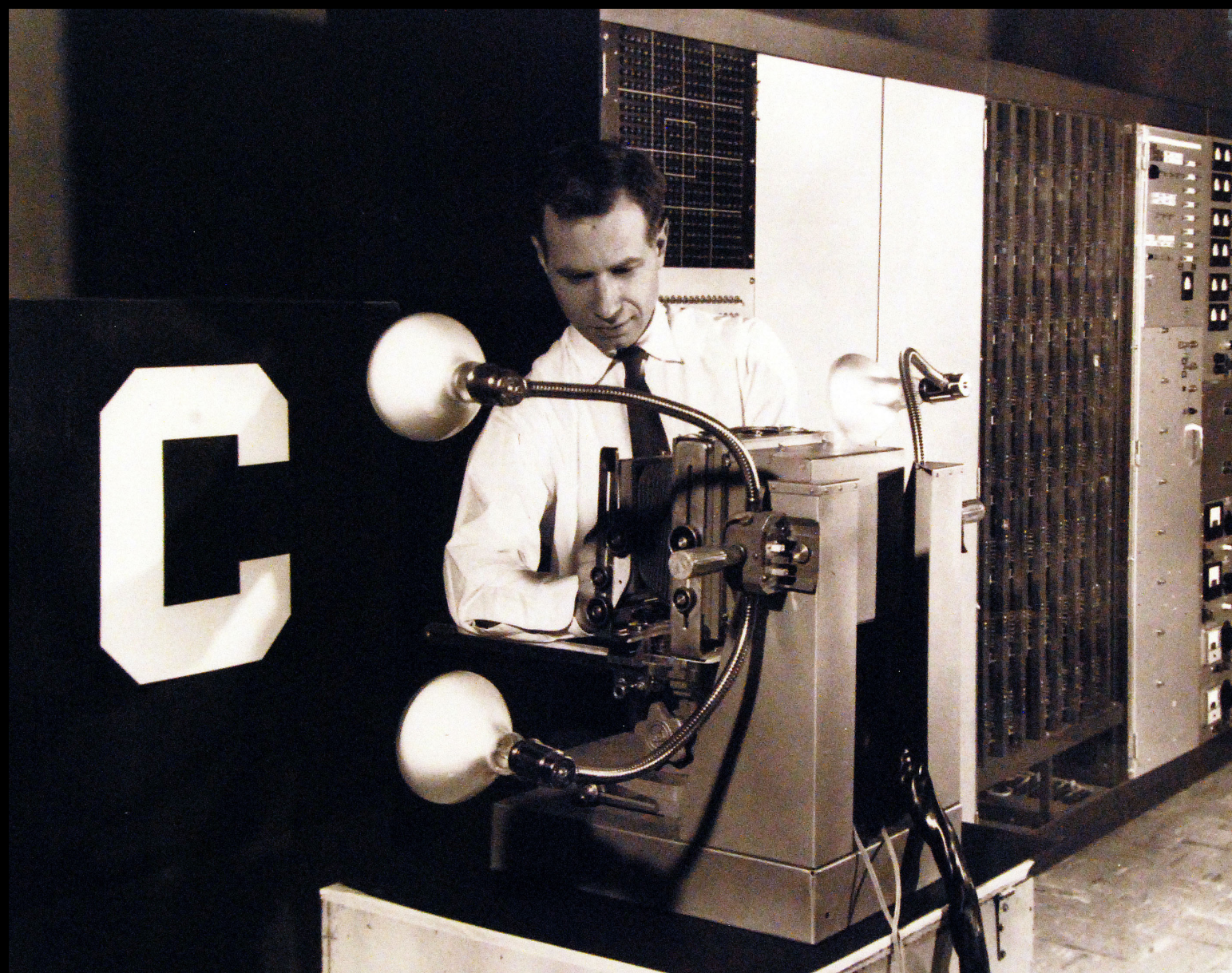# The Perceptron Algorithm

Announcements

1. P2 (Perceptron) will be out tmr

Recap on PCA

T/F: we need to center the dataset before we run PCA

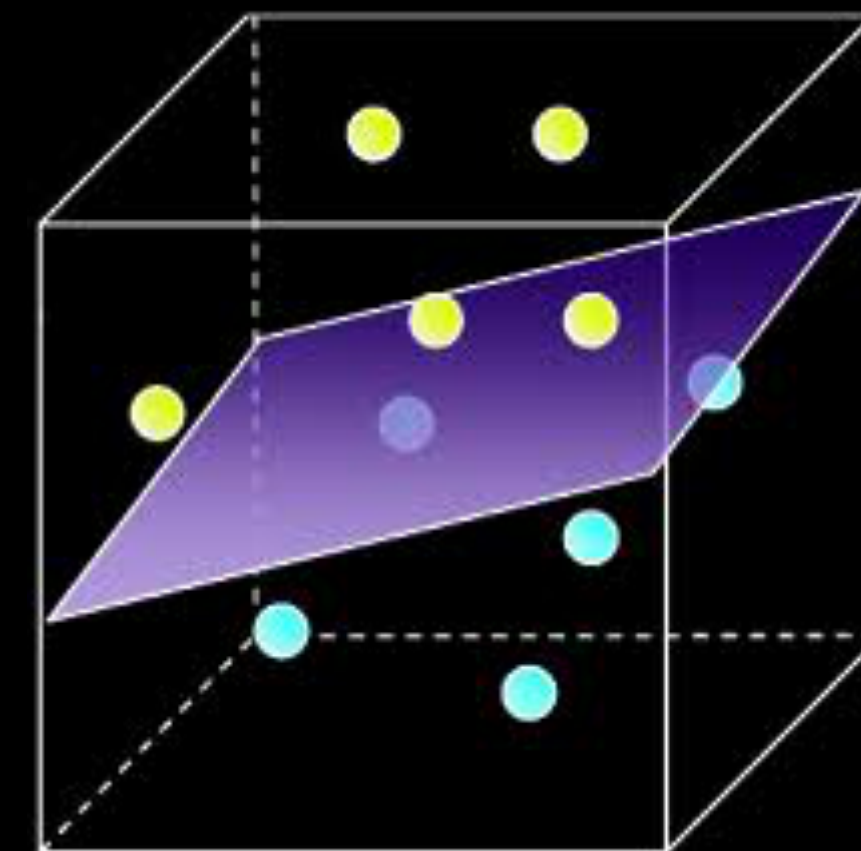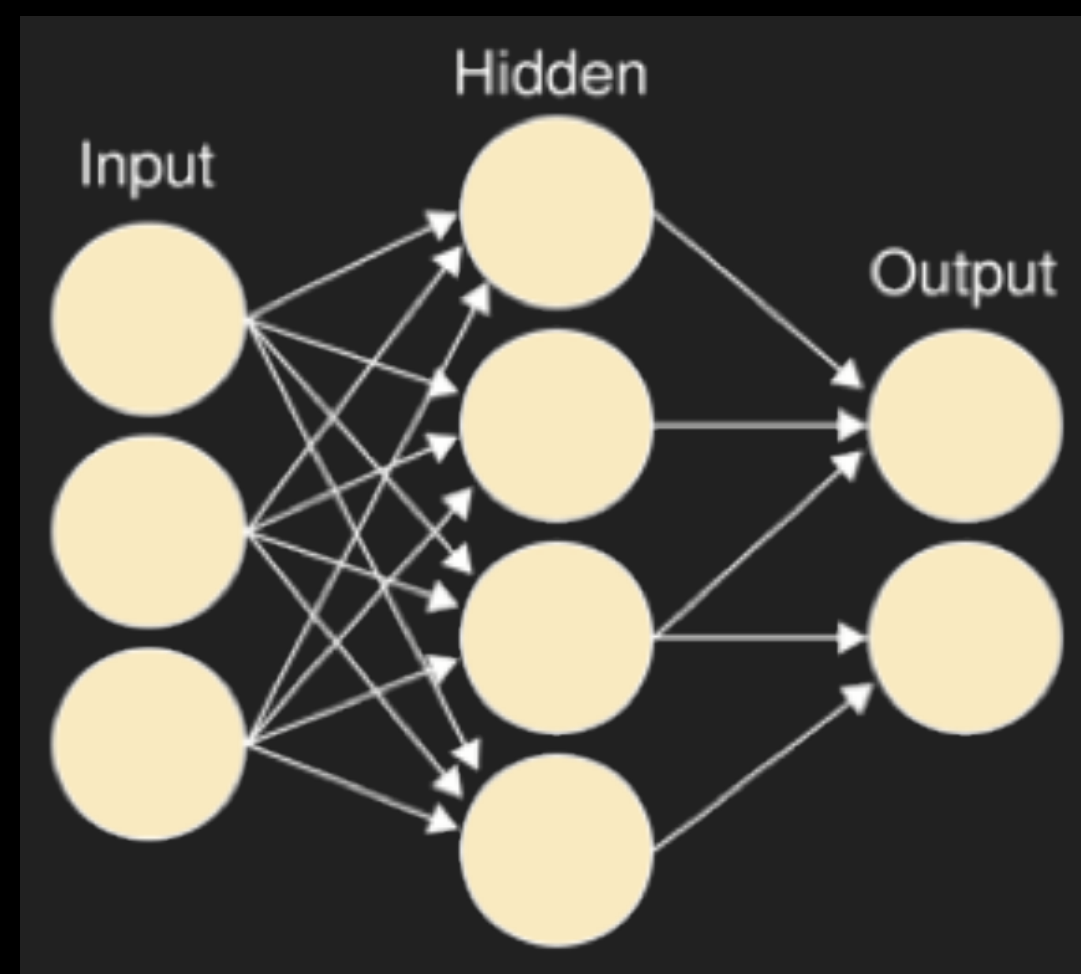Q: How to pick the parameter K in PCA?

Perceptron, 1957

Frank Rosenblatt
@ Cornell!

Predecessor of deep networks.

Separating two classes of objects using a linear threshold classifier.

Perceptron, 1957

New Navy Device Learns by Doing

- The New York Times (July 8, 1958)

"Later perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted."

https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

Today

Objective: learn our first (binary) classification algorithm
and understand why it works

Outline

1. Linear binary Classifier

2. Algorithm

3. Proof of why it works

Linear classifier

Binary classification setting: $x \in \mathbb{R}^d, y = \{-1, +1\}$

Hyperplane $H = \{x : w^\top x + b = 0\}$

w: weight vector, wlog assume $\|w\|_2 = 1$

b: bias term; $|b|$ determines the distance of the hyperplane to origin
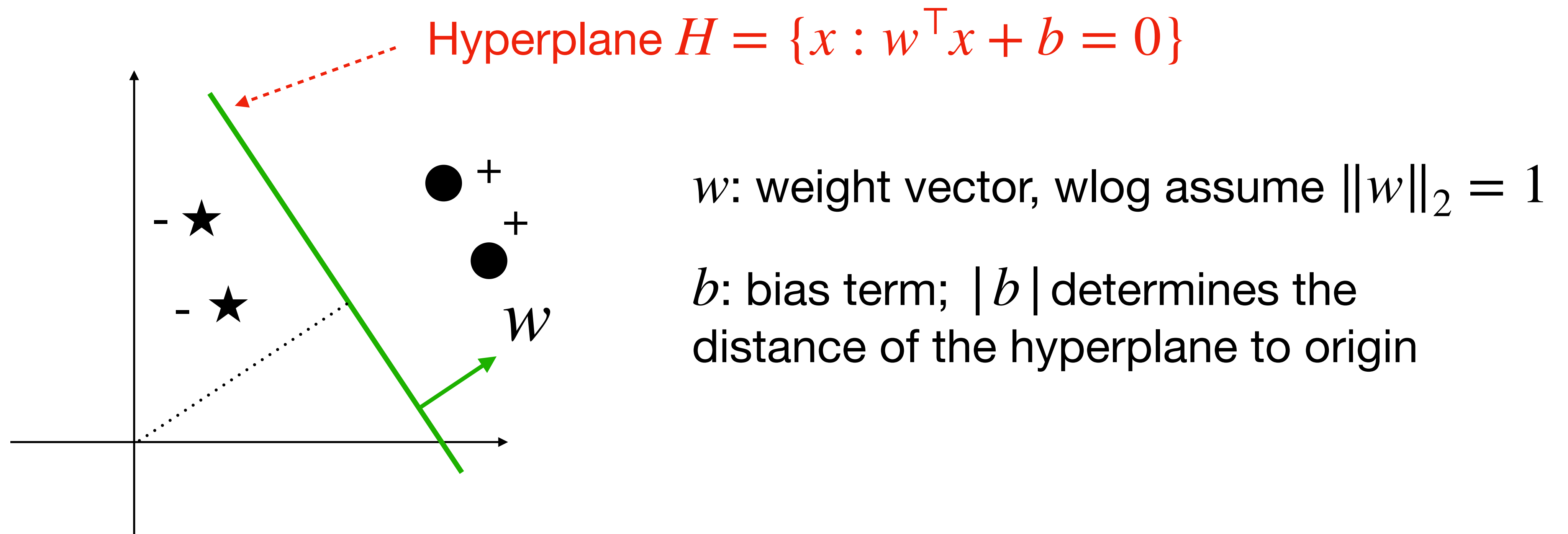
Linear classifier

Binary classification setting: $x \in \mathbb{R}^d, y = \{-1, +1\}$

A Hyperplane defines a binary linear classifier

$$\text{sign}(w^\top x + b)$$
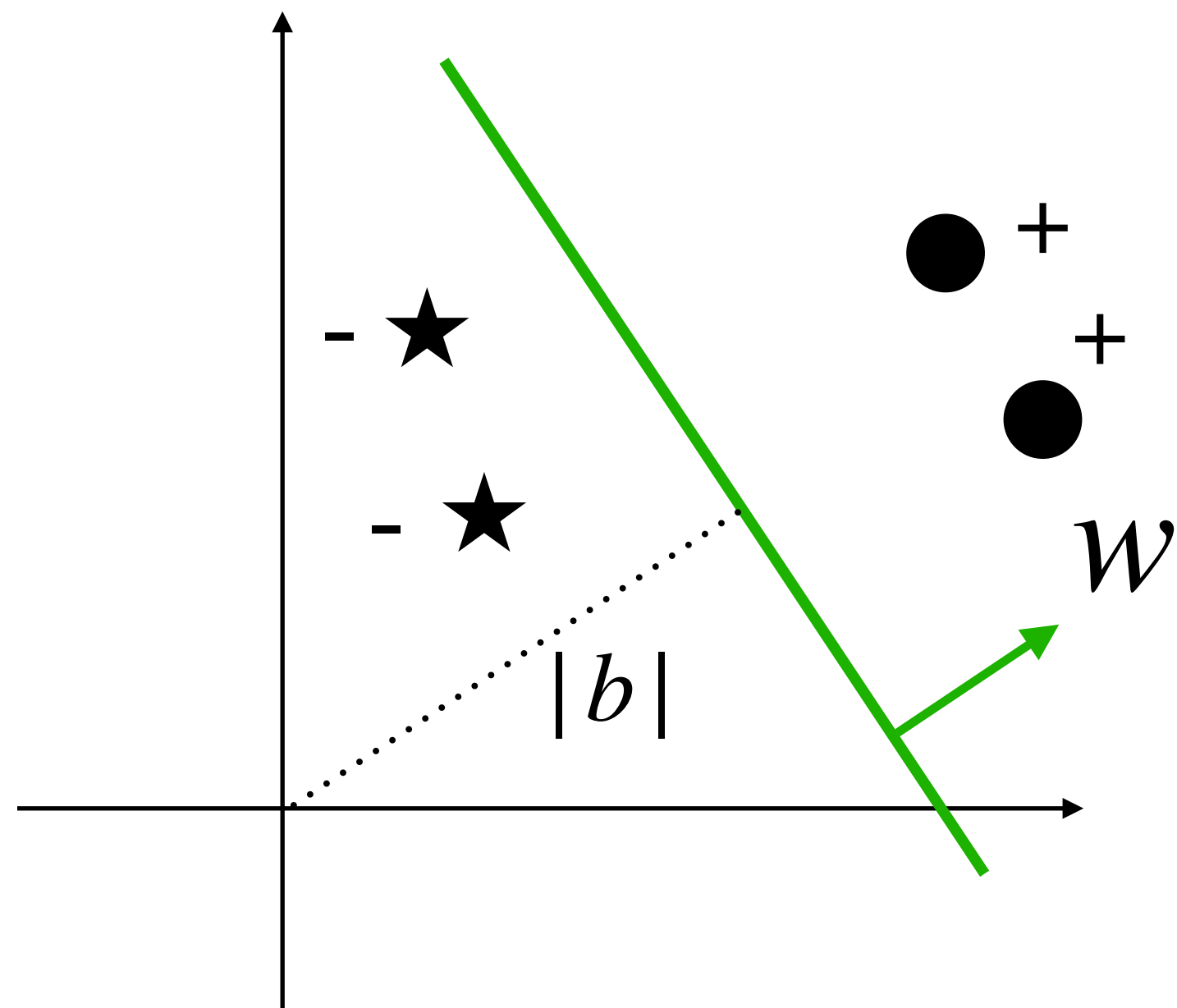
Setting

We often assume data $\{x_i, y_i\}_{i=1}^n$ is linearly separable,

i.e., $\exists w^\star, b^\star$, such that

$$\text{sign}((w^\star)^\top x_i + b^\star) = \text{sign}(y_i), \forall i$$

Or equivalently,

$$y_i((w^\star)^\top x_i + b^\star) > 0, \forall i$$
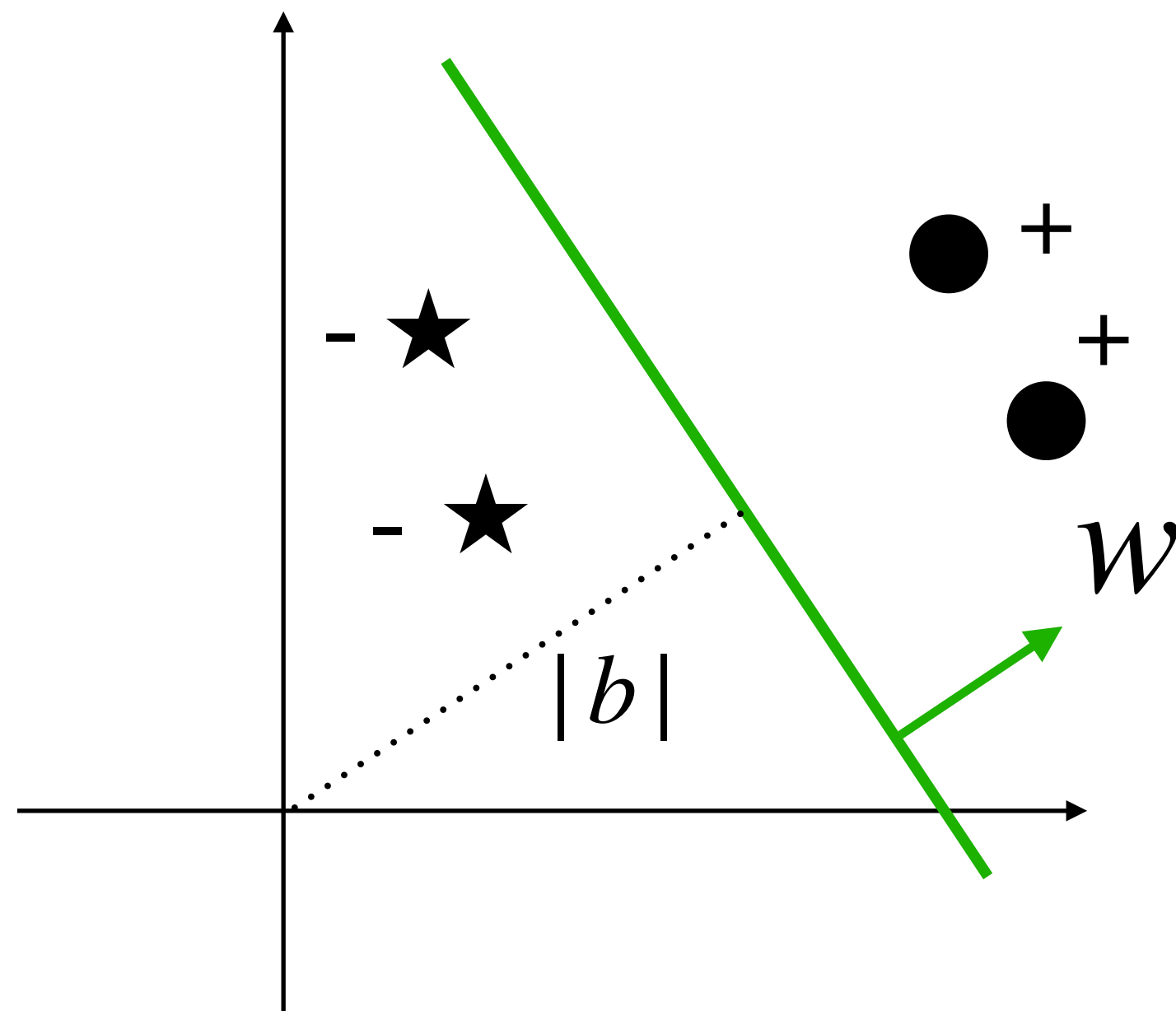
Linear classifier

Absorbing the bias term into the feature vector

$$w^\top x + b = \begin{bmatrix} w \\ b \end{bmatrix}^\top \begin{bmatrix} x \\ 1 \end{bmatrix}$$

Throughout the semester, we will assume feature x in default contains the constant 1

Outline

1. Linear binary Classifier

2. Algorithm

3. Proof of why it works

The learning protocol

Consider the **online learning** setting where every iteration t, a pair (x_t, y_t) shows up

For $t = 0 \rightarrow \infty$

New feature x_t shows up

Alg makes a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$

Check if $\hat{y}_t = y_t$

Alg updates w_{t+1} Perceptron tells us how to do this update!

Goal: make # of mistakes $\displaystyle\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t)$ as small as possible

The Algorithm

Initialize $w_0 = \mathbf{0}$

For $t = 0 \rightarrow \infty$

New feature x_t shows up

Alg makes a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$

Check if $\hat{y}_t = y_t$

Alg updates $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t)y_t x_t$

Case 1: $\hat{y}_t = y_t$, $w_{t+1} = w_t$

Case 2: $\hat{y}_t \neq y_t$ (e.g., $\hat{y}_t = -1, y_t = 1$)

$$w_{t+1}^\top x_t - w_t^\top x_t = (x_t^\top x_t)$$

Value of $w_{t+1}^\top x_t$ is increased
(the correct progress)

Q: what happens when
$\hat{y}_t = 1, y_t = -1$

A Geometric explanation

When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1$, $w_t^\top x_t > 0$)
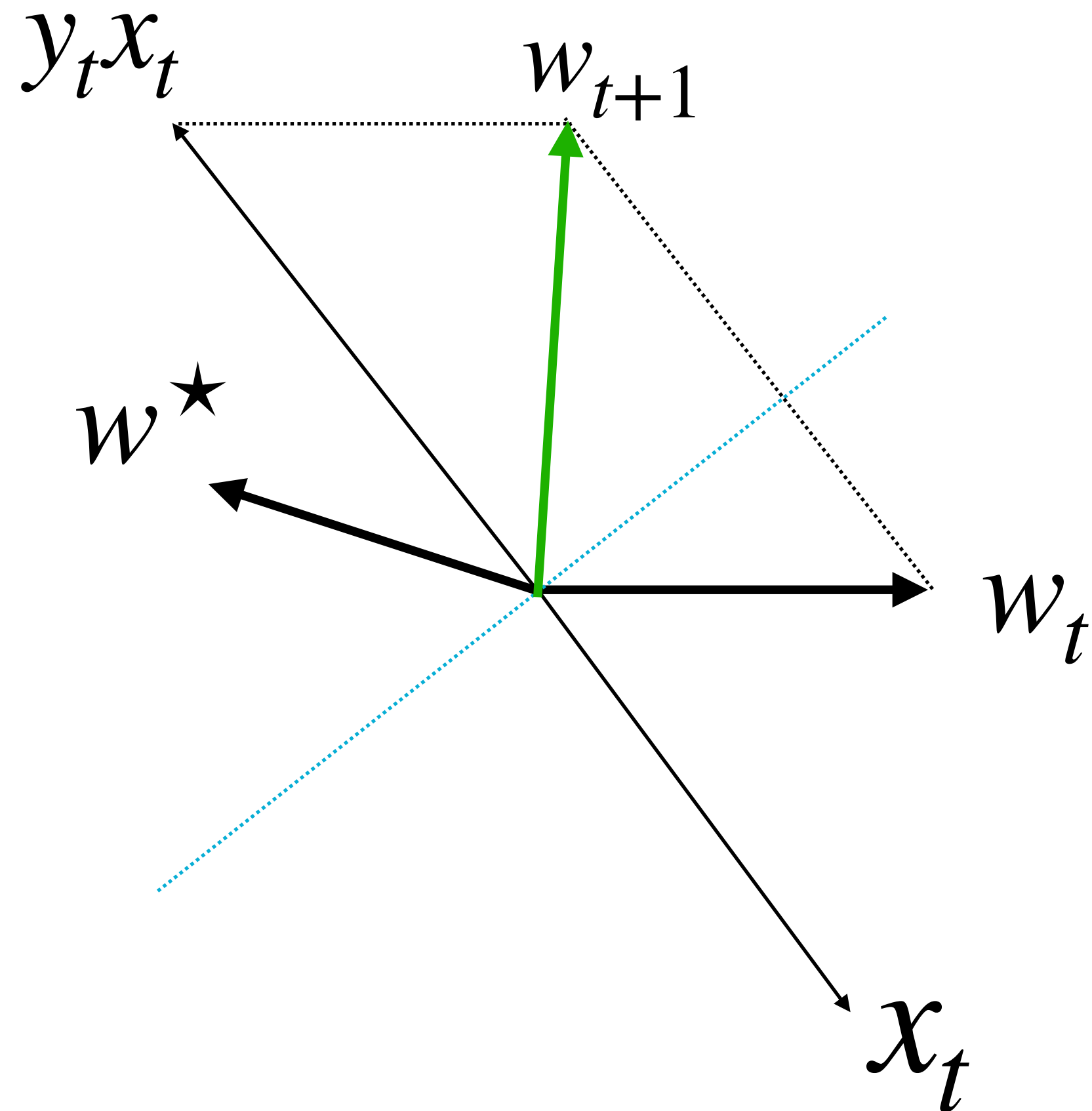
We should track how the $\cos(\theta_t)$ is changing:

$$\cos(\theta_t) = \frac{w_t^\top w^\star}{\|w_t\|_2}$$

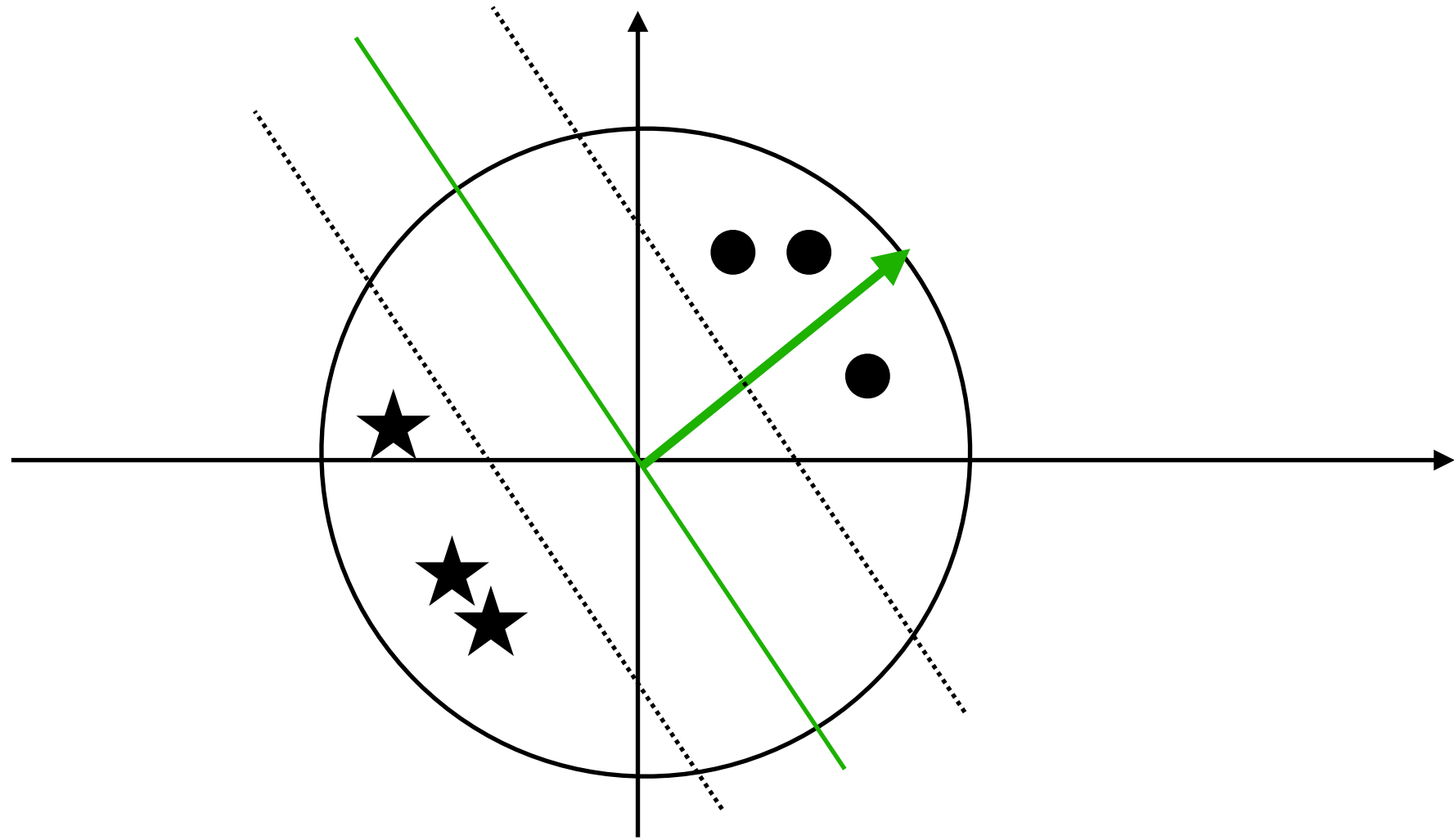Q: What does $w^\star$ look like?

Outline

1. Linear binary Classifier

2. Algorithm

3. Proof of why it works

Main theorem

Theorem of Perceptron:

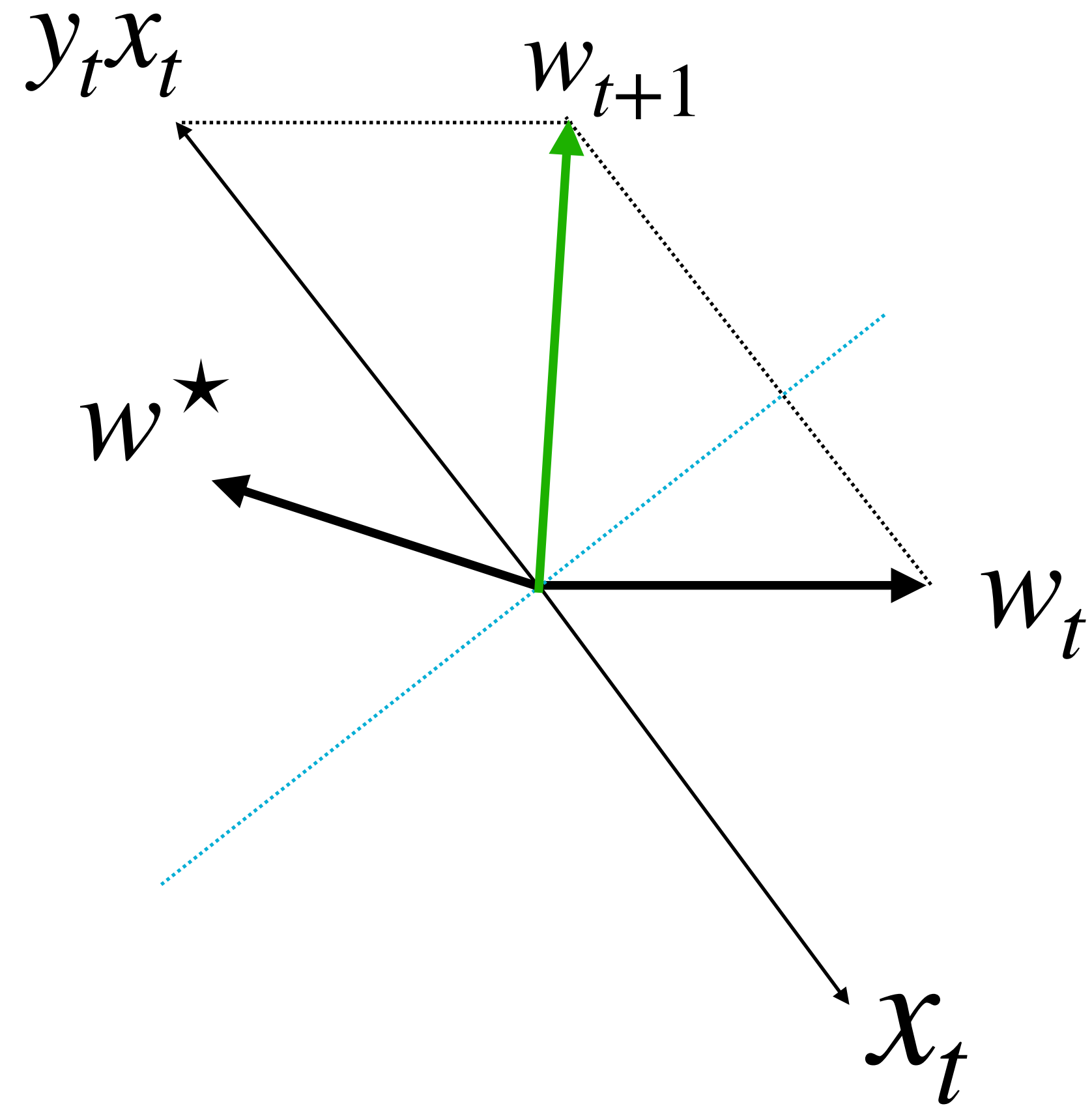Assume $\|x_t\|_2 \leq 1, \forall t$. If there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that

$$y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$$

then:

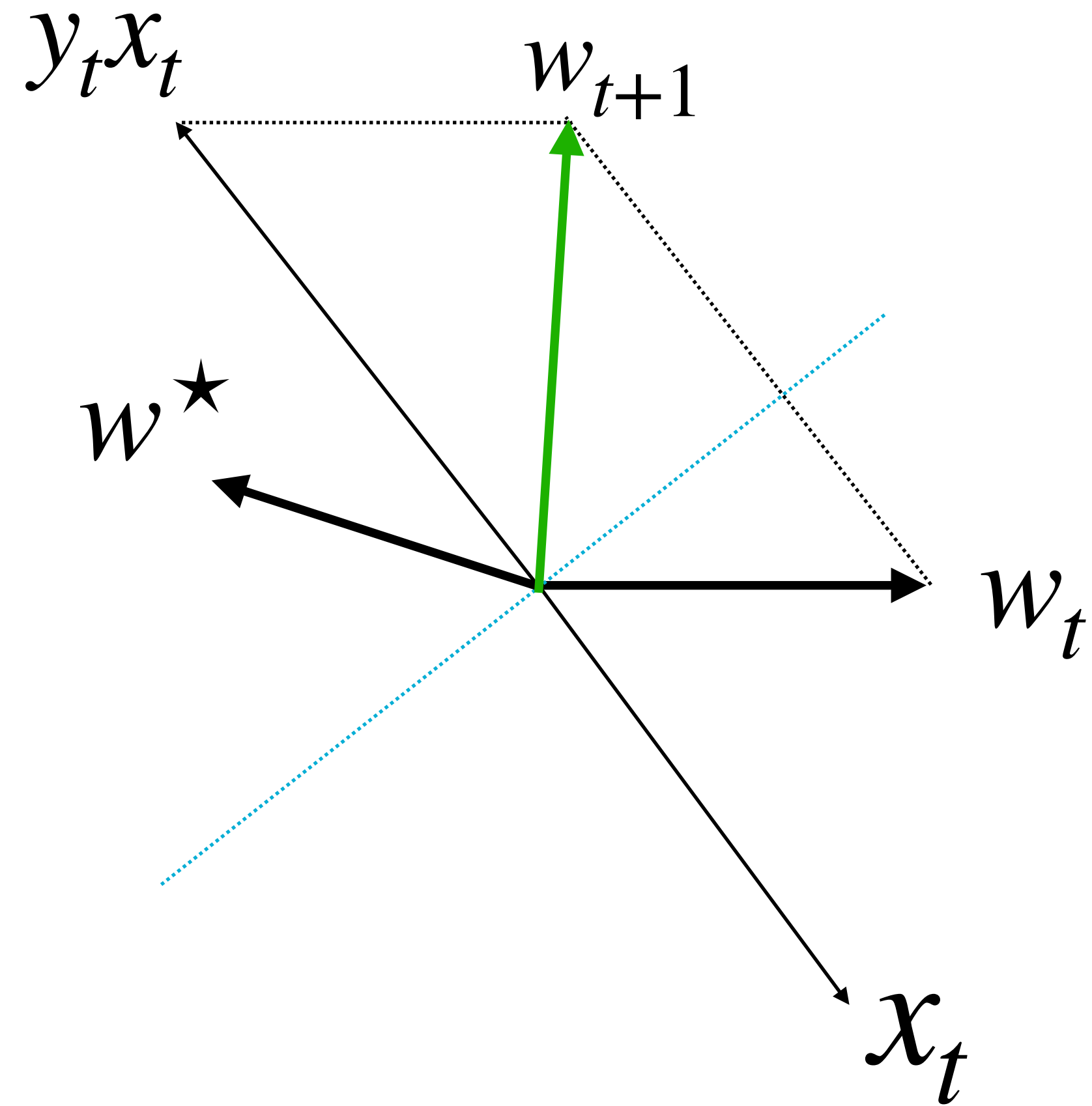$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

Proof of the theorem

$$\cos(\theta_t) = \frac{w_t^\top w^\star}{\|w_t\|_2}$$

Assume we make a mistake at x_t, track how the denominator and numerator change

Proof of the theorem

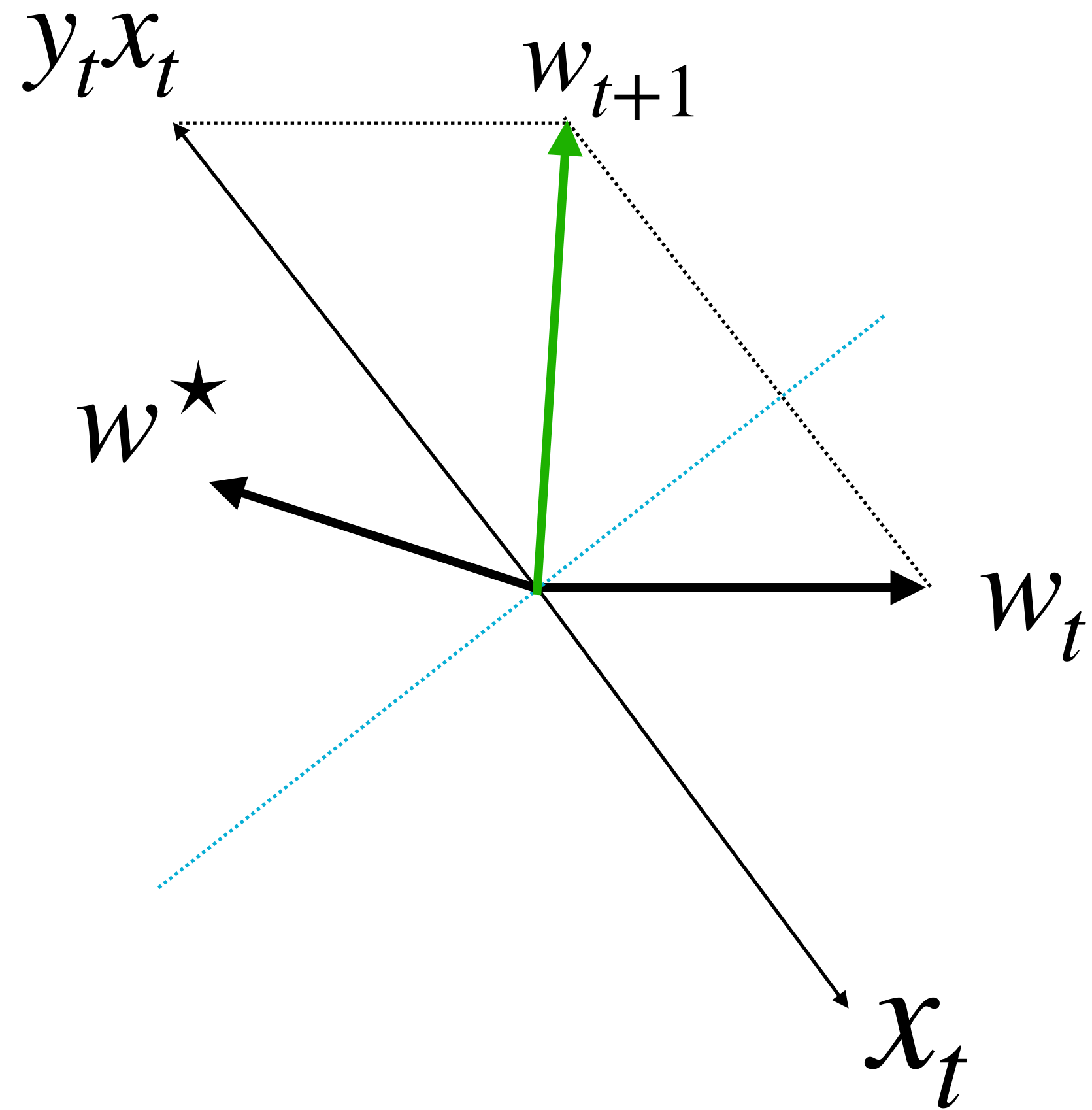**1. Track** $w_t^\top w^\star$

$$w_{t+1}^\top w^\star = (w_t + y_t x_t)^\top w^\star$$

$$= w_t^\top w^\star + y_t x_t^\top w^\star$$

$$\geq w_t^\top w^\star + \gamma$$

Whenever we make a mistake, $w_t^\top w^\star$ at least increased by γ

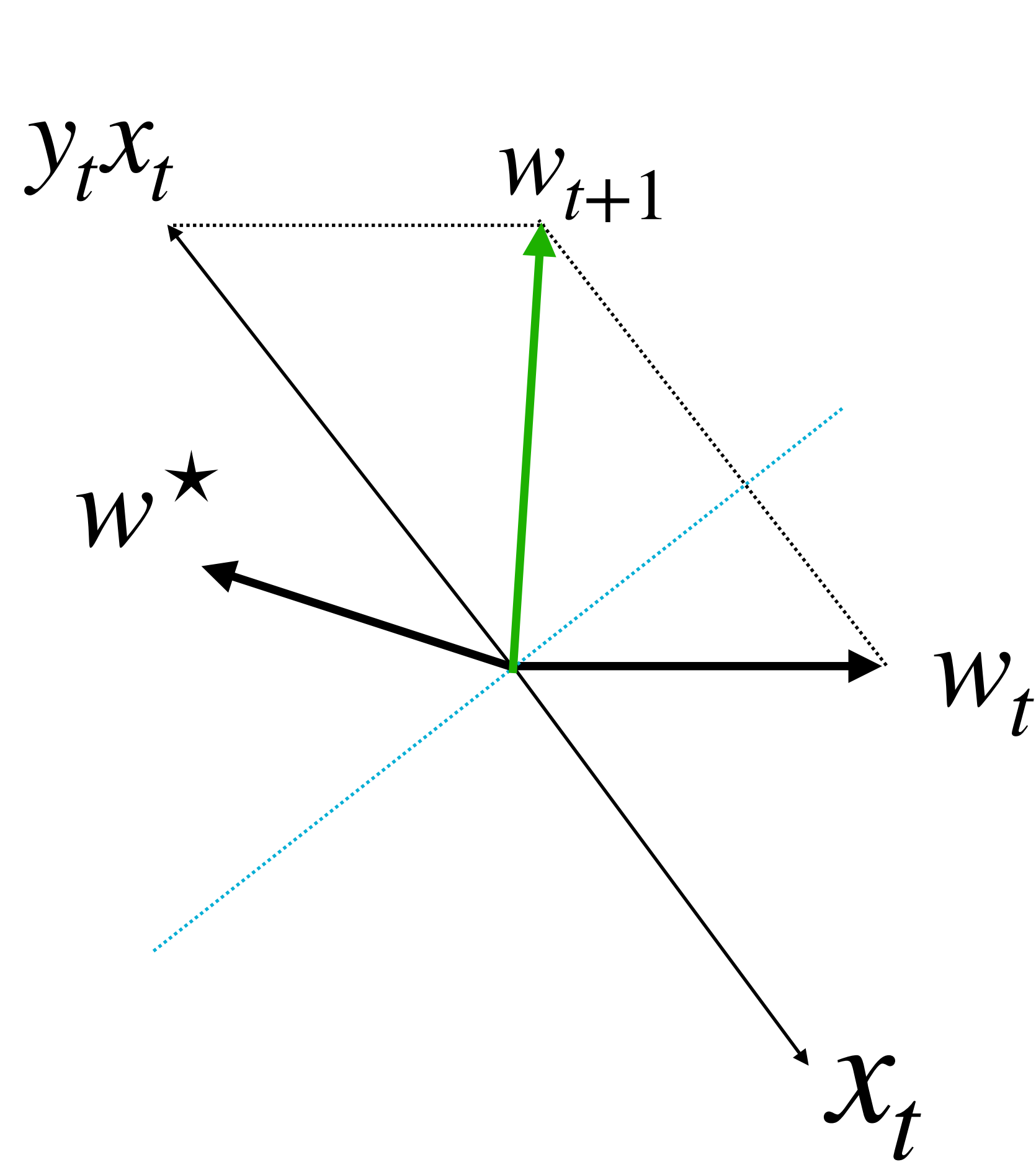# Proof of the theorem

2. Track $w_t^\top w_t$

$$w_{t+1}^\top w_{t+1} = (w_t + y_t x_t)^\top (w_t + y_t x_t)$$

$$= w_t^\top w_t + 2 w_t^\top (x_t y_t) + x_t^\top x_t$$

$$\leq w_t^\top w_t + 1$$

Discuss this derivation in small group for 5 minutes!

Proof of the theorem

3. What is $cos(\theta_t) = w_t^\top w^\star / \sqrt{w_t^\top w_t}$ **if we have made M mistakes?**

$y_t x_t$

w_{t+1}

$w^\star$

w_t

x_t

After make M mistakes:

$$w_t^\top w^\star \geq M\gamma$$

$$w_t^\top w_t \leq M$$

$$1 \geq \cos(\theta_t) \geq (M\gamma)/\sqrt{M} = \sqrt{M}\gamma$$

$$\Rightarrow M \leq 1/\gamma^2$$

Summary

The Perceptron algorithm:

1. Binary classification algorithm, runs in online mode, makes update when makes a mistake

(See lecture note for how to apply Perceptron on a static dataset)

2. Total # of mistakes is bounded by a constant $(1/\gamma^2)$