Support Vector Machine

Announcements

1. Prelim Conflict form is going out soon

2. Prelim practice: we will release previous semesters' prelims w/ solutions

3. HW4 will be out today, P4 will be out Thursday

Goal for today

Understand the Support Vector Machine (SVM) — a turnkey classification algorithm

Outline for Today

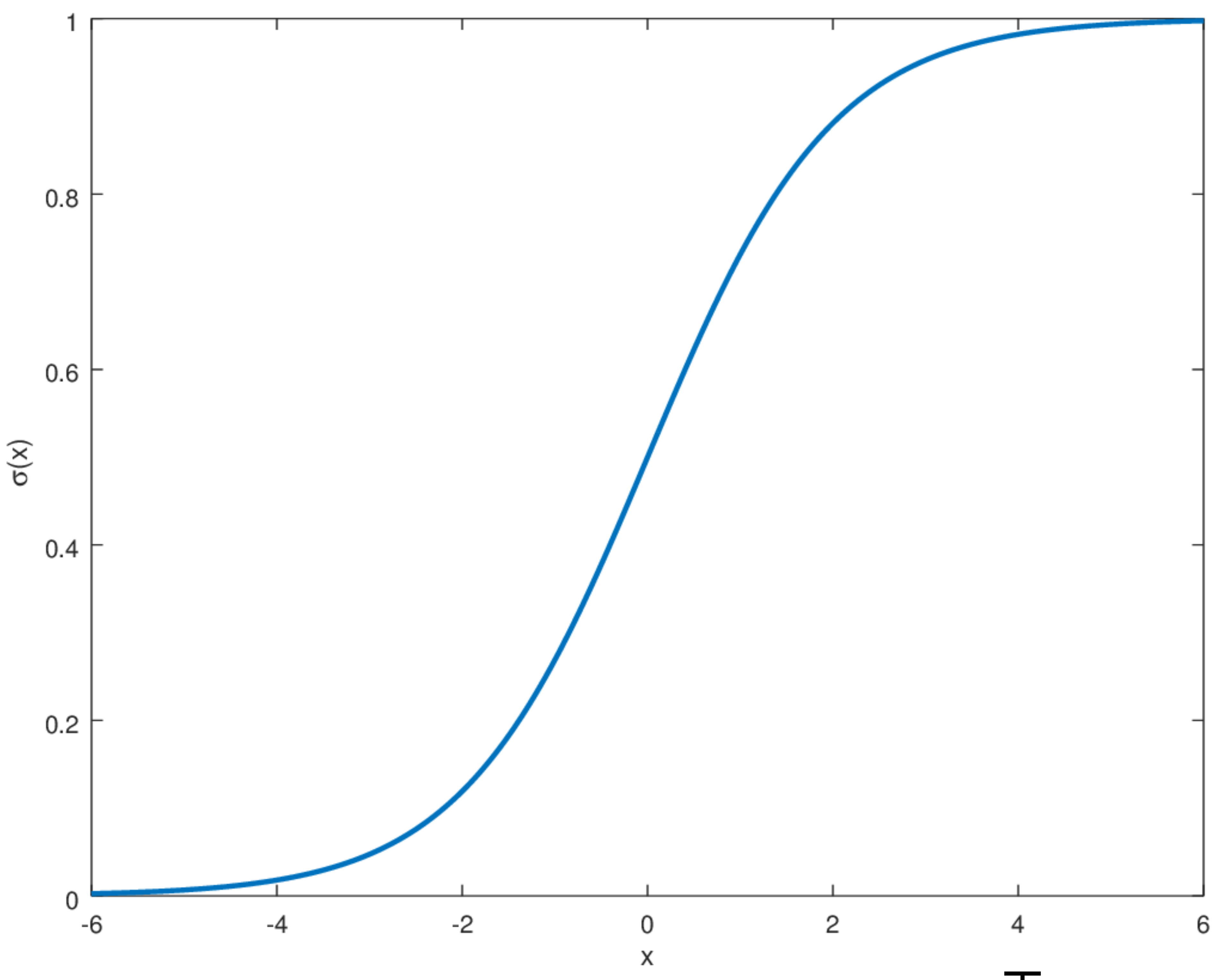
1. Functional Margin & Geometric Margin

2. Support Vector Machine for separable data

3. SVM for non-separable data

Recall Logistic Regression

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



$z := y(w^\top x + b)$

Given (x, y), our model predict label y, if

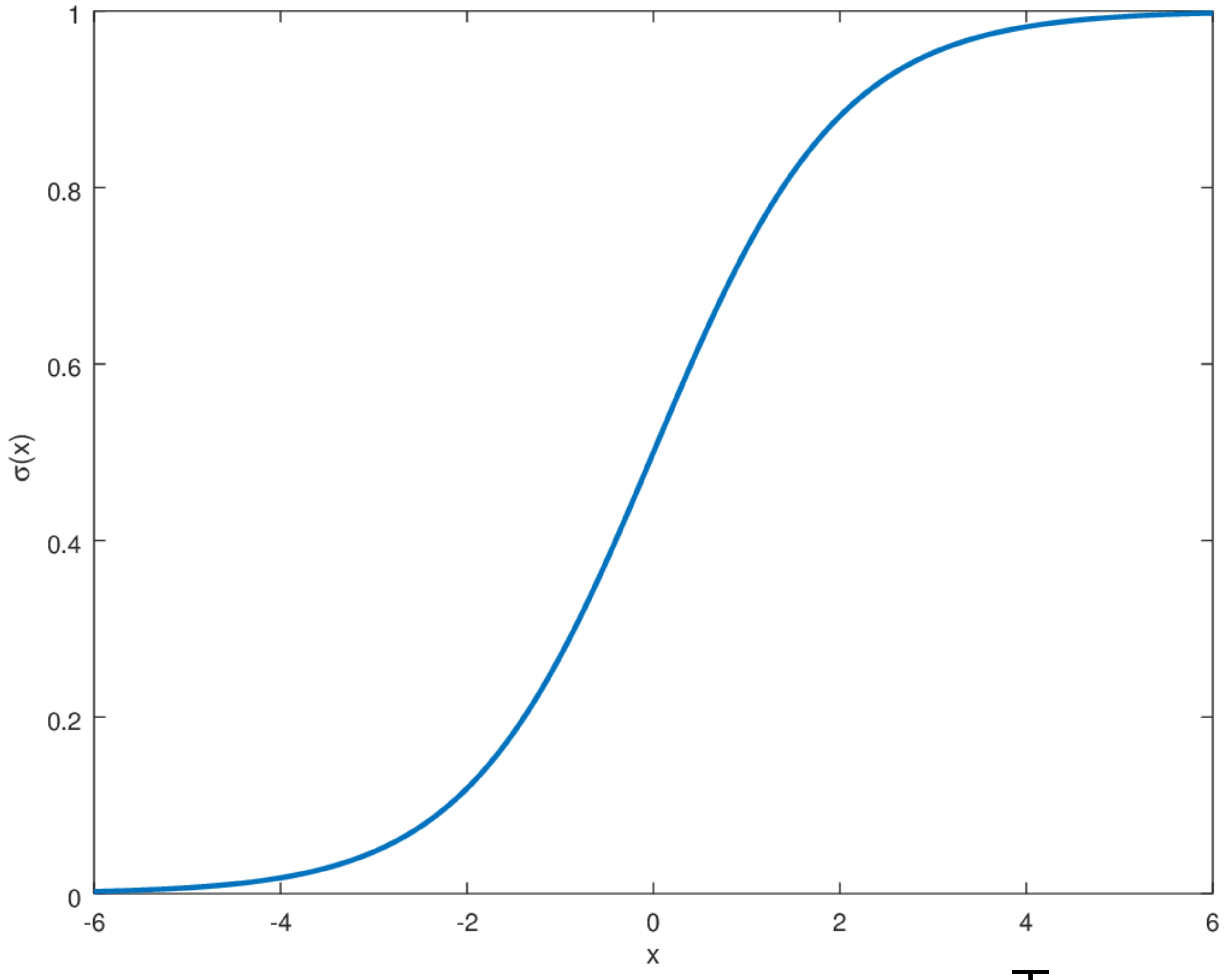
$P(y \mid x; w, b) > 0.5$, or equivalently $y(w^\top x + b) > 0$

Larger $y(w^\top x + b)$ $\rightarrow$ larger $P(y \mid x; w, b)$

Functional margin
"confidence"

Recall Logistic Regression

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



$z := y(w^\top x + b)$

A good classifier should have large functional margin on training examples:

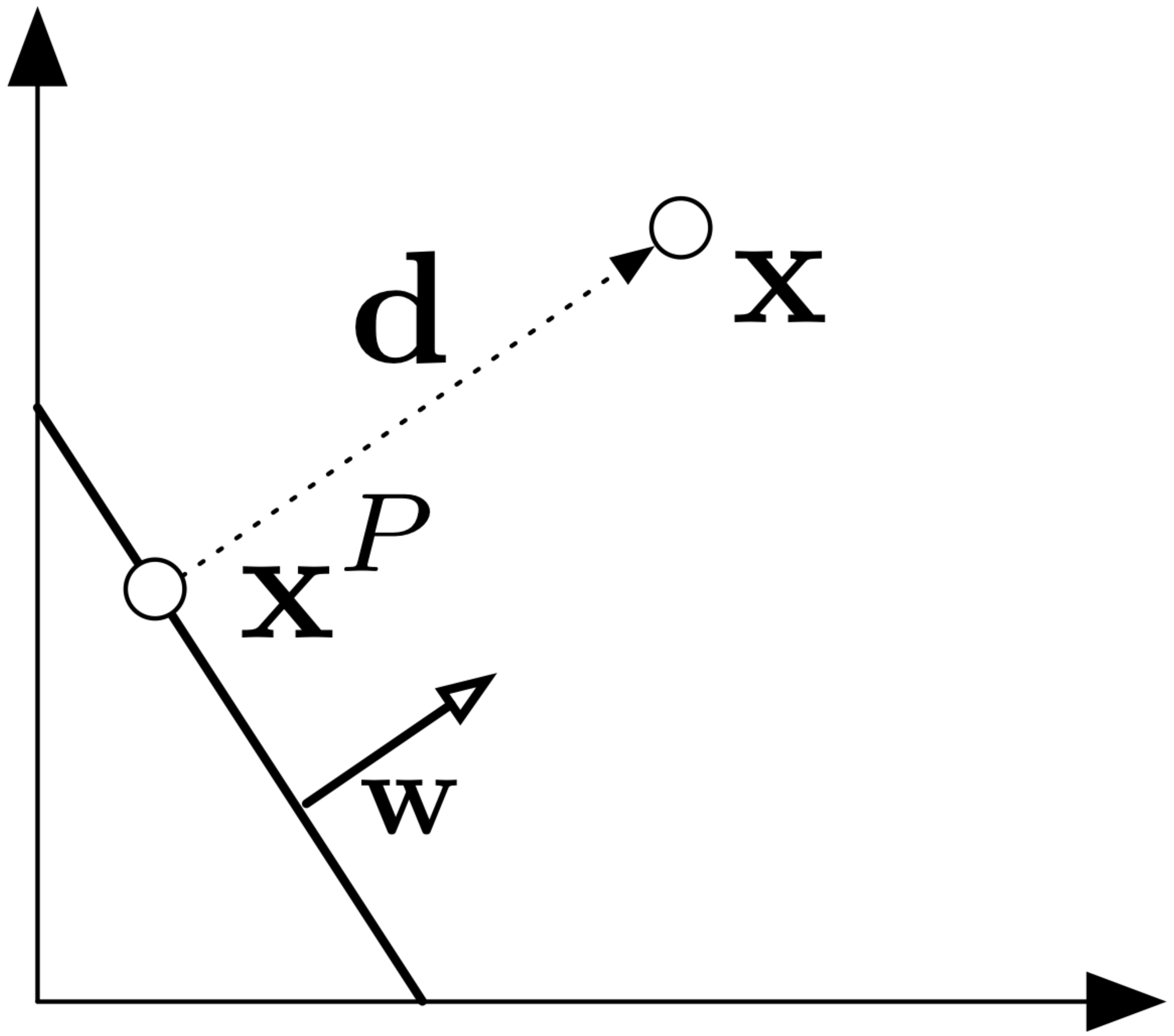
For all (x_i, y_i), $y_i(w^\top x_i + b) \gg 0$

However, functional margin is NOT scale-invariant:

Consider $(2w, 2b)$: functional margin is doubled

Geometric Margin

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

Fact 2. x^p is on the hyperplane:

$$w^\top x^P + b = 0$$

Fact 1 + fact 2 implies:

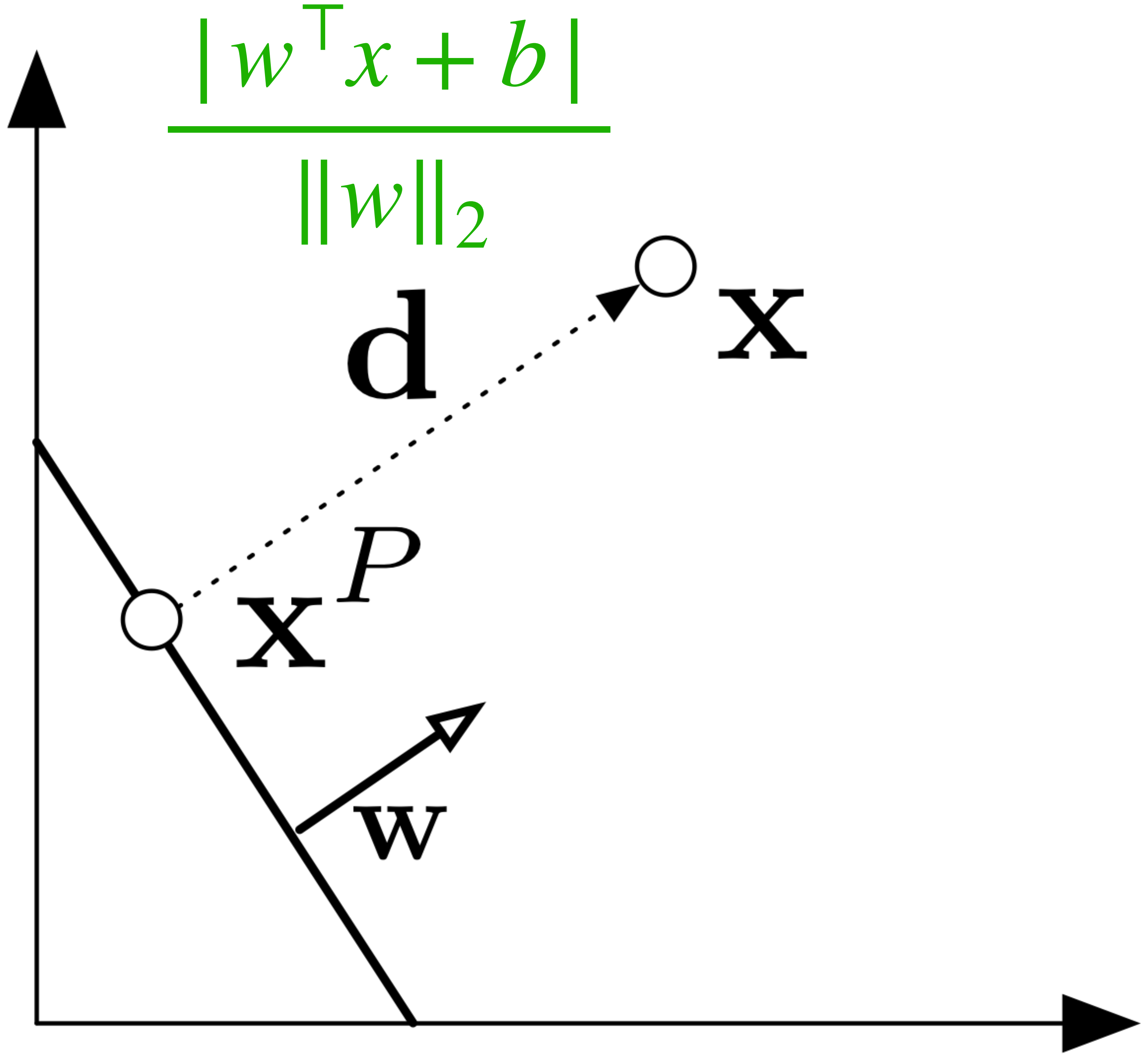
$$w^\top (x - \alpha w) + b = 0 \;\rightarrow\; \alpha = (w^\top x + b)/\|w\|_2^2$$

Final step:

$$d = \|x - x^p\|_2 = \|\alpha w\|_2 \qquad = \frac{|w^\top x + b|}{\|w\|_2}$$

Geometric Margin is Scale Invariant

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



We scale (w, b) by a constant $\gamma \in \mathbb{R}^+$

Q: is the hyperplane defined by $(\gamma w, \gamma b)$ different?

Q: does the margin change?

Hyperplane & Geometric margin are **scale invariant**!

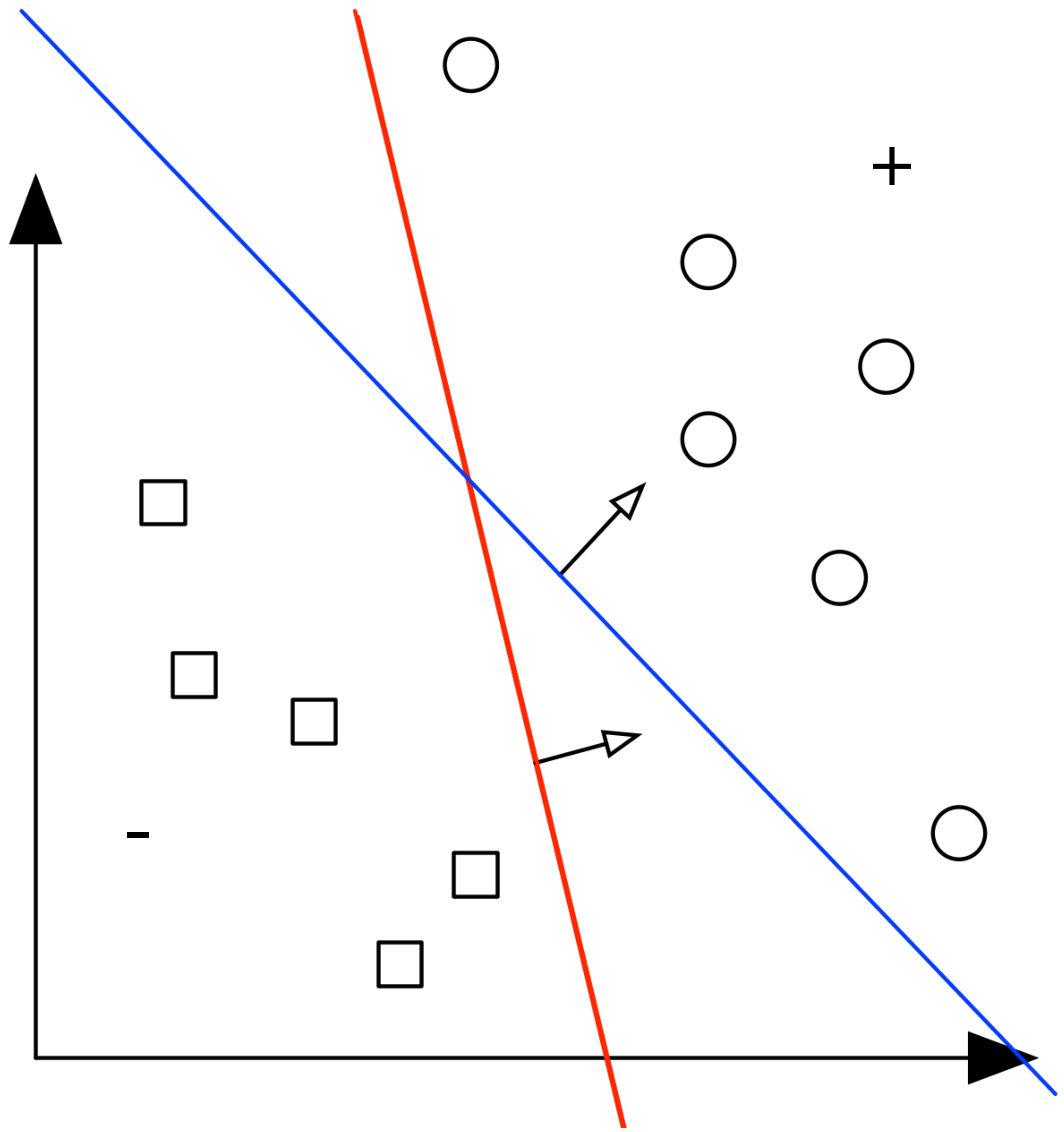
Outline for Today

1. Functional Margin & Geometric Margin

2. Support Vector Machine for separable data

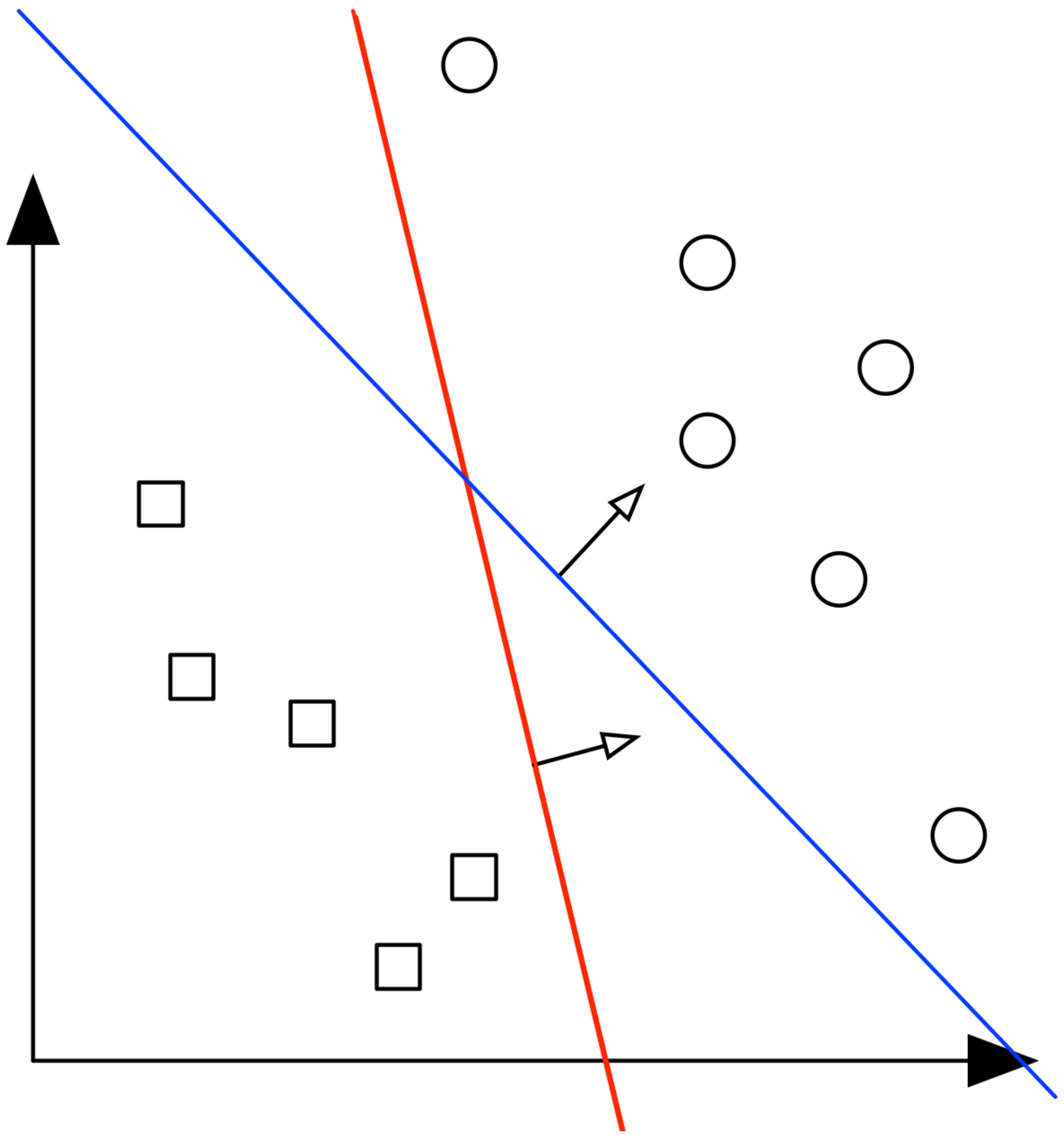
3. SVM for non-separable data

Which linear classifier is Better?



Both hyperplanes correctly separate the data

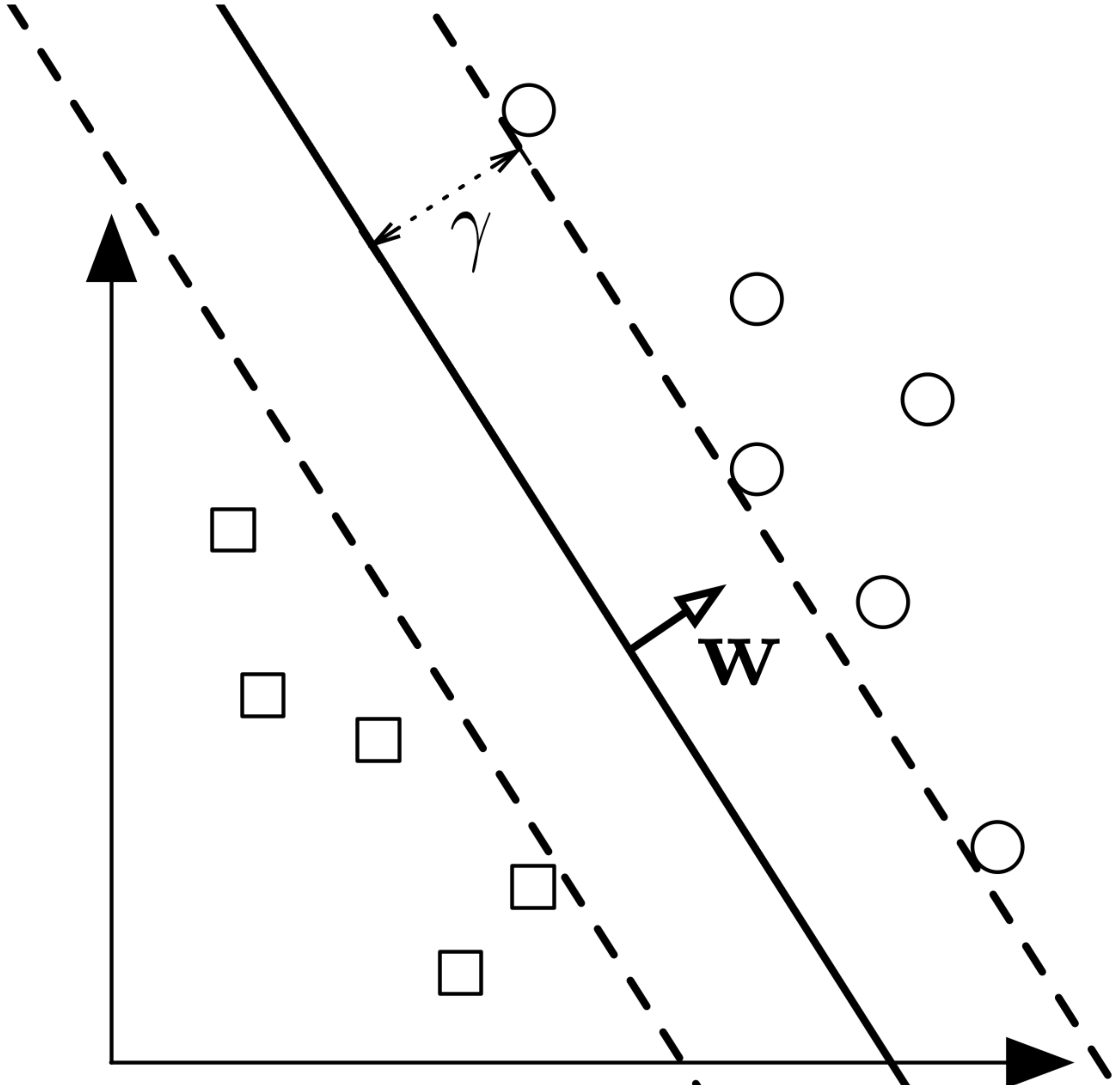
Max Margin Classifier



The Goal of SVM:

Find a hyperplane that has the largest Geometric margin

Max Margin Classifier

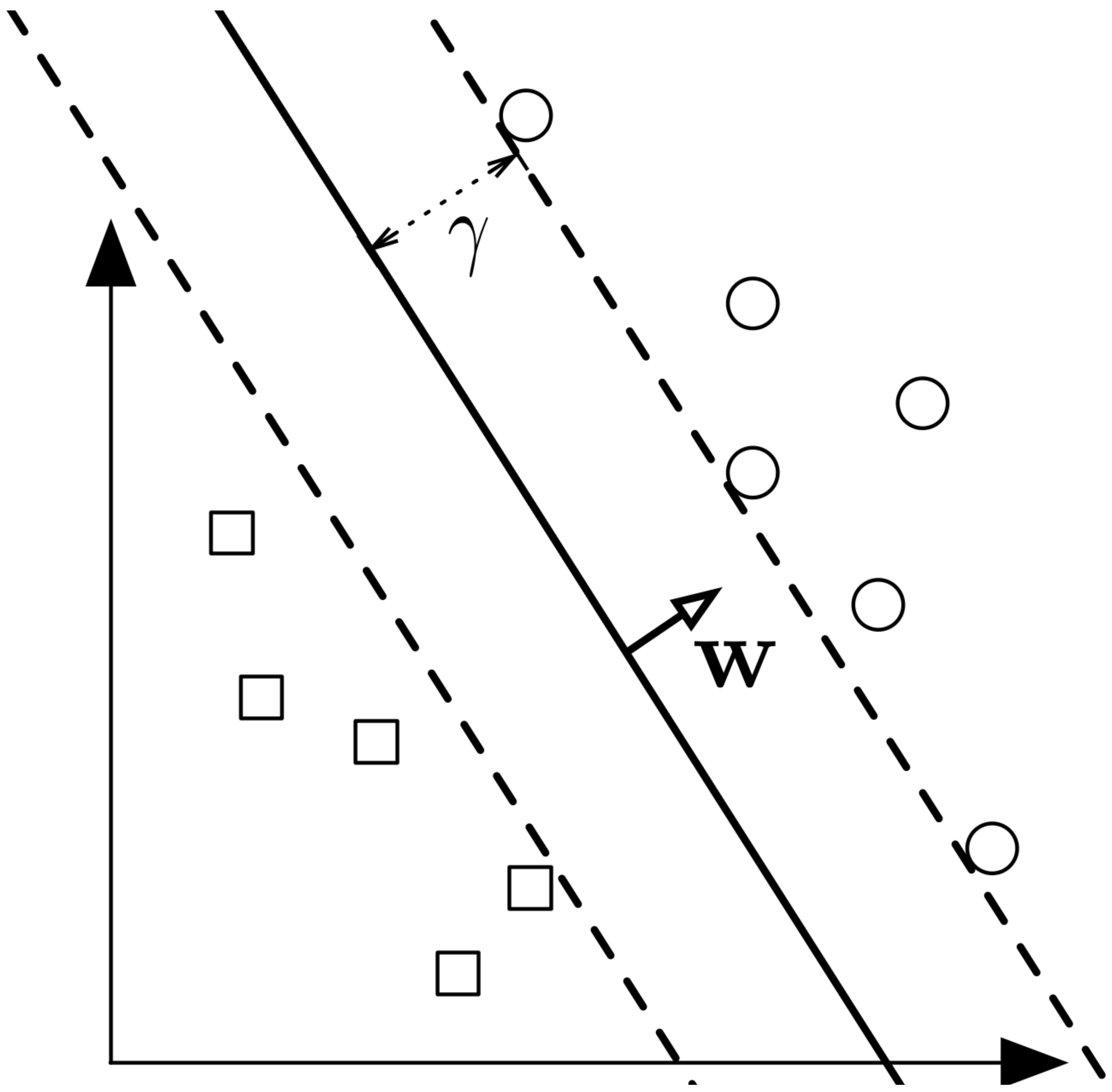


Given a linearly separable dataset $\{x_i, y_i\}_{i=1}^n$, the minimum geometric margin is defined as

$$\gamma(w, b) := \min_{x_i \in \mathcal{D}} \frac{|x_i^\top w + b|}{\|w\|_2}$$

Goal: we want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

Max Margin Classifier



We want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

$$\max_{w,b} \gamma(w, b)$$

$$\text{s.t. } \forall i,\ y_i(w^\top x_i + b) \geq 0$$

Plug in the def of $\gamma(w, b)$:

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} |w^\top x_i + b|$$

$$\text{s.t. } \forall i,\ y_i(w^\top x_i + b) \geq 0$$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} |w^\top x_i + b|$$

s.t. $\forall i, \; y_i(w^\top x_i + b) \geq 0$

$$\min_i |w^\top x_i + b| = 1$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have $\min_{x_i} |w^\top x_i + b| = 1$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} |w^\top x_i + b| = 1$

SVM for separable data: Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

$$\text{s.t.} \, \forall i, \; y_i(w^\top x_i + b) \geq 0$$

$$\min_i |w^\top x_i + b| = 1$$

We can further simplify the constraint

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \;\; y_i(w^\top x_i + b) \geq 1$$

You will prove that in HW4!

Summary for Max Margin Classifier

Avoids "cheating" (i.e., scale w, b up by large constant)

$$\boxed{\min_{w,b} \|w\|_2^2}$$

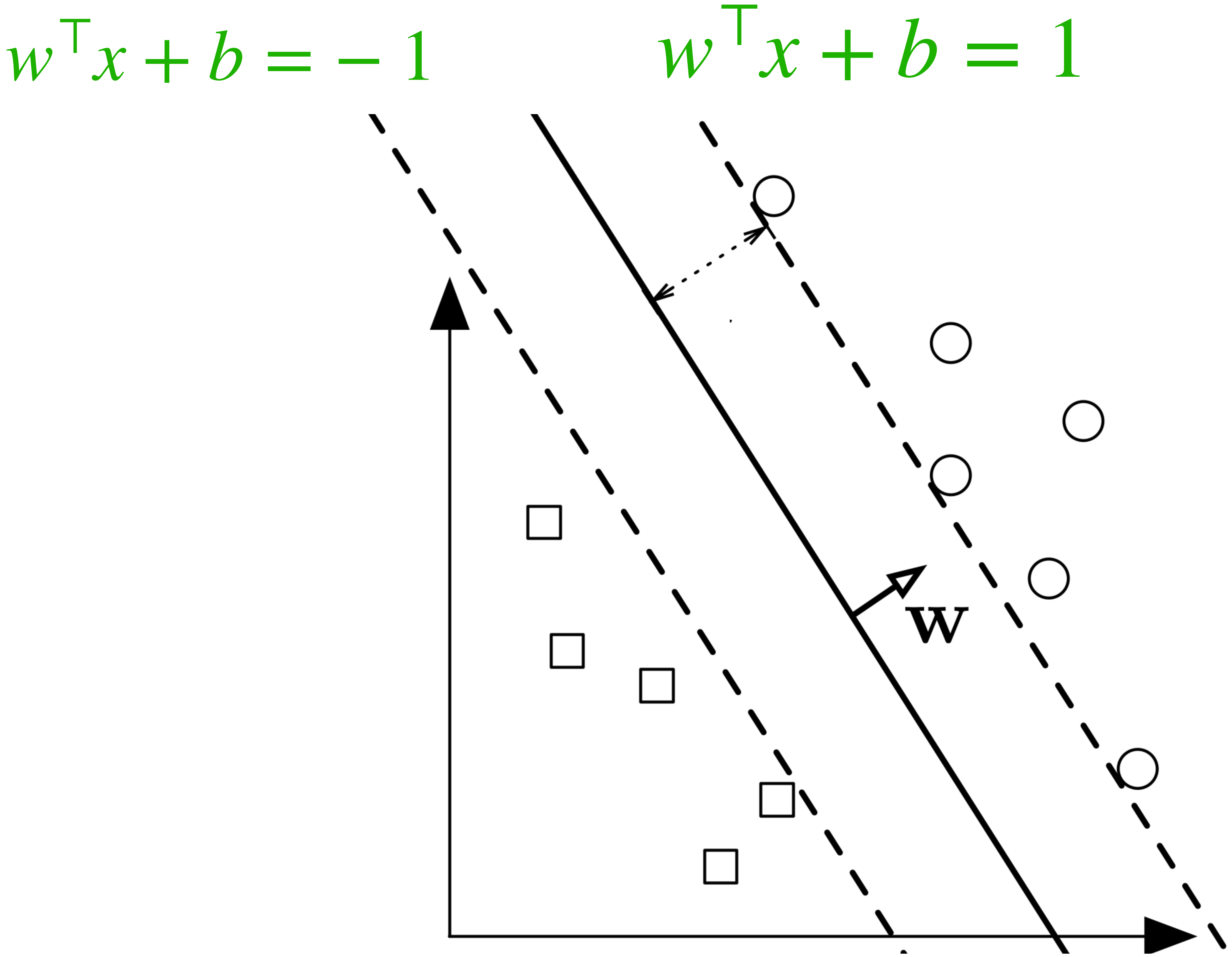
$$\boxed{\forall i: \ y_i(w^\top x_i + b) \geq 1}$$

Not only linearly separable, but also has functional margin no less than 1

Always remember **where we started**:

We want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

Support Vectors



for the optimal (w, b) pair, points x_i such that $y_i(w^\top x_i + b) = 1$ are called **support vectors**

Outline for Today

1. Functional Margin & Geometric Margin

2. Support Vector Machine for separable data

3. SVM for non-separable data

SVM for non-separable data

If data is not linearly separable, then **there is no** (w, b) can satisfy $\forall i: \; y_i(w^\top x_i + b) \geq 1$

Idea: introducing slack variables to relax the constraint, i.e., find (w, b, ξ_i), s.t,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \forall i$$

Q: does this always has a feasible solution?

SVM for non-separable data

Idea: introducing slack variables to relax the constraint, i.e., find (w, b, ξ_i), st,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

We still want our margin to be somewhat large, i.e., we want slack variables to be as small as possible

$$\min_{w,b,\xi} \|w\|_2^2 + c\sum_{i=1}^{n} \xi_i$$

Penalizing large slacks

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

SVM for non-separable data

Penalizing large slacks

$$\min_{w,b,\xi_i} \|w\|_2^2 + c\sum_{i=1}^{n} \xi_i$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

We can turn this constrained opt to a unconstraint opt w/ a single objective.

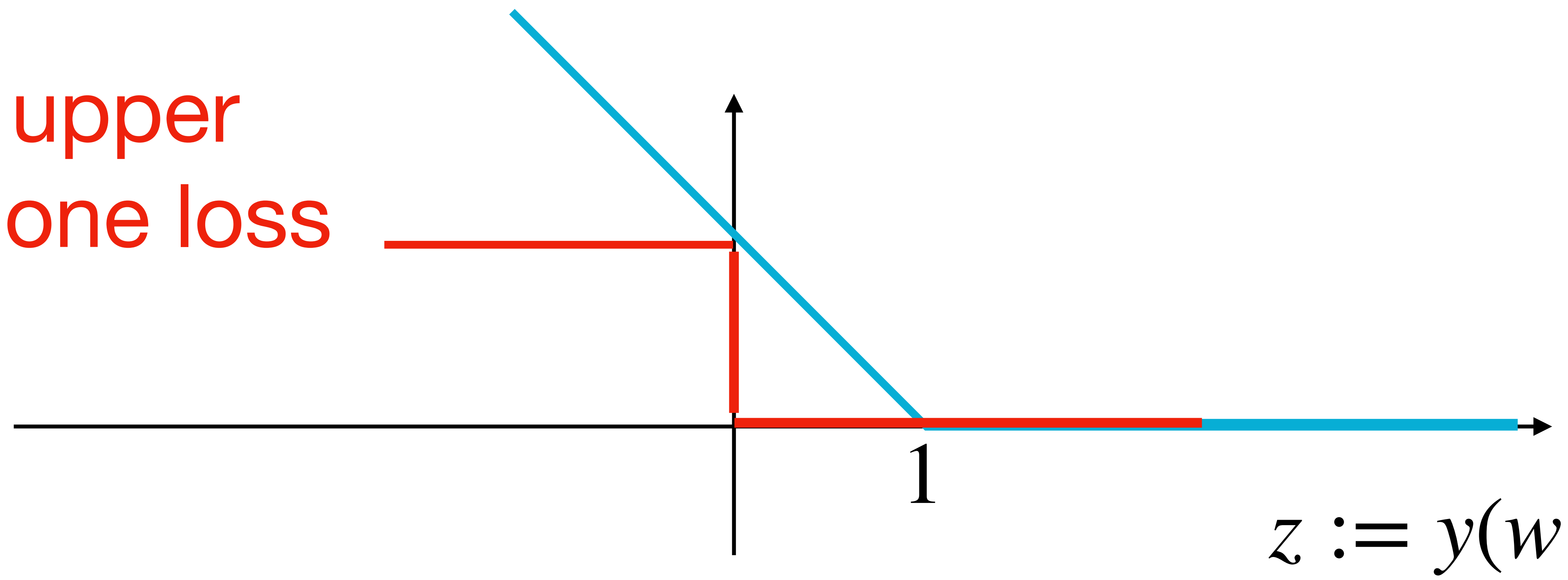
Q: For any fixed (w, b) pair, how to set ξ_i, such that the obj is minimized?

A: set $\xi_i = \max\{0, 1 - y_i(w^\top x_i + b)\}$

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \boxed{\max\left\{0, 1 - y_i(w^\top x_i + b)\right\}}$$

Hinge loss



Hinge loss upper bounds zero-one loss

Hinge loss starts penalizing when functional margin falls below 1

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

When $c \to +\infty$:

forcing $y_i(w^\top x_i + b) \geq 1$ for as many data points as possible

When $c \to 0^+$:

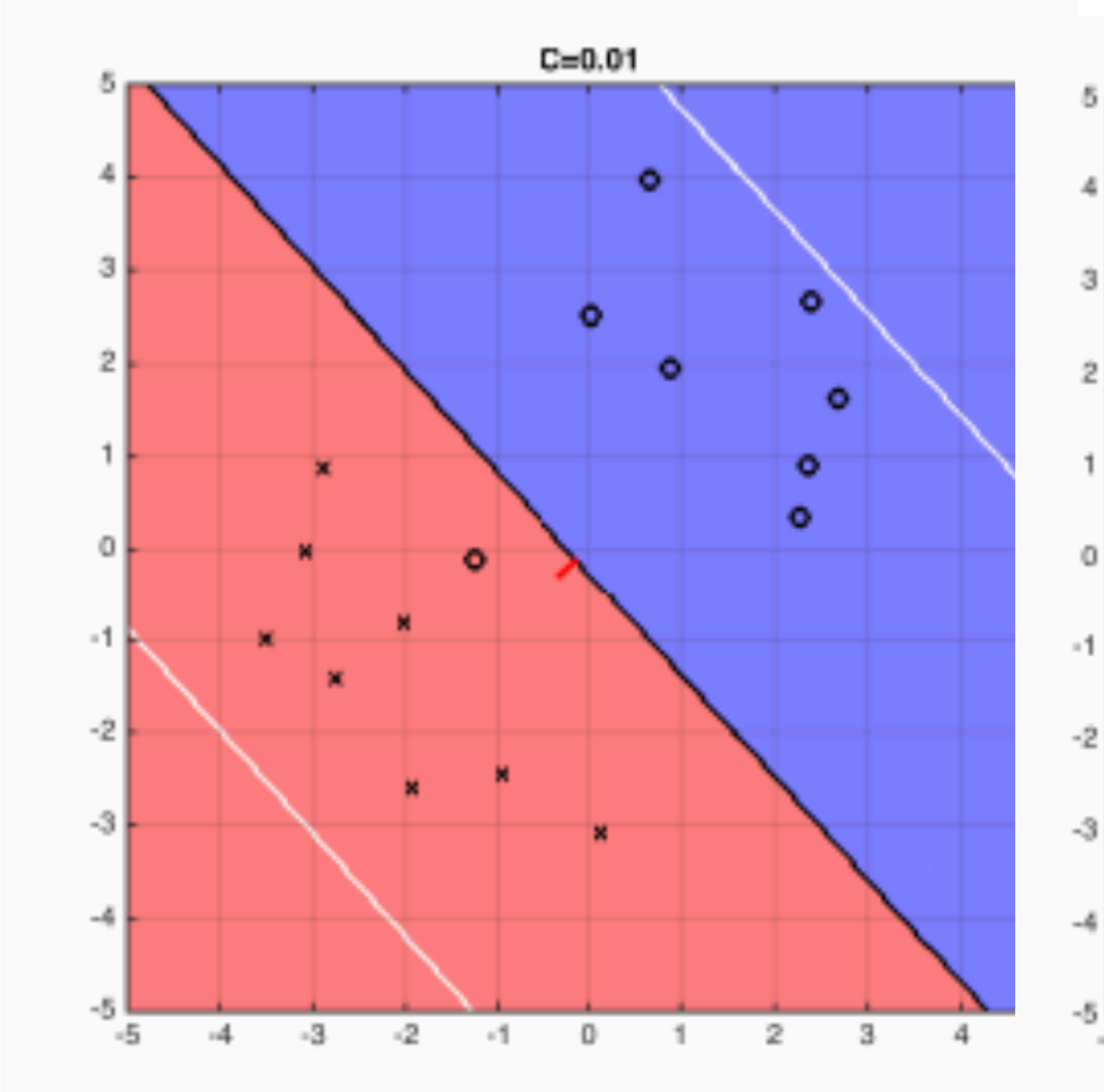
The solution $w \to \mathbf{0}$ (i.e., we do not care about hinge loss part)

SVM for non-separable data

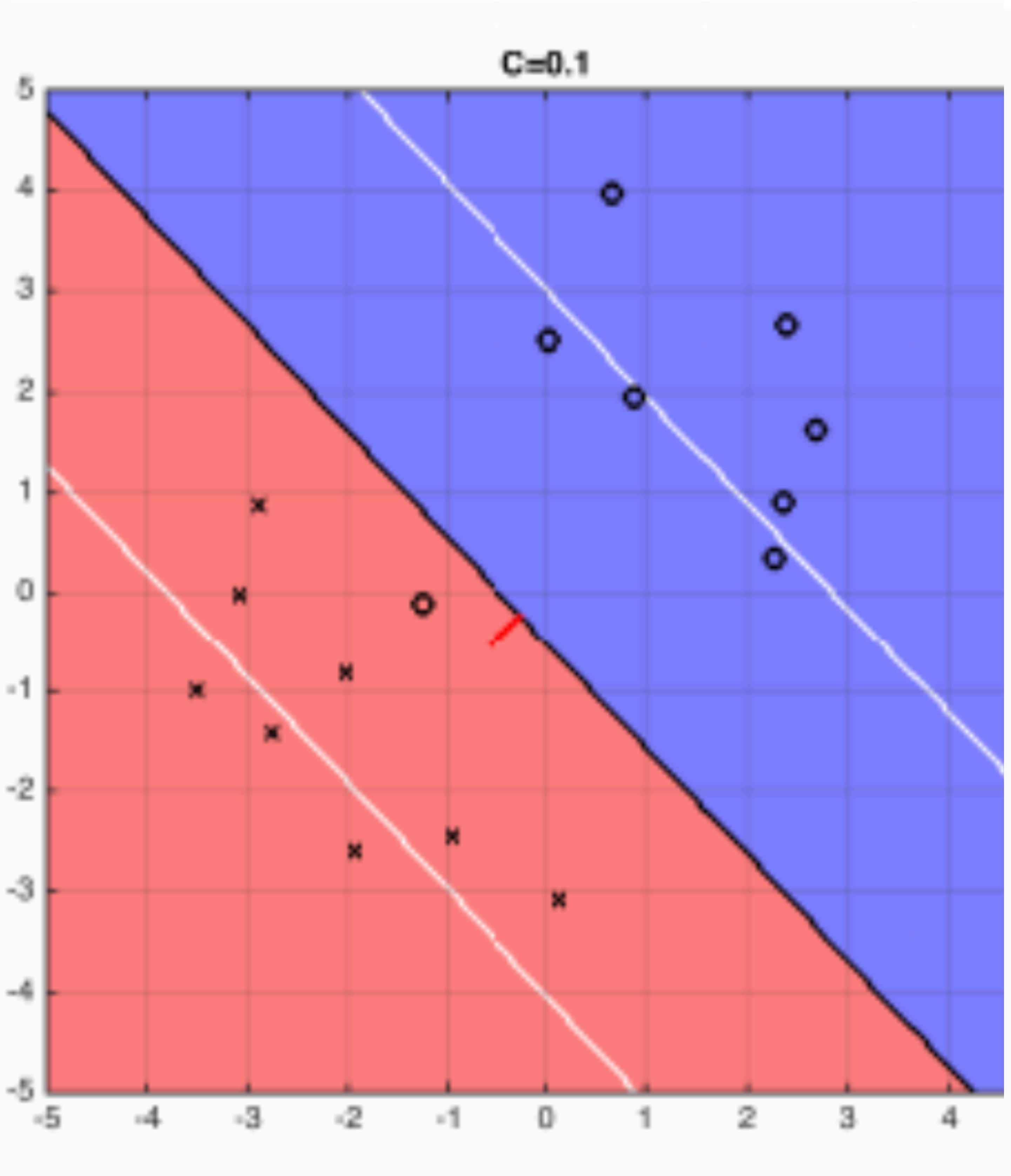
$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

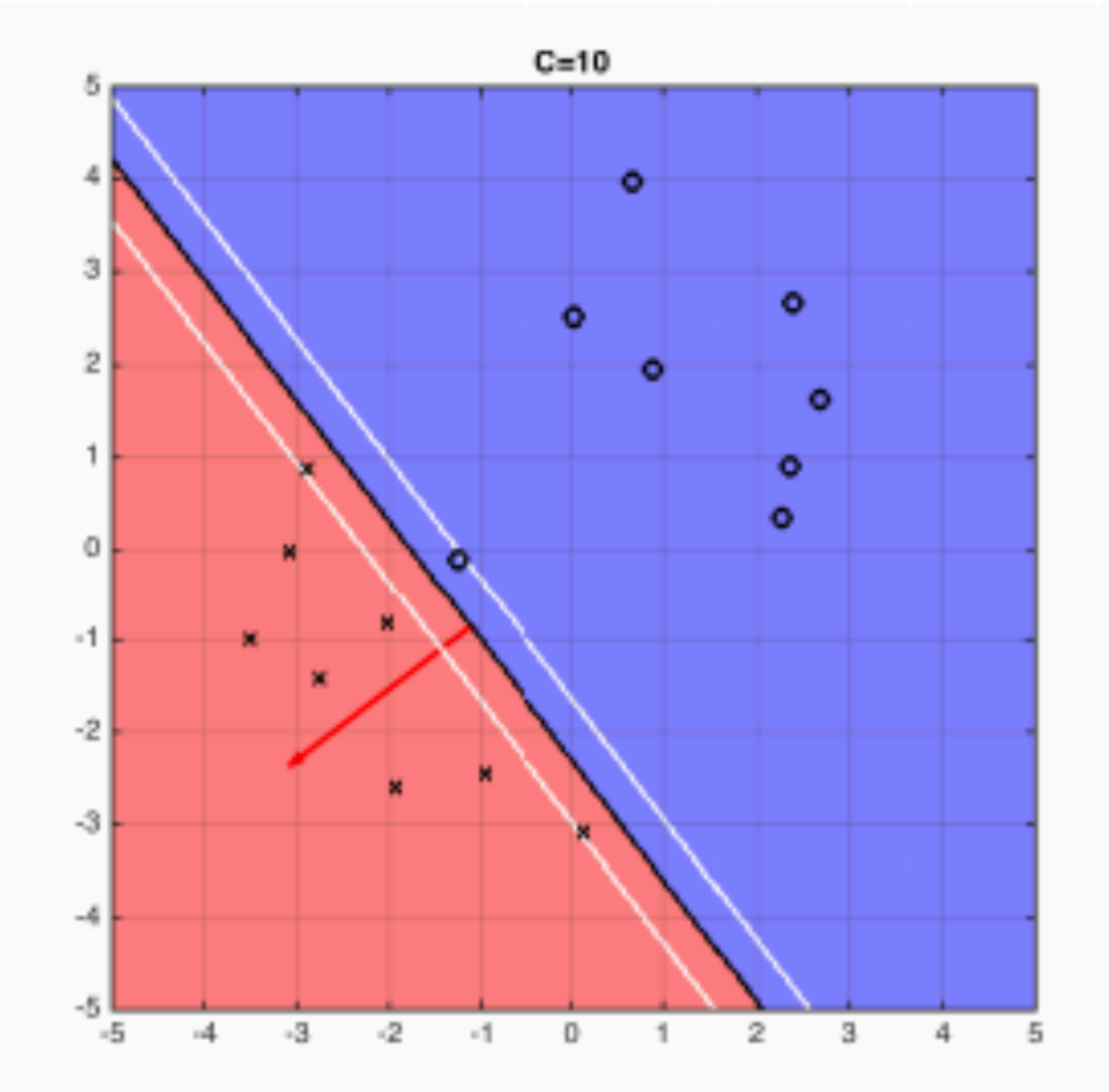
C = 0.01



C = 1



C =10

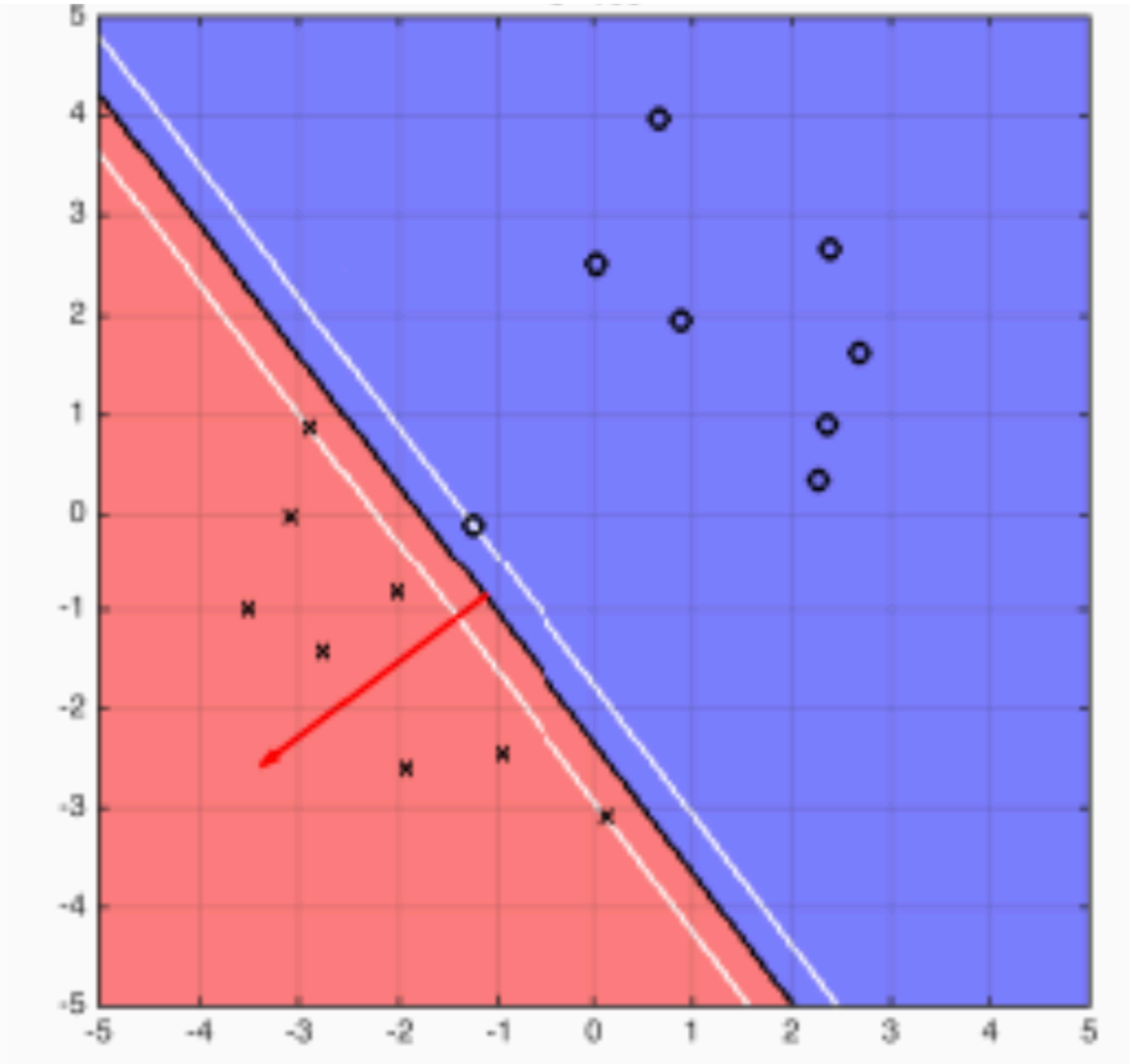


SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 100



all examples have zero Hinge loss, but w has large norm

Bad geometric margin but good functional margin (achieved by "cheating")

Potentially overfitting to the noise, not a good classifier in test time maybe

Summary for today

1. SVM for linearly separable data

$$\min_{w,b} \|w\|_2^2$$

$$\forall i : \; y_i(w^\top x_i + b) \geq 1$$

2. SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \underbrace{\boxed{\max\left\{0, 1 - y_i(w^\top x_i + b)\right\}}}_{\text{Hinge loss}}$$