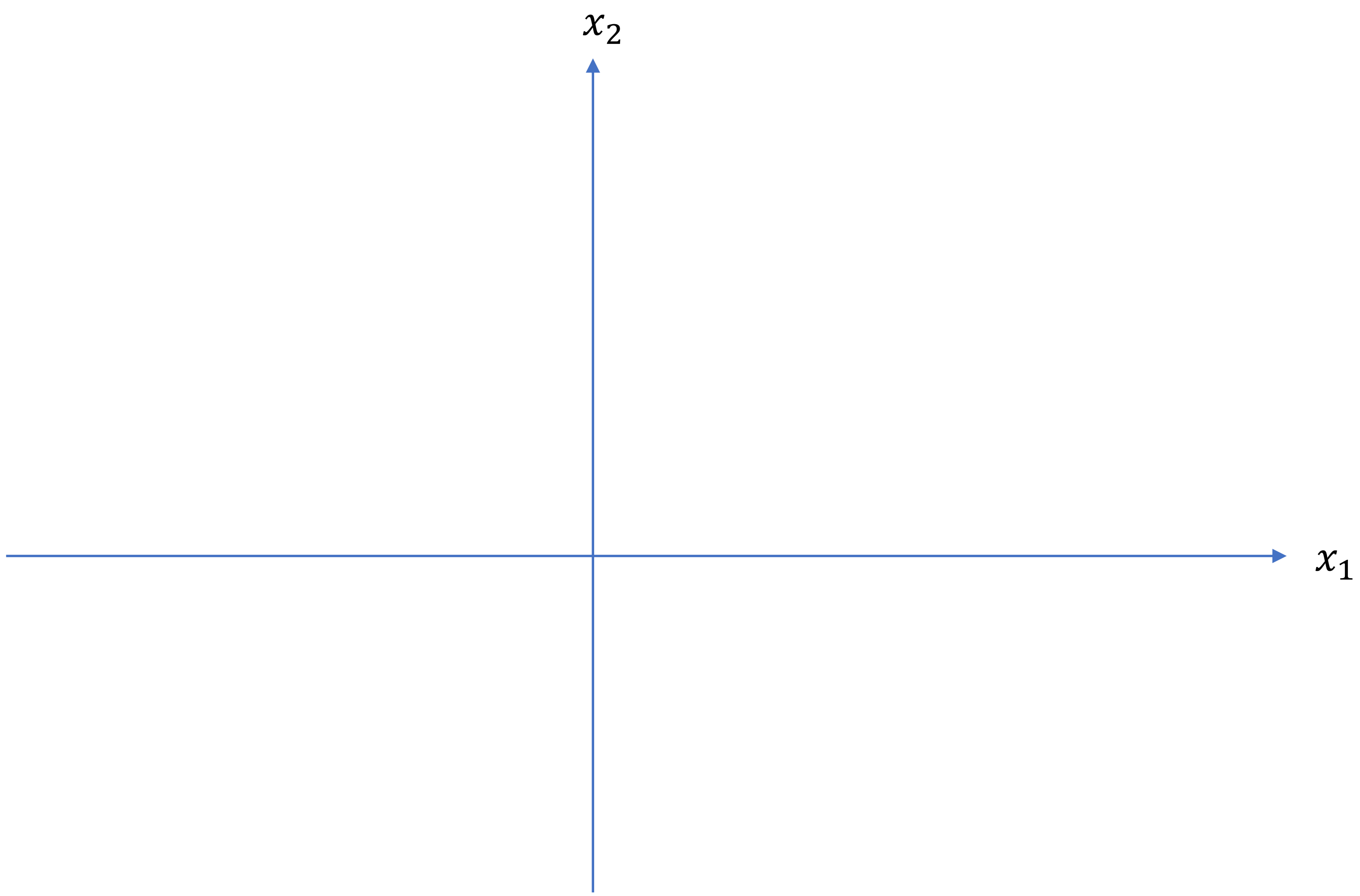
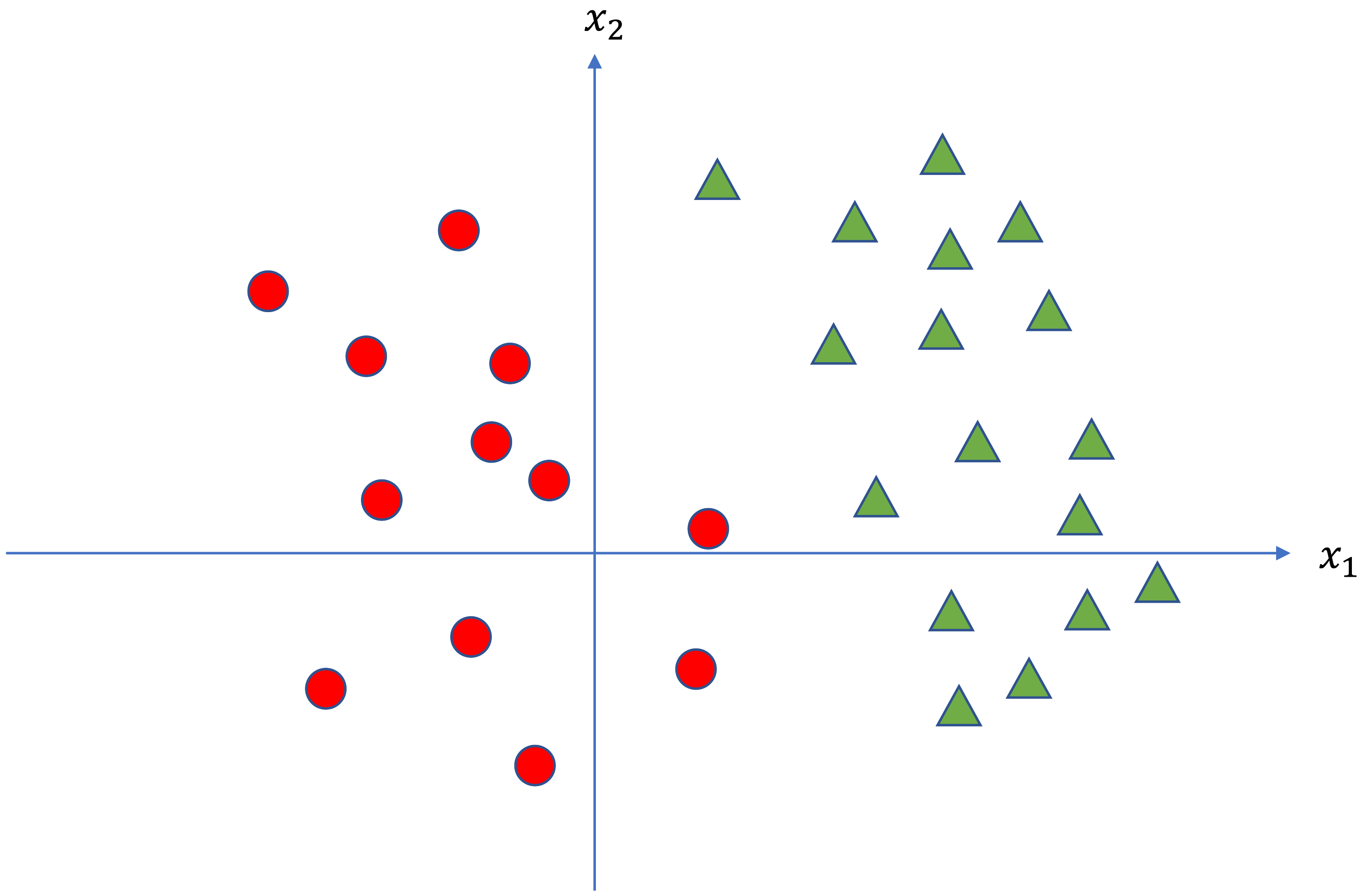
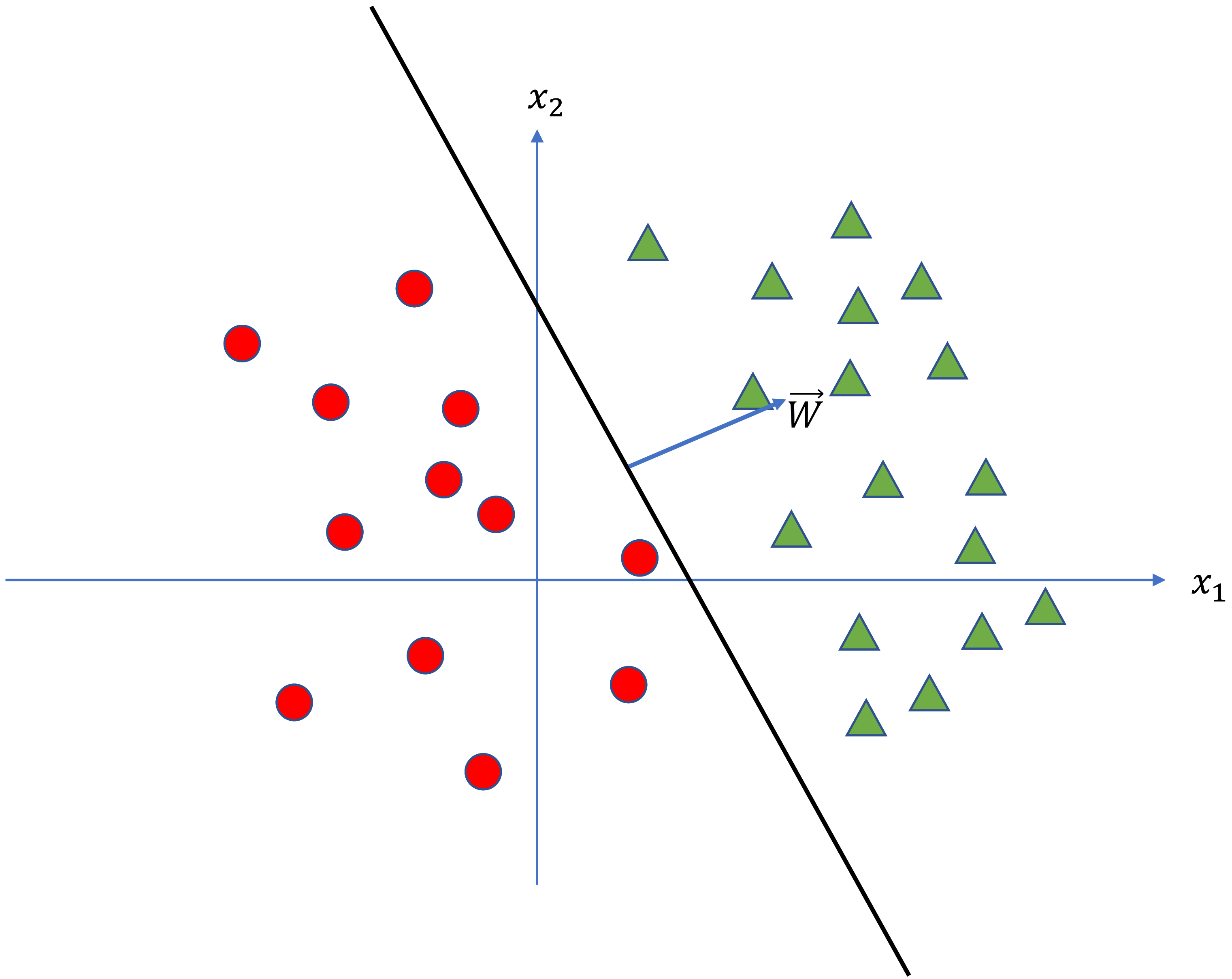
Logistic Regression

Motivation

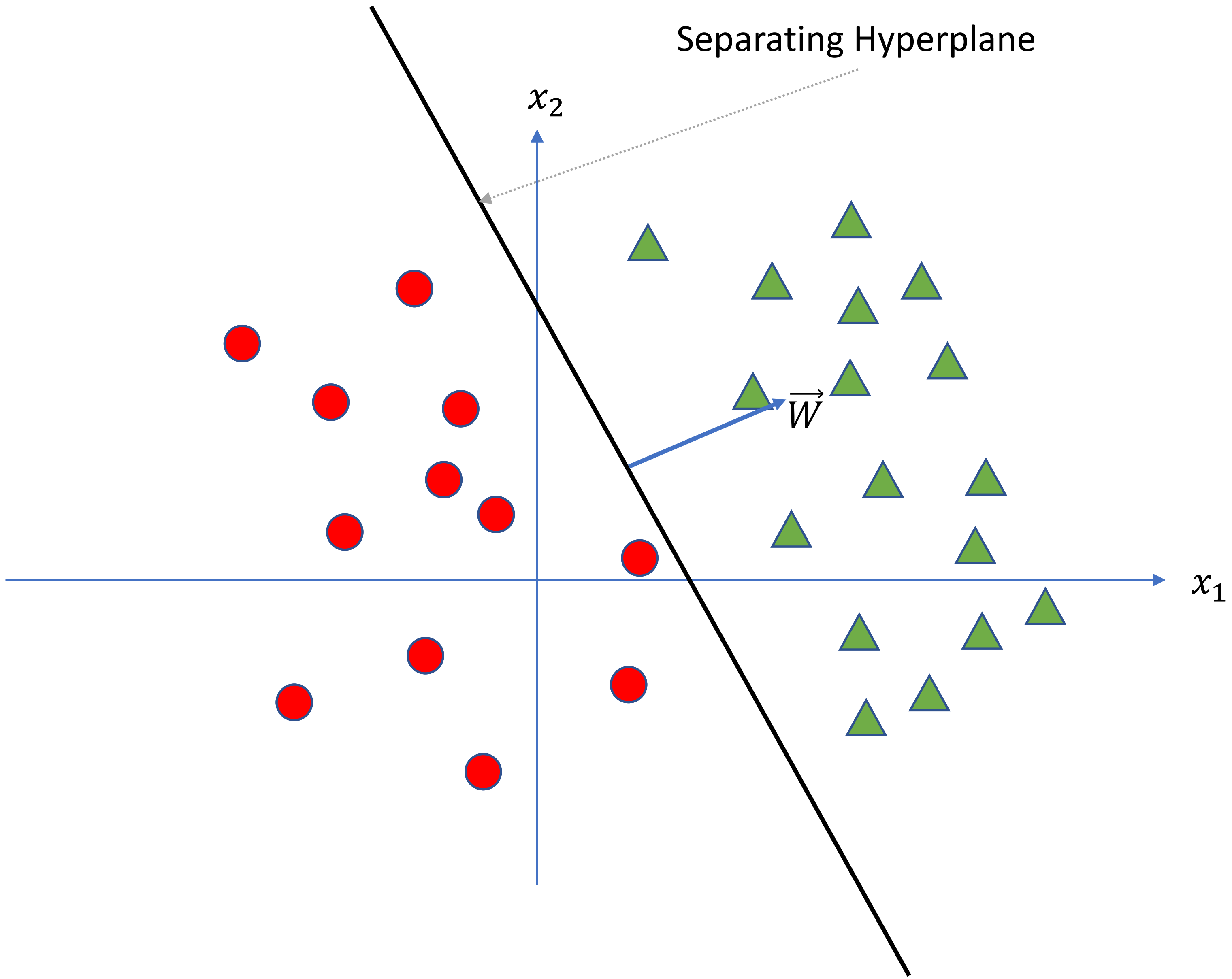

x_2
x_1



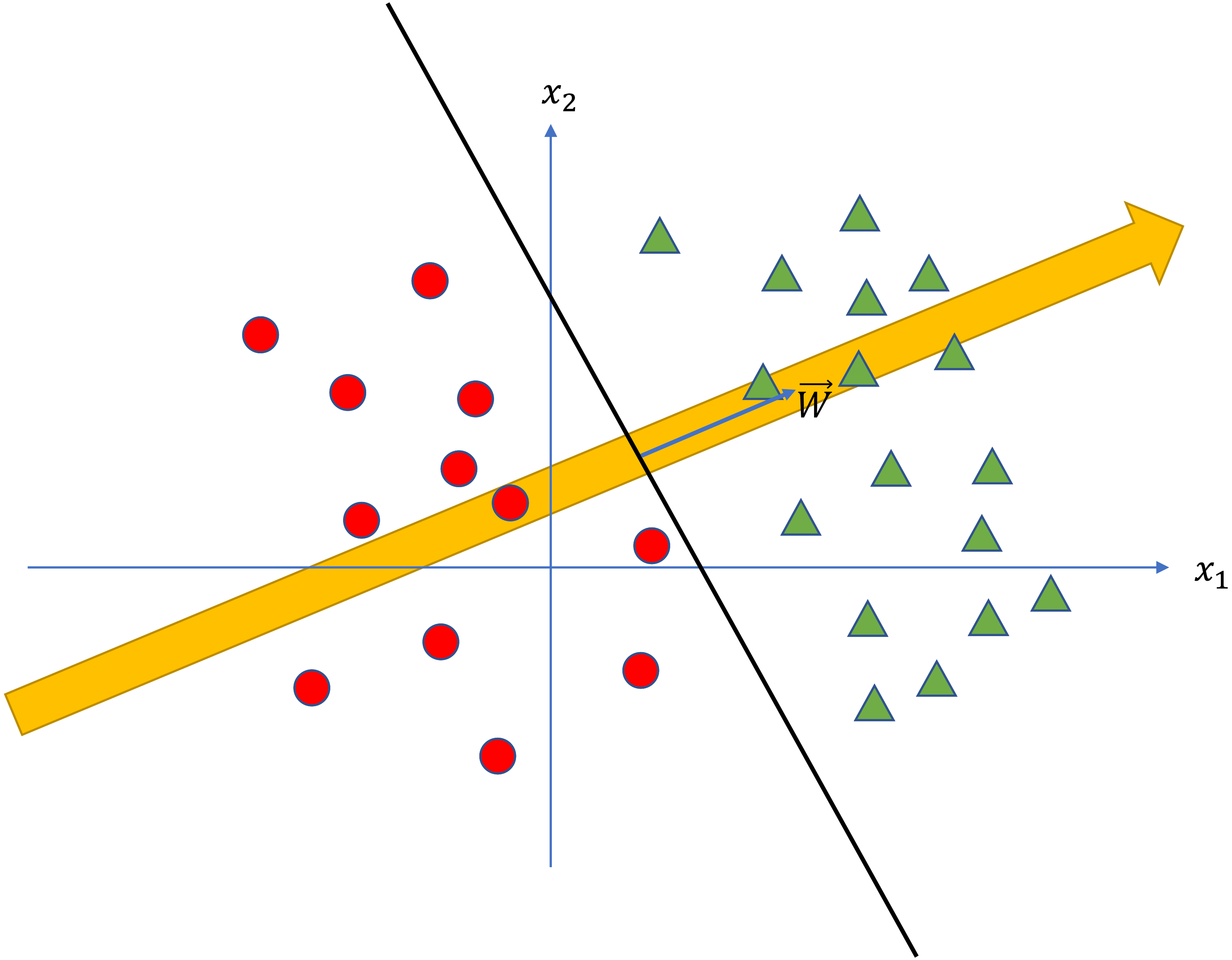
x_2
x_1

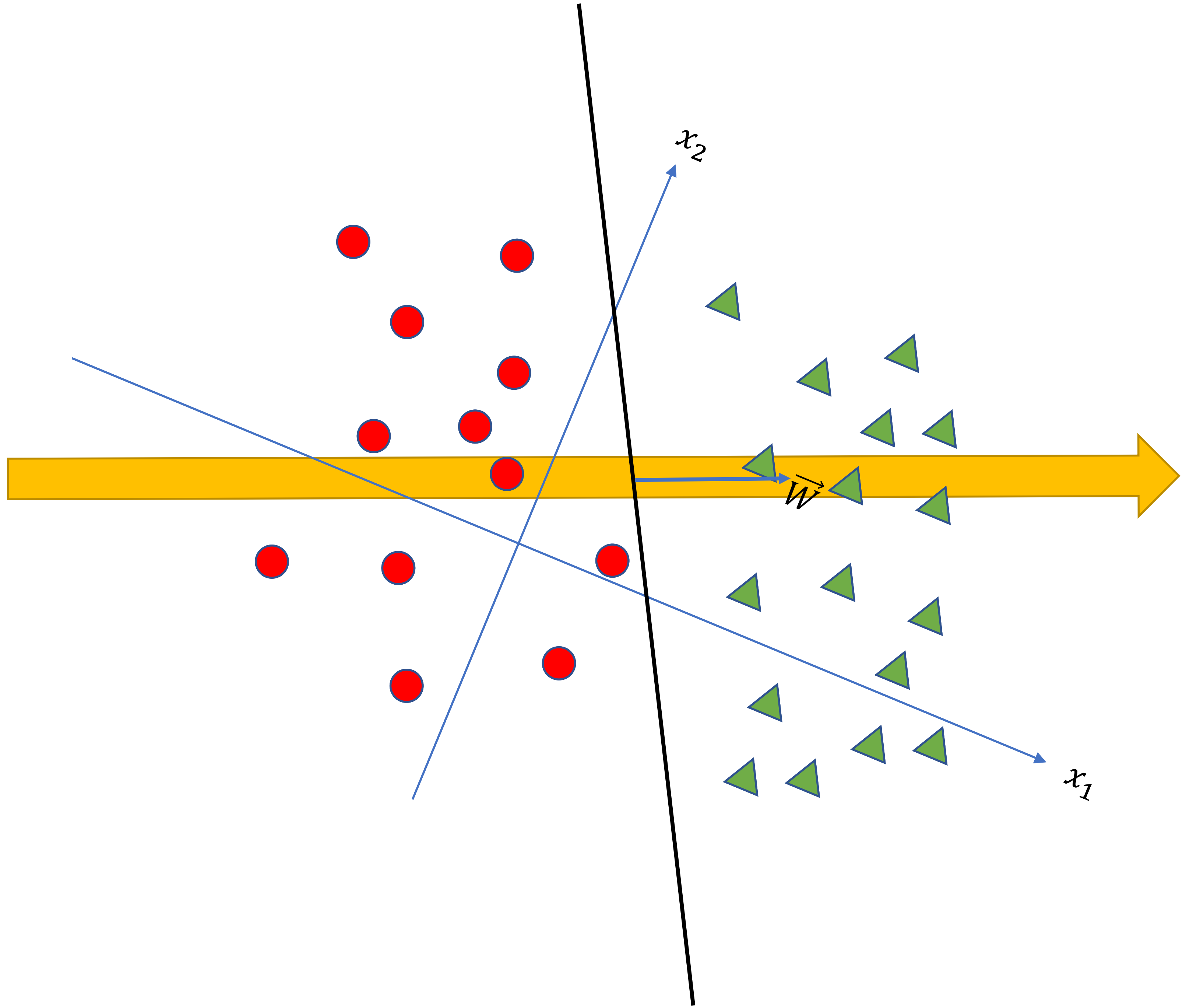


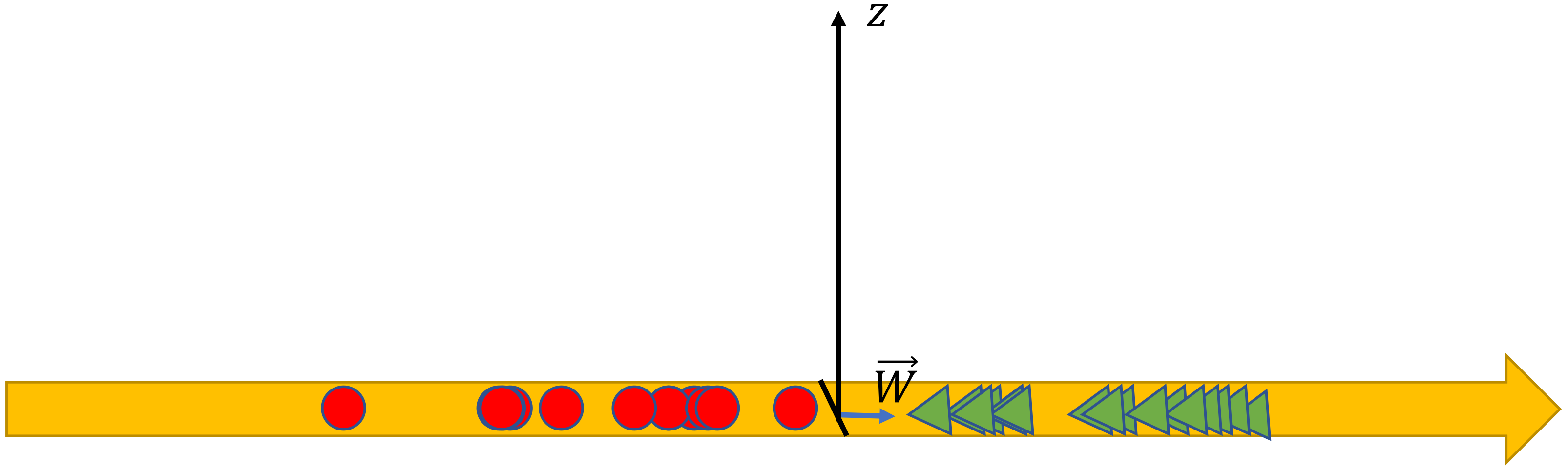
x_2
$\vec{W}$
x_1



Separating Hyperplane
x_2
$\vec{W}$
x_1

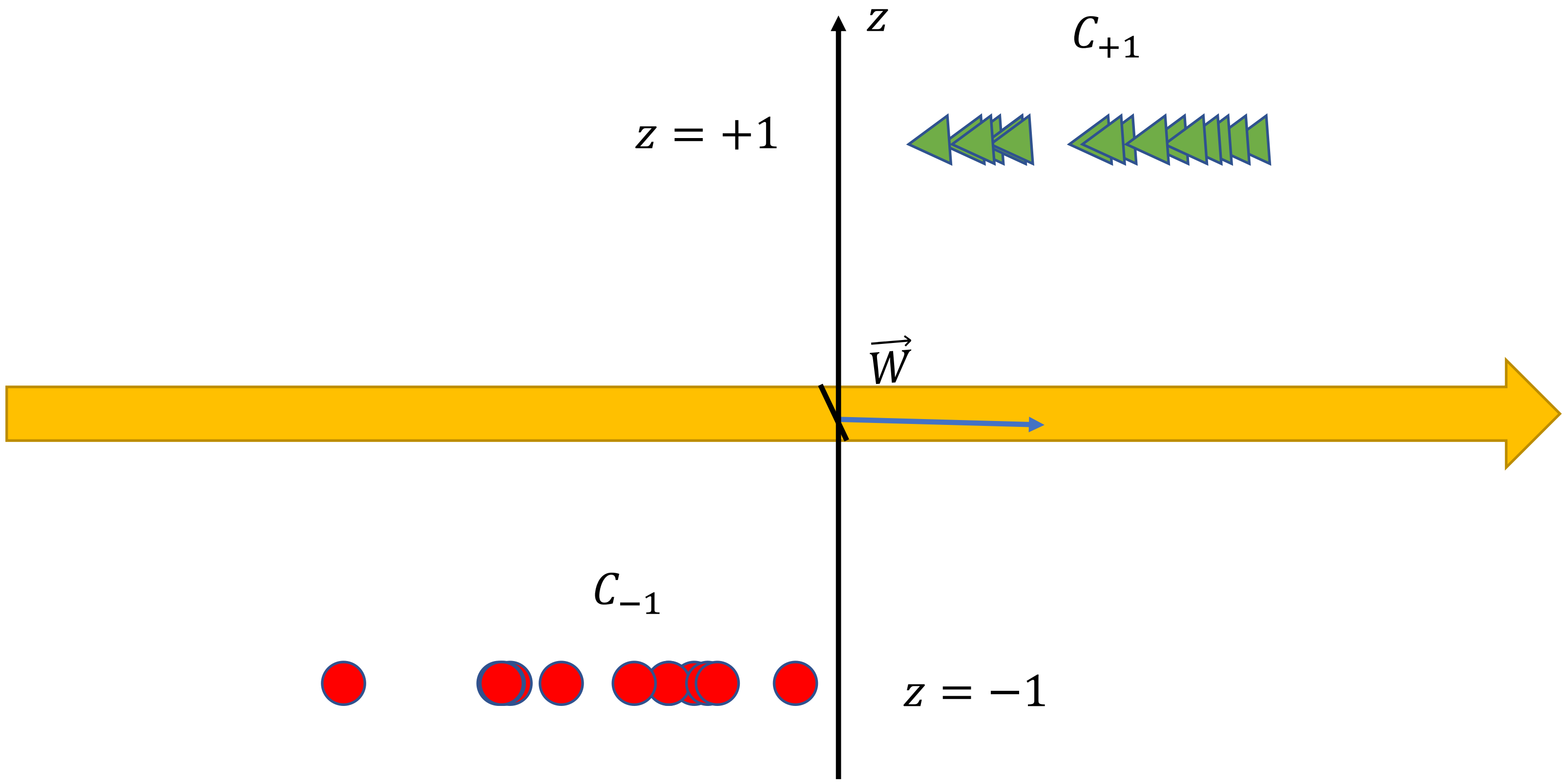

x_2
x_1
$\overrightarrow{W}$


x_2
$\vec{W}$
x_1



z
$\overrightarrow{W}$

Perceptron's prediction:

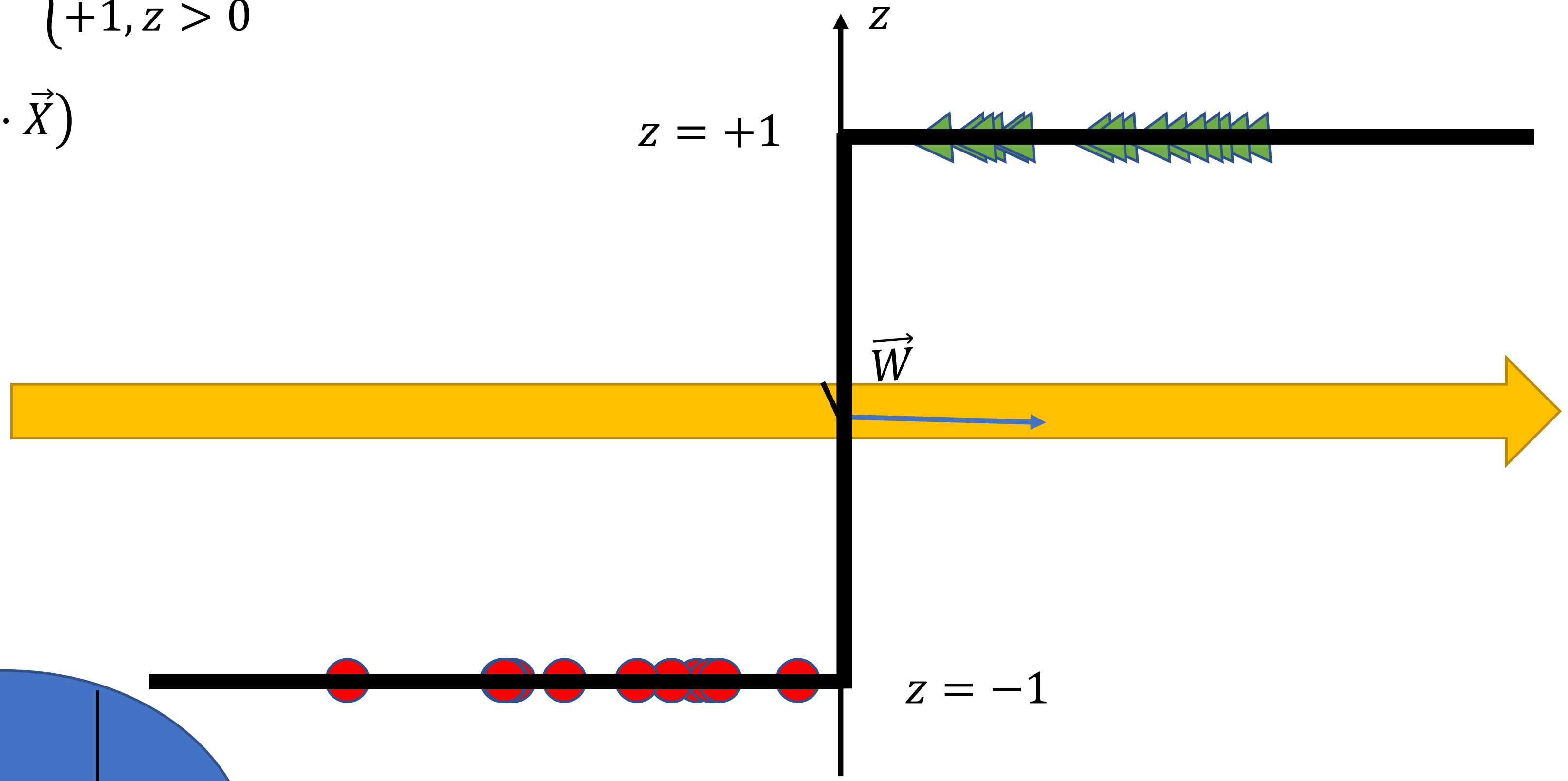


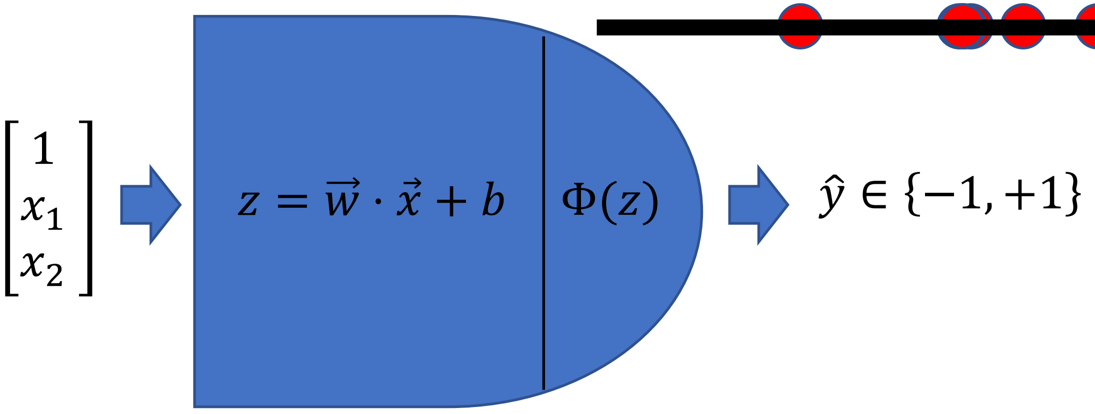
Perceptron's prediction:

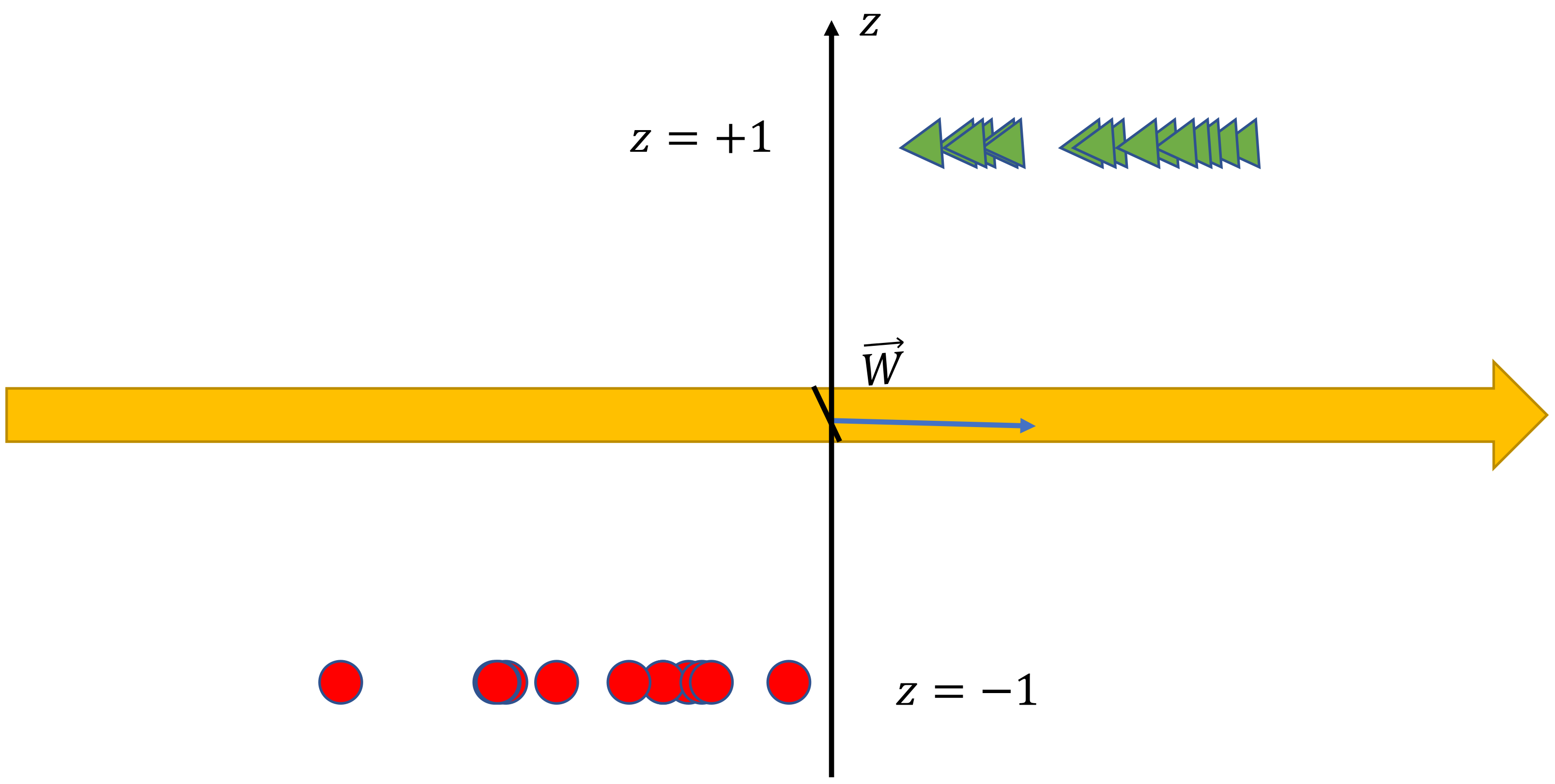
$$z = \vec{w} \cdot \vec{x} + b$$

$$\hat{y} = \Phi(z) = \operatorname{sgn}(z) = \begin{cases} -1, z \leq 0 \\ +1, z > 0 \end{cases}$$

$$\hat{y} = \Phi(z) = \operatorname{sgn}\left(\vec{W} \cdot \vec{X}\right)$$

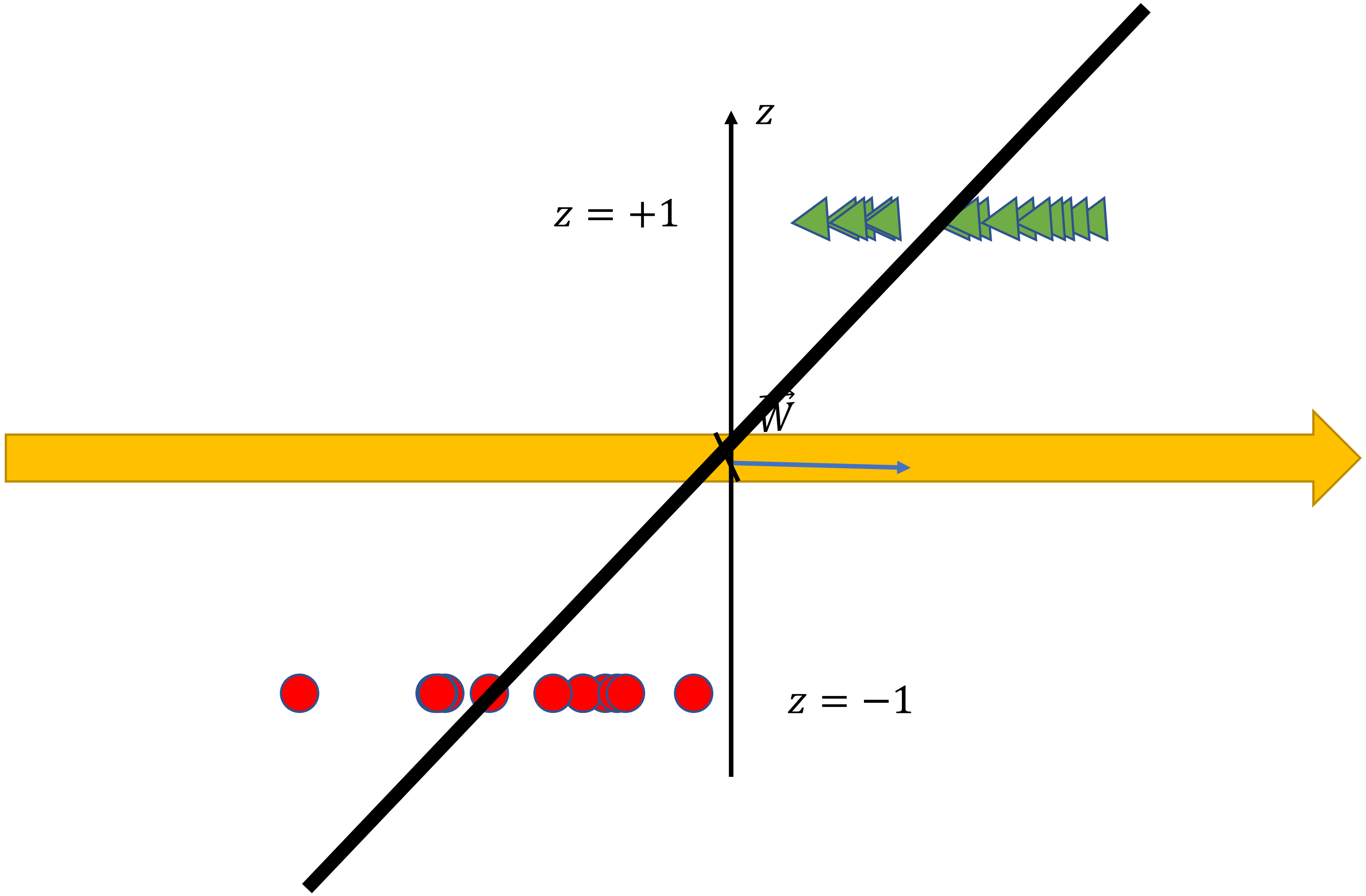




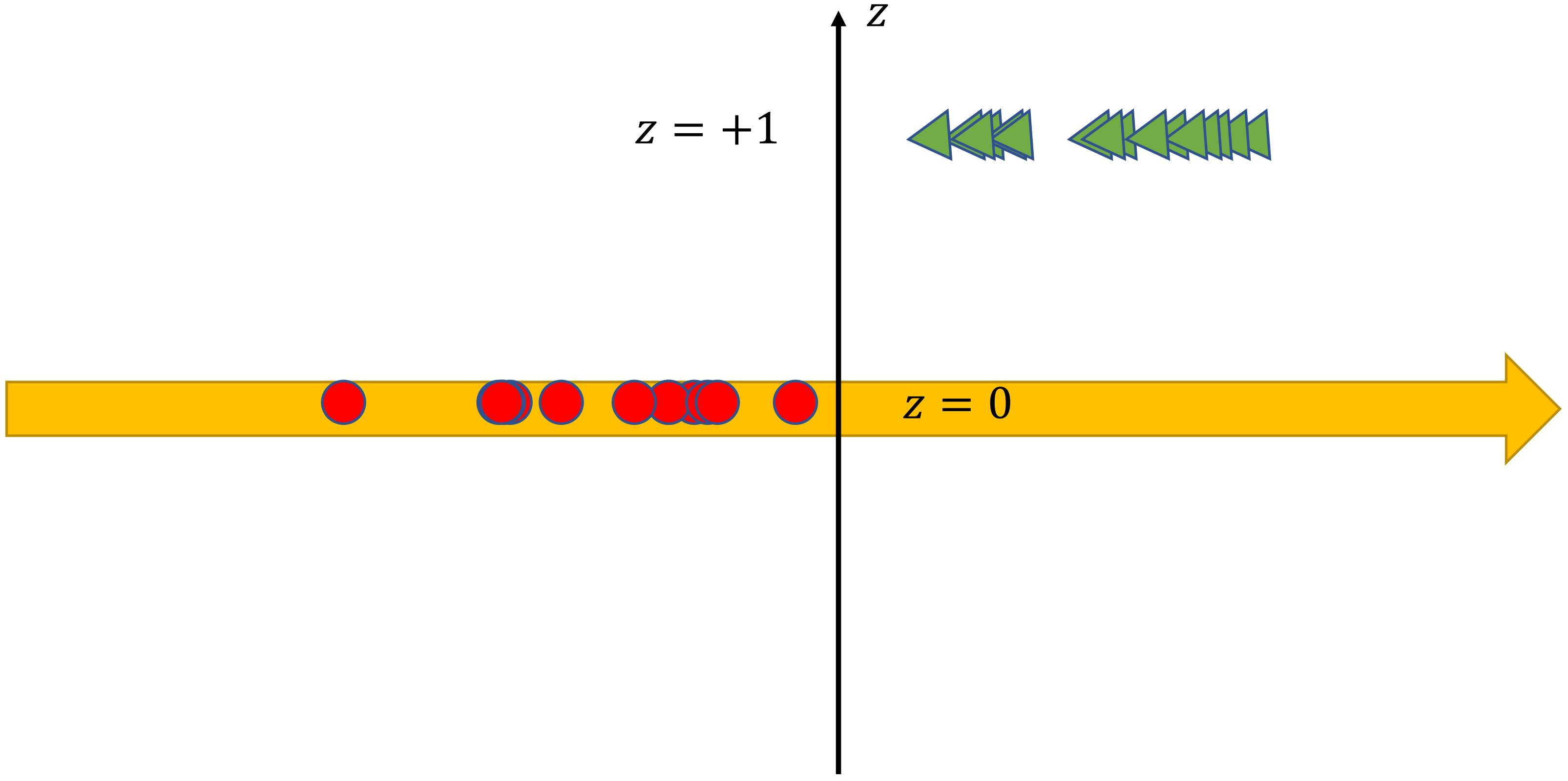


z
$z = +1$
$\vec{W}$
$z = -1$

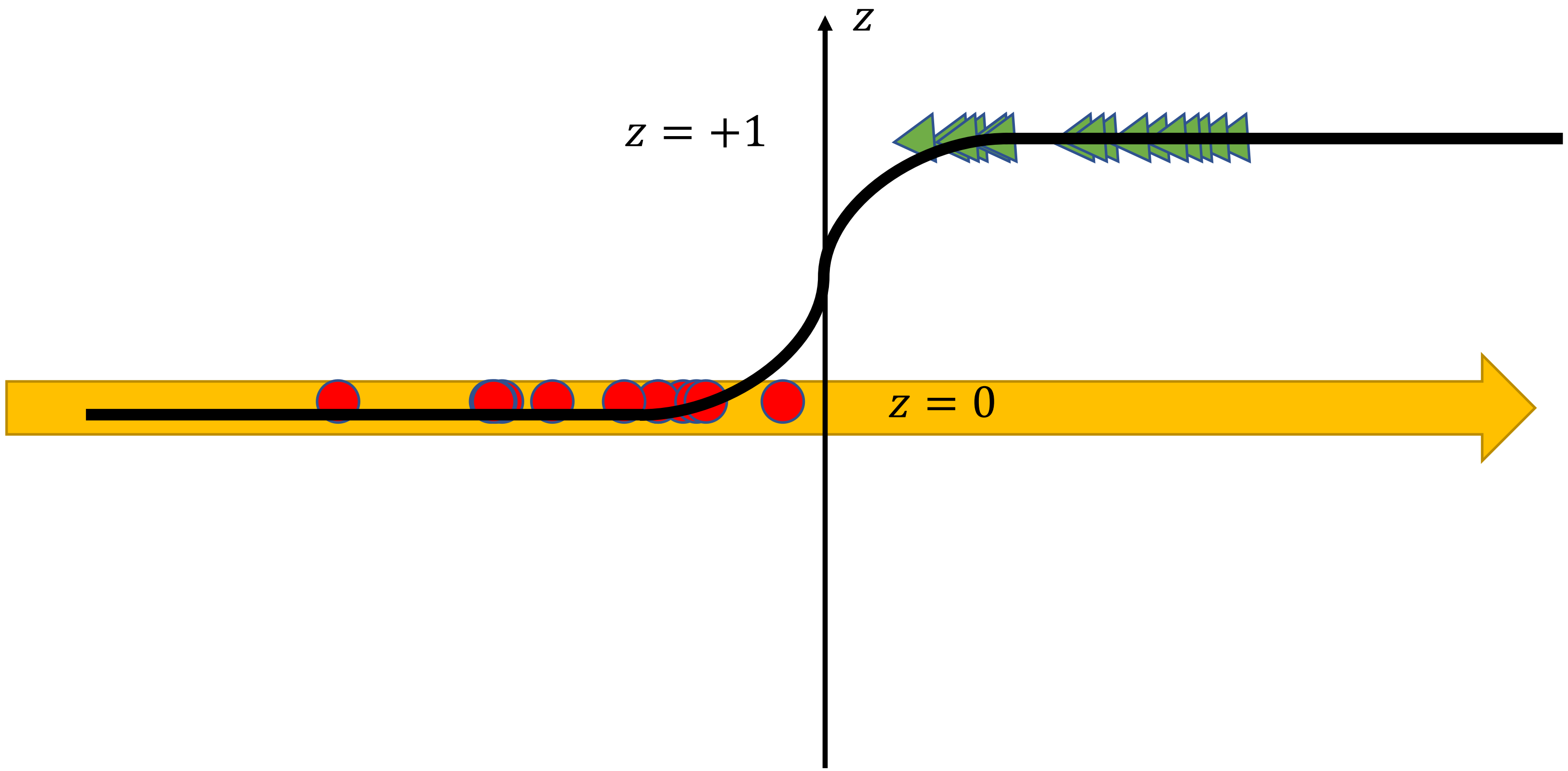
Linear regression?



Probabilistic view



Logistic Regression!



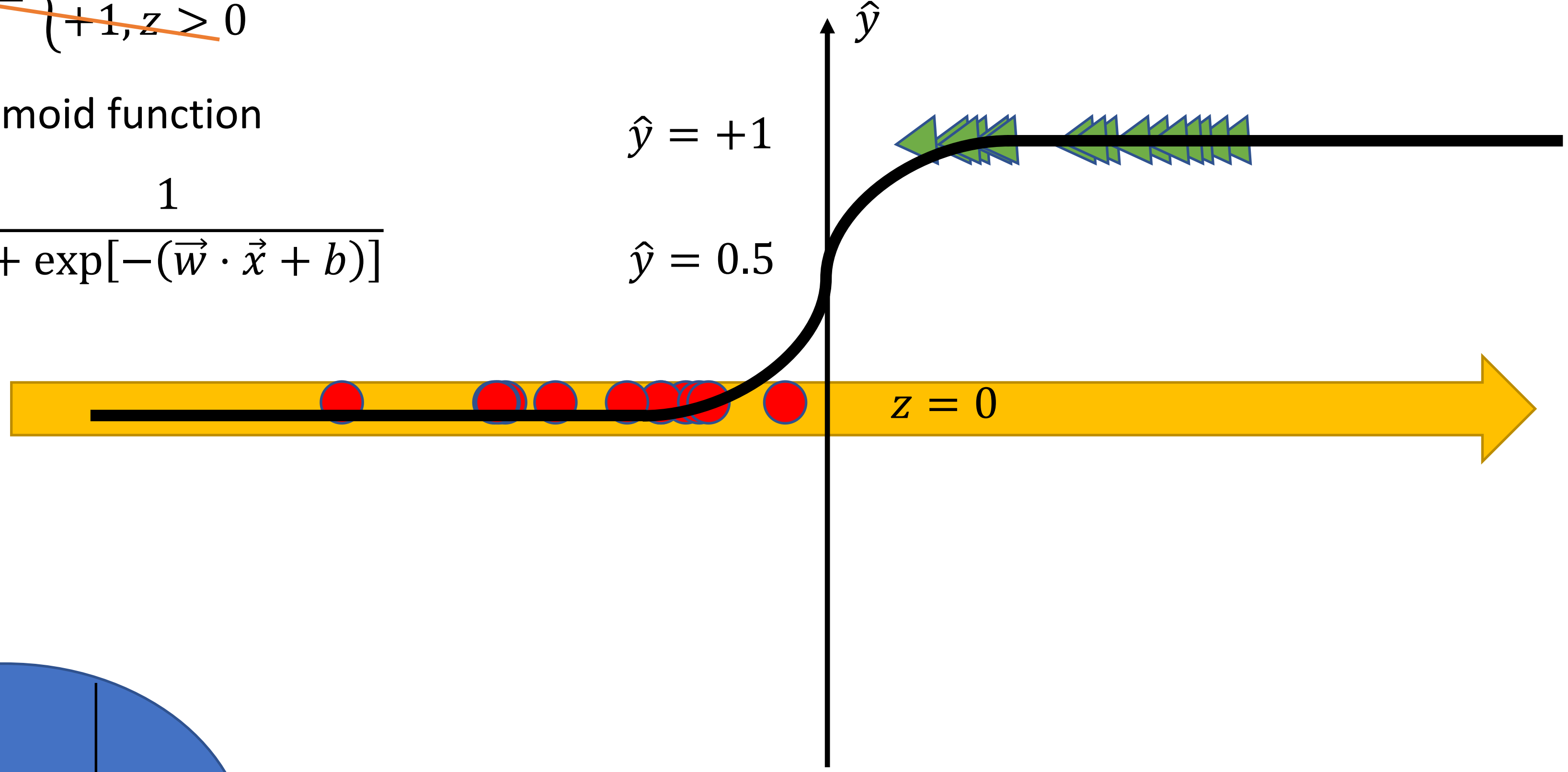
Logistic Regression!

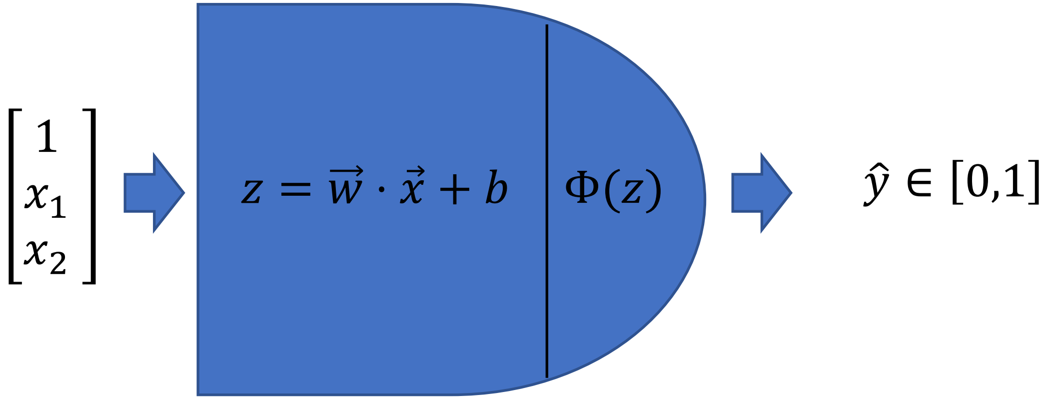
$$z = \vec{w} \cdot \vec{x} + b$$

$$\cancel{\hat{y} = \Phi(z) = \mathrm{sgn}(z) = \begin{cases} -1, z \leq 0 \\ +1, z > 0 \end{cases}}$$

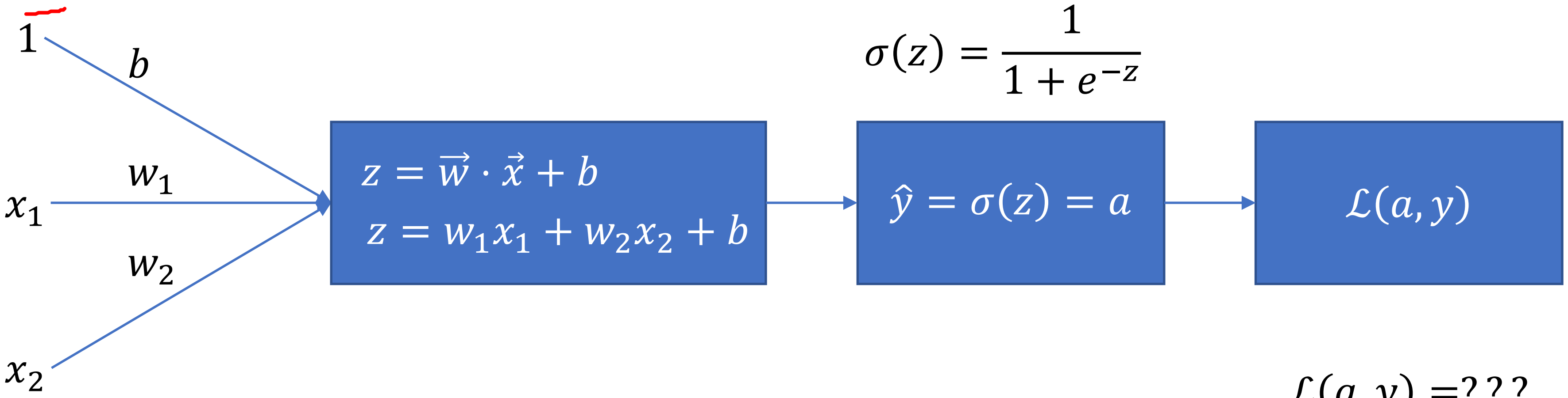
$\hat{y} = \Phi(z) = \sigma(z)$: Sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + \exp[-(\vec{w} \cdot \vec{x} + b)]}$$



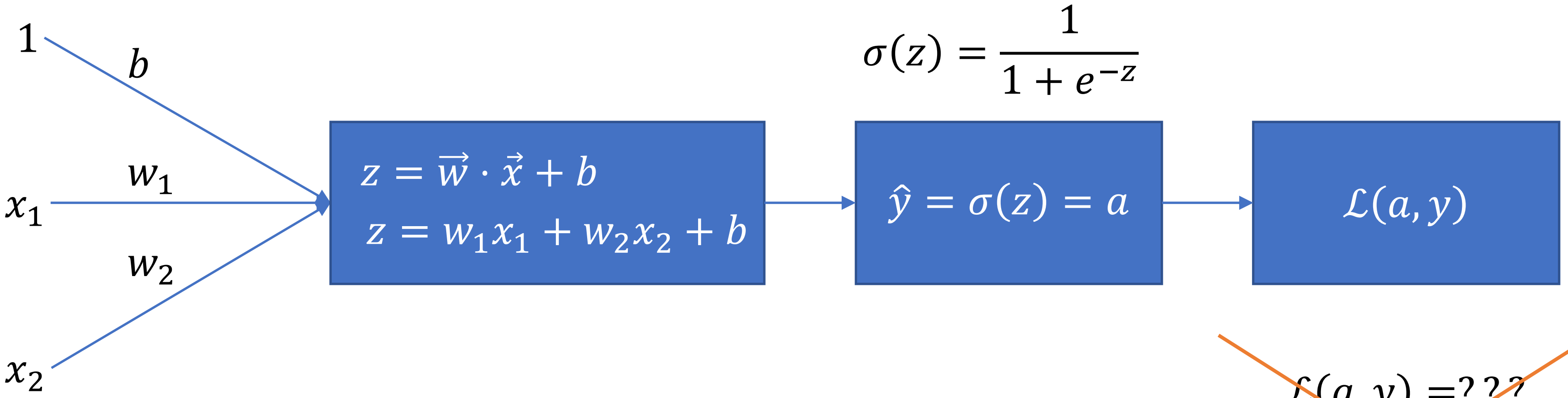


Training



$$\mathcal{L}(a, y) = ???$$

$$\mathcal{L}(a, y) = \text{MSE}??$$



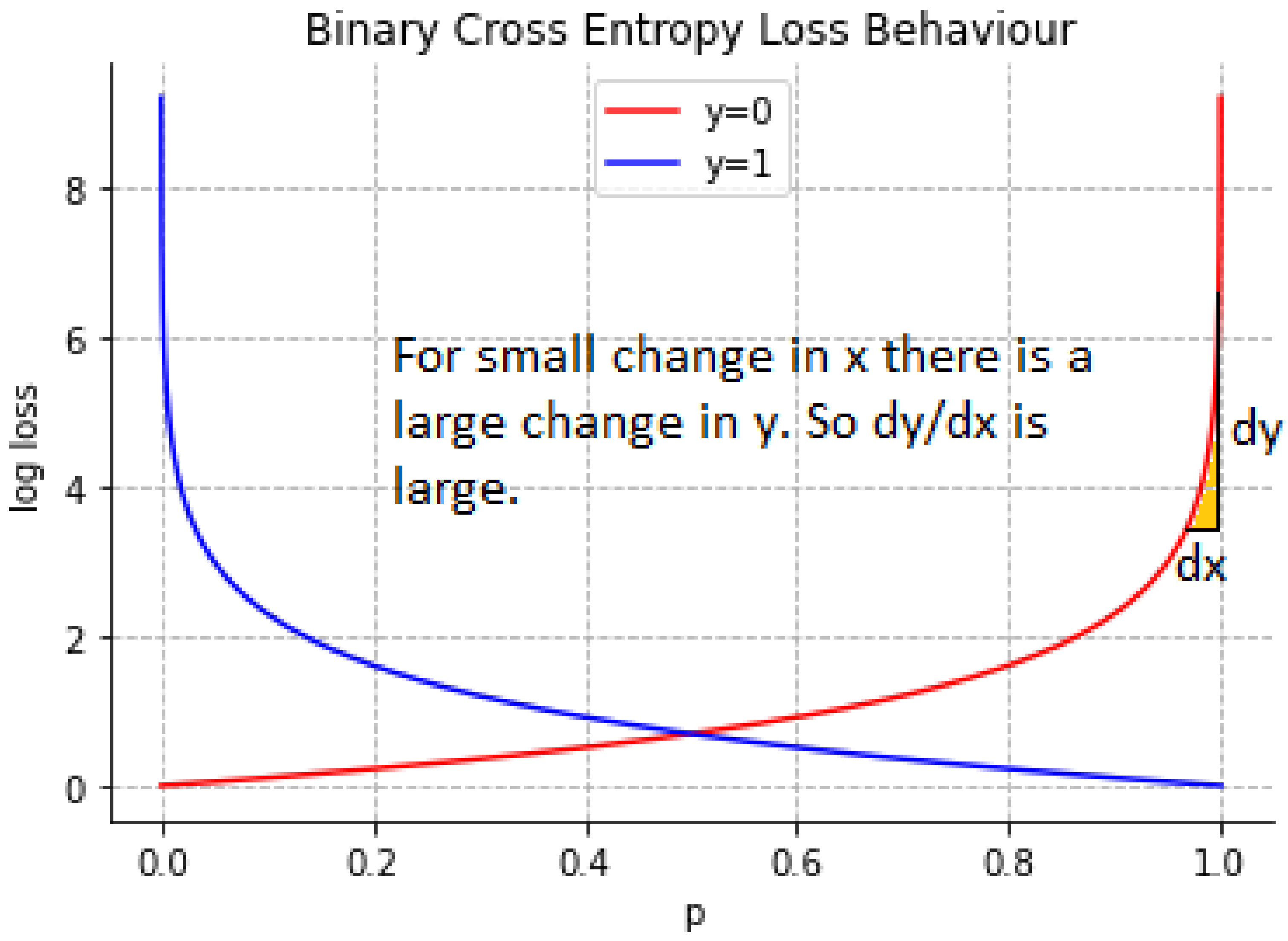
$$\cancel{\mathcal{L}(a, y) = ???}$$

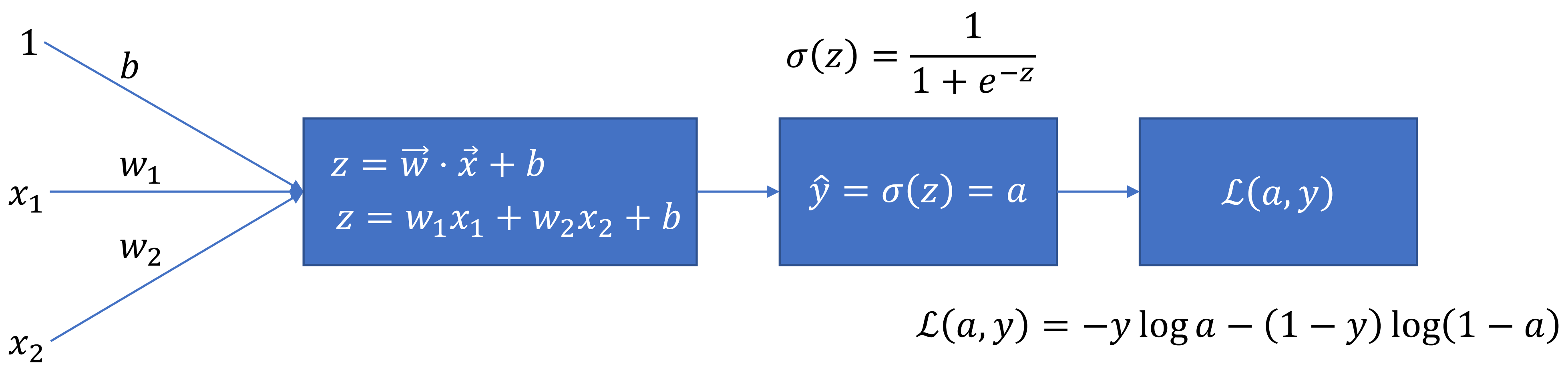
$$\cancel{\mathcal{L}(a, y) = \text{MSE}??}$$

LOG Loss!!!

Log loss

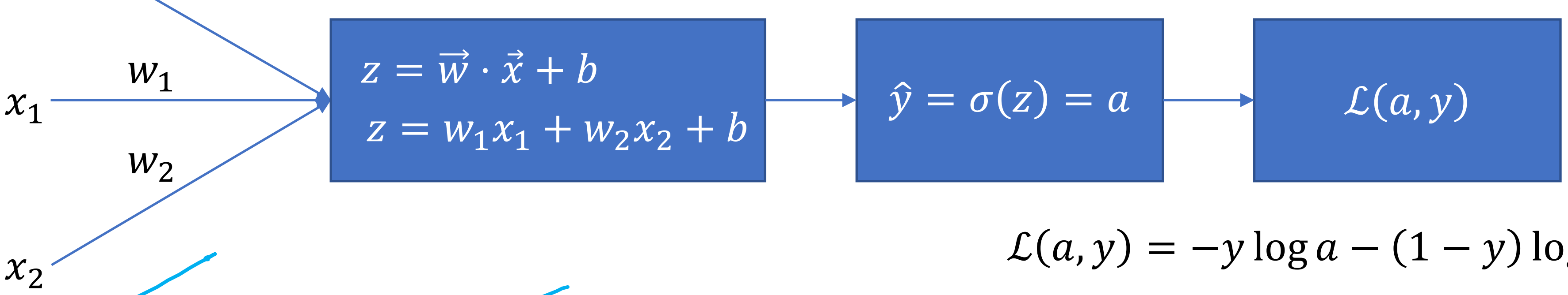
$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$





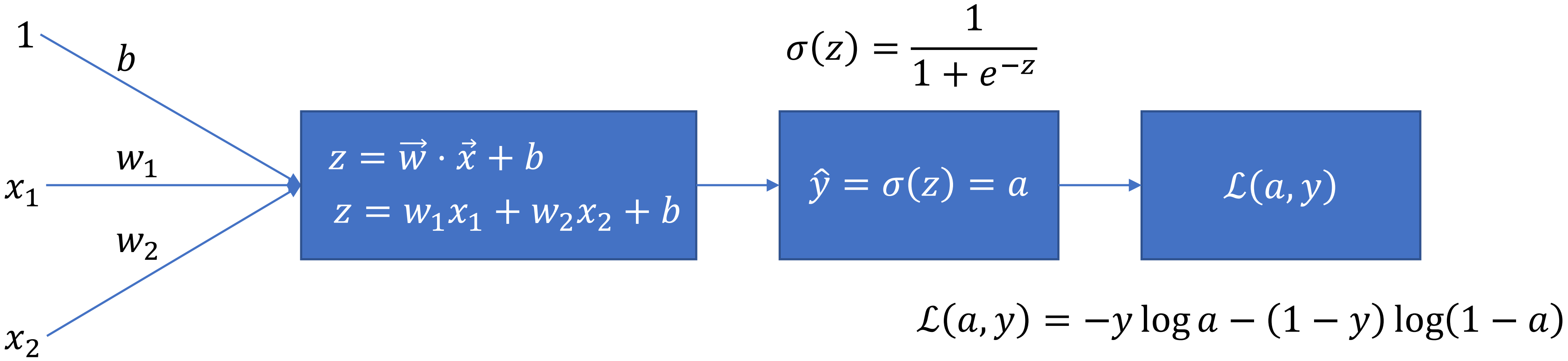
1
b
x_1
w_1
w_2
x_2
$z = \vec{w} \cdot \vec{x} + b$
$z = w_1 x_1 + w_2 x_2 + b$
$\sigma(z) = \frac{1}{1 + e^{-z}}$
$\hat{y} = \sigma(z) = a$
$\mathcal{L}(a, y)$
$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$

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

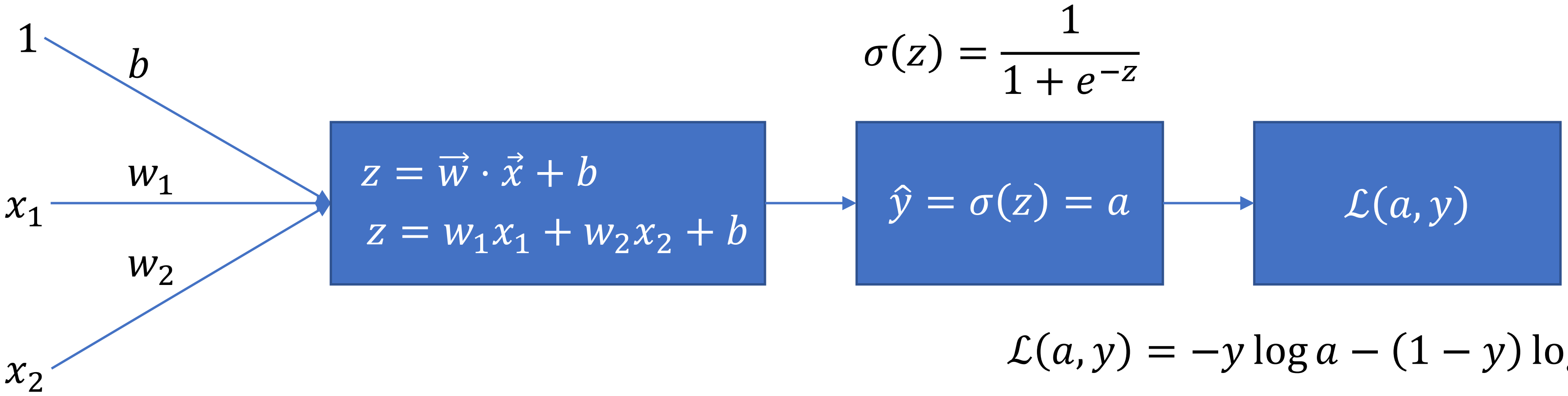


$$\mathcal{L}(a, y) = -y \log a - (1-y) \log(1-a)$$

- $\dfrac{\partial \mathcal{L}(a,y)}{\partial a} = -\dfrac{y}{a} + \dfrac{1-y}{1-a}$

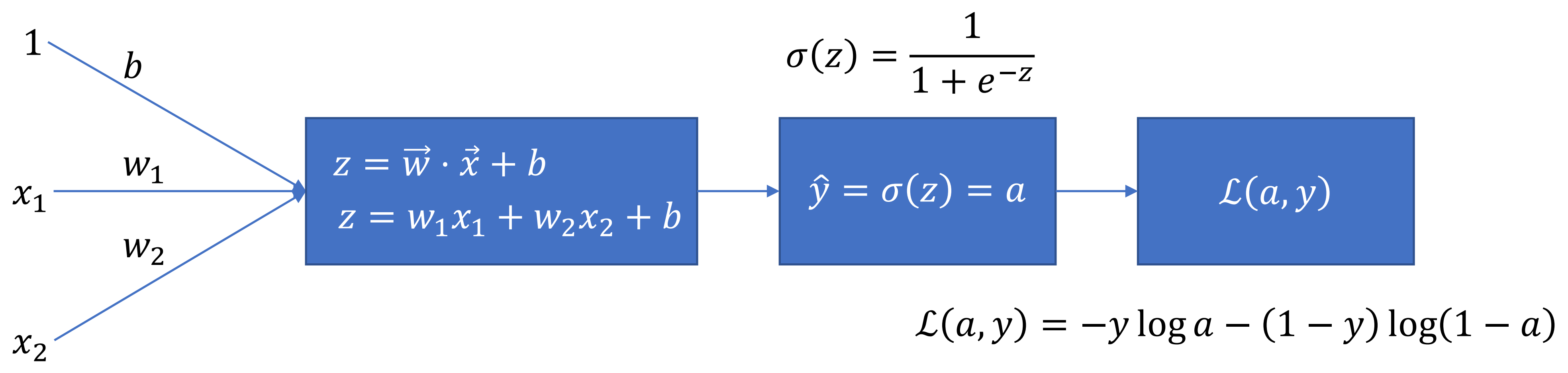


- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1+e^{-z})^{-1}\right) = a[1-a]$

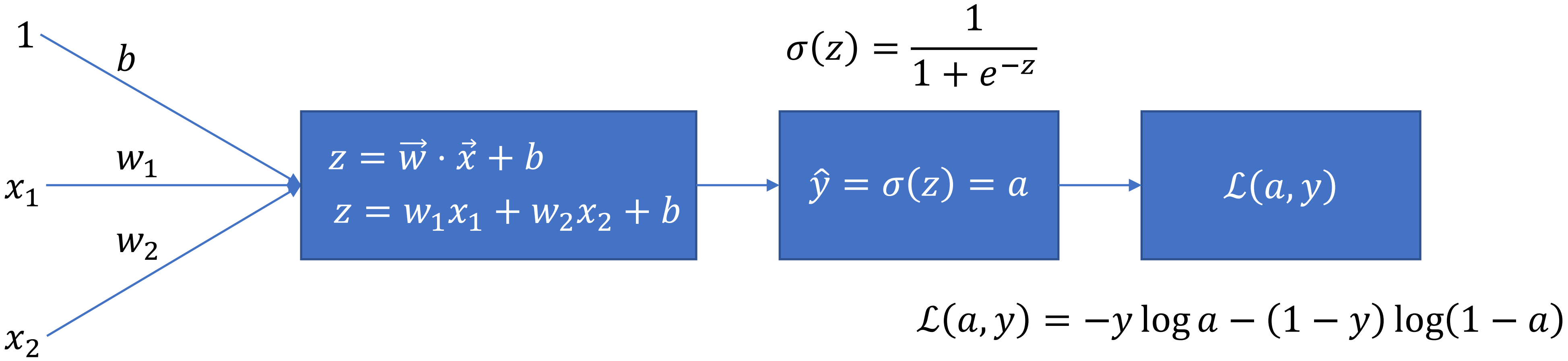


$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

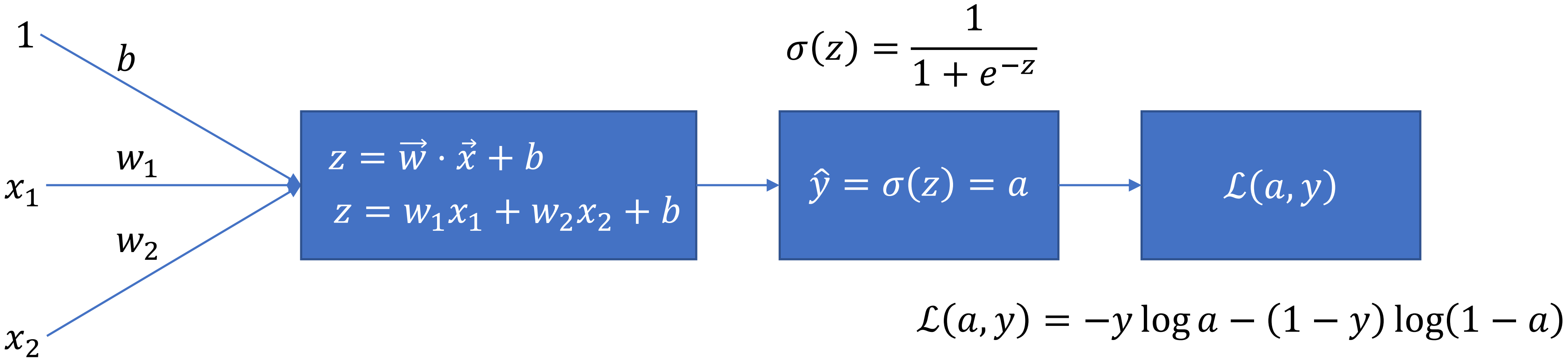
- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1 + e^{-z})^{-1}\right) = a[1 - a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a}\frac{\partial a(z)}{\partial z} = a - y$



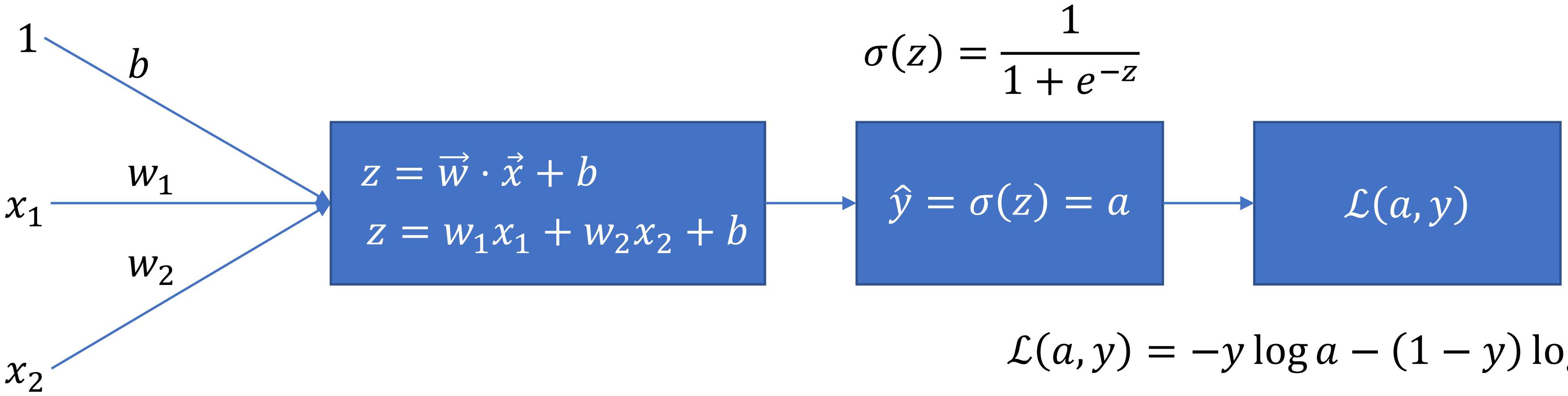
- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1 + e^{-z})^{-1}\right) = a[1 - a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a} \frac{\partial a(z)}{\partial z} = a - y$
- $\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial w_1} = (a - y) x_1$



- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1 + e^{-z})^{-1}\right) = a[1 - a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a}\frac{\partial a(z)}{\partial z} = a - y$
- $\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_1} = (a - y)x_1$
- $\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_2} = (a - y)x_2$



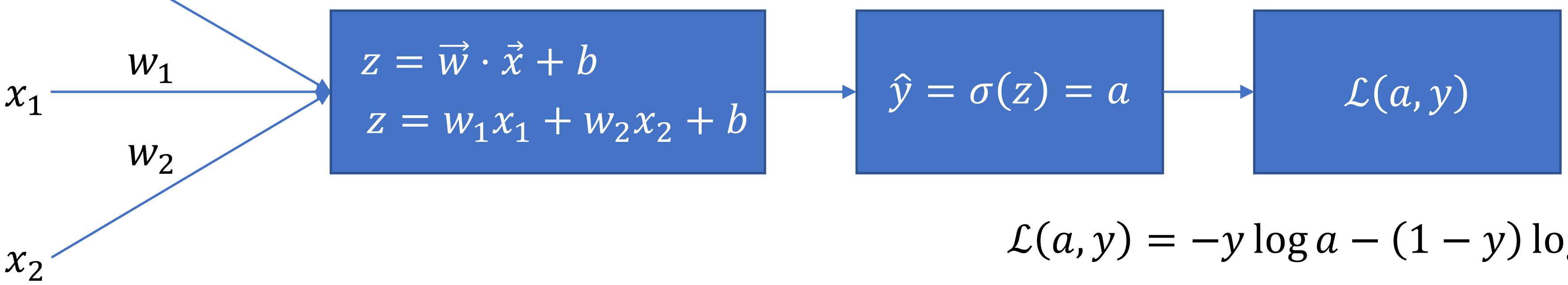
- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1+e^{-z})^{-1}\right) = a[1-a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a}\frac{\partial a(z)}{\partial z} = a - y$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_1} = (a-y)x_1}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_2} = (a-y)x_2}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial b} = (a-y)}$



$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1 + e^{-z})^{-1}\right) = a[1 - a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a} \frac{\partial a(z)}{\partial z} = a - y$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial w_1} = (a - y) x_1}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial w_2} = (a - y) x_2}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial b} = (a - y)}$

- $w_{1,i} \leftarrow w_{1,i-1} - \eta\, \vec{\boldsymbol{\nabla}}_{\vec{\boldsymbol{w}}} \boldsymbol{\mathcal{L}(a, y)}$
- $w_{2,i} \leftarrow w_{2,i-1} - \eta \vec{\boldsymbol{\nabla}}_{\vec{\boldsymbol{w}}} \boldsymbol{\mathcal{L}(a, y)}$
- $b_i \leftarrow b_{i-1} - \eta \vec{\boldsymbol{\nabla}}_{\vec{\boldsymbol{w}}} \boldsymbol{\mathcal{L}(a, y)}$

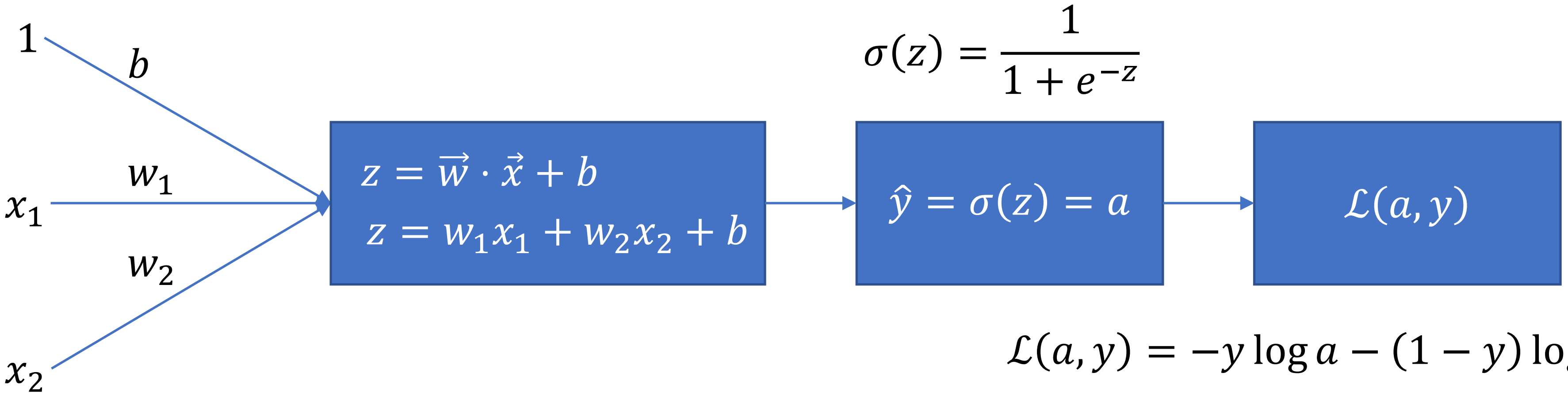


$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1 + e^{-z})^{-1}\right) = a[1 - a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a}\frac{\partial a(z)}{\partial z} = a - y$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_1} = (a - y)x_1}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_2} = (a - y)x_2}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial b} = (a - y)}$

- $w_{1,i} \leftarrow w_{1,i-1} - \eta(\boldsymbol{a} - \boldsymbol{y})\boldsymbol{x_1}$
- $w_{2,i} \leftarrow w_{2,i-1} - \eta(\boldsymbol{a} - \boldsymbol{y})\boldsymbol{x_2}$
- $b_i \leftarrow b_{i-1} - \eta(\boldsymbol{a} - \boldsymbol{y})$

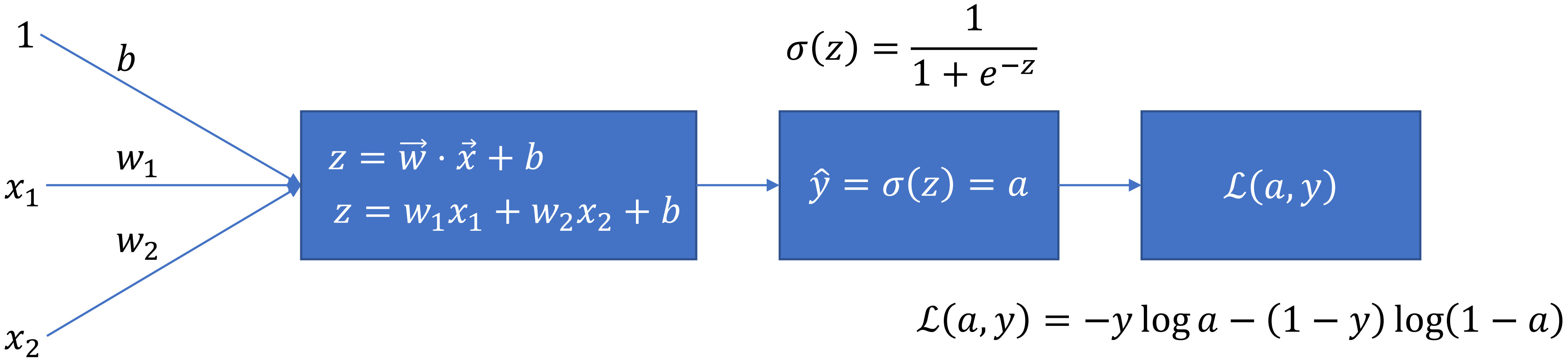


$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

- $\frac{\partial \mathcal{L}(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$
- $\frac{\partial a(z)}{\partial z} = \frac{\partial}{\partial z}\left((1 + e^{-z})^{-1}\right) = a[1 - a]$
- $\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a}\frac{\partial a(z)}{\partial z} = a - y$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_1} = (a - y)x_1}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_2} = (a - y)x_2}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial b} = (a - y)}$

- $w_{1,i} \leftarrow w_{1,i-1} - \eta(\boldsymbol{a} - \boldsymbol{y})\boldsymbol{x_1}$
- $w_{2,i} \leftarrow w_{2,i-1} - \eta(\boldsymbol{a} - \boldsymbol{y})\boldsymbol{x_2}$
- $b_i \leftarrow b_{i-1} - \eta(\boldsymbol{a} - \boldsymbol{y})$

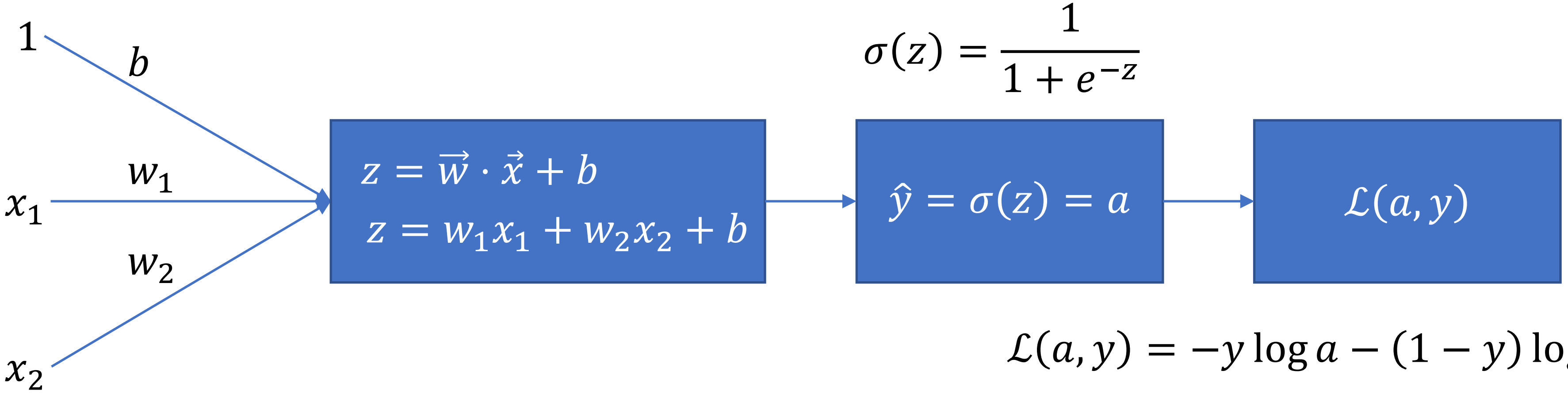
This is the loss related to 1 sample!



- Cost of *m* samples:

$$P(\cdot) = \prod_{i=1}^{m} P\left(y^{(i)} | \vec{x}^{(i)}\right) \qquad \text{(Maximum Likelihood Estimation)}$$

- The MLE estimates the ser of parameters $\vec{\theta} = (\vec{w}, b)$ so that they are as close as possible to the real set of parameters $\vec{\theta}$ that generated such data samples by minimizing the loss $\mathcal{L}(y, \hat{y})$.



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

$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

- Cost of *m* samples:

$$\log P(\cdot) = \log \prod_{i=1}^{m} P\left(y^{(i)} | \vec{x}^{(i)}\right)$$

$$\log P(\cdot) = \sum_{i=1}^{m} \log P\left(y^{(i)} | \vec{x}^{(i)}\right)$$

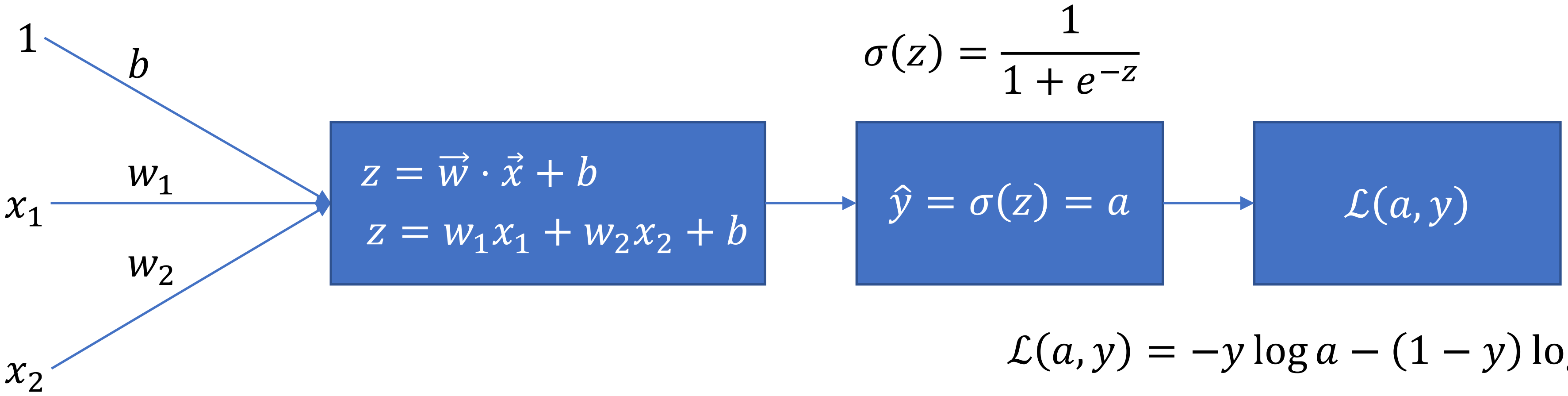
$$\log P(\cdot) = -\sum_{i=1}^{m} \mathcal{L}\left(y^{(i)} | \vec{x}^{(i)}\right)$$

- Cost:

$$\boxed{J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(y^{(i)}, \hat{y}^{(i)}\right)}$$

Minimize the cost =

= Maximize the likelihood.

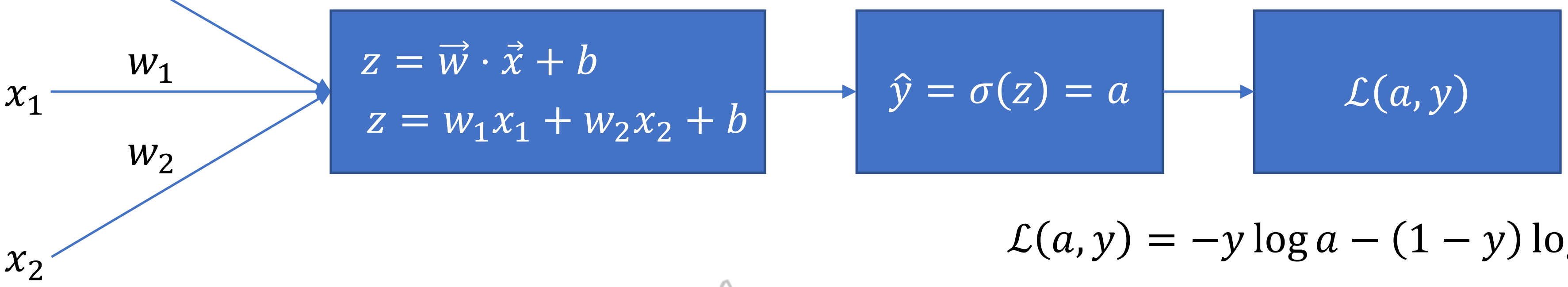


$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

- Objective:
  - Find $\vec{w}, b$ that minimize

$$\boxed{J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(y^{(i)}, \hat{y}^{(i)}\right)}$$

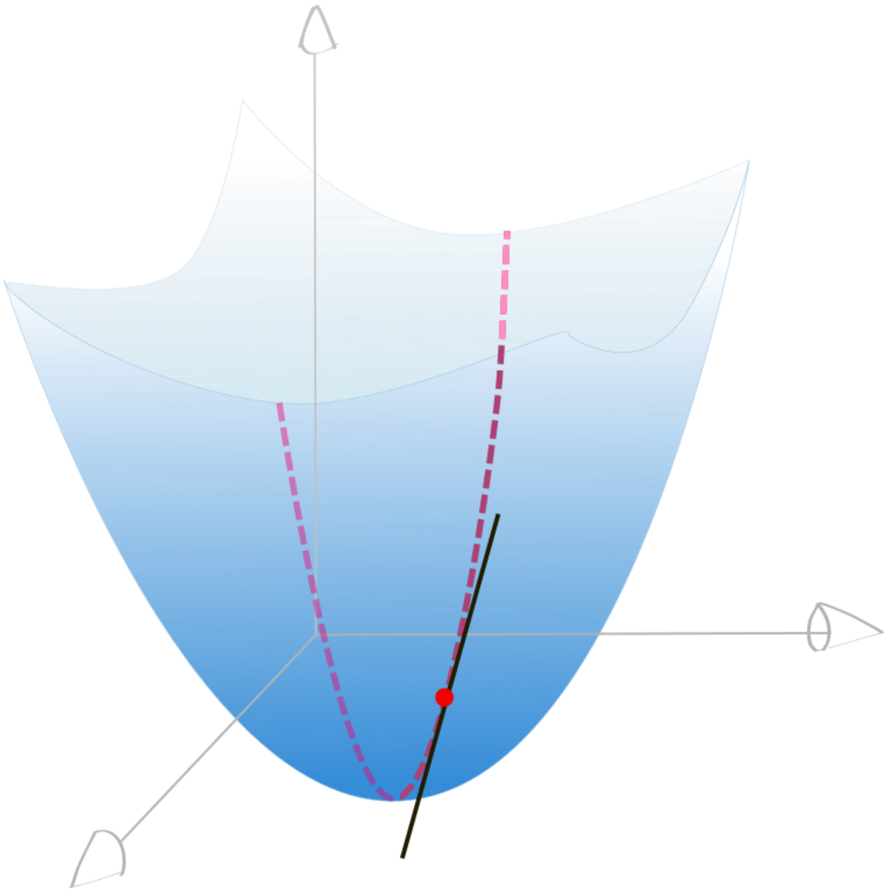
$$\sigma(z) = \frac{1}{1+e^{-z}}$$



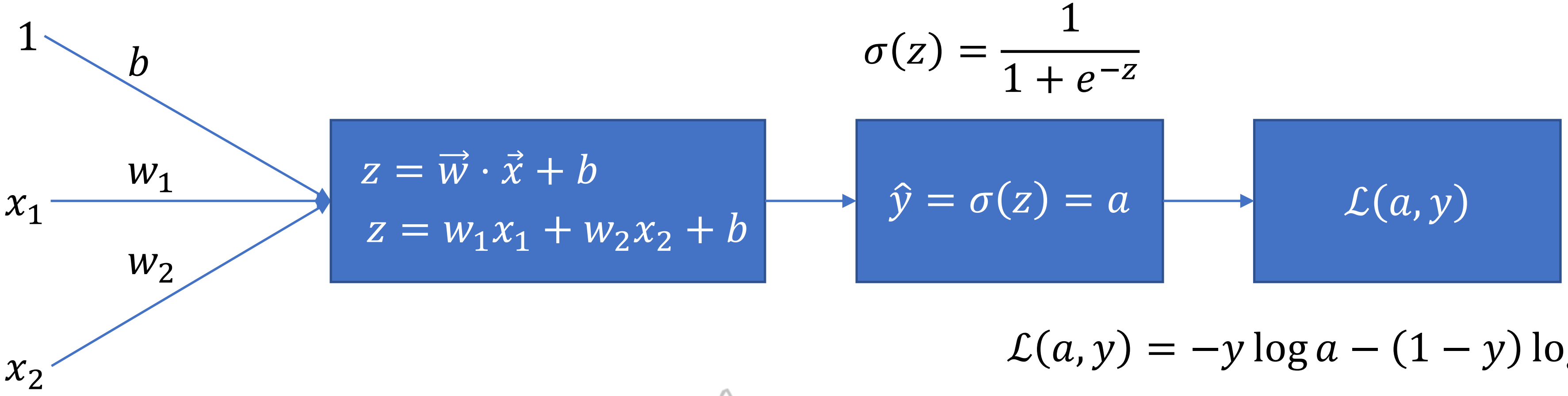
$$\mathcal{L}(a, y) = -y \log a - (1-y) \log(1-a)$$

- Objective:
  - Find $\vec{w}, b$ that minimize

$$\boxed{J(\vec{w}, b) = \sum \mathcal{L}(y, \hat{y})}$$



1. Start at any point.
2. Walk in the direction of steepest descent, ie.,



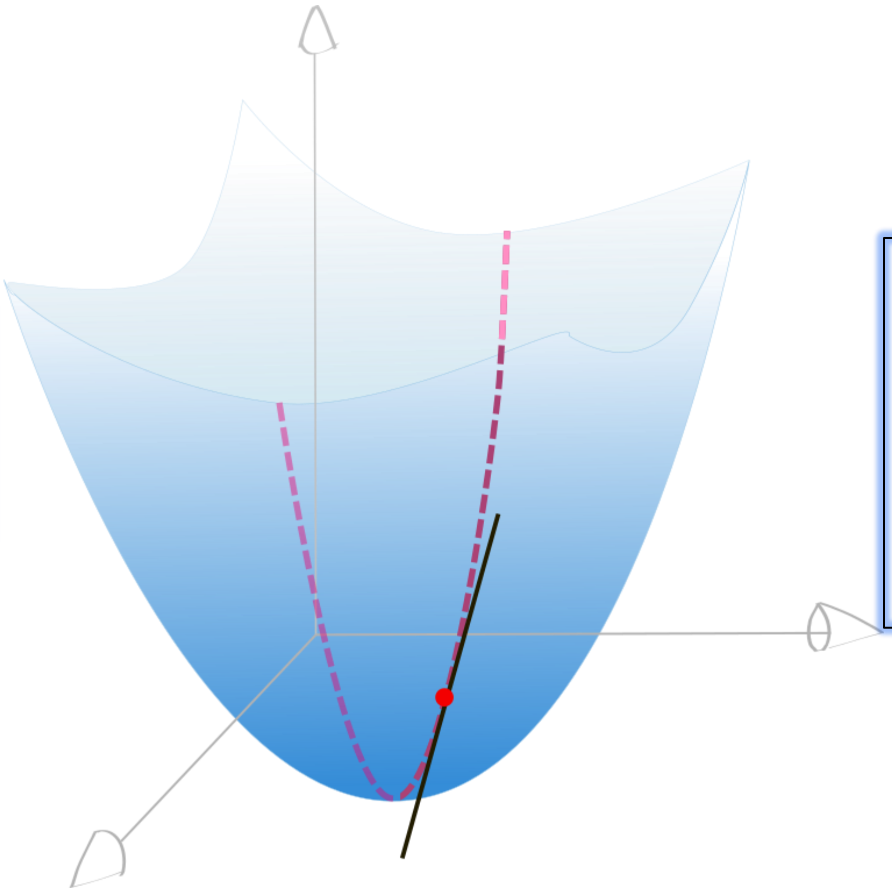
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

- Objective:
  - Find $\vec{w}, b$ that minimize

$$\boxed{J(\vec{w}, b) = \sum \mathcal{L}(y, \hat{y})}$$

1. Start at any point.
2. Walk in the direction of steepest descent, ie.,



$$\vec{w} \leftarrow \vec{w} - \alpha \vec{\nabla}_{\vec{w}} J(\vec{w}, b)$$

$$b \leftarrow b - \alpha \frac{d}{db} J(\vec{w}, b)$$

Notation

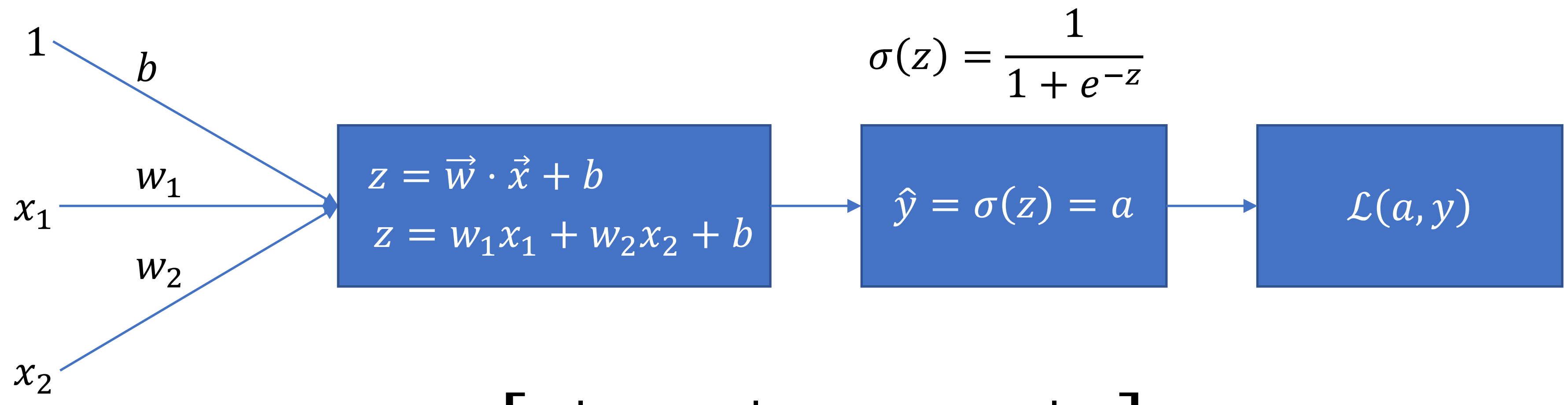
- Single Training Sample: $(\vec{x}, y)$, where:
  - $\vec{x} \in \Re^{n_x}$.
  - $y \in \{0,1\}$.

- Single Training Sample: $(\vec{x}, y)$, where:
  - $\vec{x} \in \Re^{n_x}$.
  - $y \in \{0,1\}$.
- m training samples: $\left\{\left(\vec{x}^{(1)}, y^{(1)}\right), \left(\vec{x}^{(2)}, y^{(2)}\right), \dots, \left(\vec{x}^{(m)}, y^{(m)}\right)\right\} = \left\{\vec{x}^{(i)}, y^{(i)}\right\}_{i=1}^{m}$, where:
  - m_{train} = Number of training samples.
  - m_{test} = Number of test samples.

- Single Training Sample: $(\vec{x}, y)$, where:
  - $\vec{x} \in \Re^{n_x}$.
  - $y \in \{0,1\}$.
- m training samples: $\{(\vec{x}^{(1)}, y^{(1)}), (\vec{x}^{(2)}, y^{(2)}), \ldots, (\vec{x}^{(m)}, y^{(m)})\} = \{\vec{x}^{(i)}, y^{(i)}\}_{i=1}^{m}$, where:
  - m_{train} = Number of training samples.
  - m_{test} = Number of test samples.
- $$X = \underbrace{\begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \ldots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix}}_{m} \Bigg\} n_x \qquad Y = \begin{bmatrix} y^{(1)} & y^{(2)} & \ldots & y^{(m)} \end{bmatrix}$$
- $X \in \Re^{(n_x \times m)}$, i.e., `X.shape=(n_x, m)`.
- $Y \in \Re^{(1 \times m)}$, i.e., `Y.shape = (1, m)`.

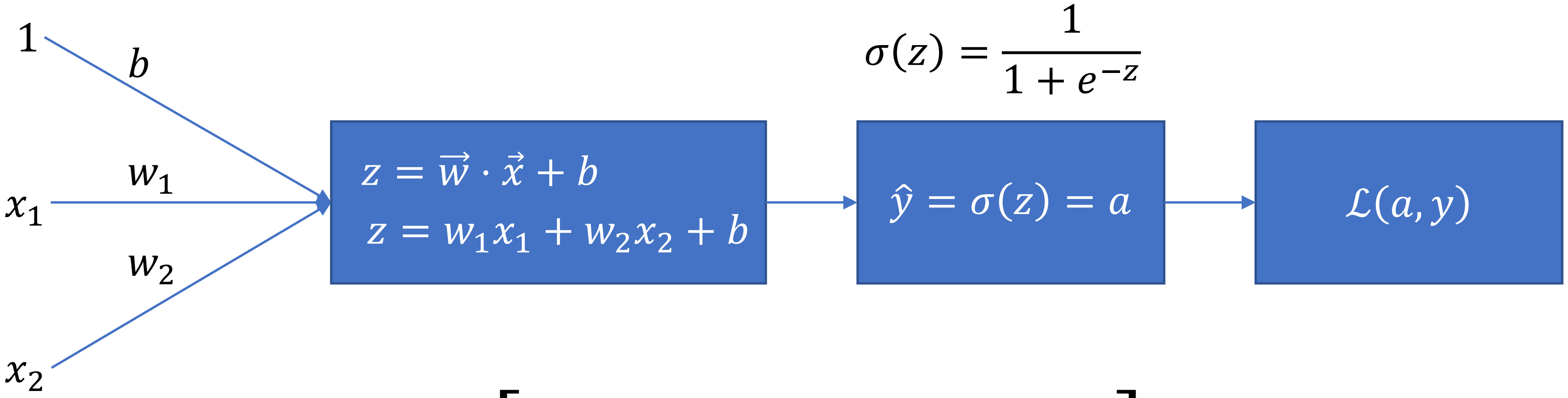
Gradient Descent on Logistic Regression

Forward Propagation for m examples



$$[w_1 \quad w_2 \quad \cdots \quad w_{n_x}] \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \dots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix} + [b \quad b \quad \cdots \quad b] =$$

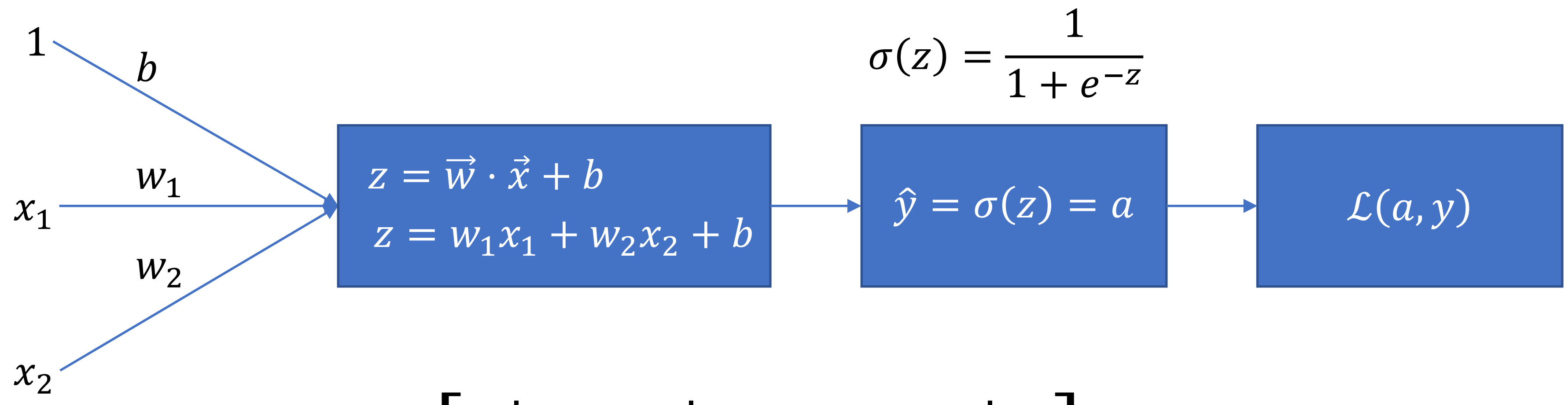
Forward Propagation for m examples



$$[w_1 \quad w_2 \quad \cdots \quad w_{n_x}] \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \dots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix} + [b \quad b \quad \cdots \quad b] =$$

$$= [z^{(1)} \quad z^{(2)} \quad \dots \quad z^{(m)}]$$

Forward Propagation for m examples



$$[w_1 \quad w_2 \quad \cdots \quad w_{n_x}]_{1\times n_x} \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \dots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix}_{n_x\times m} + [b \quad b \quad \cdots \quad b]_{1\times m} =$$

$$= \left[z^{(1)} \quad z^{(2)} \quad \dots \quad z^{(m)}\right]$$

Forward Propagation for *m* examples

$$[w_1 \quad w_2 \quad \cdots \quad w_{n_x}]_{1\times n_x} \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \dots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix}_{n_x \times m} + [b \quad b \quad \cdots \quad b]_{1\times m} =$$

$$= Z = \left[z^{(1)} \quad z^{(2)} \quad \dots \quad z^{(m)}\right]$$

Forward Propagation for *m* examples

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_{n_x} \end{bmatrix}_{1\times n_x} \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \dots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix}_{n_x\times m} + \begin{bmatrix} b & b & \cdots & b \end{bmatrix}_{1\times m} =$$

$$= Z = \begin{bmatrix} z^{(1)} & z^{(2)} & \dots & z^{(m)} \end{bmatrix}$$

$$A\sigma(Z) = \begin{bmatrix} a^{(1)} & a^{(2)} & \dots & a^{(m)} \end{bmatrix}$$

$$\mathcal{L}(a, y)$$

$$\text{A} = \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix}$$

Forward Propagation for *m* examples

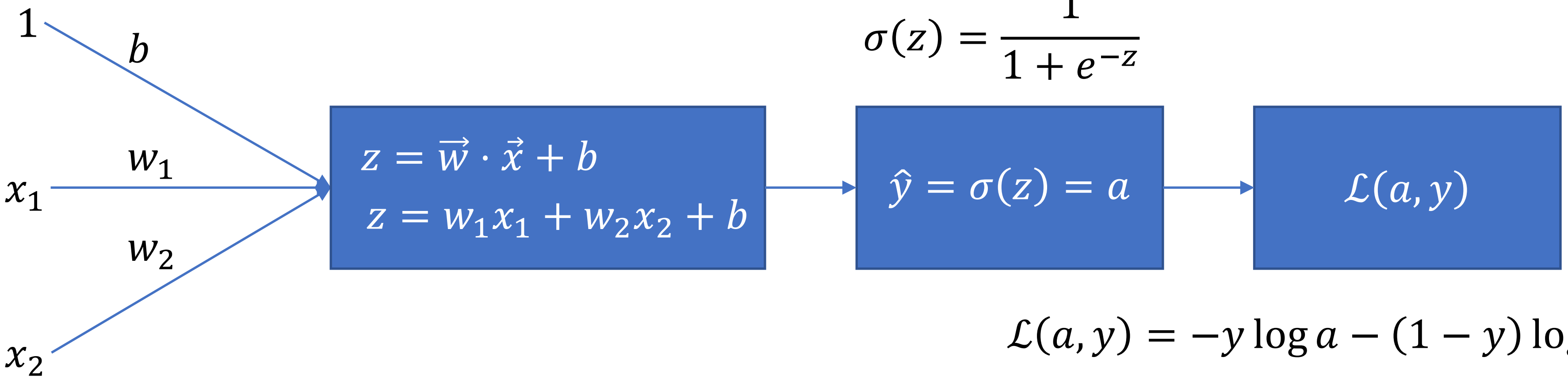
$$[w_1 \quad w_2 \quad \cdots \quad w_{n_x}]_{1\times n_x} \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \cdots & \vec{x}^{(m)} \\ | & | & & | \end{bmatrix}_{n_x\times m} + [b \quad b \quad \cdots \quad b]_{1\times m} =$$

$$= Z = \left[z^{(1)} \quad z^{(2)} \quad \cdots \quad z^{(m)}\right]$$

$$\frac{\partial \mathcal{L}(a,y)}{\partial z} = \frac{\partial \mathcal{L}(a,y)}{\partial a}\frac{\partial a(z)}{\partial z} = a - y$$

$$dZ = A - Y = \left[a^{(1)} - y^{(1)} \quad a^{(2)} - y^{(2)} \quad \cdots \quad a^{(m)} - y^{(m)}\right]$$

$$\left\{ \begin{array}{l} A\sigma(Z) = \left[a^{(1)} \quad a^{(2)} \quad \cdots \quad a^{(m)}\right] \\ \Updownarrow \mathcal{L}(a,y) \\ \mathrm{A} = \left[y^{(1)} \quad y^{(2)} \quad \cdots \quad y^{(m)}\right] \end{array} \right.$$



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

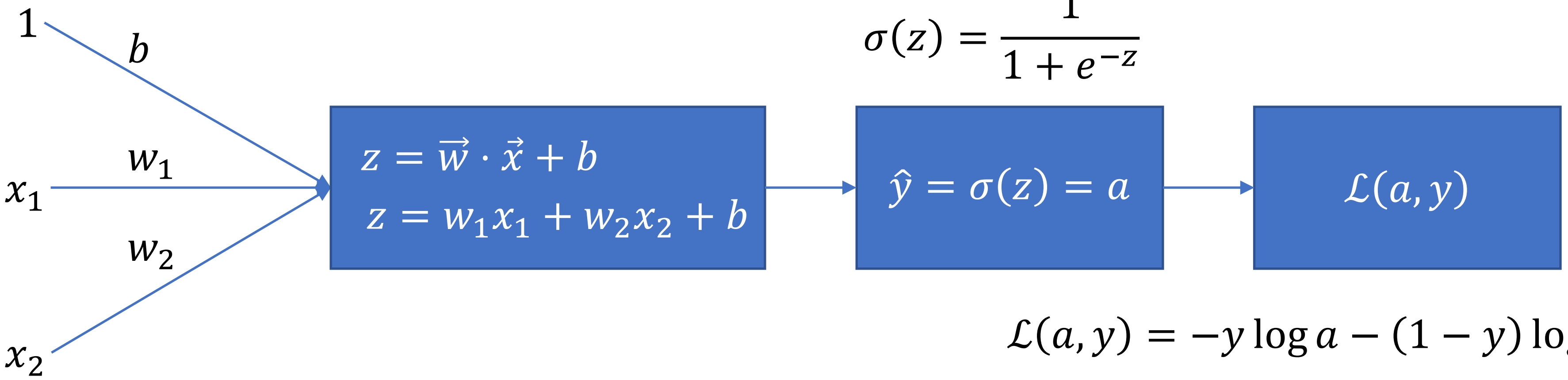
$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

Recall the gradient descent for 1 sample:

- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial w_1} = (a - y) x_1}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial w_2} = (a - y) x_2}$
- $\boldsymbol{\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z} \frac{\partial z}{\partial b} = (a - y)}$

- $w_{1,i} \leftarrow w_{1,i-1} - \eta(\boldsymbol{a - y})\boldsymbol{x_1}$
- $w_{2,i} \leftarrow w_{2,i-1} - \eta(\boldsymbol{a - y})\boldsymbol{x_2}$
- $b_i \leftarrow b_{i-1} - \eta(\boldsymbol{a - y})$

This is the loss related to 1 sample!



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

$$\mathcal{L}(a, y) = -y \log a - (1 - y) \log(1 - a)$$

Gradient descent for m samples:

$$\boxed{J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(y^{(i)}, \hat{y}^{(i)}\right)}$$

$$\frac{\partial}{\partial w_k} J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(y^{(i)}, \hat{y}^{(i)}\right)$$

Update:

$$\left[w_k \leftarrow w_k - \eta \frac{\partial}{\partial w_k} J(\vec{w}, b) \right]_{k=1}^{n_x}$$

$$b \leftarrow b - \eta \frac{\partial}{\partial b} J(\vec{w}, b)$$

Thank you