

What is Deep Learning?

ARTIFICIAL INTELLIGENCE

Any technique that enables computers to mimic human behavior

MACHINE LEARNING

Ability to learn without explicitly being programmed



DEEP LEARNING

Learn underlying features in data using neural networks



Deep Learning Success: Vision

Image Recognition

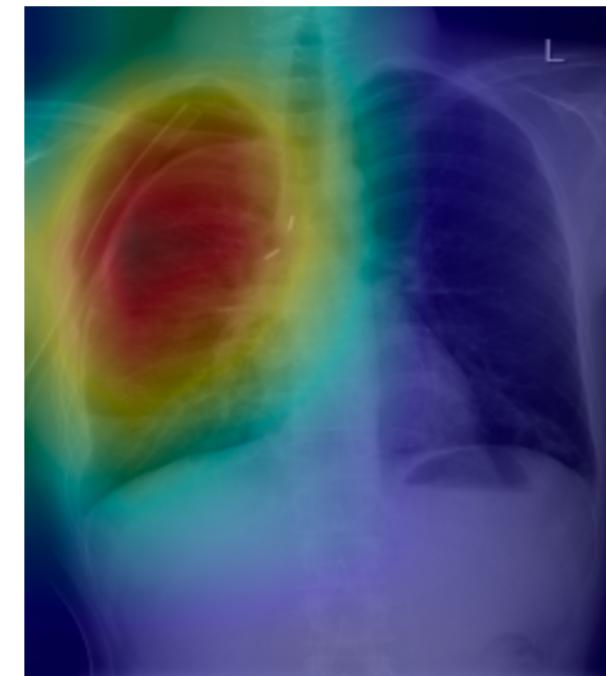
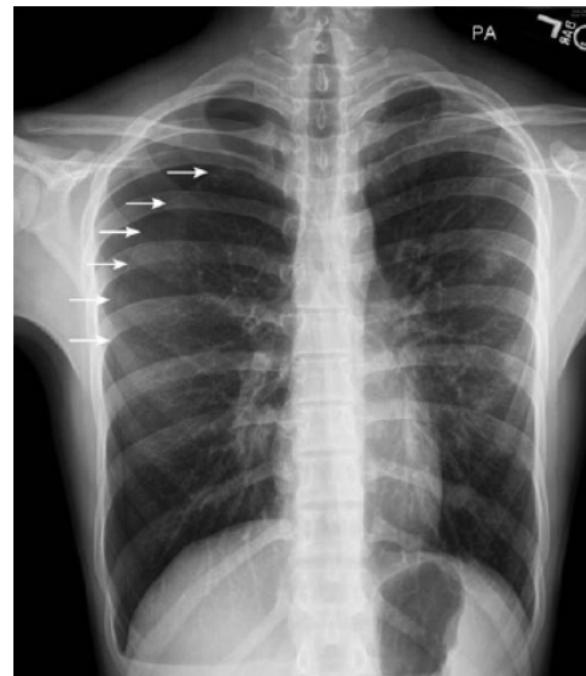


Deep Learning Success: Vision



6.S191 Lab!

Detect pneumothorax in real X-Ray scans



Deep Learning Success: Audio

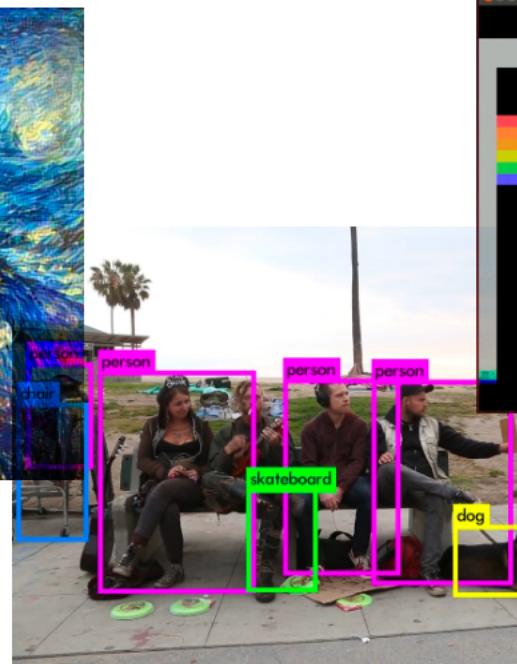


Music Generation

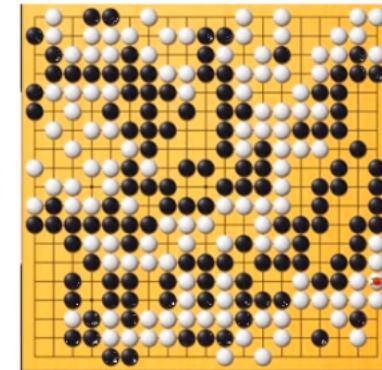


Deep Learning Success

And so many more...



 AlphaGo
ZERO



6.S191 Goals

Fundamentals

Practical Skills

Advancements

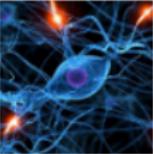
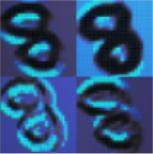
Community @
MIT

Knowledge, intuition, know-how, and community to do
deep learning research and development

Administrative Information

- Meets Mon Jan 29 – Fri Feb 2
- 10:30 am – 1:30pm
- Lecture + Lab Breakdown
- Graded P/D/F; 3 Units
- 1 Final Assignment



| Session | Part 1 | Part 2 | Lab |
|---------|--|--|--|
| 1 |  <p>Introduction to Deep Learning [Slides] [Video] <i>coming soon!</i></p> |  <p>Deep Sequence Modeling [Slides] [Video] <i>coming soon!</i></p> |  <p>Intro to TensorFlow, Music Generation with RNNs [Code] <i>coming soon!</i></p> |
| 2 |  <p>Deep Computer Vision [Slides] [Video] <i>coming soon!</i></p> |  <p>Deep Generative Models [Slides] [Video] <i>coming soon!</i></p> |  <p>Disease Detection from Human X-Ray Scans [Code] <i>coming soon!</i></p> |
| 3 |  <p>Deep Reinforcement Learning [Slides] [Video] <i>coming soon!</i></p> |  <p>Open Challenges and New Frontiers [Slides] [Video] <i>coming soon!</i></p> |  <p>Work time for paper reviews/project proposals</p> |
| 4 |  <p>Guest Lecture: Google [Info] [Slides] [Video] <i>coming soon!</i></p> |  <p>Guest Lecture: NVIDIA [Info] [Slides] [Video] <i>coming soon!</i></p> |  <p>Sponsor Booths + Work time for paper reviews/project proposals</p> |
| 5 |  <p>Guest Lecture: IBM [Info] [Slides] [Video] <i>coming soon!</i></p> |  <p>Guest Lecture: Tencent [Info] [Slides] [Video] <i>coming soon!</i></p> |  <p>Project Proposal Presentations, Judging and Awards</p> |

Final Class Project

Option I: Proposal Presentation

- Groups of 3 or 4
- Present a novel deep learning research idea or application
- 1 slide, 1 minute (strict)
- List of example proposals on website: introtodeeplearning.com
- Presentations on **Friday, Feb 2**
- Submit groups by **Wednesday 5pm** to be eligible
- Submit slide by **Thursday 9pm** to be eligible

- Judged by a panel of industry judges
- Top winners are awarded:



1x NVIDIA Titan Xp
MSRP: \$1200



2x NVIDIA Titan TX2
MSRP: \$600



3x Google Home
MSRP: \$300

Final Class Project

Option 1: Proposal Presentation

- Groups of 3 or 4
- Present a novel deep learning research idea or application
- 1 slide, 1 minute (strict)
- List of example proposals on website: introtodeeplearning.com
- Presentations on **Friday, Feb 2**
- Submit groups by **Wednesday 5pm** to be eligible
- Submit slide by **Thursday 9pm** to be eligible

Option 2: Write a 1-page review of a deep learning paper

- Suggested papers listed on website: introtodeeplearning.com
- Grade is based on clarity of writing and technical communication of main ideas
- Due **Friday 10:30am** (before lecture)

Class Support

- Piazza: <https://piazza.com/class/iwmlwep2fnd5uu>
- Course Website: <http://introtodeeplearning.com>
- Lecture slides: <http://introtodeeplearning.com/#schedule>
- Email us: introtodeeplearning-staff@mit.edu
- Office Hours by request



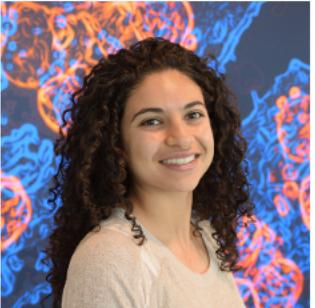
Course Staff



Alexander Amini
Lead Organizer



Harini Suresh
Co-Chair



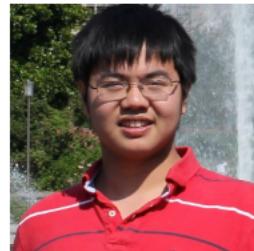
Ava Soleimany
Lead Organizer



Lex Fridman
Co-Chair



Shraman



Yilun



Abhimanyu



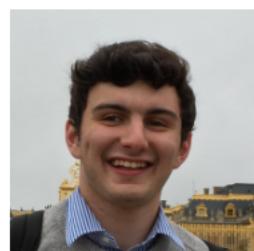
Jose Javier



Alex



Faraaz



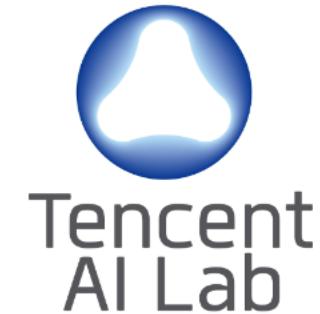
Tim



Bhaktipriya

introtodeeplearning-staff@mit.edu

Thanks to Sponsors!

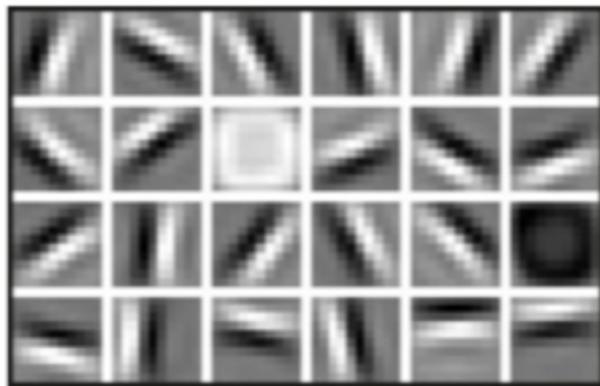


Why Deep Learning?

Hand engineered features are time consuming, brittle and not scalable in practice

Can we learn the **underlying features** directly from data?

Low Level Features



Lines & Edges

Mid Level Features



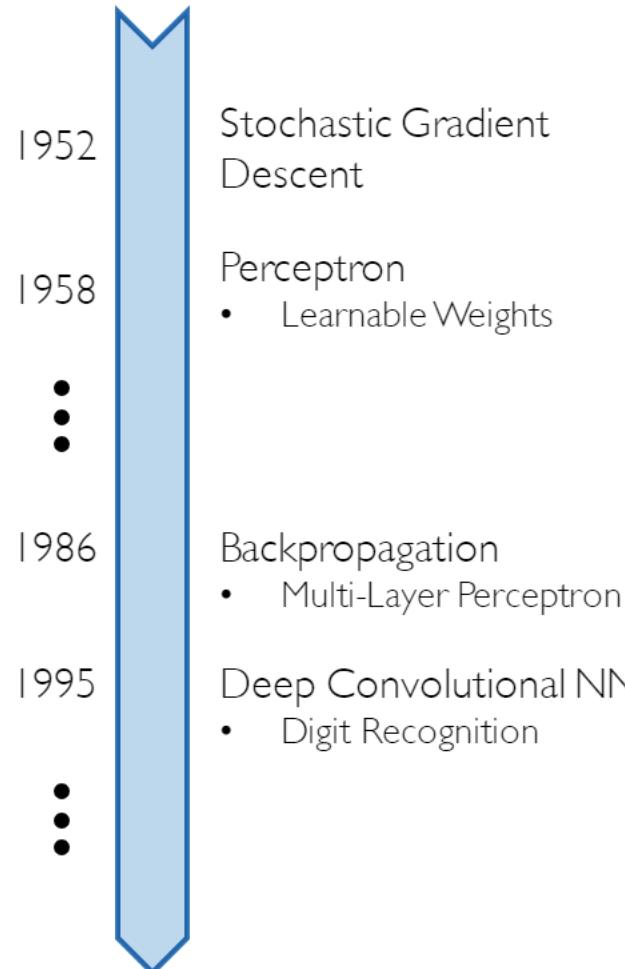
Eyes & Nose & Ears

High Level Features



Facial Structure

Why Now?



Neural Networks date back decades, so why the resurgence?

1. Big Data

- Larger Datasets
- Easier Collection & Storage



WIKIPEDIA
The Free Encyclopedia



2. Hardware

- Graphics Processing Units (GPUs)
- Massively Parallelizable



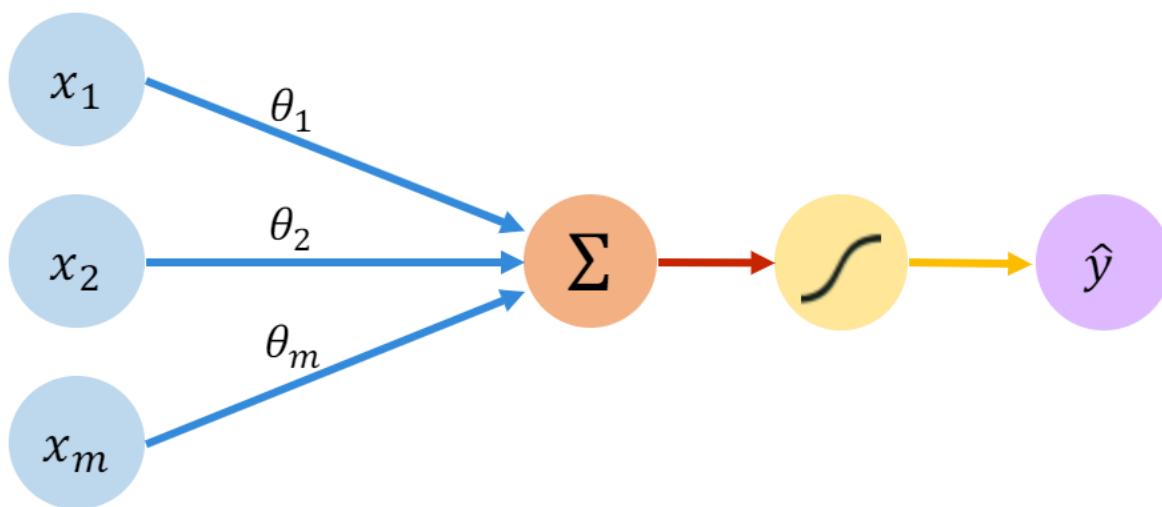
3. Software

- Improved Techniques
- New Models
- Toolboxes



TensorFlow

The Perceptron: Forward Propagation



Inputs Weights Sum Non-Linearity Output

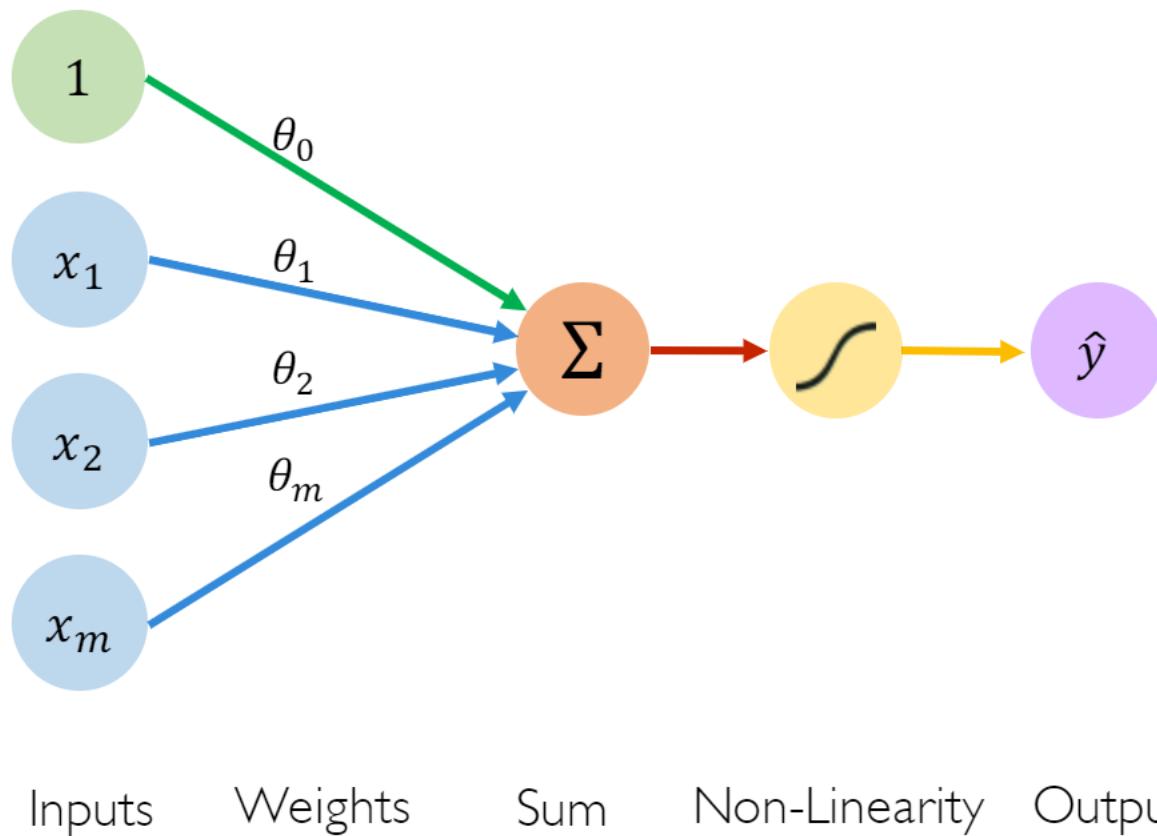
Linear combination
of inputs

Output

$$\hat{y} = g \left(\sum_{i=1}^m x_i \theta_i \right)$$

Non-linear
activation function

The Perceptron: Forward Propagation



Linear combination of inputs

Output

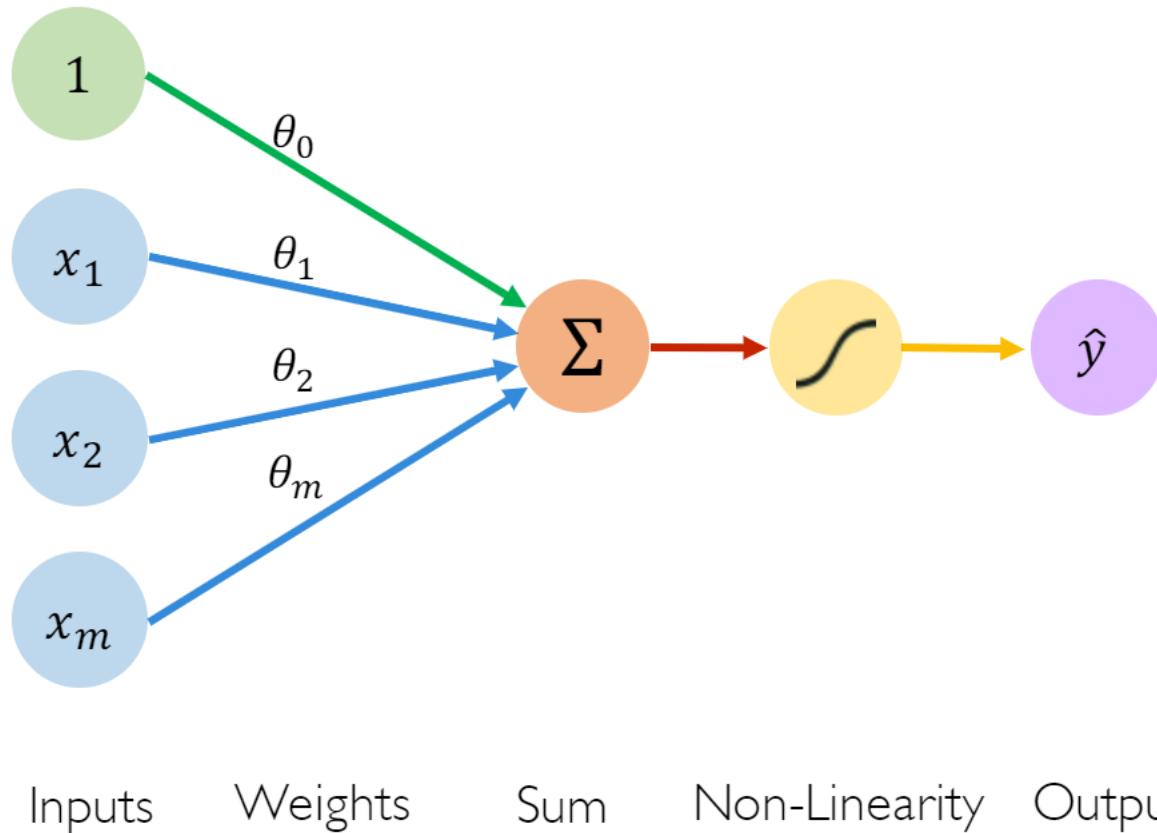
$\hat{y} = g \left(\theta_0 + \sum_{i=1}^m x_i \theta_i \right)$

Non-linear activation function

Bias

Diagram illustrating the mathematical formula for the perceptron's output. The output \hat{y} is the result of applying the activation function g to the linear combination of inputs and weights. The linear combination is formed by adding the bias θ_0 to the sum of the products of each input x_i and its corresponding weight θ_i .

The Perceptron: Forward Propagation

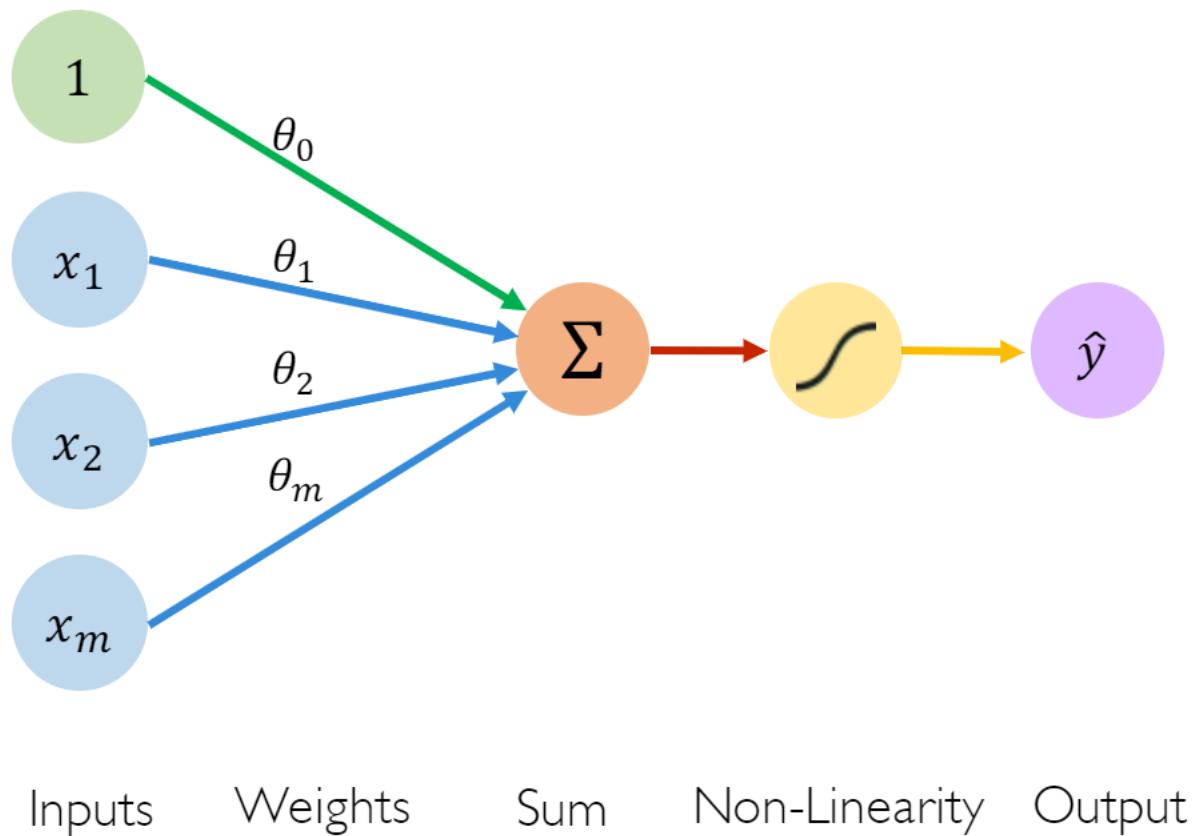


$$\hat{y} = g \left(\theta_0 + \sum_{i=1}^m x_i \theta_i \right)$$

$$\hat{y} = g (\theta_0 + \mathbf{X}^T \boldsymbol{\theta})$$

where: $\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$ and $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$

The Perceptron: Forward Propagation

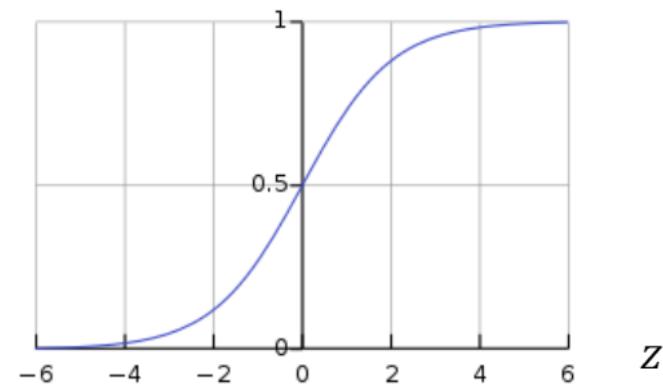


Activation Functions

$$\hat{y} = g(\theta_0 + X^T \theta)$$

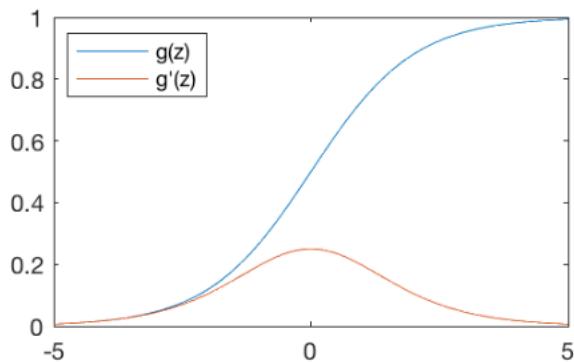
- Example: sigmoid function

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



Common Activation Functions

Sigmoid Function

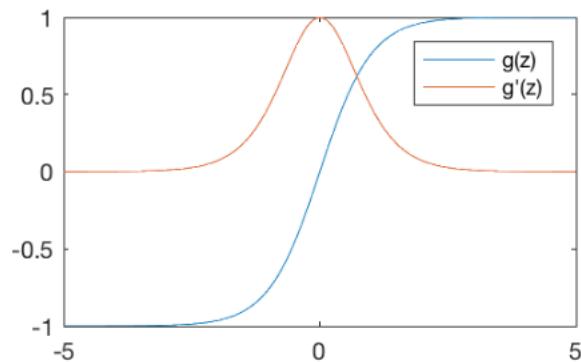


$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(z) = g(z)(1 - g(z))$$



Hyperbolic Tangent

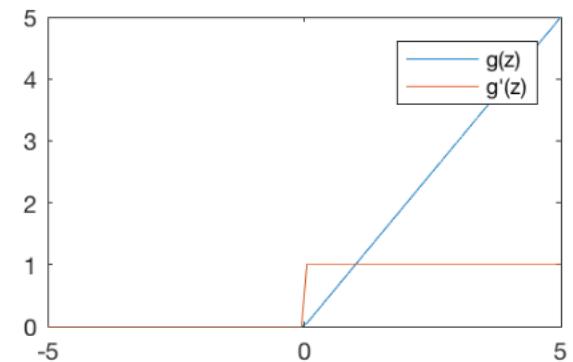


$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g'(z) = 1 - g(z)^2$$



Rectified Linear Unit (ReLU)



$$g(z) = \max(0, z)$$

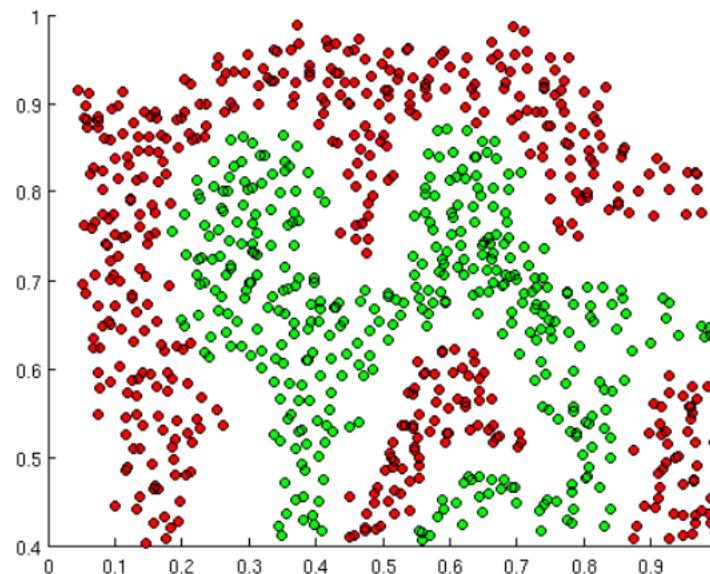
$$g'(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$



NOTE: All activation functions are non-linear

Importance of Activation Functions

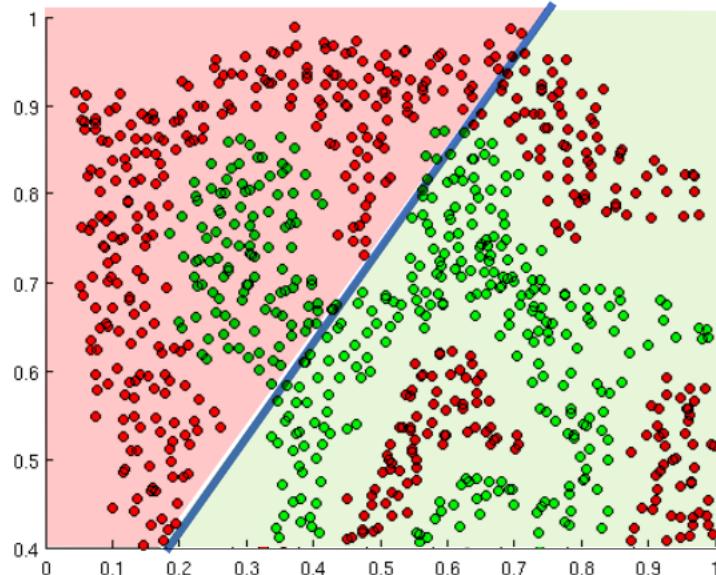
The purpose of activation functions is to **introduce non-linearities** into the network



What if we wanted to build a Neural Network to
distinguish green vs red points?

Importance of Activation Functions

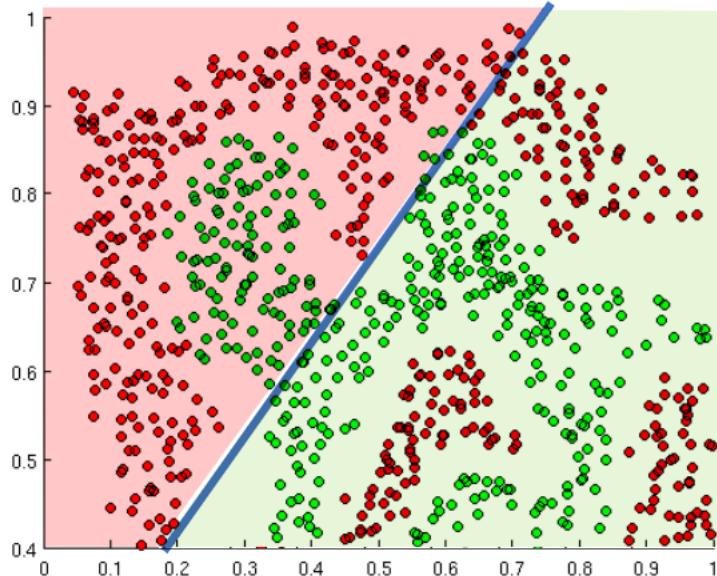
The purpose of activation functions is to **introduce non-linearities** into the network



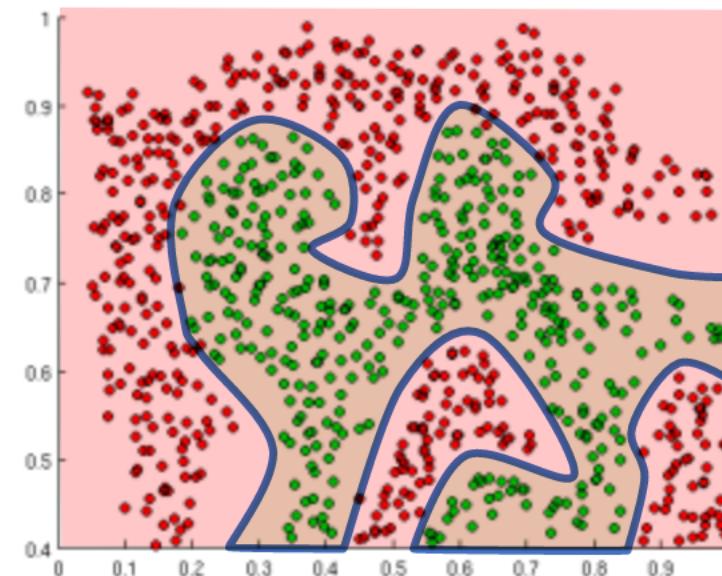
Linear Activation functions produce linear decisions no matter the network size

Importance of Activation Functions

The purpose of activation functions is to **introduce non-linearities** into the network

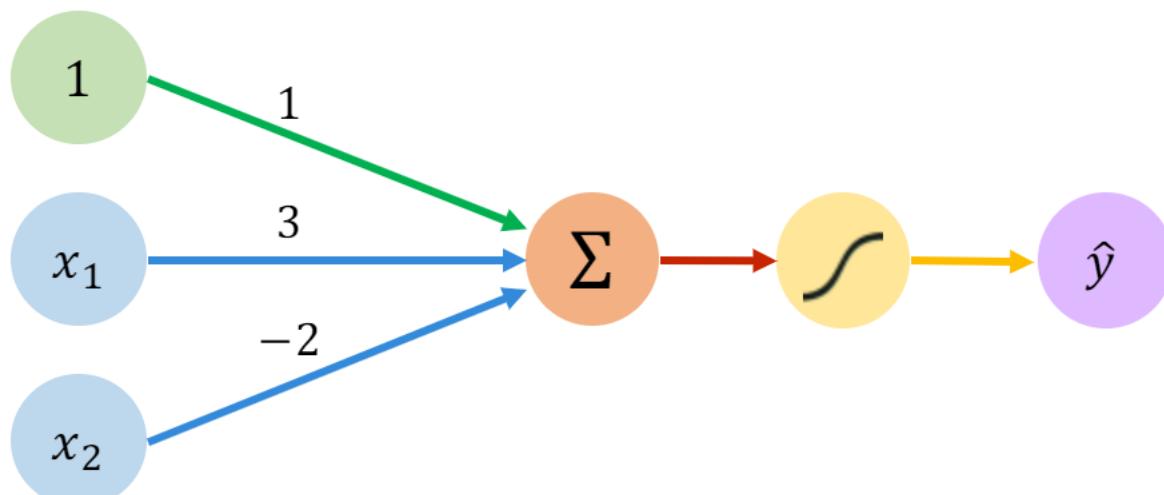


Linear Activation functions produce linear decisions no matter the network size



Non-linearities allow us to approximate arbitrarily complex functions

The Perceptron: Example



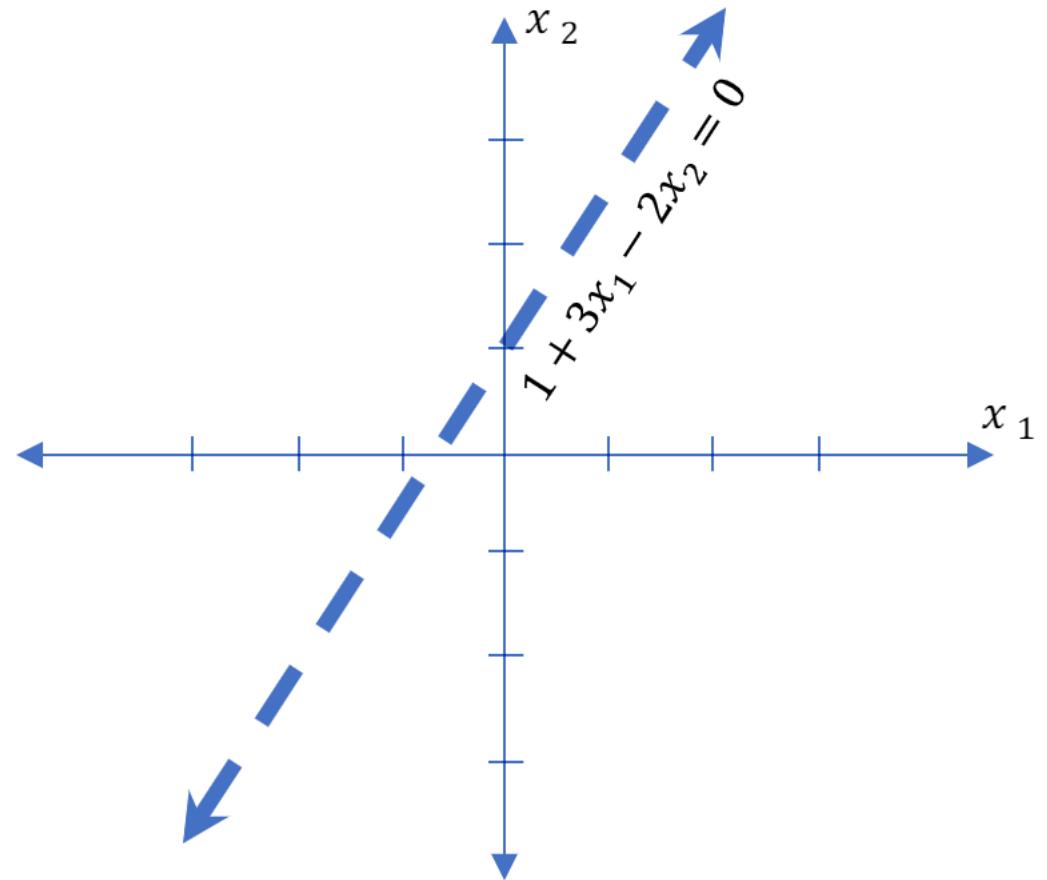
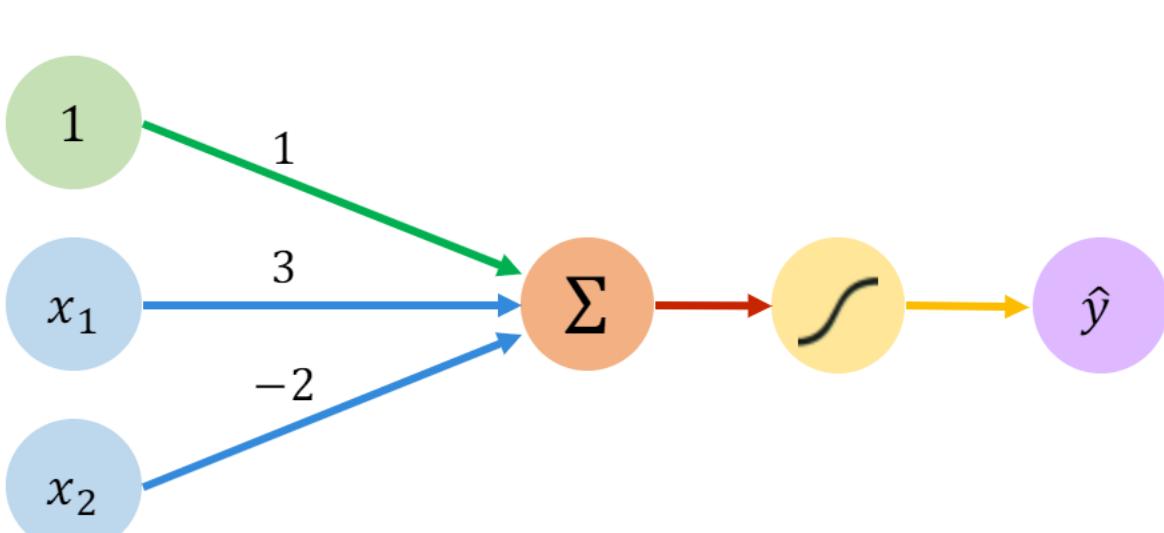
We have: $\theta_0 = 1$ and $\boldsymbol{\theta} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$

$$\begin{aligned}\hat{y} &= g(\theta_0 + \mathbf{X}^T \boldsymbol{\theta}) \\ &= g\left(1 + \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 3 \\ -2 \end{bmatrix}\right) \\ \hat{y} &= g\left(1 + 3x_1 - 2x_2\right)\end{aligned}$$

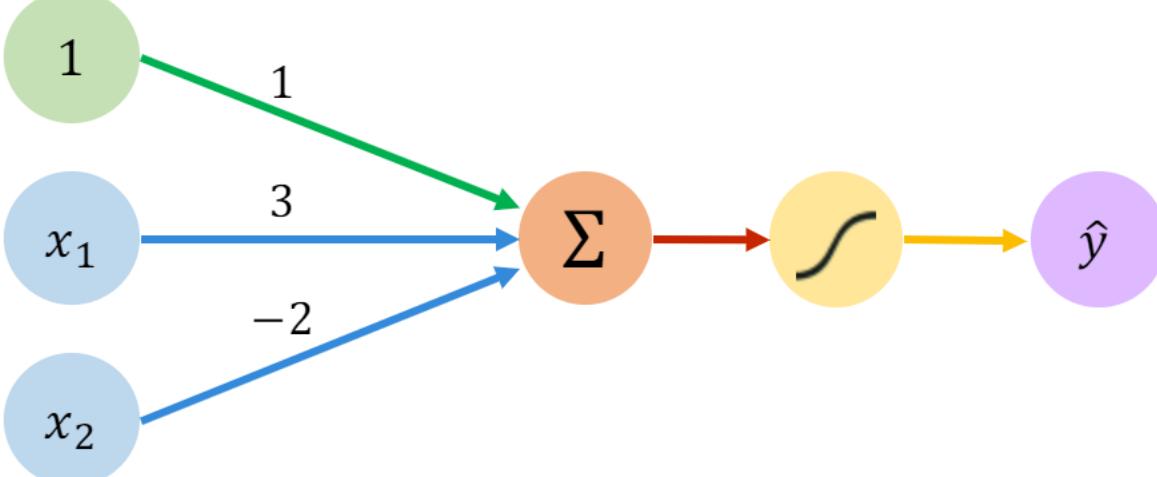
This is just a line in 2D!

The Perceptron: Example

$$\hat{y} = g(1 + 3x_1 - 2x_2)$$



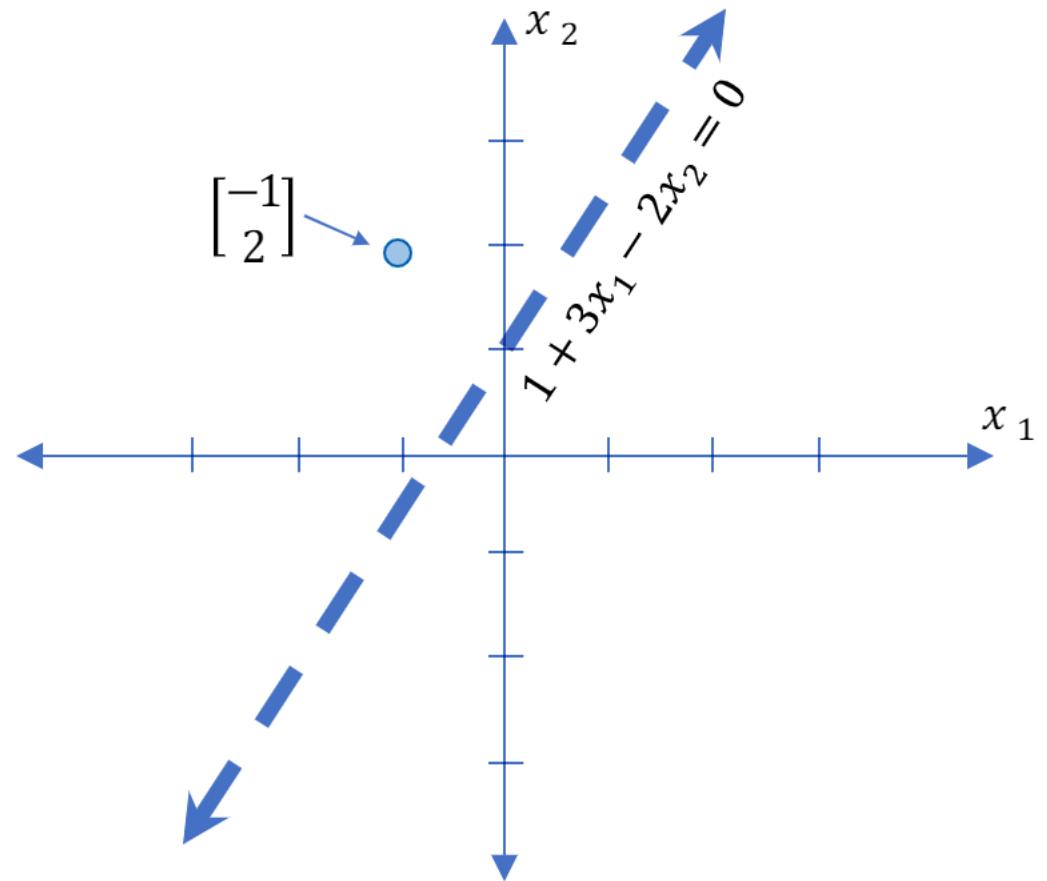
The Perceptron: Example



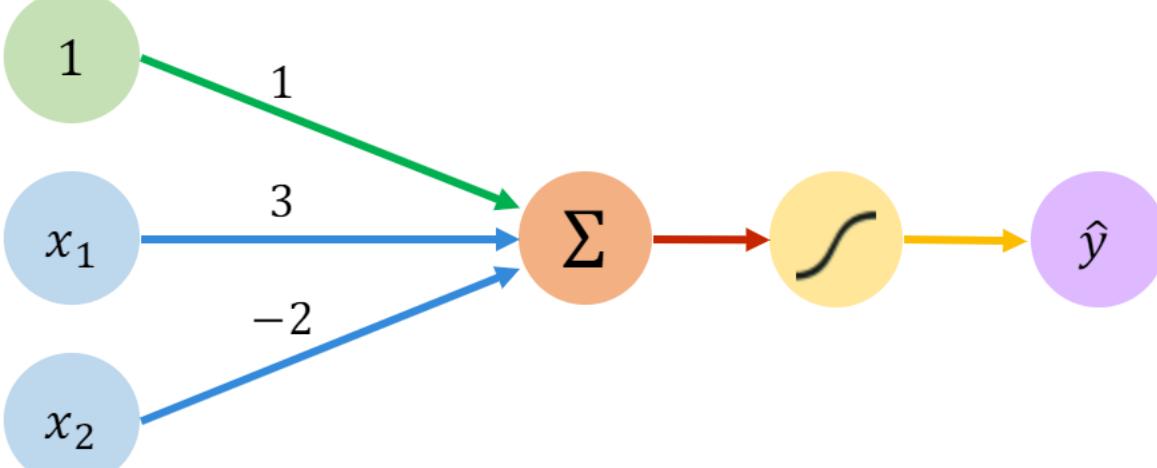
Assume we have input: $X = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

$$\begin{aligned}\hat{y} &= g(1 + (3 * -1) - (2 * 2)) \\ &= g(-6) \approx 0.002\end{aligned}$$

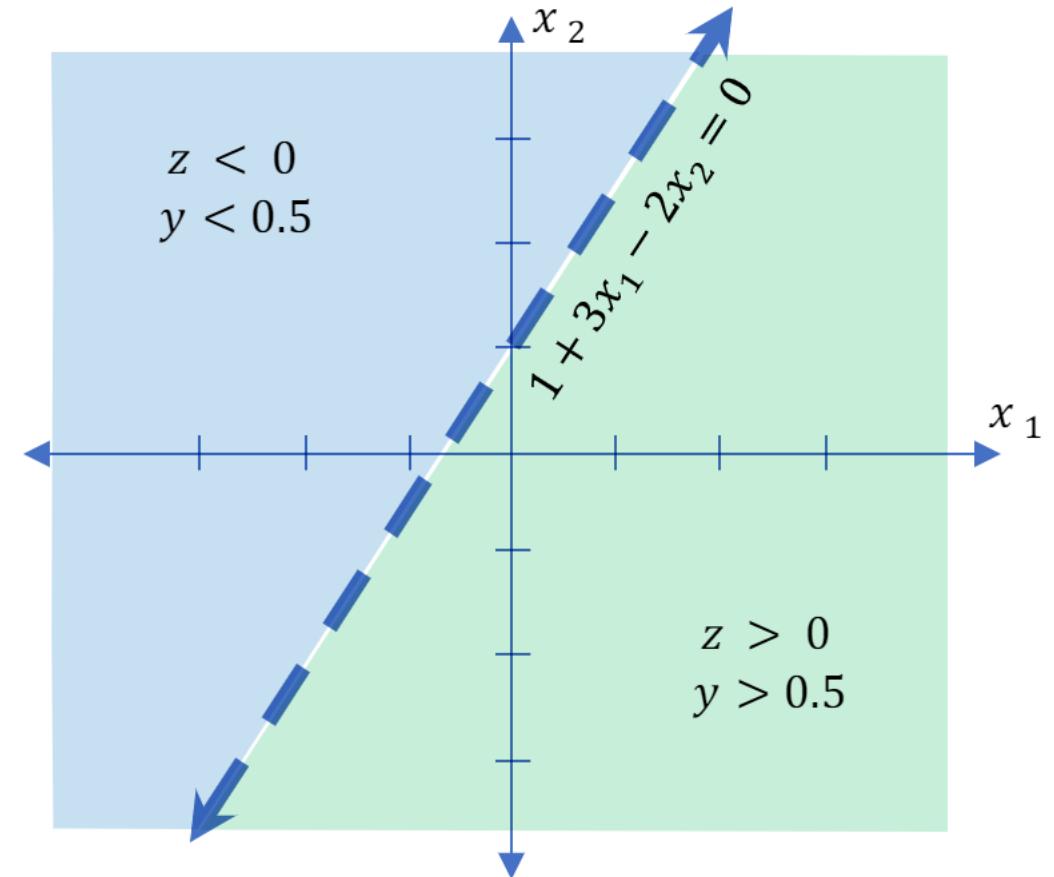
$$\hat{y} = g(1 + 3x_1 - 2x_2)$$



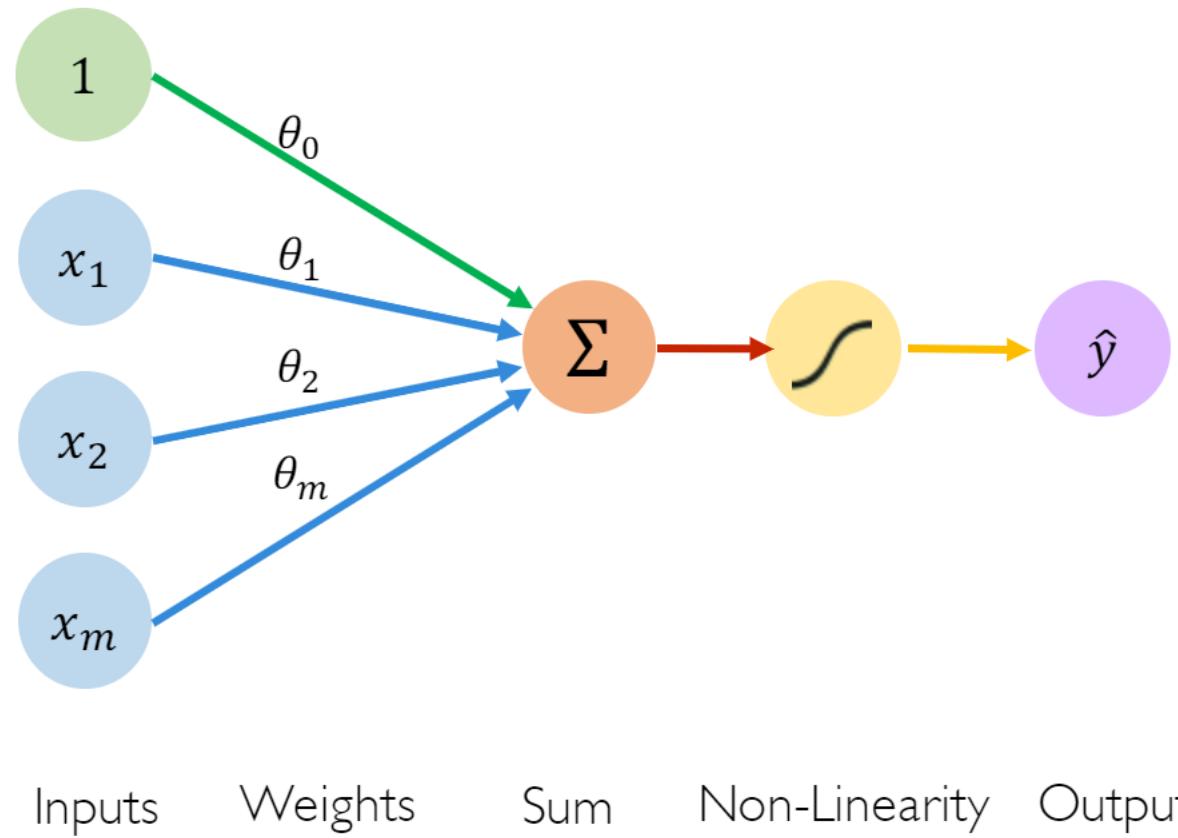
The Perceptron: Example



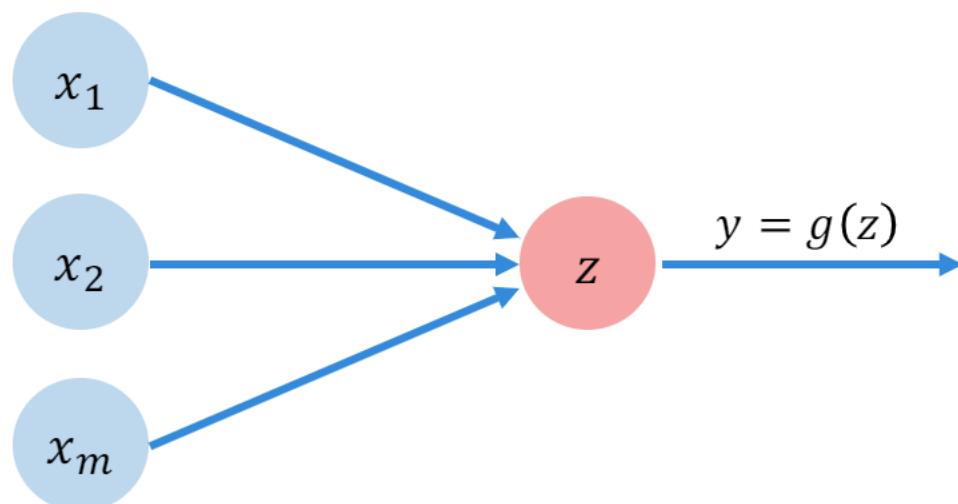
$$\hat{y} = g(1 + 3x_1 - 2x_2)$$



The Perceptron: Simplified

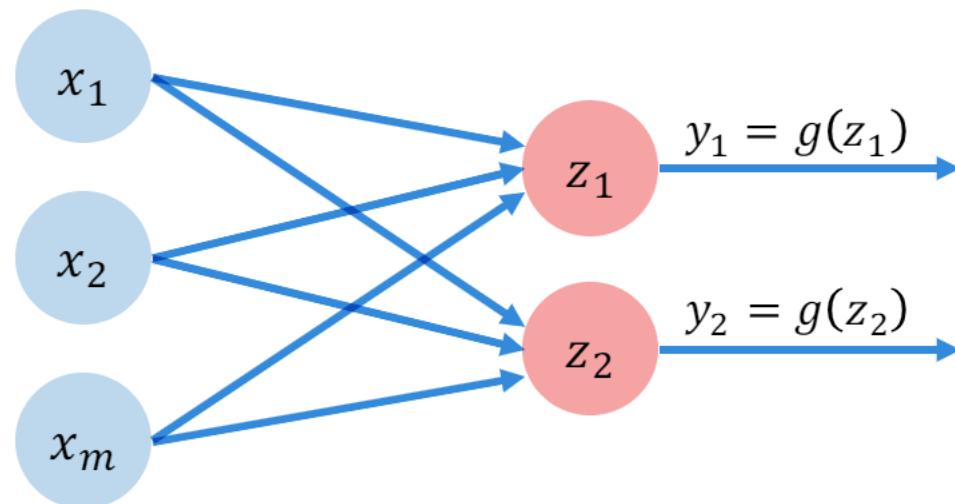


The Perceptron: Simplified



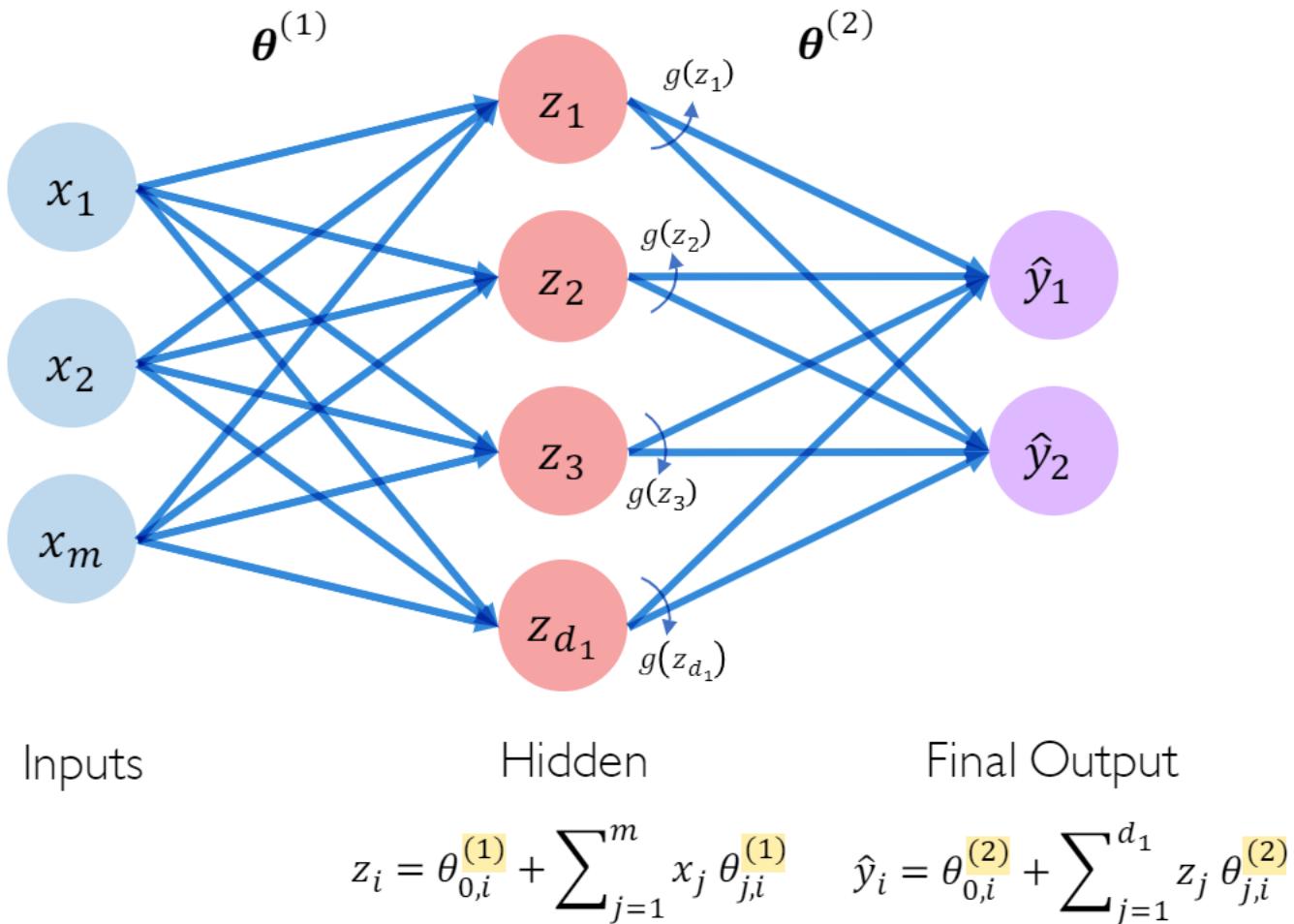
$$z = \theta_0 + \sum_{j=1}^m x_j \theta_j$$

Multi Output Perceptron

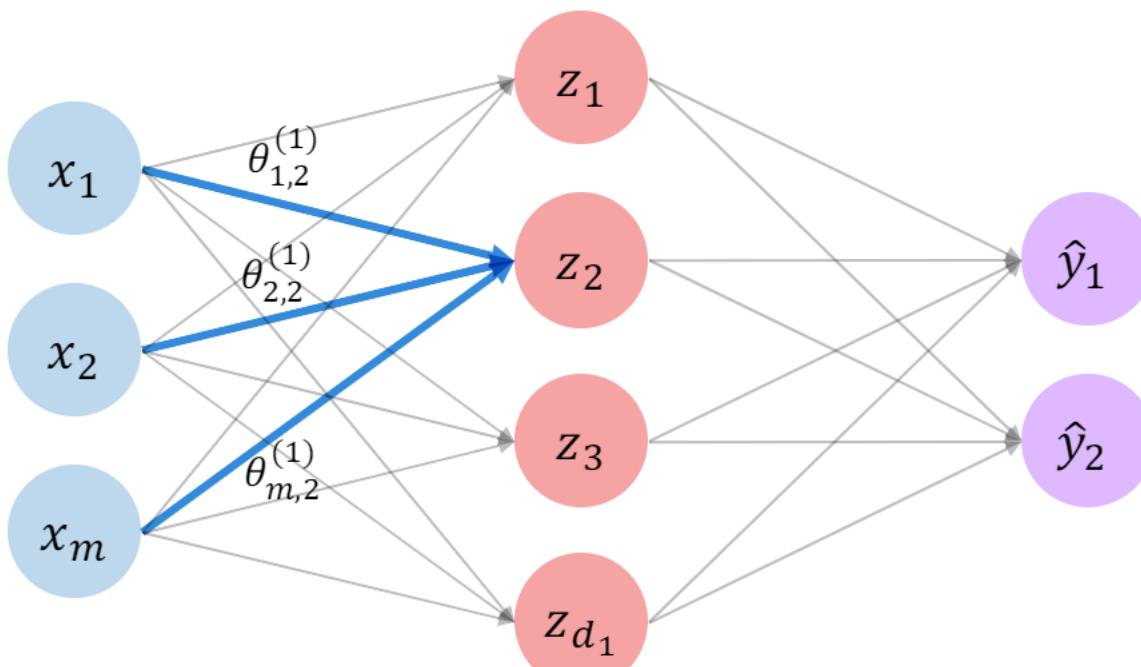


$$z_i = \theta_{0,i} + \sum_{j=1}^m x_j \theta_{j,i}$$

Single Layer Neural Network

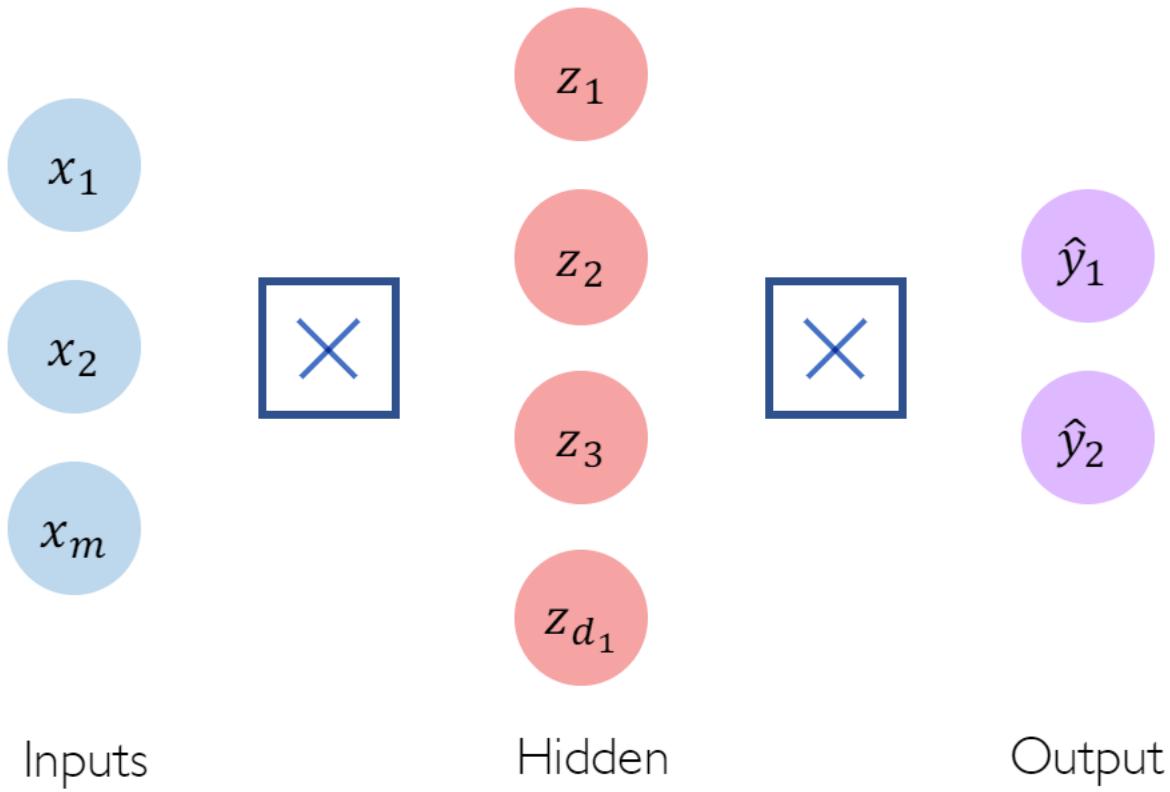


Single Layer Neural Network

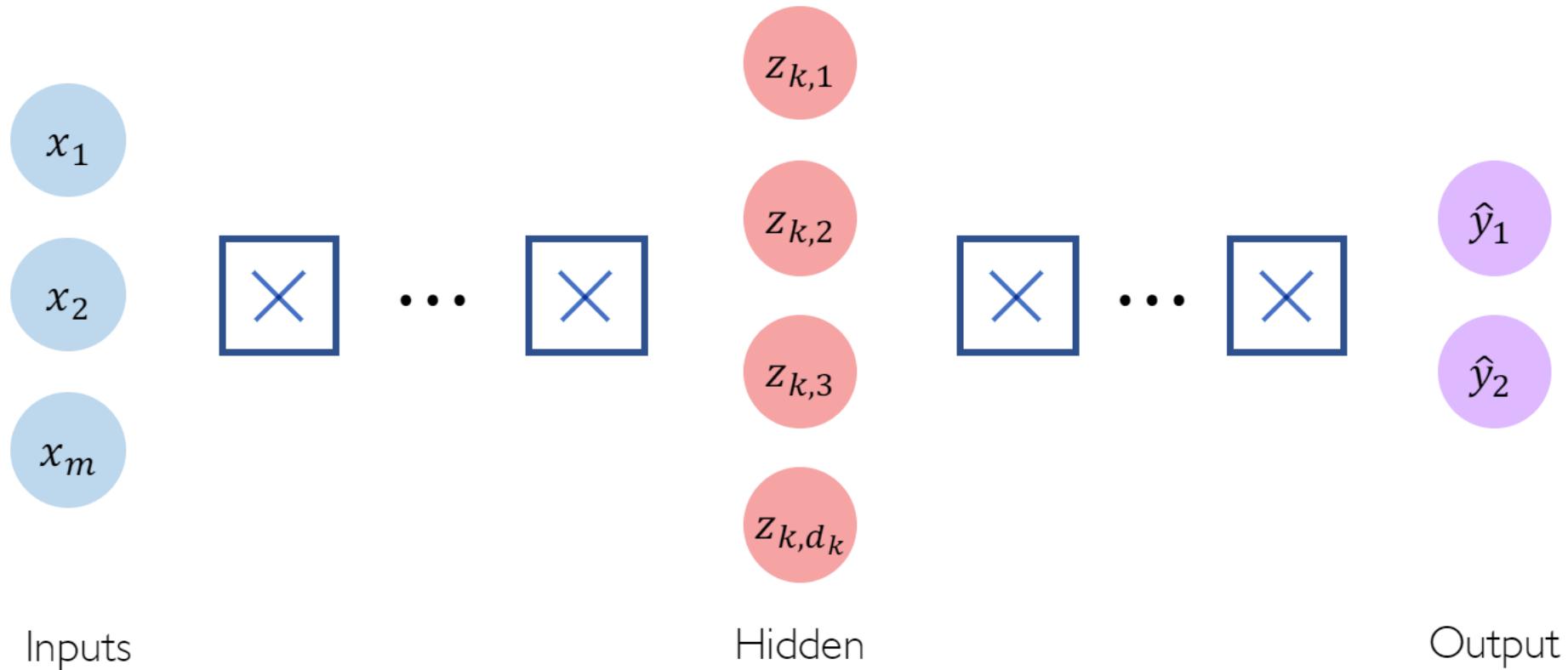


$$\begin{aligned} z_2 &= \theta_{0,2}^{(1)} + \sum_{j=1}^m x_j \theta_{j,2}^{(1)} \\ &= \theta_{0,2}^{(1)} + x_1 \theta_{1,2}^{(1)} + x_2 \theta_{2,2}^{(1)} + x_m \theta_{m,2}^{(1)} \end{aligned}$$

Multi Output Perceptron



Deep Neural Network



$$z_{k,i} = \theta_{0,i}^{(k)} + \sum_{j=1}^{d_{k-1}} g(z_{k-1,j}) \theta_{j,i}^{(k)}$$

Example Problem

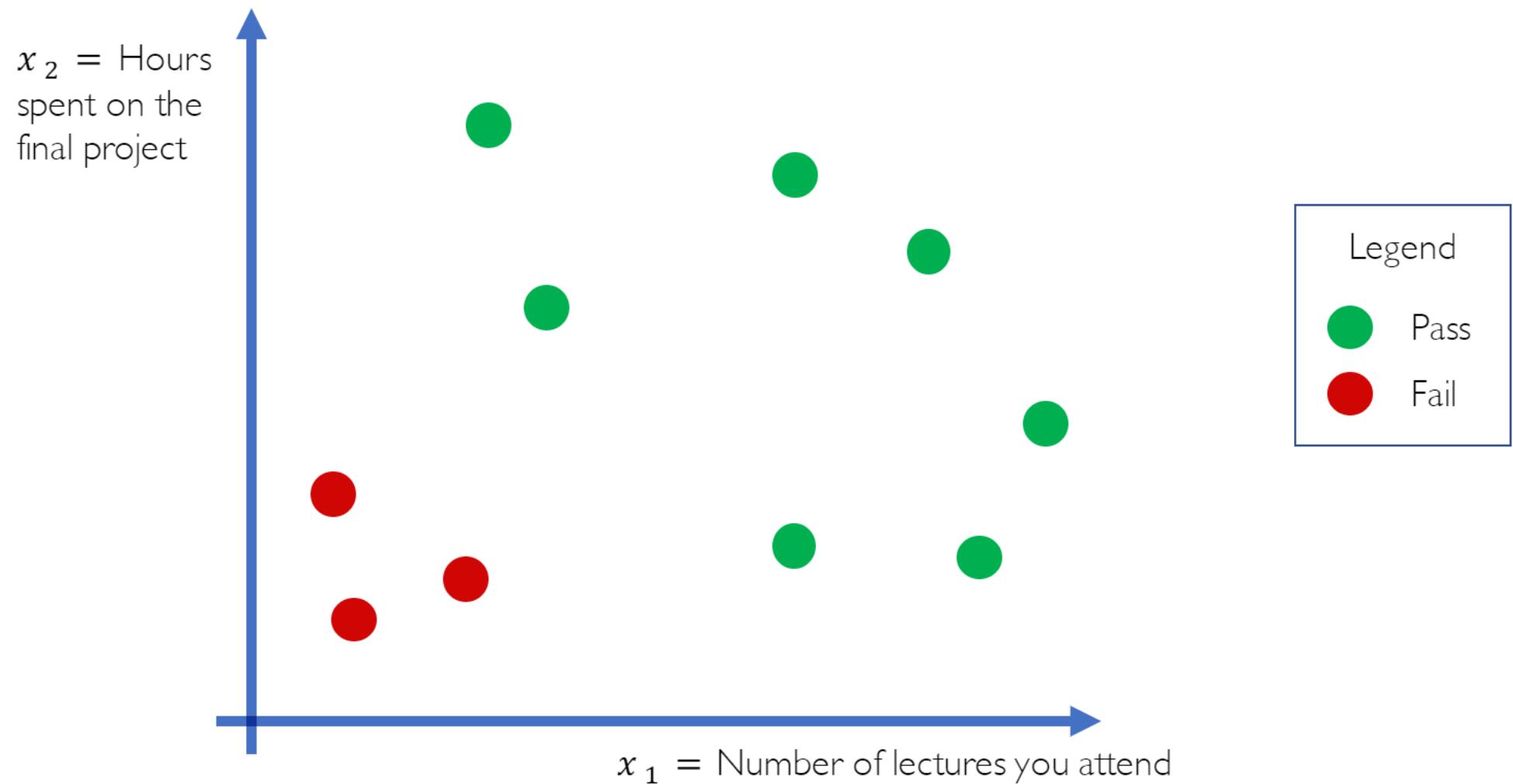
Will I pass this class?

Let's start with a simple two feature model

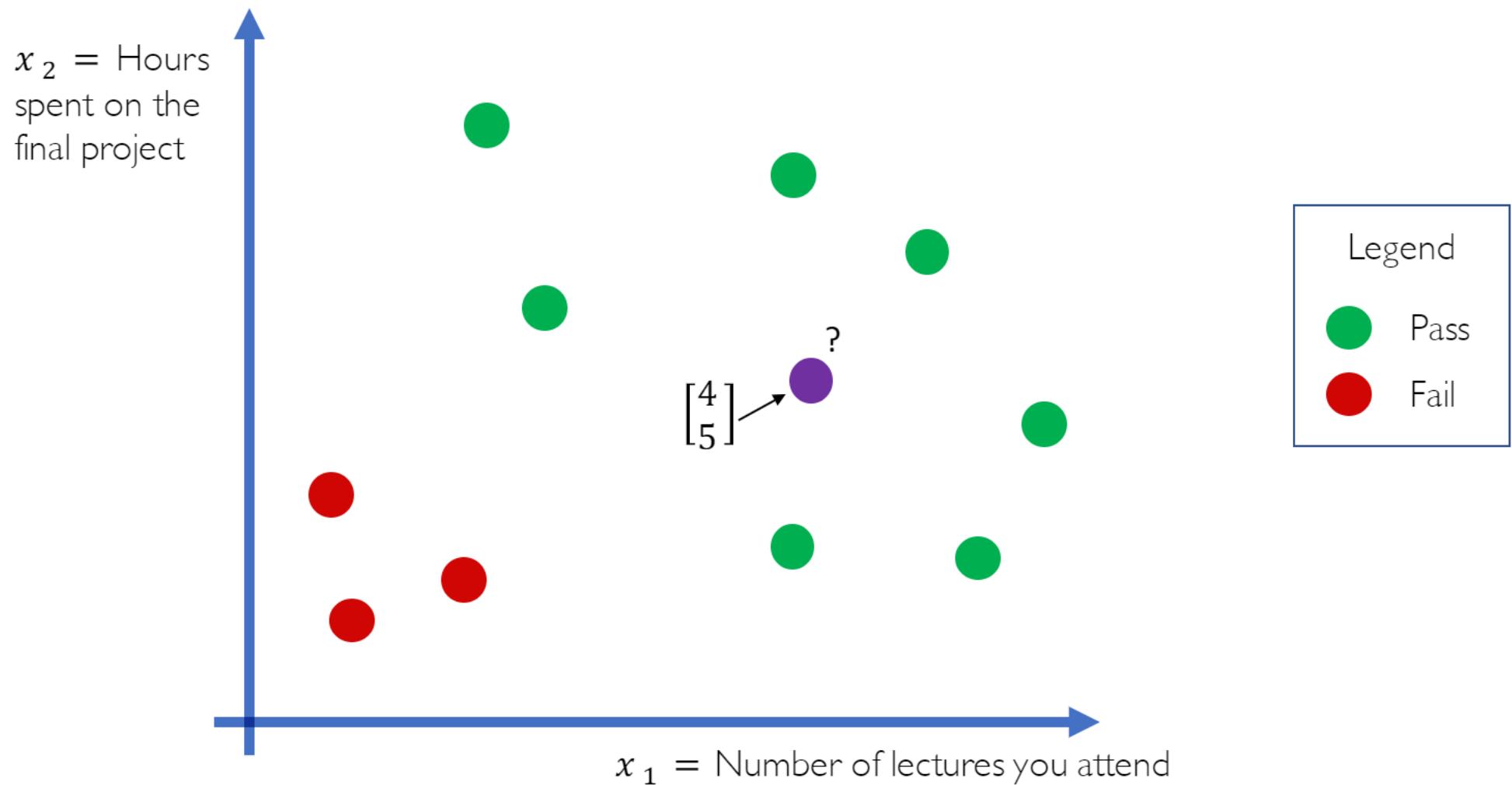
x_1 = Number of lectures you attend

x_2 = Hours spent on the final project

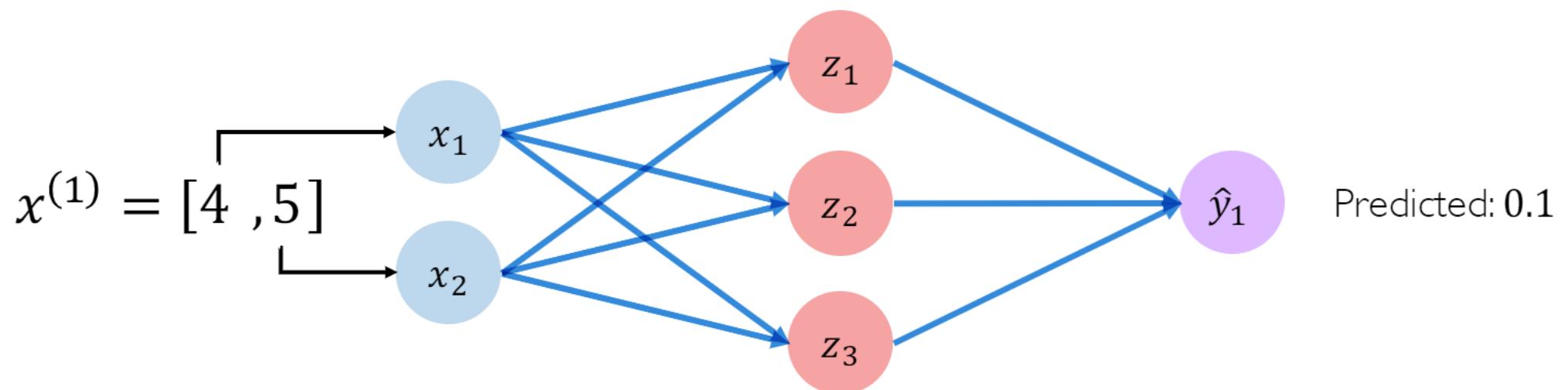
Example Problem: Will I pass this class?



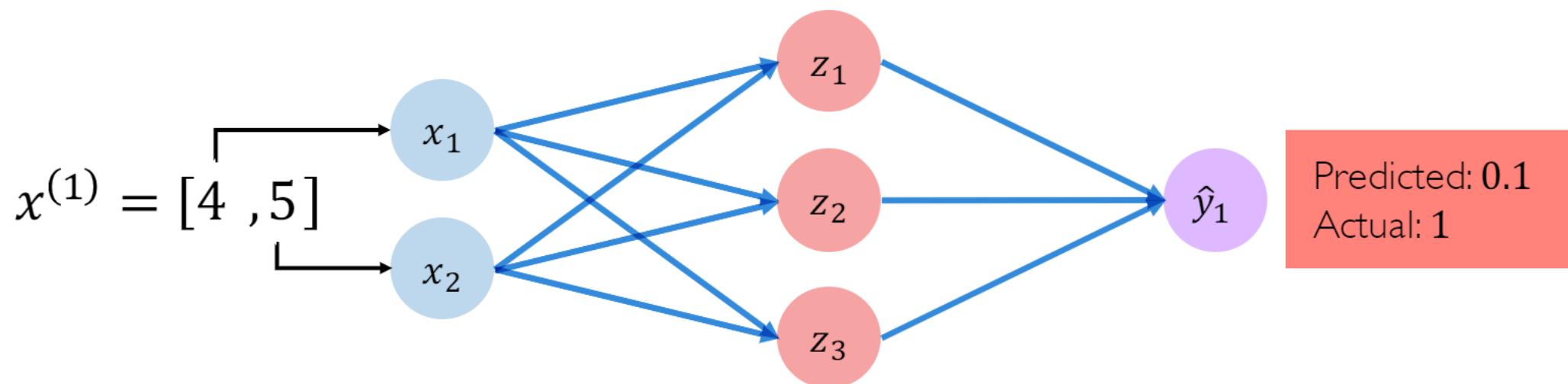
Example Problem: Will I pass this class?



Example Problem: Will I pass this class?

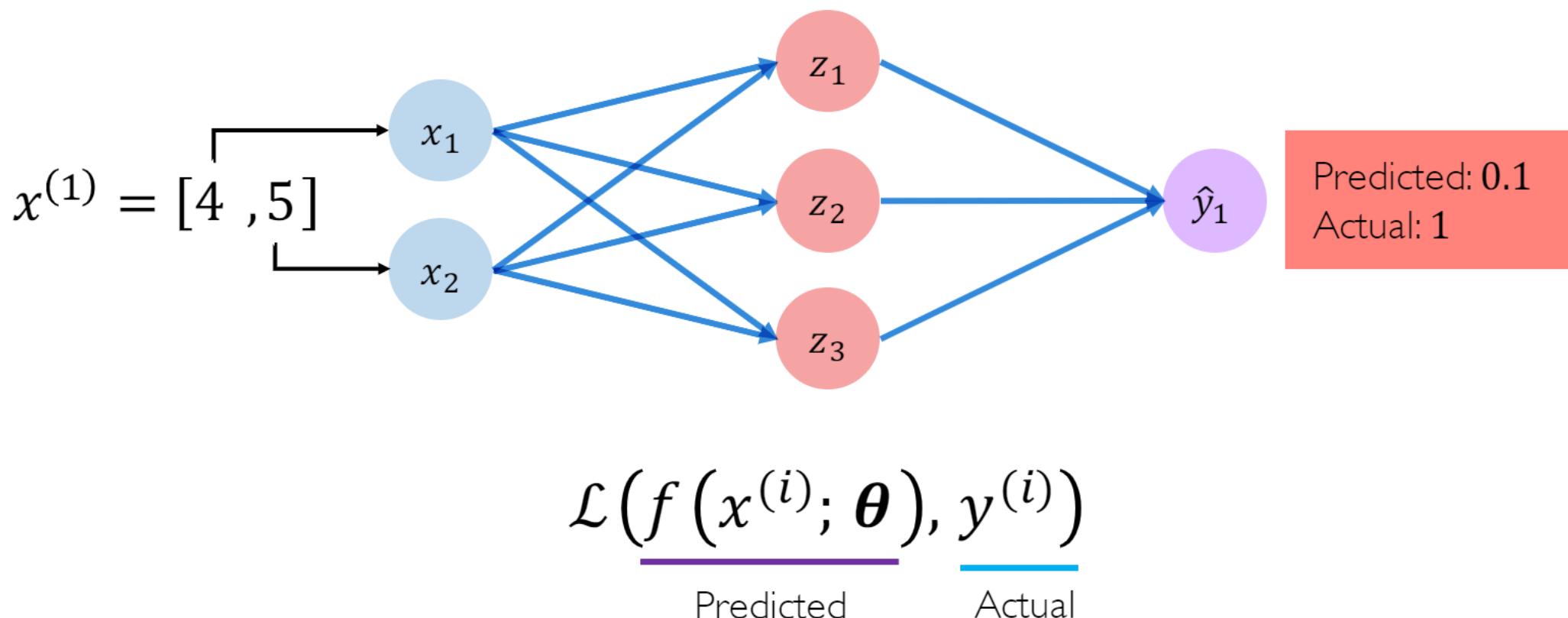


Example Problem: Will I pass this class?



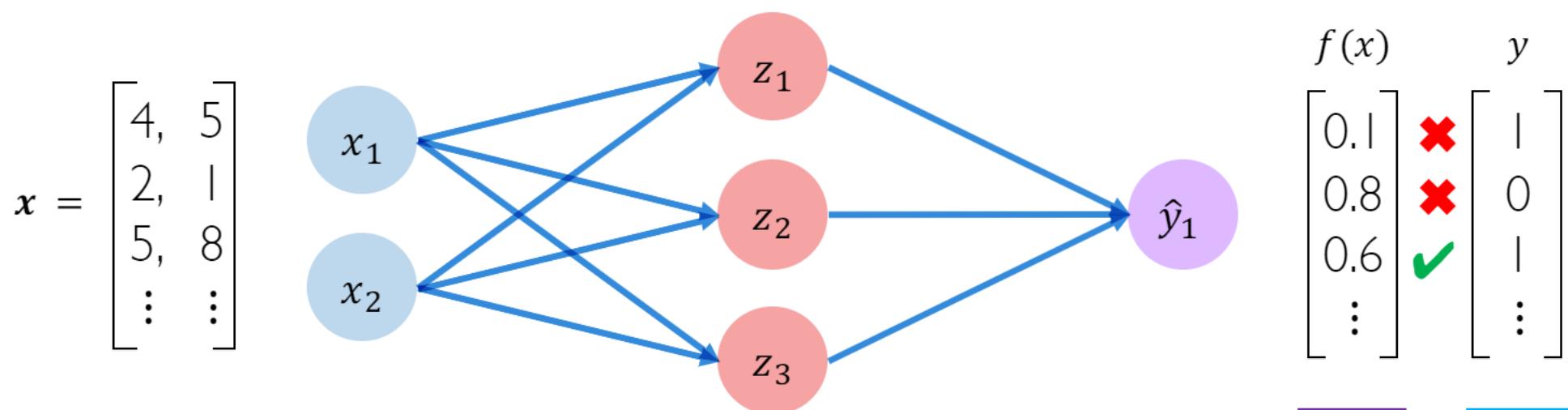
Quantifying Loss

The *loss* of our network measures the cost incurred from incorrect predictions



Empirical Loss

The **empirical loss** measures the total loss over our entire dataset



Also known as:

- Objective function
- Cost function
- Empirical Risk

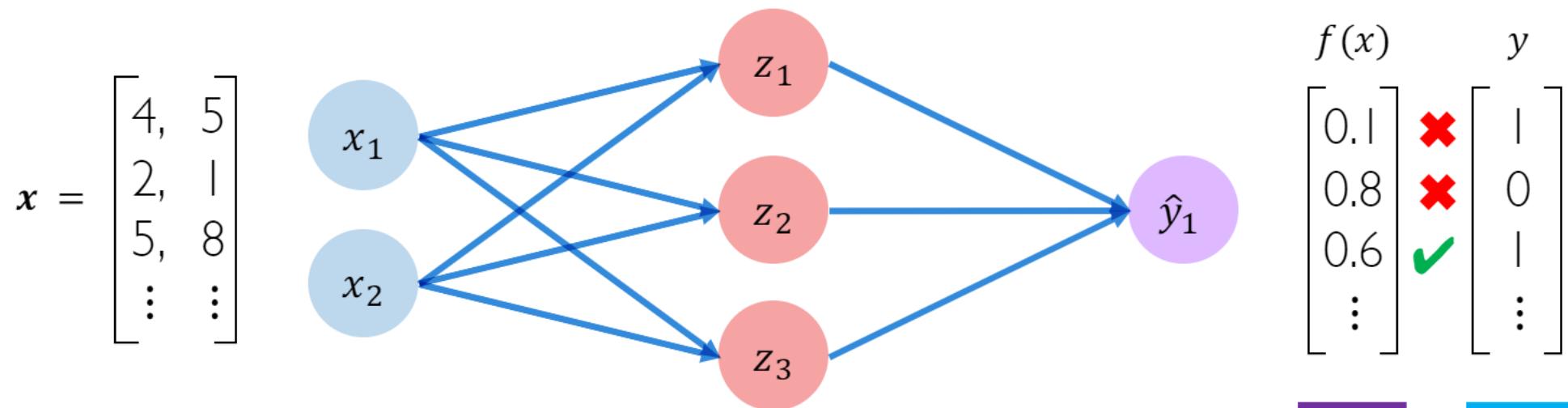


$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x^{(i)}; \theta), y^{(i)})$$

Predicted Actual

Binary Cross Entropy Loss

Cross entropy loss can be used with models that output a probability between 0 and 1



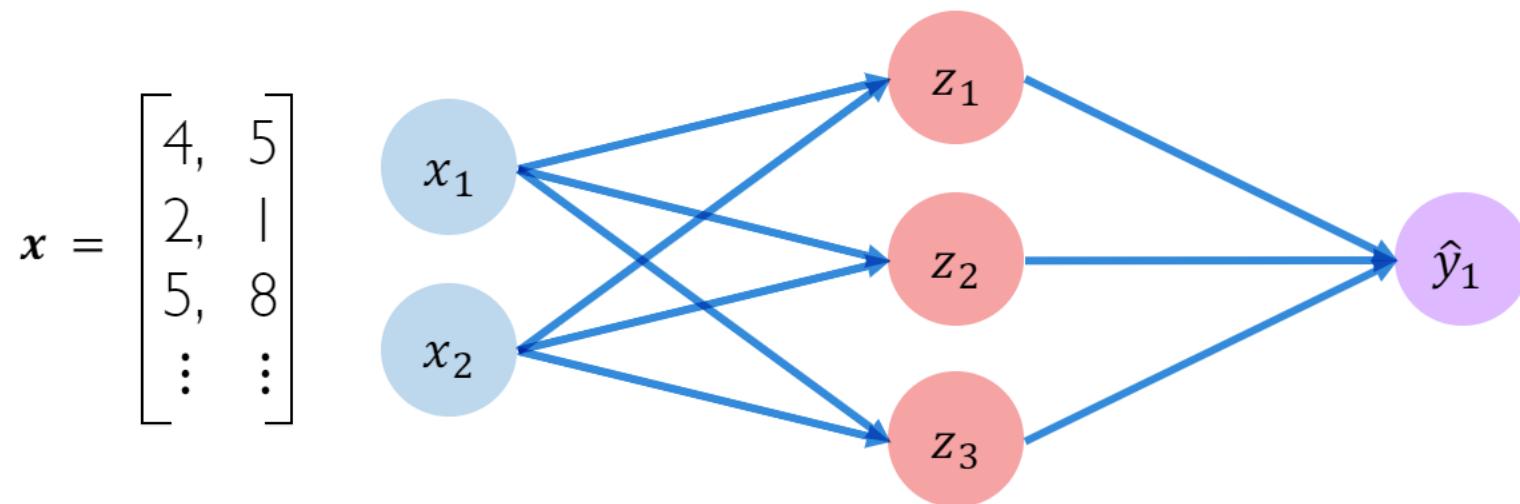
$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \underbrace{y^{(i)} \log(f(x^{(i)}; \theta))}_{\text{Actual}} + \underbrace{(1 - y^{(i)}) \log(1 - f(x^{(i)}; \theta))}_{\text{Actual}}$$

Predicted Actual Predicted Actual



Mean Squared Error Loss

Mean squared error loss can be used with regression models that output continuous real numbers



$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \left(\underline{\text{Actual}} \underbrace{y^{(i)} - f(x^{(i)}; \theta)}_{\text{Predicted}} \right)^2$$

| $f(x)$ | y |
|----------|--|
| 30 | 90 ✗ |
| 80 | 20 ✗ |
| 85 | 95 ✓ |
| \vdots | \vdots |

Final Grades (percentage)



Loss Optimization

We want to find the network weights that **achieve the lowest loss**

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x^{(i)}; \boldsymbol{\theta}), y^{(i)})$$

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Loss Optimization

We want to find the network weights that **achieve the lowest loss**

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x^{(i)}; \boldsymbol{\theta}), y^{(i)})$$

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

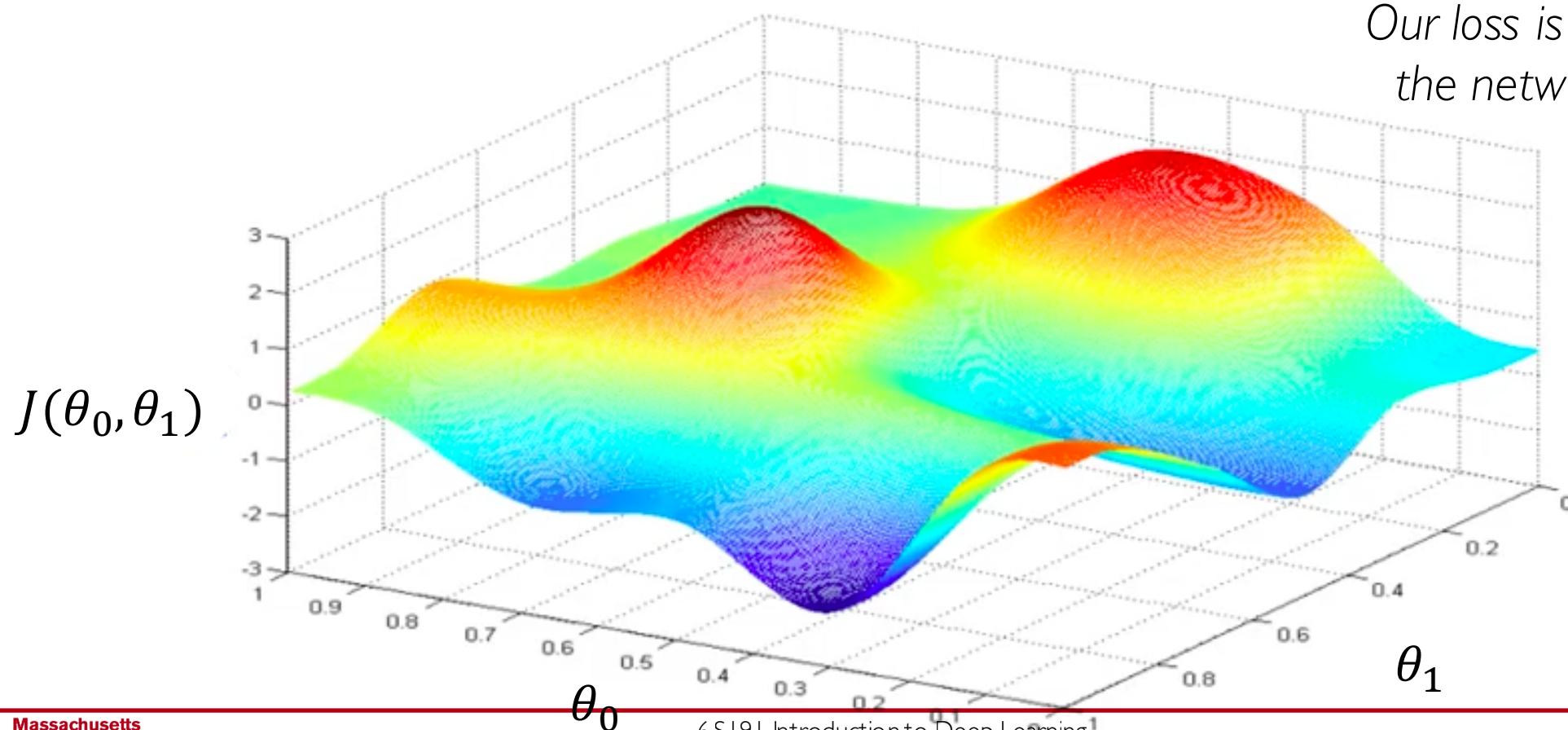


Remember:

$$\boldsymbol{\theta} = \{\theta^{(0)}, \theta^{(1)}, \dots\}$$

Loss Optimization

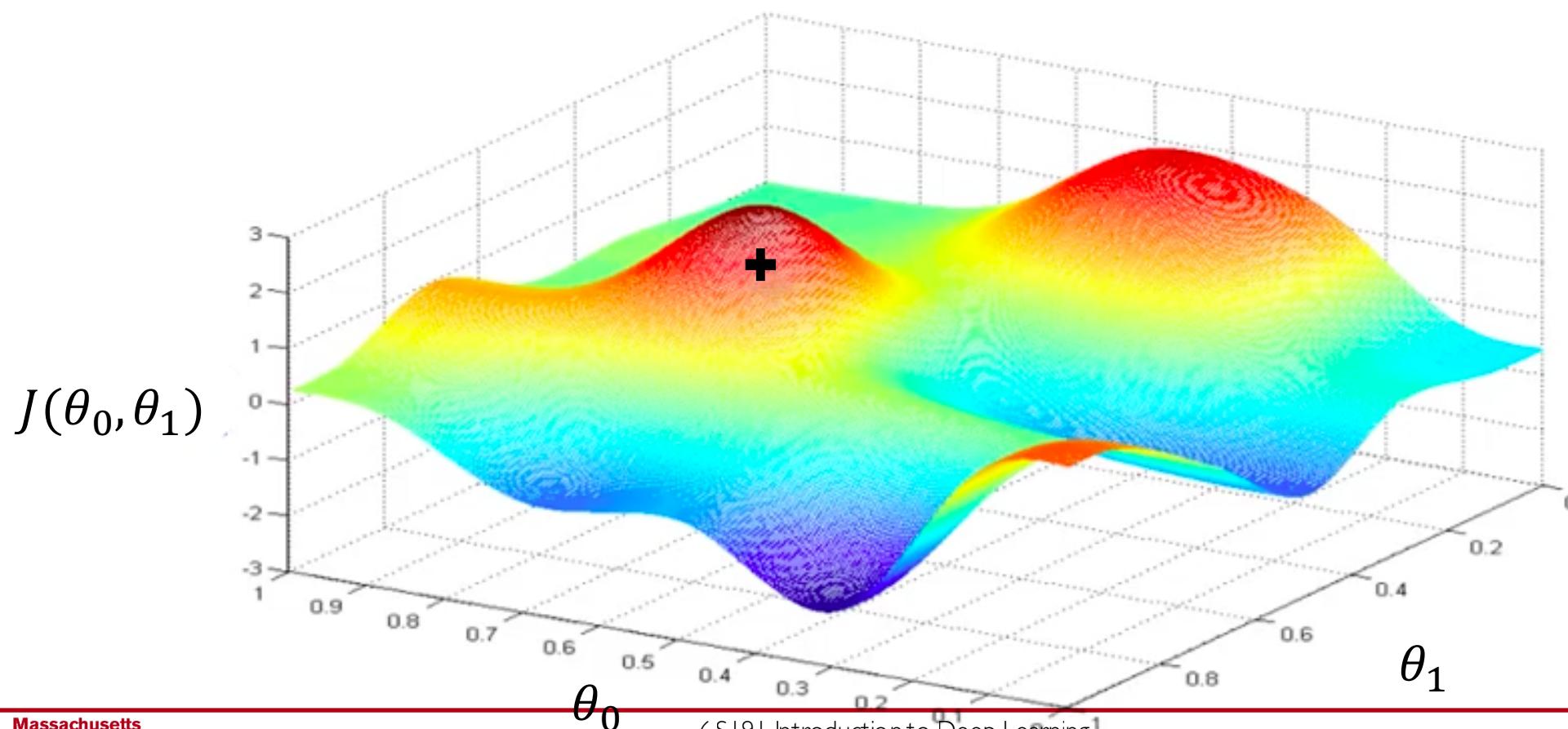
$$\theta^* = \underset{\theta}{\operatorname{argmin}} J(\theta)$$



Remember:
Our loss is a function of
the network weights!

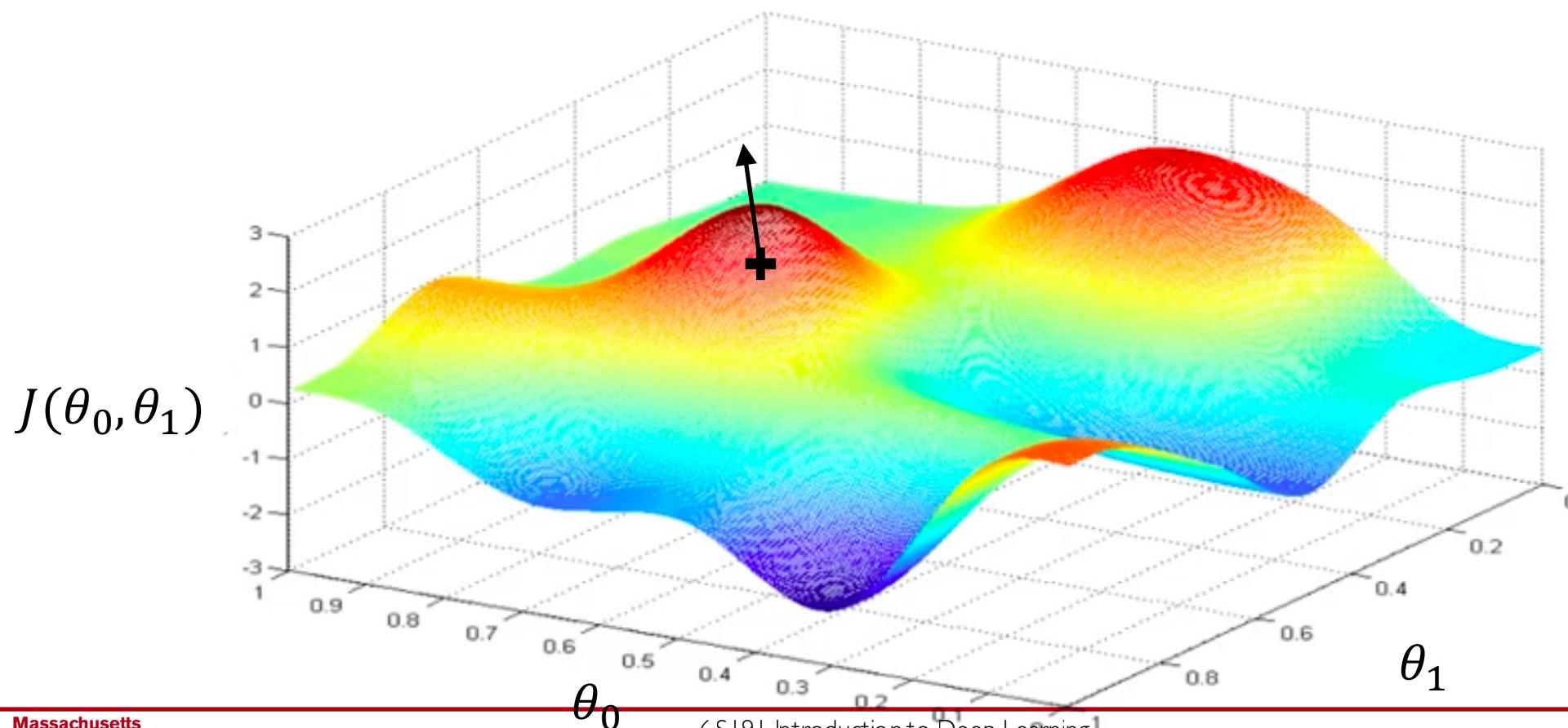
Loss Optimization

Randomly pick an initial (θ_0, θ_1)



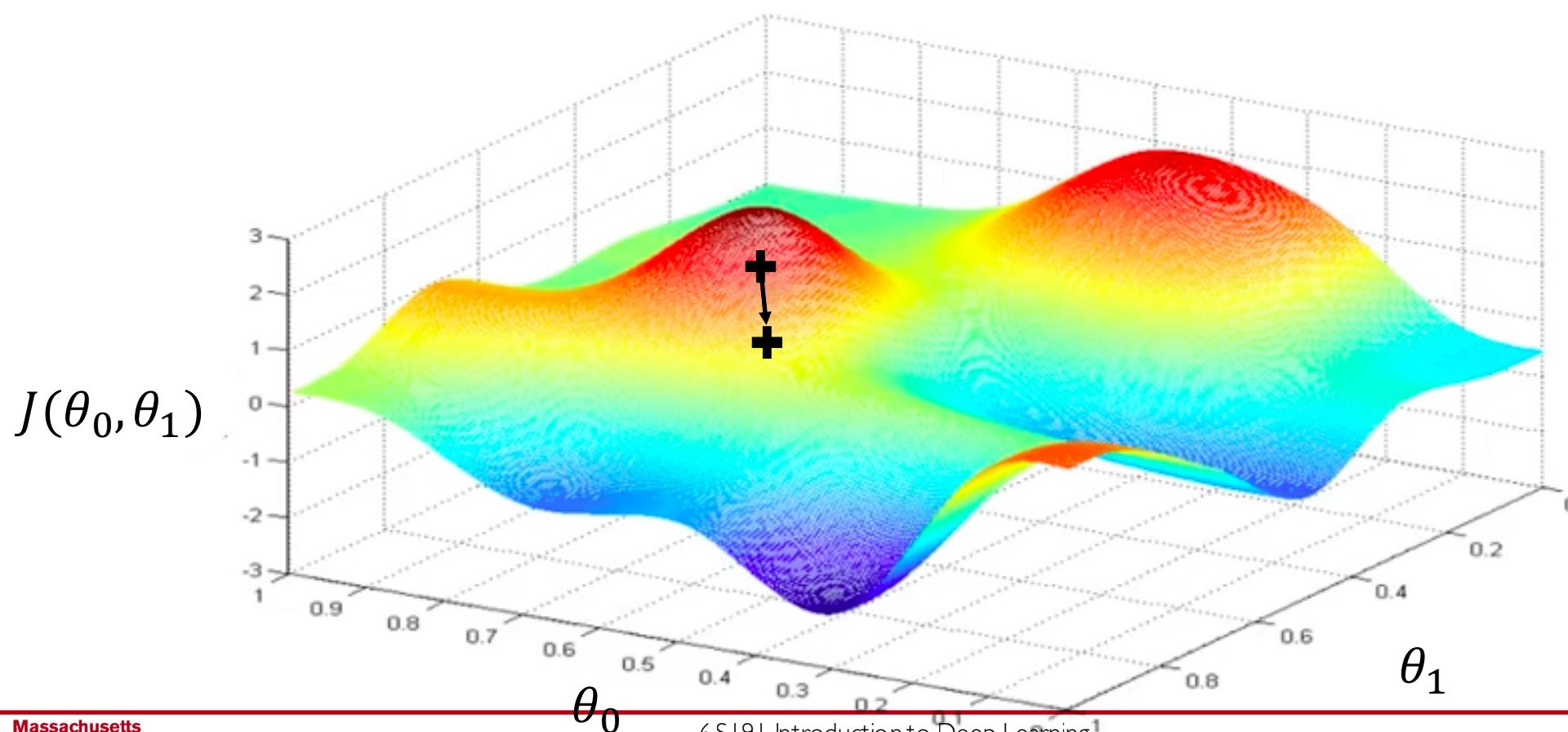
Loss Optimization

Compute gradient, $\frac{\partial J(\theta)}{\partial \theta}$



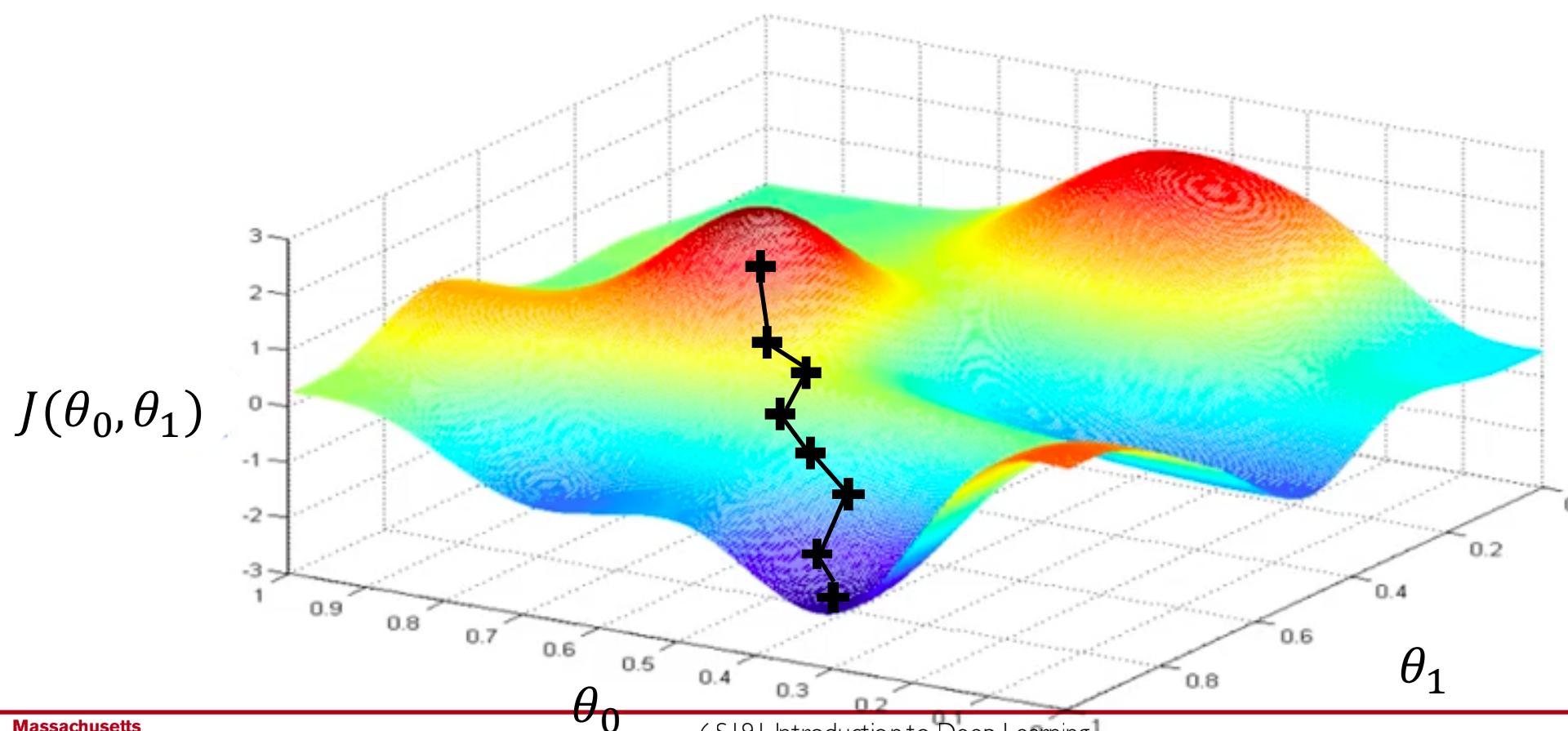
Loss Optimization

Take small step in opposite direction of gradient



Gradient Descent

Repeat until convergence



Gradient Descent

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Compute gradient, $\frac{\partial J(\theta)}{\partial \theta}$
4. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
5. Return weights



Gradient Descent

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$



2. Loop until convergence:

3. Compute gradient, $\frac{\partial J(\theta)}{\partial \theta}$

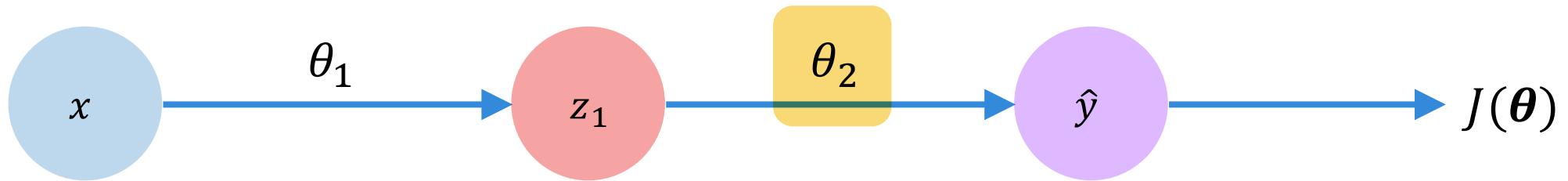


4. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$



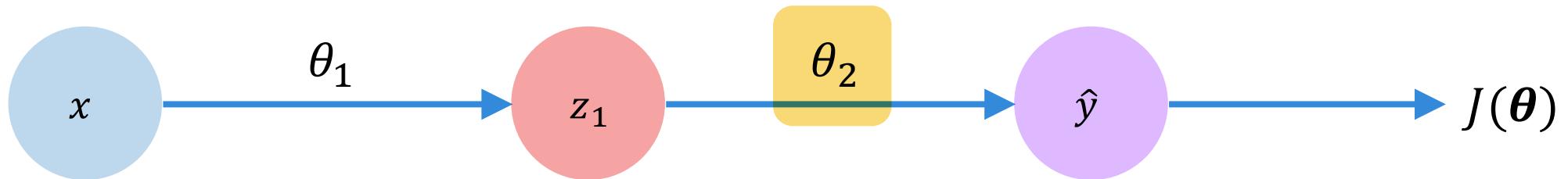
5. Return weights

Computing Gradients: Backpropagation



How does a small change in one weight (ex. θ_2) affect the final loss $J(\boldsymbol{\theta})$?

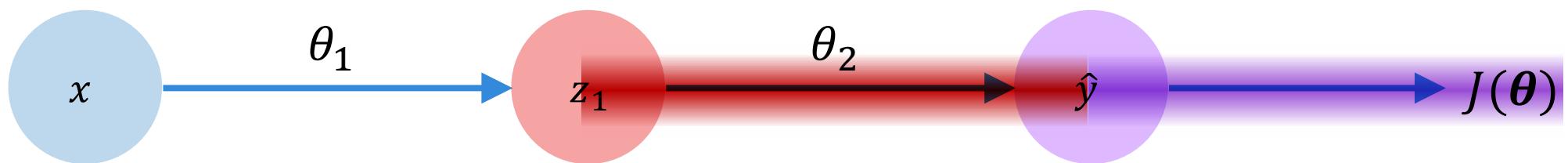
Computing Gradients: Backpropagation



$$\frac{\partial J(\theta)}{\partial \theta_2} =$$

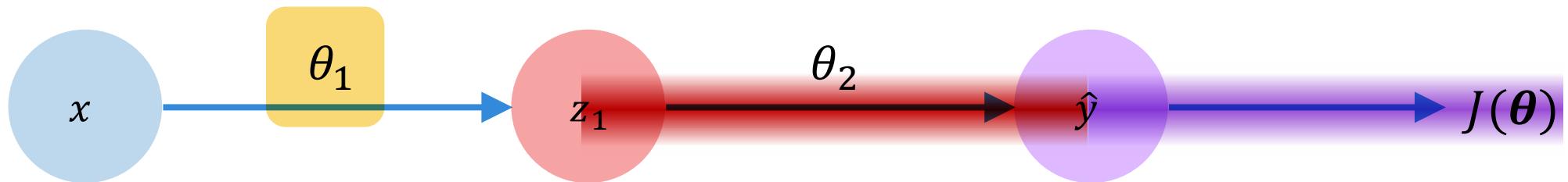
Let's use the chain rule!

Computing Gradients: Backpropagation



$$\frac{\partial J(\theta)}{\partial \theta_2} = \underline{\frac{\partial J(\theta)}{\partial \hat{y}}} * \underline{\frac{\partial \hat{y}}{\partial \theta_2}}$$

Computing Gradients: Backpropagation

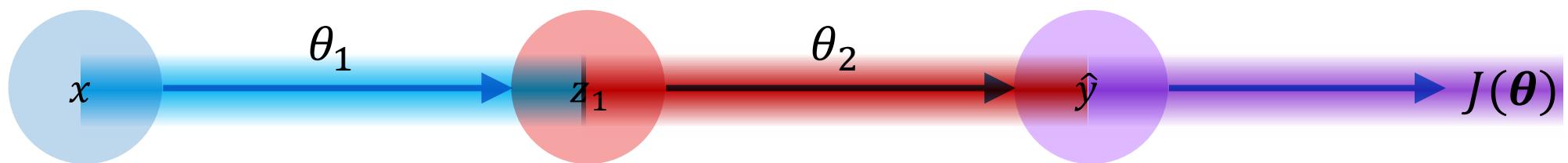


$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{\partial J(\theta)}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial \theta_1}$$

Apply chain rule!

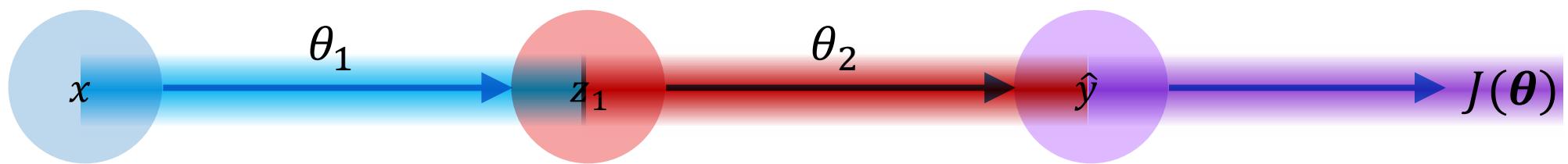
Apply chain rule!

Computing Gradients: Backpropagation



$$\frac{\partial J(\theta)}{\partial \theta_1} = \underline{\frac{\partial J(\theta)}{\partial \hat{y}}} * \underline{\frac{\partial \hat{y}}{\partial z_1}} * \underline{\frac{\partial z_1}{\partial \theta_1}}$$

Computing Gradients: Backpropagation

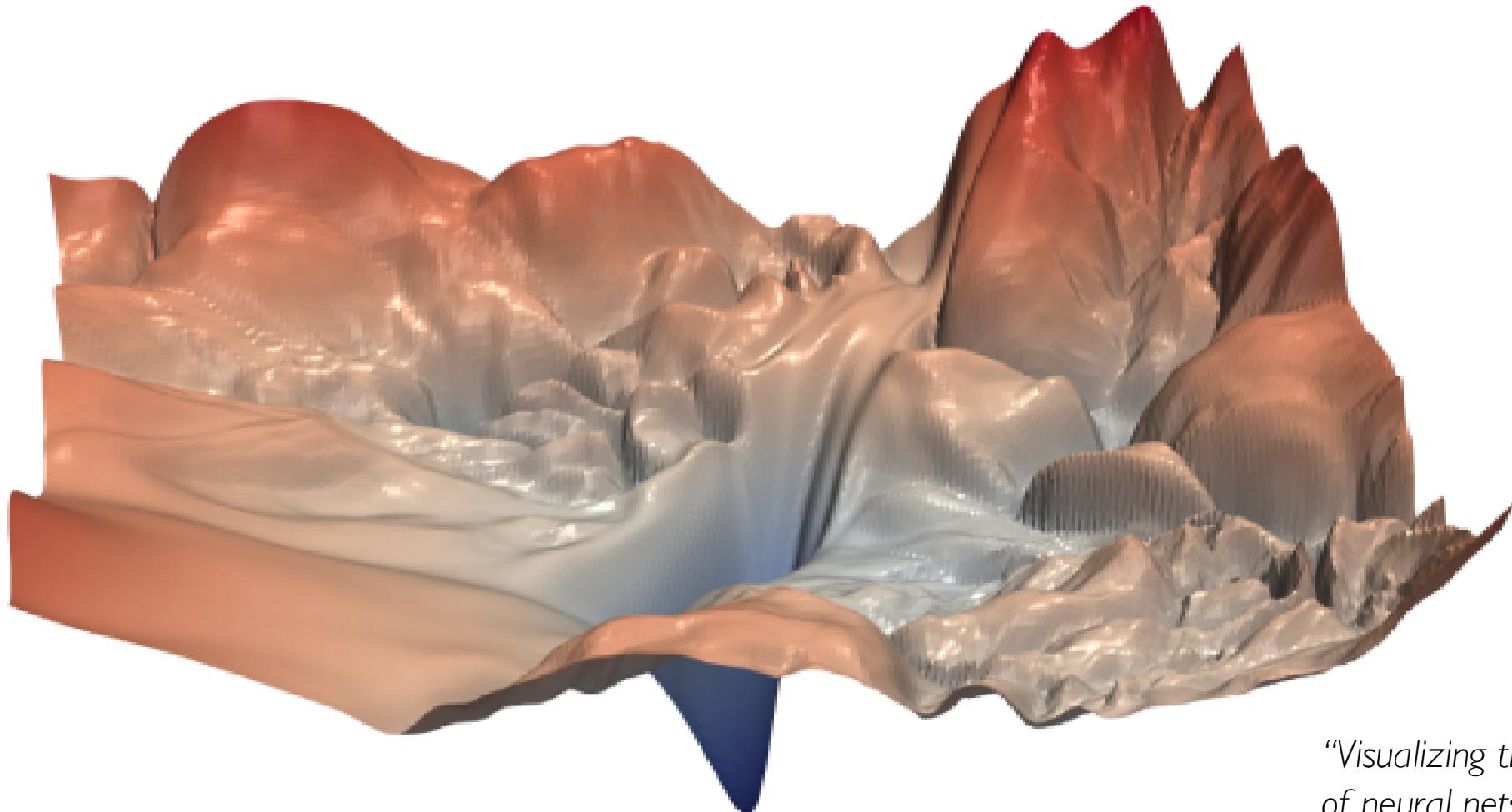


$$\frac{\partial J(\theta)}{\partial \theta_1} = \underbrace{\frac{\partial J(\theta)}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z_1}}_{\text{red bar}} * \underbrace{\frac{\partial z_1}{\partial \theta_1}}_{\text{blue bar}}$$

Repeat this for **every weight in the network** using gradients from later layers



Training Neural Networks is Difficult



*"Visualizing the loss landscape
of neural nets". Dec 2017.*

Loss Functions Can Be Difficult to Optimize

Remember:

Optimization through gradient descent

$$\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$$

Loss Functions Can Be Difficult to Optimize

Remember:

Optimization through gradient descent

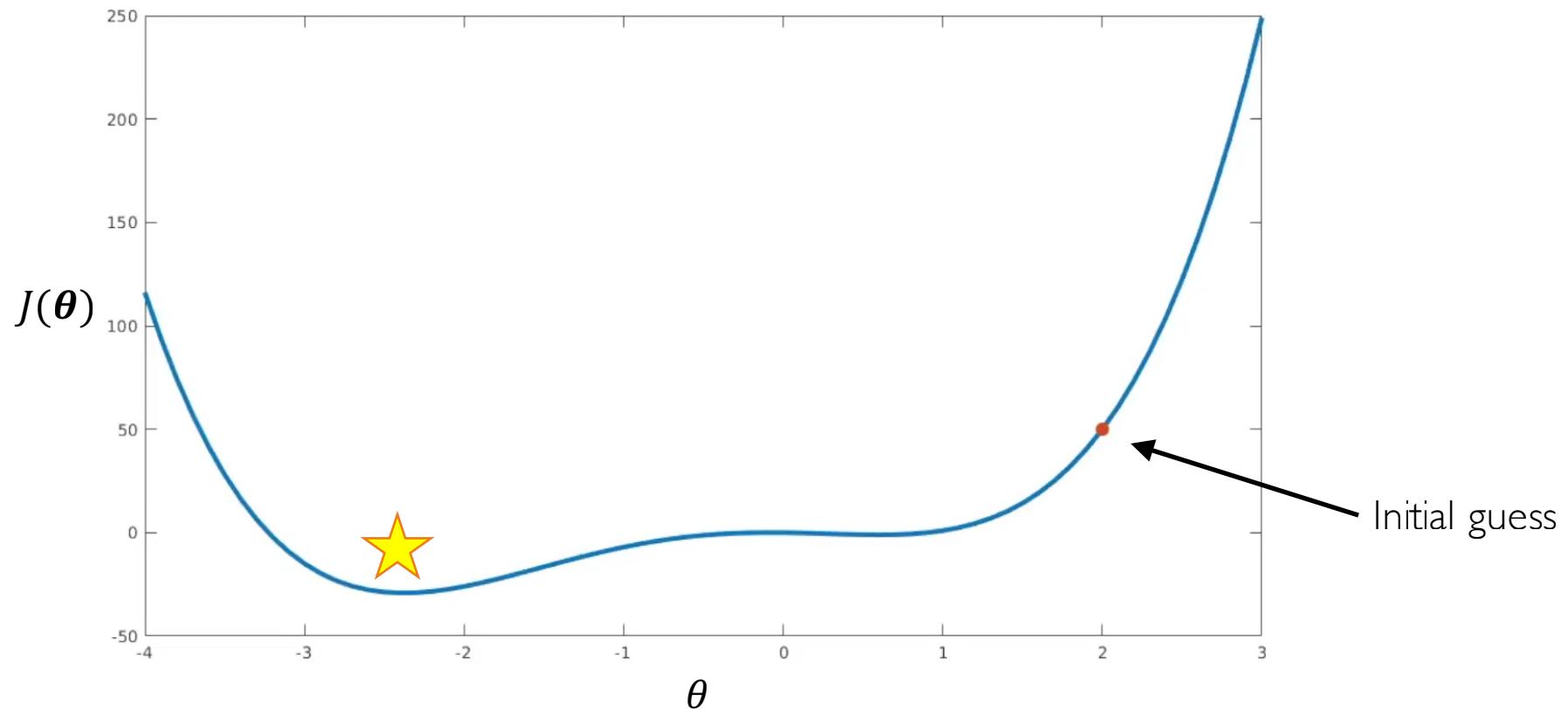
$$\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$$



How can we set the
learning rate?

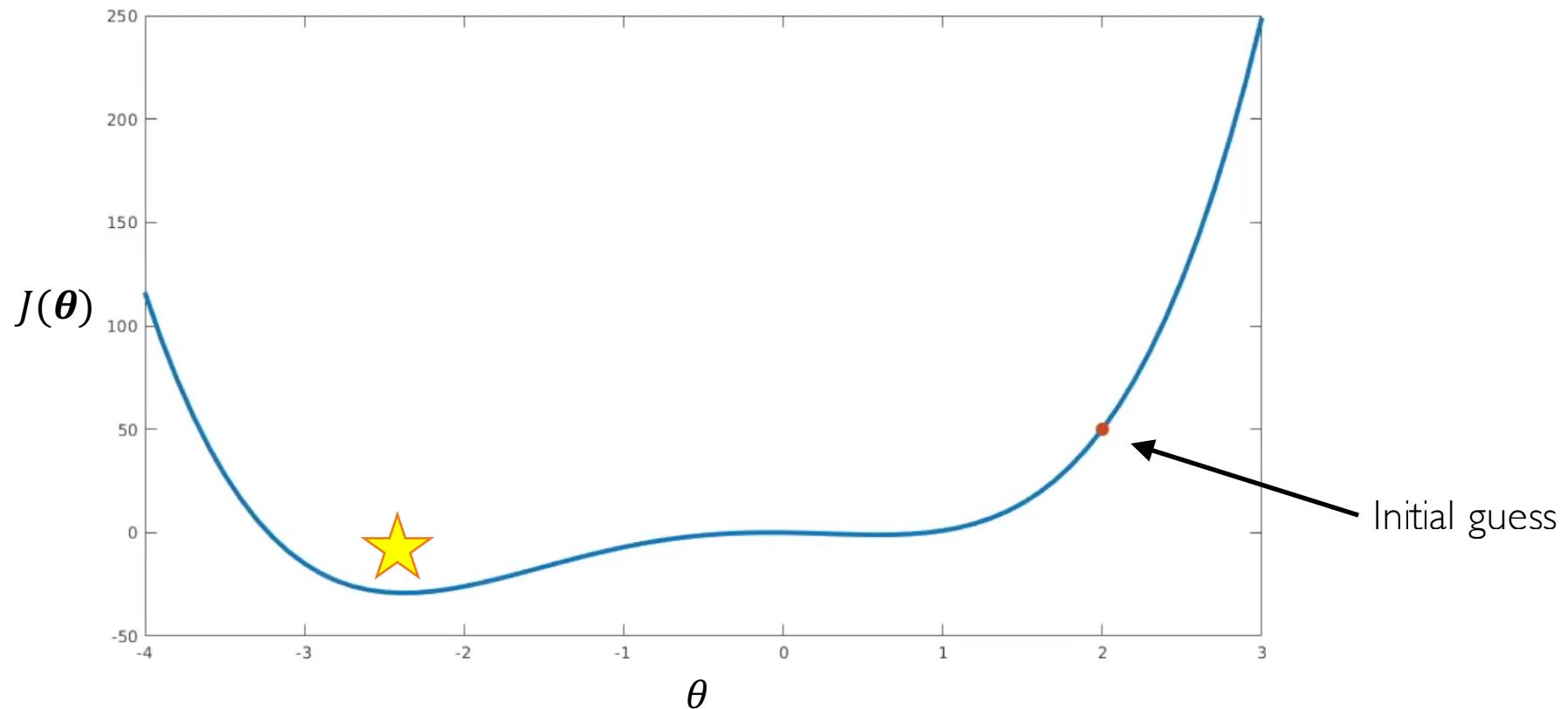
Setting the Learning Rate

Small learning rate converges slowly and gets stuck in false local minima



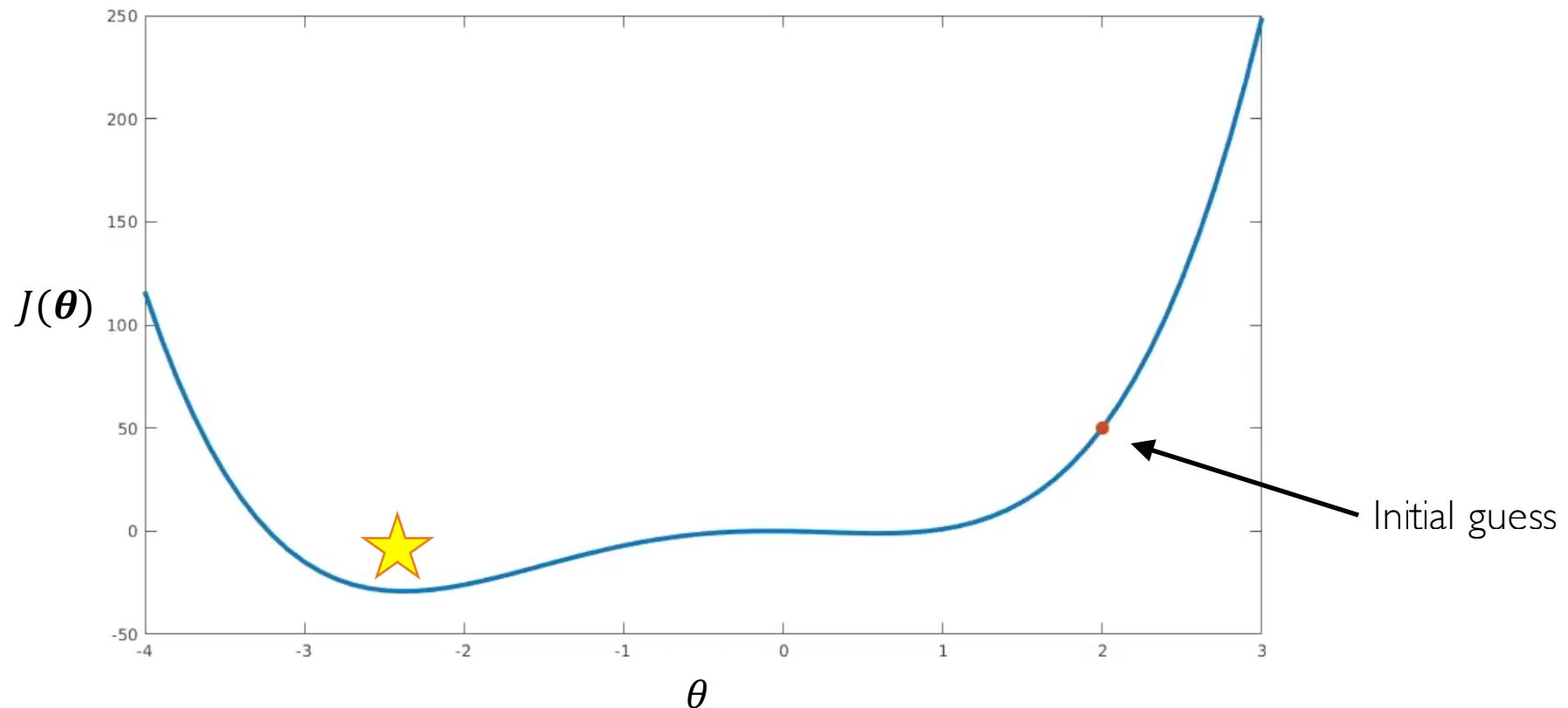
Setting the Learning Rate

Large learning rates overshoot, become unstable and diverge



Setting the Learning Rate

Stable learning rates converge smoothly and avoid local minima



How to deal with this?

Idea I:

Try lots of different learning rates and see what works “just right”

How to deal with this?

Idea 1:

Try lots of different learning rates and see what works “just right”

Idea 2:

Do something smarter!

Design an adaptive learning rate that “adapts” to the landscape

Adaptive Learning Rates

- Learning rates are no longer fixed
- Can be made larger or smaller depending on:
 - how large gradient is
 - how fast learning is happening
 - size of particular weights
 - etc...

Adaptive Learning Rate Algorithms

- Momentum
- Adagrad
- Adadelta
- Adam
- RMSProp



Qian et al. "On the momentum term in gradient descent learning algorithms." 1999.

Duchi et al. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." 2011.

Zeiler et al. "ADADELTA: An Adaptive Learning Rate Method." 2012.

Kingma et al. "Adam: A Method for Stochastic Optimization." 2014.

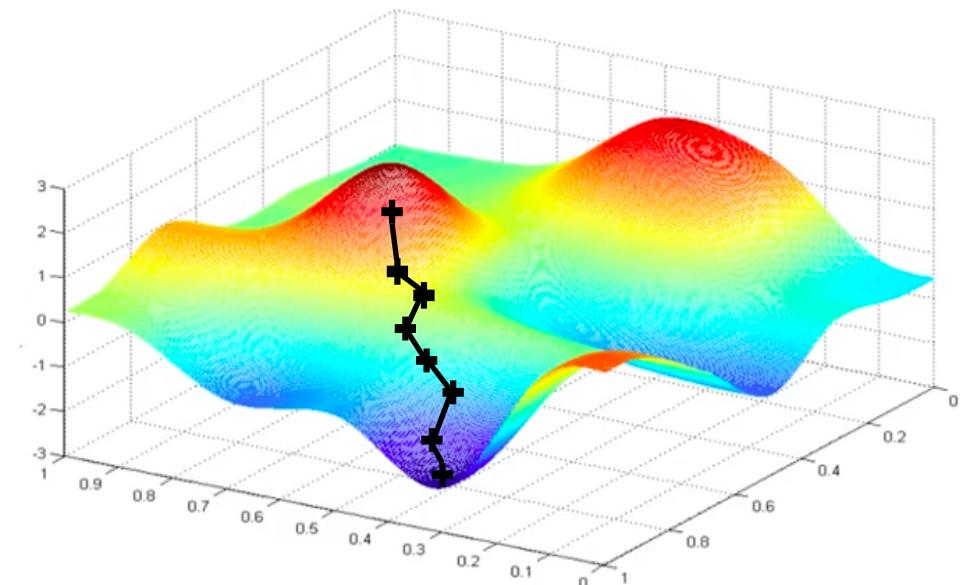
Additional details: <http://ruder.io/optimizing-gradient-descent/>



Gradient Descent

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Compute gradient, $\frac{\partial J(\theta)}{\partial \theta}$
4. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
5. Return weights

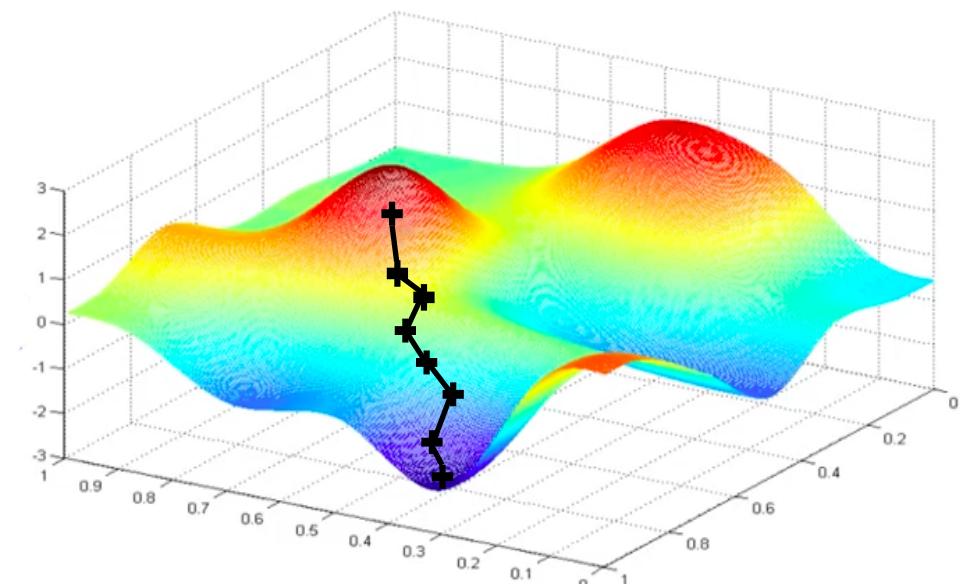


Gradient Descent

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Compute gradient, $\frac{\partial J(\theta)}{\partial \theta}$
4. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
5. Return weights

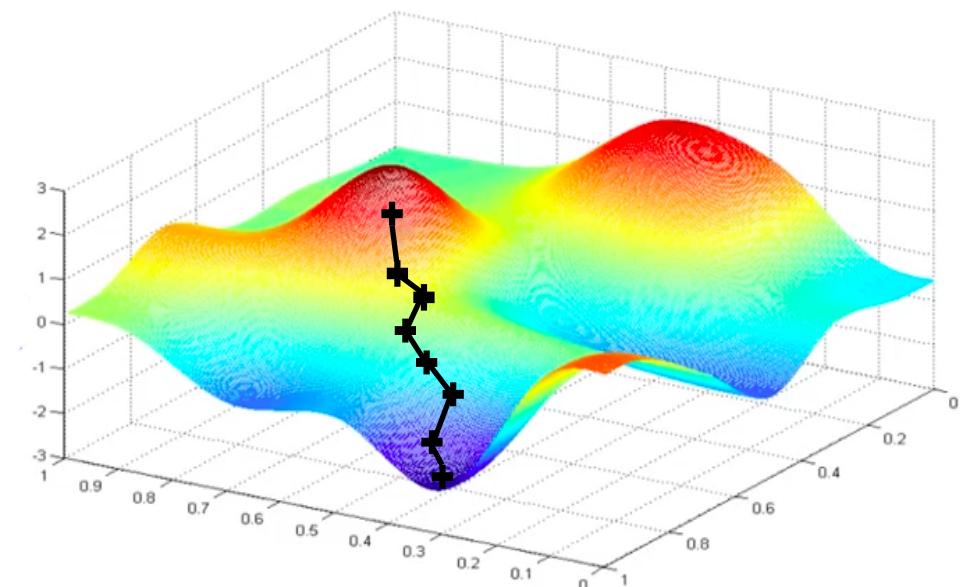
Can be very computational to compute!



Stochastic Gradient Descent

Algorithm

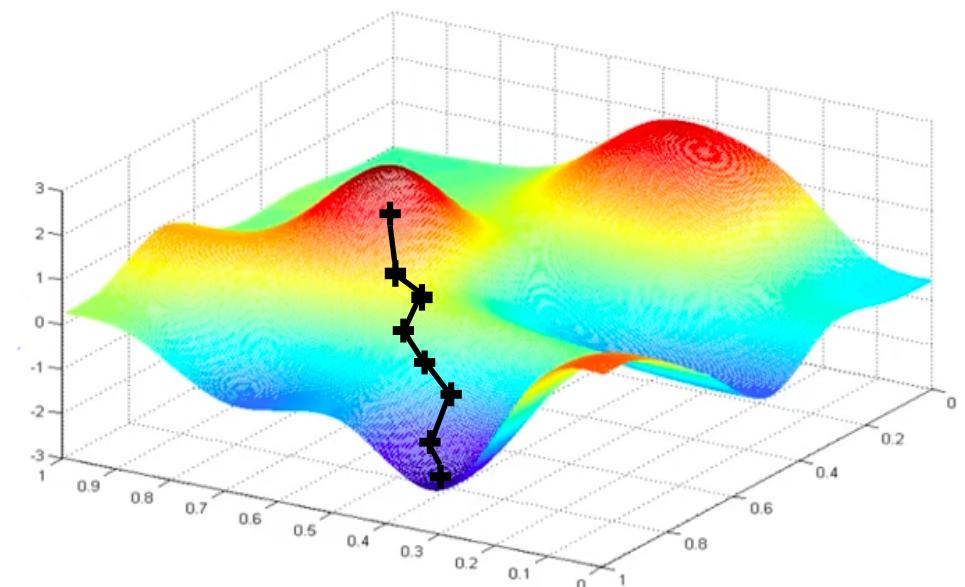
1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Pick single data point i
4. Compute gradient, $\frac{\partial J_i(\theta)}{\partial \theta}$
5. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
6. Return weights



Stochastic Gradient Descent

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Pick single data point i
4. Compute gradient, $\frac{\partial J_i(\theta)}{\partial \theta}$
5. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
6. Return weights

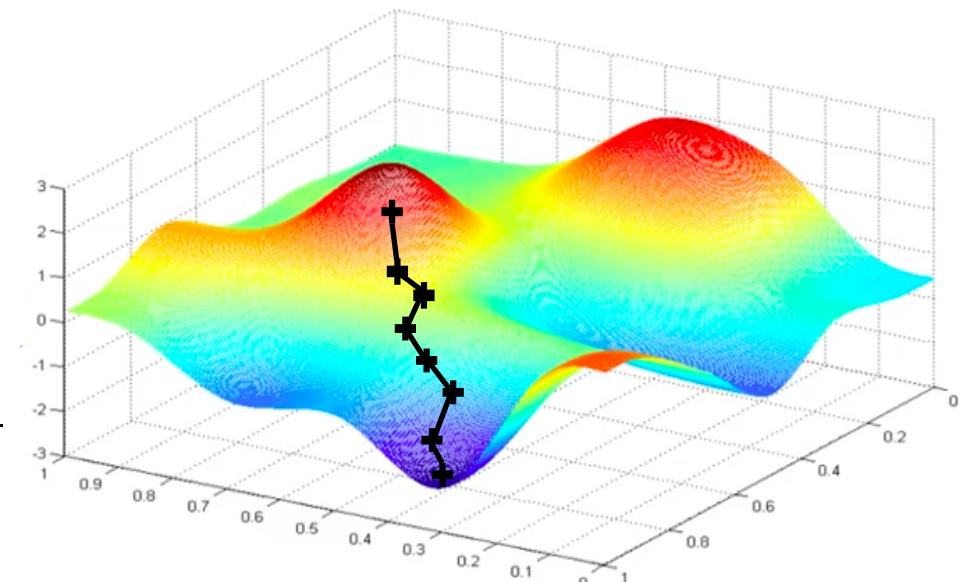


Easy to compute but
very noisy
(stochastic)!

Stochastic Gradient Descent

Algorithm

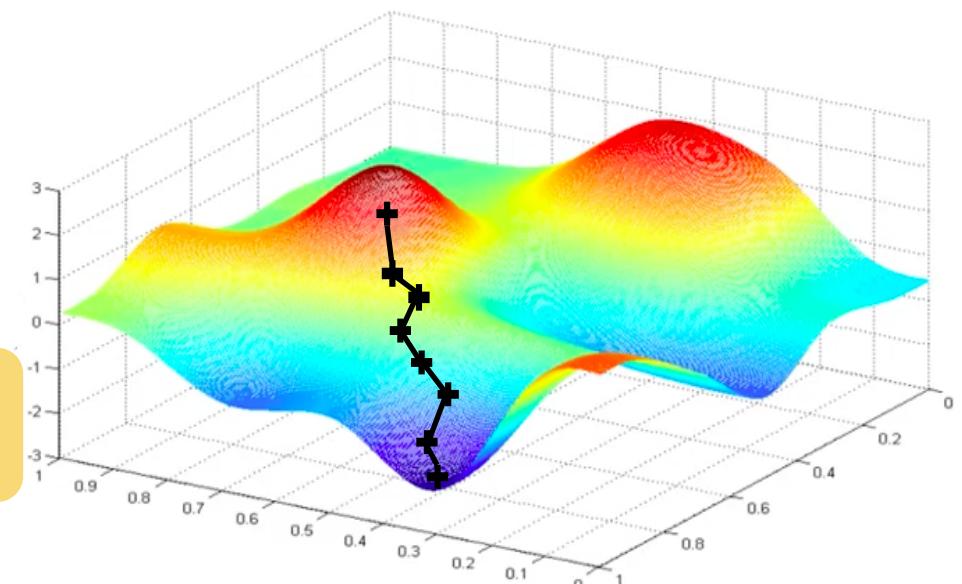
1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Pick batch of B data points
4. Compute gradient, $\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{B} \sum_{k=1}^B \frac{\partial J_k(\theta)}{\partial \theta}$
5. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
6. Return weights



Stochastic Gradient Descent

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Pick batch of B data points
4. Compute gradient,
$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{B} \sum_{k=1}^B \frac{\partial J_k(\theta)}{\partial \theta}$$
5. Update weights, $\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$
6. Return weights



Fast to compute and a much better
estimate of the true gradient!

Mini-batches while training

More accurate estimation of gradient

Smoother convergence

Allows for larger learning rates

Mini-batches while training

More accurate estimation of gradient

Smoother convergence

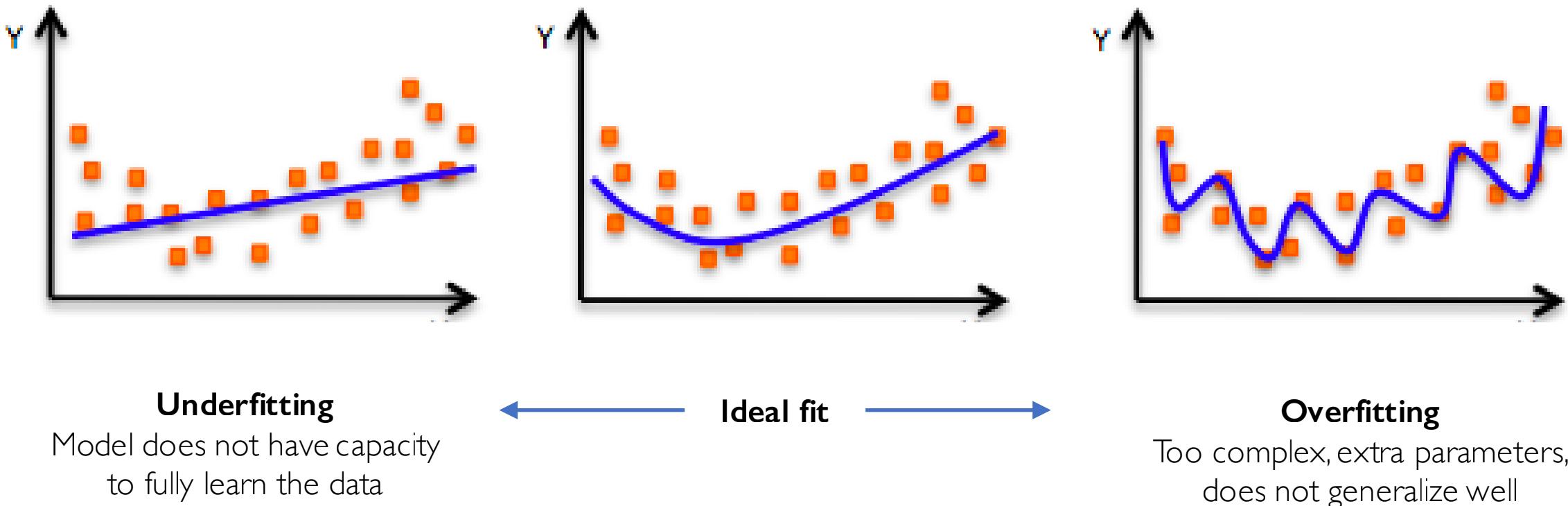
Allows for larger learning rates

Mini-batches lead to fast training!

Can parallelize computation + achieve significant speed increases on GPU's



The Problem of Overfitting



Regularization

What is it?

Technique that constrains our optimization problem to discourage complex models

Regularization

What is it?

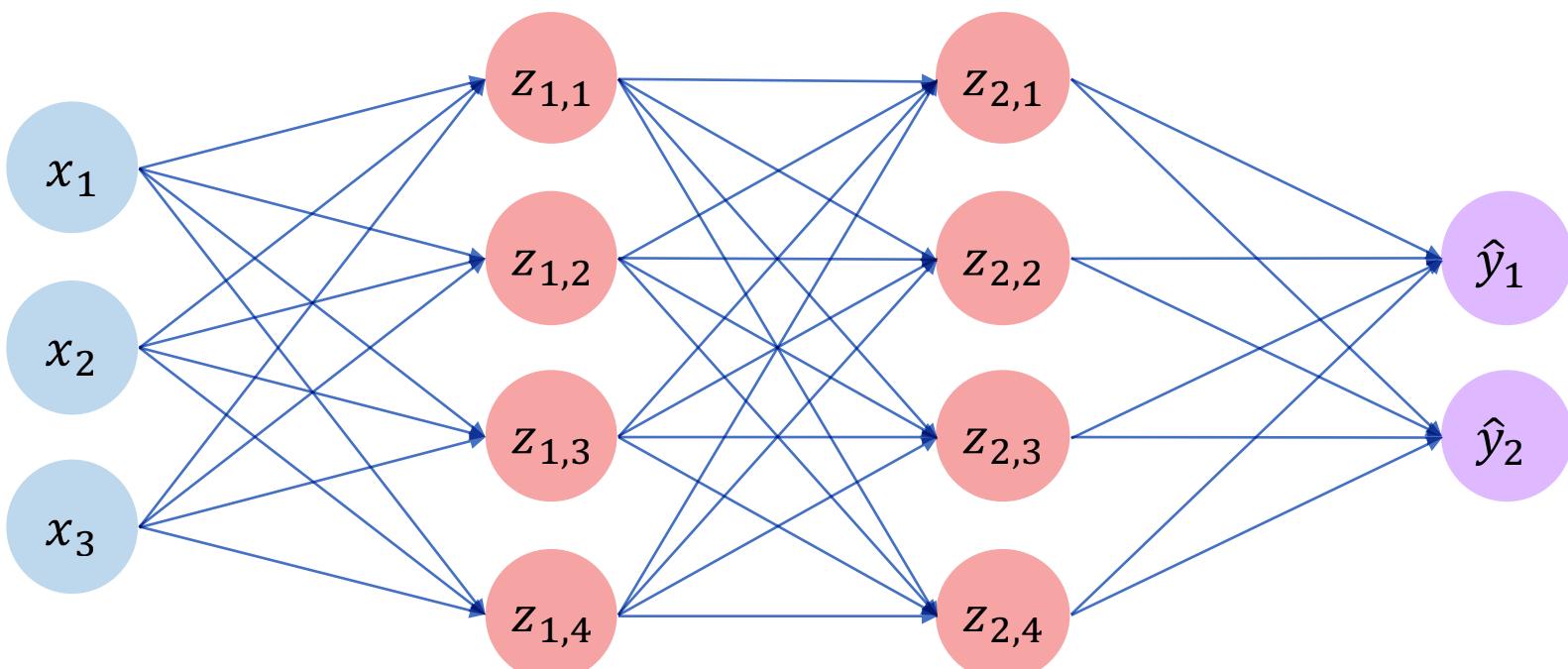
Technique that constrains our optimization problem to discourage complex models

Why do we need it?

Improve generalization of our model on unseen data

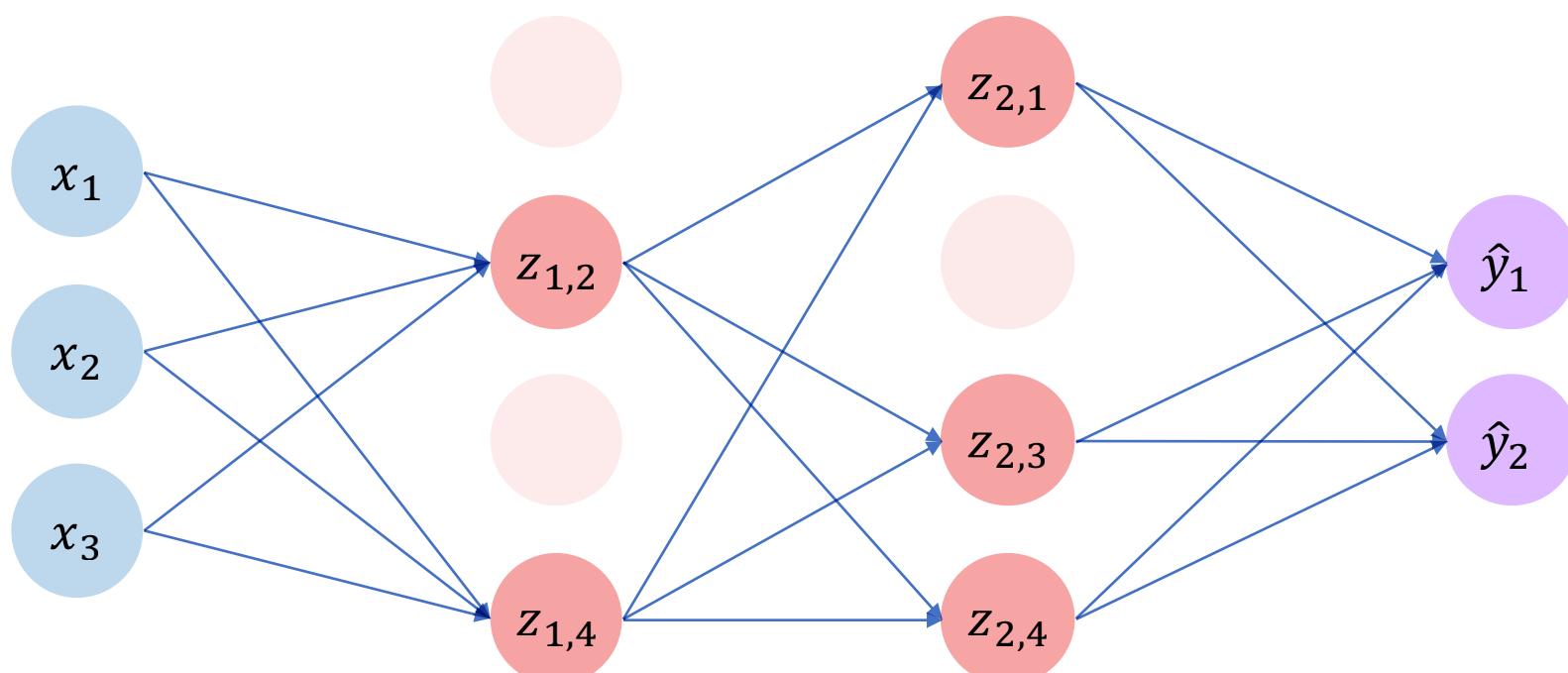
Regularization I: Dropout

- During training, randomly set some activations to 0



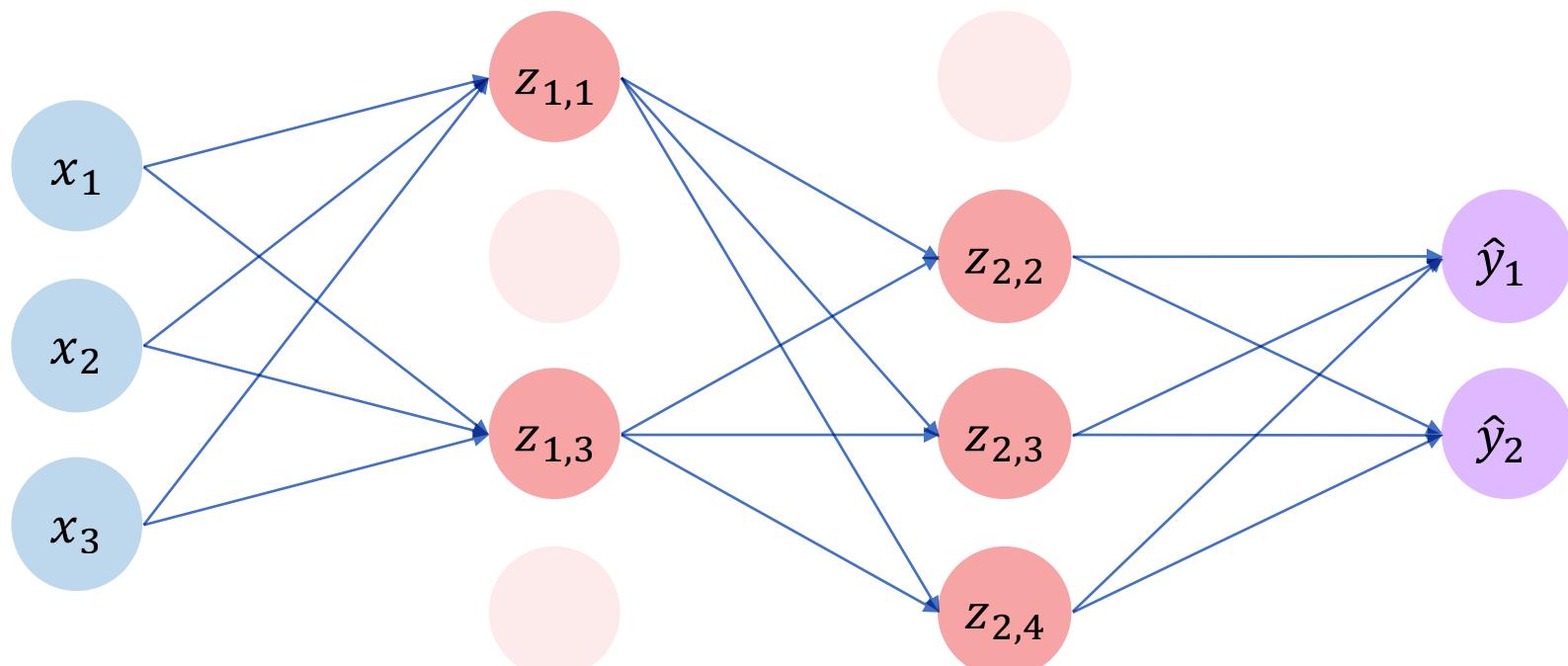
Regularization I: Dropout

- During training, randomly set some activations to 0
 - Typically ‘drop’ 50% of activations in layer
 - Forces network to not rely on any 1 node



Regularization I: Dropout

- During training, randomly set some activations to 0
 - Typically ‘drop’ 50% of activations in layer
 - Forces network to not rely on any 1 node



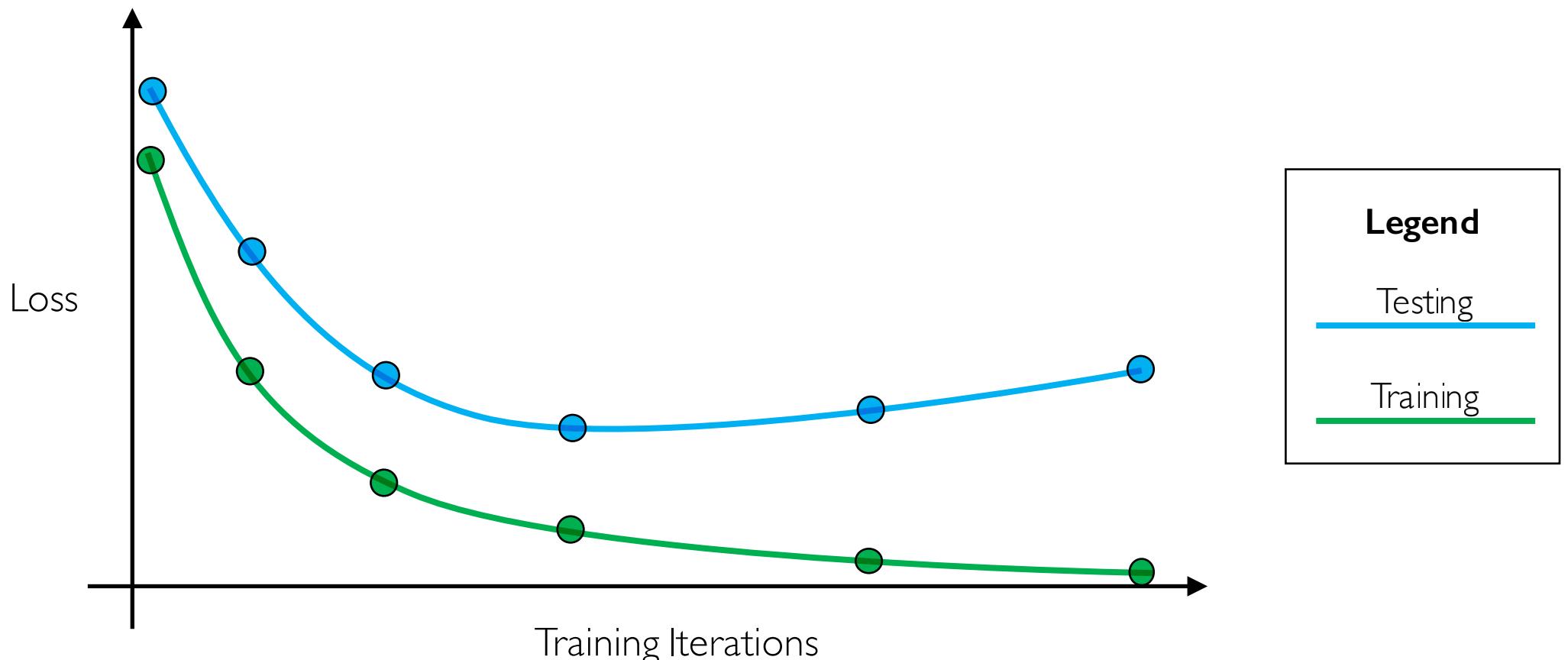
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



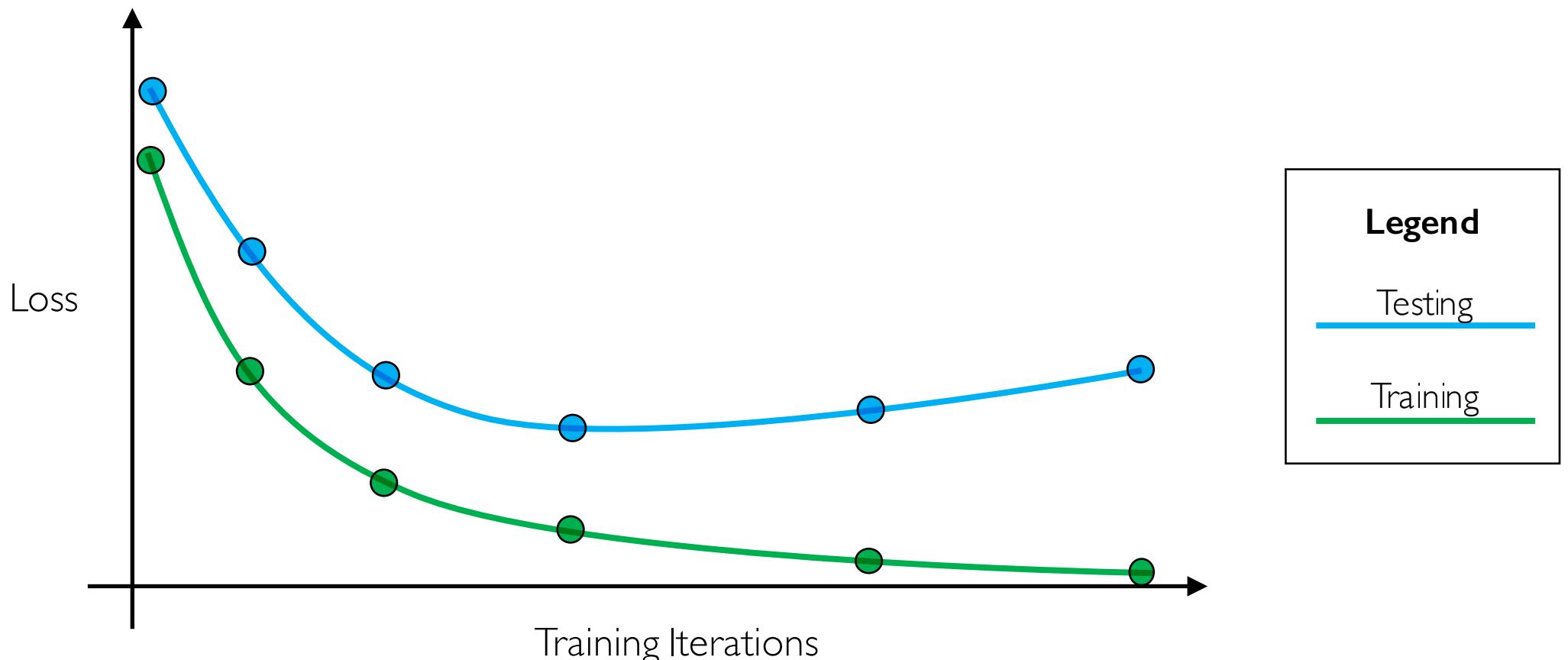
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



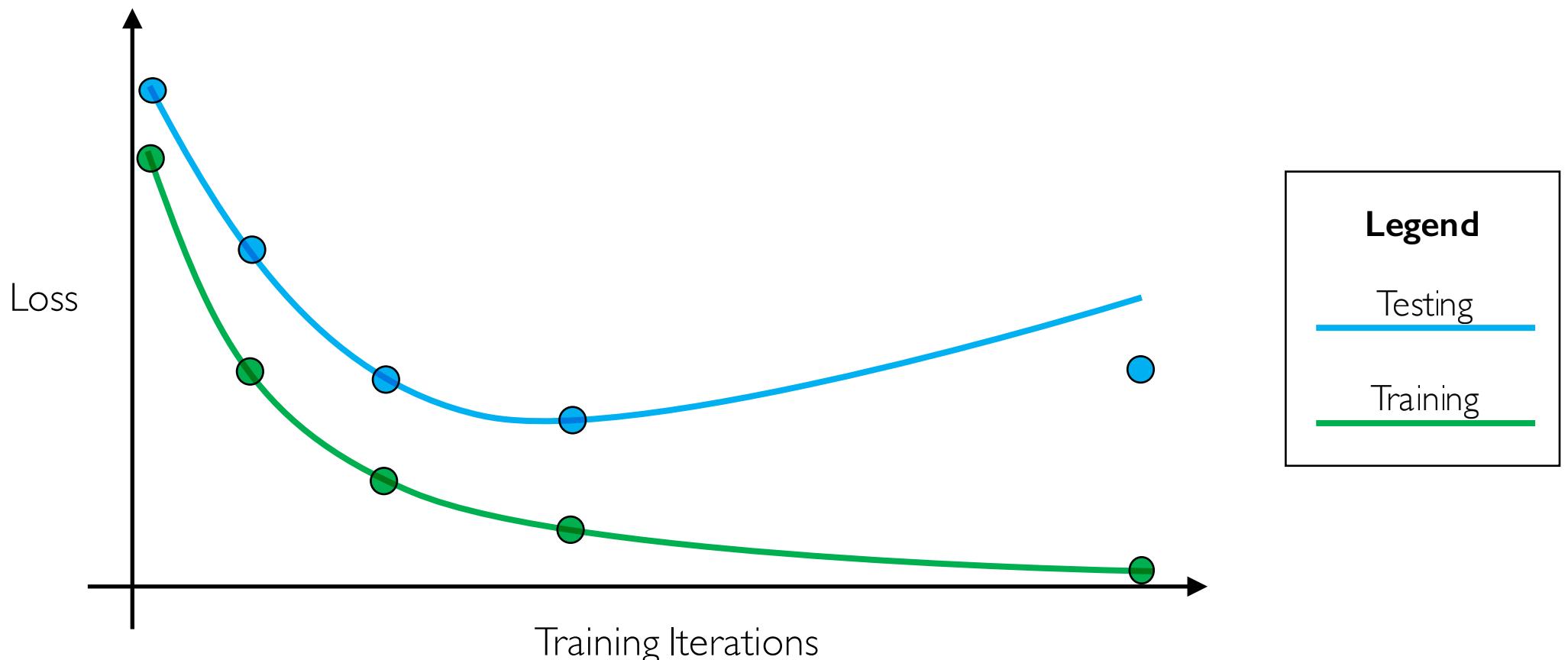
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



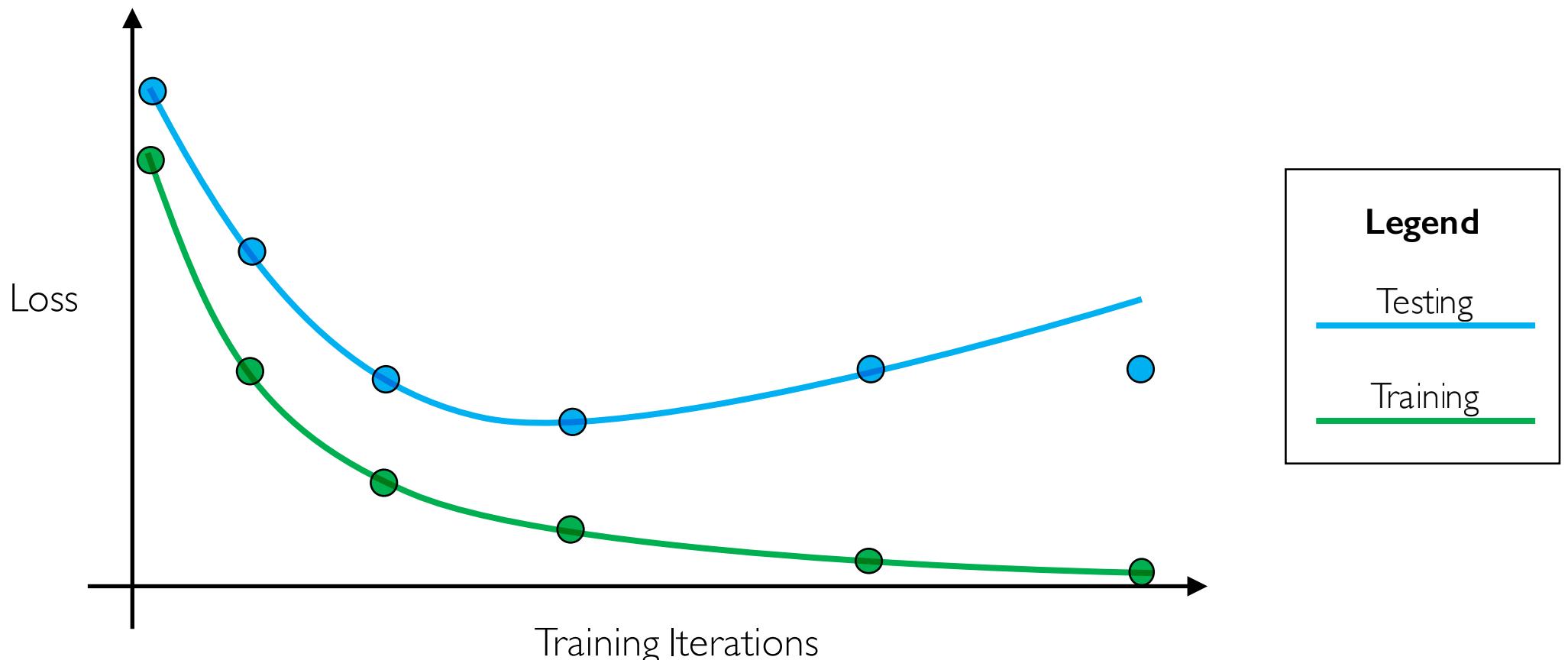
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



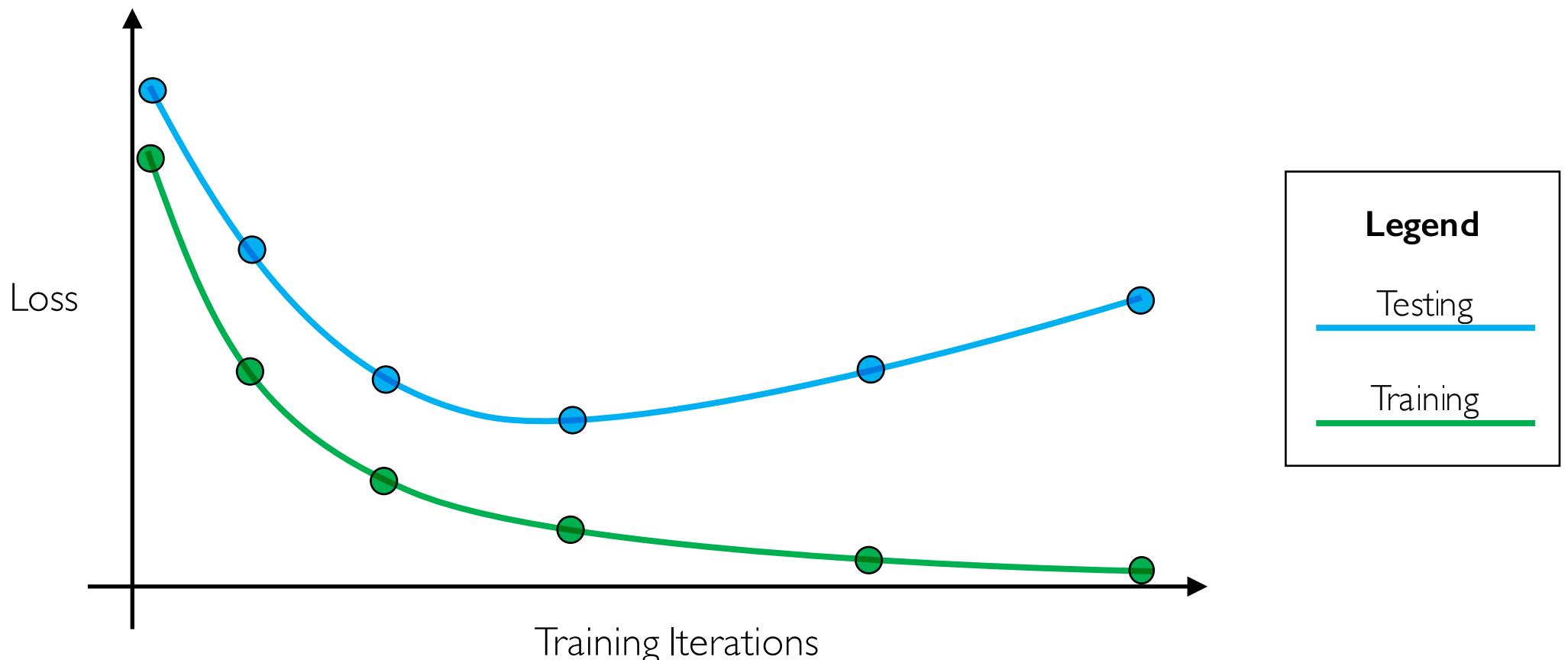
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



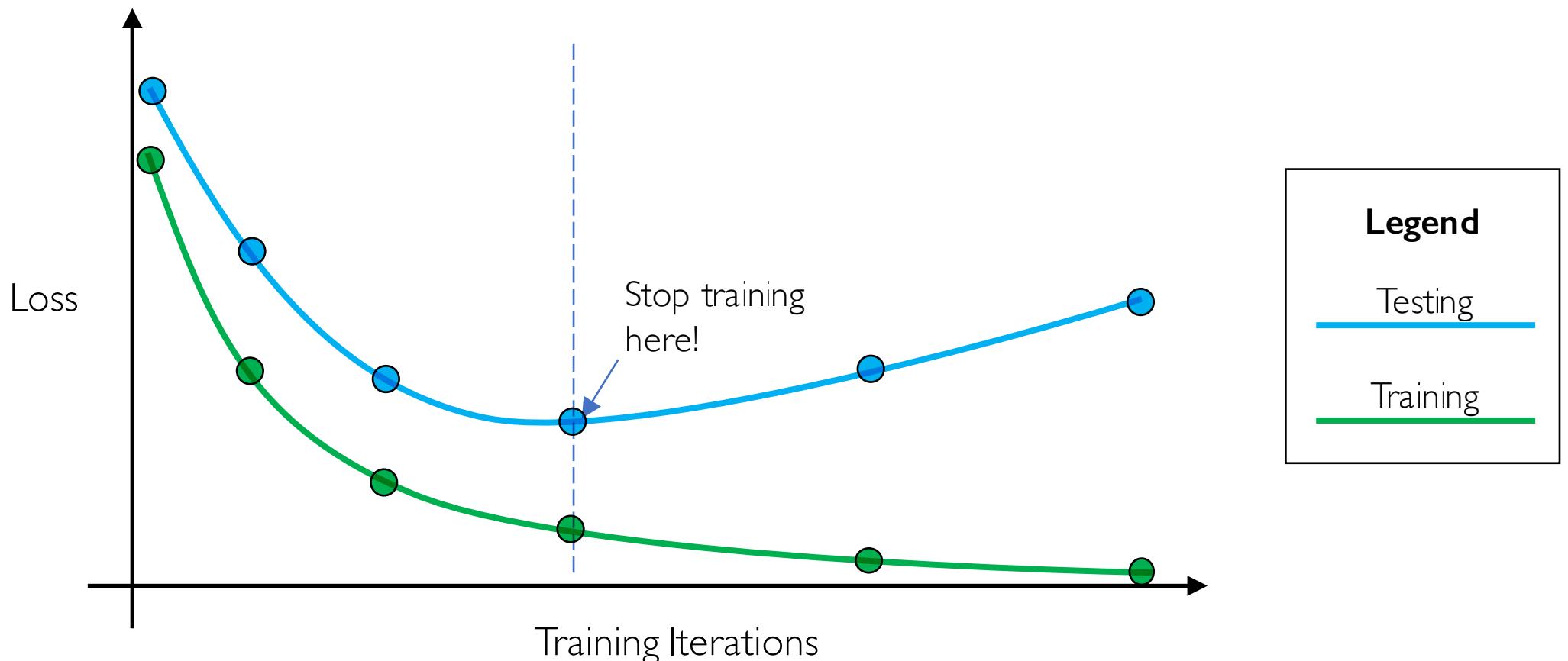
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



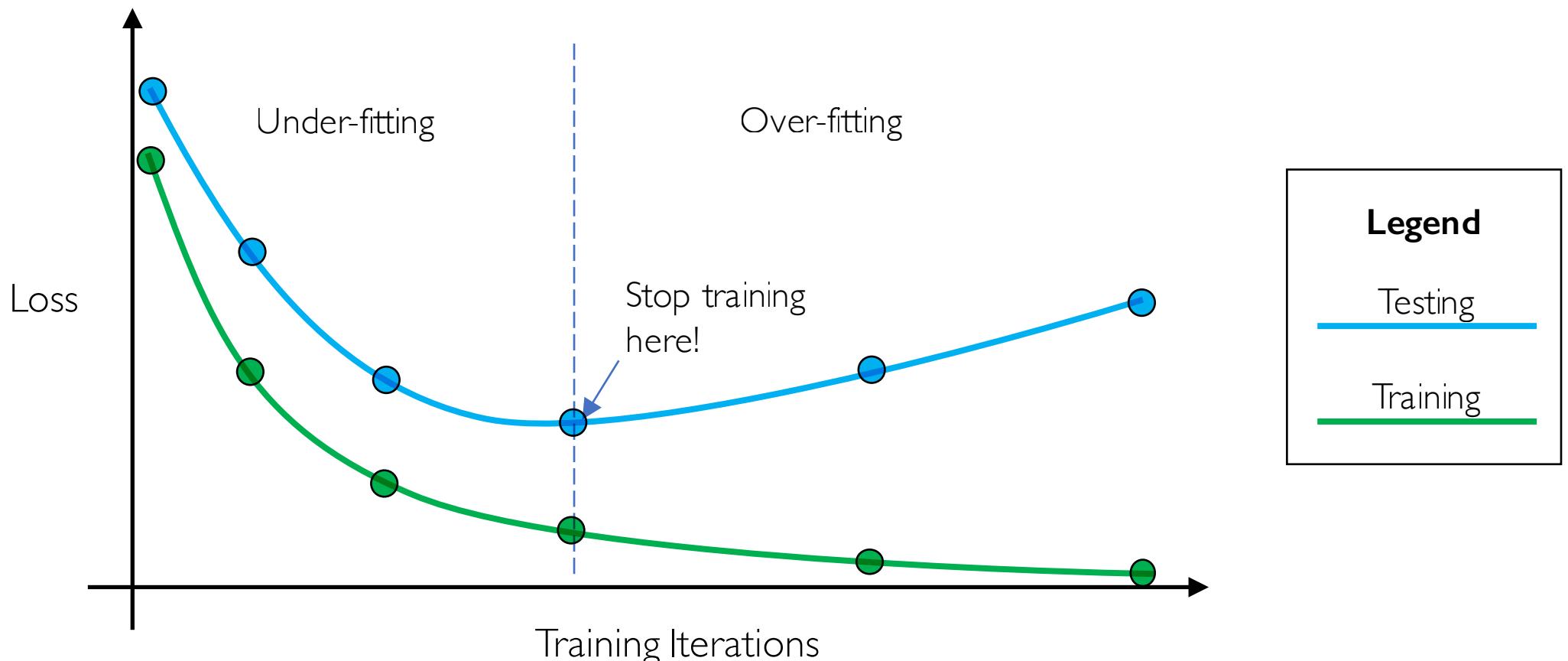
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit



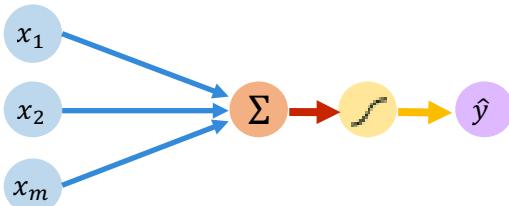
Regularization 2: Early Stopping

- Stop training before we have a chance to overfit

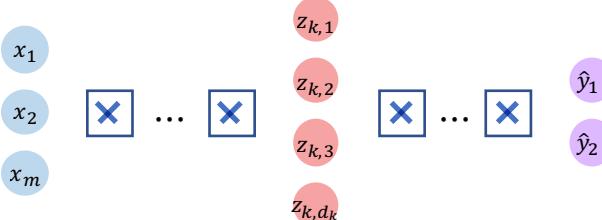


Core Foundation Review

- Structural building blocks
- Nonlinear activation functions



- Stacking Perceptrons to form neural networks
- Optimization through backpropagation



- Adaptive learning
- Batching
- Regularization

