Deep Learning (1470)

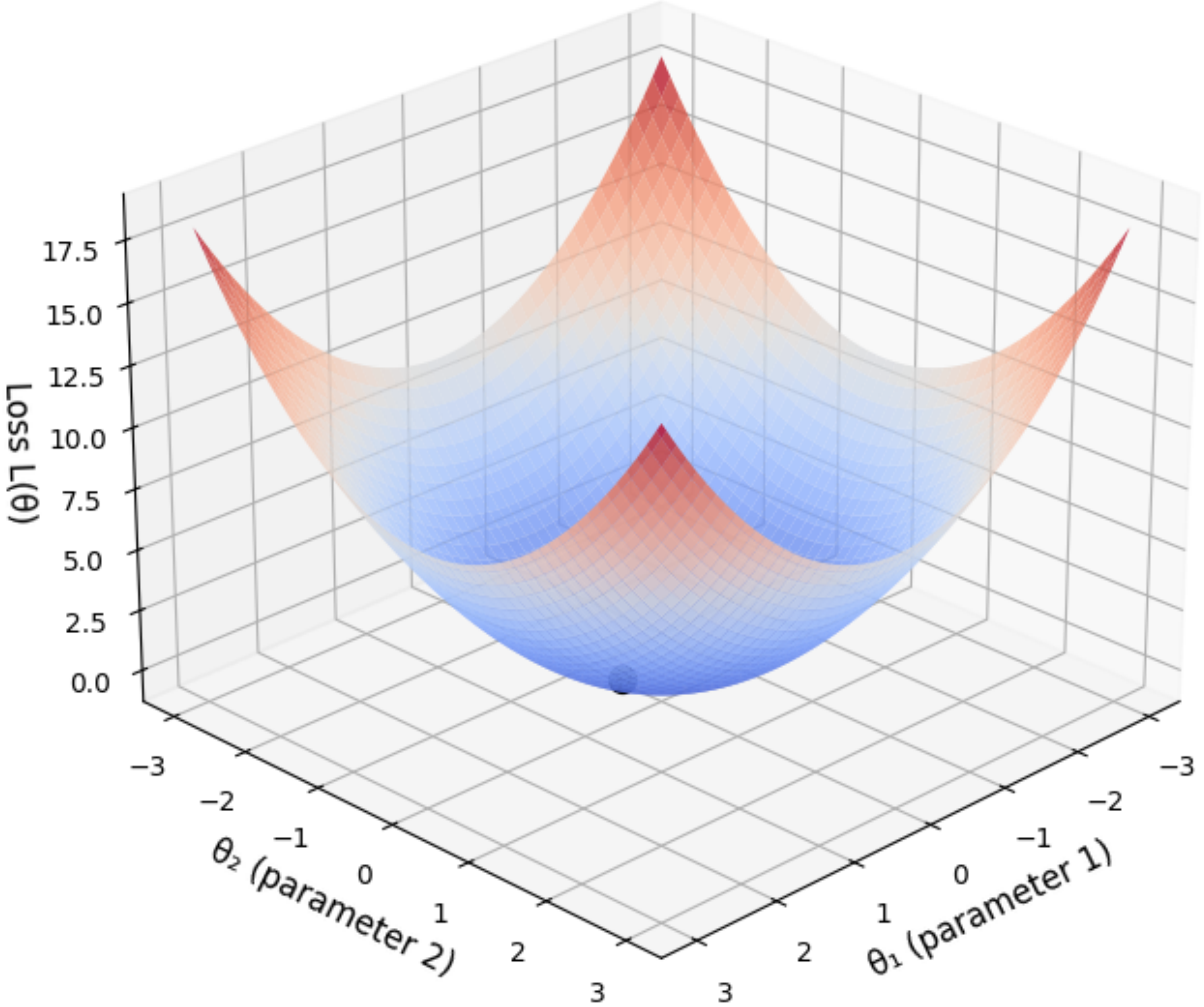
Randall Balestriero

Class 4

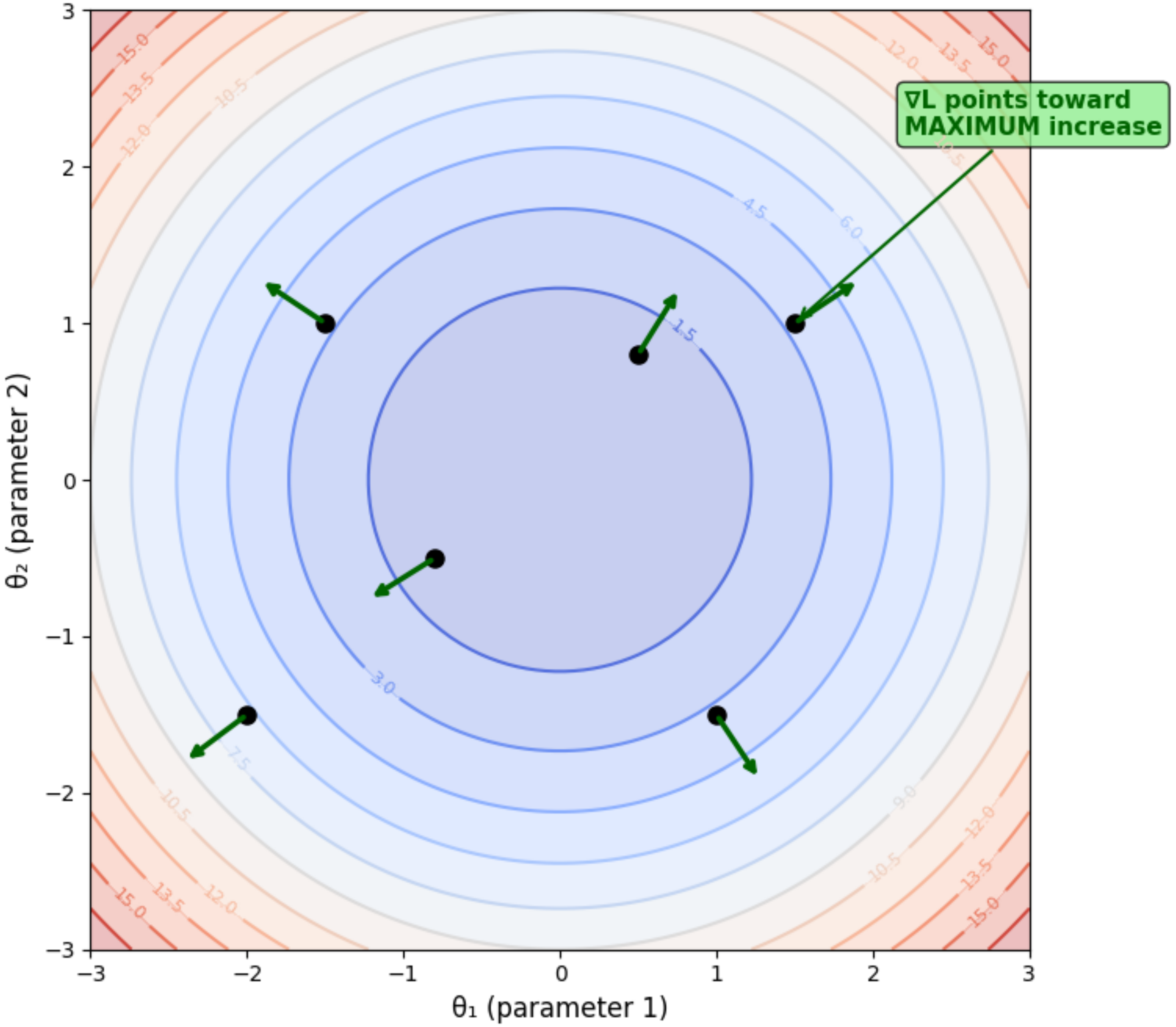
Rewind!

Gradient: Intuition

3D Loss Surface

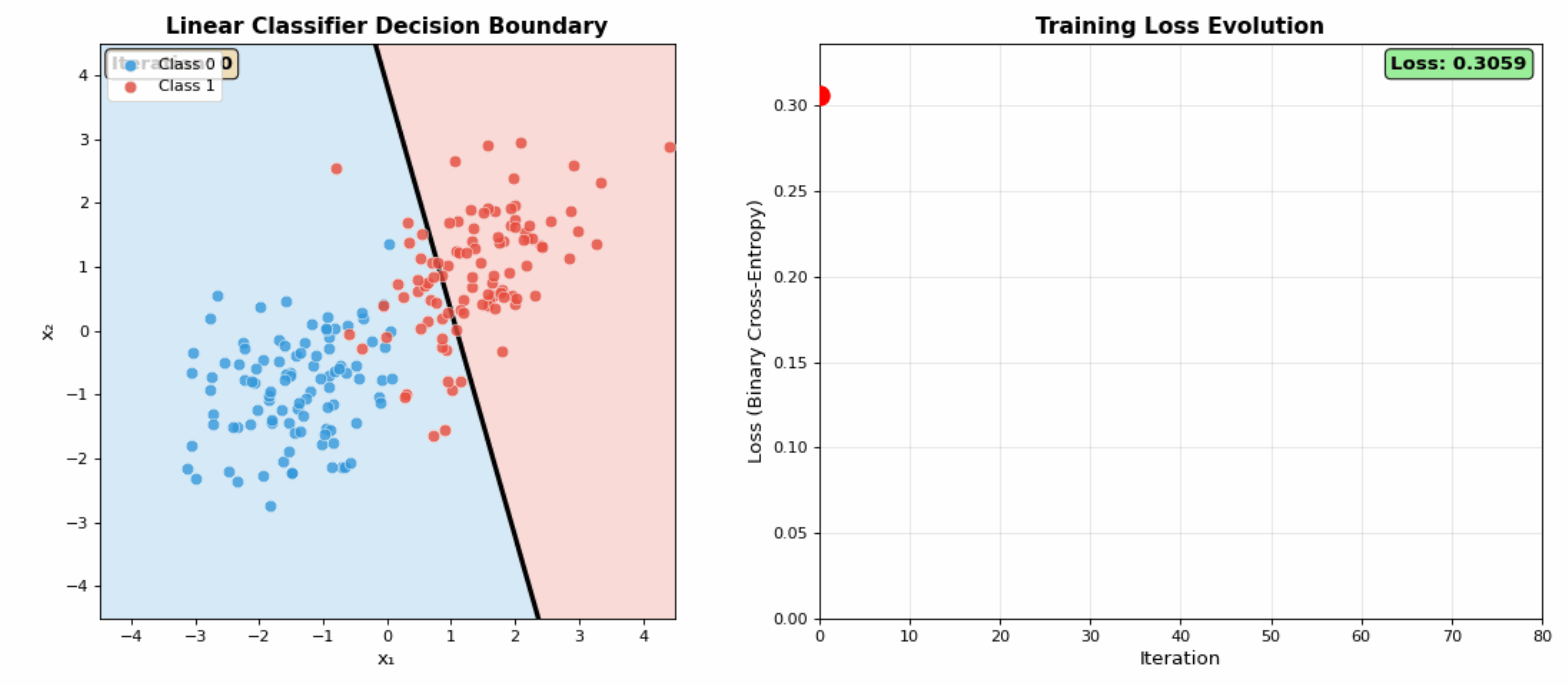


Gradient Field on Contour Plot

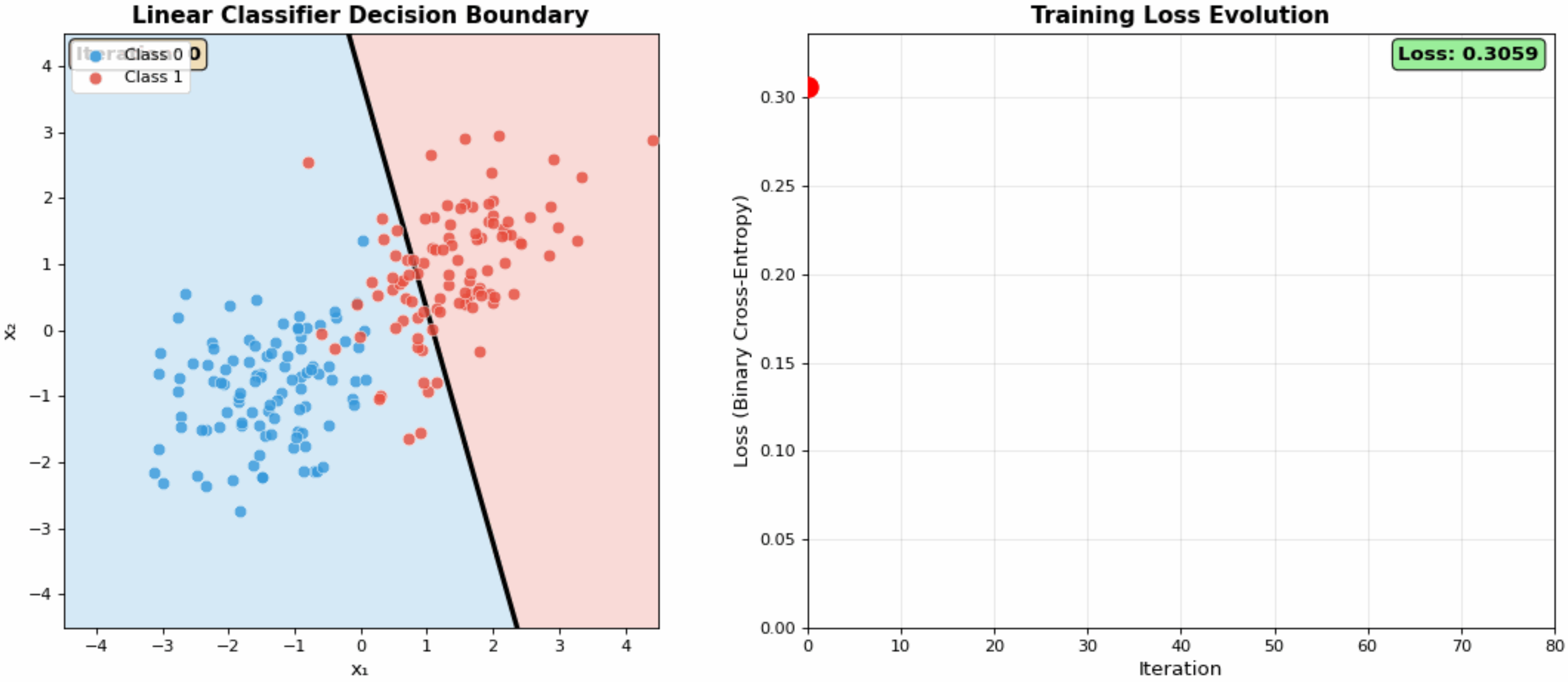


∇L points toward MAXIMUM increase

Gradient: Intuition



Gradient: Intuition



MNIST

The most famous dataset in Deep Learning

Modified **N**ational **I**nstitute of **S**tandards and **T**echnology database

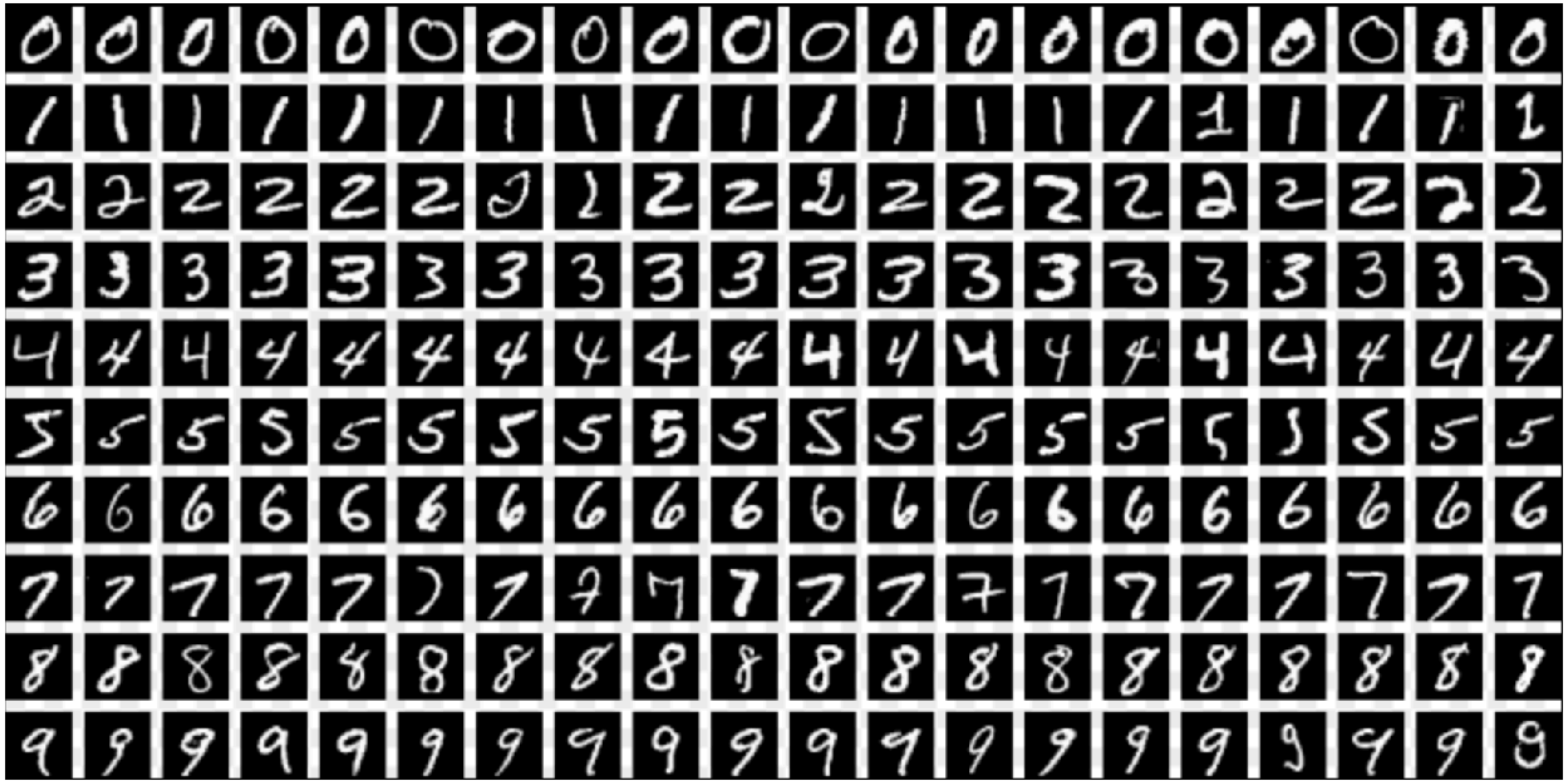


Image courtesy of Wikipedia

MNIST

MNIST

- What is D (dimension) of the samples?

MNIST

- What is D (dimension) of the samples?
- How many classes do we have?

MNIST

- What is D (dimension) of the samples?
- How many classes do we have?
- Do you think the classes are linearly separable?

MNIST

- What is D (dimension) of the samples?
- How many classes do we have?
- Do you think the classes are linearly separable?

A linear model reaches 94% accuracy!

Your First Deep Network!

Your First Deep Network!

- Why do we want to use a Deep Network?

Your First Deep Network!

- Why do we want to use a Deep Network?
- Could you think of “features” to extract by-hand that would improve linear model accuracy?

Your First Deep Network!

- Why do we want to use a Deep Network?
- Could you think of “features” to extract by-hand that would improve linear model accuracy?
- Which accuracy do you think we can reach?

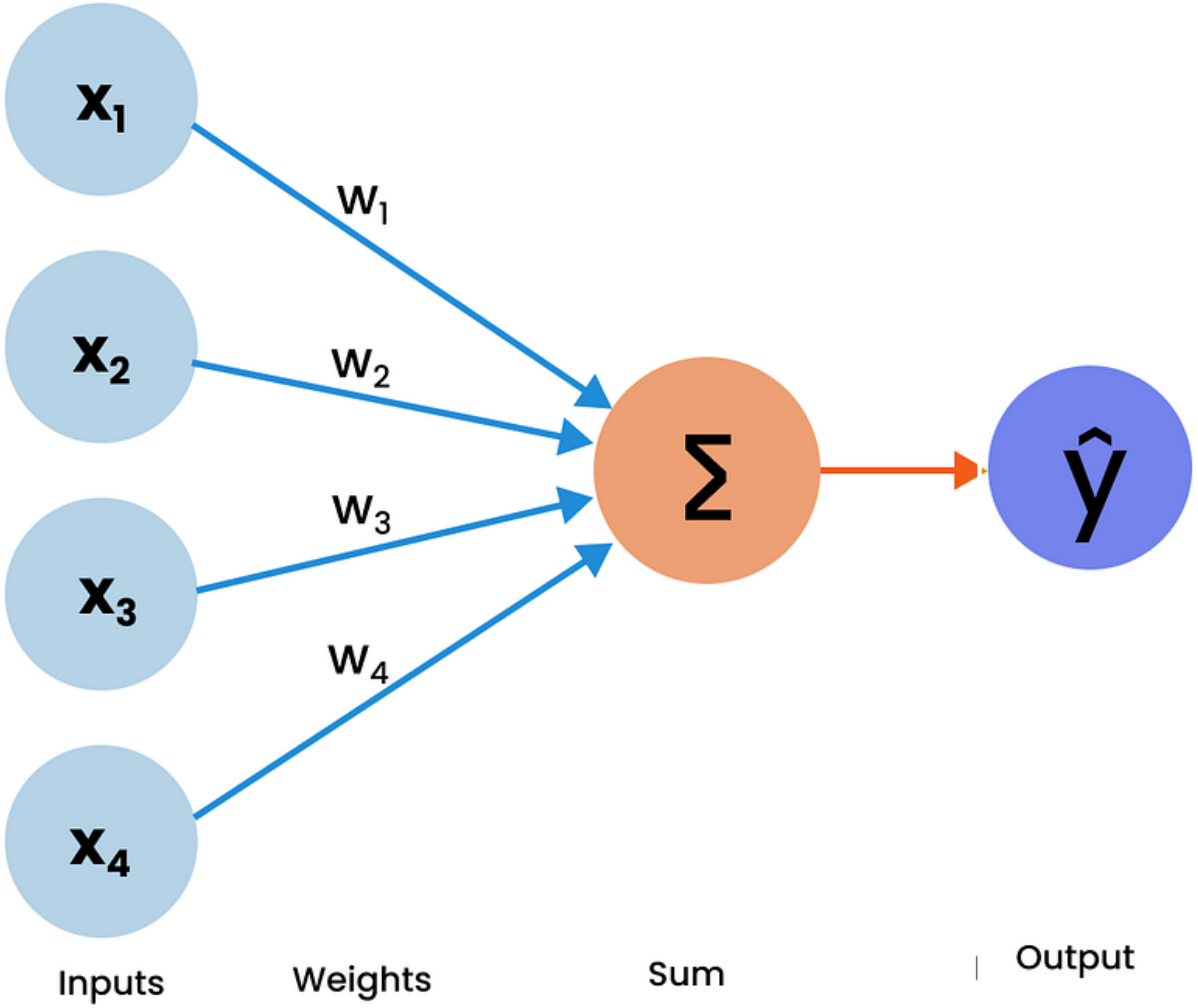
Your First Deep Network!

- Why do we want to use a Deep Network?
- Could you think of “features” to extract by-hand that would improve linear model accuracy?
- Which accuracy do you think we can reach?

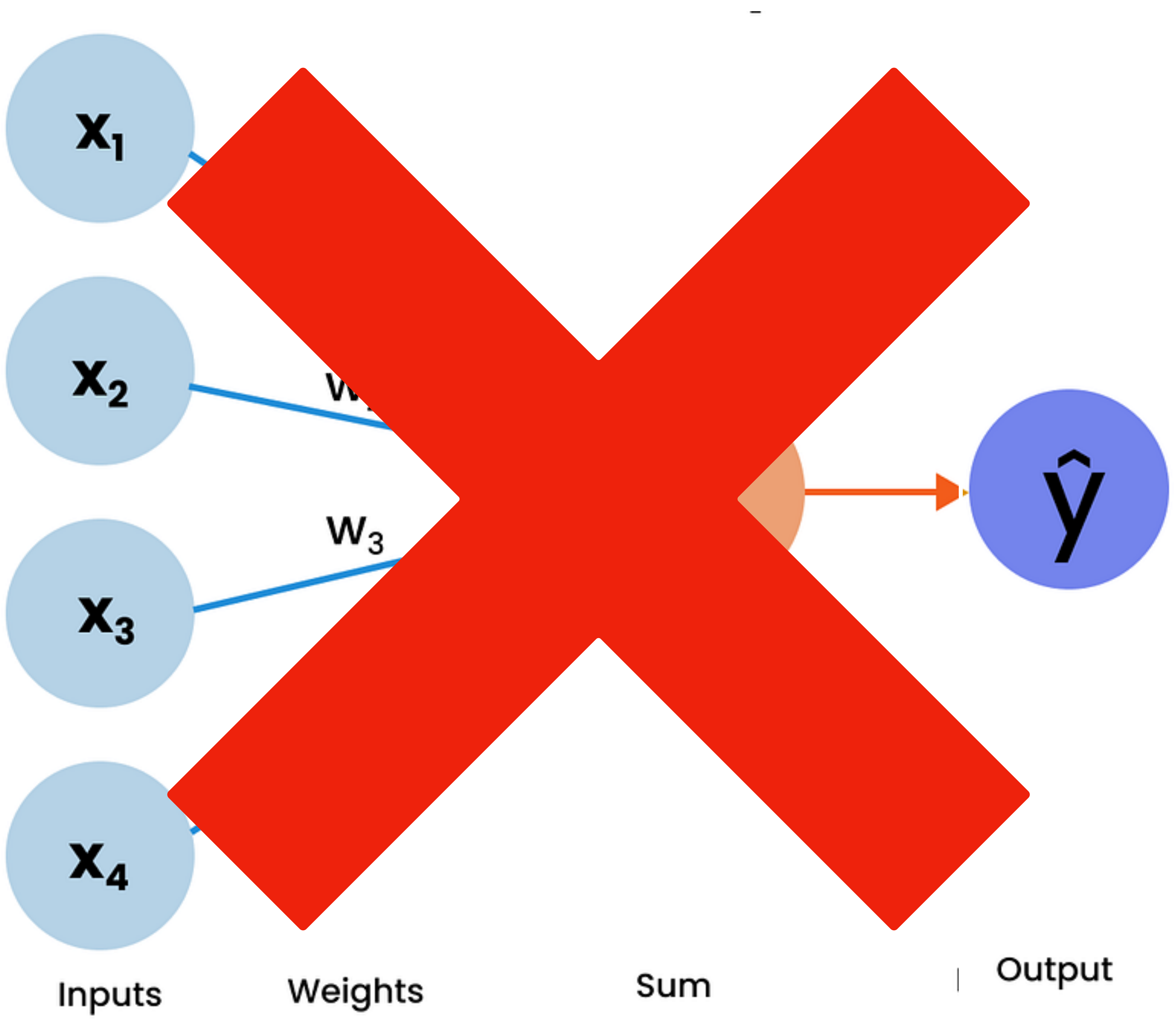
A Deep Network reaches 99.5% accuracy!

Your First Deep Network!

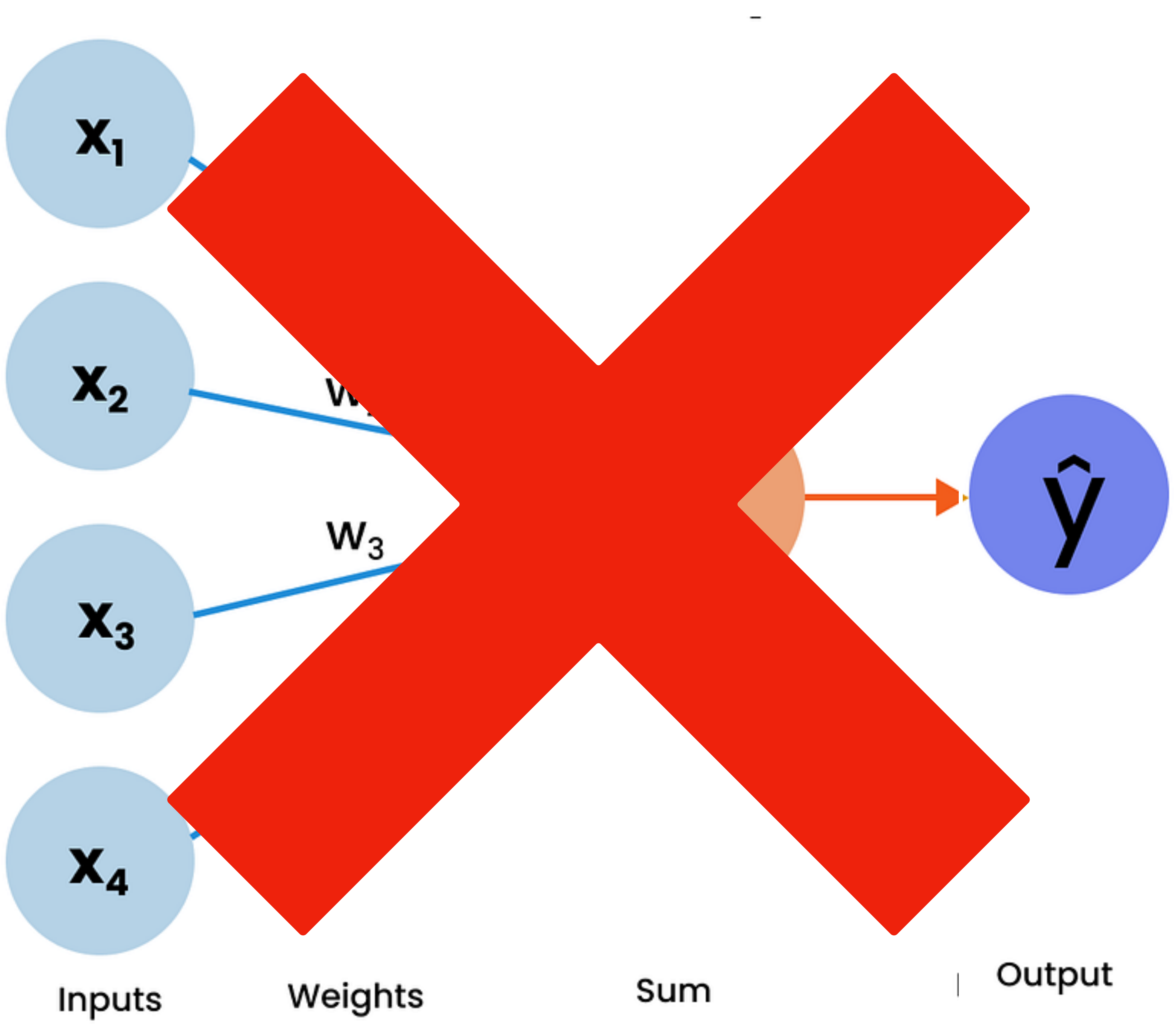
Your First Deep Network!

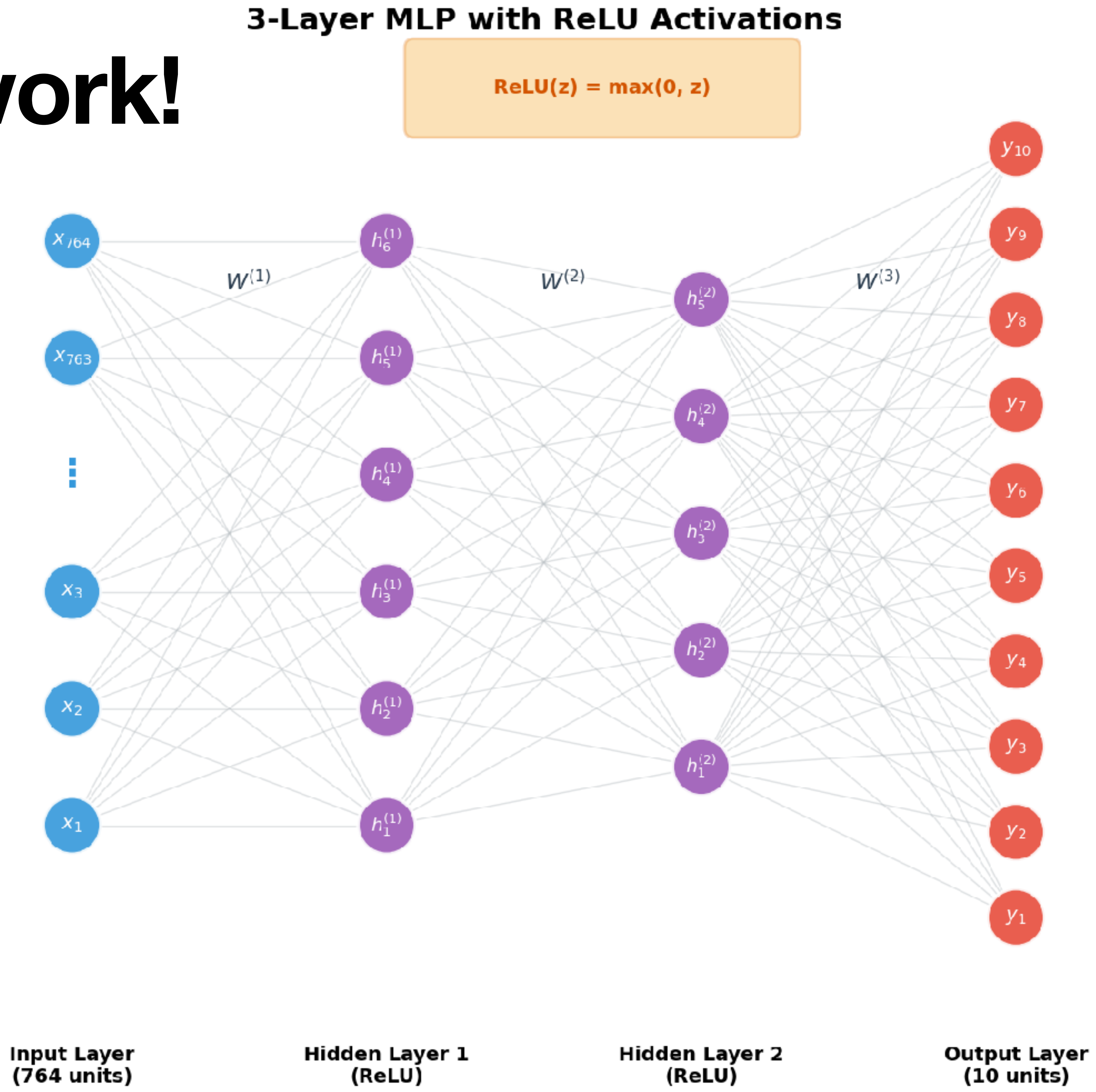


Your First Deep Network!

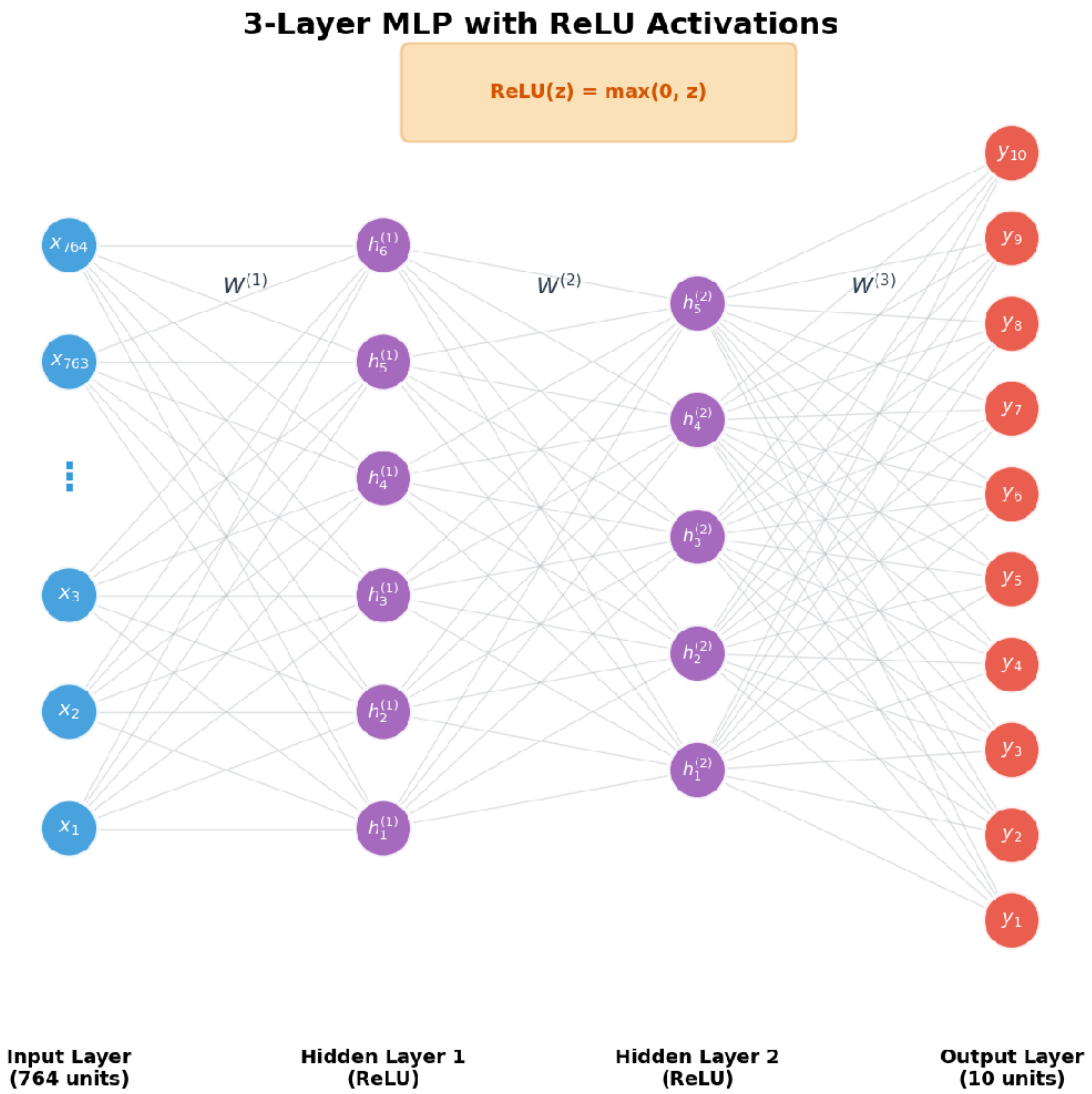


Your First Deep Network!

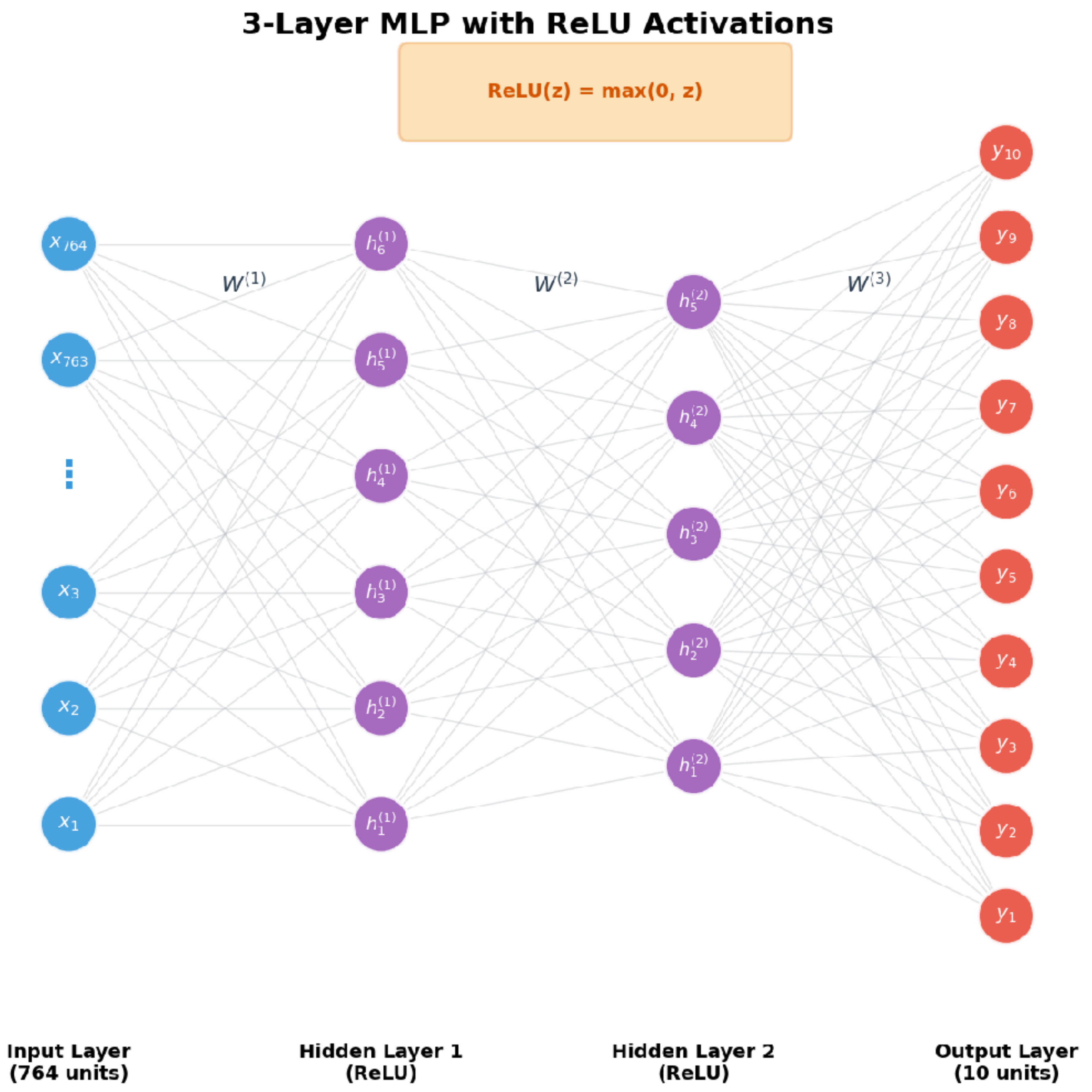




Your First Deep Network: Why?

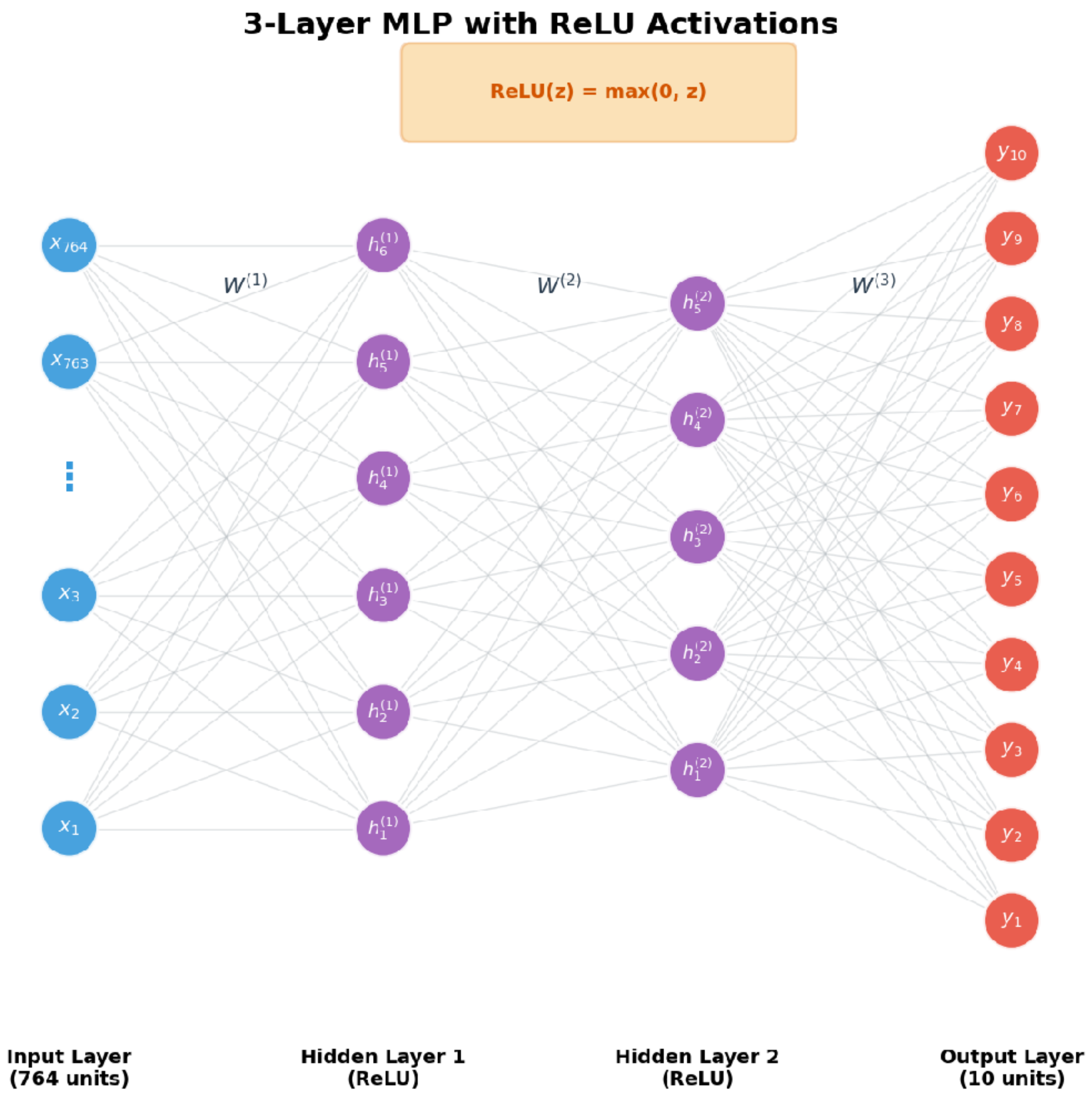


Your First Deep Network: Why?



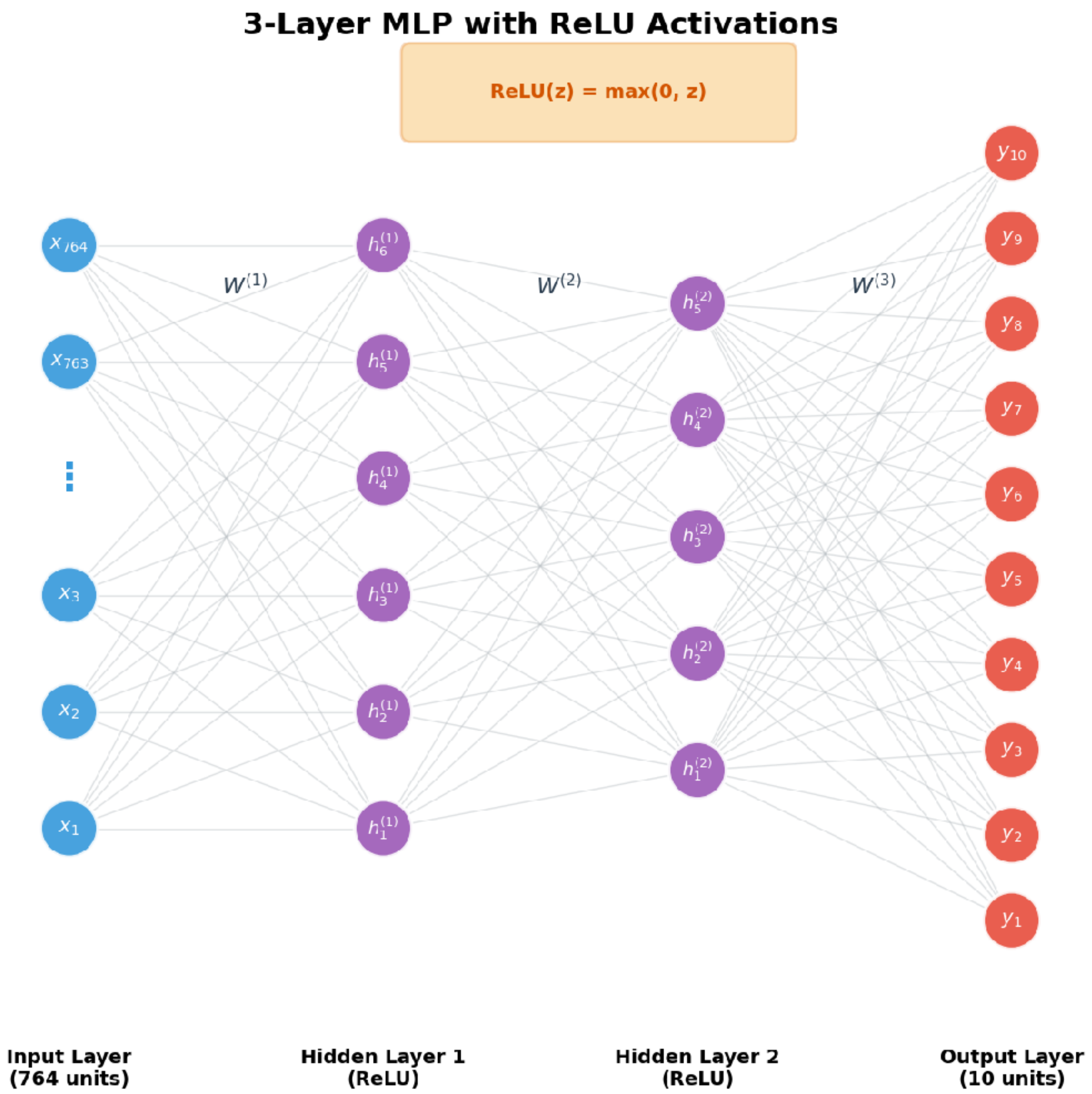
- Why 2 hidden layers?

Your First Deep Network: Why?



- Why 2 hidden layers?
- Why do we need a “ReLU”?

Your First Deep Network: Why?



- Why 2 hidden layers?
- Why do we need a “ReLU”?
- What else besides a “ReLU”?

Your First Deep Network: Why?

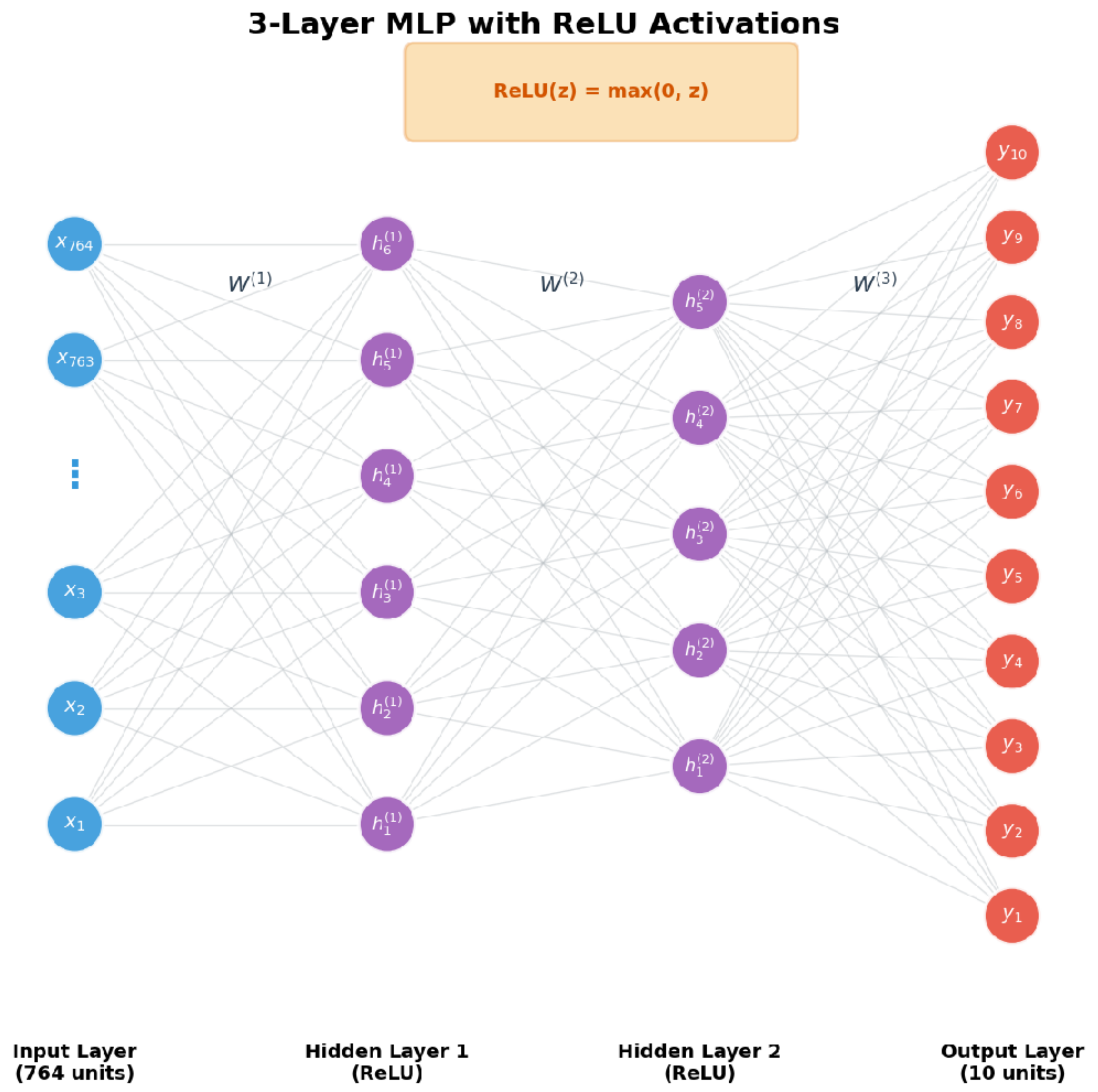
Common Activation Functions in Neural Networks
Sigmoid-like
(Bounded Output)
Sigmoid
Tanh
Hard Sigmoid
Softsign
f(z)
z
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−3
−2
−1
0
1
2
3
4
ReLU Family
(Piecewise Linear)
ReLU
Leaky ReLU (α=0.1)
ELU
SELU
f(z)
z
4
3
2
1
0
−1
−2
Smooth Modern
(Differentiable Everywhere)
GELU
Swish / SiLU
Softplus
Mish
f(z)
z
4
3
2
1
0
−1

Your First Deep Network: Why?

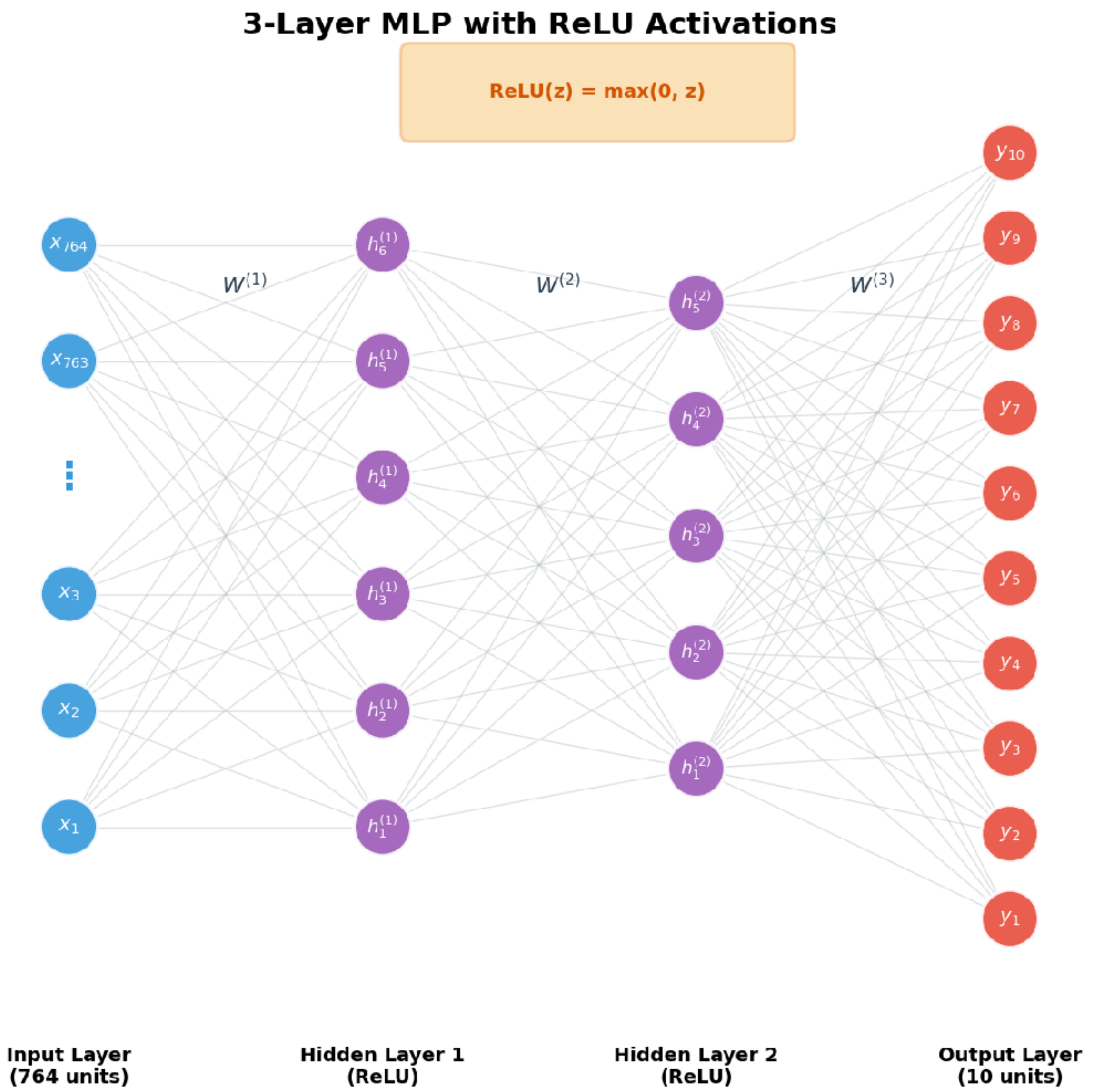
Common Activation Functions in Neural Networks
Sigmoid-like
(Bounded Output)
Sigmoid
Tanh
Hard Sigmoid
Softsign
f(z)
z
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−3
−2
−1
0
1
2
3
4
ReLU Family
(Piecewise Linear)
ReLU
Leaky ReLU (α=0.1)
ELU
SELU
4
3
2
1
0
−1
−2
Smooth Modern
(Differentiable Everywhere)
GELU
Swish / SiLU
Softplus
Mish
4
3
2
1
0
−1

Can you think of “problems” for some of them?

Your First Deep Network!

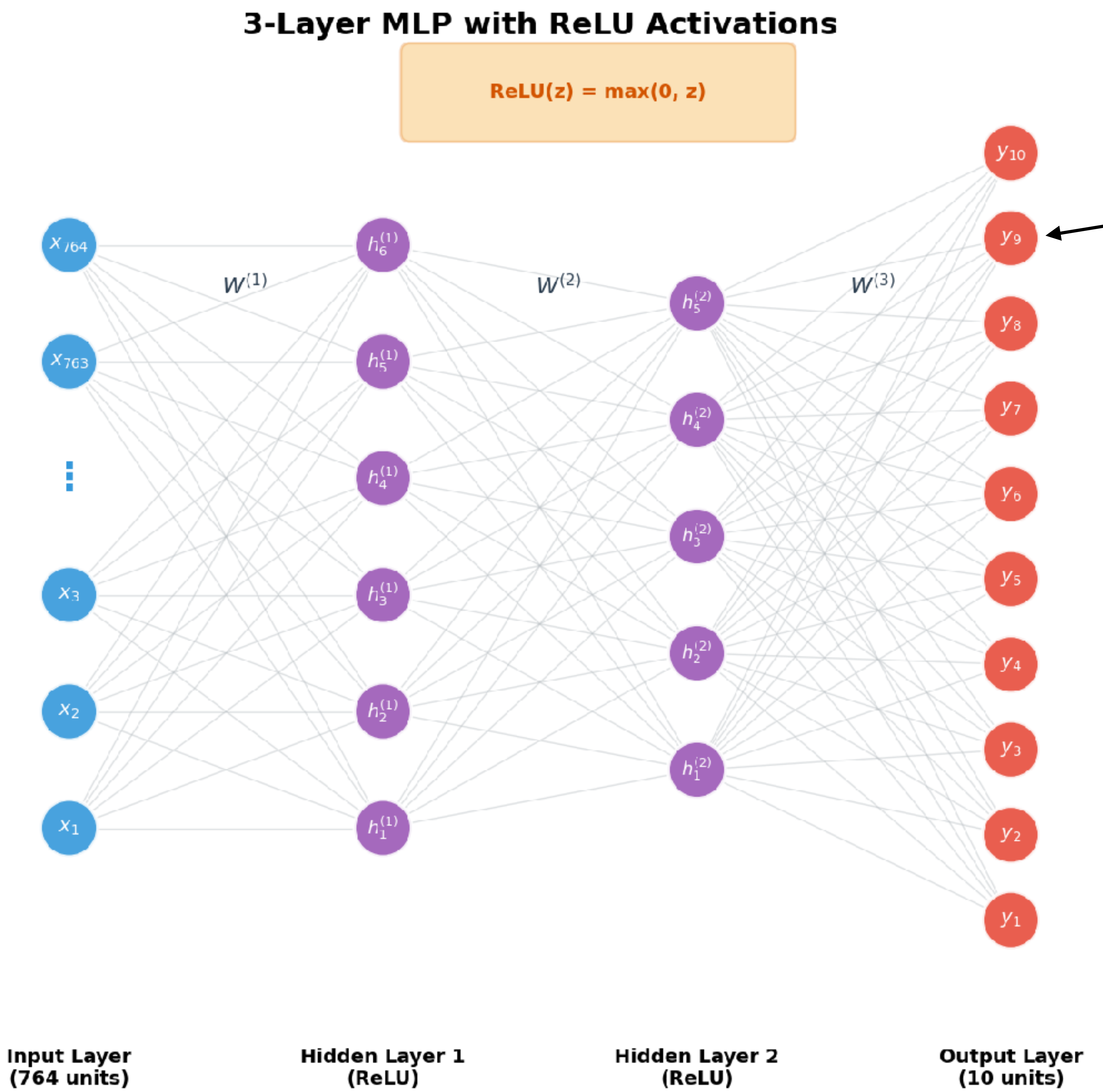


Your First Deep Network!



What training loss?

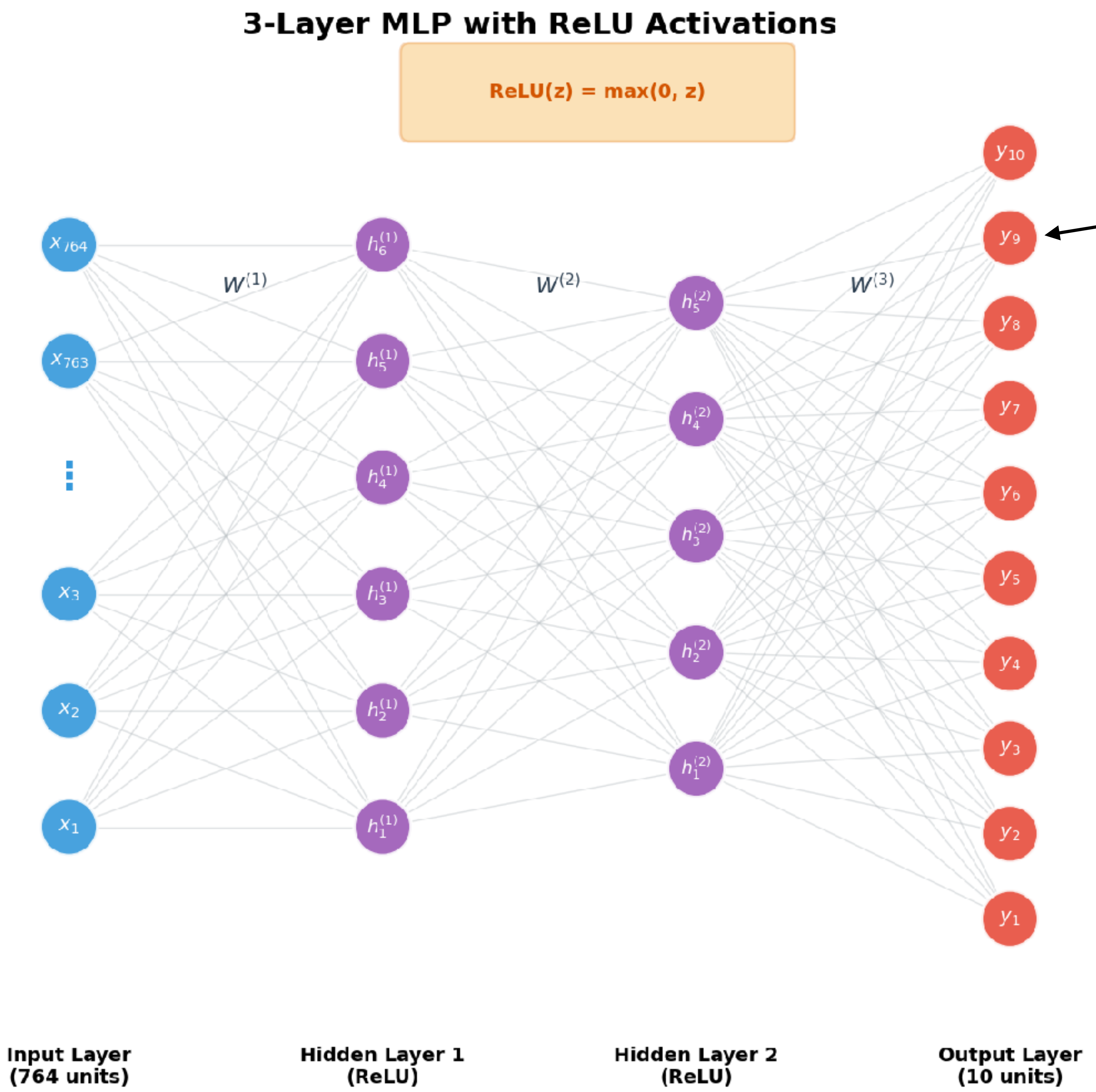
Your First Deep Network!



What training loss?

$$p(y = 9 \mid x_n)$$

Your First Deep Network!

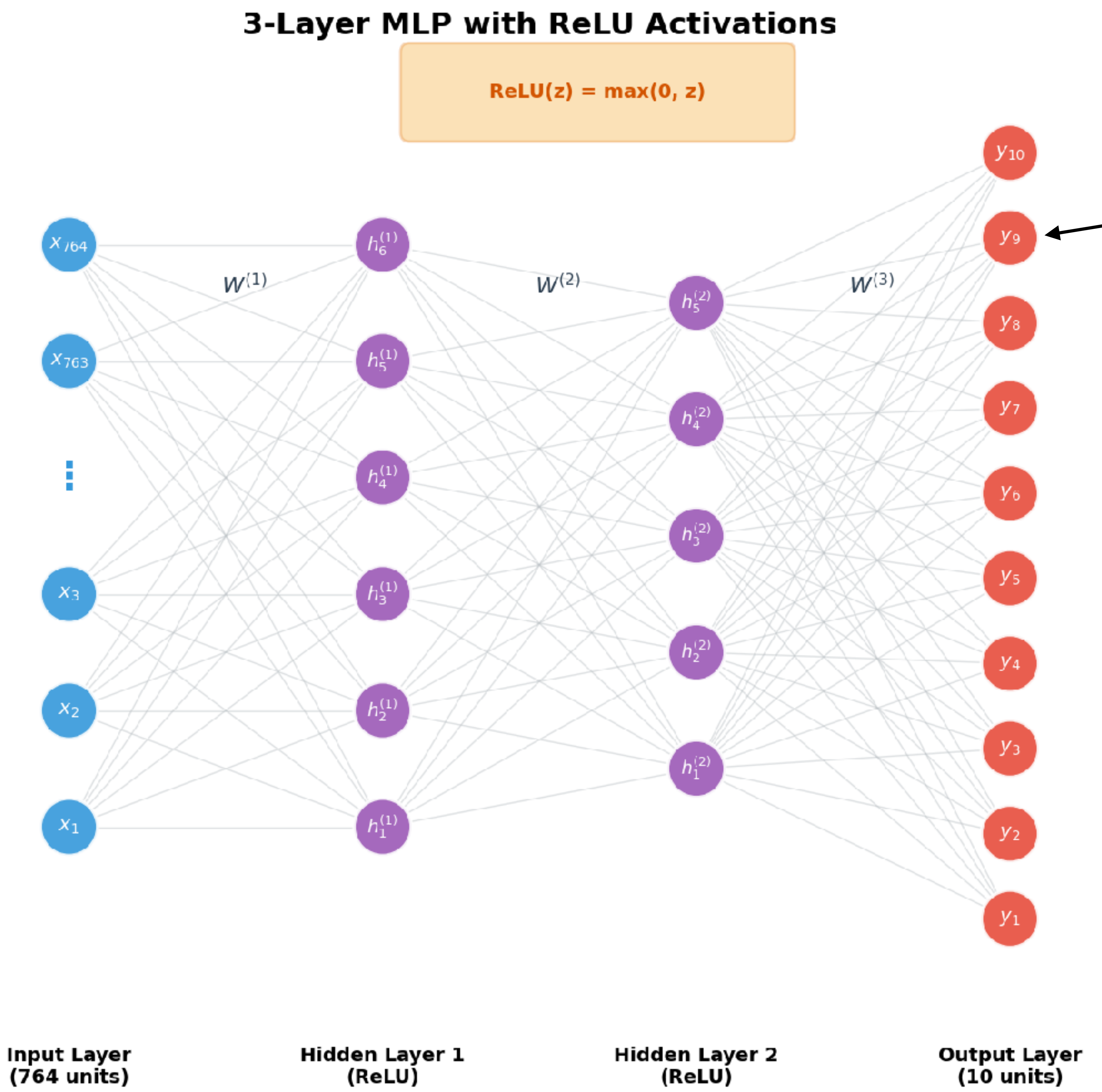


What training loss?

$p(y = 9 \mid x_n)$

$$\mathcal{L} = -\sum_{n=1}^{N} \sum_{c=1}^{10} 1_{\{y_n = c\}} \log(p(y = c \mid x_n))$$

Your First Deep Network!



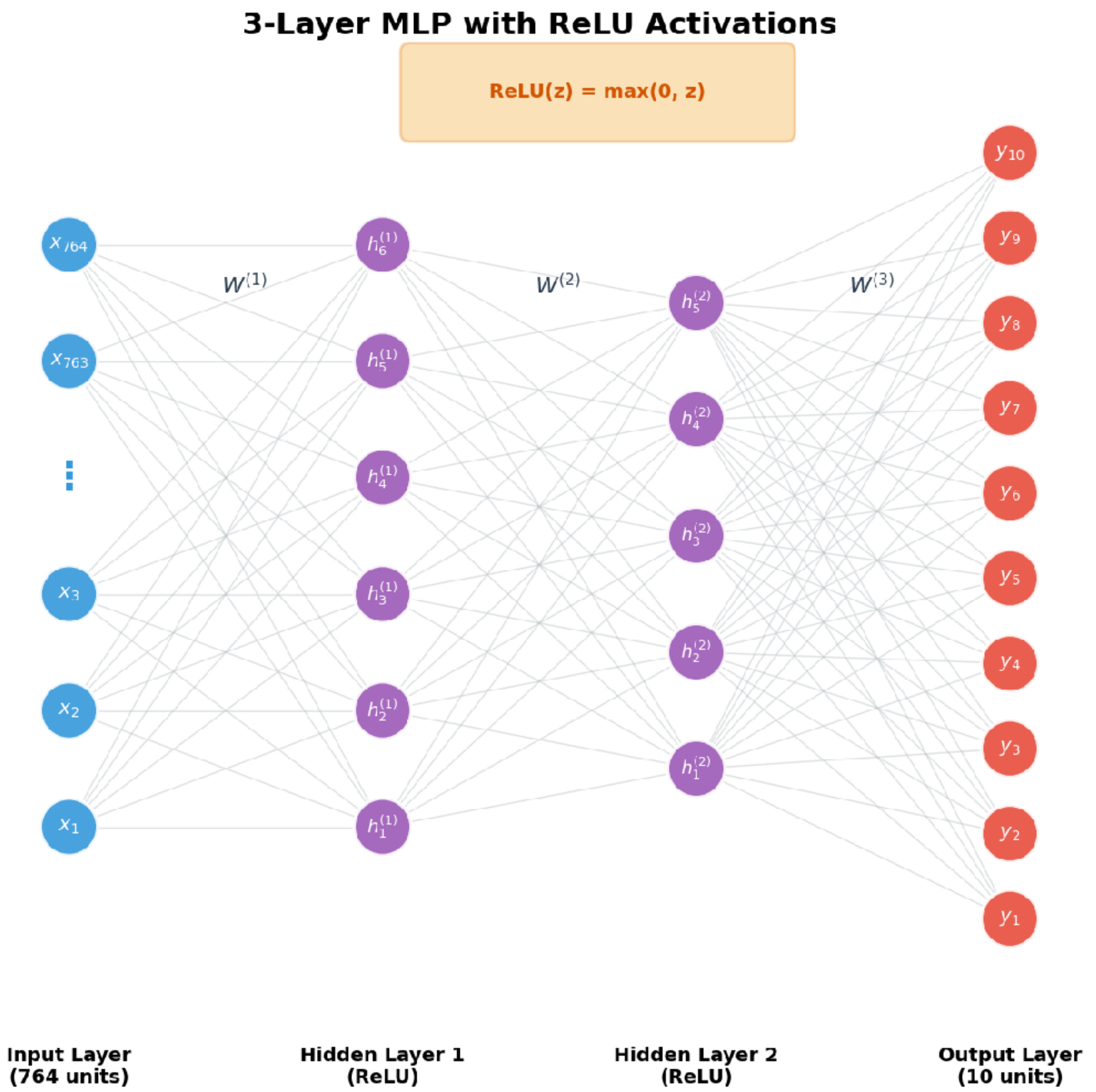
What training loss?

$p(y = 9 \mid x_n)$

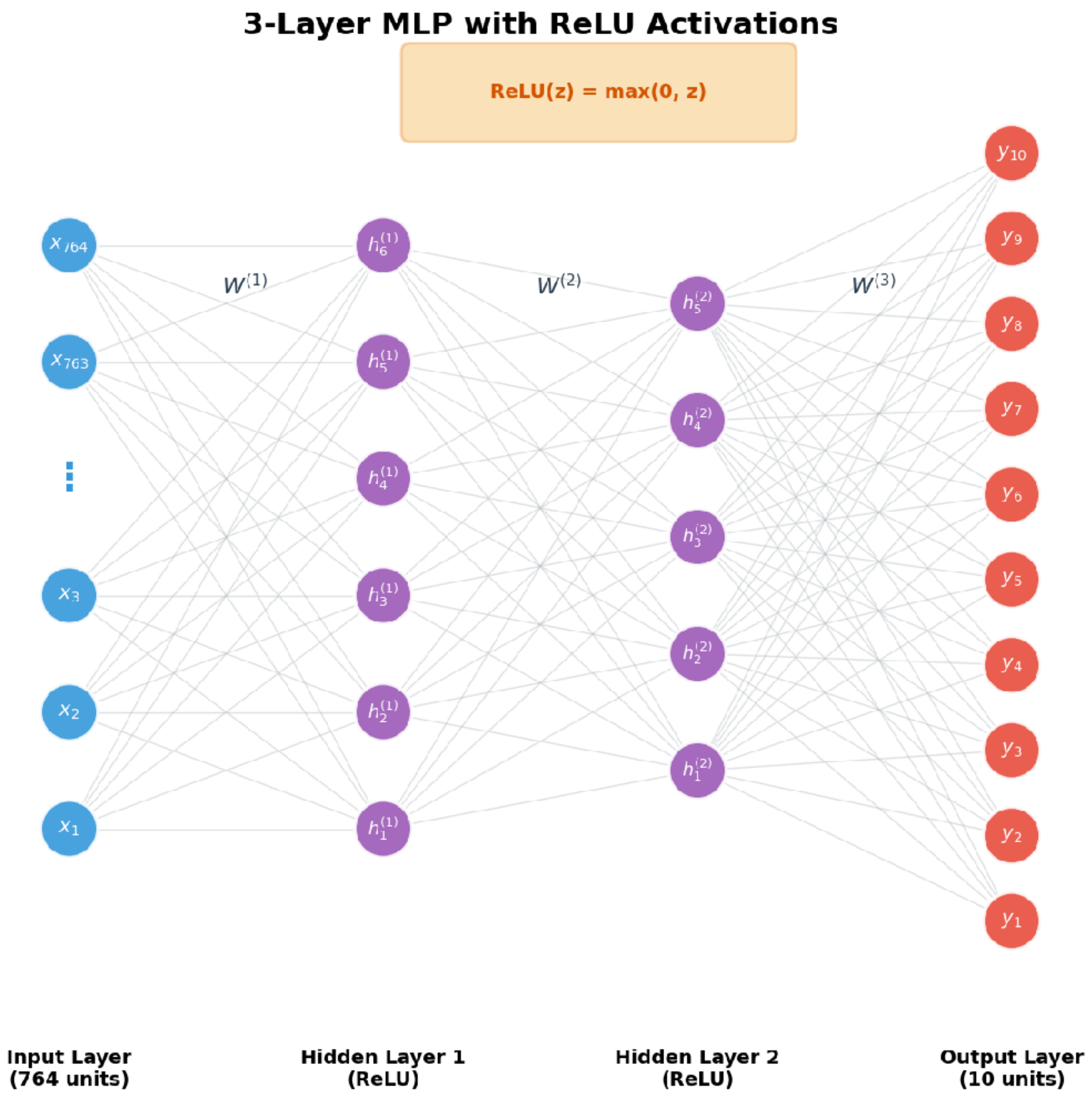
$$\mathscr{L} = -\sum_{n=1}^{N} \sum_{c=1}^{10} 1_{\{y_n = c\}} \log(p(y = c \mid x_n))$$

$$\hat{y}_n = \arg \max_{c=1,\ldots,10} p(y = c \mid x_n)$$

Your First Deep Network: Training

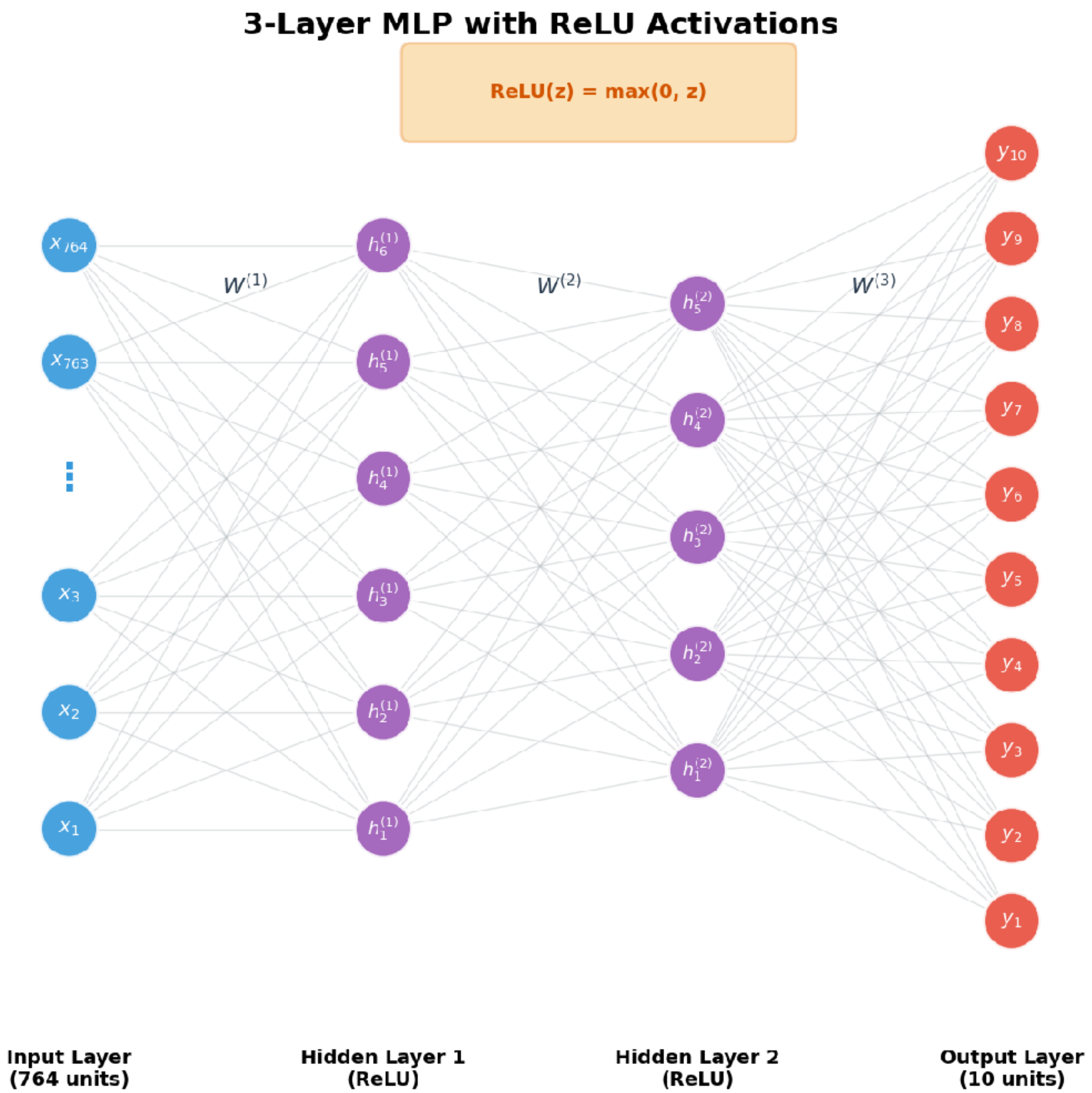


Your First Deep Network: Training



How to train the parameters $W^{(1)}, W^{(2)}, W^{(3)}$?

Your First Deep Network: Training

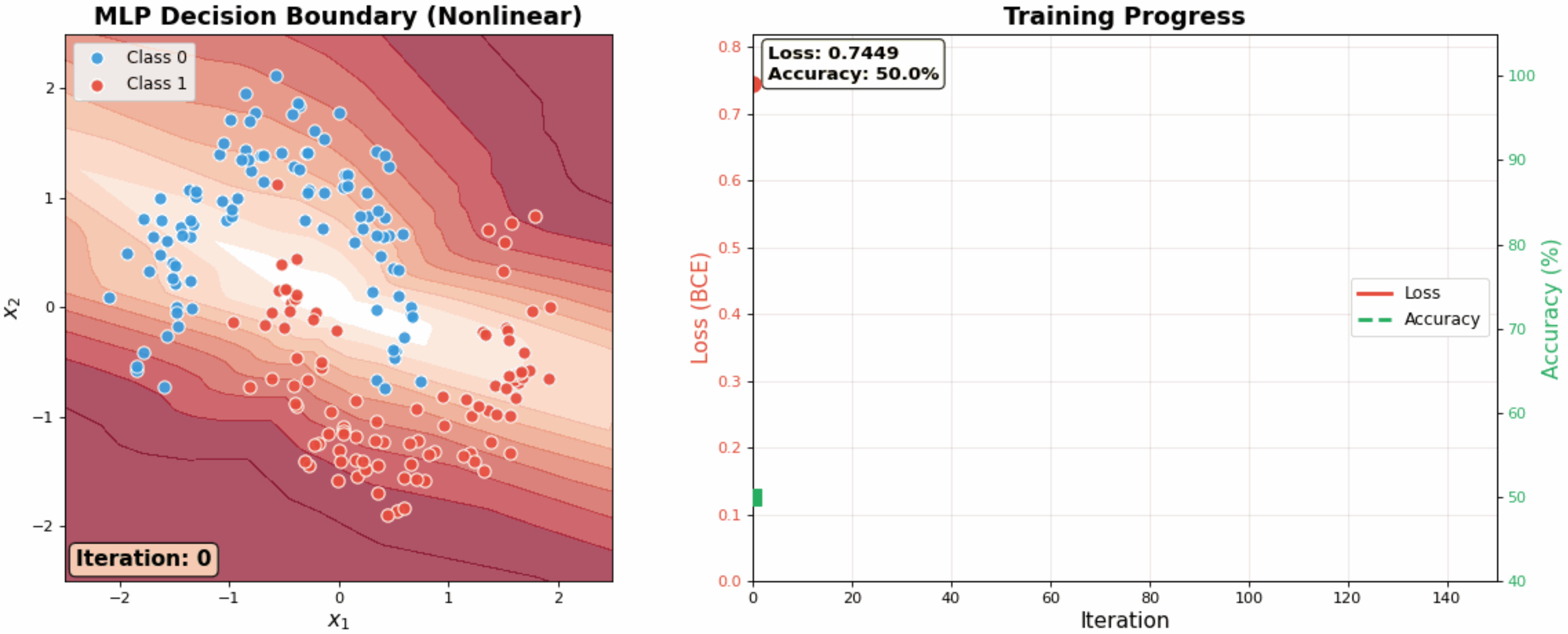


How to train the parameters $W^{(1)}, W^{(2)}, W^{(3)}$?

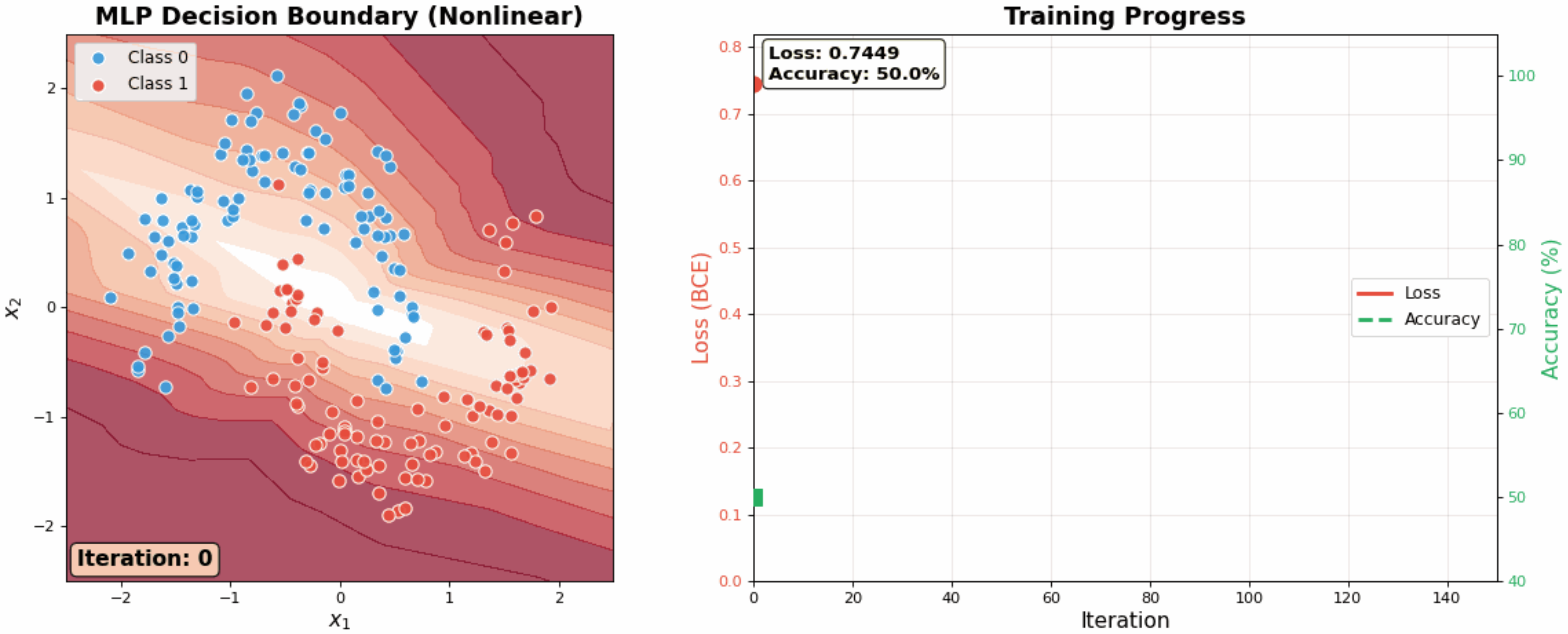
Same old gradient descent!

At home: try to derive the gradient for those 3 matrices

Your First Deep Network: Action!



Your First Deep Network: Action!



Open Discussion

Open Discussion

- Can a Deep Network solve anything?

Open Discussion

- Can a Deep Network solve anything?
- How to search for the right “architecture”?

Open Discussion

- Can a Deep Network solve anything?
- How to search for the right “architecture”?
- How many lines of codes to implement the MNIST model and reach 99.5%?

Questions?