

Week 2: Linear Models

Matthew Caldwell

COMP0088 Introduction to Machine Learning • UCL Computer Science • Autumn 2023

Admin

- Pulse surveys!
- Triage?

Week 2 Recap

This One Weird Trick

$$\theta^* = \operatorname*{argmin}_{\theta} L(f, \theta, \{X[, Y]\})$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} L(\mathbf{f}, \mathbf{w}, \{\mathbf{X}, \mathbf{Y}\})$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} L(\mathbf{f}, \mathbf{w}, \{\mathbf{X}, \mathbf{Y}\})$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} L(\mathbf{f}, \mathbf{w}, \mathbf{X}, \mathbf{y})$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} L(\mathbf{w}, \mathbf{X}, \mathbf{y})$$

↓ Feature dimensions ↓

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix}$$

↑ Samples ↑

“Design Matrix”



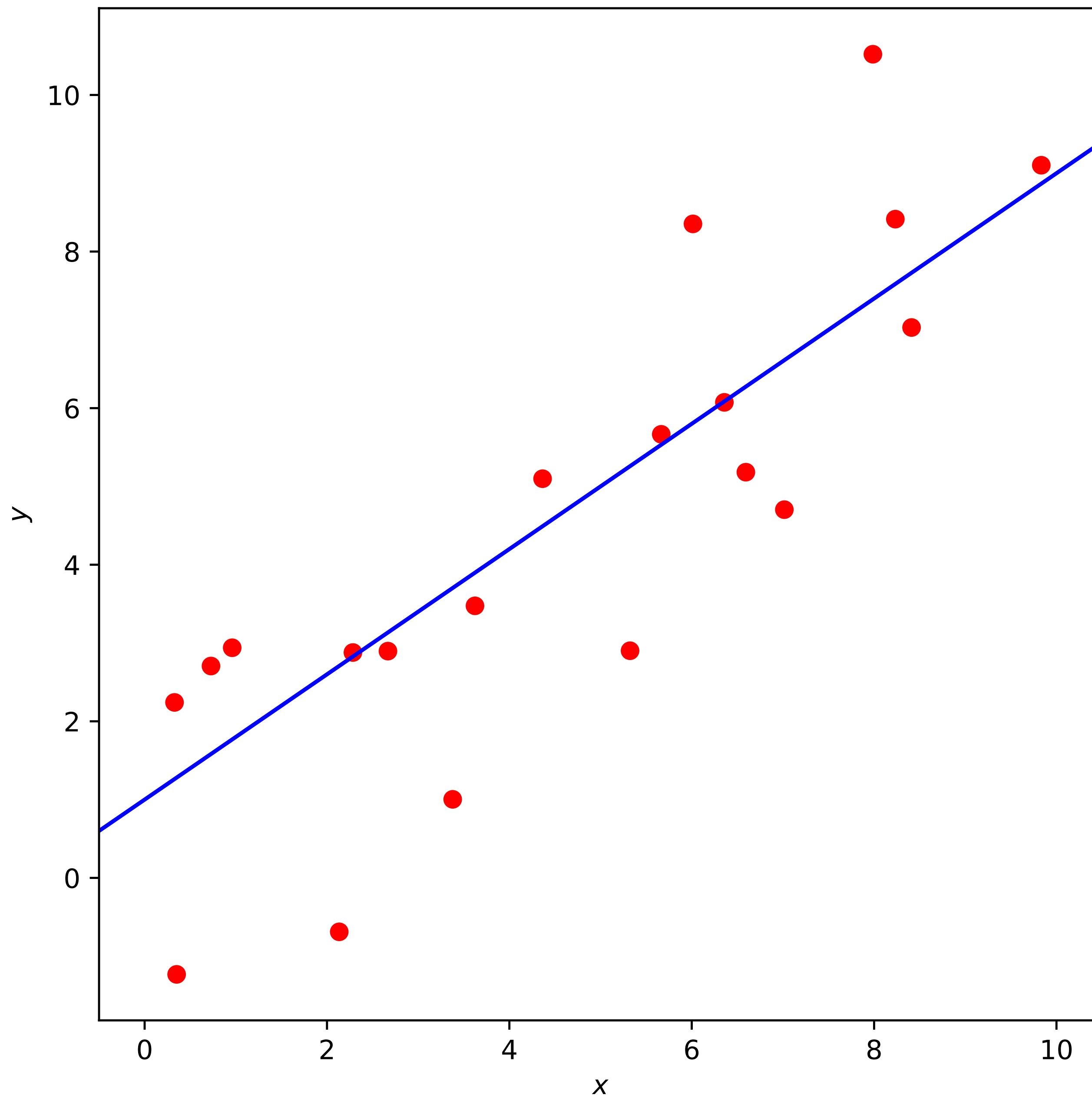
Meerkats

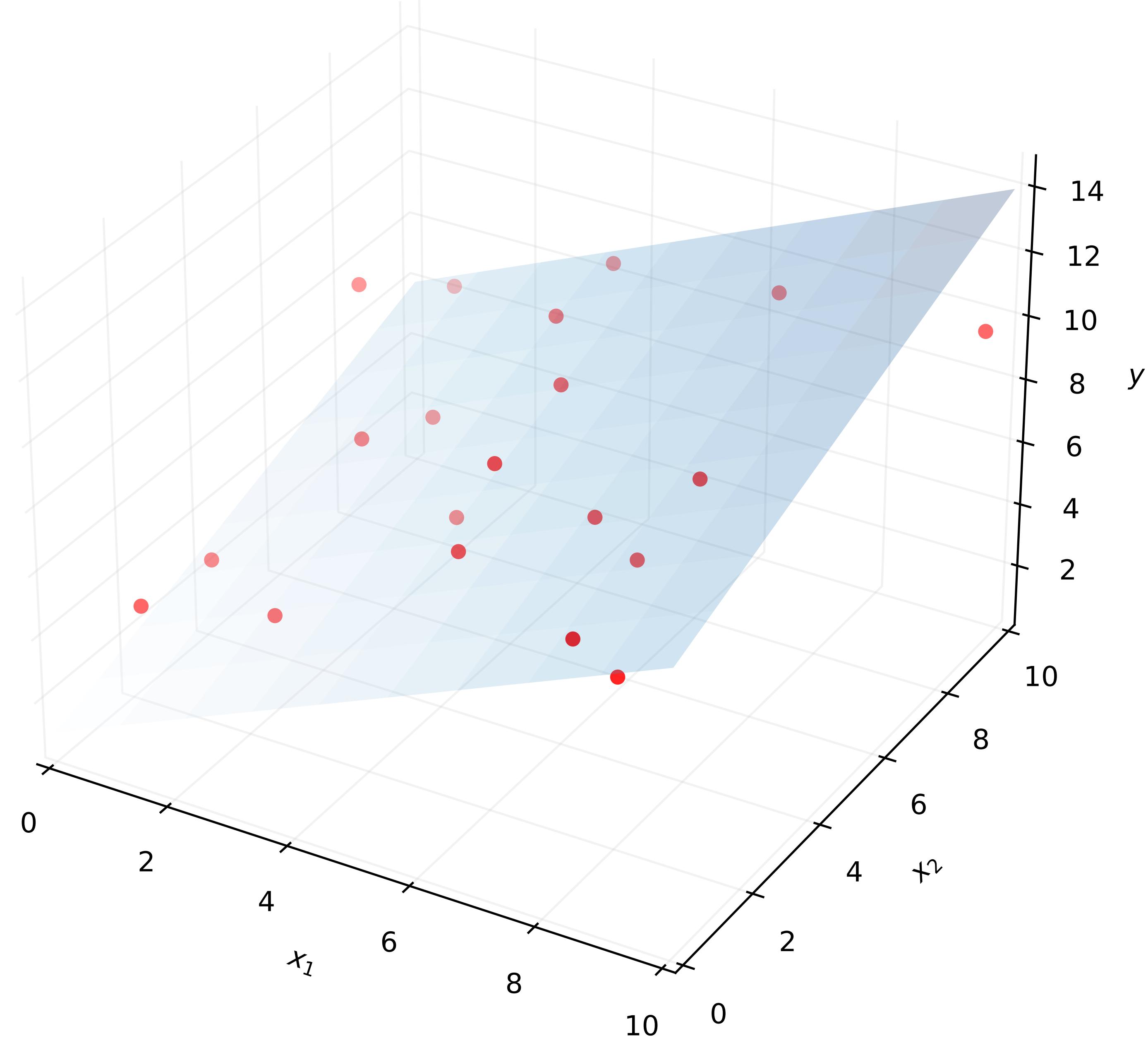
#	Age	Weight	Sex	Captive	...
1	11.8	815	1	0	
2	9.7	672	1	0	
3	8.9	446	1	0	
4	10.8	761	0	0	
5	8.3	1035	0	1	
6	11.7	930	1	1	
7	8.5	1027	0	1	
8	7.6	1234	0	1	
9	15.5	1461	0	1	
10	8.8	720	1	0	
11	12.8	1223	0	1	
12	7.7	711	1	0	
13	7.9	586			

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$\hat{y} \equiv xw$





Linear models are the best models!

$\hat{y} \equiv xw$

**Comprehensible
Interpretable
Tractable
Versatile**

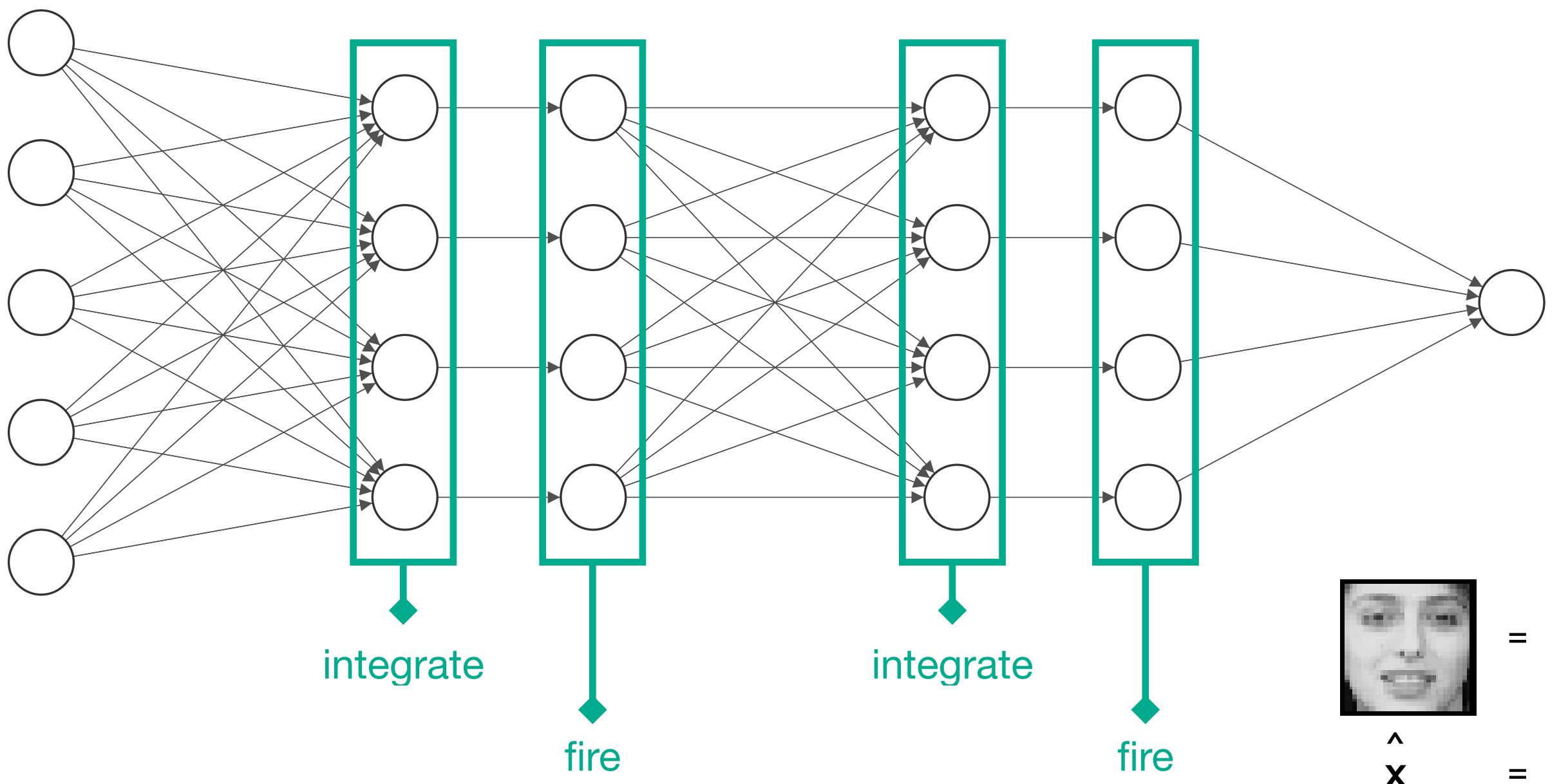
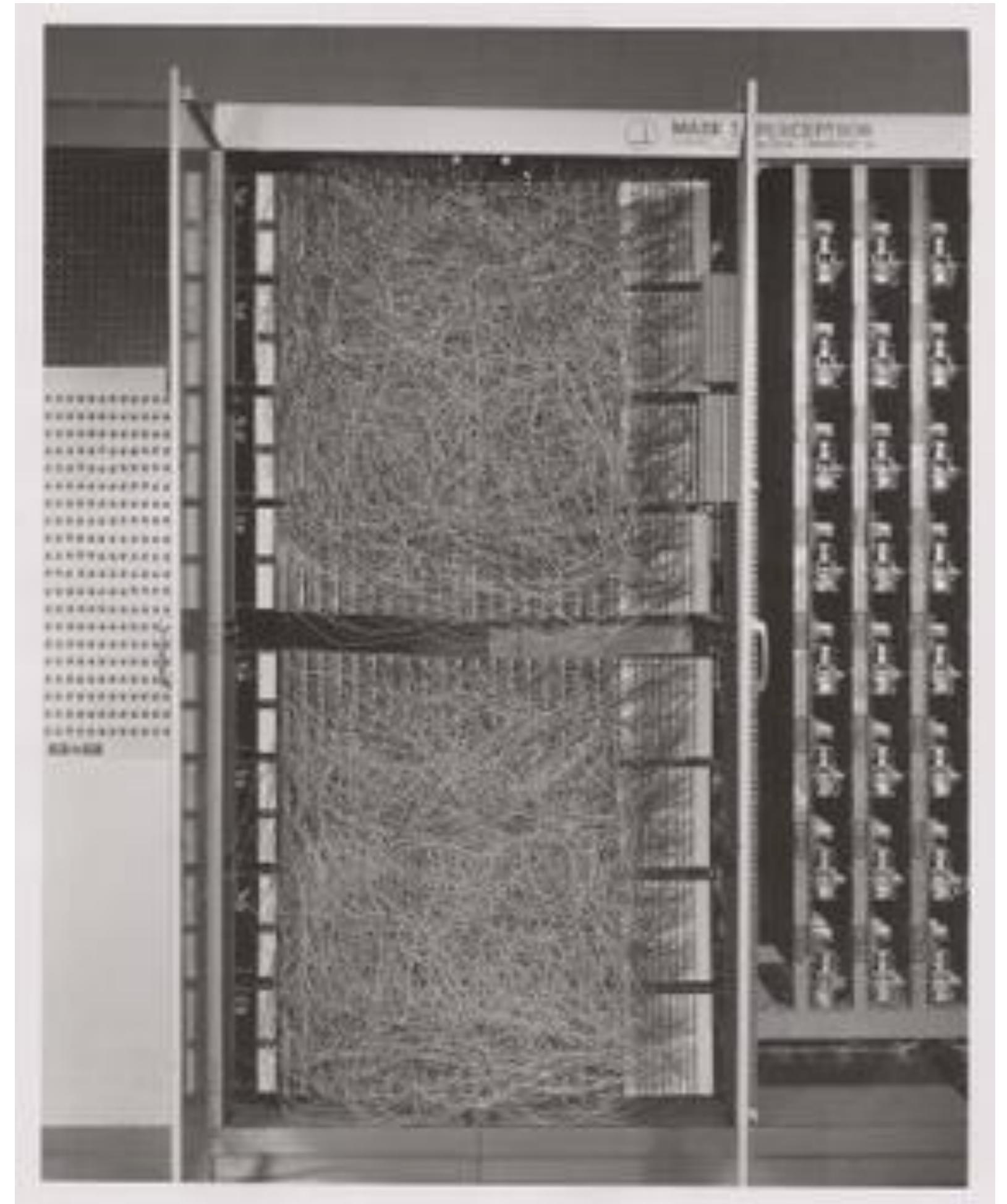
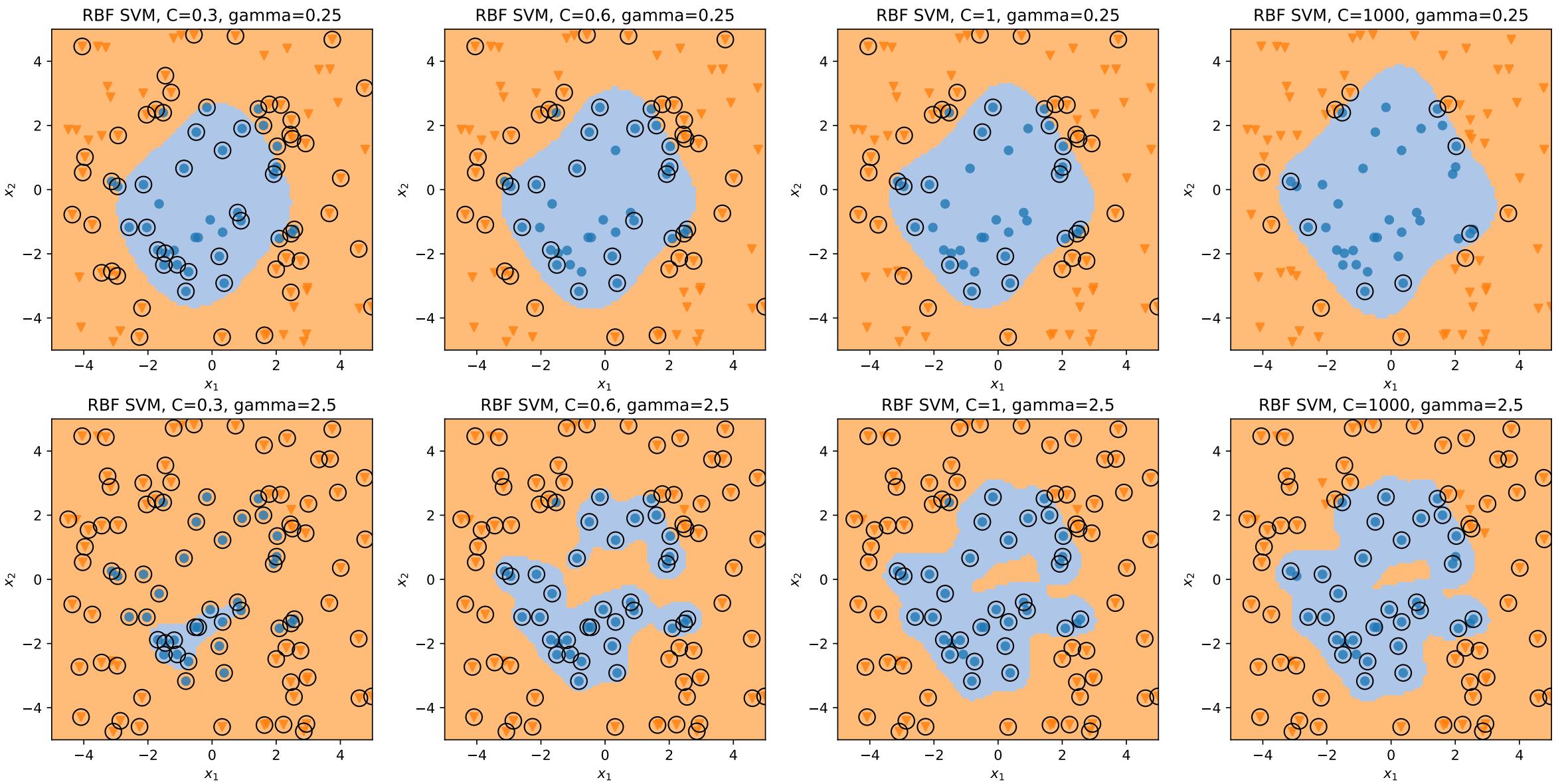
$$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} ? \\ ? \\ ? \\ ? \end{bmatrix}$$



$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} L(\mathbf{w}, \mathbf{X}, \mathbf{y})$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{Xw} - \mathbf{y}\|^2$$

$$\mathbf{X}^T \mathbf{X} \mathbf{w}^* = \mathbf{X}^T \mathbf{y}$$

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

“Normal Equations”

$$(X^T X + \lambda I) w^* = X^T y$$

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

“Revised Normal Equations”

#1

$$y \equiv f(x, \theta)$$

$$\mathbf{y} \equiv \mathbf{f}_\theta(\mathbf{x})$$

$$\theta^* = \operatorname*{argmin}_{\theta} L(f, \theta, \{X[, Y]\})$$

$$\theta^* = \operatorname*{argmin}_{\theta} L_{f,X,Y}(\theta)$$

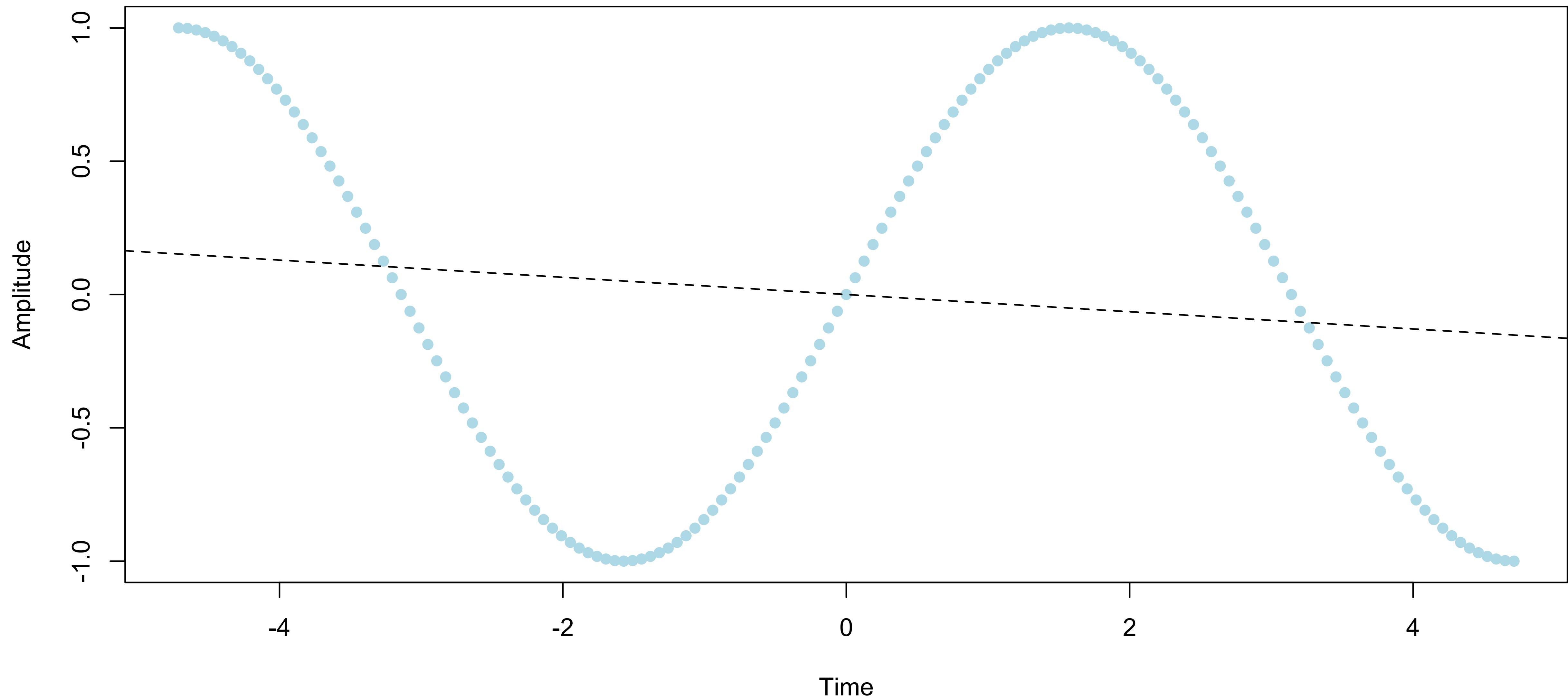
$$P(y|x; \theta) = f(x, y, \theta)$$

$$L(\theta) = f(x, y, \theta)$$

$$\theta^* = \operatorname*{argmax}_{\theta} L_{f,X,Y}(\theta)$$

training \neq test

**Given the training set, loss (or likelihood)
is solely a function of the parameters**



Don't fit a straight line to curvy data

#2

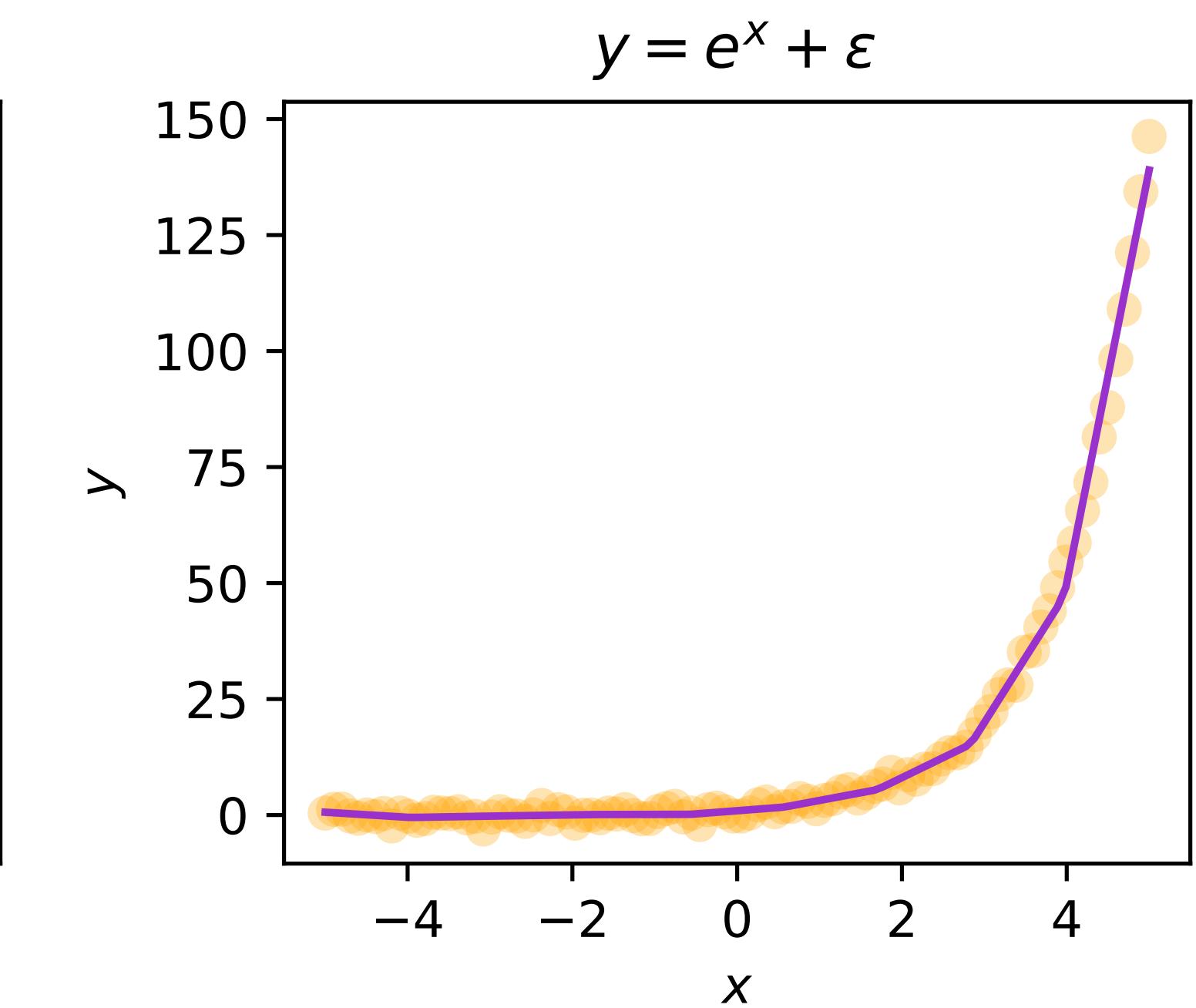
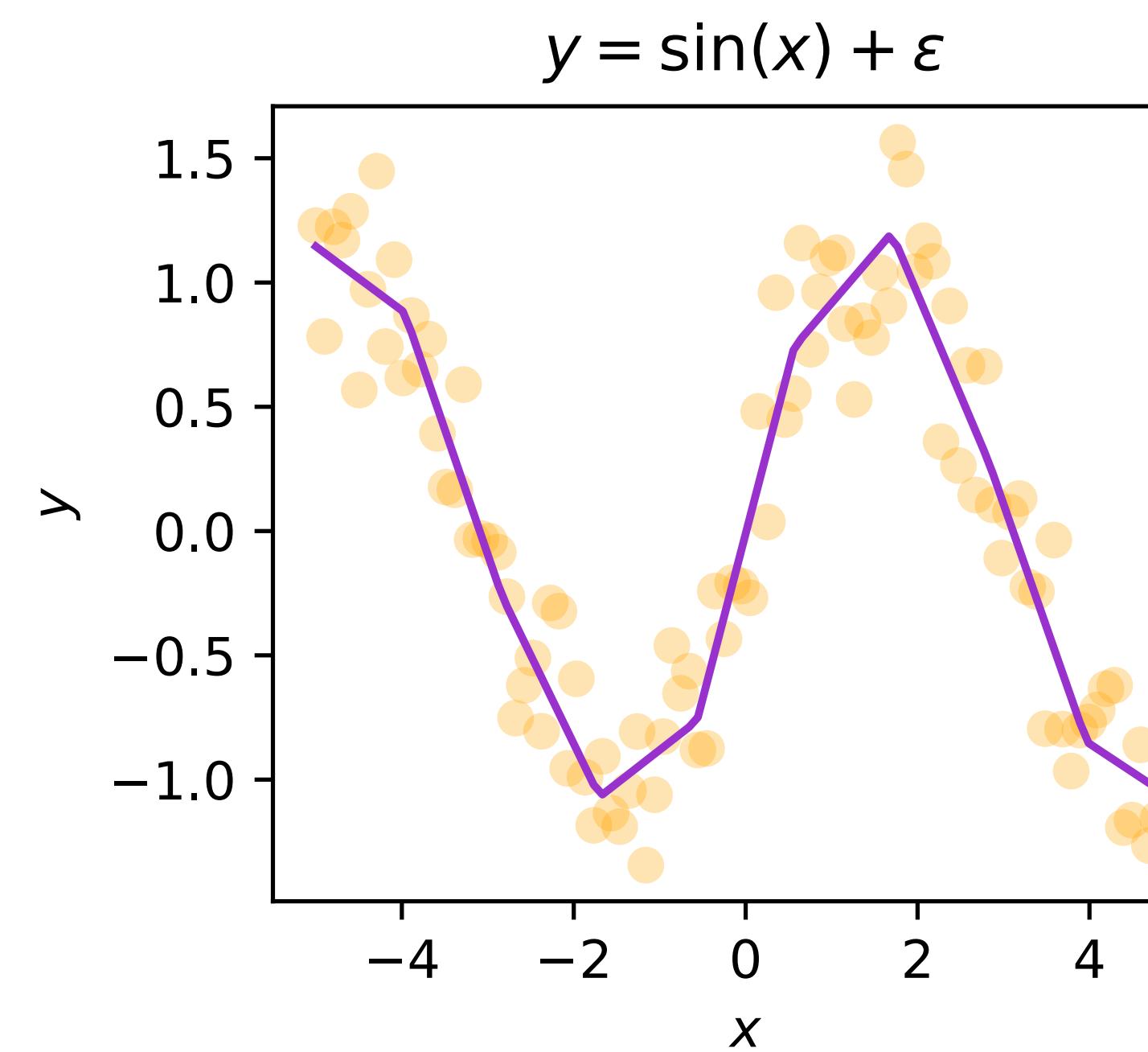
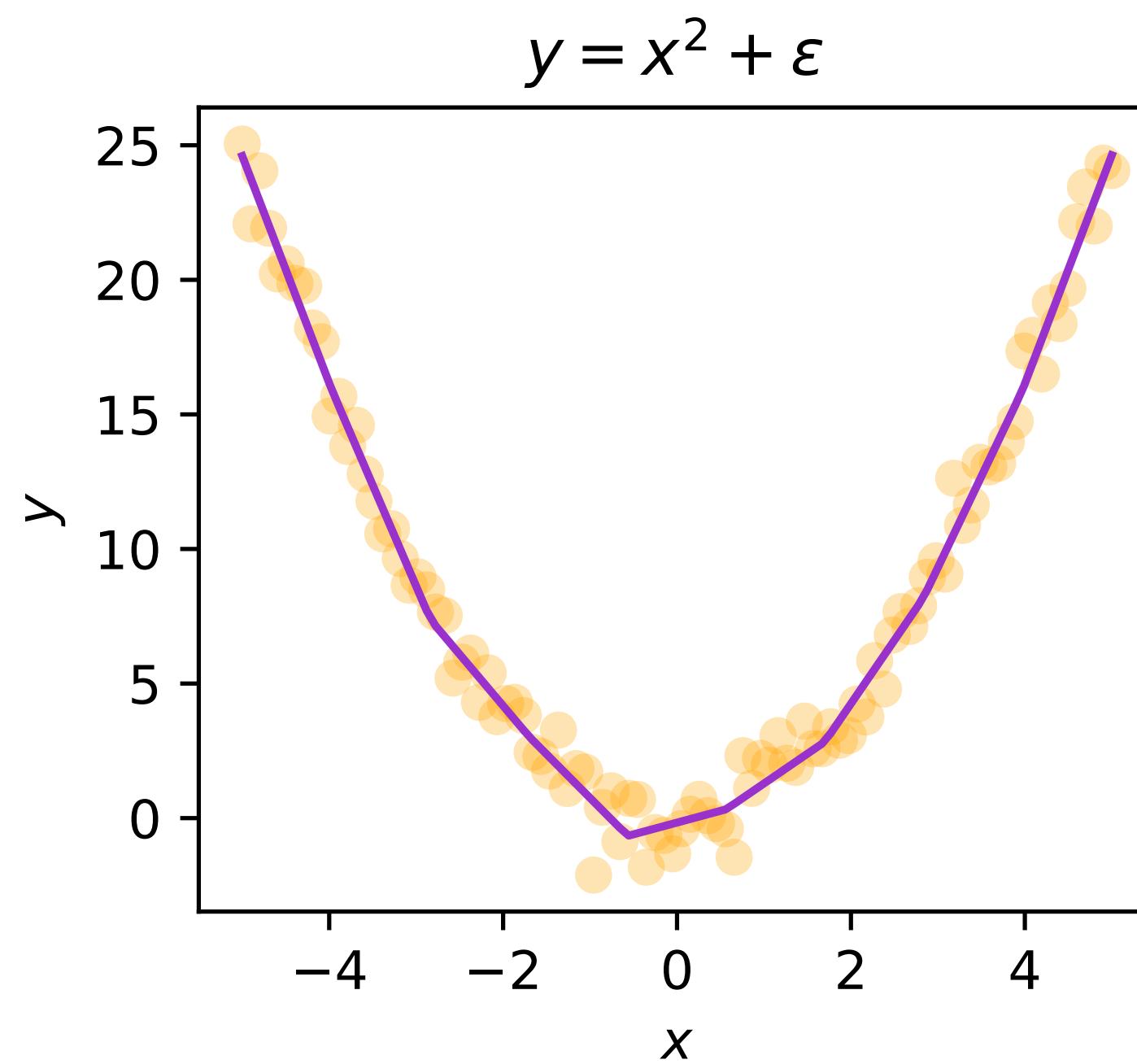
$$\hat{y} = \sum_i^k w_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x}) \cdot \mathbf{w} = \mathbf{x}' \cdot \mathbf{w}$$

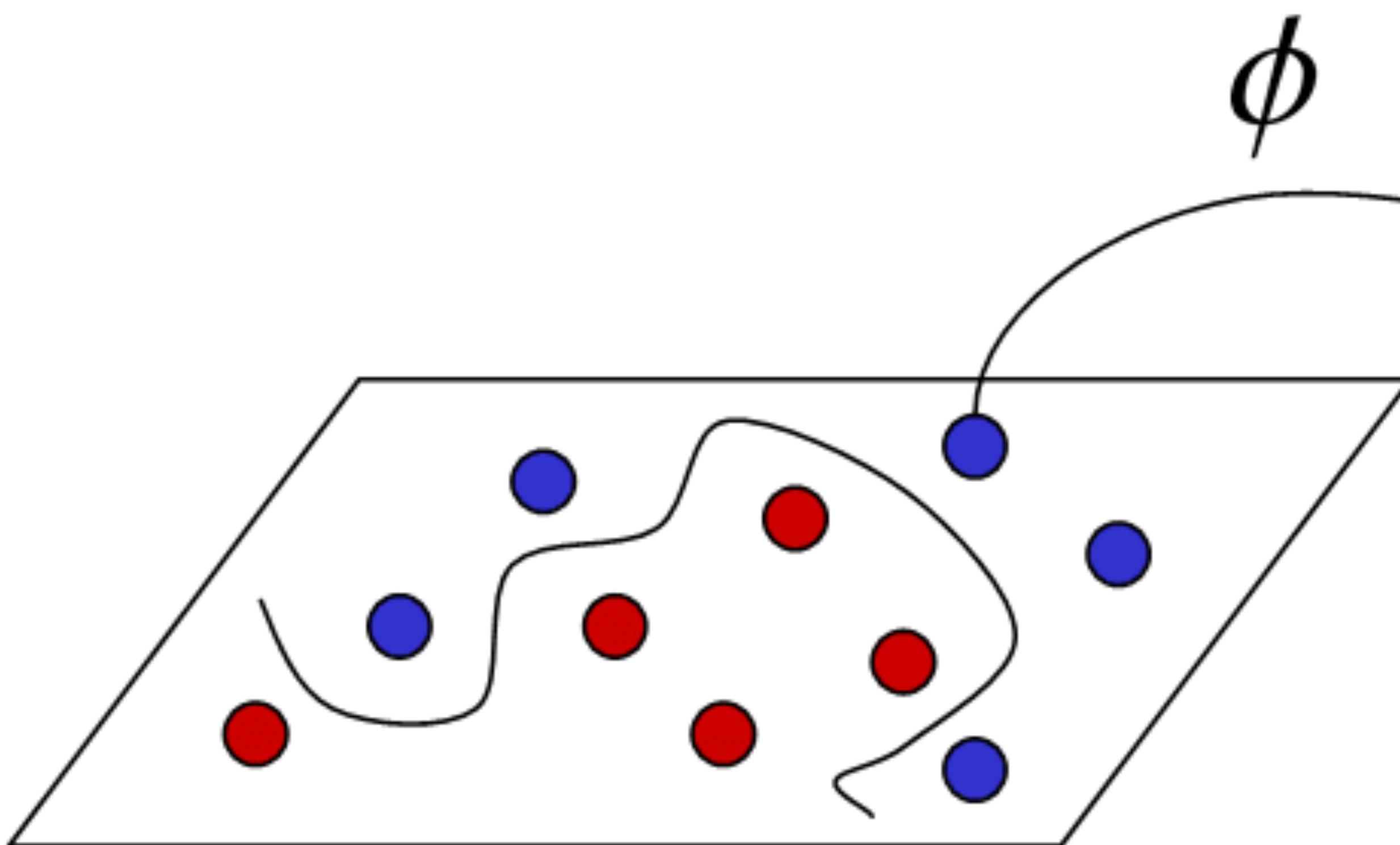
$$\mathbf{x}' = \mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_k(\mathbf{x})]$$

$$h_1(x) = 1$$

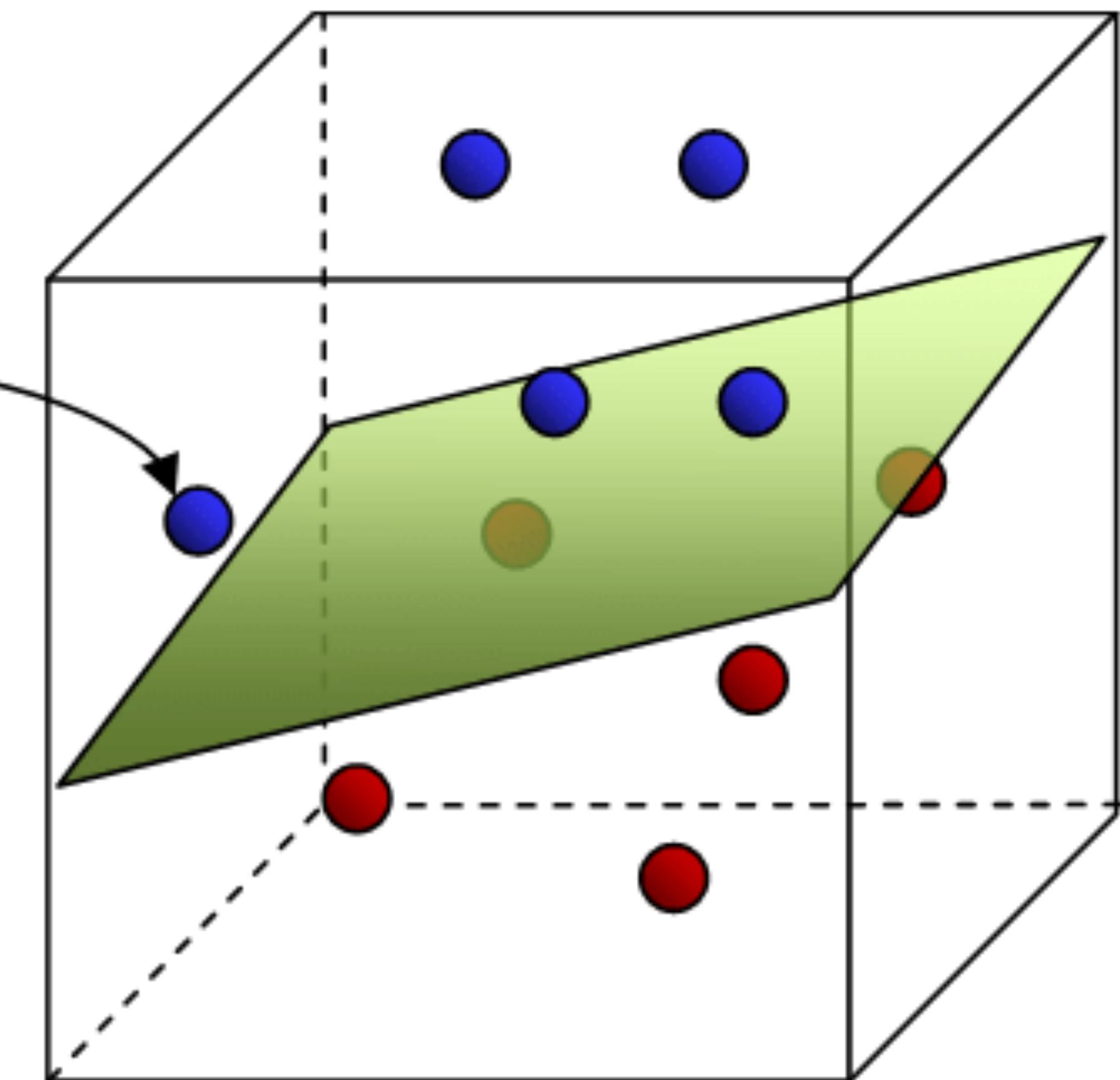
$$h_2(x) = x$$

$$h_{i+2}(x) = \max(0, x - t_i), \quad i \in \{1, 2, \dots, m\}$$





a) Input Space



b) Feature Space

#3

DEFEND · THE · CHILDREN · OF · THE ·
POOR · & · PUNISH · THE · WRONGDOER ·

$$L(\mathbf{X}, \mathbf{y}; \theta) = D(f_{\theta}(\mathbf{X}, \mathbf{y}), \mathbf{y}) + \lambda P(\theta)$$

$$L(\mathbf{X}, \mathbf{y}; \mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda P(\mathbf{w})$$

$$P(\mathbf{w}) = \|\mathbf{w}\|^2$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{Xw} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$$

“Ridge Regression”

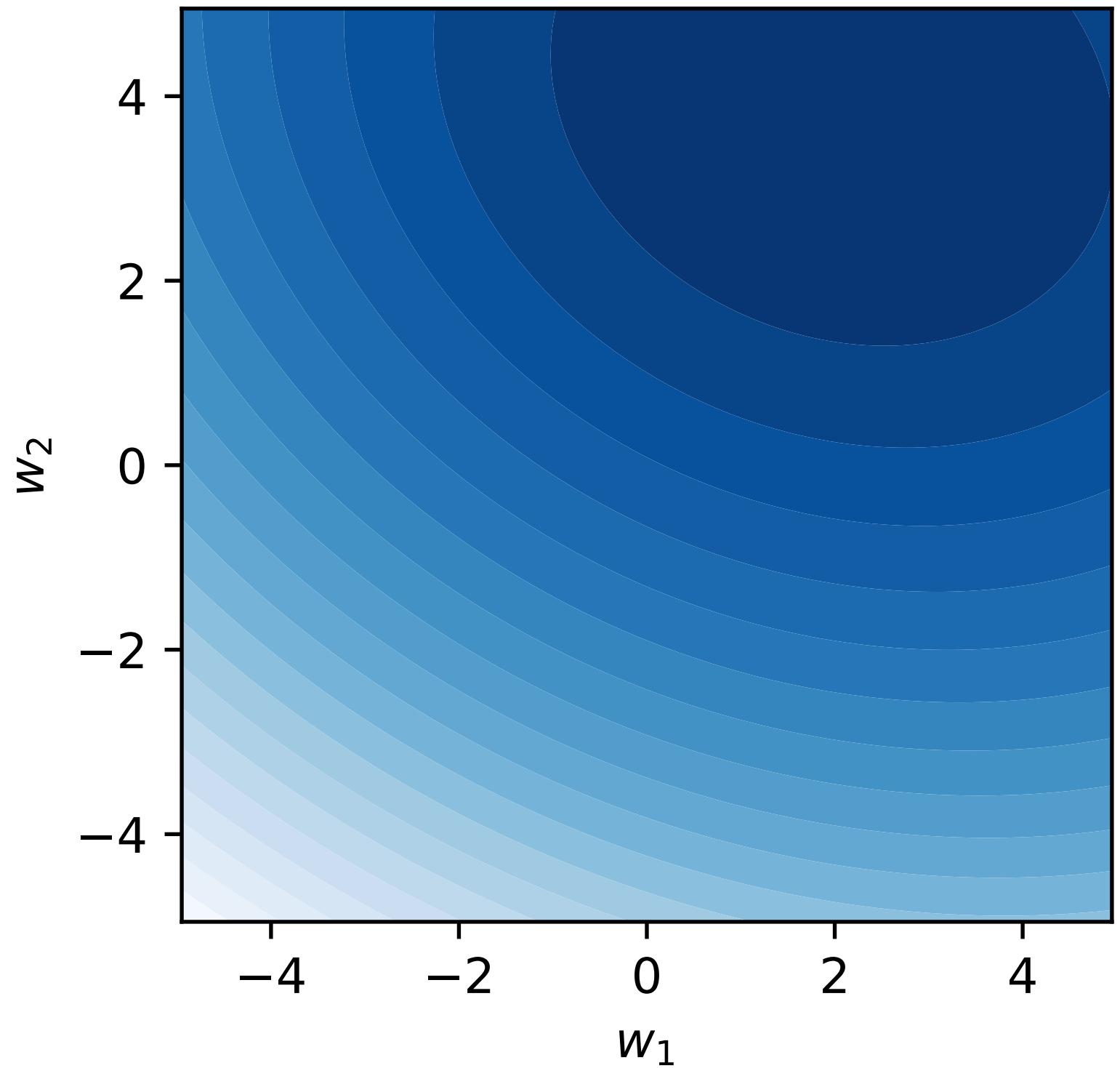
$$P(\mathbf{w}) = \|\mathbf{w}\|_1$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{Xw} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_1$$

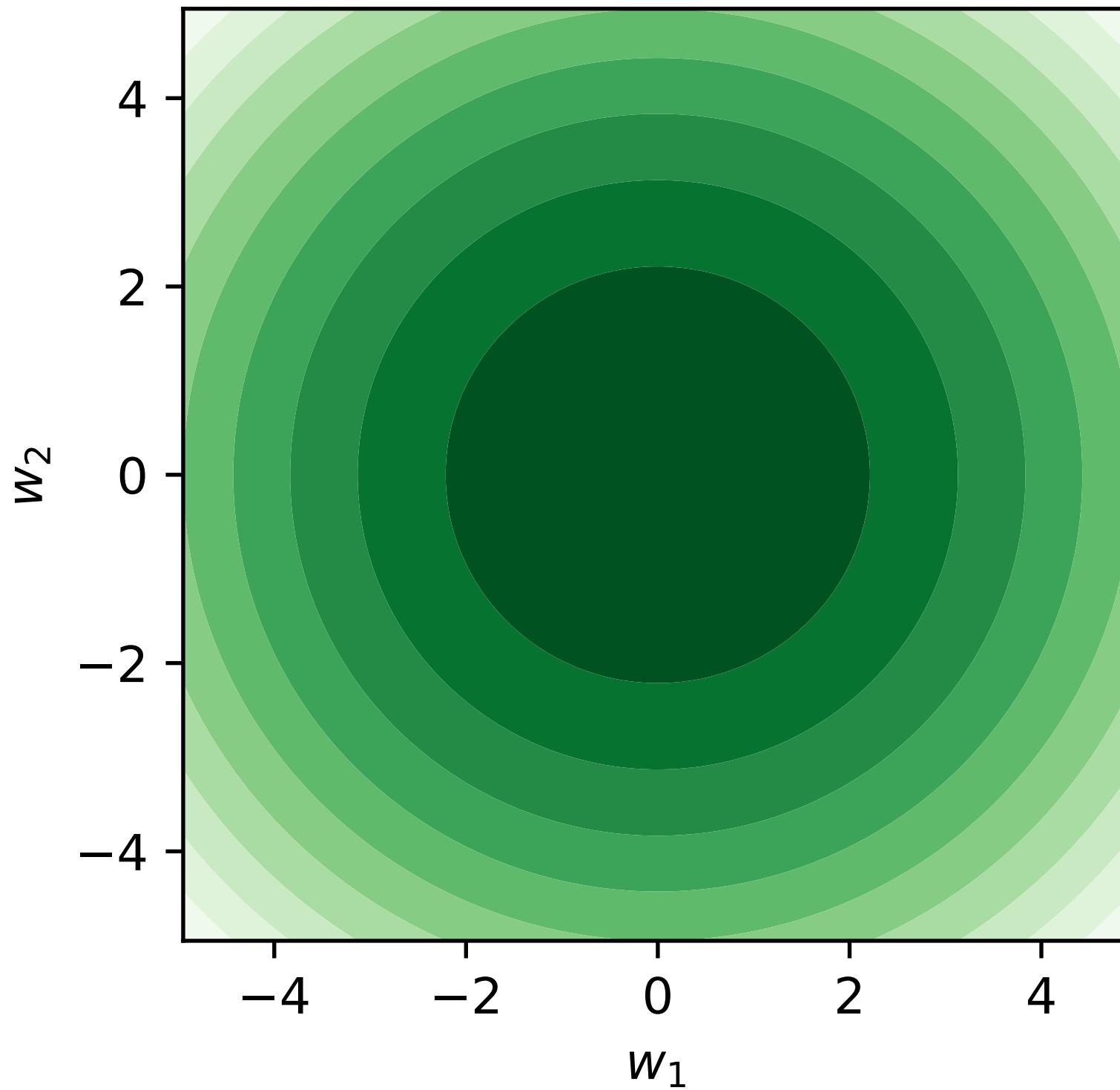
“Lasso”



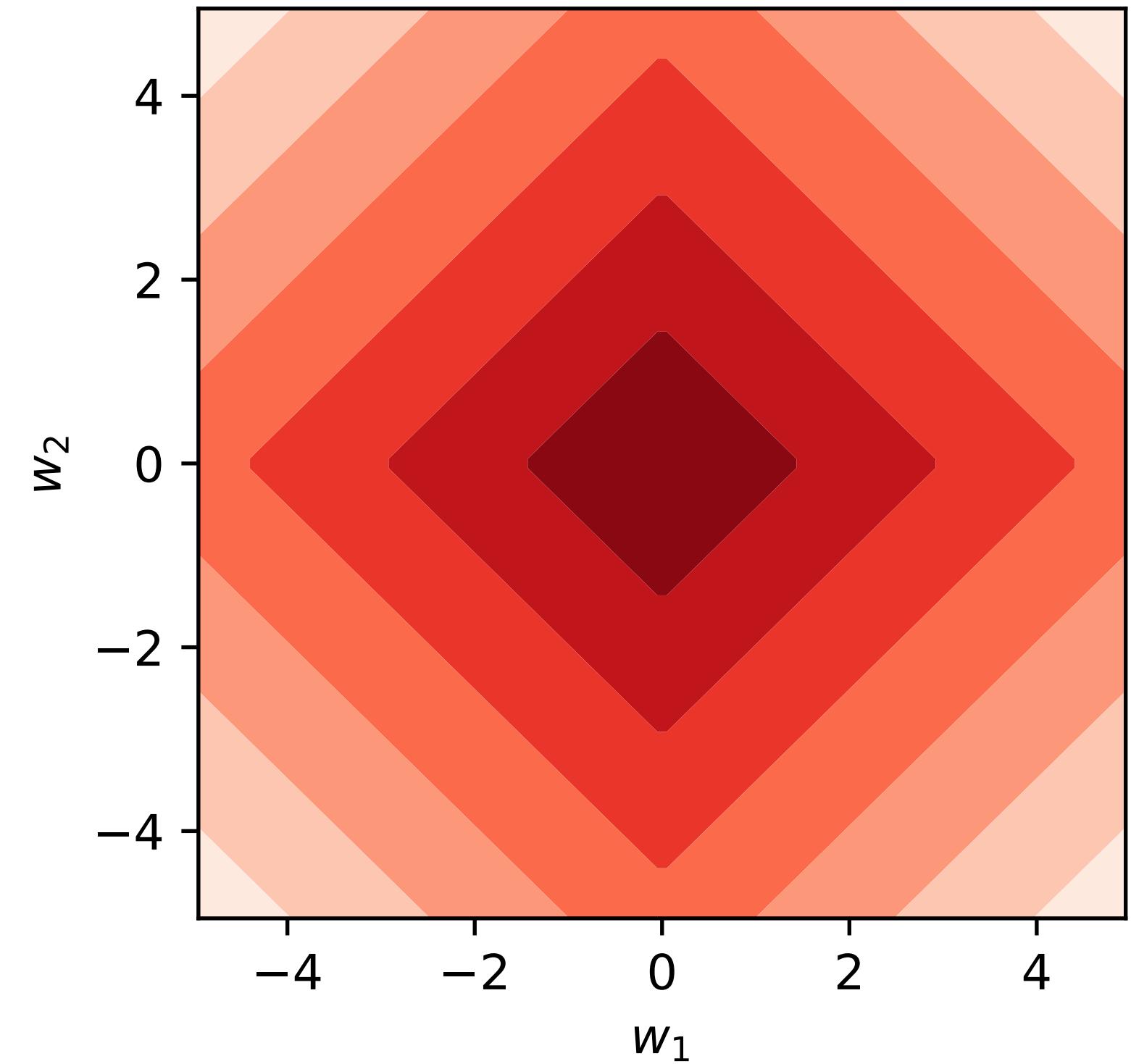
OLS Loss: $\|\hat{\mathbf{y}} - \mathbf{y}\|^2$

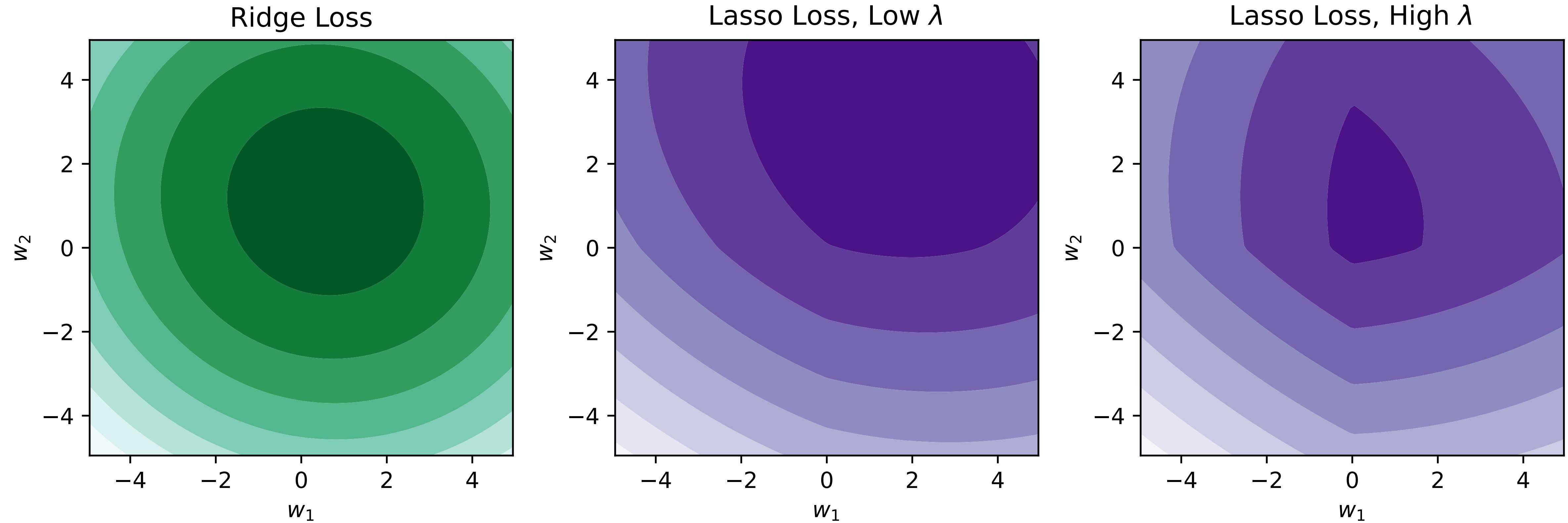


Ridge Penalty: $\|\mathbf{w}\|^2$



Lasso Penalty: $\|\mathbf{w}\|_1$





$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_1$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$$

#4



$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} L(\mathbf{f}, \mathbf{X}, \mathbf{y}, \mathbf{w})$$

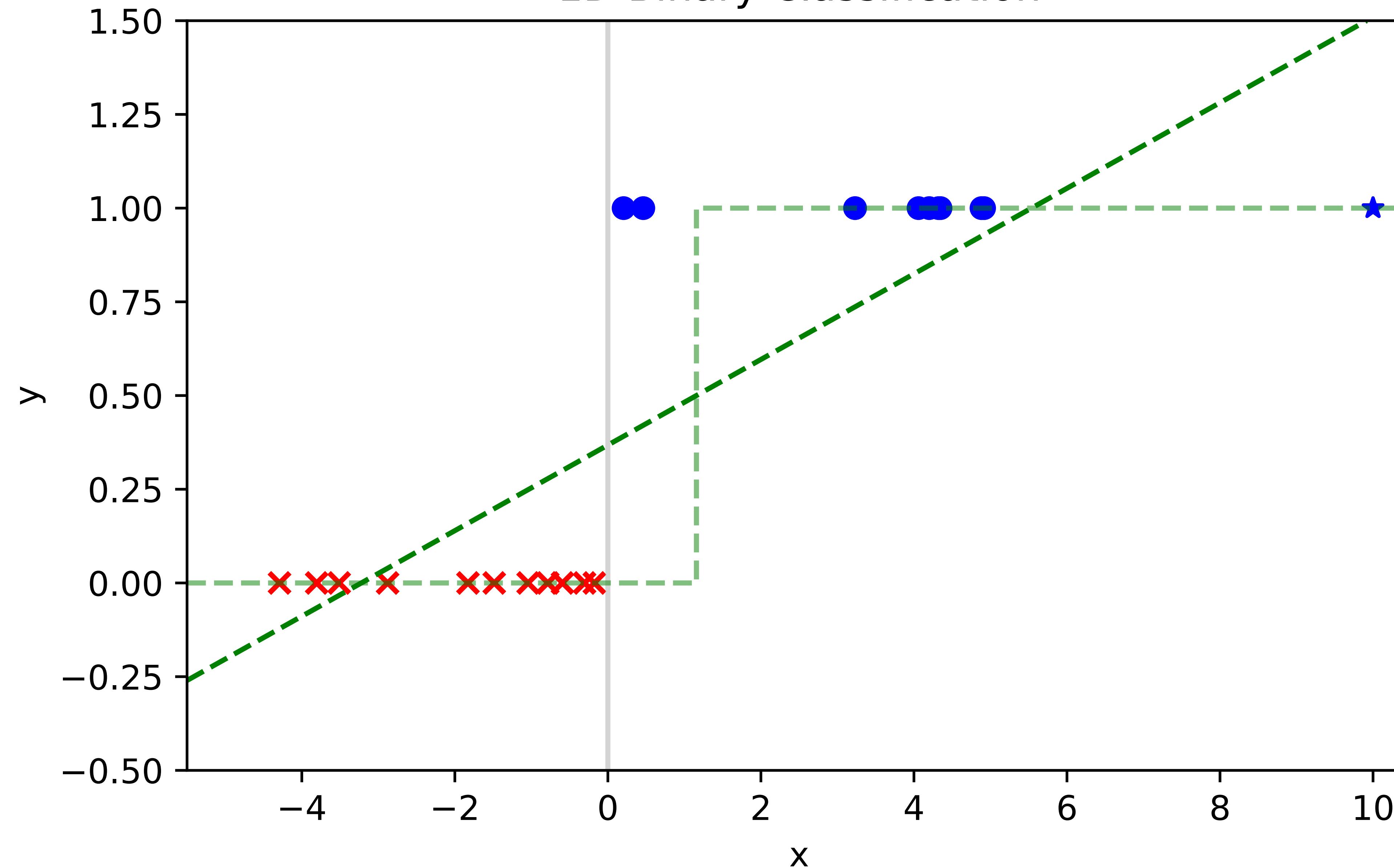
“Gradient Descent”

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} L(\mathbf{f}, \mathbf{x}, y, \mathbf{w})$$

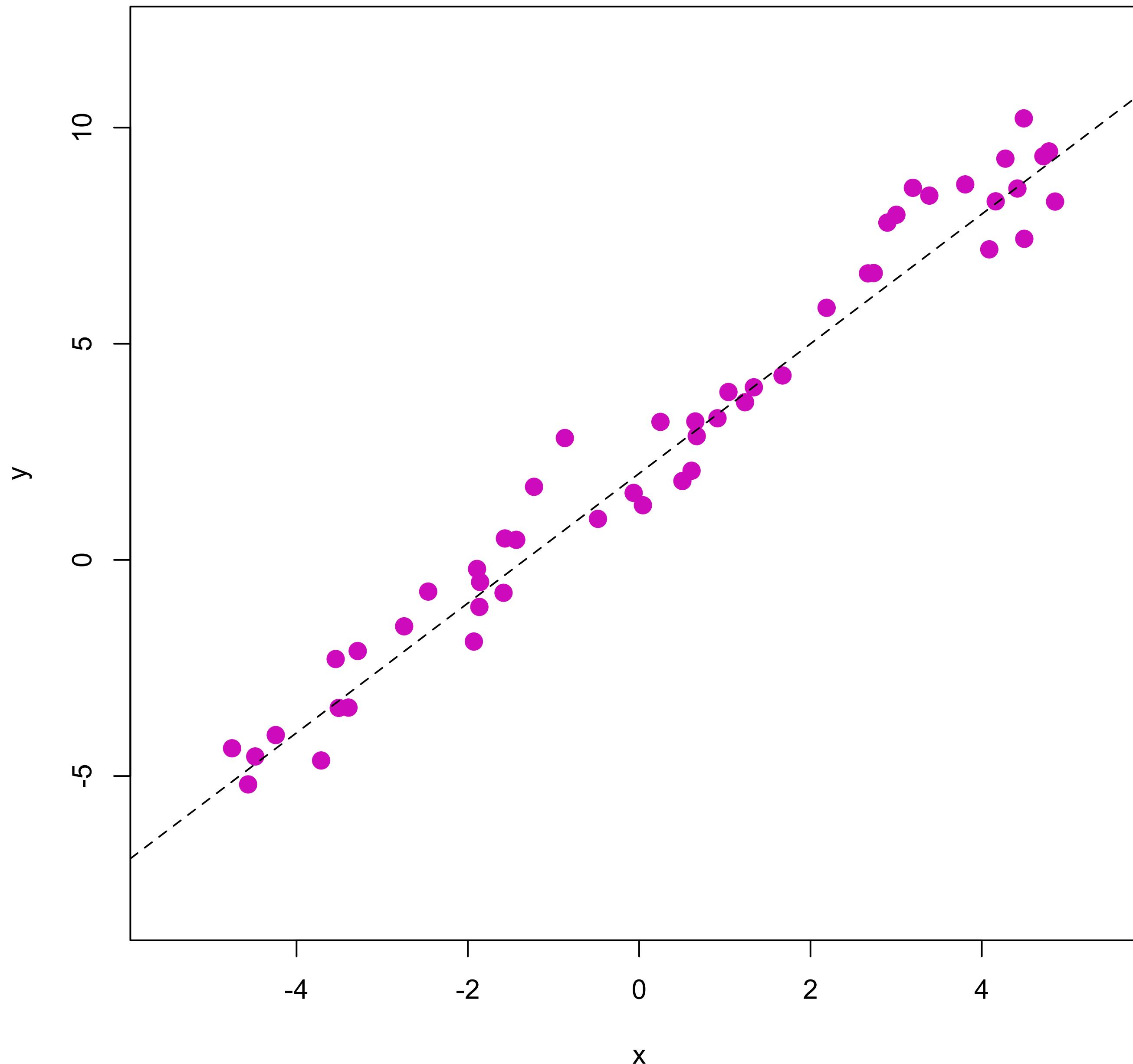
“Stochastic Gradient Descent”

#5

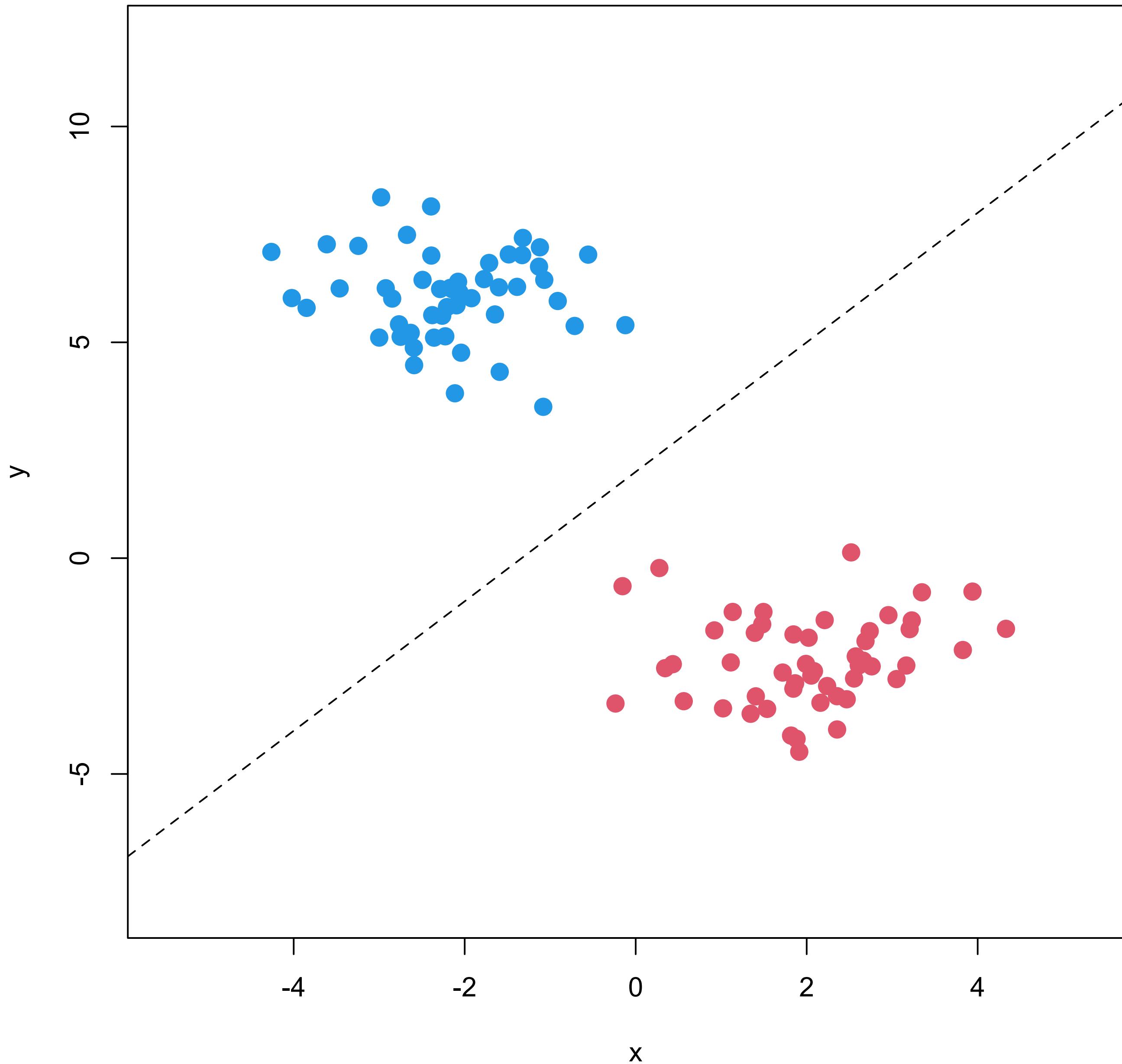
1D Binary Classification



Regression

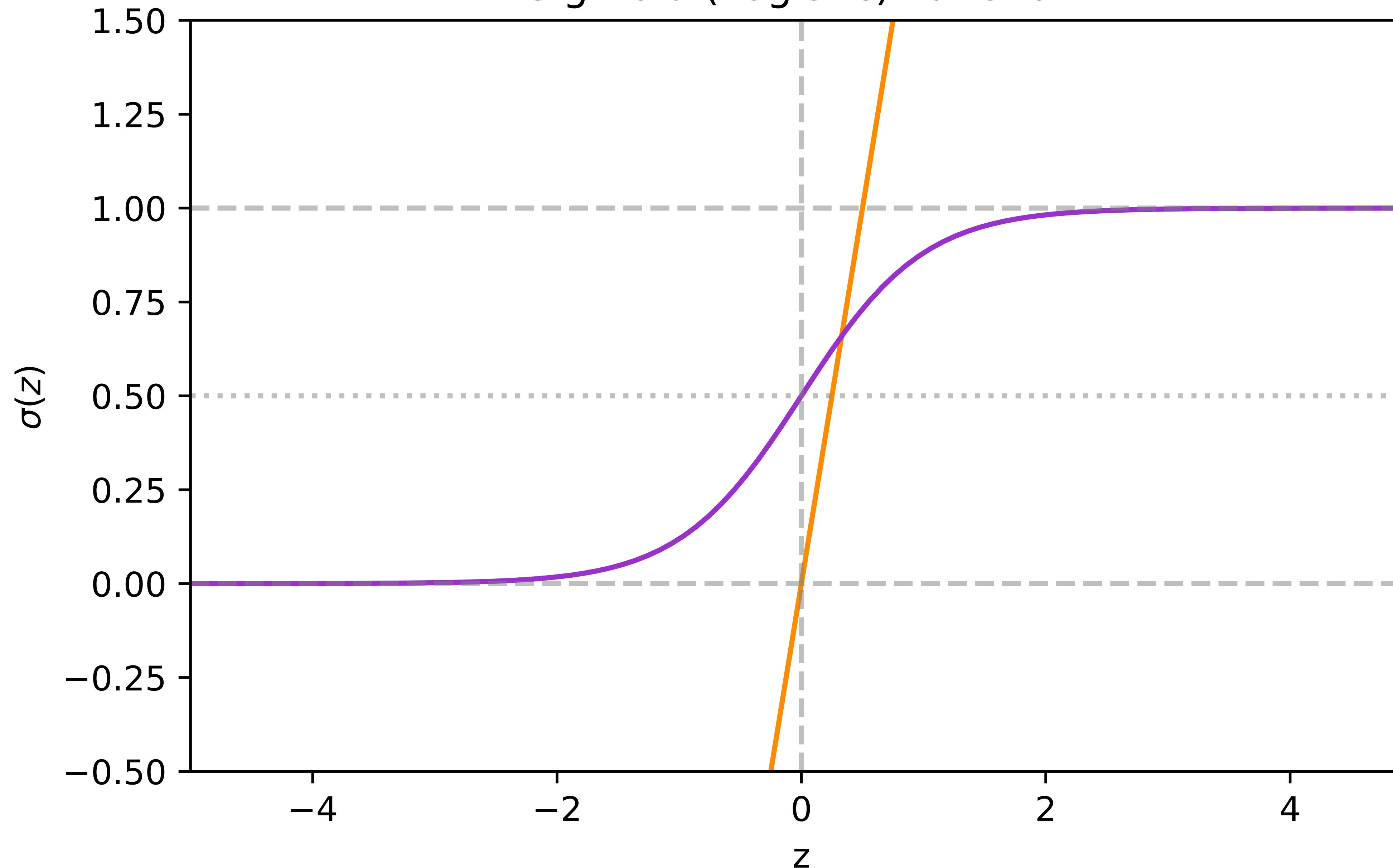


Classification



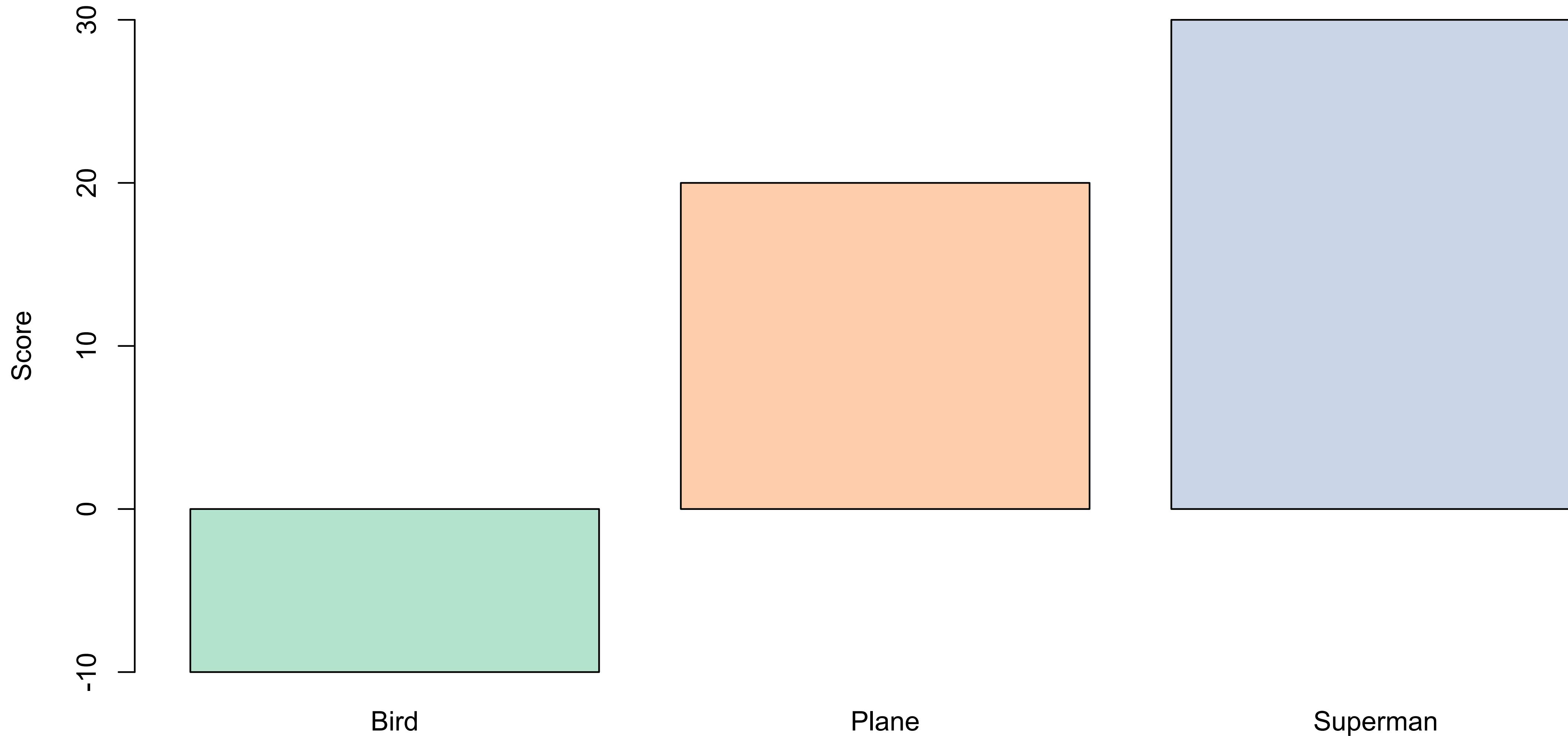
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid (Logistic) Function



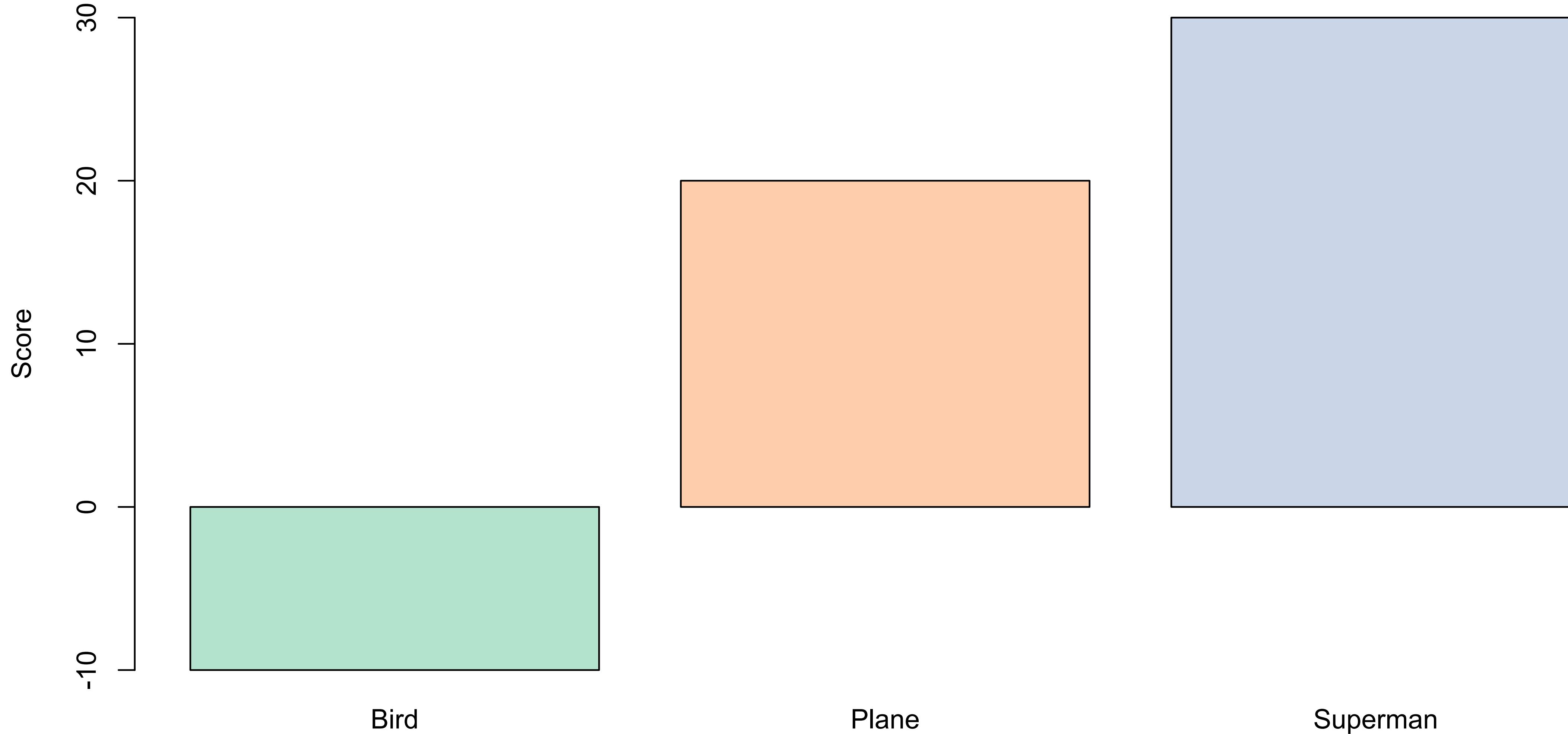
$$\begin{aligned}
\nabla_{\mathbf{w}} l &= \mathbf{x}^T (\mathbf{y} - \hat{\mathbf{y}}) \\
&= \mathbf{x}^T (\mathbf{y} - \sigma(\mathbf{Xw})) \\
&= \mathbf{x}^T \left(\mathbf{y} - \frac{1}{1 + e^{-\mathbf{Xw}}} \right)
\end{aligned}$$

Class Scores (Logits)

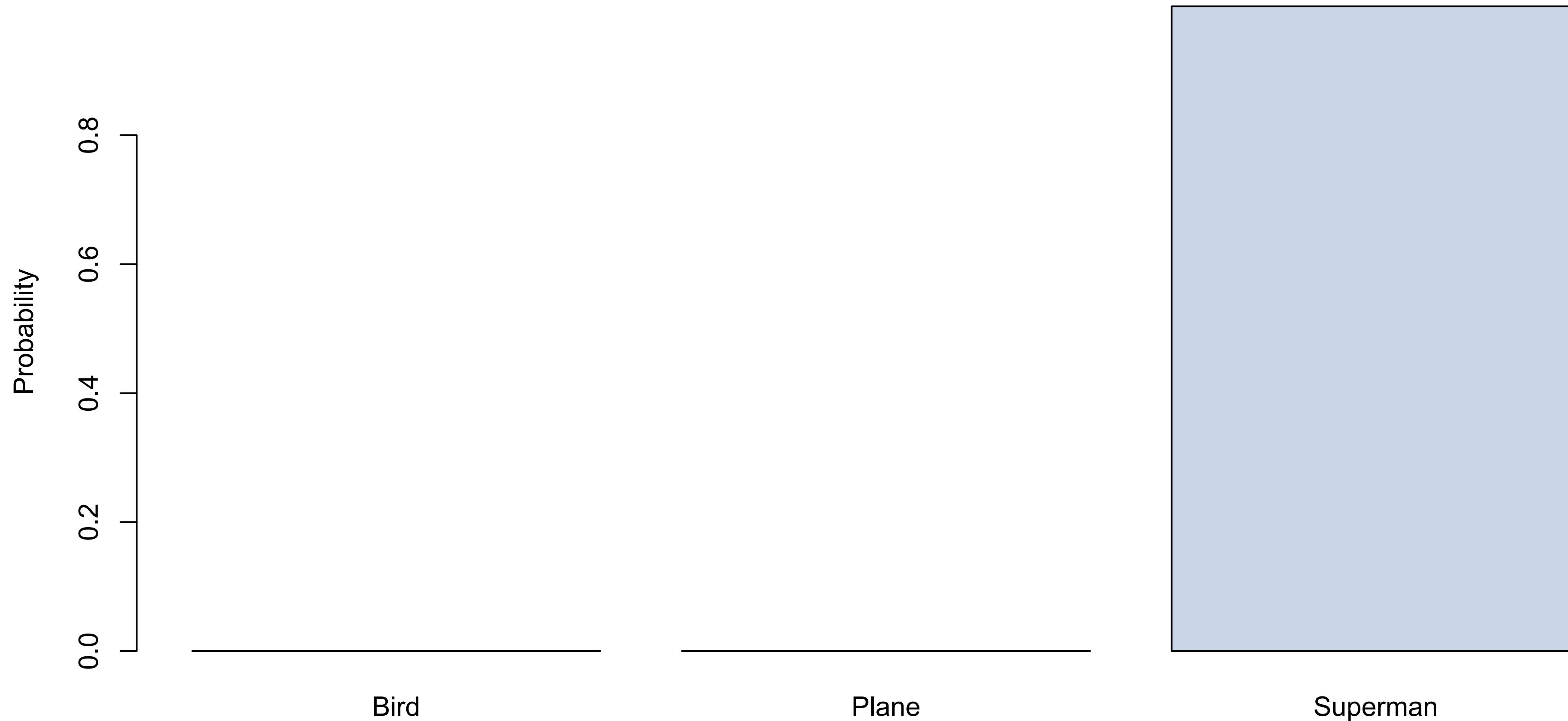


$$\varsigma(z) = \frac{e^z}{\sum e^z}$$

Class Scores (Logits)



Class Probabilities



$$\nabla_{\mathbf{W}} L = \mathbf{X}^T (\hat{\mathbf{Y}} - \mathbf{Y})$$

Questions?

Next week: Non-Parametric Models

