# ENGR 3321: Introduction to Deep Learning for Robotics

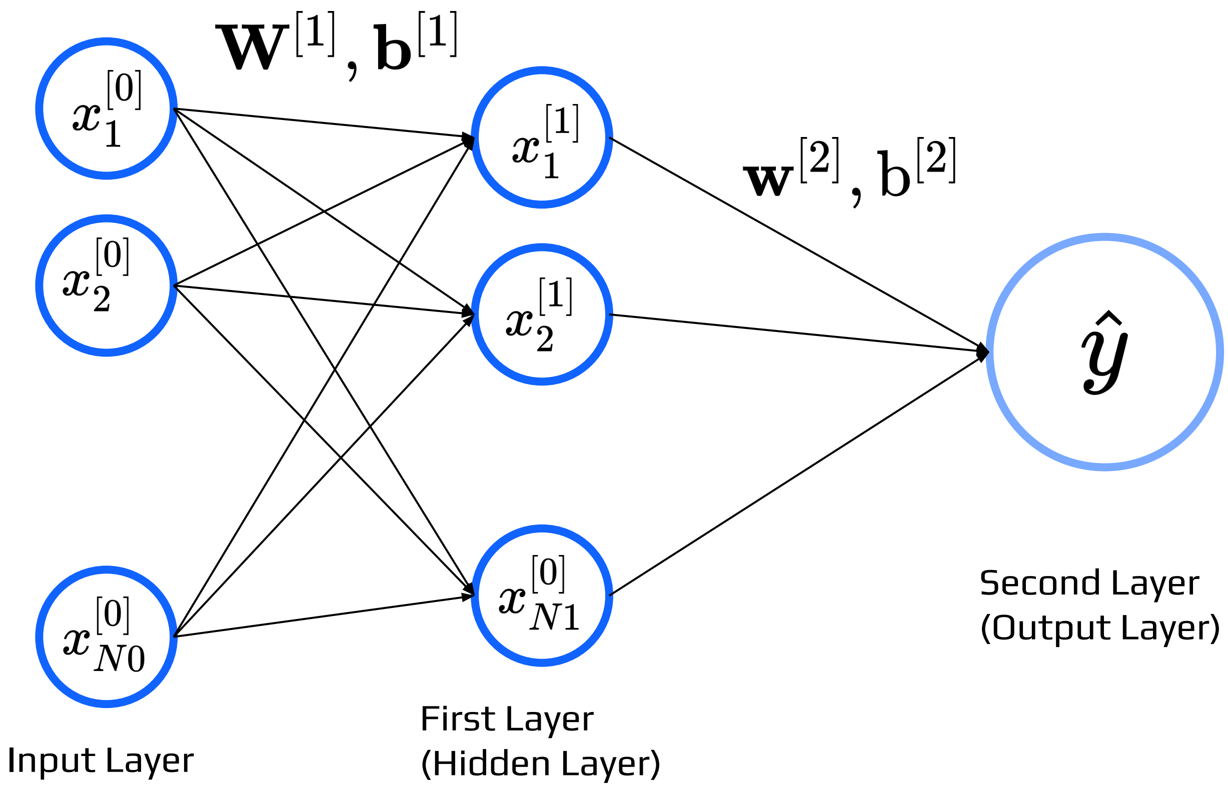
## Neural Network N11: 1-Hidden Layer Model

10/30/2023

# Outline

- Representations
- Layers
- Back-Propagation
- Gradient Descent

# 1 Hidden Layer Neural Network



# Individual Representation

$$\hat{y} = \sigma\left(w_1^{[2]} x_1^{[1]} + w_2^{[2]} x_2^{[1]} + \ldots + w_{N_1}^{[2]} x_{N_1}^{[1]} + b^{[2]}\right)$$

Where

$$x_1^{[1]} = \sigma\left(w_{11}^{[1]} x_1^{[0]} + w_{21}^{[1]} x_2^{[0]} + \ldots + w_{N_0 1}^{[1]} x_{N_0}^{[0]} + b_1^{[1]}\right)$$

$$x_2^{[1]} = \sigma\left(w_{12}^{[1]} x_1^{[0]} + w_{22}^{[1]} x_2^{[0]} + \ldots + w_{N_0 2}^{[1]} x_{N_0}^{[0]} + b_2^{[1]}\right)$$

.

.

.

$$x_{N_1}^{[1]} = \sigma\left(w_{1N_1}^{[1]} x_1^{[0]} + w_{2N_1}^{[1]} x_2^{[0]} + \ldots + w_{N_0 N_1}^{[1]} x_{N_0}^{[0]} + b_{N_1}^{[1]}\right)$$

# Matrix Form

$$
\begin{aligned}
\hat{\mathbf{y}} &= \sigma\left(\mathbf{X}^{[1]} \cdot \mathbf{w}^{[2]\mathrm{T}} + b^{[2]}\right) \\
&= \sigma\left(\sigma\left(\mathbf{X}^{[0]} \cdot \mathbf{W}^{[1]\mathrm{T}} + \mathbf{b}^{[1]}\right) \cdot \mathbf{w}^{[2]\mathrm{T}} + b^{[2]}\right)
\end{aligned}
$$

# Feature (Input) Matrix

$$\mathbf{X}^{[0]} = \begin{bmatrix} {}^{(1)}x_1^{[0]} & {}^{(1)}x_2^{[0]} & \dots & {}^{(1)}x_{N_0}^{[0]} \\ {}^{(2)}x_1^{[0]} & {}^{(2)}x_2^{[0]} & \dots & {}^{(2)}x_{N_0}^{[0]} \\ & & \dots & \\ {}^{(M)}x_1^{[0]} & {}^{(1)}x_2^{[0]} & \dots & {}^{(M)}x_{N_0}^{[0]} \end{bmatrix}_{(M,N_0)}$$

# First-Layer Parameters

$$\mathbf{W}^{[\mathbf{1}]} = \begin{bmatrix} w_{11}^{[1]} & w_{21}^{[1]} & \dots & w_{N_01}^{[1]} \\ w_{12}^{[1]} & w_{22}^{[1]} & \dots & w_{N_02}^{[1]} \\ & & \dots & \\ w_{1N_1}^{[1]} & w_{2N_1}^{[1]} & \dots & w_{N_0N_1}^{[1]} \end{bmatrix}_{(N_1,N_0)}$$

$$\mathbf{b}^{[\mathbf{1}]} = \begin{bmatrix} b_1^{[1]} & b_2^{[1]} & \dots & b_{N_1}^{[1]} \end{bmatrix}_{(1,N_1)}$$

# Second-Layer Parameters

$$\mathbf{w}^{[2]} = \begin{bmatrix} w_1^{[2]} & w_2^{[2]} & \dots & w_{N_1}^{[2]} \end{bmatrix}_{(1,N_1)}$$

$b^{[2]}$, scalar

# Forward Propagation

$$\mathbf{X}^{[1]} = \sigma\left(\mathbf{X} \cdot \mathbf{W}^{[1]\mathrm{T}} + \mathbf{b}^{[1]}\right) = \sigma\left(\mathbf{Z}^{[1]}\right)$$

$$\mathbf{X}^{[1]} = \sigma\left(\begin{bmatrix} {}^{(1)}x_1^{[0]} & {}^{(1)}x_2^{[0]} & \dots & {}^{(1)}x_{N_0}^{[0]} \\ {}^{(2)}x_1^{[0]} & {}^{(2)}x_2^{[0]} & \dots & {}^{(2)}x_{N_0}^{[0]} \\ & & \dots & \\ {}^{(M)}x_1^{[0]} & {}^{(M)}x_2^{[0]} & \dots & {}^{(M)}x_{N_0}^{[0]} \end{bmatrix} \cdot \begin{bmatrix} w_{11}^{[1]} & w_{21}^{[1]} & \dots & w_{N_1 1}^{[1]} \\ w_{12}^{[1]} & w_{22}^{[1]} & \dots & w_{N_1 2}^{[1]} \\ & & \dots & \\ w_{1N_0}^{[1]} & w_{2N_0}^{[1]} & \dots & w_{N_1 N_0}^{[1]} \end{bmatrix} + \begin{bmatrix} b_1^{[1]} & b_2^{[1]} & \dots & b_{N_1}^{[1]} \\ b_1^{[1]} & b_2^{[1]} & \dots & b_{N_1}^{[1]} \\ & & \dots & \\ b_1^{[1]} & b_2^{[1]} & \dots & b_{N_1}^{[1]} \end{bmatrix}\right)$$

$$\hat{\mathbf{y}} = \sigma\left(\mathbf{X}^{[1]} \mathbf{w}^{[\mathbf{2}]\mathrm{T}} + b^{[2]}\right) = \sigma\left(\mathbf{Z}^{[2]}\right)$$

$$\hat{\mathbf{y}} = \sigma\left(\begin{bmatrix} {}^{(1)}x_1^{[0]} & {}^{(1)}x_2^{[1]} & \dots & {}^{(1)}x_{N_1}^{[1]} \\ {}^{(2)}x_1^{[1]} & {}^{(2)}x_2^{[1]} & \dots & {}^{(2)}x_{N_1}^{[1]} \\ & & \dots & \\ {}^{(M)}x_1^{[1]} & {}^{(M)}x_2^{[1]} & \dots & {}^{(M)}x_{N_1}^{[1]} \end{bmatrix} \cdot \begin{bmatrix} w_1^{[2]} \\ w_2^{[2]} \\ \dots \\ w_{N_1}^{[2]} \end{bmatrix} + \begin{bmatrix} b^{[2]} \\ b^{[2]} \\ \dots \\ b^{[2]} \end{bmatrix}\right)$$

$$J\left(\mathbf{W}^{[\mathbf{1}]}, \mathbf{b}^{[\mathbf{1}]}, \mathbf{w}^{[\mathbf{2}]}, \mathrm{b}^{[2]}\right) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$$

# Target and Prediction

$$\mathbf{y} = \begin{bmatrix} {}^{(1)}y \\ {}^{(2)}y \\ \cdot \\ \cdot \\ \cdot \\ {}^{(M)}y \end{bmatrix}_{(M,1)}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} {}^{(1)}\hat{y} \\ {}^{(2)}\hat{y} \\ \cdot \\ \cdot \\ \cdot \\ {}^{(M)}\hat{y} \end{bmatrix}_{(M,1)}$$

# Binary Cross Entropy Loss

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{M} \Sigma(-\mathbf{y} \log_e(\hat{\mathbf{y}}) - (1 - \mathbf{y}) \log_e(1 - \hat{\mathbf{y}}))$$

# Back-Propagation

$$\nabla \mathcal{L} = \left[ \frac{\partial \mathcal{L}}{\partial {w_{11}}^{[1]}} \quad \cdots \quad \frac{\partial \mathcal{L}}{\partial {w_{\mathrm{N_1 N_0}}}^{[1]}} \quad \frac{\partial \mathcal{L}}{\partial {b_1}^{[1]}} \quad \cdots \quad \frac{\partial \mathcal{L}}{\partial {b_{\mathrm{N_1}}}^{[1]}} \quad \frac{\partial \mathcal{L}}{\partial {w_1}^{[2]}} \quad \cdots \quad \frac{\partial \mathcal{L}}{\partial {w_{\mathrm{N_1}}}^{[2]}} \quad \frac{\partial \mathcal{L}}{\partial b^{[2]}} \right]$$

$$= \left[ \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} \quad \frac{\partial \mathcal{L}}{\partial \mathbf{w}^{[2]}} \quad \frac{\partial \mathcal{L}}{\partial \mathrm{b}^{[2]}} \right]$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{Z}^{[2]}} \cdot \frac{\partial \mathbf{Z}^{[2]}}{\partial \mathbf{w}^{[2]}} = \frac{1}{M} (\hat{\mathbf{y}} - \mathbf{y})^T \cdot \mathbf{X}^{[1]}$$

$$\frac{\partial \mathcal{L}}{\partial b^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{Z}^{[2]}} \cdot \frac{\partial \mathbf{Z}^{[2]}}{\partial b^{[2]}} = \frac{1}{M} \sum (\hat{\mathbf{y}} - \mathbf{y})$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial \mathbf{X}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{Z}^{[2]}} \cdot \frac{\partial \mathbf{Z}^{[2]}}{\partial \mathbf{X}^{[1]}} = (\hat{\mathbf{y}} - \mathbf{y}) \cdot \mathbf{w}^{[2]}}$$

# Back-Propagation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{X}^{[1]}} \cdot \frac{\partial \mathbf{X}^{[1]}}{\partial \mathbf{Z}^{[1]}} \cdot \frac{\partial \mathbf{Z}^{[1]}}{\partial \mathbf{W}^{[1]}} = \frac{1}{M}\left[(\hat{\mathbf{y}} - \mathbf{y}) \cdot \mathbf{w}^{[2]} * \mathbf{X}^{[1]} * \left(1 - \mathbf{X}^{[1]}\right)\right]^T \cdot \mathbf{X}^{[0]}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{X}^{[1]}} \cdot \frac{\partial \mathbf{X}^{[1]}}{\partial \mathbf{Z}^{[1]}} \cdot \frac{\partial \mathbf{Z}^{[1]}}{\partial \mathbf{b}^{[1]}} = \frac{1}{M}\Sigma\left[(\hat{\mathbf{y}} - \mathbf{y}) \cdot \mathbf{w}^{[2]} * \mathbf{X}^{[1]} * \left(1 - \mathbf{X}^{[1]}\right)\right]^T, \text{ axis=0}$$

# Gradient Descent Optimization

Given dataset: $\left\{ \left( {}^{(1)}\mathbf{x}, {}^{(1)}y \right), \left( {}^{(2)}\mathbf{x}, {}^{(2)}y \right), \ldots, \left( {}^{(M)}\mathbf{x}, {}^{(M)}y \right) \right\}$

Initialize $\mathbf{W}^{[1]}$, $\mathbf{w}^{[\mathbf{2}]}$, $\mathbf{b}^{[1]}\, and\, b^{[2]}$

Repeat until converge {

$$\mathbf{W}^{[1]} := \mathbf{W}^{[1]} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}}$$

$$\mathbf{w}^{[2]} := \mathbf{w}^{[2]} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{w}^{[2]}}$$

$$\mathbf{b}^{[1]} := \mathbf{b}^{[1]} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}}$$

$$b^{[2]} := b^{[2]} - \alpha \frac{\partial \mathcal{L}}{\partial b^{[2]}}$$

}

where $\alpha$ is learning rate