



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Regression with a perceptron

# Regression Problem Motivation

# Regression Problem Motivation

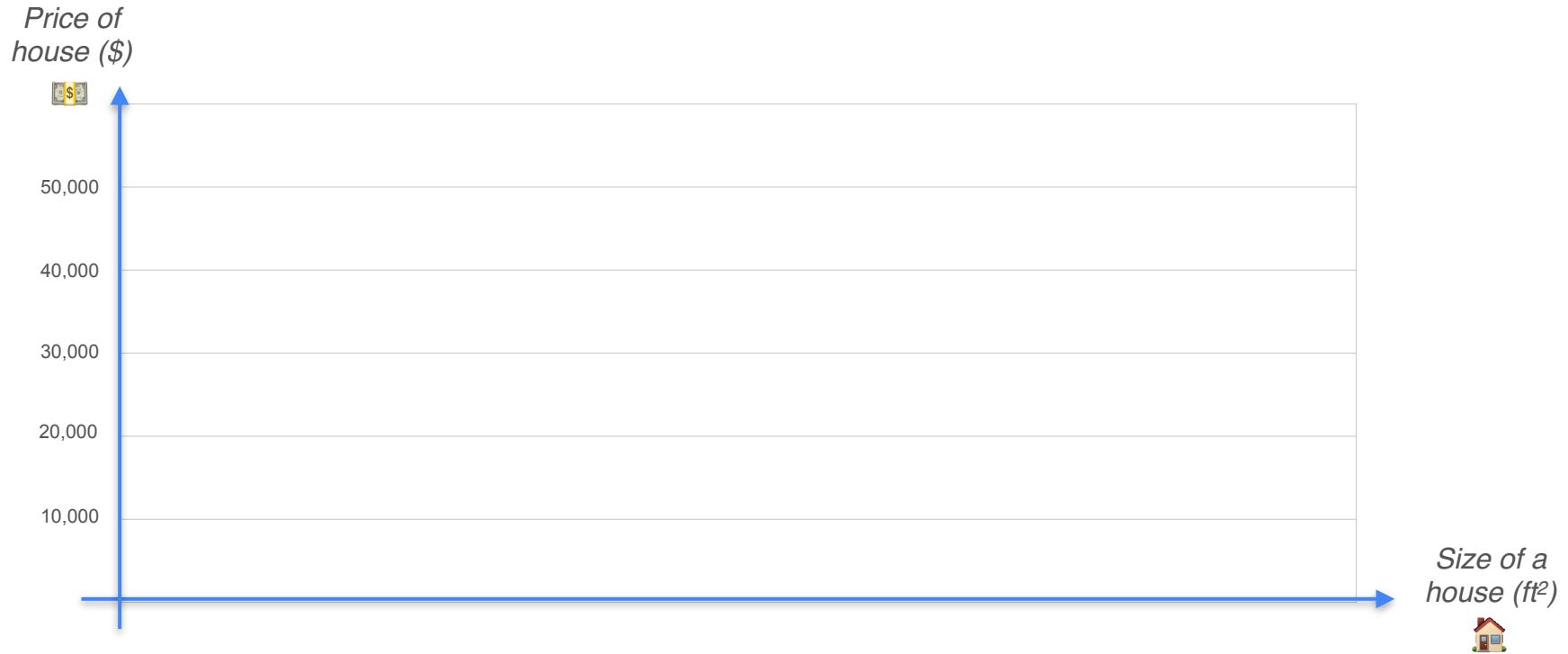
*Predicting*

***the price of a house***

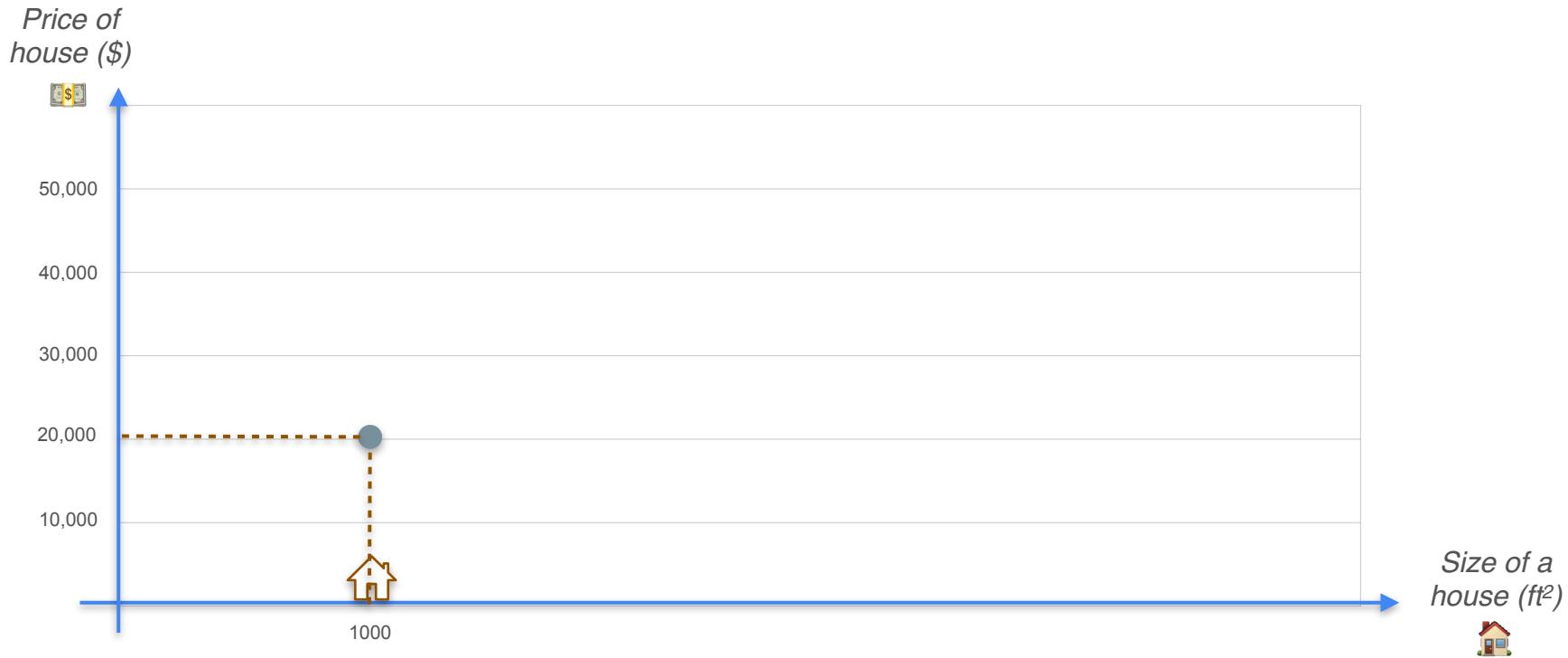
*from*

***the size of the house***

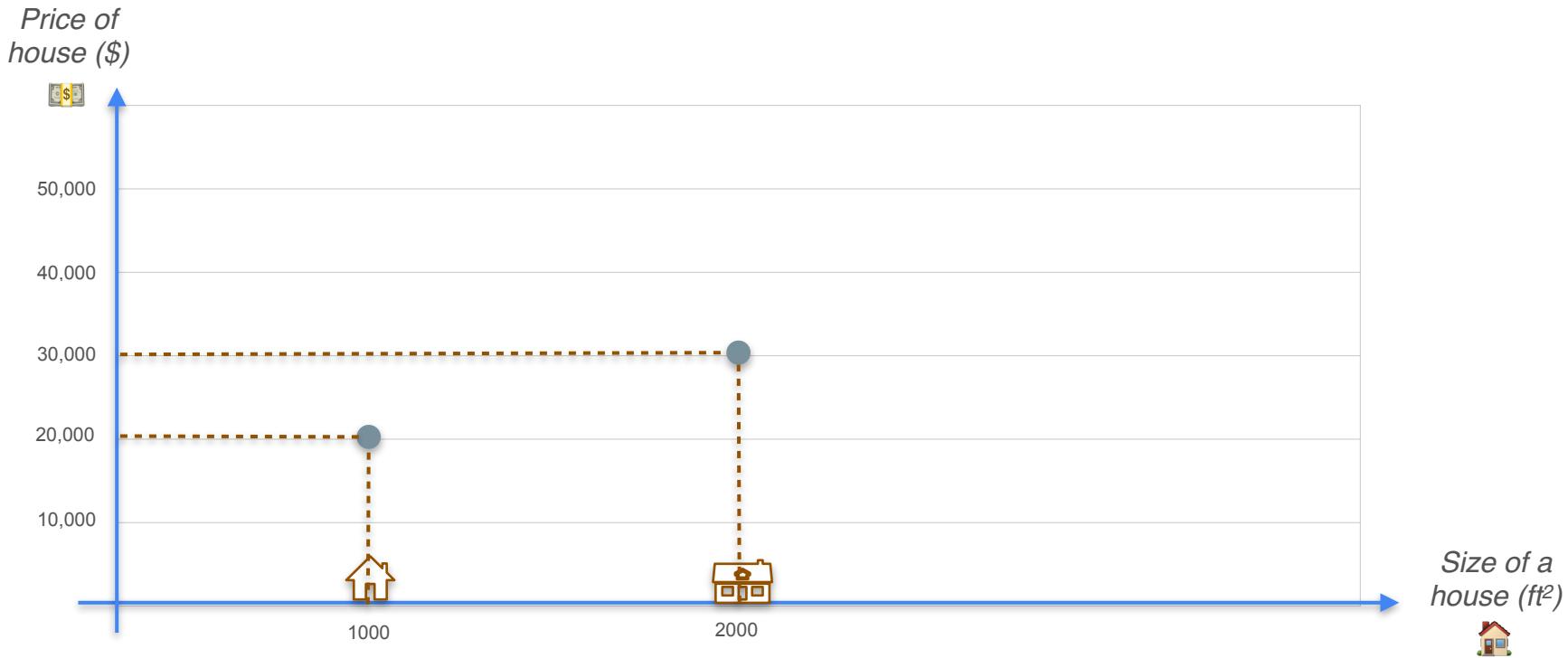
# Regression Problem Motivation



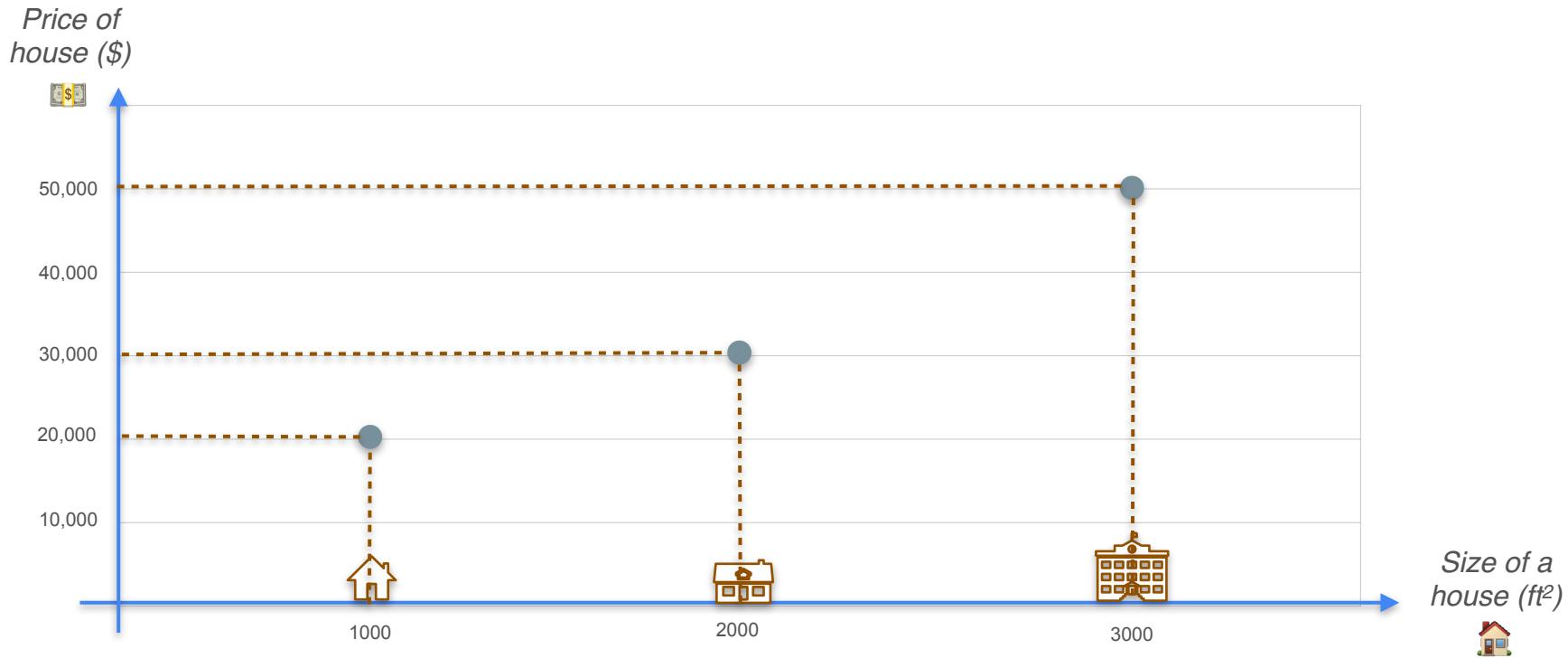
# Regression Problem Motivation



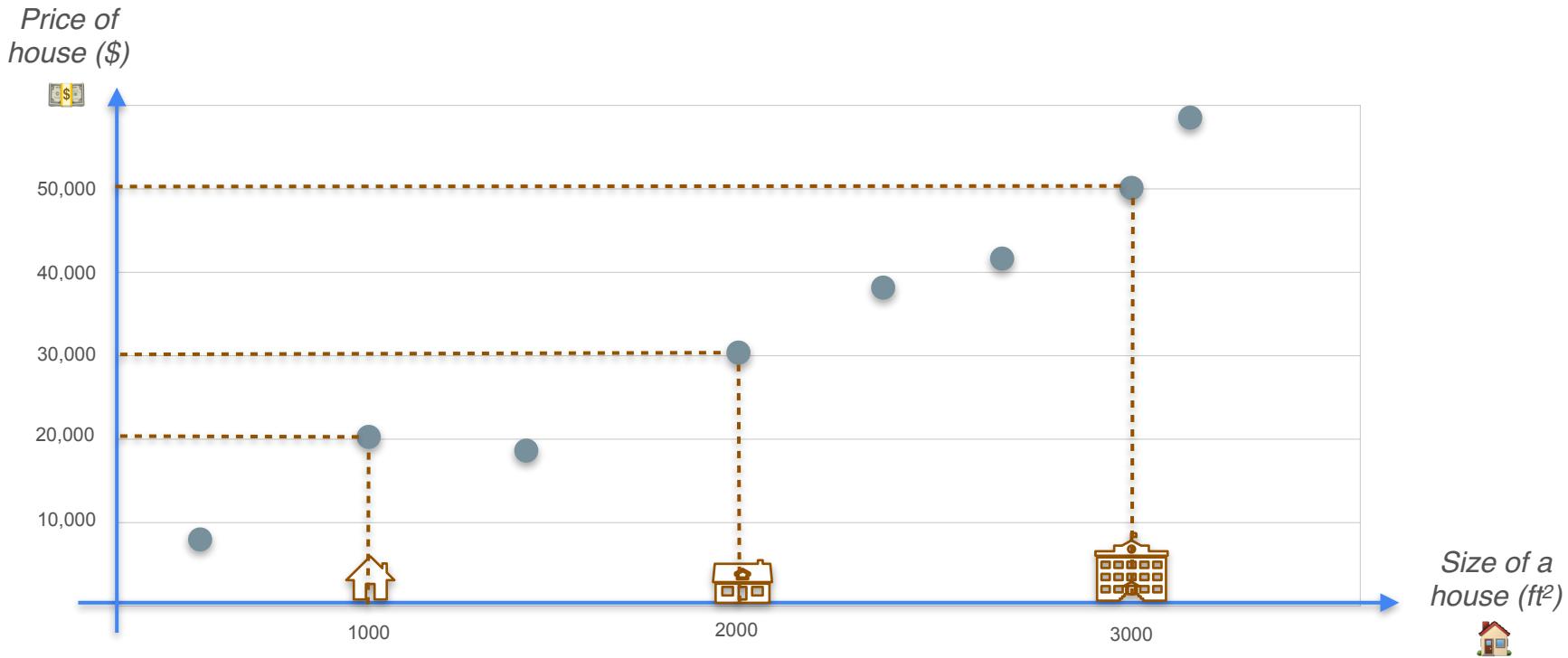
# Regression Problem Motivation



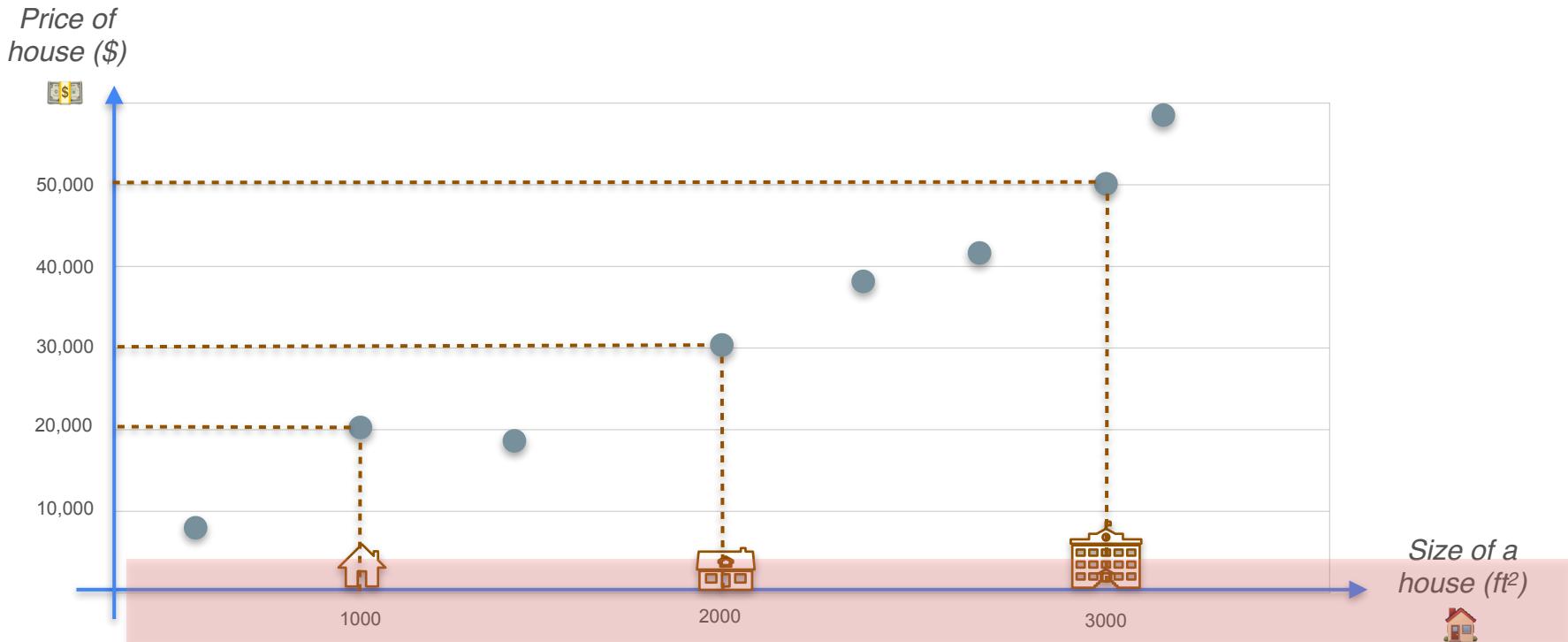
# Regression Problem Motivation



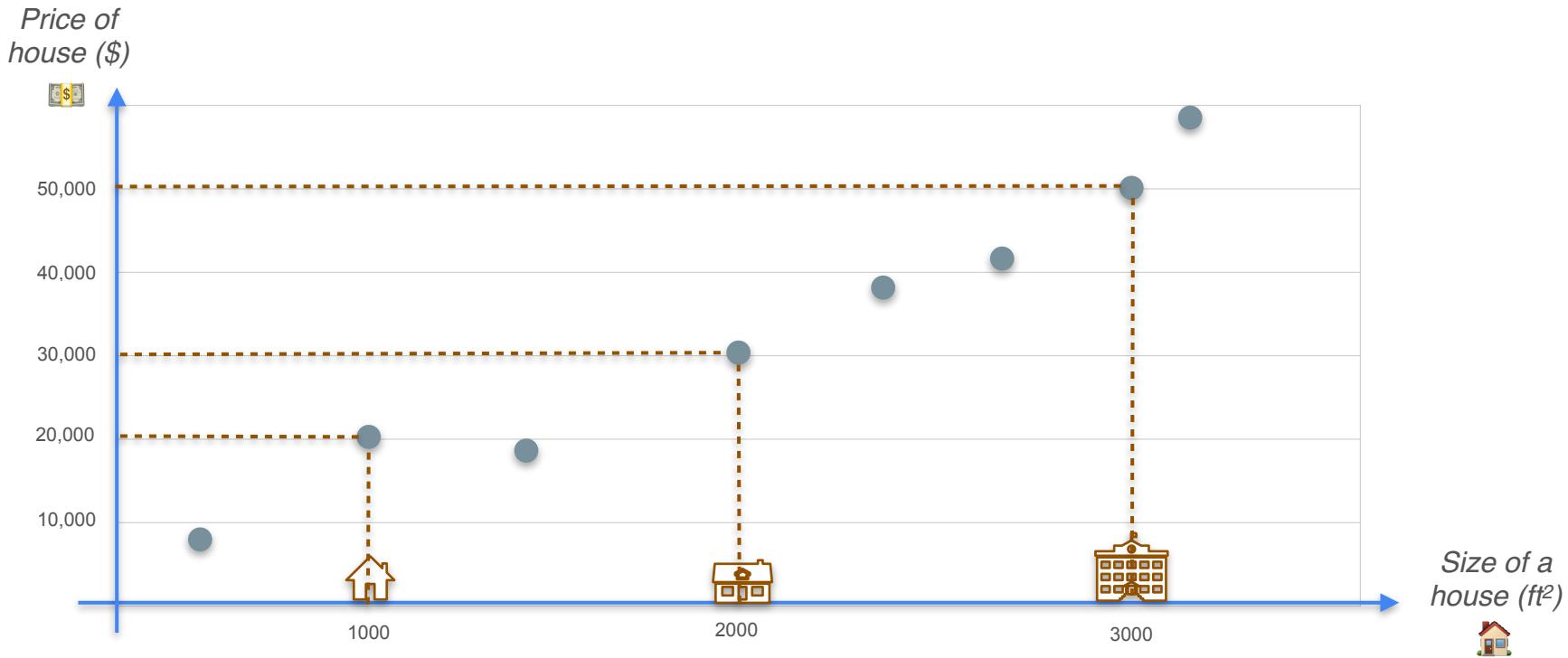
# Regression Problem Motivation



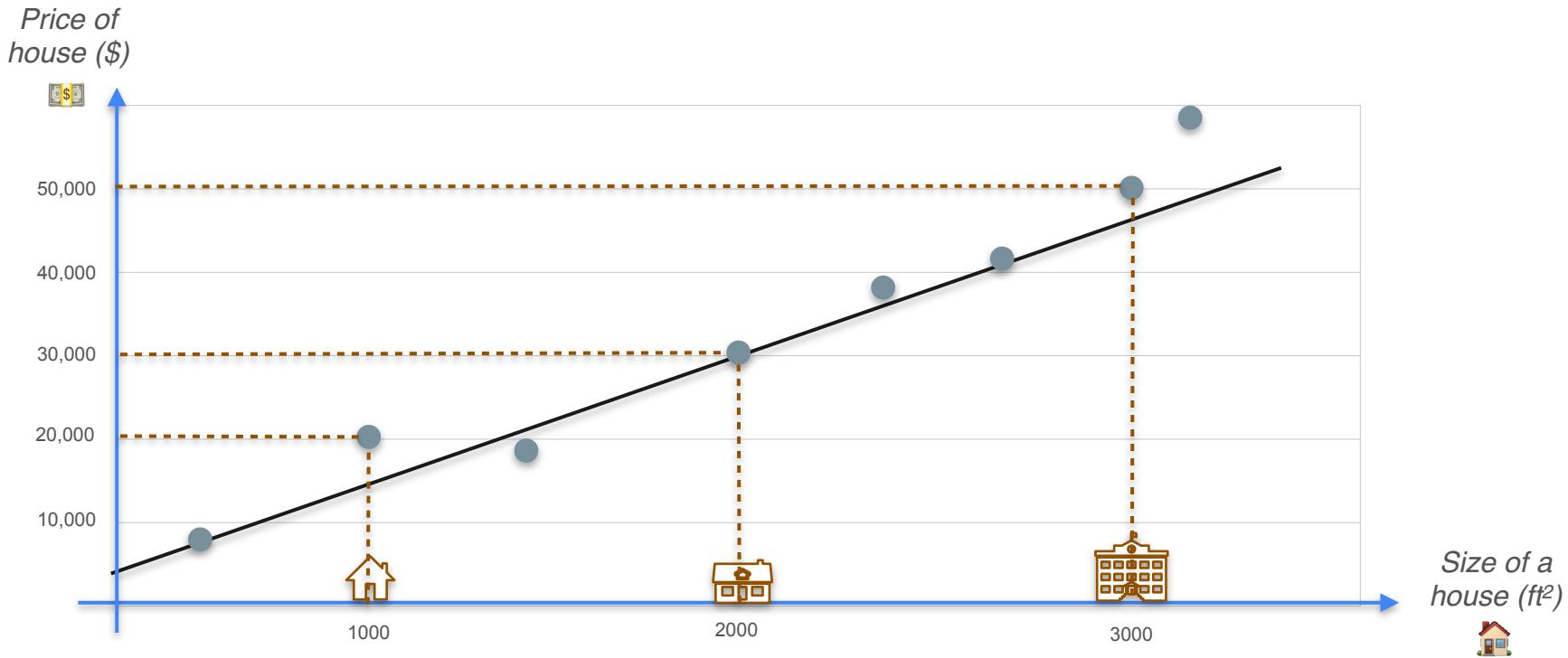
# Regression Problem Motivation



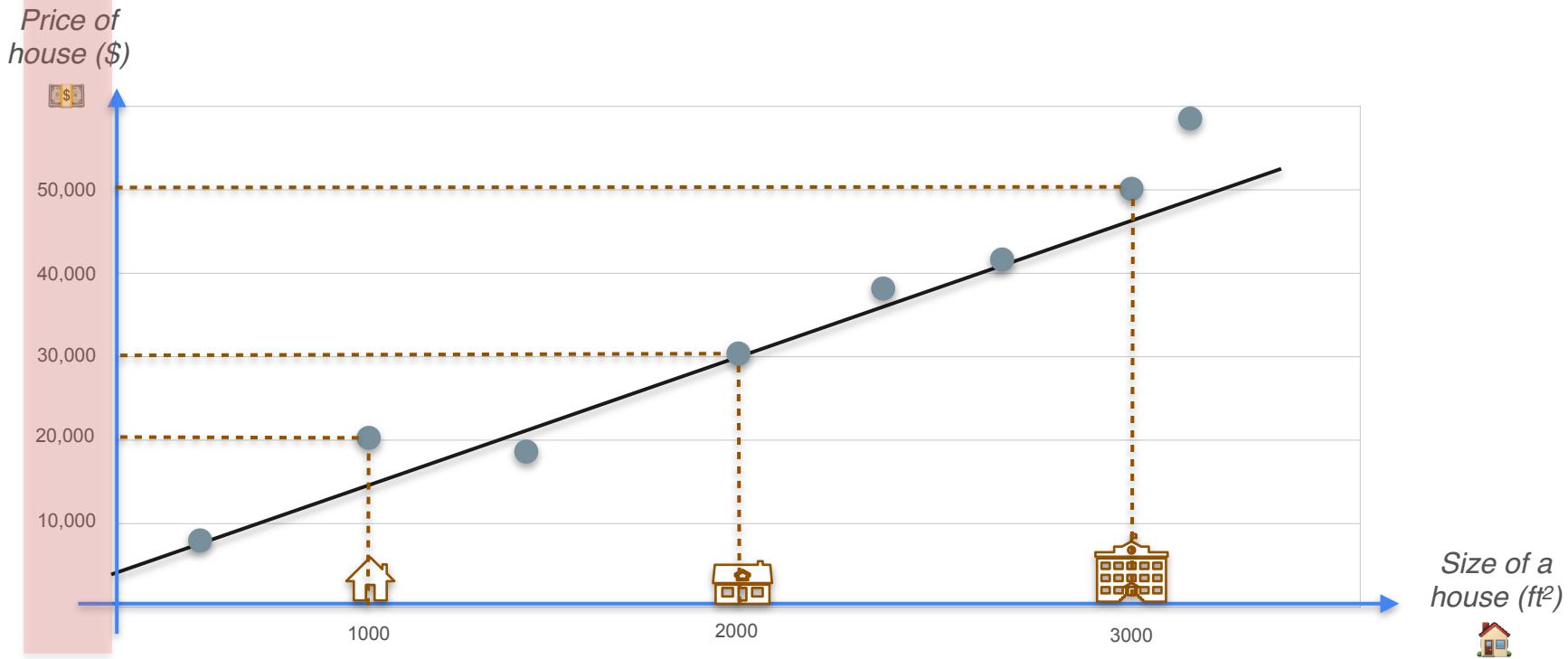
# Regression Problem Motivation



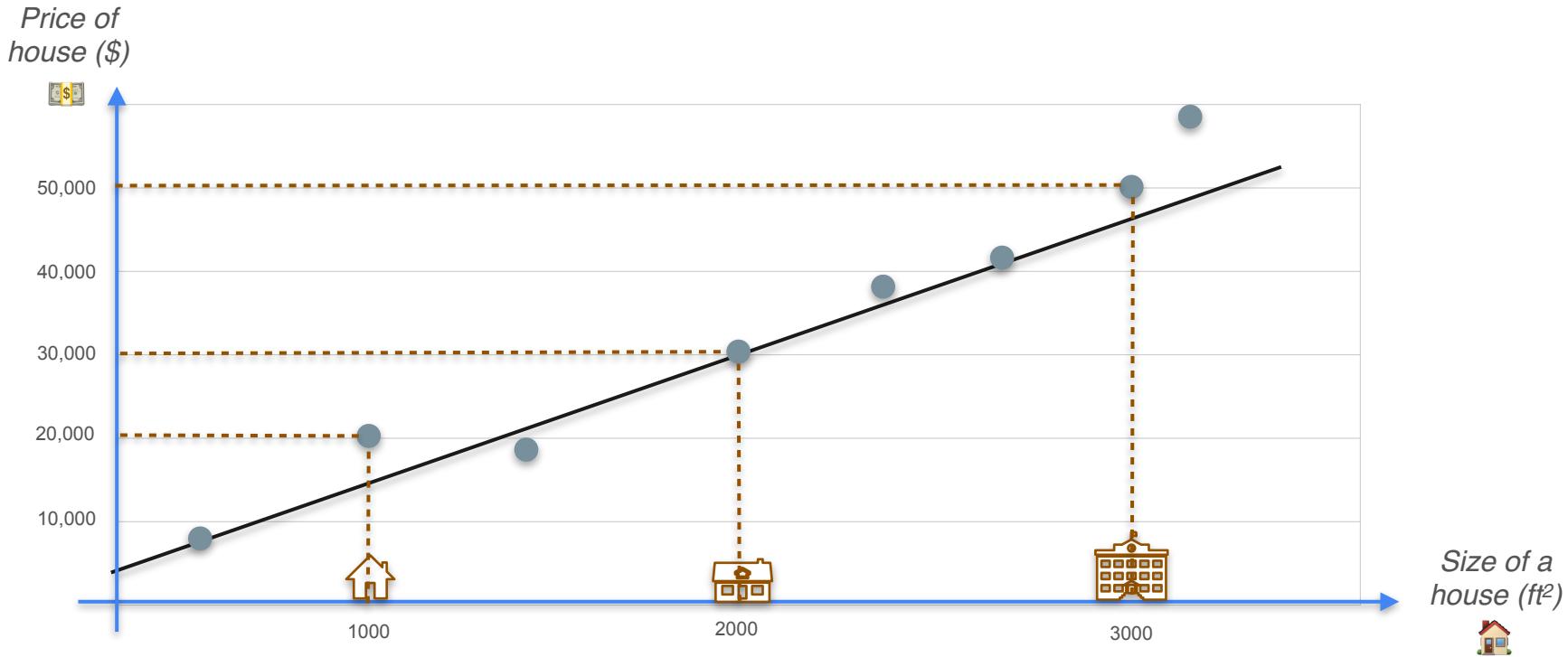
# Regression Problem Motivation



# Regression Problem Motivation



# Regression Problem Motivation



# Regression With a Perceptron

	<i>Size of a house (ft<sup>2</sup>)</i> 		<i>Price of house (\$)</i> 
			
			
			

# Regression With a Perceptron

	<i>Size of a house (ft<sup>2</sup>)</i> 		<i>Price of house (\$)</i> 
	$1000\text{ft}^2$		\$20,000
	$2000\text{ft}^2$		\$30,000
	$3000\text{ft}^2$		\$50,000

# Regression With a Perceptron

	<i>Size of a house (ft<sup>2</sup>)</i> 	<i>Number of rooms</i> 	<i>Price of house (\$)</i> 
	1000ft <sup>2</sup>	2	\$20,000
	2000ft <sup>2</sup>	4	\$30,000
	3000ft <sup>2</sup>	7	\$50,000

# Regression With a Perceptron

*Inputs*

*Size of a  
house (ft<sup>2</sup>)*



*Number of  
rooms*

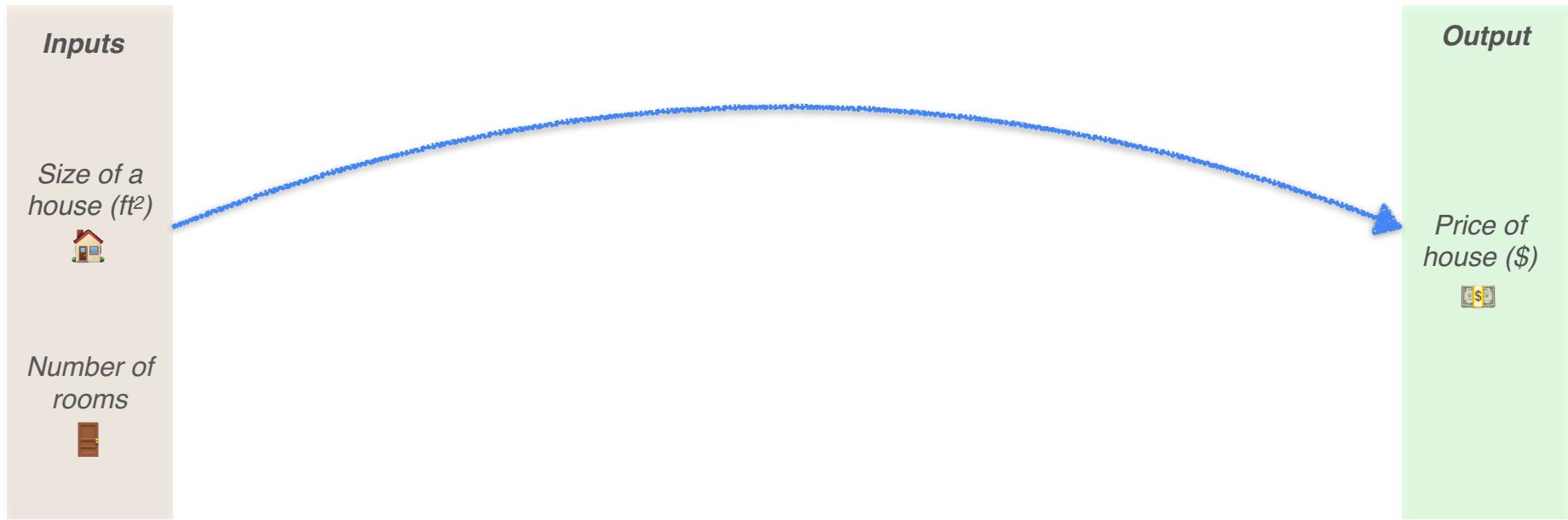


*Output*

*Price of  
house (\$)*



# Regression With a Perceptron



# Regression With a Perceptron

Single Layer Neural Network Perceptron

*Inputs*

*Size of a house (ft<sup>2</sup>)*



*Number of rooms*



*Output*

*Price of house (\$)*



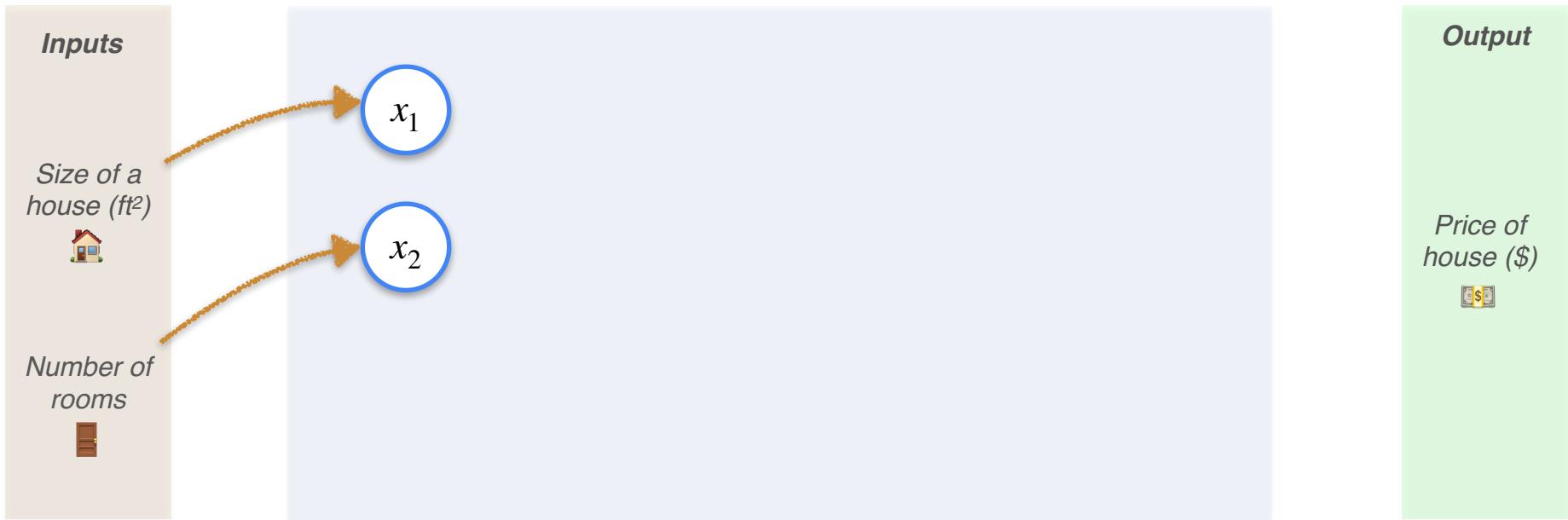
# Regression With a Perceptron

Single Layer Neural Network Perceptron



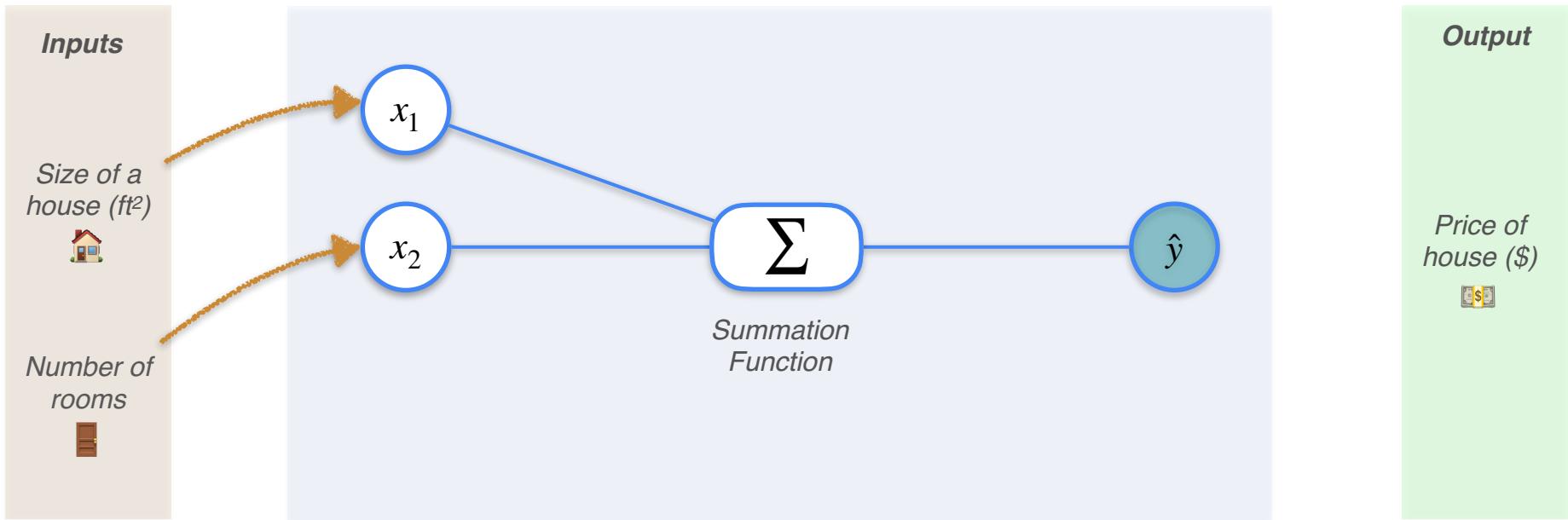
# Regression With a Perceptron

Single Layer Neural Network Perceptron



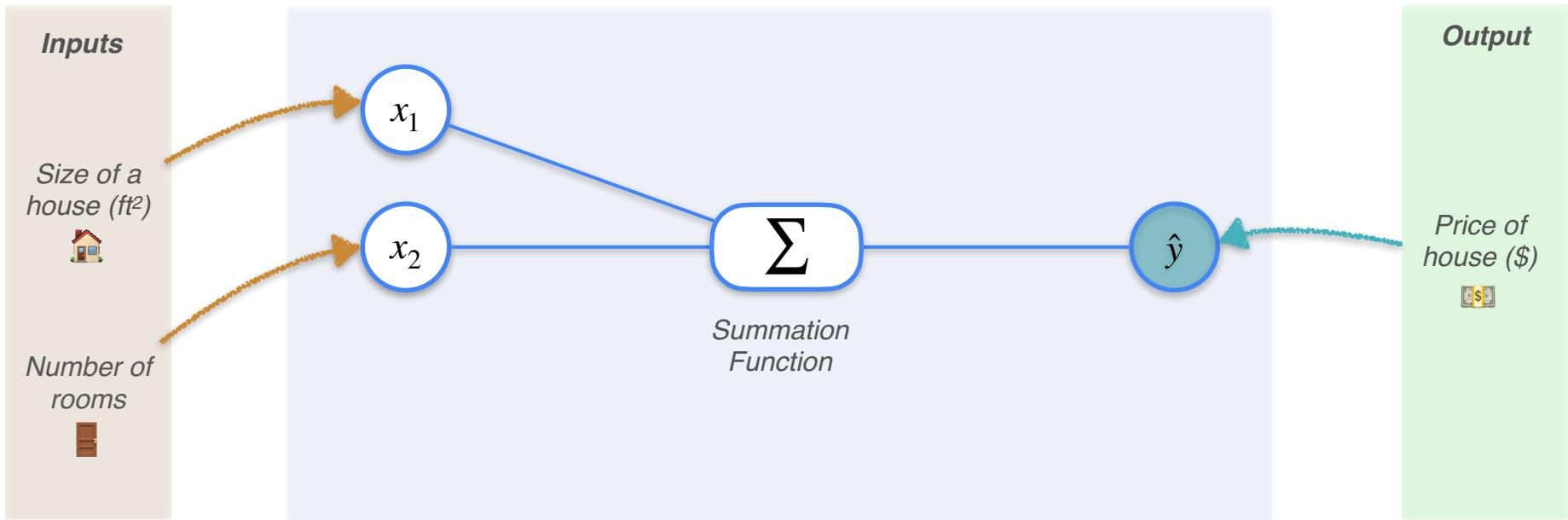
# Regression With a Perceptron

Single Layer Neural Network Perceptron



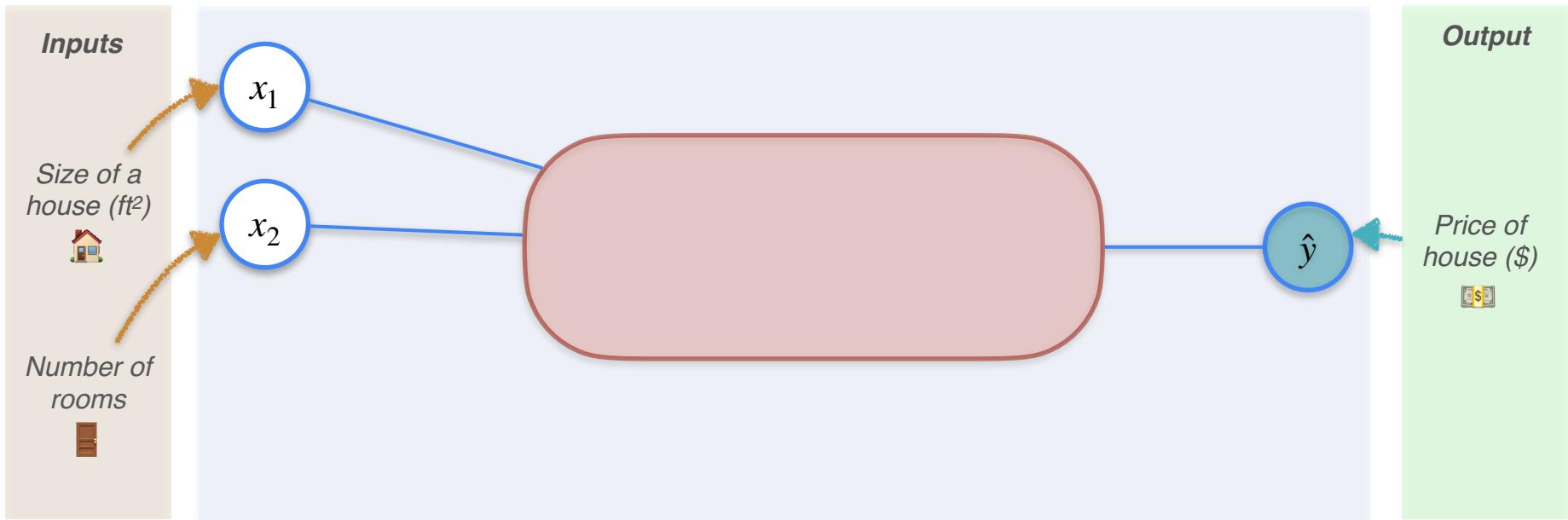
# Regression With a Perceptron

Single Layer Neural Network Perceptron



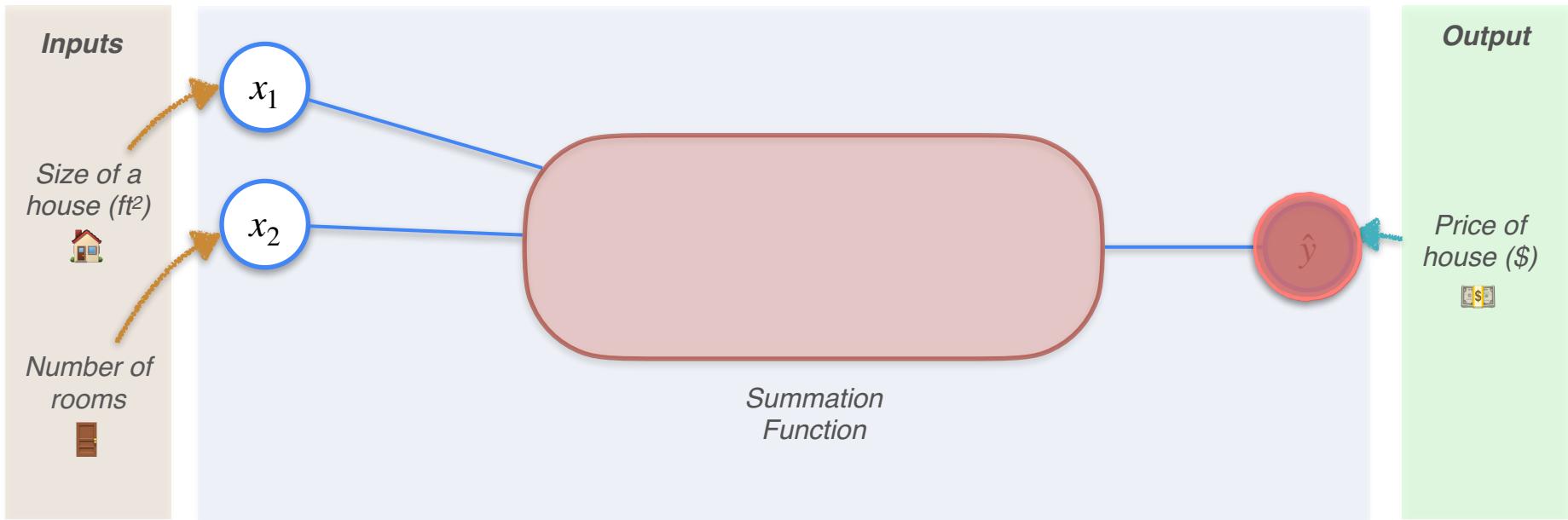
# Regression With a Perceptron

Single Layer Neural Network Perceptron



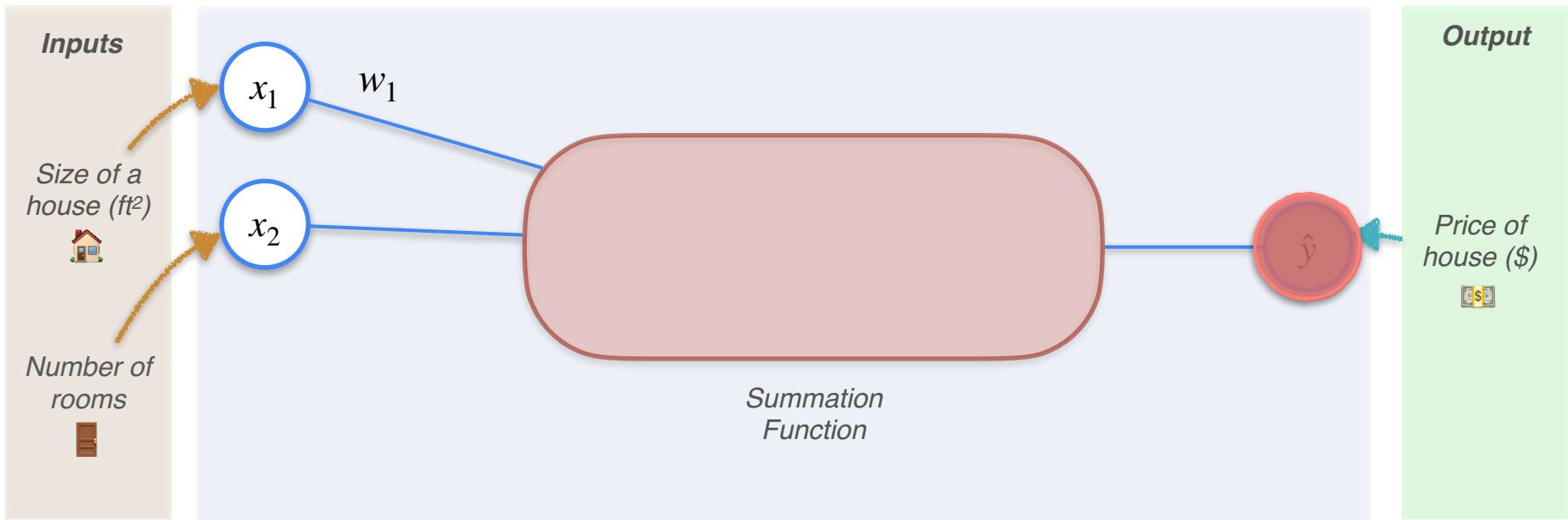
# Regression With a Perceptron

Single Layer Neural Network Perceptron



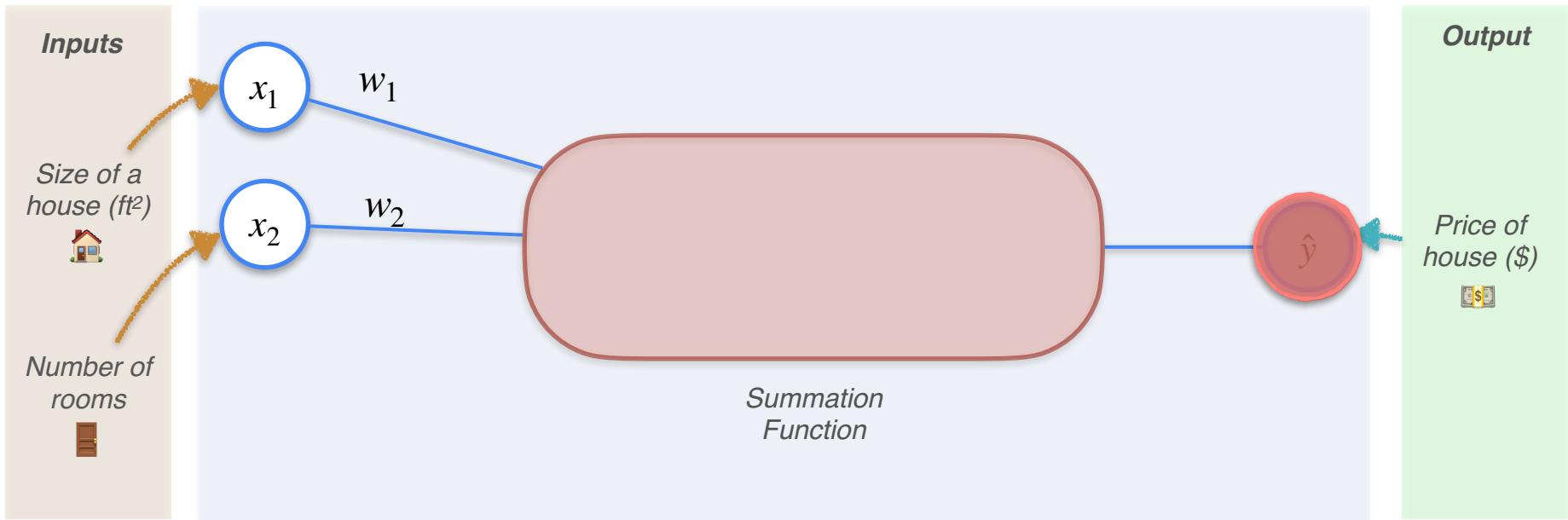
# Regression With a Perceptron

Single Layer Neural Network Perceptron



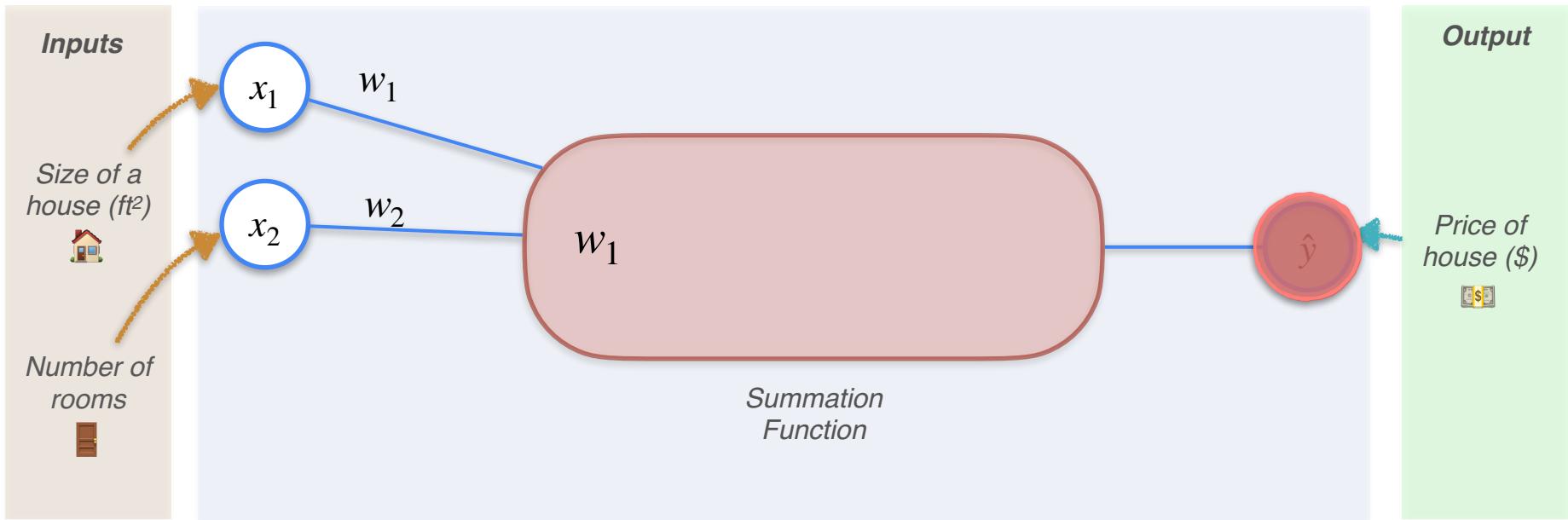
# Regression With a Perceptron

Single Layer Neural Network Perceptron



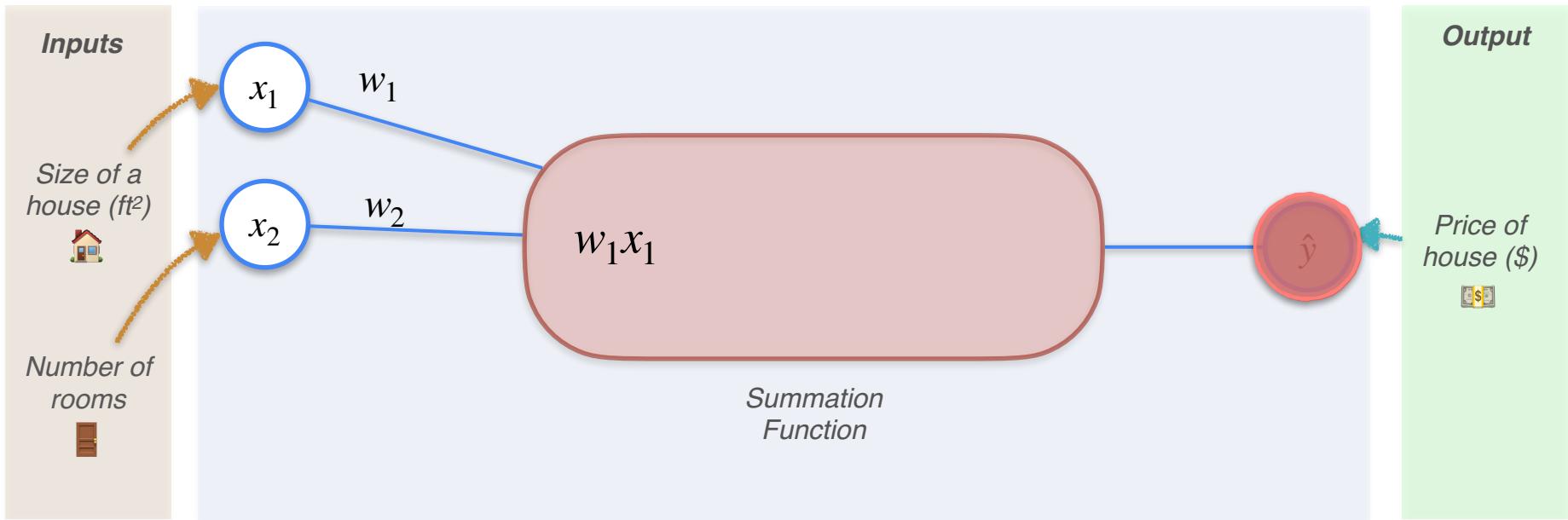
# Regression With a Perceptron

Single Layer Neural Network Perceptron



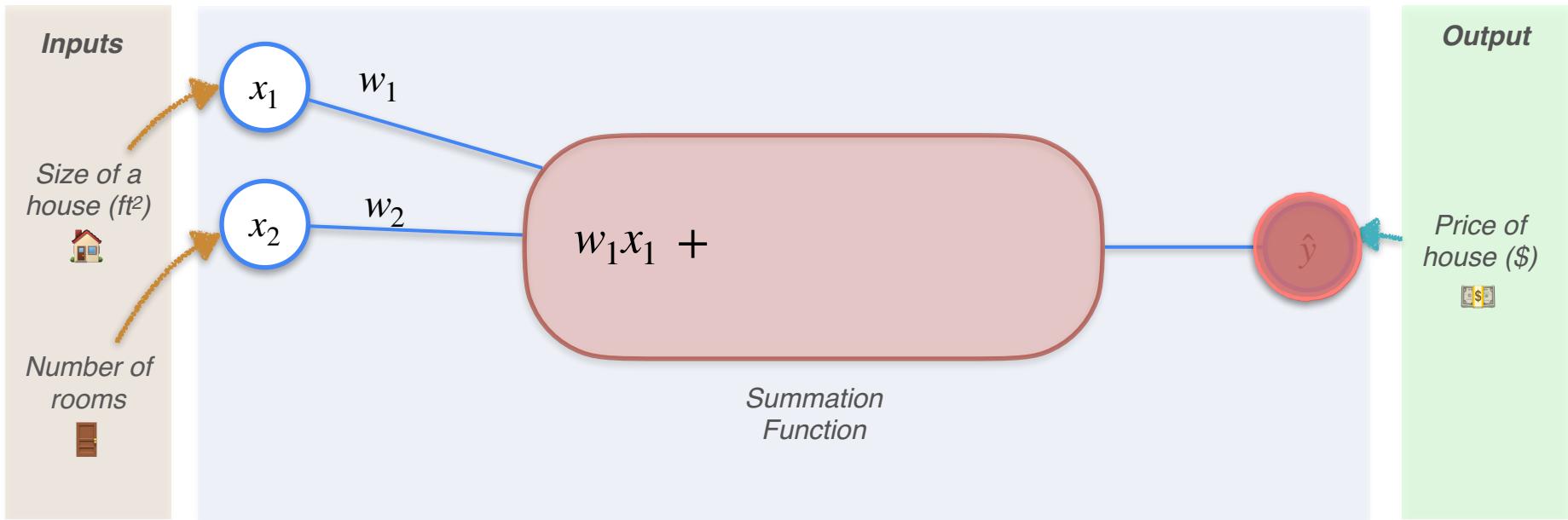
# Regression With a Perceptron

Single Layer Neural Network Perceptron



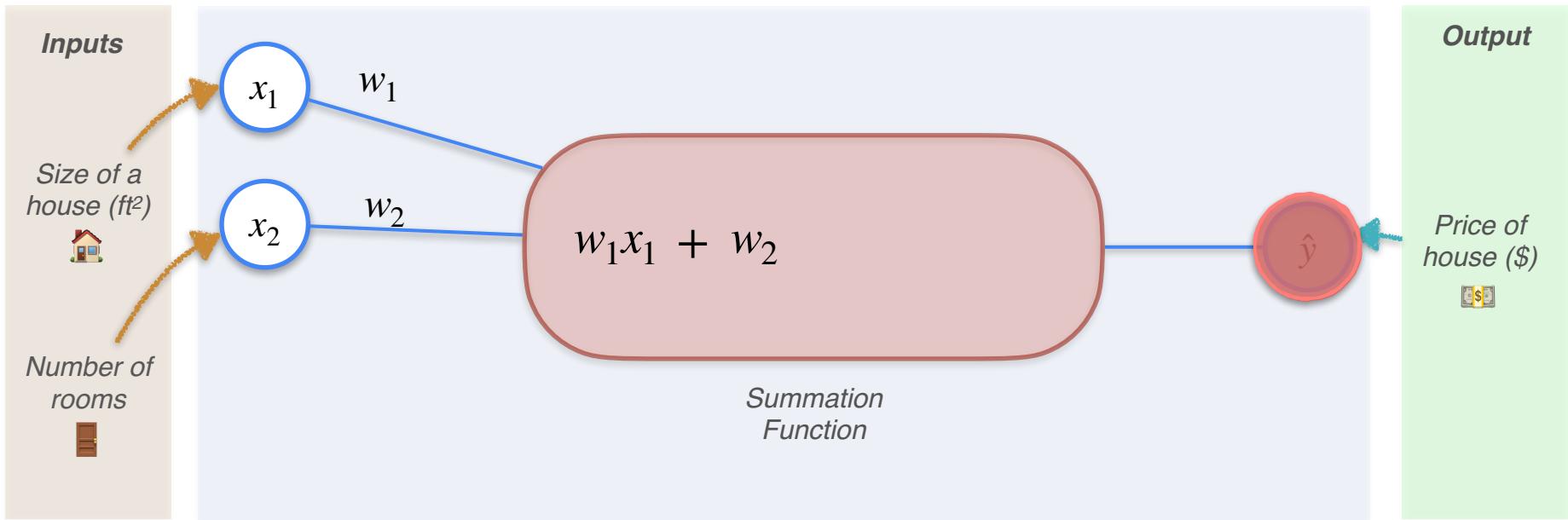
# Regression With a Perceptron

Single Layer Neural Network Perceptron



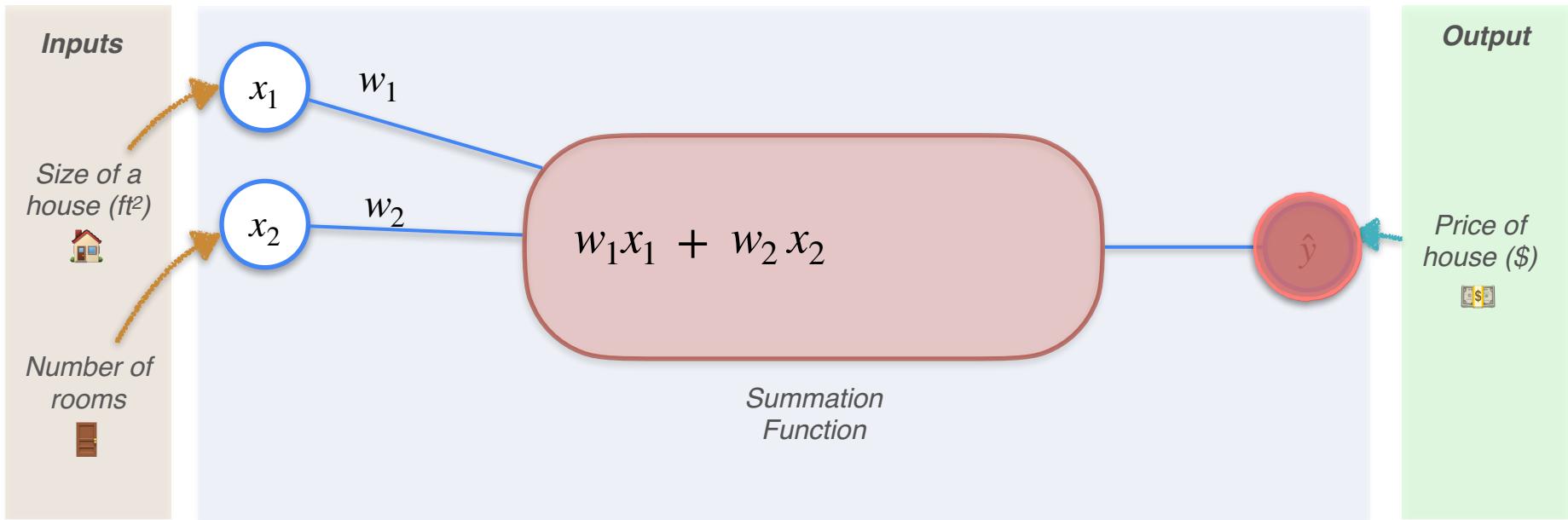
# Regression With a Perceptron

Single Layer Neural Network Perceptron



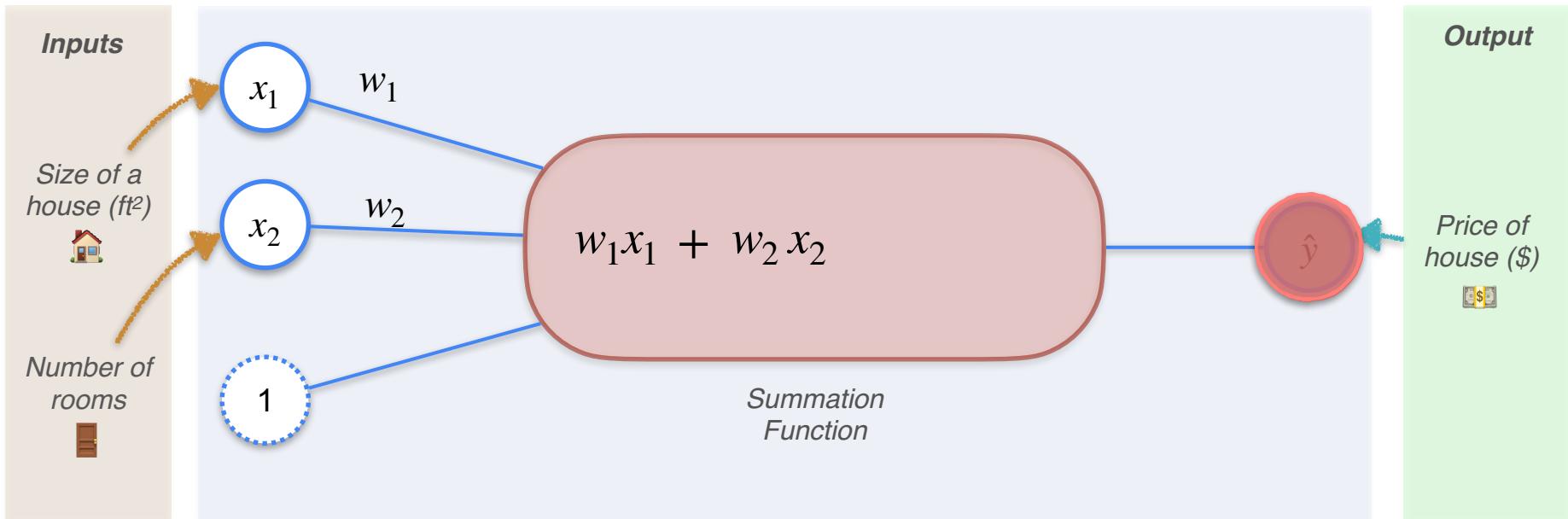
# Regression With a Perceptron

Single Layer Neural Network Perceptron



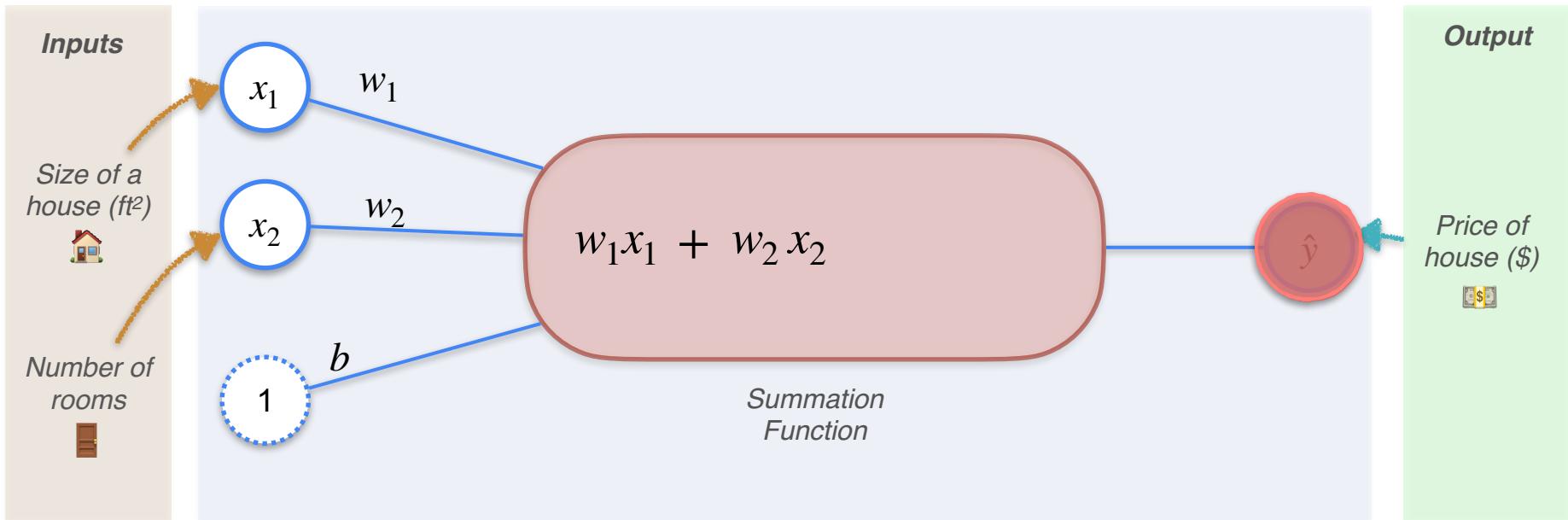
# Regression With a Perceptron

Single Layer Neural Network Perceptron



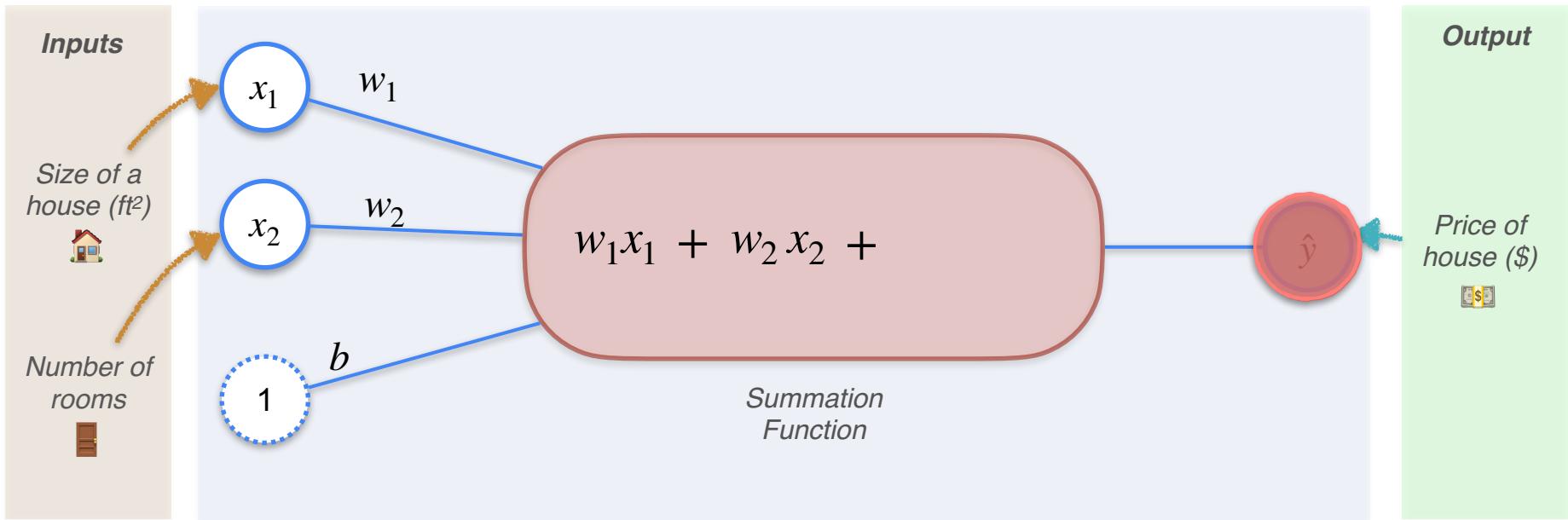
# Regression With a Perceptron

Single Layer Neural Network Perceptron



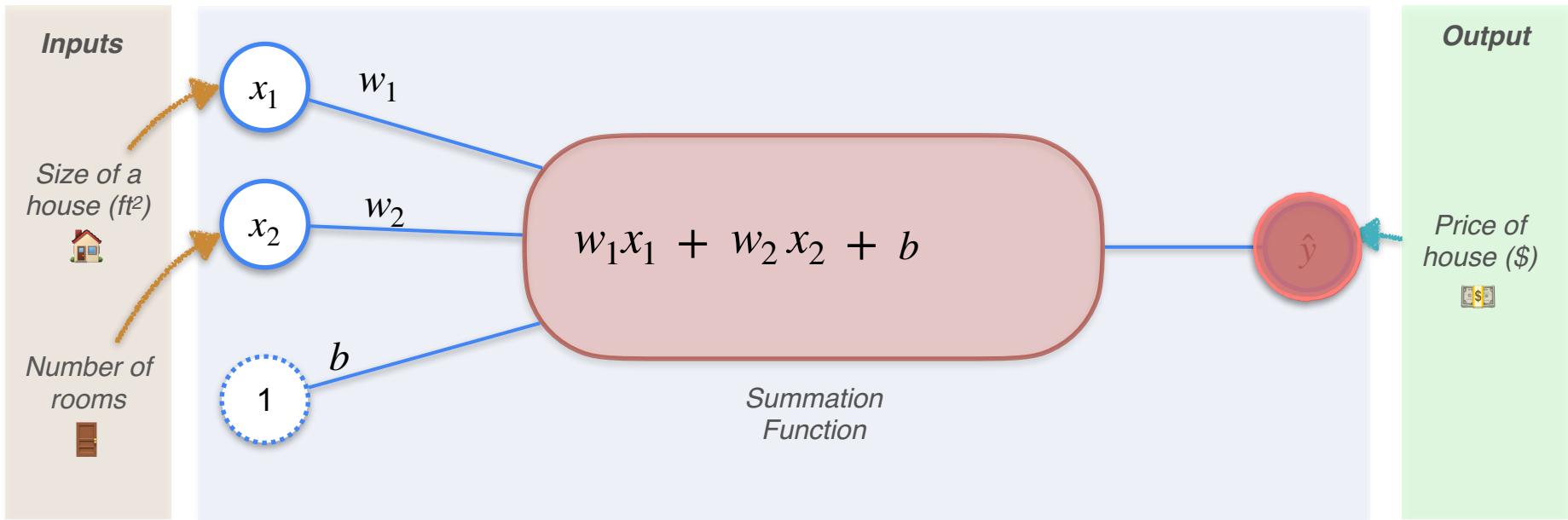
# Regression With a Perceptron

Single Layer Neural Network Perceptron



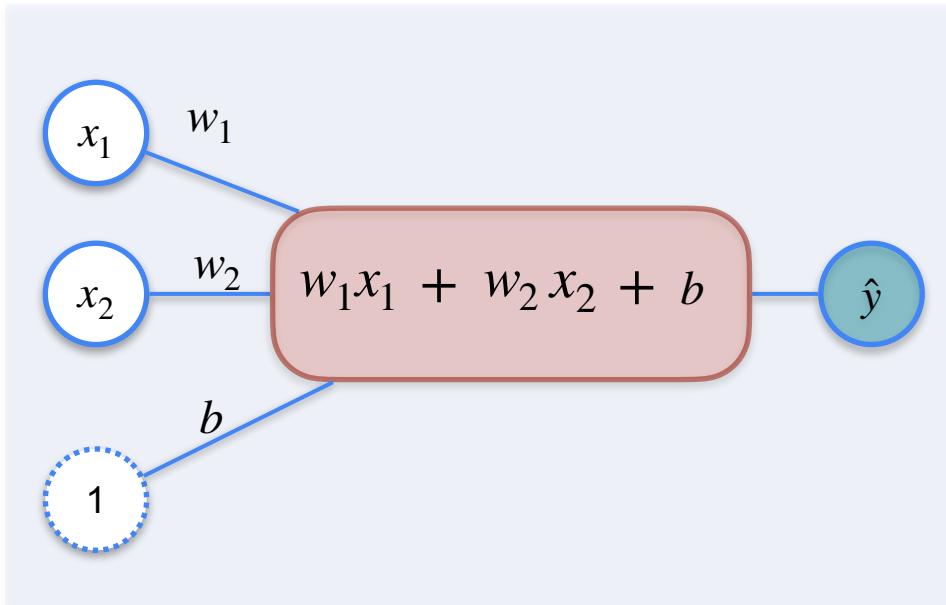
# Regression With a Perceptron

Single Layer Neural Network Perceptron



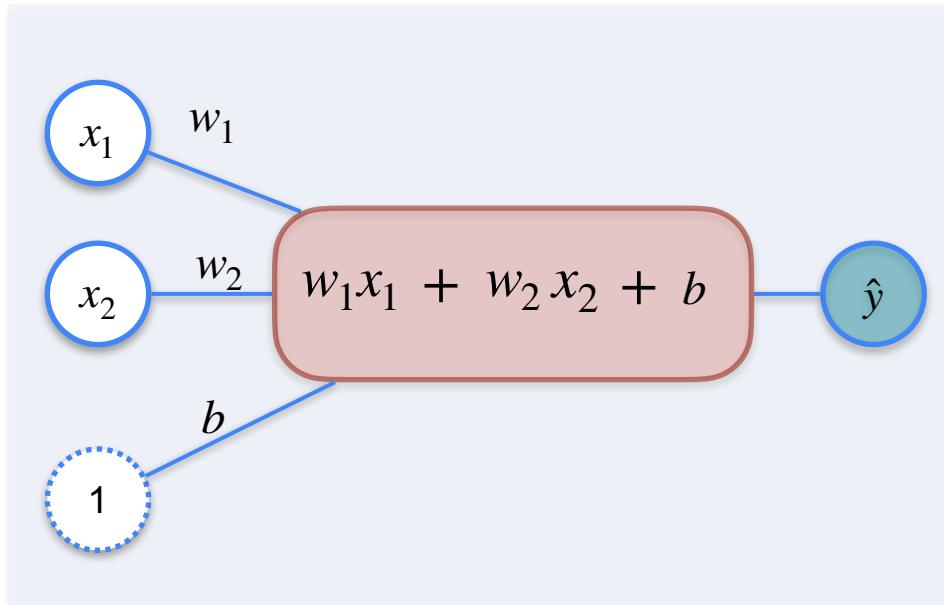
# Regression With a Perceptron

Single Layer Neural Network Perceptron



# Regression With a Perceptron

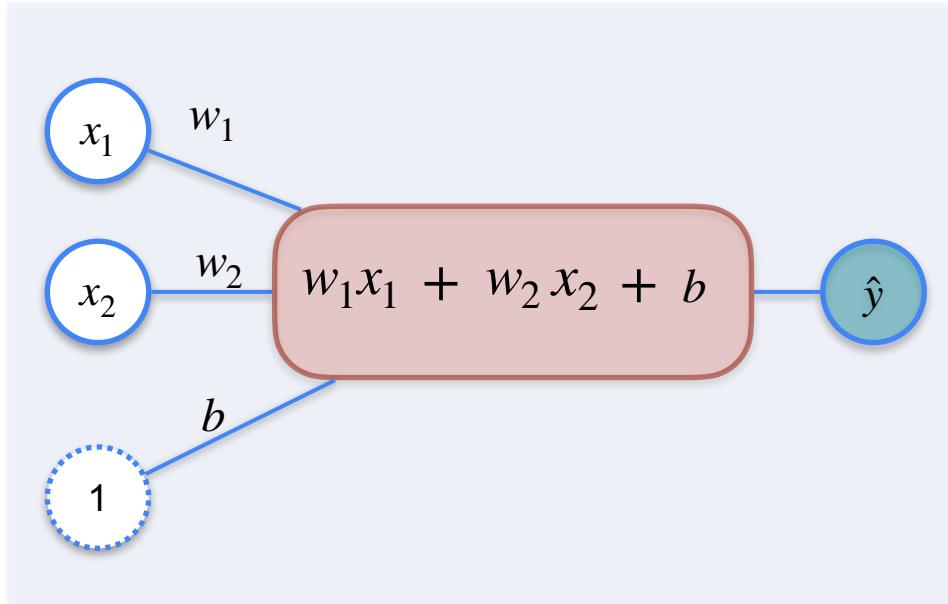
Single Layer Neural Network Perceptron



$\hat{y}$

# Regression With a Perceptron

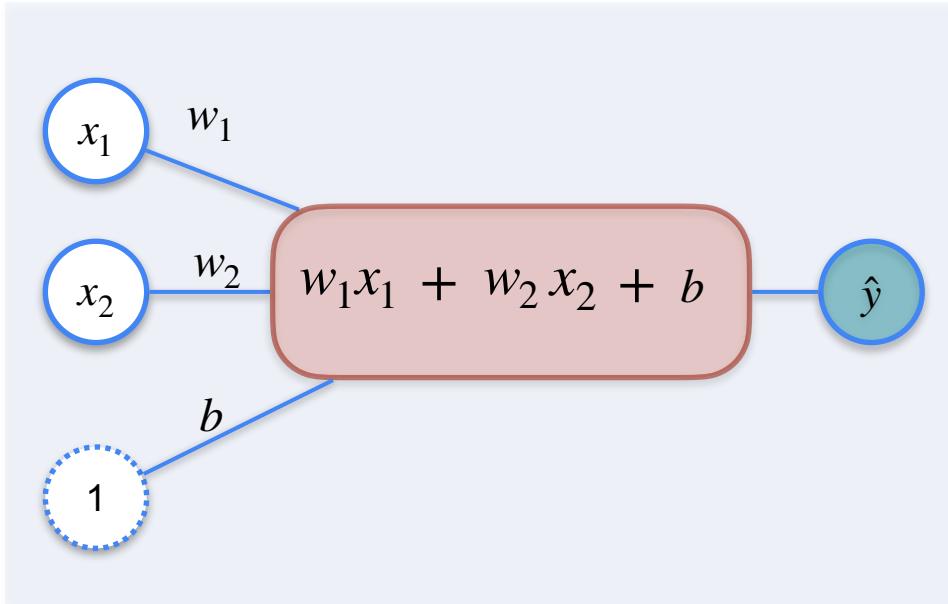
Single Layer Neural Network Perceptron



$$\hat{y} = w_1x_1 + w_2x_2 + b$$

# Regression With a Perceptron

Single Layer Neural Network Perceptron

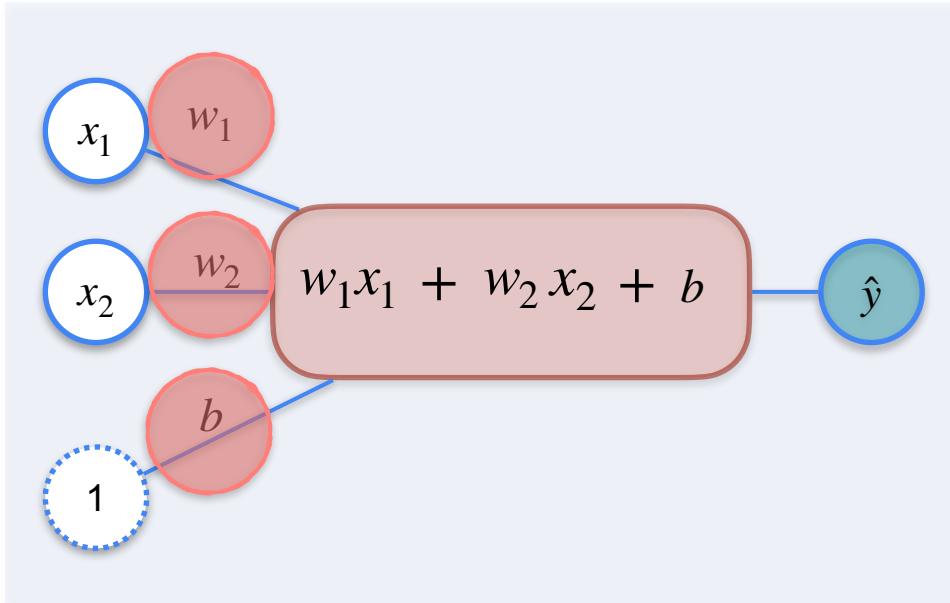


$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Main Goal:**

# Regression With a Perceptron

Single Layer Neural Network Perceptron

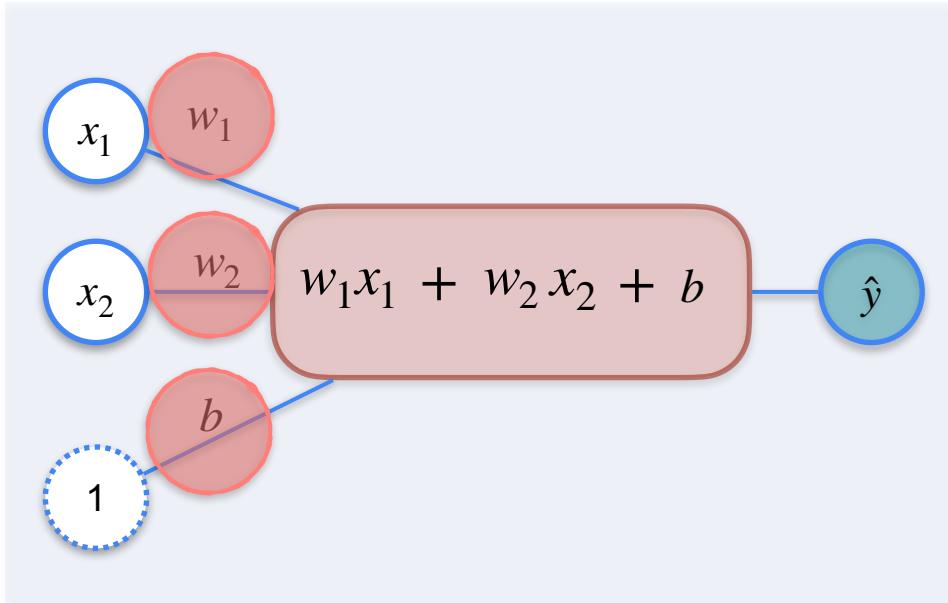


$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Main Goal:

# Regression With a Perceptron

Single Layer Neural Network Perceptron



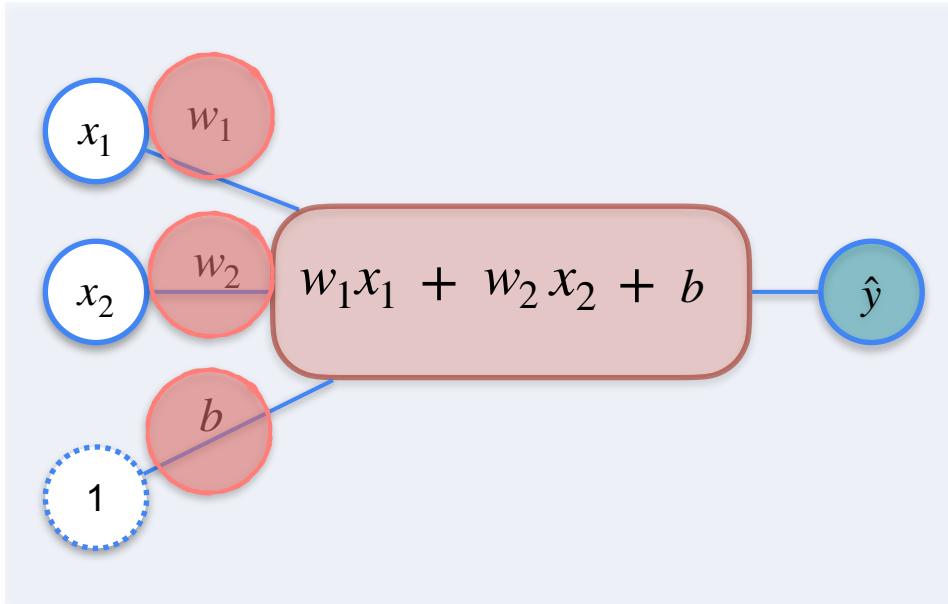
$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Main Goal:**

Find weights and bias that will optimise the predictions.

# Regression With a Perceptron

Single Layer Neural Network Perceptron



$$\hat{y} = w_1x_1 + w_2x_2 + b$$

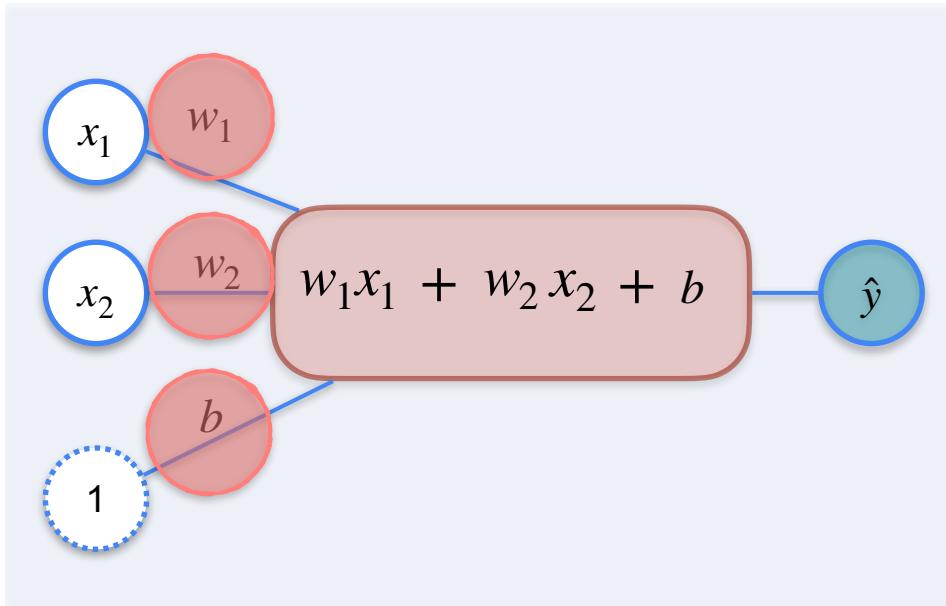
**Main Goal:**

Find weights and bias that will optimise the predictions.

i.e. Reduce the errors in the predictions

# Regression With a Perceptron

Single Layer Neural Network Perceptron



$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Main Goal:**

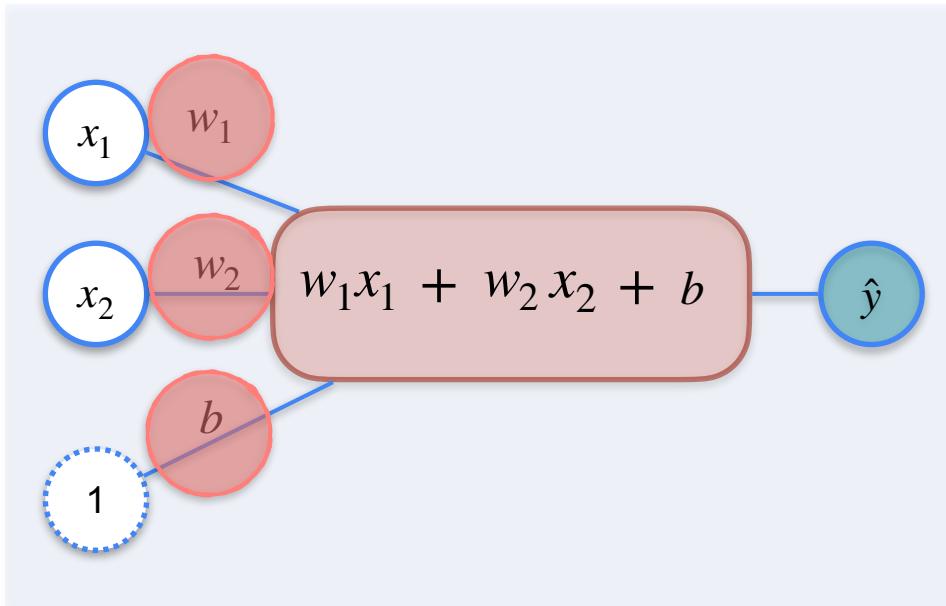
Find weights and bias that will optimise the predictions.

i.e. Reduce the errors in the predictions



# Regression With a Perceptron

Single Layer Neural Network Perceptron



$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Main Goal:**

Find weights and bias that will optimise the predictions.

i.e. Reduce the errors in the predictions



**The  
Loss  
Function**



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Regression with a perceptron: Loss function

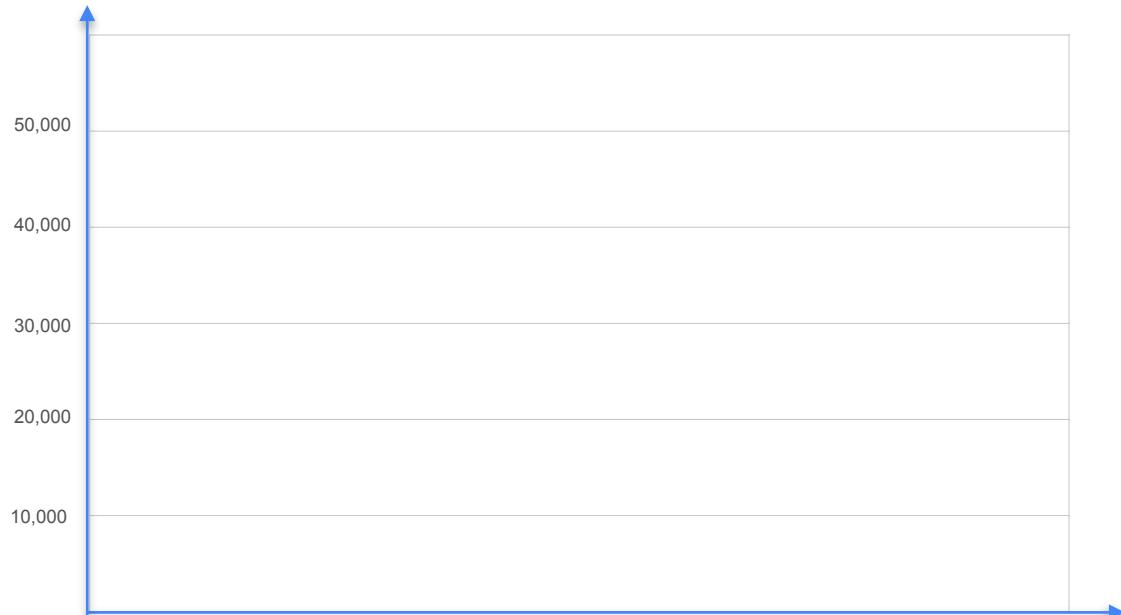
# Mean Squared Error

# Mean Squared Error

	$y$		
	\$20,000		
	\$30,000		
	\$50,000		

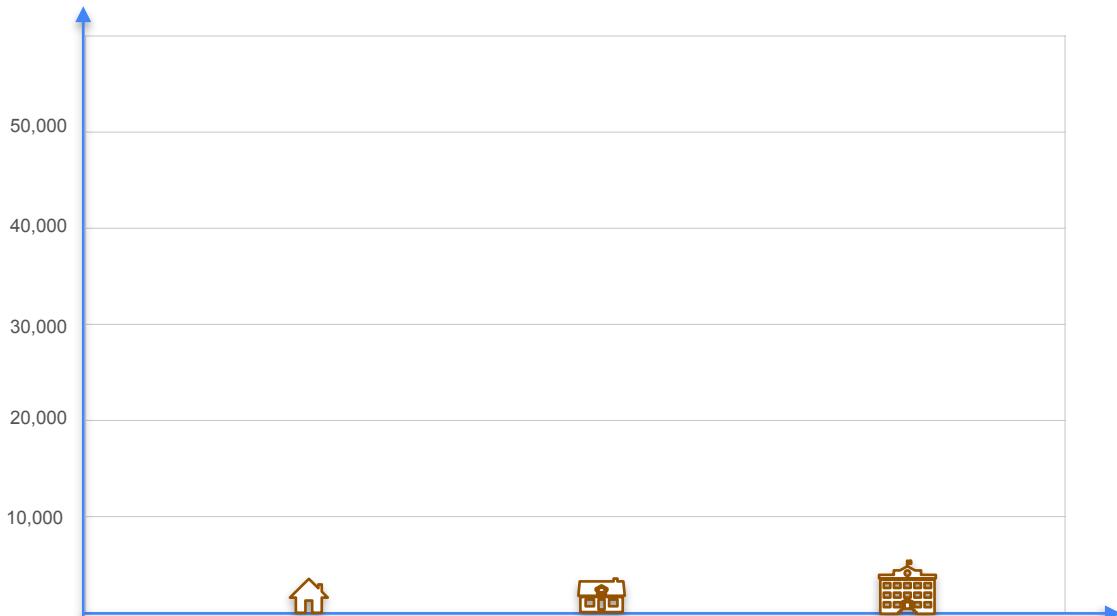
# Mean Squared Error

	$y$		
	\$20,000		
	\$30,000		
	\$50,000		



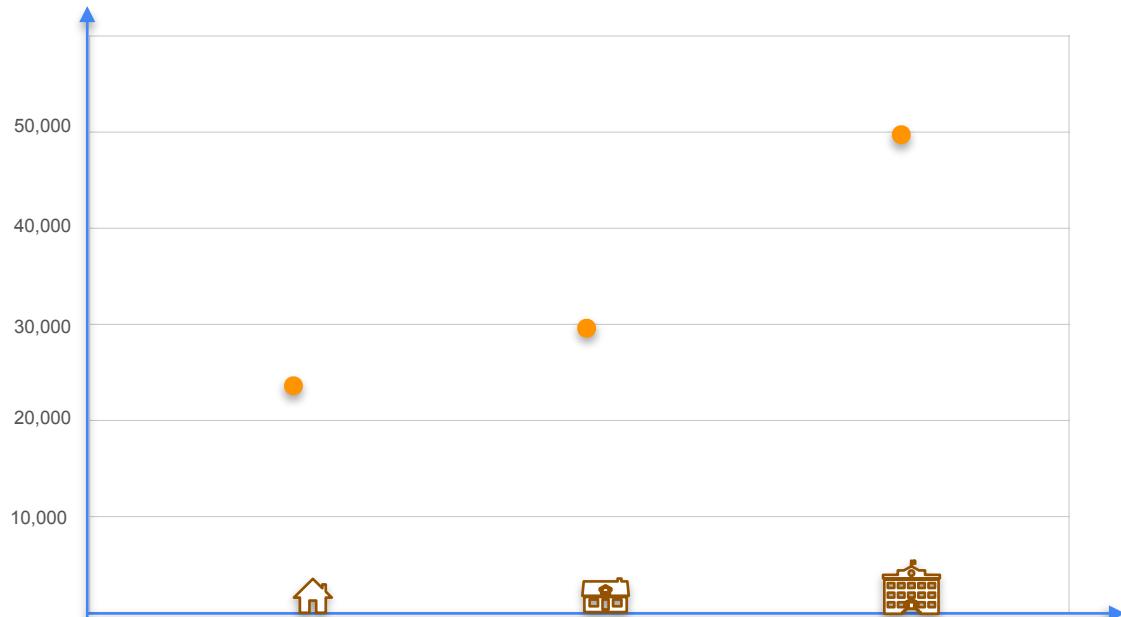
# Mean Squared Error

	$y$		
	\$20,000		
	\$30,000		
	\$50,000		



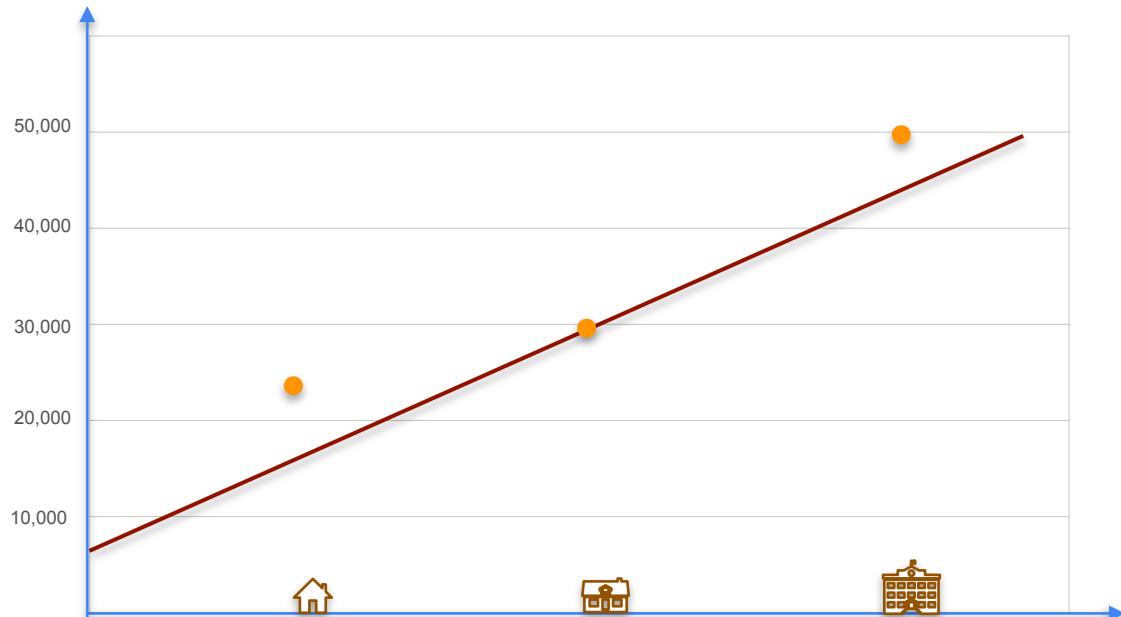
# Mean Squared Error

	$y$		
	\$20,000		
	\$30,000		
	\$50,000		



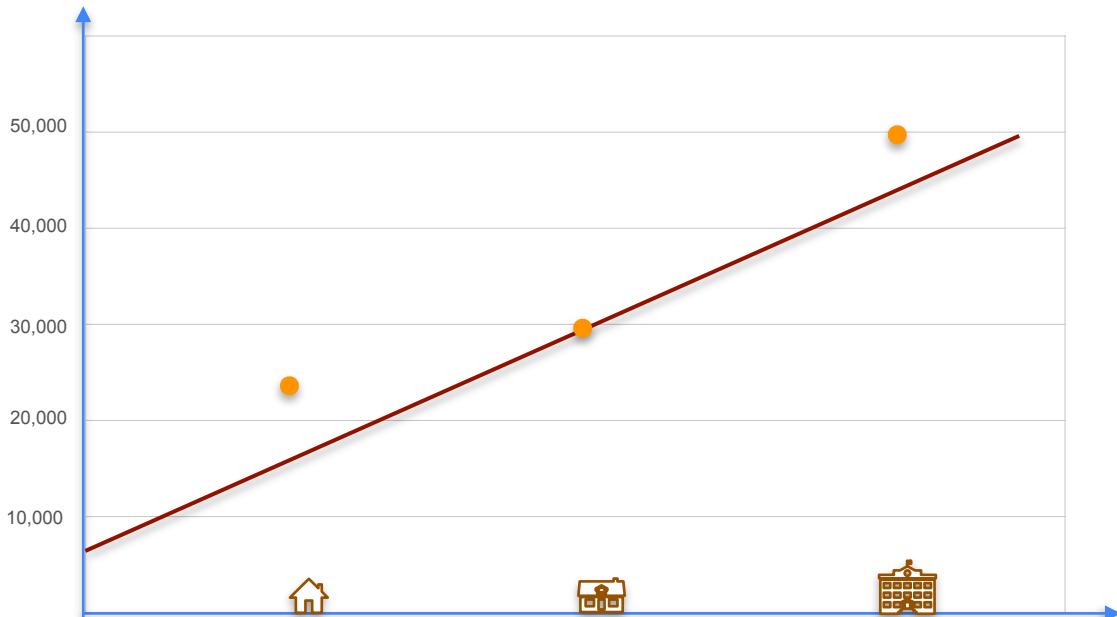
# Mean Squared Error

	$y$		
	\$20,000		
	\$30,000		
	\$50,000		



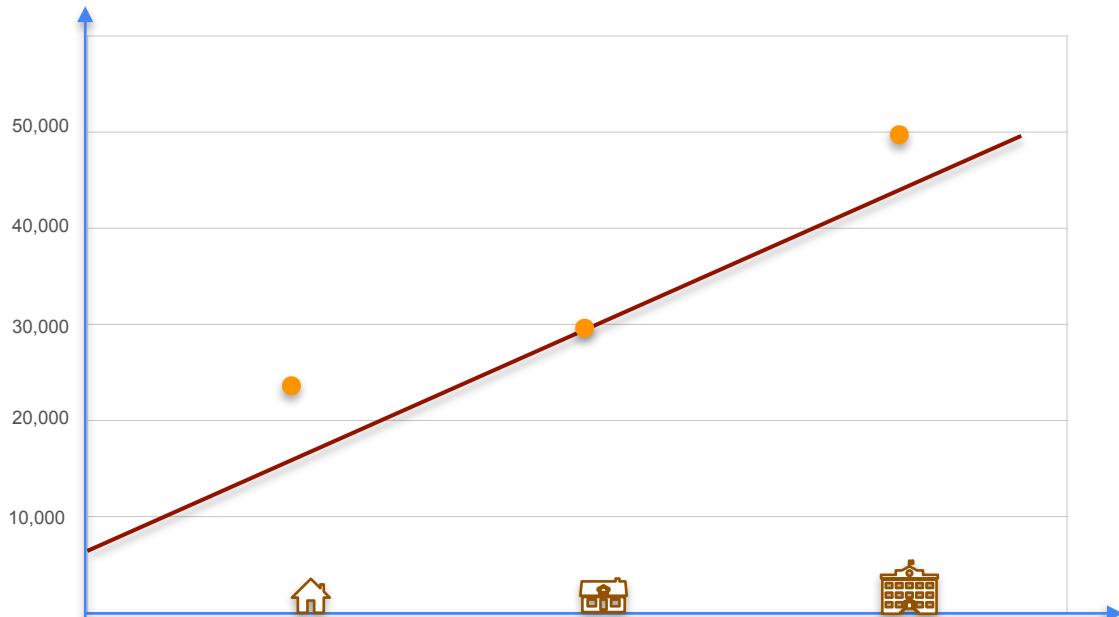
# Mean Squared Error

	$y$	$\hat{y}$	
	\$20,000	\$15,000	
	\$30,000	\$30,000	
	\$50,000	\$45,000	



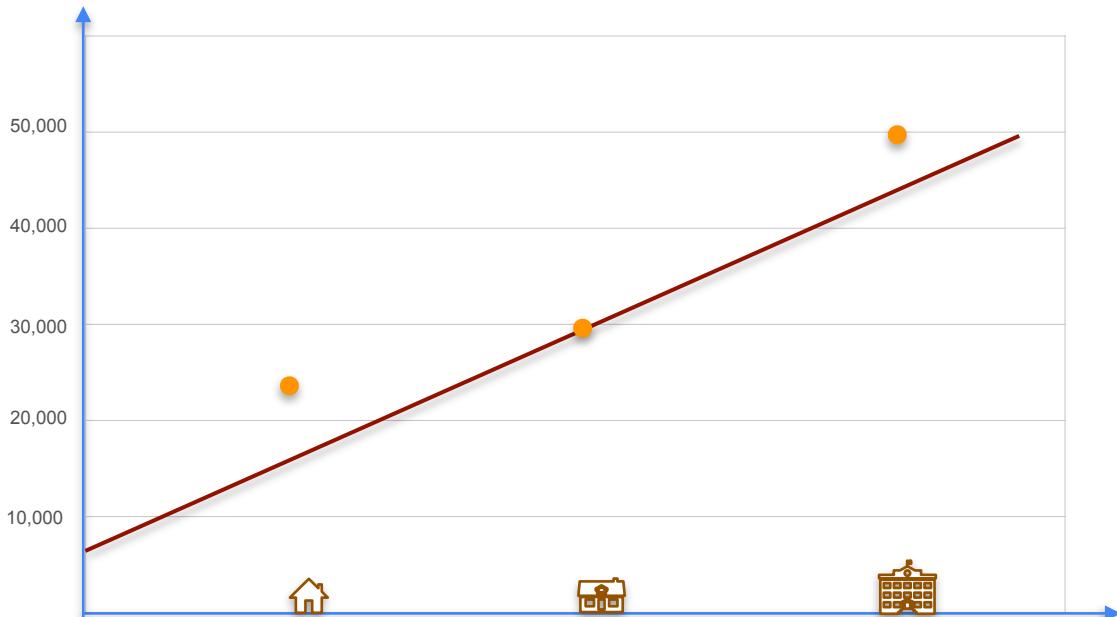
# Mean Squared Error

	$y$	$\hat{y}$	$y - \hat{y}$
	\$20,000	\$15,000	
	\$30,000	\$30,000	
	\$50,000	\$45,000	



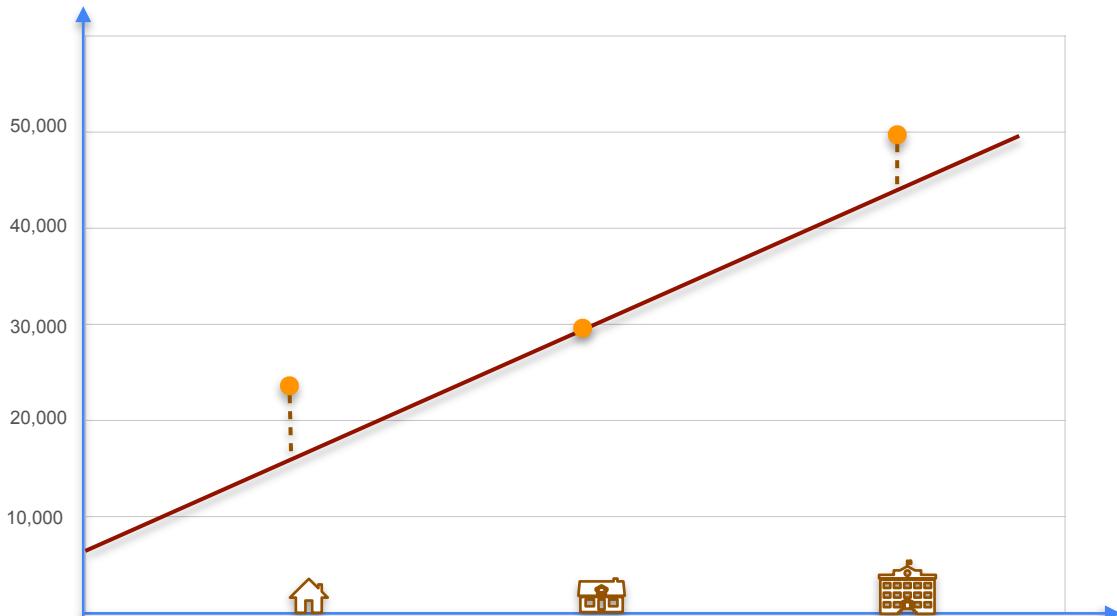
# Mean Squared Error

	$y$	$\hat{y}$	$y - \hat{y}$
	\$20,000	\$15,000	<b>Error</b>
	\$30,000	\$30,000	<b>Error</b>
	\$50,000	\$45,000	<b>Error</b>



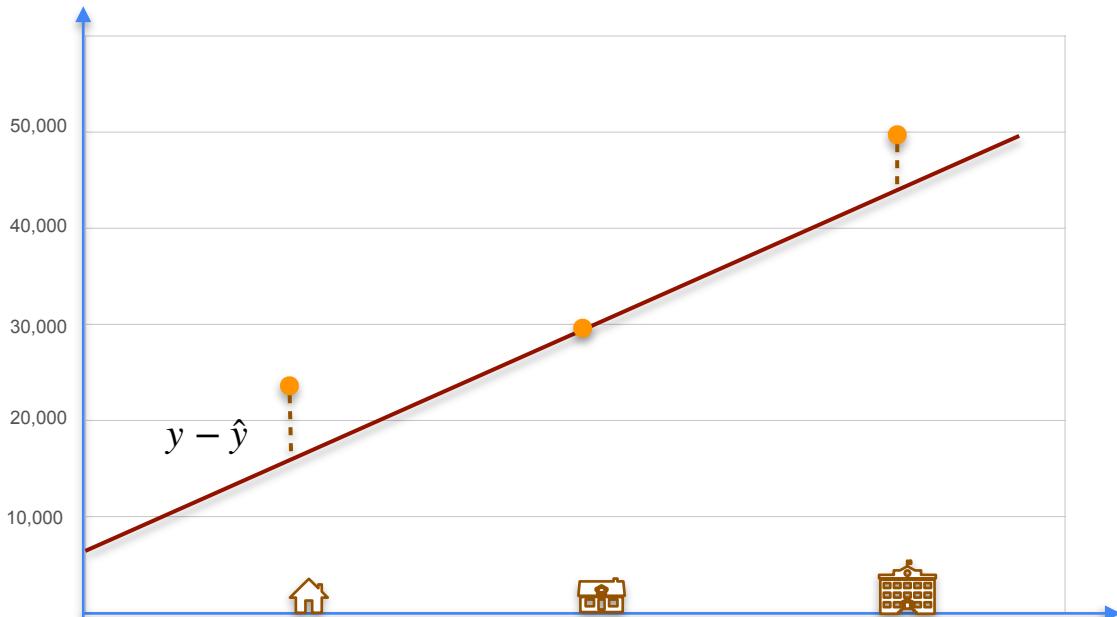
# Mean Squared Error

	$y$	$\hat{y}$	$y - \hat{y}$
	\$20,000	\$15,000	<b>Error</b>
	\$30,000	\$30,000	<b>Error</b>
	\$50,000	\$45,000	<b>Error</b>



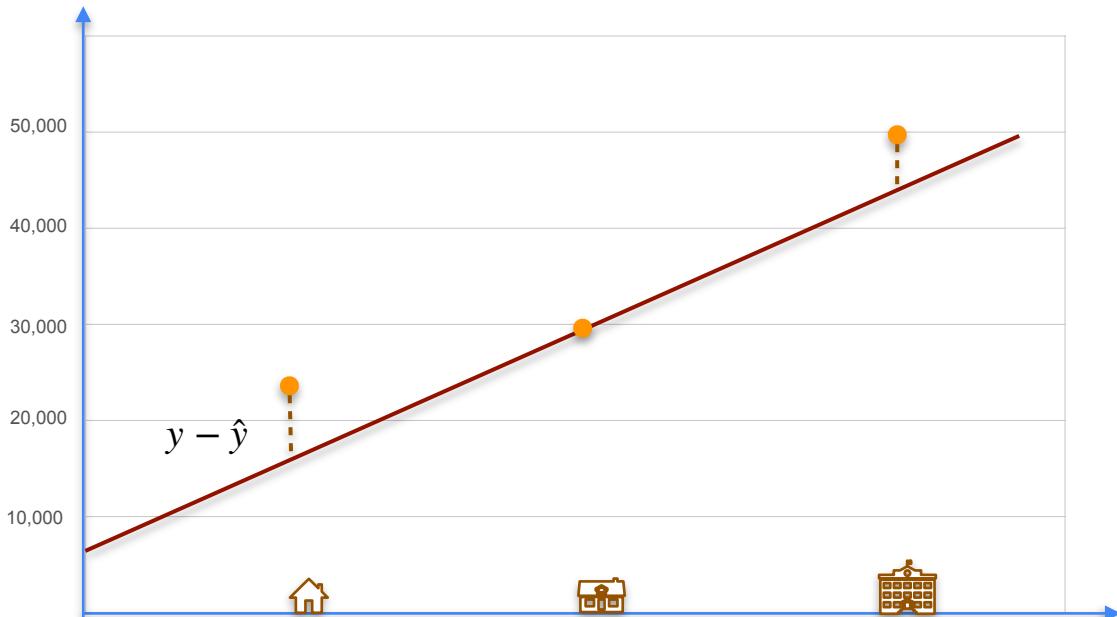
# Mean Squared Error

	$y$	$\hat{y}$	$y - \hat{y}$
	\$20,000	\$15,000	<b>Error</b>
	\$30,000	\$30,000	<b>Error</b>
	\$50,000	\$45,000	<b>Error</b>



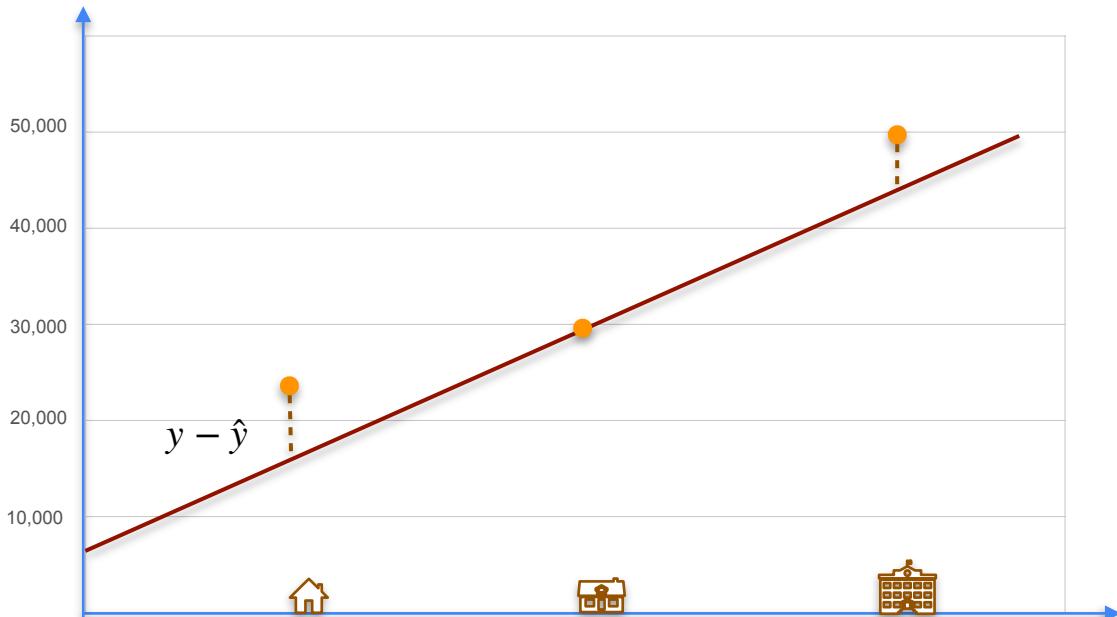
# Mean Squared Error

	$y$	$\hat{y}$	$(y - \hat{y})^2$
	\$20,000	\$15,000	<b>Error</b>
	\$30,000	\$30,000	<b>Error</b>
	\$50,000	\$45,000	<b>Error</b>



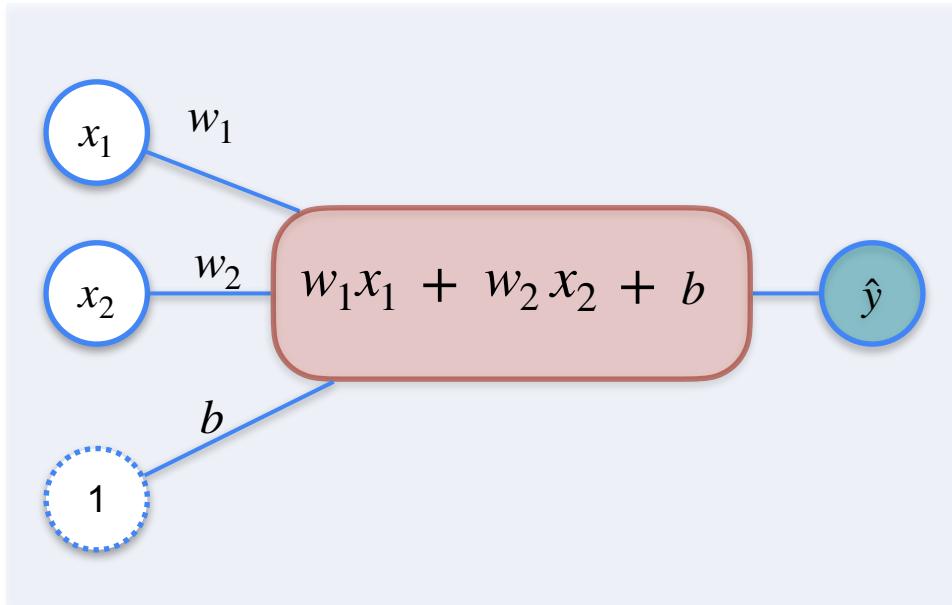
# Mean Squared Error

	$y$	$\hat{y}$	$\frac{1}{2}(y - \hat{y})^2$
	\$20,000	\$15,000	<b>Error</b>
	\$30,000	\$30,000	<b>Error</b>
	\$50,000	\$45,000	<b>Error</b>



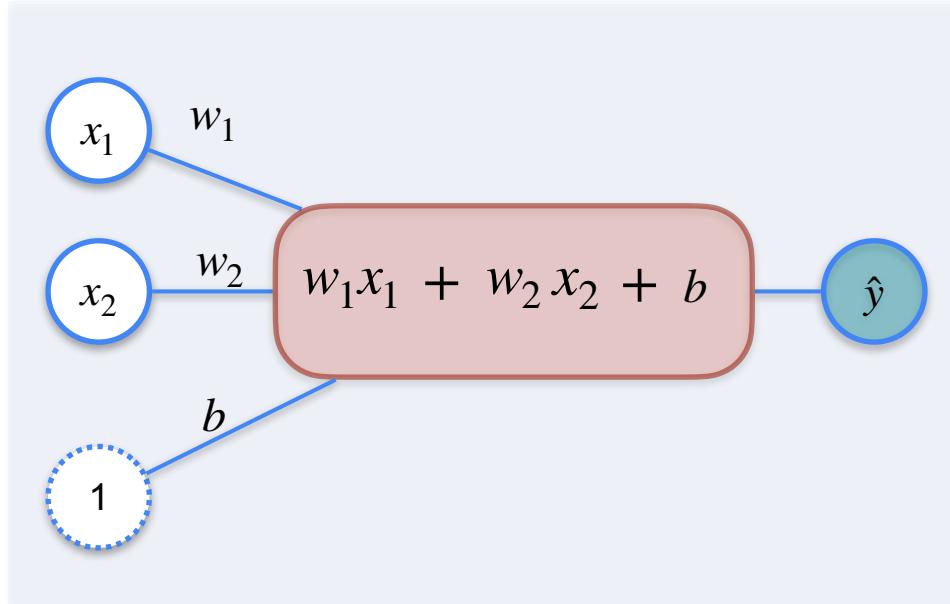
# Regression With a Perceptron

Single Layer Neural Network Perceptron



# Regression With a Perceptron

Single Layer Neural Network Perceptron

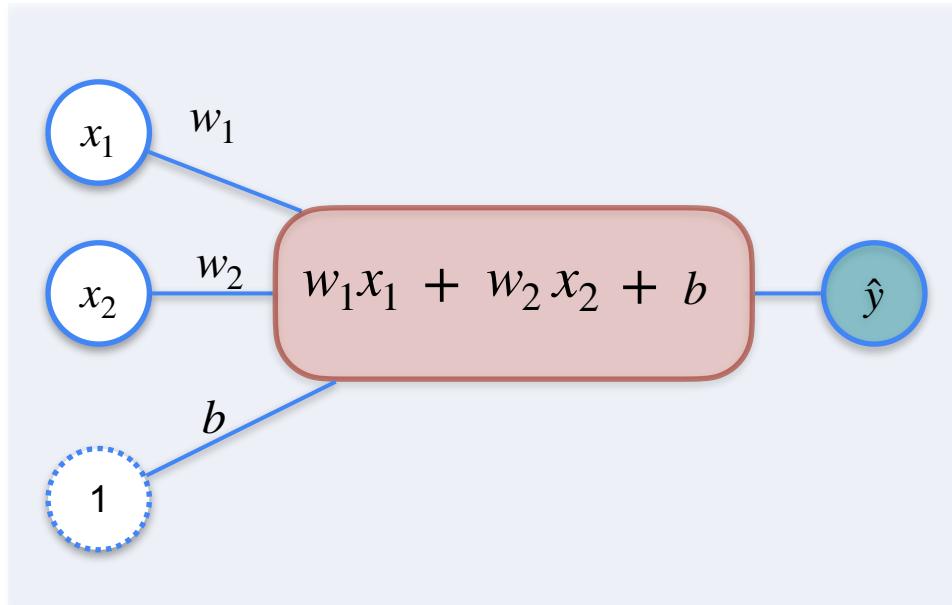


**Prediction Function:**

$$\hat{y}$$

# Regression With a Perceptron

Single Layer Neural Network Perceptron

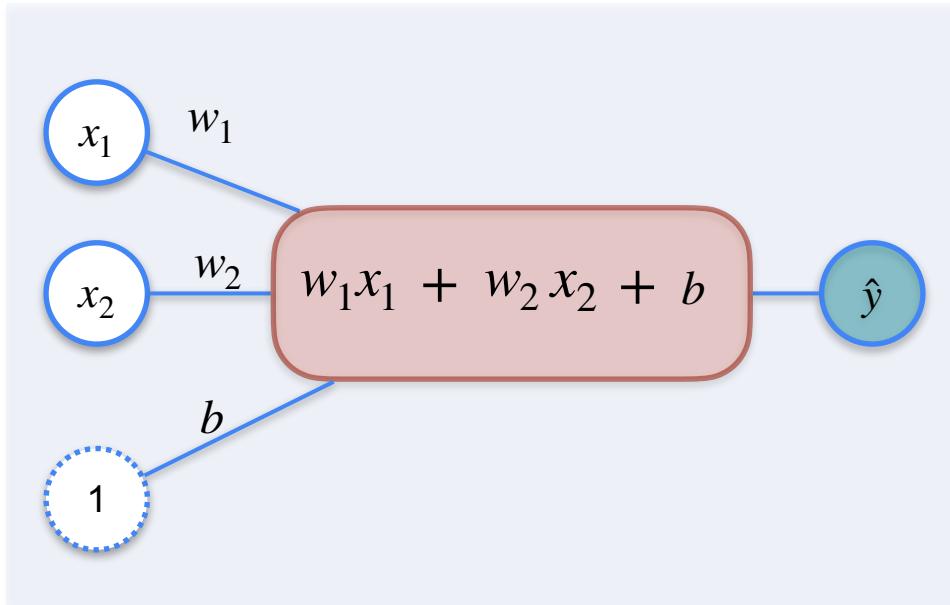


**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

# Regression With a Perceptron

Single Layer Neural Network Perceptron



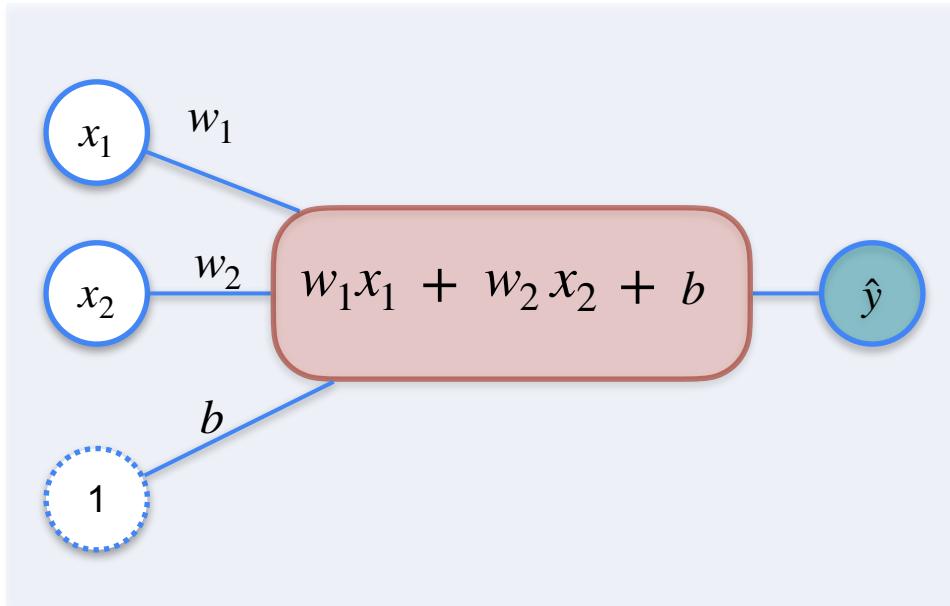
**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

# Regression With a Perceptron

Single Layer Neural Network Perceptron



**Prediction Function:**

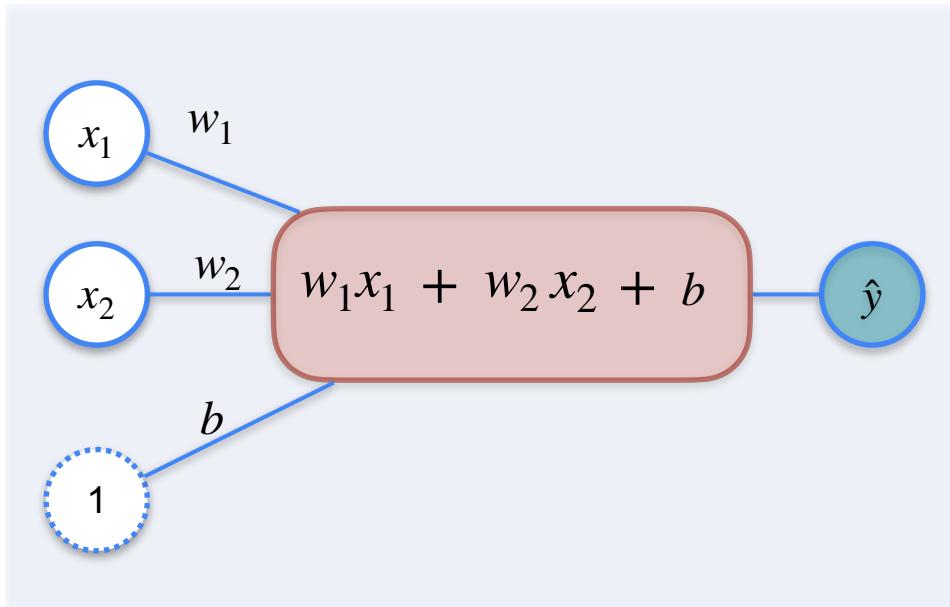
$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$= \frac{1}{2}(y - \hat{y})^2$$

# Regression With a Perceptron

Single Layer Neural Network Perceptron



**Prediction Function:**

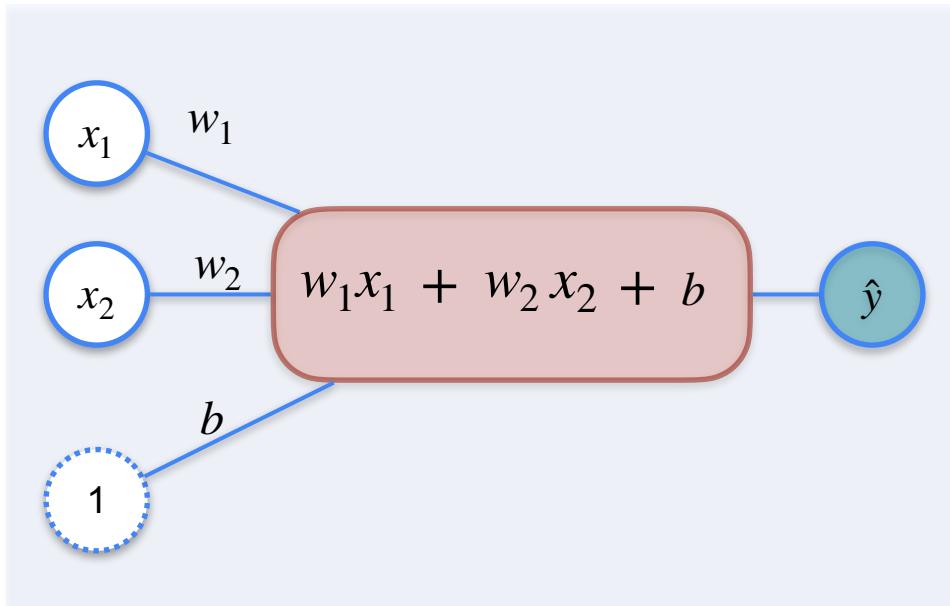
$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

# Regression With a Perceptron

Single Layer Neural Network Perceptron



**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

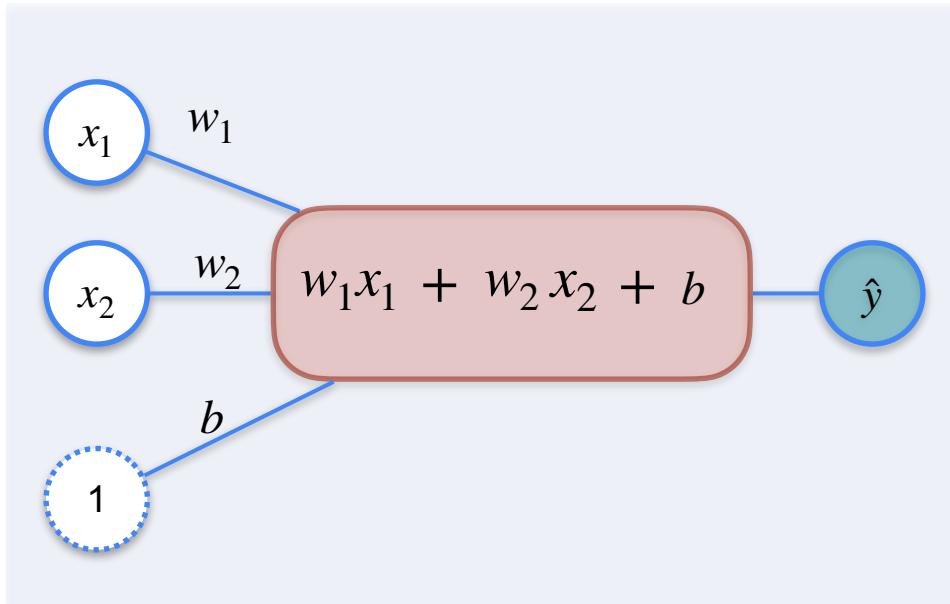
**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

# Regression With a Perceptron

Single Layer Neural Network Perceptron



**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1$ ,  $w_2$ ,  $b$  that give  $\hat{y}$  with the least error



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Regression with a perceptron: Gradient Descent

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1$ ,  $w_2$ ,  $b$  that give  $\hat{y}$  with the least error

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1$ ,  $w_2$ ,  $b$  that give  $\hat{y}$  with the least error

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1$ ,  $w_2$ ,  $b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

$$w_1, w_2, b$$

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

$$w_1, w_2, b$$

*You need gradient descent*

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

$$w_1, w_2, b$$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

$$w_1, w_2, b$$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

$$w_1, w_2, b$$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

# Regression With a Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

**To find optimal values for:**

$$w_1, w_2, b$$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

# Regression With a Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

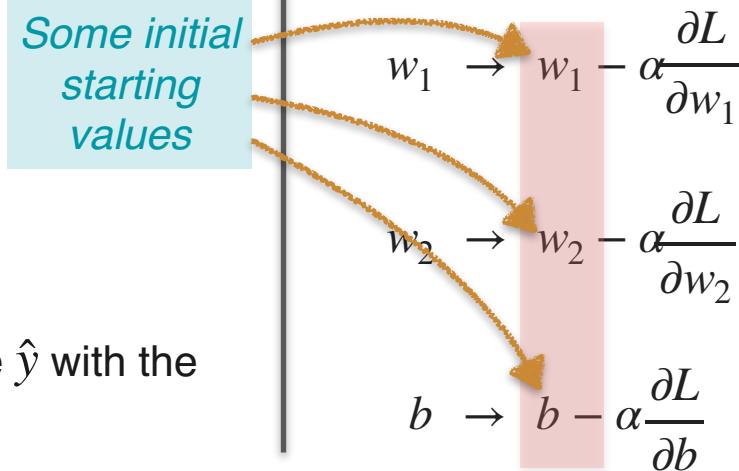
Main Goal:

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

To find optimal values for:

$$w_1, w_2, b$$

*You need gradient descent*



# Regression With a Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

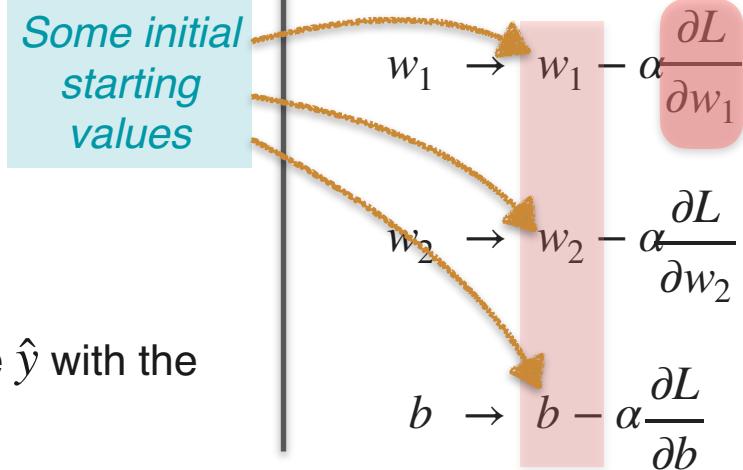
Main Goal:

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

To find optimal values for:

$$w_1, w_2, b$$

*You need gradient descent*



# Regression With a Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Main Goal:

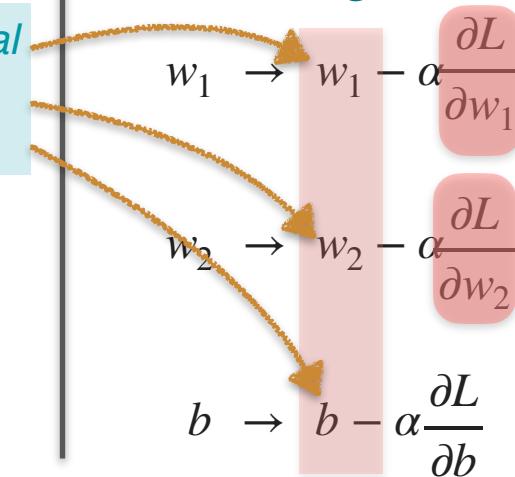
Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

To find optimal values for:

$$w_1, w_2, b$$

*You need gradient descent*

Some initial starting values



# Regression With a Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Main Goal:

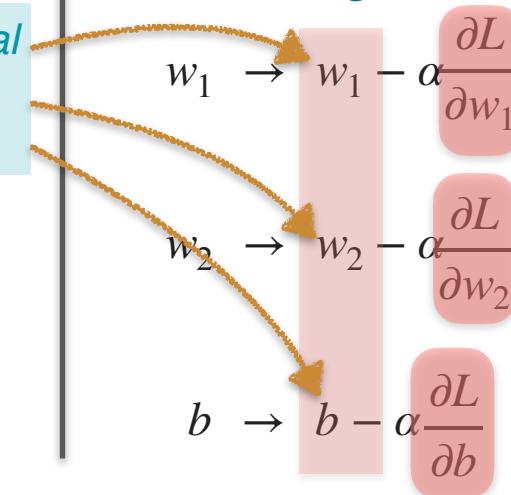
Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

To find optimal values for:

$$w_1, w_2, b$$

*You need gradient descent*

Some initial starting values



# Regression With a Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

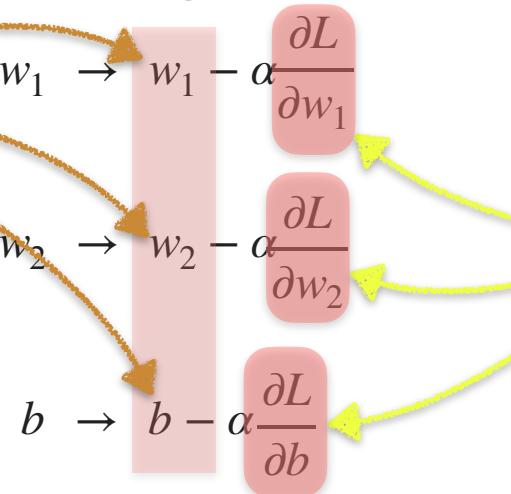
Main Goal:

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

*Some initial starting values*

To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*



**SUB-TASK**

Find the following partial derivatives

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

$$\frac{\partial L}{\partial b}$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} =$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} =$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} =$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} =$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \partial \hat{y}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} =$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} =$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} =$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

Prediction Function:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Loss Function:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

Using chain rule:

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= (y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= (y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	$= x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	$= x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	$= x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	$= x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	$= x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	$= x_2$

# Regression With Perceptron

**Prediction Function:**

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

**Loss Function:**

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$= -(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial b}$	$= 1$
$\frac{\partial \hat{y}}{\partial w_1}$	$= x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	$= x_2$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y})$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$
$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$
$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$
$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$
$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y})$$

$$= -(y - \hat{y}) x_1$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$
$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$
$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$
$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$
$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$
$$= -(y - \hat{y})$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y})$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y})$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y}) x_2$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y}) x_2$$

# Regression With Perceptron

$$\frac{\partial L}{\partial \hat{y}} = -(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b} = 1$$

$$\frac{\partial \hat{y}}{\partial w_1} = x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$= -(y - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y}) x_2$$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

*Perform Gradient Descent*

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

$$w_1 = w_1 - \alpha \frac{\partial L}{\partial w_1}$$

*Perform Gradient Descent*

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

$$w_1 = w_1 - \alpha$$

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

*Perform Gradient Descent*

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

*Perform Gradient Descent*

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

*Perform Gradient Descent*

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 = w_2 - \alpha \frac{\partial L}{\partial w_2}$$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

*Perform Gradient Descent*

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 = w_2 - \alpha$$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

***Perform Gradient Descent***

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 = w_2 - \alpha(-x_2(y - \hat{y}))$$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

***Perform Gradient Descent***

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 = w_2 - \alpha(-x_2(y - \hat{y}))$$

$$b = b - \alpha \frac{\partial L}{\partial b}$$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

***Perform Gradient Descent***

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 = w_2 - \alpha(-x_2(y - \hat{y}))$$

$$b = b - \alpha$$

# Regression With a Perceptron

**Main Goal:**

Find  $w_1$  ,  $w_2$  ,  $b$  that give  $\hat{y}$  with the least error

**ie. optimal values for:**

$w_1$  ,  $w_2$  ,  $b$

***Perform Gradient Descent***

$$w_1 = w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 = w_2 - \alpha(-x_2(y - \hat{y}))$$

$$b = b - \alpha(-(y - \hat{y}))$$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Classification with a perceptron

# Classification Problem Motivation

# Classification Problem Motivation



# Classification Problem Motivation

<i>Sentence</i>			

# Classification Problem Motivation

<i>Sentence</i>			
<i>Aack aack aack!</i>			

# Classification Problem Motivation

<i>Sentence</i>			
<i>Aack aack aack!</i>			
<i>Beep beep!</i>			

# Classification Problem Motivation

<i>Sentence</i>			
<i>Aack aack aack!</i>			
<i>Beep beep!</i>			
<i>Aack beep beep beep!</i>			

# Classification Problem Motivation

<i>Sentence</i>			
<i>Aack aack aack!</i>			
<i>Beep beep!</i>			
<i>Aack beep beep beep!</i>			
<i>Aack beep aack!</i>			

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			

*Beep beep!*

*Aack beep beep beep!*

*Aack beep aack!*

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			

*Beep beep!*

*Aack beep beep beep!*

*Aack beep aack!*

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊

*Beep beep!*

<i>Aack beep beep beep!</i>			
-----------------------------	--	--	--

*Aack beep aack!*

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			😔
<i>Aack beep beep beep!</i>			
<i>Aack beep aack!</i>			

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			
<i>Aack beep aack!</i>			

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			😔
<i>Aack beep aack!</i>			

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			😊

# Classification Problem Motivation

Sentence			Mood
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			<i>Happy</i> 😊

# Classification Problem Motivation

<i>Sentence</i>	<i>Aack</i>	<i>Beep</i>	<i>Mood</i>
<i>Aack aack aack!</i>			<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			<i>Happy</i> 😊

# Classification Problem Motivation

<i>Sentence</i>	<i>Aack</i>	<i>Beep</i>	<i>Mood</i>
<i>Aack aack aack!</i>	3		<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			<i>Happy</i> 😊

# Classification Problem Motivation

<i>Sentence</i>	<i>Aack</i>	<i>Beep</i>	<i>Mood</i>
<i>Aack aack aack!</i>	3	0	<i>Happy</i> 😊
<i>Beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			<i>Happy</i> 😊

# Classification Problem Motivation

<i>Sentence</i>	<i>Aack</i>	<i>Beep</i>	<i>Mood</i>
<i>Aack aack aack!</i>	3	0	<i>Happy</i> 😊
<i>Beep beep!</i>	0		<i>Sad</i> 😞
<i>Aack beep beep beep!</i>			<i>Sad</i> 😞
<i>Aack beep aack!</i>			<i>Happy</i> 😊

# Classification Problem Motivation

Sentence	Aack	Beep	Mood
<i>Aack aack aack!</i>	3	0	Happy 😊
<i>Beep beep!</i>	0	2	Sad 😞
<i>Aack beep beep beep!</i>			Sad 😞
<i>Aack beep aack!</i>			Happy 😊

# Classification Problem Motivation

Sentence	Aack	Beep	Mood
<i>Aack aack aack!</i>	3	0	Happy 😊
<i>Beep beep!</i>	0	2	Sad 😞
<i>Aack beep beep beep!</i>	1		Sad 😞
<i>Aack beep aack!</i>			Happy 😊

# Classification Problem Motivation

Sentence	Aack	Beep	Mood
<i>Aack aack aack!</i>	3	0	Happy 😊
<i>Beep beep!</i>	0	2	Sad 😞
<i>Aack beep beep beep!</i>	1	3	Sad 😞
<i>Aack beep aack!</i>			Happy 😊

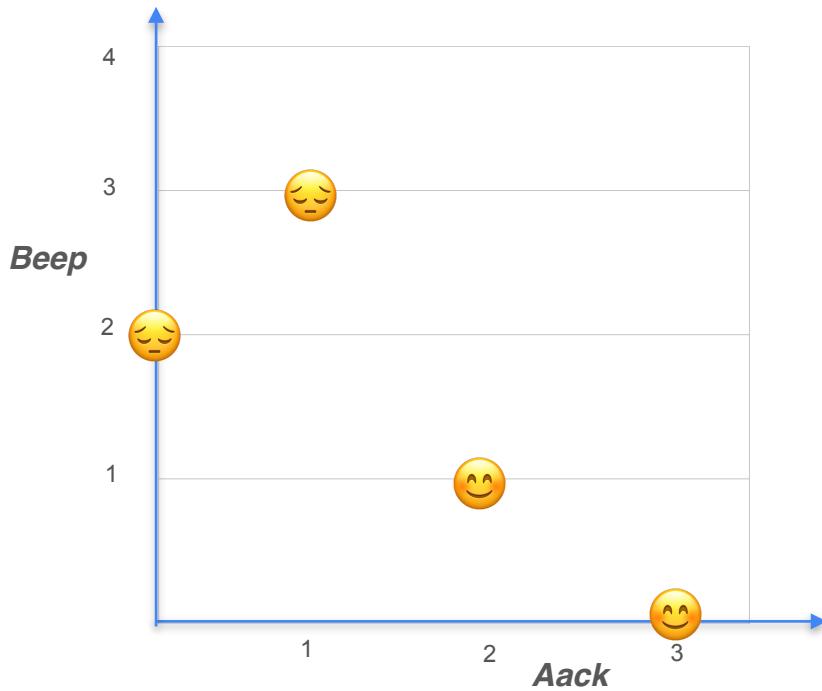
# Classification Problem Motivation

Sentence	Aack	Beep	Mood
<i>Aack aack aack!</i>	3	0	Happy 😊
<i>Beep beep!</i>	0	2	Sad 😞
<i>Aack beep beep beep!</i>	1	3	Sad 😞
<i>Aack beep aack!</i>	2		Happy 😊

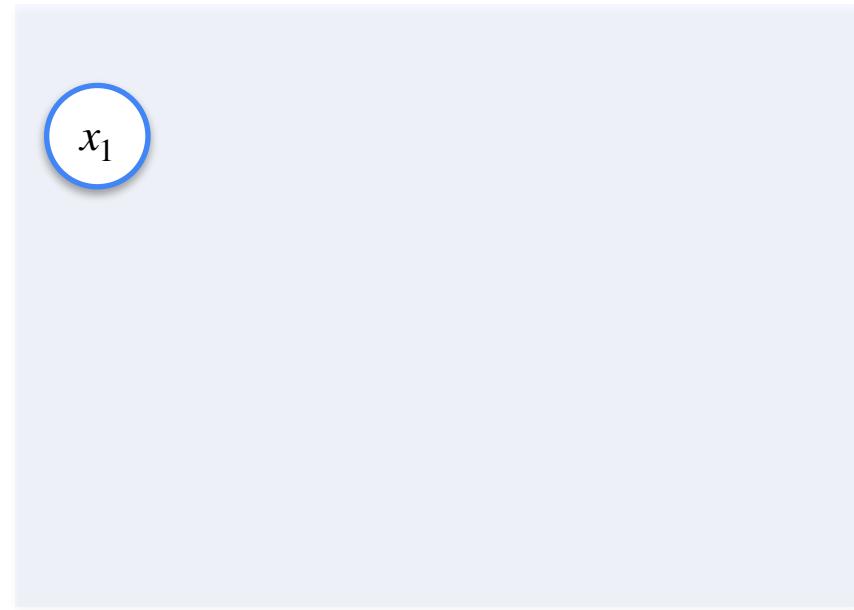
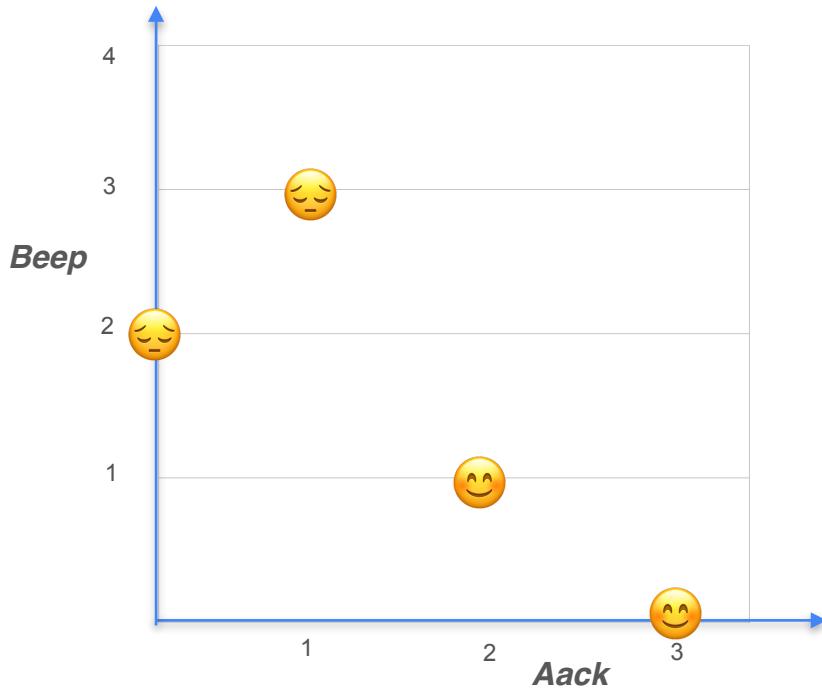
# Classification Problem Motivation

<i>Sentence</i>	<i>Aack</i>	<i>Beep</i>	<i>Mood</i>
<i>Aack aack aack!</i>	3	0	<i>Happy</i> 😊
<i>Beep beep!</i>	0	2	<i>Sad</i> 😞
<i>Aack beep beep beep!</i>	1	3	<i>Sad</i> 😞
<i>Aack beep aack!</i>	2	1	<i>Happy</i> 😊

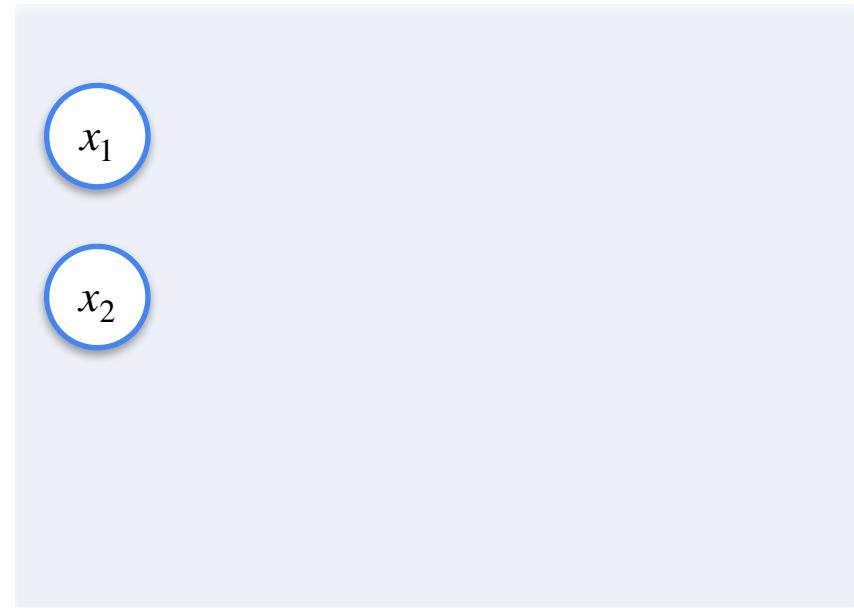
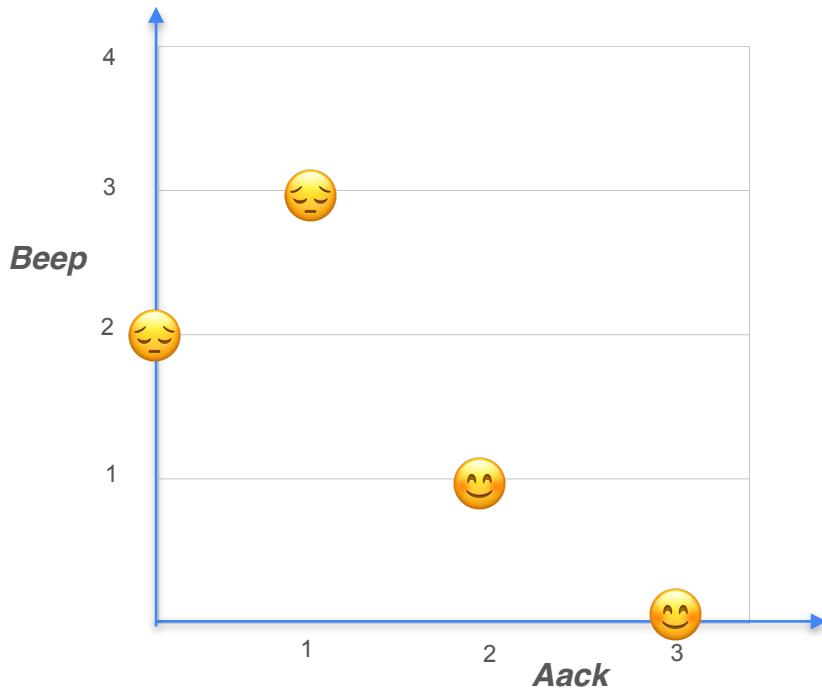
# Classification Problem Motivation



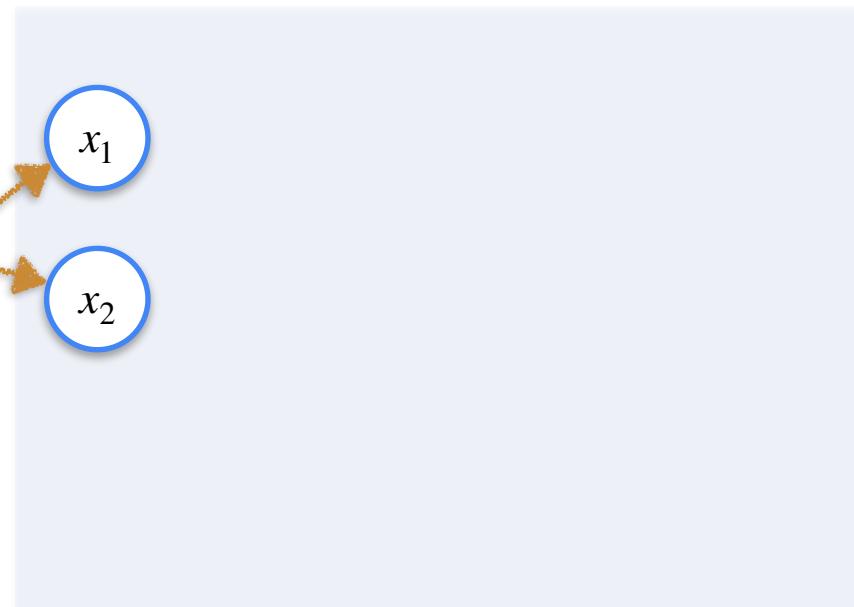
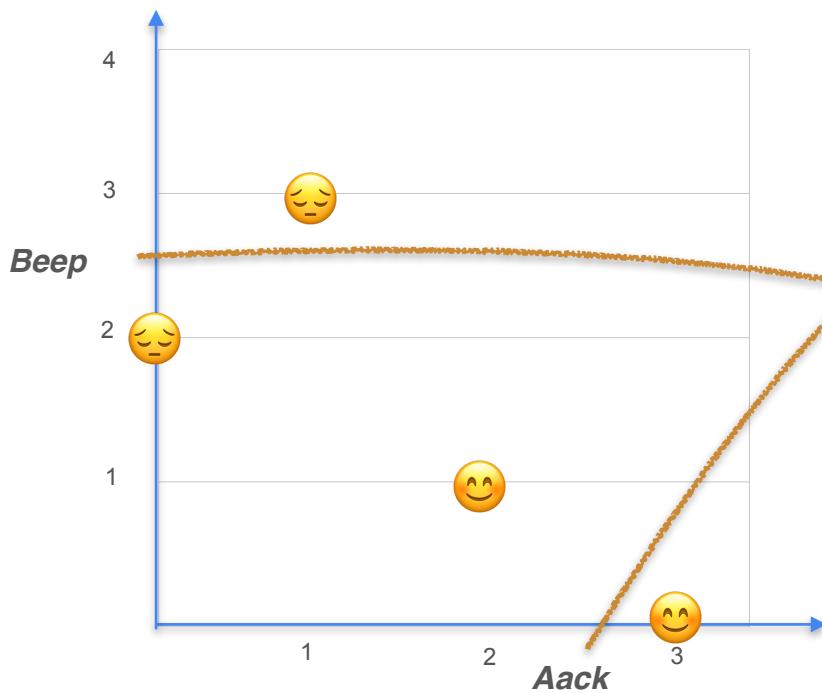
# Classification Problem Motivation



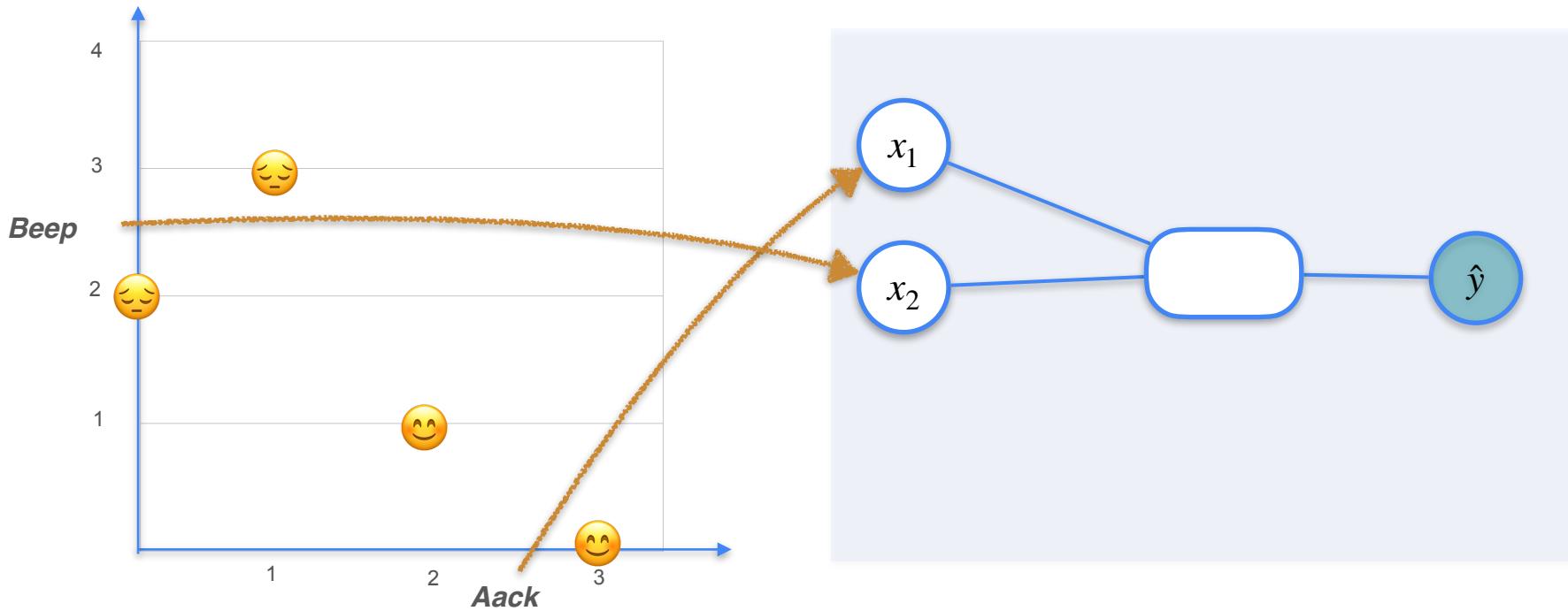
# Classification Problem Motivation



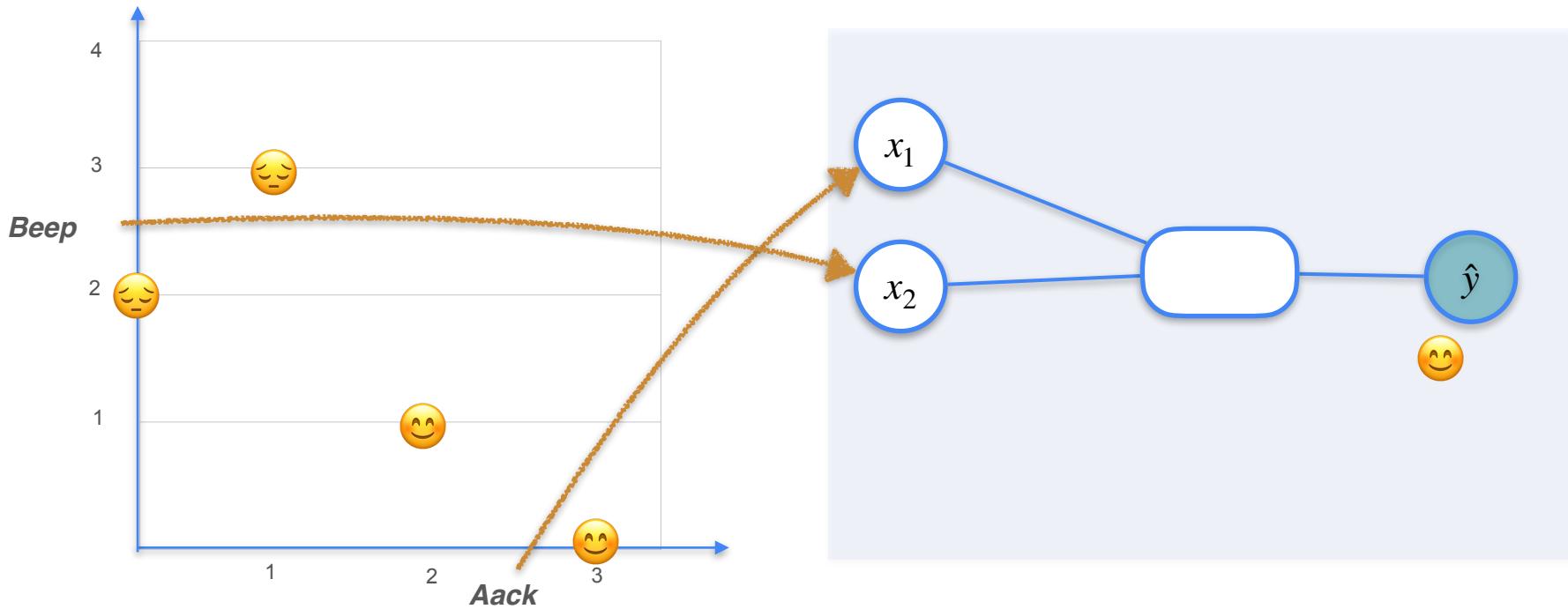
# Classification Problem Motivation



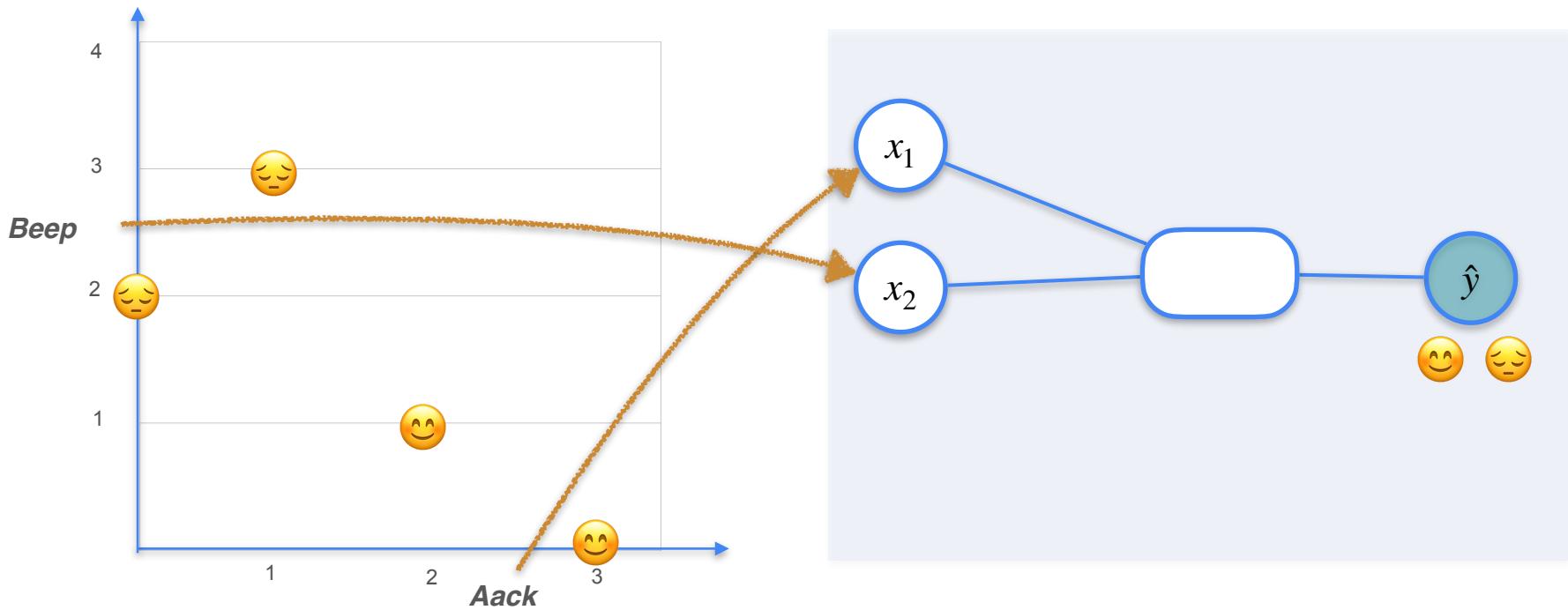
# Classification Problem Motivation



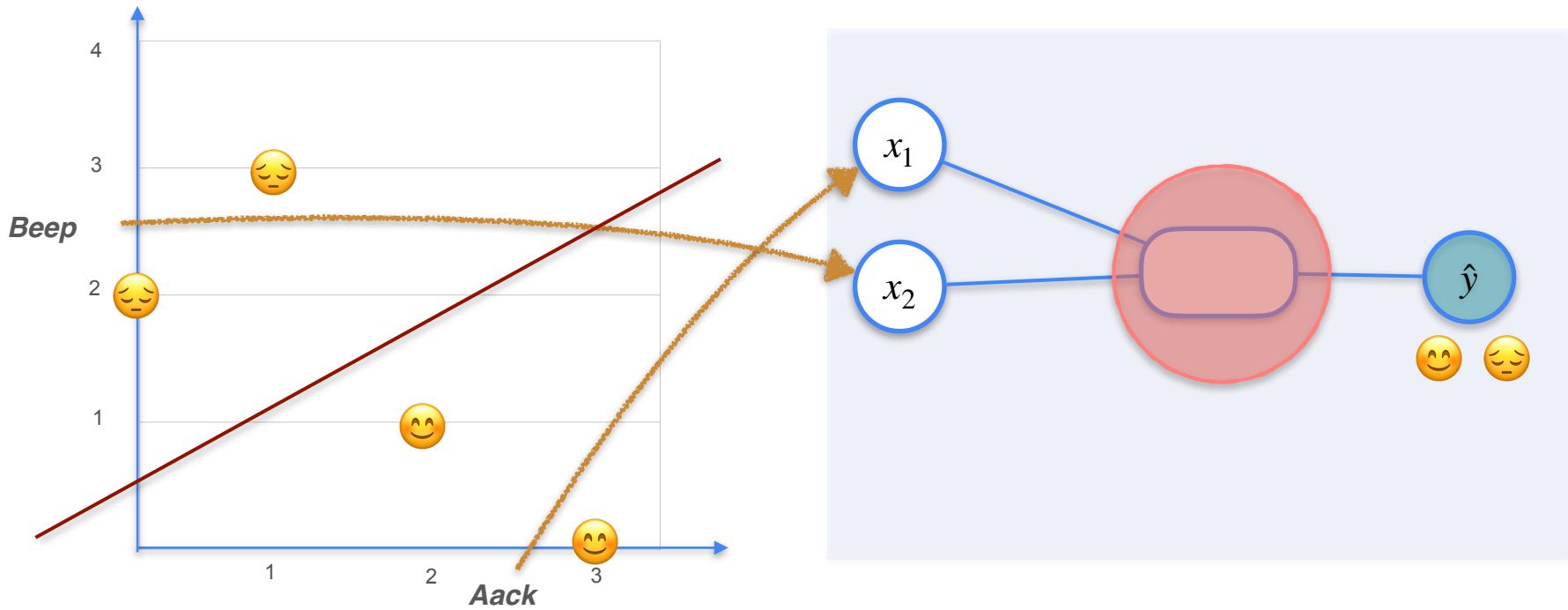
# Classification Problem Motivation



# Classification Problem Motivation

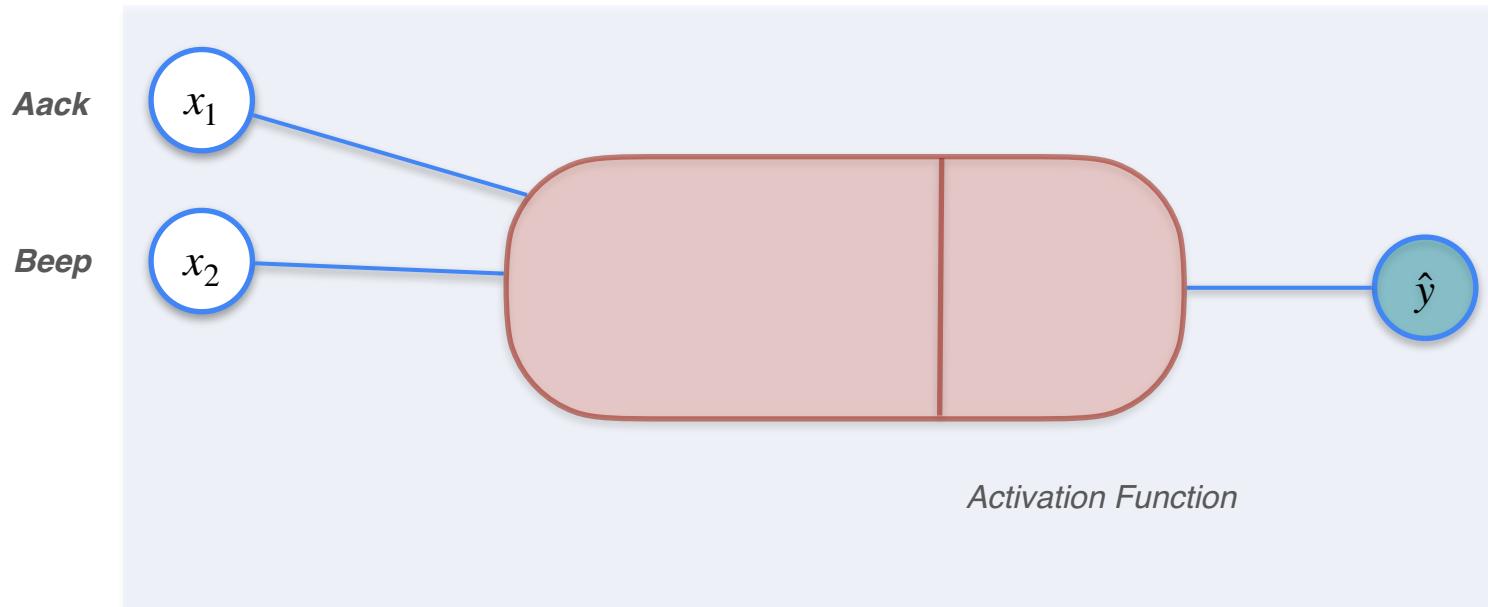


# Classification Problem Motivation



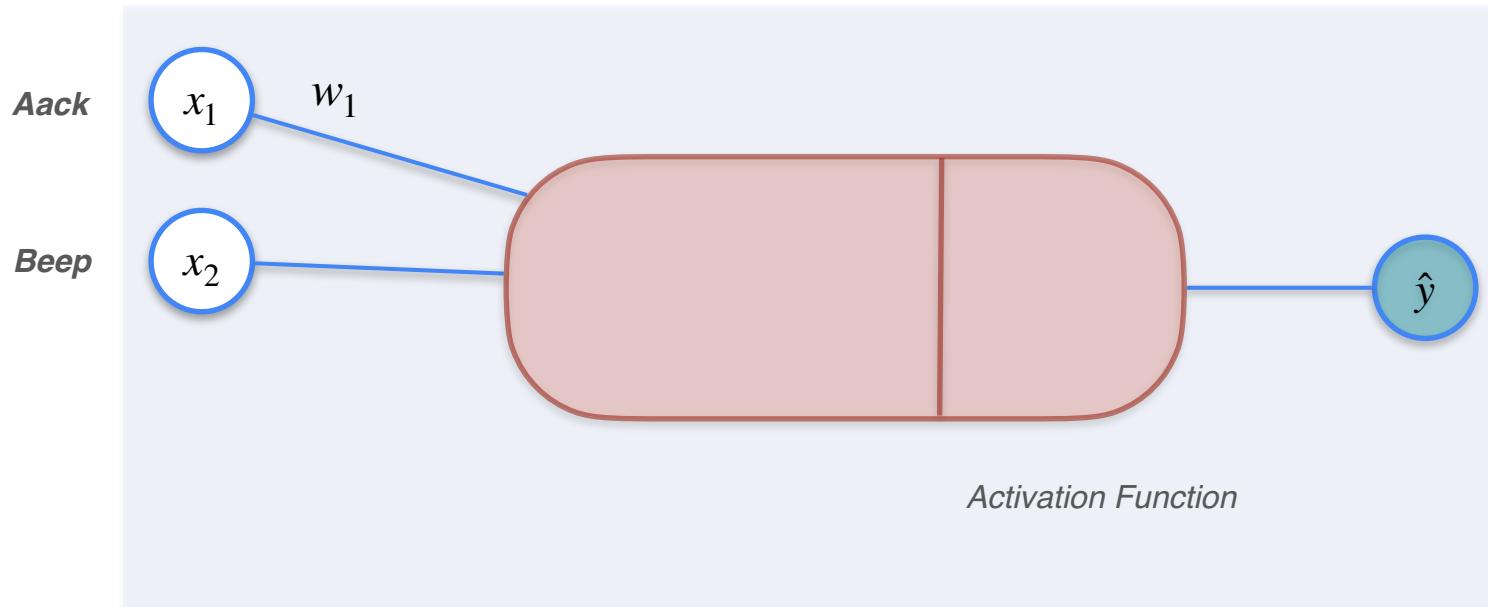
# Classification With a Perceptron

Single Layer Neural Network Perceptron



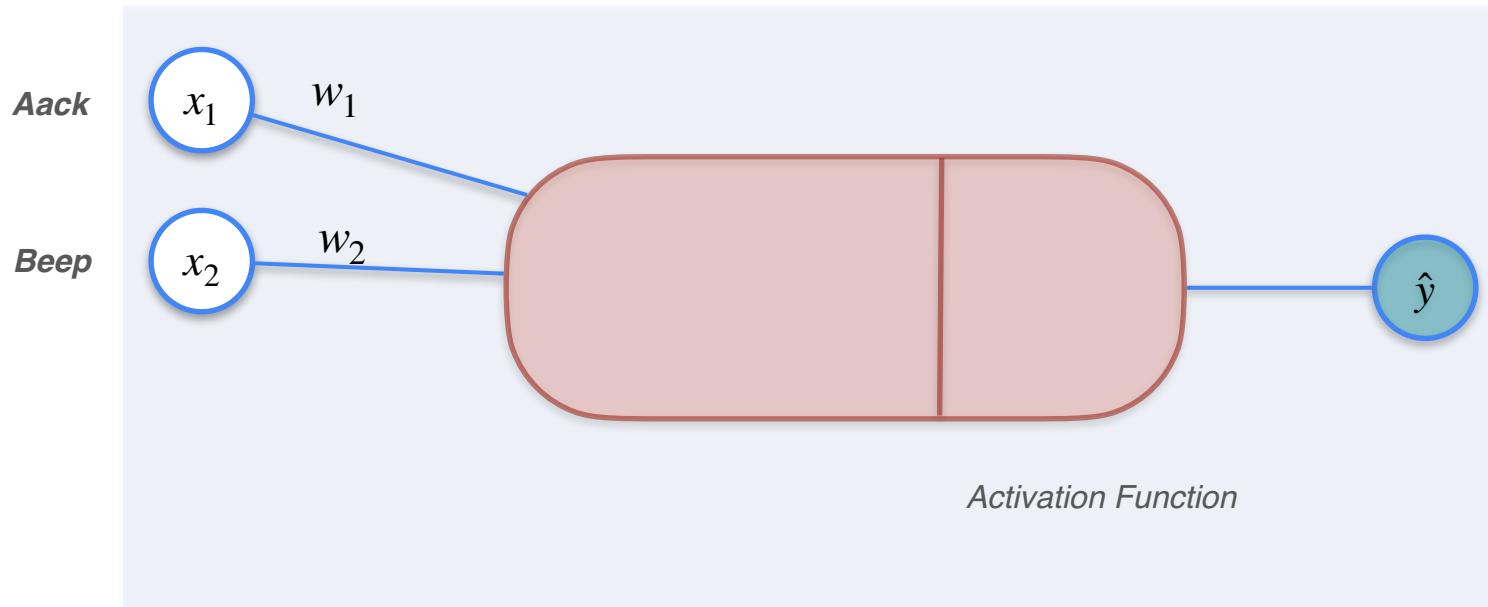
# Classification With a Perceptron

Single Layer Neural Network Perceptron



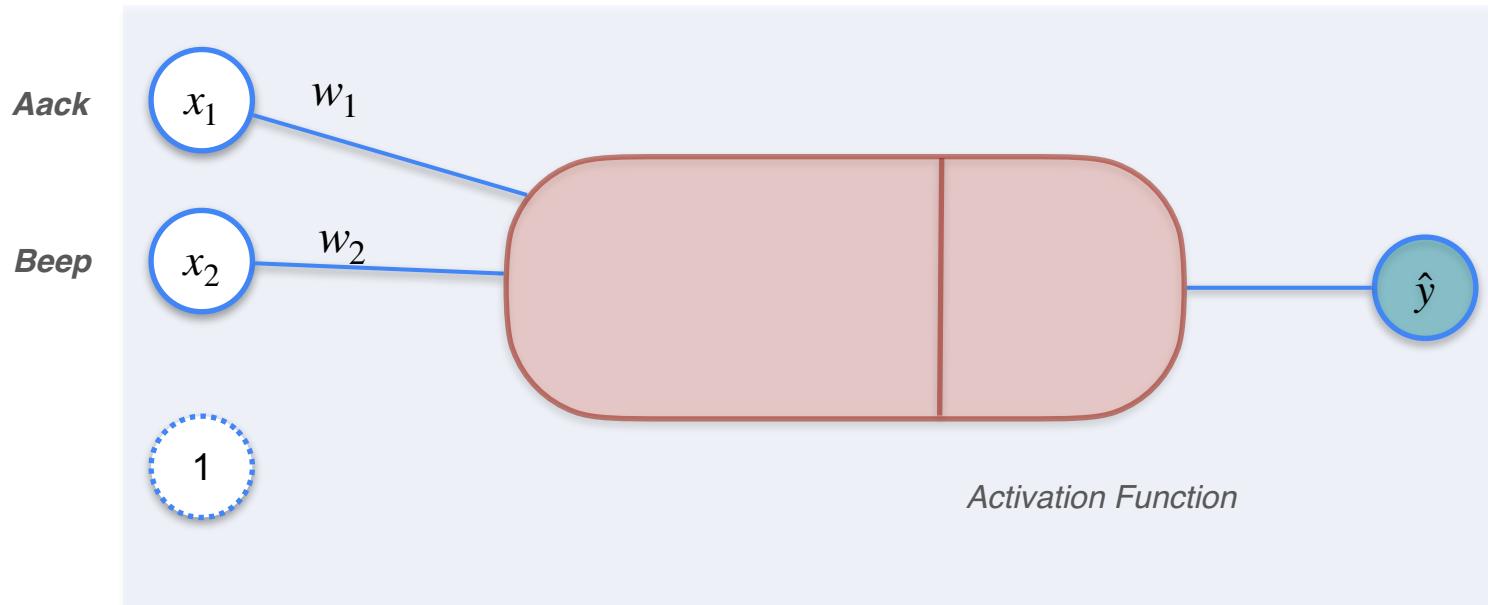
# Classification With a Perceptron

Single Layer Neural Network Perceptron



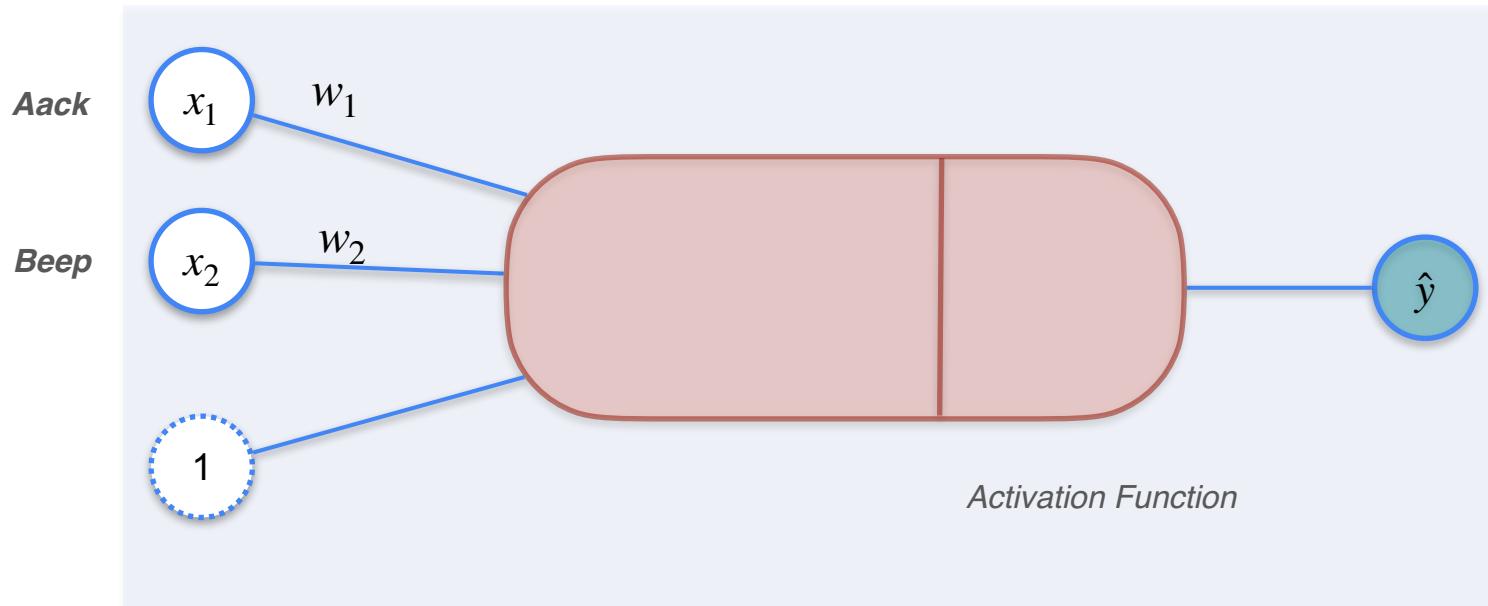
# Classification With a Perceptron

Single Layer Neural Network Perceptron



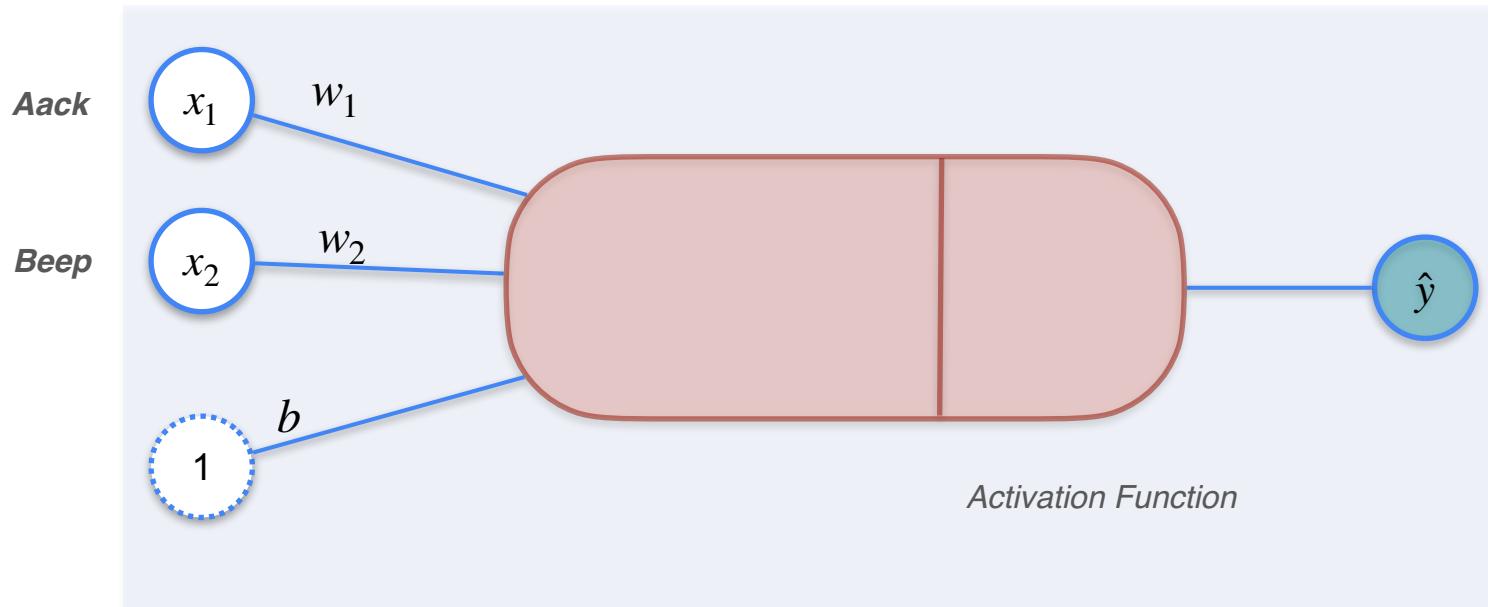
# Classification With a Perceptron

Single Layer Neural Network Perceptron



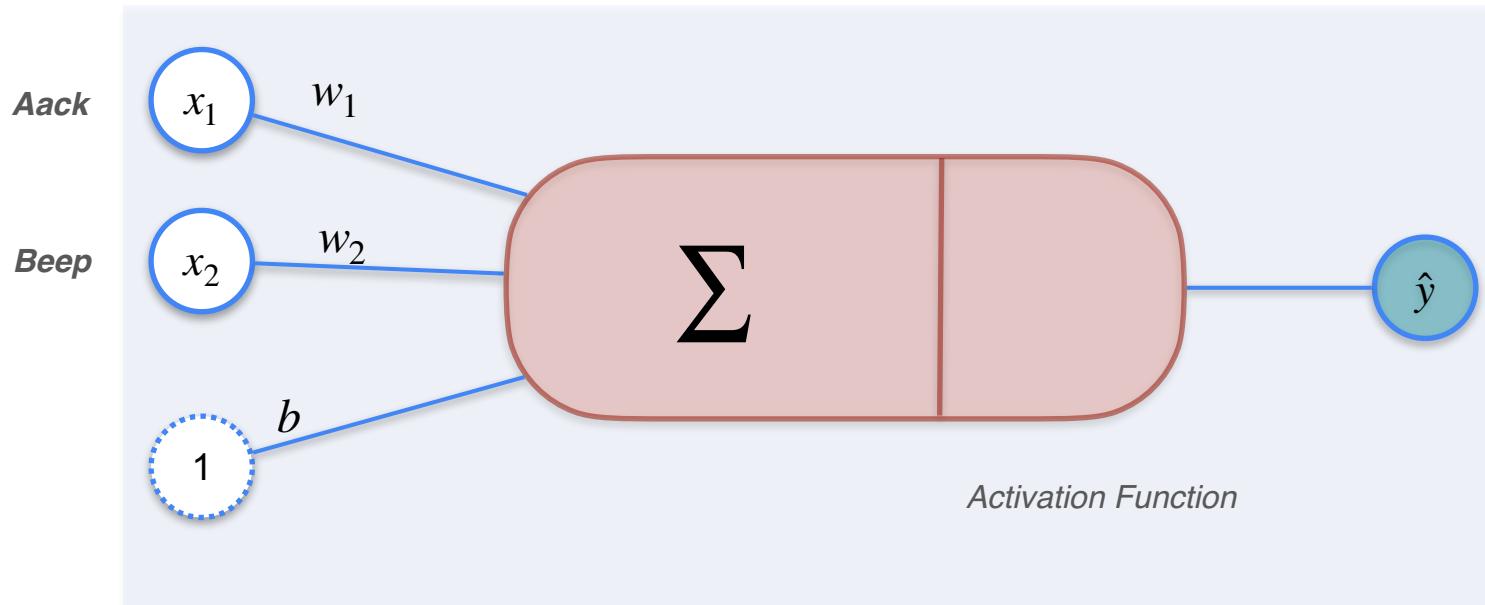
# Classification With a Perceptron

Single Layer Neural Network Perceptron



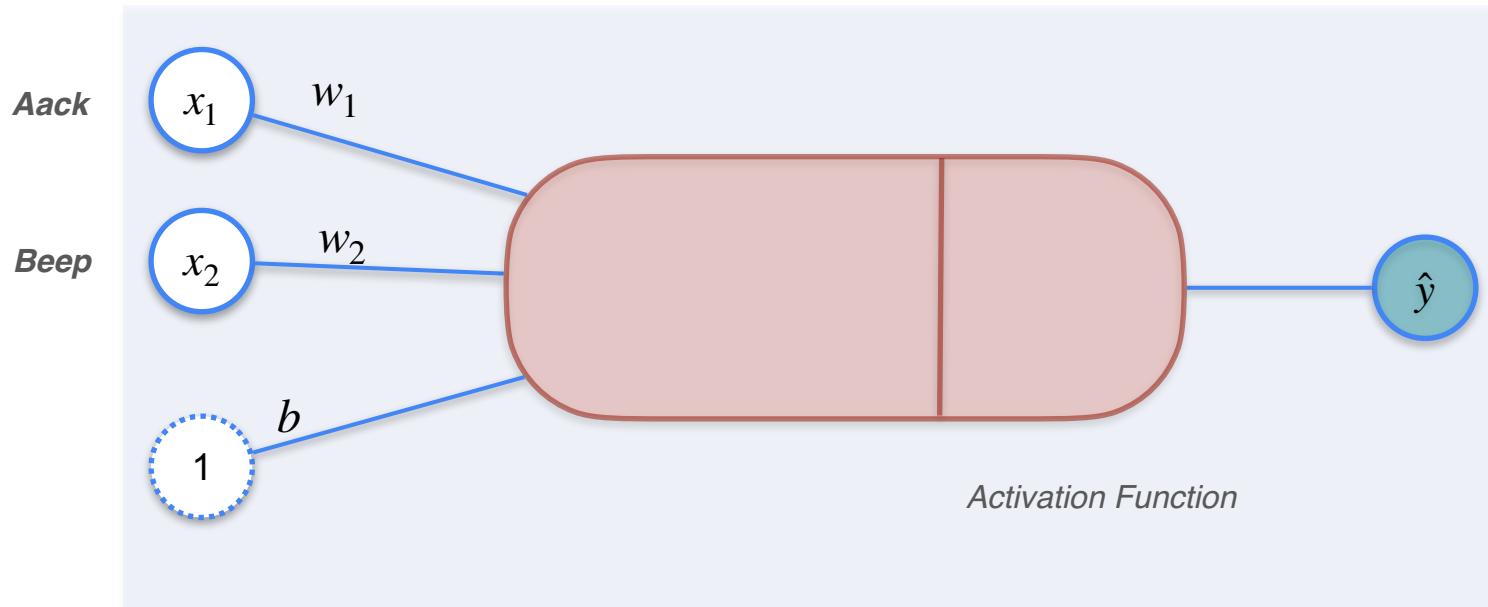
# Classification With a Perceptron

Single Layer Neural Network Perceptron



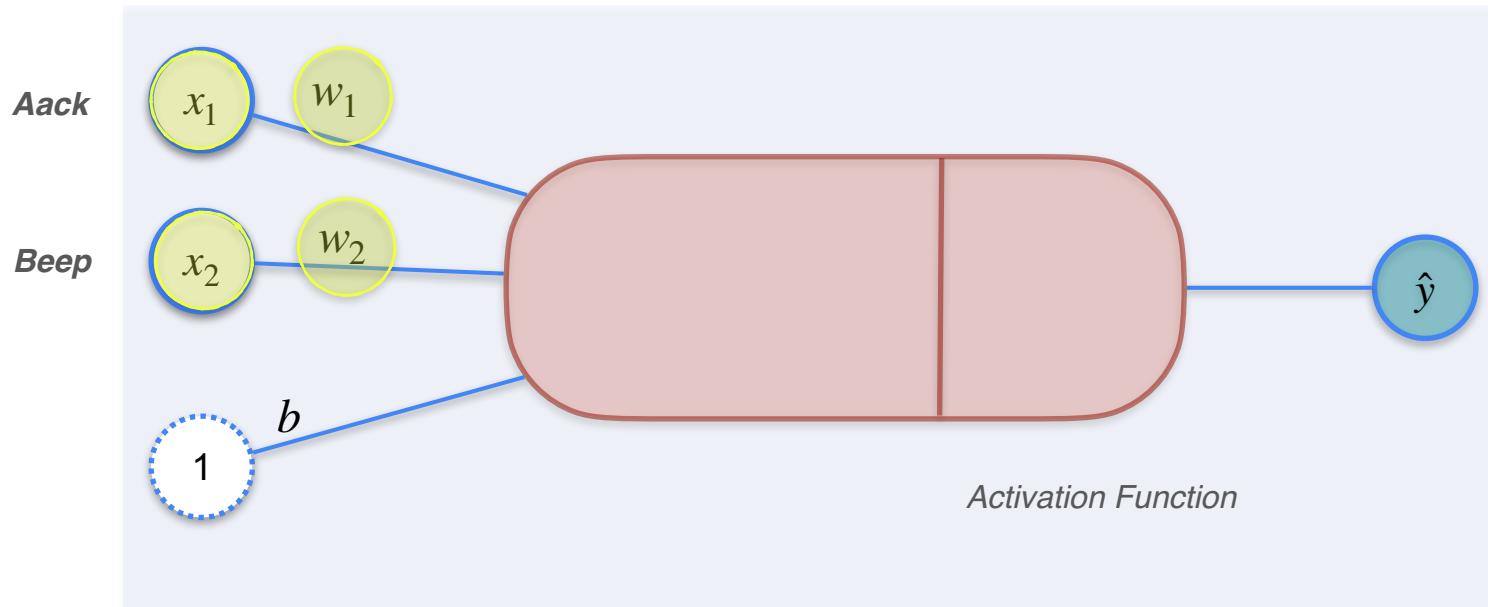
# Classification With a Perceptron

Single Layer Neural Network Perceptron



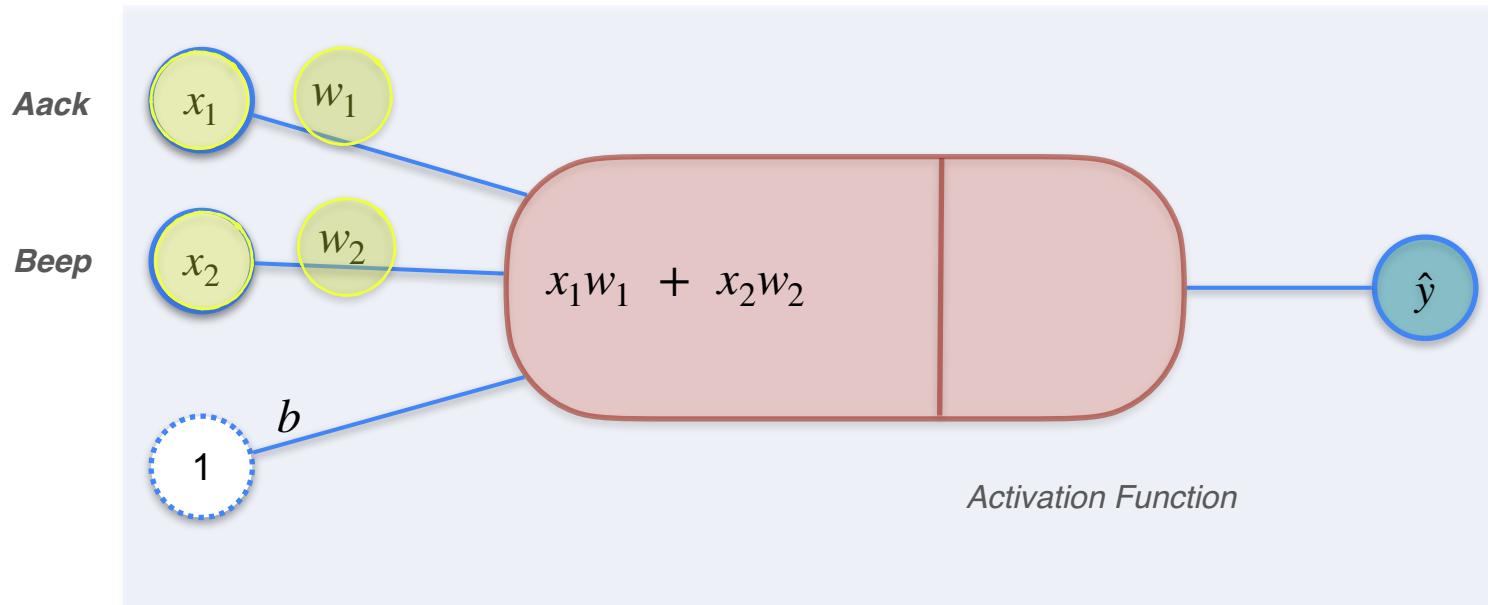
# Classification With a Perceptron

Single Layer Neural Network Perceptron



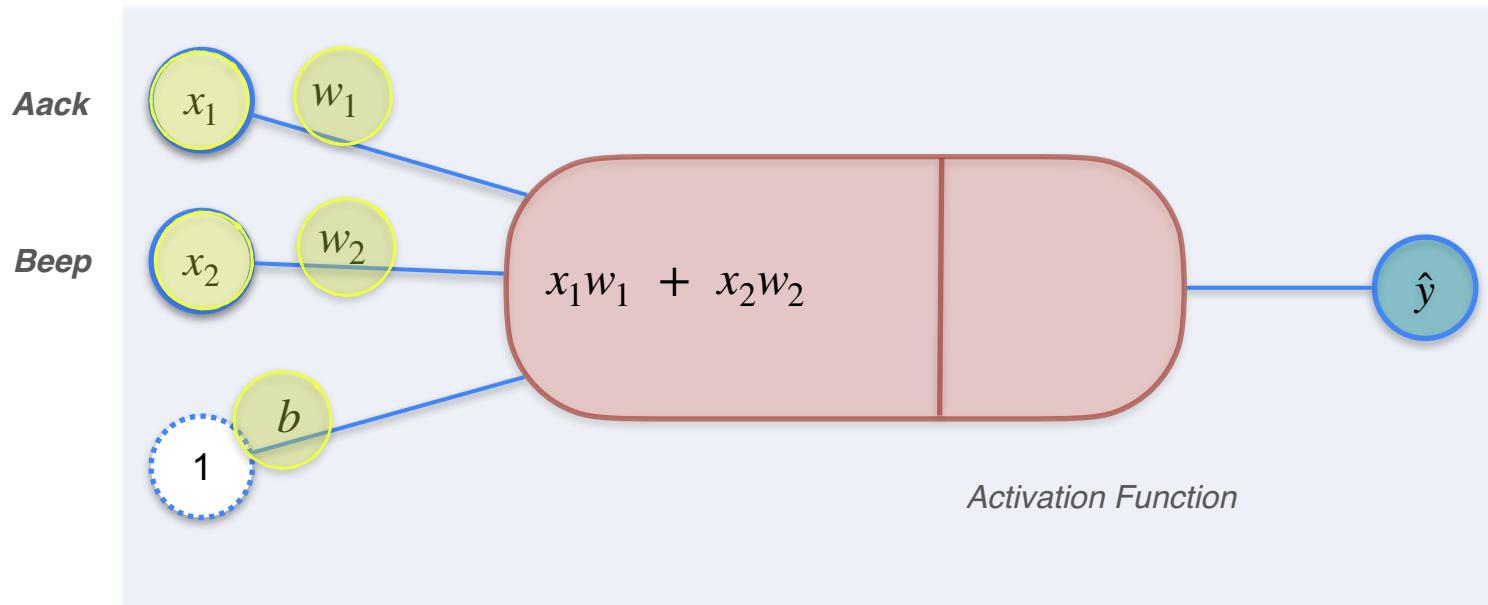
# Classification With a Perceptron

Single Layer Neural Network Perceptron



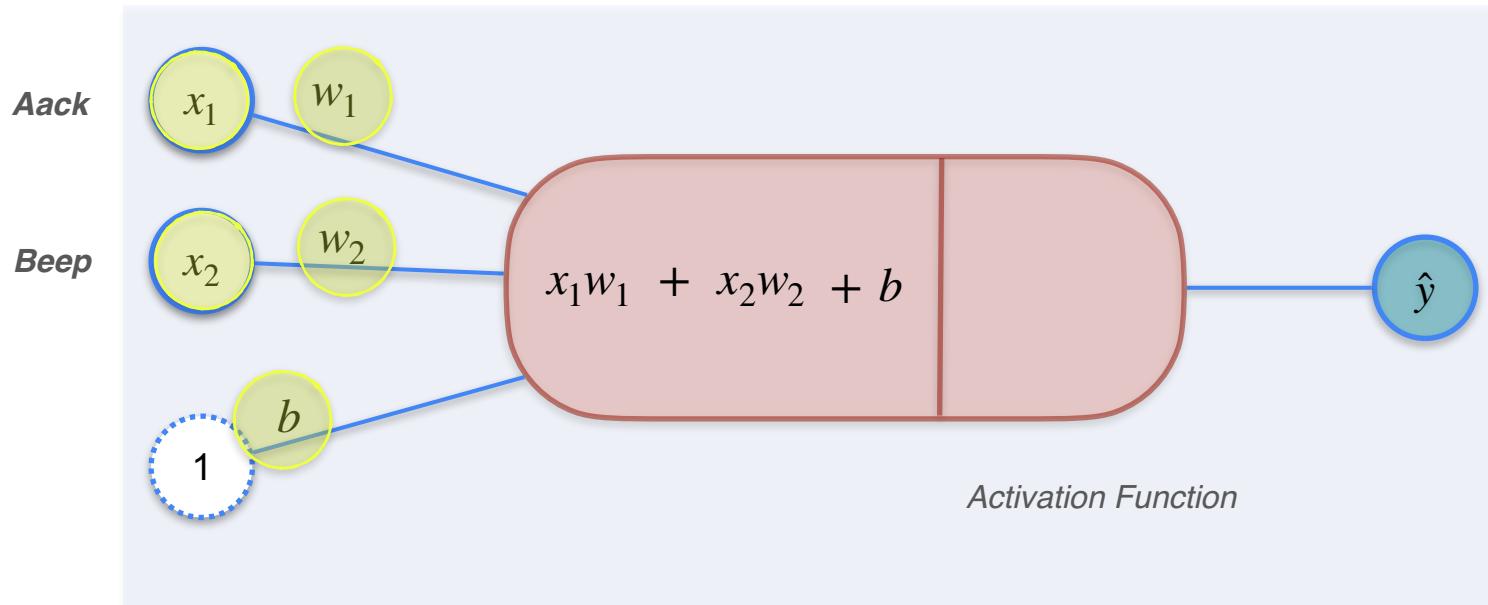
# Classification With a Perceptron

Single Layer Neural Network Perceptron



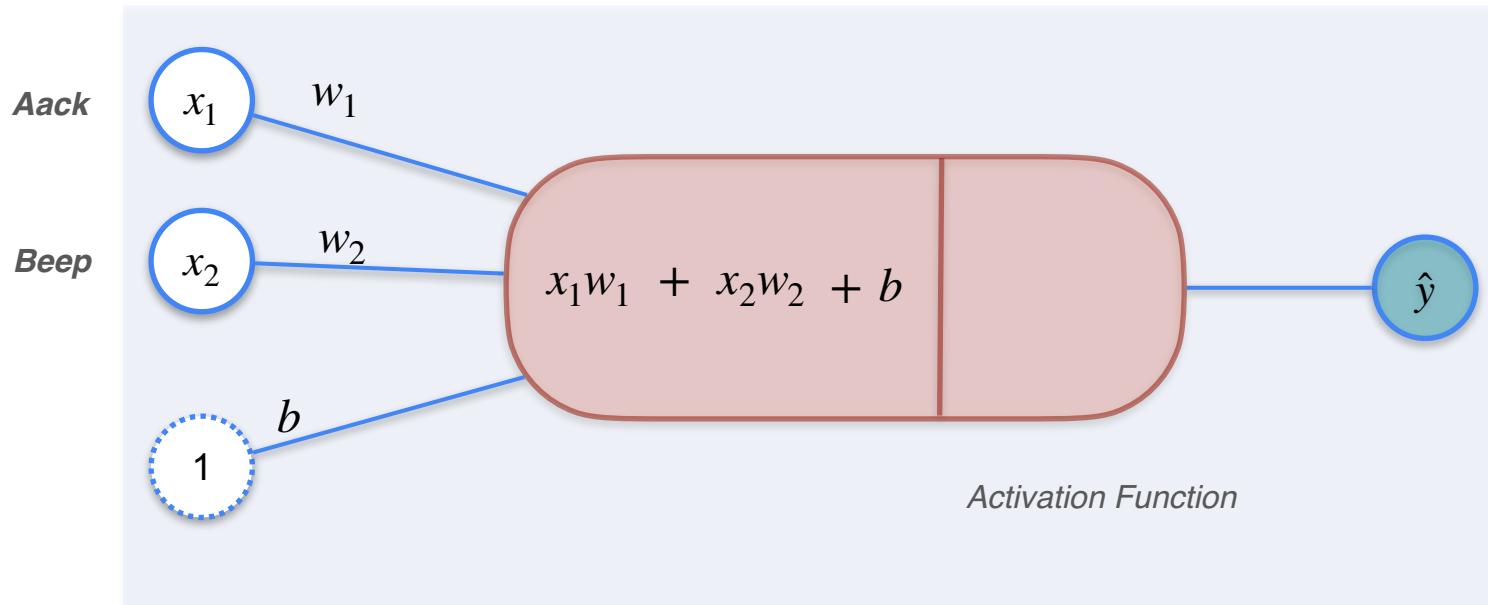
# Classification With a Perceptron

Single Layer Neural Network Perceptron



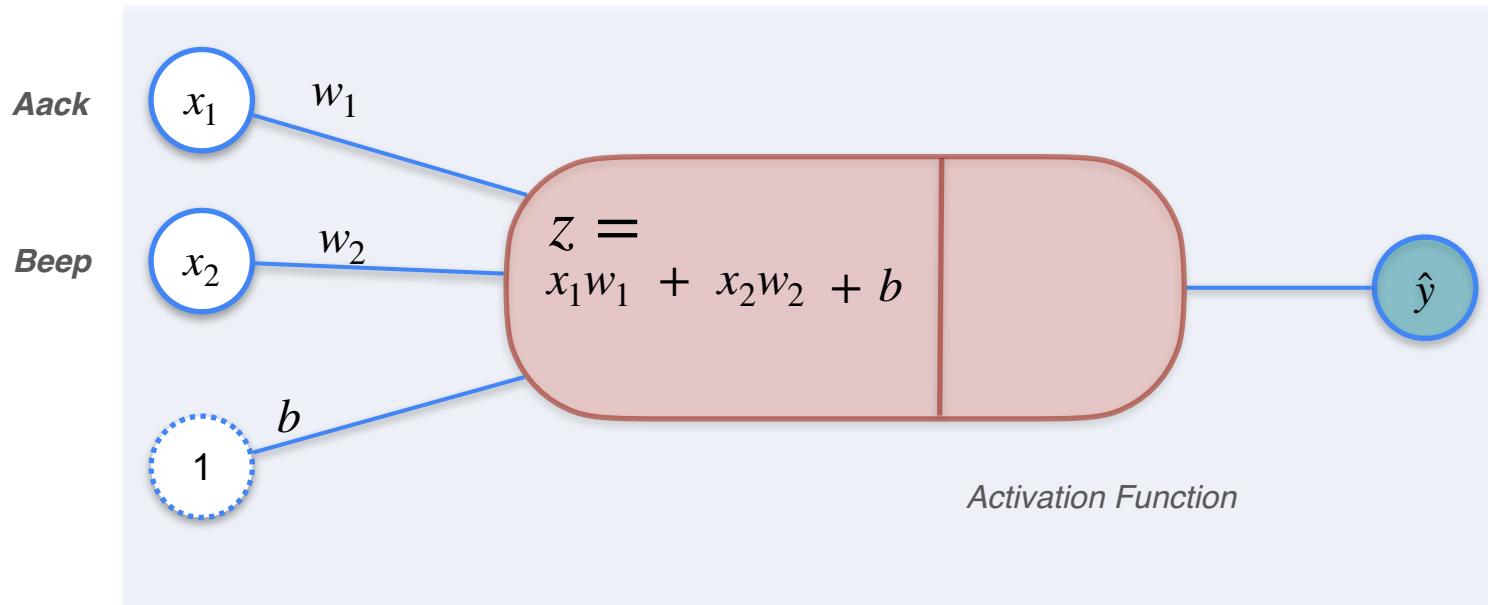
# Classification With a Perceptron

Single Layer Neural Network Perceptron



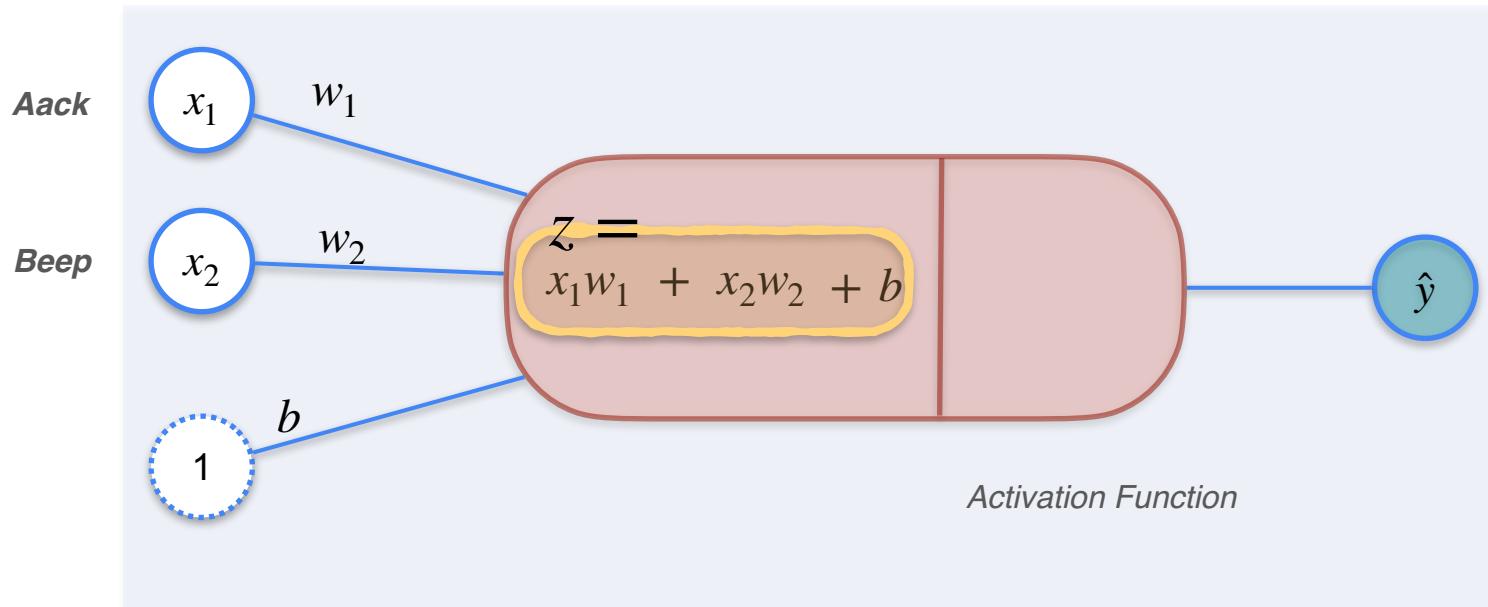
# Classification With a Perceptron

Single Layer Neural Network Perceptron



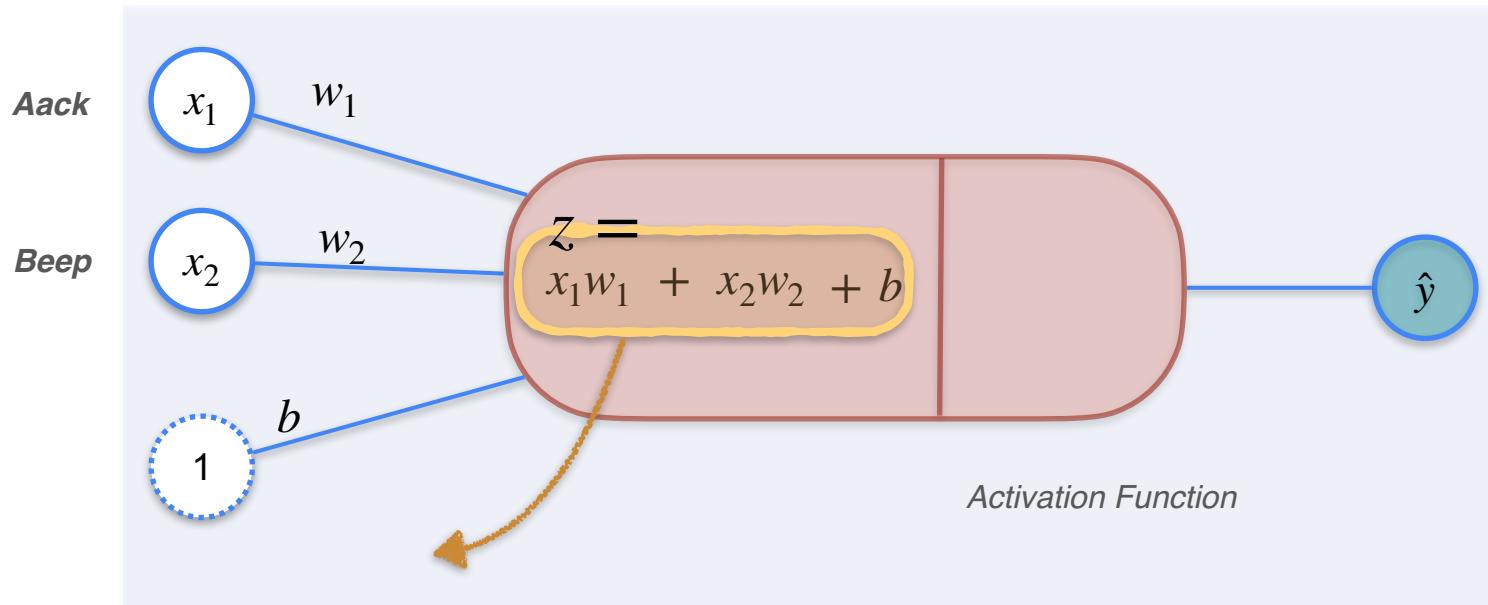
# Classification With a Perceptron

Single Layer Neural Network Perceptron



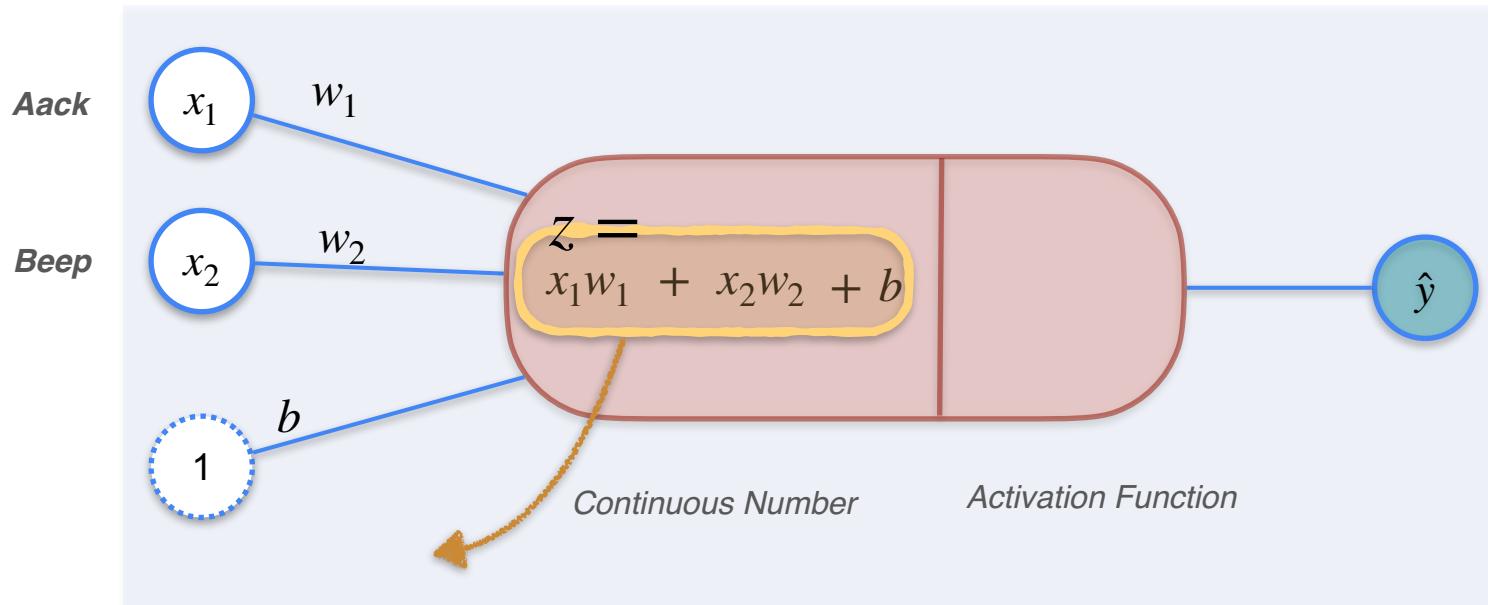
# Classification With a Perceptron

Single Layer Neural Network Perceptron



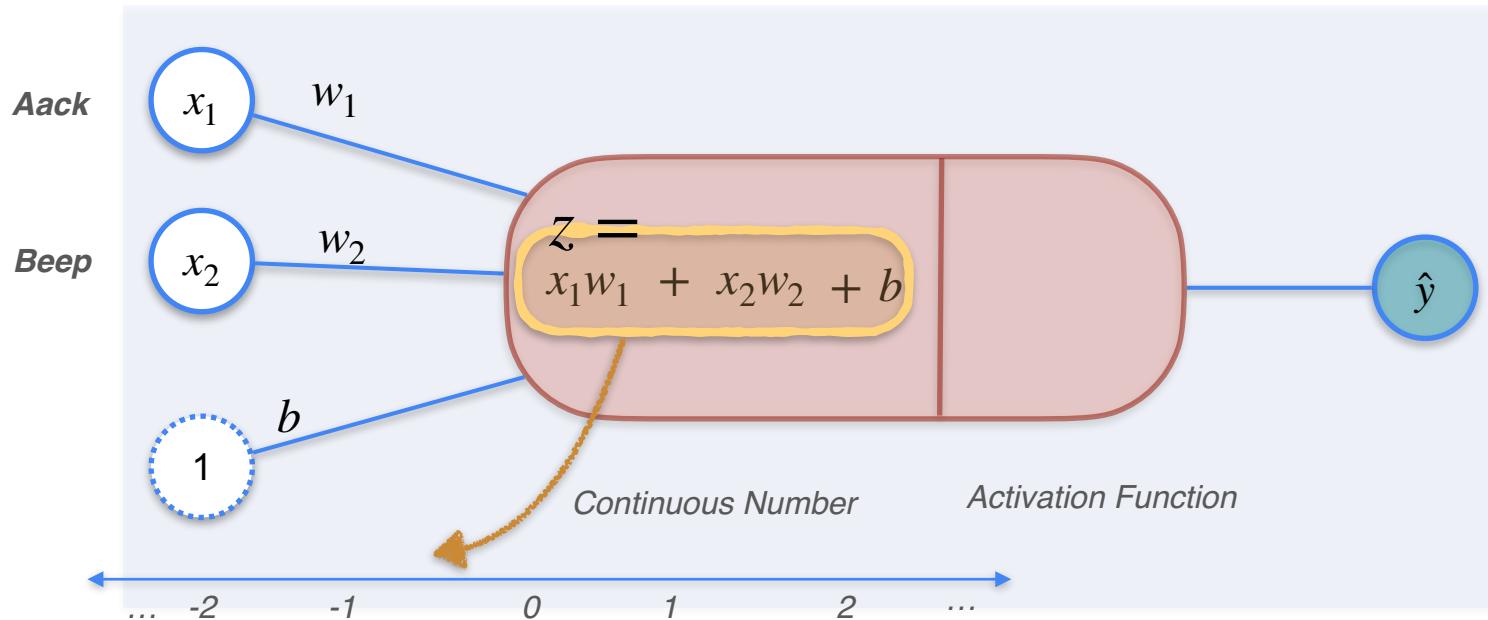
# Classification With a Perceptron

Single Layer Neural Network Perceptron



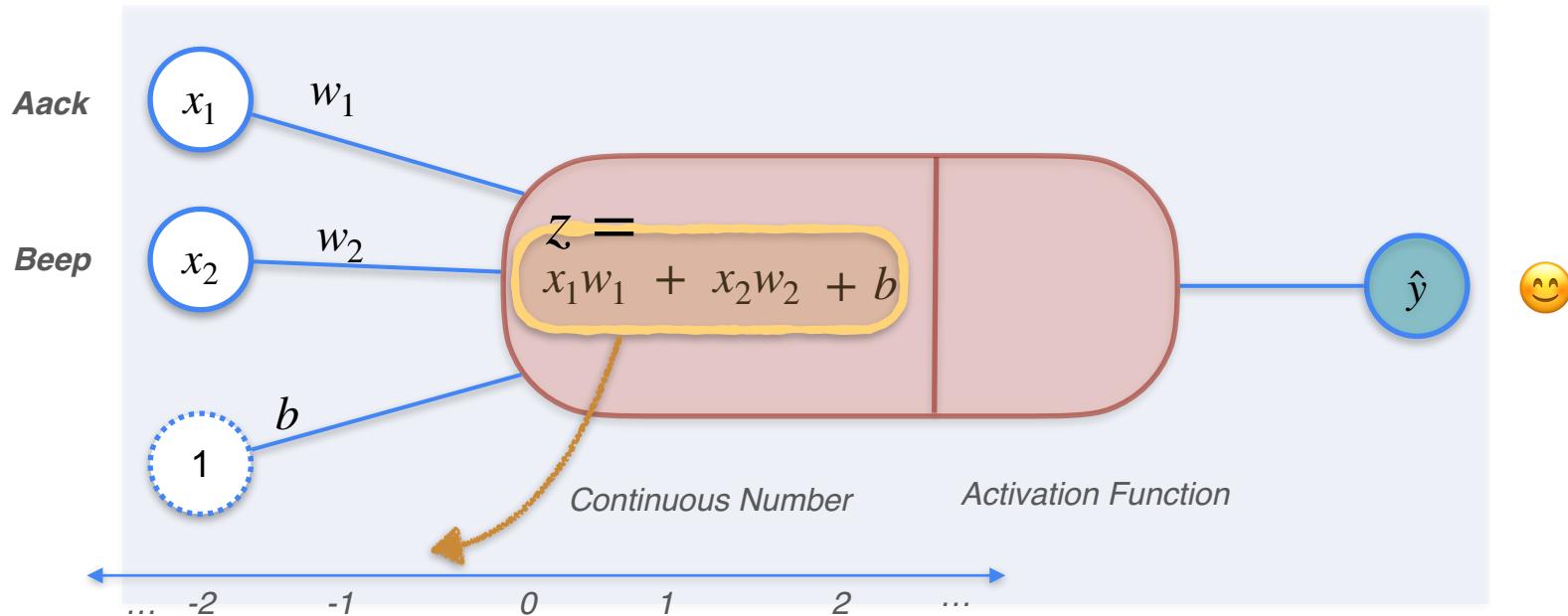
# Classification With a Perceptron

Single Layer Neural Network Perceptron



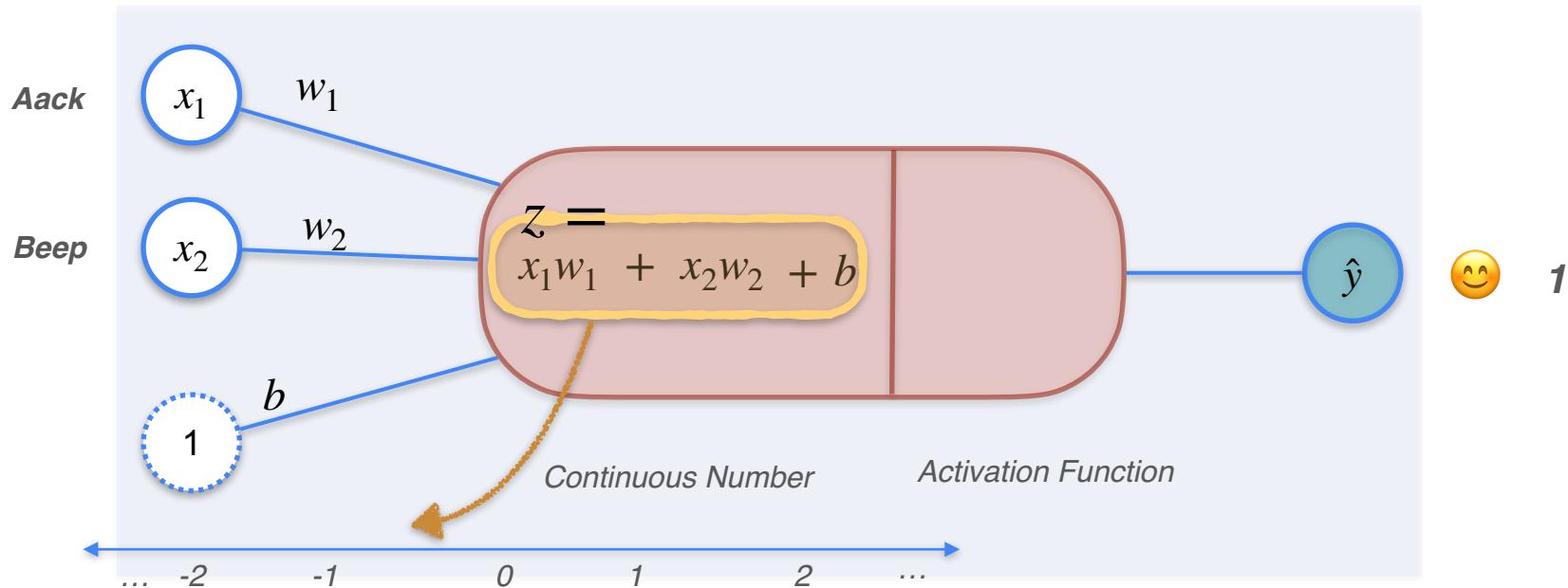
# Classification With a Perceptron

Single Layer Neural Network Perceptron



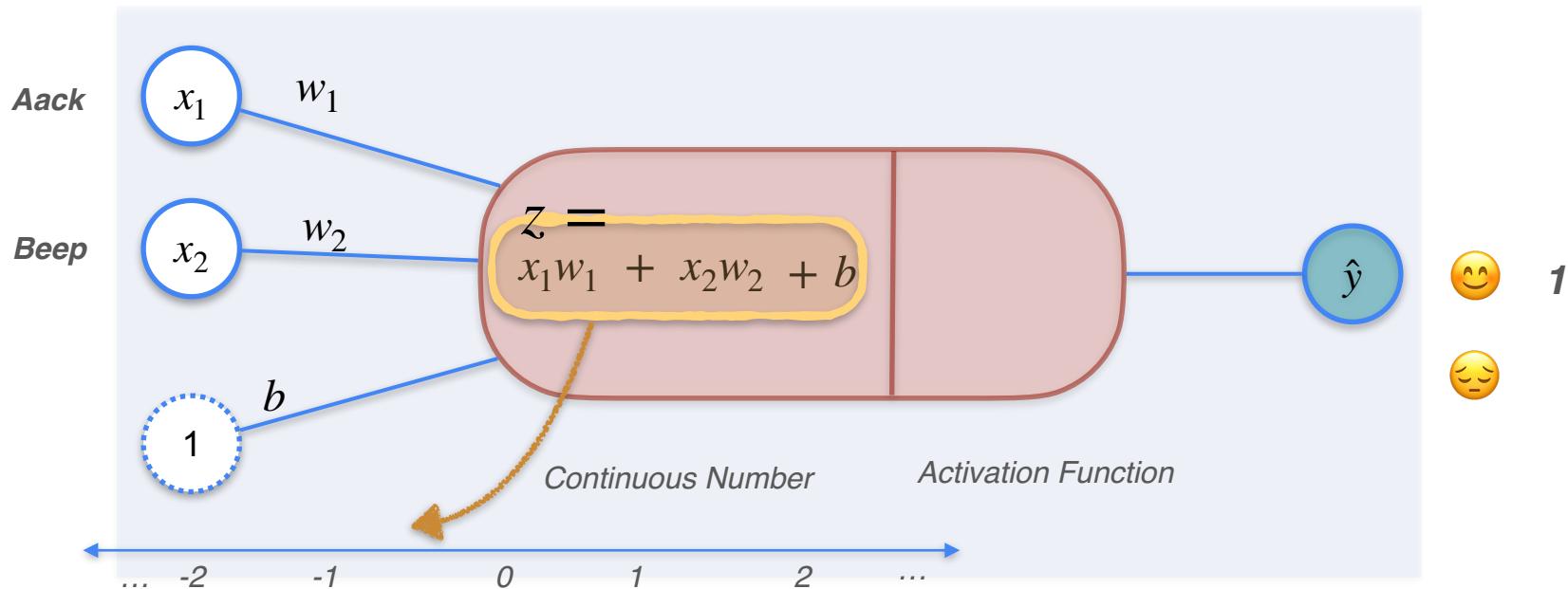
# Classification With a Perceptron

Single Layer Neural Network Perceptron



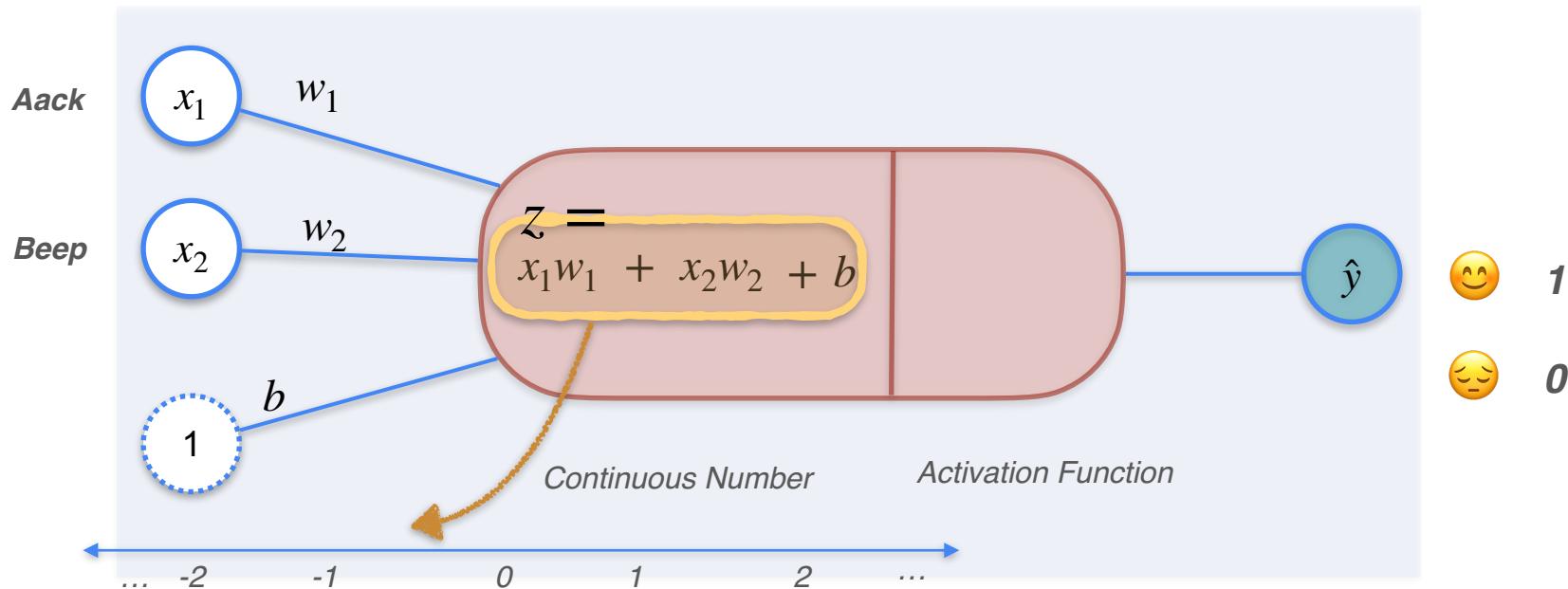
# Classification With a Perceptron

Single Layer Neural Network Perceptron



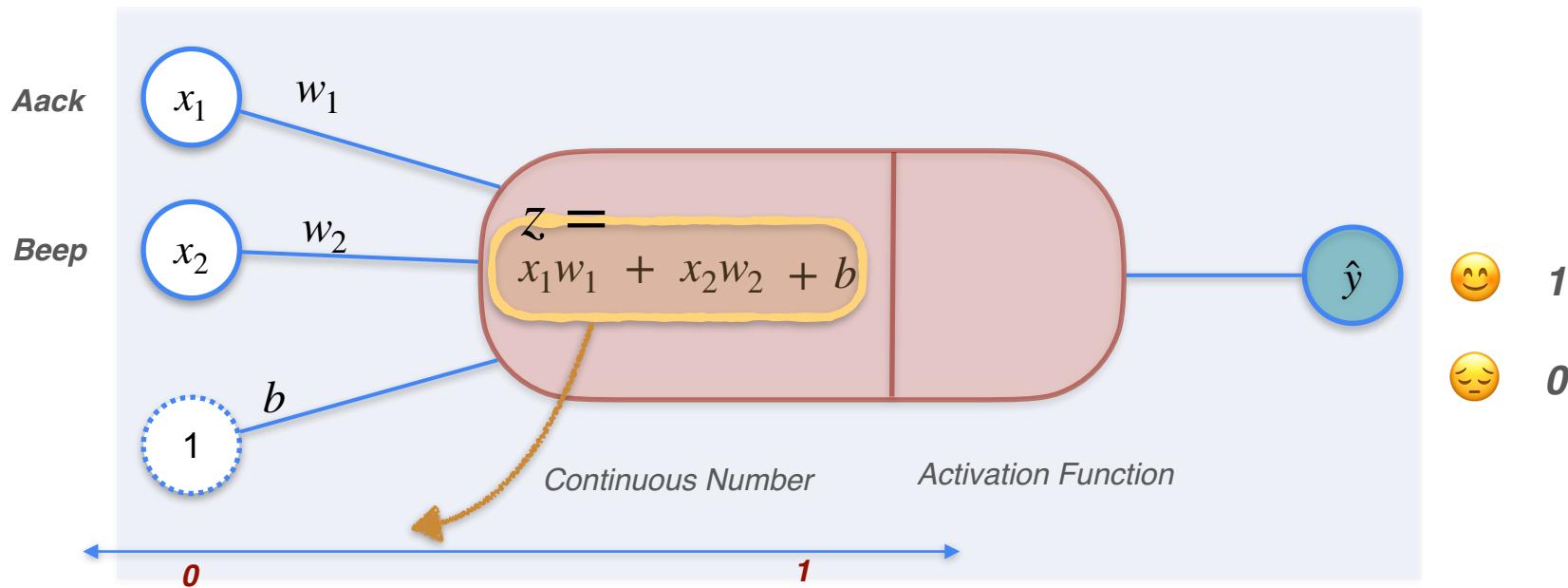
# Classification With a Perceptron

Single Layer Neural Network Perceptron



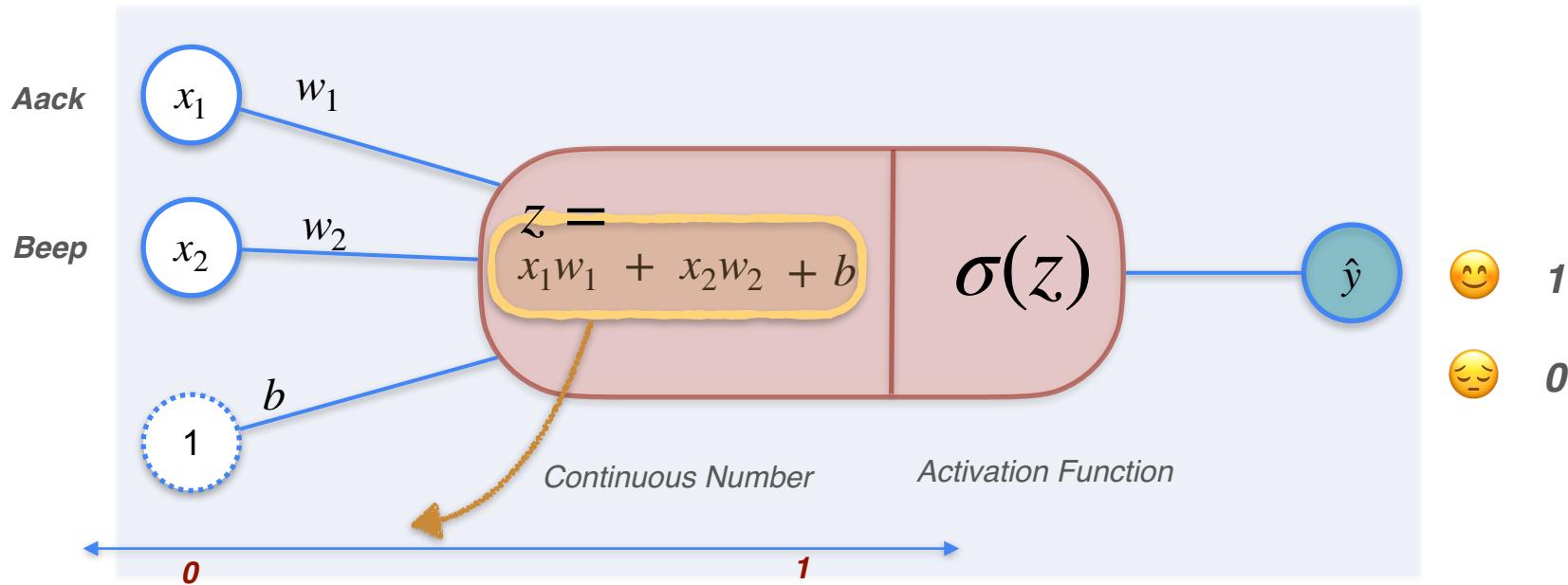
# Classification With a Perceptron

Single Layer Neural Network Perceptron



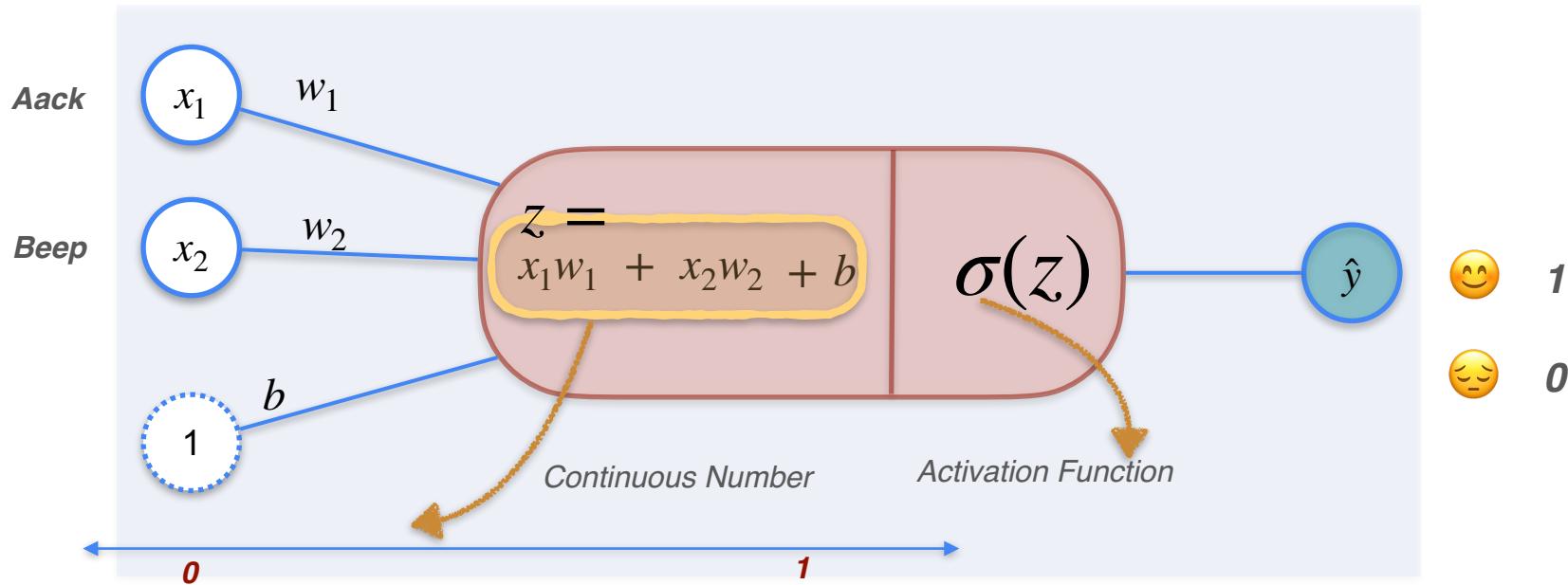
# Classification With a Perceptron

Single Layer Neural Network Perceptron



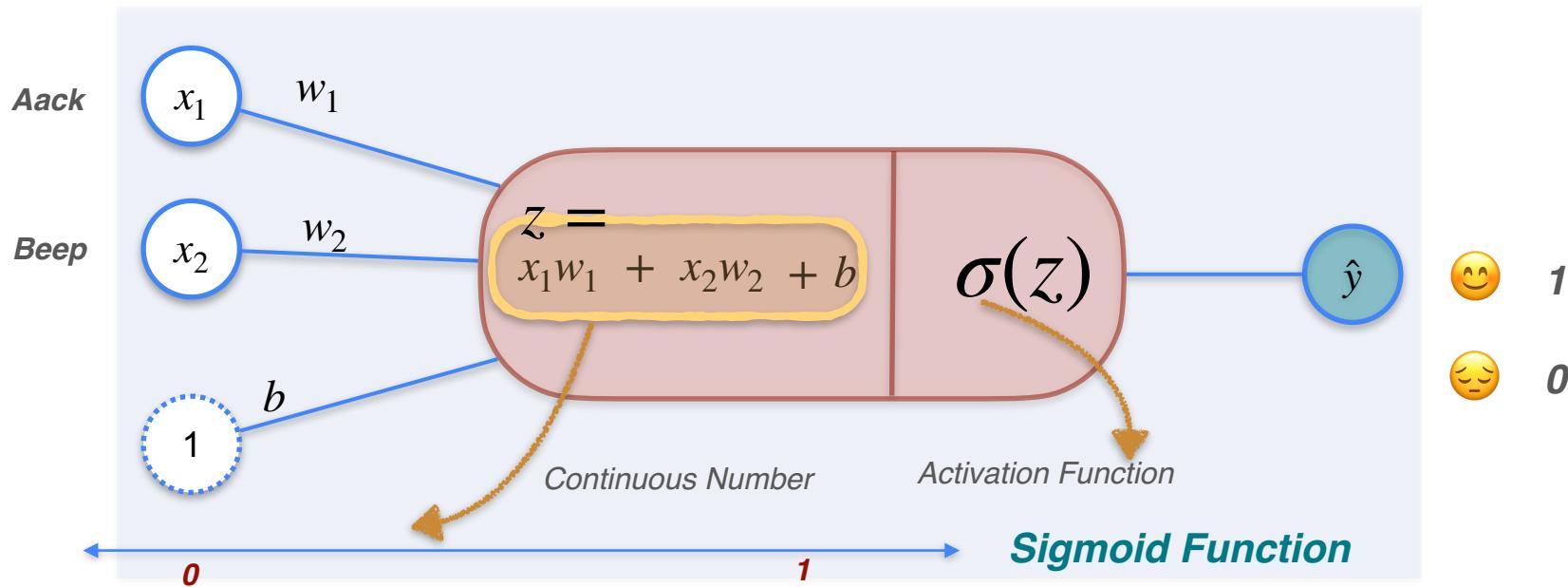
# Classification With a Perceptron

Single Layer Neural Network Perceptron

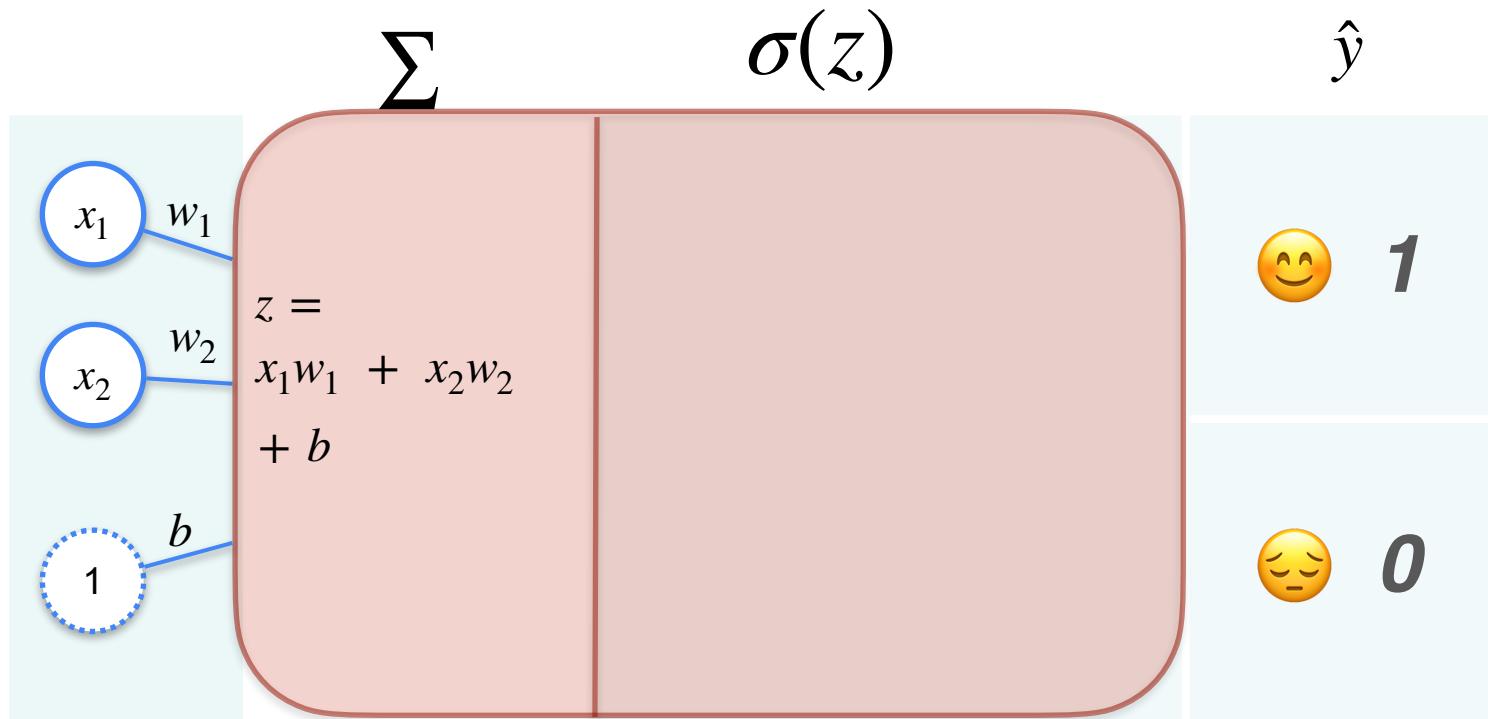


# Classification With a Perceptron

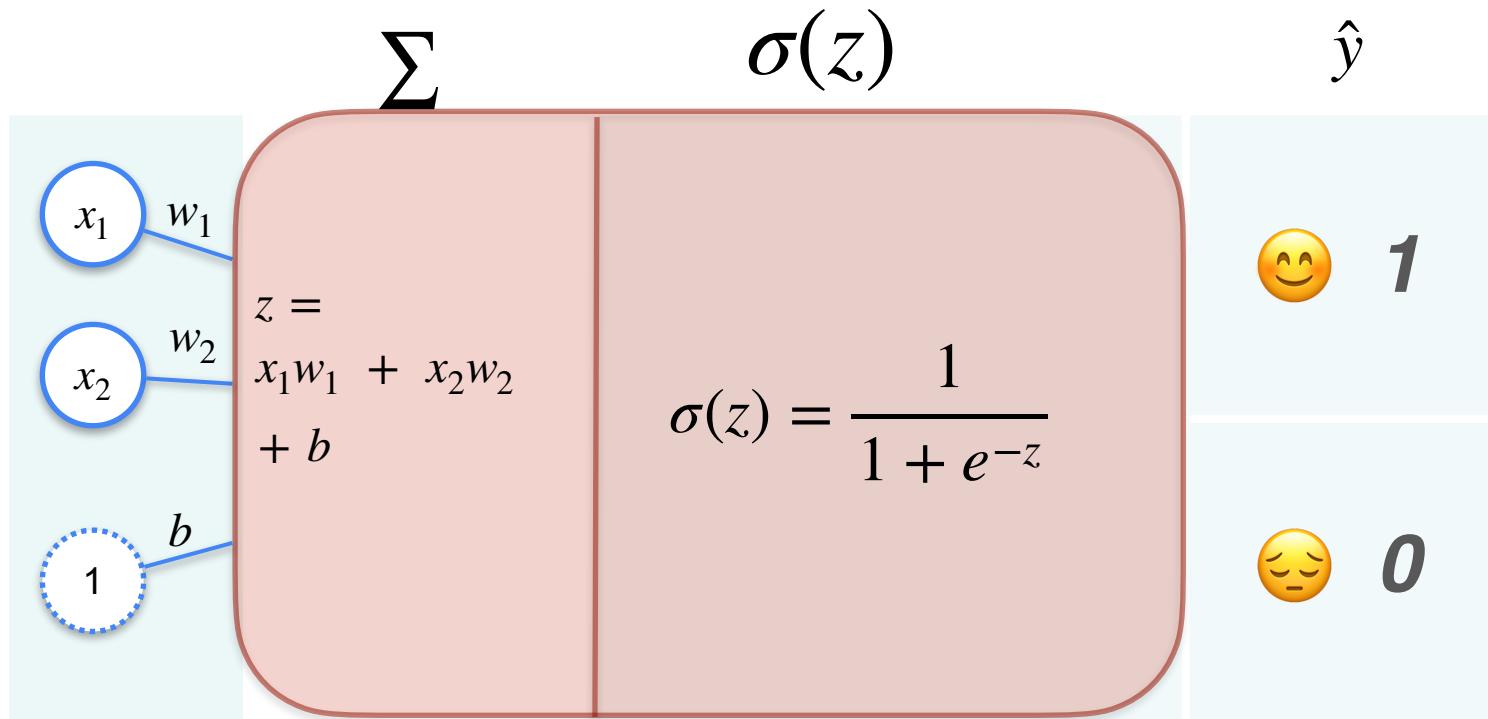
Single Layer Neural Network Perceptron



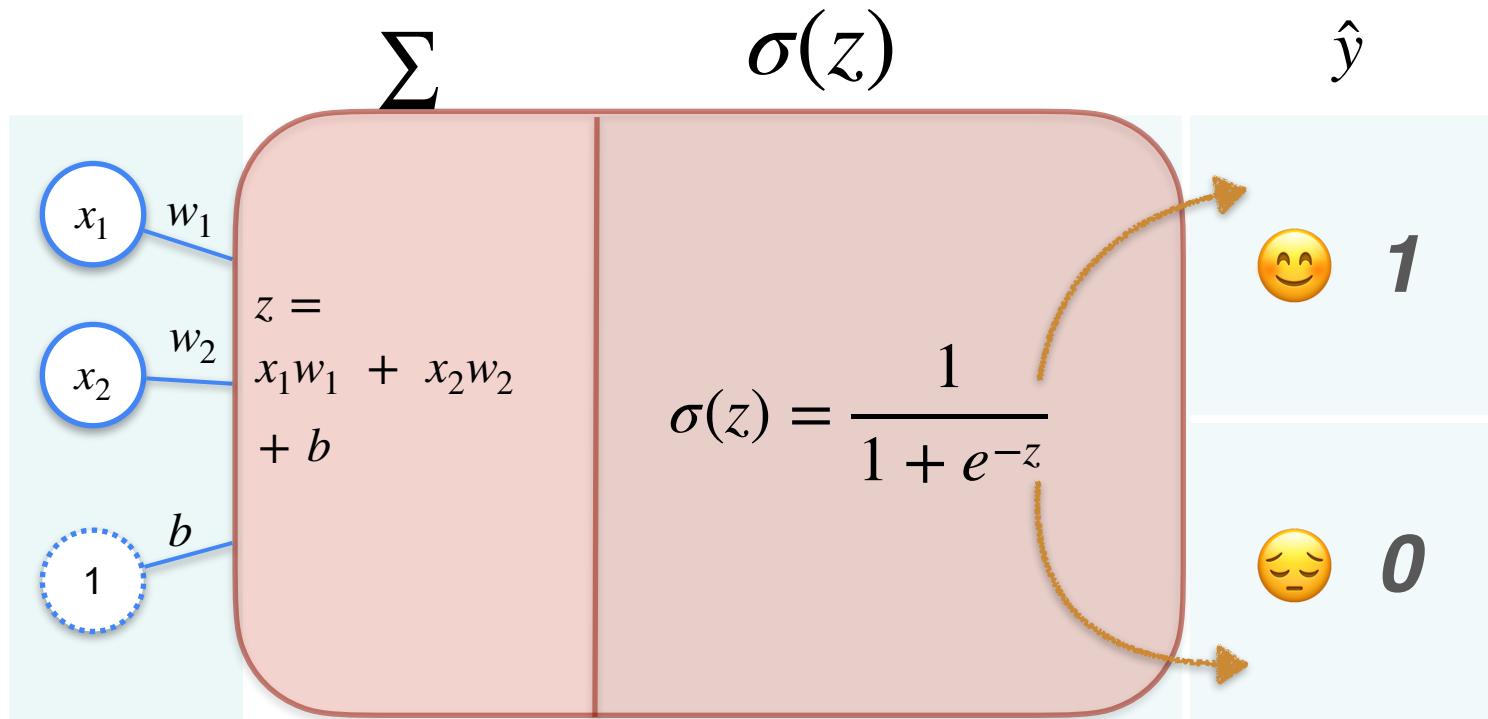
# Sigmoid Function



# Sigmoid Function



# Sigmoid Function





DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

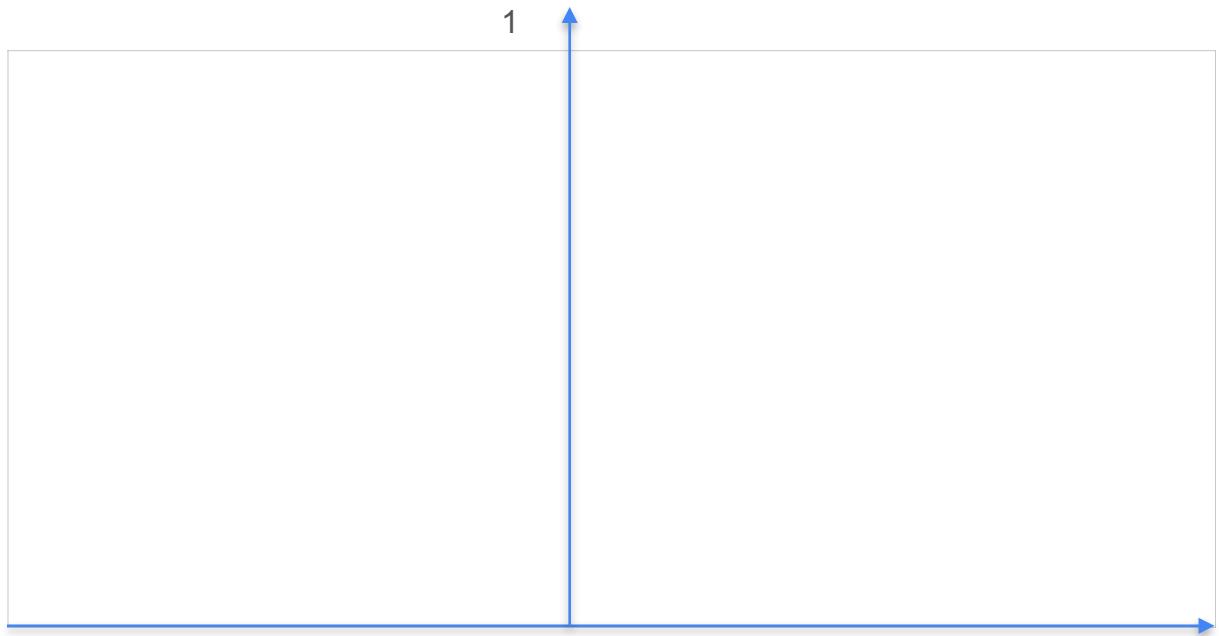
**Classification with a  
perceptron:  
The sigmoid function**

# Sigmoid Function

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

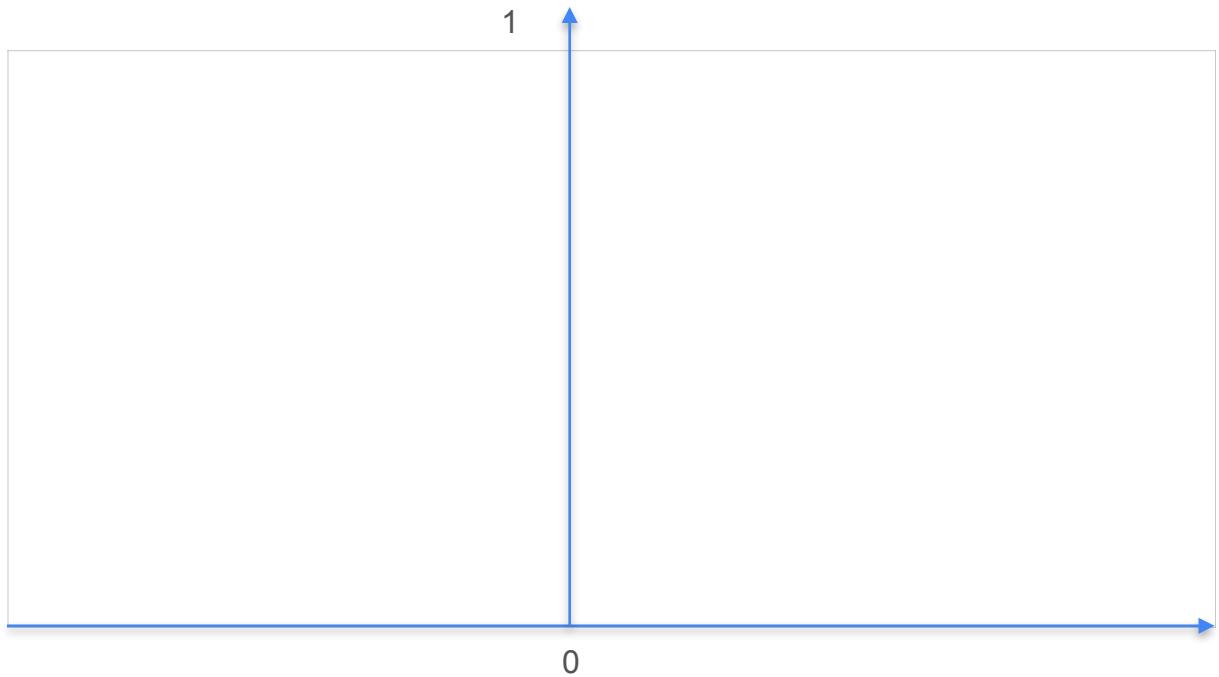
# Sigmoid Function

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



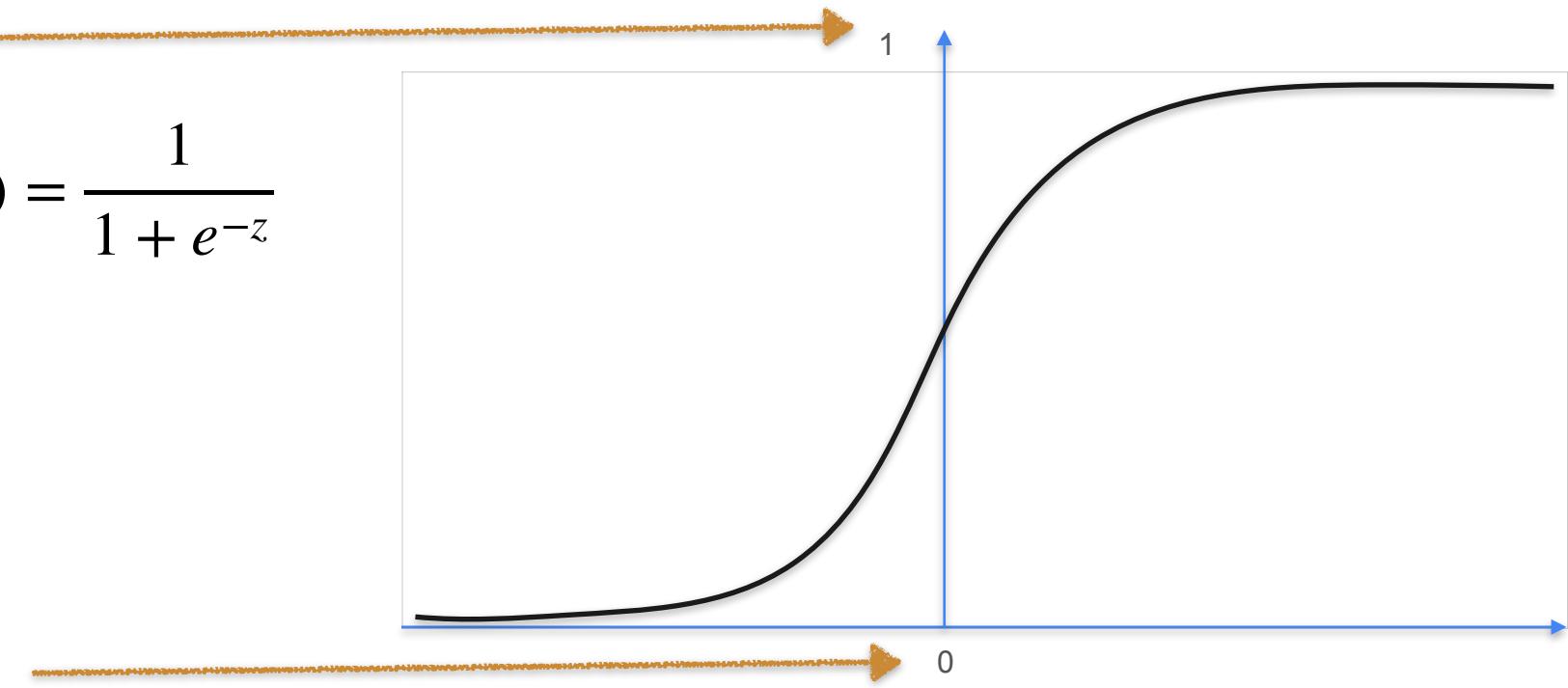
# Sigmoid Function

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



# Sigmoid Function

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



# Derivative of a Sigmoid Function

# Derivative of a Sigmoid Function

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

# Derivative of a Sigmoid Function

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

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

# Derivative of a Sigmoid Function

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

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

$$\frac{d}{dz} \sigma(z)$$

# Derivative of a Sigmoid Function

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

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

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

# Derivative of a Sigmoid Function

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

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

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



# Derivative of a Sigmoid Function

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

$$\frac{d}{dz}\sigma(z)$$

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

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

# Derivative of a Sigmoid Function

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

$$\frac{d}{dz} \sigma(z) = -1$$

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

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

# Derivative of a Sigmoid Function

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

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

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

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

# Derivative of a Sigmoid Function

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

$$\frac{d}{dz} \sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz} (1 + e^{-z}) \right)$$

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

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

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left(\frac{d}{dz}(1 + e^{-z})\right)$$

$$= -1$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left(\frac{d}{dz}(1 + e^{-z})\right)$$

$$= -1 (1 + e^{-z})^{-2}$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left(\frac{d}{dz}(1 + e^{-z})\right)$$

$$= -1 (1 + e^{-z})^{-2} \left(\frac{d}{dz}(1)\right)$$

# Derivative of a Sigmoid Function

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

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

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

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= -1 \ (1 + e^{-z})^{-1-1} \ (\frac{d}{dz}(1 + e^{-z})) \\ &= -1 \ (1 + e^{-z})^{-2} \ (\frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}))\end{aligned}$$

# Derivative of a Sigmoid Function

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

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

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

$$\begin{aligned} \frac{d}{dz}\sigma(z) &= -1 \ (1 + e^{-z})^{-1-1} \ (\frac{d}{dz}(1 + e^{-z})) \\ &= -1 \ (1 + e^{-z})^{-2} \ (\frac{d}{dz}(1) + \frac{d}{dz}(e^{-z})) \\ &= -1 \end{aligned}$$

# Derivative of a Sigmoid Function

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

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

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

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= -1 \ (1 + e^{-z})^{-1-1} \ (\frac{d}{dz}(1 + e^{-z})) \\ &= -1 \ (1 + e^{-z})^{-2} \ (\frac{d}{dz}(1) + \frac{d}{dz}(e^{-z})) \\ &= -1 \ (1 + e^{-z})^{-2}\end{aligned}$$

# Derivative of a Sigmoid Function

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

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

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

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= -1 \ (1 + e^{-z})^{-1-1} \ (\frac{d}{dz}(1 + e^{-z})) \\ &= -1 \ (1 + e^{-z})^{-2} \ (\frac{d}{dz}(1) + \frac{d}{dz}(e^{-z})) \\ &= -1 \ (1 + e^{-z})^{-2} \ (0\end{aligned}$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz}(1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \left( \frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} (0 + e^{-z}(\frac{d}{dz}(-z)))$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz}(1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \left( \frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} (0 + e^{-z}(\frac{d}{dz}(-z)))$$

$$= -1$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz}(1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \left( \frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} (0 + e^{-z}(\frac{d}{dz}(-z)))$$

$$= -1 (1 + e^{-z})^{-2}$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz}(1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \left( \frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} (0 + e^{-z}(\frac{d}{dz}(-z)))$$

$$= -1 (1 + e^{-z})^{-2} (e^{-z})$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz}(1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \left( \frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} (0 + e^{-z}(\frac{d}{dz}(-z)))$$

$$= -1 (1 + e^{-z})^{-2} (e^{-z}) (-1)$$

# Derivative of a Sigmoid Function

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

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

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

$$\frac{d}{dz}\sigma(z) = -1 (1 + e^{-z})^{-1-1} \left( \frac{d}{dz}(1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \left( \frac{d}{dz}(1) + \frac{d}{dz}(e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} (0 + e^{-z}(\frac{d}{dz}(-z)))$$

$$= -1 (1 + e^{-z})^{-2} (e^{-z}) (-1)$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz} \sigma(z) = -1 \cdot (1 + e^{-z})^{-2} \cdot (e^{-z}) \cdot (-1)$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz} \sigma(z) = \cancel{1} (1 + e^{-z})^{-2} (e^{-z}) (-1)$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz} \sigma(z) = \cancel{1} (1 + e^{-z})^{-2} (e^{-z}) \cancel{(-1)}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \cancel{1} (1 + e^{-z})^{-2} \ (e^{-z}) \cancel{(-1)} \\ &= (1 + e^{-z})^{-2}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \cancel{1} (1 + e^{-z})^{-2} \ (e^{-z}) \cancel{(-1)} \\ &= (1 + e^{-z})^{-2} \ (e^{-z})\end{aligned}$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz} \sigma(z) = \cancel{1} (1 + e^{-z})^{-2} (e^{-z}) \cancel{(-1)}$$

$$= (1 + e^{-z})^{-2} (e^{-z})$$

$$= \frac{1}{(1 + e^{-z})^2}$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz} \sigma(z) = \cancel{1} (1 + e^{-z})^{-2} (e^{-z}) \cancel{(-1)}$$

$$= (1 + e^{-z})^{-2} (e^{-z})$$

$$= \frac{1}{(1 + e^{-z})^2} (e^{-z})$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz} \sigma(z) = \cancel{1} (1 + e^{-z})^{-2} (e^{-z}) \cancel{(-1)}$$

$$= (1 + e^{-z})^{-2} (e^{-z})$$

$$= \frac{1}{(1 + e^{-z})^2} (e^{-z})$$

$$= \frac{e^{-z}}{(1 + e^{-z})^2}$$

# Derivative of a Sigmoid Function

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z)$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z}}{(1 + e^{-z})^2}$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2}$$

$$= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2}$$

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2}$$

$$= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2}$$

$$= \frac{1 + e^{-z}}{(1 + e^{-z})^2}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1 \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{1 + e^{-z}}{(1 + e^{-z})^2} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1 \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{1 + e^{-z}}{(1 + e^{-z})^2} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1 \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1 \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1 \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz} \sigma(z) &= \frac{e^{-z}}{(1 + e^{-z})^2} + 1 - 1 \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\frac{d}{dz}\sigma(z)$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

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

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2}$$

$$= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2}$$

$$= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2}$$

$$= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}$$

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

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2}$$

$$= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2}$$

$$= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2}$$

$$= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}$$

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

# Derivative of a Sigmoid Function

$$\frac{d}{dz}\sigma(z) = \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2}$$

$$= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2}$$

$$= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2}$$

$$= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}$$

$$\frac{d}{dz}\sigma(z) = \frac{1}{(1 + e^{-z})} - \left( \frac{1}{(1 + e^{-z})} \right)$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\frac{d}{dz}\sigma(z) = \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right)$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\ &= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\ &= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\ &= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\ &= \frac{1}{(1 + e^{-z})}\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\&= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\&= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\&= \frac{1}{(1 + e^{-z})} \left(1 - \frac{1}{(1 + e^{-z})}\right)\end{aligned}$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\&= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\&= \frac{\cancel{1 + e^{-z}}}{(1 + e^{-z})^2} - \frac{1}{(1 + e^{-z})^2} \\&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\&= \frac{1}{(1 + e^{-z})} \left(1 - \frac{1}{(1 + e^{-z})}\right)\end{aligned}$$

Recall that:

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\&= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\&= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\&= \frac{1}{(1 + e^{-z})} \left(1 - \frac{1}{(1 + e^{-z})}\right)\end{aligned}$$

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

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\&= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\&= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\&= \frac{1}{(1 + e^{-z})} \left(1 - \frac{1}{(1 + e^{-z})}\right)\end{aligned}$$

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

$$\frac{d}{dz}\sigma(z)$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\&= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\&= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\&= \frac{1}{(1 + e^{-z})} \left(1 - \frac{1}{(1 + e^{-z})}\right)\end{aligned}$$

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

$$\frac{d}{dz}\sigma(z) = \sigma(z)$$

# Derivative of a Sigmoid Function

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\&= \frac{1 + e^{-z} - 1}{(1 + e^{-z})^2} \\&= \frac{\cancel{1 + e^{-z}}}{(\cancel{1 + e^{-z}})^2} - \frac{1}{(1 + e^{-z})^2} \\&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2}\end{aligned}$$

$$\begin{aligned}\frac{d}{dz}\sigma(z) &= \frac{1}{(1 + e^{-z})} - \left(\frac{1}{(1 + e^{-z})}\right)\left(\frac{1}{(1 + e^{-z})}\right) \\&= \frac{1}{(1 + e^{-z})} \left(1 - \frac{1}{(1 + e^{-z})}\right)\end{aligned}$$

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

$$\frac{d}{dz}\sigma(z) = \sigma(z) (1 - \sigma(z))$$



DeepLearning.AI

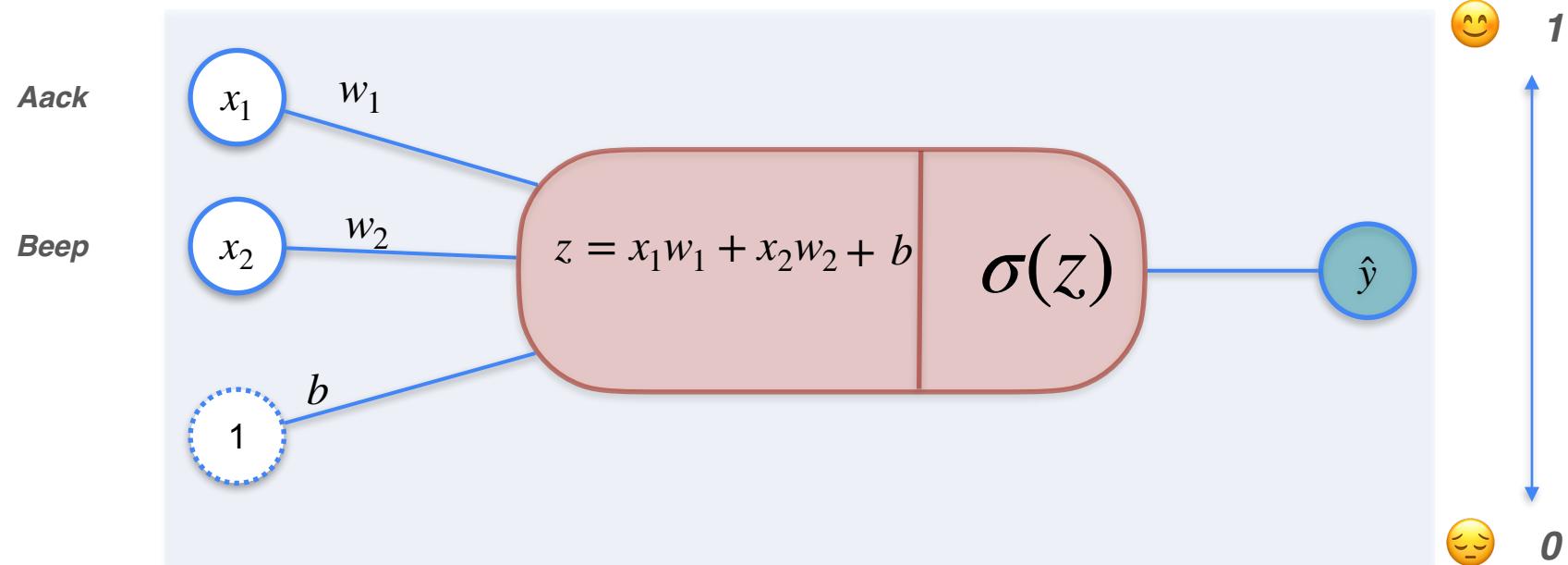
# Optimization in Neural Networks and Newton's Method

---

**Classification with a  
perceptron:  
Gradient Descent**

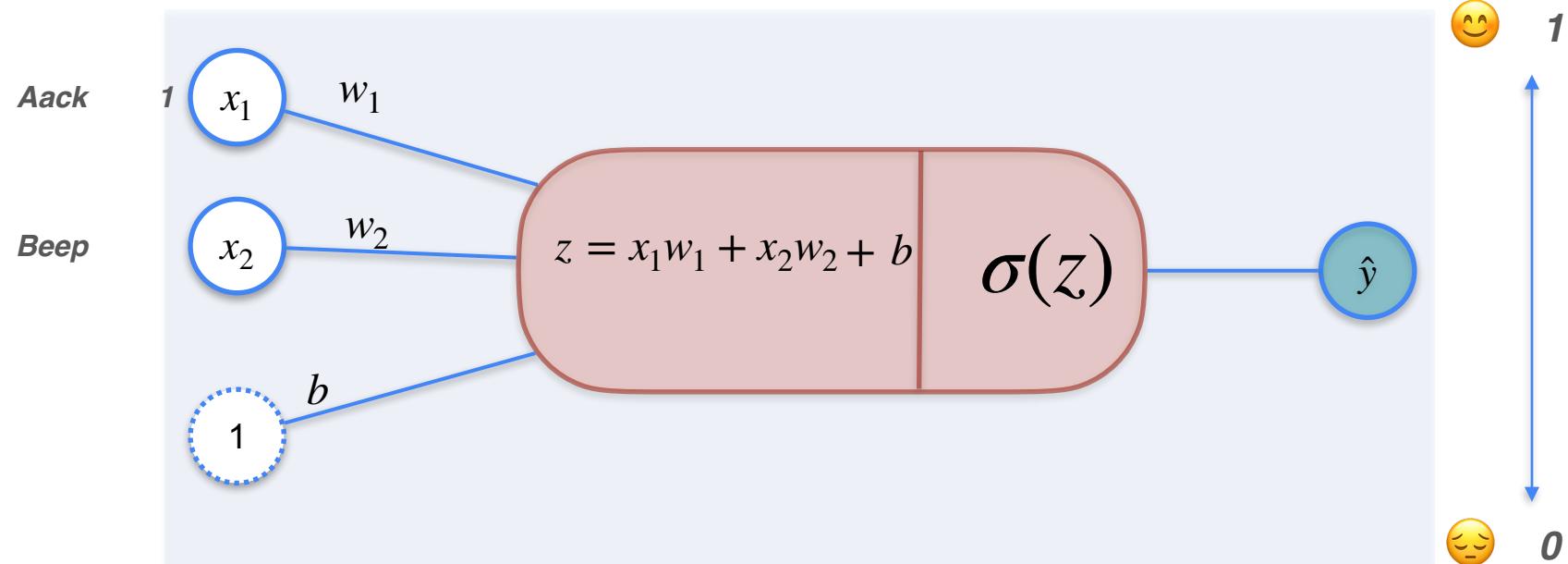
# Classification With a Perceptron

Aack beep beep beep



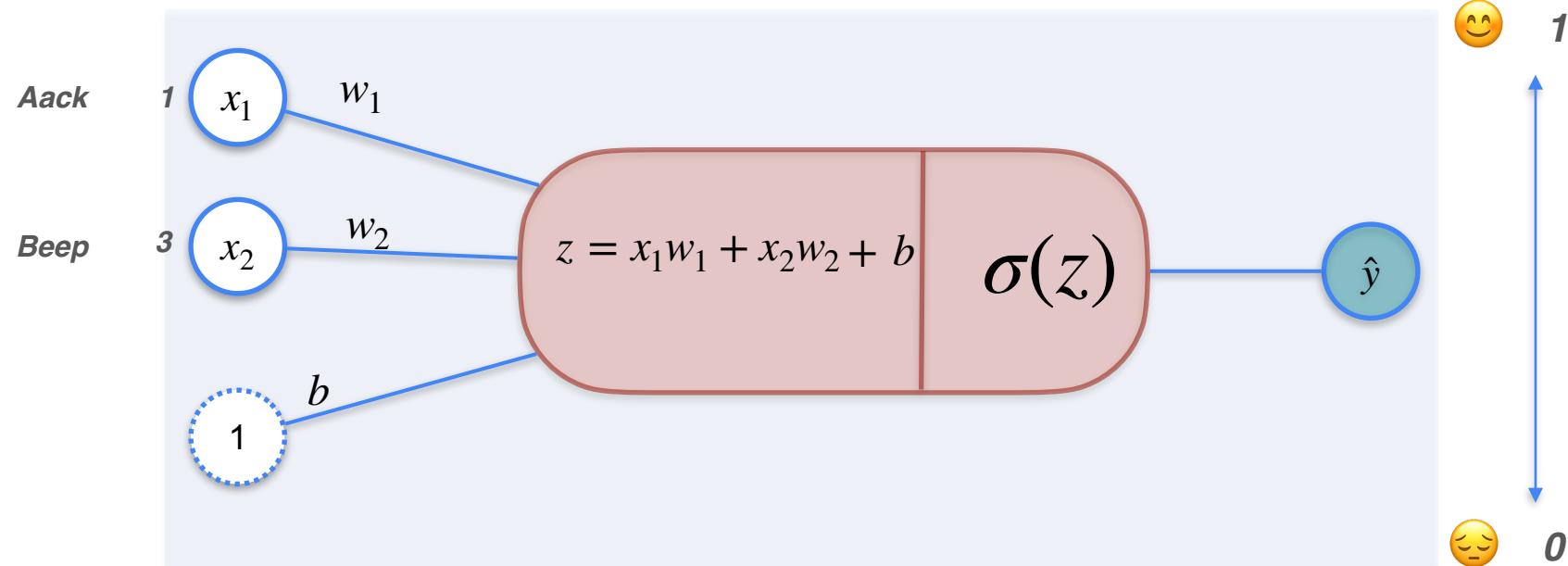
# Classification With a Perceptron

Aack beep beep beep



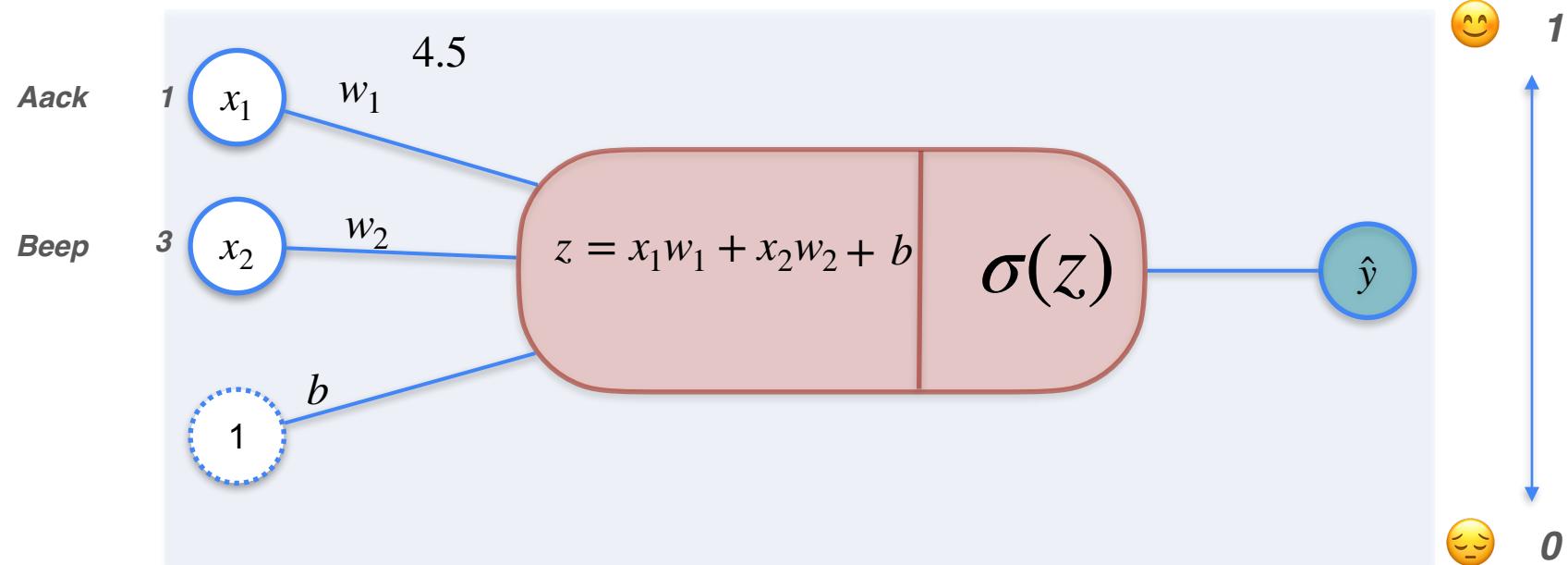
# Classification With a Perceptron

Aack beep beep beep



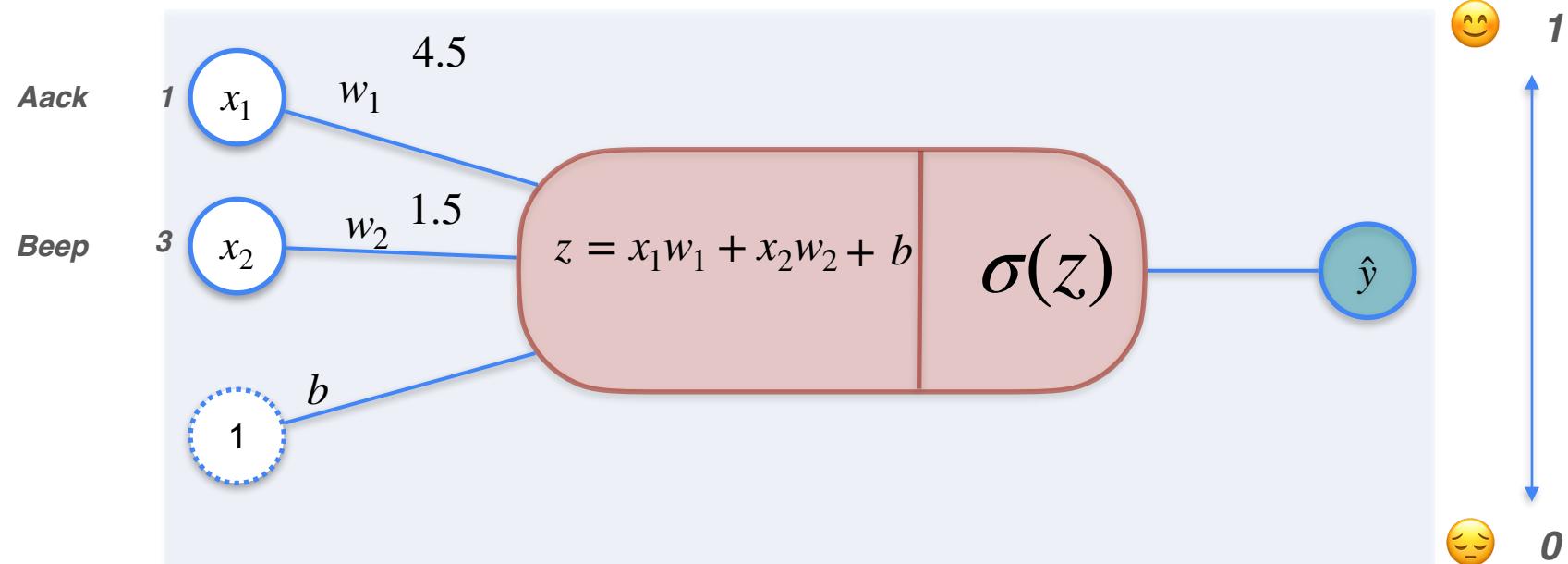
# Classification With a Perceptron

Aack beep beep beep



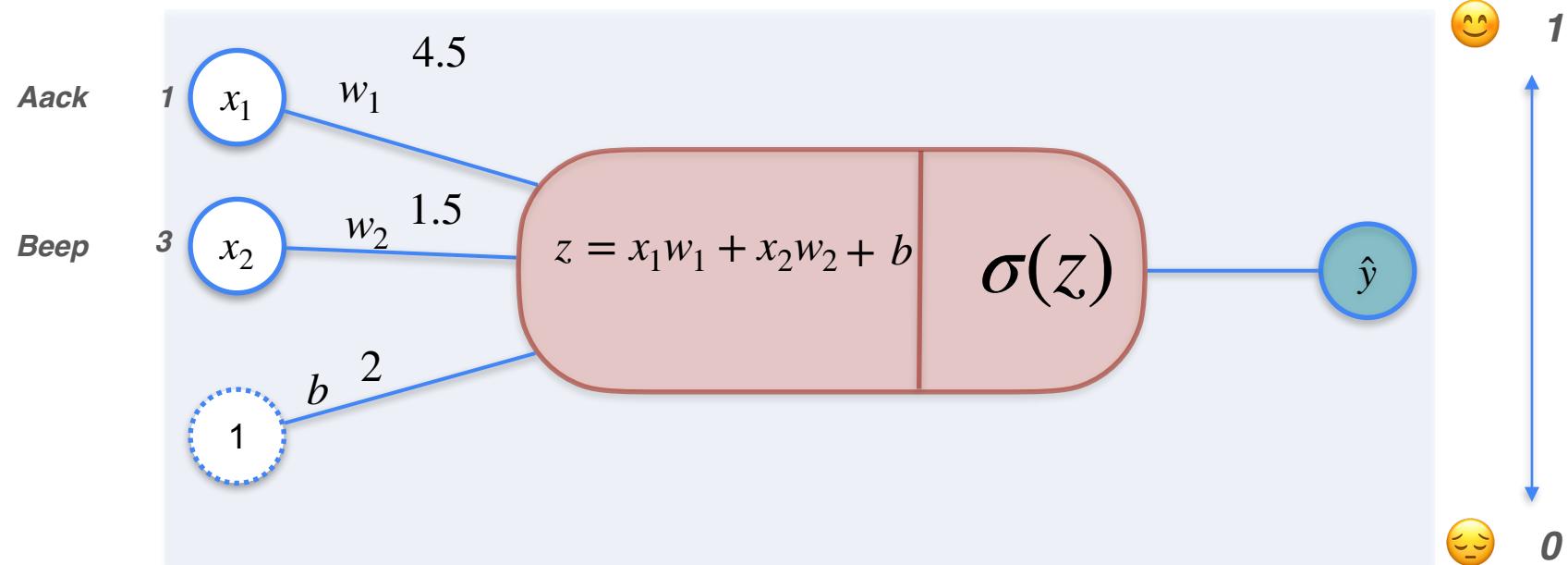
# Classification With a Perceptron

Aack beep beep beep



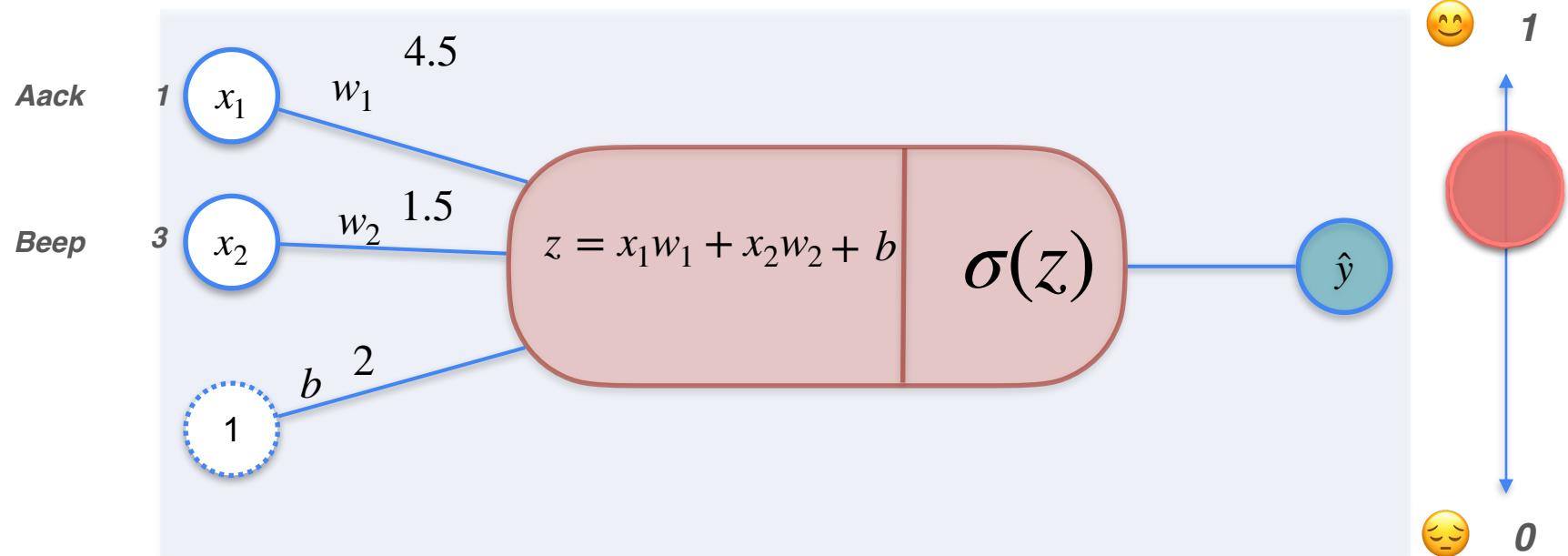
# Classification With a Perceptron

Aack beep beep beep



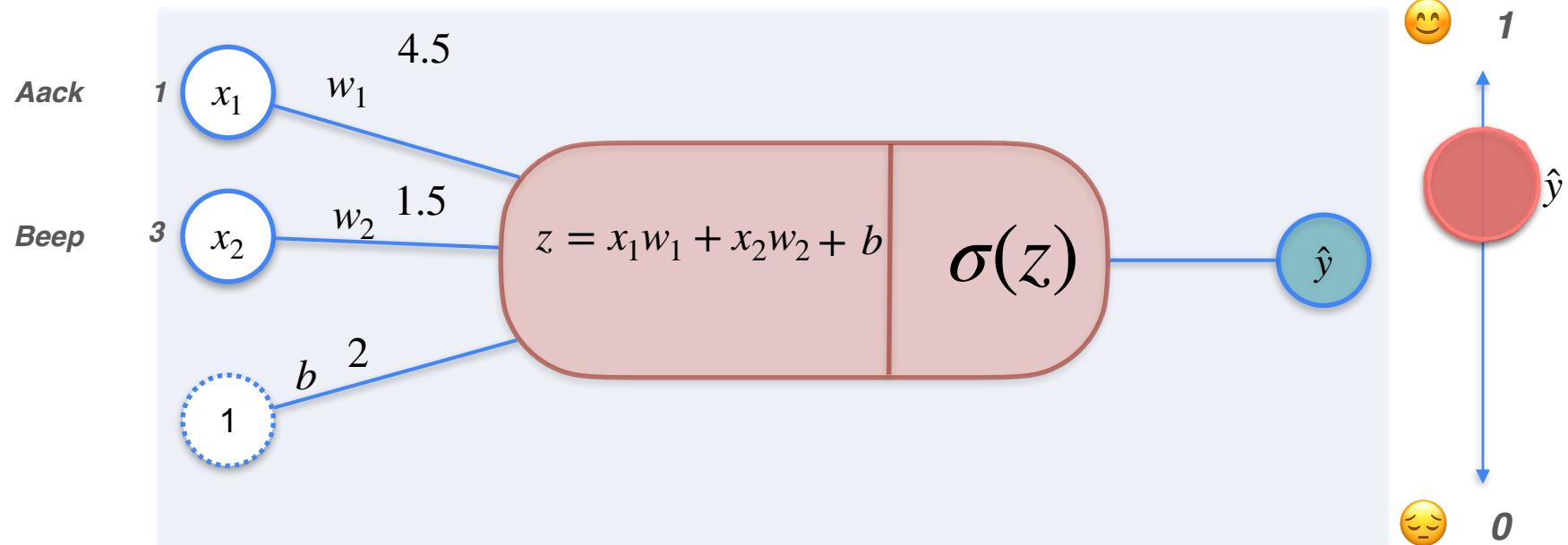
# Classification With a Perceptron

Aack beep beep beep



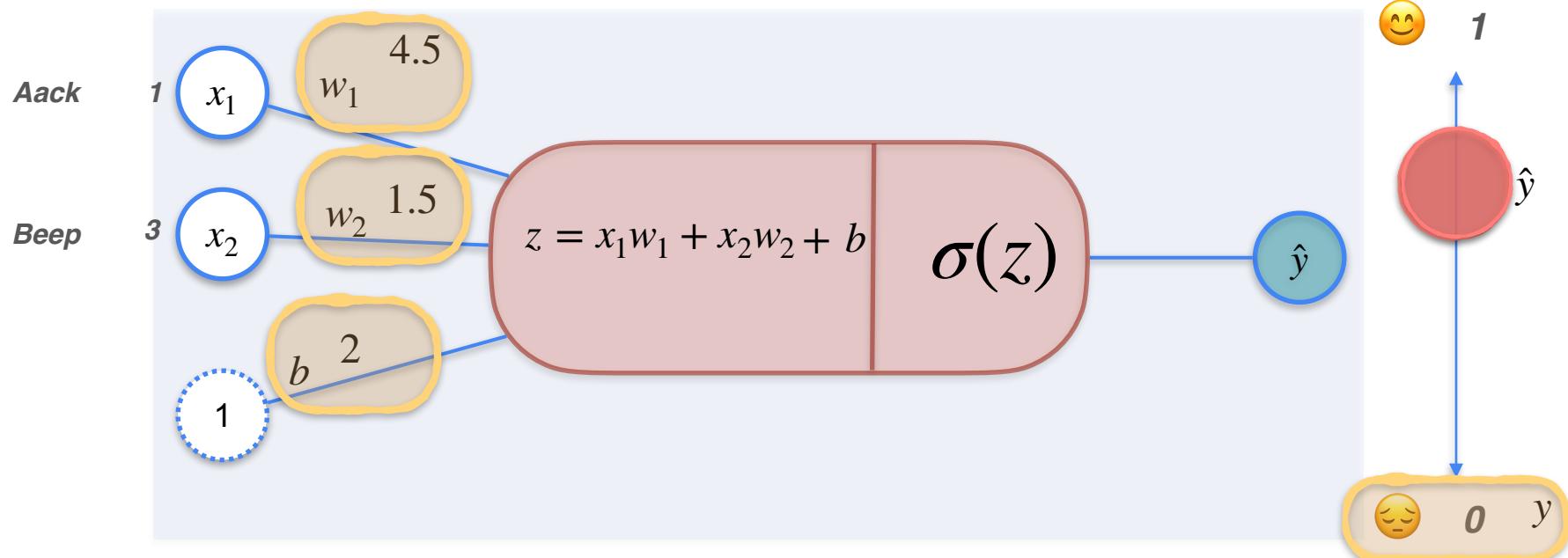
# Classification With a Perceptron

Aack beep beep beep



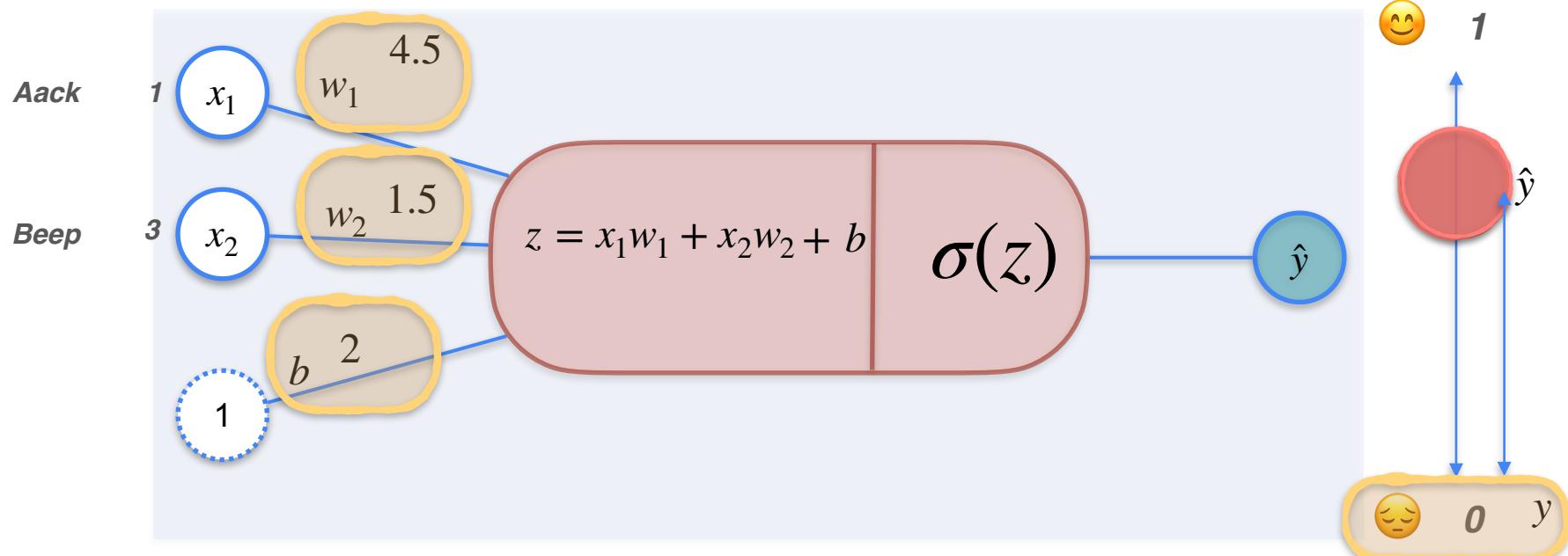
# Classification With a Perceptron

Aack beep beep beep



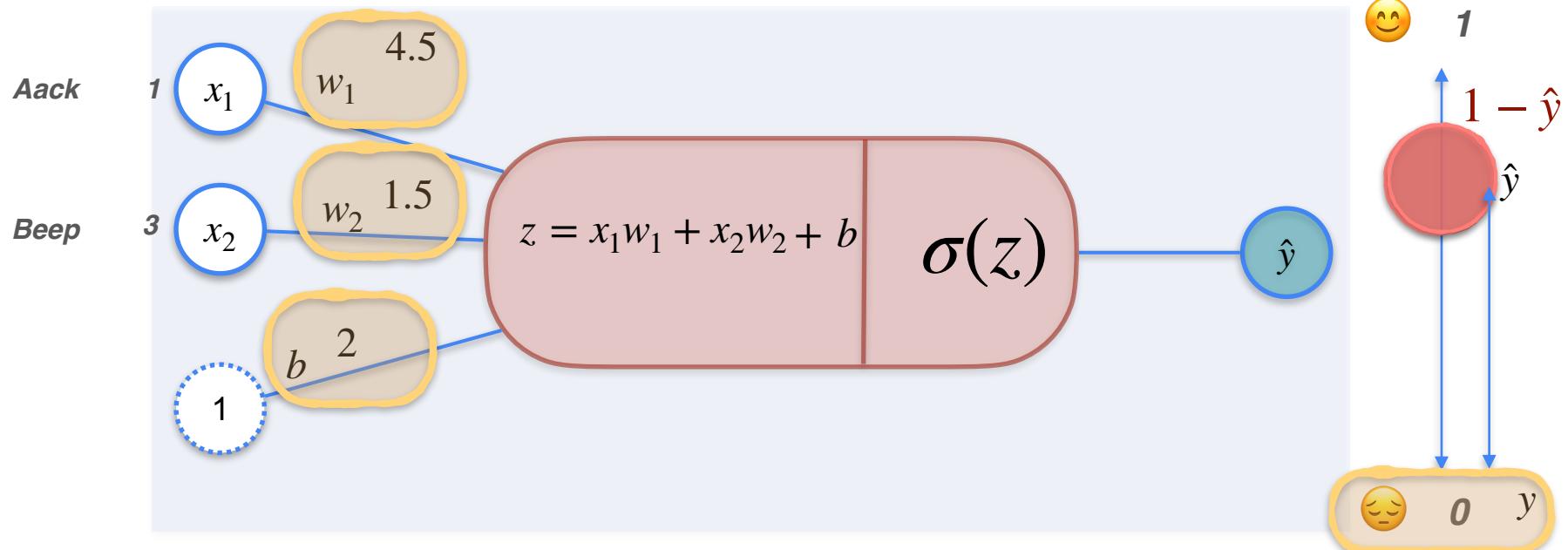
# Classification With a Perceptron

Aack beep beep beep



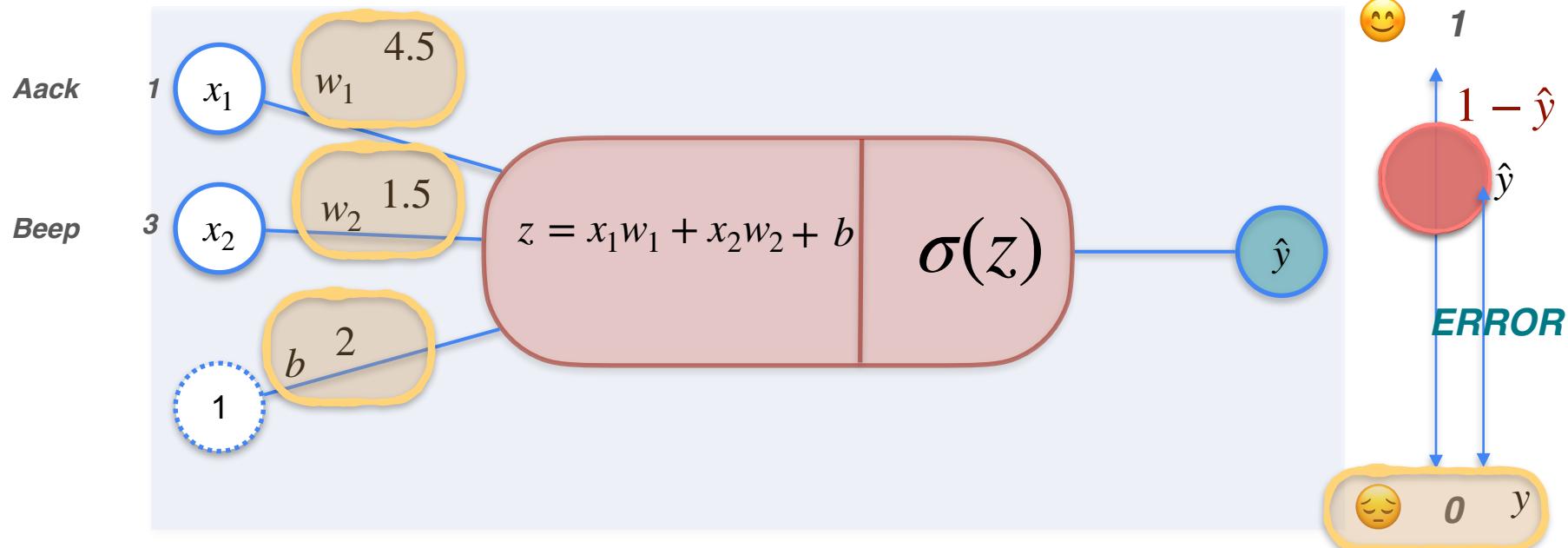
# Classification With a Perceptron

Aack beep beep beep



# Classification With a Perceptron

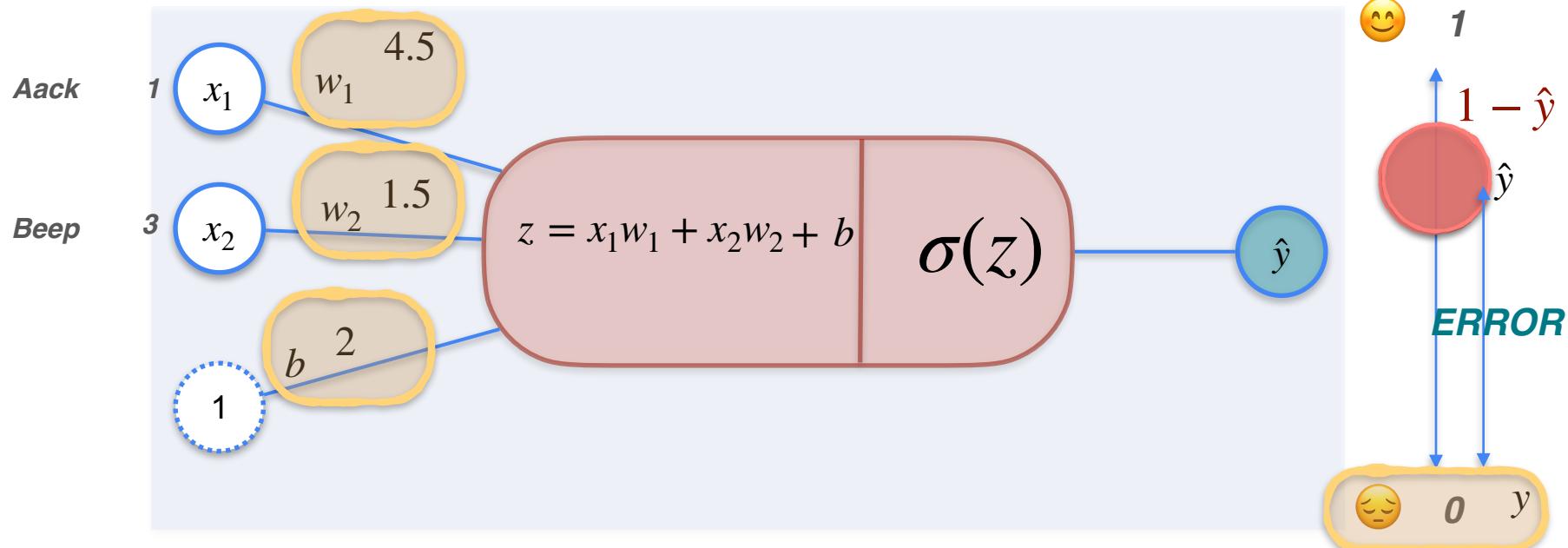
Aack beep beep beep



# Classification With a Perceptron

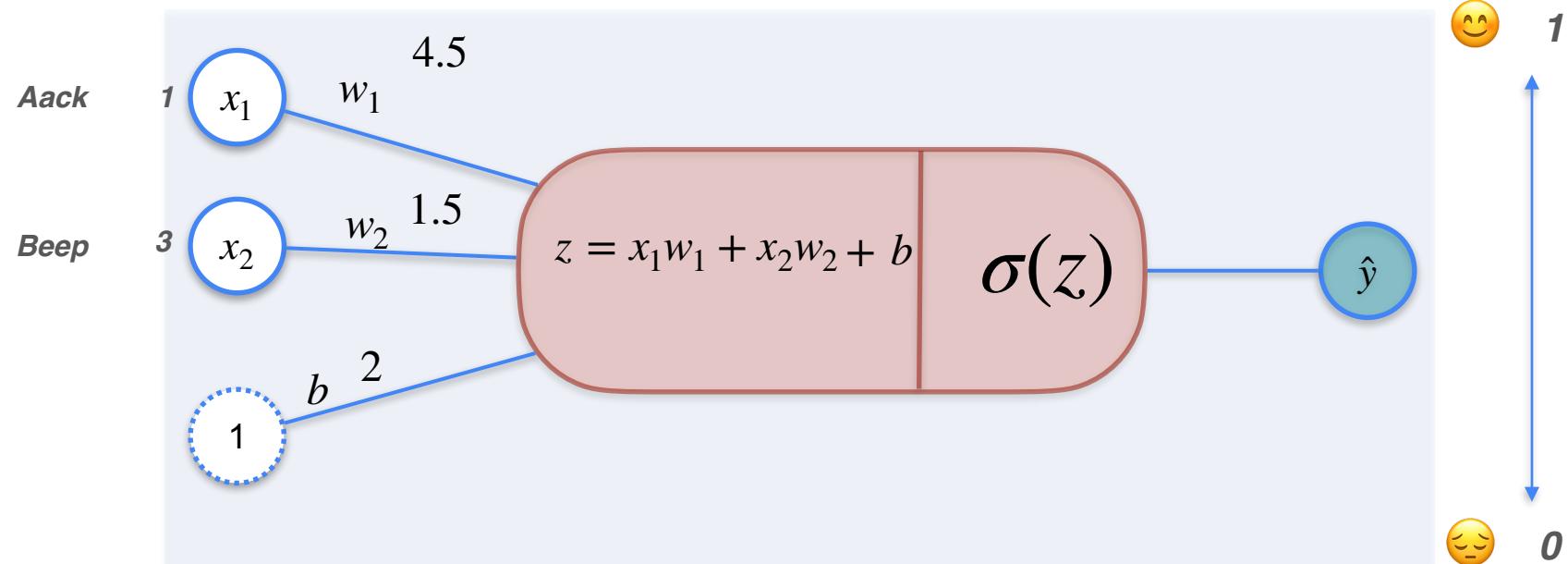
Aack beep beep beep

**LOG LOSS**



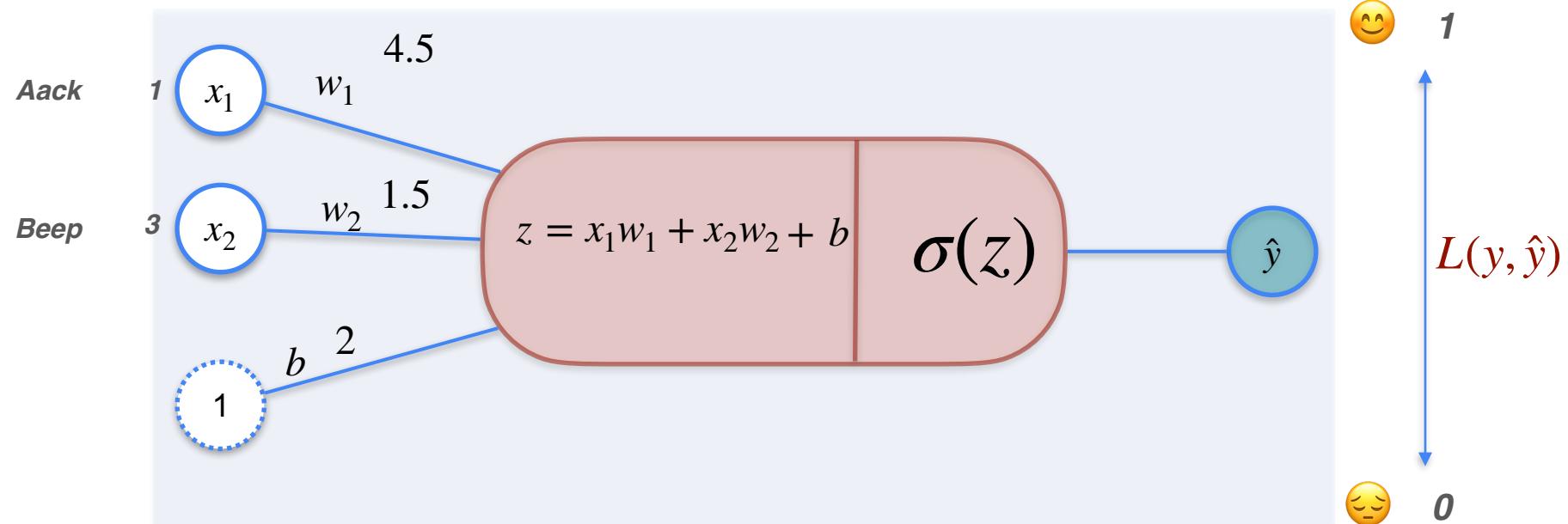
# Classification With a Perceptron

Aack beep beep beep



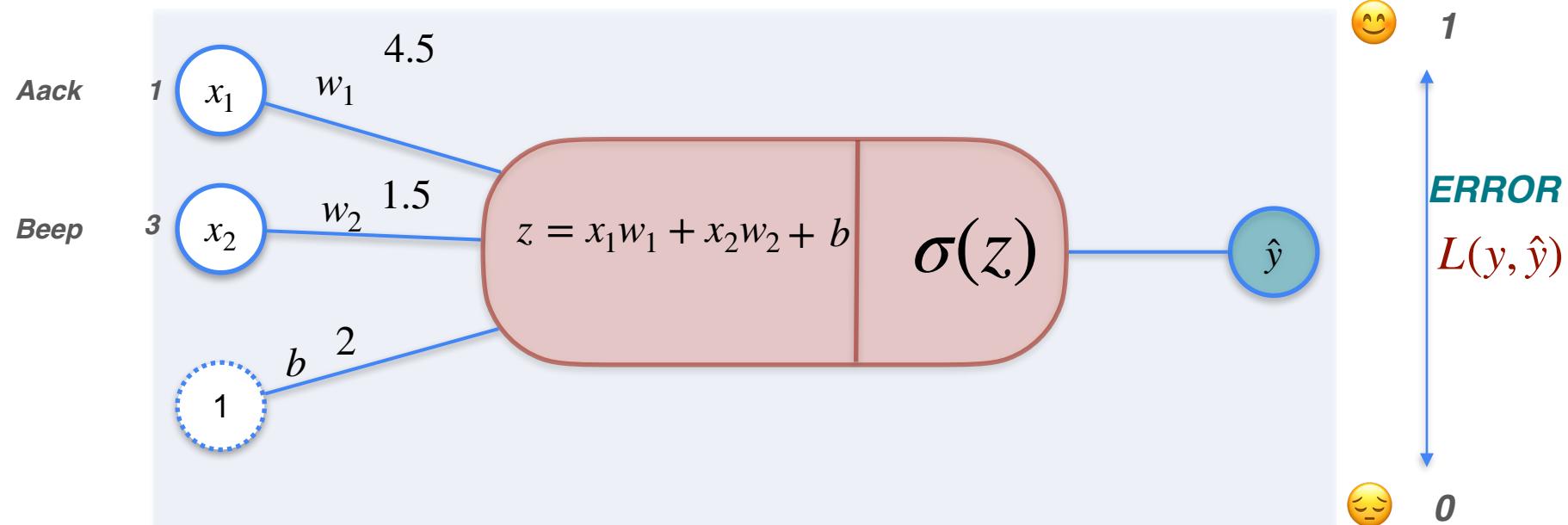
# Classification With a Perceptron

Aack beep beep beep



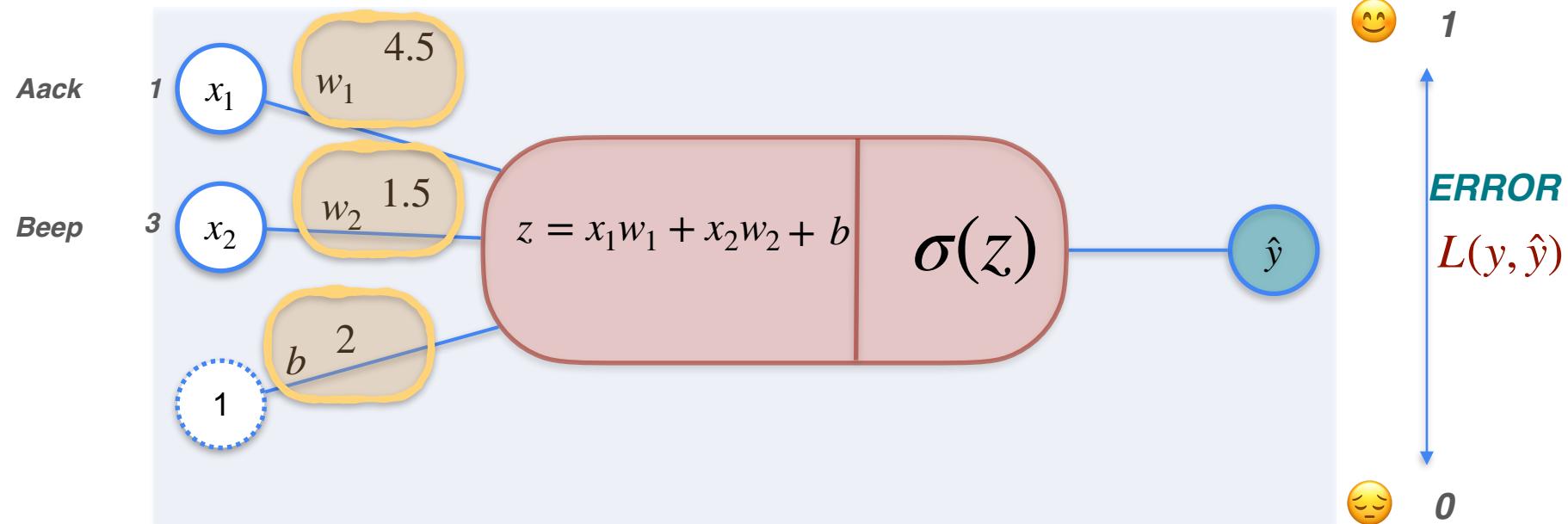
# Classification With a Perceptron

Aack beep beep beep



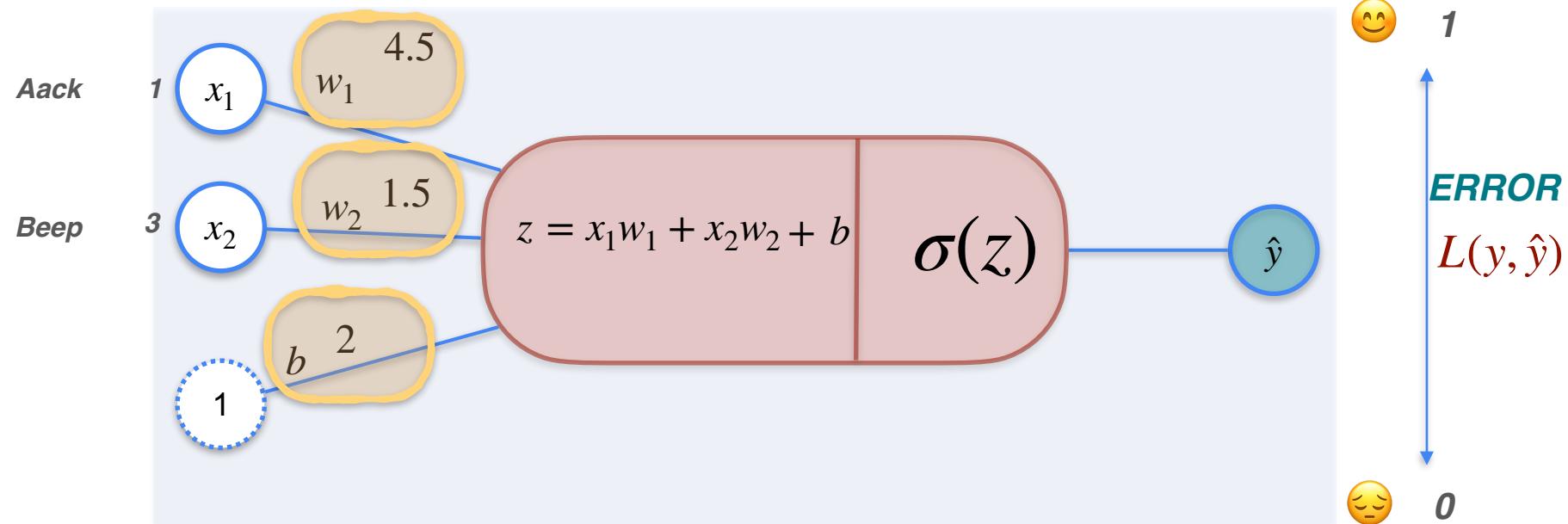
# Classification With a Perceptron

Aack beep beep beep

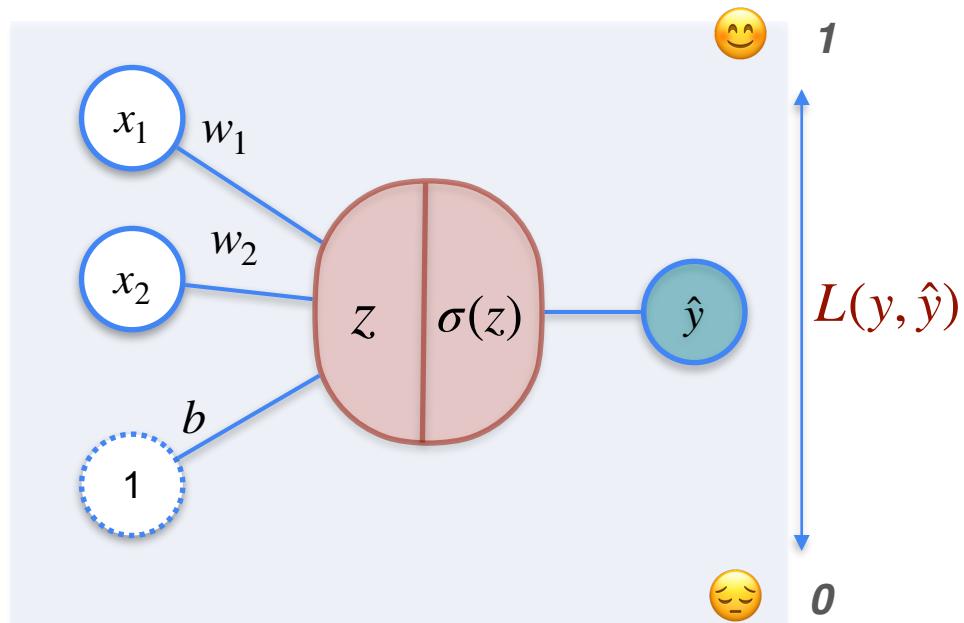


# Classification With a Perceptron

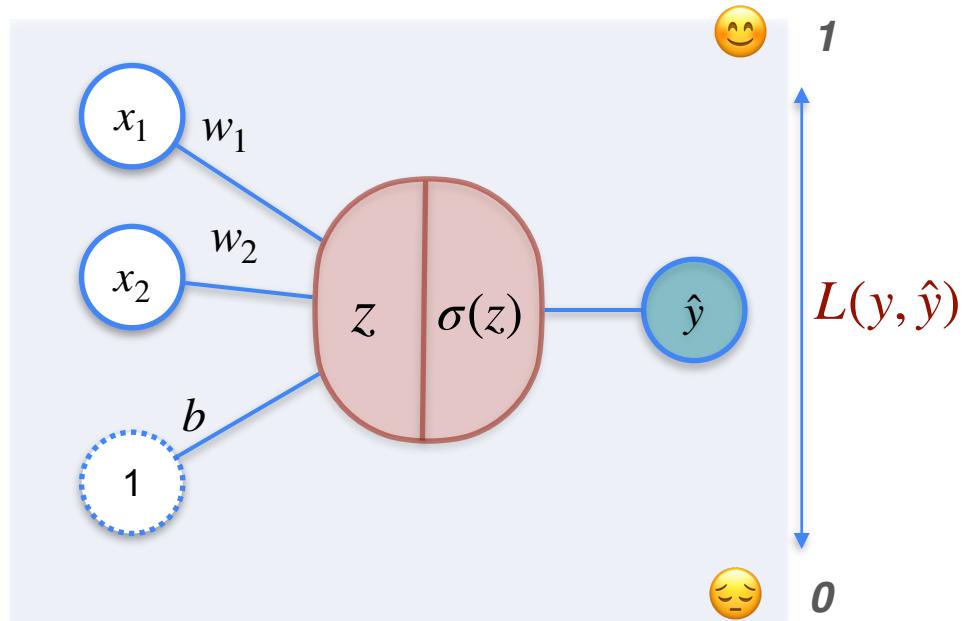
Aack beep beep beep



# Classification With a Perceptron



# Classification With a Perceptron



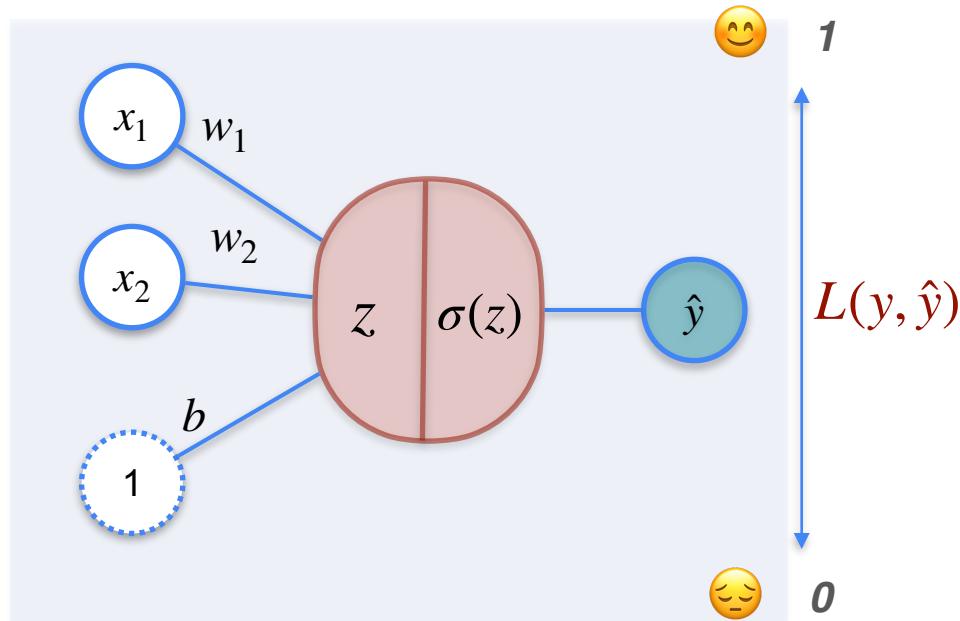
Prediction Function:

$$\hat{y}$$

$$L(y, \hat{y})$$

0

# Classification With a Perceptron

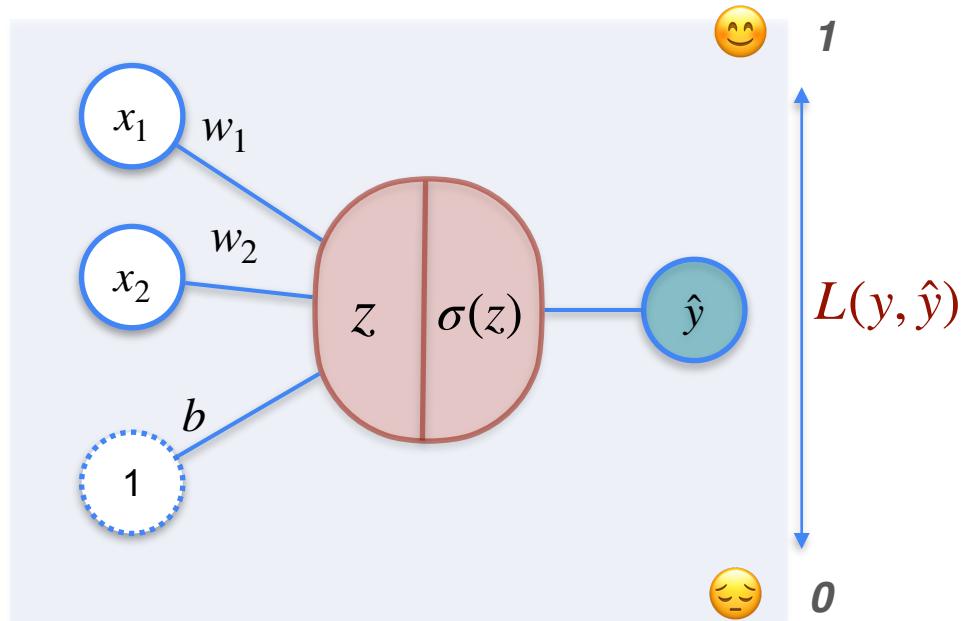


**Prediction Function:**

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$$L(y, \hat{y})$$

# Classification With a Perceptron



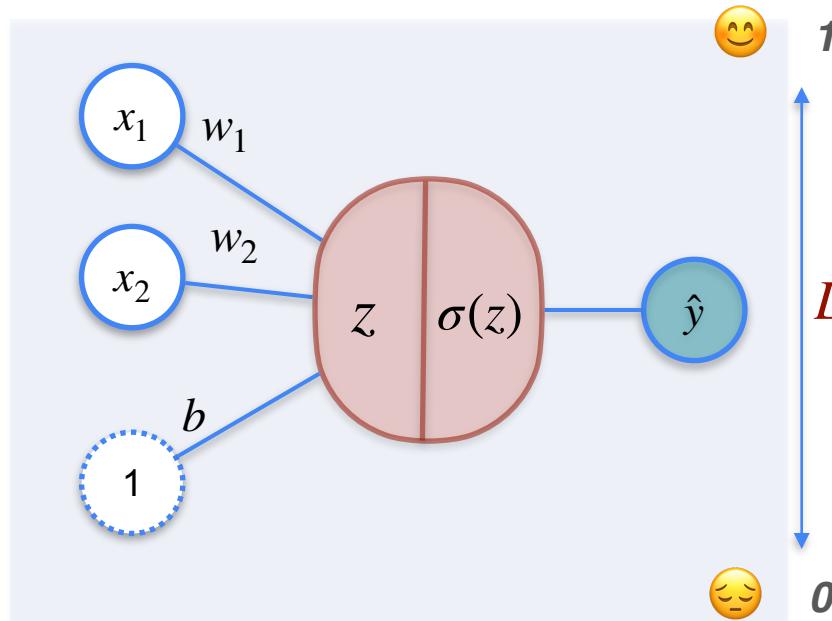
**Prediction Function:**

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

**Loss Function:**

$$L(y, \hat{y})$$

# Classification With a Perceptron



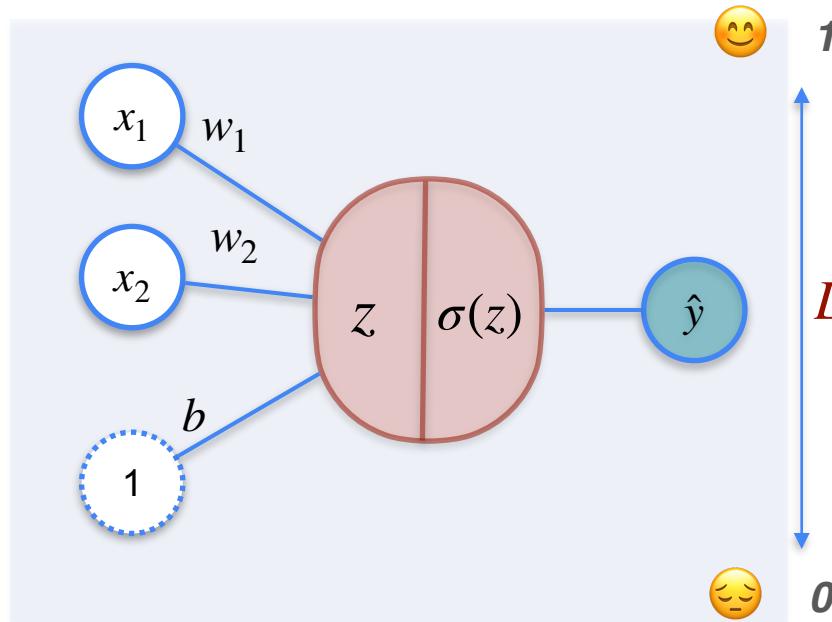
**Prediction Function:**

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

**Loss Function:**

$$L(y, \hat{y}) \quad L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

# Classification With a Perceptron



**Prediction Function:**

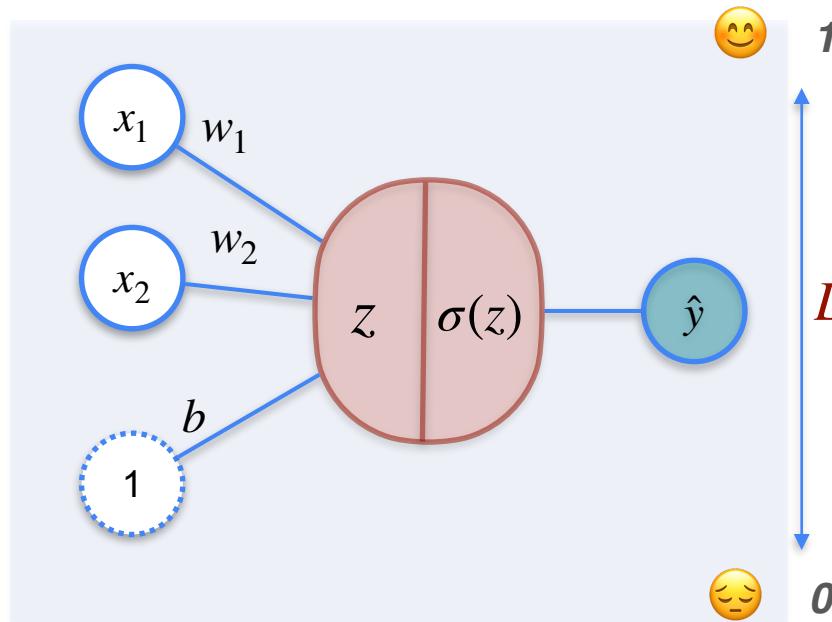
$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

**Loss Function:**

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

**Main Goal:**

# Classification With a Perceptron



**Prediction Function:**

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

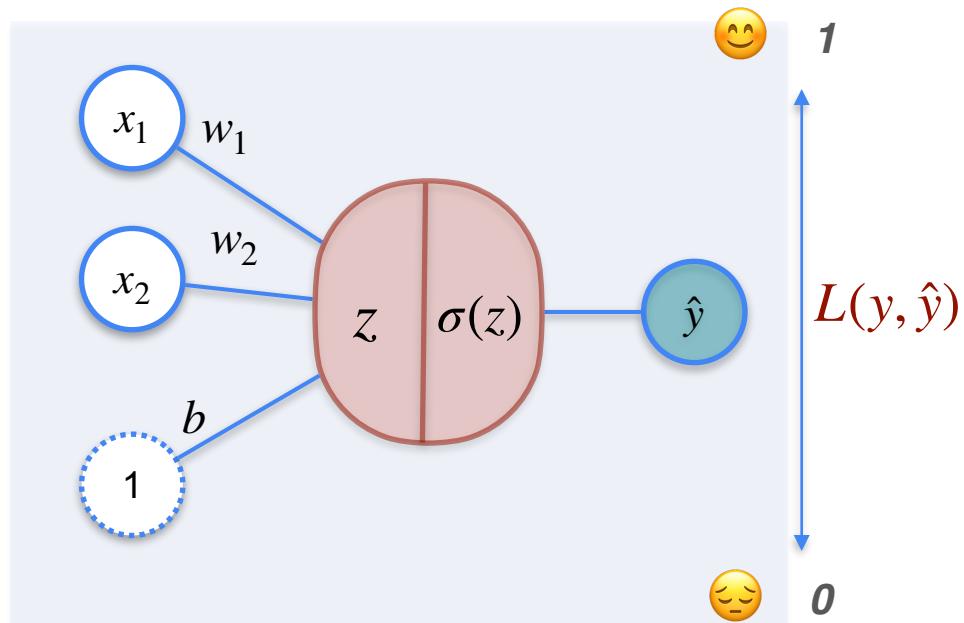
**Loss Function:**

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

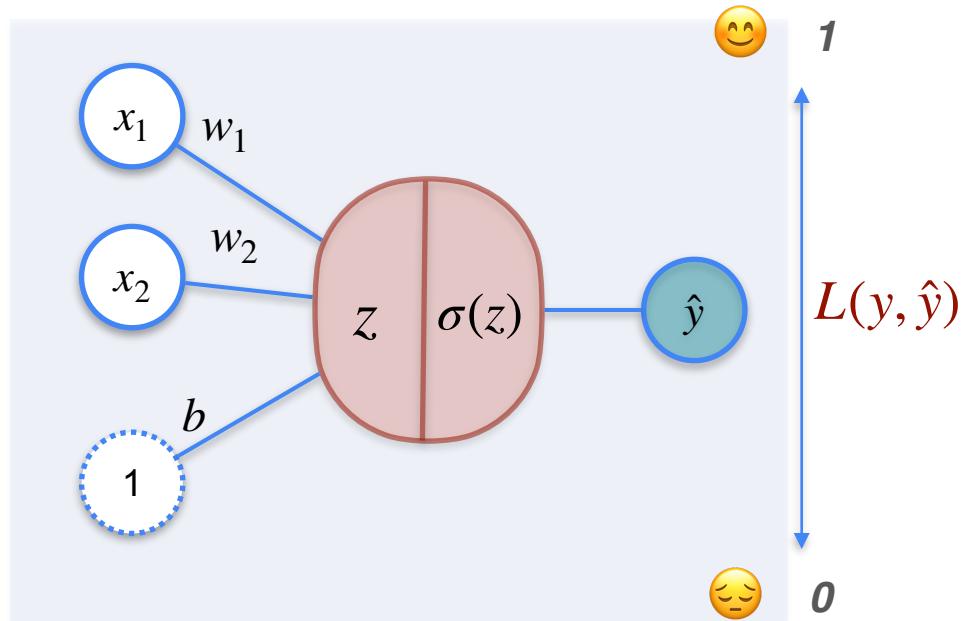
**Main Goal:**

Find  $w_1, w_2, b$  that give  $\hat{y}$  with the least error

# Classification With a Perceptron

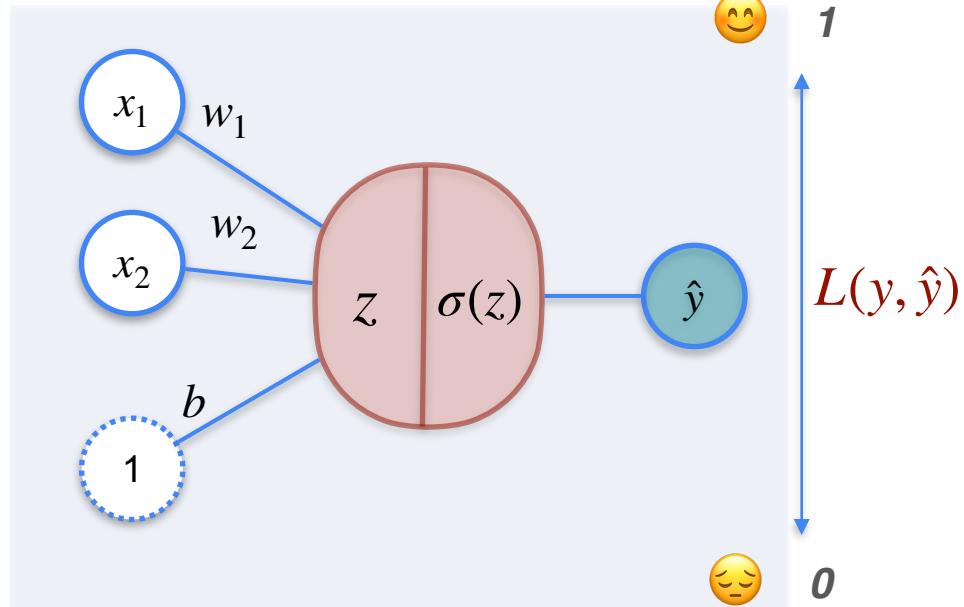


# Classification With a Perceptron



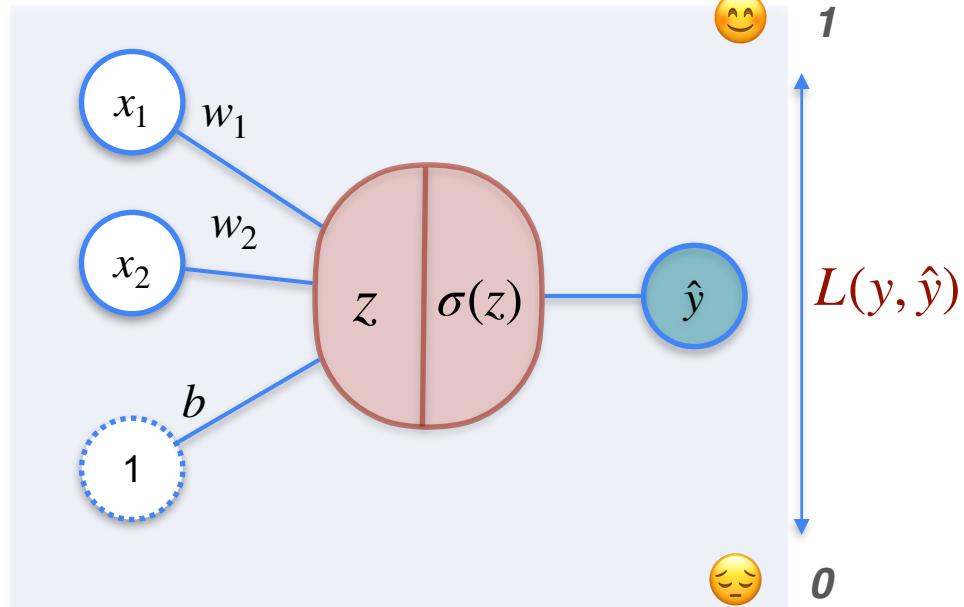
To find optimal values for:

# Classification With a Perceptron



To find optimal values for:  
 $w_1$  ,  $w_2$  ,  $b$

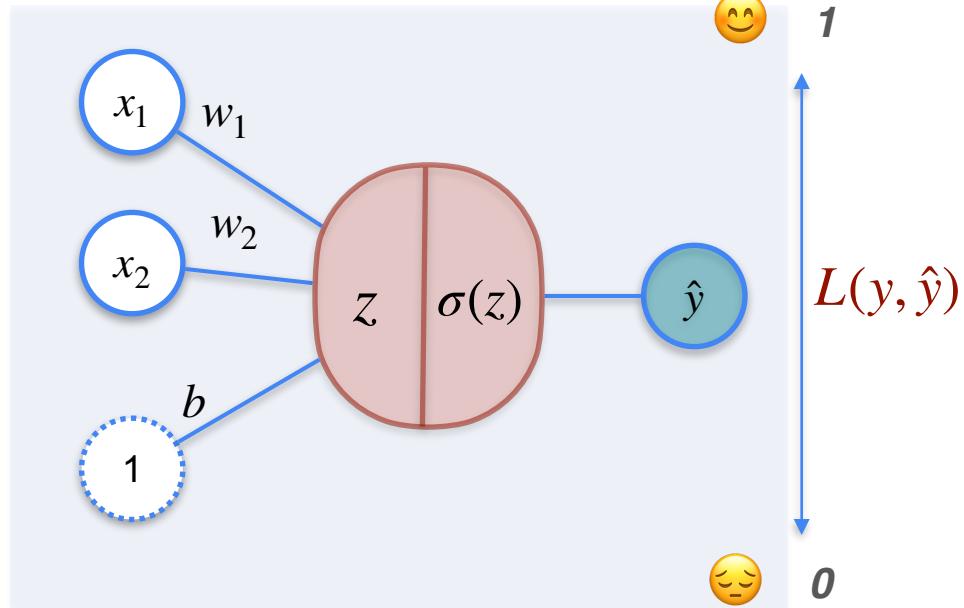
# Classification With a Perceptron



To find optimal values for:  
 $w_1$  ,  $w_2$  ,  $b$

*You need gradient descent*

# Classification With a Perceptron

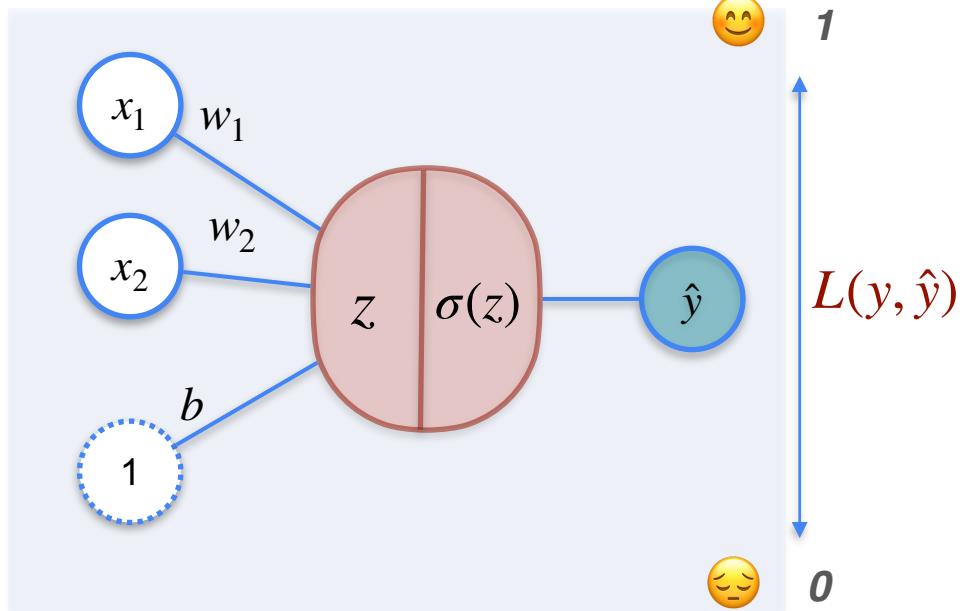


To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

# Classification With a Perceptron



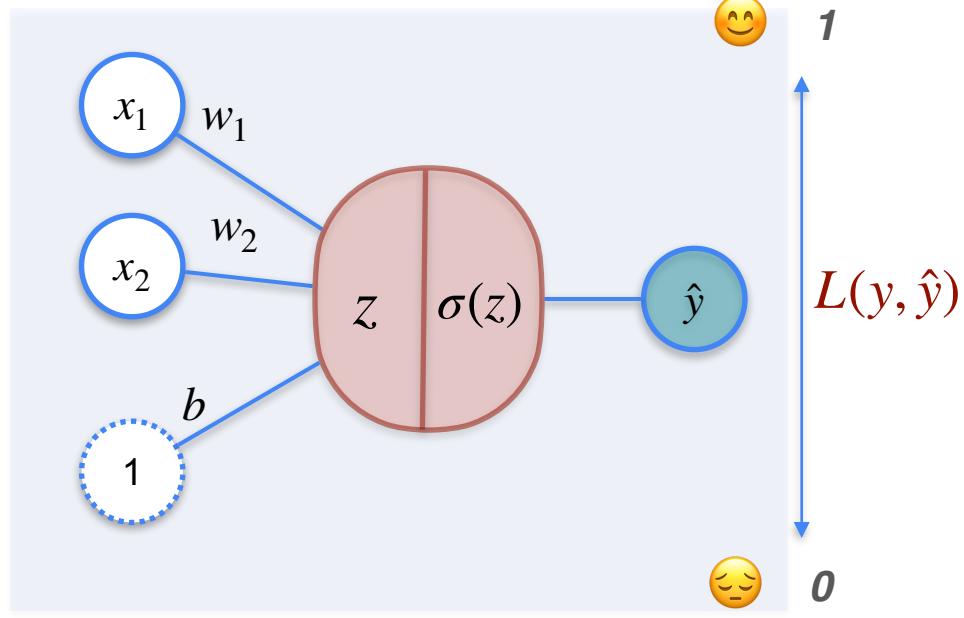
To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

# Classification With a Perceptron



To find optimal values for:  
 $w_1, w_2, b$

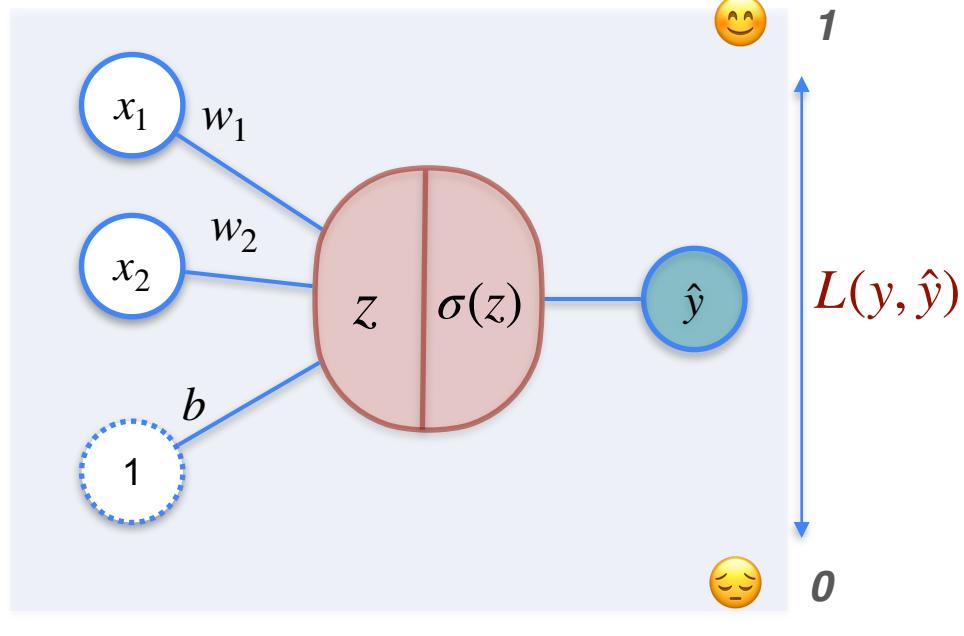
*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

# Classification With a Perceptron



To find optimal values for:  
 $w_1, w_2, b$

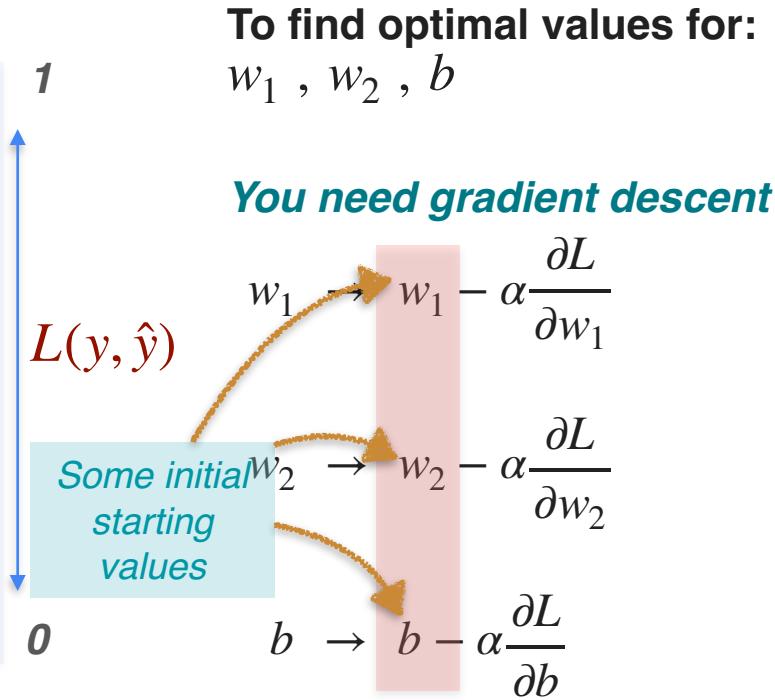
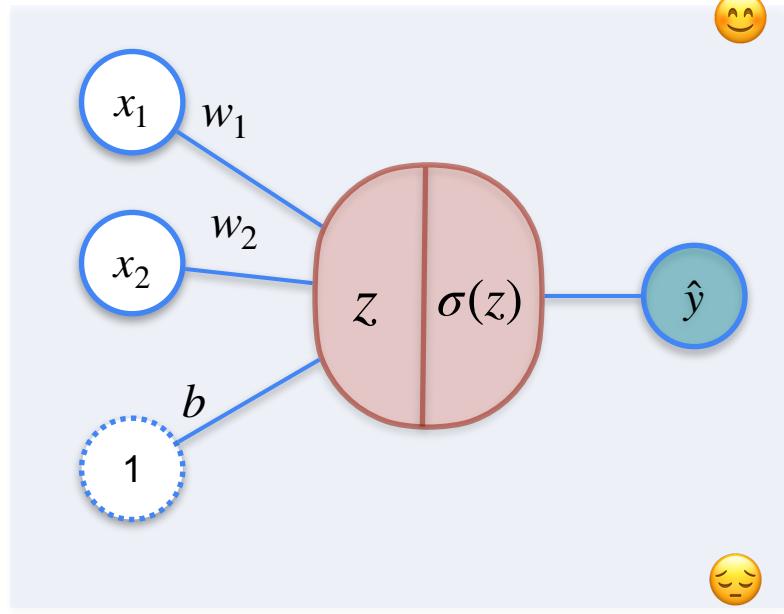
*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

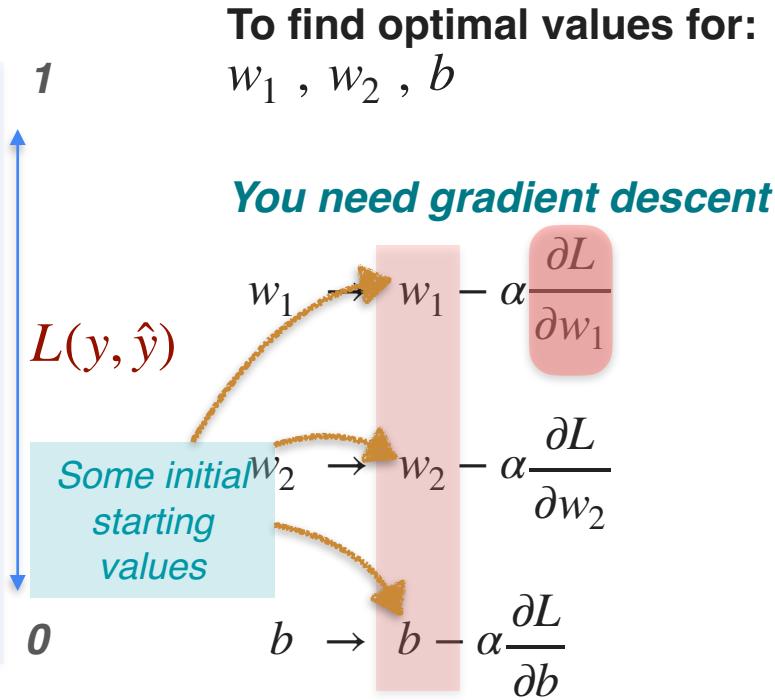
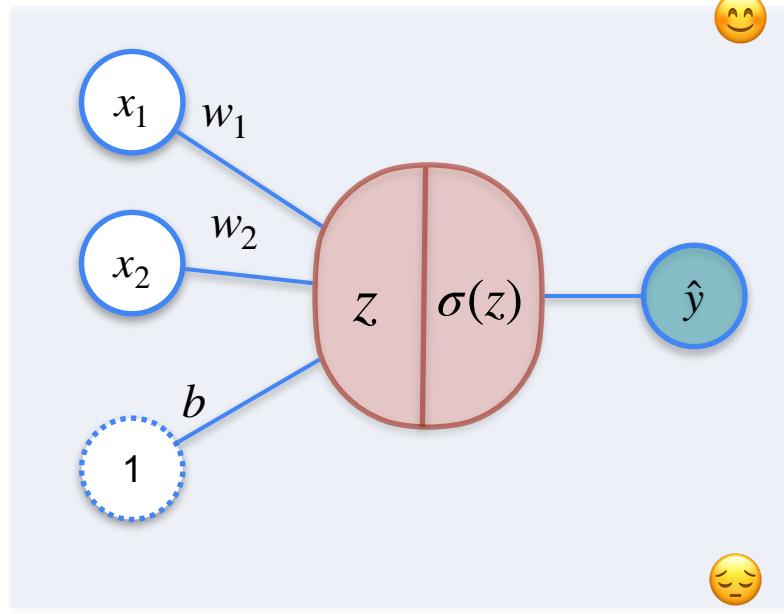
$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

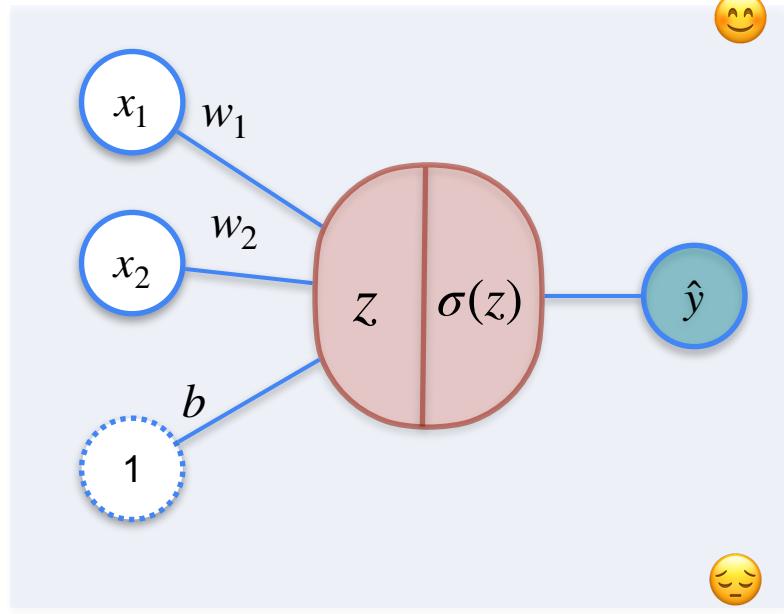
# Classification With a Perceptron



# Classification With a Perceptron

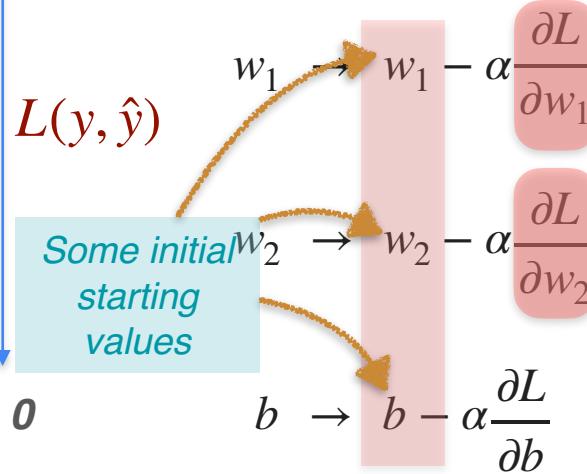


# Classification With a Perceptron

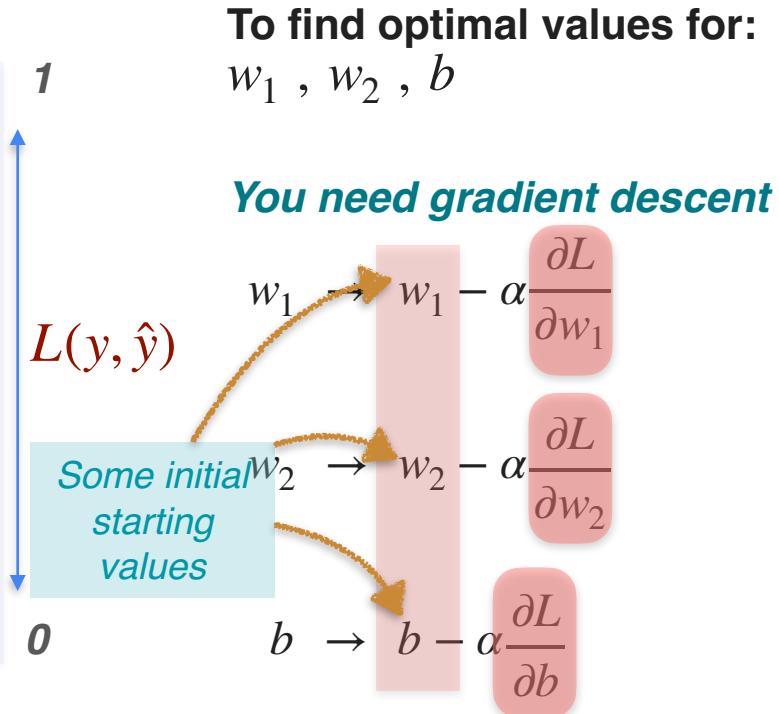
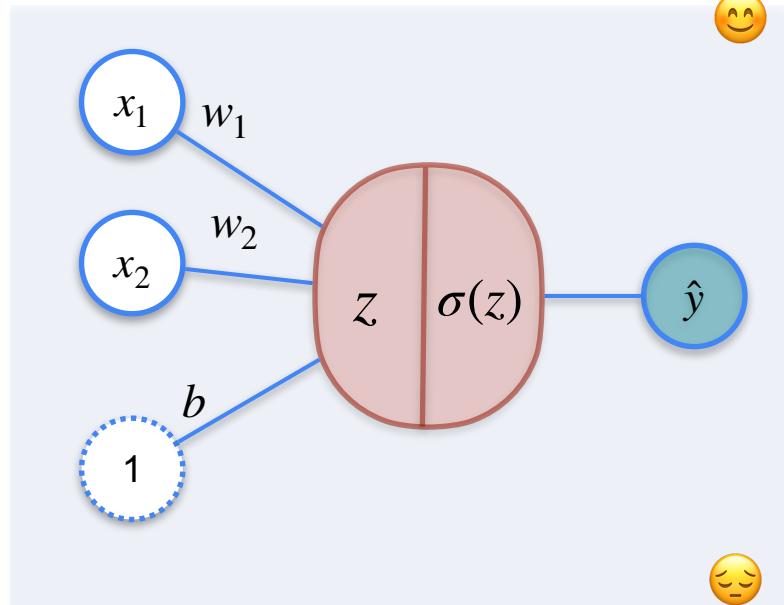


To find optimal values for:  
 $w_1, w_2, b$

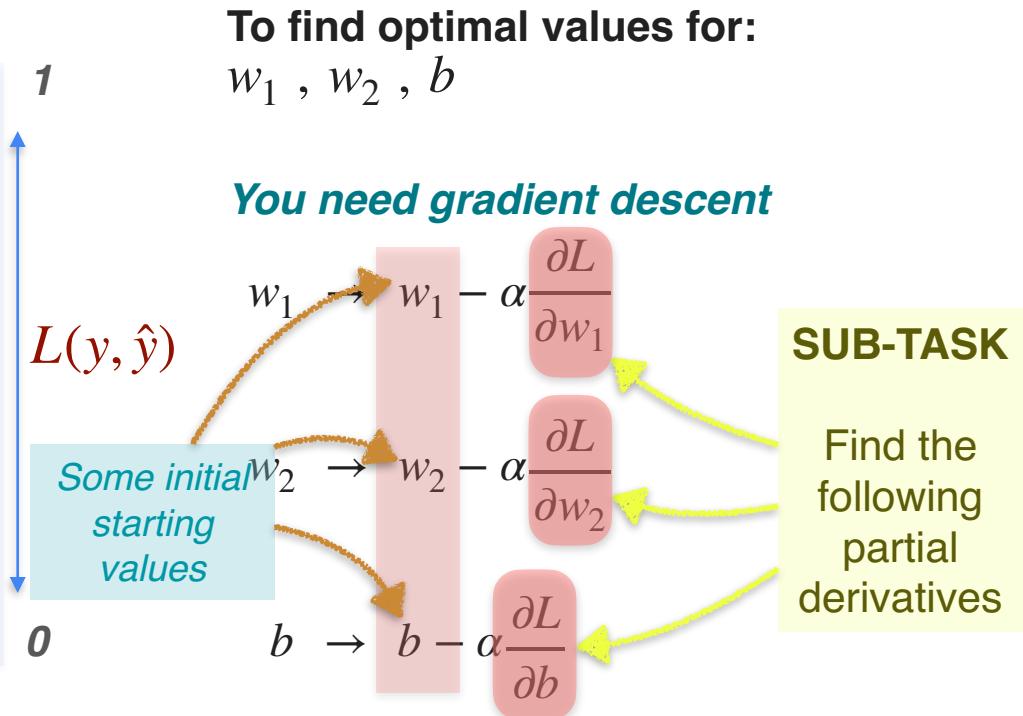
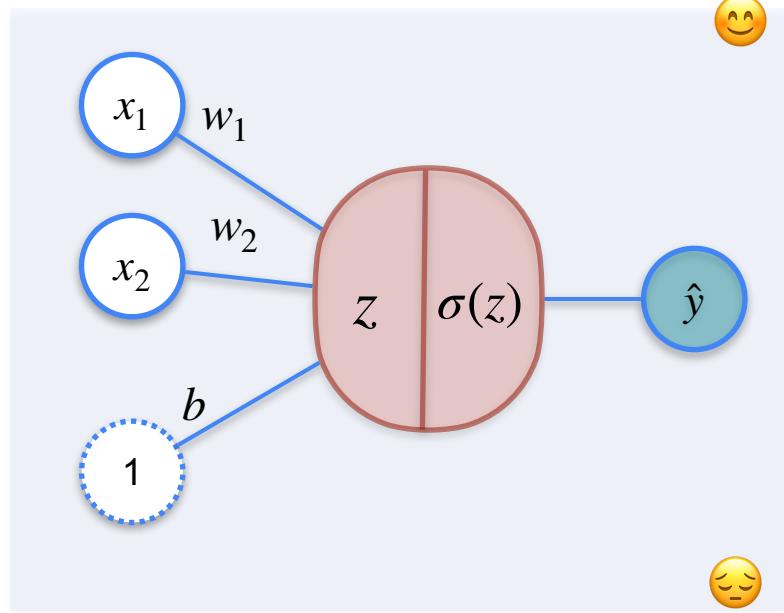
*You need gradient descent*



# Classification With a Perceptron



# Classification With a Perceptron





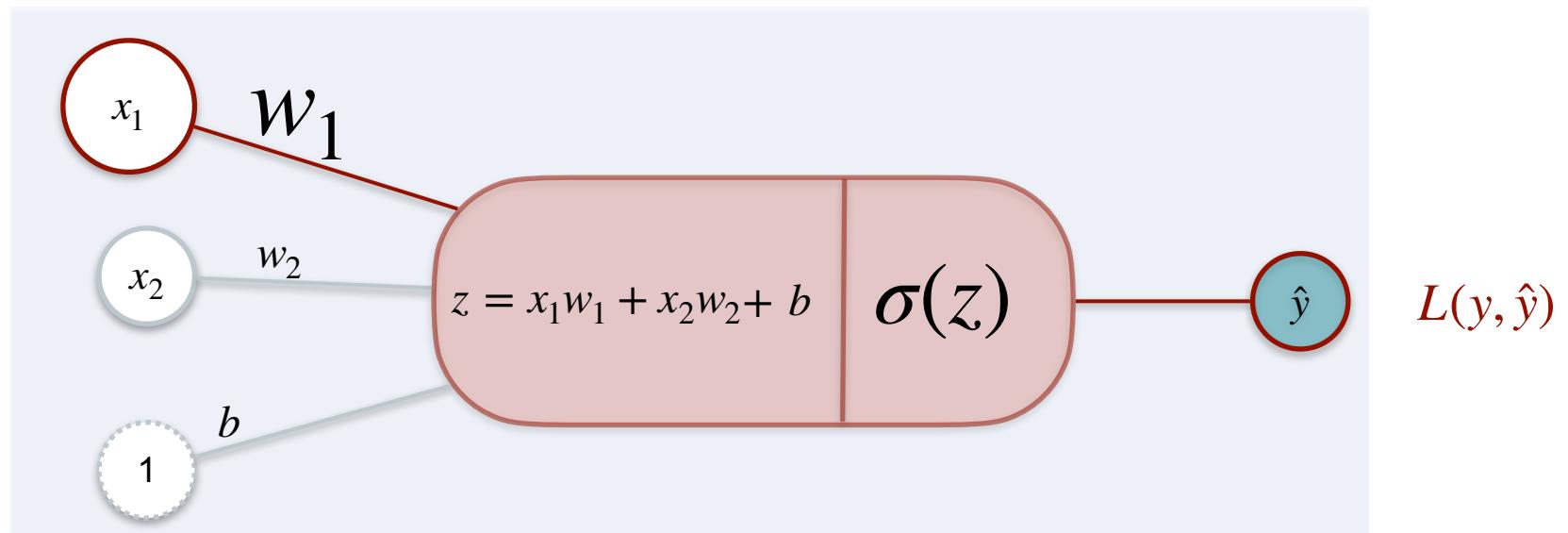
DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

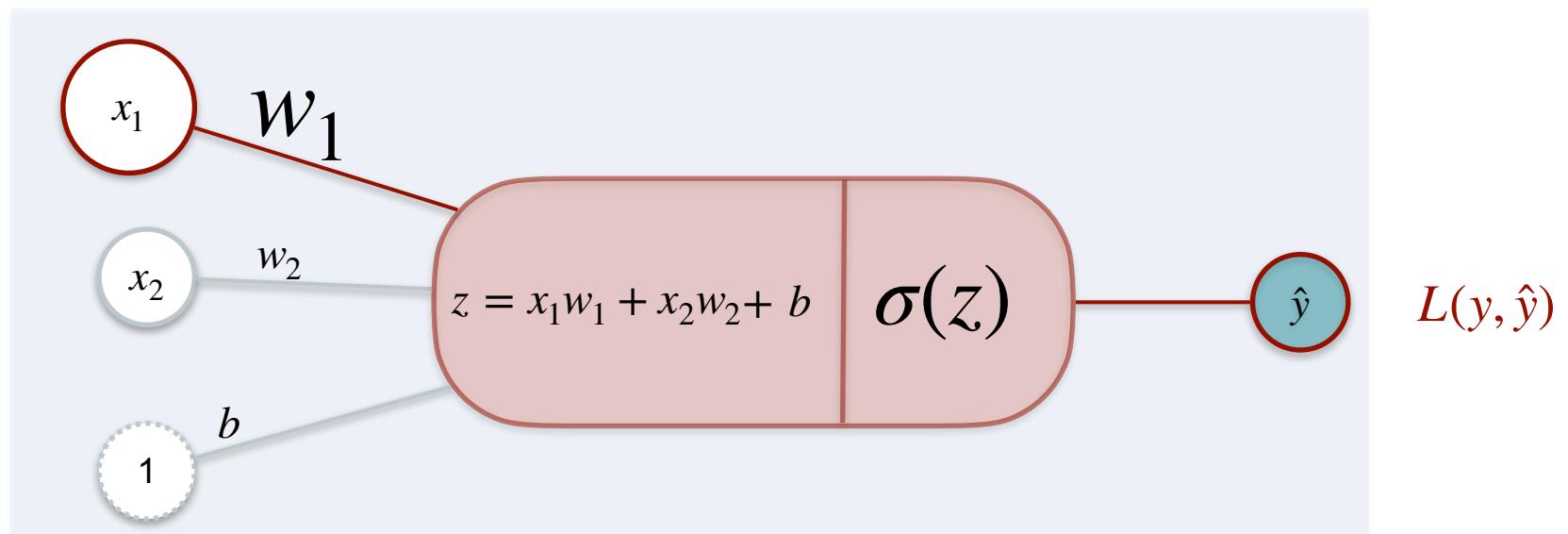
---

**Classification with a  
perceptron:  
Calculating the derivatives**

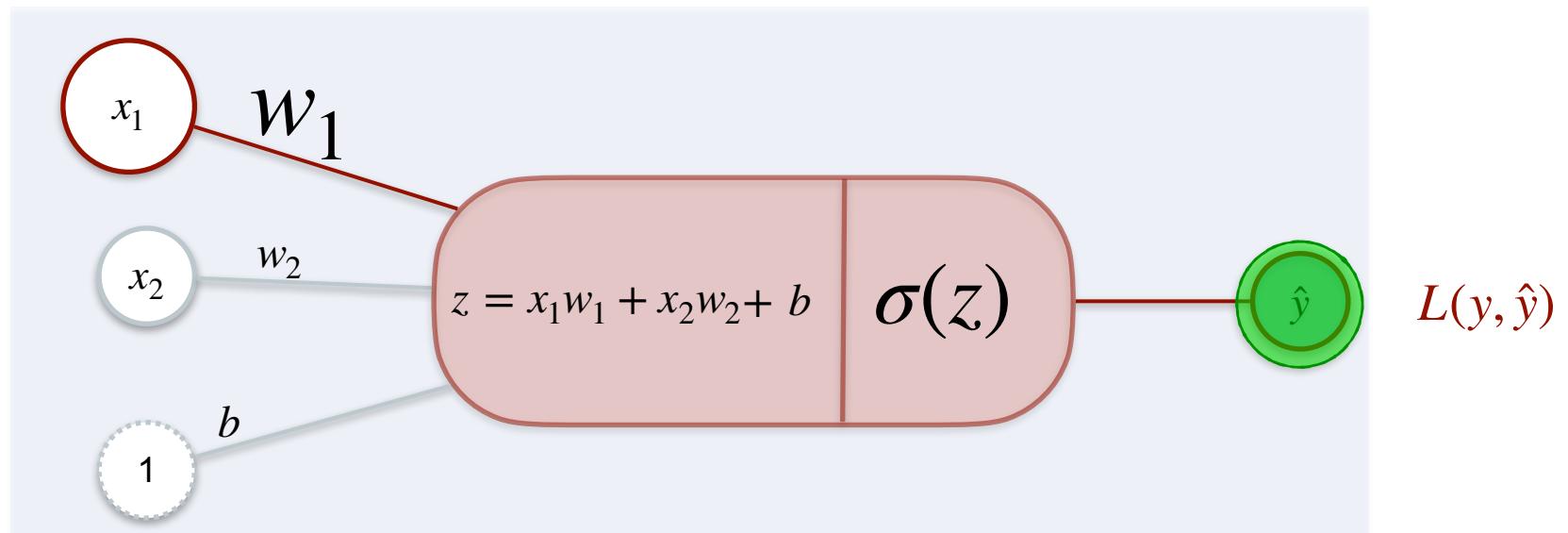
# Classification With a Perceptron



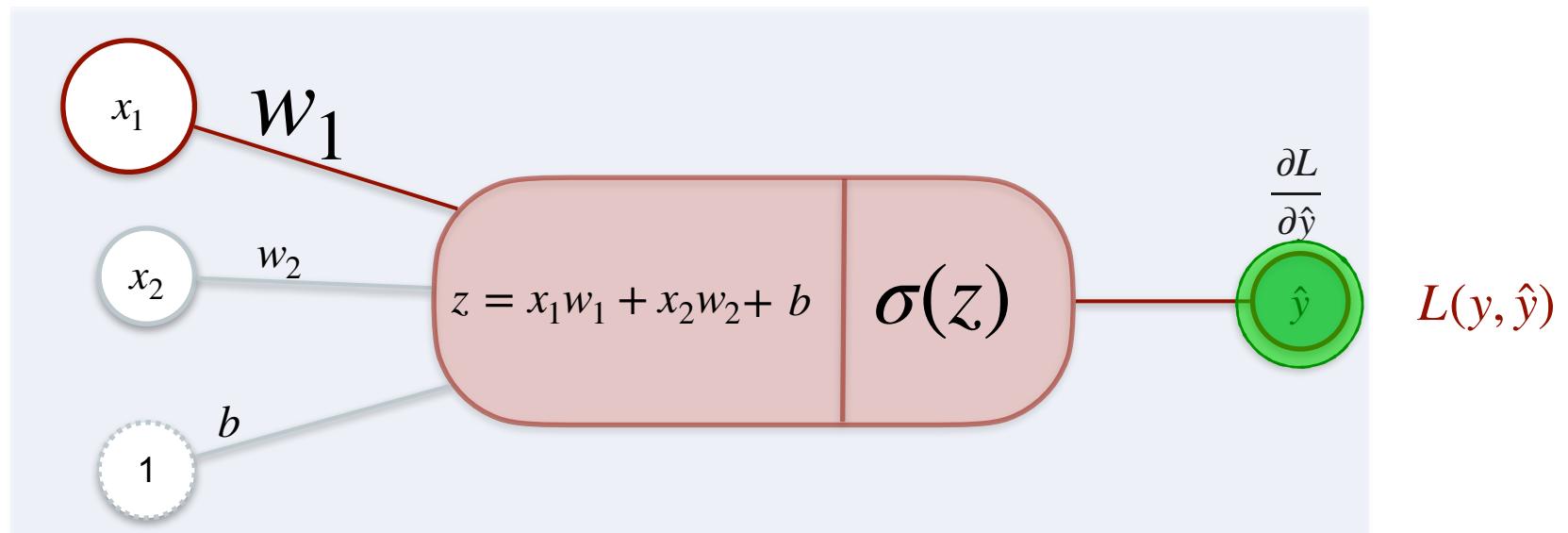
# Classification With a Perceptron



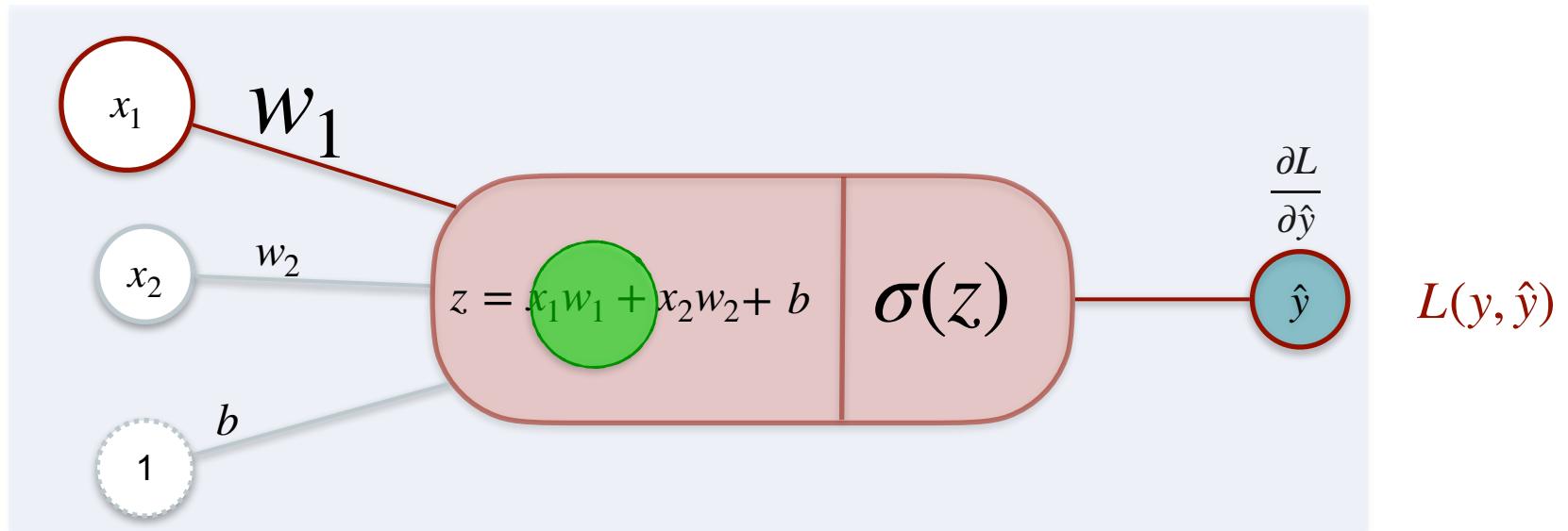
# Classification With a Perceptron



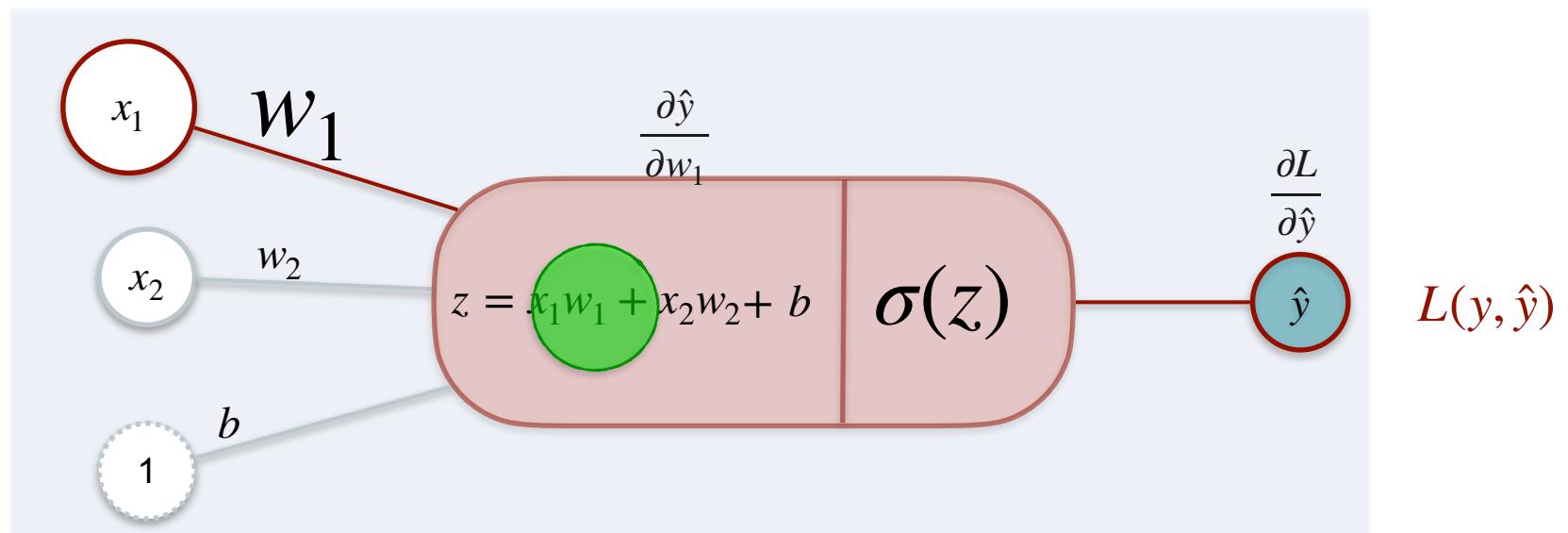
# Classification With a Perceptron



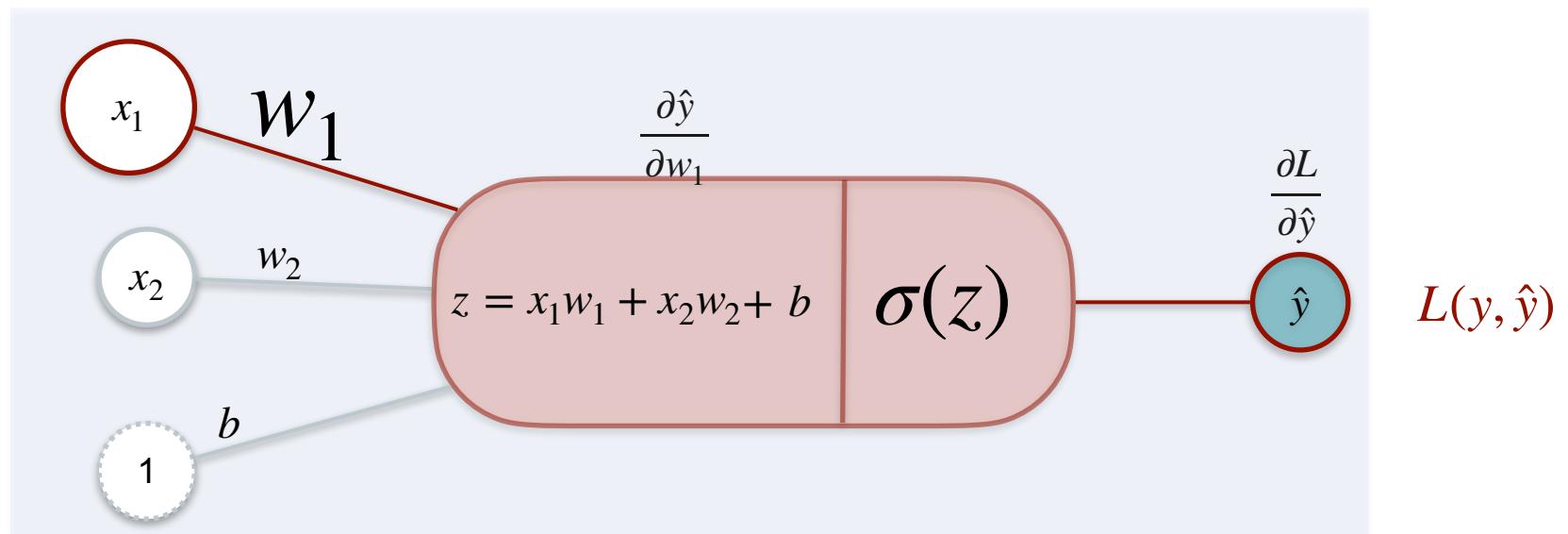
# Classification With a Perceptron



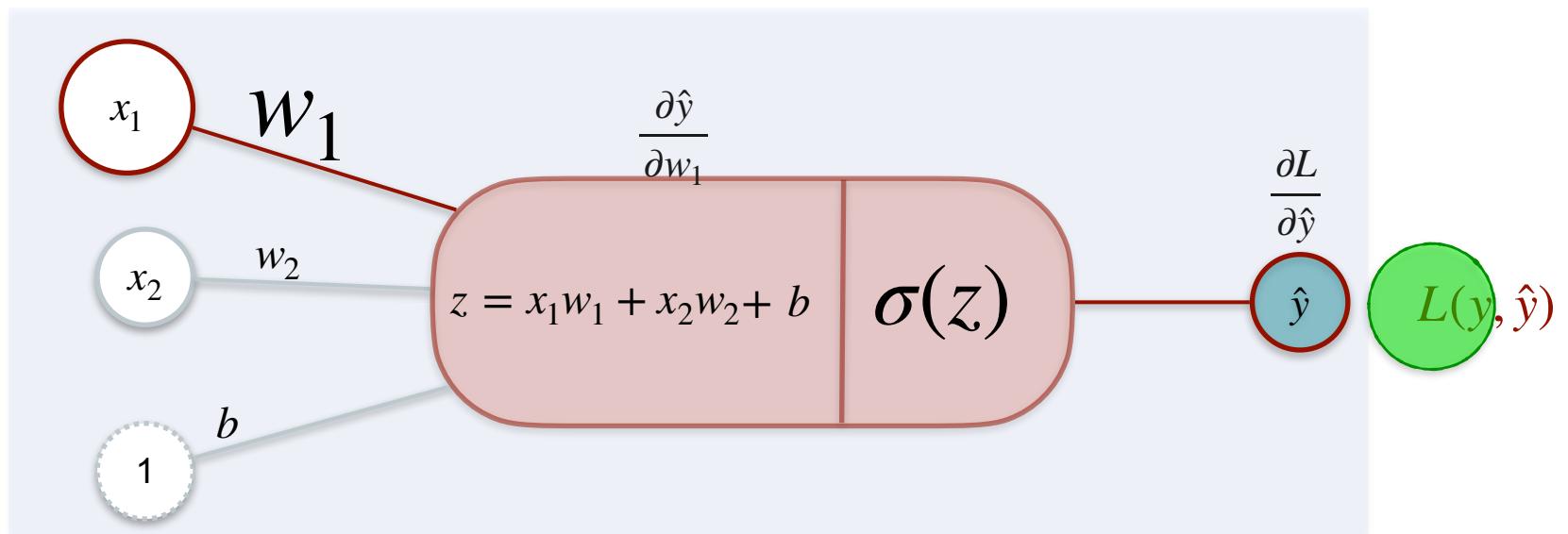
# Classification With a Perceptron



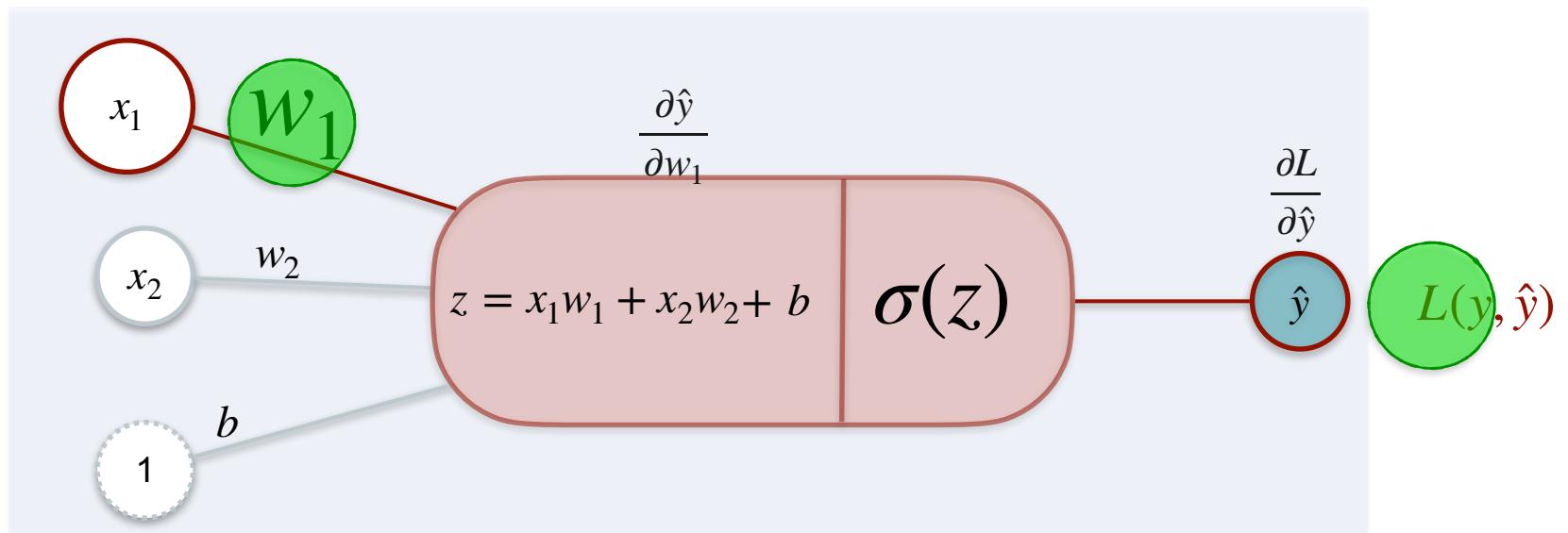
# Classification With a Perceptron



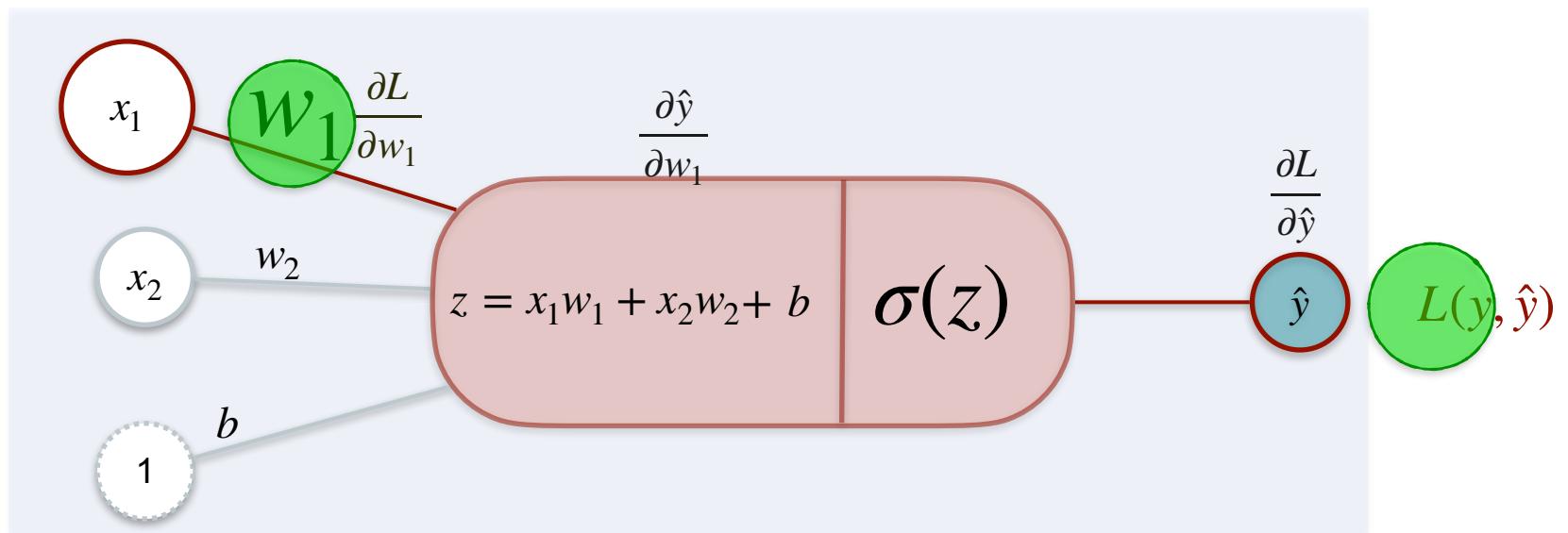
# Classification With a Perceptron



# Classification With a Perceptron

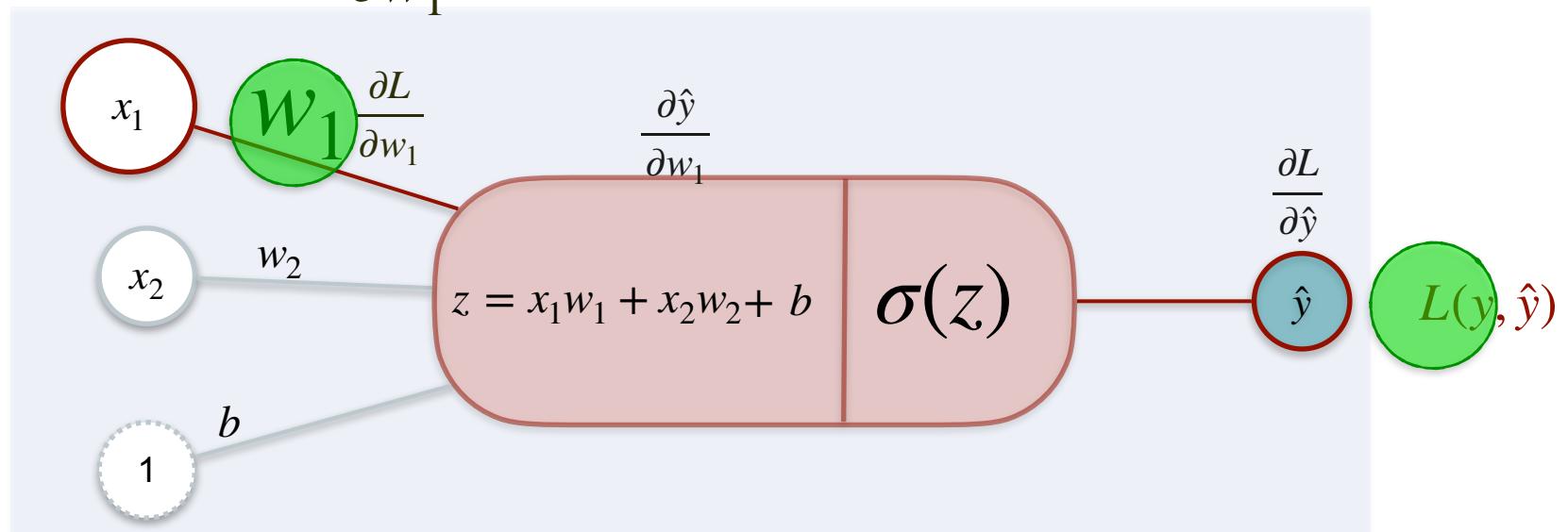


# Classification With a Perceptron



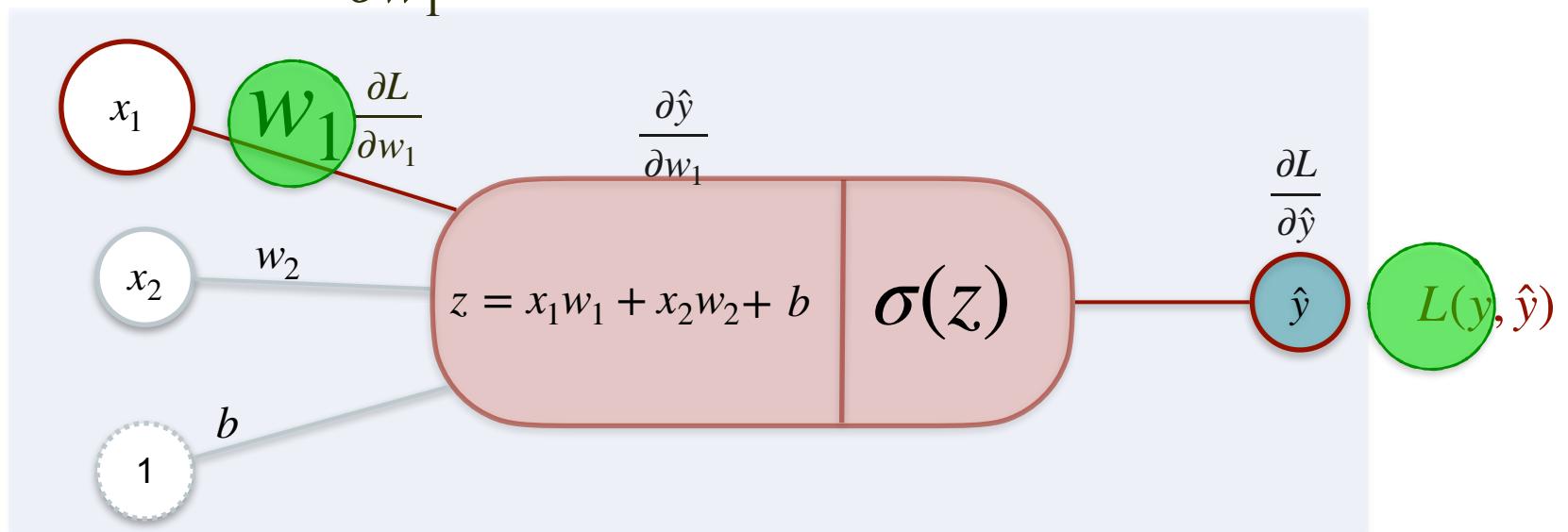
# Classification With a Perceptron

$$\frac{\partial L}{\partial w_1}$$



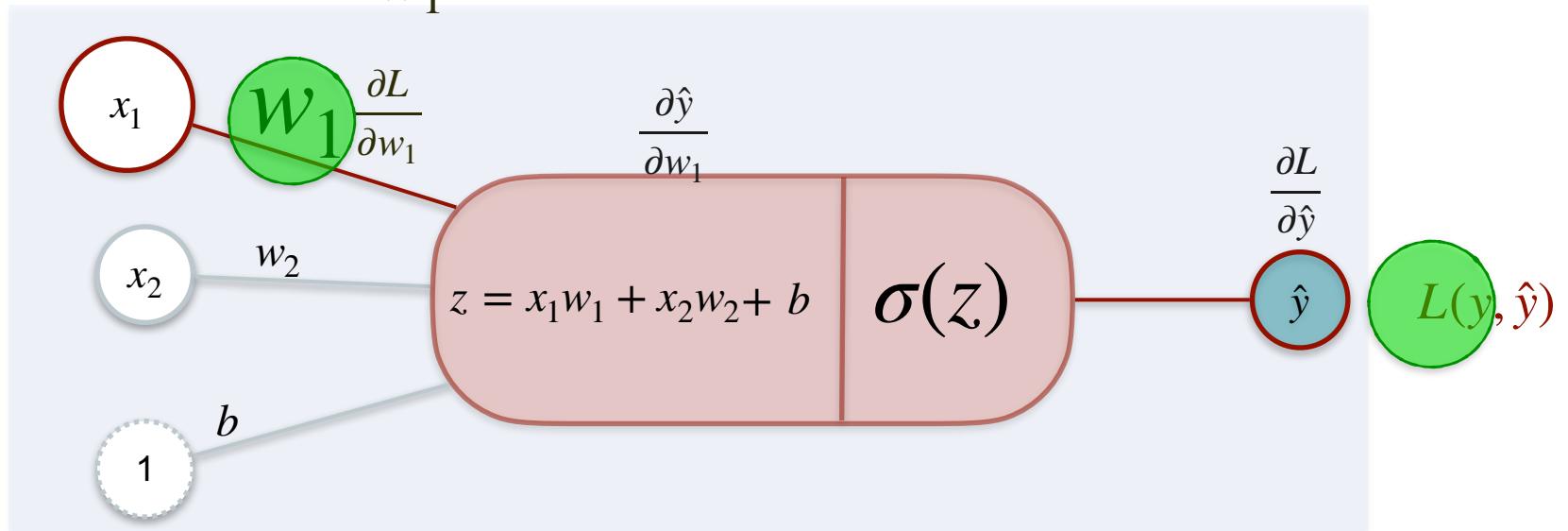
# Classification With a Perceptron

$$\frac{\partial L}{\partial w_1} =$$



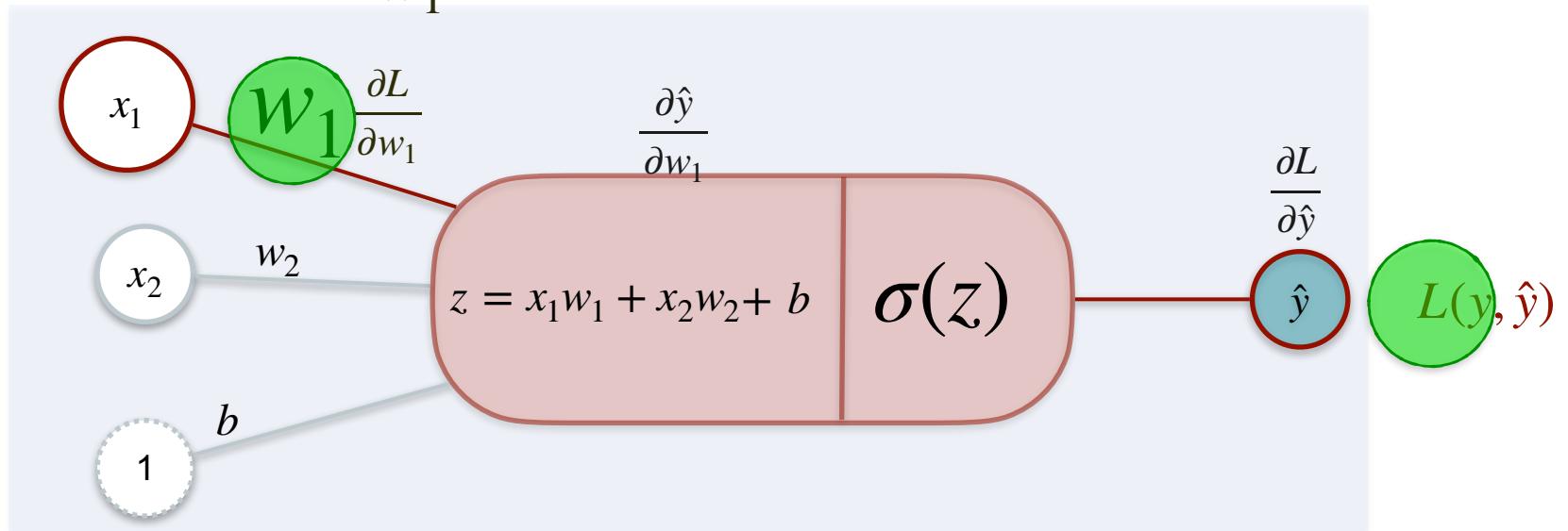
# Classification With a Perceptron

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}}$$



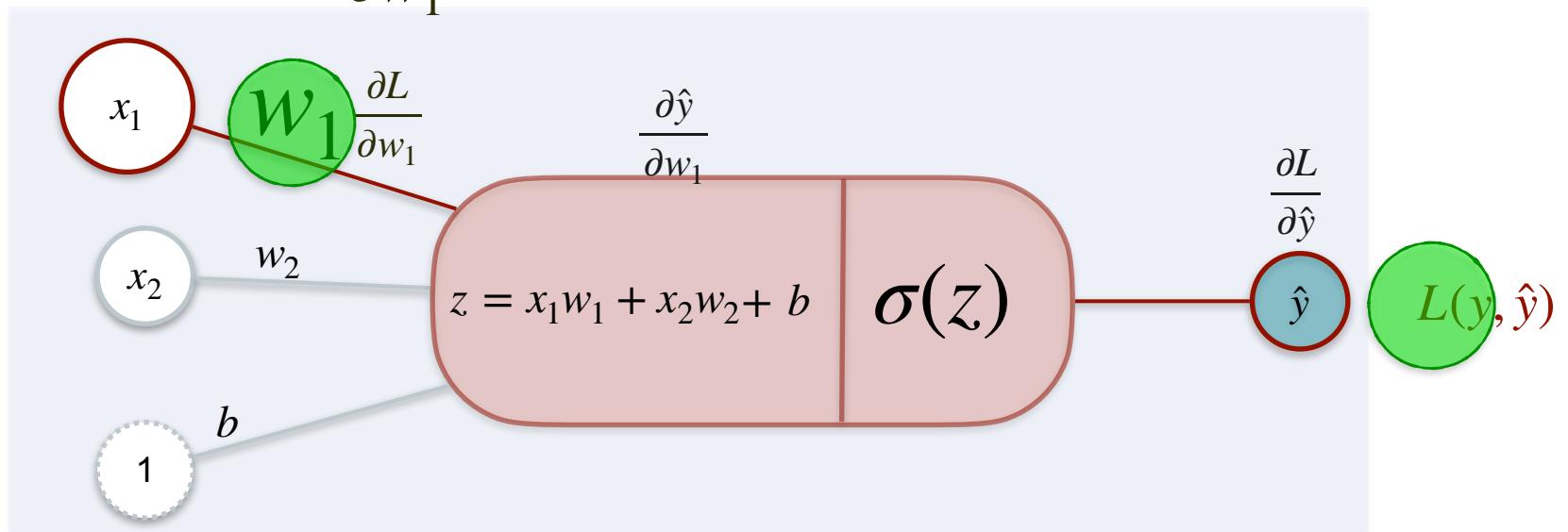
# Classification With a Perceptron

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot$$

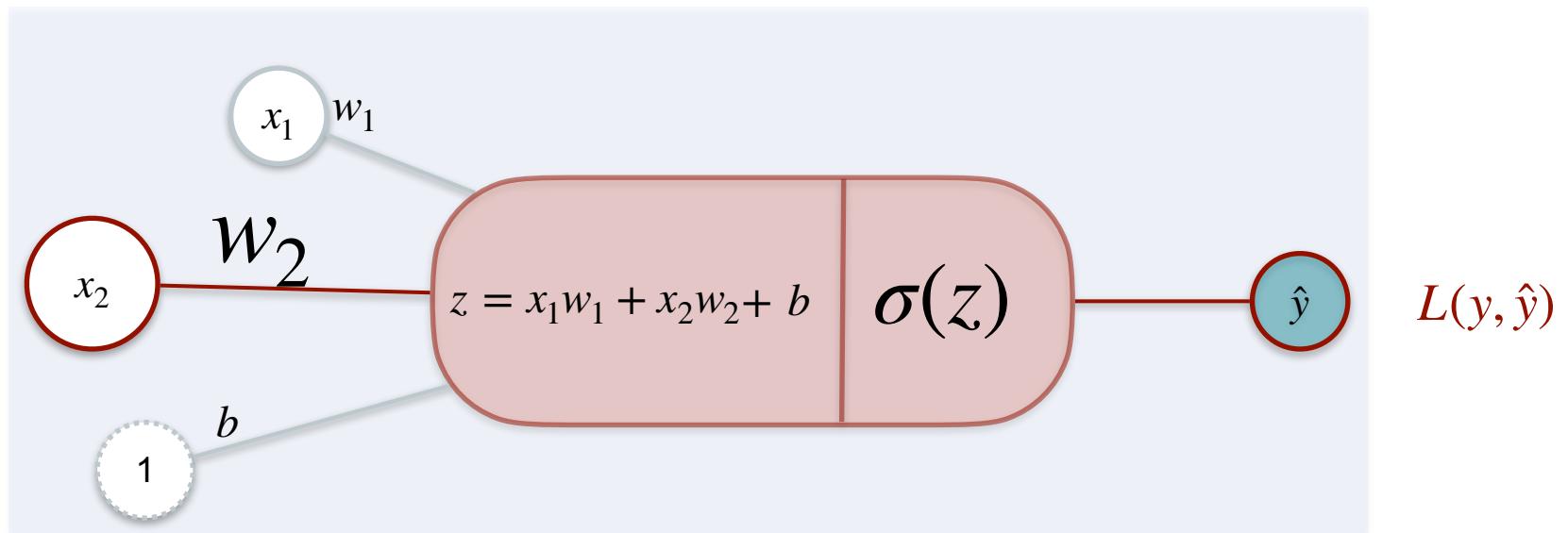


# Classification With a Perceptron

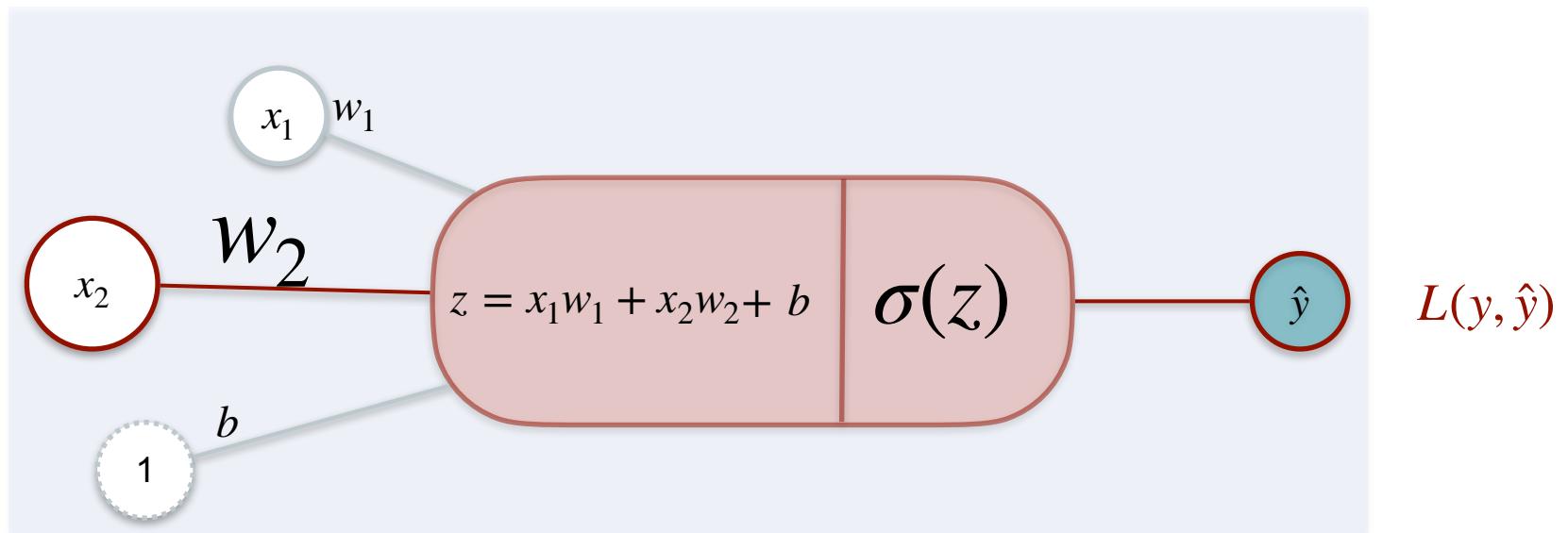
$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$



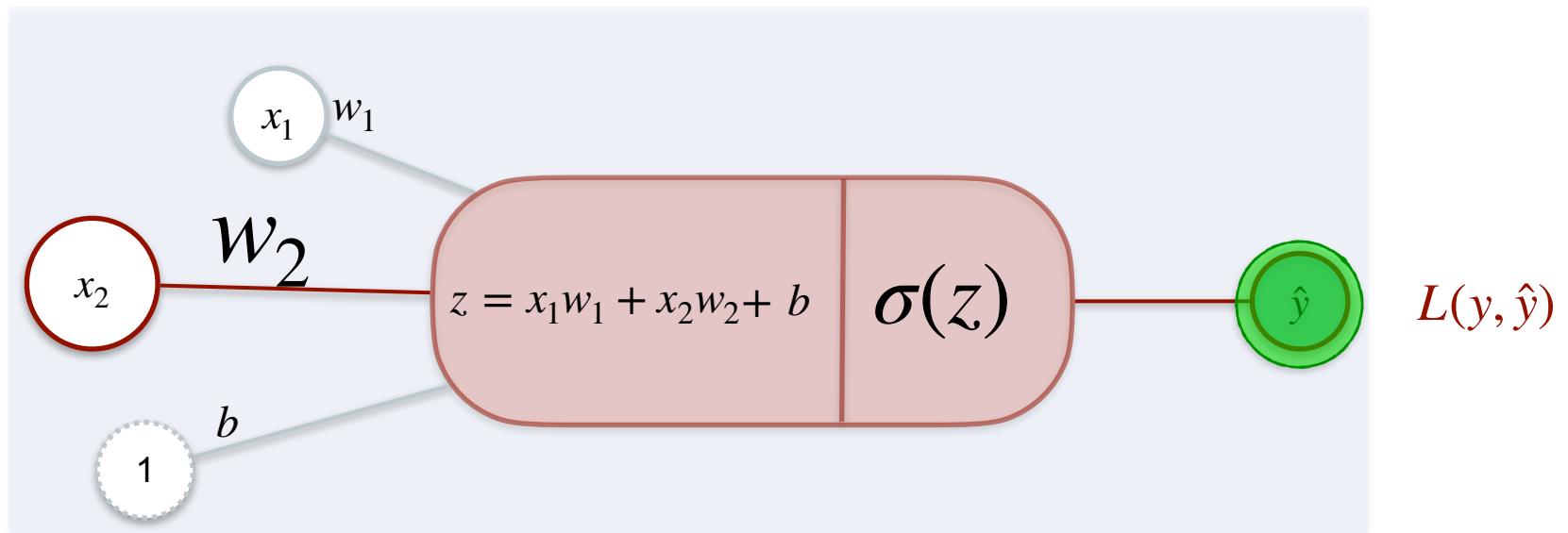
# Classification With a Perceptron



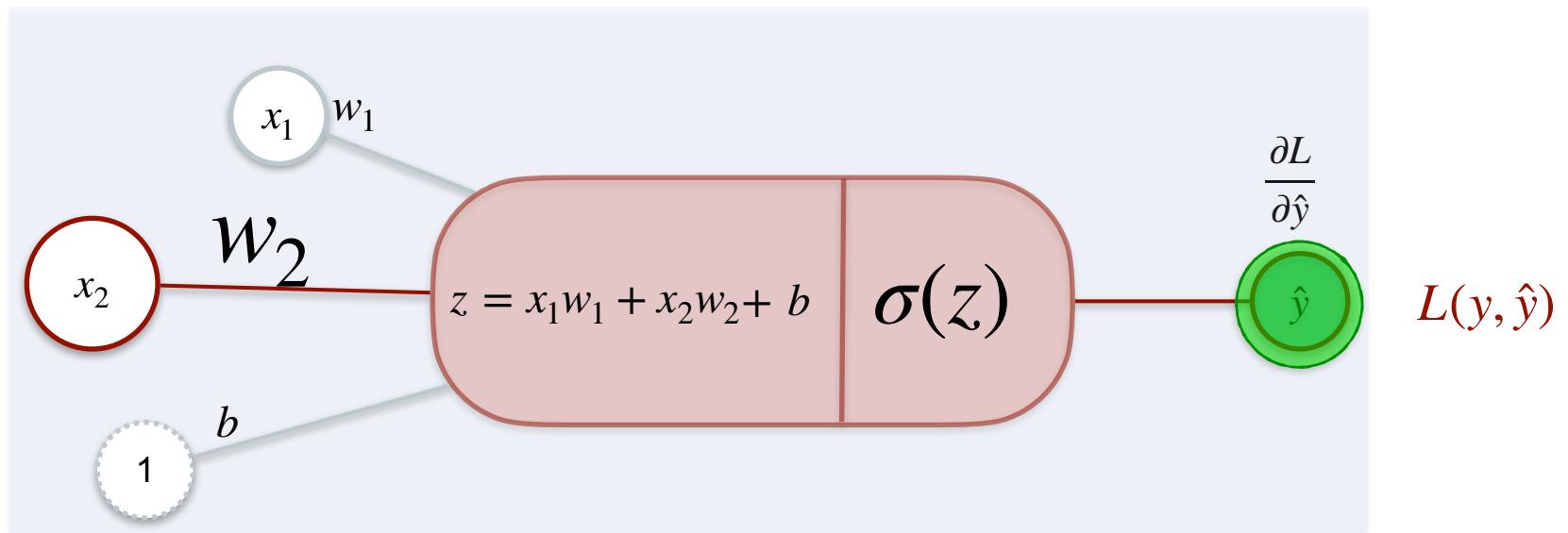
# Classification With a Perceptron



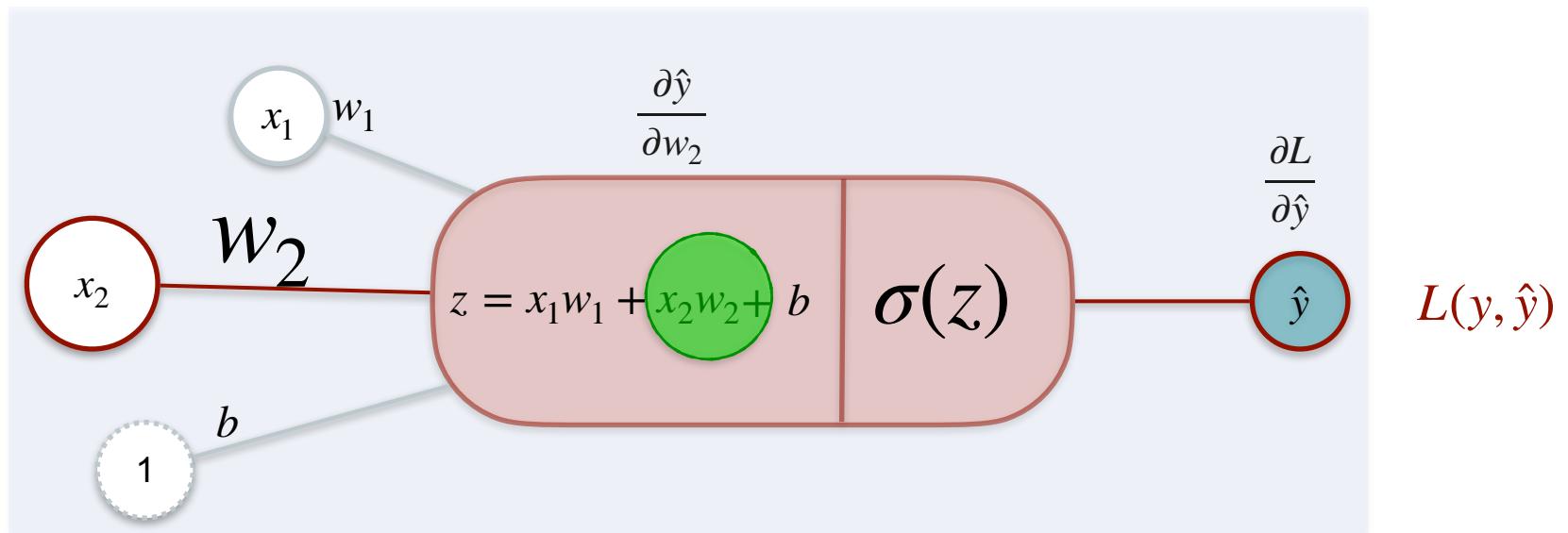
# Classification With a Perceptron



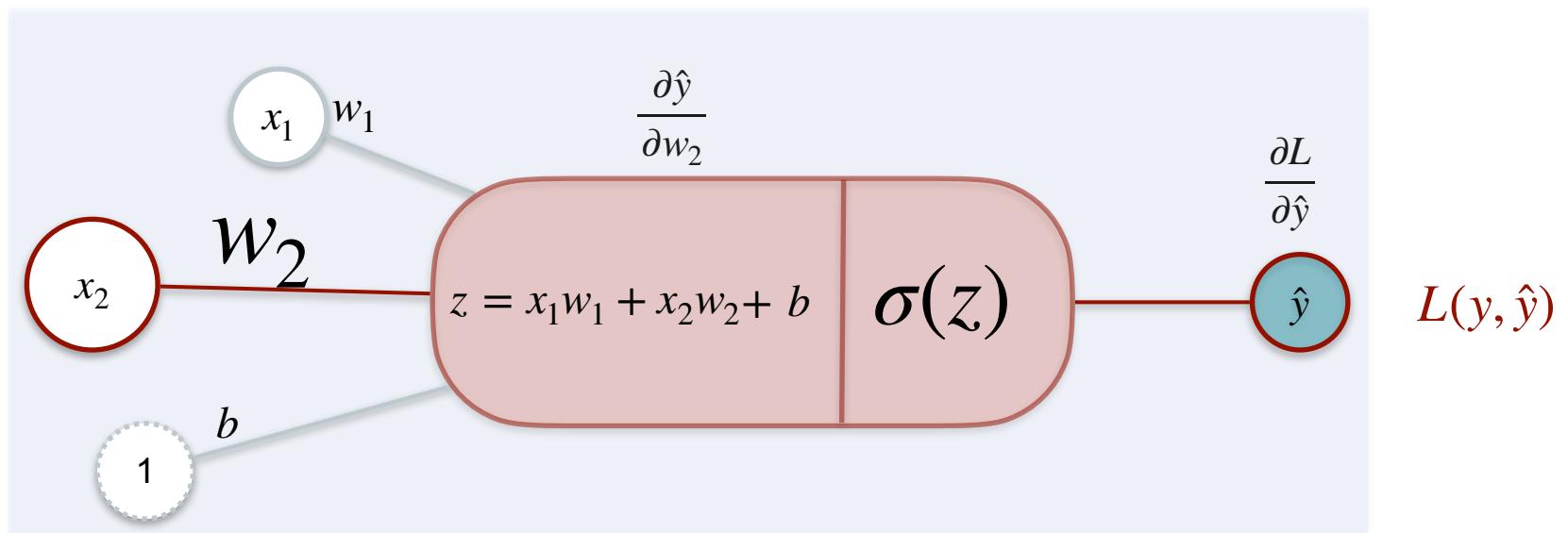
# Classification With a Perceptron



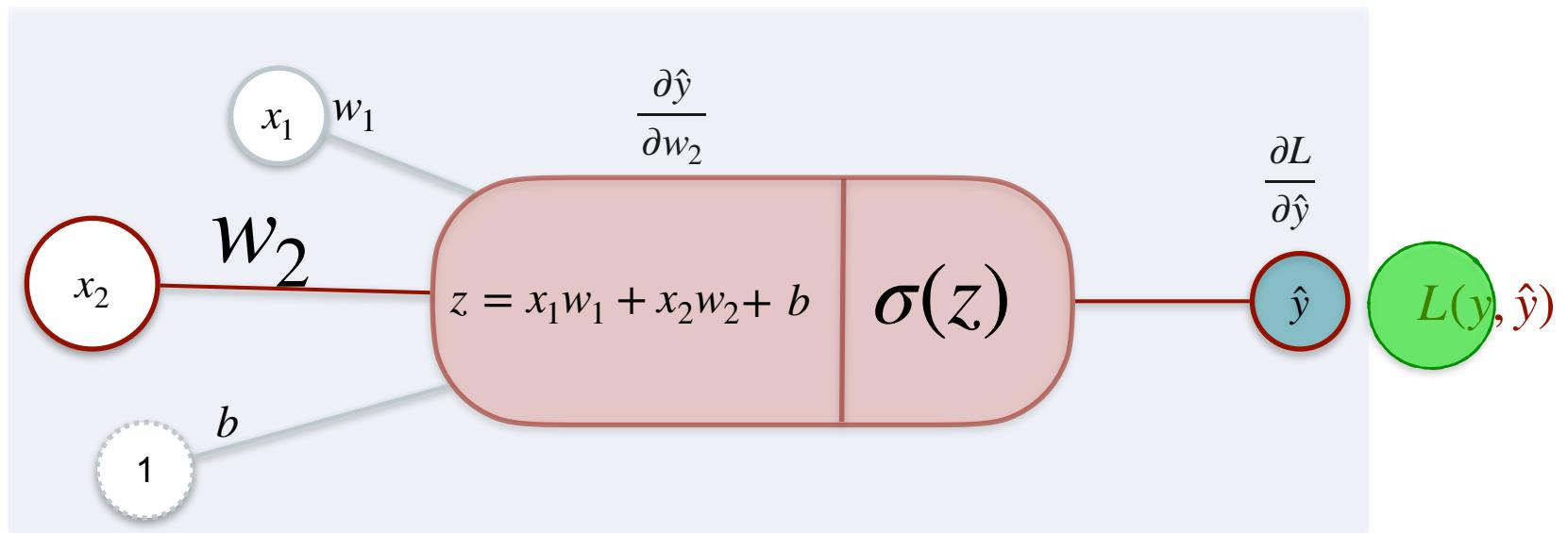
# Classification With a Perceptron



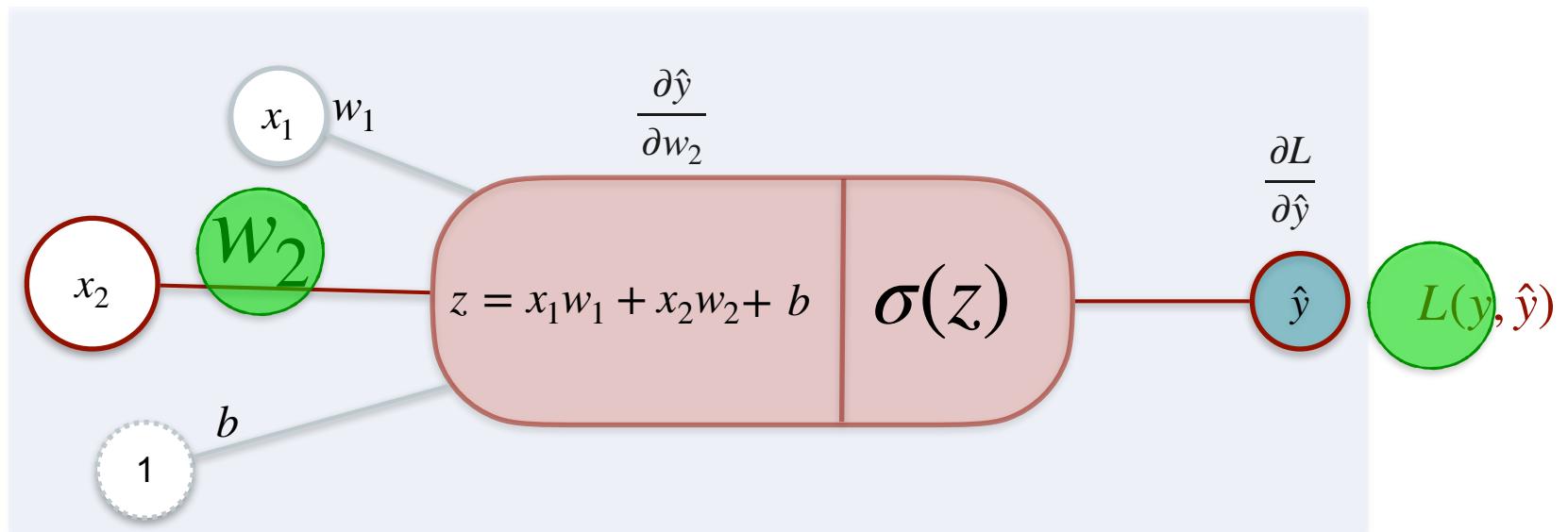
# Classification With a Perceptron



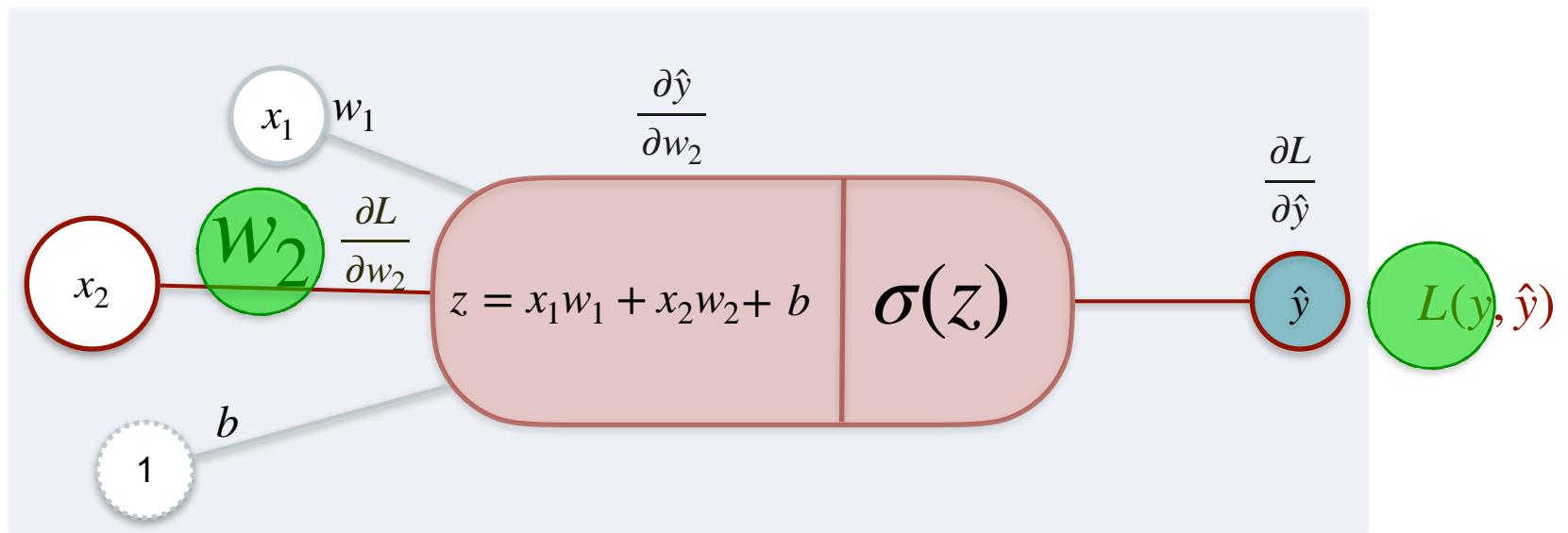
# Classification With a Perceptron



# Classification With a Perceptron

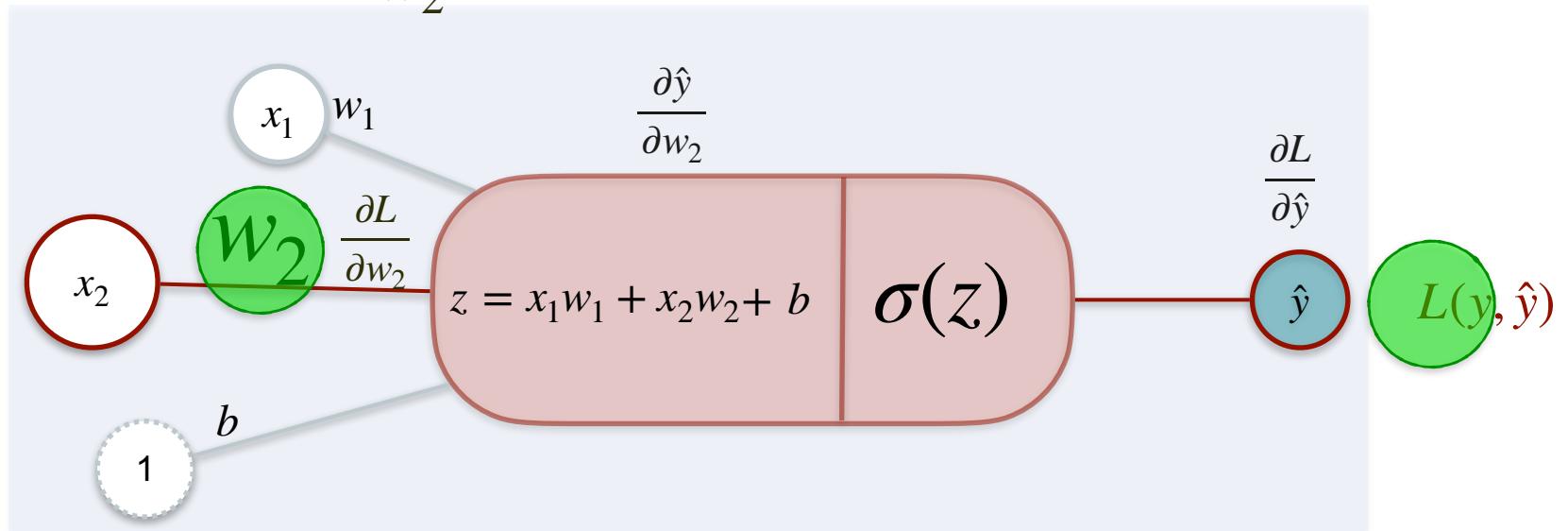


# Classification With a Perceptron

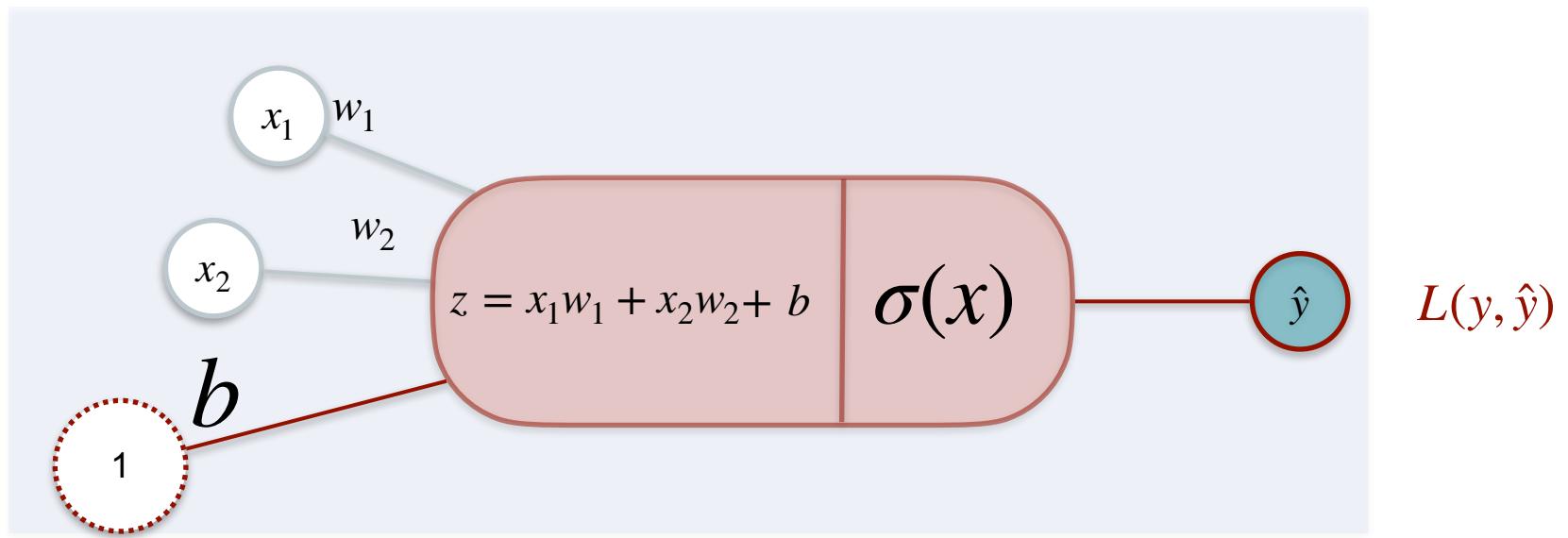


# Classification With a Perceptron

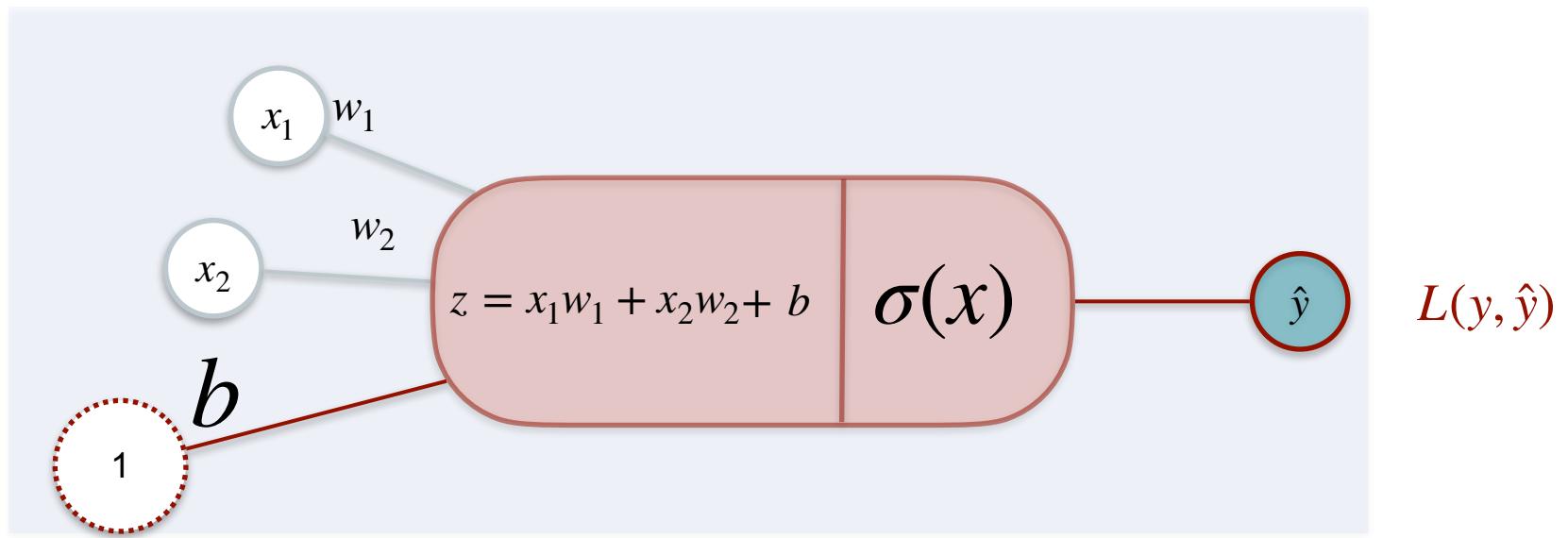
$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$



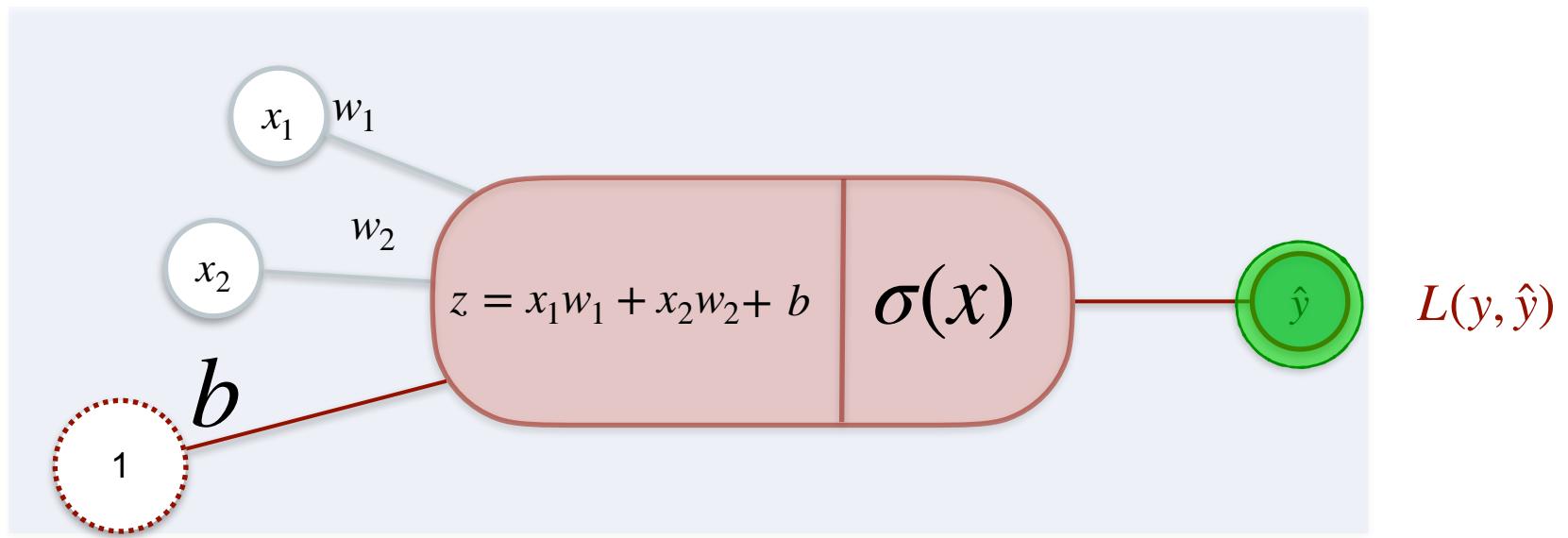
# Classification With a Perceptron



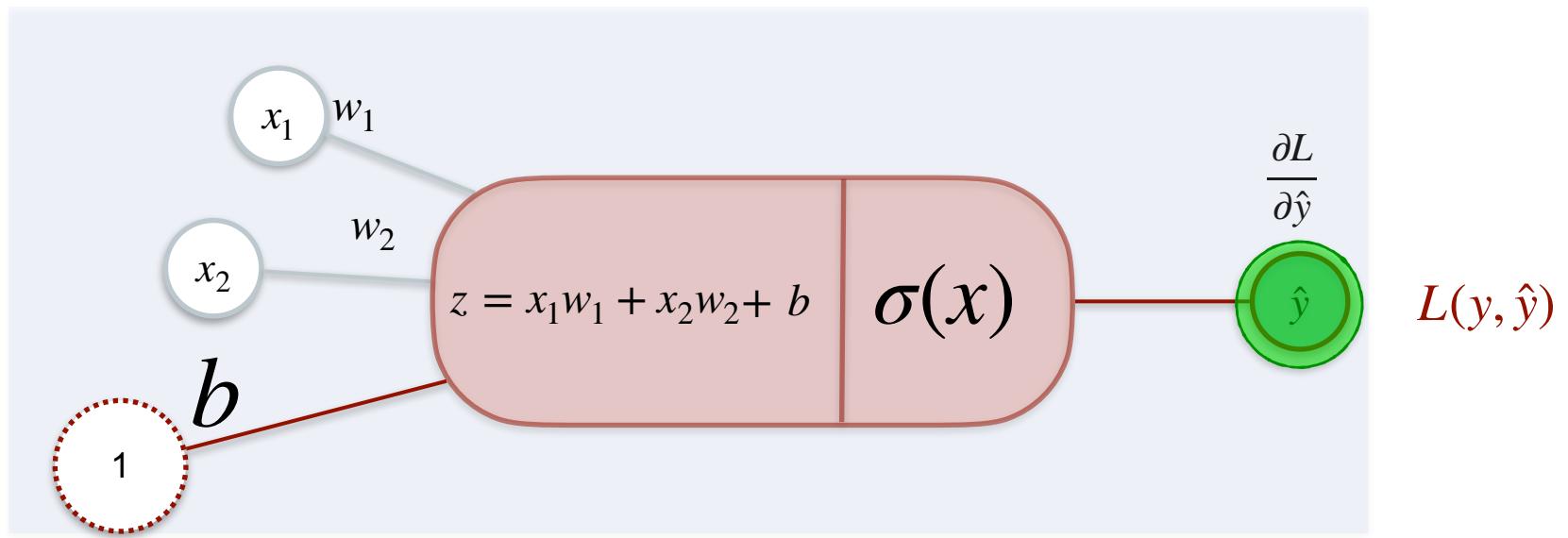
# Classification With a Perceptron



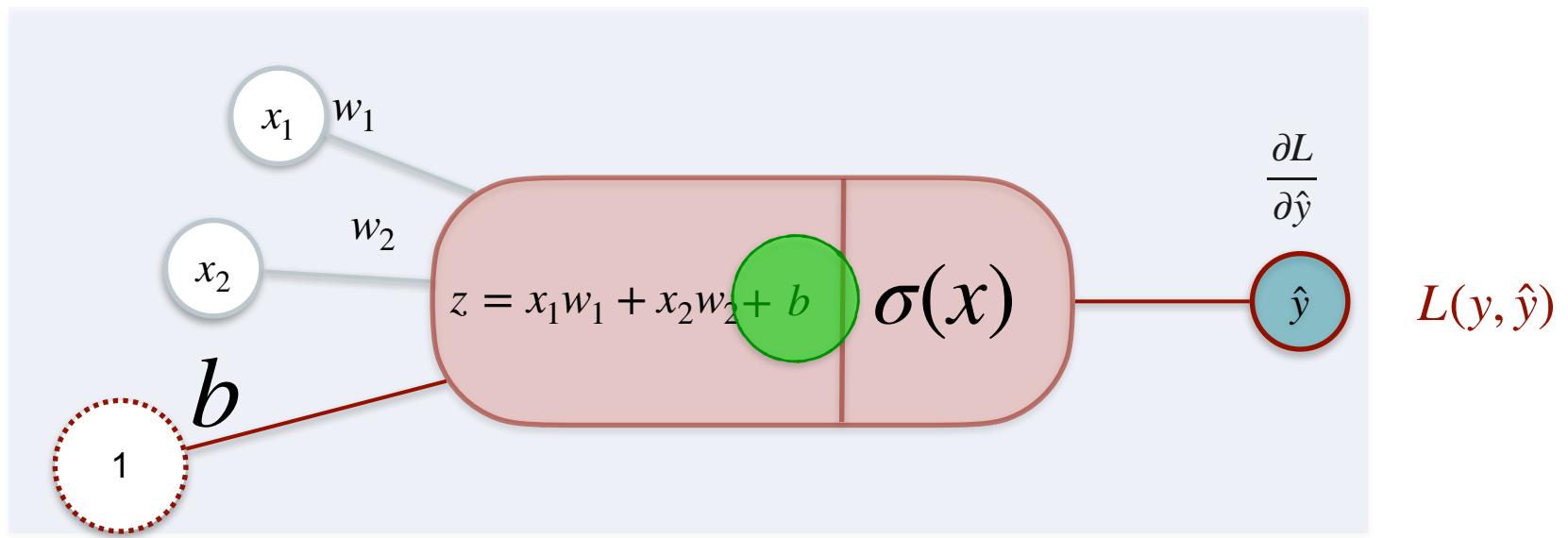
# Classification With a Perceptron



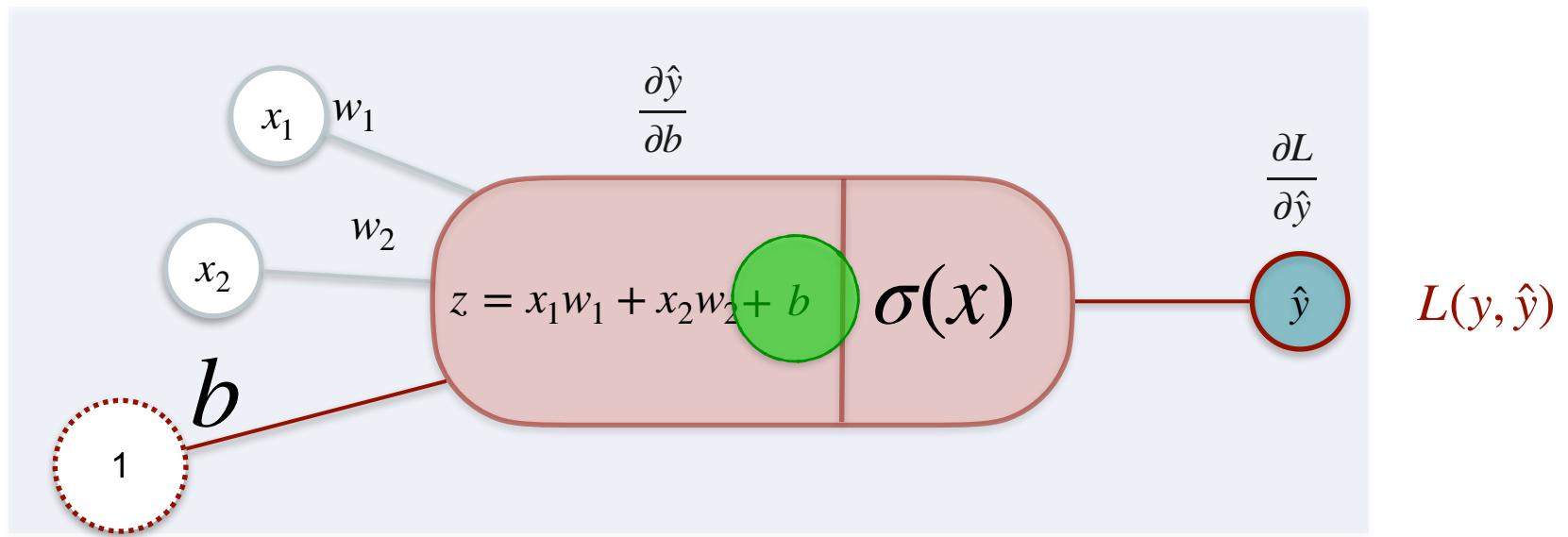
# Classification With a Perceptron



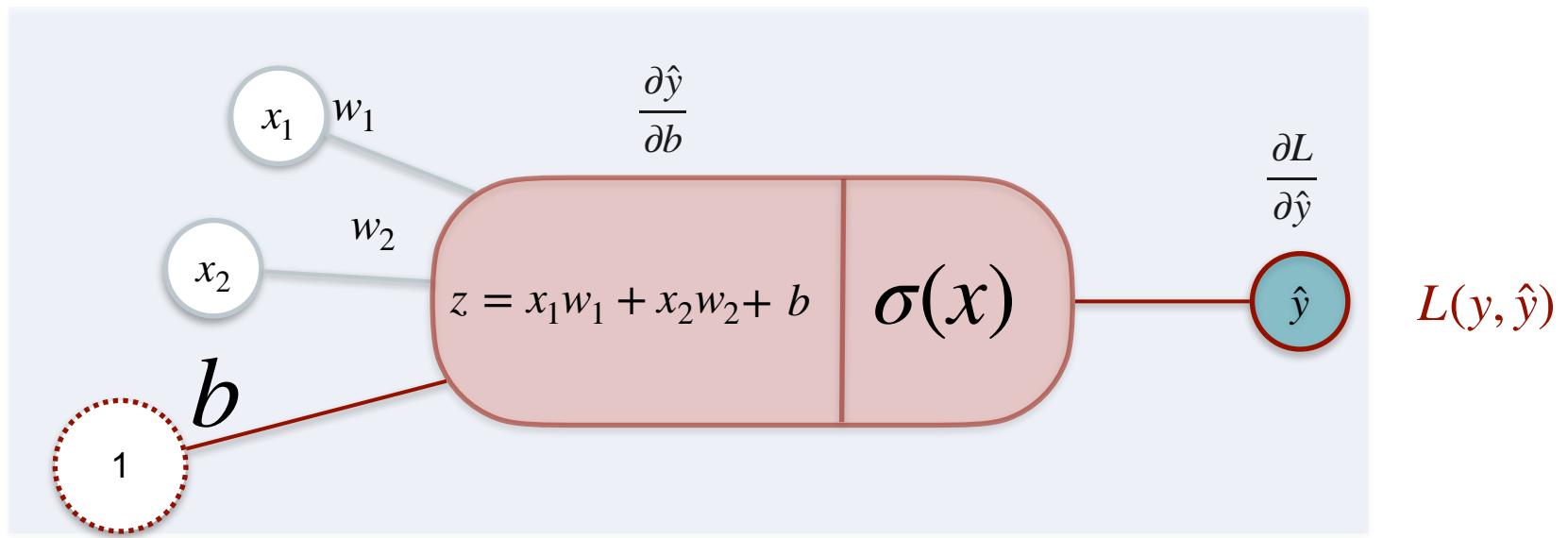
# Classification With a Perceptron



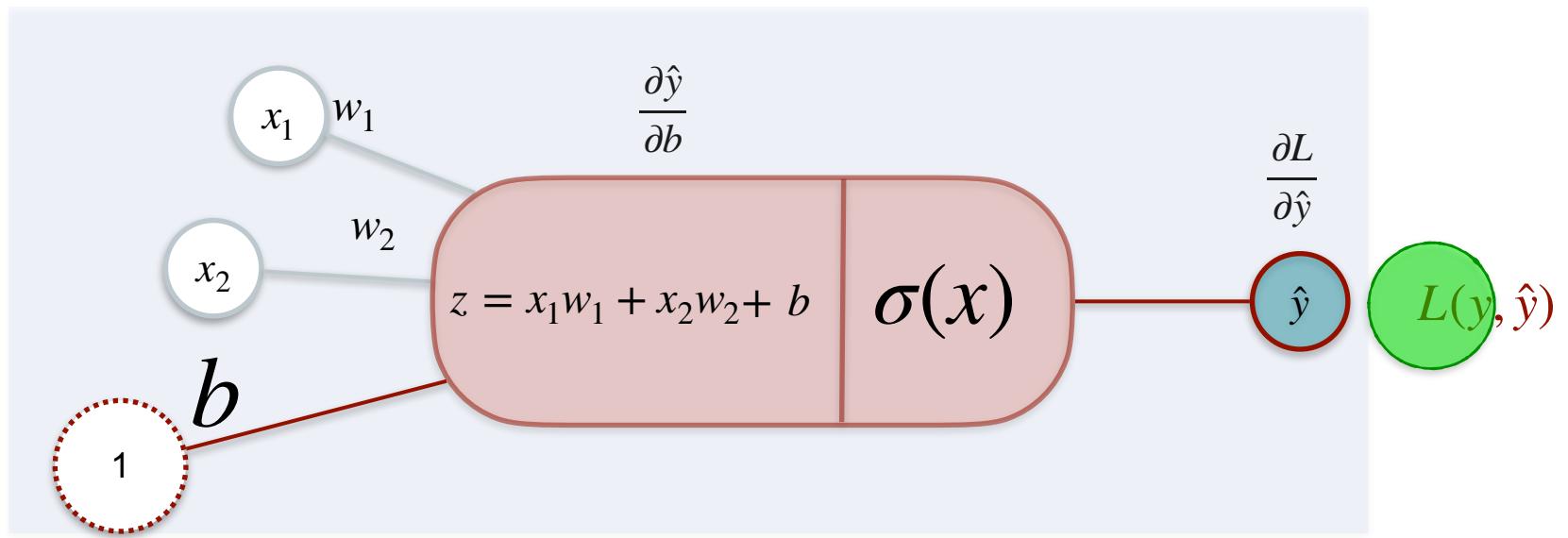
# Classification With a Perceptron



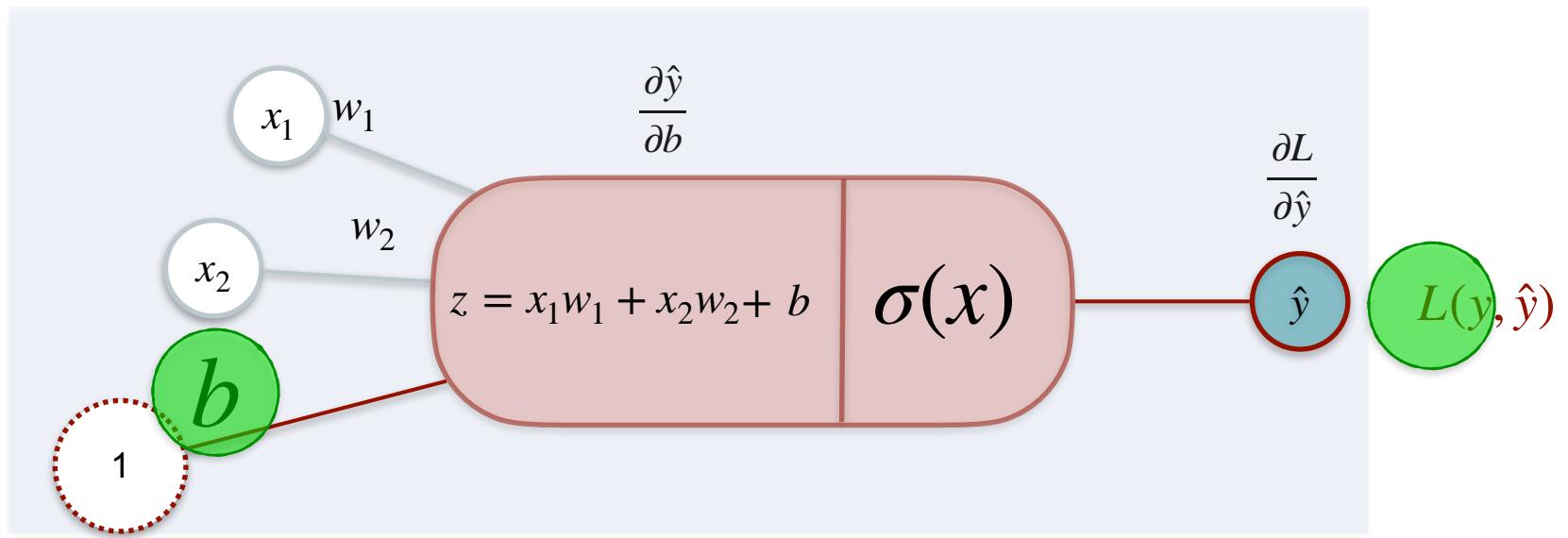
# Classification With a Perceptron



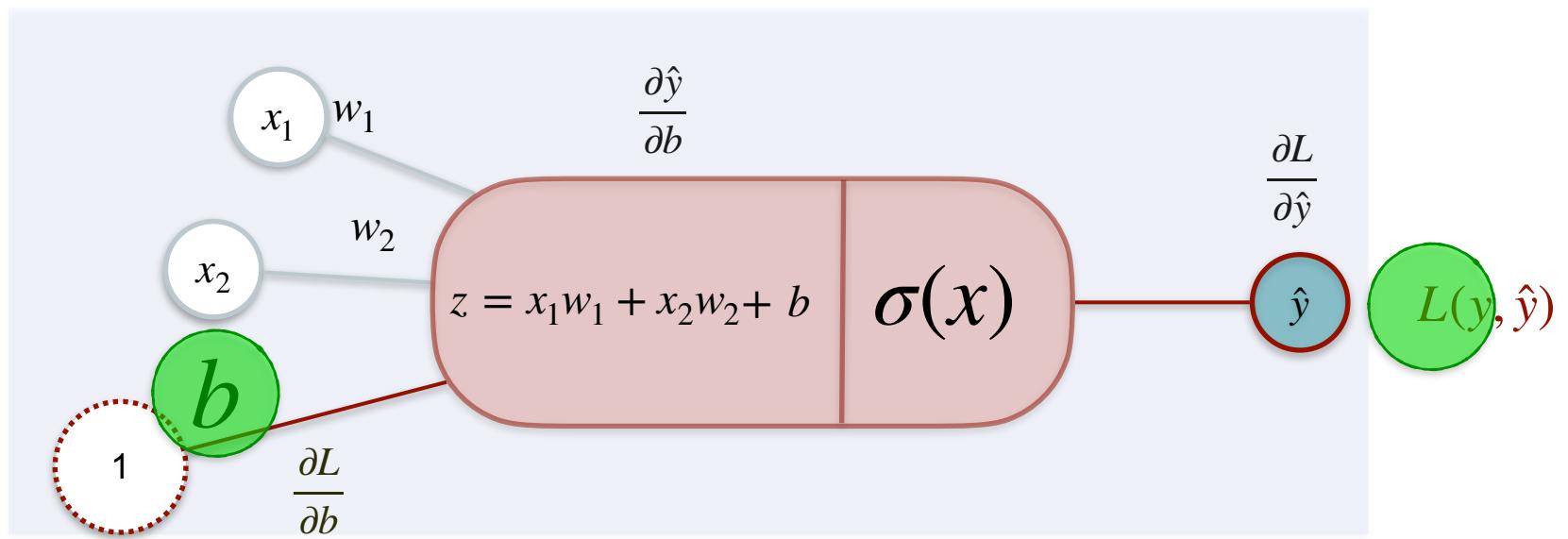
# Classification With a Perceptron



# Classification With a Perceptron

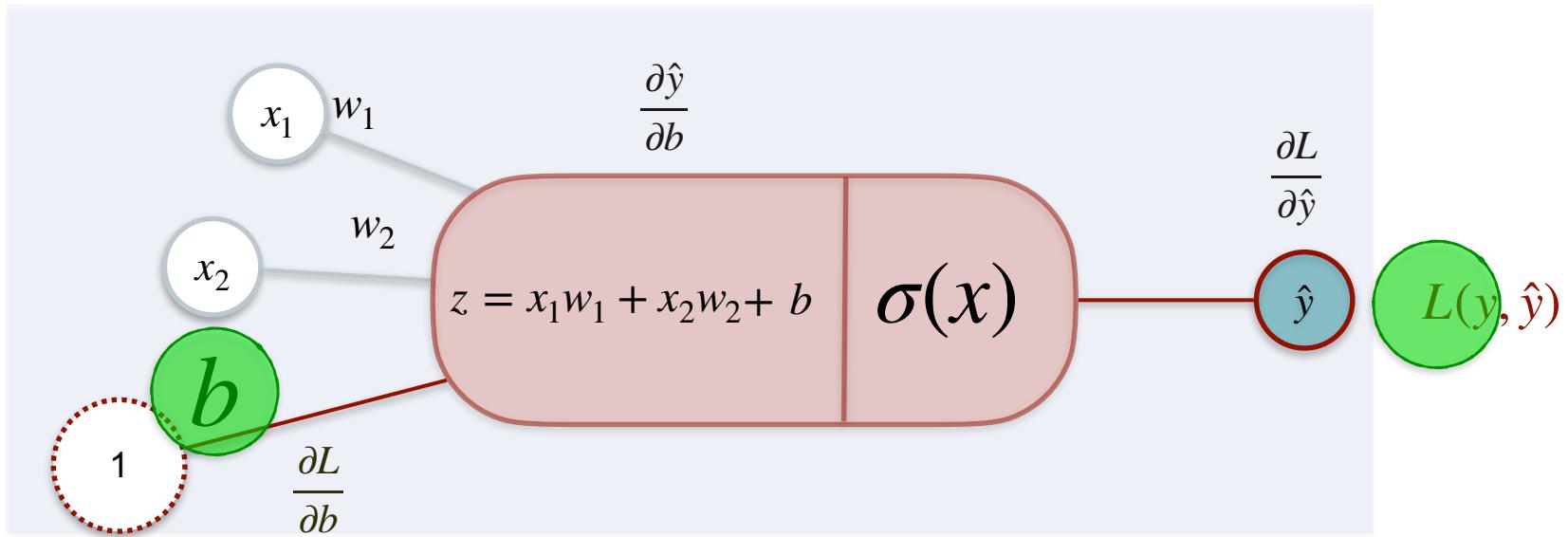


# Classification With a Perceptron



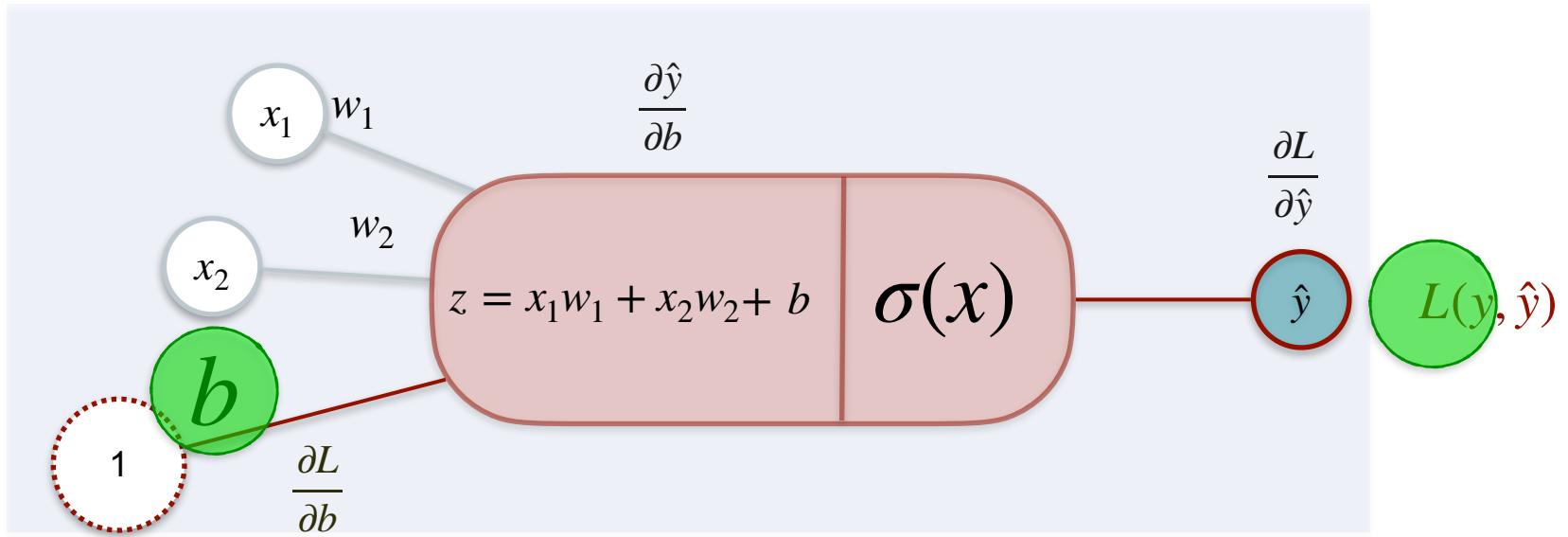
# Classification With a Perceptron

$$\frac{\partial L}{\partial b} =$$



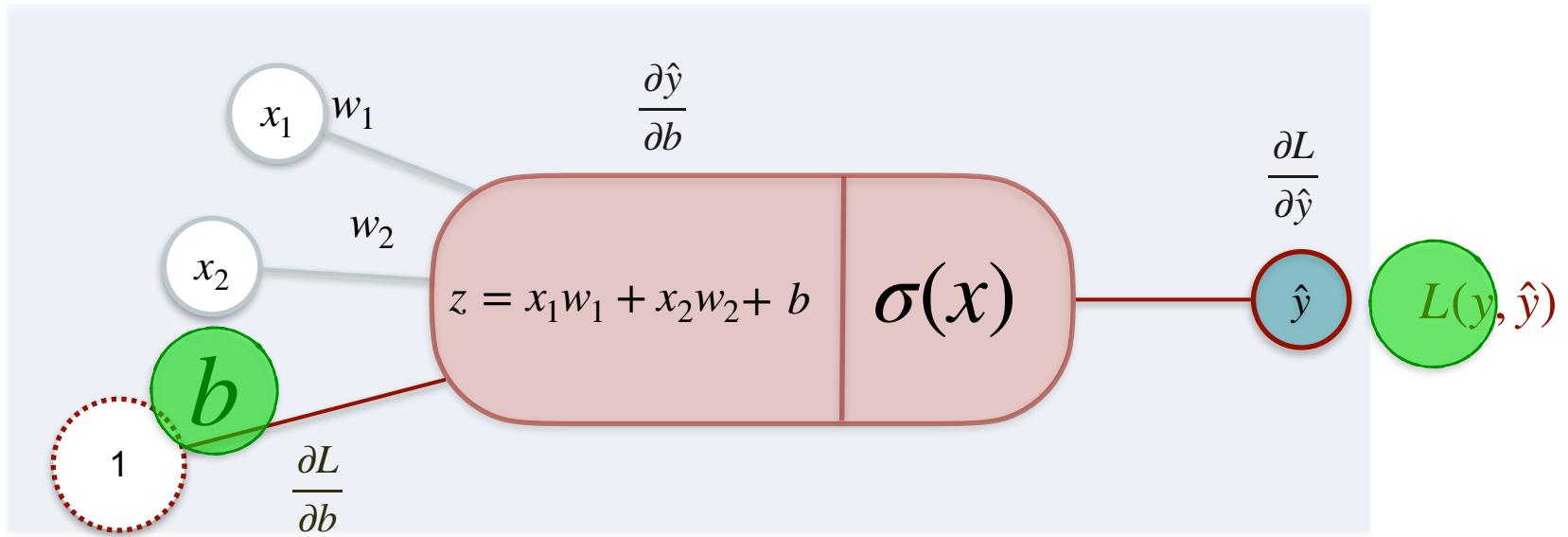
# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}}$$



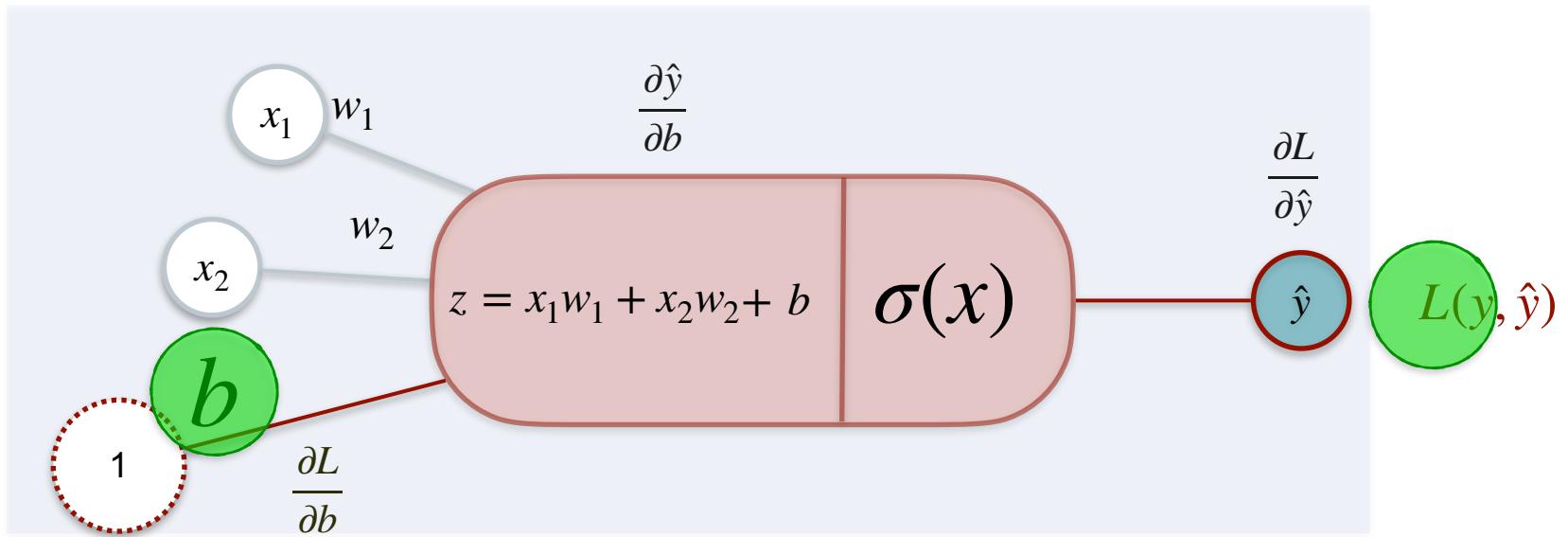
# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot$$



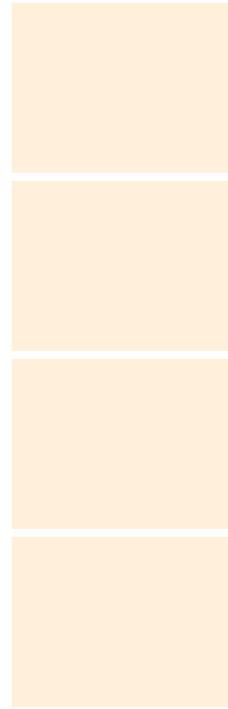
# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$



# Classification With a Perceptron

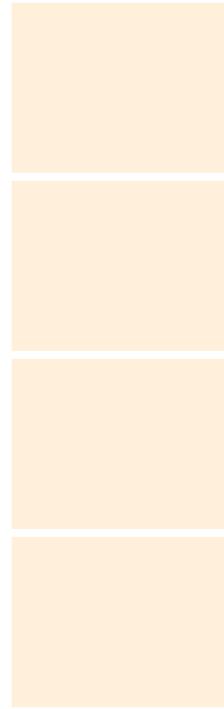
$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$



# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

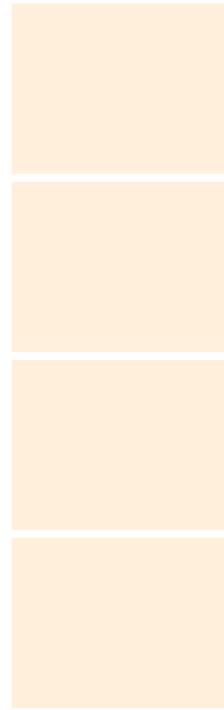


# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$



# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$\frac{\partial L}{\partial \hat{y}} =$$

$$\frac{\partial \hat{y}}{\partial b} =$$

$$\frac{\partial \hat{y}}{\partial w_1} =$$

$$\frac{\partial \hat{y}}{\partial w_2} =$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$\frac{\partial L}{\partial \hat{y}} =$$

$$\frac{\partial \hat{y}}{\partial b} =$$

?

$$\frac{\partial \hat{y}}{\partial w_1} =$$

$$\frac{\partial \hat{y}}{\partial w_2} =$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$\frac{\partial L}{\partial \hat{y}} =$$

$$\frac{\partial \hat{y}}{\partial b} =$$

$$\frac{\partial \hat{y}}{\partial w_1} =$$

$$\frac{\partial \hat{y}}{\partial w_2} =$$

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$\frac{\partial L}{\partial \hat{y}} =$$

$$\frac{\partial \hat{y}}{\partial b} =$$

$$\frac{\partial \hat{y}}{\partial w_1} =$$

$$\frac{\partial \hat{y}}{\partial w_2} =$$

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$

$$= \frac{-y + y\hat{y} + \hat{y} - y\hat{y}}{\hat{y}(1 - \hat{y})}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$

$$= \frac{-y + \cancel{y\hat{y}} + \hat{y} - y\hat{y}}{\hat{y}(1 - \hat{y})}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$

$$= \frac{-y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}}{\hat{y}(1 - \hat{y})}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$

$$= \frac{-y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}}{\hat{y}(1 - \hat{y})}$$

$$= \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

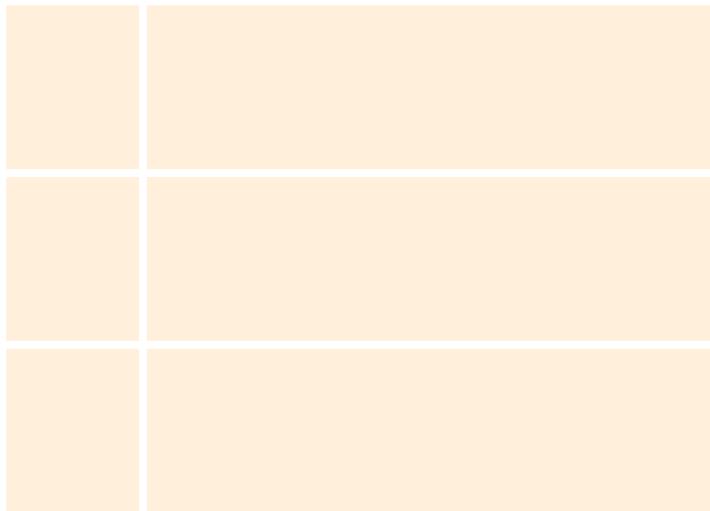
# Classification With a Perceptron

# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

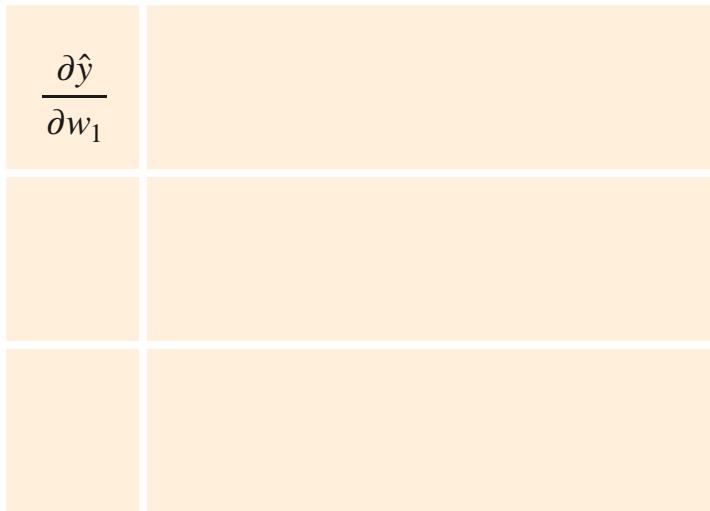
# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$



# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$



# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$\frac{\partial \hat{y}}{\partial w_1}$	$= \hat{y}(1 - \hat{y})x_1$

# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$\frac{\partial \hat{y}}{\partial w_1}$	$= \hat{y}(1 - \hat{y})x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	

# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$\frac{\partial \hat{y}}{\partial w_1}$	$= \hat{y}(1 - \hat{y})x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	$= \hat{y}(1 - \hat{y})x_2$

# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$\frac{\partial \hat{y}}{\partial w_1}$	$= \hat{y}(1 - \hat{y})x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	$= \hat{y}(1 - \hat{y})x_2$
$\frac{\partial \hat{y}}{\partial b}$	

# Classification With a Perceptron

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

$\frac{\partial \hat{y}}{\partial w_1}$	$= \hat{y}(1 - \hat{y})x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	$= \hat{y}(1 - \hat{y})x_2$
$\frac{\partial \hat{y}}{\partial b}$	$= \hat{y}(1 - \hat{y})$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_2$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_2$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial b} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w_1} = \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial \hat{y}}{\partial w_2} = \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_2$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_2$$

# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \hat{y}(1 - \hat{y})x_2$$

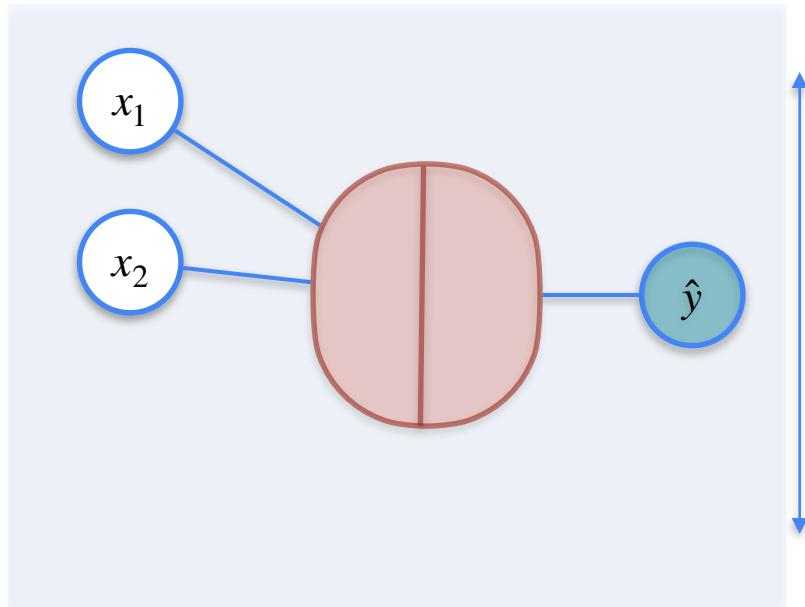
# Classification With a Perceptron

$$\frac{\partial L}{\partial b} = -(y - \hat{y})$$

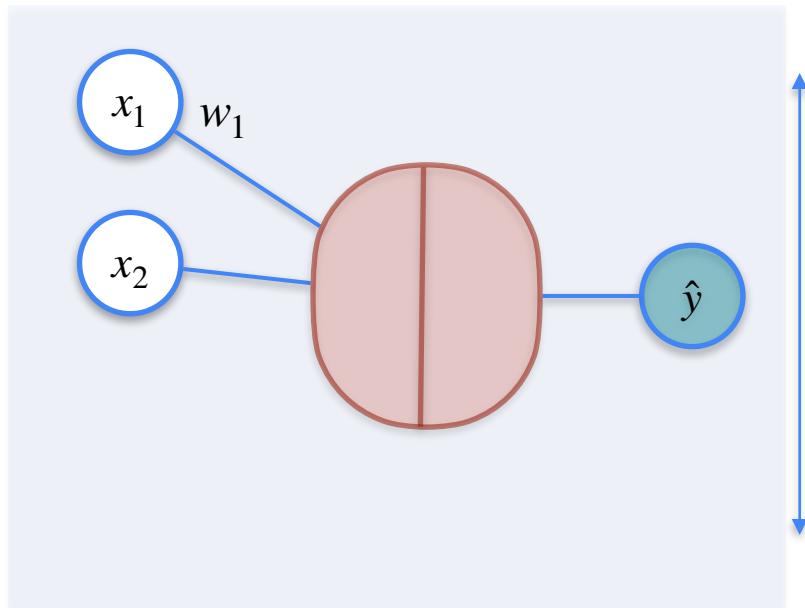
$$\frac{\partial L}{\partial w_1} = -(y - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = -(y - \hat{y})x_2$$

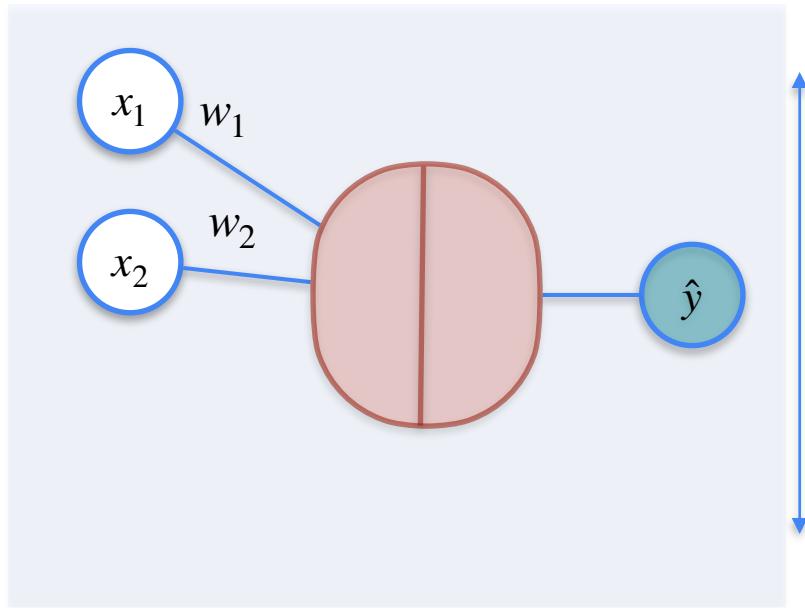
# Classification With a Perceptron



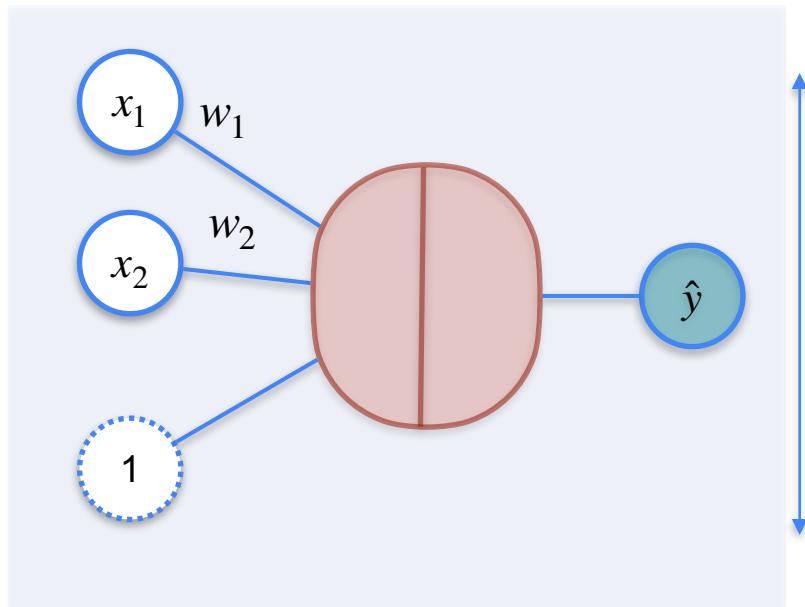
# Classification With a Perceptron



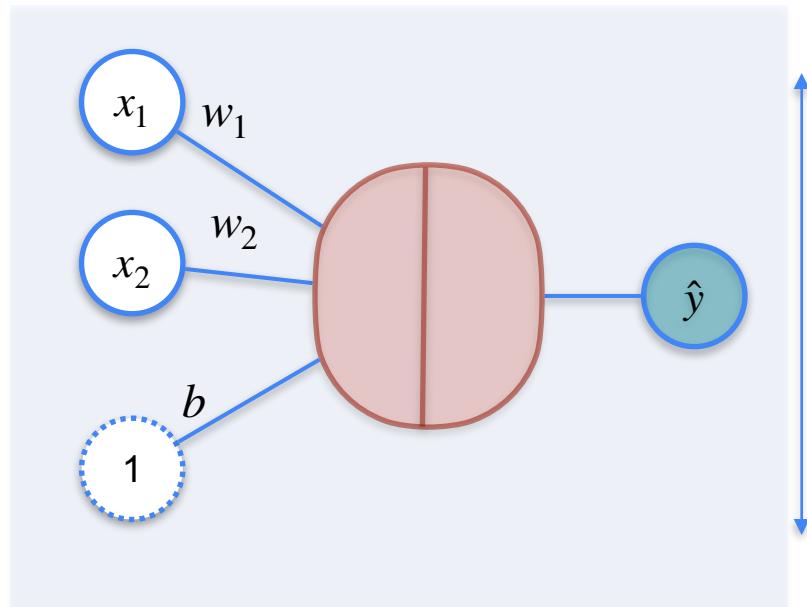
# Classification With a Perceptron



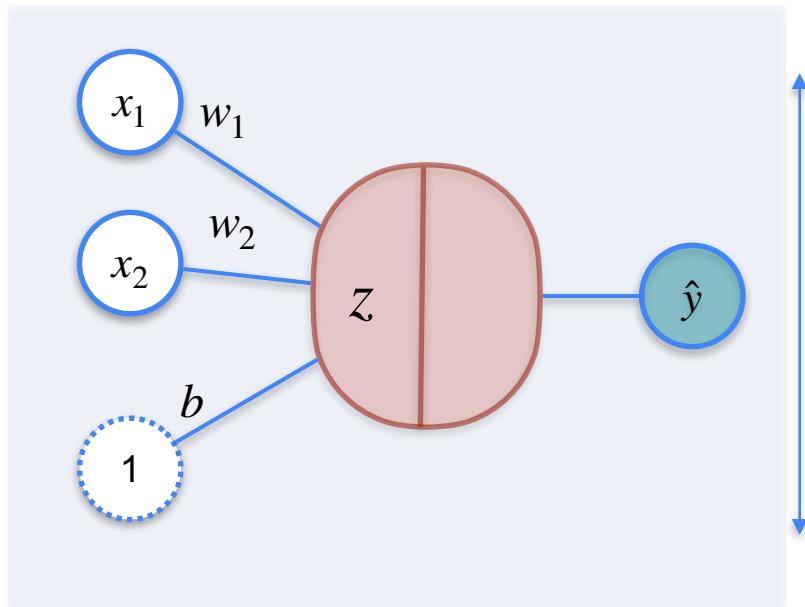
# Classification With a Perceptron



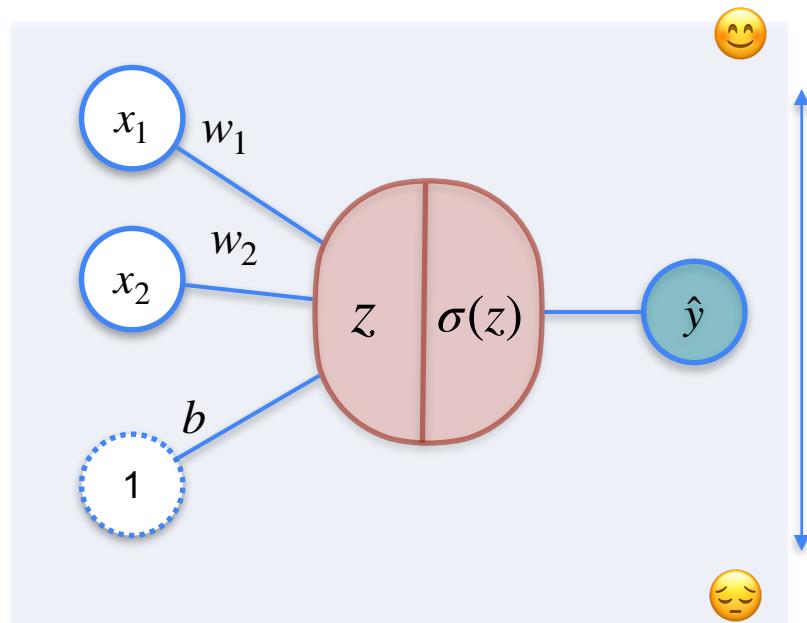
# Classification With a Perceptron



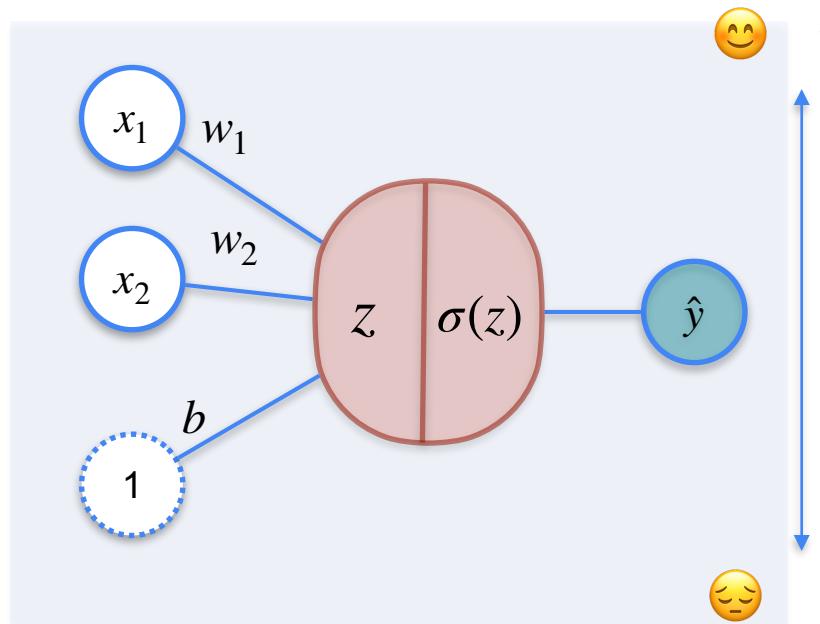
# Classification With a Perceptron



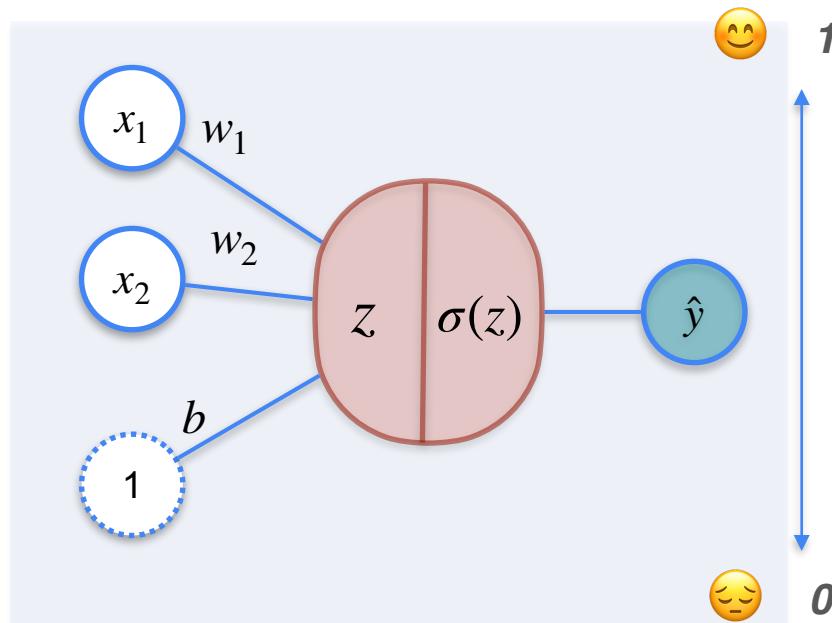
# Classification With a Perceptron



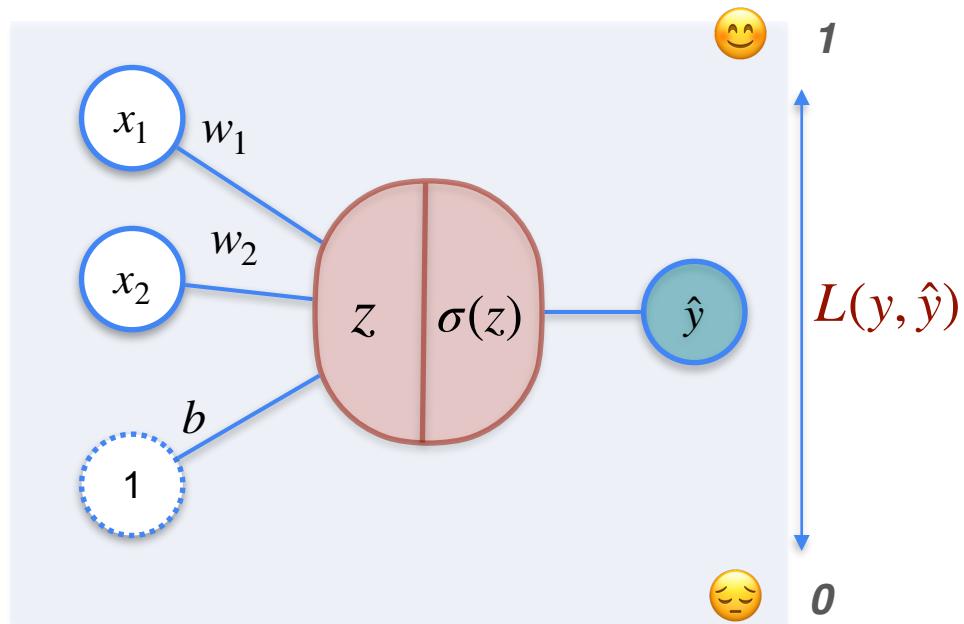
# Classification With a Perceptron



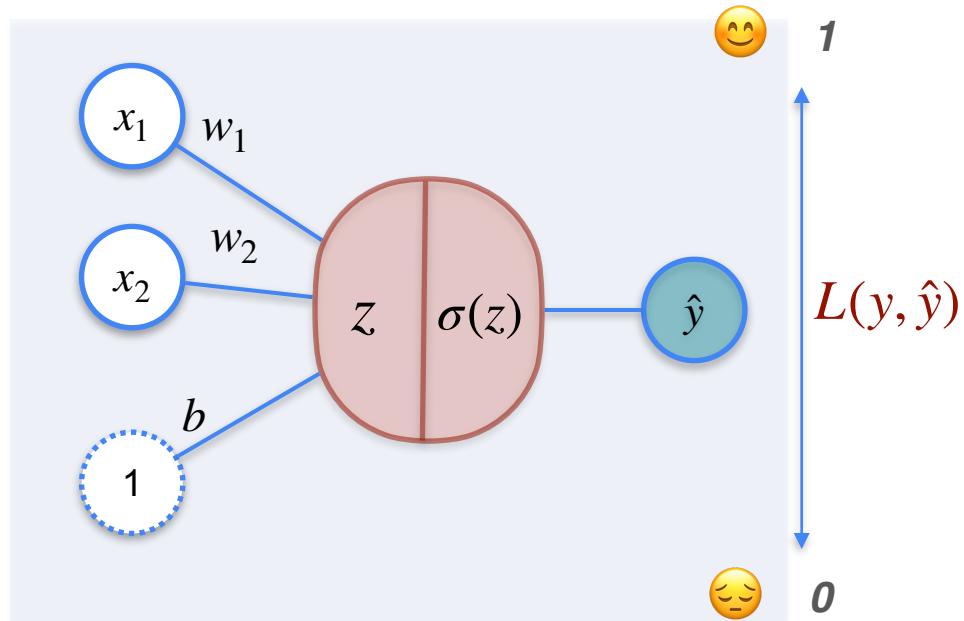
# Classification With a Perceptron



# Classification With a Perceptron

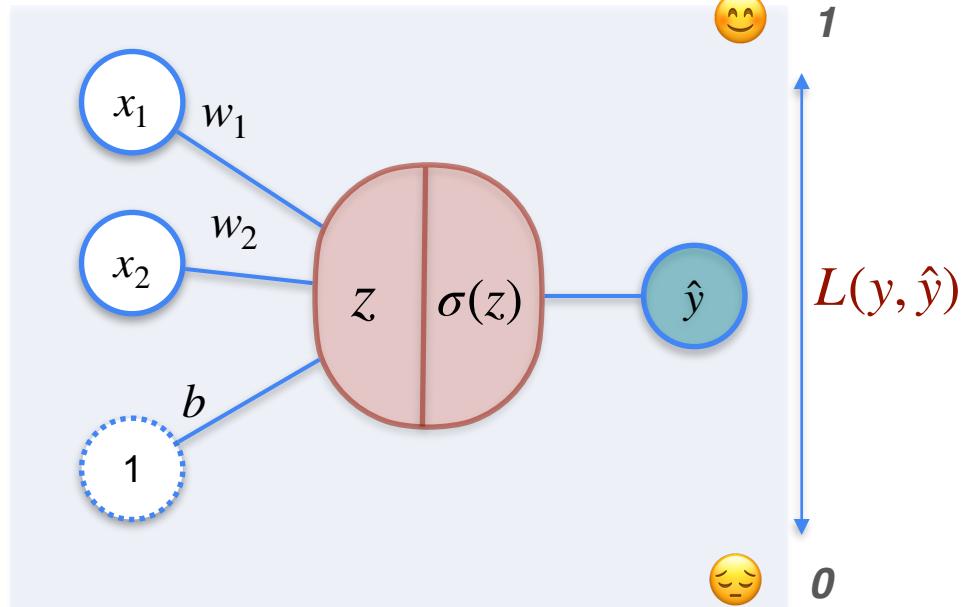


# Classification With a Perceptron



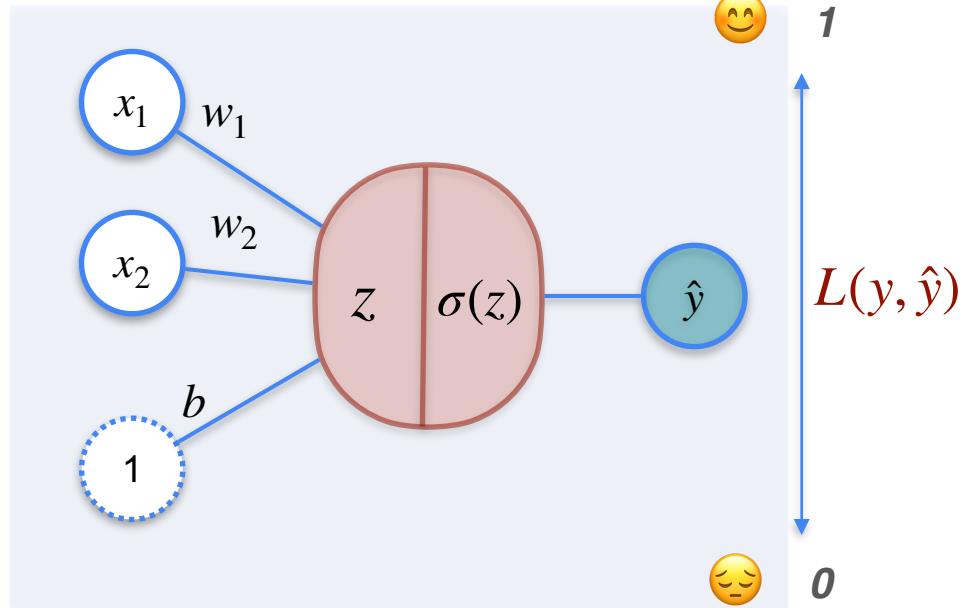
To find optimal values for:

# Classification With a Perceptron



To find optimal values for:  
 $w_1$  ,  $w_2$  ,  $b$

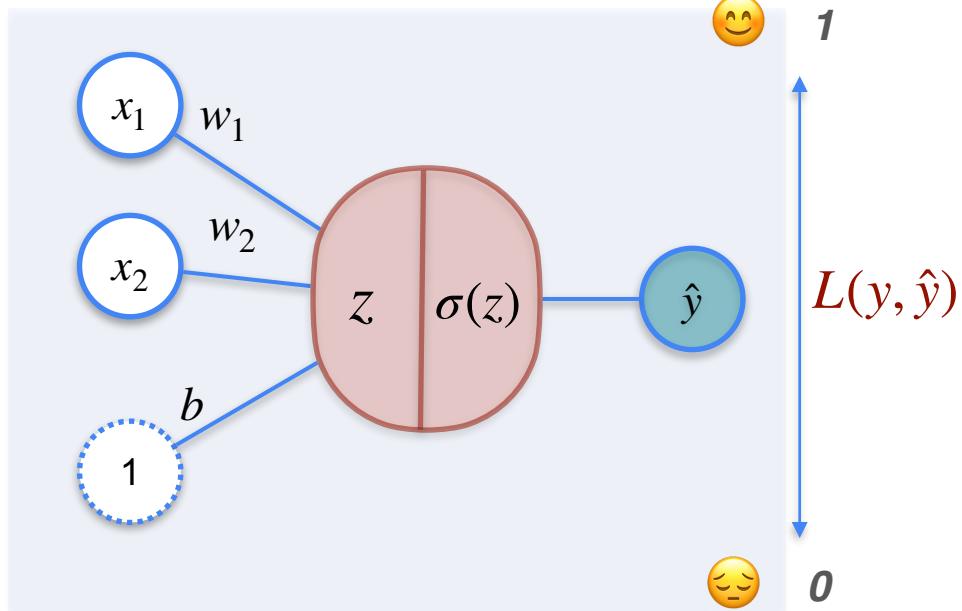
# Classification With a Perceptron



To find optimal values for:  
 $w_1$  ,  $w_2$  ,  $b$

*You need gradient descent*

# Classification With a Perceptron

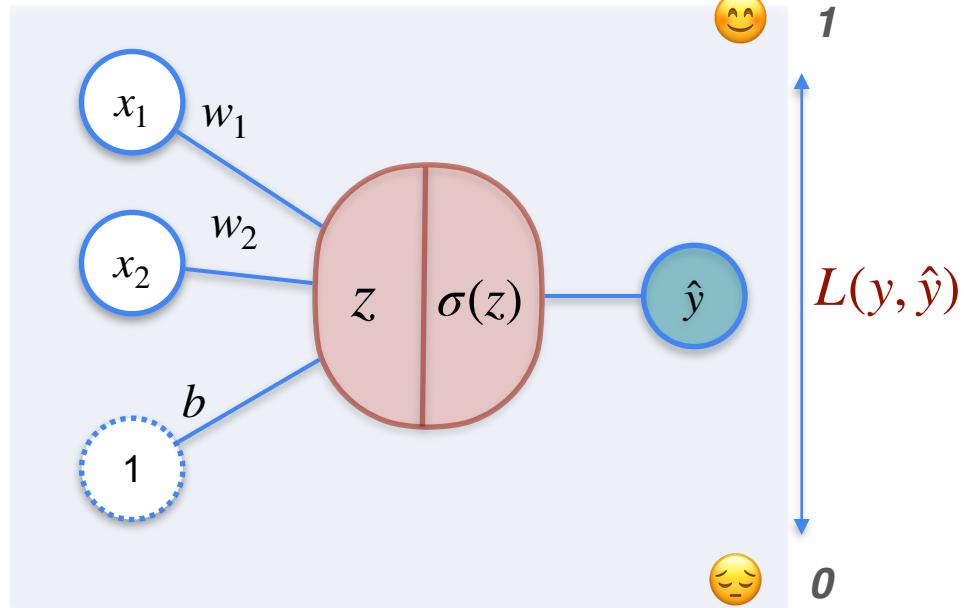


To find optimal values for:  
 $w_1$  ,  $w_2$  ,  $b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

# Classification With a Perceptron

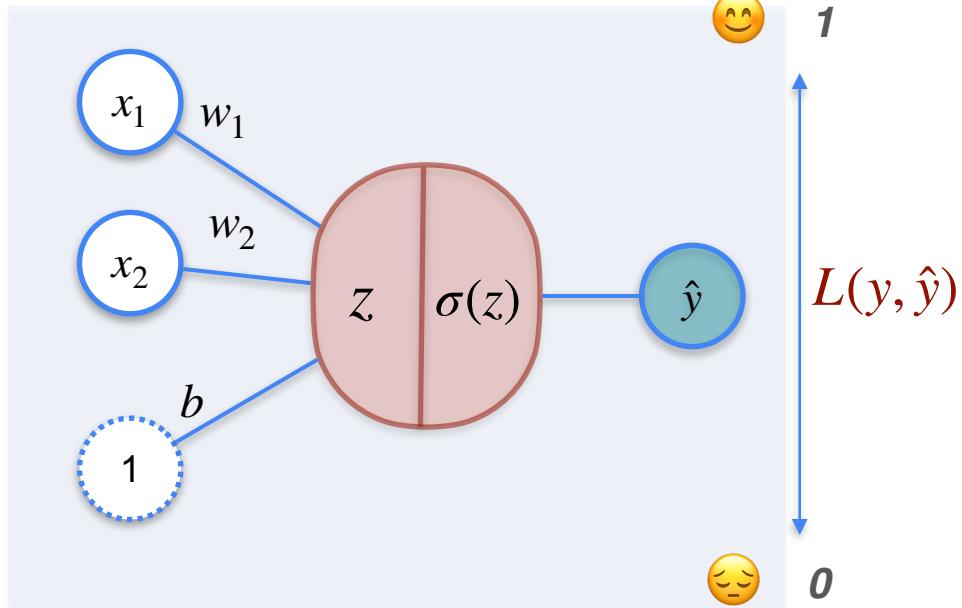


To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha$$

# Classification With a Perceptron

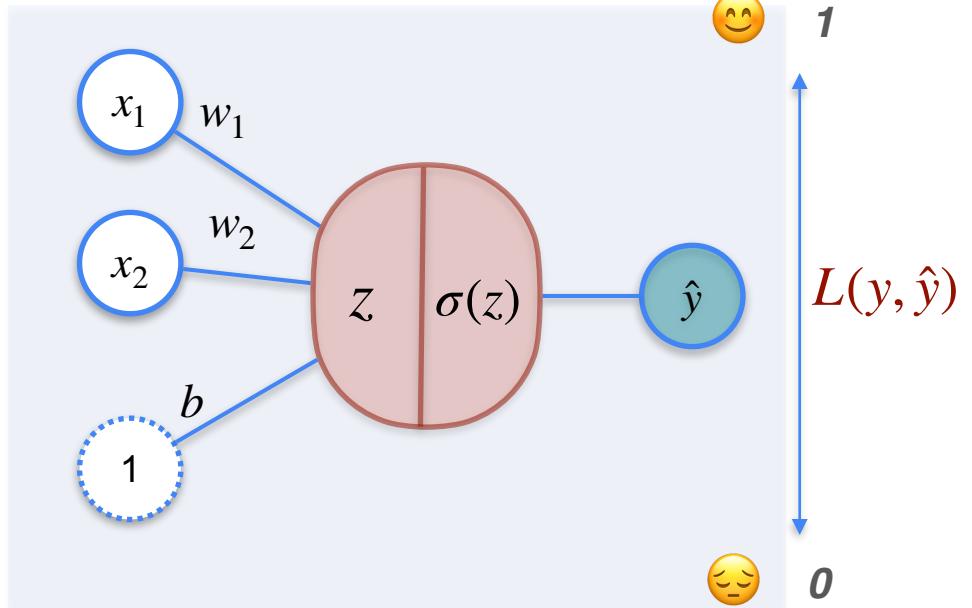


To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

# Classification With a Perceptron



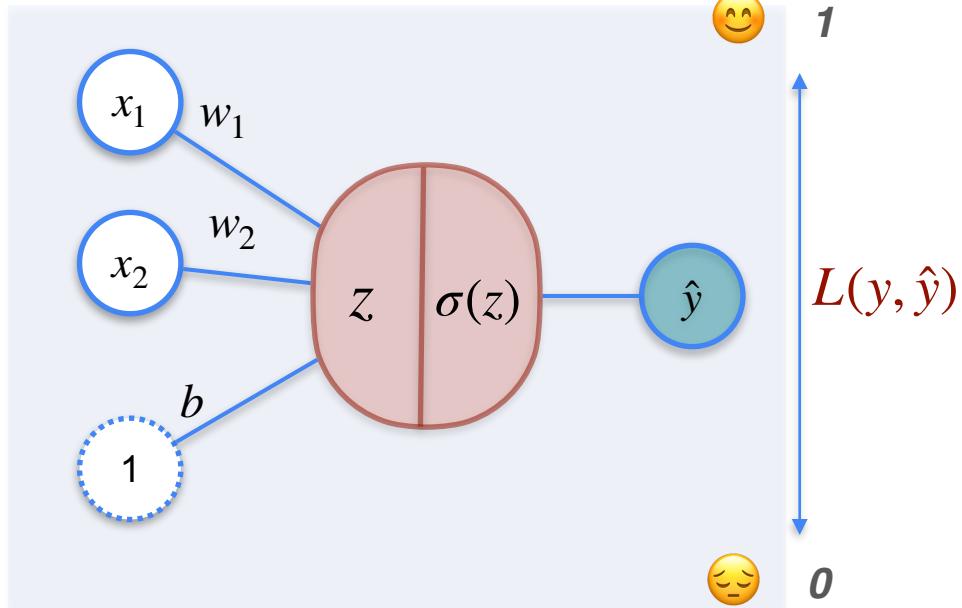
To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

# Classification With a Perceptron



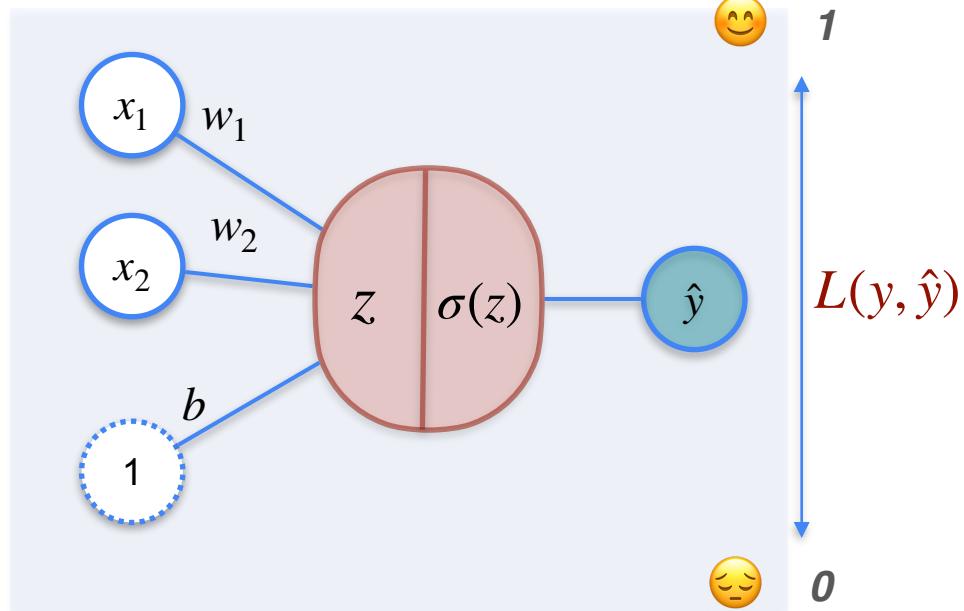
To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 \rightarrow w_2 - \alpha$$

# Classification With a Perceptron



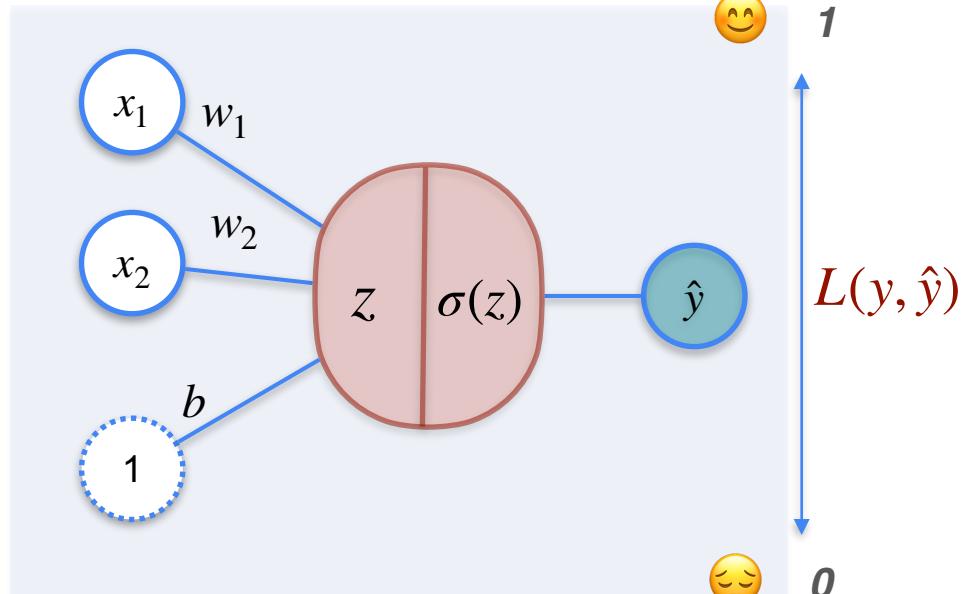
To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 \rightarrow w_2 - \alpha(-x_2(y - \hat{y}))$$

# Classification With a Perceptron



To find optimal values for:  
 $w_1, w_2, b$

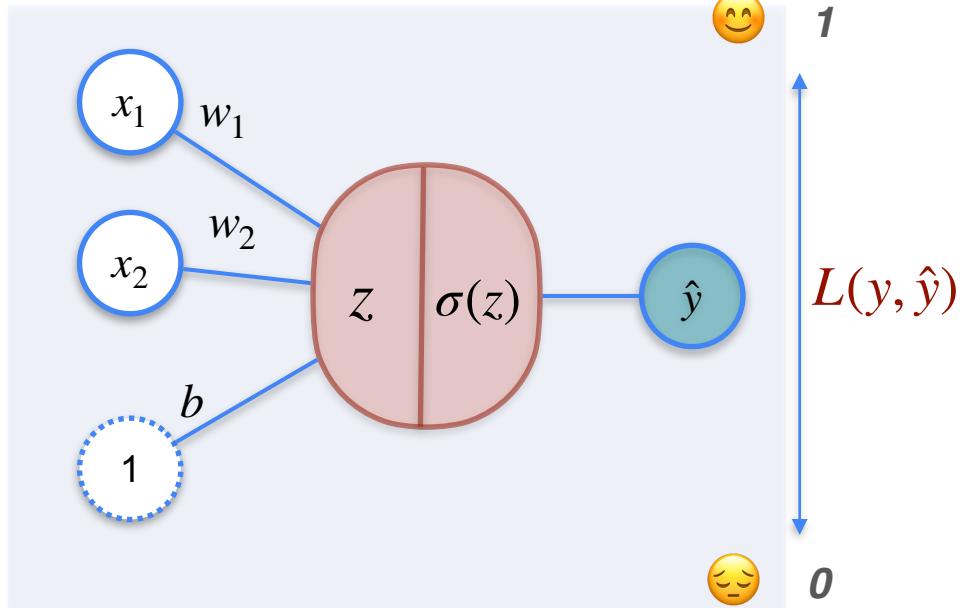
*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 \rightarrow w_2 - \alpha(-x_2(y - \hat{y}))$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

# Classification With a Perceptron



To find optimal values for:  
 $w_1, w_2, b$

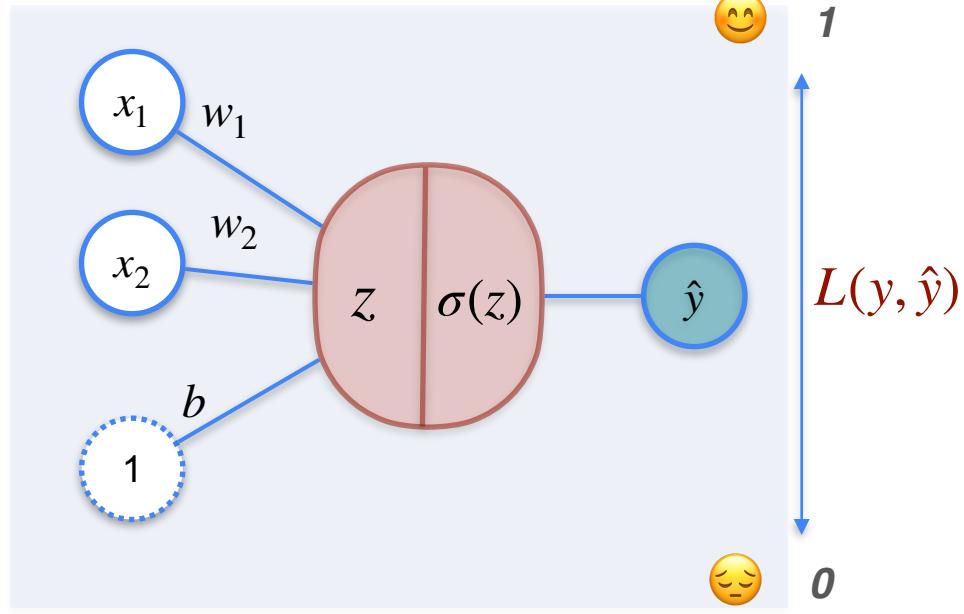
*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 \rightarrow w_2 - \alpha(-x_2(y - \hat{y}))$$

$$b \rightarrow b - \alpha$$

# Classification With a Perceptron



To find optimal values for:  
 $w_1, w_2, b$

*You need gradient descent*

$$w_1 \rightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

$$w_2 \rightarrow w_2 - \alpha(-x_2(y - \hat{y}))$$

$$b \rightarrow b - \alpha(-(y - \hat{y}))$$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Classification with a Neural Network

# Classification Problem Motivation

# Classification Problem Motivation

<i>Sentence</i>	<i>Aack</i>	<i>Beep</i>	<i>Mood</i>
<i>Aack aack aack!</i>	3	0	<i>Happy</i> 😊

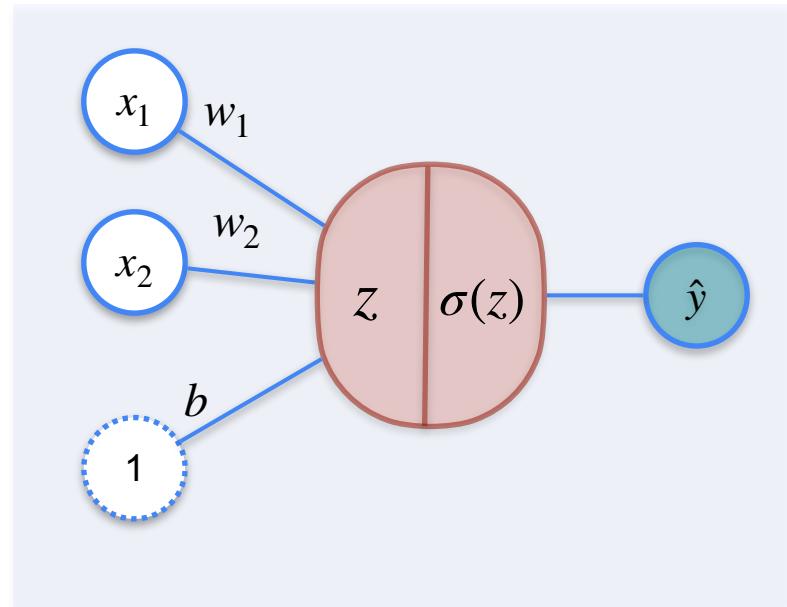
*Beep beep!*      0      2      *Sad* 😞

<i>Aack beep beep beep!</i>	1	3	<i>Sad</i> 😞
-----------------------------	---	---	--------------

*Aack beep aack!*      2      1      *Happy* 😊

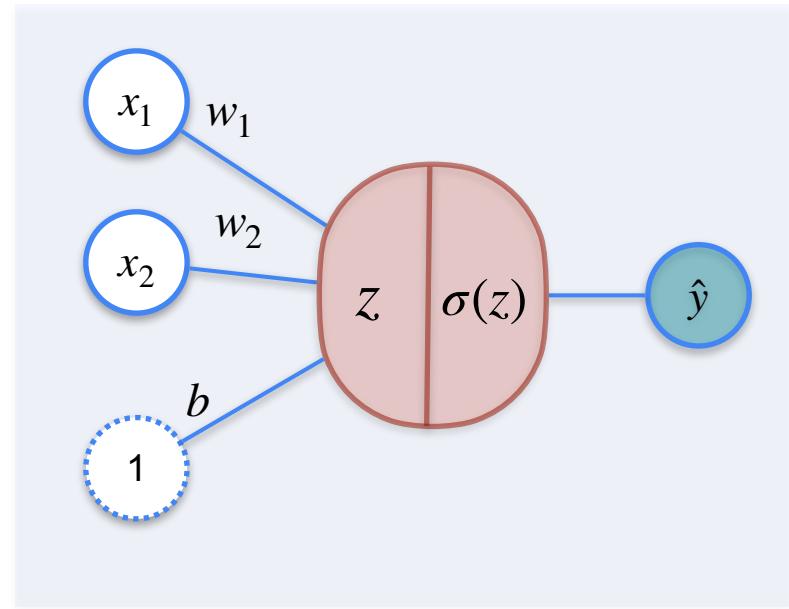
# Classification Problem Motivation

Sentence	Aack	Beep	Mood
Aack aack aack!	3	0	Happy 😊
Beep beep!	0	2	Sad 😞
Aack beep beep beep!	1	3	Sad 😞
Aack beep aack!	2	1	Happy 😊



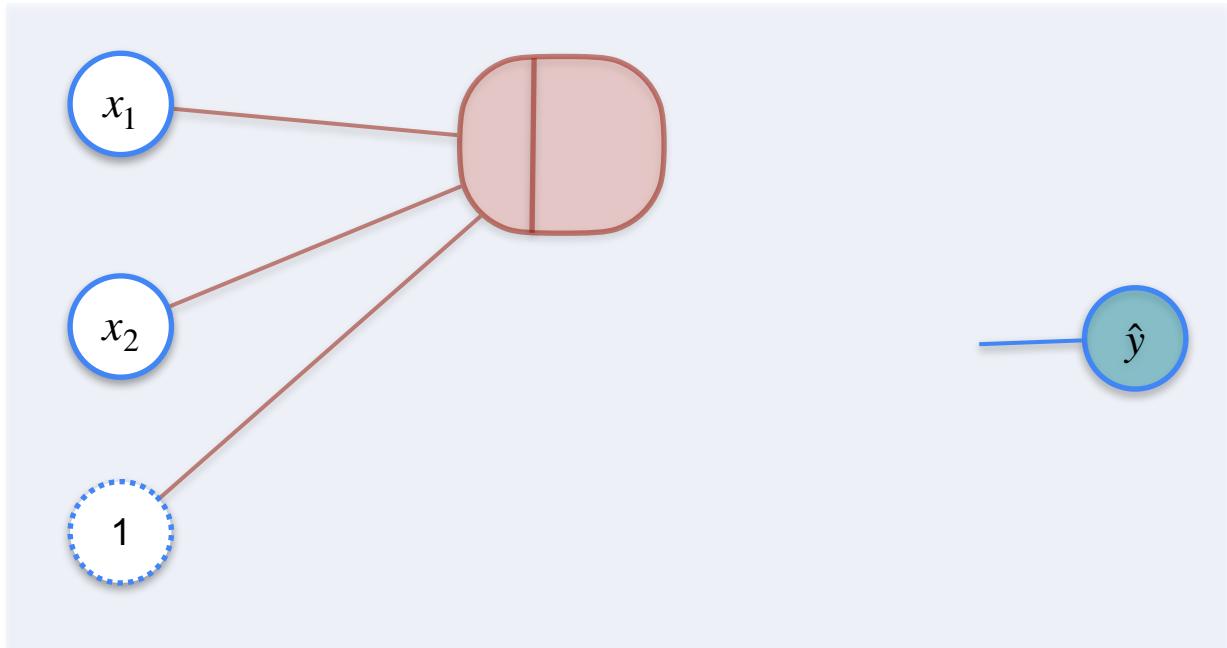
# Classification Problem Motivation

Sentence	Aack	Beep	Mood
Aack aack aack!	3	0	Happy 😊
Beep beep!	0	2	Sad 😞
Aack beep beep beep!	1	3	Sad 😞
Aack beep aack!	2	1	Happy 😊

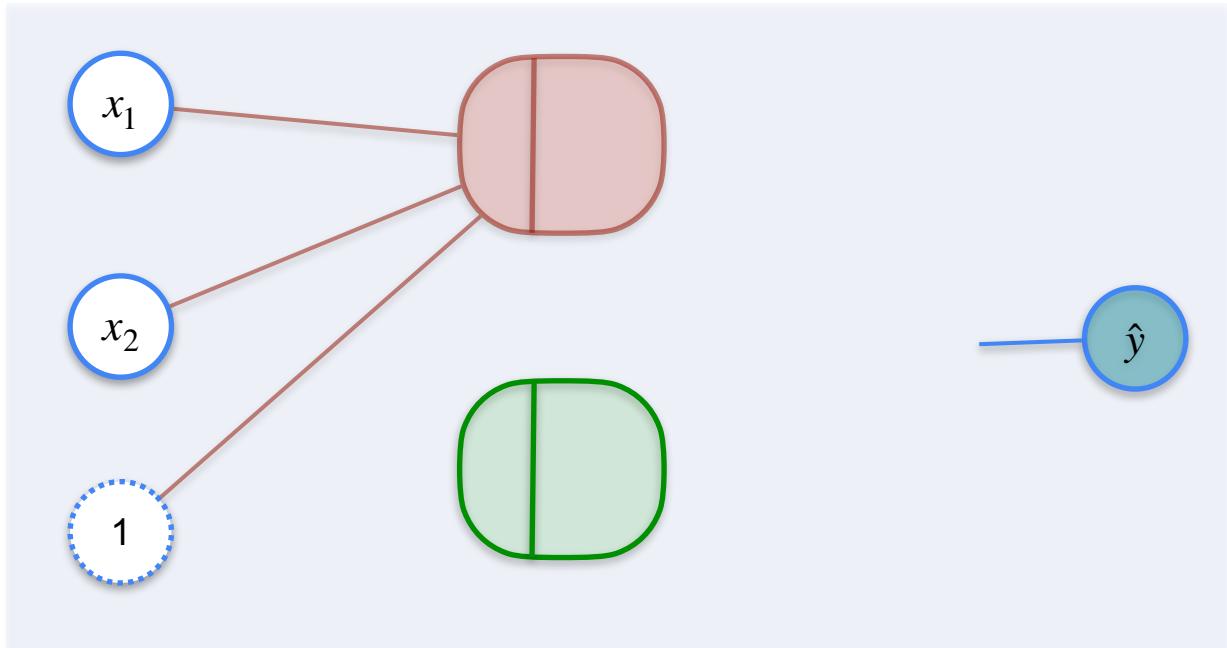


$$z = x_1 w_1 + x_2 w_2 + b$$

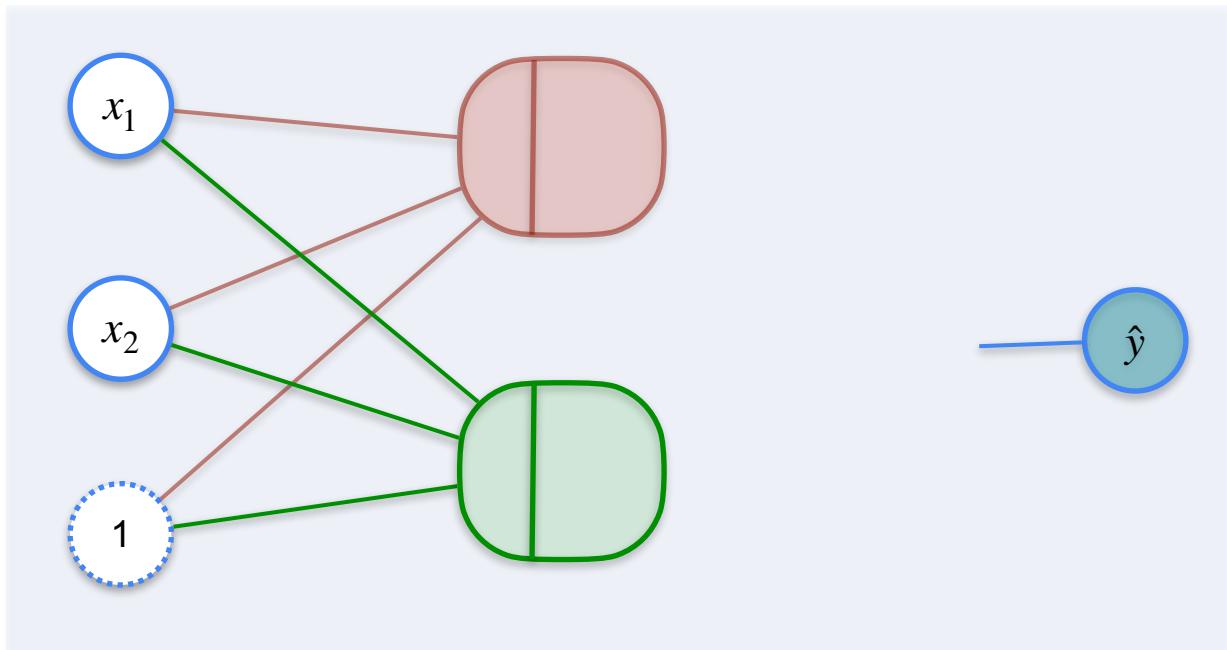
# 2,2,1 Neural Network



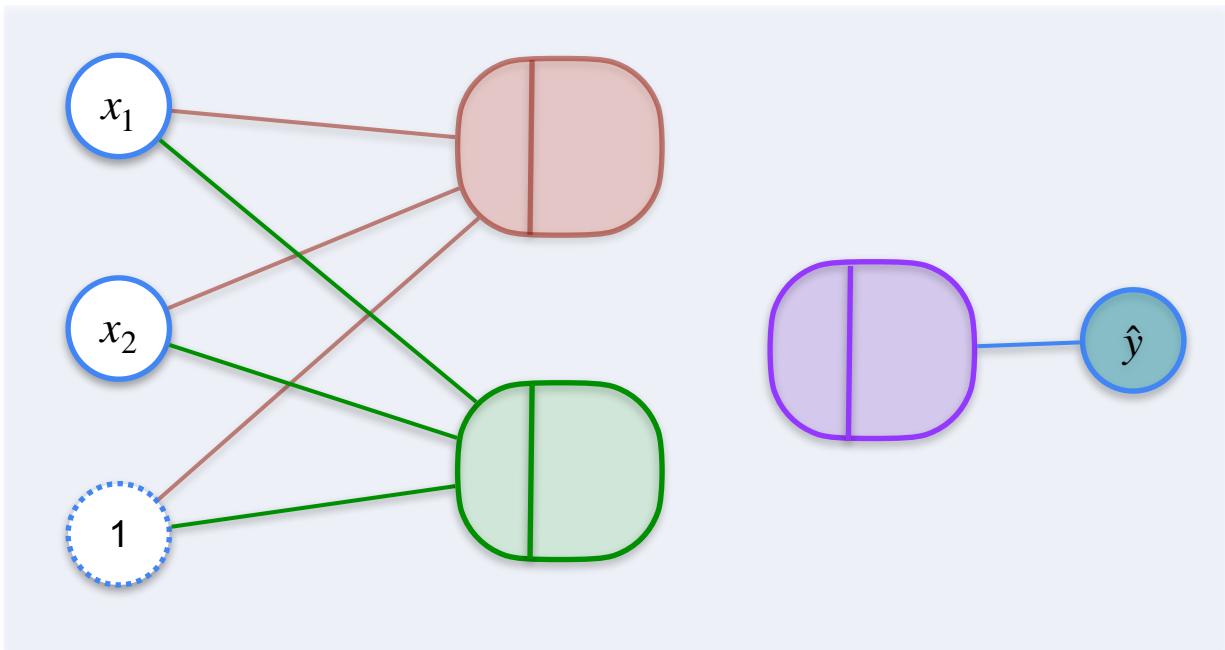
# 2,2,1 Neural Network



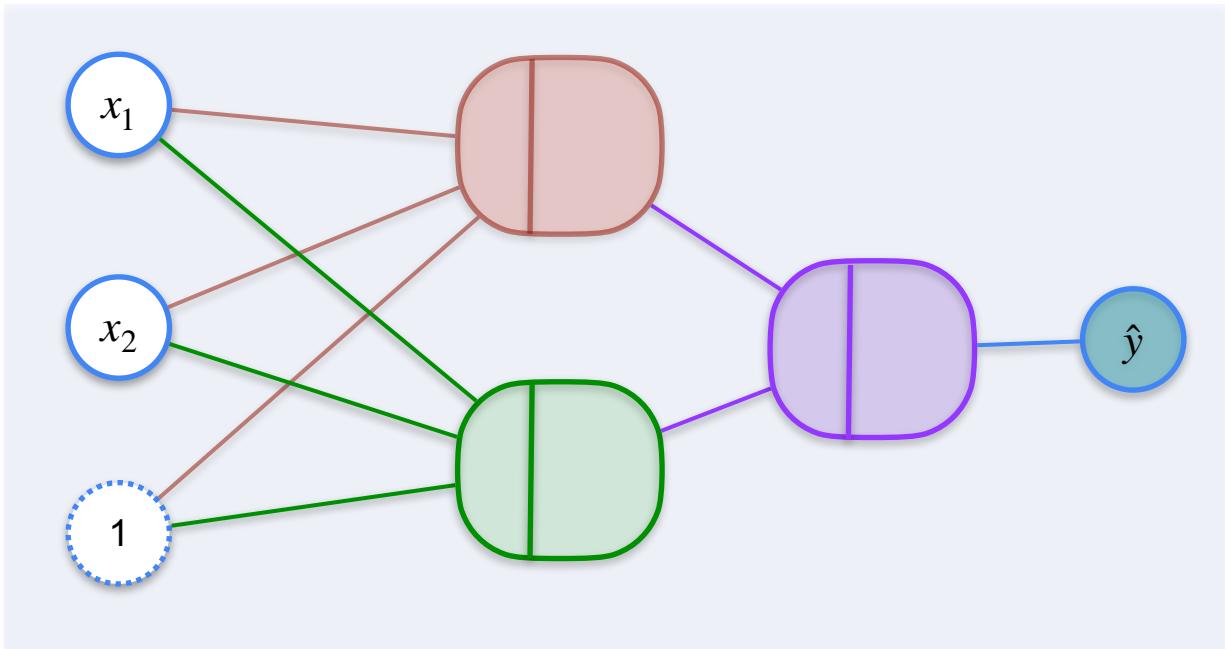
# 2,2,1 Neural Network



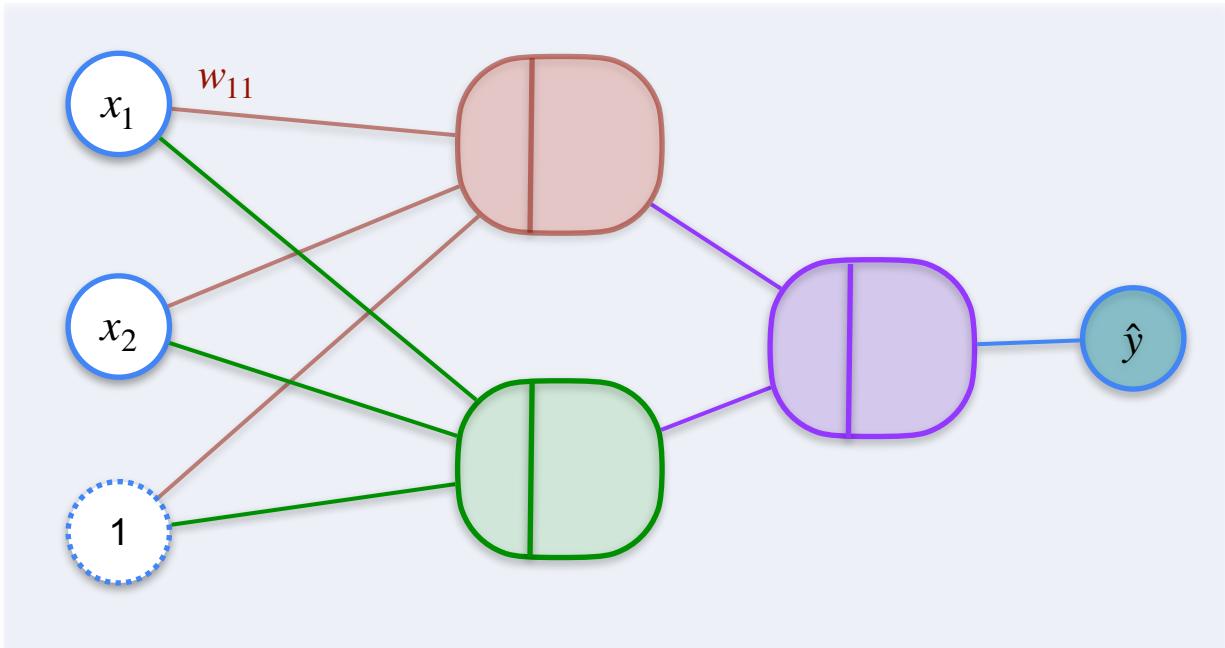
# 2,2,1 Neural Network



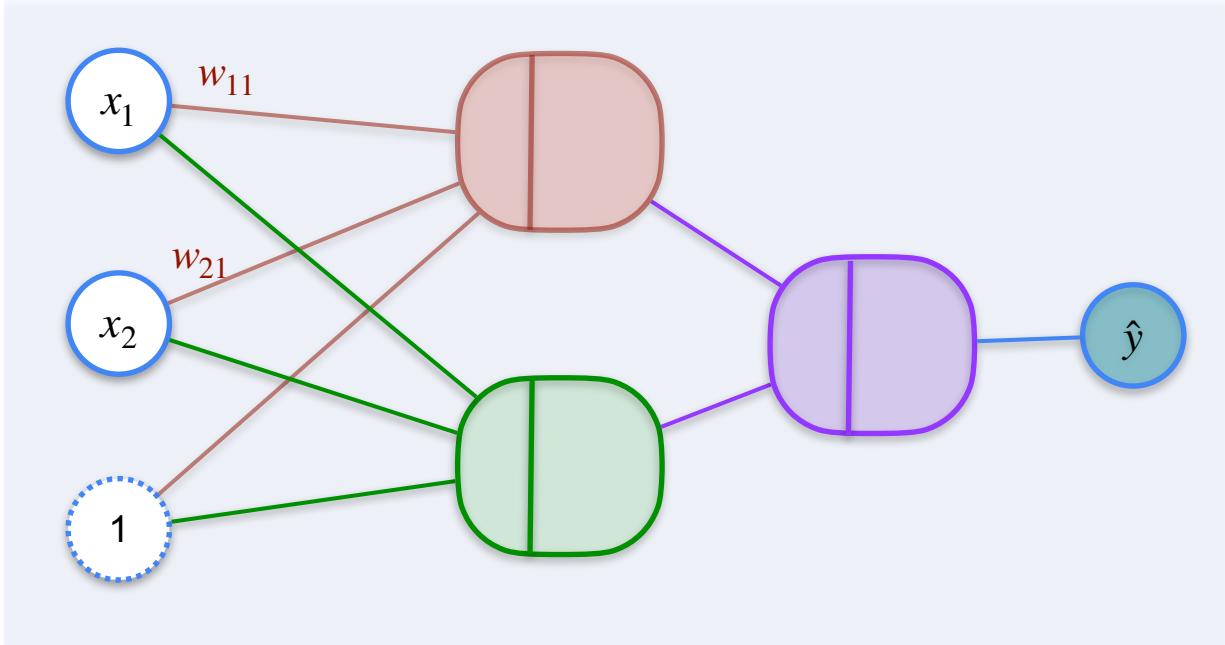
# 2,2,1 Neural Network



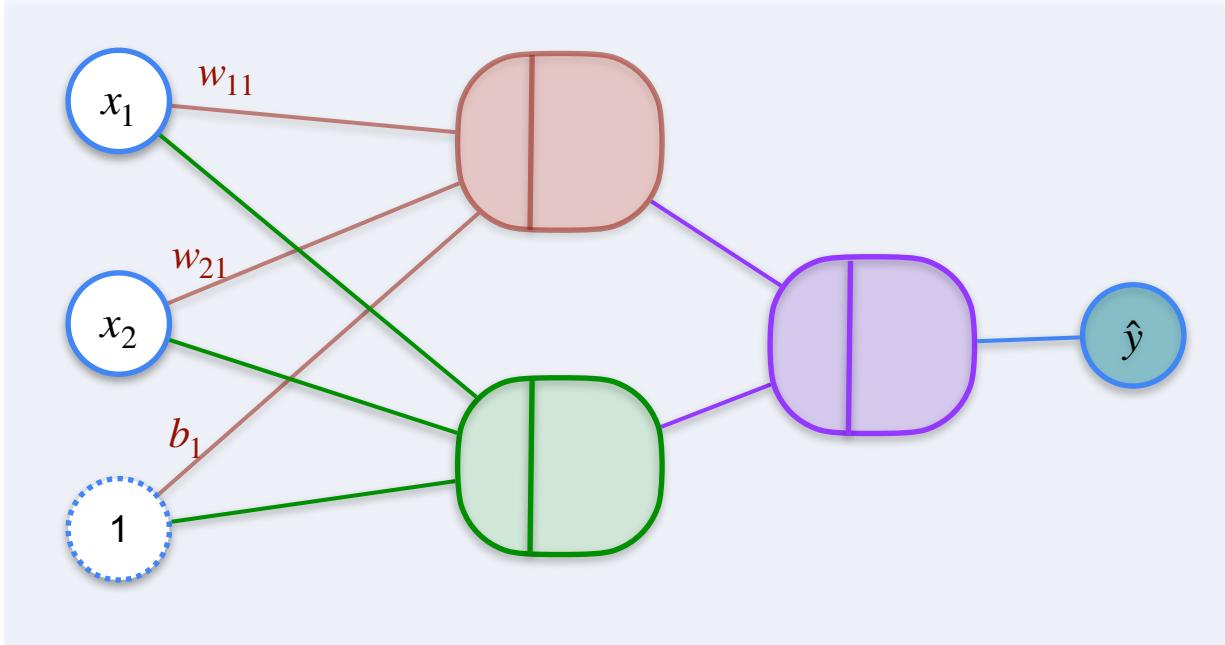
# 2,2,1 Neural Network



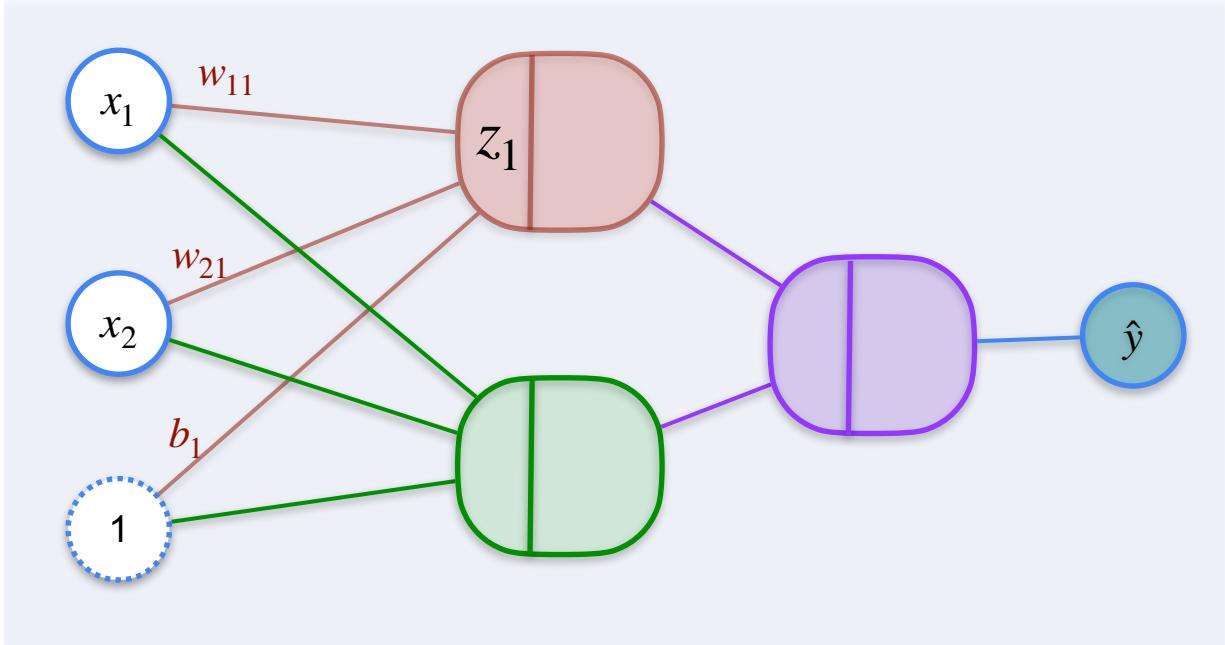
# 2,2,1 Neural Network



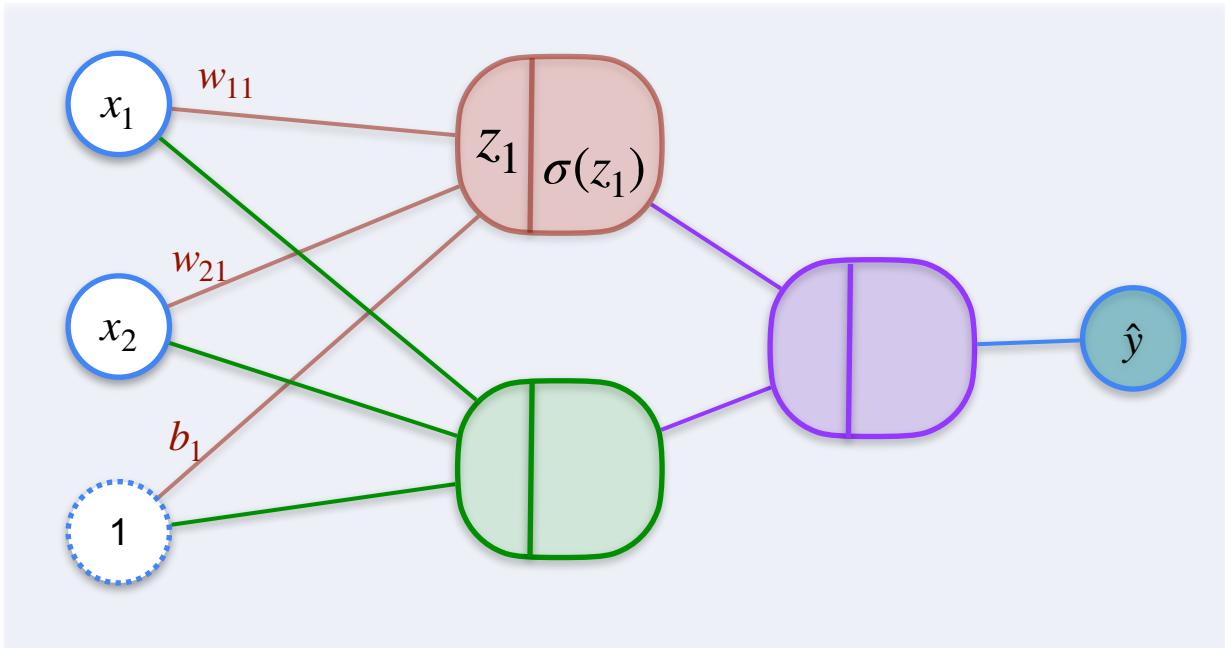
# 2,2,1 Neural Network



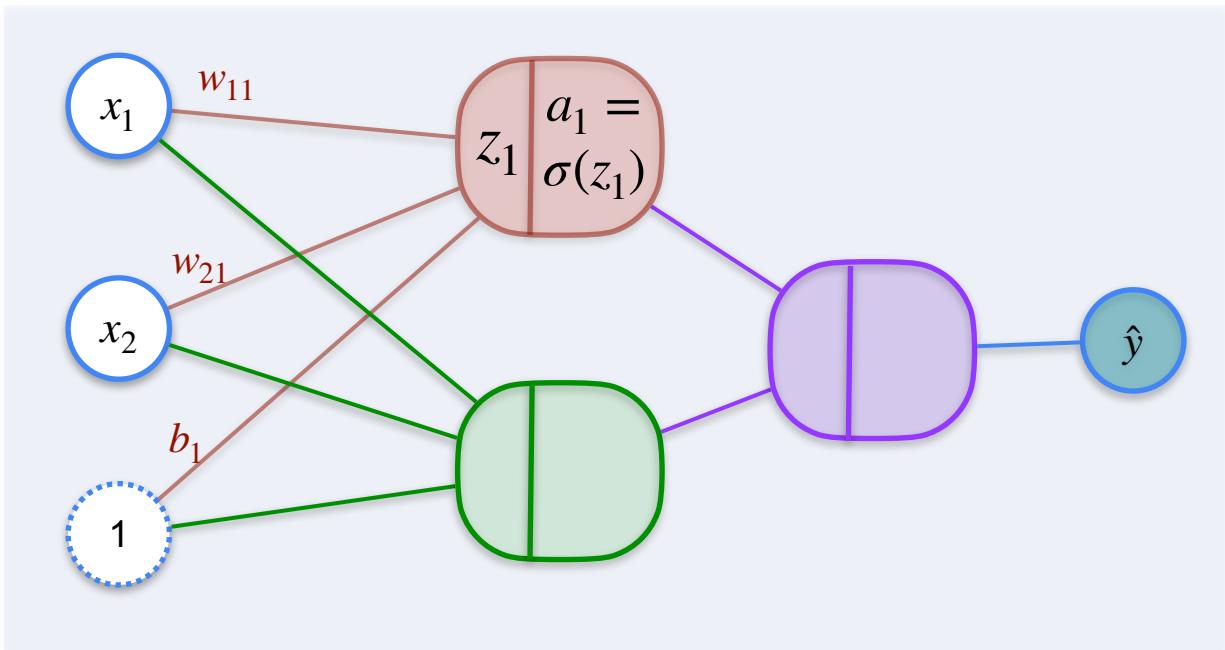
# 2,2,1 Neural Network



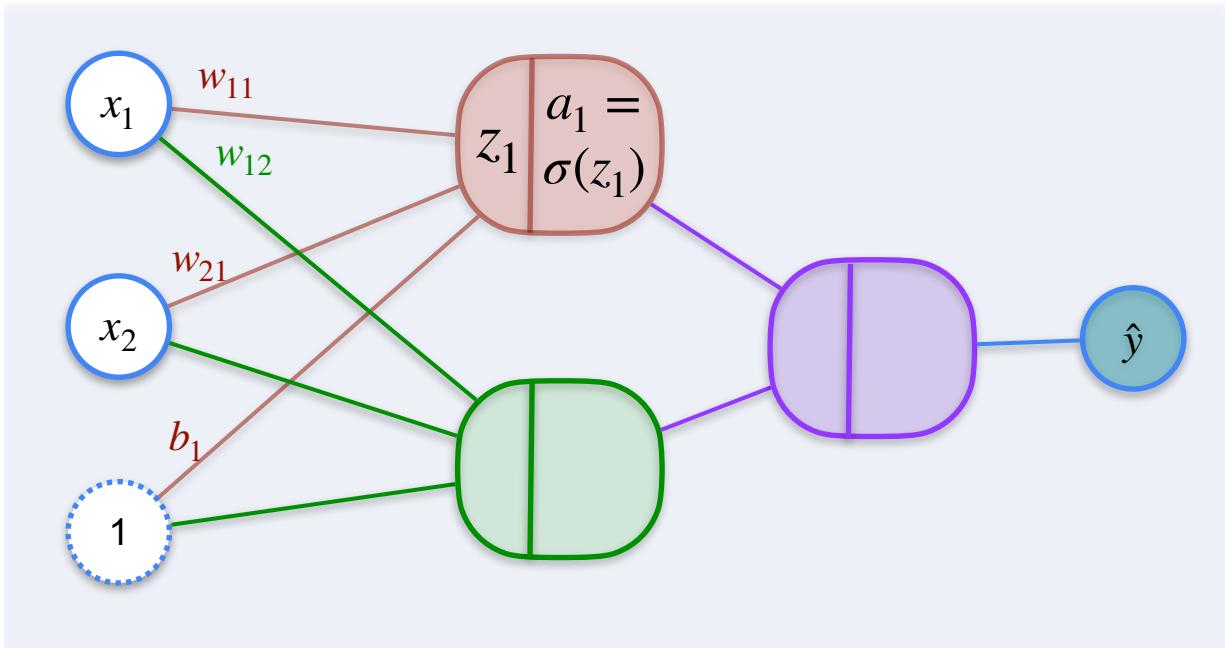
# 2,2,1 Neural Network



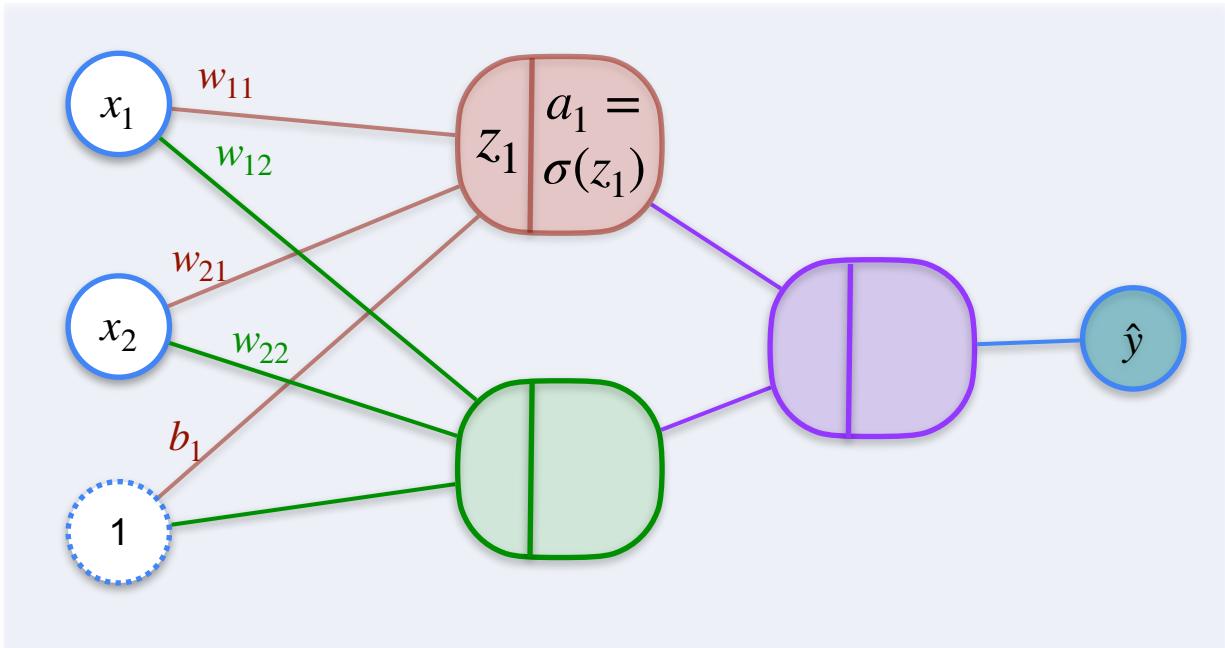
# 2,2,1 Neural Network



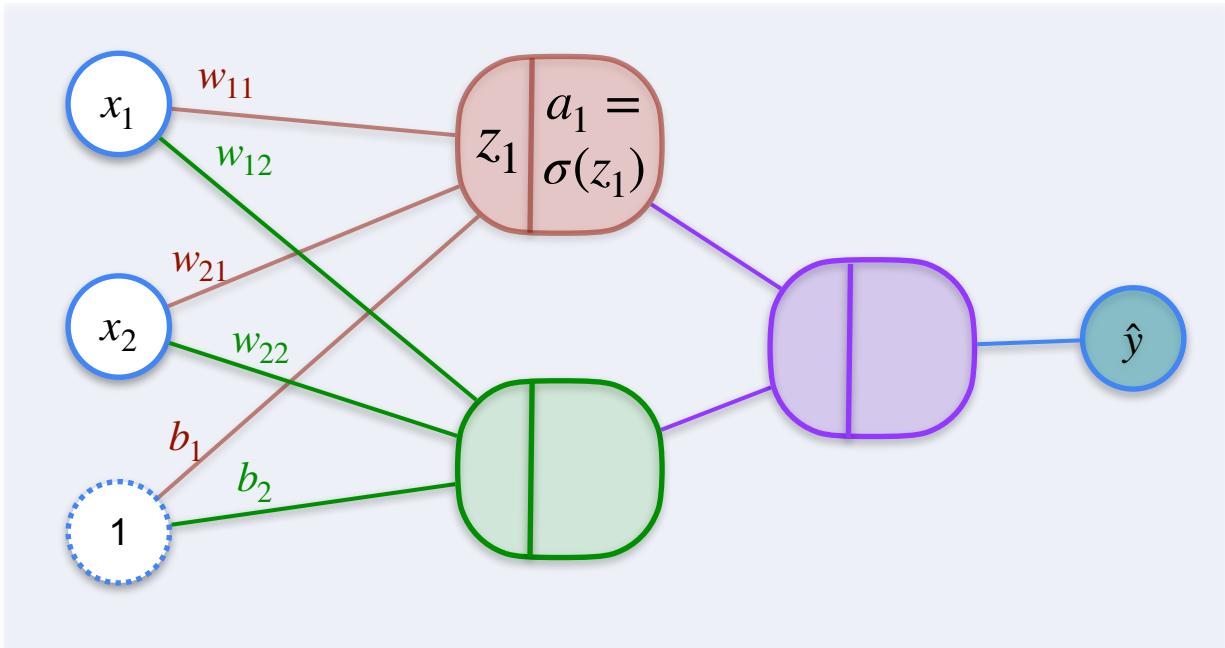
# 2,2,1 Neural Network



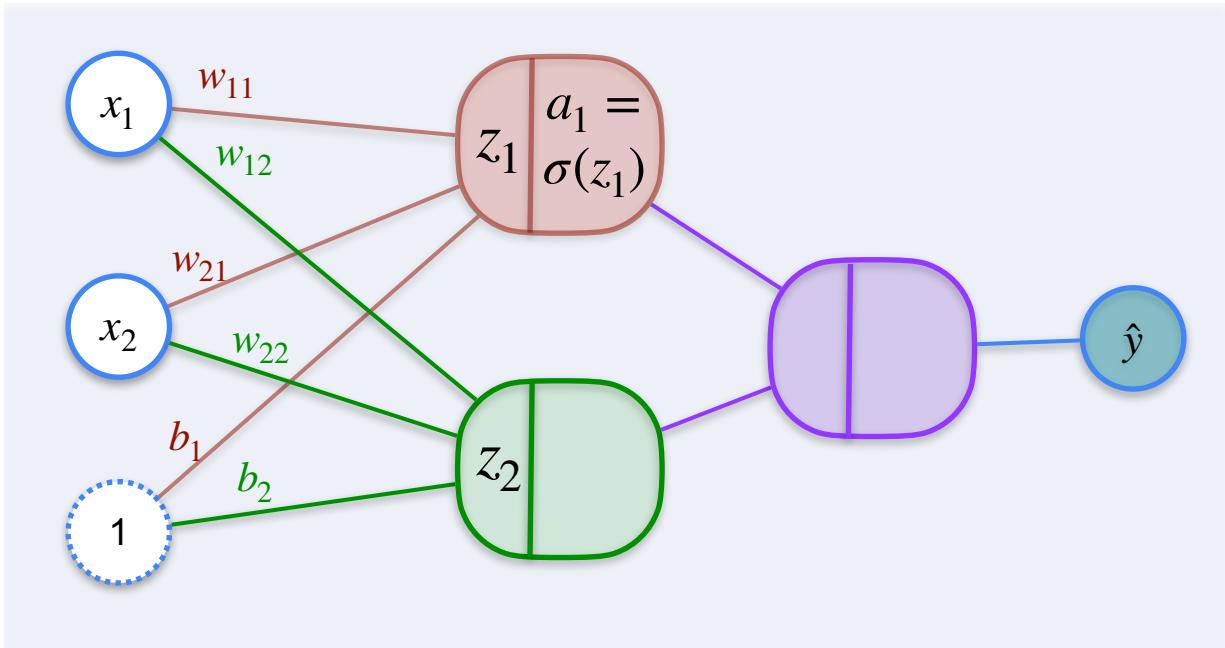
# 2,2,1 Neural Network



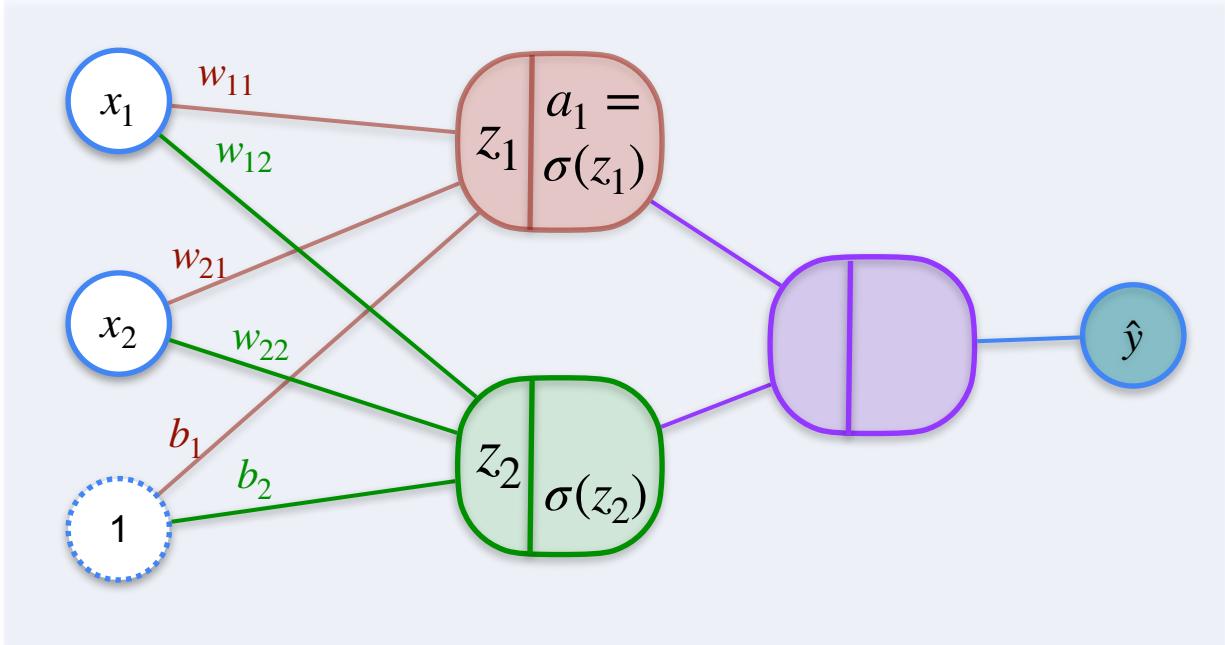
# 2,2,1 Neural Network



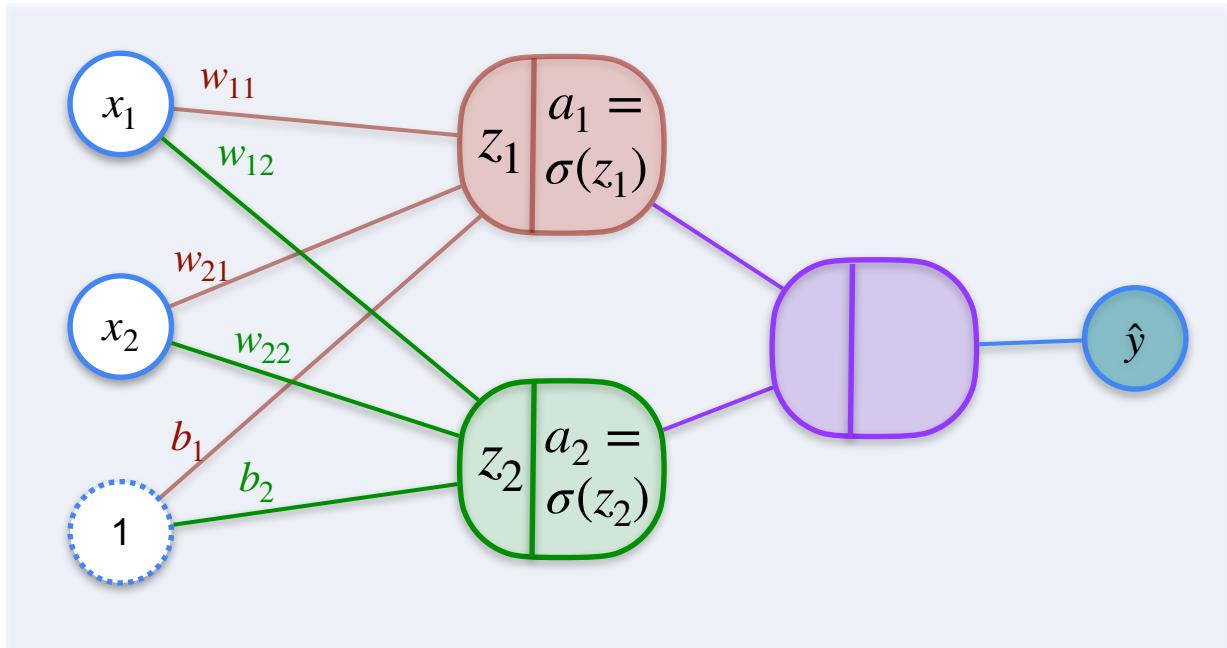
# 2,2,1 Neural Network



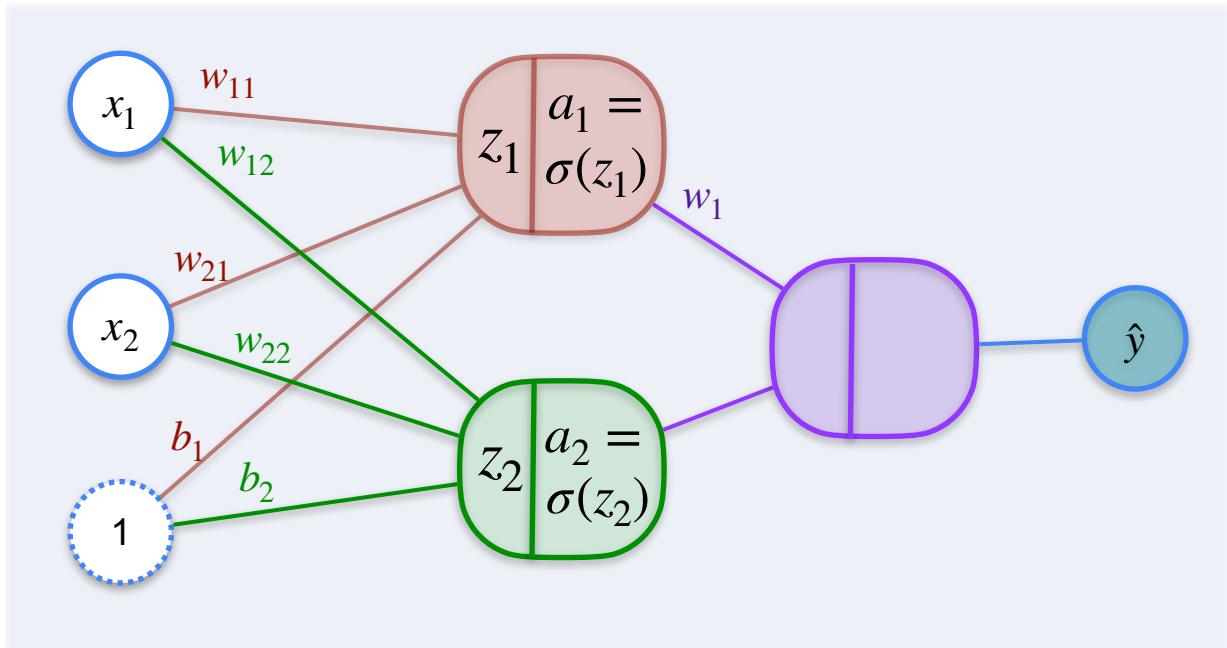
# 2,2,1 Neural Network



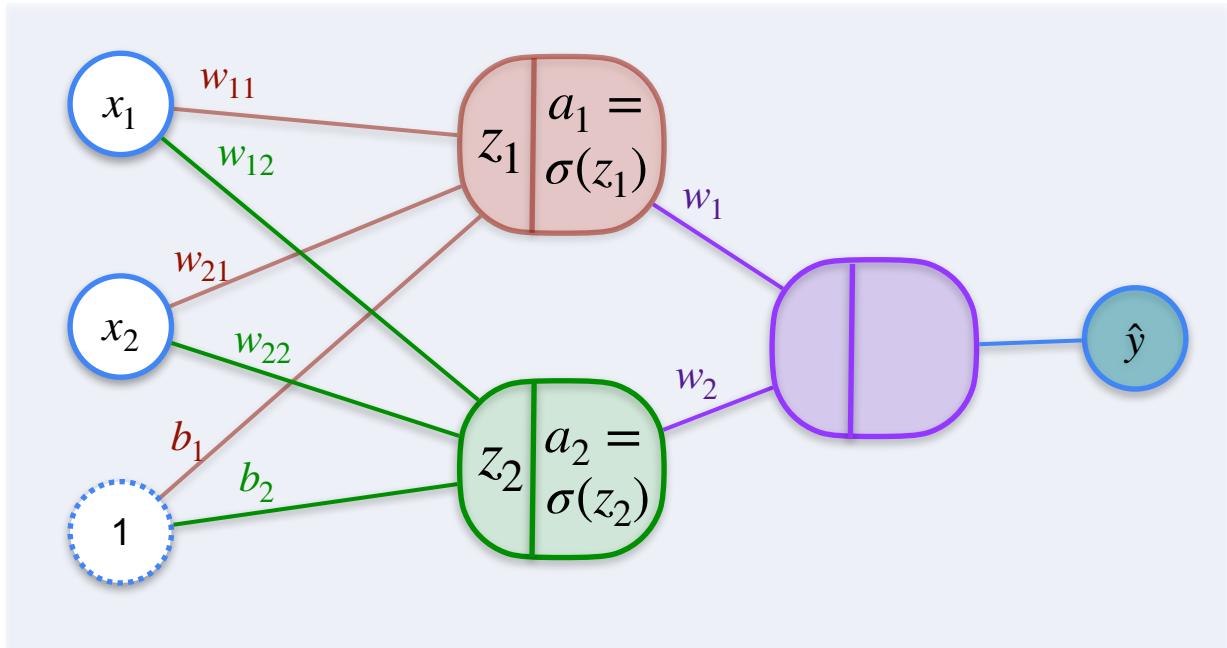
# 2,2,1 Neural Network



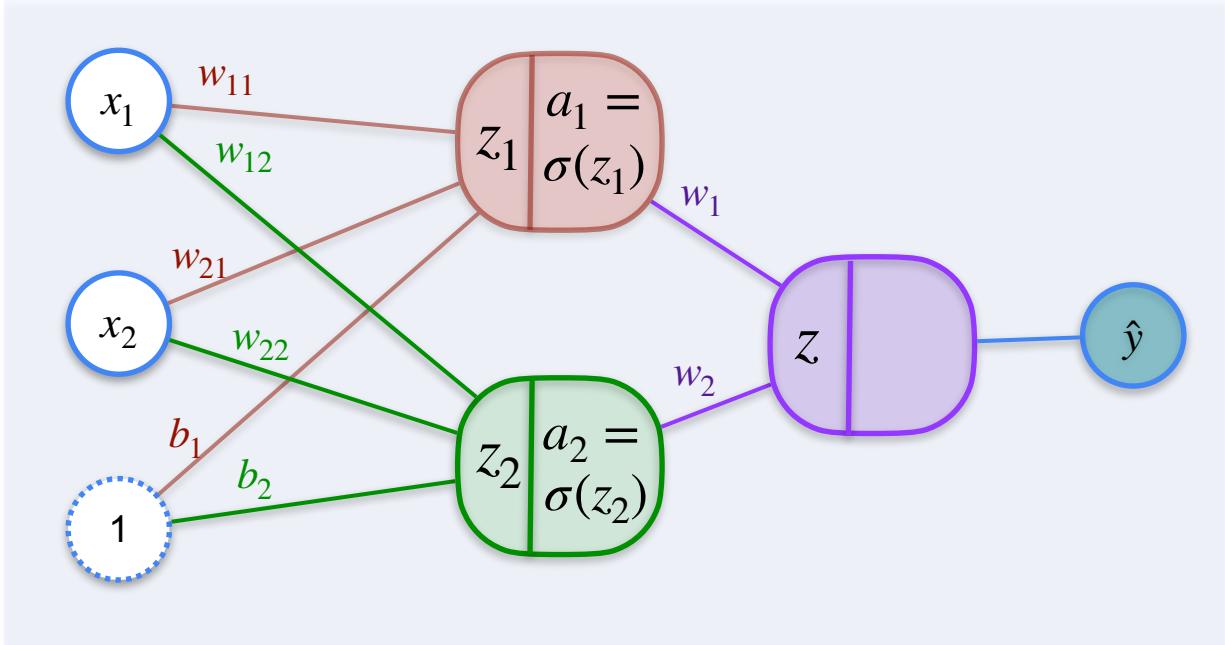
# 2,2,1 Neural Network



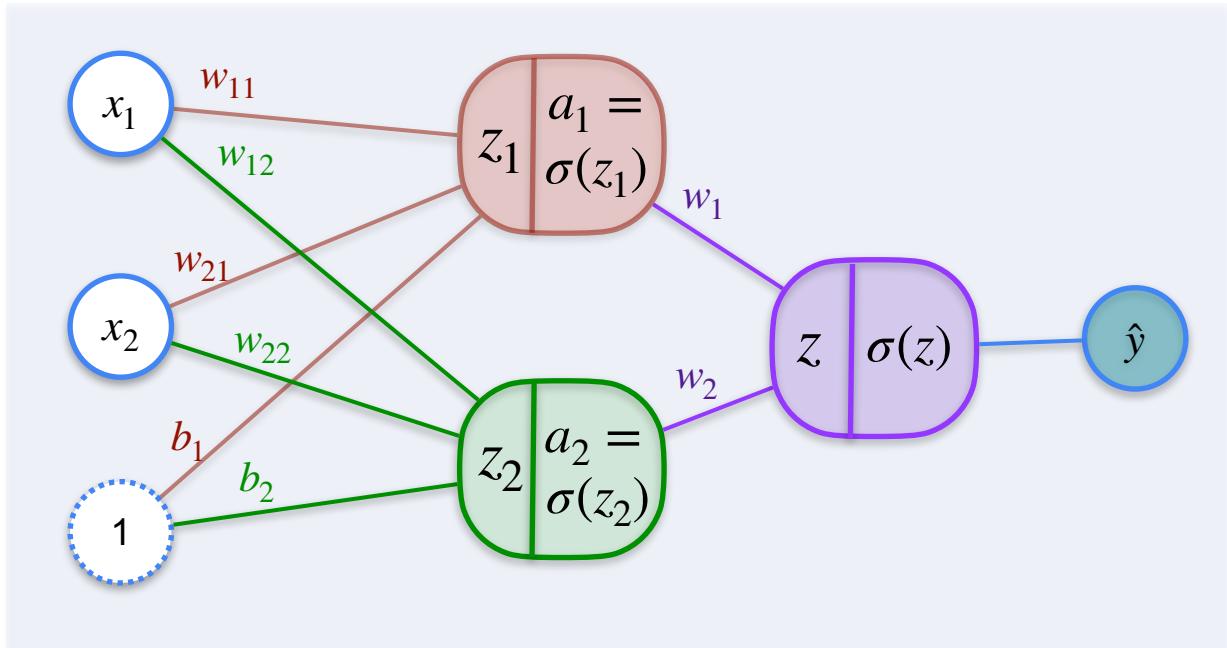
# 2,2,1 Neural Network



# 2,2,1 Neural Network



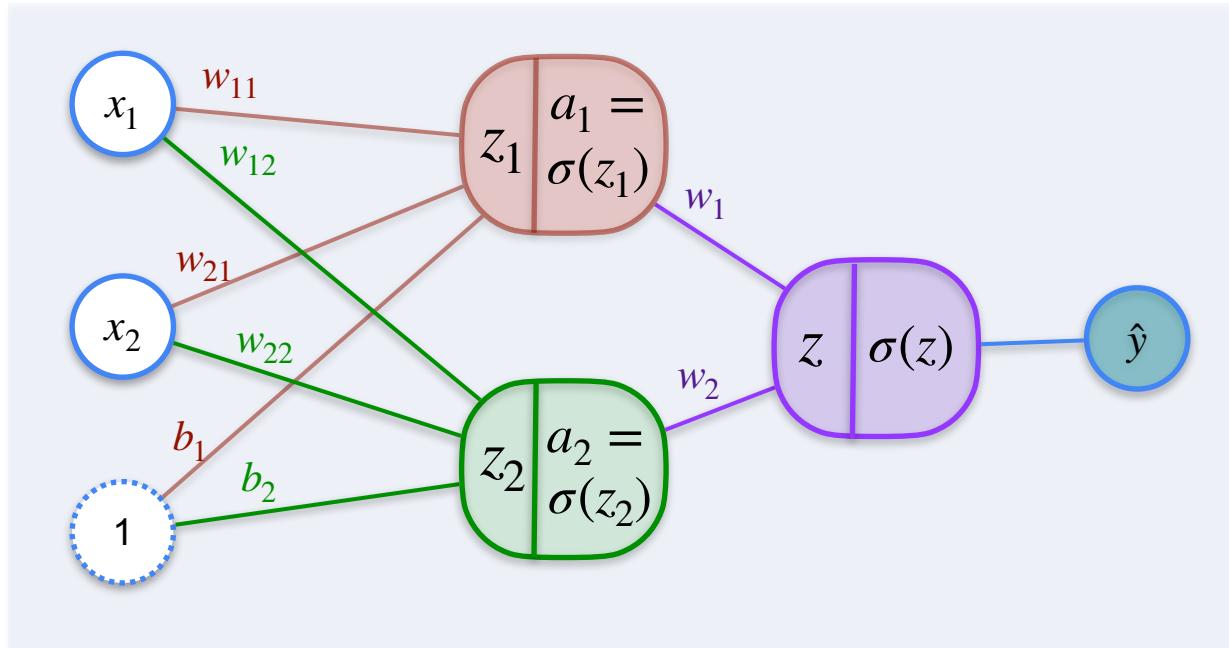
# 2,2,1 Neural Network



# 2,2,1 Neural Network

Neural network of depth 2

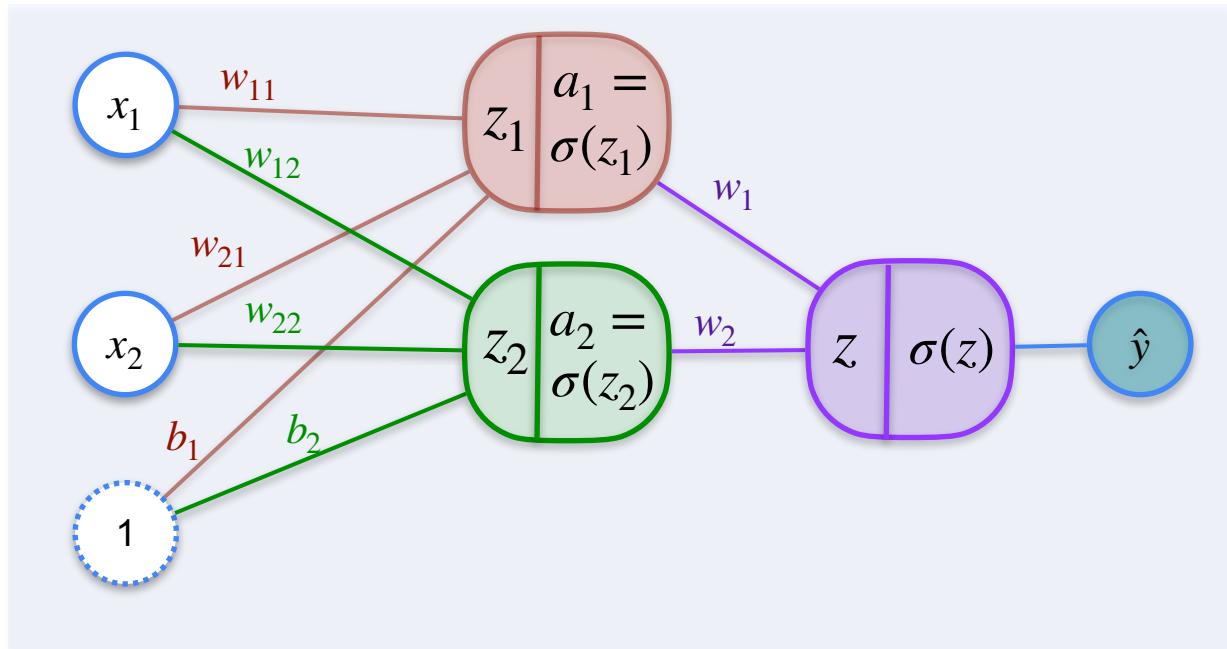
- one input layer
- one hidden layer
- one output layer



# 2,2,1 Neural Network

Neural network of depth 2

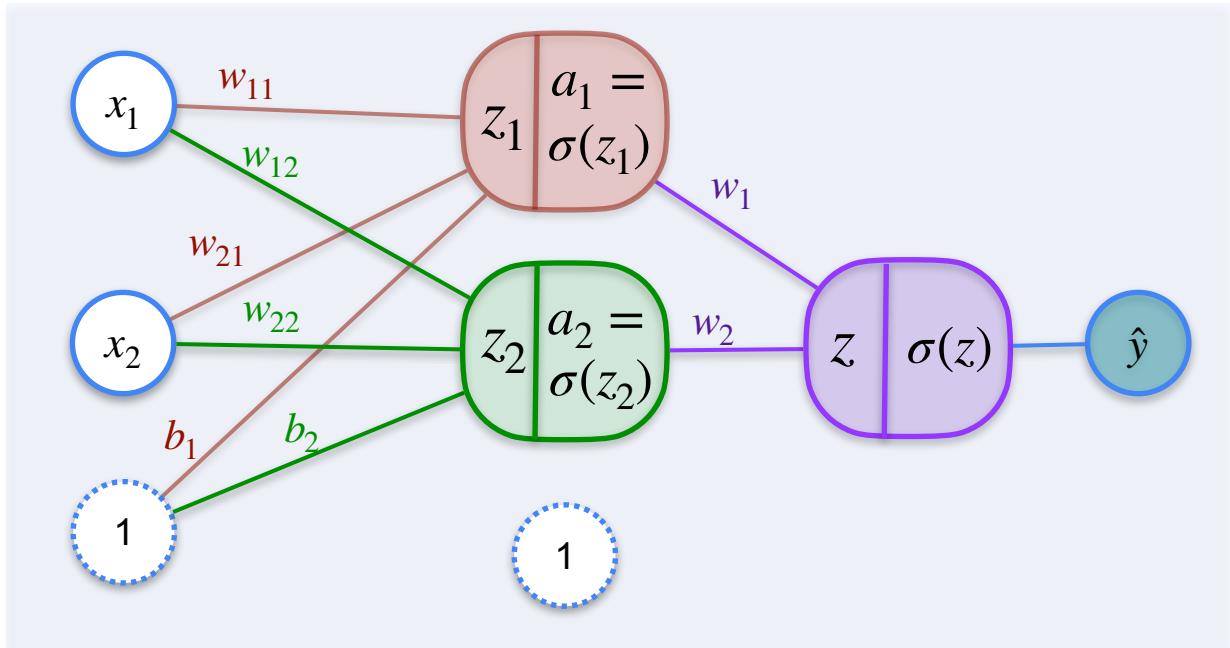
- one input layer
- one hidden layer
- one output layer



# 2,2,1 Neural Network

Neural network of depth 2

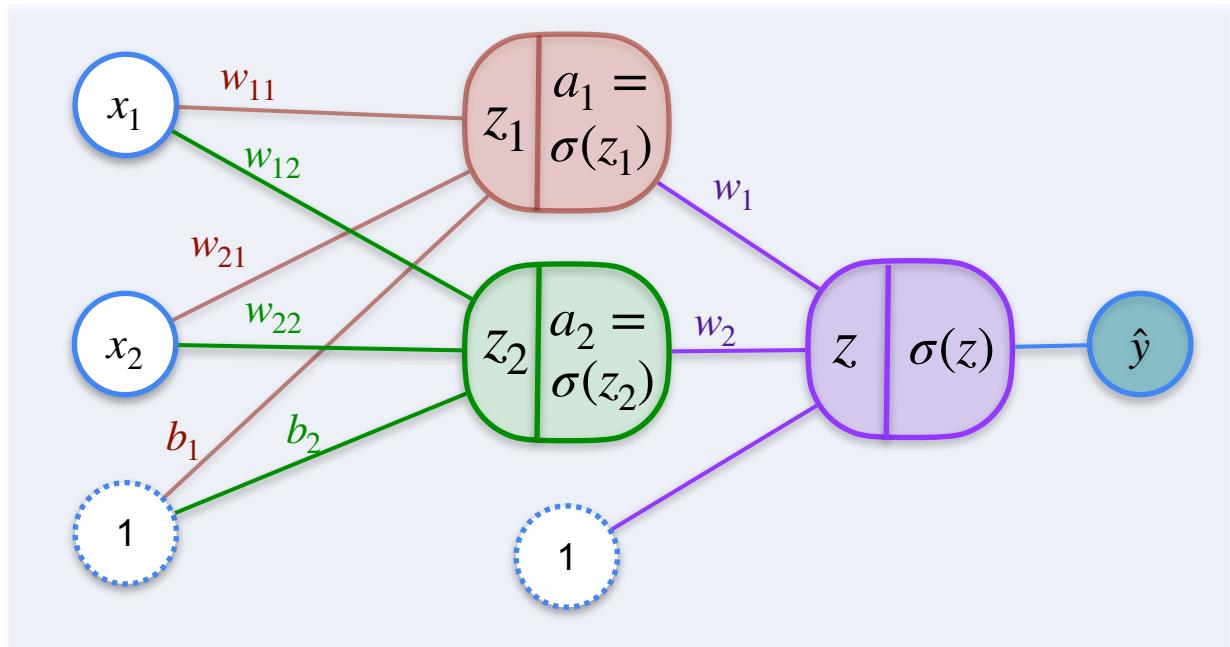
- one input layer
- one hidden layer
- one output layer



# 2,2,1 Neural Network

Neural network of depth 2

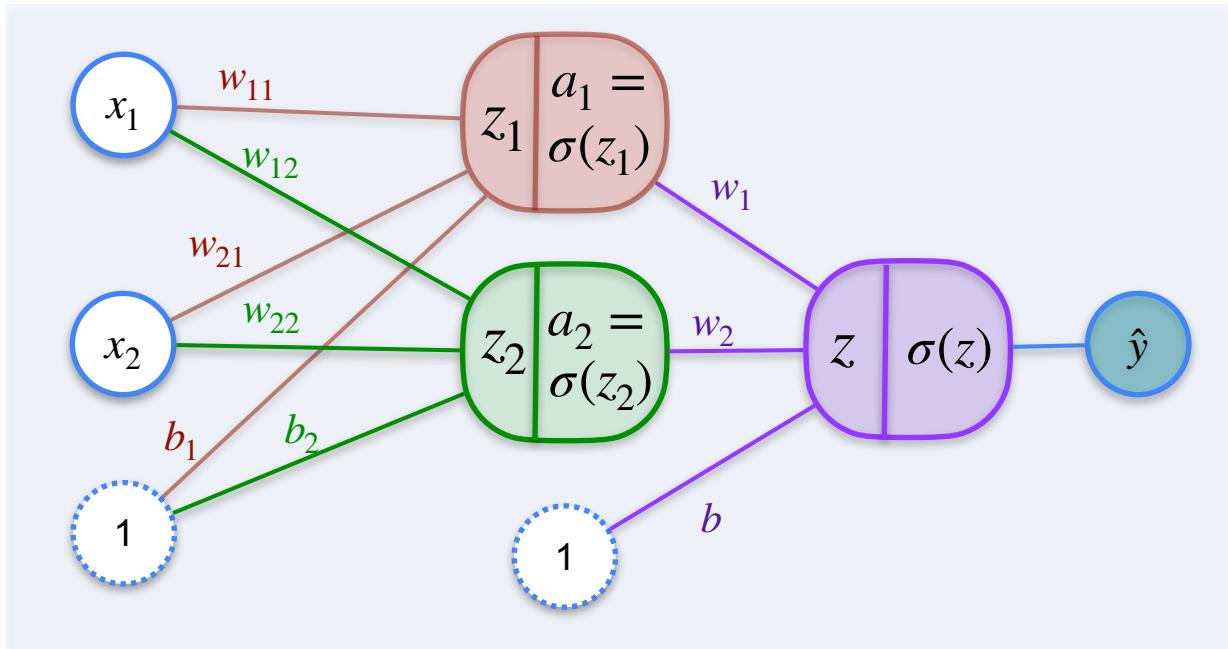
- one input layer
- one hidden layer
- one output layer



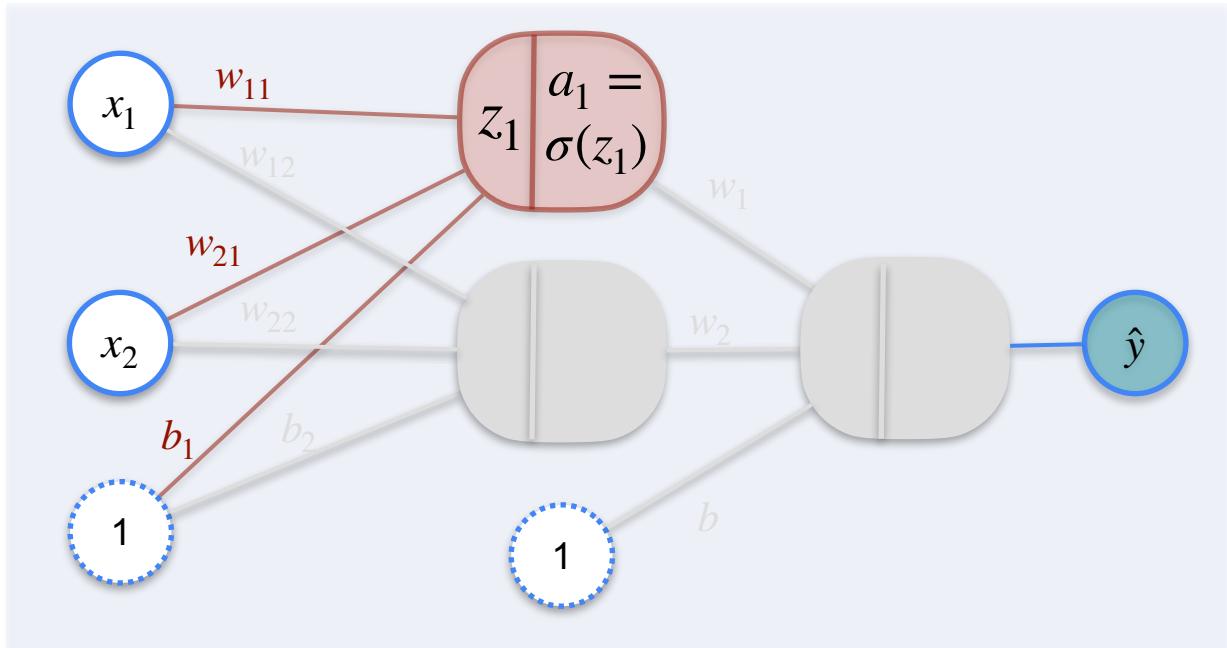
# 2,2,1 Neural Network

Neural network of depth 2

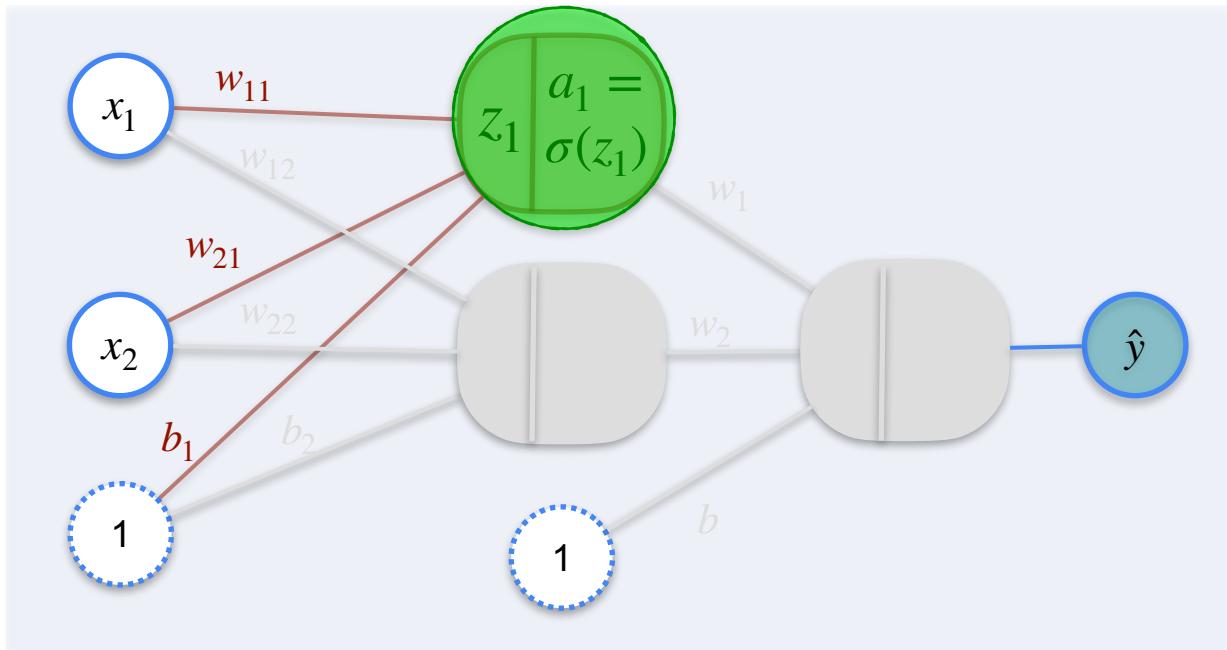
- one input layer
- one hidden layer
- one output layer



# 2,2,1 Neural Network

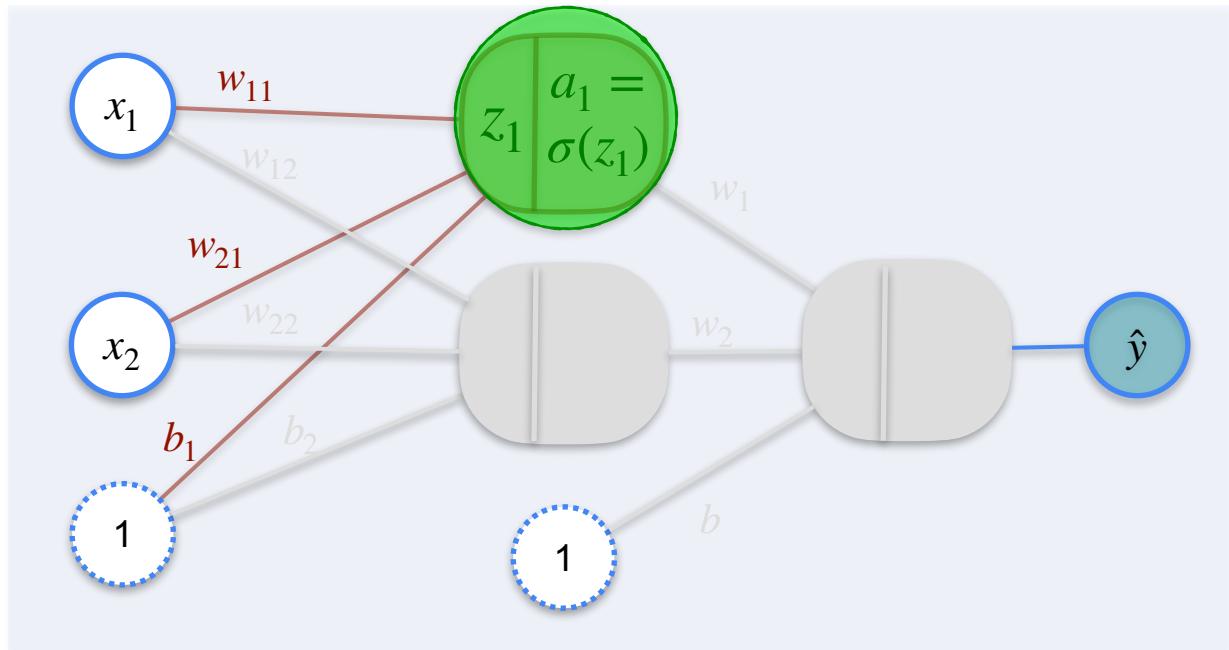


# 2,2,1 Neural Network



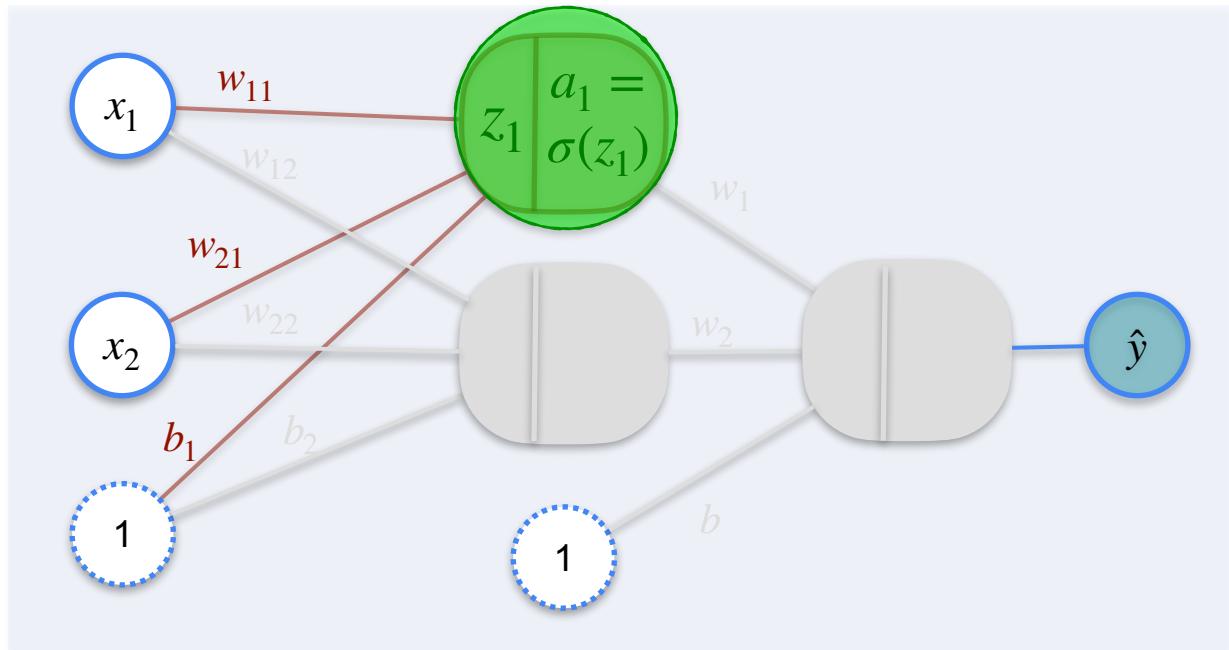
# 2,2,1 Neural Network

$a_1$



# 2,2,1 Neural Network

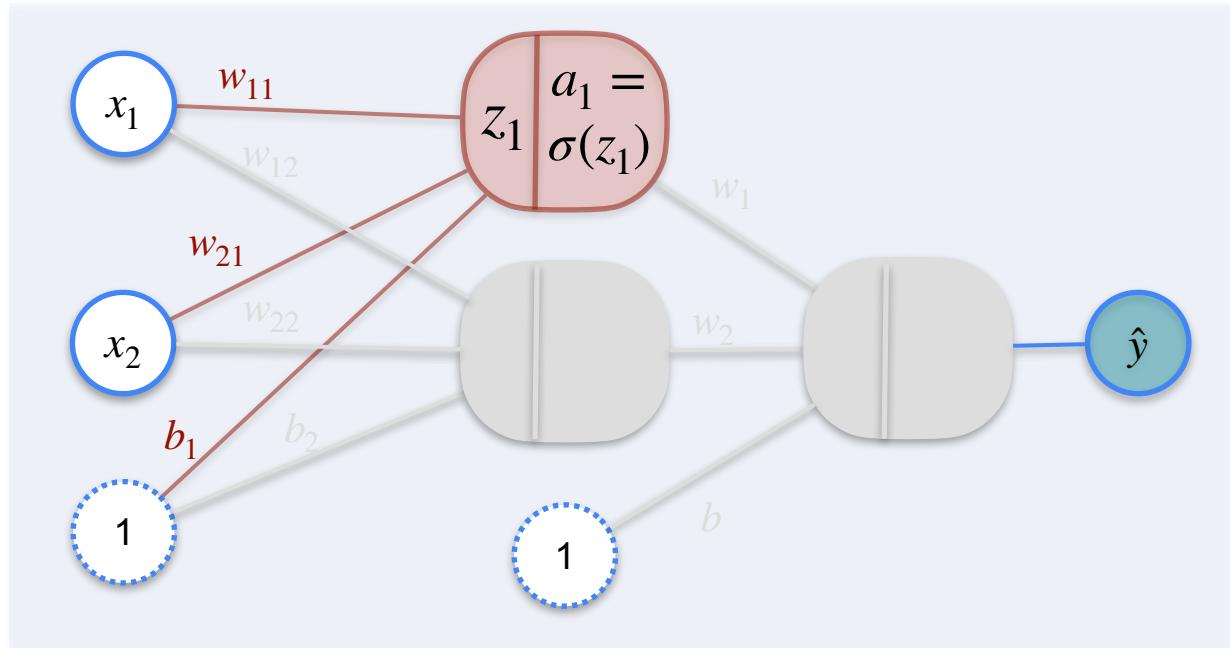
$$a_1 = \sigma(z_1)$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

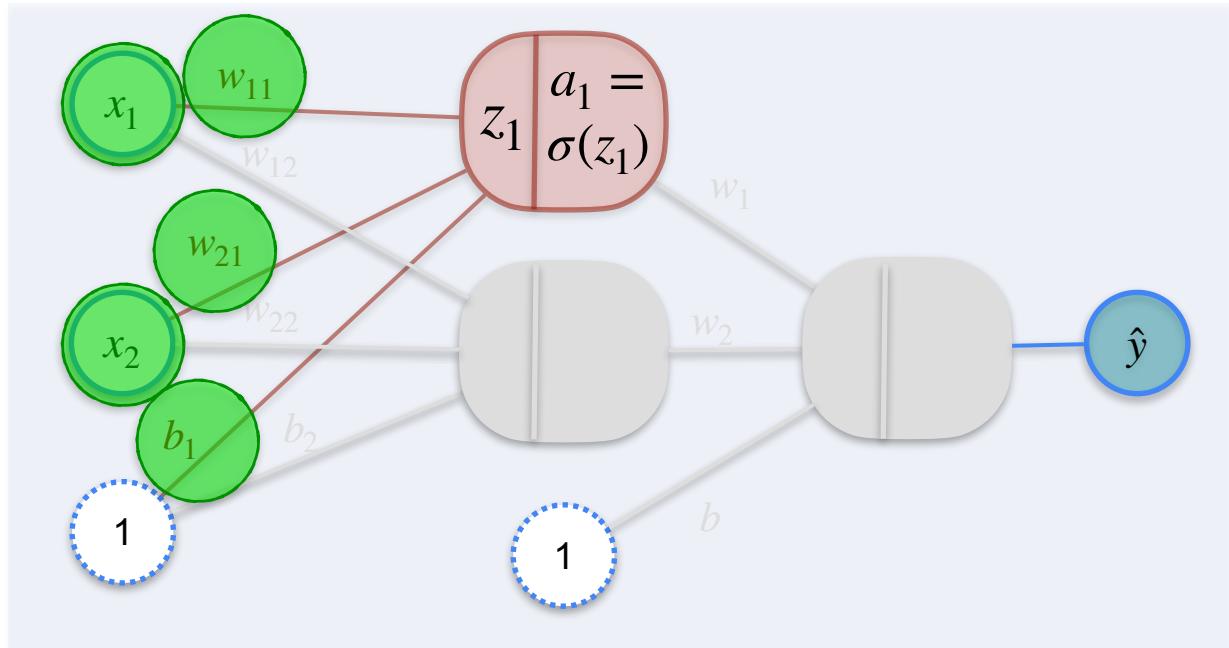
$$z_1$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

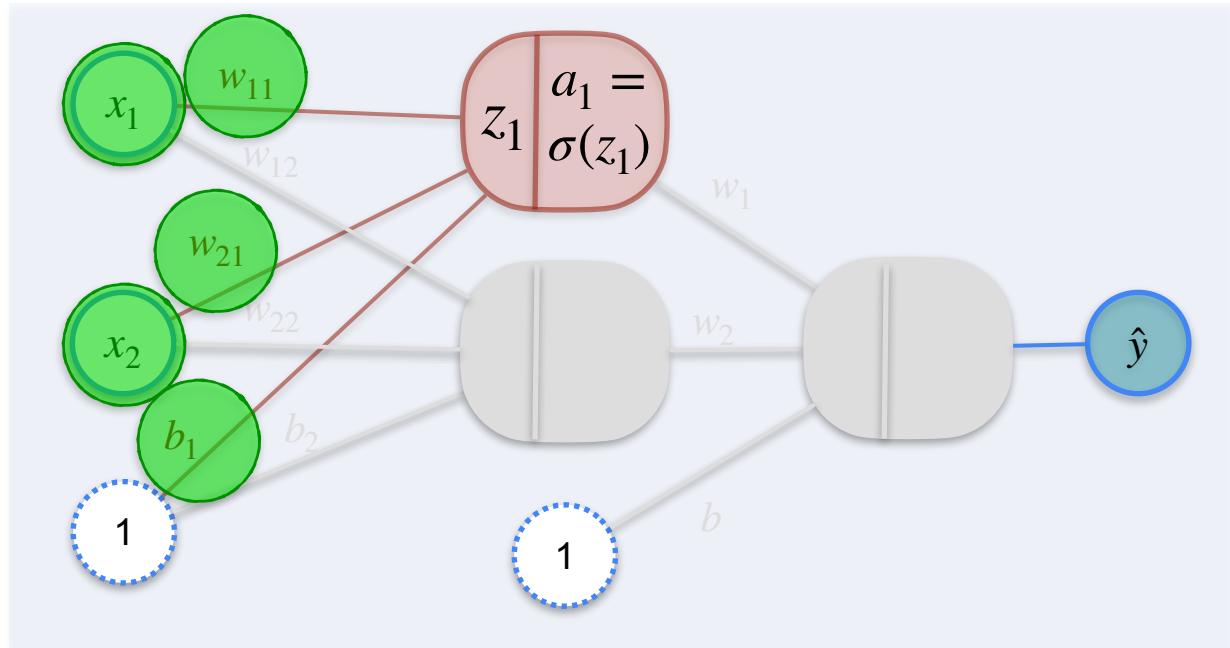
$$z_1$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

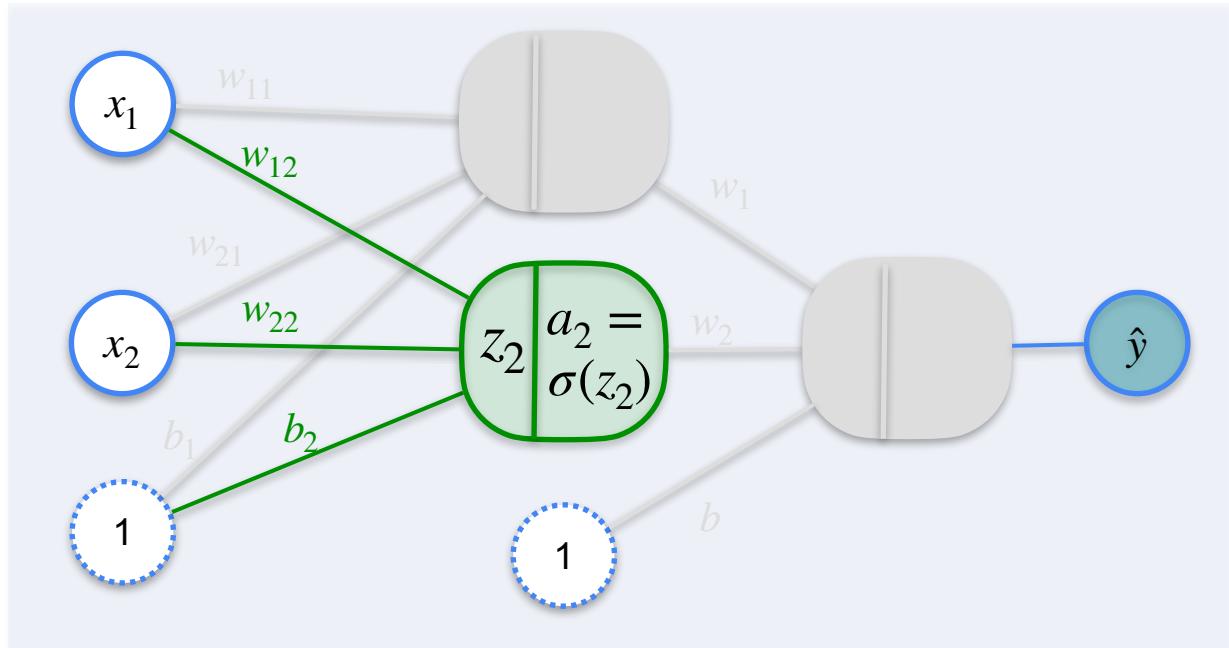
$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

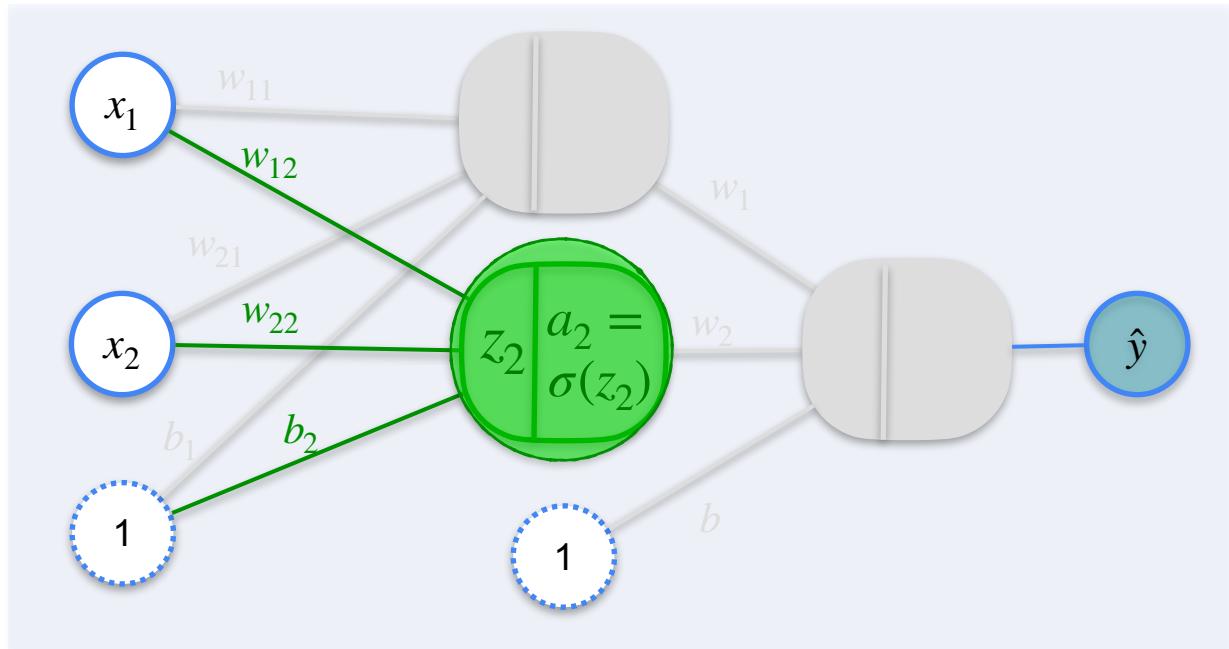
$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

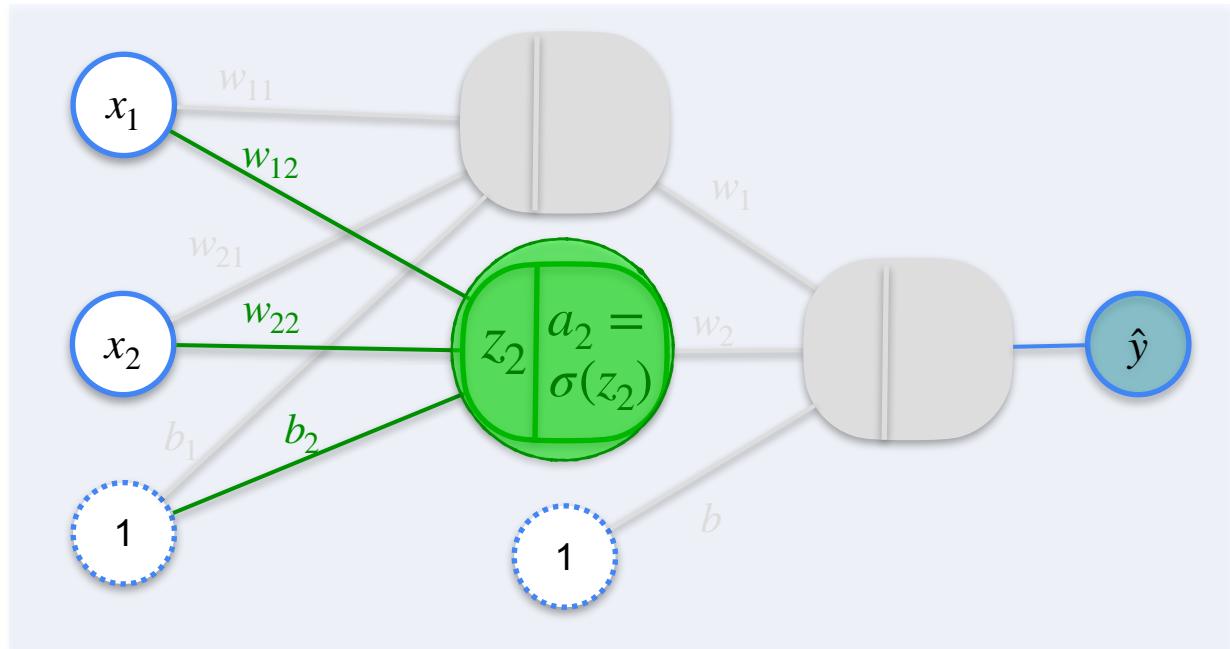


# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2$$

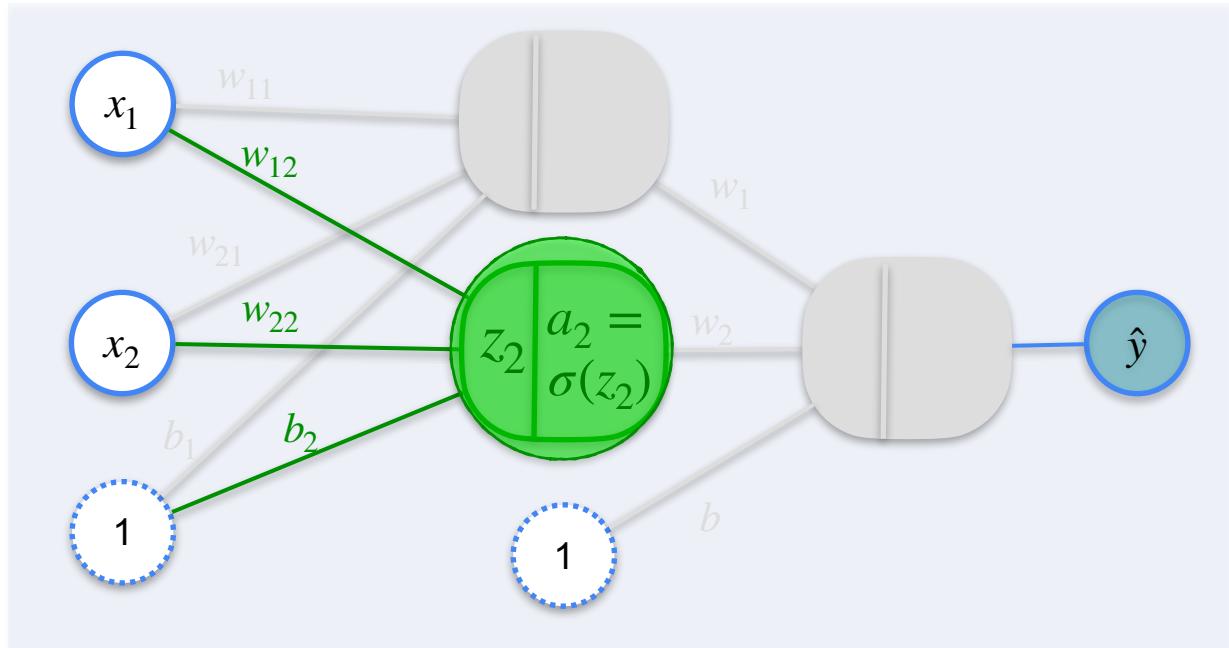


# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$



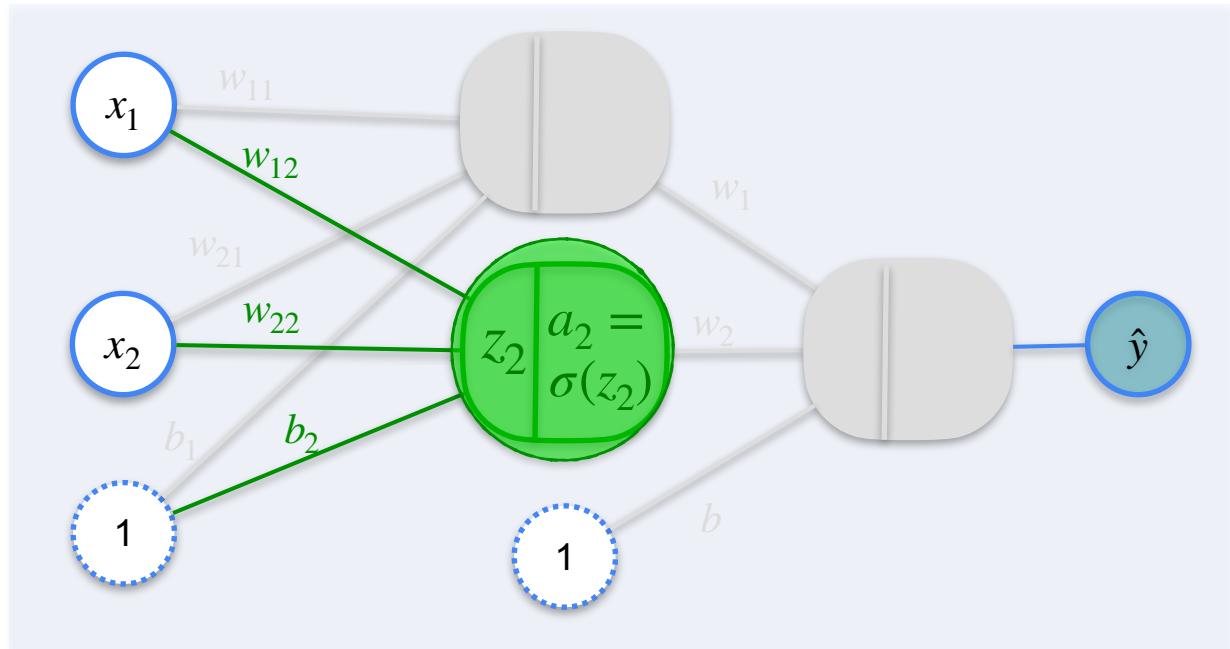
# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2$$



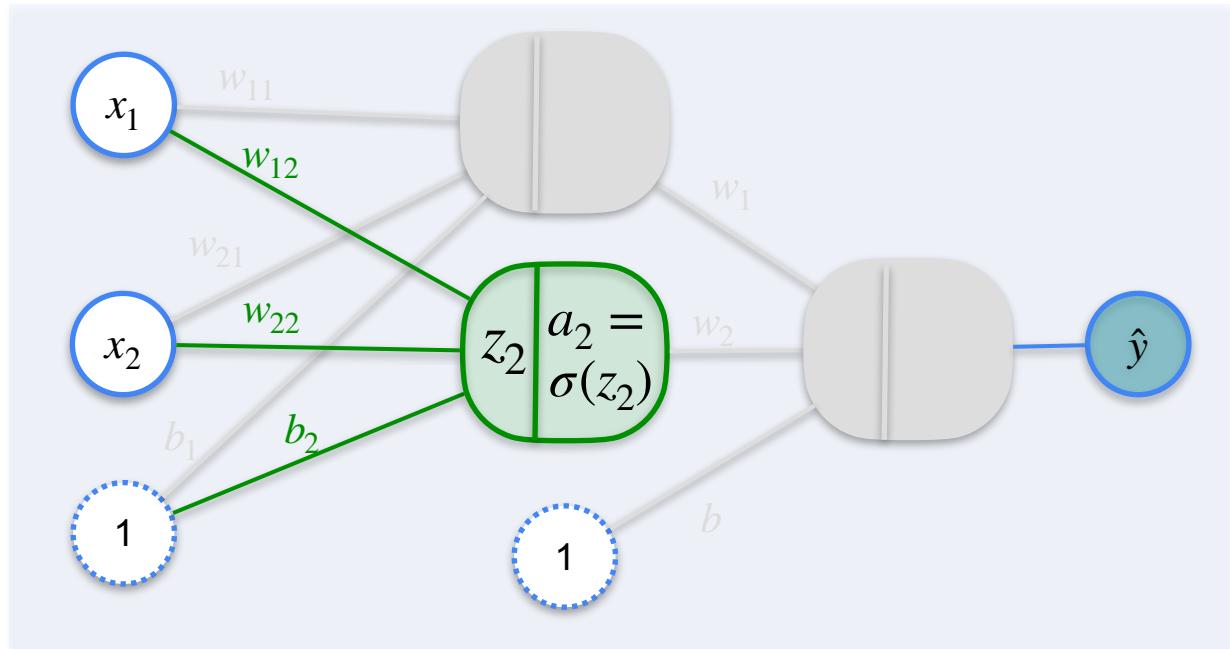
# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2$$



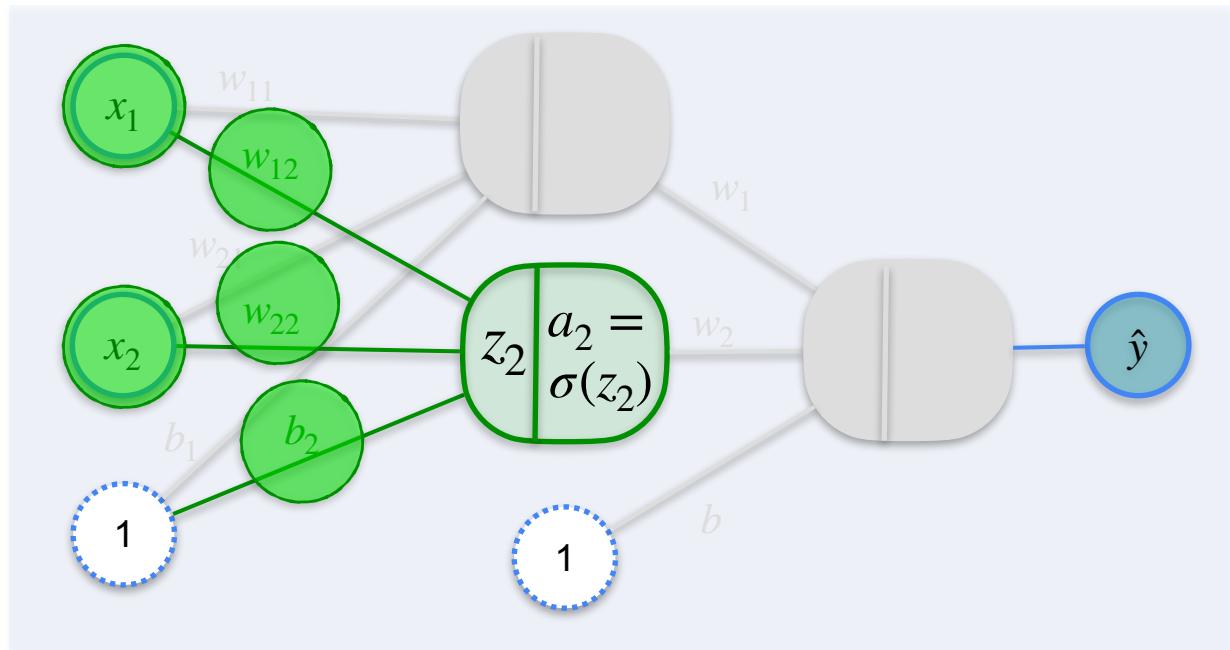
# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2$$



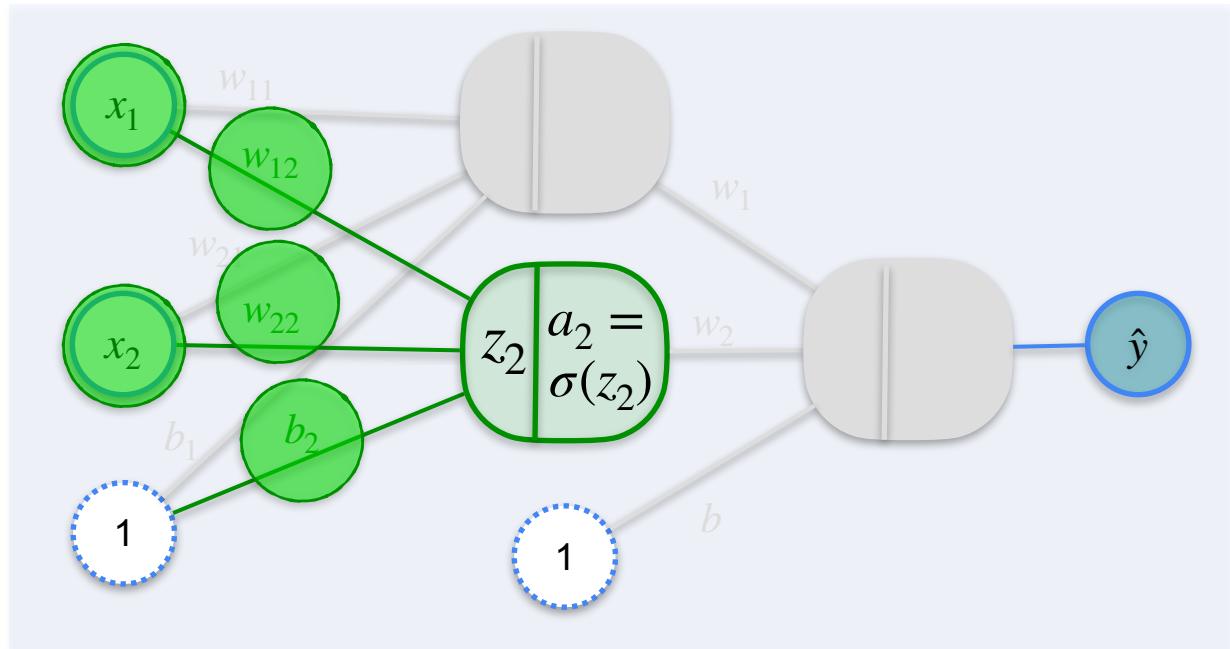
# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$



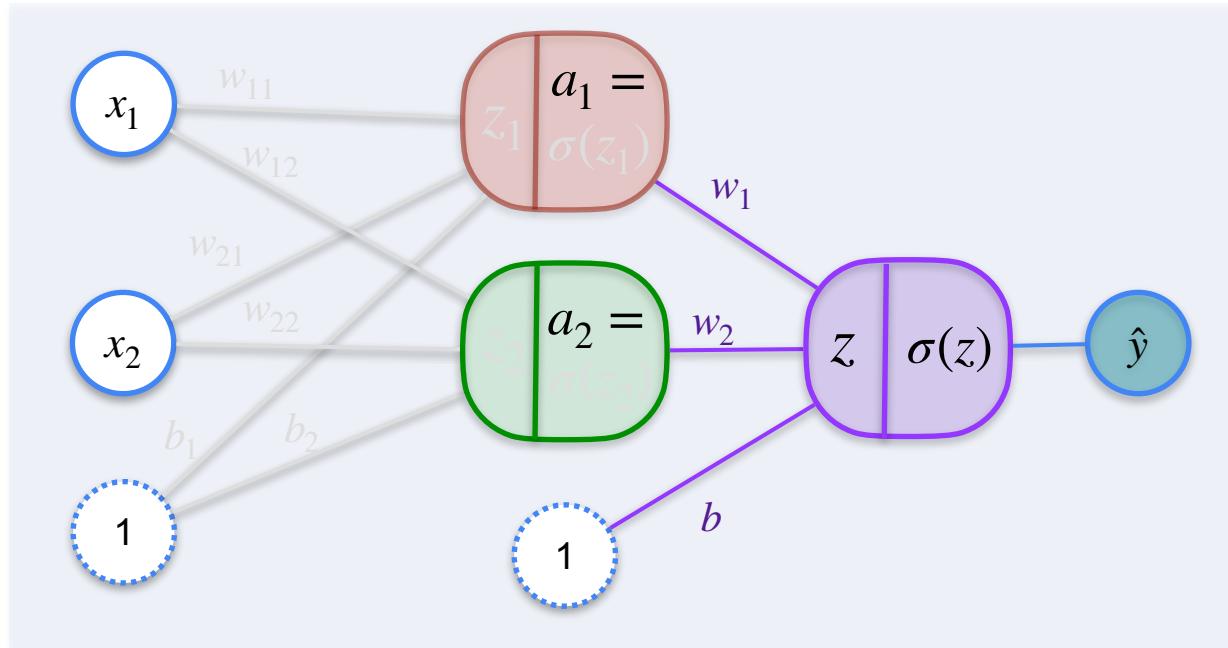
# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$



# 2,2,1 Neural Network

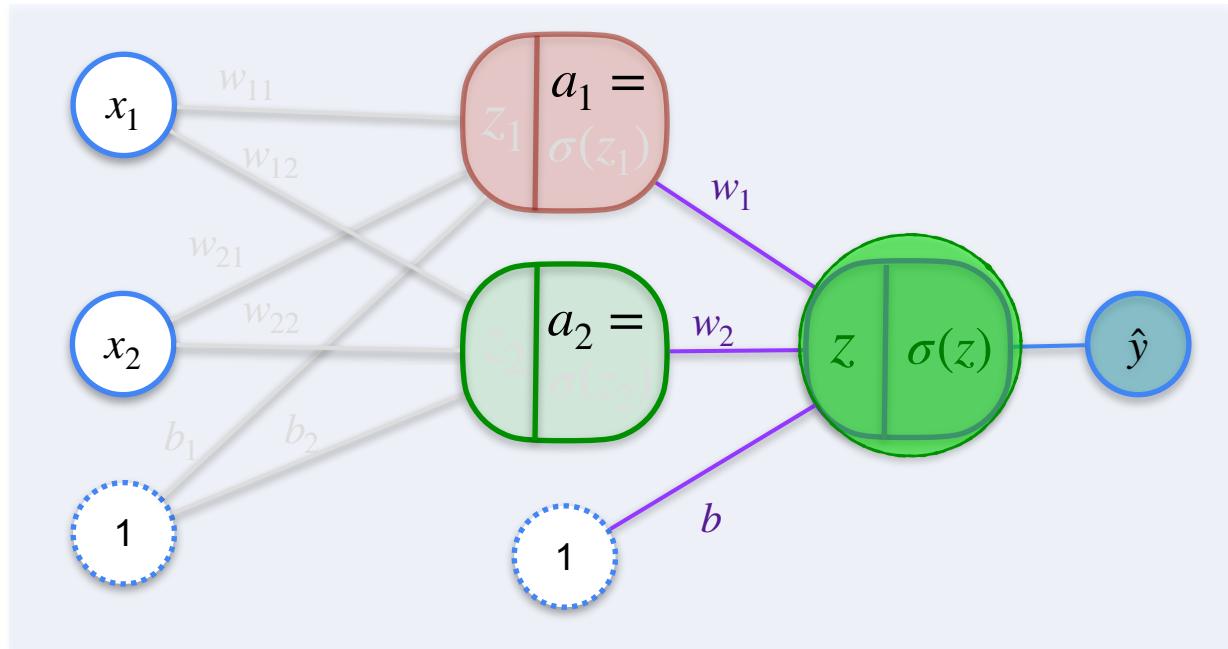
$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$



# 2,2,1 Neural Network

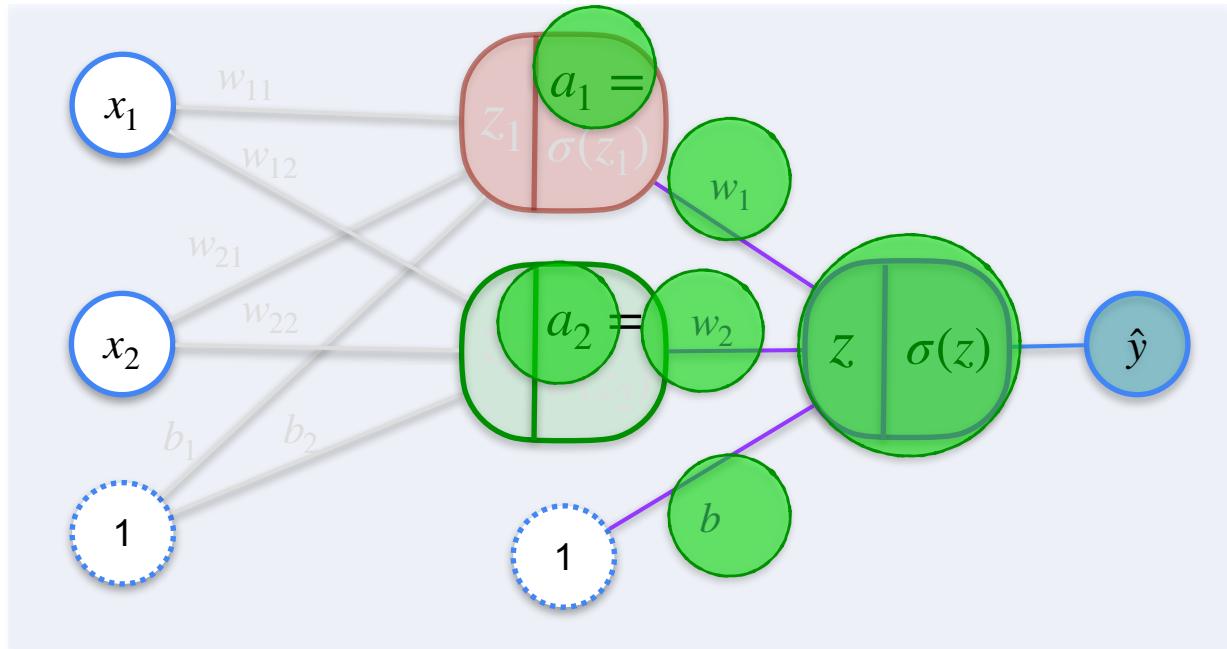
$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

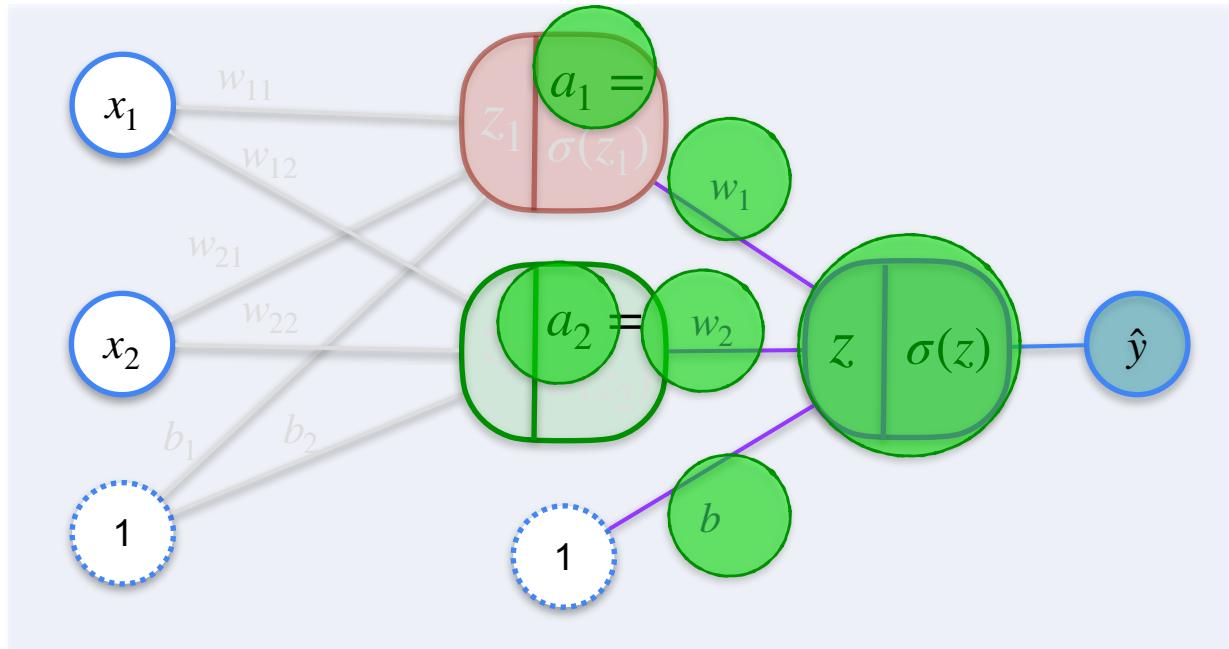
$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

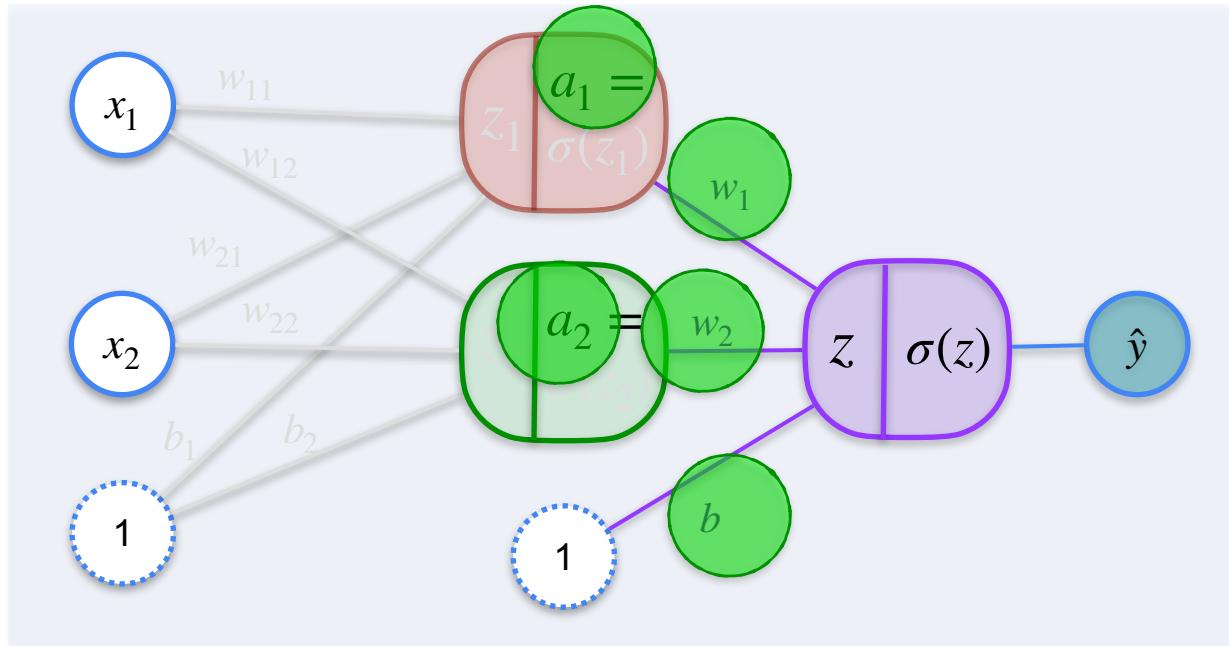
$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

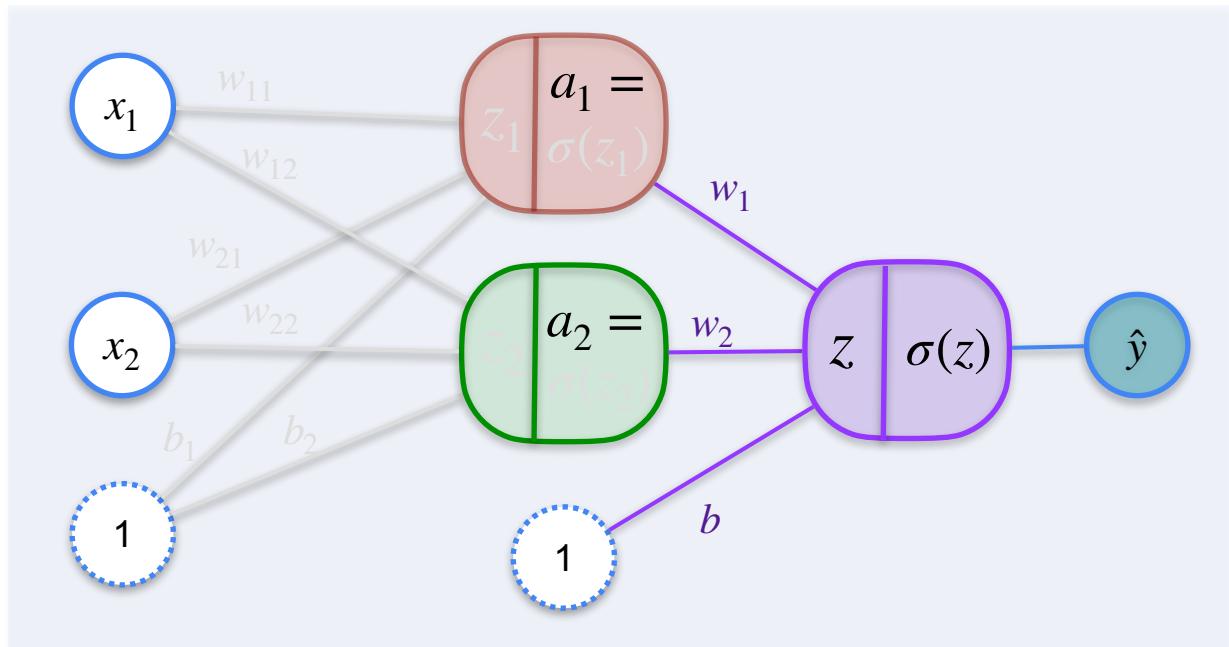
$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

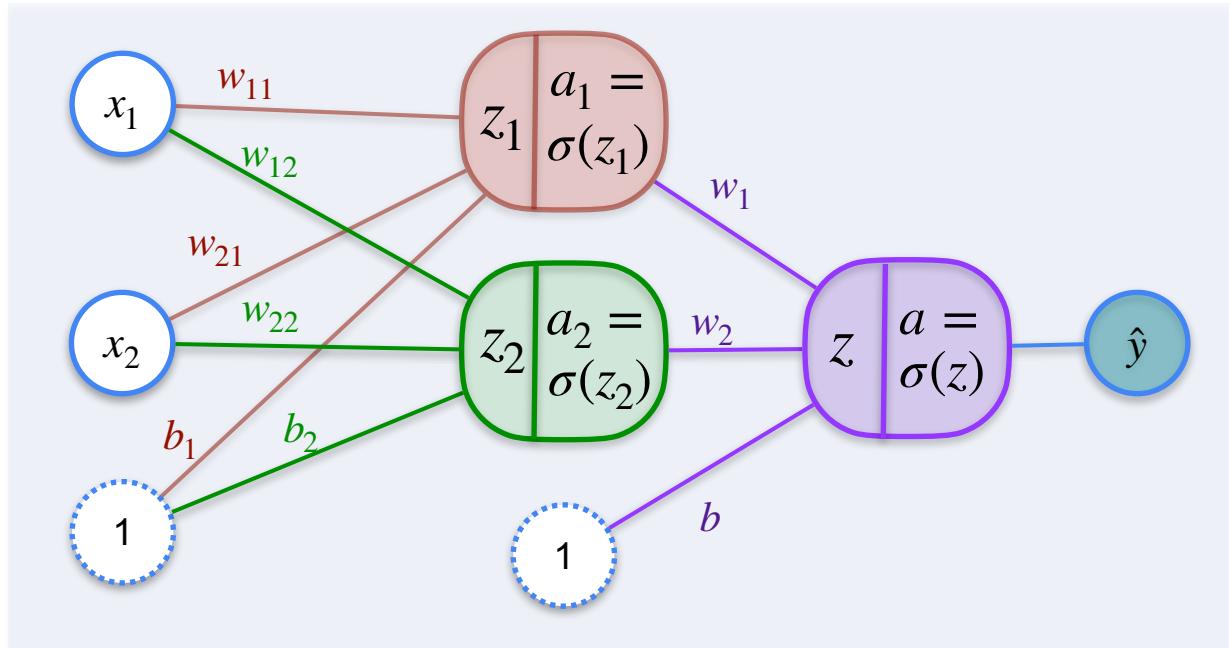
$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$



# 2,2,1 Neural Network

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

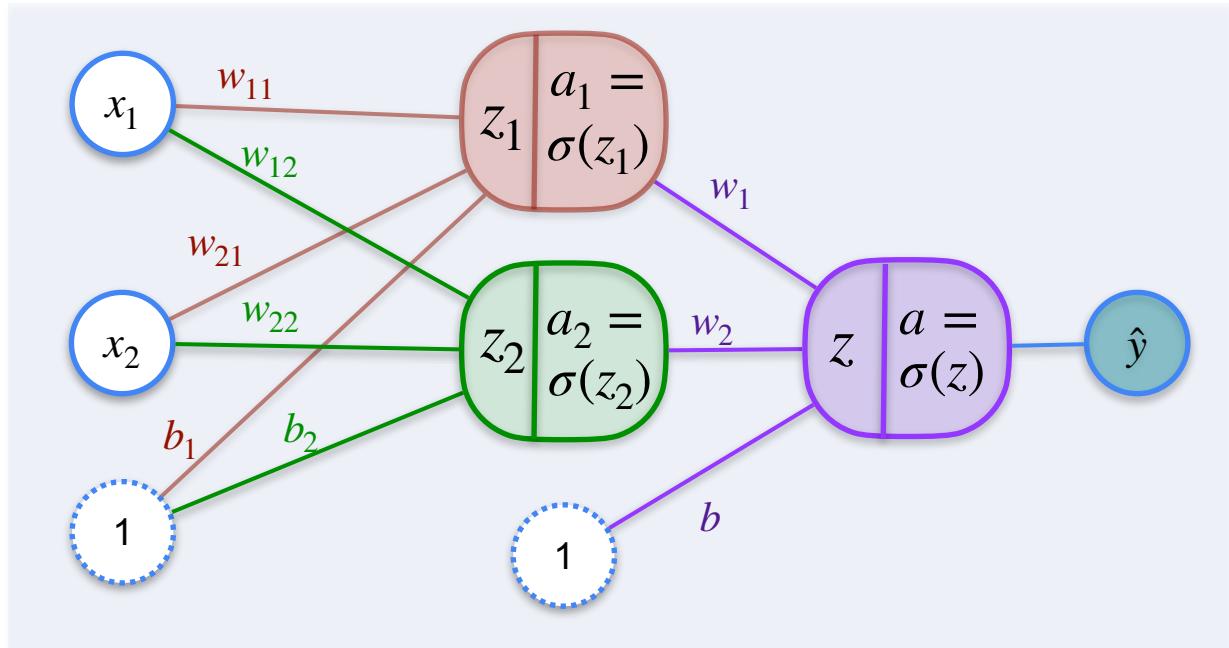
$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$





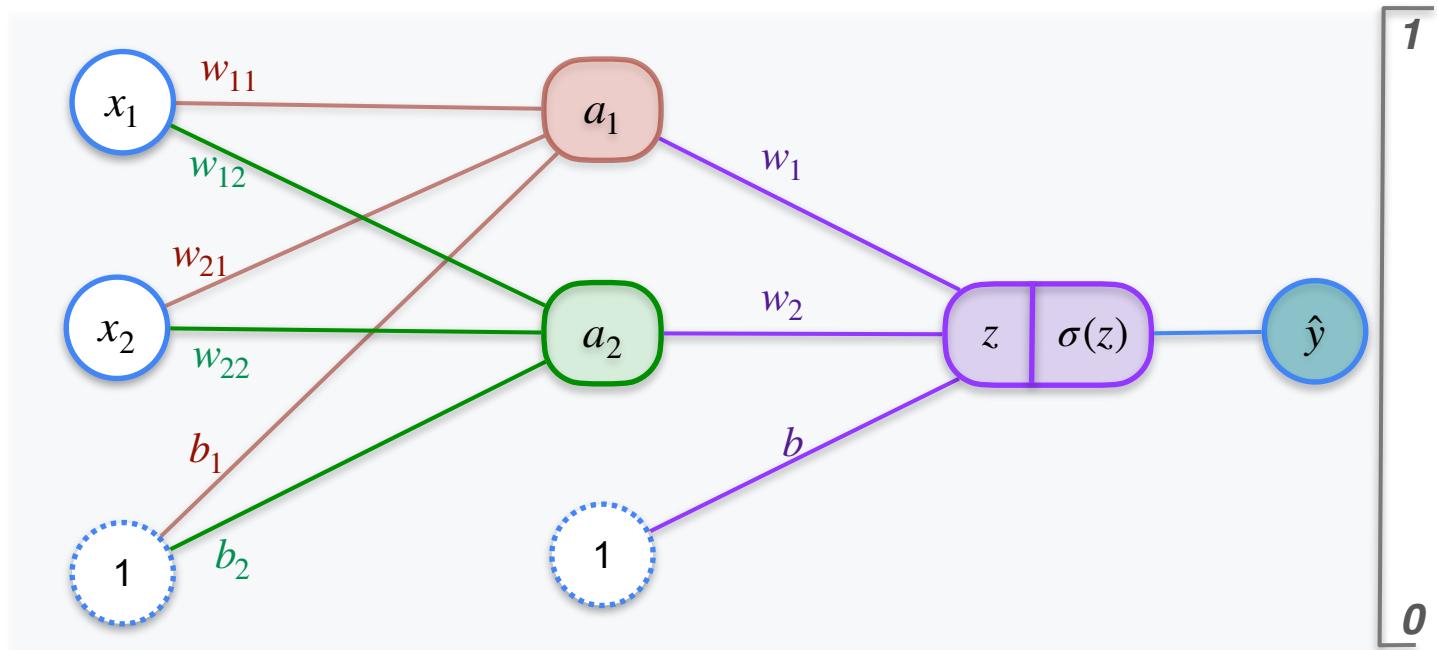
DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

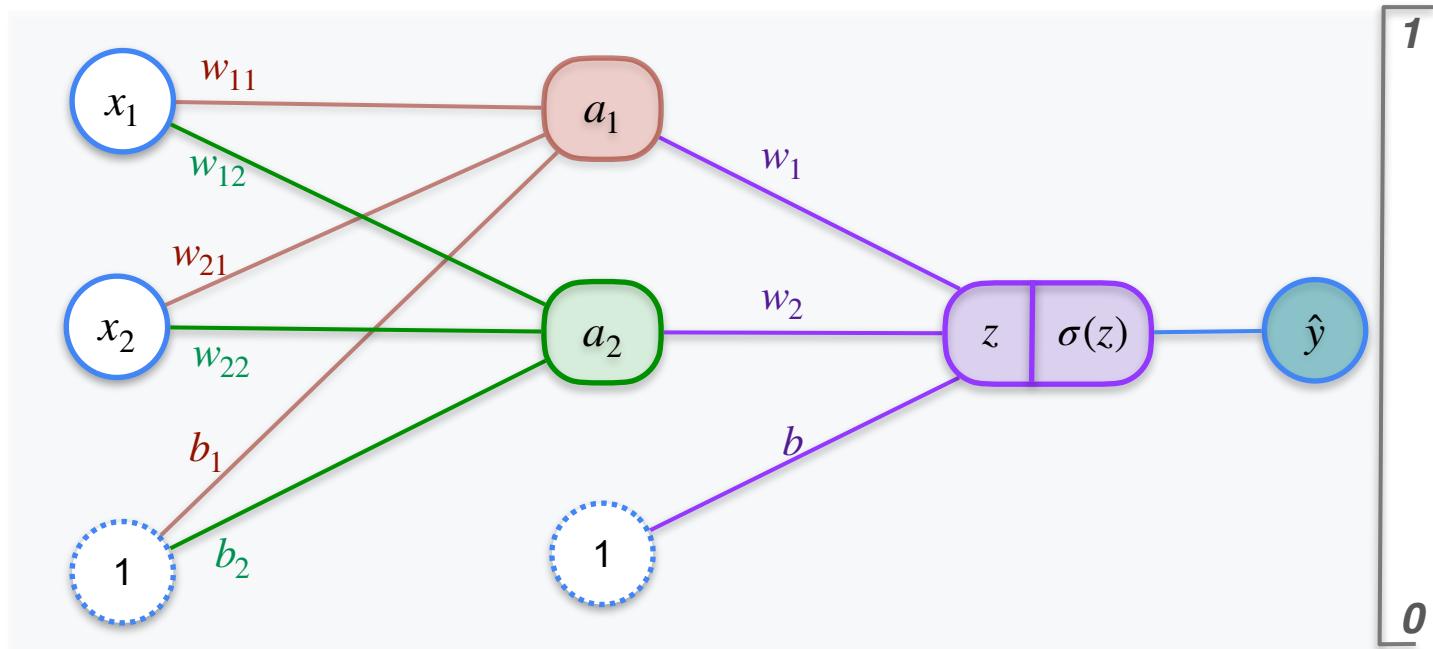
**Classification with a  
Neural Network:  
Minimizing log-loss**

# 2,2,1 Neural Network



# 2,2,1 Neural Network

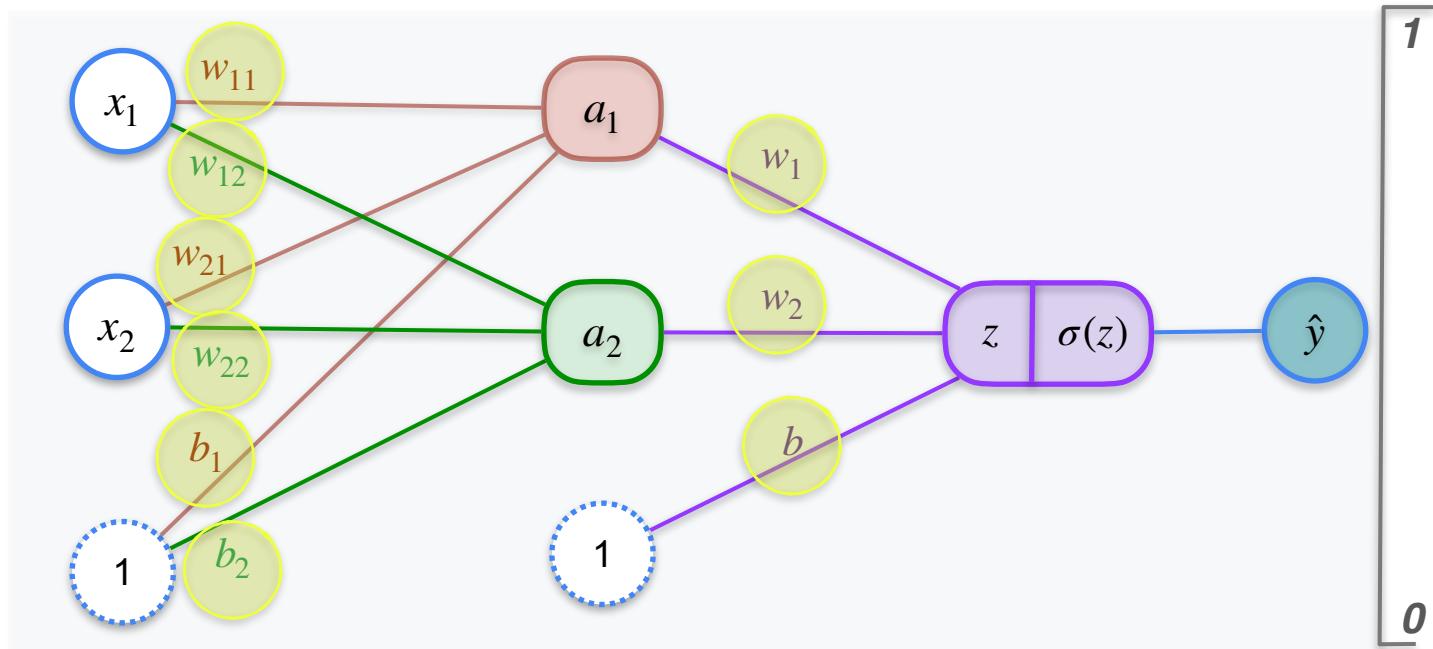
Goal



# 2,2,1 Neural Network

Goal

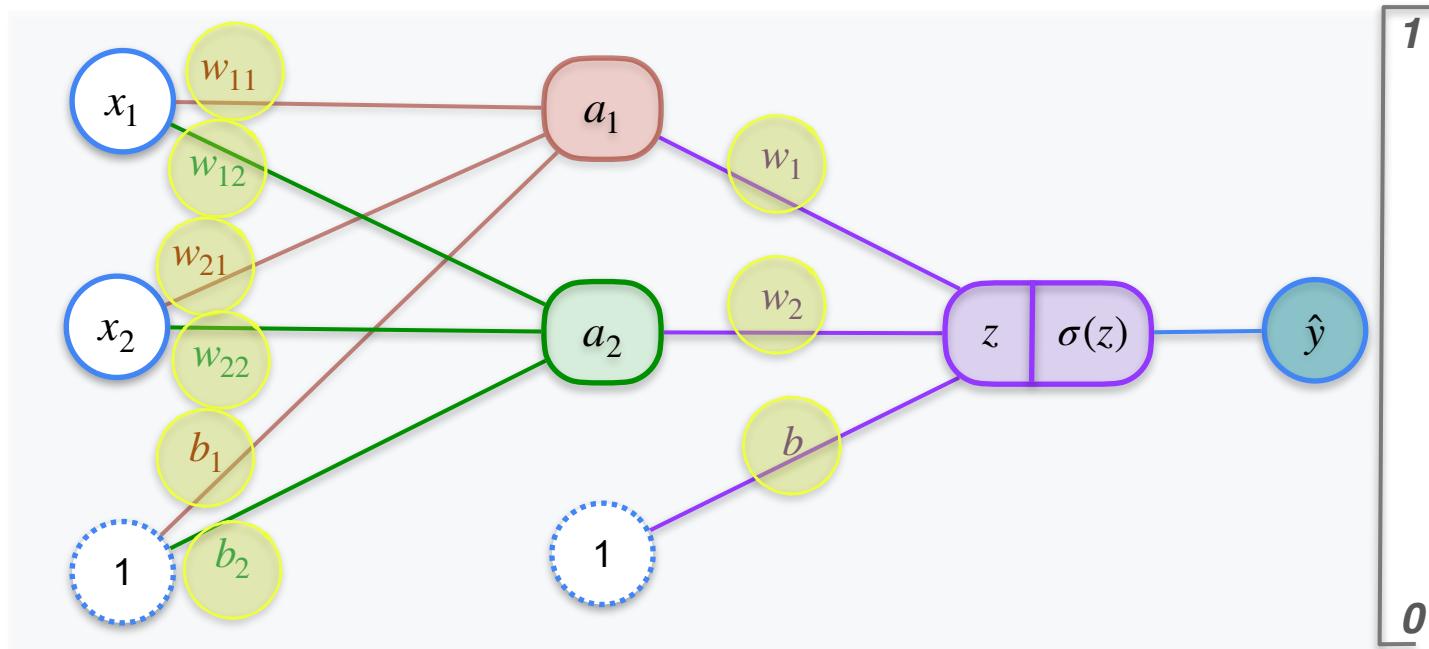
Adjust each of the highlighted weights and biases



# 2,2,1 Neural Network

Goal

Adjust each of the highlighted weights and biases

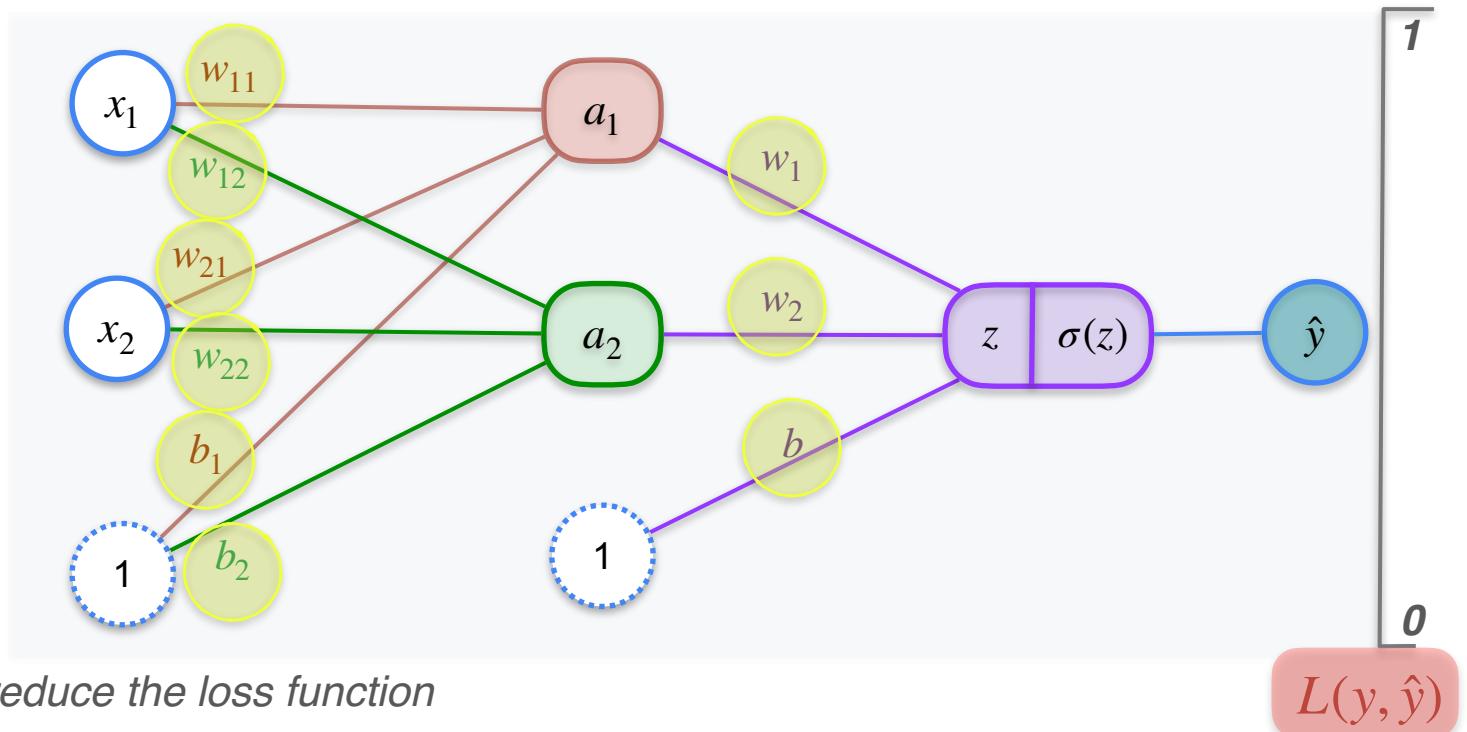


*To reduce the loss function*

# 2,2,1 Neural Network

Goal

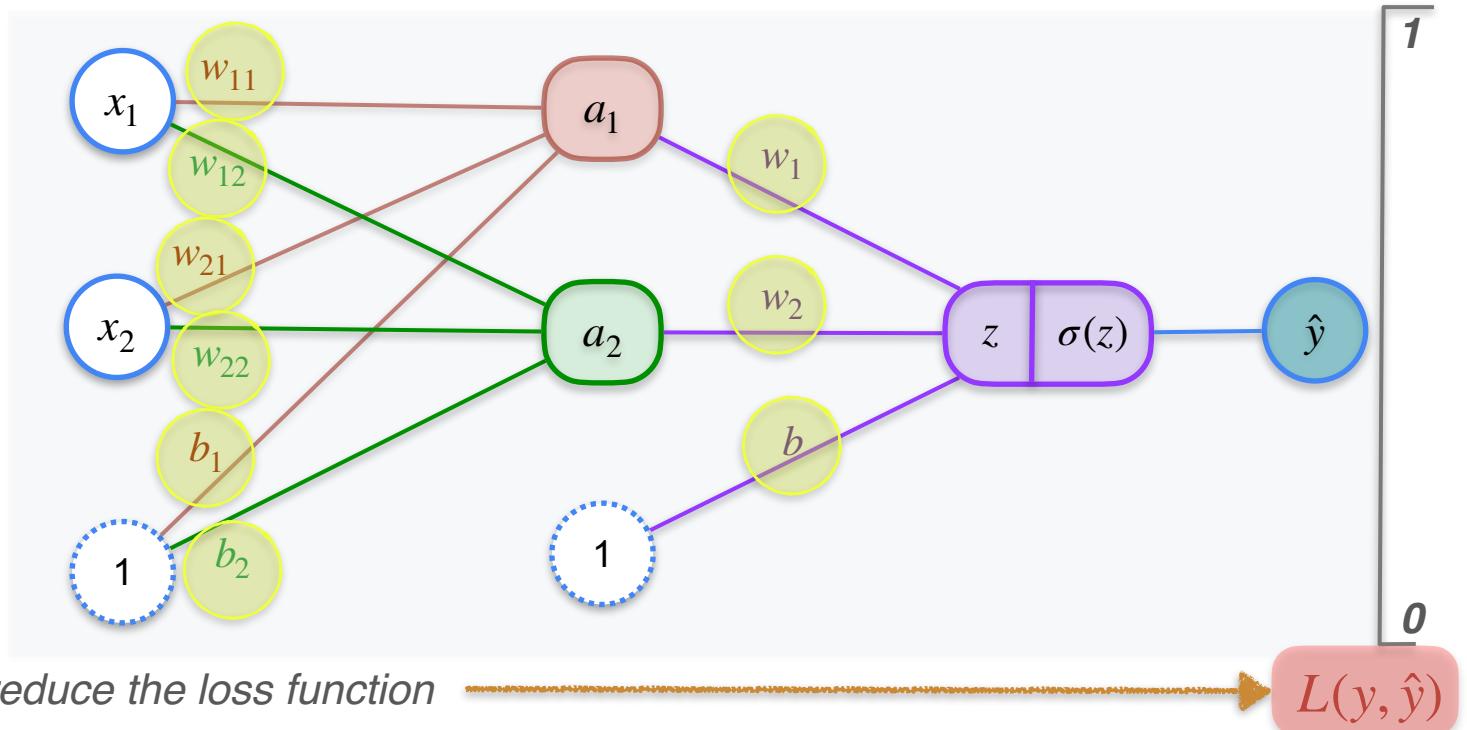
Adjust each of the highlighted weights and biases



# 2,2,1 Neural Network

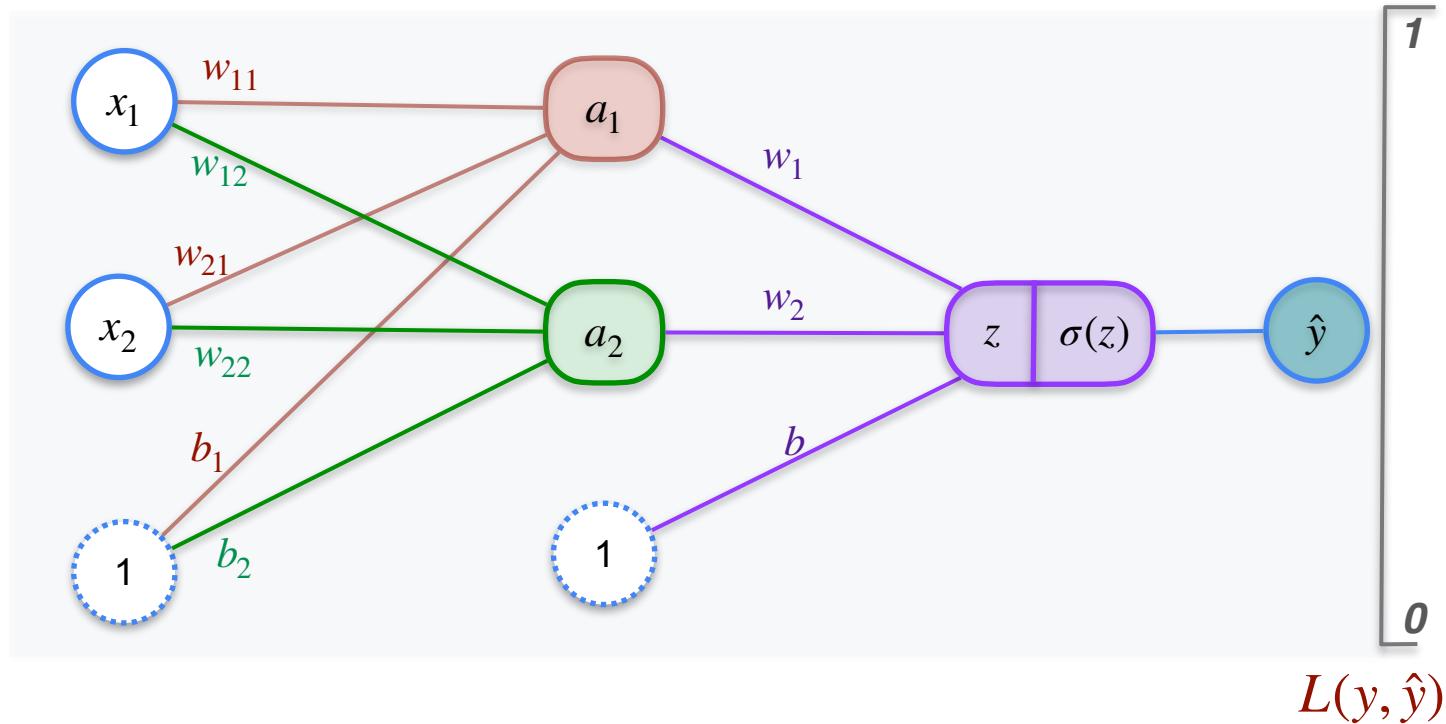
Goal

Adjust each of the highlighted weights and biases

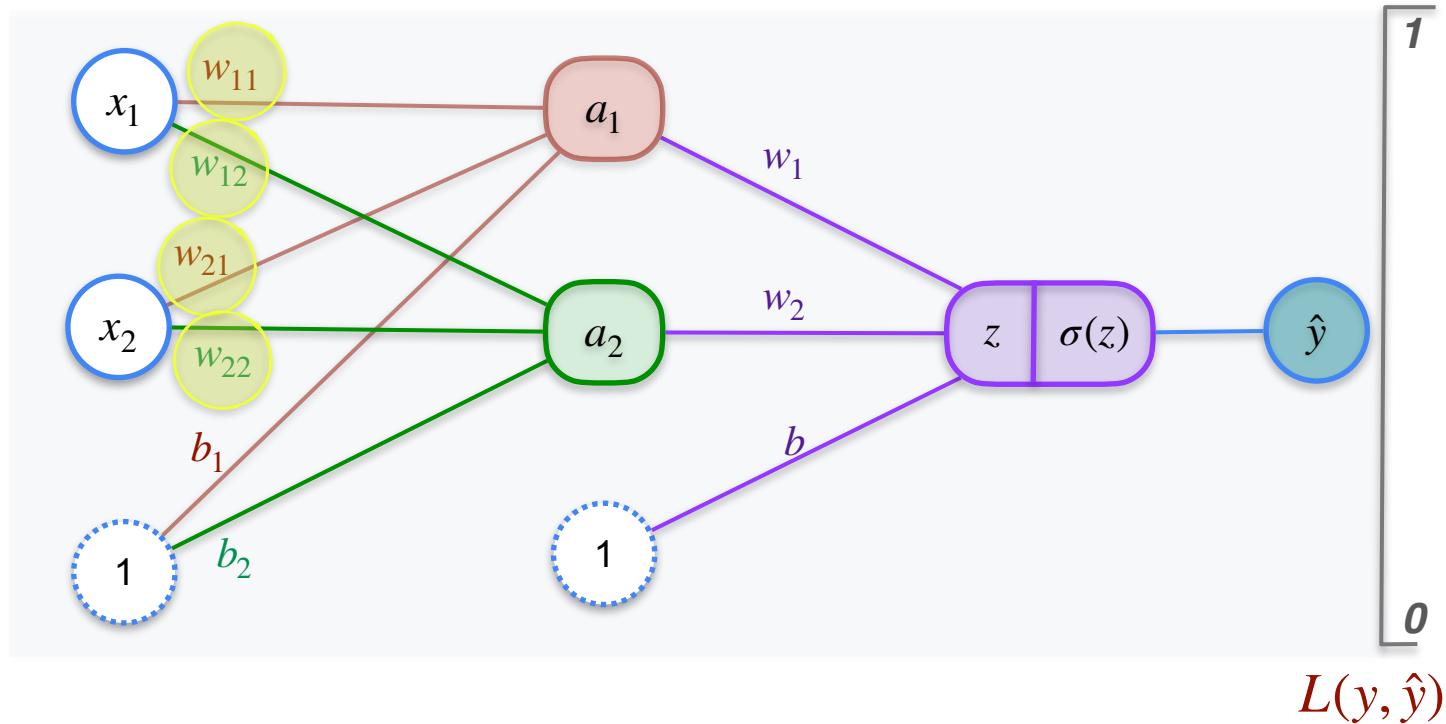


To reduce the loss function

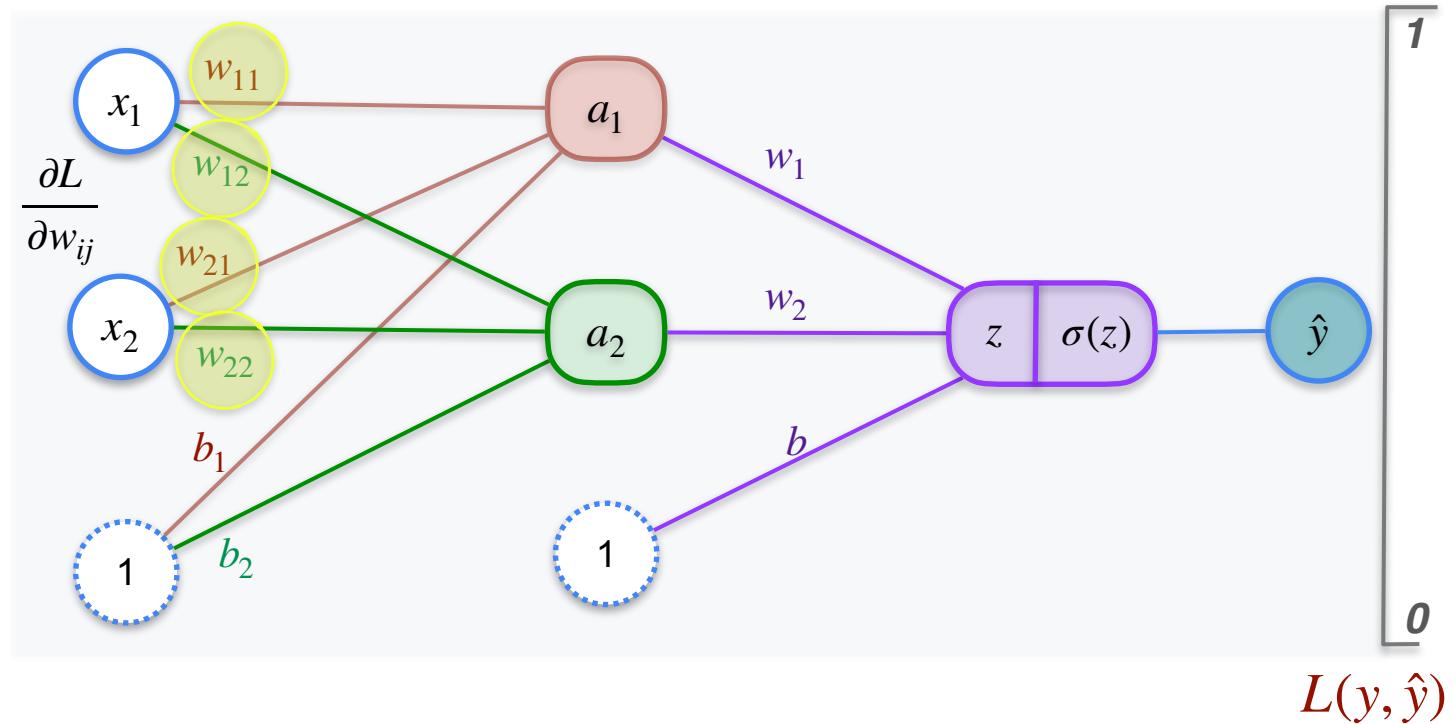
# 2,2,1 Neural Network



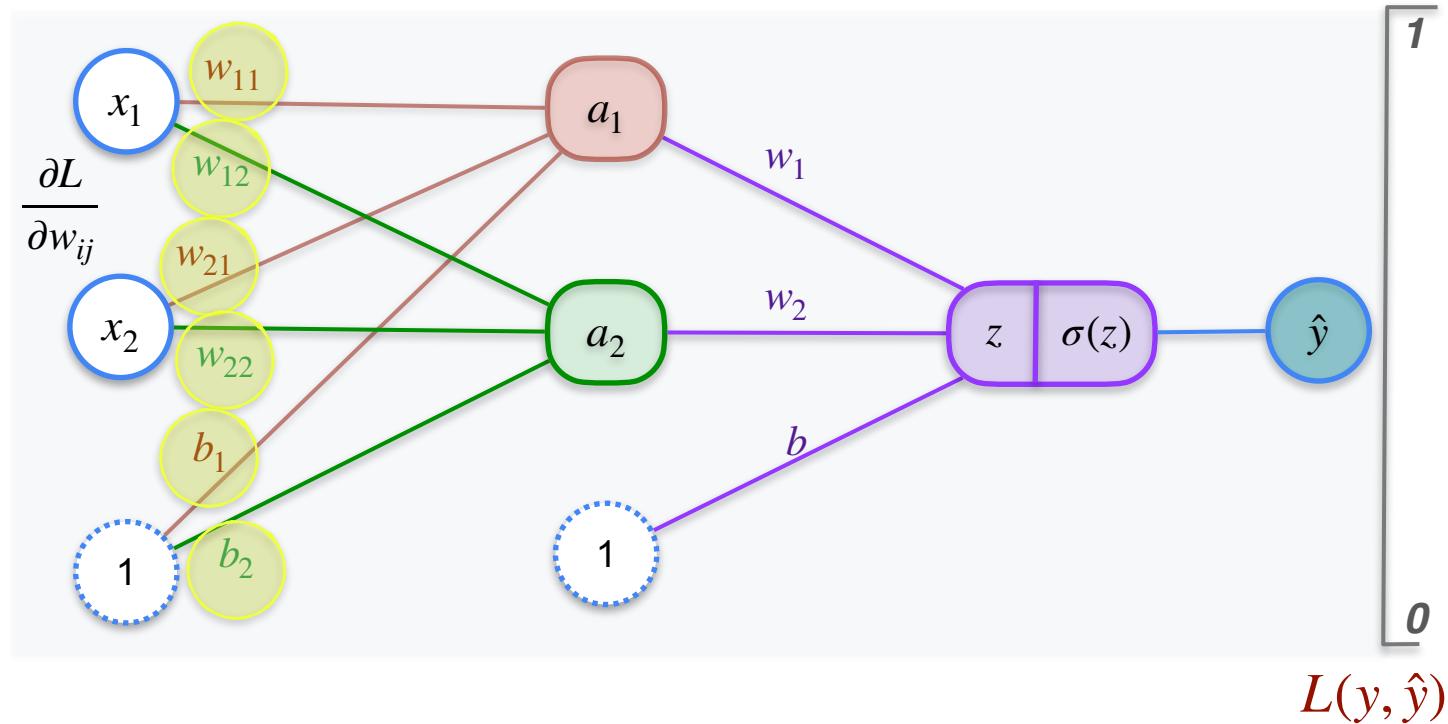
# 2,2,1 Neural Network



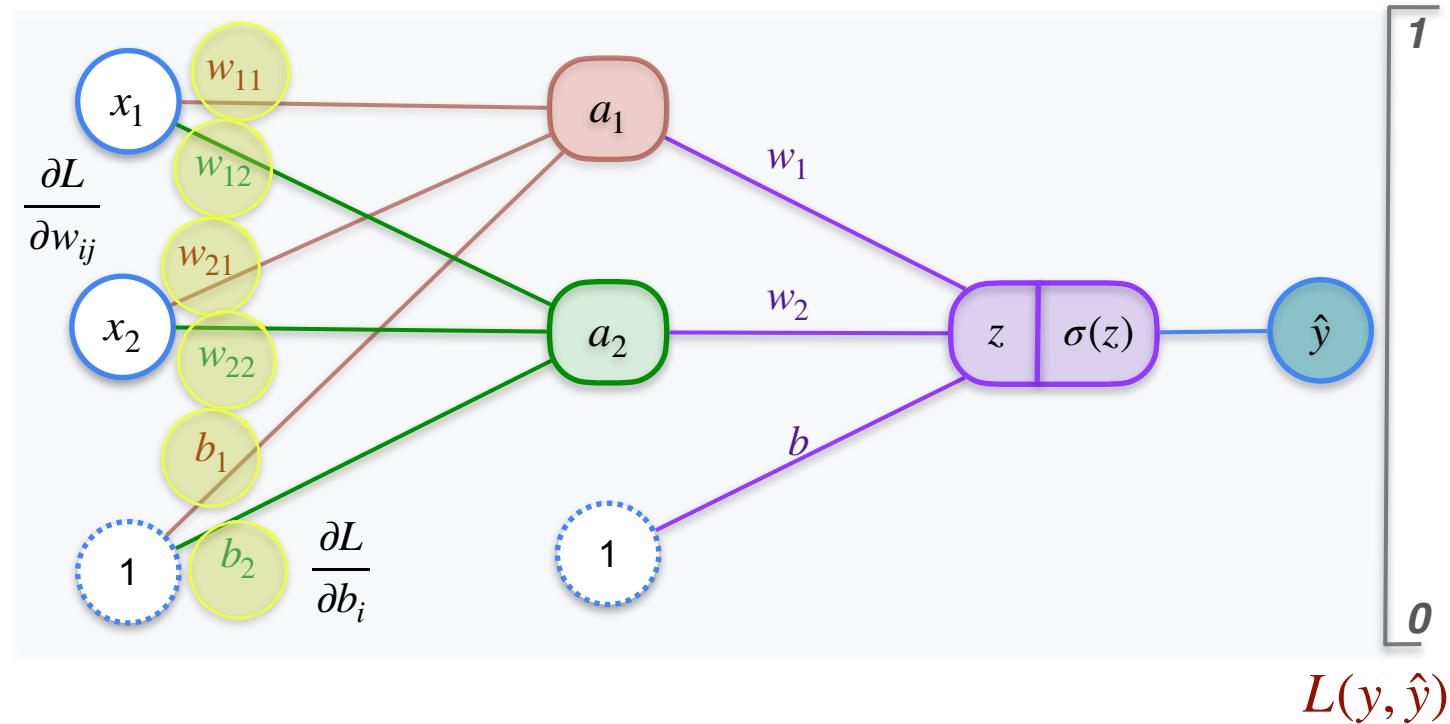
# 2,2,1 Neural Network



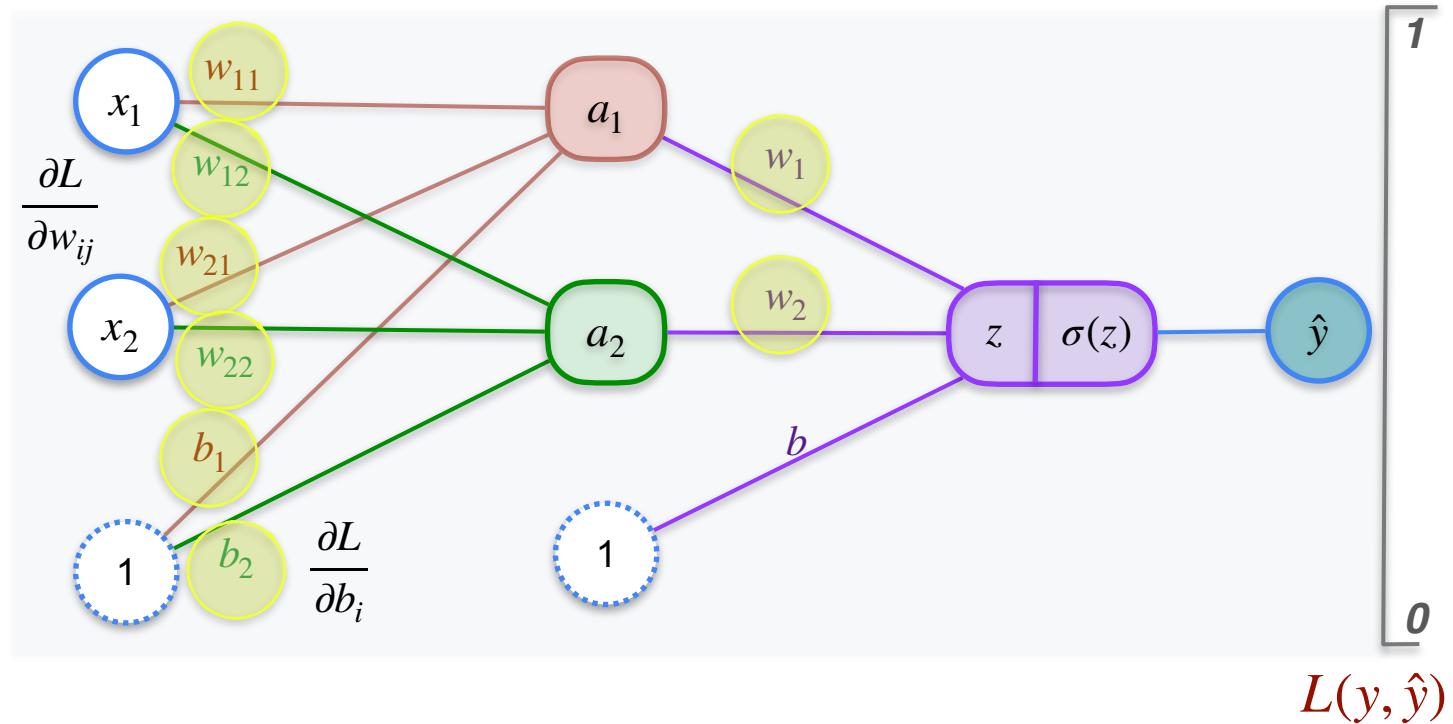
# 2,2,1 Neural Network



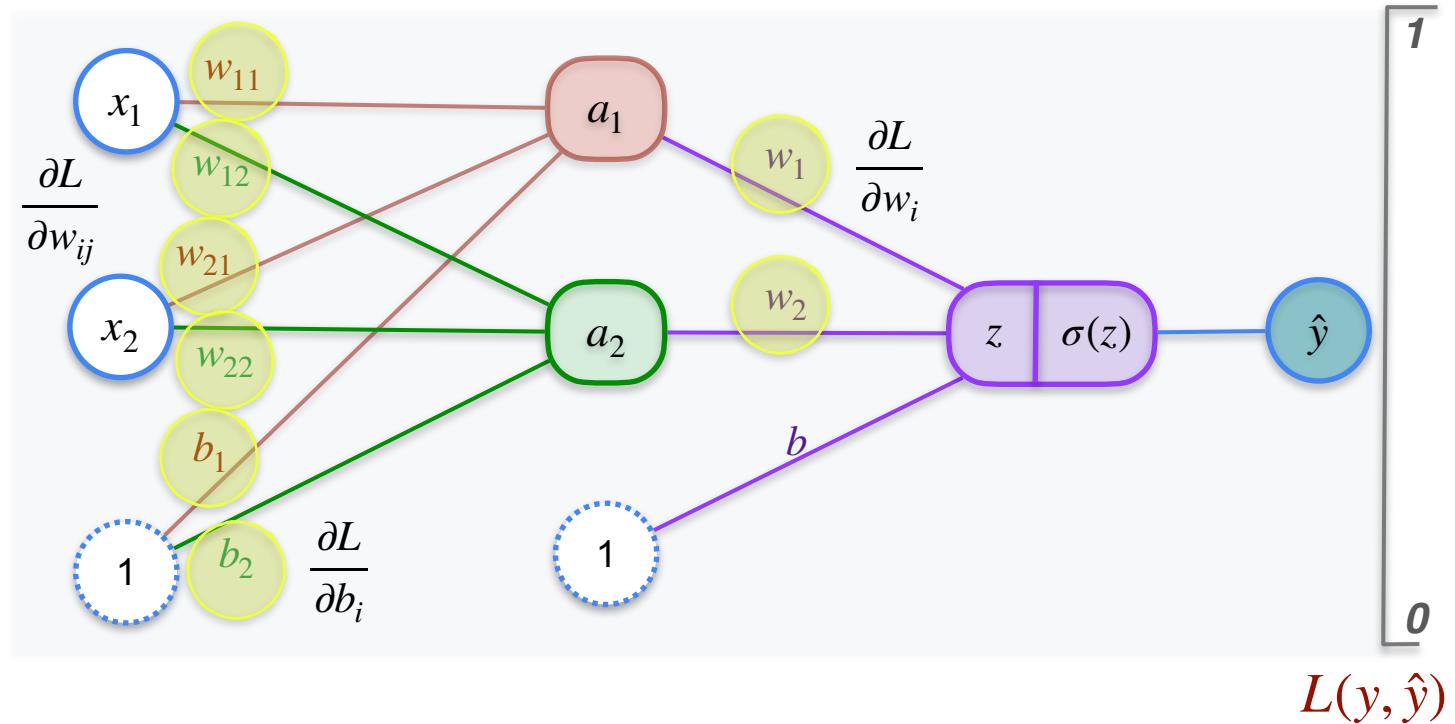
# 2,2,1 Neural Network



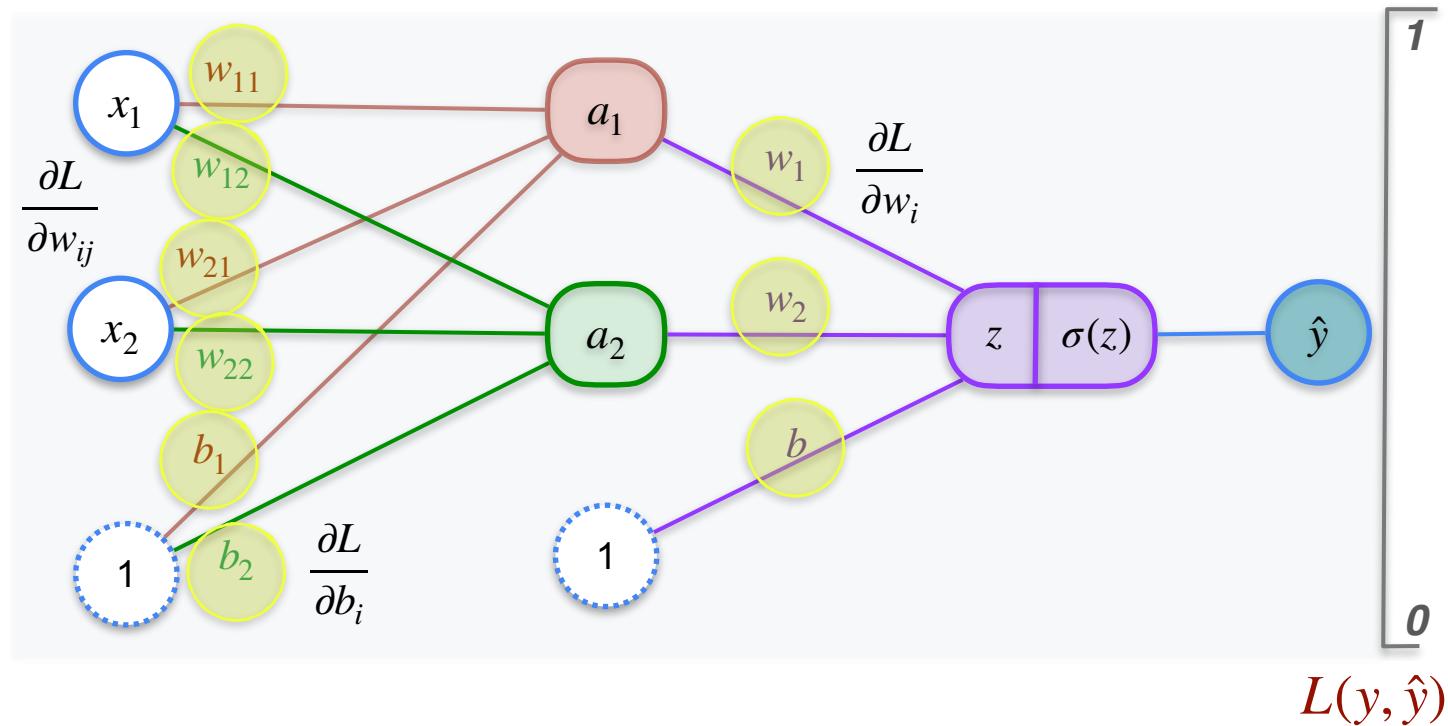
# 2,2,1 Neural Network



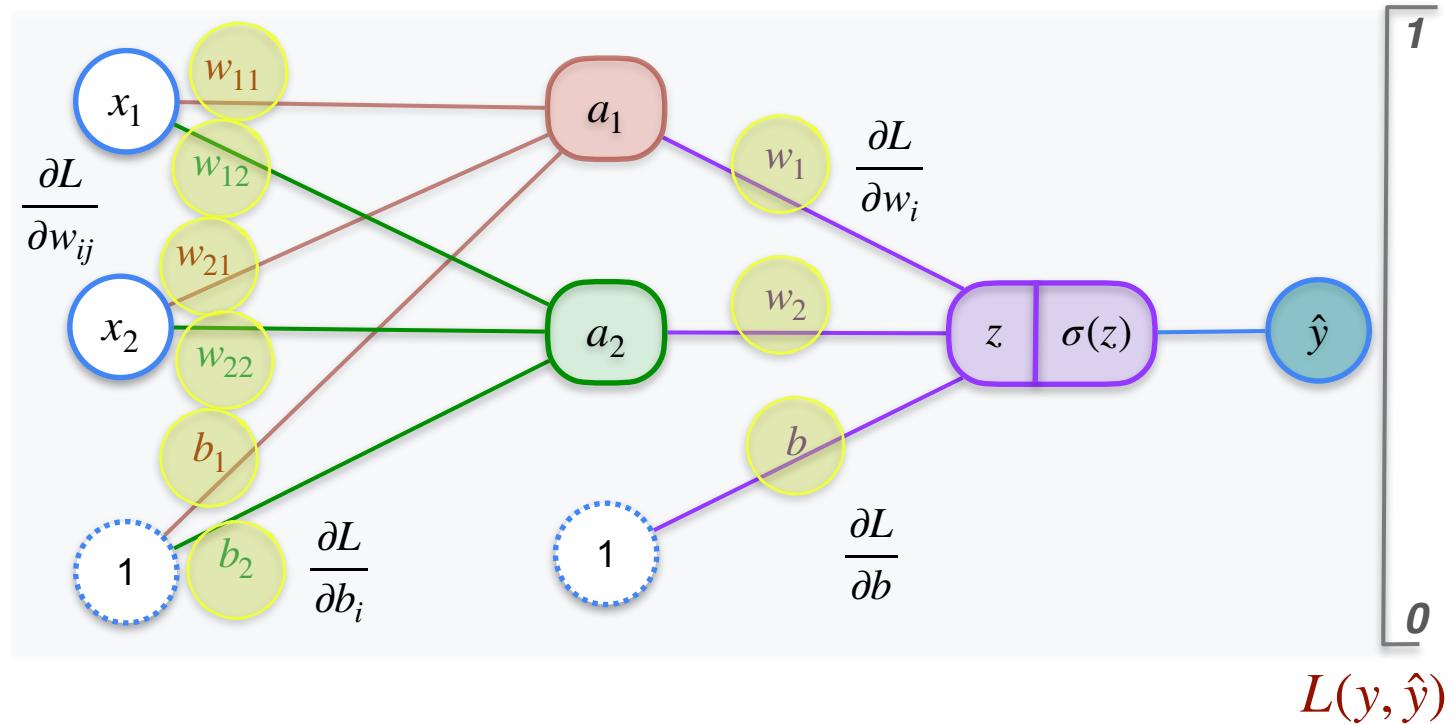
# 2,2,1 Neural Network



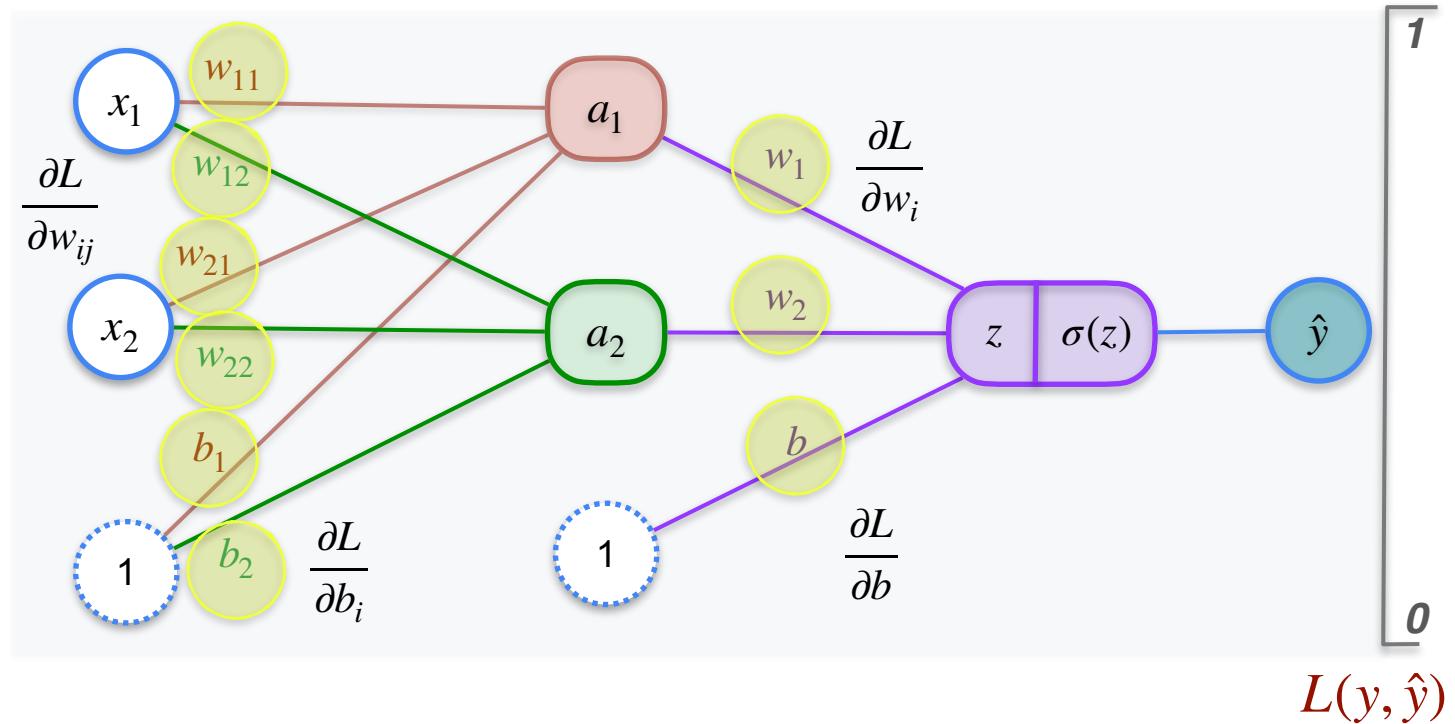
# 2,2,1 Neural Network



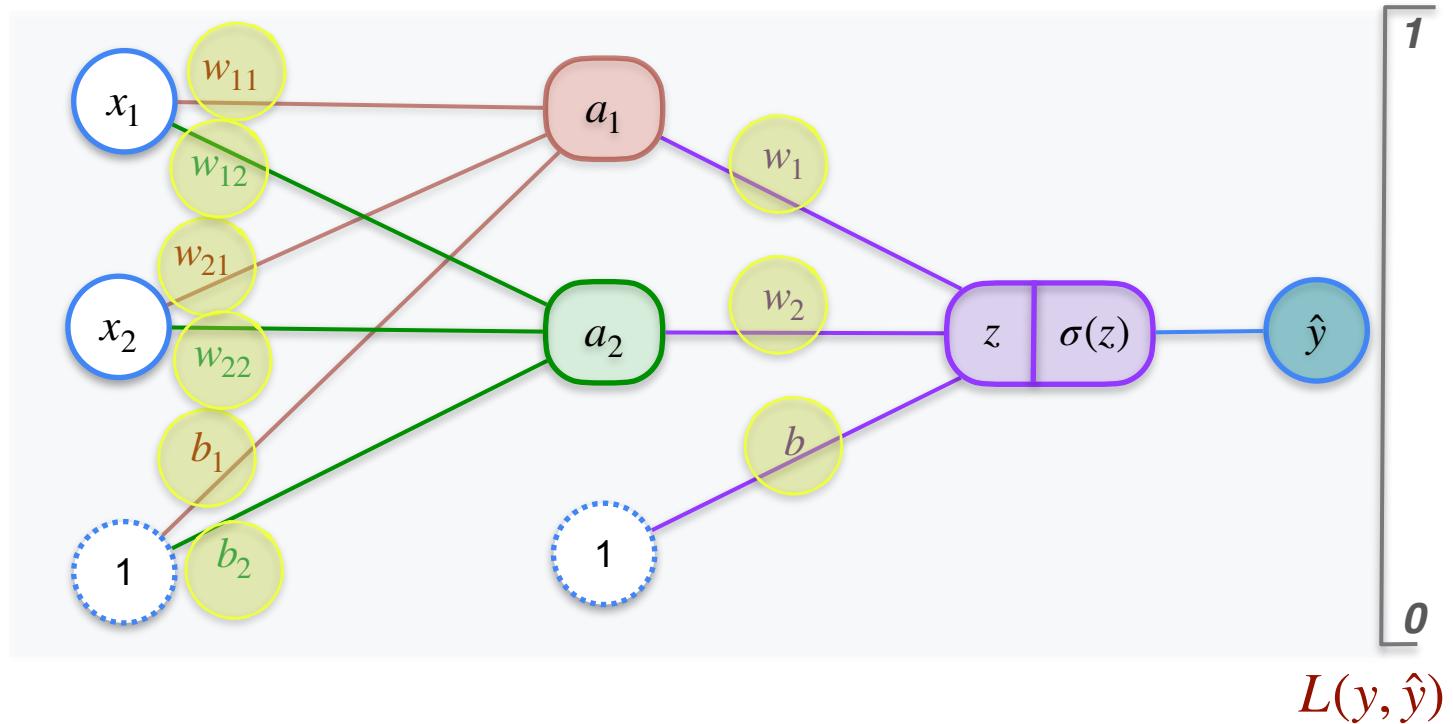
# 2,2,1 Neural Network



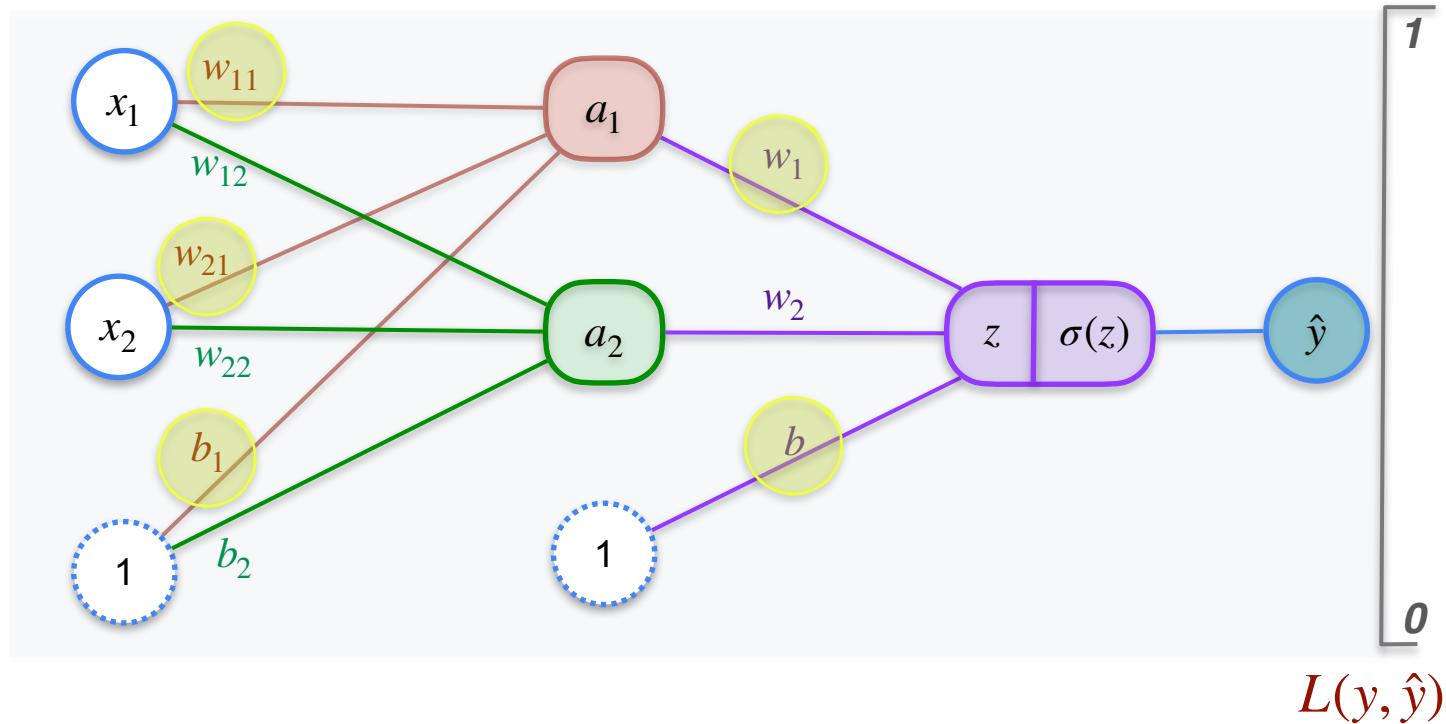
# 2,2,1 Neural Network



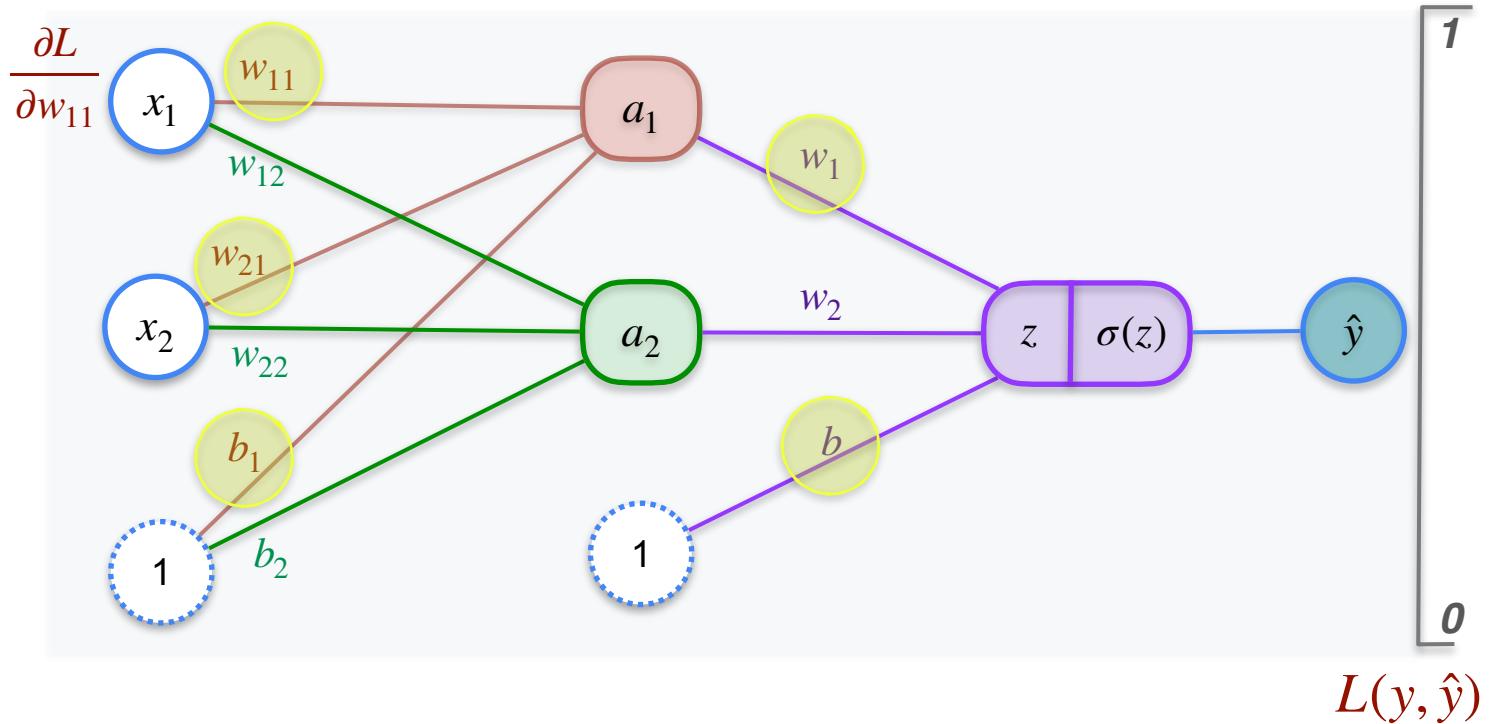
# 2,2,1 Neural Network



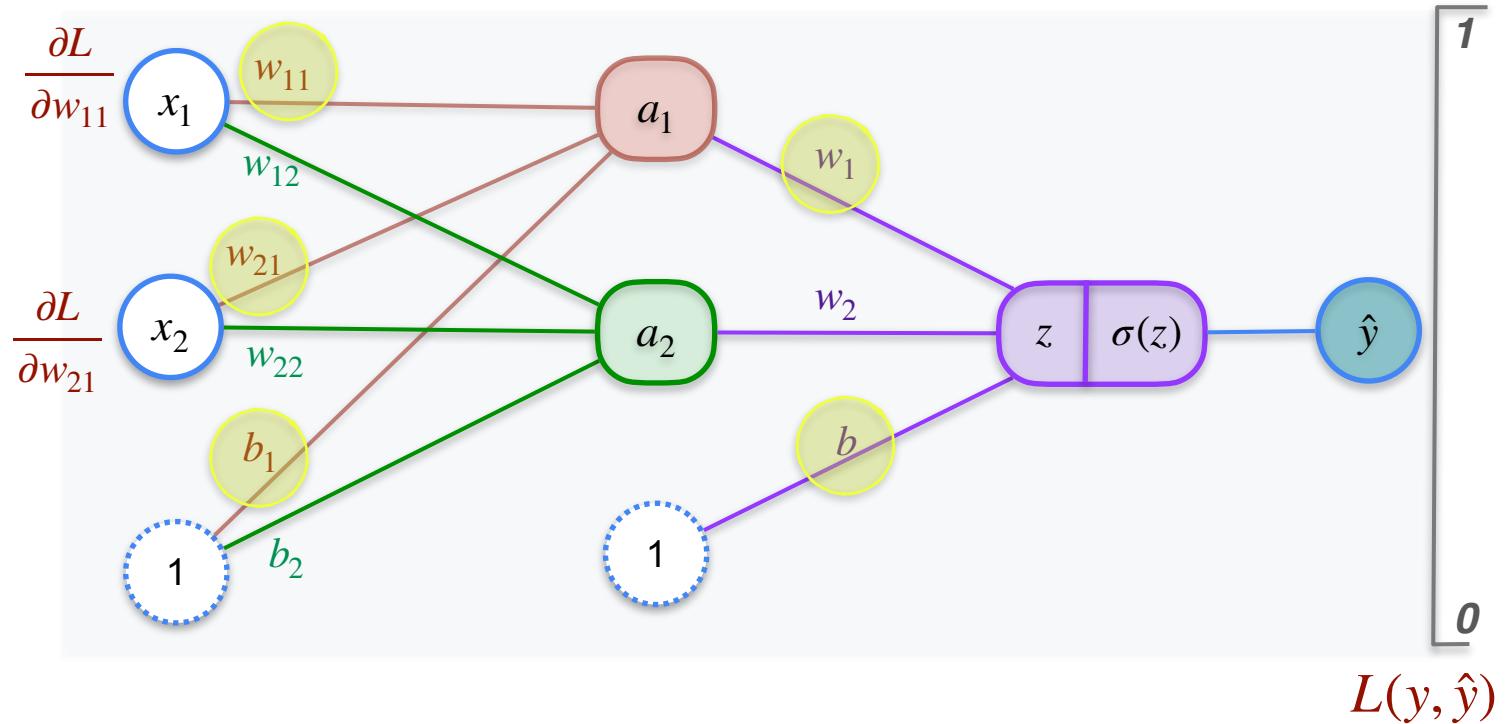
# 2,2,1 Neural Network



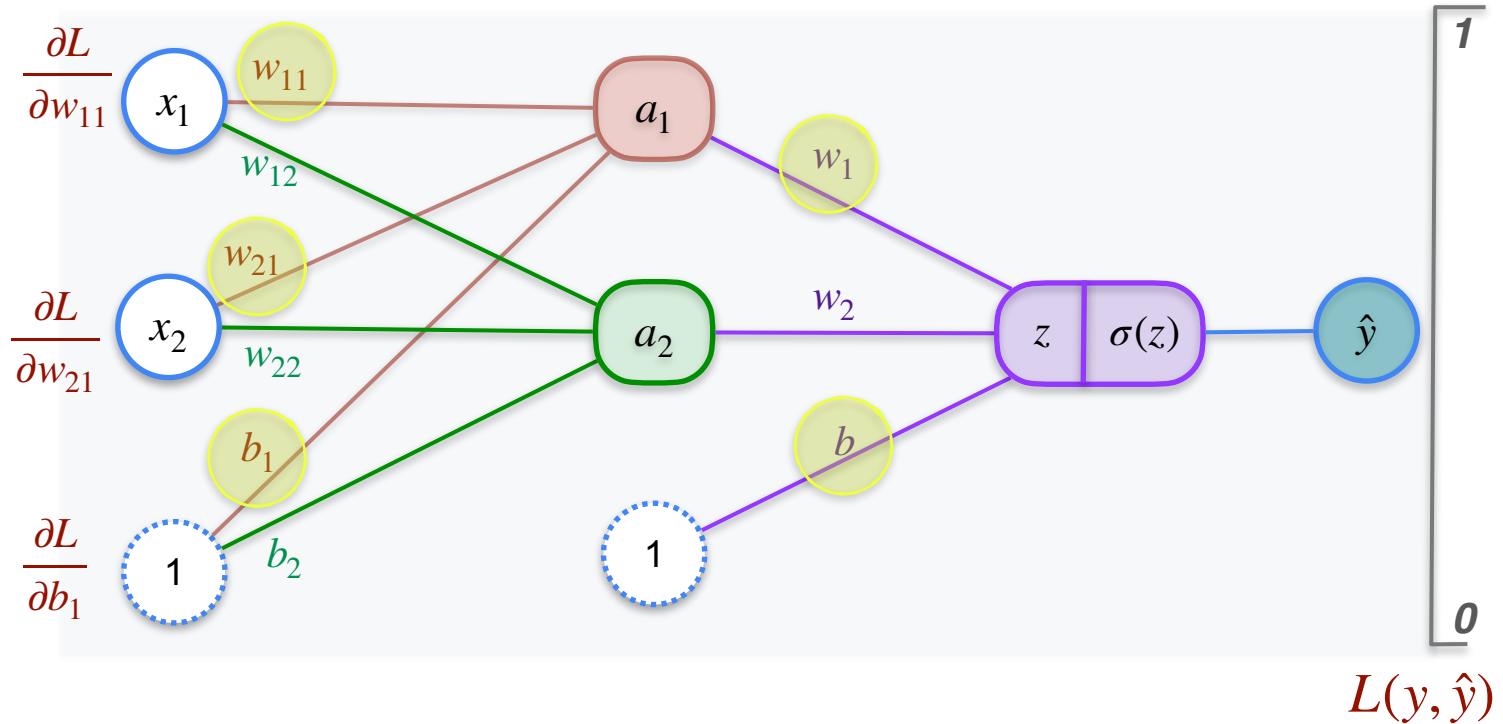
# 2,2,1 Neural Network



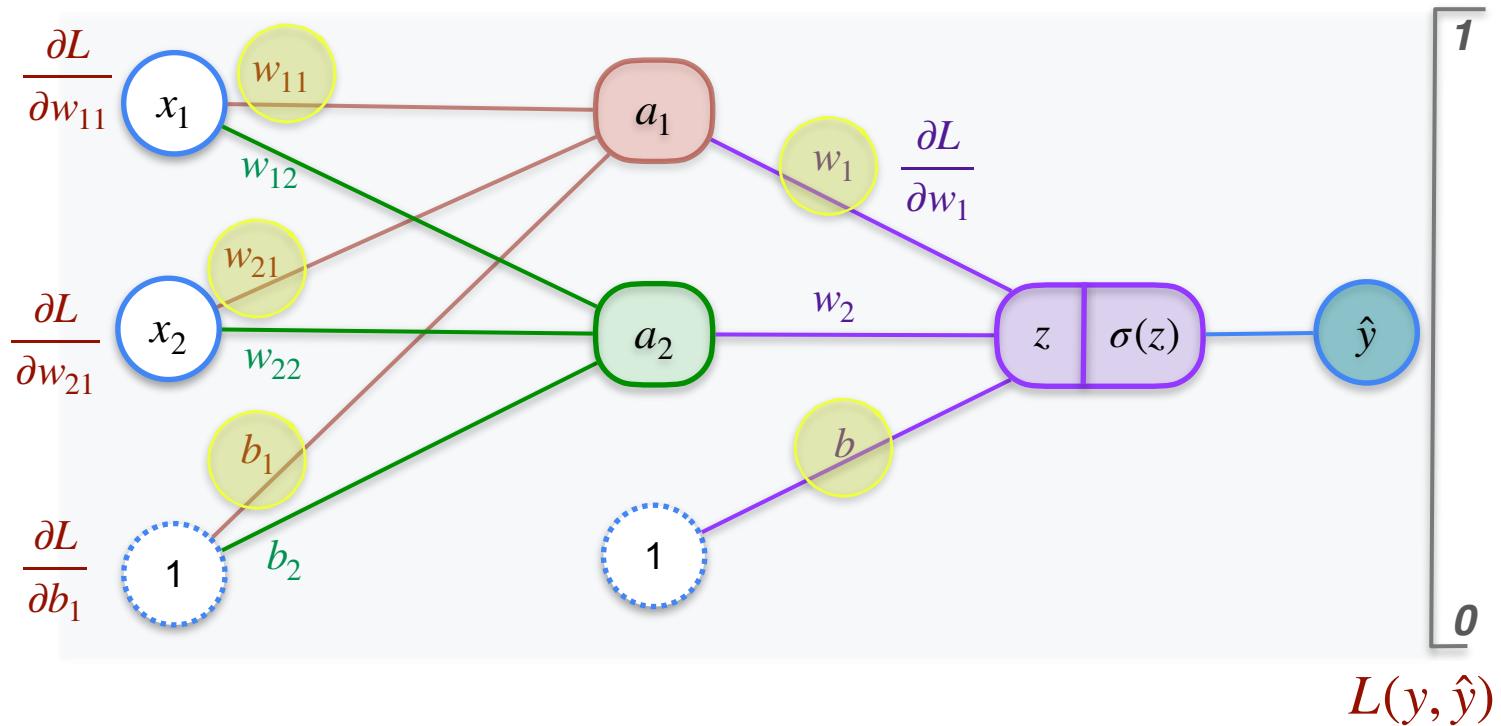
# 2,2,1 Neural Network



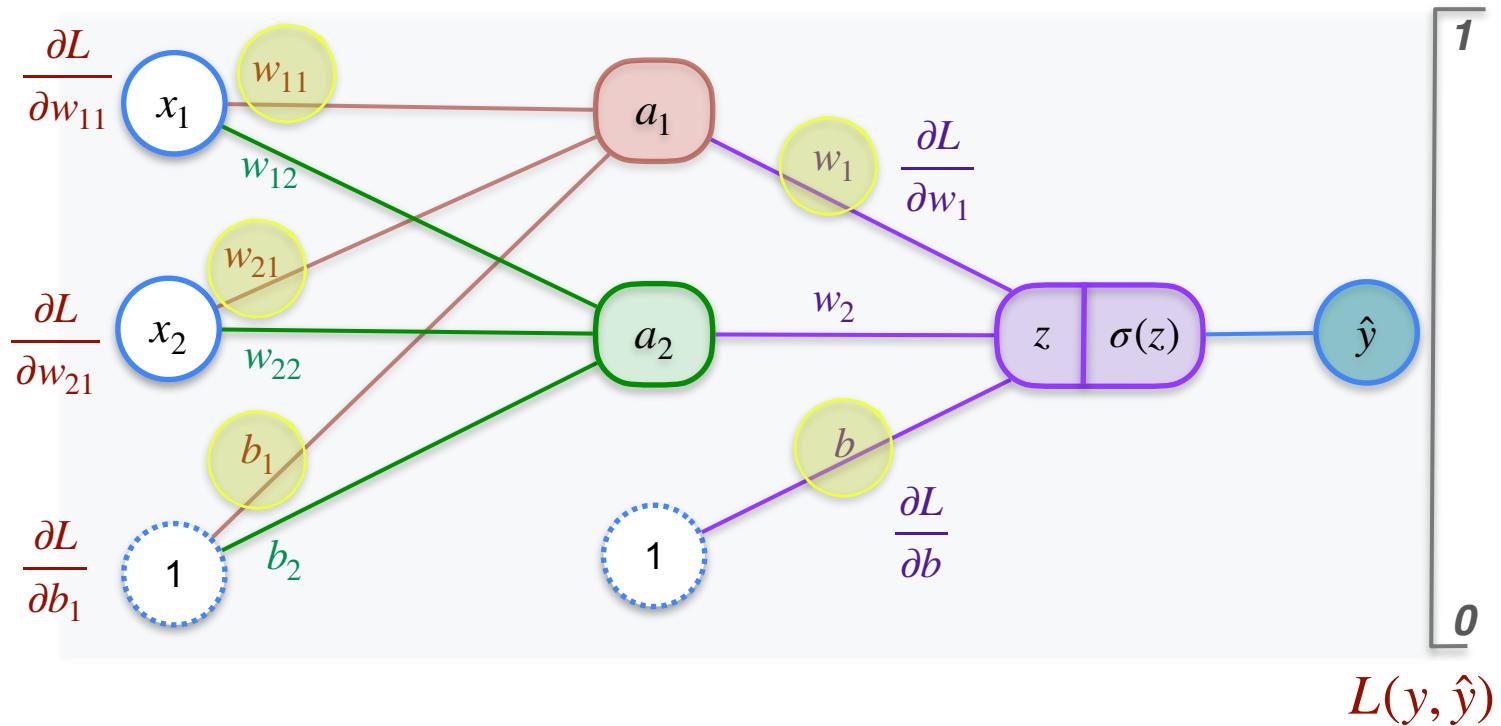
# 2,2,1 Neural Network



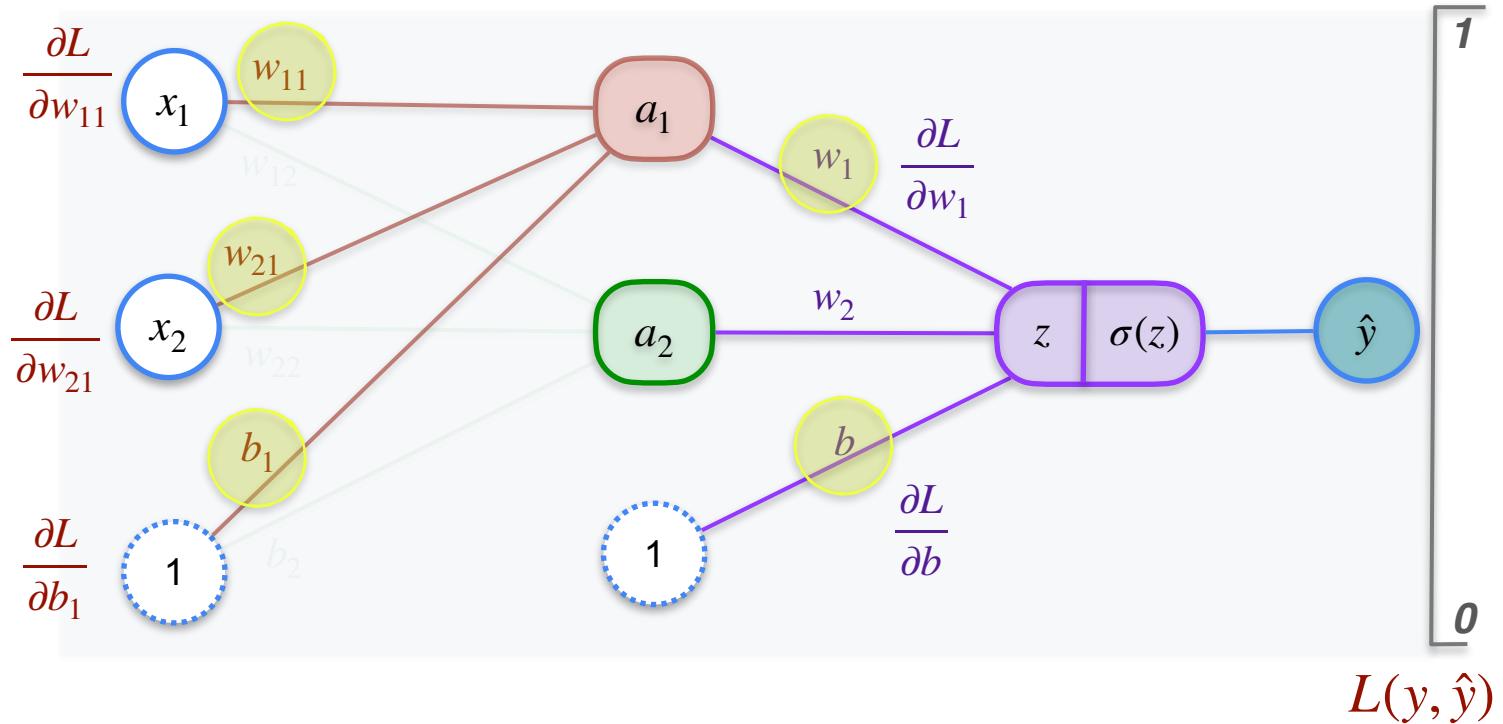
# 2,2,1 Neural Network



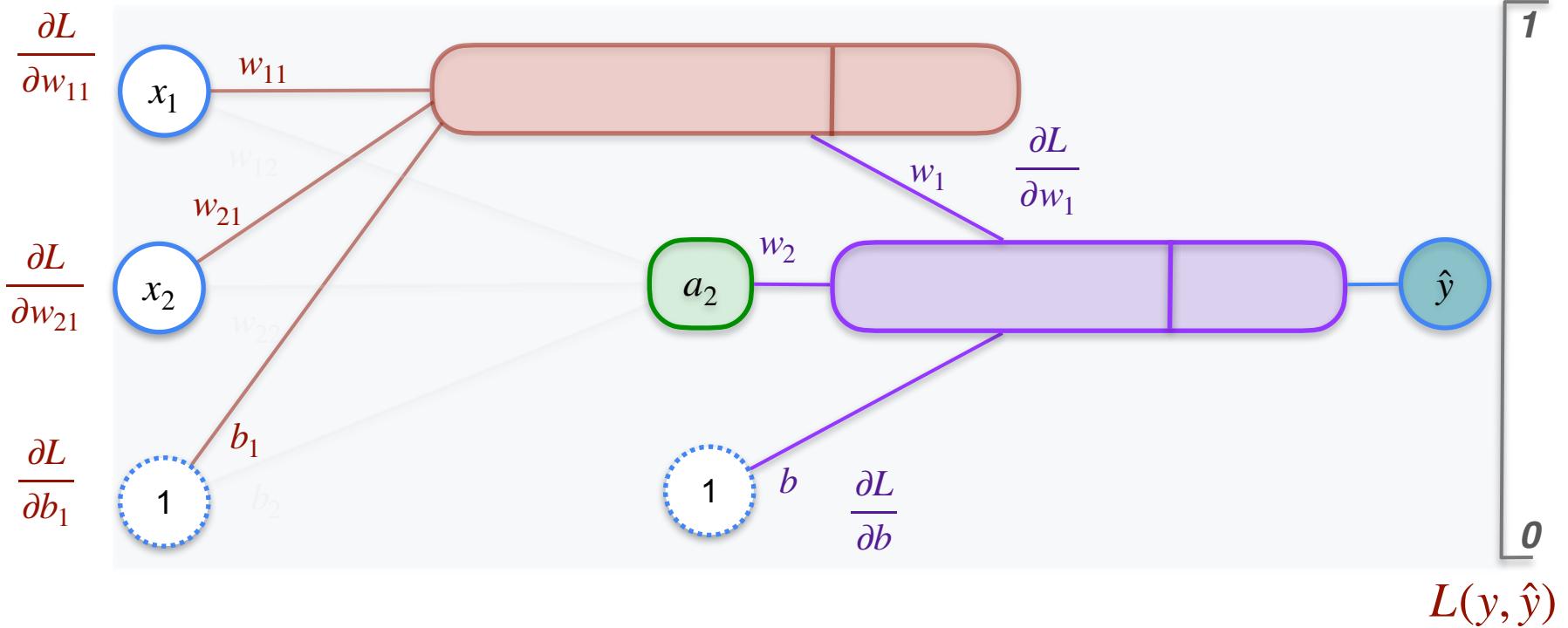
# 2,2,1 Neural Network



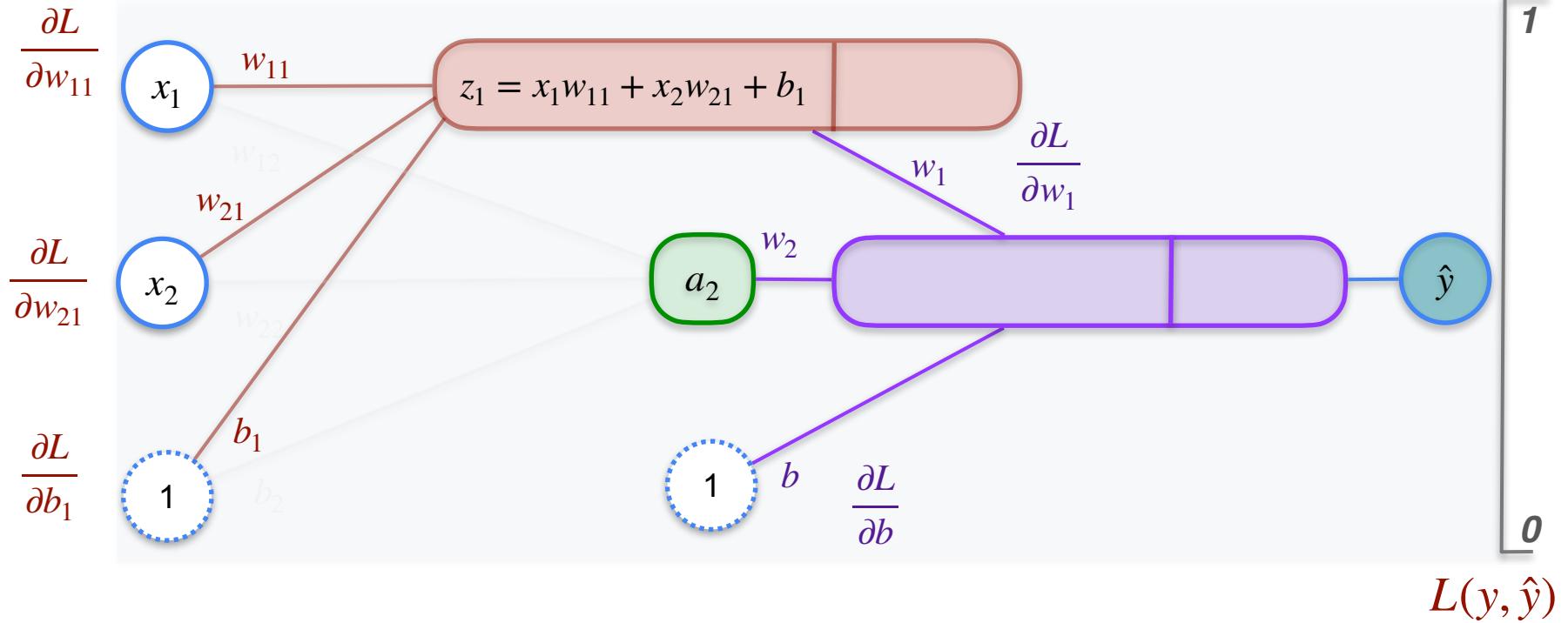
# 2,2,1 Neural Network



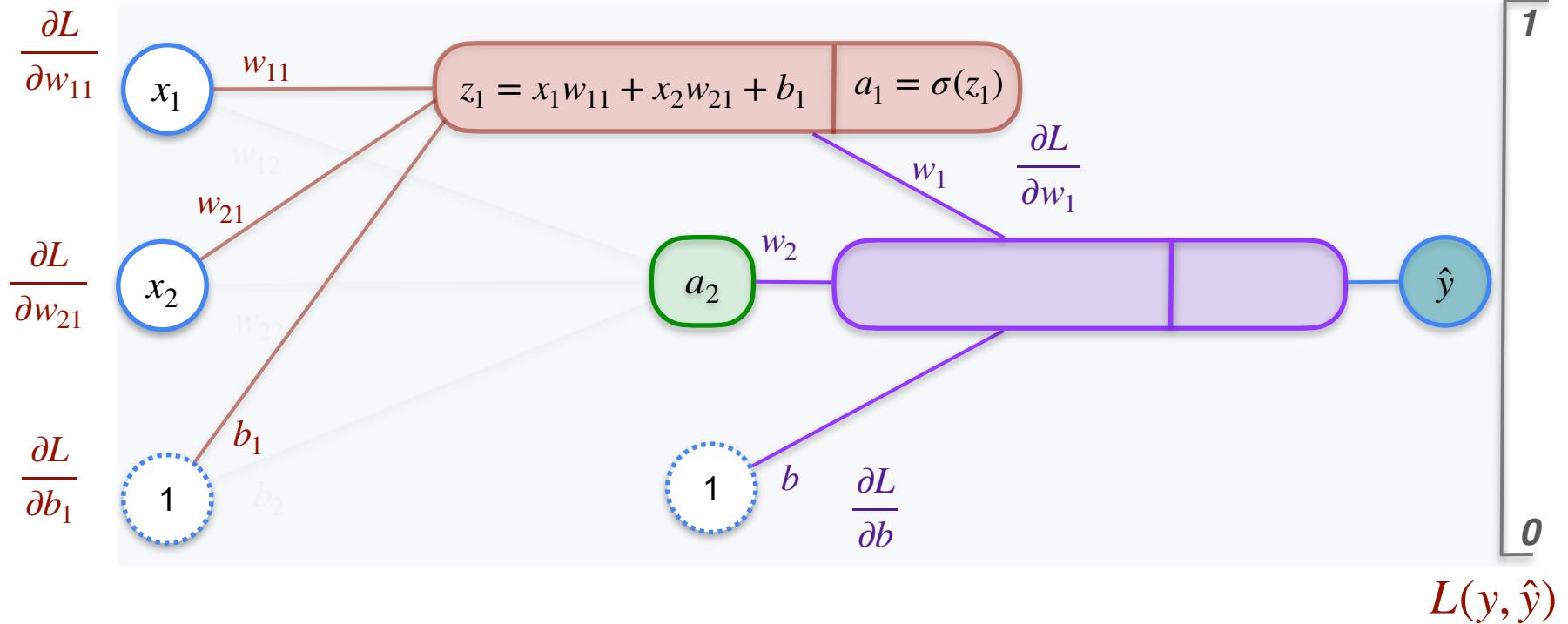
# 2,2,1 Neural Network



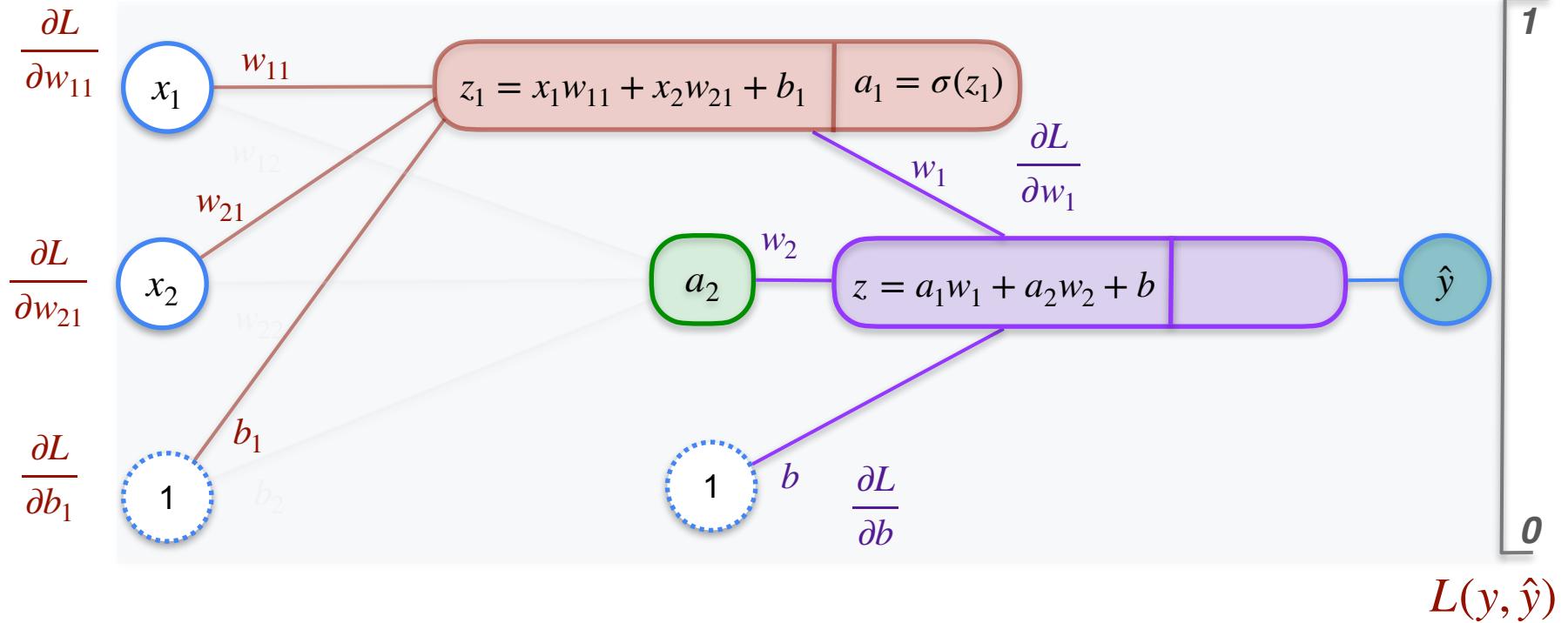
# 2,2,1 Neural Network



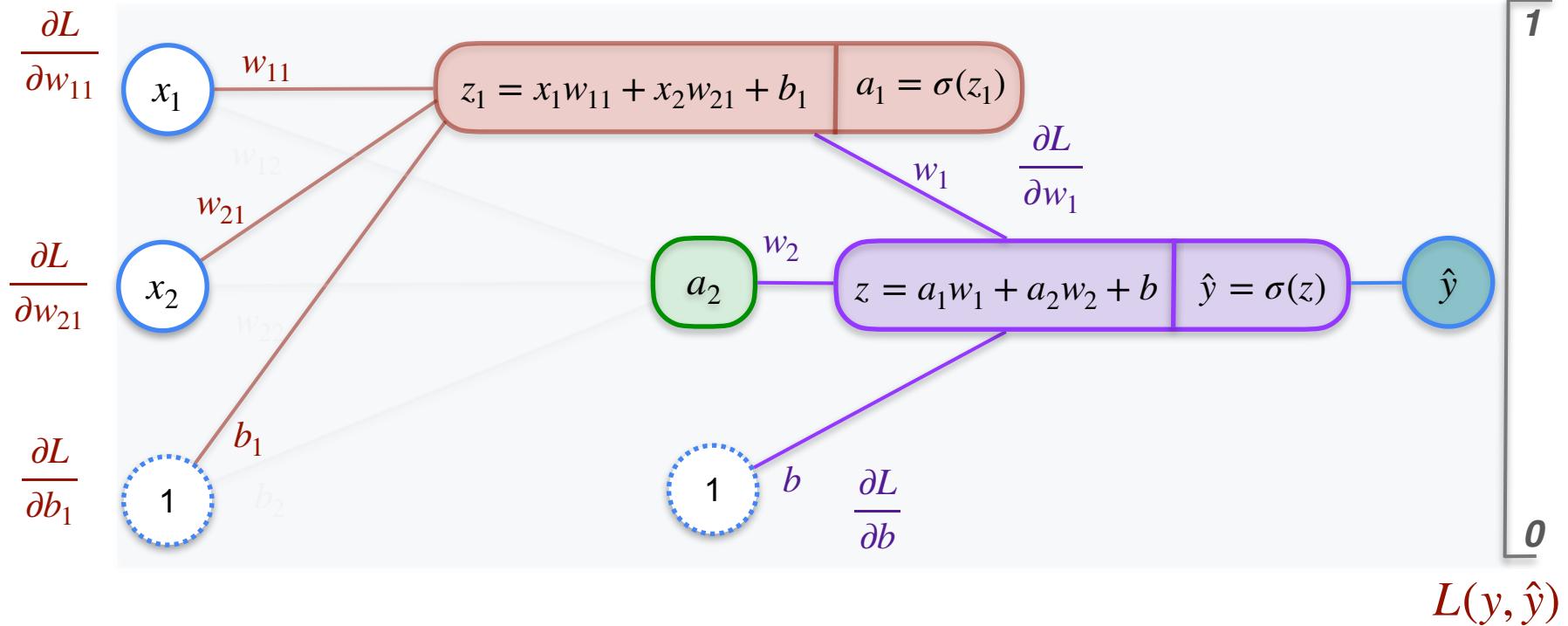
# 2,2,1 Neural Network



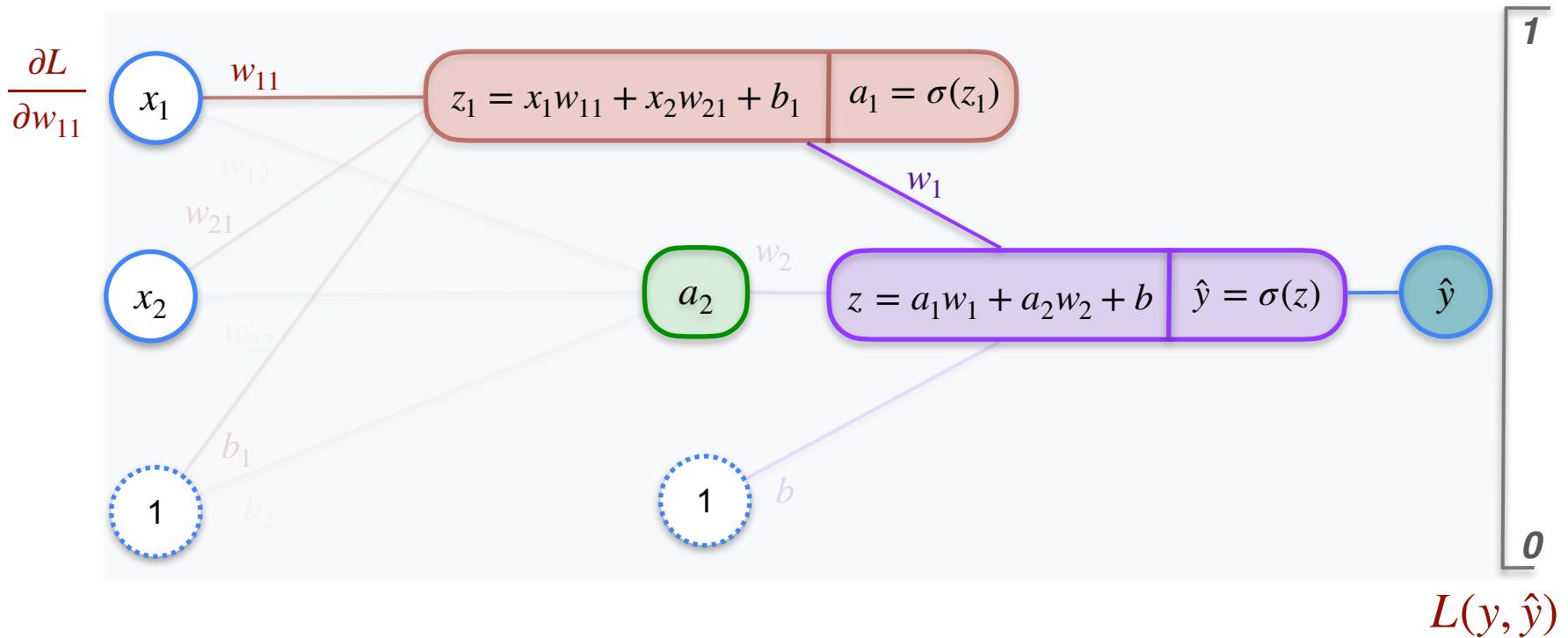
# 2,2,1 Neural Network



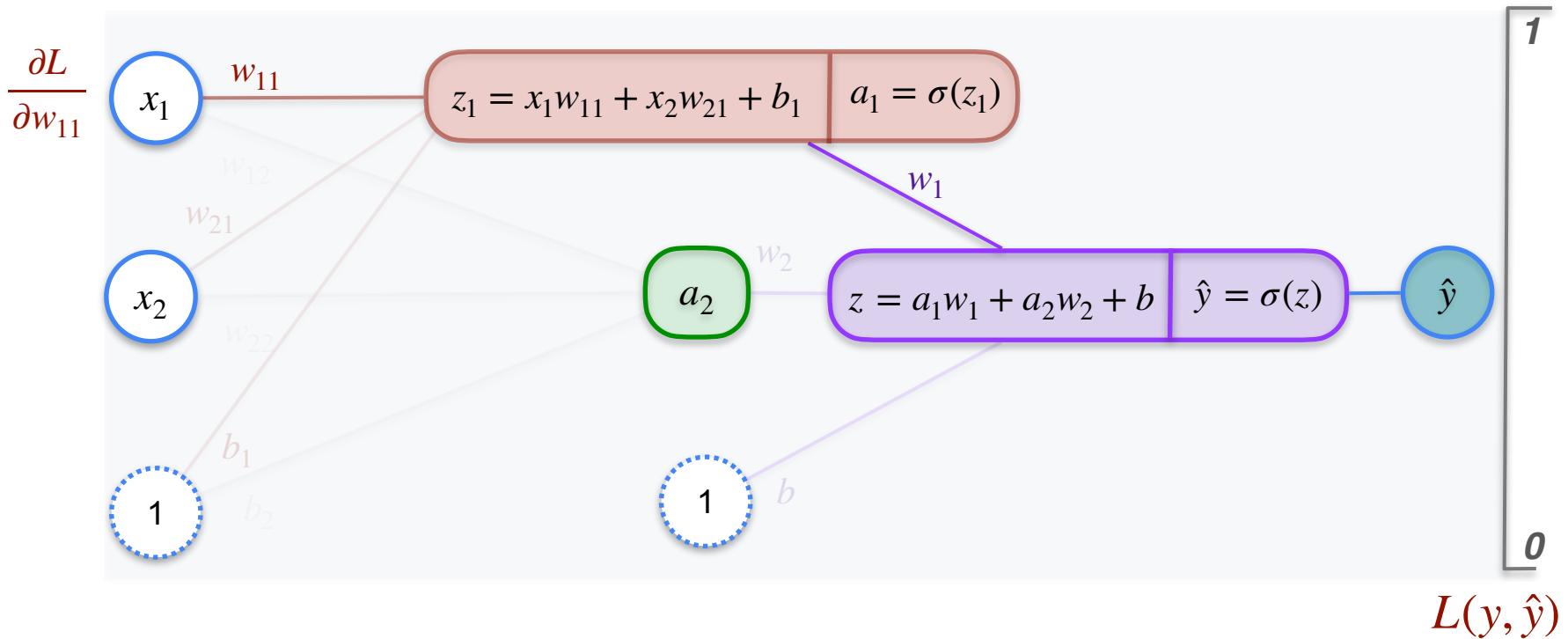
# 2,2,1 Neural Network



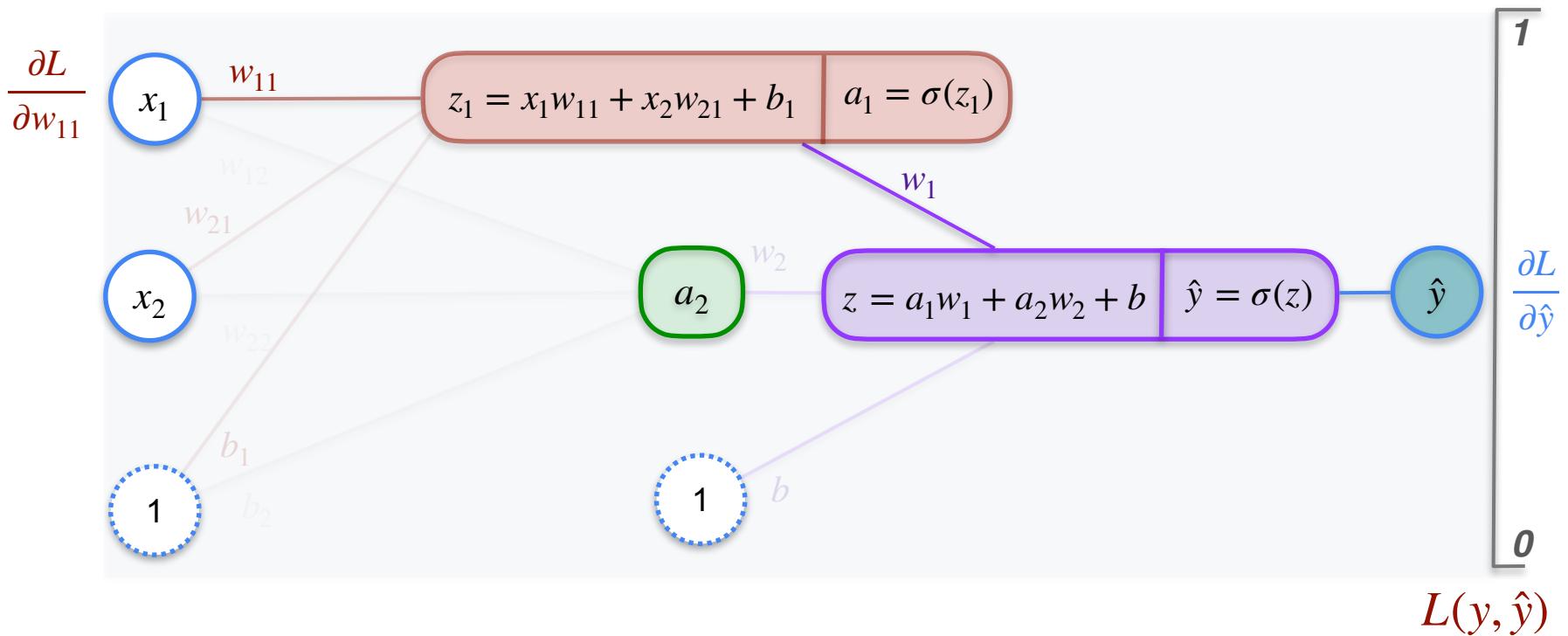
# 2,2,1 Neural Network



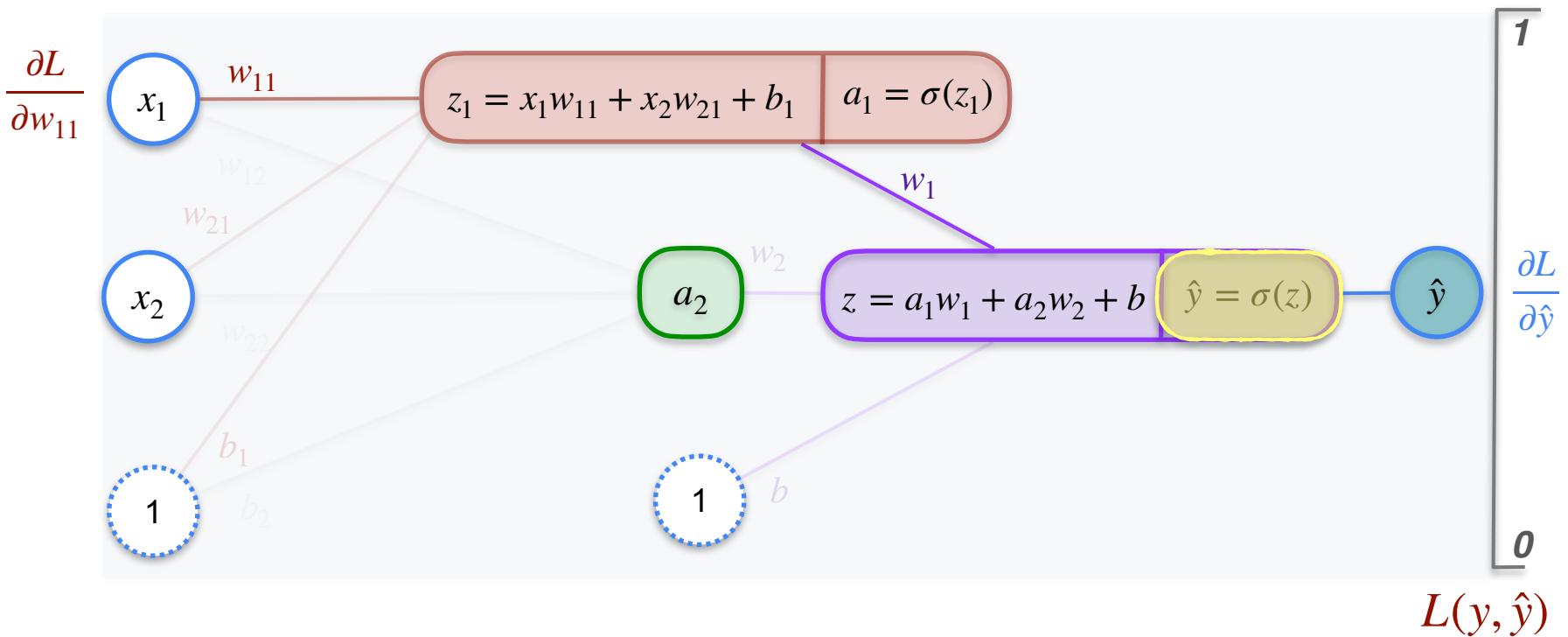
# 2,2,1 Neural Network



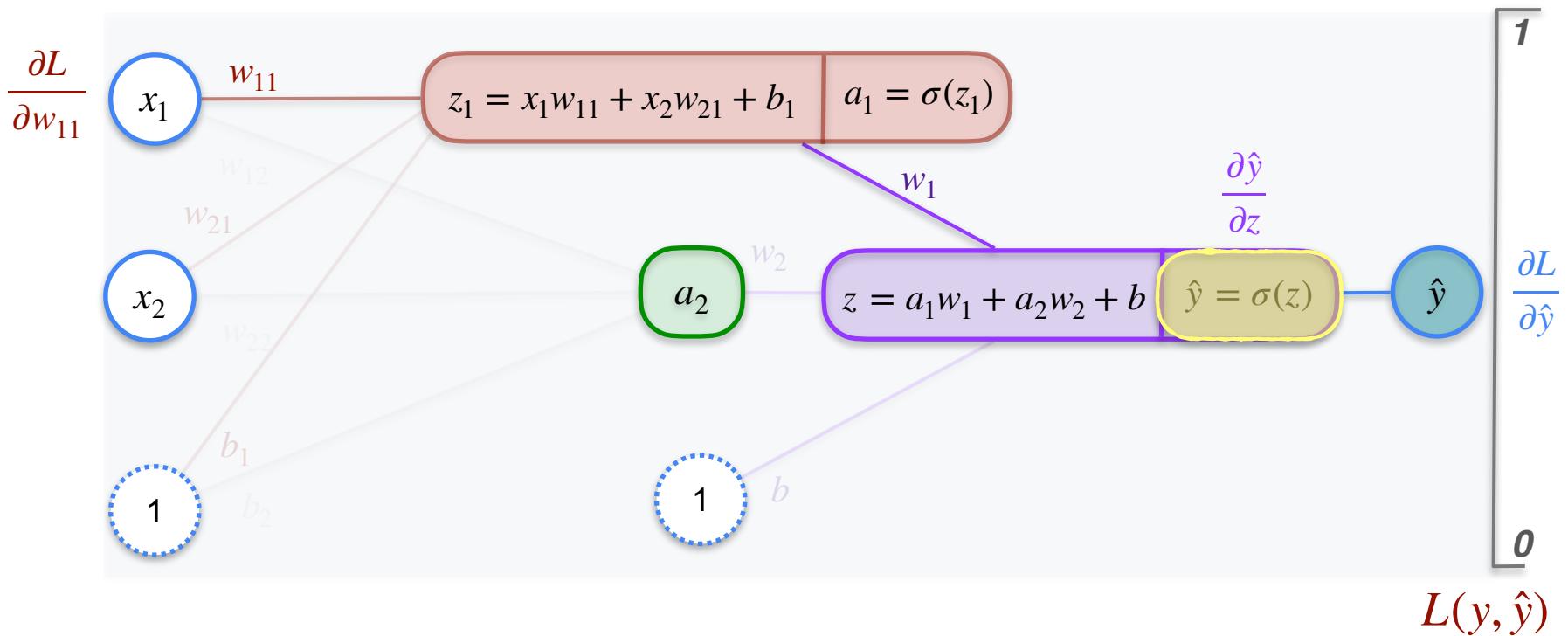
# 2,2,1 Neural Network



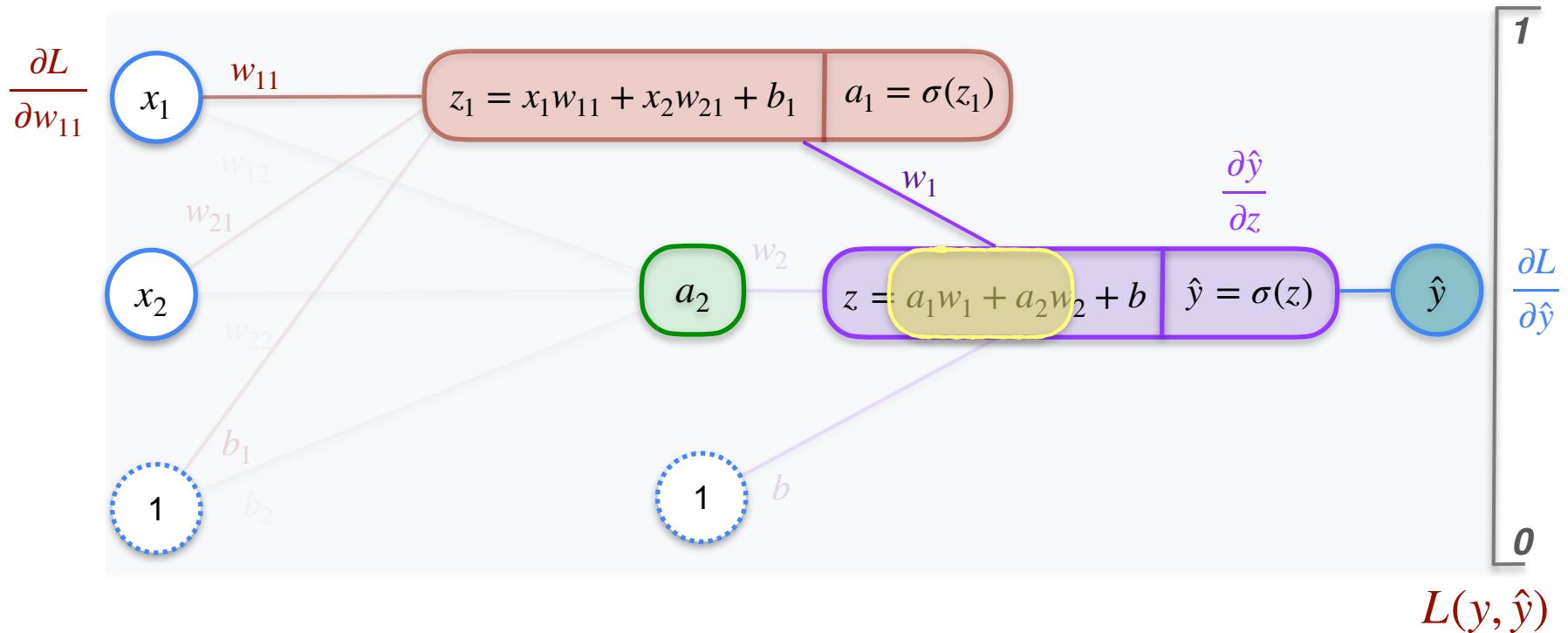
# 2,2,1 Neural Network



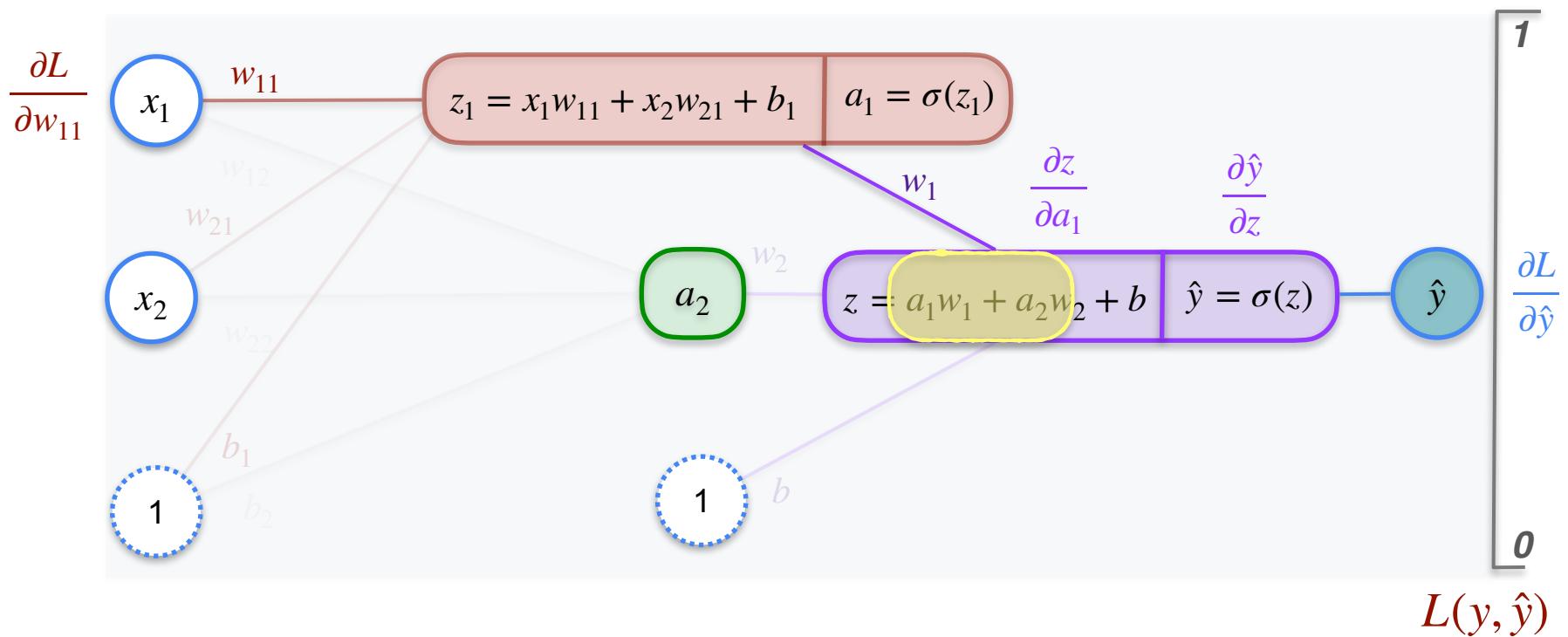
# 2,2,1 Neural Network



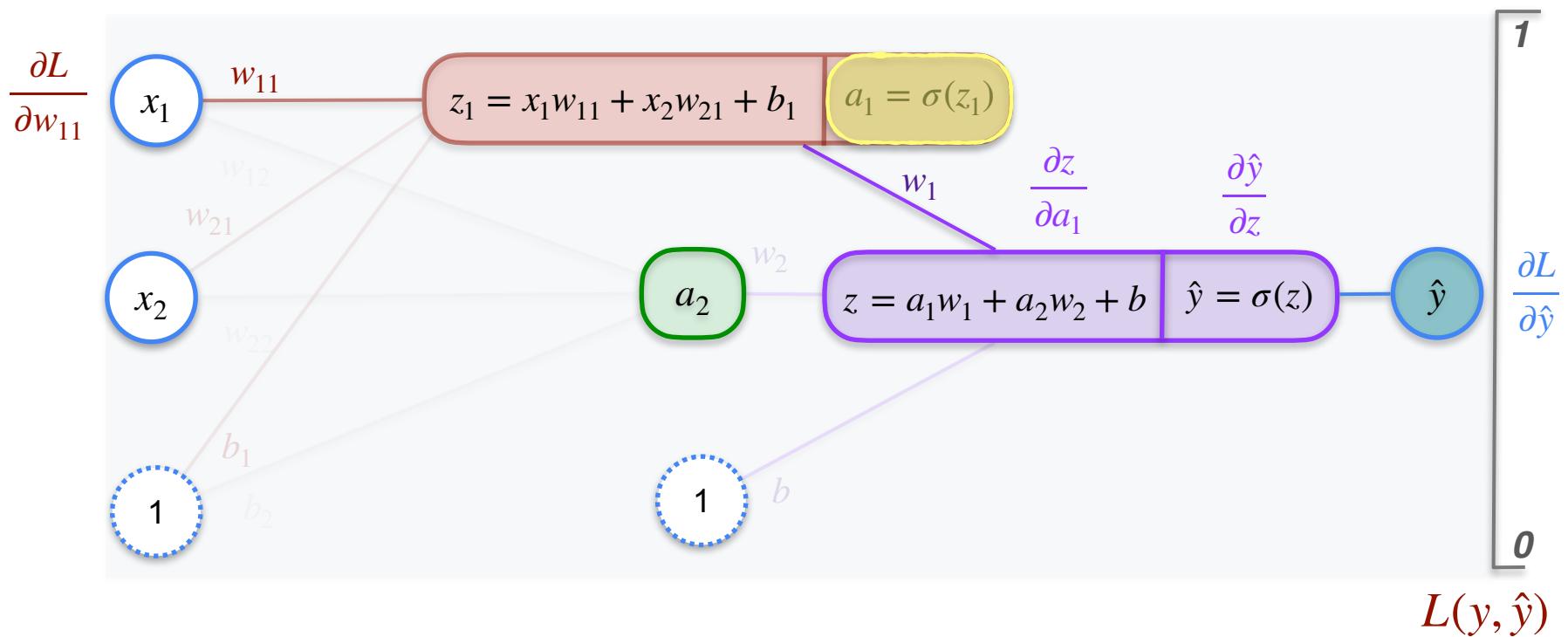
# 2,2,1 Neural Network



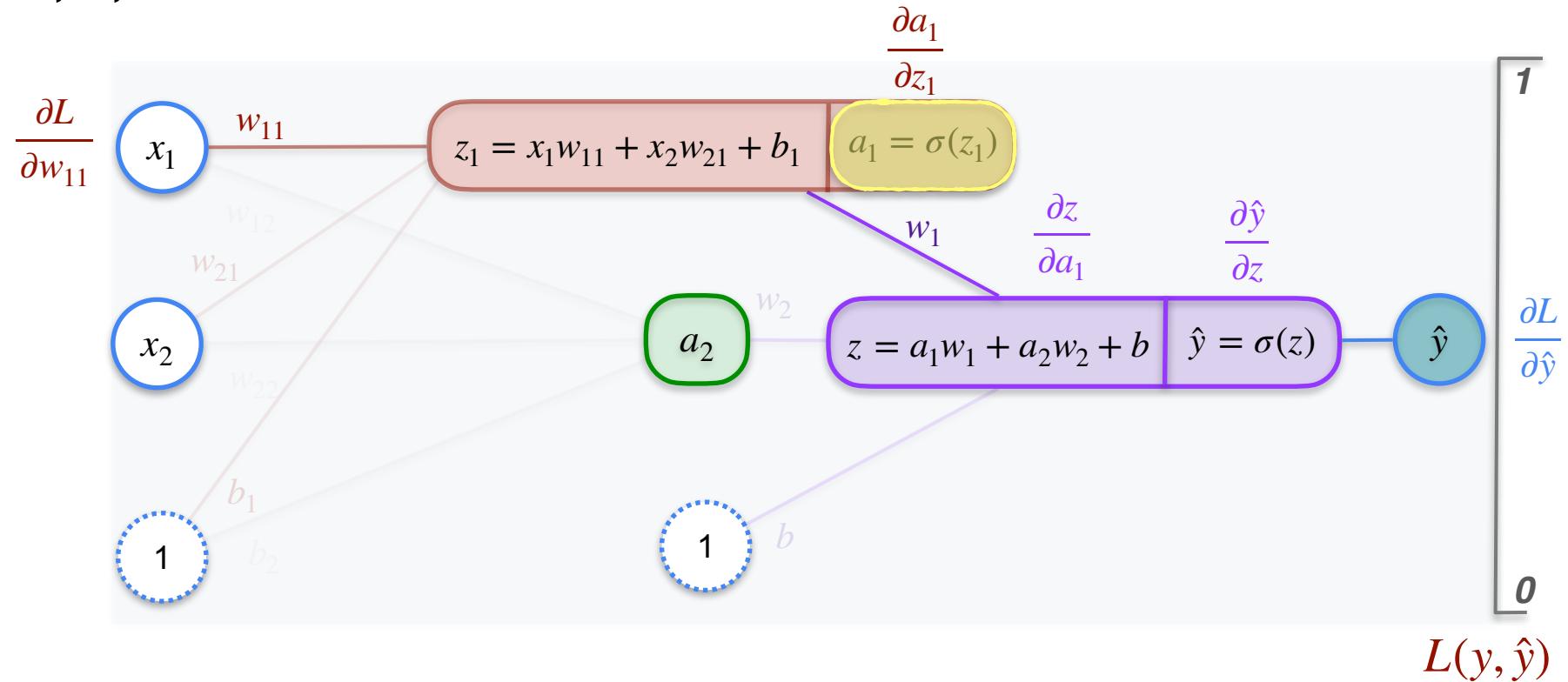
# 2,2,1 Neural Network



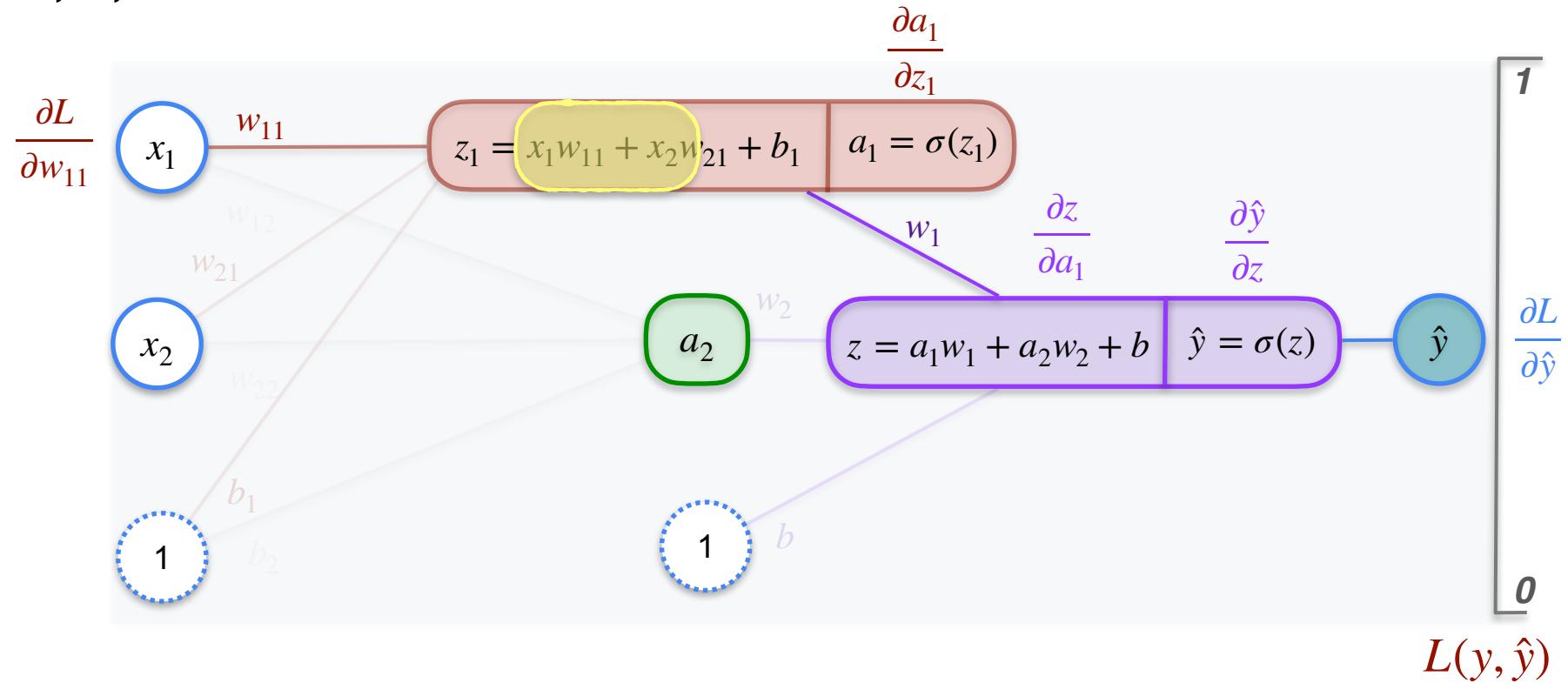
# 2,2,1 Neural Network



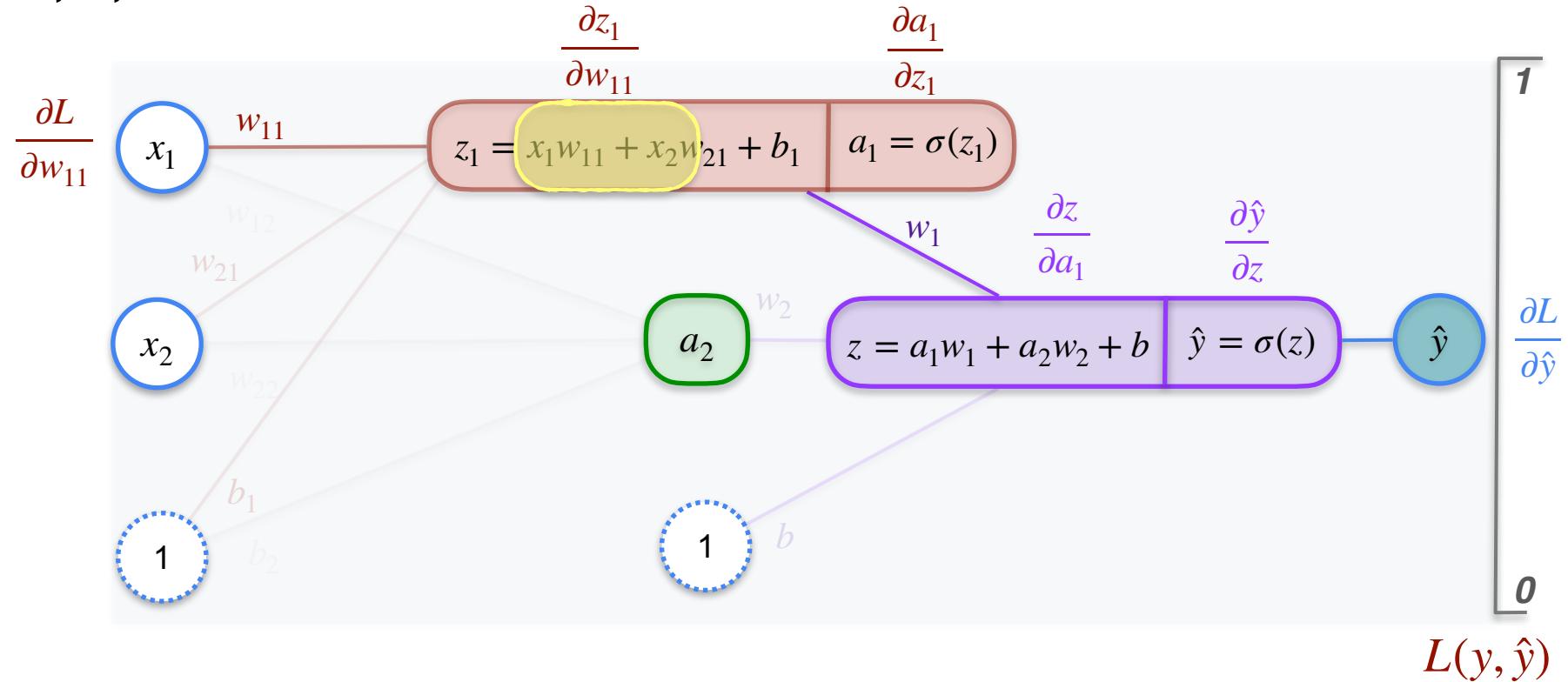
# 2,2,1 Neural Network



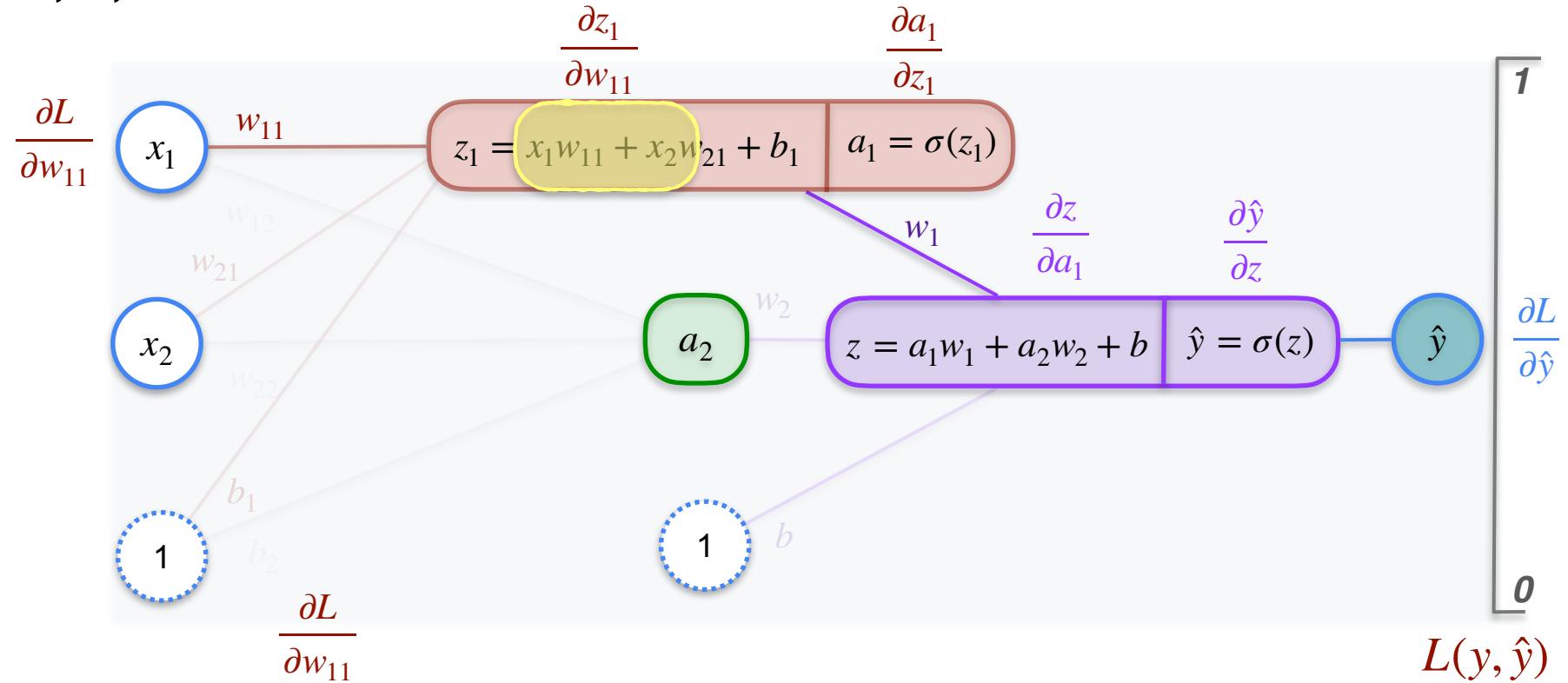
# 2,2,1 Neural Network



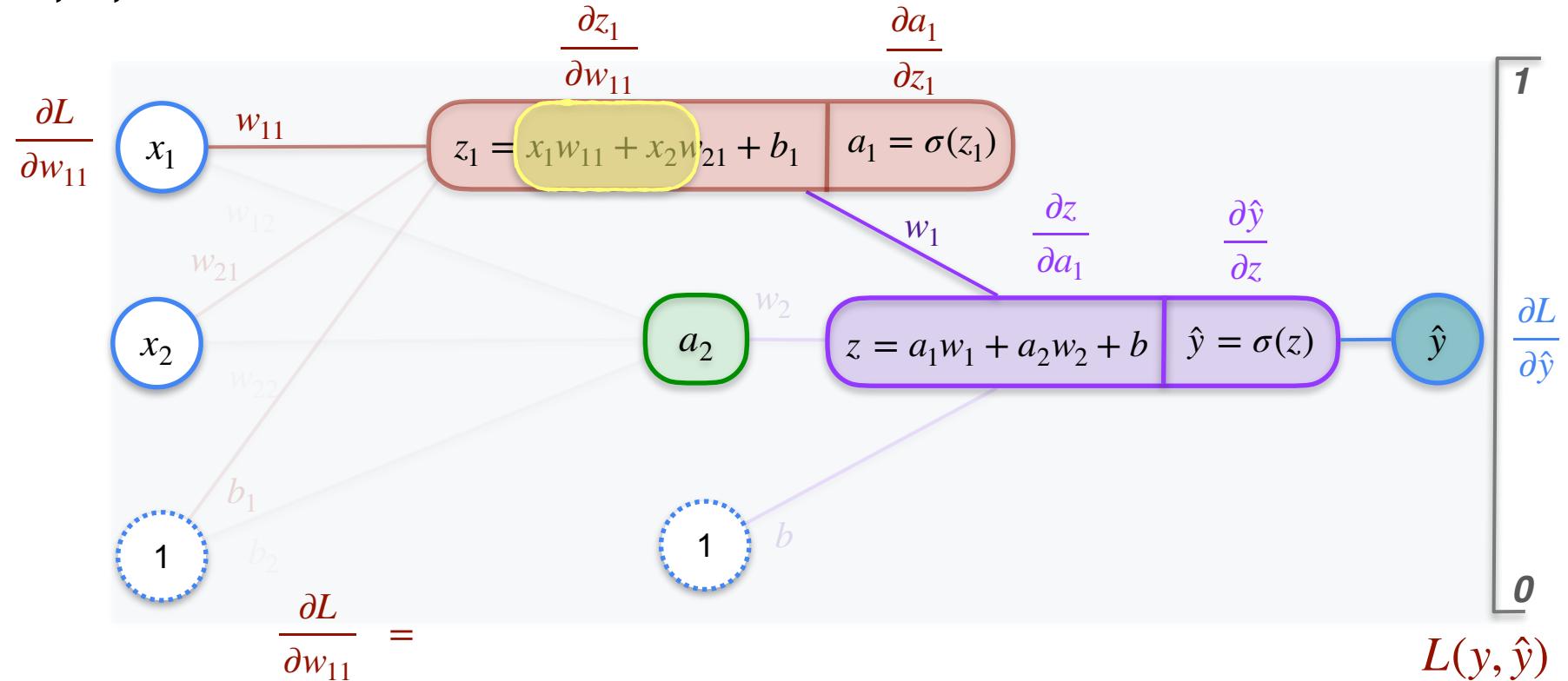
# 2,2,1 Neural Network



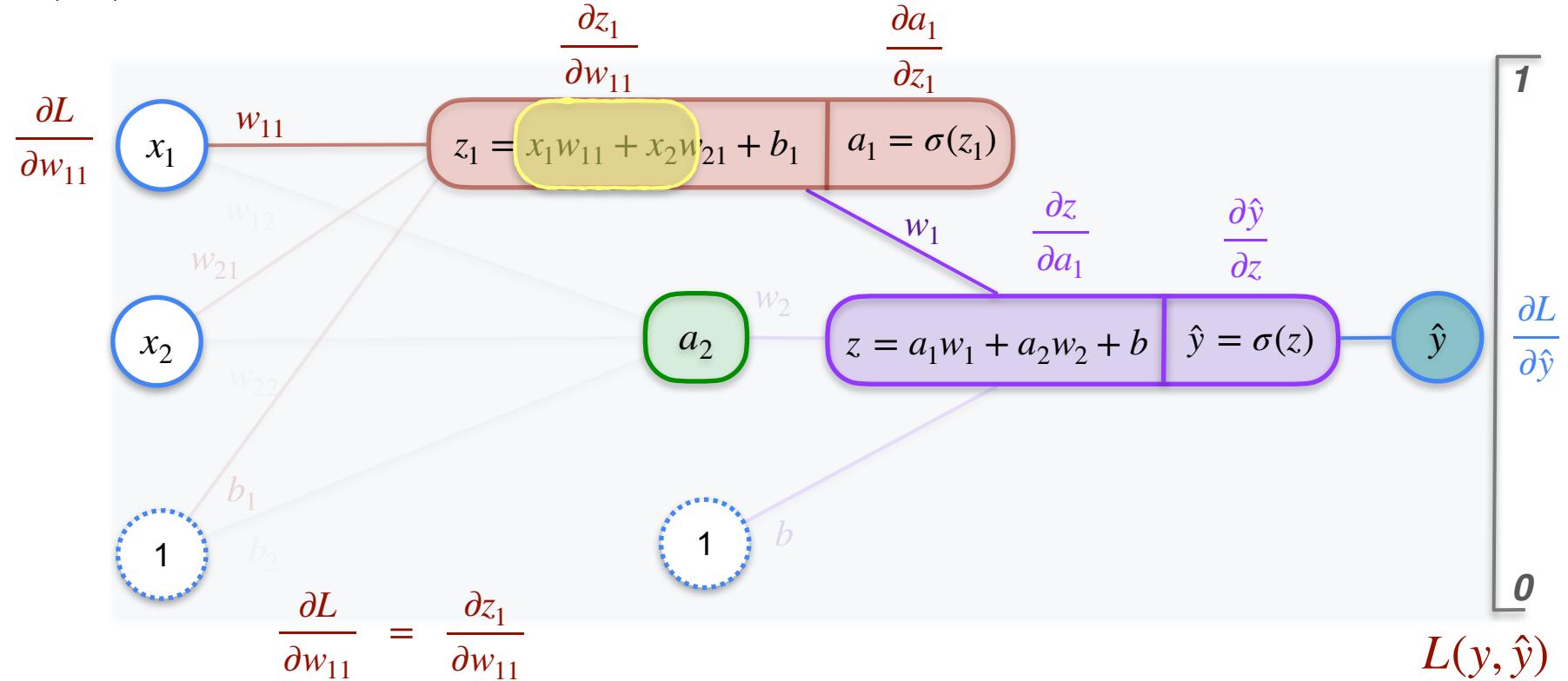
# 2,2,1 Neural Network



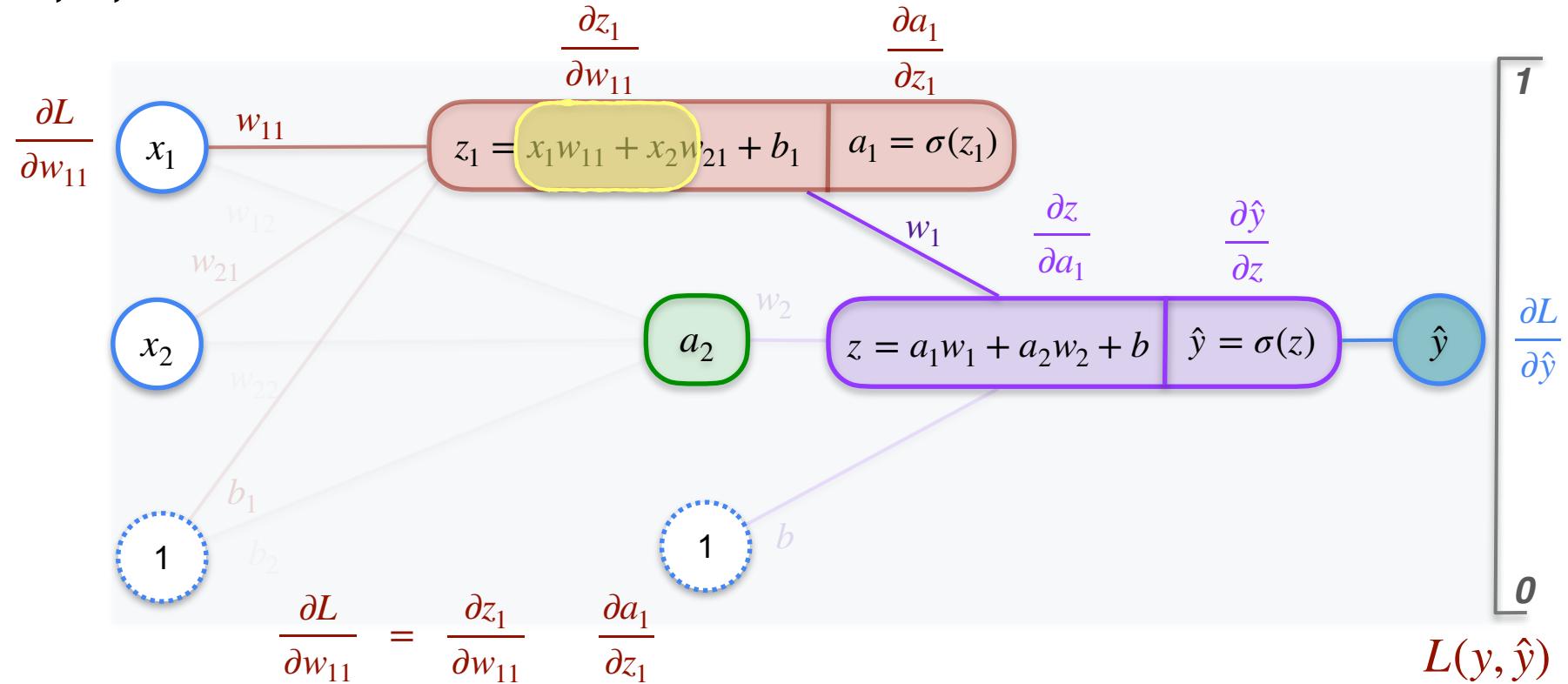
# 2,2,1 Neural Network



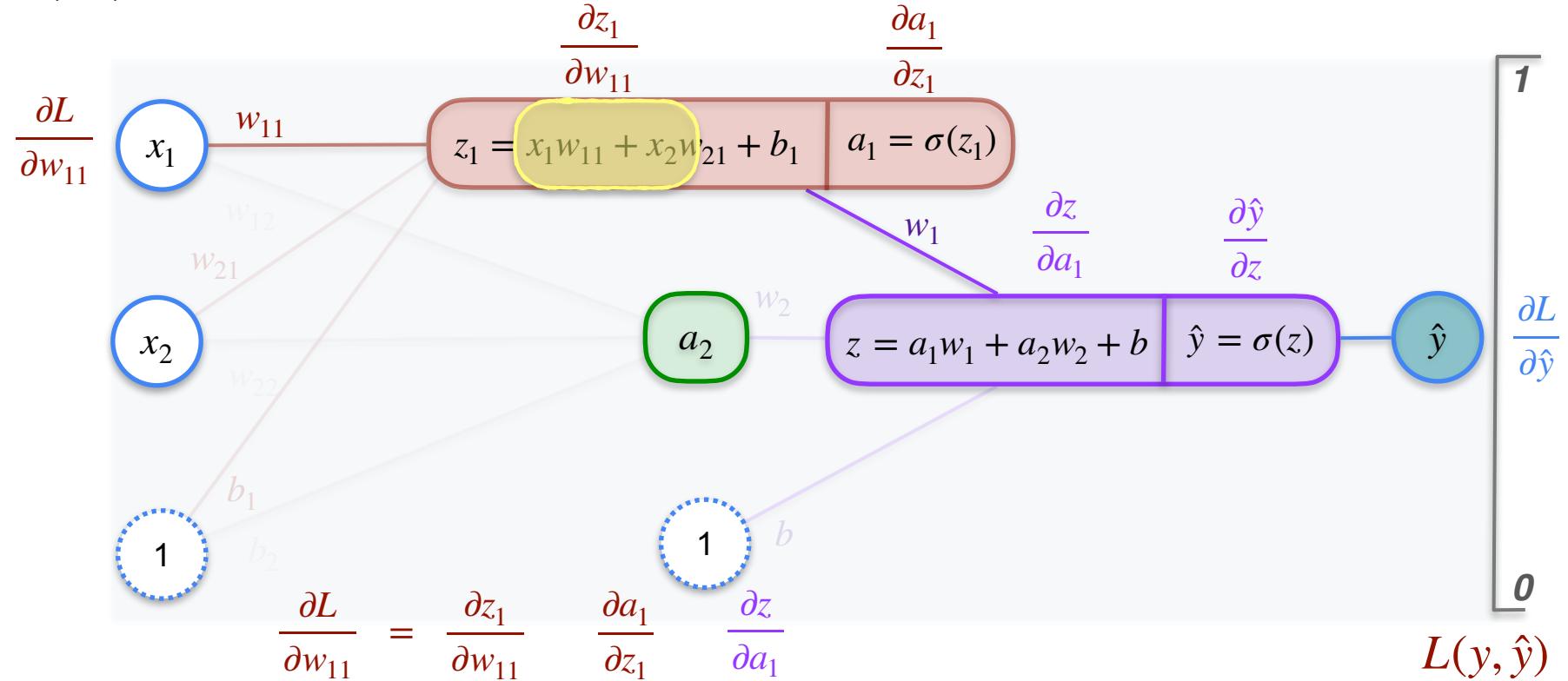
# 2,2,1 Neural Network



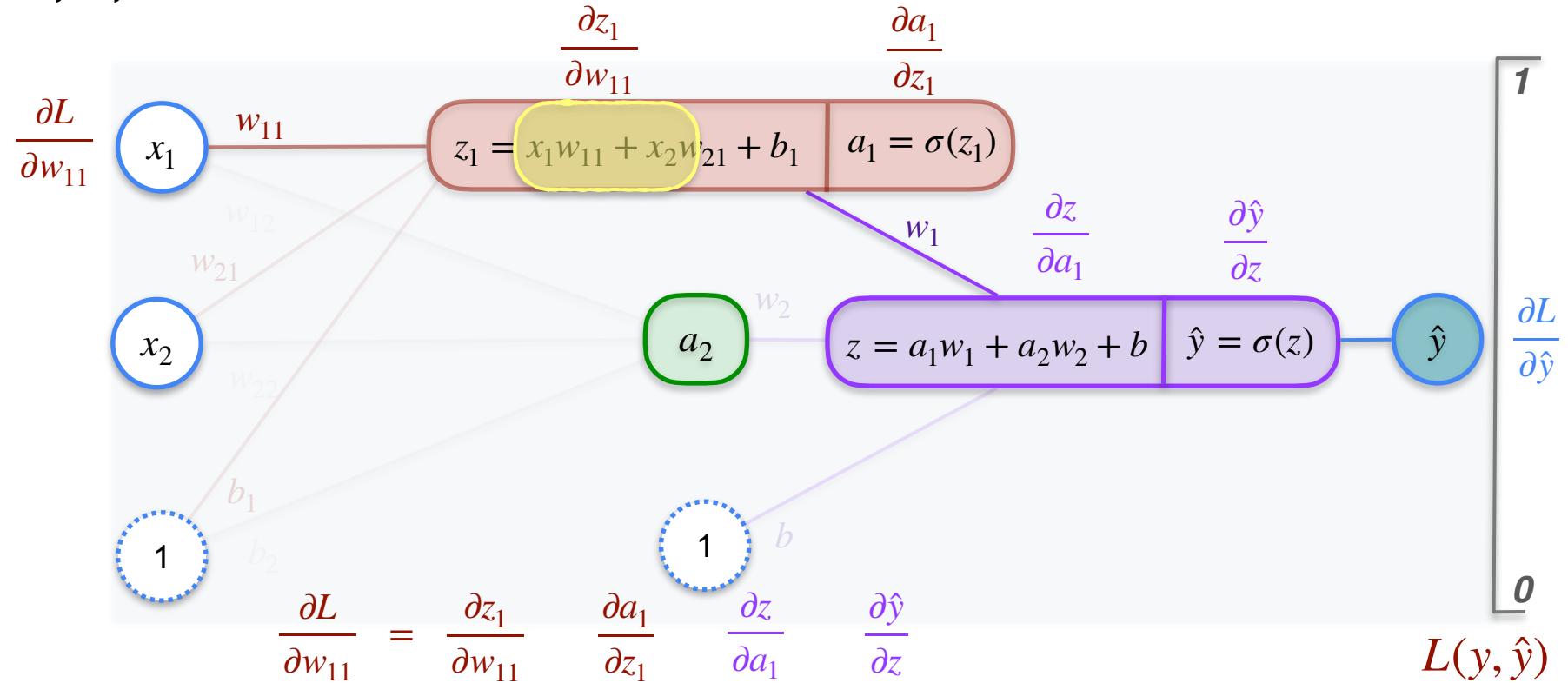
# 2,2,1 Neural Network



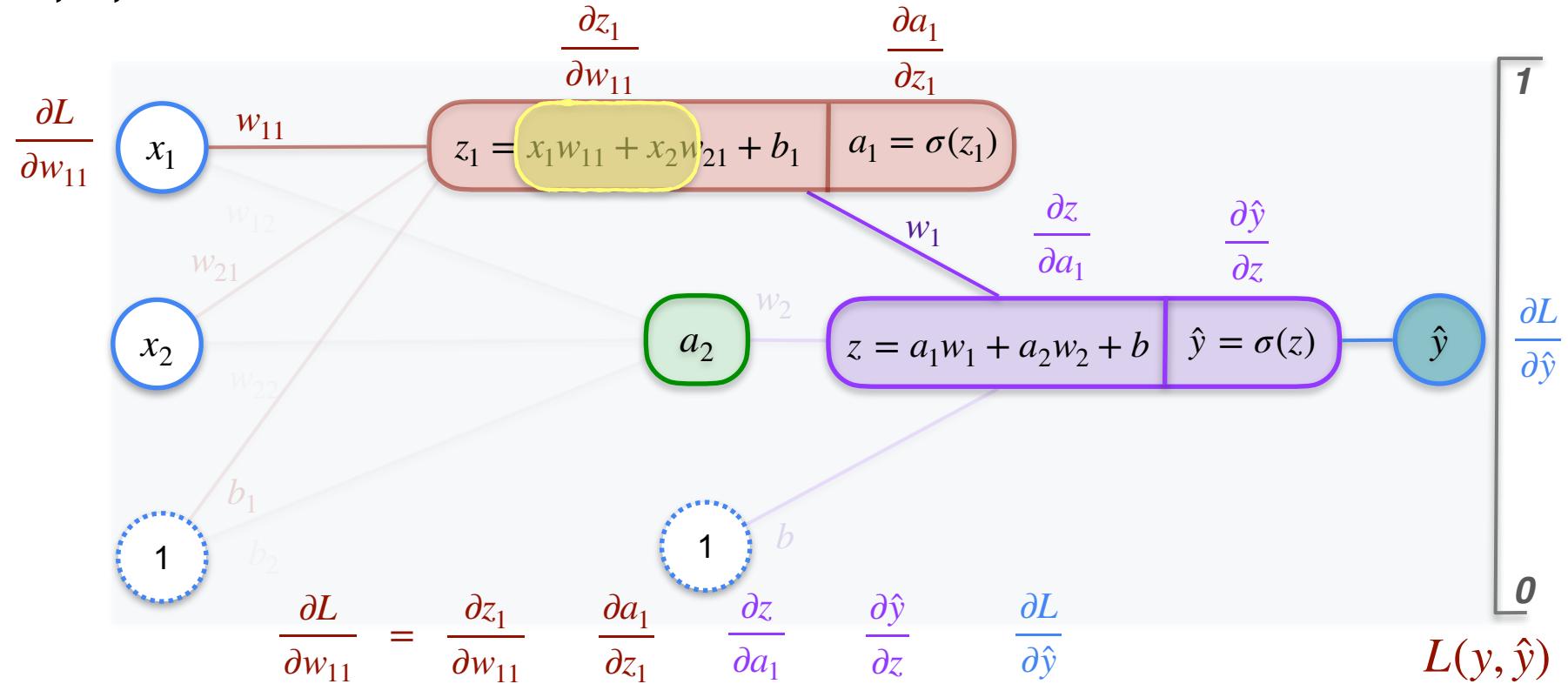
# 2,2,1 Neural Network



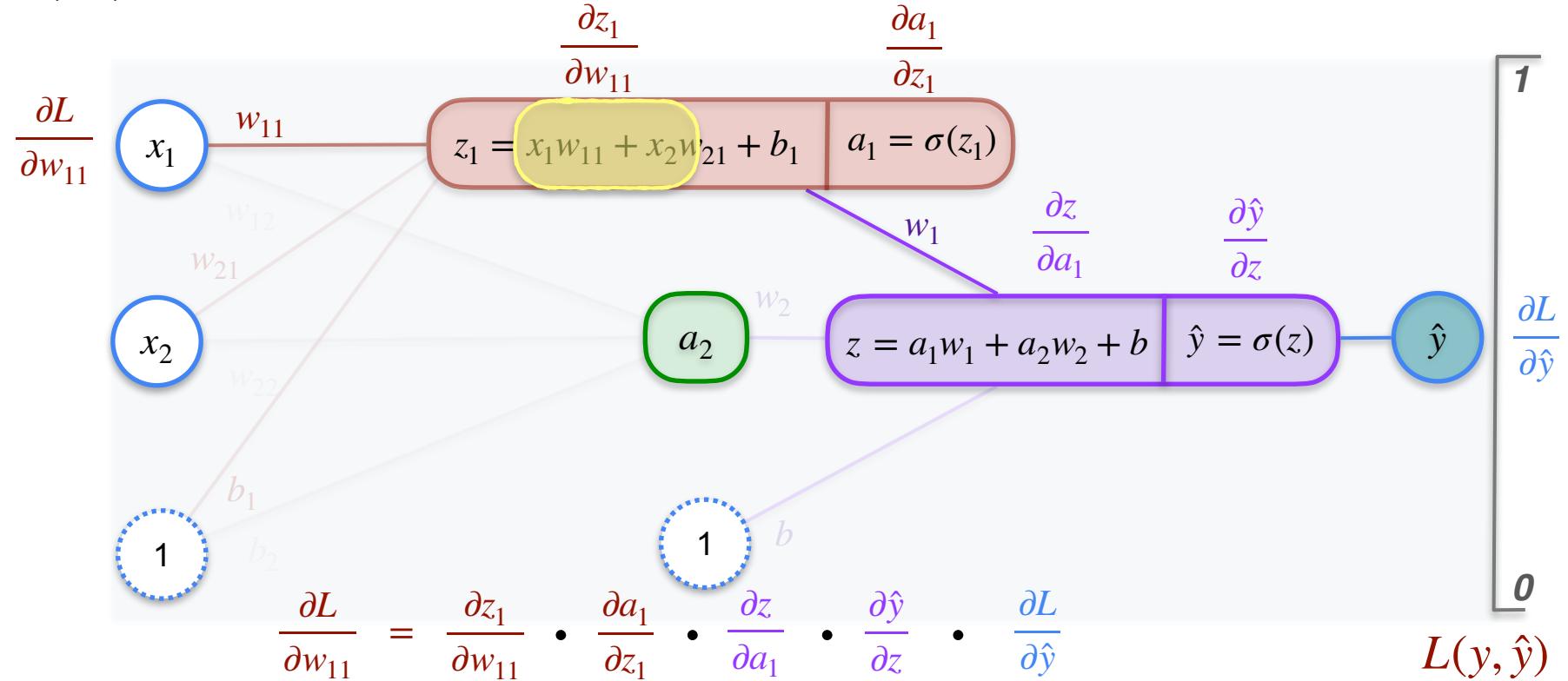
# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \bullet \frac{\partial a_1}{\partial z_1} \bullet \frac{\partial z}{\partial a_1} \bullet \frac{\partial \hat{y}}{\partial z} \bullet \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_{11}}$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \boxed{\frac{\partial z_1}{\partial w_{11}}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \boxed{\frac{\partial z_1}{\partial w_{11}}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1-a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1-a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1-a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1-a_1) w_1 - \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1-a_1) w_1 - \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = x_1 - a_1(1-a_1) w_1 - \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{11}} = x_1 \quad a_1(1-a_1) \quad w_1 \quad \hat{y}(1-\hat{y}) \quad \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{11}} = x_1 \cdot a_1(1-a_1) \cdot w_1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{11}} = x_1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{11}} = x_1 \cdot a_1 (1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{11}} &= \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{11}} &= x_1 \cdot a_1 (1 - a_1) \cdot w_1 \cdot \cancel{\hat{y}(1 - \hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1 - \hat{y})}} \\ &= -x_1 w_1 a_1 (1 - a_1) (y - \hat{y})\end{aligned}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{11}} &= \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{11}} &= x_1 \cdot a_1 (1 - a_1) \cdot w_1 \cdot \cancel{\hat{y}(1 - \hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1 - \hat{y})}} \\ &= -x_1 w_1 a_1 (1 - a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

*to find optimal  
value of  $w_{11}$  that  
gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{11}} &= \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{11}} &= x_1 \cdot a_1 (1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -x_1 w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

$$w_{11} \rightarrow w_{11} - \alpha \frac{\partial L}{\partial w_{11}}$$

*to find optimal value of  $w_{11}$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{11}} &= \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{11}} &= x_1 \cdot a_1 (1 - a_1) \cdot w_1 \cdot \cancel{\hat{y}(1 - \hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1 - \hat{y})}} \\ &= -x_1 w_1 a_1 (1 - a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

$$w_{11} \rightarrow w_{11} - \alpha$$

*to find optimal value of  $w_{11}$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

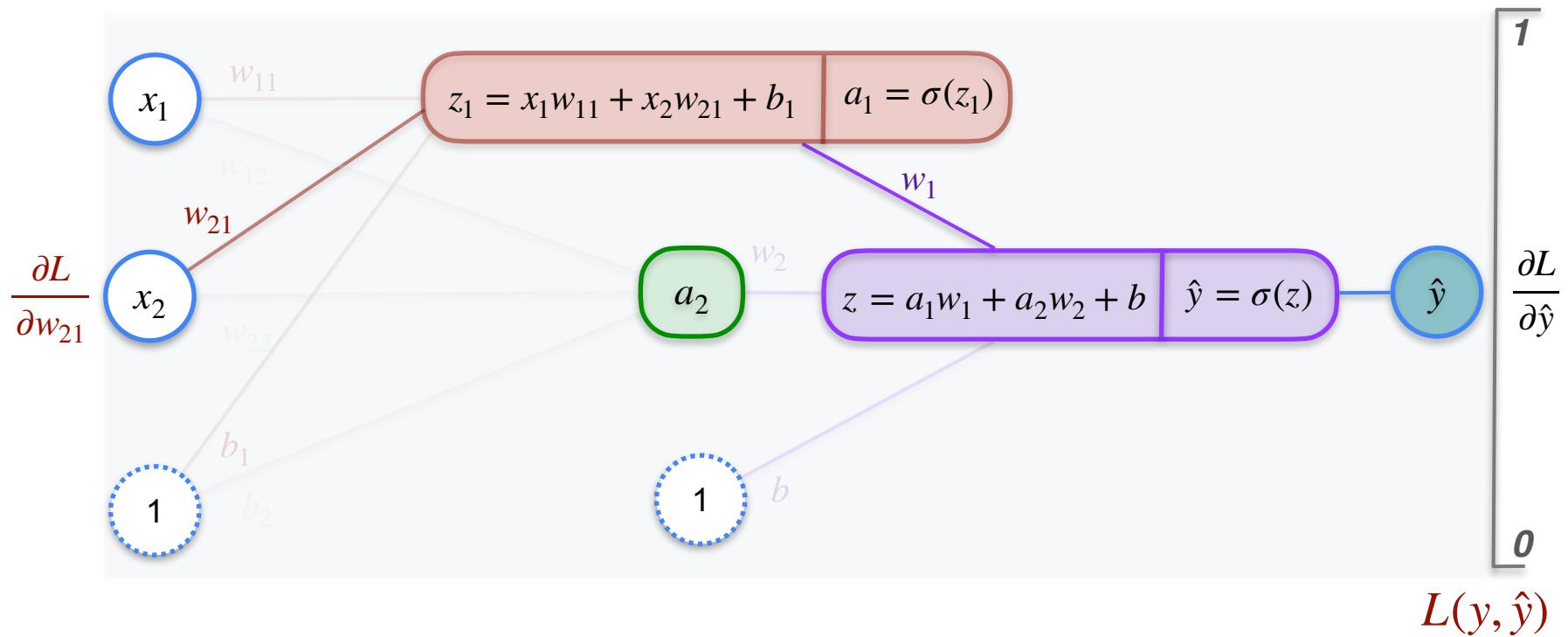
$$\begin{aligned}\frac{\partial L}{\partial w_{11}} &= \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{11}} &= x_1 \cdot a_1 (1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -x_1 w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

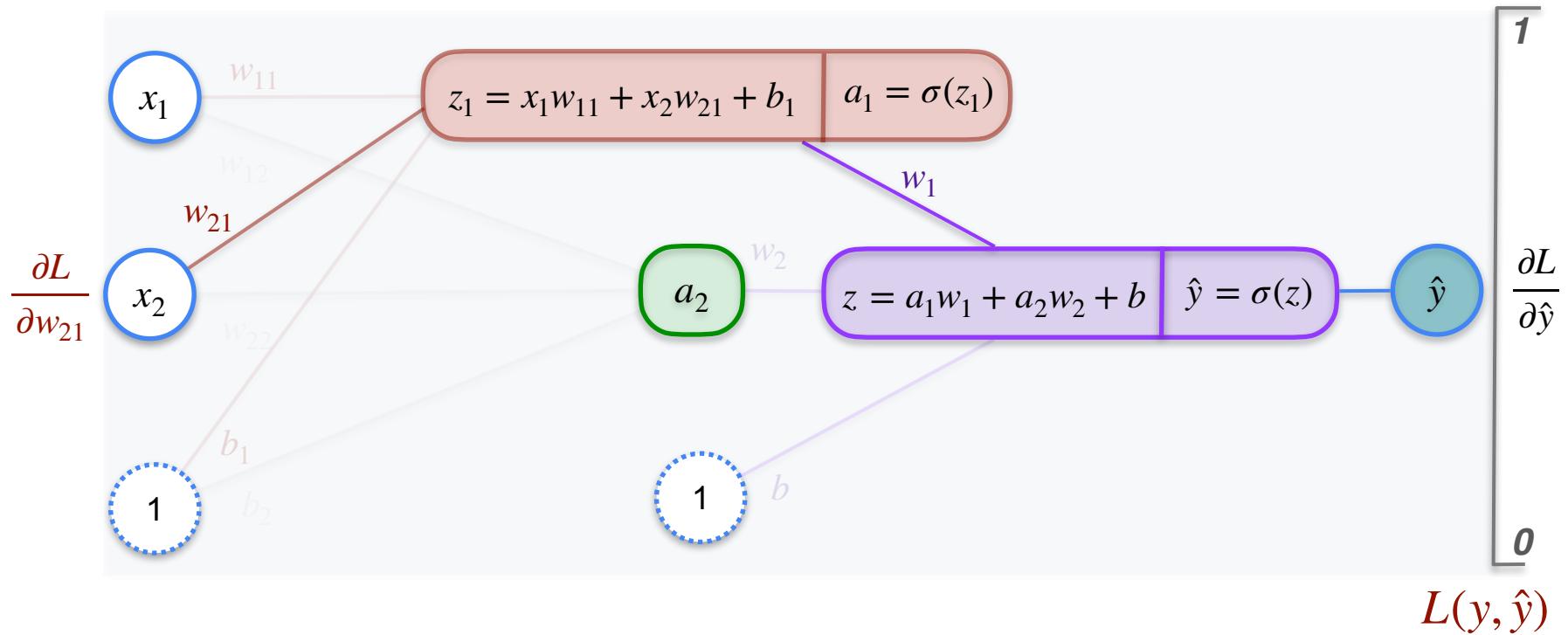
$$w_{11} \rightarrow w_{11} - \alpha \cdot x_1 w_1 a_1 (1-a_1) (y - \hat{y})$$

*to find optimal value of  $w_{11}$  that gives the least error*

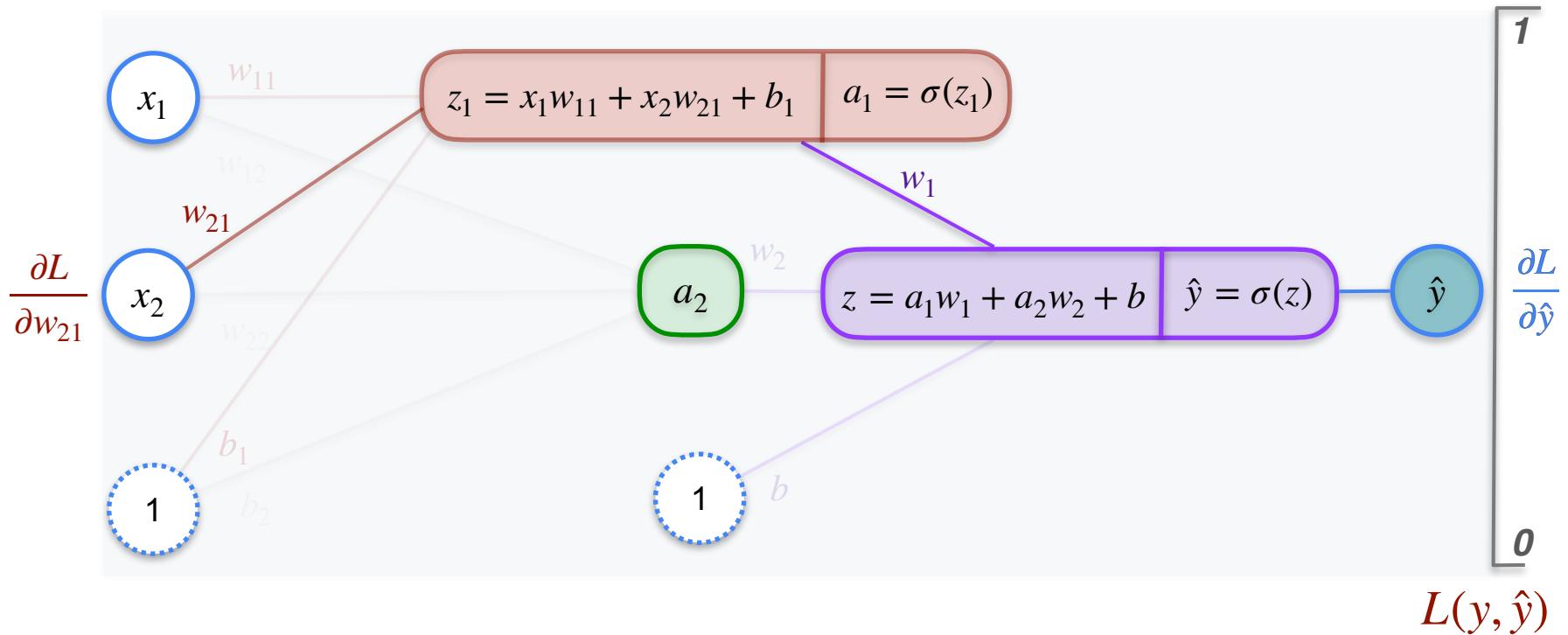
# 2,2,1 Neural Network



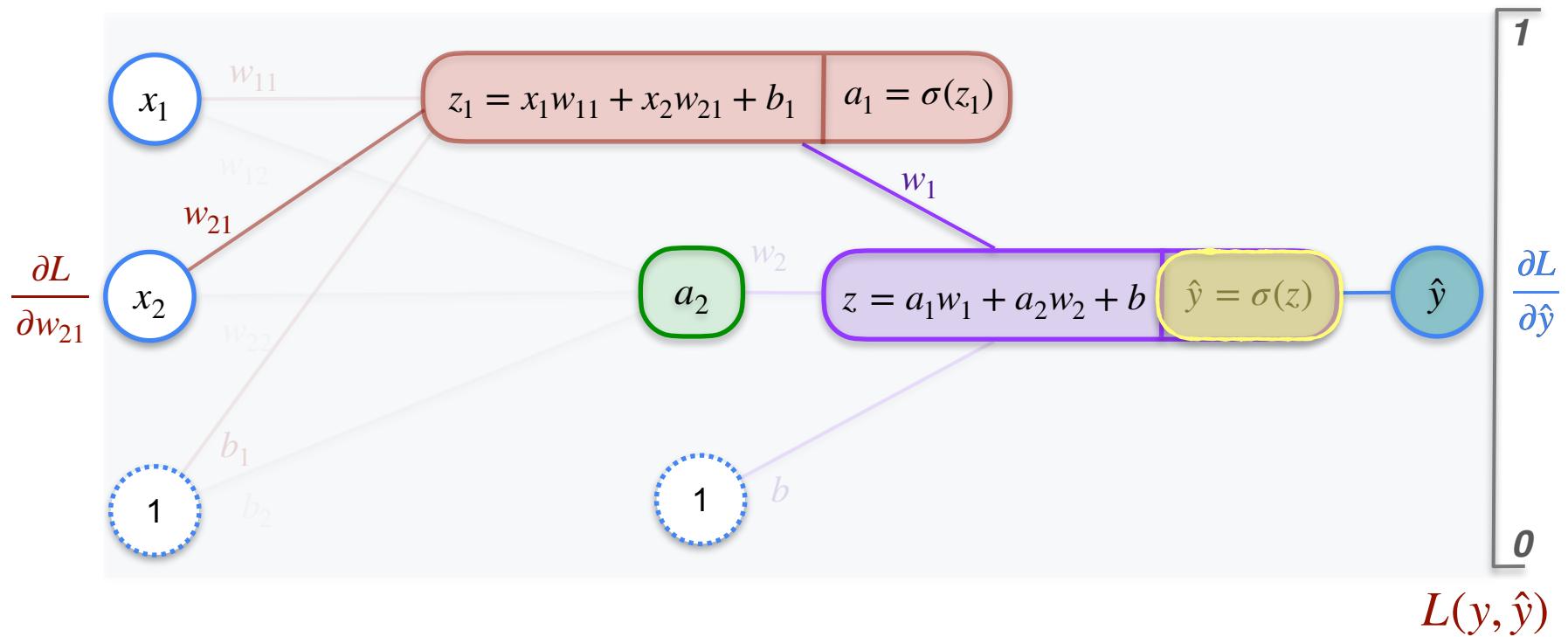
# 2,2,1 Neural Network



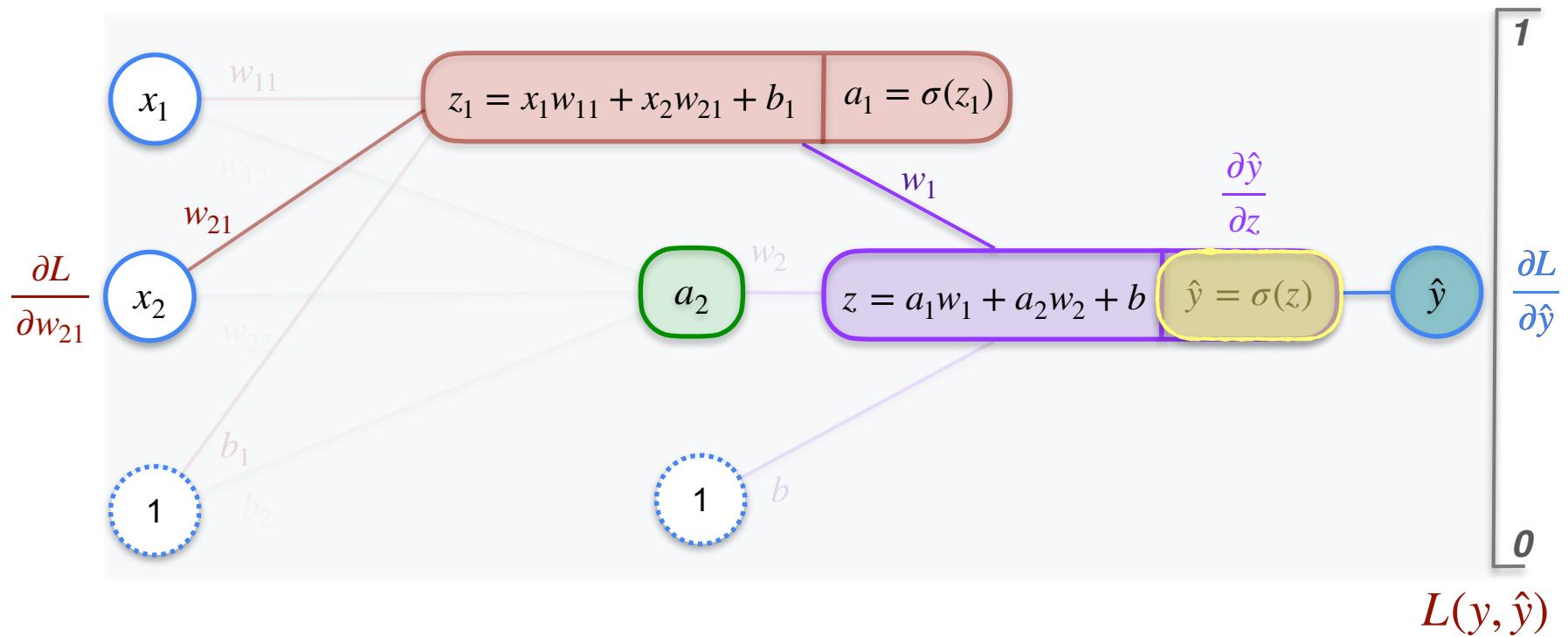
# 2,2,1 Neural Network



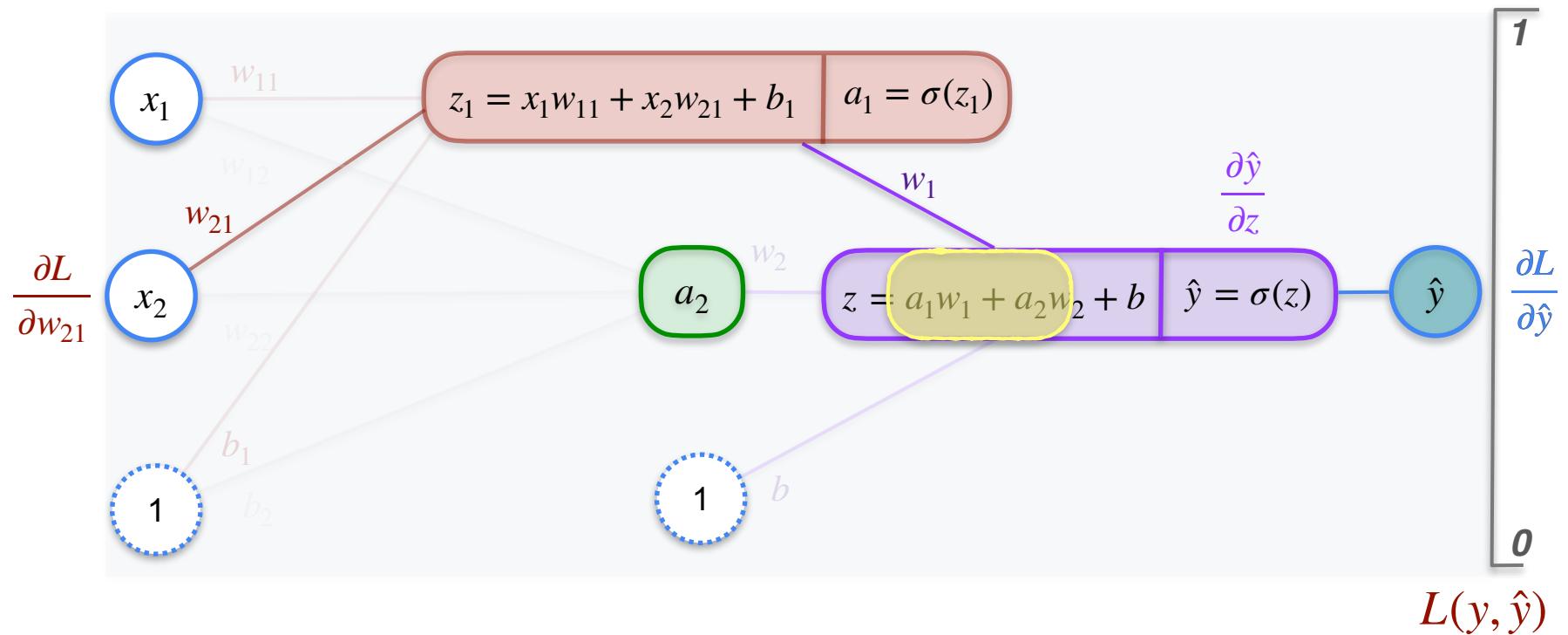
# 2,2,1 Neural Network



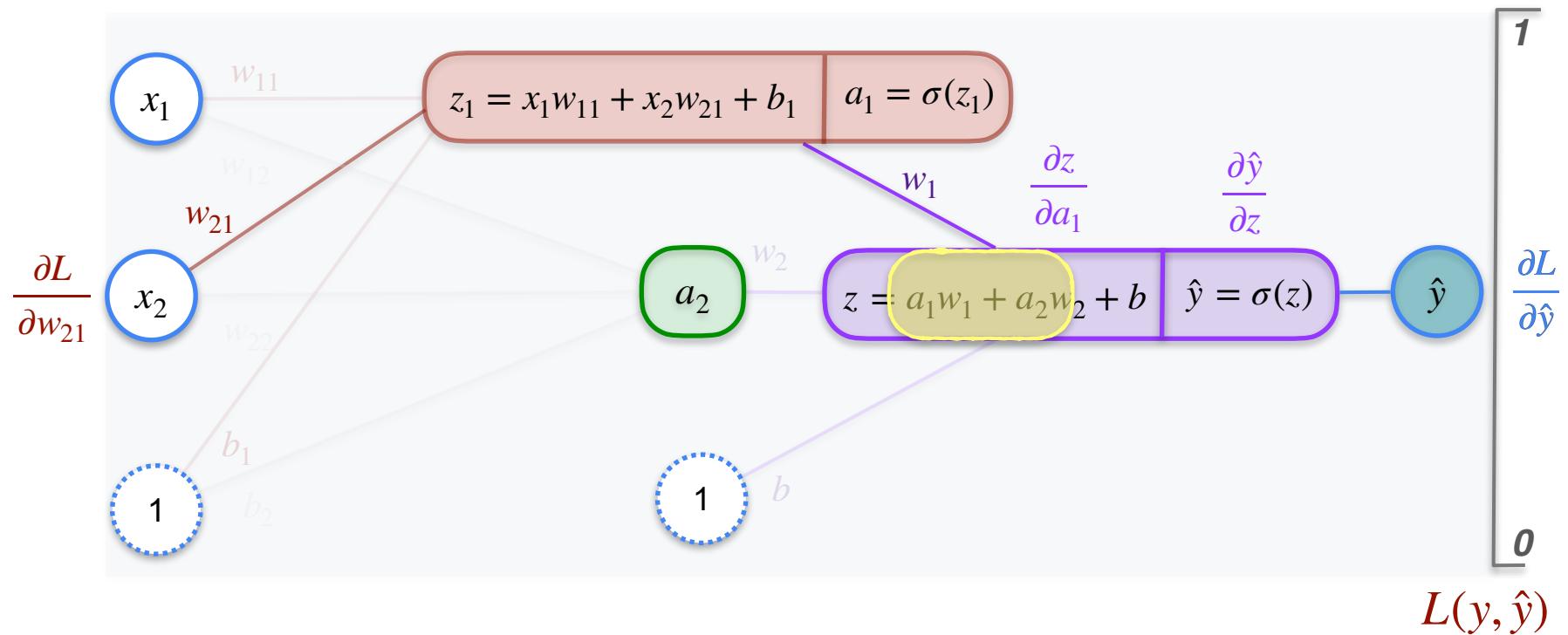
# 2,2,1 Neural Network



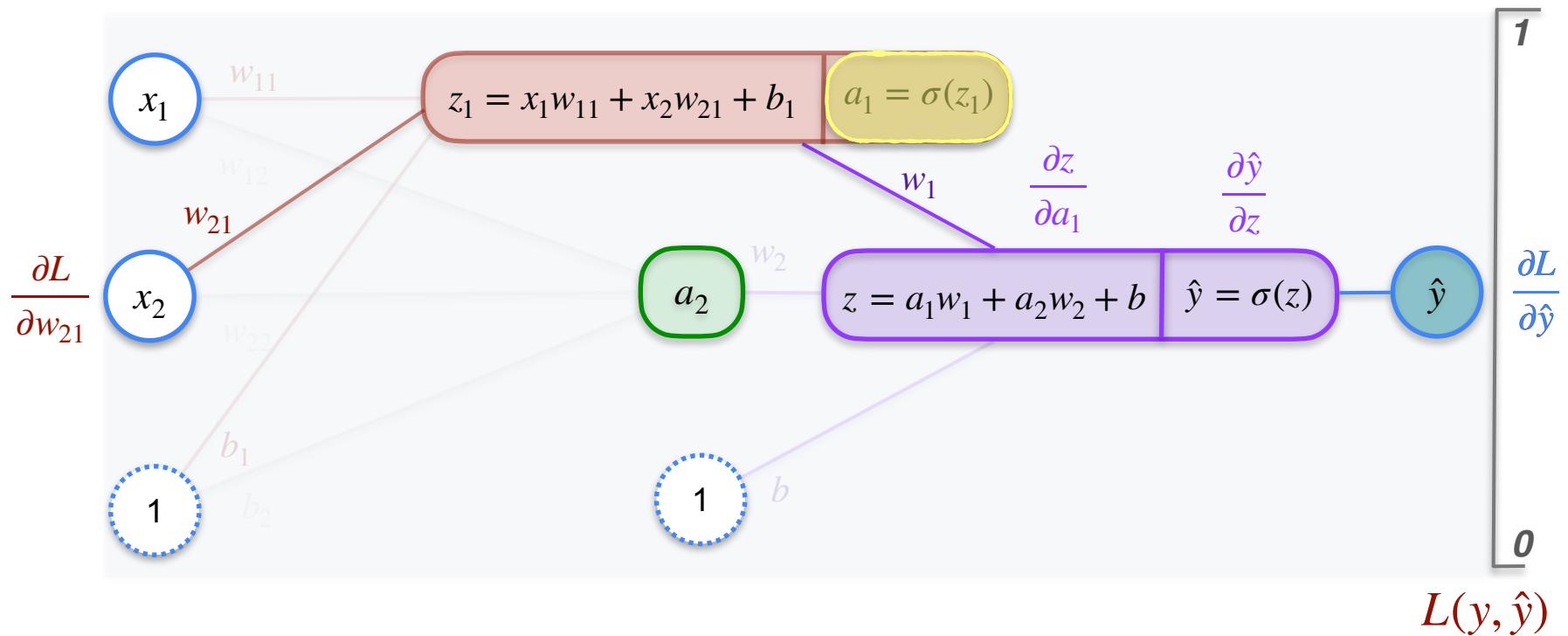
# 2,2,1 Neural Network



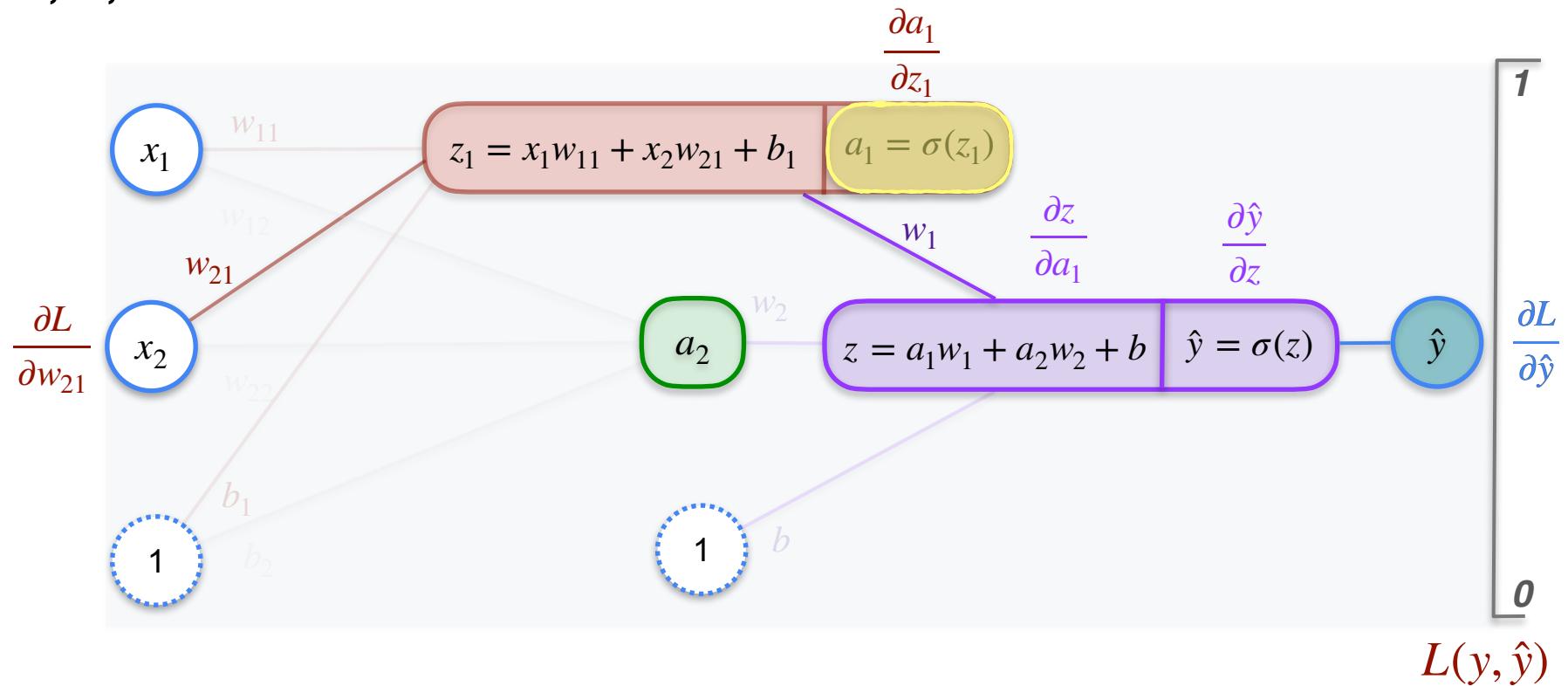
# 2,2,1 Neural Network



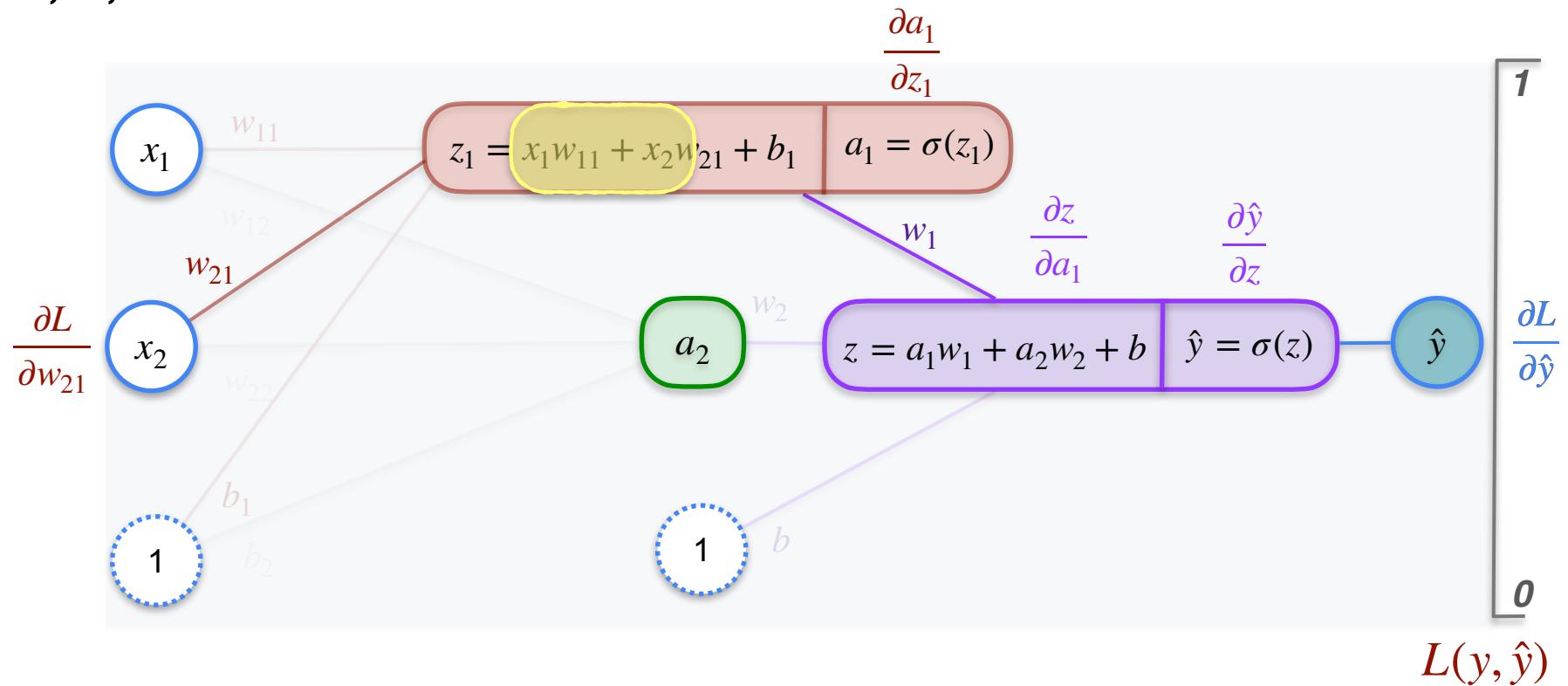
# 2,2,1 Neural Network



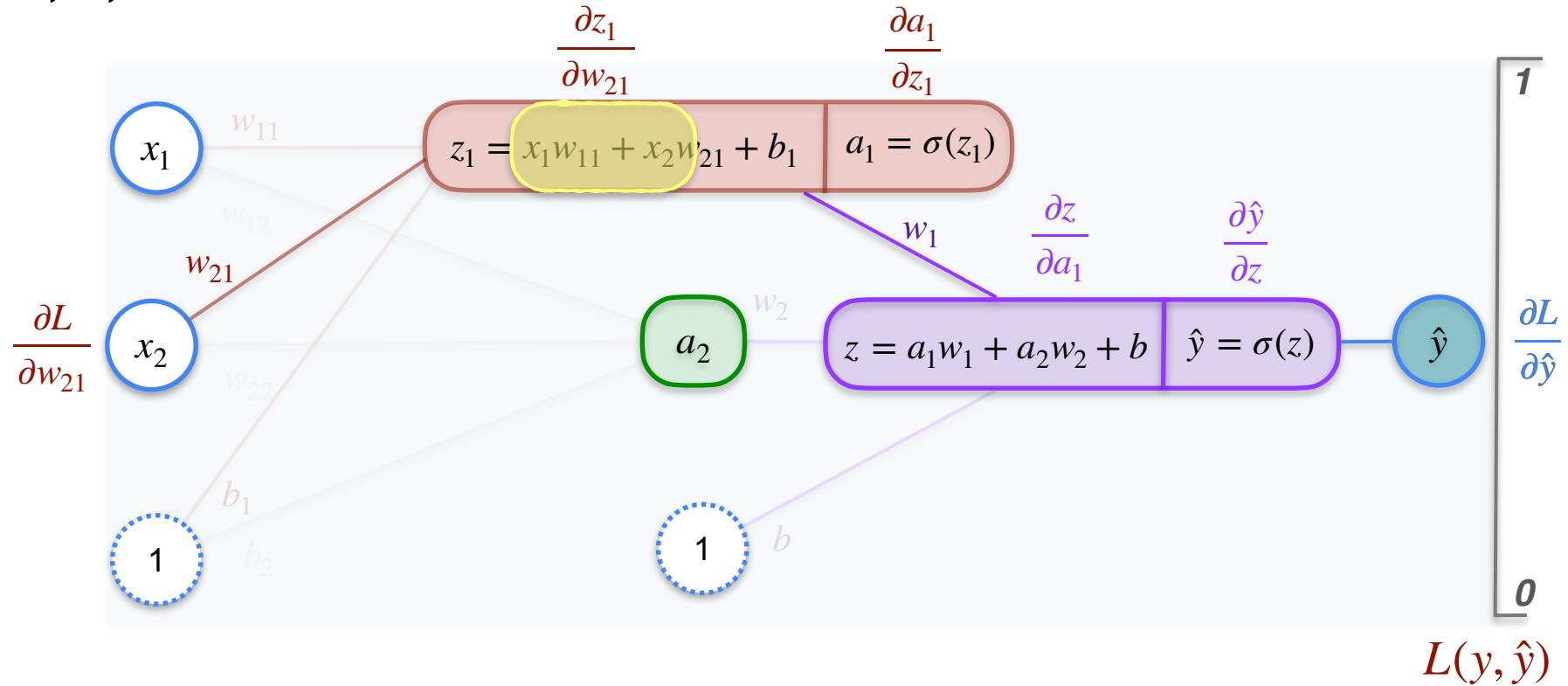
# 2,2,1 Neural Network



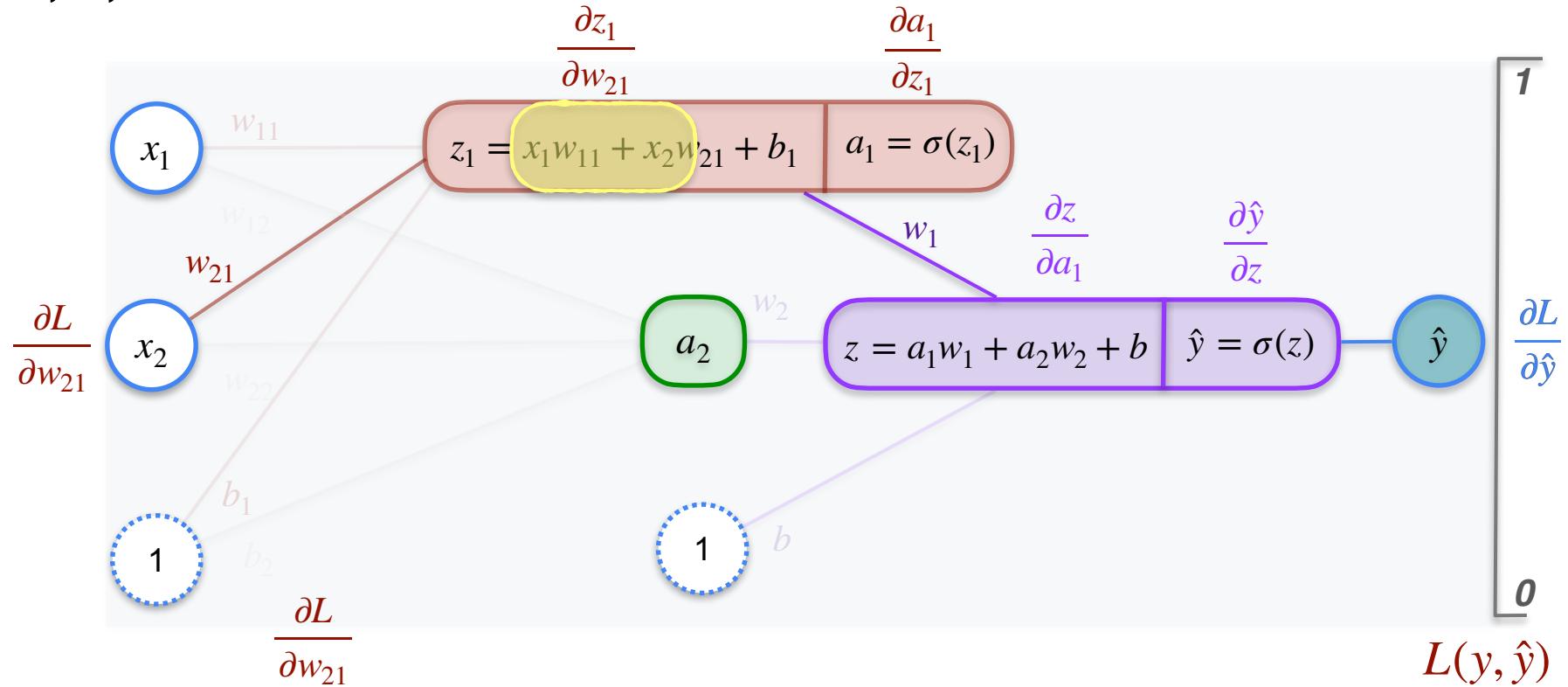
# 2,2,1 Neural Network



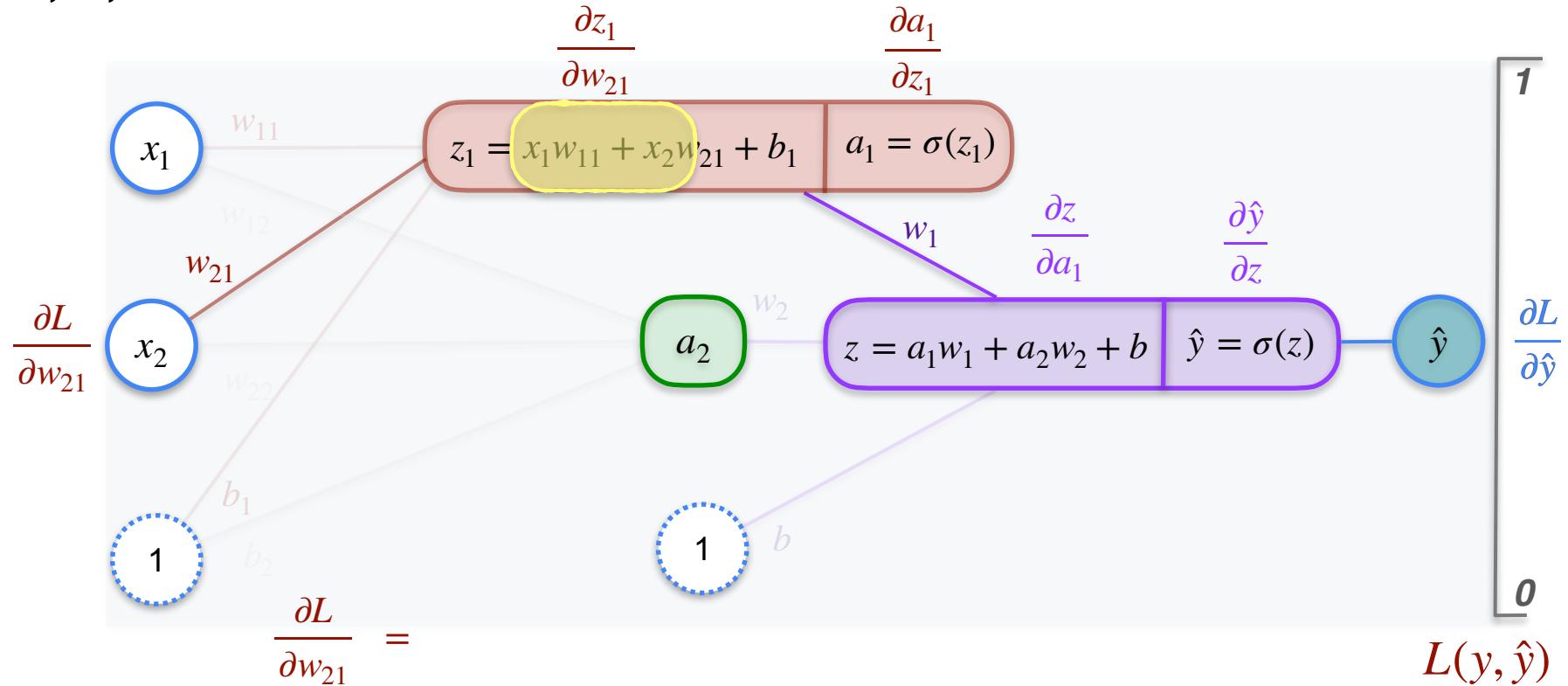
# 2,2,1 Neural Network



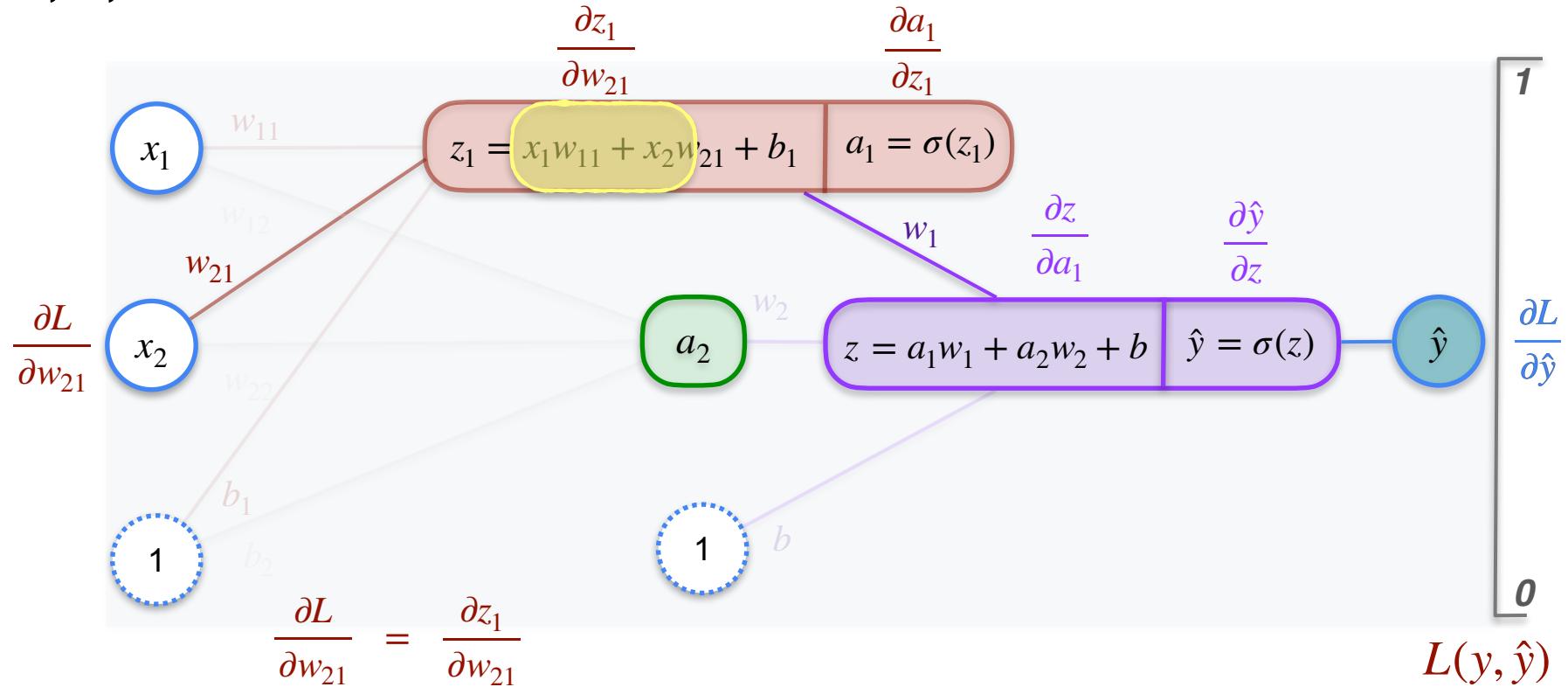
# 2,2,1 Neural Network



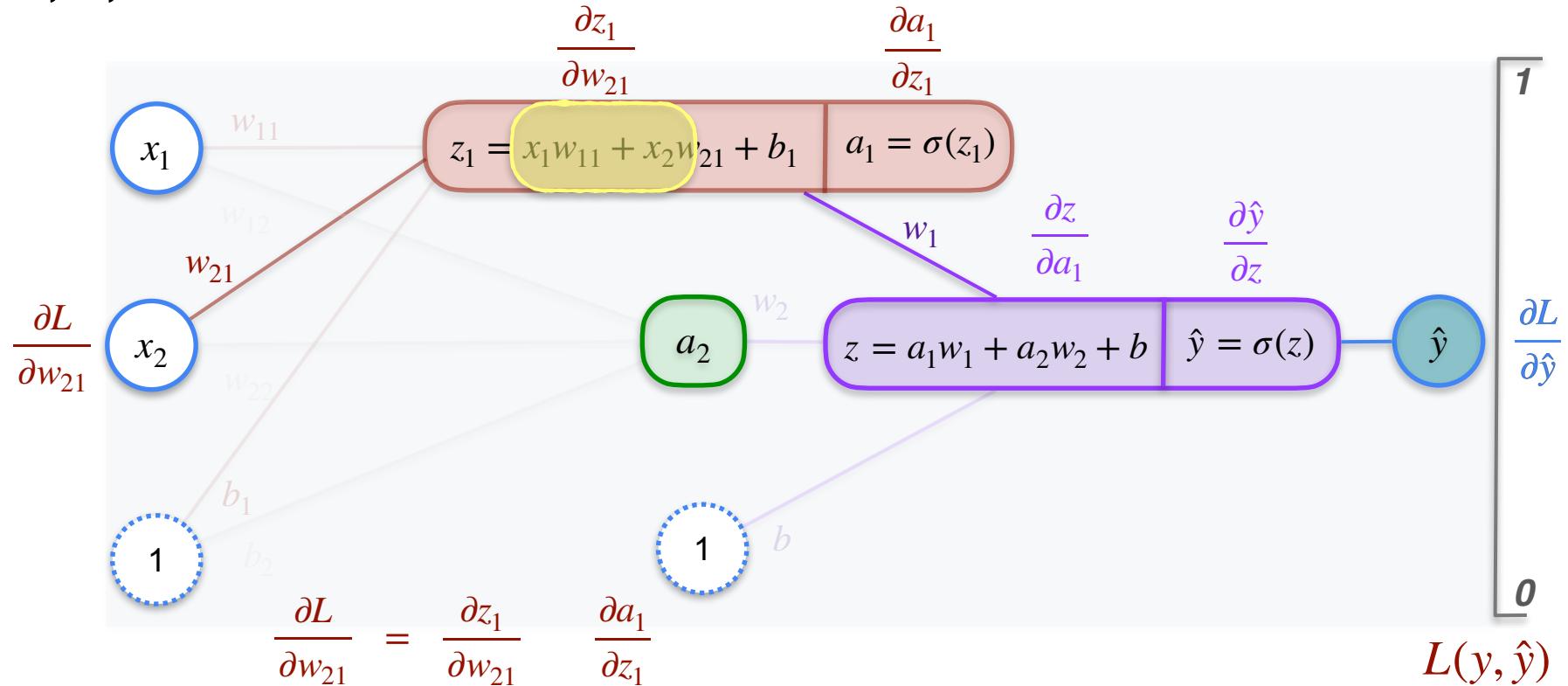
# 2,2,1 Neural Network



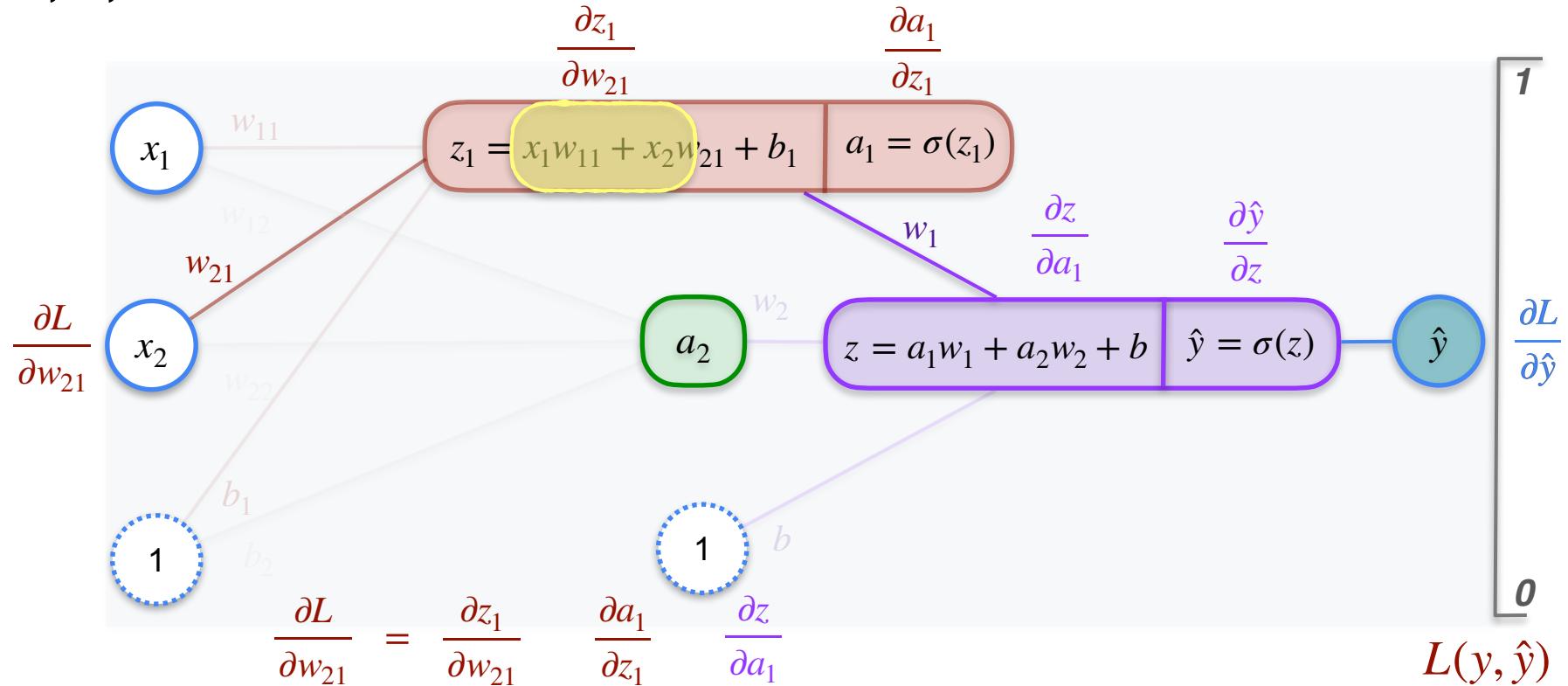
# 2,2,1 Neural Network



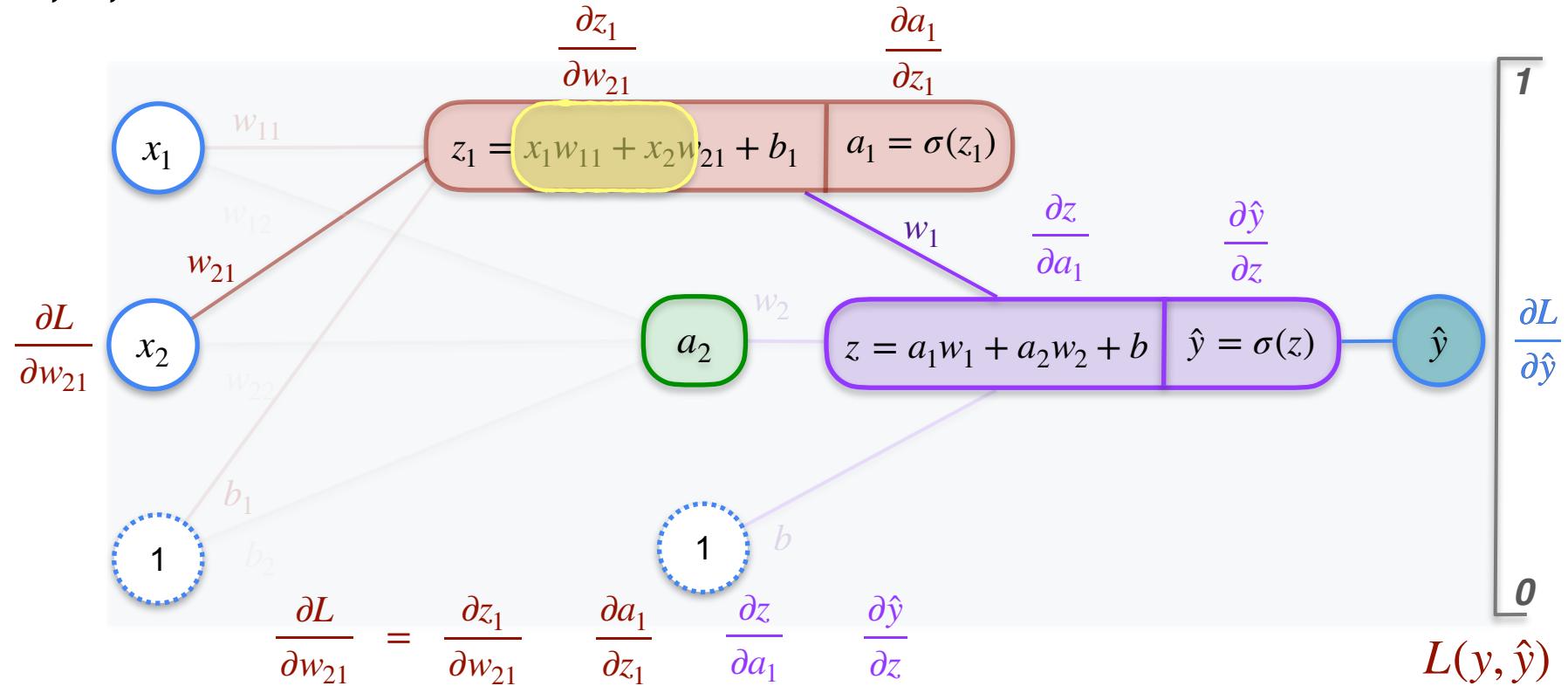
# 2,2,1 Neural Network



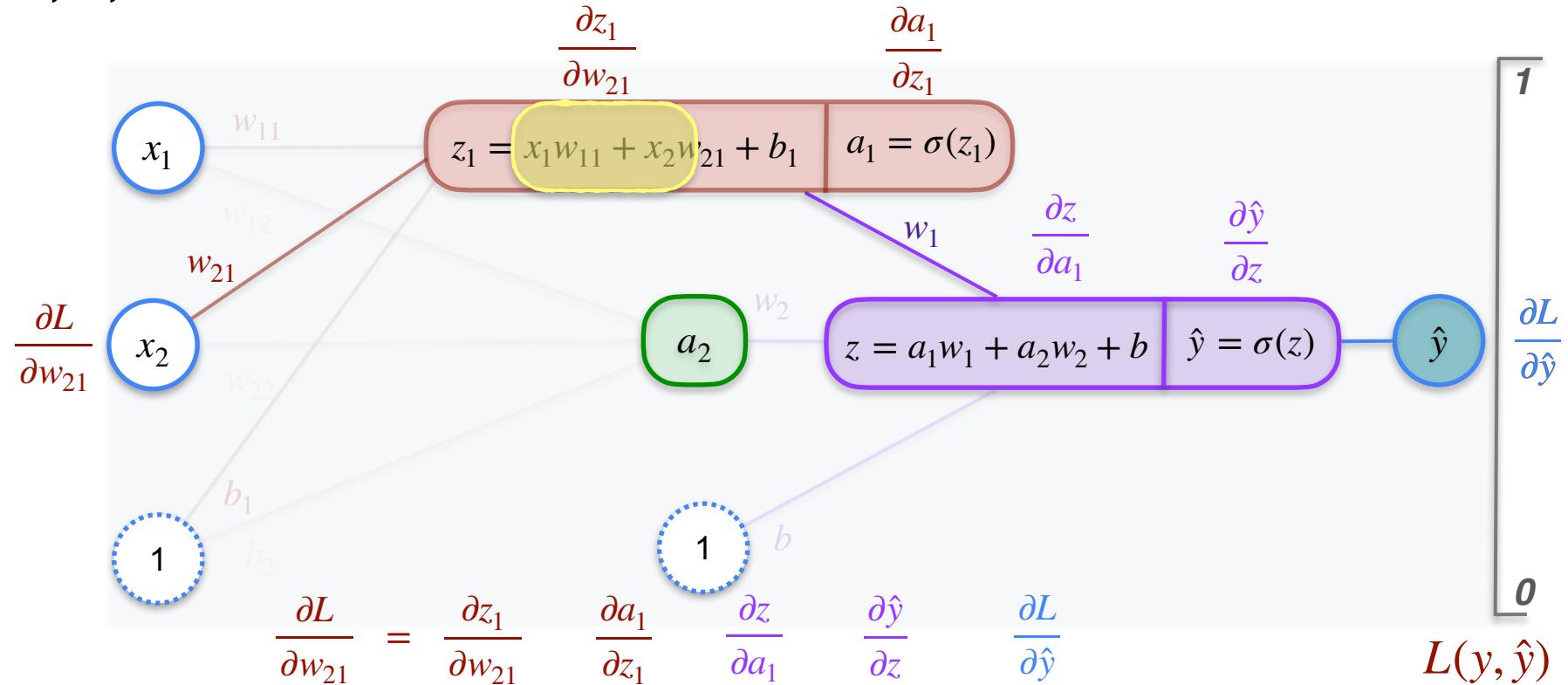
# 2,2,1 Neural Network



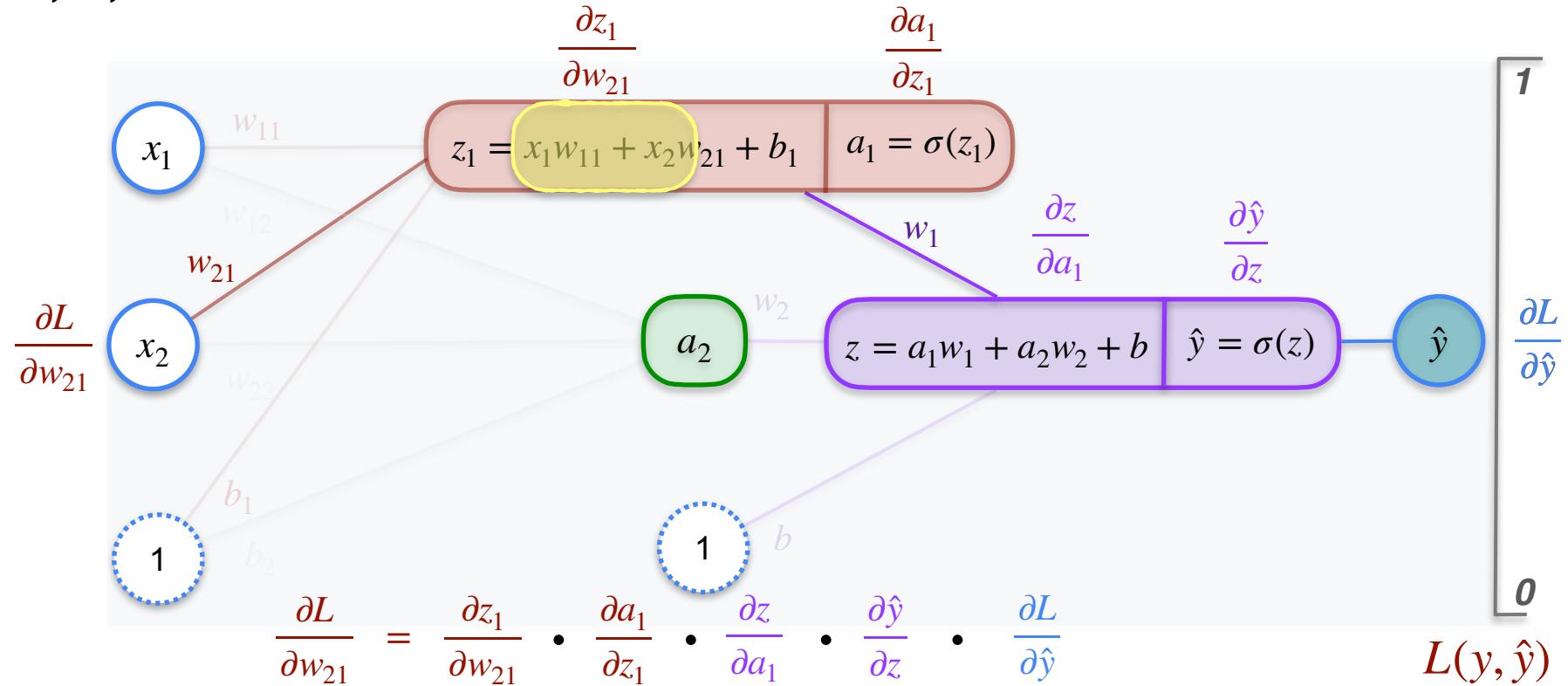
# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \bullet \frac{\partial a_1}{\partial z_1} \bullet \frac{\partial z}{\partial a_1} \bullet \frac{\partial \hat{y}}{\partial z} \bullet \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_{21}}$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \boxed{\frac{\partial z_1}{\partial w_{21}}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \boxed{\frac{\partial z_1}{\partial w_{21}}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1-a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1-a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1-a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1-a_1) w_1 - \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \boxed{\frac{\partial L}{\partial \hat{y}}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 \quad a_1(1-a_1) \quad w_1 \quad \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = x_2 - a_1(1-a_1) w_1 - \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{21}} = x_2 \quad a_1(1-a_1) \quad w_1 \quad \hat{y}(1-\hat{y}) \quad \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{21}} = x_2 \cdot a_1(1-a_1) \cdot w_1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{21}} = x_2 \cdot a_1 (1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial w_{21}} = x_2 \cdot a_1 (1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{21}} &= \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{21}} &= x_2 \cdot a_1 (1 - a_1) \cdot w_1 \cdot \cancel{\hat{y}(1 - \hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1 - \hat{y})}} \\ &= -x_2 w_1 a_1 (1 - a_1) (y - \hat{y})\end{aligned}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{21}} &= \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{21}} &= x_2 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -x_2 w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

*to find optimal  
value of  $w_{21}$  that  
gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial w_{21}} &= \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{21}} &= x_2 \cdot a_1 (1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -x_2 w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

**Perform gradient descent with**

$$w_{21} \rightarrow w_{21} - \alpha \frac{\partial L}{\partial w_{21}}$$

**to find optimal value of  $w_{21}$  that gives the least error**

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\begin{aligned} \frac{\partial L}{\partial w_{21}} &= x_2 \cdot a_1 (1 - a_1) \cdot w_1 \cdot \cancel{\hat{y}(1 - \hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1 - \hat{y})}} \\ &= -x_2 w_1 a_1 (1 - a_1) (y - \hat{y}) \end{aligned}$$

*Perform gradient descent with*

$$w_{21} \rightarrow w_{21} - \alpha$$

*to find optimal value of  $w_{21}$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

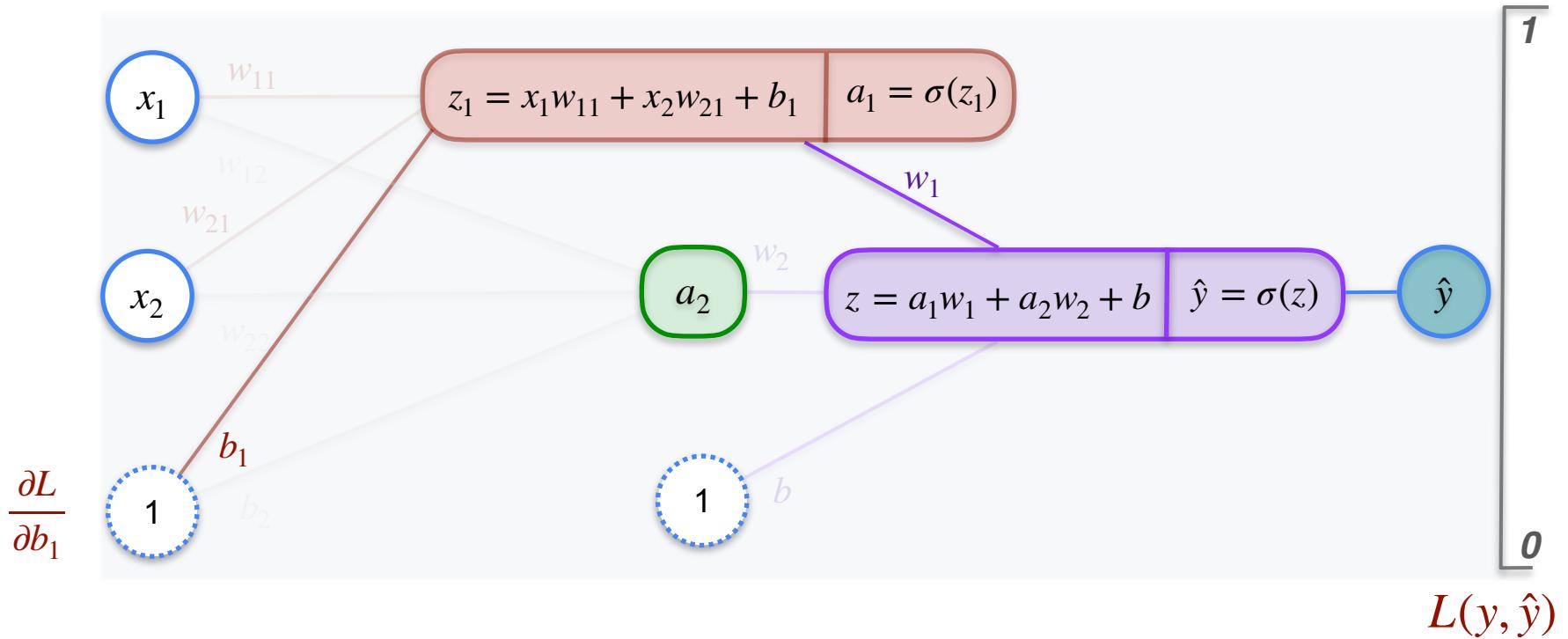
$$\begin{aligned}\frac{\partial L}{\partial w_{21}} &= \frac{\partial z_1}{\partial w_{21}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_{21}} &= x_2 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -x_2 w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

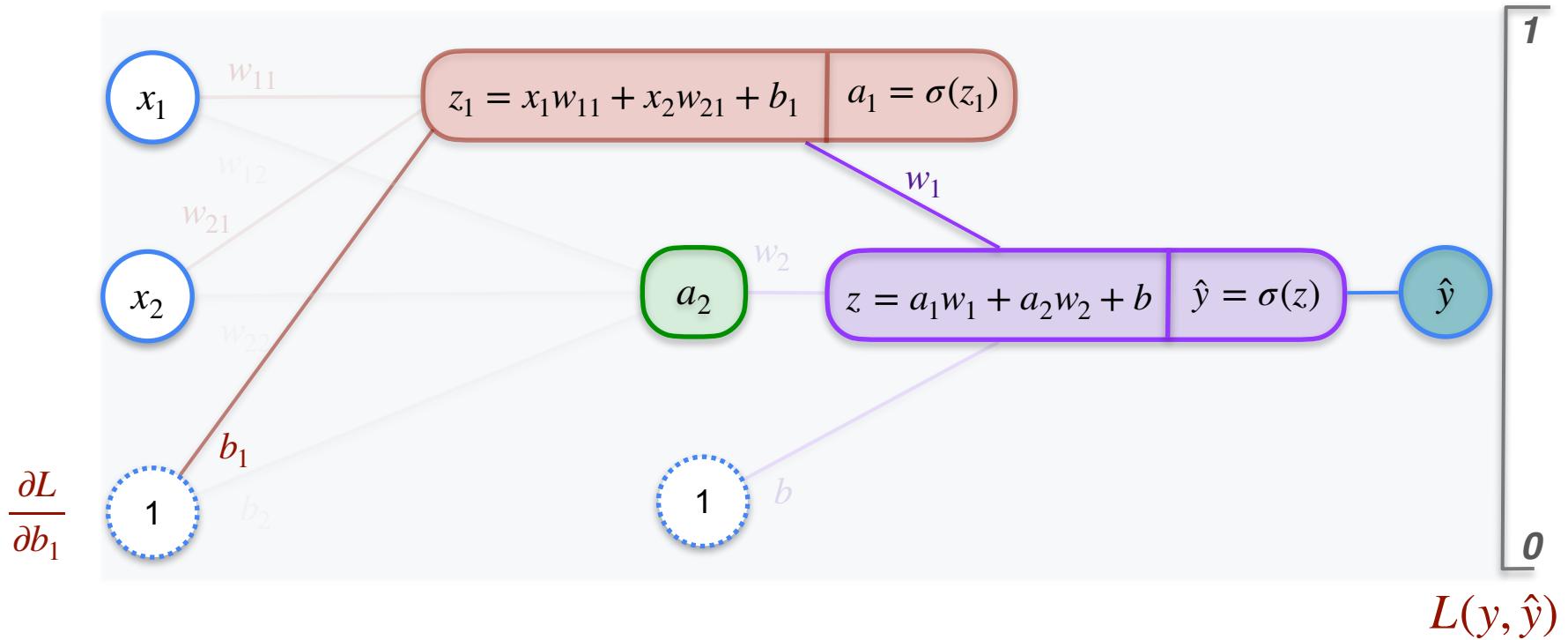
$$w_{21} \rightarrow w_{21} - \alpha \cdot x_2 w_1 a_1 (1-a_1) (y - \hat{y})$$

*to find optimal value of  $w_{21}$  that gives the least error*

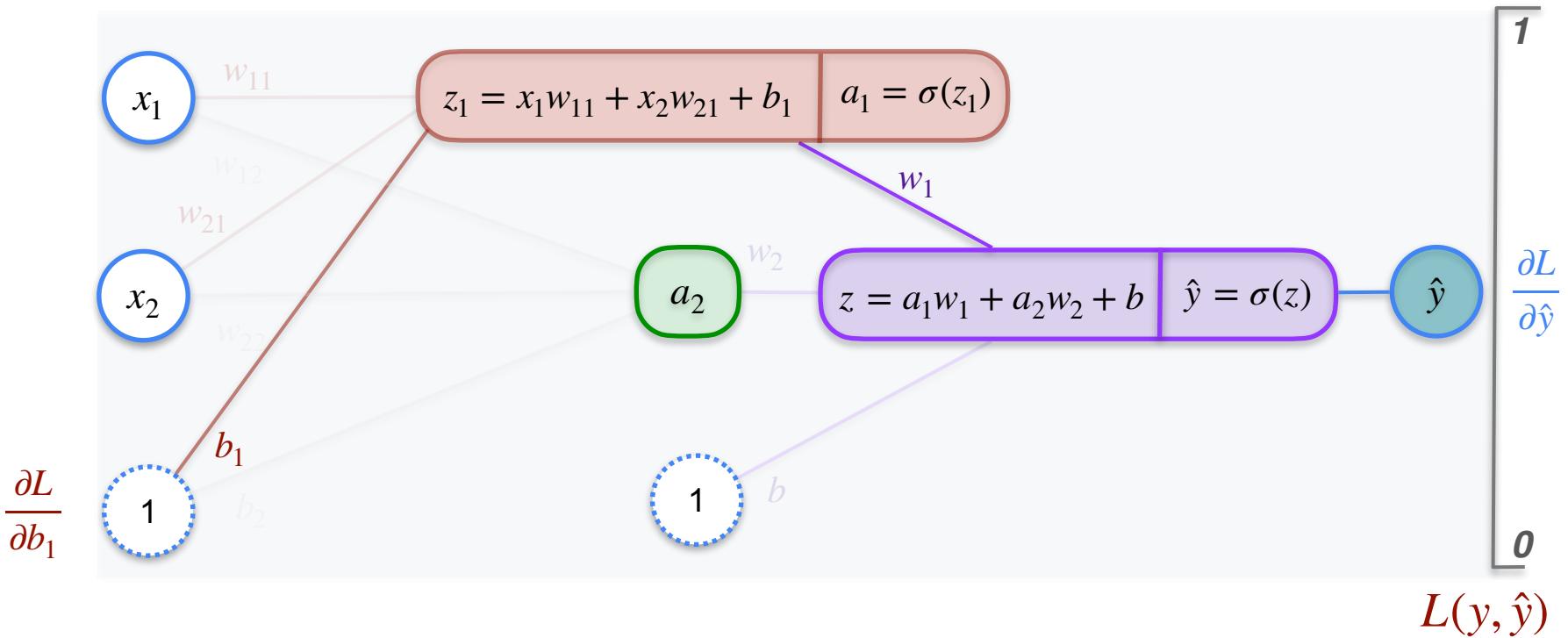
# 2,2,1 Neural Network



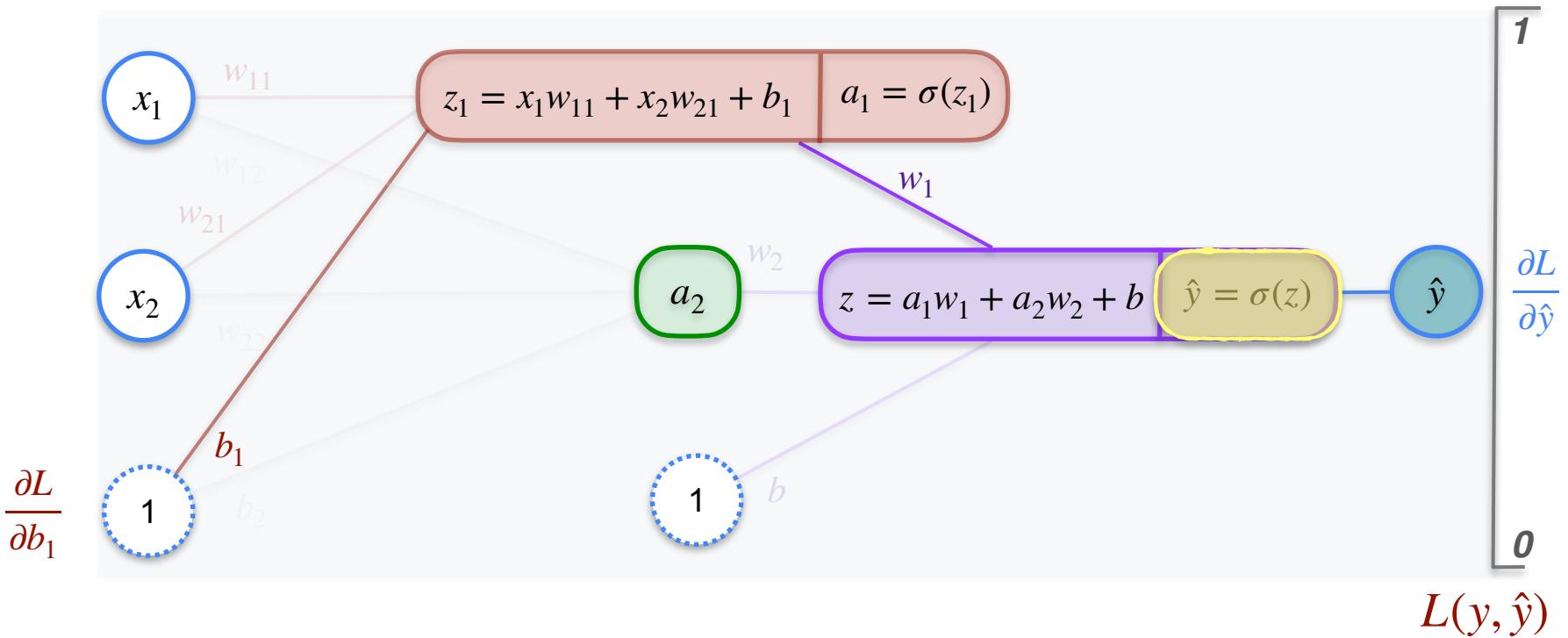
# 2,2,1 Neural Network



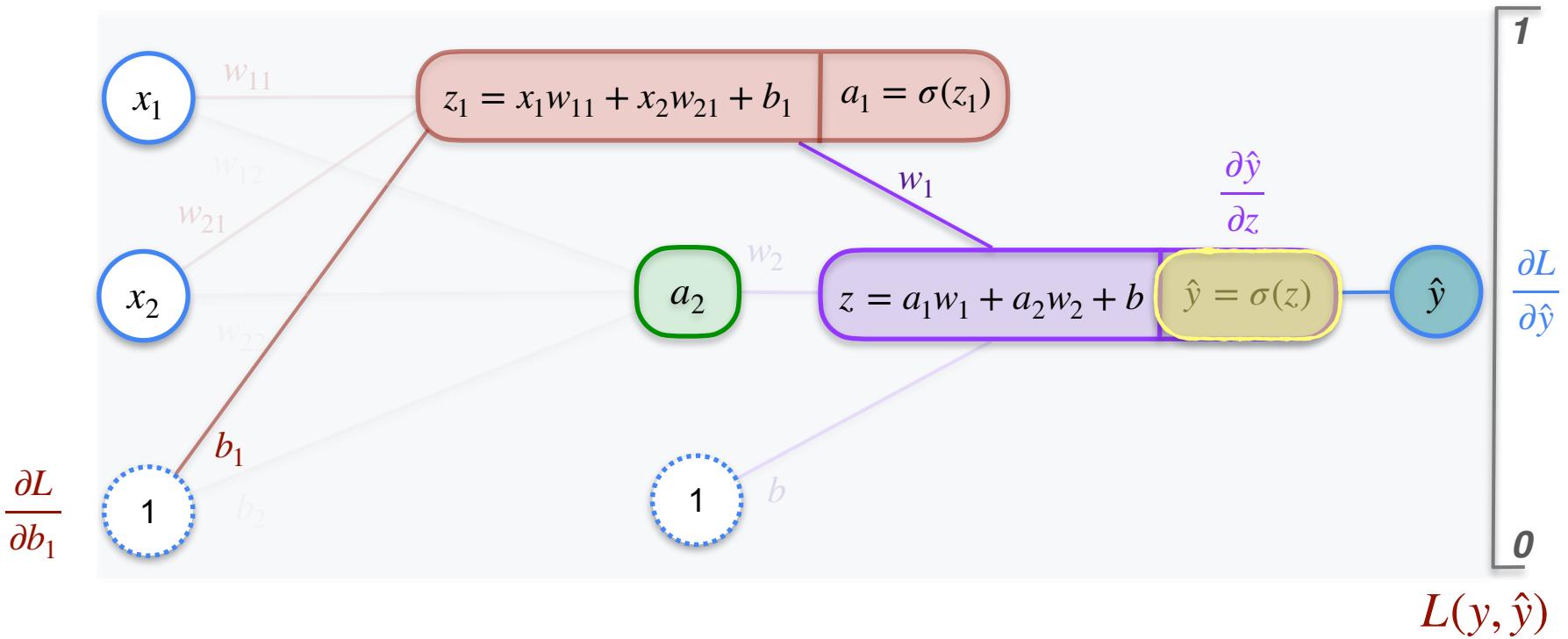
# 2,2,1 Neural Network



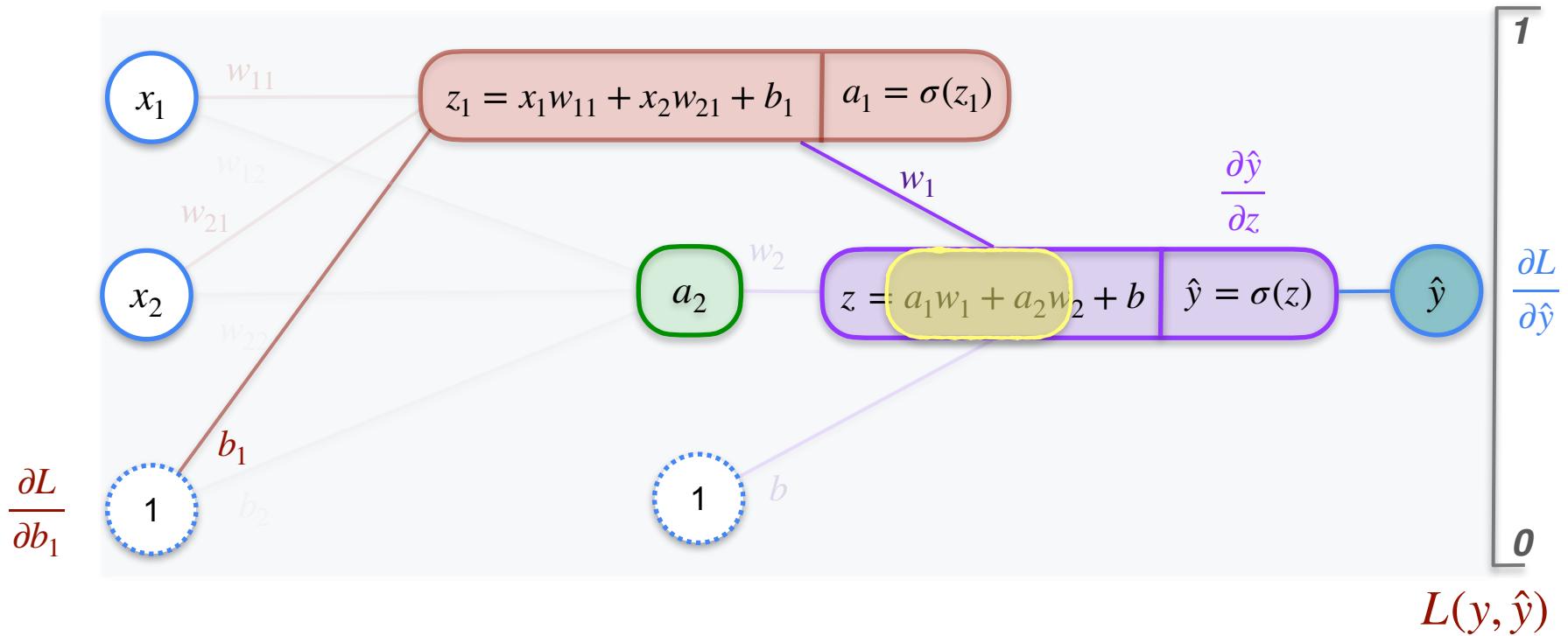
# 2,2,1 Neural Network



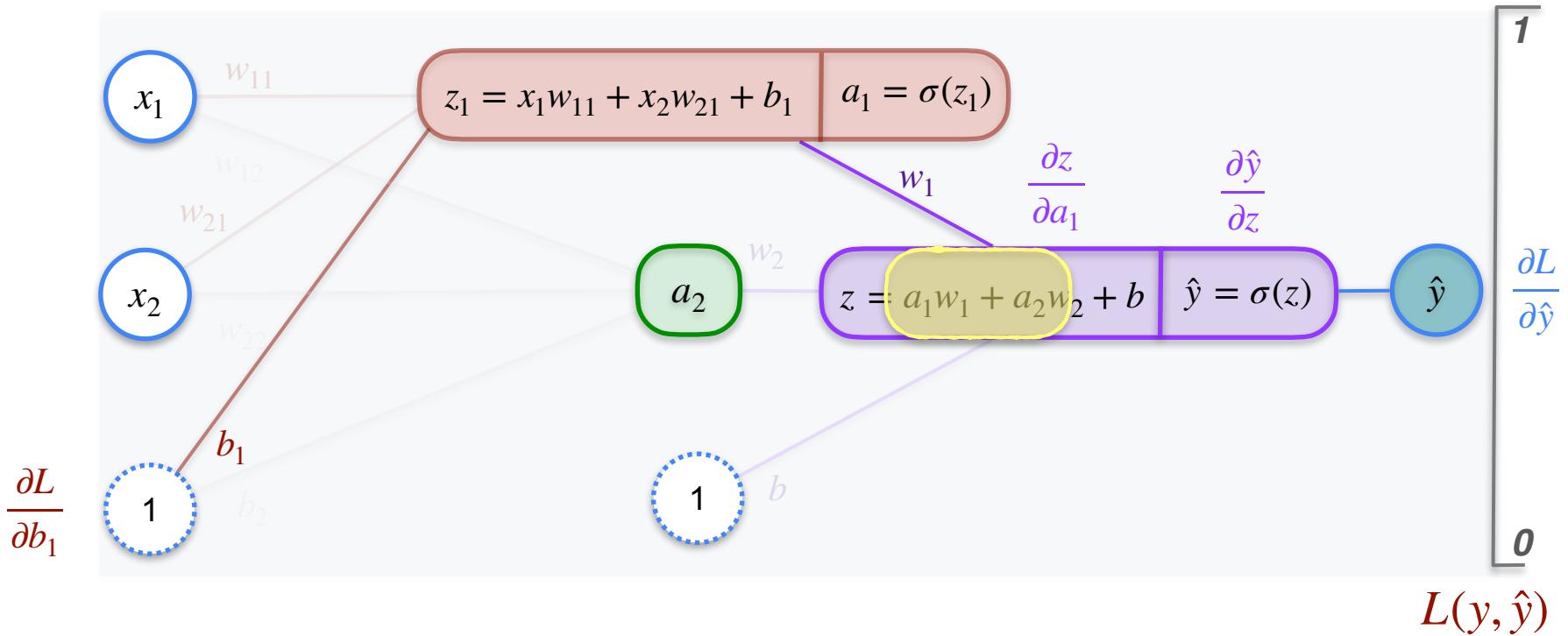
# 2,2,1 Neural Network



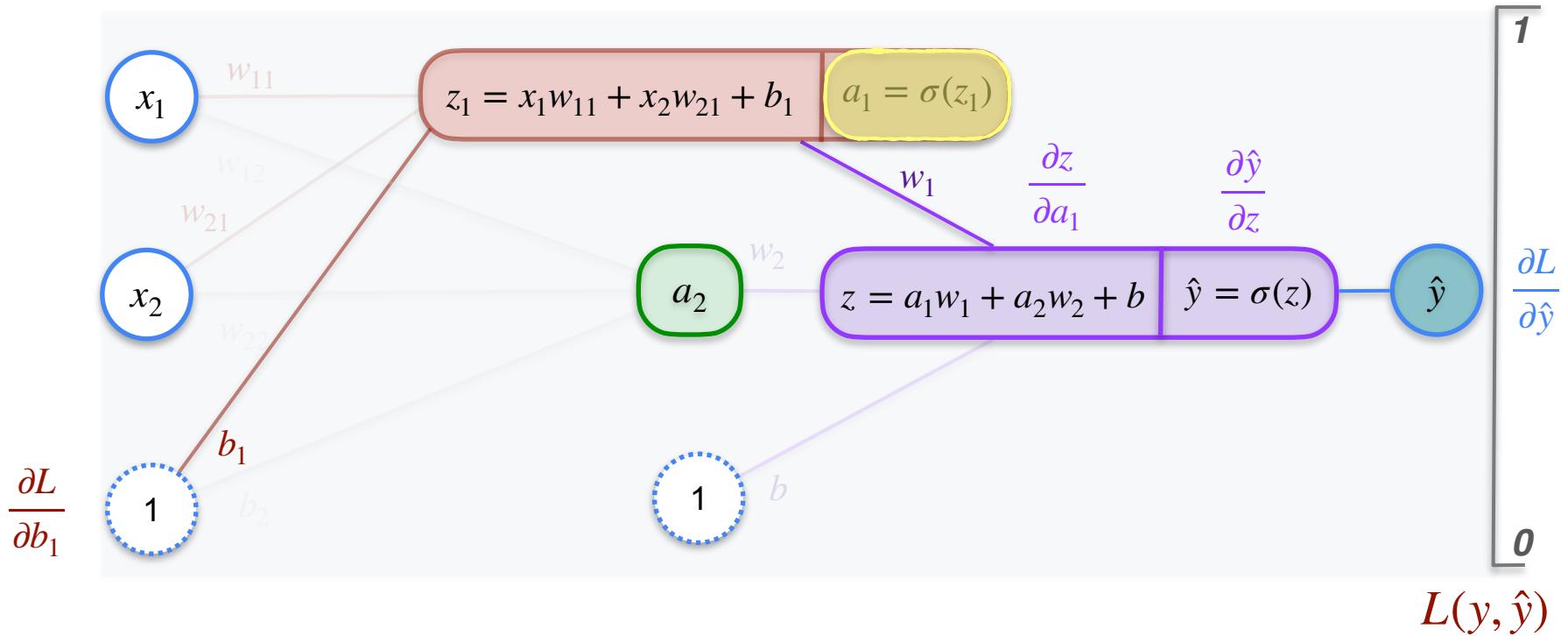
# 2,2,1 Neural Network



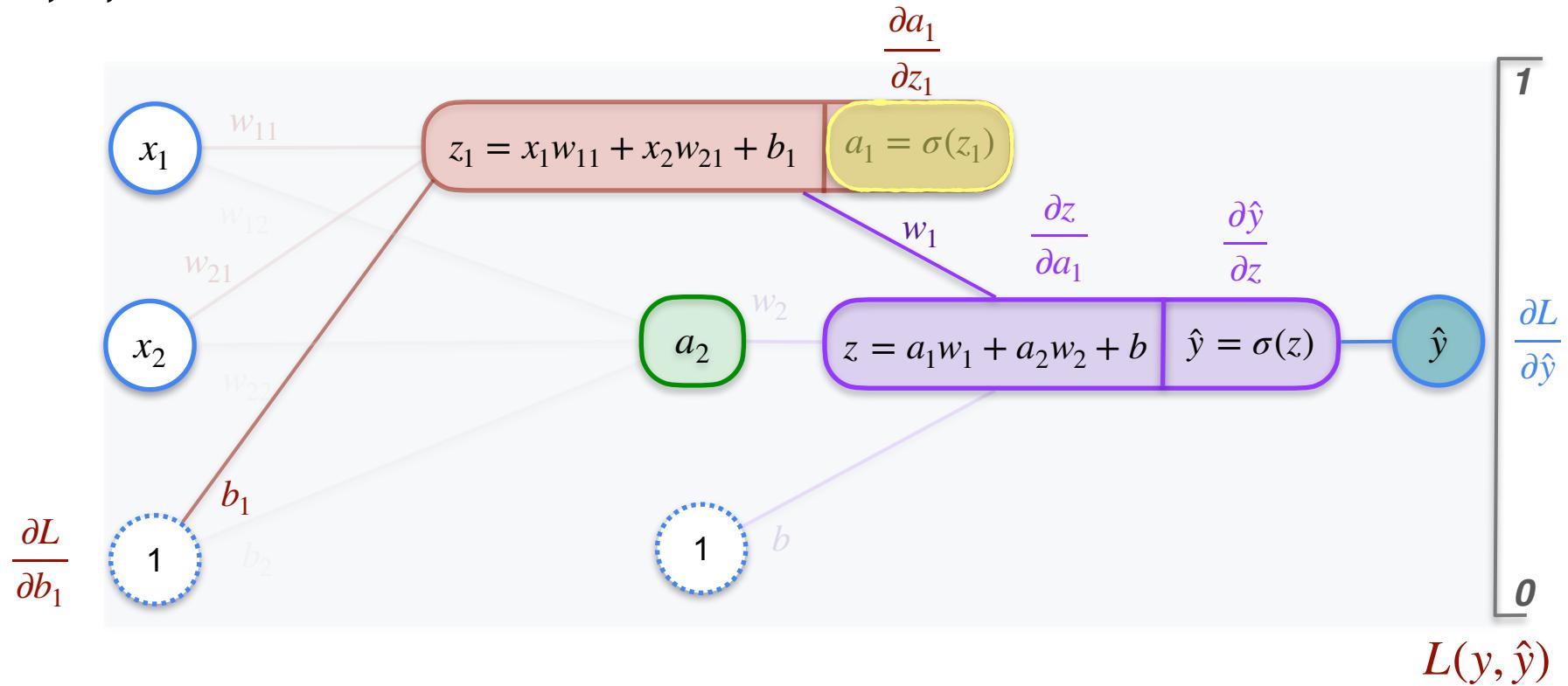
# 2,2,1 Neural Network



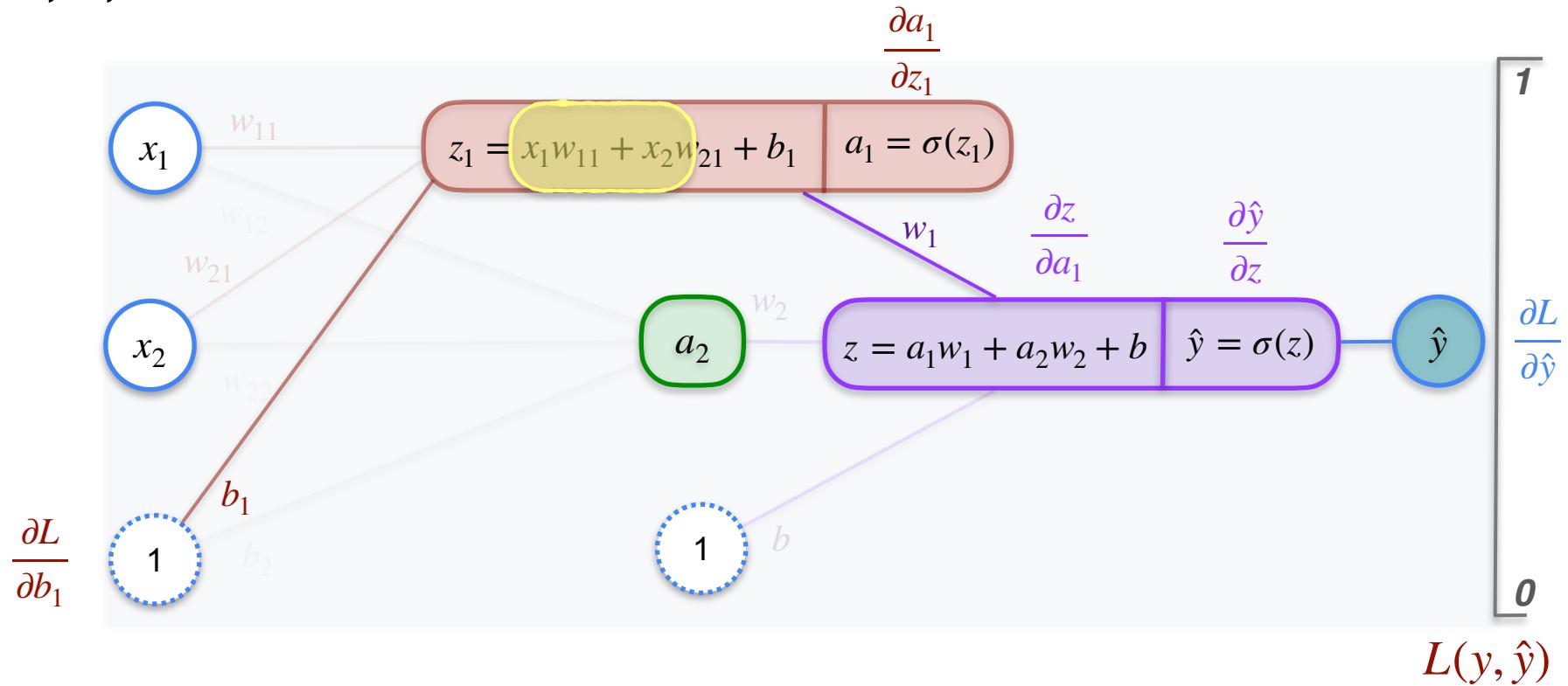
# 2,2,1 Neural Network



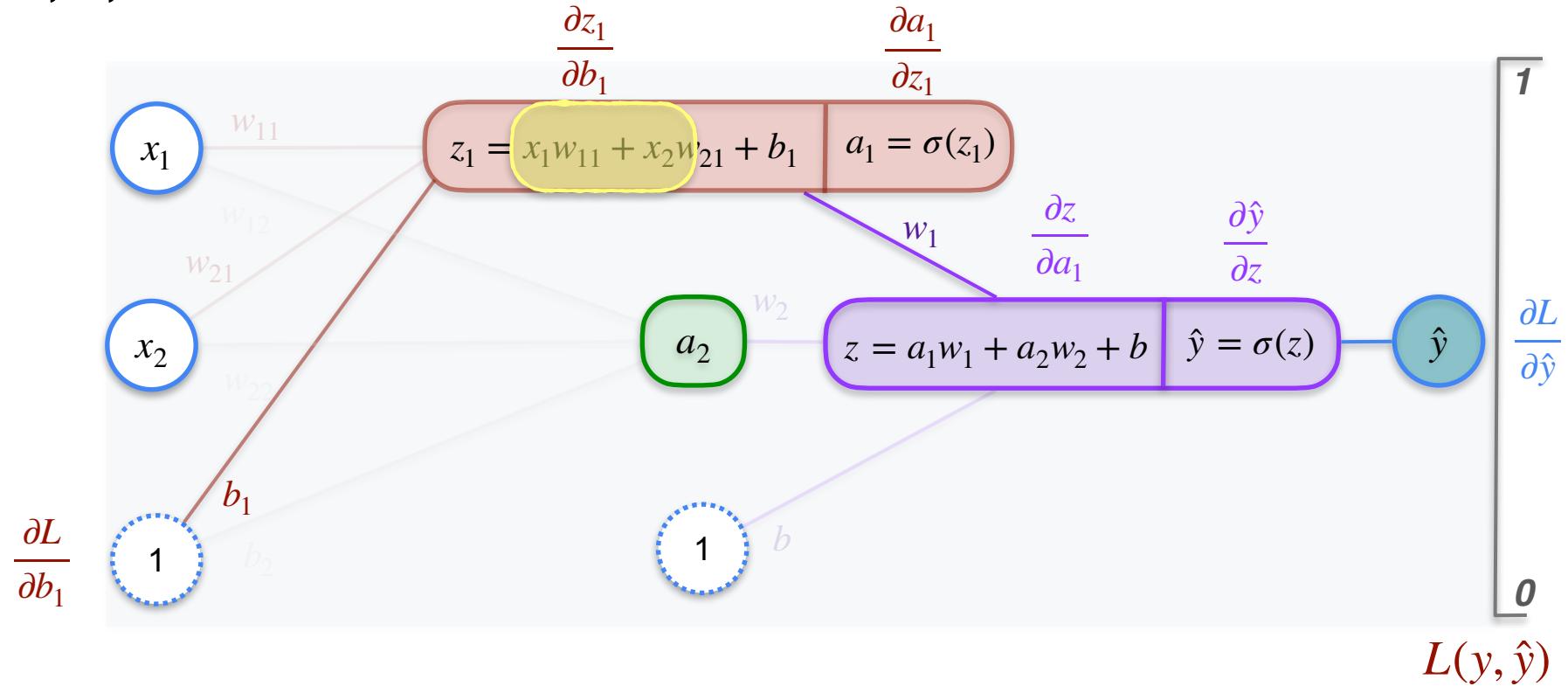
# 2,2,1 Neural Network



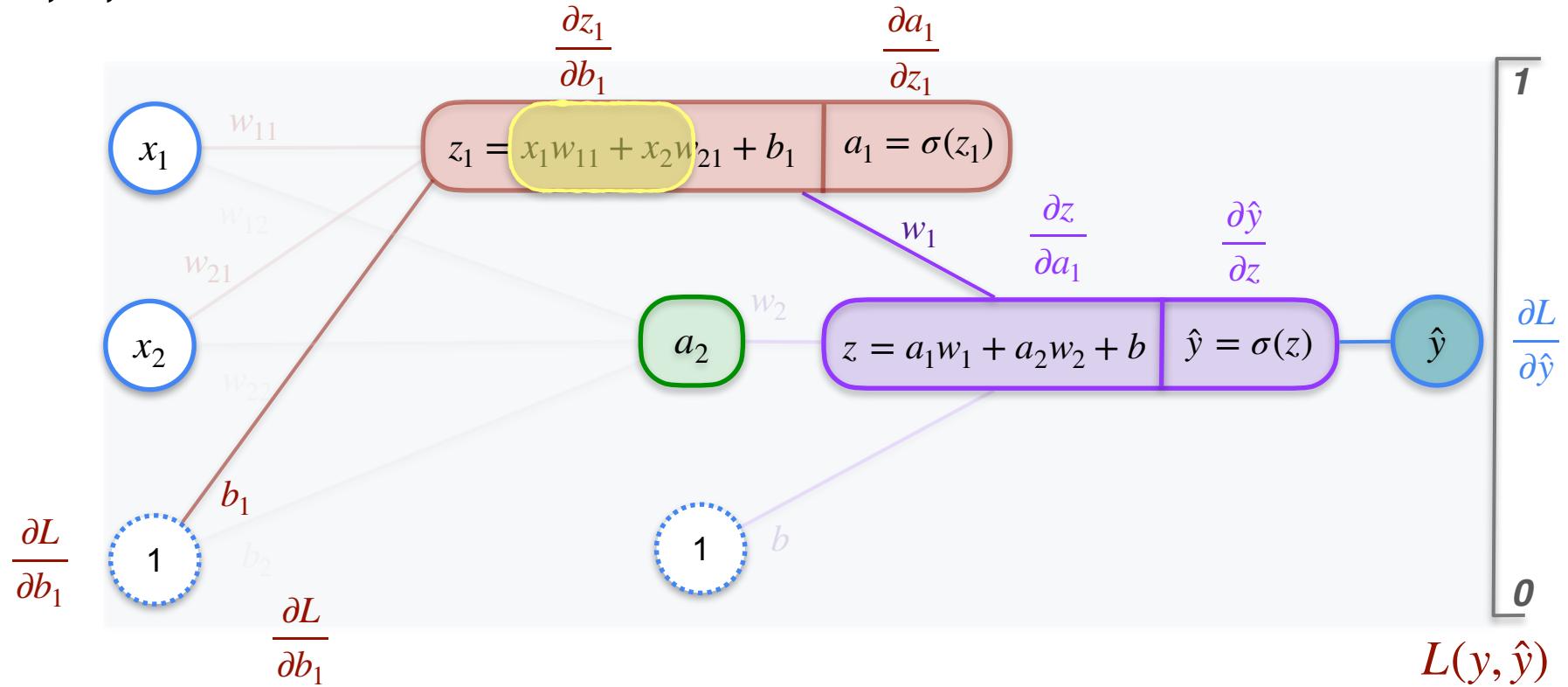
# 2,2,1 Neural Network



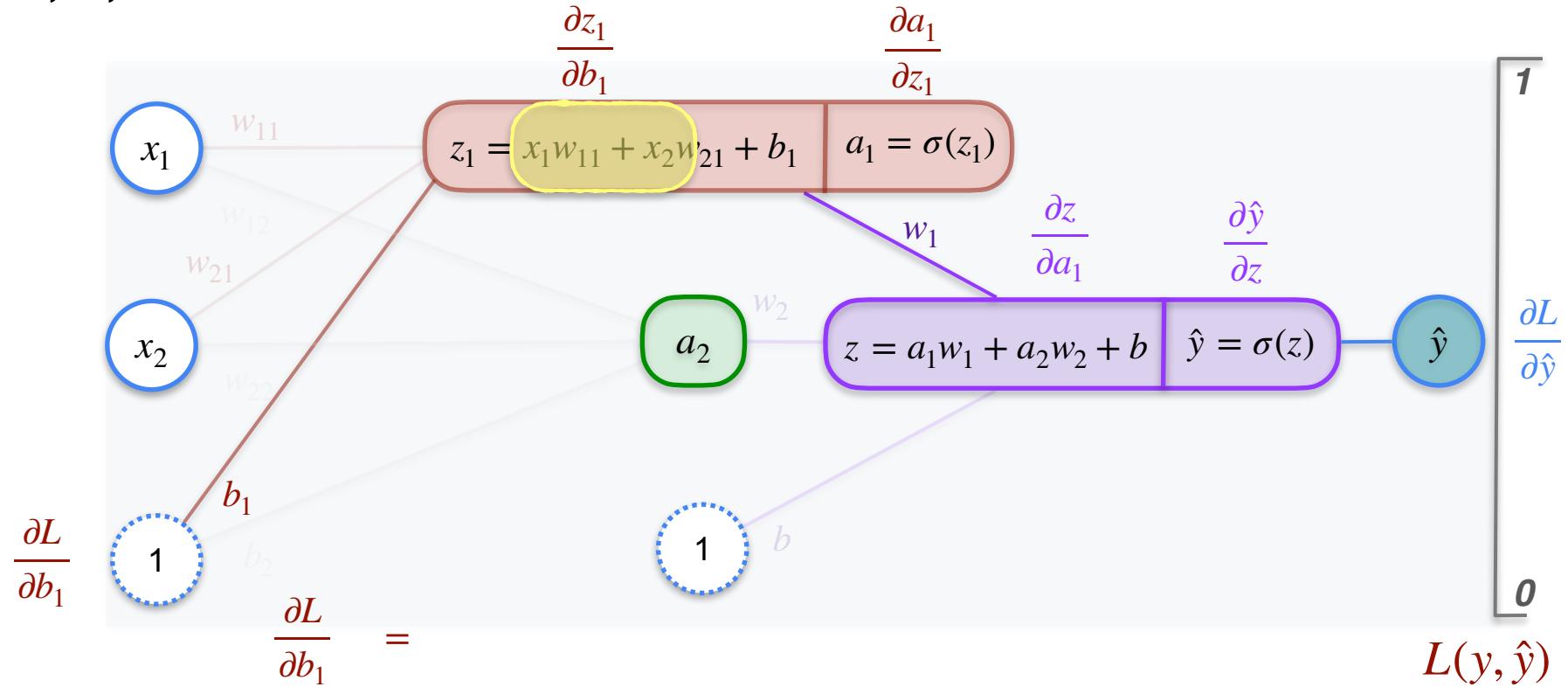
# 2,2,1 Neural Network



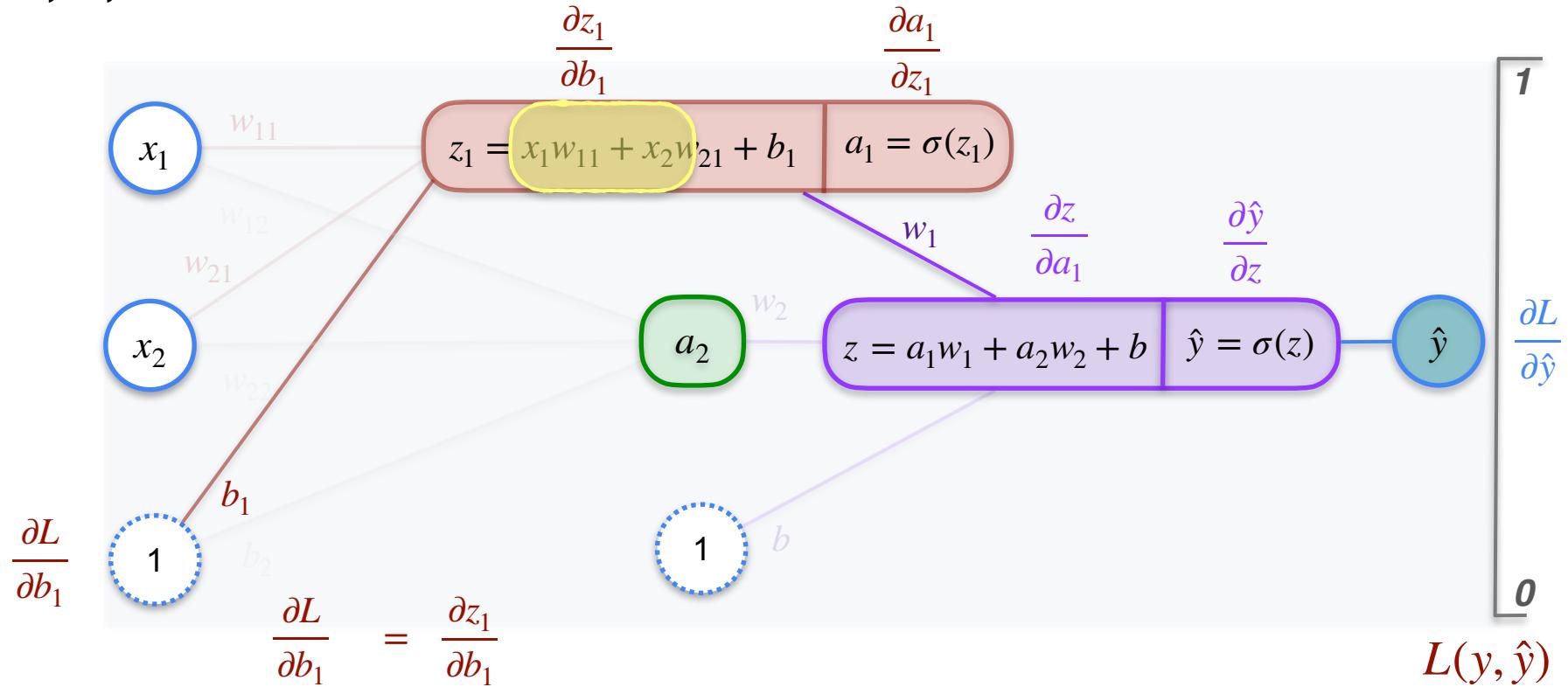
# 2,2,1 Neural Network



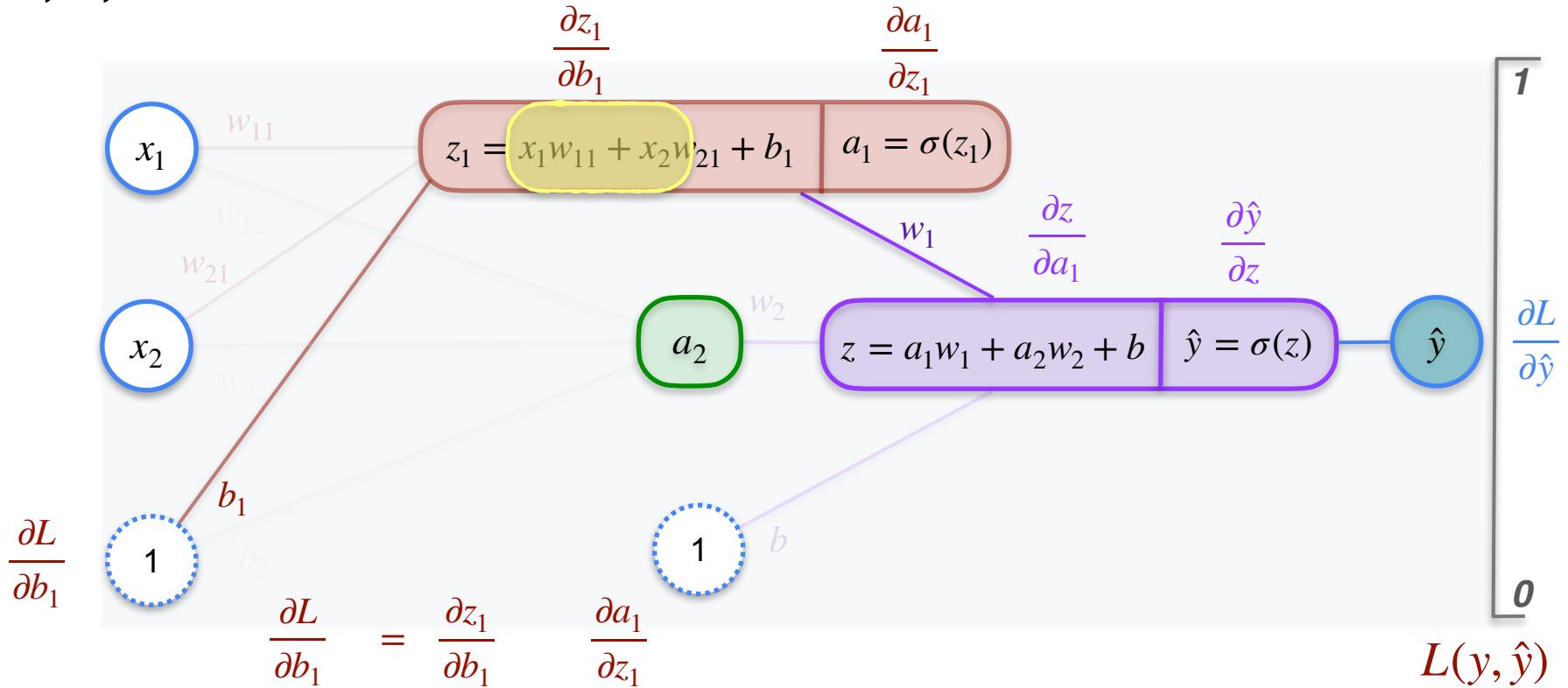
# 2,2,1 Neural Network



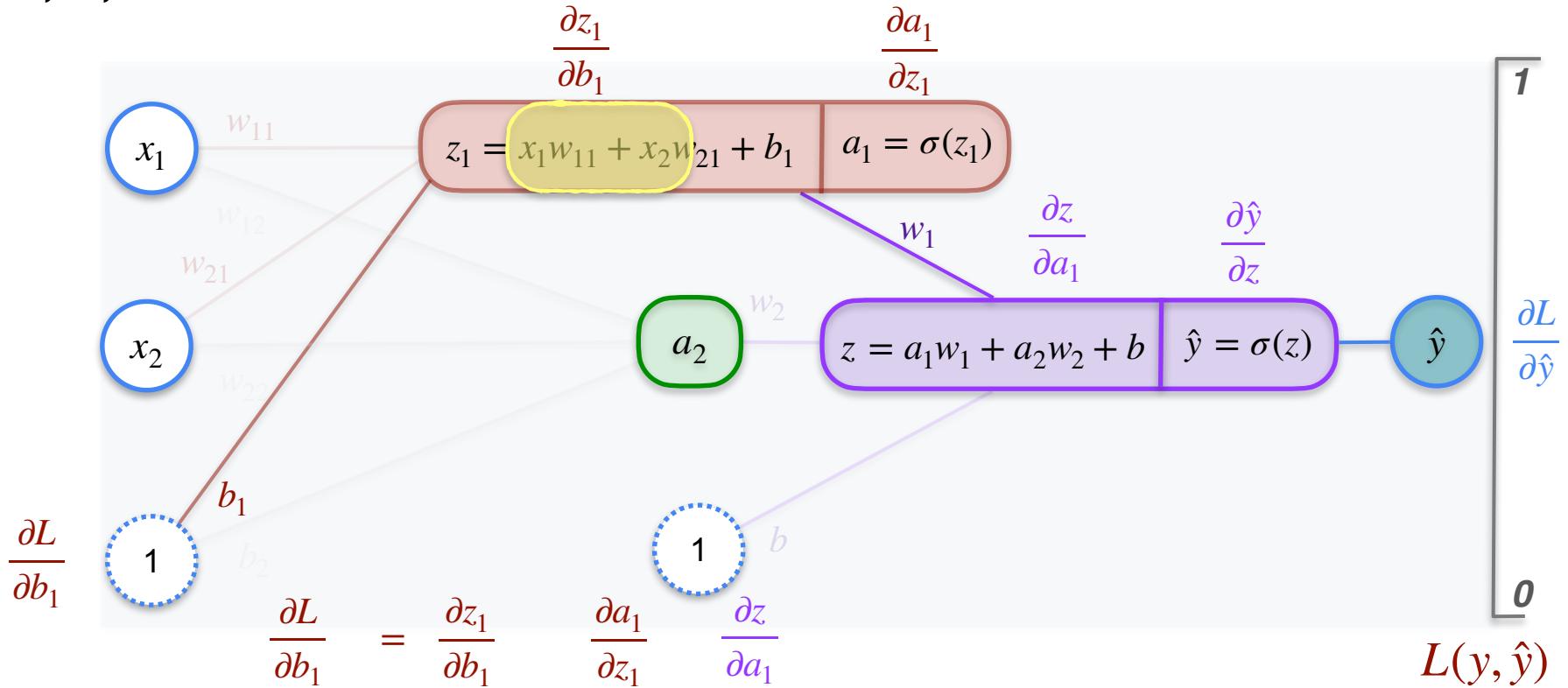
# 2,2,1 Neural Network



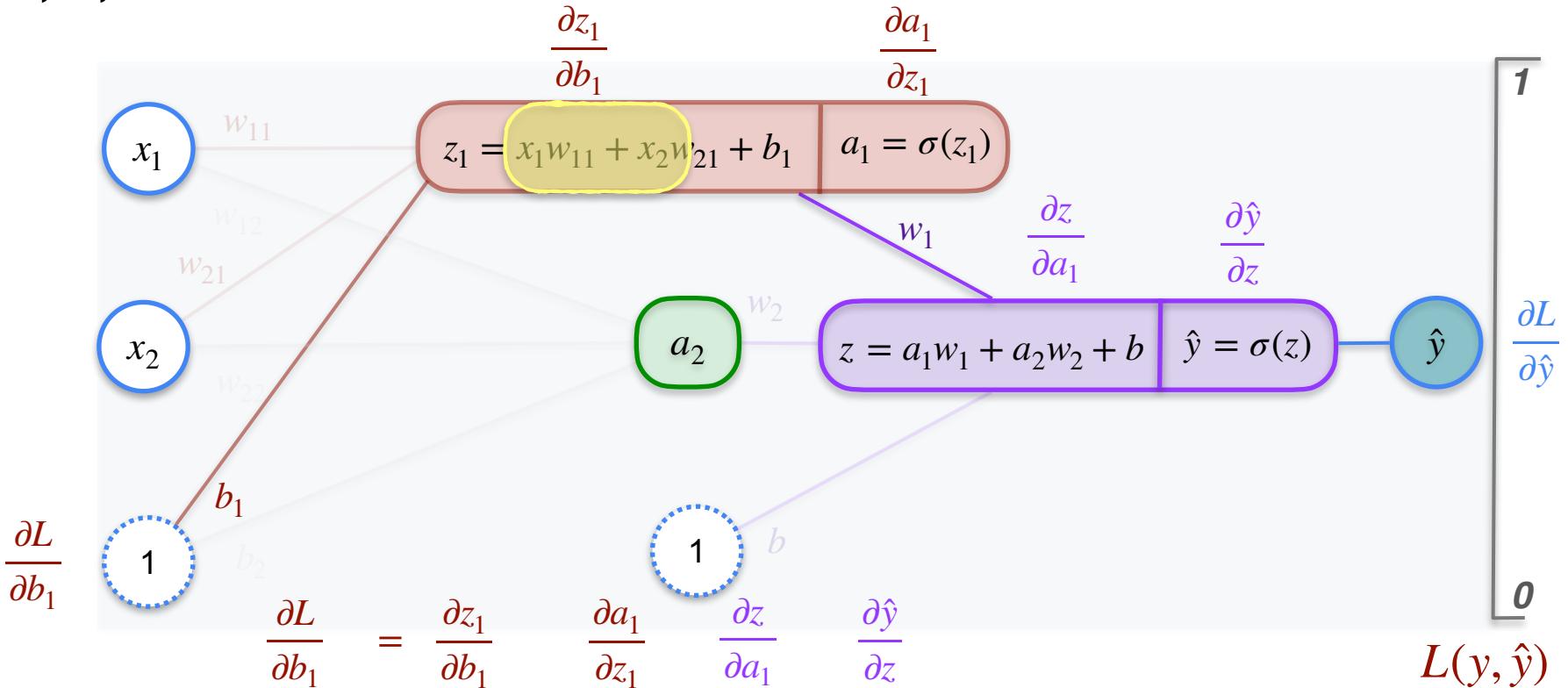
# 2,2,1 Neural Network



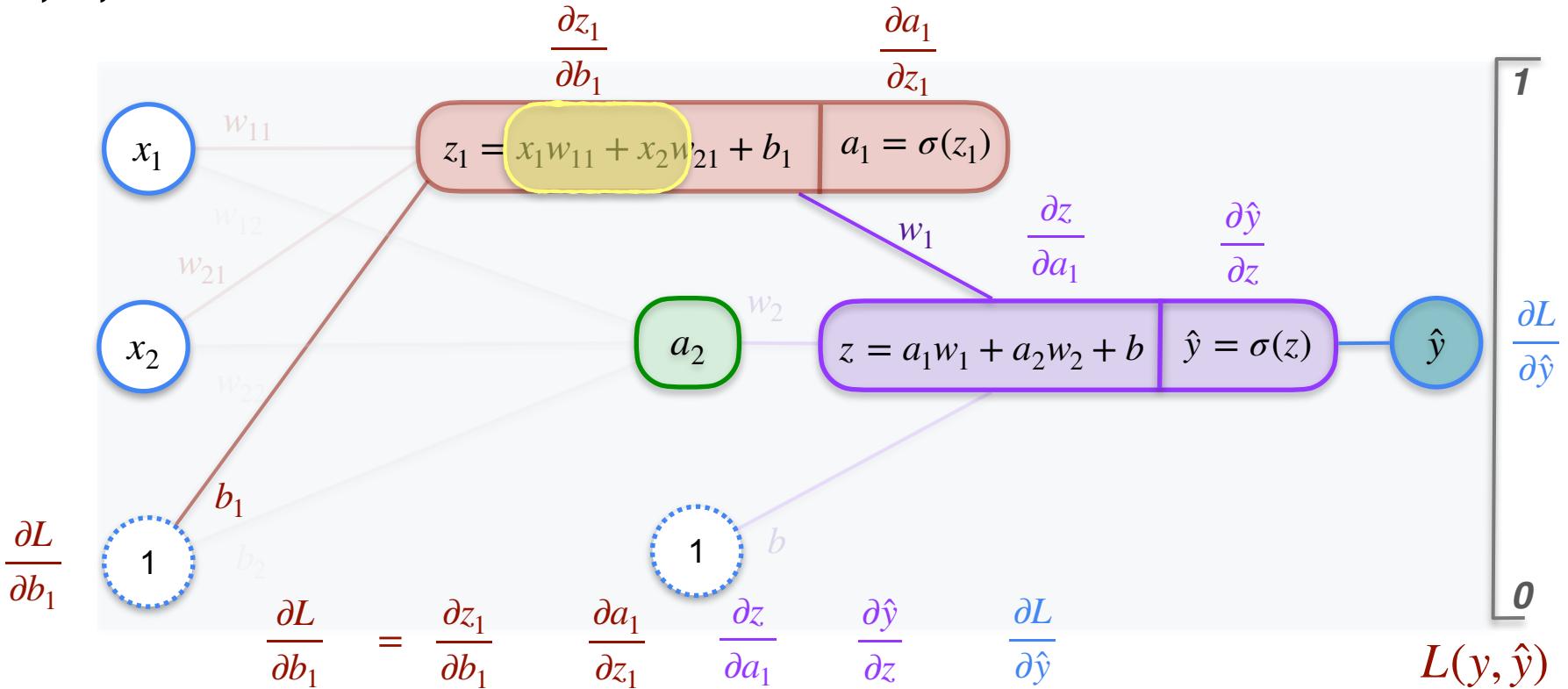
# 2,2,1 Neural Network



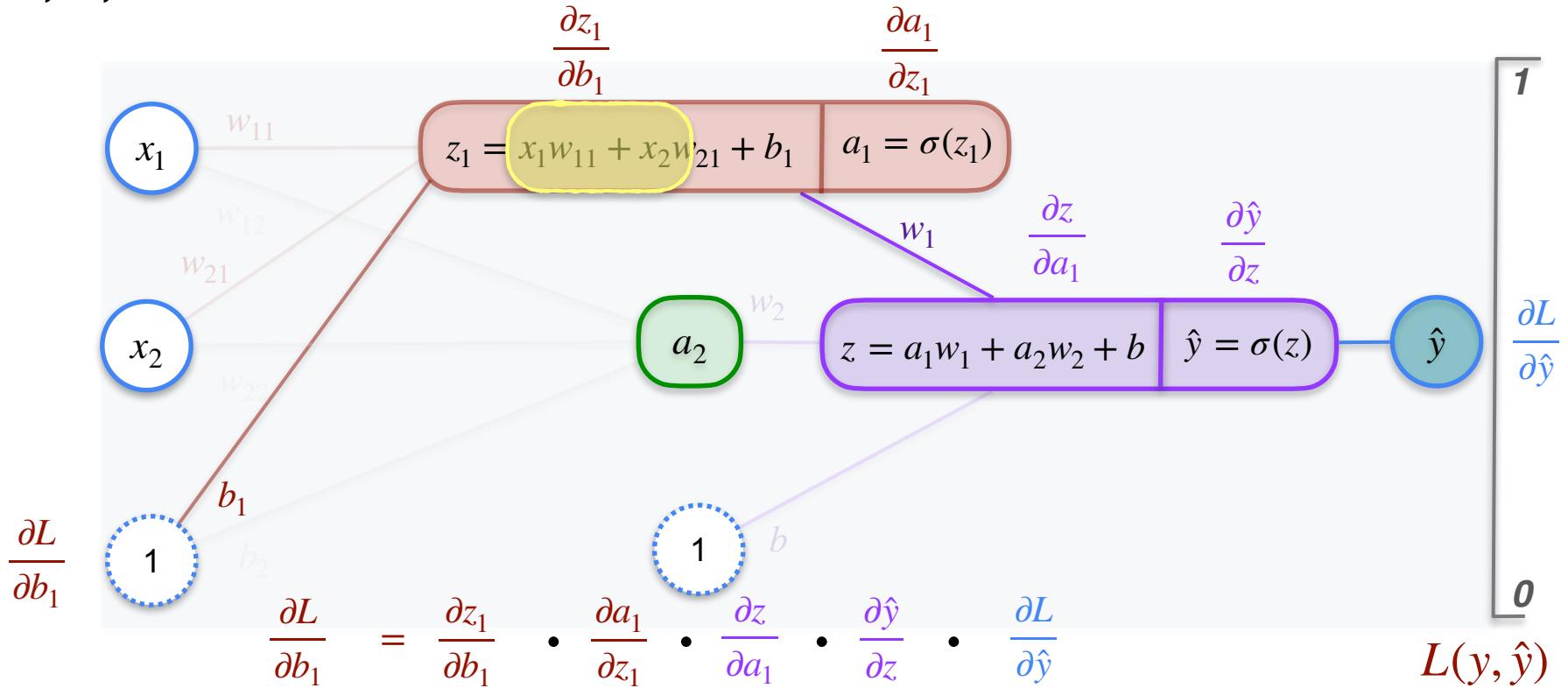
# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b_1}$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b_1} =$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \boxed{\frac{\partial z_1}{\partial b_1}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \boxed{\frac{\partial z_1}{\partial b_1}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \boxed{\frac{\partial z_1}{\partial b_1}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$\frac{\partial L}{\partial b_1} = 1$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1 - a_1)$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1 - a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1 - a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1 - a_1) w_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1-a_1) w_1 \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1-a_1) w_1 \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b_1} = 1 - a_1(1-a_1) w_1 \hat{y}(1-\hat{y})$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial b_1} = 1 - a_1(1-a_1) w_1 \hat{y}(1-\hat{y}) \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial b_1} = 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial b_1} = 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$
$$\frac{\partial L}{\partial b_1} = 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial b_1} &= \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial b_1} &= 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial b_1} &= \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial b_1} &= 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

*to find optimal  
value of  $b_1$  that  
gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial b_1} &= \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial b_1} &= 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

**Perform gradient descent with**

$$b_1 \rightarrow b_1 - \alpha \frac{\partial L}{\partial b_1}$$

**to find optimal value of  $b_1$  that gives the least error**

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$\begin{aligned}\frac{\partial L}{\partial b_1} &= \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial b_1} &= 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

$$b_1 \rightarrow b_1 - \alpha$$

*to find optimal value of  $b_1$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

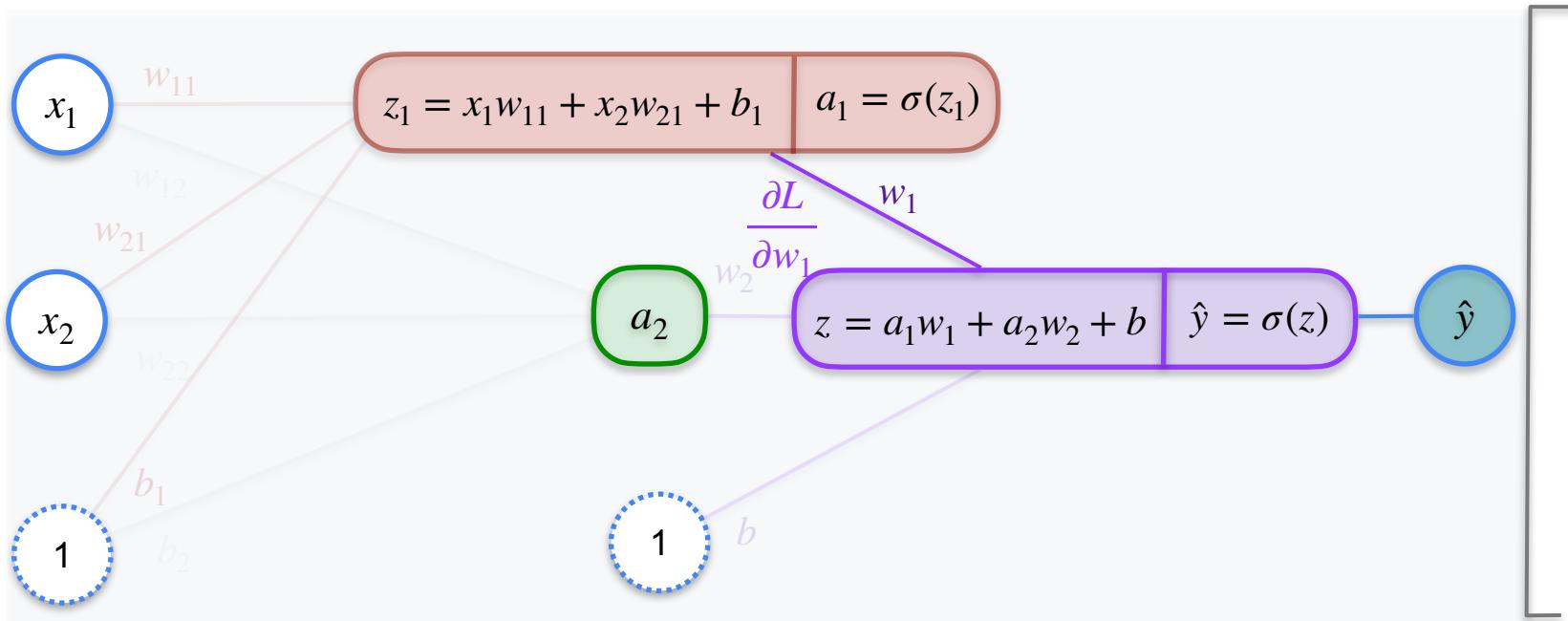
$$\begin{aligned}\frac{\partial L}{\partial b_1} &= \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial b_1} &= 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -w_1 a_1 (1-a_1) (y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

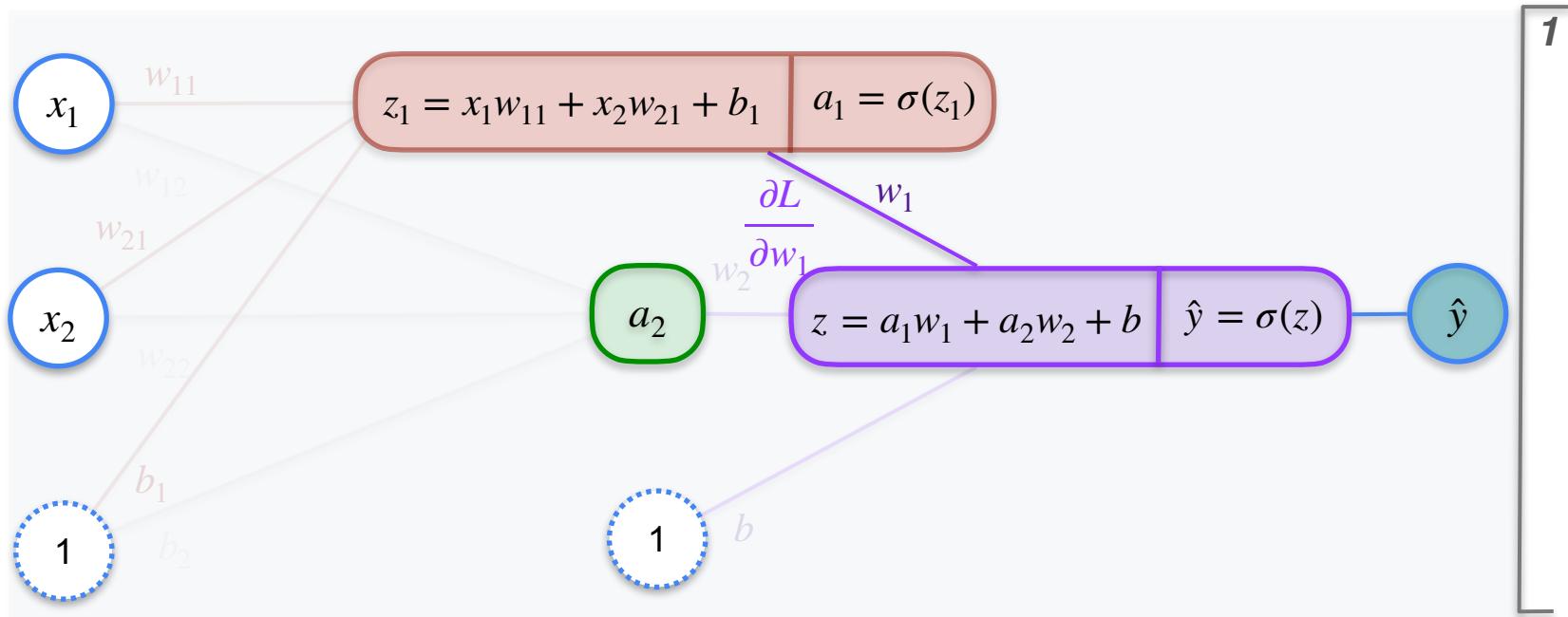
$$b_1 \rightarrow b_1 - \alpha (-w_1 a_1 (1-a_1) (y - \hat{y}))$$

*to find optimal value of  $b_1$  that gives the least error*

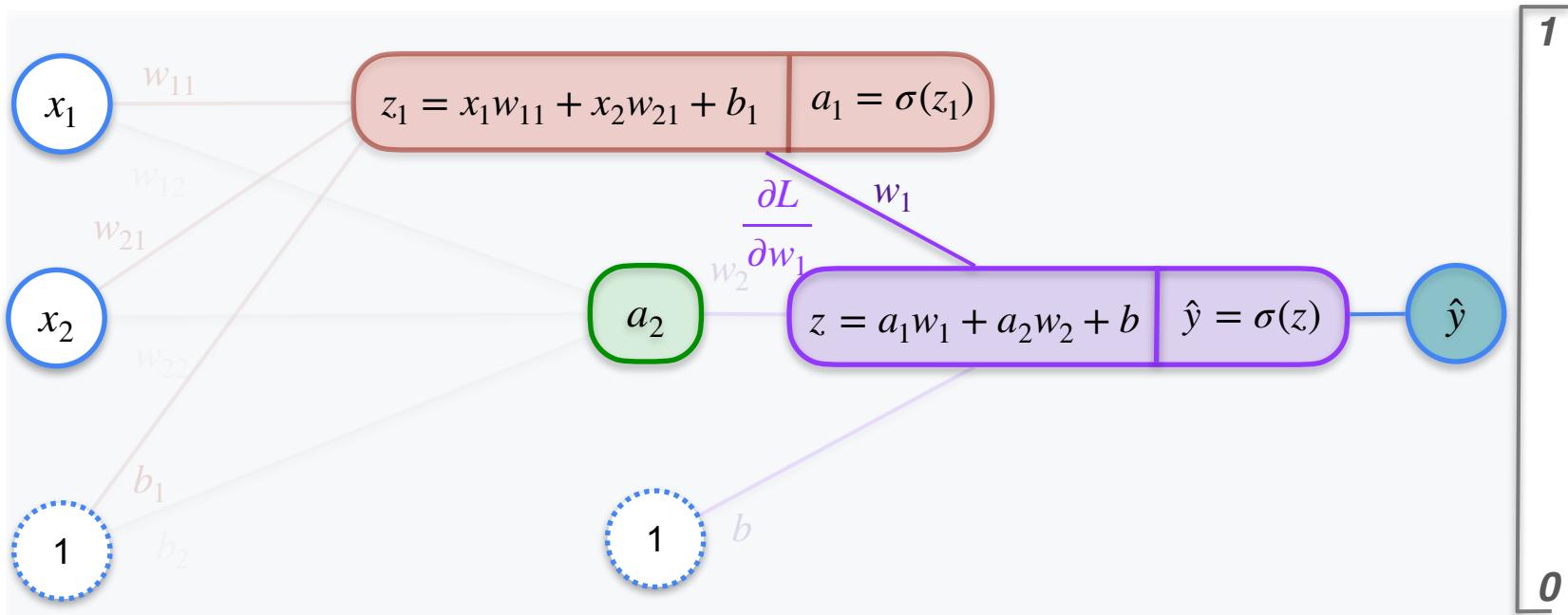
# 2,2,1 Neural Network



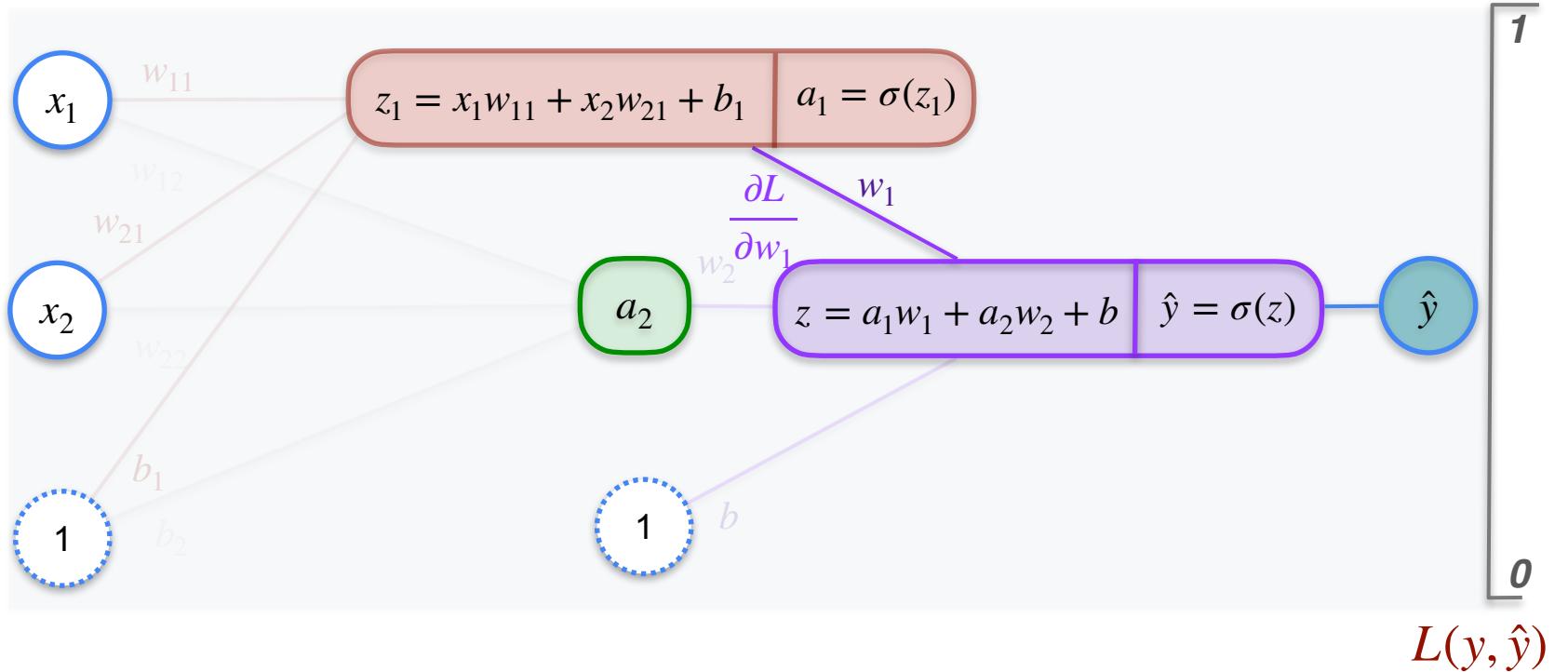
# 2,2,1 Neural Network



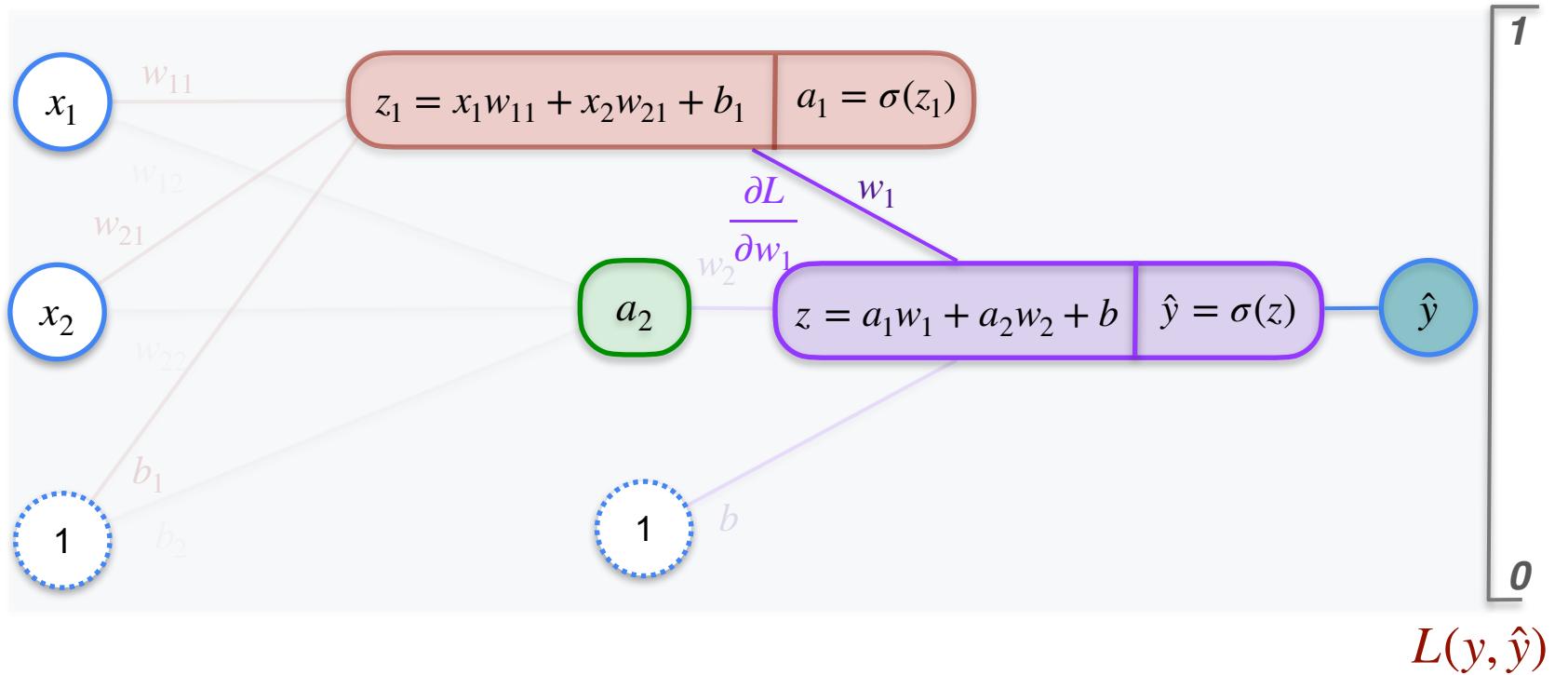
# 2,2,1 Neural Network



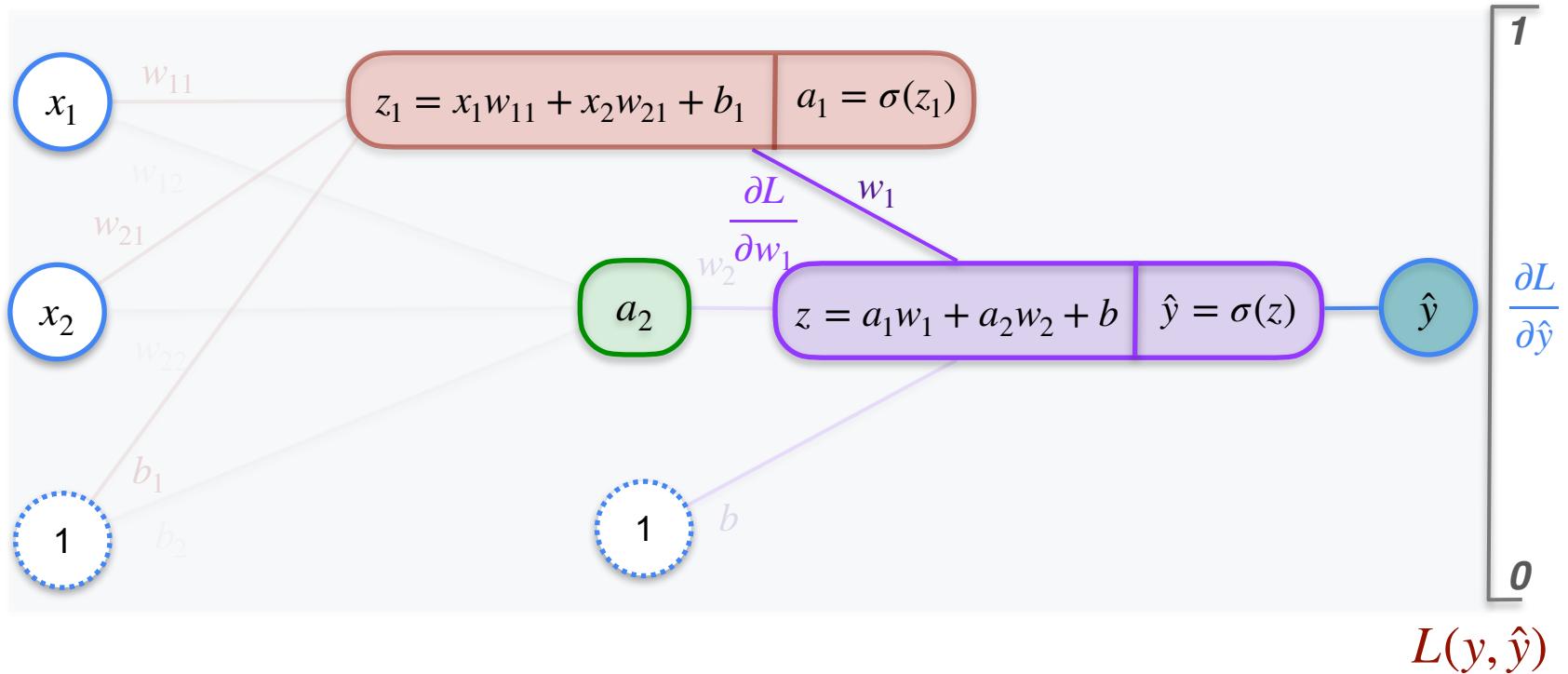
# 2,2,1 Neural Network



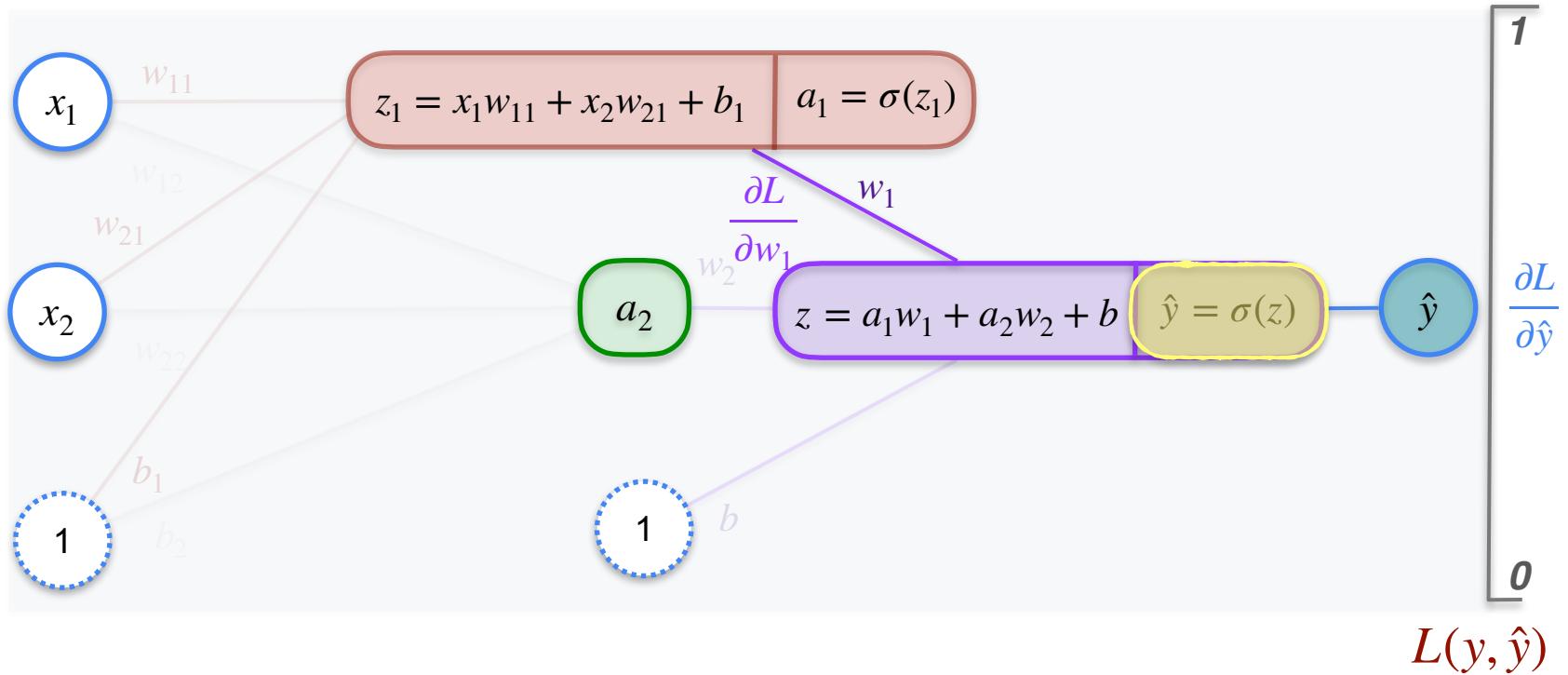
# 2,2,1 Neural Network



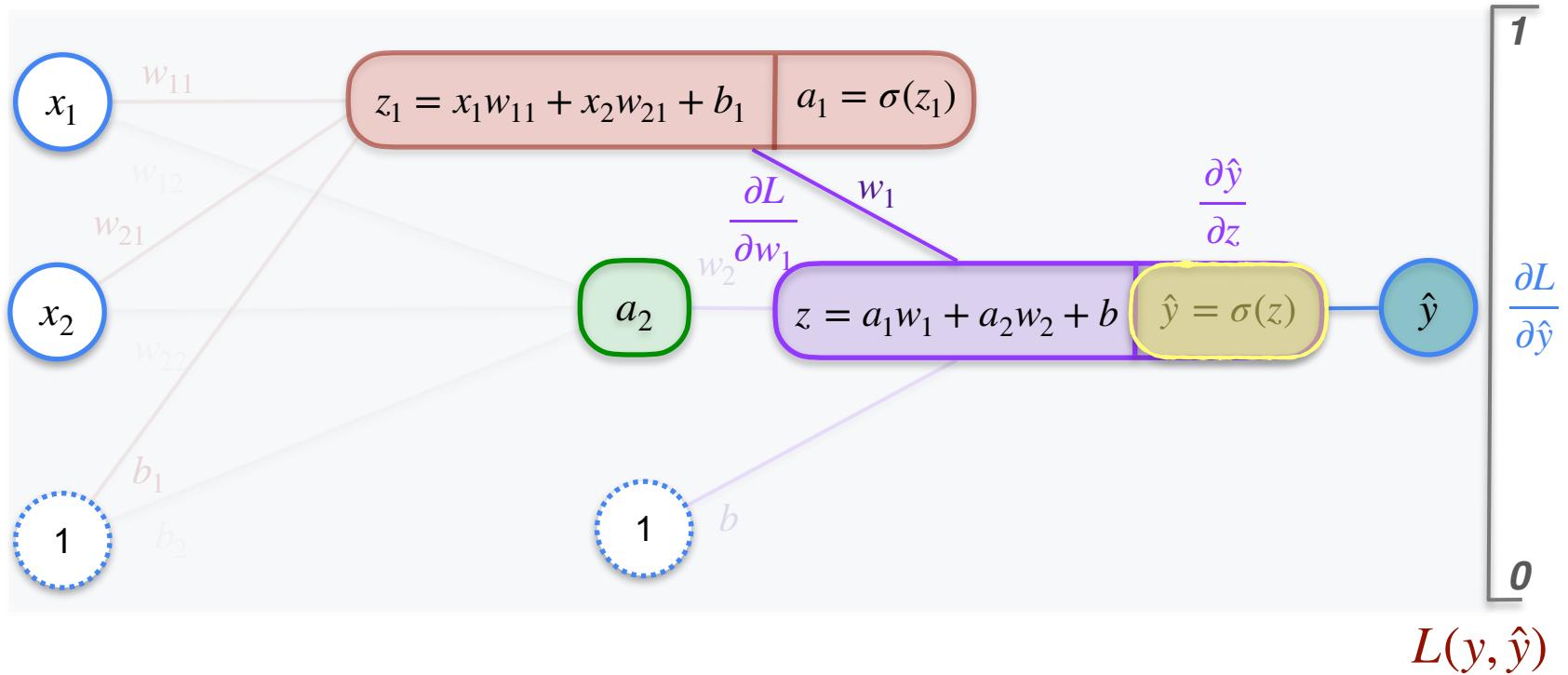
# 2,2,1 Neural Network



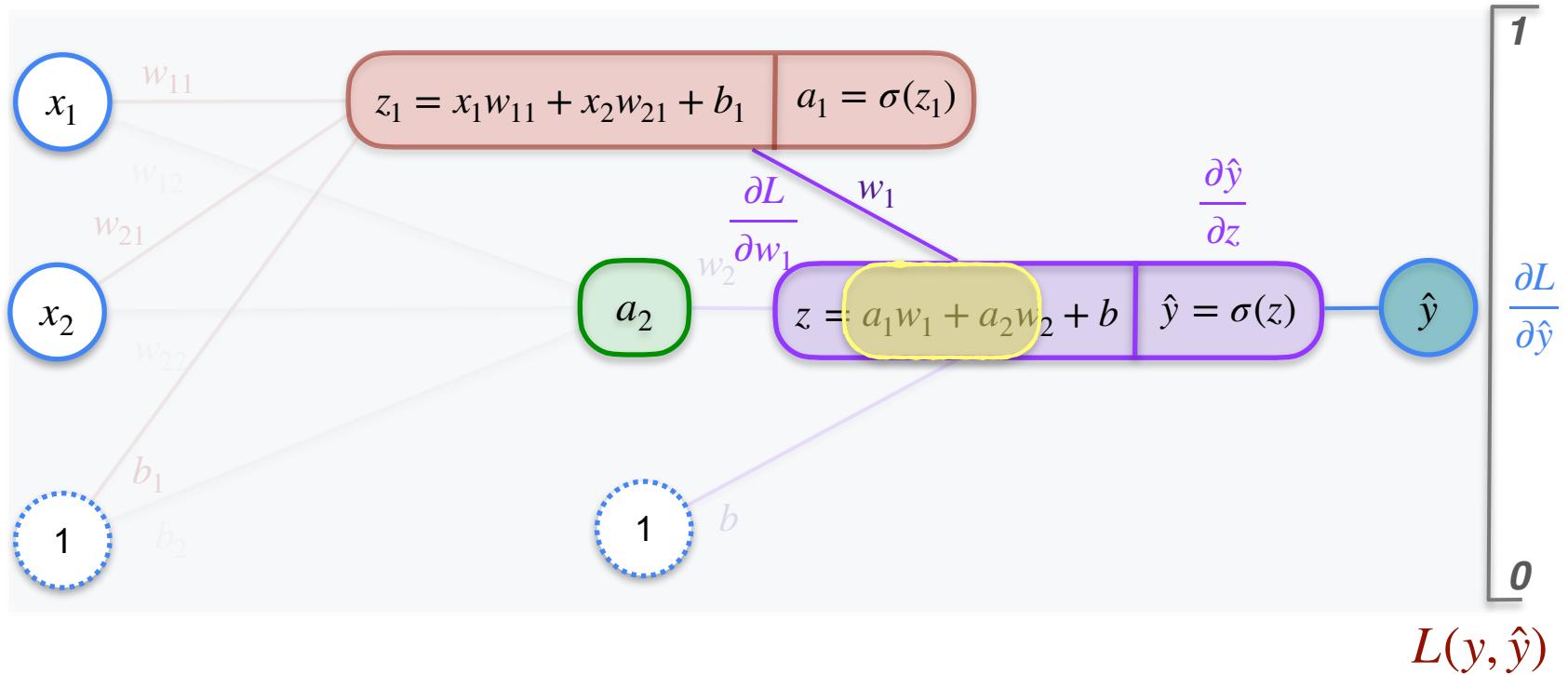
# 2,2,1 Neural Network



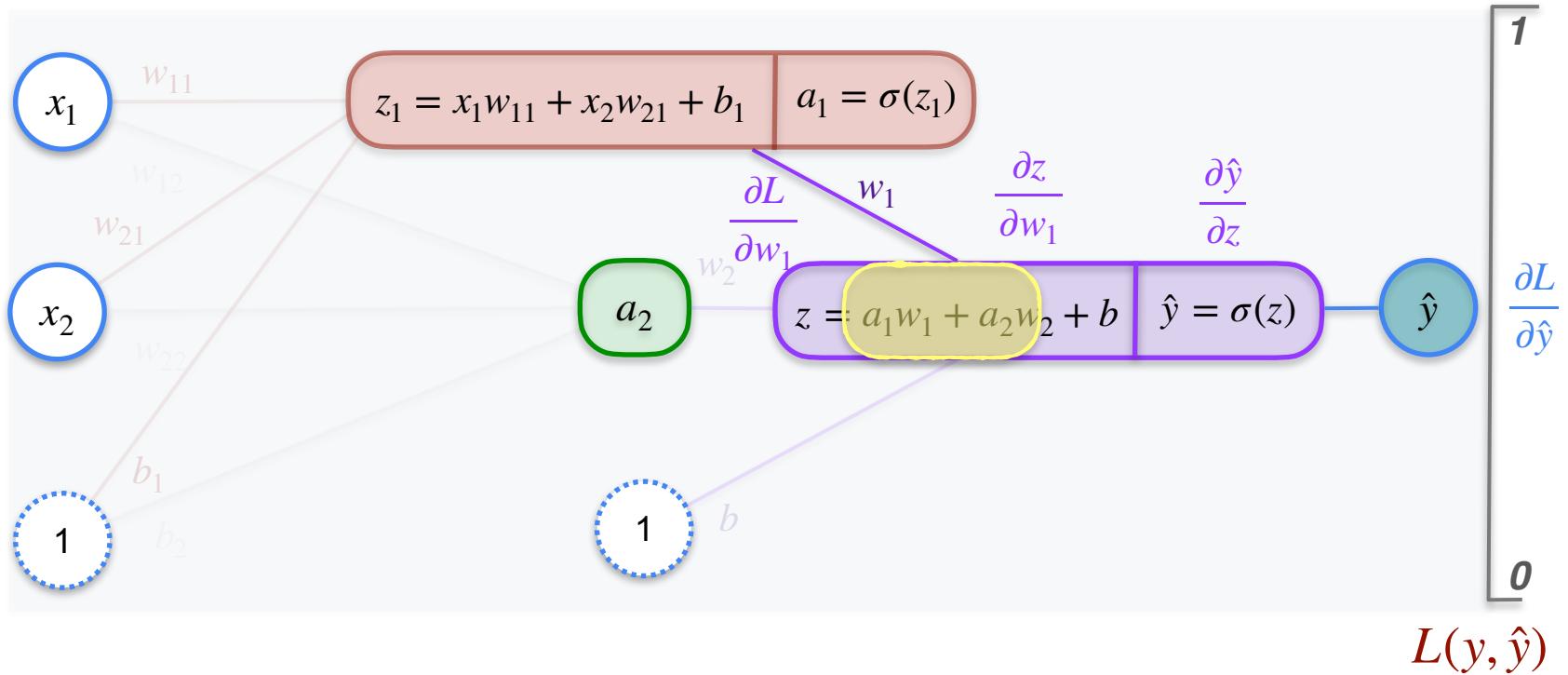
# 2,2,1 Neural Network



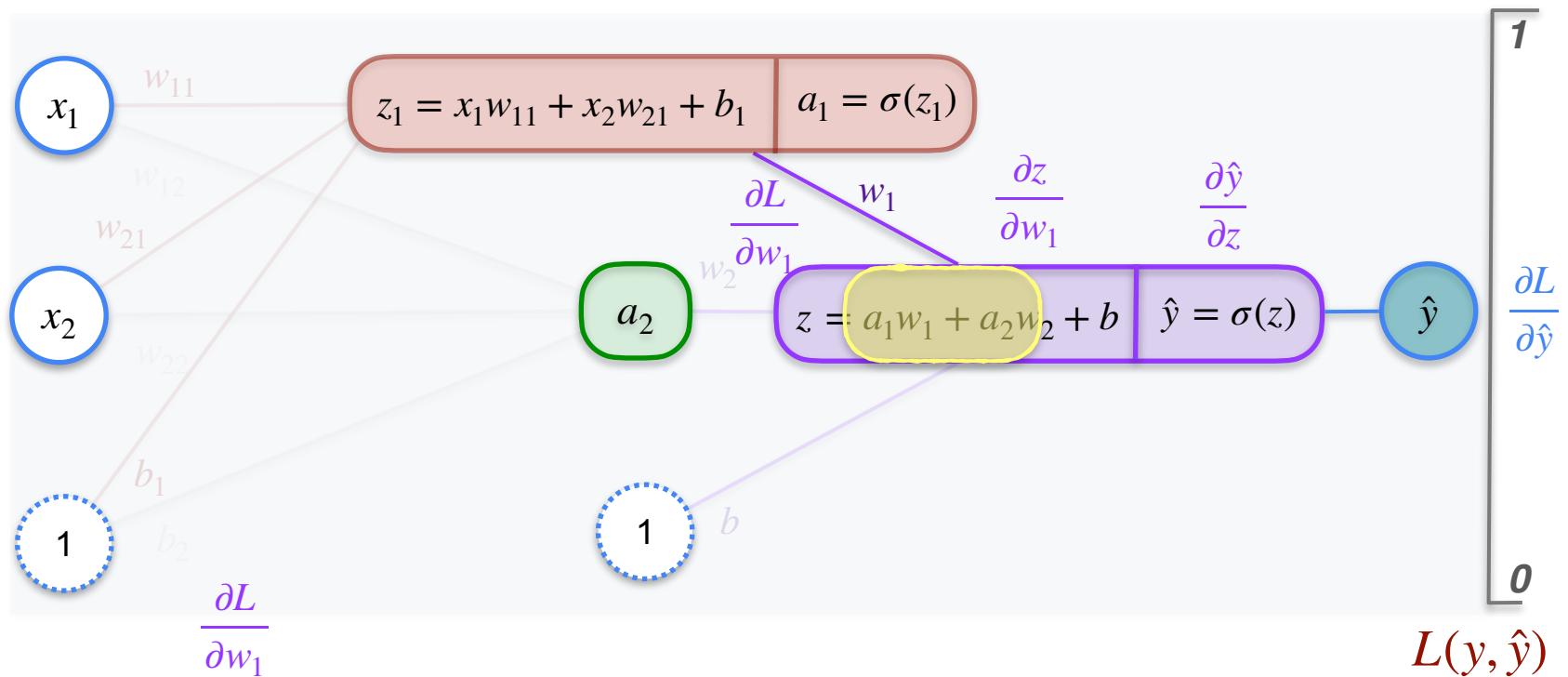
# 2,2,1 Neural Network



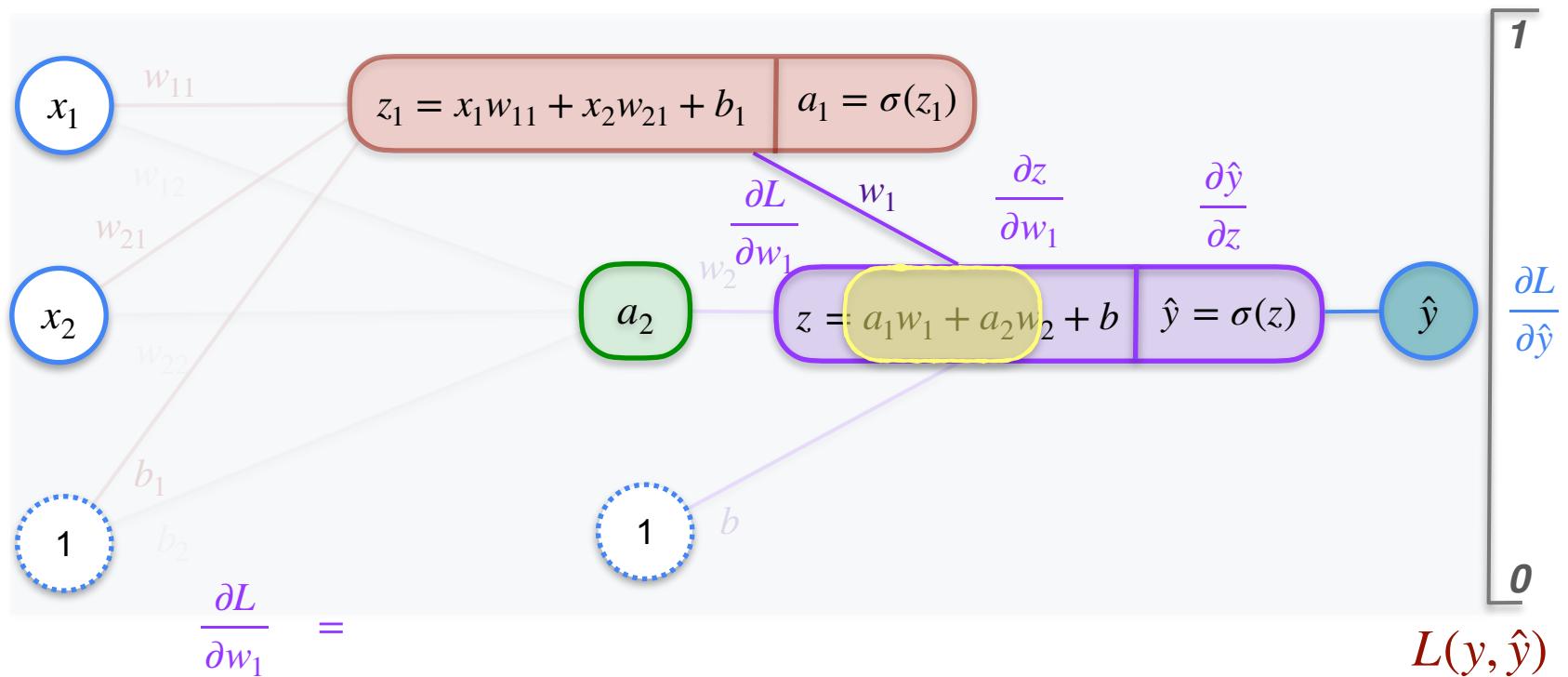
# 2,2,1 Neural Network



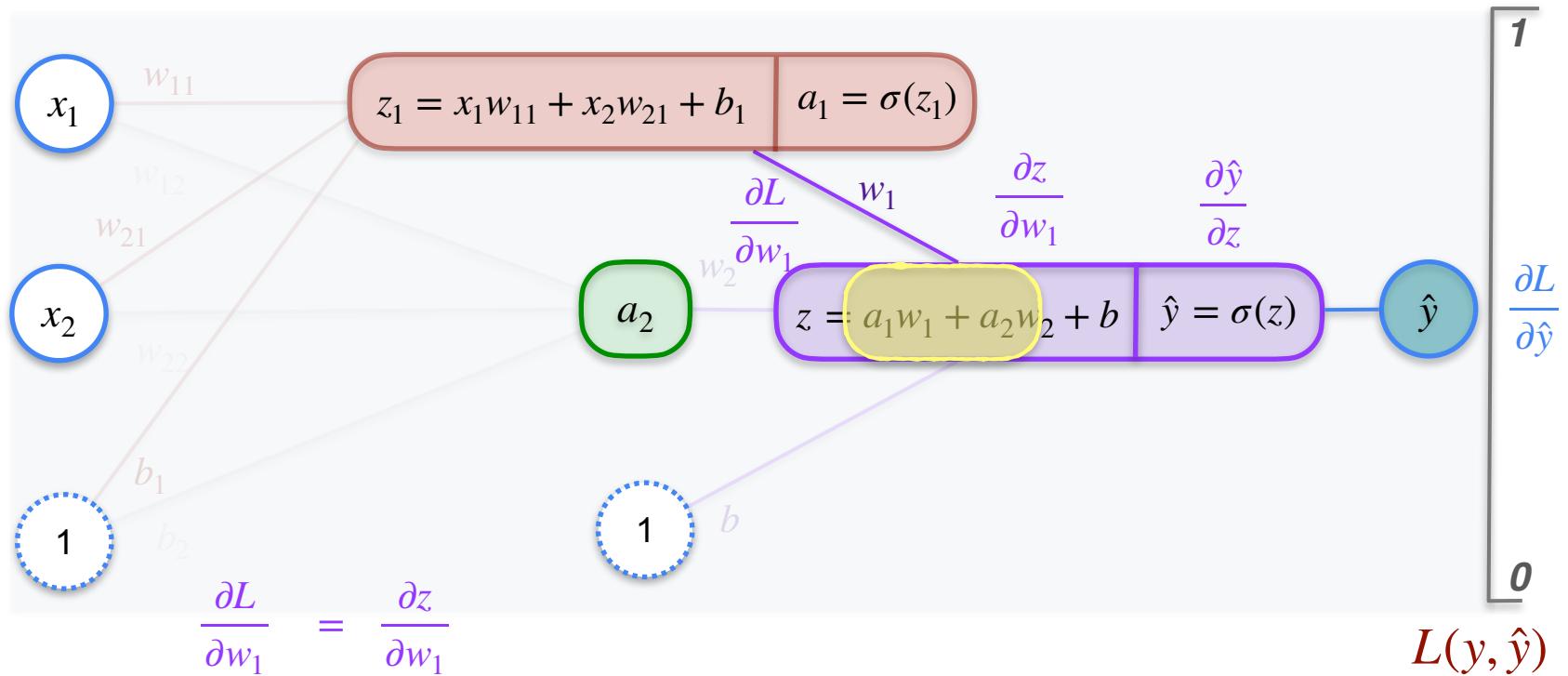
# 2,2,1 Neural Network



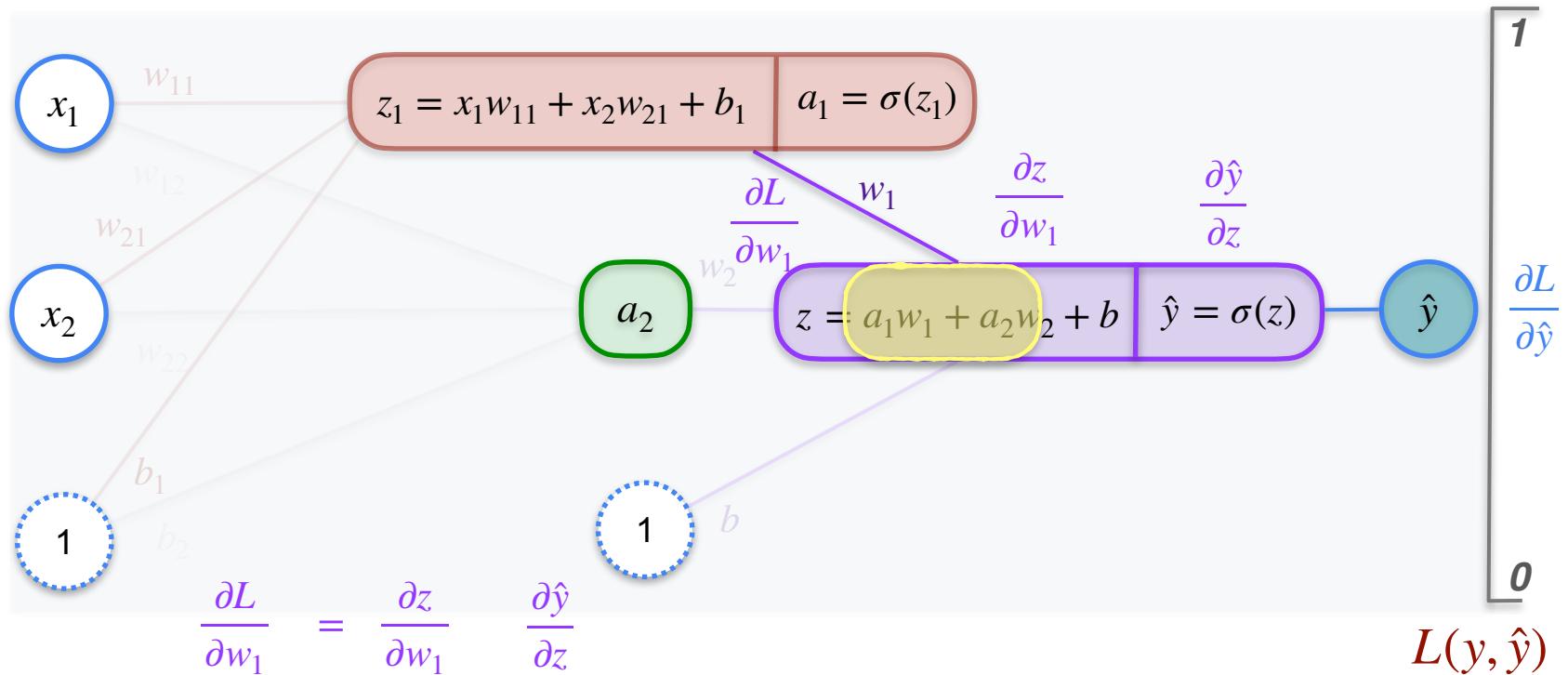
# 2,2,1 Neural Network



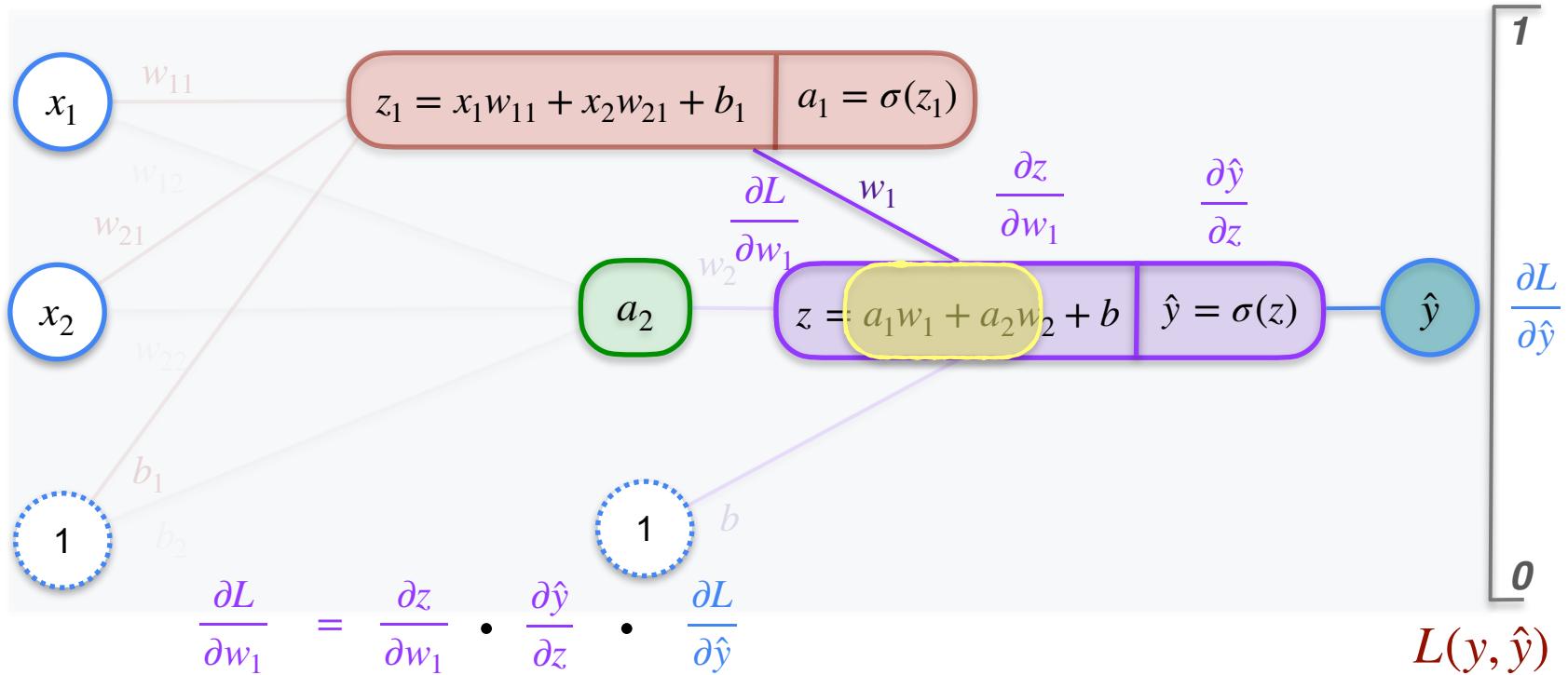
# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1w_1 + a_2w_2 + b$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \boxed{\frac{\partial z}{\partial w_1}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} = a_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} = a_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} = a_1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} = a_1 \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} = a_1 \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = a_1 \hat{y}(1-\hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = a_1 \hat{y}(1-\hat{y}) \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = a_1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = a_1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial w_1} = a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -a_1(y - \hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_1} &= a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -a_1(y - \hat{y})\end{aligned}$$

*to find optimal value of  $w_1$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_1} &= a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -a_1(y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

*to find optimal  
value of  $w_1$  that  
gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_1} &= a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -a_1(y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

*to find optimal value of  $w_1$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_1} &= a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -a_1(y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

$$w_1 \rightarrow w_1 - \alpha$$

*to find optimal value of  $w_1$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

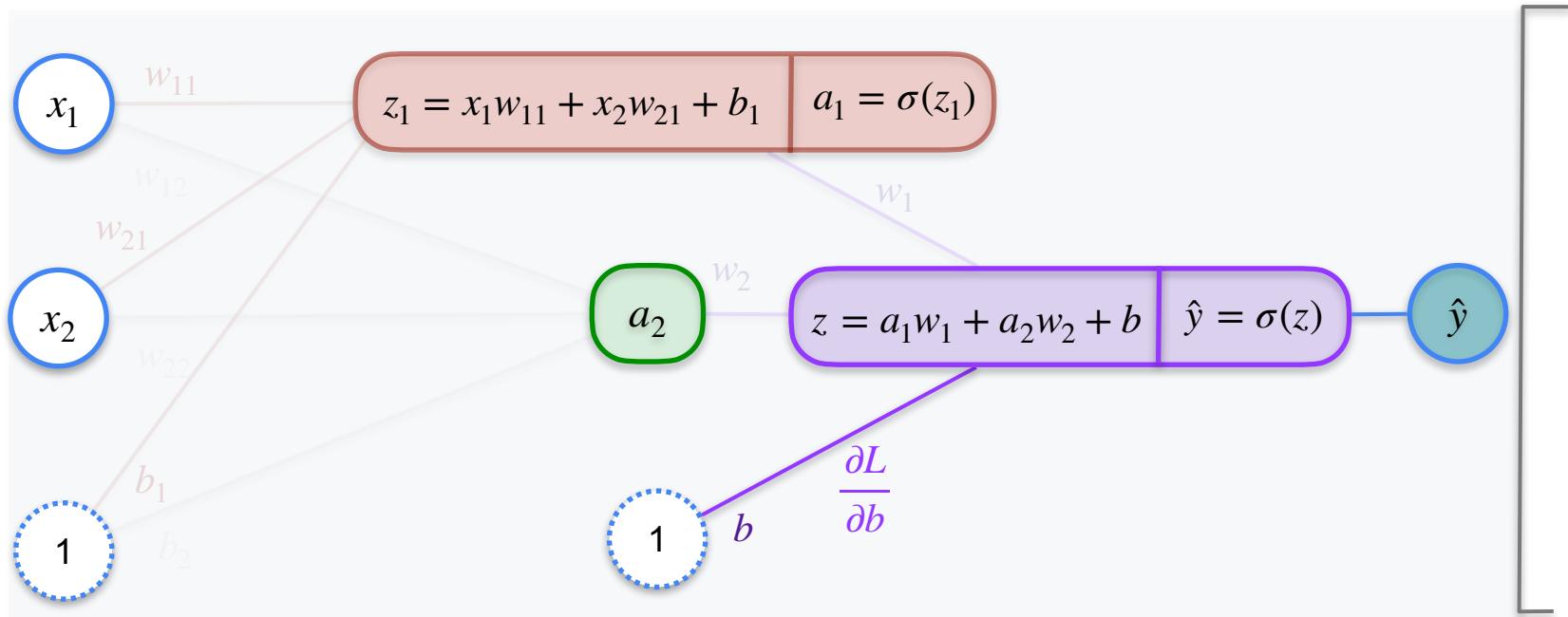
$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}} \\ \frac{\partial L}{\partial w_1} &= a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \\ &= -a_1(y - \hat{y})\end{aligned}$$

*Perform gradient descent with*

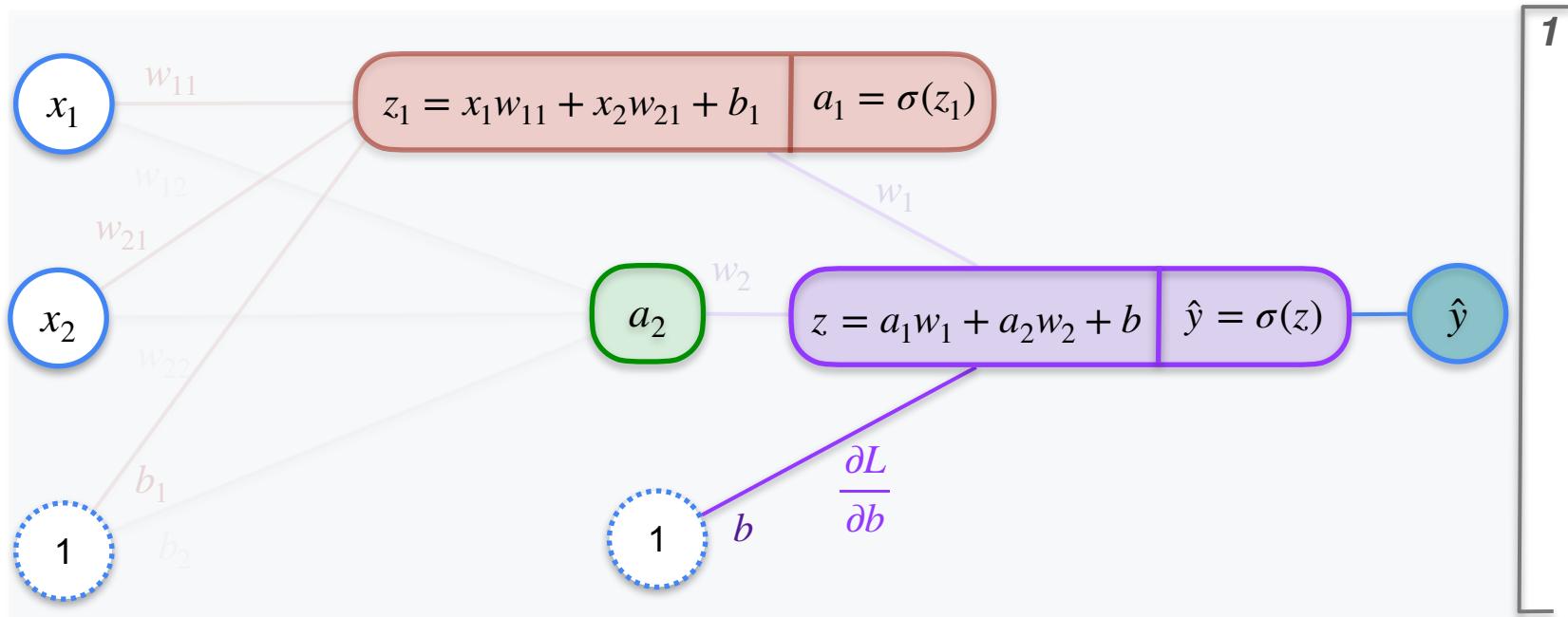
$$w_1 \rightarrow w_1 - \alpha(-a_1(y - \hat{y}))$$

*to find optimal value of  $w_1$  that gives the least error*

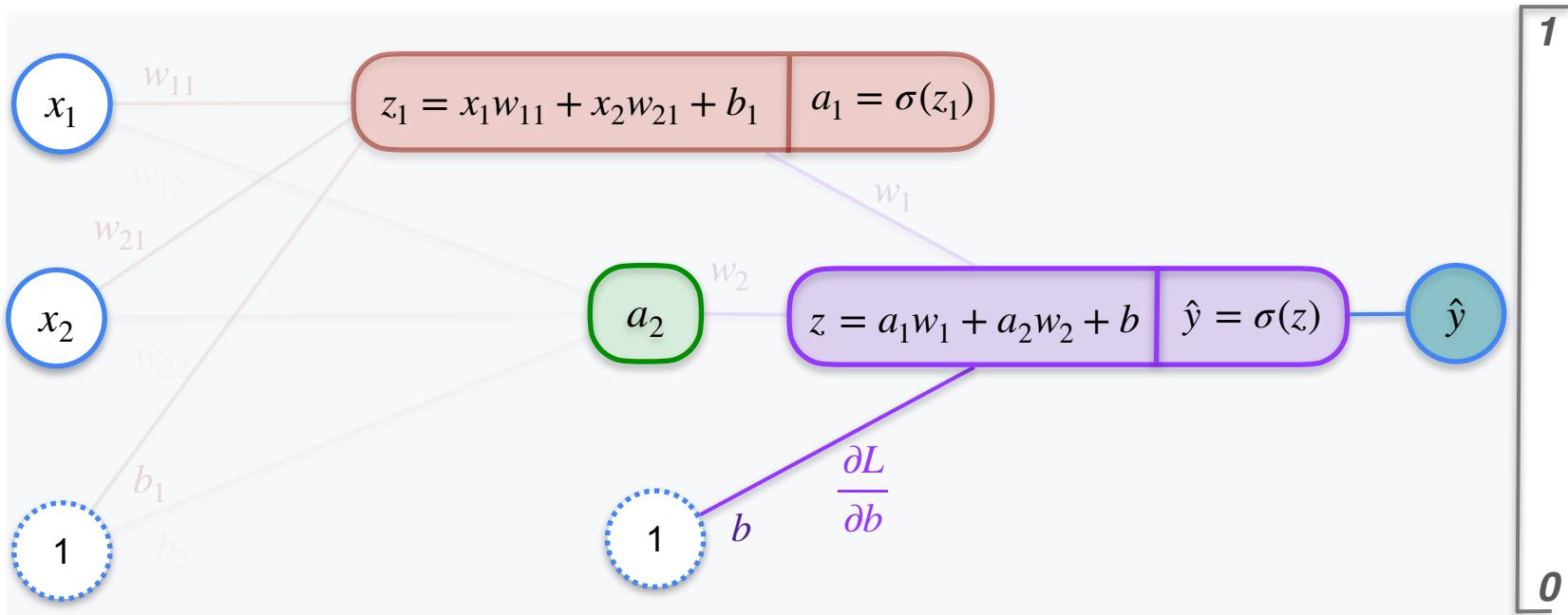
# 2,2,1 Neural Network



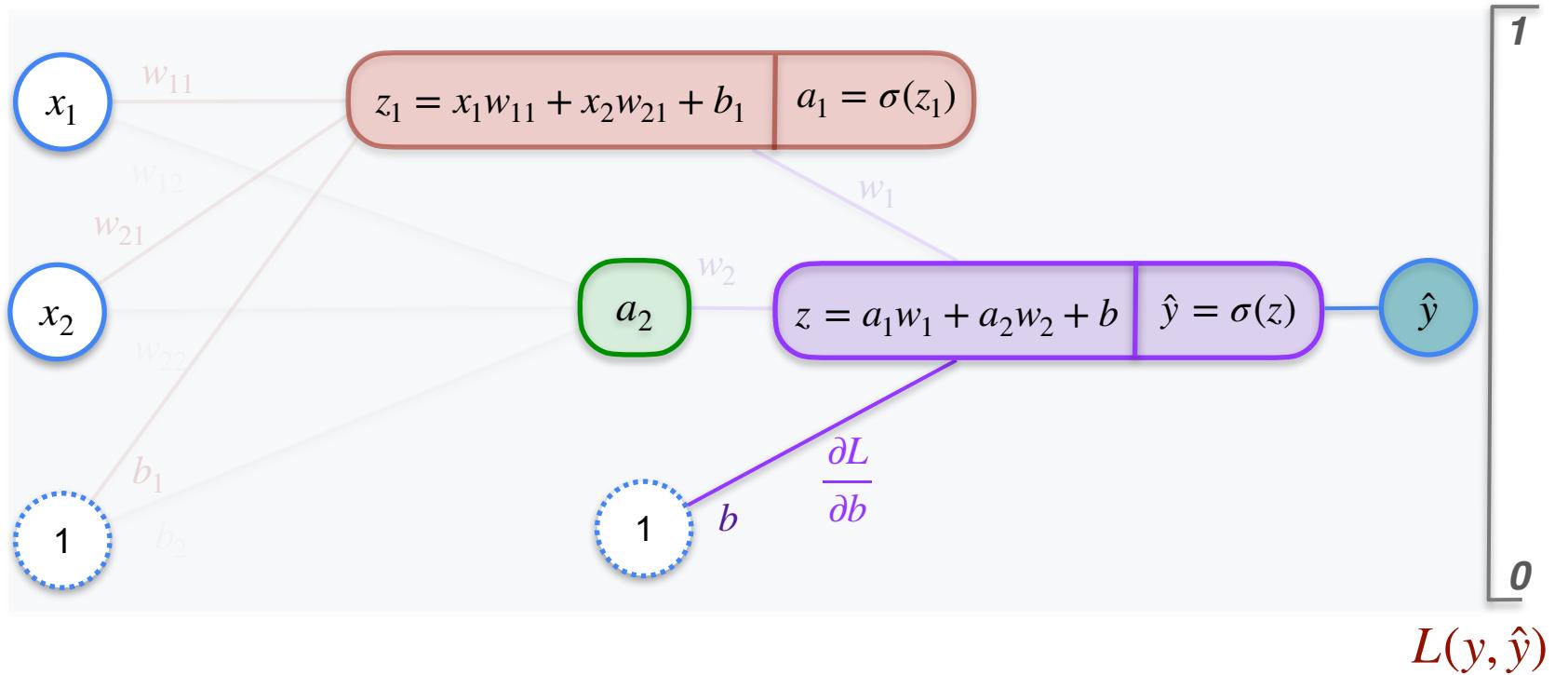
# 2,2,1 Neural Network



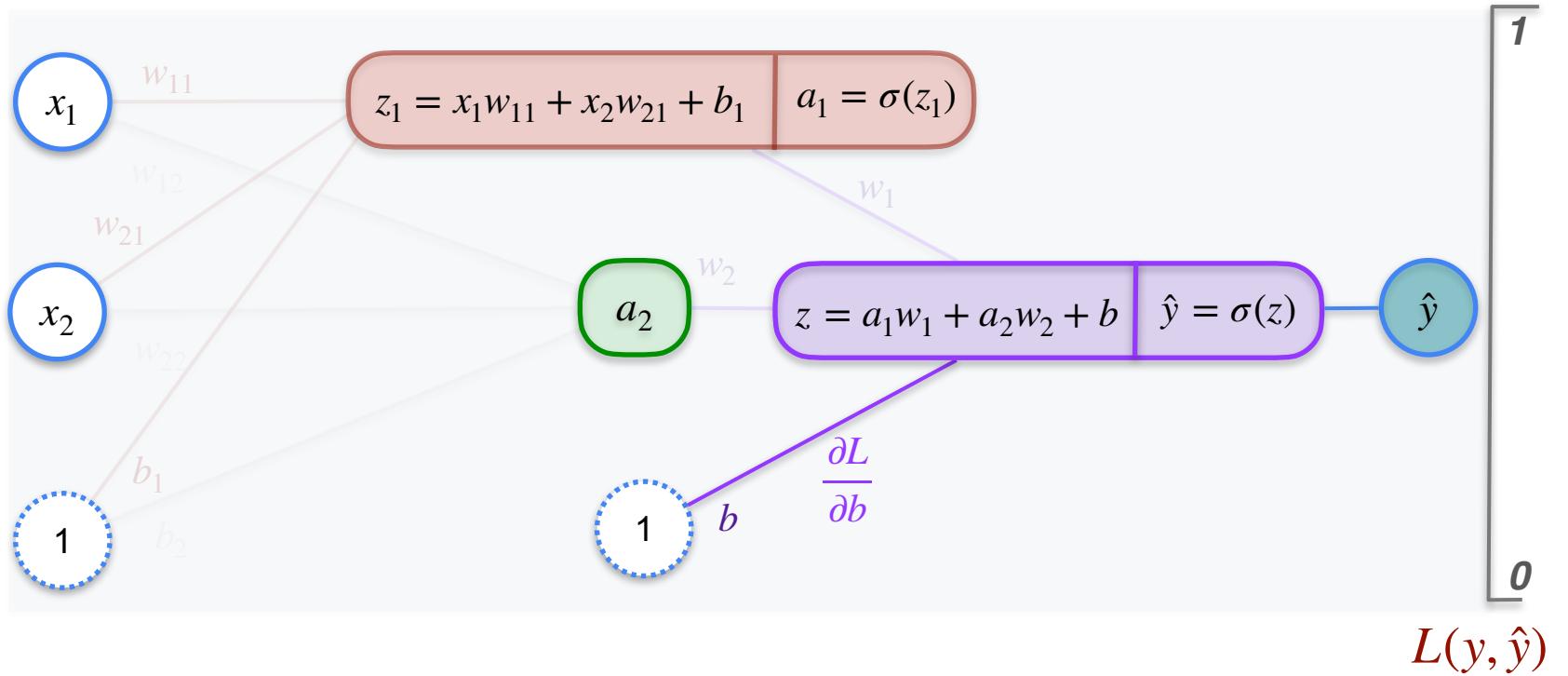
# 2,2,1 Neural Network



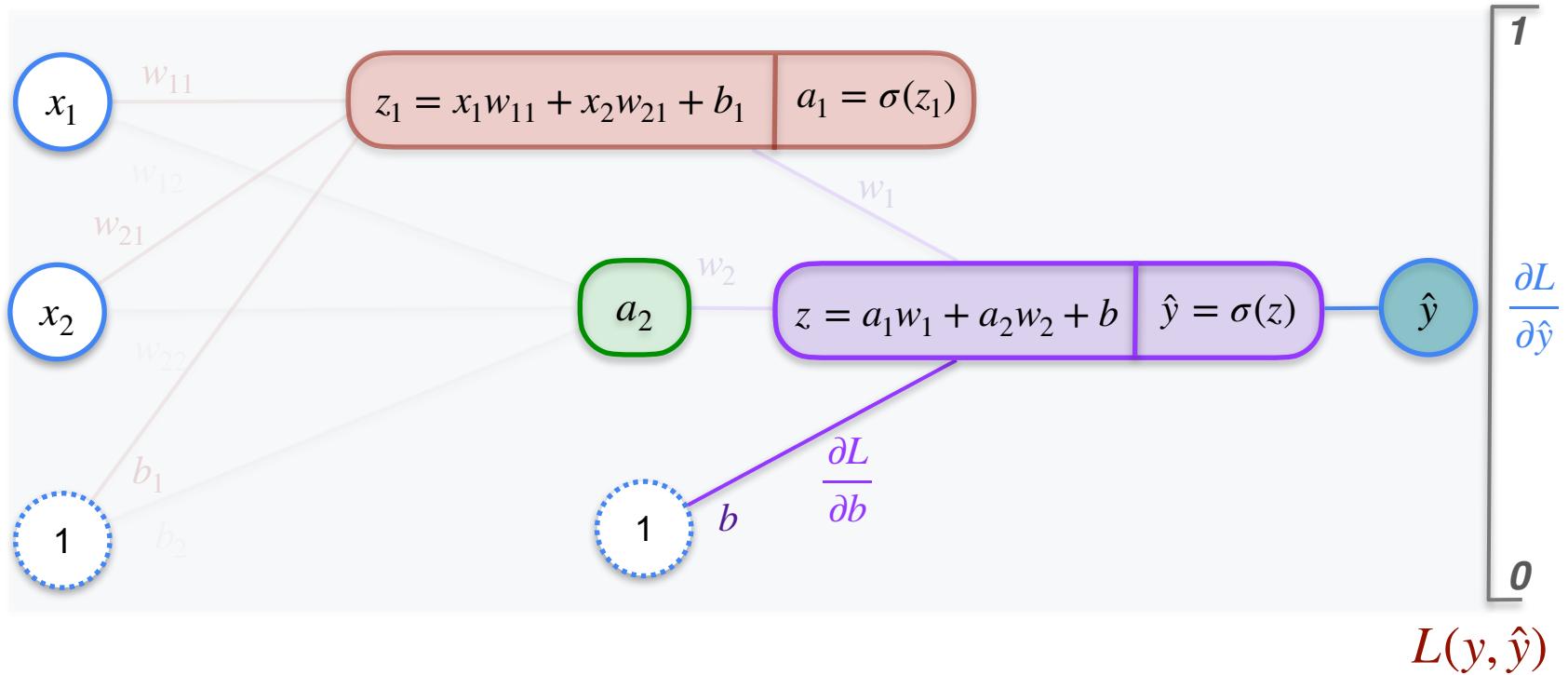
# 2,2,1 Neural Network



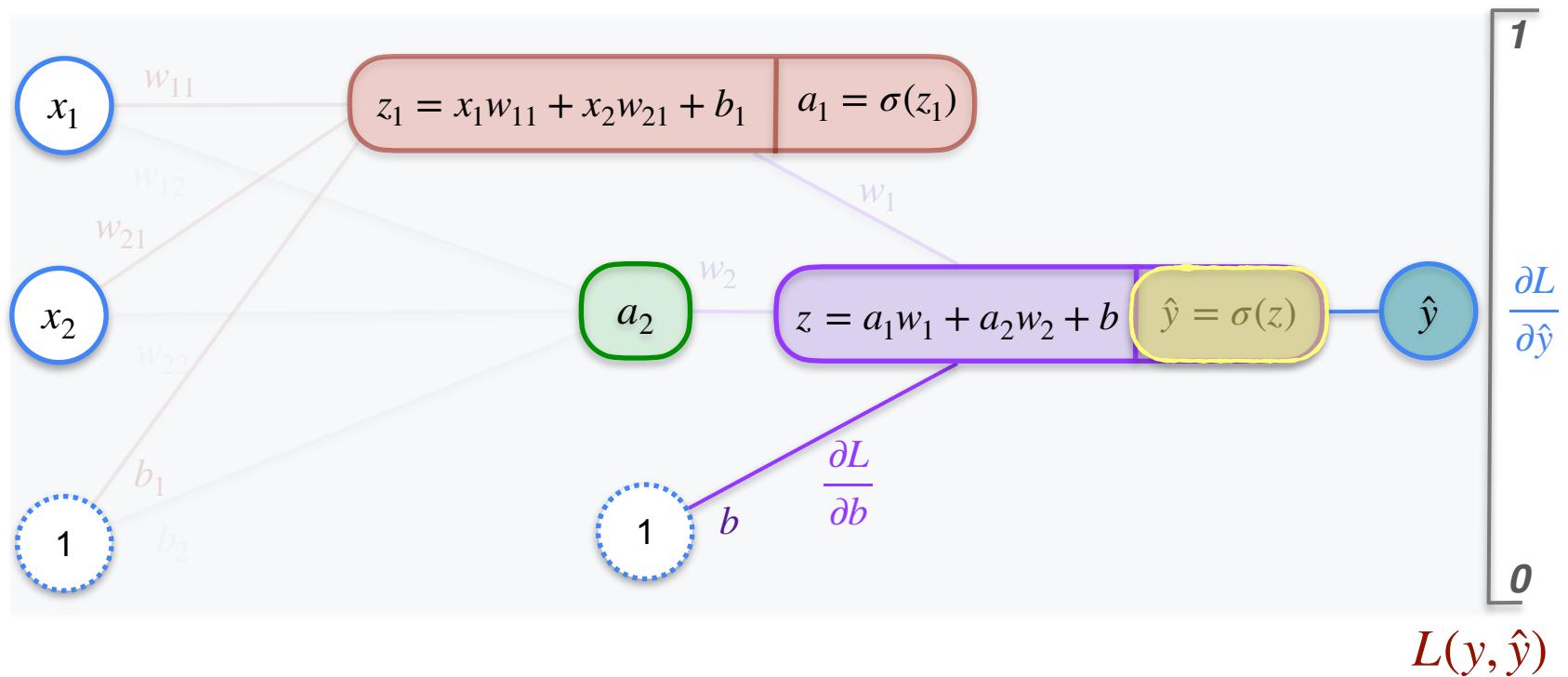
# 2,2,1 Neural Network



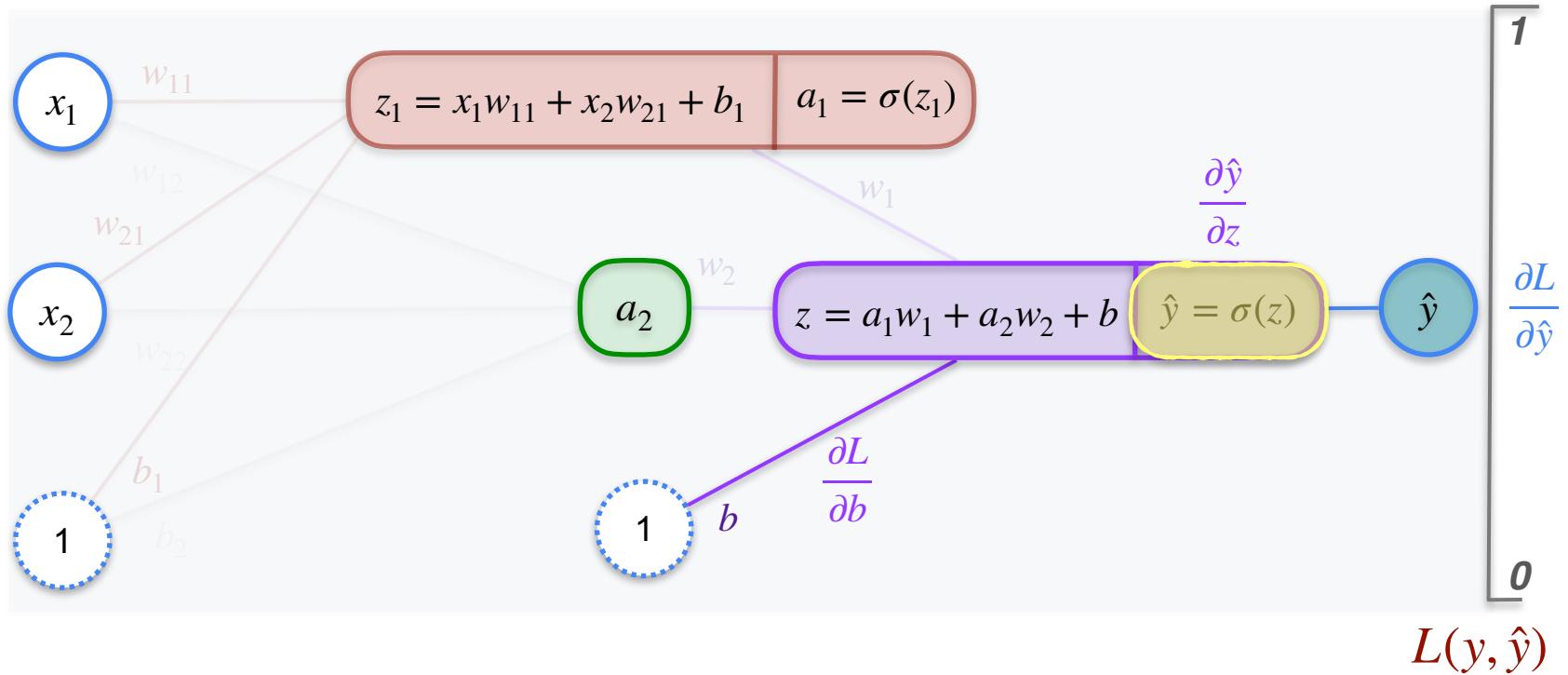
# 2,2,1 Neural Network



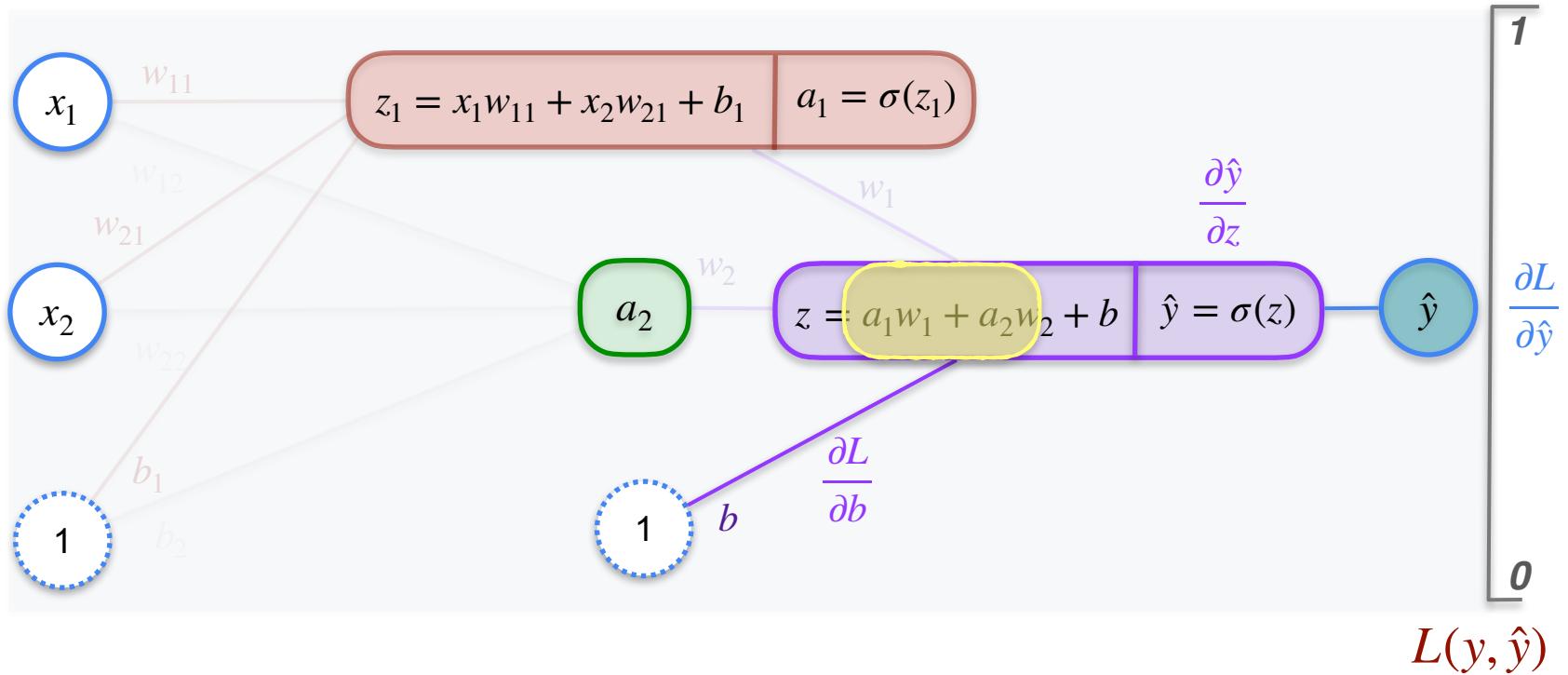
# 2,2,1 Neural Network



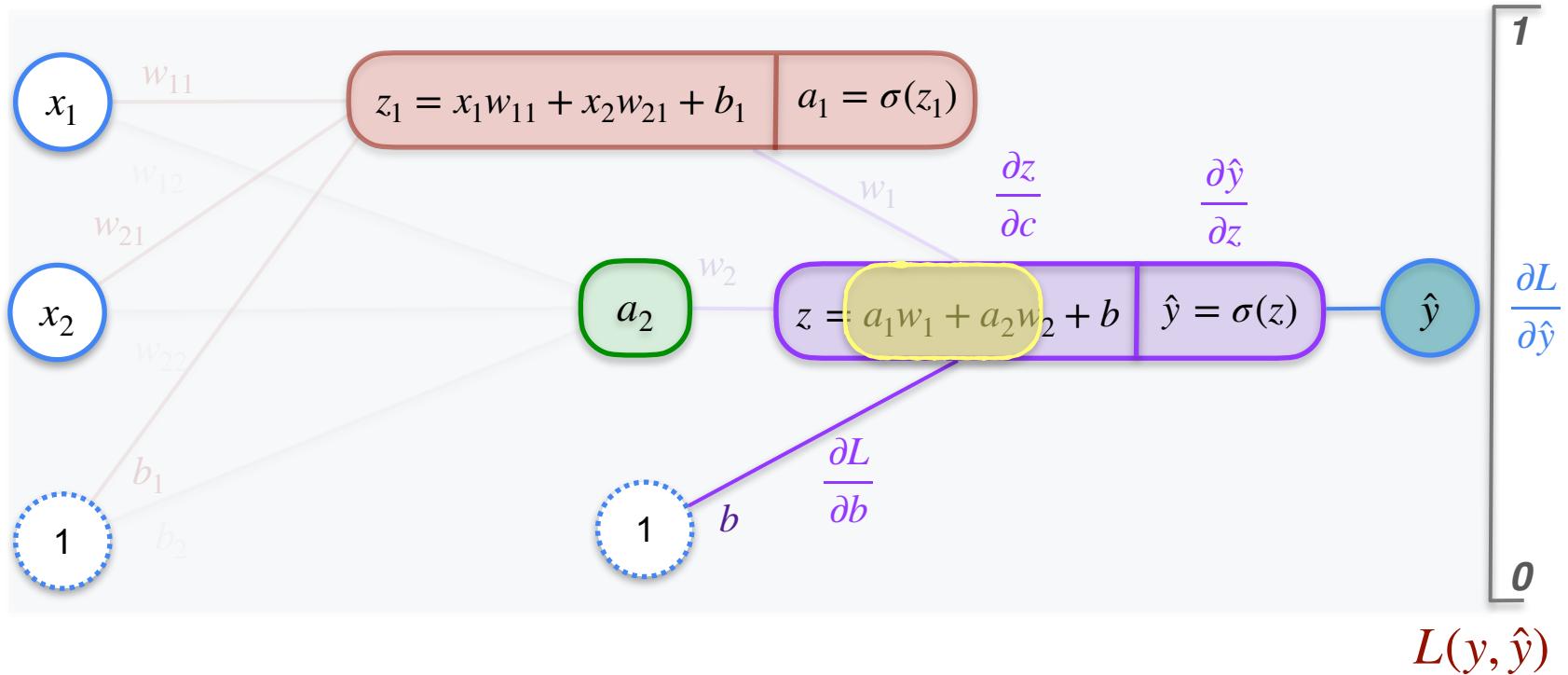
# 2,2,1 Neural Network



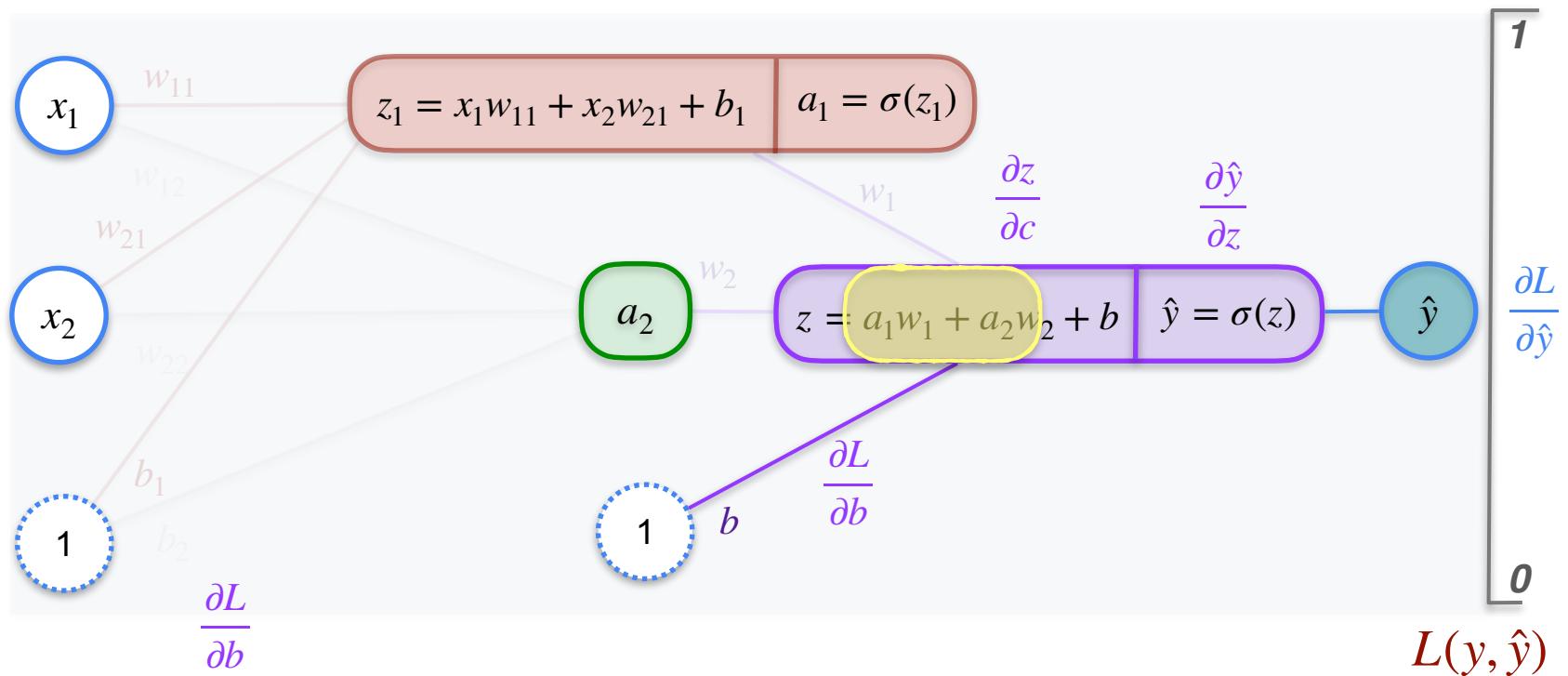
# 2,2,1 Neural Network



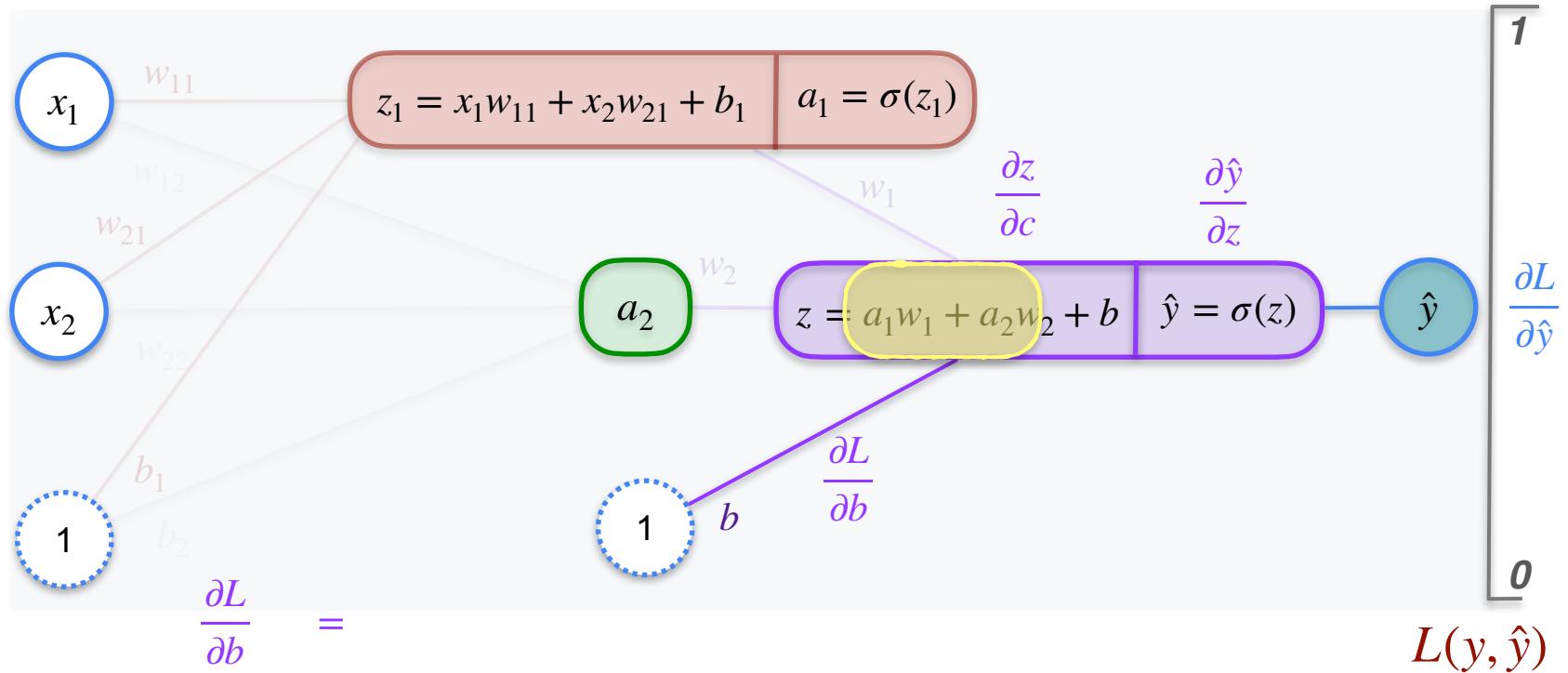
# 2,2,1 Neural Network



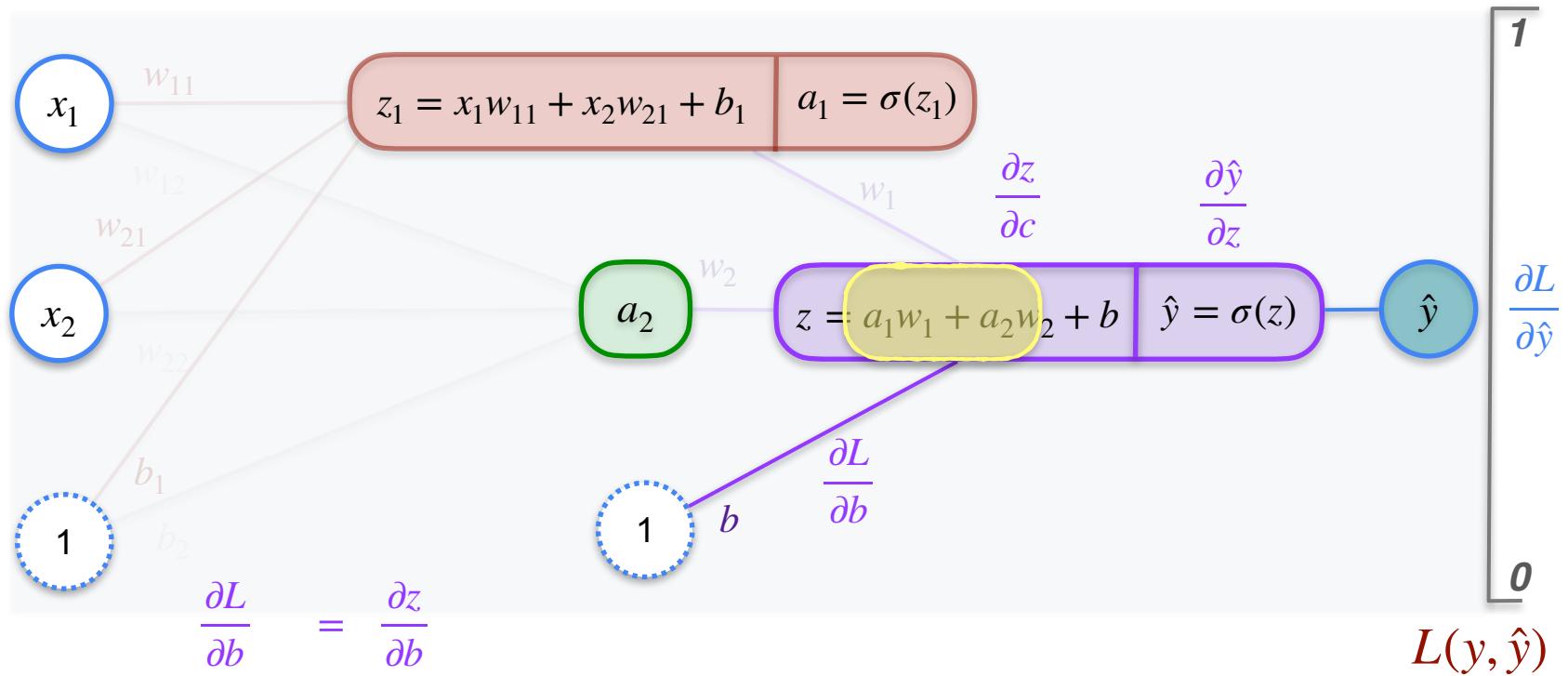
# 2,2,1 Neural Network



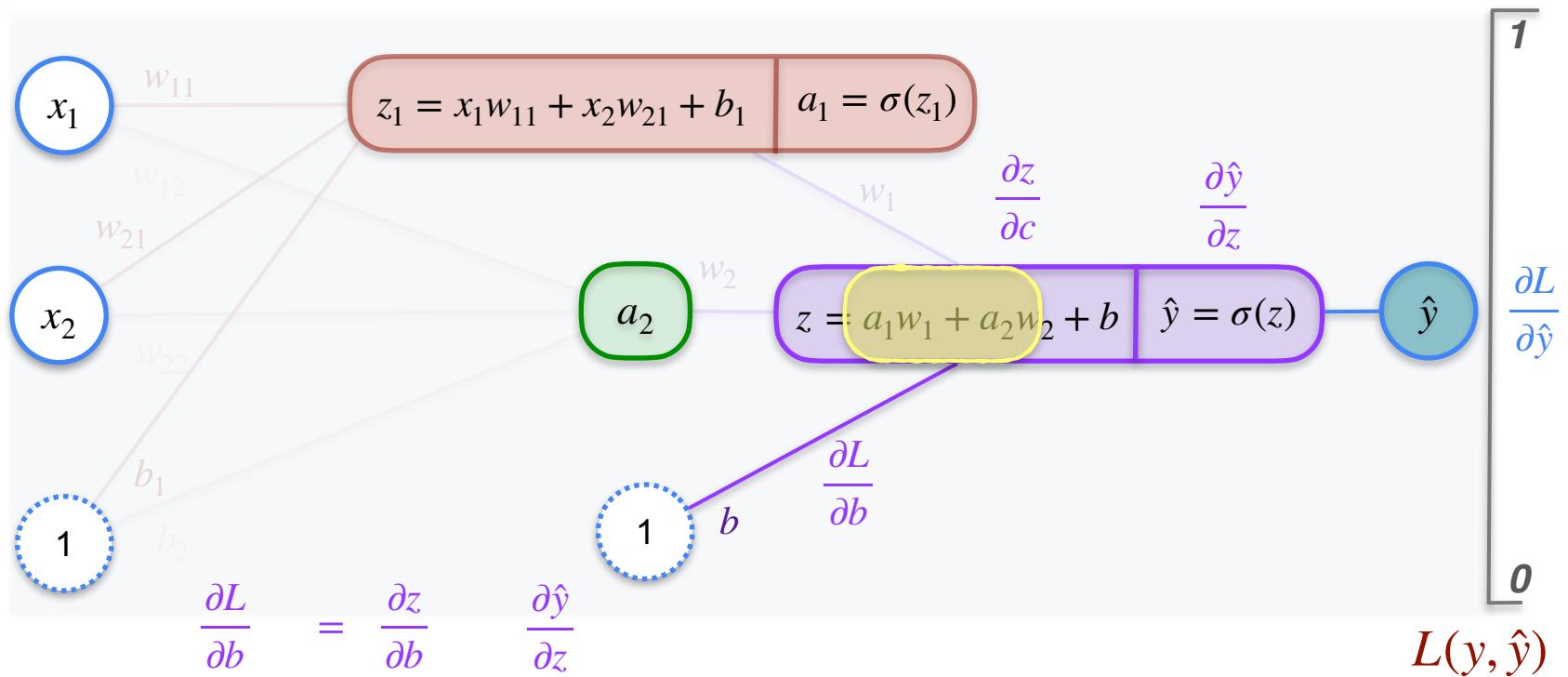
# 2,2,1 Neural Network



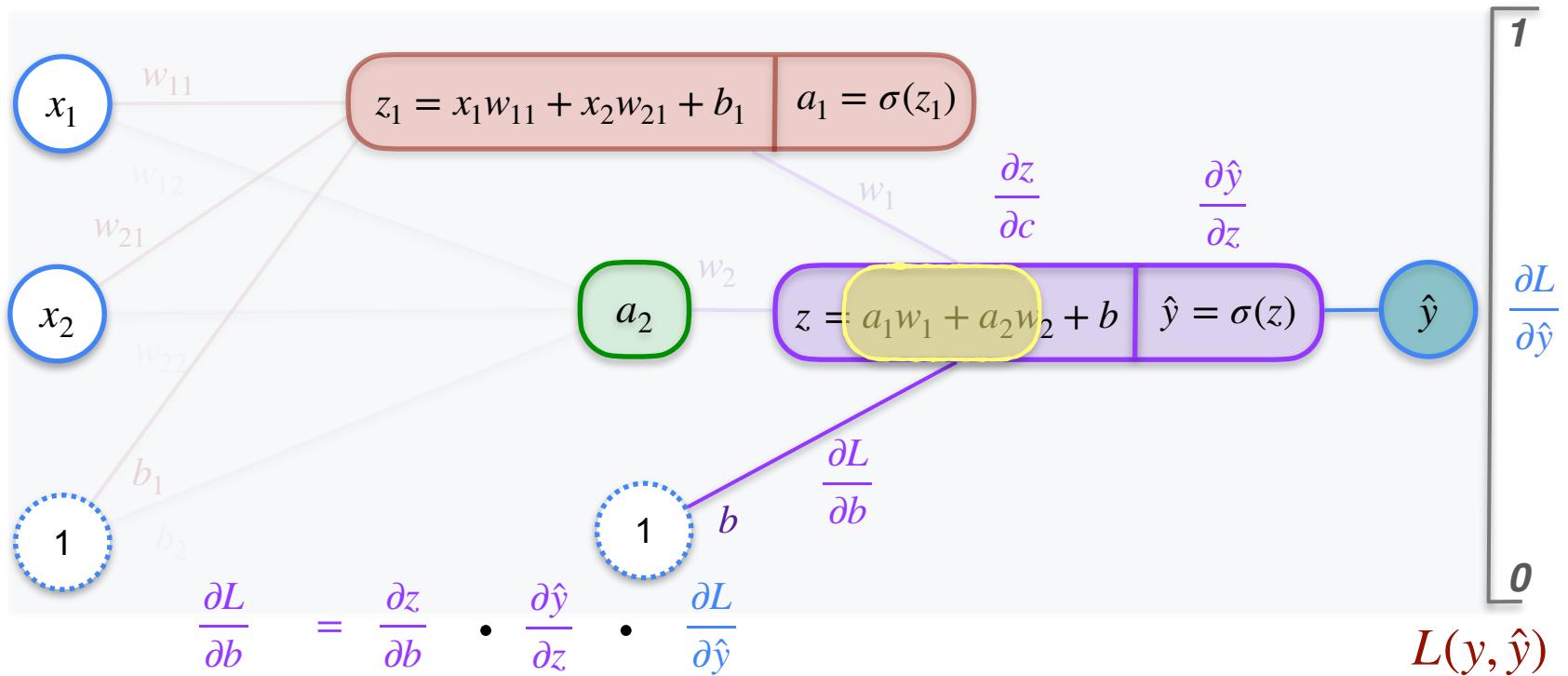
# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network



# 2,2,1 Neural Network

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1w_1 + a_2w_2 + b$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \quad \frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b} =$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b} = 1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b} = 1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b} = 1$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b} = 1 - \hat{y}(1 - \hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b} = 1 - \hat{y}(1 - \hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 - \hat{y}(1 - \hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 - \hat{y}(1 - \hat{y}) \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\hat{y}(1-\hat{y})}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -(y - \hat{y})$$

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -(y - \hat{y})$$

*to find optimal value of  $b$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -(y - \hat{y})$$

*Perform gradient descent with*

*to find optimal  
value of  $b$  that gives  
the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -(y - \hat{y})$$

*Perform gradient descent with*

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

*to find optimal value of  $b$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -(y - \hat{y})$$

*Perform gradient descent with*

$$b \rightarrow b - \alpha$$

*to find optimal value of  $b$  that gives the least error*

# 2,2,1 Neural Network

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

$$\frac{\partial L}{\partial b} = \frac{\partial z}{\partial b} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$\hat{y} = \sigma(z)$$

$$z = a_1 w_1 + a_2 w_2 + b$$

$$\frac{\partial L}{\partial b} = 1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y - \hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -(y - \hat{y})$$

*Perform gradient descent with*

$$b \rightarrow b - \alpha(-(y - \hat{y}))$$

*to find optimal value of  $b$  that gives the least error*



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Gradient Descent and Backpropagation

# Back Propagation Introduction

# Back Propagation Introduction



# Back Propagation Introduction



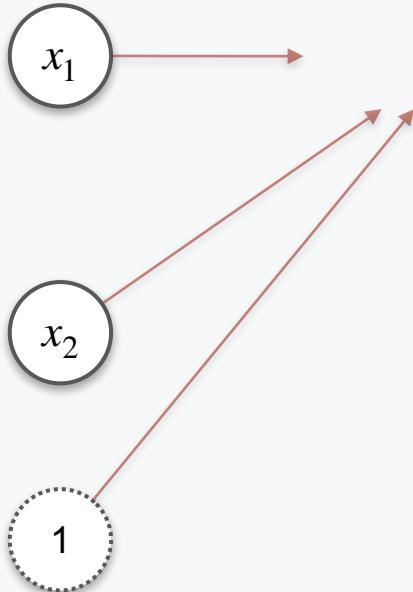
# Back Propagation Introduction

$x_1$

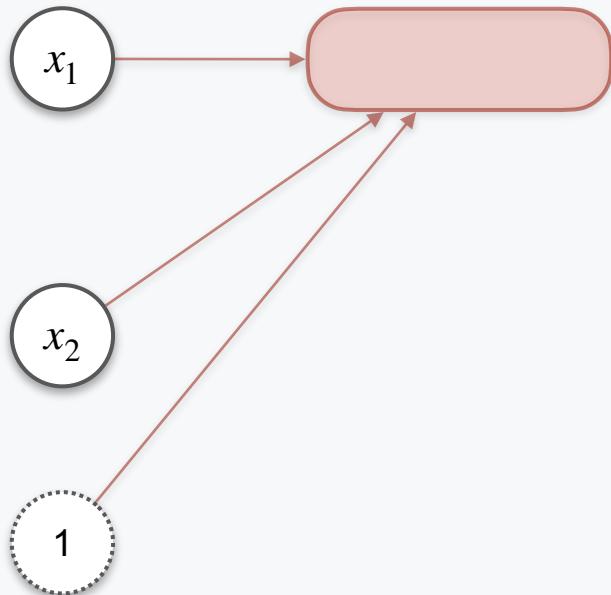
$x_2$

1

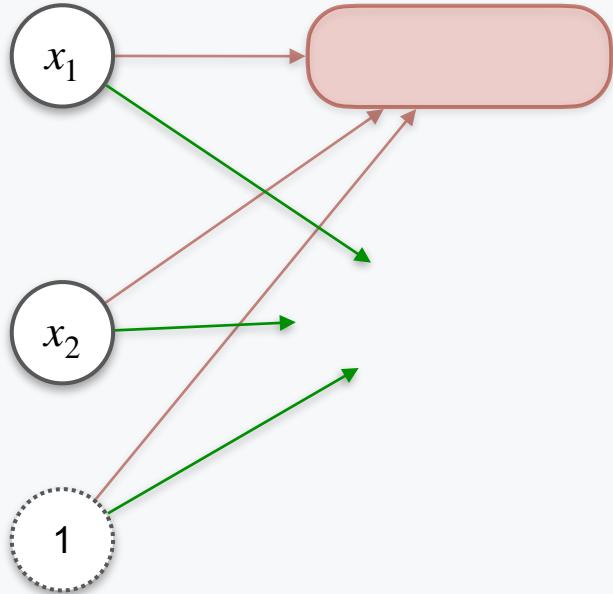
# Back Propagation Introduction



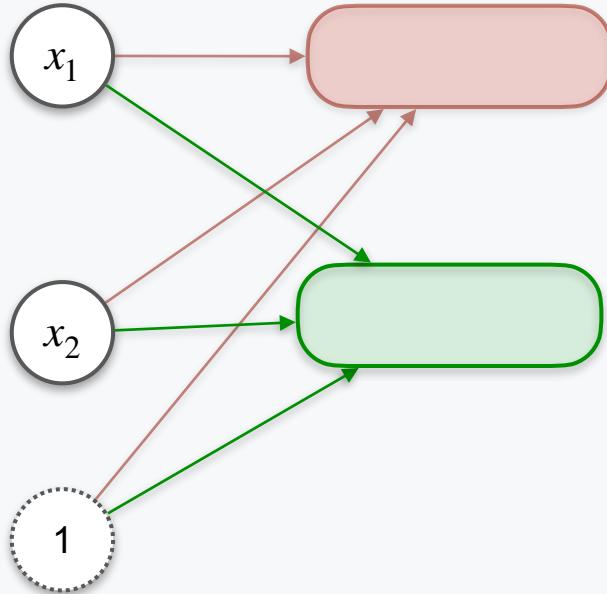
# Back Propagation Introduction



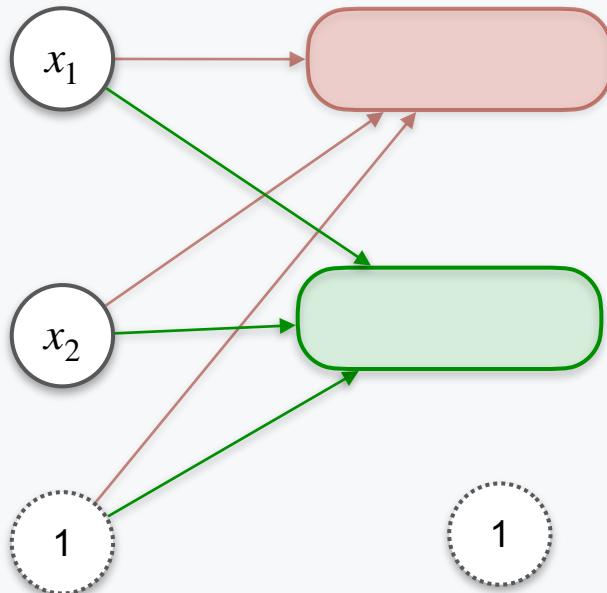
# Back Propagation Introduction



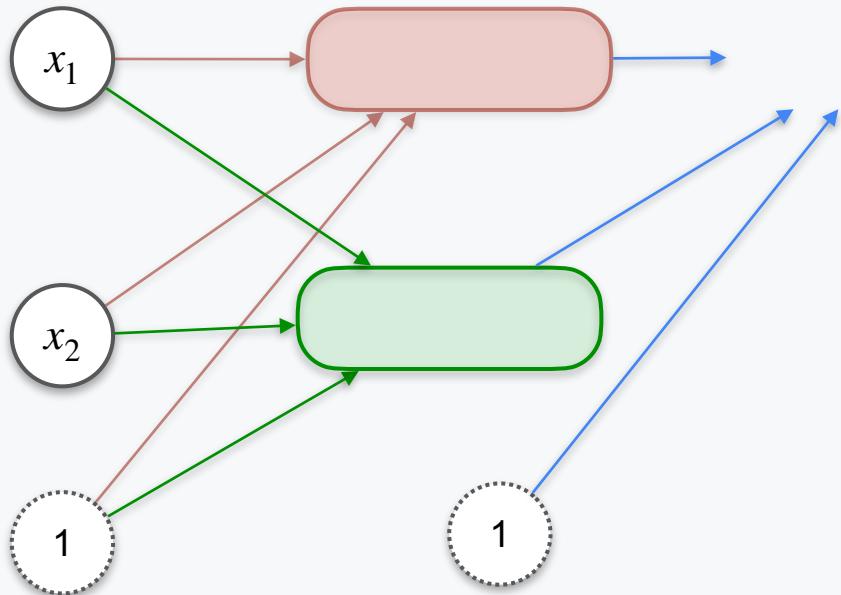
# Back Propagation Introduction



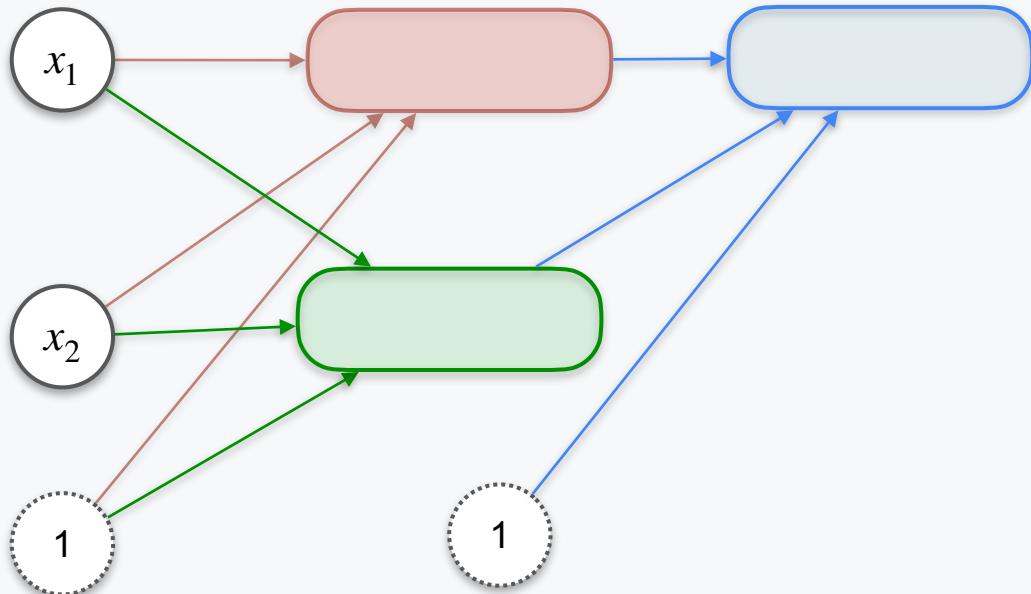
# Back Propagation Introduction



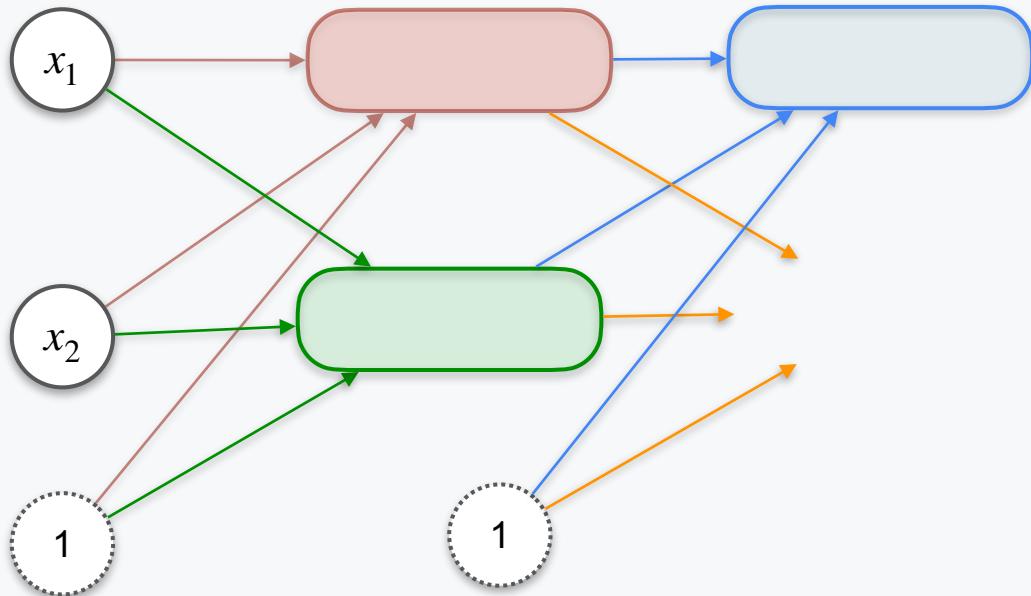
# Back Propagation Introduction



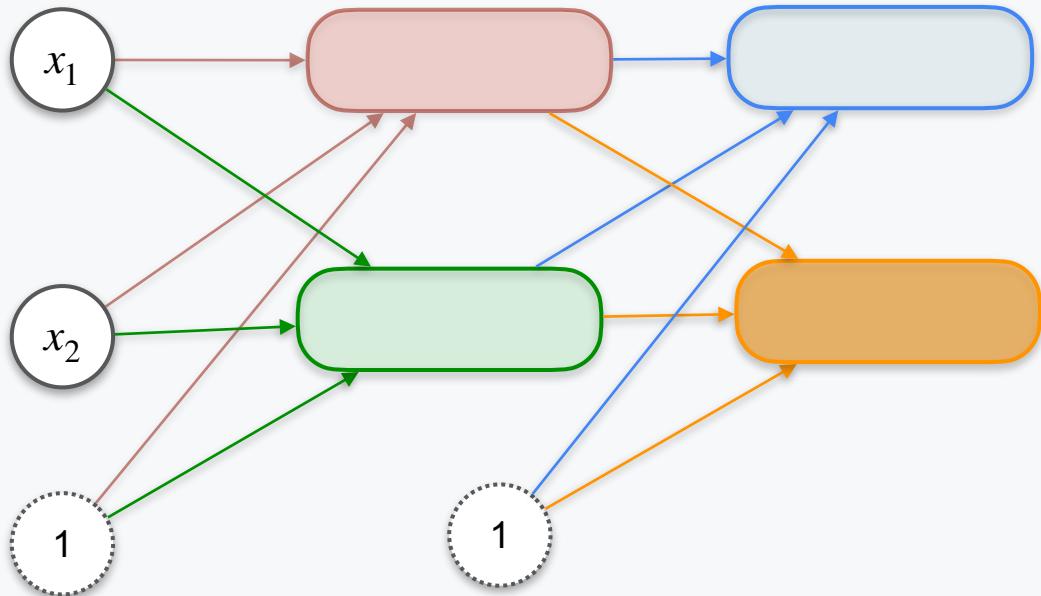
# Back Propagation Introduction



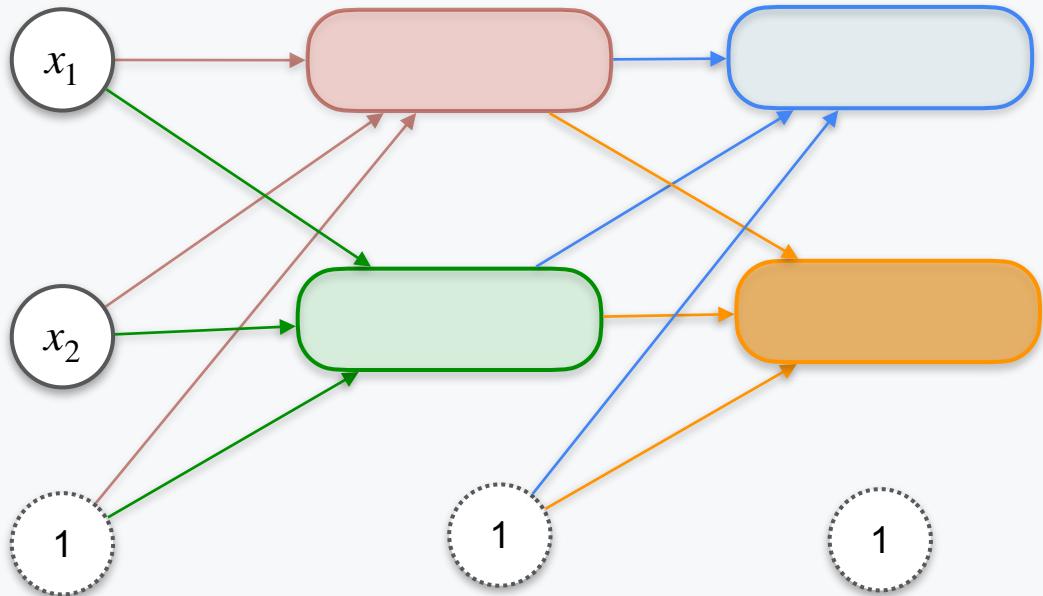
# Back Propagation Introduction



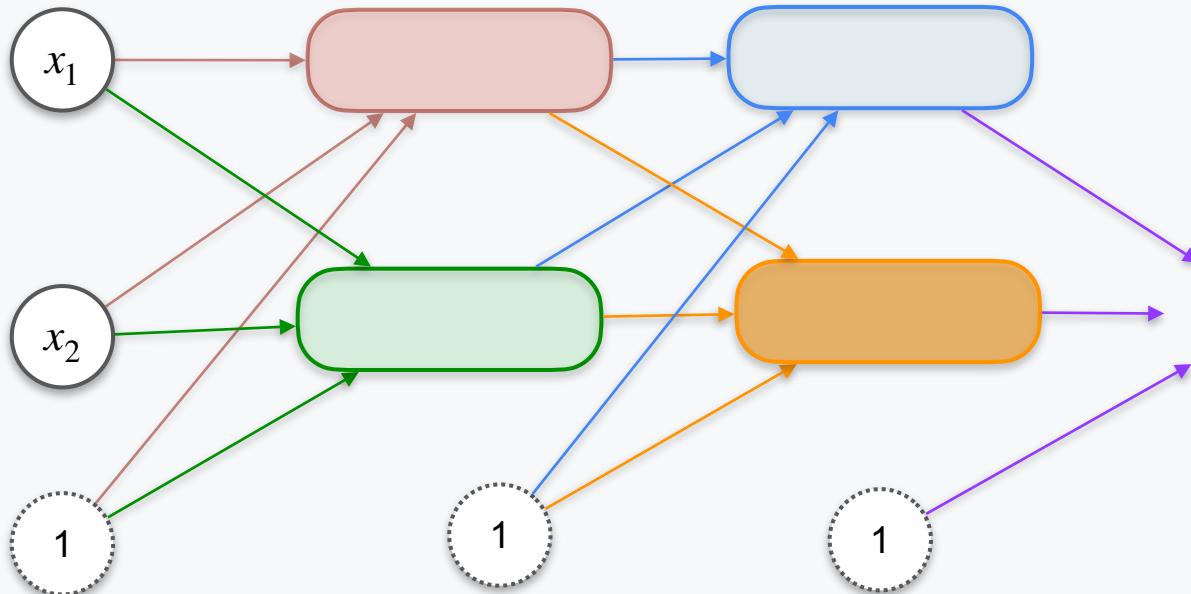
# Back Propagation Introduction



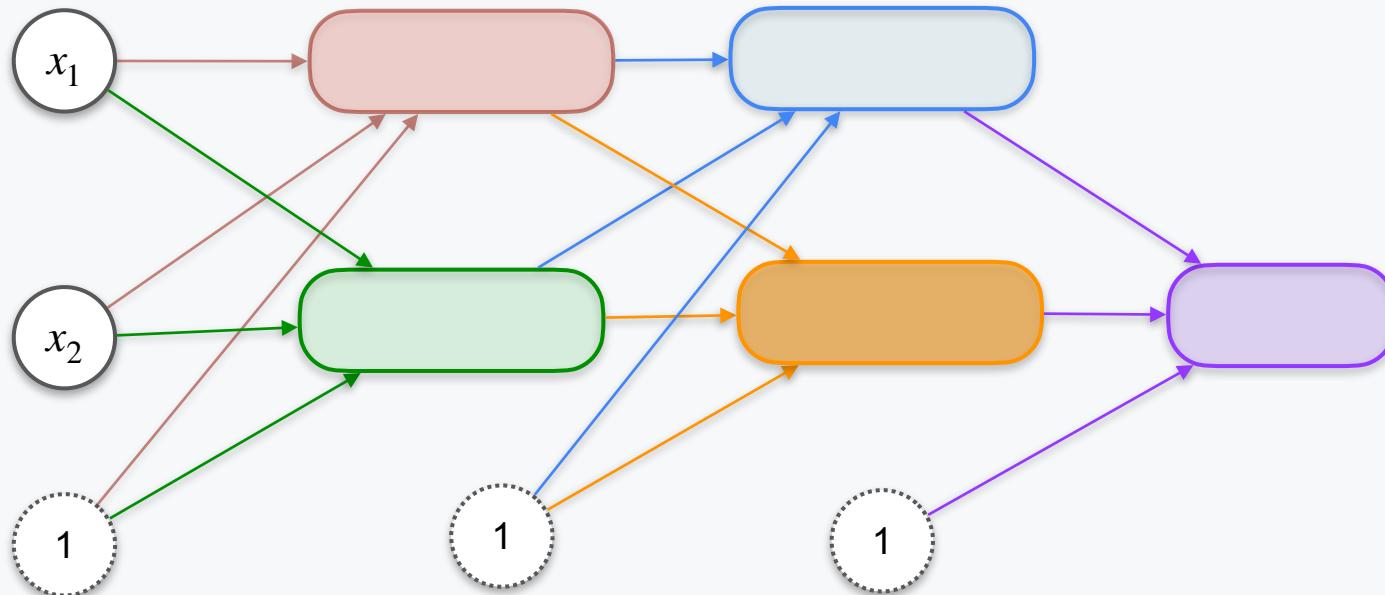
# Back Propagation Introduction



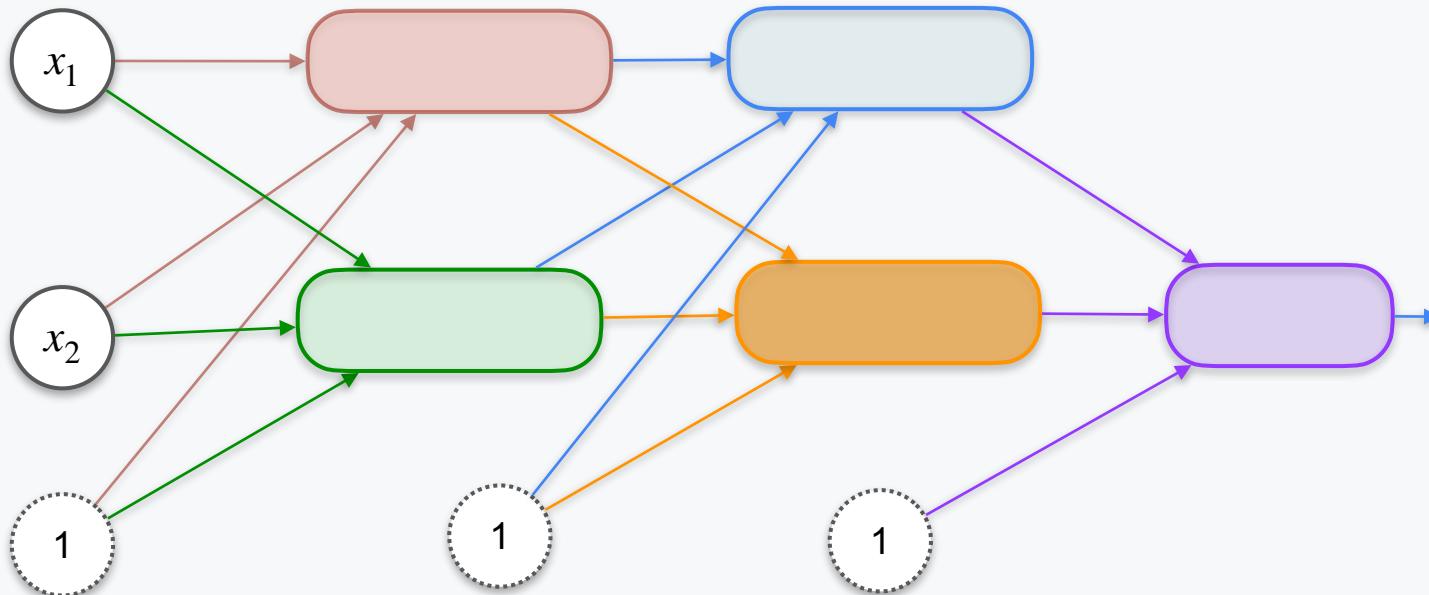
# Back Propagation Introduction



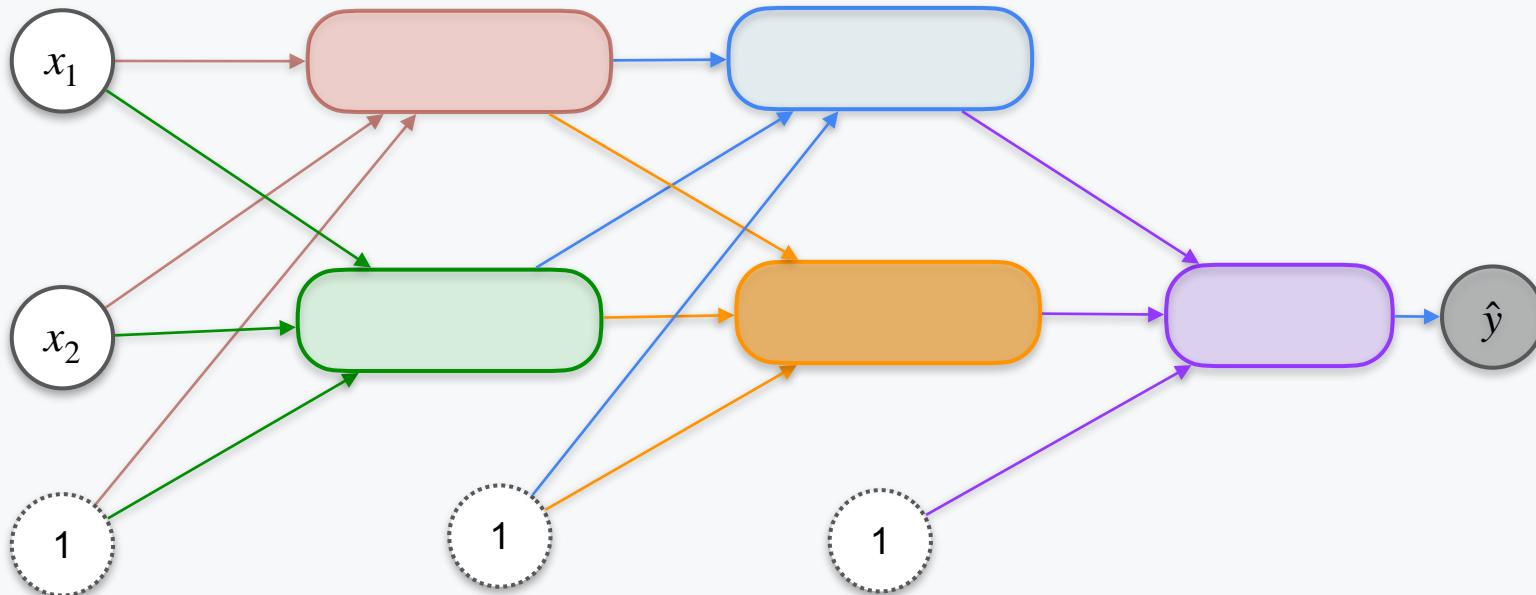
# Back Propagation Introduction



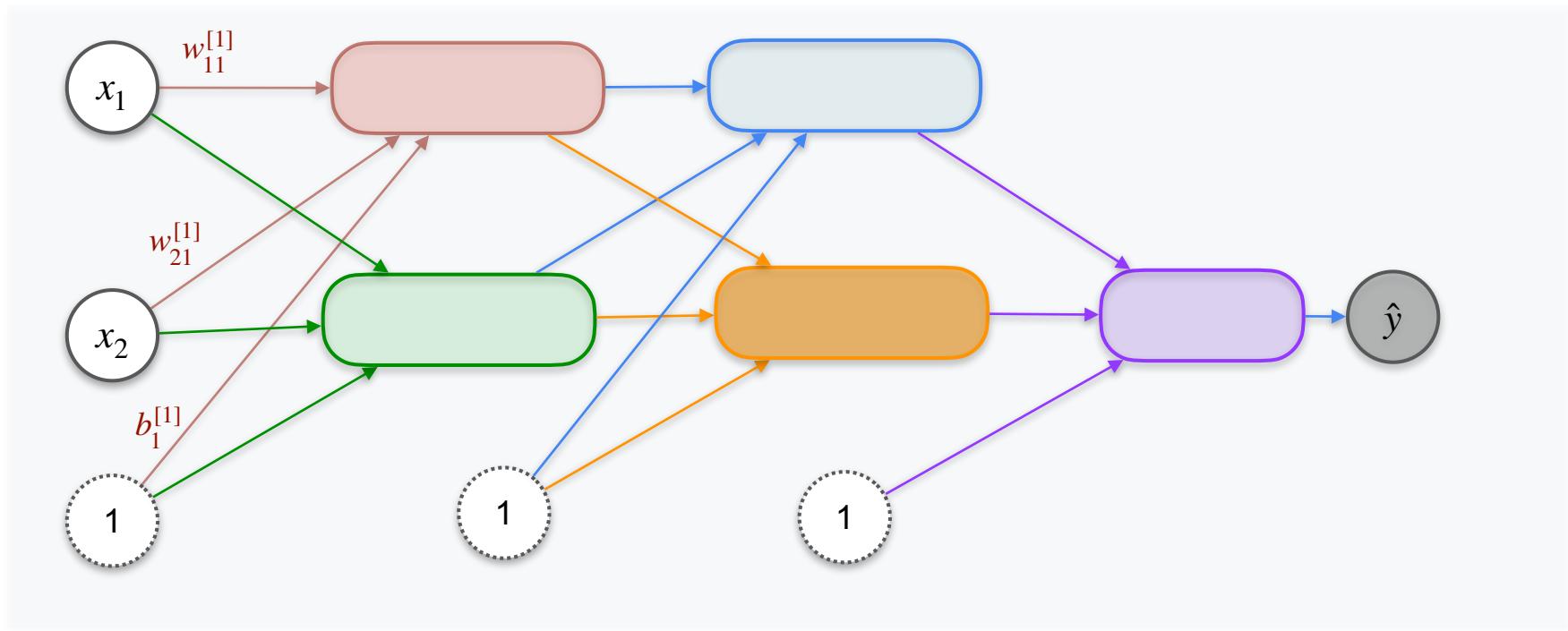
# Back Propagation Introduction



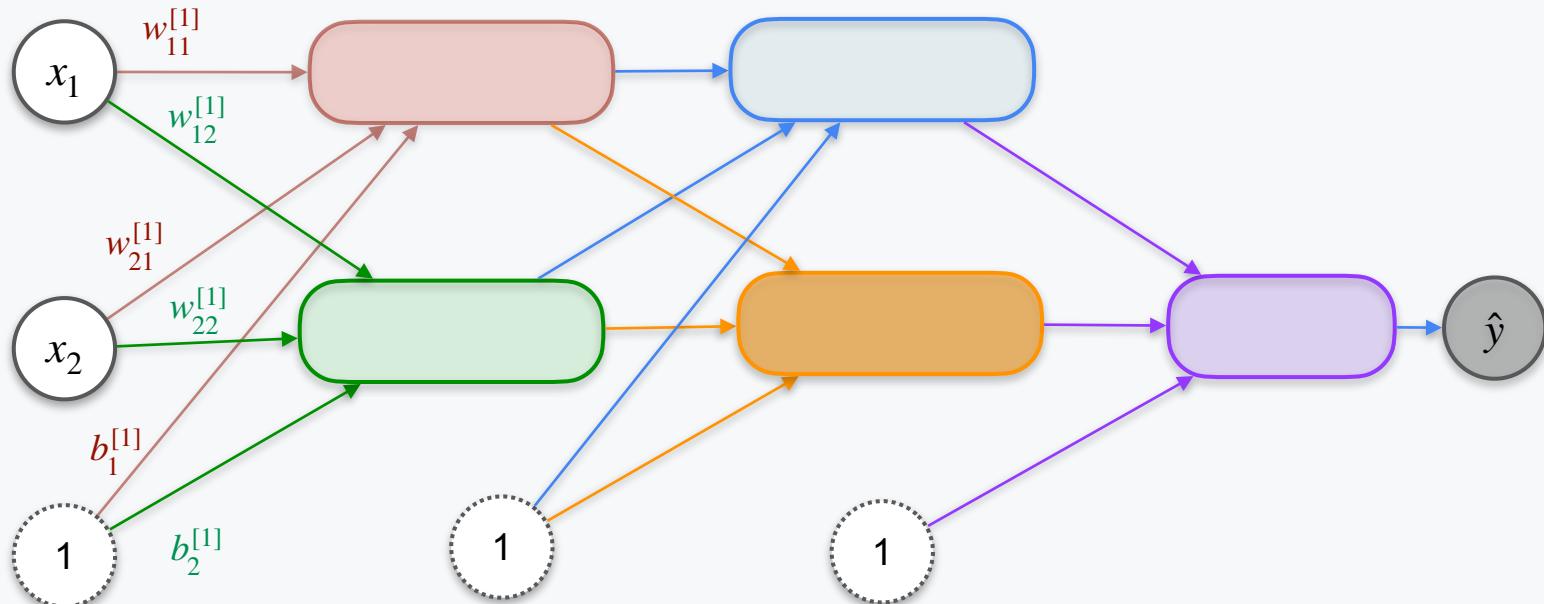
# Back Propagation Introduction



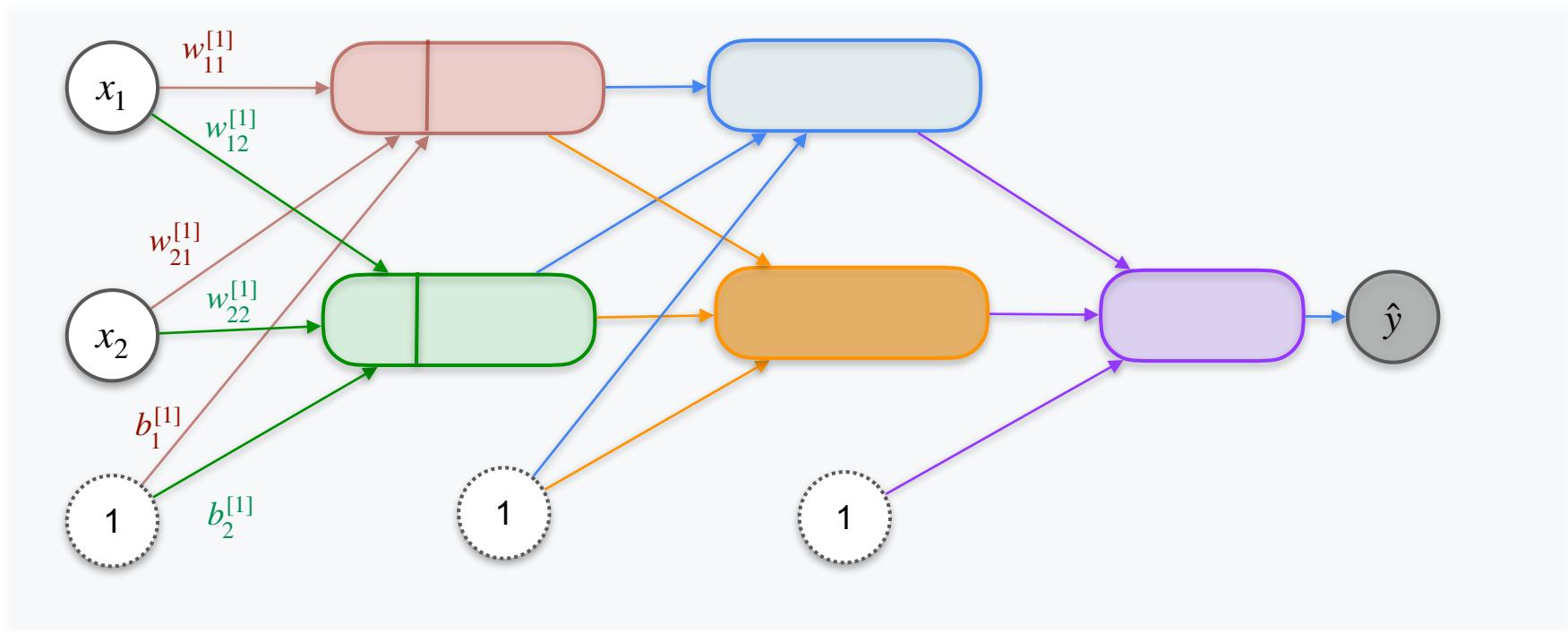
# Back Propagation Introduction



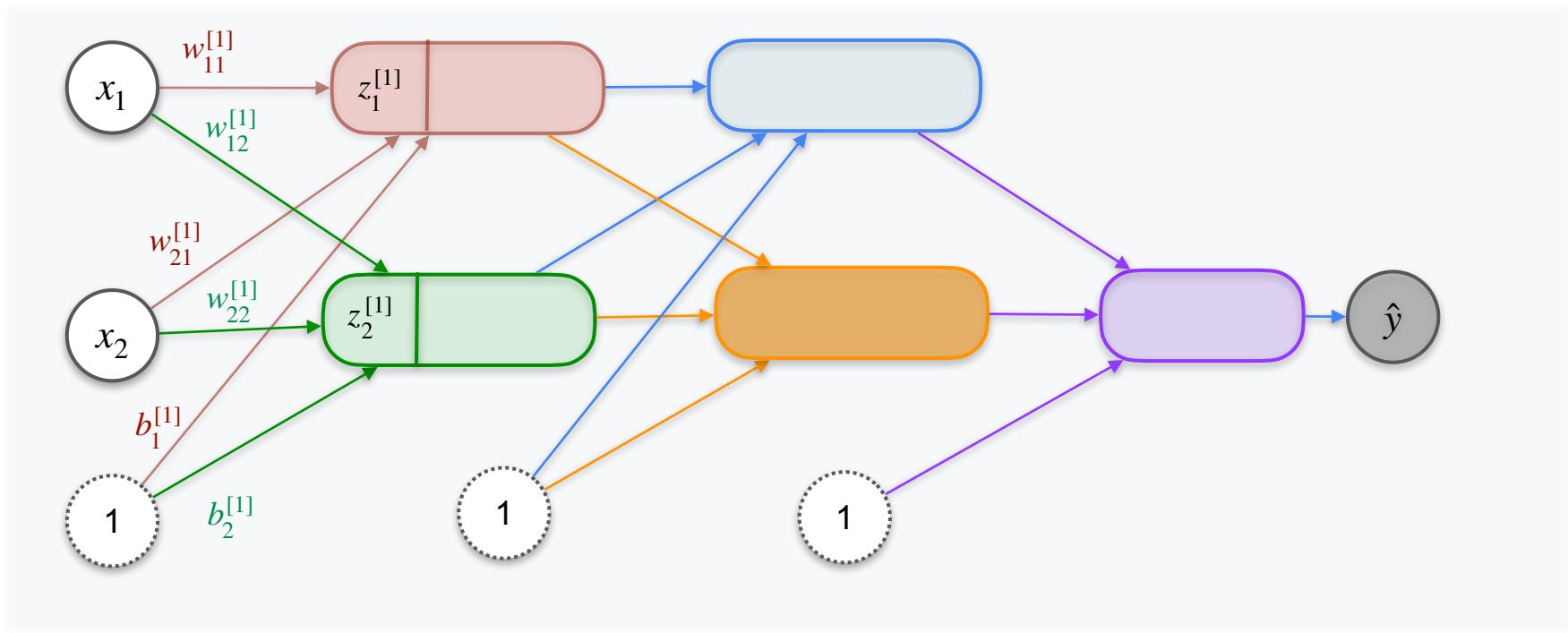
# Back Propagation Introduction



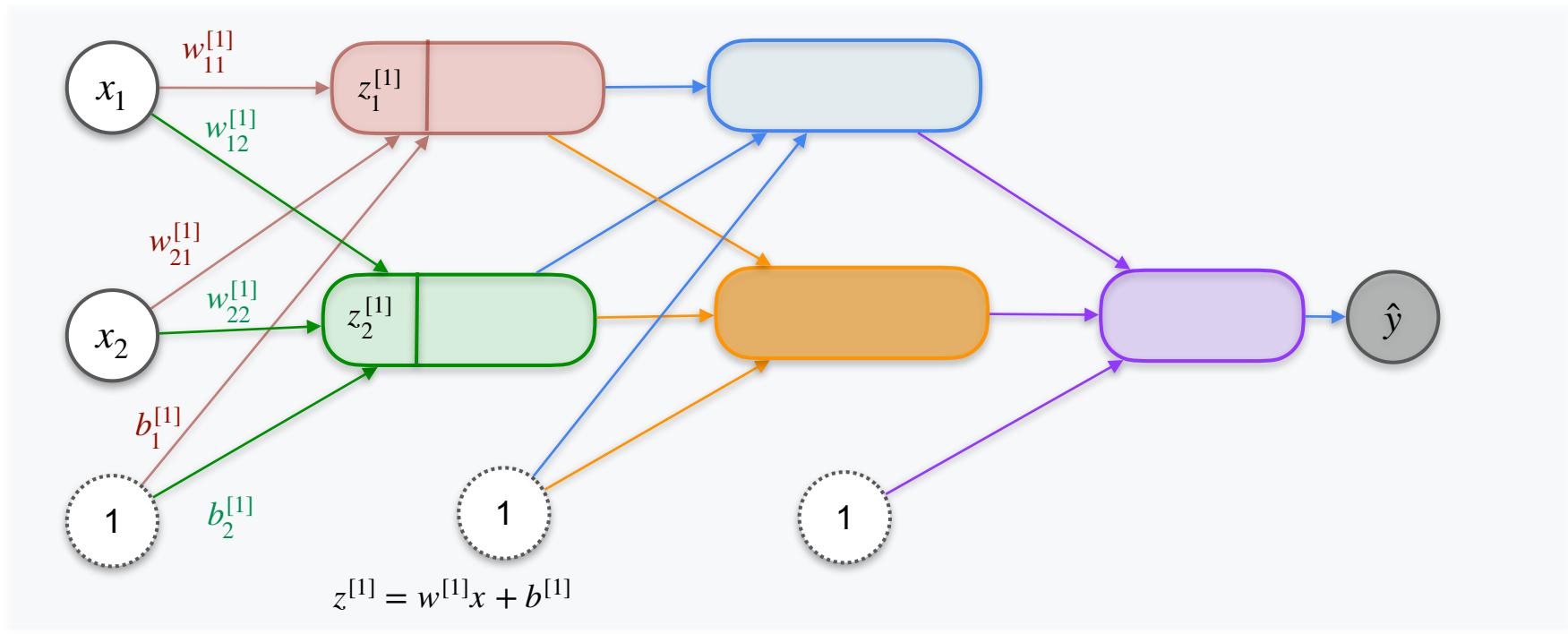
# Back Propagation Introduction



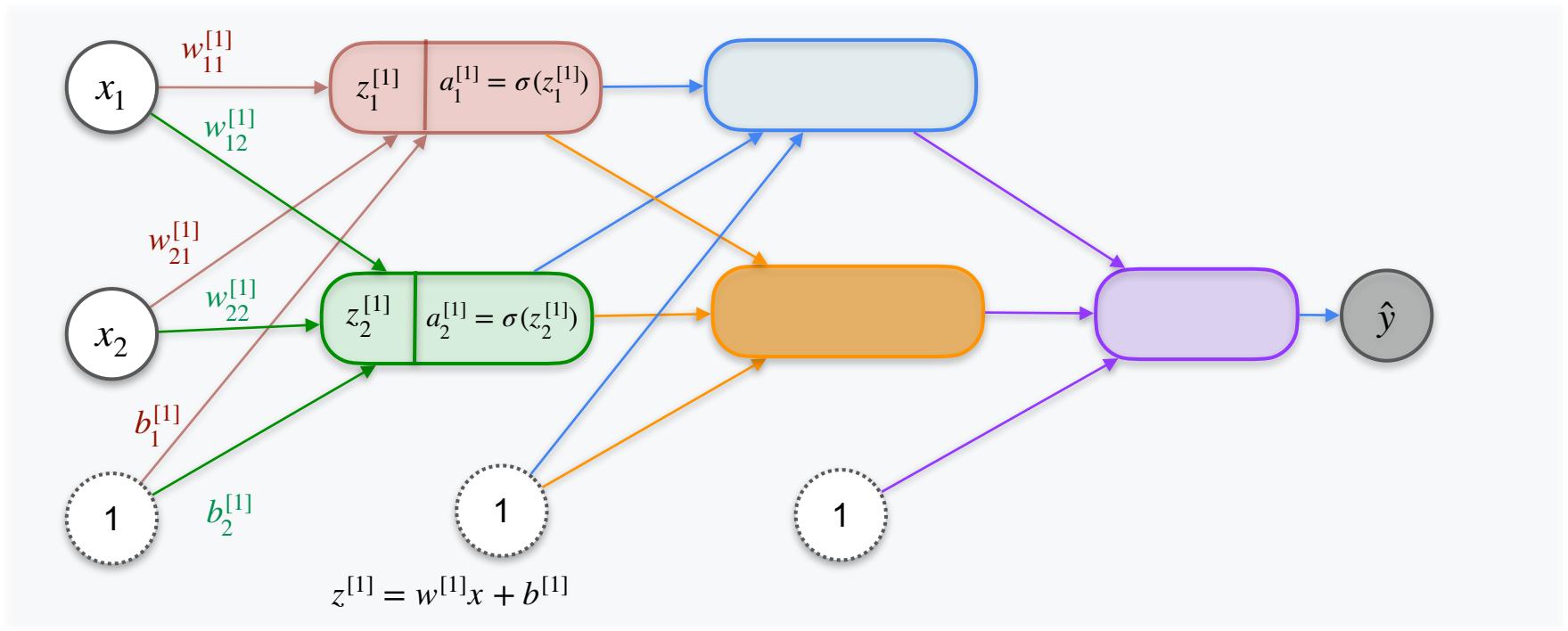
# Back Propagation Introduction



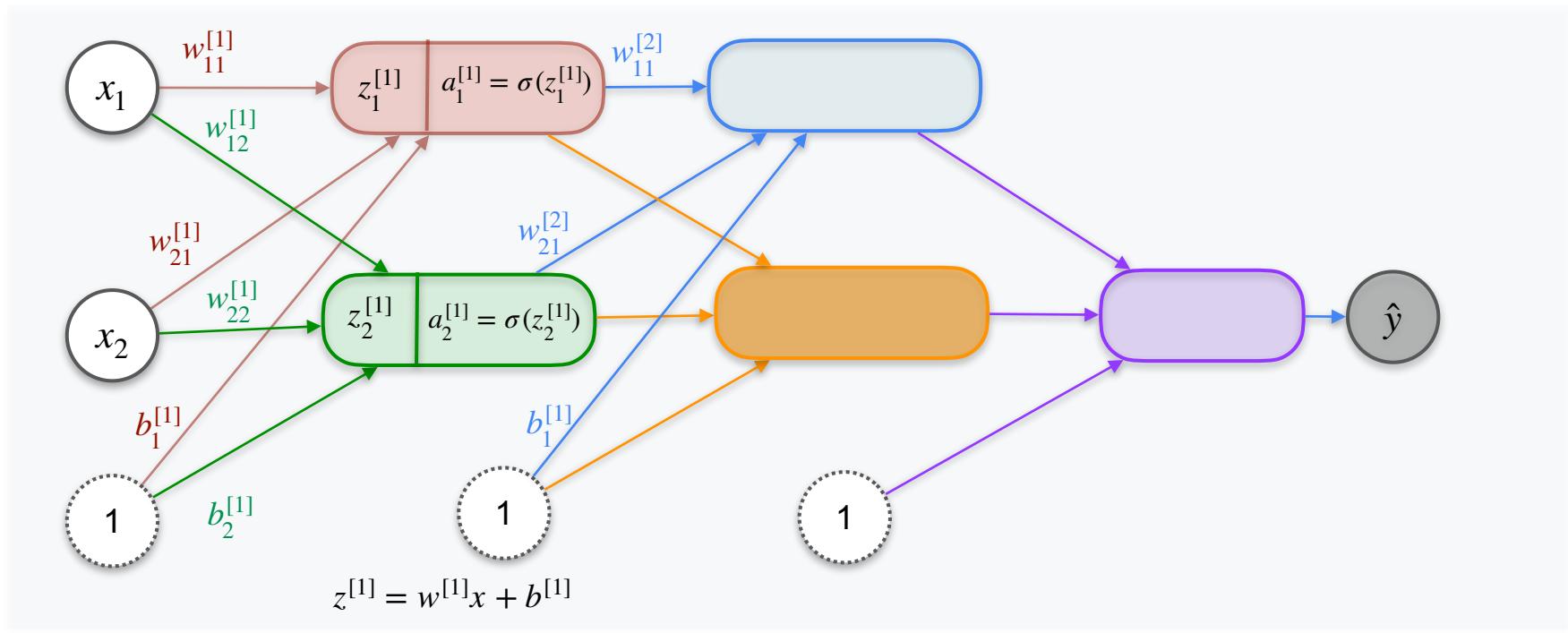
# Back Propagation Introduction



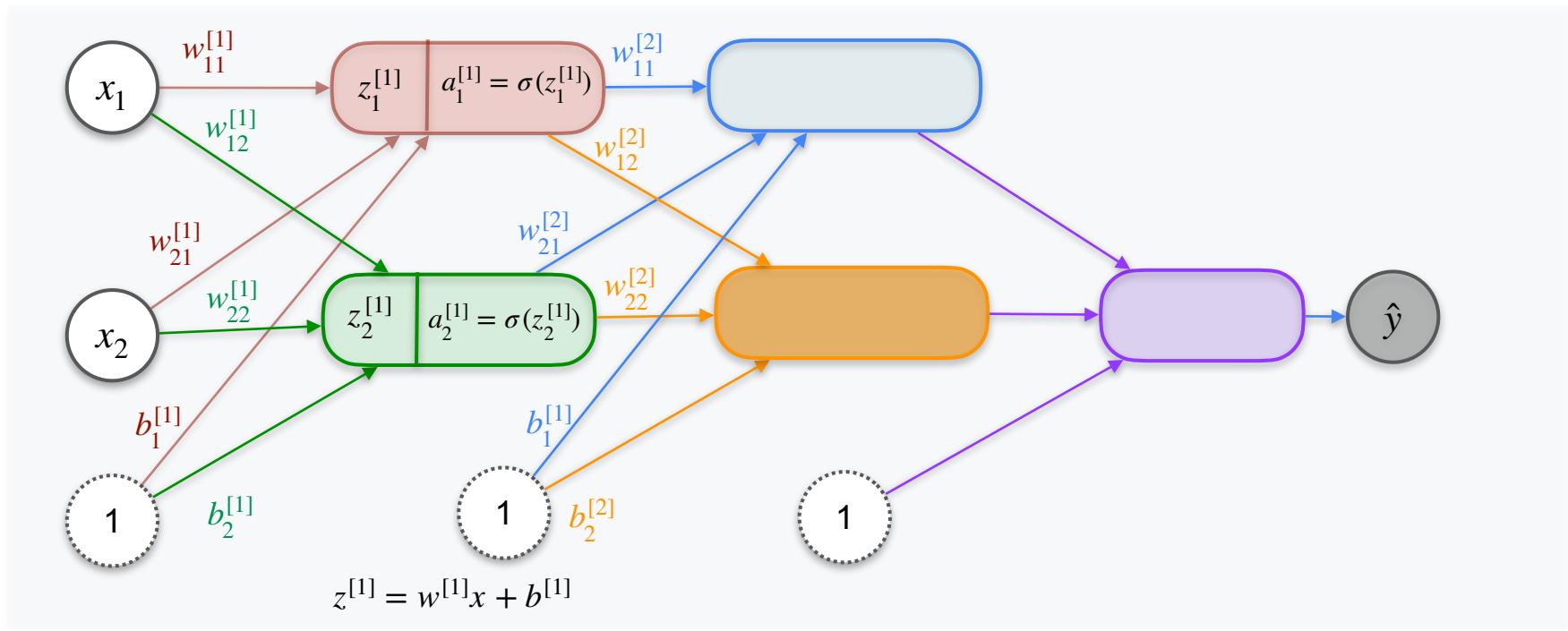
# Back Propagation Introduction



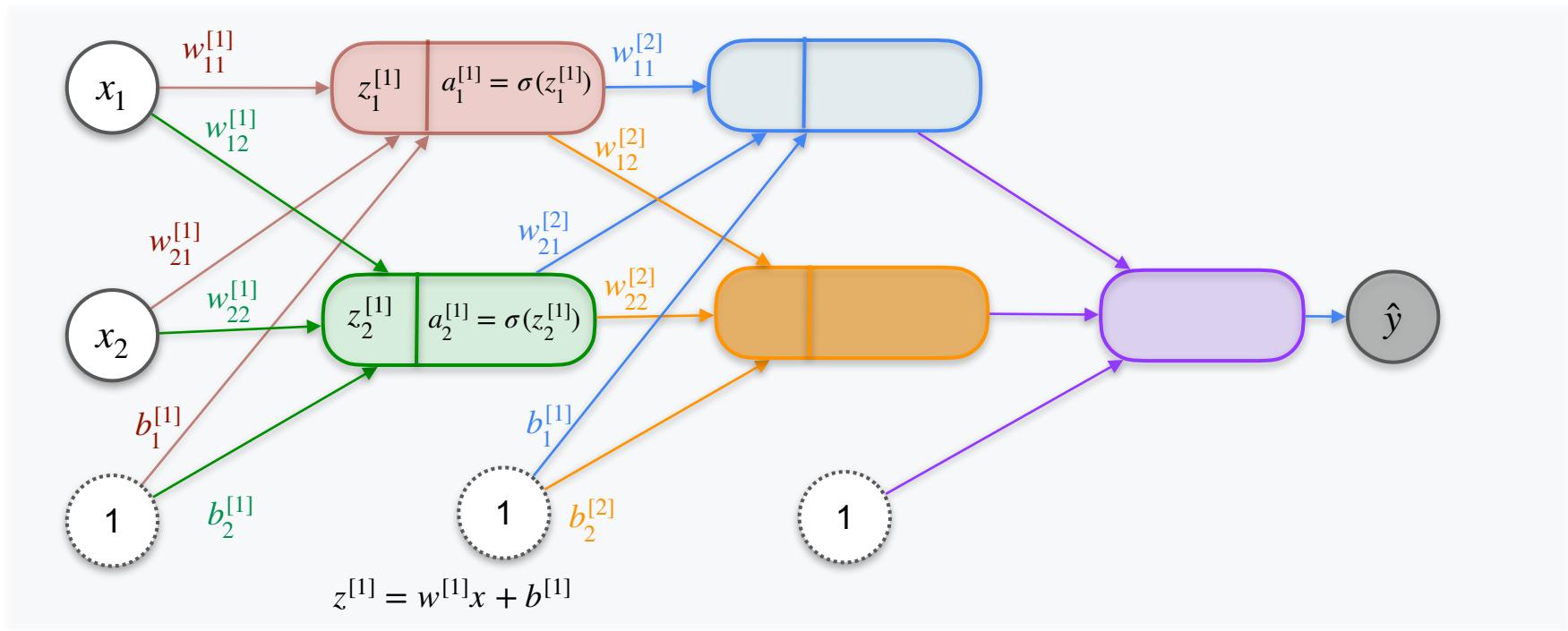
# Back Propagation Introduction



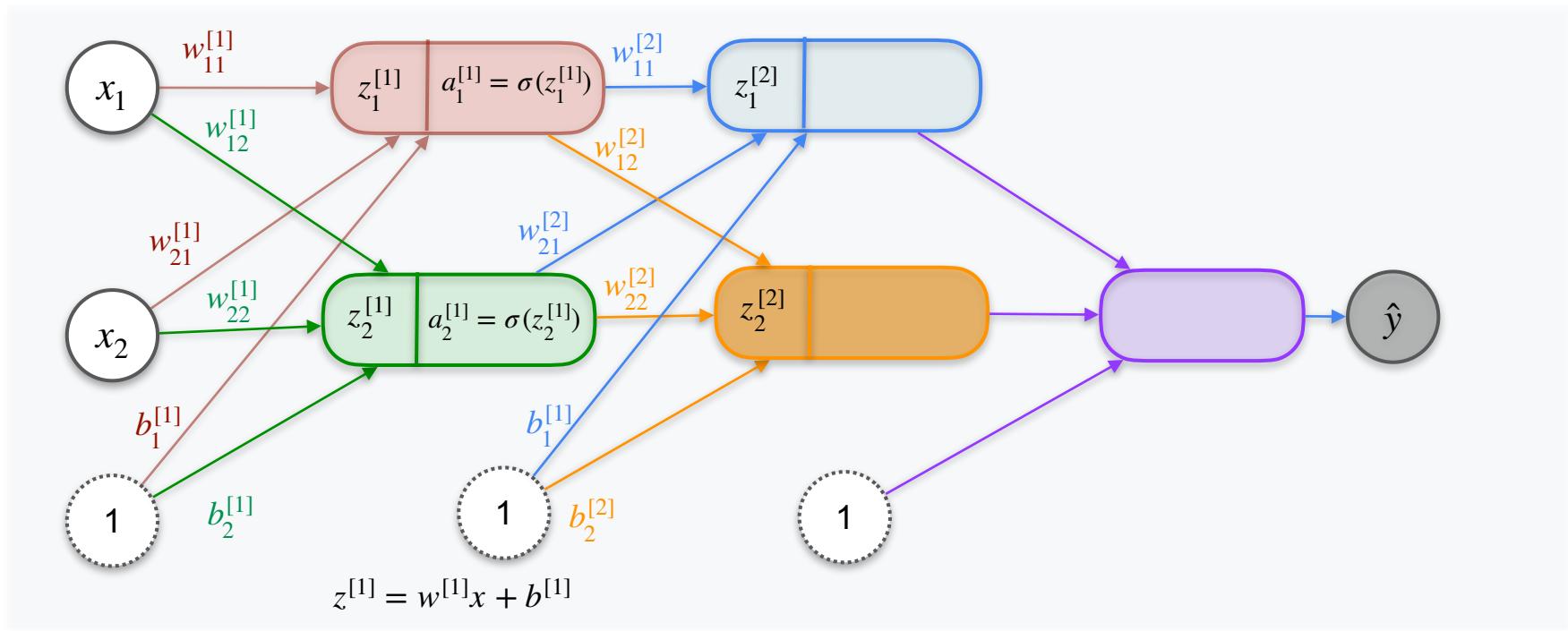
# Back Propagation Introduction



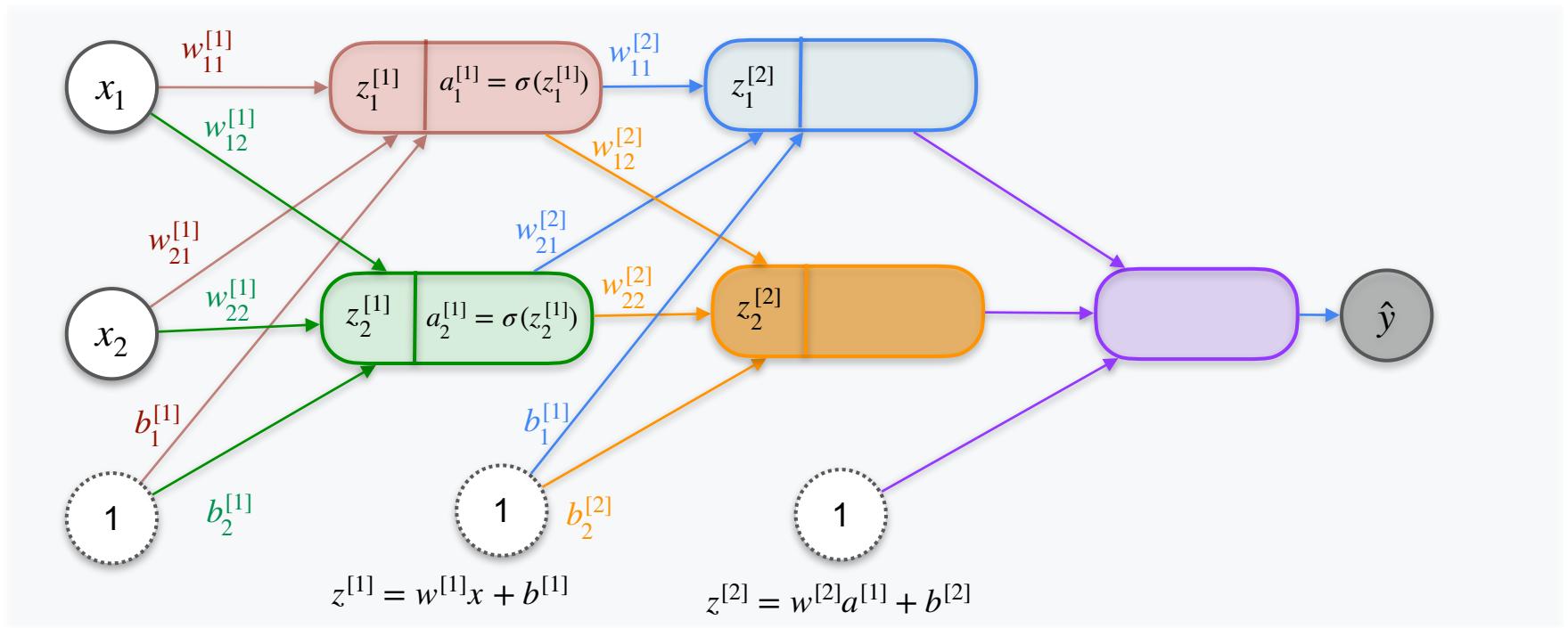
# Back Propagation Introduction



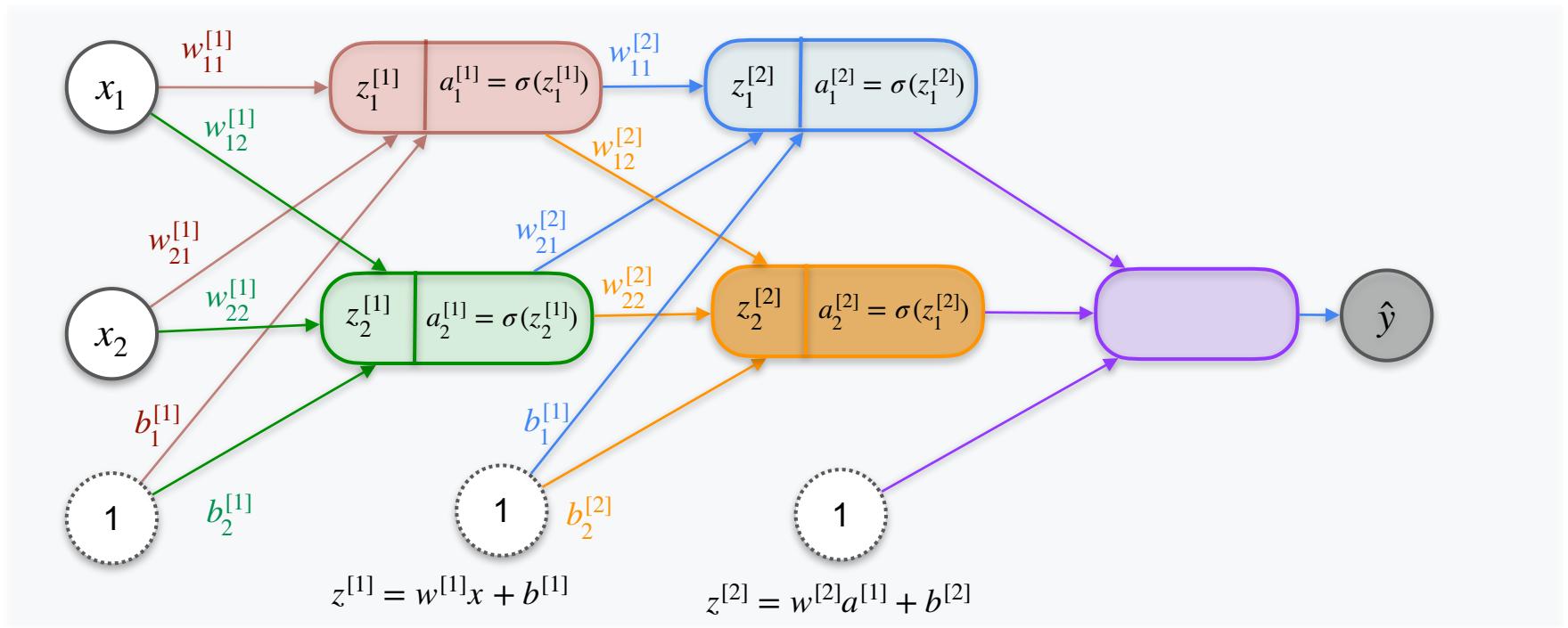
# Back Propagation Introduction



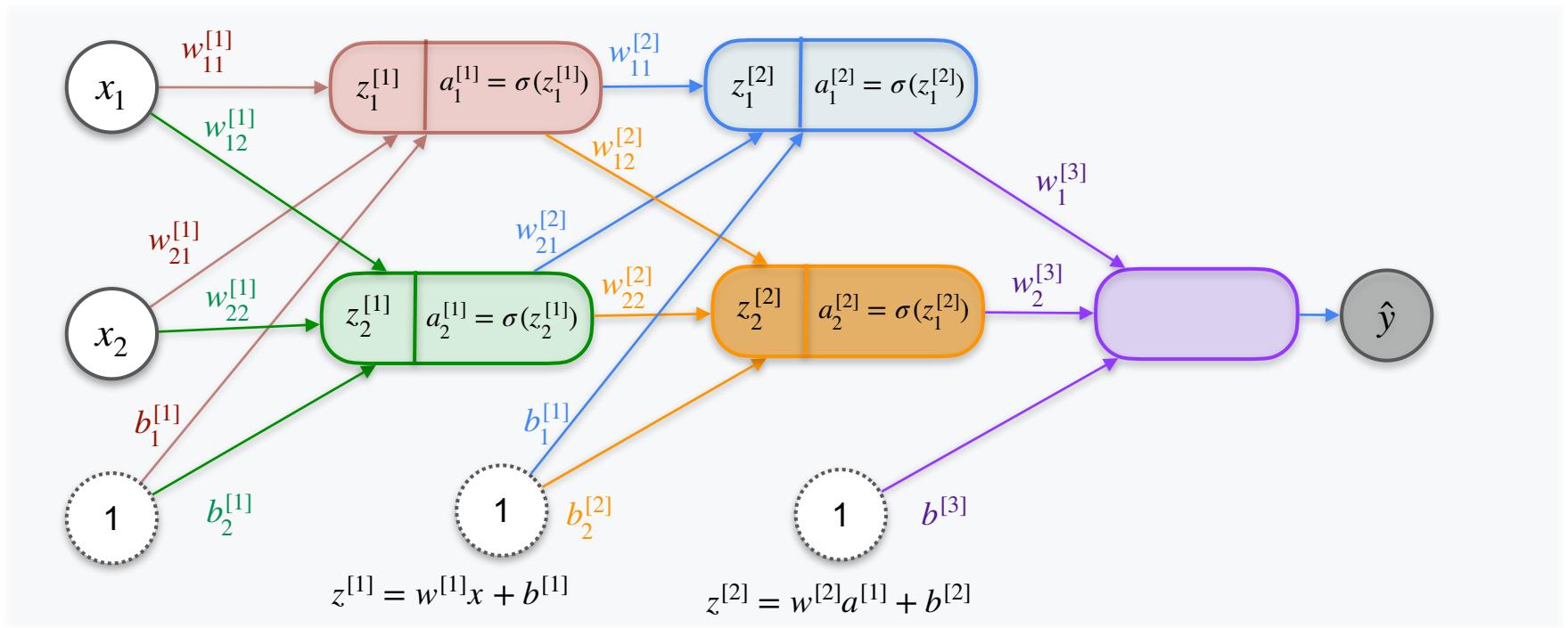
# Back Propagation Introduction



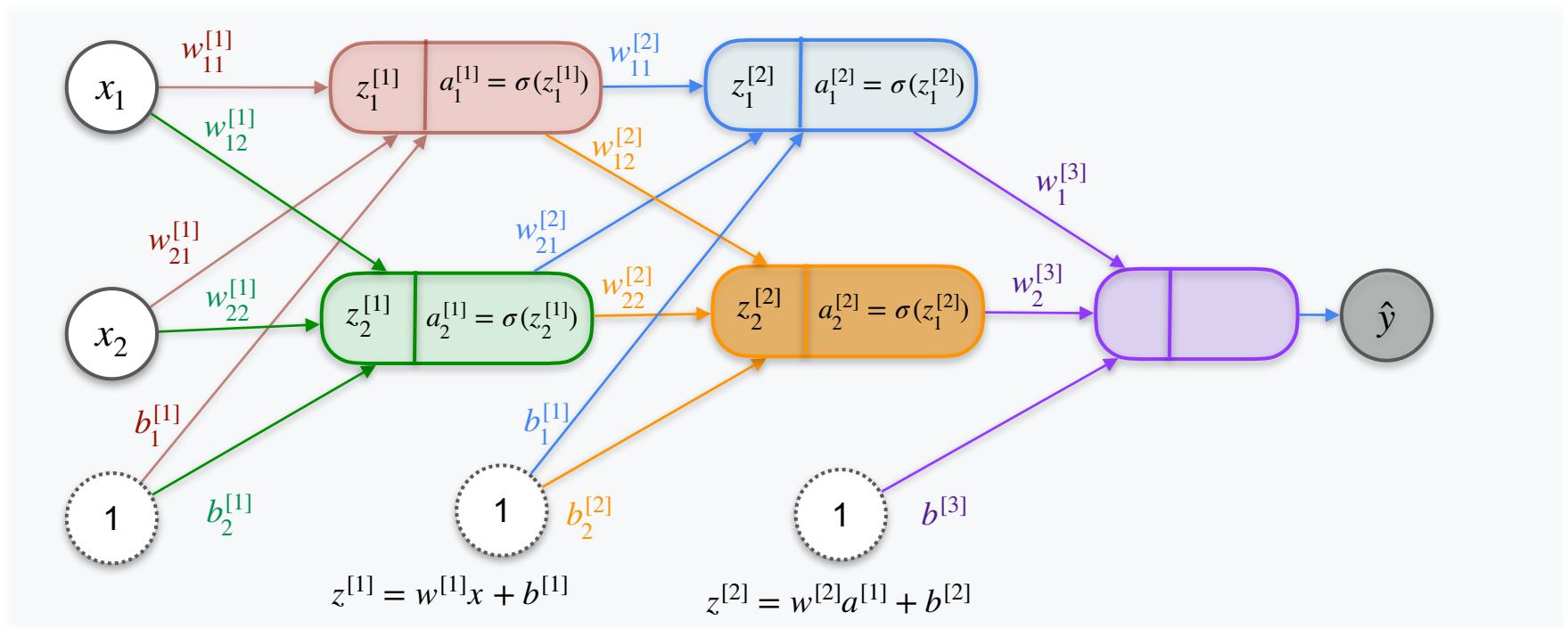
# Back Propagation Introduction



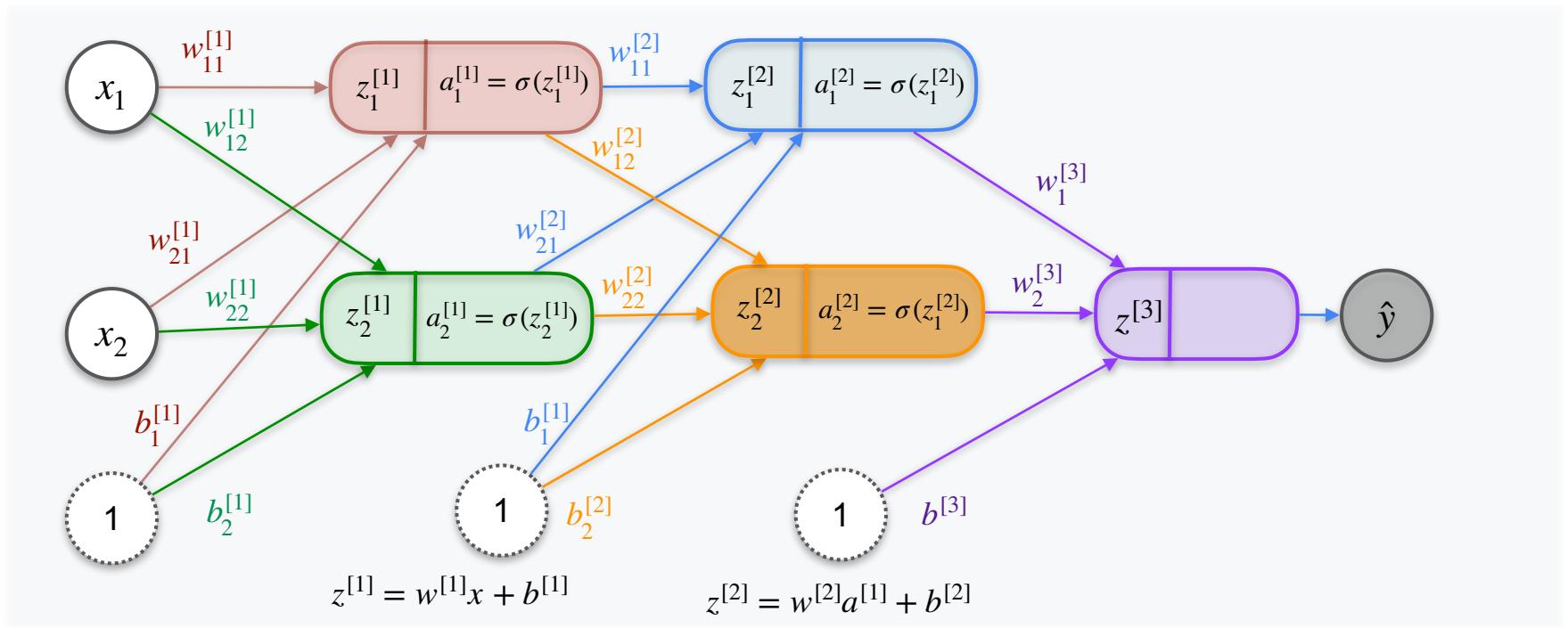
# Back Propagation Introduction



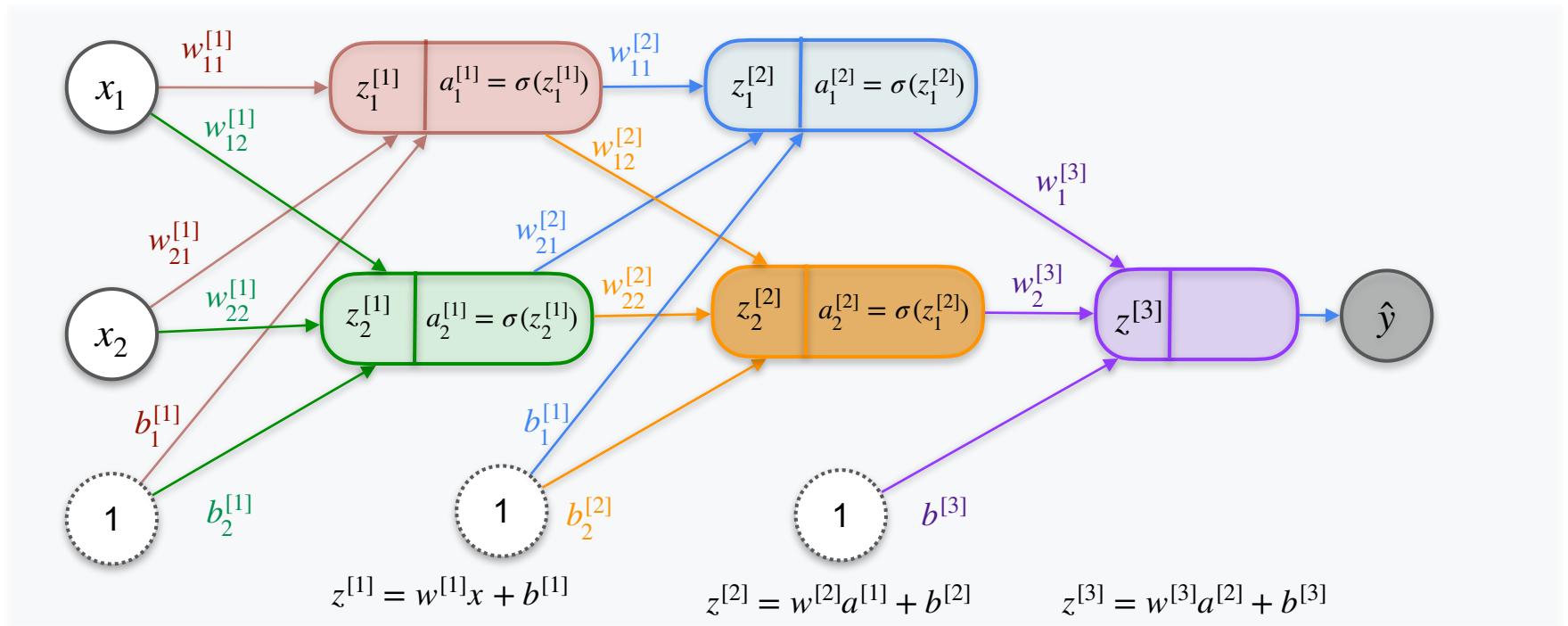
# Back Propagation Introduction



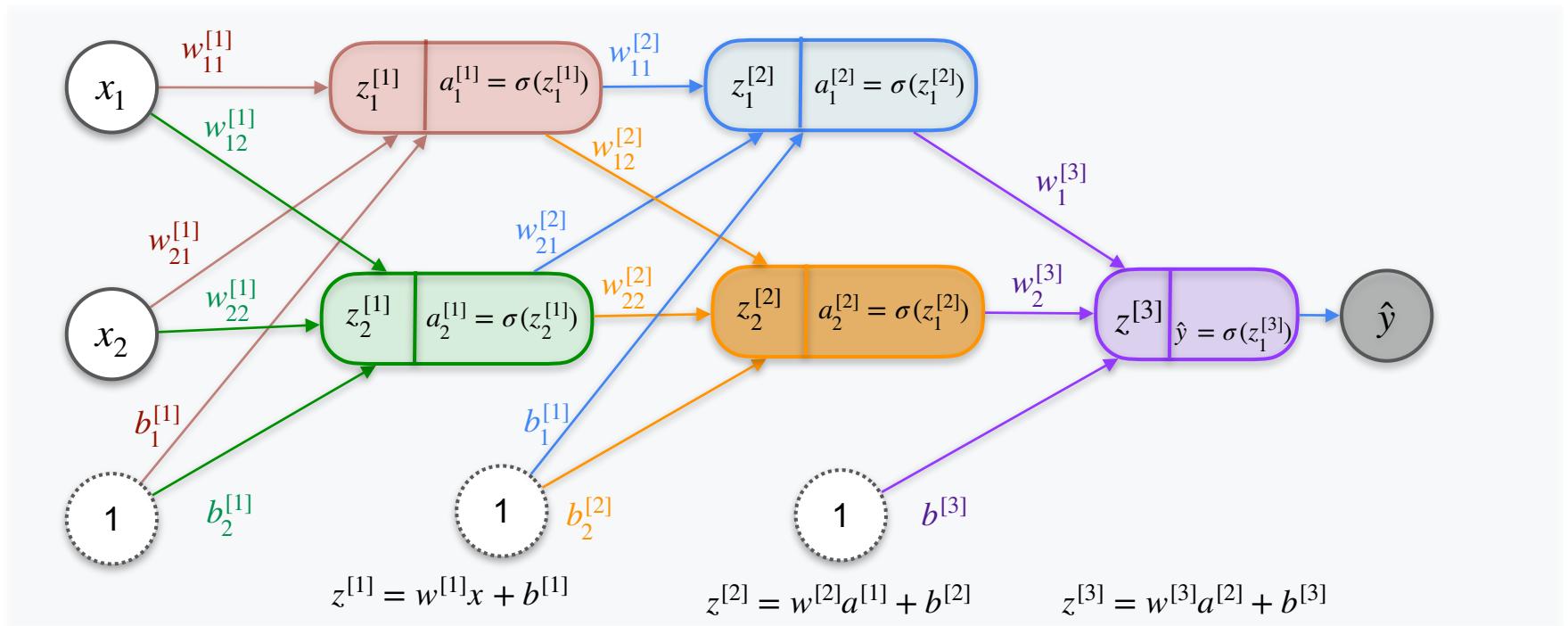
# Back Propagation Introduction



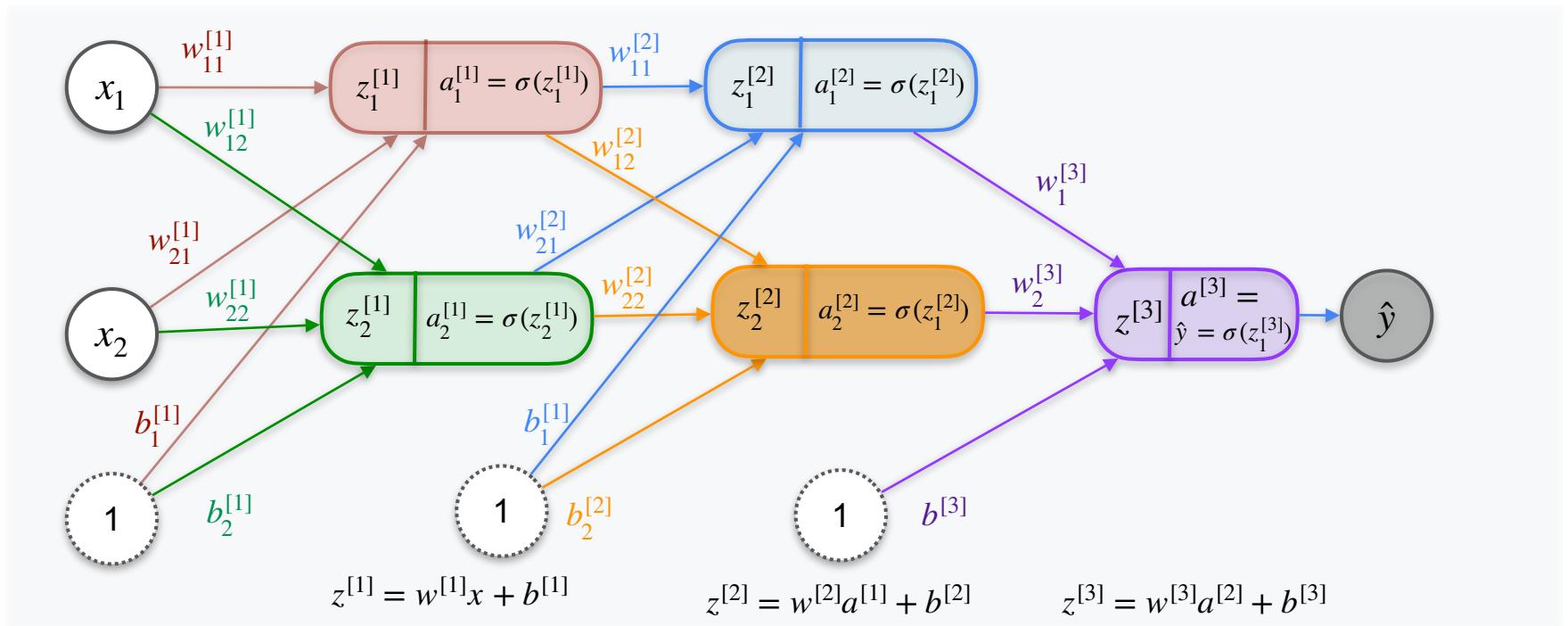
# Back Propagation Introduction



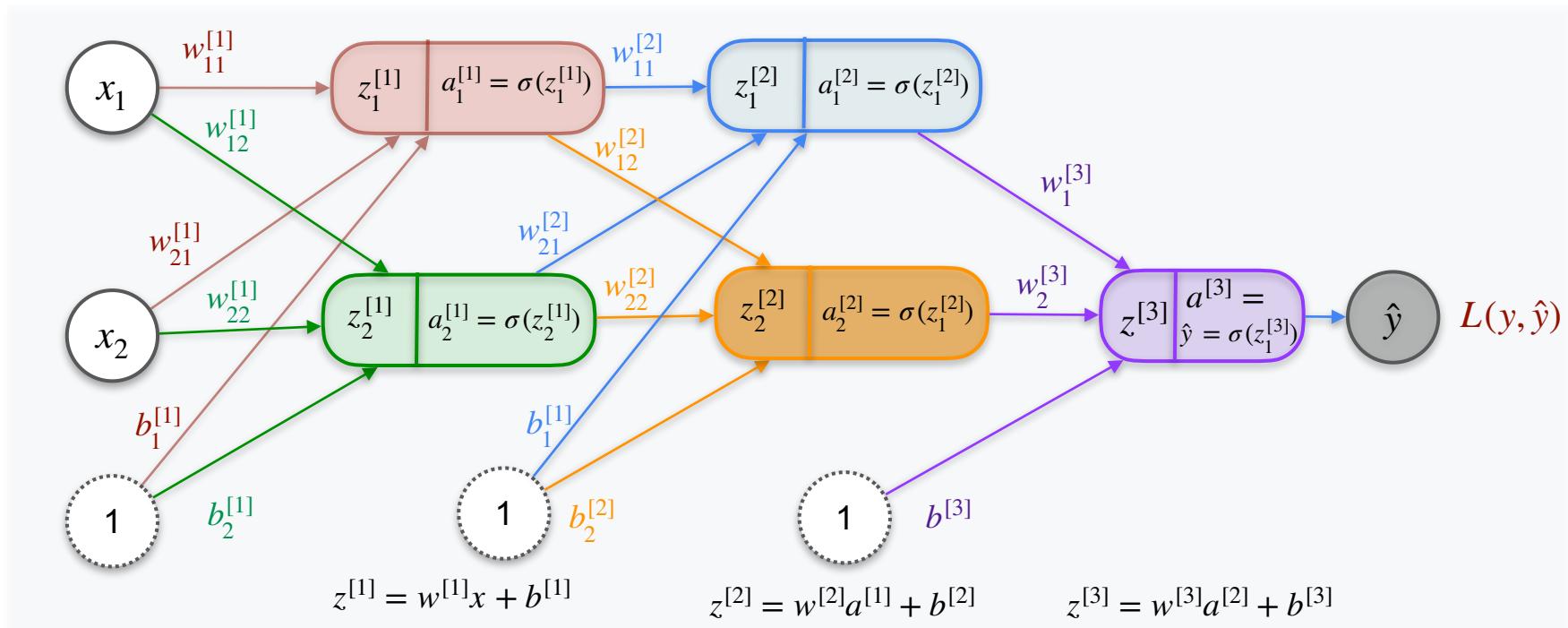
# Back Propagation Introduction



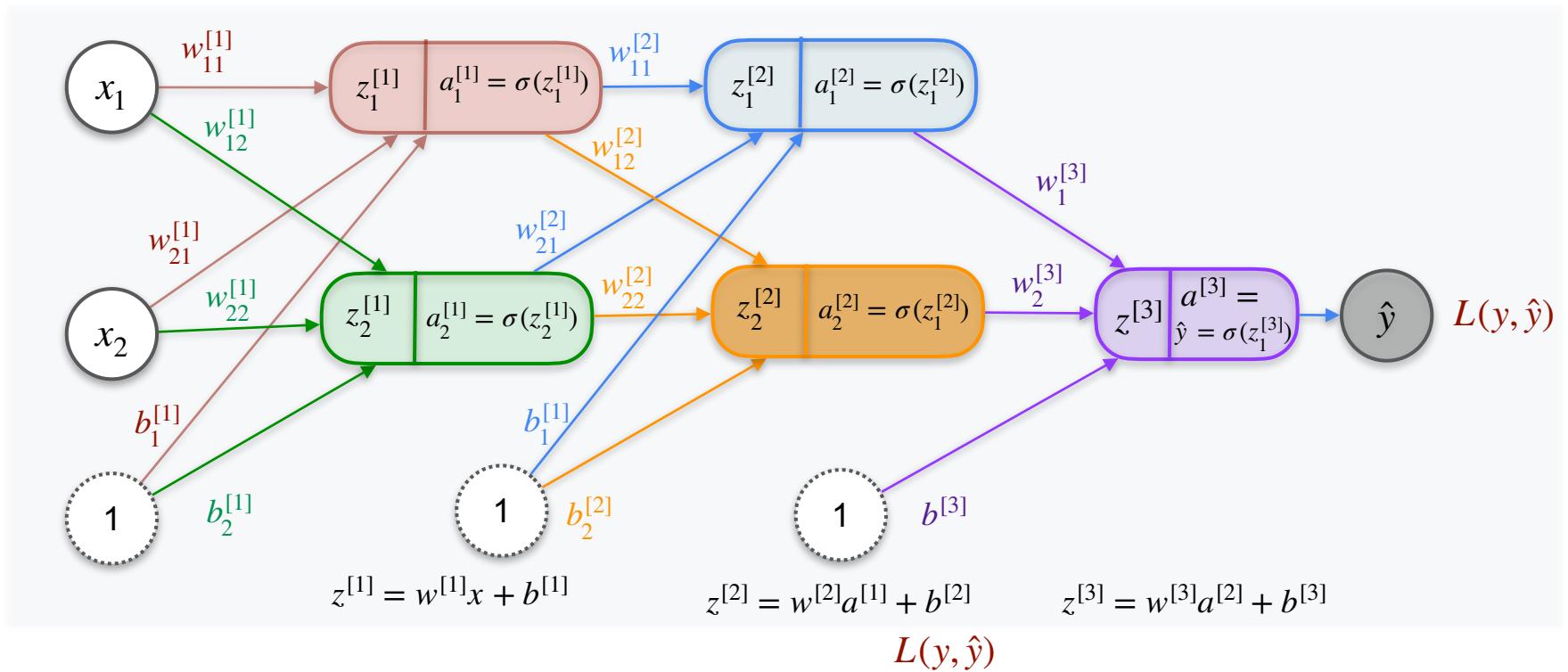
# Back Propagation Introduction



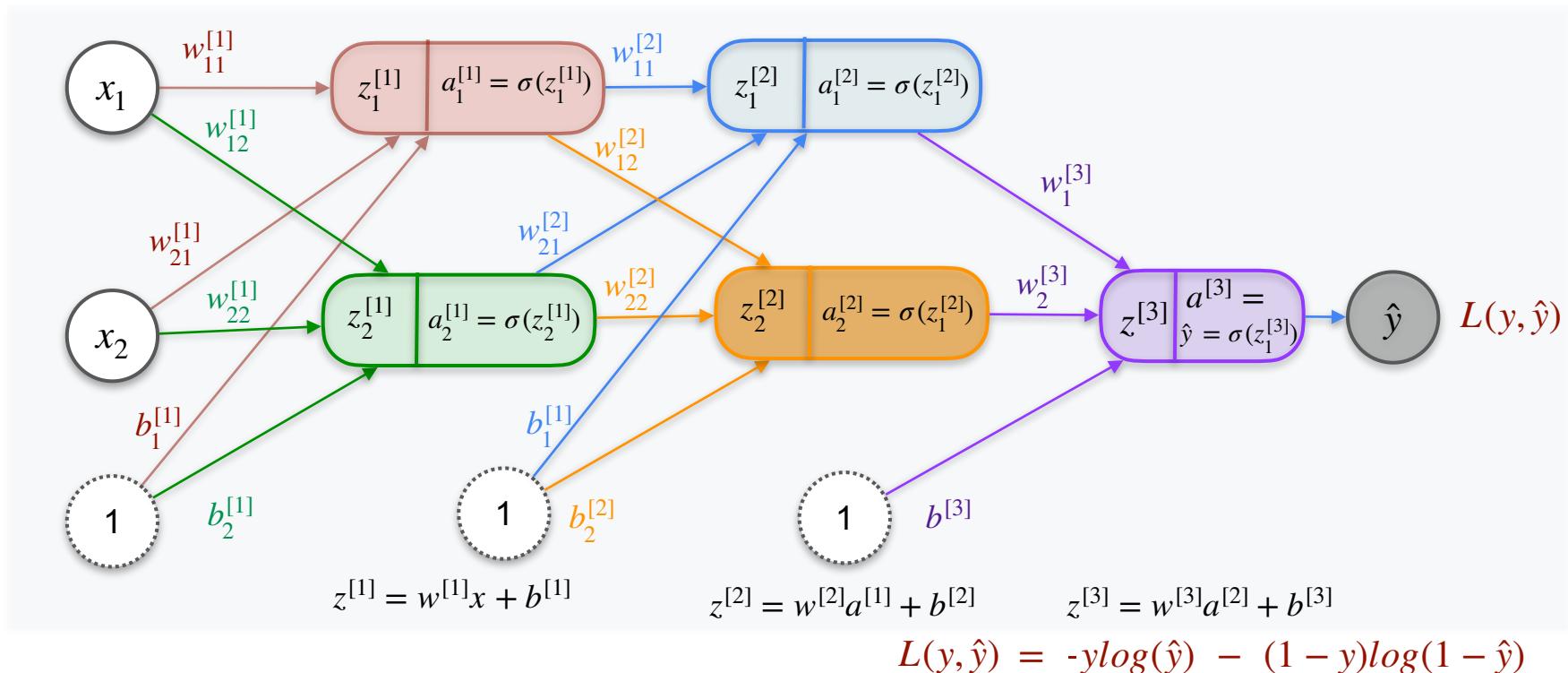
# Back Propagation Introduction



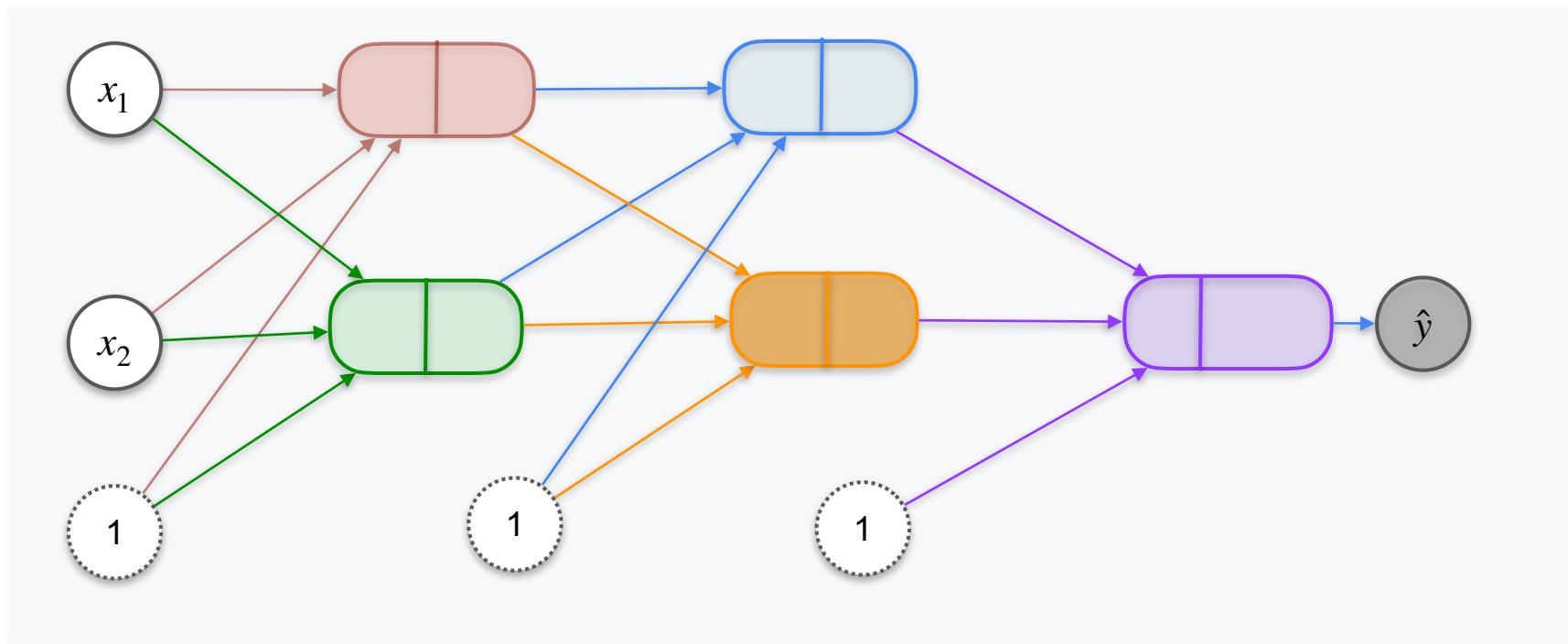
# Back Propagation Introduction



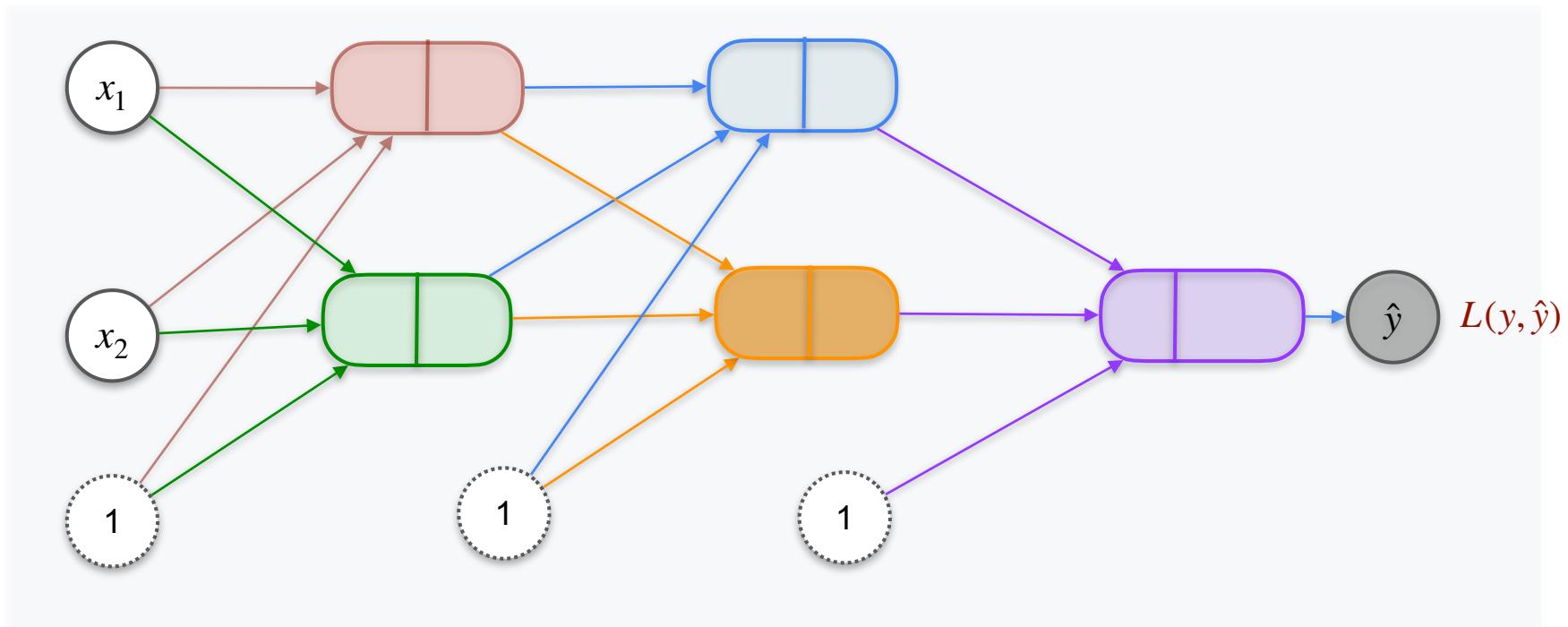
# Back Propagation Introduction



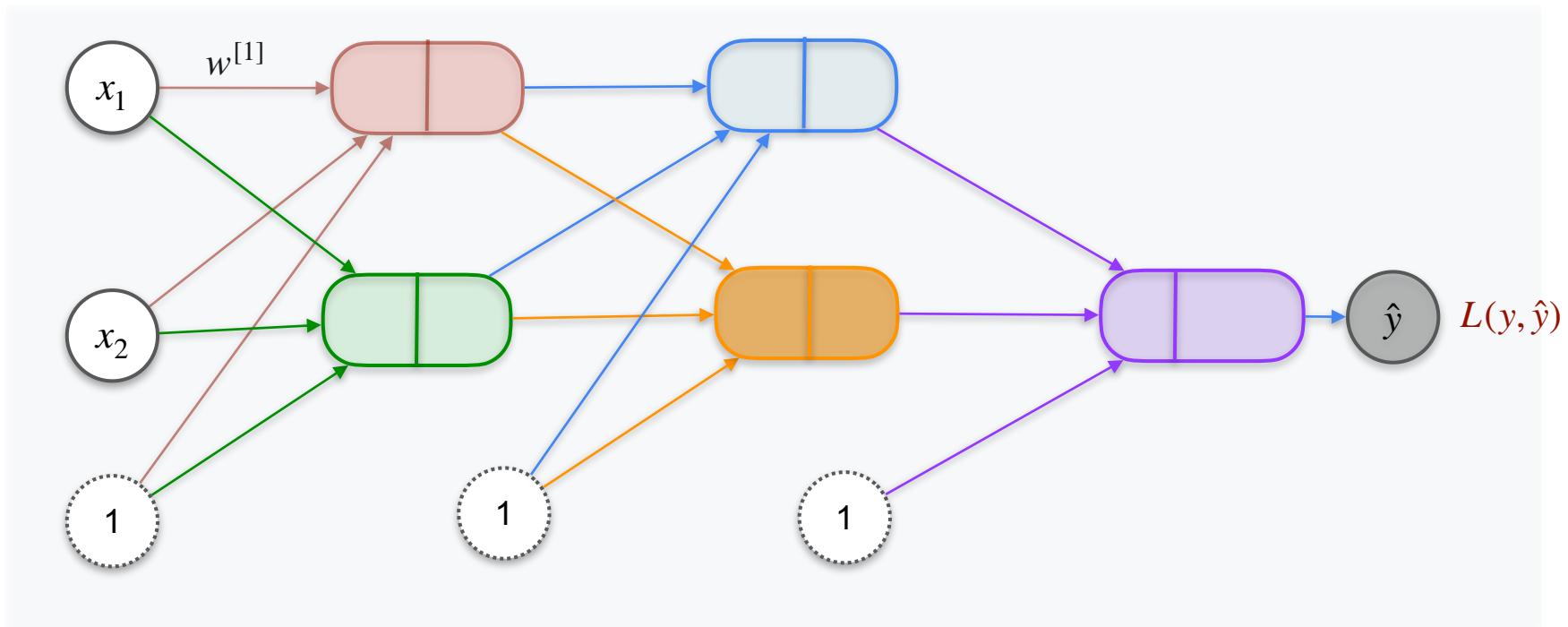
# Back Propagation Introduction



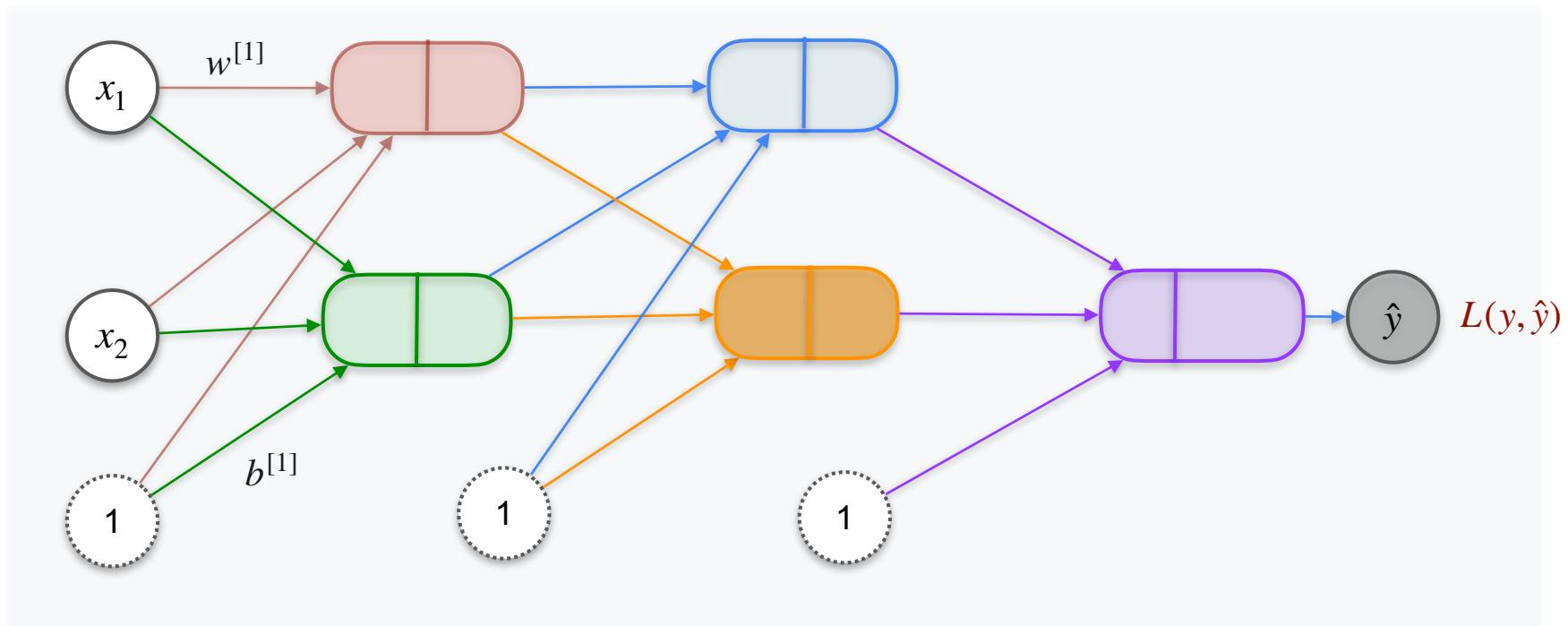
# Back Propagation Introduction



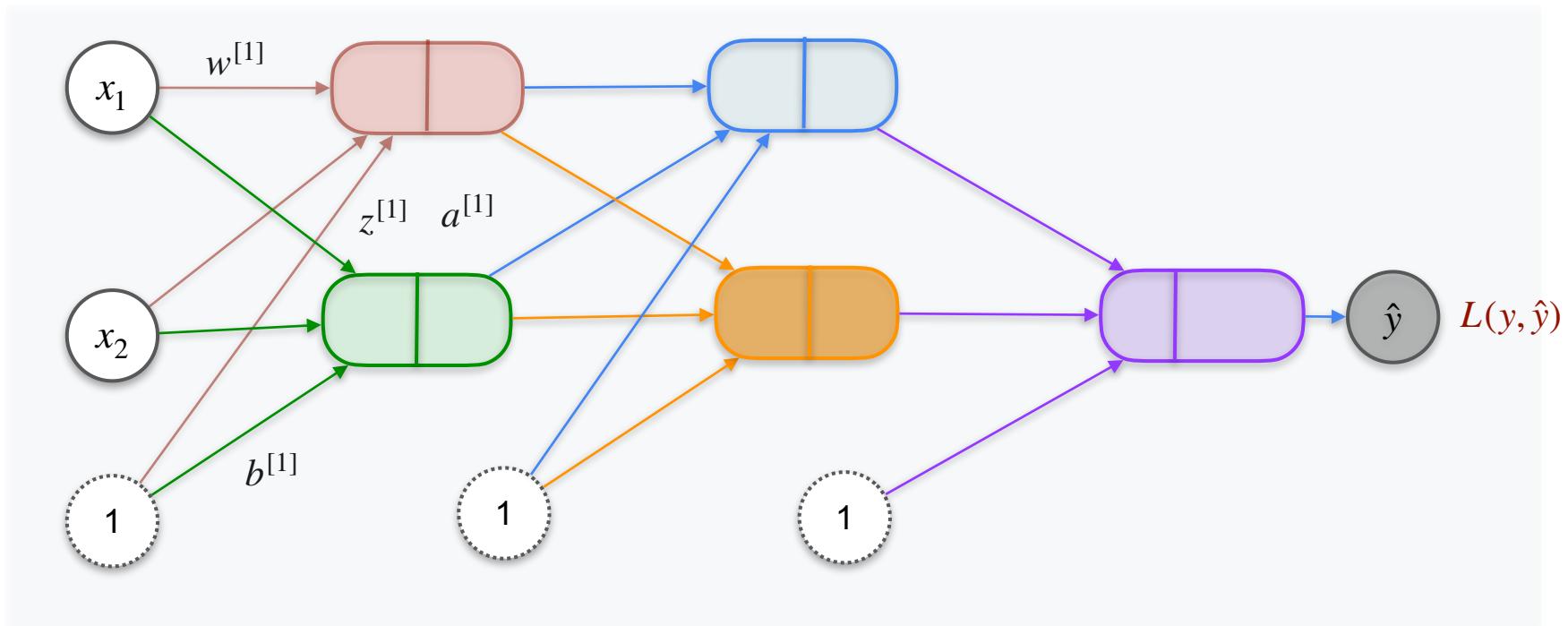
# Back Propagation Introduction



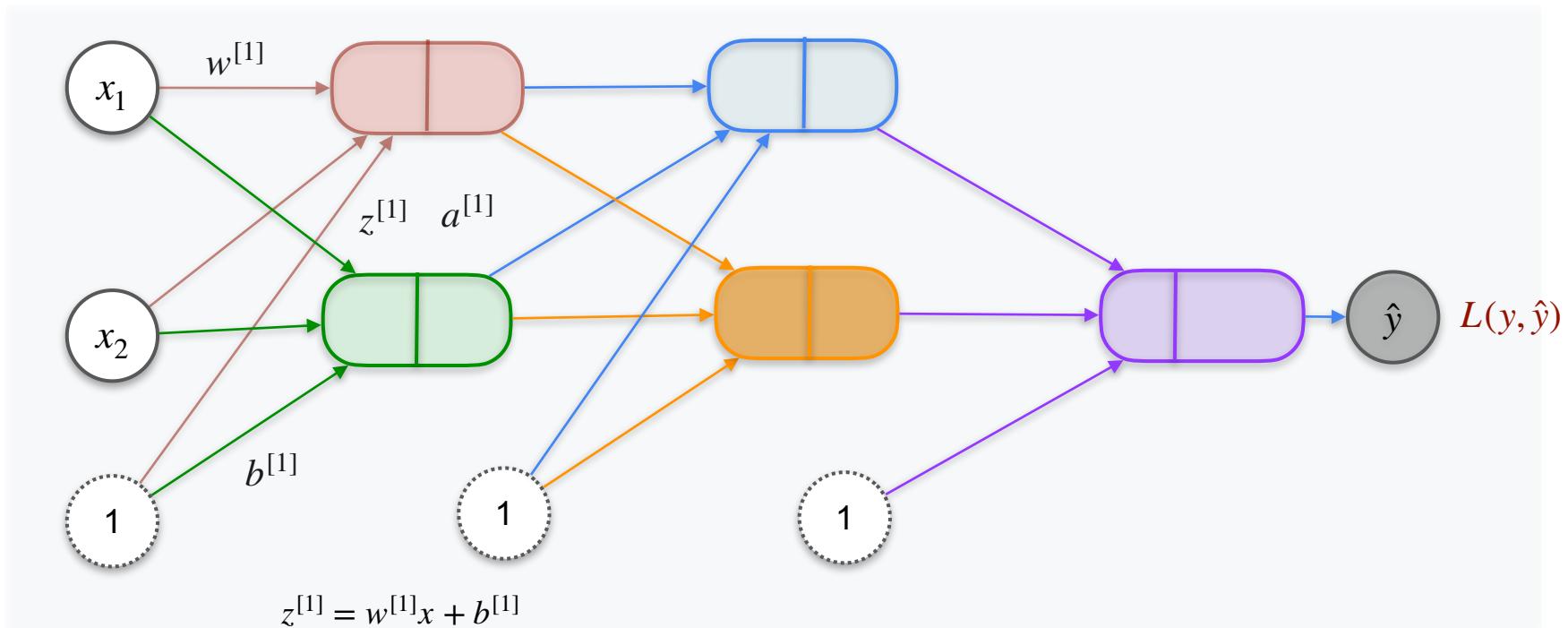
# Back Propagation Introduction



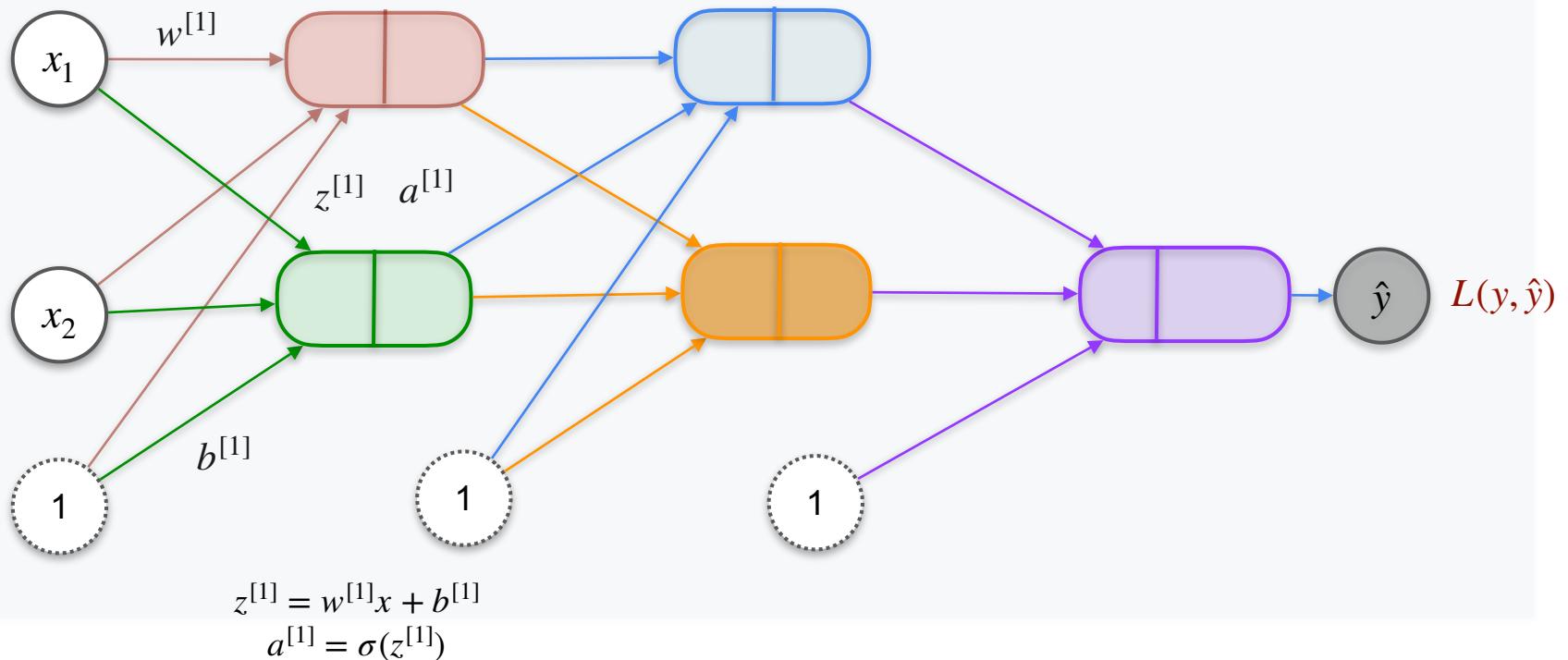
# Back Propagation Introduction



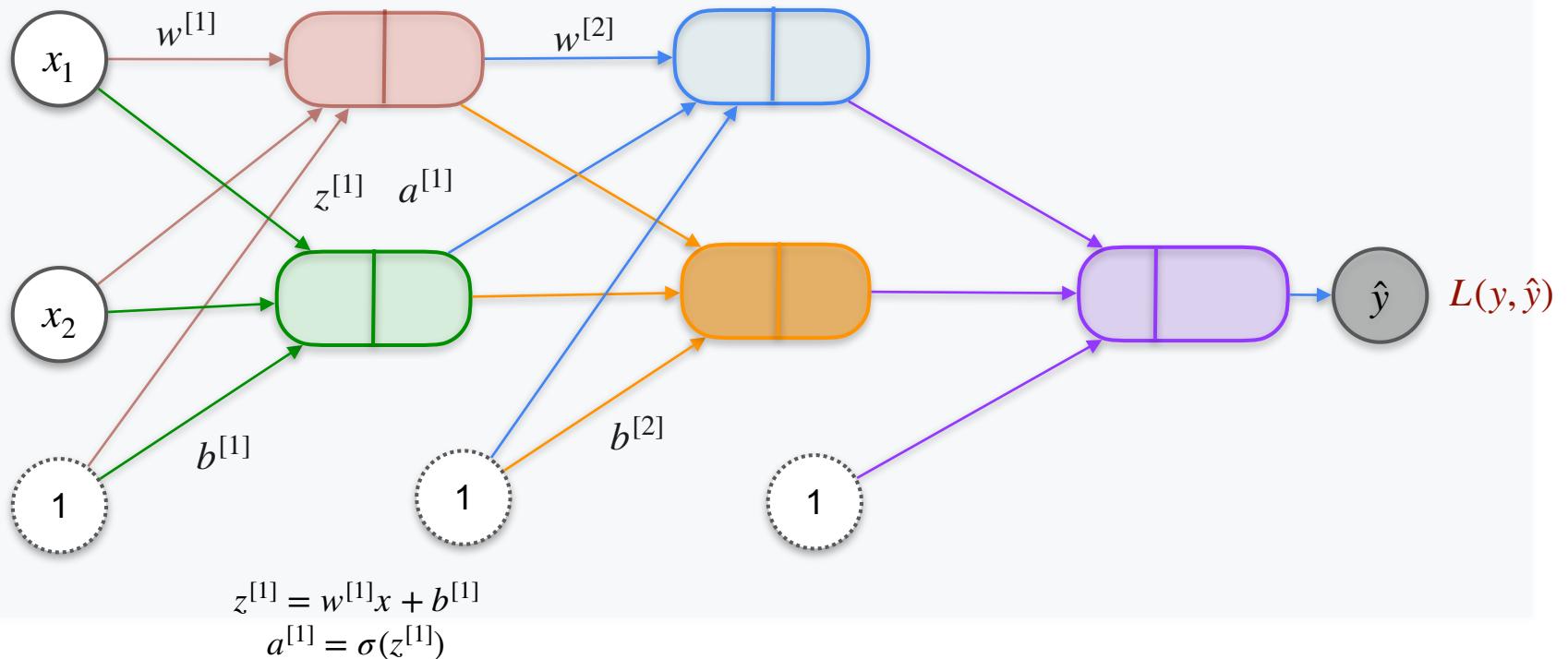
# Back Propagation Introduction



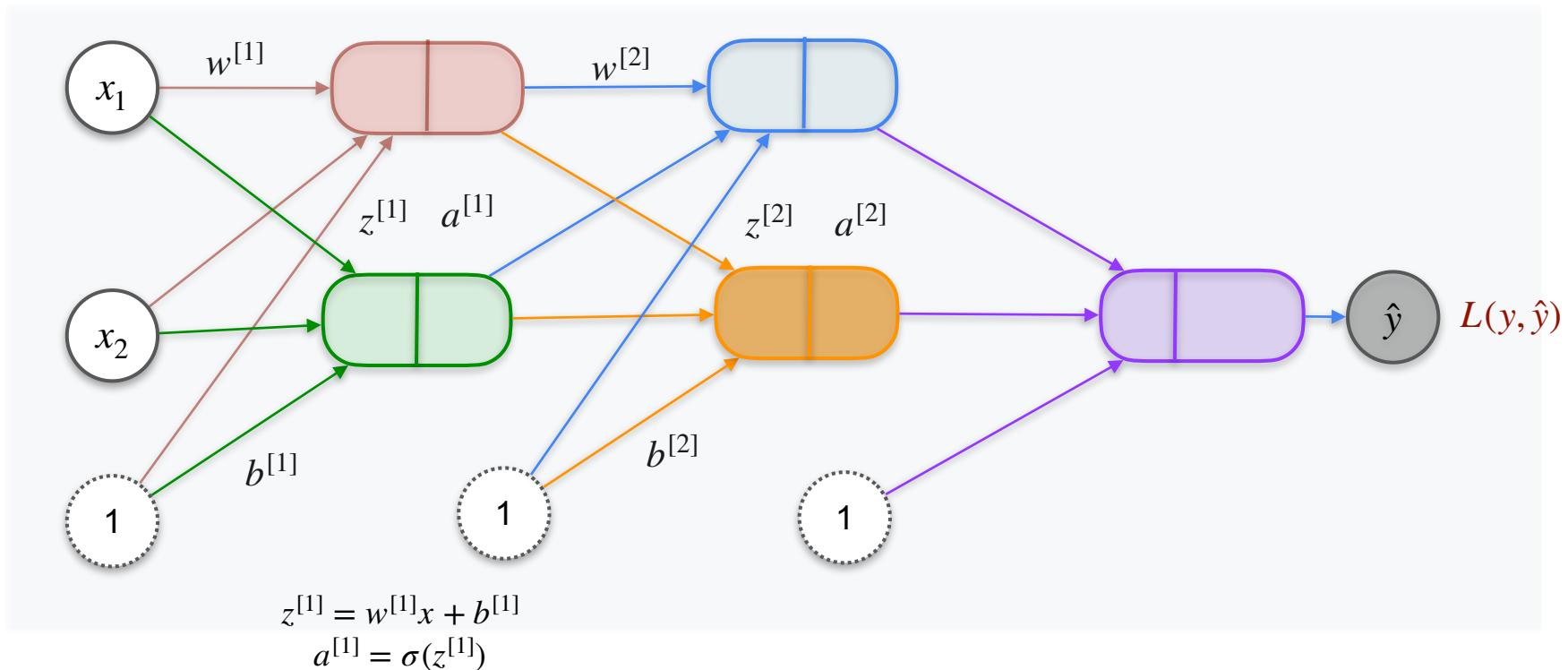
# Back Propagation Introduction



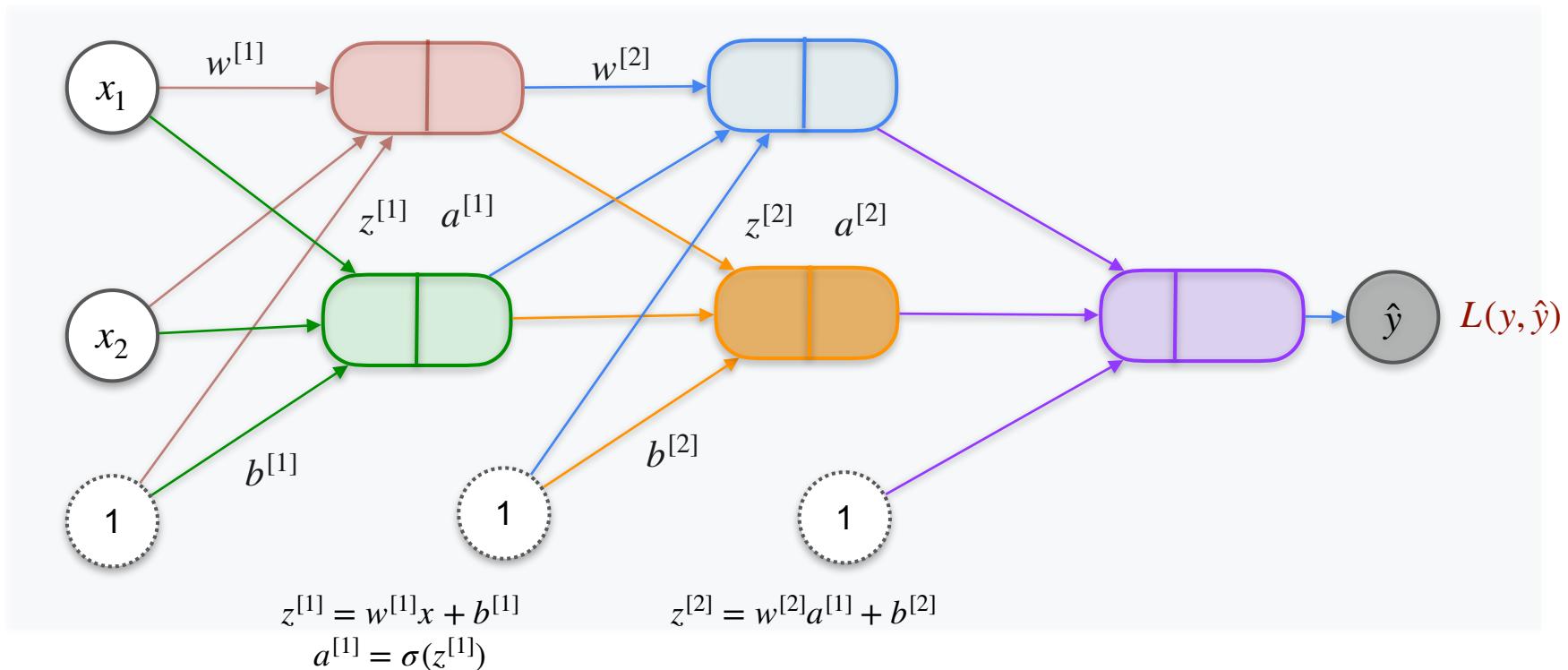
# Back Propagation Introduction



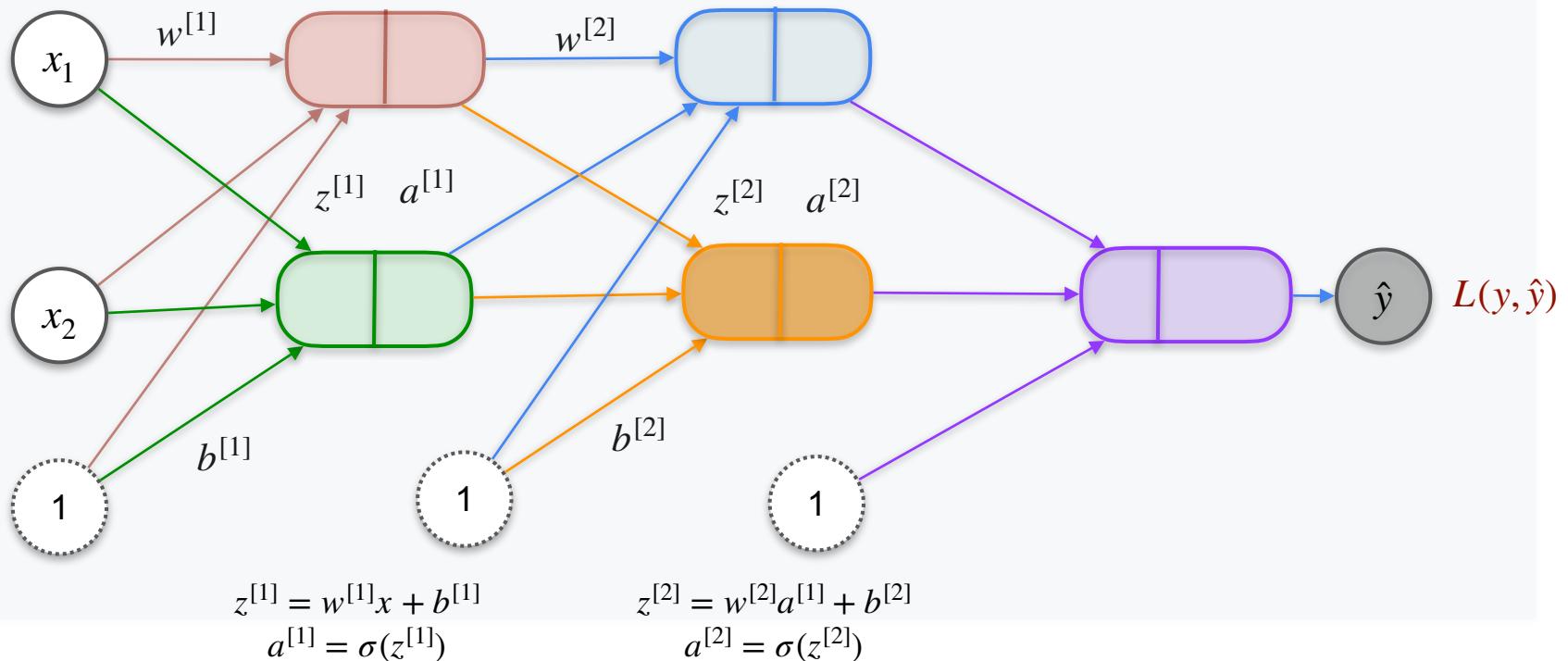
# Back Propagation Introduction



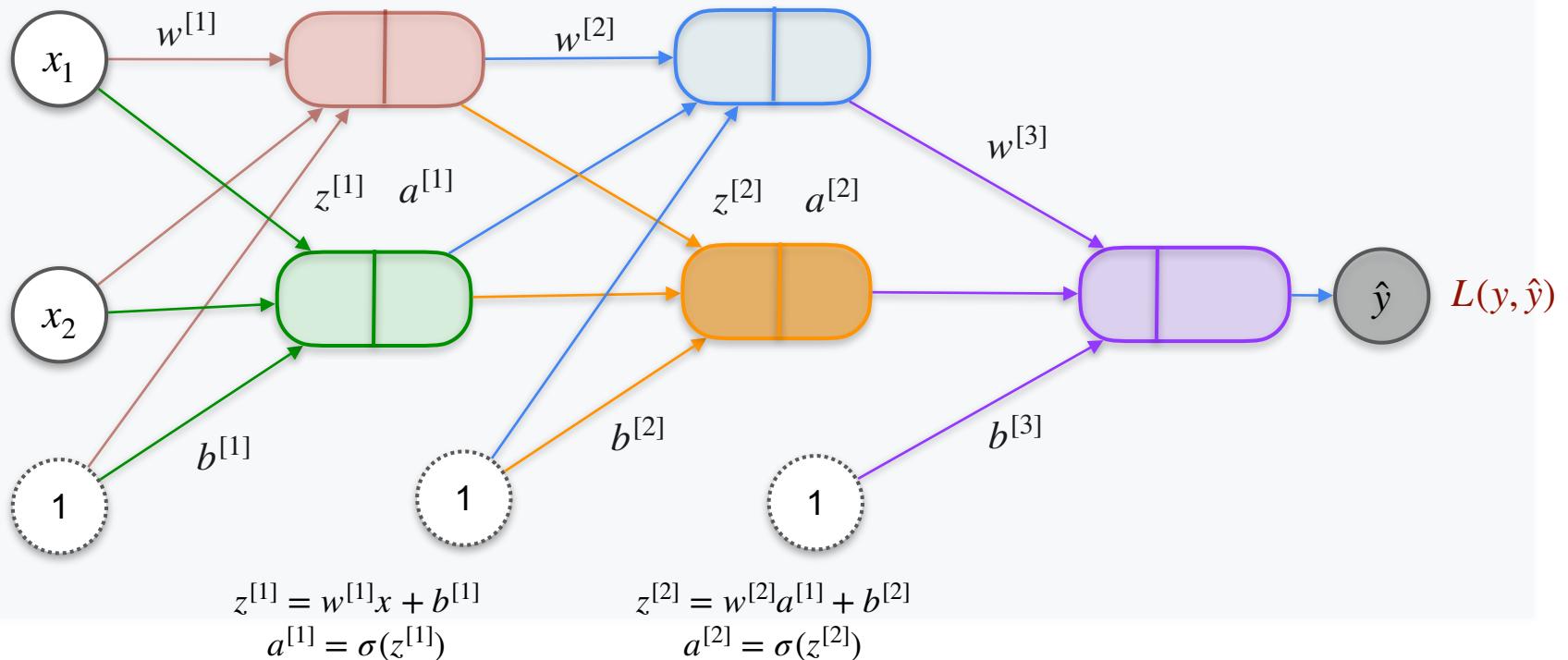
# Back Propagation Introduction



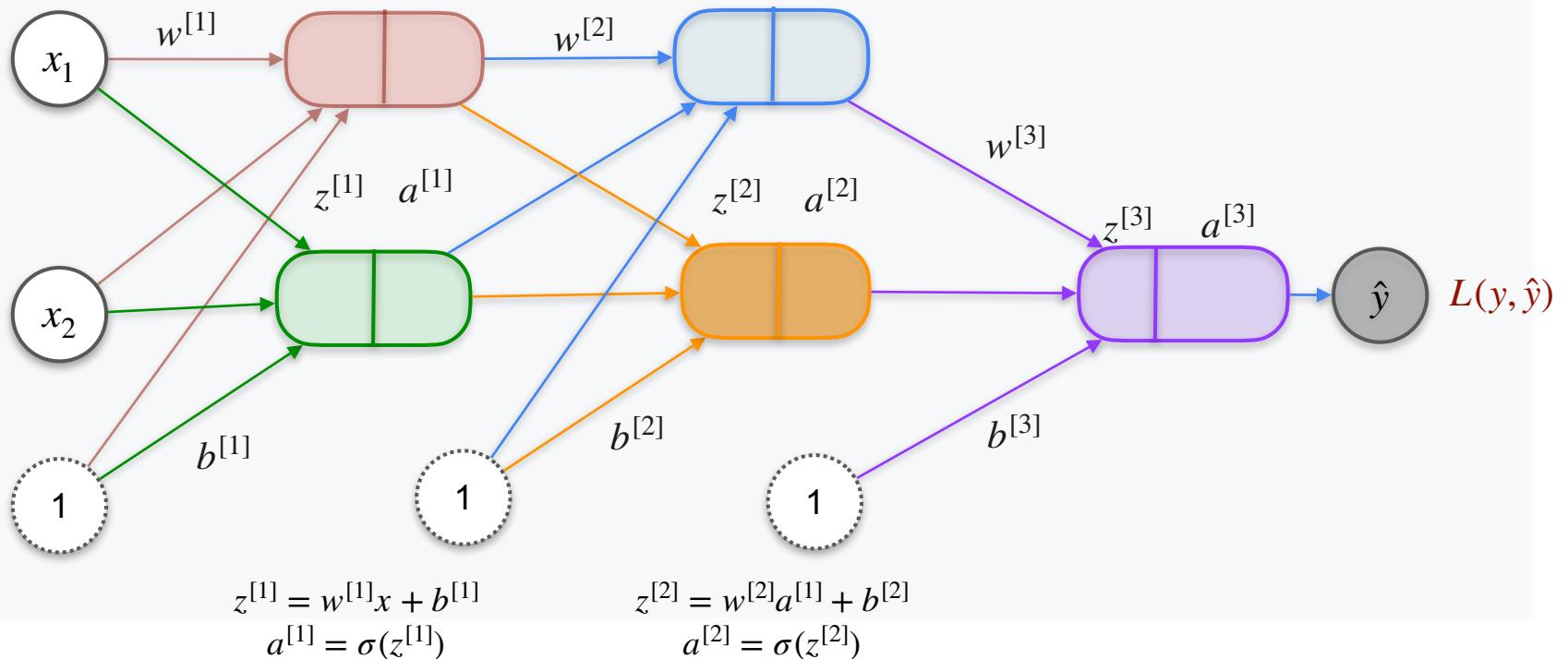
# Back Propagation Introduction



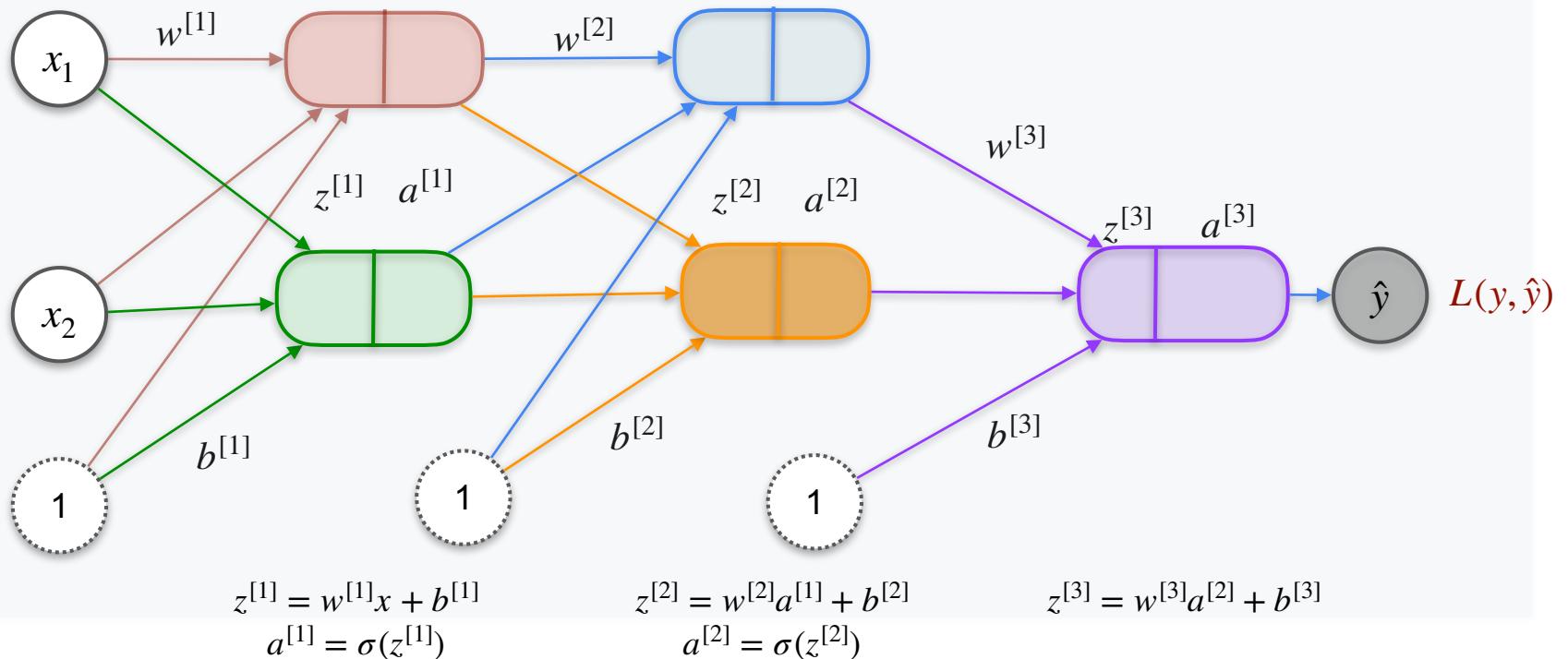
# Back Propagation Introduction



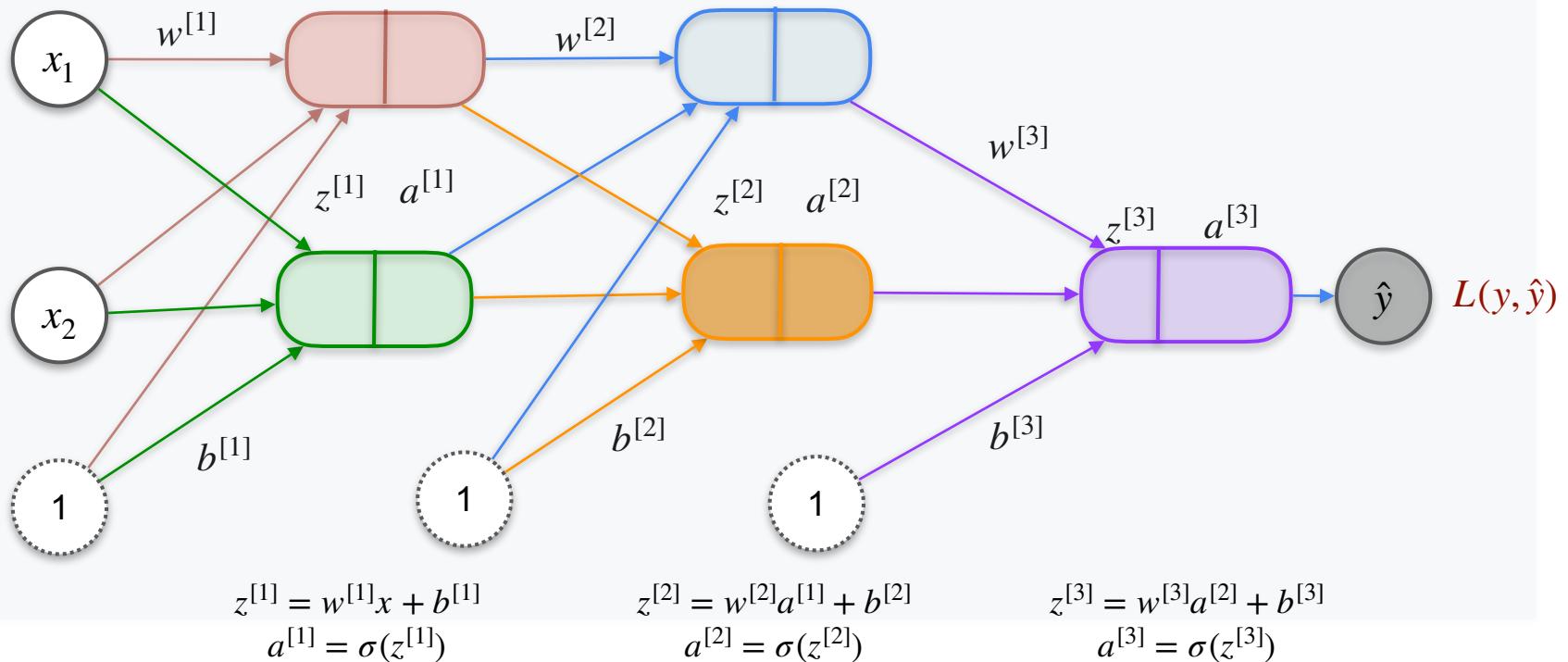
# Back Propagation Introduction



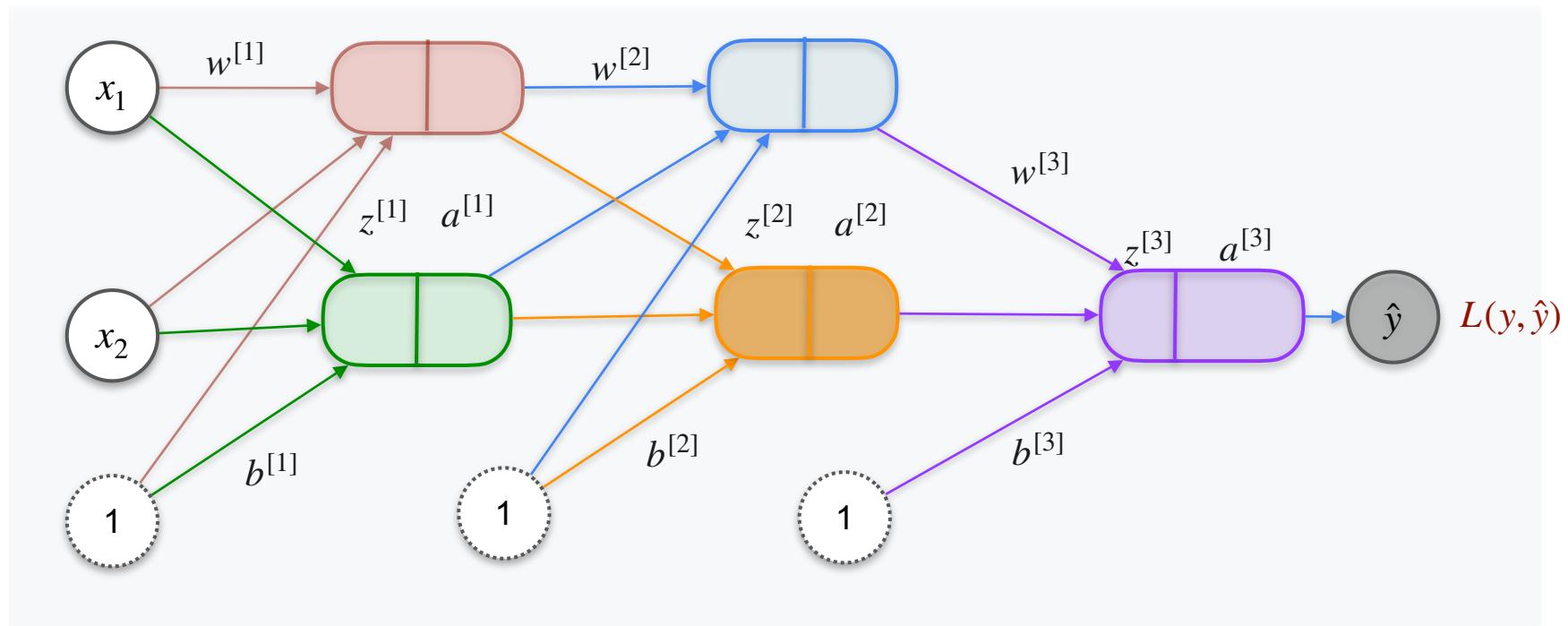
# Back Propagation Introduction



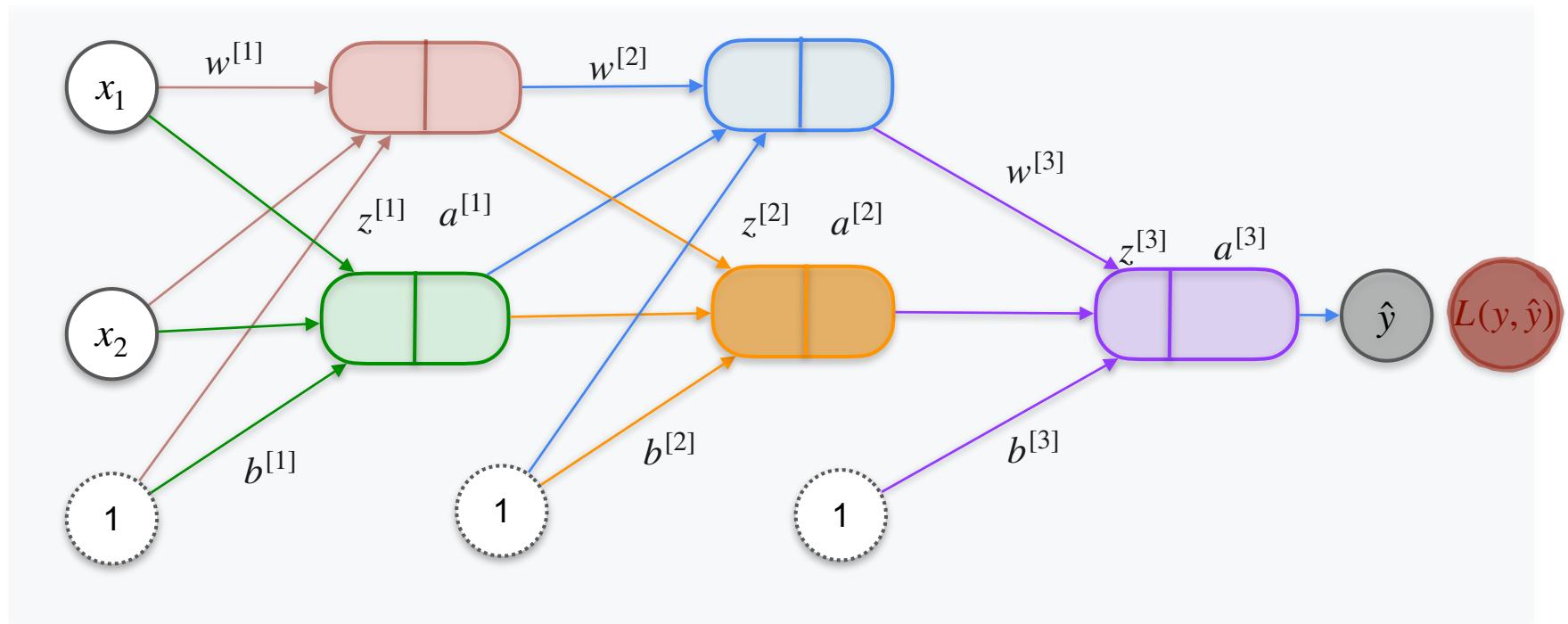
# Back Propagation Introduction



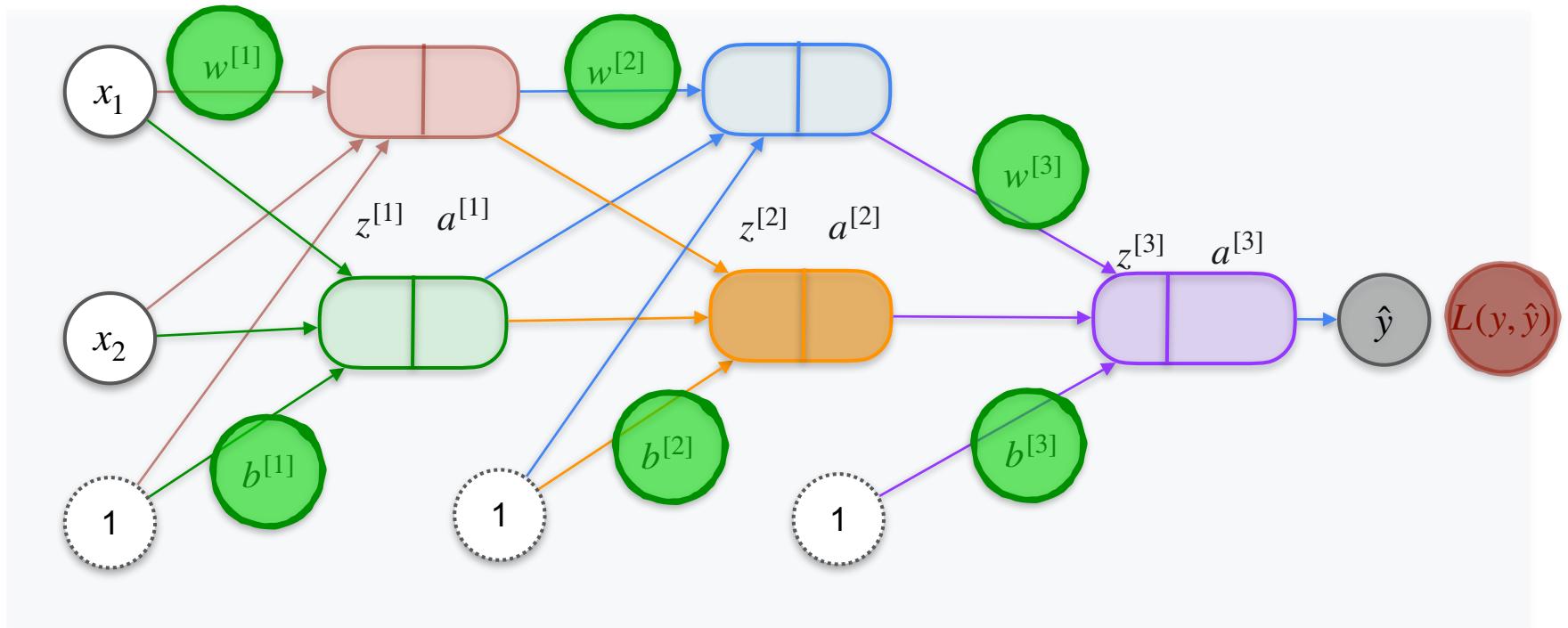
# Back Propagation Introduction



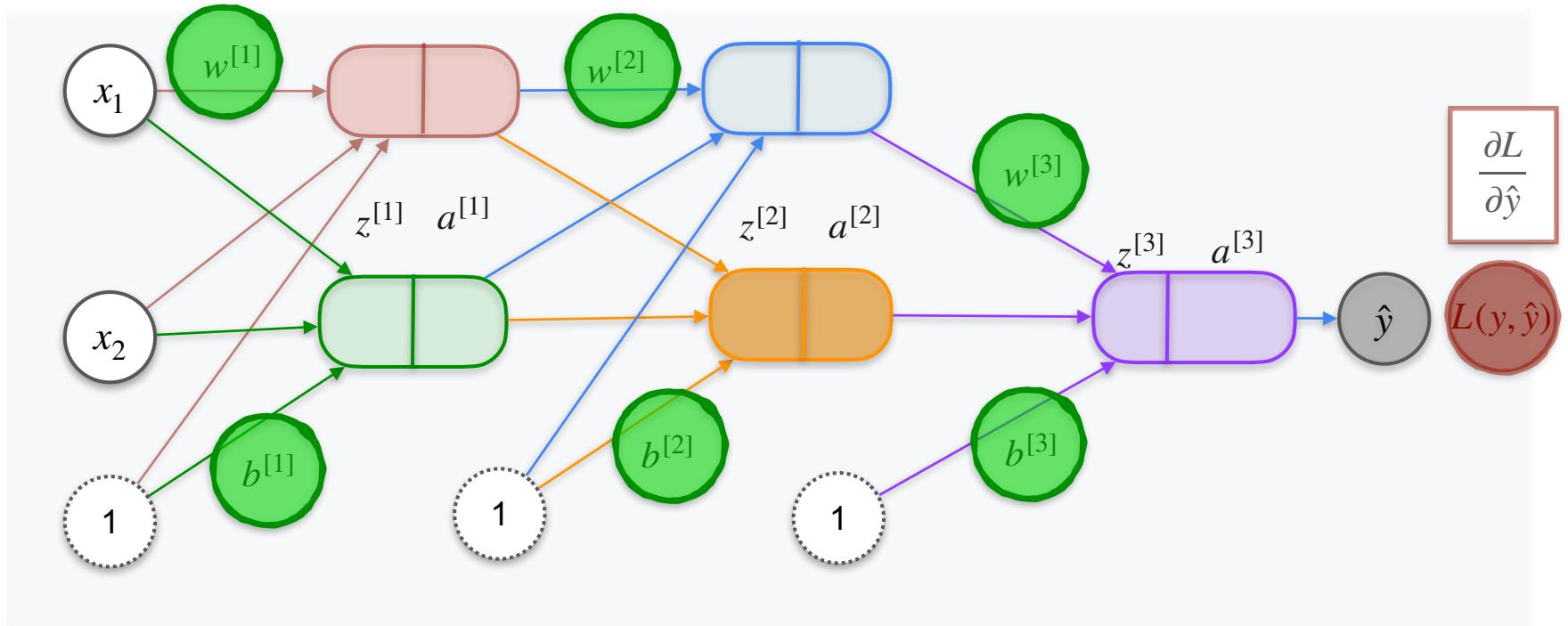
# Back Propagation Introduction



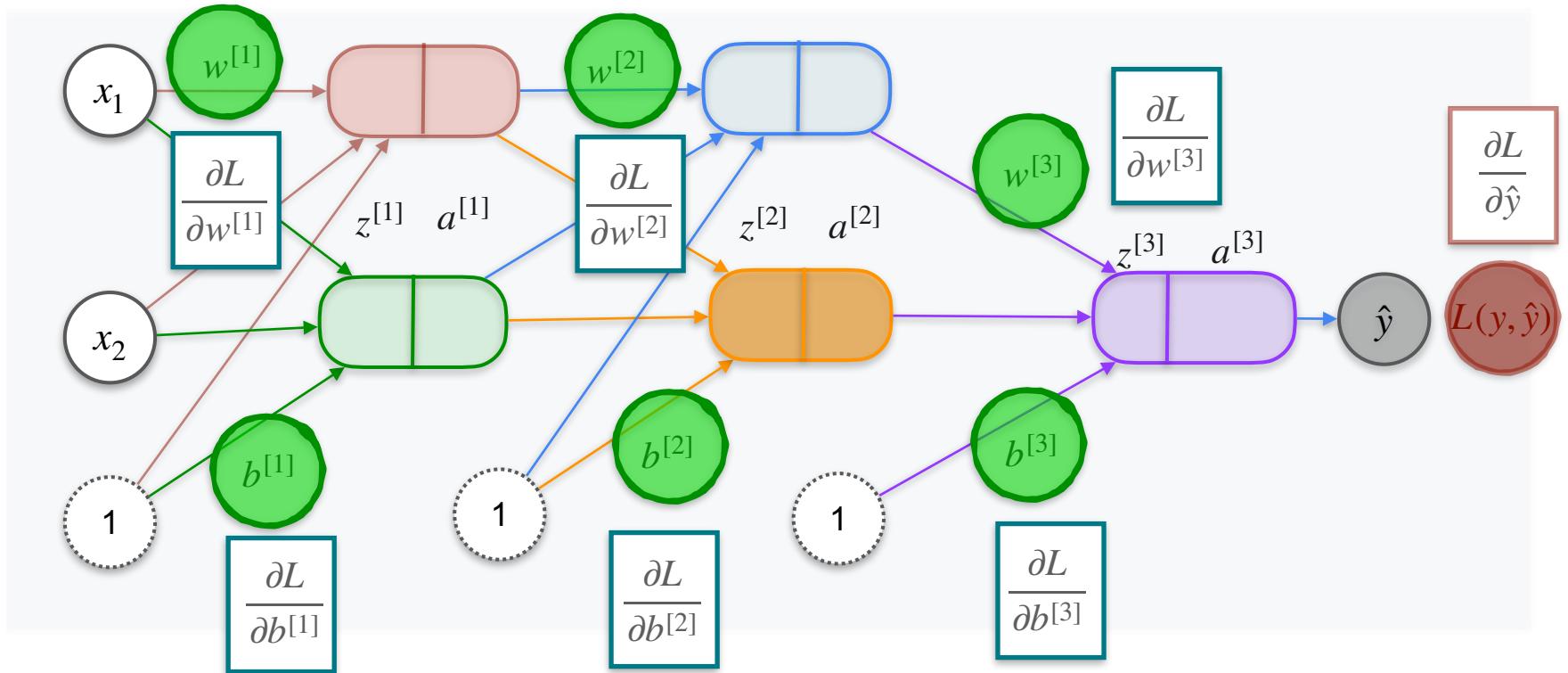
# Back Propagation Introduction



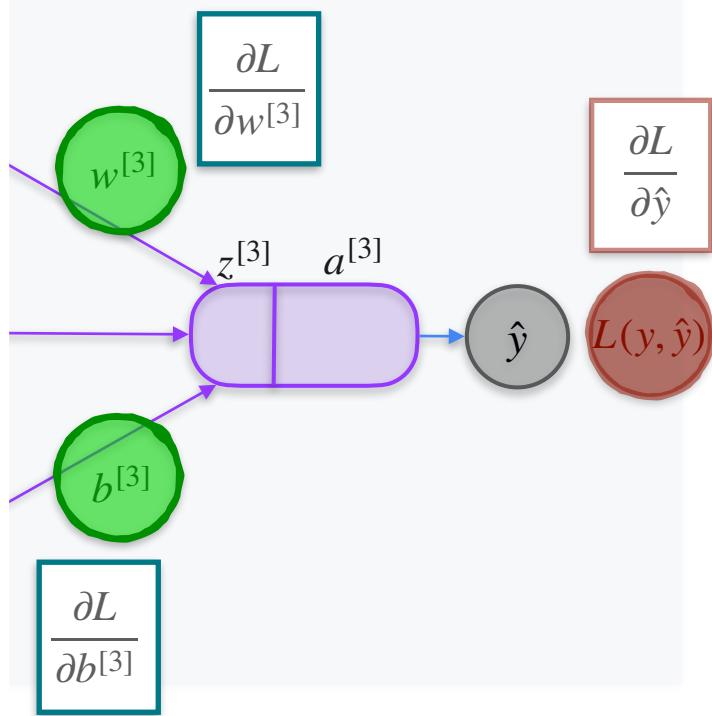
# Back Propagation Introduction



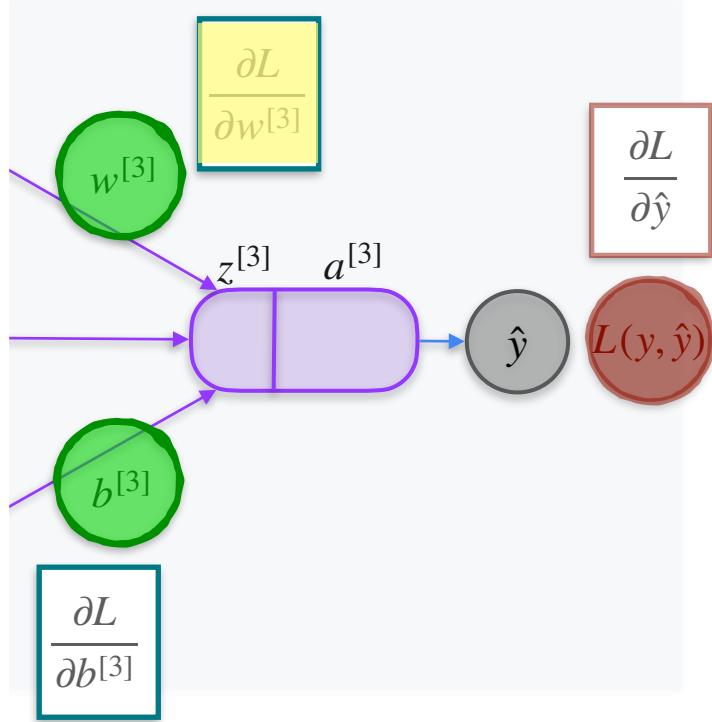
# Back Propagation Introduction



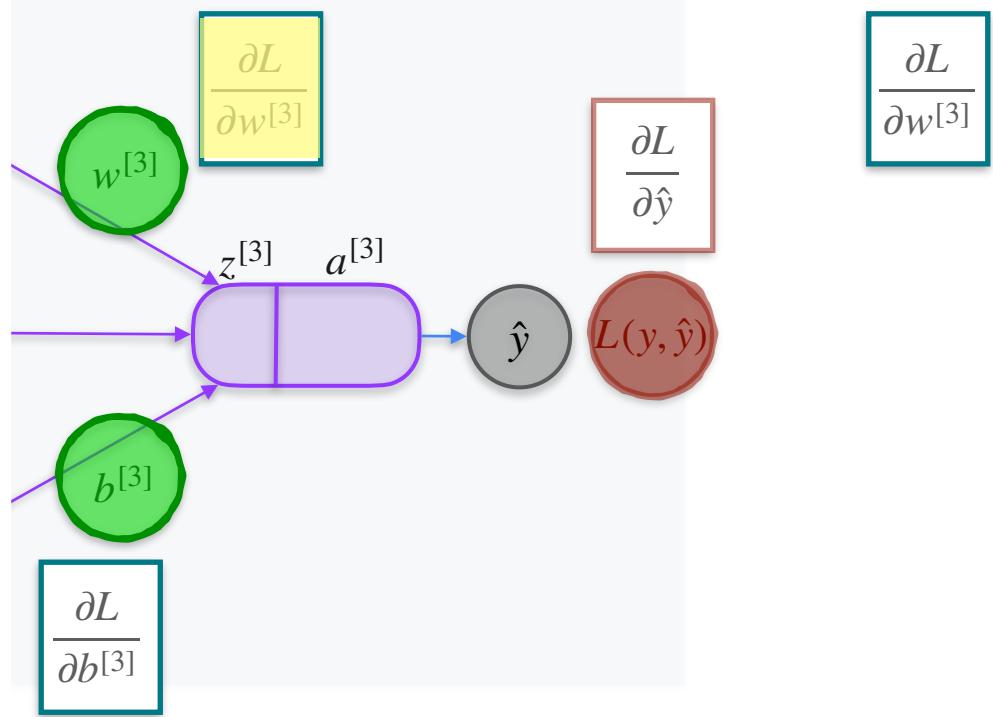
# Back Propagation Introduction



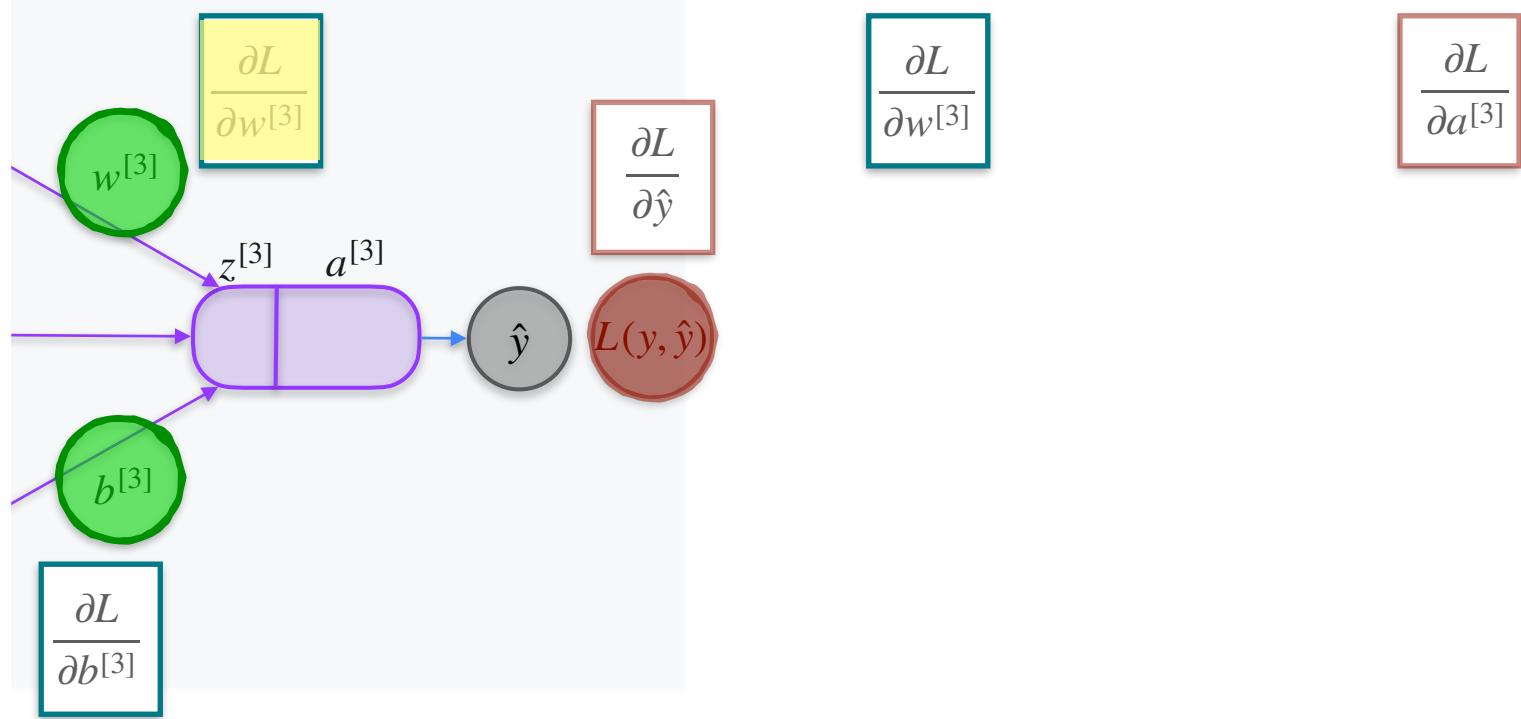
# Back Propagation Introduction



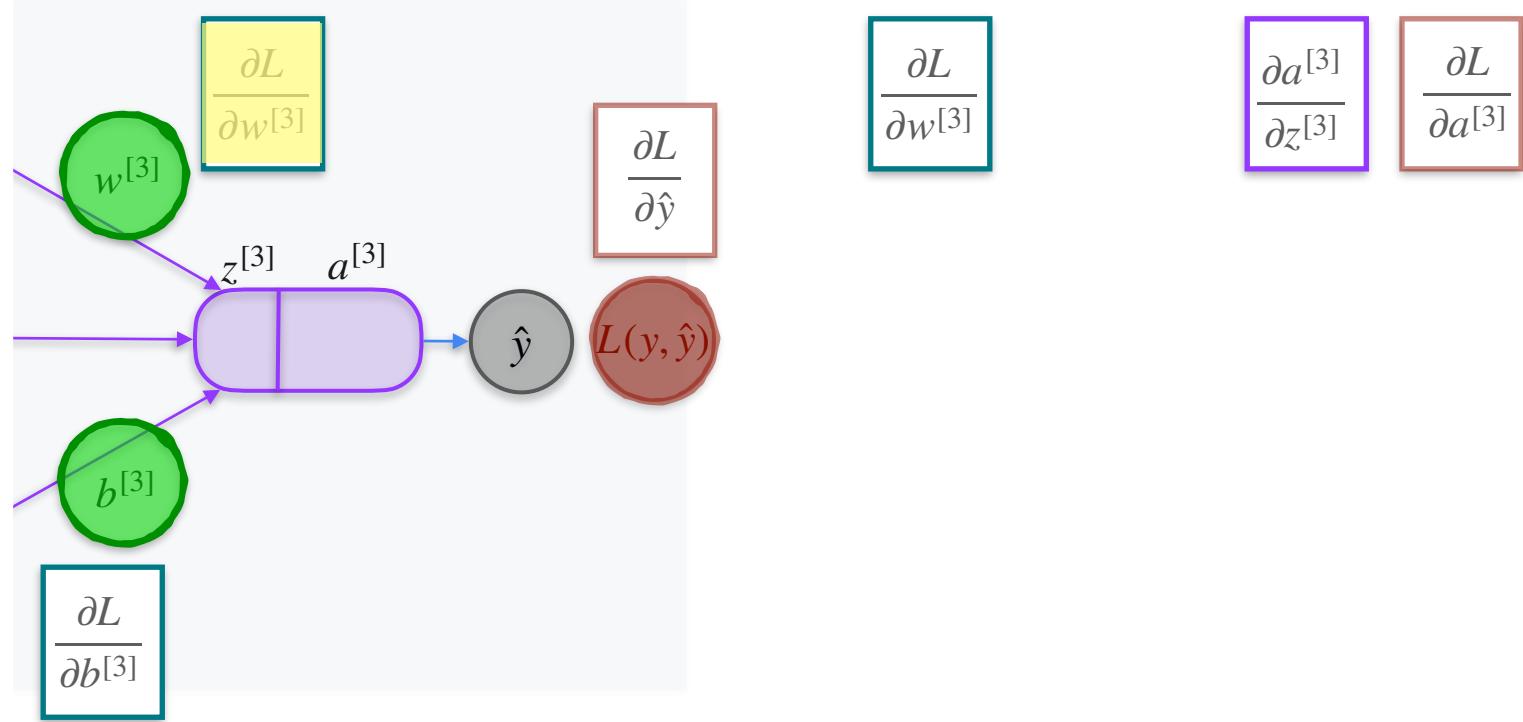
# Back Propagation Introduction



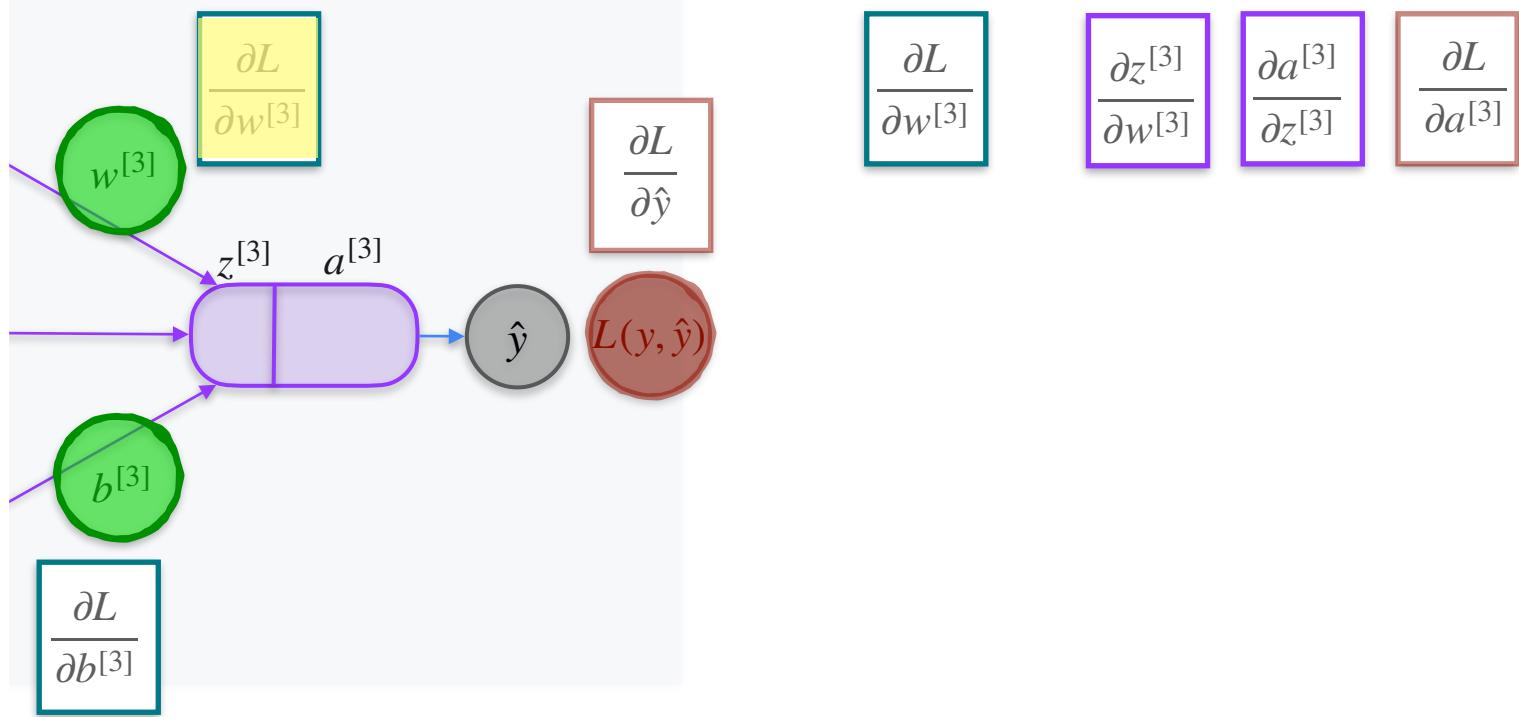
# Back Propagation Introduction



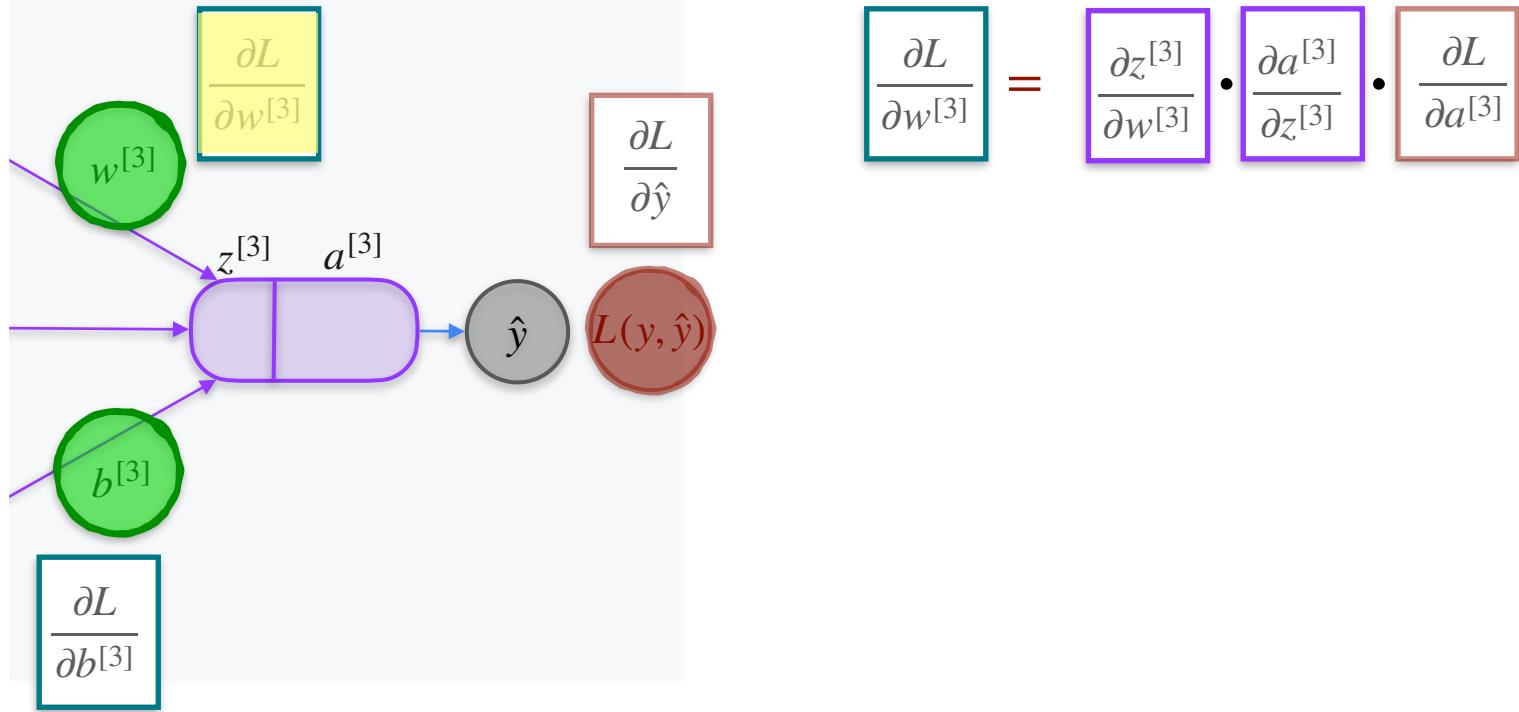
# Back Propagation Introduction



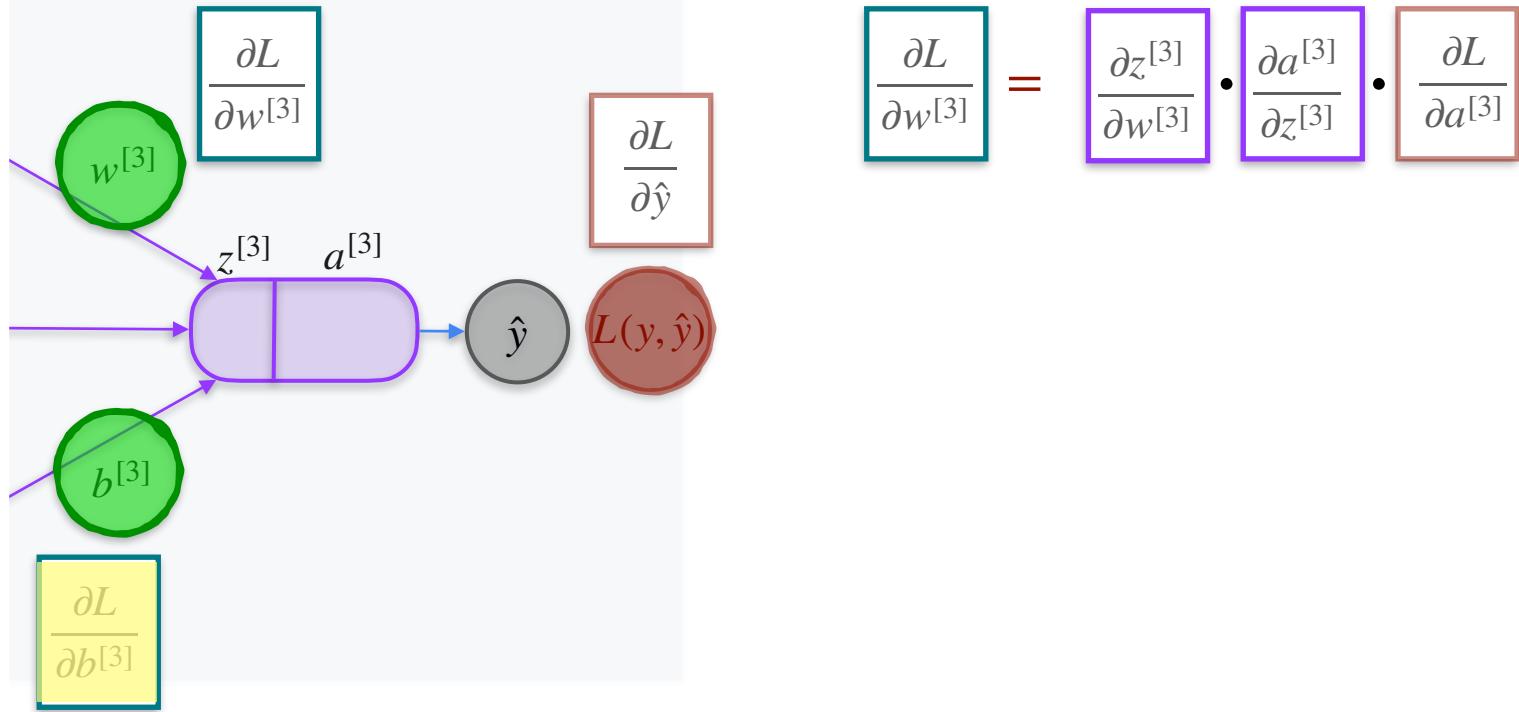
# Back Propagation Introduction



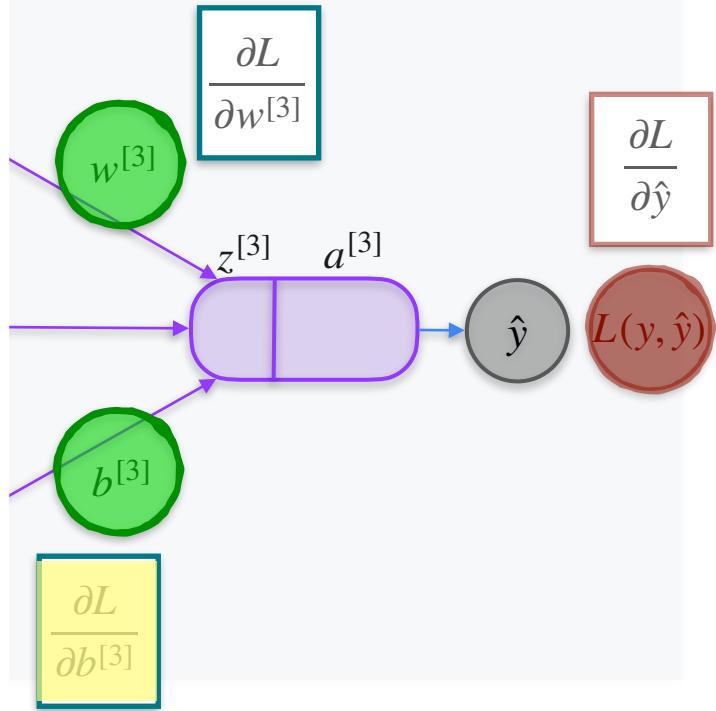
# Back Propagation Introduction



# Back Propagation Introduction



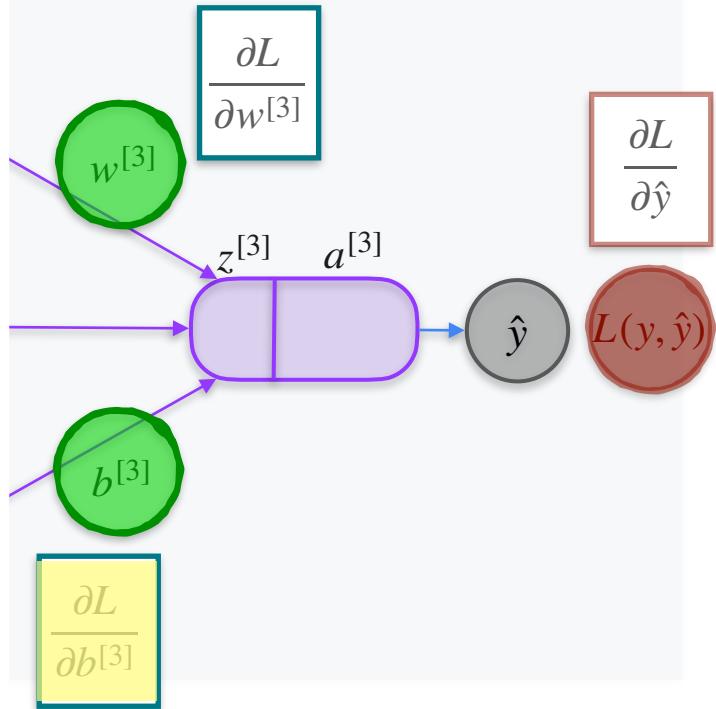
# Back Propagation Introduction



$$\frac{\partial L}{\partial w^{[3]}} = \frac{\partial z^{[3]}}{\partial w^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial L}{\partial a^{[3]}}$$

$$\frac{\partial L}{\partial b^{[3]}}$$

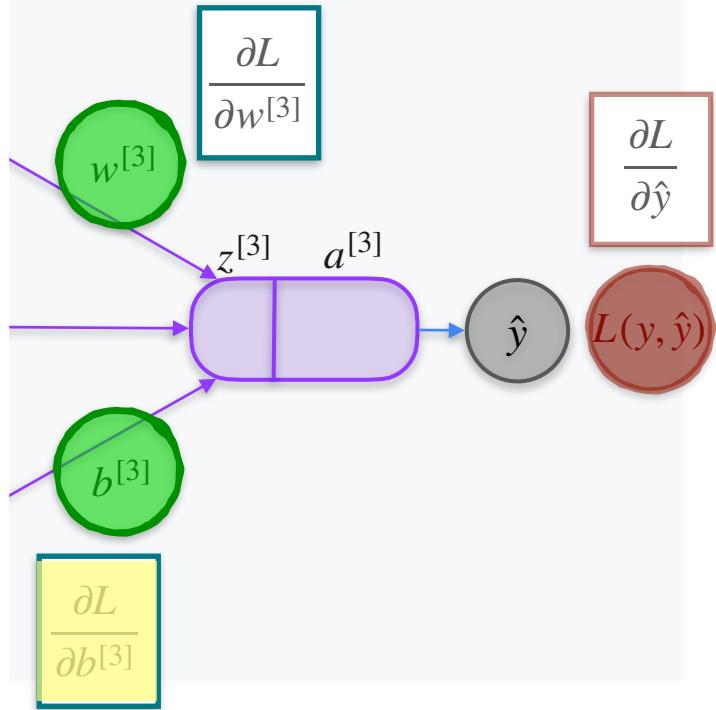
# Back Propagation Introduction



$$\frac{\partial L}{\partial w^{[3]}} = \frac{\partial z^{[3]}}{\partial w^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial L}{\partial a^{[3]}}$$

$$\frac{\partial L}{\partial b^{[3]}} \quad \frac{\partial z^{[3]}}{\partial b^{[3]}} \quad \frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

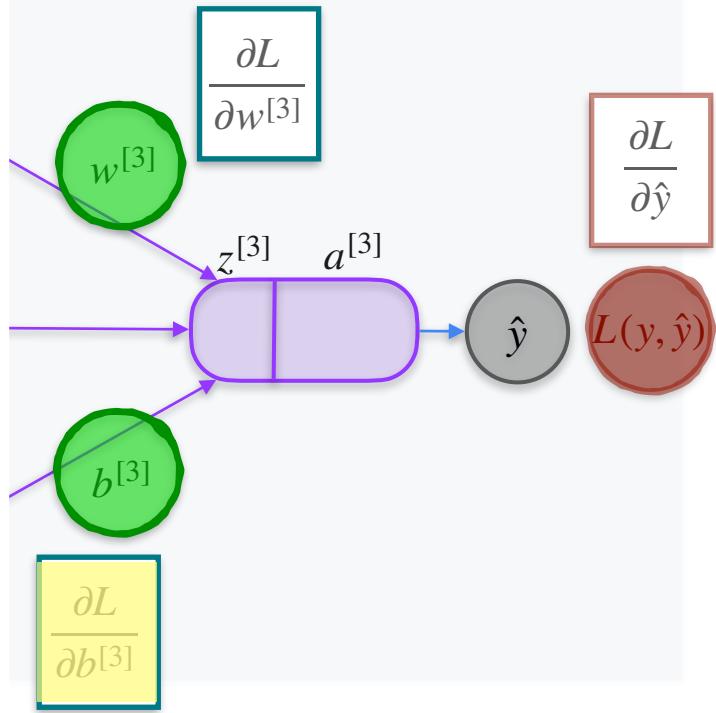
# Back Propagation Introduction



$$\frac{\partial L}{\partial w^{[3]}} = \frac{\partial z^{[3]}}{\partial w^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial L}{\partial a^{[3]}}$$

$$\frac{\partial L}{\partial b^{[3]}} = \frac{\partial z^{[3]}}{\partial b^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction

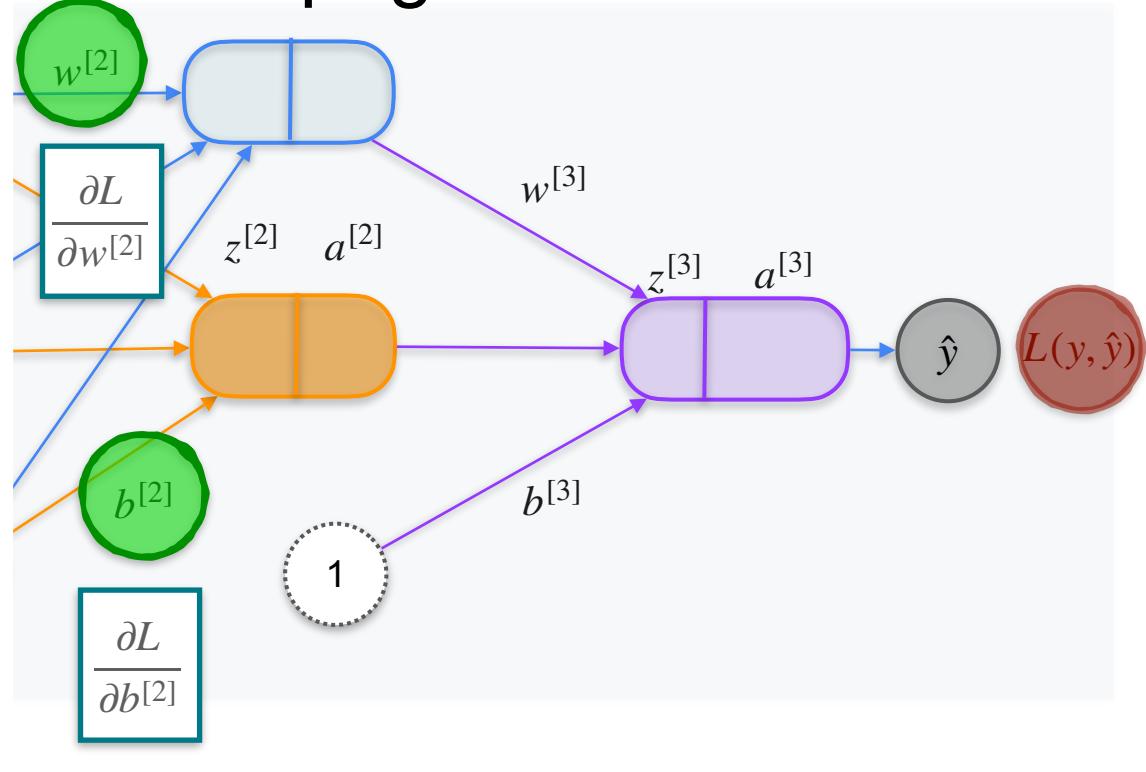


$$\frac{\partial L}{\partial w^{[3]}} = \frac{\partial z^{[3]}}{\partial w^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial L}{\partial a^{[3]}}$$

$$\frac{\partial L}{\partial b^{[3]}} = \frac{\partial z^{[3]}}{\partial b^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial L}{\partial a^{[3]}}$$

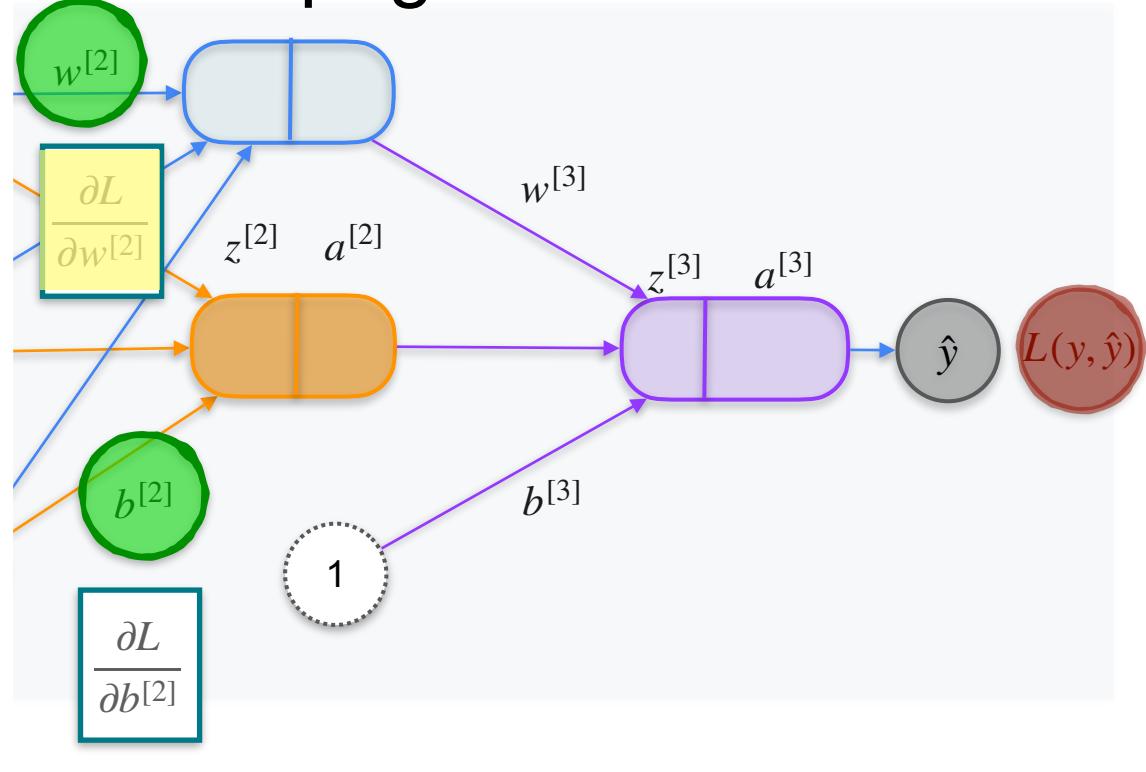
$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction



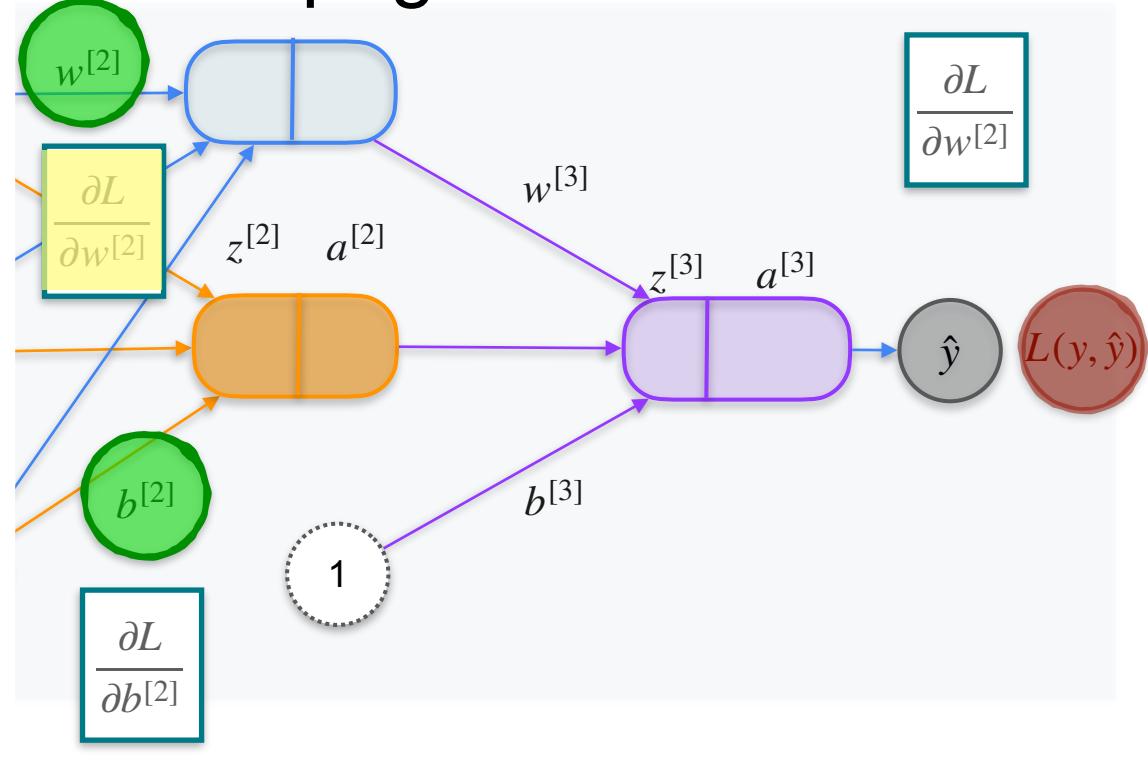
$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction



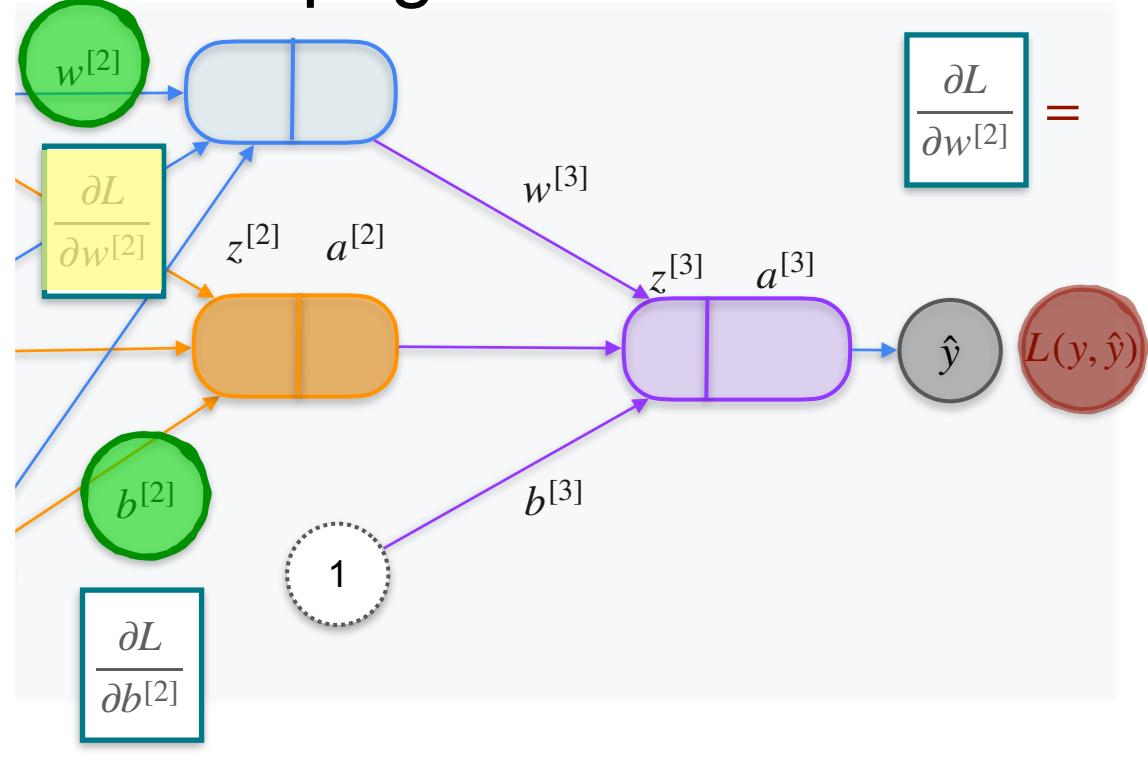
$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction



$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

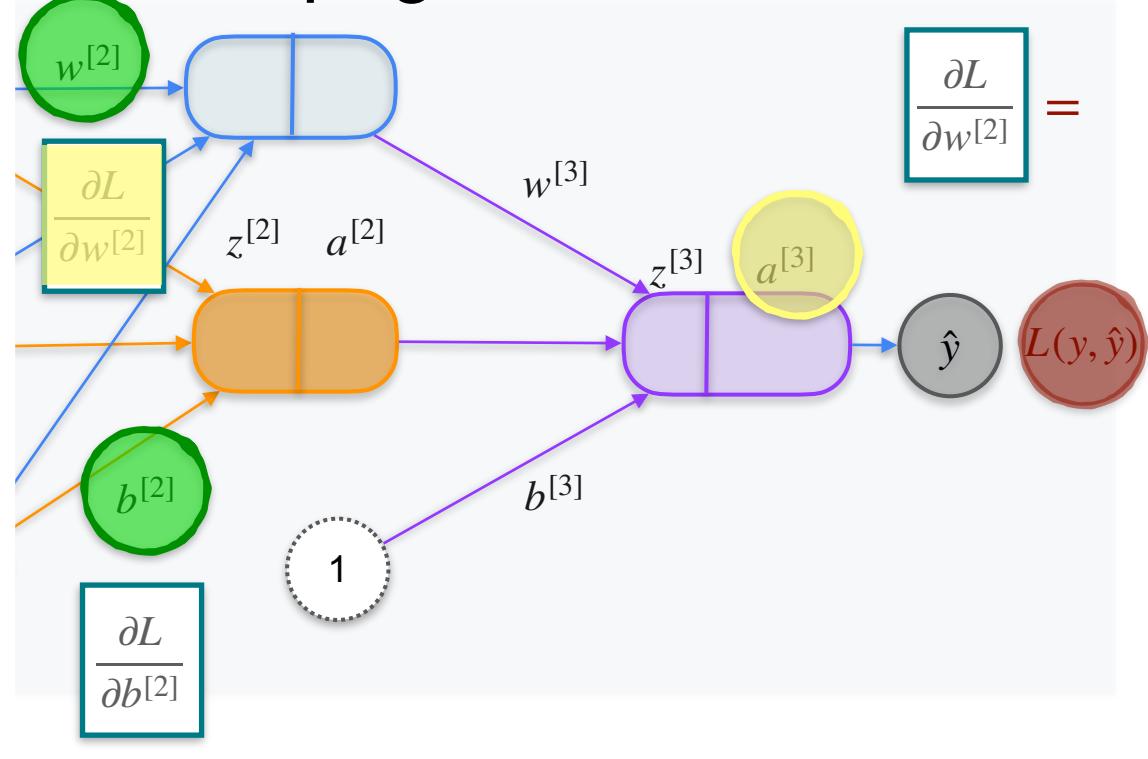
# Back Propagation Introduction



$$\frac{\partial L}{\partial w^{[2]}} =$$

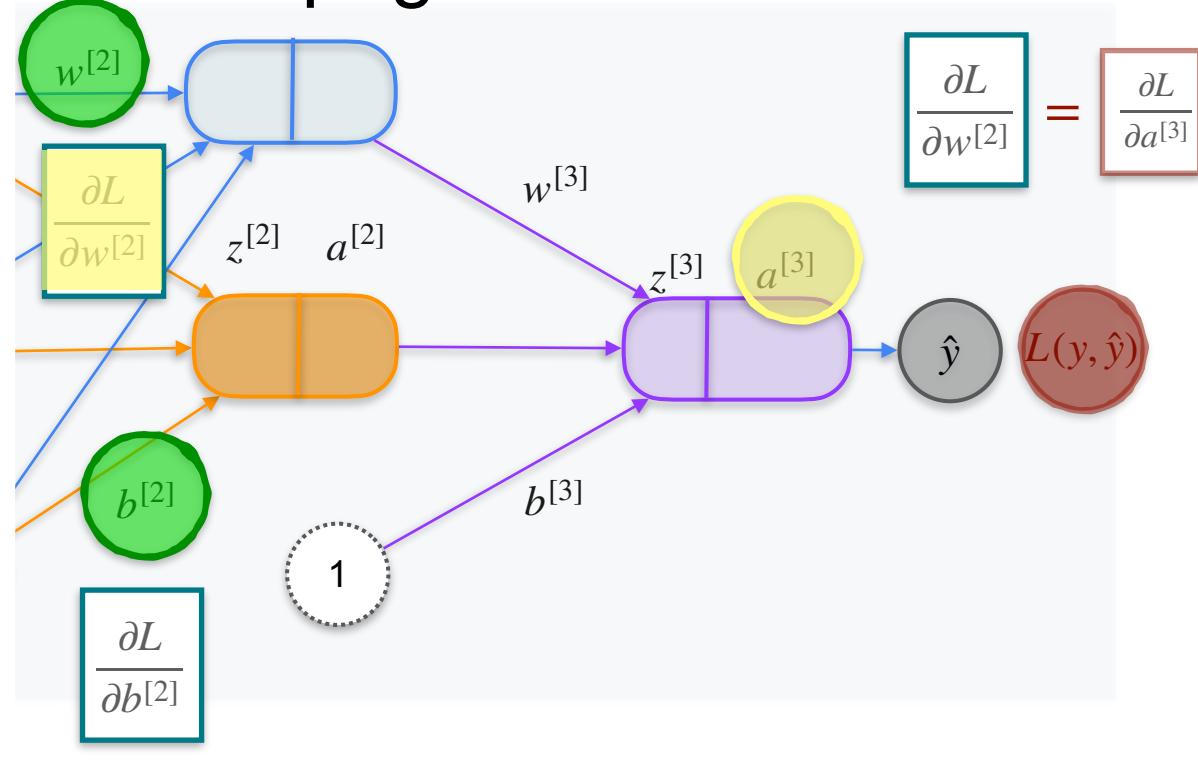
$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction



$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

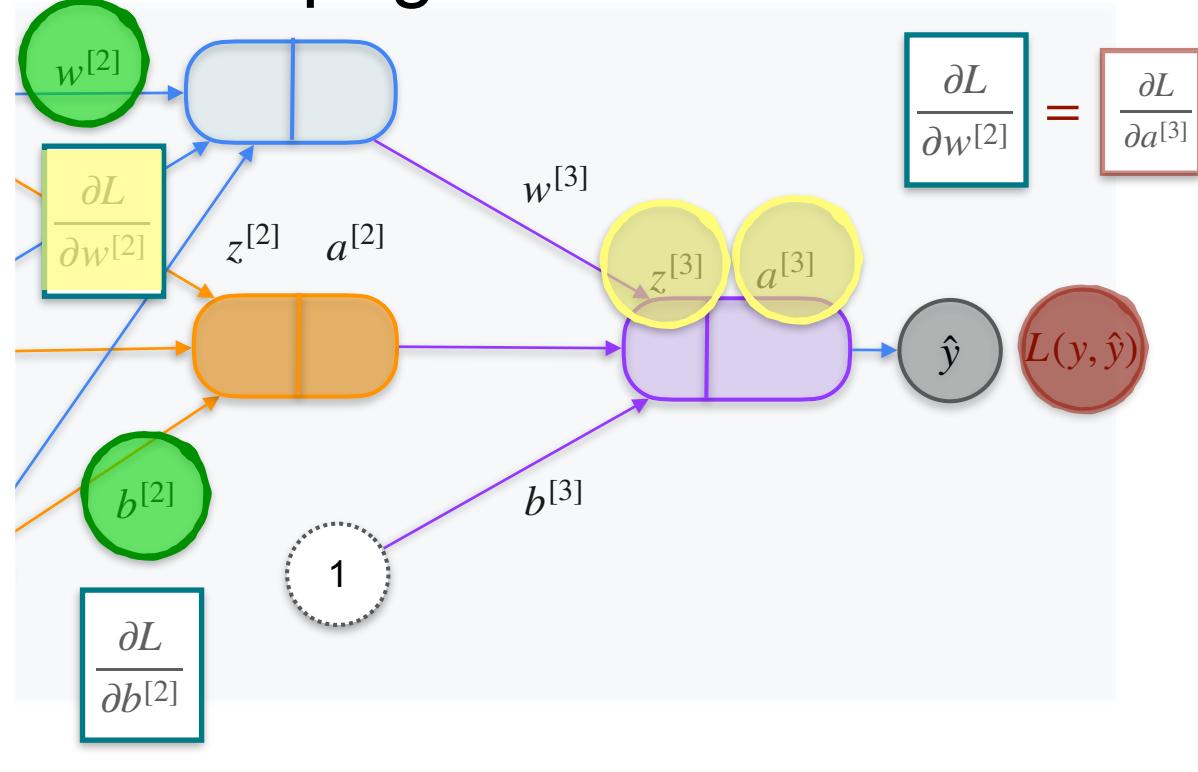
# Back Propagation Introduction



$$\frac{\partial L}{\partial w^{[2]}} = \frac{\partial L}{\partial a^{[3]}}$$

$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

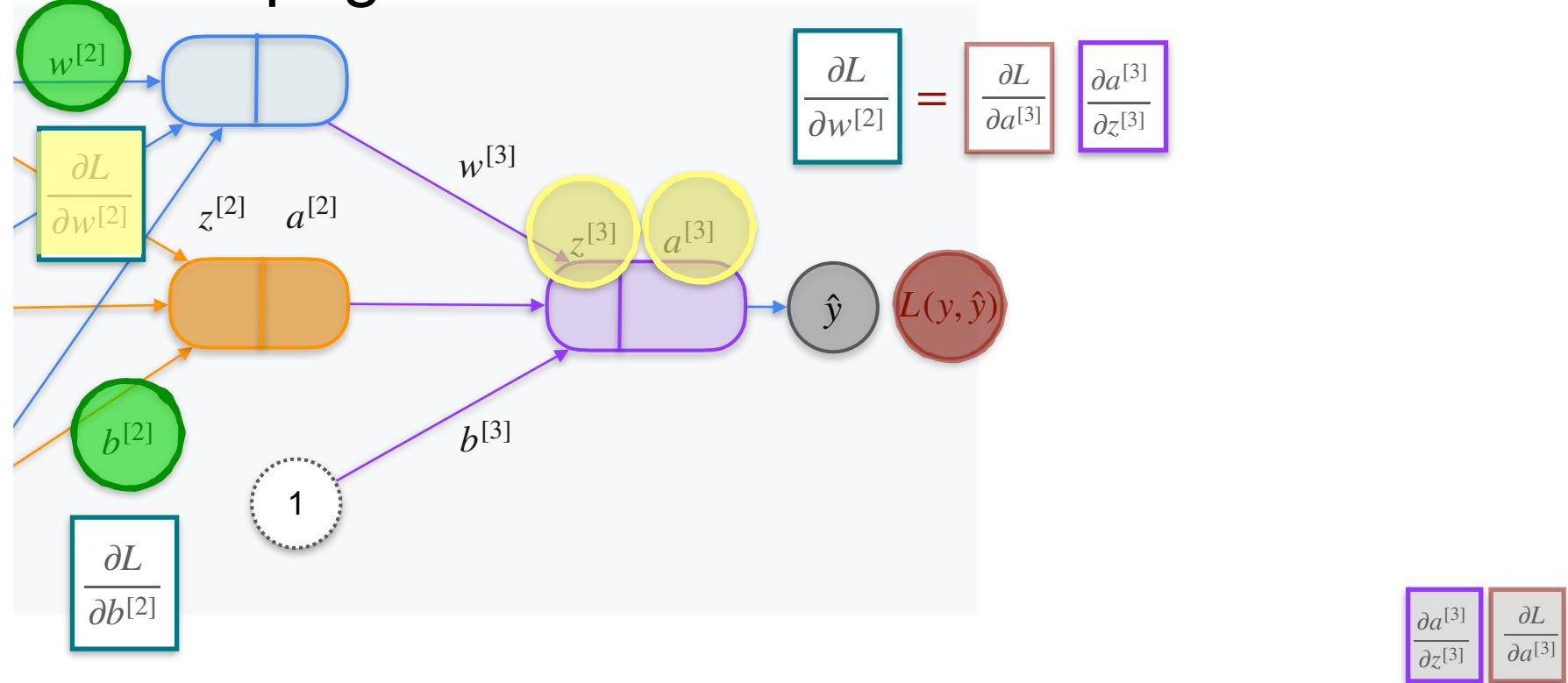
# Back Propagation Introduction



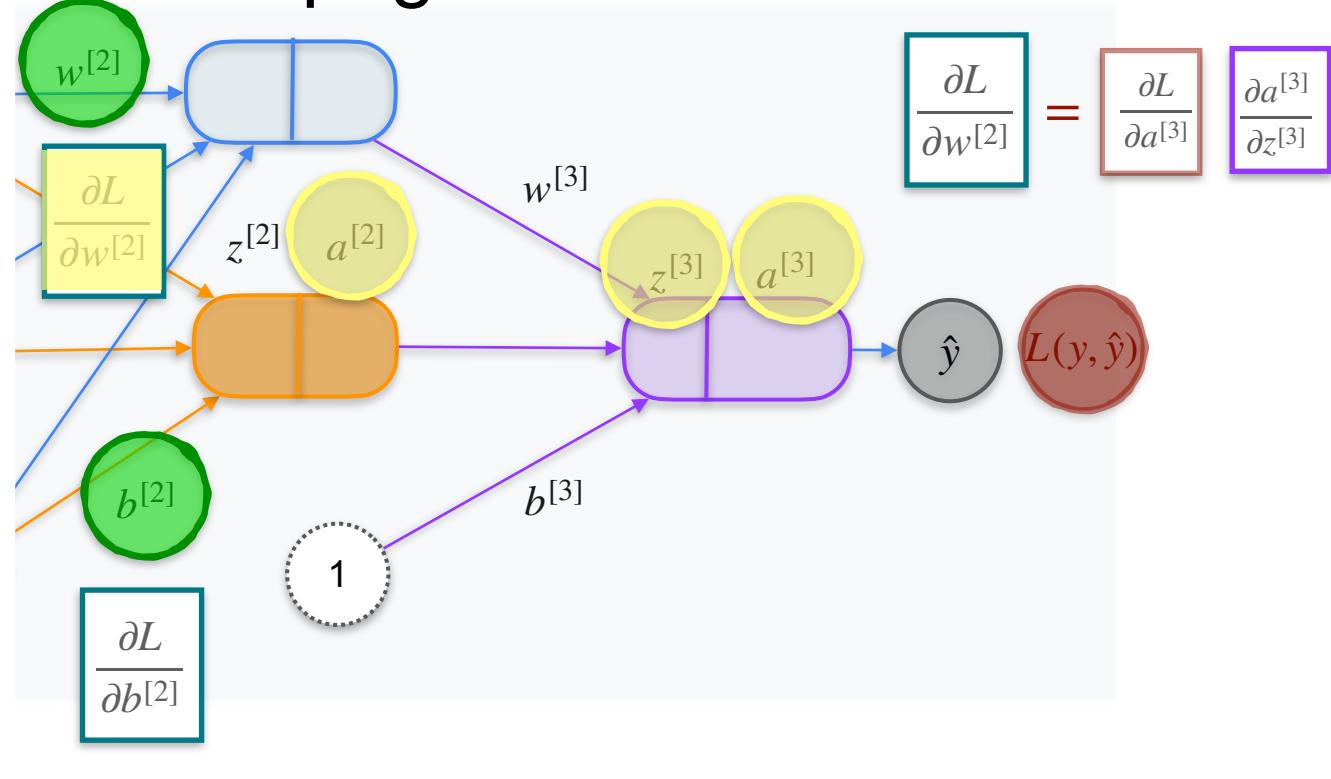
$$\frac{\partial L}{\partial w^{[2]}} = \frac{\partial L}{\partial a^{[3]}}$$

$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction



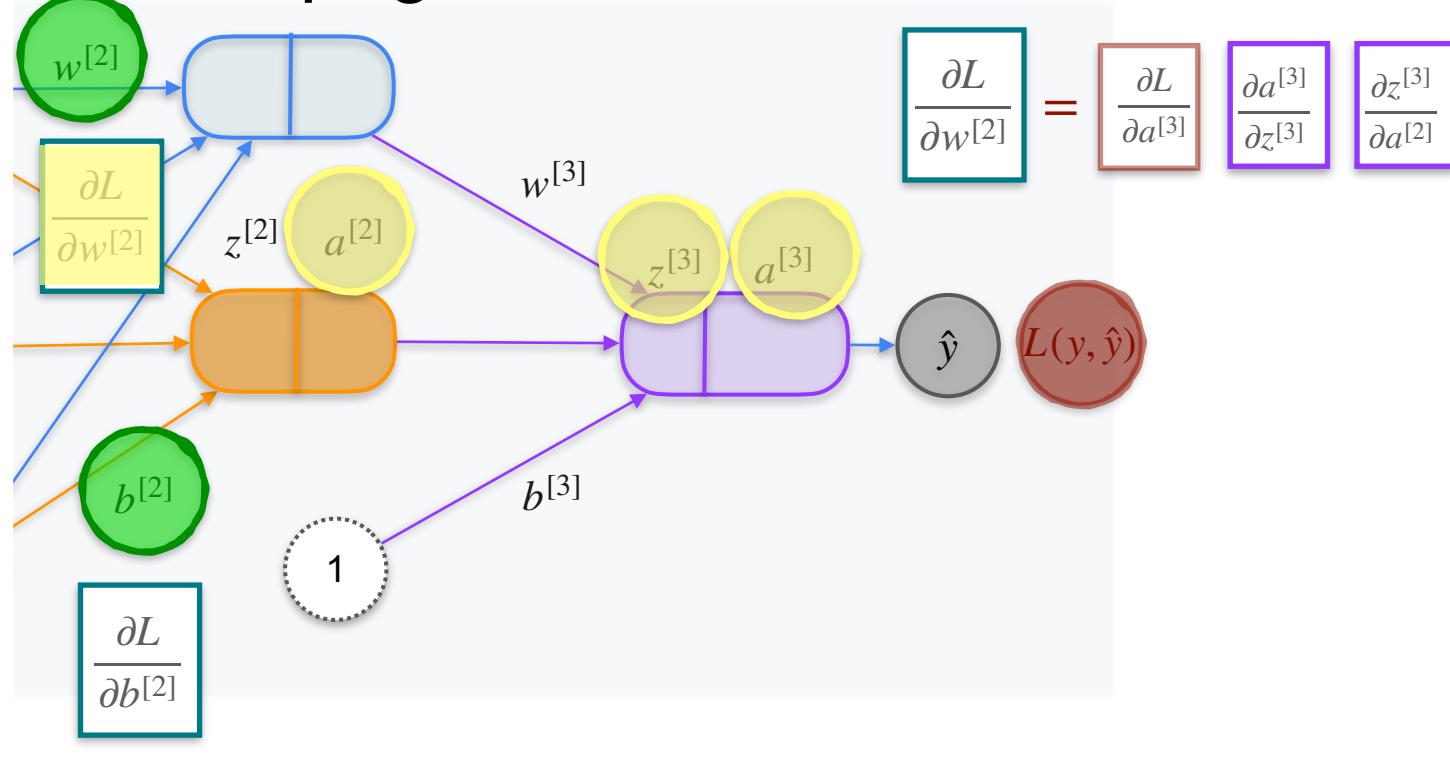
# Back Propagation Introduction



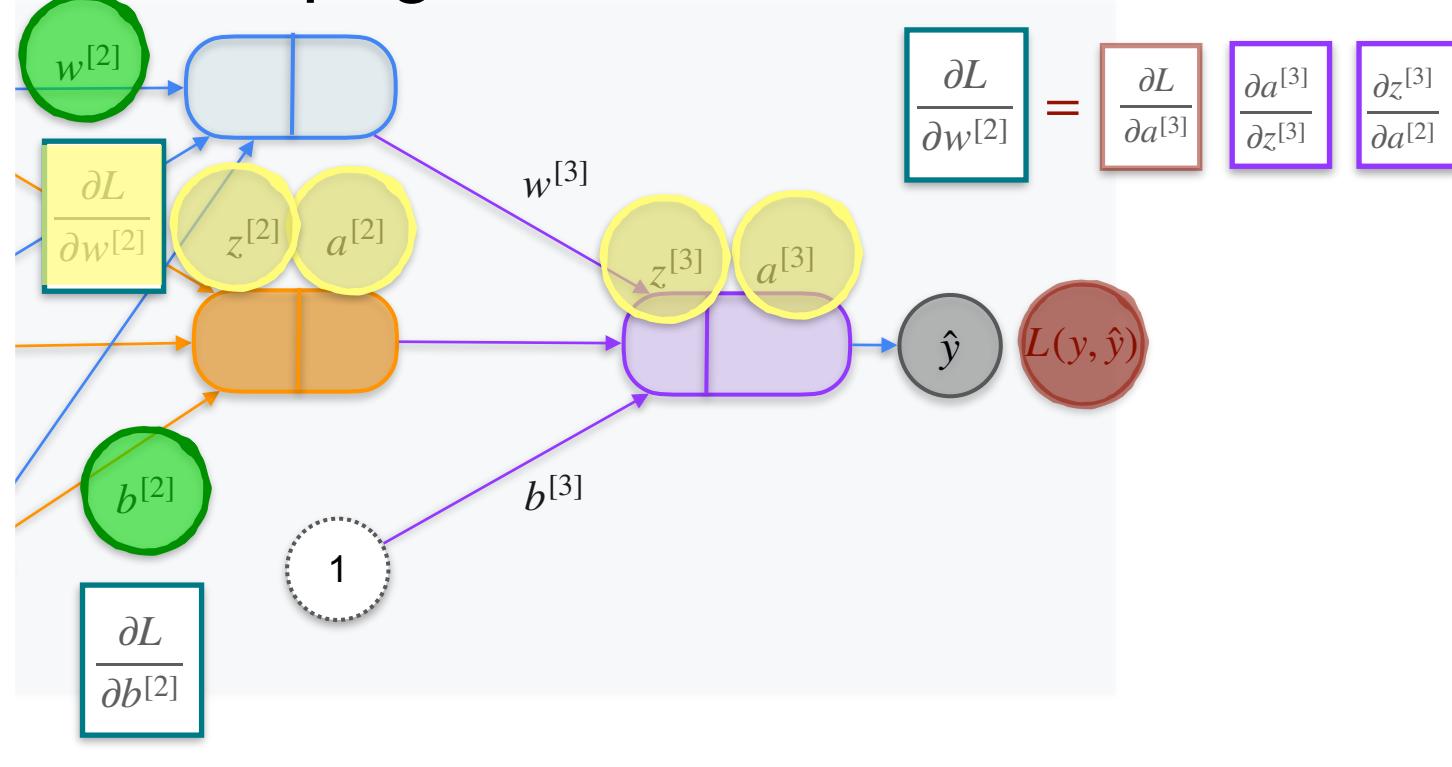
$$\frac{\partial a^{[3]}}{\partial z^{[3]}}$$

$$\frac{\partial L}{\partial a^{[3]}}$$

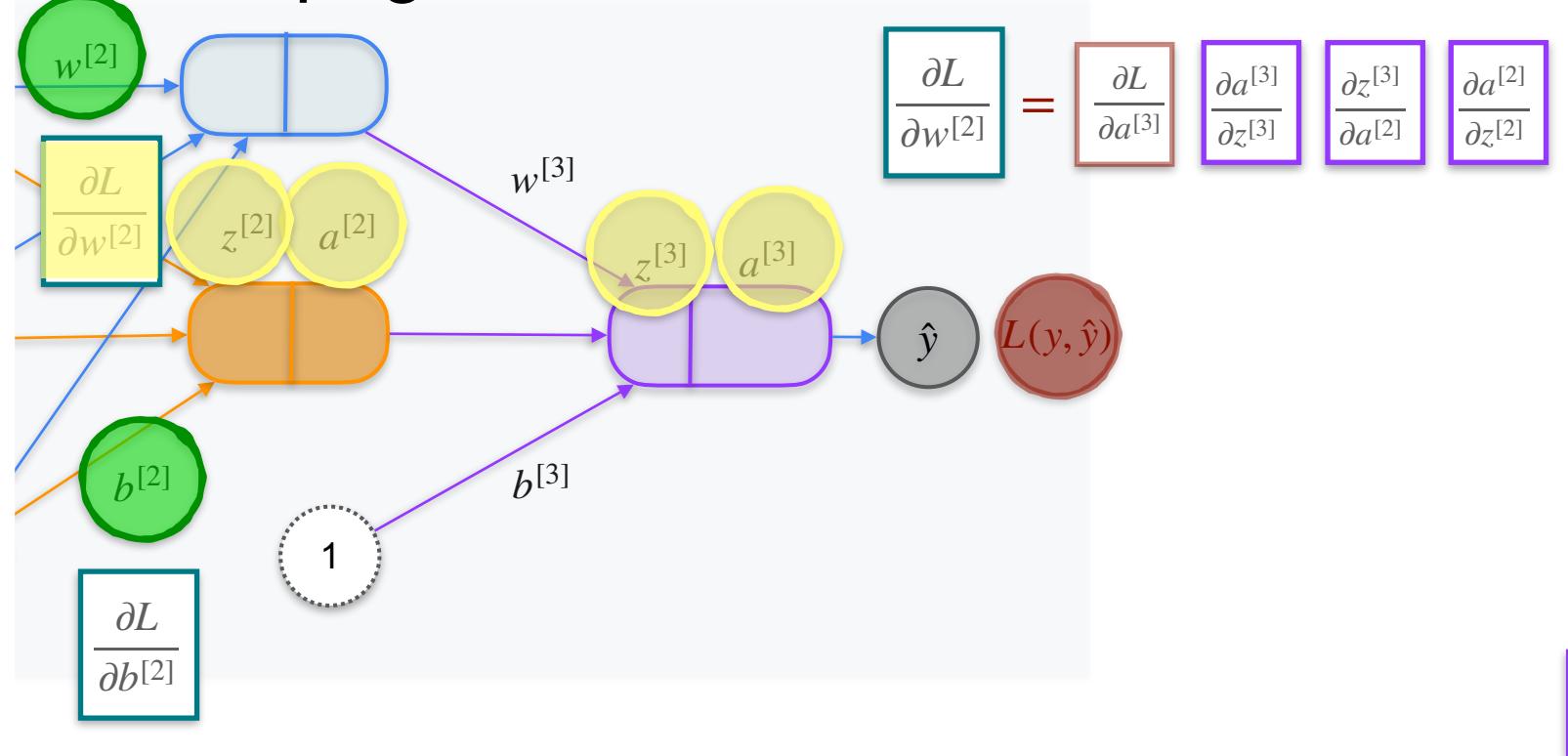
# Back Propagation Introduction



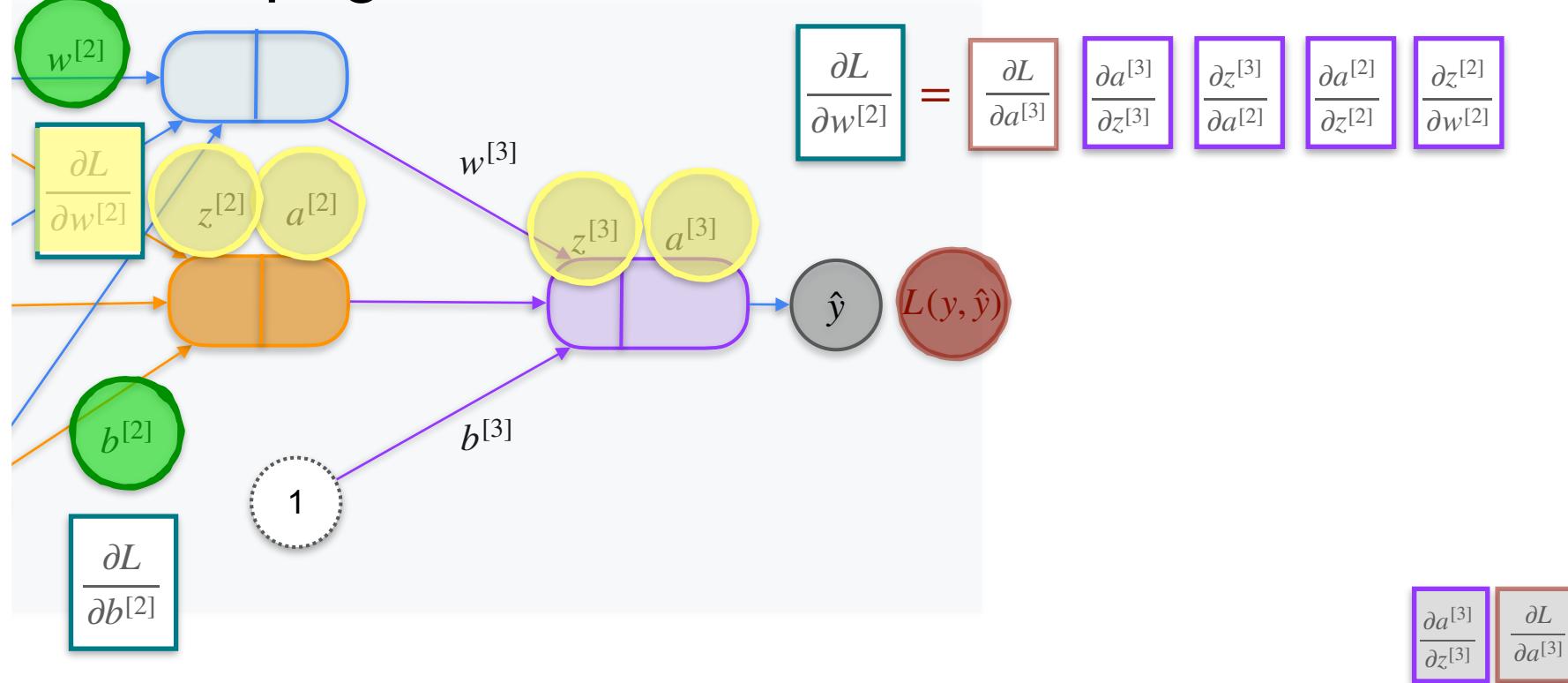
# Back Propagation Introduction



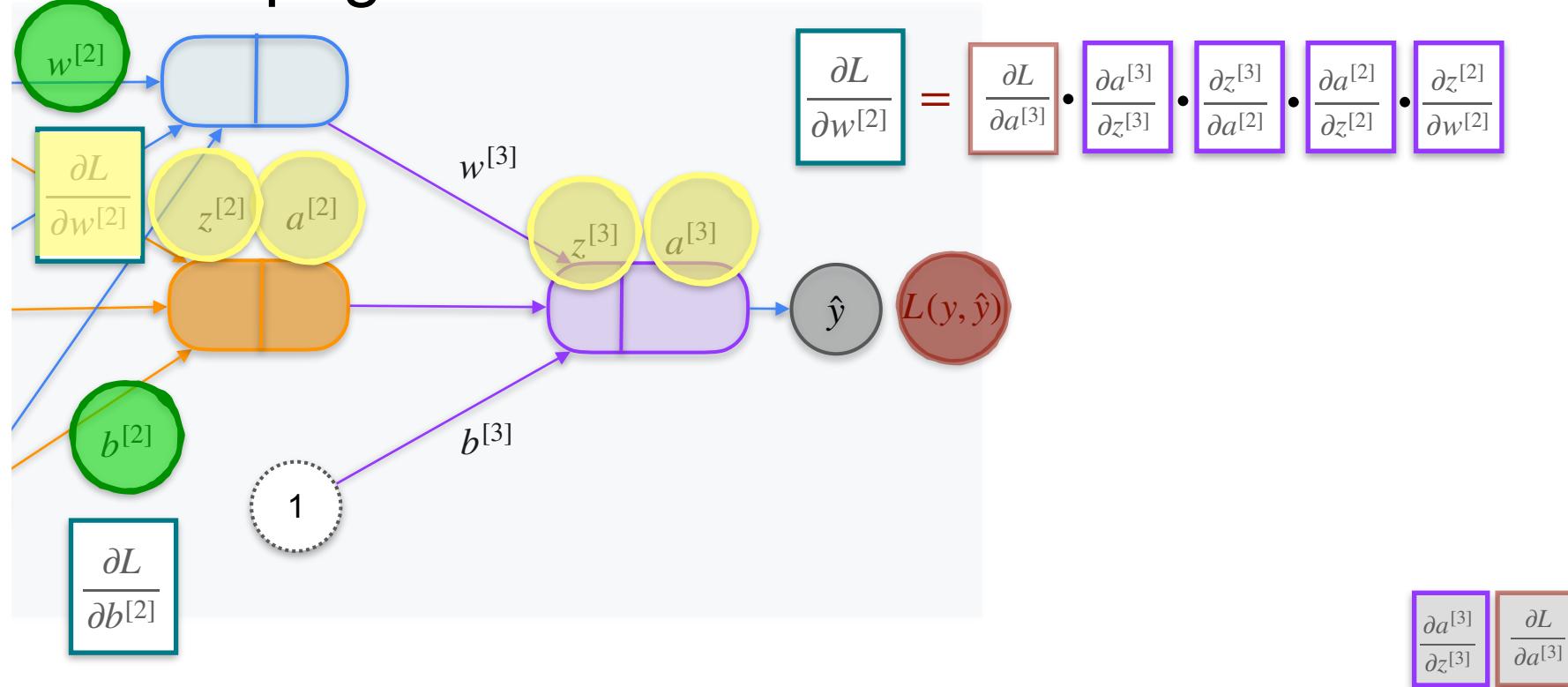
# Back Propagation Introduction



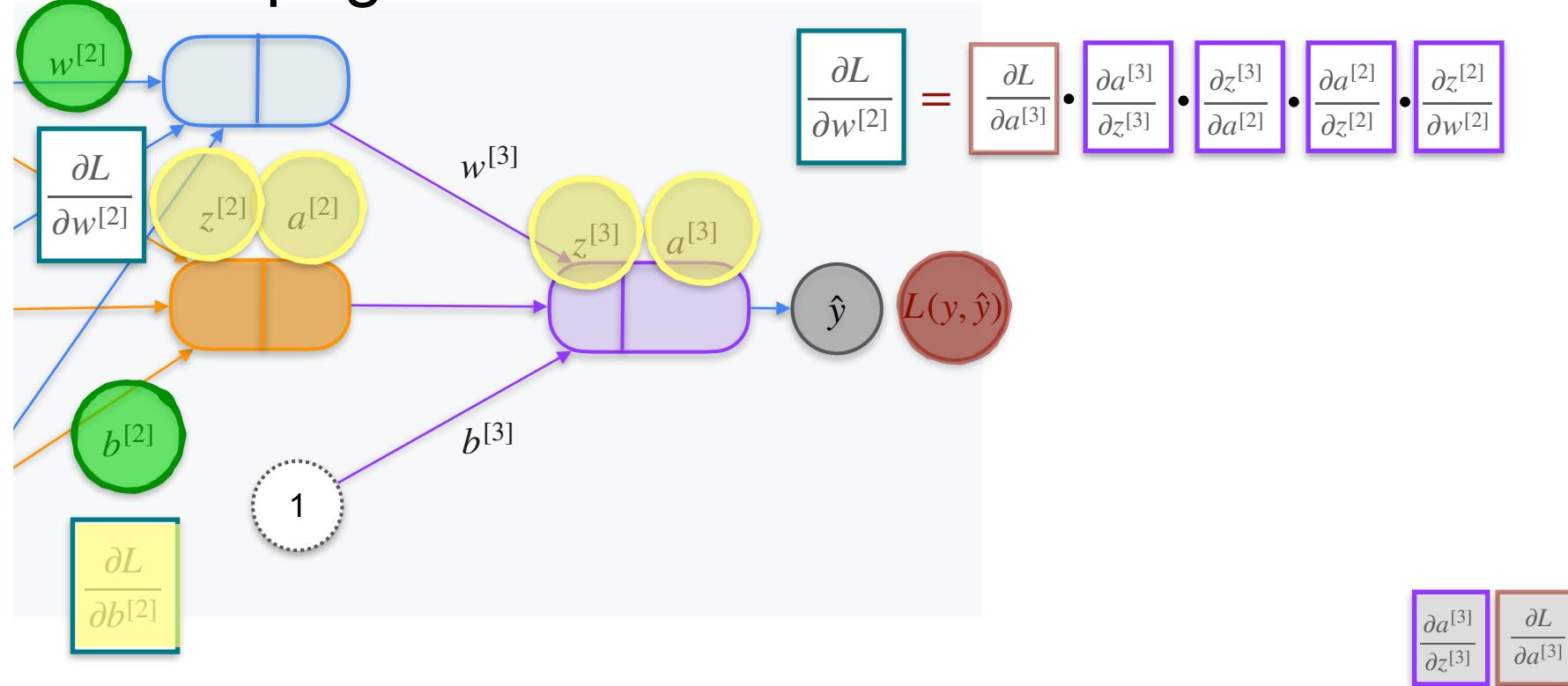
# Back Propagation Introduction



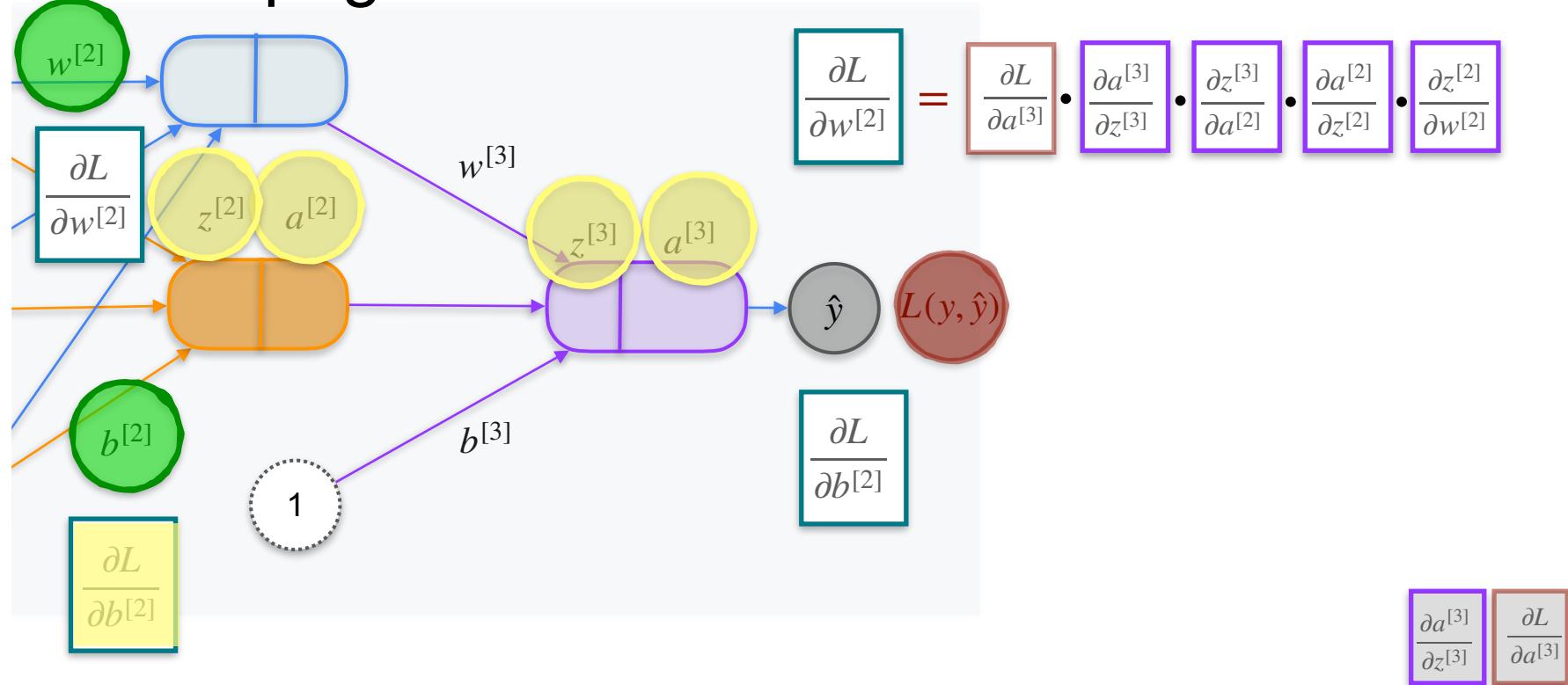
# Back Propagation Introduction



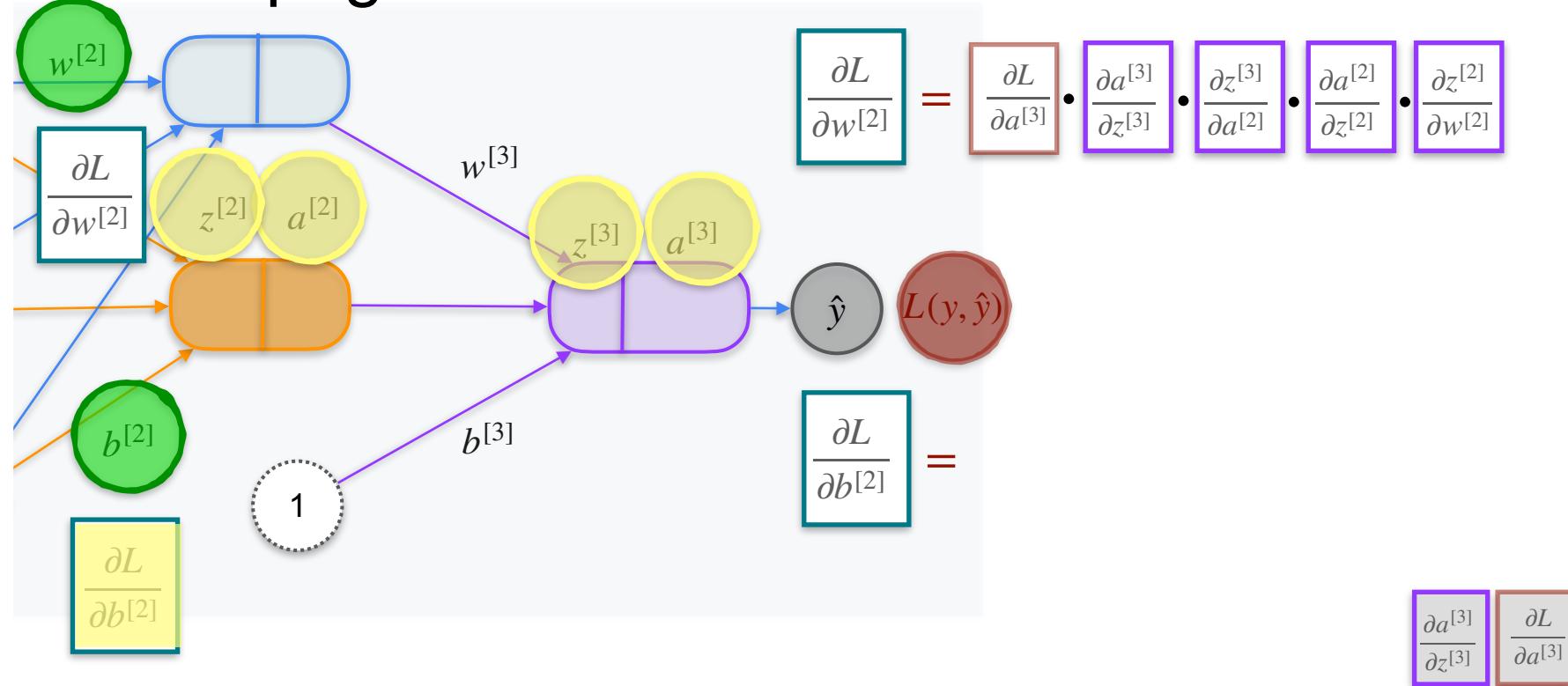
# Back Propagation Introduction



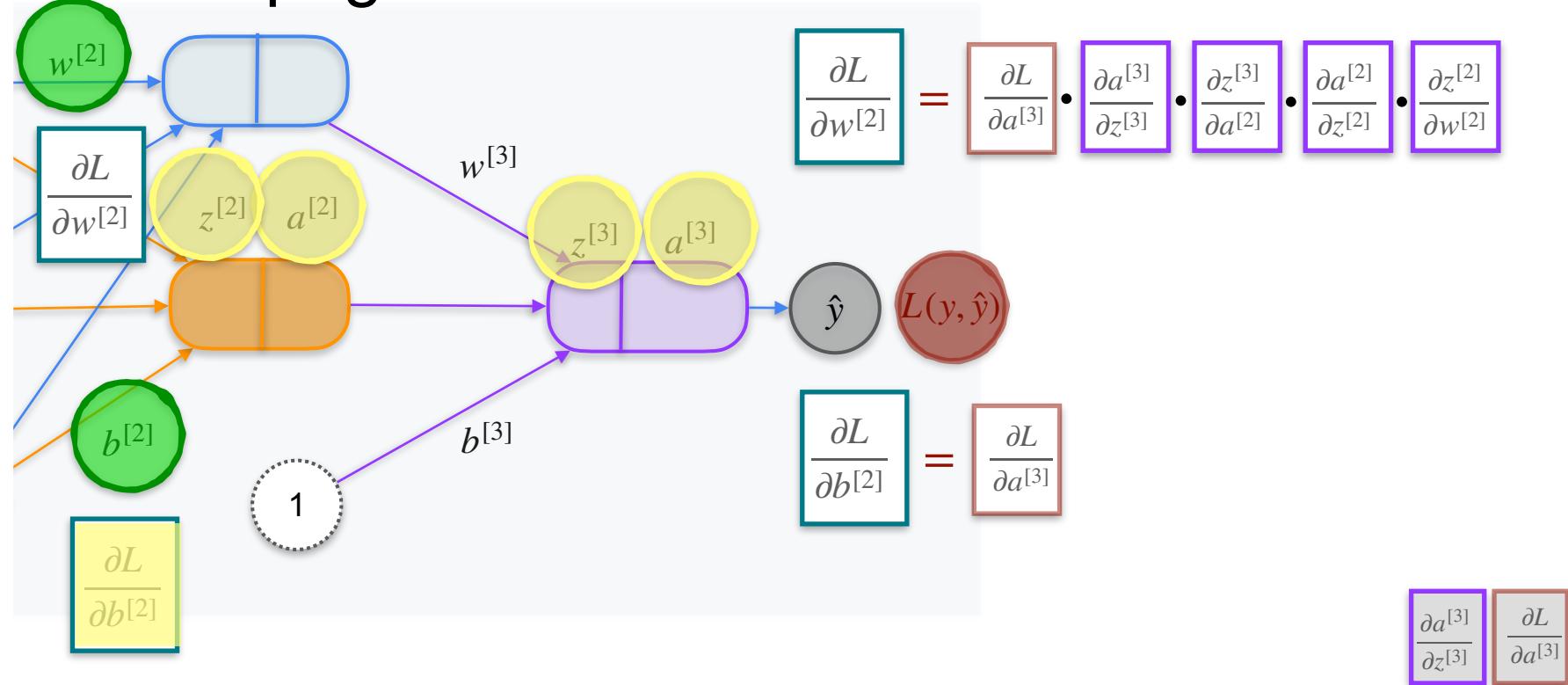
# Back Propagation Introduction



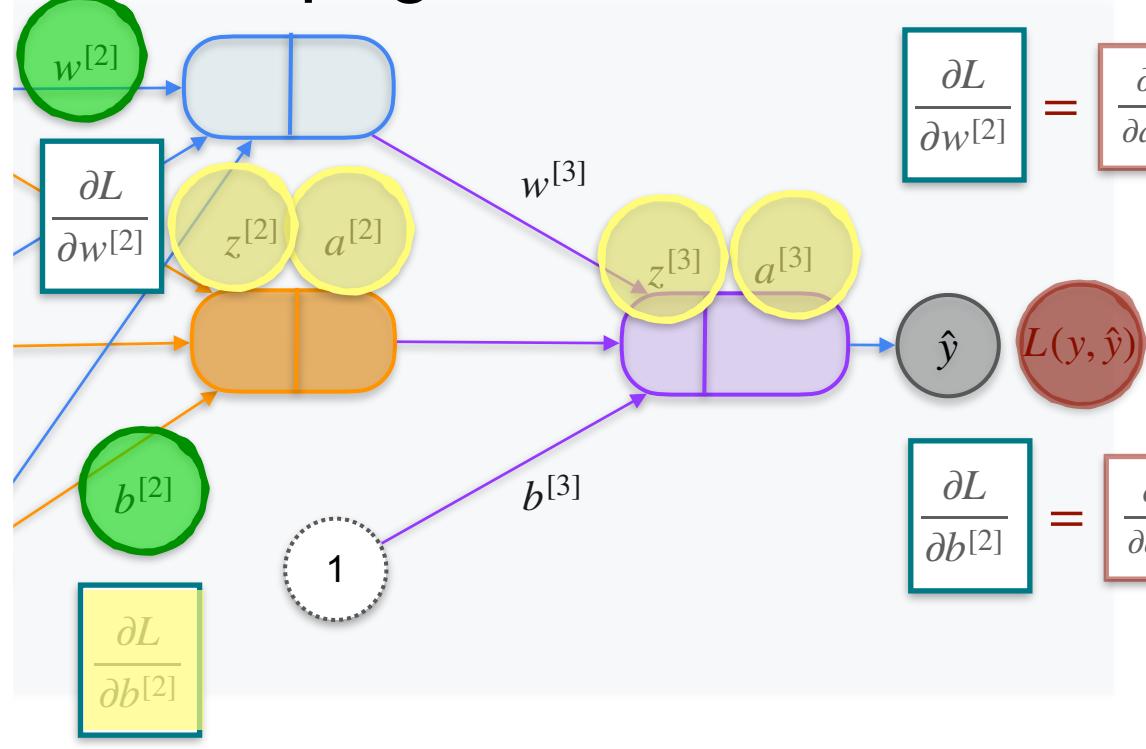
# Back Propagation Introduction



# Back Propagation Introduction



# Back Propagation Introduction

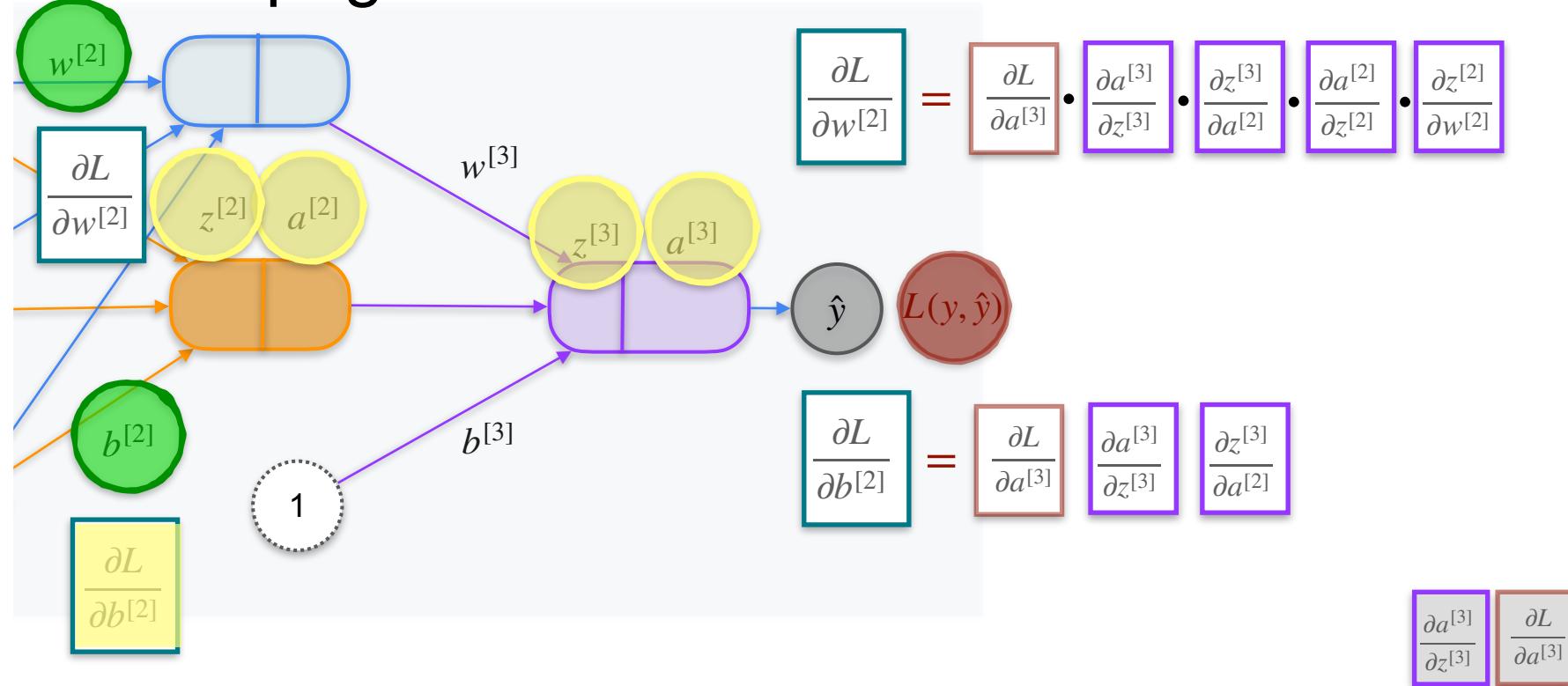


$$\frac{\partial L}{\partial w^{[2]}} = \frac{\partial L}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial w^{[2]}}$$

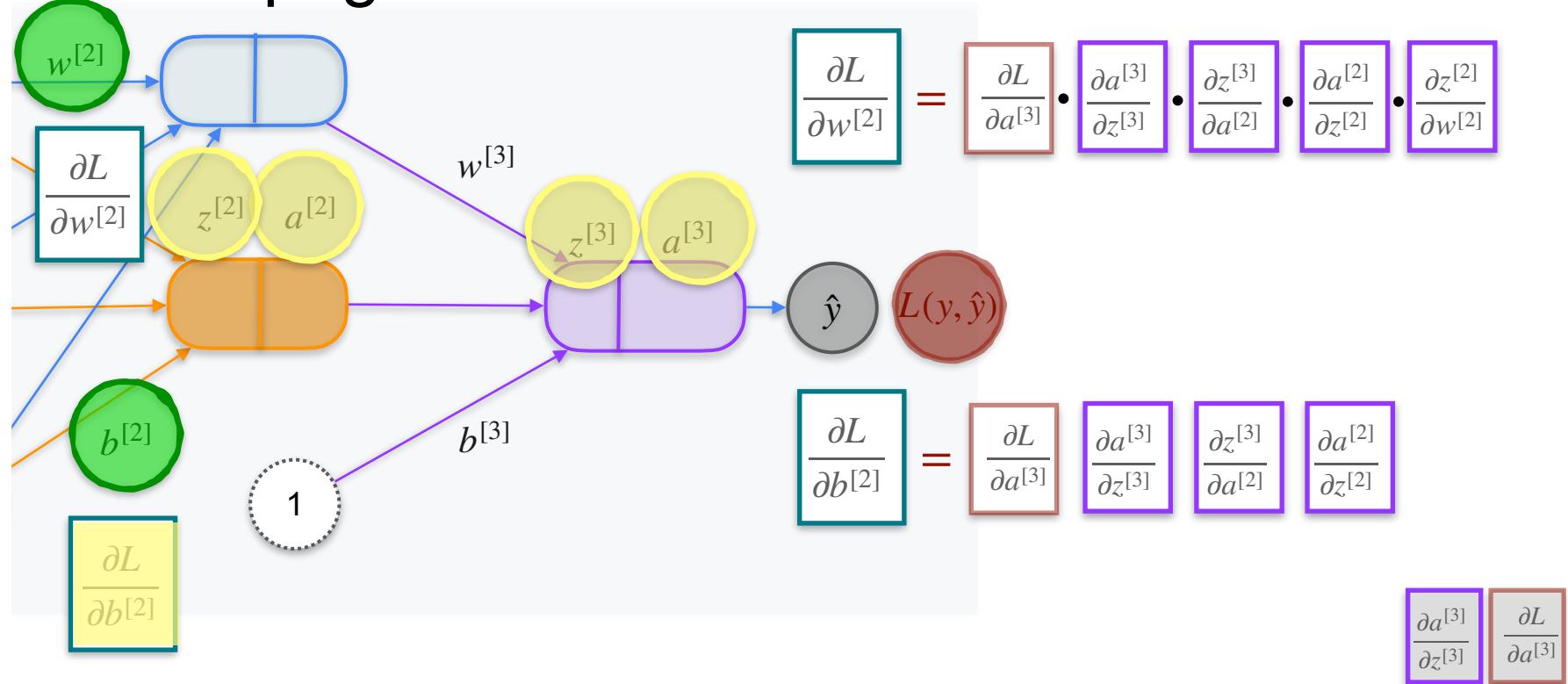
$$\frac{\partial L}{\partial b^{[2]}} = \frac{\partial L}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}}$$

$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

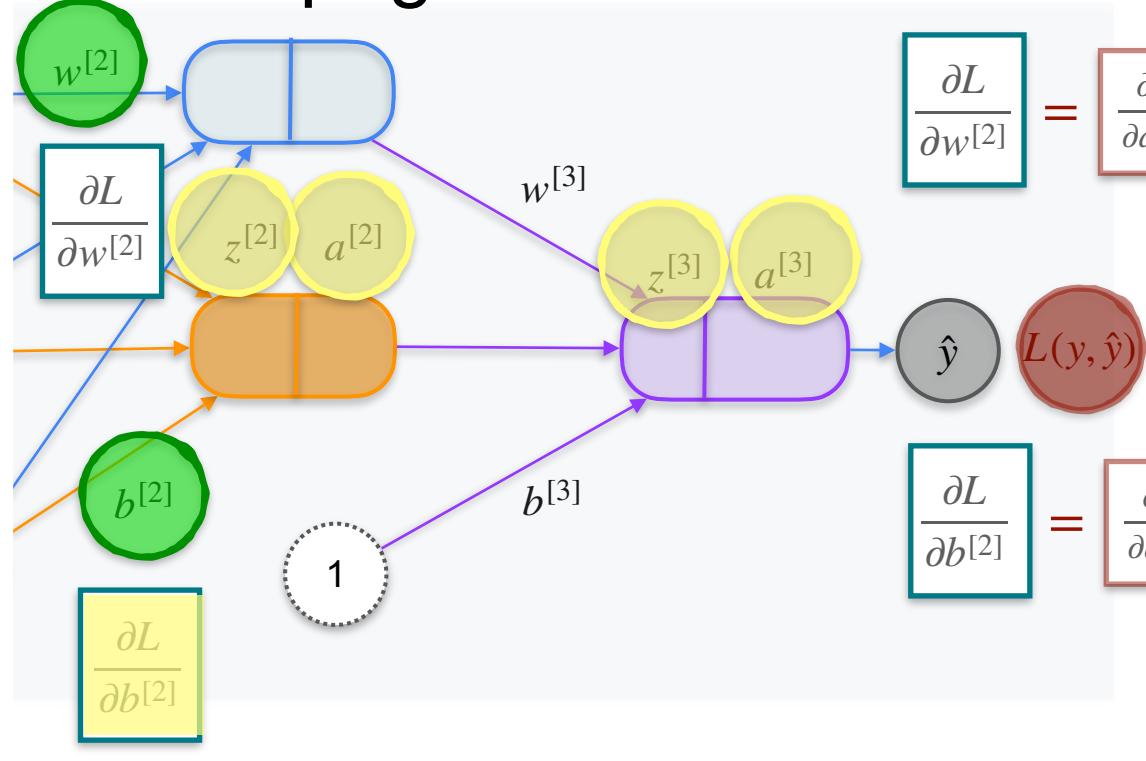
# Back Propagation Introduction



# Back Propagation Introduction

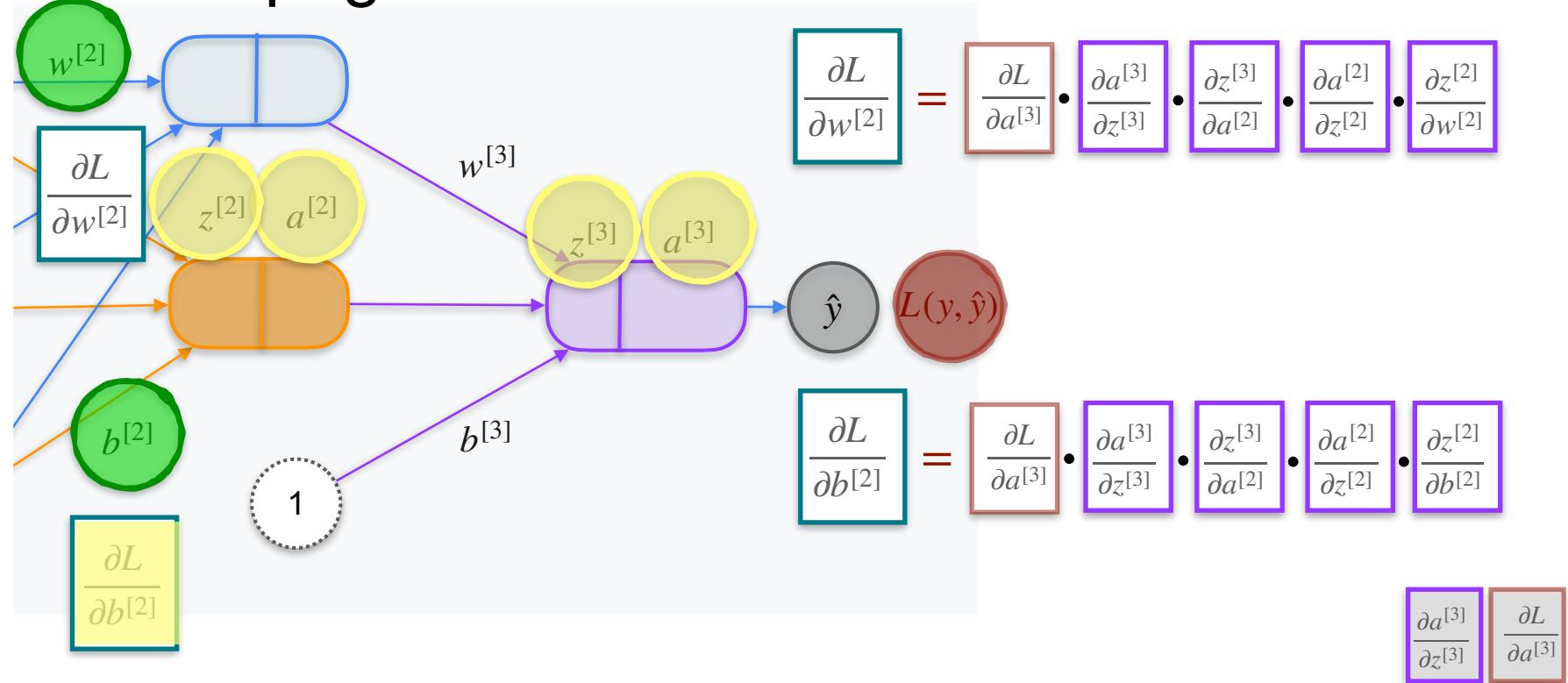


# Back Propagation Introduction

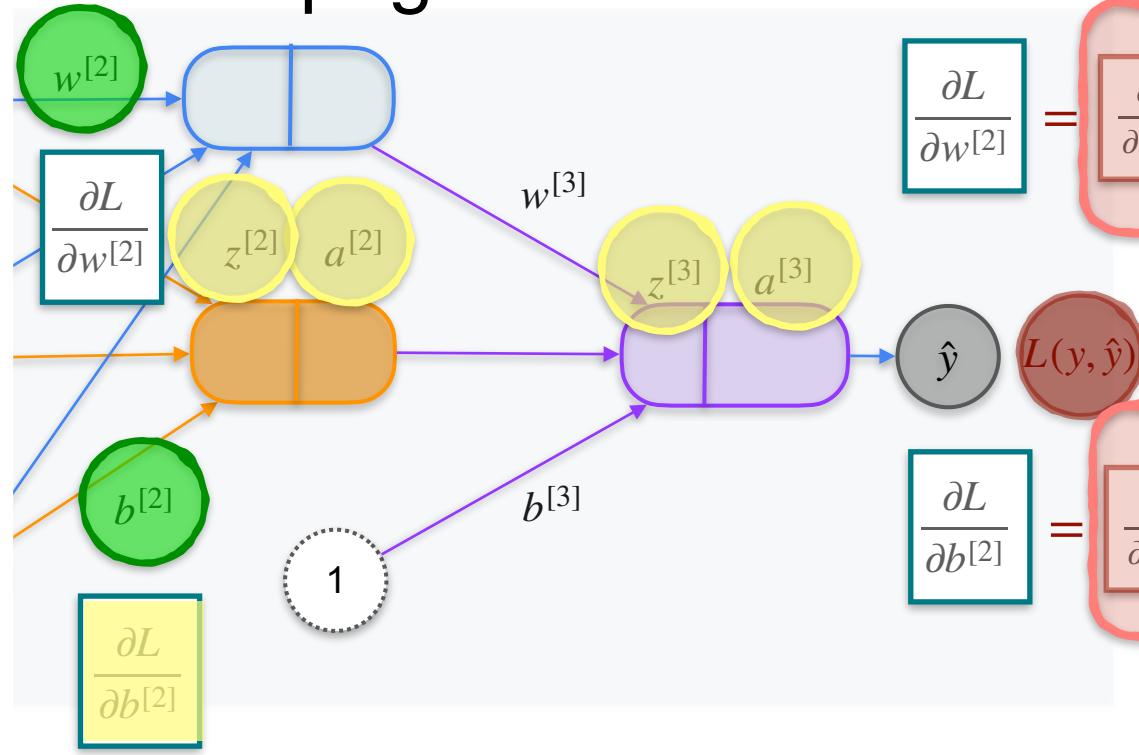


$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

# Back Propagation Introduction



# Back Propagation Introduction

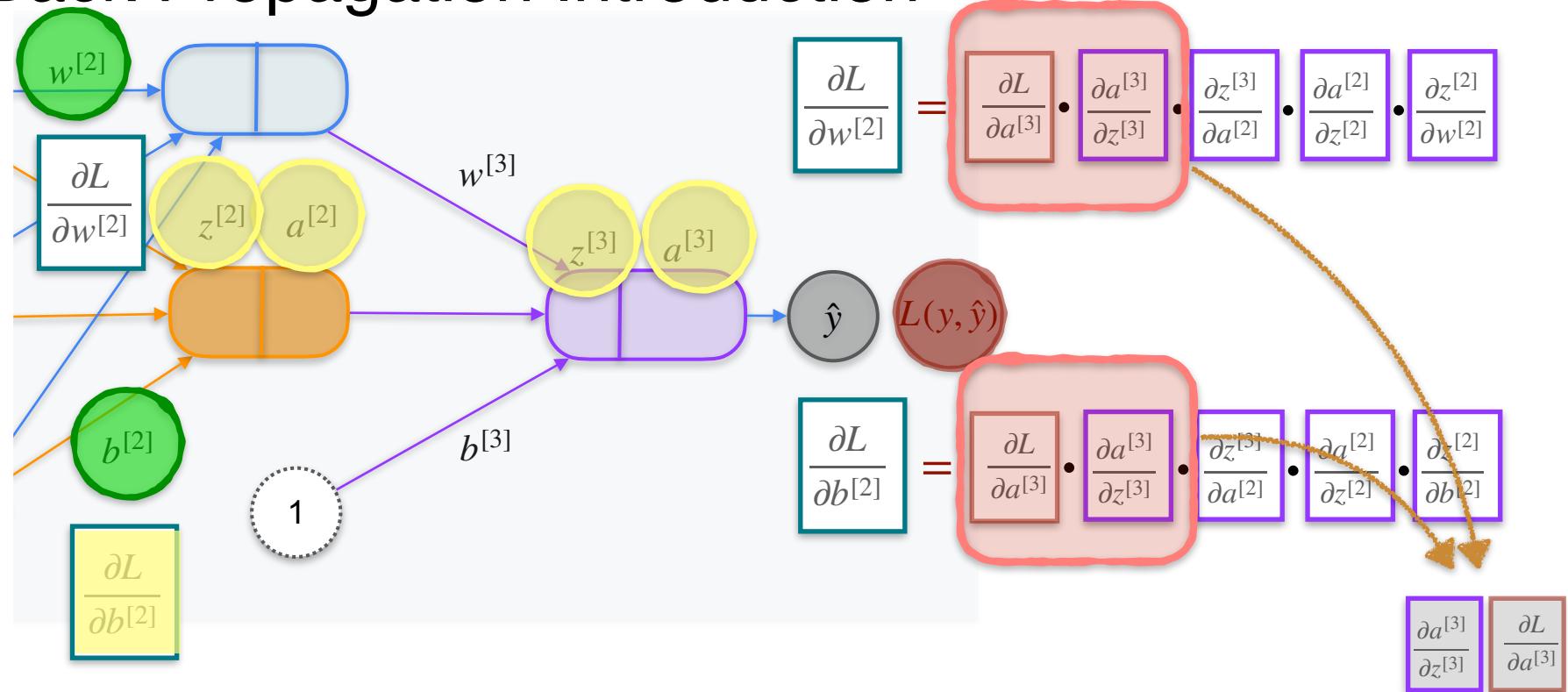


$$\frac{\partial L}{\partial w^{[2]}} = \boxed{\frac{\partial L}{\partial a^{[3]}}} \cdot \boxed{\frac{\partial a^{[3]}}{\partial z^{[3]}}} \cdot \boxed{\frac{\partial z^{[3]}}{\partial a^{[2]}}} \cdot \boxed{\frac{\partial a^{[2]}}{\partial z^{[2]}}} \cdot \boxed{\frac{\partial z^{[2]}}{\partial w^{[2]}}}$$

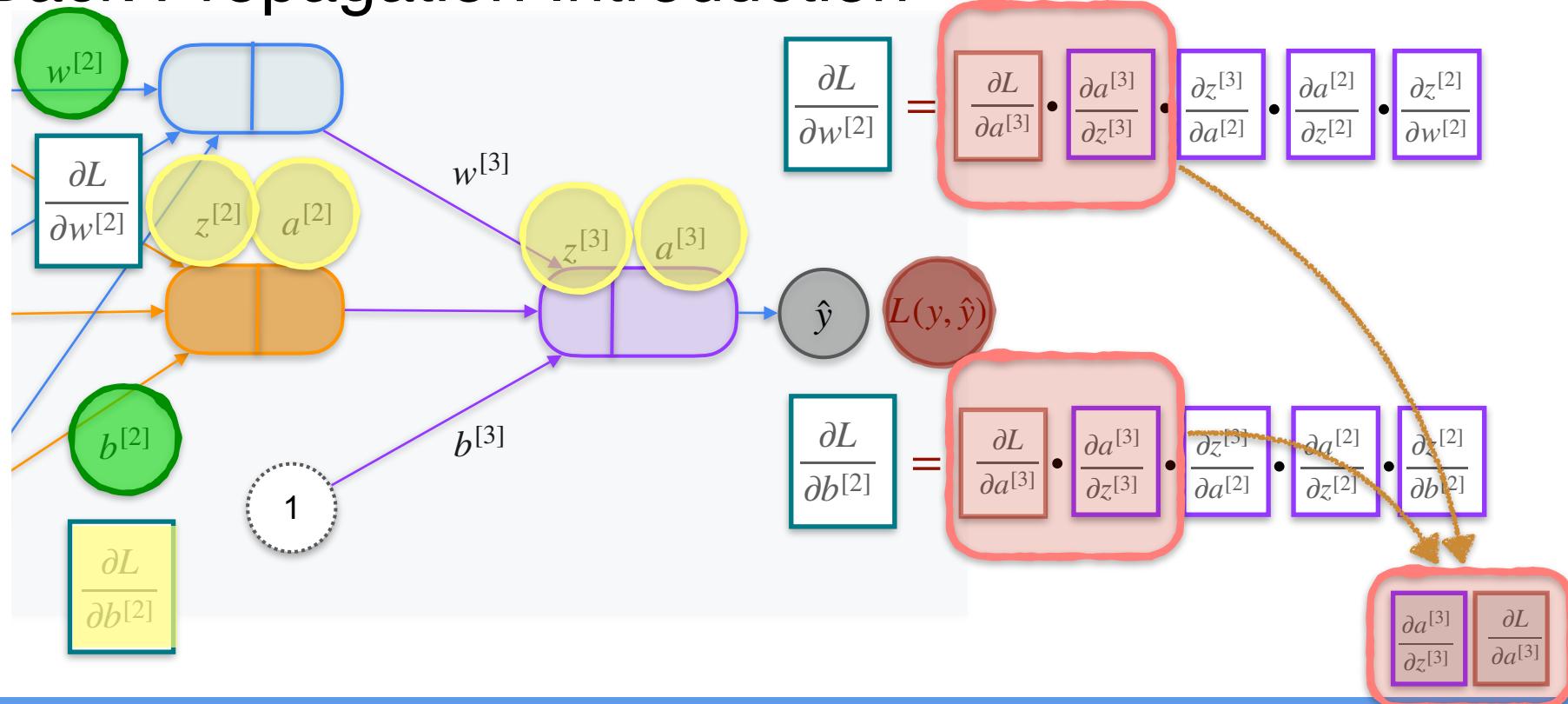
$$\frac{\partial L}{\partial b^{[2]}} = \boxed{\frac{\partial L}{\partial a^{[3]}}} \cdot \boxed{\frac{\partial a^{[3]}}{\partial z^{[3]}}} \cdot \boxed{\frac{\partial z^{[3]}}{\partial a^{[2]}}} \cdot \boxed{\frac{\partial a^{[2]}}{\partial z^{[2]}}} \cdot \boxed{\frac{\partial z^{[2]}}{\partial b^{[2]}}}$$

$$\frac{\partial a^{[3]}}{\partial z^{[3]}} \quad \frac{\partial L}{\partial a^{[3]}}$$

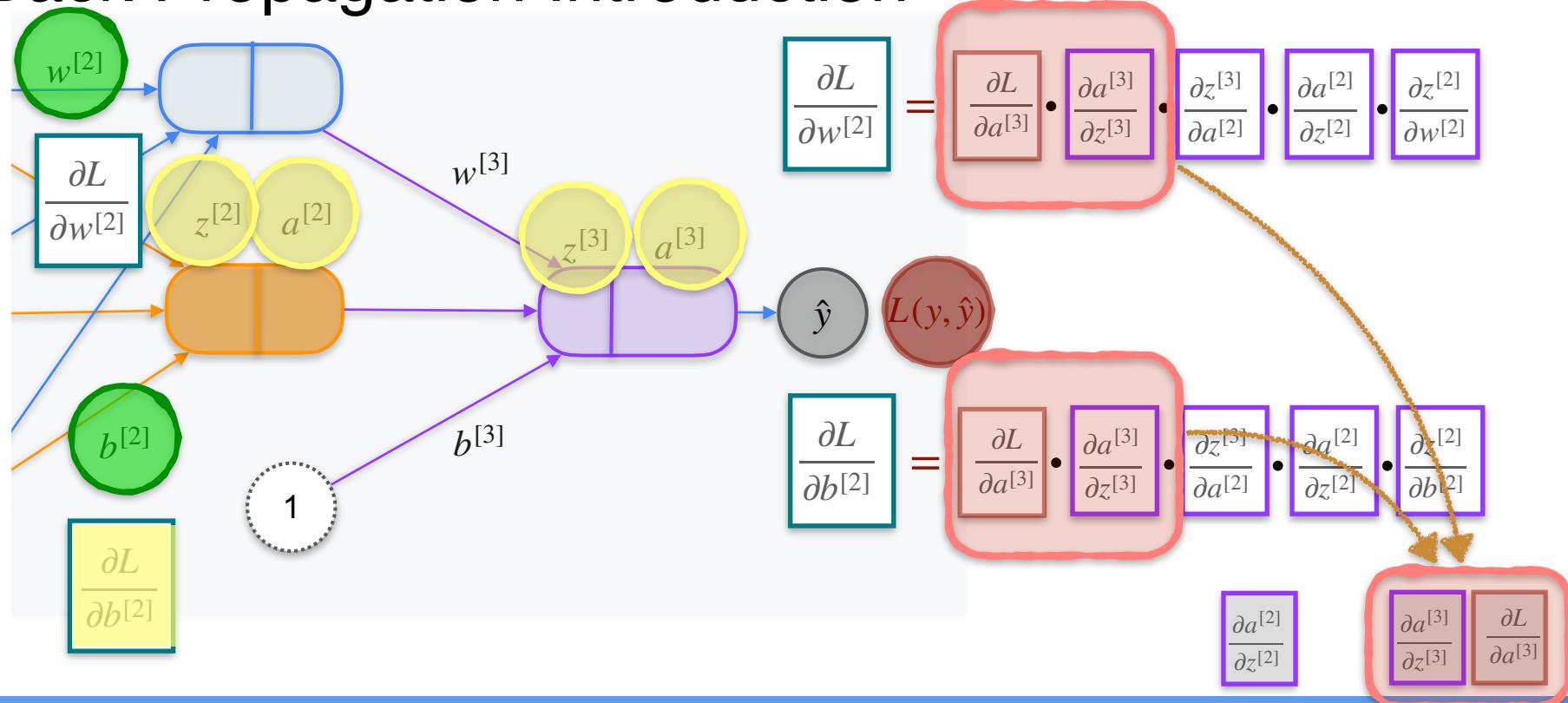
# Back Propagation Introduction



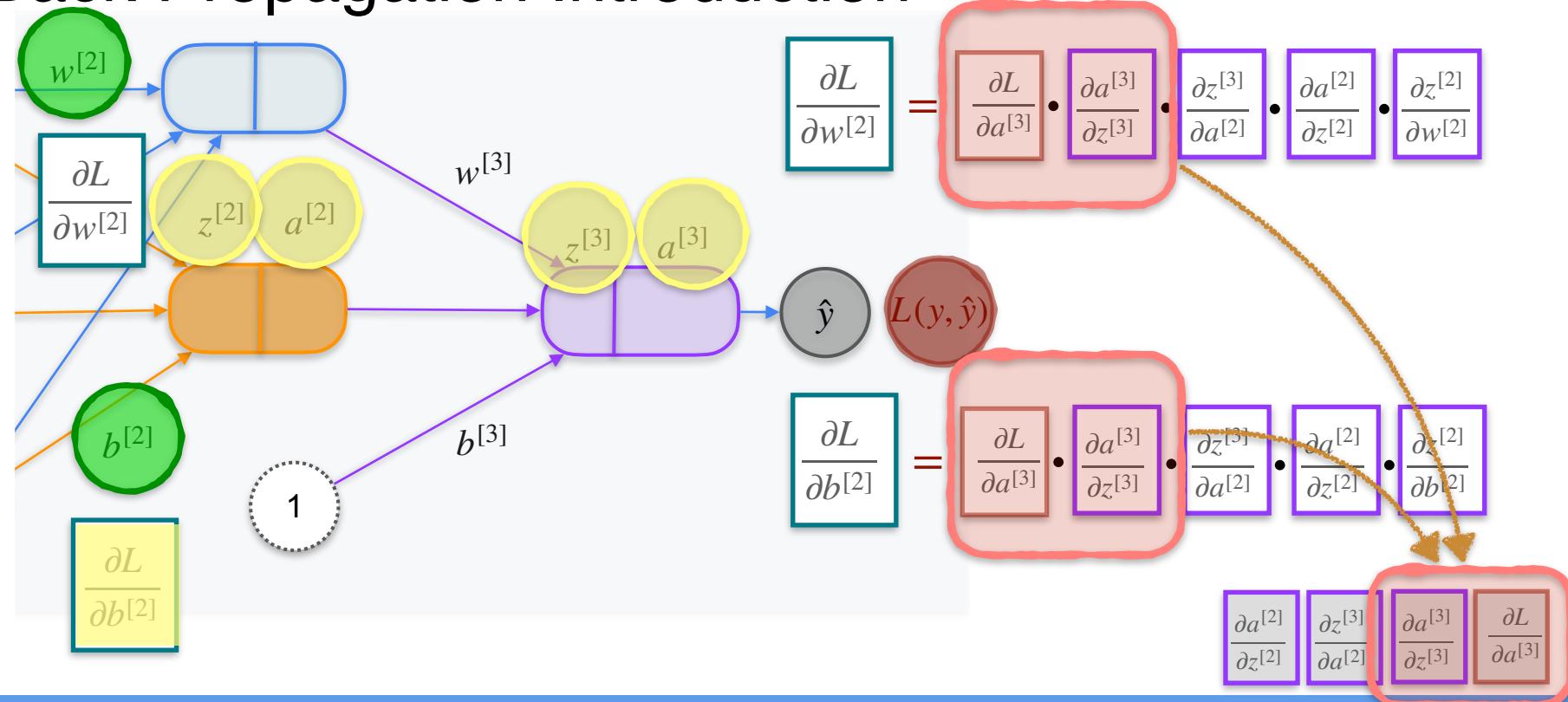
# Back Propagation Introduction



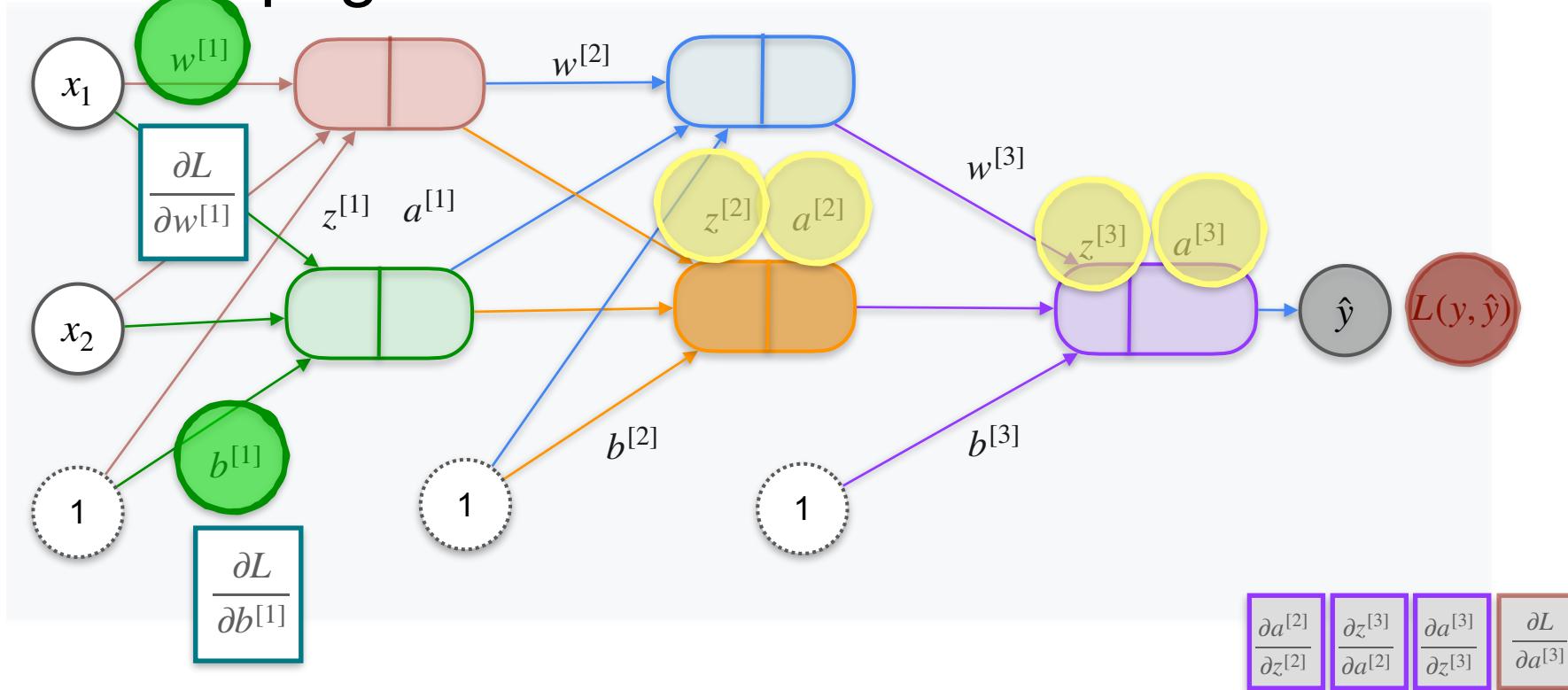
# Back Propagation Introduction



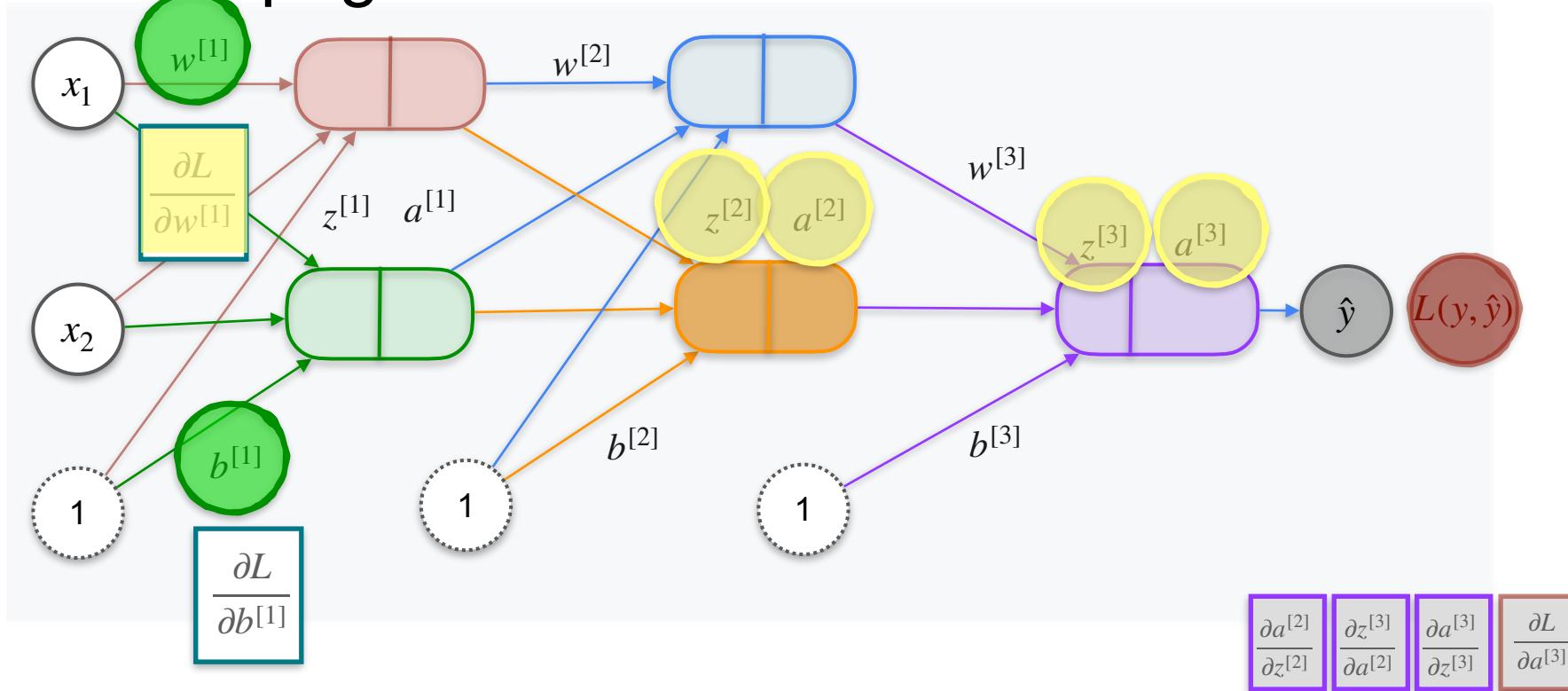
# Back Propagation Introduction



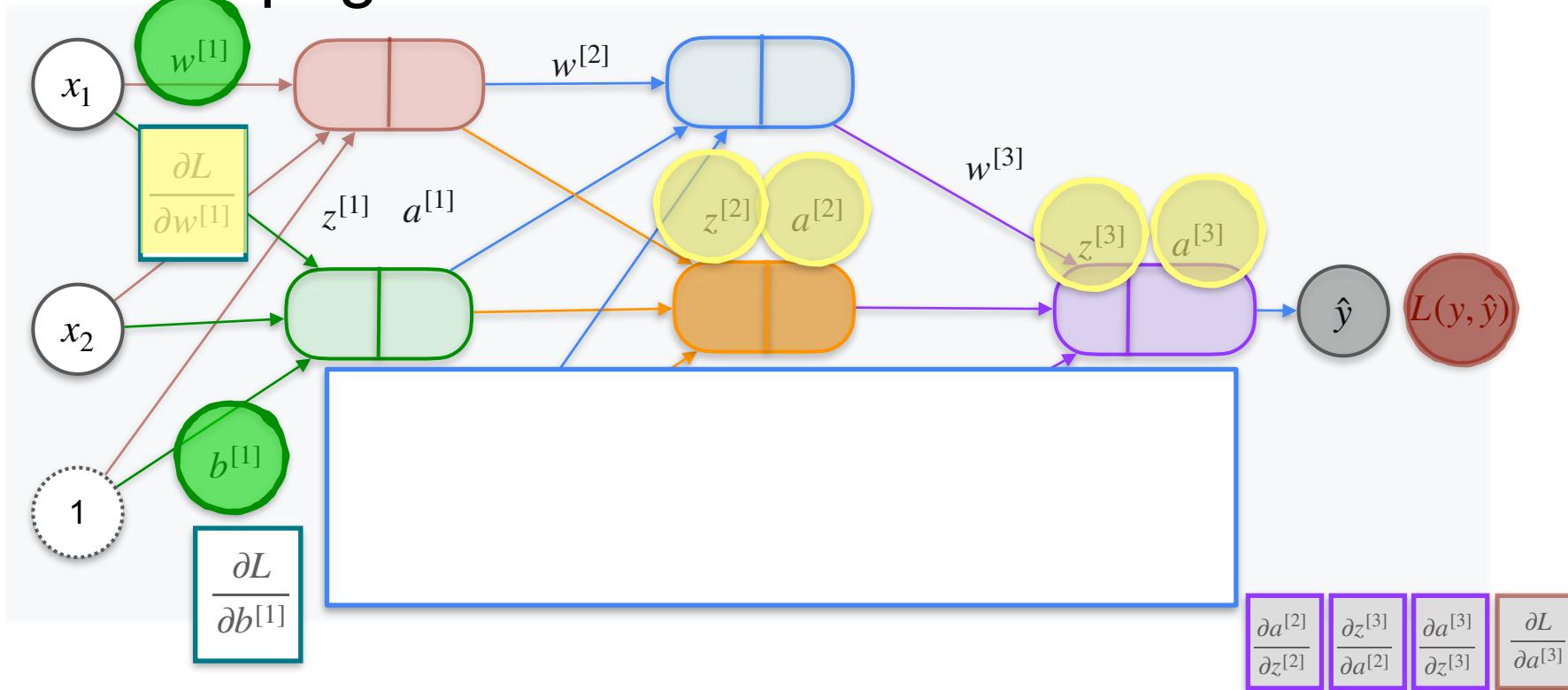
# Back Propagation Introduction



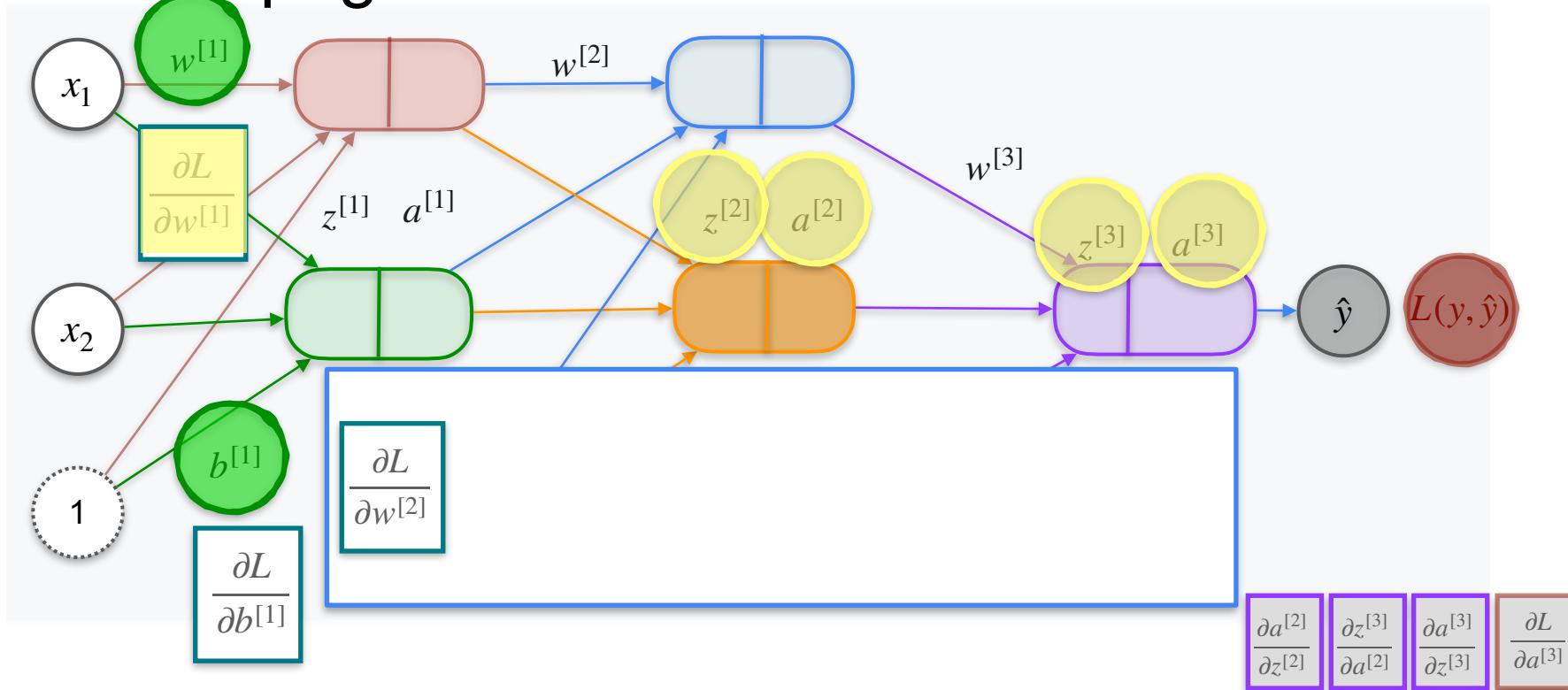
# Back Propagation Introduction



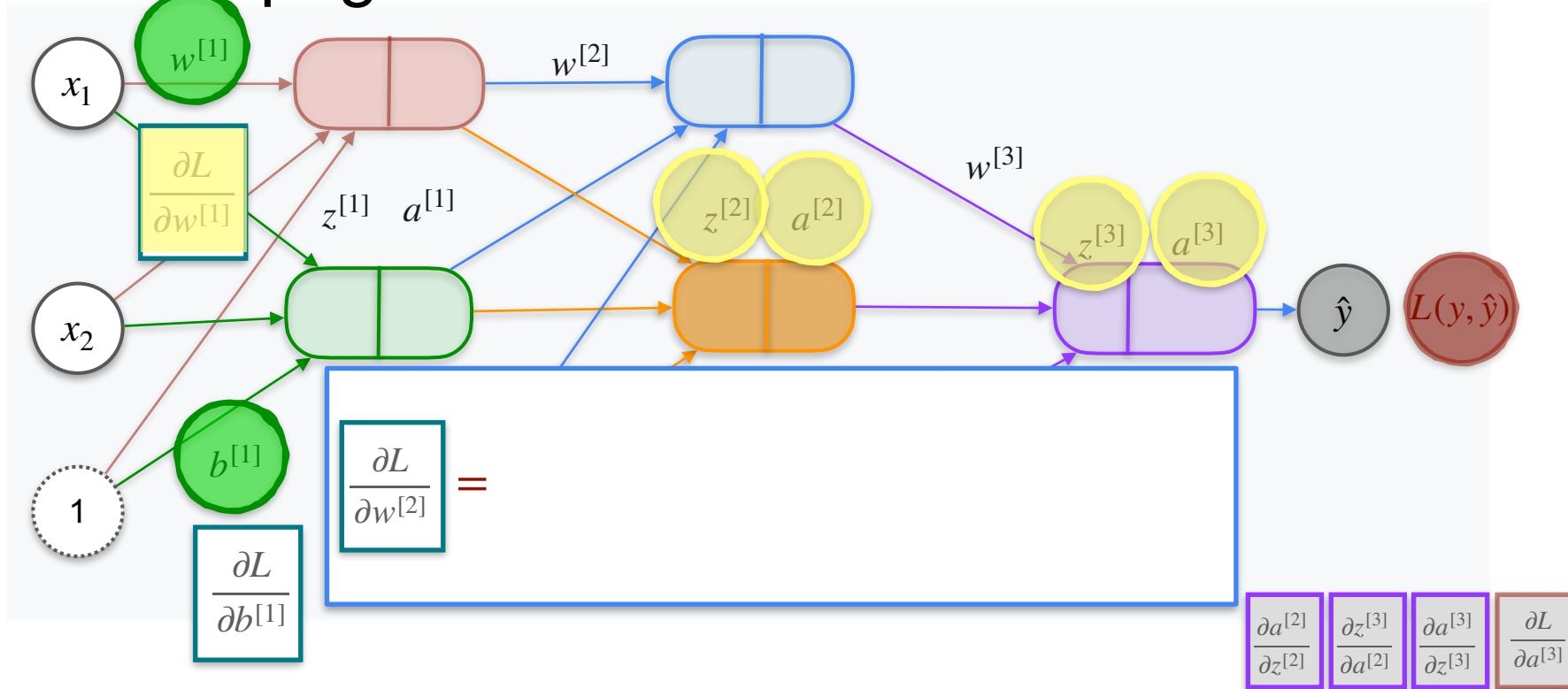
# Back Propagation Introduction



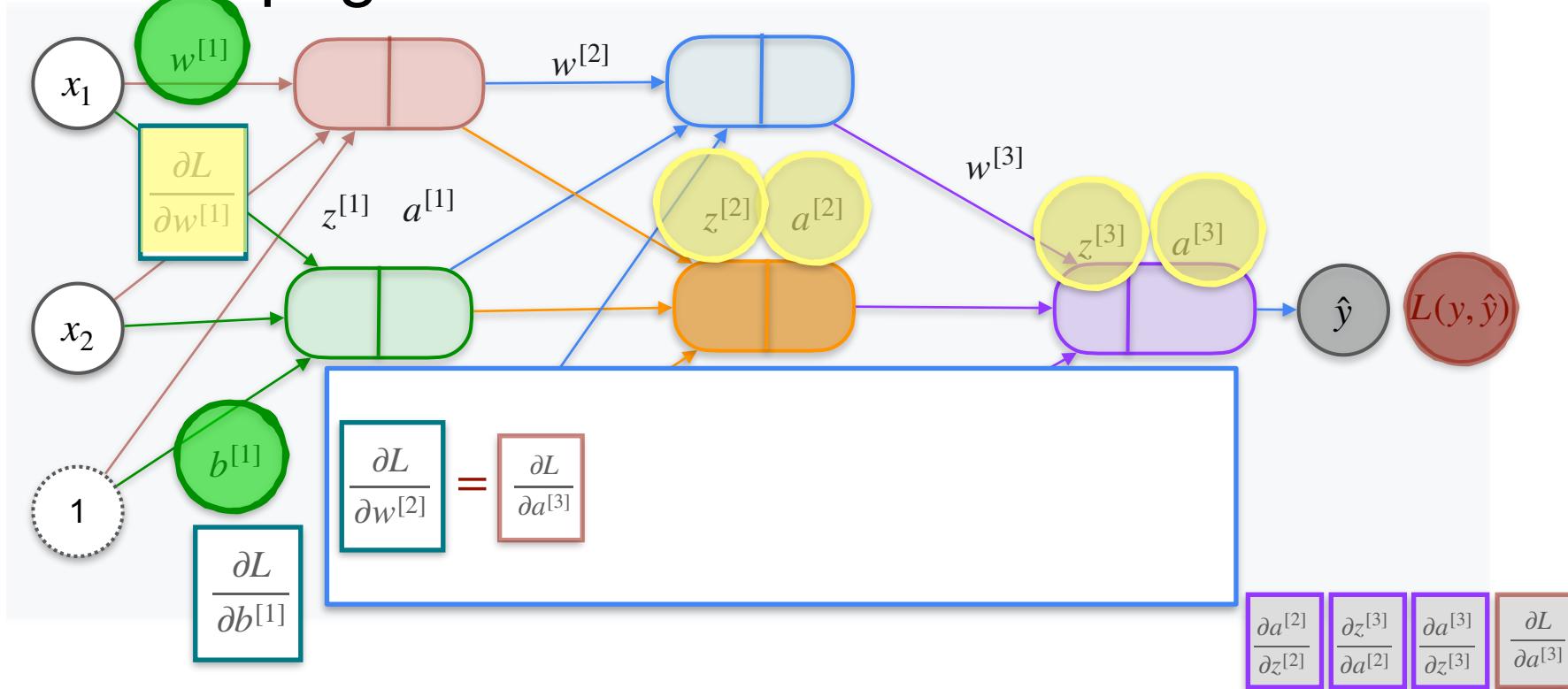
# Back Propagation Introduction



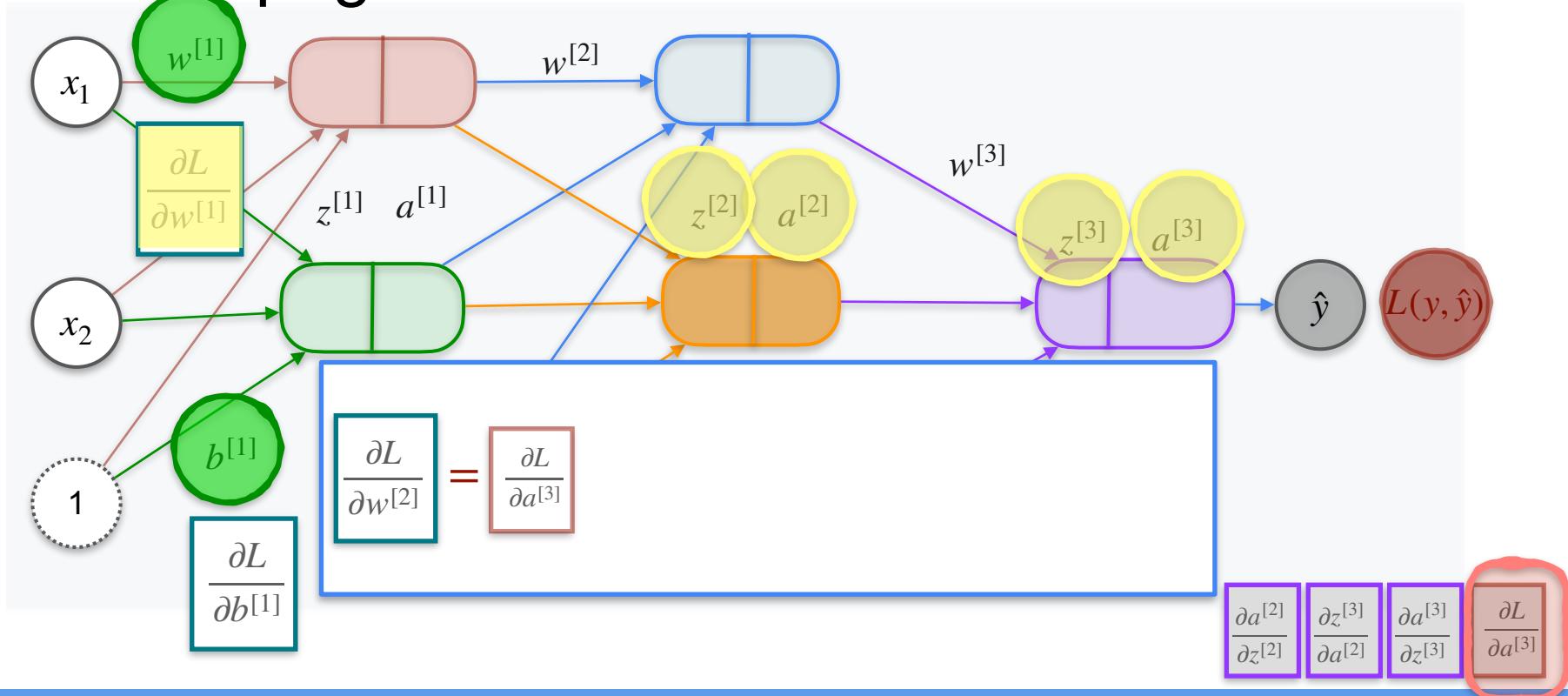
# Back Propagation Introduction



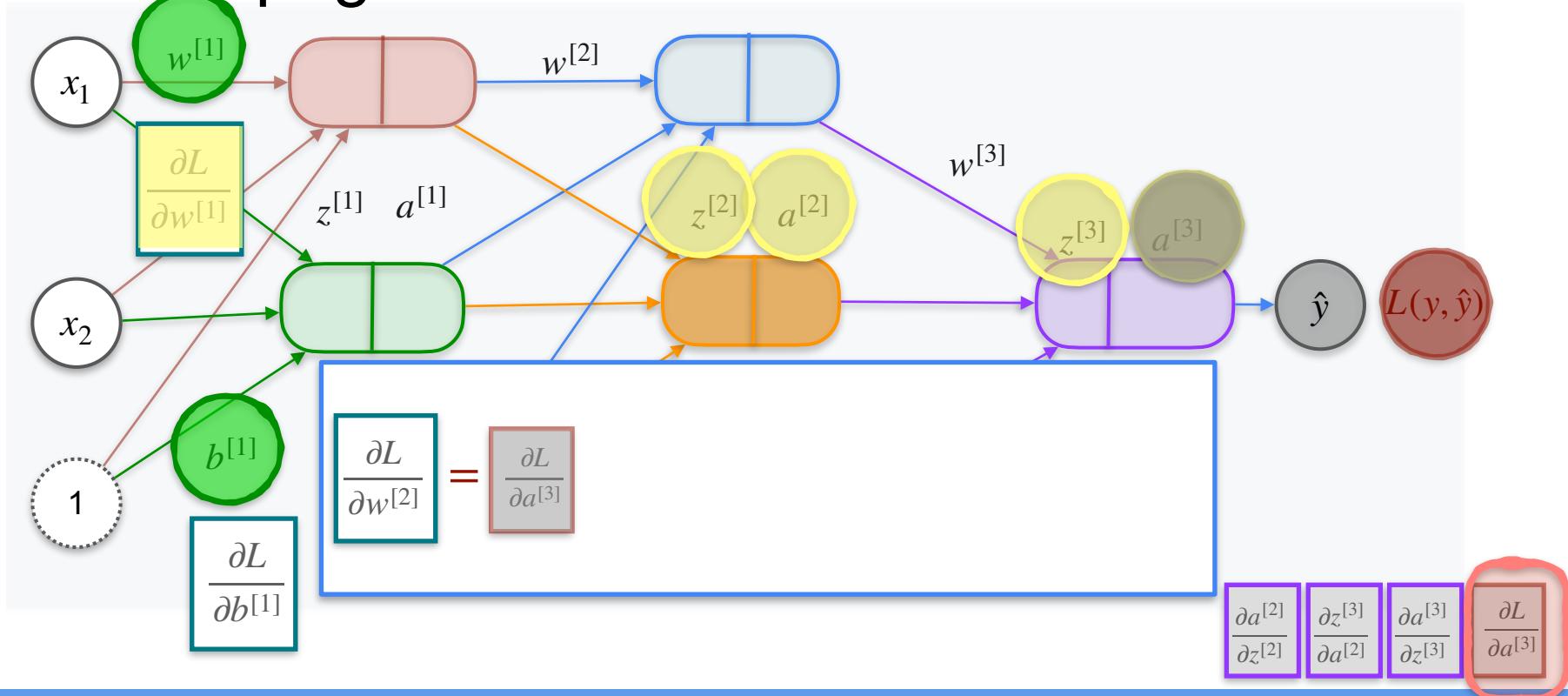
# Back Propagation Introduction



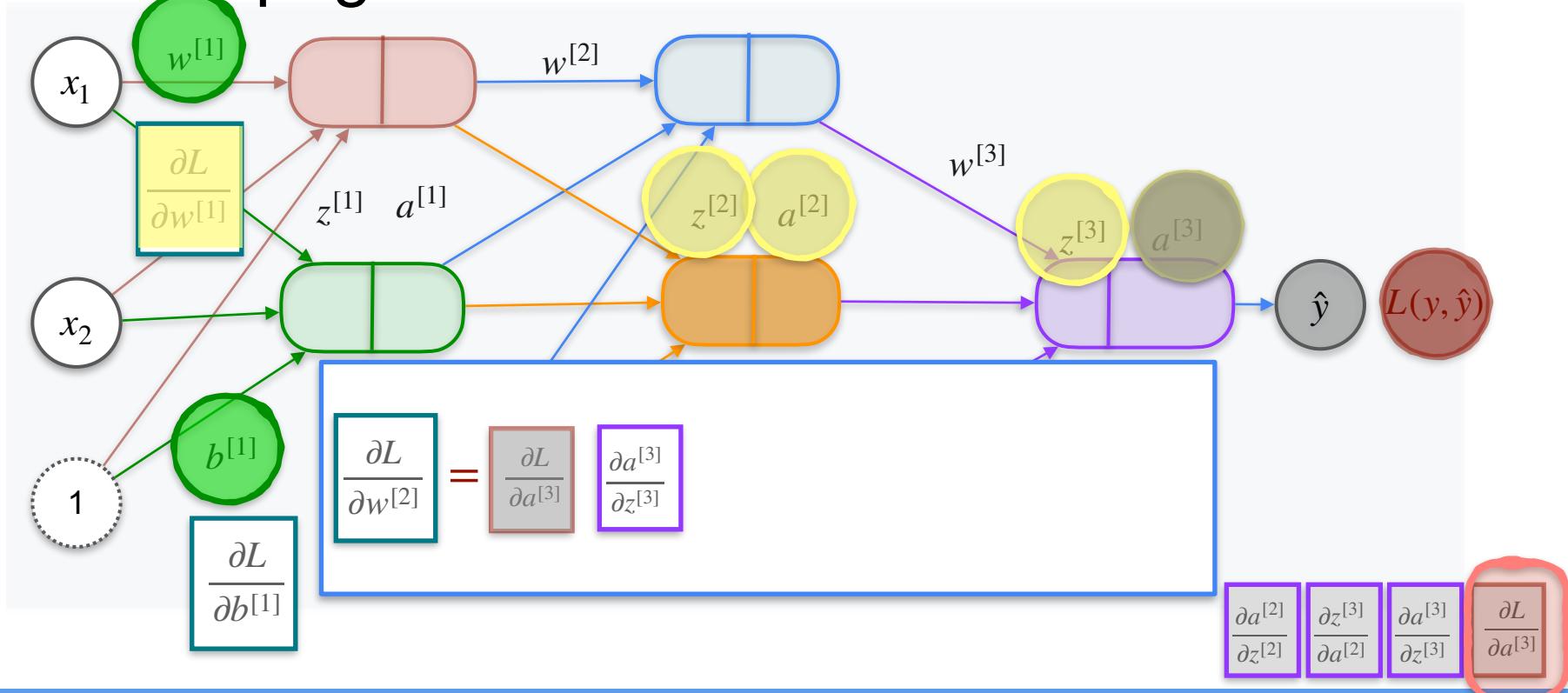
# Back Propagation Introduction



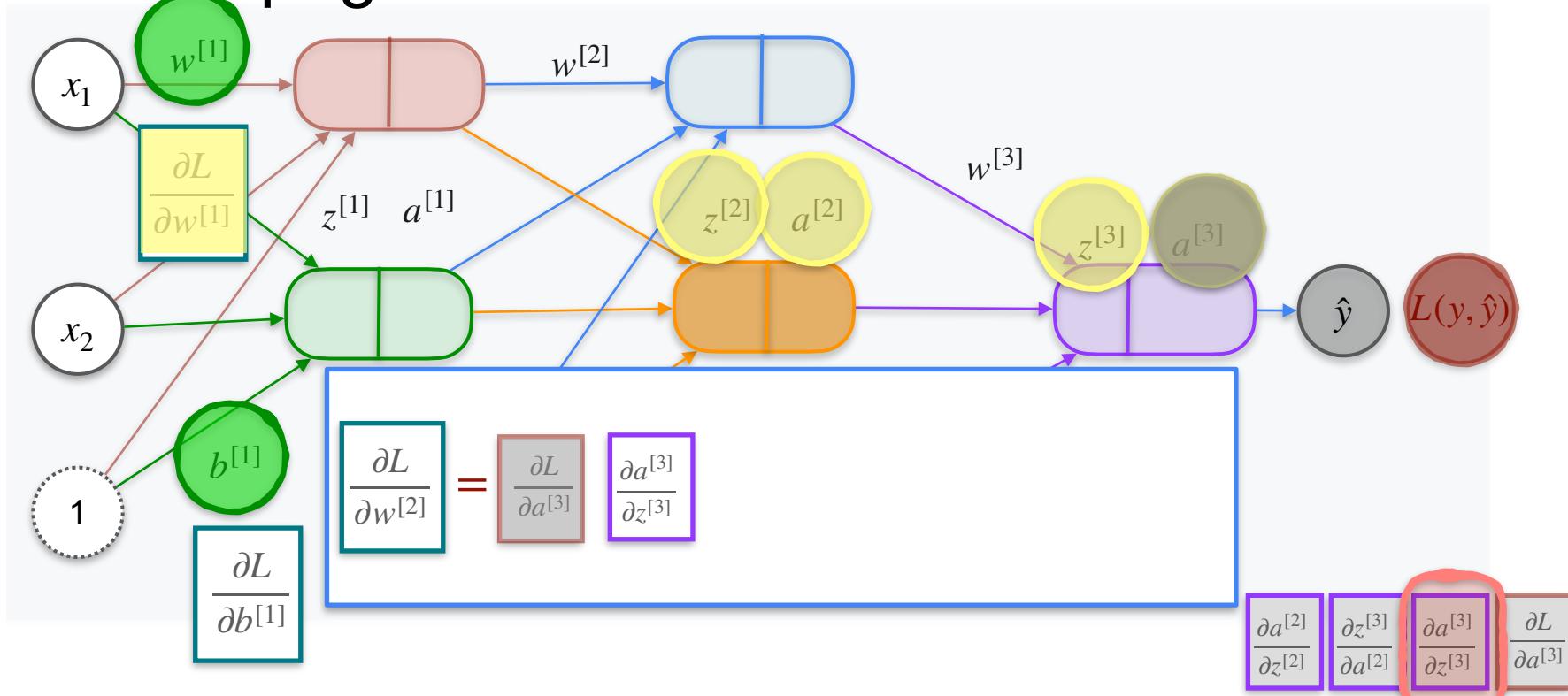
# Back Propagation Introduction



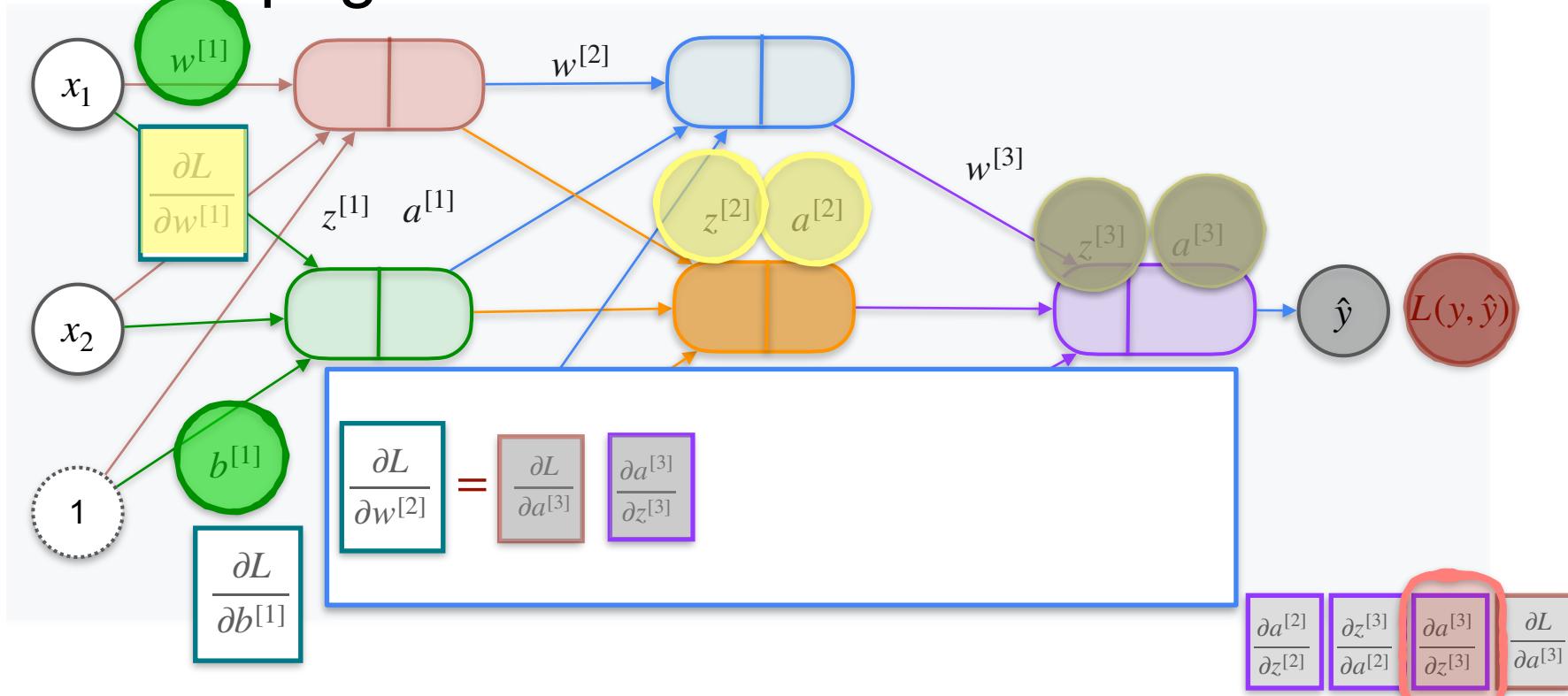
# Back Propagation Introduction



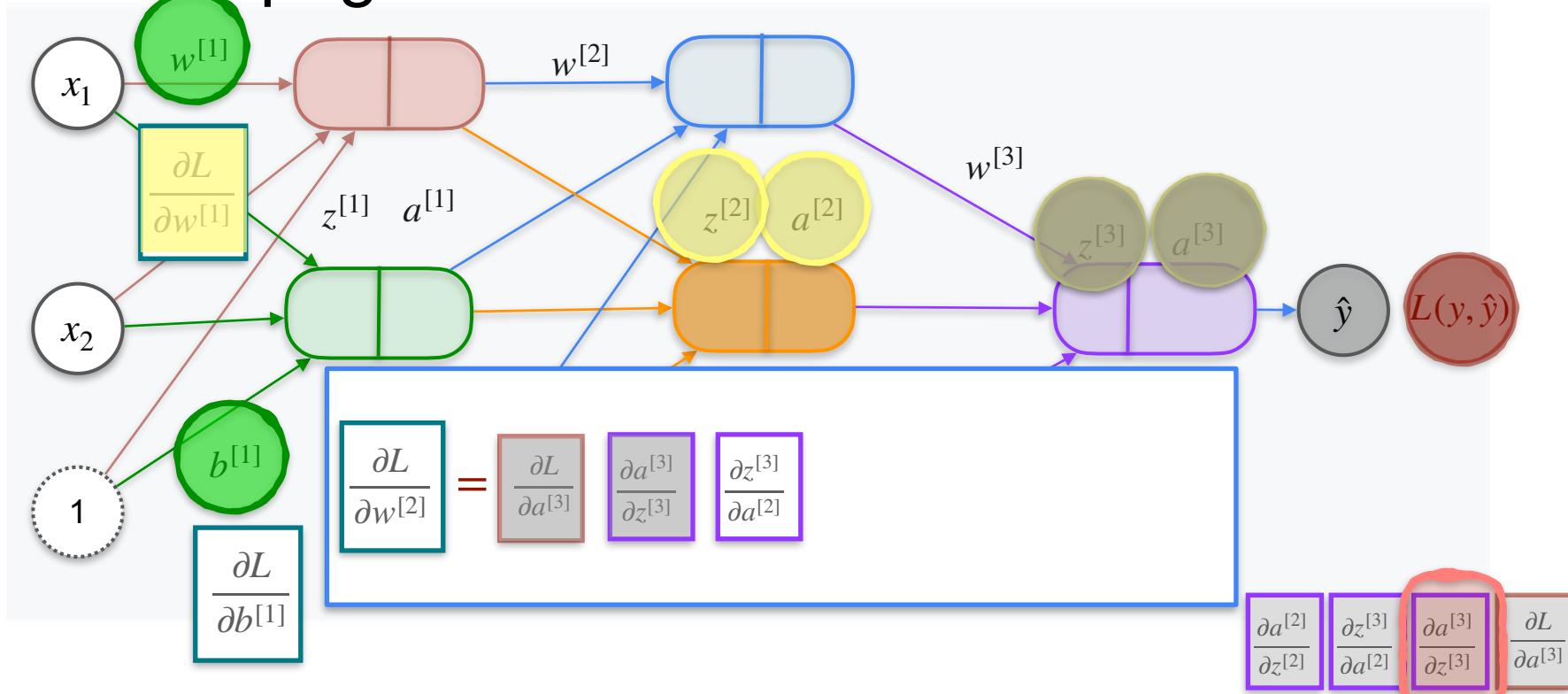
# Back Propagation Introduction



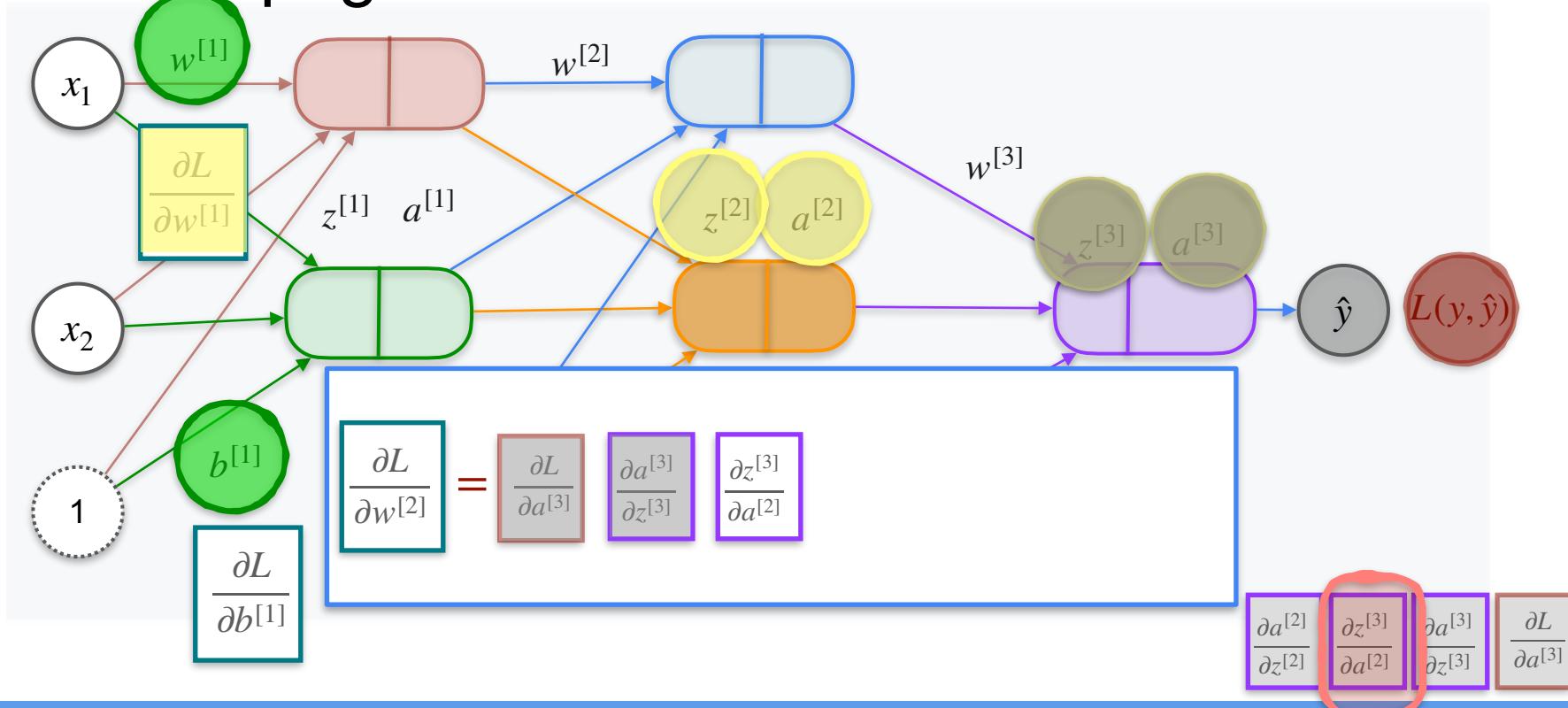
# Back Propagation Introduction



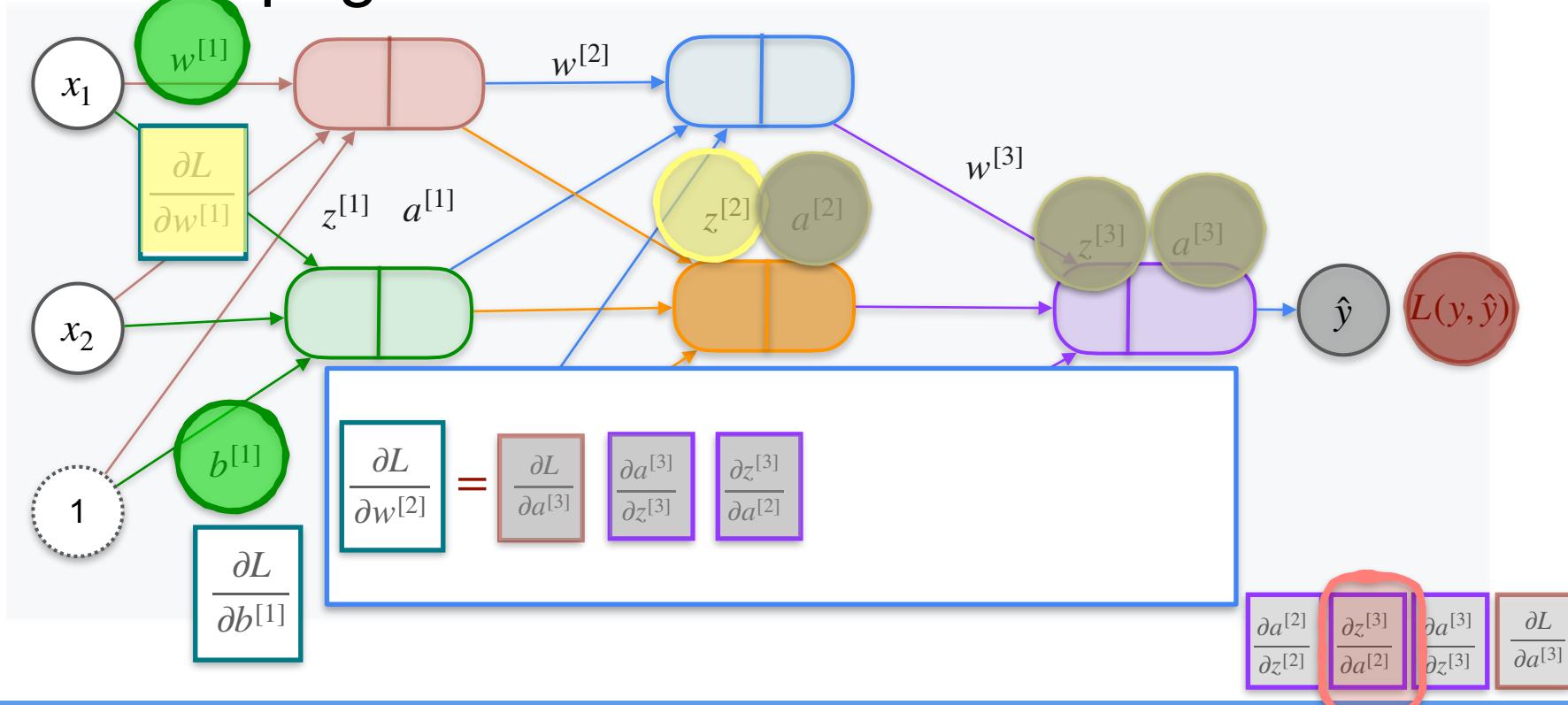
# Back Propagation Introduction



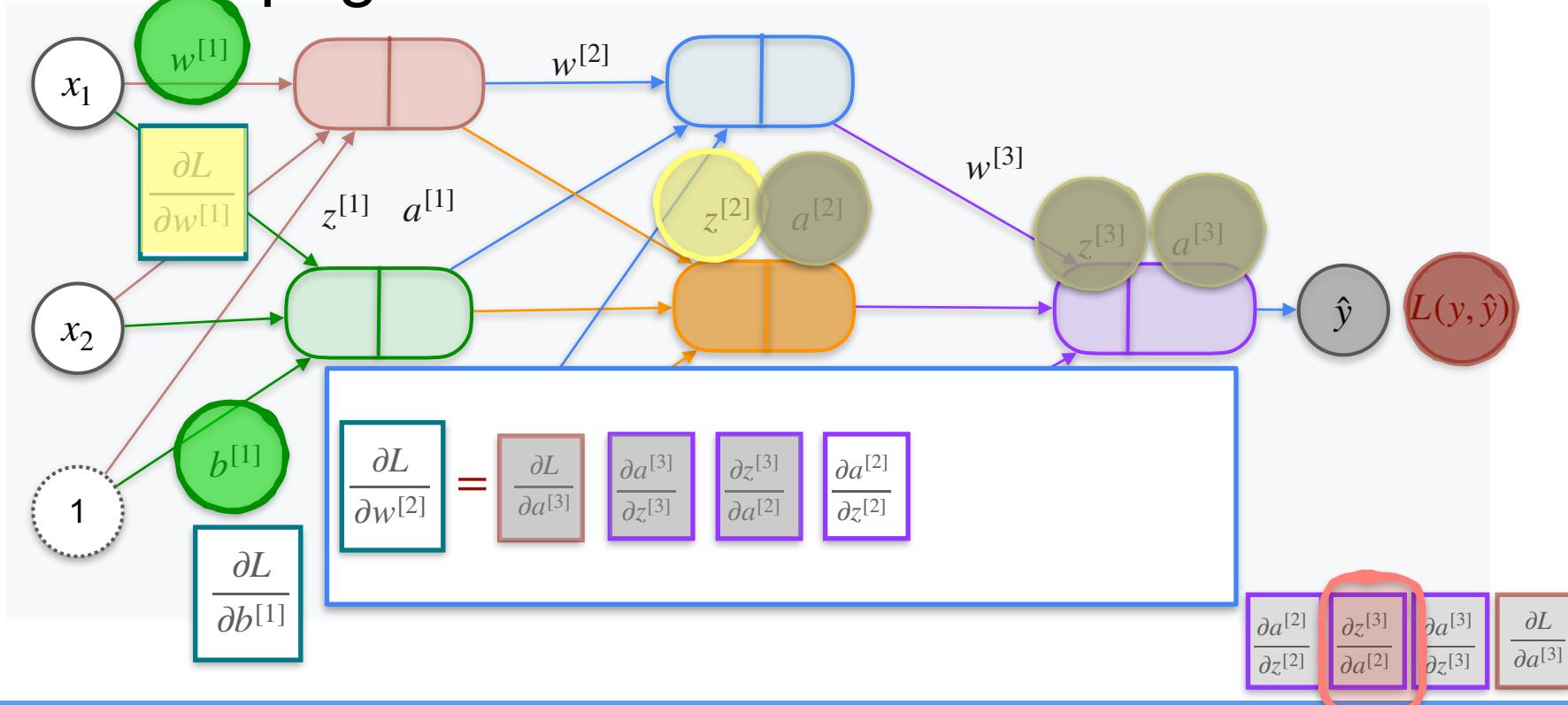
# Back Propagation Introduction



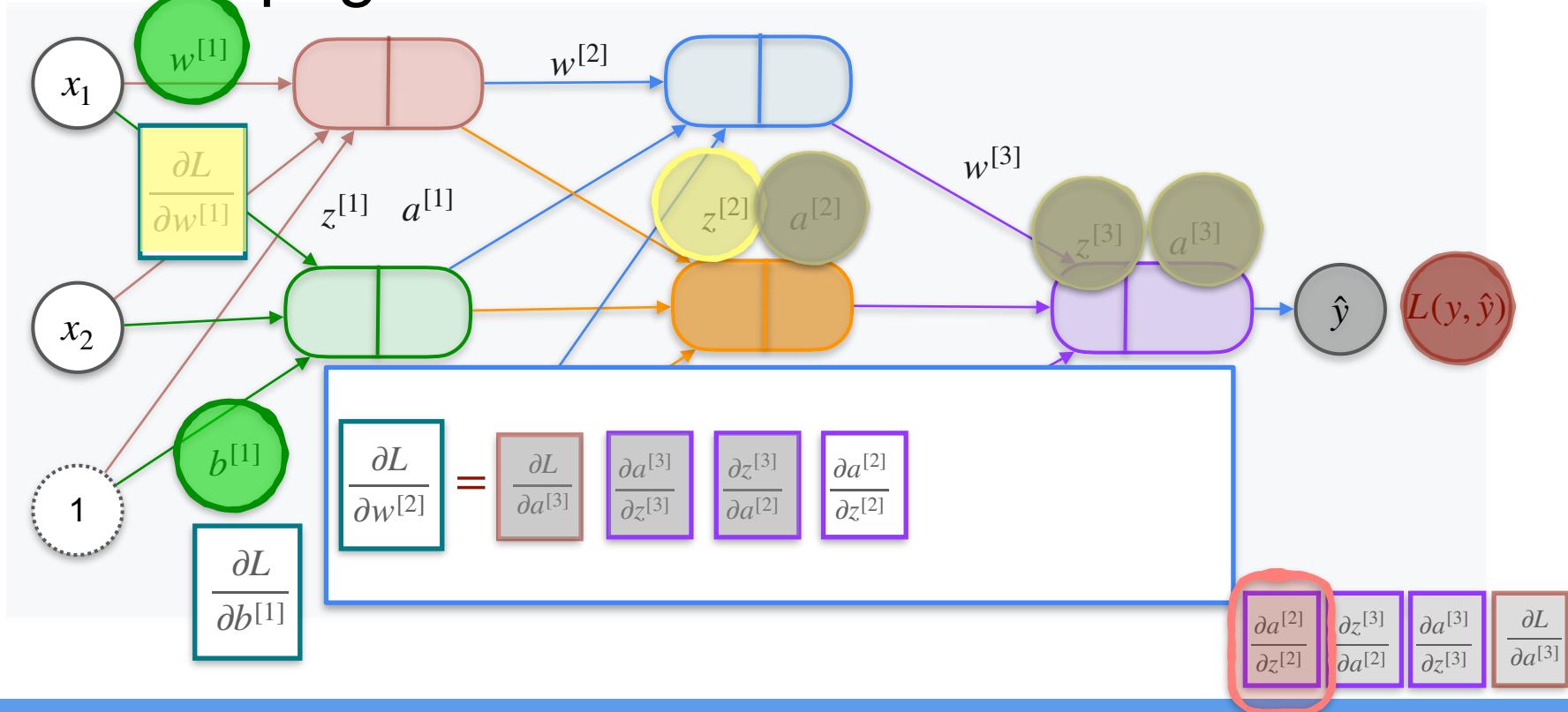
# Back Propagation Introduction



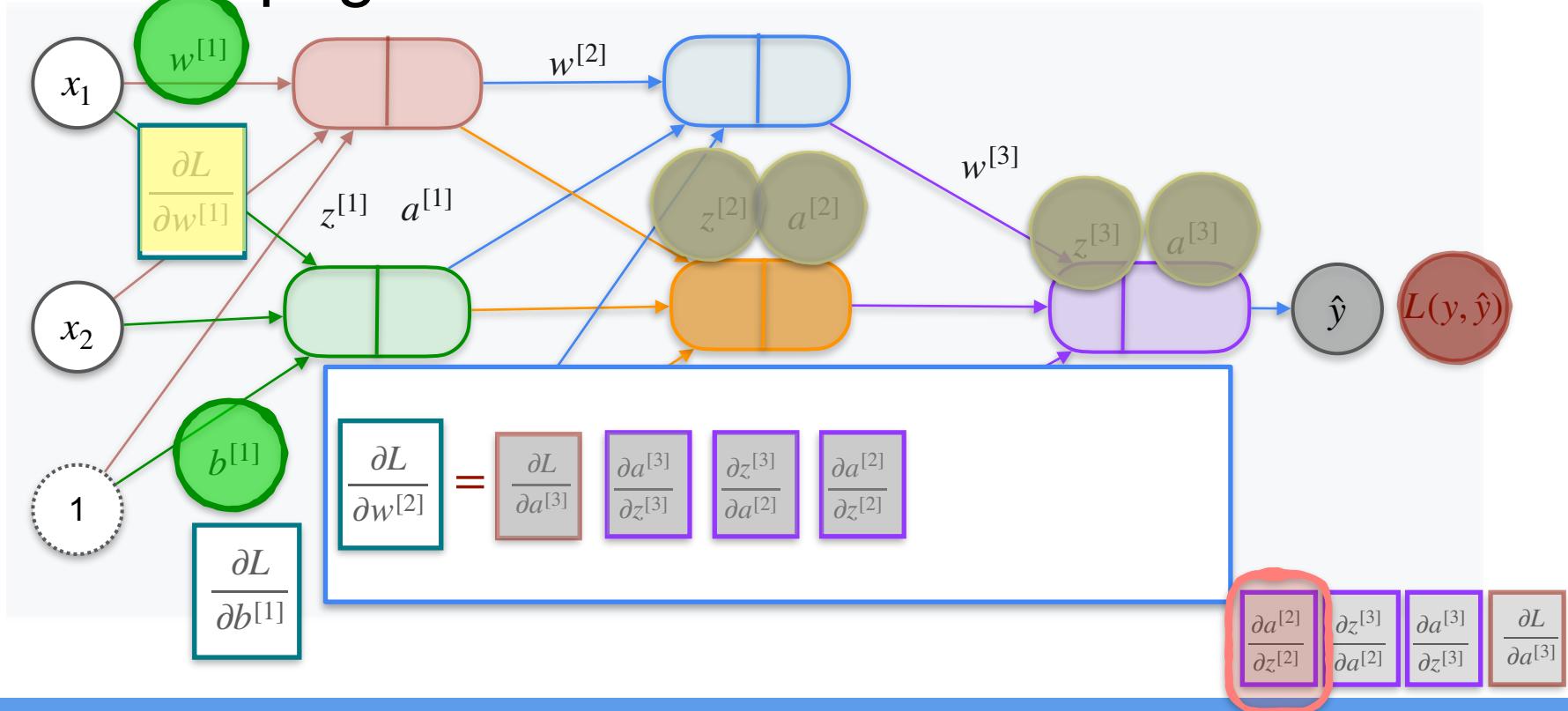
# Back Propagation Introduction



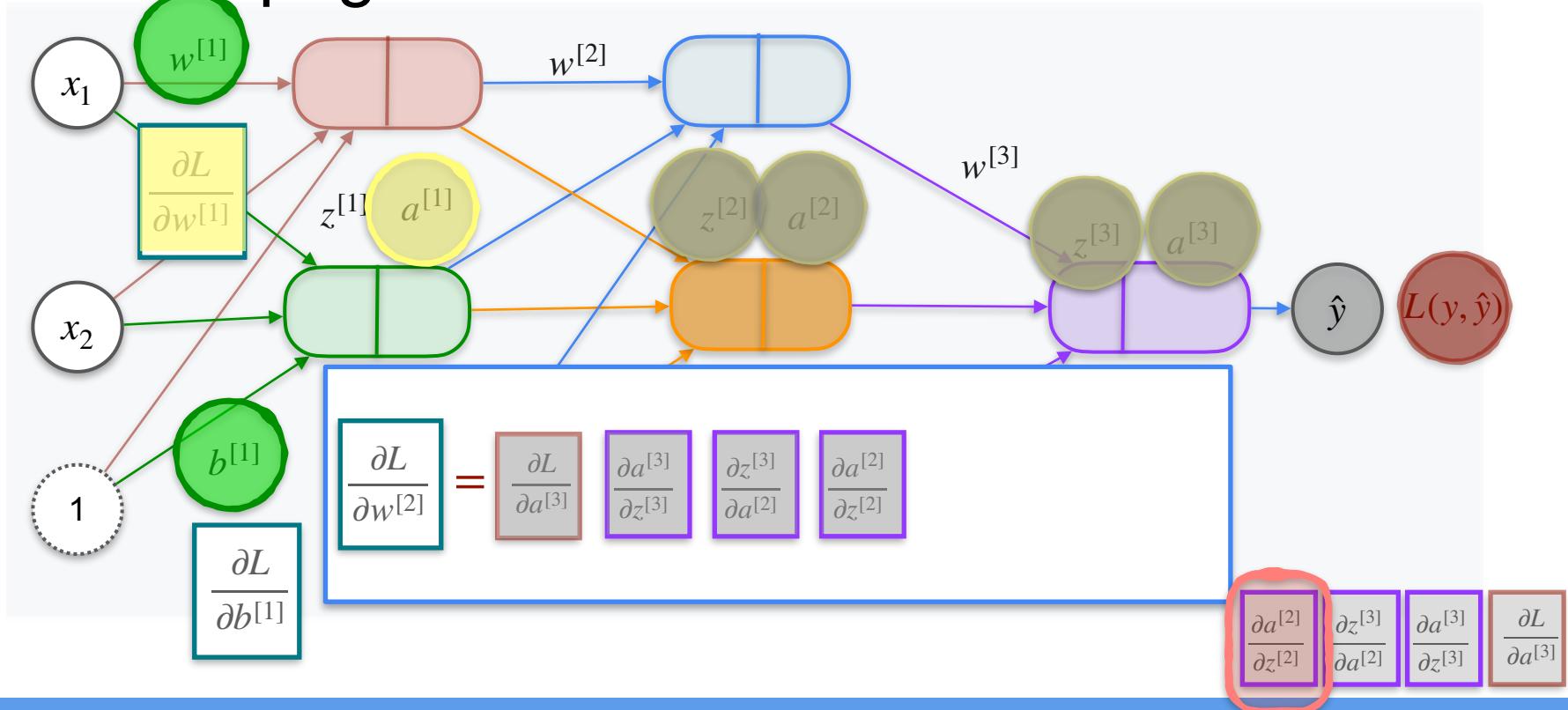
# Back Propagation Introduction



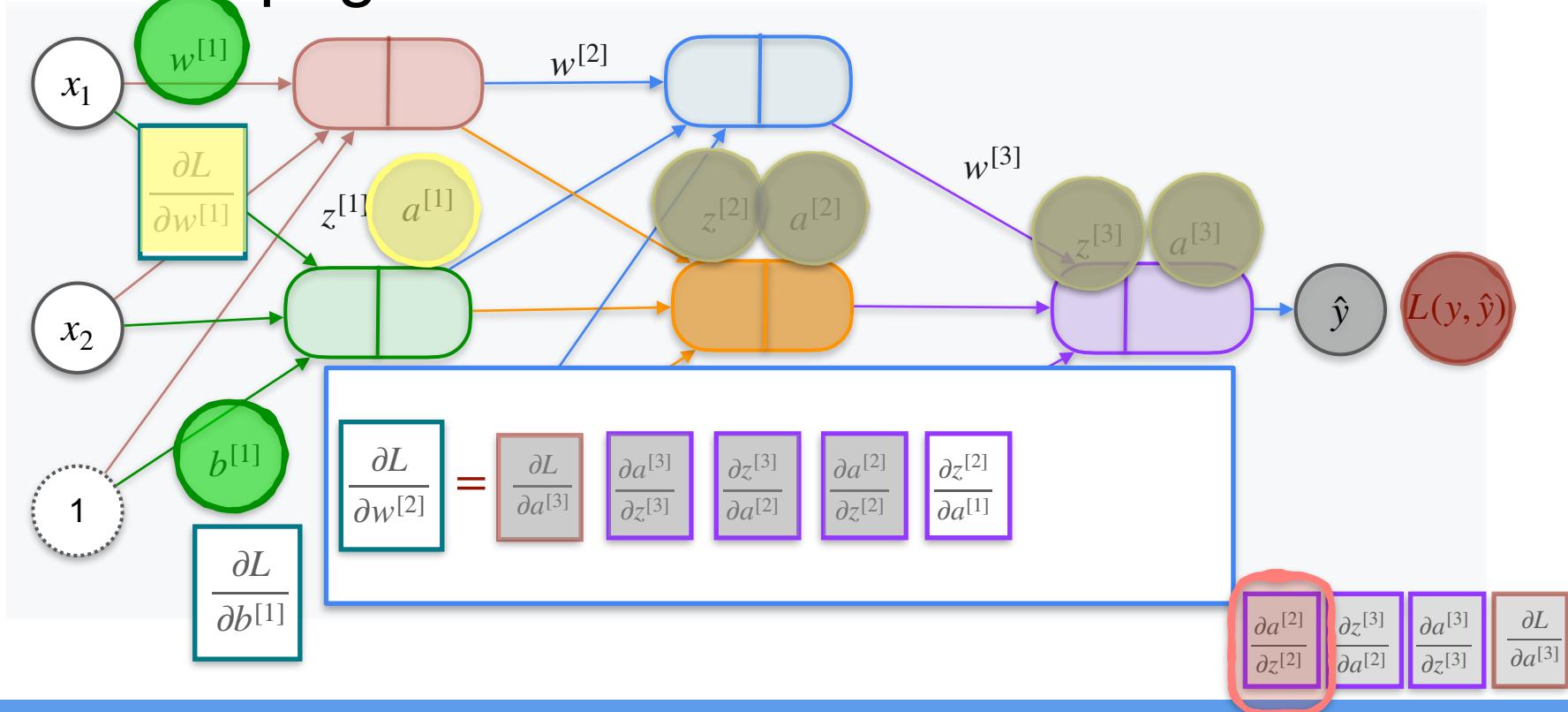
# Back Propagation Introduction



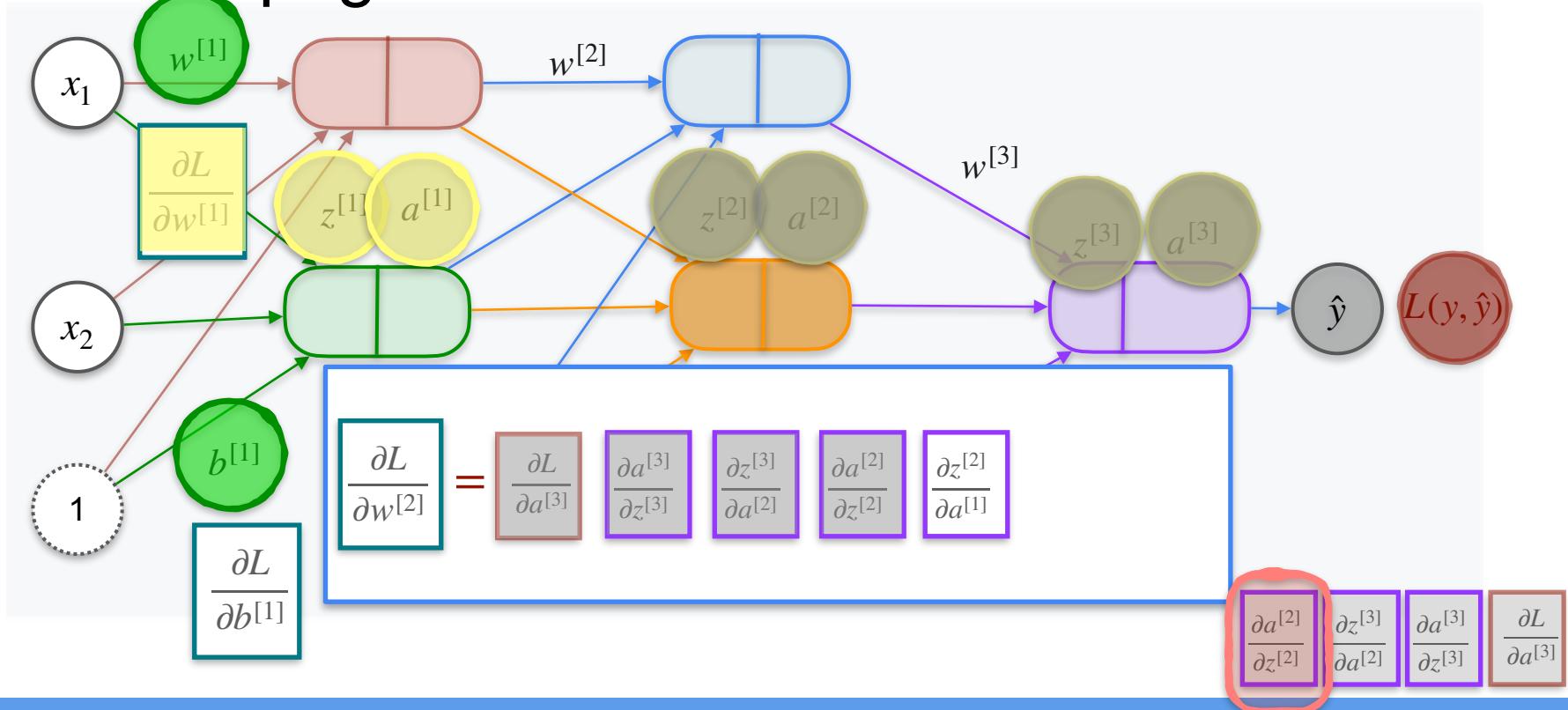
# Back Propagation Introduction



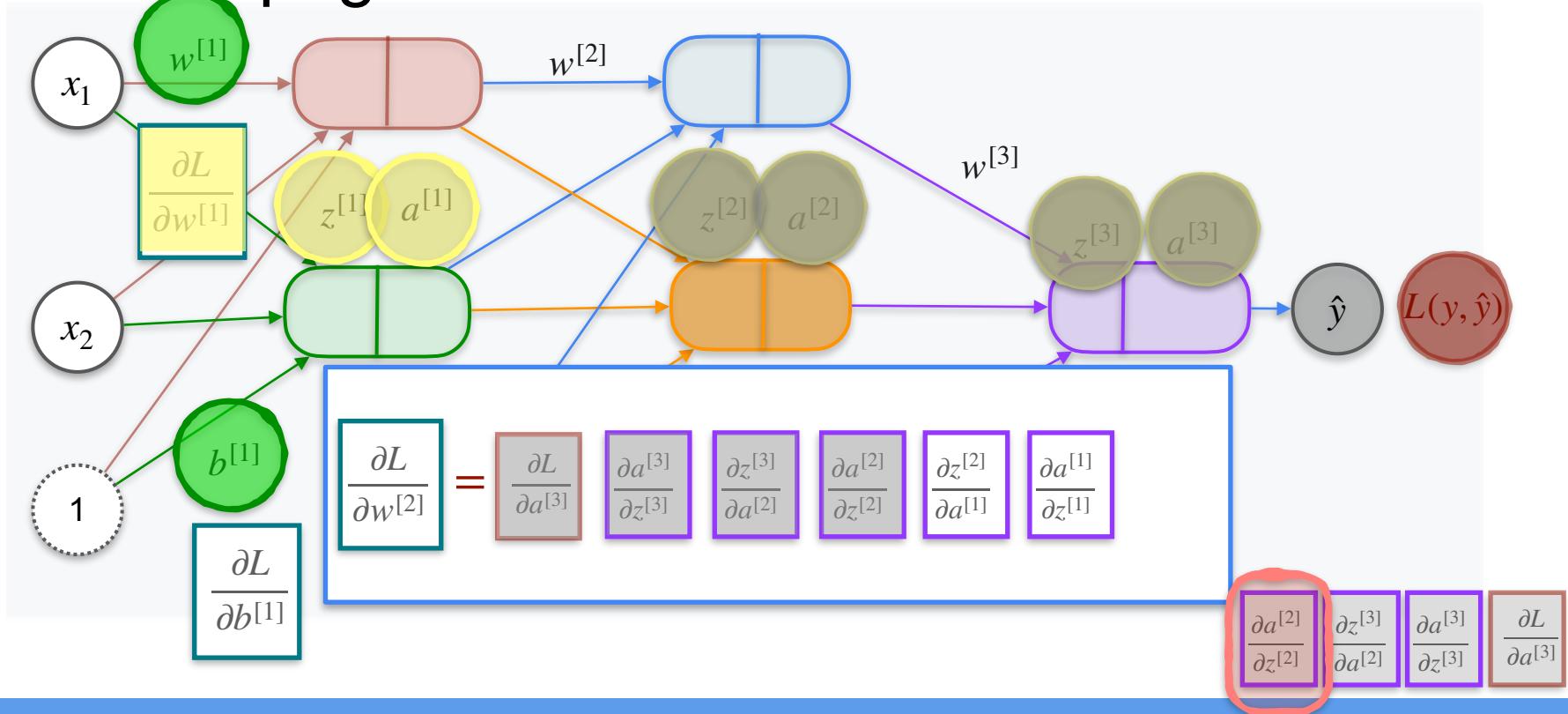
# Back Propagation Introduction



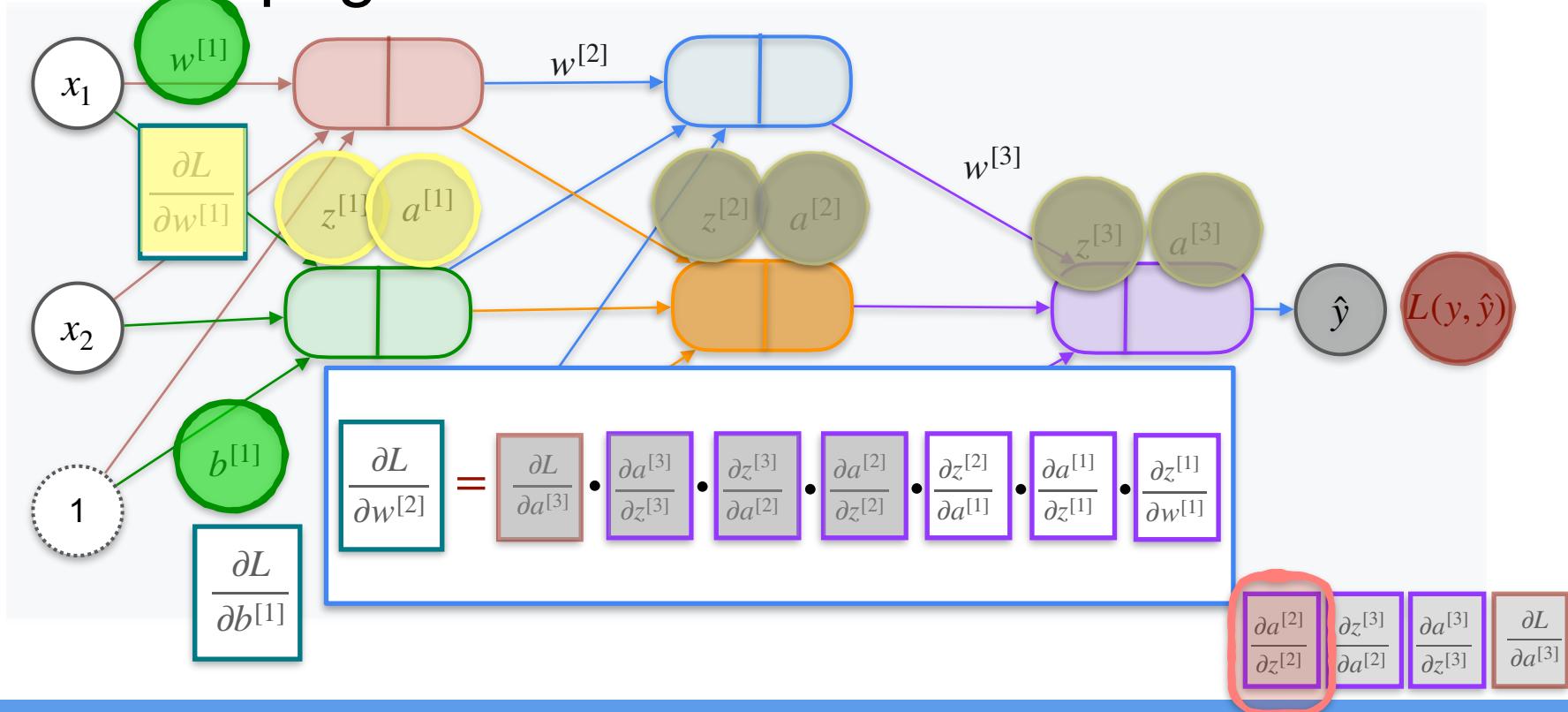
# Back Propagation Introduction



# Back Propagation Introduction



# Back Propagation Introduction





DeepLearning.AI

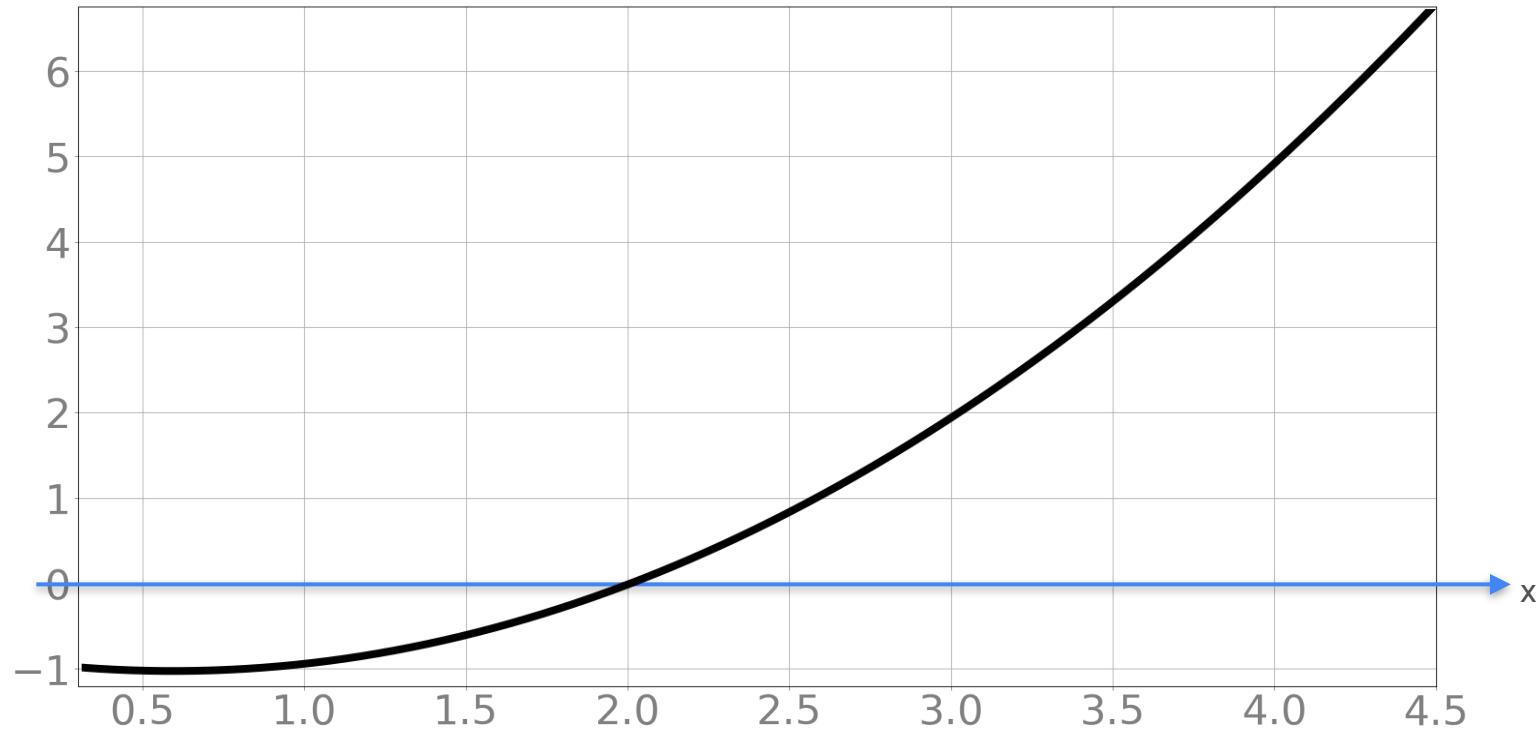
# Optimization in Neural Networks and Newton's Method

---

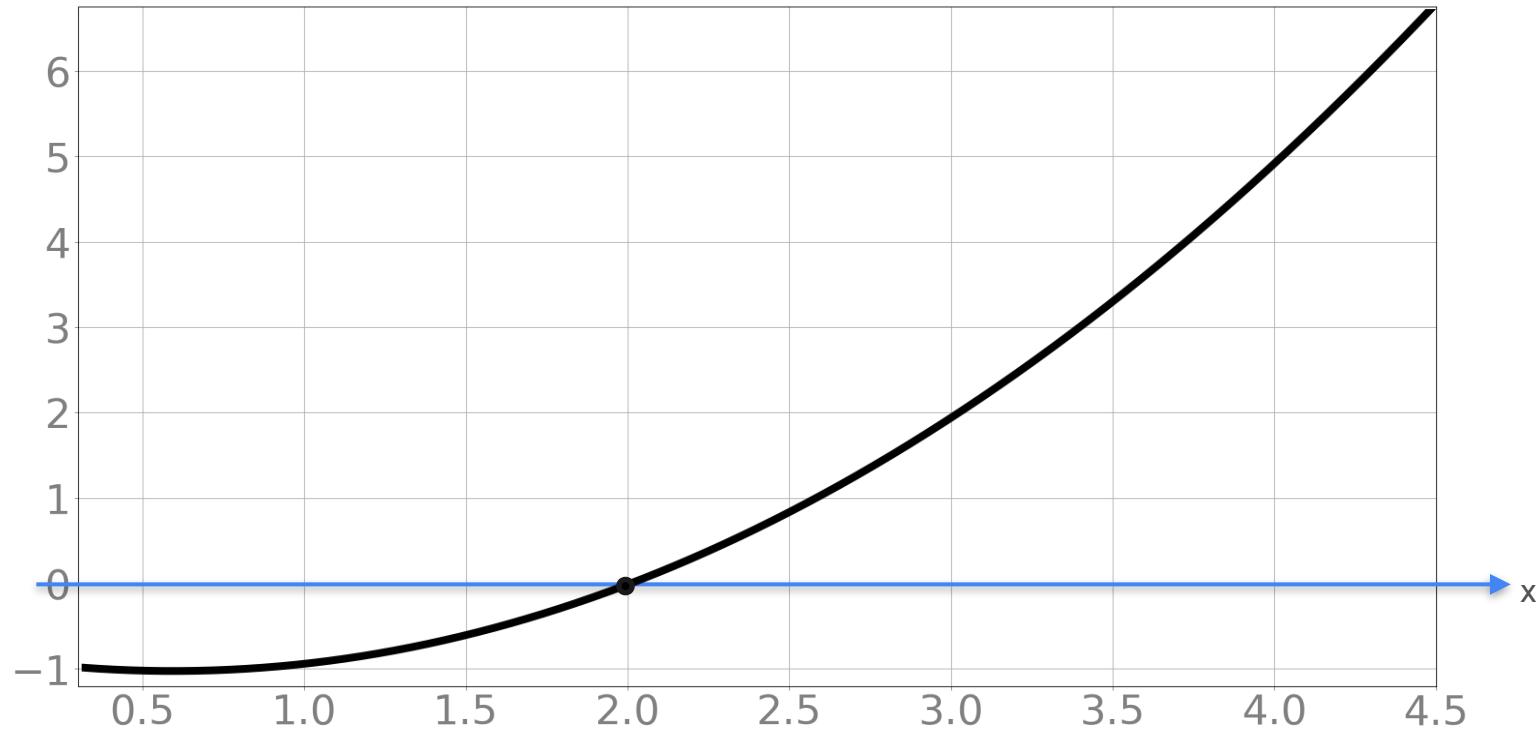
## Newton's method

# Newton's Method

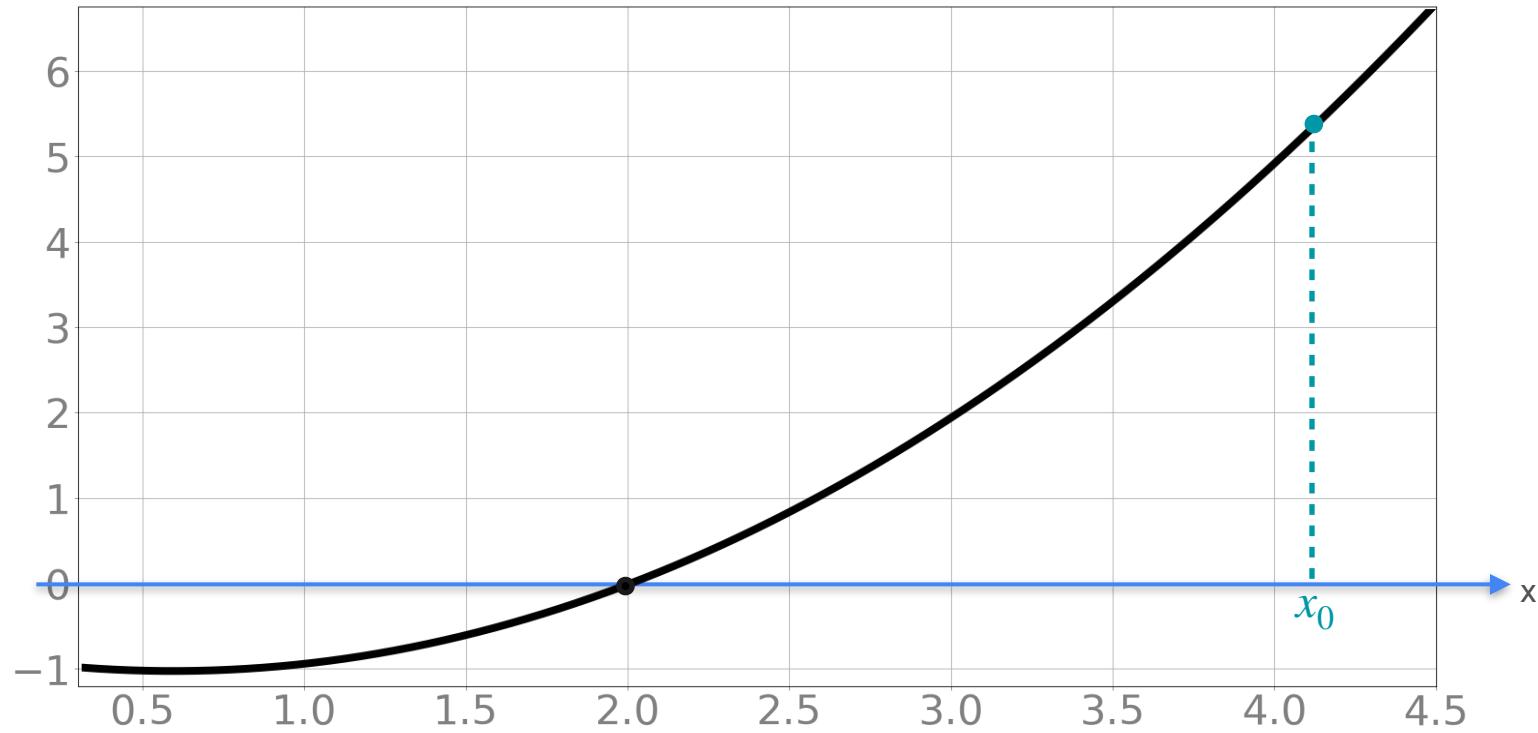
# Newton's Method



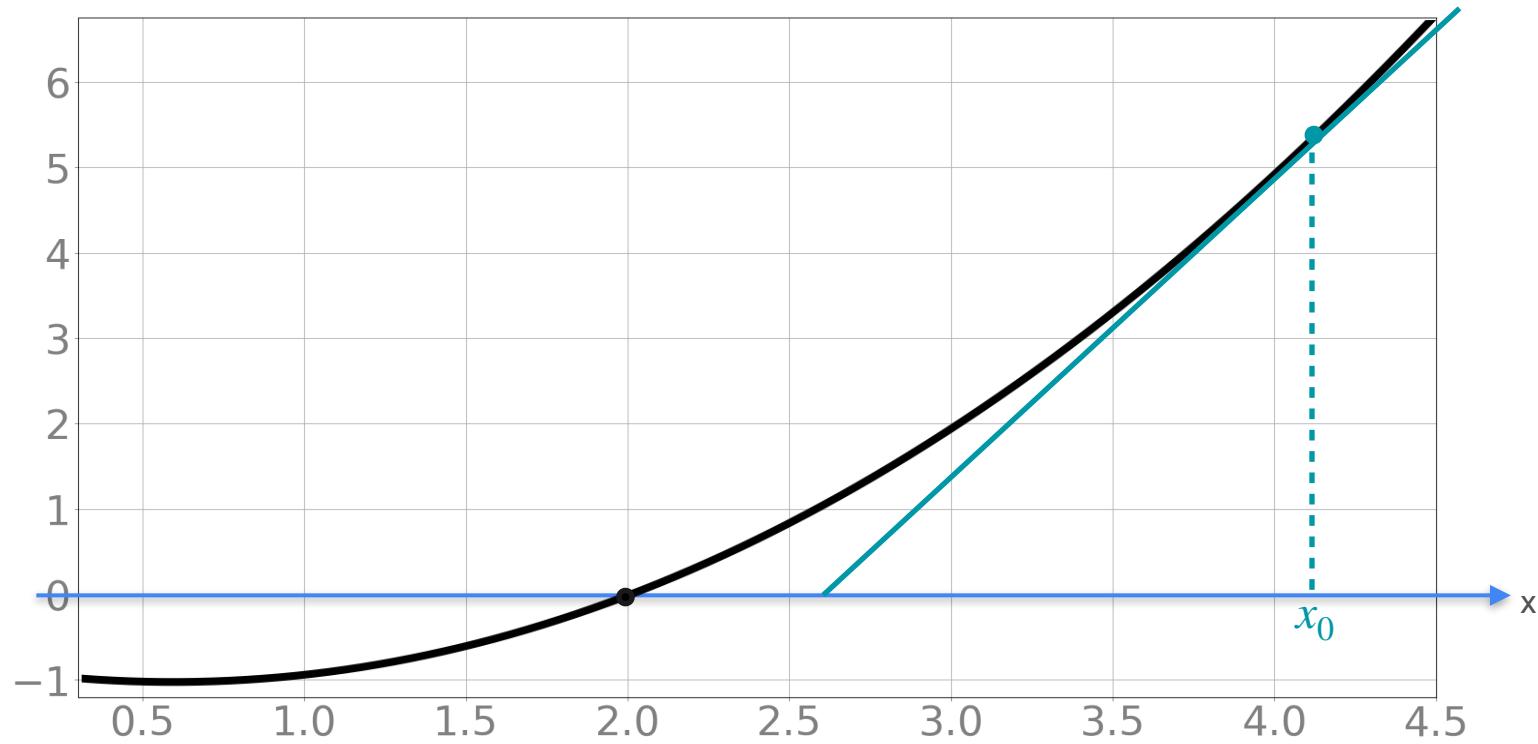
# Newton's Method



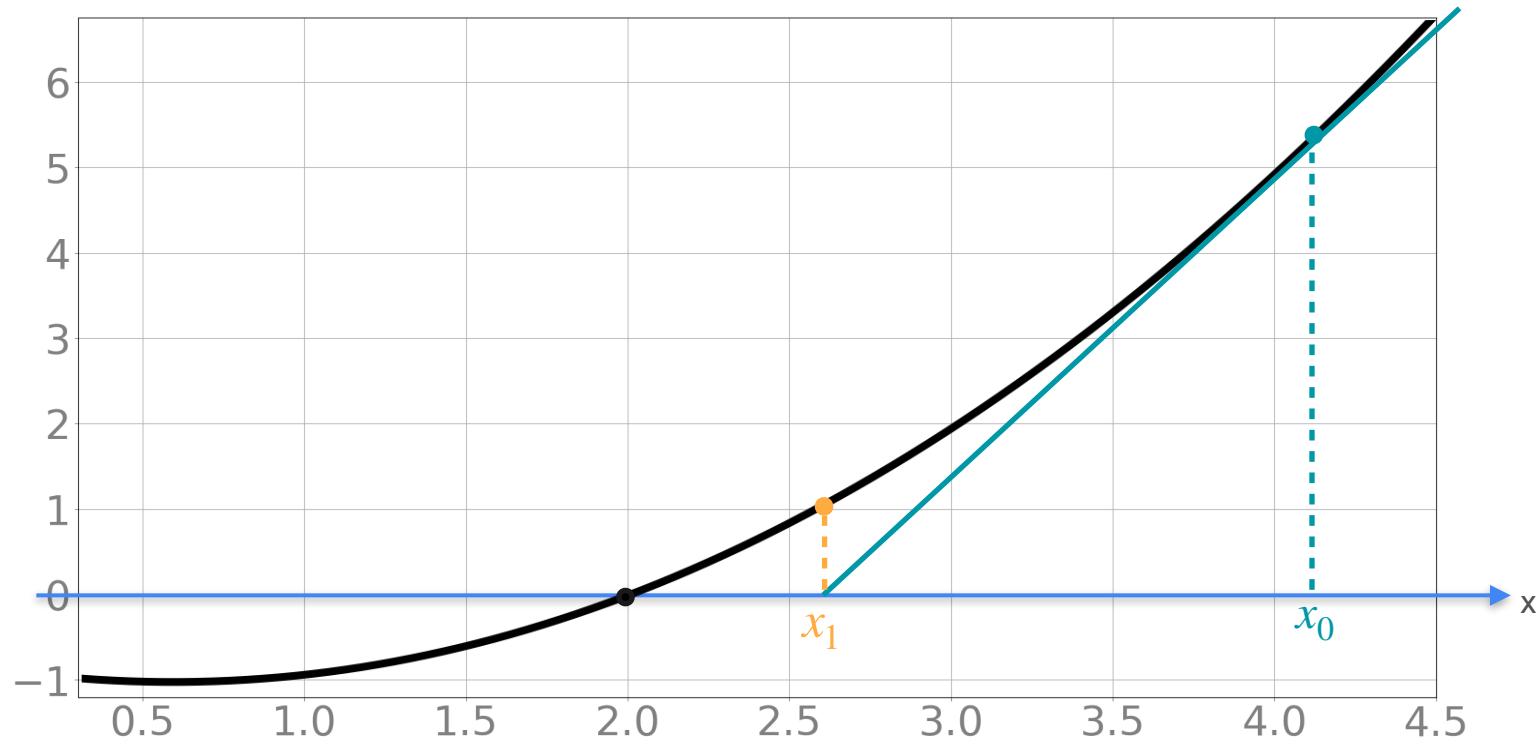
# Newton's Method



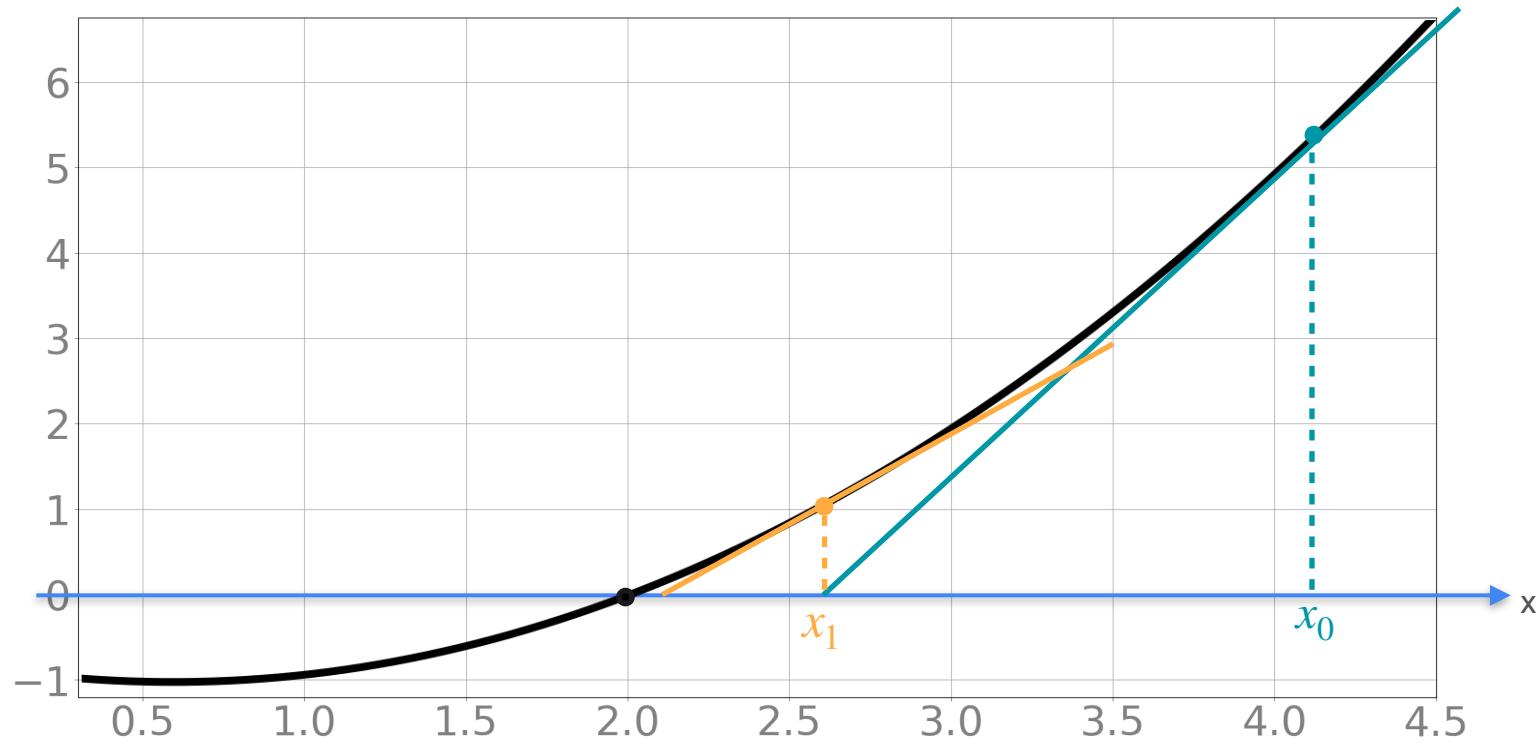
# Newton's Method



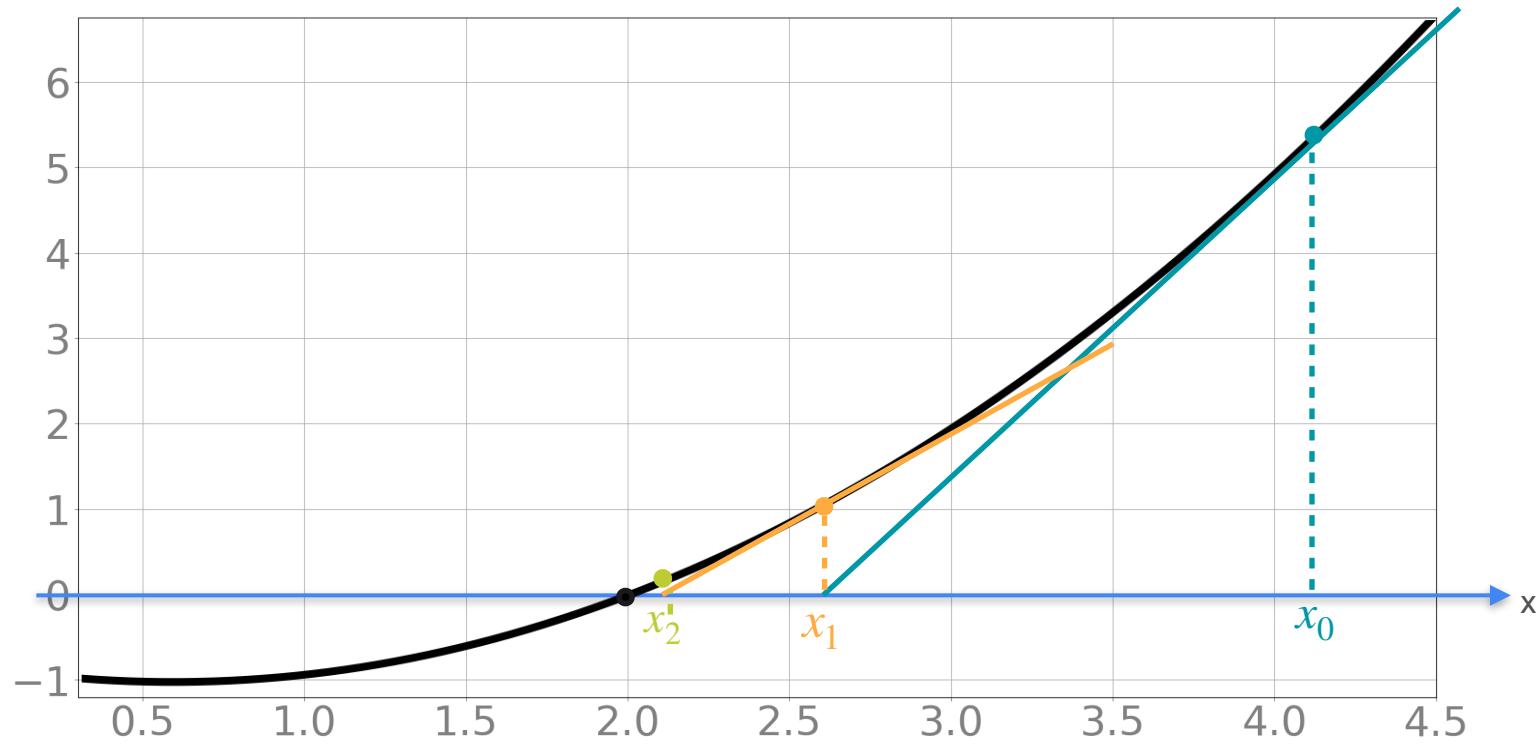
# Newton's Method



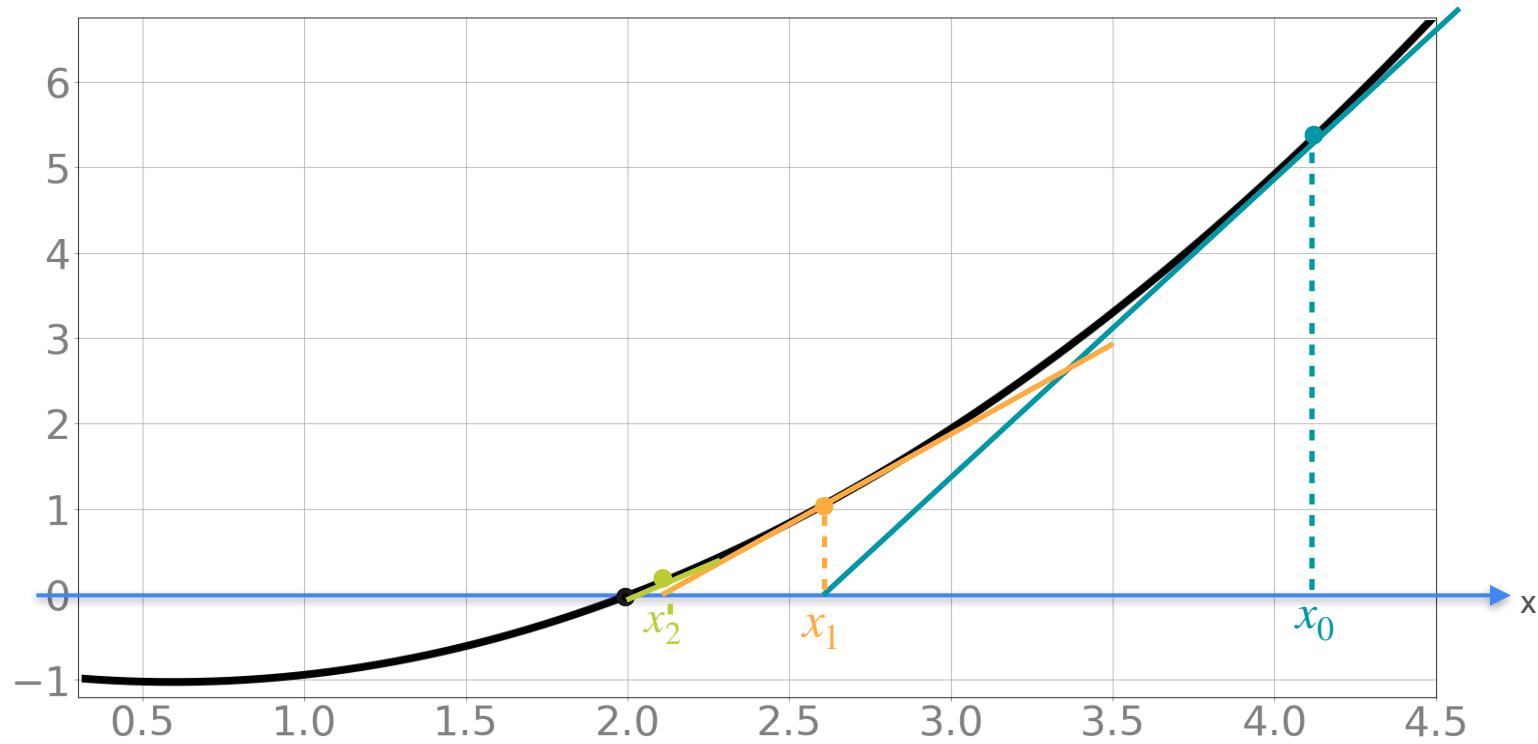
# Newton's Method



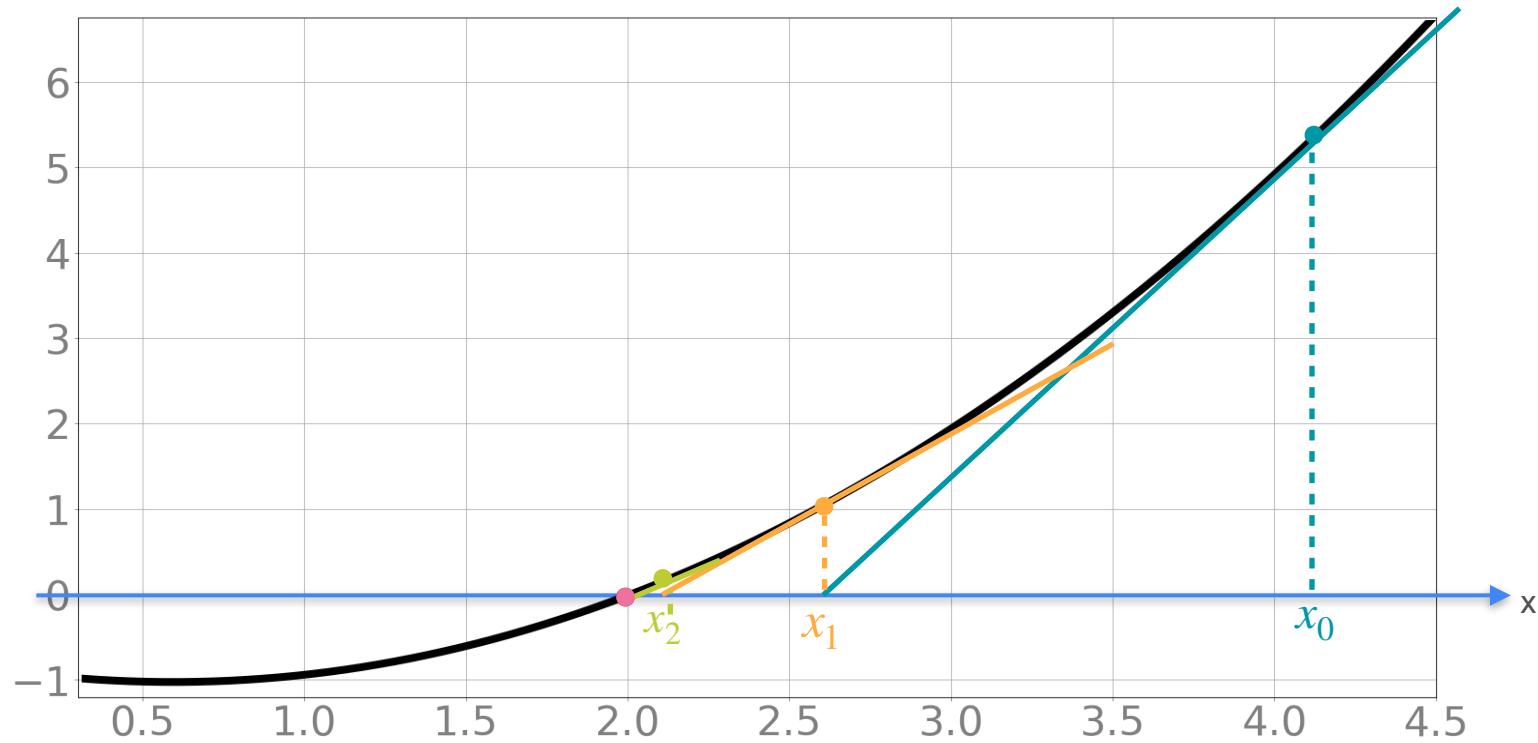
# Newton's Method



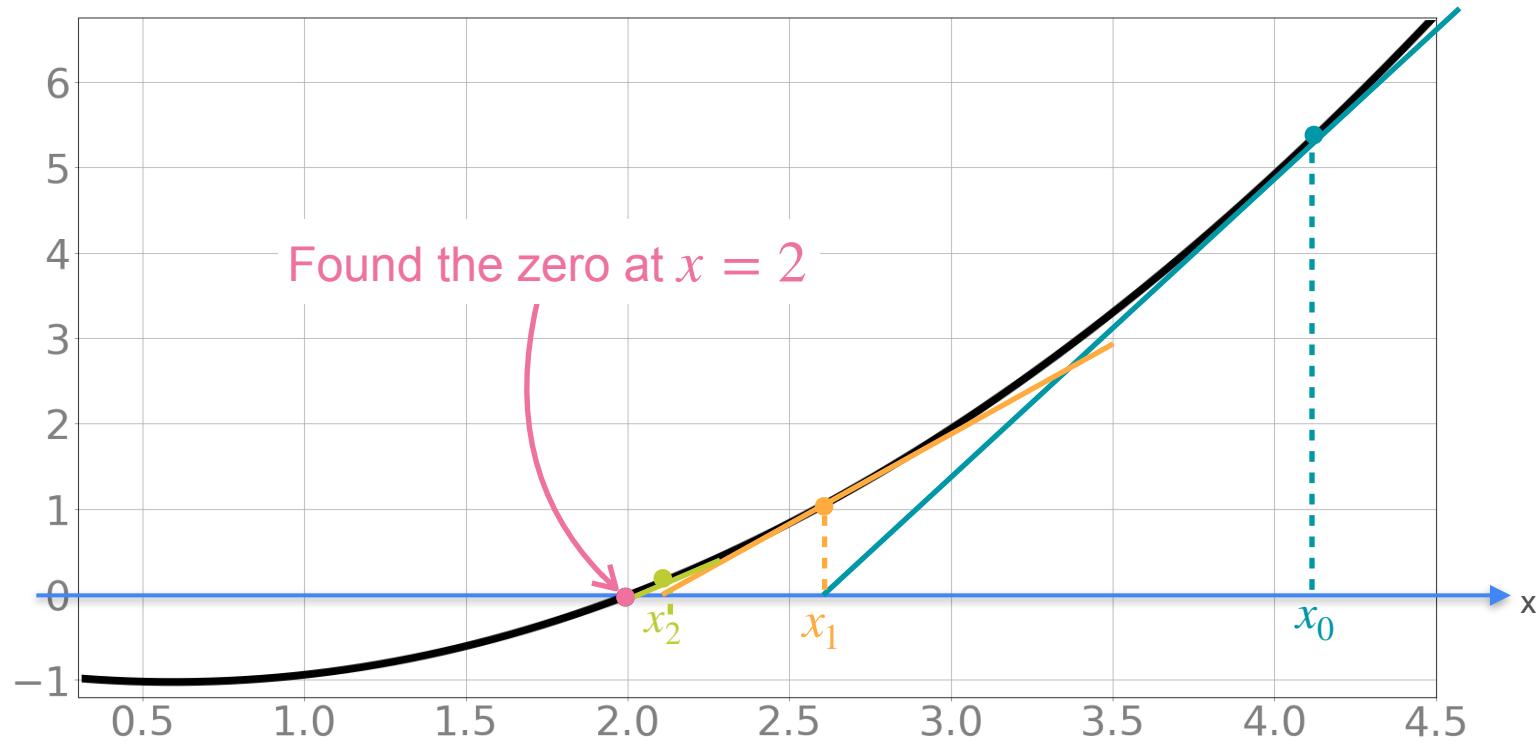
# Newton's Method



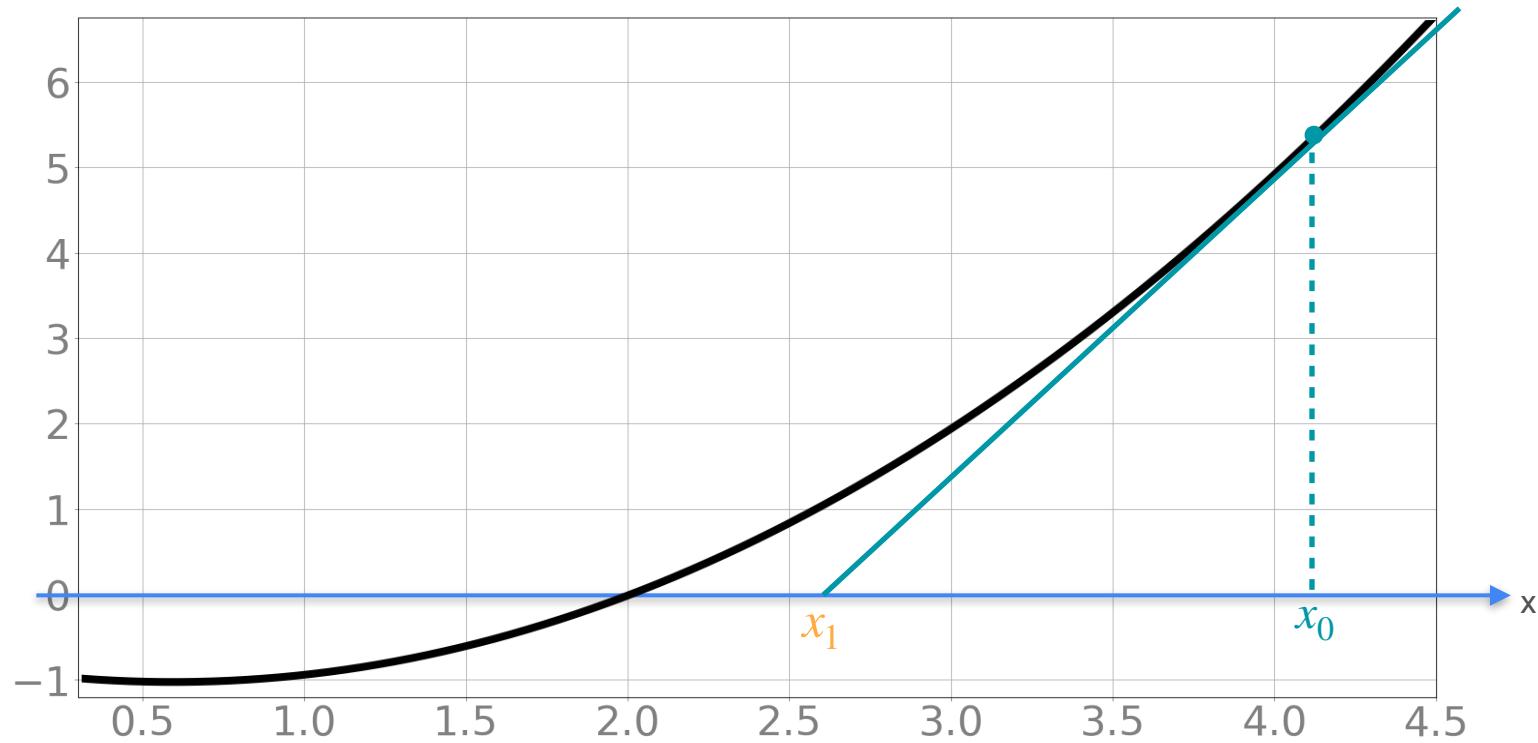
# Newton's Method



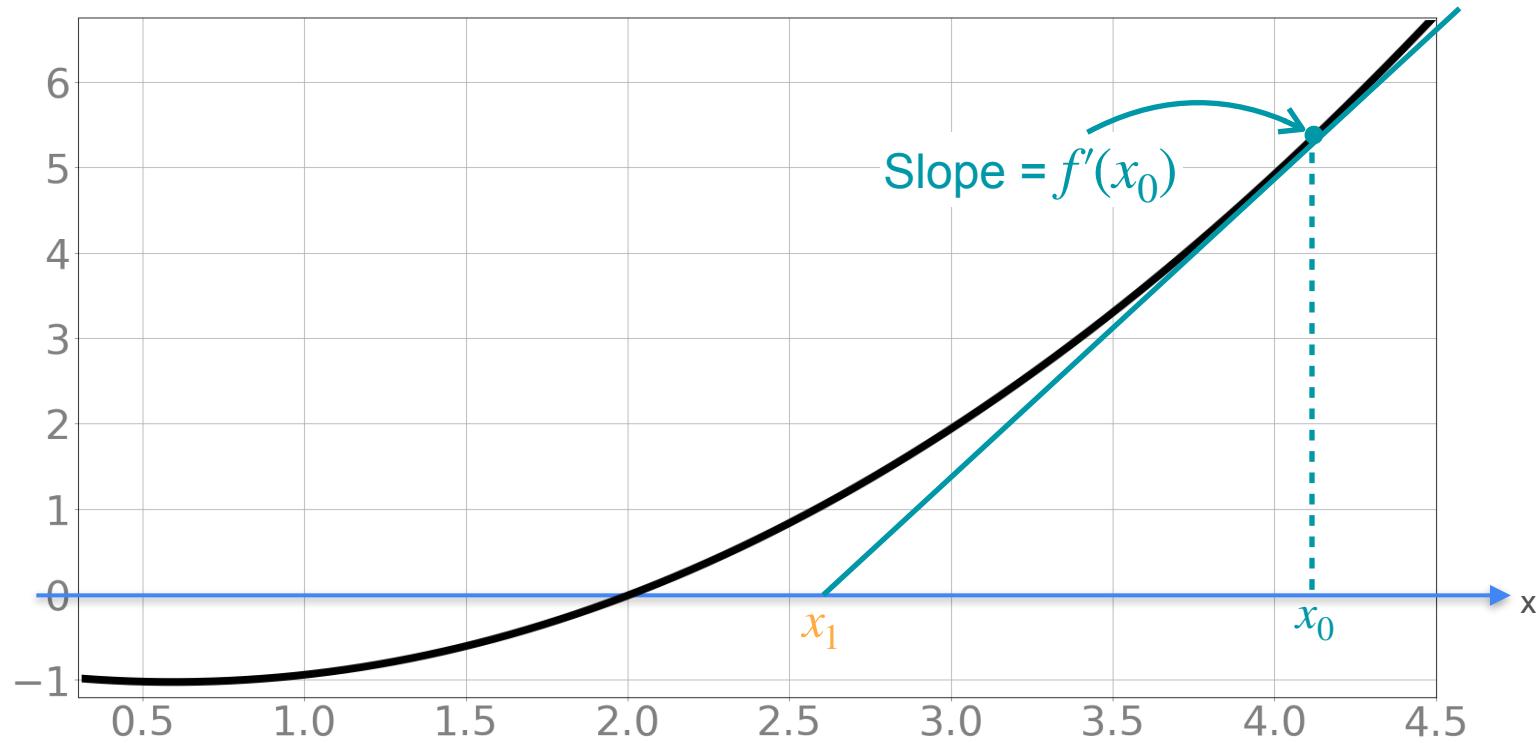
# Newton's Method



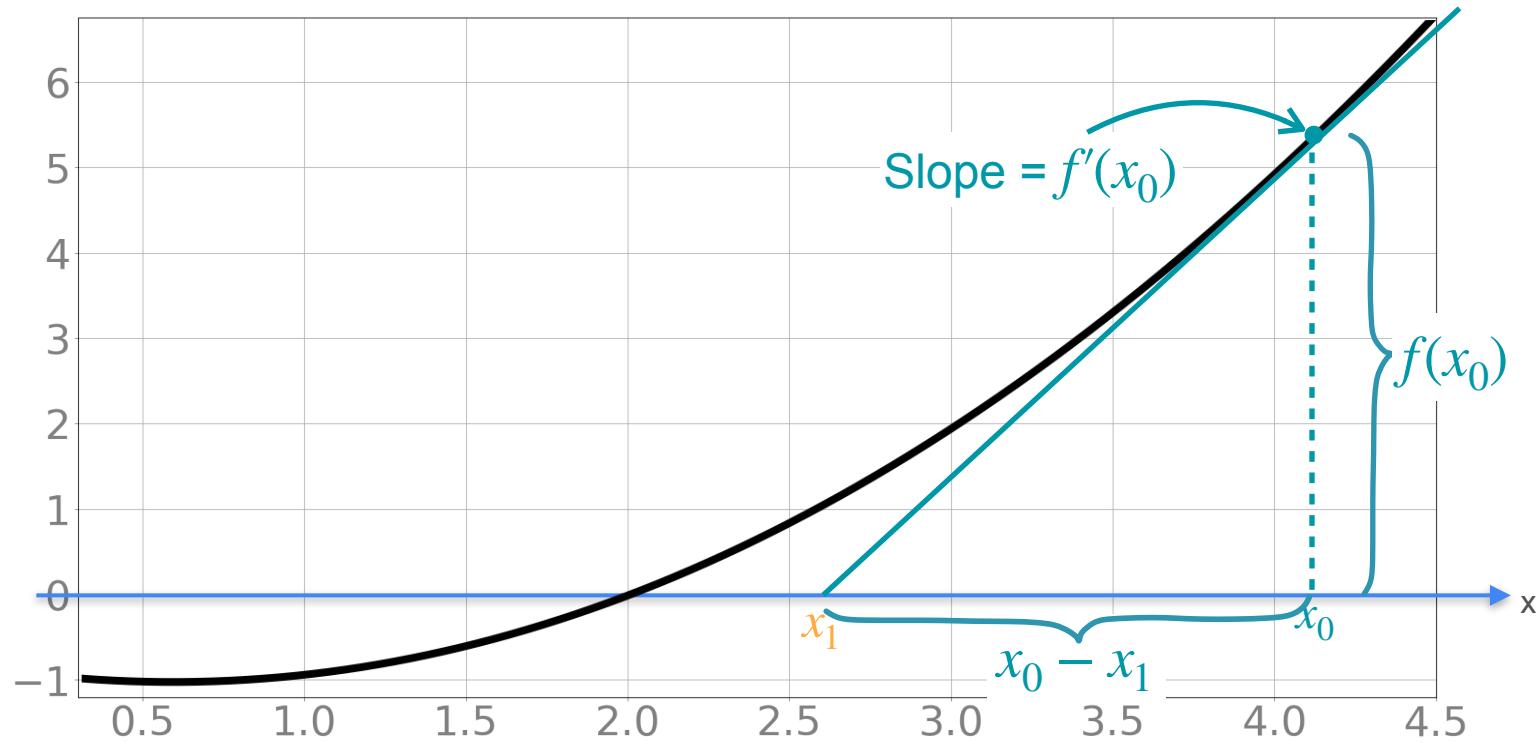
# Update Approximation



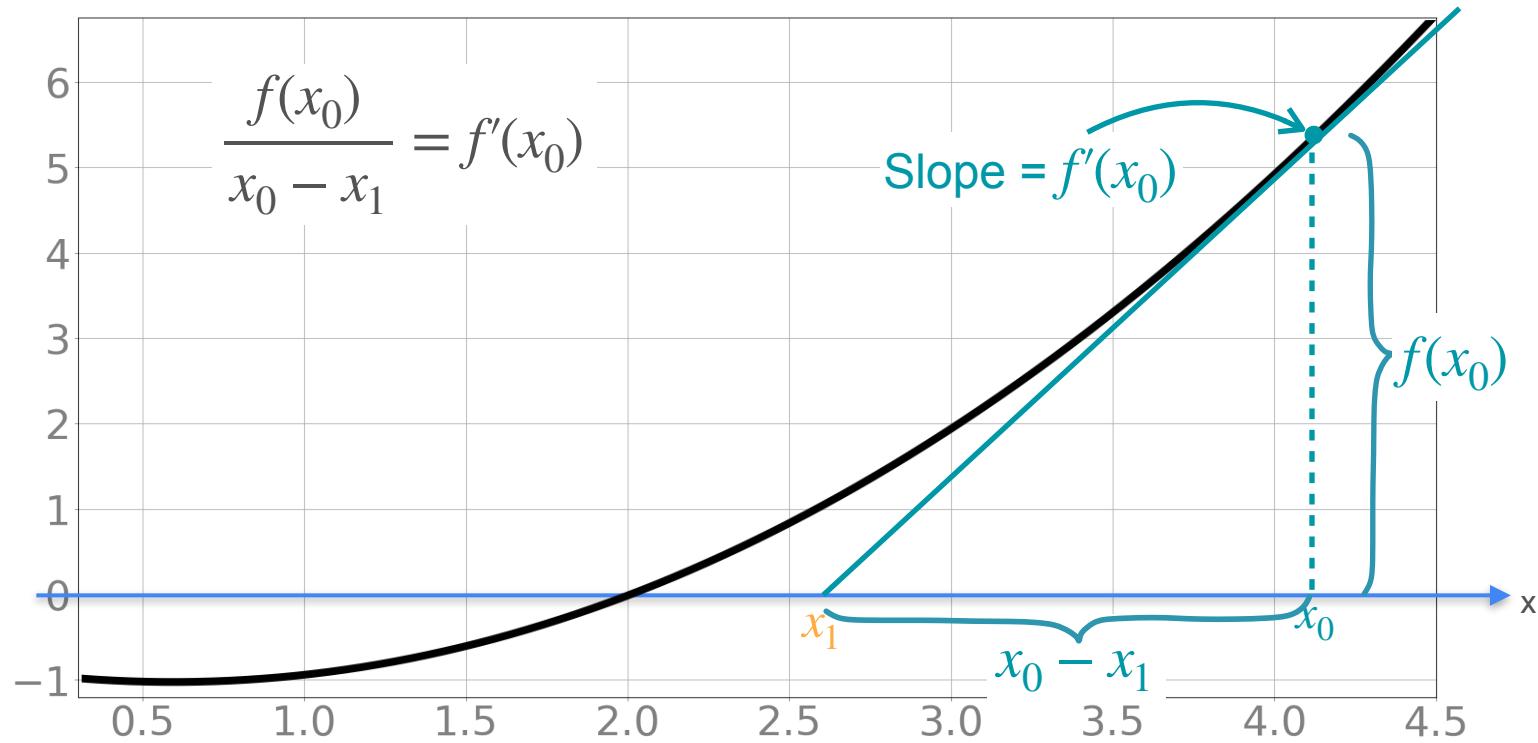
# Update Approximation



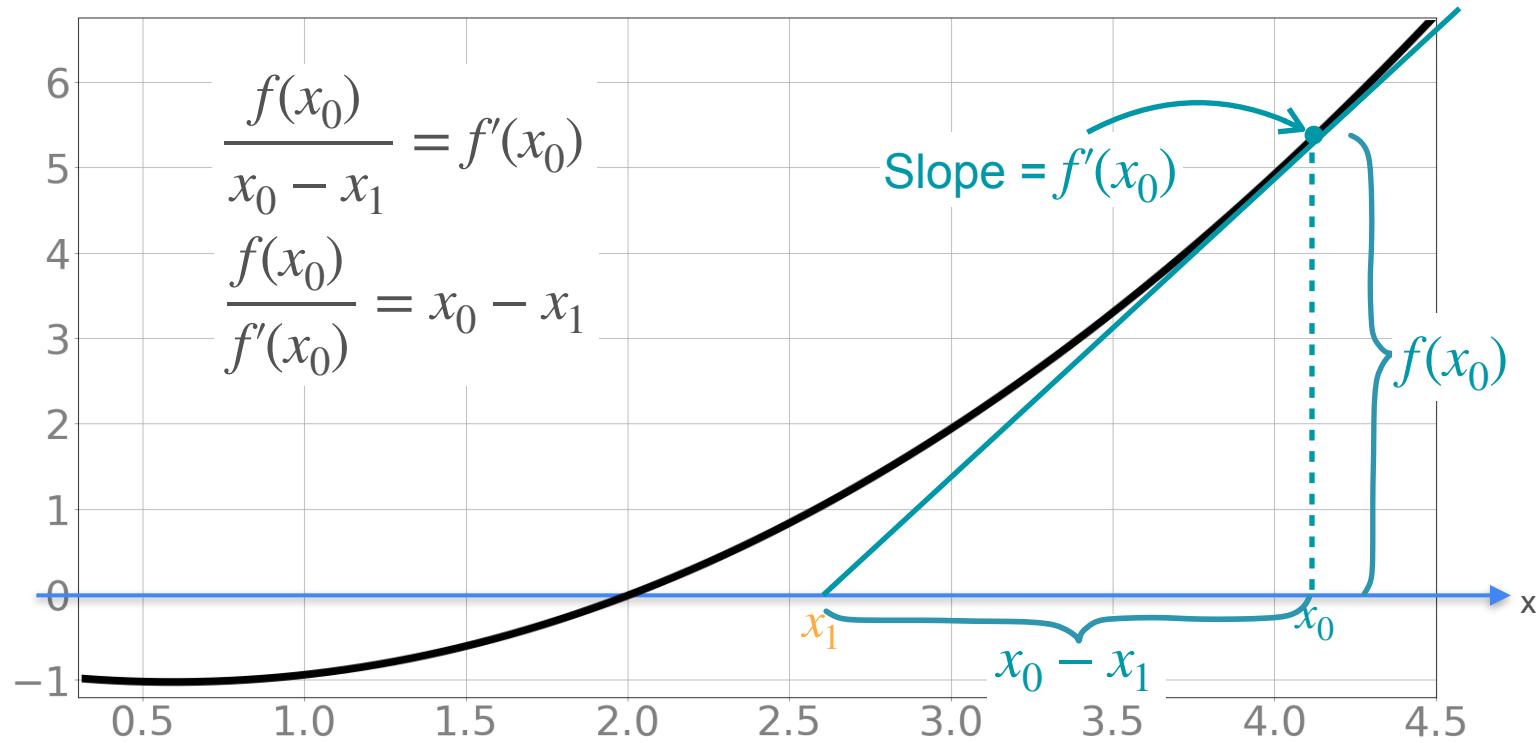
# Update Approximation



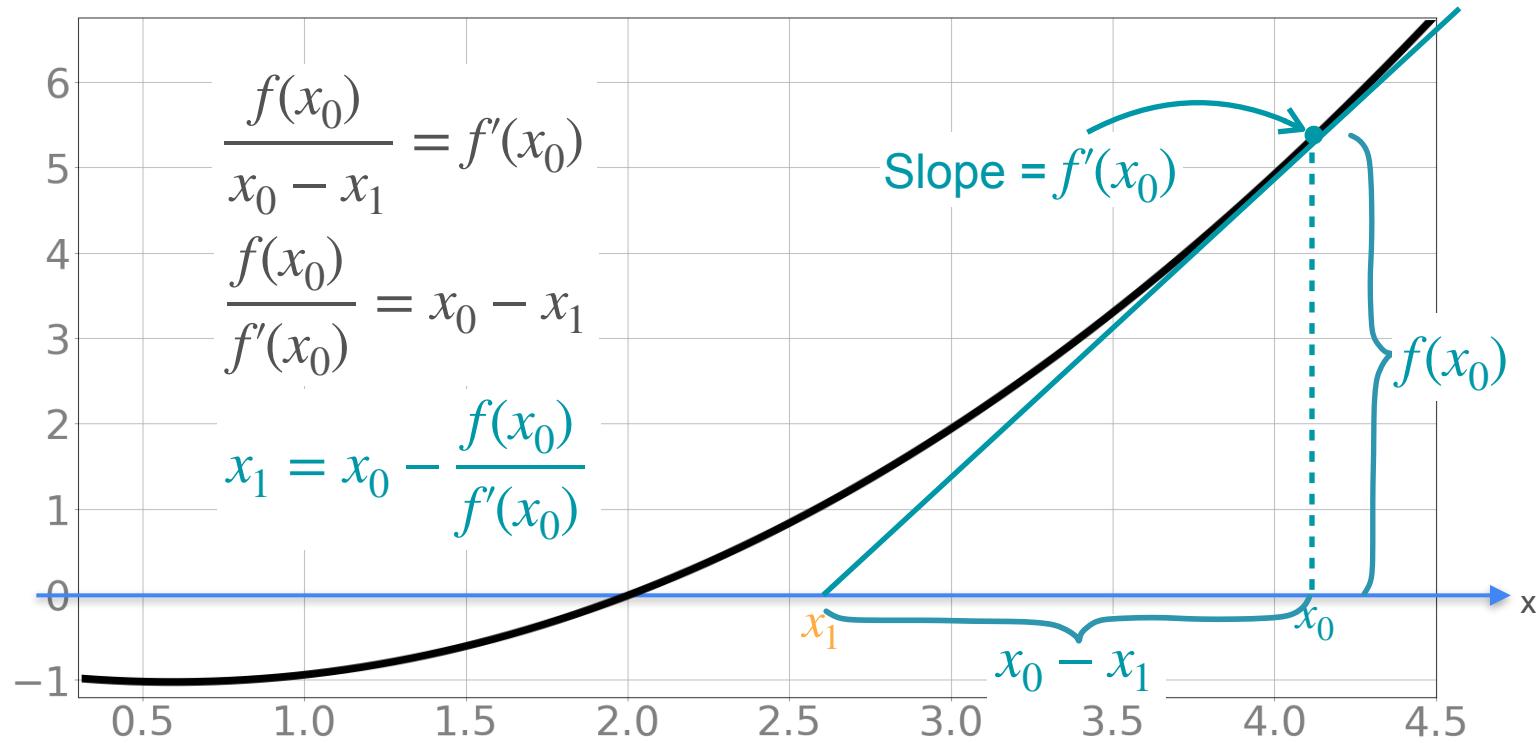
# Update Approximation



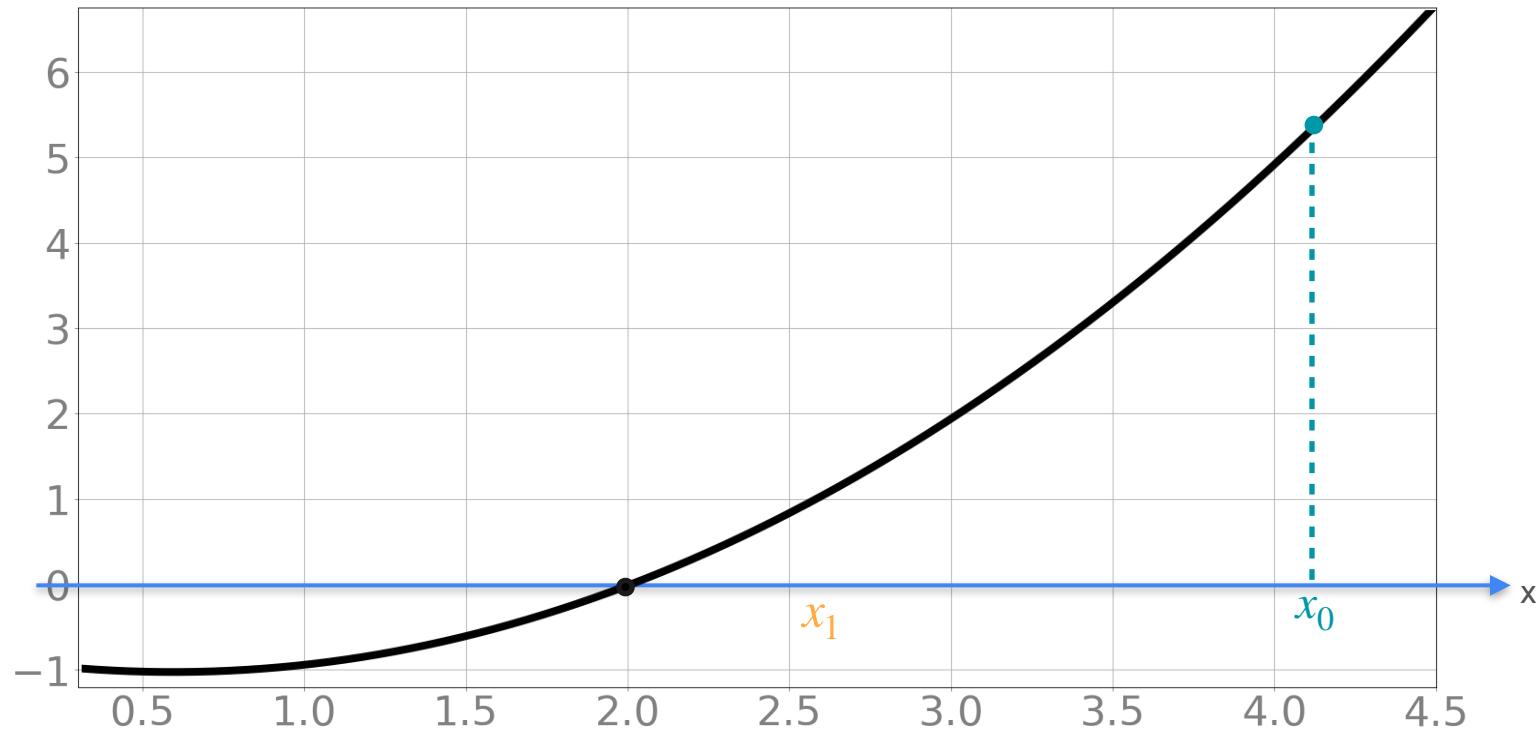
# Update Approximation



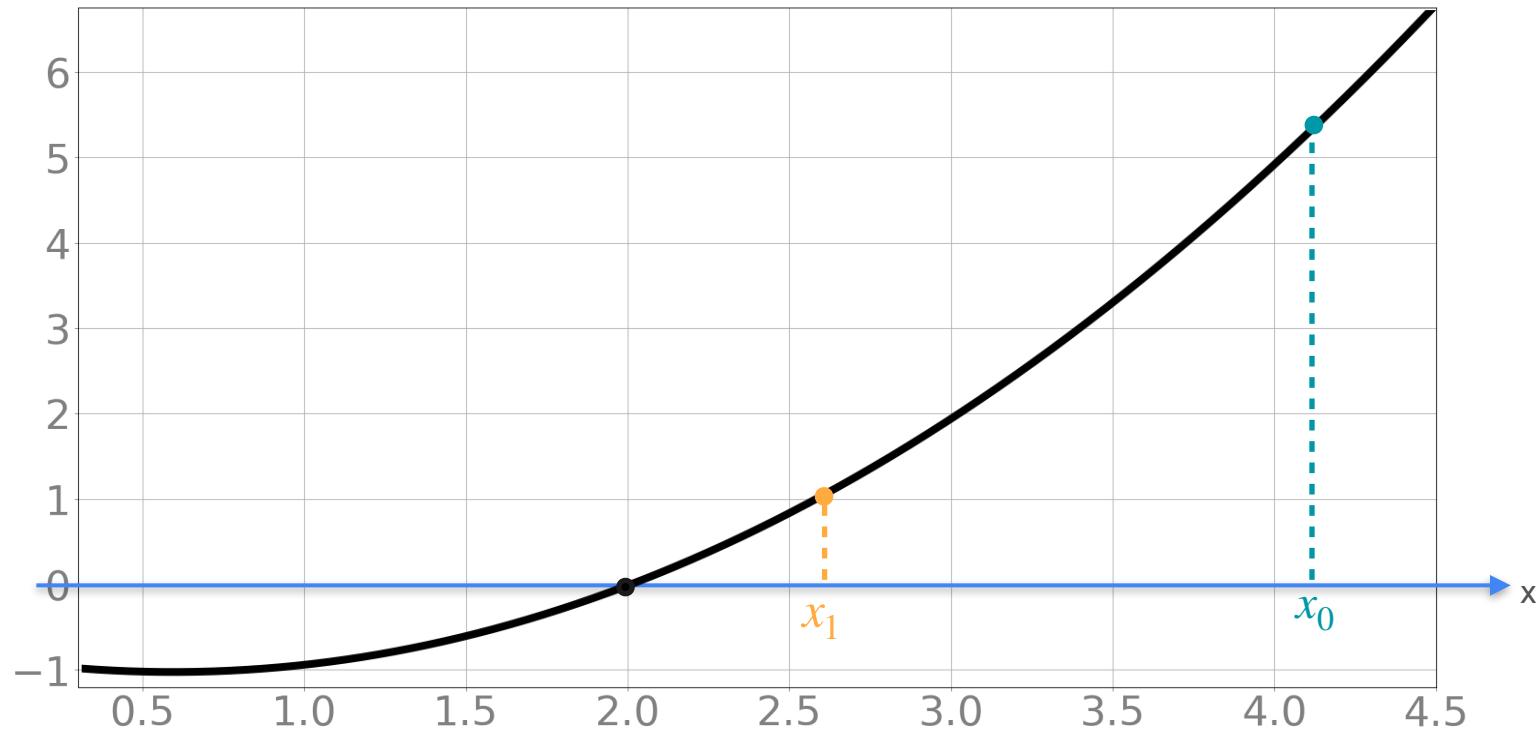
# Update Approximation



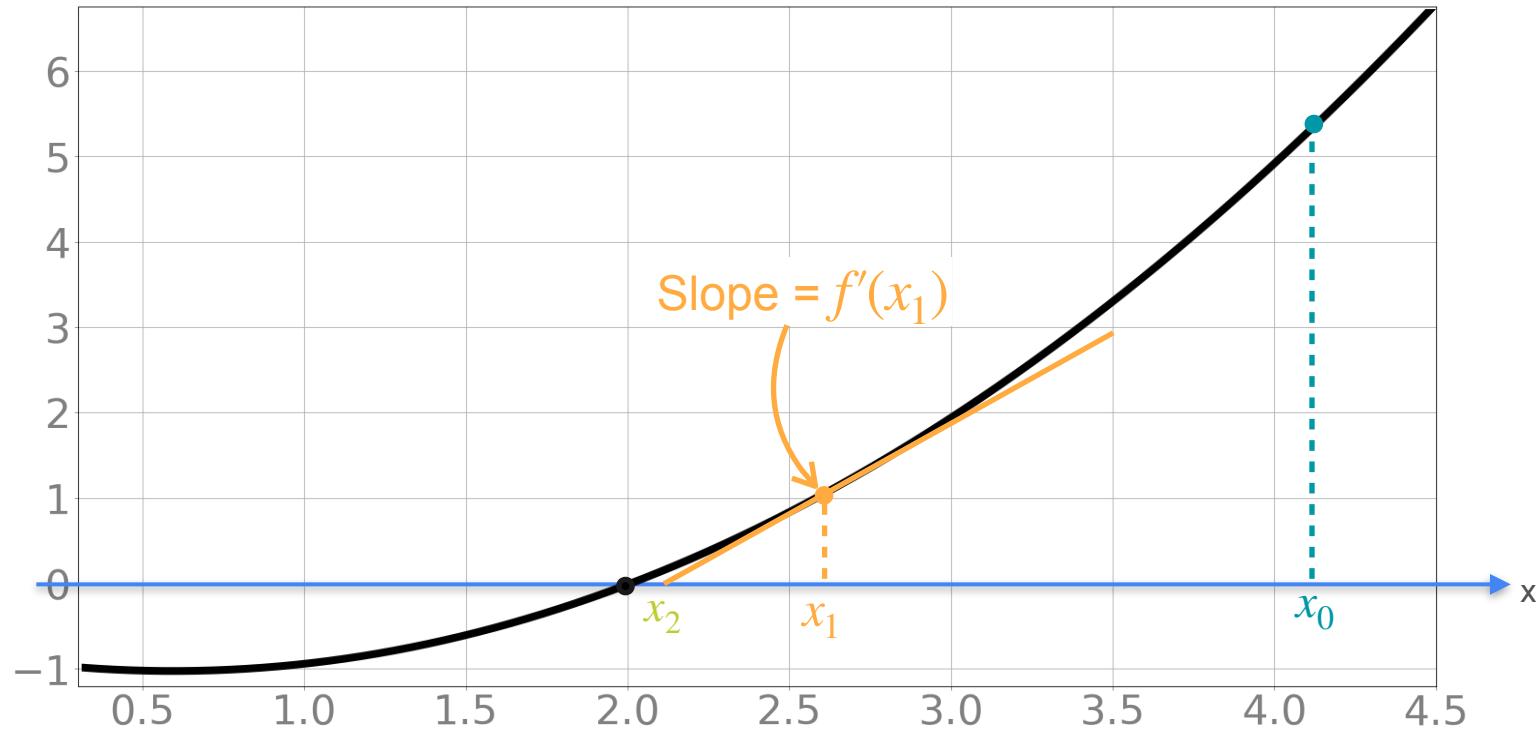
# Update Approximation



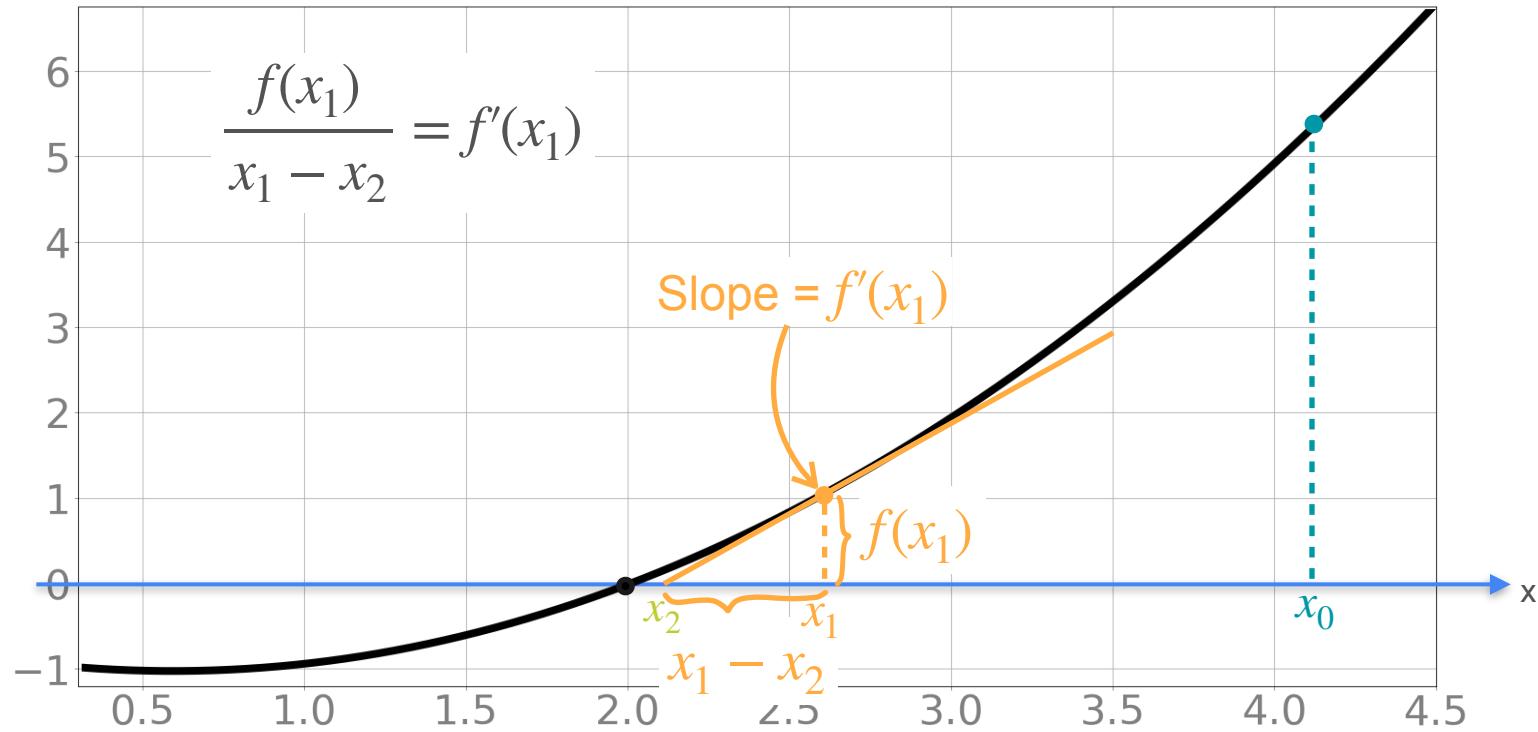
# Update Approximation



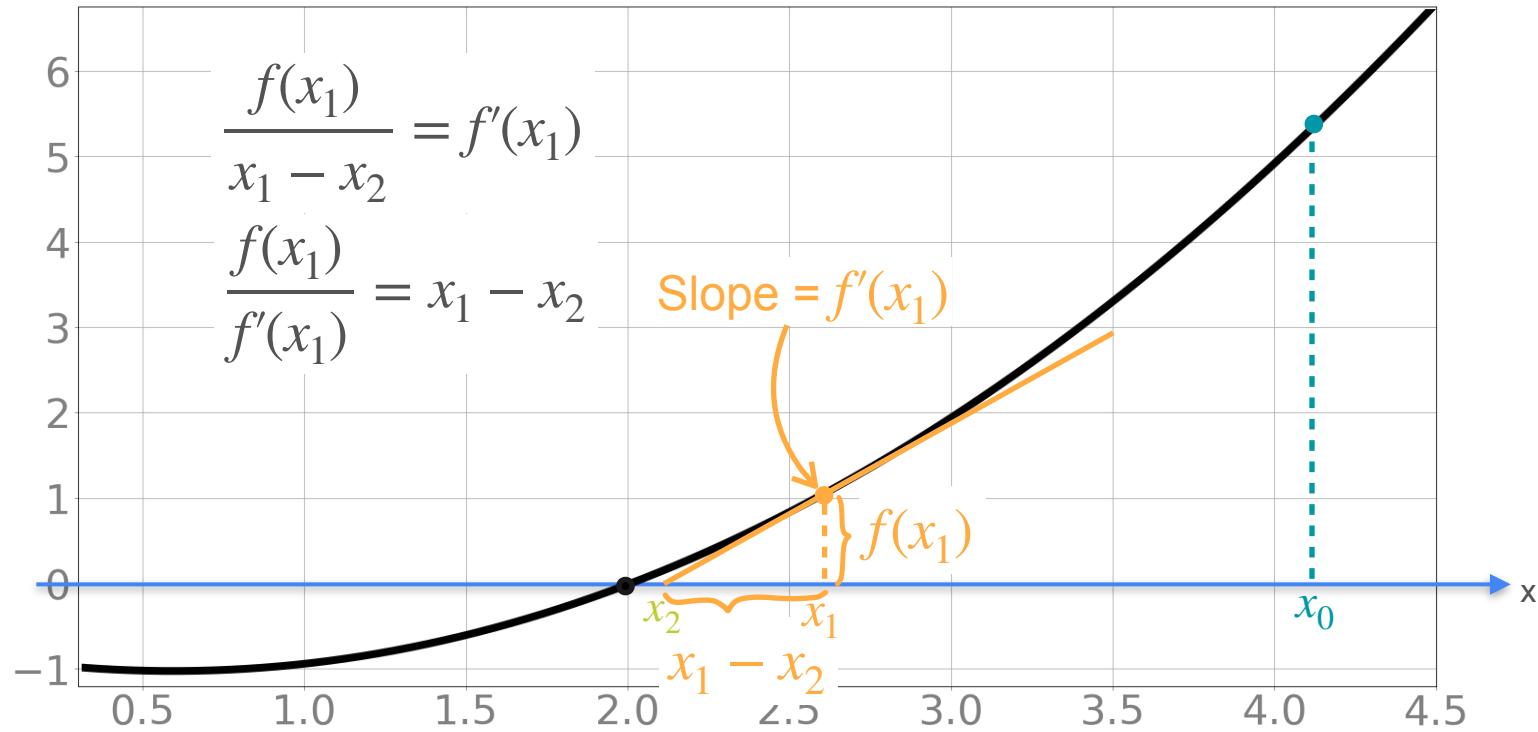
# Update Approximation



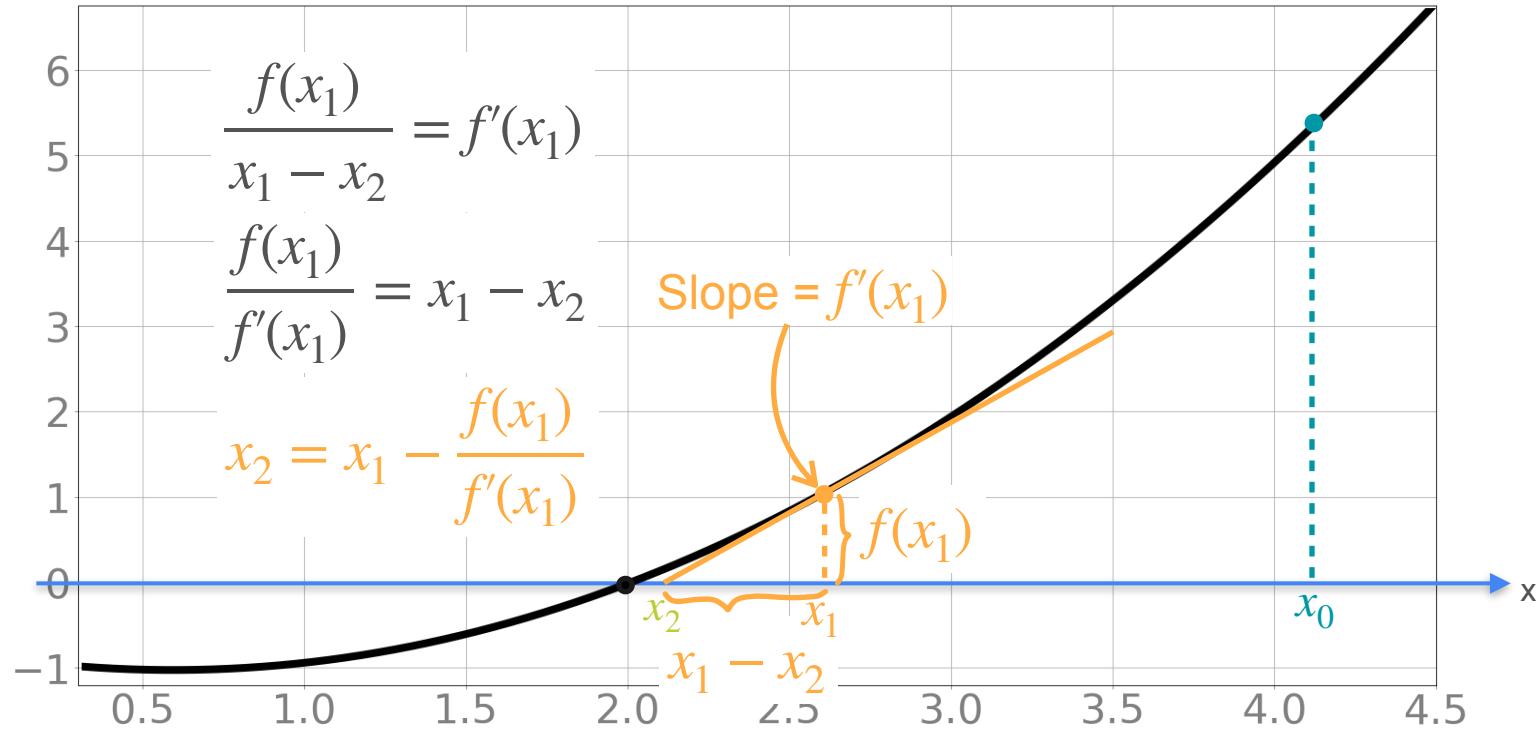
# Update Approximation



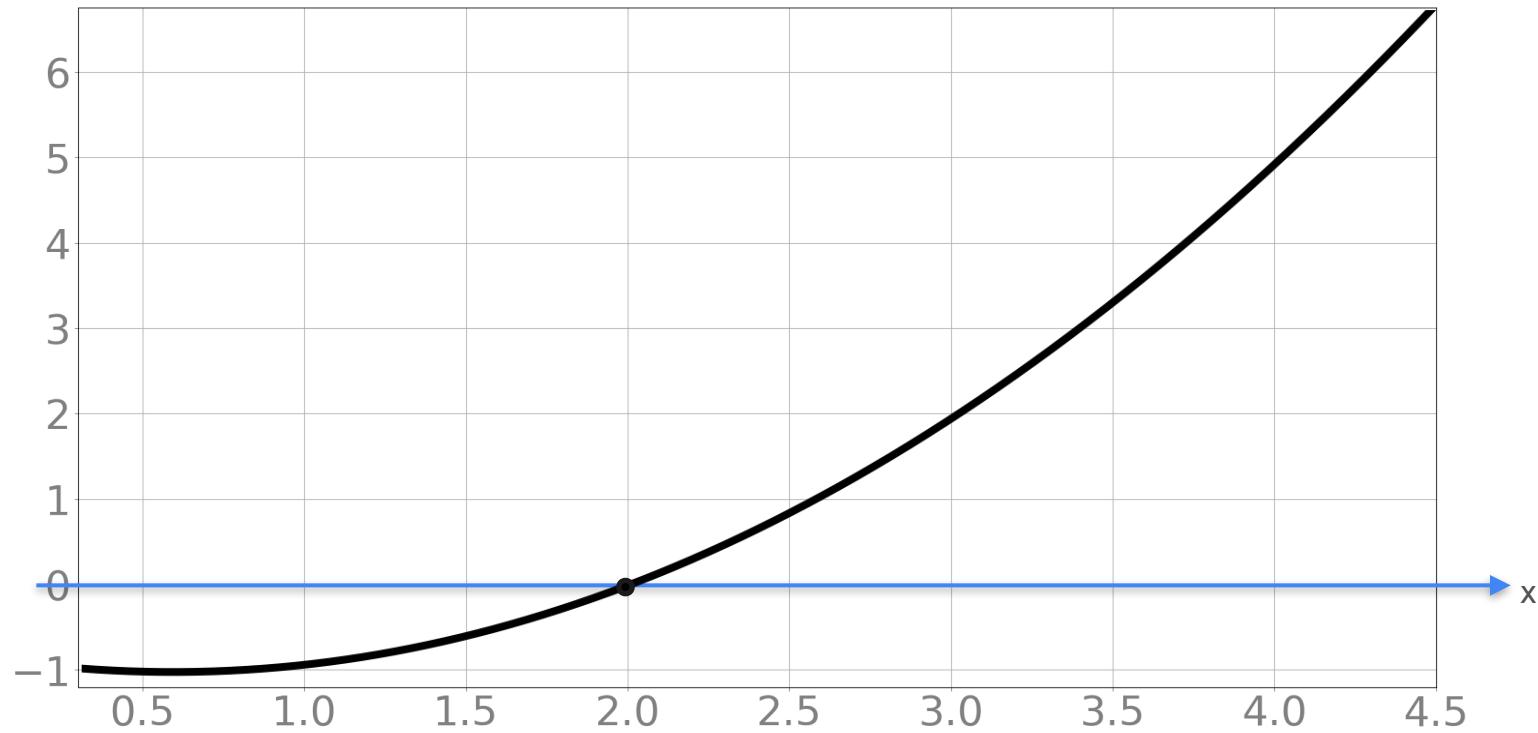
# Update Approximation



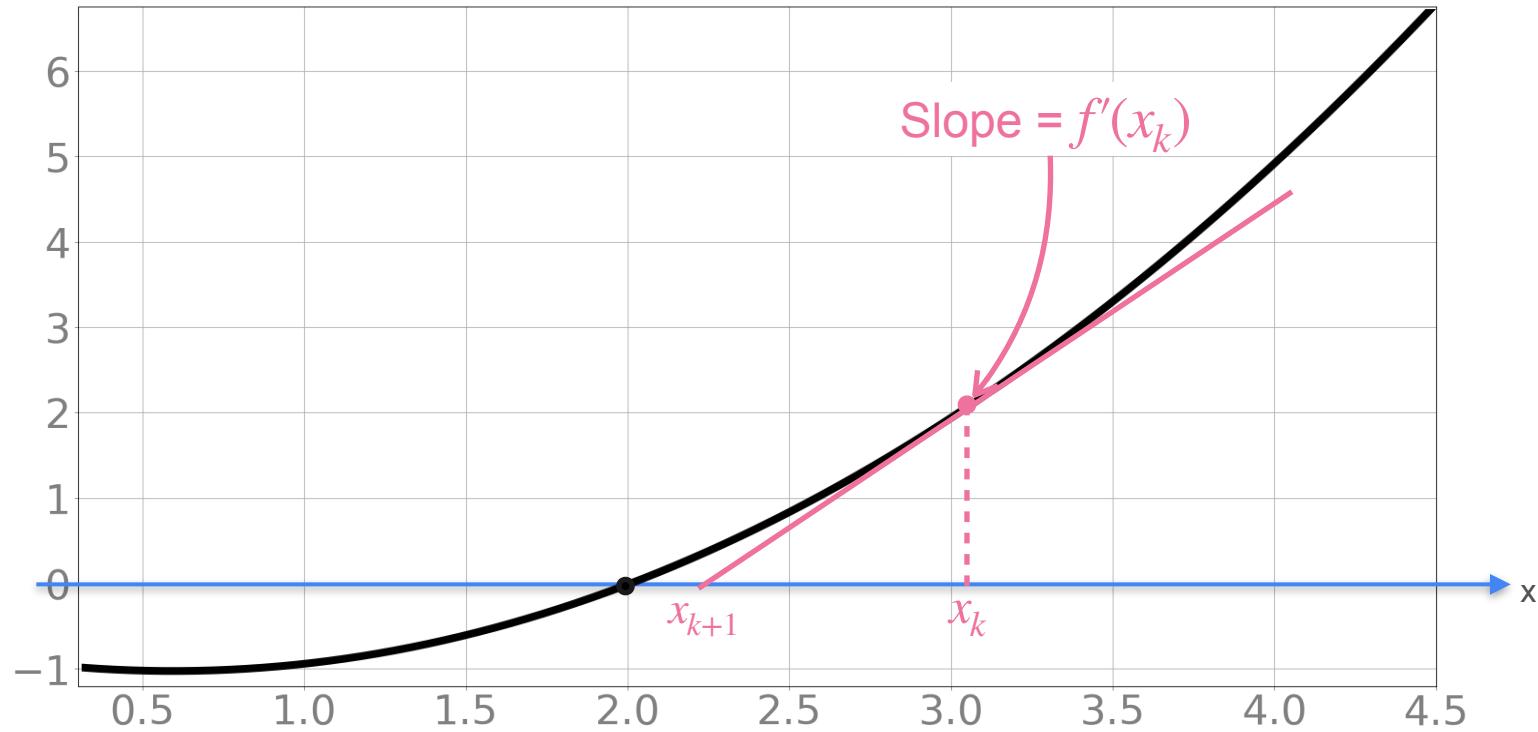
# Update Approximation



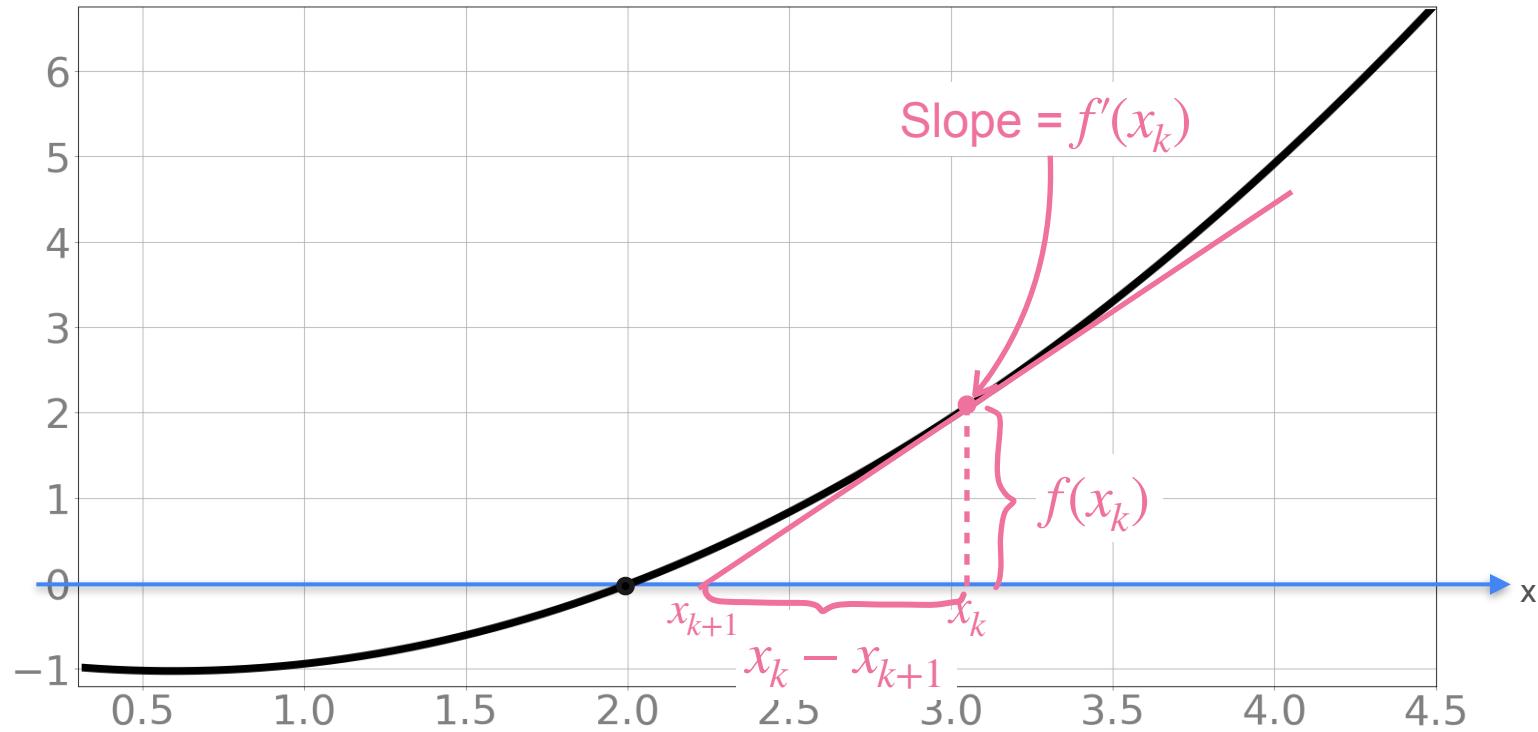
# Update Approximation



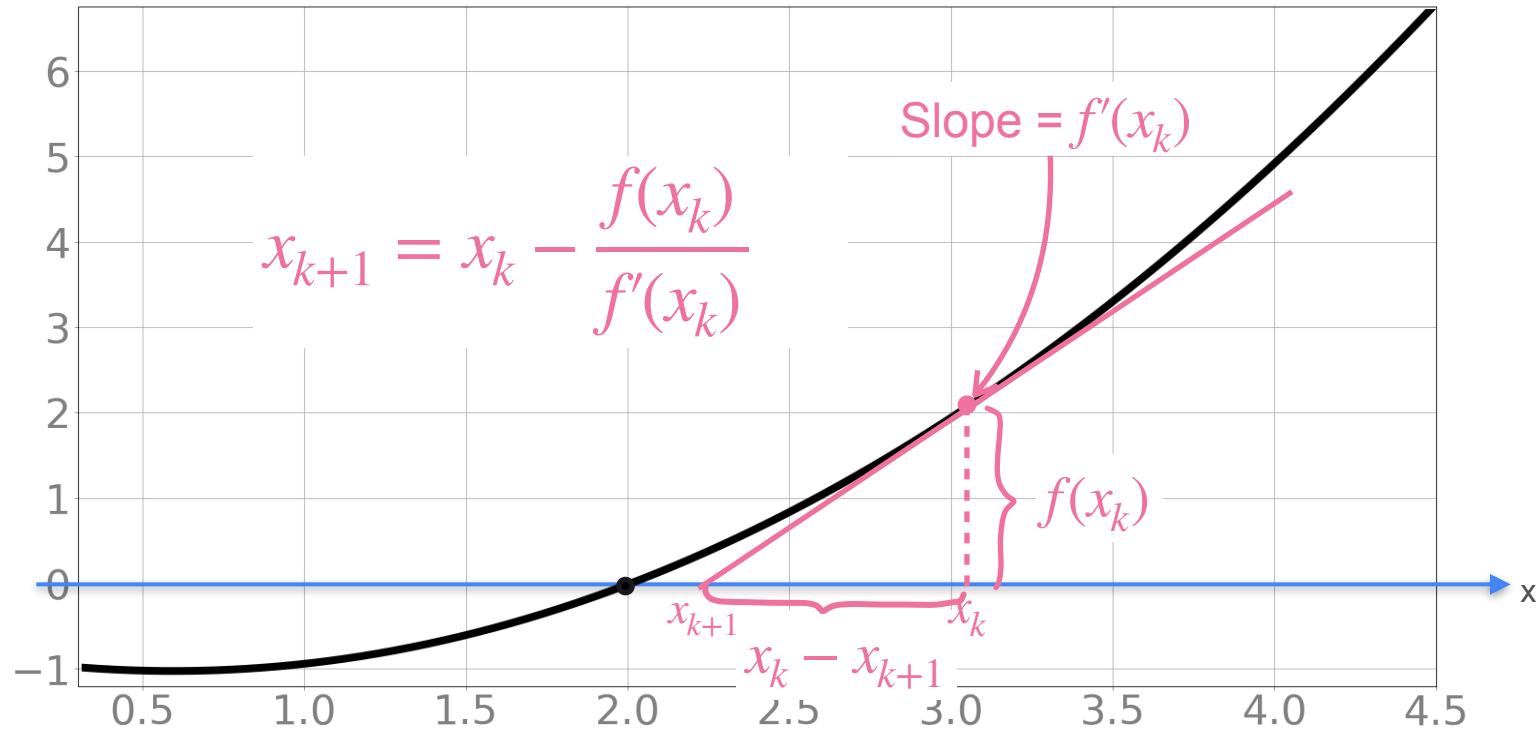
# Update Approximation



# Update Approximation



# Update Approximation



# Newton's Method for Optimization

# Newton's Method for Optimization



# Newton's Method for Optimization

Newton's method

Goal: find a zero of  $f(x)$



# Newton's Method for Optimization

Newton's method

Goal: find a zero of  $f(x)$



NM for Optimization

Goal: minimize  $g(x)$  find zeros of  $g'(x)$

# Newton's Method for Optimization

Newton's method

Goal: find a zero of  $f(x)$



NM for Optimization

Goal: minimize  $g(x)$  find zeros of  $g'(x)$

$$f(x) \mapsto g'(x)$$

$$f'(x) \mapsto (g'(x))'$$

# Newton's Method for Optimization

Newton's method

Goal: find a zero of  $f(x)$

1) Start with some  $x_0$



NM for Optimization

Goal: minimize  $g(x)$  find zeros of  $g'(x)$

$$f(x) \mapsto g'(x) \qquad f'(x) \mapsto (g'(x))'$$

1) Start with some  $x_0$

# Newton's Method for Optimization

Newton's method

Goal: find a zero of  $f(x)$

1) Start with some  $x_0$

2) Update:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$



NM for Optimization

Goal: minimize  $g(x)$  find zeros of  $g'(x)$

$$f(x) \mapsto g'(x)$$

$$f'(x) \mapsto (g'(x))'$$

1) Start with some  $x_0$

2) Update:

# Newton's Method for Optimization

Newton's method

Goal: find a zero of  $f(x)$

1) Start with some  $x_0$

2) Update:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$



NM for Optimization

Goal: minimize  $g(x)$  find zeros of  $g'(x)$

$$f(x) \mapsto g'(x)$$

$$f'(x) \mapsto (g'(x))'$$

1) Start with some  $x_0$

2) Update:

$$x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$$

# Newton's Method for Optimization

## Newton's method

Goal: find a zero of  $f(x)$

1) Start with some  $x_0$

2) Update:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

3) Repeat 2) until you find the root.



## NM for Optimization

Goal: minimize  $g(x)$  find zeros of  $g'(x)$

$$f(x) \mapsto g'(x)$$

$$f'(x) \mapsto (g'(x))'$$

1) Start with some  $x_0$

2) Update:

$$x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$$

3) Repeat 2) until you find the root.



DeepLearning.AI

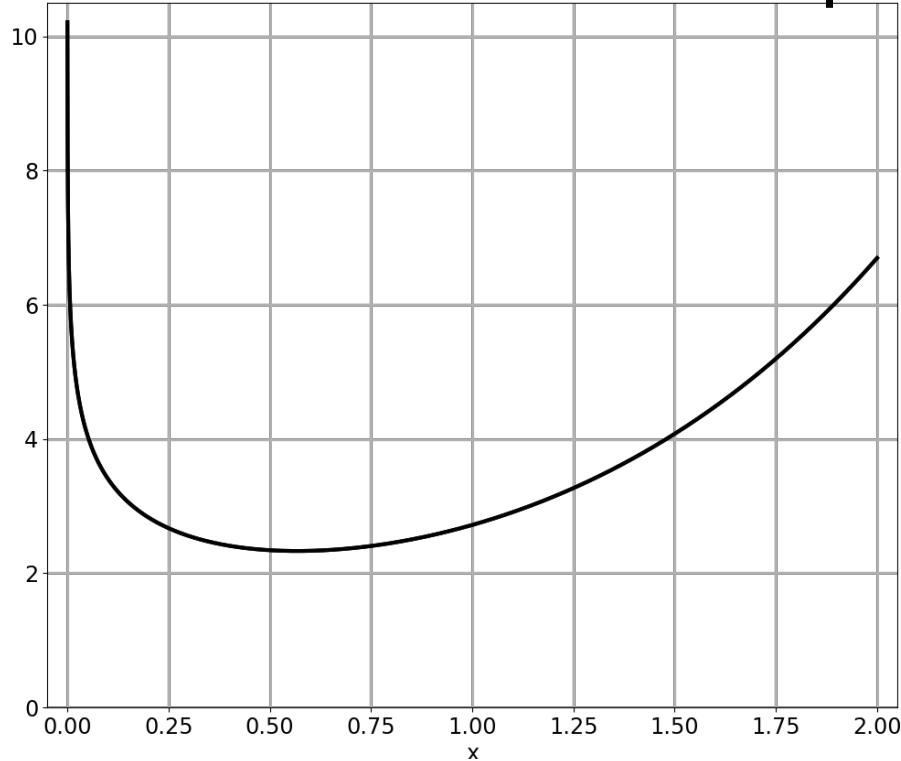
# Optimization in Neural Networks and Newton's Method

---

**Newton's method:  
An example**

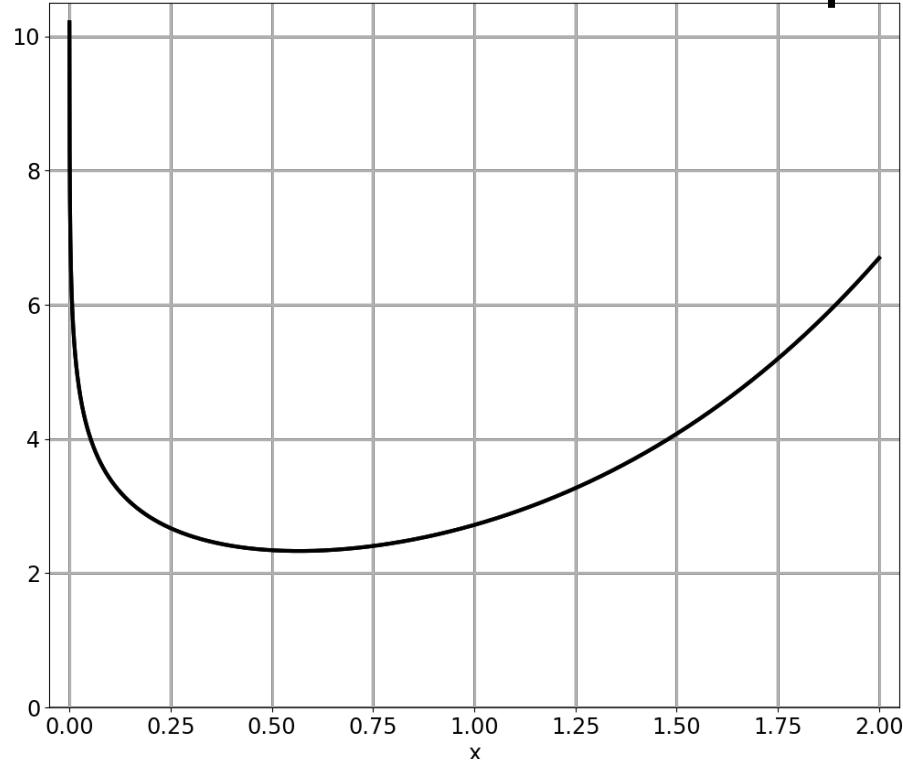
# Newton's Method for Optimization

# Newton's Method for Optimization



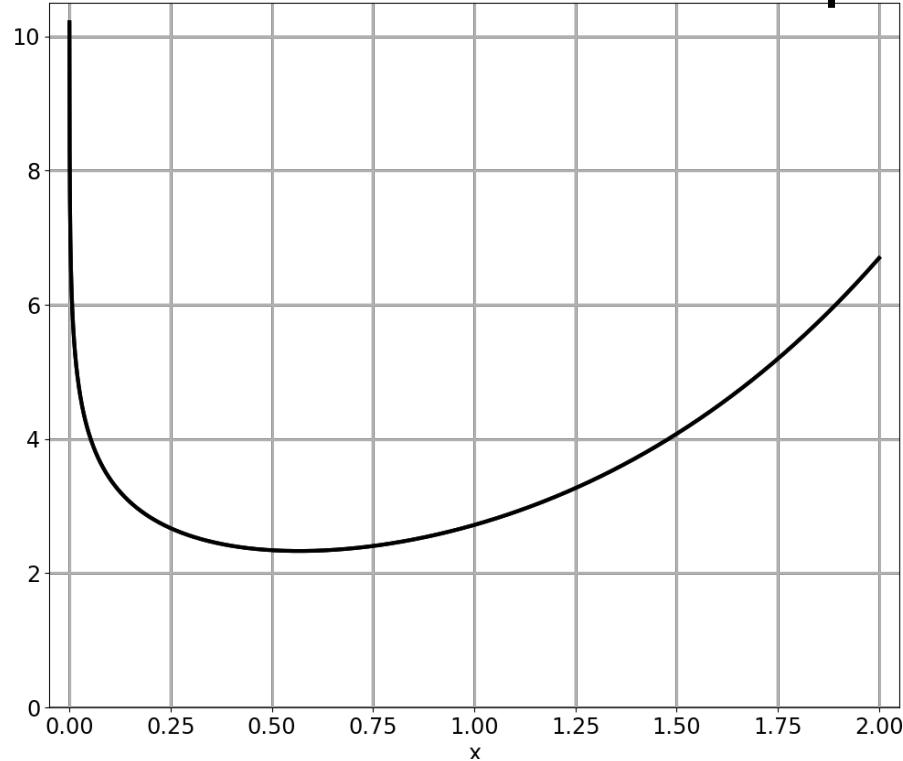
$$g(x) = e^x - \log(x)$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

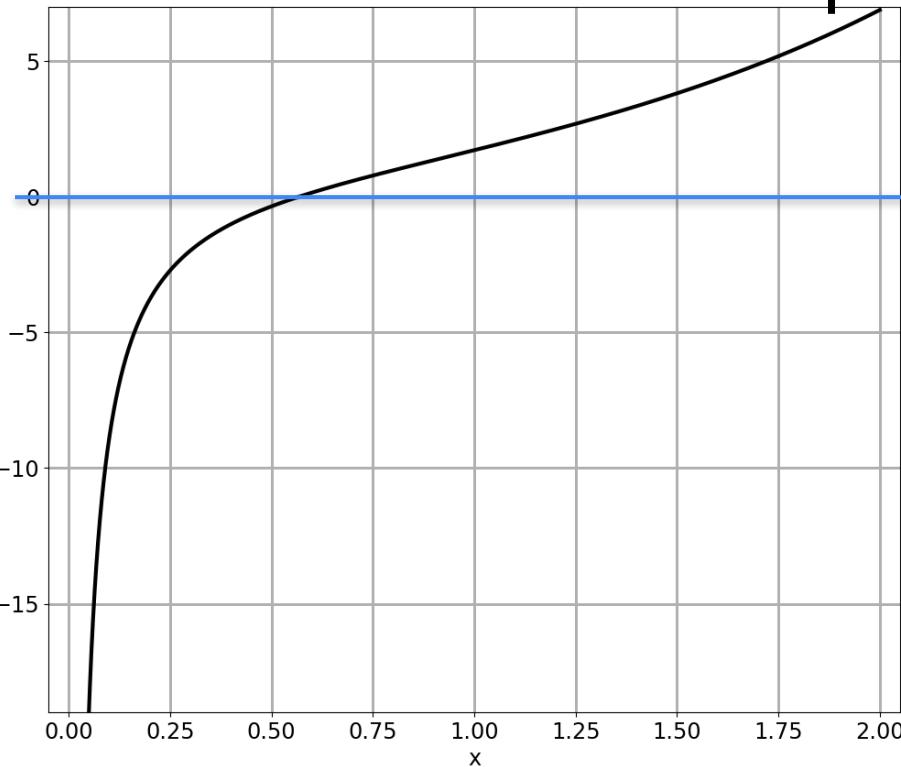
# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

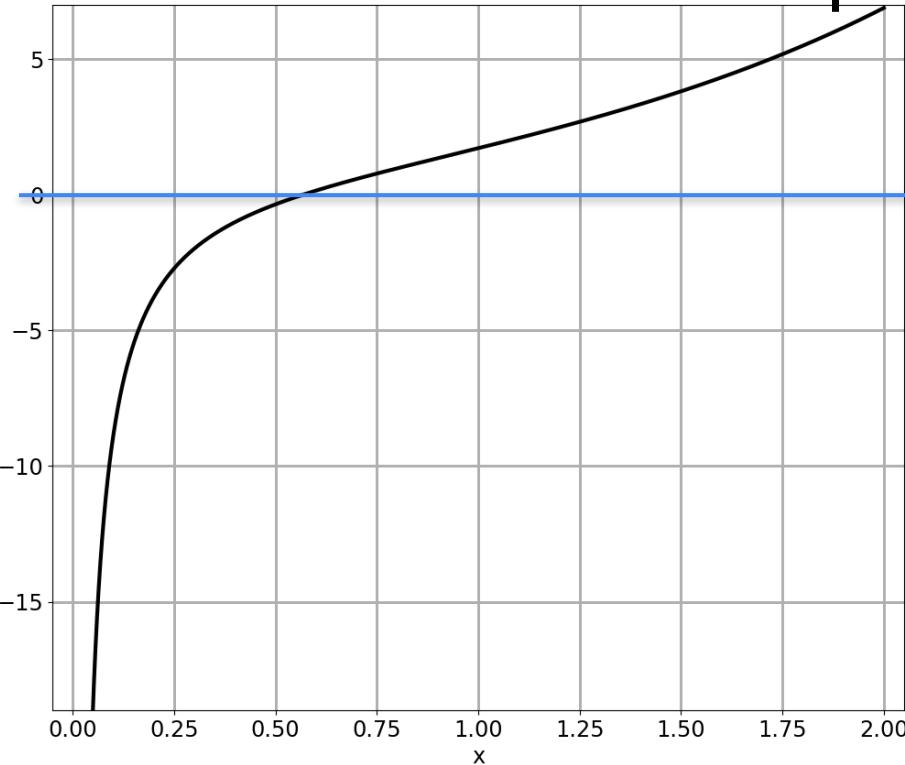
# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

# Newton's Method for Optimization

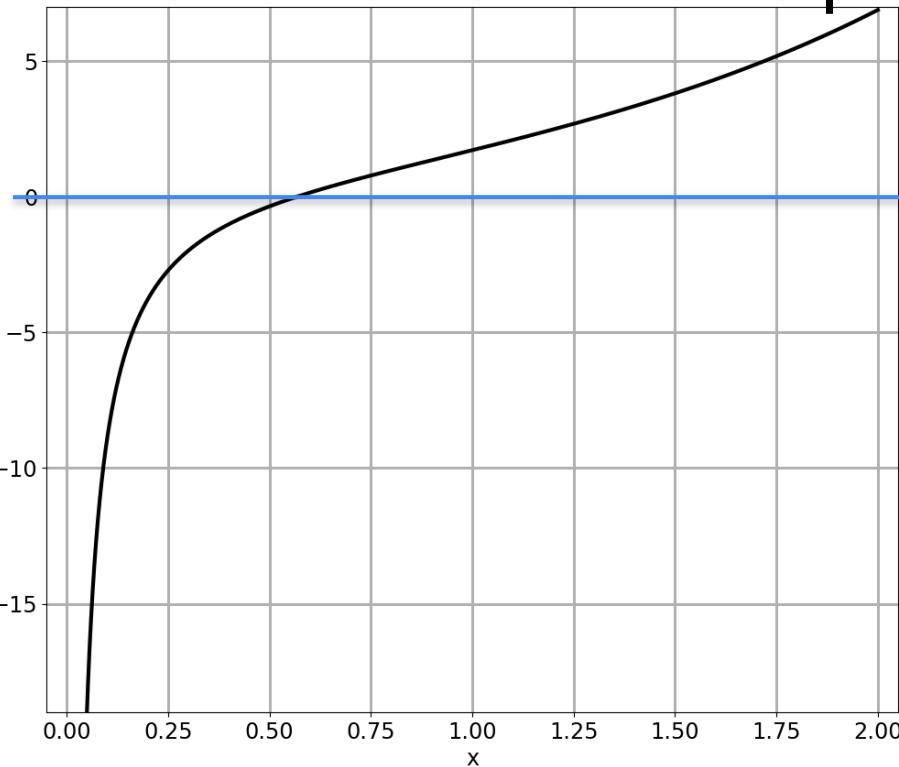


$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

# Newton's Method for Optimization

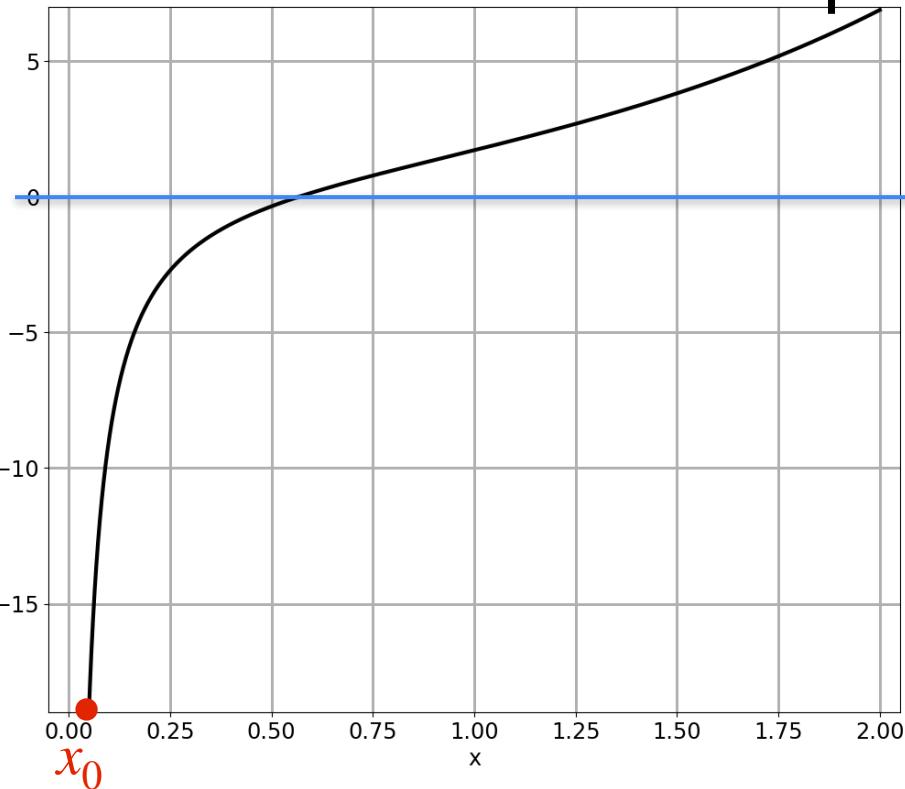


$$g(x) = e^x - \log(x) \quad \underbrace{g'(x) = e^x - 1/x}_{f(x)}$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2} \quad \underbrace{\phantom{(g'(x))'} f'(x)}_{\frac{1}{x^2}}$$

# Newton's Method for Optimization

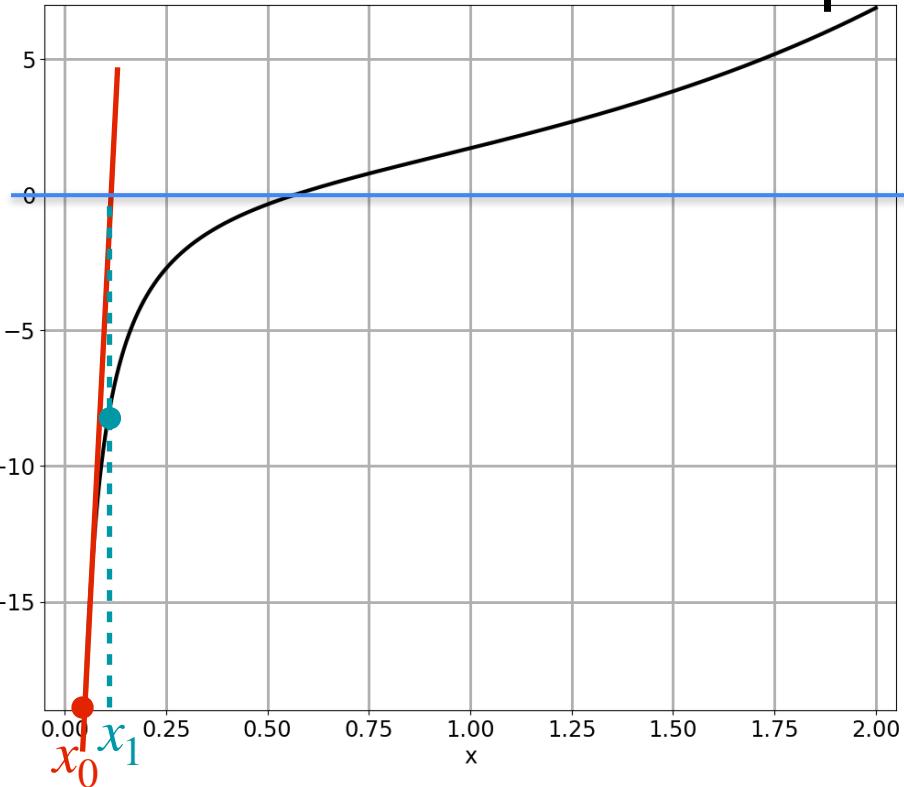


$$g(x) = e^x - \log(x)$$
$$g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$
$$x_0 = 0.05$$
$$f'(x)$$

# Newton's Method for Optimization

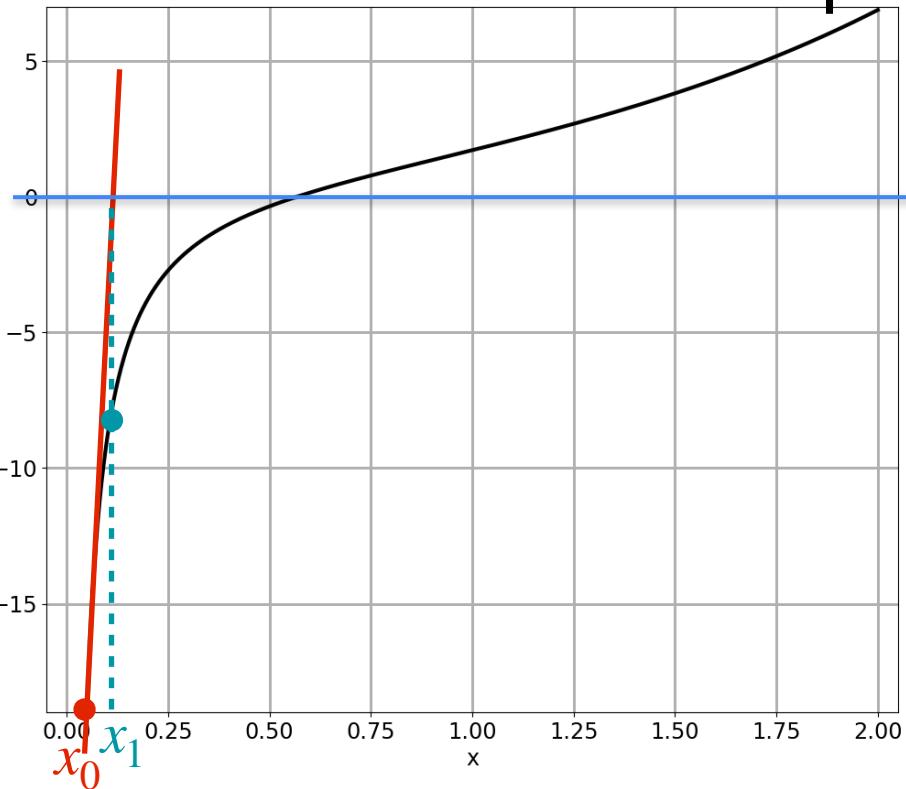


$$g(x) = e^x - \log(x)$$
$$g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$
$$f'(x)$$
$$x_0 = 0.05$$

# Newton's Method for Optimization

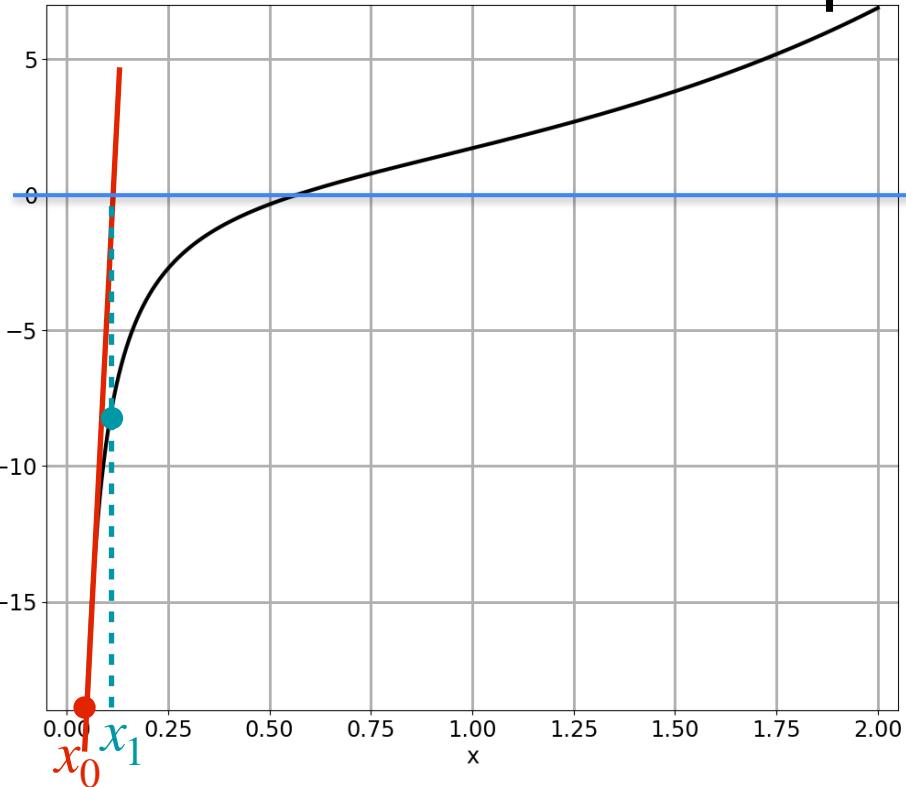


$$g(x) = e^x - \log(x) \quad \overbrace{g'(x)}^{f(x)} = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$
$$x_0 = 0.05$$
$$x_1 = x_0 - \frac{g'(x_0)}{(g'(x_0))'}$$
$$= 0.05 - \frac{\left(e^{0.05} - \frac{1}{0.05}\right)}{\left(e^{0.05} + \frac{1}{0.05^2}\right)}$$

# Newton's Method for Optimization

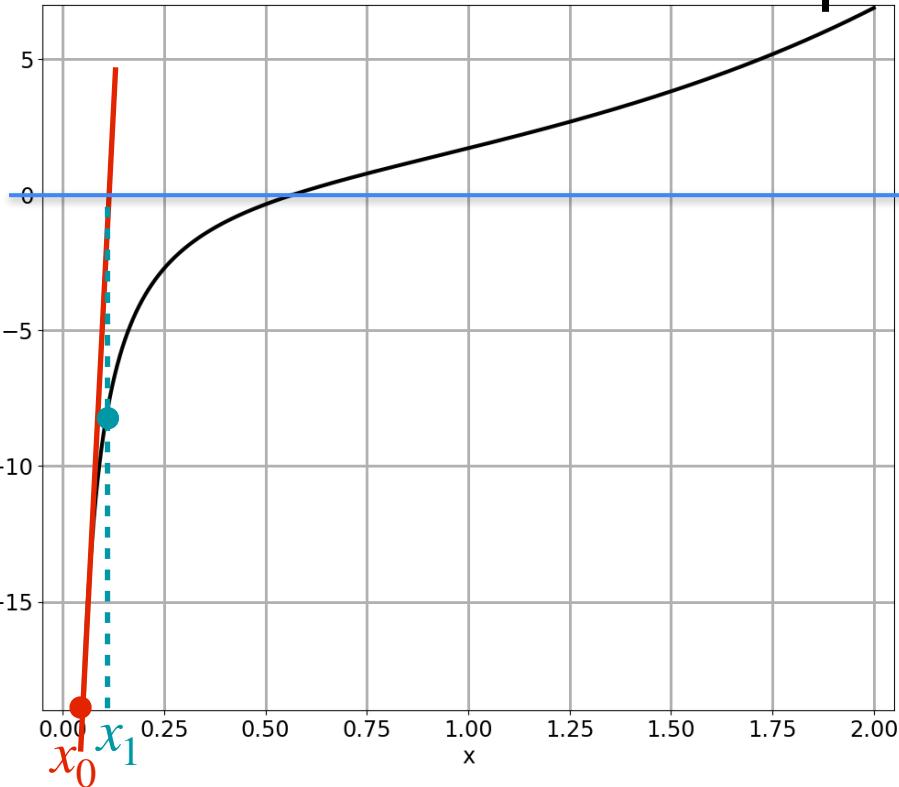


$$g(x) = e^x - \log(x) \quad \overbrace{g'(x)}^{f(x)} = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$
$$x_0 = 0.05$$
$$x_1 = x_0 - \frac{g'(x_0)}{(g'(x_0))'} \quad \overbrace{f'(x)}^{g'(x)}$$
$$= 0.05 - \frac{\left(e^{0.05} - \frac{1}{0.05}\right)}{\left(e^{0.05} + \frac{1}{0.05^2}\right)} = 0.097$$

# Newton's Method for Optimization

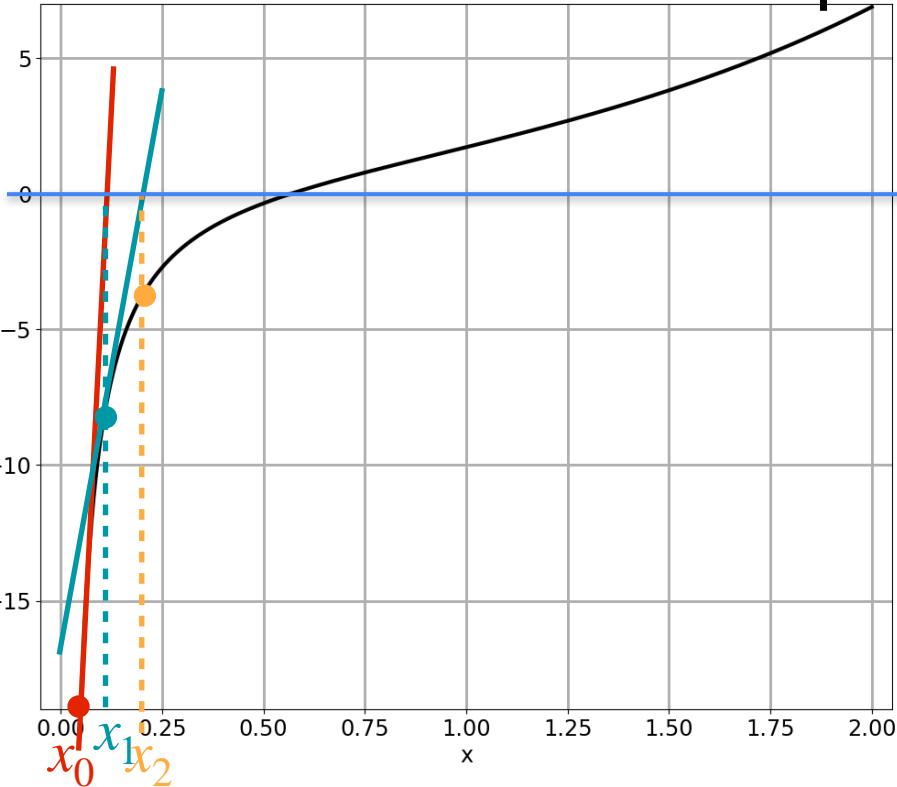


$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

# Newton's Method for Optimization



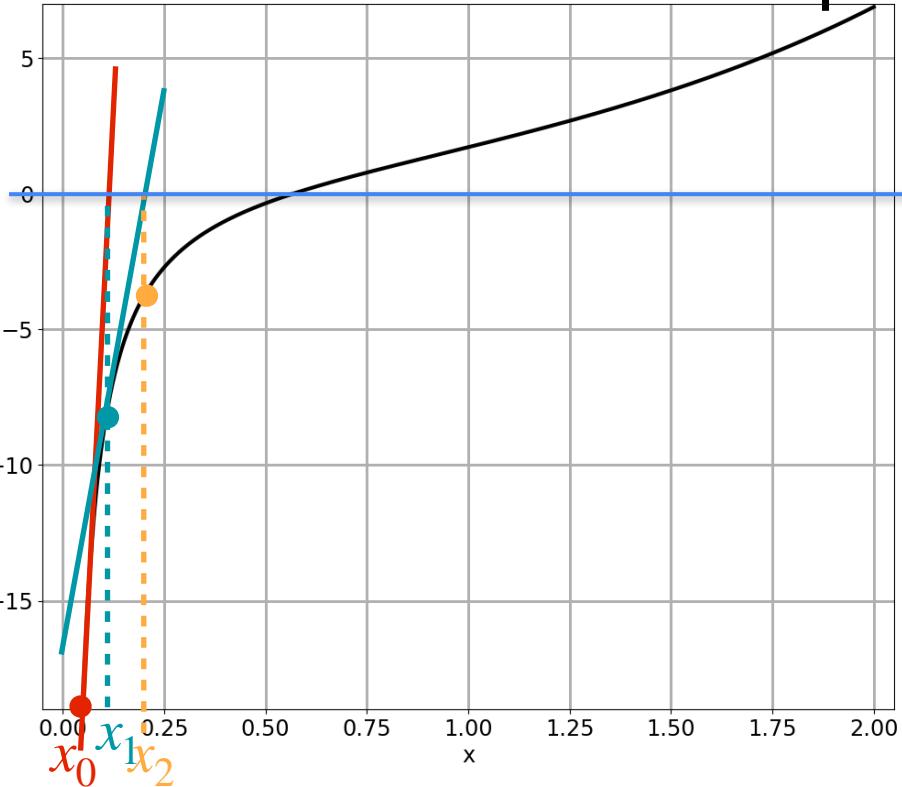
$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_1 = 0.097$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

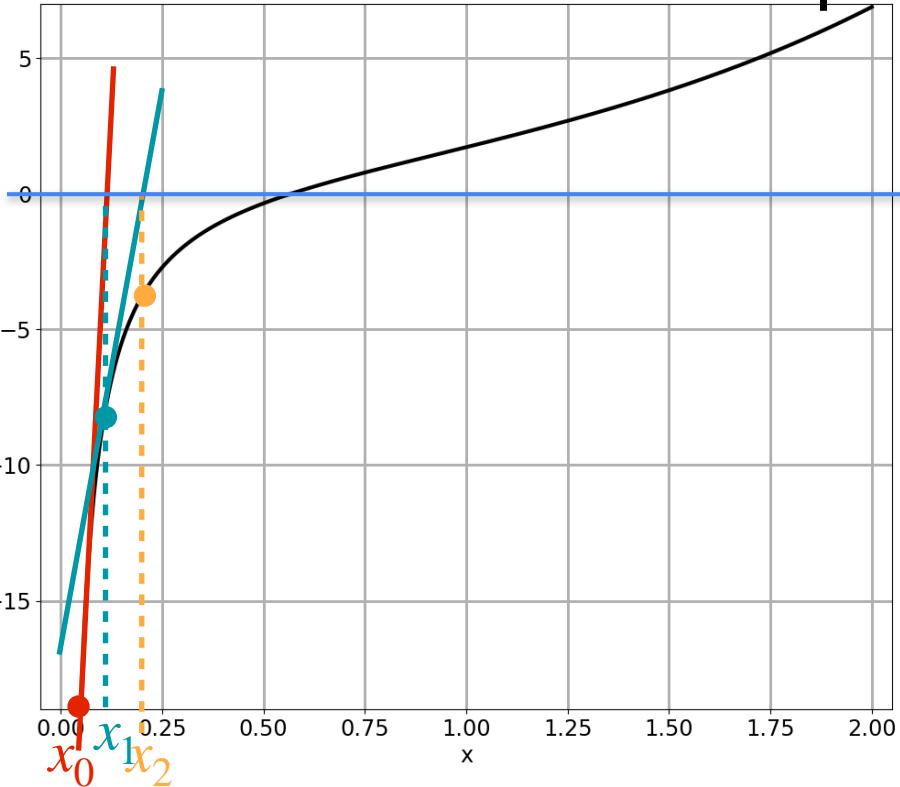
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_1 = 0.097$$

$$x_2 = x_1 - \frac{g'(x_1)}{(g'(x_1))'}$$

$$= 0.097 - \frac{\left(e^{0.097} - \frac{1}{0.097}\right)}{\left(e^{0.097} + \frac{1}{0.097^2}\right)} = 0.183$$

# Newton's Method for Optimization

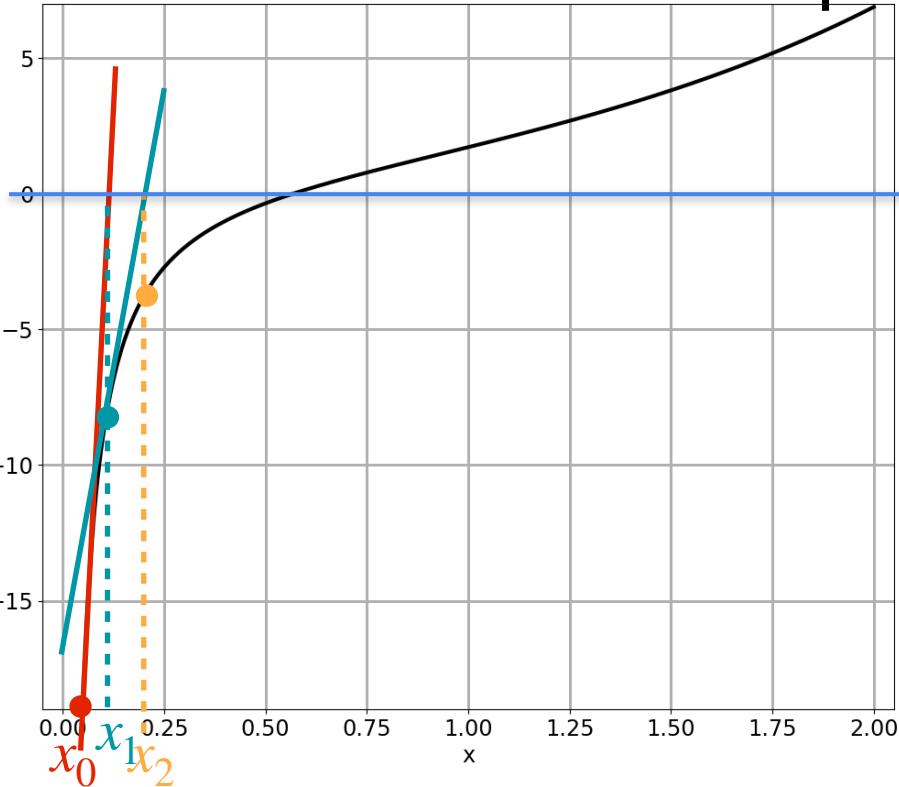


$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

# Newton's Method for Optimization



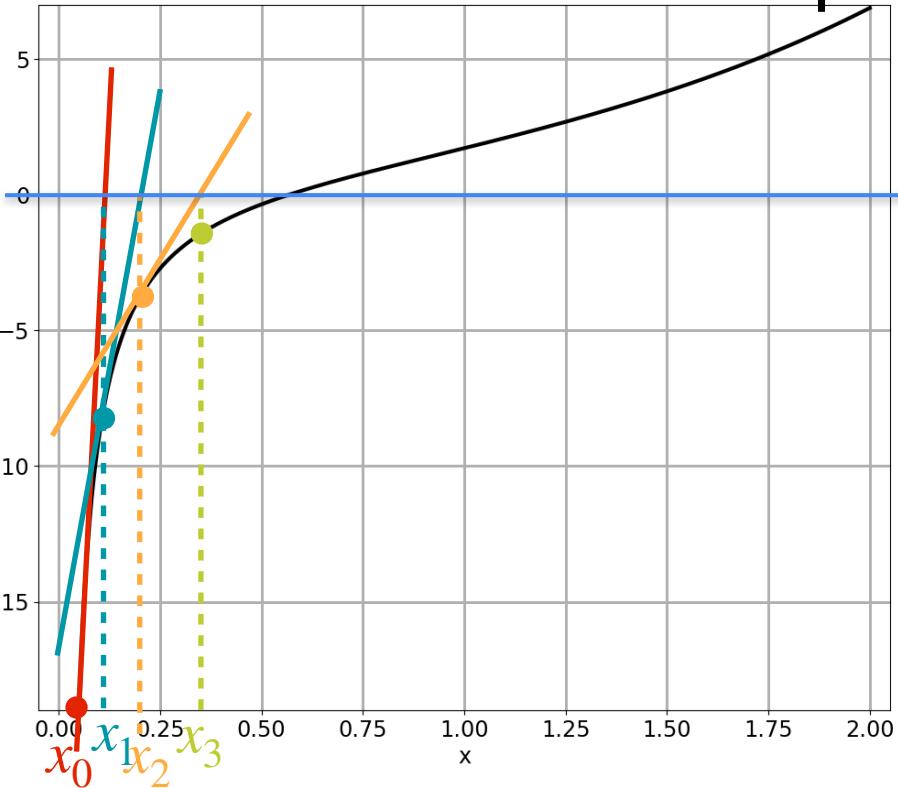
$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_2 = 0.183$$

# Newton's Method for Optimization



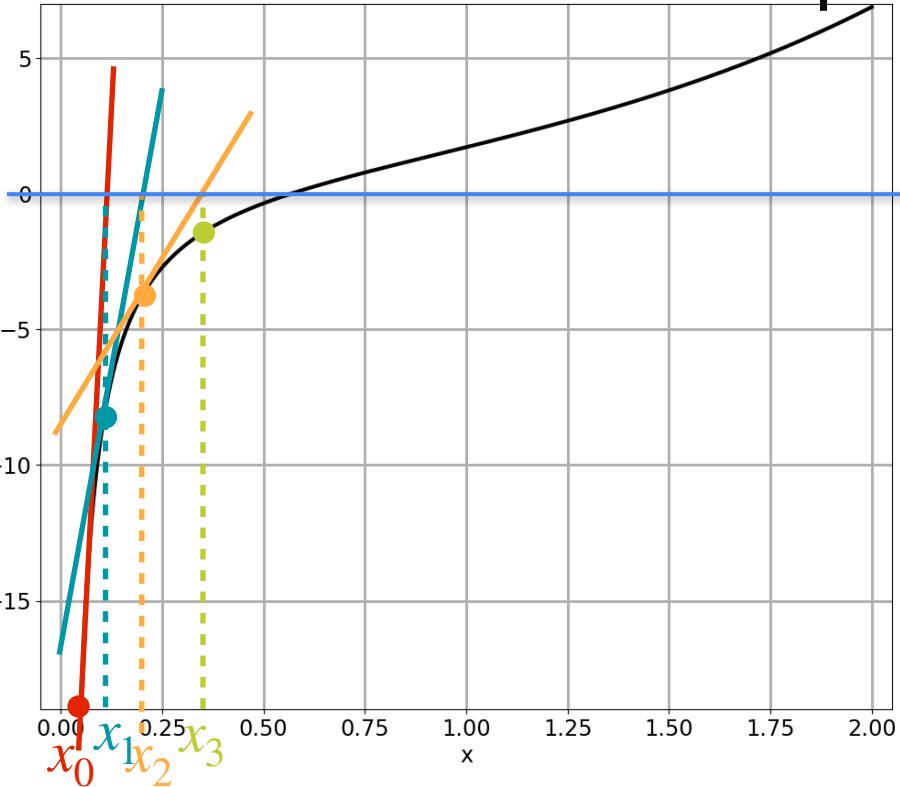
$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_2 = 0.183$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

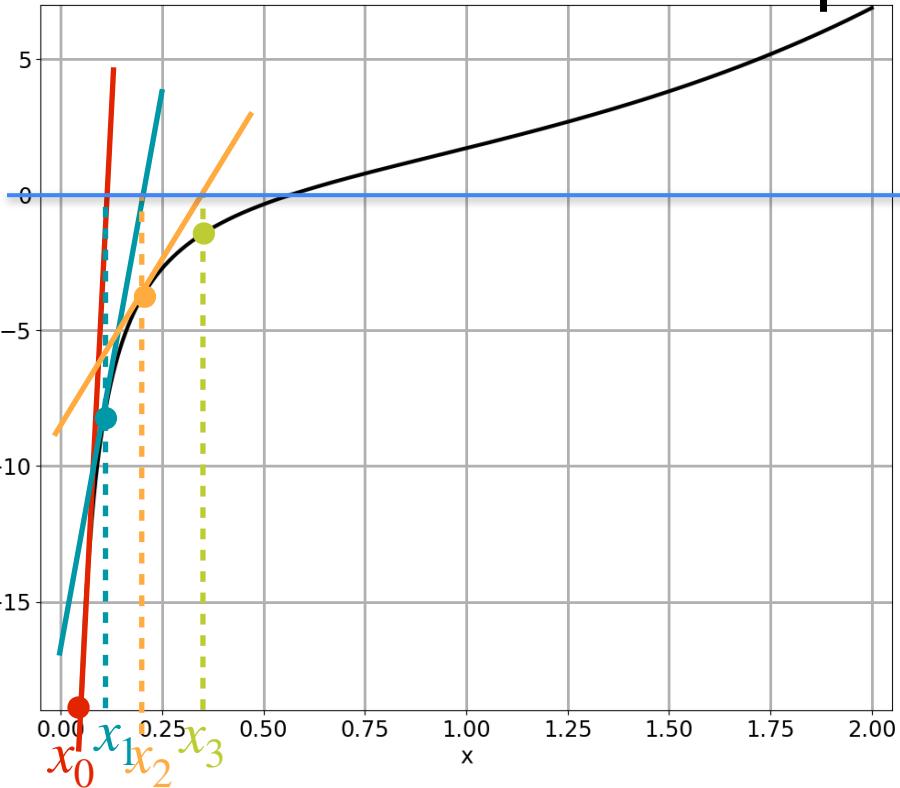
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_2 = 0.183$$

$$x_3 = x_2 - \frac{g'(x_2)}{(g'(x_2))'}$$

$$= 0.183 - \frac{\left(e^{0.183} - \frac{1}{0.183}\right)}{\left(e^{0.183} + \frac{1}{0.183^2}\right)}$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

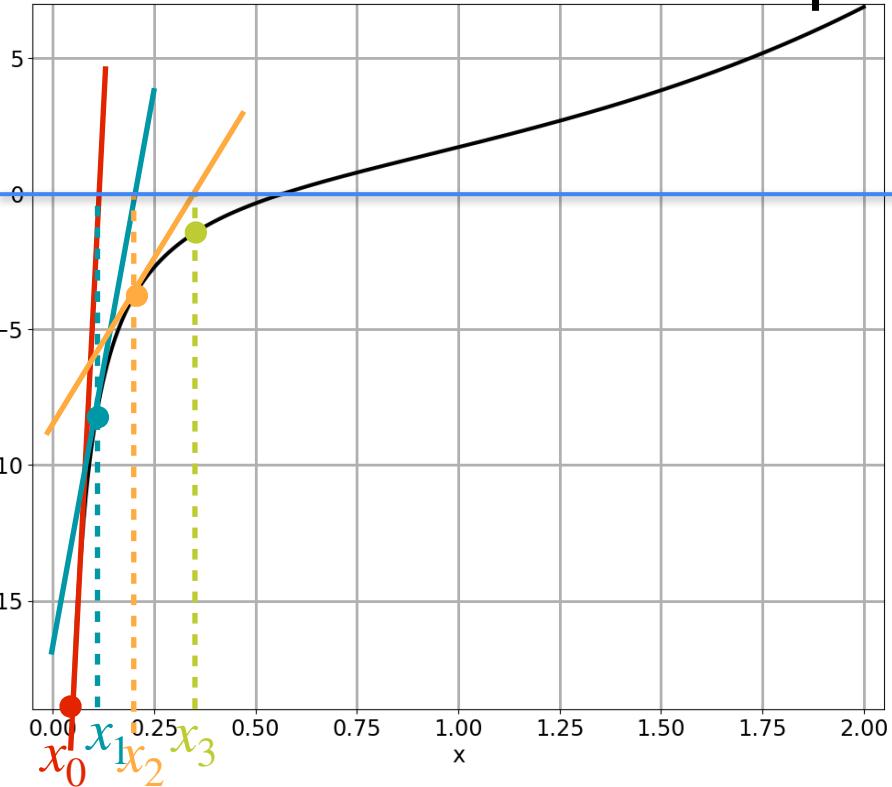
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_2 = 0.183$$

$$x_3 = x_2 - \frac{g'(x_2)}{(g'(x_2))'}$$

$$= 0.183 - \frac{\left(e^{0.183} - \frac{1}{0.183}\right)}{\left(e^{0.183} + \frac{1}{0.183^2}\right)} = 0.320$$

# Newton's Method for Optimization



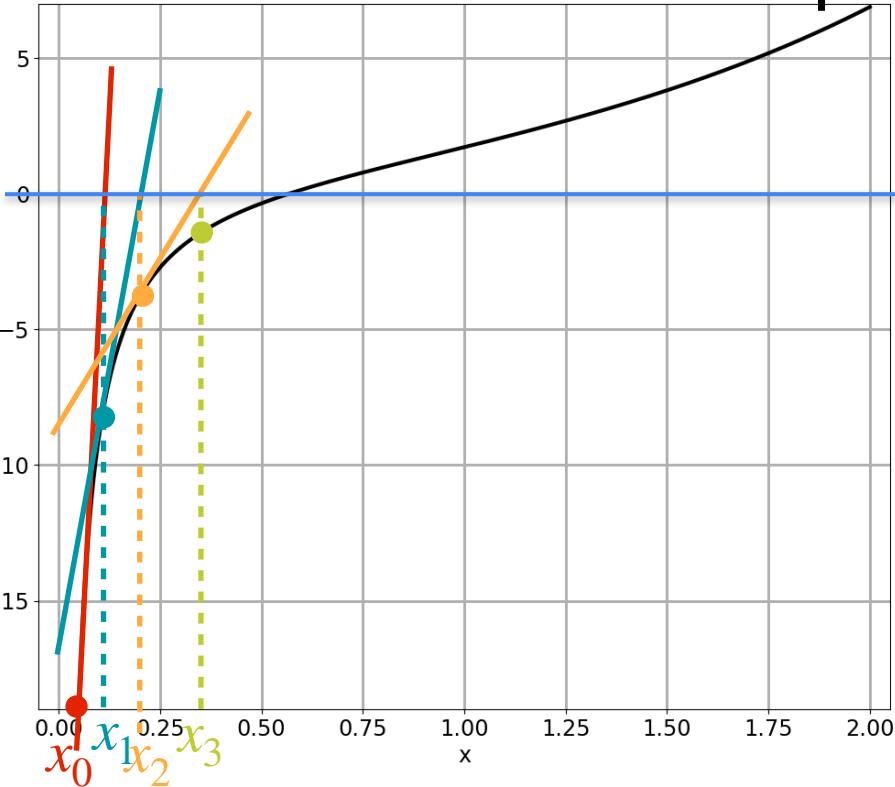
$$g(x) = e^x - \log(x)$$

$$g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

# Newton's Method for Optimization



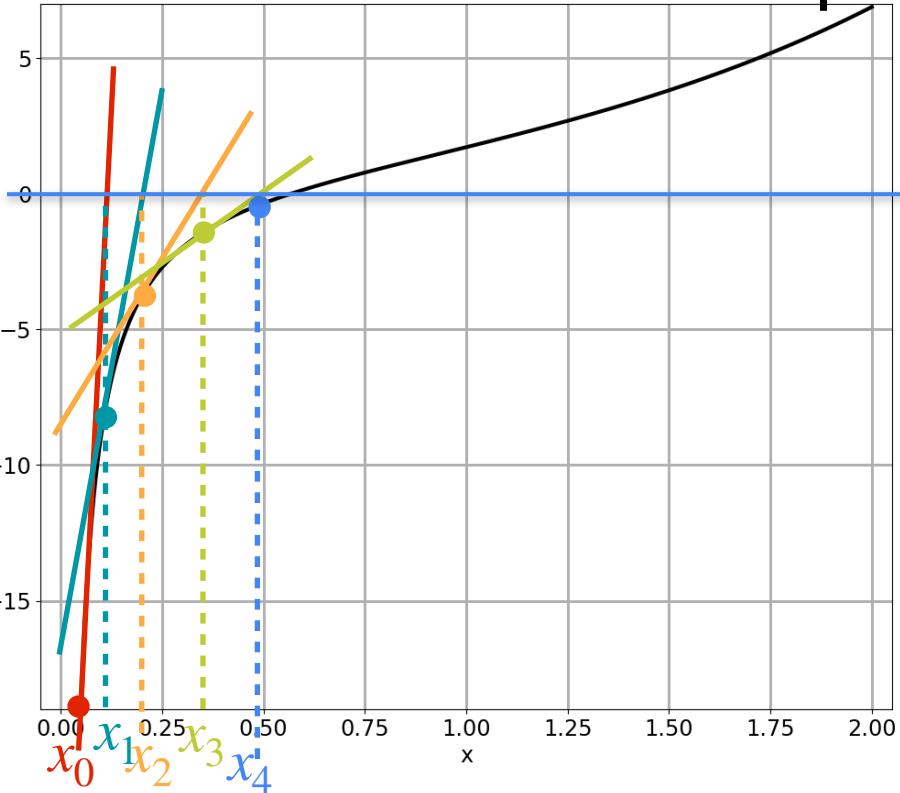
$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_3 = 0.320$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

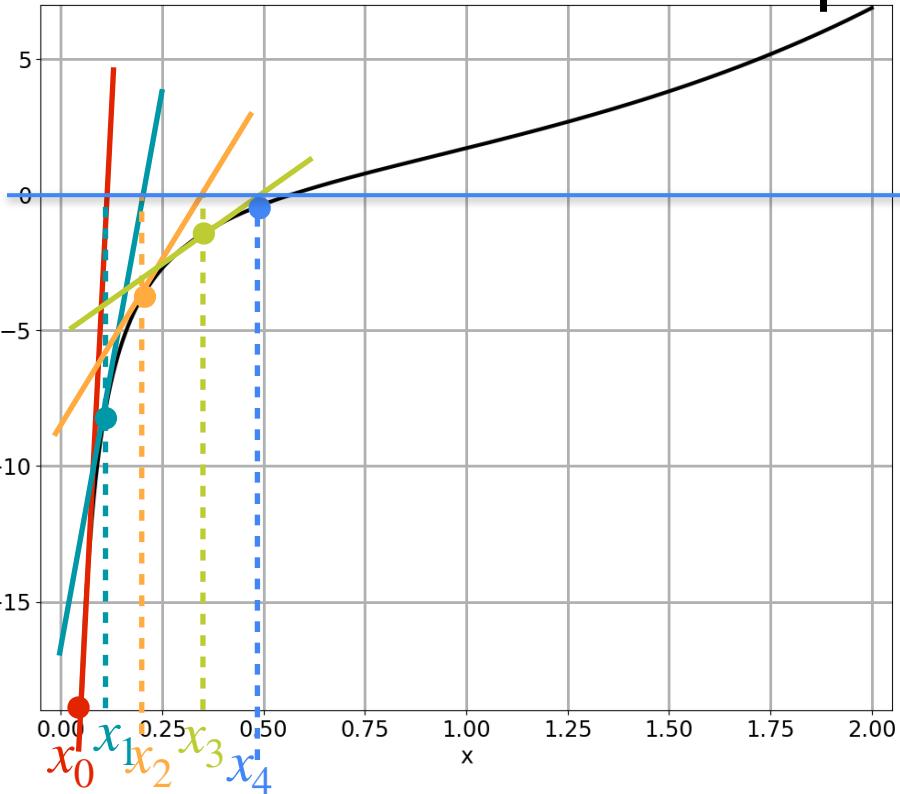
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_3 = 0.320$$

$$x_4 = x_3 - \frac{g'(x_3)}{(g'(x_3))'}$$

$$= 0.320 - \frac{\left(e^{0.320} - \frac{1}{0.320}\right)}{\left(e^{0.320} + \frac{1}{0.320^2}\right)}$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

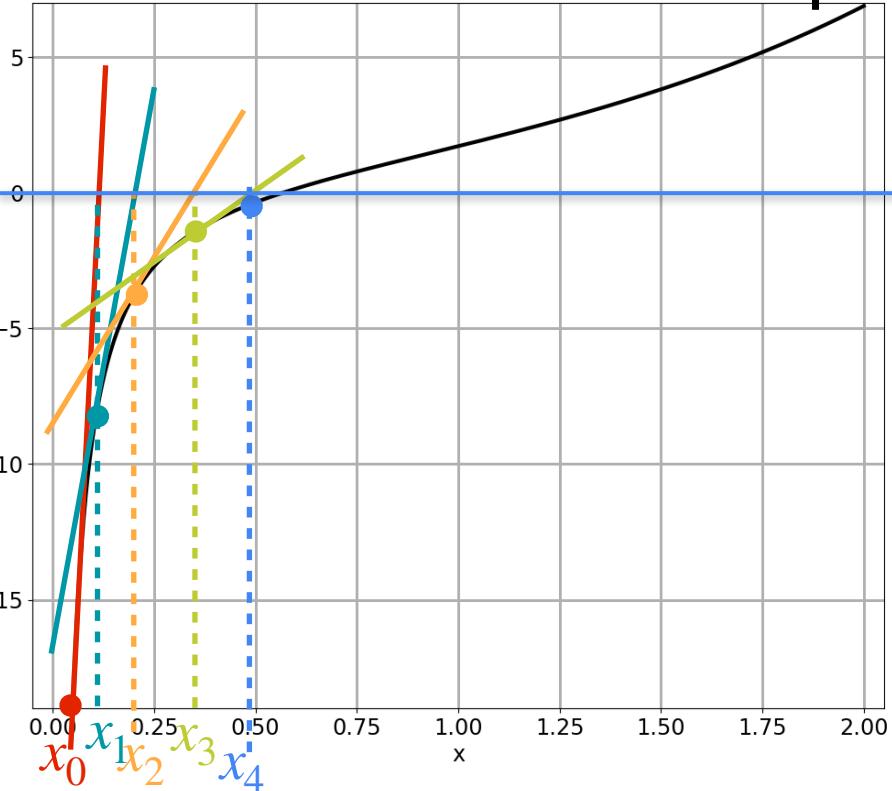
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_3 = 0.320$$

$$x_4 = x_3 - \frac{g'(x_3)}{(g'(x_3))'}$$

$$= 0.320 - \frac{\left(e^{0.320} - \frac{1}{0.320}\right)}{\left(e^{0.320} + \frac{1}{0.320^2}\right)} = 0.477$$

# Newton's Method for Optimization



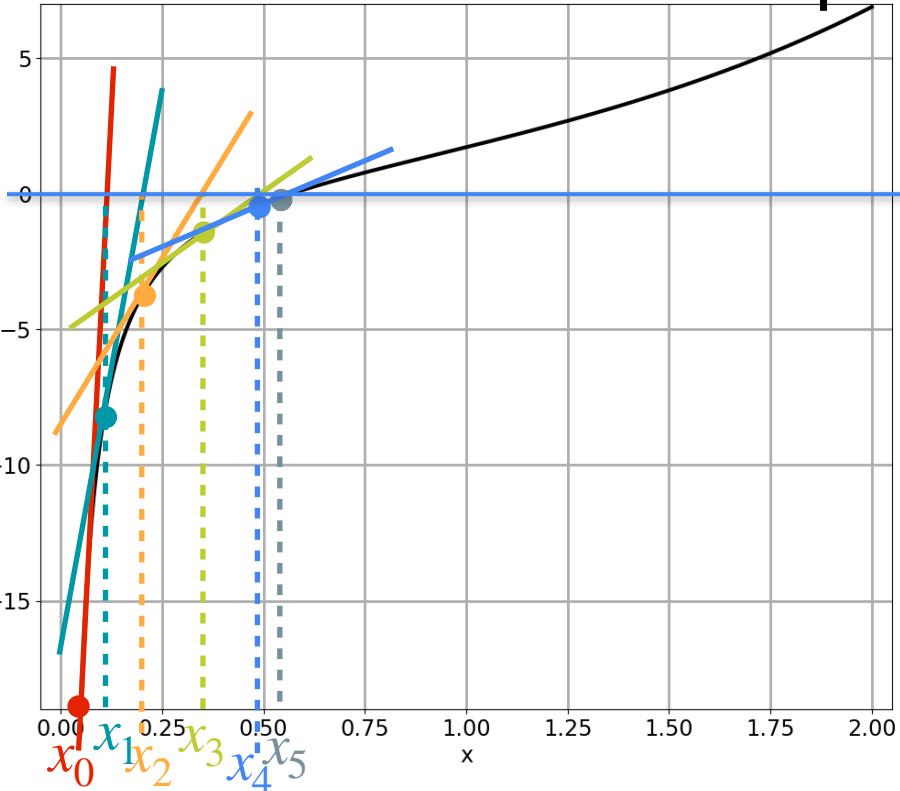
$$g(x) = e^x - \log(x)$$

$$g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.5671$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.5671$$

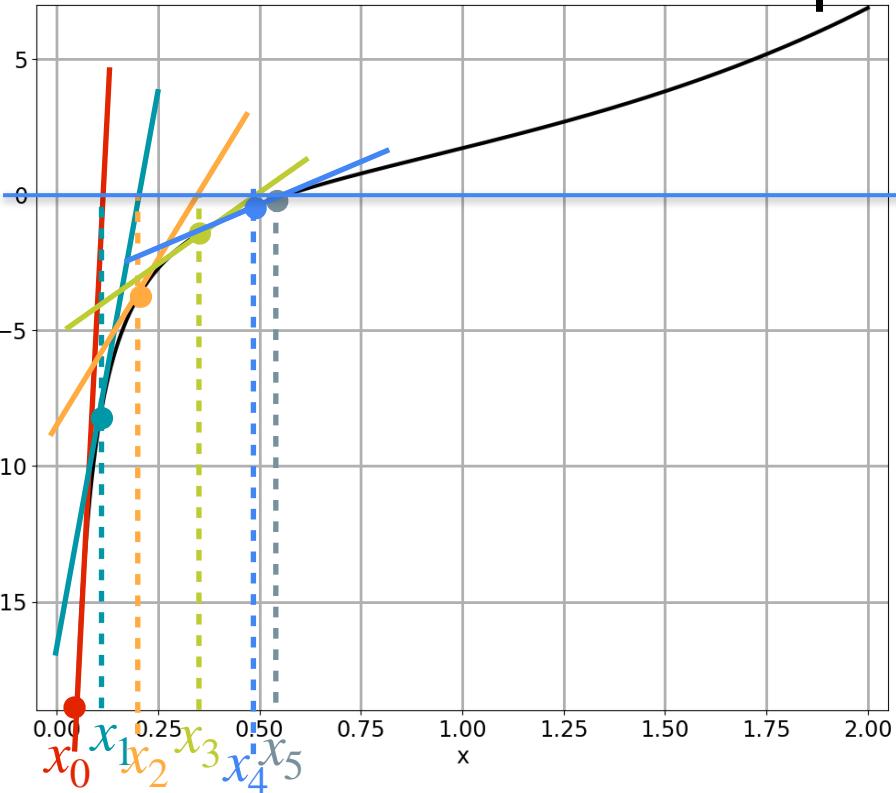
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_4 = 0.477$$

$$x_5 = x_4 - \frac{g'(x_4)}{(g'(x_4))'}$$

$$= 0.447 - \frac{\left(e^{0.447} - \frac{1}{0.447}\right)}{\left(e^{0.447} + \frac{1}{0.447^2}\right)} = 0.558$$

# Newton's Method for Optimization

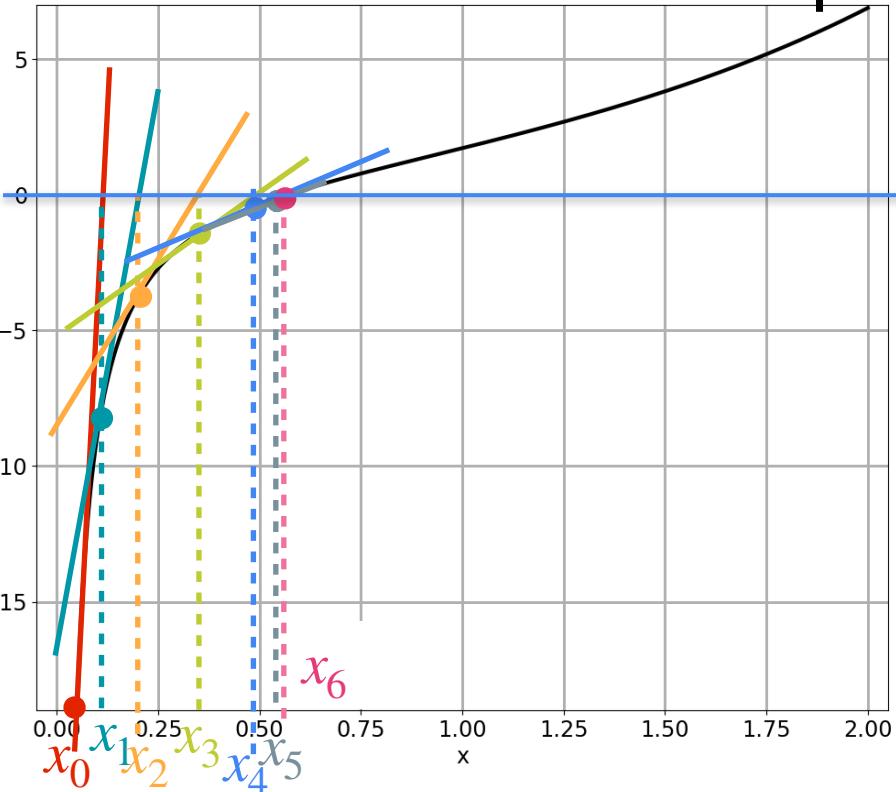


$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.567$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

# Newton's Method for Optimization

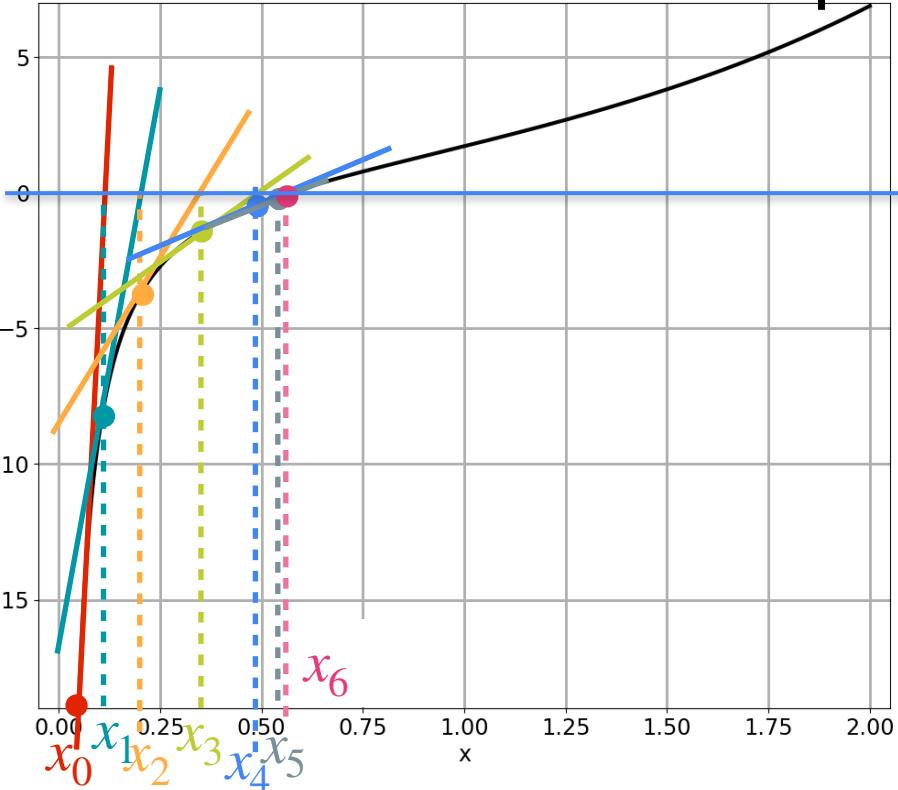


$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.567$

$$(g'(x))' = e^x + \frac{1}{x^2}$$
$$x_5 = 0.558$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

$$\text{Minimum: } x^* = 0.567$$

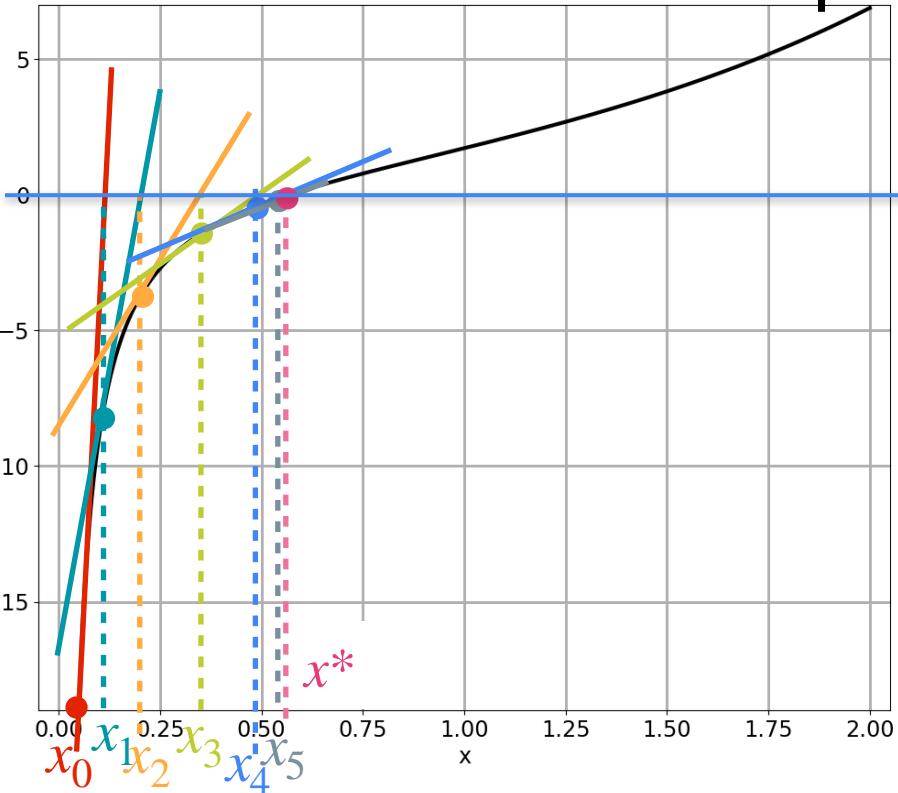
$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_5 = 0.558$$

$$x_6 = x_5 - \frac{g'(x_5)}{(g'(x_5))'}$$

$$= 0.558 - \frac{\left(e^{0.558} - \frac{1}{0.558}\right)}{\left(e^{0.558} + \frac{1}{0.558^2}\right)}$$

# Newton's Method for Optimization



$$g(x) = e^x - \log(x) \quad g'(x) = e^x - 1/x$$

Minimum:  $x^* = 0.567$

$$(g'(x))' = e^x + \frac{1}{x^2}$$

$$x_5 = 0.558$$

$$x^* = x_5 - \frac{g'(x_5)}{(g'(x_5))'}$$

$$= 0.558 - \frac{\left(e^{0.558} - \frac{1}{0.558}\right)}{\left(e^{0.558} + \frac{1}{0.558^2}\right)} = 0.567$$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## The second derivative

# Second Derivative

# Second Derivative

Newton's method:

# Second Derivative

Newton's method:  $x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$

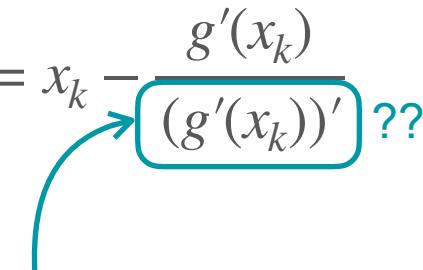
# Second Derivative

Newton's method:  $x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'} ??$

# Second Derivative

Newton's method:  $x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$  ??

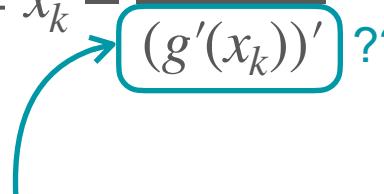
Second derivative



# Second Derivative

Newton's method:  $x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$  ??

Second derivative



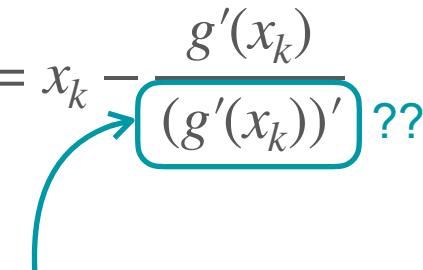
Leibniz notation:

$$\frac{d^2f(x)}{dx^2} = \frac{d}{dx} \left( \frac{df(x)}{dx} \right)$$

# Second Derivative

Newton's method:  $x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$  ??

Second derivative



Leibniz notation:

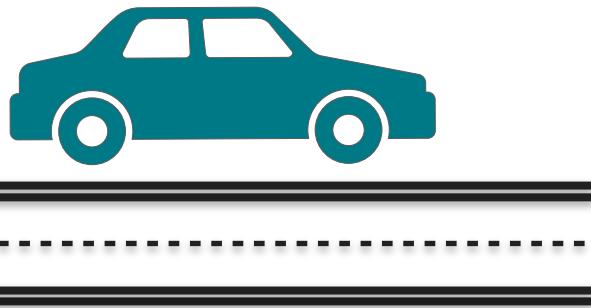
$$\frac{d^2f(x)}{dx^2} = \frac{d}{dx} \left( \frac{df(x)}{dx} \right)$$

Lagrange notation:

$$f''(x)$$

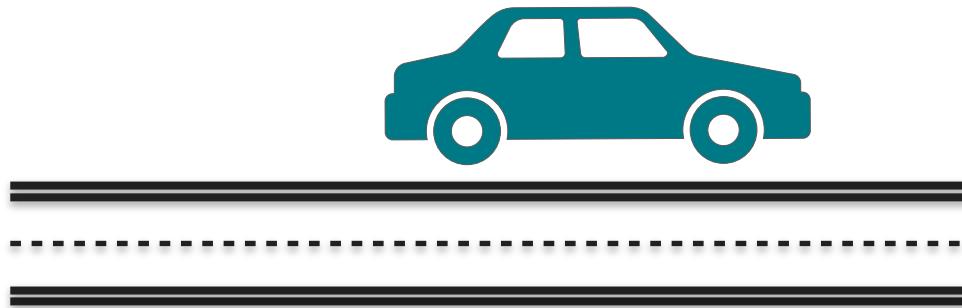
# Understanding Second Derivative

# Understanding Second Derivative

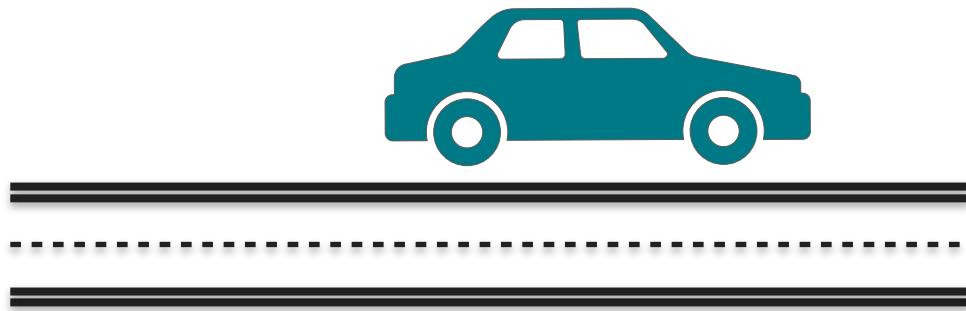


# Understanding Second Derivative

$x$  Distance



# Understanding Second Derivative



$x$

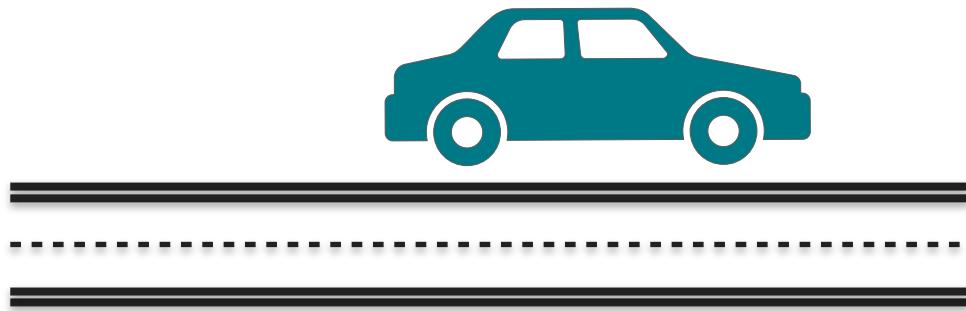
Distance

$v$

Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative



$x$  Distance

$v$  Velocity

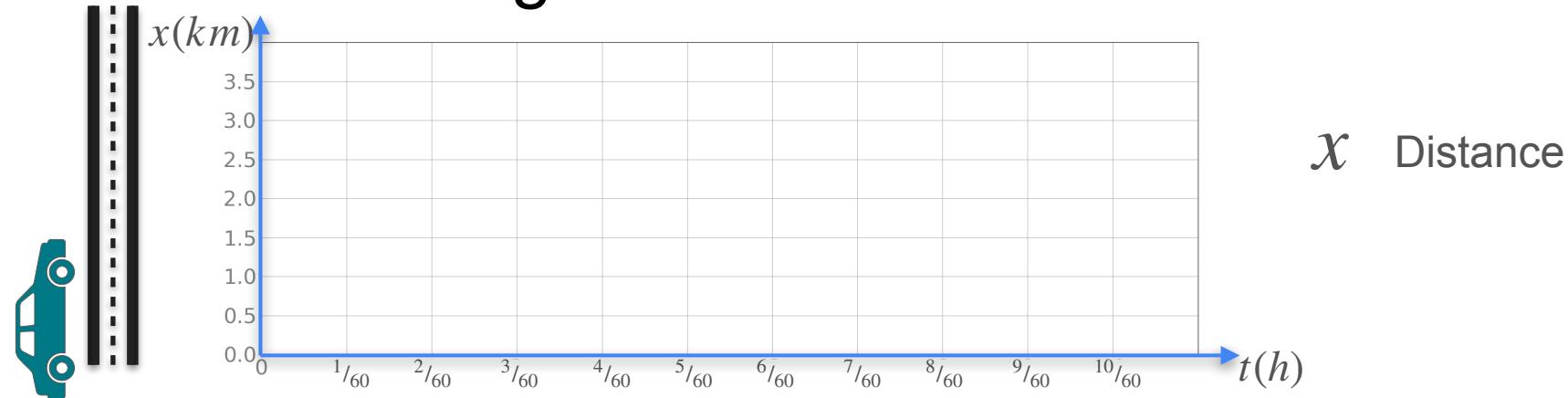
$a$  Acceleration

$$\frac{dx}{dt}$$

$$\frac{dv}{dt} = \frac{d^2x}{dt^2}$$

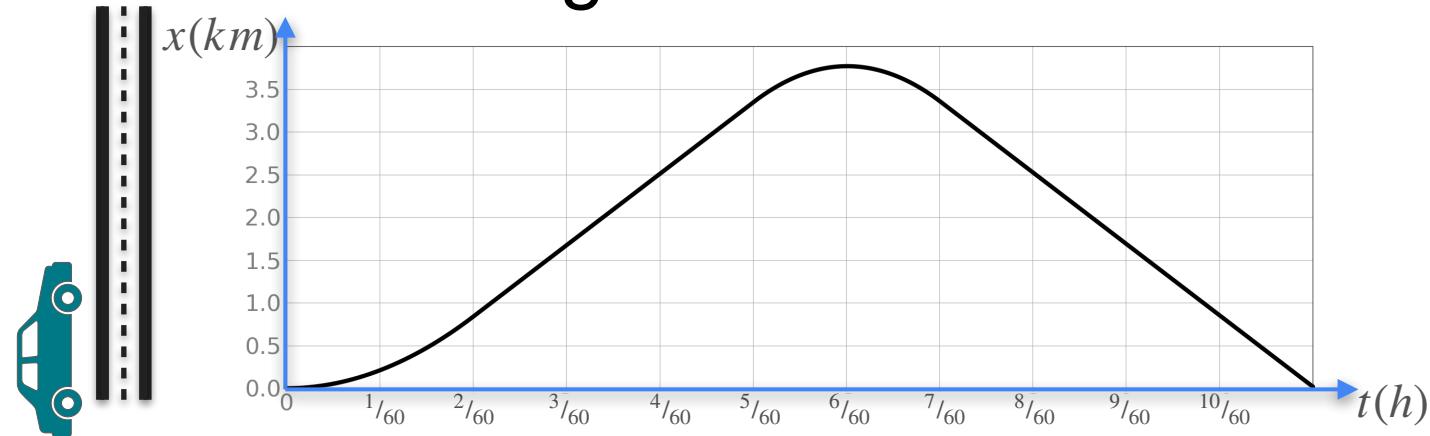
# Understanding Second Derivative

# Understanding Second Derivative

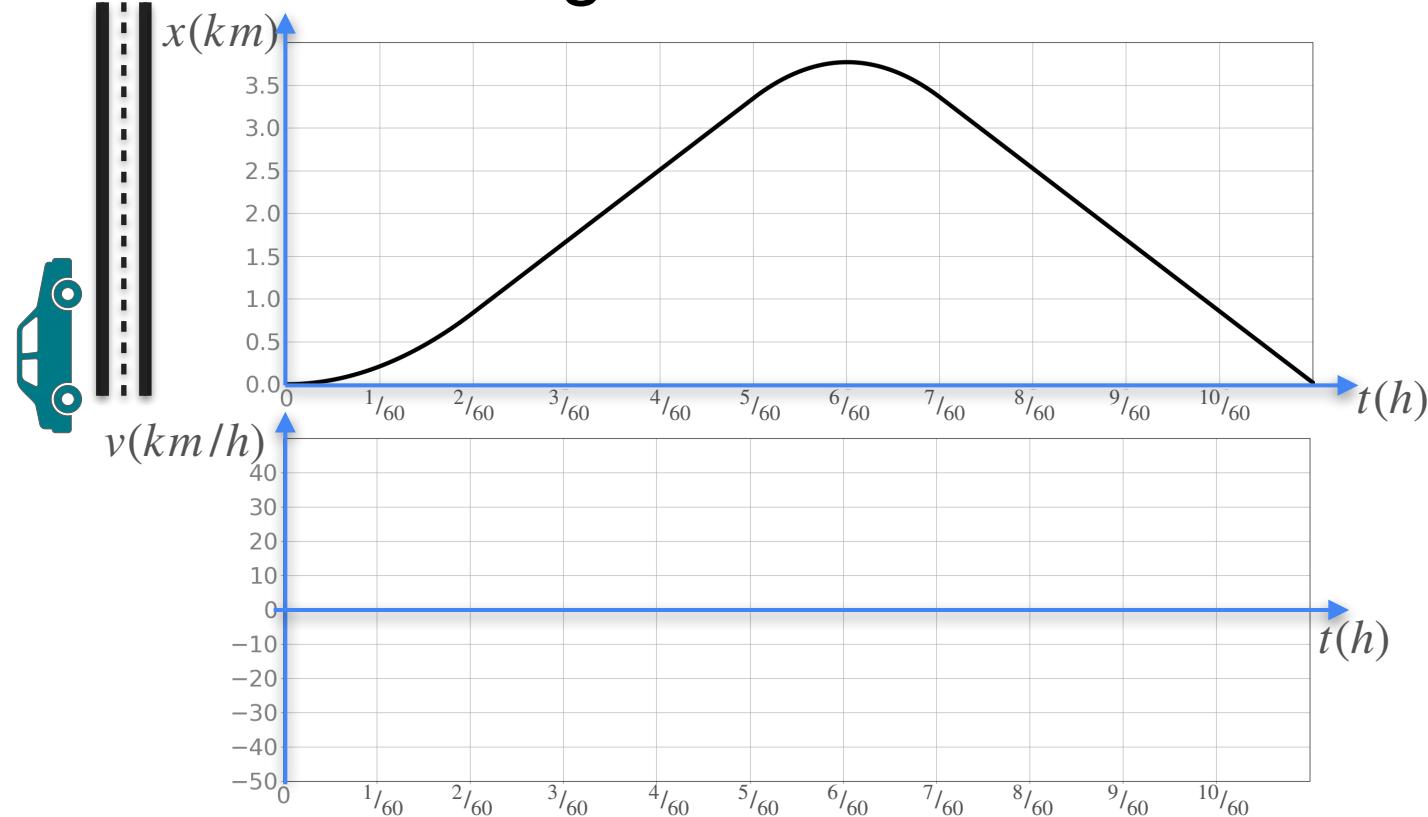


$\mathcal{X}$  Distance

# Understanding Second Derivative



# Understanding Second Derivative

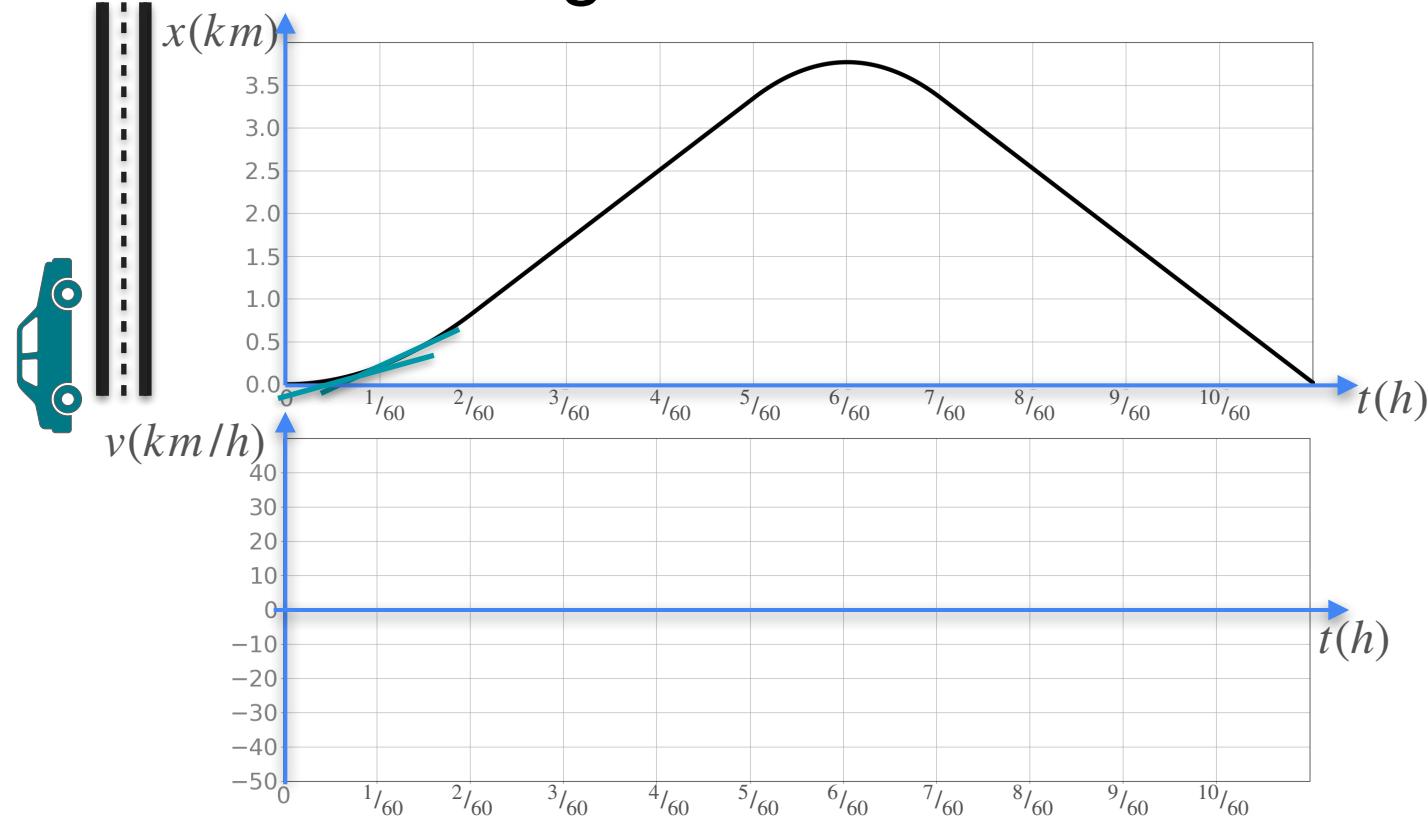


$\mathcal{X}$  Distance

$\mathcal{V}$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

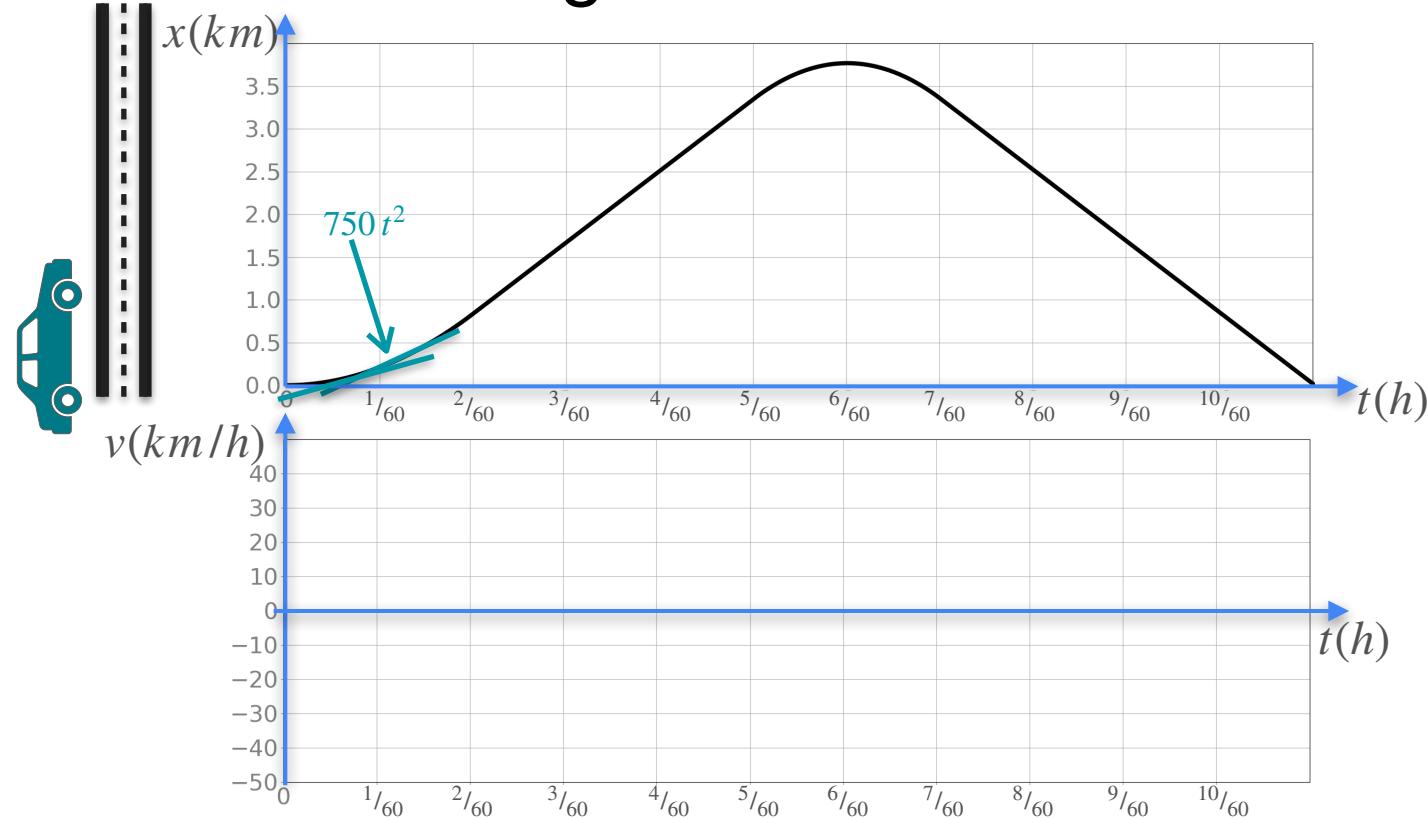


$\mathcal{X}$  Distance

$\mathcal{V}$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

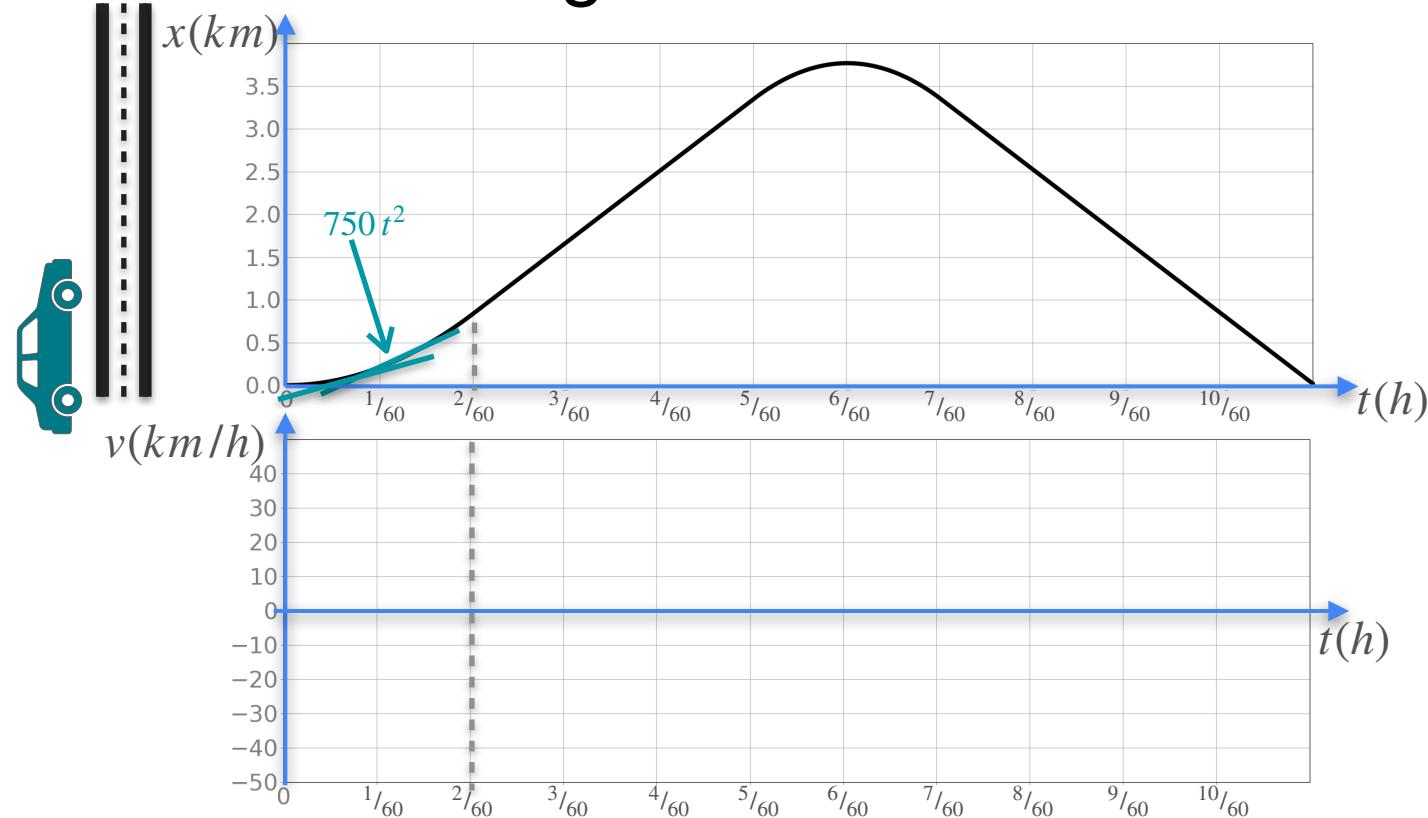


$\mathcal{X}$  Distance

$\mathcal{V}$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

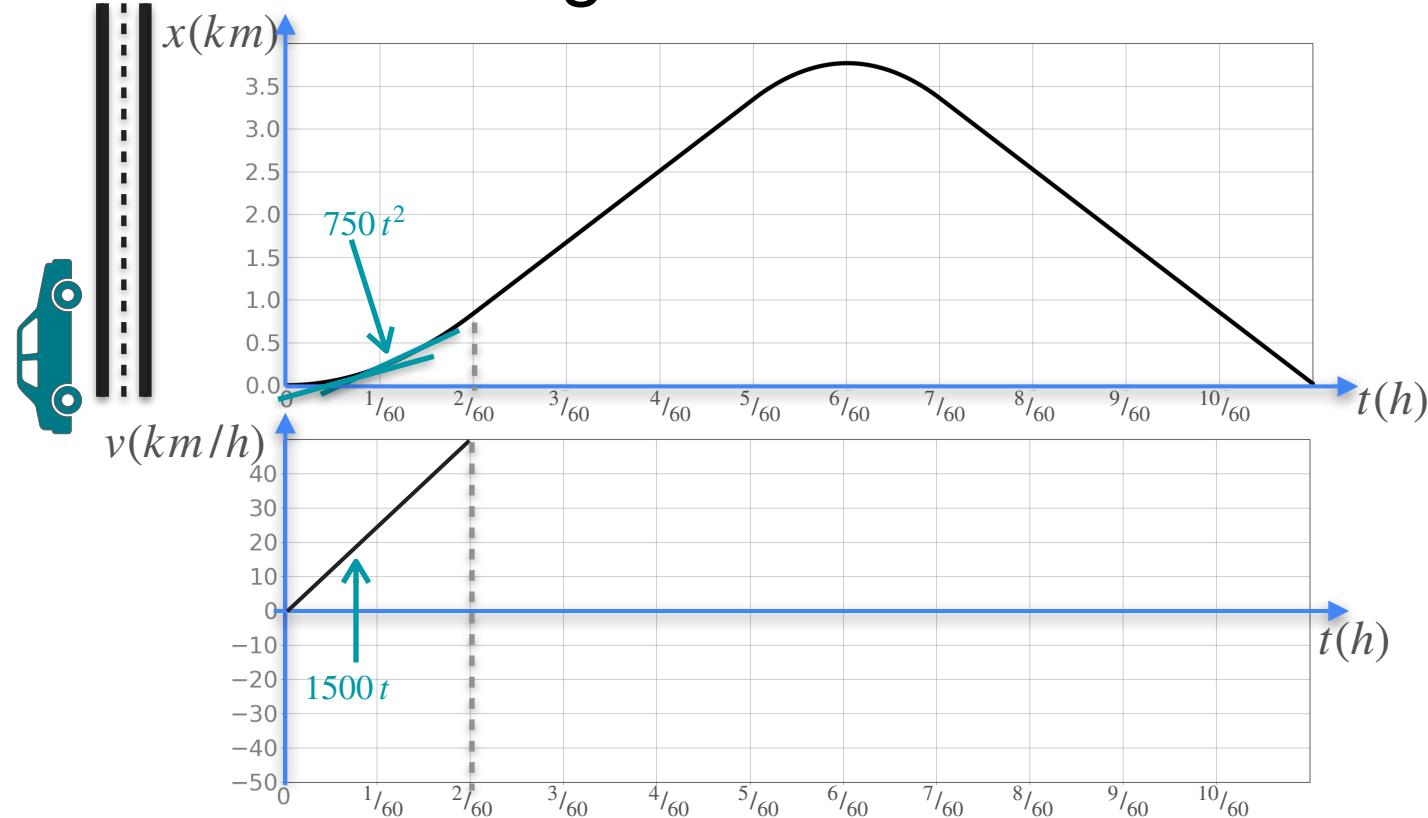


$\mathcal{X}$  Distance

$\mathcal{V}$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

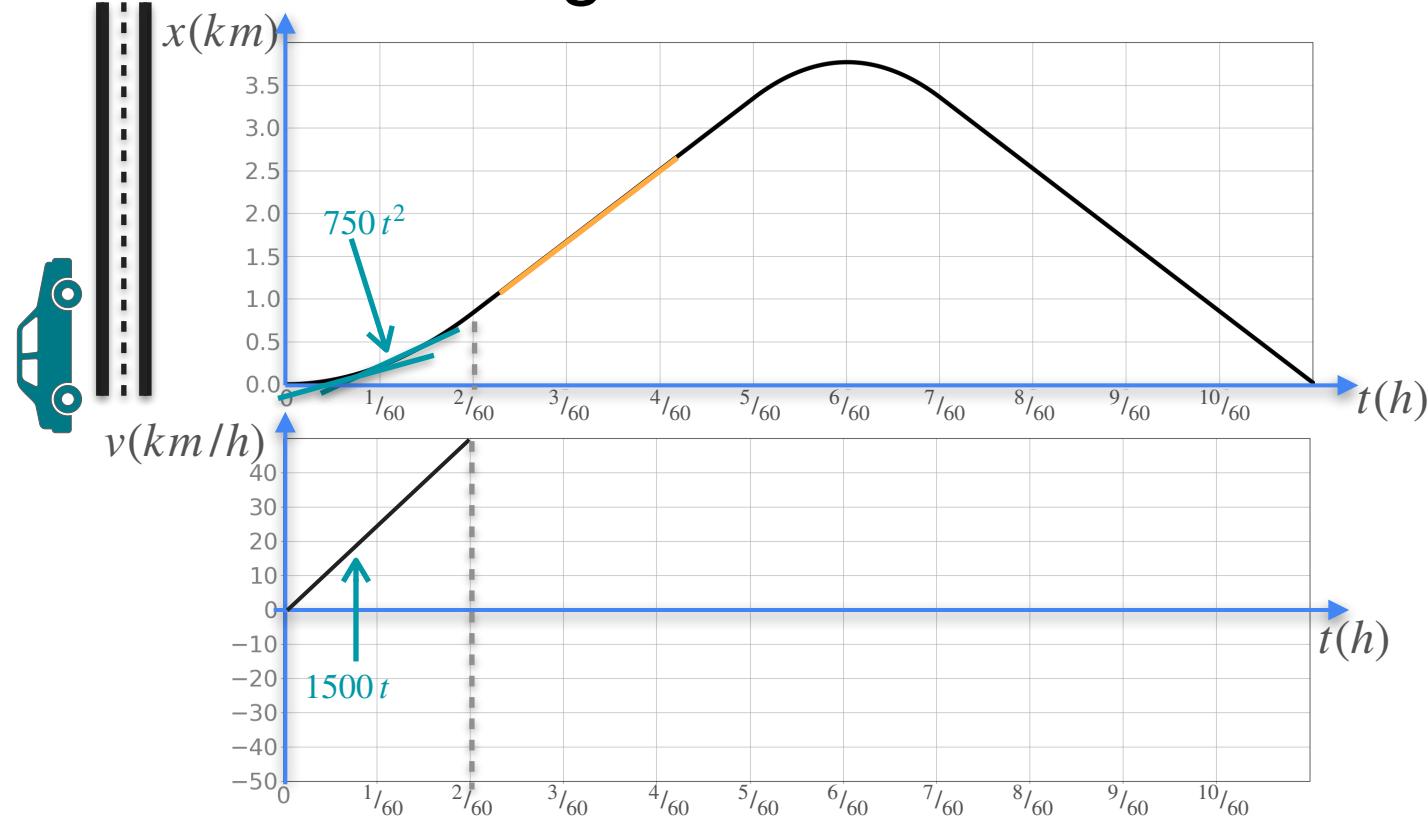


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

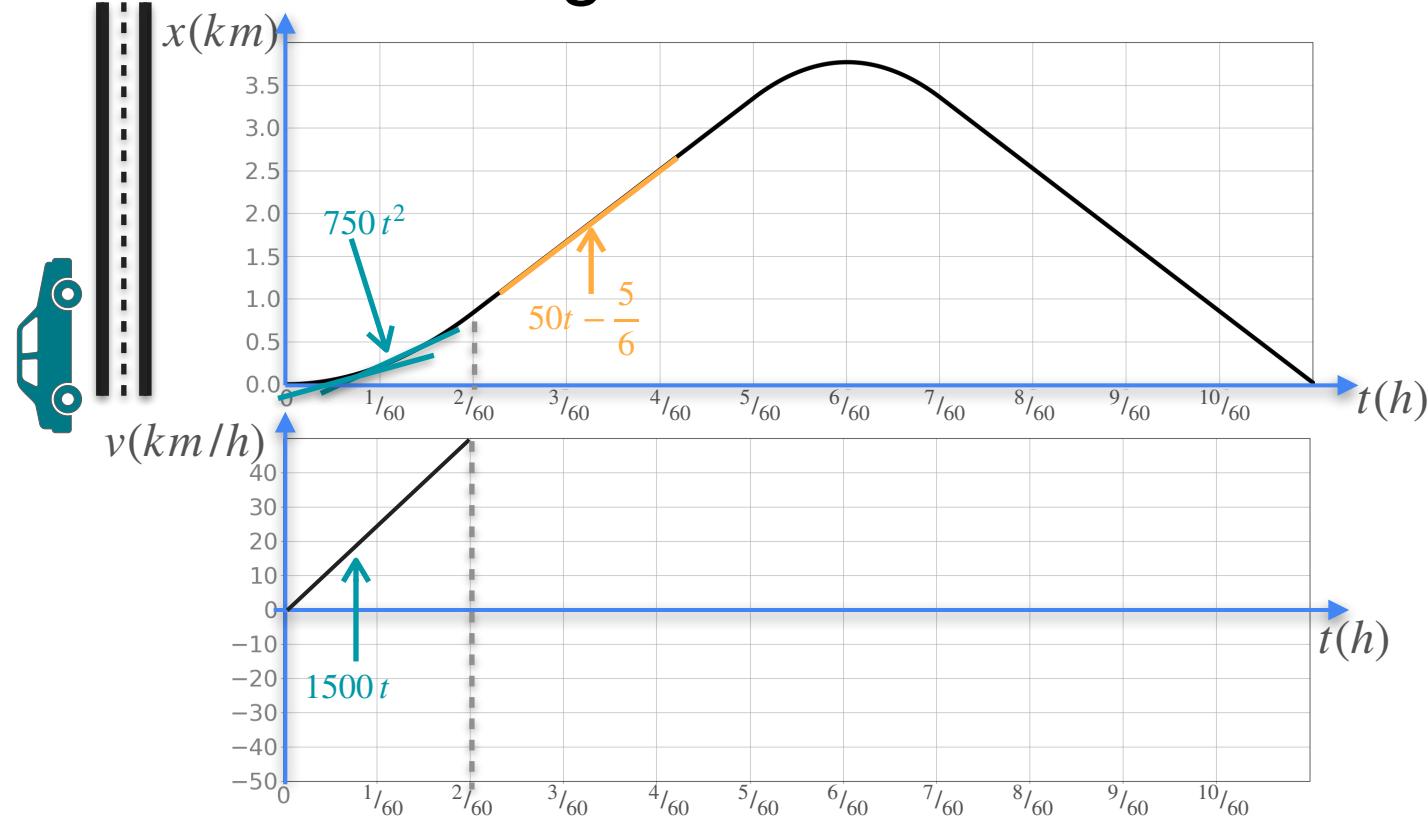


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

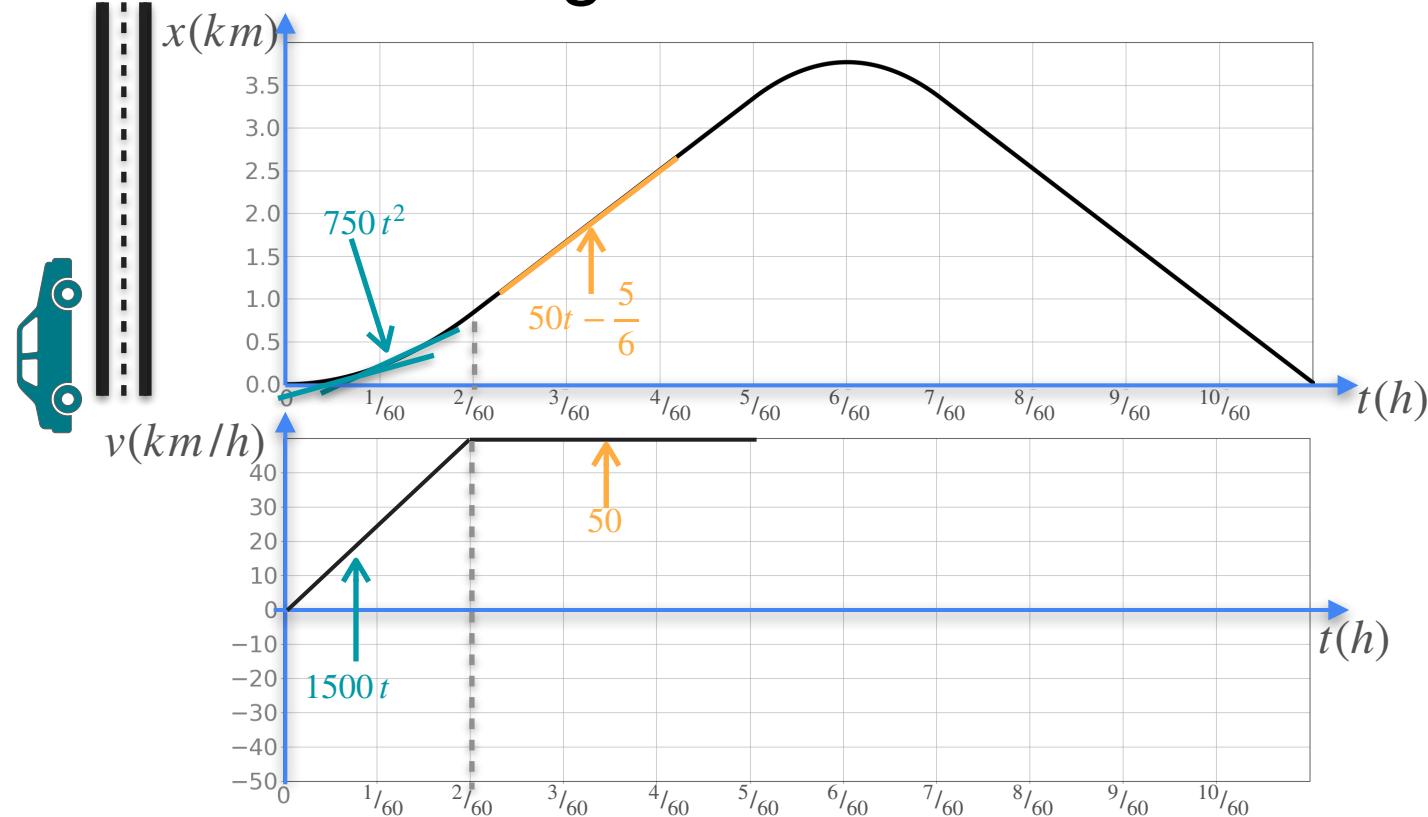


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

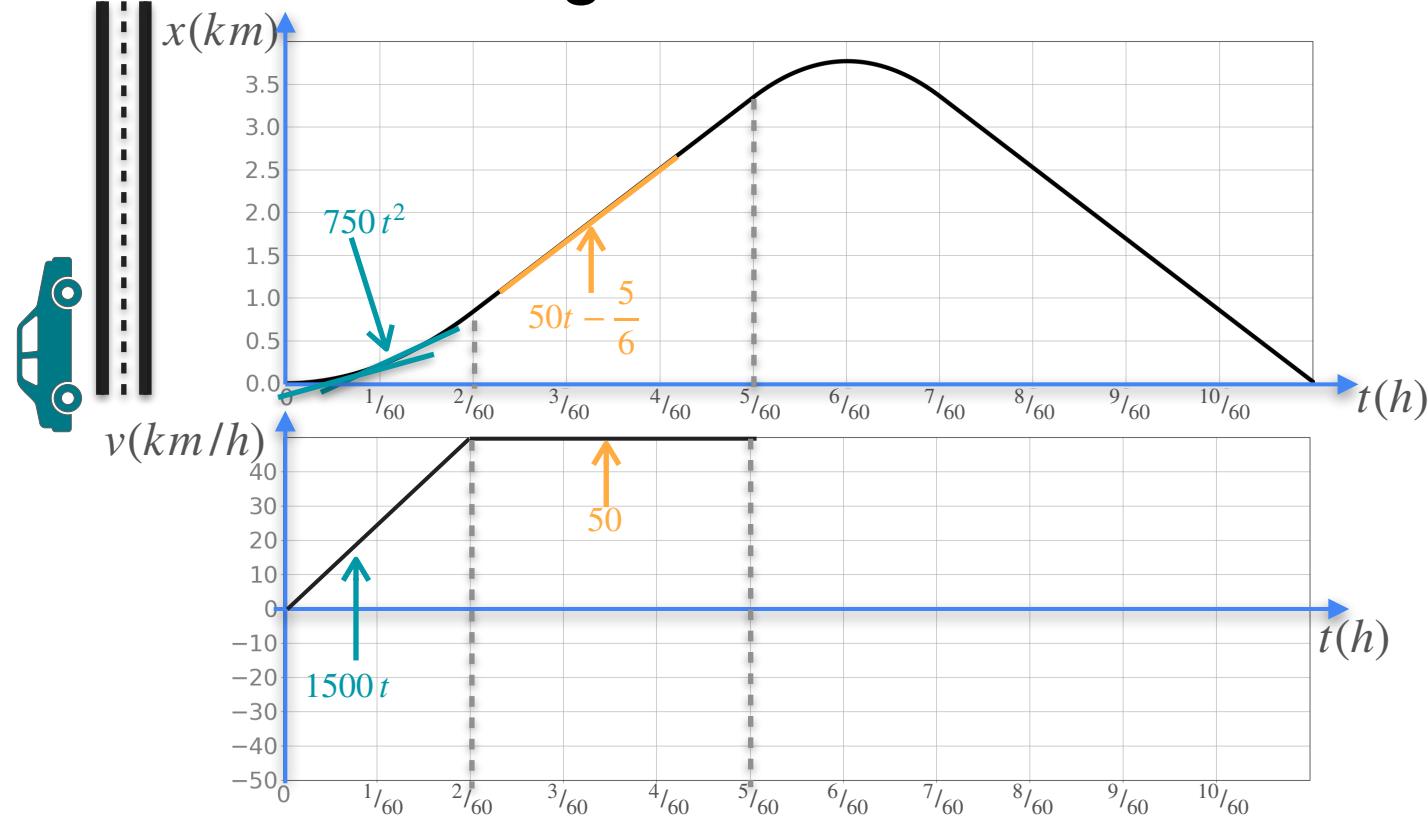


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

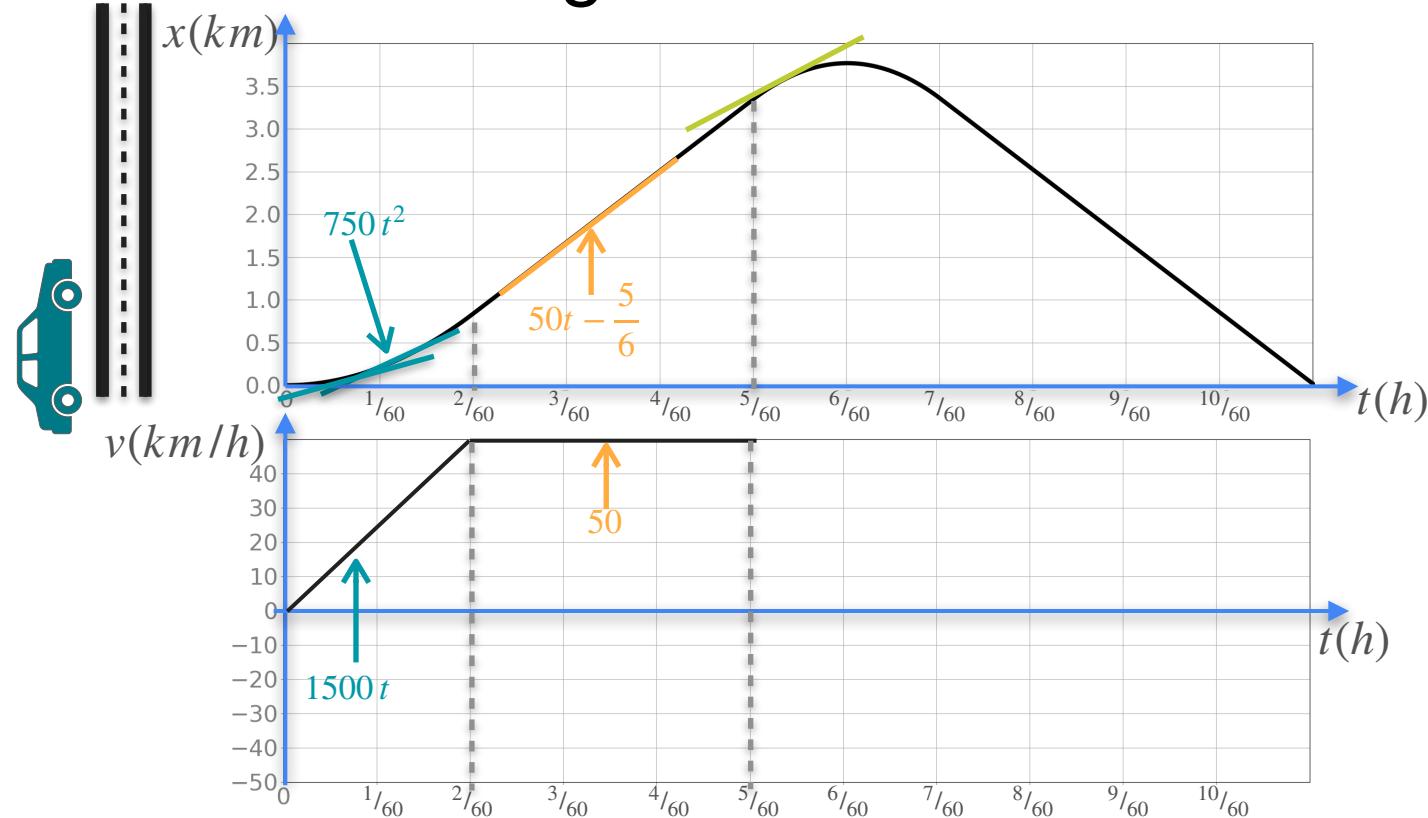


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

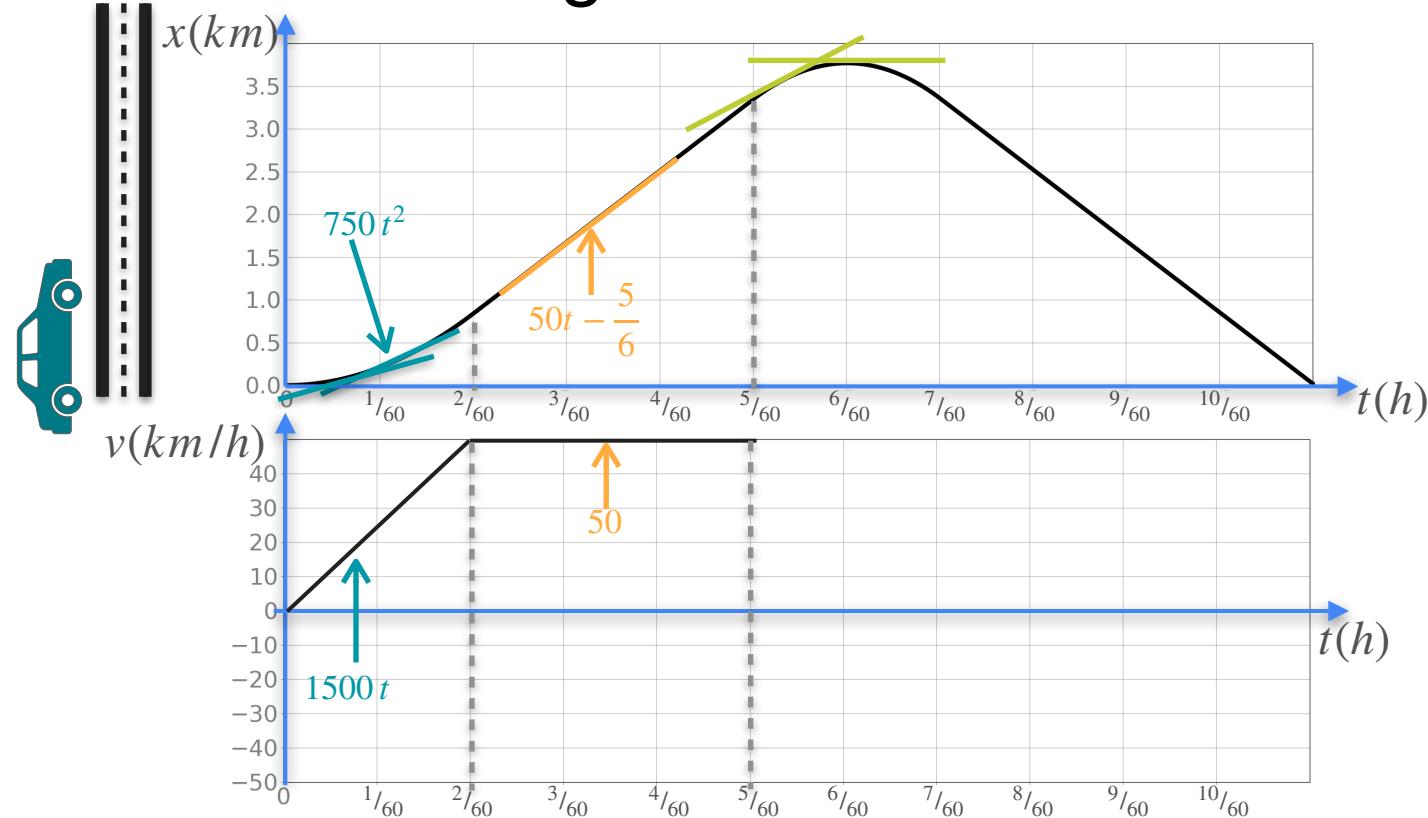


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

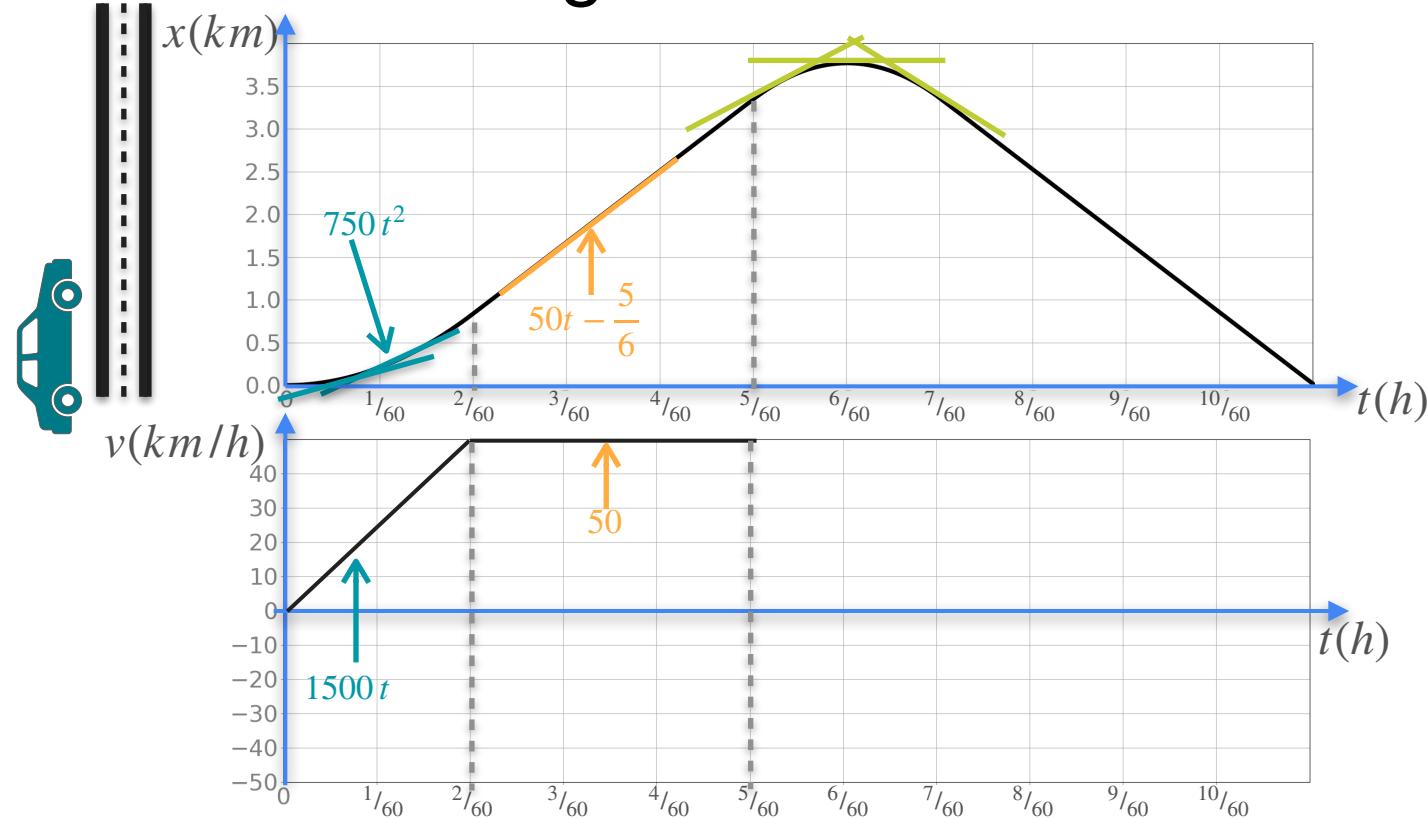


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

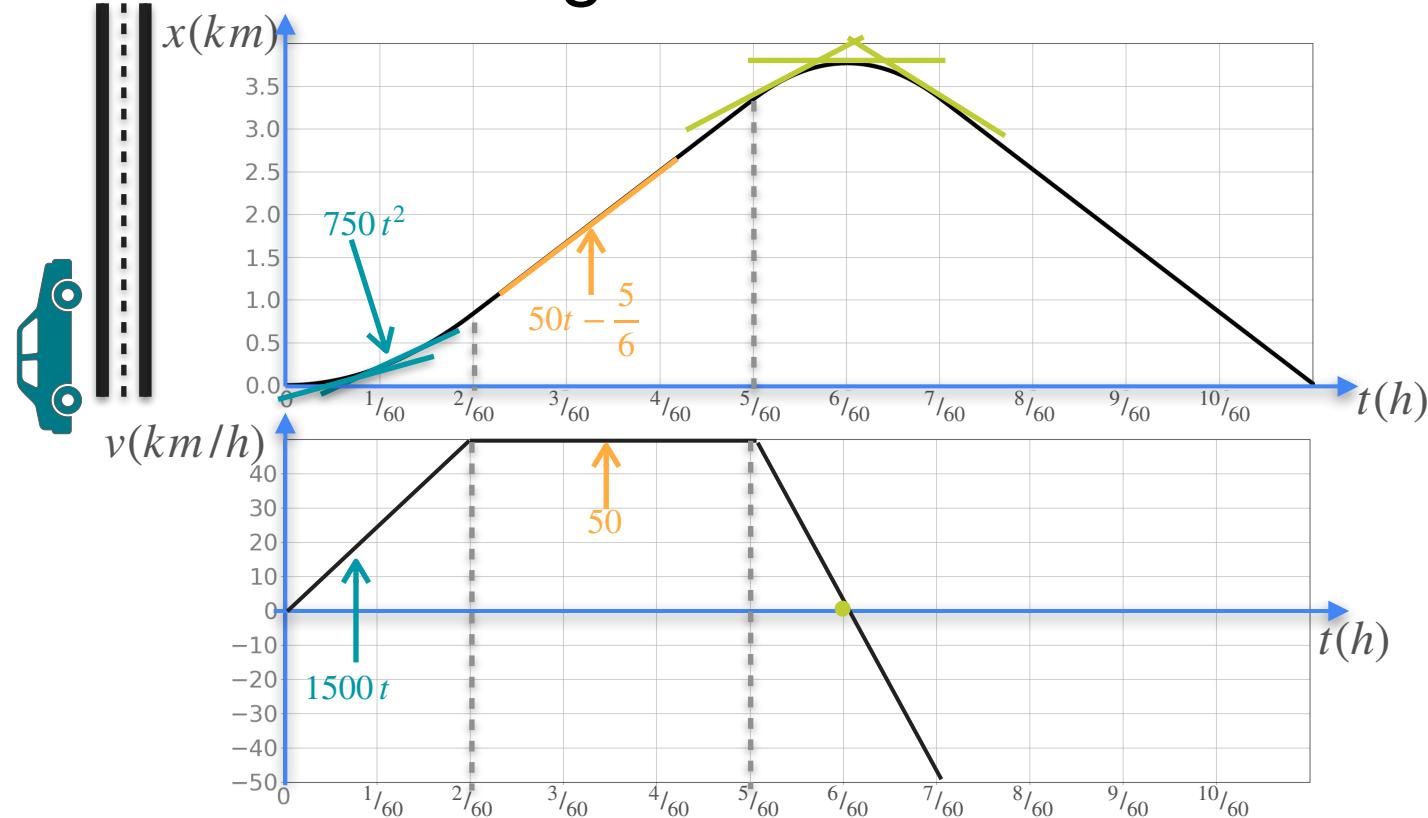


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

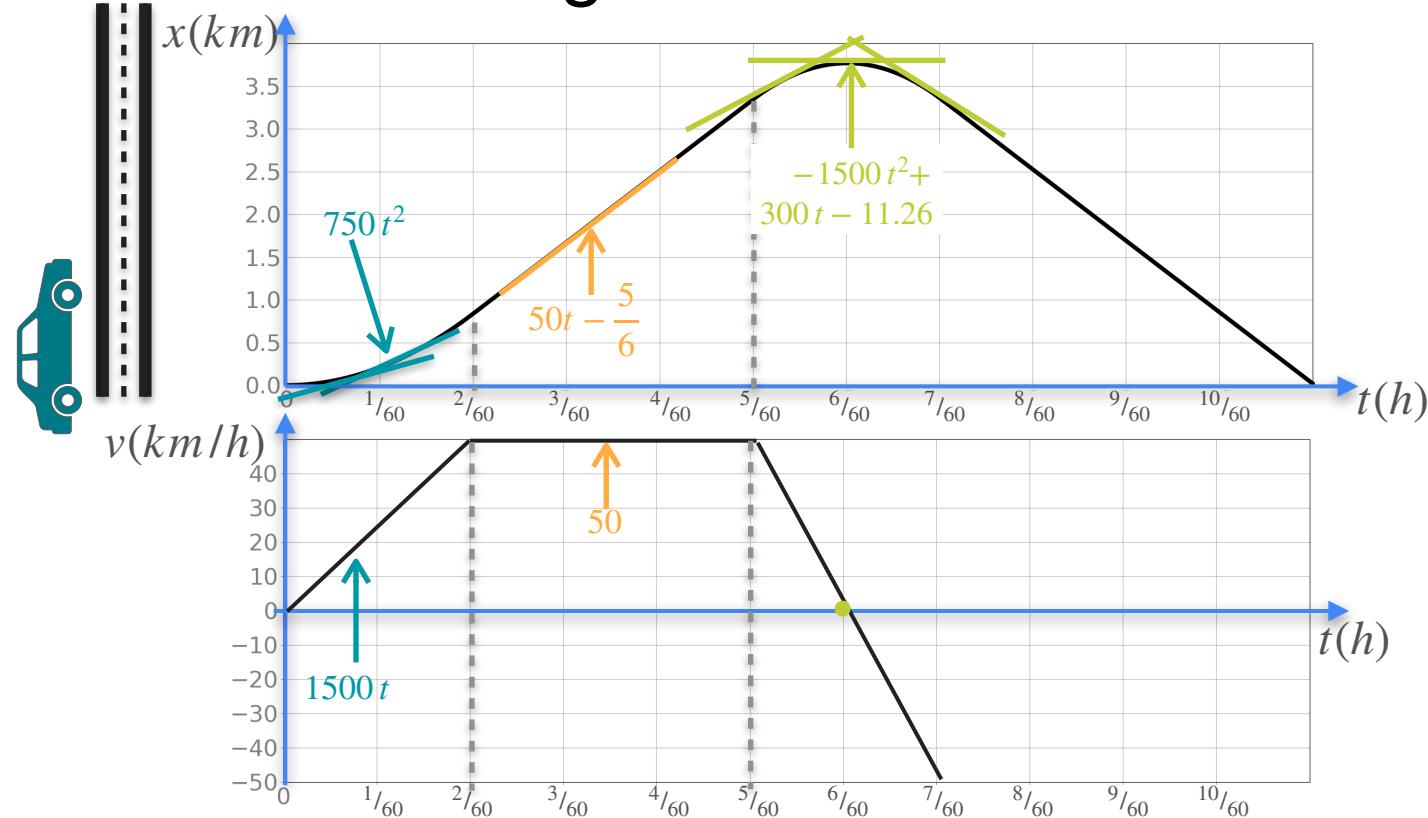


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

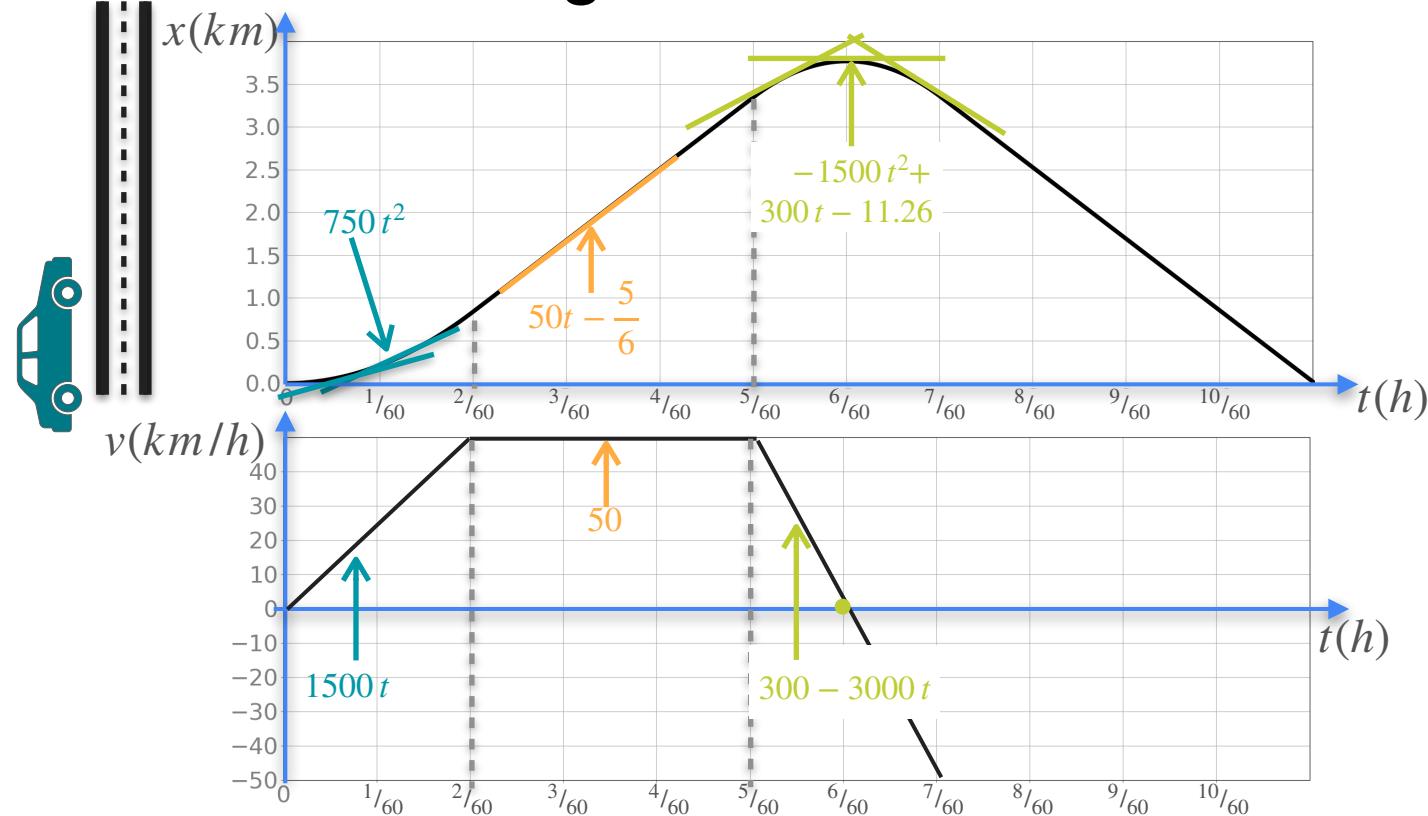


$\mathcal{X}$  Distance

$\mathcal{V}$  Velocity

$$\frac{d^2x}{dt^2}$$

# Understanding Second Derivative

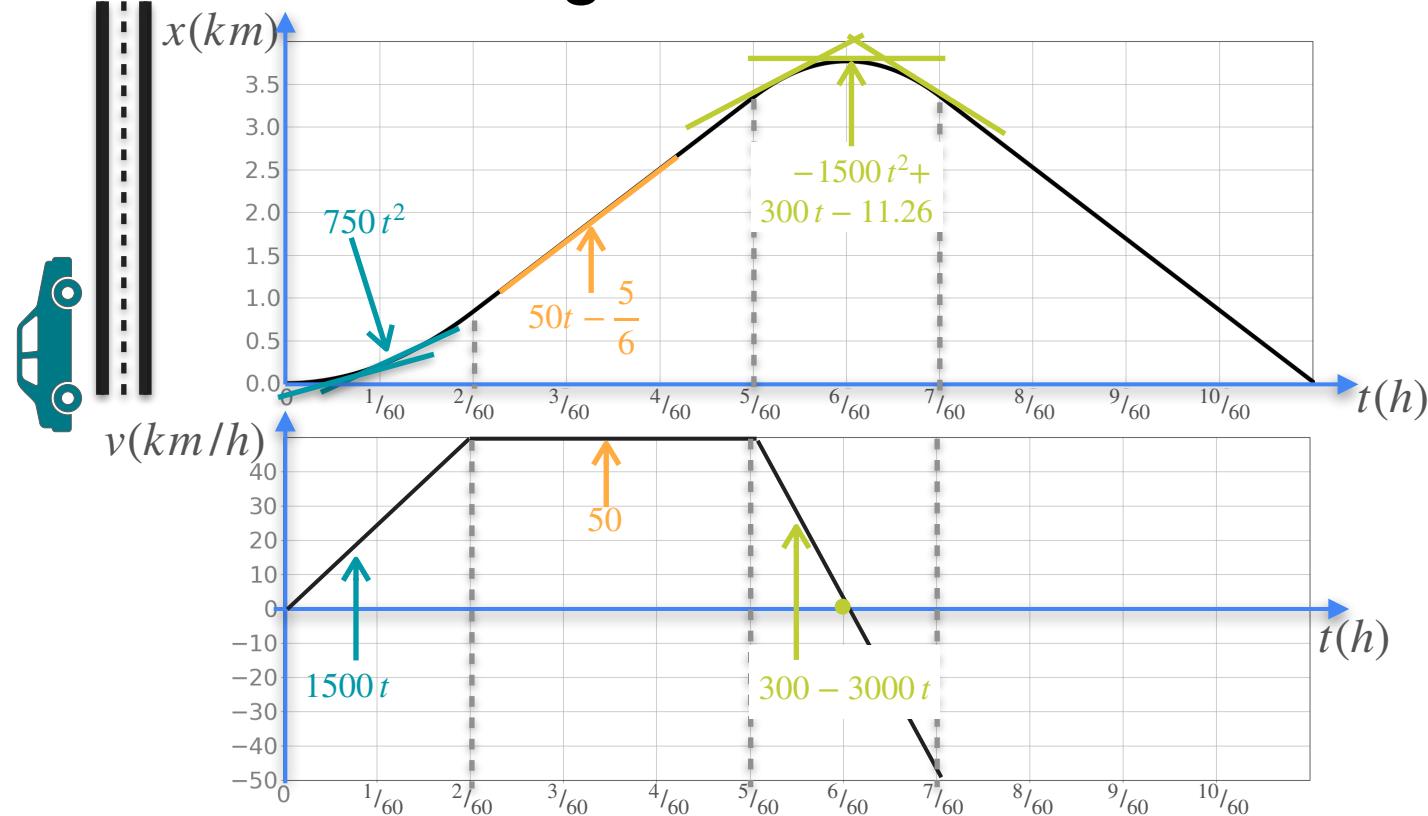


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

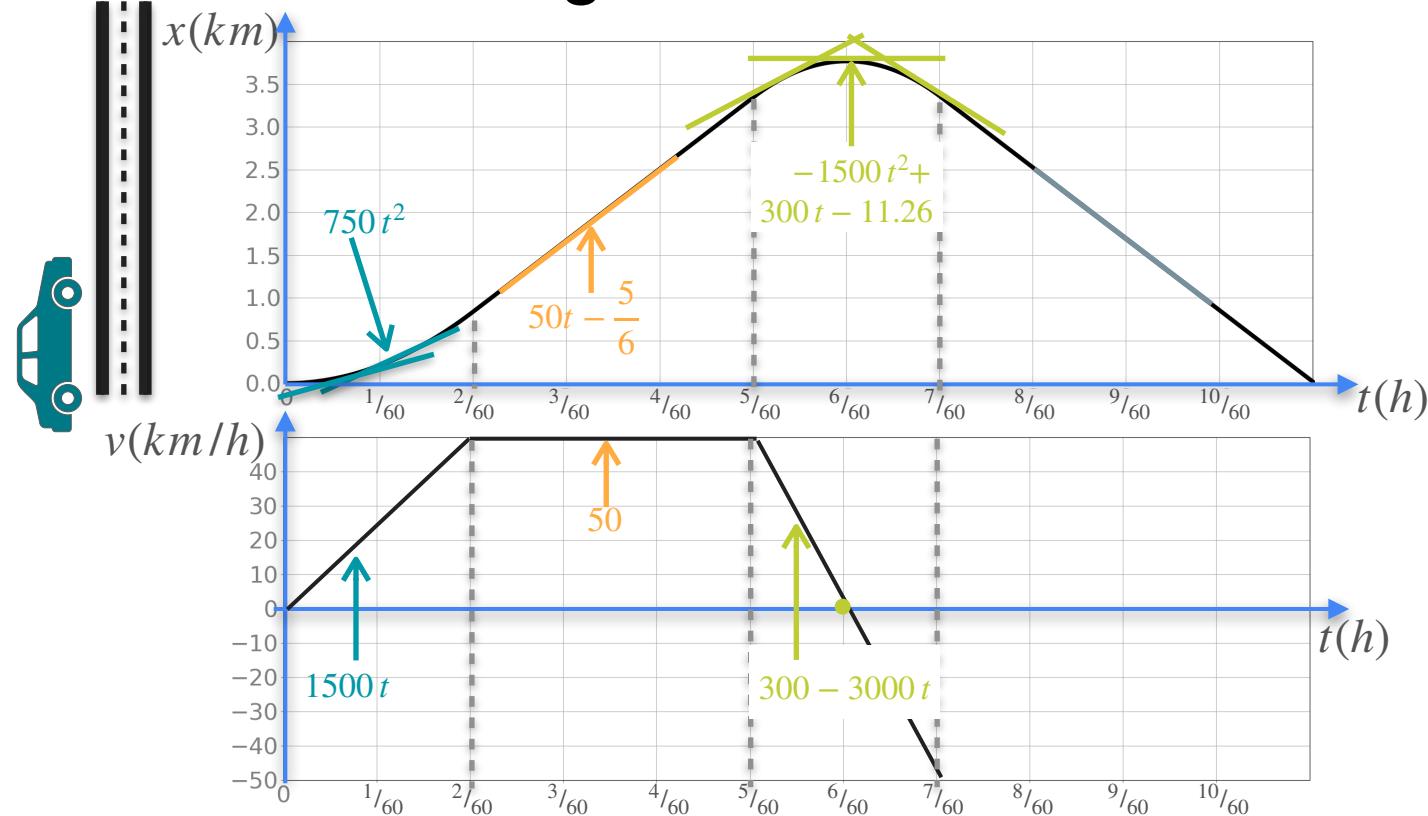


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

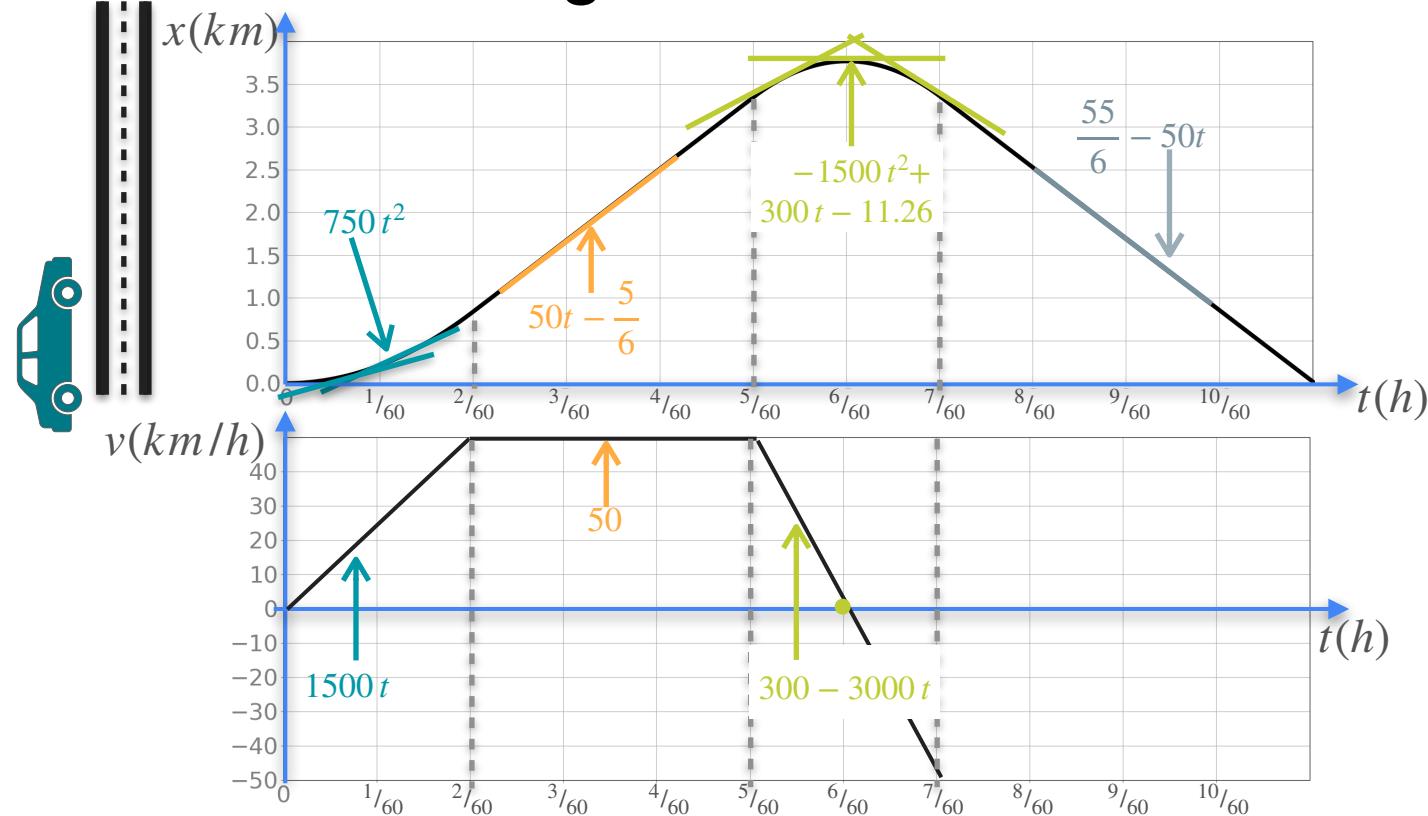


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

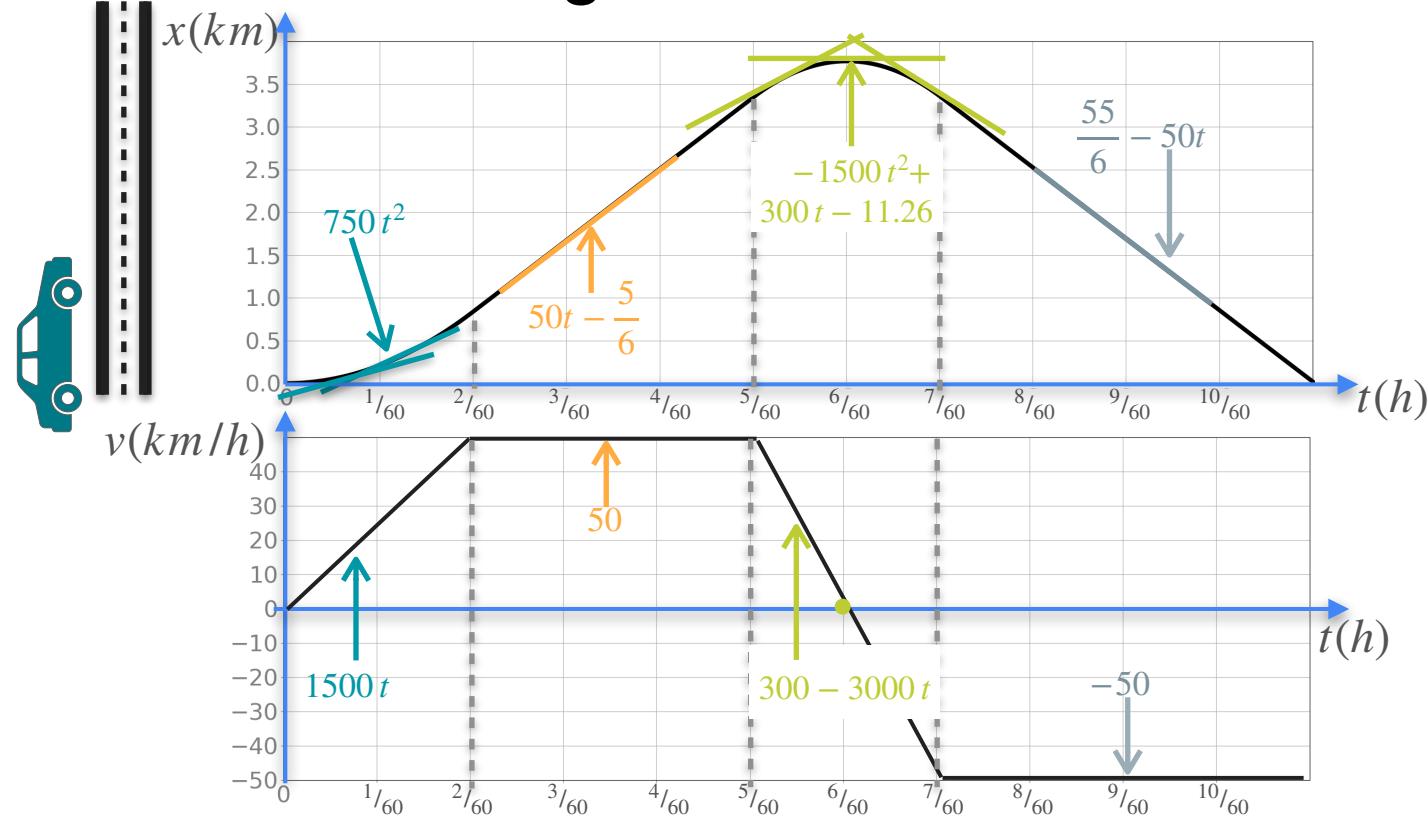


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

# Understanding Second Derivative

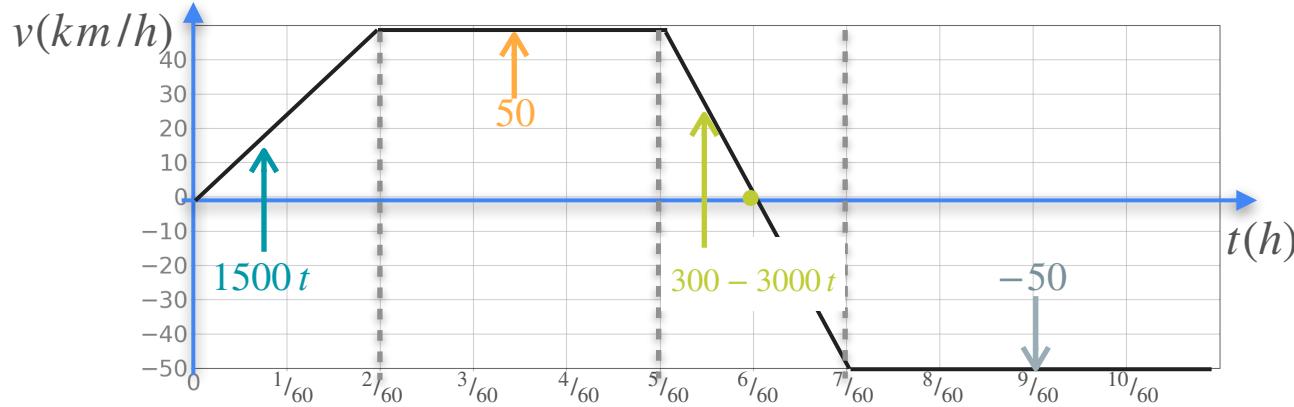


$x$  Distance

$v$  Velocity

$$\frac{dx}{dt}$$

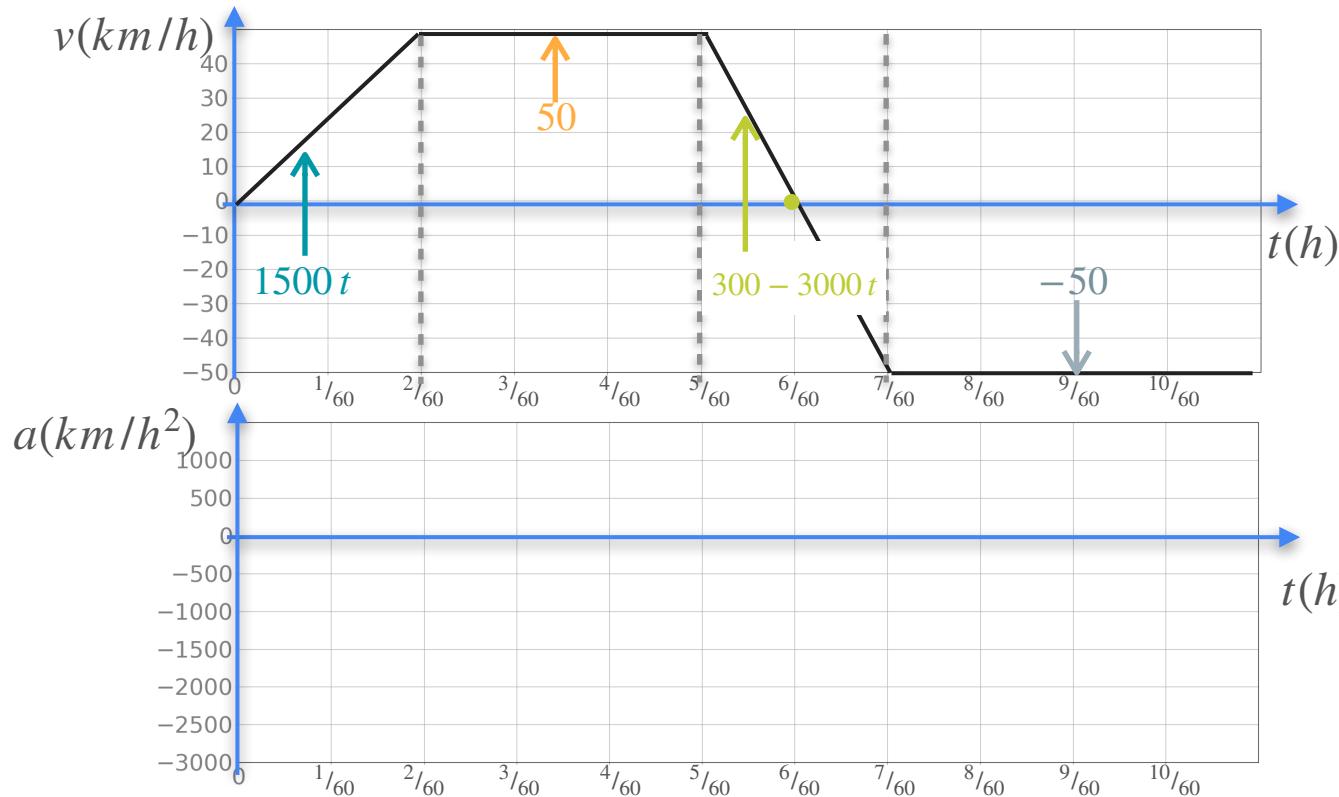
# Understanding Second Derivative



$v$  Velocity

$$\frac{dx}{dt}$$

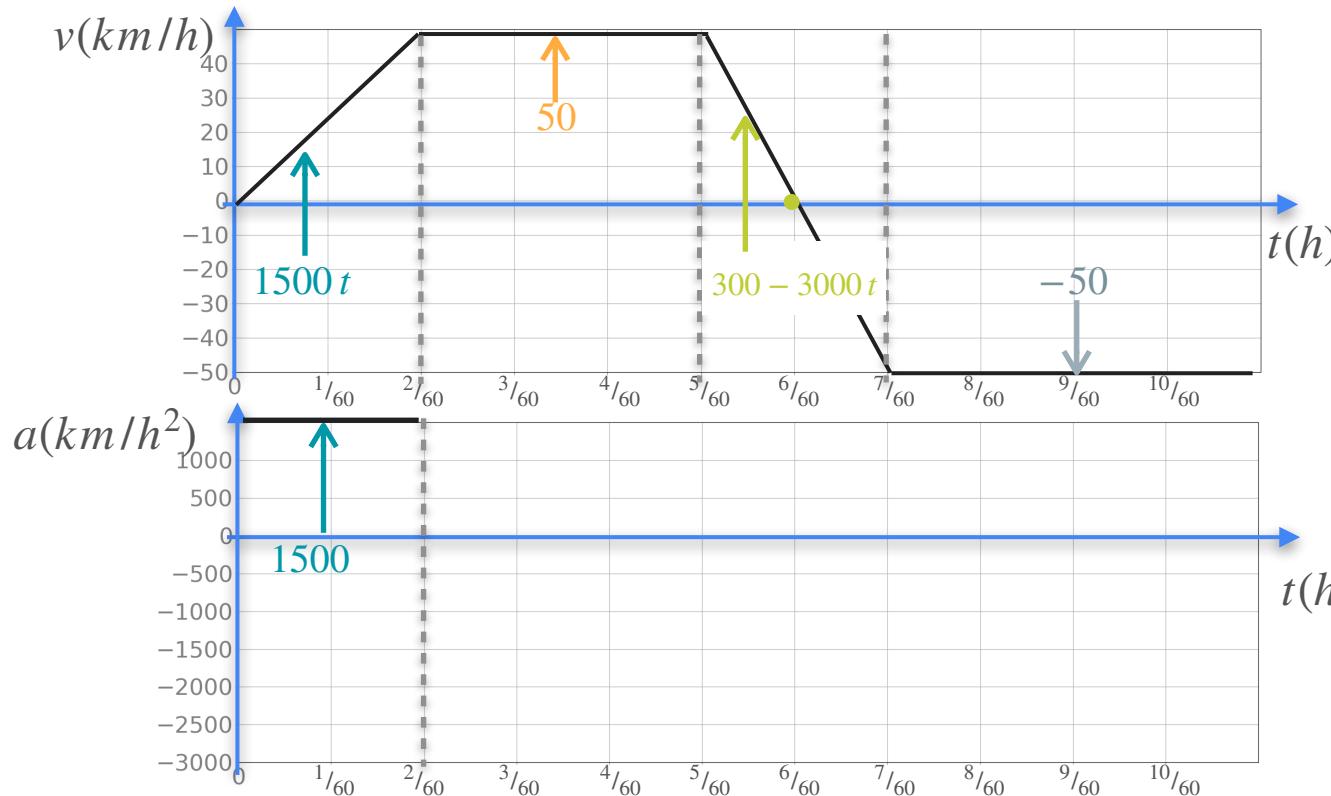
# Understanding Second Derivative



$v$  Velocity  $\frac{dx}{dt}$

$a$  Acceleration  $\frac{dv}{dt}$

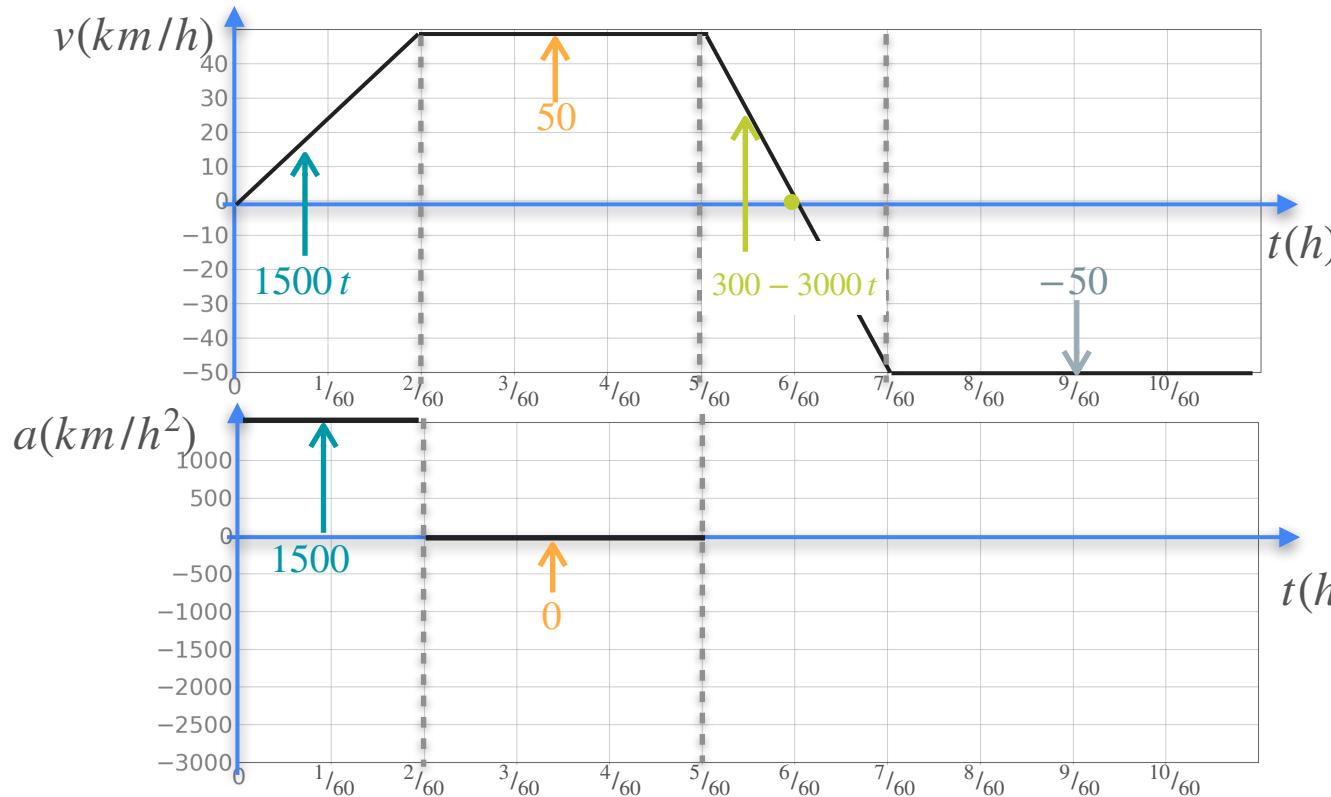
# Understanding Second Derivative



$v$  Velocity  $\frac{dx}{dt}$

$a$  Acceleration  $\frac{dv}{dt}$

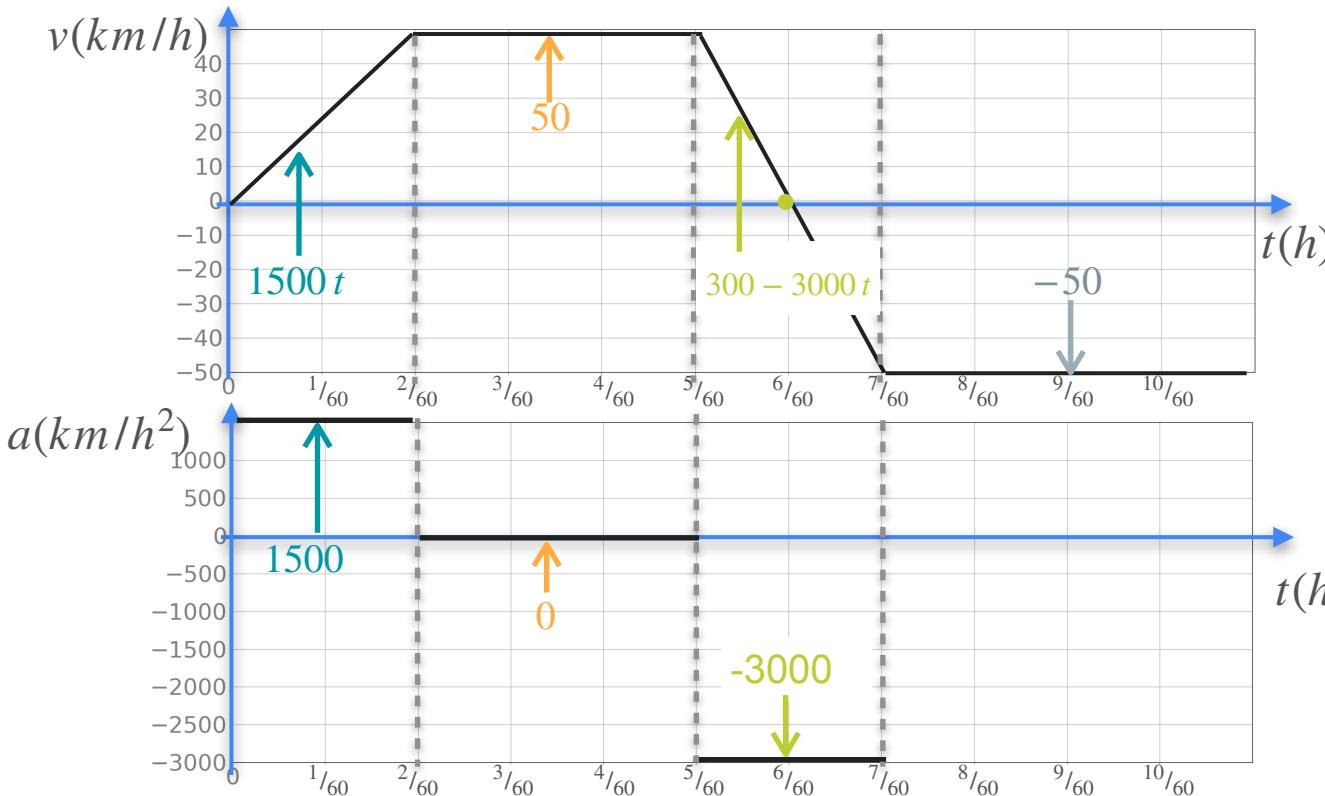
# Understanding Second Derivative



$v$  Velocity  $\frac{dx}{dt}$

$a$  Acceleration  $\frac{dv}{dt}$

# Understanding Second Derivative



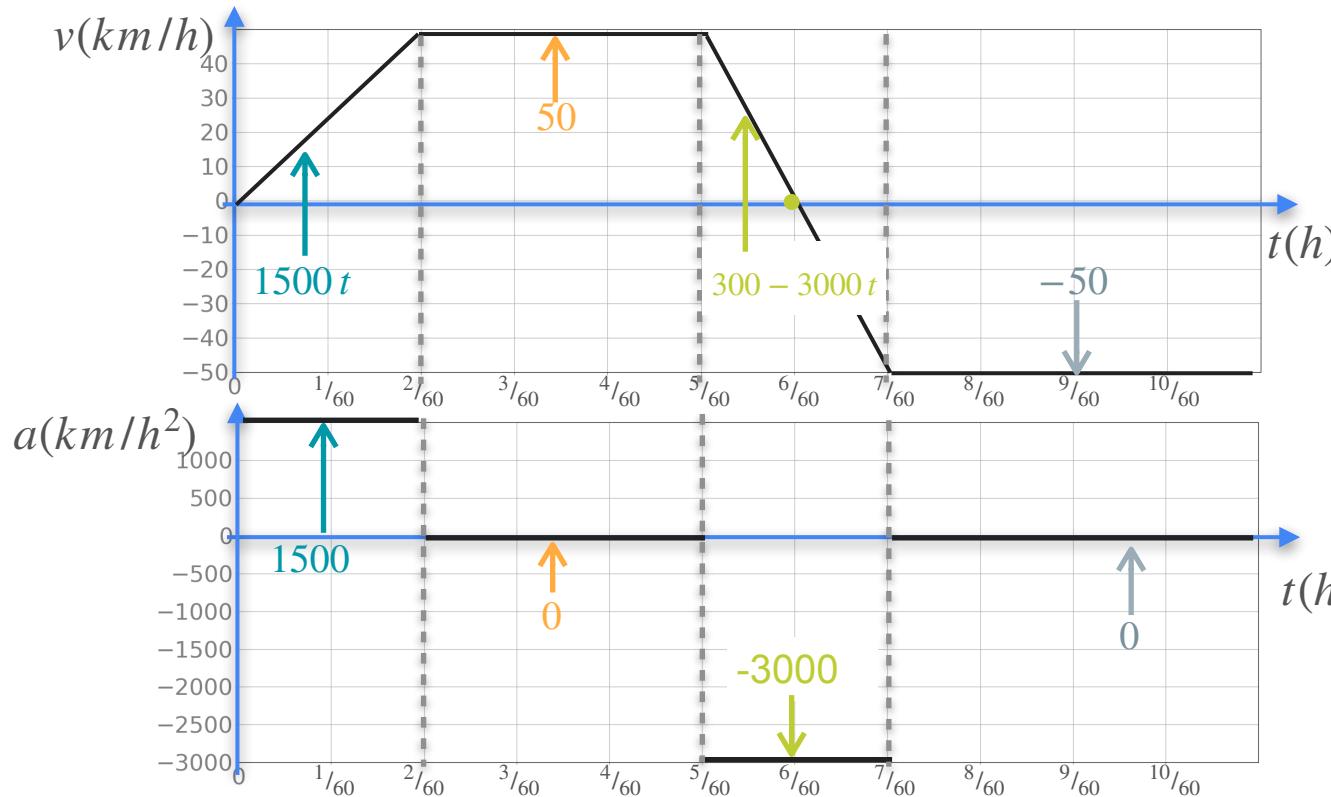
$v$  Velocity

$$\frac{dx}{dt}$$

$a$  Acceleration

$$\frac{dv}{dt}$$

# Understanding Second Derivative



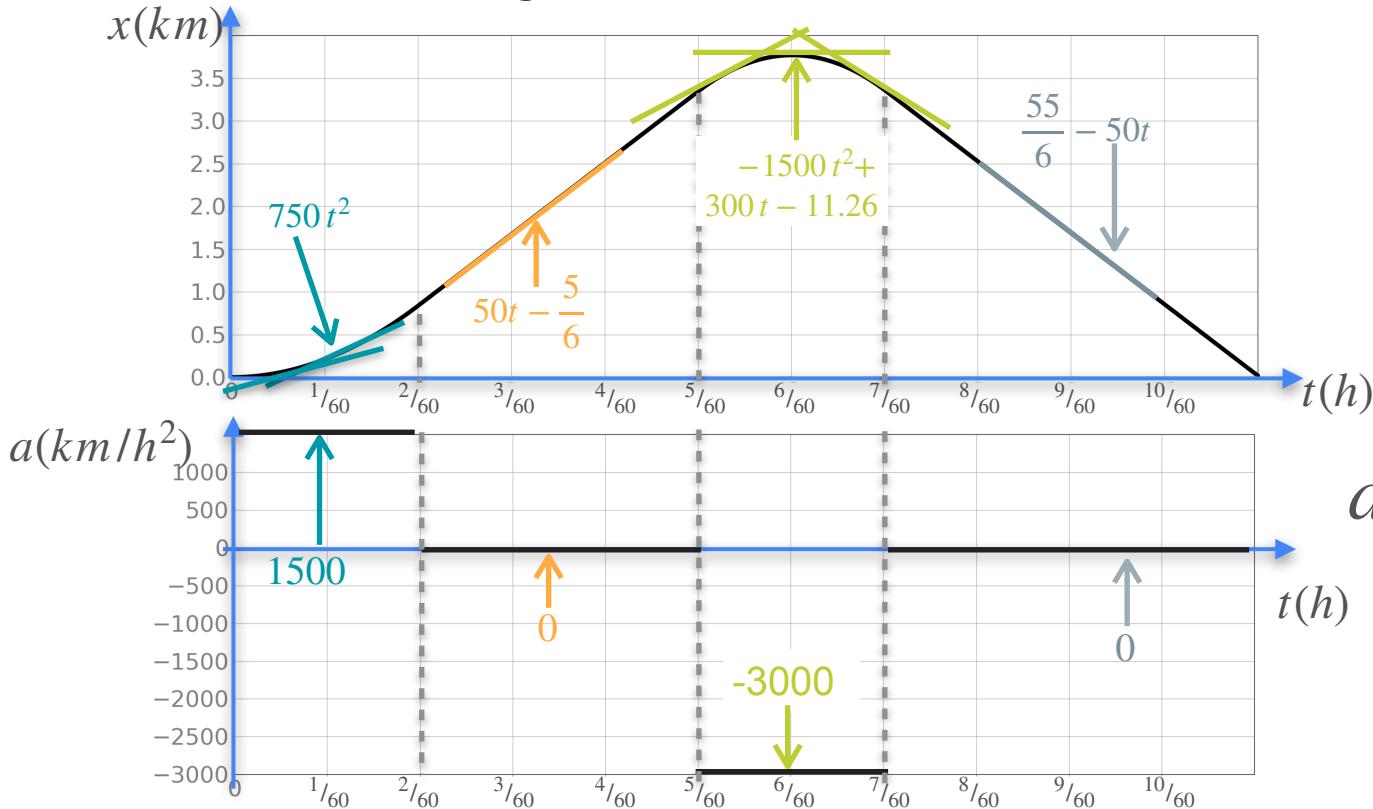
$v$  Velocity

$$\frac{dx}{dt}$$

$a$  Acceleration

$$\frac{dv}{dt}$$

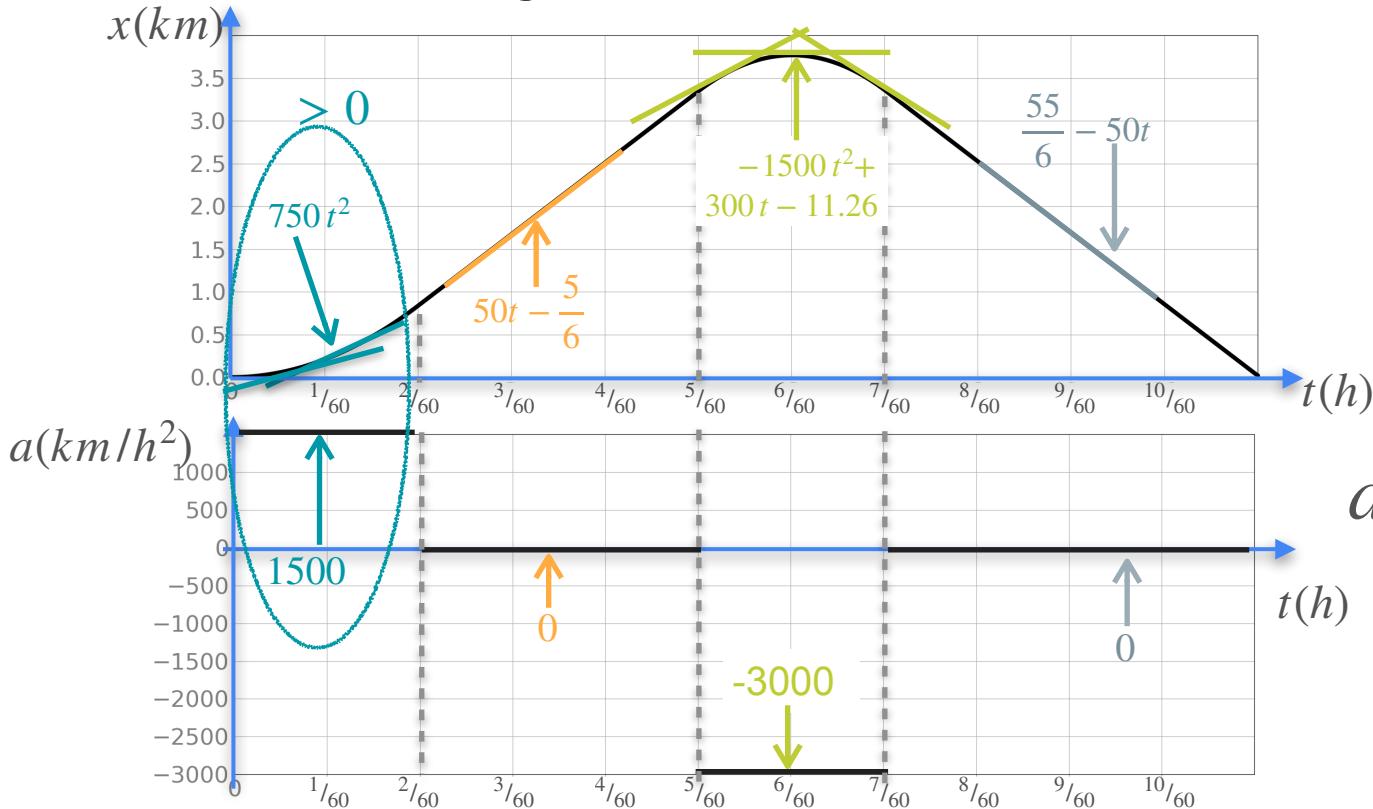
# Understanding Second Derivative



$x$  Distance

$a$  Acceleration  $\frac{d^2x}{dt^2}$

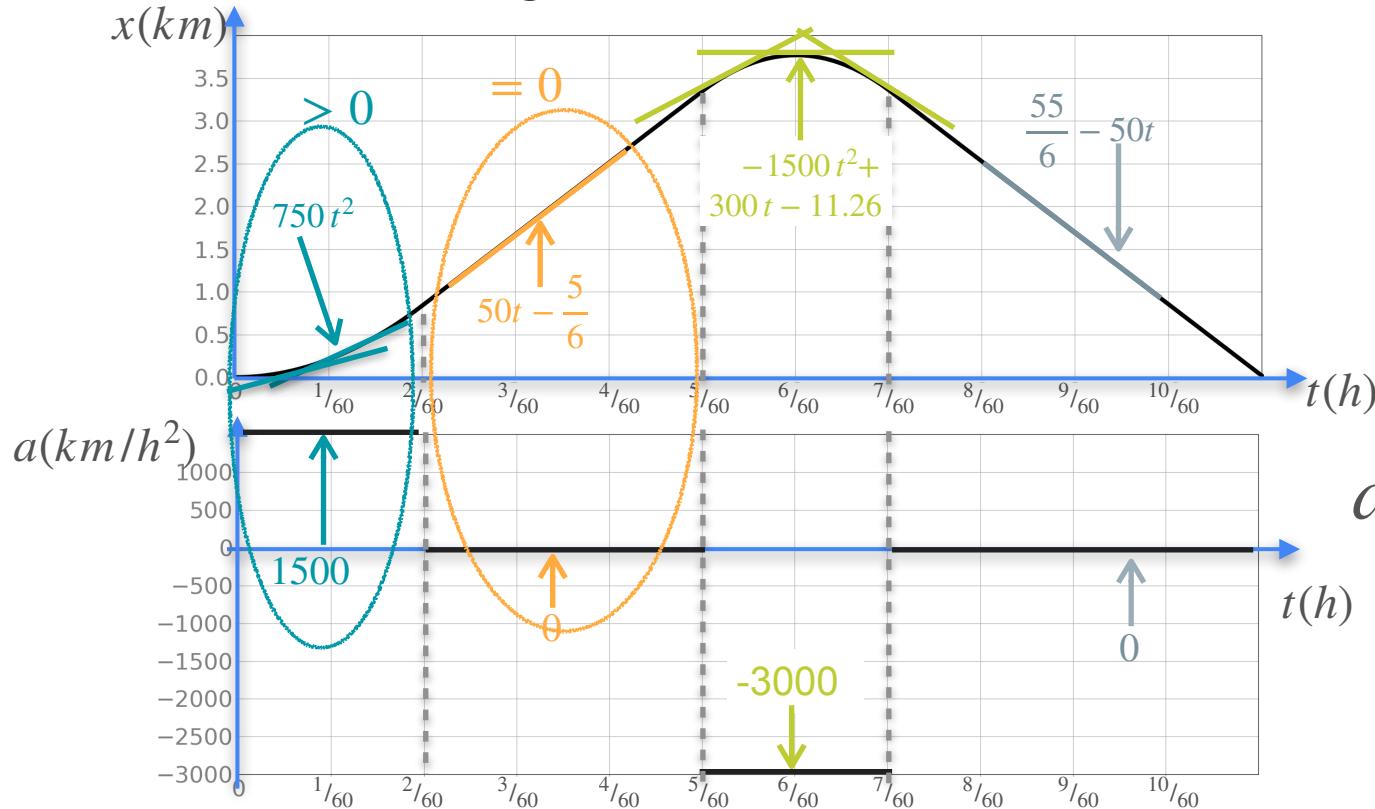
# Understanding Second Derivative



$x$  Distance

$a$  Acceleration  $\frac{d^2x}{dt^2}$

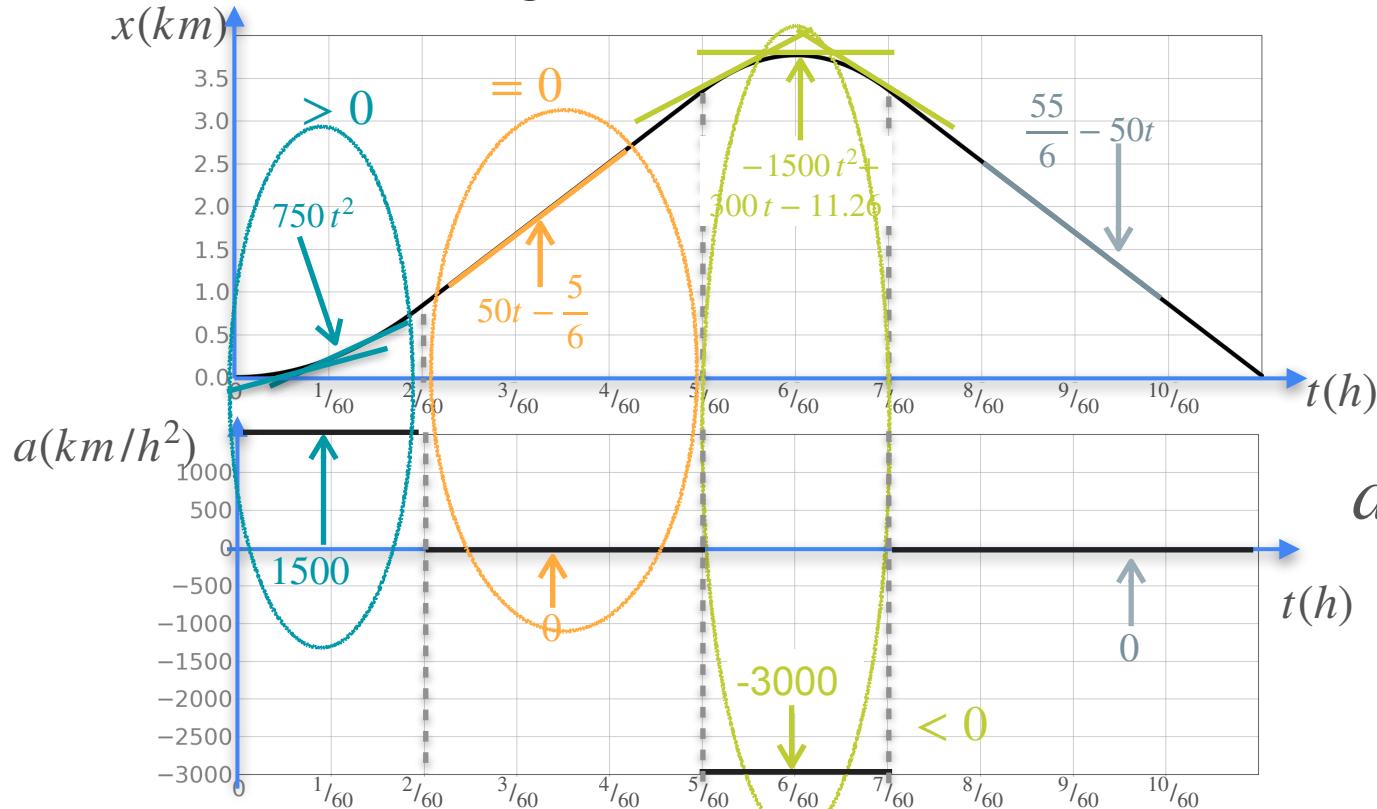
# Understanding Second Derivative



$x$  Distance

$a$  Acceleration  $\frac{d^2x}{dt^2}$

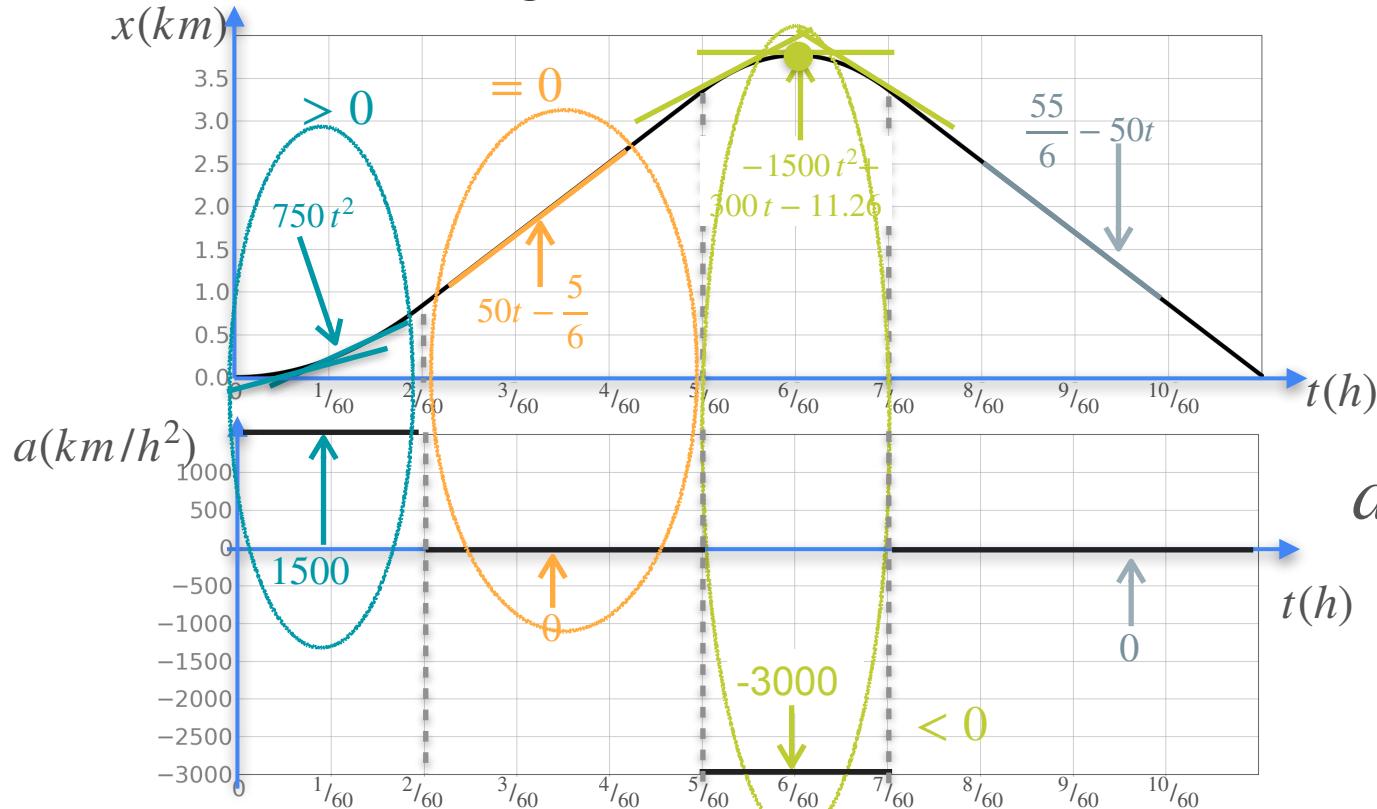
# Understanding Second Derivative



$x$  Distance

$a$  Acceleration  $\frac{d^2x}{dt^2}$

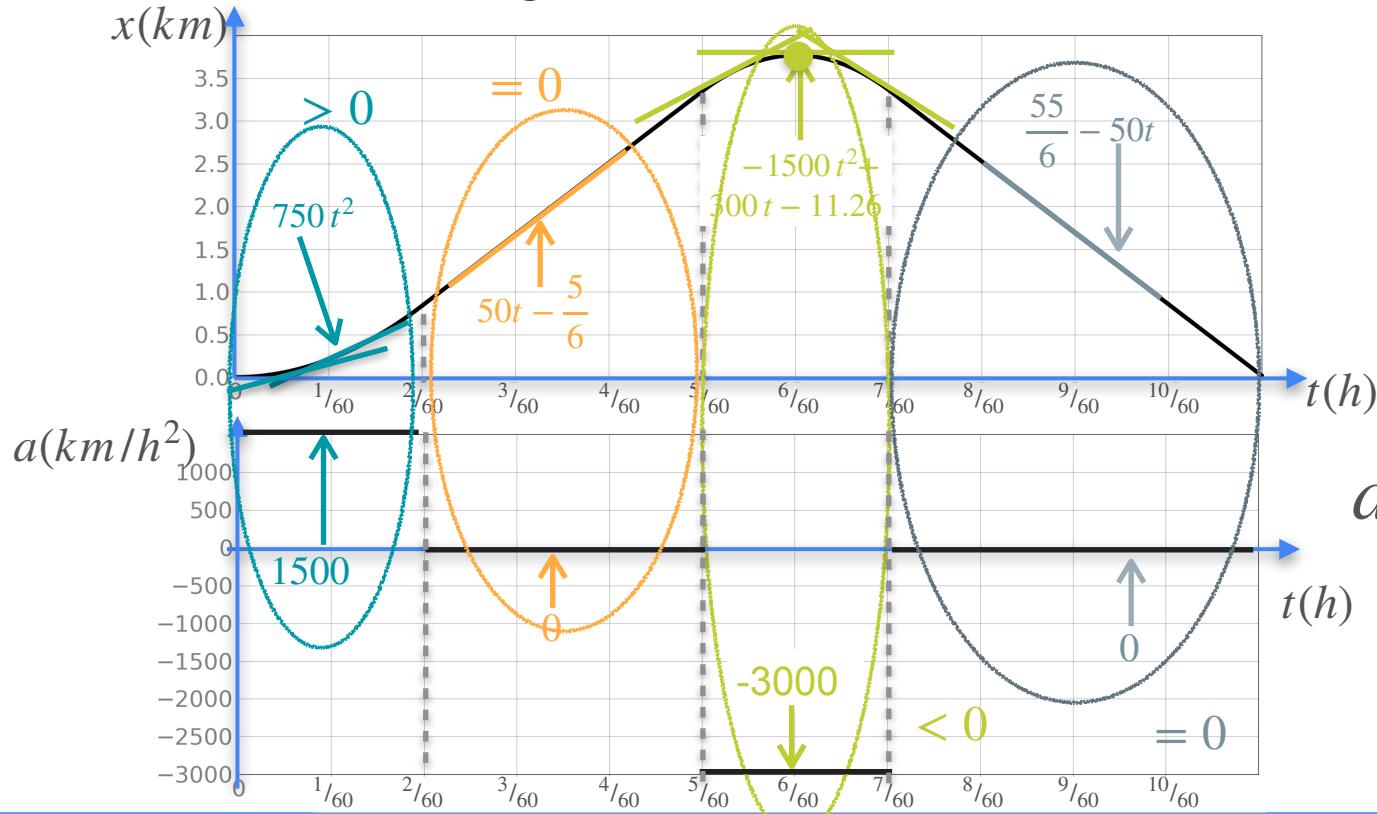
# Understanding Second Derivative



$x$  Distance

$a$  Acceleration  $\frac{d^2x}{dt^2}$

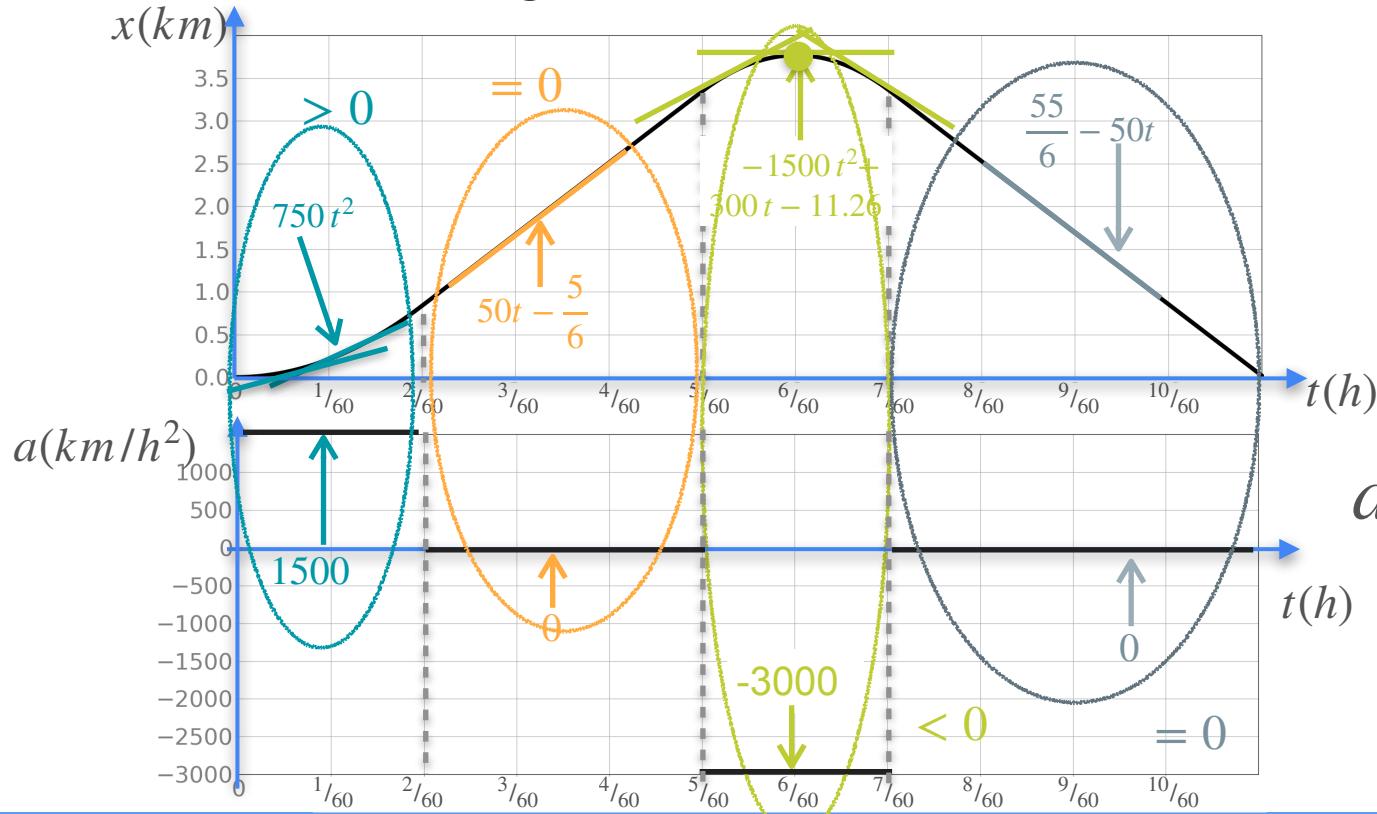
# Understanding Second Derivative



$x$  Distance

$a$  Acceleration  $\frac{d^2 x}{dt^2}$

# Understanding Second Derivative



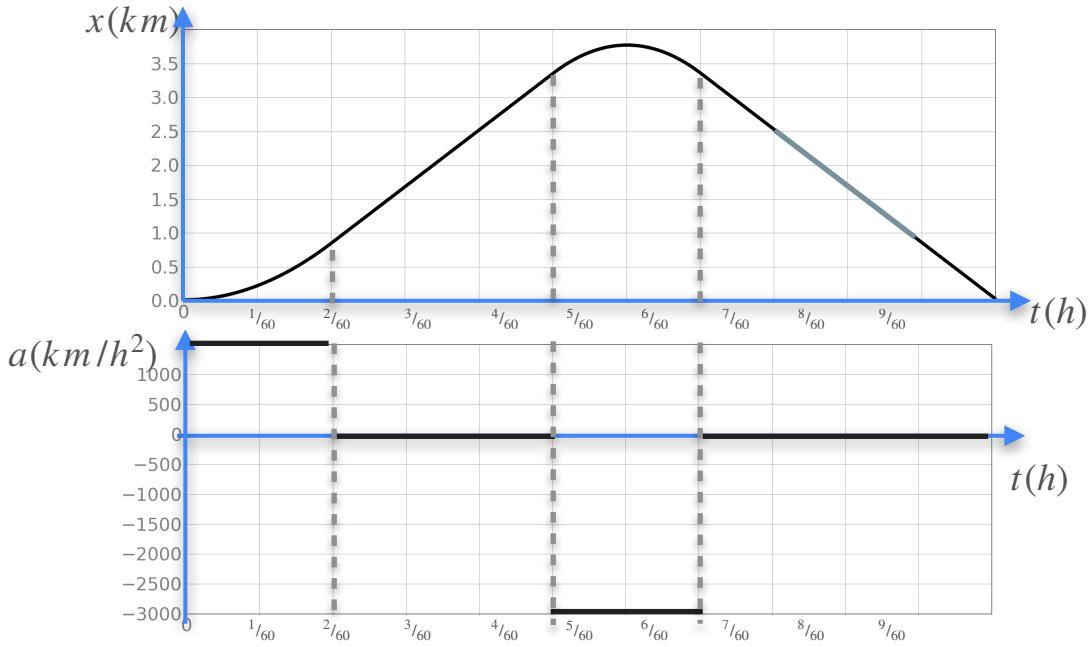
$x$  Distance

Second derivative tells us about the curvature

$a$  Acceleration  $\frac{d^2x}{dt^2}$

# Curvature

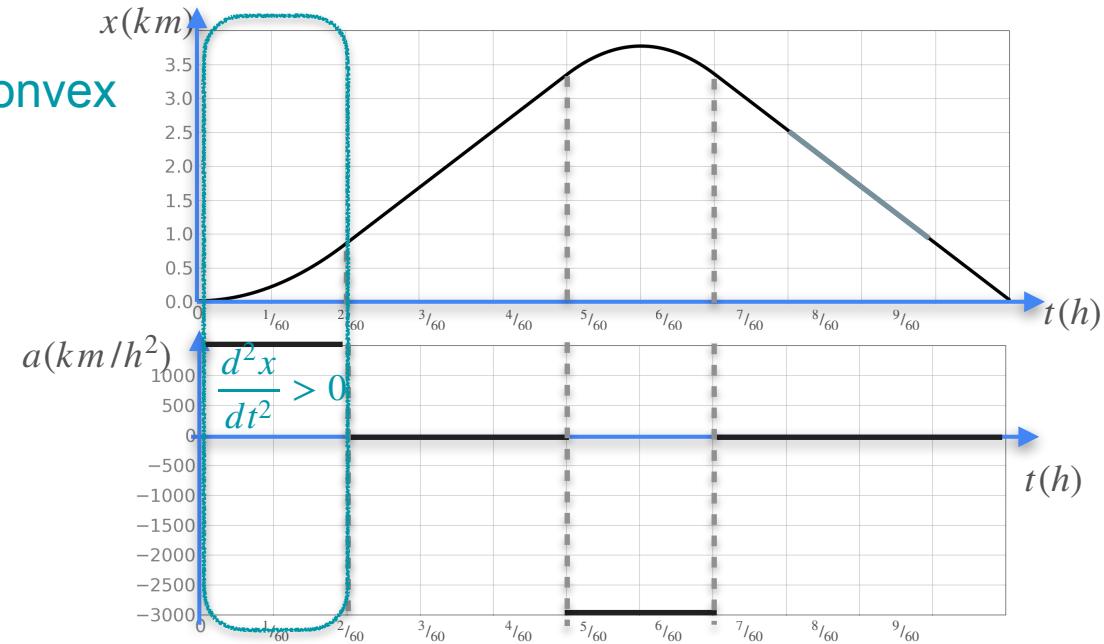
# Curvature



# Curvature

$$\frac{d^2x}{dt^2} > 0$$

Concave up or convex



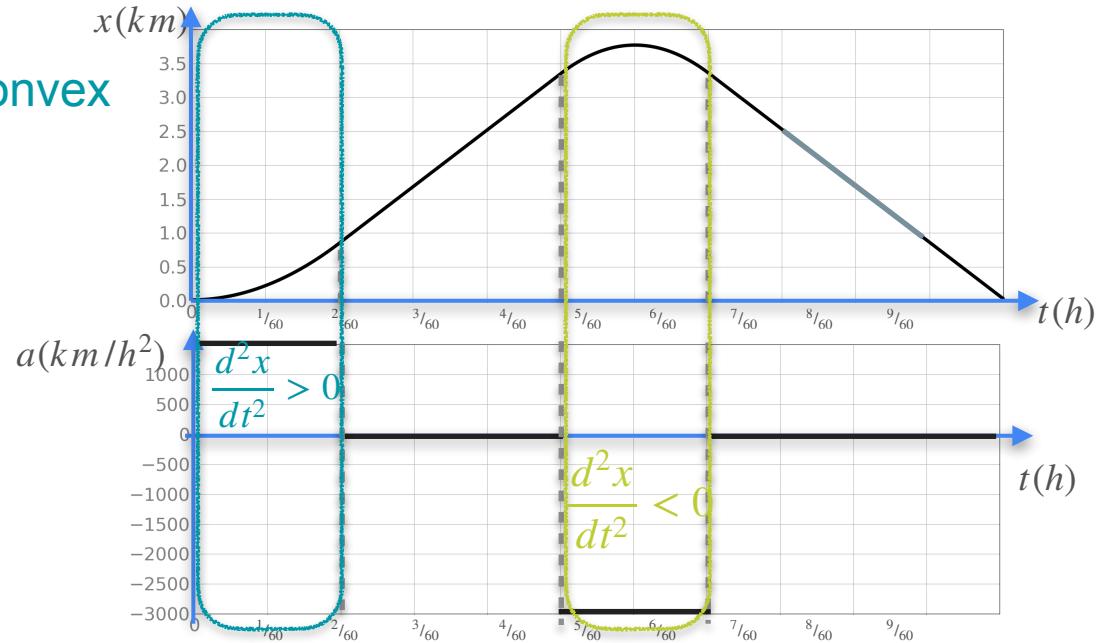
# Curvature

$$\frac{d^2x}{dt^2} > 0$$

Concave up or convex

$$\frac{d^2x}{dt^2} < 0$$

Concave down



# Curvature

$$\frac{d^2x}{dt^2} > 0$$

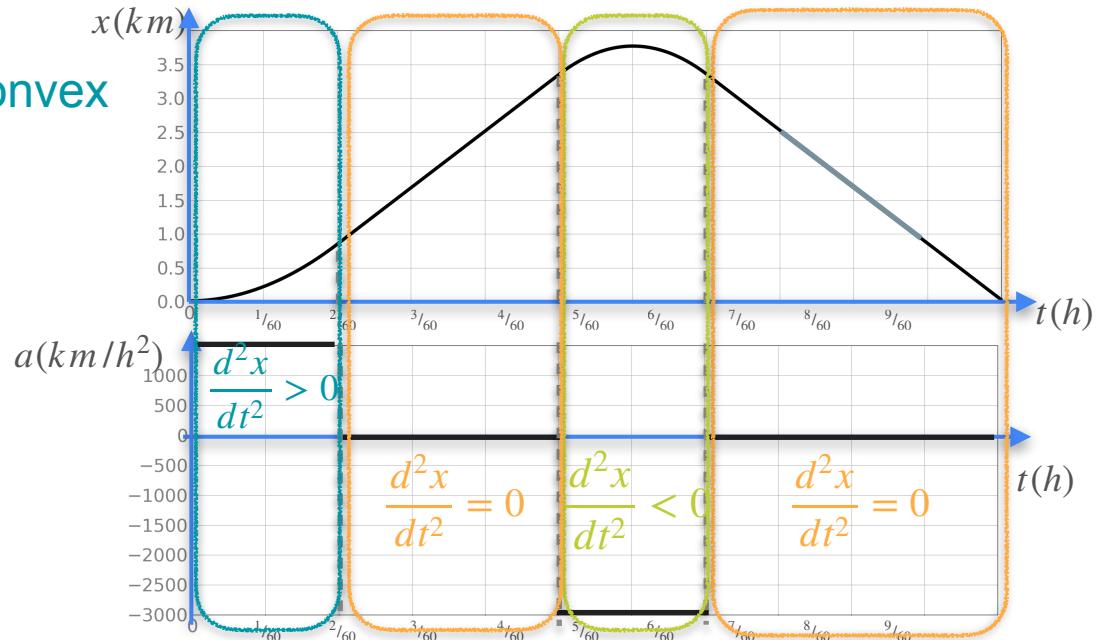
Concave up or convex

$$\frac{d^2x}{dt^2} < 0$$

Concave down

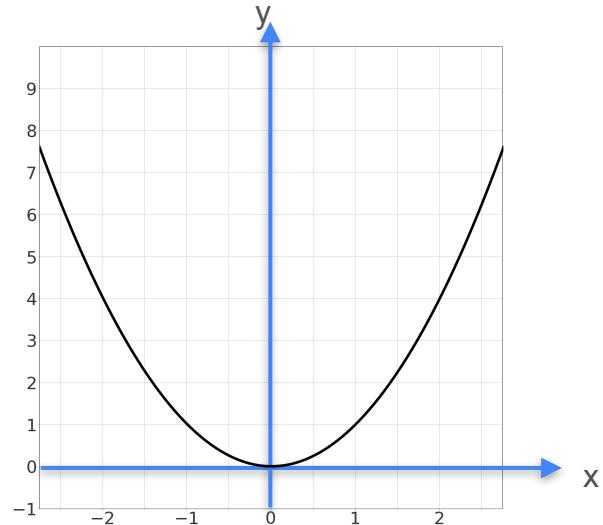
$$\frac{d^2x}{dt^2} = 0$$

Need more information

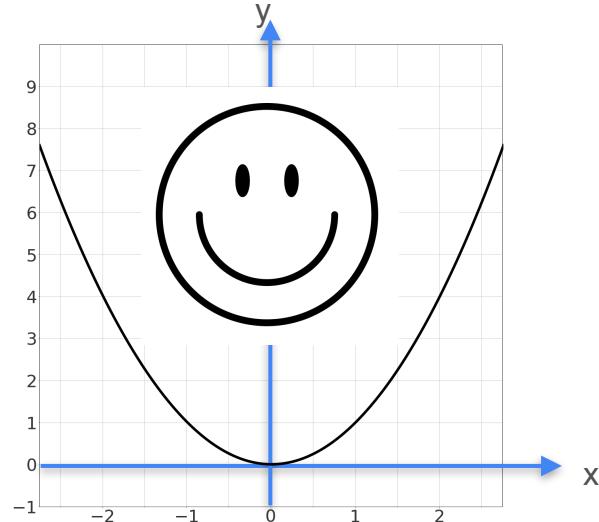


# Curvature

# Curvature



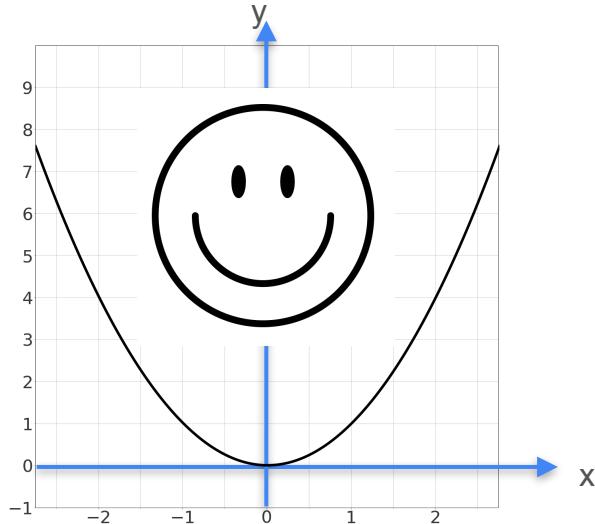
# Curvature



Concave up or convex

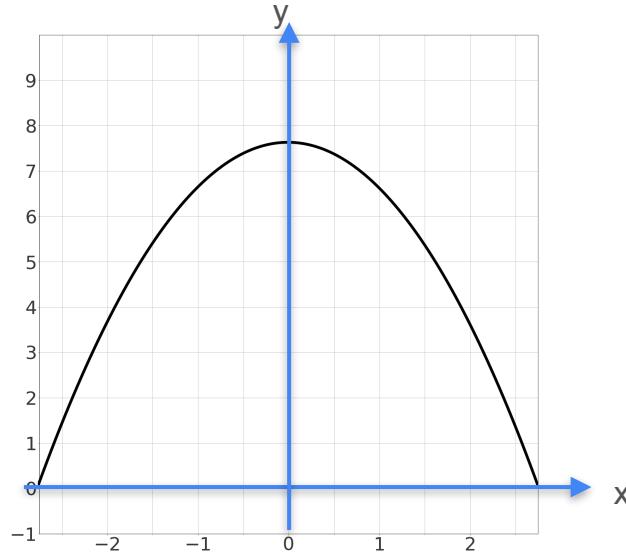
$$f''(0) > 0$$

# Curvature

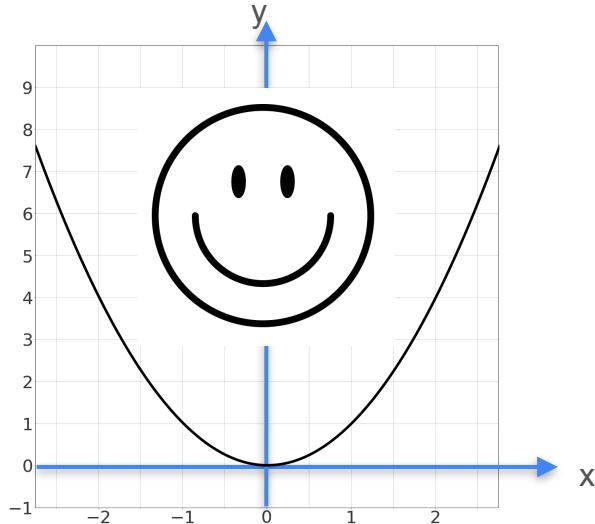


Concave up or convex

$$f''(0) > 0$$

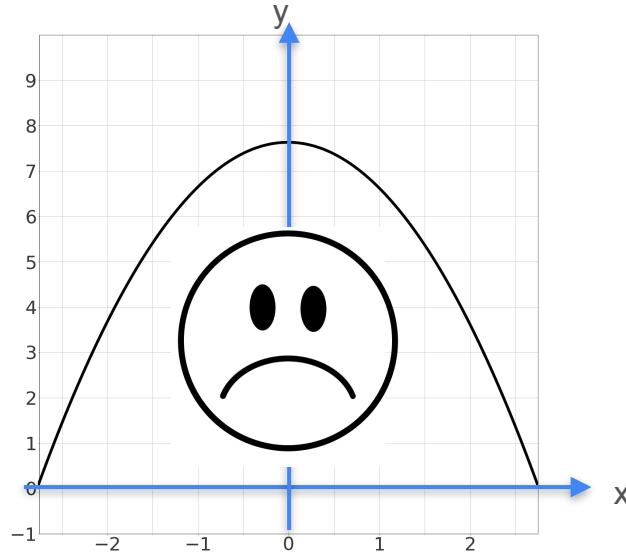


# Curvature



Concave up or convex

$$f''(0) > 0$$



Concave down

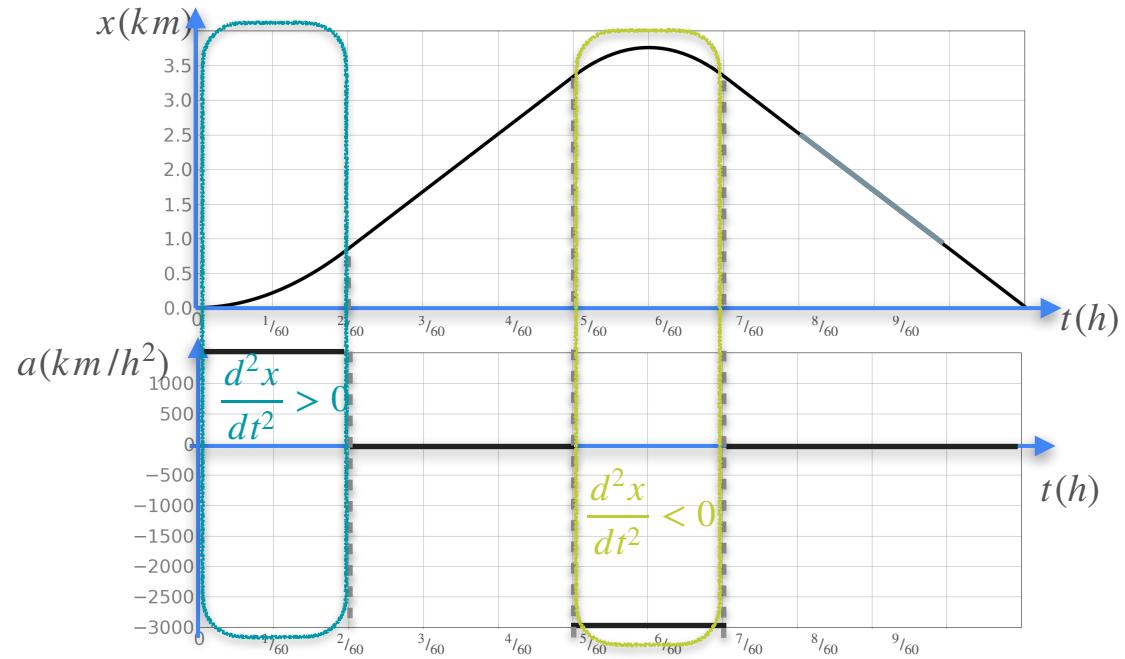
$$f''(0) < 0$$

# Second Derivative and Optimization

$$\frac{d^2x}{dt^2} > 0$$

$$\frac{d^2x}{dt^2} < 0$$

$$\frac{d^2x}{dt^2} = 0$$

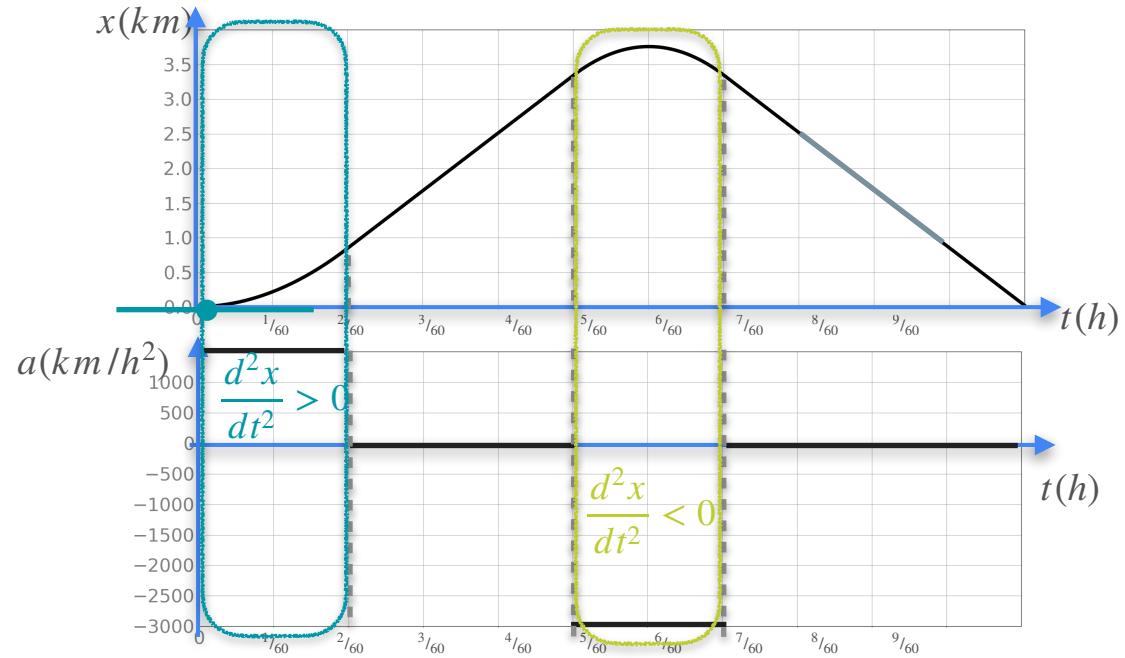


# Second Derivative and Optimization

$\frac{d^2x}{dt^2} > 0$     (Local) Minimum

$\frac{d^2x}{dt^2} < 0$

$\frac{d^2x}{dt^2} = 0$

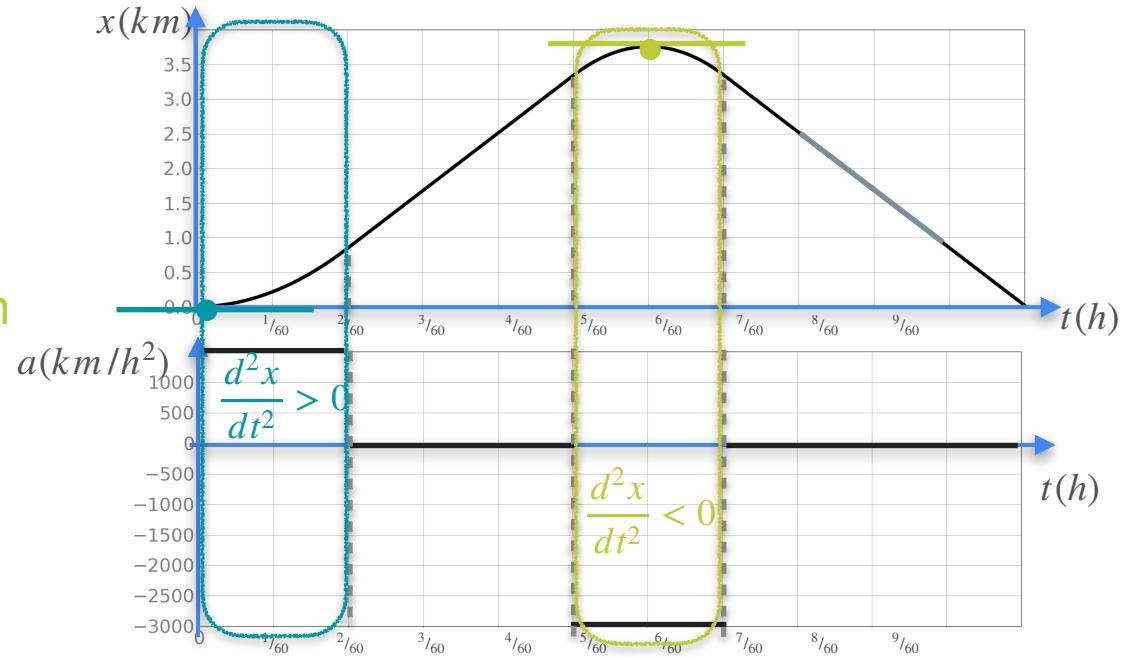


# Second Derivative and Optimization

$$\frac{d^2x}{dt^2} > 0 \quad (\text{Local}) \text{ Minimum}$$

$$\frac{d^2x}{dt^2} < 0 \quad (\text{Local}) \text{ maximum}$$

$$\frac{d^2x}{dt^2} = 0$$

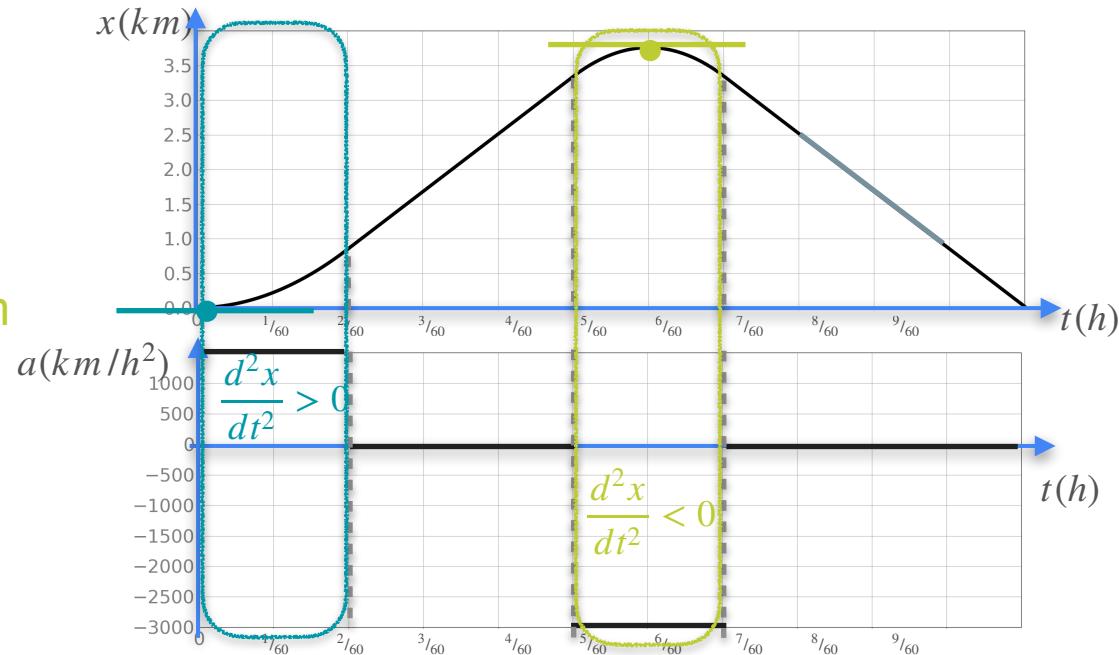


# Second Derivative and Optimization

$\frac{d^2x}{dt^2} > 0$     (Local) Minimum

$\frac{d^2x}{dt^2} < 0$     (Local) maximum

$\frac{d^2x}{dt^2} = 0$     Inconclusive



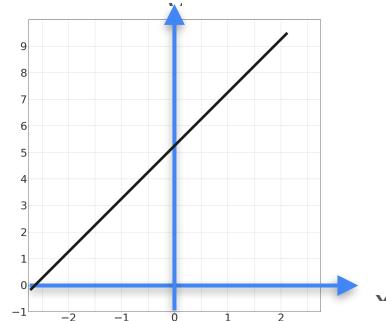
# Curvature

First derivative

Second derivative

# Curvature

First derivative



Increasing

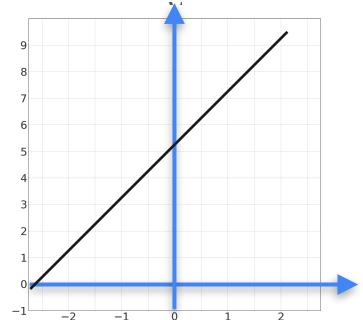
$$f'(0) > 0$$

Second derivative



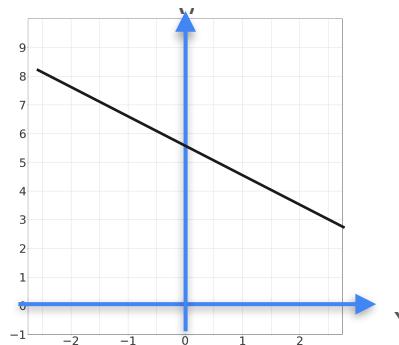
# Curvature

First derivative



Increasing

$$f'(0) > 0$$



Decreasing

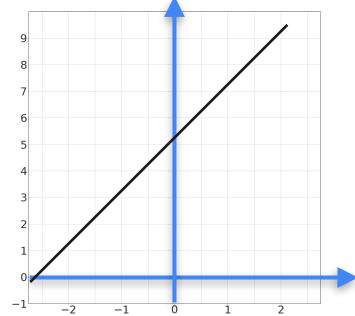
$$f'(0) < 0$$

Second derivative



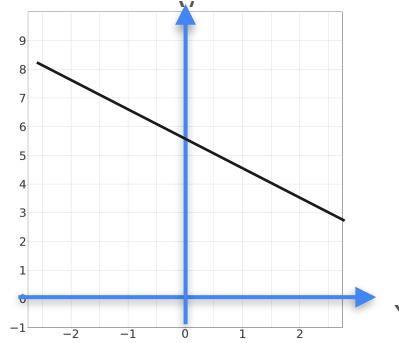
# Curvature

First derivative



Increasing

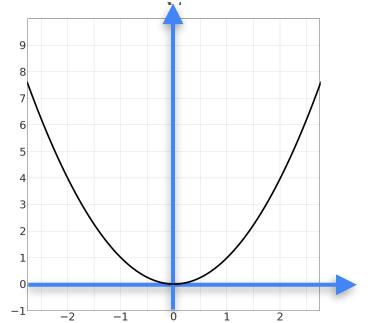
$$f'(0) > 0$$



Decreasing

$$f'(0) < 0$$

Second derivative

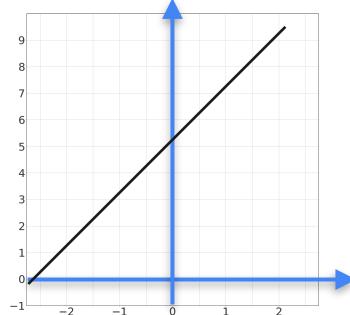


Concave up

$$f''(0) > 0$$

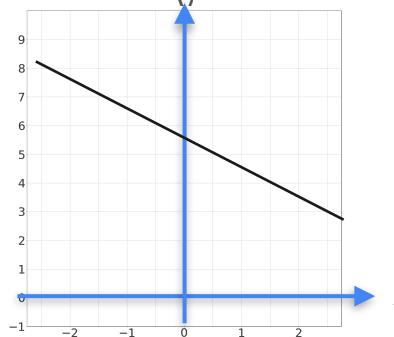
# Curvature

First derivative



Increasing

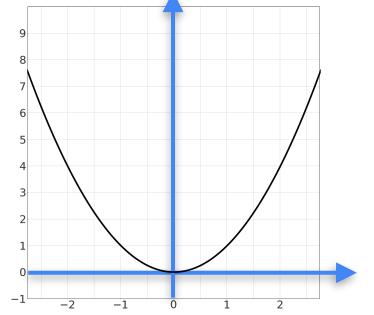
$$f'(0) > 0$$



Decreasing

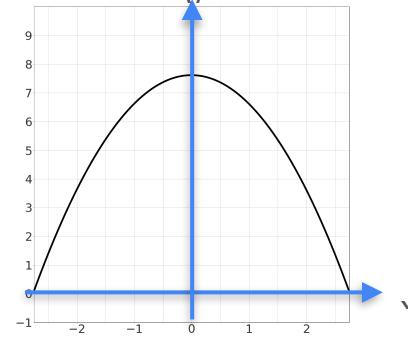
$$f'(0) < 0$$

Second derivative



Concave up

$$f''(0) > 0$$



Concave down

$$f''(0) < 0$$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## The Hessian

# Second Derivative

# Second Derivative

1 variable

2 variables

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t $x$

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t $x$ $f_y(x, y)$ Rate of change w.r.t $y$

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ $f_y(x, y)$ $\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$

# Second Derivative

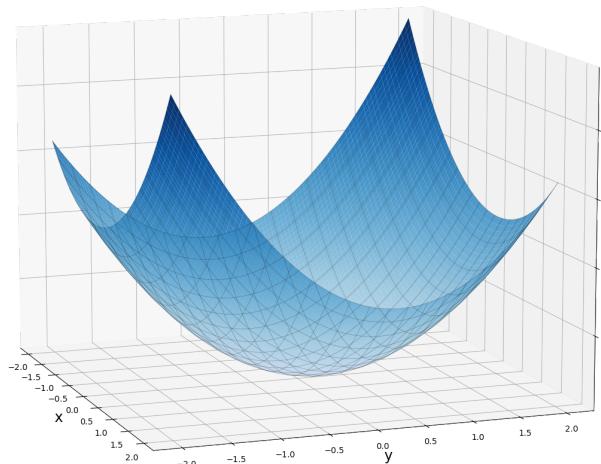
	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t $x$ $f_y(x, y)$ Rate of change w.r.t $y$ $\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$
Second derivative	$f''(x)$ Rate of change of the rate of change of $f(x)$	

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t $x$ $f_y(x, y)$ Rate of change w.r.t $y$ $\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$
Second derivative	$f''(x)$ Rate of change of the rate of change of $f(x)$	???

# Second Derivative

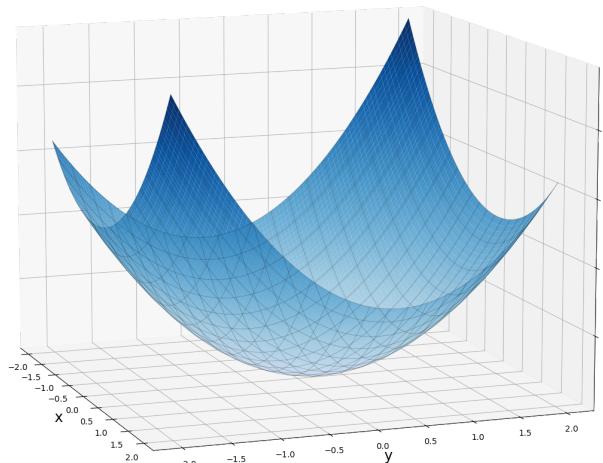
# Second Derivative



$$f(x, y) =$$

$$2x^2 + 3y^2 - xy$$

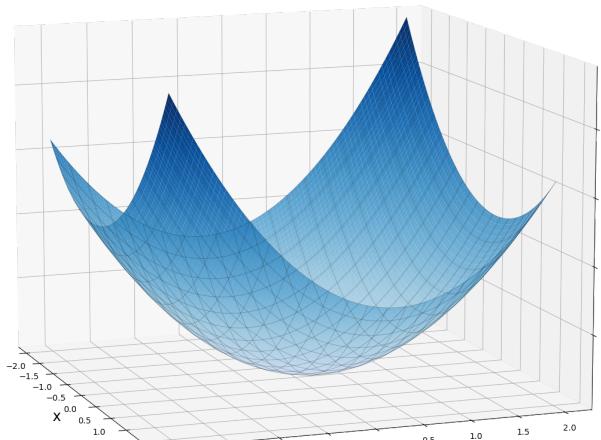
# Second Derivative



$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$\begin{matrix} 4x - y \\ x \end{matrix}$$

# Second Derivative

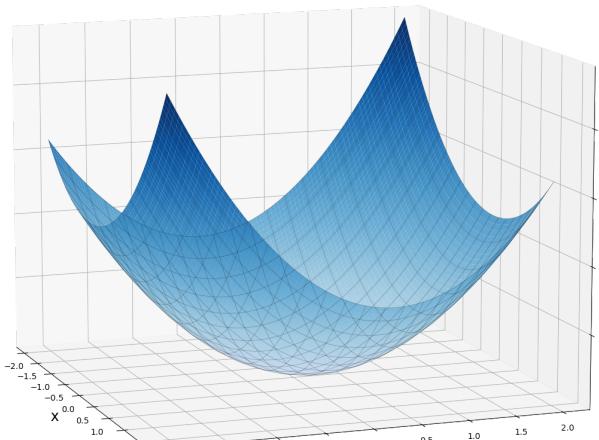


$$f(x, y) = 2x^2 + 3y^2 - xy$$

A diagram illustrating the second derivatives of the function  $f(x, y) = 2x^2 + 3y^2 - xy$ . A central point is connected by three arrows to the terms  $4x - y$ ,  $x$ , and  $6y - x$ .

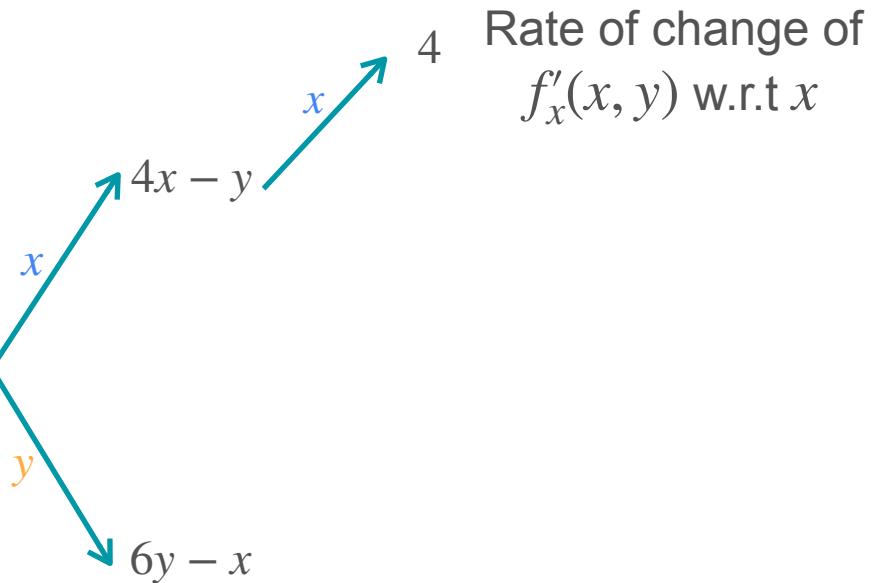
The term  $4x - y$  is associated with the  $x$ -axis direction,  $x$  is associated with the central point, and  $6y - x$  is associated with the  $y$ -axis direction.

# Second Derivative



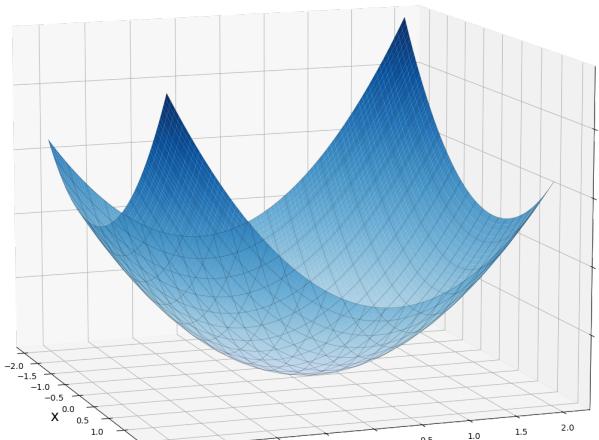
$$f(x, y) =$$

$$2x^2 + 3y^2 - xy$$



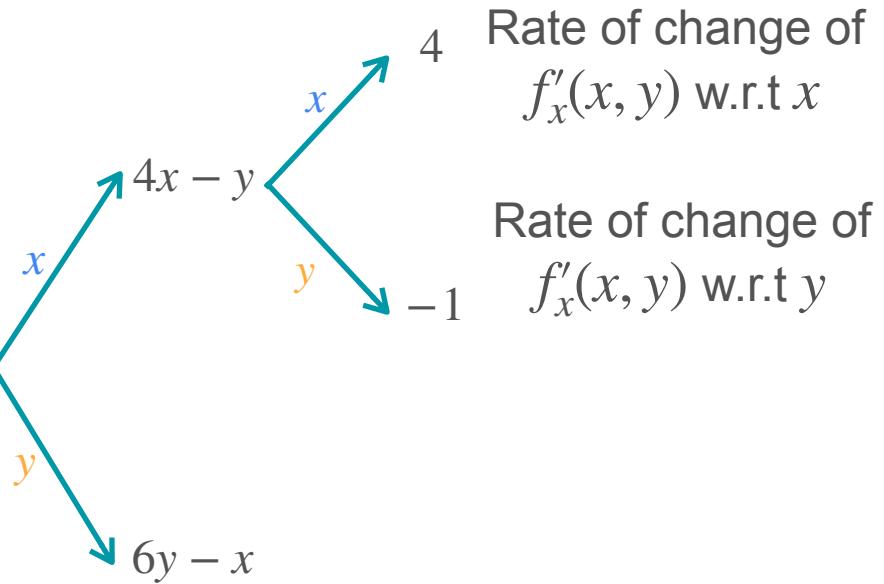
Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

# Second Derivative

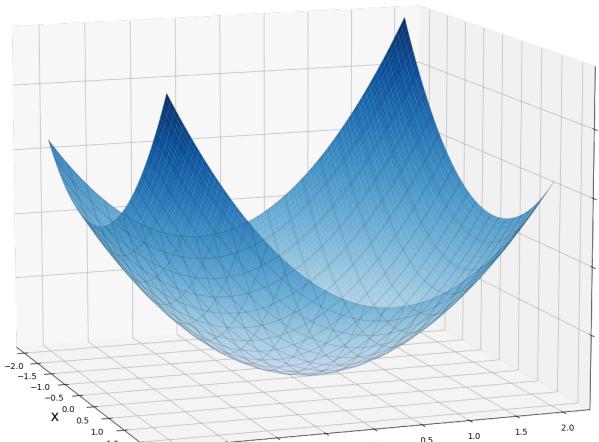


$$f(x, y) =$$

$$2x^2 + 3y^2 - xy$$

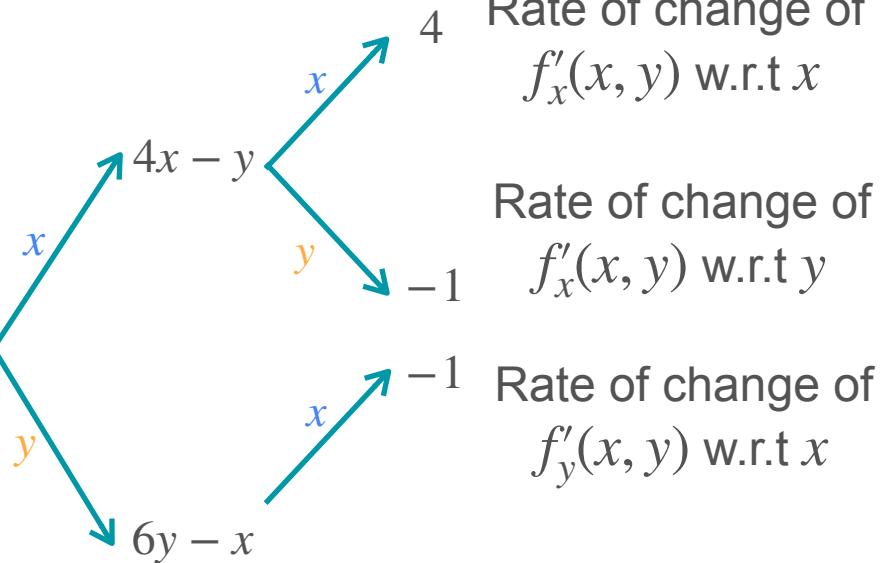


# Second Derivative

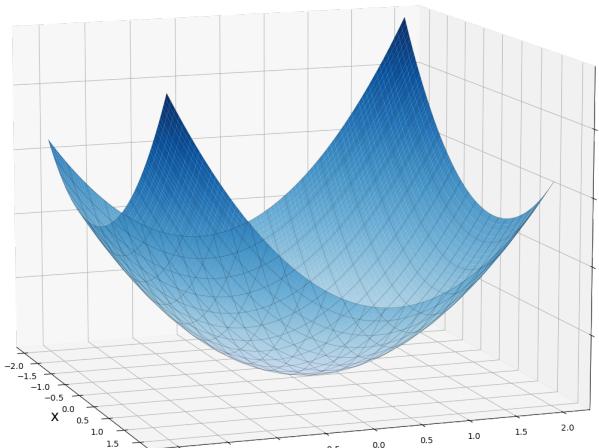


$$f(x, y) =$$

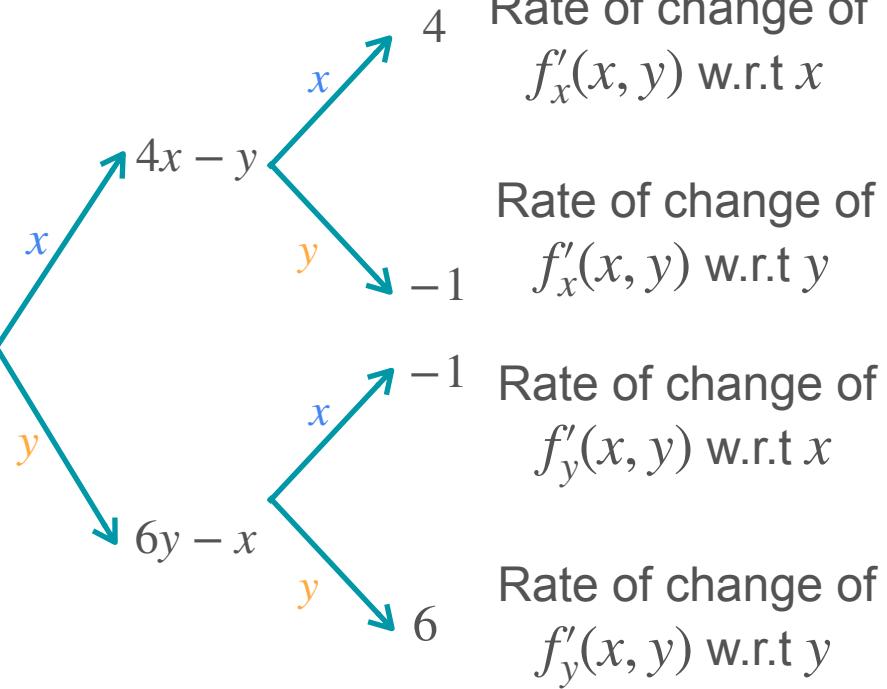
$$2x^2 + 3y^2 - xy$$



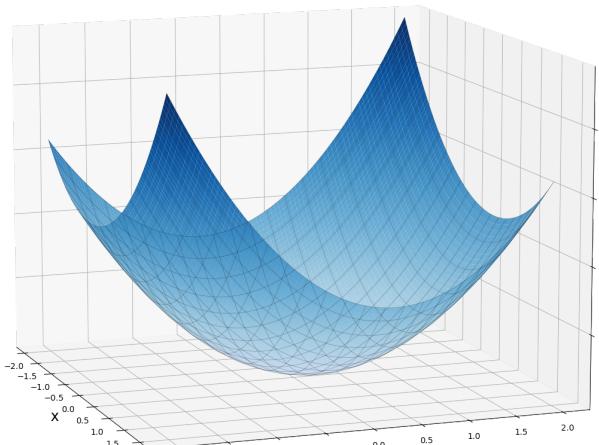
# Second Derivative



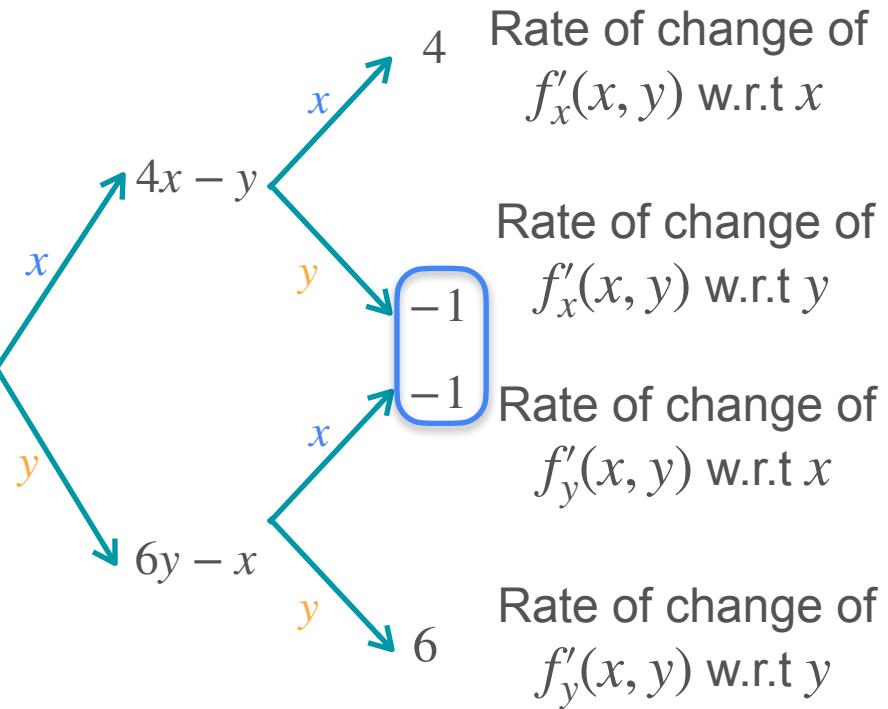
$$f(x, y) =$$
$$2x^2 + 3y^2 - xy$$



# Second Derivative



$$f(x, y) =$$
$$2x^2 + 3y^2 - xy$$



# What Do These Mean?

Rate of change of  
 $f_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f_y(x, y)$  w.r.t  $x$

# What Do These Mean?

Rate of change of  
 $f_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f_y(x, y)$  w.r.t  $y$

Change in the change in the function  
w.r.t tiny changes in  $x$  and  $y$

Rate of change of  
 $f_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f_y(x, y)$  w.r.t  $x$

# What Do These Mean?

Rate of change of

$f_x(x, y)$  w.r.t  $x$

Rate of change of

$f_y(x, y)$  w.r.t  $y$

Rate of change of

$f_x(x, y)$  w.r.t  $y$

Rate of change of

$f_y(x, y)$  w.r.t  $x$

Change in the change in the function  
w.r.t tiny changes in  $x$  and  $y$

Same idea as  
with one  
variable!

1. Change in the slope along one coordinate axis w.r.t tiny changes along an orthogonal coordinate axis

# What Do These Mean?

Rate of change of

$f_x(x, y)$  w.r.t  $x$

Rate of change of

$f_y(x, y)$  w.r.t  $y$

Rate of change of

$f_x(x, y)$  w.r.t  $y$

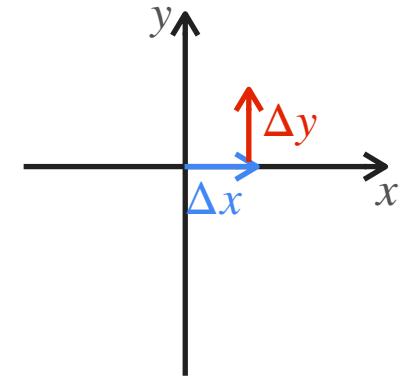
Rate of change of

$f_y(x, y)$  w.r.t  $x$

Change in the change in the function  
w.r.t tiny changes in  $x$  and  $y$

Same idea as  
with one  
variable!

1. Change in the slope along one coordinate axis w.r.t tiny changes along an orthogonal coordinate axis



# What Do These Mean?

Rate of change of

$f_x(x, y)$  w.r.t  $x$

Rate of change of

$f_y(x, y)$  w.r.t  $y$

Rate of change of

$f_x(x, y)$  w.r.t  $y$

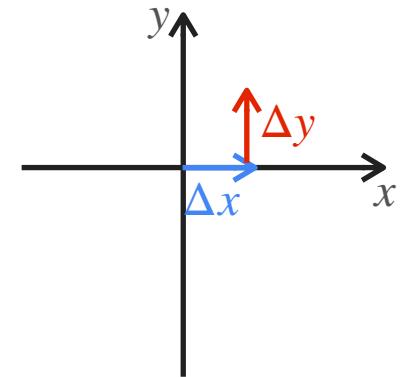
Rate of change of

$f_y(x, y)$  w.r.t  $x$

Change in the change in the function  
w.r.t tiny changes in  $x$  and  $y$

Same idea as  
with one  
variable!

1. Change in the slope along one coordinate axis w.r.t tiny changes along an orthogonal coordinate axis
2. They are the same!



# What Do These Mean?

Rate of change of  
 $f_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f_y(x, y)$  w.r.t  $y$

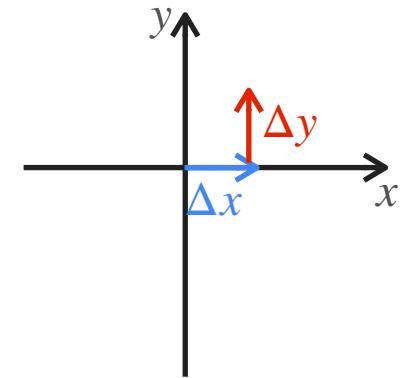
Change in the change in the function  
w.r.t tiny changes in  $x$  and  $y$

Rate of change of  
 $f_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f_y(x, y)$  w.r.t  $x$

1. Change in the slope along one coordinate axis w.r.t tiny changes along an orthogonal coordinate axis
2. They are the same!  
*(In most cases)*

Same idea as  
with one  
variable!



# Notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $x$

# Notation

## Leibniz's notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $x$

# Notation

## Leibniz's notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

$$\frac{\partial^2 f}{\partial x^2}$$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

$$\frac{\partial^2 f}{\partial y^2}$$

Rate of change of  
 $f'_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $x$

# Notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $x$

## Leibniz's notation

$$\frac{\partial^2 f}{\partial x^2}$$

$$\frac{\partial^2 f}{\partial y^2}$$

$$\frac{\partial^2 f}{\partial x \partial y}$$

$$\frac{\partial^2 f}{\partial y \partial x}$$

# Notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $x$

## Leibniz's notation

$$\frac{\partial^2 f}{\partial x^2}$$

$$\frac{\partial^2 f}{\partial y^2}$$

$$\frac{\partial^2 f}{\partial x \partial y}$$

$$\frac{\partial^2 f}{\partial y \partial x}$$

## Lagrange's notation

# Notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f''_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f''_y(x, y)$  w.r.t  $x$

## Leibniz's notation

$$\frac{\partial^2 f}{\partial x^2}$$

$$\frac{\partial^2 f}{\partial y^2}$$

$$\frac{\partial^2 f}{\partial x \partial y}$$

$$\frac{\partial^2 f}{\partial y \partial x}$$

## Lagrange's notation

$$f_{xx}(x, y)$$

$$f_{yy}(x, y)$$

# Notation

Rate of change of  
 $f'_x(x, y)$  w.r.t  $x$

Rate of change of  
 $f'_y(x, y)$  w.r.t  $y$

Rate of change of  
 $f''_x(x, y)$  w.r.t  $y$

Rate of change of  
 $f''_y(x, y)$  w.r.t  $x$

## Leibniz's notation

$$\frac{\partial^2 f}{\partial x^2}$$

$$\frac{\partial^2 f}{\partial y^2}$$

$$\frac{\partial^2 f}{\partial x \partial y}$$

$$\frac{\partial^2 f}{\partial y \partial x}$$

## Lagrange's notation

$$f_{xx}(x, y)$$

$$f_{yy}(x, y)$$

$$f_{xy}(x, y)$$

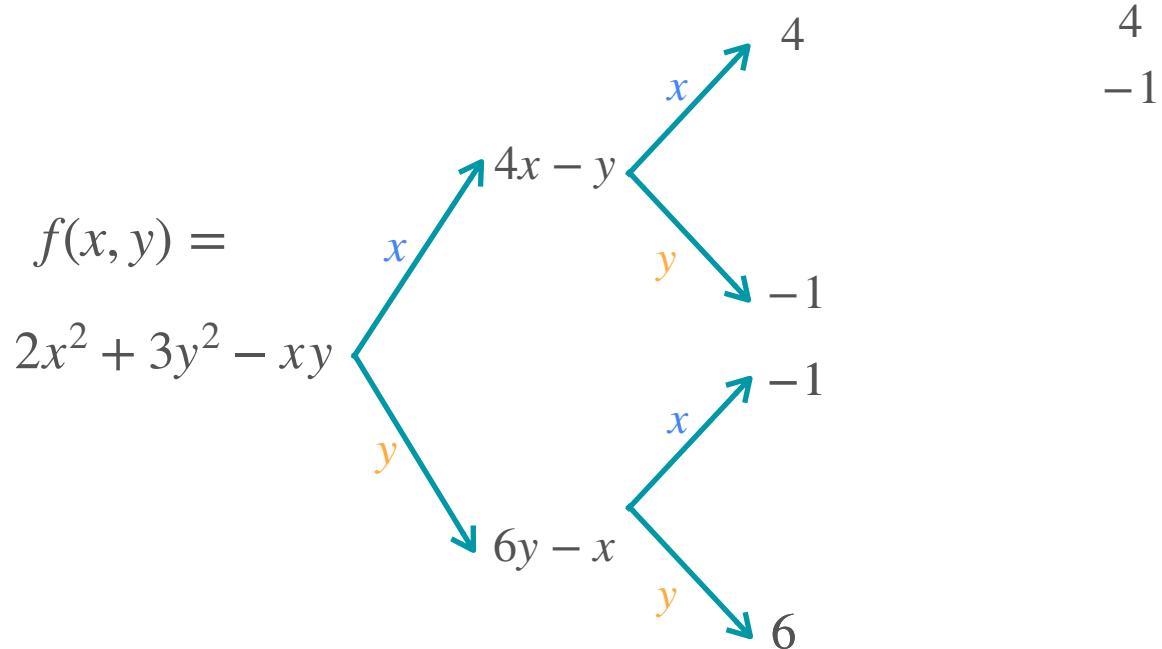
$$f_{yx}(x, y)$$

# Hessian Matrix

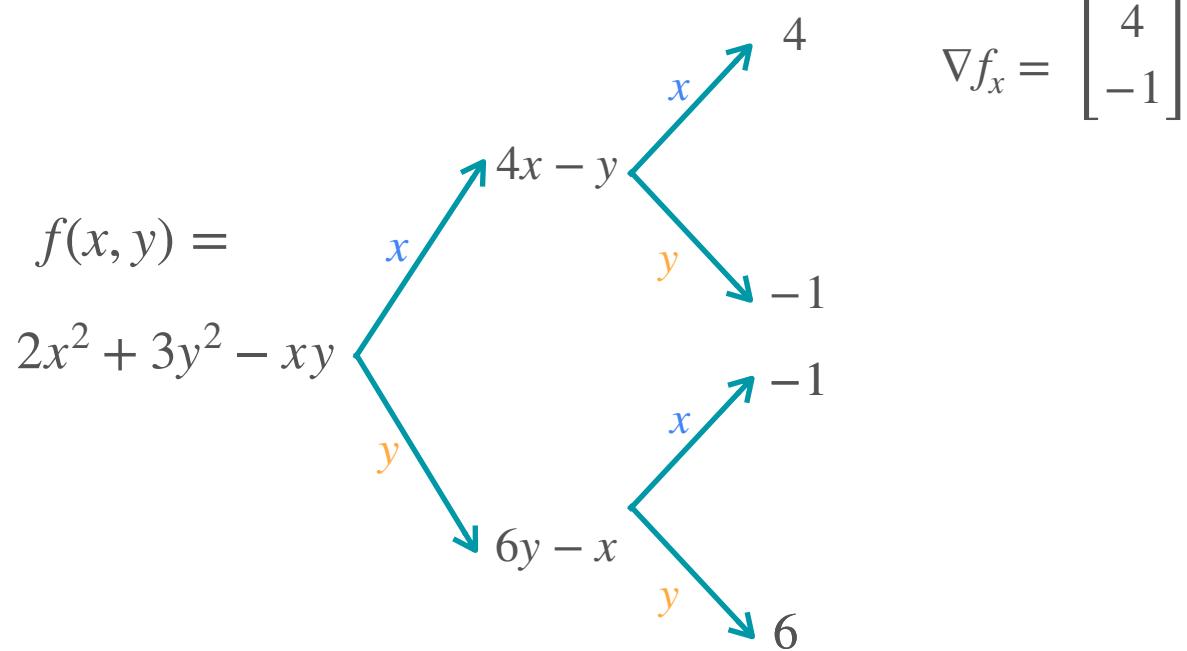
# Hessian Matrix

$$f(x, y) = 2x^2 + 3y^2 - xy$$
$$\begin{matrix} & \begin{matrix} 4 & \\ & -1 \end{matrix} \\ \begin{matrix} x \\ y \end{matrix} & \begin{pmatrix} 4x - y & \\ & 6y - x \end{pmatrix} \\ & \begin{matrix} -1 & \\ & 6 \end{matrix} \end{matrix}$$

# Hessian Matrix



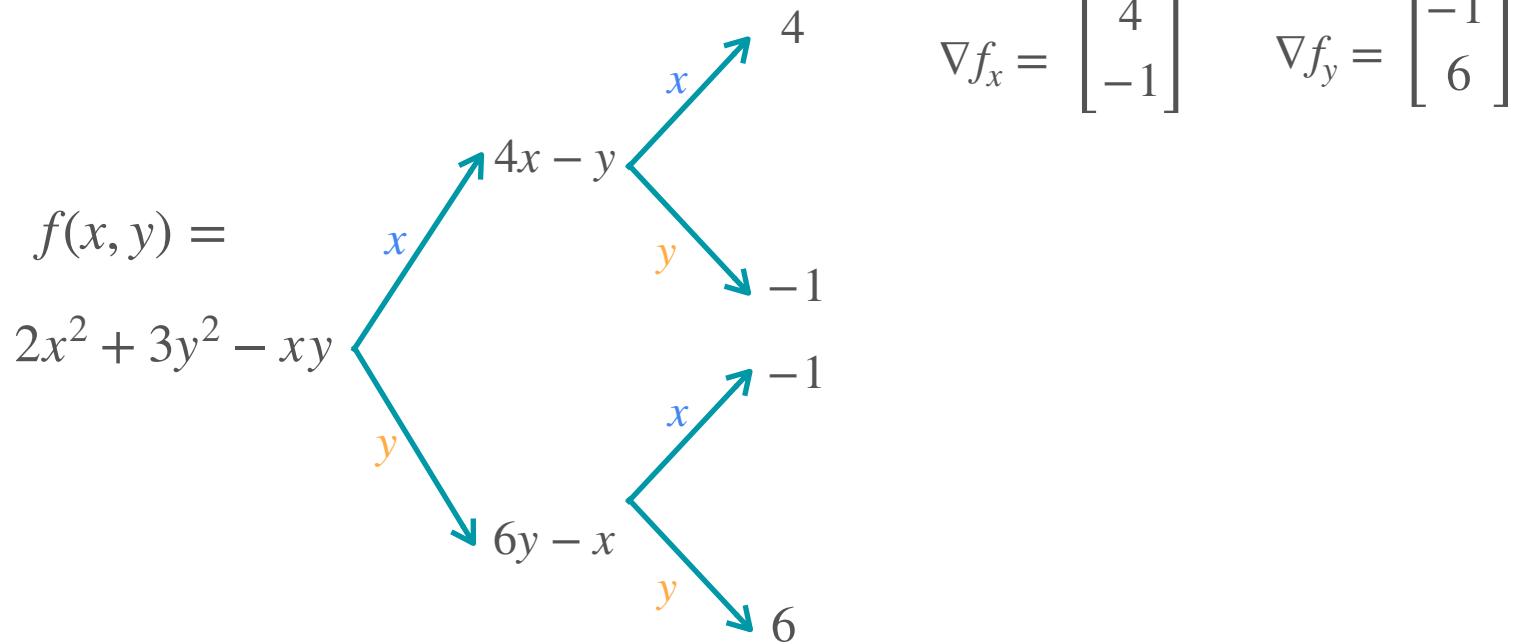
# Hessian Matrix



# Hessian Matrix

$$f(x, y) = 2x^2 + 3y^2 - xy$$
$$\nabla f_x = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$$
$$\begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

# Hessian Matrix



# Hessian Matrix

$$f(x, y) = 2x^2 + 3y^2 - xy$$

The Hessian matrix is represented by the following arrows originating from the point  $(4x-y, 6y-x)$ :

- Top-left arrow (pointing up-right):  $4$  (blue  $x$ , orange  $y$ )
- Top-right arrow (pointing down-right):  $-1$  (orange  $y$ )
- Bottom-left arrow (pointing down-left):  $-1$  (blue  $x$ , orange  $y$ )
- Bottom-right arrow (pointing down-right):  $6$  (orange  $y$ )

$$\nabla f_x = \begin{bmatrix} 4 \\ -1 \end{bmatrix} \quad \nabla f_y = \begin{bmatrix} -1 \\ 6 \end{bmatrix}$$

$$\begin{matrix} 4 & -1 \\ -1 & 6 \end{matrix}$$

# Hessian Matrix

$$f(x, y) = 2x^2 + 3y^2 - xy$$
$$\begin{matrix} & \begin{matrix} 4x - y & 4 \\ 6y - x & 6 \end{matrix} \\ \begin{matrix} x \\ y \end{matrix} & \end{matrix}$$

$$\nabla f_x = \begin{bmatrix} 4 \\ -1 \end{bmatrix} \quad \nabla f_y = \begin{bmatrix} -1 \\ 6 \end{bmatrix}$$

$$\begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix} = \begin{bmatrix} \nabla f_x^T \\ \nabla f_y^T \end{bmatrix}$$

# Hessian Matrix

$$f(x, y) = 2x^2 + 3y^2 - xy$$

Diagram illustrating the second partial derivatives of the function  $f(x, y) = 2x^2 + 3y^2 - xy$ . The function value is at the center. Four arrows point outwards from the center, each labeled with a second-order derivative:

- Top-left arrow:  $4x - y$  (x up, y down)
- Top-right arrow:  $4$  (x up, y up)
- Bottom-left arrow:  $6y - x$  (x down, y up)
- Bottom-right arrow:  $6$  (x down, y down)

$$\nabla f_x = \begin{bmatrix} 4 \\ -1 \end{bmatrix} \quad \nabla f_y = \begin{bmatrix} -1 \\ 6 \end{bmatrix}$$

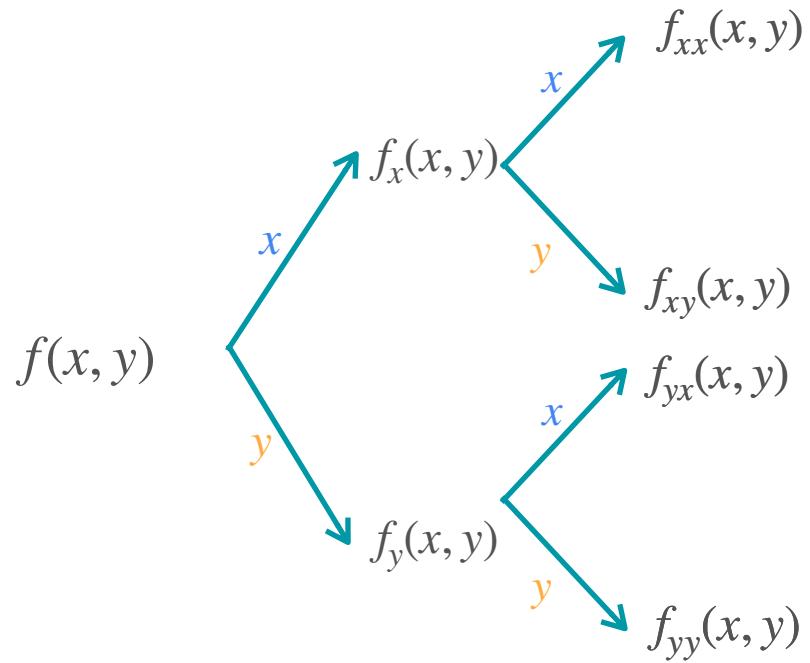
$$H = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix} = \begin{bmatrix} \nabla f_x^T \\ \nabla f_y^T \end{bmatrix}$$

**Hessian  
matrix**

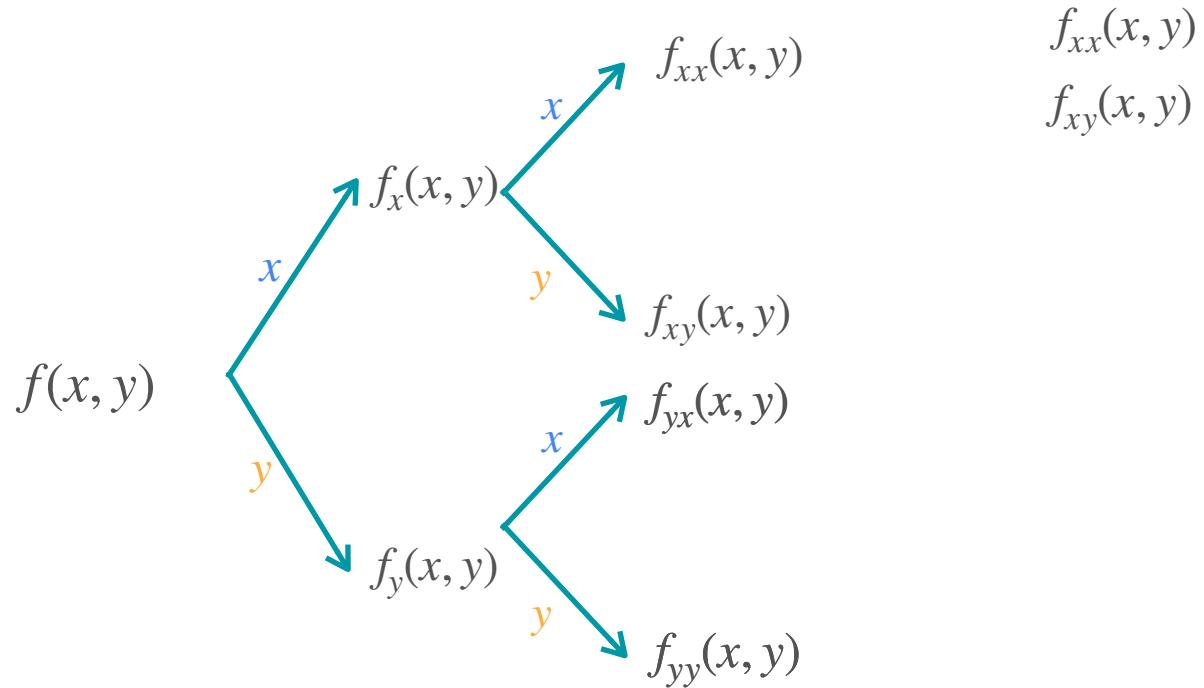
All information  
about second  
derivatives

# Hessian Matrix - General Case

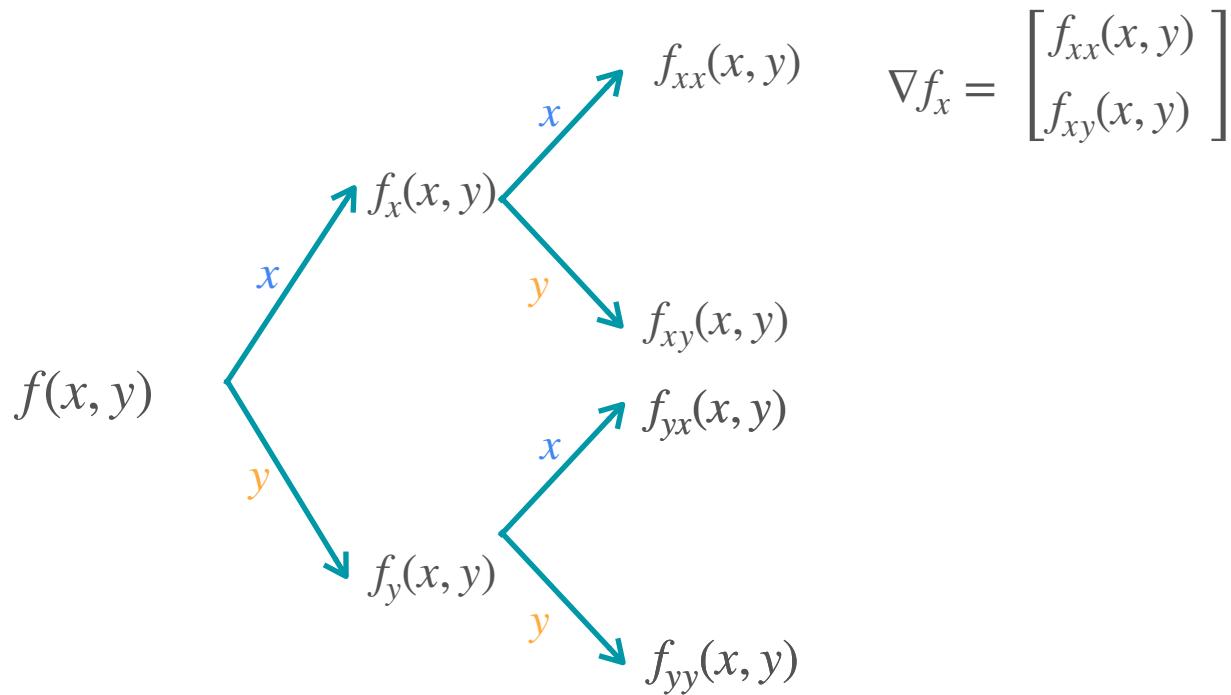
# Hessian Matrix - General Case



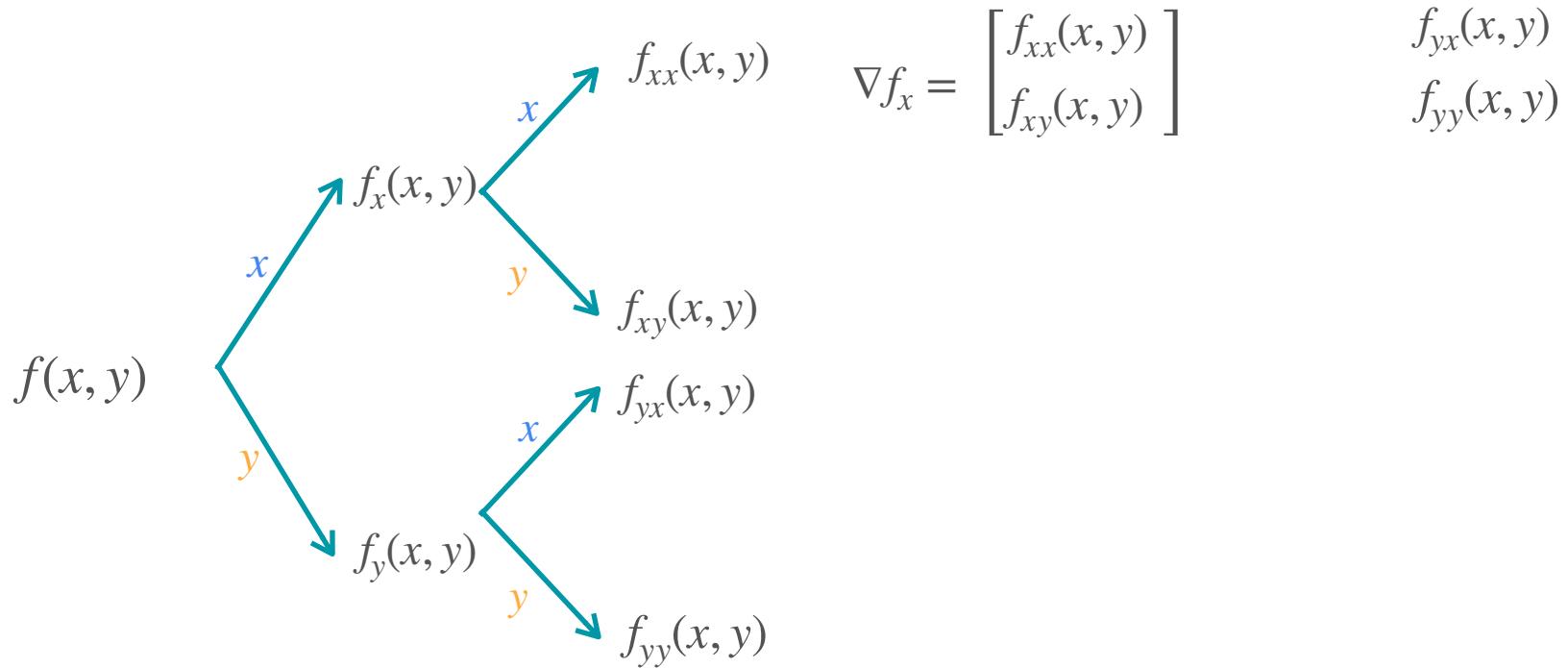
# Hessian Matrix - General Case



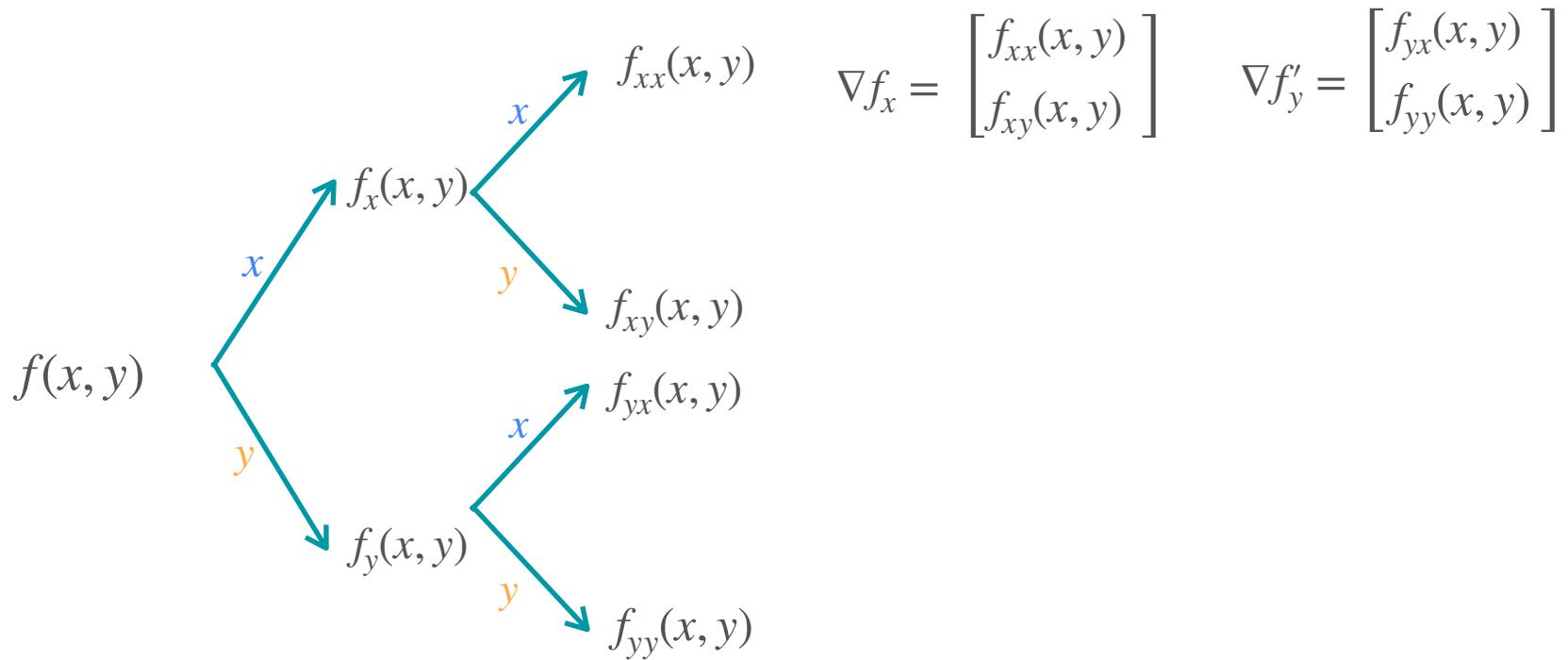
# Hessian Matrix - General Case



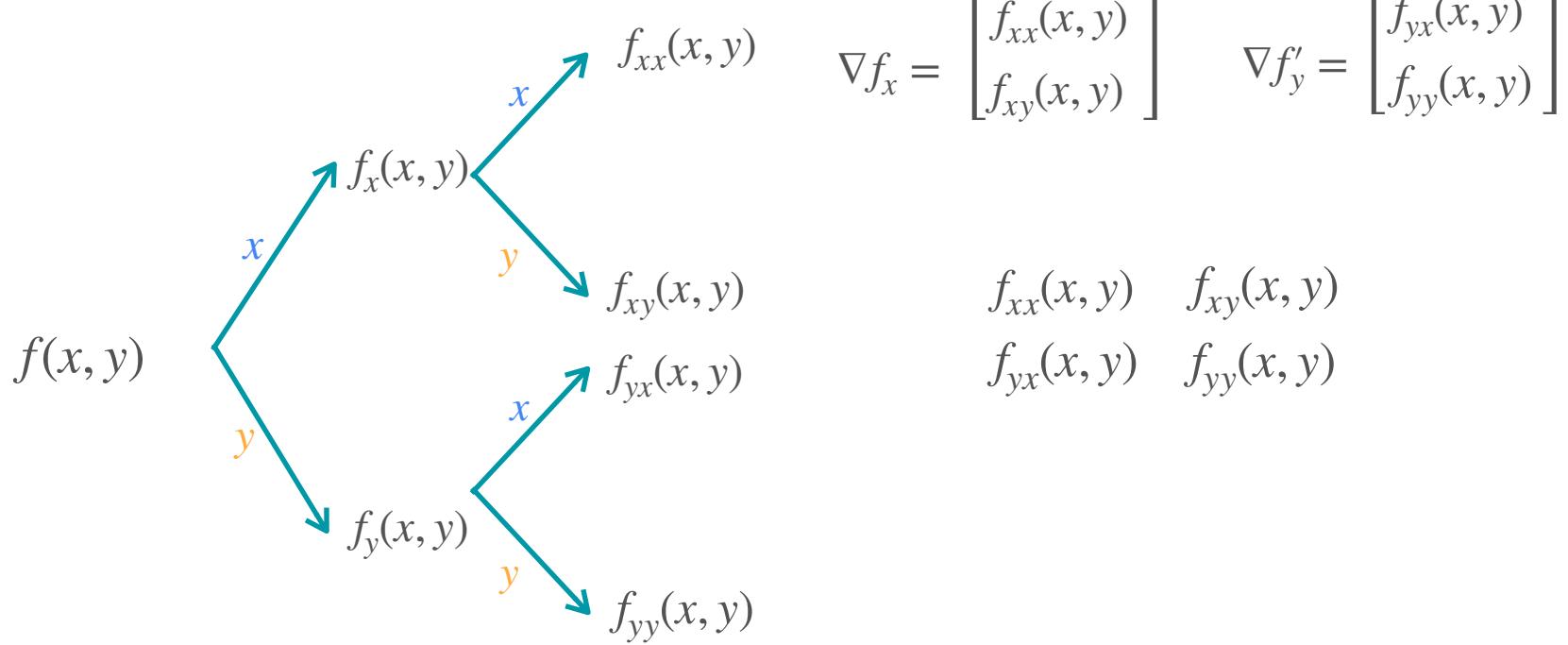
# Hessian Matrix - General Case



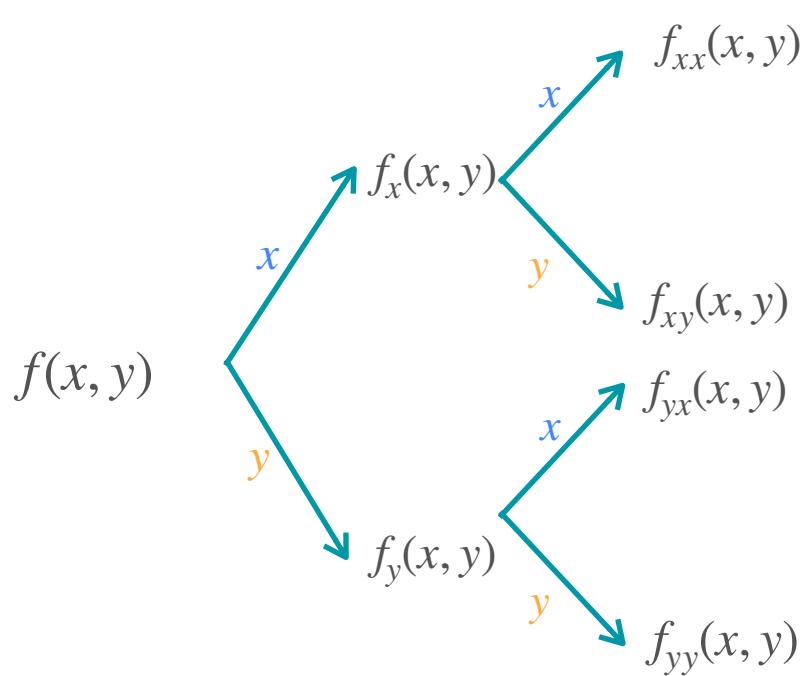
# Hessian Matrix - General Case



# Hessian Matrix - General Case



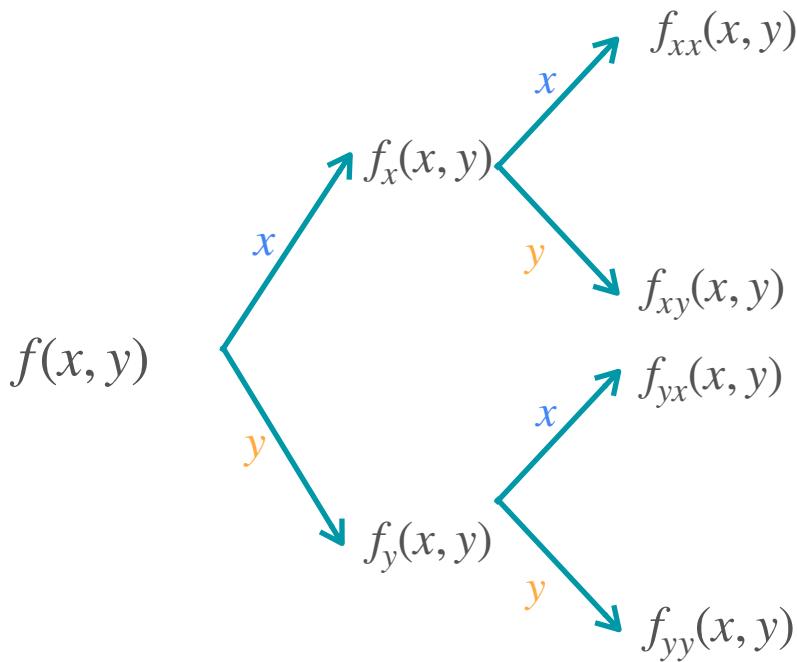
# Hessian Matrix - General Case



$$\nabla f_x = \begin{bmatrix} f_{xx}(x, y) \\ f_{xy}(x, y) \end{bmatrix} \quad \nabla f'_y = \begin{bmatrix} f_{yx}(x, y) \\ f_{yy}(x, y) \end{bmatrix}$$

$$\begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix} = \begin{bmatrix} \nabla f_x^T \\ \nabla f_y^T \end{bmatrix}$$

# Hessian Matrix - General Case



$$\nabla f_x = \begin{bmatrix} f_{xx}(x, y) \\ f_{xy}(x, y) \end{bmatrix} \quad \nabla f'_y = \begin{bmatrix} f_{yx}(x, y) \\ f_{yy}(x, y) \end{bmatrix}$$

$$H = \begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix} = \begin{bmatrix} \nabla f_x^T \\ \nabla f_y^T \end{bmatrix}$$

**Hessian  
matrix**

All information  
about second  
derivatives

# Second Derivative

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t $x$ $f_y(x, y)$ Rate of change w.r.t $y$ $\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$
Second derivative	$f''(x)$ Rate of change of the rate of change of $f(x)$	

# Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t $x$ $f_y(x, y)$ Rate of change w.r.t $y$ $\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$
Second derivative	$f''(x)$ Rate of change of the rate of change of $f(x)$	$H(x, y) = \begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix}$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

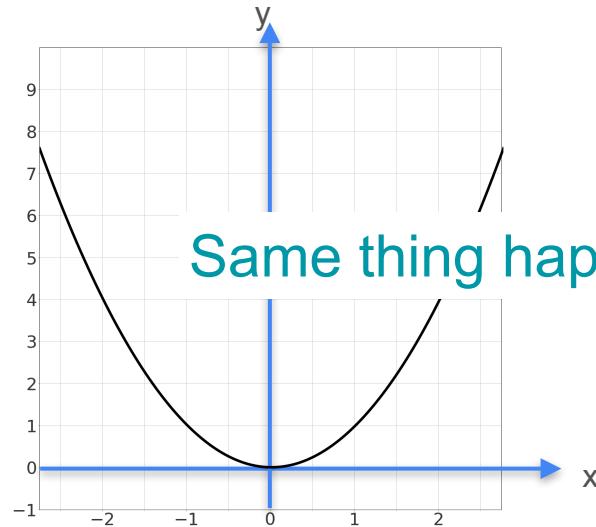
---

## Hessians and concavity

# Remember...

Same thing happens for many variables!

# Remember...

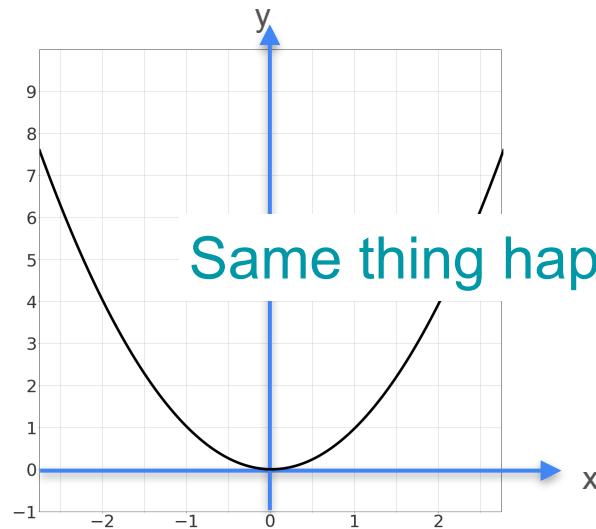


Same thing happens for many variables!

Concave up or convex

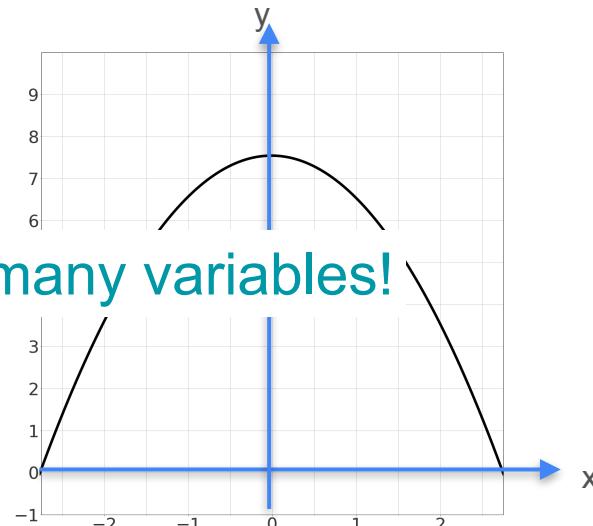
$$f''(0) > 0$$

# Remember...



Concave up or convex

$$f''(0) > 0$$

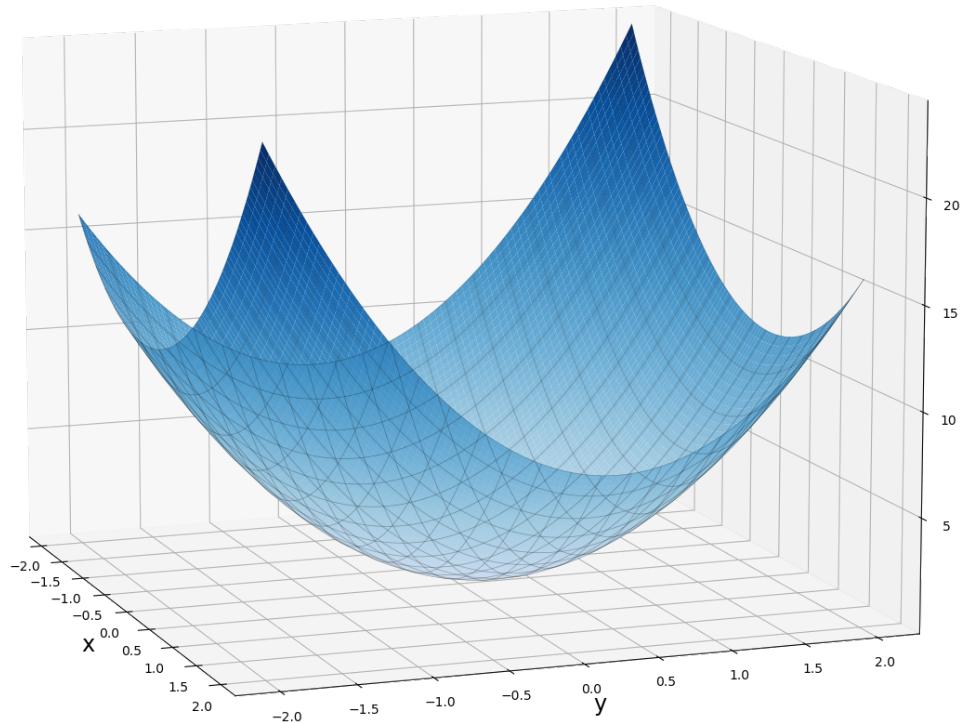


Concave down

$$f''(0) < 0$$

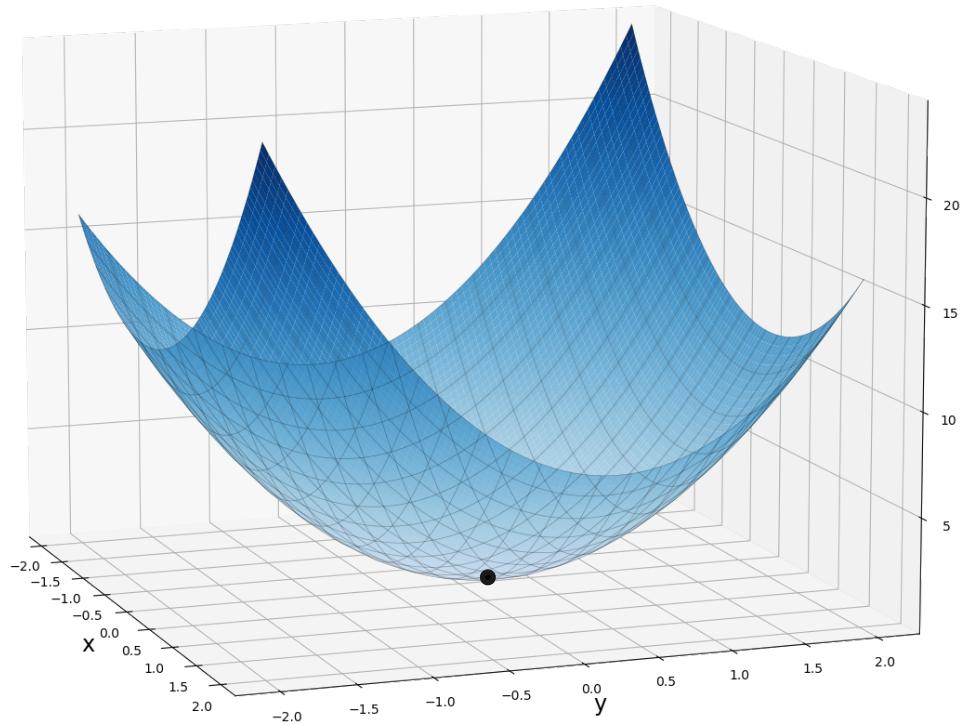
# Concave Up

# Concave Up



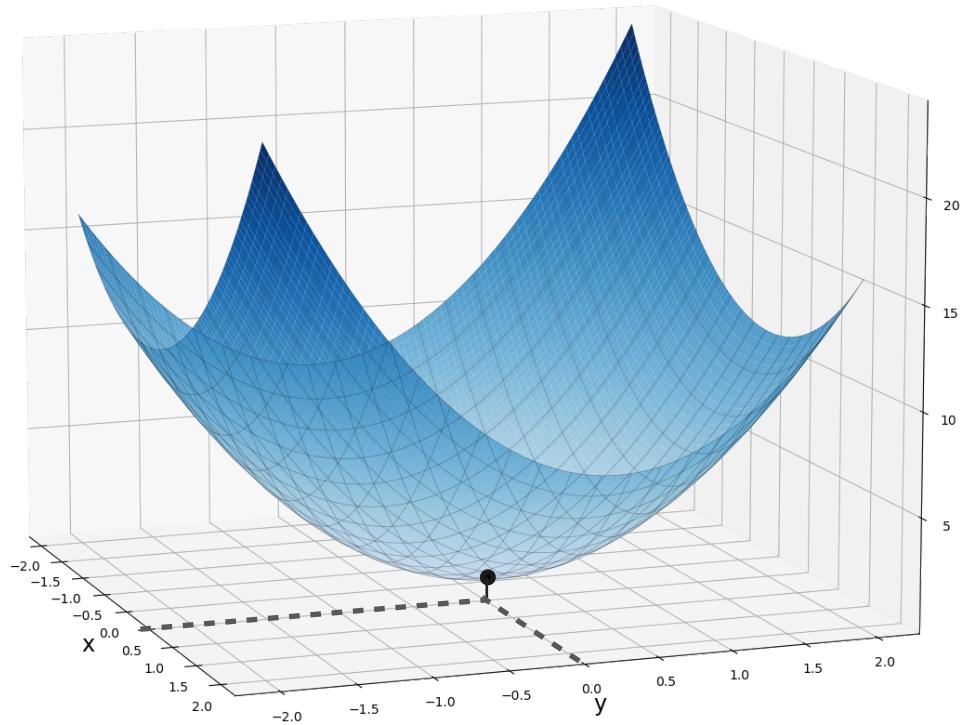
$$f(x, y) = 2x^2 + 3y^2 - xy$$

# Concave Up



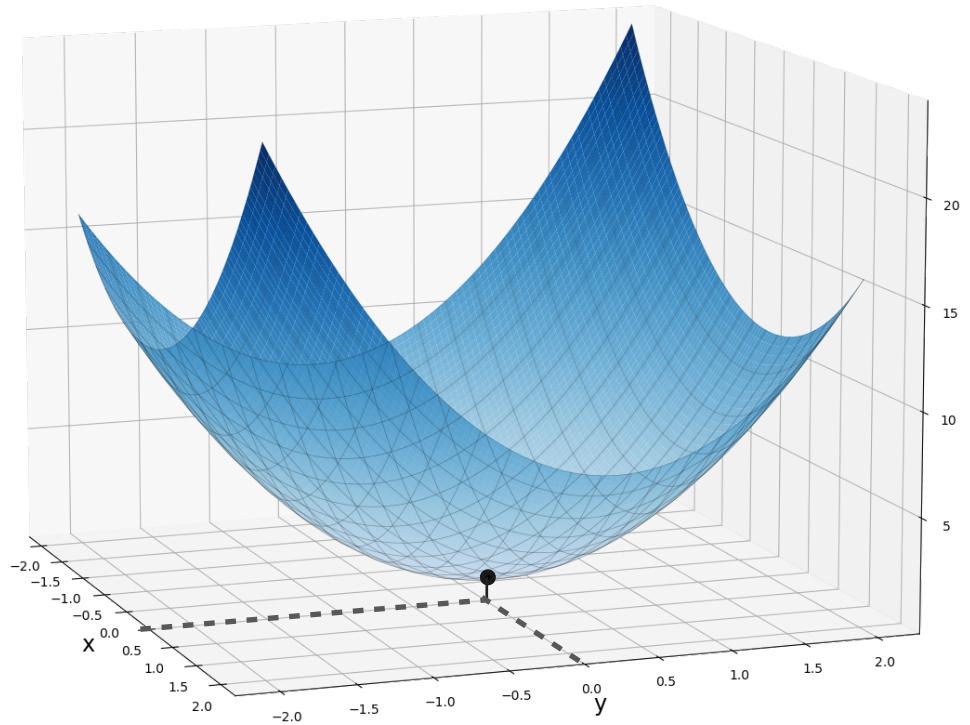
$$f(x, y) = 2x^2 + 3y^2 - xy$$

# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

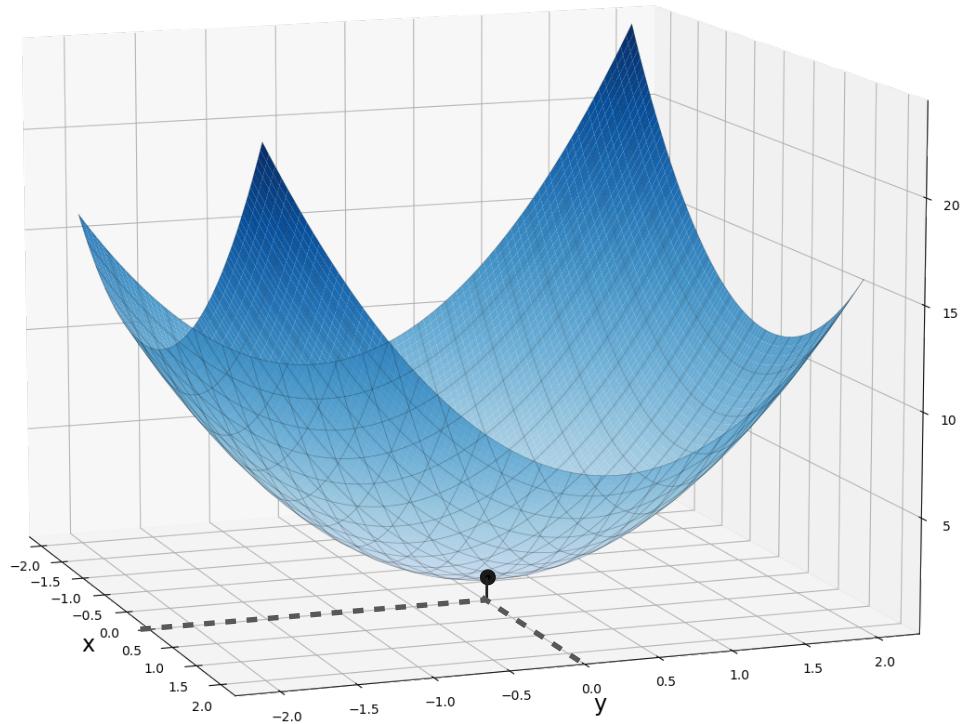
# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

# Concave Up

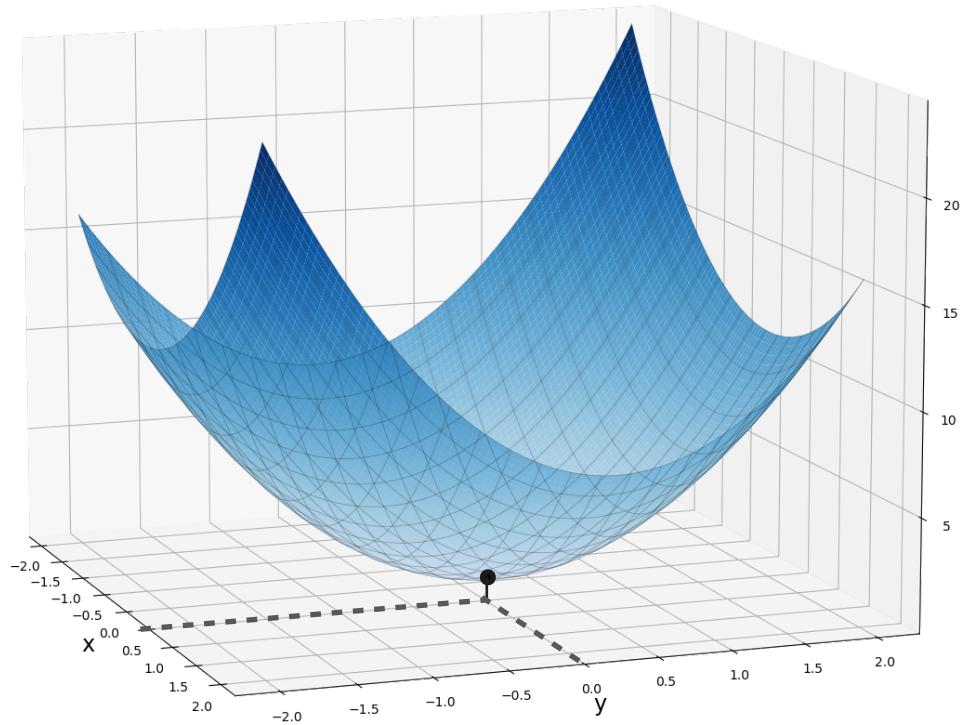


$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

# Concave Up

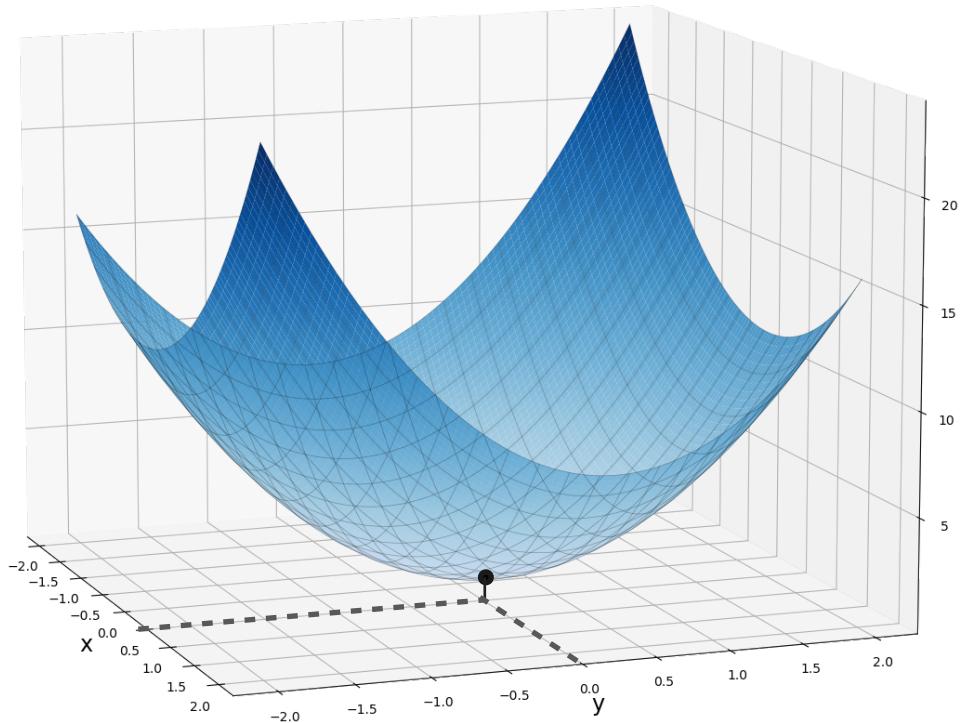


$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$

# Concave Up

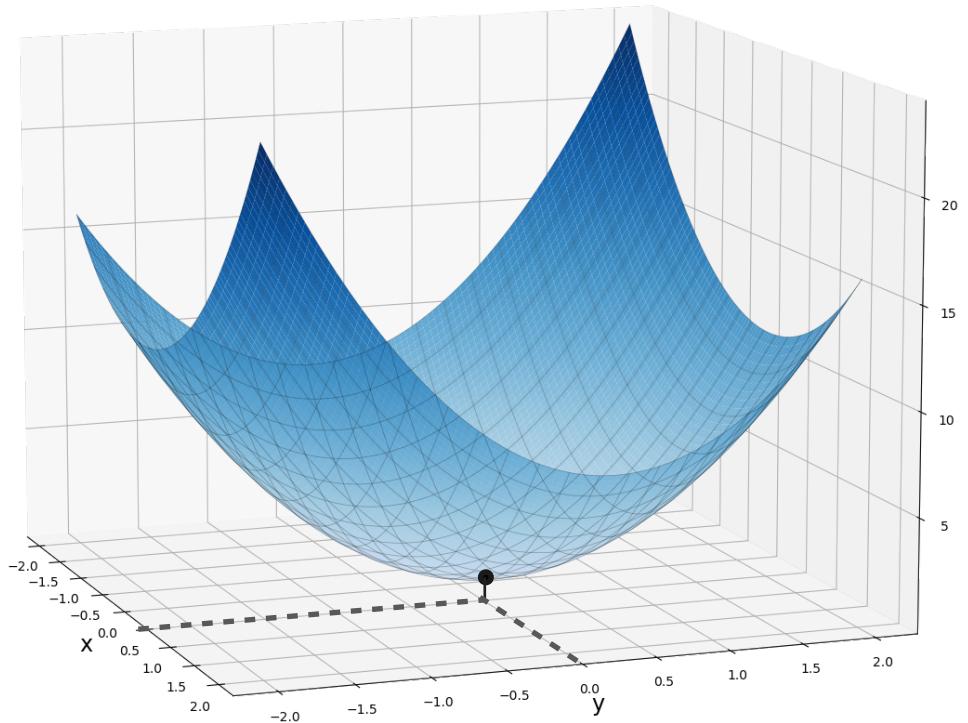


$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$
$$= (4 - \lambda)(6 - \lambda) - (-1)(-1)$$

# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

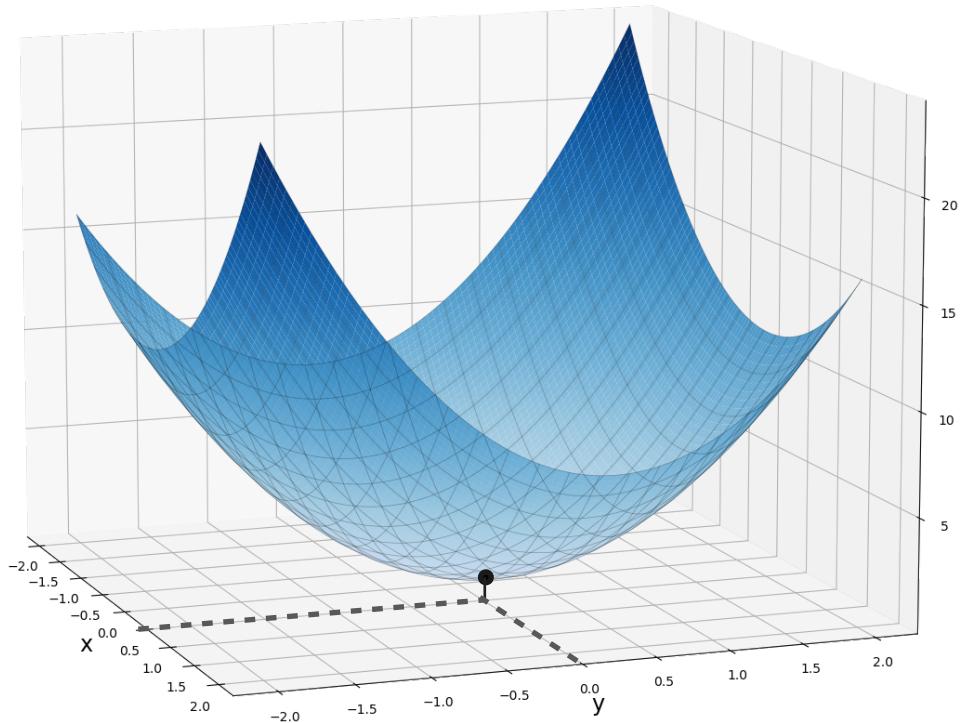
$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$

$$= (4 - \lambda)(6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 - 10\lambda + 23$$

# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

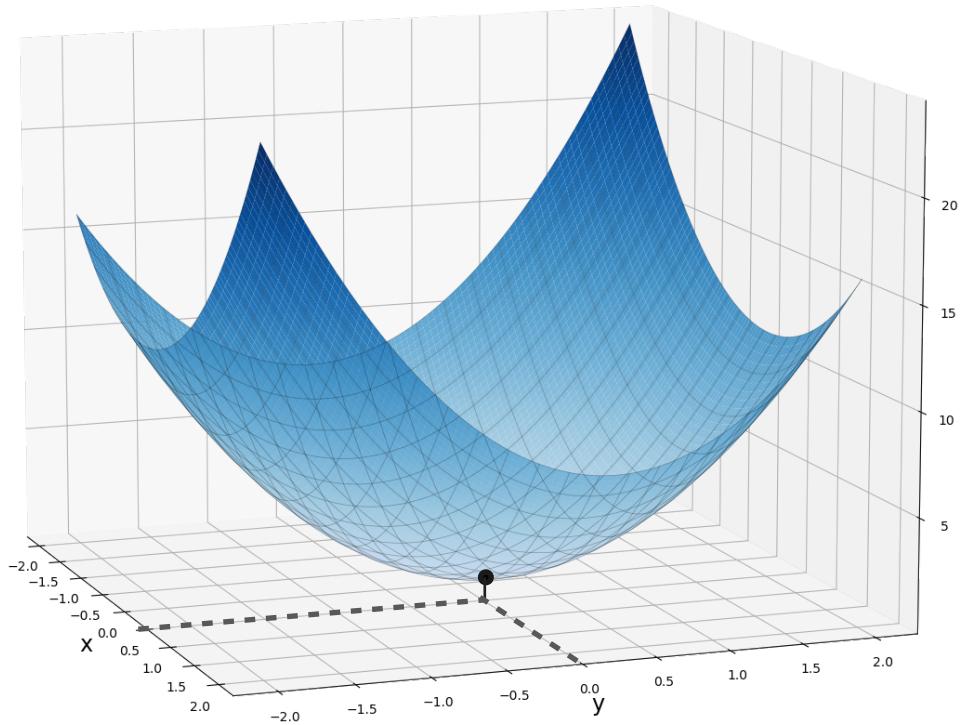
$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$

$$= (4 - \lambda)(6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 - 10\lambda + 23 \rightarrow \lambda_1 = 6.41$$

# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

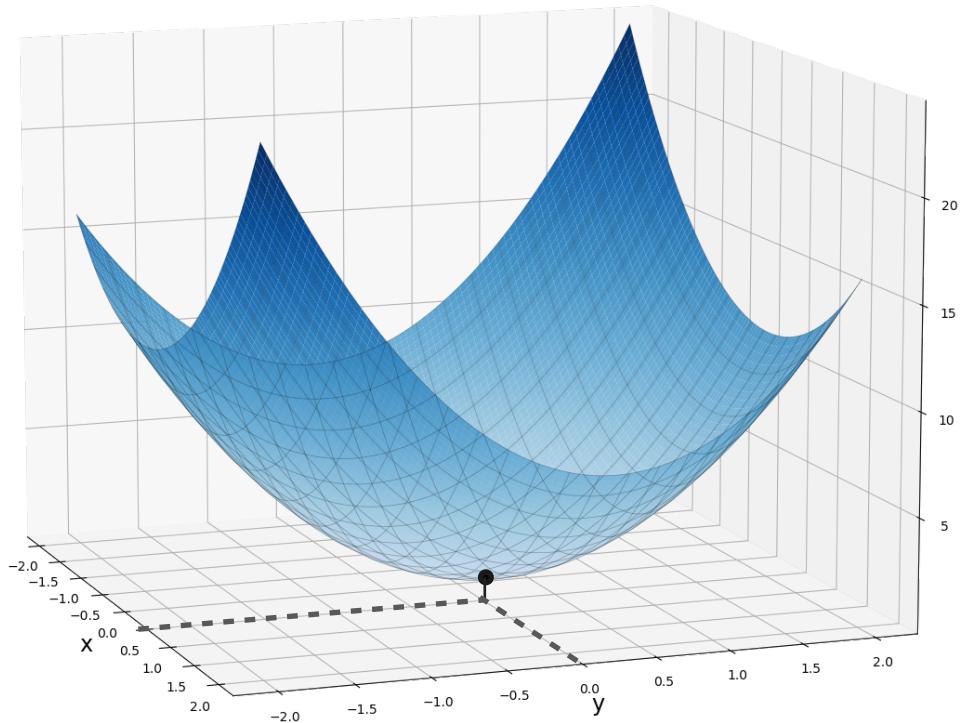
$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$

$$= (4 - \lambda)(6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 - 10\lambda + 23$$

$$\begin{array}{l} \xrightarrow{\text{blue}} \lambda_1 = 6.41 \\ \xrightarrow{\text{blue}} \lambda_2 = 3.59 \end{array}$$

# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$

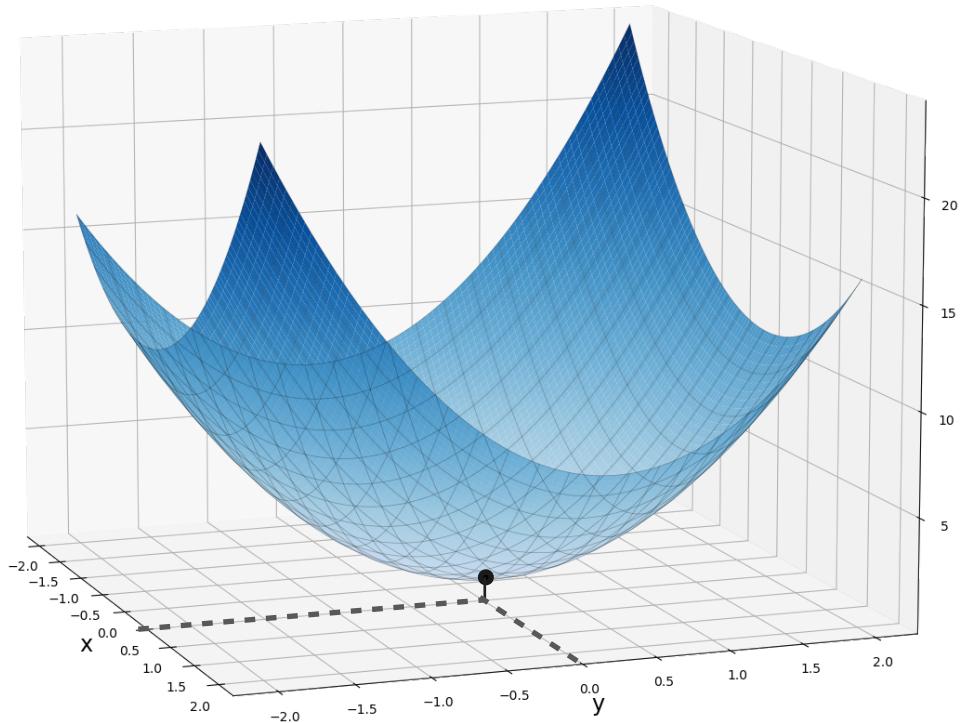
$$= (4 - \lambda)(6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 - 10\lambda + 23$$

$$\lambda_1 = 6.41$$
$$\lambda_2 = 3.59$$

$> 0$

# Concave Up



$$f(x, y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det \left( \begin{bmatrix} 4 - \lambda & -1 \\ -1 & 6 - \lambda \end{bmatrix} \right)$$

$$= (4 - \lambda)(6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 - 10\lambda + 23$$

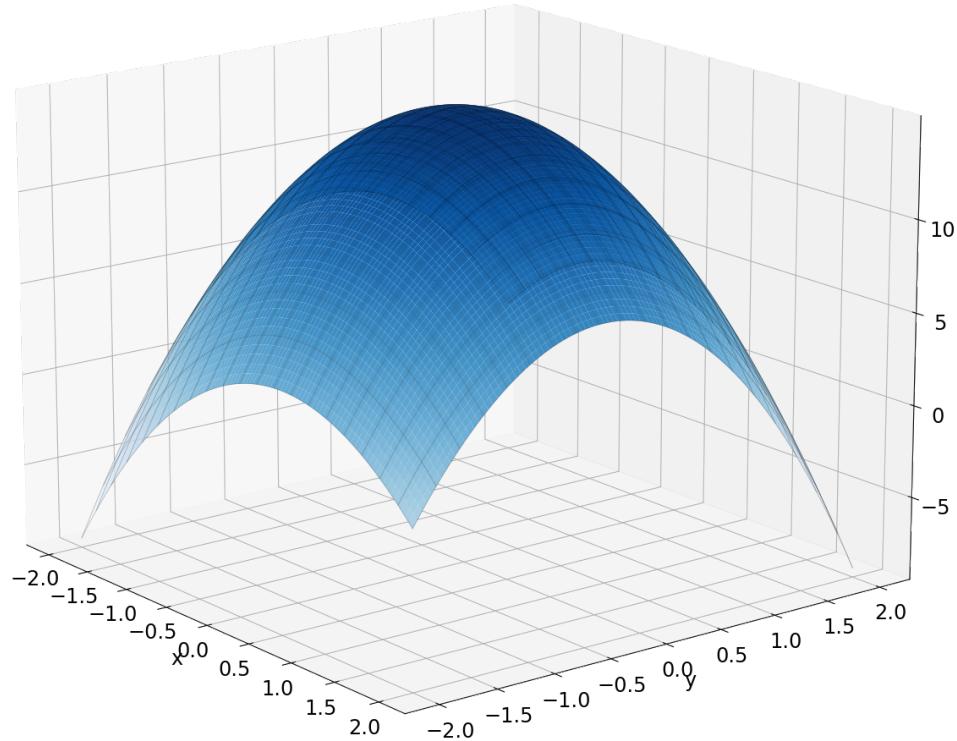
$$\lambda_1 = 6.41$$
$$\lambda_2 = 3.59$$

(0,0) is a minimum!

> 0

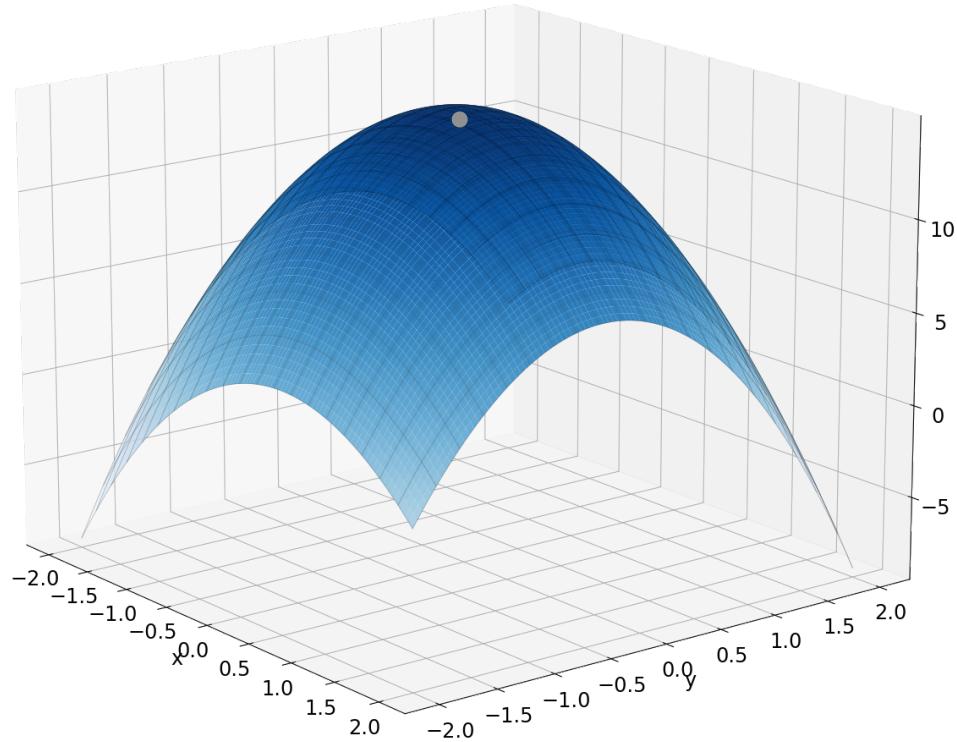
# Concave Down

# Concave Down



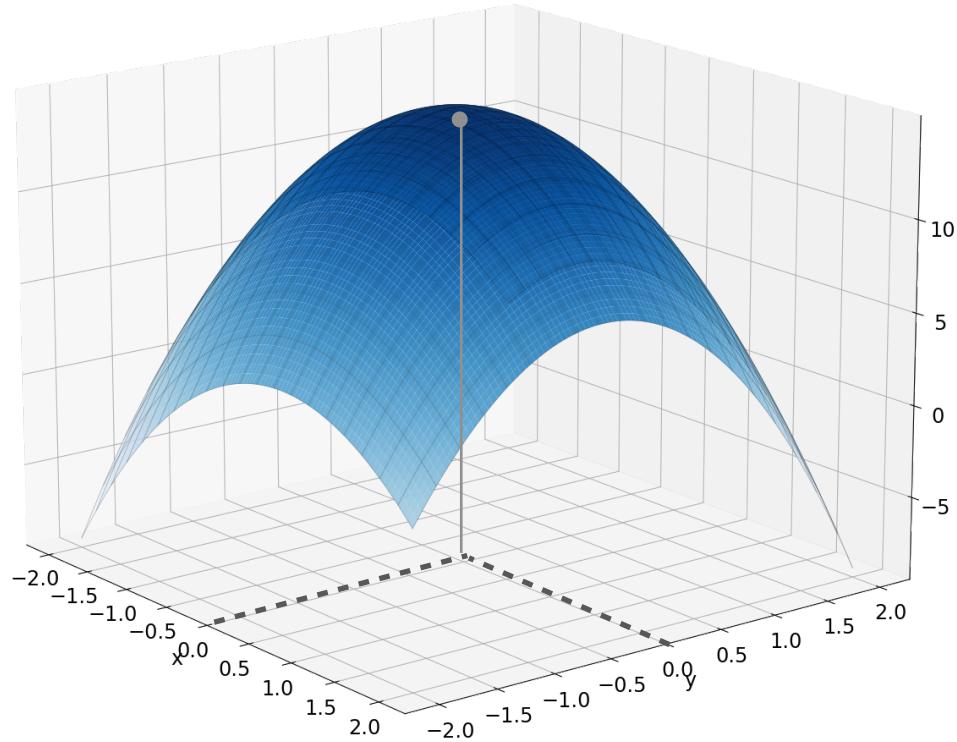
$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

# Concave Down



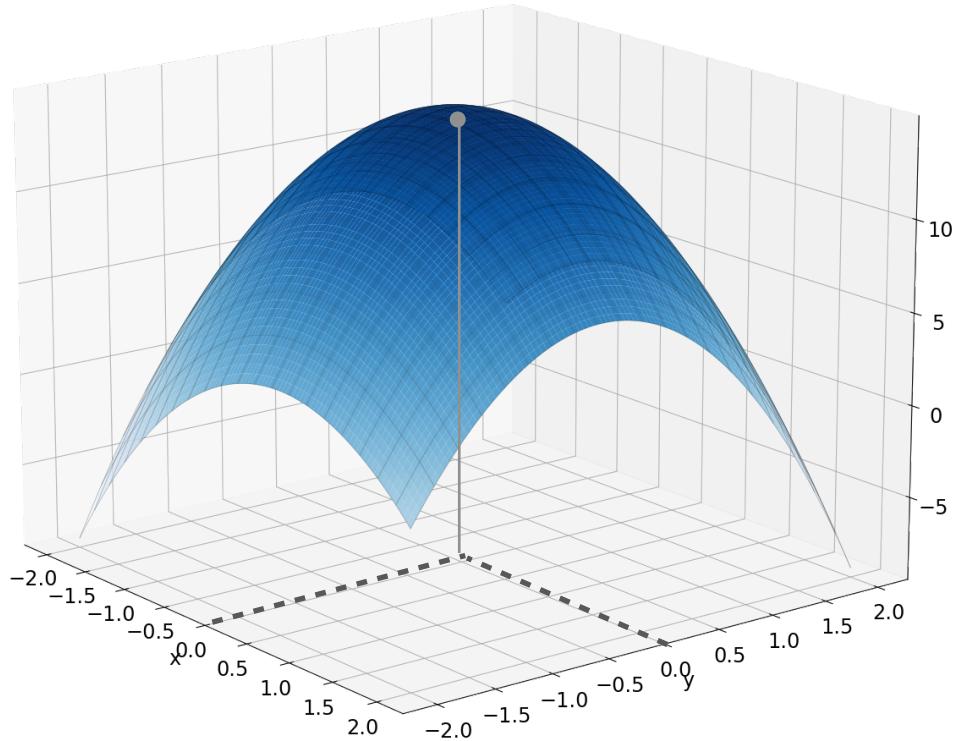
$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

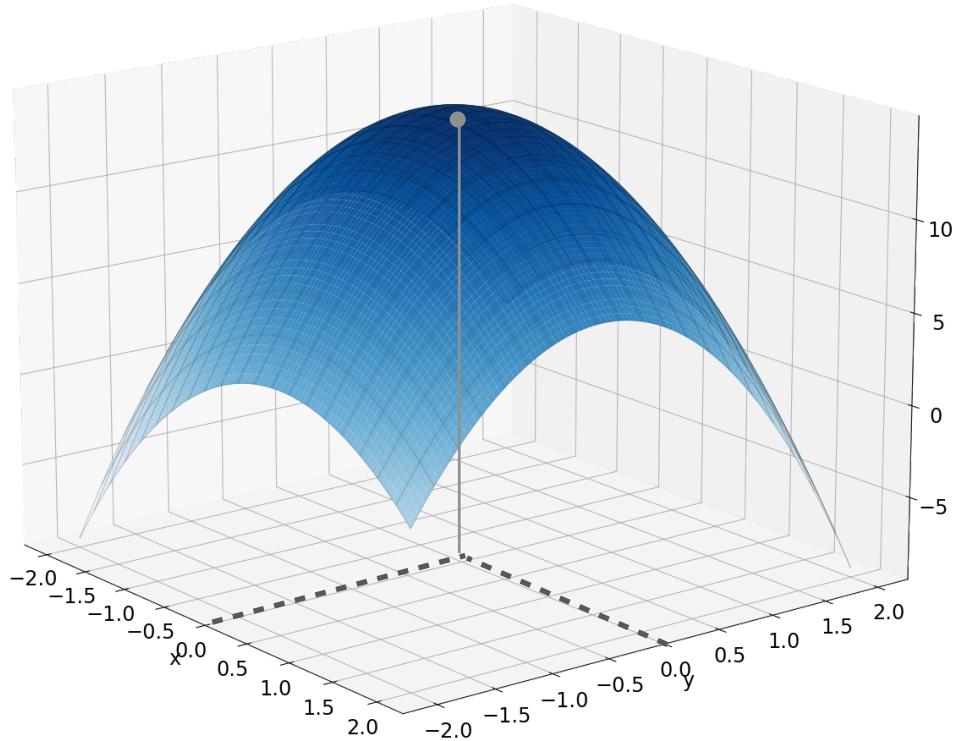
# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

# Concave Down

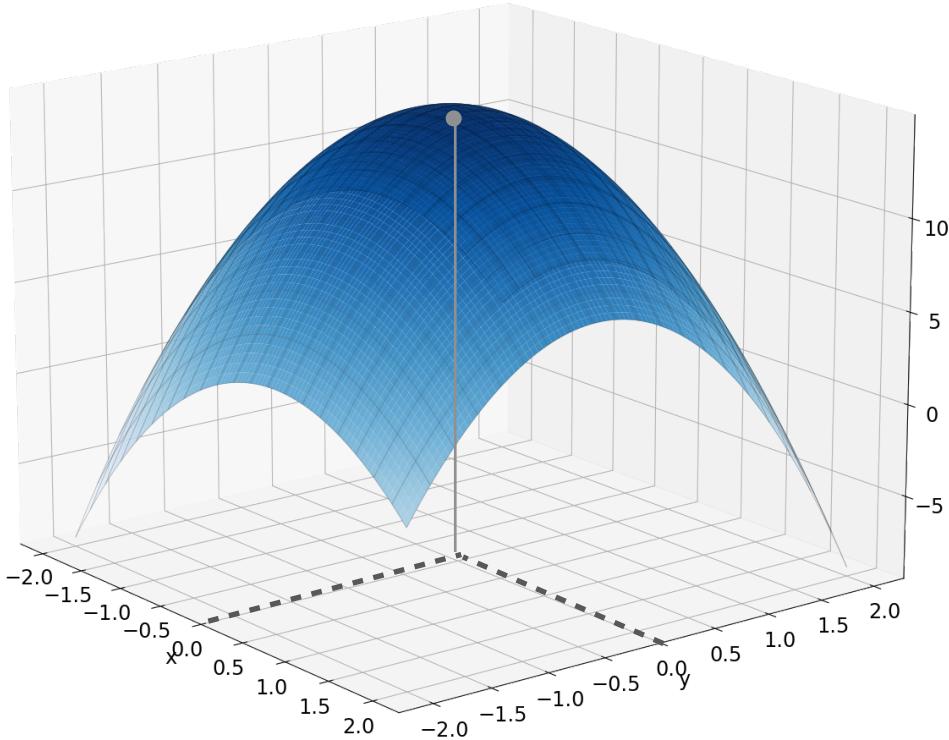


$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

# Concave Down



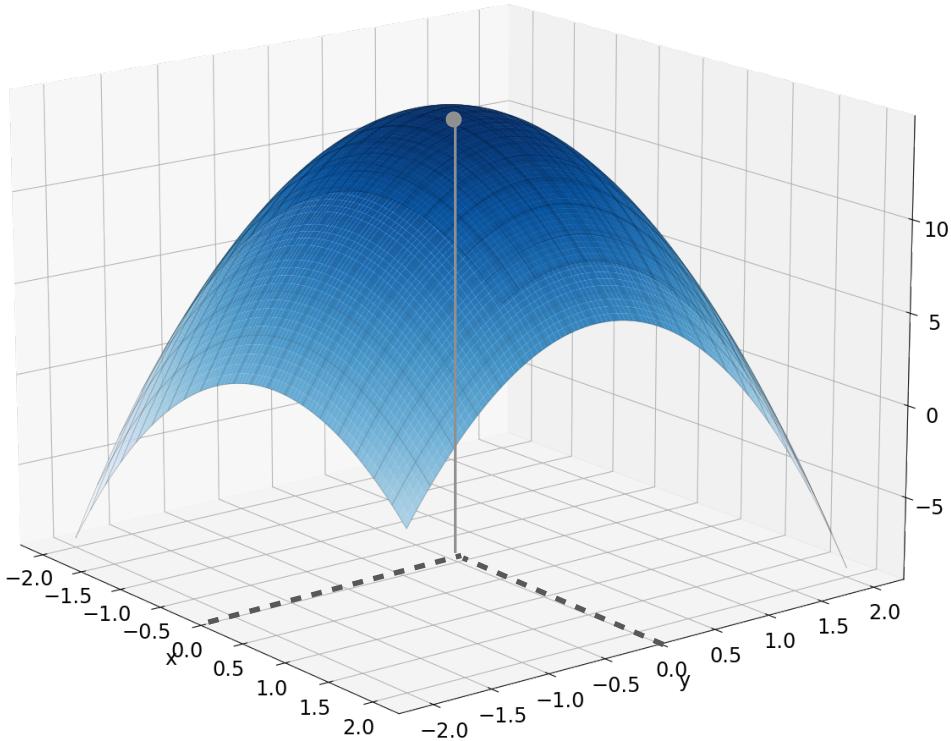
$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

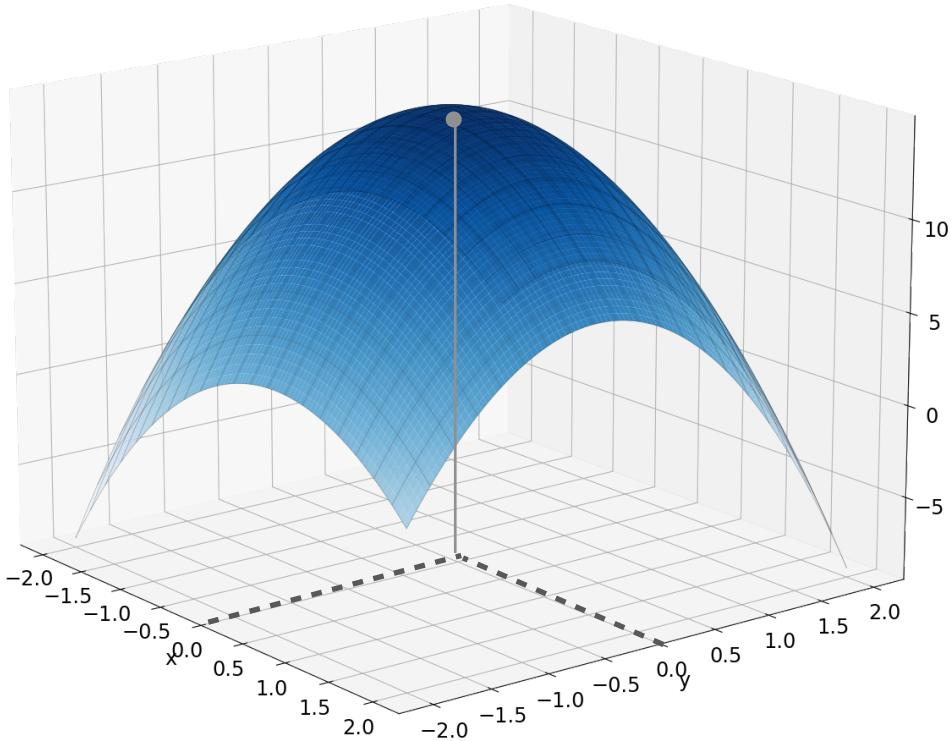
$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(-4 - \lambda)(-6 - \lambda) - (-1)(-1)$$

# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

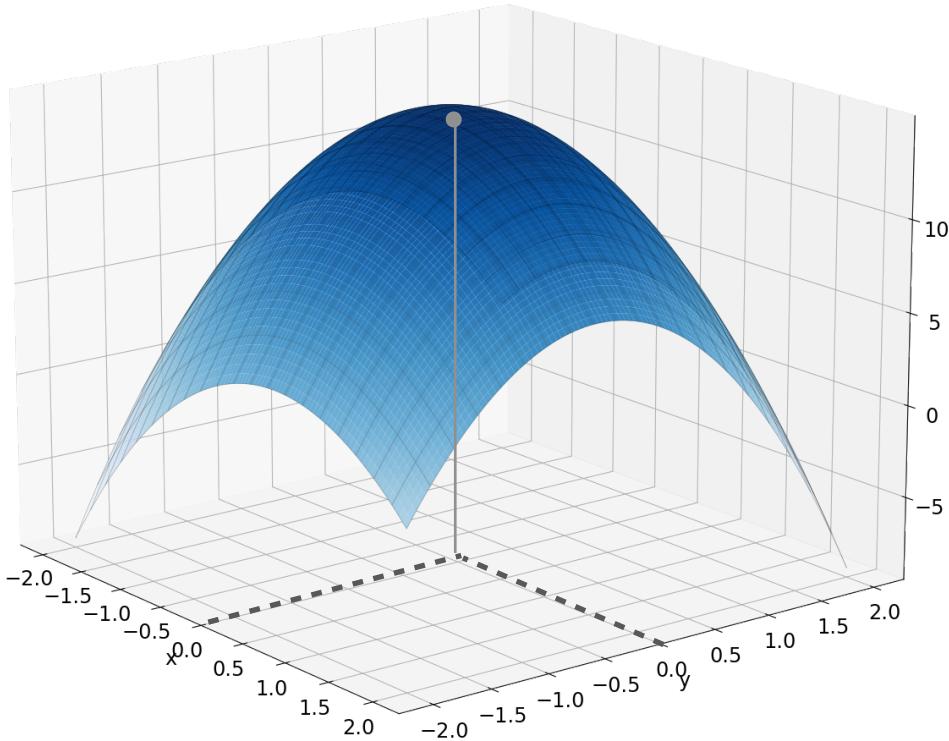
$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(-4 - \lambda)(-6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 + 10\lambda + 23$$

# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

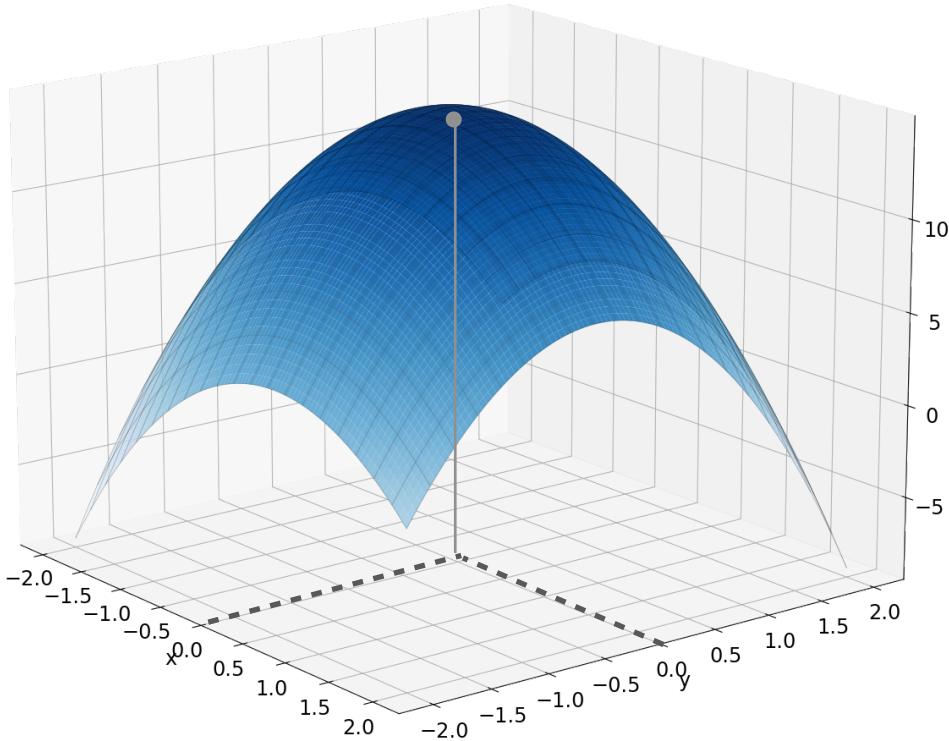
$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(-4 - \lambda)(-6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 + 10\lambda + 23 \rightarrow \lambda_1 = -3.49$$

# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

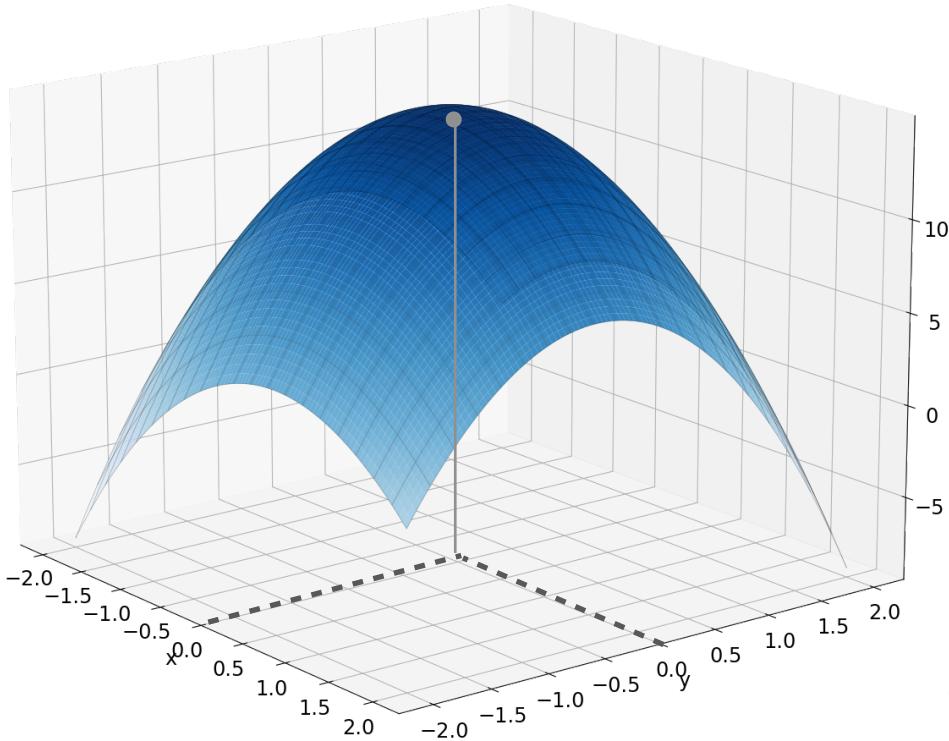
$$\det(H(0,0) - \lambda I) =$$

$$(-4 - \lambda)(-6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 + 10\lambda + 23$$

$$\begin{array}{l} \lambda_1 = -3.49 \\ \lambda_2 = -6.41 \end{array}$$

# Concave Down



$$f(x, y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x, y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(-4 - \lambda)(-6 - \lambda) - (-1)(-1)$$

$$= \lambda^2 + 10\lambda + 23$$

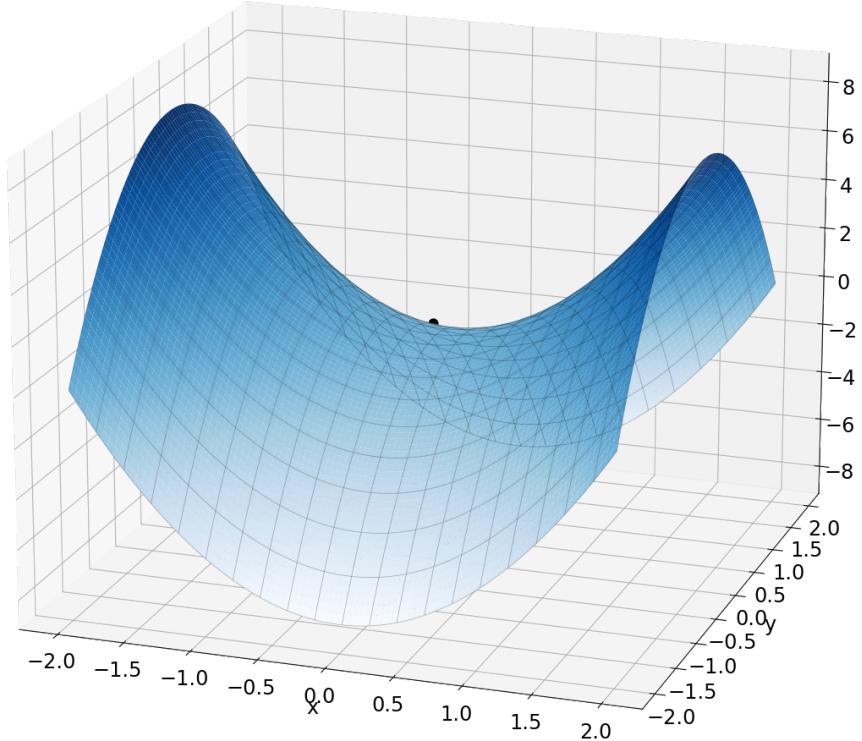
$$\lambda_1 = -3.49$$
$$\lambda_2 = -6.41$$

(0,0) is a maximum!

< 0

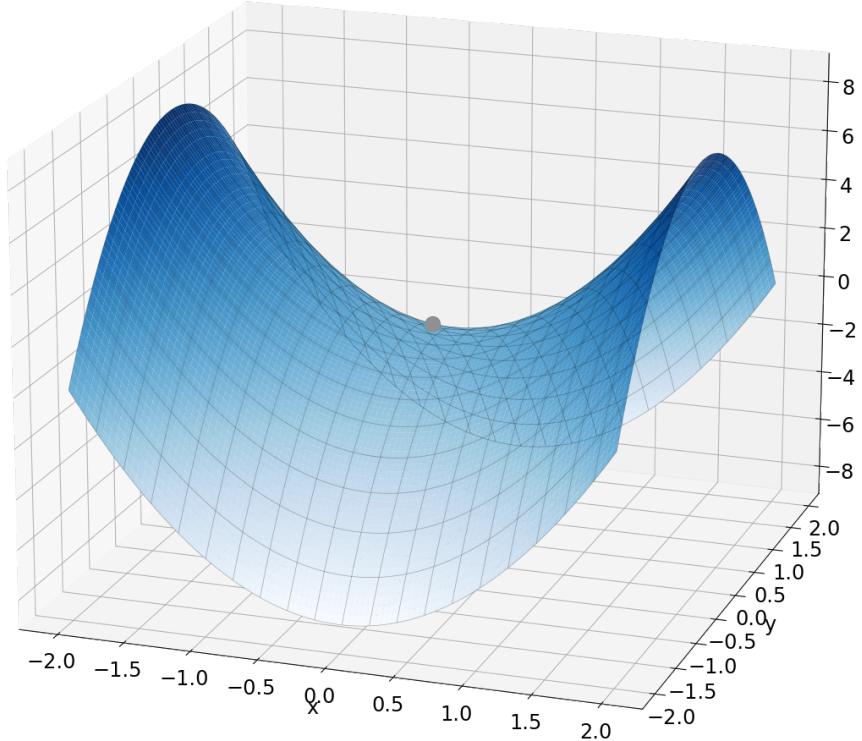
# Saddle Point

# Saddle Point



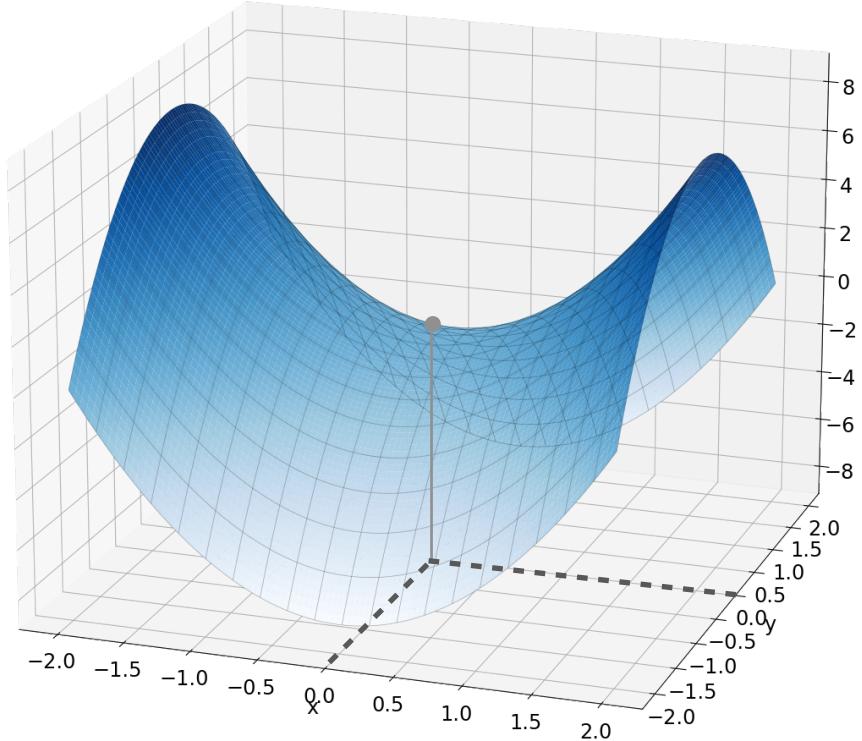
$$f(x, y) = 2x^2 - 2y^2$$

# Saddle Point



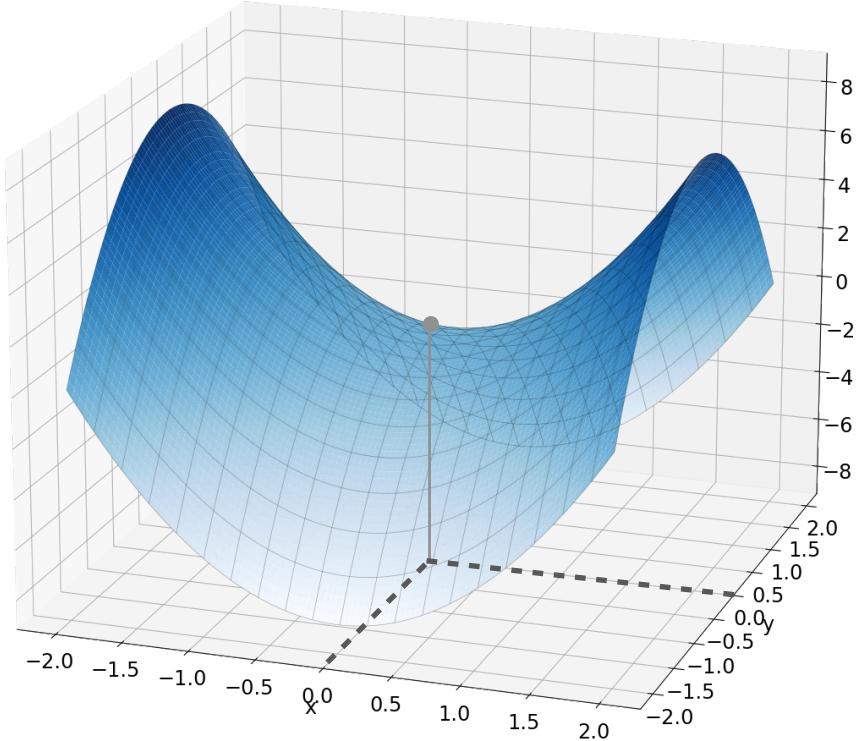
$$f(x, y) = 2x^2 - 2y^2$$

# Saddle Point



$$f(x, y) = 2x^2 - 2y^2$$

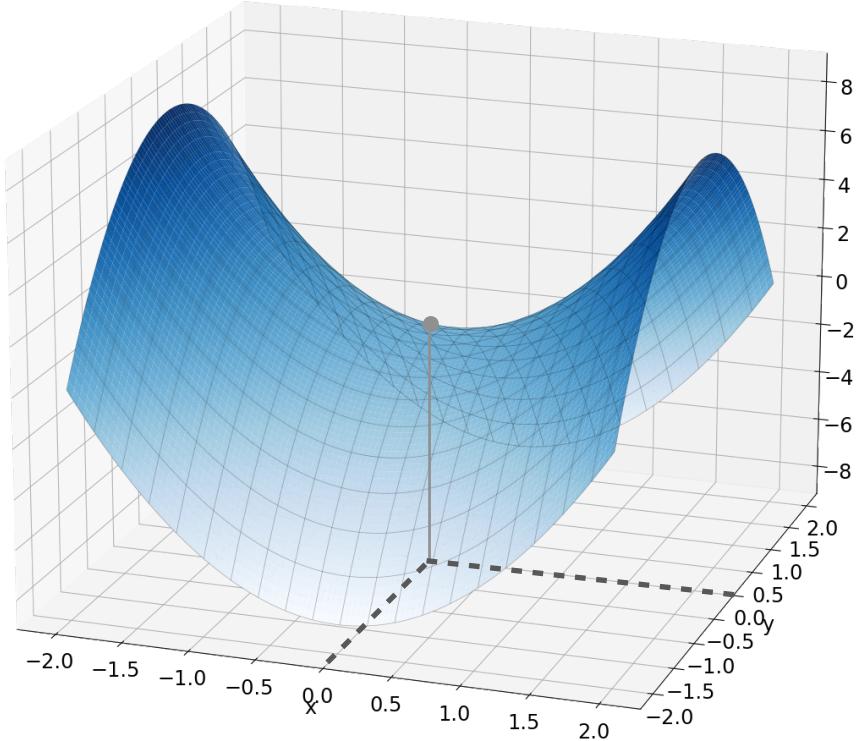
# Saddle Point



$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

# Saddle Point

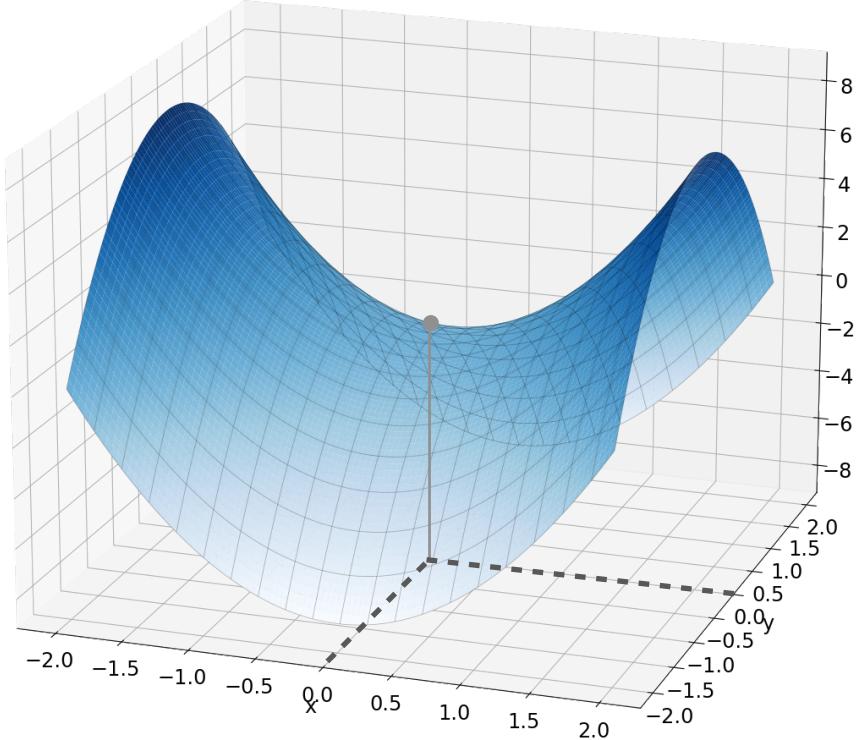


$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} 4 & 0 \\ 0 & -4 \end{bmatrix}$$

# Saddle Point



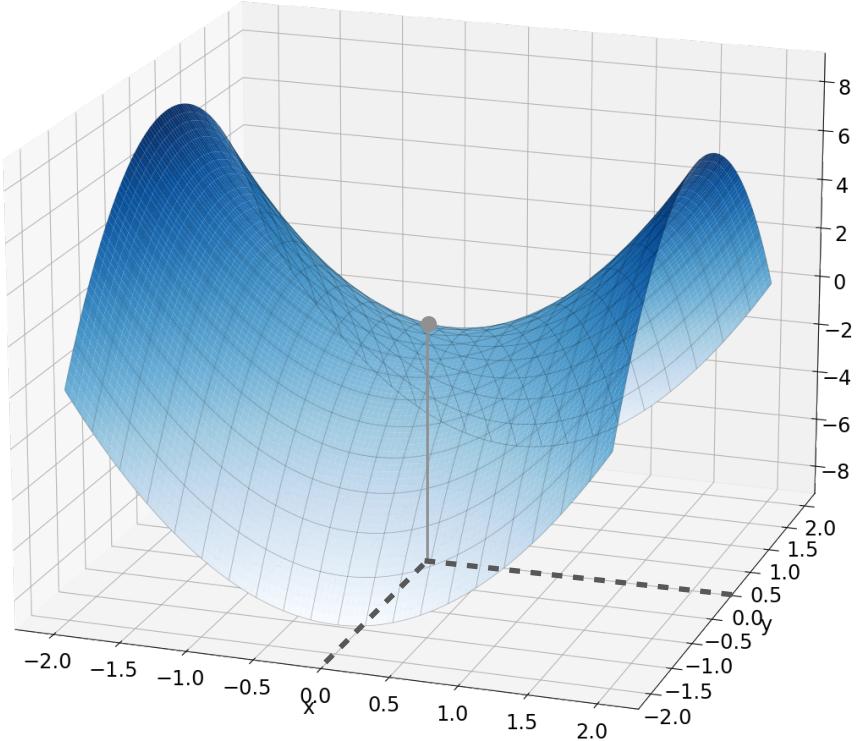
$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} 4 & 0 \\ 0 & -4 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

# Saddle Point



$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

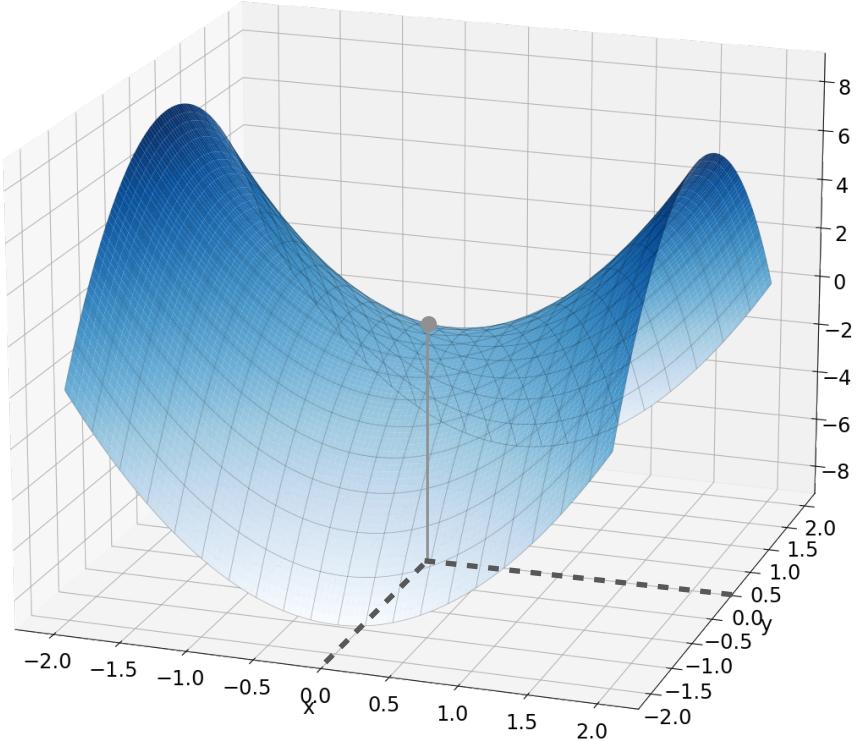
$$H(0,0) = \begin{bmatrix} 4 & 0 \\ 0 & -4 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(4 - \lambda)(-4 - \lambda) - 0$$

$$\lambda_1 = -4$$

# Saddle Point



$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

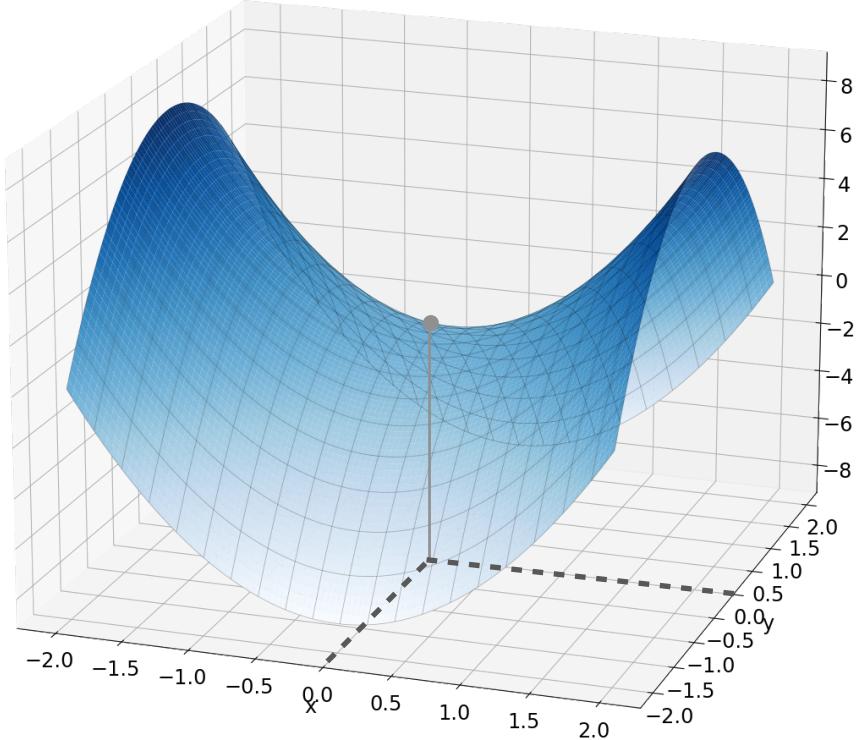
$$H(0,0) = \begin{bmatrix} 4 & 0 \\ 0 & -4 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(4 - \lambda)(-4 - \lambda) - 0$$

$$\begin{array}{l} \xrightarrow{\hspace{1cm}} \lambda_1 = -4 \\ \xrightarrow{\hspace{1cm}} \lambda_2 = 4 \end{array}$$

# Saddle Point



$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} 4 & 0 \\ 0 & -4 \end{bmatrix}$$

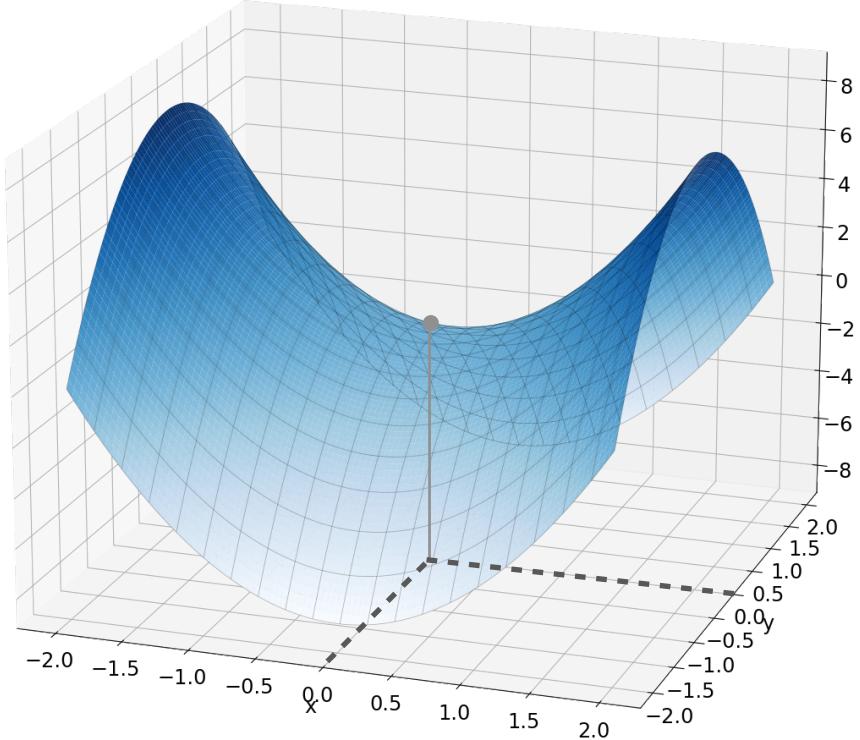
$$\det(H(0,0) - \lambda I) =$$

$$(4 - \lambda)(-4 - \lambda) - 0 < 0$$

$\lambda_1 = -4$   
 $\lambda_2 = 4$

(0,0) is saddle point

# Saddle Point



$$f(x, y) = 2x^2 - 2y^2$$

$$\nabla f(x, y) = \begin{bmatrix} 4x \\ -4y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} 4 & 0 \\ 0 & -4 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) =$$

$$(4 - \lambda)(-4 - \lambda) - 0 < 0$$

$\lambda_1 = -4$

$$> 0$$

$\lambda_2 = 4$

(0,0) is saddle point

# Summary

# Summary

1 variable  
 $f(x)$

2 variables  
 $f(x, y)$

More variables  
 $f(x_1, x_2, \dots, x_n)$

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$		

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$		

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \text{ & } \lambda_2 < 0$	

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \text{ & } \lambda_2 < 0$	All $\lambda_i < 0$

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \text{ & } \lambda_2 < 0$	All $\lambda_i < 0$
Need more information	$f''(x) = 0$		

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \text{ & } \lambda_2 < 0$	All $\lambda_i < 0$
Need more information	$f''(x) = 0$	Saddle point $\lambda_1 > 0 \text{ & } \lambda_2 < 0$ $\lambda_1 < 0 \text{ & } \lambda_2 > 0$	

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \text{ & } \lambda_2 < 0$	All $\lambda_i < 0$
Need more information	$f''(x) = 0$	Saddle point $\lambda_1 > 0 \text{ & } \lambda_2 < 0$ $\lambda_1 < 0 \text{ & } \lambda_2 > 0$ Or some $\lambda_i = 0$	

# Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \dots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \text{ & } \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \text{ & } \lambda_2 < 0$	All $\lambda_i < 0$
Need more information	$f''(x) = 0$	Saddle point $\lambda_1 > 0 \text{ & } \lambda_2 < 0$ $\lambda_1 < 0 \text{ & } \lambda_2 > 0$ Or some $\lambda_i = 0$	Some $\lambda_i > 0$ and some $\lambda_j < 0$ OR At least one $\lambda_i = 0$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

**Newton's method for two  
variables**

# Newton's Method

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

2 variables

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix}$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} -$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - H^{-1}(x_k, y_k)$$

# Newton's Method

1 variable

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - H^{-1}(x_k, y_k) \nabla f(x_k, y_k)$$

# Newton's Method

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - H^{-1}(x_k, y_k) \nabla f(x_k, y_k)$$

# Newton's Method

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \textcolor{orange}{H^{-1}(x_k, y_k)} \quad \textcolor{teal}{\nabla f(x_k, y_k)}$$

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \nabla f(x_k, y_k) \quad H^{-1}(x_k, y_k)$$

# Newton's Method

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - H^{-1}(x_k, y_k) \nabla f(x_k, y_k)$$

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \nabla f(x_k, y_k) \cancel{H^{-1}(x_k, y_k)}$$

# Newton's Method

2 variables

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \underbrace{H^{-1}(x_k, y_k)}_{2 \times 2} \underbrace{\nabla f(x_k, y_k)}_{2 \times 1}$$

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \nabla f(x_k, y_k) \cancel{- H^{-1}(x_k, y_k)}$$

# Newton's Method

2 variables

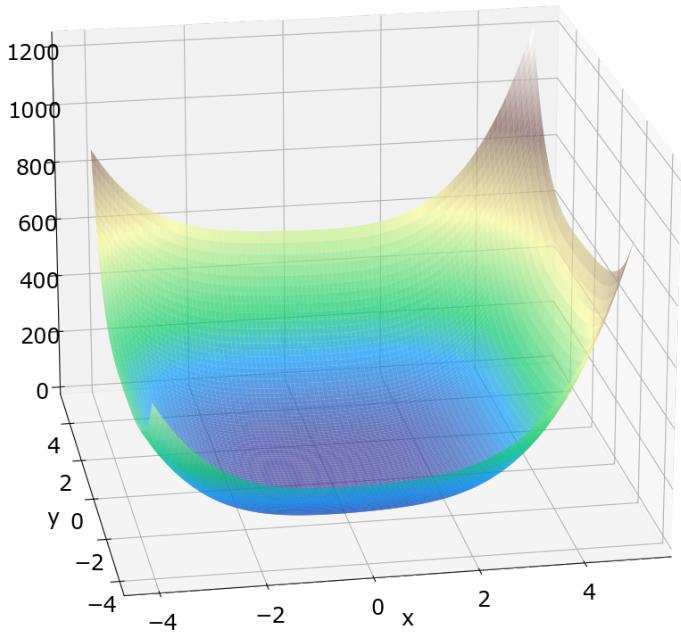
$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \underbrace{\mathbf{H}^{-1}(x_k, y_k)}_{2 \times 2} \underbrace{\nabla f(x_k, y_k)}_{2 \times 1}$$

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \nabla f(x_k, y_k) \cancel{- H^{-1}(x_k, y_k)}$$

When working with 2 variables the order is crucial!

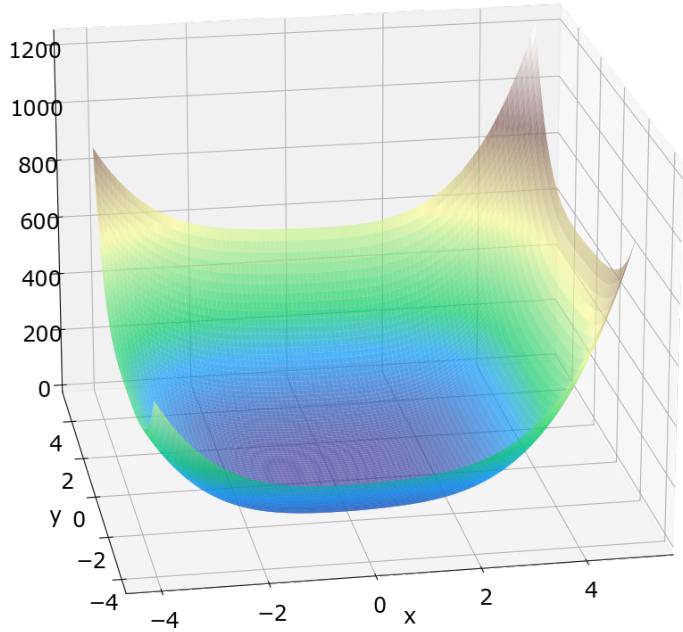
# An Example

# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

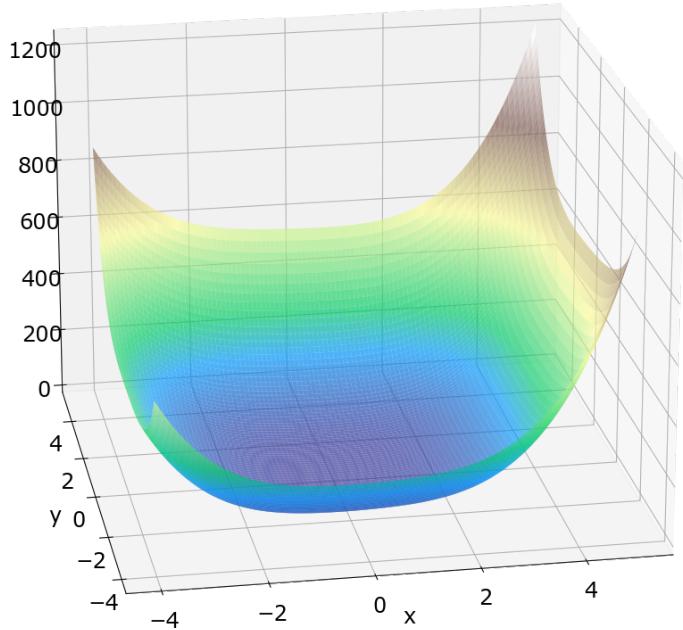
# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$$f(x, y) \rightarrow 4x^3 + 8x - y - 0.4xy$$

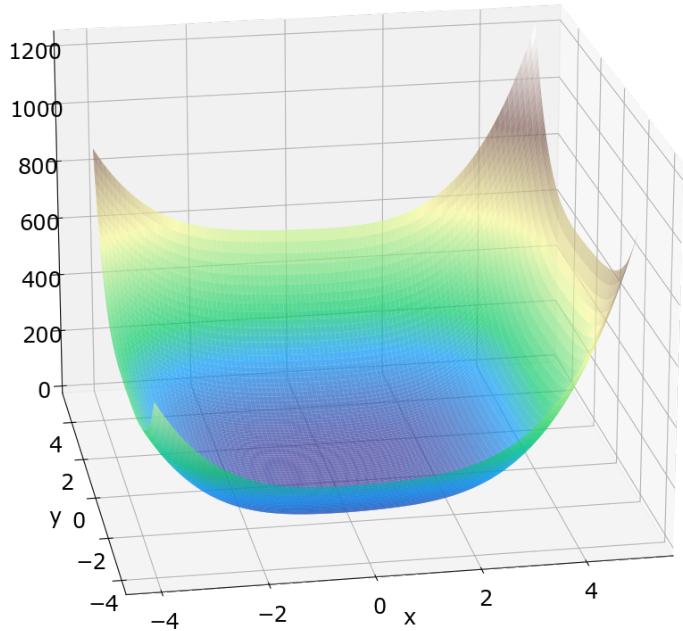
# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$$\begin{array}{l} \textcolor{blue}{x} \nearrow 4x^3 + 8x - y - 0.4xy \\ f(x, y) \\ \textcolor{orange}{y} \searrow 3.2y^3 + 4y - x - 0.2x^2 \end{array}$$

# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

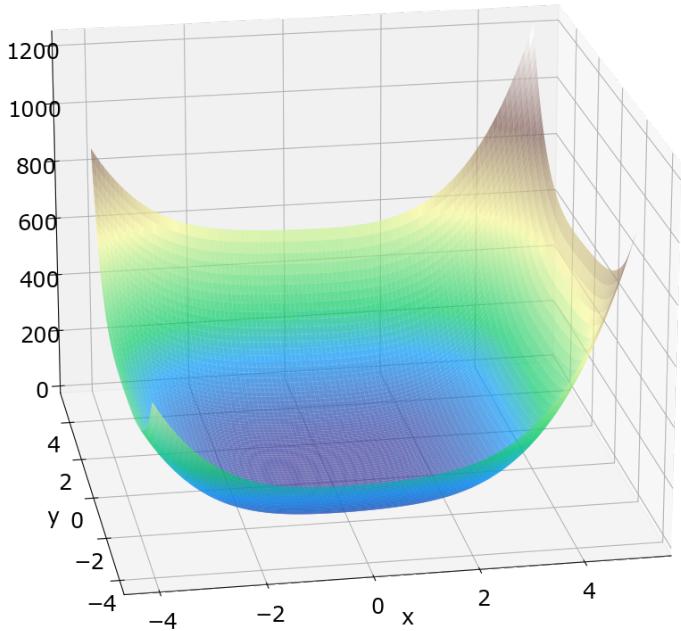
$f(x, y)$

$x$   $12x^2 + 8 - 0.4y$

$y$   $4x^3 + 8x - y - 0.4xy$

$3.2y^3 + 4y - x - 0.2x^2$

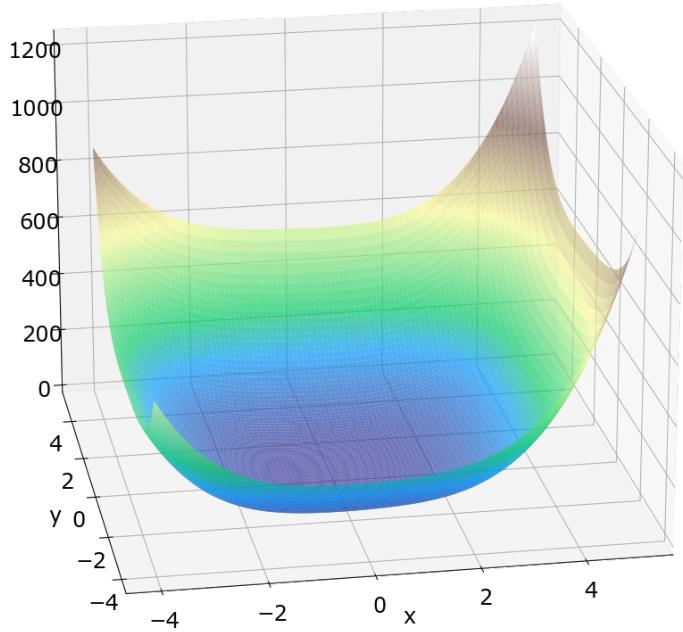
# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$$f(x, y) \begin{matrix} \nearrow x \\ \searrow y \end{matrix} \begin{array}{l} 4x^3 + 8x - y - 0.4xy \\ 3.2y^3 + 4y - x - 0.2x^2 \\ 12x^2 + 8 - 0.4y \\ -1 - 0.4x \end{array}$$

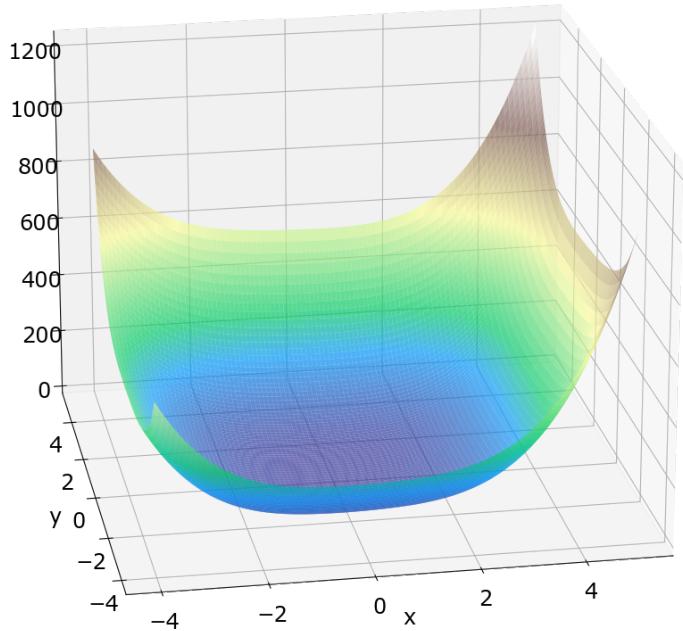
# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$$\begin{aligned} f(x, y) &\quad \begin{matrix} x \\ y \end{matrix} \quad 4x^3 + 8x - y - 0.4xy \quad \begin{matrix} x \\ y \end{matrix} \quad 12x^2 + 8 - 0.4y \\ &\quad \begin{matrix} y \\ \downarrow \end{matrix} \quad 3.2y^3 + 4y - x - 0.2x^2 \quad \begin{matrix} x \\ \downarrow \end{matrix} \quad -1 - 0.4x \end{aligned}$$

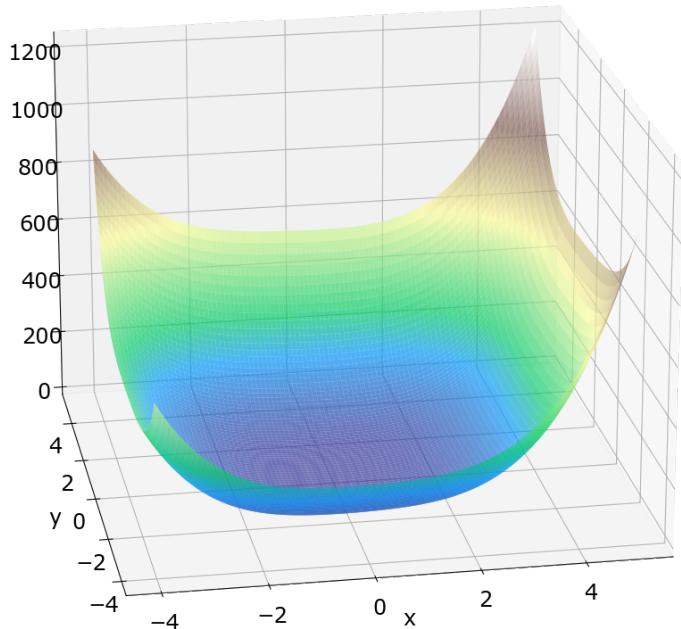
# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

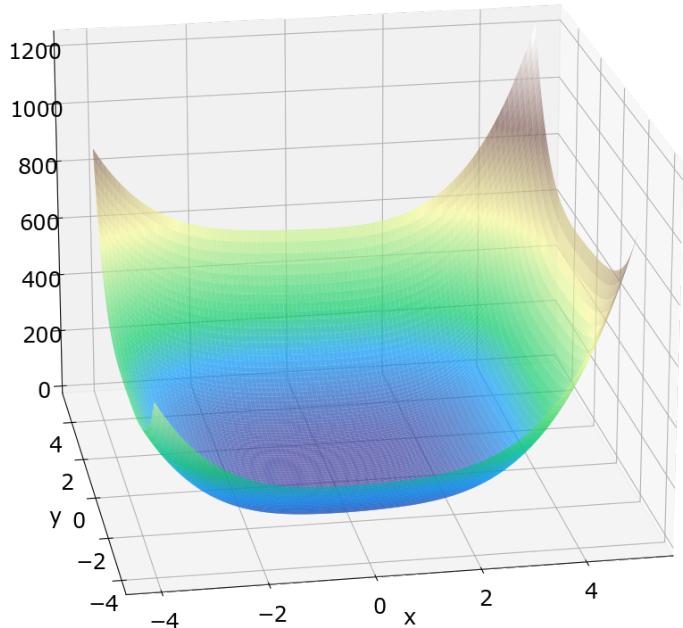
$$\begin{aligned} f(x, y) &\quad \begin{array}{l} \nearrow x \\ \searrow y \end{array} \\ &= 4x^3 + 8x - y - 0.4xy && \begin{array}{l} \nearrow x \\ \searrow y \end{array} & 12x^2 + 8 - 0.4y \\ &\quad \begin{array}{l} \nearrow x \\ \searrow y \end{array} && & -1 - 0.4x \\ &= 3.2y^3 + 4y - x - 0.2x^2 && \begin{array}{l} \nearrow x \\ \searrow y \end{array} & -1 - 0.4x \\ &\quad \begin{array}{l} \nearrow x \\ \searrow y \end{array} && & 9.6y^2 + 4 \end{aligned}$$

# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

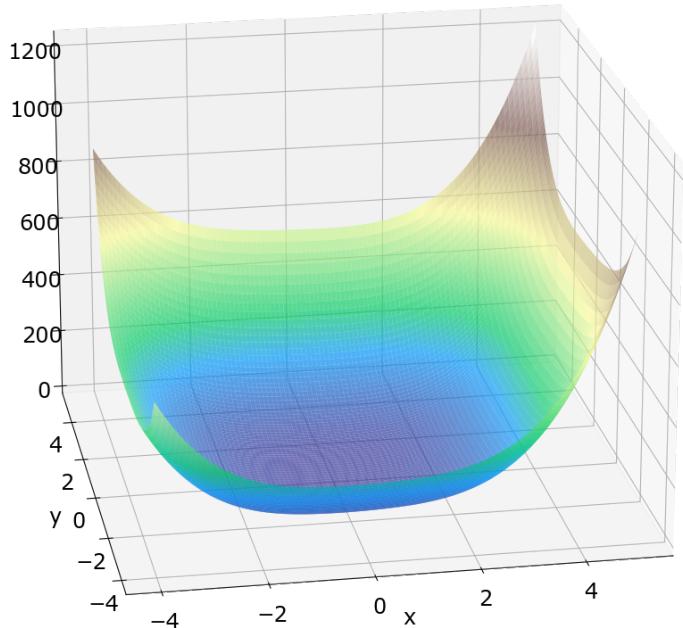
# An Example



$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$$\nabla f(x, y) = \begin{bmatrix} 4x^3 + 8x - y - 0.4xy \\ 3.2y^3 + 4y - x - 0.2x^2 \end{bmatrix}$$

# An Example

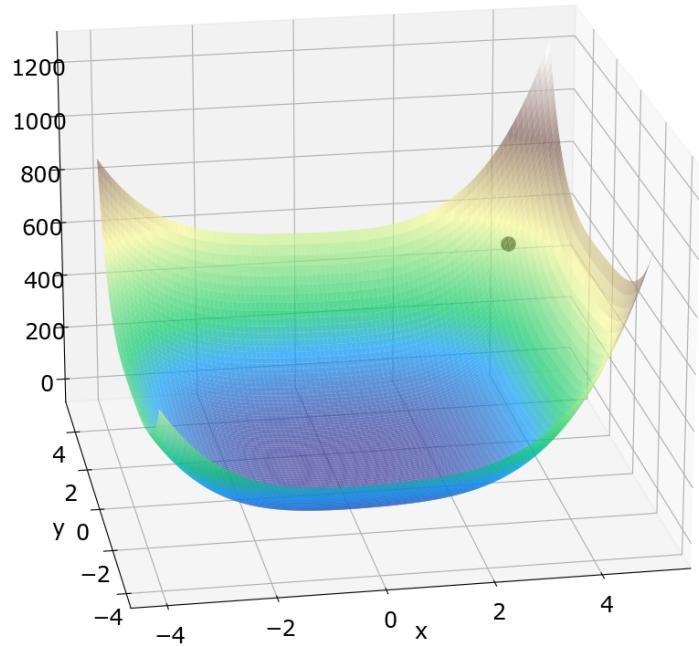


$$f(x, y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

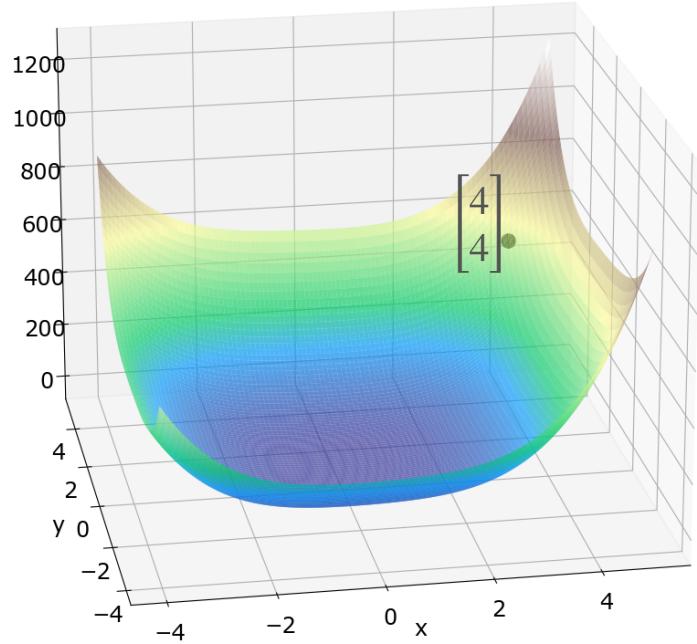
$$\nabla f(x, y) = \begin{bmatrix} 4x^3 + 8x - y - 0.4xy \\ 3.2y^3 + 4y - x - 0.2x^2 \end{bmatrix}$$

$$H(x, y) = \begin{bmatrix} 12x^2 + 8 - 0.4y & -1 - 0.4x \\ -1 - 0.4x & 9.6y^2 + 4 \end{bmatrix}$$

# An Example

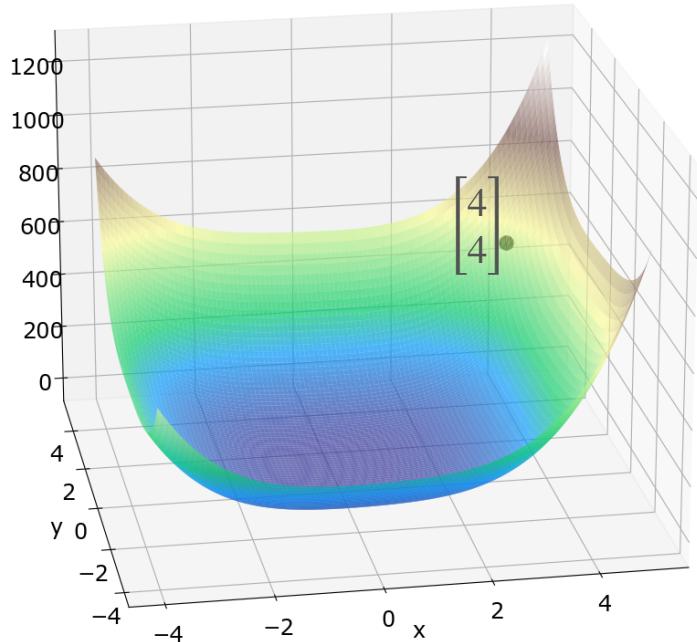


# An Example



Start at some point  $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

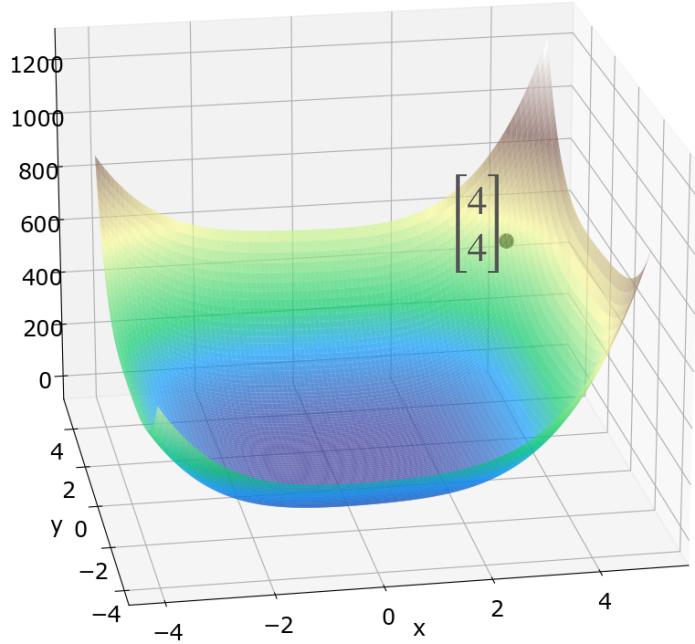
# An Example



Start at some point  $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$$\nabla f(4,4) = \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix}$$

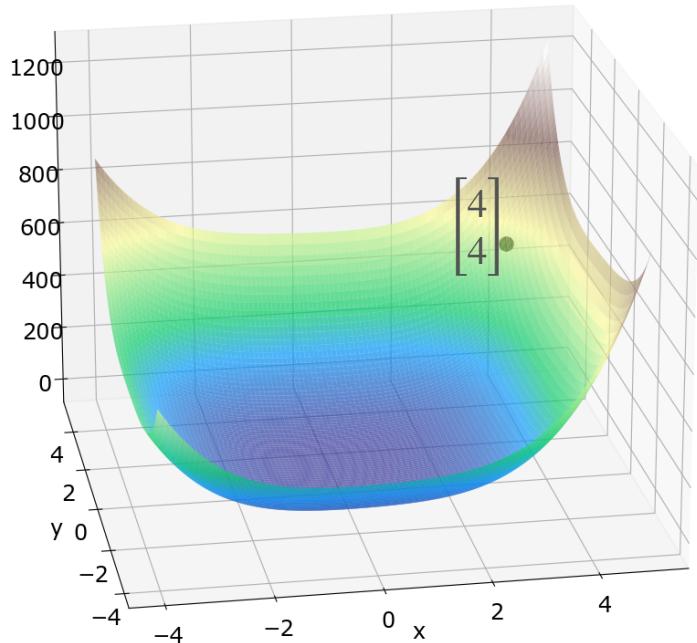
# An Example



Start at some point  $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$$\nabla f(4,4) = \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix} \quad H(4,4) = \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}$$

# An Example

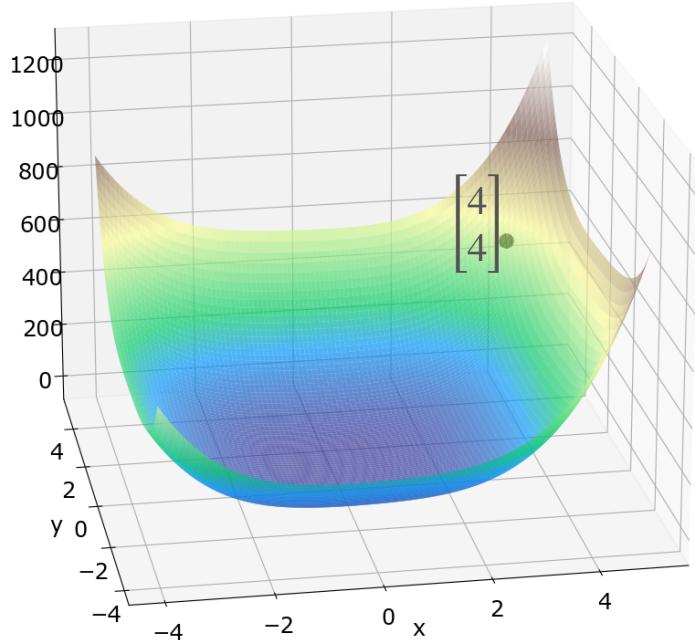


Start at some point  $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$$\nabla f(4,4) = \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix} \quad H(4,4) = \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} -$$

# An Example

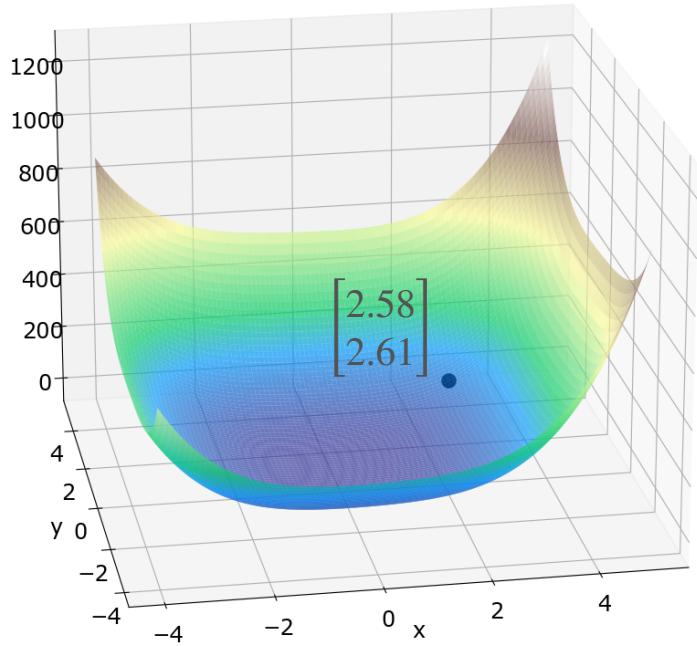


Start at some point  $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$$\nabla f(4,4) = \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix} \quad H(4,4) = \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} - \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}^{-1} \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix}$$

# An Example



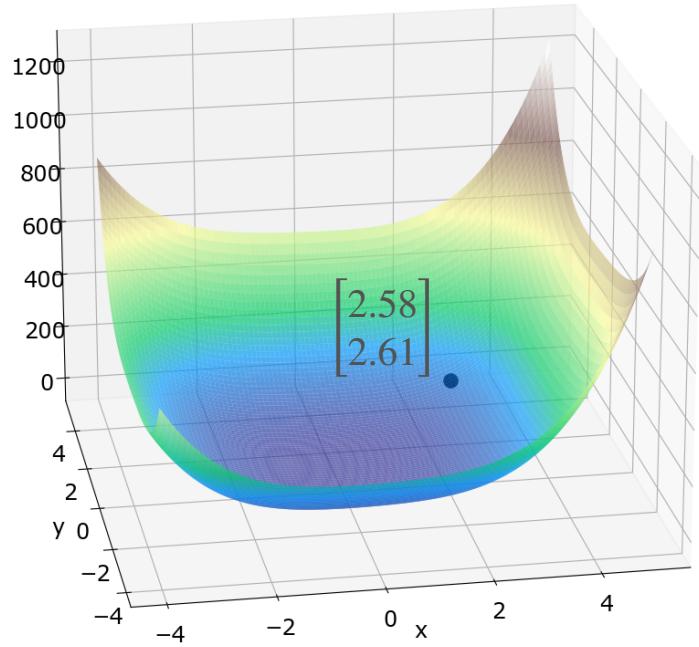
Start at some point  $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$$\nabla f(4,4) = \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix} \quad H(4,4) = \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} - \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}^{-1} \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix}$$

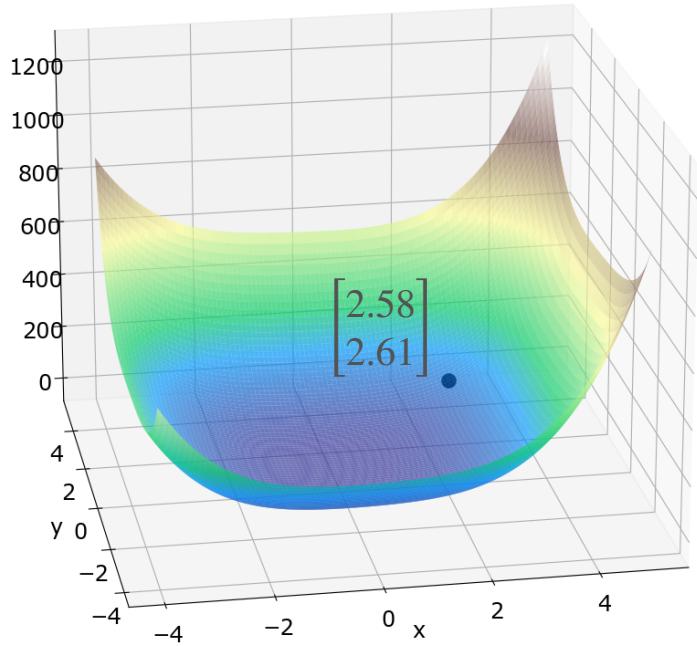
$$= \begin{bmatrix} 2.58 \\ 2.62 \end{bmatrix}$$

# An Example



$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix}$$

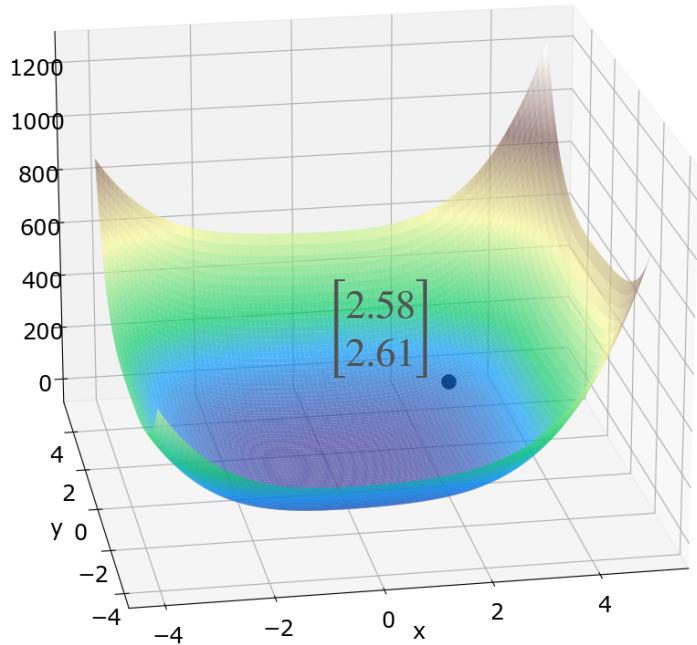
# An Example



$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix}$$

$$\nabla f(2.58, 2.61) = \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix}$$

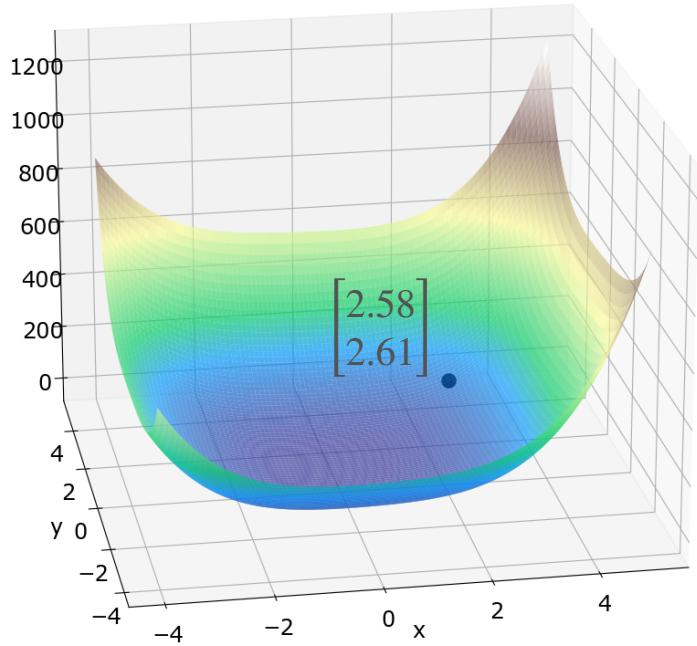
# An Example



$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix}$$

$$\nabla f(2.58, 2.61) = \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix} \quad H(2.58, 2.61) = \begin{bmatrix} 86.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}$$

# An Example

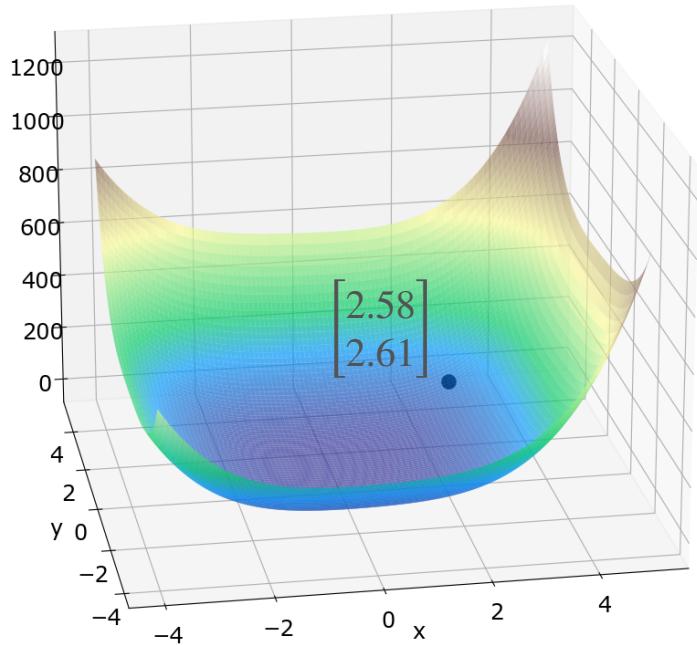


$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix}$$

$$\nabla f(2.58, 2.61) = \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix} H(2.58, 2.61) = \begin{bmatrix} 86.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix} -$$

# An Example

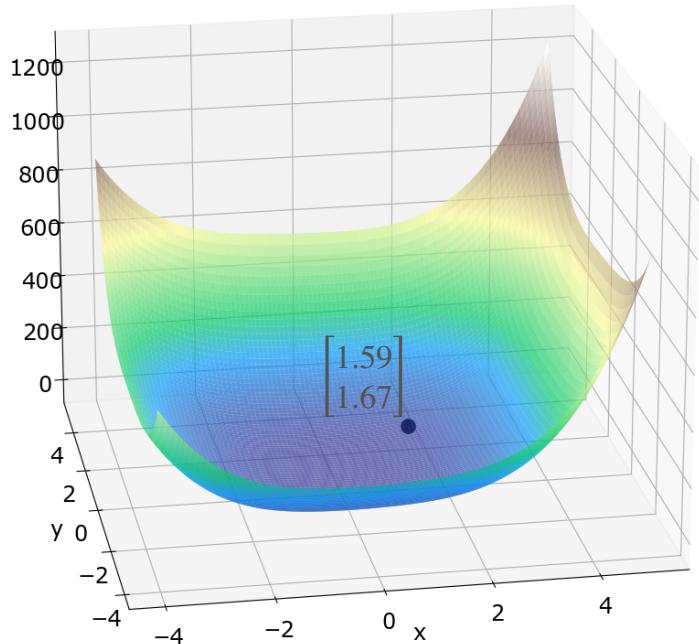


$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix}$$

$$\nabla f(2.58, 2.61) = \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix} H(2.58, 2.61) = \begin{bmatrix} 86.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix} - \begin{bmatrix} 86.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}^{-1} \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix}$$

# An Example

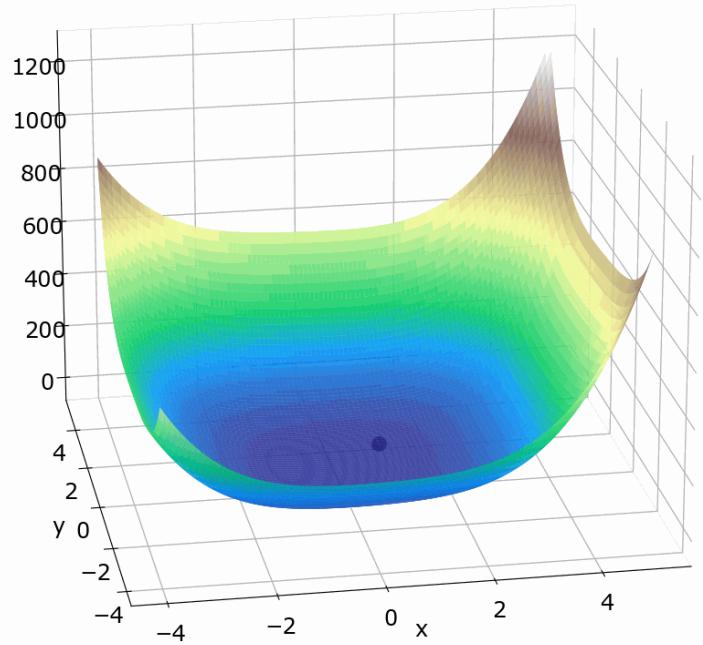


$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix}$$

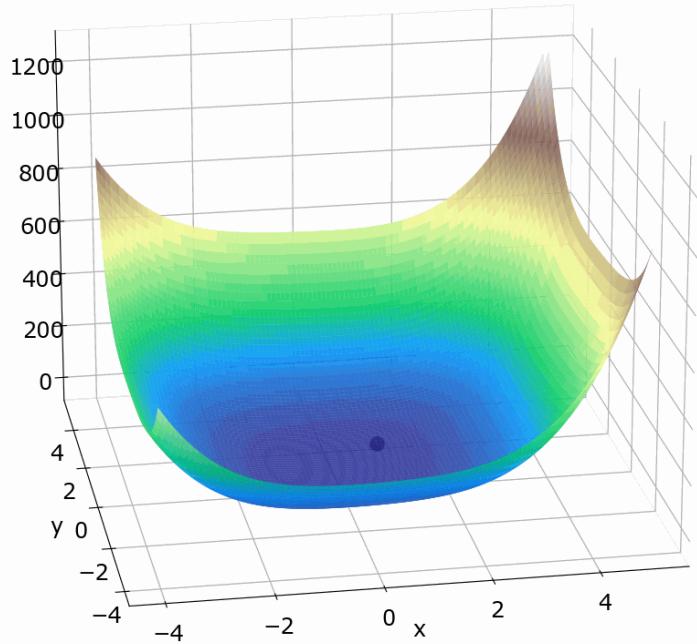
$$\nabla f(2.58, 2.61) = \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix} H(2.58, 2.61) = \begin{bmatrix} 86.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.58 \\ 2.61 \end{bmatrix} - \begin{bmatrix} 86.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}^{-1} \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix}$$
$$= \begin{bmatrix} 1.59 \\ 1.67 \end{bmatrix}$$

# An Example

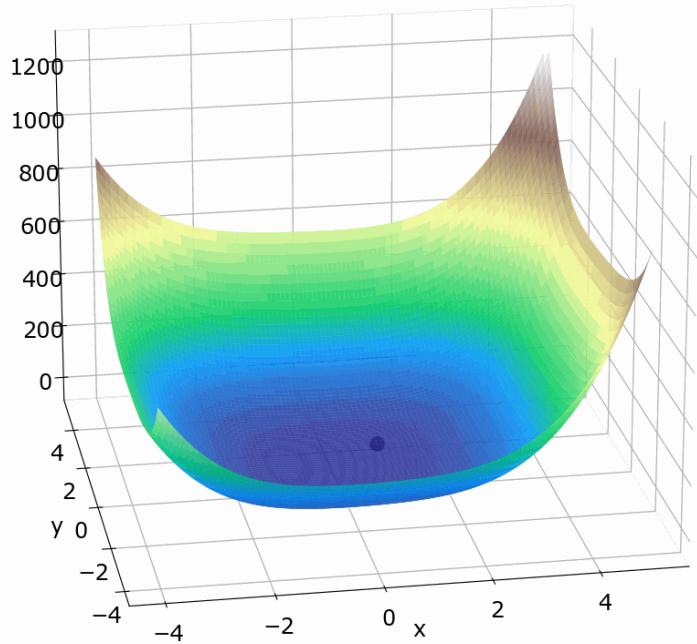


# An Example



Repeat until you are close enough to the actual zero!

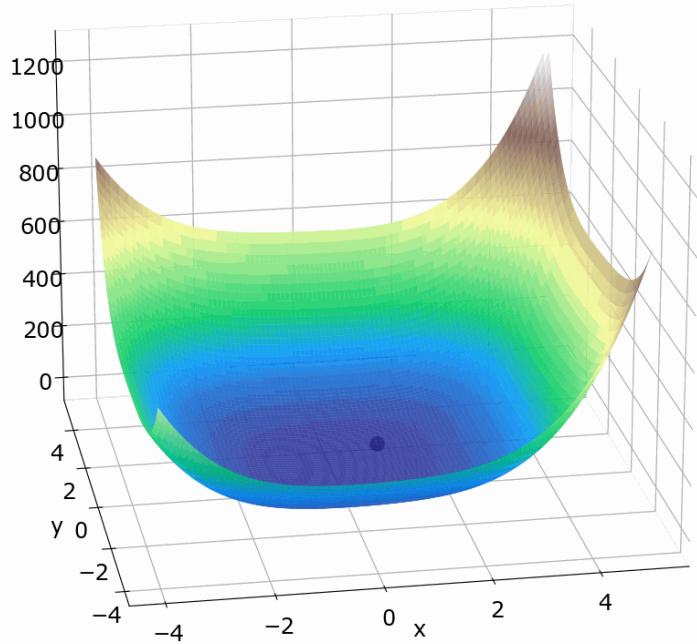
# An Example



Repeat until you are close enough to the actual zero!

Needed  $k = 8$  steps

# An Example

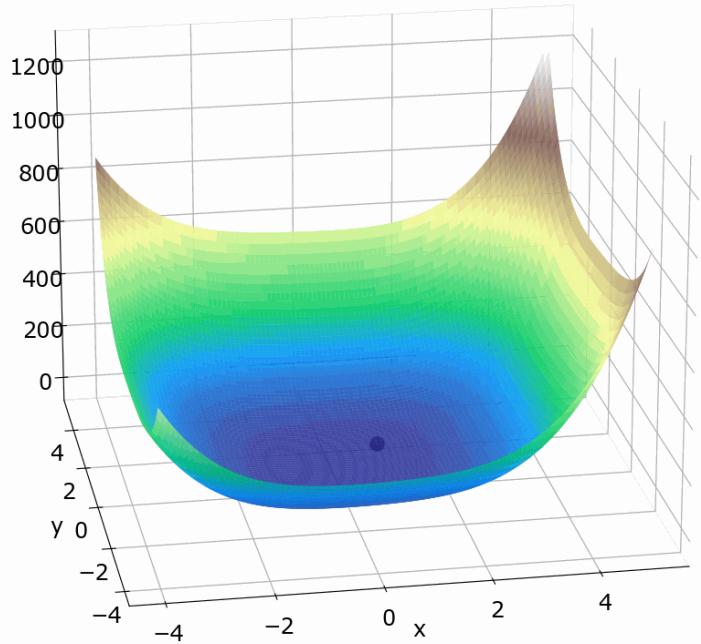


Repeat until you are close enough to the actual zero!

Needed  $k = 8$  steps

$$\begin{bmatrix} x_8 \\ y_8 \end{bmatrix} = \begin{bmatrix} 4.15 \cdot 10^{-17} \\ -2.05 \cdot 10^{-17} \end{bmatrix}$$

# An Example



Repeat until you are close enough to the actual zero!

Needed  $k = 8$  steps

$$\begin{bmatrix} x_8 \\ y_8 \end{bmatrix} = \begin{bmatrix} 4.15 \cdot 10^{-17} \\ -2.05 \cdot 10^{-17} \end{bmatrix}$$

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$



DeepLearning.AI

# Optimization in Neural Networks and Newton's Method

---

## Conclusion