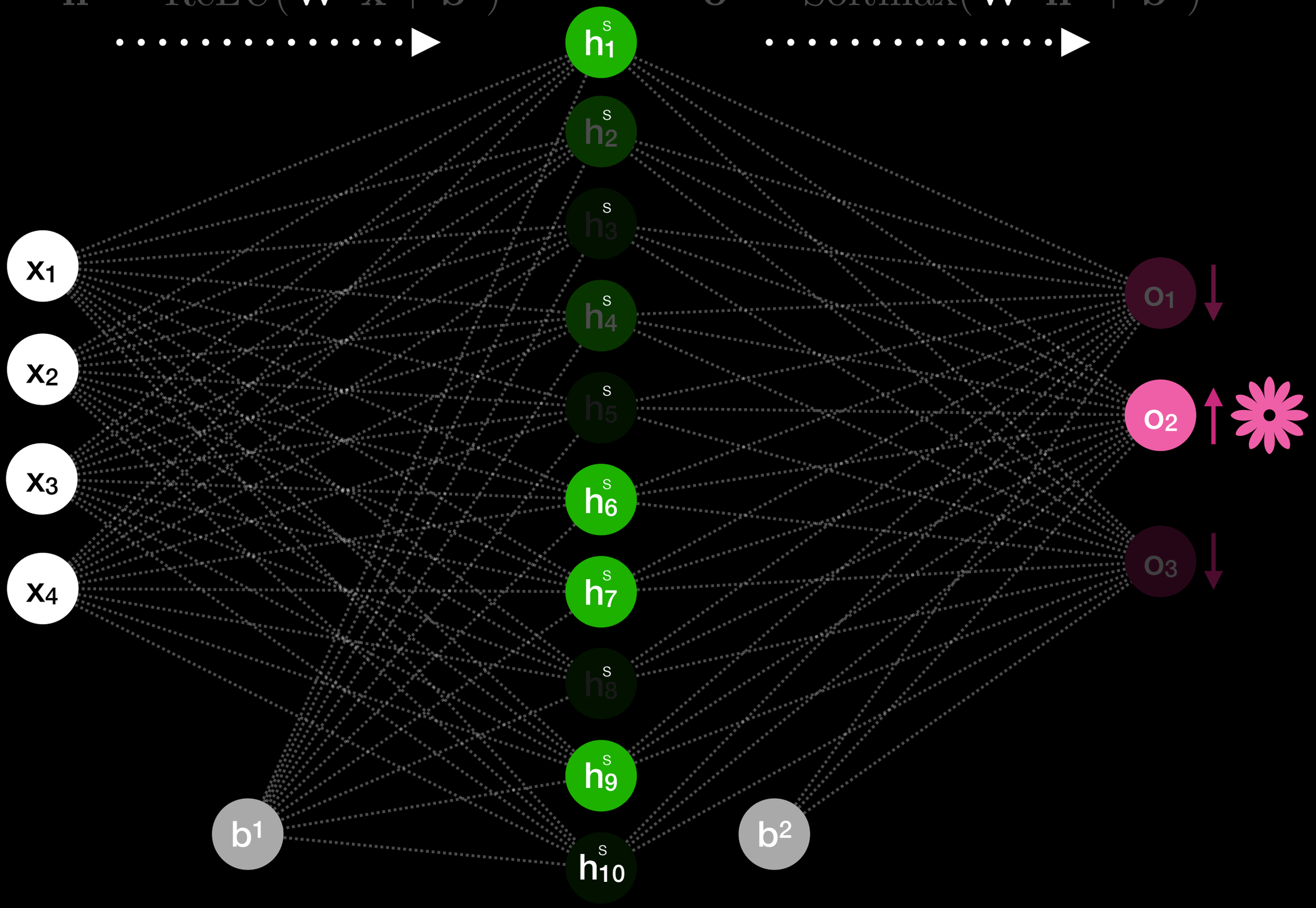
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



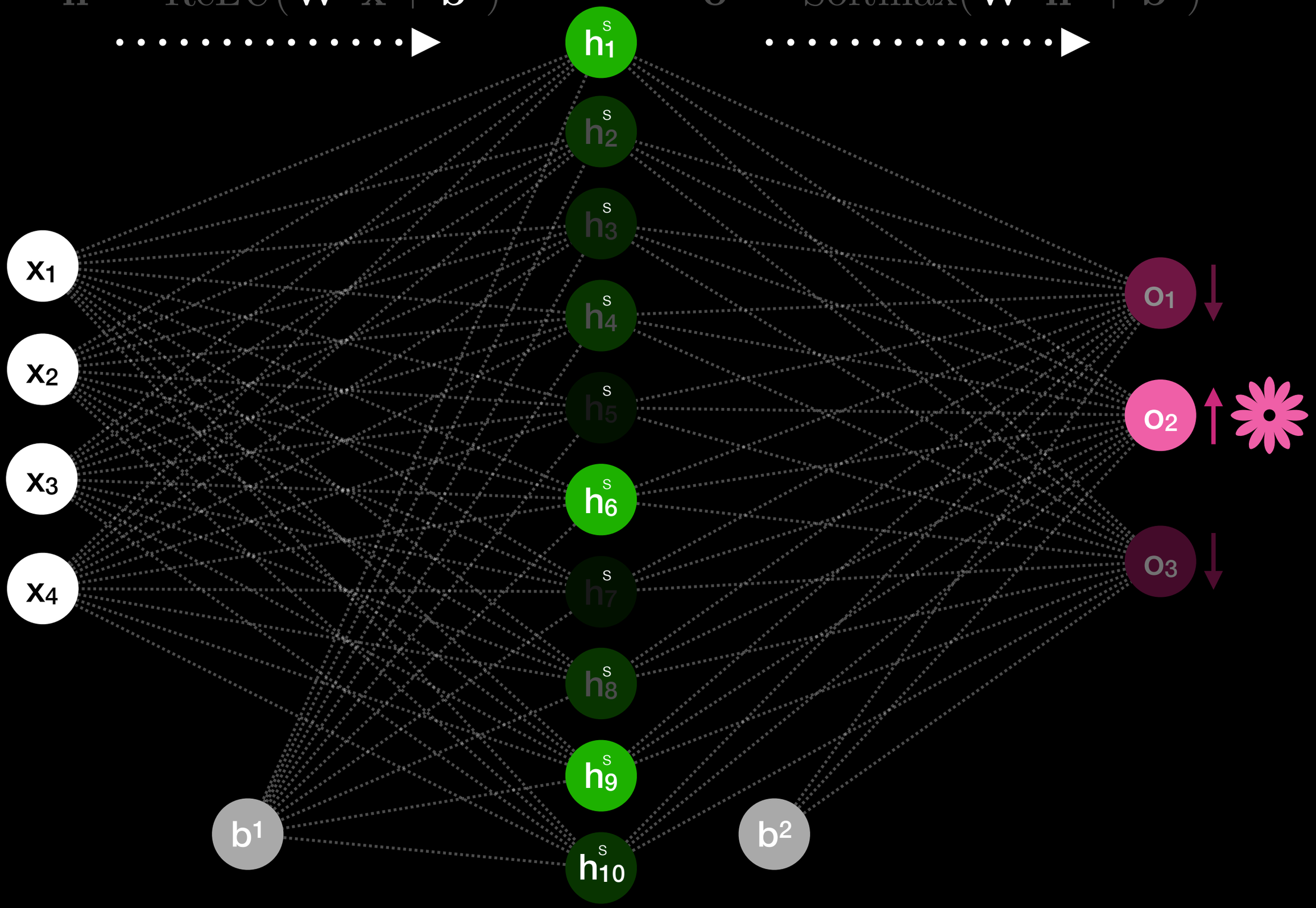
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



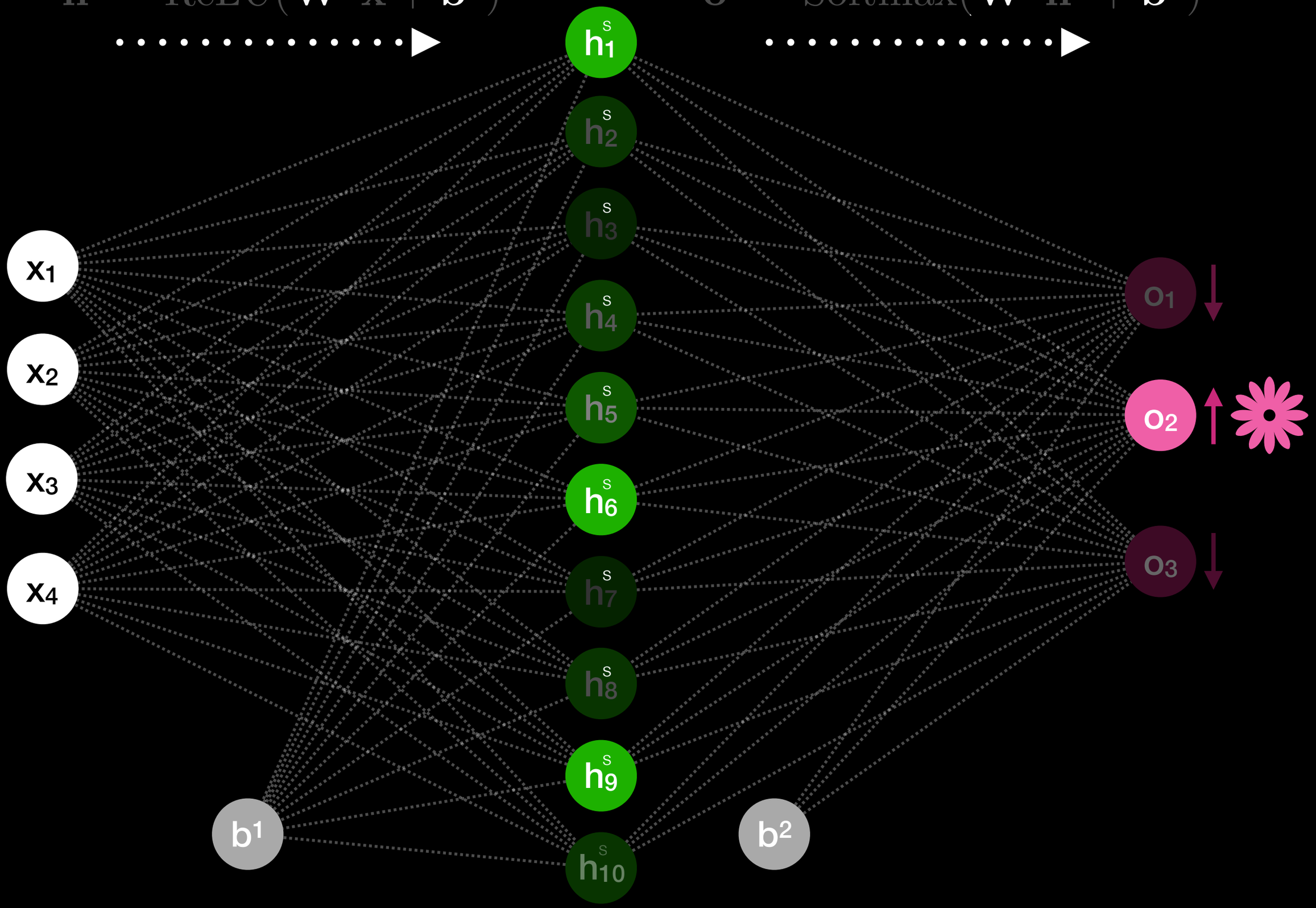
$$\mathbf{h}^s = \mathrm{ReLU}(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

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



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

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



Optimal parameters?

Backpropagation

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

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

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

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

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

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

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

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

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

$$\frac{\delta\mathcal{L}}{\delta\mathbf{o}^a} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^s}\frac{\delta\mathbf{o}^s}{\delta\mathbf{o}^a} = \mathbf{o}^s - \mathrm{onehot}_{\mathrm{m}}(\mathrm{y}) \quad \text{(matrix)}$$
$$\frac{\delta\mathcal{L}}{\delta\mathbf{W}^2} = \frac{\delta\mathcal{L}}{\delta\mathbf{o}^a}\frac{\delta\mathbf{o}^a}{\delta\mathbf{W}^2}$$

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

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

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

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

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

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

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

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

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

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

Basic matrix operations

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

$$\frac{\delta \mathcal{L}}{\delta \mathbf{b}^2} = \frac{\delta \mathcal{L}}{\delta \mathbf{o}^a} \qquad \frac{\delta \mathcal{L}}{\delta \mathbf{b}^1} = \frac{\delta \mathcal{L}}{\delta \mathbf{h}^a}$$

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

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

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

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

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

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

Gradients points along the greatest rate of increase

Objective: Minimize $\mathcal{L}$

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

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

Gradients points along the greatest rate of increase

Objective: Minimize $\mathcal{L}$

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

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

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

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

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

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

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

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

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

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

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

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

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

Optimisation problem