L25: the Perceptron Algorithm

for separable linear classification

Apr 13, 2026

EoDA

Jeff M. Phillips

Linear Classification

Input $(X, y) \subset \mathbb{R}^d \times \{-1, +1\}$

data points (x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$ label

Output linear classifier $(w, b) \in \mathbb{R}^d \times \mathbb{R}$

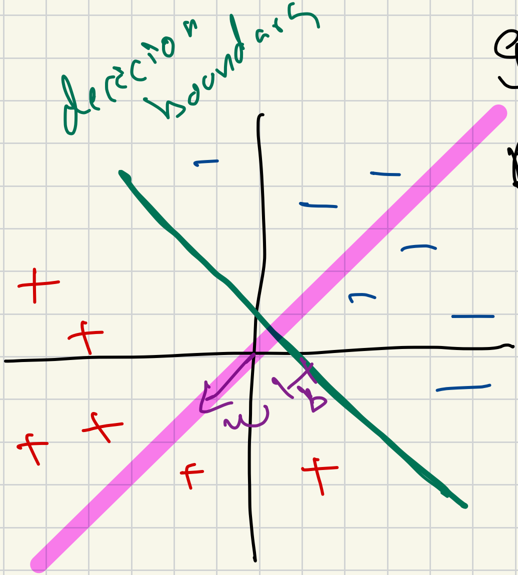
$$g_{w,b}(x) = \langle w, x \rangle + b \in \mathbb{R}$$

predict $x \Rightarrow +1$ if $g_{w,b}(x) > 0$

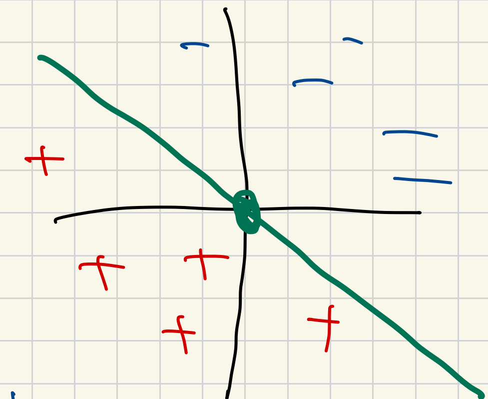
$\quad\quad\;\; -1$ if $g_{w,b}(x) < 0$

new data x

no label



Simplification for Perceptron

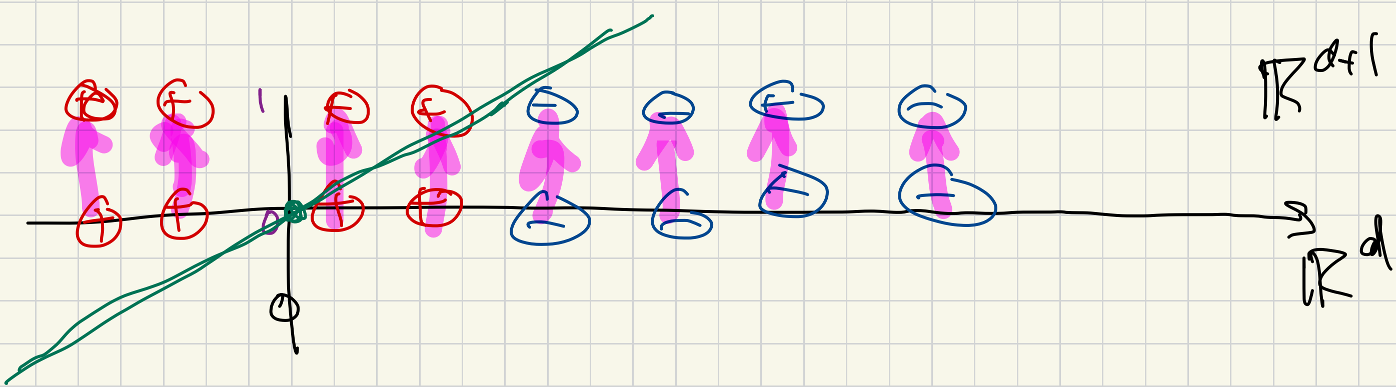


(1) Assume decision boundary passes through origin

$$g_w(x) = \langle x, w \rangle$$

convert $\quad x \to (1; x) \in \mathbb{R}^{d+1}$

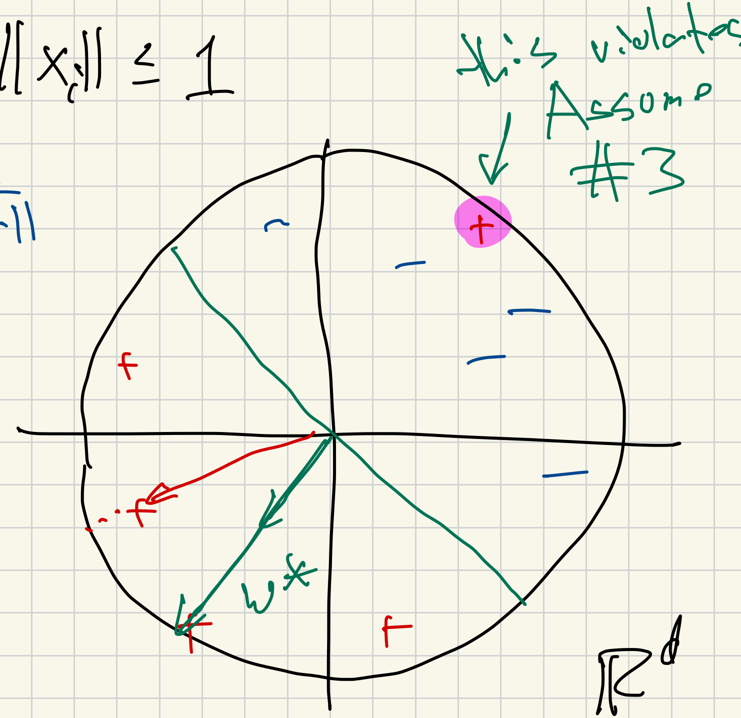
$(b, w) = (\alpha_0, \alpha_1, \dots \alpha_d) \in \mathbb{R}^d$



② Assume $\forall x_i \in X \quad \|x_i\| \le 1$

normalize data $\quad x_i \to \frac{x_i}{\|x_i\|}$

or $\quad x_i \to \frac{x_i}{\max_{x \in X} \|x\|}$



③ Assume there exists a "perfect classifier" w^*

so $\forall_{(x_i, y_i) \in (X, Y)}$

$$\text{sign}(g_{w^*}(x_i)) = y_i$$

Perceptron Algorithm

0. Initialize : set $w = y_i \overset{\in \mathbb{R}^d}{x_i}$ for any $(x_i, y_i) \in (X, y)$

1. <u>repeat</u>

   for any $(x_i, y_i) \in (X, y)$ s.t. $\underbrace{y_i \overbrace{\langle x_i, w \rangle}^{g_w(x_i)} < 0}_{\text{misclassified}}$

   <u>update</u> $w \leftarrow w + y_i x_i$

   until (no point misclassified or T steps)

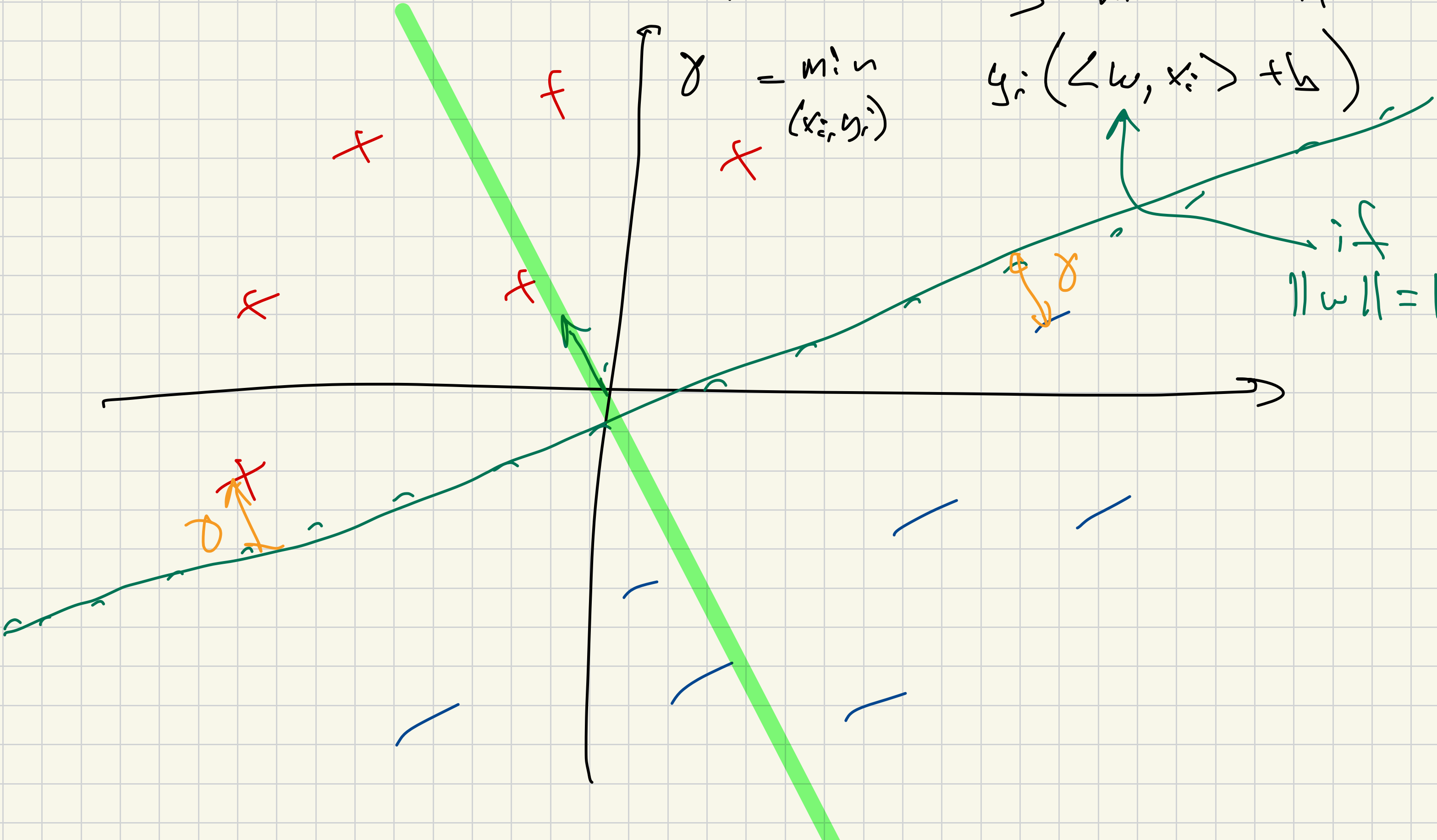
2. <u>return</u> $w \leftarrow \dfrac{w}{\|w\|}$ <u>normalize</u> w

the Margin

γ = how far is any data point from being misclassified

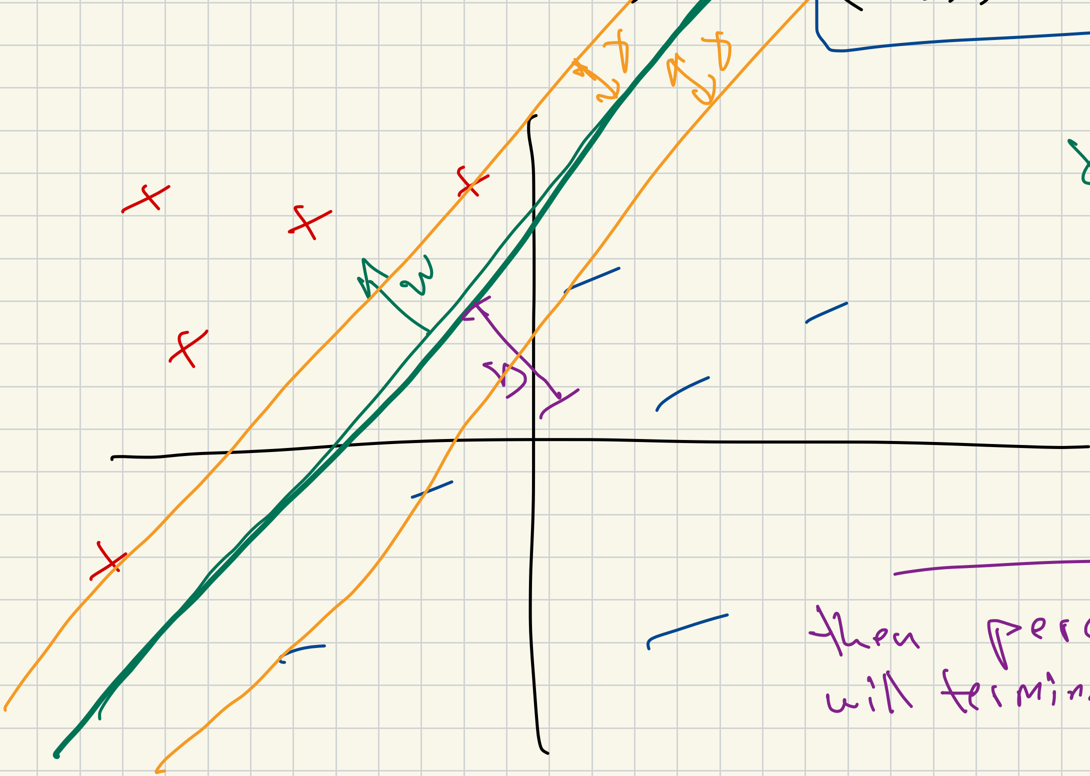
$$\gamma = \min_{(x_i, y_i)} y_i\left(\langle w, x_i \rangle + b\right)$$

if $\|w\| = 1$



Max-Margin Separator

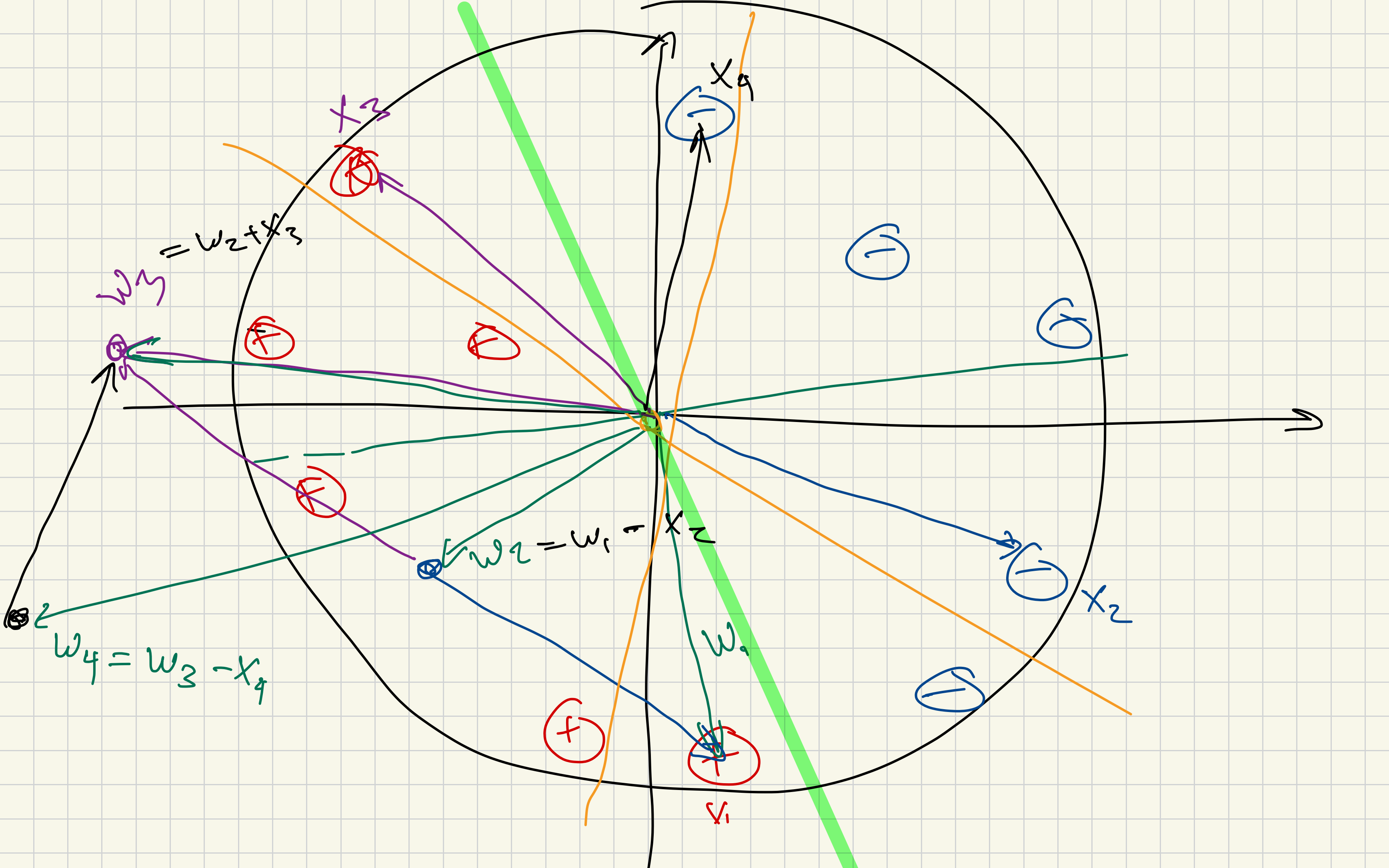
$$w^*, b^* = \underset{(w,b)}{\arg\max} \; \underbrace{\min_{(x_i, y_i)} y_i \, s_{w,b}(x_i)}_{\text{margin}}$$



γ^* = margin of the max-margin classifier

if $\|x_i\| \le 1$ ⟶

then perceptron algo will terminate in $(1/\gamma^*)^2$ steps.



x_4
x_3
$w_3 = w_2 + x_3$
$w_2 = w_1 + x_2$
w_1
x_2
x_1
$w_4 = w_3 - x_4$

Why does Perceptron work?

assuming #1, #2, #3 $\Rightarrow$ converge in $\left(1/\gamma^*\right)^2$ steps

max margin

no misclassifying points.

Analyze t steps of Perceptron

① $\|w\|^2 \le t \Rightarrow \|w\| \le \sqrt{t}$

$$\|w_{t+1}\|^2 = \langle w_t + y_i x_i, w_t + y_i x_i \rangle = \langle w_t, w_t \rangle + \underbrace{(y_i)^2}_{=1} \underbrace{\langle x_i, x_i \rangle}_{\le 1} + 2 \underbrace{y_i \langle w_t, x_i \rangle}_{<0}$$

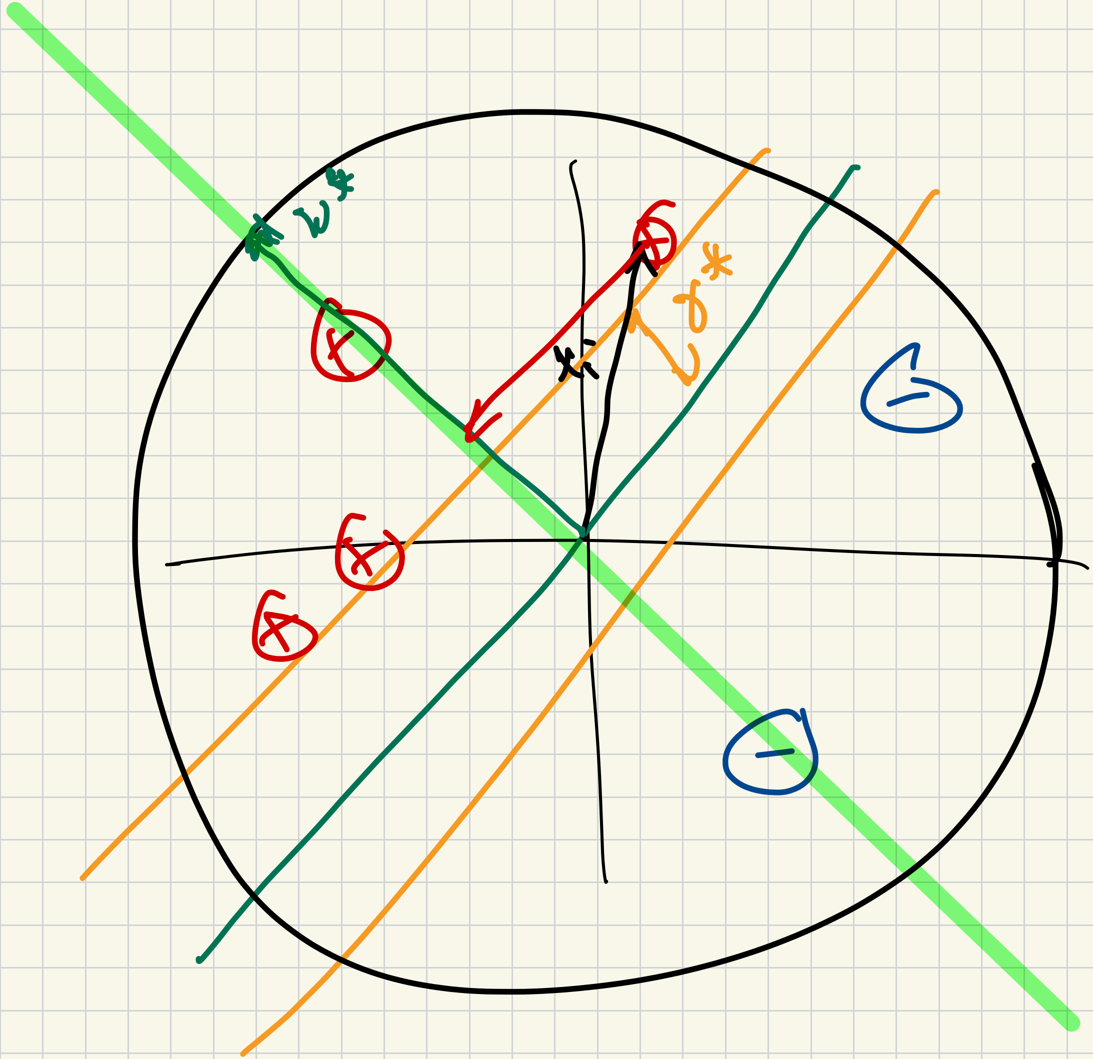
$$\le \langle w_t, w_t \rangle + 1$$

$\Rightarrow$ t steps $\|w_t\|^2 \le t$

---

② $\langle w_t, w^* \rangle > t\gamma^*$

$$\langle w_t + y_i x_i, w^* \rangle = \langle w_t, w^* \rangle + \overbrace{y_i \langle x_i, w^* \rangle}^{\ge \gamma^*}$$



w*
b*
x_i
+
+
+
+
−
−

two properties

(1) $\|w_t\| \leq \sqrt{t}$ (2) $\langle w_t, w^* \rangle > \gamma^* t$

$$t\gamma^* < \langle w_t, w^* \rangle \leq \left\langle w_t, \frac{w_t}{\|w_t\|} \right\rangle \leq \|w_t\| \leq \sqrt{t}$$

$$t\gamma^* \leq \sqrt{t} \implies \gamma^* \leq \frac{\sqrt{t}}{t} = \frac{1}{\sqrt{t}}$$

$$t \leq \left(\frac{1}{\gamma^*}\right)^2$$