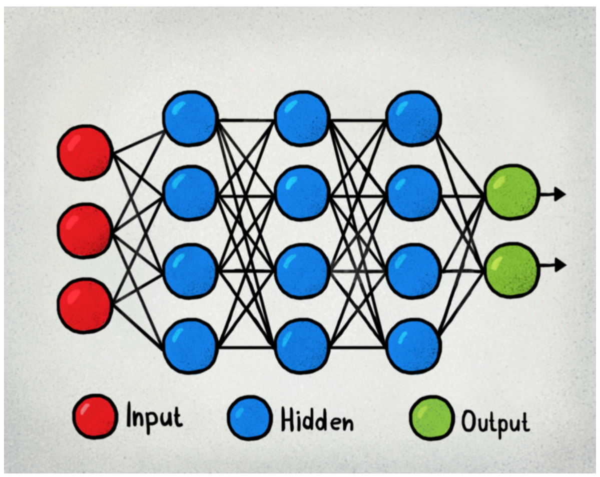
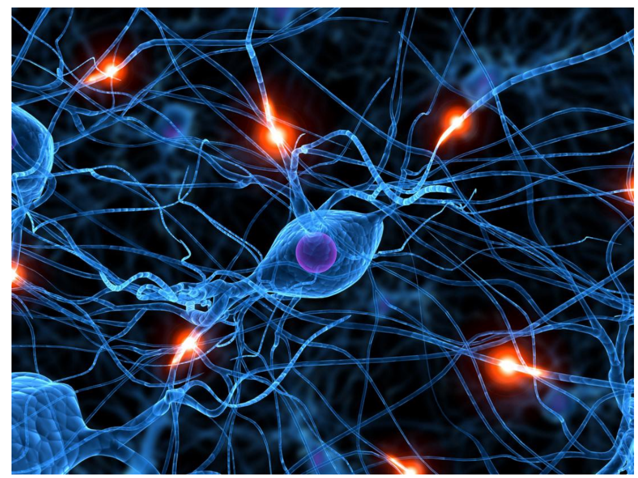
CSE 3521: Neural Networks



OR



The Ohio State University

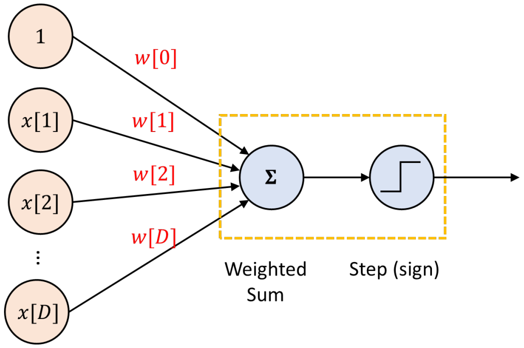
Today

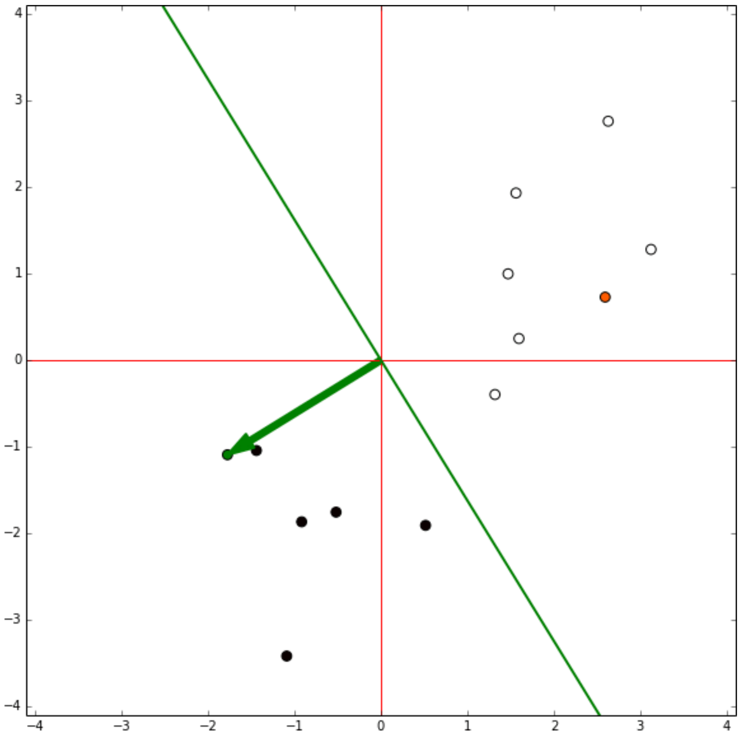
- Learning "deeper" networks beyond one-layer perceptrons
  - Losses and gradients
  - Back-propagation
  - Stochastic gradient descent
- Training particulars
  - Regularization or weight decay
- Discriminative vs. generative models

Perceptron Algorithm

- Let $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{w} \in \mathbb{R}^D$ (b is merged into $\boldsymbol{w}$), $y \in \{-1,1\}$

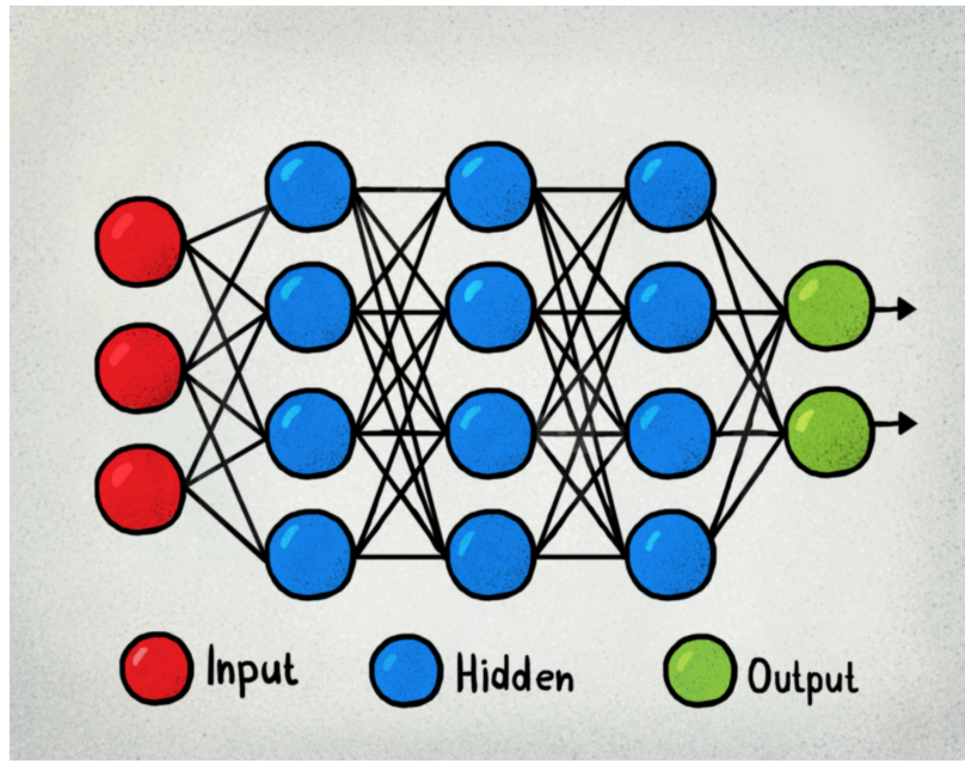
- Initialize weight vector $\boldsymbol{w} = \boldsymbol{0}$
- Loop for T iterations
  - Loop for all training examples $\boldsymbol{x}_n$ (random order!)
  - Predict $\hat{y}_n = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_n)$
  - If $\hat{y}_n \neq y_n$
    - Update: $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta(y_n \boldsymbol{x}_n)$





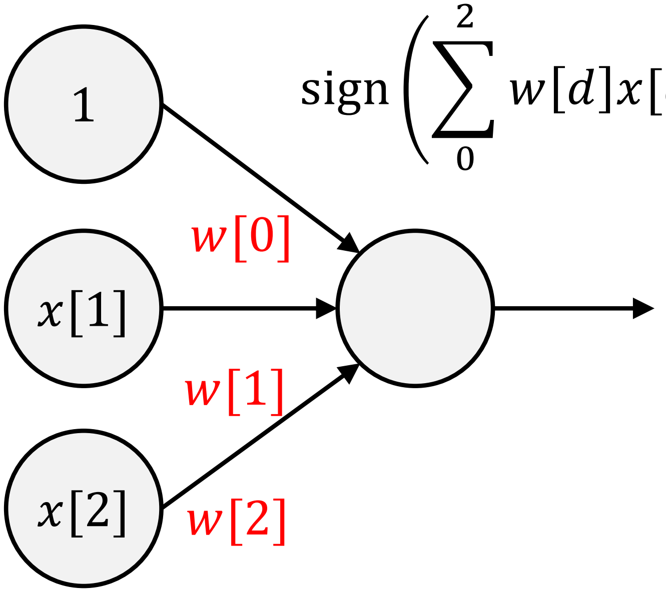
Multi-layer perceptron

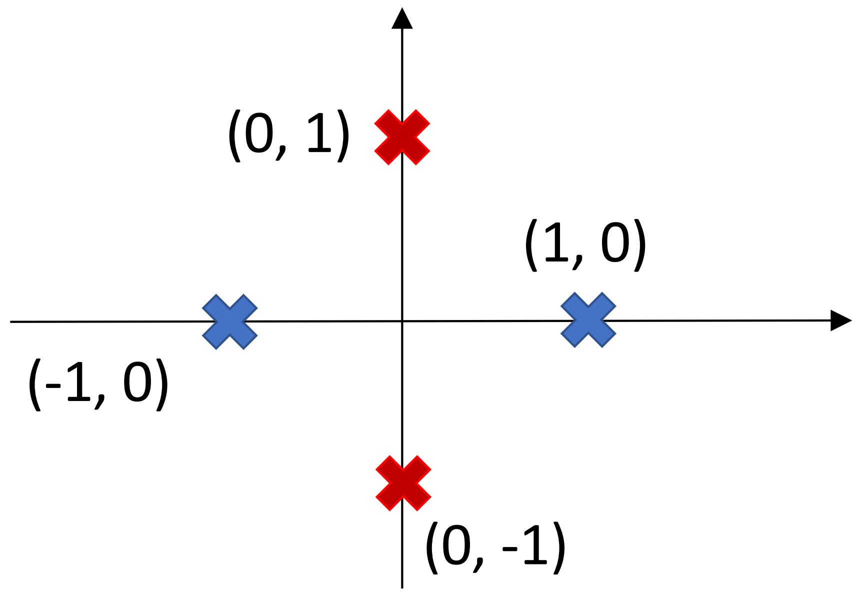
- How to learn?
- Perceptron algorithm?
  - Only look at one instance a time
  - Can update only a single perceptron
  - Not converge for not linear separable data



Why multiple layers?

$$\text{sign}\left(\sum_{0}^{2} w[d]x[d]\right)$$

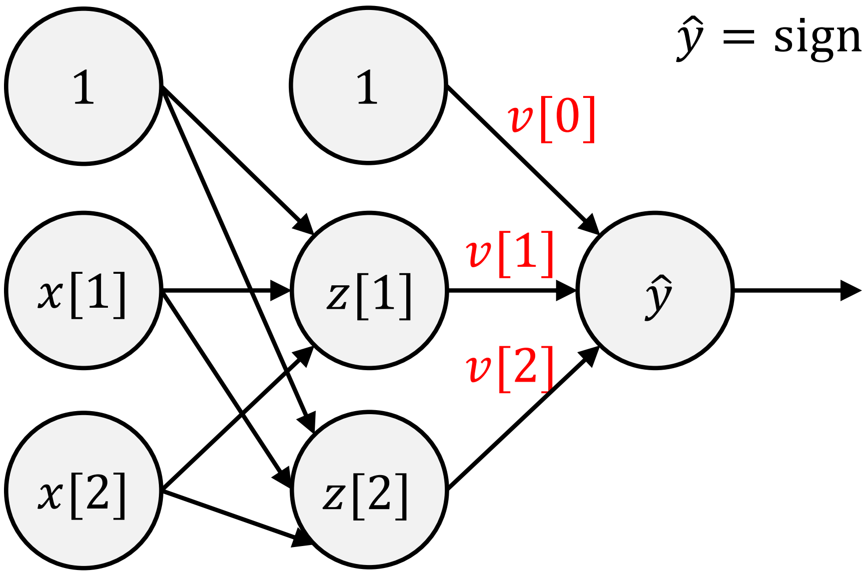


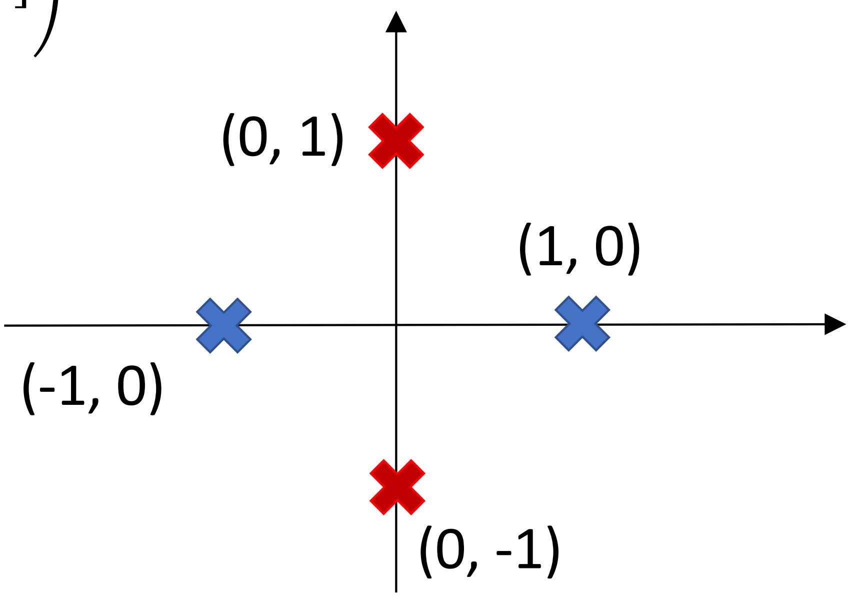


Impossible to classify the 4 points correctly!

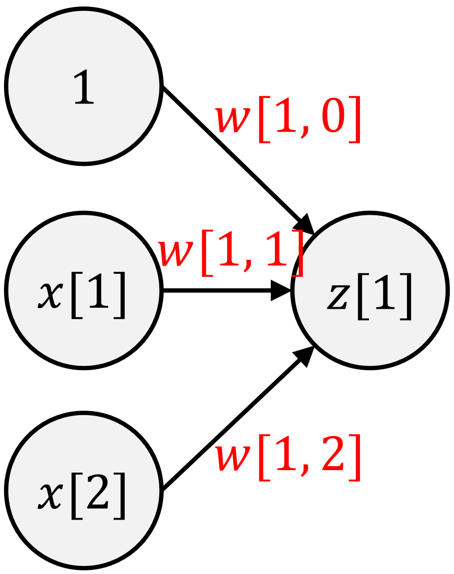
Why multiple layers?

$$\hat{y} = \text{sign}\left(\sum_{0}^{2} v[d]z[d]\right)$$

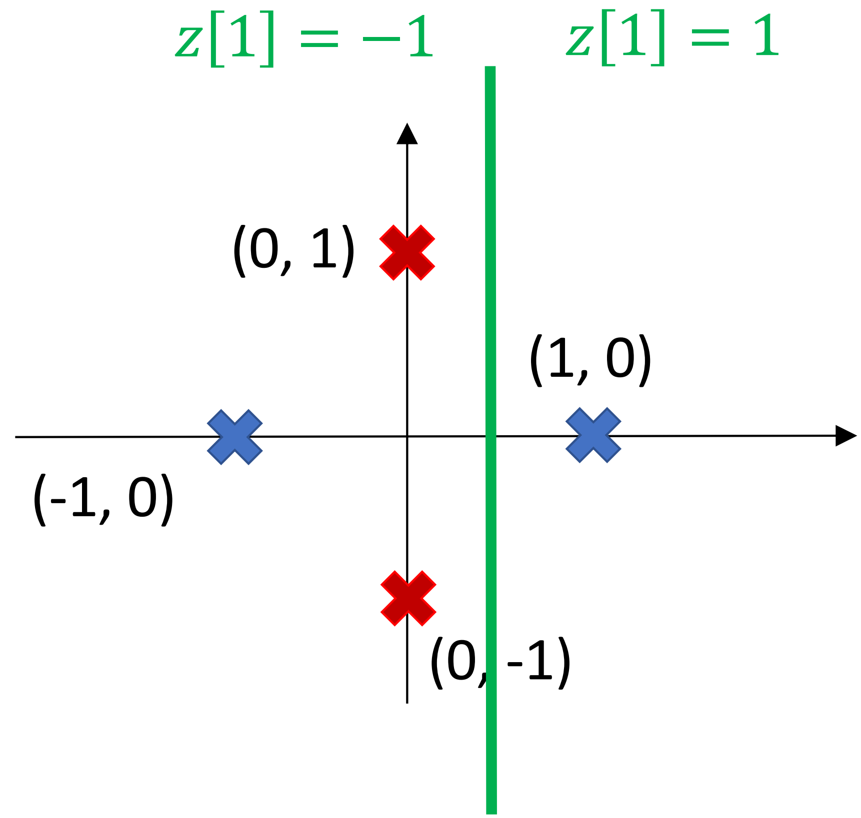




Why multiple layers?

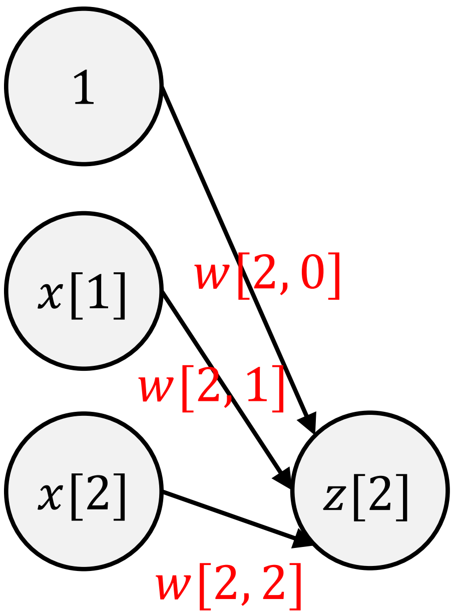


Let
$w[1,0] = -0.5$
$w[1,1] = 1$
$w[1,2] = 0$



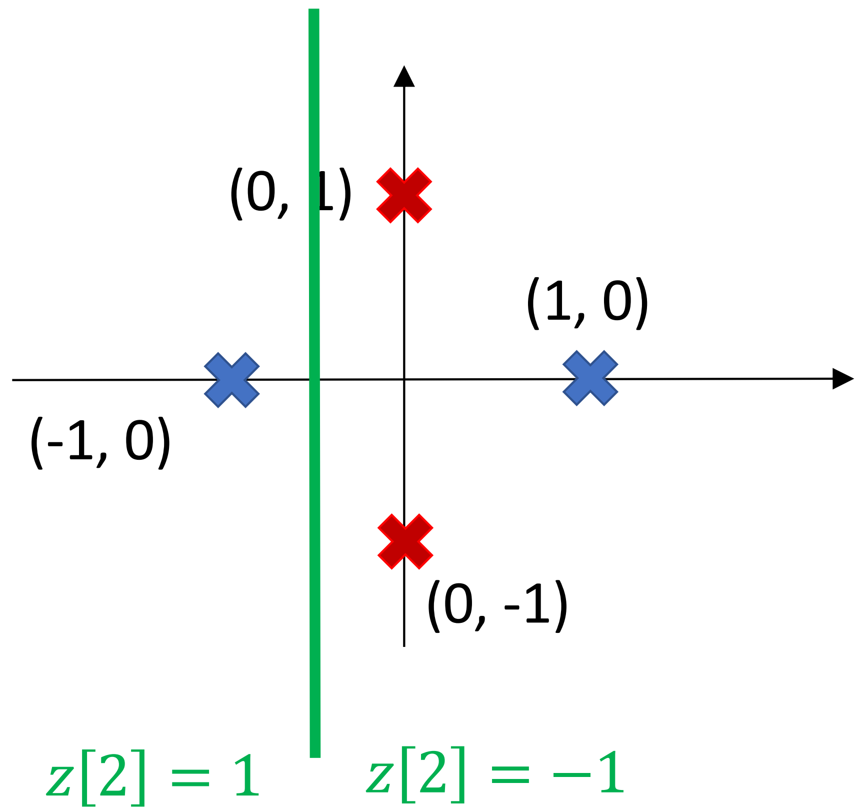
$$z[d'] = \text{sign}\left(\sum_{0}^{2} w[d', d]x[d]\right)$$

Why multiple layers?

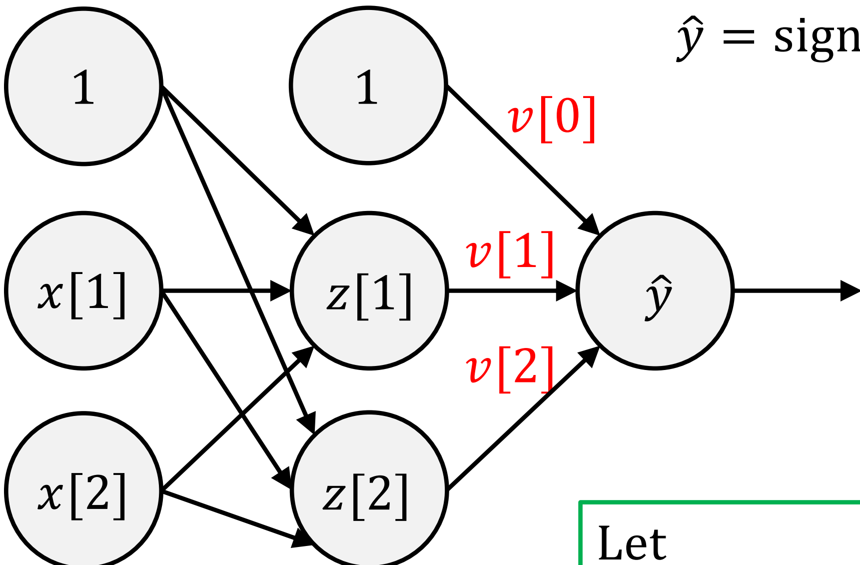


Let
$w[2,0] = -0.5$
$w[2,1] = -1$
$w[2,2] = 0$

$$z[d'] = \operatorname{sign}\left(\sum_{0}^{2} w[d',d]x[d]\right)$$

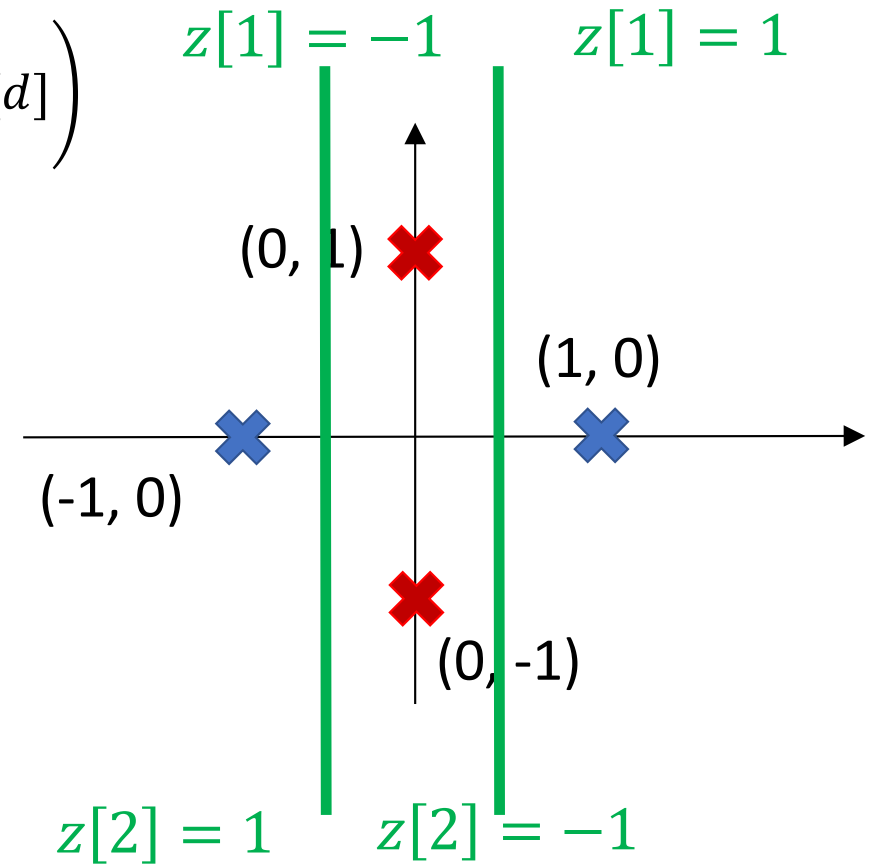


Why multiple layers?



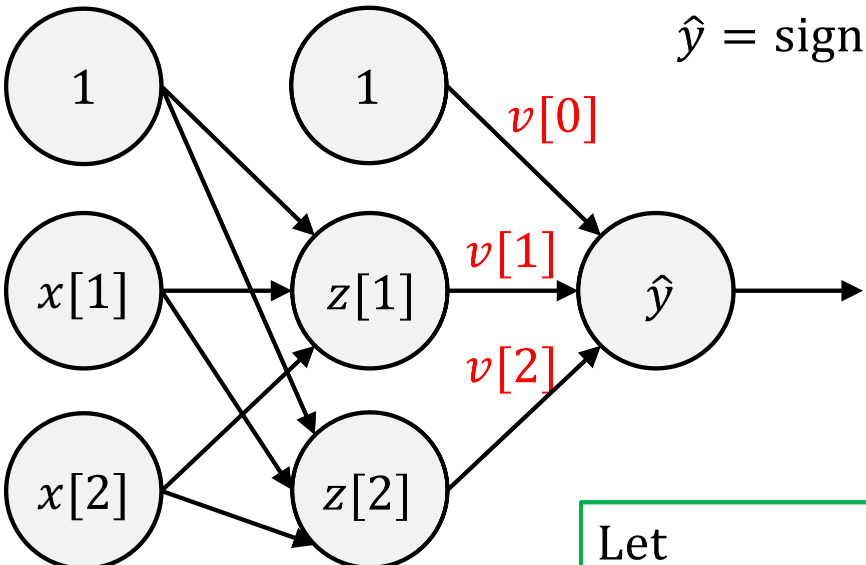
$$\hat{y} = \text{sign}\left(\sum_{0}^{2} v[d]z[d]\right)$$

Let
$v[0] = +0.5$
$v[1] = 1$
$v[2] = 1$

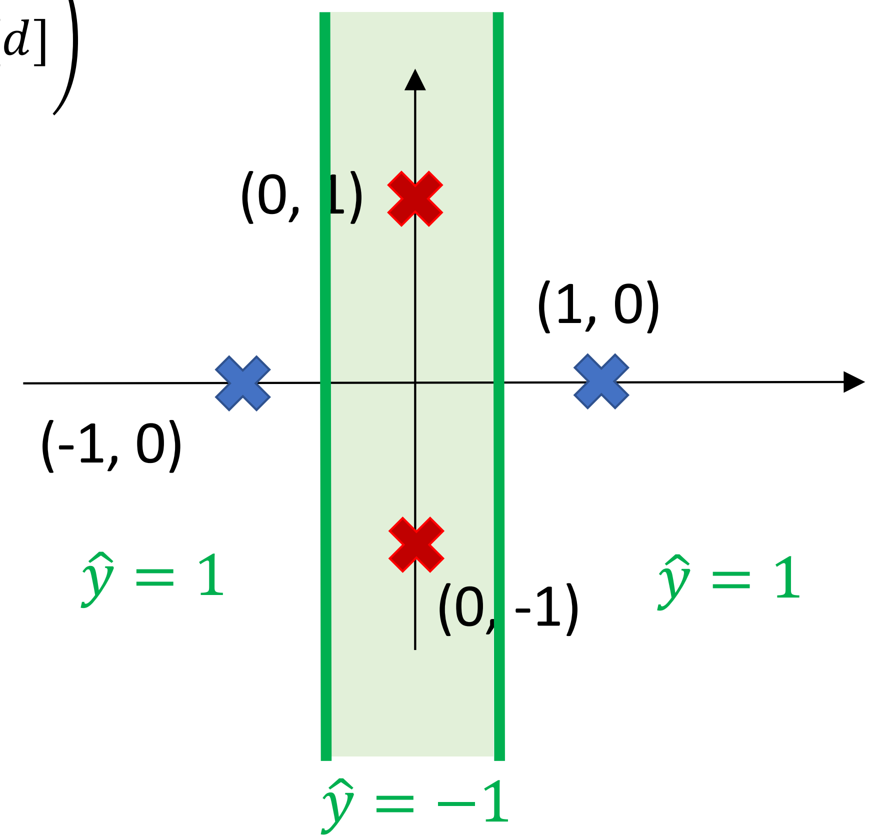


Why multiple layers?

$$\hat{y} = \text{sign}\left(\sum_{0}^{2} v[d]z[d]\right)$$



Let
$v[0] = +0.5$
$v[1] = 1$
$v[2] = 1$



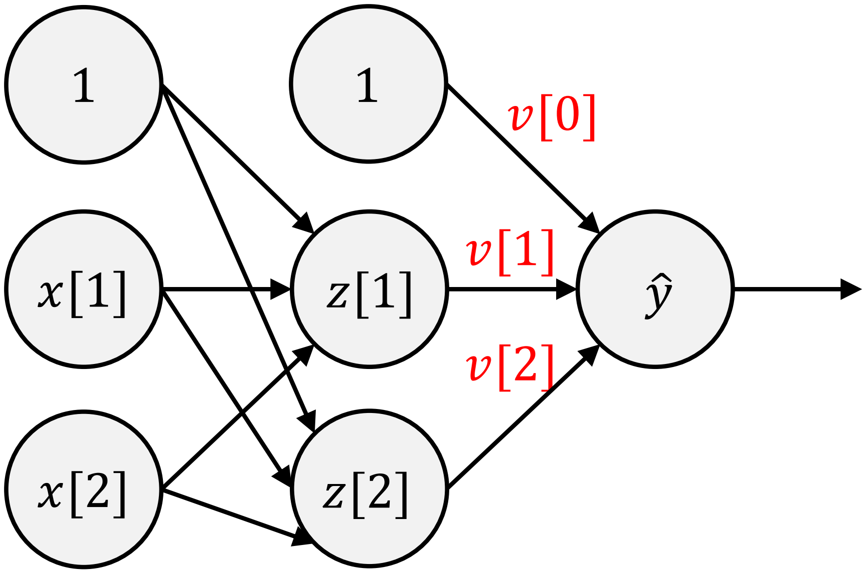
Today

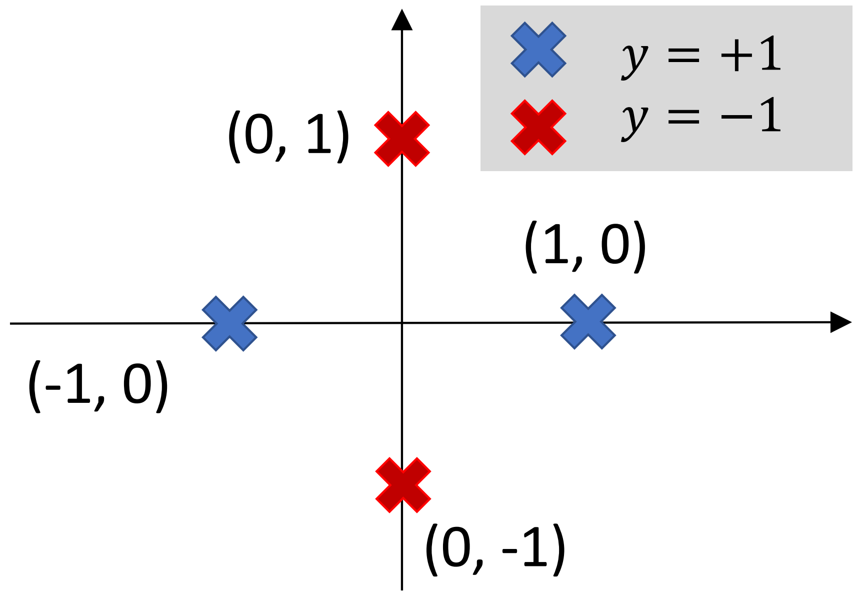
- Learning "deeper" networks beyond one-layer perceptrons
  - Losses and gradients
  - Back-propagation
  - Stochastic gradient descent

- Training particulars
  - Regularization or weight decay

- Discriminative vs. generative models

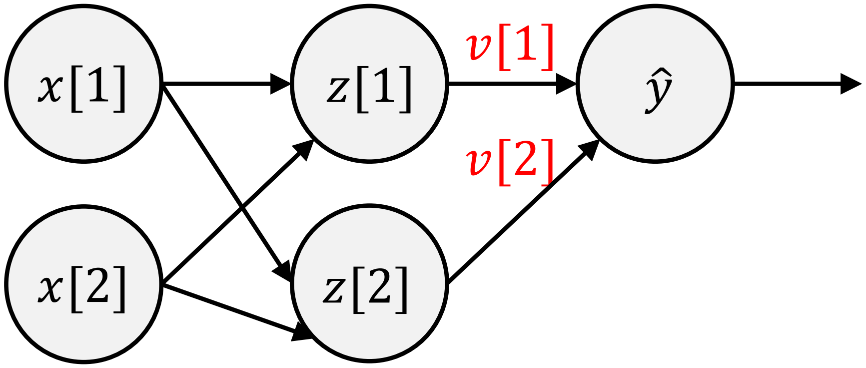
Losses: $\hat{y}$ vs. y

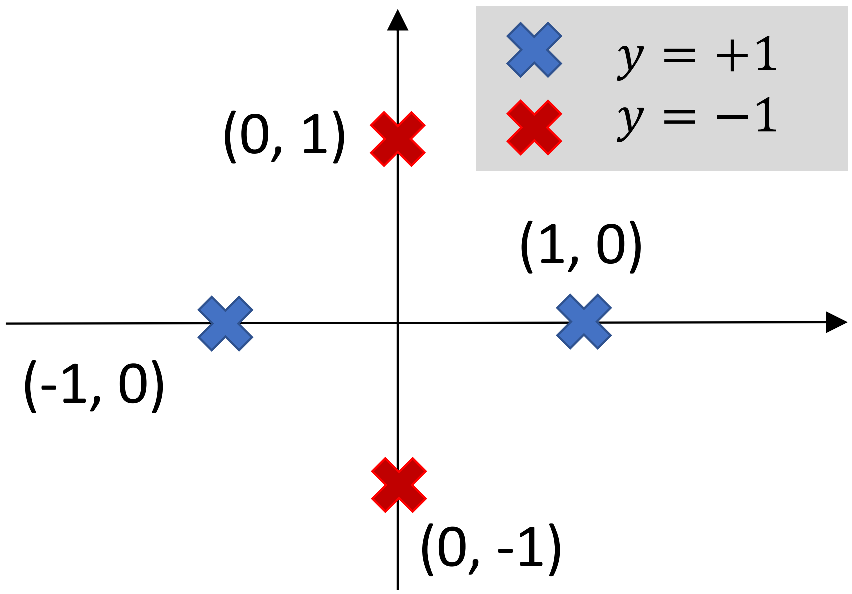




$$\hat{y} = \text{sign}\left(\sum_{0}^{2} v[d]z[d]\right)$$

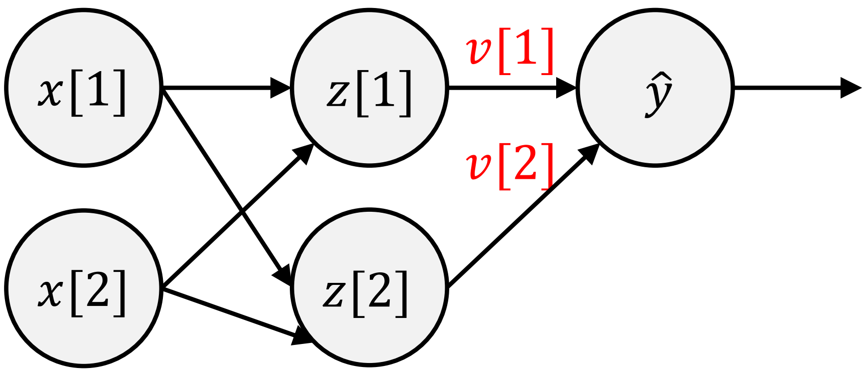
Temporally take out “1” for simplicity

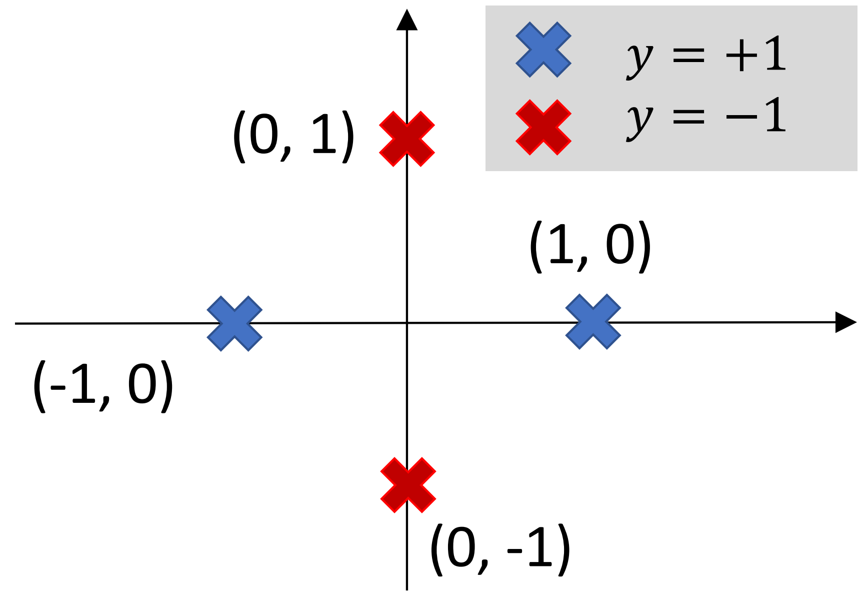




$$\hat{y} = \text{sign}\left(\sum_{1}^{2} v[d]z[d]\right)$$

Re-written with linear algebra

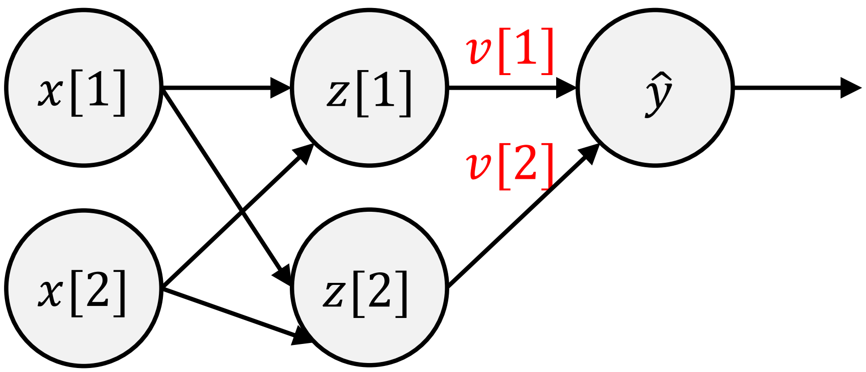


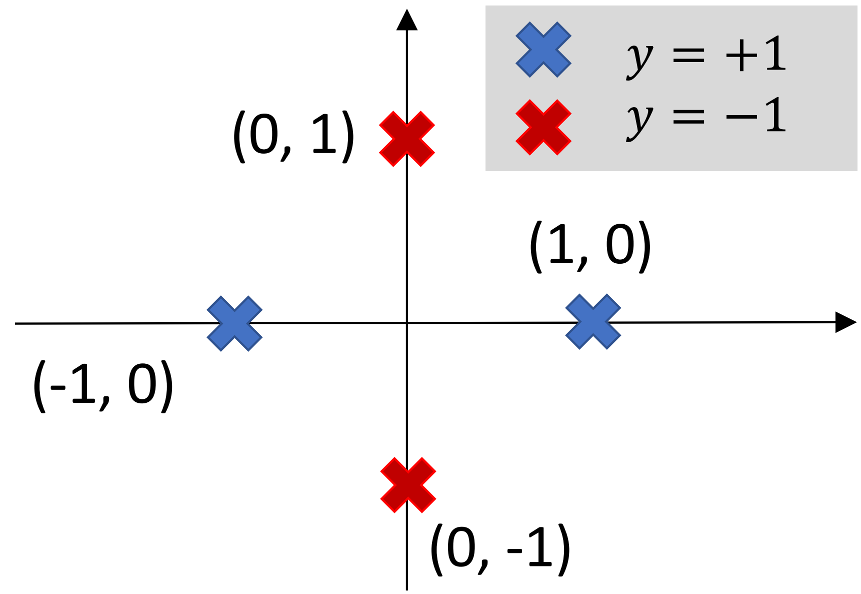


$$\boxed{\hat{y} = \text{sign}\left(\sum_{1}^{2} v[d]z[d]\right) = \text{sign}(\boldsymbol{v}^T\boldsymbol{z}) = \text{sign}(\boldsymbol{v}^T\text{sign}(\boldsymbol{W}\boldsymbol{x}))}$$

$$\begin{bmatrix} z[1] \\ z[2] \end{bmatrix} = \text{sign}\left(\begin{bmatrix} w[1,1] & w[1,2] \\ w[2,1] & w[2,2] \end{bmatrix}\begin{bmatrix} x[1] \\ x[2] \end{bmatrix}\right)$$

Using sigmoid for inner layers





$$\hat{y} = \text{sign}\left(\sum_{1}^{2} v[d]z[d]\right) = \text{sign}(\boldsymbol{v}^T \boldsymbol{z}) = \text{sign}(\boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x}))$$

$$\begin{bmatrix} z[1] \\ z[2] \end{bmatrix} = \rho\left(\begin{bmatrix} w[1,1] & w[1,2] \\ w[2,1] & w[2,2] \end{bmatrix} \begin{bmatrix} x[1] \\ x[2] \end{bmatrix}\right)$$

What are the parameters?

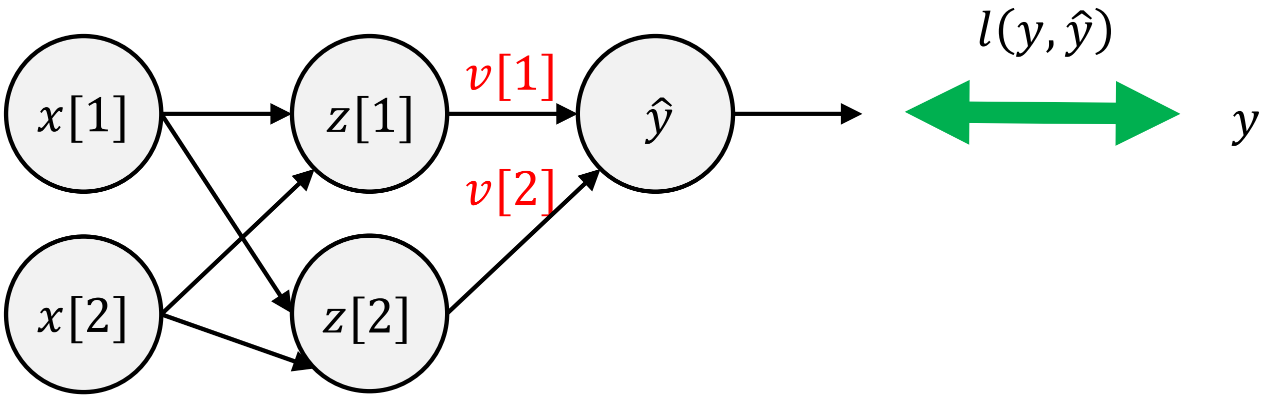
$$\hat{y} = \operatorname{sign}\left(\sum_{1}^{2} v[d]z[d]\right) = \operatorname{sign}(\boldsymbol{v}^T \boldsymbol{z}) = \operatorname{sign}(\boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x}))$$

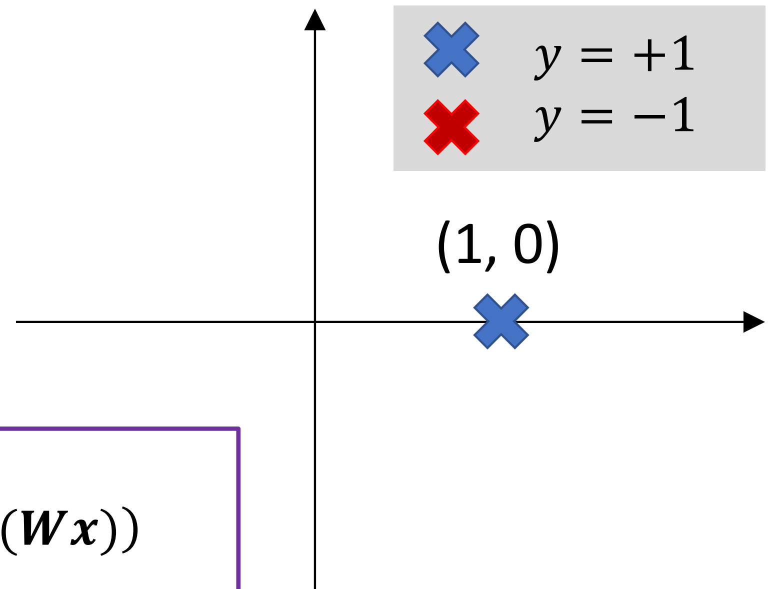
What are the parameters?

$$\hat{y} = \text{sign}\left(\sum_{1}^{2} v[d]z[d]\right) = \text{sign}(\boldsymbol{v}^T\boldsymbol{z}) = \text{sign}(\boldsymbol{v}^T\rho(\boldsymbol{W}\boldsymbol{x}))$$

Answer: $\boldsymbol{v}, \boldsymbol{W}$

Losses and gradients for one data instance





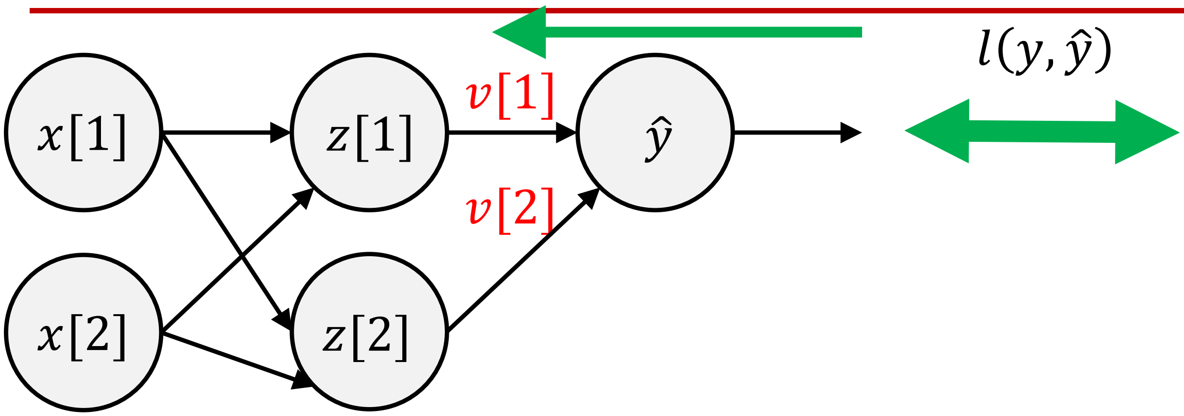
$$\hat{y} = \text{sign}\left(\sum_{1}^{2} v[d]z[d]\right) = \text{sign}(\boldsymbol{v}^T \boldsymbol{z}) = \text{sign}(\boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x}))$$

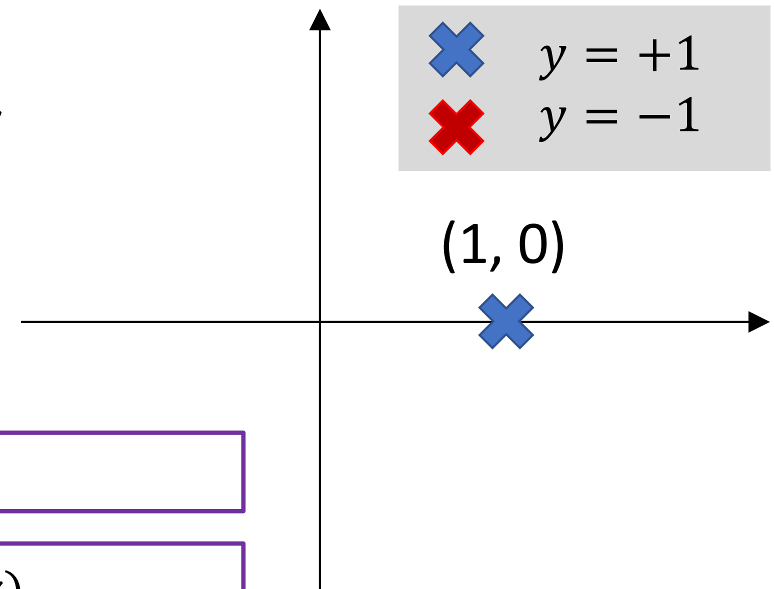
$$l(y, \hat{y}) \approx l(y, \boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x}))$$

For example, logistic loss (binary entropy loss):

$$l\big(y, \boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x})\big) = -y \times \log \rho(\boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x})) - (1 - \mathrm{y}) \times \log(1 - \rho(\boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x})))$$

Losses and gradients for one data instance



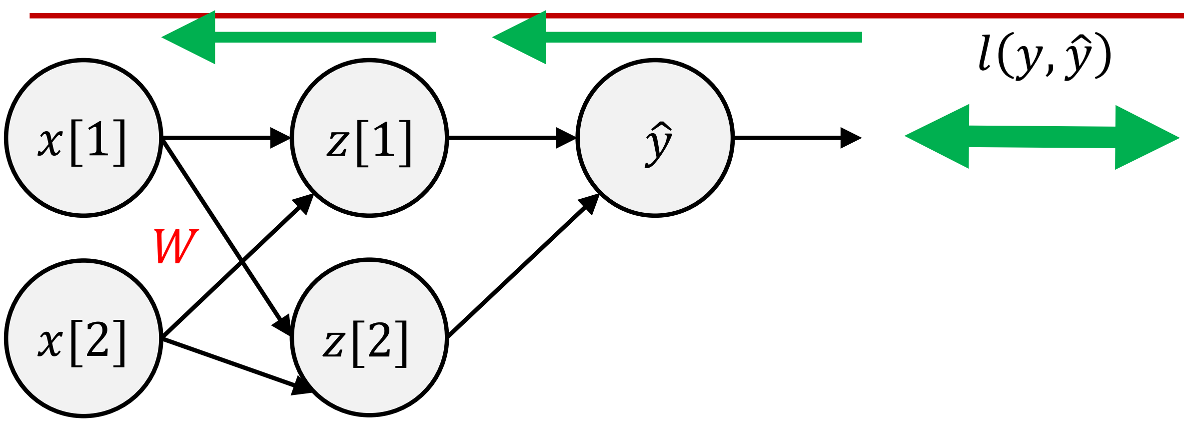


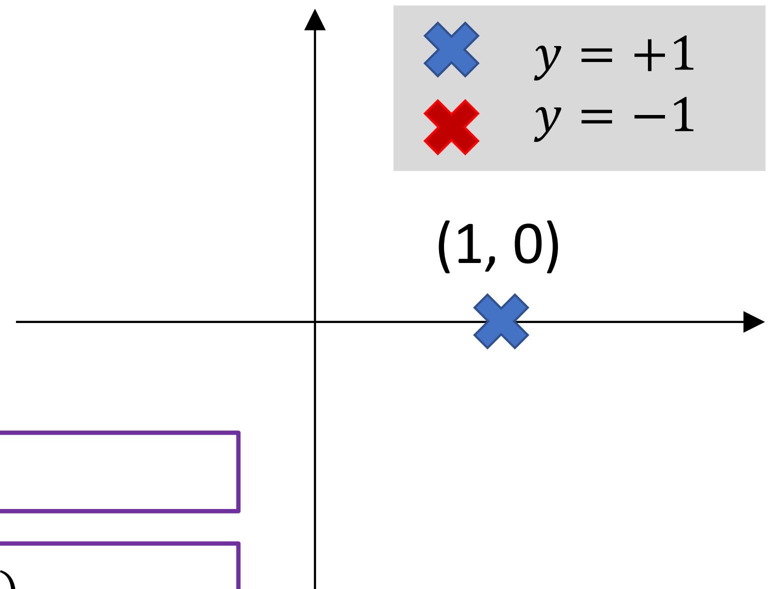
$$l(y, \hat{y}) \approx l(y, \boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x}))$$

$$\nabla_{\boldsymbol{v}} l(y, \hat{y}) = \nabla_{\boldsymbol{v}}\, l\left(y, \boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x})\right) = \nabla_{\boldsymbol{v}}\, l(y, \boldsymbol{v}^T \boldsymbol{z})$$

$$\nabla_{\boldsymbol{W}} l(y, \hat{y})?$$

Losses and gradients for one data instance





$$l(y, \hat{y}) \approx l(y, \boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x}))$$

$$\nabla_{\boldsymbol{v}} l(y, \hat{y}) = \nabla_{\boldsymbol{v}}\, l\left(y, \boldsymbol{v}^T \rho(\boldsymbol{W}\boldsymbol{x})\right) = \nabla_{\boldsymbol{v}}\, l(y, \boldsymbol{v}^T \boldsymbol{z})$$

$\nabla_{\boldsymbol{W}} l(y, \hat{y})$ by chain rules: $\nabla_{\boldsymbol{z}} l(y, \hat{y}) = \begin{bmatrix} \frac{\partial l}{\partial z[1]} \\ \frac{\partial l}{\partial z[2]} \end{bmatrix}$, $\nabla_{\boldsymbol{W}} l(y, \hat{y}) = \sum_{d=1}^{2} \frac{\partial l}{\partial z[d]} \times \nabla_{\boldsymbol{W}} z[d]$

Review chain rules

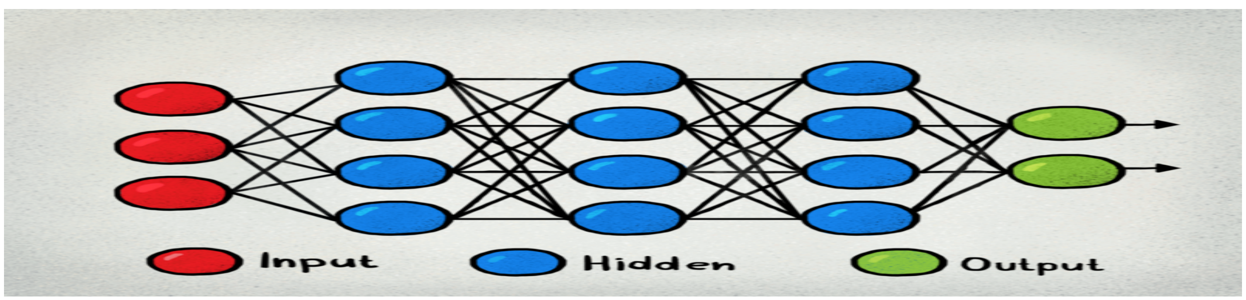
- If $f(x) = A(B(C(x, w)))$: assuming x, w and all functions output a scalar
- $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial A} \times \frac{\partial A}{\partial B} \times \frac{\partial B}{\partial C} \times \frac{\partial C}{\partial x}$
- $\frac{\partial f}{\partial w} = \frac{\partial f}{\partial A} \times \frac{\partial A}{\partial B} \times \frac{\partial B}{\partial C} \times \frac{\partial C}{\partial w}$
- Where
  - $A = A(B(C(x, w)))$
  - $B = B\big(C(x, w)\big)$
  - $C = C(x, w)$

Backpropagation

$$\boldsymbol{x} \rightarrow \boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \boldsymbol{z}^{(3)} \rightarrow \hat{\boldsymbol{y}} \longrightarrow \boldsymbol{y}$$

$$\frac{\partial l}{\partial \boldsymbol{z}^{(3)}} = \frac{\partial l}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial \boldsymbol{z}^{(3)}} \leftarrow \frac{\partial l}{\partial \hat{y}} \leftarrow$$

$$\frac{\partial l}{\partial \boldsymbol{W}^{(4)}} = \frac{\partial l}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial \boldsymbol{W}^{(4)}}$$



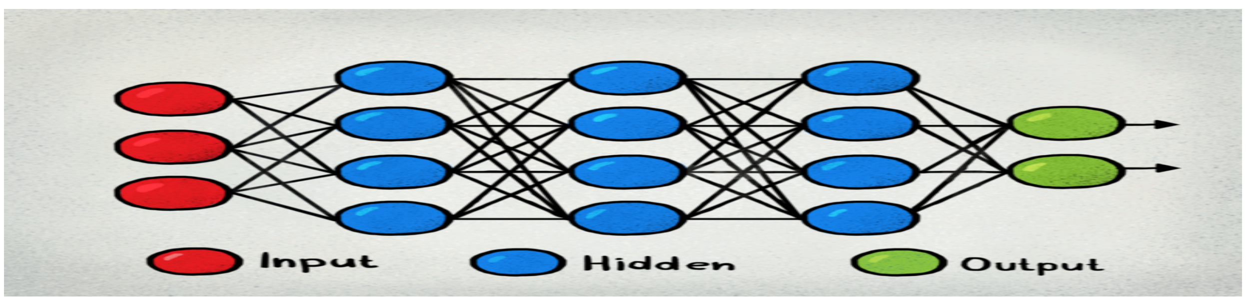
$\boldsymbol{W}^{(1)}$ $\boldsymbol{W}^{(2)}$ $\boldsymbol{W}^{(3)}$ $\boldsymbol{W}^{(4)}$

Backpropagation

$$\boldsymbol{x} \rightarrow \boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \boldsymbol{z}^{(3)} \rightarrow \widehat{\boldsymbol{y}} \rightarrow \boldsymbol{y}$$

$$\frac{\partial l}{\partial \boldsymbol{z}^{(2)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \times \frac{\partial \boldsymbol{z}^{(3)}}{\partial \boldsymbol{z}^{(2)}} \leftarrow \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \leftarrow \frac{\partial l}{\partial \hat{y}} \leftarrow$$

$$\frac{\partial l}{\partial \boldsymbol{W}^{(3)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \times \frac{\partial \boldsymbol{z}^{(3)}}{\partial \boldsymbol{W}^{(3)}} \qquad \frac{\partial l}{\partial \boldsymbol{W}^{(4)}}$$



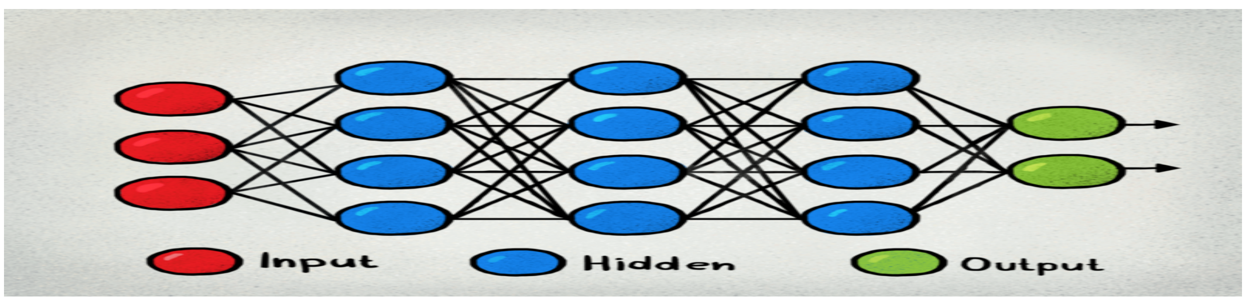
$\boldsymbol{W}^{(1)}$ $\boldsymbol{W}^{(2)}$ $\boldsymbol{W}^{(3)}$ $\boldsymbol{W}^{(4)}$

Backpropagation

$$\boldsymbol{x} \rightarrow \boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \boldsymbol{z}^{(3)} \rightarrow \hat{\boldsymbol{y}} \longrightarrow \boldsymbol{y}$$

$$\frac{\partial l}{\partial \boldsymbol{z}^{(1)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(2)}} \times \frac{\partial \boldsymbol{z}^{(2)}}{\partial \boldsymbol{z}^{(1)}} \leftarrow \frac{\partial l}{\partial \boldsymbol{z}^{(2)}} \leftarrow \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \leftarrow \frac{\partial l}{\partial \hat{y}} \longleftarrow$$

$$\frac{\partial l}{\partial \boldsymbol{W}^{(2)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(2)}} \times \frac{\partial \boldsymbol{z}^{(2)}}{\partial \boldsymbol{W}^{(2)}} \qquad \frac{\partial l}{\partial \boldsymbol{W}^{(3)}} \qquad \frac{\partial l}{\partial \boldsymbol{W}^{(4)}}$$



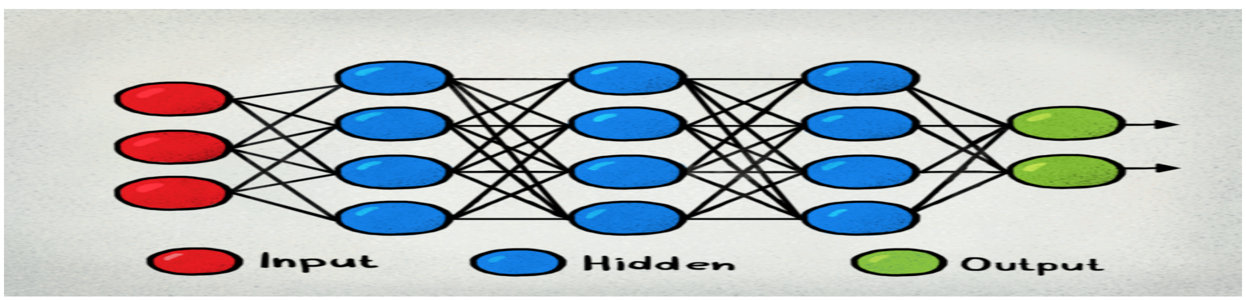
$\boldsymbol{W}^{(1)}$ $\boldsymbol{W}^{(2)}$ $\boldsymbol{W}^{(3)}$ $\boldsymbol{W}^{(4)}$

Backpropagation

$$\boldsymbol{x} \Rightarrow \boldsymbol{z}^{(1)} \Rightarrow \boldsymbol{z}^{(2)} \Rightarrow \boldsymbol{z}^{(3)} \Rightarrow \hat{\boldsymbol{y}} \Rightarrow \boldsymbol{y}$$

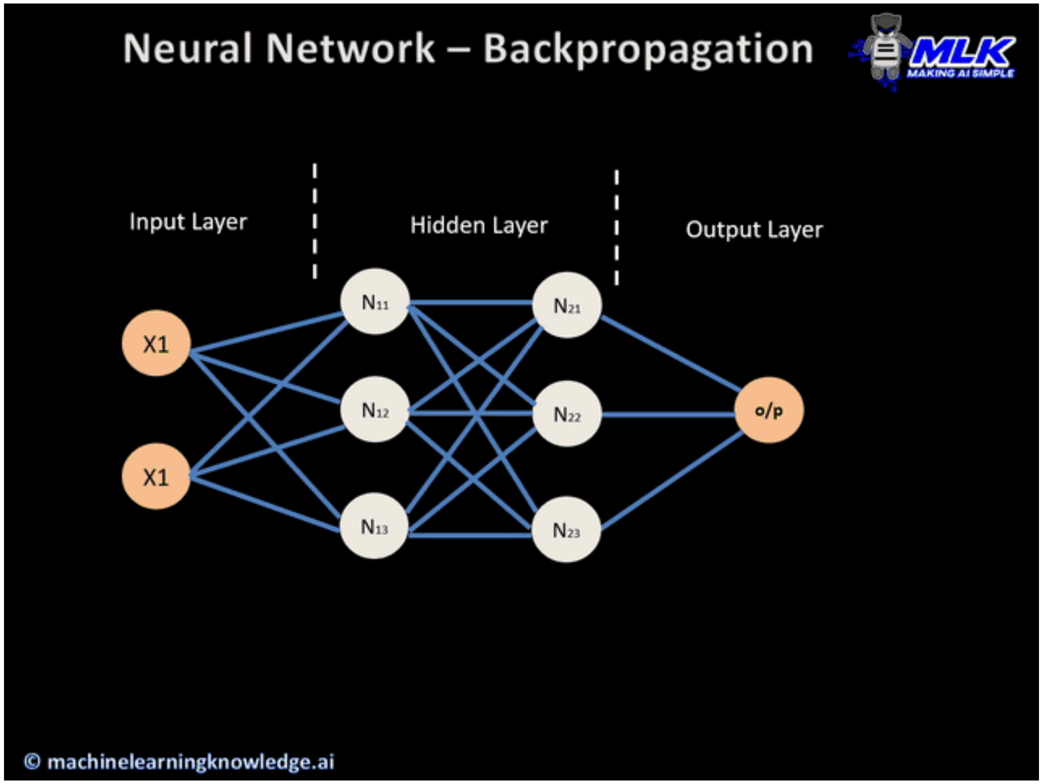
$$\frac{\partial l}{\partial \boldsymbol{z}^{(1)}} \Leftarrow \frac{\partial l}{\partial \boldsymbol{z}^{(2)}} \Leftarrow \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \Leftarrow \frac{\partial l}{\partial \hat{y}} \Leftarrow$$

$$\frac{\partial l}{\partial \boldsymbol{W}^{(1)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(1)}} \times \frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{W}^{(1)}} \qquad \frac{\partial l}{\partial \boldsymbol{W}^{(2)}} \qquad \frac{\partial l}{\partial \boldsymbol{W}^{(3)}} \qquad \frac{\partial l}{\partial \boldsymbol{W}^{(4)}}$$



$\boldsymbol{W}^{(1)}$ $\boldsymbol{W}^{(2)}$ $\boldsymbol{W}^{(3)}$ $\boldsymbol{W}^{(4)}$

Illustration



Backpropagation

- When it is not a scalar:
  - $\frac{\partial l}{\partial \boldsymbol{z}^{(2)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \times \frac{\partial \boldsymbol{z}^{(3)}}{\partial \boldsymbol{z}^{(2)}}$ → $\frac{\partial l}{\partial \boldsymbol{z}^{(2)}} = \sum_{d=1}^{D} \frac{\partial l}{\partial \boldsymbol{z}^{(3)}[\mathrm{d}]} \times \frac{\partial \boldsymbol{z}^{(3)}[\mathrm{d}]}{\partial \boldsymbol{z}^{(2)}}$
  - $\frac{\partial l}{\partial \boldsymbol{W}^{(3)}} = \frac{\partial l}{\partial \boldsymbol{z}^{(3)}} \times \frac{\partial \boldsymbol{z}^{(3)}}{\partial \boldsymbol{W}^{(3)}}$ → $\frac{\partial l}{\partial \boldsymbol{W}^{(3)}} = \sum_{d=1}^{D} \frac{\partial l}{\partial \boldsymbol{z}^{(3)}[\mathrm{d}]} \times \frac{\partial \boldsymbol{z}^{(3)}[\mathrm{d}]}{\partial \boldsymbol{W}^{(3)}}$

Training a neural network: gradient descent

- Let $\boldsymbol{x} \in \mathbb{R}^D$; y as true label, and $\{(\boldsymbol{x}_n, y_n)\}$ as the training data
- $\boldsymbol{\theta}$ as all parameters, $\hat{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as the neural network's prediction

- Initialize $\boldsymbol{\theta}$ with [with some specific methods]

- Loop for T "epochs"
  - $\nabla_{\boldsymbol{\theta}} L = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} l\,(y_n, \hat{y}_n)$
  - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \times \nabla_{\boldsymbol{\theta}} L$

Stochastic gradient descent

- Let $\boldsymbol{x} \in \mathbb{R}^D$; y as true label, and $\{(\boldsymbol{x}_n, y_n)\}$ as the training data
- $\boldsymbol{\theta}$ as all parameters, $\hat{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as the neural network's prediction

- Initialize $\boldsymbol{\theta}$ with [with some specific methods]

- Loop for T "epochs"
  - Loop for all training examples $\boldsymbol{x}_n$ (random order!)
  - $\nabla_{\boldsymbol{\theta}} L = \nabla_{\boldsymbol{\theta}} l(y_n, \hat{y}_n)$
  - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \times \nabla_{\boldsymbol{\theta}} L$

"Mini-batch" Stochastic gradient descent

- Let $\boldsymbol{x} \in \mathbb{R}^D$; y as true label, and $\{(\boldsymbol{x}_n, y_n)\}$ as the training data
- $\boldsymbol{\theta}$ as all parameters, $\hat{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as the neural network's prediction

- Initialize $\boldsymbol{\theta}$ with [with some specific methods]

- Loop for T "epochs"
  - Loop for "B sampled examples from $\{(\boldsymbol{x}_n, y_n)\}$" (called batch) without replacement (random order!)
  - $\nabla_{\boldsymbol{\theta}} L = \frac{1}{N} \sum_{b=1}^{B} \nabla_{\boldsymbol{\theta}} l\,(y_b, \hat{y}_b)$
  - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \times \nabla_{\boldsymbol{\theta}} L$