Week 4: Linear Models Revisited

Matthew Caldwell

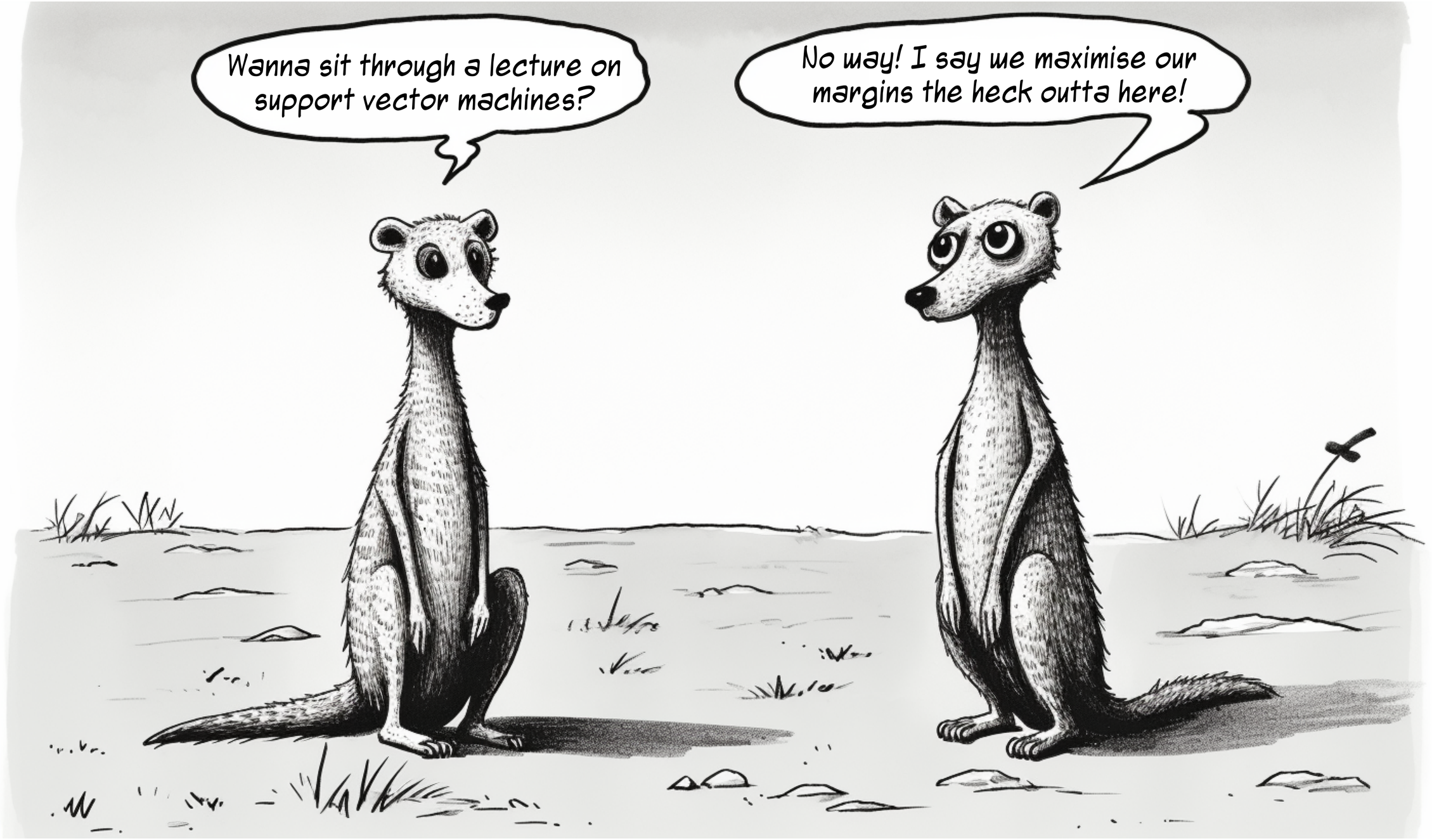
COMP0088 Introduction to Machine Learning • UCL Computer Science • Autumn 2025

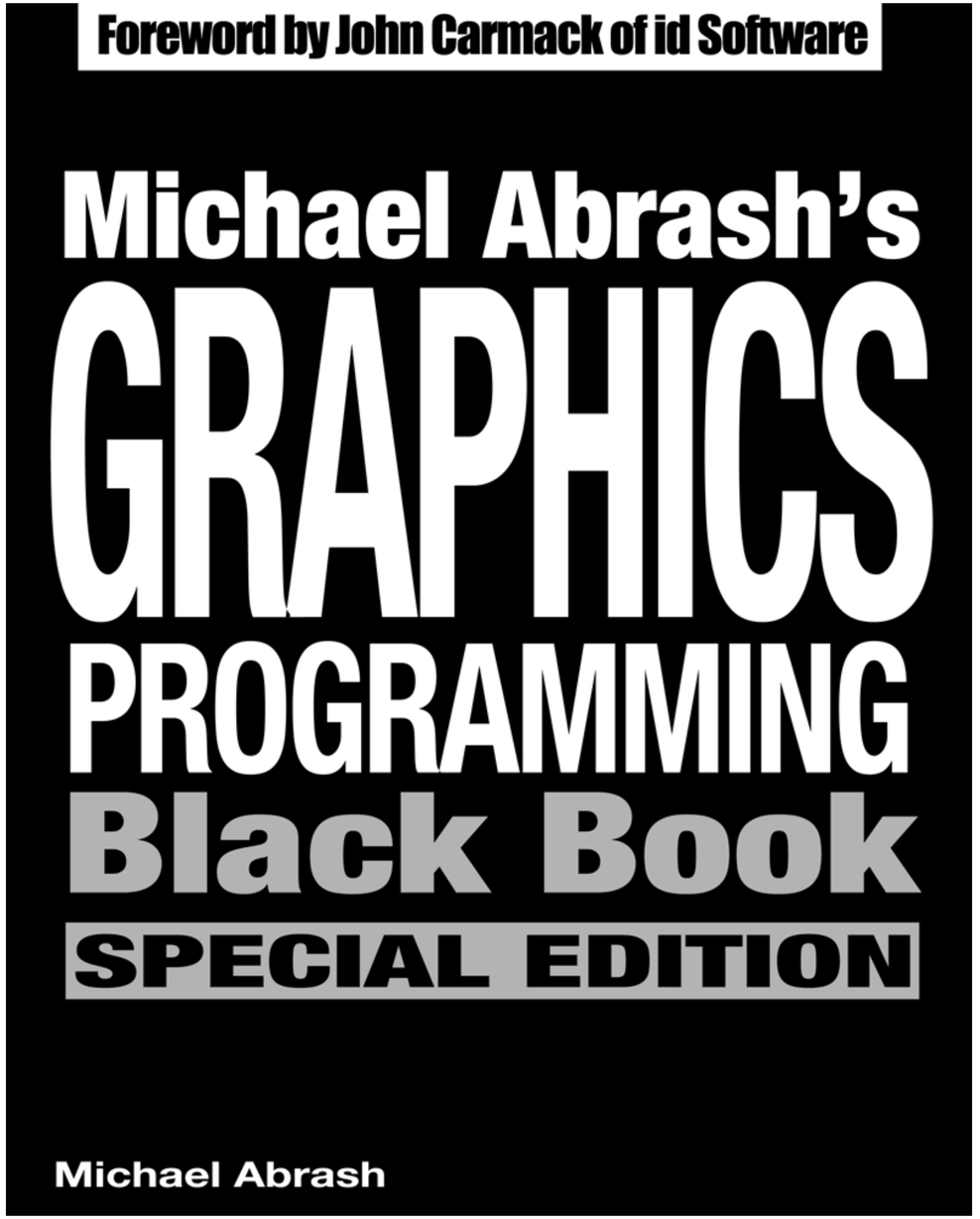
Admin

- Pulse surveys?
- Coursework

Week 4 Recap

Joining the Dots


Wanna sit through a lecture on support vector machines?
No way! I say we maximise our margins the heck outta here!


Foreword by John Carmack of id Software
Michael Abrash's
GRAPHICS
PROGRAMMING
Black Book
SPECIAL EDITION
Michael Abrash


YOU RECEIVE 15 HEALTH
271 01 01 01
99
27

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b} = \langle \mathbf{a}, \mathbf{b} \rangle = \sum_i a_i b_i$$

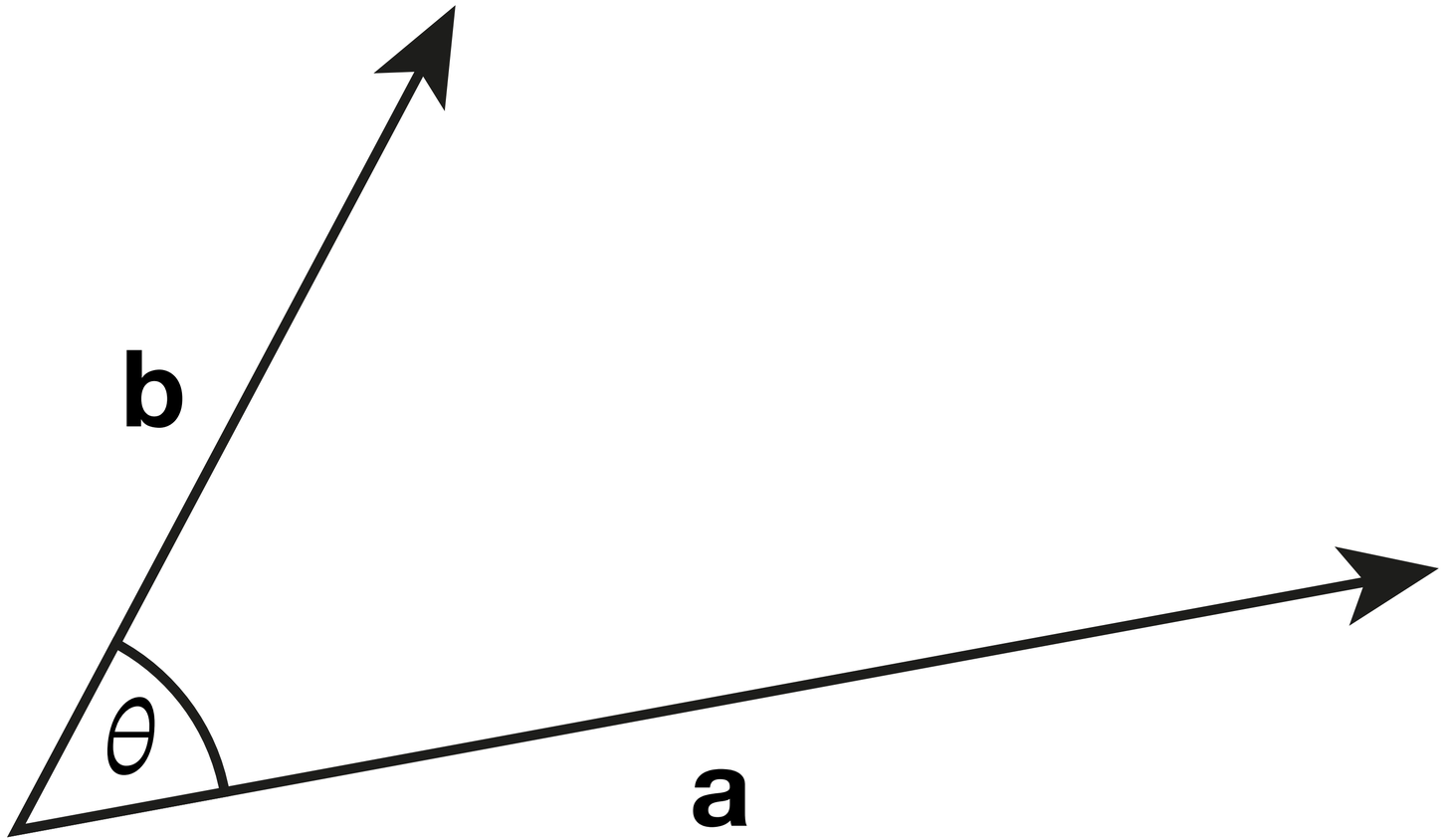
$$\mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|^2$$

$$\mathbf{w} \cdot \mathbf{x} = w_i x_1 + w_2 x_2 + ... + w_d x_d$$

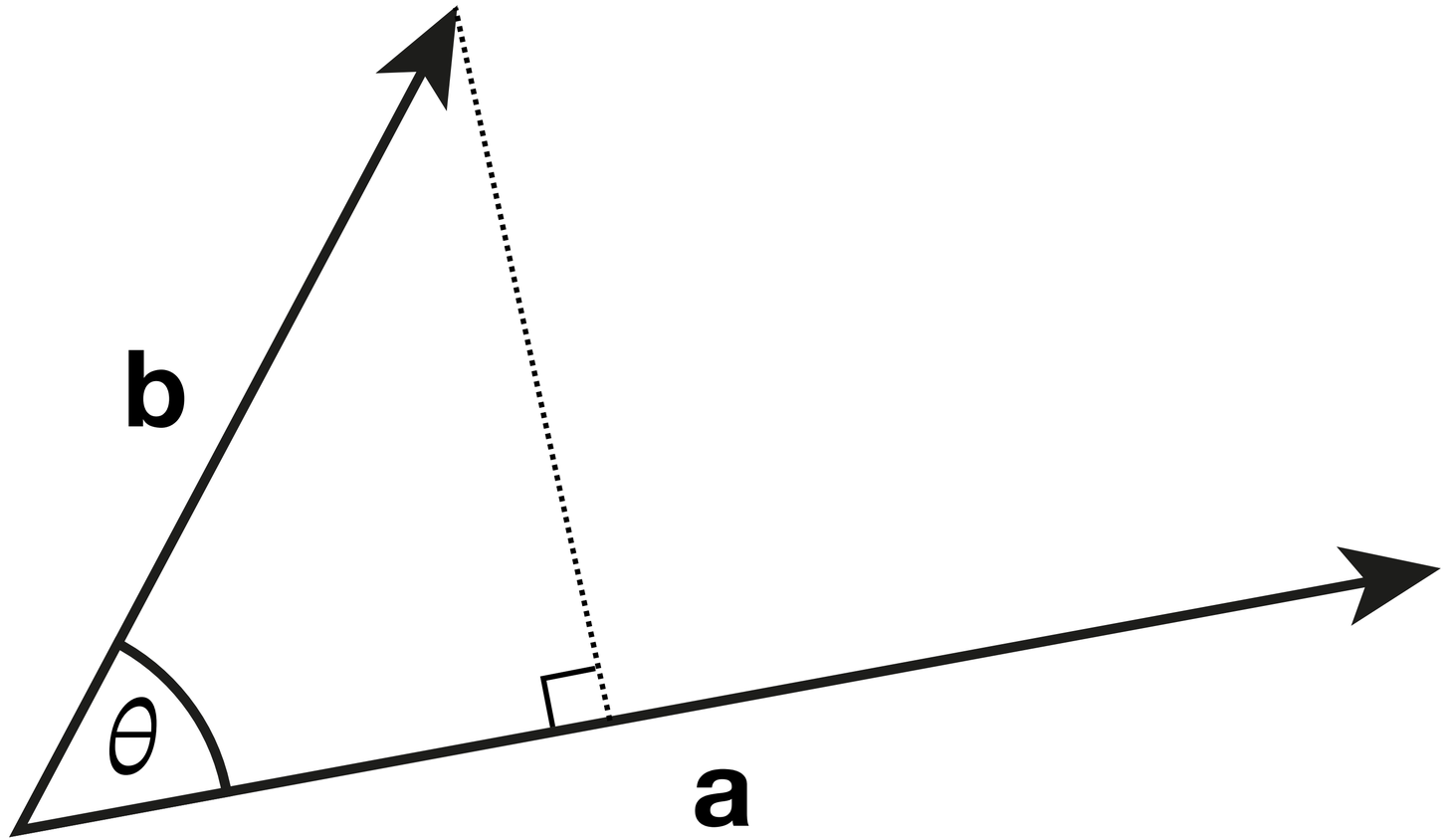
$$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

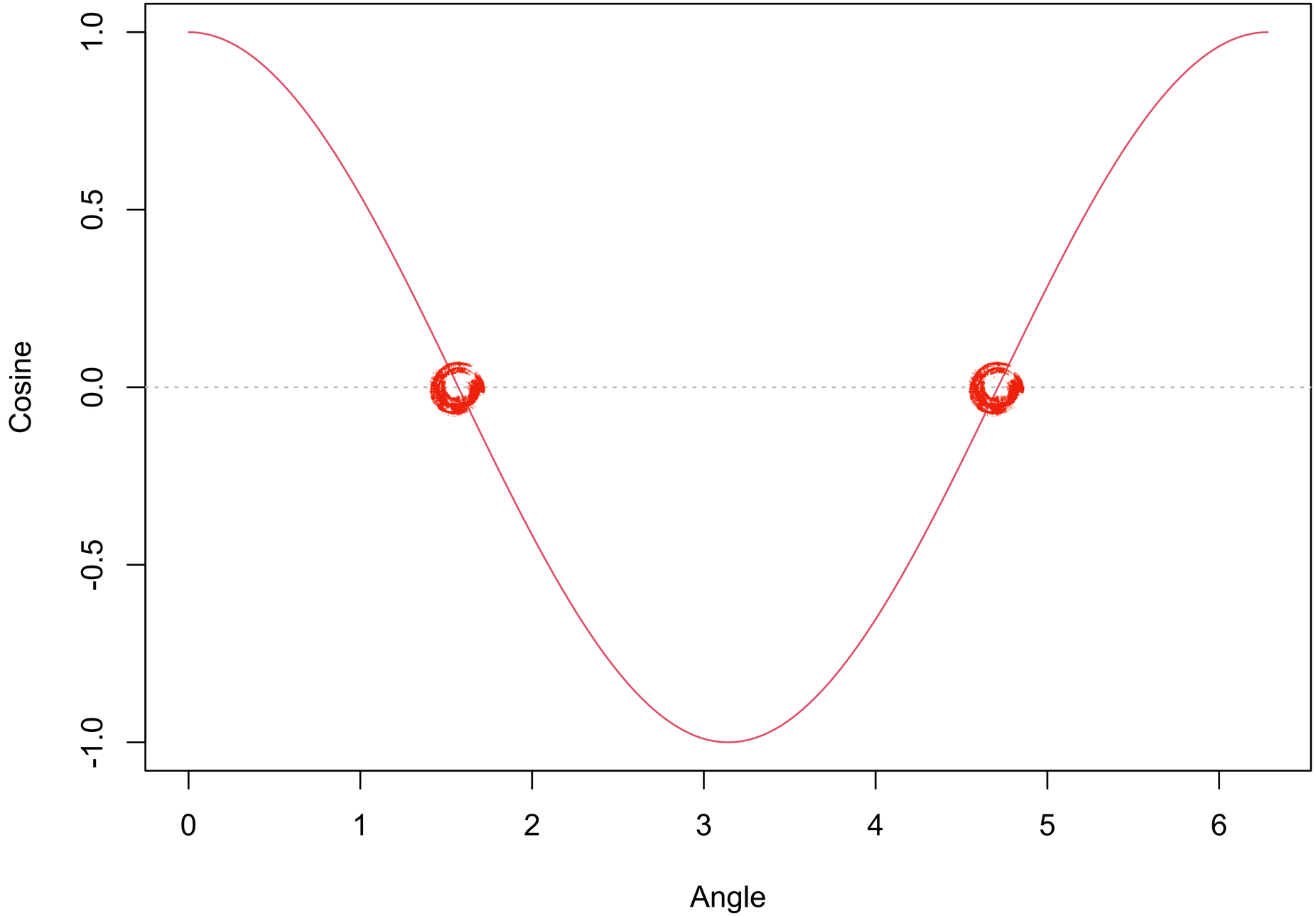
$$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_d \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 \cdot \mathbf{b}_1 & \mathbf{a}_1 \cdot \mathbf{b}_2 & \dots & \mathbf{a}_1 \cdot \mathbf{b}_d \\ \mathbf{a}_2 \cdot \mathbf{b}_1 & \mathbf{a}_2 \cdot \mathbf{b}_2 & \dots & \mathbf{a}_2 \cdot \mathbf{b}_d \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n \cdot \mathbf{b}_1 & \mathbf{a}_n \cdot \mathbf{b}_2 & \dots & \mathbf{a}_n \cdot \mathbf{b}_d \end{bmatrix}$$


b
θ
a

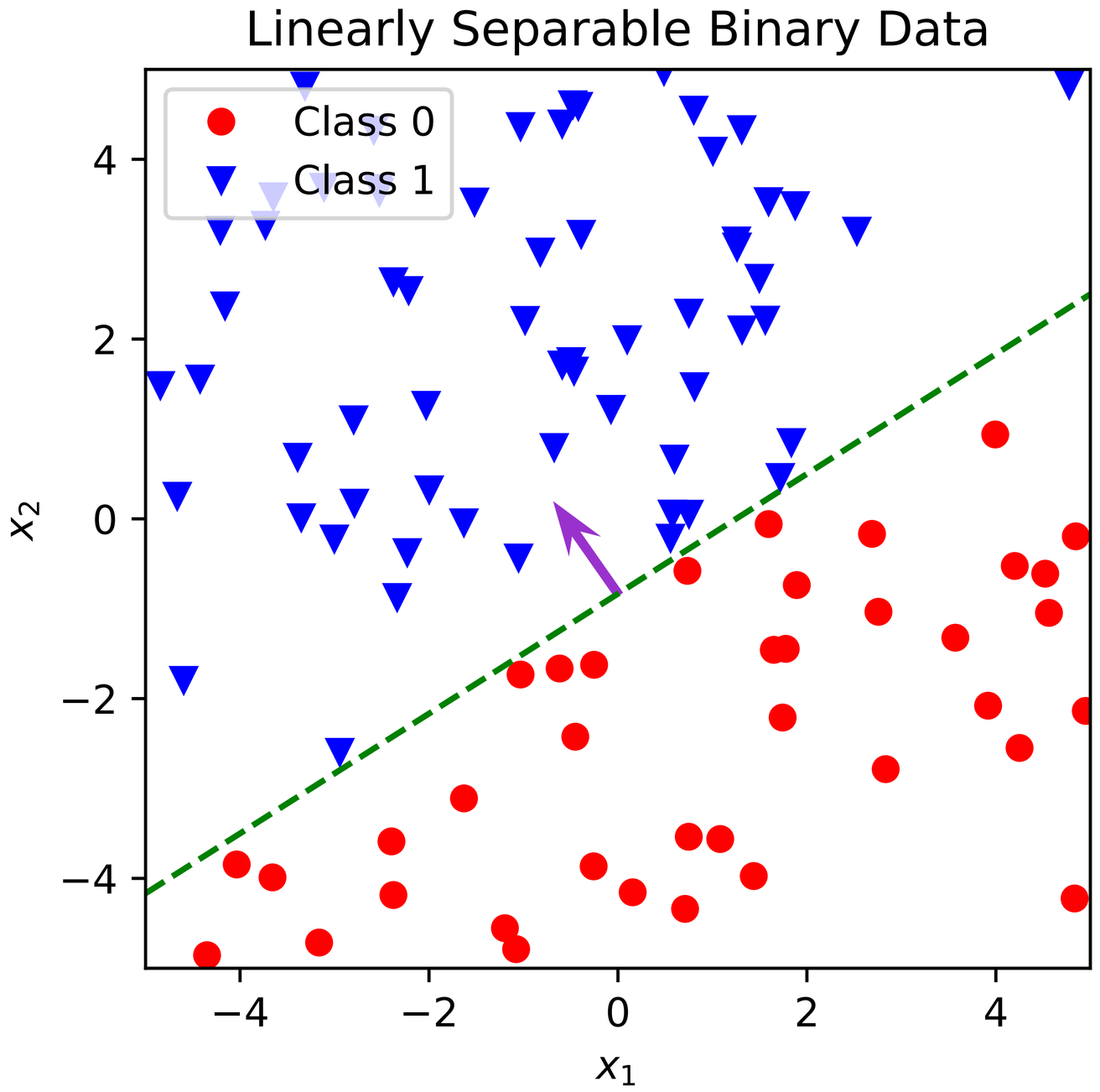
$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

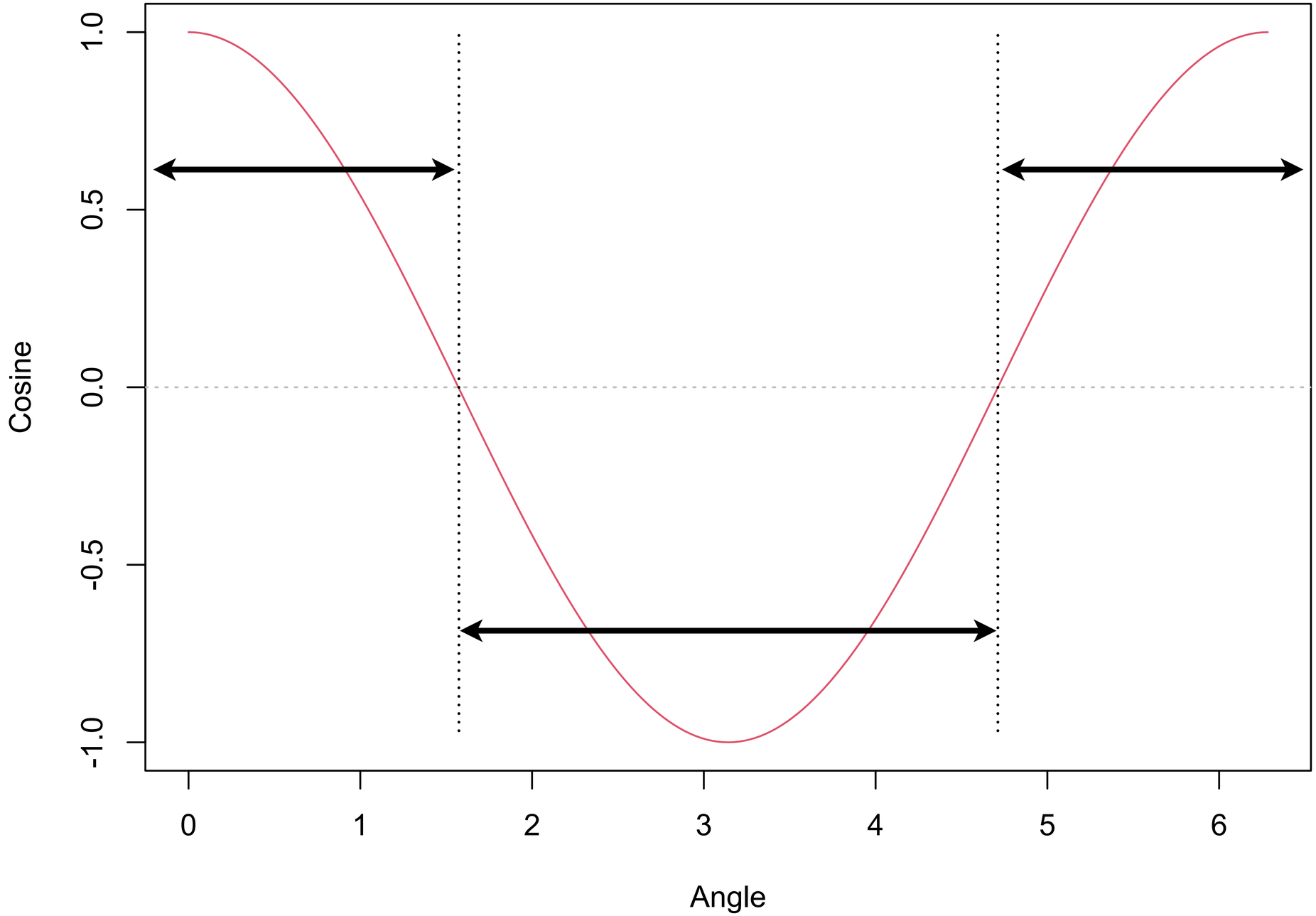

b
θ
a


Cosine
1.0
0.5
0.0
-0.5
-1.0
0
1
2
3
4
5
6
Angle

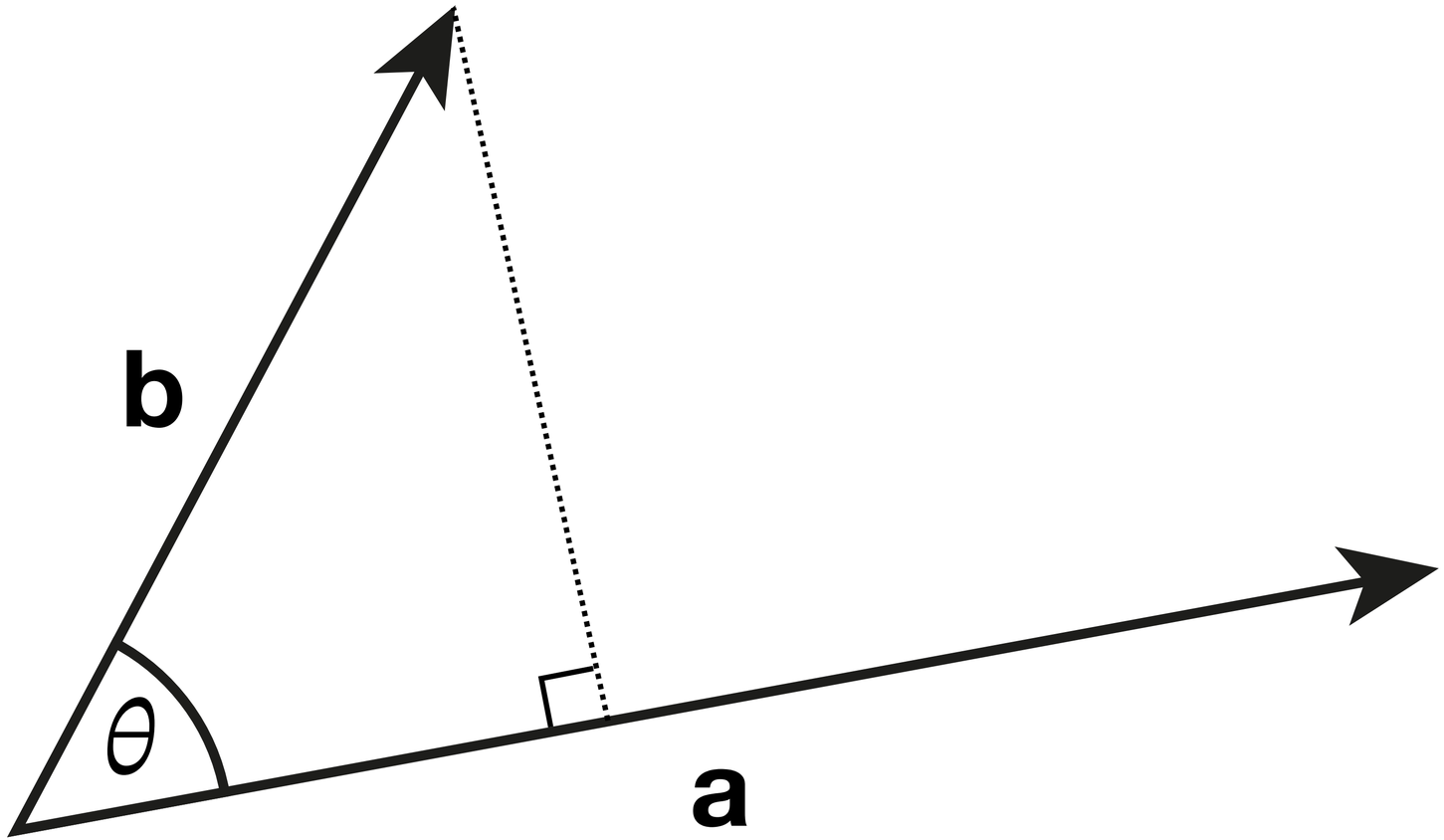
$$\mathbf{x} \cdot \mathbf{w} = 0$$

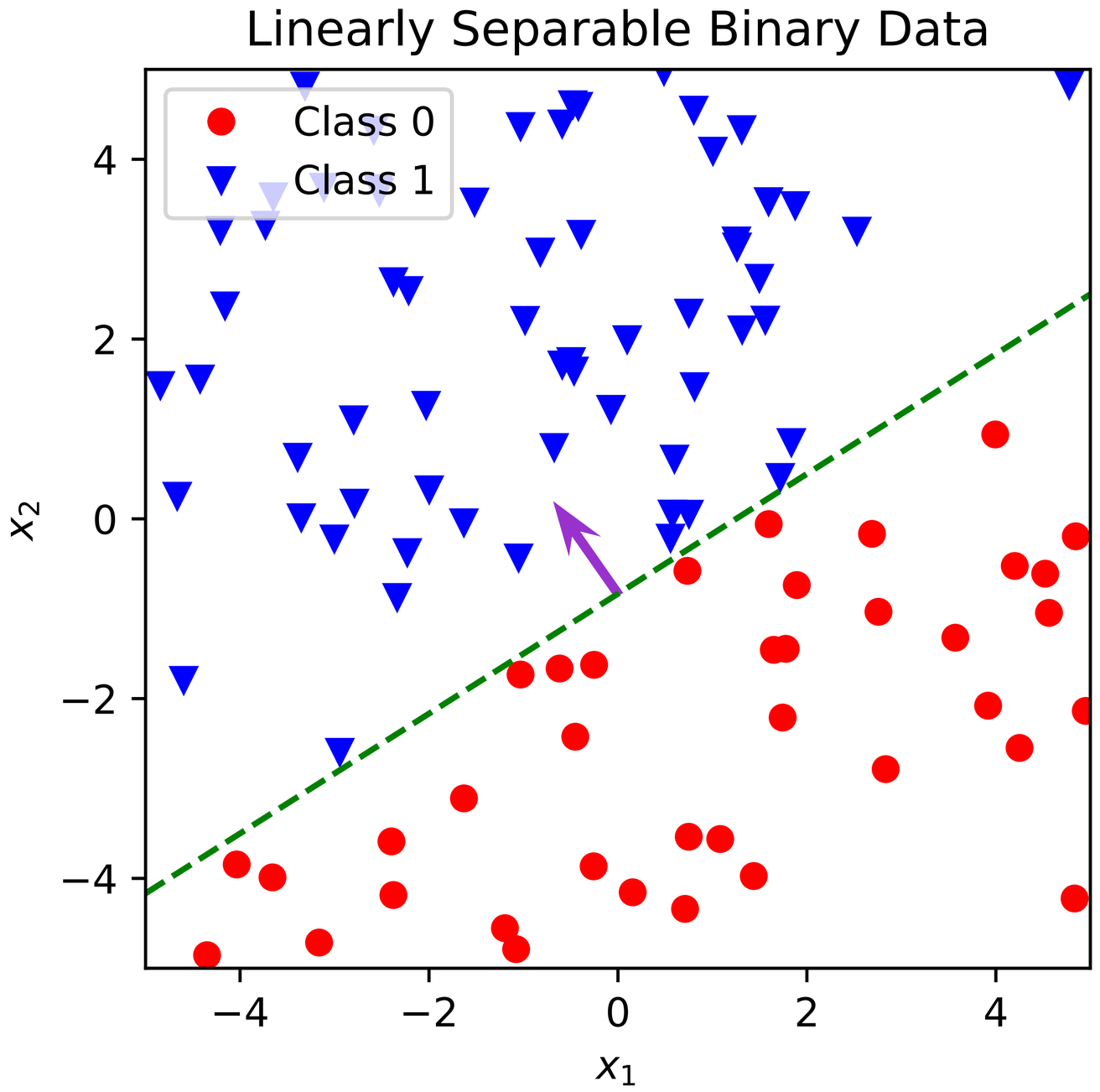
$$\Longrightarrow \mathbf{x} \perp \mathbf{w}$$


Linearly Separable Binary Data
Class 0
Class 1
4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


Cosine
1.0
0.5
0.0
-0.5
-1.0
0
1
2
3
4
5
6
Angle

$$\hat{y} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$


b
θ
a


Linearly Separable Binary Data
Class 0
Class 1
4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

$$\mathbf{x} \cdot \mathbf{w}$$

$$\frac{\mathbf{x} \cdot \mathbf{w}}{\|\mathbf{w}\|}$$

$$y \in \{0,1\}$$

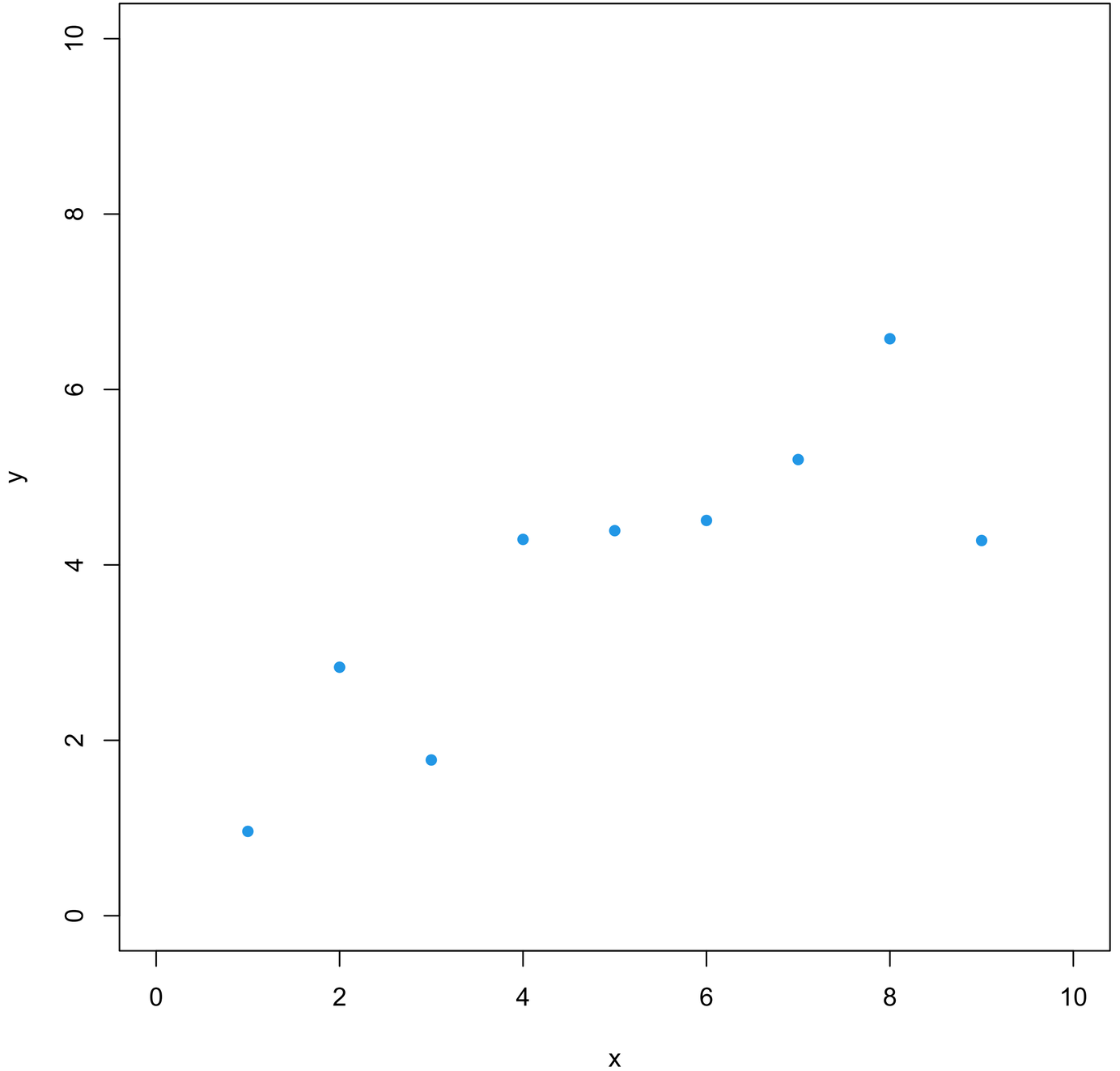
$$y \in \{-1,1\}$$

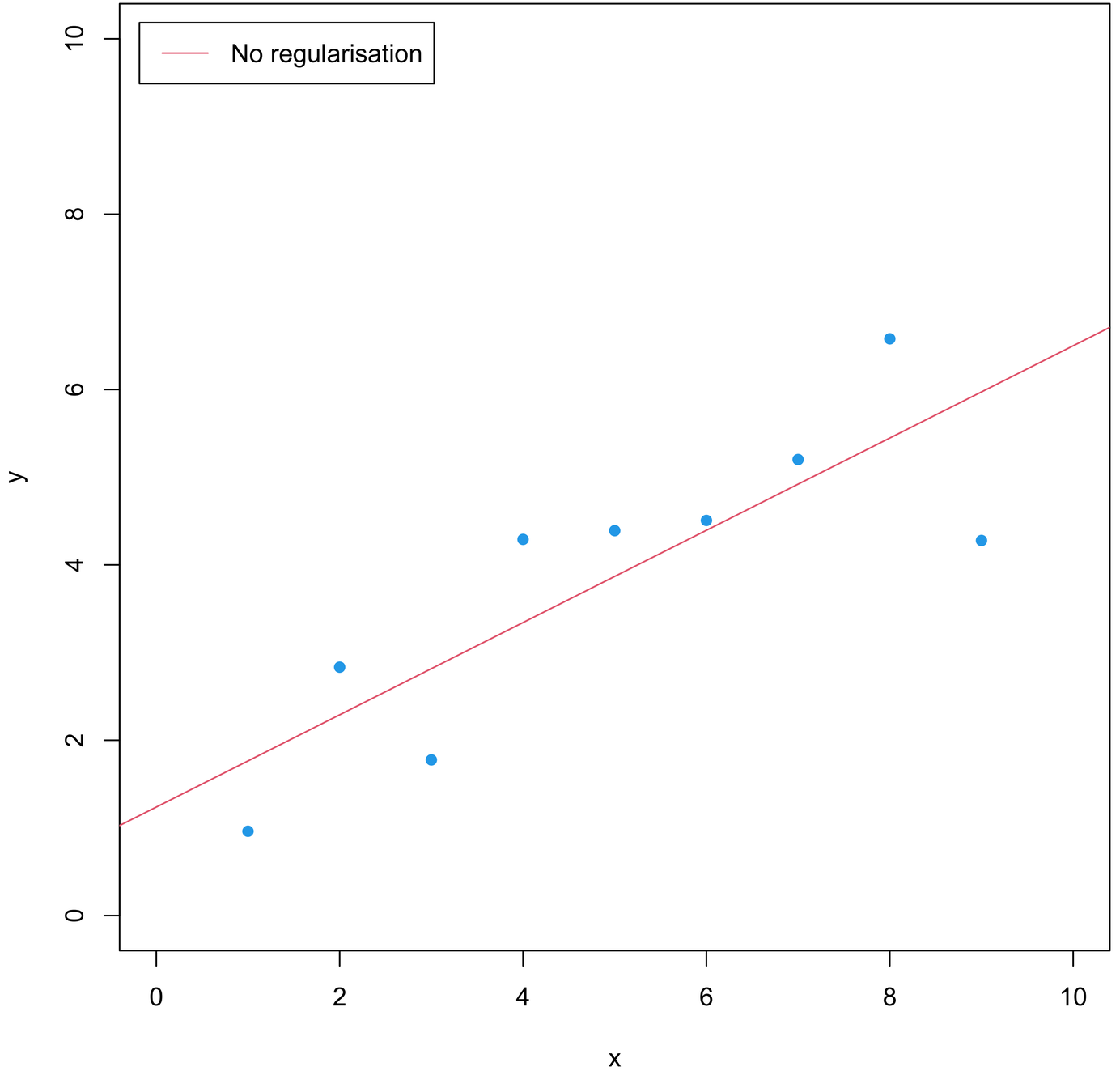
Functional Margin

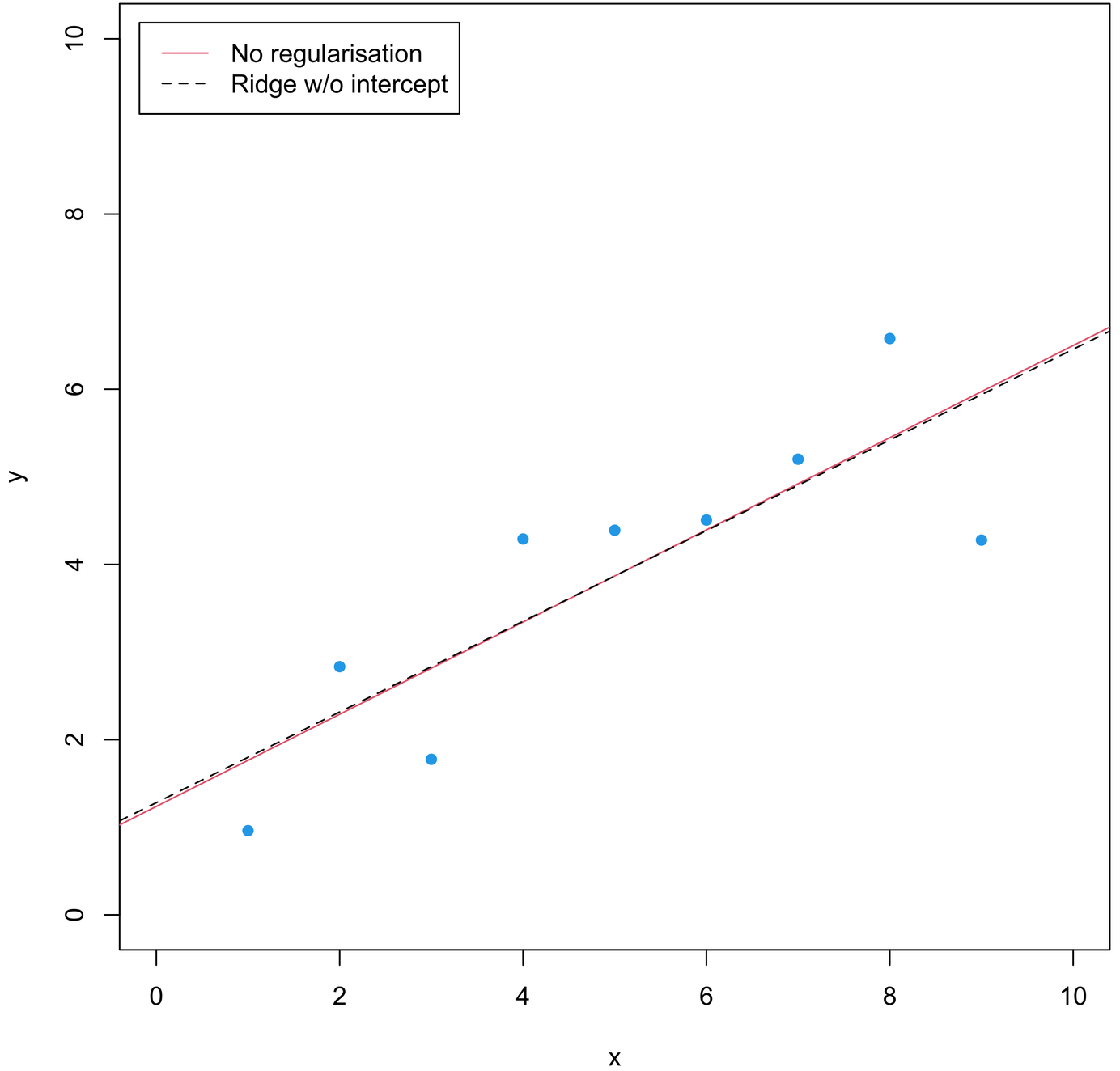
$$y\hat{y} = yf(\mathbf{x}) = y\mathbf{x} \cdot \mathbf{w}$$

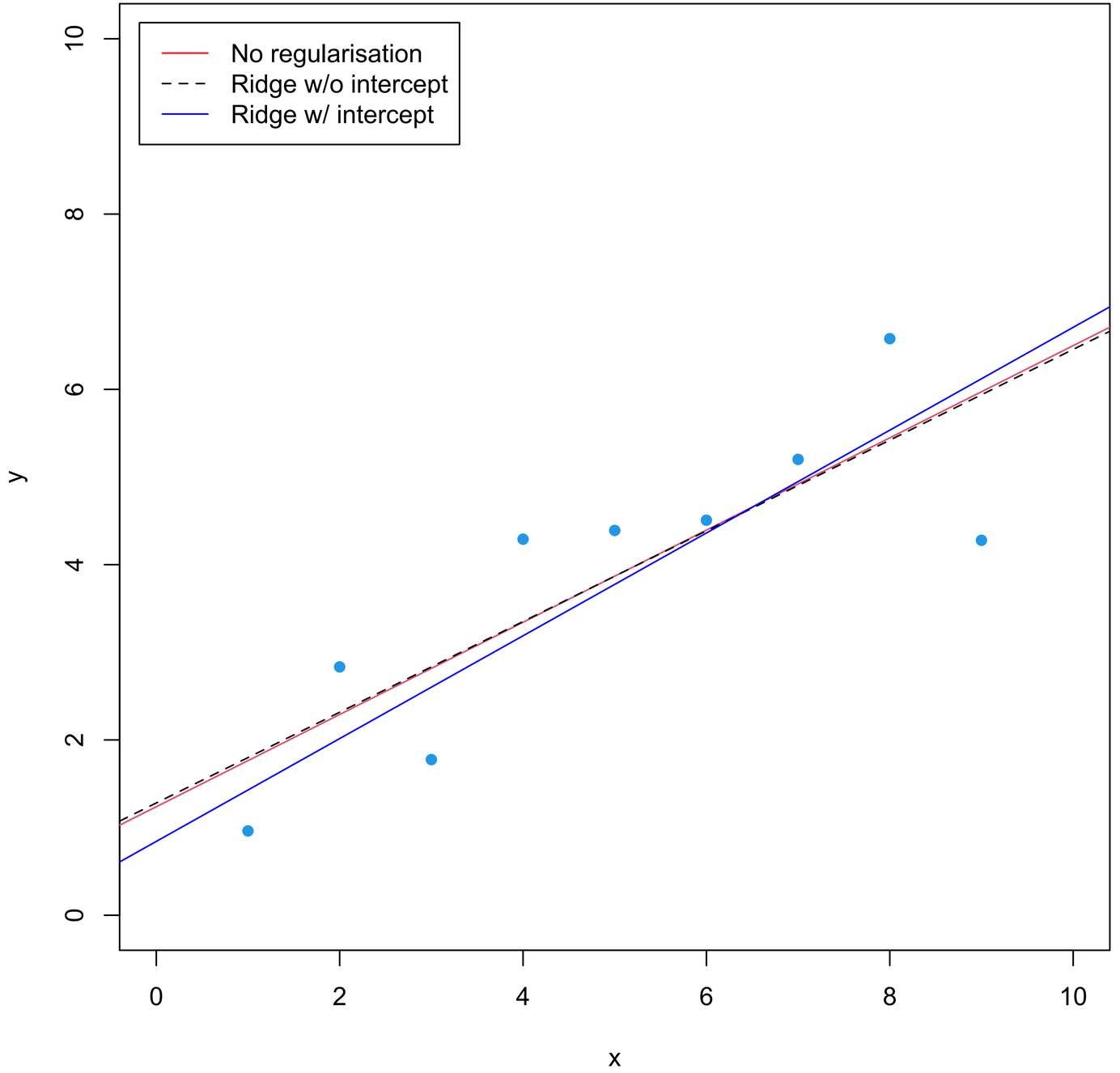
Geometric Margin

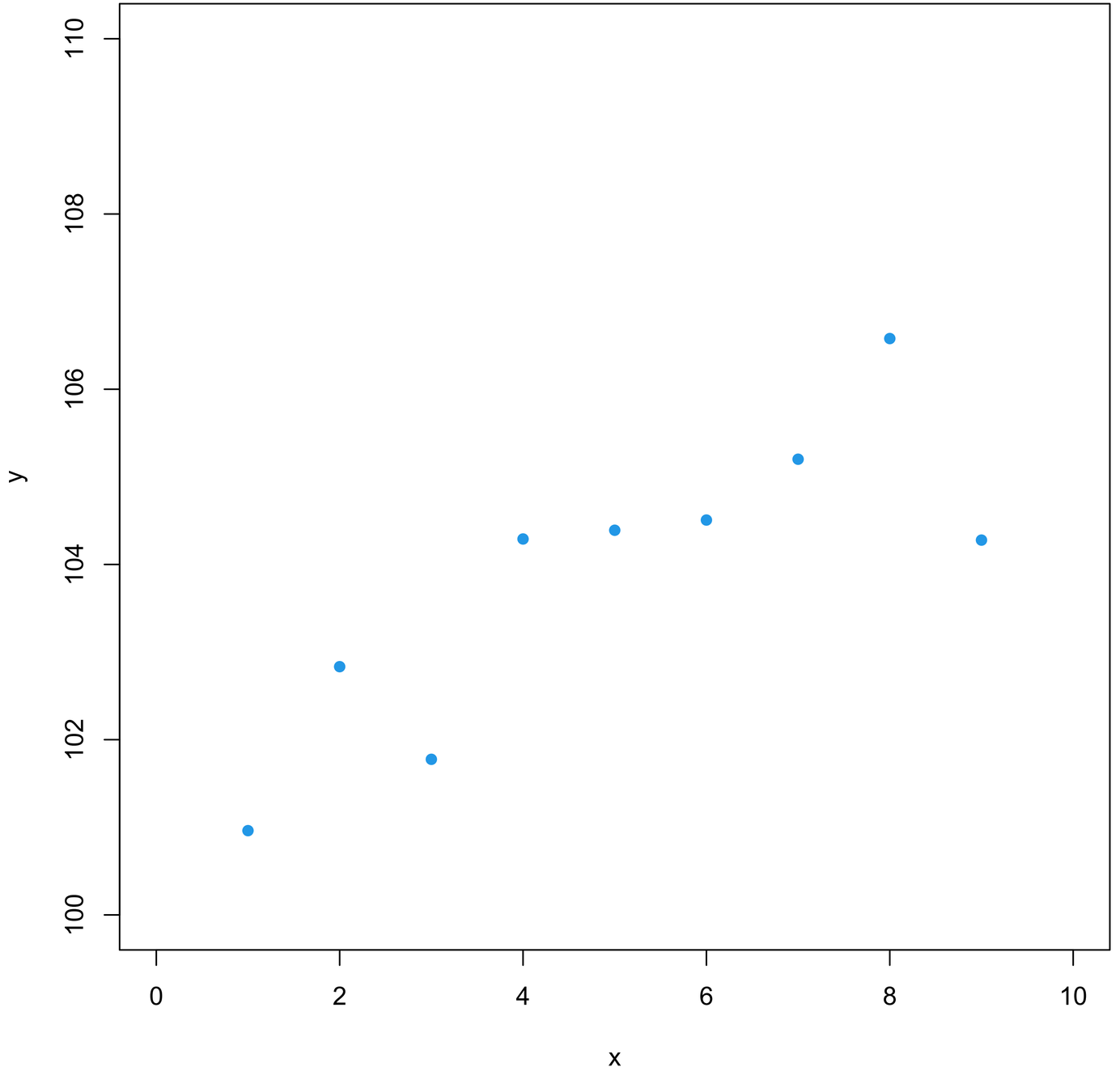
$$\frac{y\,\mathbf{x}\cdot\mathbf{w}}{\|\mathbf{w}\|}$$

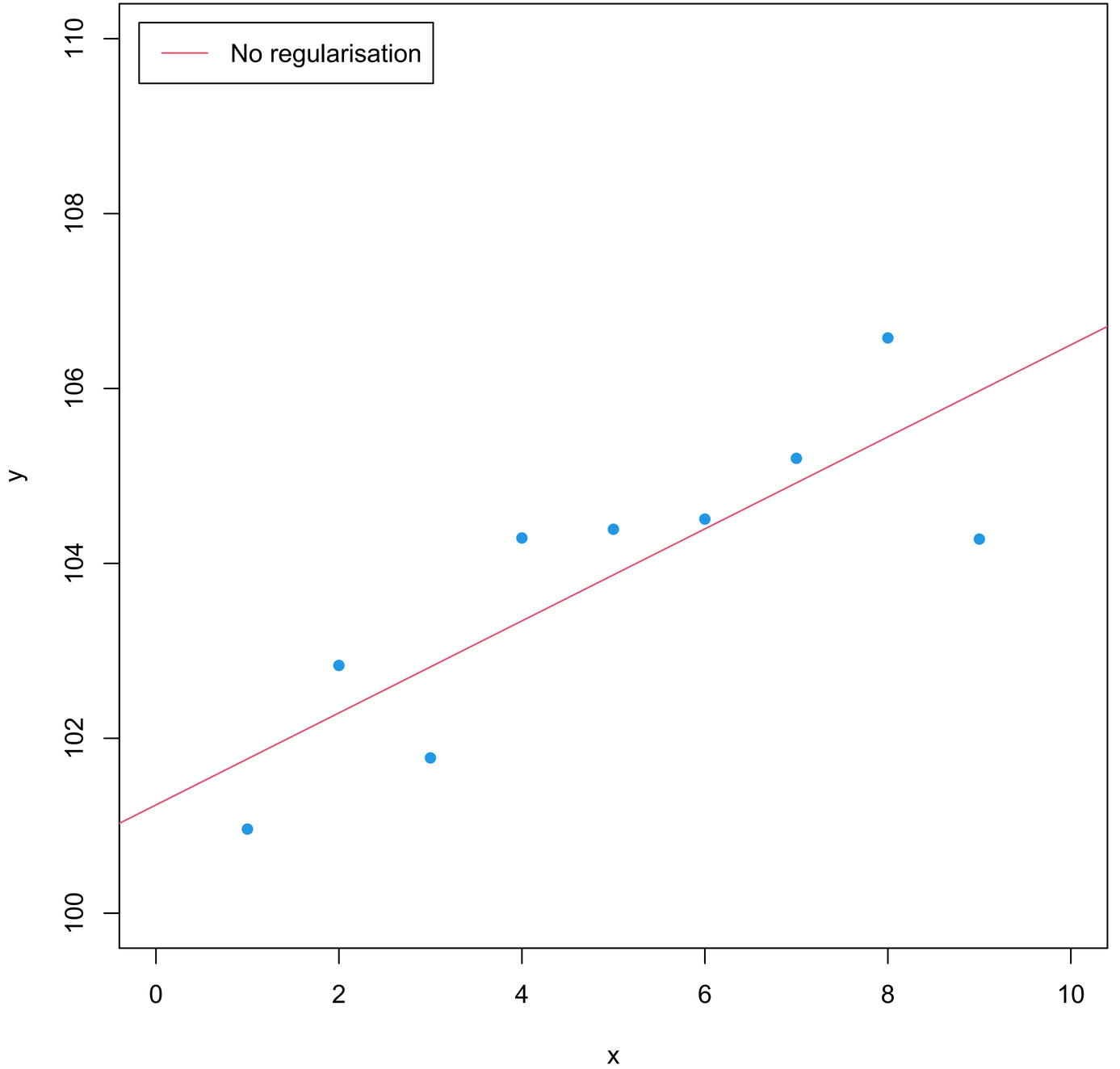

y
x
0
2
4
6
8
10

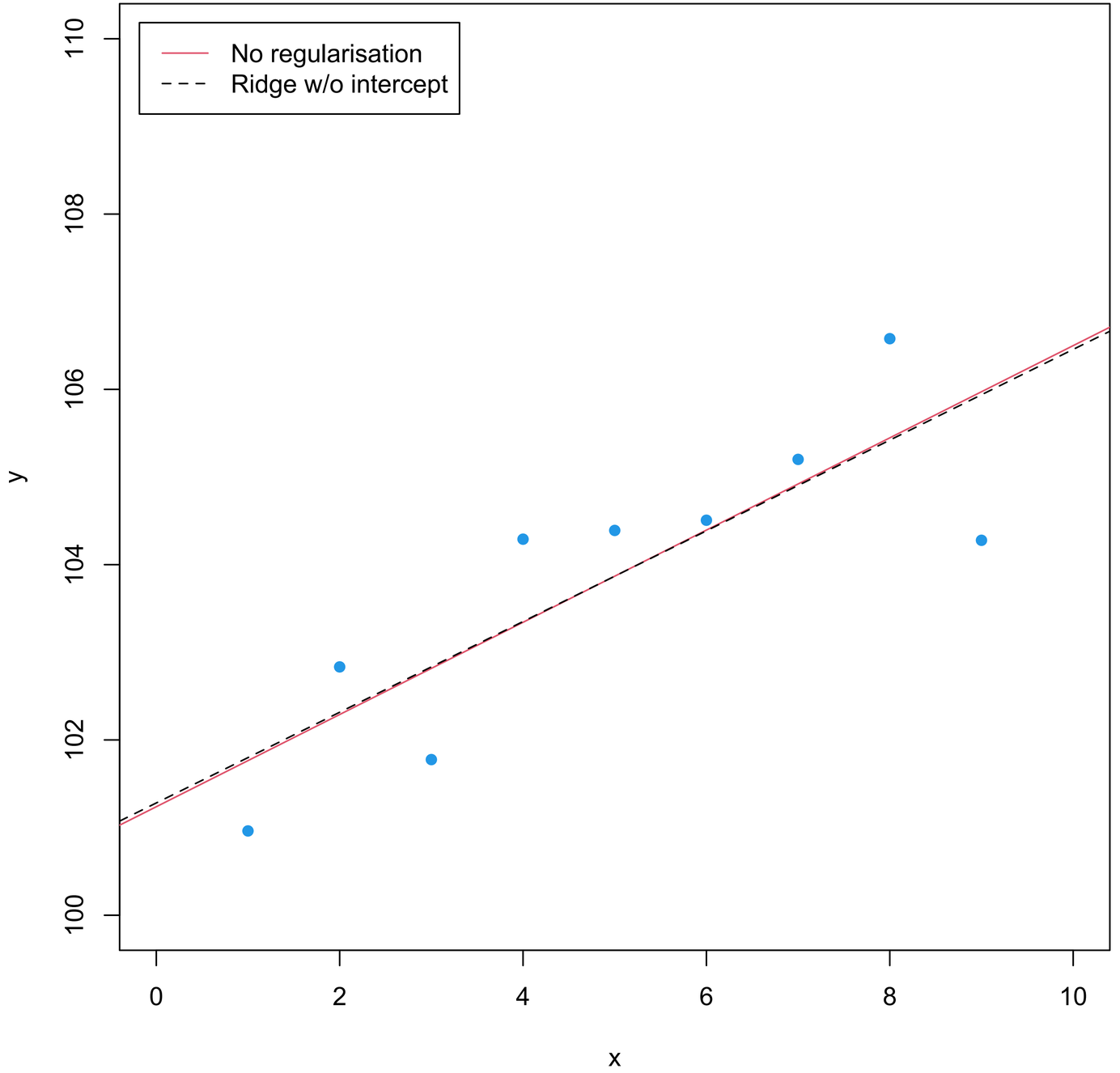

No regularisation
y
x
0
2
4
6
8
10

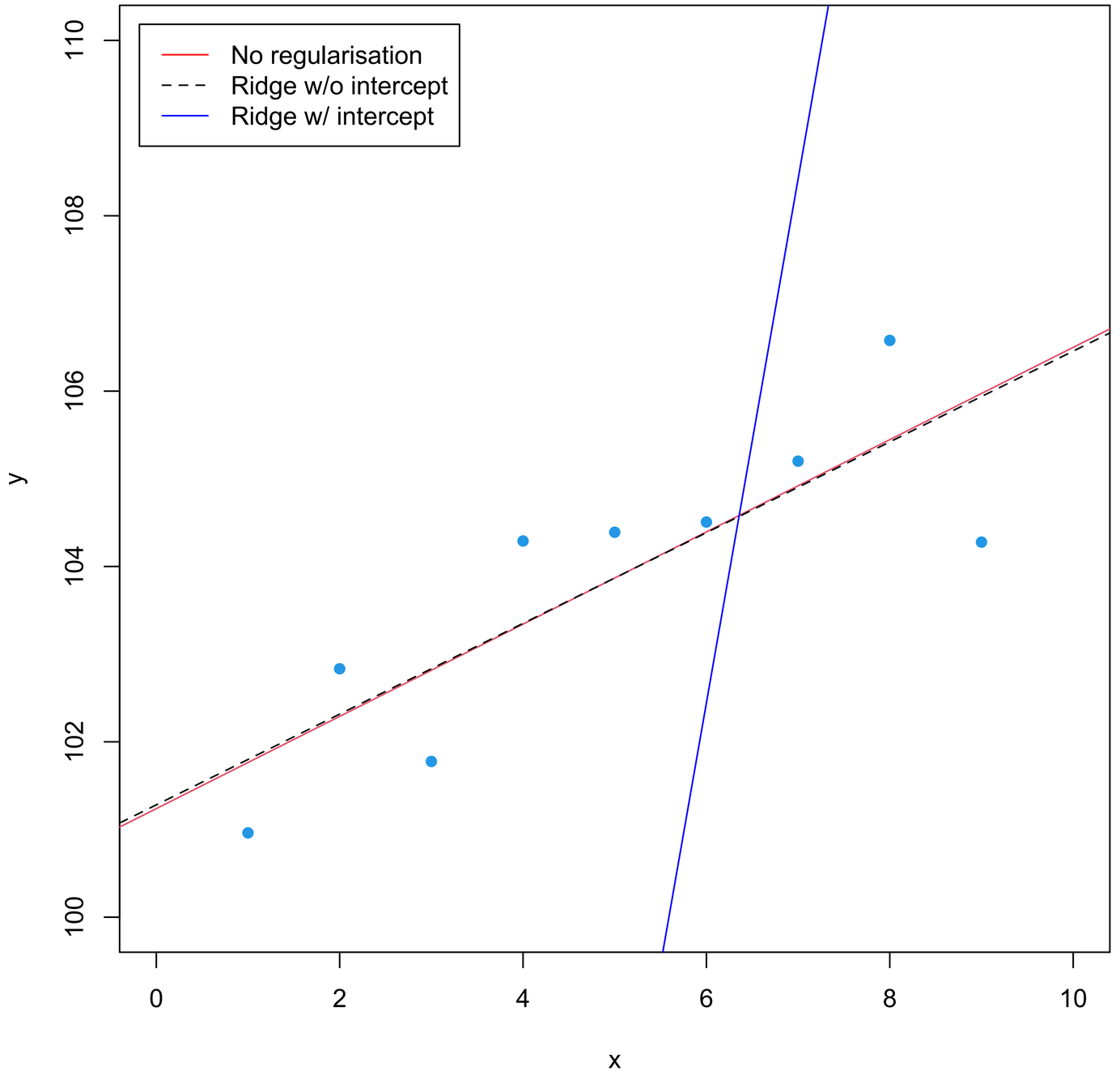

No regularisation
Ridge w/o intercept
y
x
0
2
4
6
8
10

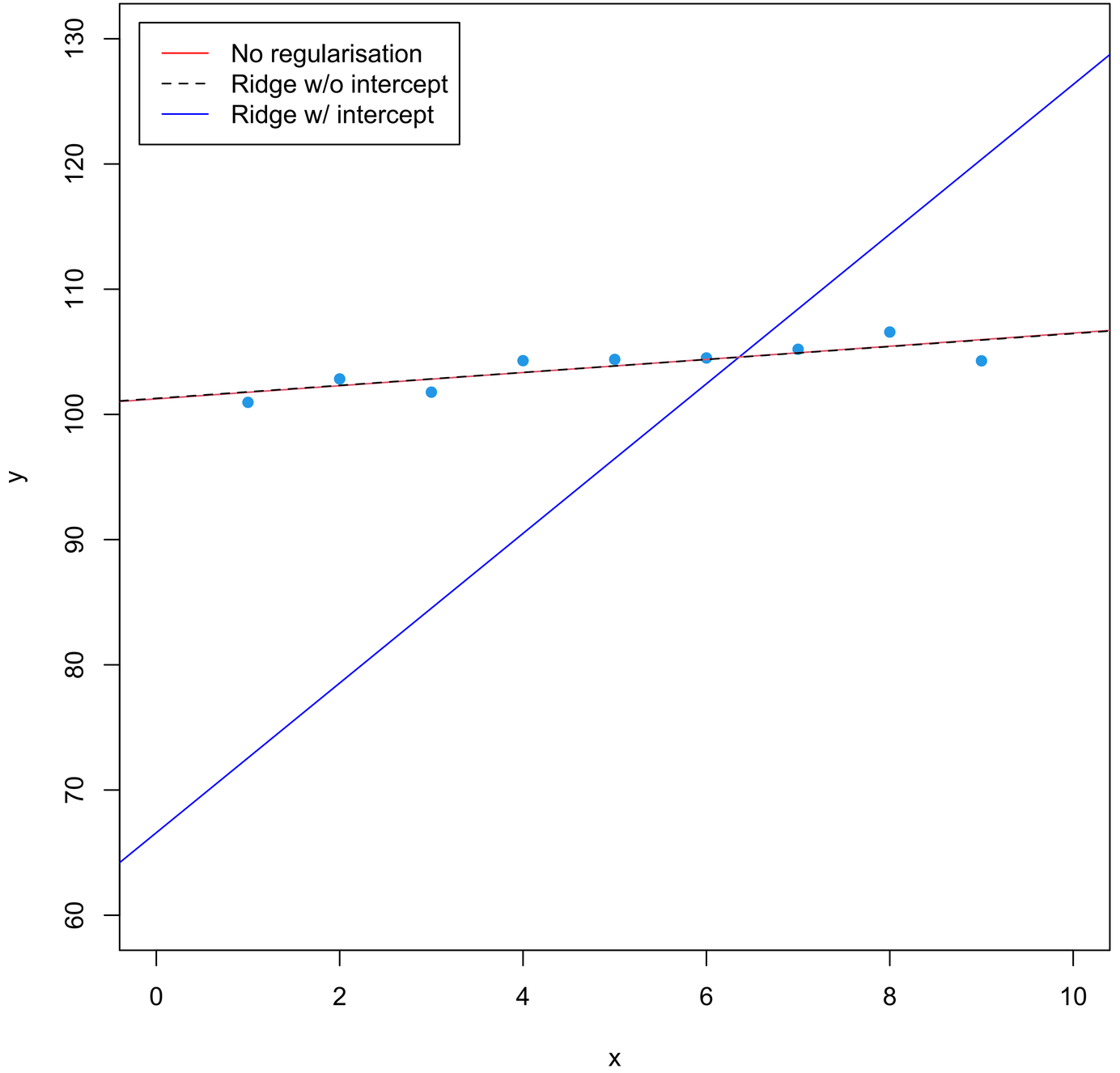

No regularisation
Ridge w/o intercept
Ridge w/ intercept
y
x
0
2
4
6
8
10


y
110
108
106
104
102
100
0
2
4
6
8
10
x

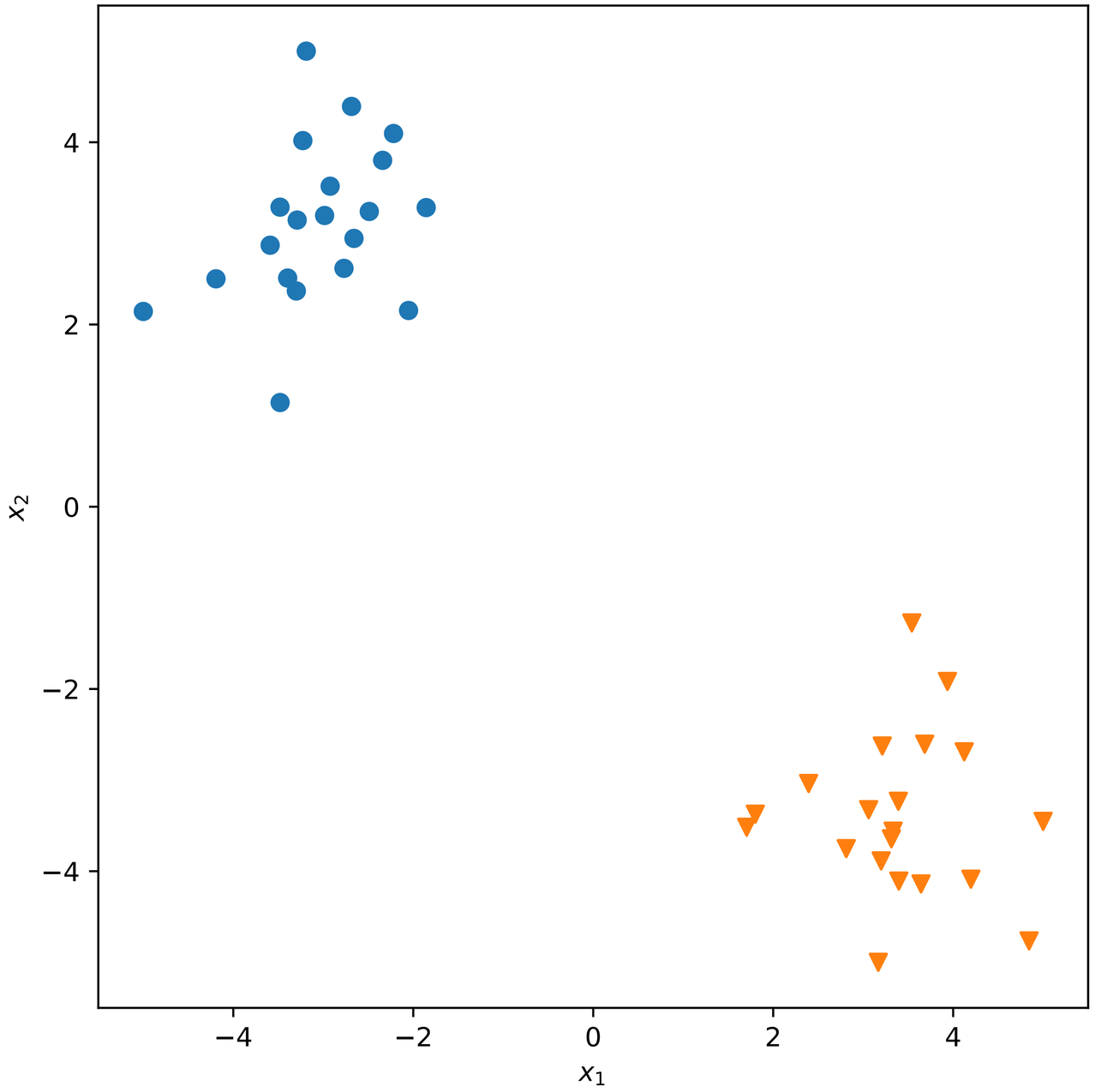

No regularisation
y
x
100
102
104
106
108
110
0
2
4
6
8
10

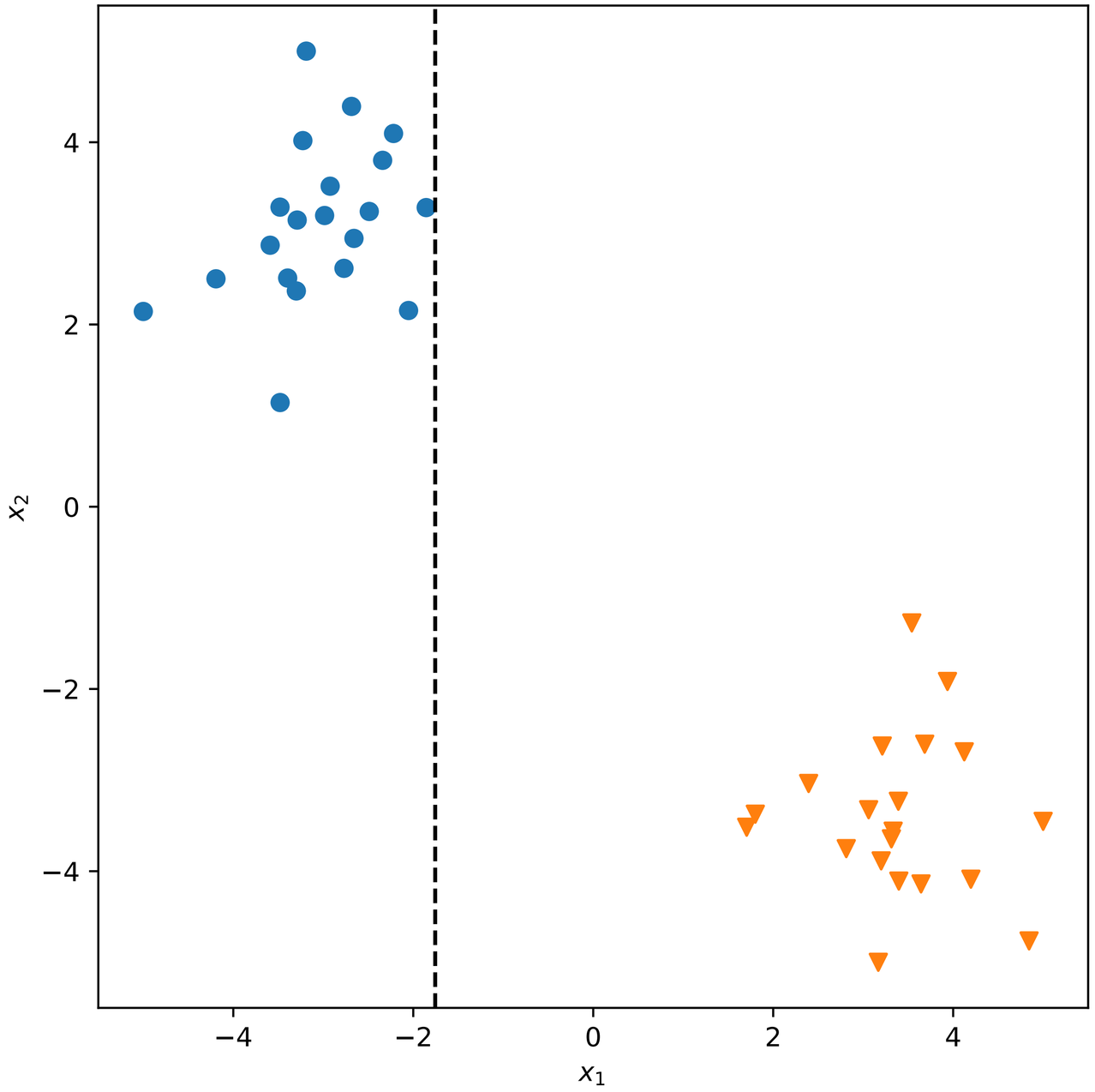

No regularisation
Ridge w/o intercept
y
x
100
102
104
106
108
110
0
2
4
6
8
10

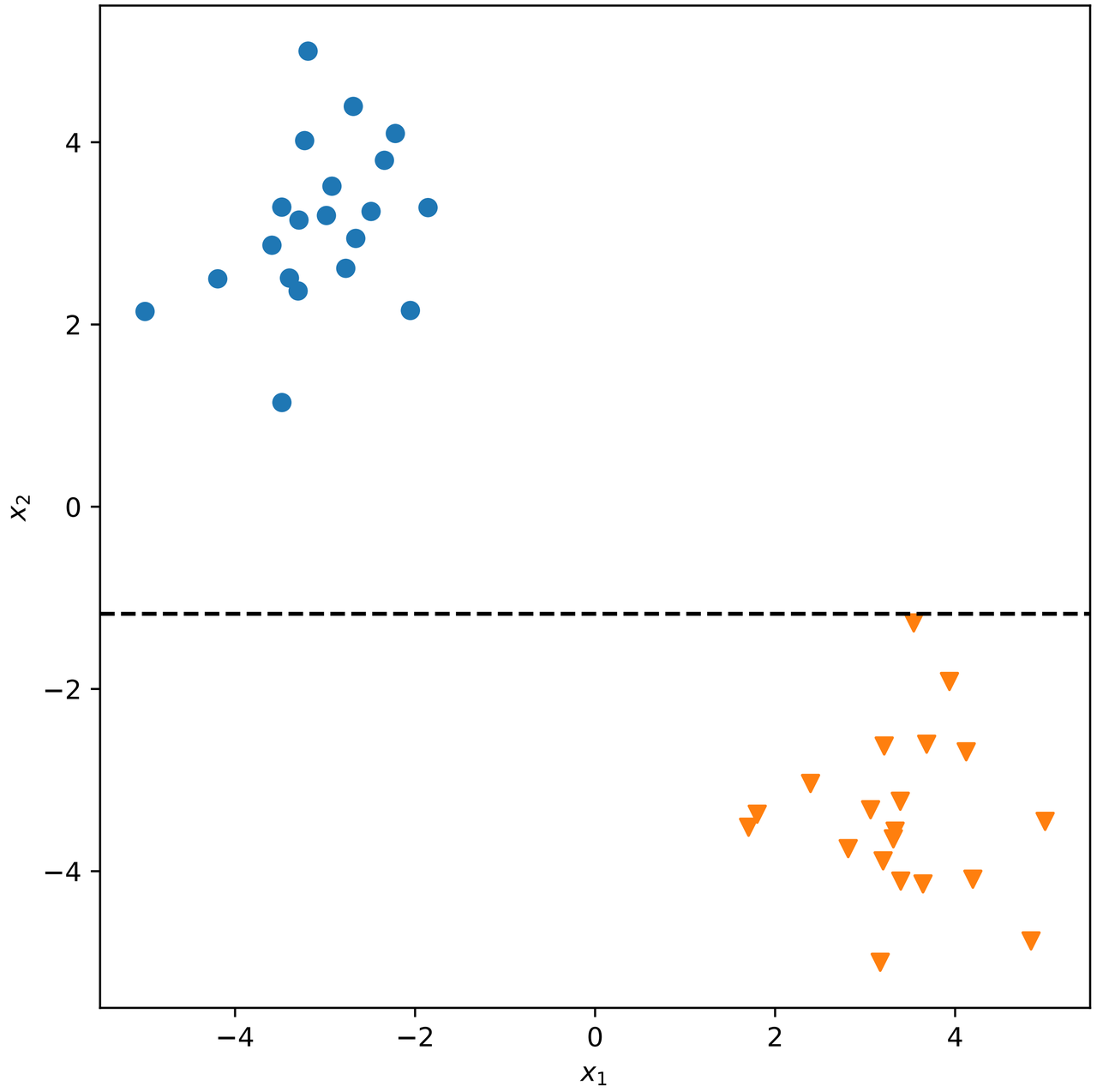

No regularisation
Ridge w/o intercept
Ridge w/ intercept
y
x
100
102
104
106
108
110
0
2
4
6
8
10

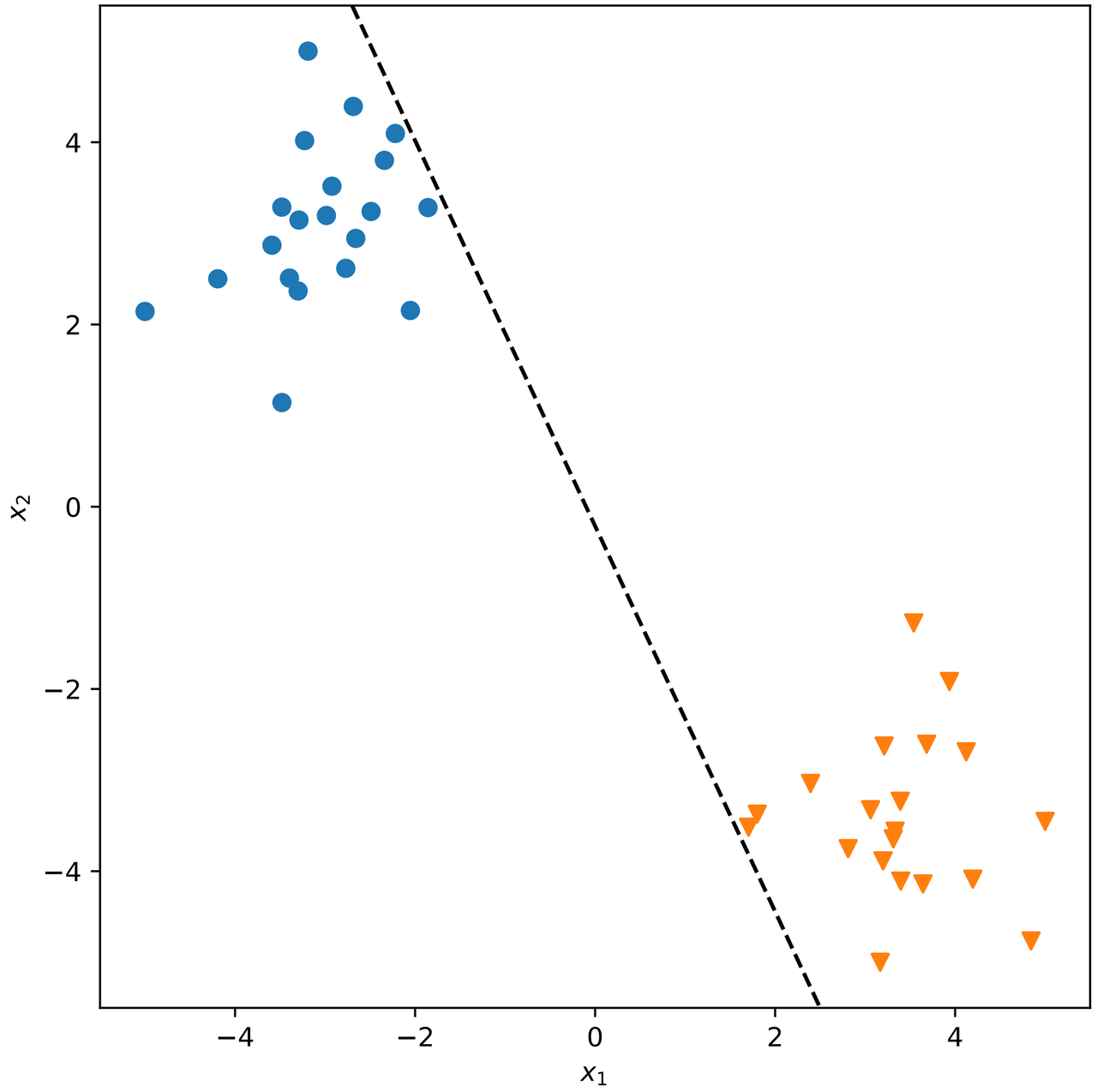

No regularisation
Ridge w/o intercept
Ridge w/ intercept
y
x
130
120
110
100
90
80
70
60
0
2
4
6
8
10

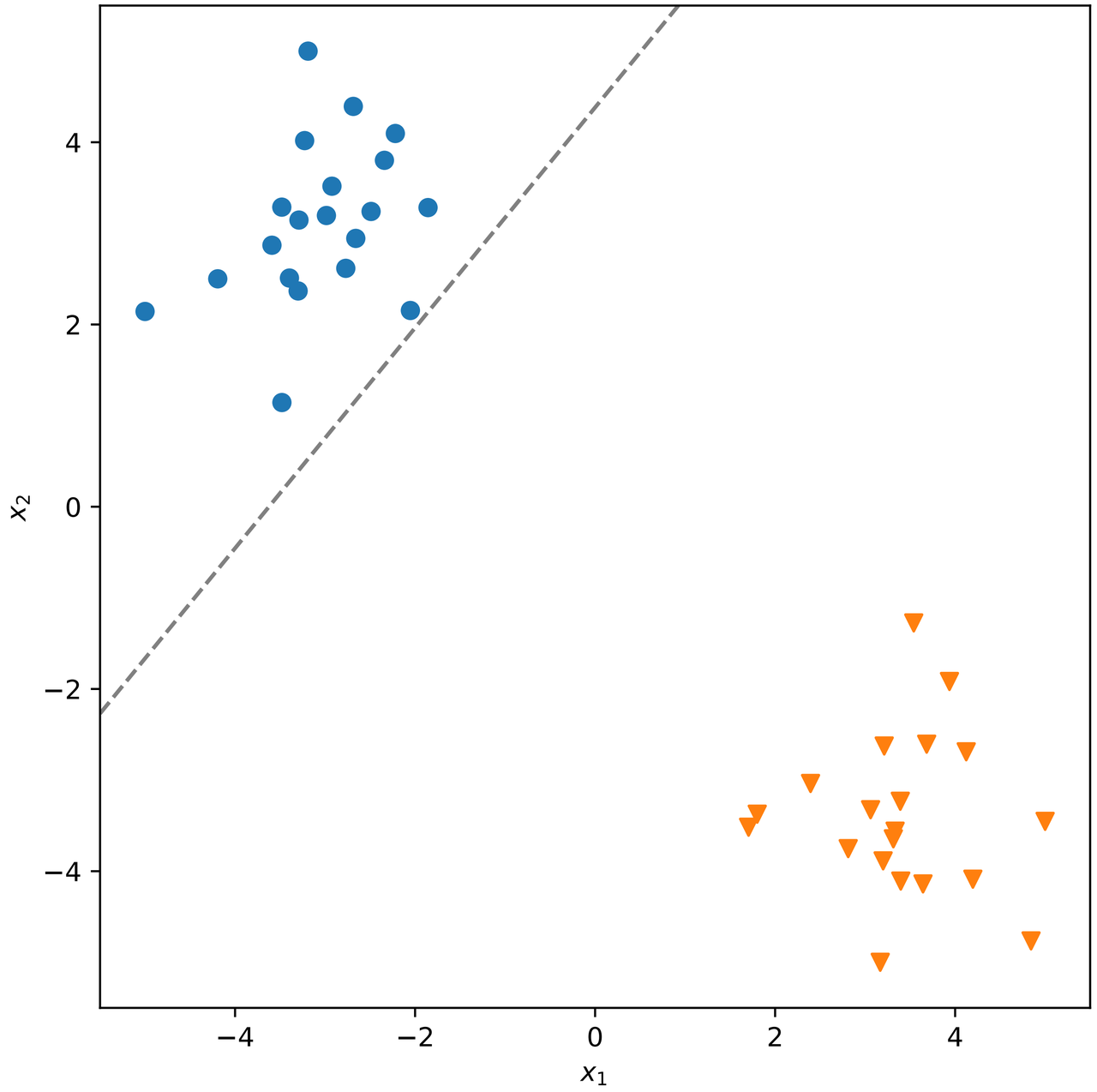
$$\hat{y} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

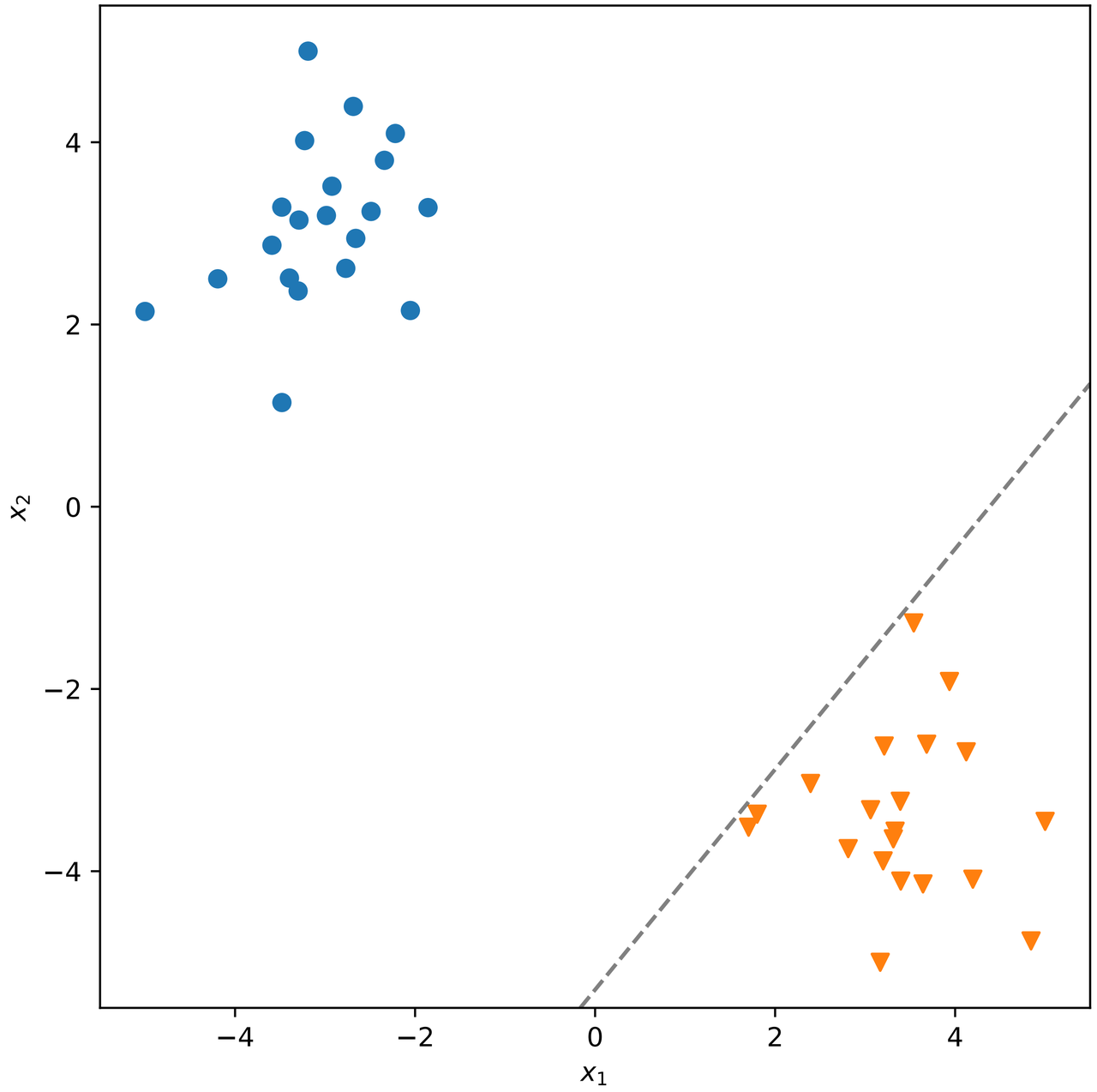

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

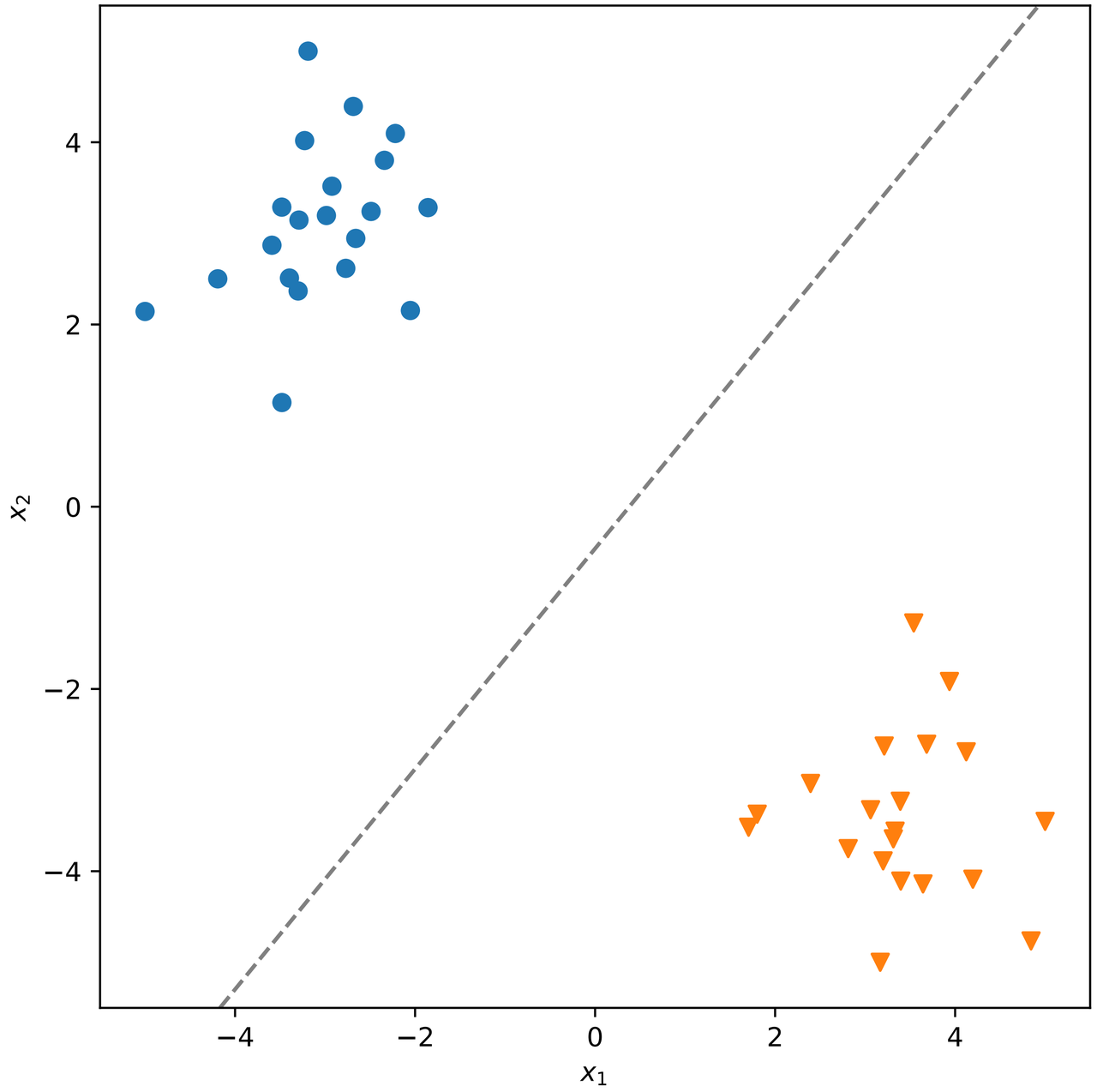

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

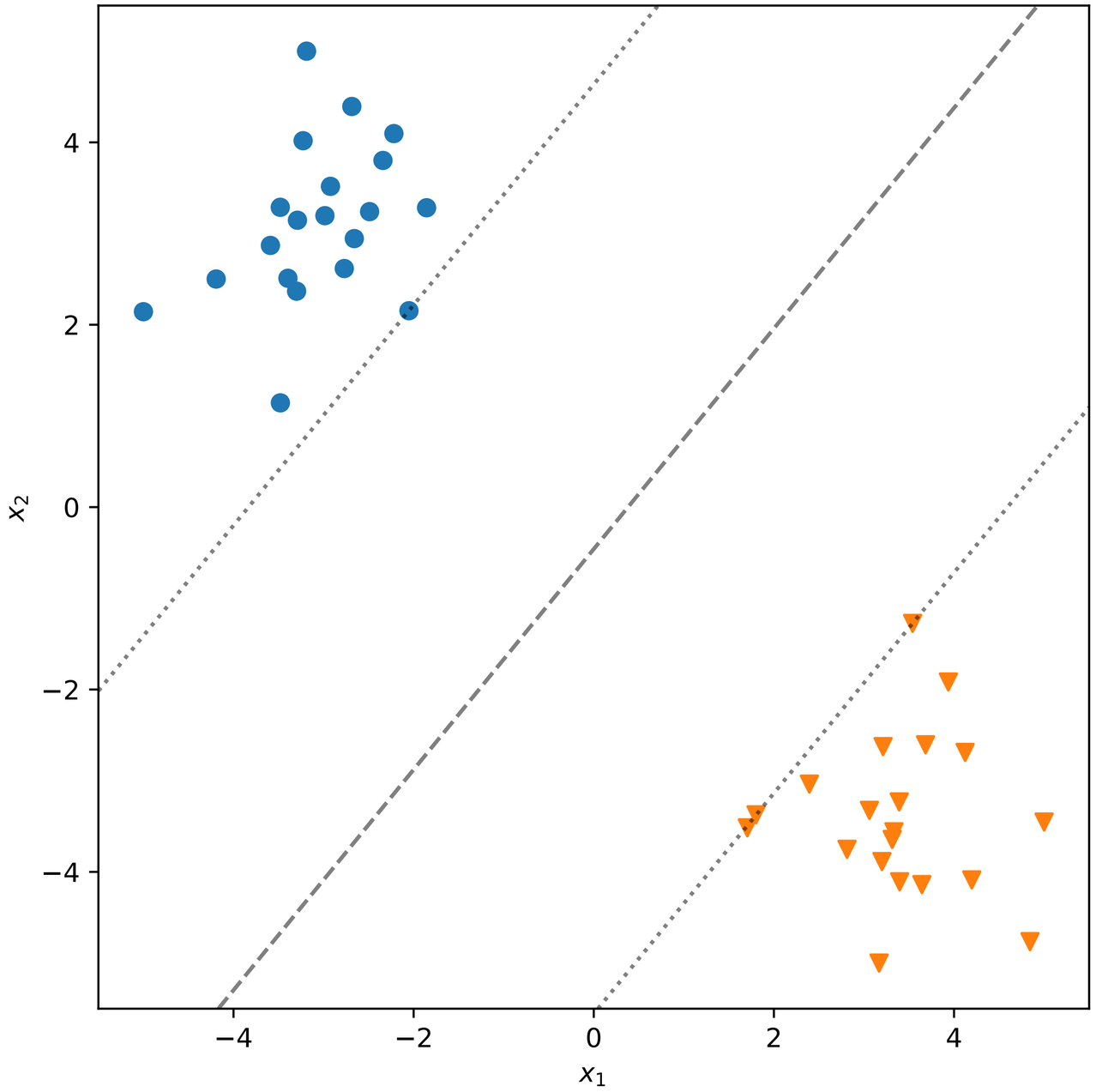

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

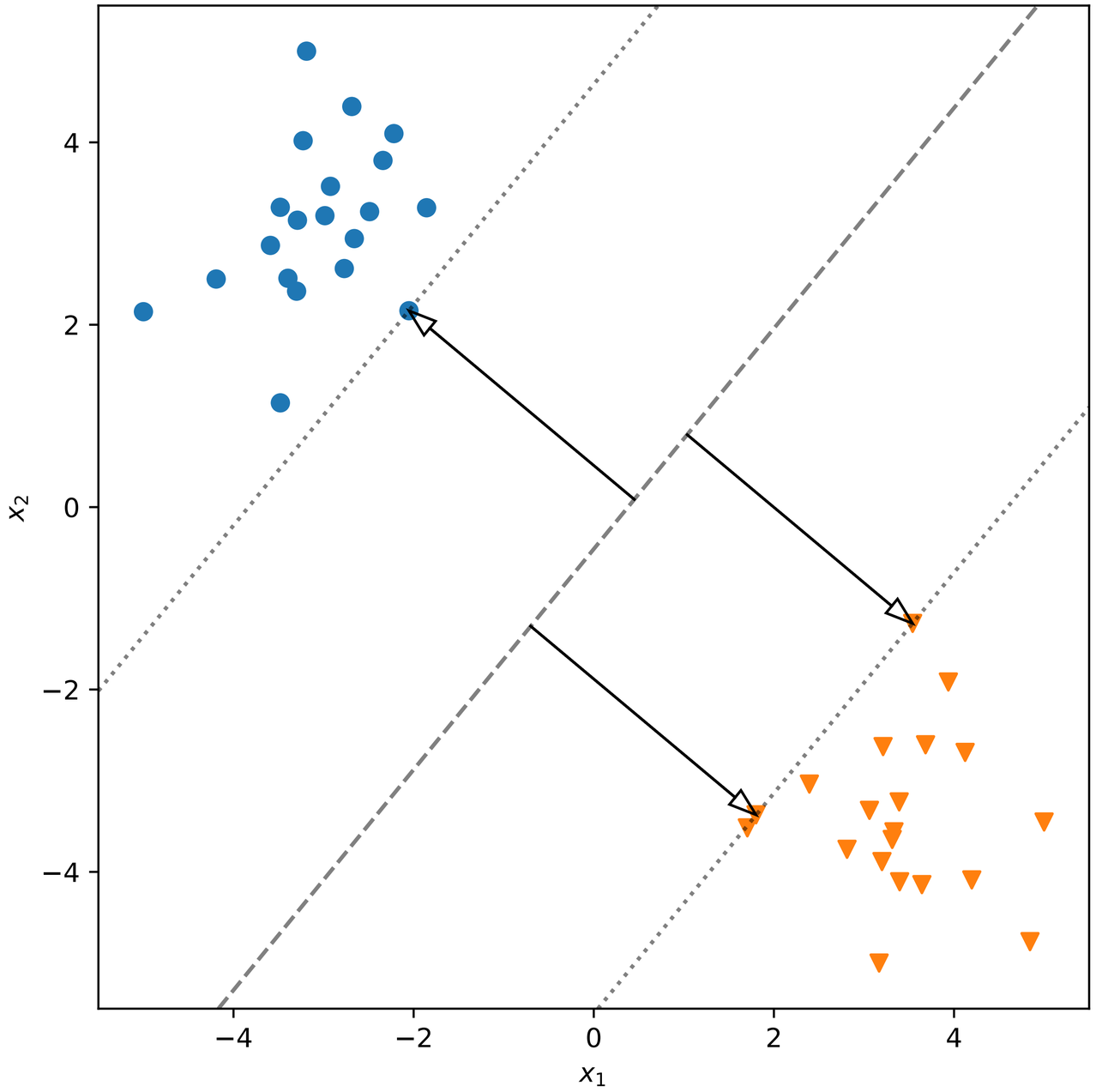

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

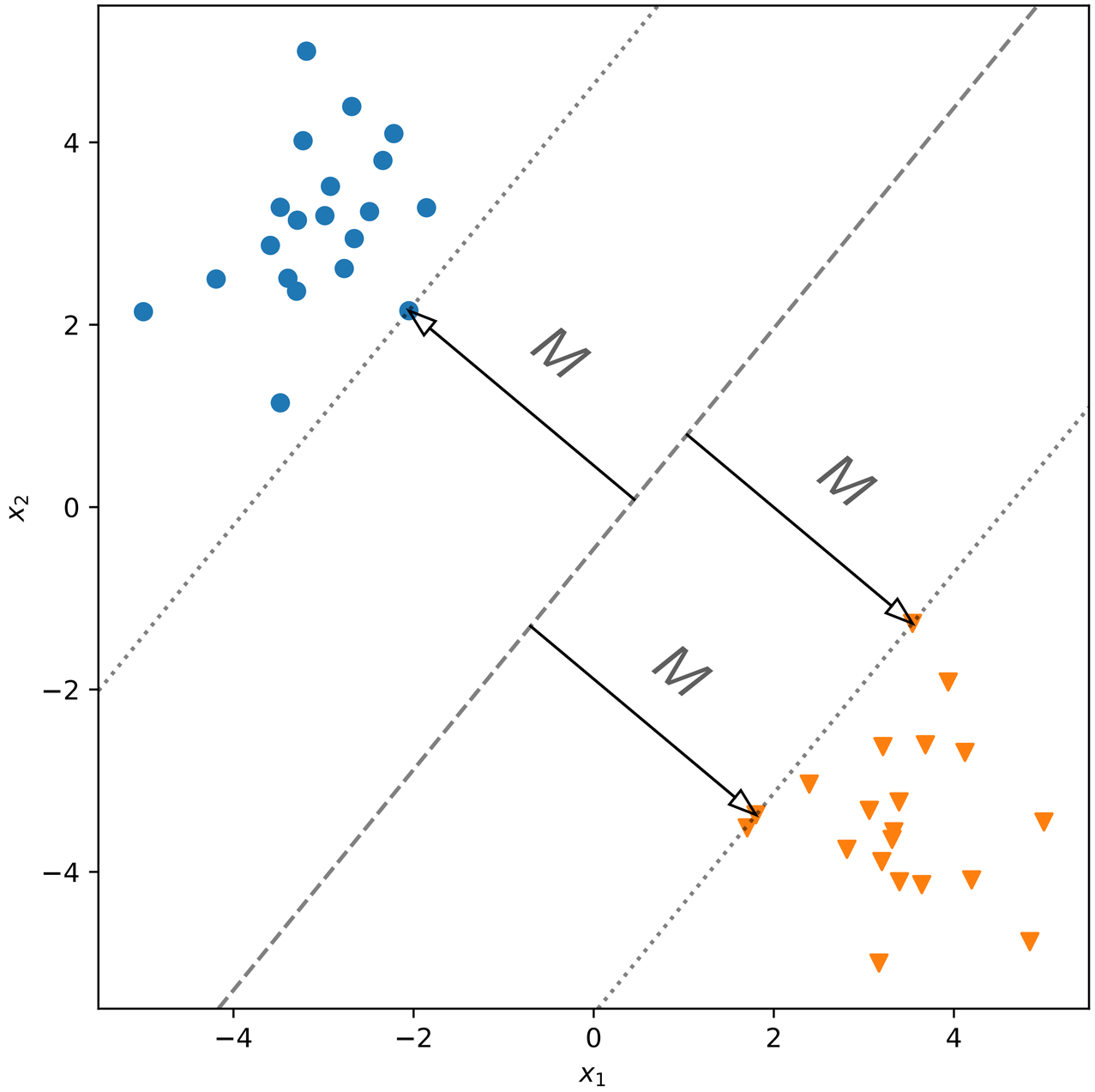

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1
M
M
M

$$M = yf(\mathbf{x}) = y(\mathbf{x} \cdot \mathbf{w} + b)$$

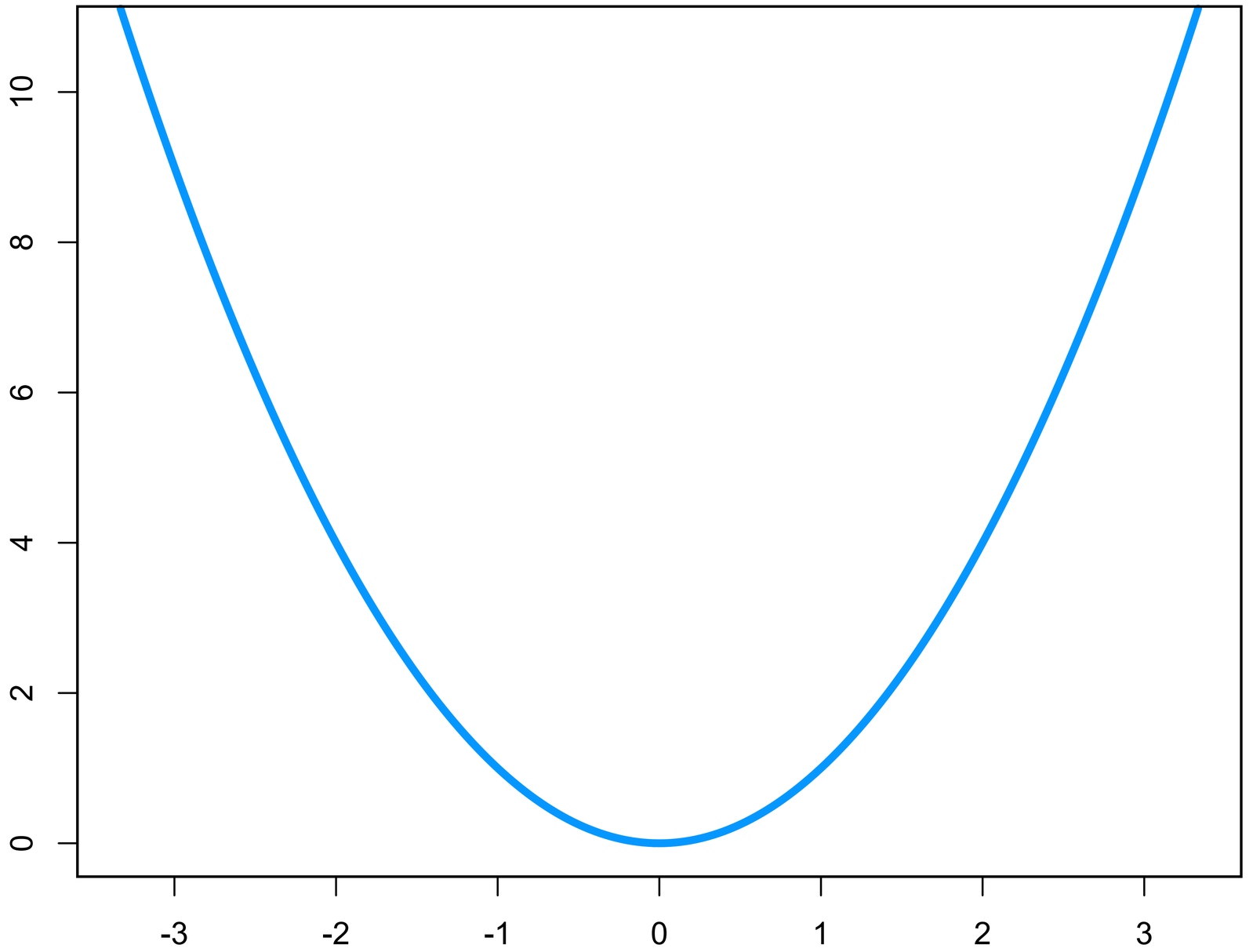
$$\|\mathbf{w}\| = 1$$

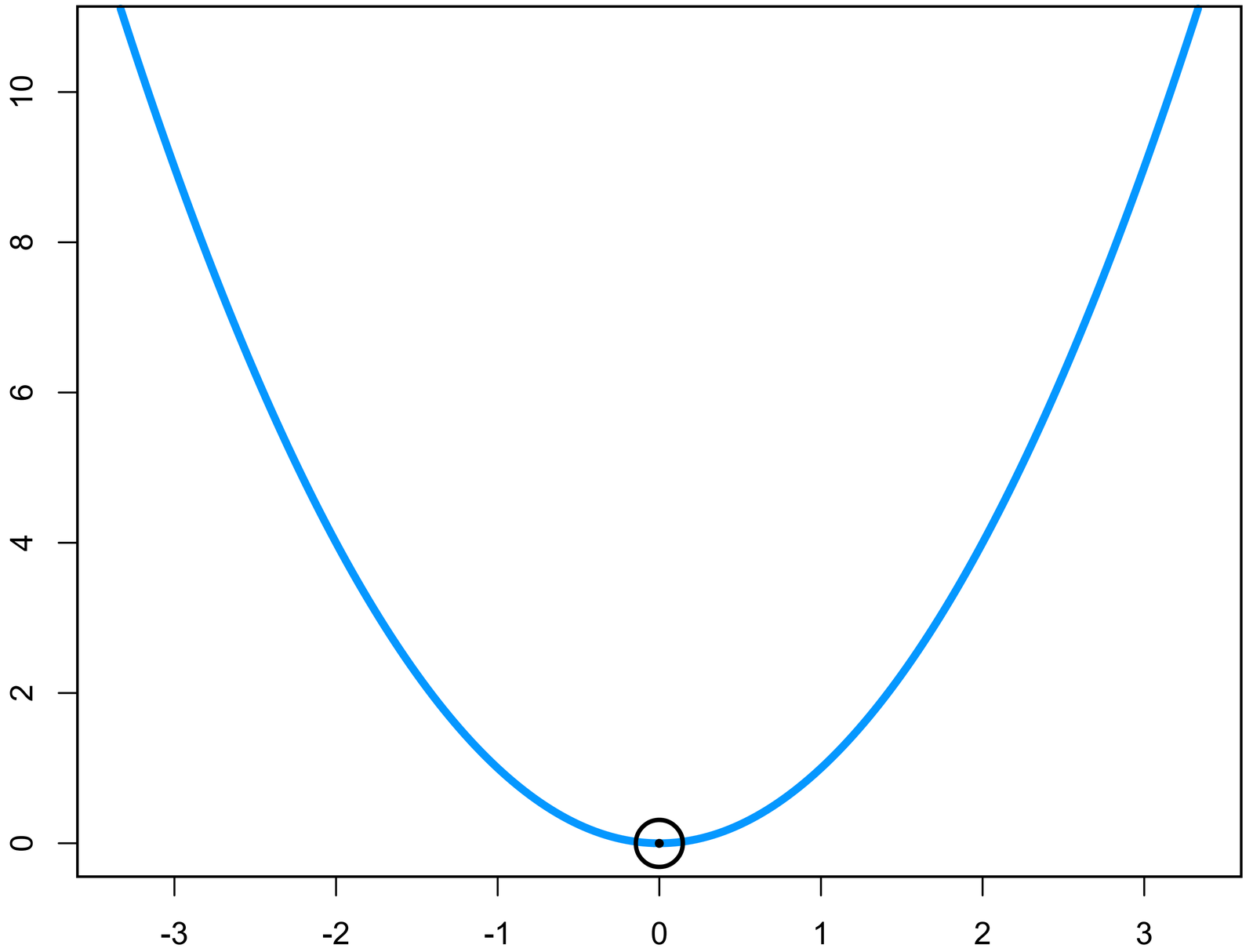
$$\underset{\mathbf{w},b}{\text{maximise}} \;\; M$$

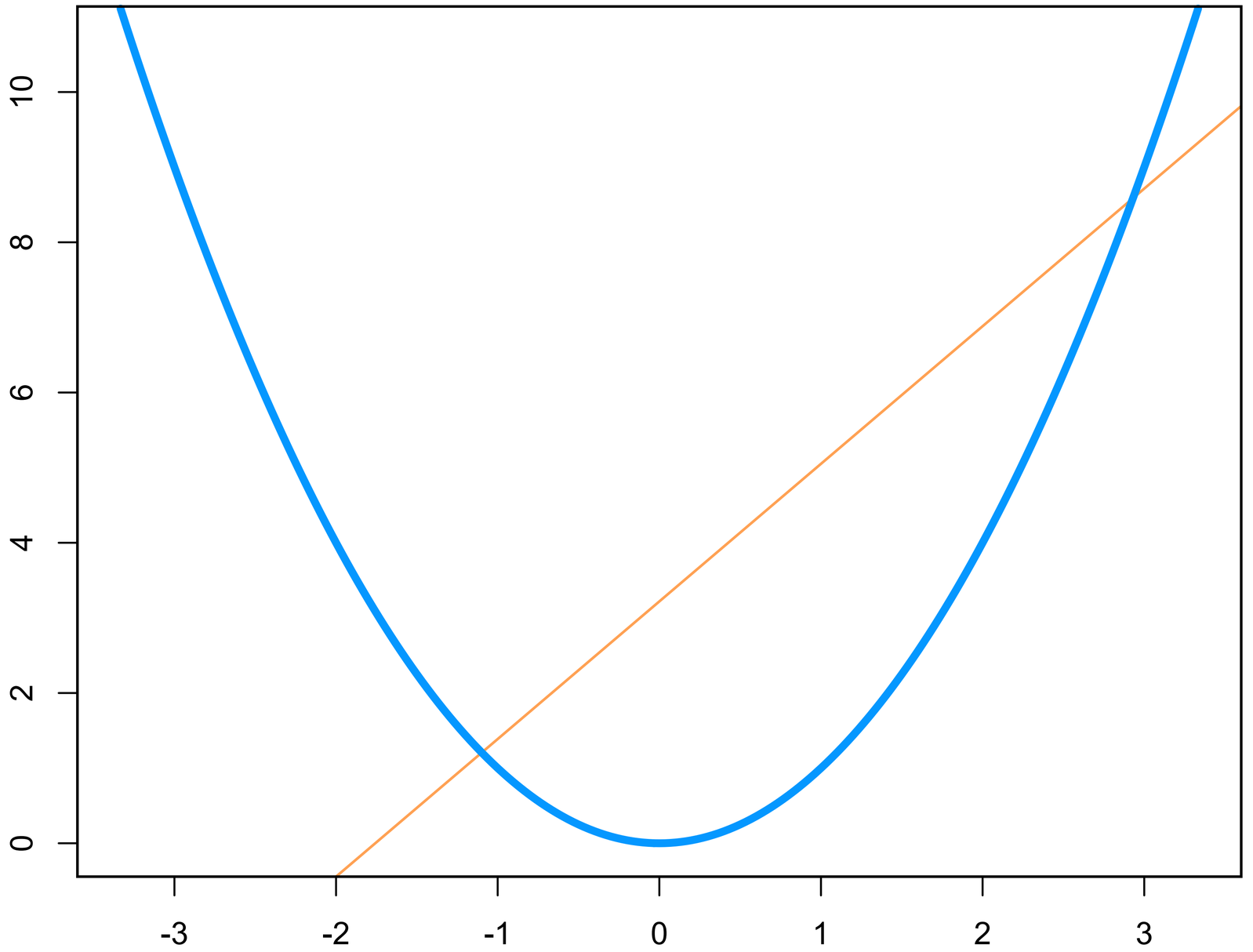
$$\text{subject to:}$$

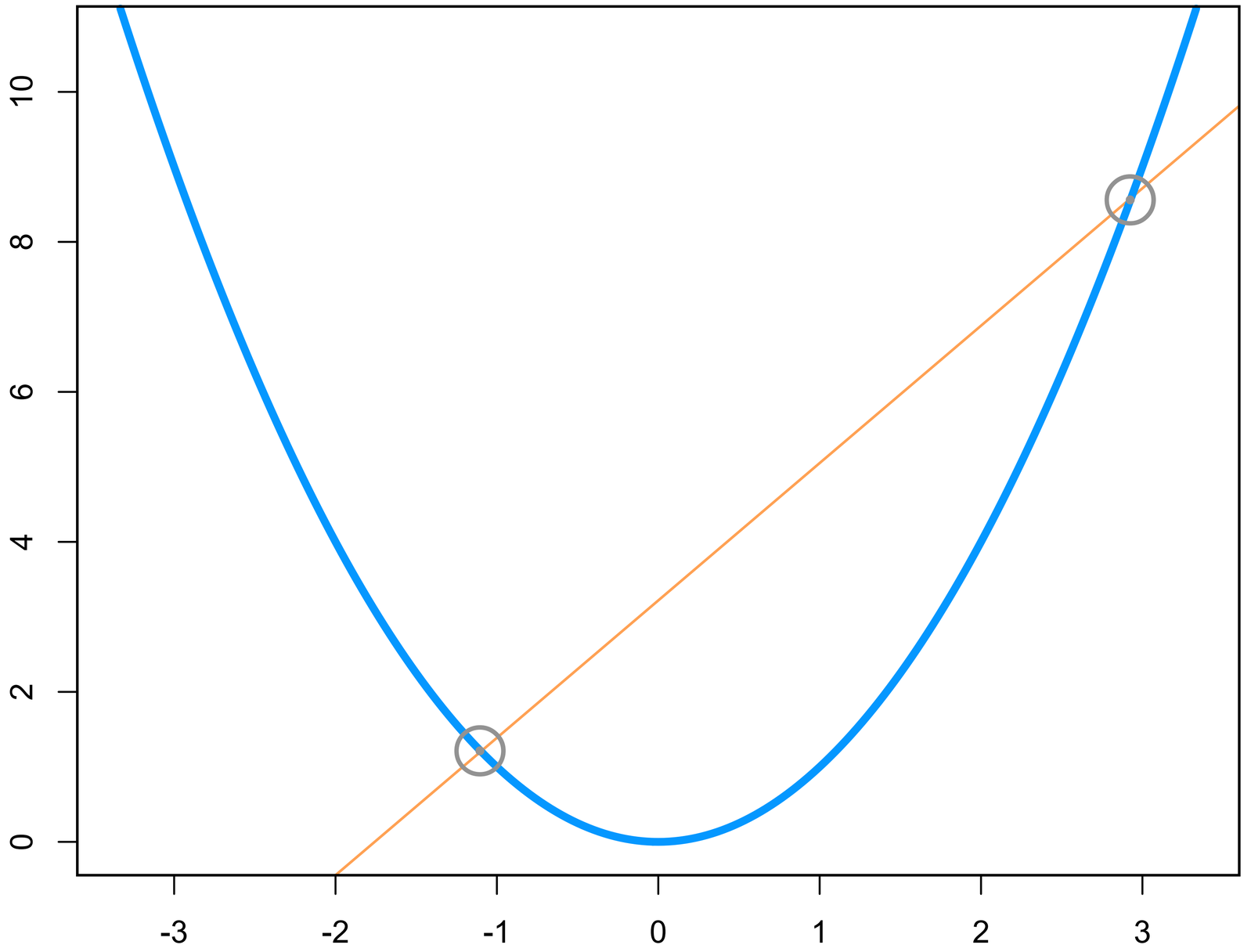
$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq M$$

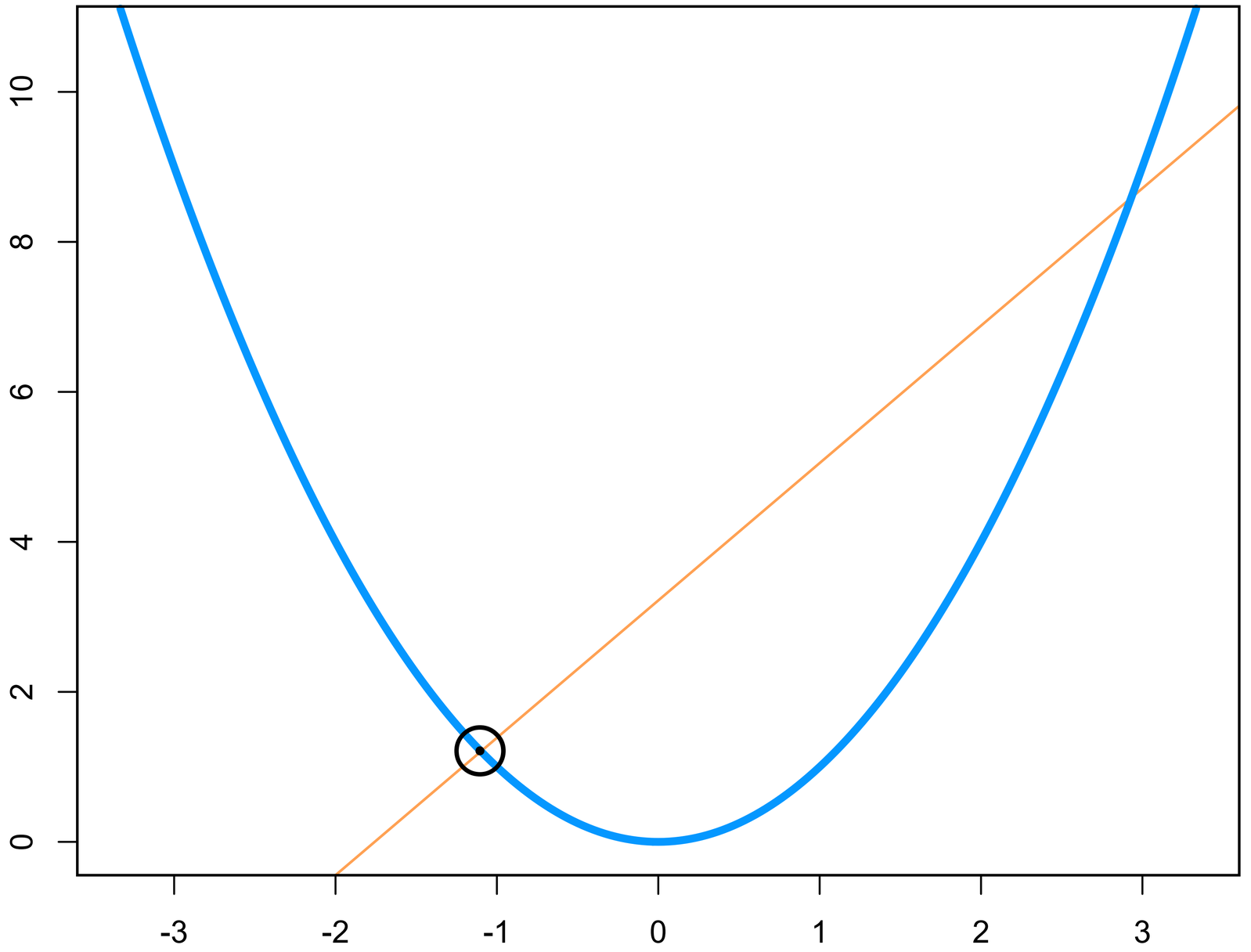
$$\|\mathbf{w}\| = 1$$

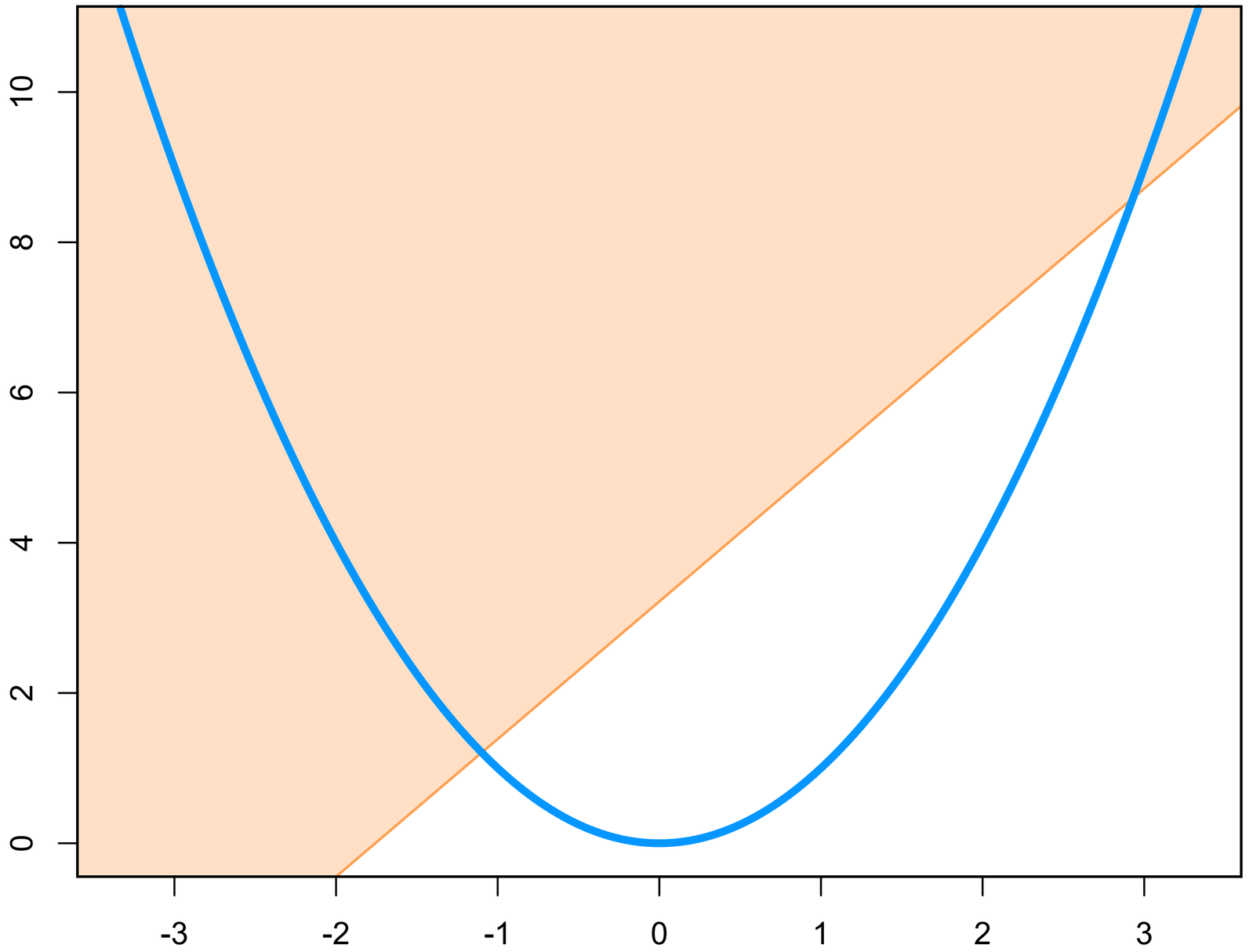

0
2
4
6
8
10
-3
-2
-1
0
1
2
3

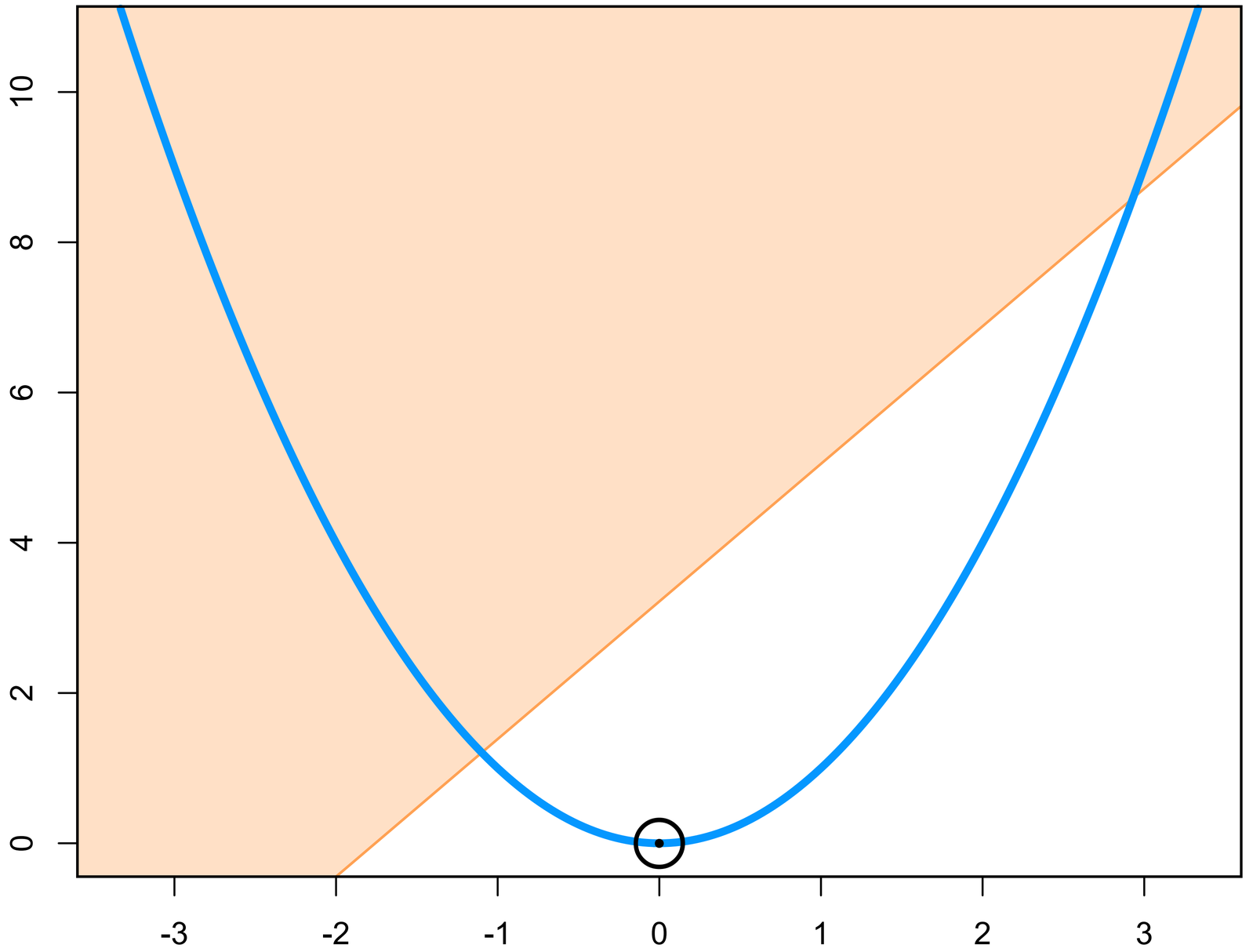

0
2
4
6
8
10
-3
-2
-1
0
1
2
3

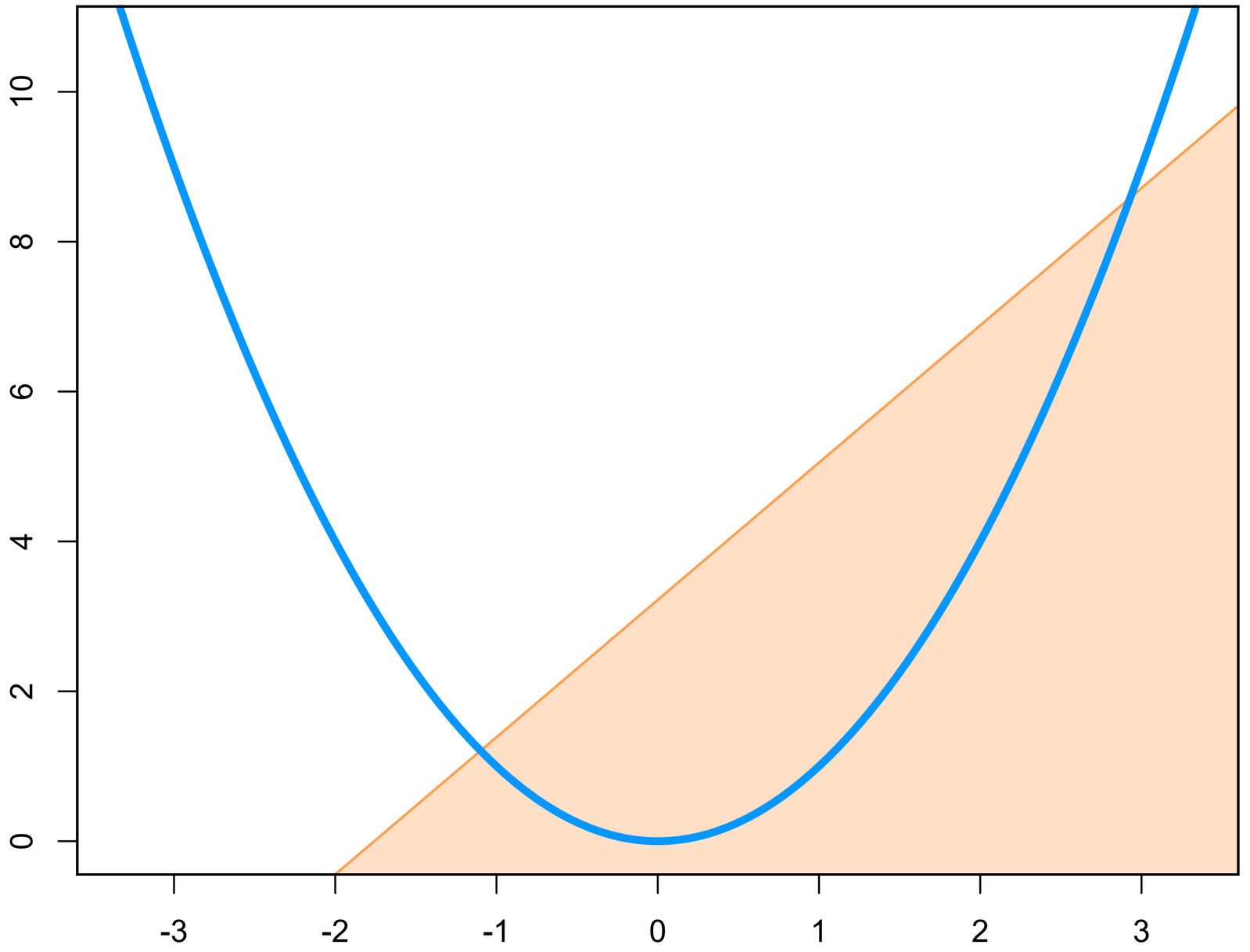

0
2
4
6
8
10
-3
-2
-1
0
1
2
3

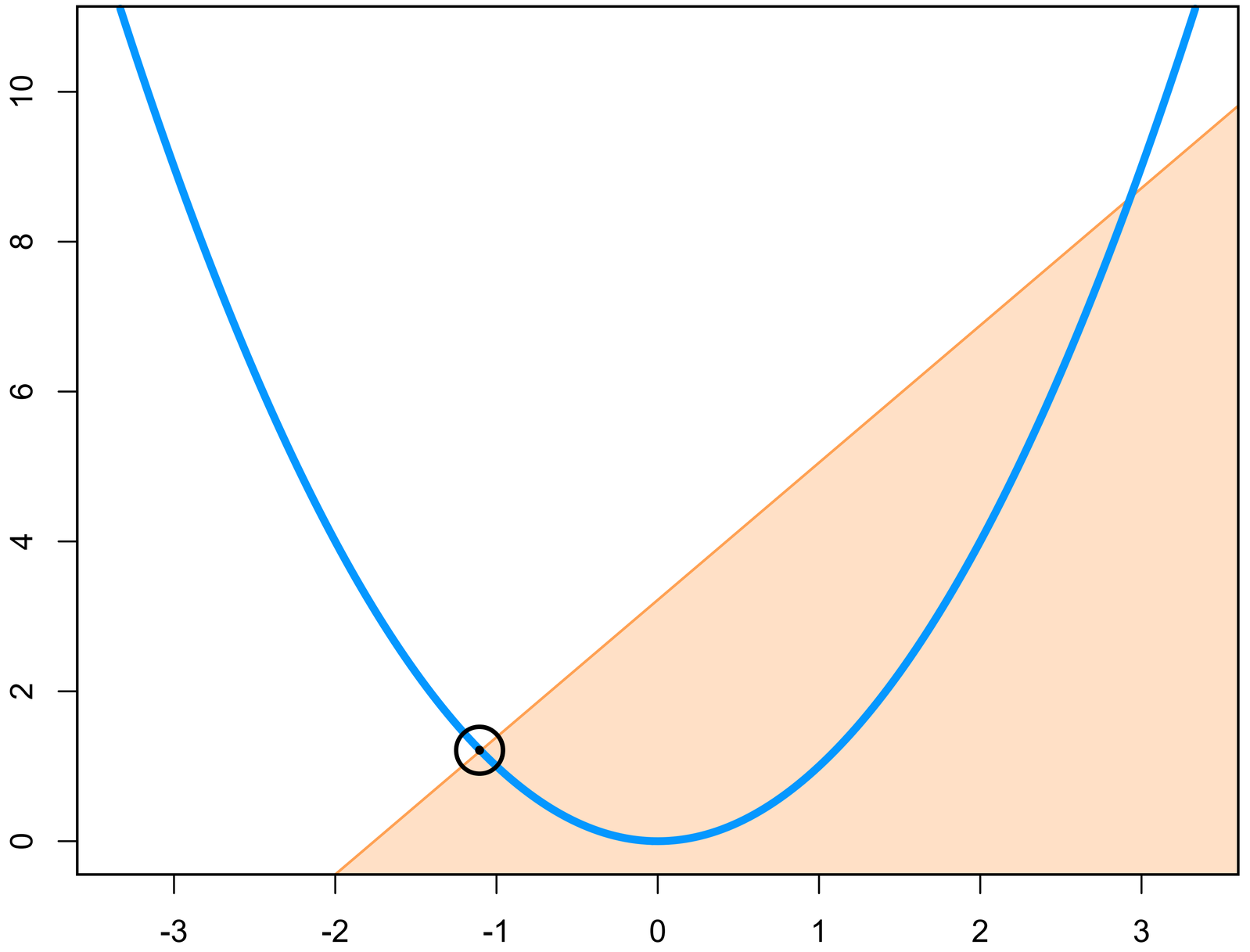

0
2
4
6
8
10
-3
-2
-1
0
1
2
3


10
8
6
4
2
0
-3
-2
-1
0
1
2
3


10
8
6
4
2
0
-3
-2
-1
0
1
2
3


10
8
6
4
2
0
-3
-2
-1
0
1
2
3


0
2
4
6
8
10
-3
-2
-1
0
1
2
3


10
8
6
4
2
0
-3
-2
-1
0
1
2
3

$$\underset{x}{\text{maximise}} \quad f(x)$$

$$\text{subject to:} \quad g(x) = 0$$

$$\mathcal{L}(x, \alpha) = f(x) - \alpha g(x)$$

$$\underset{x}{\text{maximise}} \quad f(x)$$

$$\text{subject to:} \quad g(x) = 0$$

$$\underset{x}{\text{maximise}} \quad f(x)$$

$$\text{subject to:} \quad g(x) \leq 0$$

$$\alpha_i g_i(x) = 0, \quad \forall i$$

Complementary Slackness

Inequality constraints are only active at equality

$$\underset{\mathbf{w},b}{\text{maximise}} \;\; M$$

$$\text{subject to:}$$

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq M$$

$$\|\mathbf{w}\| = 1$$

$$\underset{\mathbf{w},b}{\text{minimise}} \quad \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{subject to:} \quad y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \geq 1$$

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i^n \alpha_i \left(y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) - 1 \right)$$

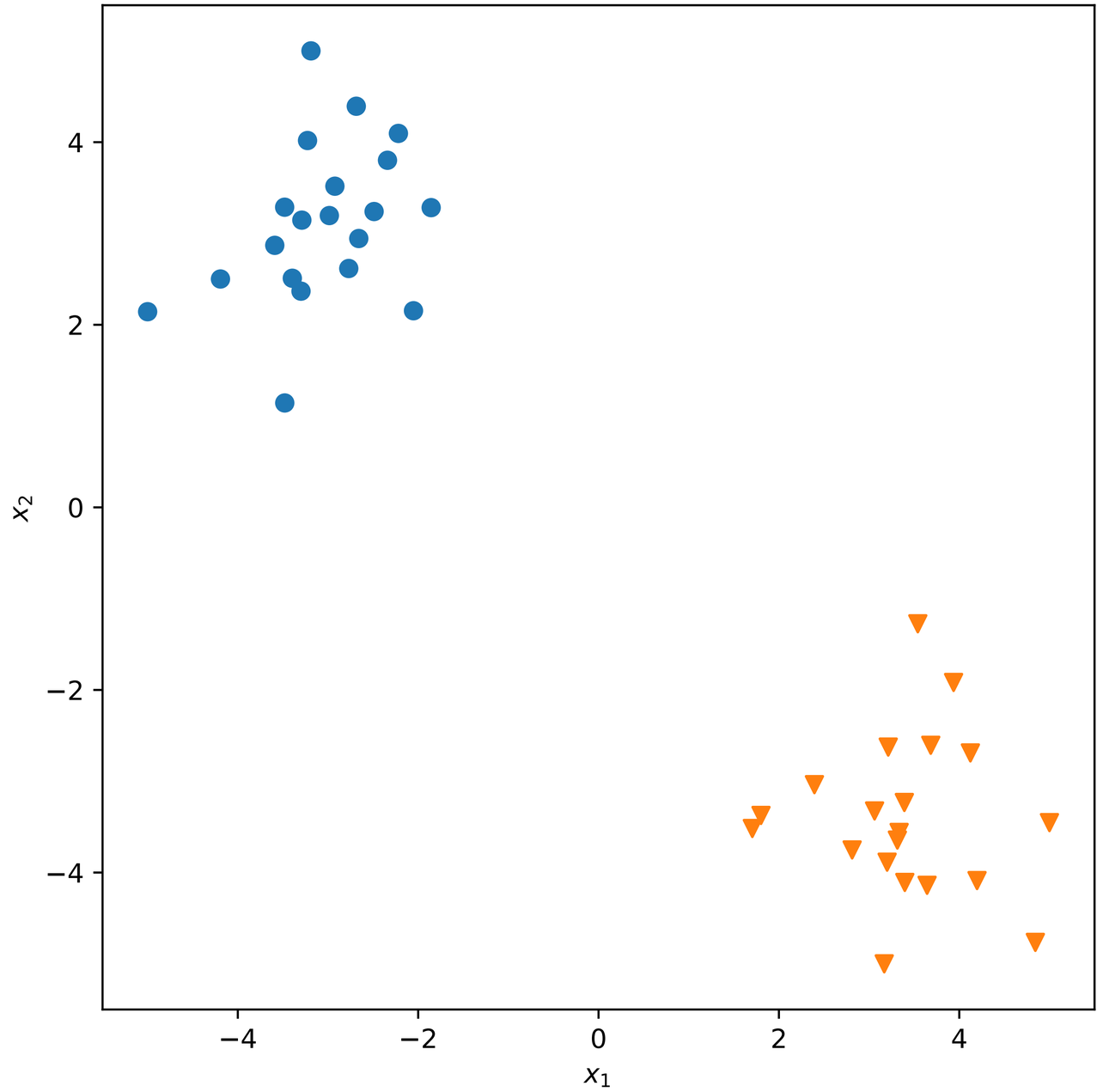
$$L_d = \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

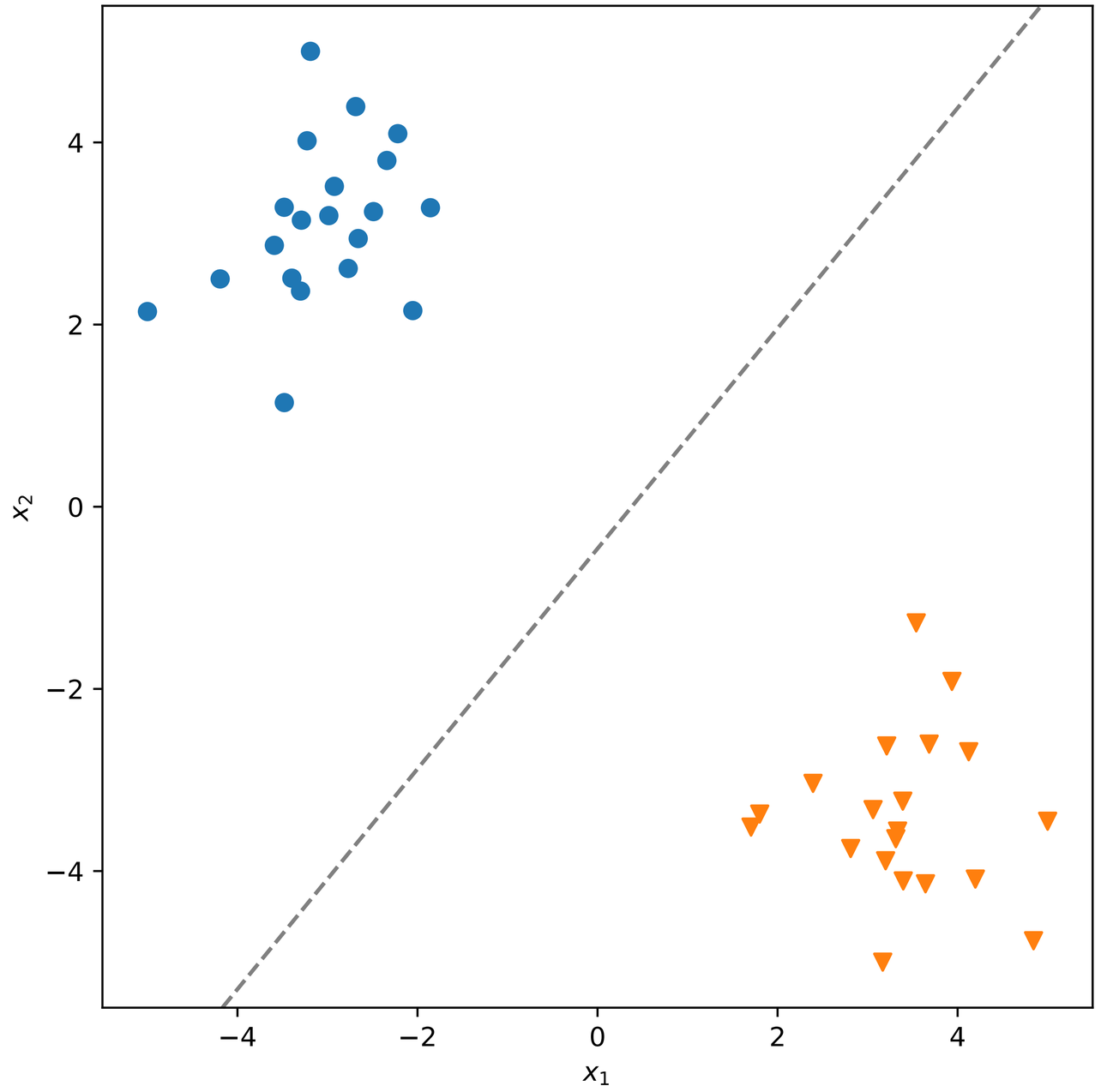
$$\mathbf{w} = \sum_{i}^{n} \alpha_i y_i \mathbf{x}_i$$

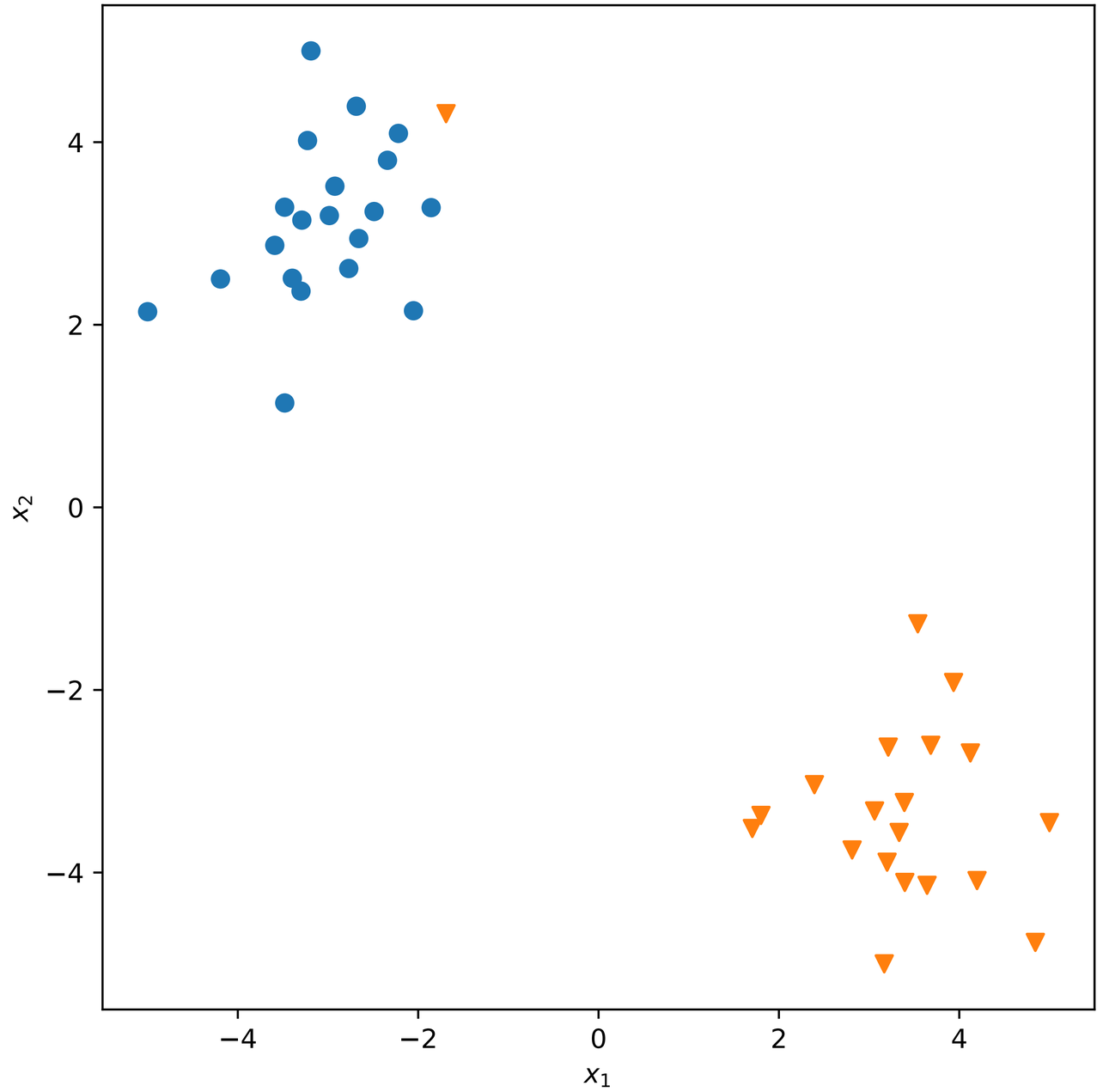
$$\alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \right] = 0 \quad \forall i$$

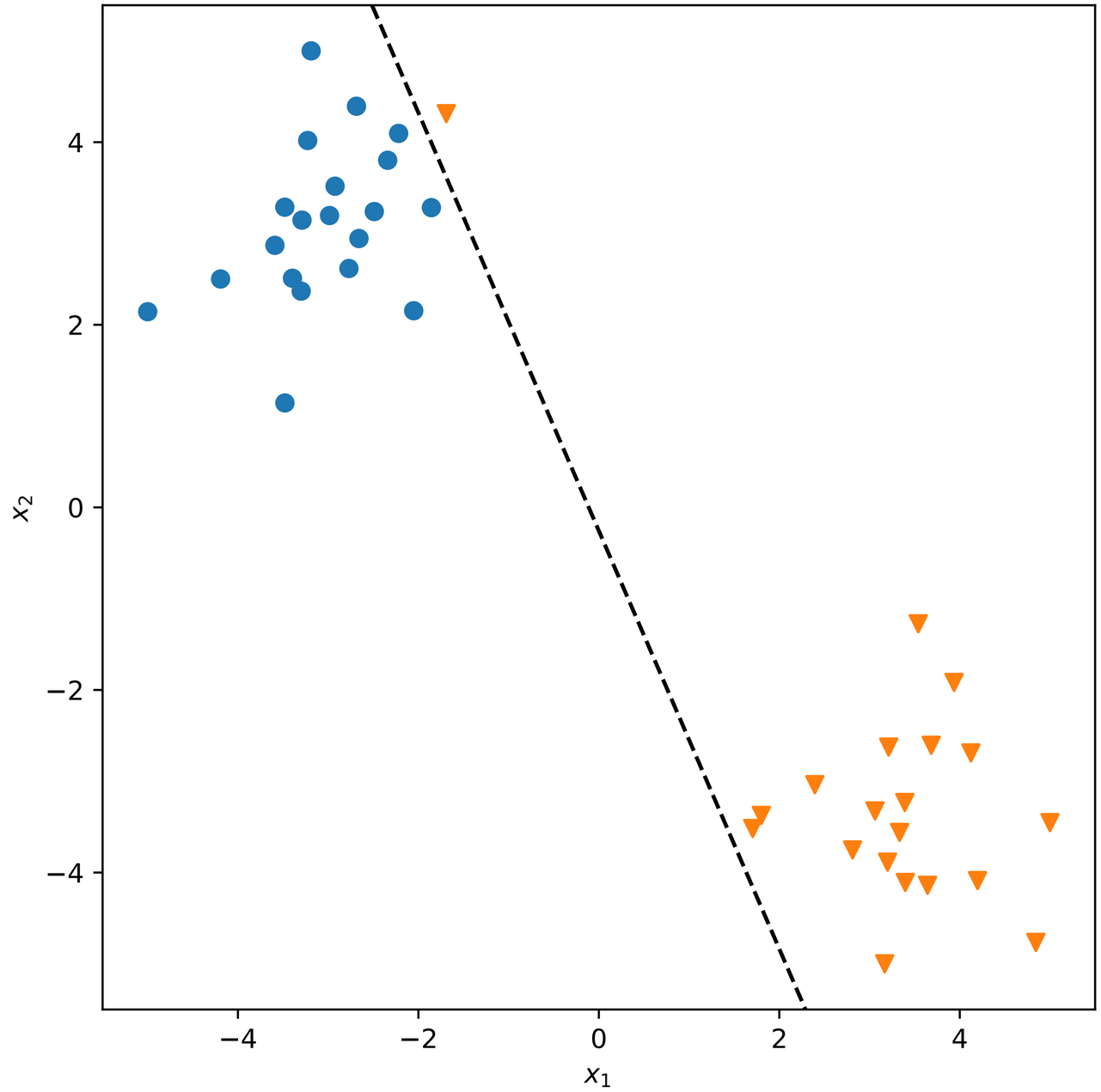
$$\underset{\alpha}{\text{maximise}} \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j \boxed{(\mathbf{x}_i \cdot \mathbf{x}_j)}$$

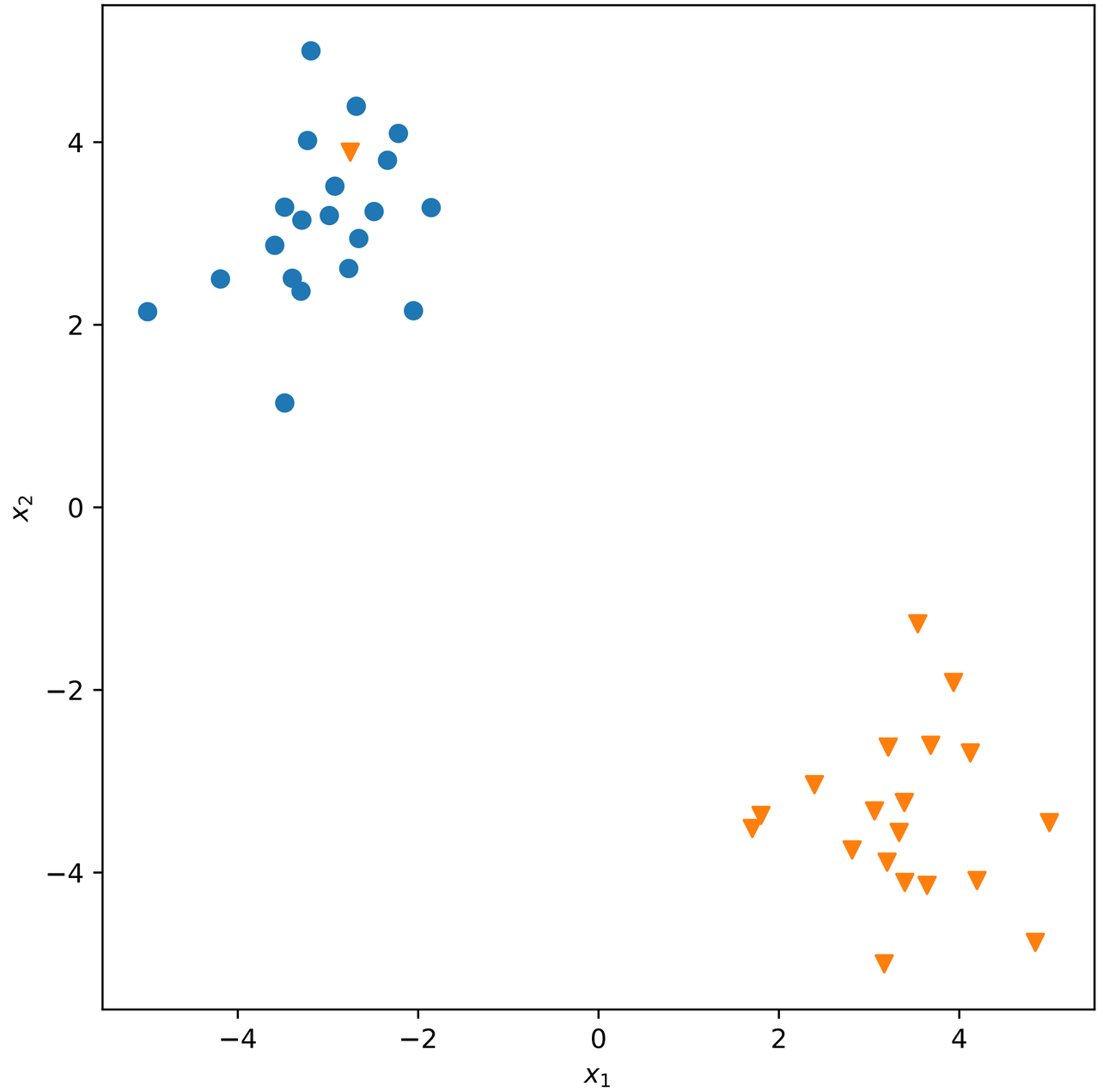
$$\hat{y} = \begin{cases} 1 & \text{if } \sum_i^n \alpha_i y_i \boxed{(\mathbf{x}_i \cdot \mathbf{x})} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

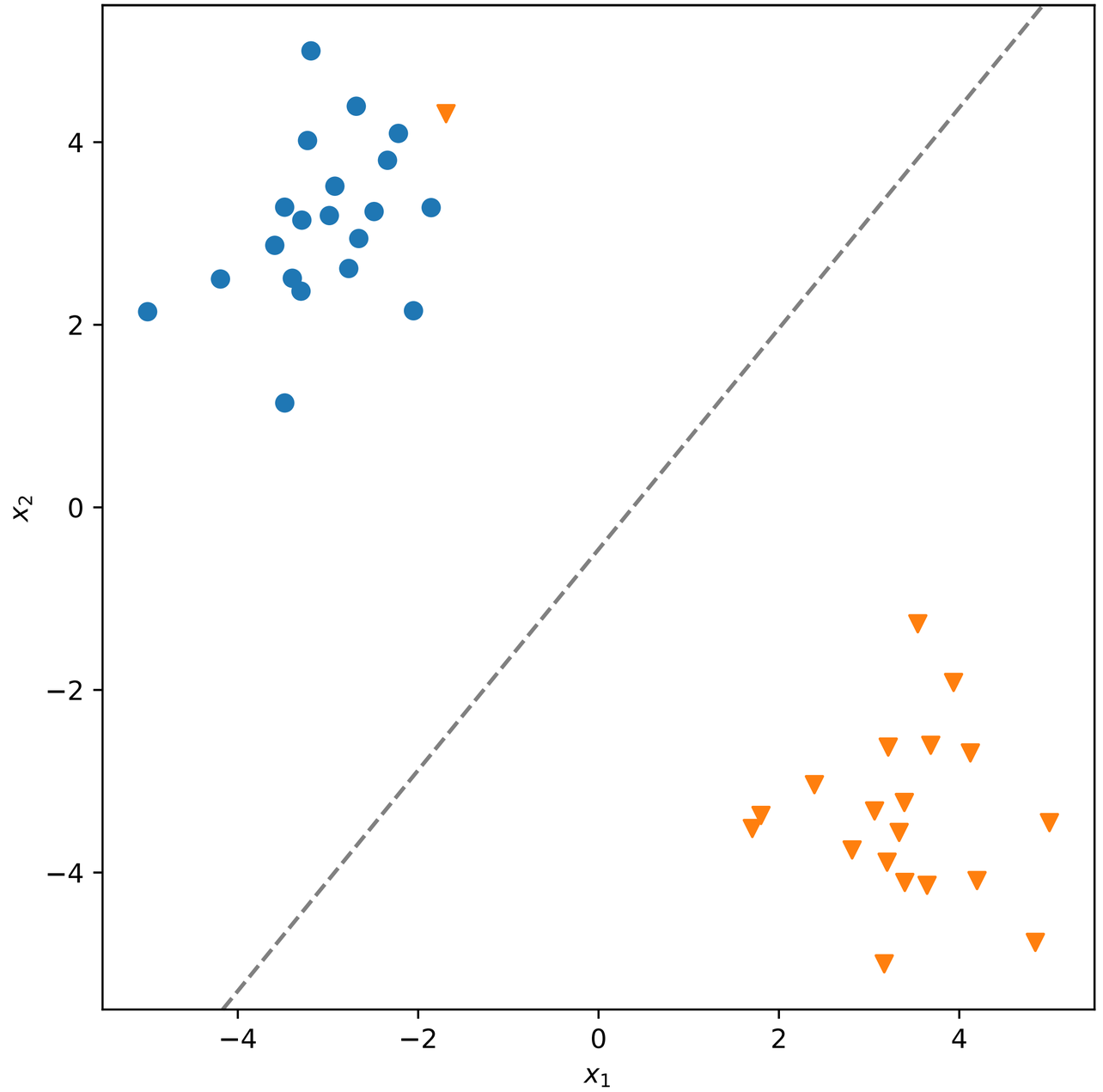

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

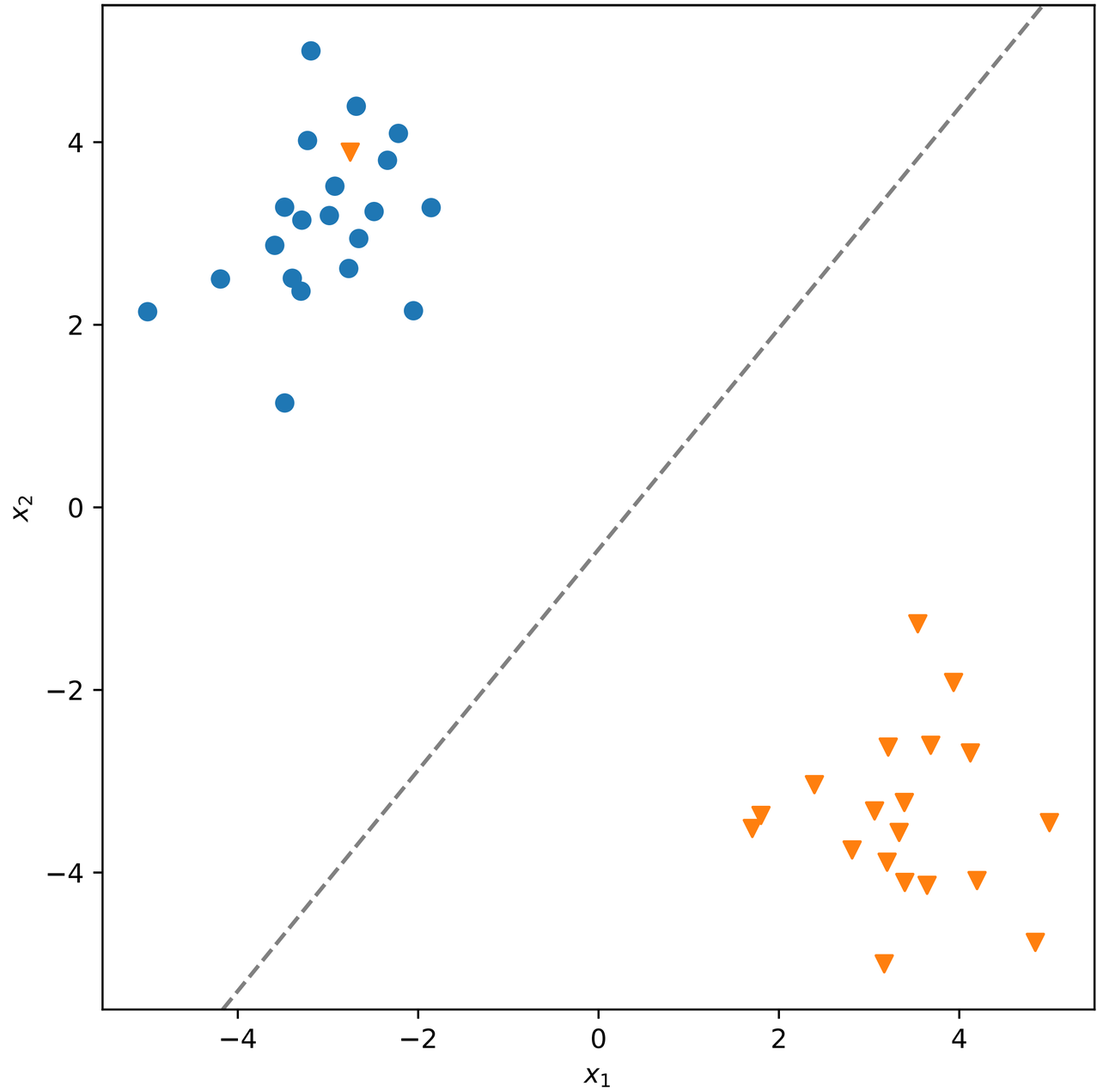

4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

$$\underset{\mathbf{w},b,\xi}{\text{maximise}} \;\; M$$

subject to:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq M(1 - \xi_i)$$

$$\|\mathbf{w}\| = 1$$

$$\xi_i \geq 0 \;\; \forall i$$

$$\sum_i \xi_i \leq \text{ constant}$$

$$\underset{\mathbf{w},b,\xi}{\text{minimise}}\ \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i}^{n}\xi_i$$

subject to:

$$y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i^n \xi_i - \sum_i^n \alpha_i \left[y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) - \left(1 - \xi_i\right)\right] - \sum_i^n \mu_i \xi_i$$

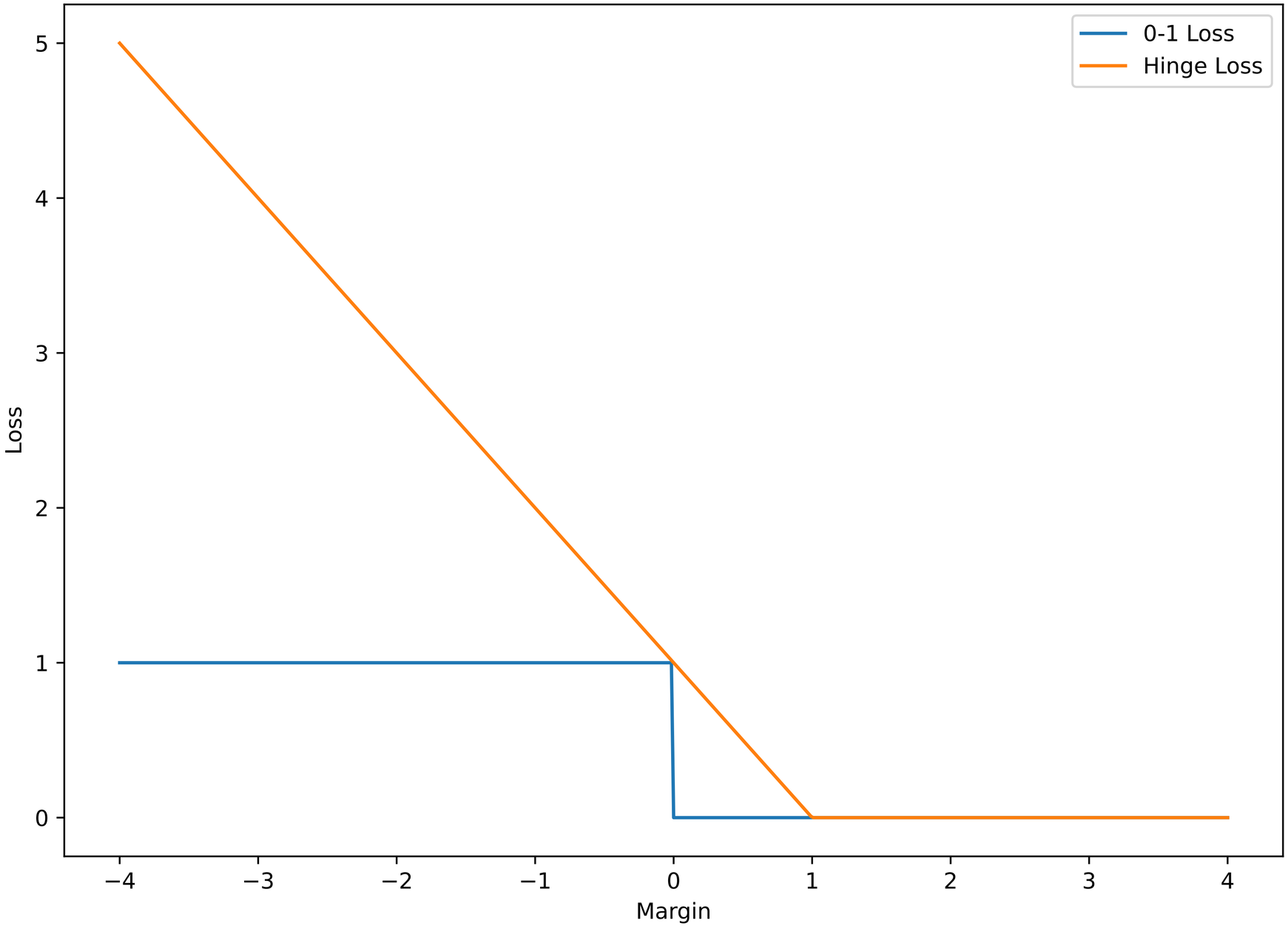
$$L_d = \sum_i^n \alpha_i - \frac{1}{2}\sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

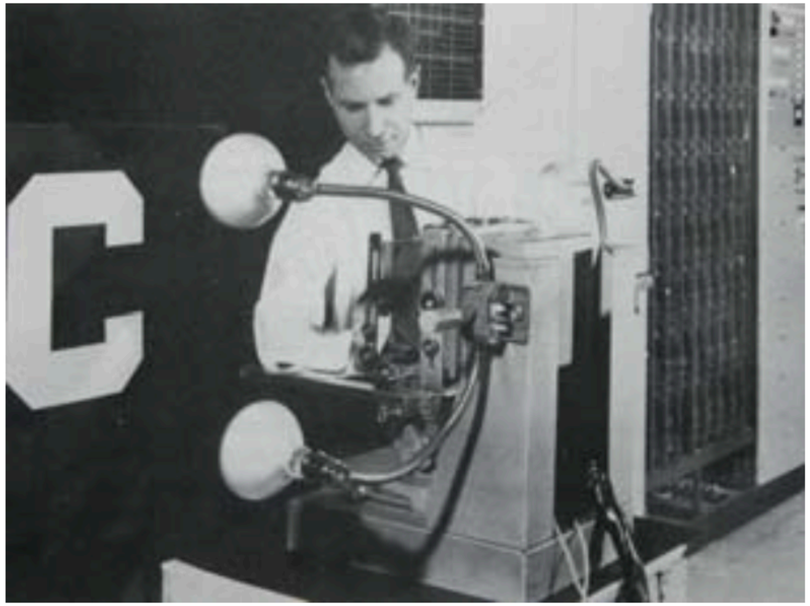
$$0 \leq \alpha_i \leq C$$

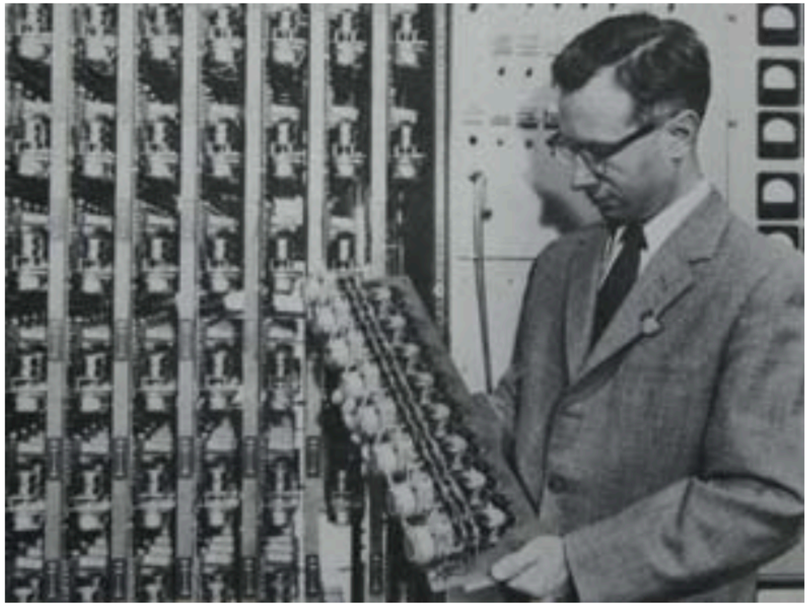
$$\alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w}) - (1 - \xi_i) \right] = 0$$

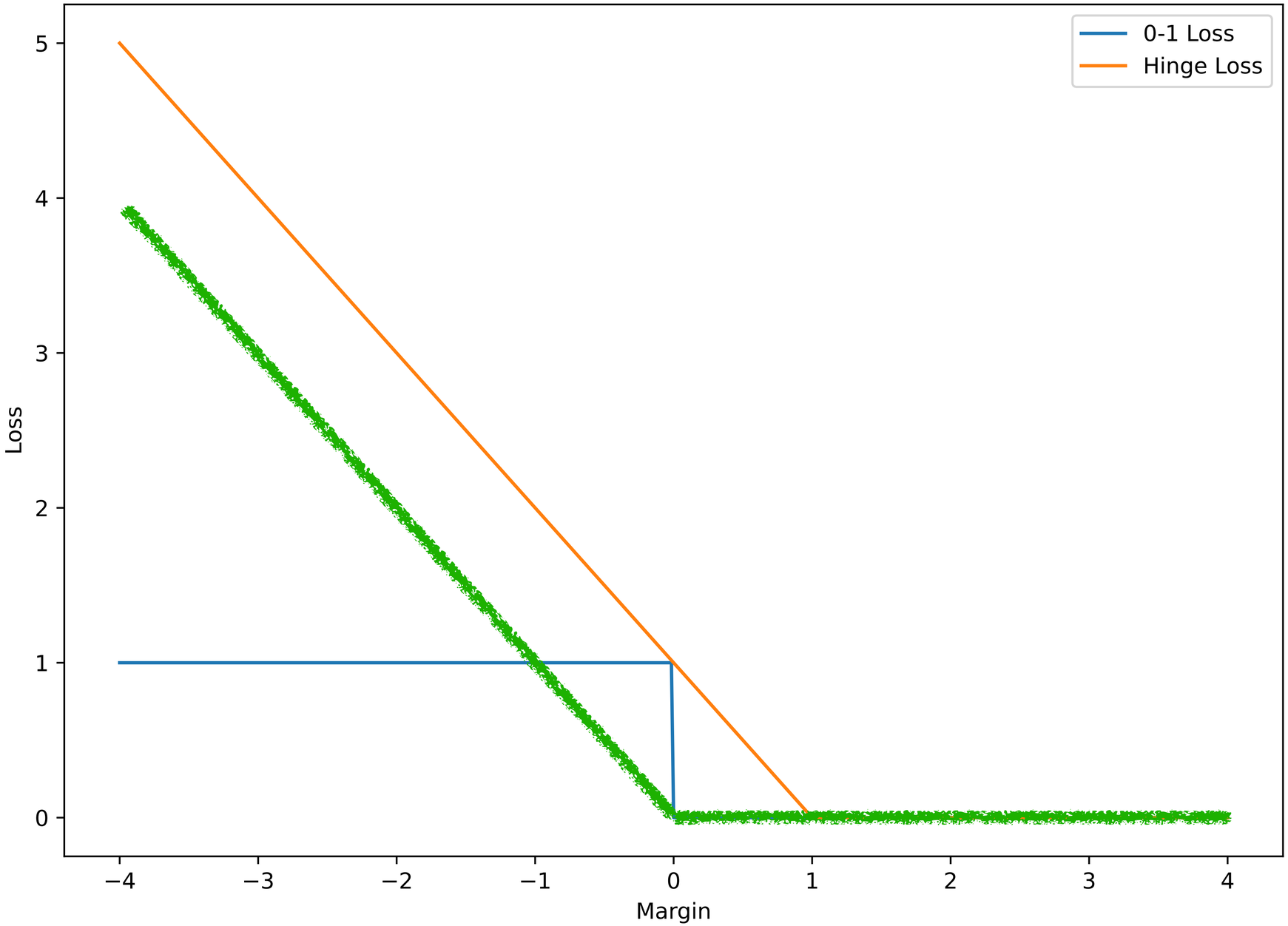
$$y_i(\mathbf{x}_i \cdot \mathbf{w}) - (1 - \xi_i) \geq 0$$

$$
\begin{aligned}
L(y, \mathbf{x}) &= \max(0, 1 - yf(\mathbf{x})) \\
&= \max(0, 1 - y(\mathbf{x} \cdot \mathbf{w} + b))
\end{aligned}
$$


0-1 Loss
Hinge Loss
Loss
Margin
5
4
3
2
1
0
−4
−3
−2
−1
0
1
2
3
4


C

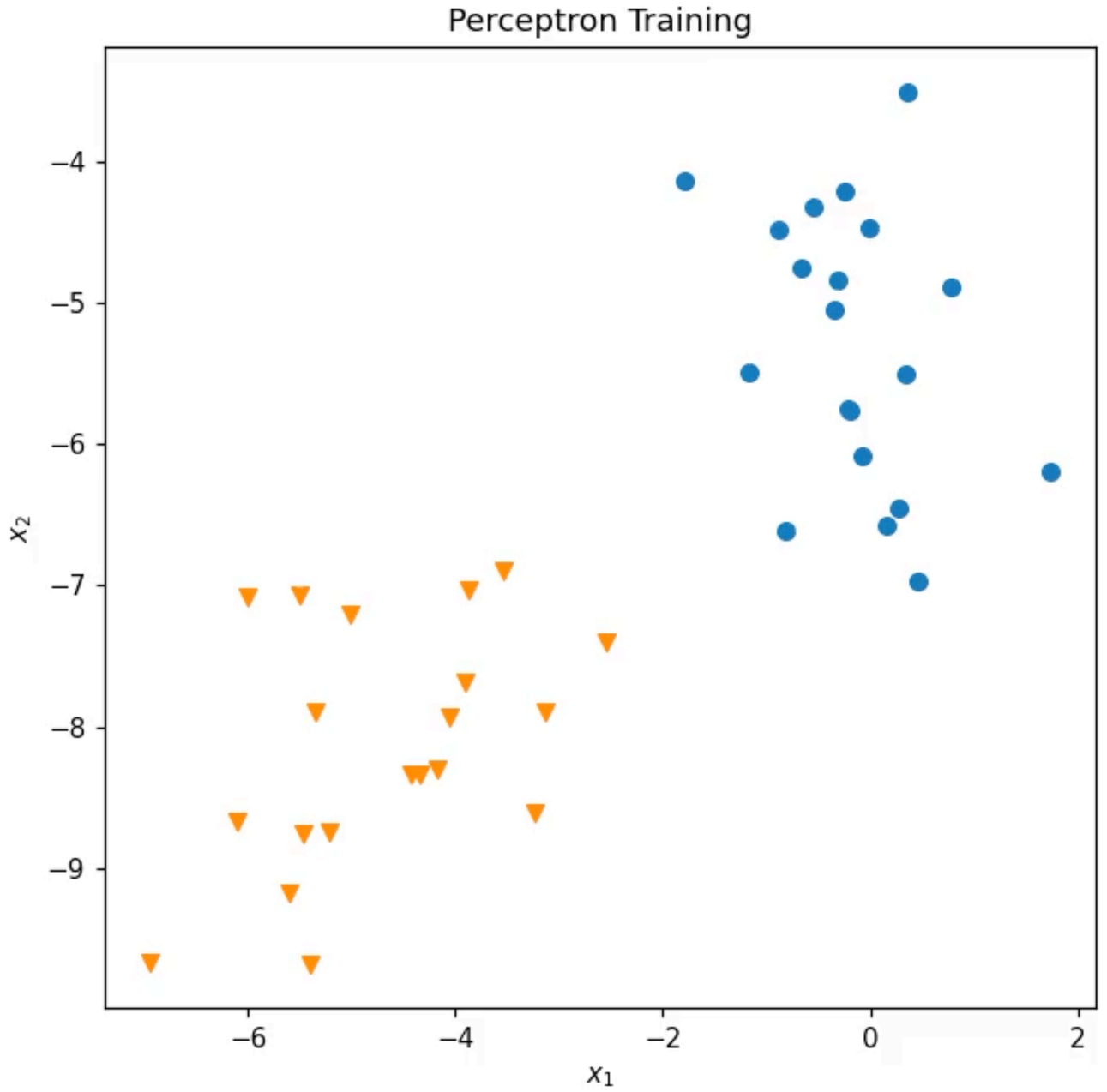



0-1 Loss
Hinge Loss
Loss
Margin
5
4
3
2
1
0
−4
−3
−2
−1
0
1
2
3
4

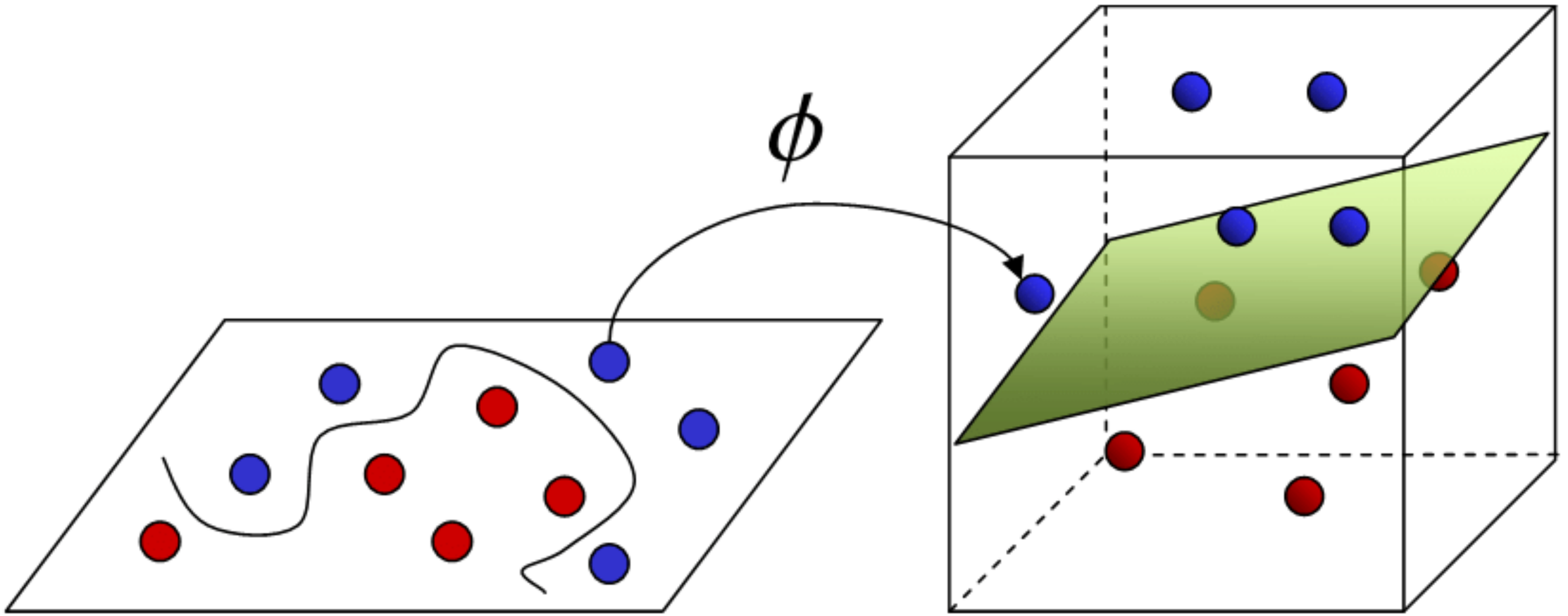
$$
\begin{aligned}
L(y, \mathbf{w}) &= \max(0, -yf(\mathbf{x})) \\
&= \max(0, -y(\mathbf{x} \cdot \mathbf{w} + b))
\end{aligned}
$$

The Algorithm

- Initialise: $\mathbf{w} \leftarrow \mathbf{0}$
  - → (Could also be any other initial value)
- For each sample $\mathbf{x}$ in the training set:
  - Predict: $\hat{y} = \mathbf{1}(\mathbf{x} \cdot \mathbf{w} \geq 0)$
  - Update: $\mathbf{w} \leftarrow \mathbf{w} + \alpha(y - \hat{y})\mathbf{x}$
    - → NB: only updates weights when prediction is wrong
- Repeat until there are no errors


Perceptron Training
−4
−5
−6
−7
−8
−9
x_2
−6
−4
−2
0
2
x_1

This One Weird Trick


ϕ
a) Input Space
b) Feature Space

$$\mathbf{x}' = \phi(\mathbf{x})$$

$$\underset{\alpha}{\text{maximise}} \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j \boxed{(\mathbf{x}_i \cdot \mathbf{x}_j)}$$

$$\hat{y} = \begin{cases} 1 & \text{if } \sum_i^n \alpha_i y_i \boxed{(\mathbf{x}_i \cdot \mathbf{x})} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\underset{\alpha}{\text{maximise}} \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i}^{n} \sum_{j}^{n} \alpha_i \alpha_j y_i y_j (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$$

$$\hat{y} = \begin{cases} 1 & \text{if } \sum_i^n \alpha_i y_i (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x})) + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\phi(\mathbf{x}) \mapsto \mathbb{R}^{10{,}000{,}000}$$

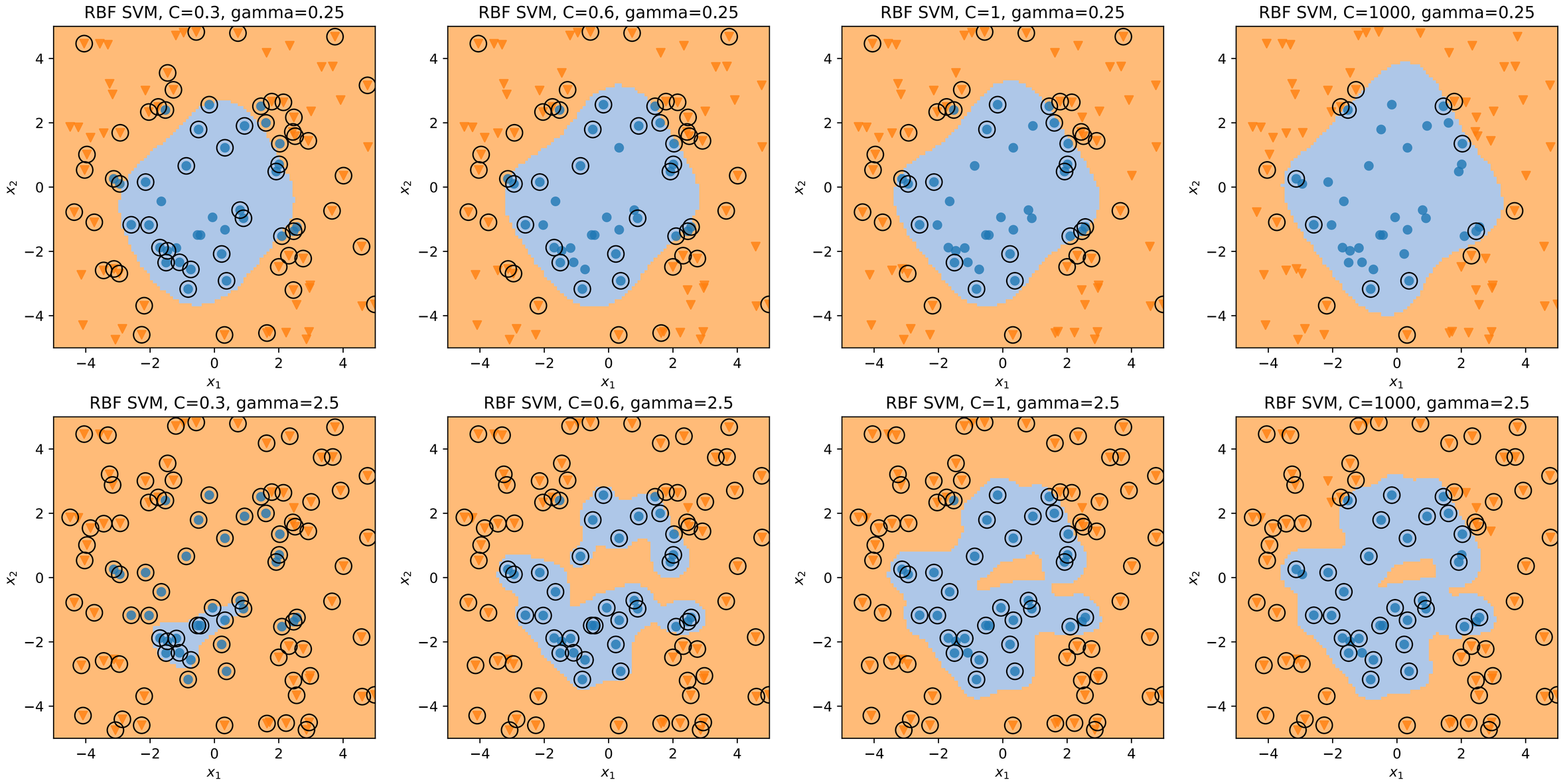
$$K(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u}) \cdot \phi(\mathbf{v})$$

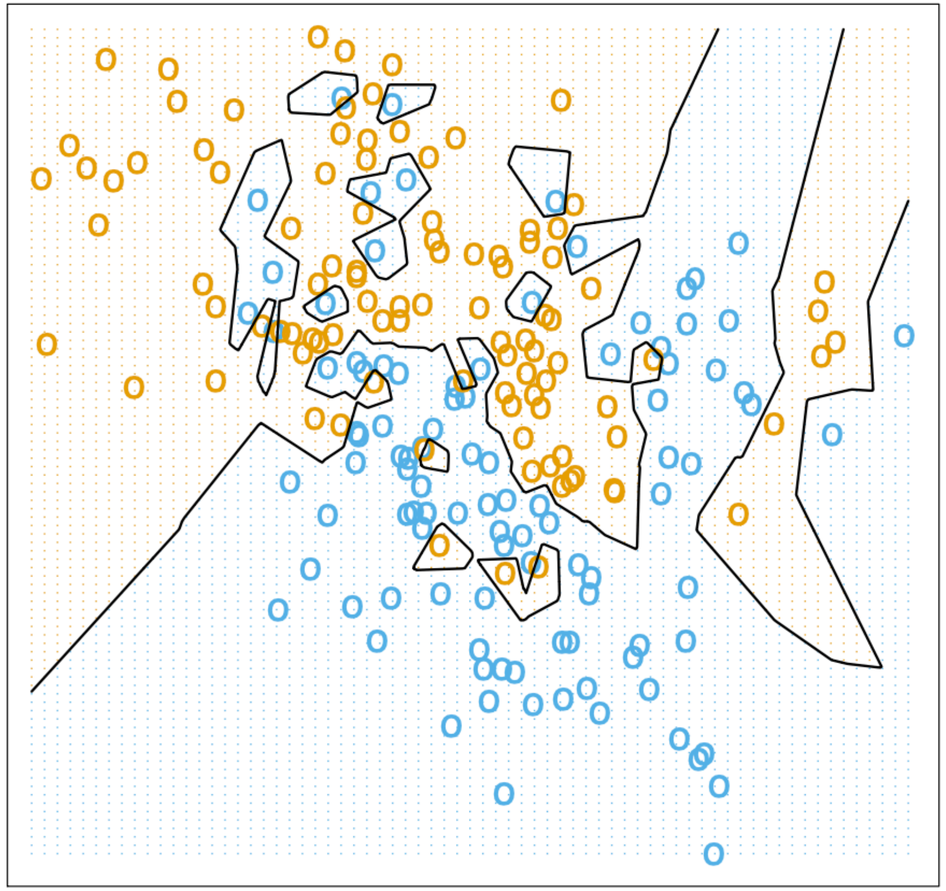
$$\underset{\alpha}{\text{maximise}} \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i}^{n} \sum_{j}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\hat{y} = \begin{cases} 1 & \text{if } \sum_i^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

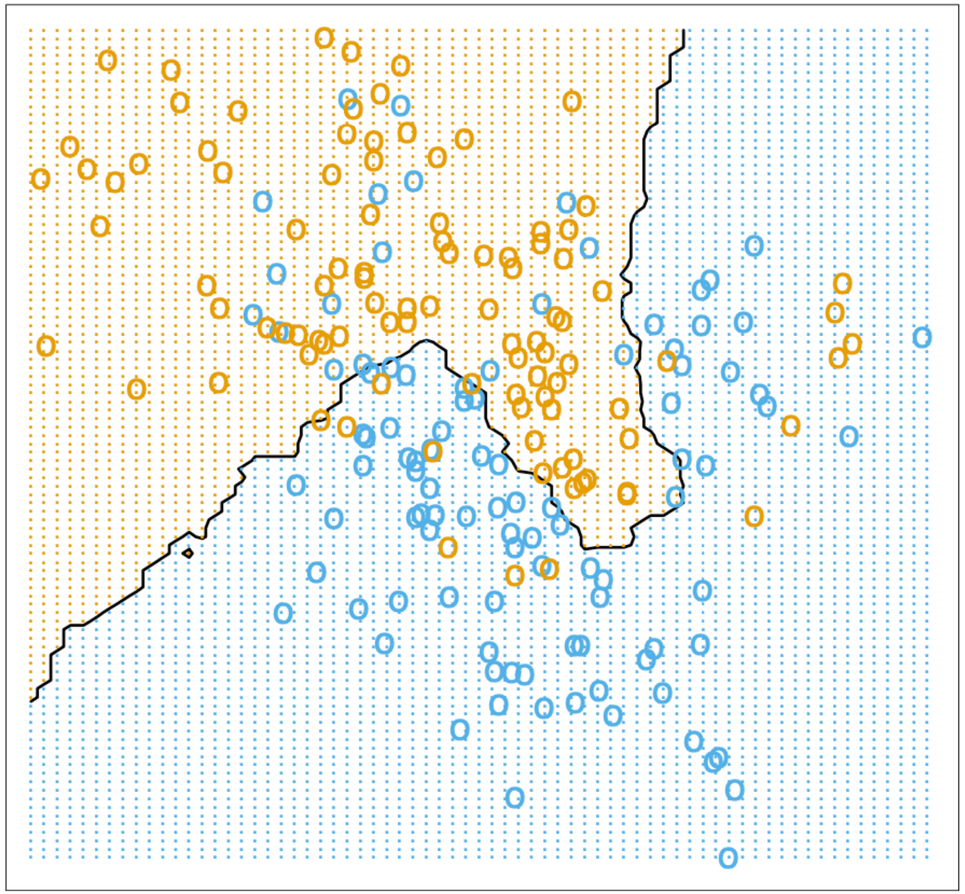
$$K(\mathbf{u}, \mathbf{v}) = e^{-\gamma \|\mathbf{u}-\mathbf{v}\|^2}$$

$$= e^{-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}}$$


RBF SVM, C=0.3, gamma=0.25
RBF SVM, C=0.6, gamma=0.25
RBF SVM, C=1, gamma=0.25
RBF SVM, C=1000, gamma=0.25
RBF SVM, C=0.3, gamma=2.5
RBF SVM, C=0.6, gamma=2.5
RBF SVM, C=1, gamma=2.5
RBF SVM, C=1000, gamma=2.5
x_1
x_2



k=1



k=15

$$K(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u}) \cdot \phi(\mathbf{v})$$

$$\mathbf{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n, \mathbf{x}_2) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

Mercer's Condition

$$K(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u}) \cdot \phi(\mathbf{v}) \iff \mathbf{K} \text{ is PSD}$$

Questions?

Next week: Neural Networks