10-701: Introduction to Machine Learning Lecture 12 – Neural Networks

Henry Chai & Zack Lipton

10/09/23

Front Matter

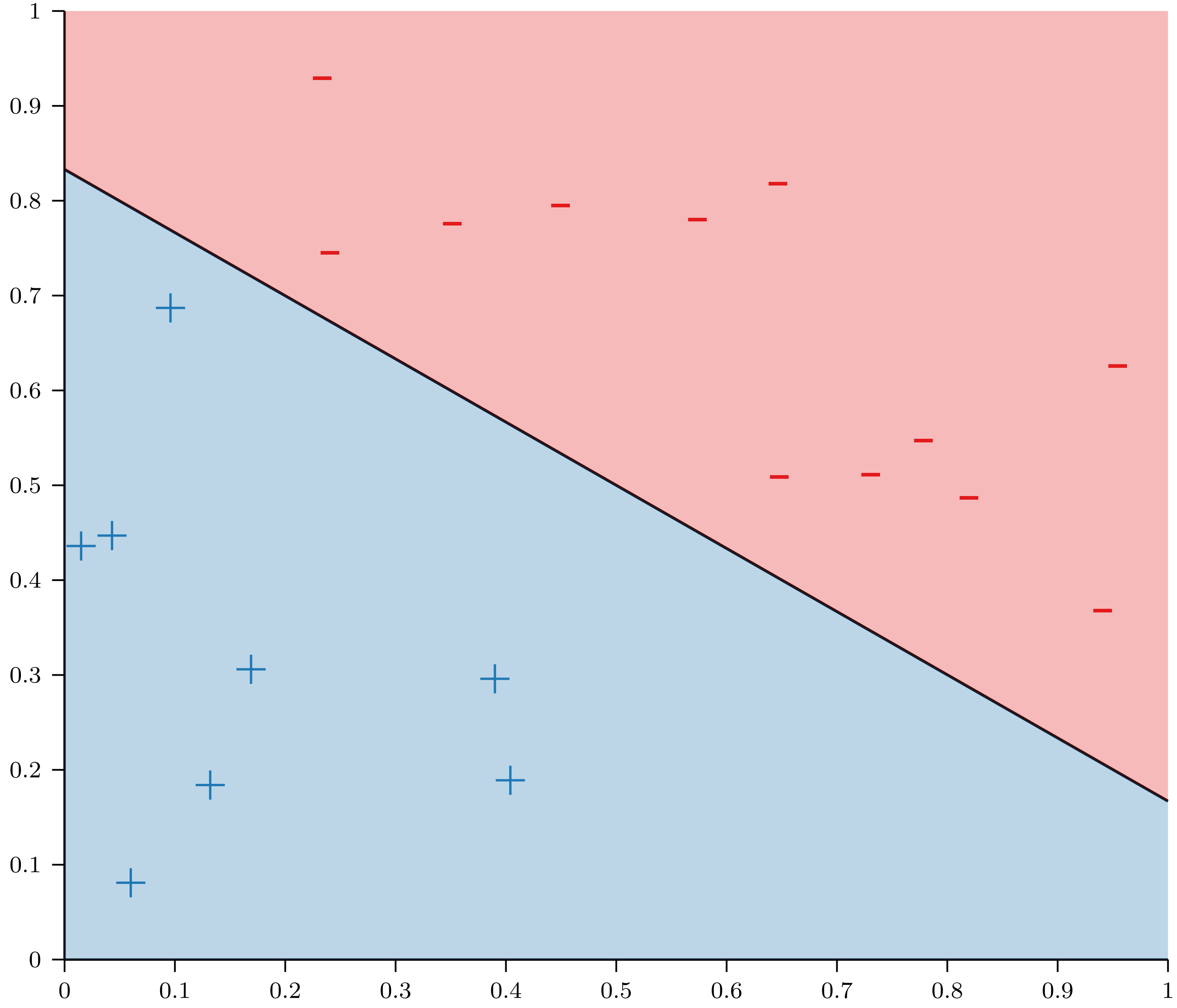
- Announcements
  - HW3 released 10/4, due 10/11 (Wednesday) at 11:59 PM
  - HW4 released 10/11 (Wednesday), due 10/25 (**after fall break**) at 11:59 PM
  - Project details will be released on 10/13 (Friday)
  - Midterm exam on 10/31 from 6:30 – 8:30 PM
    - If you have a conflict with this date/time fill out the conflict on Piazza ASAP
- Recommended Readings
  - Mitchell, Chapters 4.1 – 4.6
  - Zhang, Lipton, Li & Smola, Chapters 5.1 – 5.3

Biological Neural Network

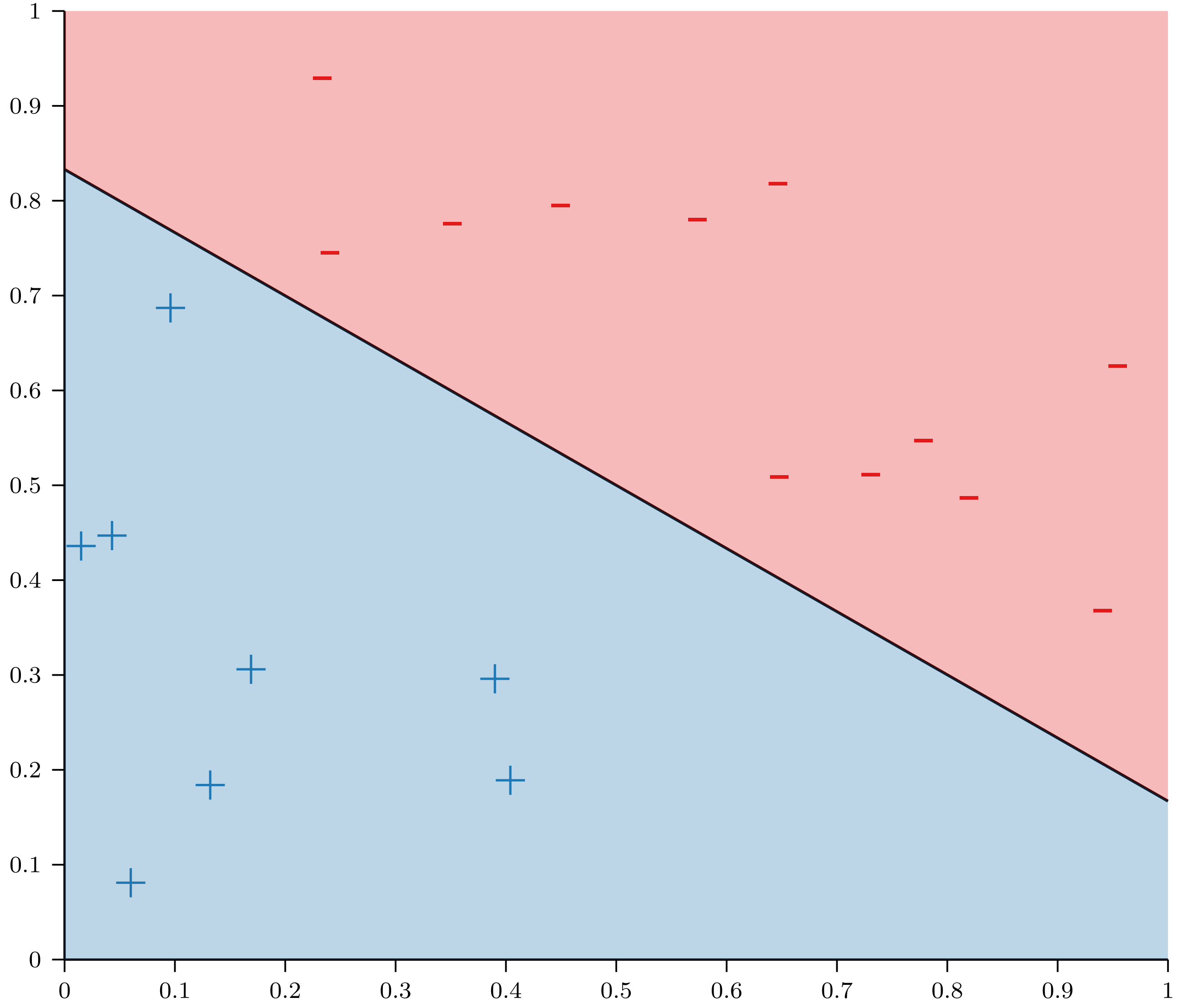
Source: https://science-art.com/image/?id=2971&m=168&pagename=neural_network_3d#.W_wrzJNKhUM

Biological Neural Network(s)

Recall: Linear Models



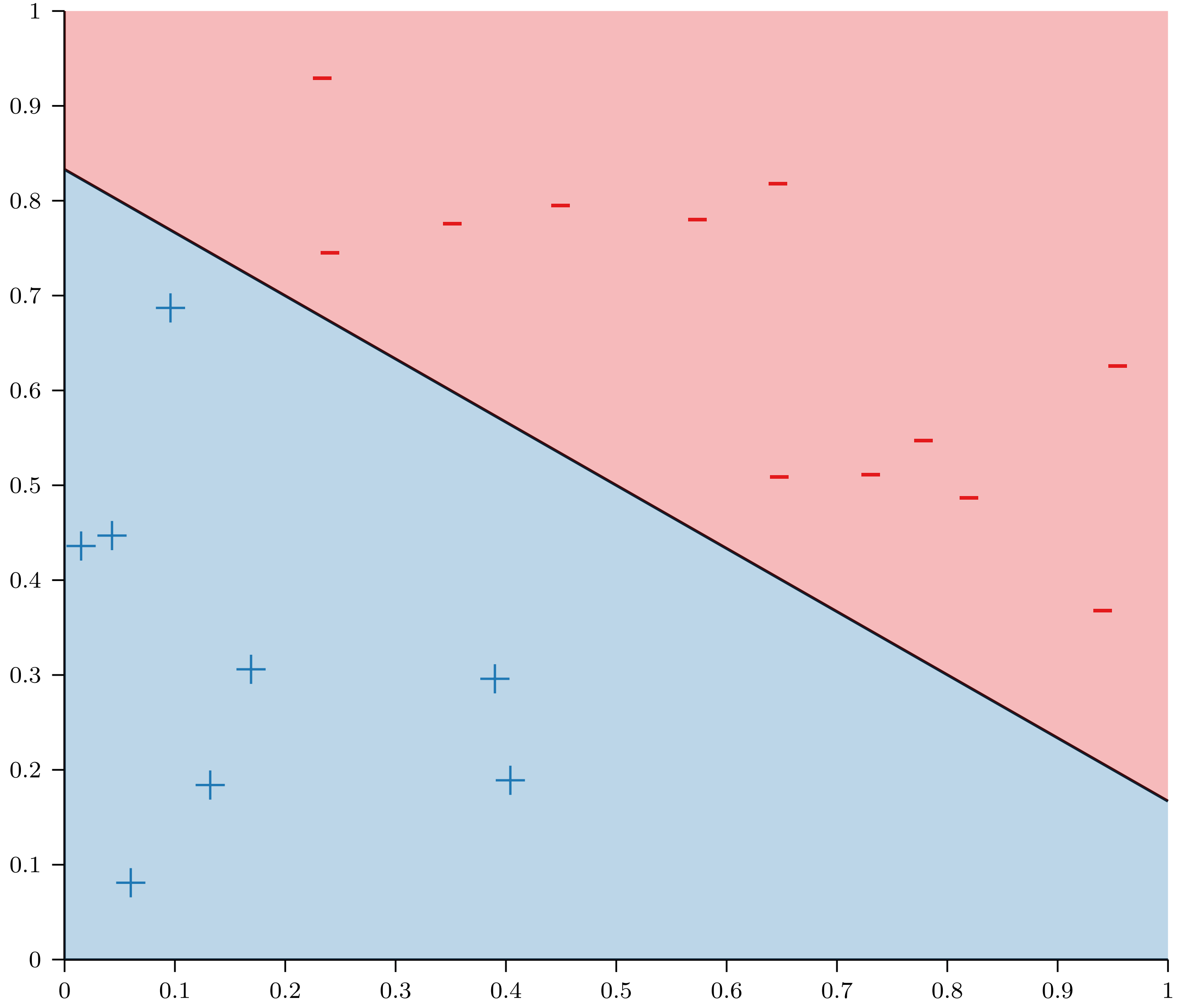
Where do linear decision boundaries come from?



The equation of a line is

$$\boldsymbol{w}^T\boldsymbol{x} = b \rightarrow w^T x - b = 0$$

$$w'^T \begin{bmatrix} 1 \\ x \end{bmatrix} = 0$$



The equation of a line is

$\boldsymbol{w}^T\boldsymbol{x} = 0$

(bias term prepended to $\boldsymbol{w}$)

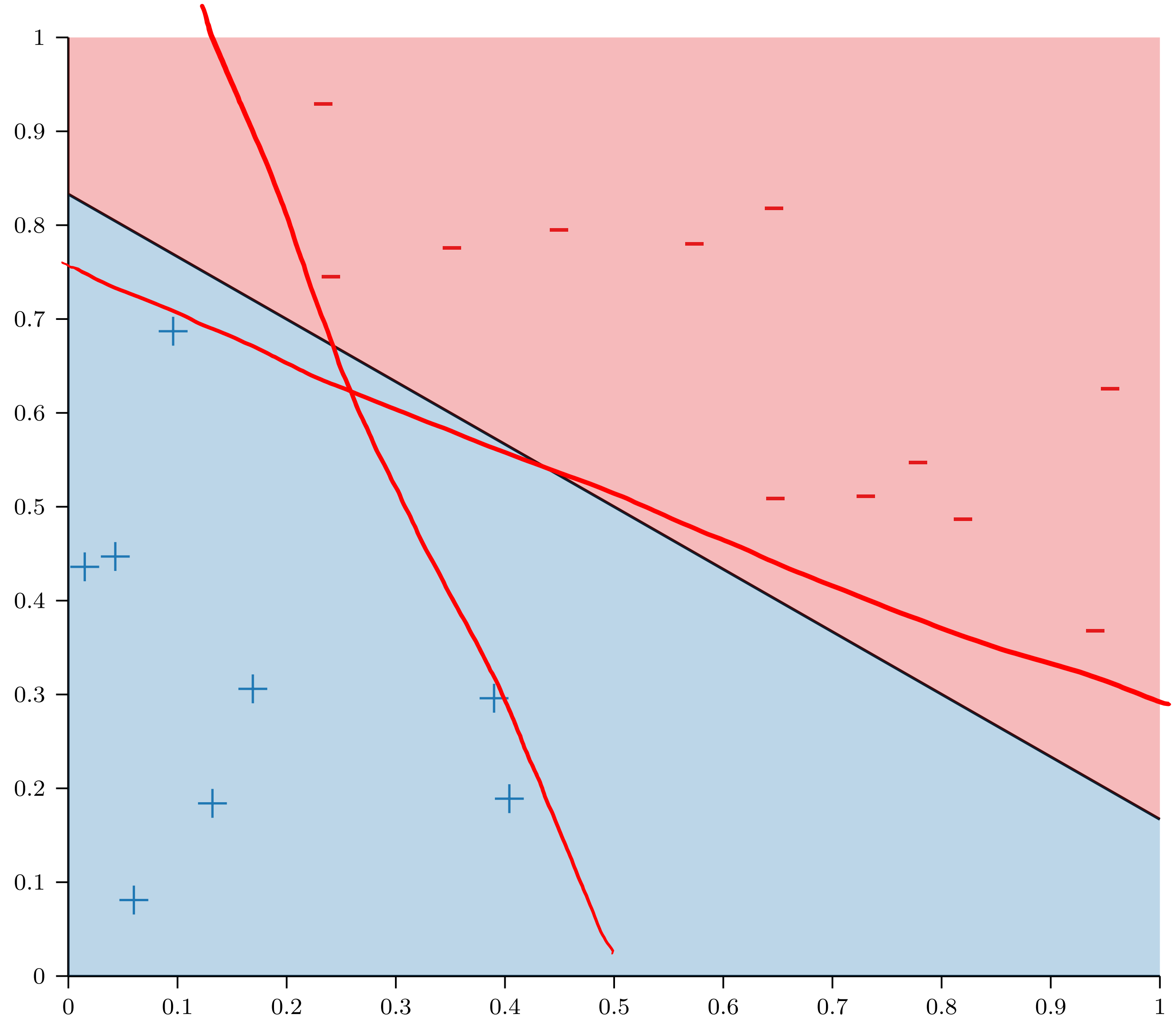
The line defines two half-spaces in $\mathbb{R}^D$:

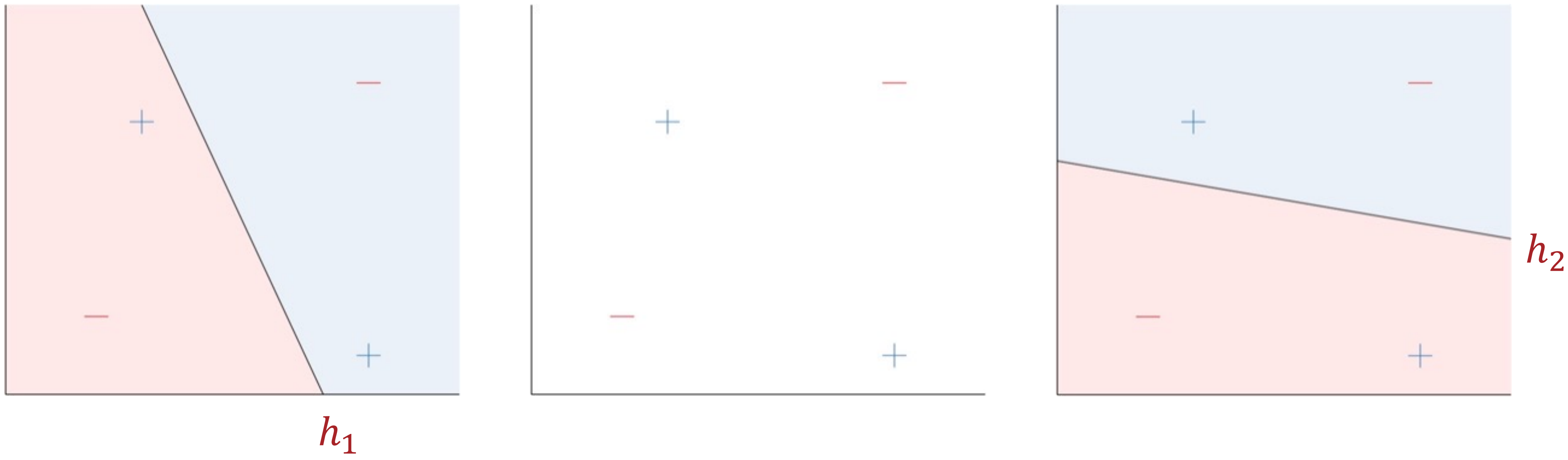
- $\mathcal{S}_+ = \{\boldsymbol{x}\colon \boldsymbol{w}^T\boldsymbol{x} > 0\}$
- $\mathcal{S}_- = \{\boldsymbol{x}\colon \boldsymbol{w}^T\boldsymbol{x} < 0\}$

So the model

$$h(x) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x})$$

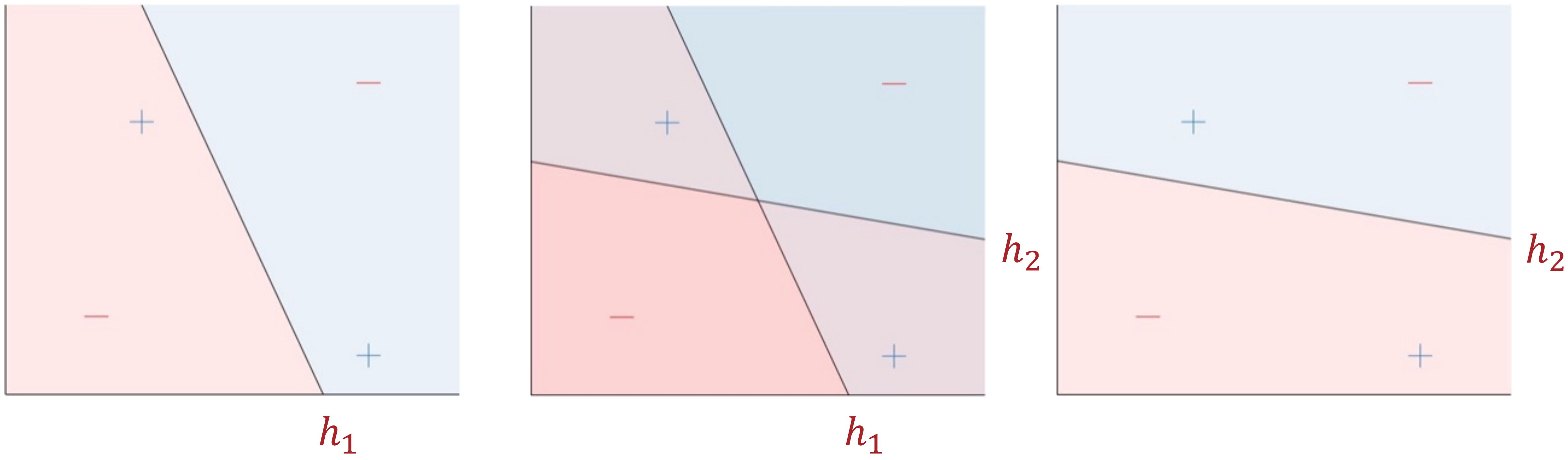
gives rise to linear decision boundaries!



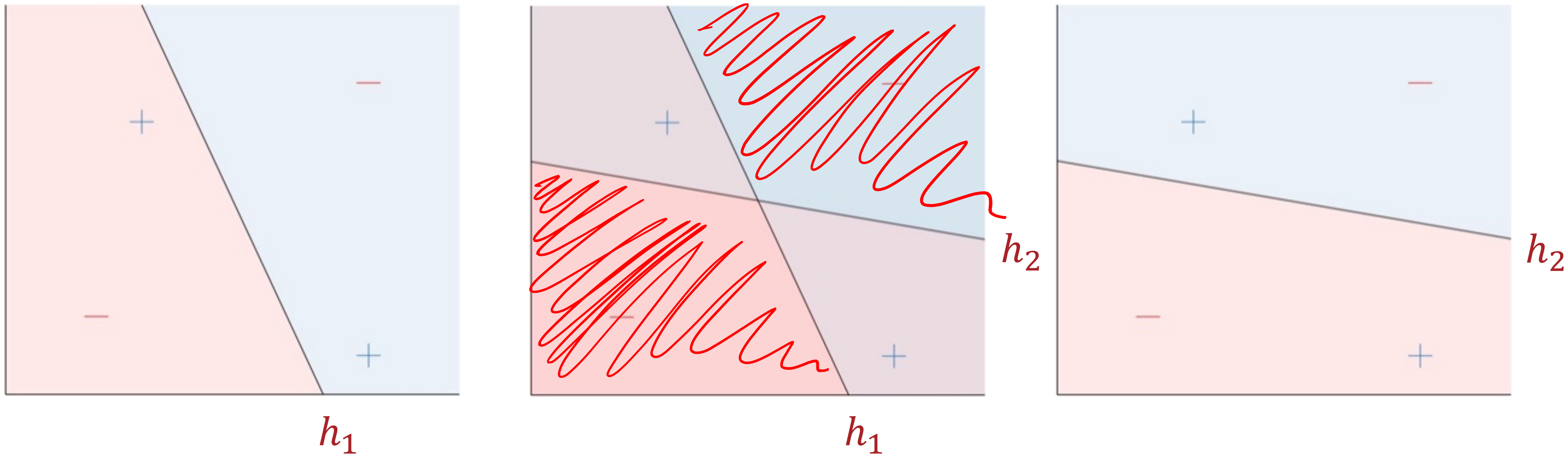


Perceptrons

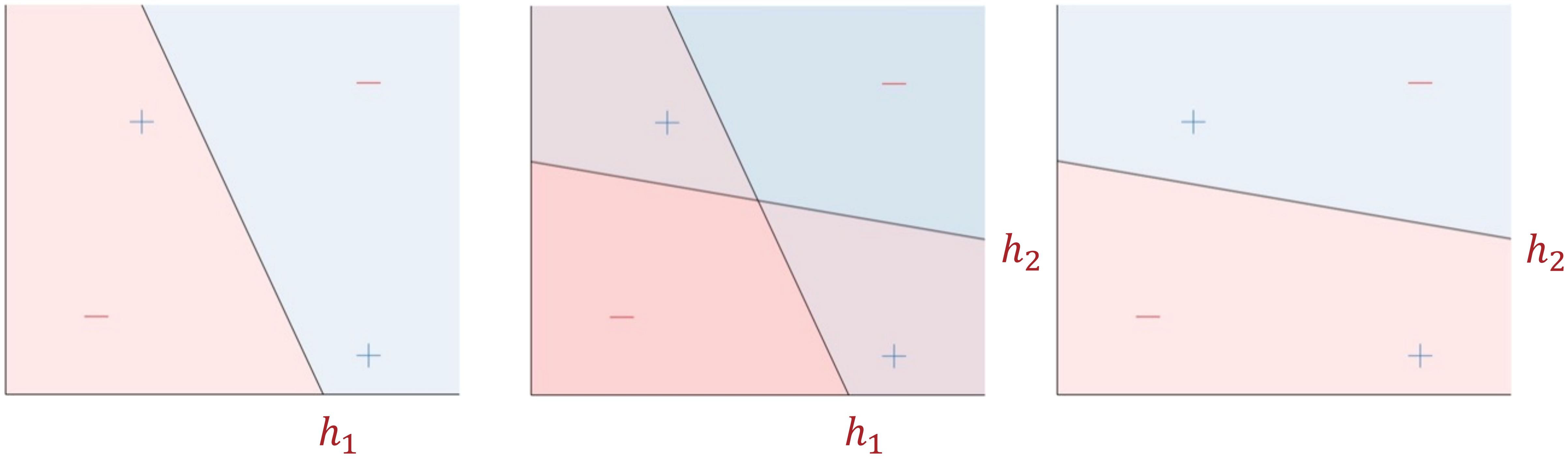
- Linear model for classification
- $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x})$
- Predictions are $+1$ or -1



Combining Perceptrons



$$h(\boldsymbol{x}) = \begin{cases} +1 \text{ if } (h_1(\boldsymbol{x}) = +1 \text{ and } h_2(\boldsymbol{x}) = -1) \text{ or } (h_1(\boldsymbol{x}) = -1 \text{ and } h_2(\boldsymbol{x}) = +1) \\ -1 \text{ otherwise} \end{cases}$$



$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x}))\right)$$

Boolean Algebra

- Boolean variables are either $+1$ ("true") or -1 ("false")
- Basic Boolean operations
  - Negation: $\neg z = -1 * z$
  - And: $AND(z_1, z_2) = \begin{cases} +1 \text{ if both } z_1 \text{ and } z_2 \text{ equal } + 1 \\ -1 \text{ otherwise} \end{cases}$
  - Or: $OR(z_1, z_2) = \begin{cases} +1 \text{ if either } z_1 \text{ or } z_2 \text{ equals } + 1 \\ -1 \text{ otherwise} \end{cases}$

Boolean Algebra

- Boolean variables are either $+1$ (”true”) or -1 (“false”)
- Basic Boolean operations
  - Negation: $\neg z = -1 * z$
  - And: $AND(z_1, z_2) = \text{sign}(z_1 + z_2 - 1.5)$
  - Or: $OR(z_1, z_2) = \text{sign}(z_1 + z_2 + 1.5)$

Boolean Algebra

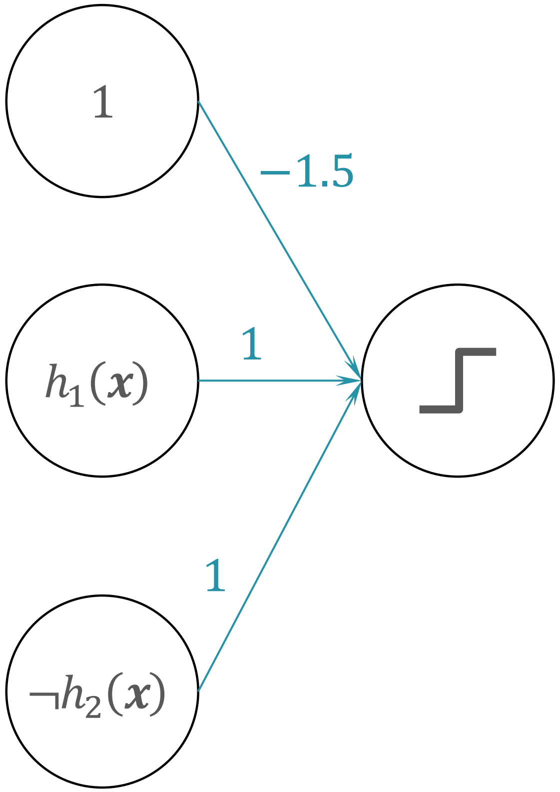
- Boolean variables are either $+1$ (”true”) or -1 (“false”)
- Basic Boolean operations
  - Negation: $\neg z = -1 * z$
  - And: $AND(z_1, z_2) = \text{sign}\left([-1.5, 1, 1] \begin{bmatrix} 1 \\ z_1 \\ z_2 \end{bmatrix} \right)$
  - Or: $OR(z_1, z_2) = \text{sign}\left(\underline{[1.5, 1, 1]} \begin{bmatrix} 1 \\ z_1 \\ z_2 \end{bmatrix} \right)$

Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$

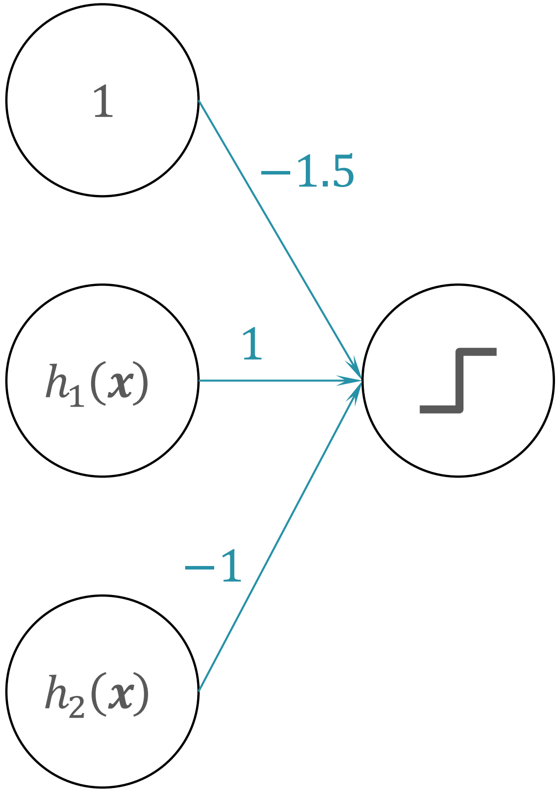
Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



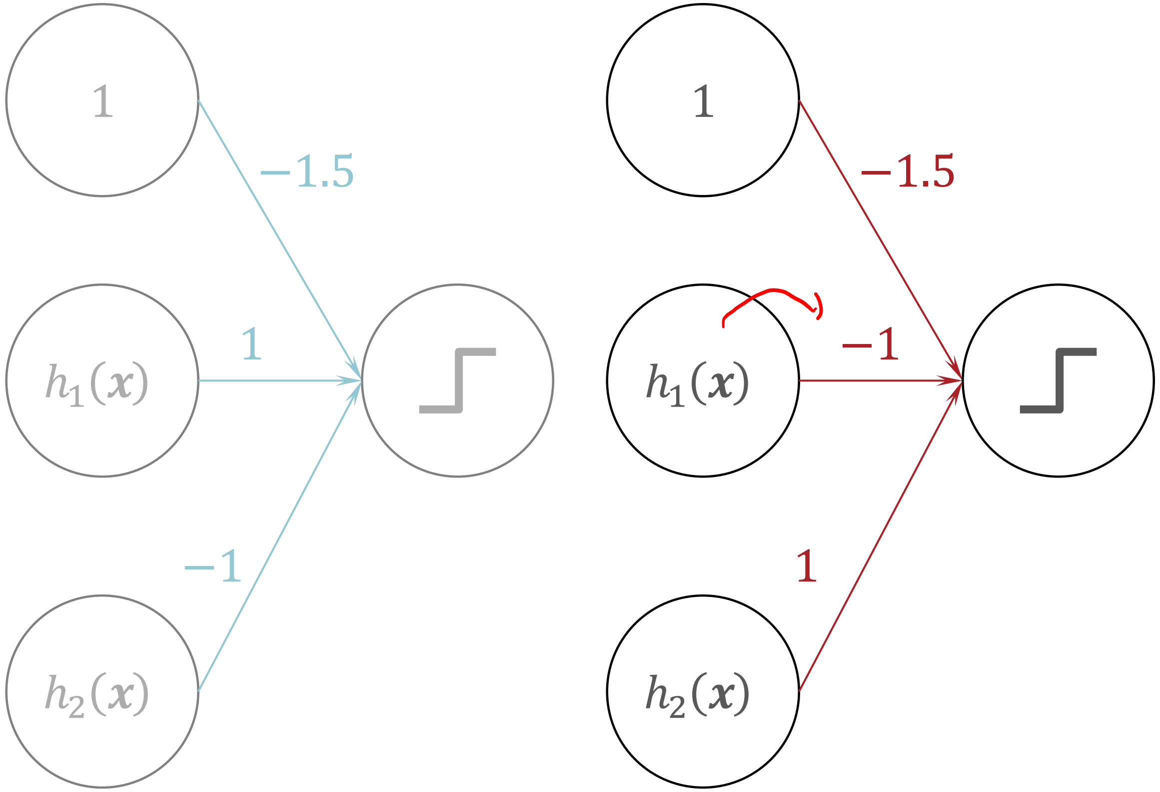
Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



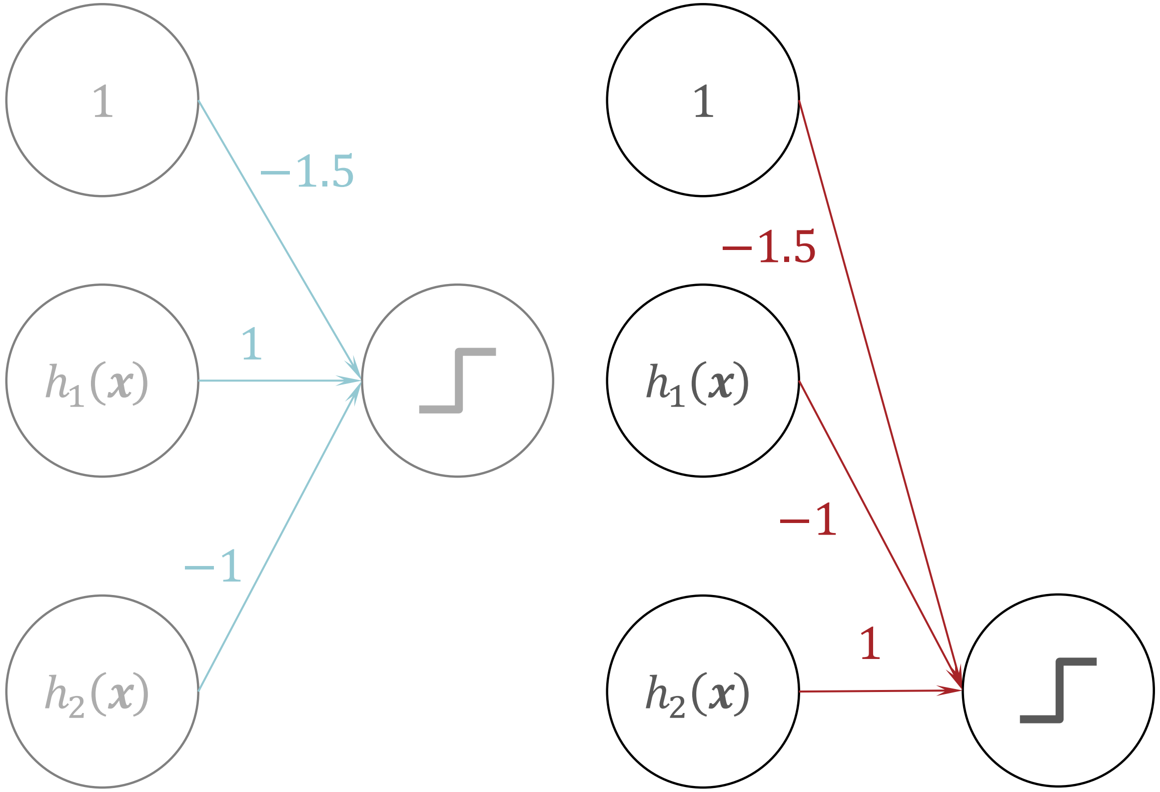
Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



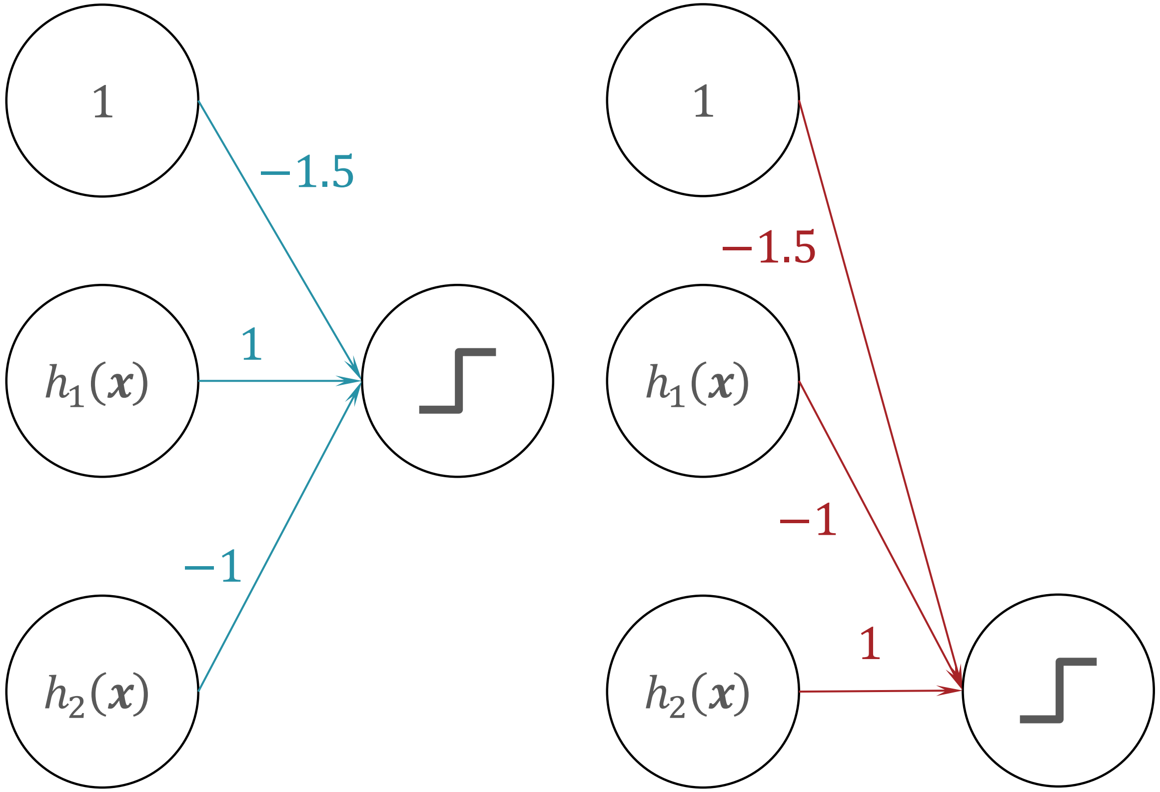
Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



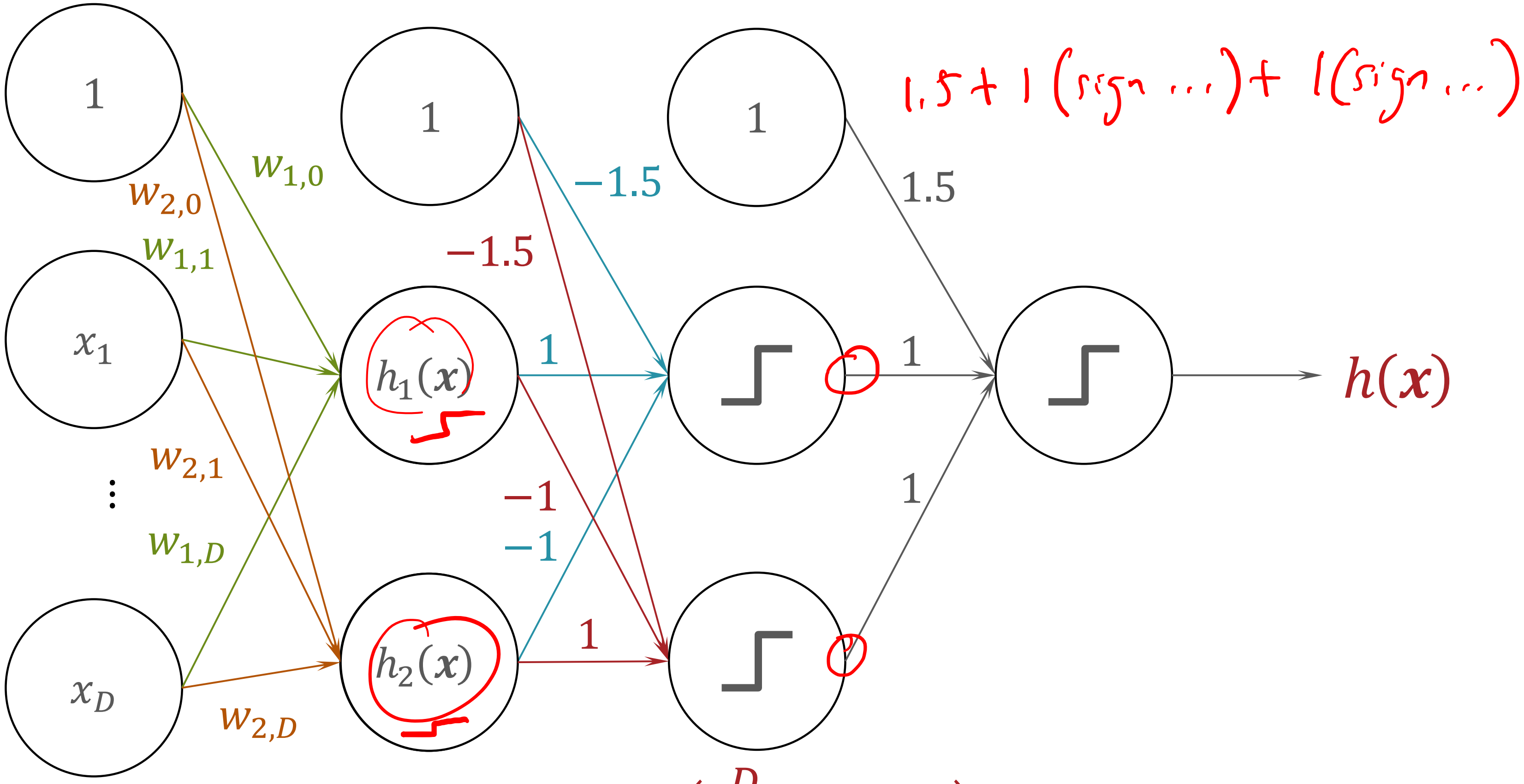
Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



Building a Network

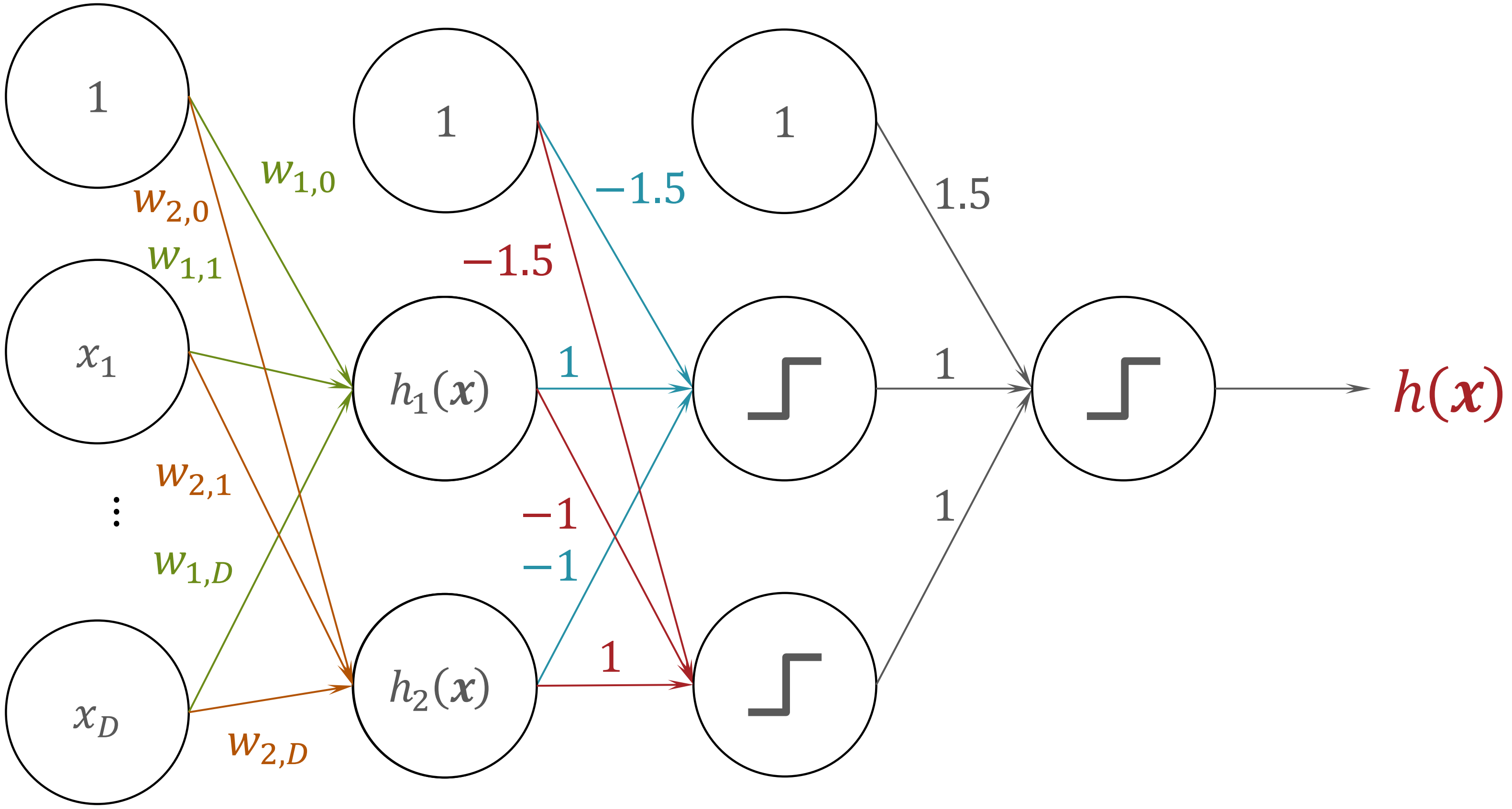
$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



$$h_i(\boldsymbol{x}) = \text{sign}\left(\boldsymbol{w}_i^T \boldsymbol{x}\right) = \text{sign}\left(\sum_{d=0}^{D} w_{i,d}\, x_d\right)$$

Building a Network

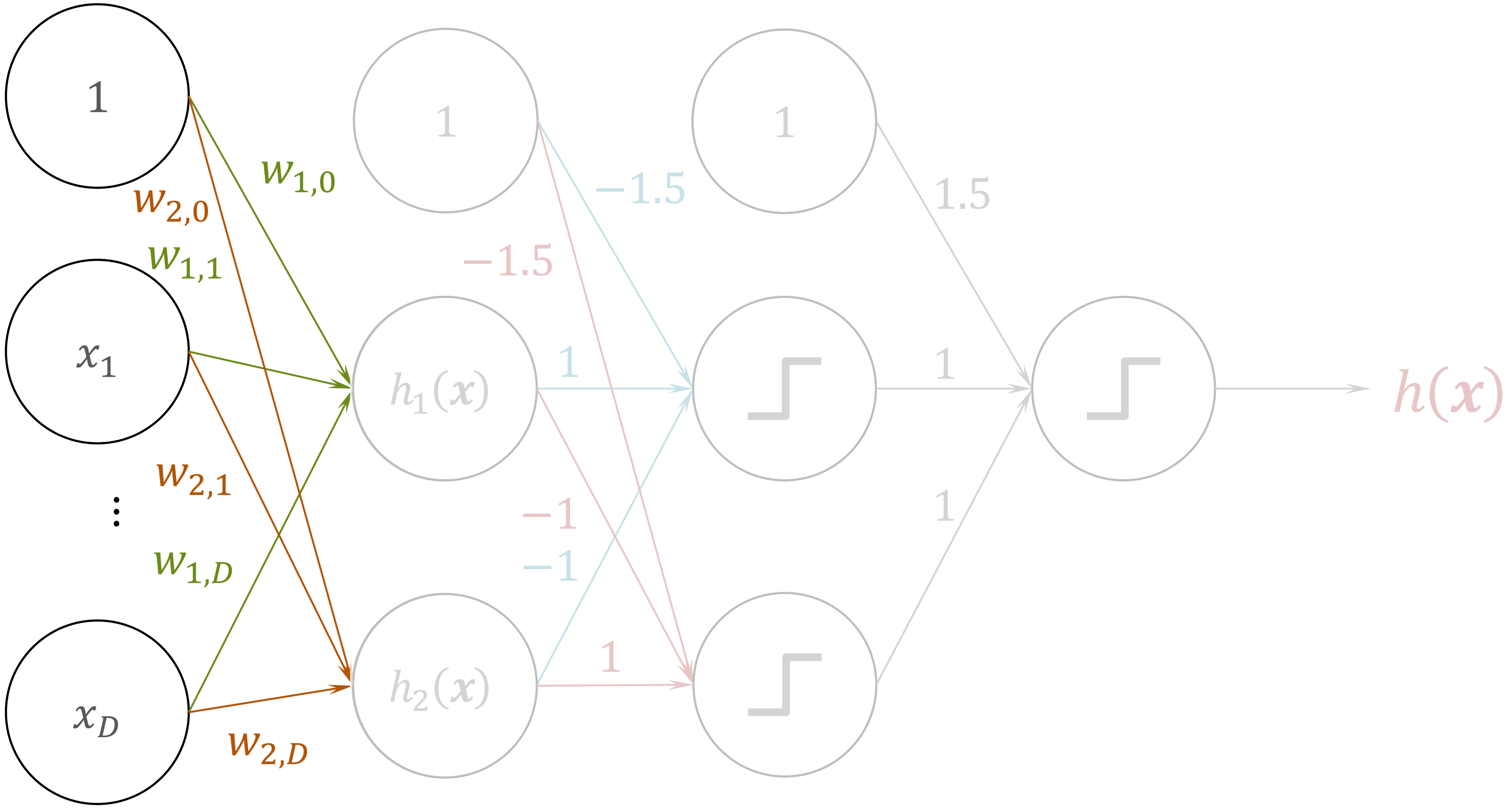
$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



$$h(\boldsymbol{x}) = \text{sign}(\text{sign}(\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) - \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + \text{sign}(-\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) + \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + 1.5)$$

Building a Network

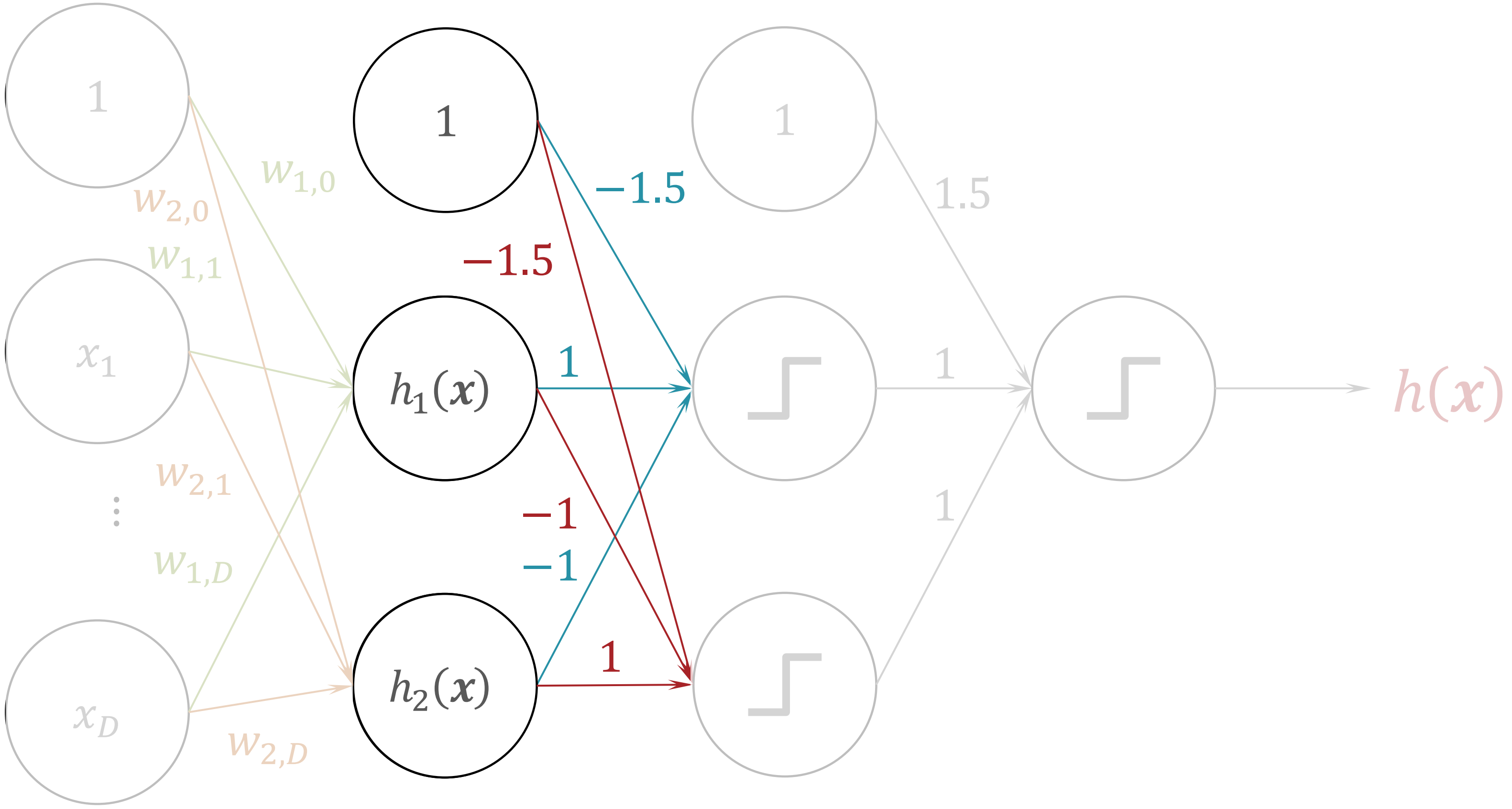
$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



$$h(\boldsymbol{x}) = \text{sign}(\text{sign}(\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) - \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + \text{sign}(-\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) + \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + 1.5)$$

Building a Network

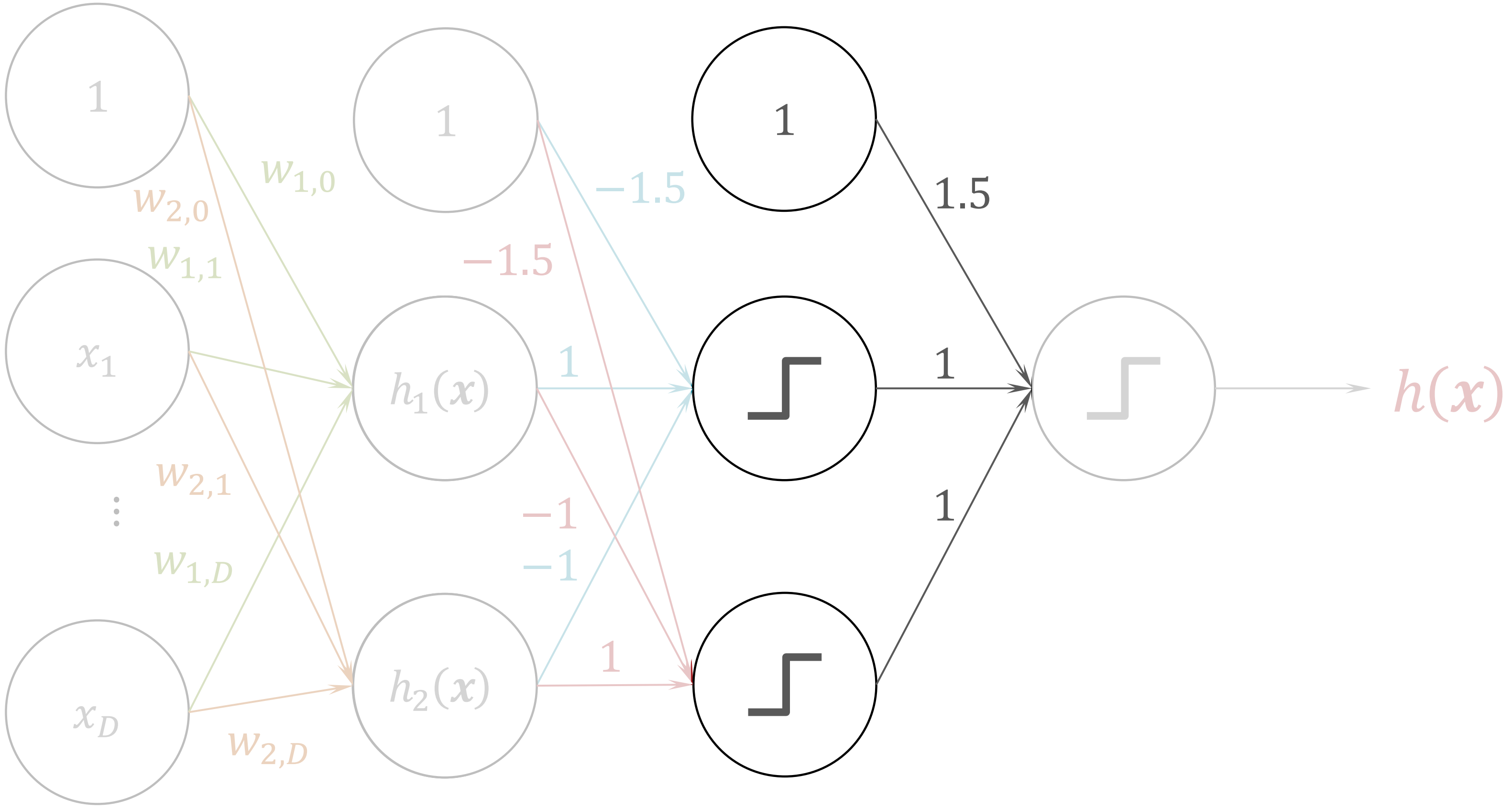
$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$



$$h(\boldsymbol{x}) = \text{sign}(\text{sign}(\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) - \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + \text{sign}(-\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) + \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + 1.5)$$

Building a Network

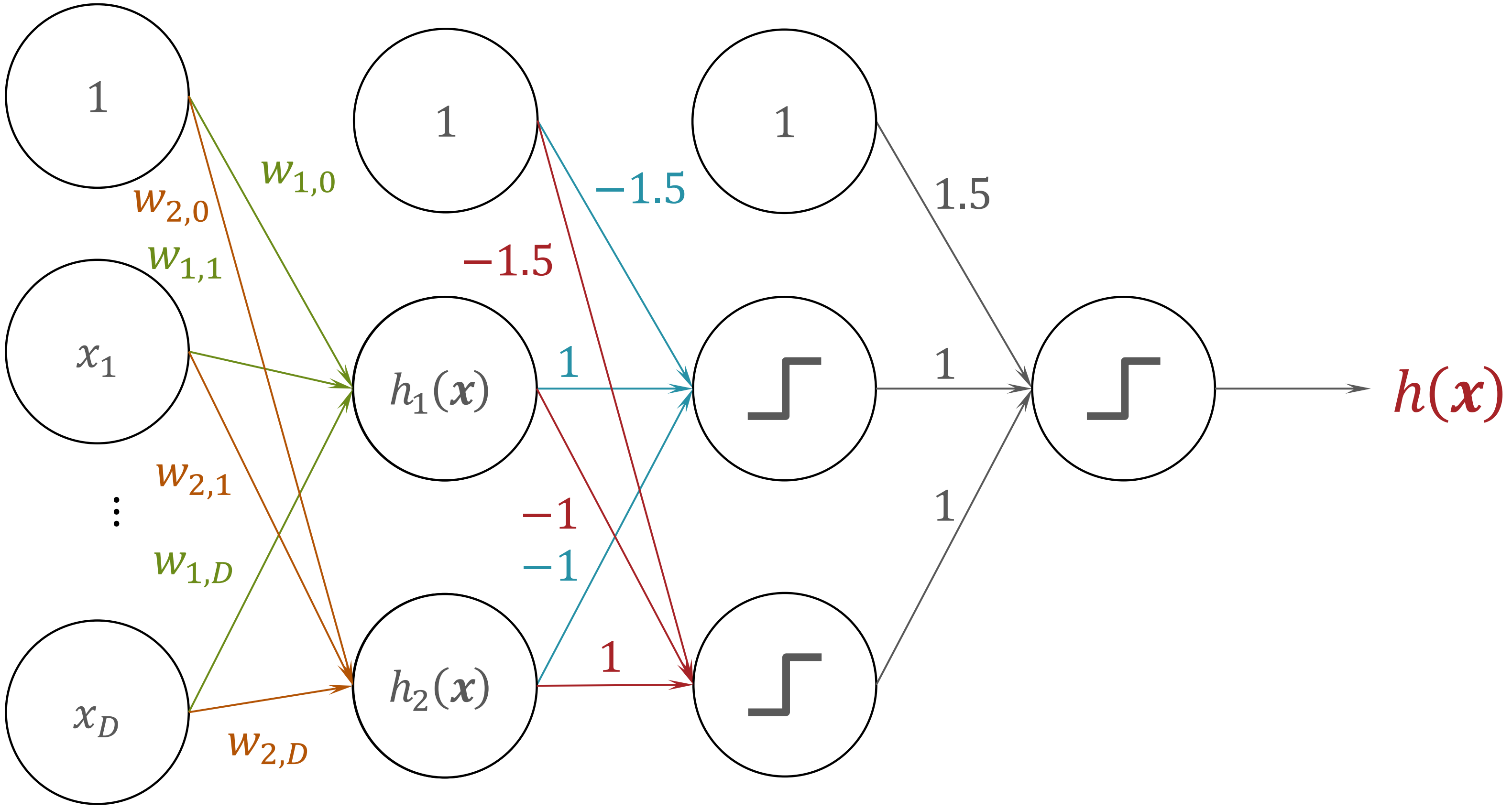
$$h(\boldsymbol{x}) = OR\Big(AND\big(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\big), AND\big(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\big)\Big)$$



$$h(\boldsymbol{x}) = \text{sign}\big(\text{sign}(\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) - \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + \text{sign}(-\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) + \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + 1.5\big)$$

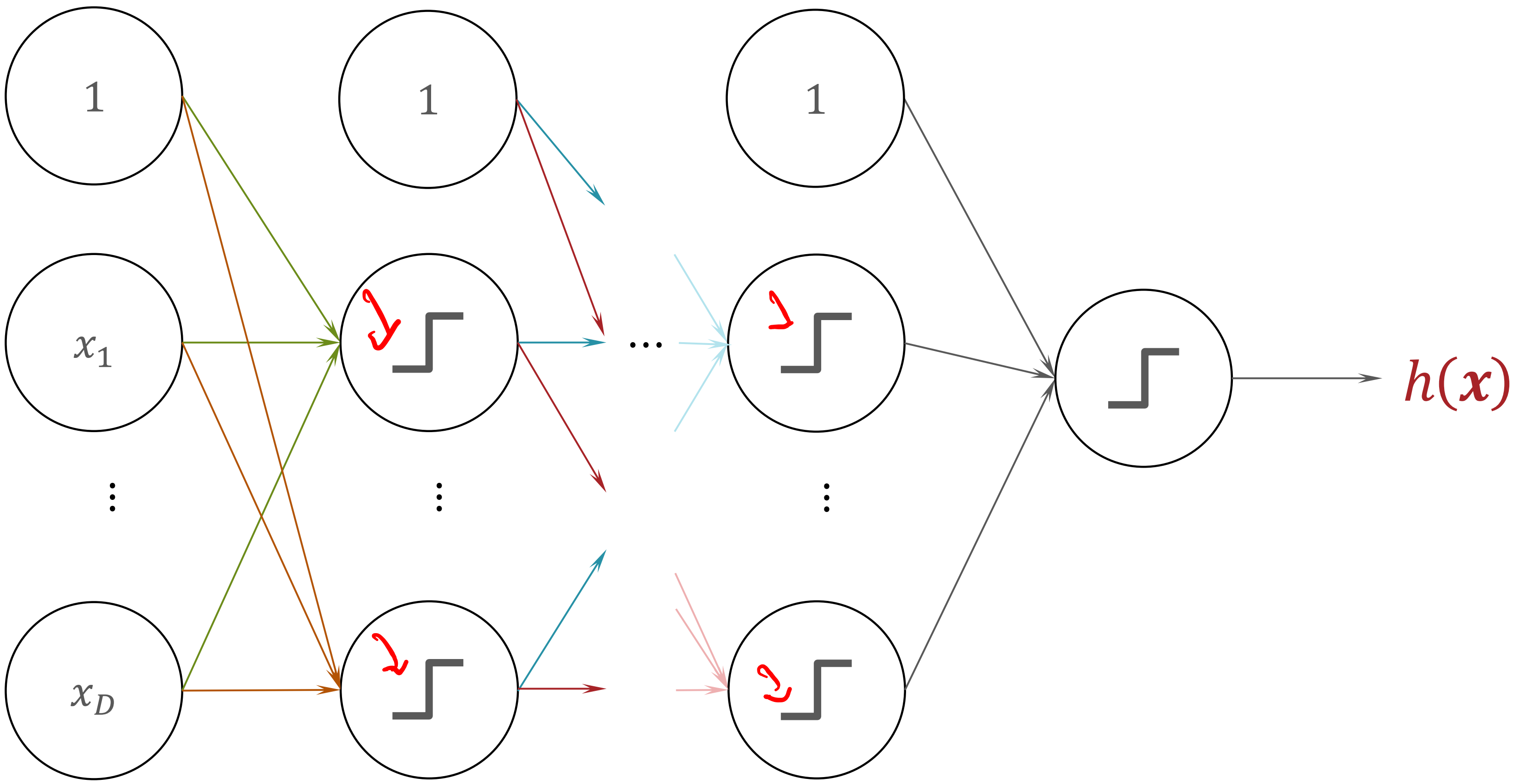
Building a Network

$$h(\boldsymbol{x}) = OR\left(AND\left(h_1(\boldsymbol{x}), \neg h_2(\boldsymbol{x})\right), AND\left(\neg h_1(\boldsymbol{x}), h_2(\boldsymbol{x})\right)\right)$$

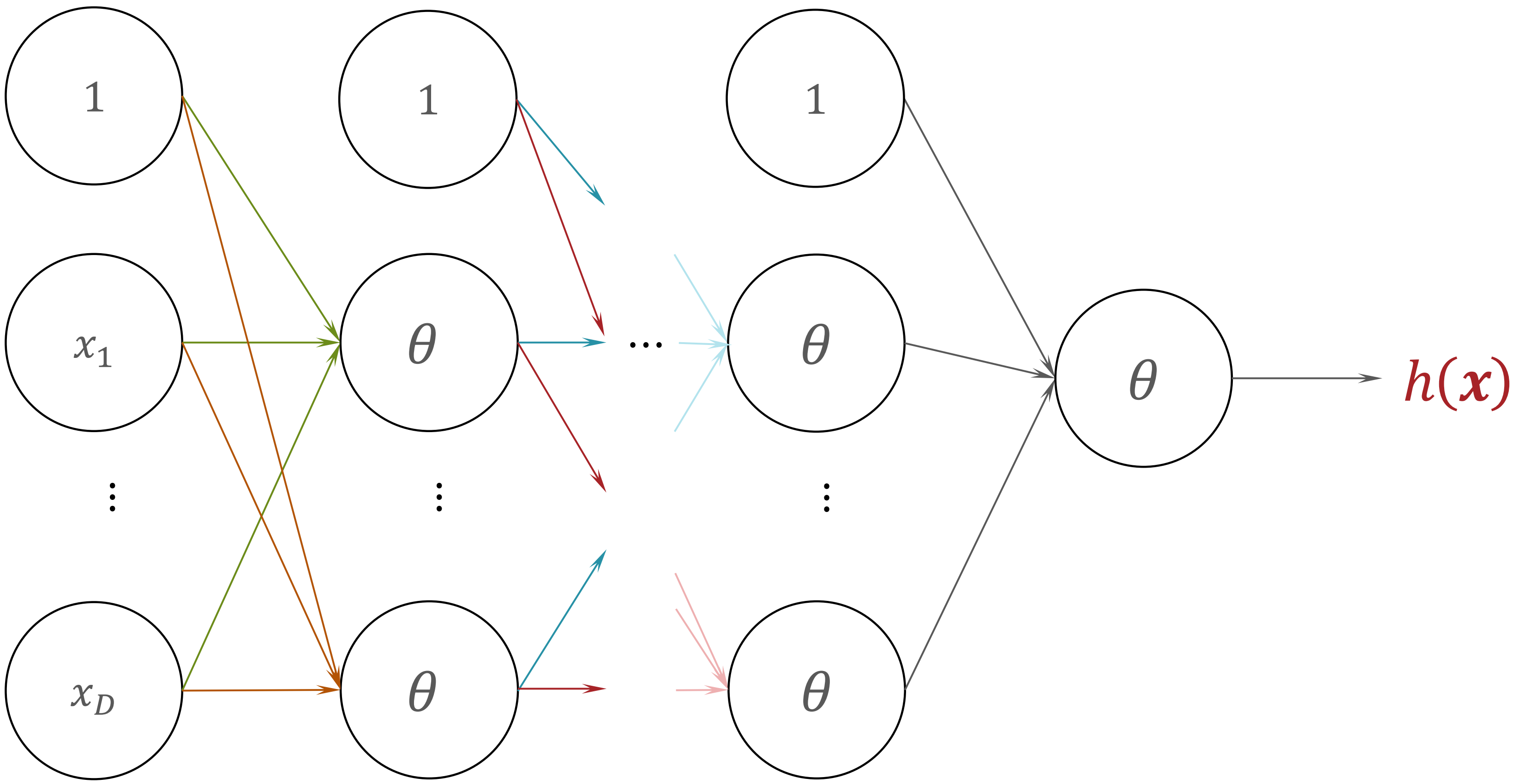


$$h(\boldsymbol{x}) = \text{sign}(\text{sign}(\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) - \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + \text{sign}(-\text{sign}(\boldsymbol{w}_1^T\boldsymbol{x}) + \text{sign}(\boldsymbol{w}_2^T\boldsymbol{x}) - 1.5) + 1.5)$$

Multi-Layer Perceptron (MLP)



(Fully-Connected) Feed Forward Neural Network

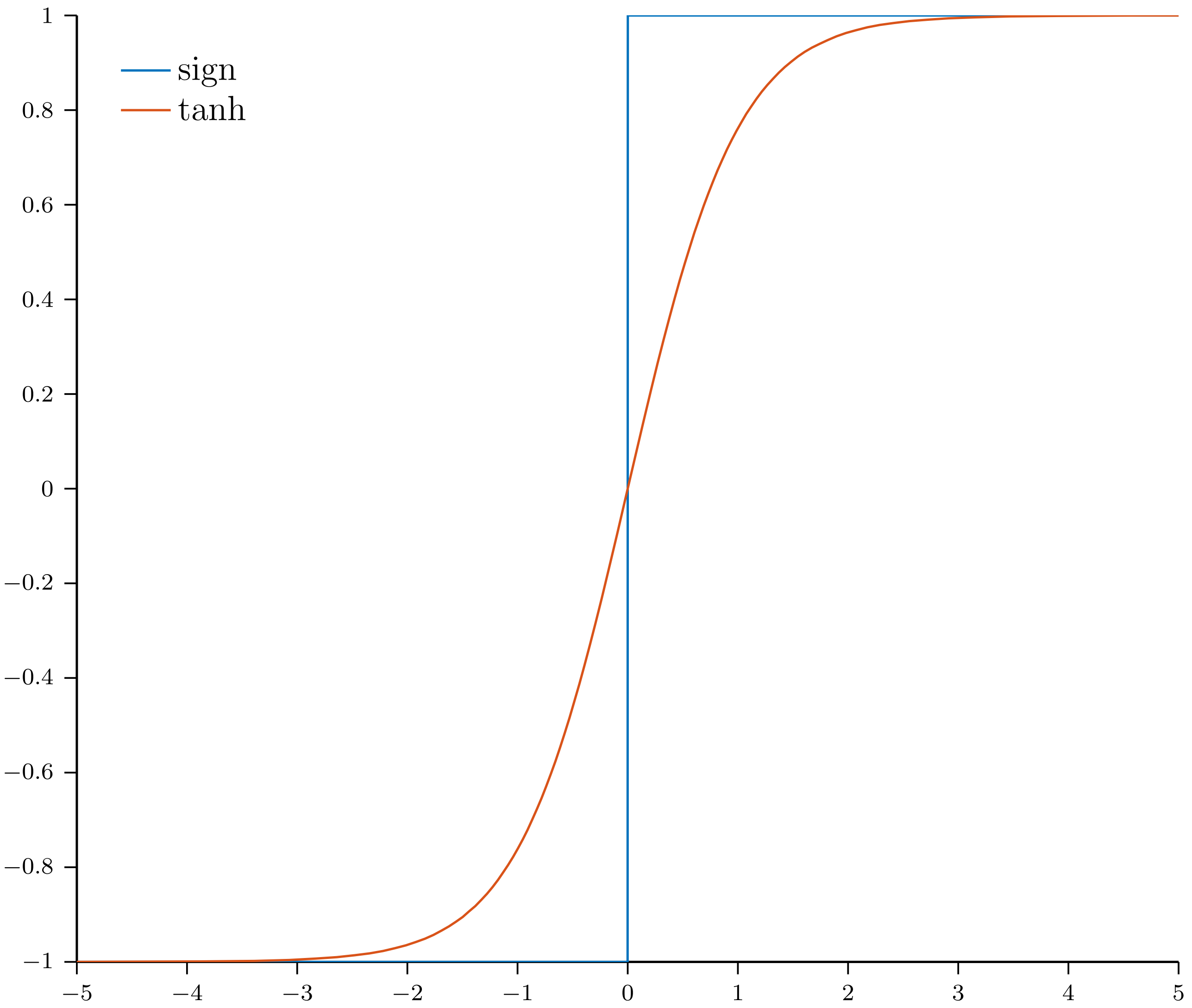


$\theta(\cdot)$

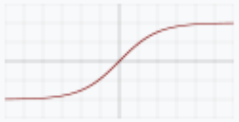
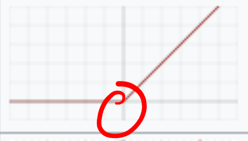
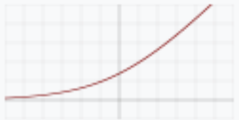
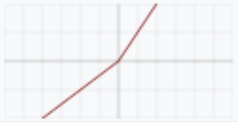
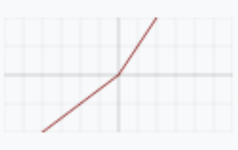
- Hyperbolic tangent:

$$\tanh(z) = \frac{\sinh(z)}{\cosh(z)} = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$

- $\frac{\partial \tanh(z)}{\partial z} = 1 - \tanh(z)^2$

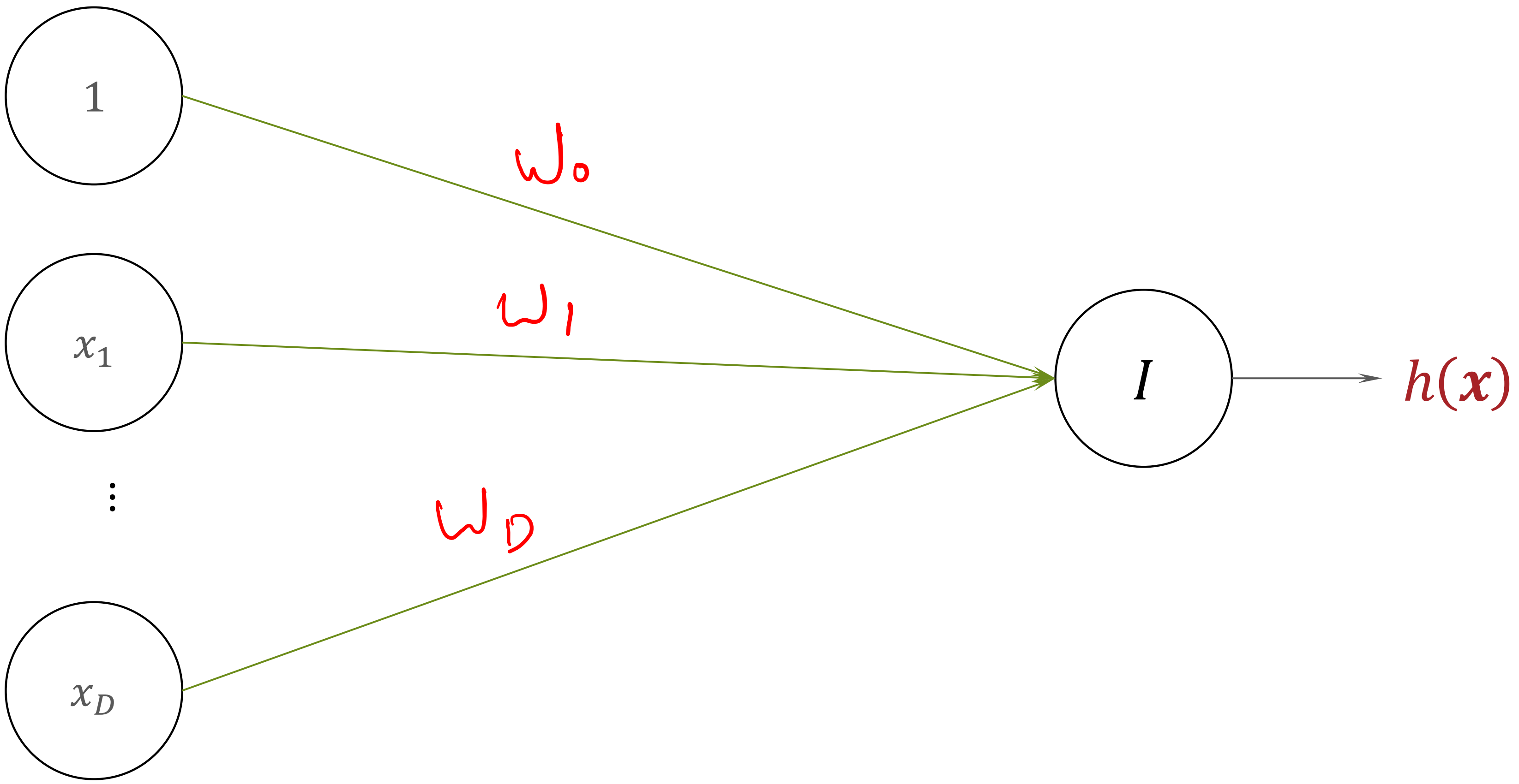


Other Activation Functions

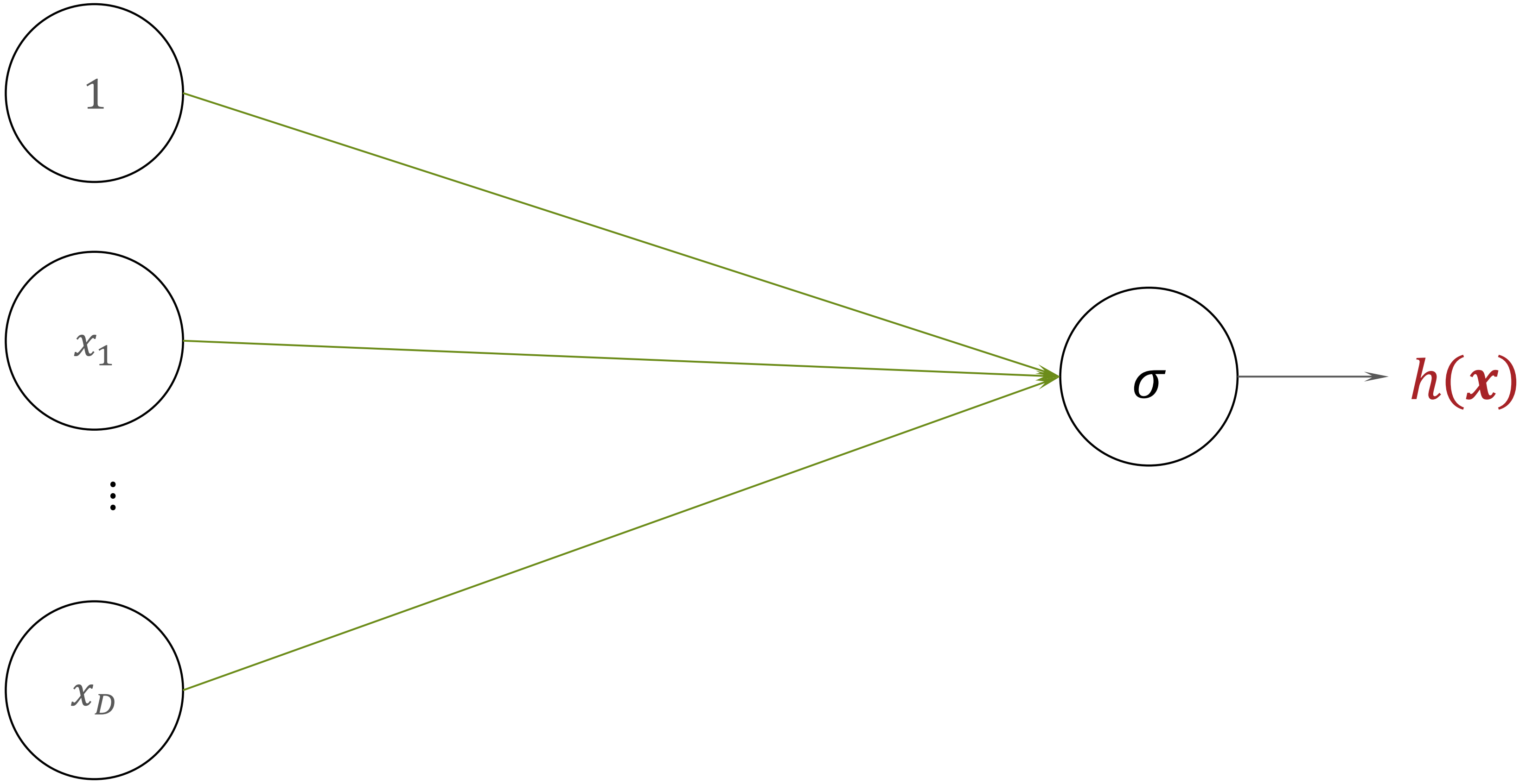
Logistic, sigmoid, or soft step		$\sigma(x) = \frac{1}{1+e^{-x}}$
Hyperbolic tangent (tanh)		$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
Rectified linear unit (ReLU)[7]		$\begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$ $= \max\{0, x\} = x\mathbf{1}_{x>0}$
Gaussian Error Linear Unit (GELU)[4]		$\frac{1}{2}x\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$ $= x\Phi(x)$
Softplus[8]		$\ln(1+e^x)$
Exponential linear unit (ELU)[9]		$\begin{cases} \alpha(e^x - 1) & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$ with parameter α
Leaky rectified linear unit (Leaky ReLU)[11]		$\begin{cases} 0.01x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$
Parametric rectified linear unit (PReLU)[12]		$\begin{cases} \alpha x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$ with parameter α

Source: https://en.wikipedia.org/wiki/Activation_function

Linear Regression as a Neural Network



Logistic Regression as a Neural Network



Recall: Building a Probabilistic Classifier

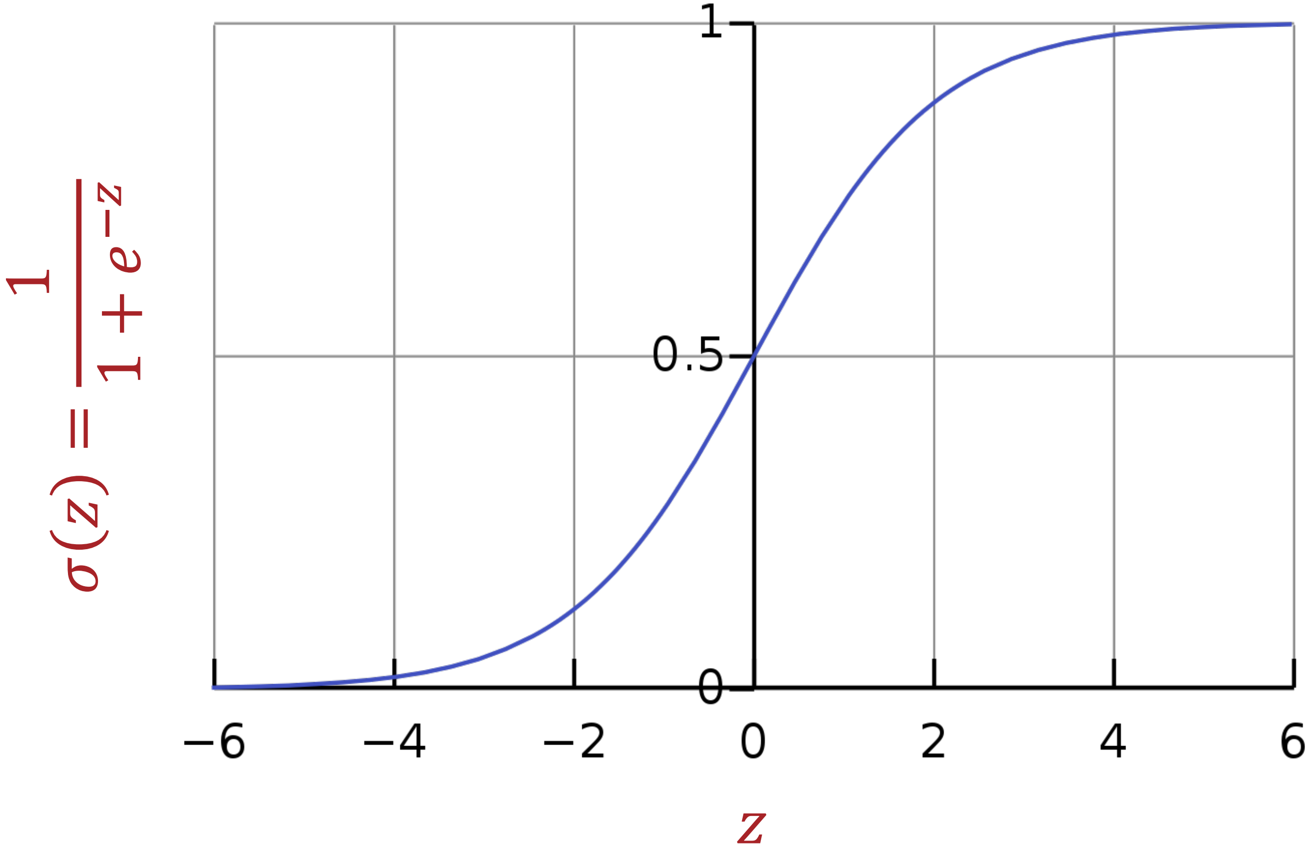
- Define a decision rule
  - Given a test data point $\boldsymbol{x}'$, predict its label $\hat{y}$ using the *posterior distribution* $P(Y = y|X = \boldsymbol{x}')$
  - Common choice: $\hat{y} = \underset{y}{\operatorname{argmax}} P(Y = y|X = \boldsymbol{x}')$
- Model the posterior distribution
  - Option 1 - Model $P(Y|X)$ directly as some function of X: given binary labels $y \in \{0,1\}$ **assume**

$$P(Y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})}$$

  - Option 2 - Use Bayes' rule (Naïve Bayes):

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)} \propto P(X|Y)\,P(Y)$$

Logistic Function



Source: https://en.wikipedia.org/wiki/Logistic_function#/media/File:Logistic-curve.svg

Why use the Logistic Function?

- Differentiable everywhere
- $\sigma : \mathbb{R} \to [0,1]$ ← $\sigma(\boldsymbol{w}^T\boldsymbol{x})$
- The decision boundary is linear in $\boldsymbol{x}$!

$\boldsymbol{w}^T\boldsymbol{x}$

- Differentiable everywhere
- $\sigma : \mathbb{R} \to [0,1]$
- The decision boundary is linear in $\boldsymbol{x}$!

$\sigma : \mathbb{R} \to [0,1]$

Reasonably smooth → differentiable everywhere

Linear decision boundaries

Source: https://en.wikipedia.org/wiki/Logistic_function#/media/File:Logistic-curve.svg

Logistic Regression Decision Boundary

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(w^T x) = P(Y=1 \mid x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{1}{1+\exp(-w^T x)} \geq \frac{1}{2}$$

$$2 \geq 1+\exp(-w^T x)$$

$$\exp(-w^T x) \leq 1$$

$$-w^T x \leq 0$$

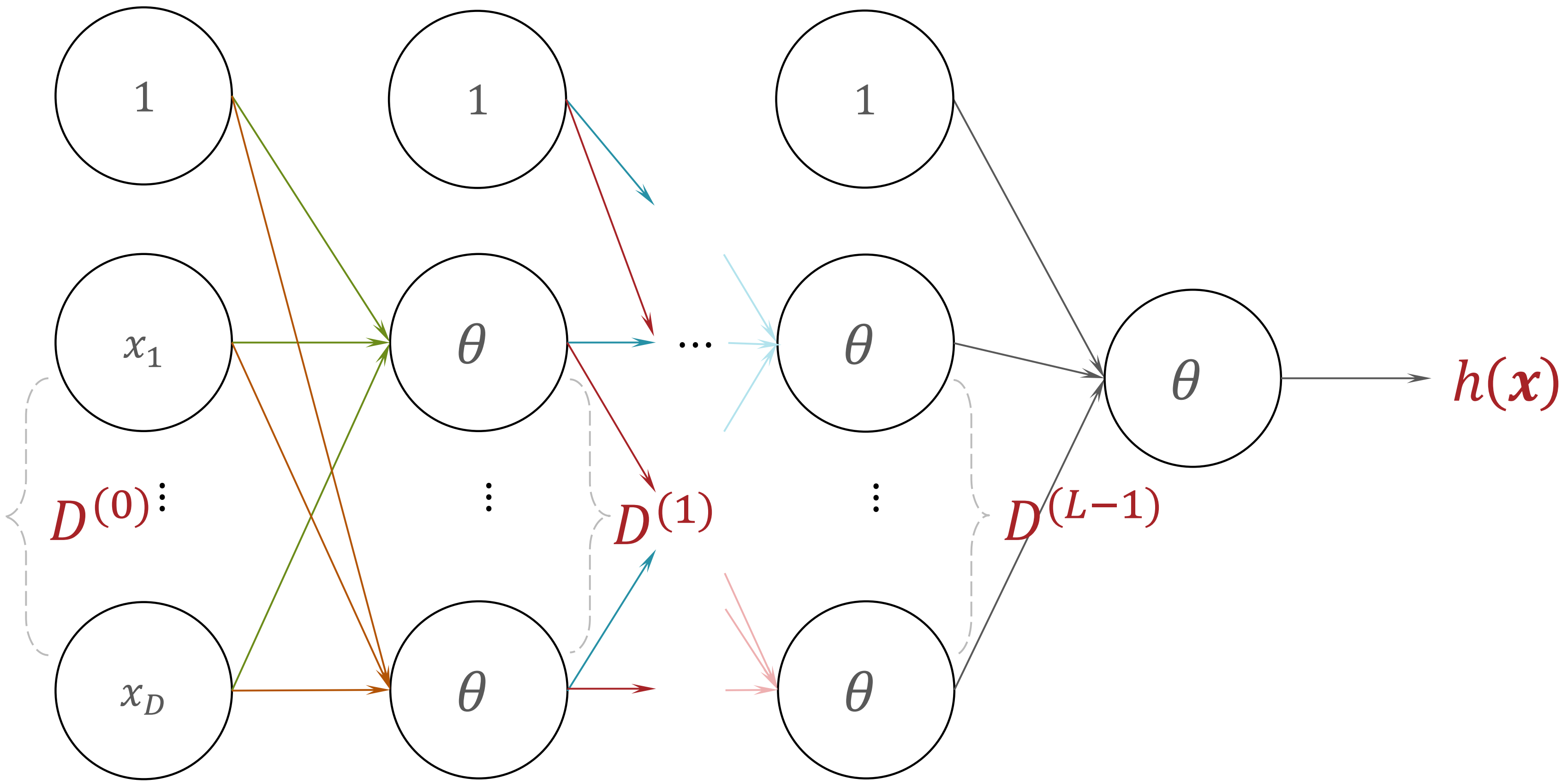
$$w^T x \geq 0$$

(Fully-Connected) Feed Forward Neural Network

Input layer: $l = 0$

Hidden layers: $l \in \{1, \dots, L-1\}$

Output layer: $l = L$

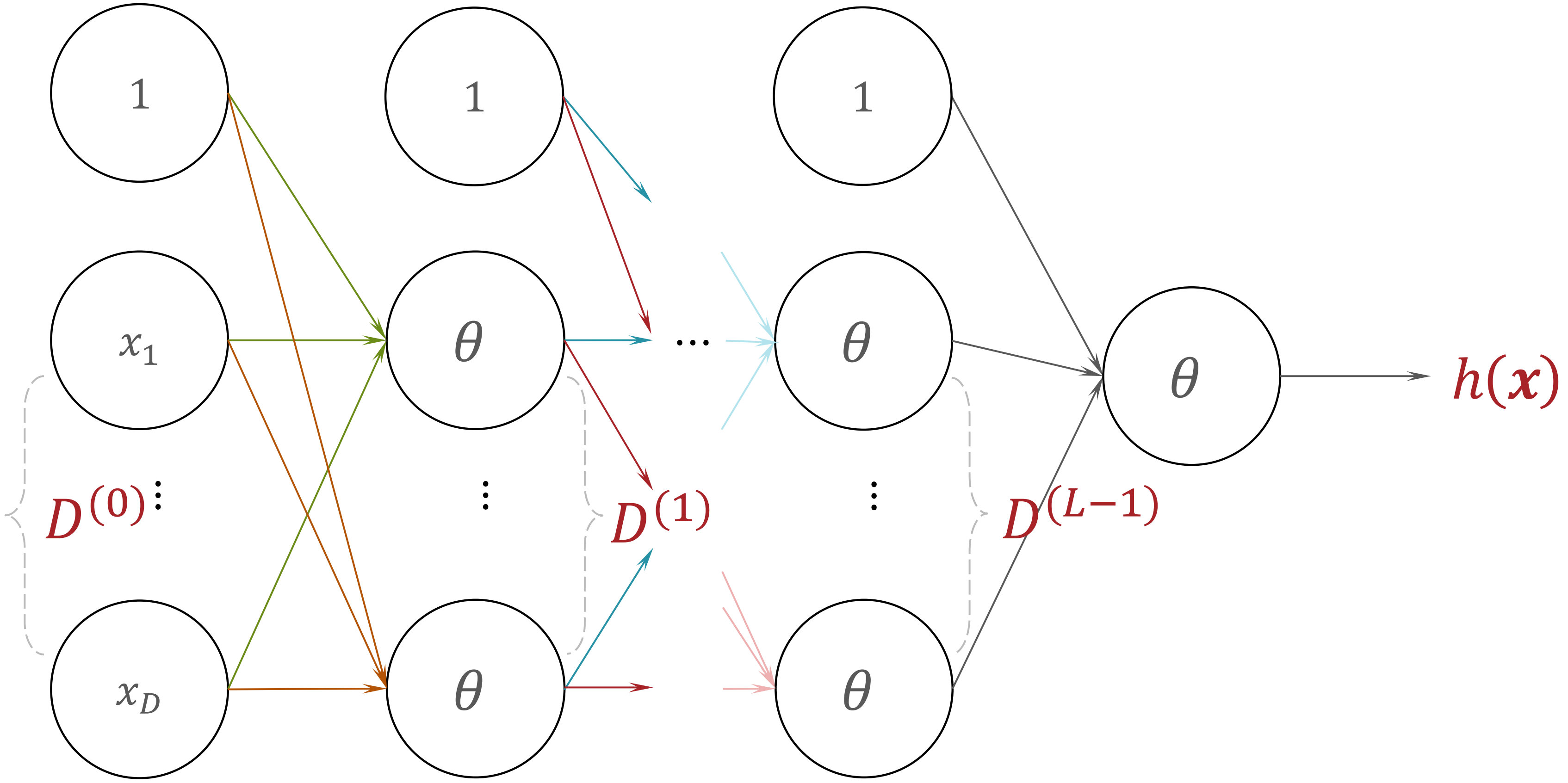


Layer l has dimension $D^{(l)}$ → Layer l has $D^{(l)} + 1$ nodes, counting the bias node

(Fully-Connected) Feed Forward Neural Network

The weights between layer $l-1$ and layer l are a matrix:

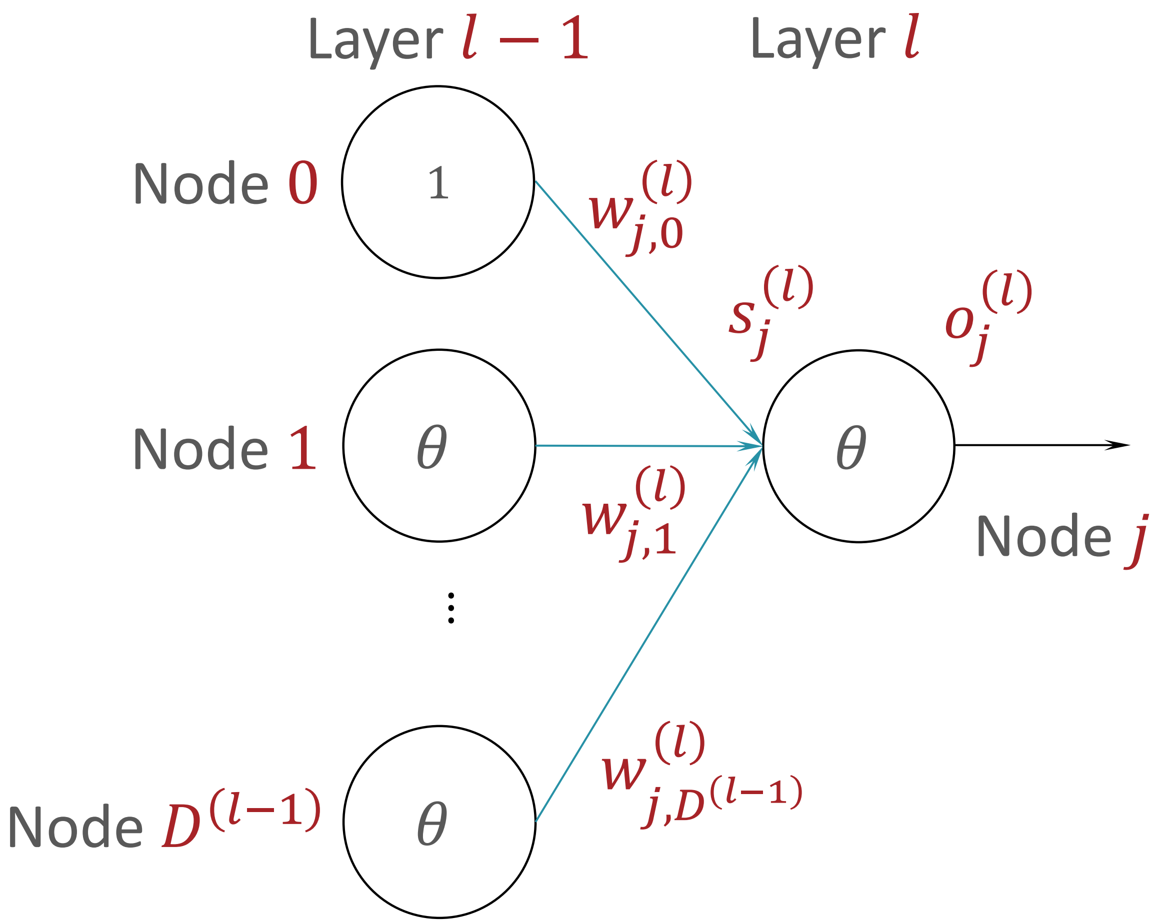
$W^{(l)} \in \mathbb{R}^{D^{(l)} \times \left(D^{(l-1)}+1\right)}$



$w_{j,i}^{(l)}$ is the weight between node i in layer $l-1$ and node j in layer l

Signal and Outputs

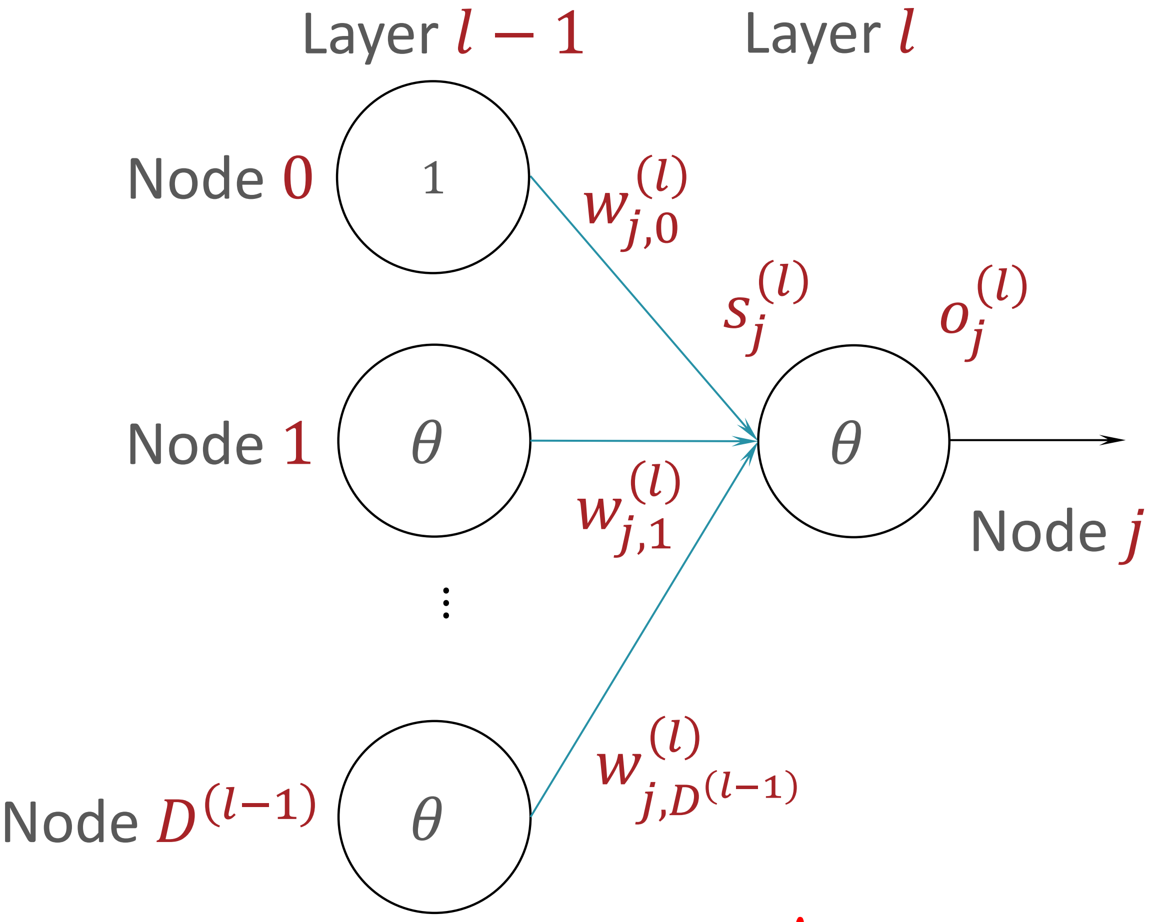
Every node has an incoming *signal* and outgoing *output*



$$s_j^{(l)} = \sum_{i=0}^{D^{(l-1)}} w_{j,i}^{(l)} o_i^{(l-1)} \text{ and } o_j^{(l)} = \theta\left(s_j^{(l)}\right)$$

Signal and Outputs

Every node has an incoming *signal* and outgoing *output*



$$\mathbf{s}^{(l)} = W^{(l)}\boldsymbol{o}^{(l-1)} \text{ and } \boldsymbol{o}^{(l)} = \left[1, \theta\left(\mathbf{s}^{(l)}\right)\right]^T$$

Forward Propagation for Making Predictions

- Input: weights $W^{(1)}, \ldots, W^{(L)}$ and a query data point $\boldsymbol{x}$
- Initialize $\boldsymbol{o}^{(0)} = \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix}$
- For $l = 1, \ldots, L$
  - $\boldsymbol{s}^{(l)} = W^{(l)} \boldsymbol{o}^{(l-1)}$
  - $\boldsymbol{o}^{(l)} = \begin{bmatrix} 1 \\ \theta\left(\boldsymbol{s}^{(l)}\right) \end{bmatrix}$
- Output: $h_{W^{(1)}, \ldots, W^{(L)}}(\boldsymbol{x}) = \boldsymbol{o}^{(L)}$

Stochastic Gradient Descent for Learning

- Input: $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}, \eta^{(0)}$
- Initialize all weights $W_{(0)}^{(1)}, \dots, W_{(0)}^{(L)}$ to small, random numbers and set $t = 0$
- While TERMINATION CRITERION is not satisfied
  - For $i \in \text{shuffle}(\{1, \dots, N\})$
    - For $l = 1, \dots, L$
      - Compute $G^{(l)} = \nabla_{W^{(l)}} \ell^{(i)}\left(W_{(t)}^{(1)}, \dots, W_{(t)}^{(L)}\right)$
      - Update $W^{(l)}$: $W_{(t+1)}^{(l)} = W_{(t)}^{(l)} - \eta^{(0)}_{\cancel{0}} G^{(l)}$
    - Increment t: $t = t + 1$
- Output: $W_{(t)}^{(1)}, \dots, W_{(t)}^{(L)}$

Two questions:

1. What is this loss function $\ell^{(i)}$?

2. How on earth do we take these gradients?

- Input: $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}, \eta^{(0)}$
- Initialize all weights $W_{(0)}^{(1)}, \dots, W_{(0)}^{(L)}$ to small, random numbers and set $t = 0$ (???)
- While TERMINATION CRITERION is not satisfied (???)
  - For $i \in \text{shuffle}(\{1, \dots, N\})$
    - For $l = 1, \dots, L$
      - Compute $G^{(l)} = \nabla_{W^{(l)}} \ell^{(i)} \left(W_{(t)}^{(1)}, \dots, W_{(t)}^{(L)}\right)$
      - Update $W^{(l)}$: $W_{(t+1)}^{(l)} = W_{(t)}^{(l)} - \eta_0 G^{(l)}$
    - Increment t: $t = t + 1$
- Output: $W_{(t)}^{(1)}, \dots, W_{(t)}^{(L)}$

$$\omega = \{W^{(1)}, \dots, W^{(L)}\}$$

Loss Functions for Neural Networks

- Regression - squared error (same as linear regression!)

$$\ell^{(i)}\left(W_{(t)}^{(1)}, \dots, W_{(t)}^{(L)}\right) = \left(h_{W^{(1)},\dots,W^{(L)}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2$$

- Binary classification - cross-entropy loss $(Y \in \{0,1\})$
  - Assume $P\left(Y = 1 \middle| \boldsymbol{x}, W^{(1)}, \dots, W^{(L)}\right) = h_{W^{(1)},\dots,W^{(L)}}(\boldsymbol{x})$

$$\ell^{(i)}(W^{(1)}, \dots, W^{(L)}) = -\log P(y^{(i)} \mid x^{(i)}, W^{(1)}, \dots, W^{(L)})$$

$$= -\log\left(\underbrace{h_w(x^{(i)})}^{y^{(i)}} \underbrace{(1 - h_w(x^{(i)}))}^{1-y^{(i)}}\right)$$

$$= -\left(y^{(i)} \log(h_w(x^{(i)})) + (1-y^{(i)}) \log(1 - h_w(x^{(i)}))\right)$$

Loss Functions for Neural Networks

- Multi-class classification - also the cross-entropy loss!
  - Express the label as a one-hot or one-of-C vector e.g.,

$$y = [0 \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0]$$

  - Assume the neural network output is also a vector of length C

$$P\left(y[c] = 1 \middle| \boldsymbol{x}, W^{(1)}, \dots, W^{(L)}\right) = h_{W^{(1)}, \dots, W^{(L)}}(\boldsymbol{x})[c]$$

  - Then the cross-entropy loss is

$$\ell^{(i)}\left(W_{(t)}^{(1)}, \dots, W_{(t)}^{(L)}\right) = -\log P\left(y^{(i)} \middle| \boldsymbol{x}^{(i)}, W^{(1)}, \dots, W^{(L)}\right)$$

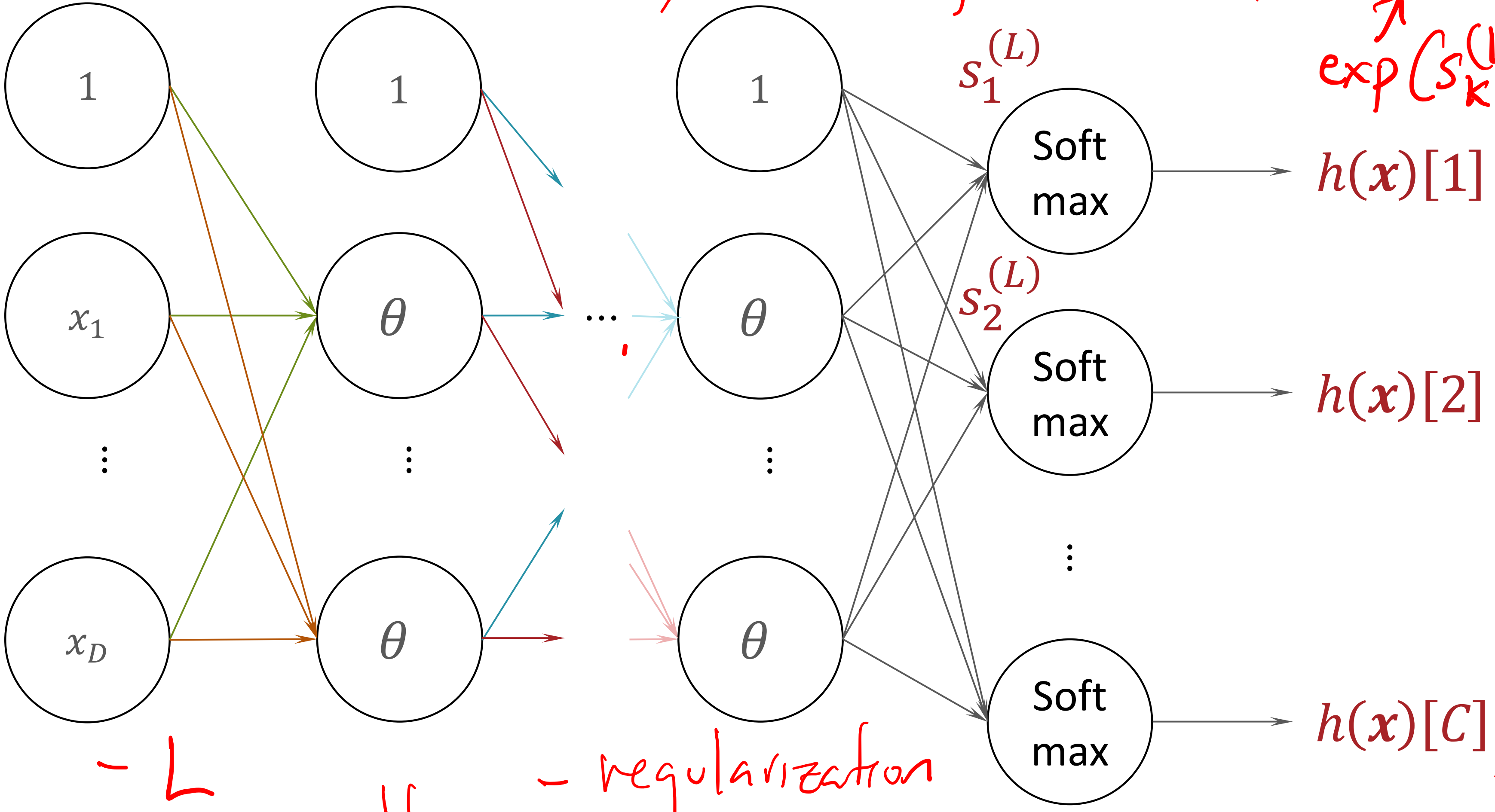
$$= -\sum_{c=1}^{C} y[c] \log h_w(x^{(i)})[c]$$

Multi-dimensional Outputs

hyperparameters?
- learning rate
- activation function(s)
- loss functions
- initialization
- stopping criterion

$$h(\boldsymbol{x})[c] = \frac{\exp s_c^{(L)}}{\sum_{k=1}^{C} \cancel{s_k^{(L)}}}$$

$\exp(s_k^{(L)})$



$s_{D^{(L)}}^{(L)}$

- L
- dimensionality
- connections
- regularization
- pre-processing

Key Takeaways

- Many common machine learning models can be represented as neural networks.
- Perceptrons can be combined to achieve non-linear decision boundaries
- Feed-forward neural network model:
  - Activation function
  - Layers: input, hidden & output
  - Weight matrices
  - Signals & outputs
- Forward propagation for making predictions
- Neural networks can use the same loss functions as other machine learning models