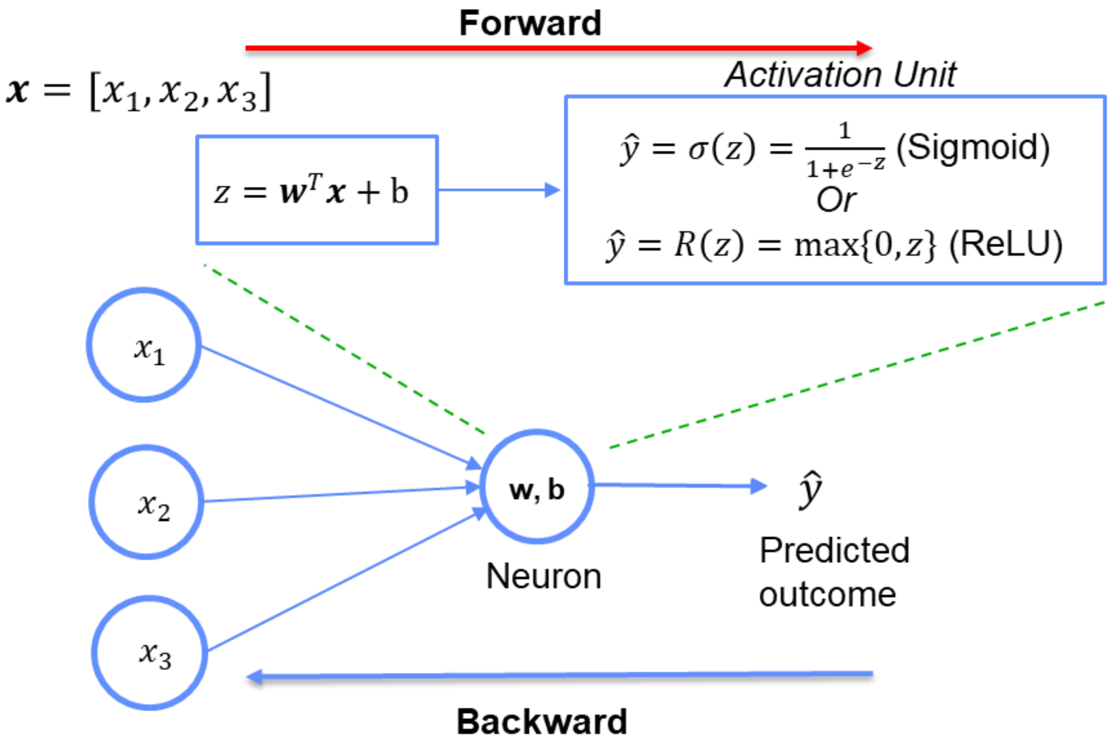
Discussion Neural Network

ECE 498 Data Science & Engr Spring 2020

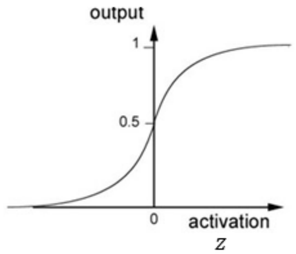
04/24/2020

Perceptron

- The core of the neural network is perceptron model

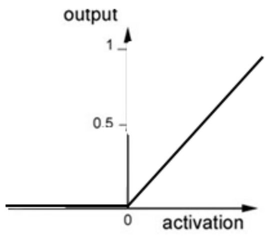


Sigmoid Function



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

ReLU Function



$$\mathrm{R}(z) = \max\{0, z\}$$

Update Rule (Backward):

$$w_{t+1} = w_t - \eta \nabla J(w_t)$$

η : Learning rate

Loss

$$J(w) = \frac{1}{N}\sum_{i=1}^{N} L(w.x^{(i)}, y^{(i)})$$

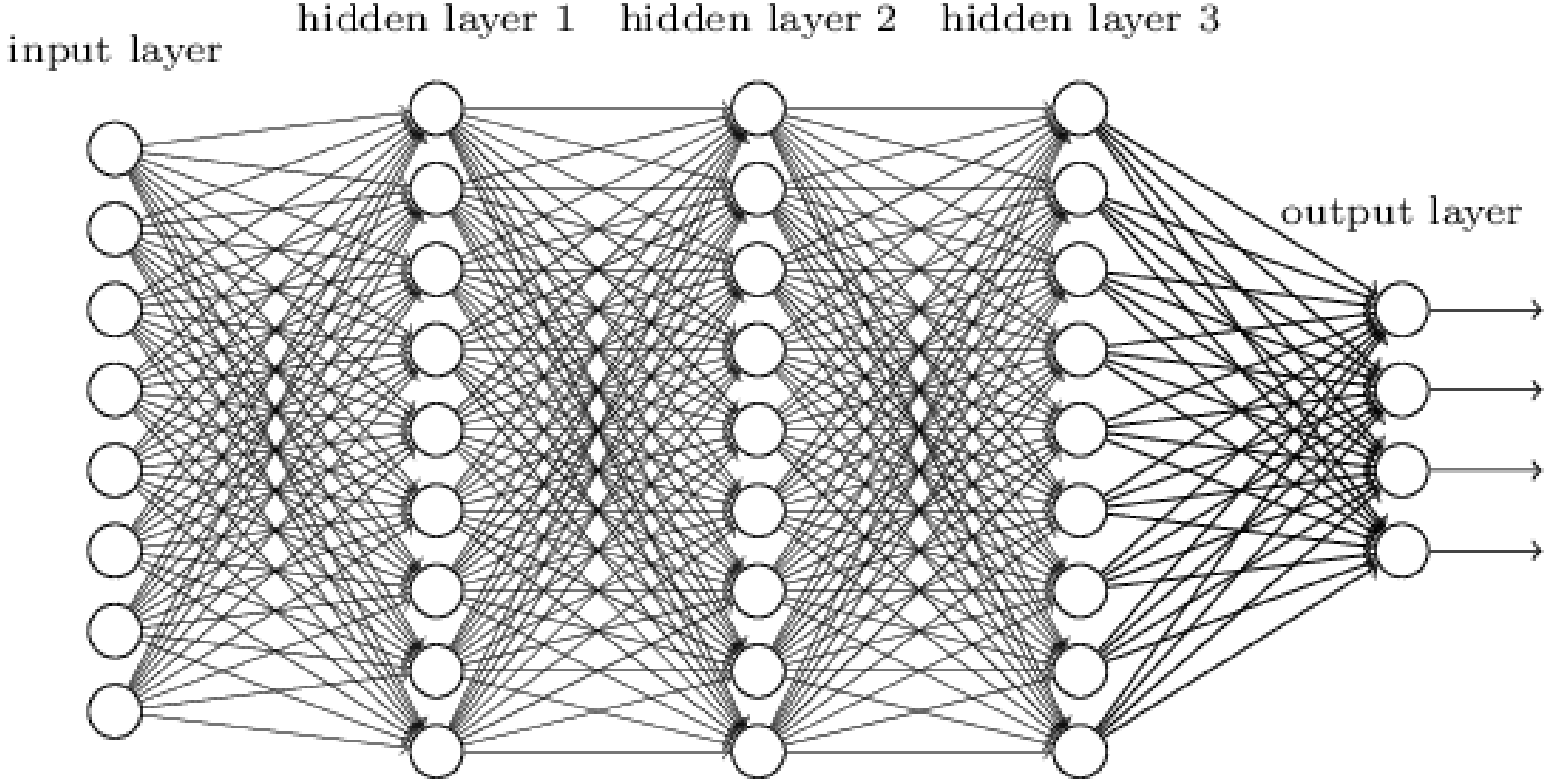
Computing Gradient

$$\nabla J(\mathbf{w}_0) = \left(\frac{\partial J(\mathbf{w})}{\partial w_0}, \frac{\partial J(\mathbf{w})}{\partial w_1}, \cdots, \frac{\partial J(\mathbf{w})}{\partial w_n}\right)_{\mathbf{w}_0}$$

N: number of samples $x^{(i)}$: feature of i^{th} sample

Neural Network

- When we stack perceptron together (universal function approximator)
- we can have many hidden layers, and each layer of arbitrary size (# of neurons)



Why Activation Function?

- Recall that at each layer:

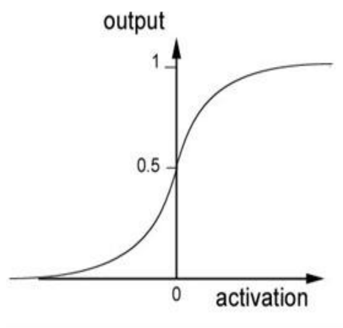
$$z = \boldsymbol{w}^T\boldsymbol{x} + \mathrm{b}$$

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \text{ (Sigmoid)}$$

Or

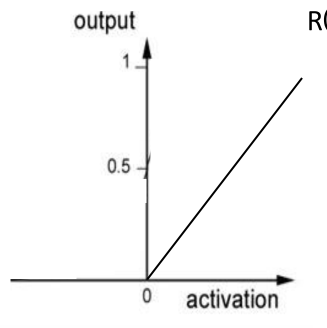
$$\hat{y} = R(z) = \max\{0, z\} \text{ (ReLU)}$$

Sigmoid Function



$$\sigma(z) = \frac{1}{1+e^{-z}}$$

ReLU Function

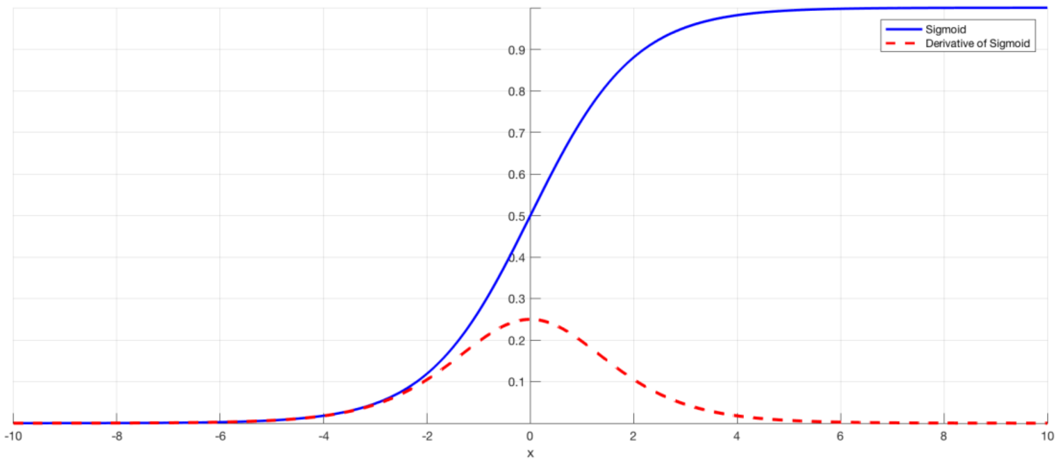


$$\mathrm{R}(z) = \max\{0, z\}$$

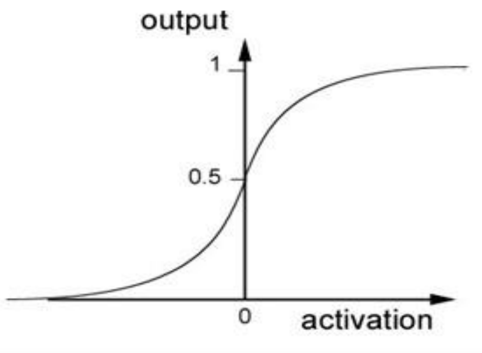
- Activation function adds non-linearity into the network.

ReLU?

- ReLU vs Sigmoid



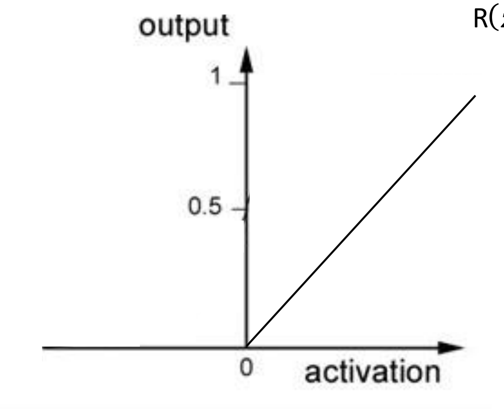
Sigmoid Function



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

ReLU Function

$$\mathrm{R}(z) = \max\{0, z\}$$

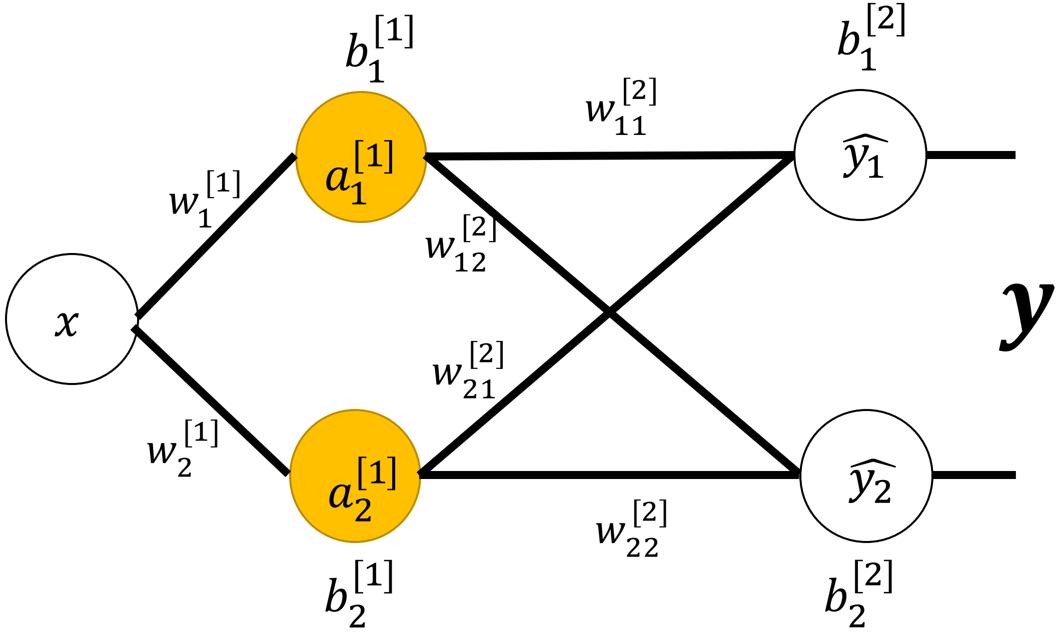


- Vanishing gradient problem in deep networks.
- But could run into numerical issue

Validation vs Test set

- Validation == Test Set? Not really strictly speaking
- Validation set: Dataset you validate your model on during training for model selection
- Test set: Dataset you test your model on after everything is fixed for unbiased-generalization error, **test set should not affect training procedure, including model selection**

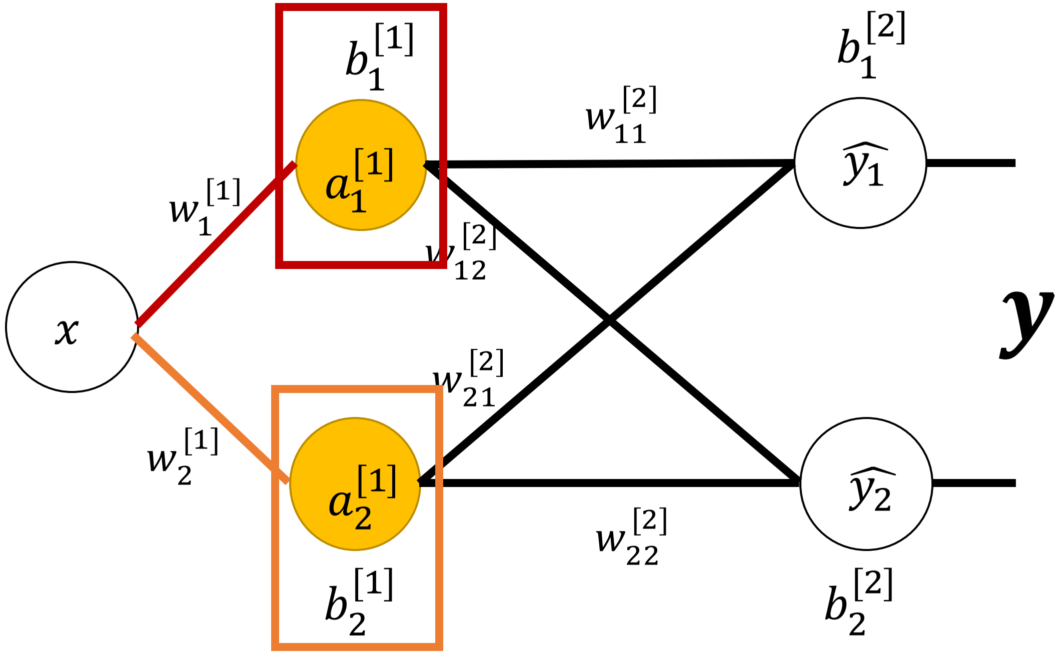
Example



- Yellow --- with activation function
- Activation --- Sigmoid --- $\sigma(z) = a$
- Loss --- L_2 squared loss, also known as the squared L-2 norm
- Squared $L_2(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\widehat{\boldsymbol{y}} - \boldsymbol{y})||_2^2$

Forward Pass 1

- Yellow --- with activation function
- Activation --- Sigmoid
- $\sigma(z_i) = (1 + e^{-z_i})^{-1} = a_i$
- Loss --- L_2 squared loss, also known as the squared L-2 norm
- Squared $L_2(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2$

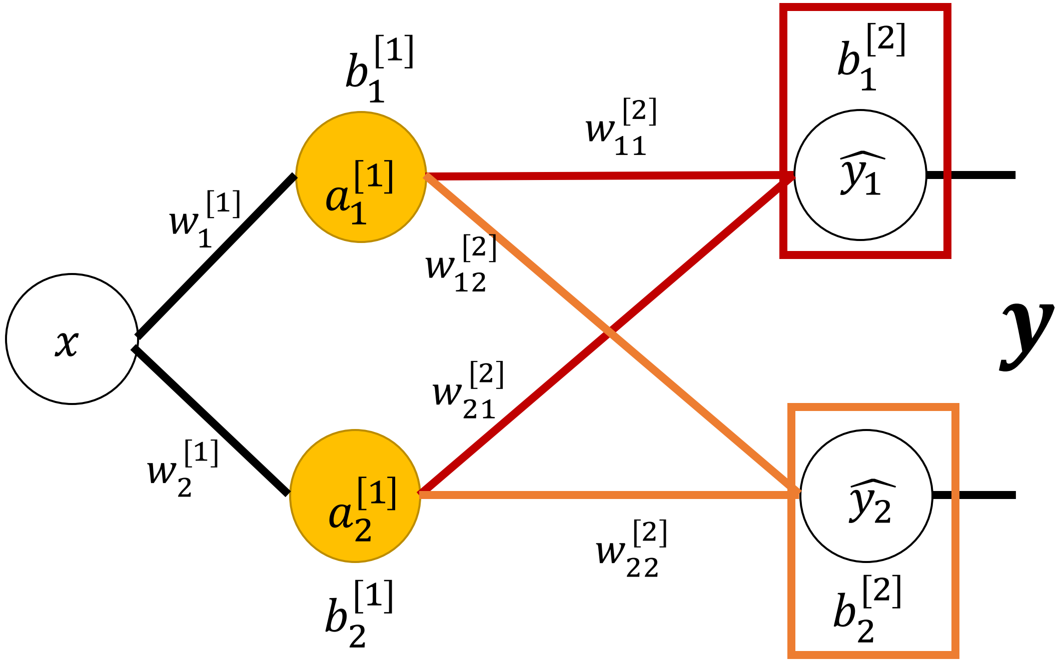


$$z_1^{[1]} = x\, w_1^{[1]} + b_1^{[1]}$$
$$a_1^{[1]} = \sigma(z_1^{[1]})$$

$$z_2^{[1]} = x\, w_2^{[1]} + b_2^{[1]}$$
$$a_2^{[1]} = \sigma(z_2^{[1]})$$

Forward Pass 2

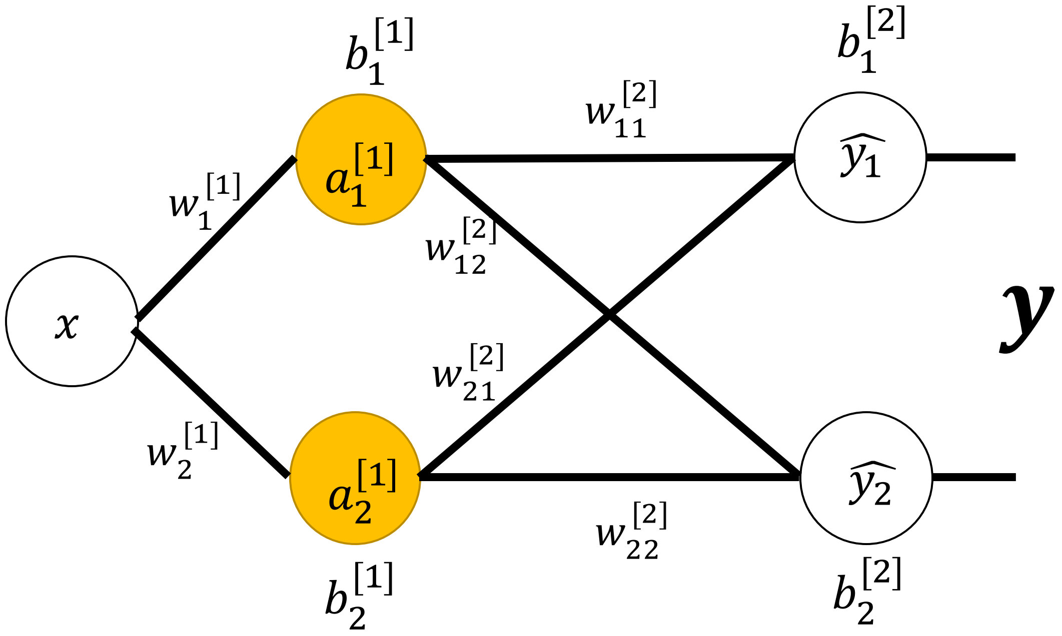
- Yellow --- with activation function
- Activation --- Sigmoid
- $\sigma(z_i) = (1 + e^{-z_i})^{-1} = a_i$
- Loss --- L_2 squared loss, also known as the squared L-2 norm
- Squared $L_2(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2$



$$\widehat{y_1} = w_{11}^{[2]} a_1^{[1]} + w_{21}^{[2]} a_2^{[1]} + b_1^{[2]}$$

$$\widehat{y_2} = w_{12}^{[2]} a_1^{[1]} + w_{22}^{[2]} a_2^{[1]} + b_2^{[2]}$$

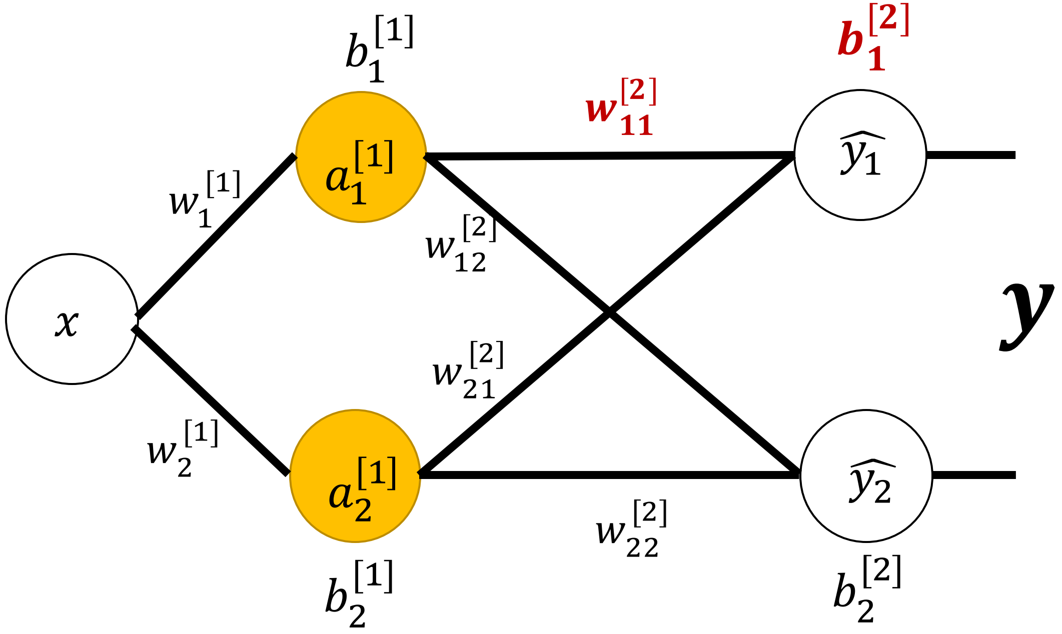
Calculate Loss



- Yellow --- with activation function
- Activation --- Sigmoid
- $\sigma(z_i) = (1 + e^{-z_i})^{-1} = a_i$
- Loss --- L_2 squared loss, also known as the squared L-2 norm
- Squared $L_2(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\widehat{\boldsymbol{y}} - \boldsymbol{y})||_2^2$

$$L = ||(\widehat{\boldsymbol{y}} - \boldsymbol{y})||_2^2 = (\widehat{y_1} - y_1)^2 + (\widehat{y_2} - y_2)^2$$

Back Propagation 1



- Yellow --- with activation function
- Activation --- Sigmoid
- $\sigma(z_i) = (1 + e^{-z_i})^{-1} = a_i$
- Loss --- L_2 squared loss, also known as the squared L-2 norm
- Squared $L_2(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\widehat{\boldsymbol{y}} - \boldsymbol{y})||_2^2$

$$L = ||(\widehat{\boldsymbol{y}} - \boldsymbol{y})||_2^2 = (\widehat{y_1} - y_1)^2 + (\widehat{y_2} - y_2)^2$$

Lets update $w_{11}^{[2]}$ and $b_1^{[2]}$ using **gradient descent**

Compute the gradients:

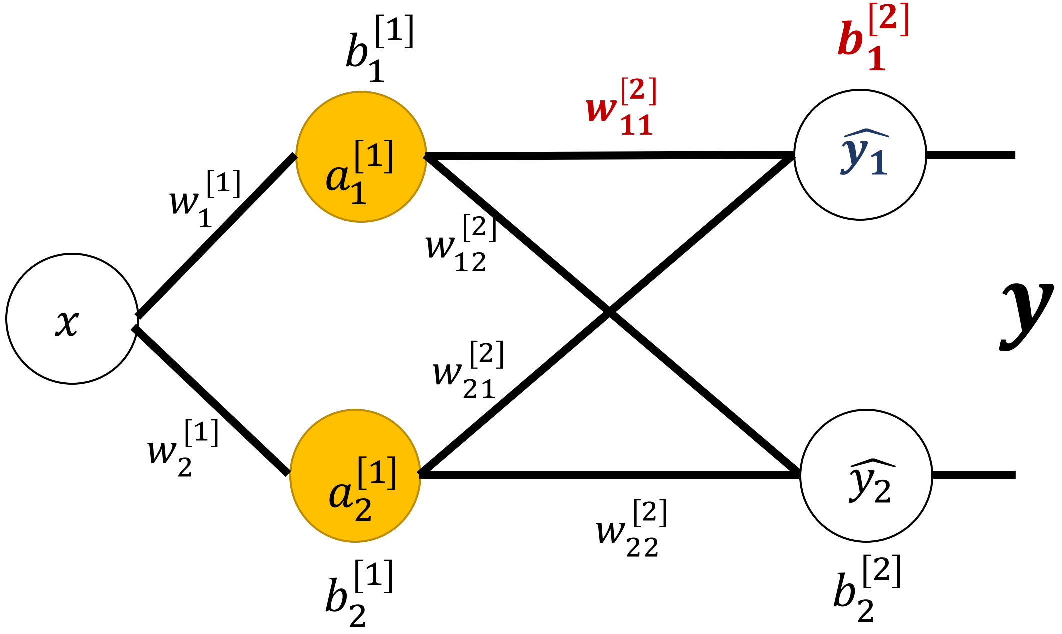
$$\frac{\partial L}{\partial w_{11}^{[2]}}$$

and

$$\frac{\partial L}{\partial b_1^{[2]}}$$

Back Propagation 1

- Yellow --- with activation function
- Activation --- Sigmoid
- $\sigma(z_i) = (1 + e^{-z_i})^{-1} = a_i$
- Loss --- L_2 squared loss, also known as the squared L-2 norm
- Squared $L_2(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2$



$$L = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2 = (\widehat{y_1} - y_1)^2 + (\widehat{y_2} - y_2)^2$$

$$\boldsymbol{\widehat{y_1}} = \boldsymbol{w_{11}^{[2]}} a_1^{[1]} + w_{21}^{[2]} a_2^{[1]} + \boldsymbol{b_1^{[2]}}$$

$$\widehat{y_2} = w_{12}^{[2]} a_1^{[1]} + w_{22}^{[2]} a_2^{[1]} + b_2^{[2]}$$

$$\frac{\partial L}{\partial \widehat{y_1}} = 2(\widehat{y_1} - y_1)$$

$$\frac{\partial \widehat{y_1}}{\partial w_{11}^{[2]}} = a_1^{[1]}$$

(We get $a_1^{[1]}$ from the forward pass)

$$\boldsymbol{\frac{\partial L}{\partial w_{11}^{[2]}} = \frac{\partial L}{\partial \widehat{y_1}} \frac{\partial \widehat{y_1}}{\partial w_{11}^{[2]}} = 2(\widehat{y_1} - y_1)\, a_1^{[1]}}$$

$$\boldsymbol{\frac{\partial L}{\partial b_1^{[2]}} = \frac{\partial L}{\partial \widehat{y_1}} \frac{\partial \widehat{y_1}}{\partial b_1^{[2]}} = 2(\widehat{y_1} - y_1)}$$

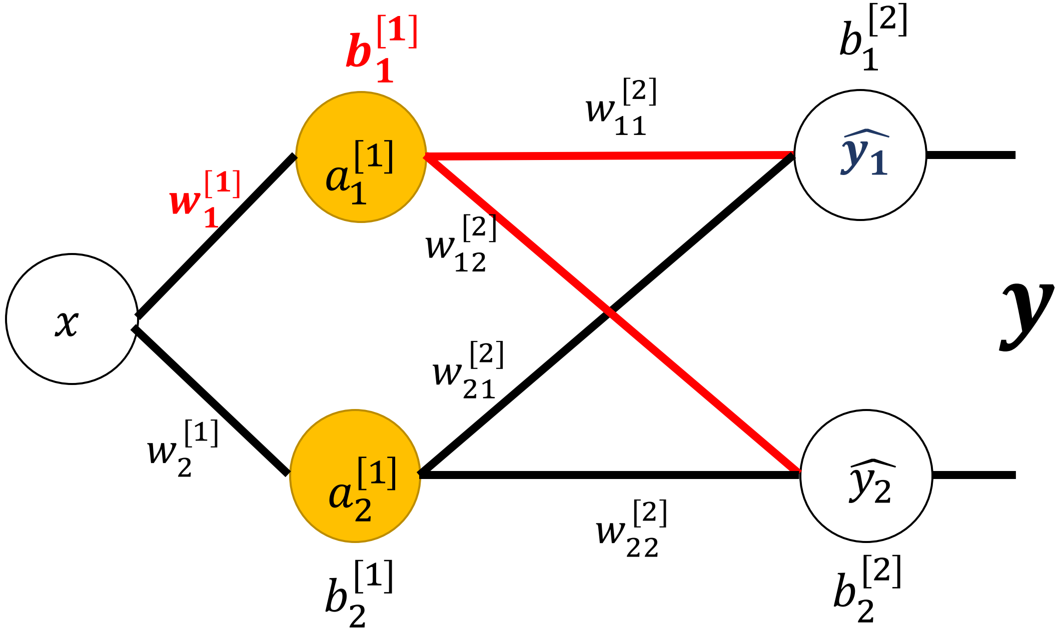
Compute the gradients with chain rule:

$$\frac{\partial L}{\partial w_{11}^{[2]}} = \frac{\partial L}{\partial \widehat{y_1}} \frac{\partial \widehat{y_1}}{\partial w_{11}^{[2]}}$$

Similarly

$$\frac{\partial L}{\partial b_1^{[2]}} = \frac{\partial L}{\partial \widehat{y_1}} \frac{\partial \widehat{y_1}}{\partial b_1^{[2]}}$$

Back Propagation 2



$$z_1^{[1]} = x\, \boldsymbol{w_1^{[1]}} + \boldsymbol{b_1^{[1]}}$$
$$a_1^{[1]} = \sigma(z_1^{[1]})$$

$$z_2^{[1]} = x\, w_2^{[1]} + b_2^{[1]}$$
$$a_2^{[1]} = \sigma(z_2^{[1]})$$

$$\widehat{y_1} = w_{11}^{[2]} a_1^{[1]} + w_{21}^{[2]} a_2^{[1]} + b_1^{[2]}$$

$$\widehat{y_2} = w_{12}^{[2]} a_1^{[1]} + w_{22}^{[2]} a_2^{[1]} + b_2^{[2]}$$

$$L = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2 = (\widehat{y_1} - y_1)^2 + (\widehat{y_2} - y_2)^2$$

$$\frac{\partial L}{\partial \hat{\boldsymbol{y}}} = \left[\frac{\partial L}{\partial \widehat{y_1}} = 2(\widehat{y_1} - y_1) \quad \frac{\partial L}{\partial \widehat{y_2}} = 2(\widehat{y_2} - y_2) \right]$$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial a_1} = \begin{bmatrix} \dfrac{\partial\, \widehat{y_1}}{\partial a_1} = w_{11}^{[2]} \\ \dfrac{\partial\, \widehat{y_2}}{\partial a_1} = w_{12}^{[2]} \end{bmatrix}$$

$$\frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} = \sigma\left(z_1^{[1]}\right)\left(1 - \sigma\left(z_1^{[1]}\right)\right)$$

$$\frac{\partial z_1^{[1]}}{\partial w_1^{[1]}} = x \quad \boldsymbol{and} \quad \frac{\partial z_1^{[1]}}{\partial b_1^{[1]}} = 1$$

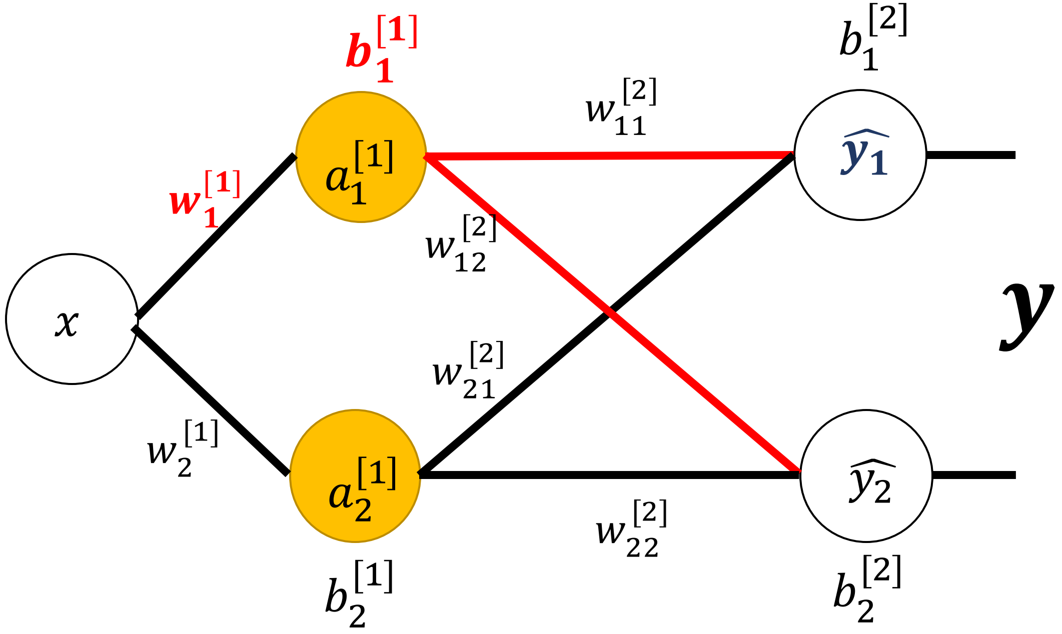
Compute the gradients with chain rule:

$$\frac{\partial L}{\partial w_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_1^{[1]}}$$

Similarly

$$\frac{\partial L}{\partial b_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial b_1^{[1]}}$$

Back Propagation 2



$$z_1^{[1]} = x\, \boldsymbol{w_1^{[1]}} + \boldsymbol{b_1^{[1]}}$$
$$a_1^{[1]} = \sigma(z_1^{[1]})$$
$$z_2^{[1]} = x\, w_2^{[1]} + b_2^{[1]}$$
$$a_2^{[1]} = \sigma(z_2^{[1]})$$
$$\hat{y_1} = w_{11}^{[2]} a_1^{[1]} + w_{21}^{[2]} a_2^{[1]} + b_1^{[2]}$$
$$\hat{y_2} = w_{12}^{[2]} a_1^{[1]} + w_{22}^{[2]} a_2^{[1]} + b_2^{[2]}$$
$$L = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2 = (\hat{y_1} - y_1)^2 + (\hat{y_2} - y_2)^2$$

$$\frac{\partial L}{\partial w_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_1^{[1]}}$$
$$= [2(\hat{y_1} - y_1) \quad 2(\hat{y_2} - y_2)] \begin{bmatrix} w_{11}^{[2]} \\ w_{12}^{[2]} \end{bmatrix} \sigma\left(z_1^{[1]}\right)\left(1 - \sigma\left(z_1^{[1]}\right)\right) x$$
$$= (2(\hat{y_1} - y_1) w_{11}^{[2]} + 2(\hat{y_2} - y_2) w_{12}^{[2]})\, \sigma\left(z_1^{[1]}\right)\left(1 - \sigma\left(z_1^{[1]}\right)\right) x$$

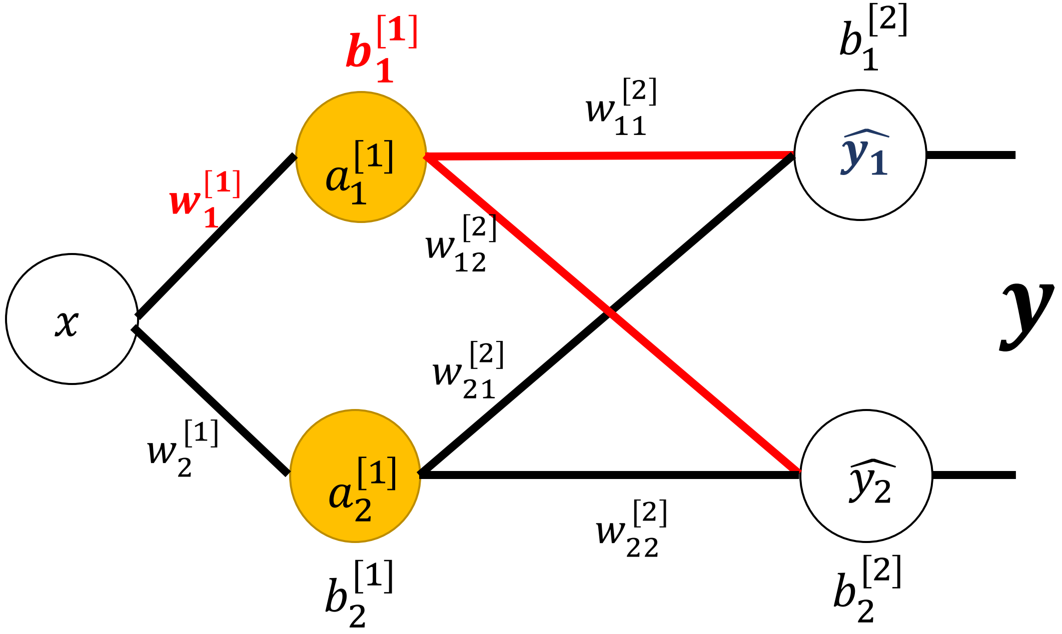
Compute the gradients with chain rule:

$$\frac{\partial L}{\partial w_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_1^{[1]}}$$

Similarly

$$\frac{\partial L}{\partial b_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial b_1^{[1]}}$$

Back Propagation 2



$$z_1^{[1]} = x\, \boldsymbol{w_1^{[1]}} + \boldsymbol{b_1^{[1]}}$$
$$a_1^{[1]} = \sigma(z_1^{[1]})$$
$$z_2^{[1]} = x\, w_2^{[1]} + b_2^{[1]}$$
$$a_2^{[1]} = \sigma(z_2^{[1]})$$
$$\widehat{y_1} = w_{11}^{[2]} a_1^{[1]} + w_{21}^{[2]} a_2^{[1]} + b_1^{[2]}$$
$$\widehat{y_2} = w_{12}^{[2]} a_1^{[1]} + w_{22}^{[2]} a_2^{[1]} + b_2^{[2]}$$
$$L = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||_2^2 = (\widehat{y_1} - y_1)^2 + (\widehat{y_2} - y_2)^2$$

Compute the gradients with chain rule:

$$\frac{\partial L}{\partial w_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_1^{[1]}}$$

Similarly

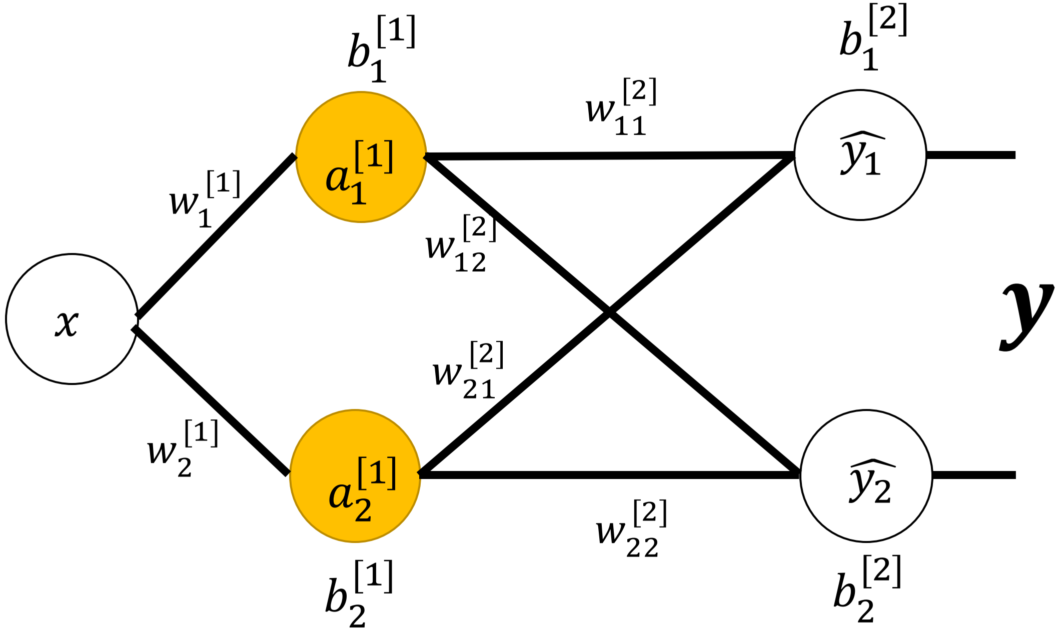
$$\frac{\partial L}{\partial b_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial b_1^{[1]}}$$

$$\frac{\partial L}{\partial b_1^{[1]}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial b_1^{[1]}}$$

$$= [2(\widehat{y_1} - y_1) \quad 2(\widehat{y_2} - y_2)] \begin{bmatrix} w_{11}^{[2]} \\ w_{12}^{[2]} \end{bmatrix} \sigma\left(z_1^{[1]}\right)\left(1 - \sigma\left(z_1^{[1]}\right)\right) 1$$

$$= (2(\widehat{y_1} - y_1) w_{11}^{[2]} + 2(\widehat{y_2} - y_2) w_{12}^{[2]})\, \sigma\left(z_1^{[1]}\right)\left(1 - \sigma\left(z_1^{[1]}\right)\right)$$

Update Rule Gradient Descent



$$w_{k+1} = w_k - \eta \frac{\partial L}{\partial w}$$

Walking in the opposite direction of the gradient