

FOD A

L26

: the
Perceptron

Algorithm

Linear Classification (review)

Input $X \subset \mathbb{R}^d$
 $x_1, x_2, \dots, x_n \in \mathbb{R}^d$

Labels $y \in \{-1, +1\}^n$
 $y_1, \dots, y_n \in \{-1, +1\}$

Goal

linear model

$$g_\alpha(x) \rightarrow \mathbb{R}$$

$\in \mathbb{R}^d$

want
 $\text{sign}(g_\alpha(x_i)) = y_i$

$$\langle \alpha, x \rangle + \alpha_0$$

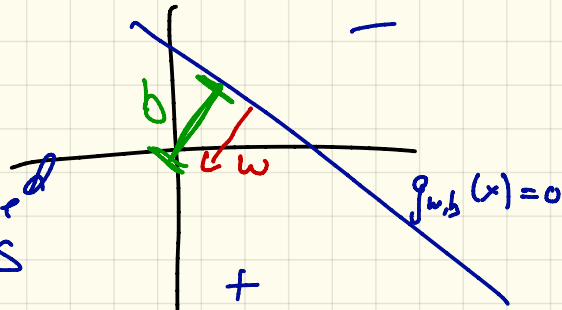
$$\alpha = (b, w)$$

$\in \mathbb{R}$ $\in \mathbb{R}^d$

$$\langle w, x \rangle + b$$

minimize
 w, b

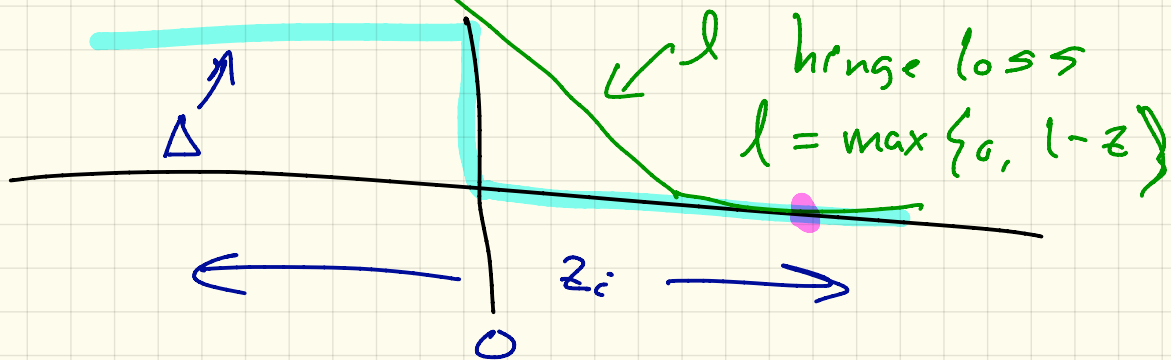
$\Delta(g_{w,b}, (x, y))$
 $= \# \text{ misclassified points}$



$$\min_{w, b} \Delta(g_{w,b}, (X, y)) = \# \text{ misclass pts.}$$

$$= \sum_{i=1}^n \Delta(g_i, g_{w,b}(x_i)) = \begin{cases} 1 & \text{if } y_i \neq g_{w,b}(x_i) \\ 0 & \text{if } y_i = g_{w,b}(x_i) \end{cases}$$

$$f(w, b) = \sum_{i=1}^n \Delta(y_i (\langle w, x_i \rangle + b))$$



Cross-Validation

$$g_{\alpha} = \text{linear} \langle \alpha, x \rangle$$

loss function

$$f(x) = \sum_{i=1}^n f_i(x) + \eta \cdot r(\alpha)$$

$$r(\alpha) = \|\alpha\|_2^2 \text{ or } \|\alpha\|_1$$

How well will this work on
new data?



evaluate on
accuracy $\frac{\Delta(g, \text{test}(x_i))}{\text{size}(\text{test})}$
build model
 w, b
 α

Perceptron Algorithm

for linear model $g_{\alpha}(x) = \langle \alpha, (1, x) \rangle$

Simplification

$$\alpha = (\alpha_0, \alpha_1, \dots, \alpha_d) \in \mathbb{R}^{d+1}$$

1. assume w, b $g_{\alpha}(x) = \langle w, x \rangle + b \implies b^* = 0$

map $x_i \in \mathbb{R}^d \rightarrow (1, x_i) \in \mathbb{R}^{d+1}$

solve for $\alpha \in \mathbb{R}^{d+1}$ instead of w, b

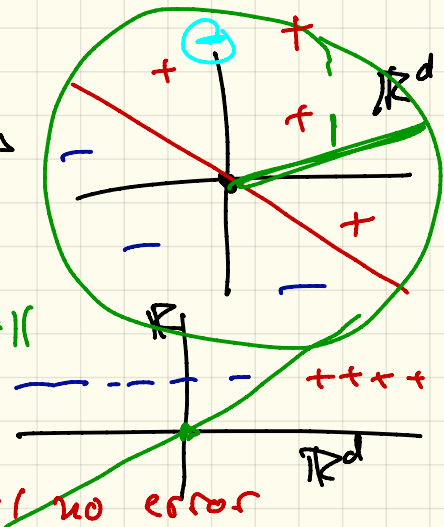
2. Assume all x_i have $\|x_i\| \leq 1$

$$x_i \leftarrow \frac{x_i}{\|x_{\max}\|}$$

$$x_{\max} = \arg \max_{x_i \in X} \|x_i\|$$

3. **Linearly separable**:

exists a classifier w with no error



Perceptron (X, y)

- assume $g_w(x) = \langle w, x \rangle$
- assume $\|x_i\| \leq 1$
- assume exist perfect w

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

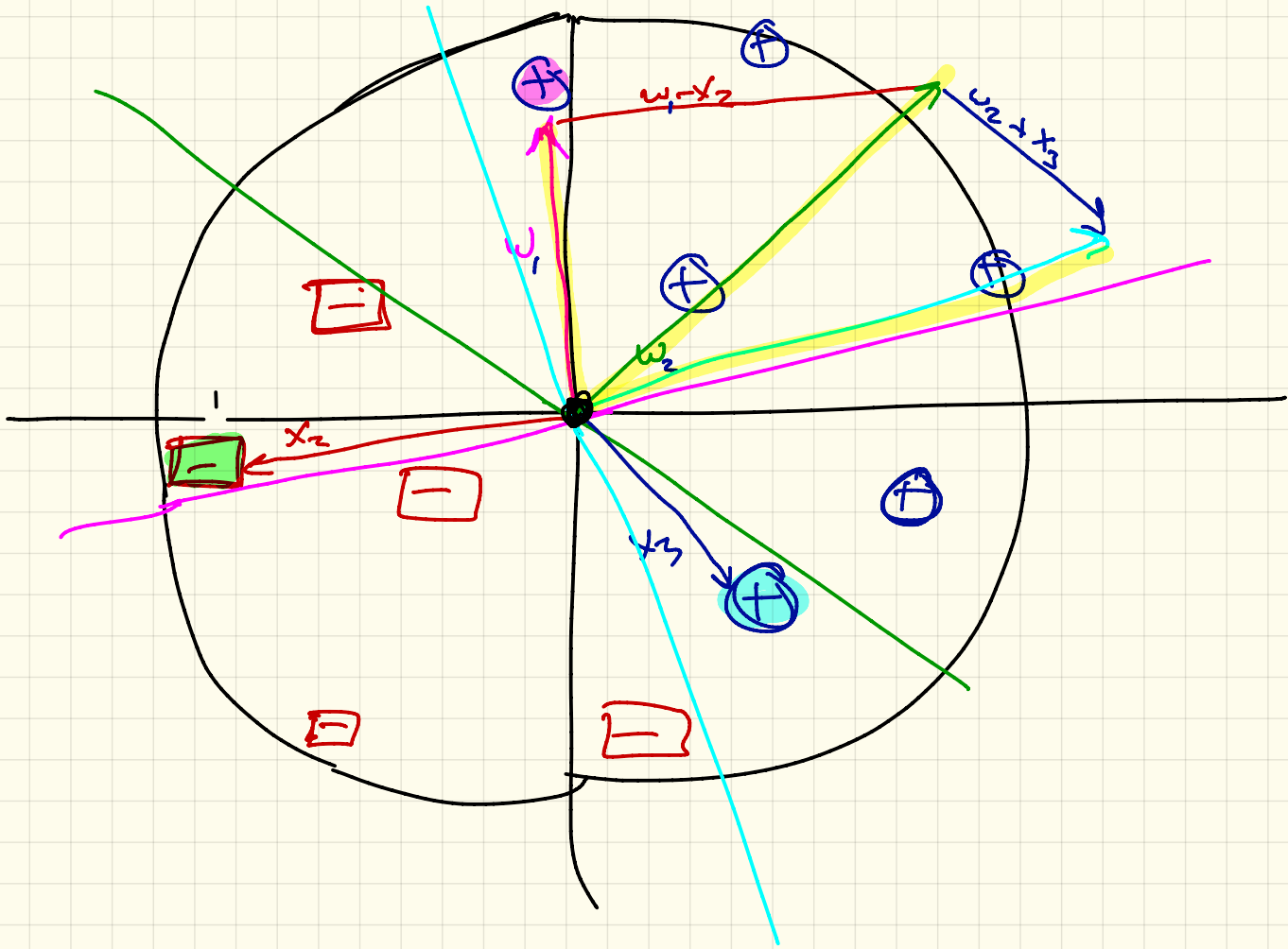
1. repeat

For any (x_i, y_i) s.t. misclassified $y_i \langle x_i, w \rangle < 0$

update $w \leftarrow w + y_i x_i$

until (no misclassified pts for T steps)

2. return $w \leftarrow \frac{w}{\|w\|}$



The Margin

the margin γ of w, b

$$\gamma = \min_{(x_i, y_i) \in X, y_i} g_i(\langle x_i, w \rangle + b)$$
$$y_i g_i(x_i) = z_i$$

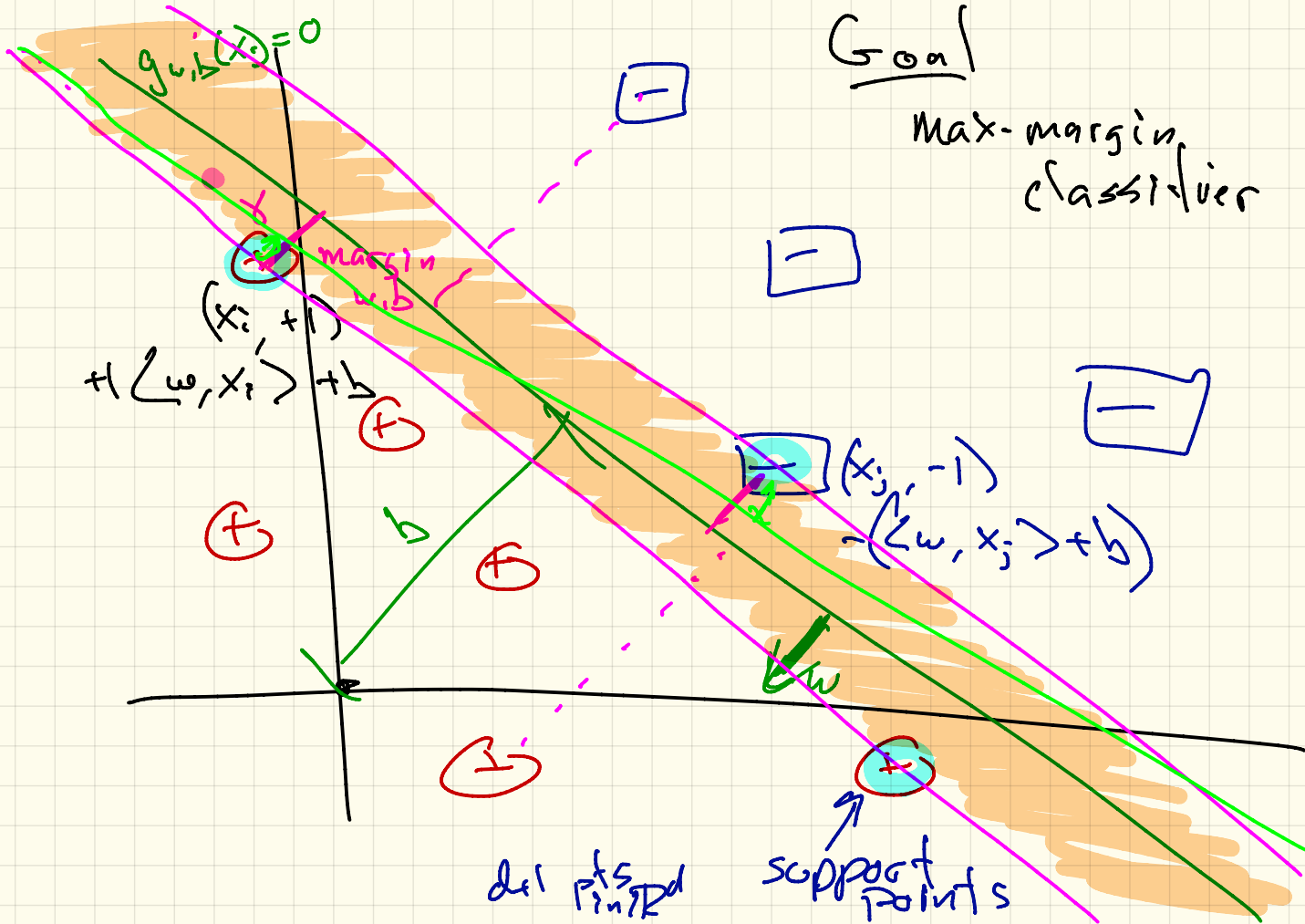
optimal

max-margin classifier :

$$(w^*, b^*) = \arg \max_{w, b} \min_{(x_i, y_i) \in X, y_i} g_i(x_i)$$

max-margin \rightarrow

$$\gamma^* = \min_{(x_i, y_i) \in X, y_i} g_i(w^*, b^*(x_i))$$




Why perceptron works?

Claim in $T \leq \left(\frac{1}{\gamma^*}\right)^2$ steps, will converge.
max-margin

t steps

(a) $\|w\|^2 \leq t$ (b) $\langle w, w^* \rangle \geq t \gamma^*$

$\|w\|^2 = \langle w, w \rangle$ 

$$t \gamma^* \leq \langle w, w^* \rangle \leq \langle w, \frac{w}{\|w\|} \rangle = \|w\| \leq \sqrt{t}$$
$$t \leq \left(\frac{1}{\gamma^*}\right)^2$$

(a) $\|w\|^2 = \langle w, w \rangle \Rightarrow \langle w + x_i y_i, w + x_i y_i \rangle =$
increase by norm² $\leq \dagger$ $= \langle w, w \rangle + \underbrace{(y_i)^2}_{\leq 1} \langle x_i, x_i \rangle + 2 y_i \langle w, x_i \rangle \leq \langle w, w \rangle + 1 + 0$ ≤ 0

(b) $\langle w, w^* \rangle \Rightarrow \langle w + y_i x_i, w^* \rangle = \langle w, w^* \rangle + y_i \langle x_i, w^* \rangle \geq \langle w, w^* \rangle + \gamma^*$