Week 2: Linear Models

Matthew Caldwell

COMP0088 Introduction to Machine Learning • UCL Computer Science • Autumn 2025

Admin

- Labs
- Triage

Week 2 Recap

Drawing the Line

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ L\left(\mathbf{f}, \boldsymbol{\theta}, \{\mathbf{X}[, \mathbf{Y}]\}\right)$$

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

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

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

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

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}$$

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

↓ Feature dimensions ↓

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

↑ Samples ↑

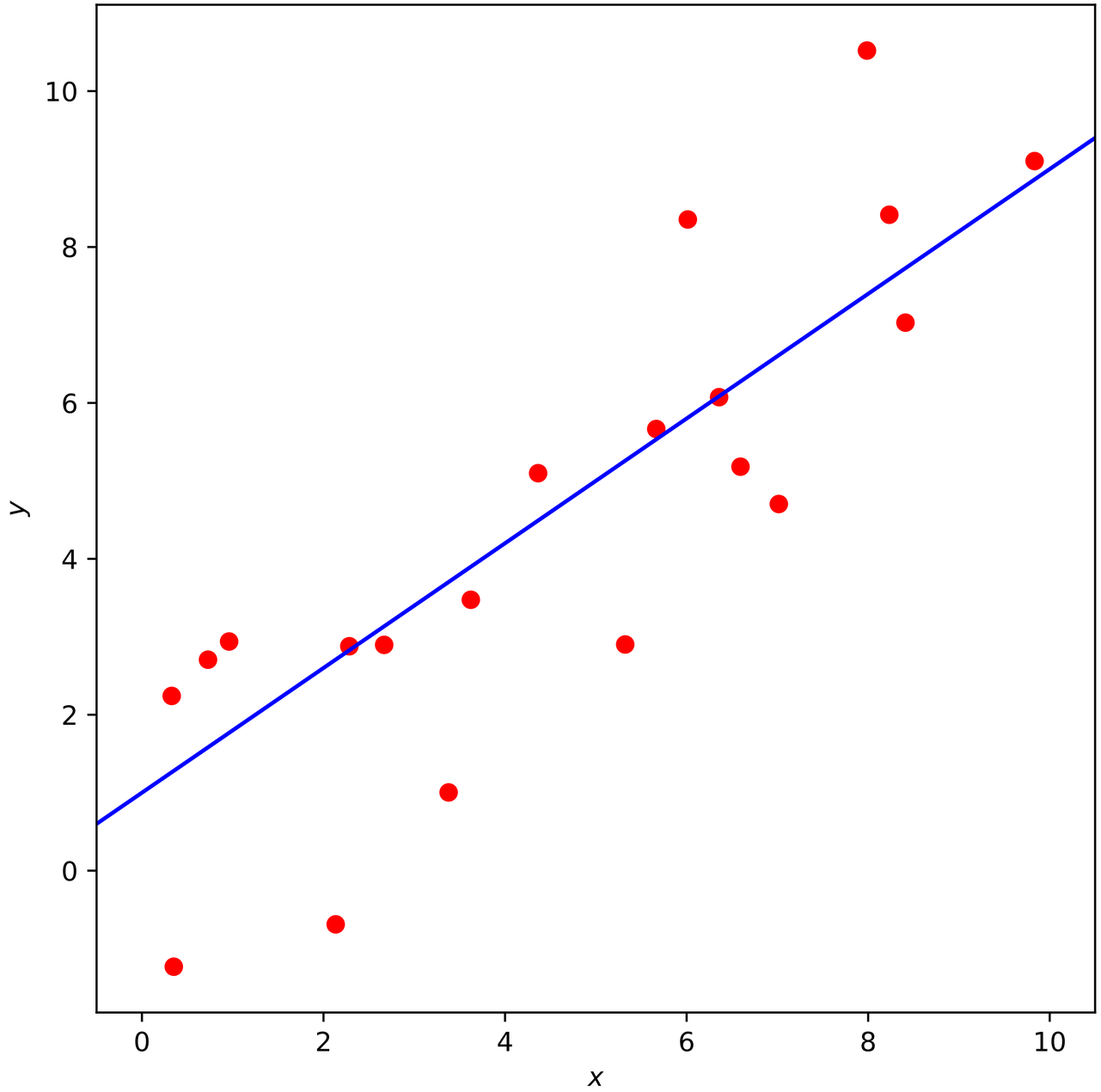
“Design Matrix”

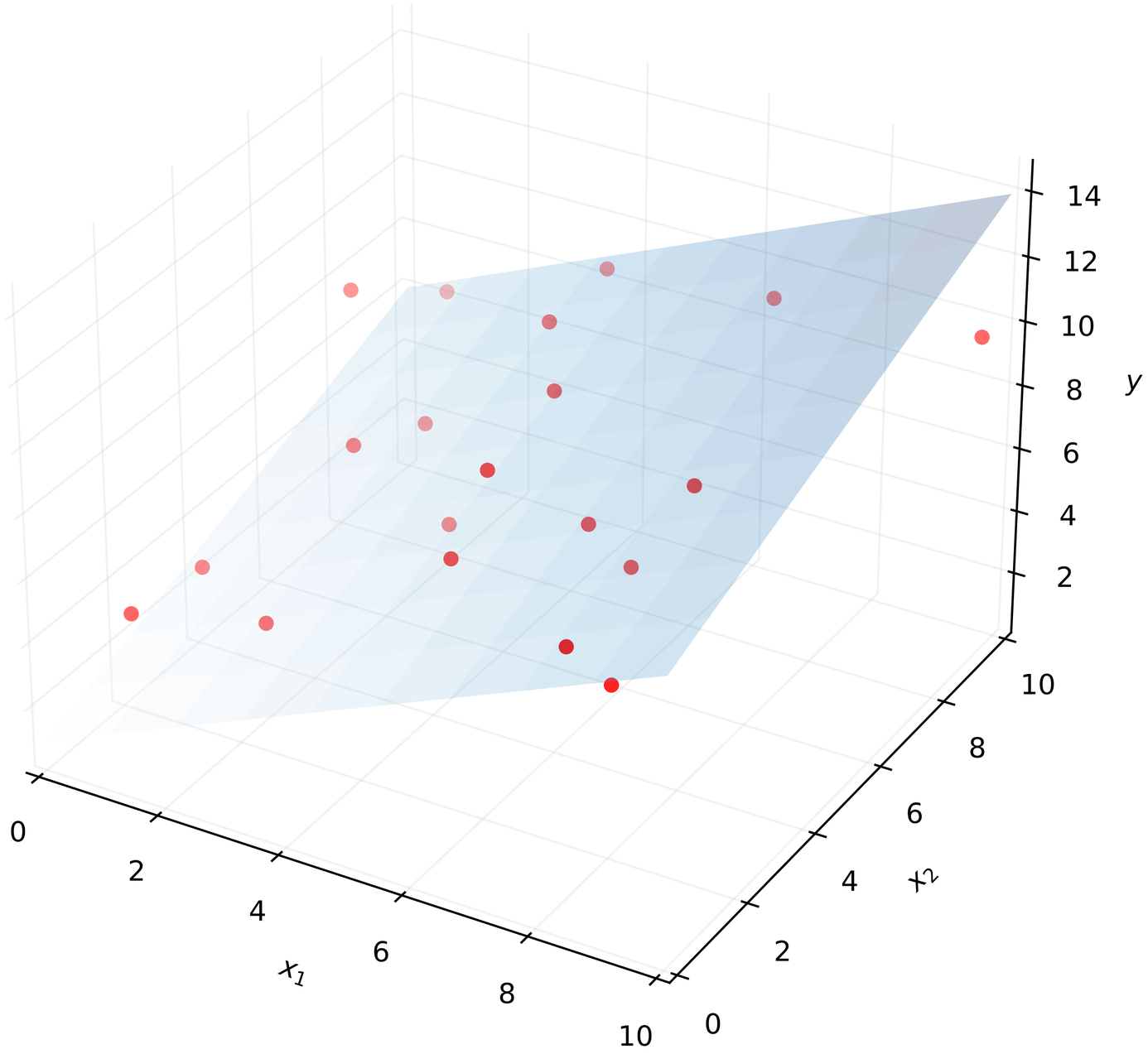
Meerkats

#	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	7.2	526			

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$


10
8
6
4
2
0
y
0
2
4
6
8
10
x


14
12
10
8
6
4
2
y
0
2
4
6
8
10
x_1
0
2
4
6
8
10
x_2

Linear models are the best models!

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

Comprehensible

Interpretable

Tractable

Versatile

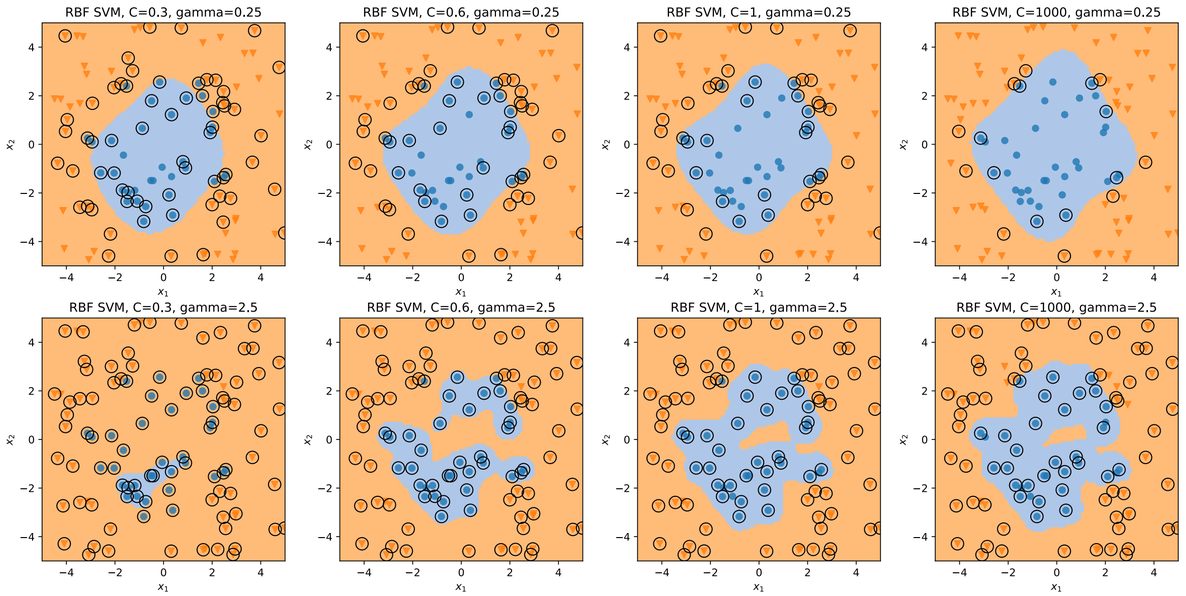
$$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

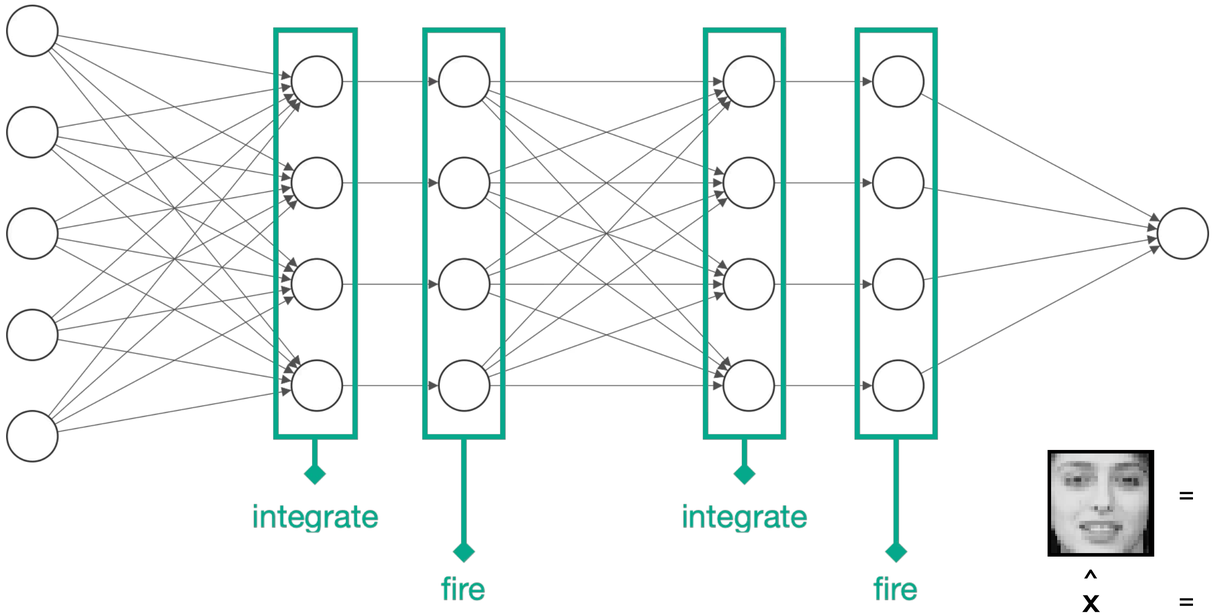
$$\mathbf{w} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

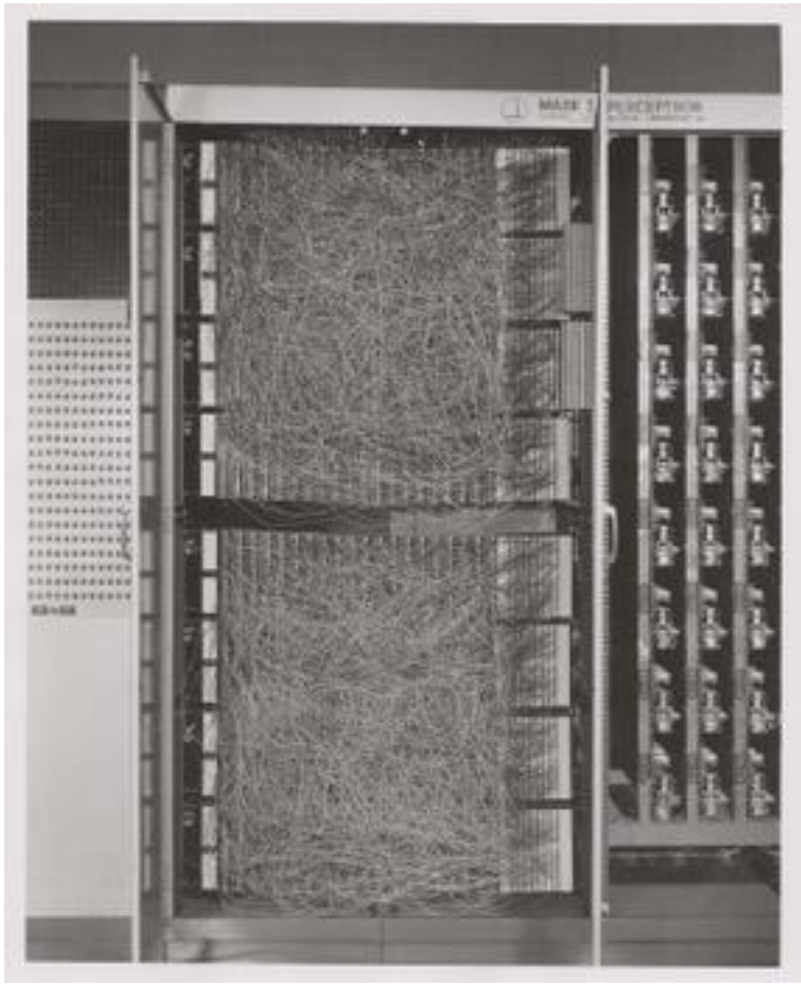
$$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

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

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



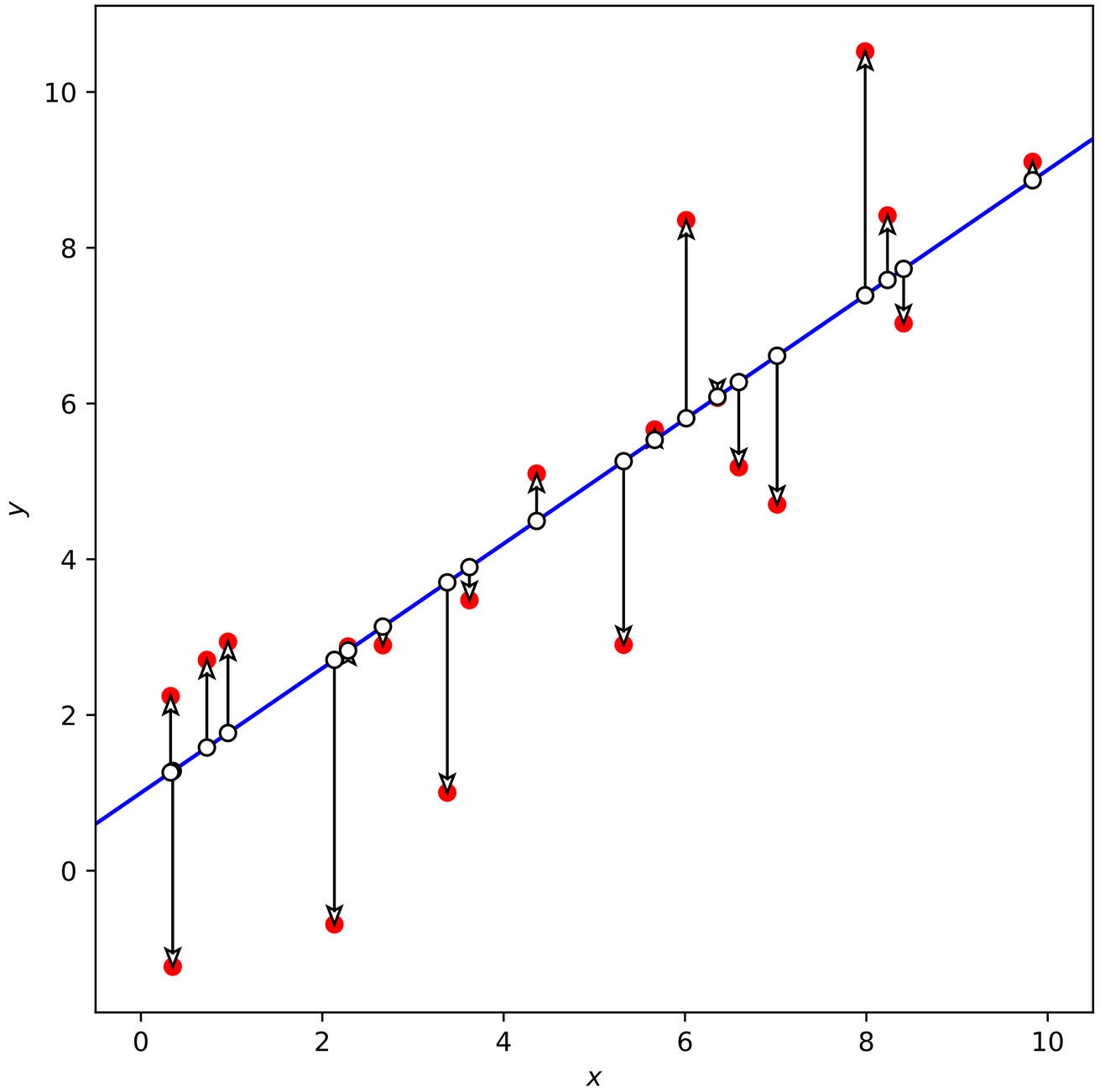




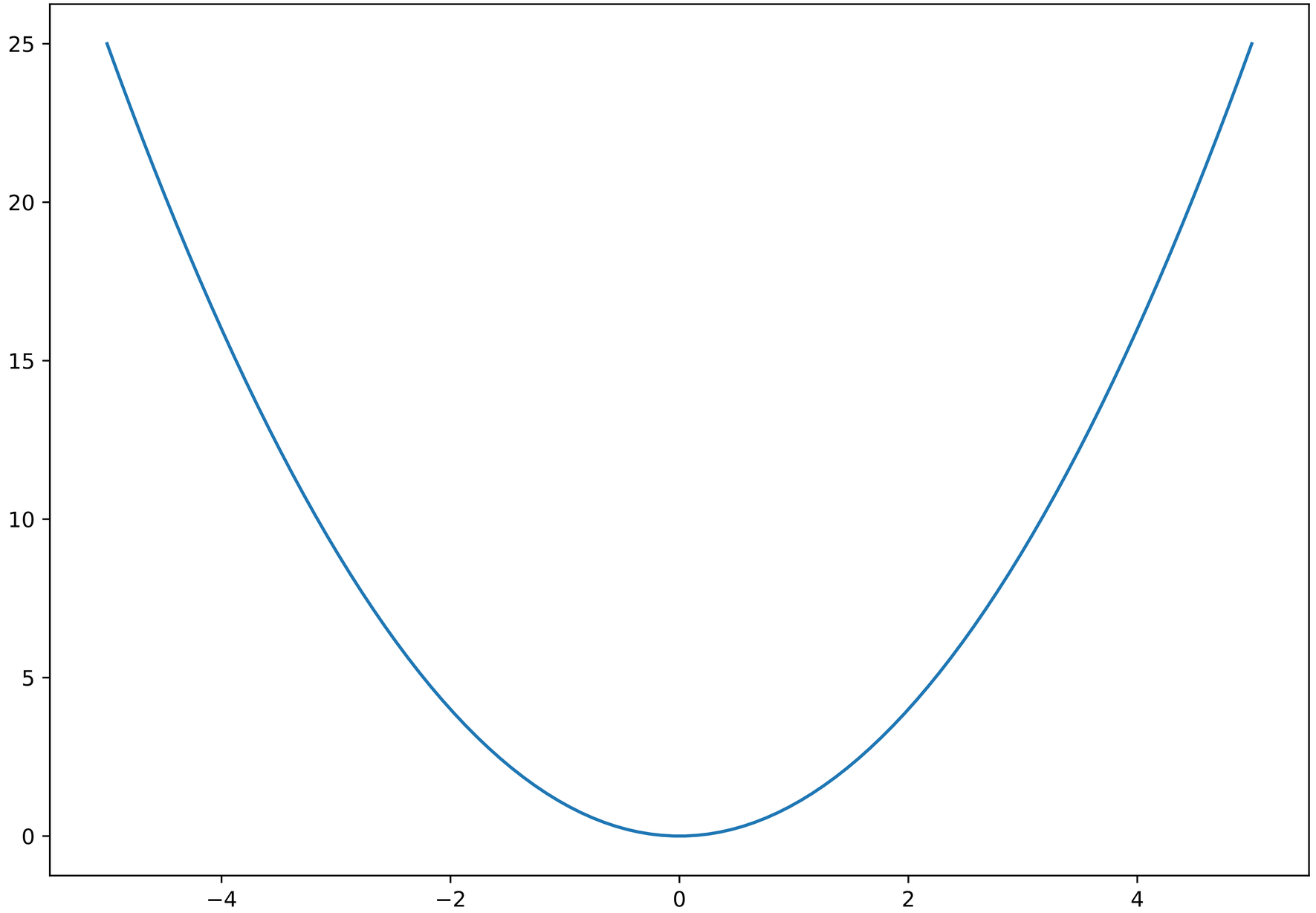
$\hat{x} = \mu + w_1u_1+w_2u_2+w_3u_3+w_4u_4+ \ldots$

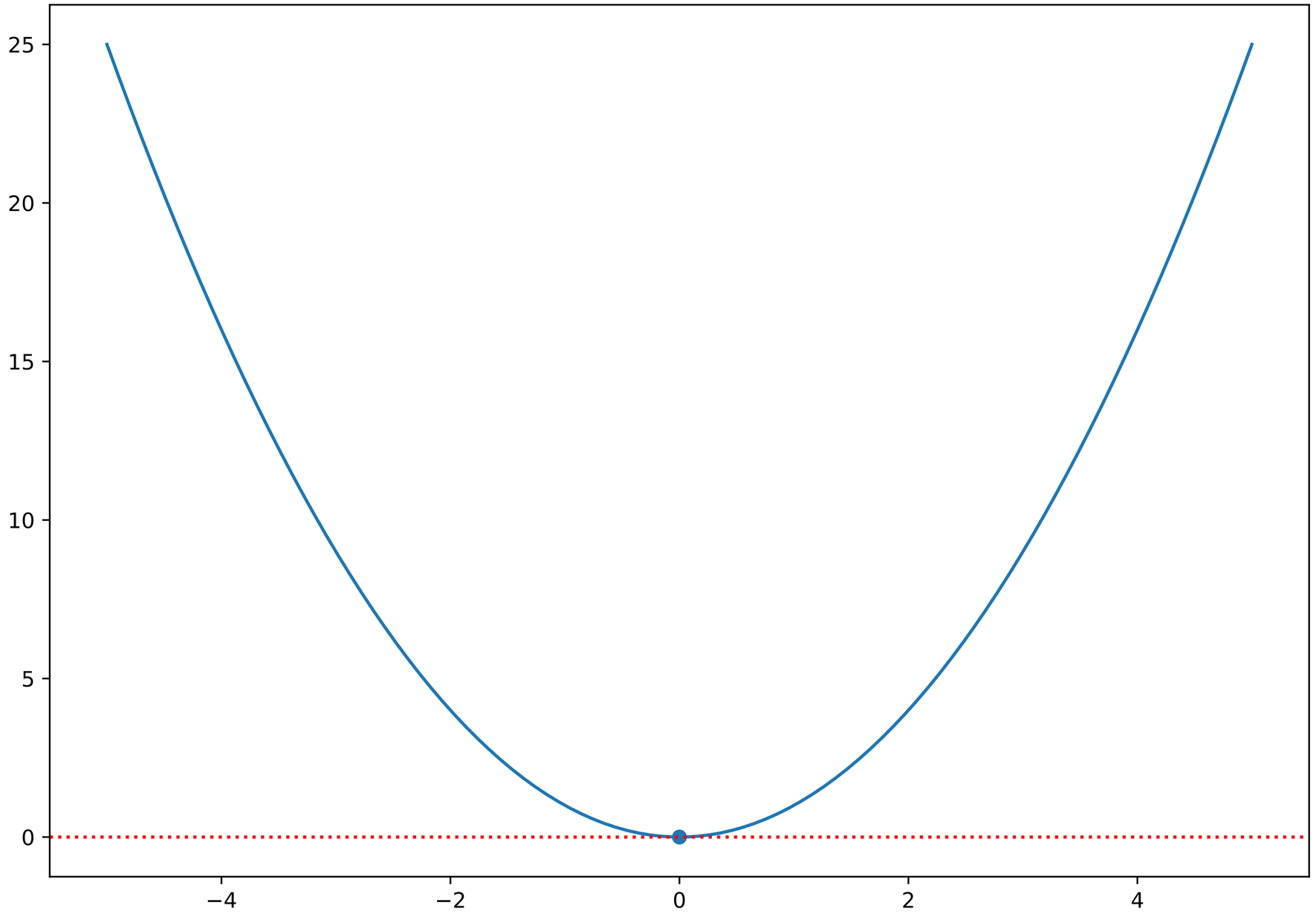
$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}}\ L(\mathbf{w}, \mathbf{X}, \mathbf{y})$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$


10
8
6
4
2
0
y
0
2
4
6
8
10
x

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}} \, \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$


25
20
15
10
5
0
−4
−2
0
2
4


25
20
15
10
5
0
−4
−2
0
2
4

$$\frac{dy}{dx} = 0$$

$$\frac{dL}{d\mathbf{w}} = 0$$

$$\nabla_{\mathbf{w}} L = 0$$

$$\nabla_{\mathbf{w}} L = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \\ \vdots \\ \frac{\partial L}{\partial w_d} \end{bmatrix}$$

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}} \, \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

$$\nabla_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 = 0$$

$$
\begin{aligned}
\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 &= (\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) \\
&= (\mathbf{X}\mathbf{w})^\mathsf{T}\mathbf{X}\mathbf{w} - (\mathbf{X}\mathbf{w})^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}(\mathbf{X}\mathbf{w}) + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= (\mathbf{X}\mathbf{w})^\mathsf{T}\mathbf{X}\mathbf{w} - 2\mathbf{y}^\mathsf{T}(\mathbf{X}\mathbf{w}) + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= (\mathbf{X}\mathbf{w})^\mathsf{T}\mathbf{X}\mathbf{w} - 2(\mathbf{y}^\mathsf{T}\mathbf{X})\mathbf{w} + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= \mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} - 2(\mathbf{X}^\mathsf{T}\mathbf{y})^\mathsf{T}\mathbf{w} + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= \mathbf{w}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{w} - 2(\mathbf{X}^\mathsf{T}\mathbf{y})^\mathsf{T}\mathbf{w} + \mathbf{y}^\mathsf{T}\mathbf{y}
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\mathbf{w}} L &= \nabla_{\mathbf{w}} \left(\mathbf{w}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X}) \mathbf{w} - 2 (\mathbf{X}^{\mathsf{T}} \mathbf{y})^{\mathsf{T}} \mathbf{w} + \mathbf{y}^{\mathsf{T}} \mathbf{y} \right) \\
&= \nabla_{\mathbf{w}} \left(\mathbf{w}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X}) \mathbf{w} - 2 (\mathbf{X}^{\mathsf{T}} \mathbf{y})^{\mathsf{T}} \mathbf{w} \right) \\
&= 2\mathbf{X}^{\mathsf{T}} \mathbf{X} \mathbf{w} - 2\mathbf{X}^{\mathsf{T}} \mathbf{y}
\end{aligned}
$$

$$\mathbf{X}^\top \mathbf{X} \mathbf{w}^\star = \mathbf{X}^\top \mathbf{y}$$

$$\mathbf{w}^\star = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

“Normal Equations”

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})\mathbf{w}^\star = \mathbf{X}^\mathsf{T}\mathbf{y}$$

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

“Revised Normal Equations”

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})\mathbf{w}^\star = \mathbf{X}^\mathsf{T}\mathbf{y}$$

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

“Revised Normal Equations”

$$\mathbf{y} == \mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$$

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

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ L\left(\mathbf{f}, \boldsymbol{\theta}, \{\mathbf{X}[, \mathbf{Y}]\}\right)$$

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ L_{\mathbf{f},\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})$$

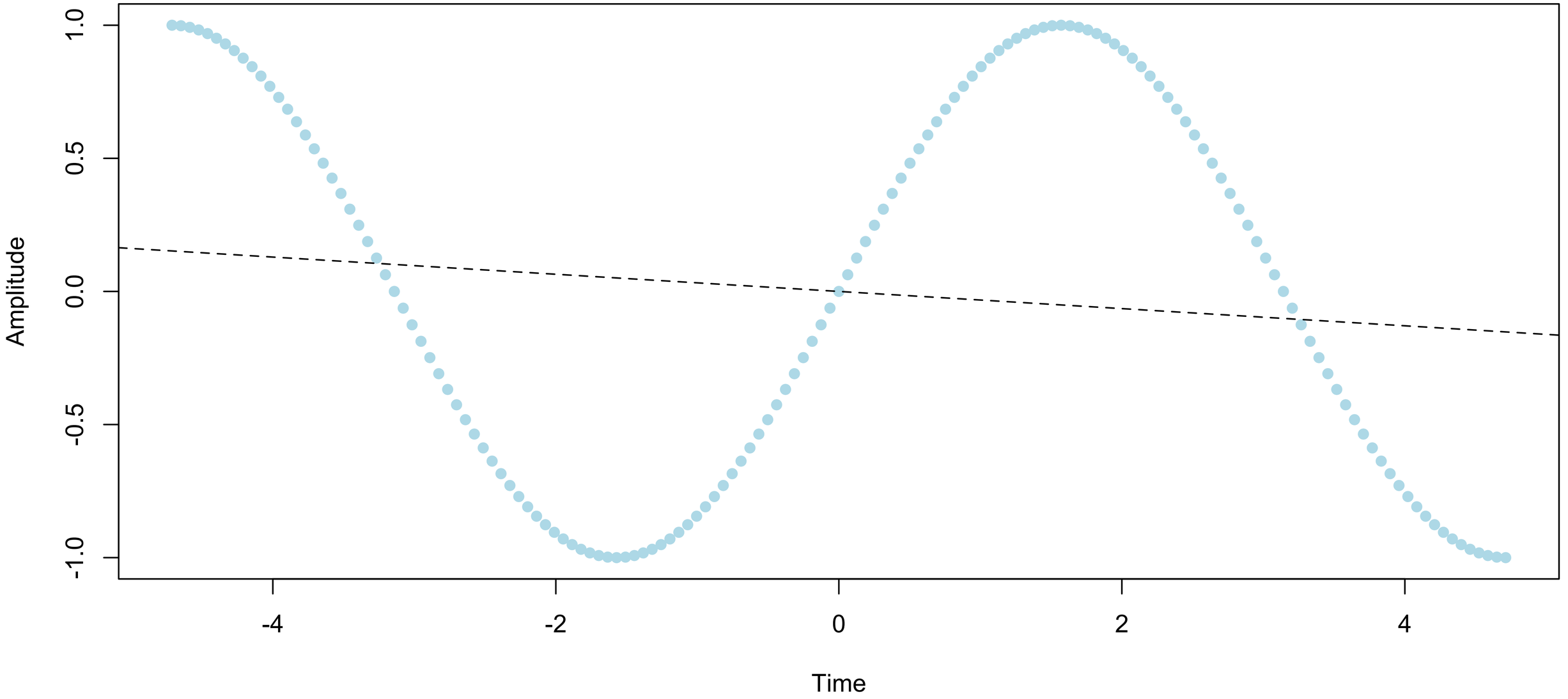
$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x},\mathbf{y},\boldsymbol{\theta})$$

$$L(\boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ L_{\mathbf{f},\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})$$

training ≠ test

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


Amplitude
1.0
0.5
0.0
-0.5
-1.0
-4
-2
0
2
4
Time

Don’t fit a straight line to curvy data

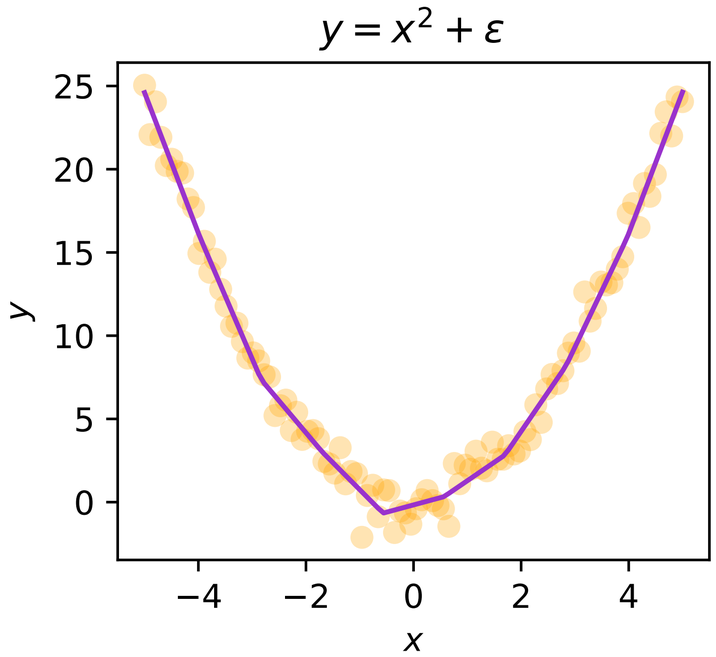
$$\hat{y} = \sum_{i}^{k} w_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x}) \cdot \mathbf{w} = \mathbf{x}' \cdot \mathbf{w}$$

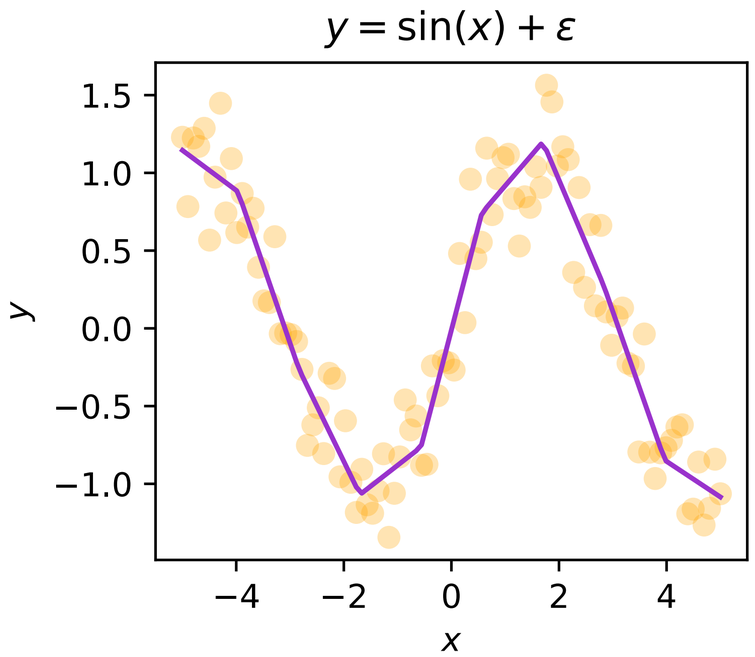
$$\mathbf{x}' = \mathbf{h}(\mathbf{x}) = [\, h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_k(\mathbf{x}) \,]$$

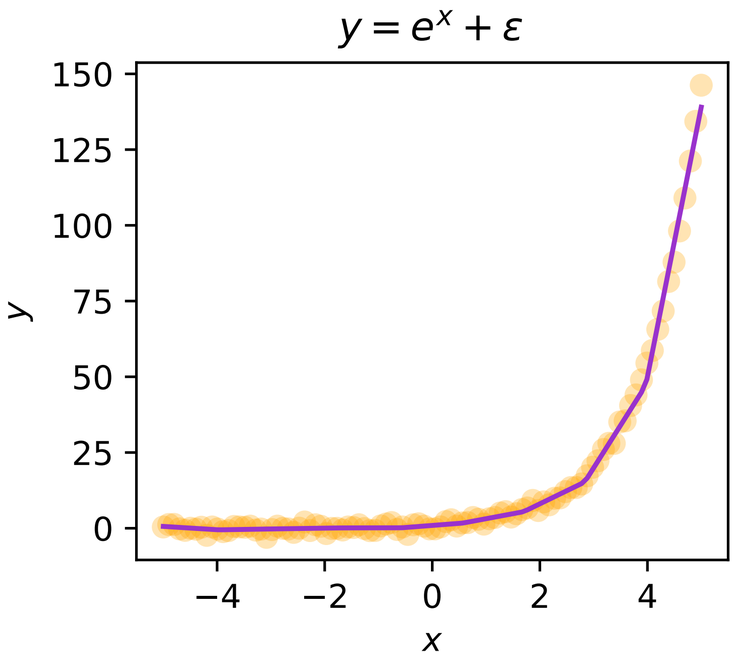
$$h_1(x) = 1$$

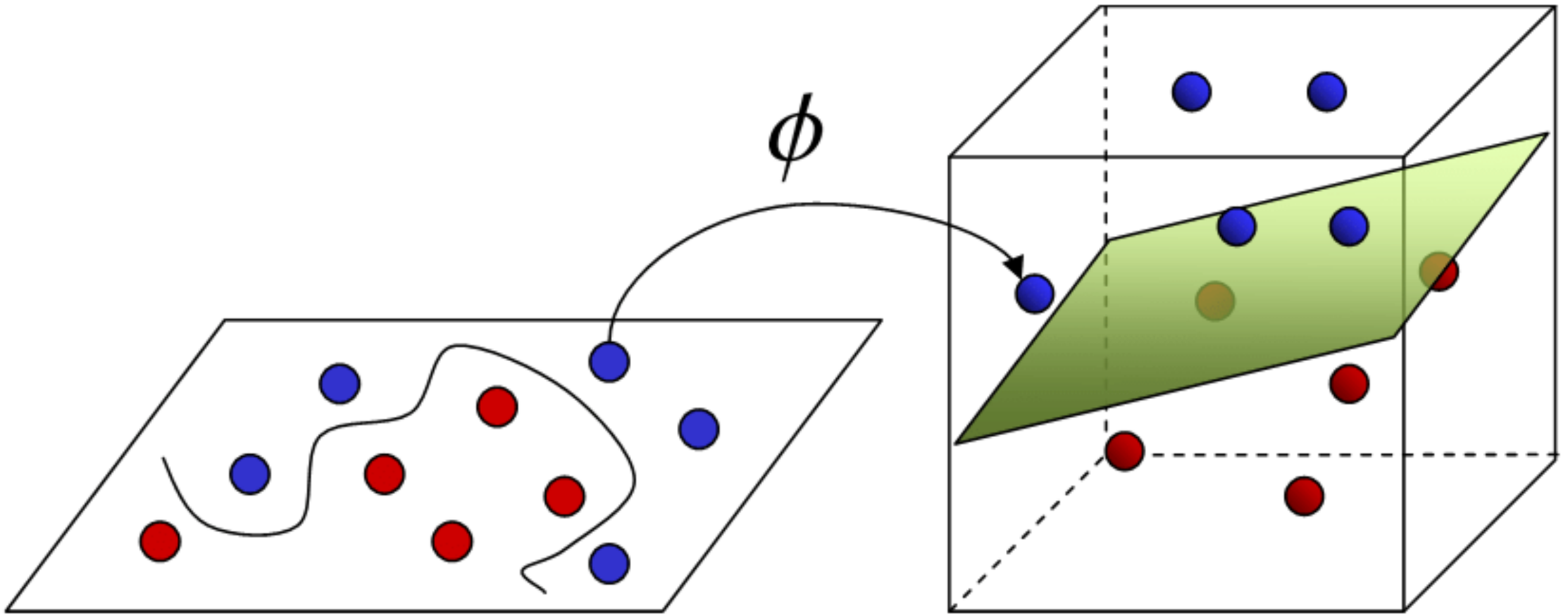
$$h_2(x) = x$$

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








ϕ
a) Input Space
b) Feature Space


DEFEND THE CHILDREN OF THE
POOR & PVNISH THE WRONGDOER

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

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

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

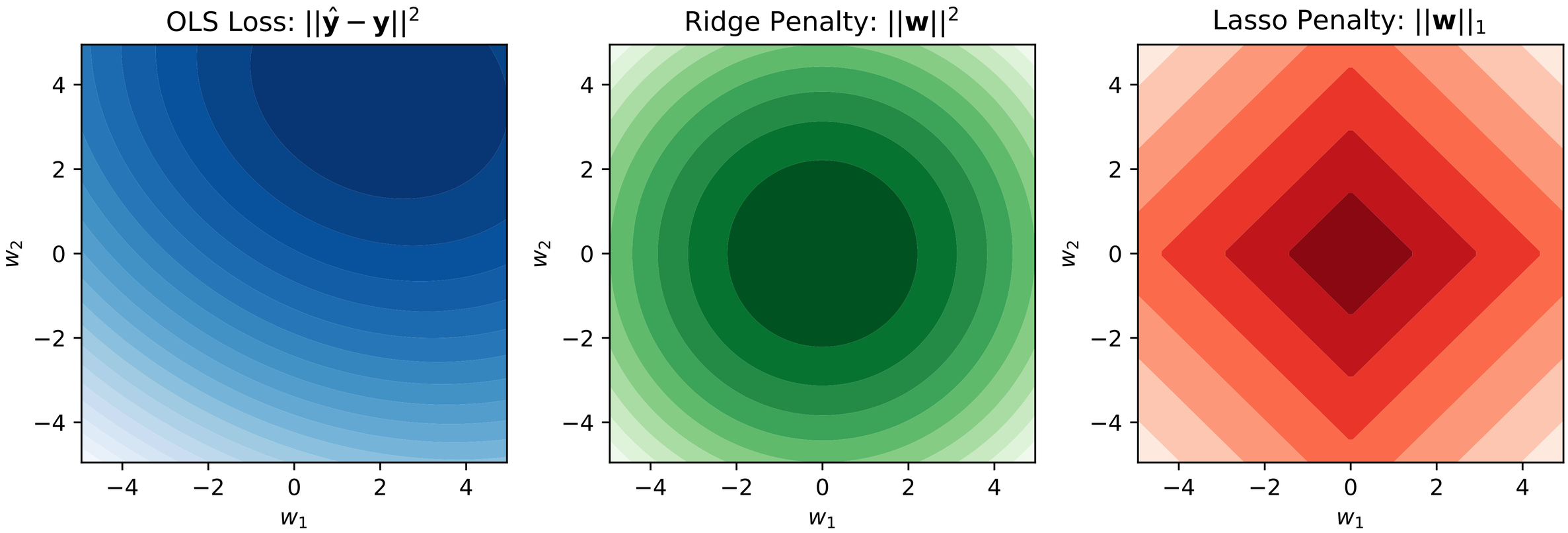
$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2$$

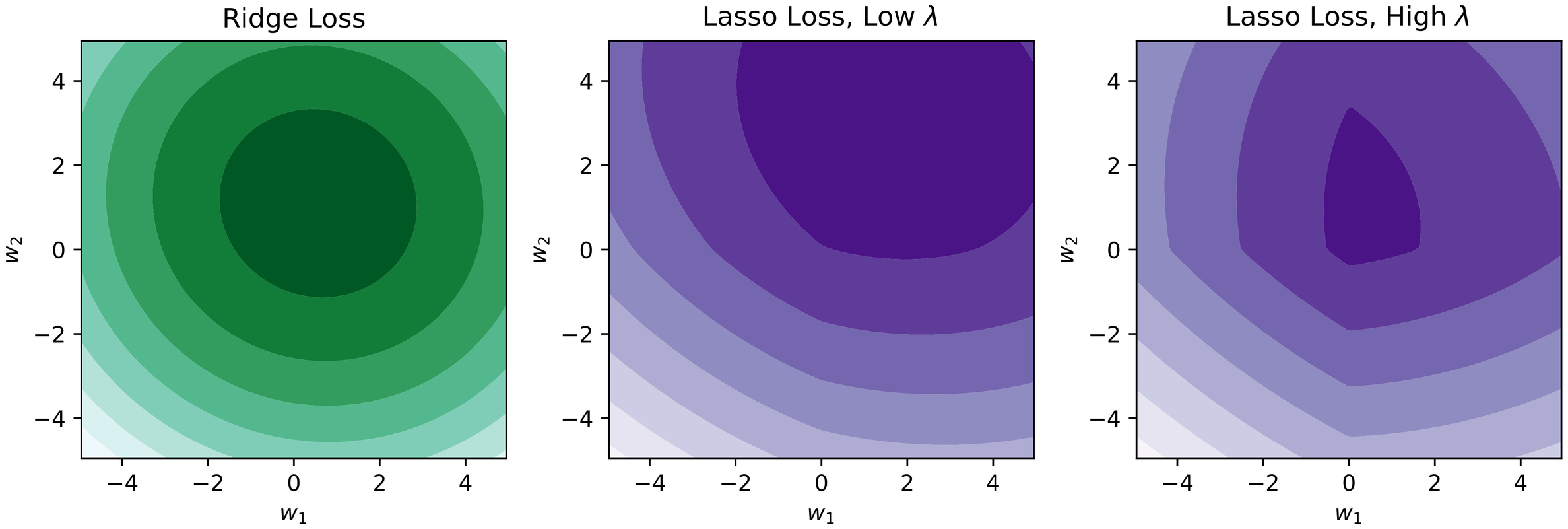
“Ridge Regression”

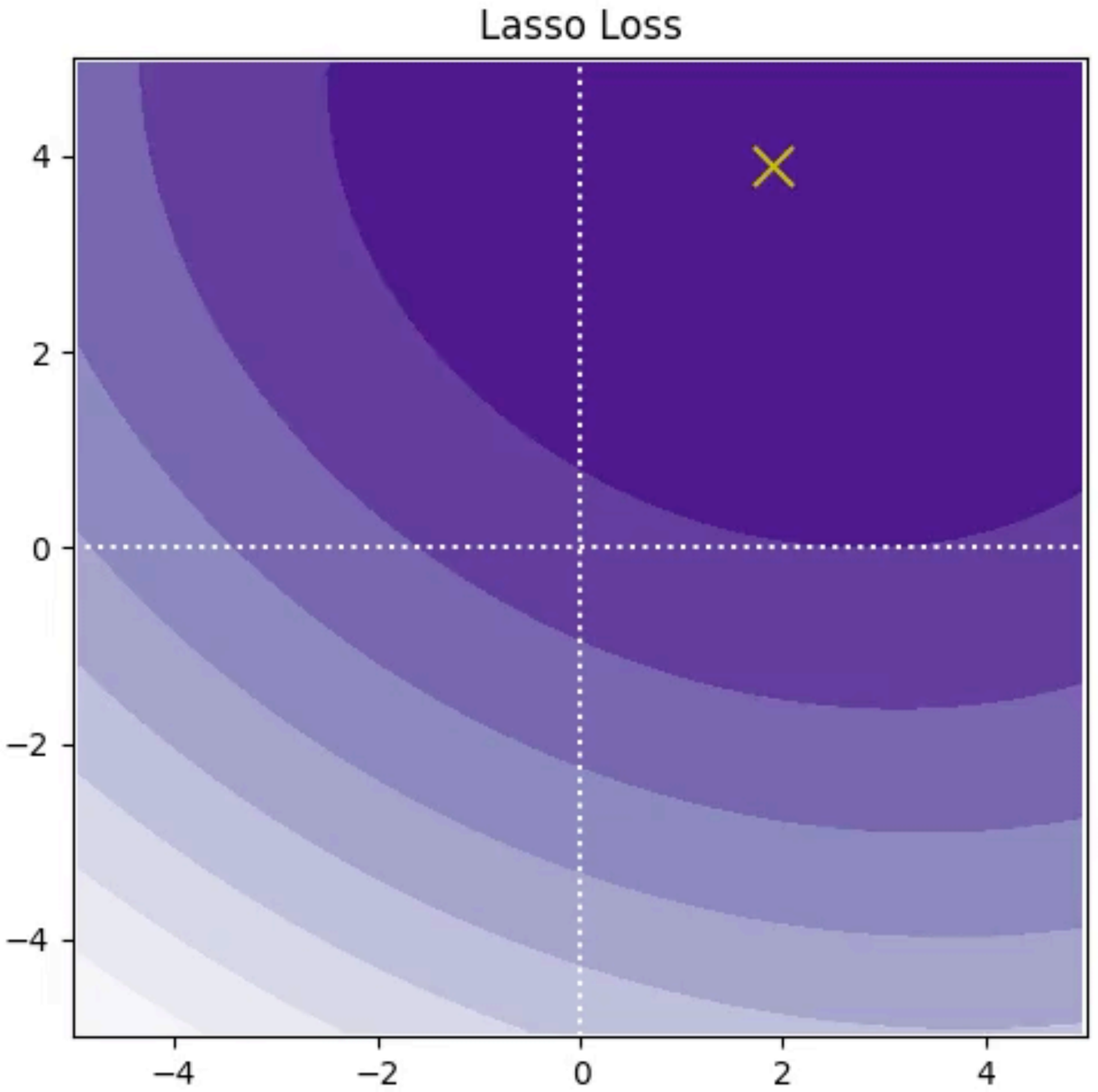
$$P(\mathbf{w}) = \|\mathbf{w}\|_1$$

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

“Lasso”


OLS Loss: $||\hat{\mathbf{y}} - \mathbf{y}||^2$
Ridge Penalty: $||\mathbf{w}||^2$
Lasso Penalty: $||\mathbf{w}||_1$
w_1
w_2
4
2
0
−2
−4


Ridge Loss
Lasso Loss, Low λ
Lasso Loss, High λ
w1
w2
4
2
0
−2
−4


Lasso Loss
4
2
0
−2
−4
−4
−2
0
2
4

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

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

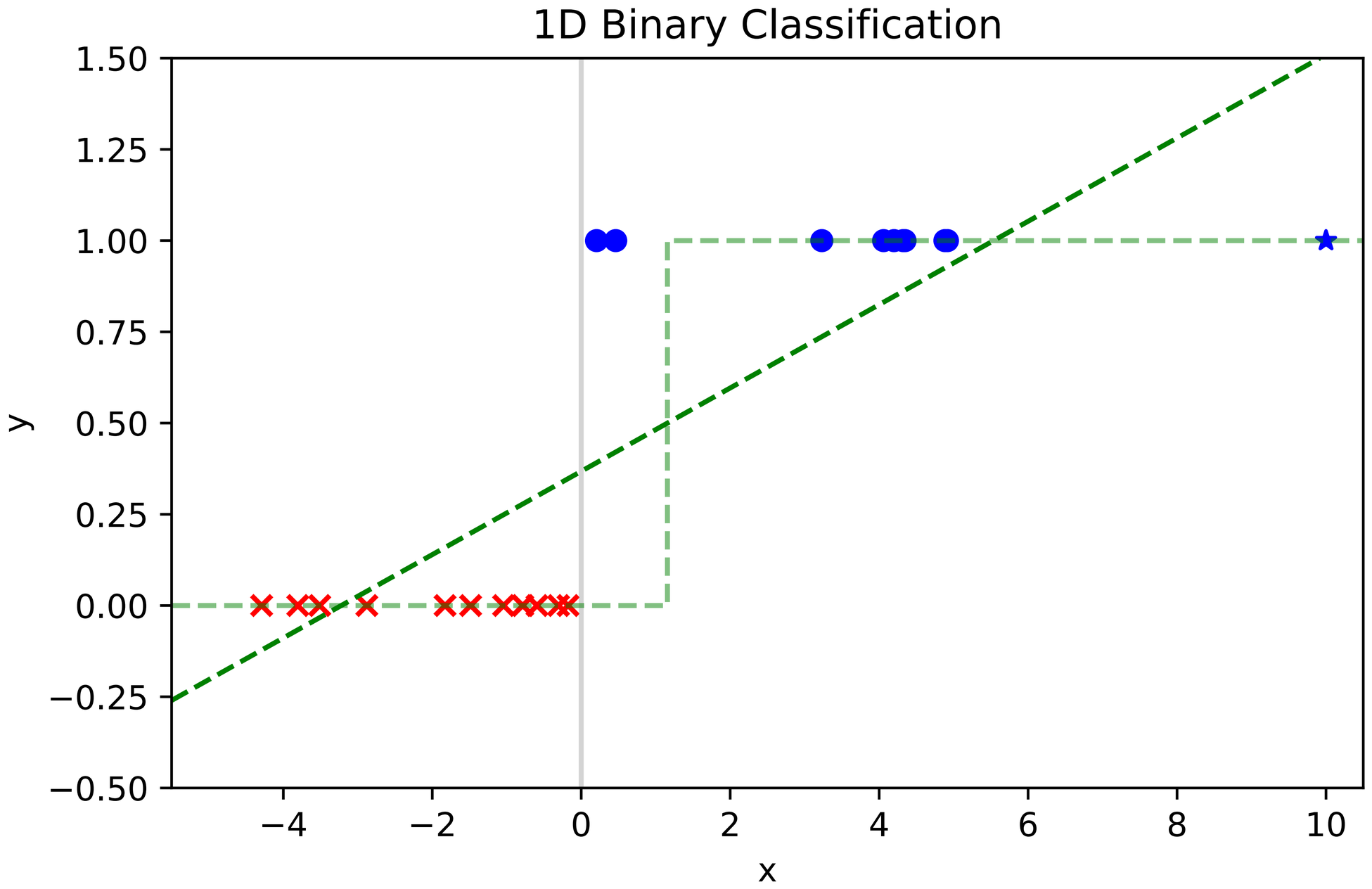
$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} L(\mathbf{f}, \mathbf{X}, \mathbf{y}, \mathbf{w})$$

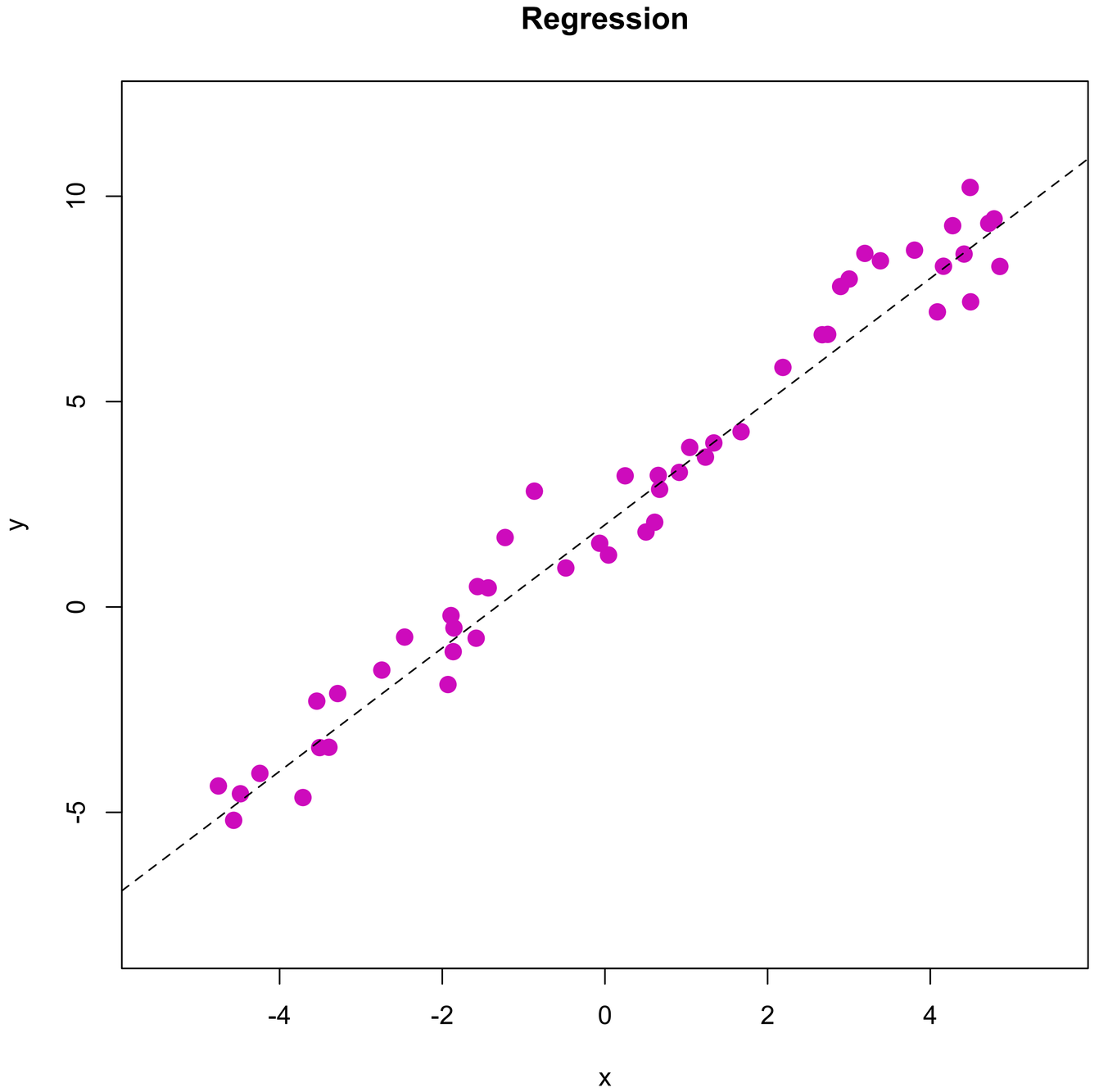
“Gradient Descent”

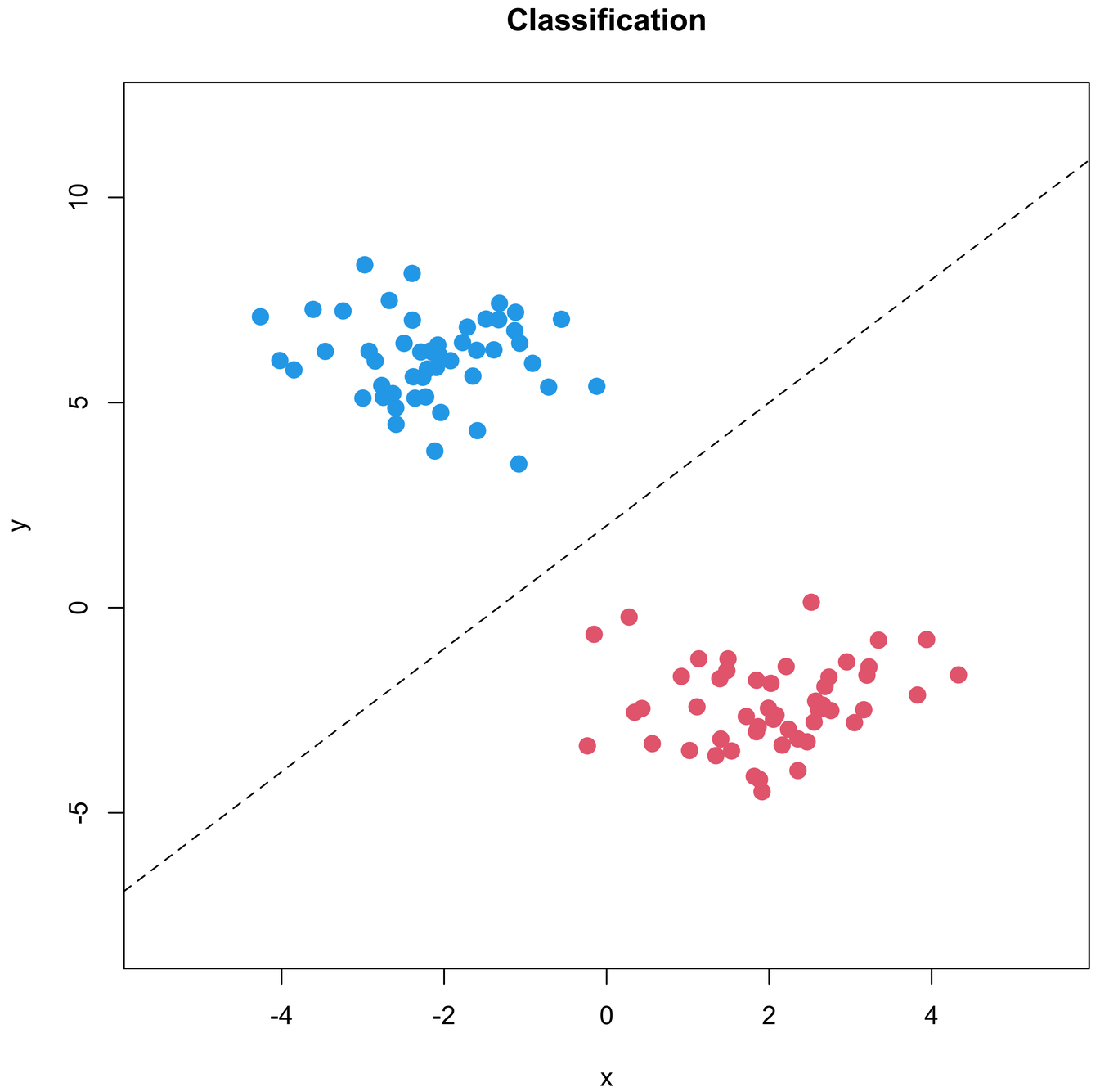
$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} L(\mathbf{f}, \mathbf{X}, \mathbf{y}, \mathbf{w})$$

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

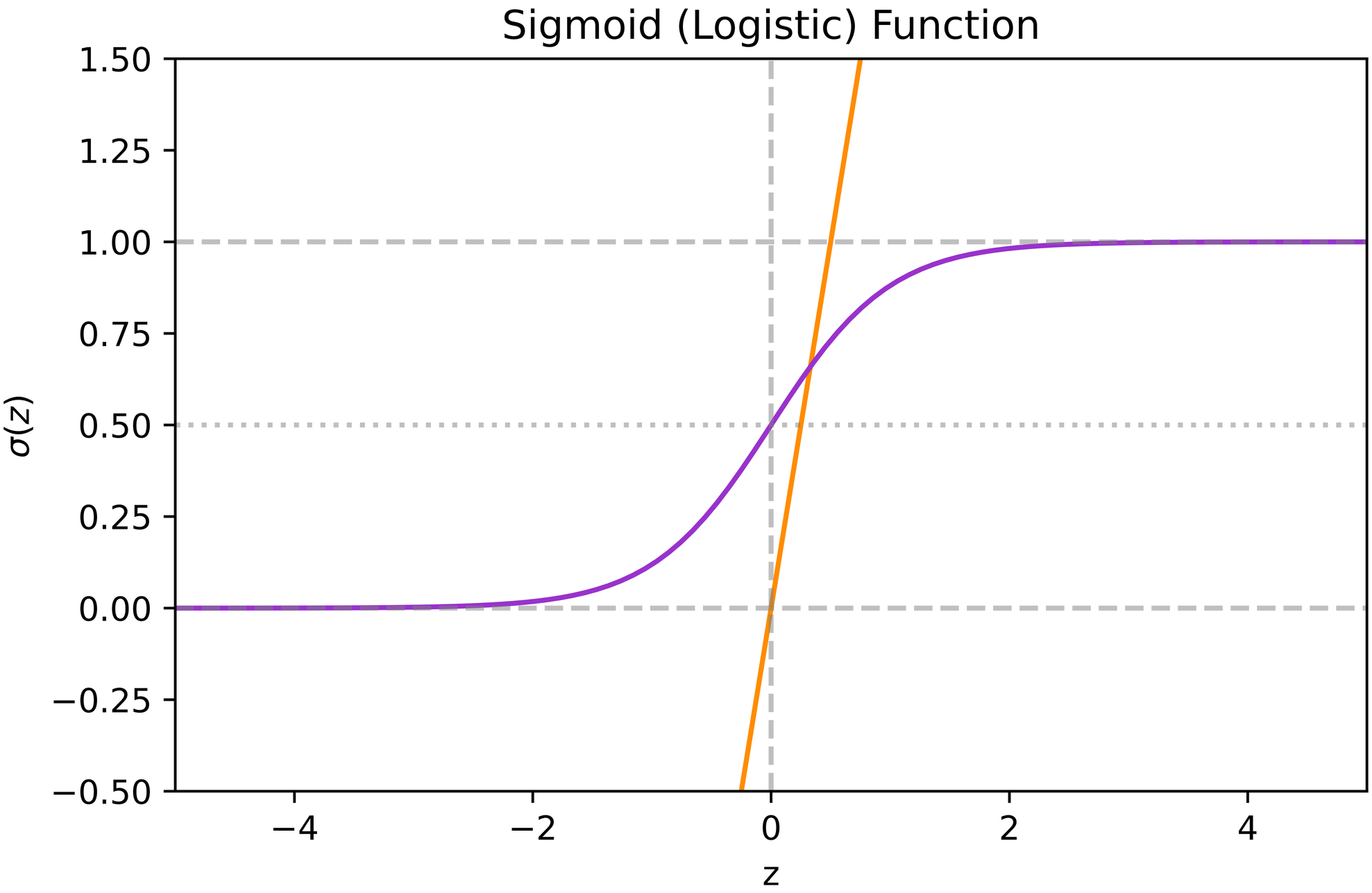
“Stochastic Gradient Descent”


1D Binary Classification
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
y
−4
−2
0
2
4
6
8
10
x


Regression
y
x
10
5
0
-5
-4
-2
0
2
4


Classification
y
10
5
0
-5
-4
-2
0
2
4
x

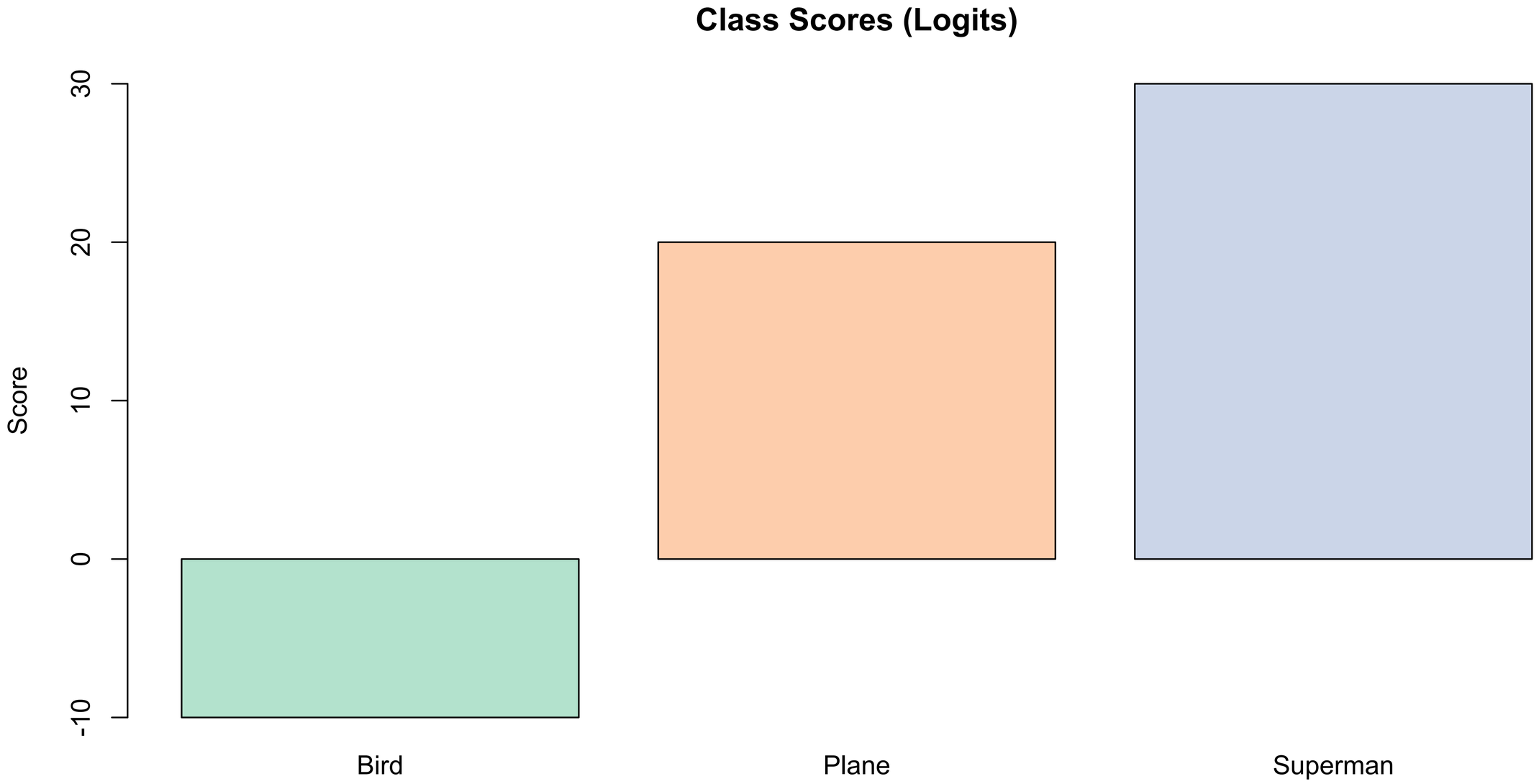
$$\sigma(z) = \frac{1}{1+e^{-z}}$$

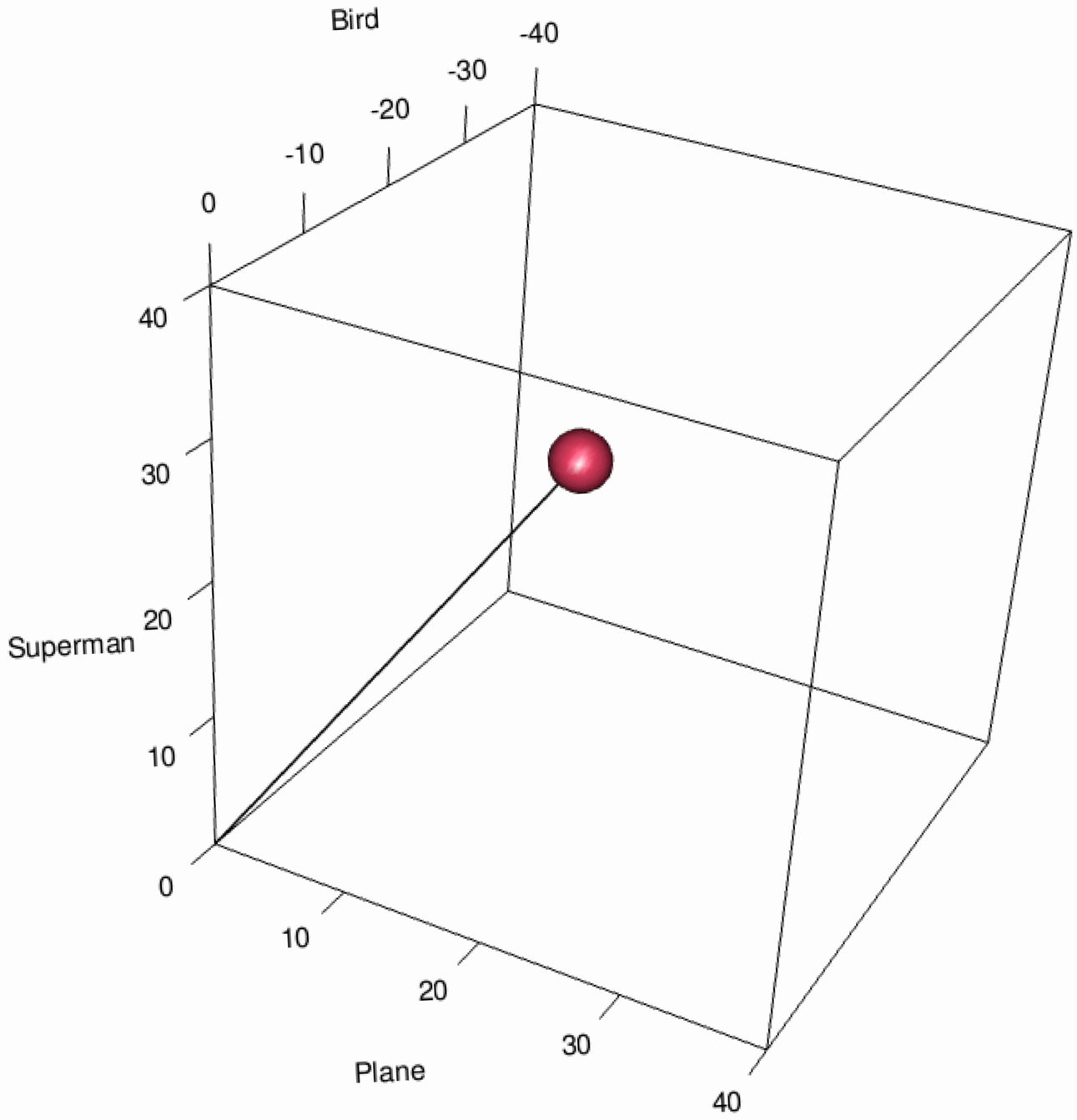

Sigmoid (Logistic) Function
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
σ(z)
−4
−2
0
2
4
z

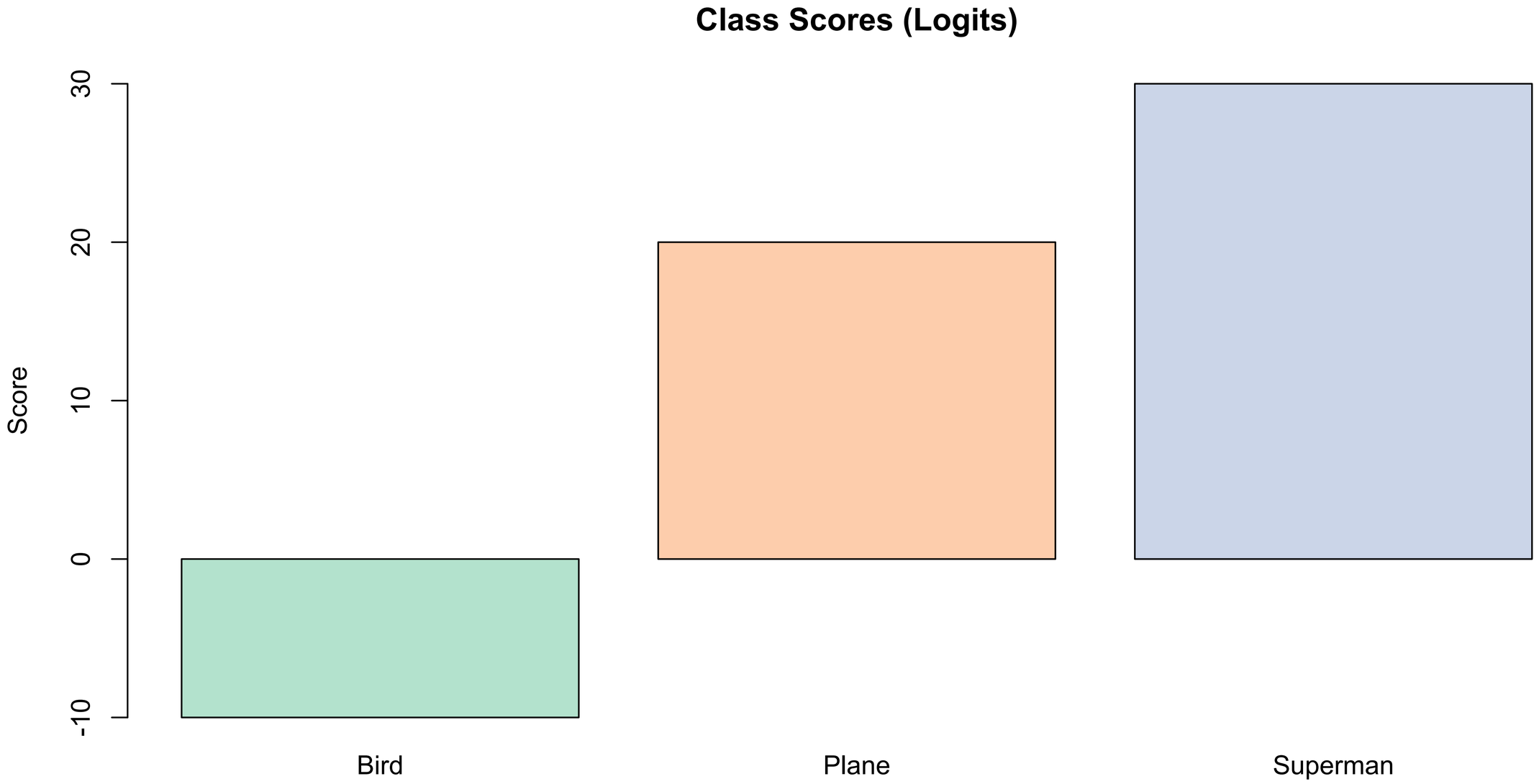
$$\nabla_{\mathbf{w}} l = \mathbf{X}^{\top}(\mathbf{y} - \hat{\mathbf{y}})$$

$$= \mathbf{X}^{\top}(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}))$$

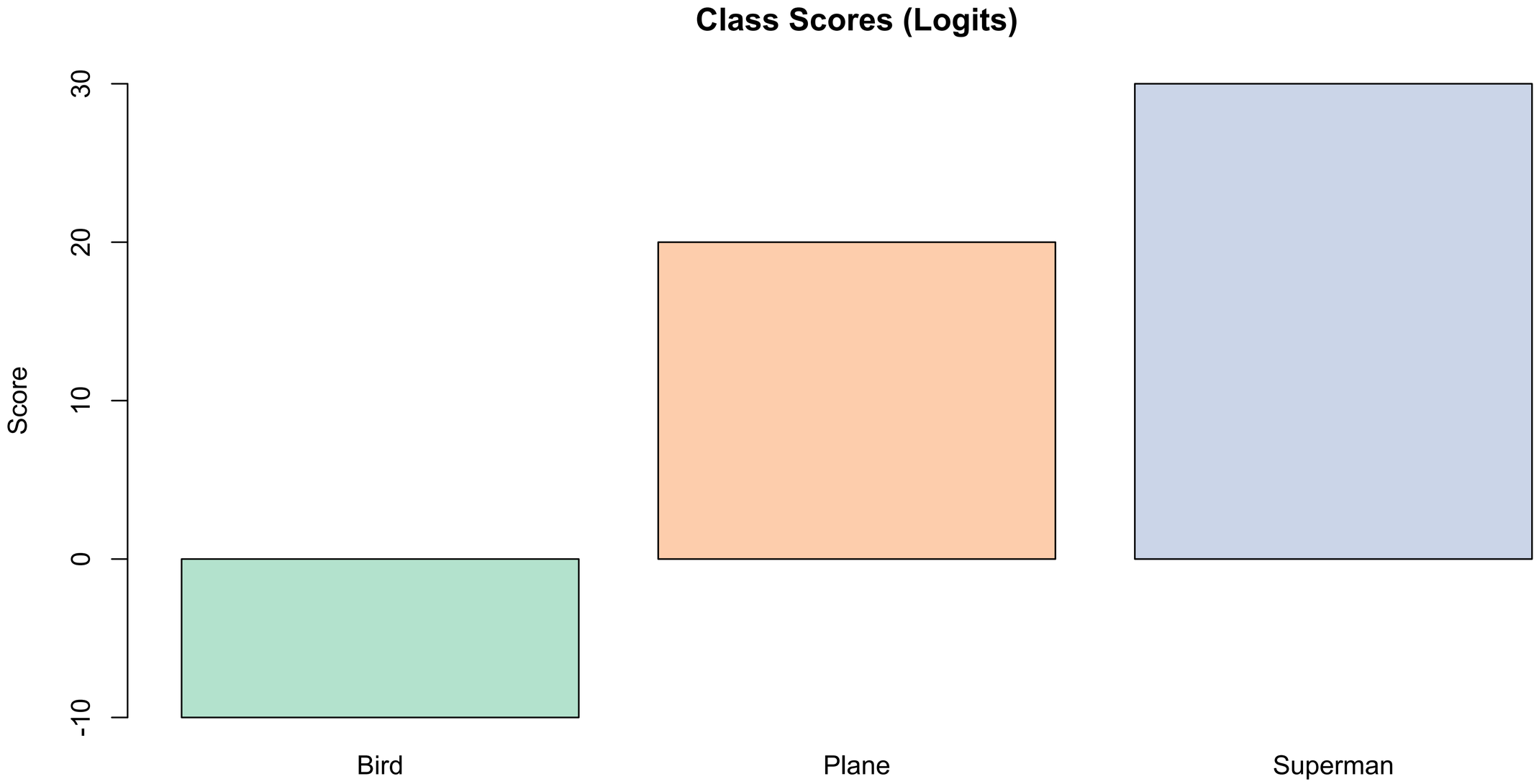
$$= \mathbf{X}^{\top}\left(\mathbf{y} - \frac{1}{1 + e^{-\mathbf{X}\mathbf{w}}}\right)$$

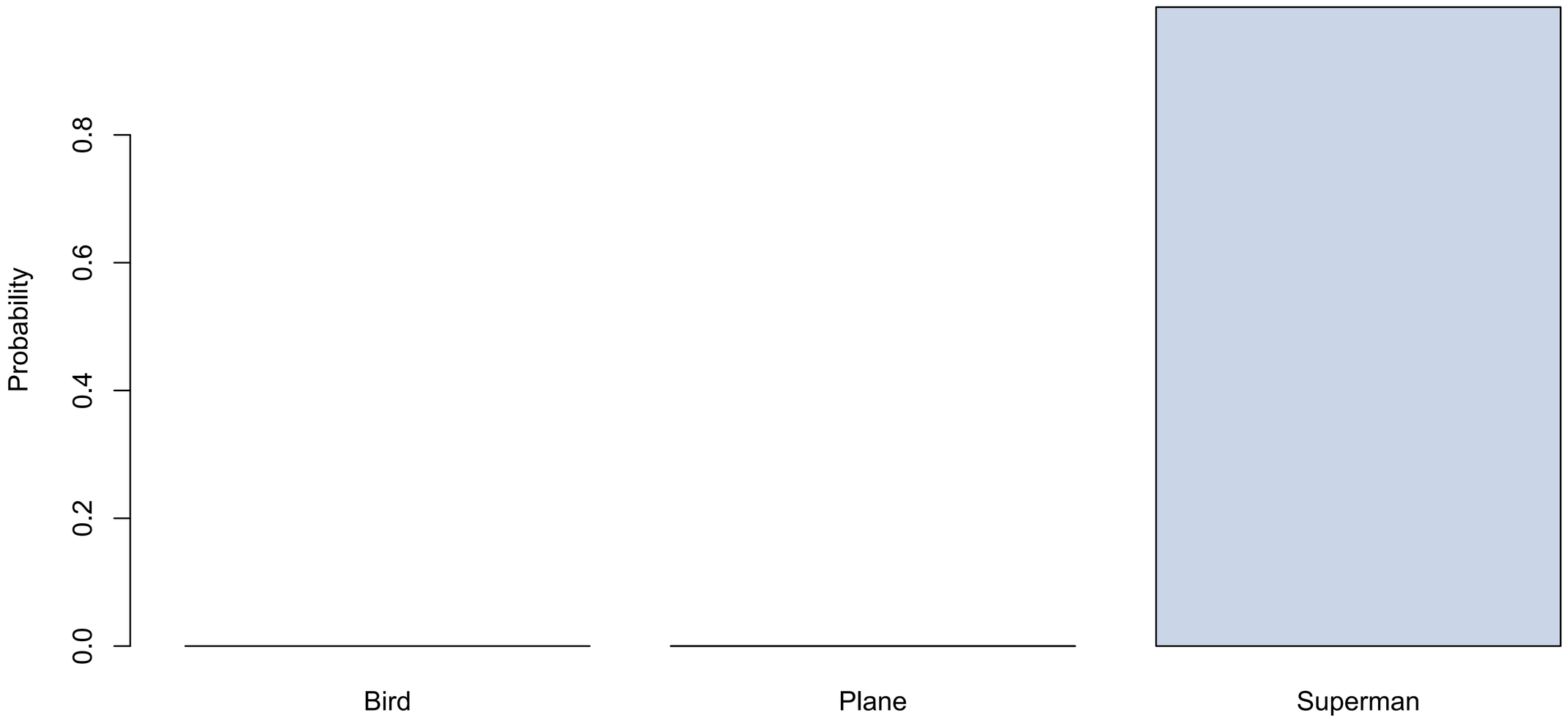

Class Scores (Logits)
Score
30
20
10
0
-10
Bird
Plane
Superman


Bird
-40
-30
-20
-10
0
40
30
20
Superman
10
0
10
20
30
Plane
40


Class Scores (Logits)
Score
30
20
10
0
-10
Bird
Plane
Superman

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


Class Scores (Logits)
Score
30
20
10
0
-10
Bird
Plane
Superman


Class Probabilities
Probability
0.0
0.2
0.4
0.6
0.8
Bird
Plane
Superman

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

Questions?

Next: Non-Parametric Models