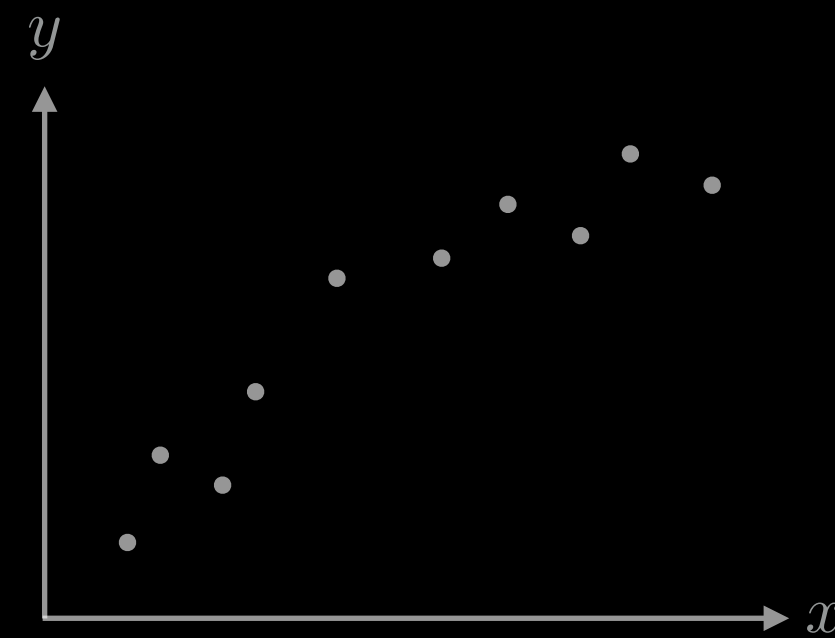


Deep Learning

The ML workflow

Training Set Validation Set Test Set



How do I choose the feature vector?

$$\phi(x) = [1, x, \dots, ?]$$

not satisfied

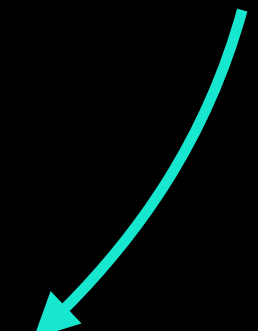
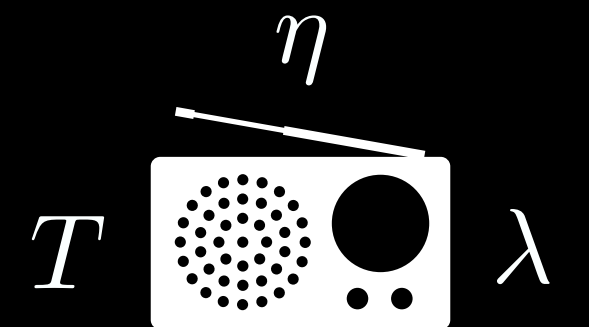
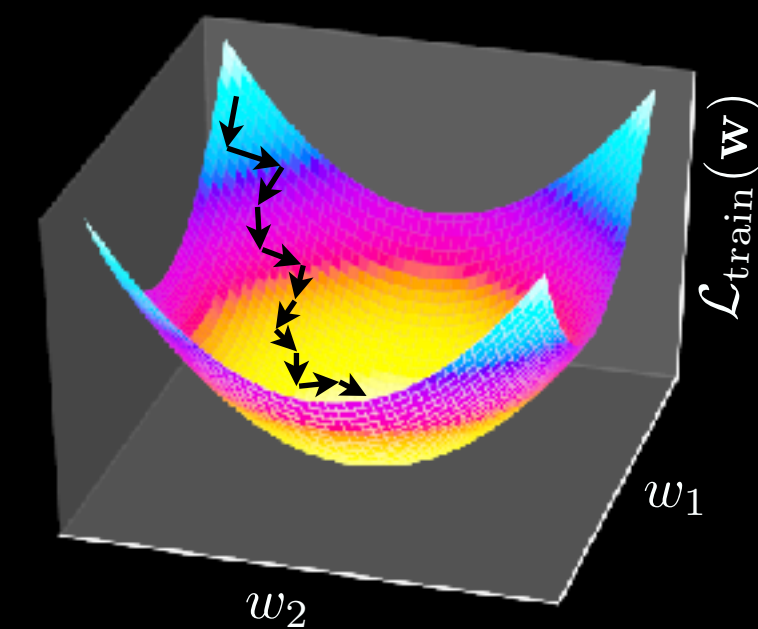
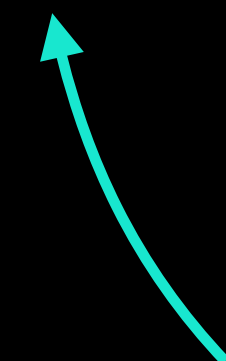


$\mathcal{L}_{\text{test}}$



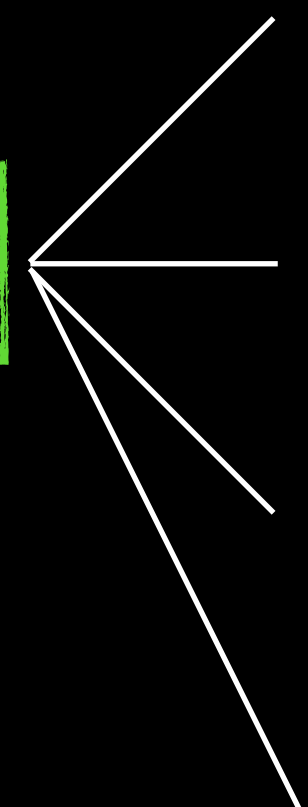
satisfied

$\mathcal{L}_{\text{val}}$



How do I choose the feature vector?

$$\phi(x) = [1, x, \dots, ?]$$

$$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \underbrace{\phi(x)}_{\mathbf{x}}$$


- $\phi(x) = [1, x]$
- $\phi(x) = [1, x, x^2, x^3]$
- $\phi(x) = [1, x, \sin(3x)]$
- ????????????????

How do I choose the feature vector?

$$\phi(x) = [1, x, \dots, ?]$$



Decision Boundary

$$\phi(x) \cdot \mathbf{w} = 0$$



Boat

Linear Predictor

$$f_{\mathbf{w}}(x) = \mathbf{x} \cdot \mathbf{w}$$

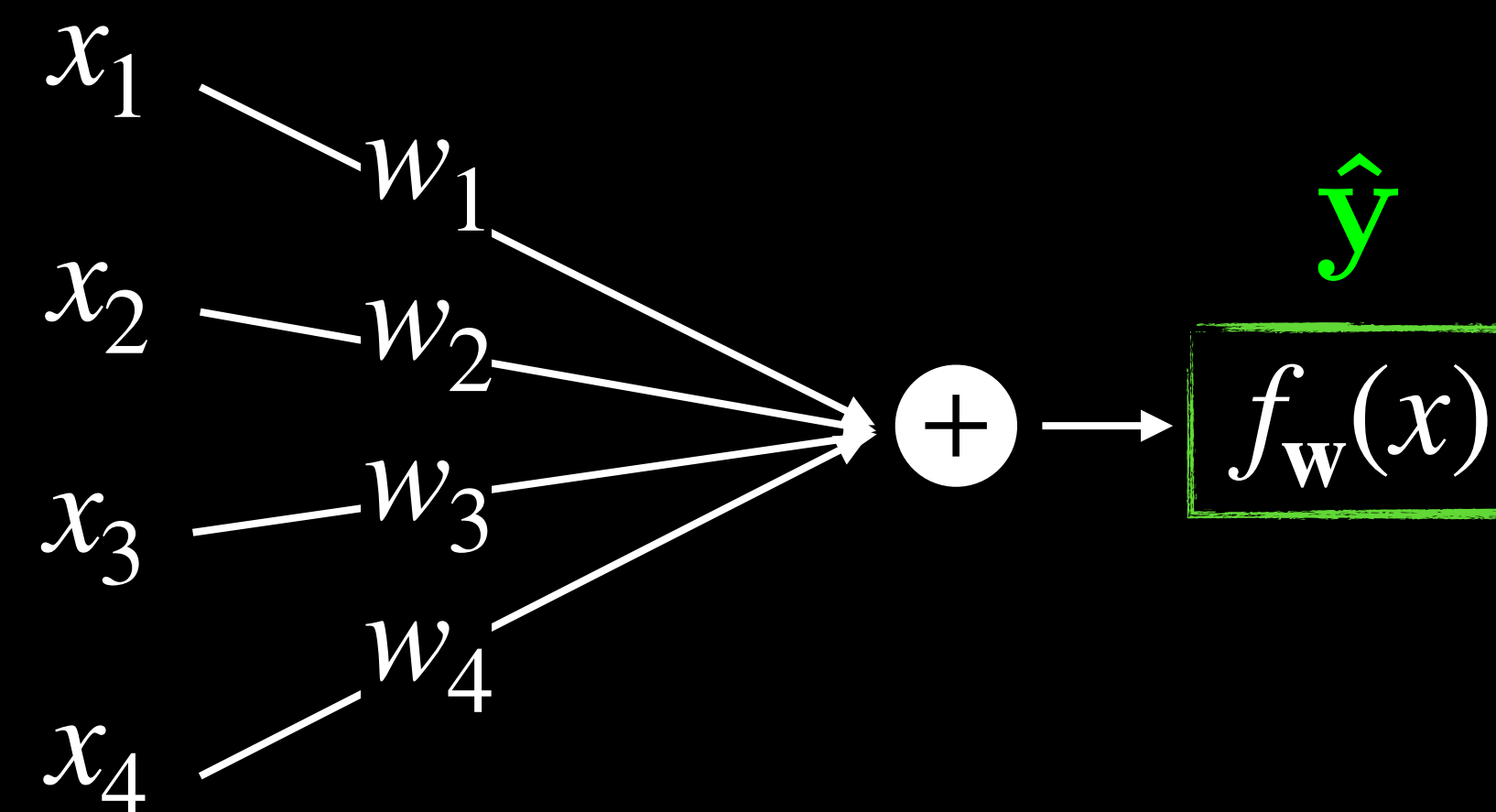
$$\mathbf{w} = [w_1, w_2, w_3, w_4]$$

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

$$f_{\mathbf{w}}(x) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$$

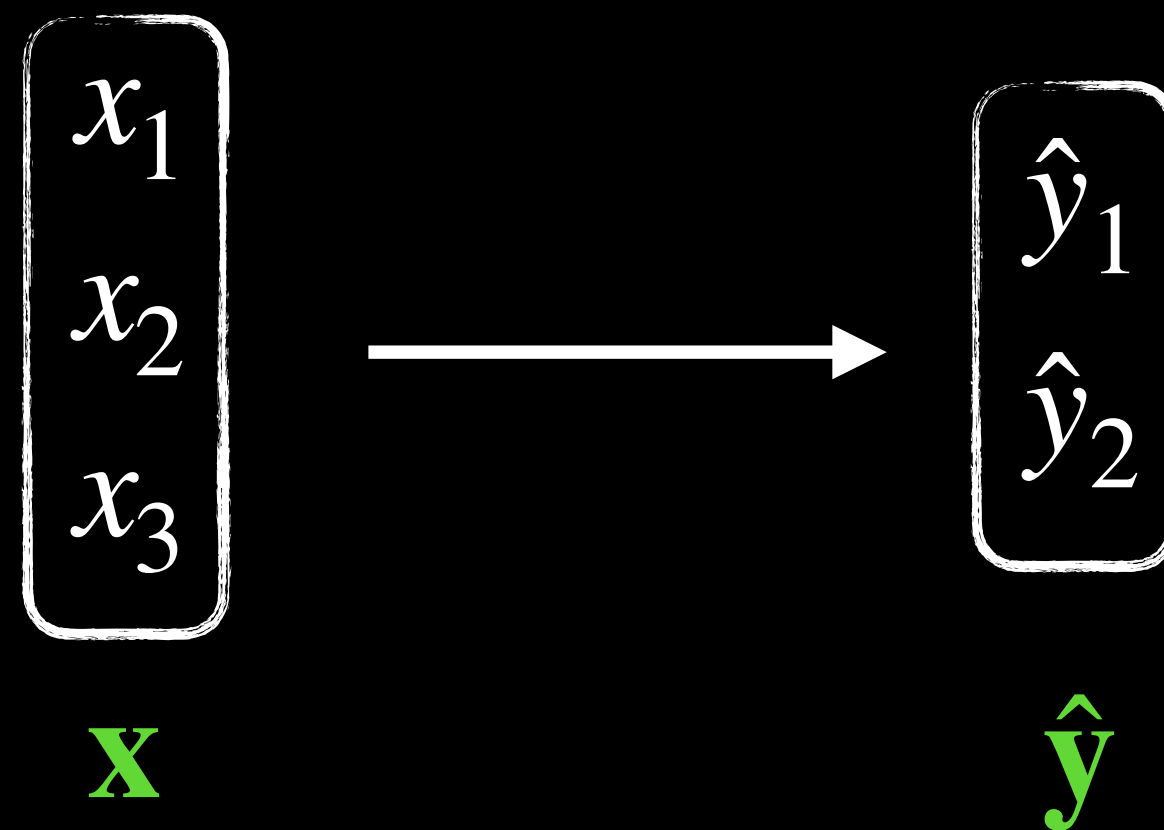


Network Representation



Linear Predictor

2 outputs?



3 * 2 fitting parameters

$$\hat{y}_1 = \mathbf{w}_1 \cdot \mathbf{x} = w_{11}x_1 + w_{12}x_2 + w_{13}x_3$$

$$\hat{y}_2 = \mathbf{w}_2 \cdot \mathbf{x} = w_{21}x_1 + w_{22}x_2 + w_{23}x_3$$

Matrix form

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

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

From Matrix to Network

Matrix Representation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

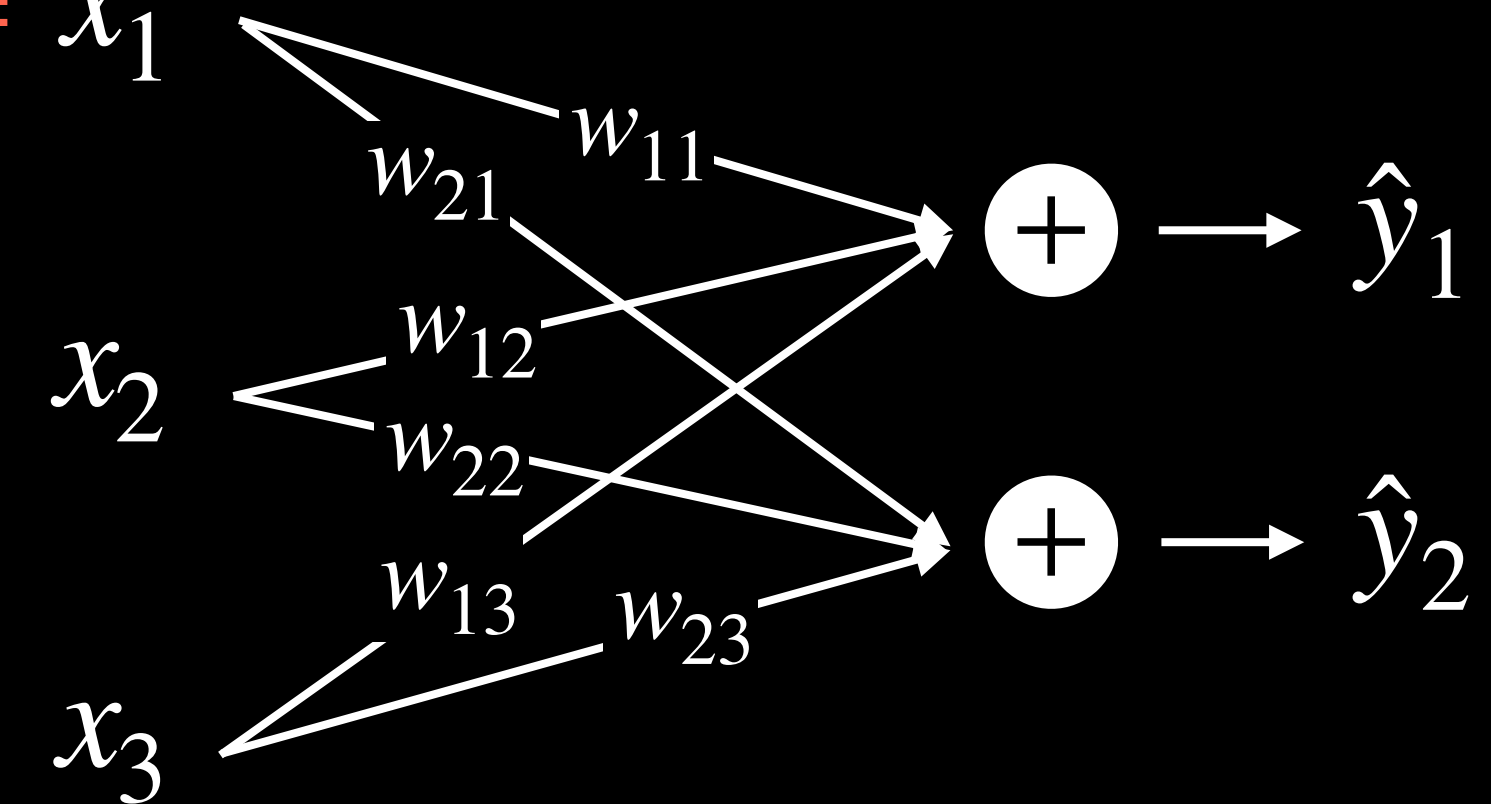
$\hat{\mathbf{y}} = \mathbf{W} \mathbf{x}$

Index notation

$$\hat{y}_i = \sum_{j=1}^n w_{ij} x_j$$

Network Representation

Account for Bias: $1 = x_1$



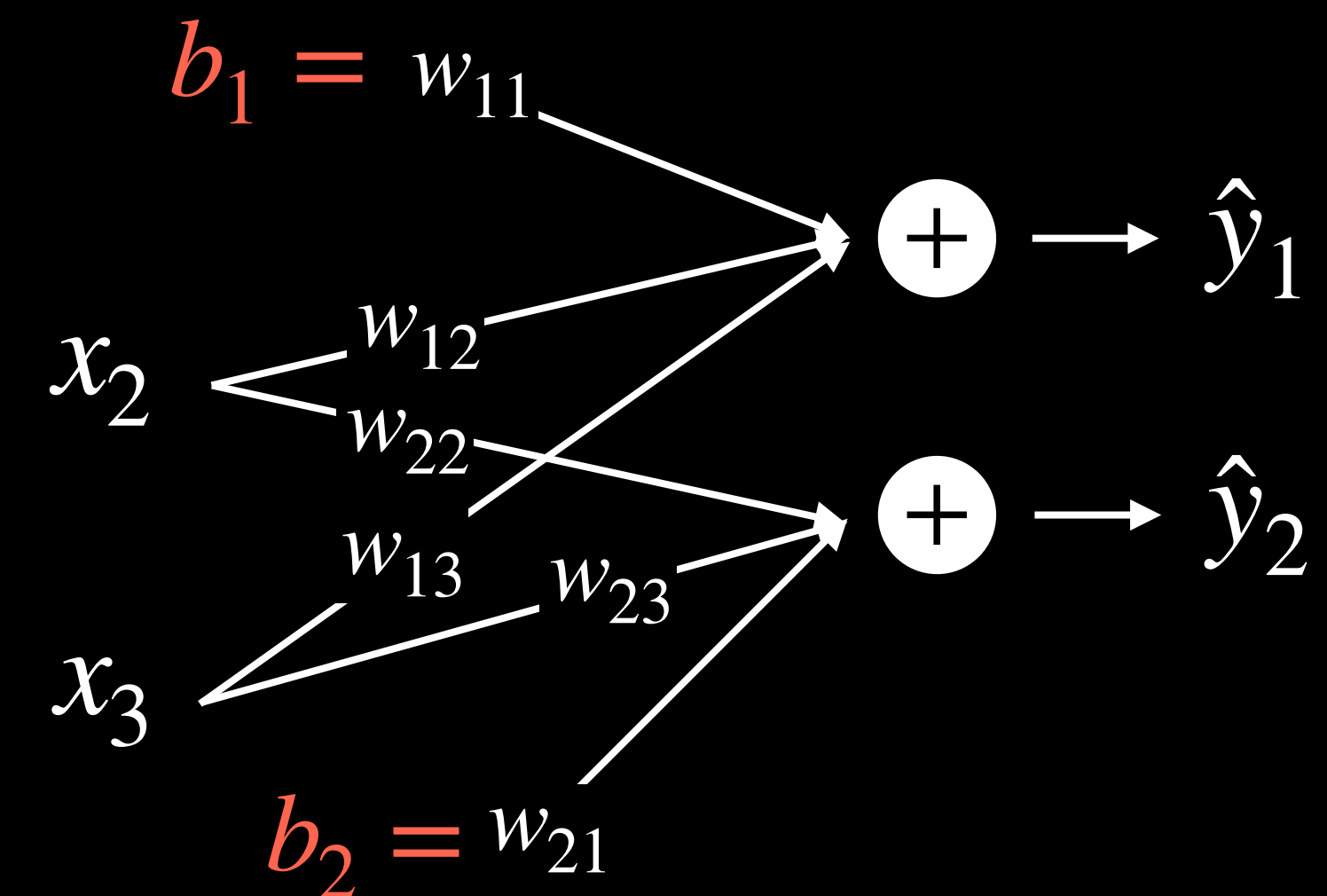
Linear Predictor - Explicit Bias

Matrix Representation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} w_{12} & w_{13} \\ w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} w_{11} \\ w_{21} \end{bmatrix}$$

$$\hat{\mathbf{y}} = \mathbf{W} \mathbf{x} + \mathbf{b}$$

Network Representation



Linear Predictor

$$\hat{\mathbf{y}} = \mathbf{W} \mathbf{x} + \mathbf{b}$$

$m \times 1$
Number of outputs

$m \times n$

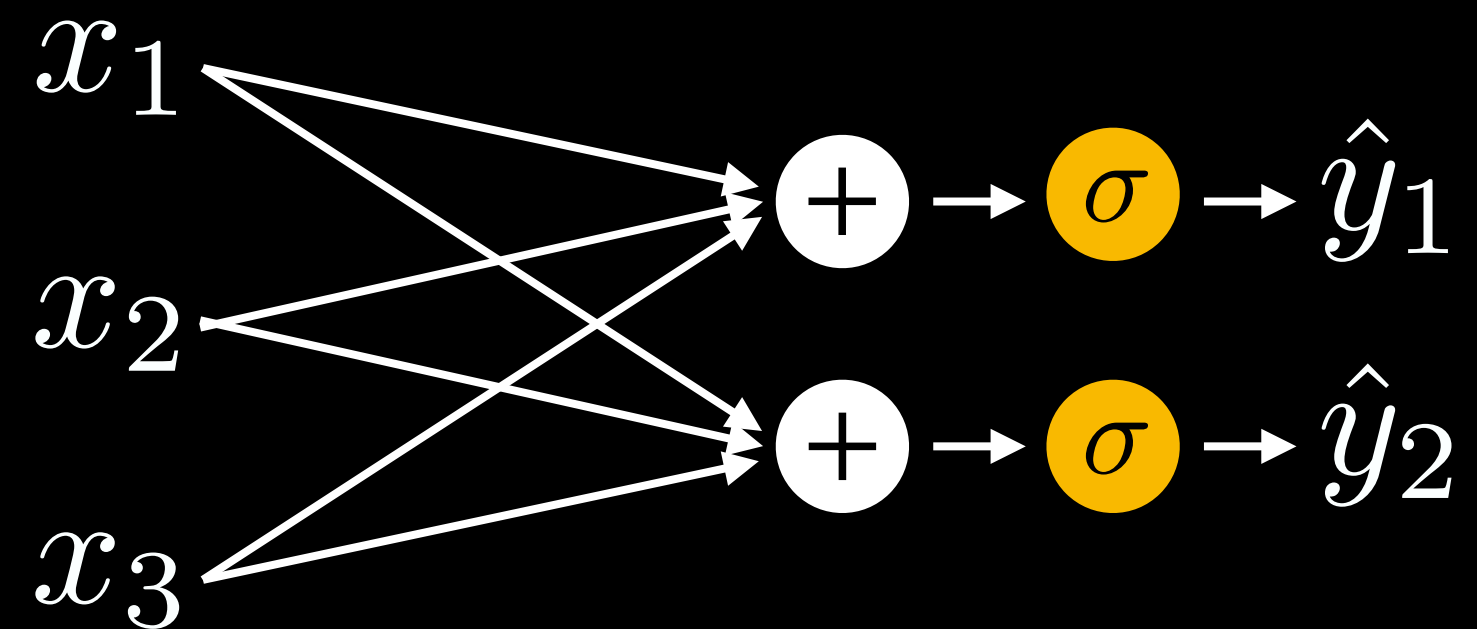
$n \times 1$
Number of inputs

$m \times 1$
Some formulations explicitly account for $\mathbf{b}$, while others include the bias as part of $\mathbf{x}$

Here we omit $\mathbf{b}$ for simplicity of representation

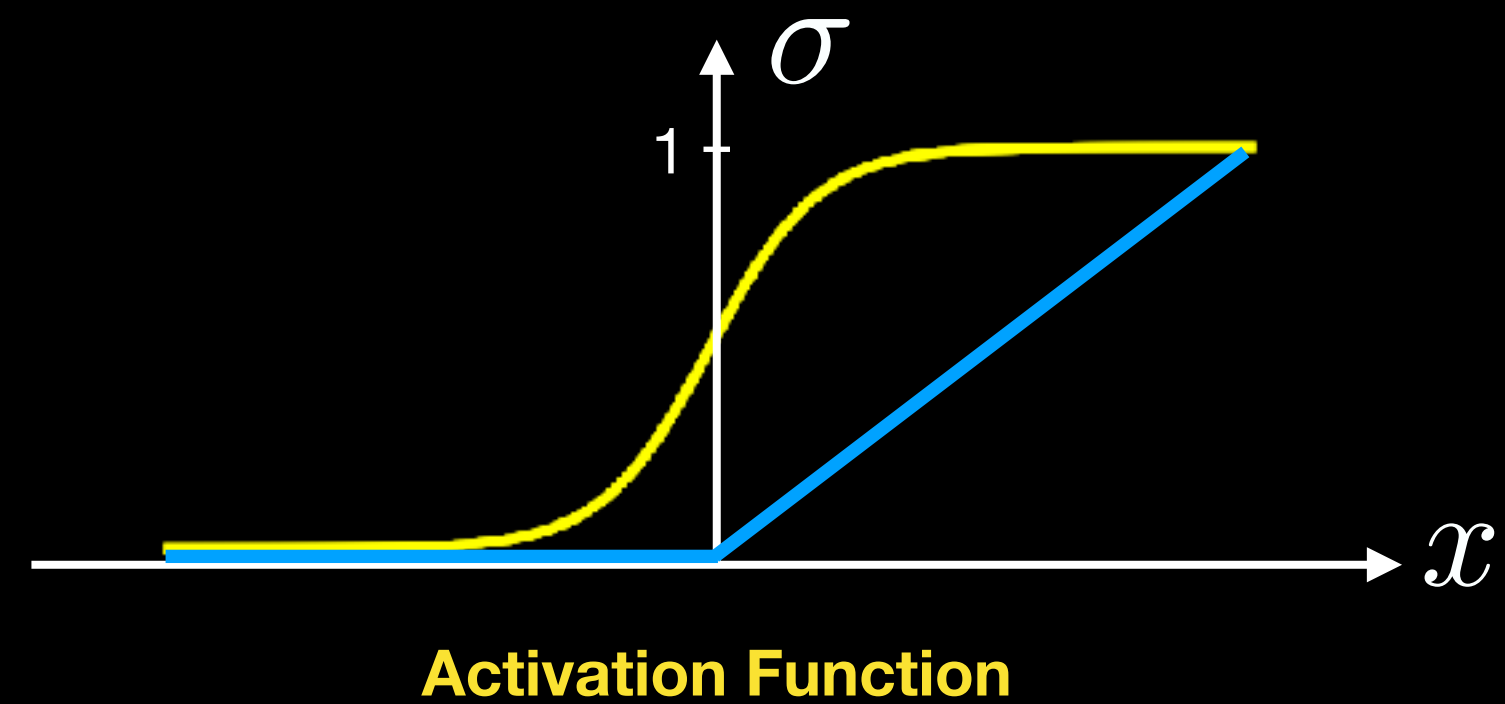
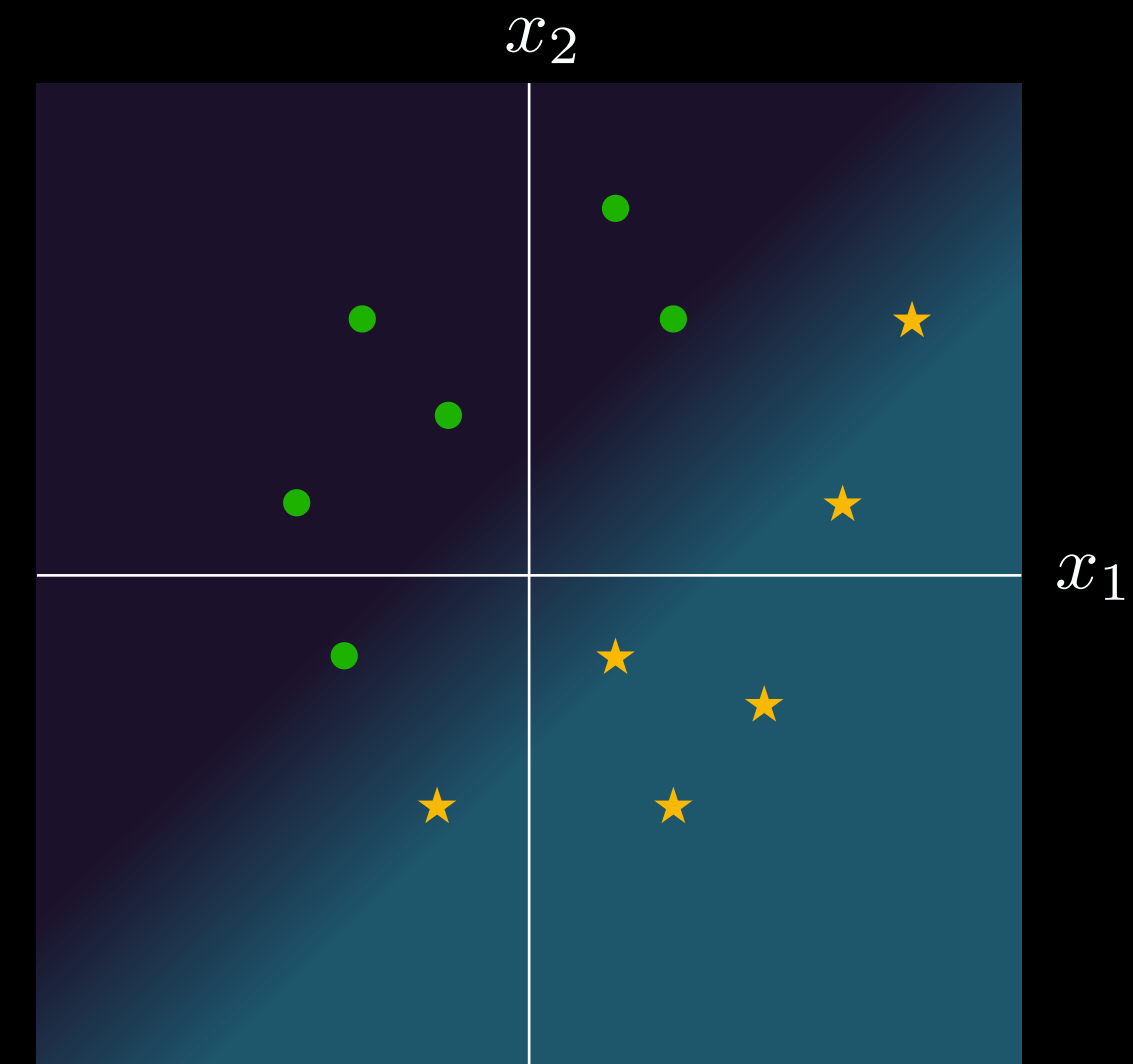
Nonlinear Predictor

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}\mathbf{x})$$



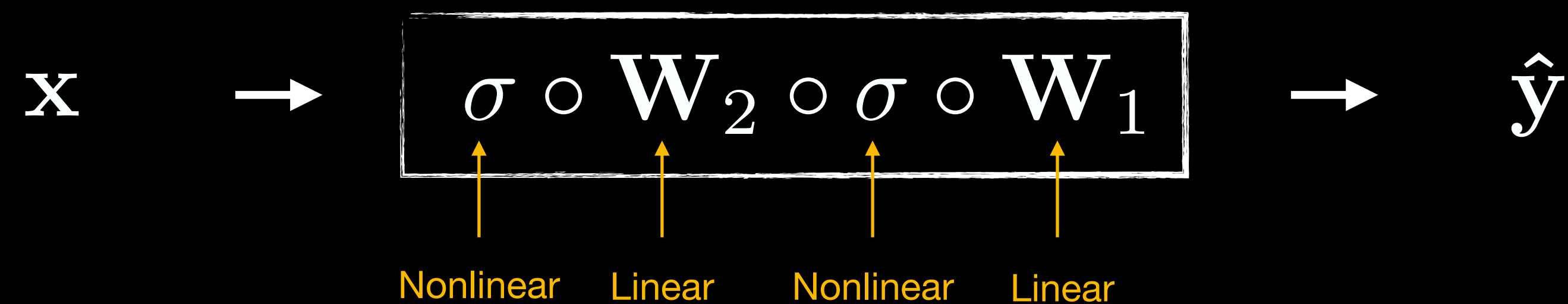
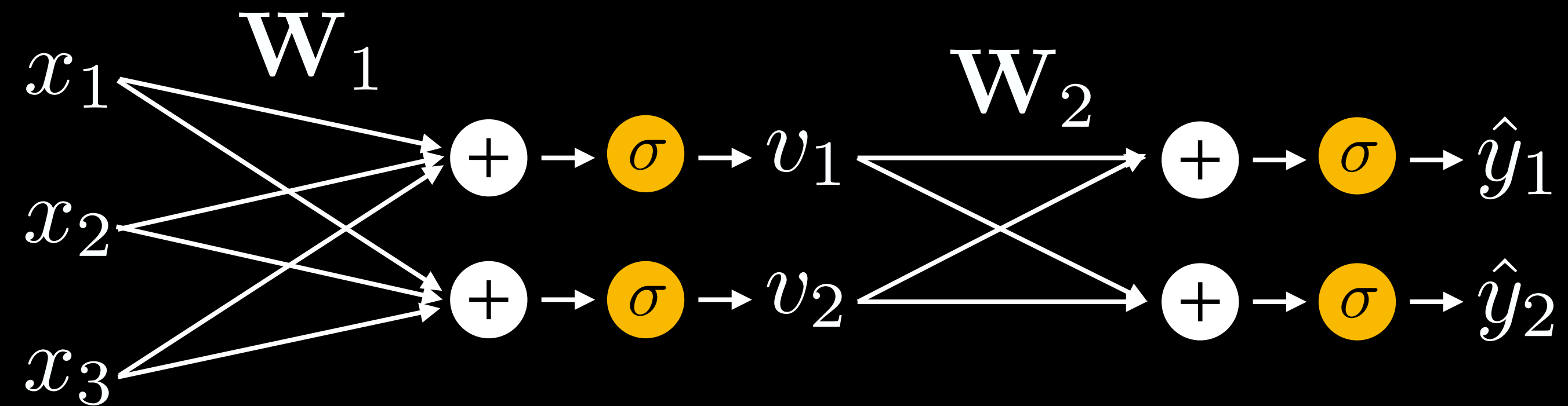
Logistic function: $\sigma(x) = \frac{1}{1 + e^{-x}}$

ReLU: $\sigma(x) = xH(x)$



Neural network

$$\hat{y} = \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))$$

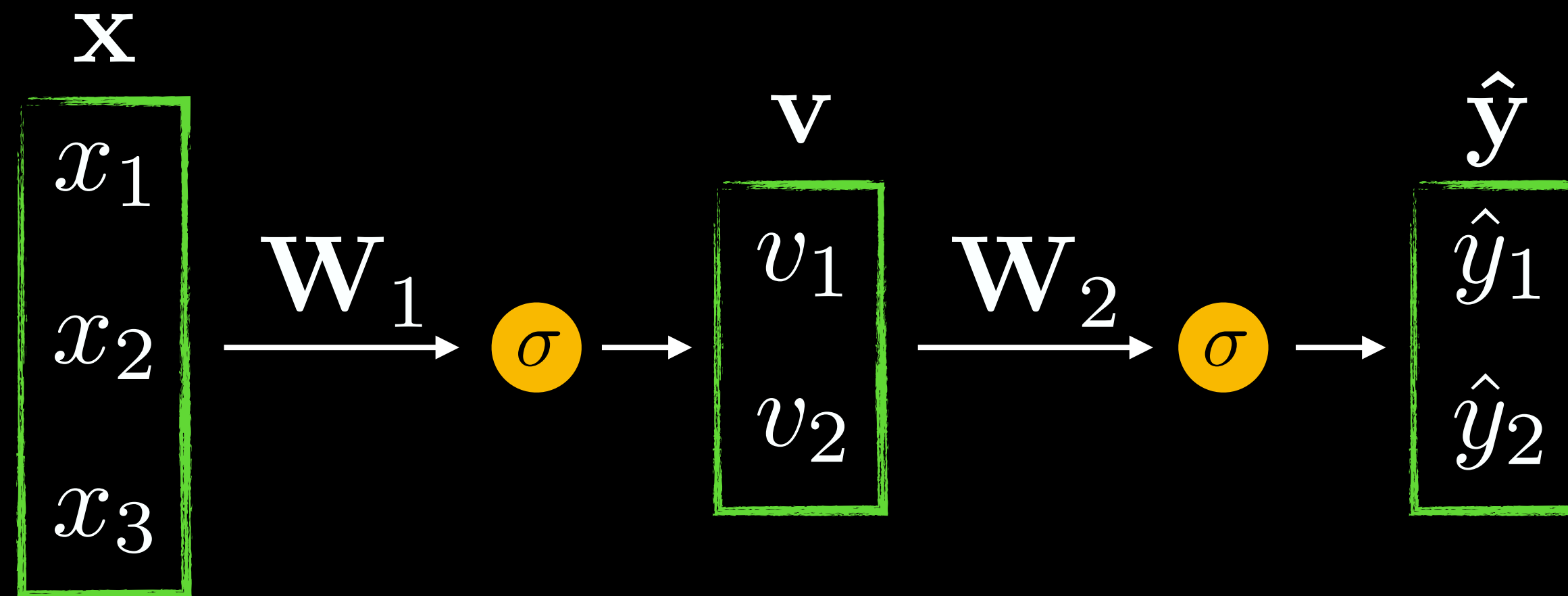


Neural network

$$\hat{y} = \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))$$

$$\mathbf{v} = \sigma(\mathbf{W}_1 \mathbf{x})$$

$$\hat{y} = \sigma(\mathbf{W}_2 \mathbf{v})$$

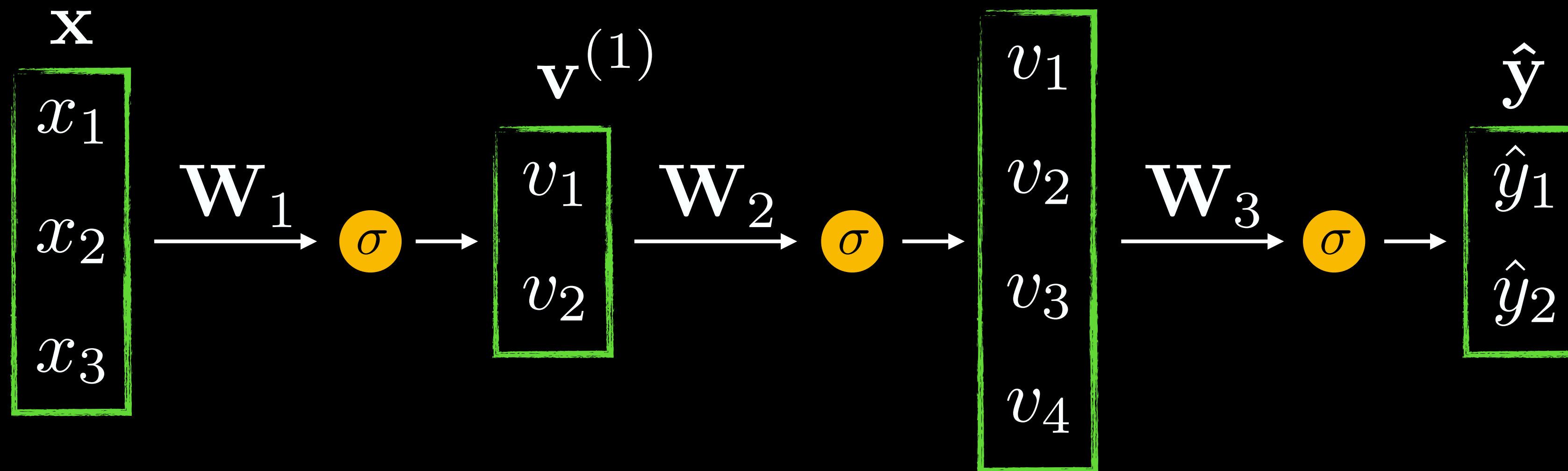


Hidden layer

Can be interpreted
as a learned $\phi(\mathbf{x})$

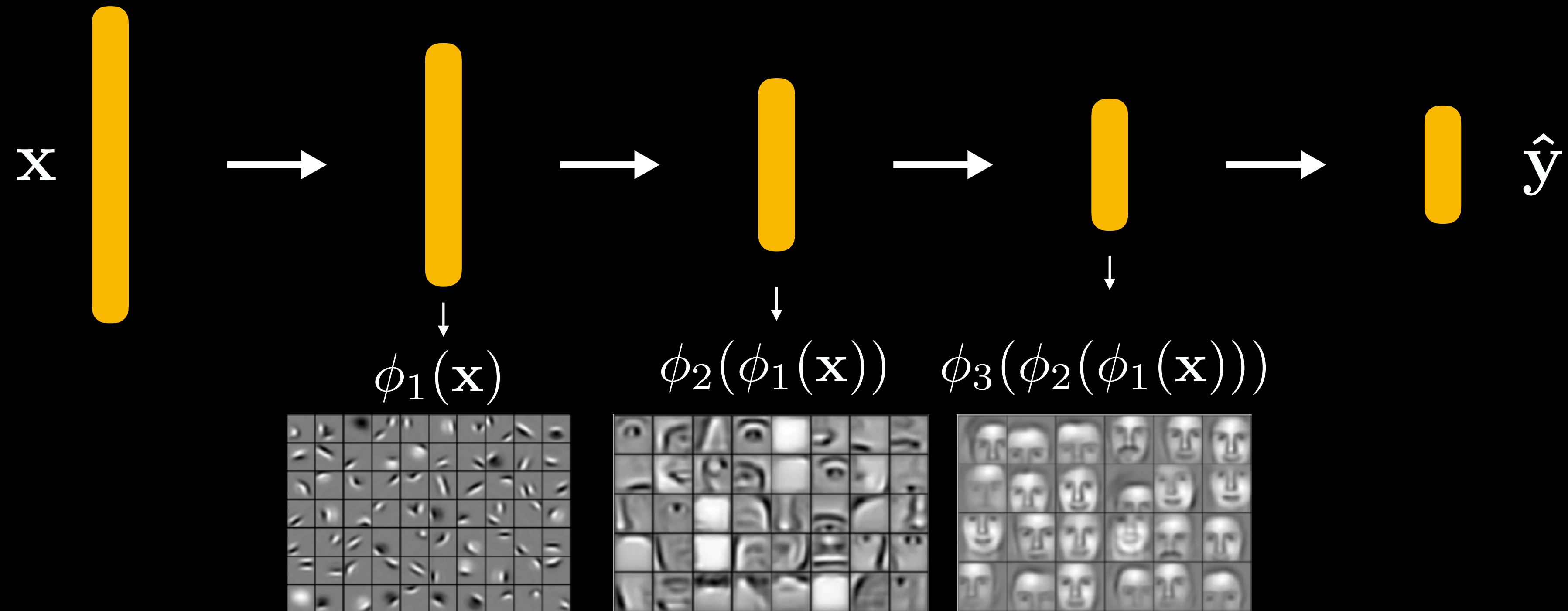
Deep network

$$\hat{y} = \sigma \left(\mathbf{W}_3 \sigma \left(\mathbf{W}_2 \sigma \left(\mathbf{W}_1 \mathbf{x} \right) \right) \right)$$



$$\mathbf{v}^{(1)} = \sigma \left(\mathbf{W}_1 \mathbf{x} \right) \quad \mathbf{v}^{(2)} = \sigma \left(\mathbf{W}_2 \mathbf{v}^{(1)} \right) \quad \hat{y} = \sigma \left(\mathbf{W}_3 \mathbf{v}^{(2)} \right)$$

Why *deep* learning?



Feature learning

Loss function

$$f_{\mathbf{W}_1 \mathbf{W}_2}(\mathbf{x}) = \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x})$$

$$\mathcal{L}(\mathbf{x}, y, \mathbf{W}_1, \mathbf{W}_2) = \left\| f_{\mathbf{W}_1 \mathbf{W}_2}(\mathbf{x}) - y \right\|^2$$

Stochastic gradient descent update

$$\mathbf{W}_1 \leftarrow \mathbf{W}_1 - \alpha \nabla_{\mathbf{W}_1} \mathcal{L}(\mathbf{x}, y, \mathbf{W}_1, \mathbf{W}_2)$$

$$\mathbf{W}_2 \leftarrow \mathbf{W}_2 - \alpha \nabla_{\mathbf{W}_2} \mathcal{L}(\mathbf{x}, y, \mathbf{W}_1, \mathbf{W}_2)$$

How do we calculate the gradients?

Approach

Training loss

$$\mathcal{L}(\mathbf{W}_1, \mathbf{W}_2) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{W}_1, \mathbf{W}_2)$$

Objective

$$\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2 = \arg \min_{\mathbf{W}_1, \mathbf{W}_2} \mathcal{L}(\mathbf{W}_1, \mathbf{W}_2)$$

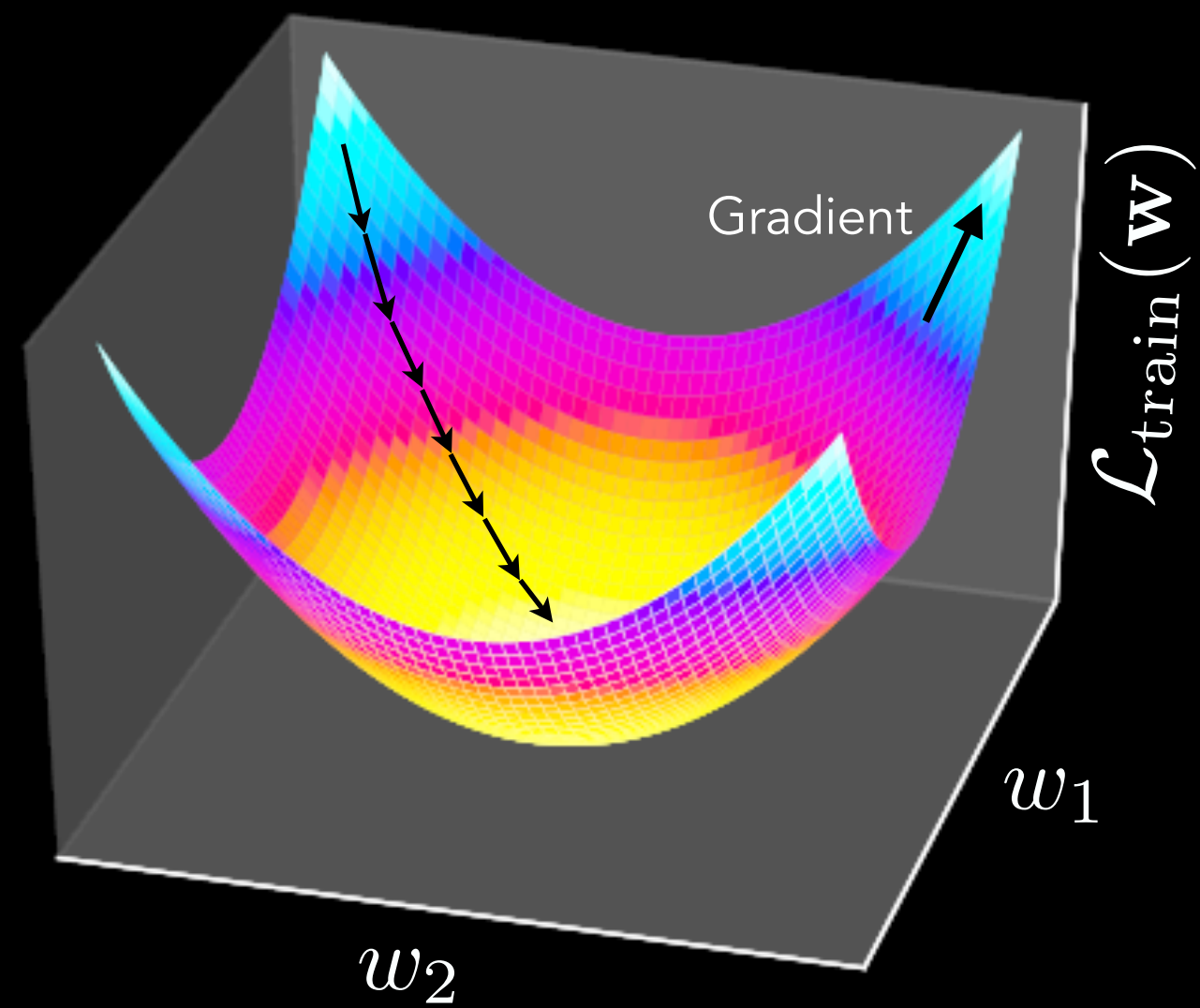
Optimal predictor

$$f_{\hat{\mathbf{W}}_1 \hat{\mathbf{W}}_2}(\mathbf{x}) = \hat{\mathbf{W}}_2 \sigma(\hat{\mathbf{W}}_1 \mathbf{x})$$

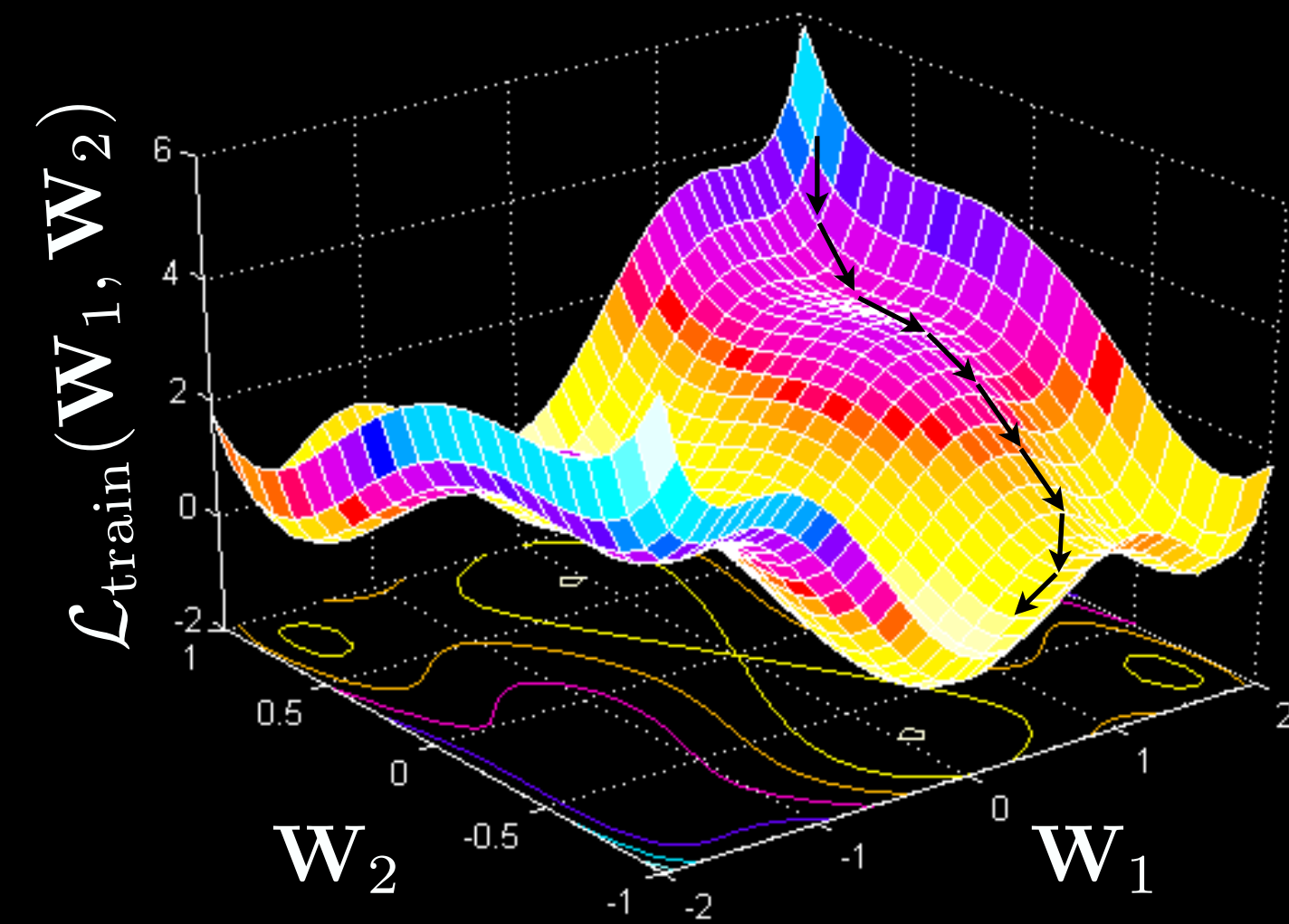
Non-convexity

$$\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2 = \arg \min_{\mathbf{W}_1, \mathbf{W}_2} \mathcal{L}(\mathbf{W}_1, \mathbf{W}_2)$$

Linear predictor loss



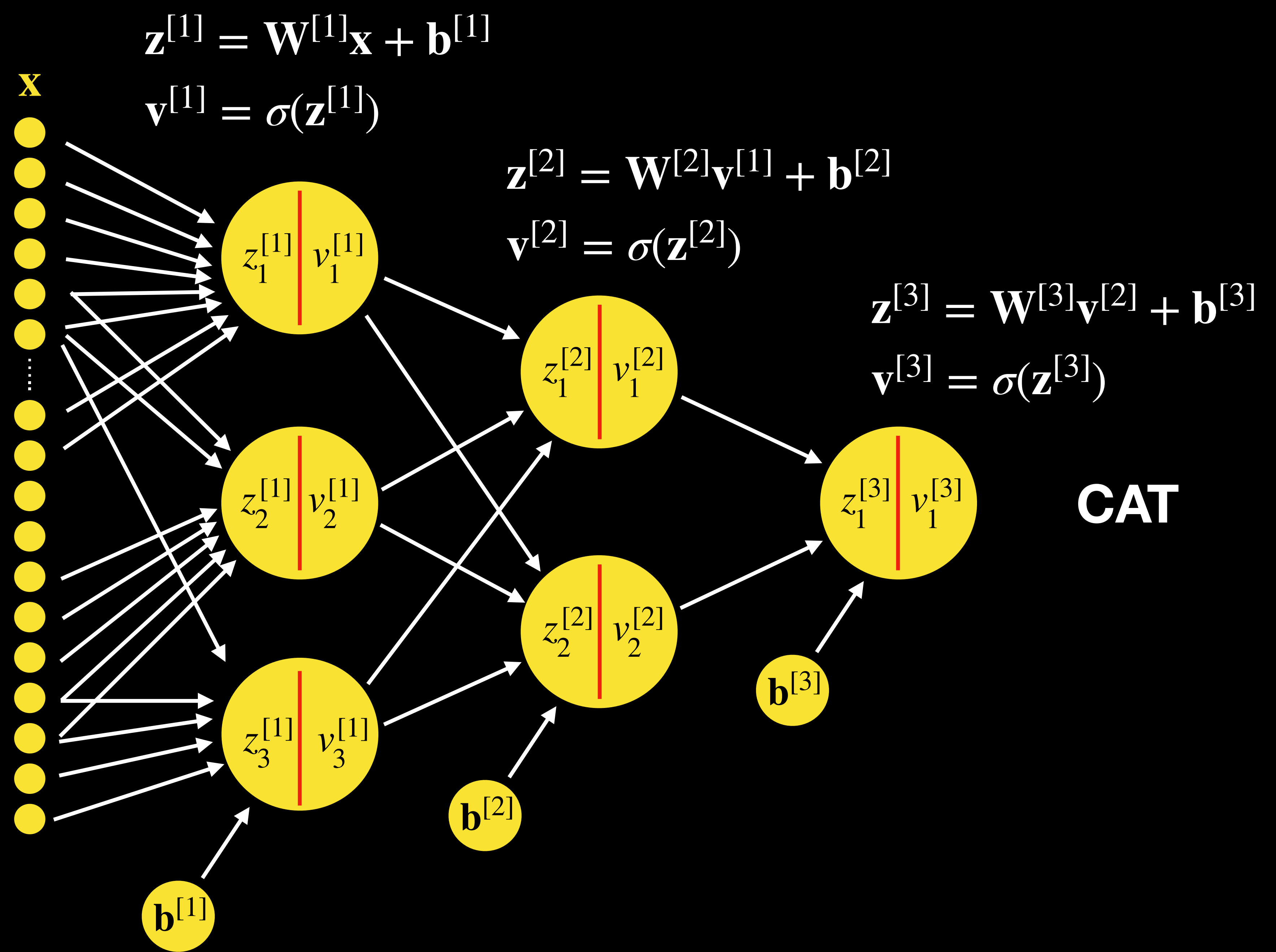
Neural network loss







Flatten



Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

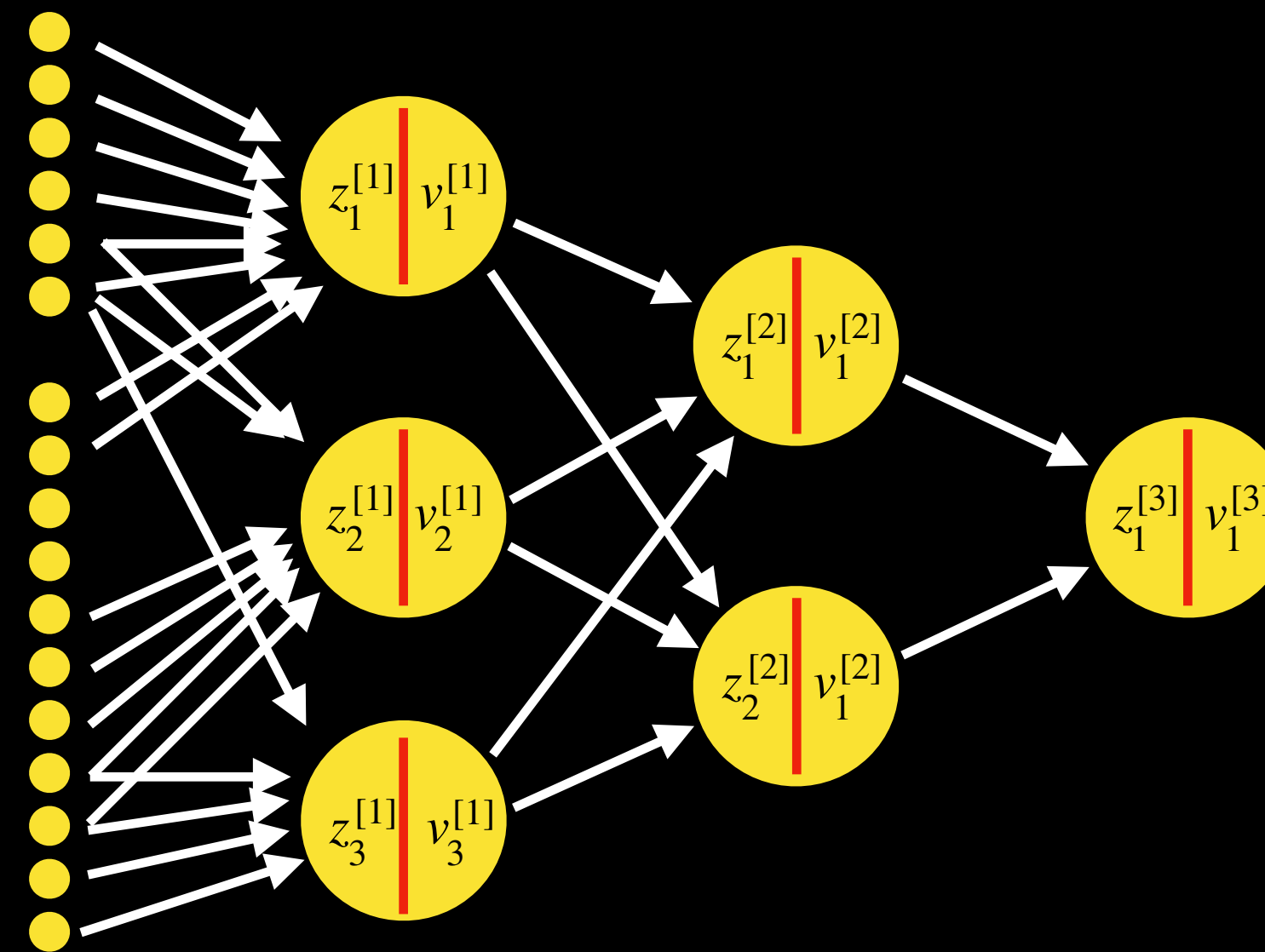
$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\mathbf{v}^{[3]} = \sigma(\mathbf{z}^{[3]})$$



$$\hat{y} \equiv \mathbf{v}^{[3]}$$

$$\hat{y} = f_{\mathbf{W}^{[1]}\mathbf{W}^{[2]}\mathbf{W}^{[3]}}(\mathbf{x})$$

Loss (Binary output):

$$\mathcal{L}(\hat{y}, y) = - \sum_{i=1}^n y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

Gradient Descent

Update the i^{th} layer:

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

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

What are the gradients?

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

What are the gradients?

$$\mathcal{L}(\hat{y}, y) = - \sum_{i=1}^n y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{v}^{[2]}} \frac{\partial \mathbf{v}^{[2]}}{\partial \mathbf{z}^{[2]}} \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{v}^{[2]}} \frac{\partial \mathbf{v}^{[2]}}{\partial \mathbf{z}^{[2]}} \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}$$

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]} \mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]} \mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]} \mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

What are the gradients?

$$\mathcal{L}(\hat{y}, y) = - \sum_{i=1}^n y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = -y^{(i)} \frac{1}{\hat{y}^{(i)}} + (1 - y^{(i)}) \frac{1}{1 - \hat{y}^{(i)}}$$

$$\frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} = \sigma'(\mathbf{z}^{[3]}) = \sigma(\mathbf{z}^{[3]})(1 - \sigma(\mathbf{z}^{[3]}))$$

$$\frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}} = \mathbf{v}^{[2]\top}$$

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = -y^{(i)} \frac{1}{\hat{y}^{(i)}} + (1 - y^{(i)}) \frac{1}{1 - \hat{y}^{(i)}}$$

$$\frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} = \sigma'(\mathbf{z}^{[3]}) = \sigma(\mathbf{z}^{[3]})(1 - \sigma(\mathbf{z}^{[3]}))$$

$$\frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}} = \mathbf{v}^{[2]\top}$$

⋮

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[3]}} = (y^{(i)} - \hat{y}^{(i)}) \mathbf{v}^{[2]\top}$$

SGD Update



$$\mathbf{W}^{[3]} = \mathbf{W}^{[3]} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[3]}}$$

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]} \mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]} \mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]} \mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}} = (y^{(i)} - \hat{y}^{(i)}) \mathbf{v}^{[2]\top}$$

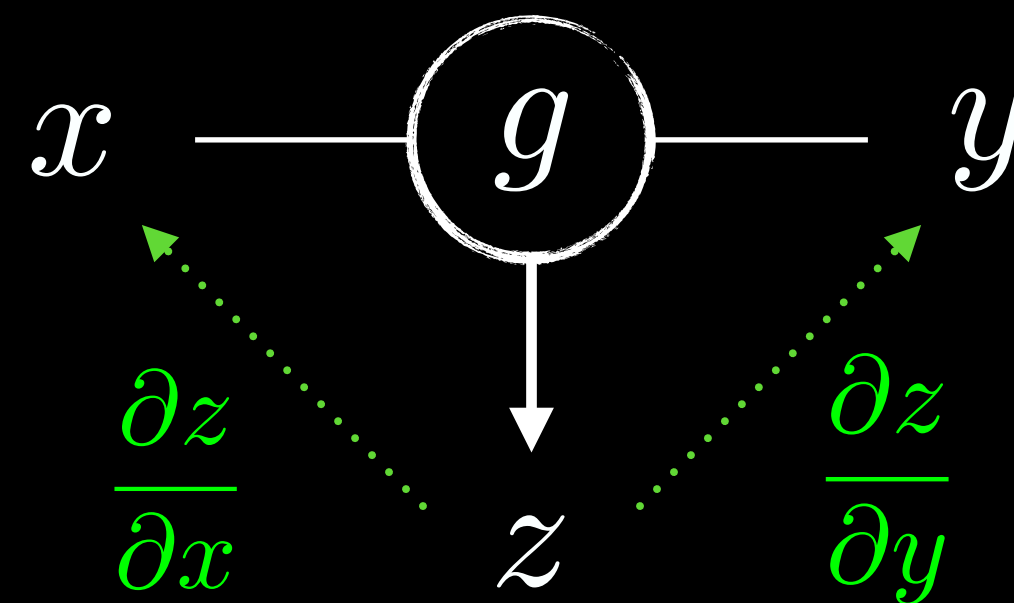
$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{v}^{[2]}} \frac{\partial \mathbf{v}^{[2]}}{\partial \mathbf{z}^{[2]}} \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}$$

common terms across gradients

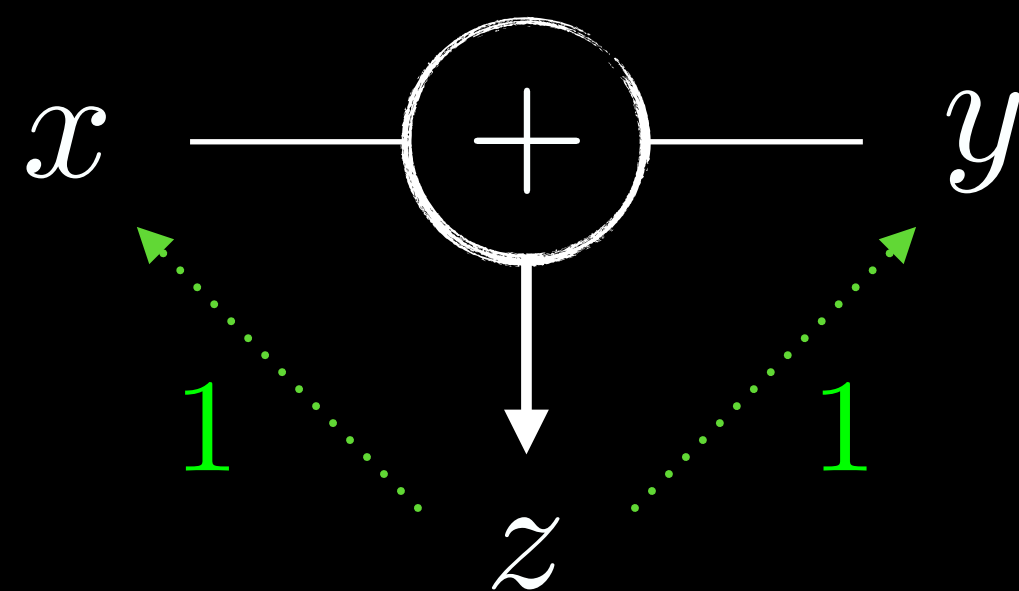
**Can we save the gradients to be reused
for computing other gradients?**

Computation graphs

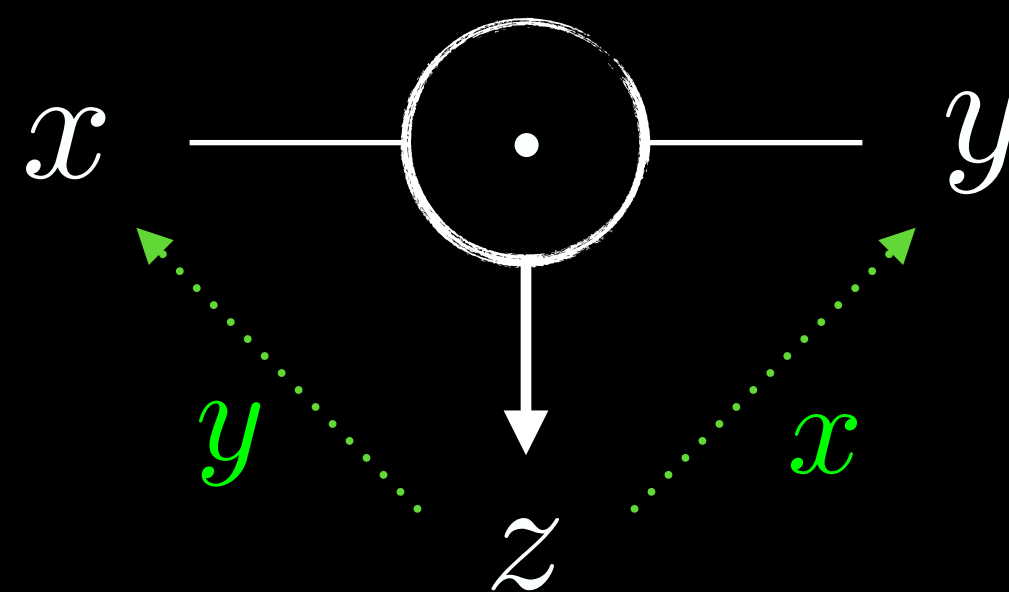
$$z = g(x, y)$$



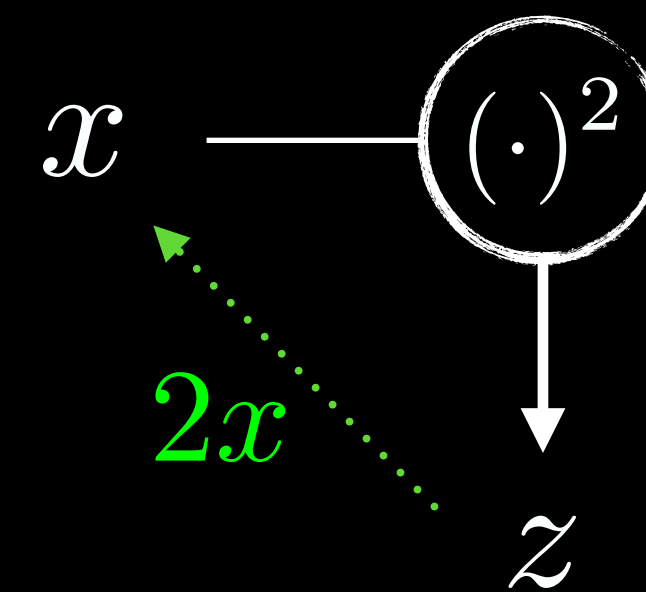
$$z = x + y$$



$$z = xy$$

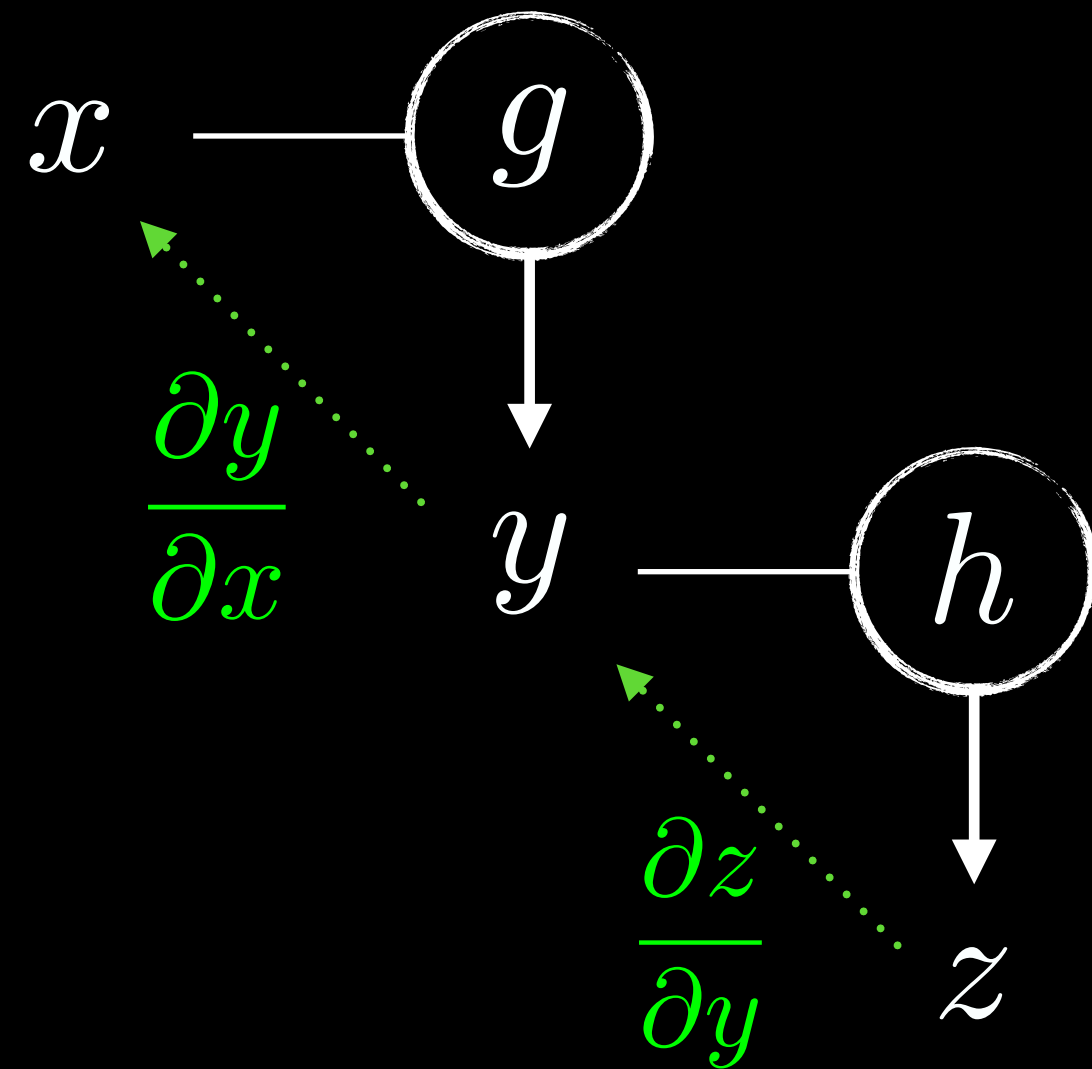


$$z = x^2$$



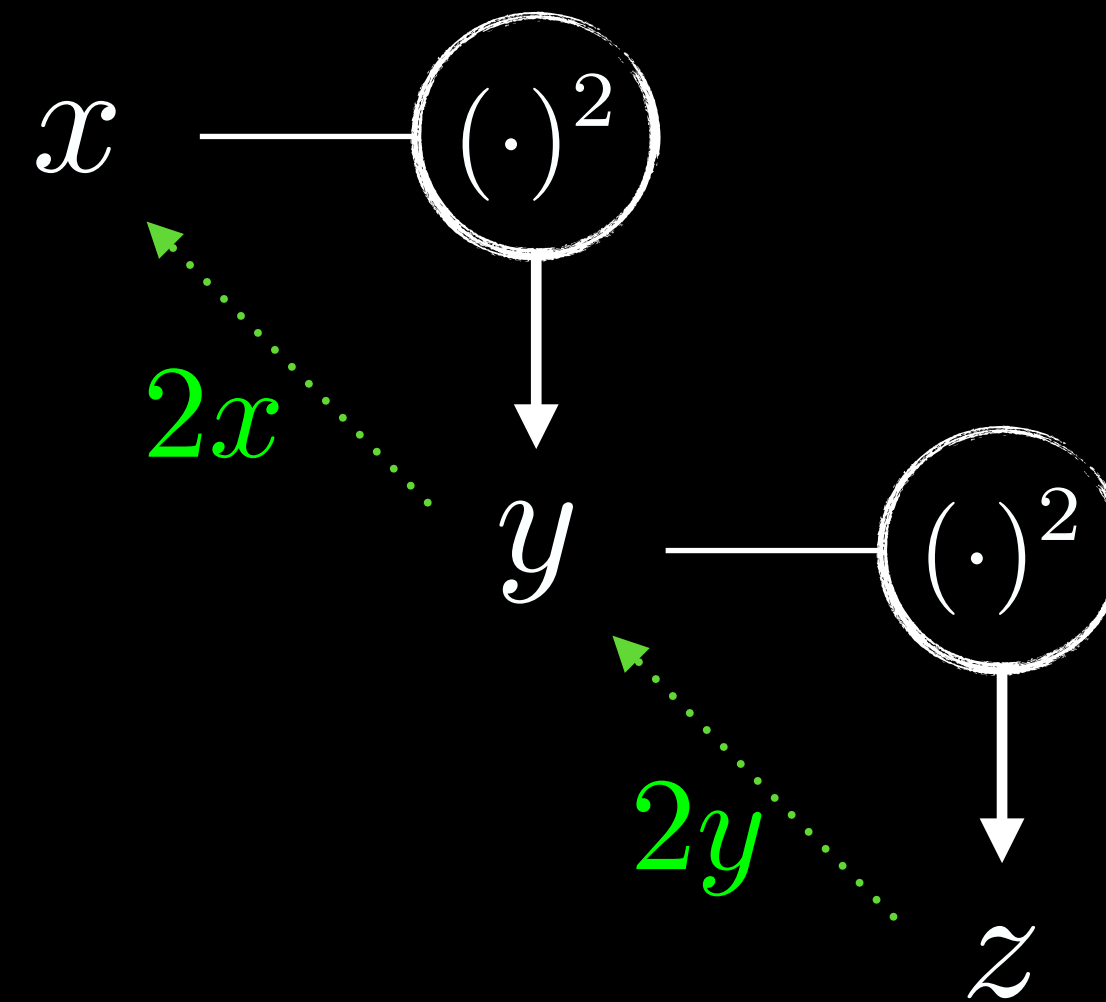
Chain rule

$$y = g(x)$$
$$z = h(y)$$



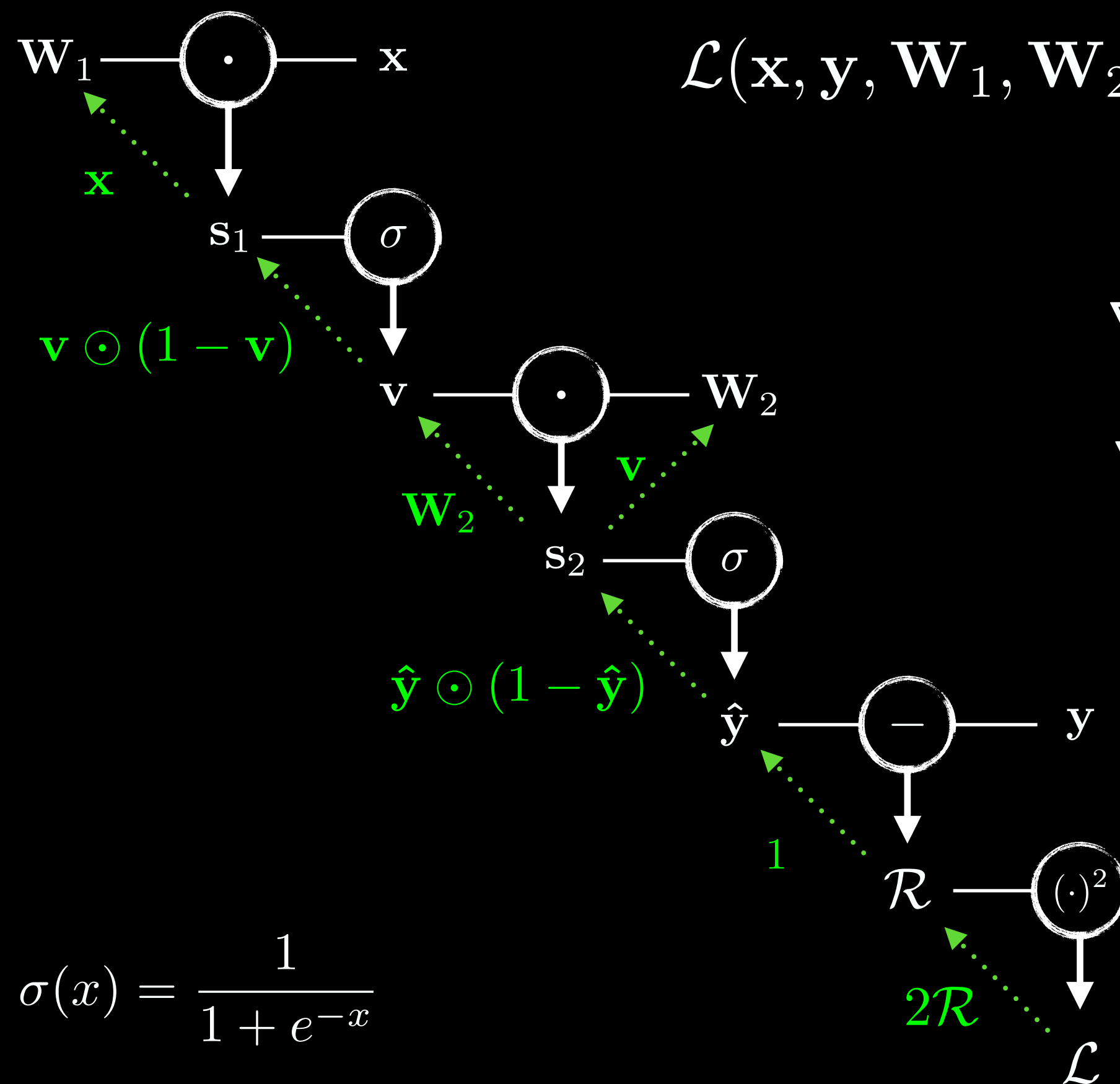
$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$$

$$y = x^2$$
$$z = y^2$$



$$\frac{\partial z}{\partial x} = 4xy = 4x^3$$

Backpropagation



$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \mathbf{W}_1, \mathbf{W}_2) = \|\sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x})) - \mathbf{y}\|^2$$

$$\nabla_{\mathbf{W}_1} \mathcal{L} = 2\mathbf{W}_2^\top \mathcal{R} \odot \hat{\mathbf{y}} \odot (1 - \hat{\mathbf{y}}) \odot \mathbf{v} \odot (1 - \mathbf{v}) \mathbf{x}^\top$$

$$\nabla_{\mathbf{W}_2} \mathcal{L} = 2\mathcal{R} \odot \hat{\mathbf{y}} \odot (1 - \hat{\mathbf{y}}) \mathbf{v}^\top$$