AI VIETNAM
All-in-One Course

Multi-layer Perception

Activation and Initialization

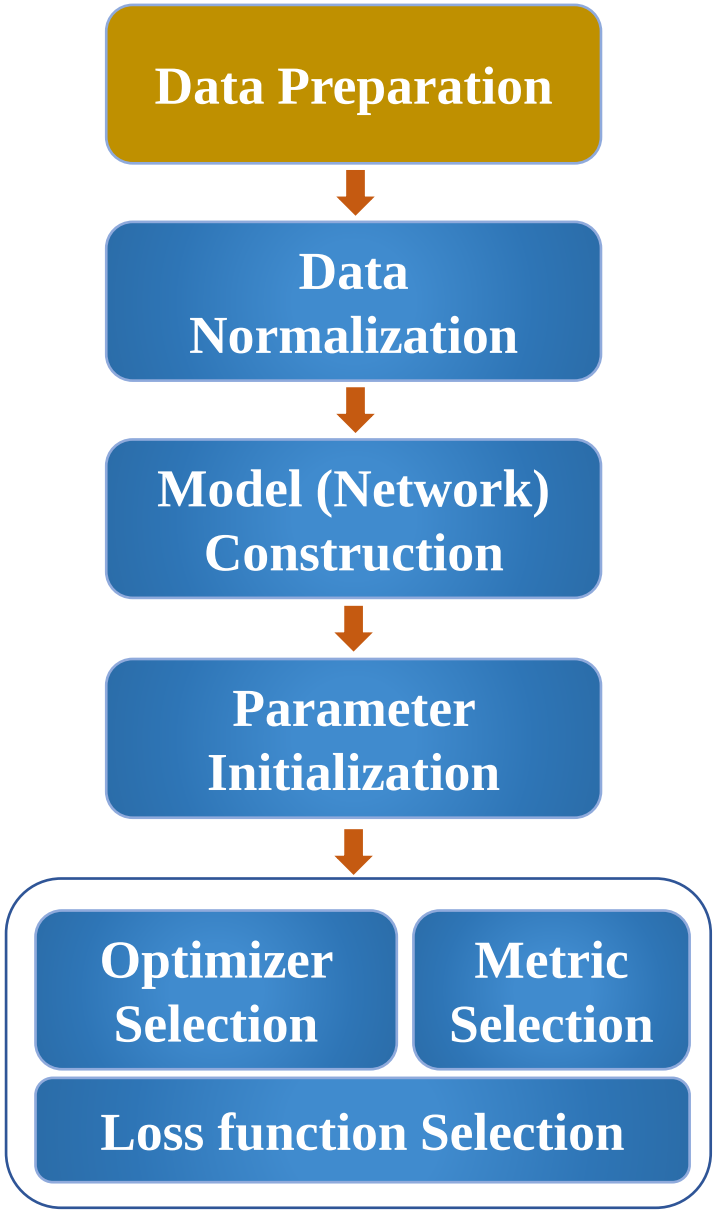
Quang-Vinh Dinh
Ph.D. in Computer Science

Year 2023

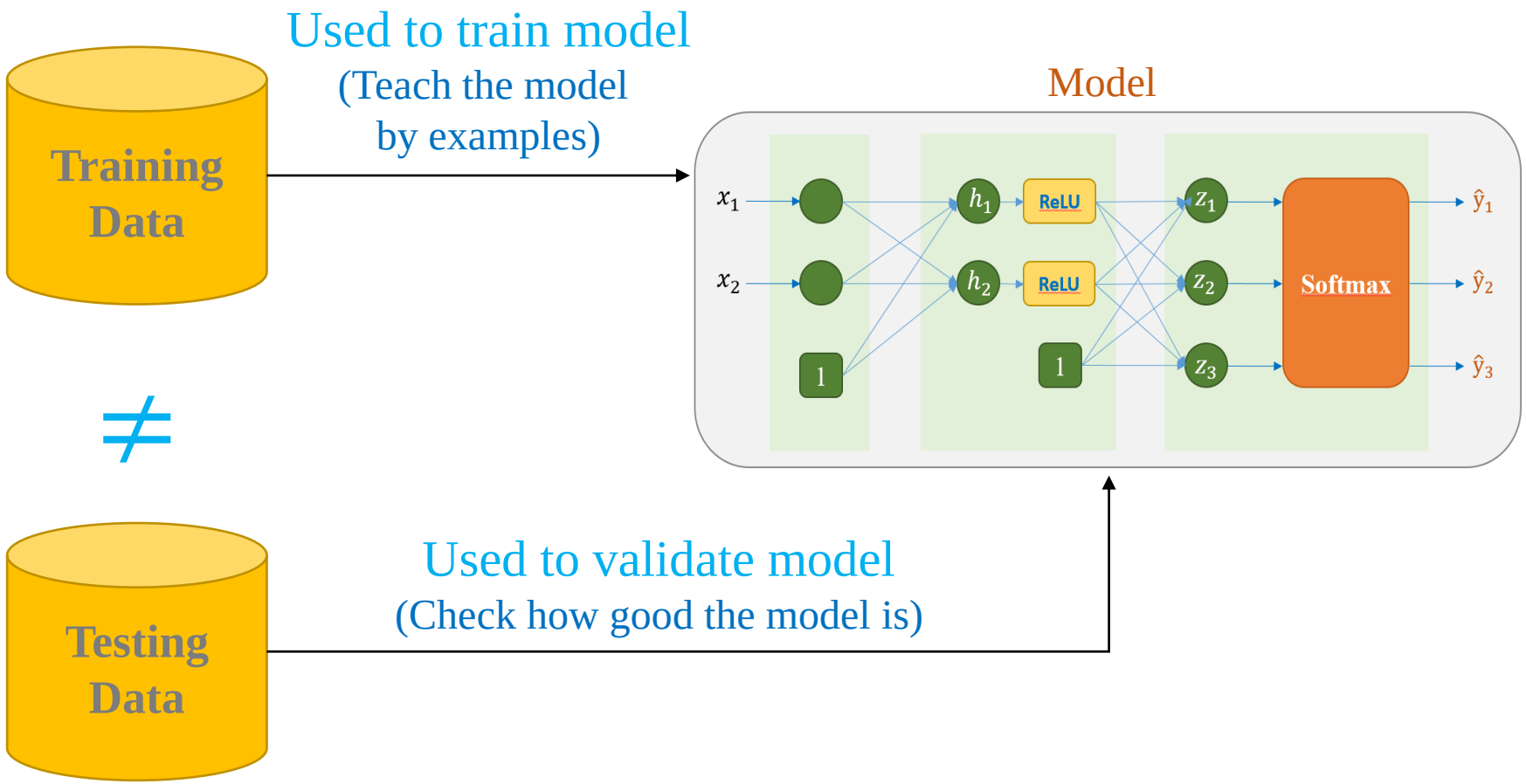
Outline

- Pipeline Recommendation
- Data Normalization
- Activation Functions
- MLP Examples
- Initialization Methods

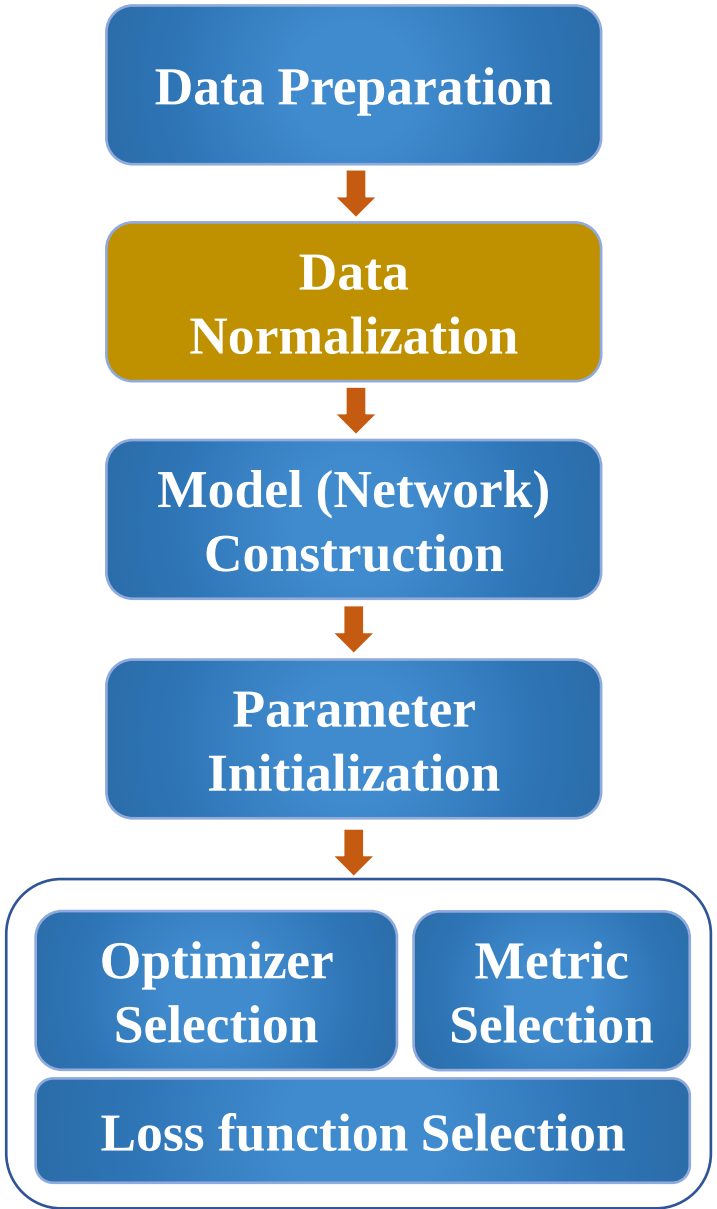
To-do List for Training



Data Preparation



Data Normalization



In Theory

$$X \in [0, 255]$$

Convert to the range [0,1]

$$\text{Image} = \frac{\text{Image}}{255}$$

Convert to the range [-1,1]

$$\text{Image} = \frac{\text{Image}}{127.5} - 1$$

Z-score normalization

$$\text{Image} = \frac{\text{Image} - \mu}{\sigma}$$

In Pytorch

$$X \in [0, 1]$$

Normalize($mean$, std)

$$\text{Image} = \frac{\text{Image} - mean}{\text{std}}$$

[0,1]	mean = 0 ; std = 1
[-1,1]	mean = 0.5; std = 0.5

Compute mean and std from data

```python
transform = transforms.Compose([transforms.ToTensor(), transforms.Normalize((0.5,), (0.5,))])

trainset = torchvision.datasets.FashionMNIST(root='data', train=True, download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=1024, num_workers=10, shuffle=True)

testset = torchvision.datasets.FashionMNIST(root='data', train=False, download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=1024, num_workers=10, shuffle=False)
```

1

(a) [0, 1] Normalization

```python
transform = transforms.Compose([transforms.ToTensor(), transforms.Normalize((0,), (1.0,))])

trainset = torchvision.datasets.FashionMNIST(root='data', train=True, download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=1024, num_workers=10, shuffle=True)

testset = torchvision.datasets.FashionMNIST(root='data', train=False, download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=1024, num_workers=10, shuffle=False)
```

2

(b) [-1, 1] Normalization

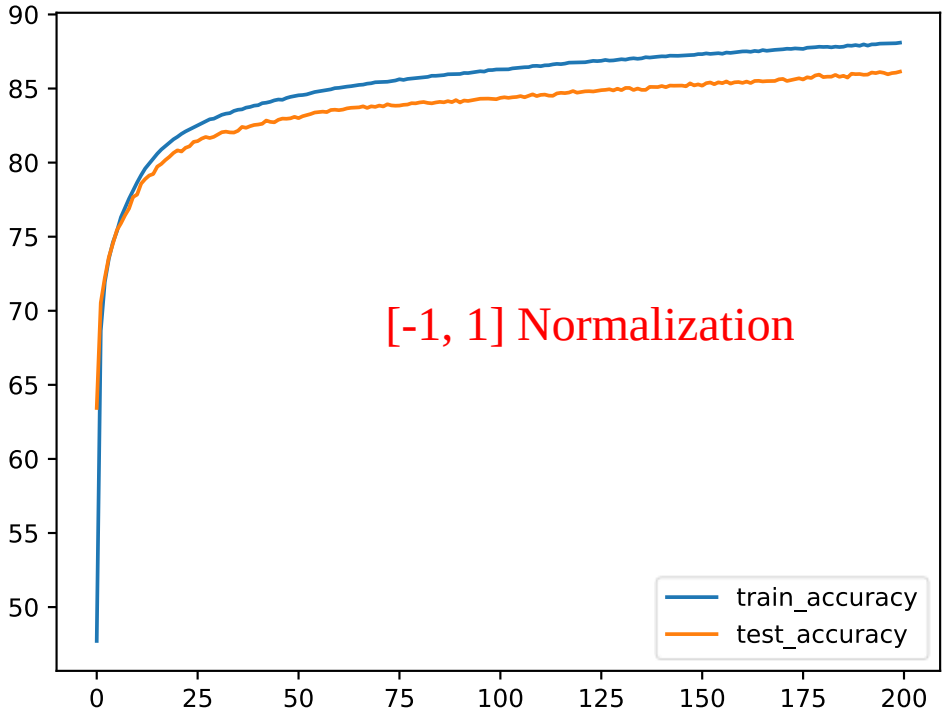
```python
# computed mean and std in advance
transform = transforms.Compose([transforms.ToTensor(), transforms.Normalize((mean,), (std,))])

trainset = torchvision.datasets.FashionMNIST(root='data', train=True, download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=1024, num_workers=10, shuffle=True)

testset = torchvision.datasets.FashionMNIST(root='data', train=False, download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=1024, num_workers=10, shuffle=False)
```

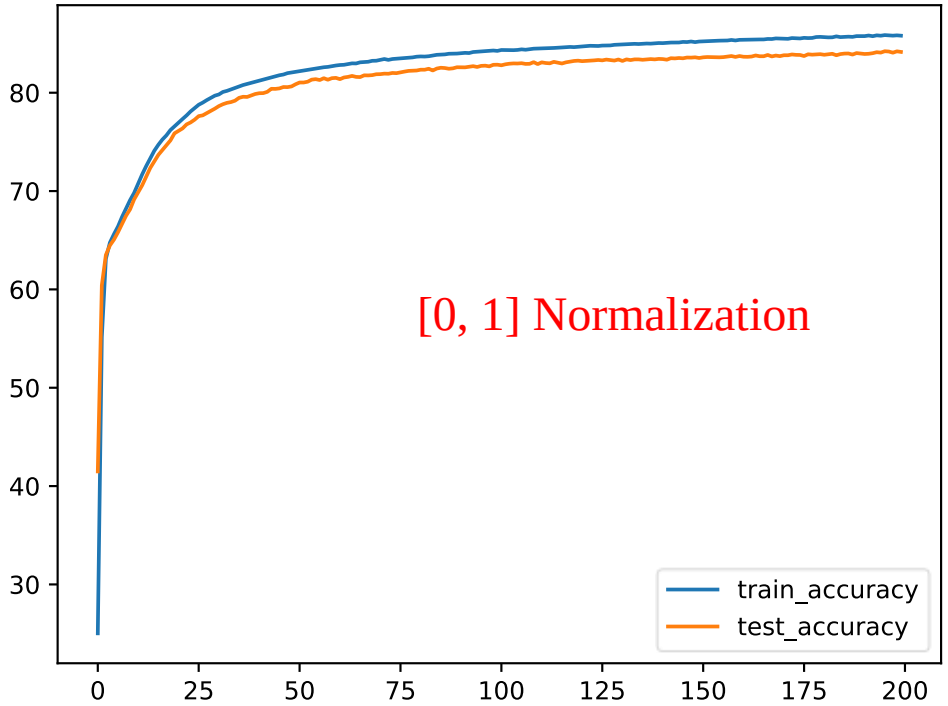
3

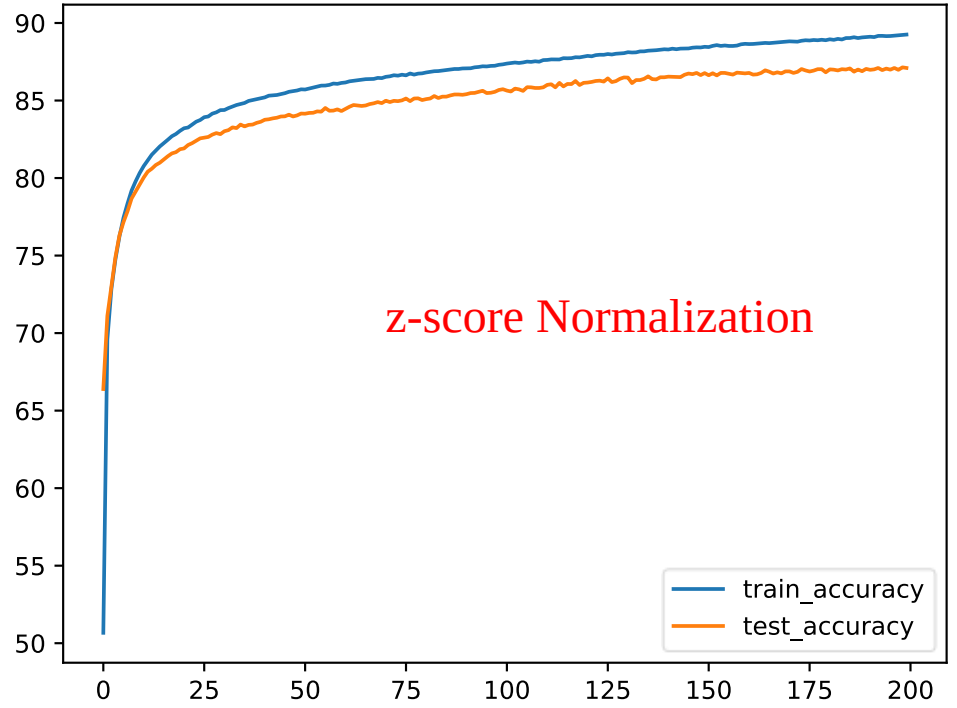
(c) z-score Normalization

Data Normalization

```
model = nn.Sequential(
    nn.Flatten(), nn.Linear(784, 256),
    nn.ReLU(), nn.Linear(256, 10)
)
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(),
                      lr=0.01)
```







Training Pipeline

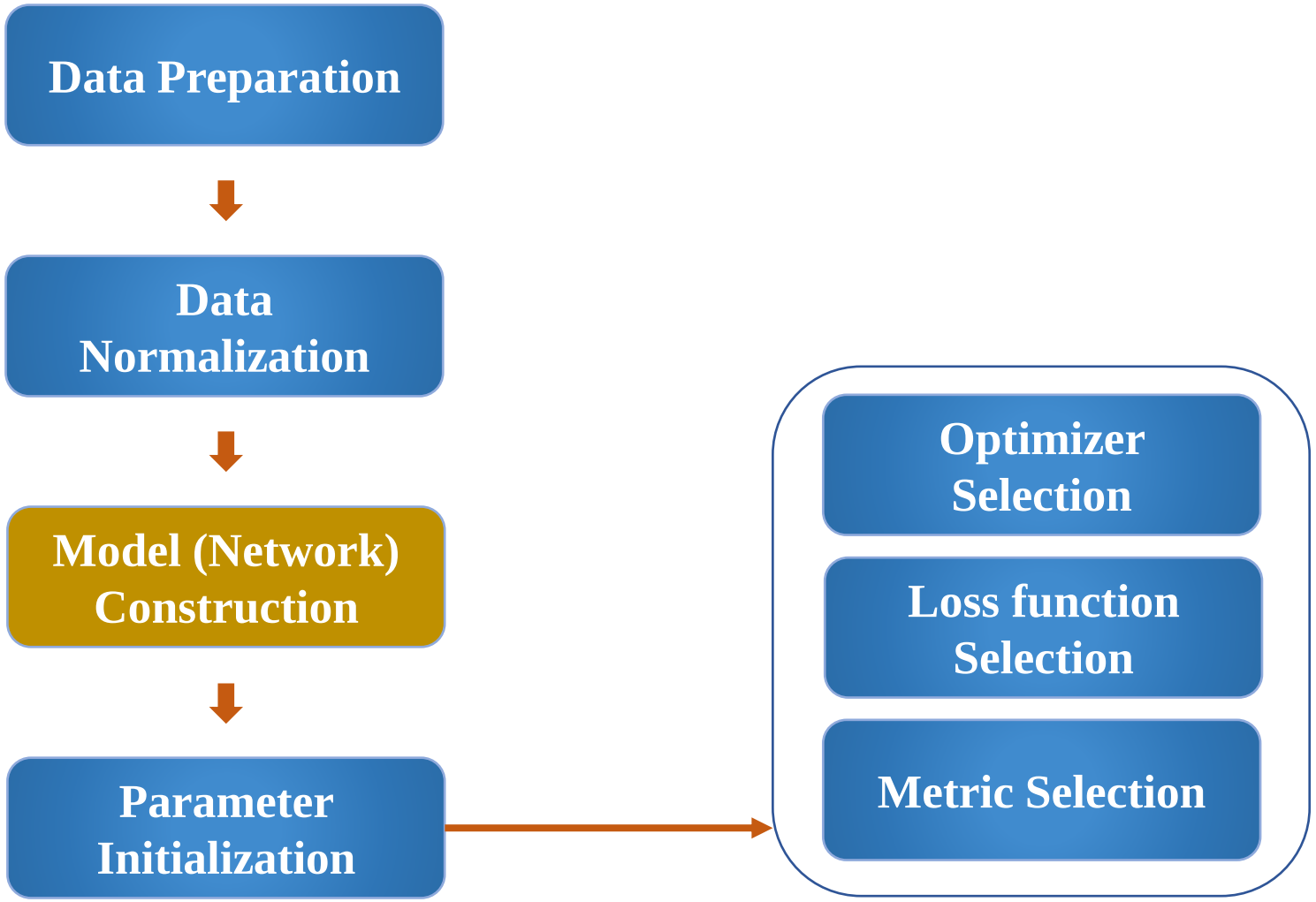
Multi-layer Perceptron

1) #Hidden Layers?

2) #Nodes in a Hidden Layers?

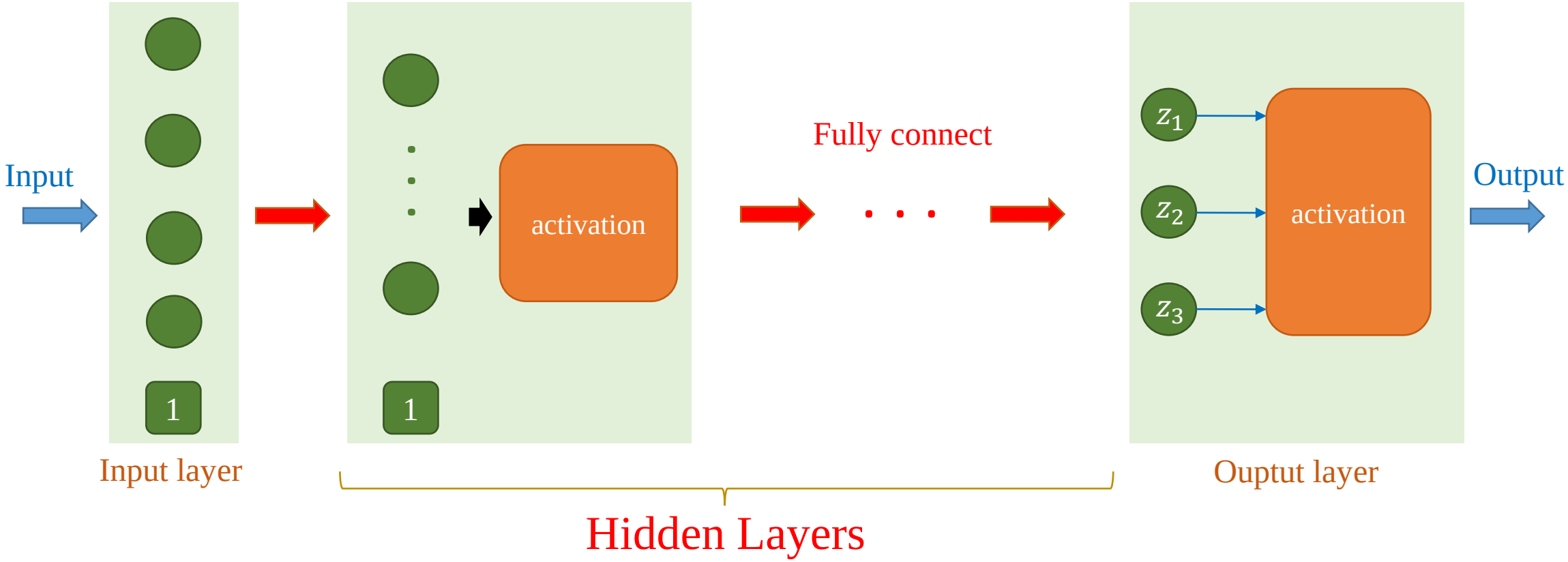
3) Which activation functions?

4) Which Initializers?



Training Pipeline

Model (Network) Construction

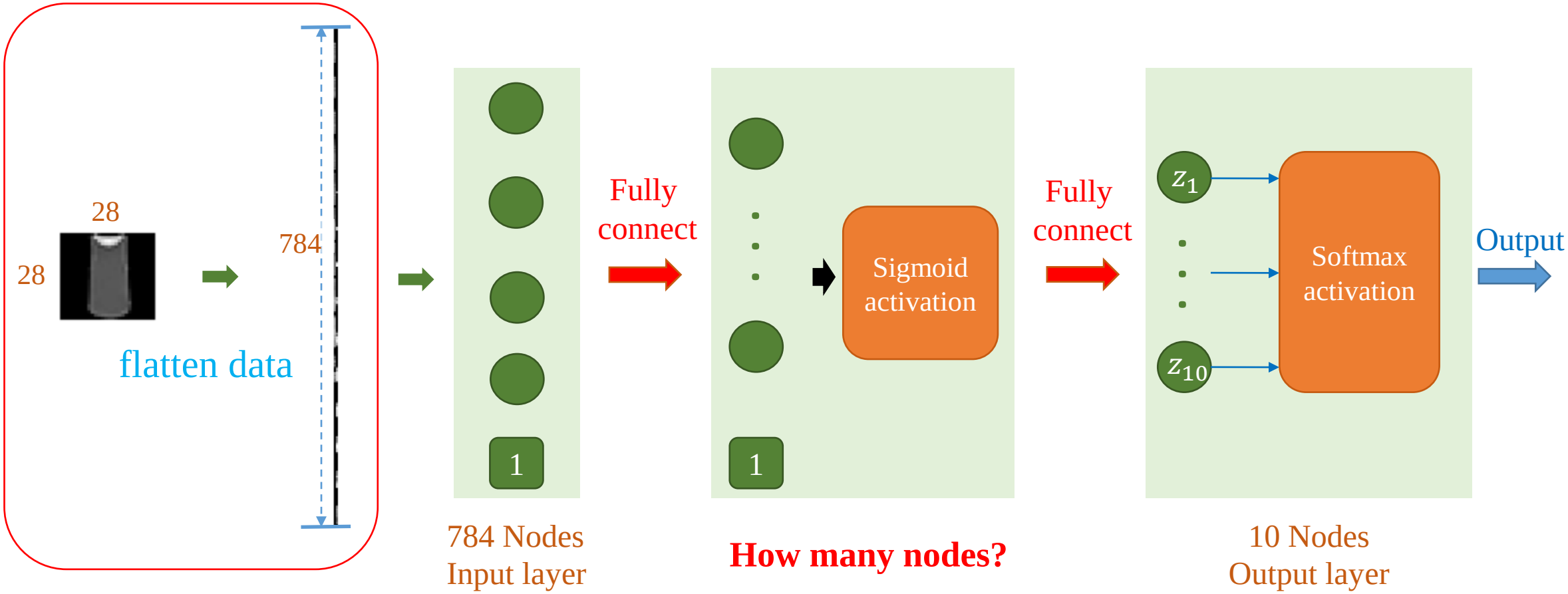


How many hidden layers?
How many nodes in a hidden layer?

Which activation function?
Which network components?

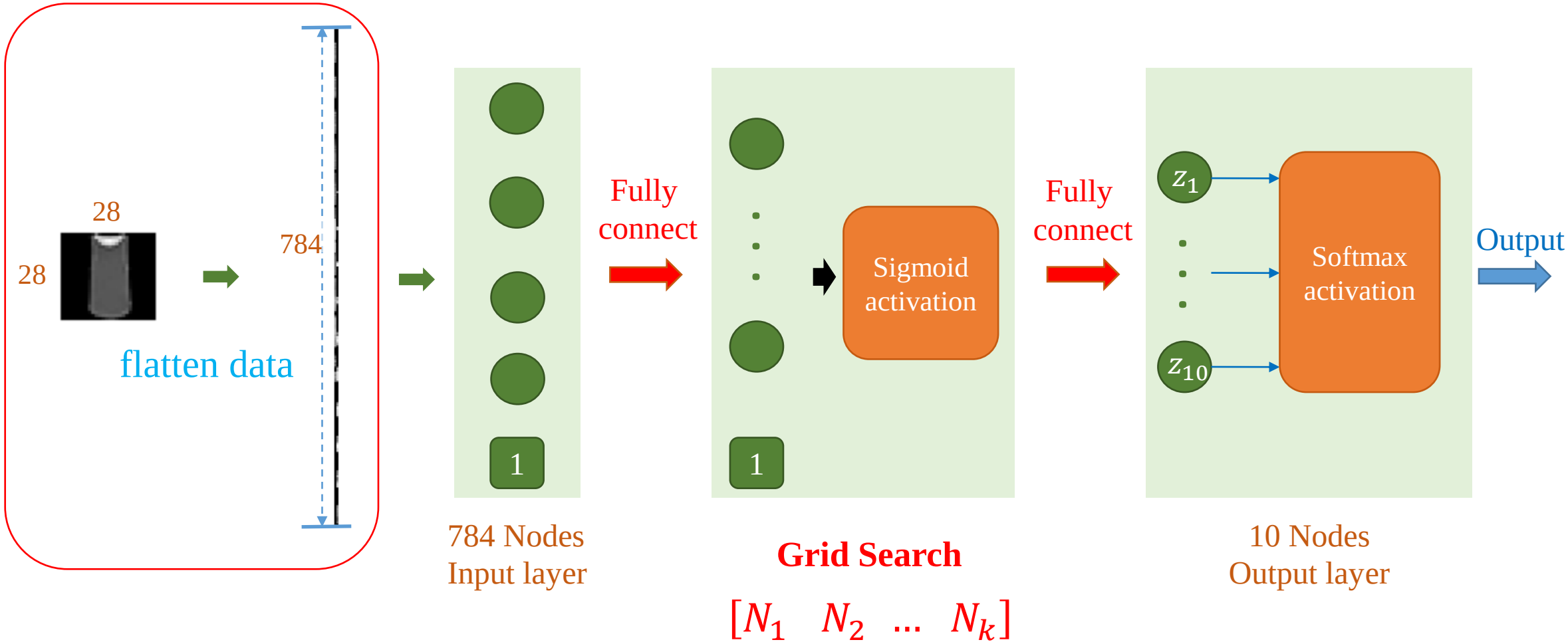
How many nodes?

Model (Network) Construction



How many nodes?

Model (Network) Construction

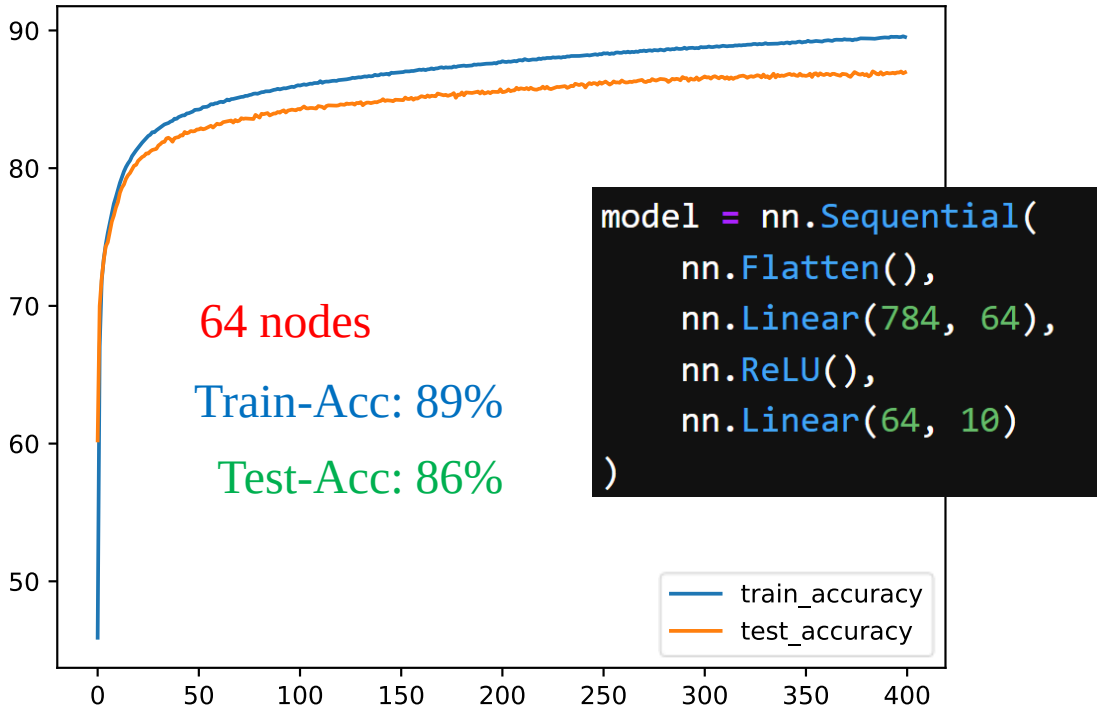


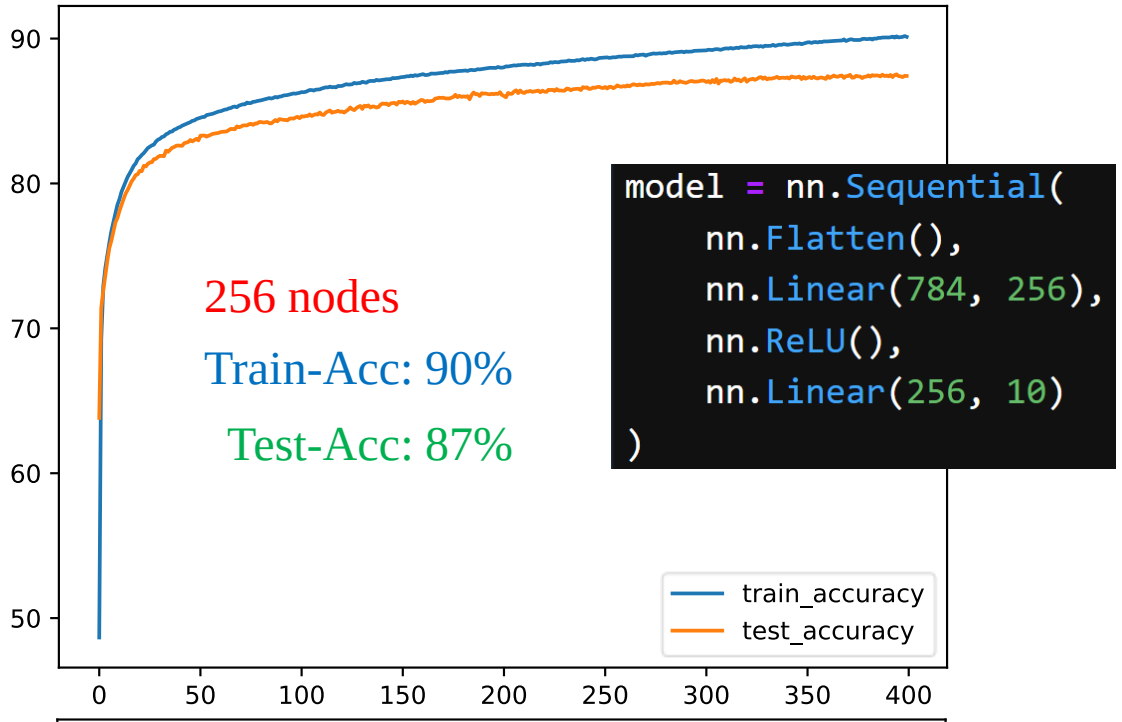
How many nodes?

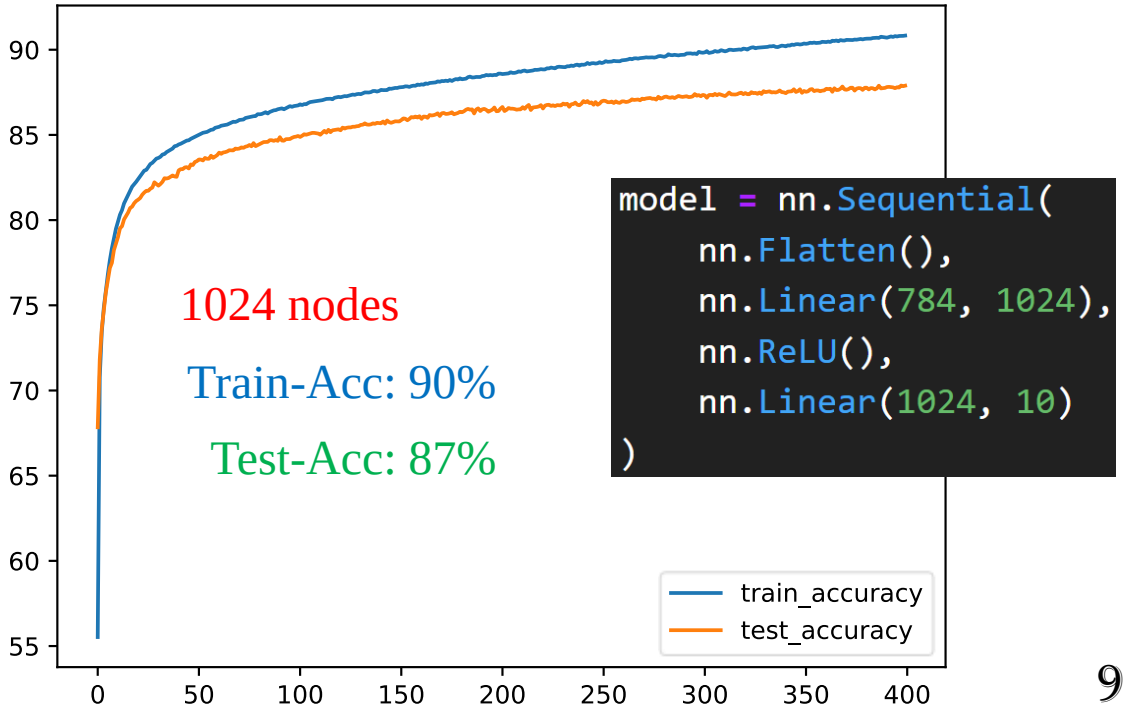
[-1, 1] Normalization

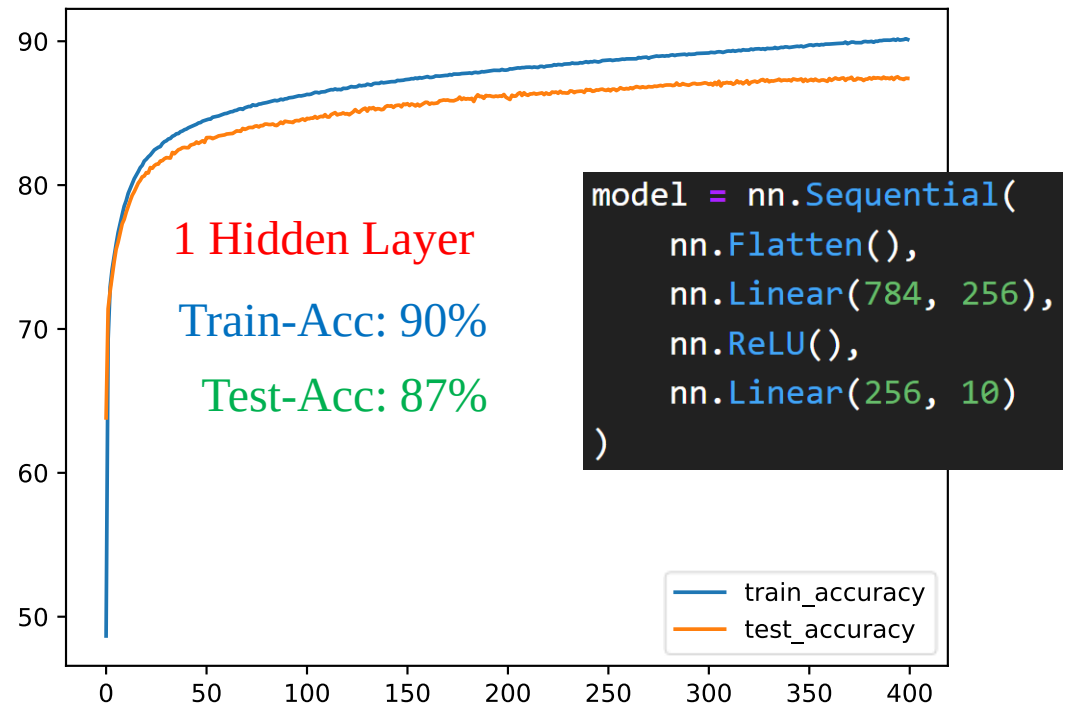
Cross-entropy Loss

SGD with lr=0.01







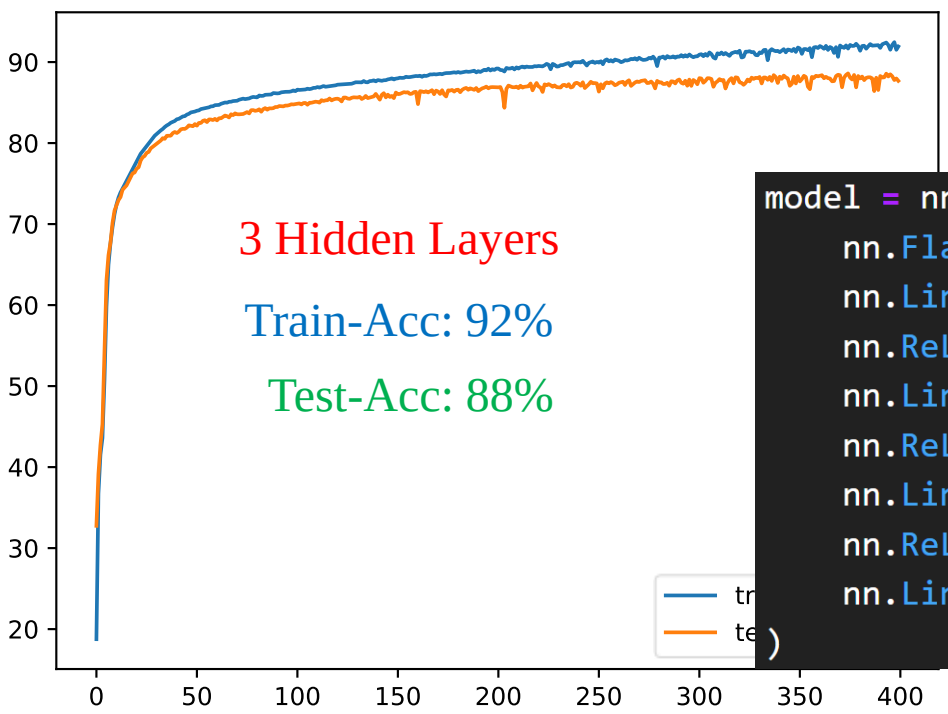


1 Hidden Layer

Train-Acc: 90%

Test-Acc: 87%

```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 256),
    nn.ReLU(),
    nn.Linear(256, 10)
)
```

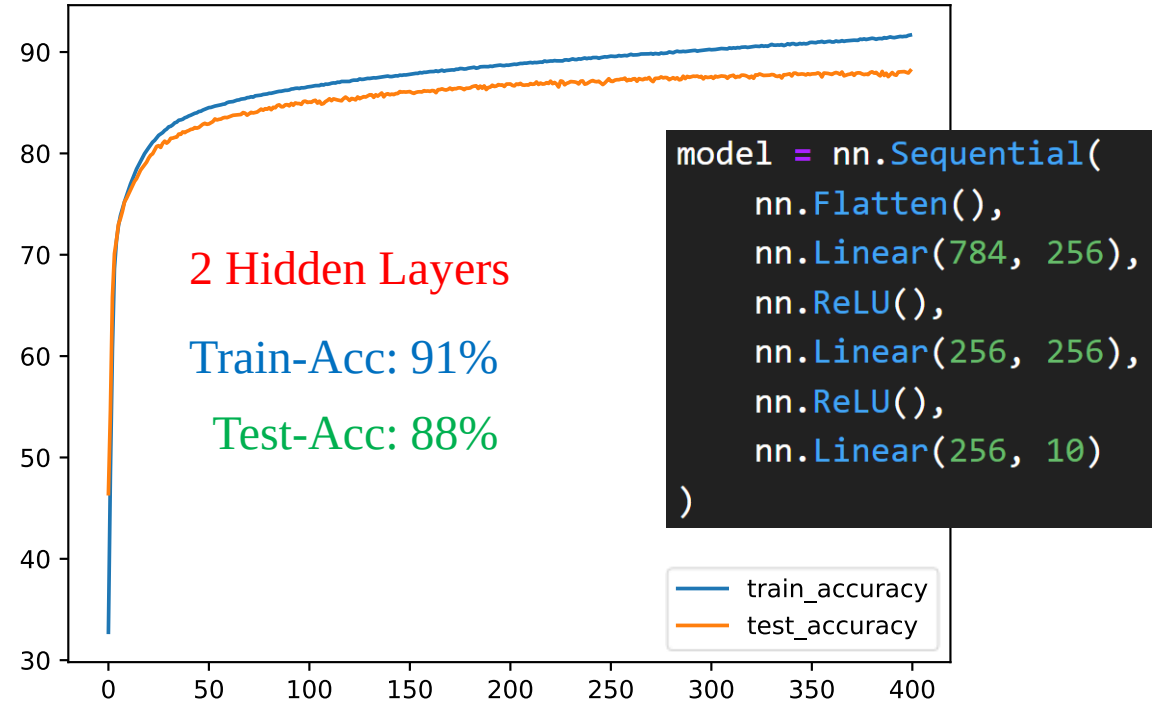


3 Hidden Layers

Train-Acc: 92%

Test-Acc: 88%

```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 256),
    nn.ReLU(),
    nn.Linear(256, 256),
    nn.ReLU(),
    nn.Linear(256, 256),
    nn.ReLU(),
    nn.Linear(256, 10)
)
```

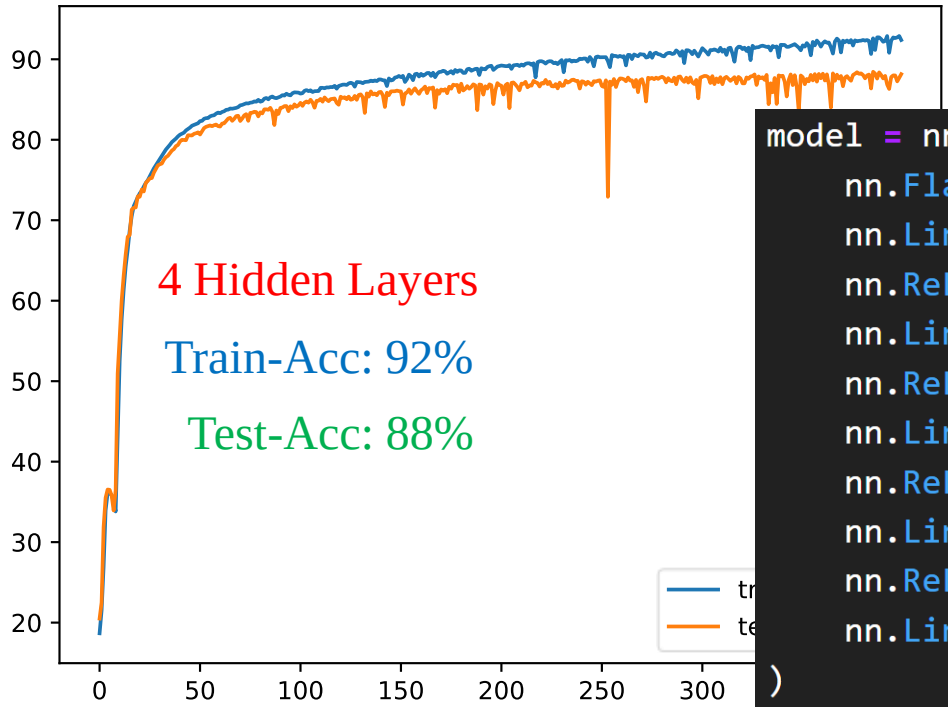


2 Hidden Layers

Train-Acc: 91%

Test-Acc: 88%

```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 256),
    nn.ReLU(),
    nn.Linear(256, 256),
    nn.ReLU(),
    nn.Linear(256, 10)
)
```



4 Hidden Layers

Train-Acc: 92%

Test-Acc: 88%

```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 256),
    nn.ReLU(),
    nn.Linear(256, 256),
    nn.ReLU(),
    nn.Linear(256, 256),
    nn.ReLU(),
    nn.Linear(256, 10)
)
```

Activation Functions

Model (Network) Construction

Which activation function?

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

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

2001

$$\text{softplus}(x) = \log(1 + e^{x})$$

2010

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

2015

$$\text{ELU}(x) = \begin{cases} \alpha(e^{x} - 1) & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

2015

$$\text{PReLU}(x) = \begin{cases} \alpha x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

2017

$$SELU(x) = \begin{cases} \lambda x & \text{if } x \geq 0 \\ \lambda\alpha(e^{x} - 1) & \text{if } x < 0 \end{cases}$$

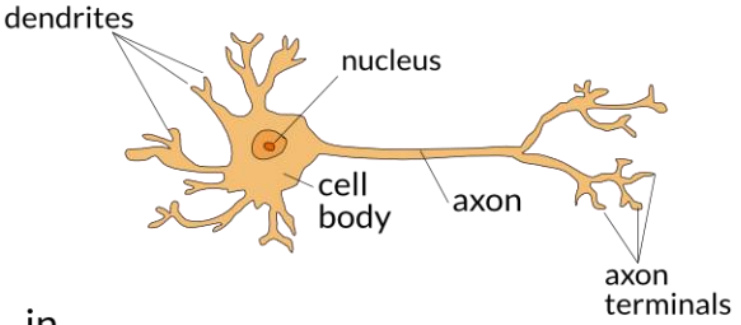
$$\lambda \approx 1.0507$$

$$\alpha \approx 1.6733$$

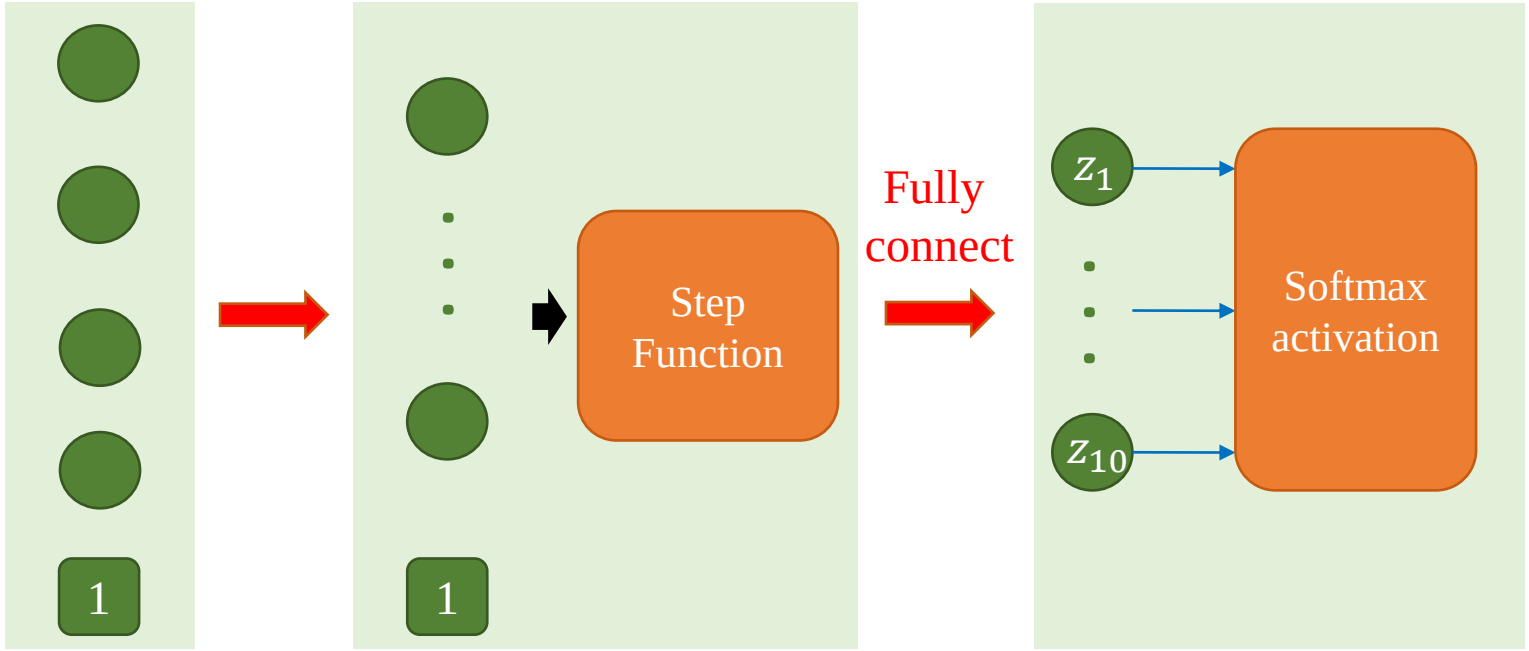
2017

$$swish(x) = x * \frac{1}{1 + e^{-x}}$$

Activation Functions

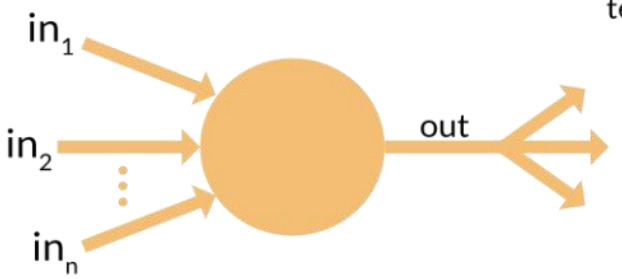


❖ Step function

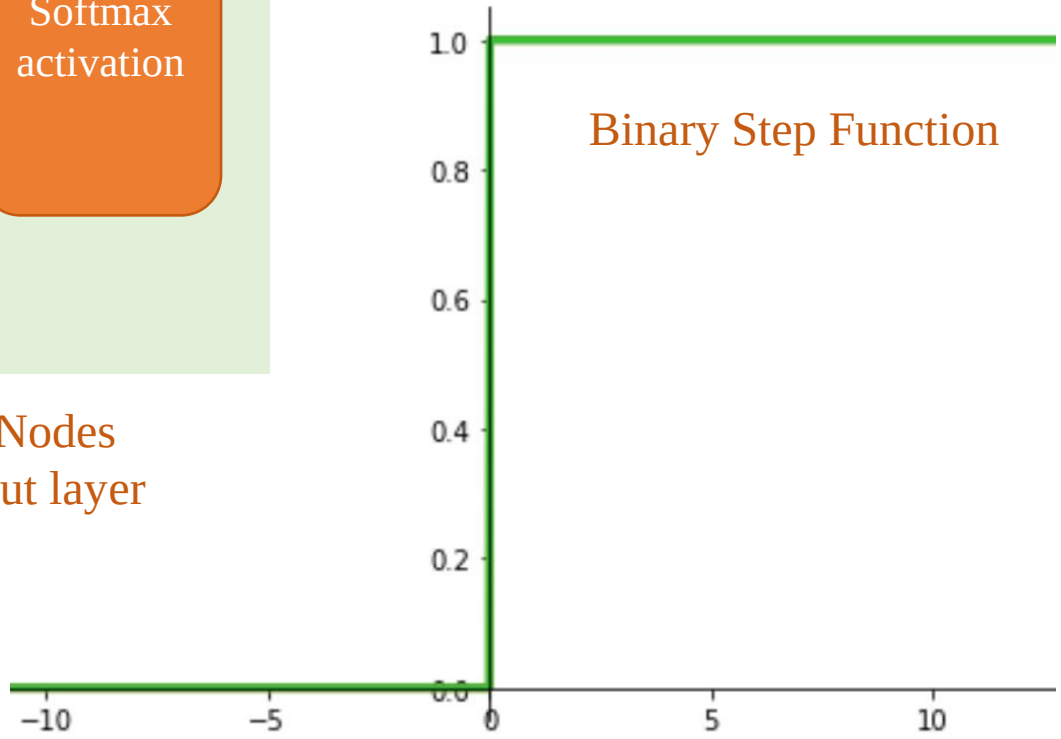


Input layer

10 Nodes
Output layer



Binary Step Function



$$f(x) = \begin{cases} 0 & if \ x < 0 \\ 1 & if \ x \geq 0 \end{cases}$$

https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/

Activation Functions

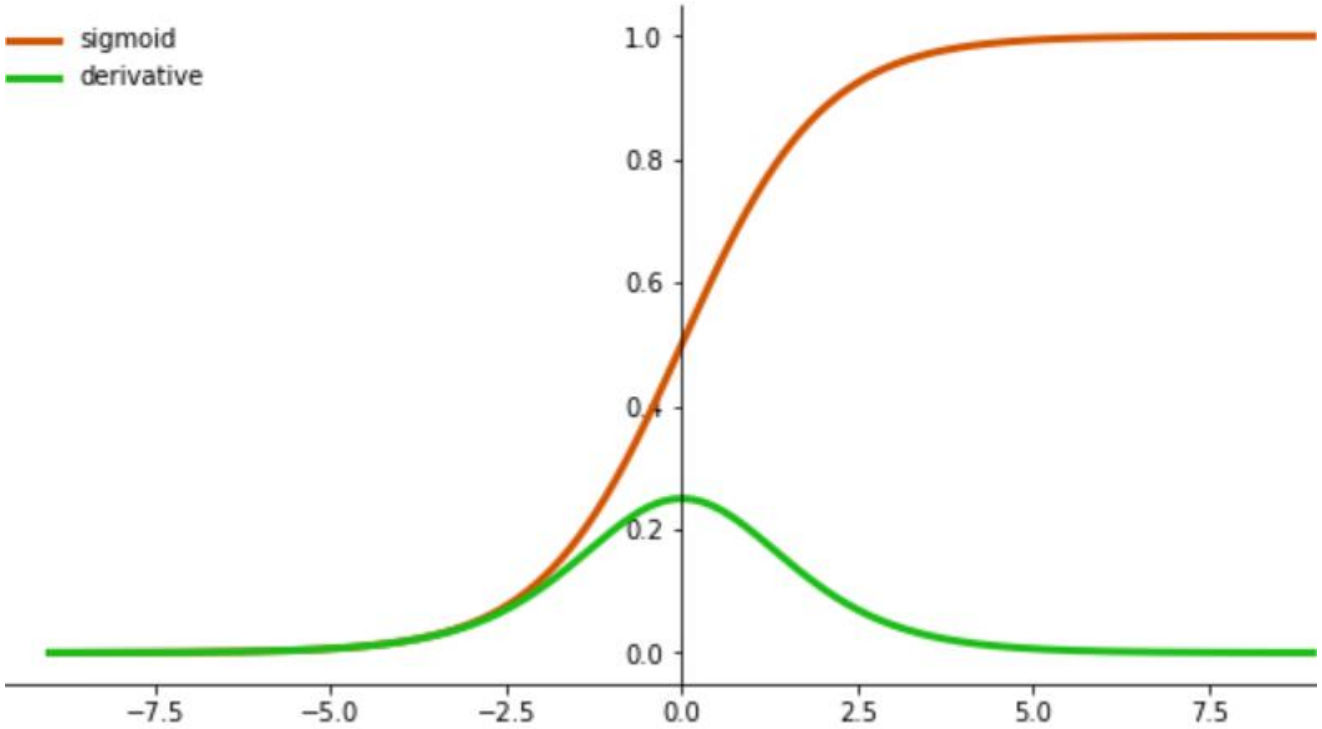
❖ Sigmoid function

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

data =	1	5	-4	3	-2

data_a = sigmoid(data)

data_a =	0.731	0.993	0.017	0.95	0.119



$$\text{sigmoid}'(x) = \text{sigmoid}(x)\,(1 - \text{sigmoid}(x))$$

Activation Functions

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

$$\text{sigmoid}'(x) = \left(\frac{1}{1 + e^{-x}}\right)' = \frac{-1}{(1 + e^{-x})^2}(-e^{-x})$$

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

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

$$= \frac{1}{1 + e^{-x}}\left(1 - \frac{1}{1 + e^{-x}}\right)$$

$$= \text{sigmoid}(x)\,(1 - \text{sigmoid}(x))$$

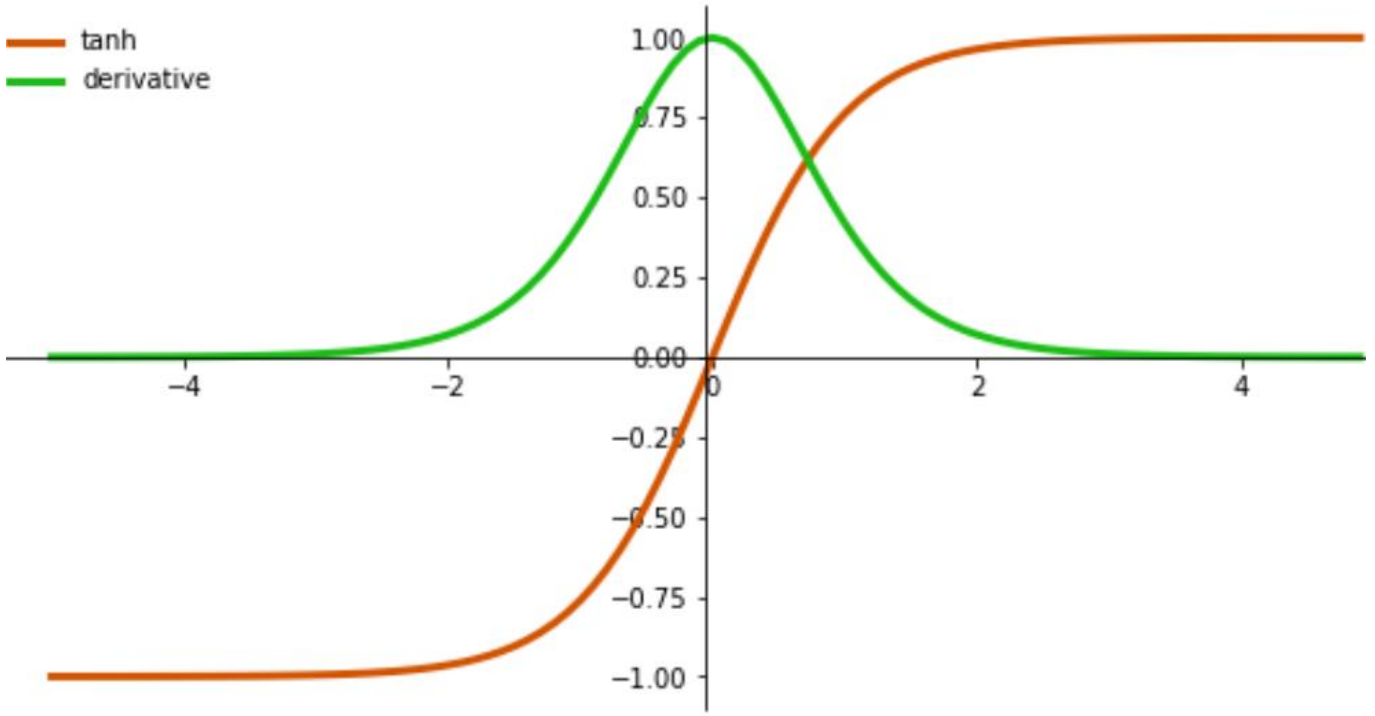
Activation Functions

❖ Tanh function

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$= \frac{2}{1 + e^{-2x}} - 1$$

$$= 1 - \frac{2}{e^{2x} + 1}$$



data =

1	5	-4	3	-2

data_a = tanh(data)

data_a =

0.761	0.999	-0.999	0.995	-0.964

$$tanh'(x) = 1 - tanh^{2}(x)$$

Activation Functions

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

$$\tanh'(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)' = \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

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

$$= 1 - \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)^2 = 1 - tanh^2(x)$$

Activation Functions

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

$$tanh'(x) = \left(\frac{2}{e^{-2x} + 1} - 1\right)' = \frac{4e^{-2x}}{(e^{-2x} + 1)^2} = 4\left(\frac{e^{-2x} + 1 - 1}{(e^{-2x} + 1)^2}\right)$$

$$= 4\left(\frac{1}{e^{-2x} + 1} - \frac{1}{(e^{-2x} + 1)^2}\right) = -\left(\frac{4}{(e^{-2x} + 1)^2} - \frac{4}{e^{-2x} + 1}\right)$$

$$= -\left(\frac{4}{(e^{-2x} + 1)^2} - \frac{4}{e^{-2x} + 1} + 1 - 1\right) = 1 - \left(\frac{2}{e^{-2x} + 1} - 1\right)^2 = 1 - tanh^2(x)$$

Activation Functions

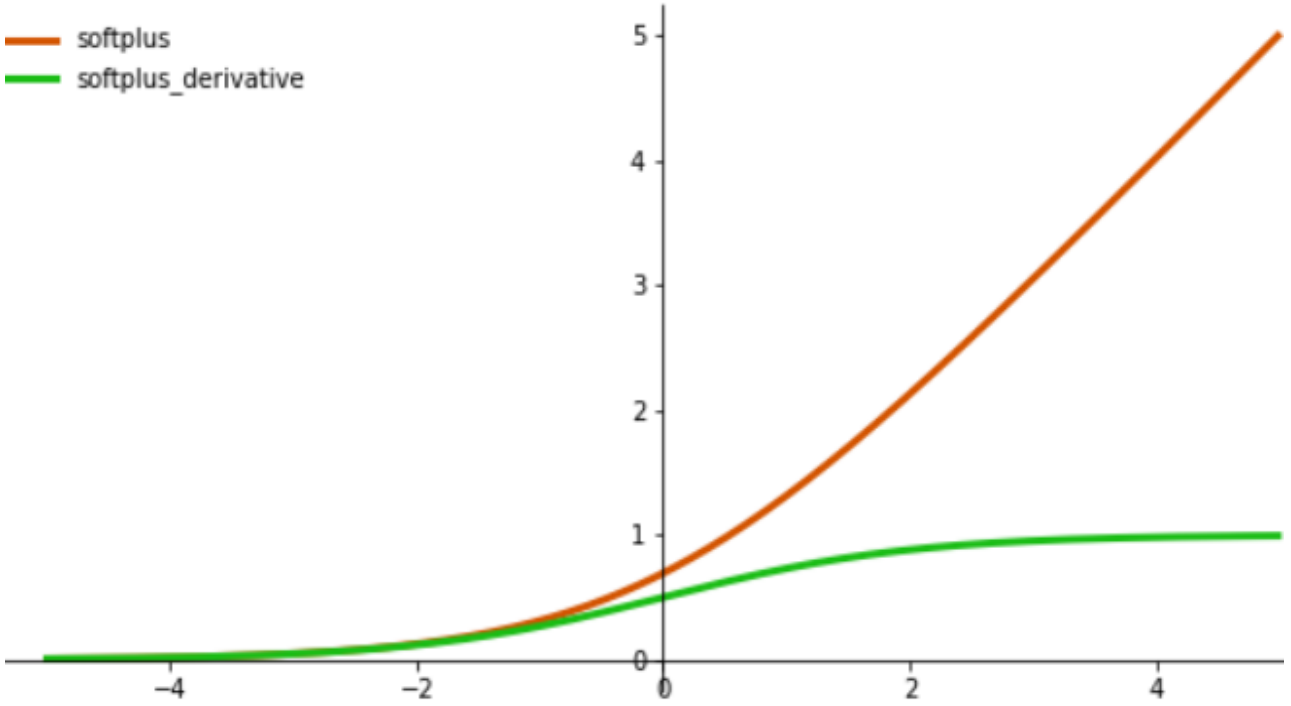
❖ Softplus function

$$\text{softplus}(x) = \log(1 + e^x)$$

data = | 1 | 5 | -4 | 3 | -2 |

data_a = softplus(data)

data_a = | 1.313 | 5.006 | 0.018 | 3.048 | 0.126 |



$$\text{softplus}'(x) = \frac{1}{1 + e^{-x}}$$

Activation Functions

❖ ReLU function

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

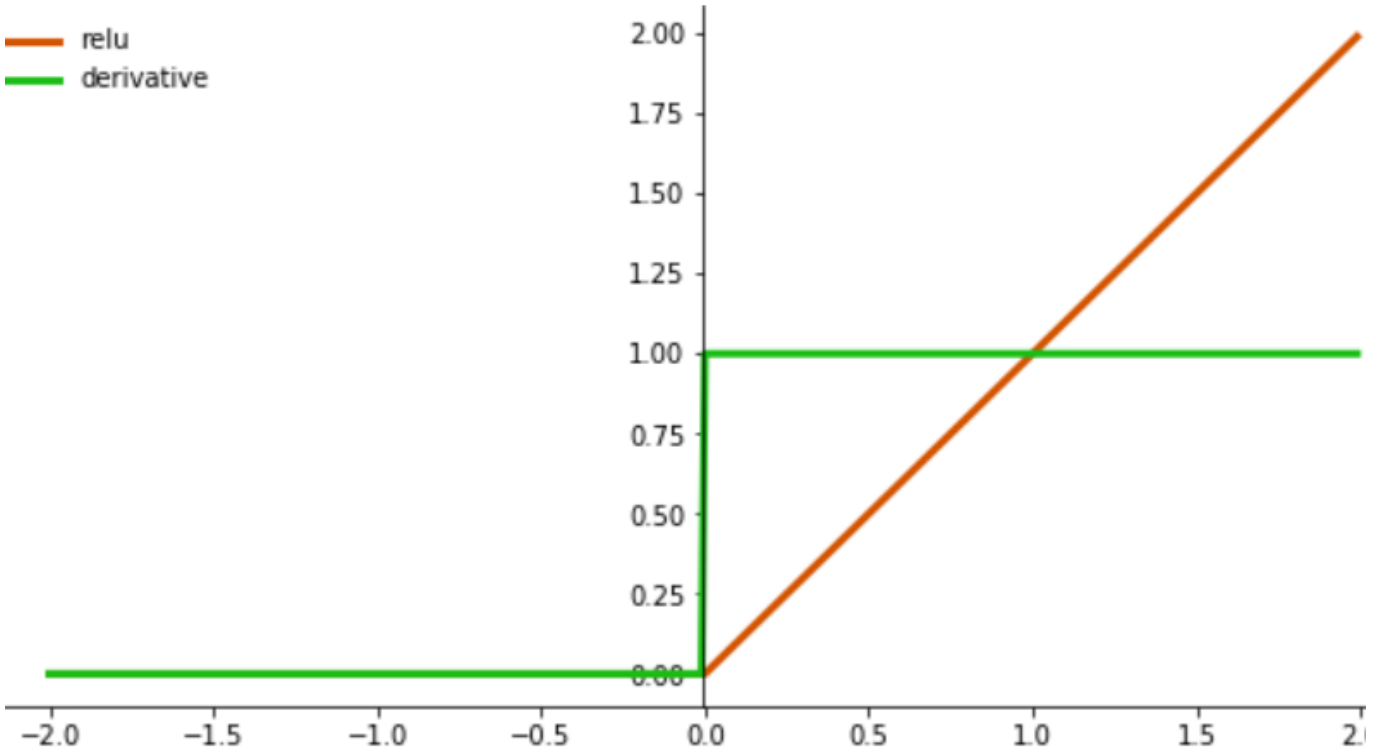
data =

1	5	-4	3	-2

data_a = ReLU(data)

data_a =

1	5	0	3	0

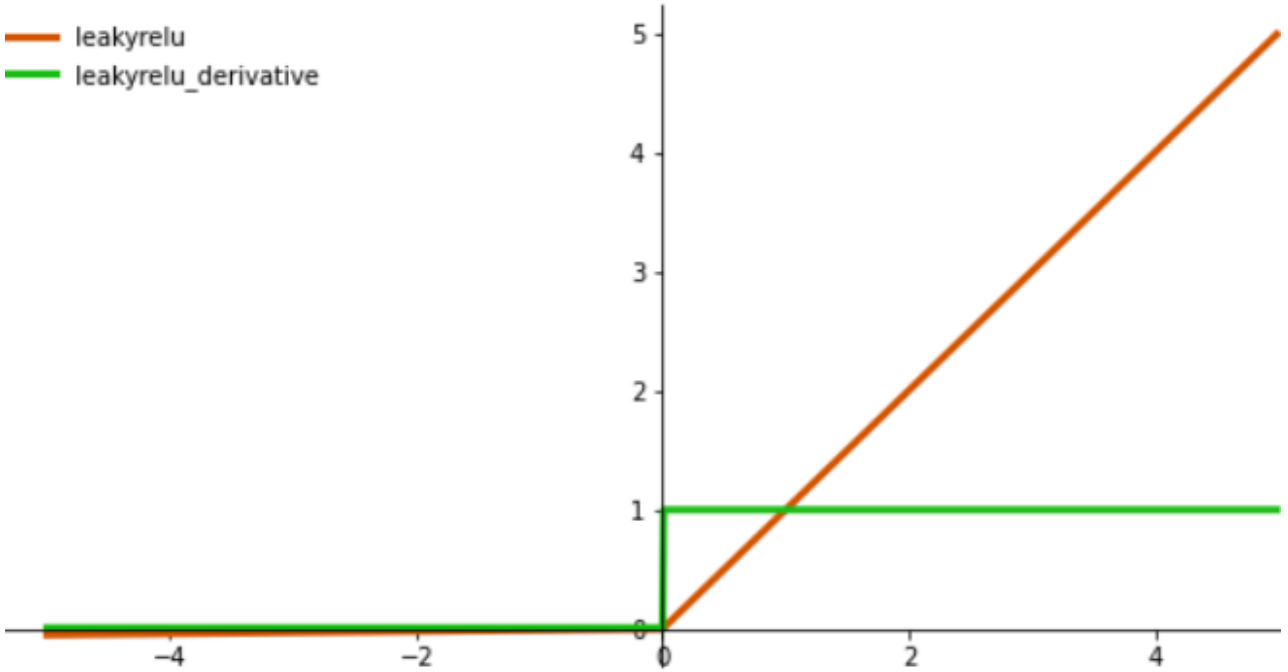


$$\text{ReLU}'(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ LeakyReLU function

$$\text{LeakyReLU}(x) = \begin{cases} 0.01x & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$



data =

1	5	-4	3	-2

data_a = leakyrelu(data)

data_a =

1	5	-0.04	3	-0.02

$$\text{LeakyReLU}'(x) = \begin{cases} 0.01 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ ELU function

$$\text{ELU}(x) = \begin{cases} \alpha(e^x - 1) & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

$\alpha = 0.1$

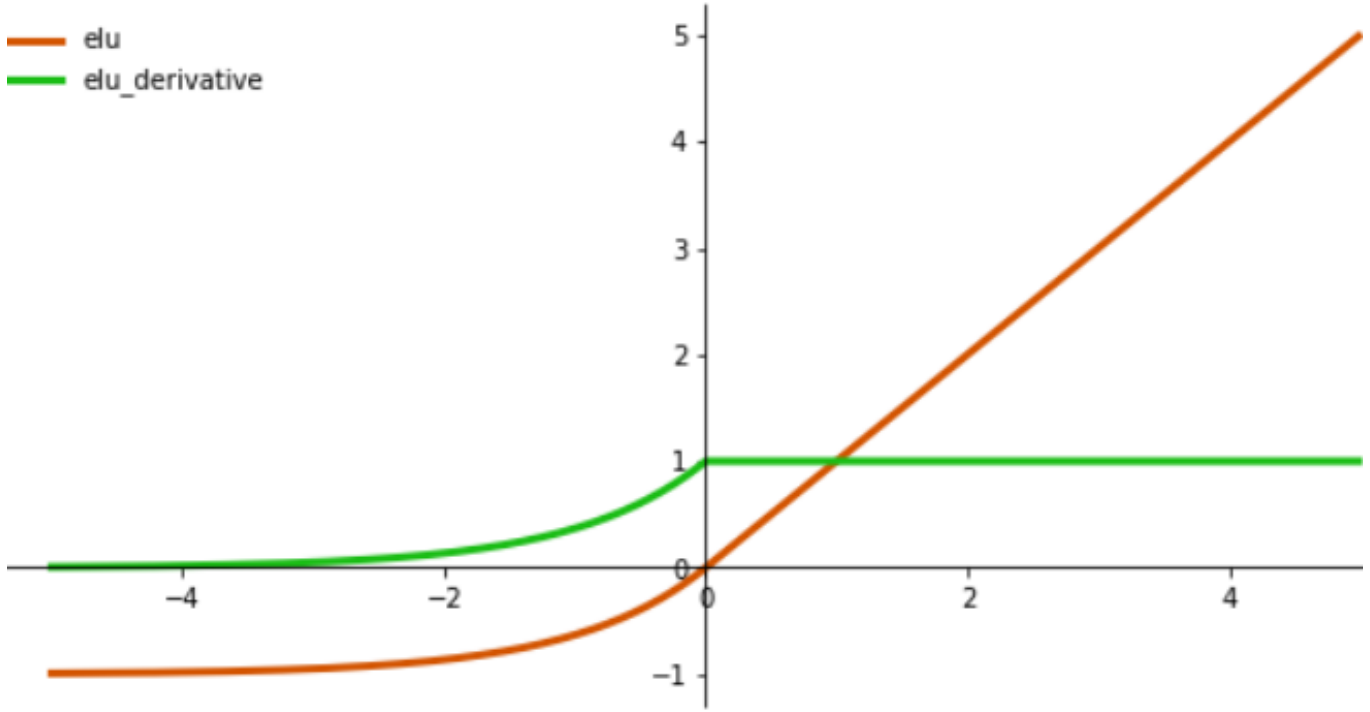
data =

1	5	-4	3	-2

data_a = ELU(data)

data_a =

1	5	-0.098	3	-0.086



$$\text{ELU}'(x) = \begin{cases} \alpha e^x & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ PReLU function

$$\text{PReLU}(x) = \begin{cases} \alpha x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

$\alpha = 0.1$

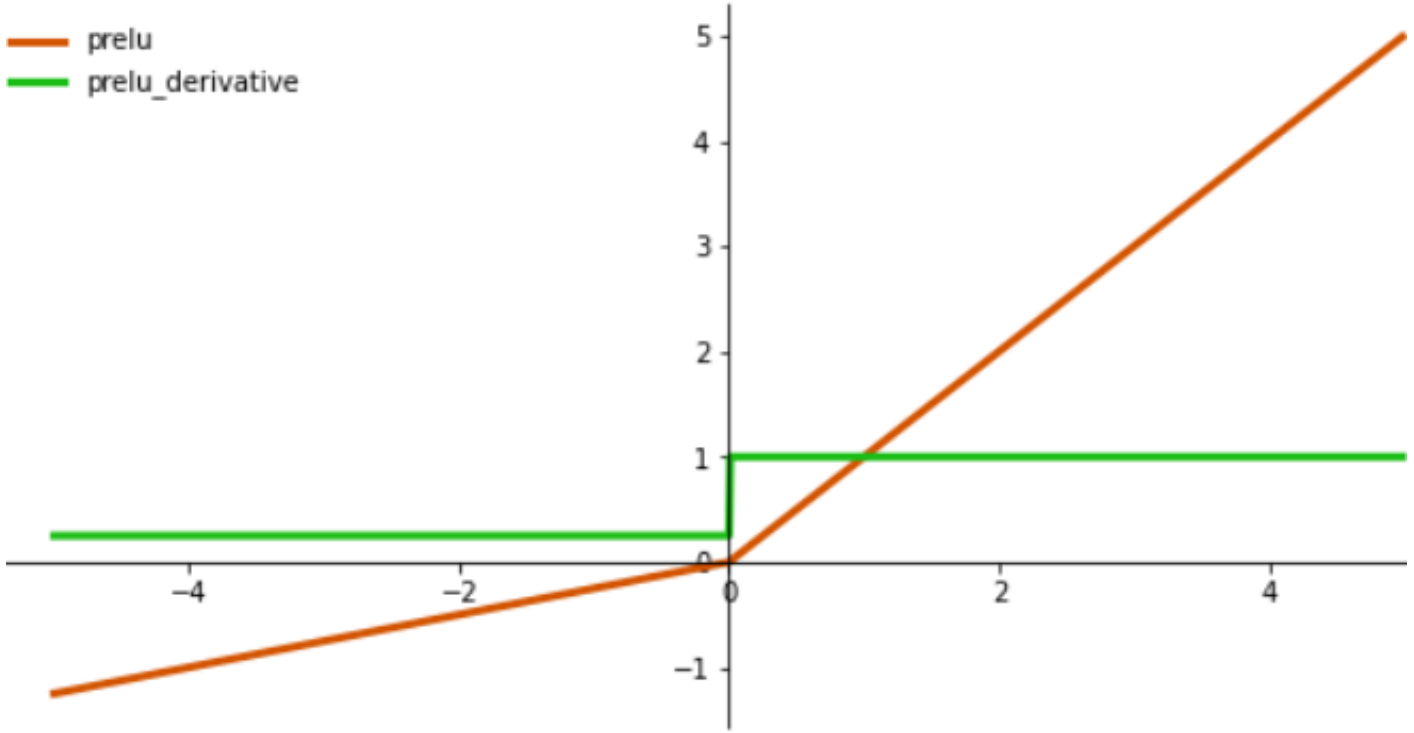
data =

1	5	-4	3	-2

data_a = PRELU(data)

data_a =

1	5	-0.4	3	-0.2



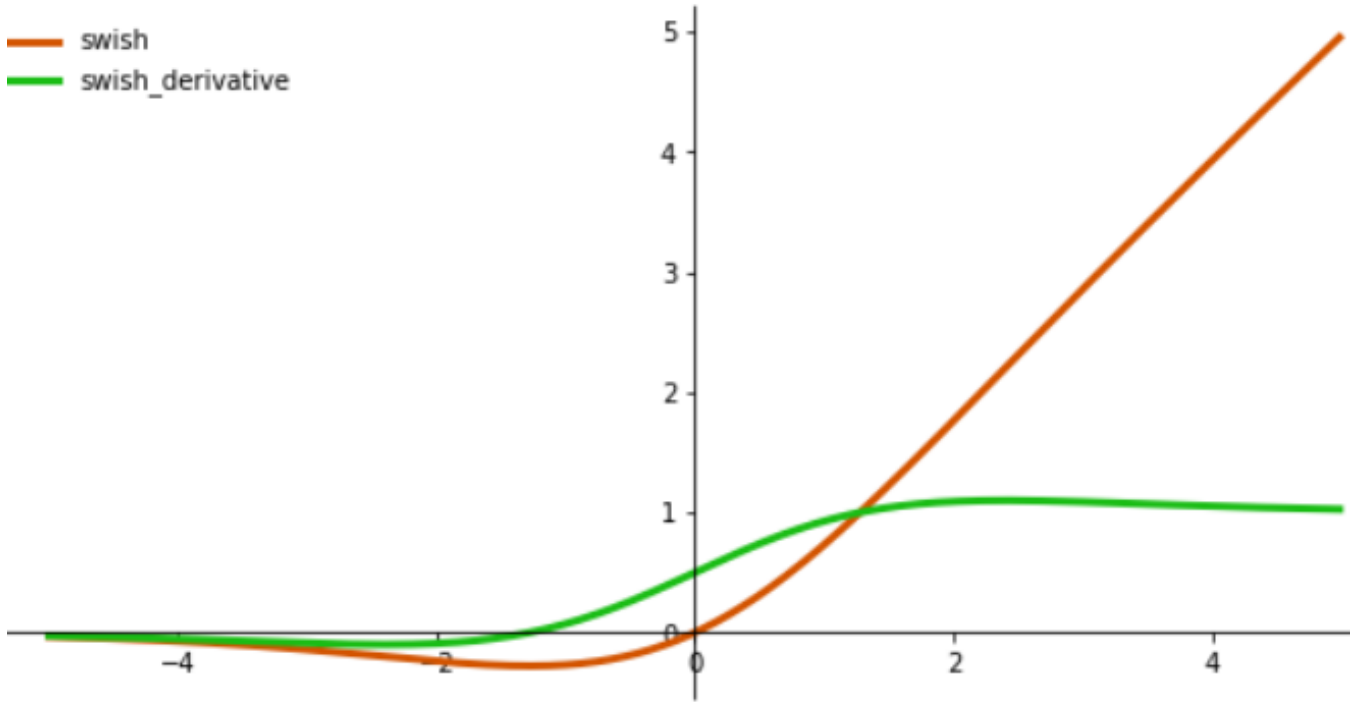
$$\text{PReLU}'(x) = \begin{cases} \alpha & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ Swish function

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

$$swish(x) = \frac{x}{1 + e^{-x}} = x\,\sigma(x)$$



data =

1	5	-4	3	-2

data_a = swish(data)

data_a =

0.731	4.966	-0.071	2.857	-0.238

$$swish'(x) = swish(x) + \sigma(x)\,(1 - swish(x))$$

Activation Functions

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

$$swish(x) = \frac{x}{1 + e^{-x}} = x\,\sigma(x)$$

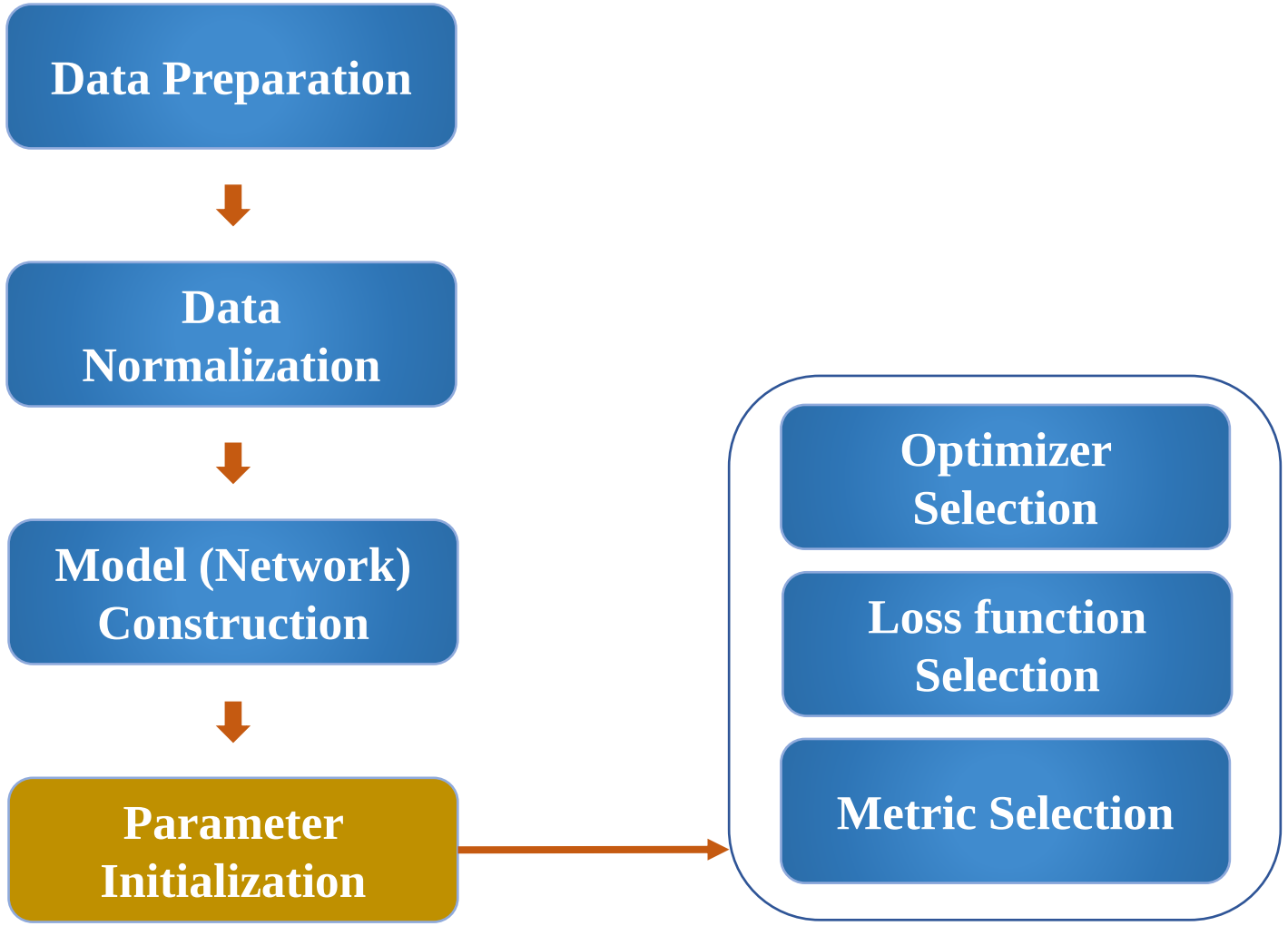
$$\begin{aligned} swish'(x) &= (x\,\sigma(x))' = (x)'\,\sigma(x) + x(\sigma(x))' \\ &= \sigma(x) + x\,\sigma(x)\,(1 - \sigma(x)) \\ &= \sigma(x) + x\,\sigma(x) - x\,\sigma(x)^2 \\ &= x\,\sigma(x) + \sigma(x)(1 - x\,\sigma(x)) \\ &= swish(x) + \sigma(x)\,(1 - swish(x)) \end{aligned}$$

Outline

- Pipeline Recommendation
- Data Normalization
- Activation Functions
- MLP Examples
- Initialization Methods

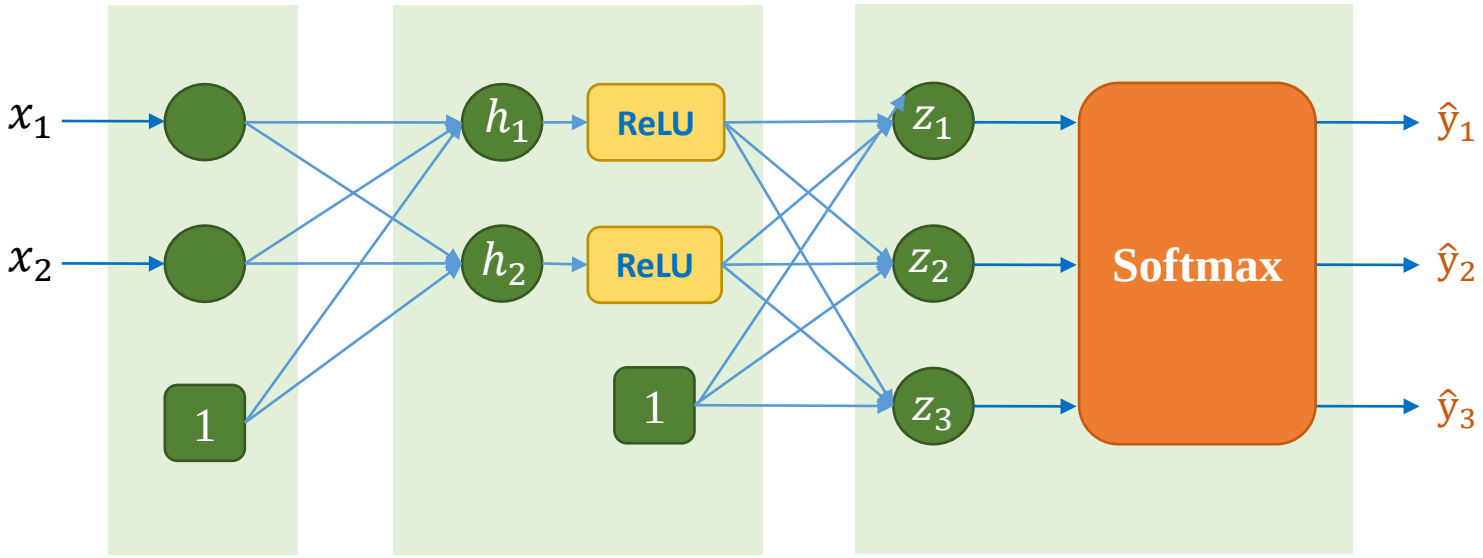
To-do List for Training

Train a model



MLP Example 1

Feature		Label
Petal Length	Petal Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

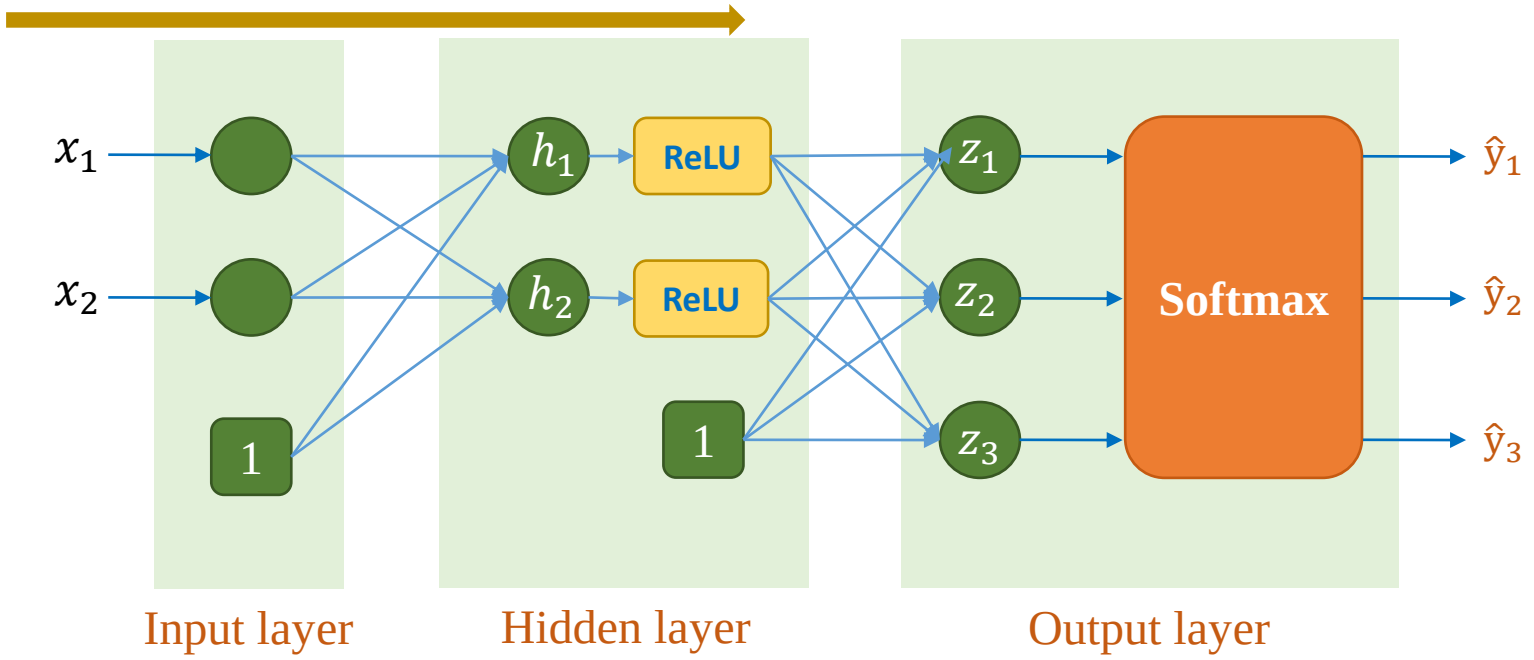
$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{h} = \boldsymbol{x}\boldsymbol{W}_h = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} = \begin{bmatrix} 1.373 & -1.696 \\ 4.708 & -5.951 \\ 5.731 & -7.281 \end{bmatrix}$$

$$\mathrm{ReLU}(\boldsymbol{h}) = \begin{bmatrix} 1.373 & 0 \\ 4.708 & 0 \\ 5.731 & 0 \end{bmatrix}$$

Feature		Label
Petal Length	Petal Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

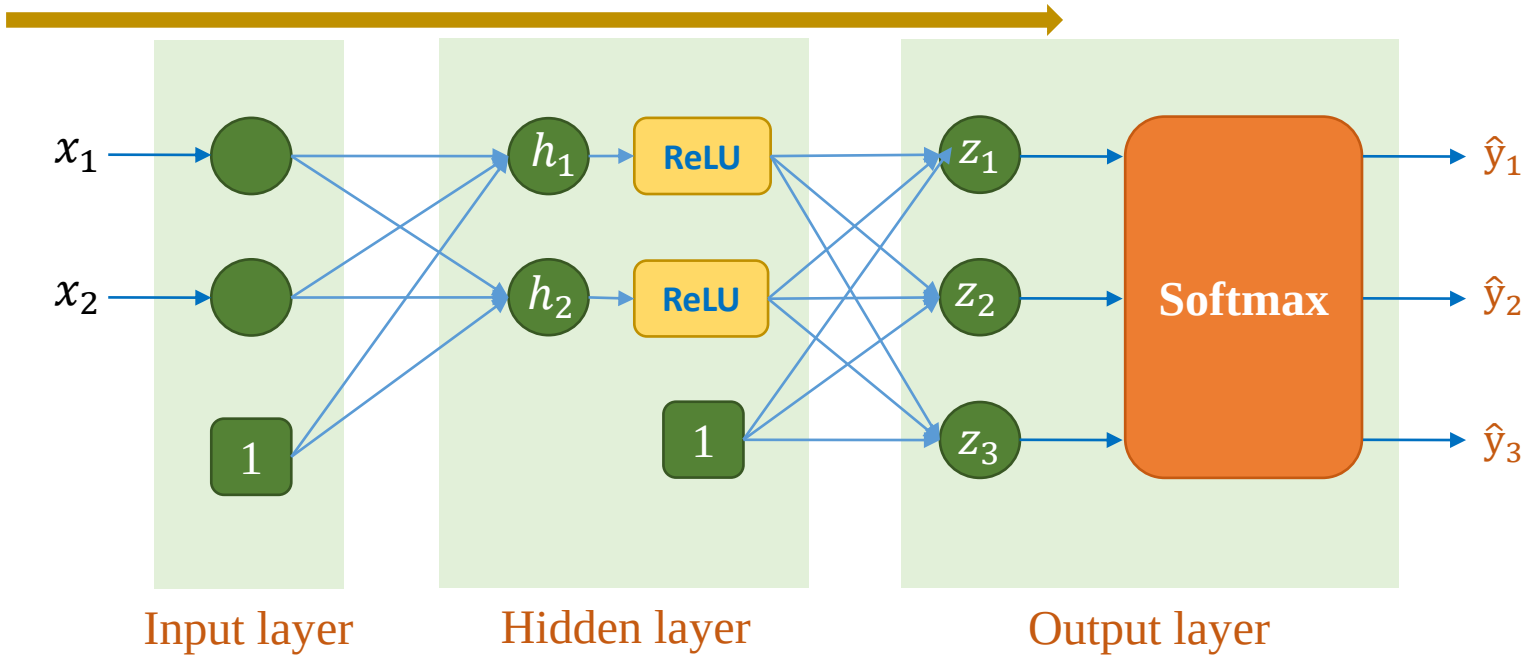
$$\text{ReLU}(\boldsymbol{h}) = \begin{bmatrix} 1.373 & 0 \\ 4.708 & 0 \\ 5.731 & 0 \end{bmatrix}$$

$$[\mathbf{1} \quad \text{ReLU}(\boldsymbol{h})] = \begin{bmatrix} 1 & 1.373 & 0 \\ 1 & 4.708 & 0 \\ 1 & 5.731 & 0 \end{bmatrix}$$

$$\boldsymbol{z} = [\mathbf{1} \quad \text{ReLU}(\boldsymbol{h})]\boldsymbol{W}_z = \begin{bmatrix} 1 & 1.373 & 0 \\ 1 & 4.708 & 0 \\ 1 & 5.731 & 0 \end{bmatrix} \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$= \begin{bmatrix} 0.439 & 0.356 & 0.195 \\ 1.507 & 1.220 & 0.670 \\ 1.835 & 1.485 & 0.816 \end{bmatrix}$$

Feature		Label
Petal Length	Petal Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

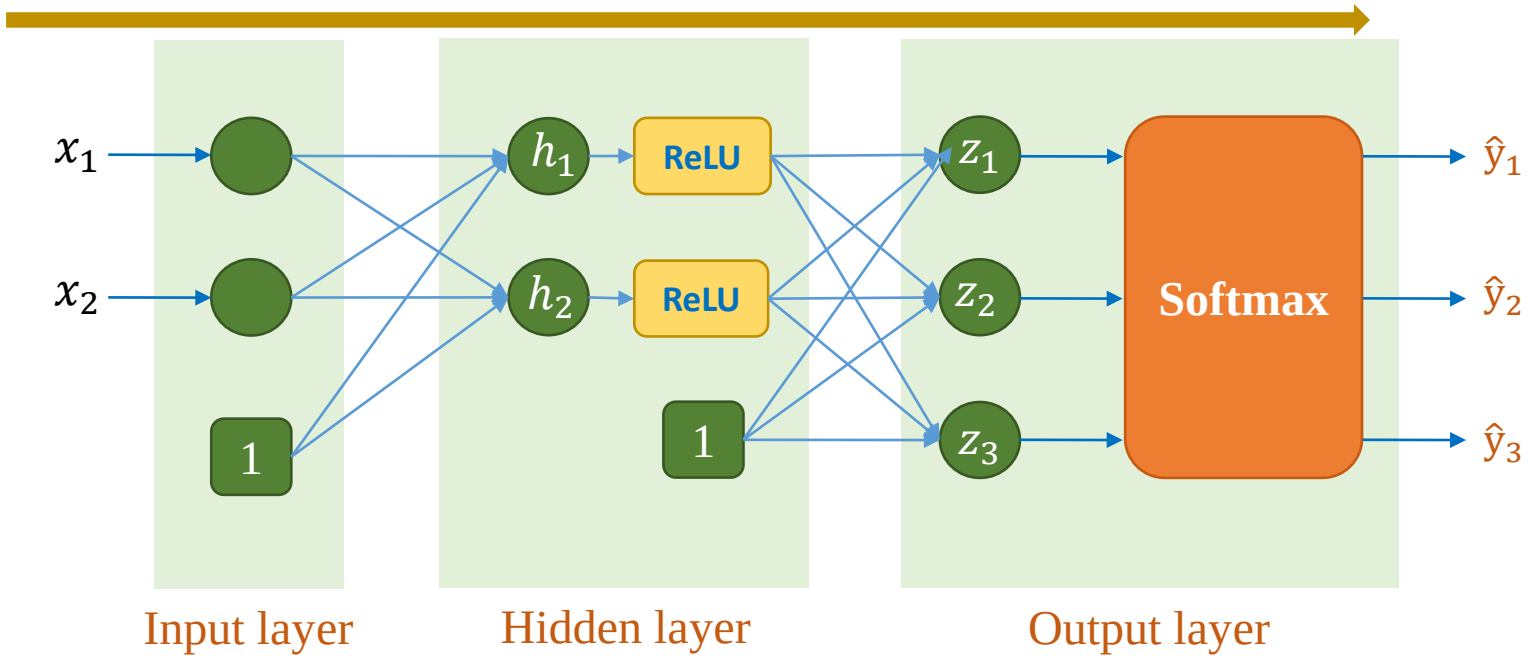
$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] \qquad \boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}]$$

$$= \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} \qquad = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$z = \begin{bmatrix} 0.439 & 0.356 & 0.195 \\ 1.507 & 1.220 & 0.670 \\ 1.835 & 1.485 & 0.816 \end{bmatrix} \qquad \hat{y} = \text{softmax}(z) = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \begin{bmatrix} 0.369 & 0.340 & 0.289 \\ 0.458 & 0.343 & 0.198 \\ 0.484 & 0.341 & 0.174 \end{bmatrix}$$

$$\text{loss} = 1.269$$

Feature		Label
Petal Length	Petal Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



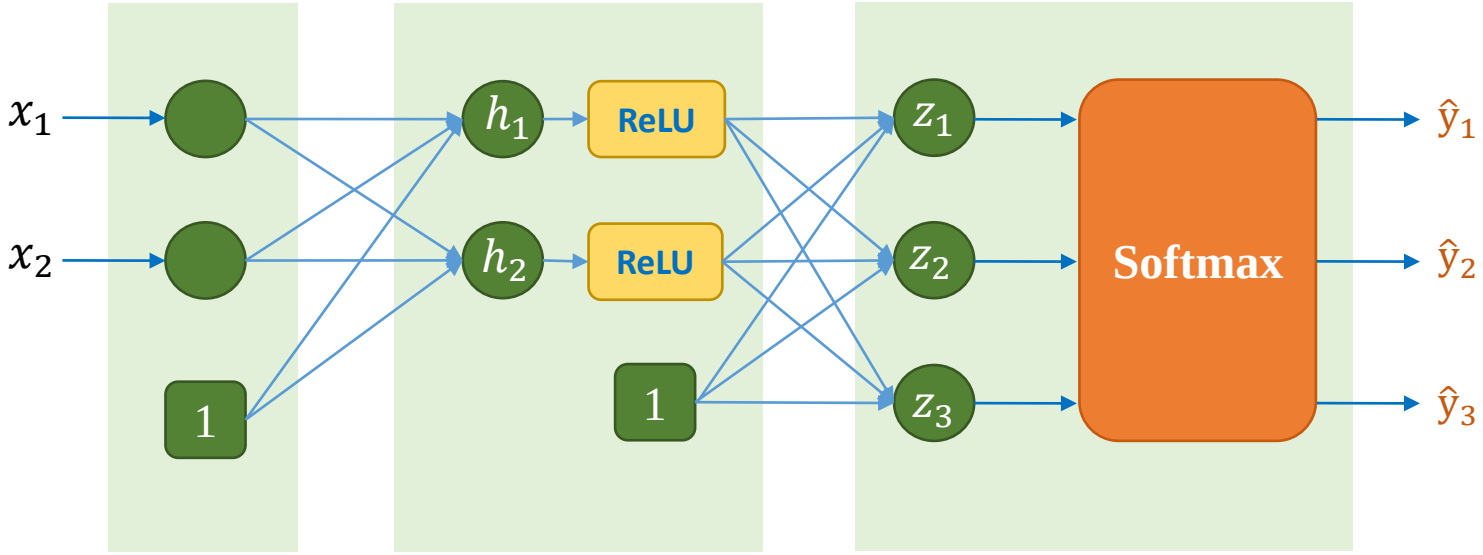
$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \qquad y = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$W_h = [W_{h1} \quad W_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$W_z = [W_{z1} \quad W_{z2} \quad W_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

Example 2 - Dying ReLU

Feature		Label
Petal Length	Petal Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2

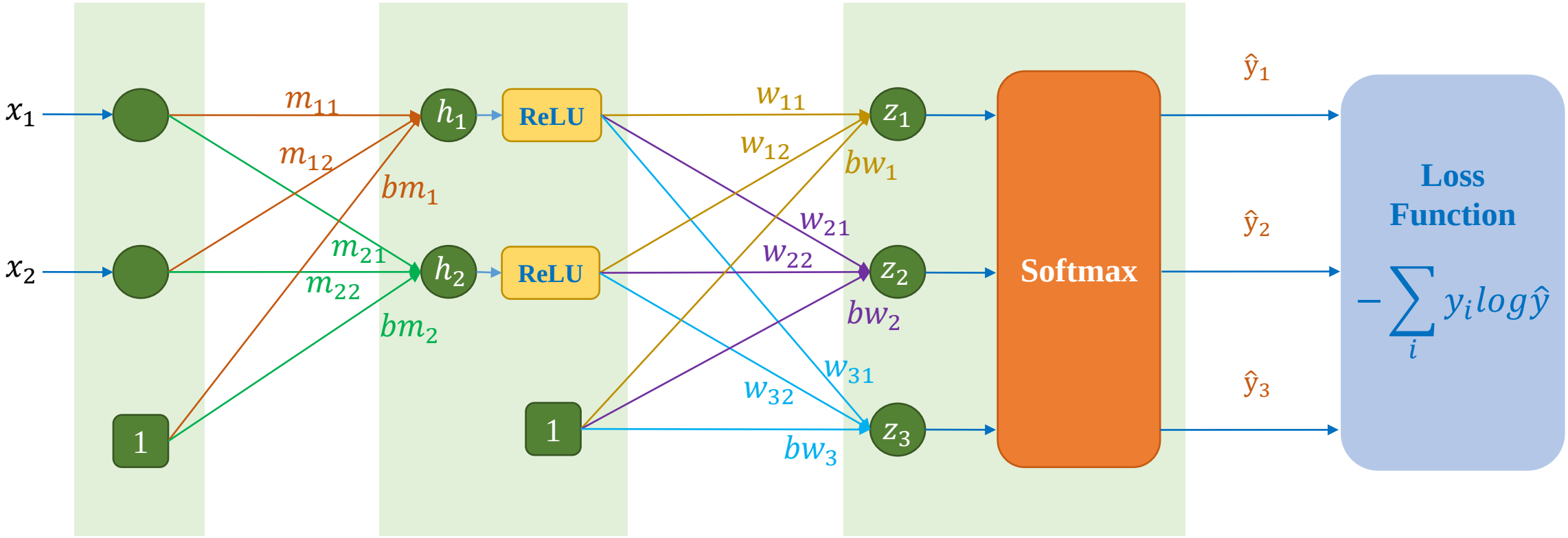


$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix} \qquad y = 0$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \qquad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$



$y = 0$

$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\rightarrow \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

Forward pass

$$\boldsymbol{h} = \begin{bmatrix} 1.372 \\ -1.68 \end{bmatrix}$$

zero value

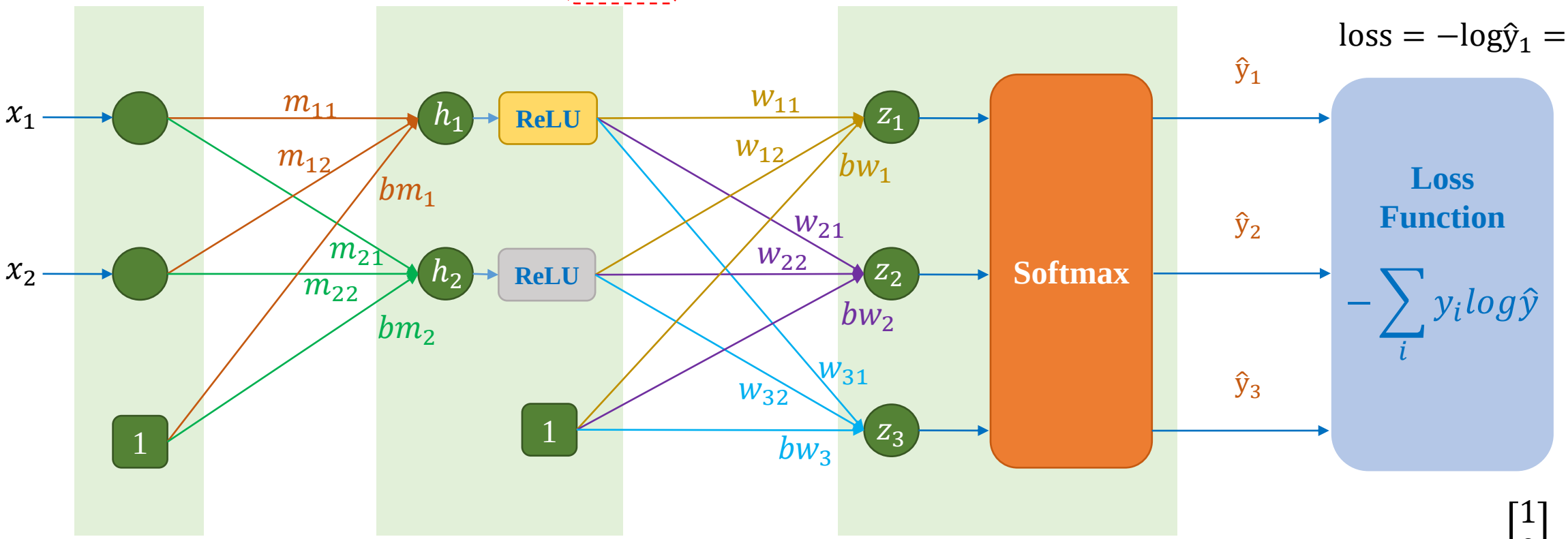
$$\boldsymbol{z} = \begin{bmatrix} 0.439 \\ 0.343 \\ 0.192 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \begin{bmatrix} 0.372 \\ 0.338 \\ 0.290 \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix}$$

$$\textbf{ReLU} = \begin{bmatrix} 1.372 \\ 0.0 \end{bmatrix}$$

$$\text{loss} = -\log\hat{y}_1 = 0.989$$



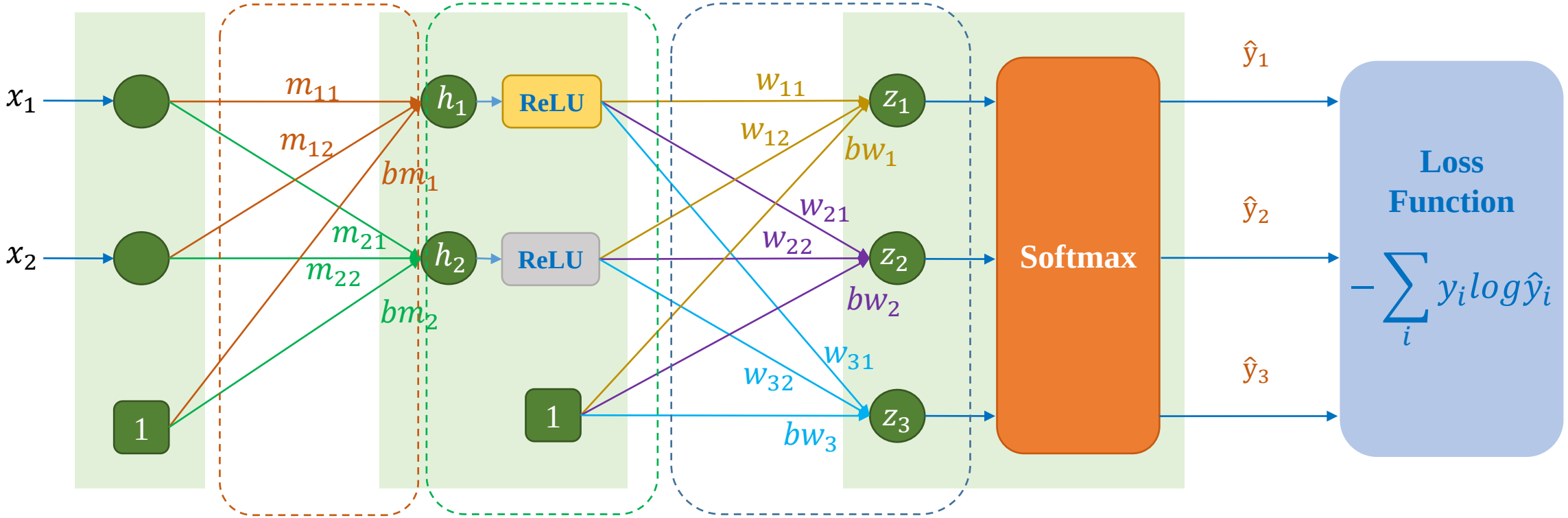
$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$



Backward pass

$$\frac{\partial L}{\partial m_{jk}} = x_k \frac{\partial L}{\partial h_j}$$

$$\frac{\partial L}{\partial bm_j} = \frac{\partial L}{\partial h_j}$$

$$\frac{\partial L}{\partial relu_j} = \sum_i w_{ij} \frac{\partial L}{\partial z_i}$$

$$\text{ReLU}'(h_j) = \begin{cases} 0 & \text{if } h_j \le 0 \\ 1 & \text{if } h_j > 0 \end{cases}$$

$$\frac{\partial L}{\partial h_j} = \begin{cases} 0 & \text{if } h_j \le 0 \\ \frac{\partial L}{\partial relu_j} & \text{if } h_j > 0 \end{cases}$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

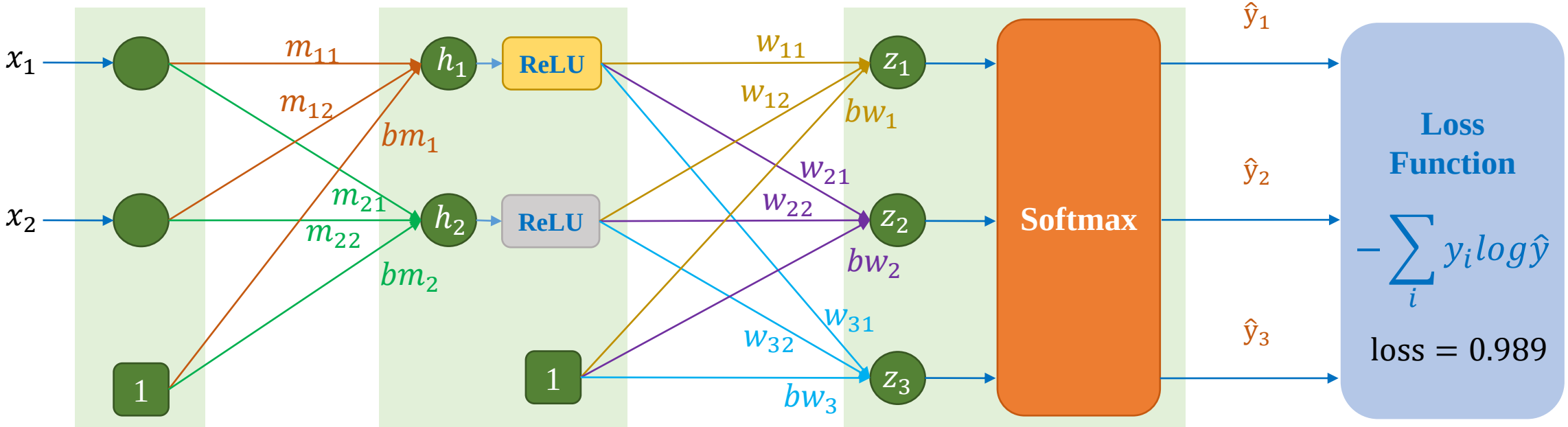
$$\frac{\partial L}{\partial w_{ij}} = x_j \frac{\partial L}{\partial z_i}$$

$$\frac{\partial L}{\partial bw_i} = \frac{\partial L}{\partial z_i}$$

Backward pass

$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix} \quad \boldsymbol{h} = \begin{bmatrix} 1.372 \\ -1.68 \end{bmatrix} \quad \textbf{ReLU} = \begin{bmatrix} 1.372 \\ 0.0 \end{bmatrix} \quad \boldsymbol{z} = \begin{bmatrix} 0.439 \\ 0.343 \\ 0.192 \end{bmatrix} \quad \hat{\boldsymbol{y}} = \begin{bmatrix} 0.372 \\ 0.338 \\ 0.290 \end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$



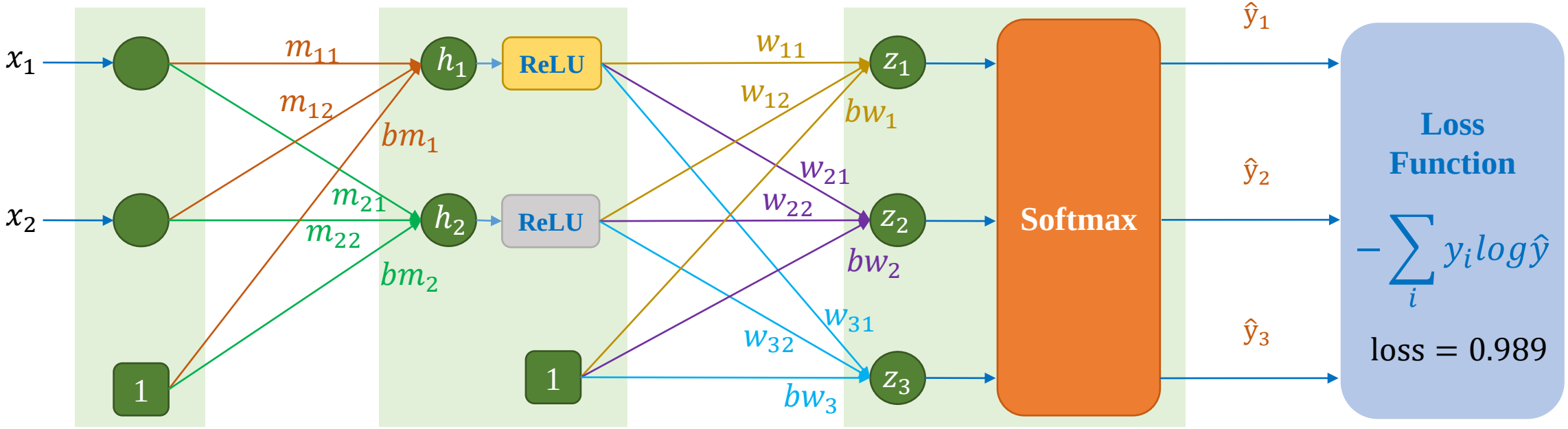
$$\frac{\partial L}{\partial relu_j} = \sum_i w_{ij} \frac{\partial L}{\partial z_i} \qquad \frac{\partial L}{\partial w_{ij}} = x_j \frac{\partial L}{\partial z_i} \qquad \frac{\partial L}{\partial bw_i} = \frac{\partial L}{\partial z_i} \qquad \frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\nabla_{\textbf{ReLU}} L = \begin{bmatrix} -0.0759 \\ -0.0445 \end{bmatrix} \quad \nabla_{\boldsymbol{w}} L = \begin{bmatrix} -0.628 & 0.338 & 0.29 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{bw}} L = \begin{bmatrix} -0.628 \\ 0.338 \\ 0.290 \end{bmatrix} \quad \nabla_{\boldsymbol{z}} L = \begin{bmatrix} -0.628 \\ 0.338 \\ 0.290 \end{bmatrix}$$

Backward pass

$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix} \quad \boldsymbol{h} = \begin{bmatrix} 1.372 \\ -1.68 \end{bmatrix} \quad \textbf{ReLU} = \begin{bmatrix} 1.372 \\ 0.0 \end{bmatrix} \quad \boldsymbol{z} = \begin{bmatrix} 0.439 \\ 0.343 \\ 0.192 \end{bmatrix} \quad \hat{\boldsymbol{y}} = \begin{bmatrix} 0.372 \\ 0.338 \\ 0.290 \end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$



$$\frac{\partial L}{\partial m_{jk}} = x_k \frac{\partial L}{\partial h_j}$$

$$\nabla_m L = \begin{bmatrix} -0.114 & 0.0 \\ -0.015 & 0.0 \end{bmatrix}$$

$$\frac{\partial L}{\partial bm_j} = \frac{\partial L}{\partial h_j}$$

$$\nabla_{bm} L = \begin{bmatrix} -0.0759 \\ 0.0 \end{bmatrix}$$

$$\frac{\partial L}{\partial h_j} = \begin{cases} 0 & \text{if } h_j \leq 0 \\ \dfrac{\partial L}{\partial relu_j} & \text{if } h_j > 0 \end{cases}$$

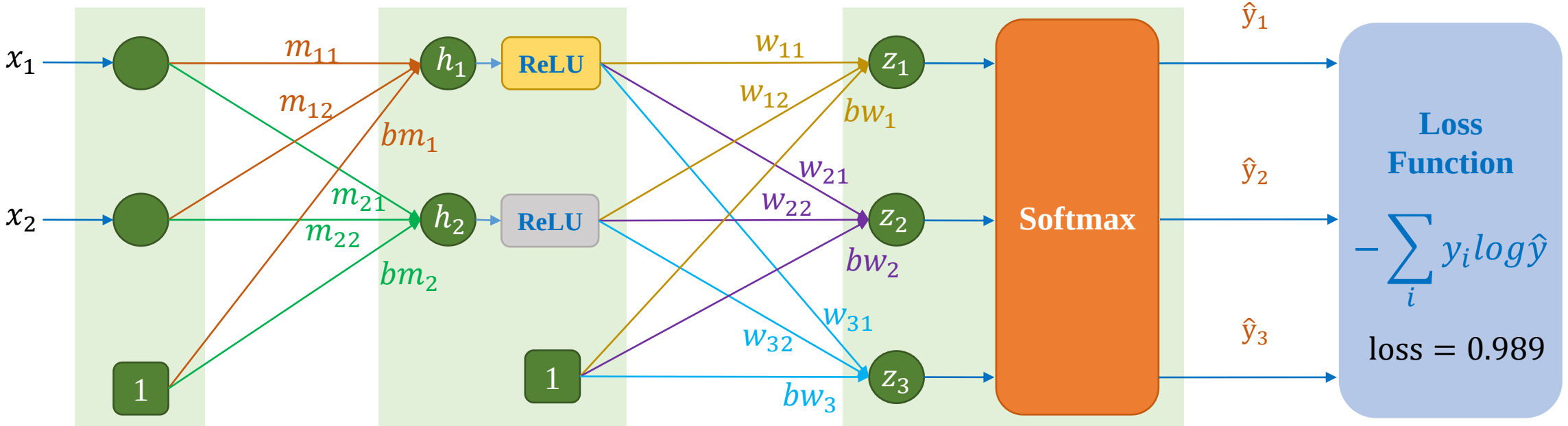
$$\nabla_h L = \begin{bmatrix} -0.0759 \\ 0.0 \end{bmatrix}$$

$$\frac{\partial L}{\partial relu_j} = \sum_i w_{ij} \frac{\partial L}{\partial z_i}$$

$$\nabla_{\text{ReLU}} L = \begin{bmatrix} -0.0759 \\ -0.0445 \end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\nabla_{\boldsymbol{m}} L = \begin{bmatrix} -0.114 & 0.0 \\ -0.015 & 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{bm}} L = \begin{bmatrix} -0.0759 \\ 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{w}} L = \begin{bmatrix} -0.628 & 0.338 & 0.29 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{bw}} L = \begin{bmatrix} -0.628 \\ 0.338 \\ 0.290 \end{bmatrix}$$



Update the parameters with $\eta = 0.01$

$$\boldsymbol{m} = \begin{bmatrix} 0.861 & -1.04 \\ 0.4105 & -0.65 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.000759 \\ 0.0 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} 0.328 & 0.245 & 0.136 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} 0.0062 \\ -0.0033 \\ -0.0029 \end{bmatrix}$$

Forward pass again

$$\boldsymbol{h} = \begin{bmatrix} 1.374 \\ -1.68 \end{bmatrix}$$

still zero value

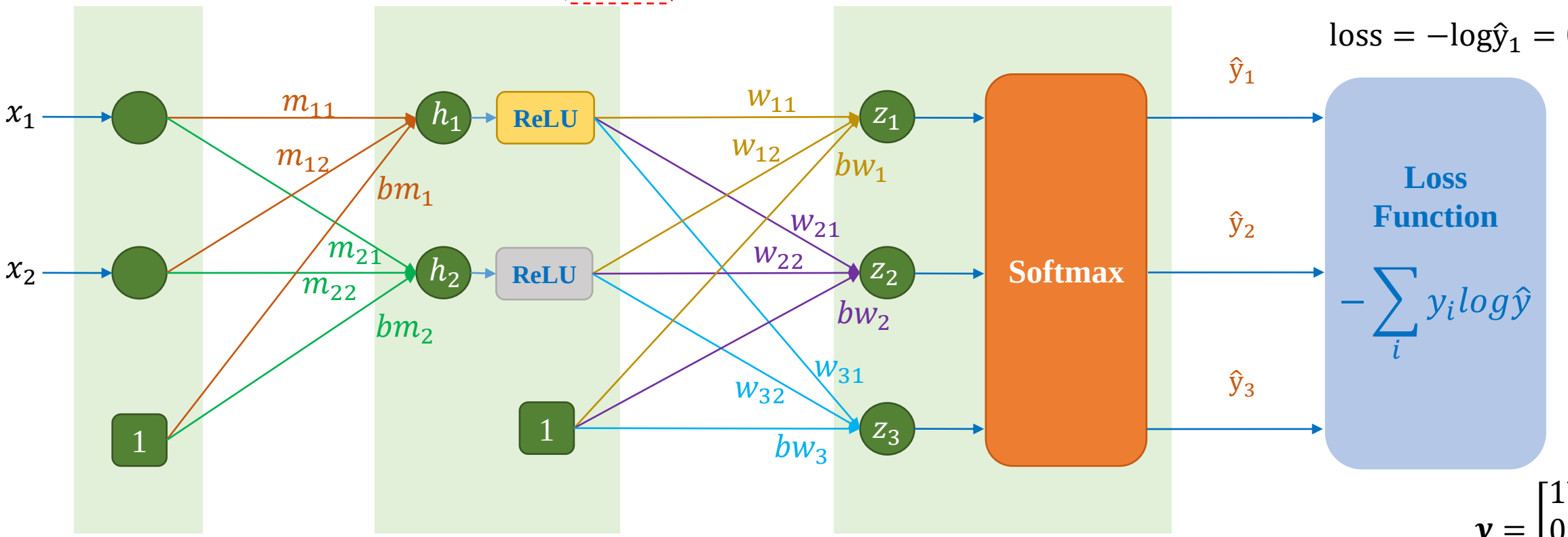
$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix}$$

$$\textbf{ReLU} = \begin{bmatrix} 1.374 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{z} = \begin{bmatrix} 0.458 \\ 0.334 \\ 0.184 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \begin{bmatrix} 0.378 \\ 0.334 \\ 0.287 \end{bmatrix}$$

$$\text{loss} = -\log\hat{y}_1 = 0.972$$



$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.861 & -1.04 \\ 0.4105 & -0.65 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.000759 \\ 0.0 \end{bmatrix}$$

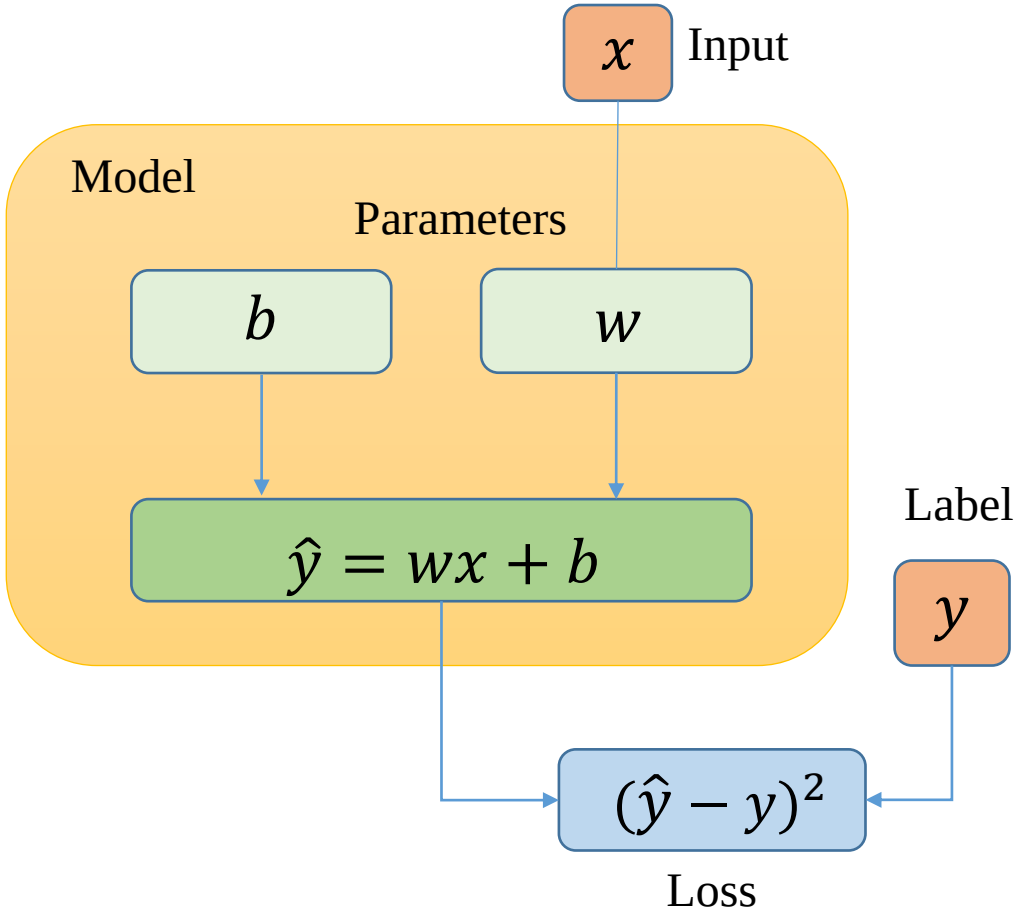
$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.328 & 0.245 & 0.136 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{bw} = \begin{bmatrix} 0.0062 \\ -0.0033 \\ -0.0029 \end{bmatrix}$$

Example 3 - Zero Initialization

❖ Linear regression

Diagram



Cheat sheet

Compute the output $\hat{y}$

$$\hat{y} = wx + b$$

Compute the loss

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

Compute derivative

$$L'_w = 2x(\hat{y} - y)$$

$$L'_b = 2(\hat{y} - y)$$

Update parameters

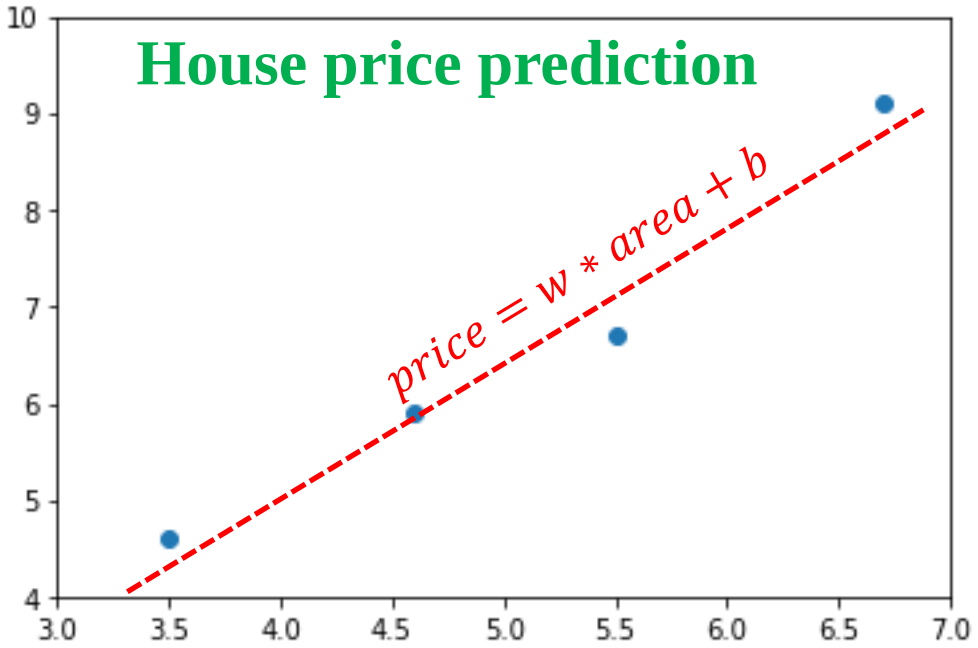
$$w = w - \eta L'_w$$

$$b = b - \eta L'_b$$

Example 3 - Zero Initialization

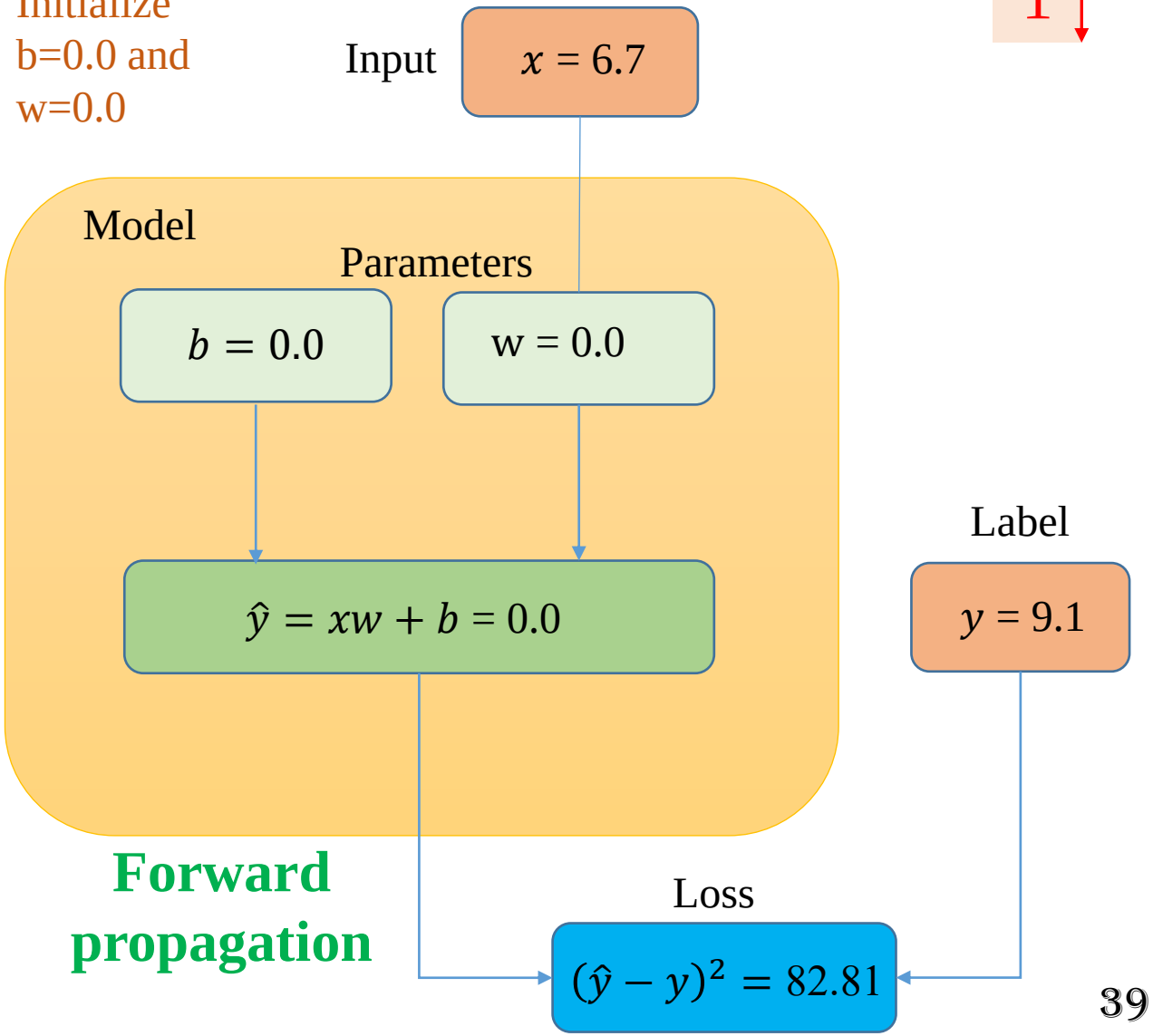
Given sample data

Feature	Label
area	price
6.7	9.1
4.6	5.9
3.5	4.6
5.5	6.7

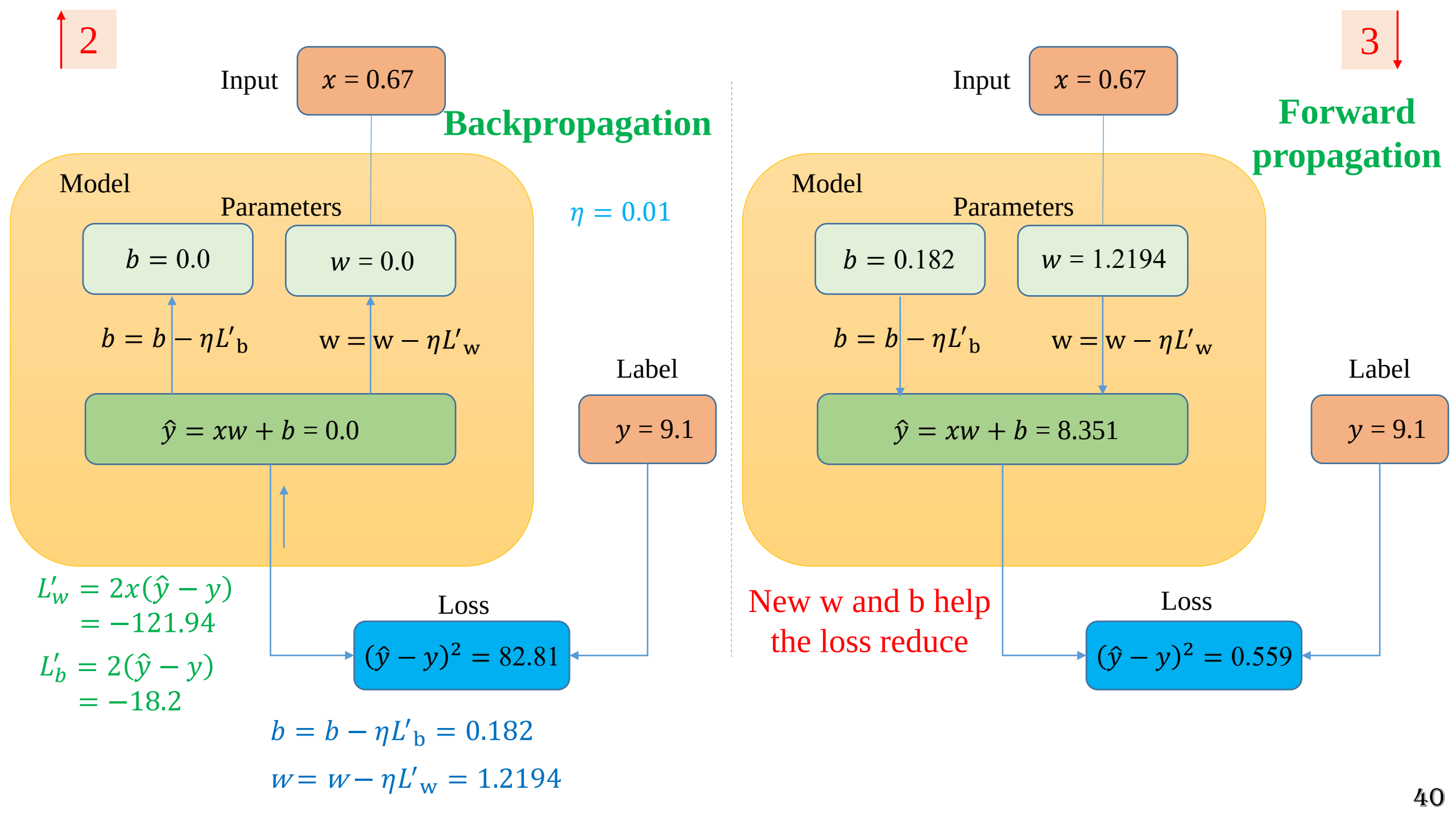


Initialize b=0.0 and w=0.0

1



Forward propagation


2
Input
$x = 0.67$
Backpropagation
Model
Parameters
$b = 0.0$
$w = 0.0$
$\eta = 0.01$
$b = b - \eta L'_b$
$w = w - \eta L'_w$
Label
$y = 9.1$
$\hat{y} = xw + b = 0.0$
$L'_w = 2x(\hat{y} - y)$
$= -121.94$
$L'_b = 2(\hat{y} - y)$
$= -18.2$
Loss
$(\hat{y} - y)^2 = 82.81$
$b = b - \eta L'_b = 0.182$
$w = w - \eta L'_w = 1.2194$
3
Input
$x = 0.67$
Forward propagation
Model
Parameters
$b = 0.182$
$w = 1.2194$
$b = b - \eta L'_b$
$w = w - \eta L'_w$
Label
$y = 9.1$
$\hat{y} = xw + b = 8.351$
New w and b help the loss reduce
Loss
$(\hat{y} - y)^2 = 0.559$

Example 4 - Zero Initialization

❖ Logistic regression

$$\boldsymbol{\theta}^T = [b \quad w_1 \quad w_2]$$

$$\boldsymbol{x}^T = [1 \quad x_1 \quad x_2]$$

1) Pick a sample (x, y) from training data

2) Compute output $\hat{y}$

$$z = \boldsymbol{\theta}^T \boldsymbol{x}$$

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

3) Compute loss

$$L(\boldsymbol{\theta}) = (-y\log\hat{y} - (1-y)\log(1-\hat{y}))$$

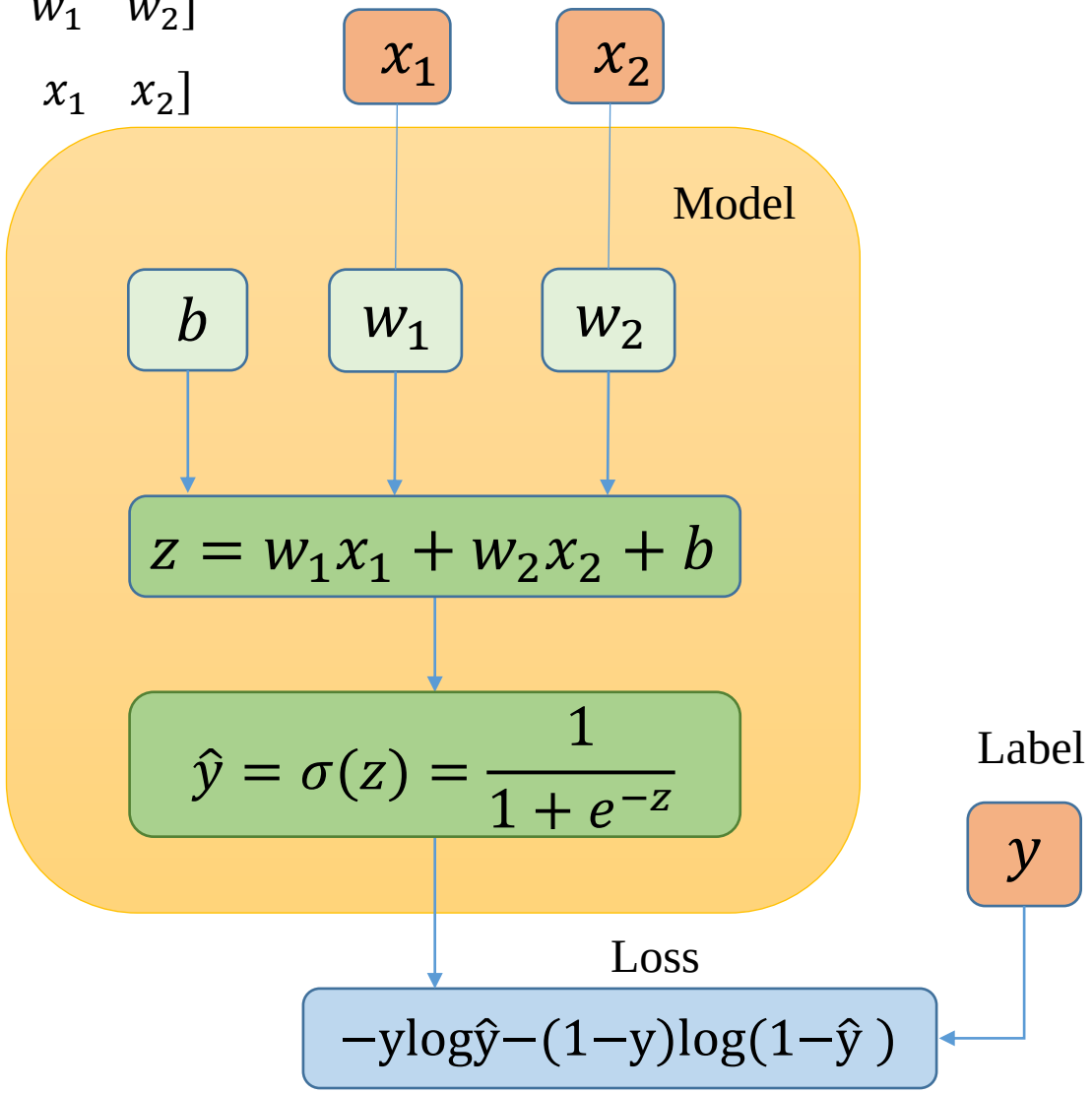
4) Compute derivative

$$\nabla_{\boldsymbol{\theta}} L = \mathbf{x}(\hat{y} - y)$$

5) Update parameters

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta L'_{\boldsymbol{\theta}}$$

η is learning rate

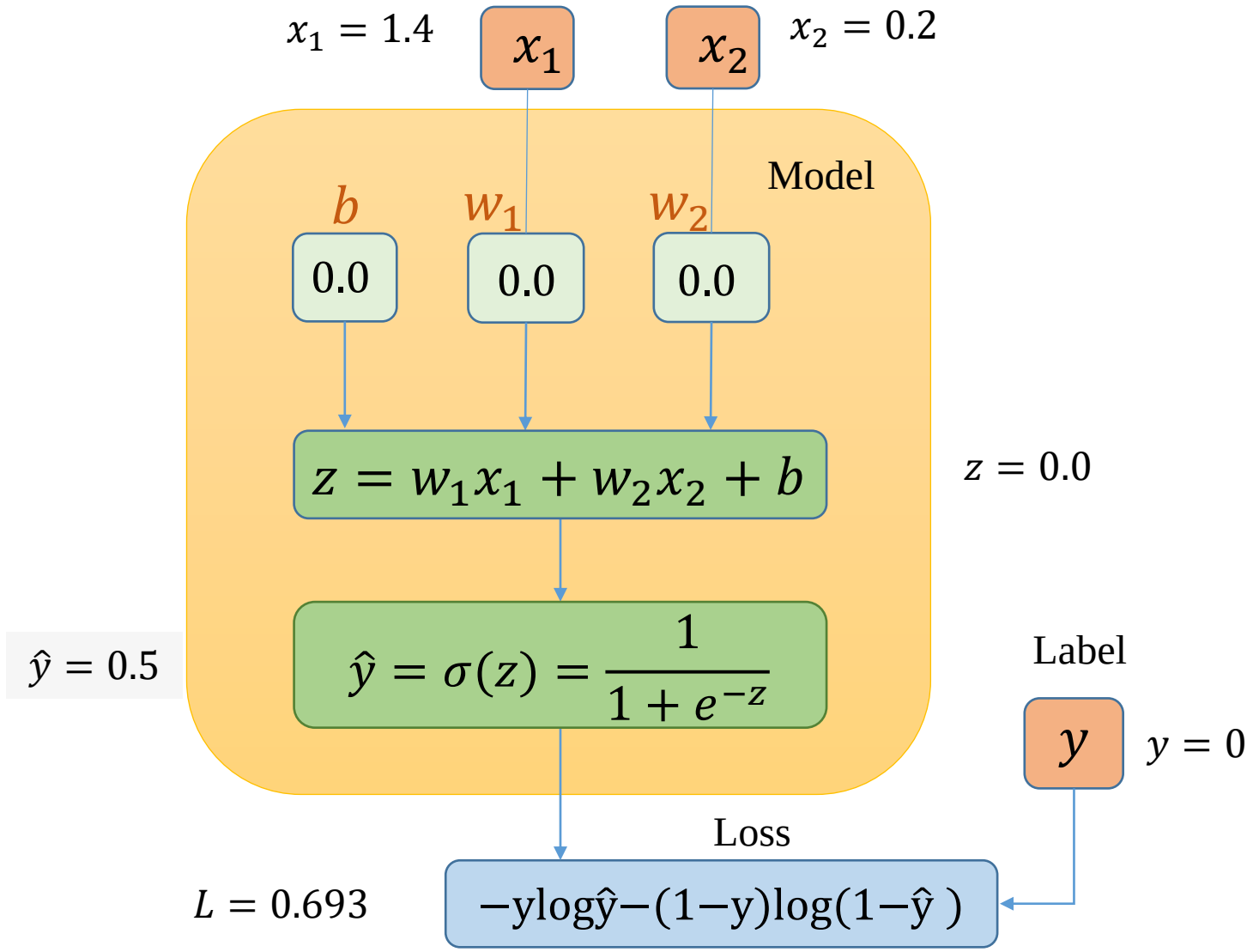


Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$x = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad y = [0]$$



Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad \boldsymbol{y} = [0]$$

$\eta = 0.01$

$b = 0.005$

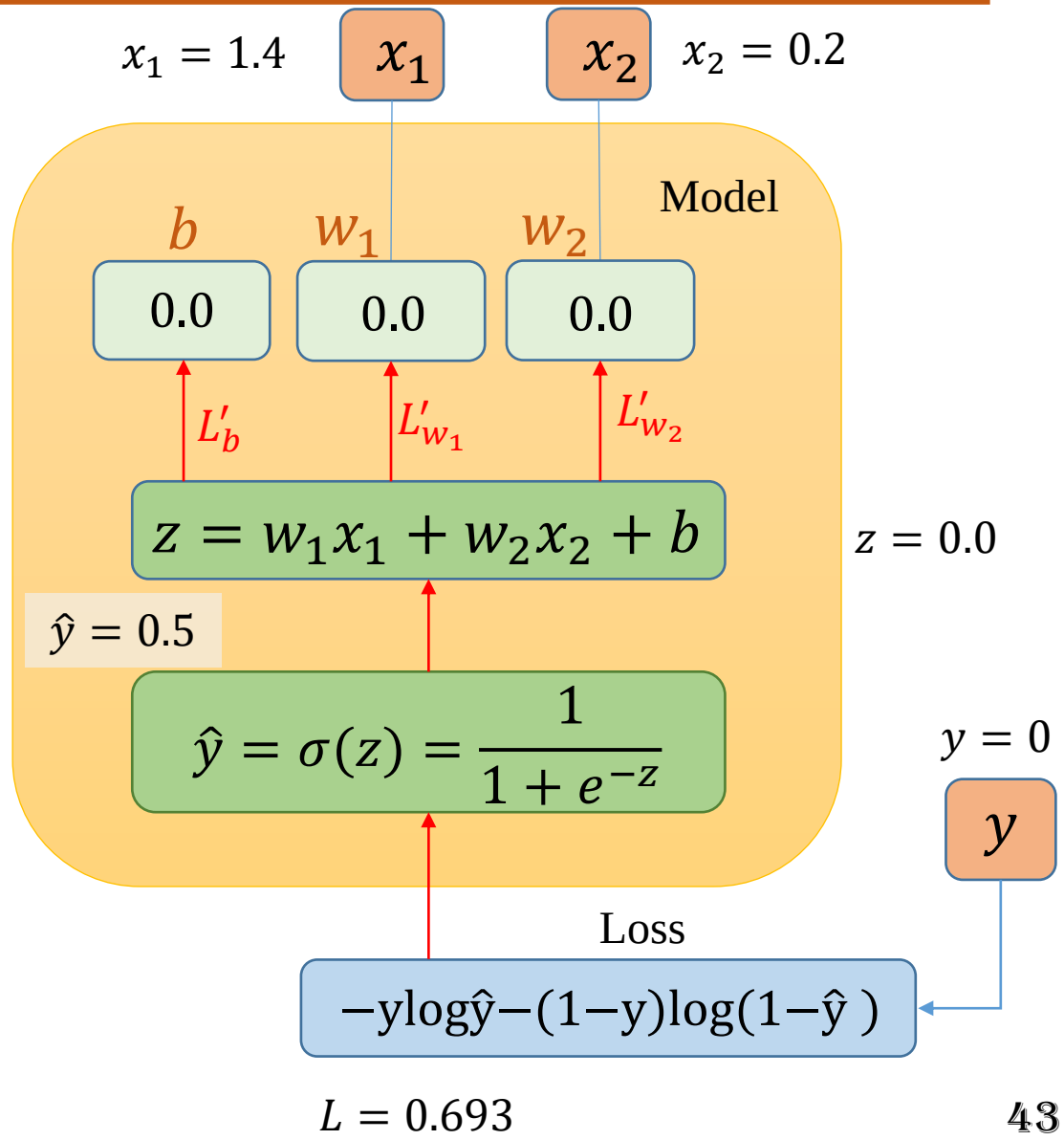
$w_1 = 0.007$

$w_2 = 0.001$

$$L'_{\boldsymbol{\theta}} = \mathbf{x}(\hat{y} - y)$$

$$= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} [0.5]$$

$$= \begin{bmatrix} 0.5 \\ 0.7 \\ 0.1 \end{bmatrix} = \begin{bmatrix} L'_b \\ L'_{w_1} \\ L'_{w_2} \end{bmatrix}$$



Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad \boldsymbol{y} = [0]$$

$\eta = 0.01$

$b = -0.005$

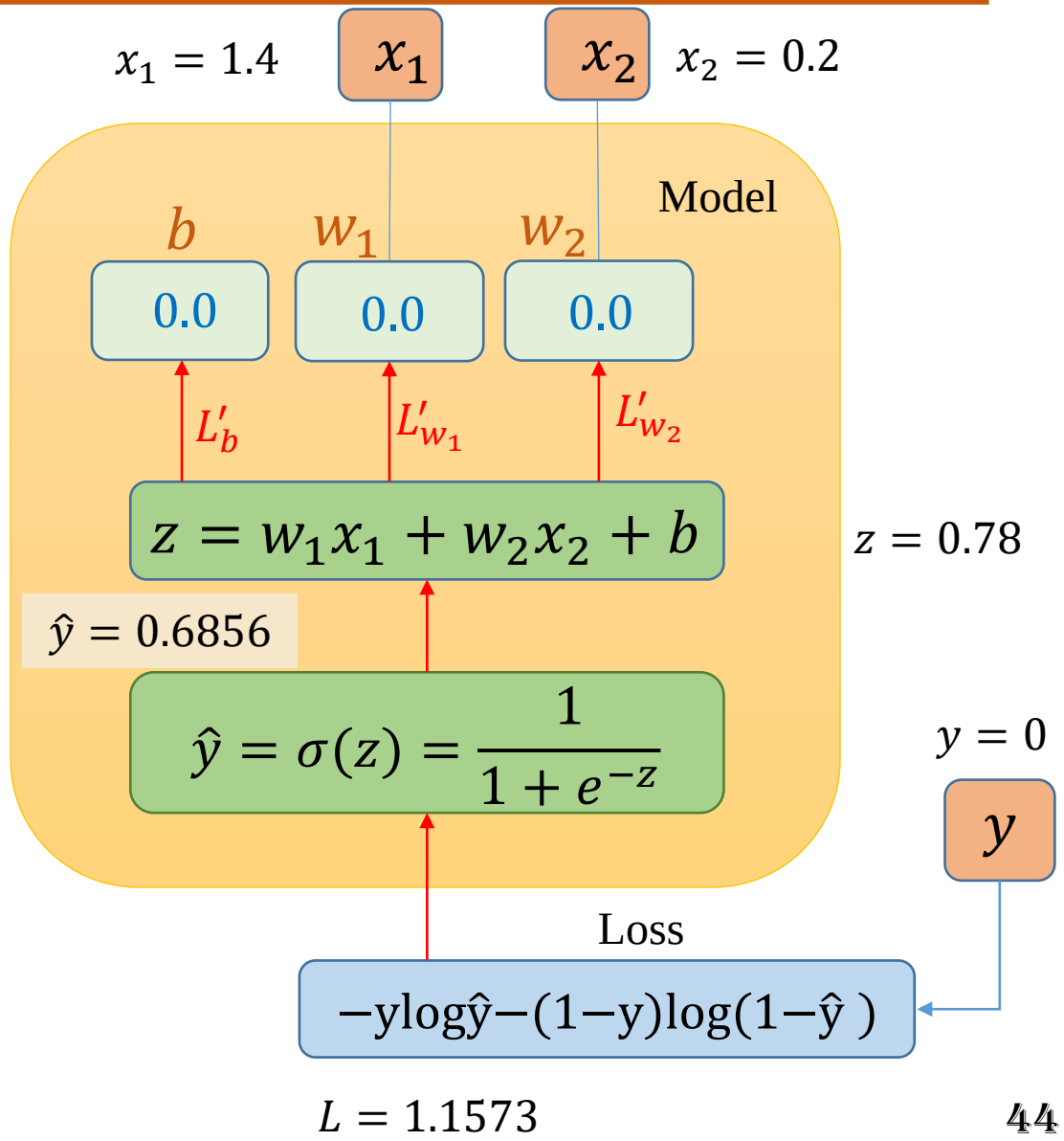
$w_1 = -0.007$

$w_2 = -0.001$

$$L'_{\boldsymbol{\theta}} = \mathbf{x}(\hat{y} - y)$$

$$= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} [0.5]$$

$$= \begin{bmatrix} 0.5 \\ 0.7 \\ 0.1 \end{bmatrix} = \begin{bmatrix} L'_b \\ L'_{w_1} \\ L'_{w_2} \end{bmatrix}$$

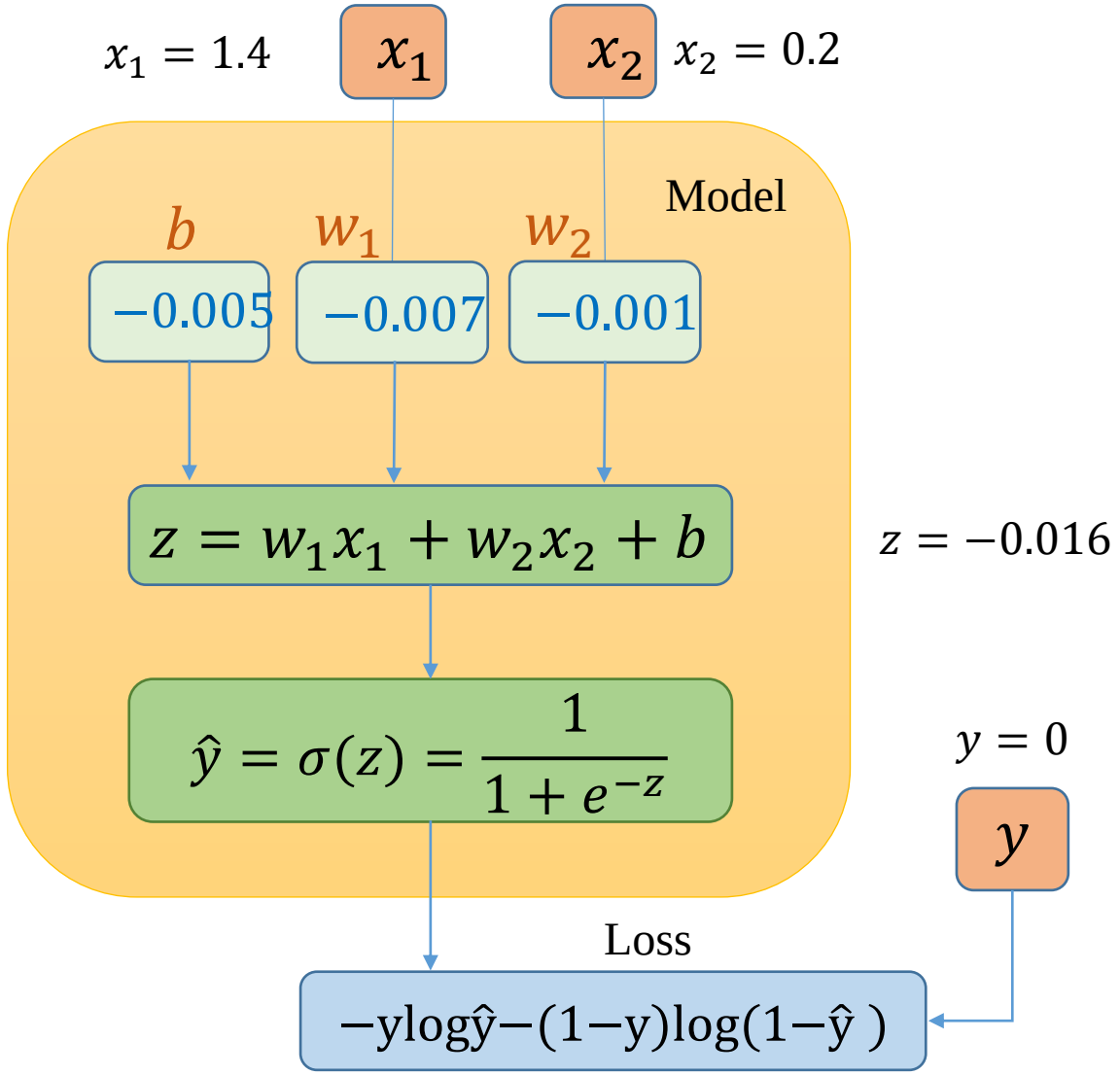


Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$x = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad y = [0]$$



previous $L = 1.1573$

$L = 0.68$

Example 5 - Zero Initialization

❖ Softmax regression

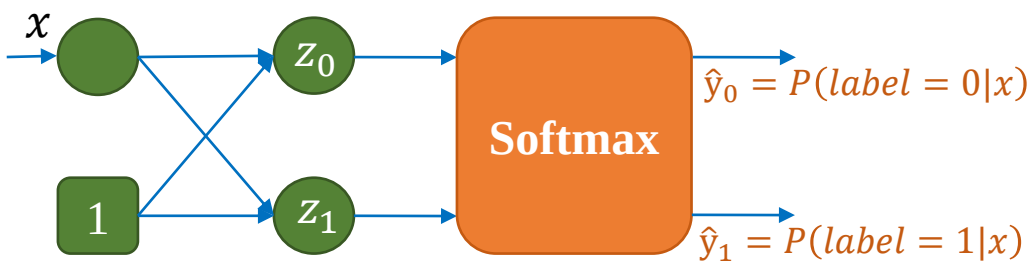
Training data

Feature	Label	
Petal_Length	Label	
1.4	0	Category A
1.3	0	
1.5	0	
4.5	1	Category B
4.1	1	
4.6	1	

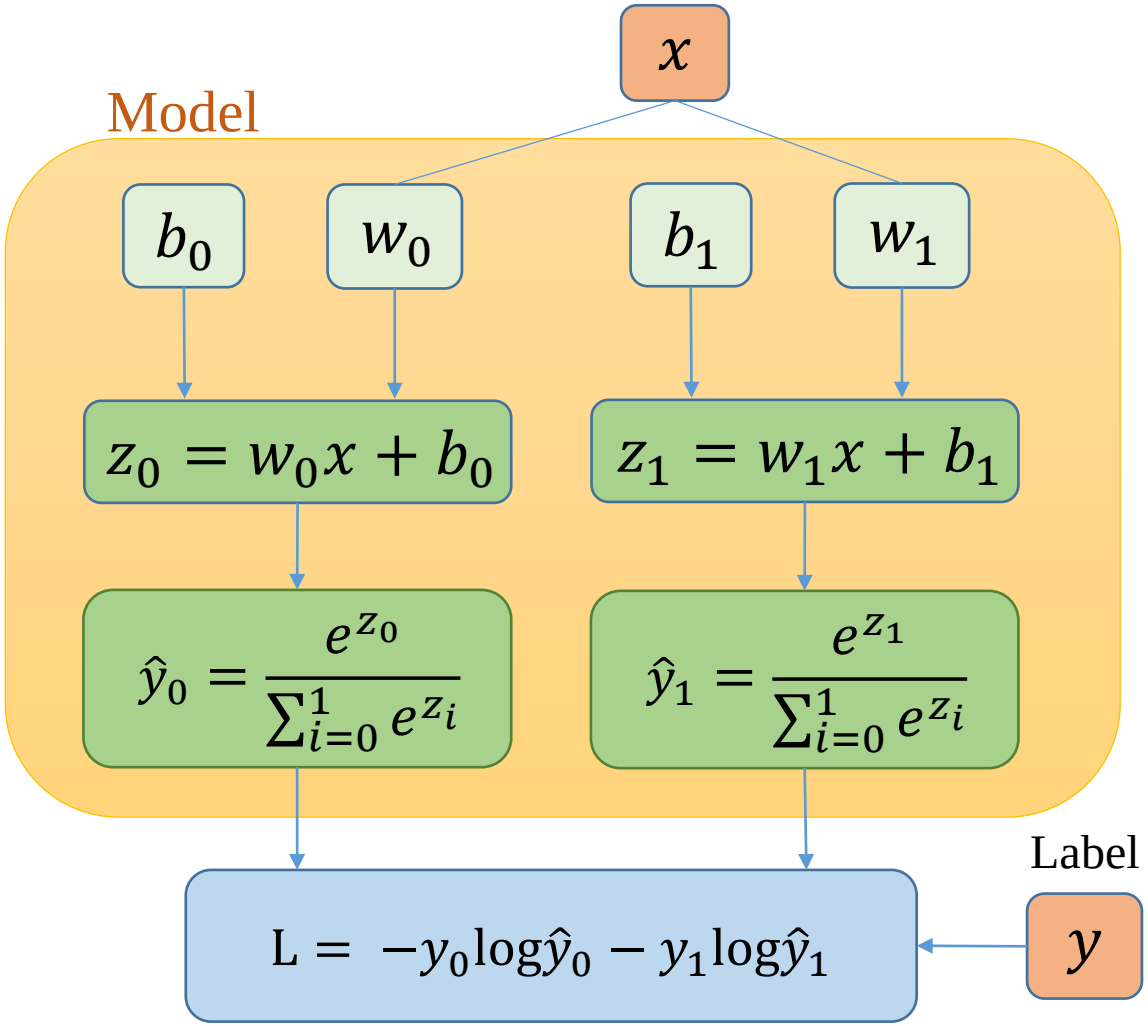
One-hot encoding for labels

index 0 1

$$y = 0 \rightarrow \boldsymbol{y}^T = [1, 0]$$
$$y = 1 \rightarrow \boldsymbol{y}^T = [0, 1]$$



Model



Example 5 - Zero Initialization

Training data

Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1

#class=2

#feature=1

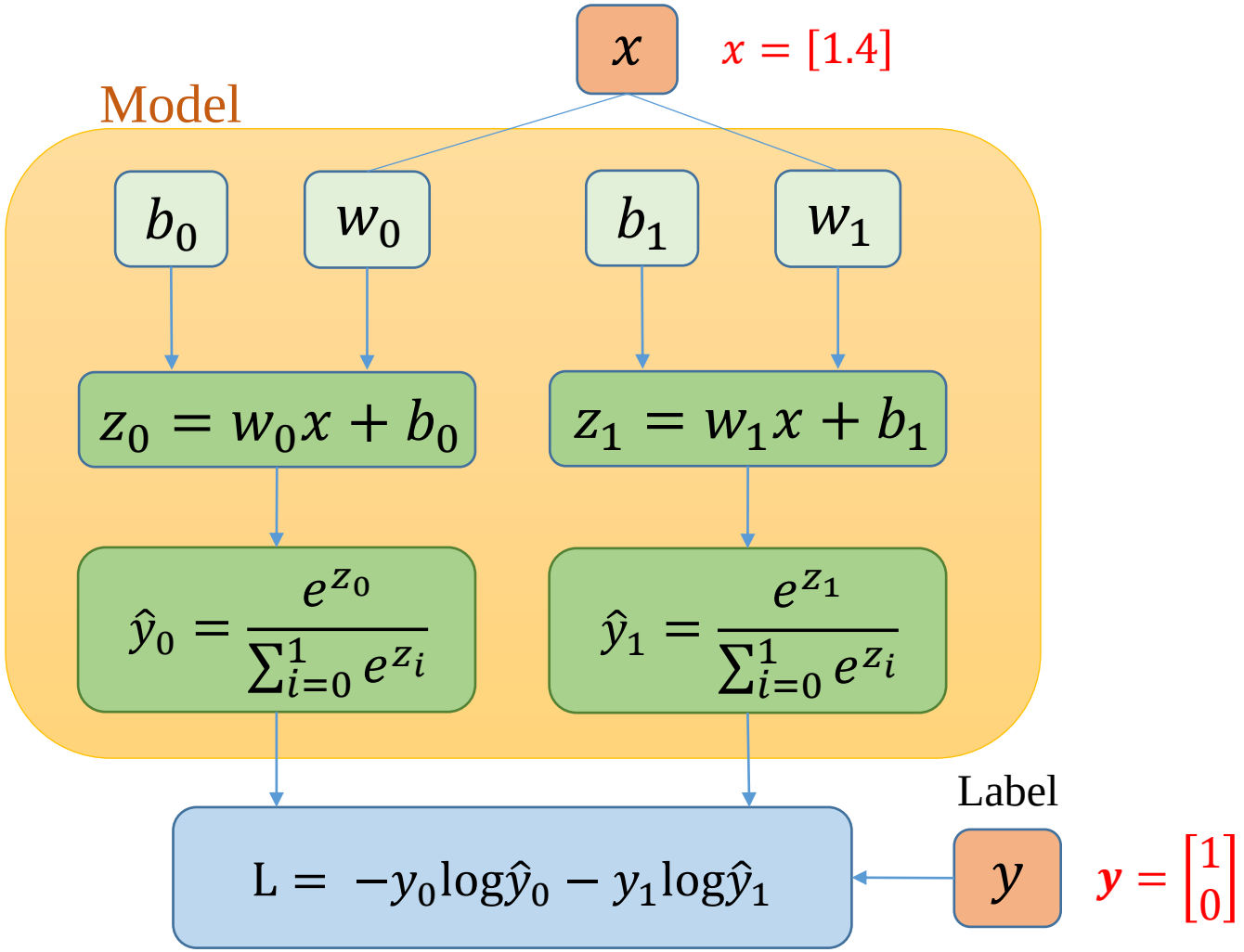
One-hot encoding for label

$$y = 0 \rightarrow \boldsymbol{y}^T = \begin{matrix} y_0 & y_1 \\ [1 & 0] \end{matrix}$$

$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

Training example

$$(x, y) = (1.4, 0)$$



Example 5 - Zero Initialization

Training data

Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1

#class=2

#feature=1

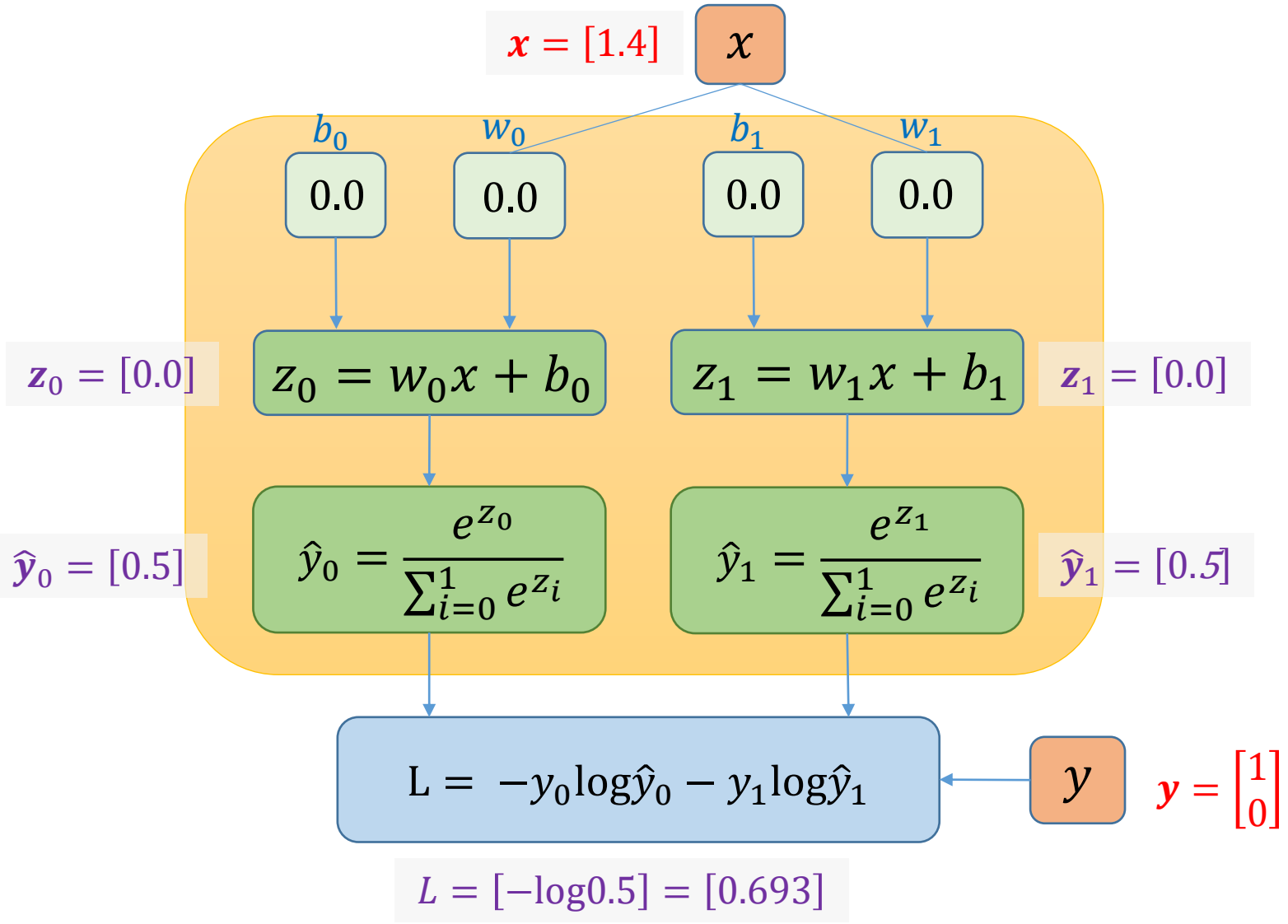
One-hot encoding for label

$$y = 0 \rightarrow \boldsymbol{y}^T = \begin{matrix} y_0 & y_1 \\ [1 & 0] \end{matrix}$$

$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

Training example

$$(x, y) = (1.4, 0)$$



Example 5 - Zero Initialization

Derivative

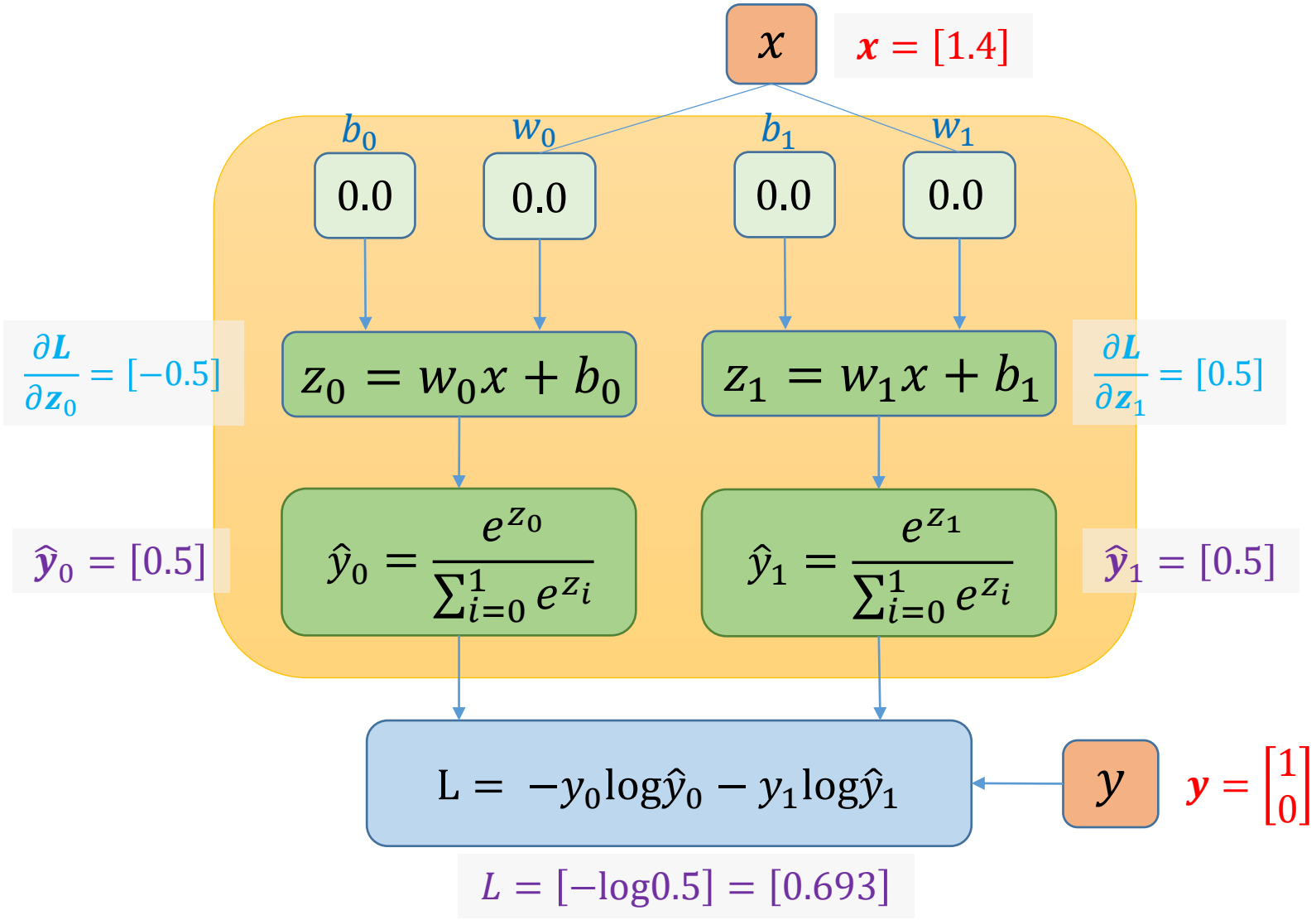
$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$y = 0 \rightarrow \boldsymbol{y}^T = \overset{y_0}{[1} \quad \overset{y_1}{0]}$$
$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

$$\frac{\partial L}{\partial \boldsymbol{z}_0} = \hat{y}_0 - 1$$
$$= 0.5 - 1 = -0.5$$
$$\frac{\partial L}{\partial \boldsymbol{z}_1} = \hat{y}_1 - 0 = 0.5$$



Example 5 - Zero Initialization

Derivative

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

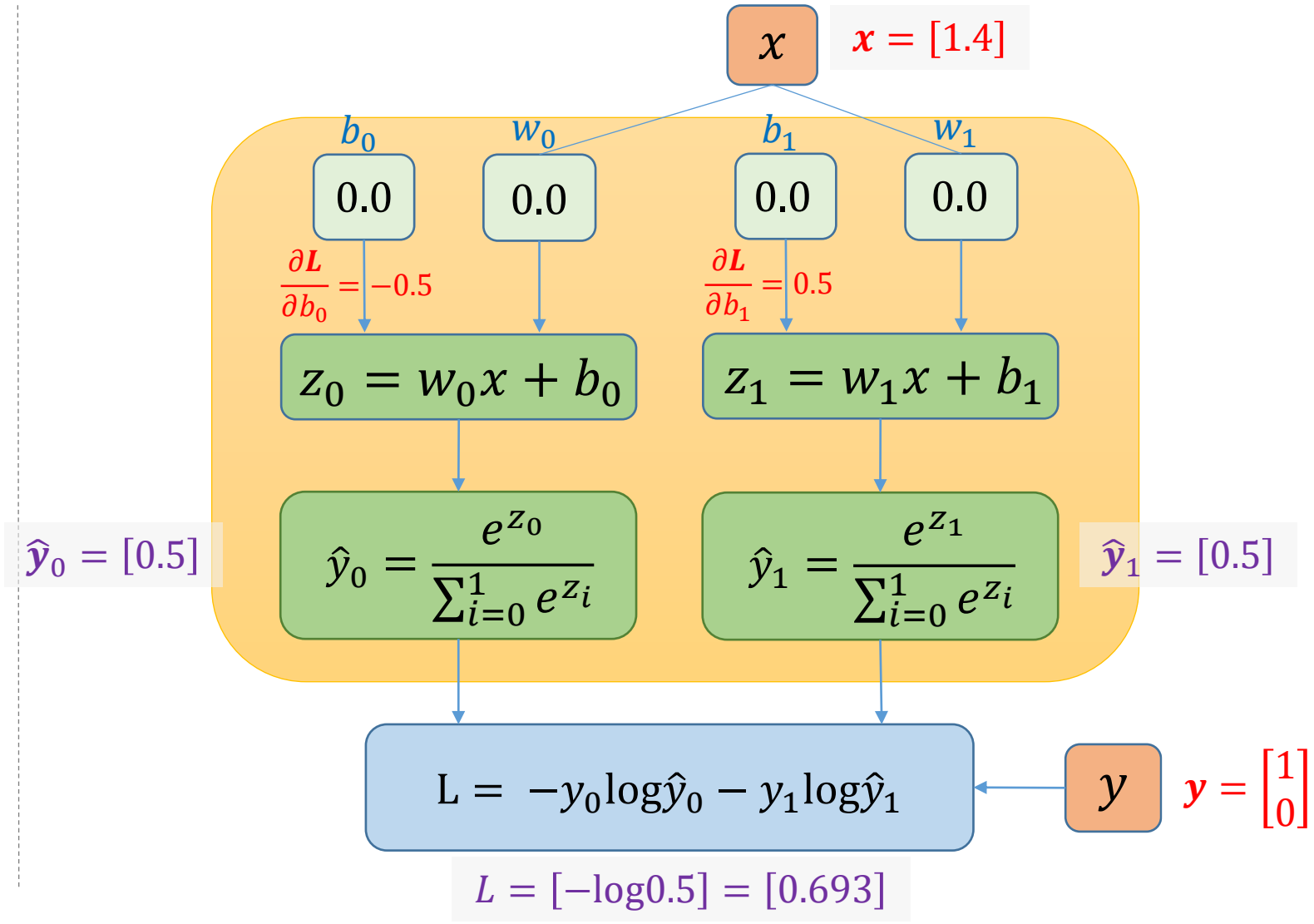
$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$y = 0 \rightarrow \boldsymbol{y}^T = \overset{y_0}{[1} \quad \overset{y_1}{0]}$$
$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

$$\frac{\partial \boldsymbol{L}}{\partial b_0} = (\hat{y}_0 - 1) = -0.5$$

$$\frac{\partial \boldsymbol{L}}{\partial b_1} = (\hat{y}_1 - 0) = 0.5$$



Example 5 - Zero Initialization

Derivative

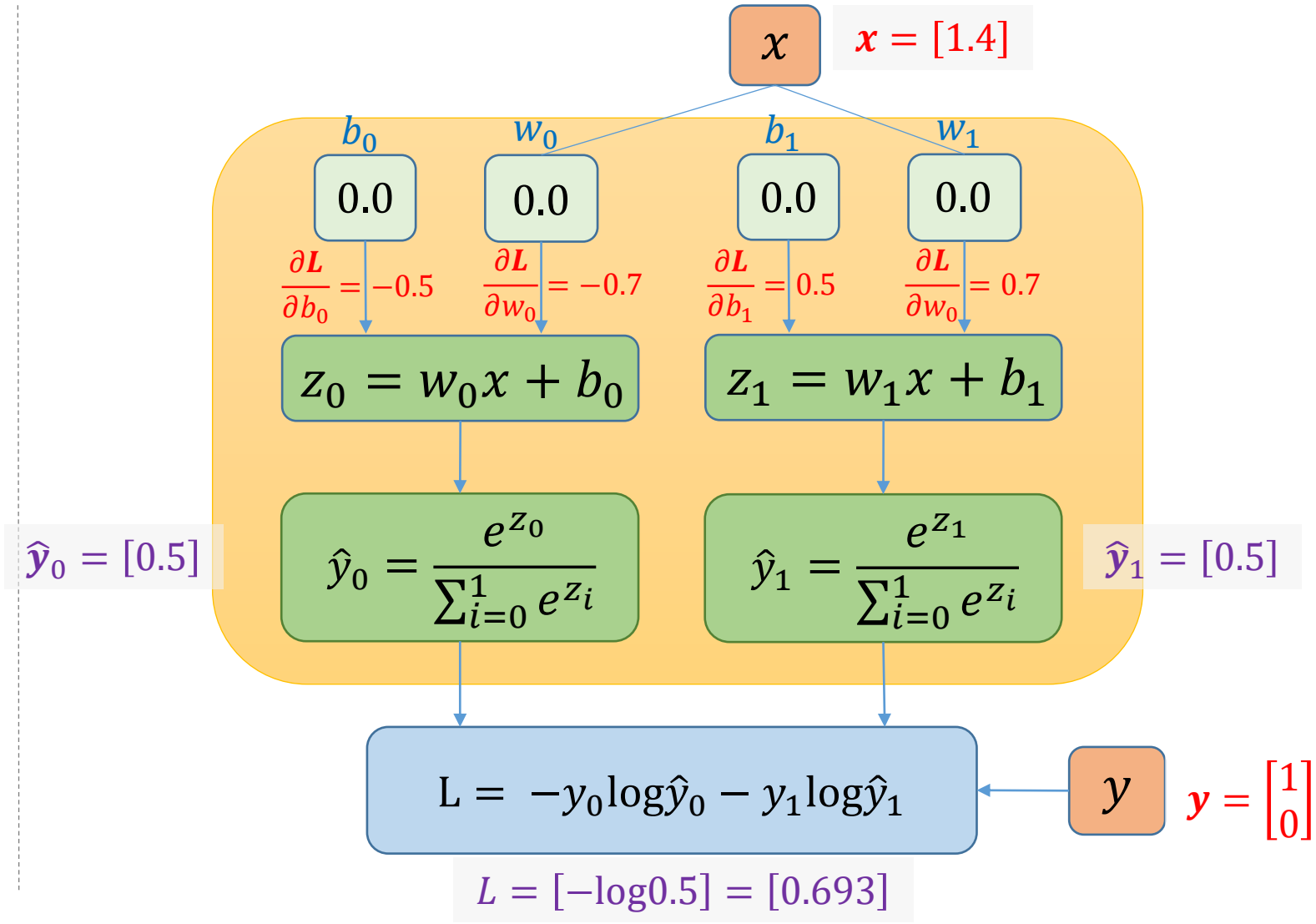
$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$y = 0 \rightarrow \boldsymbol{y}^T = \overset{y_0}{[1} \quad \overset{y_1}{0]}$$
$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

$$\frac{\partial \boldsymbol{L}}{\partial w_0} = x(\hat{y}_0 - 1)$$
$$= -0.5*1.4 = -0.7$$
$$\frac{\partial \boldsymbol{L}}{\partial w_1} = x(\hat{y}_1 - 0)$$
$$= 0.5*1.4 = 0.7$$



Example 5 - Zero Initialization

Update parameters

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta L'_{\boldsymbol{\theta}}$$

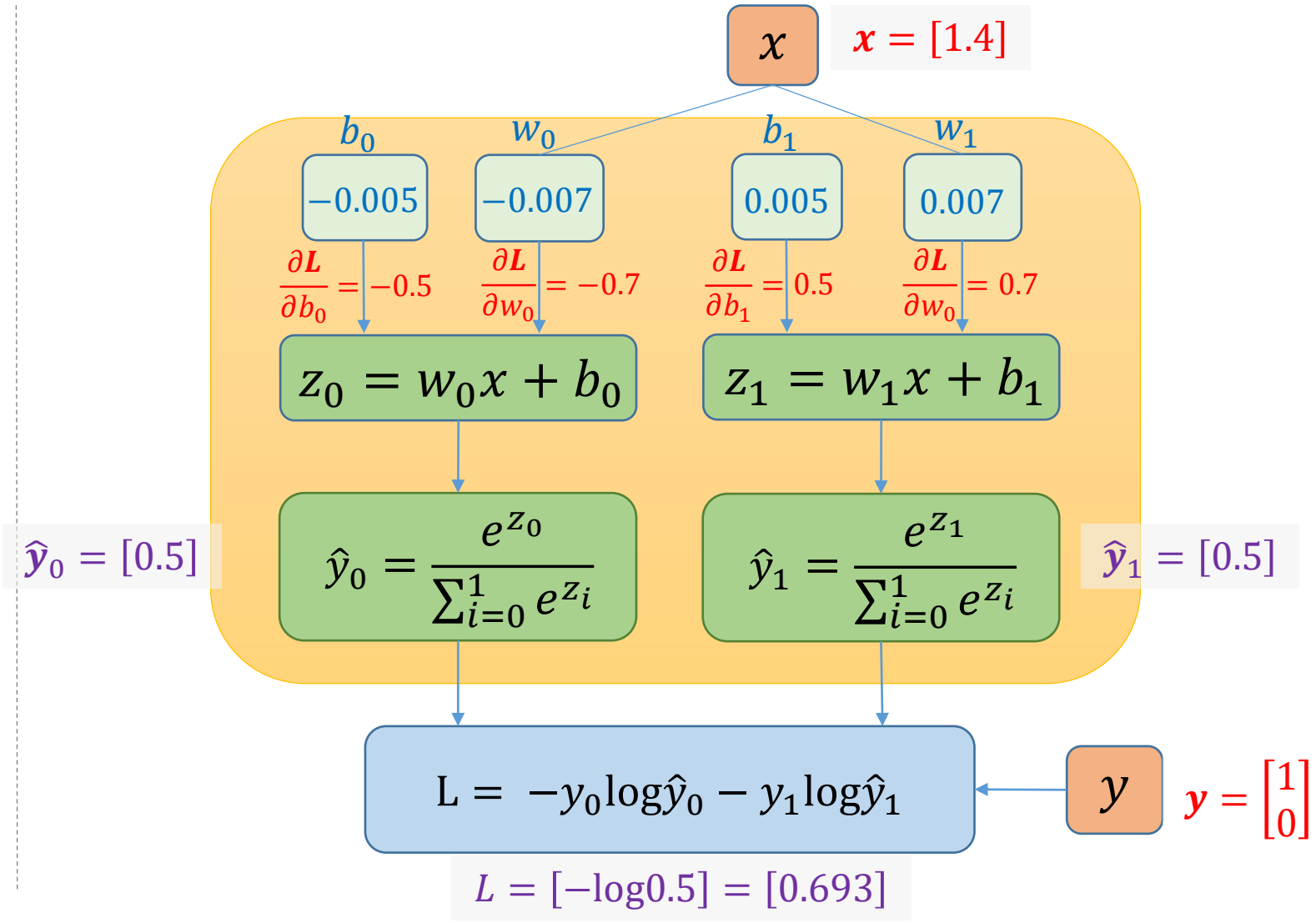
η is learning rate

$$\boldsymbol{\theta} = \begin{bmatrix} b_0 & b_1 \\ w_0 & w_1 \end{bmatrix} \qquad L'_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial L}{\partial b_0} & \frac{\partial L}{\partial b_1} \\ \frac{\partial L}{\partial w_0} & \frac{\partial L}{\partial w_1} \end{bmatrix}$$

$$\eta = 0.1$$

$$\boldsymbol{\theta} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix} - 0.01 \begin{bmatrix} -0.5 & 0.5 \\ -0.7 & 0.7 \end{bmatrix}$$

$$= \begin{bmatrix} -0.005 & 0.005 \\ -0.007 & 0.007 \end{bmatrix}$$



Example 5 - Zero Initialization

Training data

Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1

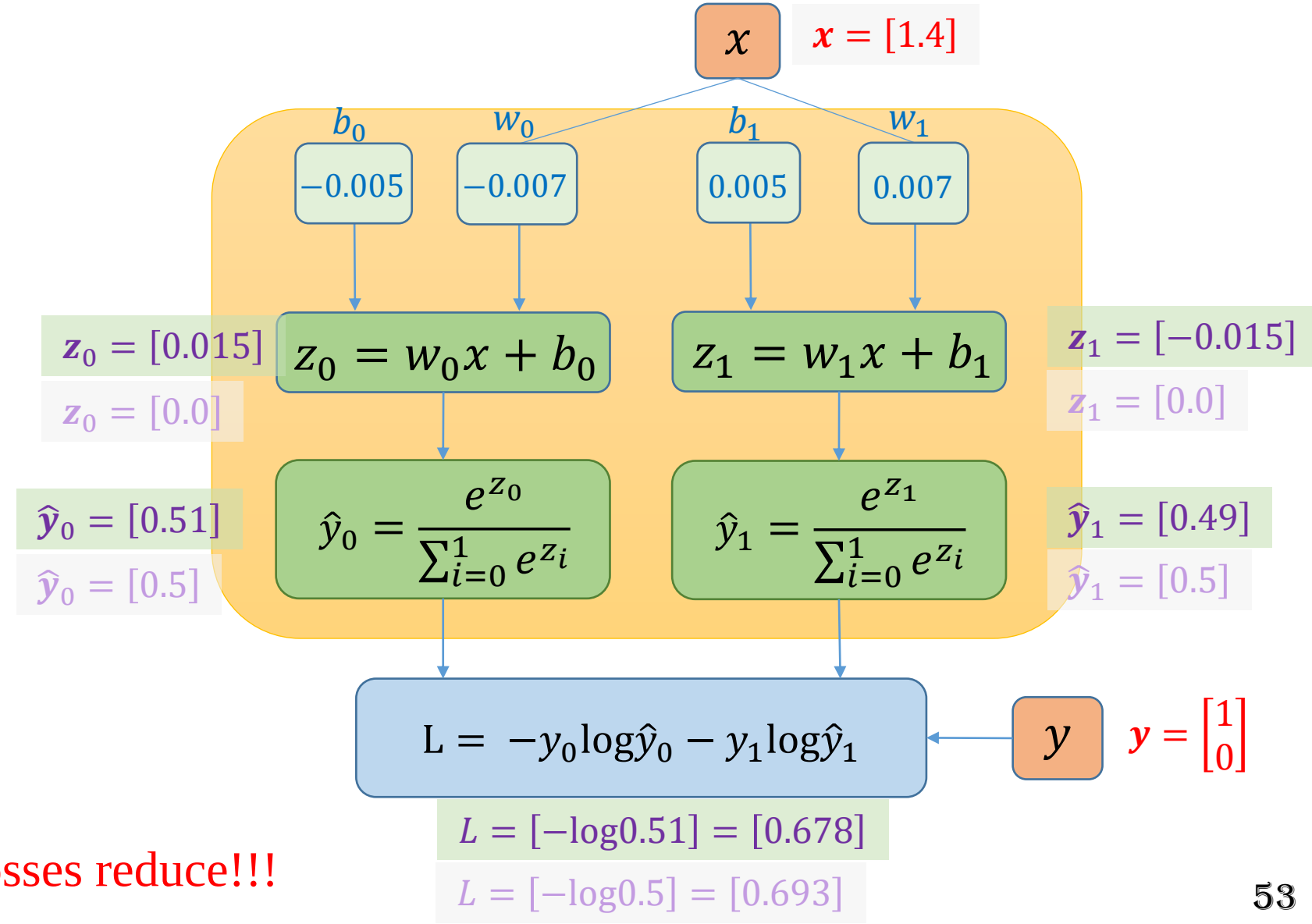
One-hot encoding for label

$$y = 0 \rightarrow \boldsymbol{y}^T = \begin{matrix} y_0 & y_1 \\ [1 & 0] \end{matrix}$$

$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

Training example

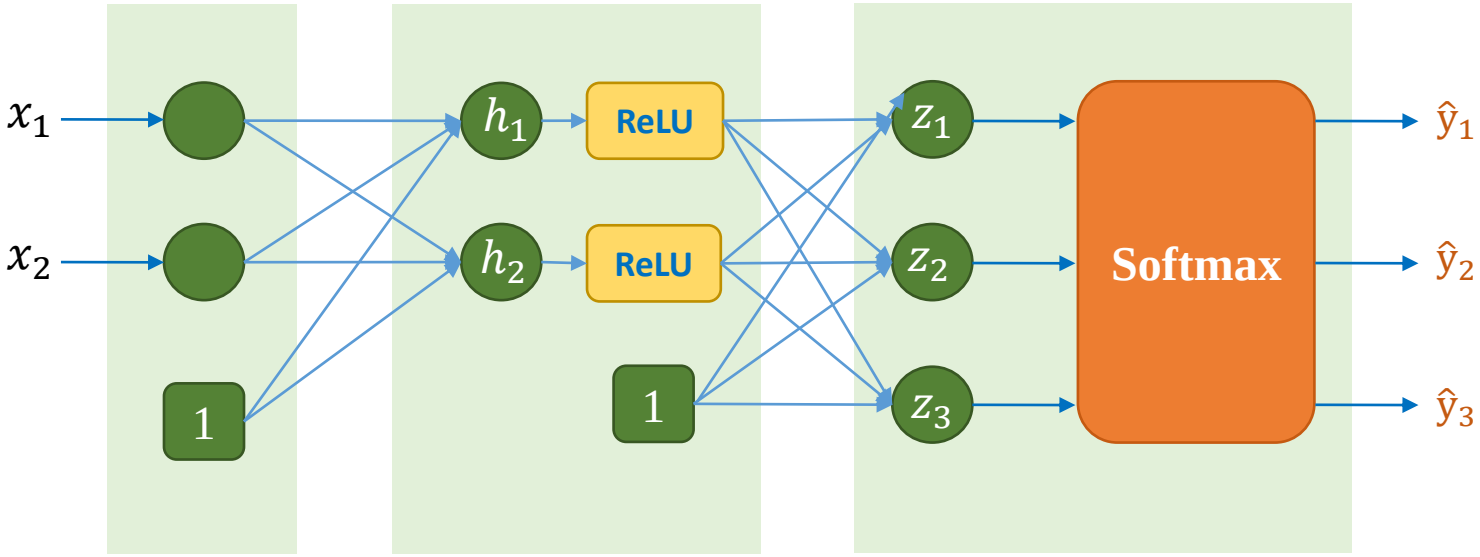
$$(x, y) = (1.4, 0)$$



losses reduce!!!

Example 6 - Zero Initialization

Feature		Label
Petal Length	Petal Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{bmatrix} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix}$$

$$y = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$h = [h_1 \quad h_2] = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$w = [w_1 \quad w_2 \quad w_3] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$b_h = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$b_w = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

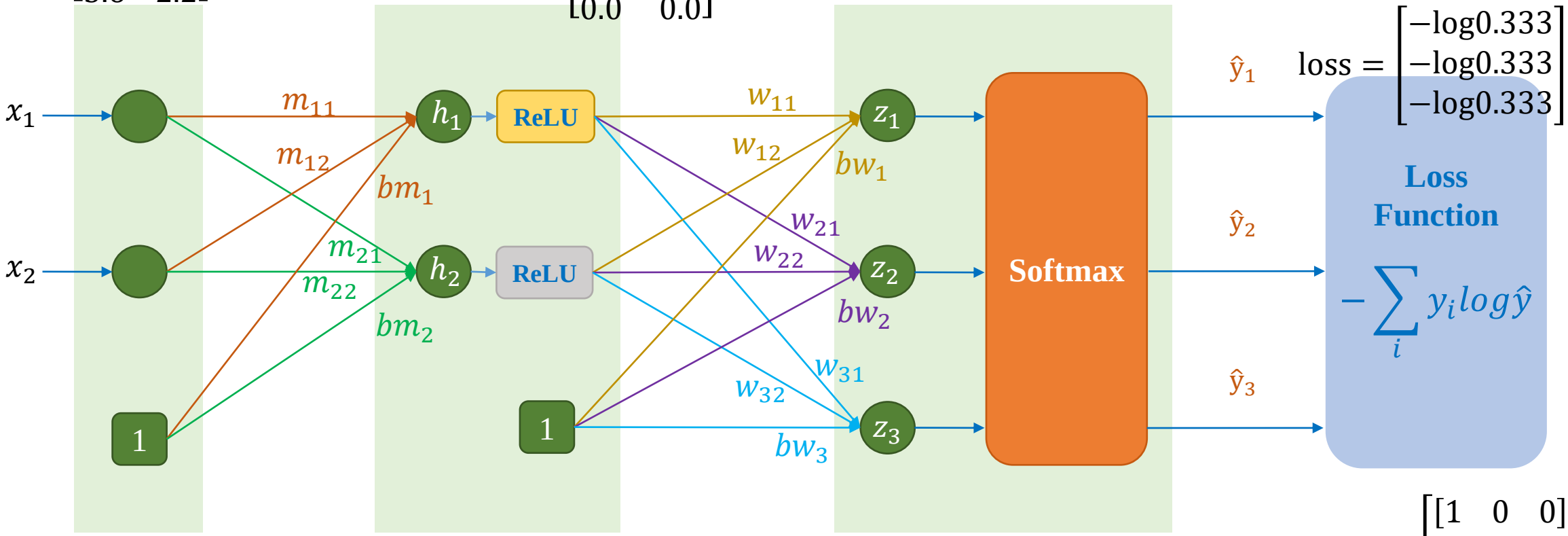
$$\boldsymbol{x} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix}$$

$$\boldsymbol{h} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\mathbf{ReLU} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{z} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \begin{bmatrix} 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}$$



$$\text{loss} = \begin{bmatrix} -\log 0.333 \\ -\log 0.333 \\ -\log 0.333 \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

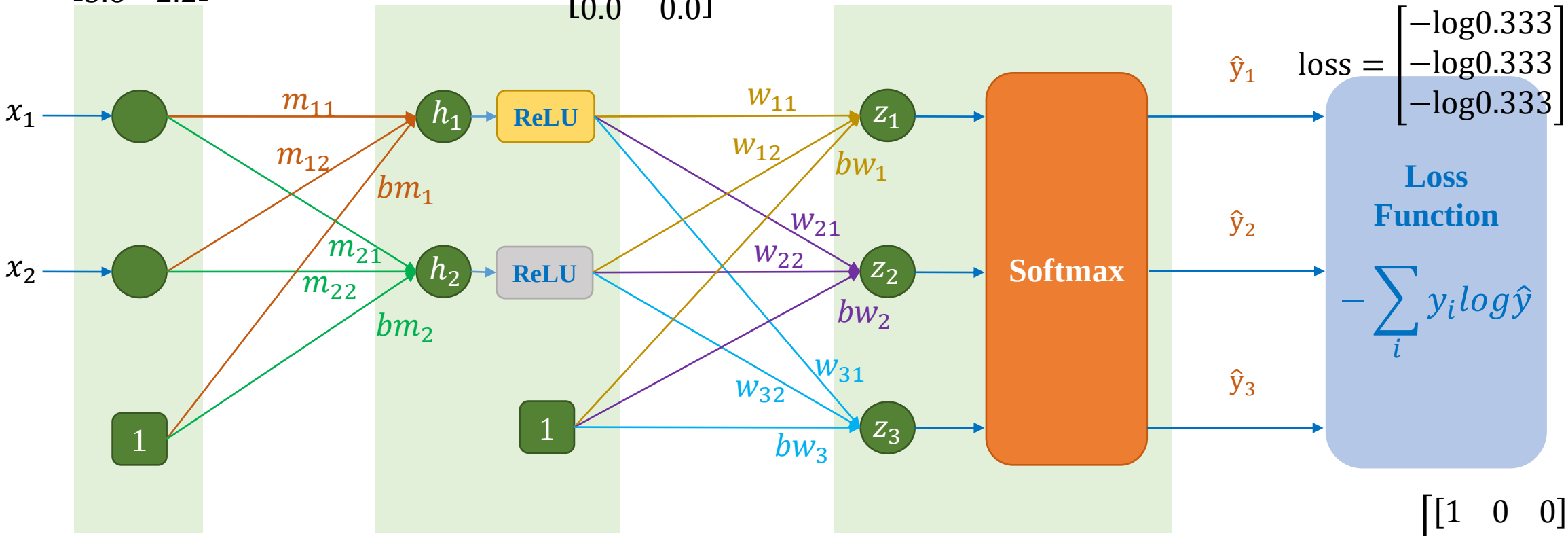
$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix} \quad \boldsymbol{h} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix} \quad \mathbf{ReLU} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{z} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \hat{\boldsymbol{y}} = \begin{bmatrix} 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}$$

$$\text{loss} = \begin{bmatrix} -\log 0.333 \\ -\log 0.333 \\ -\log 0.333 \end{bmatrix}$$



$$\boldsymbol{y} = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} v \\ v \\ v \end{bmatrix}$$

Optimizers

Optimizer Selection

Data Preparation

⬇

Data Normalization

⬇

Model (Network) Construction

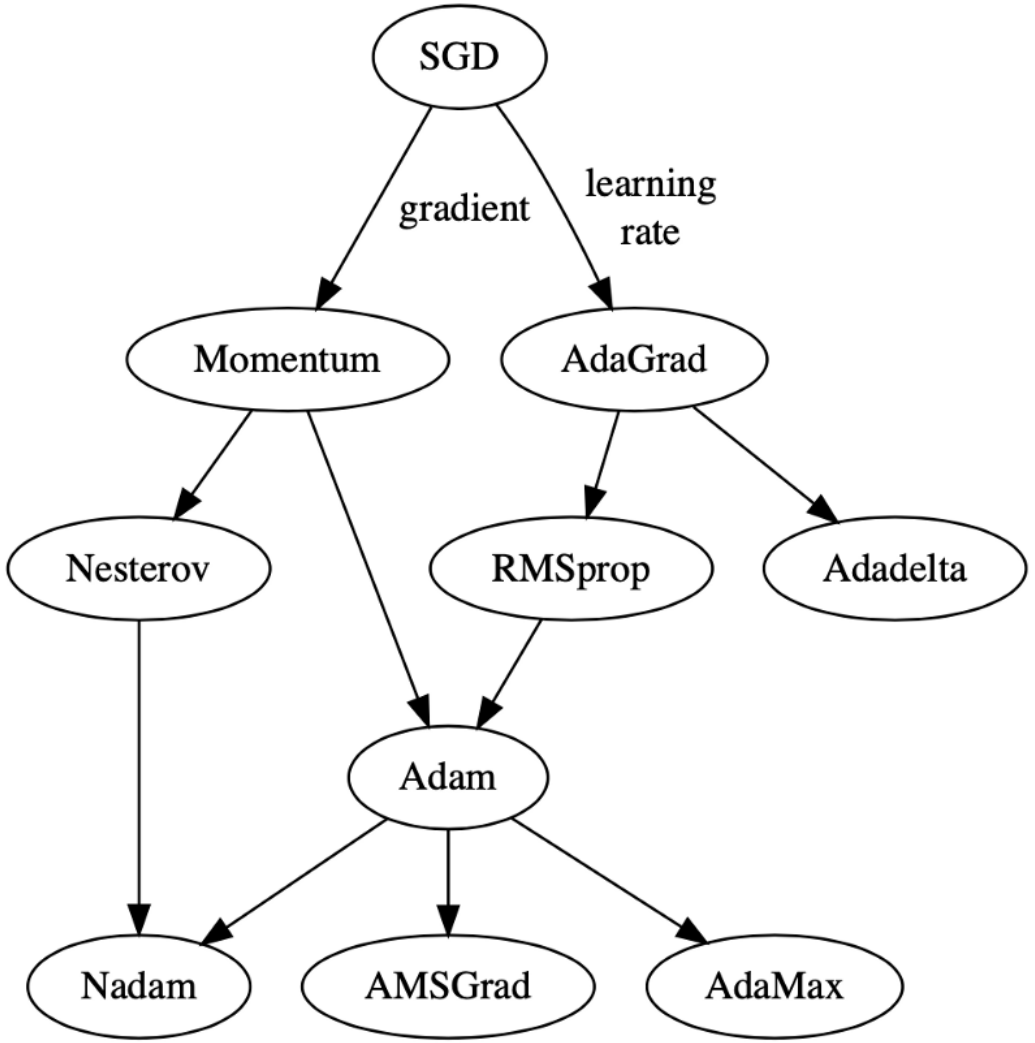
⬇

Parameter Initialization

→

Define a way to update parameters

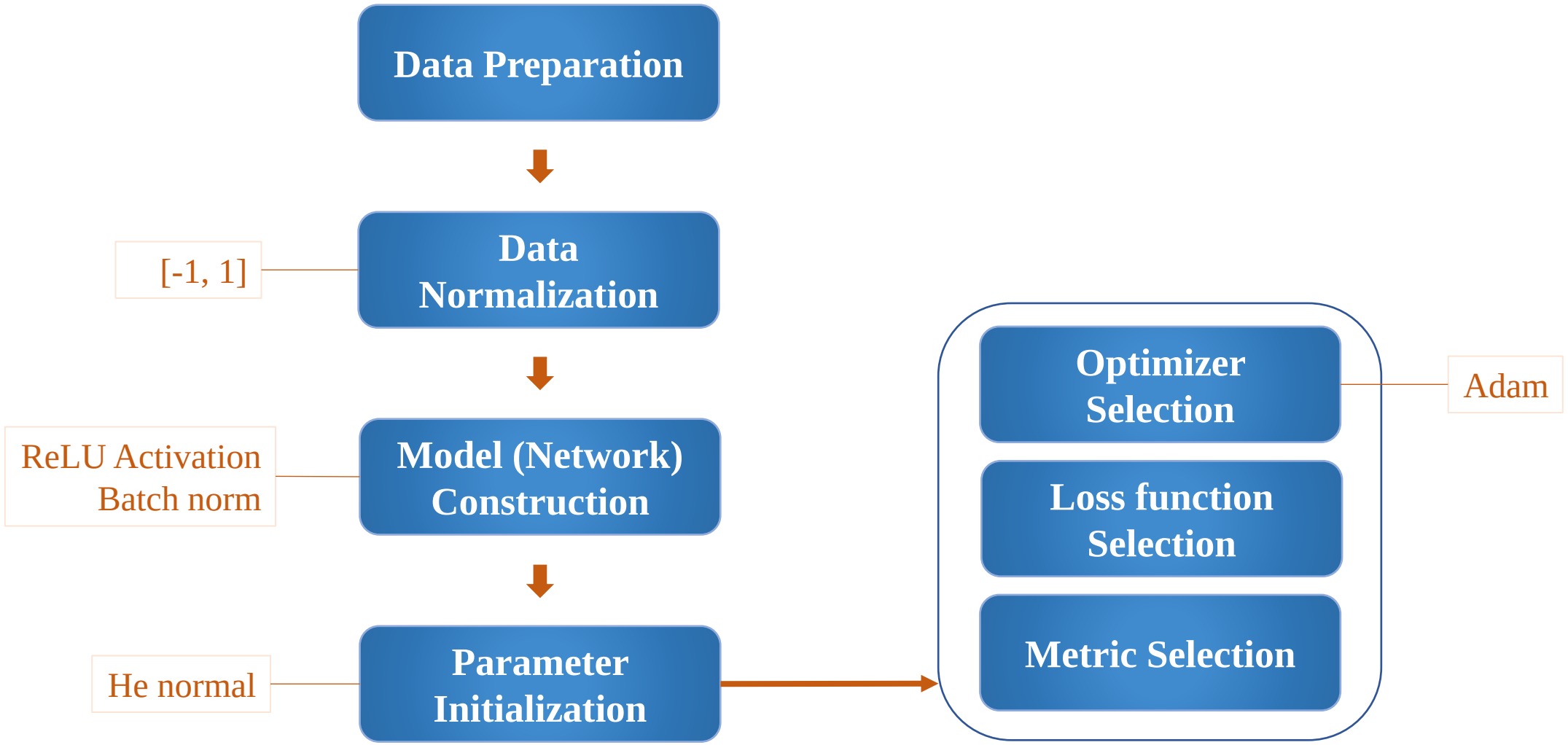
- **Optimizer Selection**
- Loss function Selection
- Metric Selection



https://www.kdnuggets.com/2019/06/gradient-descent-algorithms-cheat-sheet.html

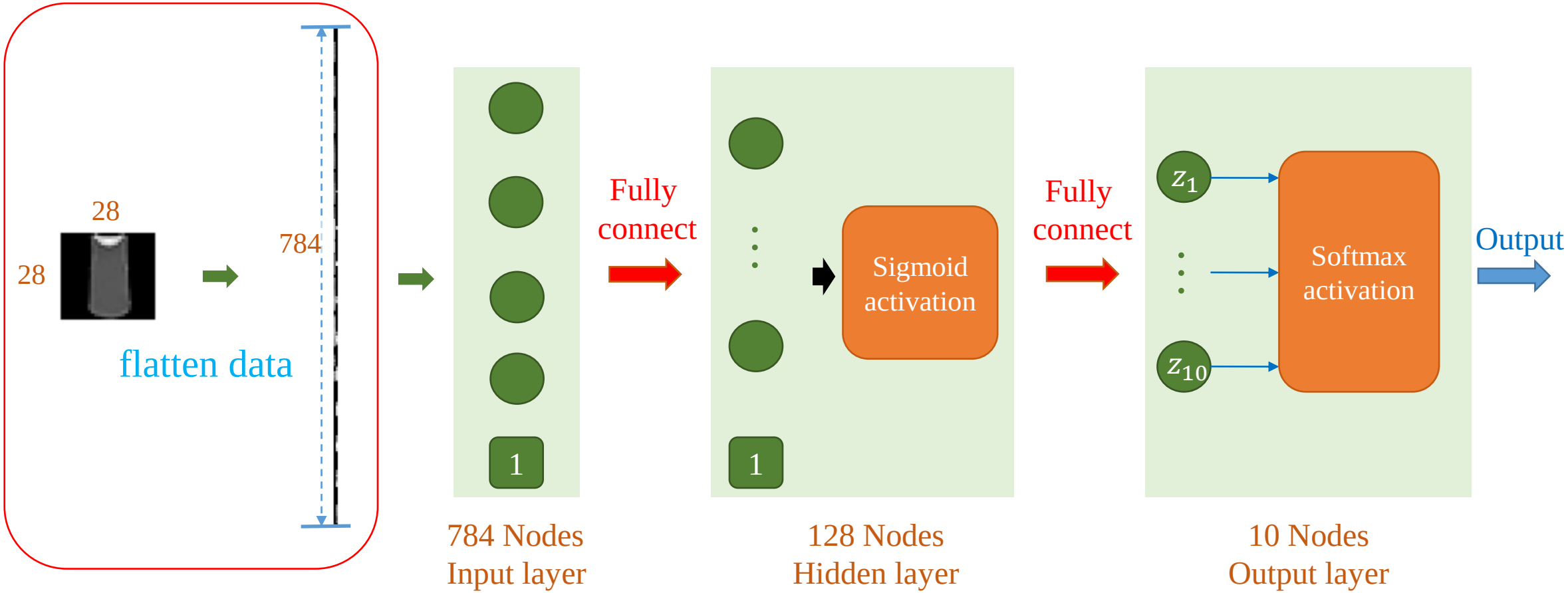
Summary

Recommendation



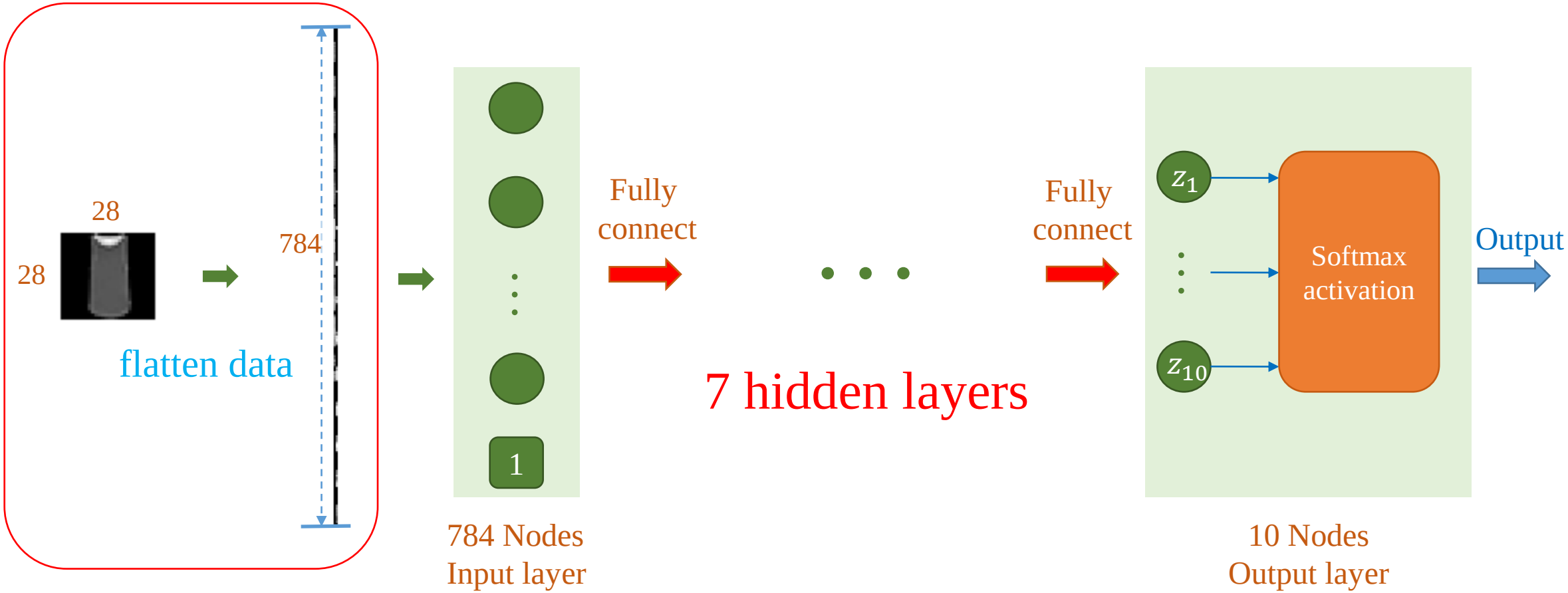
Discussion

- **Sigmoid and SGD**
- **W/o using normalization**



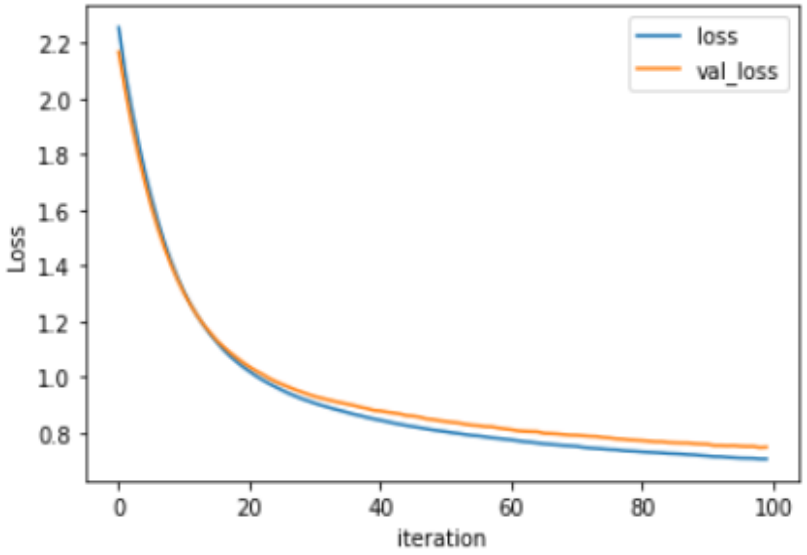
Discussion

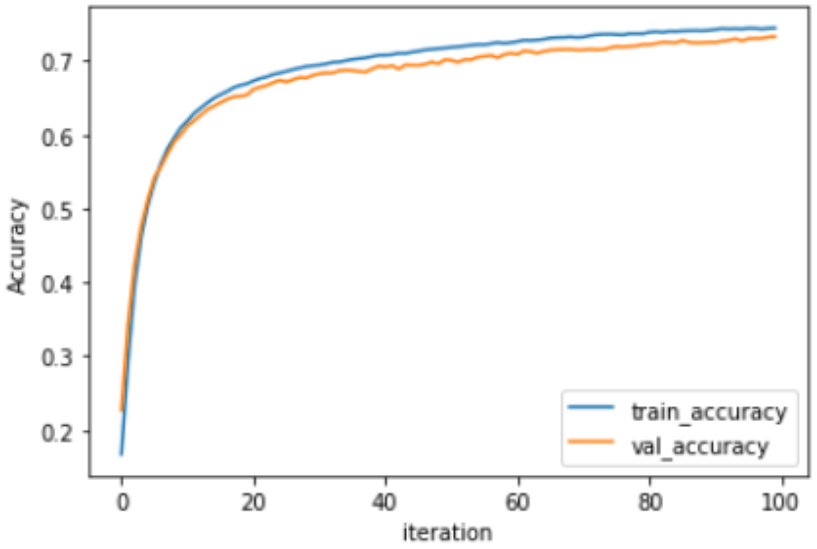
- **Sigmoid and SGD**
- **W/o using normalization**



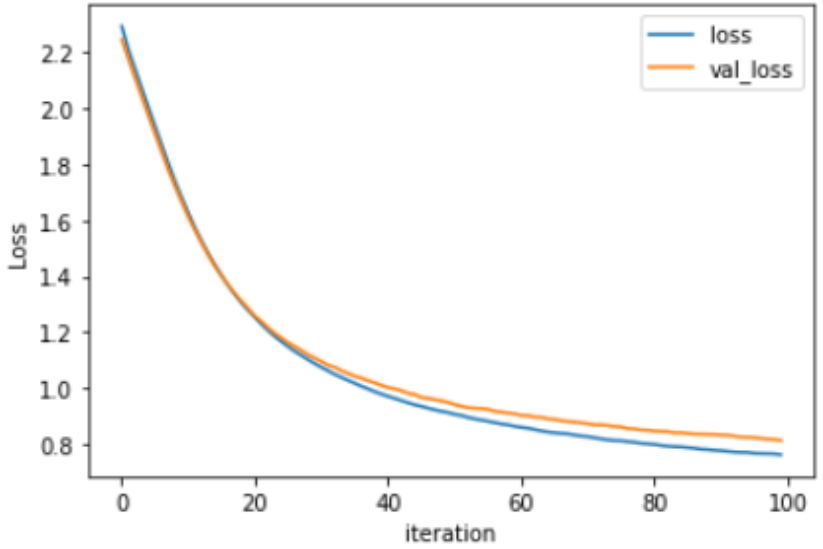
Discussion

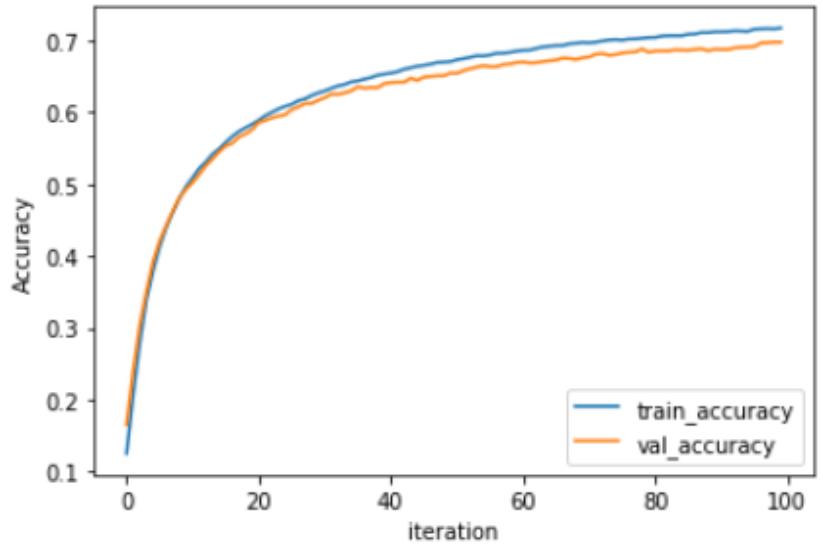
2 hidden layers



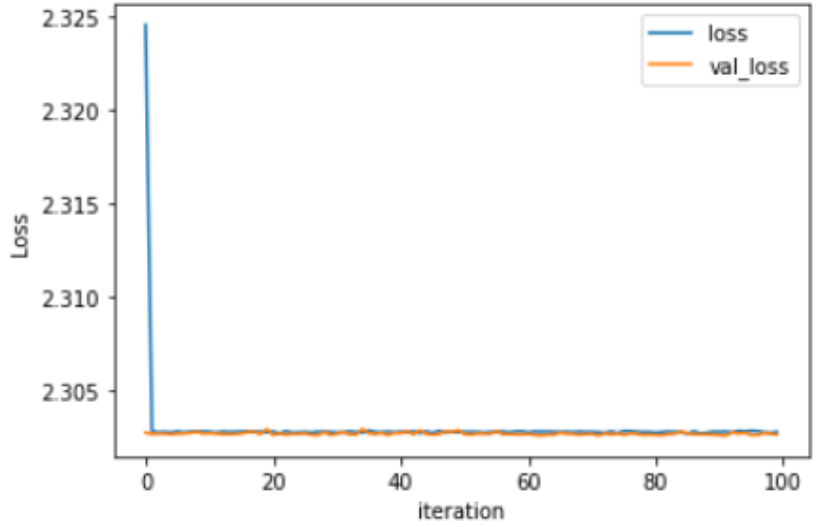


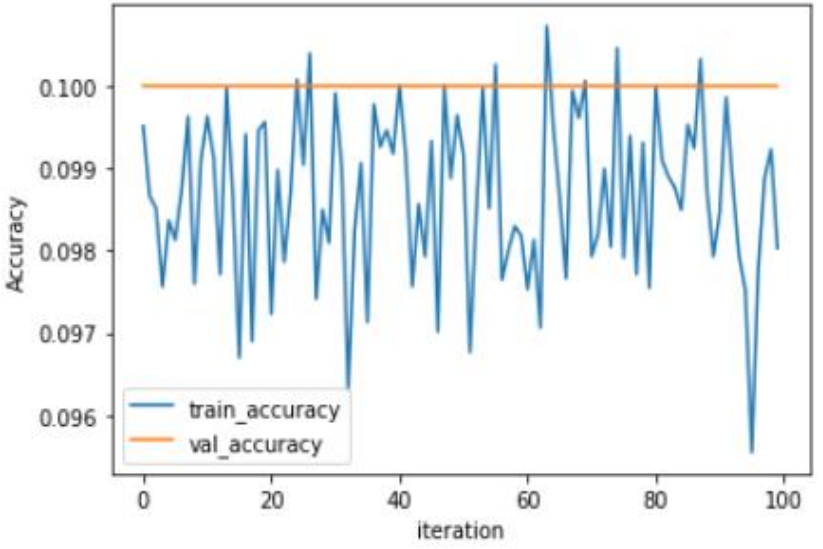
5 hidden layers (!)





7 hidden layers

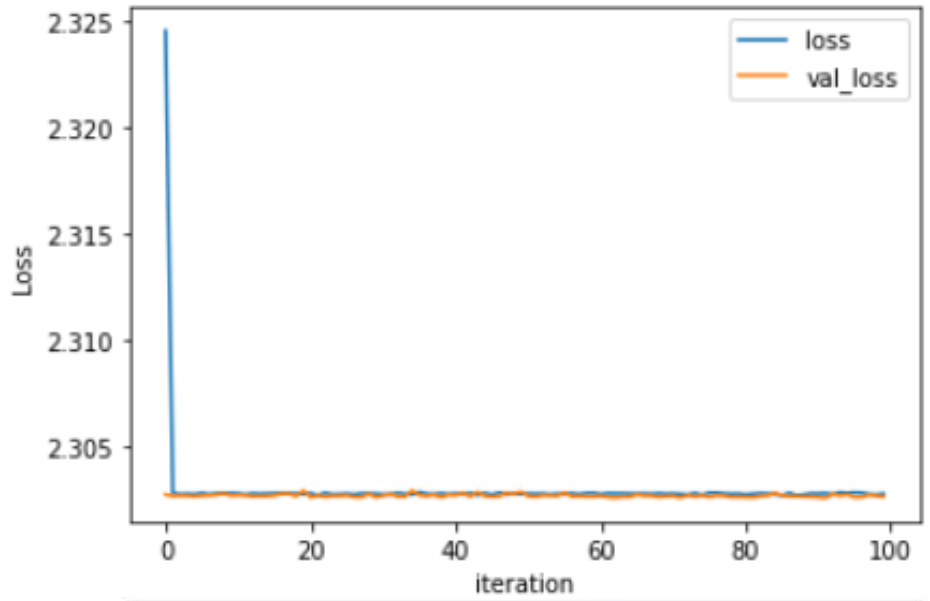


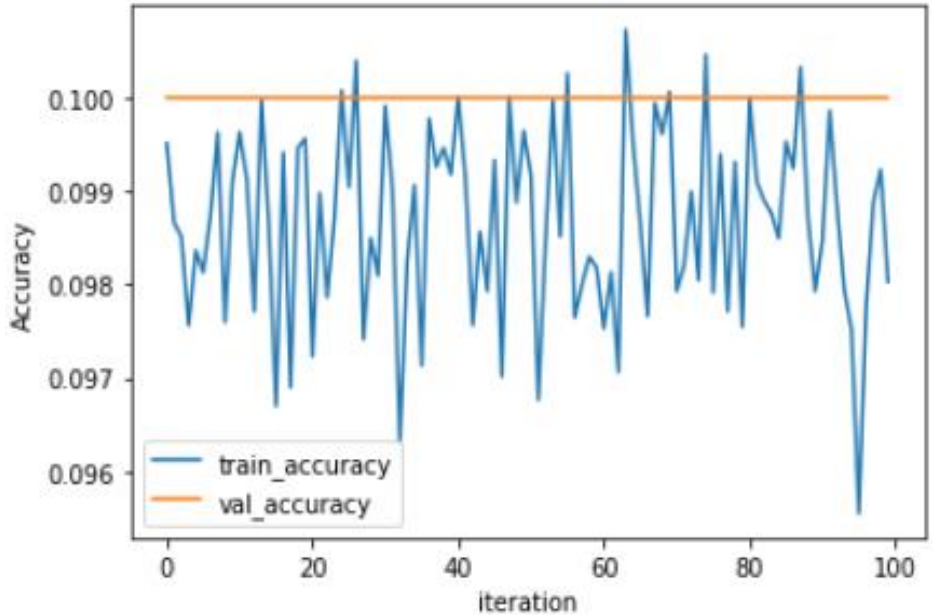


Discussion

Tensorflow

```python
import tensorflow as tf
import tensorflow.keras as keras

# set seed
tf.random.set_seed(1234)
initializer = tf.keras.initializers.RandomNormal()

# create model
model = keras.Sequential()
model.add(keras.Input(shape=(784,)))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(128, activation='sigmoid',
                             kernel_initializer=initializer))
model.add(keras.layers.Dense(10, activation='softmax'))
```





Further Reading

Dying ReLU

https://towardsdatascience.com/the-dying-relu-problem-clearly-explained-42d0c54e0d24

Initialization

https://www.deeplearning.ai/ai-notes/initialization/index.html