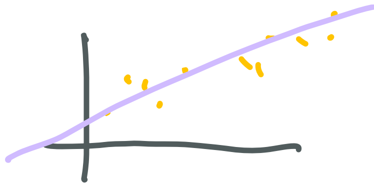
Neural Networks

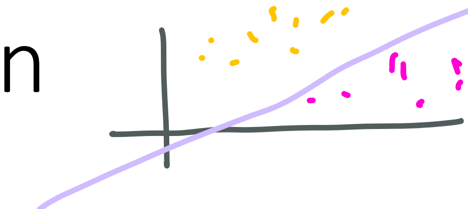
Outline

- Questions about projects?
- Recap of a singular neuron model
- Notation and terminology
- Compute graphs
- Optimization
- Backpropagation

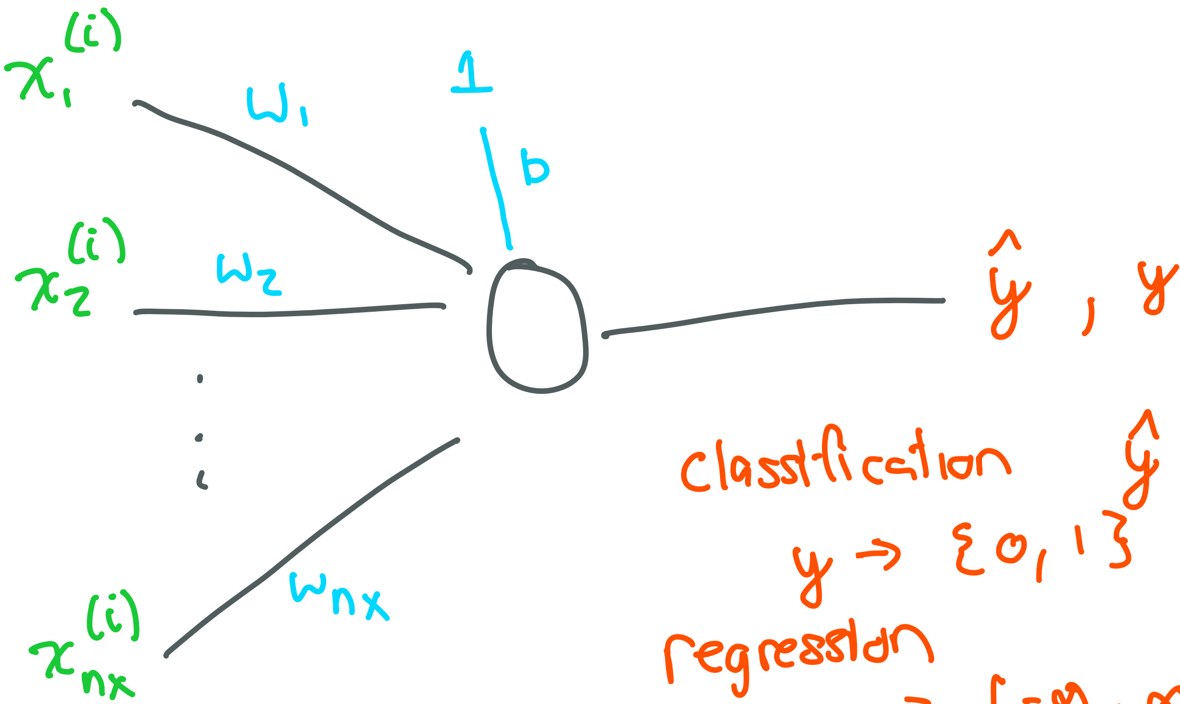
- This will be our most math heavy week
- Next week we’ll rely on PyTorch to compute all derivatives

Recap: A Single Neuron





- Take five minutes to draw
  - Whatever will help you remember (no correct or incorrect drawings)
  - You'll keep a running drawing log the rest of the semester



$\mathcal{L} = ...$

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

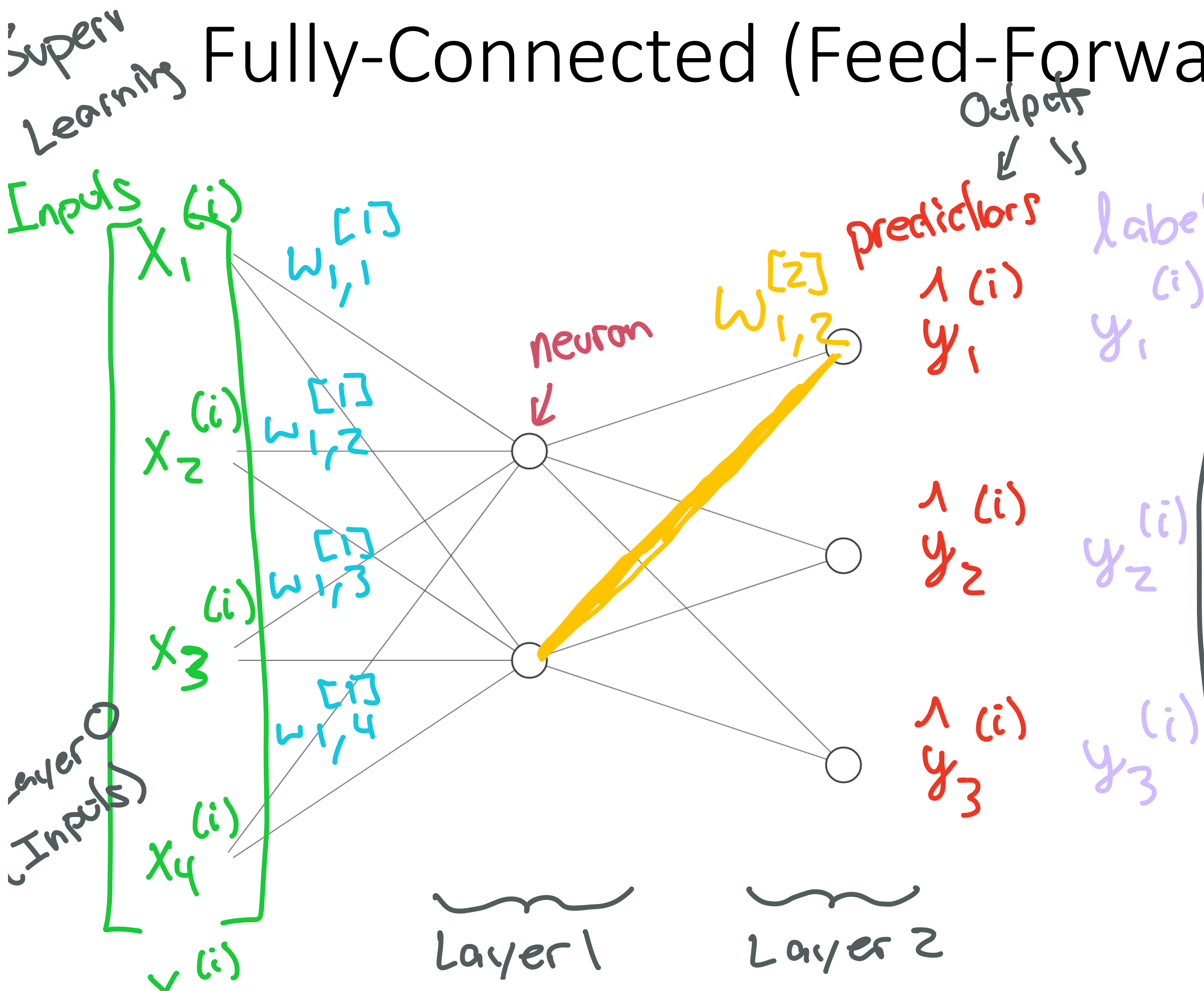
classification $\hat{y} = \sigma(w^T x + b)$

$y \rightarrow \{0, 1\}$

regression $\hat{y} = w^T x + b$

$y \rightarrow [-\infty, \infty], [-1, 1]$

Fully-Connected (Feed-Forward) Network

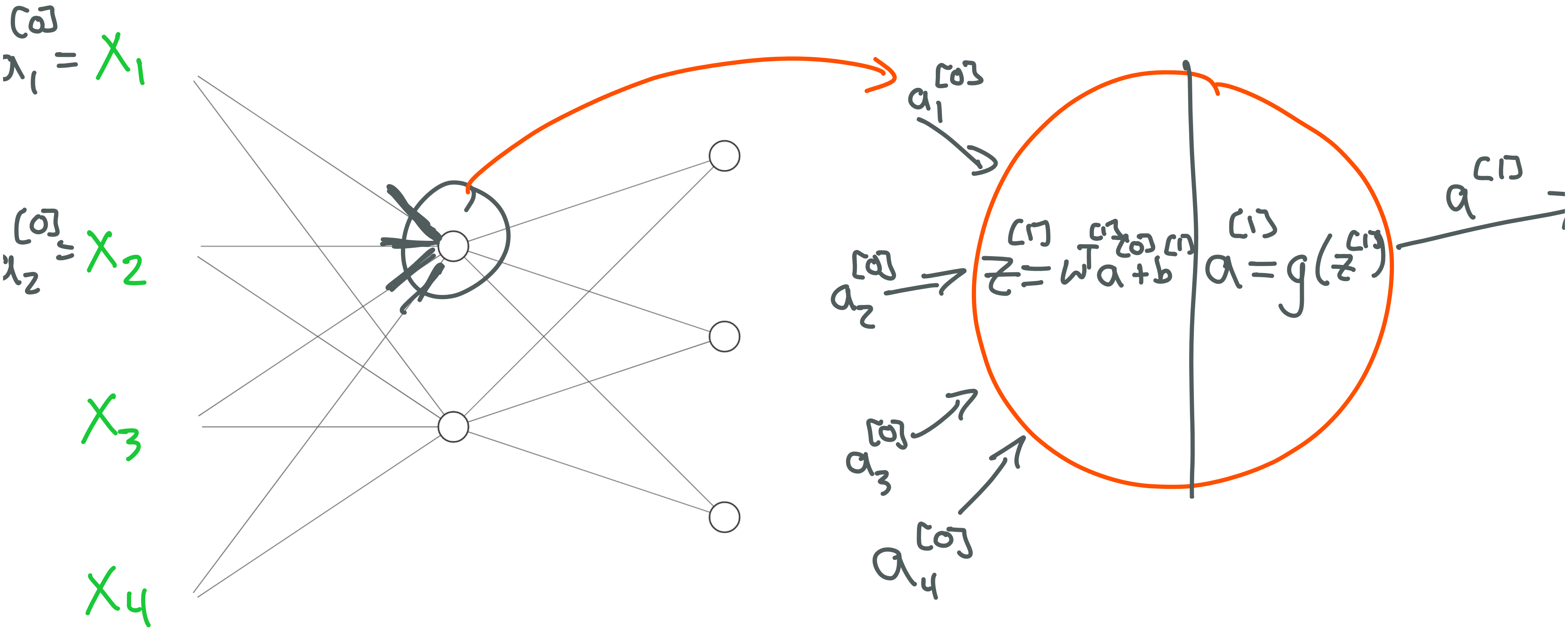


$$z_1^{[1](i)} = \underbrace{w_1^{[1]T} \cdot x^{(i)}}_{\text{dot-product}} + b_1^{[1]}$$

$$a_1^{[1](i)} = g\left(z_1^{[1](i)}\right)$$

2-Layer Network

Fully-Connected (Feed-Forward) Network



2-Layer Network

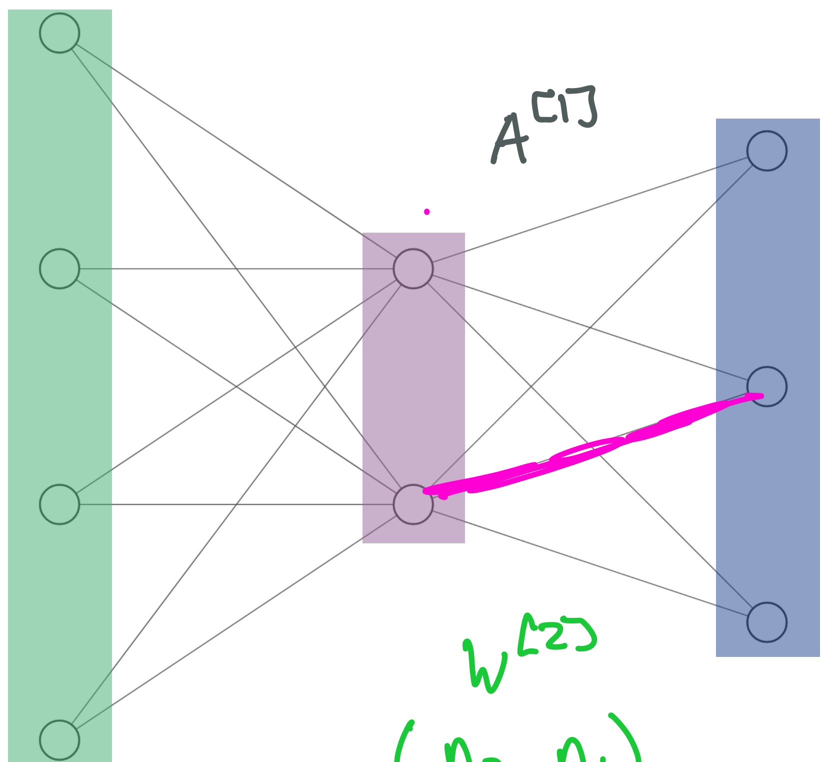
Fully-Connected (Feed-Forward) Network

$A^{[0]} = X$

$\hat{Y} = A^{[2]}$

$A^{[0]}$

$A^{[1]}$



$W^{[2]}$

(n_2, n_1)

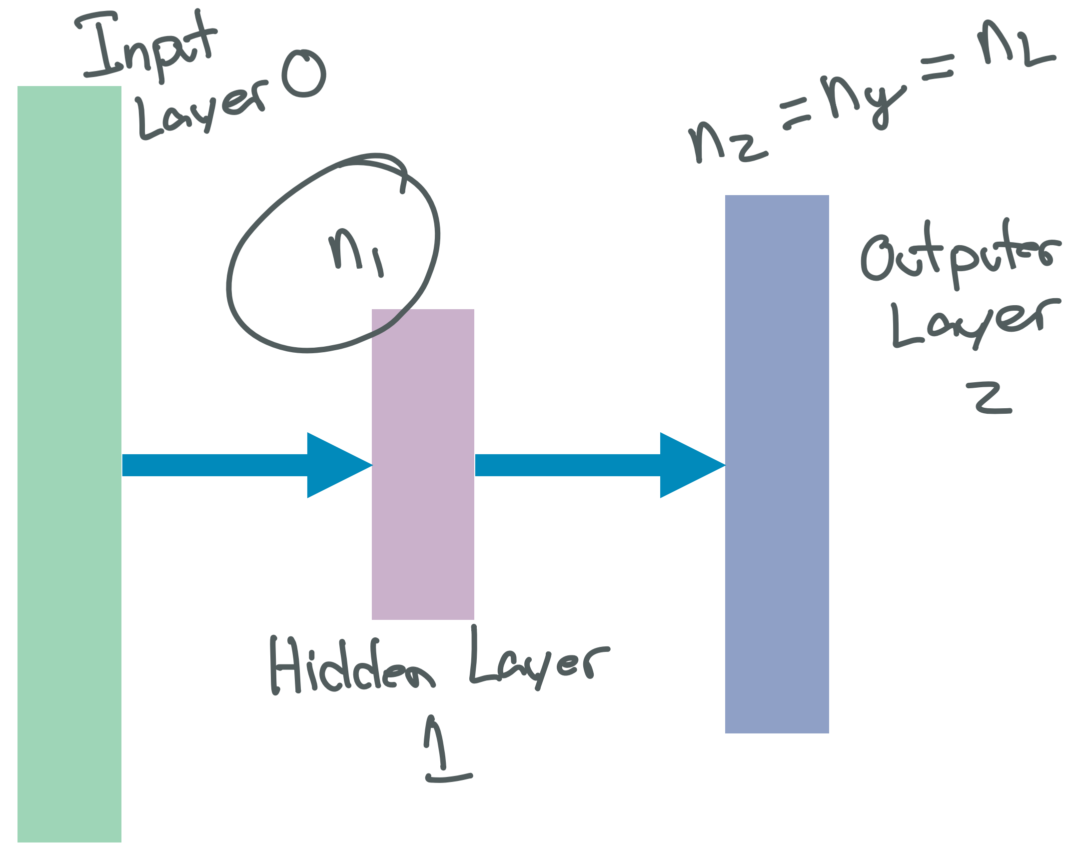
$\dfrac{\partial \mathcal{L}}{\partial W^{[2]}}$

(n_2, n_1)

Fully-Connected (Feed-Forward) Network

(n_x, n_1, n_y)

(n_0, n_1, n_2)



Simplified (easier to draw) diagram

Vectorized Equations

16,000 examples

$z_1^{[1](i)}$

$z_2^{[1](i)}$

$\vdots$

$z_{100}^{[1](i)}$

these are scalars

$\underbrace{z^{[1]}}_{\text{matrix}}$

$(100, \cancel{1})$

n_1 ✓

$(100, 16{,}000)$

$(16{,}000, 100)$

(N, n_1)

$z_1^{[1](i)} = w_1^{[1]T} \cdot x^{(i)} + b_1^{[}$

$$\underset{(N,\, n_1)}{Z^{[1]}} = \underset{(N,\, n_x)}{X}\, \underset{(n_x,\, n_1)}{W^{[1]T}} + \underset{(1,\, n_1)}{b^{[1]T}}$$

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_{n_x}^{(1)} \\ \vdots & & & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_{n_x}^{(N)} \end{bmatrix} \quad N \text{ rows}, \; n_x \text{ columns}$$

$W^{[l]} \rightarrow (n_l, n_{l-1})$

Vectorized Equations (for any layer)

$$\underbrace{Z^{[l]}}_{(N,\ n_l)} = \underbrace{\underbrace{A^{[l-1]}}_{(N,\ n_{l-1})} \underbrace{W^{[l]T}}_{(n_{l-1},\ n_l)}}_{(N,\ n_l)} + \underbrace{b^{[l]T}}_{(1,\ n_l)}$$

broadcasted to (N, n_l)

$$Z^{[1]} = A^{[\overset{0}{1-1}]} W^{[1]T} + b^{[1]T} \qquad A^{[1]} = g(Z^{[1]})$$

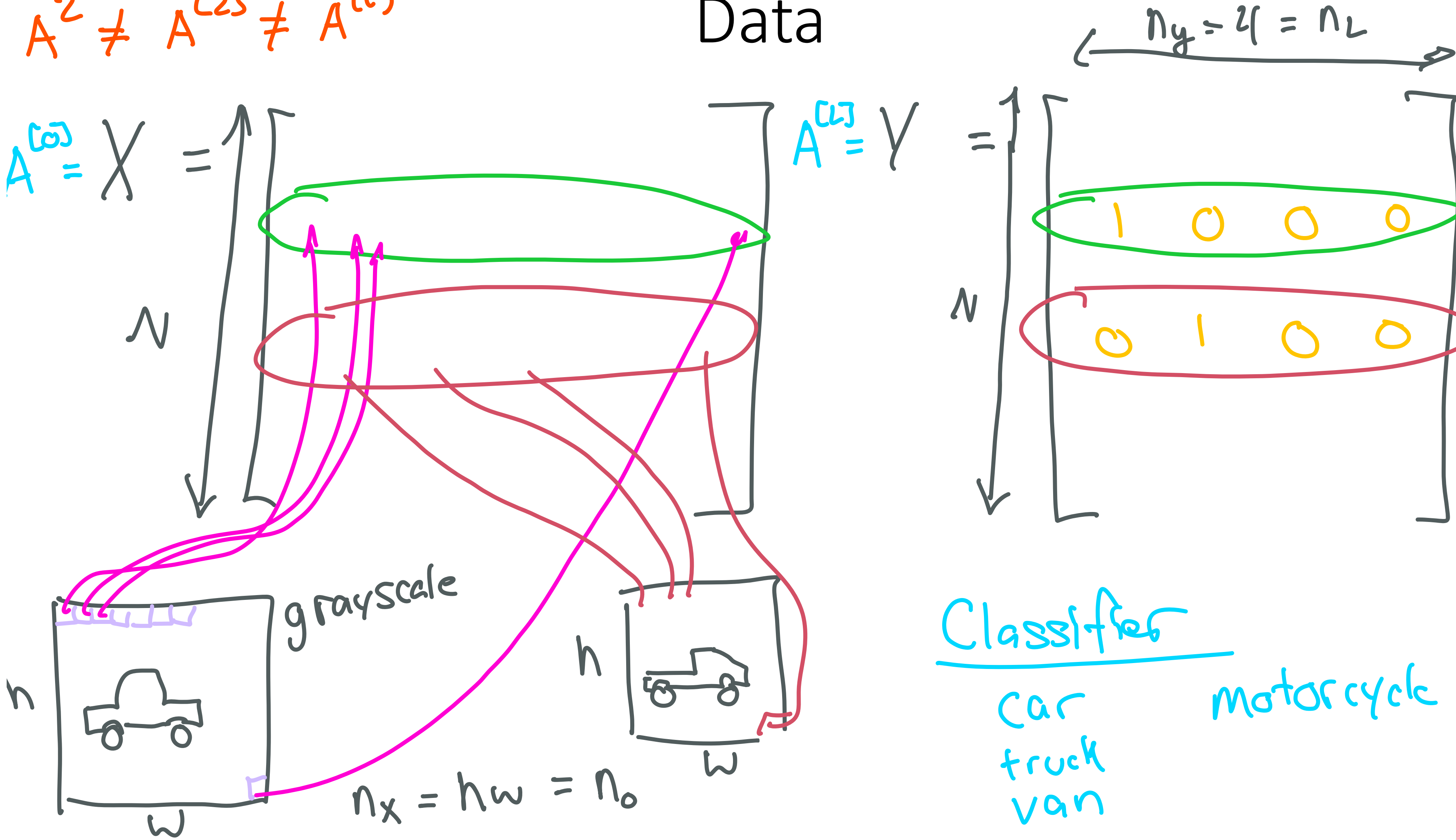
$$Z^{[2]} = A^{[\overset{1}{2-1}]} W^{[2]T} + b^{[2]T} \qquad A^{[2]} = g(Z^{[2]}) = \hat{y}$$

element-wise

$$\underbrace{A^{[l]}}_{(N,\ n_l)} = g(\underbrace{Z^{[l]}}_{(N,\ n_l)}) = \begin{bmatrix} g(z_{1,1}^{[l]}) & g(z_{1,2}^{[l]}) & \cdots \\ \vdots & \ddots & \\ g(z_{n_l,1}^{[l]}) & \cdots & \end{bmatrix}$$

Data

$A^2 \neq A^{[2]} \neq A^{(i)}$



MNIST Dataset Example

- MNIST includes 60,000 training images
- Each image is grayscale and 28x28 pixels in size

$28 \cdot 28 = 784$

- Each output is a one-hot encoding of the digits 0 through 9

- What is the shape of X?

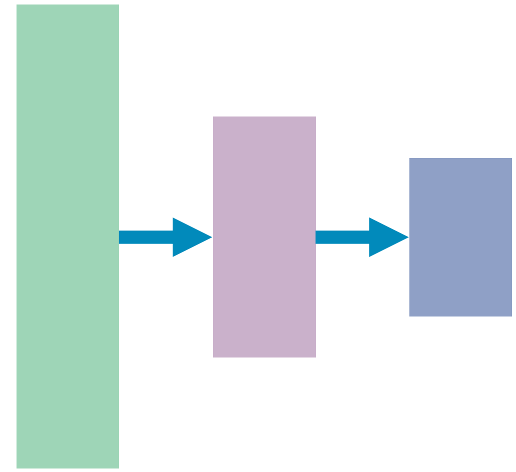
$(N, n_x) \rightarrow (60{,}000, 784)$

$\underbrace{(60{,}000, \overbrace{28, 28}^{\text{1-byte}})}_{\text{How much memory?}}$

- What is the shape of Y?

$(N, n_y) \rightarrow (60{,}000, 10)$

MNIST Neural Network



- What is the shape of $Z^{[1]}$?

$(N, n_x) \Rightarrow (N, n_1) \Rightarrow (60{,}000, 17)$

- What is the shape of $A^{[1]}$?

$(60{,}000, 17)$

- What is the shape of $Z^{[2]}$?

$(60{,}000, 10)$

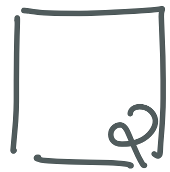
- What is the shape of $A^{[2]}$?

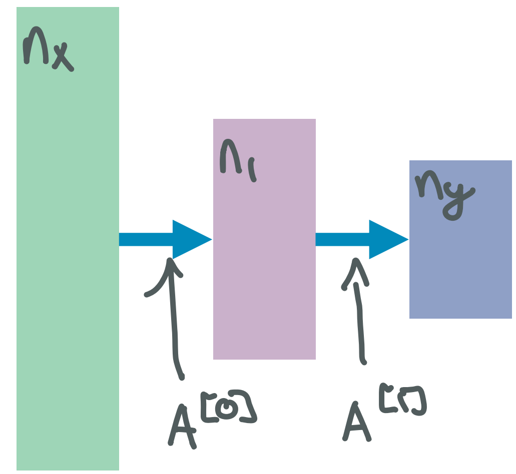
$(60{,}000, 10)$

$\boxed{2}$ $\boxed{1}$

MNIST Neural Network

- Imagine we have a two-layer network
- The hidden layer has 17 neurons





$$Z^{[l]} = A^{[l-1]} W^{[l]T} + b^{[l]}$$

- What is the shape of $W^{[1]}$?

  $(n_l, n_{l-1}) \Rightarrow (n_1, n_0) \Rightarrow (17, 784)$

- What is the shape of $b^{[1]}$?

  $(n_l, 1) \Rightarrow (n_1, 1) \Rightarrow (17, 1)$

- What is the shape of $W^{[2]}$?

  $(n_l, n_{l-1}) \Rightarrow (n_2, n_1) \Rightarrow (10, 17)$

- What is the shape of $b^{[2]}$?

  $(n_l, n_{l-1}) \Rightarrow (n_2, 1) \Rightarrow (10, 1)$

Compute Graph

$W^{[1]}, b^{[1]}$

$A^{[0]} \rightarrow$

$\boxed{Z^{[1]} = A^{[0]} W^{[1]T} + b^{[1]T}}$

$\boxed{A^{[1]} = g(Z^{[1]})}$

Neurons layer 1

$W^{[2]}, b^{[2]}$

$\boxed{Z^{[2]} = A^{[1]} W^{[2]T} + b^{[2]T}}$

$\boxed{A^{[2]} = g(Z^{[2]})} \rightarrow \boxed{\mathcal{L} = (Y - A^{[2]})^2}$

Optimization with Binary Cross Entropy Loss

$$\mathcal{L}(\hat{Y}, Y) = -\| Y \log \hat{Y} + (1-Y)\log(1-\hat{Y}) \|$$

$\hat{y}$	y	$\log \hat{y}$	$\log(1-\hat{y})$	$\mathcal{L}$
0.1	0	−1	−0.046	−0.046
0.1	1	−1	−0.046	−1
0.9	0	−0.046	−1	−1
0.9	1	−0.046	−1	−0.046

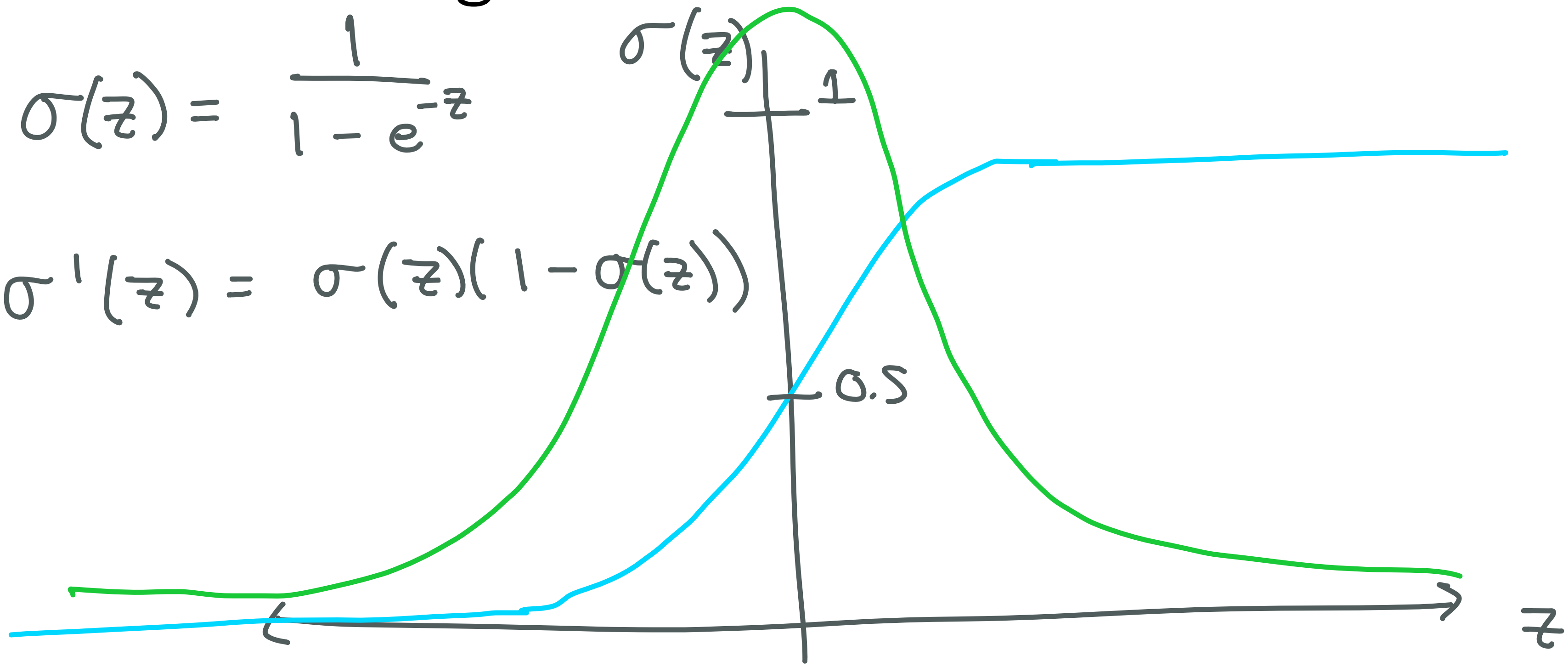
$\hat{y} = [0, 1]$

$y \in \{0, 1\}$

Sigmoid Activation Functions

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

$$\sigma'(z) = \sigma(z)(1 - \sigma(z))$$



$$\sigma(0) = 0.5$$

Backpropagation $W^{[2]}$

$$\frac{\partial \mathcal{L}}{\partial W^{[2]}} = \frac{\partial}{\partial W^{[2]}} - || Y \log \hat{Y} + (1-Y) \log (1-\hat{Y}) ||$$

$$= \overset{①}{\frac{\partial \mathcal{L}}{\partial \hat{Y}}} \cdot \overset{②}{\frac{\partial \hat{Y}}{\partial Z^{[2]}}} \cdot \overset{③}{\frac{\partial Z^{[2]}}{\partial W^{[2]}}} = -\left(\frac{Y}{\hat{Y}} - \frac{(1-Y)}{1-\hat{Y}}\right) \cdot \hat{Y}(1-\hat{Y}) A$$

$$= (\hat{Y} - Y) A^{[1]}$$

① $$\frac{\partial \mathcal{L}}{\partial \hat{Y}} = -\left(\frac{Y}{\hat{Y}} - \frac{(1-Y)}{(1-\hat{Y})}\right)$$

② $$\frac{\partial \hat{Y}}{\partial Z^{[2]}} = \sigma(Z^{[2]})(1-\sigma(Z^{[2]})) = \hat{Y}(1-\hat{Y})$$

③ $$\frac{\partial Z^{[2]}}{\partial W^{[2]}} = \frac{\partial}{\partial W^{[2]}} A^{[1]} W^{[2]T} + b^{[2]T} = A^{[1]}$$

Backpropagation $b^{[2]}$

$$\frac{\partial \mathcal{L}}{\partial b^{[2]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}}}_{\text{Already Computed}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} \text{ (3)} = (\hat{y} - y)$$

(3) $$\frac{\partial z^{[2]}}{\partial b^{[2]}} = \frac{\partial}{\partial b^{[2]}} A^{[1]} W^{[2]T} + b^{[2]T} = \underline{1}$$

Backpropagation $W^{[1]}$

$$\frac{\partial \mathcal{L}}{\partial W^{[1]}} = \underbrace{\overset{①}{\frac{\partial \mathcal{L}}{\partial \hat{Y}}} \cdot \overset{②}{\frac{\partial \hat{Y}}{\partial Z^{[2]}}}}_{\text{Already Computed (Back propagate)}} \cdot \underbrace{\overset{③}{\frac{\partial Z^{[2]}}{\partial A^{[1]}}} \cdot \overset{④}{\frac{\partial A^{[1]}}{\partial Z^{[1]}}} \cdot \overset{⑤}{\frac{\partial Z^{[1]}}{\partial W^{[1]}}}}$$

Practice run (not completed in class)

Update for $W^{[1]}$

$$\frac{d}{dx}\log_{10} x = \frac{1}{x}\cdot\frac{1}{\log 10}$$

$$\frac{\partial \mathcal{L}}{\partial W^{[1]}} = \frac{\partial}{\partial W^{[1]}} - \| y \log \hat{y} + (1-y)\log(1-\hat{y}) \|$$

$$= -y \frac{\partial}{\partial W^{[1]}} \log \hat{y} - (1-y)\frac{\partial}{\partial W^{[1]}} \log(1-\hat{y})$$ distribute & log

$$= -y \frac{1}{\hat{y}} \frac{\partial}{\partial W^{[1]}} \hat{y} - (1-y)\frac{1}{1-\hat{y}} \frac{\partial}{\partial W^{[1]}} (\cancel{1} - \hat{y})$$ chain rule

$$= -\frac{y}{\hat{y}} \frac{\partial}{\partial W^{[1]}} \sigma(z^{[2]}) + \frac{1-y}{1-\hat{y}} \frac{\partial}{\partial W^{[1]}} \sigma(z^{[2]})$$ regroup

$$= \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right) \frac{\partial}{\partial W^{[1]}} \sigma(z^{[2]})$$ Apply sigmoid derivative

$$= \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right) \sigma(z^{[2]})(1-\sigma(z^{[2]})) \frac{\partial}{\partial W^{[1]}} z^{[2]}$$ substitute

$$= \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right) \hat{y}(1-\hat{y}) \frac{\partial}{\partial W^{[1]}} (A^{[1]} W^{[2]T} + \cancel{b^{[2]T}})$$

$$= (\hat{y}(1-y) - y(1-\hat{y}))\, W^{[2]} \frac{\partial}{\partial W^{[1]}} A^{[1]}$$

$$= (\hat{y} - \cancel{\hat{y}y} - y + \cancel{\hat{y}y})\, W^{[2]} \frac{\partial}{\partial W^{[1]}} \sigma(z^{[1]})$$

$$= (\hat{y} - y)\, W^{[2]} \sigma(z^{[1]})(1-\sigma(z^{[1]})) \frac{\partial}{\partial W^{[1]}} z^{[1]}$$

$$= (\hat{y} - y)\, W^{[2]} A^{[1]}(1-A^{[1]}) \frac{\partial}{\partial W^{[1]}} (A^{[0]} W^{[1]T} + \cancel{b^{[1]T}})$$

$$= (\hat{y} - y)\, W^{[2]} A^{[1]}(1-A^{[1]})\, A^{[0]}$$

$$X \to \boxed{A} \to \boxed{A} \to Y$$

$$W^{[1]}, b^{[1]} \qquad W^{[2]}, b^{[2]}$$

$$\frac{d}{dx} \log_{10} X = \frac{1}{X} \frac{1}{\log(10)}$$

$$\sigma'(z) = \sigma(z)(1-\sigma(x))$$

$$\frac{\partial \mathcal{L}}{\partial W^{[1]}} = \frac{\partial}{\partial W^{[1]}} -\| Y \log \hat{Y} + (1-Y)\log(1-\hat{Y}) \|$$

$$= -\left(\frac{\partial}{\partial W^{[1]}} \left(Y \log \hat{Y} \right) + \frac{\partial}{\partial W^{[1]}} \left((1-Y)\log(1-\hat{Y}) \right) \right)$$

$$= -\left(Y \frac{1}{\hat{Y}} \frac{\partial}{\partial W^{[1]}} \hat{Y} + (1-Y) \frac{1}{1-\hat{Y}} \frac{\partial}{\partial W^{[1]}} (\cancel{1} - \hat{Y}) \right)$$

$$= -\left(\frac{Y}{\hat{Y}} - \frac{1-Y}{1-\hat{Y}} \right) \frac{\partial}{\partial W^{[1]}} \hat{Y}$$

$$= \left(\frac{1-Y}{1-\hat{Y}} - \frac{Y}{\hat{Y}} \right) \frac{\partial}{\partial W^{[1]}} \sigma(Z^{[2]})$$

$$= \left(\frac{1-Y}{1-\hat{Y}} - \frac{Y}{\hat{Y}} \right) \sigma(Z^{[2]})(1 - \sigma(Z^{[2]})) \frac{\partial}{\partial W^{[1]}} Z^{[2]}$$

$$= \left(\frac{1-Y}{1-\hat{Y}} - \frac{Y}{\hat{Y}} \right) \hat{Y}(1-\hat{Y}) \frac{\partial}{\partial W^{[1]}} \left(A^{[1]} W^{[2]T} + \cancel{b^{[2]T}} \right)$$

$$= \left(\frac{1-Y}{1-\hat{Y}} - \frac{Y}{\hat{Y}} \right) \hat{Y}(1-\hat{Y}) W^{[2]} \frac{\partial}{\partial W^{[1]}} A^{[1]}$$

$$= \left(\hat{Y}(1-Y) - Y(1-\hat{Y}) \right) W^{[2]} \frac{\partial}{\partial W^{[1]}} \sigma(Z^{[1]})$$

$$= \left(\hat{Y} \cancel{- Y\hat{Y}} - Y \cancel{+ Y\hat{Y}} \right) W^{[2]} \sigma(Z^{[1]})(1 - \sigma(Z^{[1]})) \frac{\partial}{\partial W^{[1]}} Z^{[1]}$$

$$= (\hat{Y} - Y) W^{[2]} A^{[1]} (1 - A^{[1]}) \frac{\partial}{\partial W^{[1]}} \left(X W^{[1]T} + b^{[1]T} \right)$$

$$\boxed{= (\hat{Y} - Y) W^{[2]} A^{[1]} (1 - A^{[1]}) X}$$

For e in range (num_epochs)

$$W^{[1]} = W^{[1]} - \alpha \cdot$$

$$X \to |A| \to |A| \to Y$$

$$\boxed{W^{[1]}}, b^{[1]} \qquad W^{[2]}, b^{[2]}$$

$$\frac{d}{dx}\log_{10} X = \frac{1}{X}\,\frac{1}{\log(10)}$$

$$\frac{\partial \mathcal{L}}{\partial W^{[1]}} = \frac{\partial}{\partial W^{[1]}} -\| Y \log \hat{Y} + (1-Y)\log(1-\hat{Y}) \|$$

$$= \frac{\partial}{\partial W^{[1]}}\left(Y \log \hat{Y}\right) + \frac{\partial}{\partial W^{[1]}}\left((1-Y)\log(1-\hat{Y})\right)$$

$$Y \log(\hat{Y})$$

$$Y \log(\sigma(Z^{[2]}))$$

$$Y \log(\sigma(A^{[1]} W^{[2]T} + b^{[2]T}))$$

$$Y \log(\sigma(\sigma(Z^{[1]}) W^{[2]T} + b^{[2]T}))$$

$$\frac{\partial}{\partial W^{[1]}} Y \log\Big(\underbrace{\sigma(\sigma(A^{[0]} W^{[1]T} + b^{[1]T}) W^{[2]T} + b^{[2]T})}_{\text{FF - Neural Network}}\Big)$$

$$\hat{Y}$$

Backpropagation $b^{[1]}$

Follow the previous slides

Parameter Updates

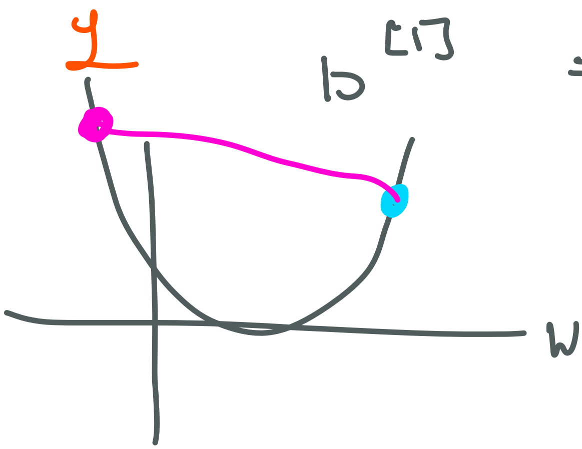
For Loop

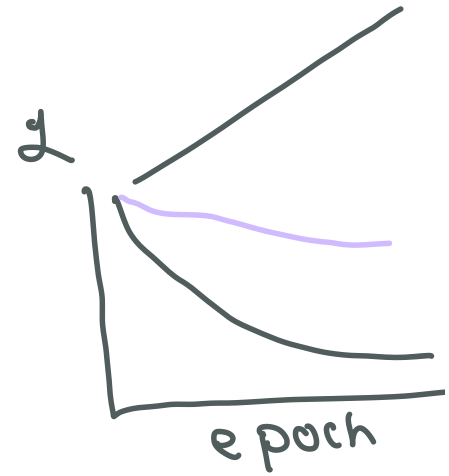
$$W^{[2]} = W^{[2]} - \alpha \overbrace{(\hat{y} - y) A^{[1]}}^{\frac{\partial \mathcal{L}}{\partial W^{[2]}}}$$

$$b^{[2]} = b^{[2]} - \dots$$

$$W^{[1]} = W^{[1]} - \dots$$

$$b^{[1]} = b^{[1]} - \dots$$





$$\alpha = \cancel{0.1}\ 0.01$$

Dropout
Regularization
Batch normalization
RMSProp

Overfitting