# Machine Learning

## Lecture 7 – Support Vector Machine

**Jialin Yu**
**PhD Student @ Durham**

# Today

- Linear Separable SVM
- Non-linear separable SVM

# Linear Separable SVM

# Linear Separable SVM

- Hard Margin Support Vector Machine
- Soft Margin Support Vector Machine

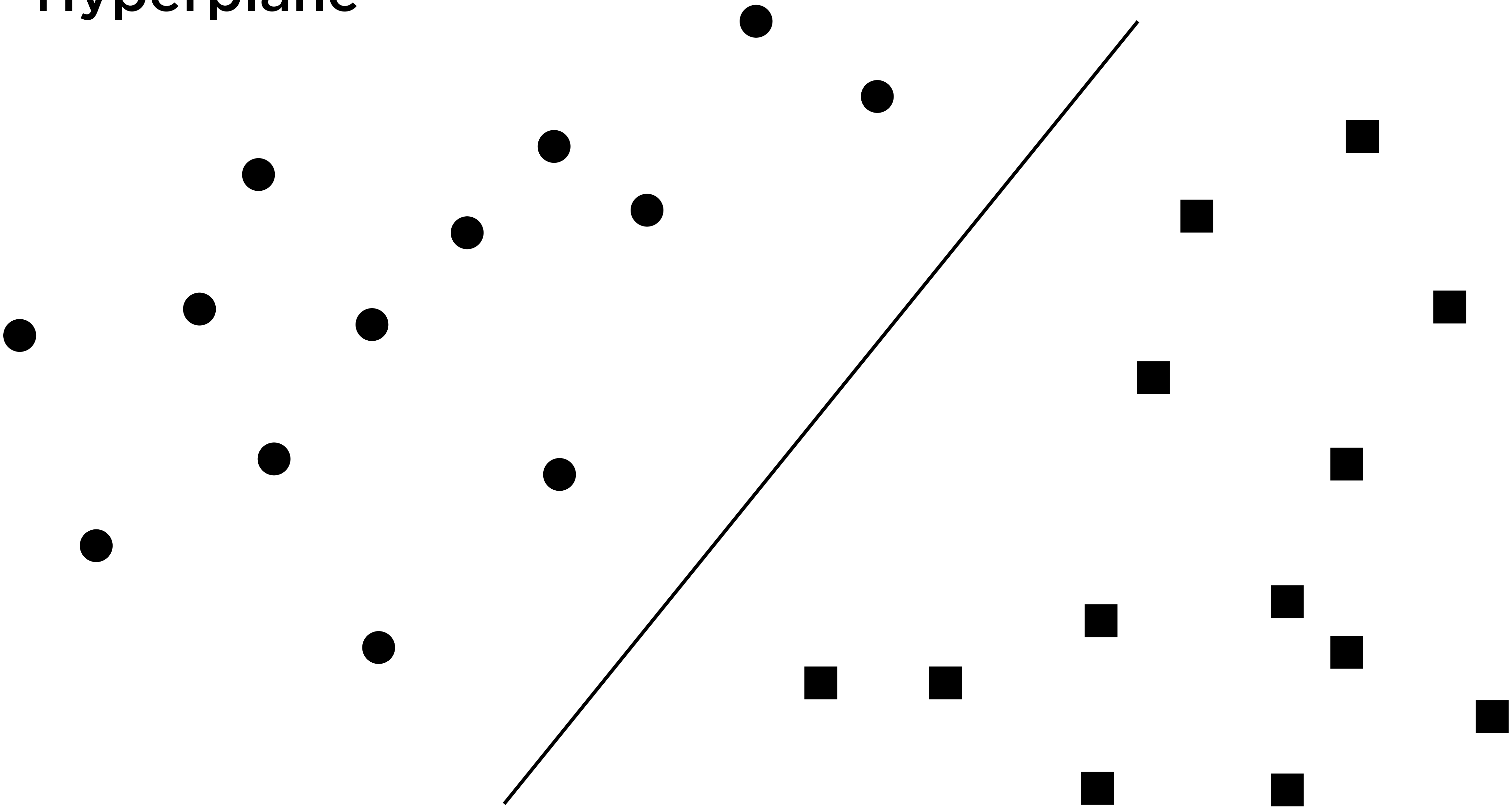
# Hard Margin Support Vector Machine

## Hyperplane

In geometry, a hyperplane is a subspace whose dimension is one less than its ambient space. For the context of this module, the ambient space is defined as the Hilbert space($\mathcal{H}$).

Hyperplane is a linear decision surface that can be used to separate and classify data points.

# Hyperplane



# Hard Margin Support Vector Machine

## Intuition: an practical problem

Given training data $(x_i, y_i)$ for $i = 1, \dots, N$ with $x_i \in \mathbb{R}^2$ and $y_i \in \{-1, +1\}$, learn a classifier $f(x)$ training such that:

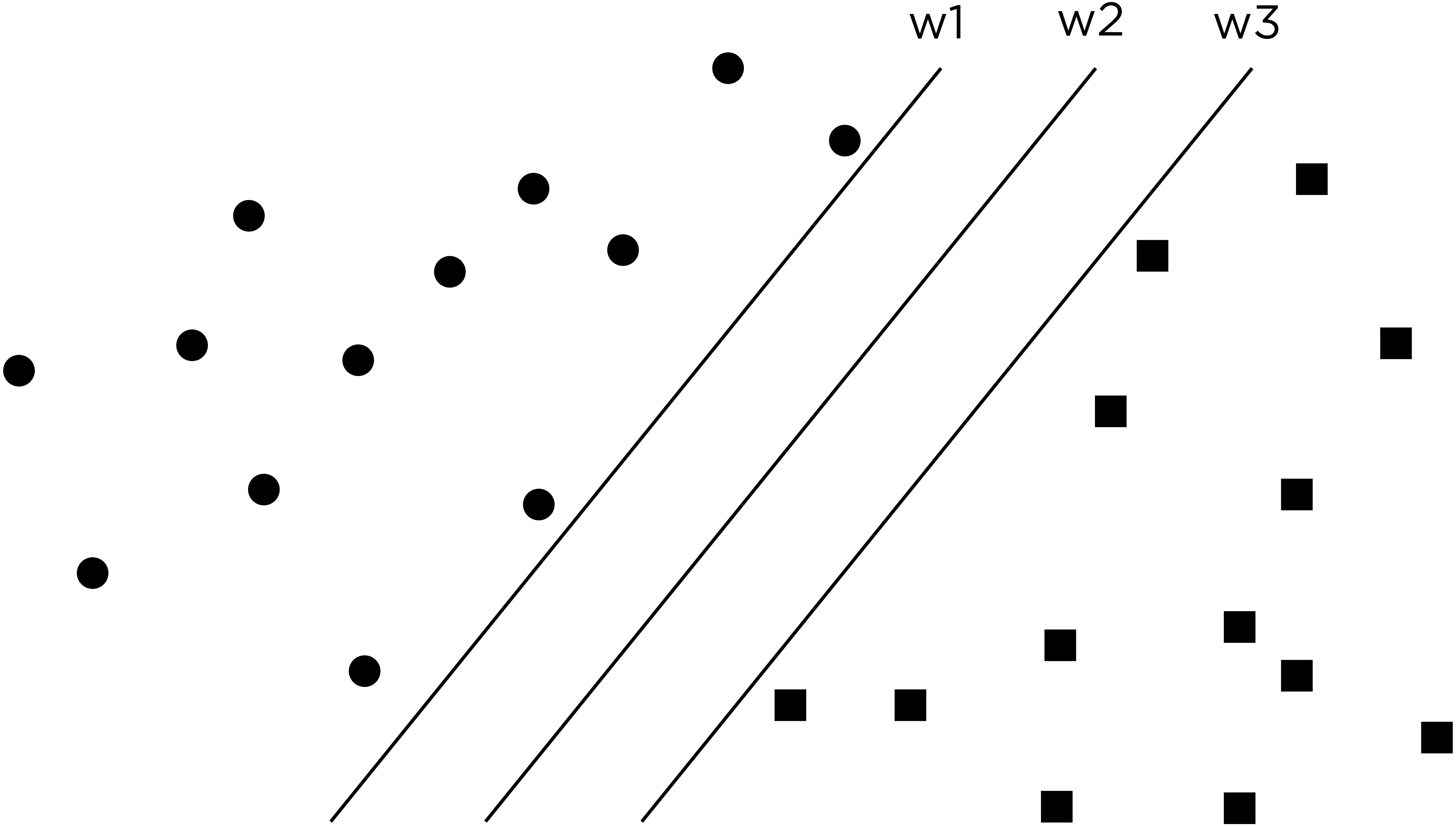
$$f(x_i) = \begin{cases} \geq 0, & y_i = +1 \\ < 0, & y_i = -1 \end{cases}$$

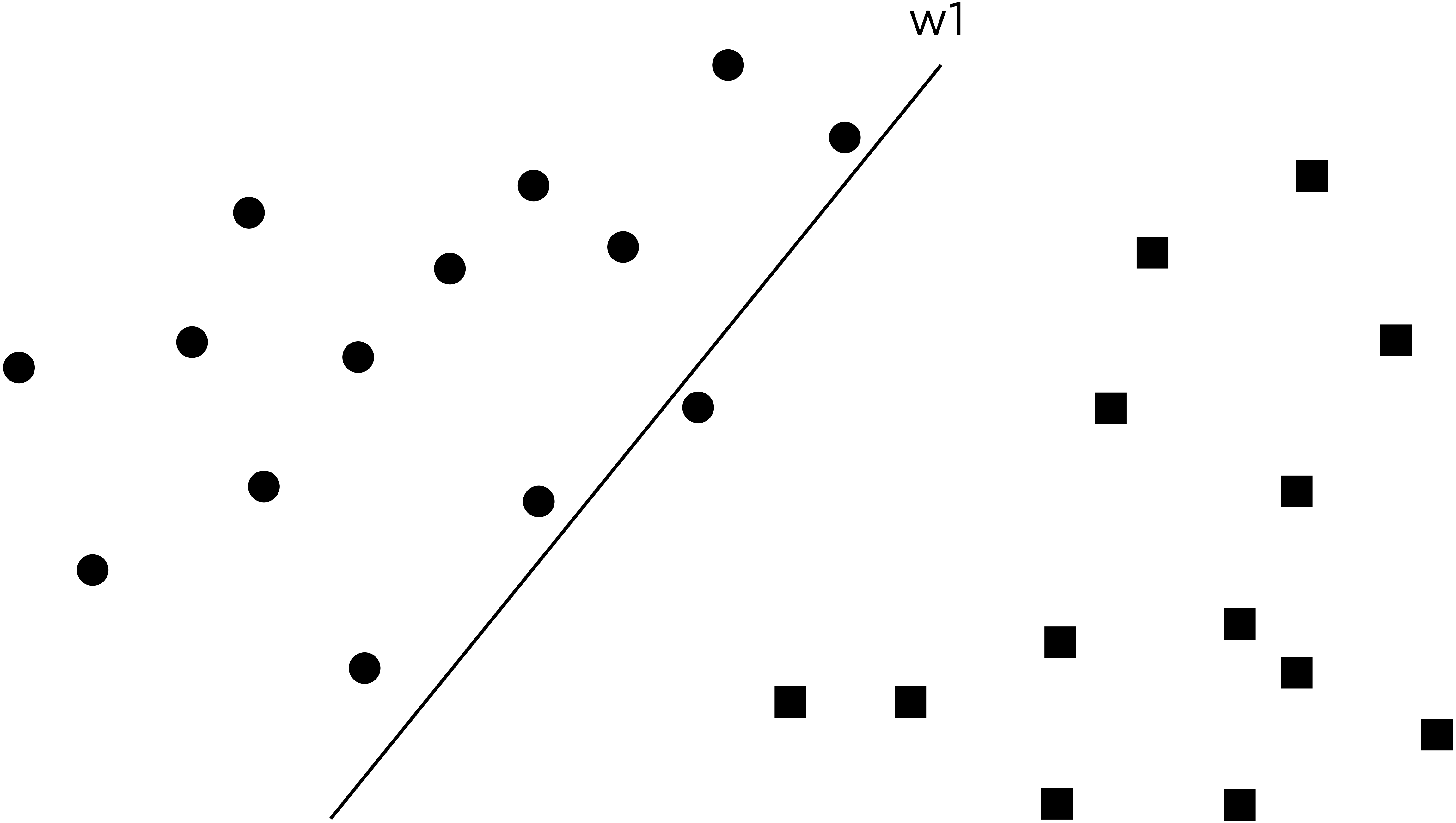
i.e. $y_i f(x_i) > 0$ for a correct classification.

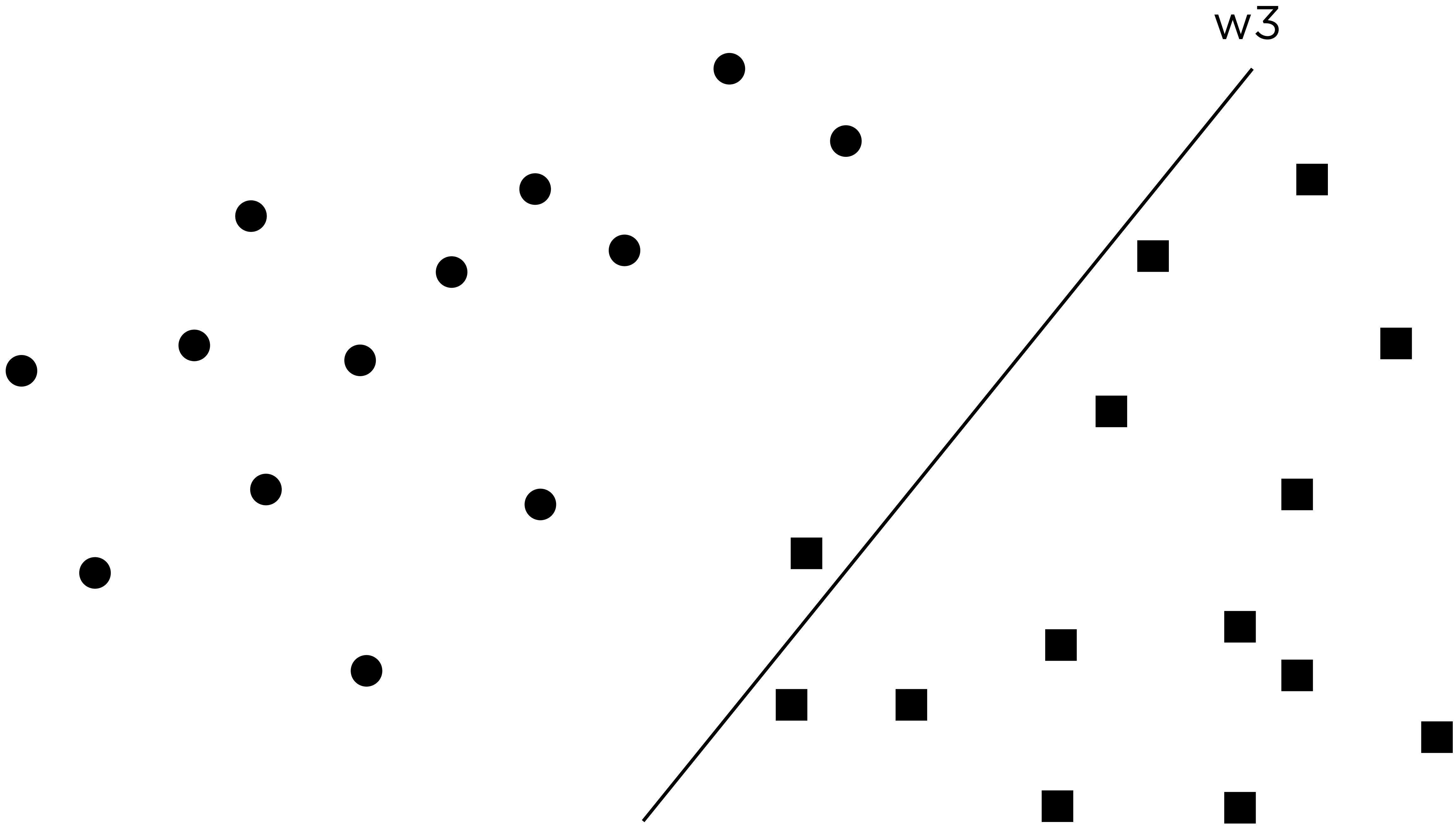
What is the problem with this solution?

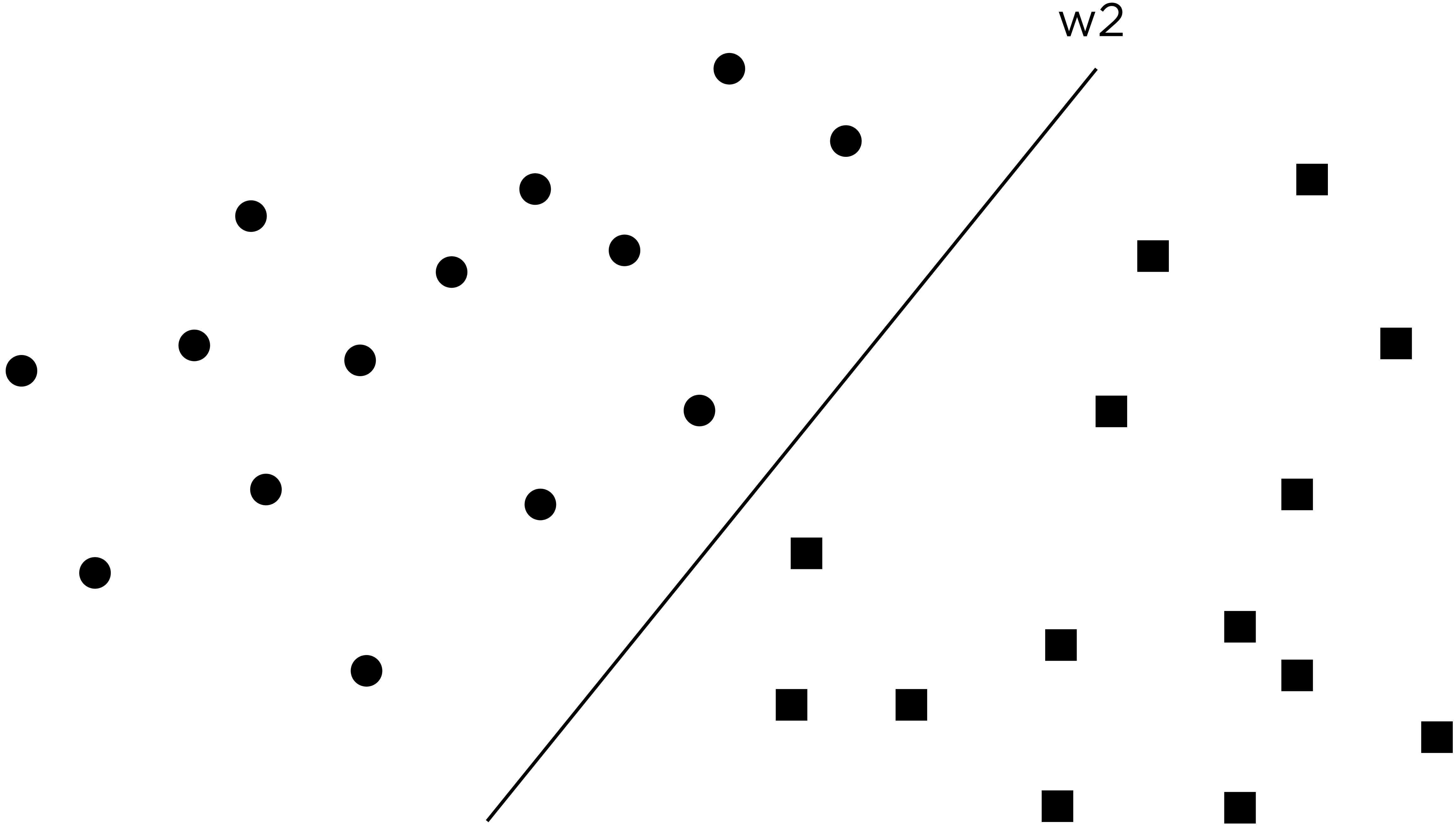
# Hard Margin Support Vector Machine

There is no optimal solution of Hyperplane given the training data points. Hyperplane here can alternatively be named **decision boundary**. An example solution could be either $w1$, $w2$, or $w3$.

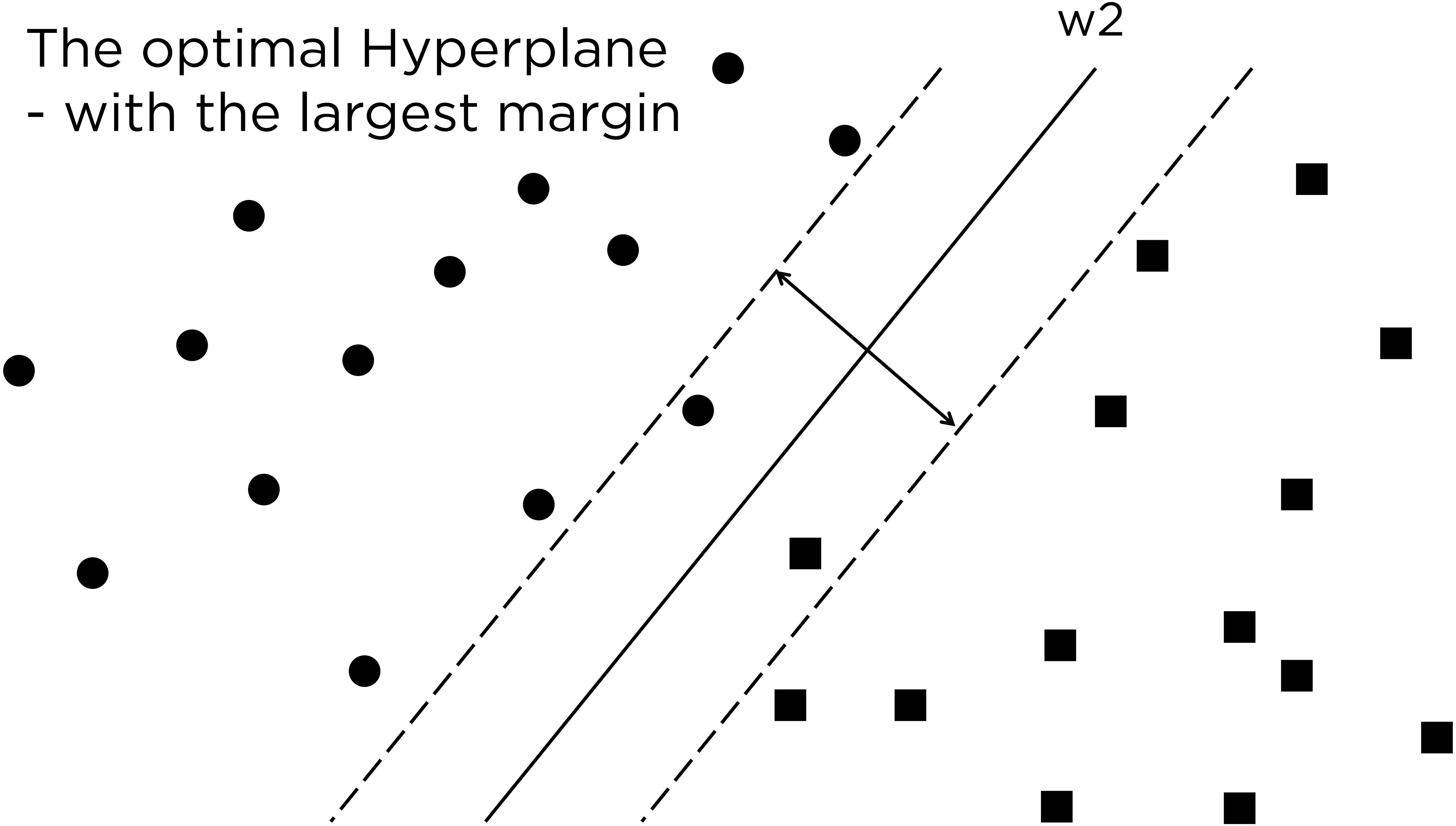

w1
w2
w3


w1


w3


w2

The optimal Hyperplane
- with the largest margin



# Hard Margin Support Vector Machine

## Definition: Separating Hyperplane

Let $S = \{(x_i, y_i)\}_{i=1}^{m} \in \mathbb{R}^d \times \{-1, +1\}$ be a training set.

By a hyperplane we mean a set of Hilbert space $\mathcal{H}_{w,b} = \{x \in \mathbb{R}^d : w^T x + b = 0\}$ parameterized by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$

We assume that the data are linearly separable, that is, there exist $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y_i(w^T x_i + b) > 0,\ i = 1, \ldots, m$

In which case we call $\mathcal{H}_{w,b}$ a **separating hyperplane**

**Note that we require the inequality to be strict (we do not admit that the data lie on a hyperplane).**

# Hard Margin Support Vector Machine

## Definition: Distance & Margin

The **distance** $\rho_x(w, b)$ of a point x from a hyperplane $\mathcal{H}_{w,b}$ is:

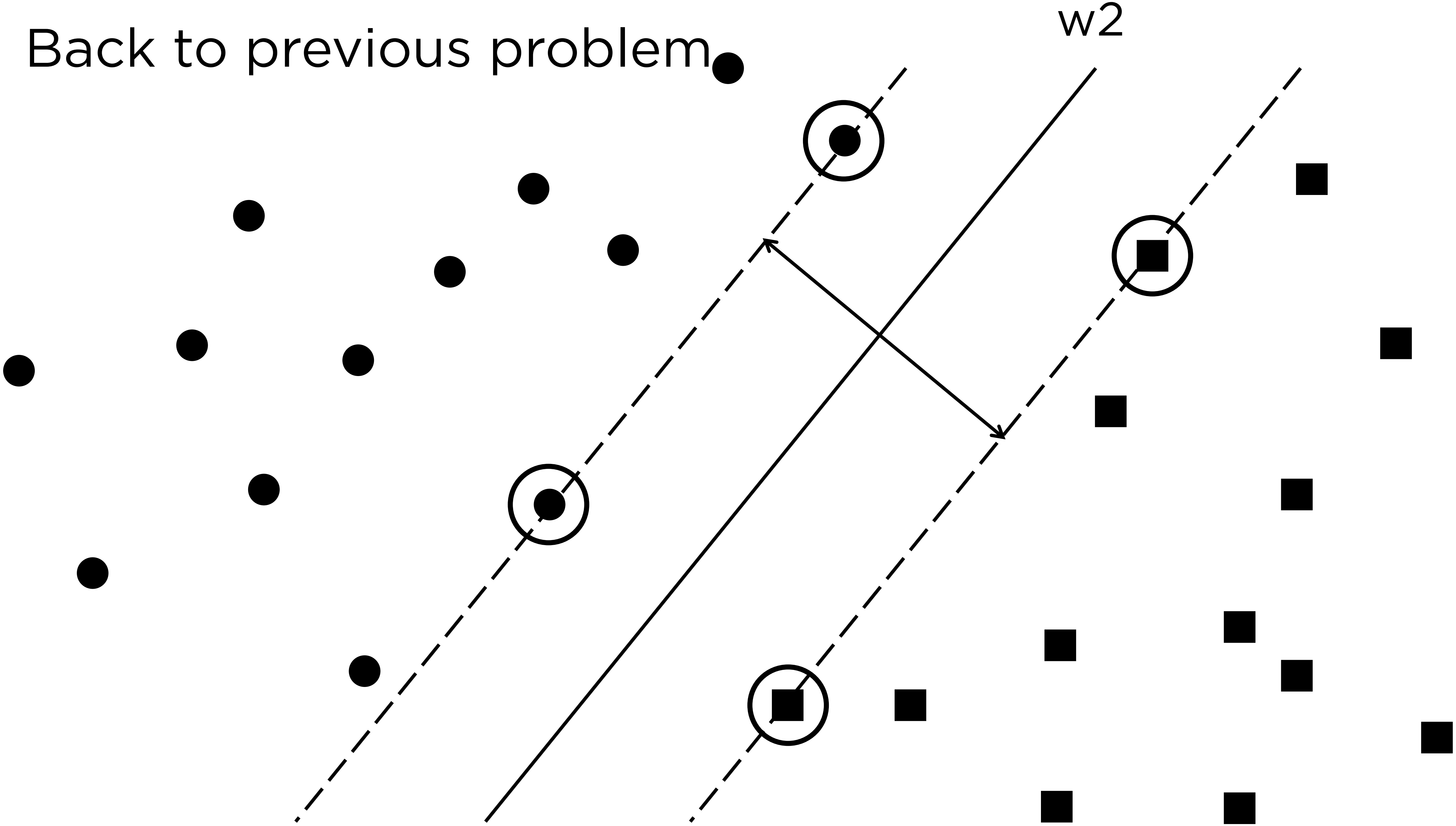
$$\rho_x(w, b) = \frac{|w^T x + b|}{||w||}$$

If $\mathcal{H}_{w,b}$ separates the training set $S$ we define its **margin** as:

$$\rho_x(w, b) = \min_{i=1:m} \rho_{x_i}(w, b)$$

If $\mathcal{H}_{w,b}$ is a hyperplane (separating or not) we also define the margin of a point $x$ as $w^T x + b$ (note that this can be positive )

Back to previous problem



# Hard Margin Support Vector Machine

## Optimal separating hyperplane (OSH)

The separating hyperplane with maximum margin can be solved with the following optimization problem

$$\rho(S) = \max_{w,b} \min_{i} \{ \frac{y_i(w^T x_i + b)}{||w||} : \ y_i(w^T x_i + b) \geq 0, \qquad i = 1, \ldots, m\} > 0$$

A separating hyperplane is parameterised by $(w, b)$, but this choice is not unique (rescaling with a positive constant gives the same separating hyperplane).

# Hard Margin Support Vector Machine

## Optimal separating hyperplane (OSH)

Two possible ways to fix the parameterisation:

- **Normalised hyperplane**: set $\|w\| = 1$, in which case $\rho_x(w, b) = |w^T x + b|$

  and $\rho_S(w, b) = \min_{i=1:m} y_i(w^T x_i + b)$

- **Canonical hyperplane**: choose $\|w\|$ such that $\rho_S(w, b) = \frac{1}{\|w\|}$, i.e. we

  require that $\min_{i=1:m} y_i(w^T x_i + b) = 1$ (a data –dependent parameterization)

We will mainly work with the second parameterisation and it is also the most common version of SVM mentioned in the literature.

# Hard Margin Support Vector Machine

## Optimal separating hyperplane (OSH)

Given the canonical hyperplane, we have

$$\rho(S) = \max_{w,b}\left\{\frac{1}{\|w\|}: \min_i\{y_i(w^T x_i + b)\} = 1, \qquad y_i(w^T x_i + b) \geq 0\right\} > 0$$

$$= \max_{w,b}\left\{\frac{1}{\|w\|}:\ y_i(w^T x_i + b) \geq 1\right\}$$

$$= \frac{1}{\min_{w,b}\{\|w\|: y_i(w^T x_i + b) \geq 1\}}$$

# Hard Margin Support Vector Machine

## Optimisation problem (primal form)

The problem thus can be defined as

Minimise $\frac{1}{2}w^T w$

Subject to $y_i(w^T x_i + b) \geq 1,\ i = 1, \dots, m$

# Hard Margin Support Vector Machine

## Saddle point

To determine the saddle point of the Lagrangian function

$$L(w, b; \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i \{y_i(w^T x_i + b) - 1\}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers

We minimise $L$ over $(w, b)$ and maximize over $\alpha$. Differentiating w.r.t $w$ and $b$ we obtain:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} y_i \alpha_i = 0$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

# Hard Margin Support Vector Machine

## Optimization problem (dual form)

Substituting the solution of $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ leads to the dual problem

Maximise $\quad \mathrm{Q}(\alpha) = -\frac{1}{2}\alpha^T A \alpha + \sum_{i=1}^{m} \alpha_i$

Subject to $\quad \sum_{i=1}^{m} y_i \alpha_i = 0$

$\alpha_i \geq 0,\ i = 1, \dots, m$

where $A$ is an $m \times m$ matrix $A = (y_i y_i x_i^T x_j : i, j = 1, \dots, m)$

**Note the complexity of this problem depends on $\boldsymbol{m}$, not on the number of data point dimension $\mathbb{R}^{\boldsymbol{d}}$.**

# Hard Margin Support Vector Machine

## Karush-Kuhn-Tucker conditions and support vectors

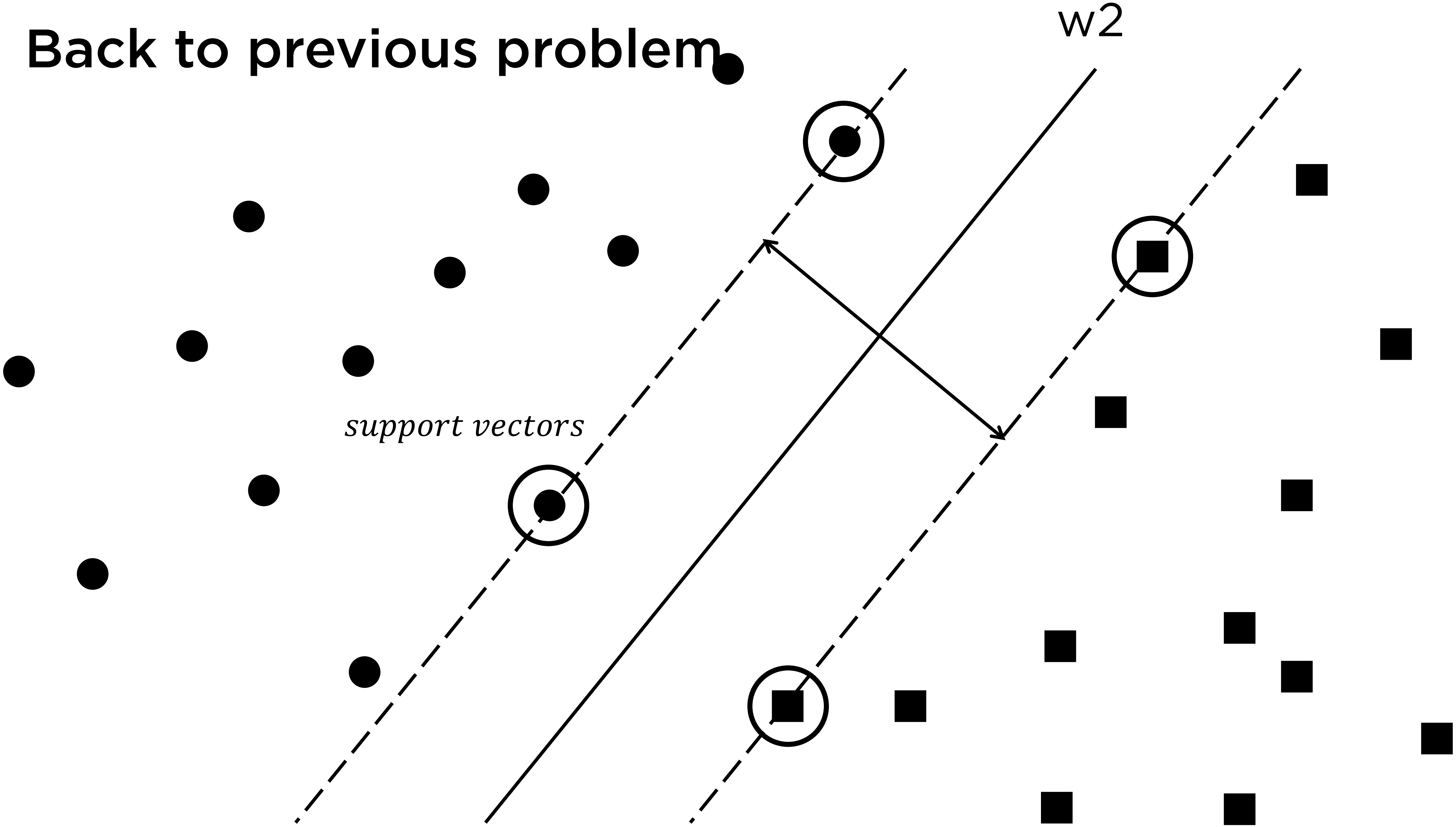
In optimisation theory, the dual form solution ($d^*$) serves as the lower bound of the primal solution ($p^*$). i.e. $d^* \leq p^*$

When $d^* = p^*$, the solution satisfy Karush-Kuhn-Tucker (KKT) condition and we have the maximum margin for SVM constructed by **support vectors.** The optimal solution is given by

$$\bar{w} = \sum_{i=1}^{m} \bar{\alpha}_i y_i x_i$$

This $\bar{w}$ is a linear combination of only the $x_i$ for which $\bar{\alpha}_i > 0$. These $x_i$ are termed **support vectors.**

# Back to previous problem



# Hard Margin Support Vector Machine

## Some conclusions

- The most remarkable fact about OSH is that it is determined only by support vectors, which is usually a subset of the training data
- All the information contained in the data points is summarized by the support vectors: The whole data set could be replaced by only these points and the same hyperplane would be found
- A new point $x$ is classified as $sgn(\sum_{i=1}^{m} y_i \bar{\alpha}_i {x_i}^T x + \bar{b})$

# Linear Separable SVM

- Hard Margin Support Vector Machine
- Soft Margin Support Vector Machine

# Soft Margin Support Vector Machine

## Motivation

- In idea cases, we show that if data is completely linearly separable without any errors (noise or outliers). Support Vector Machine is an efficient algorithm for these data points.
- However, what if the data is not strictly linearly separable?

# Soft Margin Support Vector Machine

If the data is not linearly separable, the previous analysis can be generalized as the following problem:

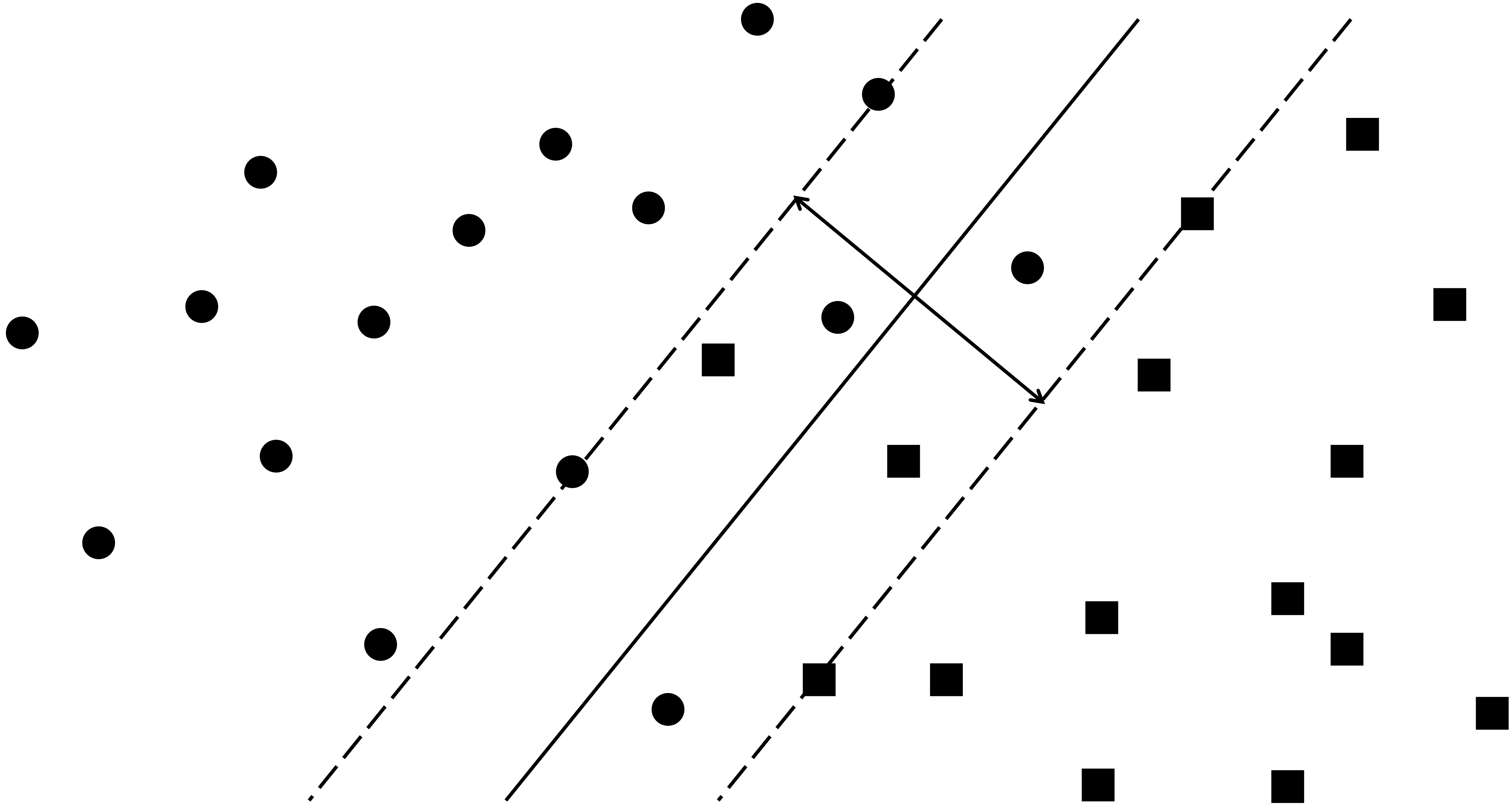
Minimise $\quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$

Subject to $\quad y_i(w^T x_i + b) \geq 1 - \xi_i,$

$$\xi_i \geq 0, \qquad i = 1, \dots, m$$

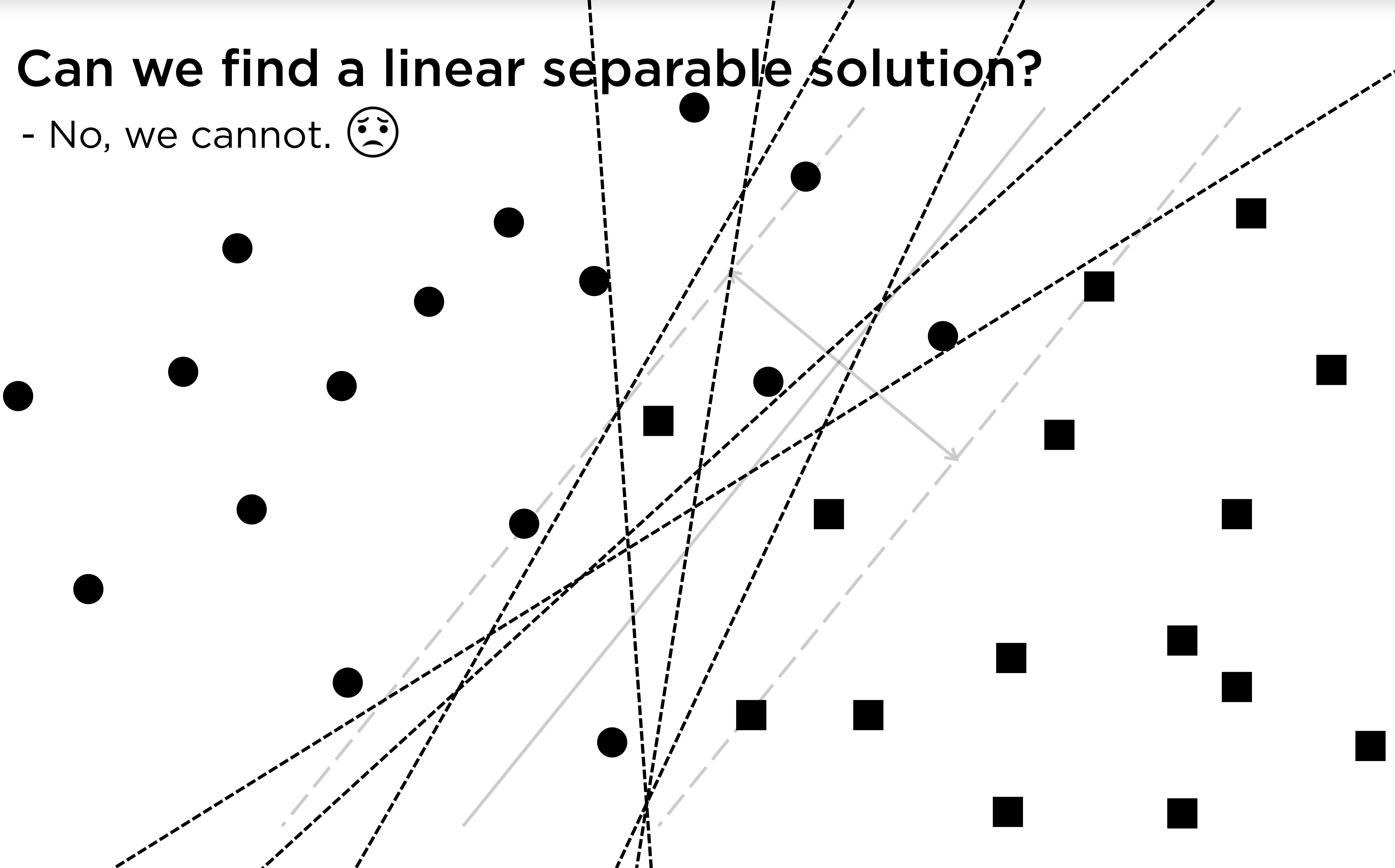
The idea is to introduce the **slack variable** $\xi_i$ to relax the separation constraints ($\xi_i > 0$)

How do we relax the constrains?

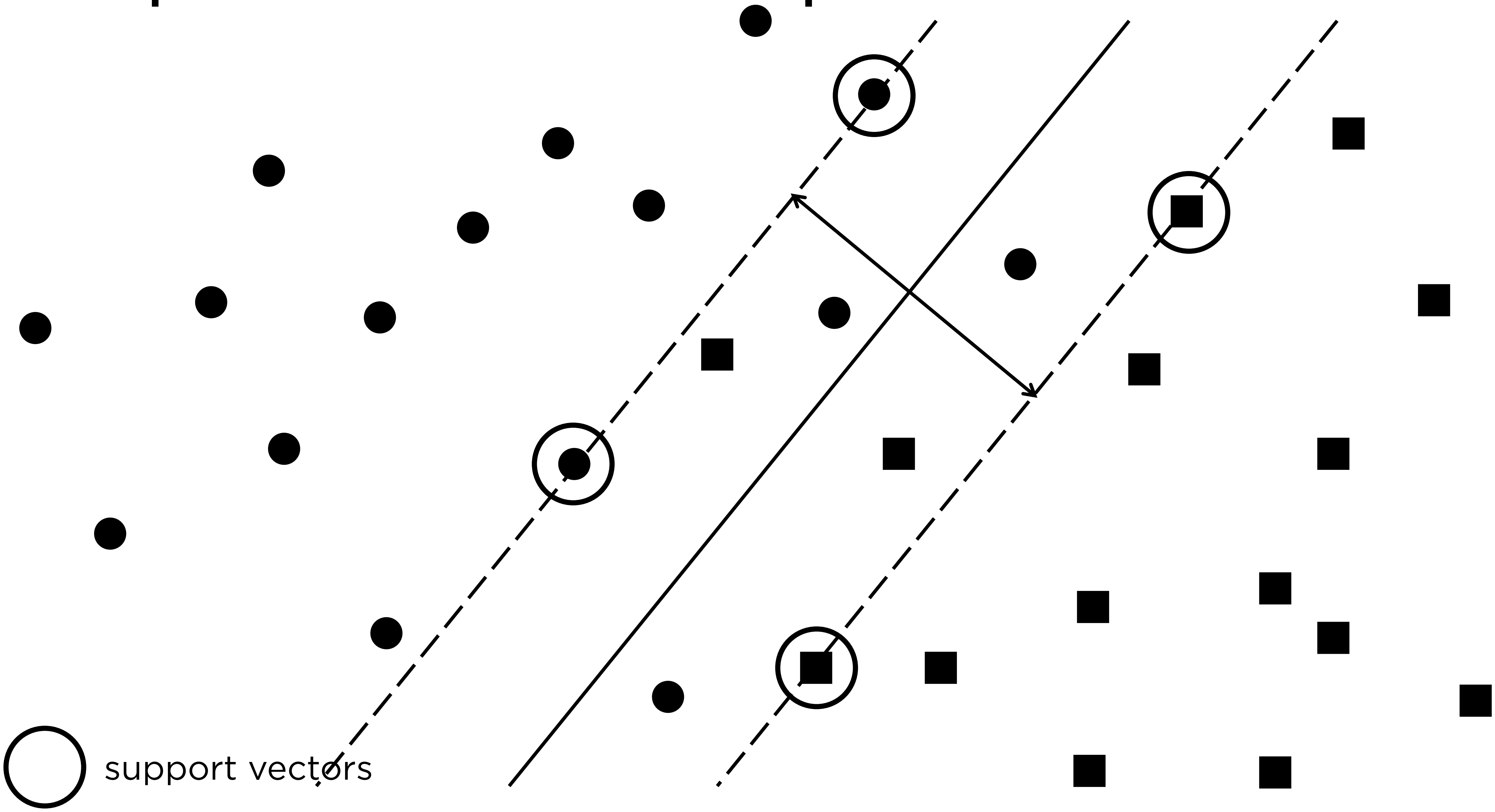


# Can we find a linear separable solution?

- No, we cannot. 😟

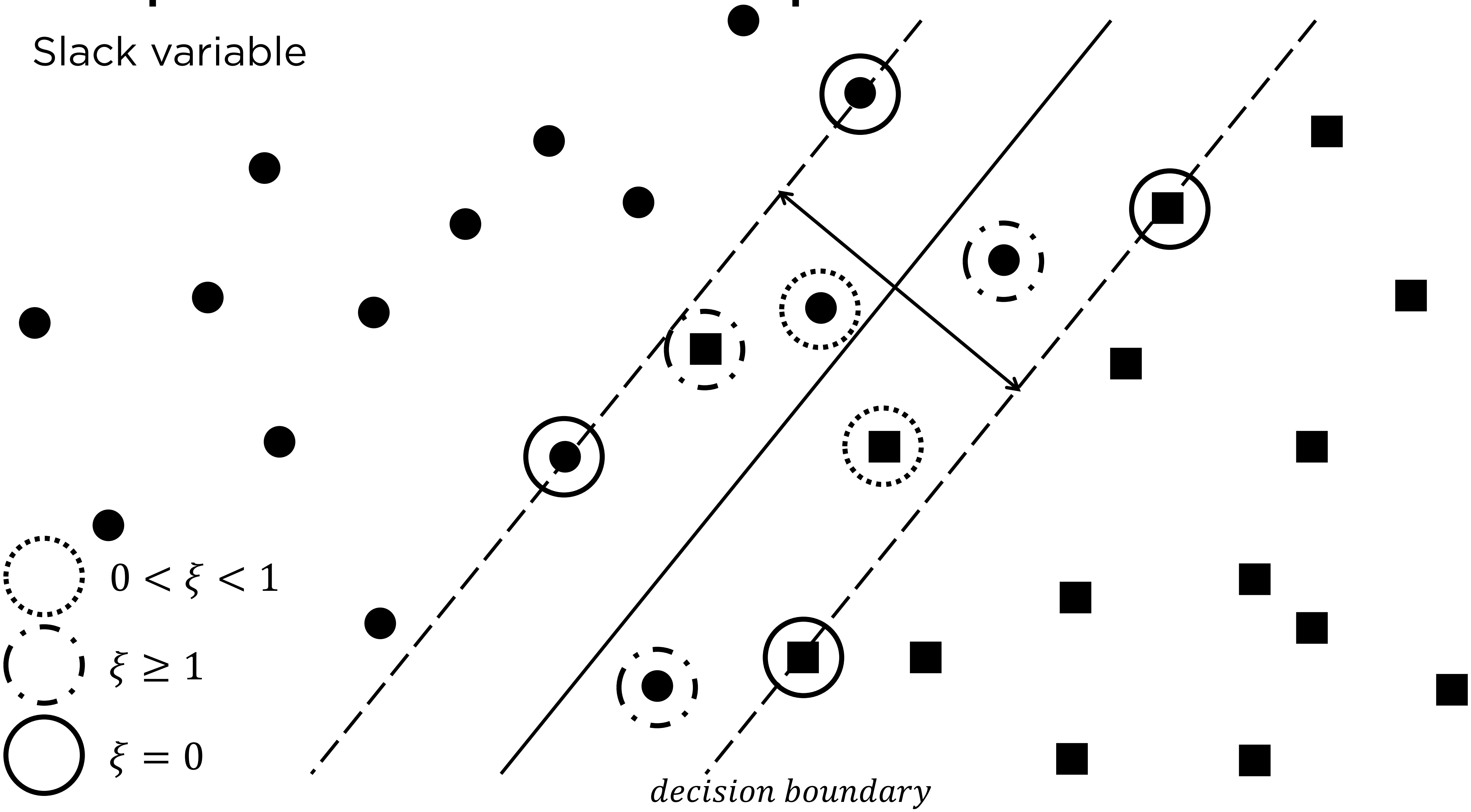


# Old problem with new data points



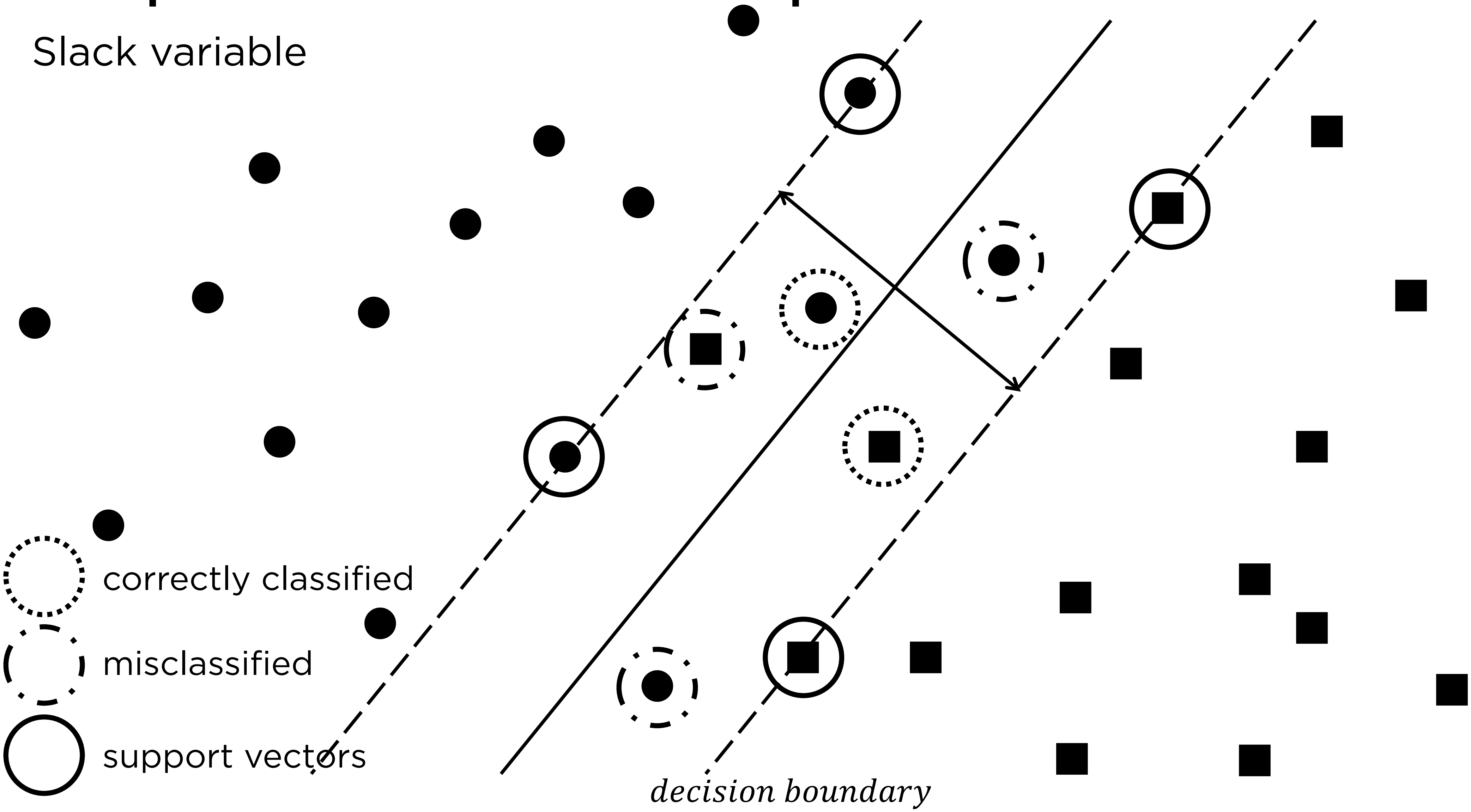
# Old problem with new data points

Slack variable



# Old problem with new data points

Slack variable



# Soft Margin Support Vector Machine

## The role of the parameter $C$

$C$ is a regularization parameter:

- Small $C$ allows constraints to be easily ignored, hence results in large margin.
- Large $C$ makes constraints hard to ignore, hence results in narrow margin.
- When $C = \infty$, it enforces all constraints to become hard margin problem.

# Today

- Linear Separable SVM
- Non-Linear Separable SVM

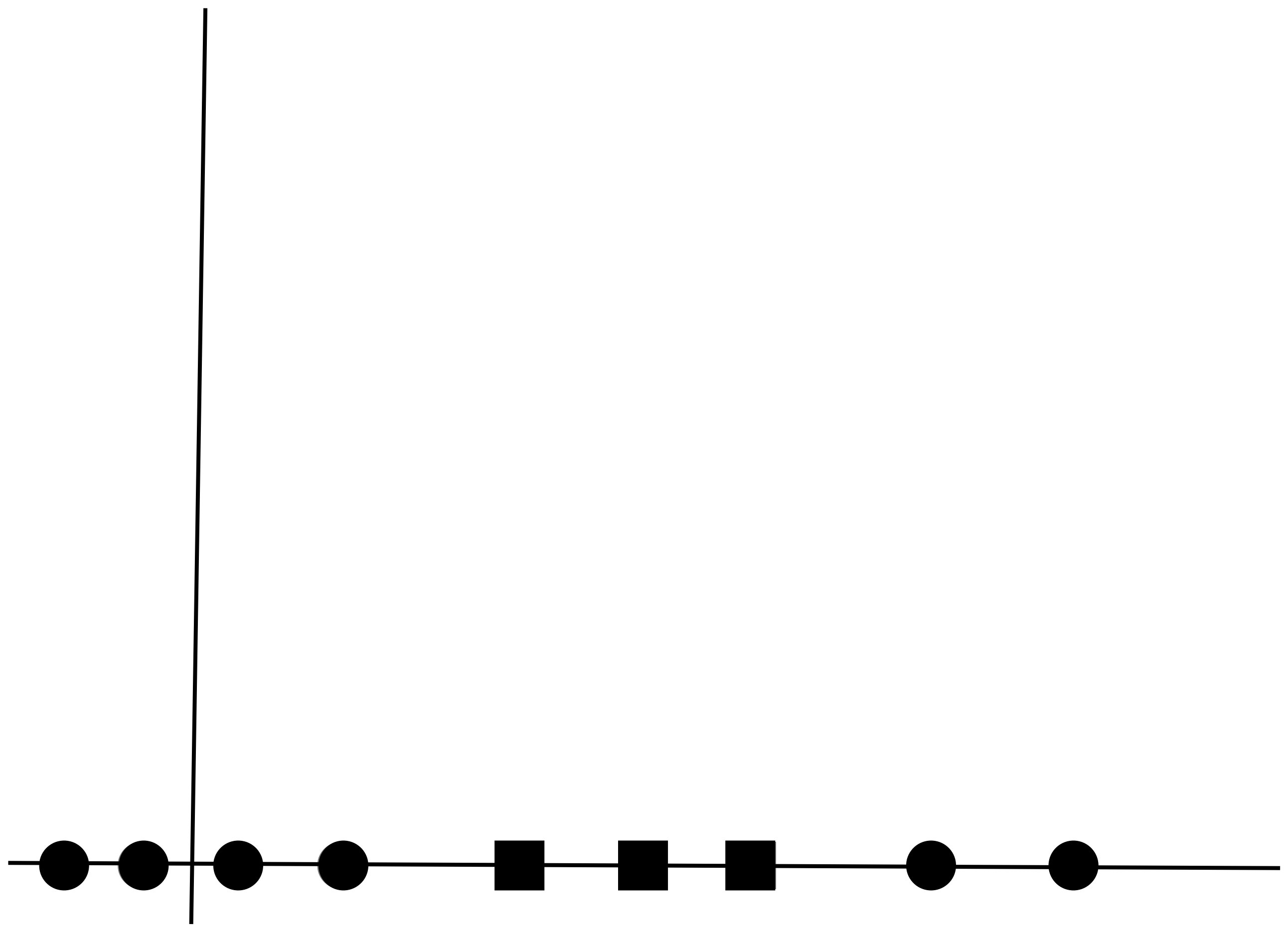
# Non-Linear Separable SVM

# Soft Margin Support Vector Machine

## Hard 1-dimensional Dataset

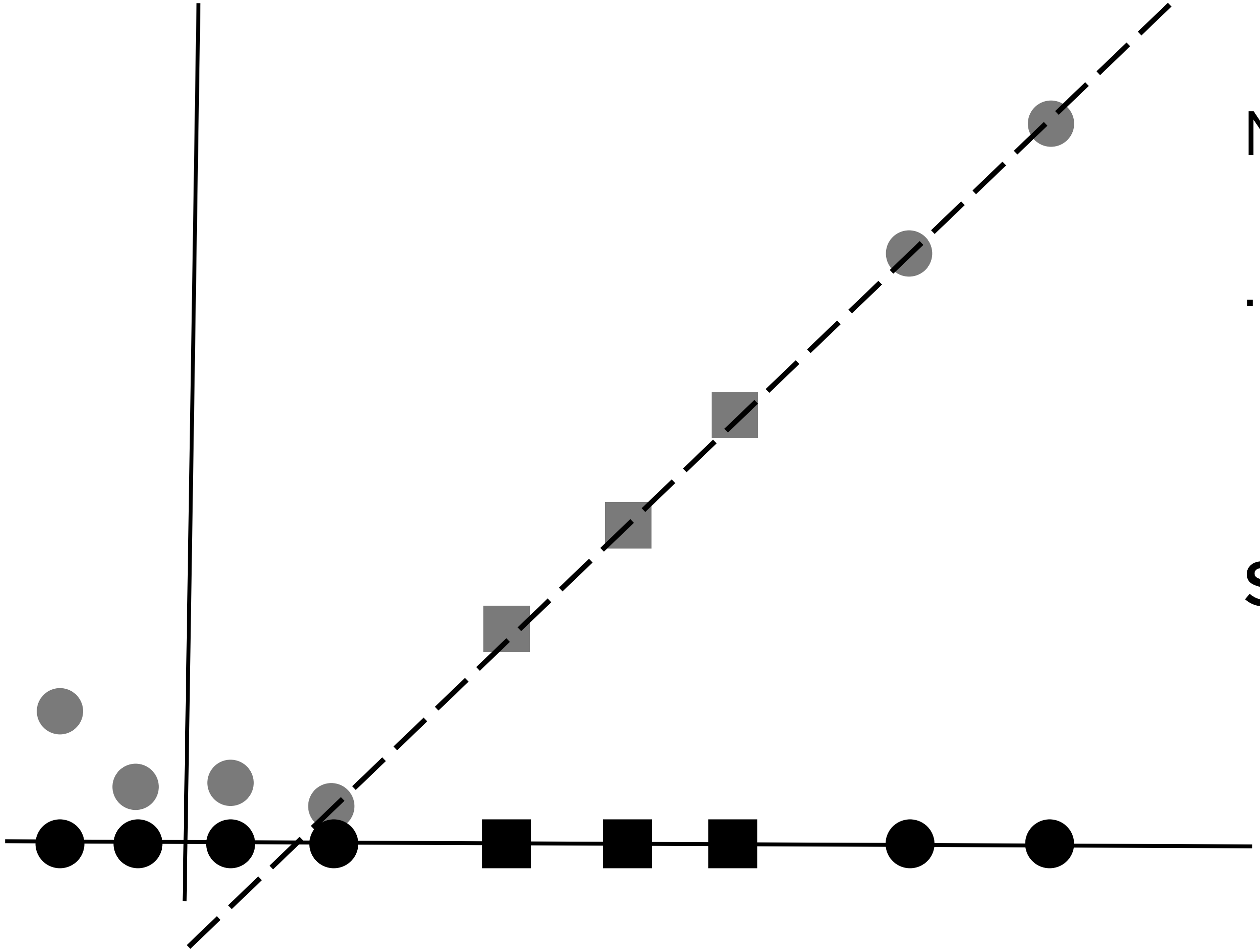
# Soft Margin Support Vector Machine

## Hard 1-dimensional Dataset Solution



# Soft Margin Support Vector Machine

## Hard 1-dimensional Dataset Solution



Make up a new feature!

…Computed from original feature(s)

$$y_k = (x_k, {x_k}^2)$$

**Separable.** ☺

# Soft Margin Support Vector Machine

## Feature Vector and Feature Space

Let $\vec{x} = [\overrightarrow{x_1}] \in \mathbb{R}$ be a vector representation of object $x \in X$

Let $\Phi: X \longrightarrow K \in \mathbb{R}^2$ feature map given by

$$\Phi(\vec{x}) = [\overrightarrow{x_1}, \overrightarrow{x_1}^2] \in \mathbb{R}^2$$

$K$ is refers to as **feature space** and $\Phi(\vec{x}) = [\Phi_1(\vec{x}), \Phi_2(\vec{x})]$, vector $\Phi(\vec{x})$ is called the **feature vector.**

# Soft Margin Support Vector Machine

## Feature Map and Kernel Function

A **feature map** refers to a function $\Phi: \mathbb{R}^n \longrightarrow \mathbb{R}^N$

$$\Phi(\vec{x}) = \left(\Phi_1(\vec{x}), \dots, \Phi_N(\vec{x})\right)^T, \vec{x} \in \mathbb{R}^n$$

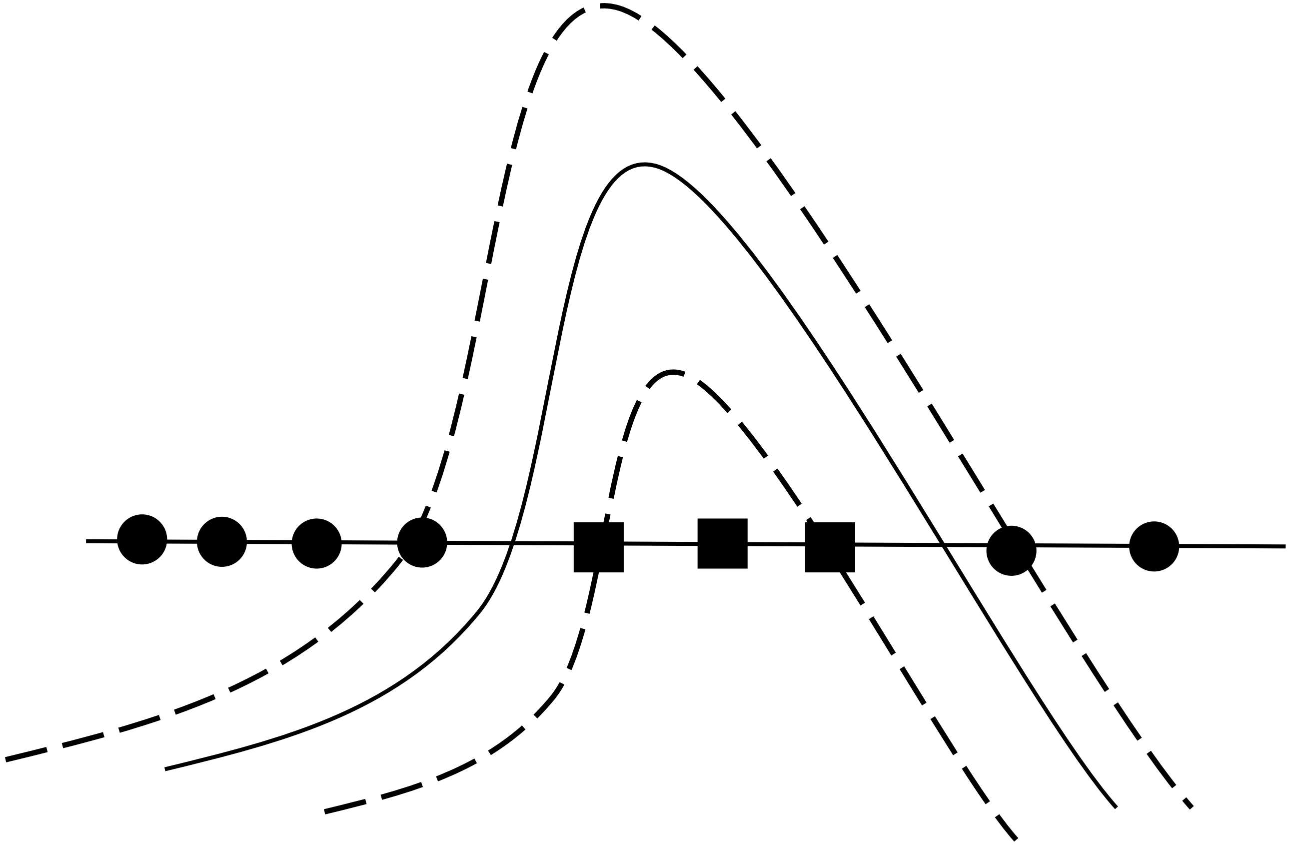
The $\Phi_1, .., \Phi_N$ are called **basis functions**, given a feature map $\Phi$

we define its associated kernel function $\mathrm{K}: \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ as

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{t}) =< \Phi(\boldsymbol{x}), \Phi(\boldsymbol{t}) >, \qquad \boldsymbol{x}, \boldsymbol{t} \in \mathbb{R}^n$$

# Soft Margin Support Vector Machine

**What is the decision boundary for this data?**



The least polynomial degree equal to 2 for this example since that is the highest polynomial degree in our basis functions, we can also called it quadratic kernels

# Today

- Linear Separable SVM
  - Hard Margin SVM
  - Soft Margin SVM
- Non-Linear Separable SVM

Questions?

# Further Reading

The Elements of Statistical Learning Chapter 6 and Chapter 12

https://web.stanford.edu/~hastie/Papers/ESLII.pdf