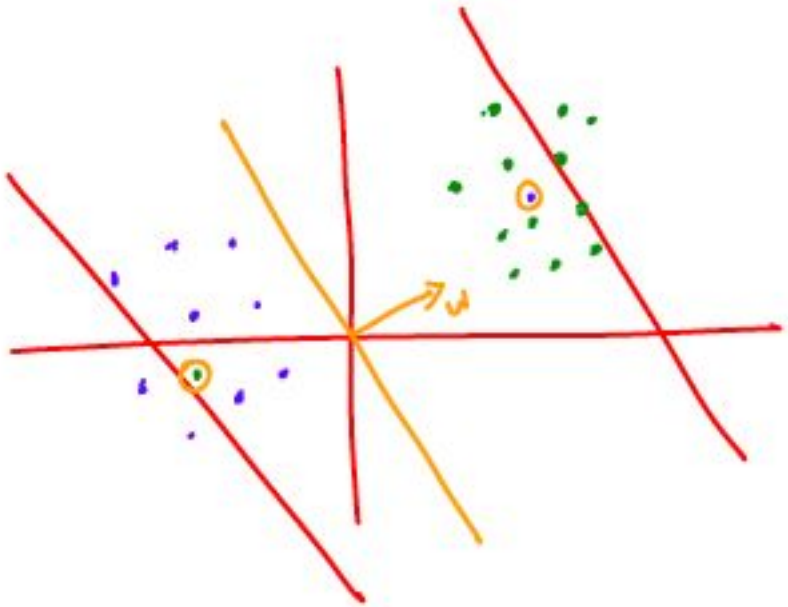
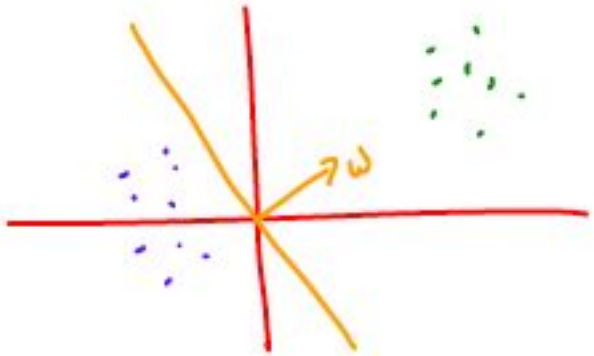
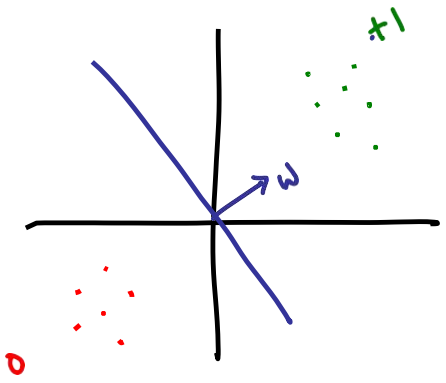
LINEAR SEPARABILITY ASSUMPTION



Dataset is not allowed in our model



Allowed.



$$P(y=1|x) = 1 \quad \text{if} \quad w^T x \geq 0$$

$$= 0 \quad \text{o/w}$$

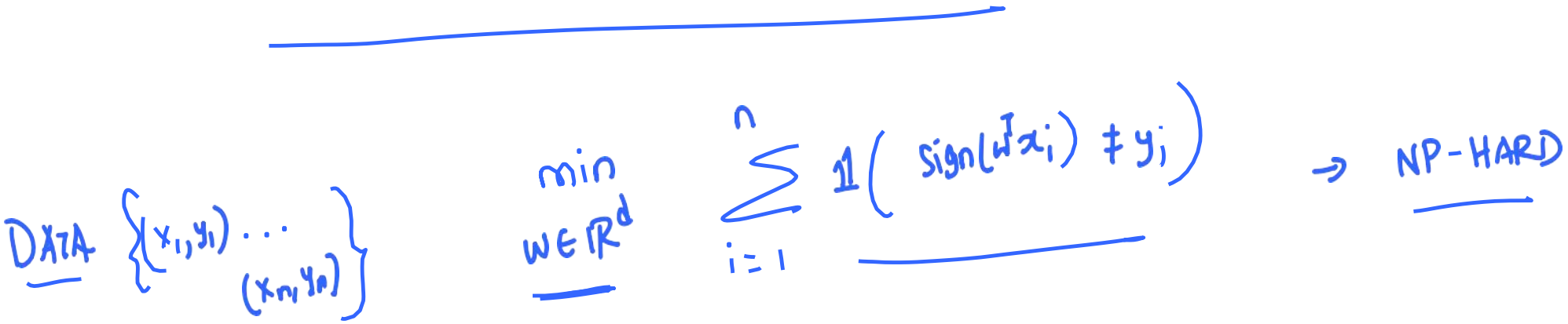
Linear separability assumption.

Lin. sep assumption :

$$\exists w \in \mathbb{R}^d \text{ s.t } \underline{\text{sign}}(w^T x_i) = y_i \quad \forall i \in [n]$$

$\{1, \ldots, n\}$.

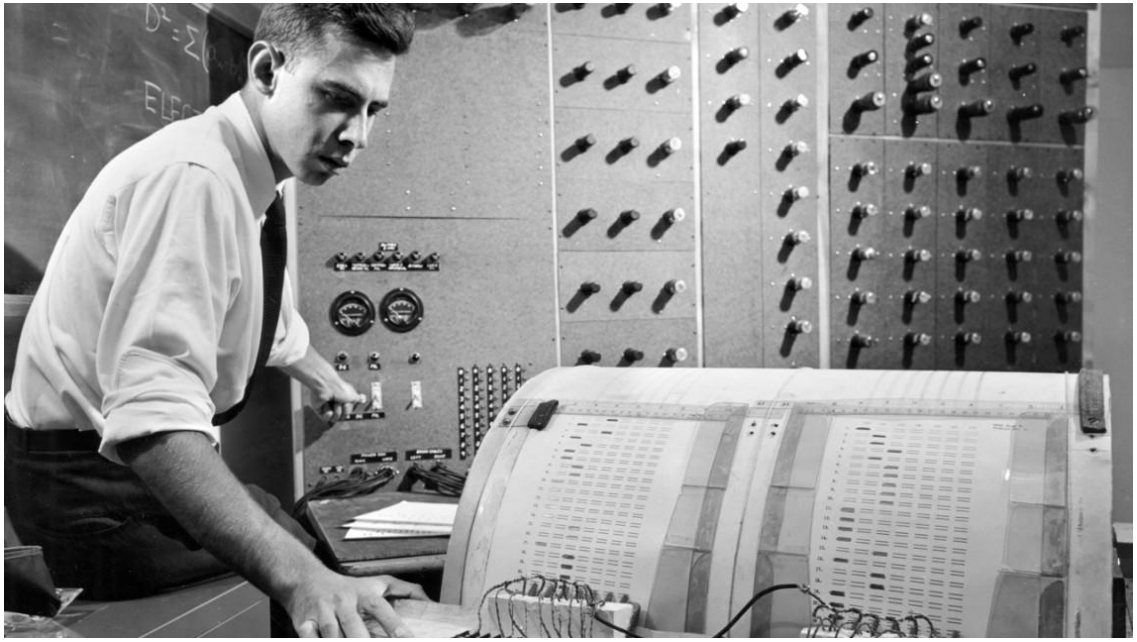
$$\text{Sign}(z) = \begin{cases} +1 & \text{if } z \geq 0 \\ -1 & \text{o/w} \end{cases}$$



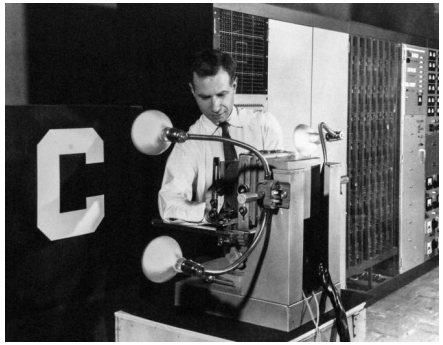
We know in general the problem of finding best w is NP-hard.

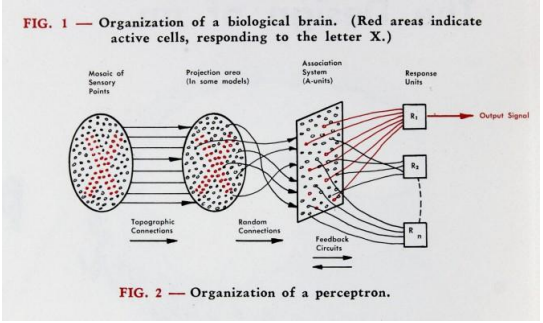
But is it still hard under linear separability assumption?

PERCEPTRON



Frank Rosenblatt ’50, Ph.D. ’56, works on the “perceptron” –
what he described as the first machine “capable of having an original idea.”





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

PERCEPTRON - ALGORITHM

Input: $\{(x_1, y_1) \dots\}$, $x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$

$w^0 = [0\ 0 \dots 0]$ $0 \in \mathbb{R}^d$. (Iteration)

until convergence

→ Pick (x_i, y_i) from dataset

IF $(\text{sign}(w^{tT} x_i) = y_i)$

do nothing

else

$\{ w^{t+1} = w^t + x_i y_i$ ($x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$) UPDATE RULE.

end.

end

UPDATE RULE

$$w^{t+1} = w^t + x_i y_i$$

Two types of mistake

Type 1

Pred → 1
act → -1

Type 2

Pred → -1
act → 1

Type -1

Pred → 1
act → -1

$w^{t^T} x_i \geq 0$ but $y_i = -1$

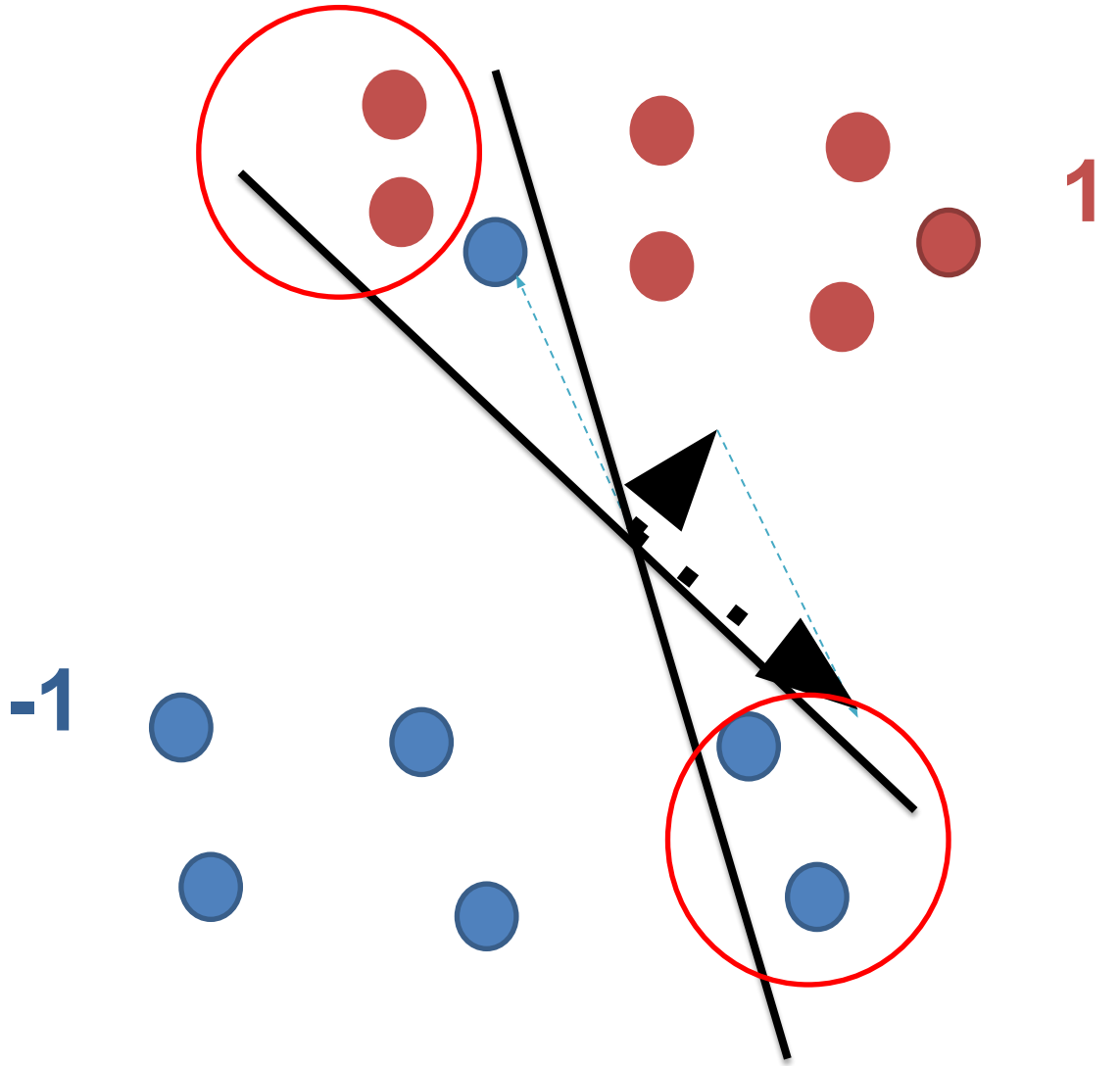
$$w^{t+1^T} x_i = \left(w^t + x_i y_i\right)^T x_i$$

$$= w^{t^T} x_i + y_i \|x_i\|^2$$

≥ 0 $\quad$ < 0

Type -2

≤ 0 $\quad$ > 0



Fixing one error might lead to more errors elsewhere

In general, does perceptron work for linearly separable data?

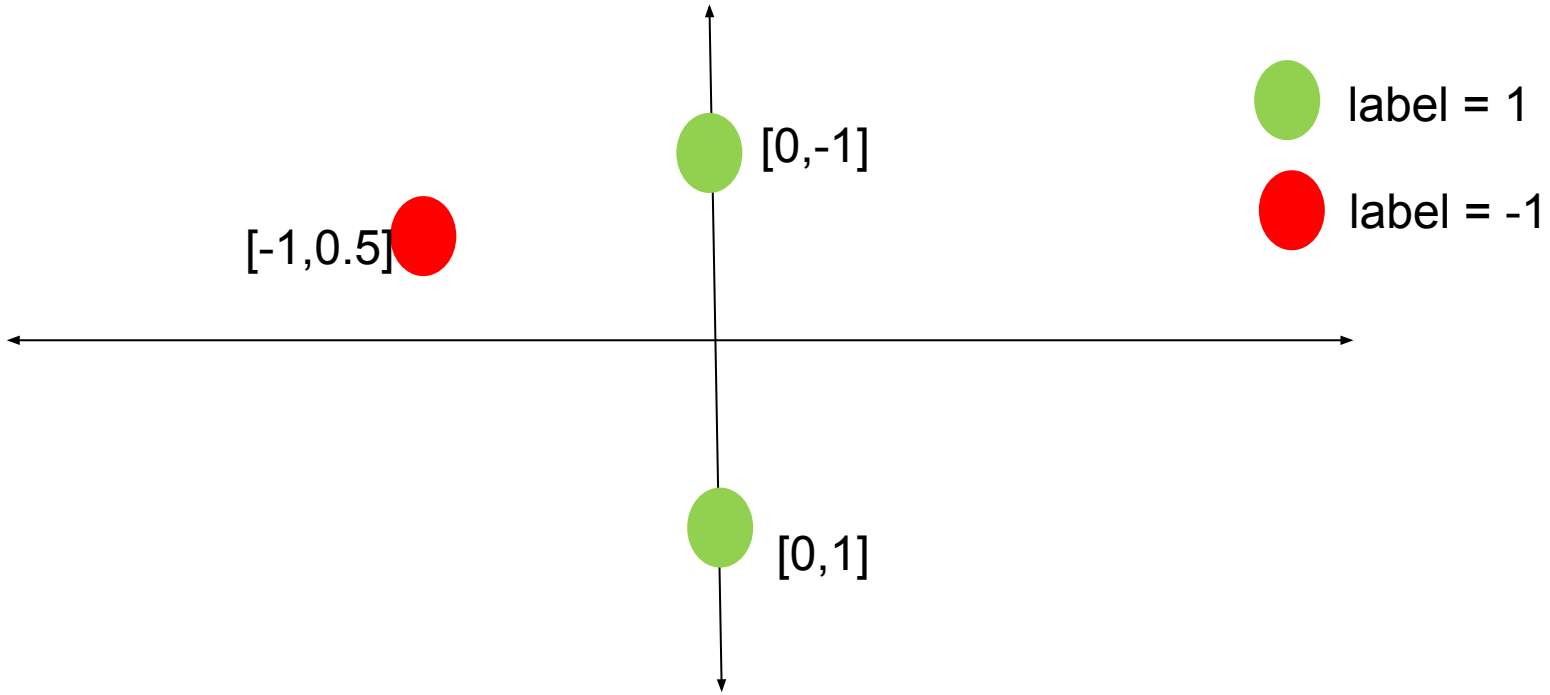
Recal
l

Lin. sep assumption :

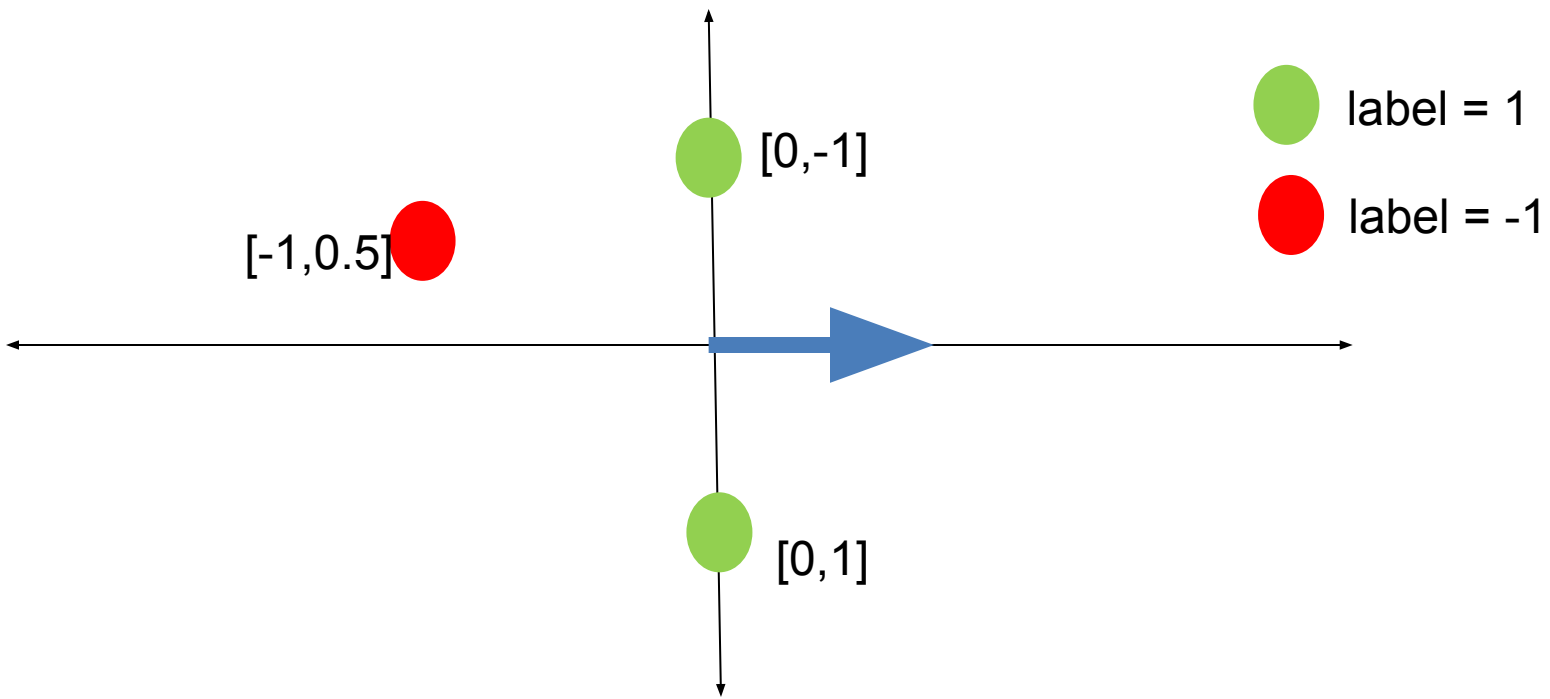
$\exists w \in \mathbb{R}^d$ s.t $\text{sign}(w^T x_i) = y_i \quad \forall i \in [n]$

$\{1, \dots, n\}$.

$\text{sign}(z) = \begin{cases} +1 & \text{if } z \geq 0 \\ -1 & \text{o/w} \end{cases}$

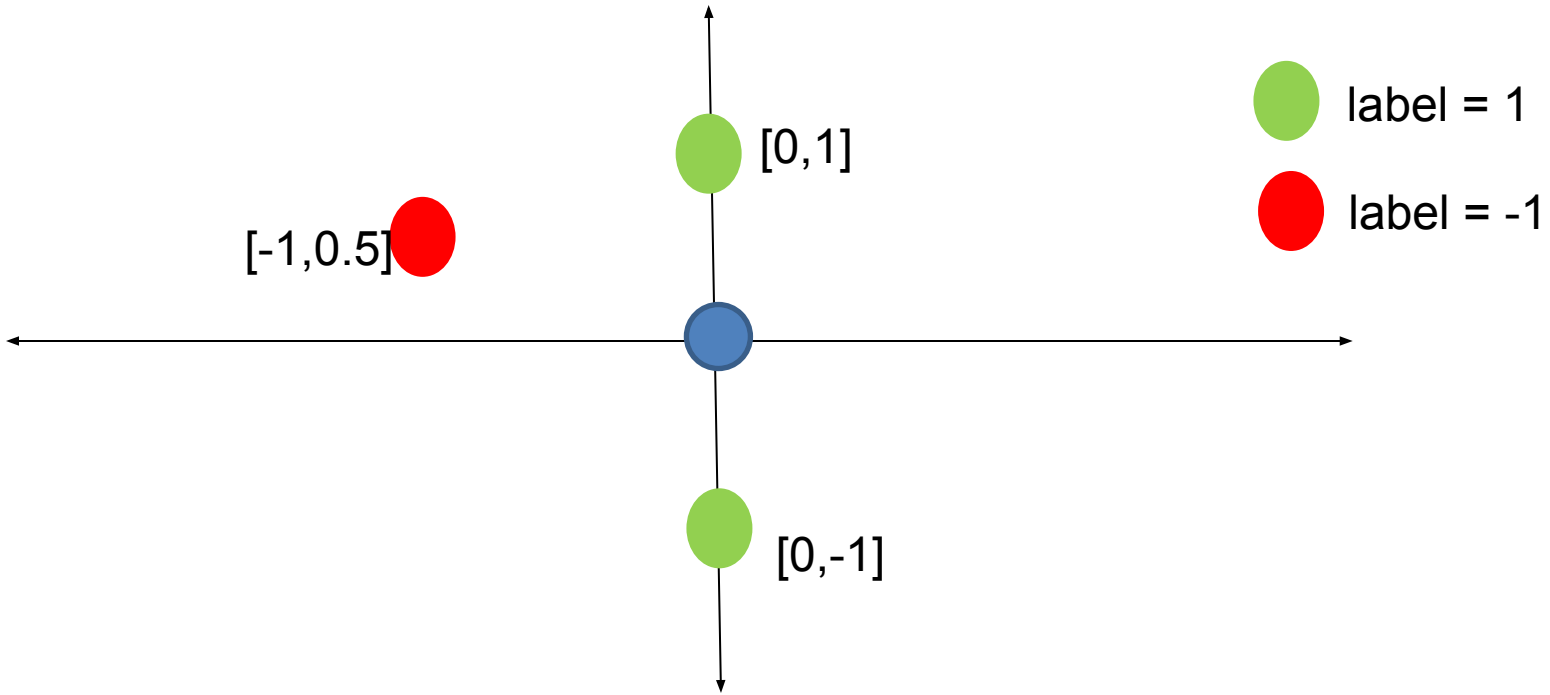


Is this a linearly separable dataset?



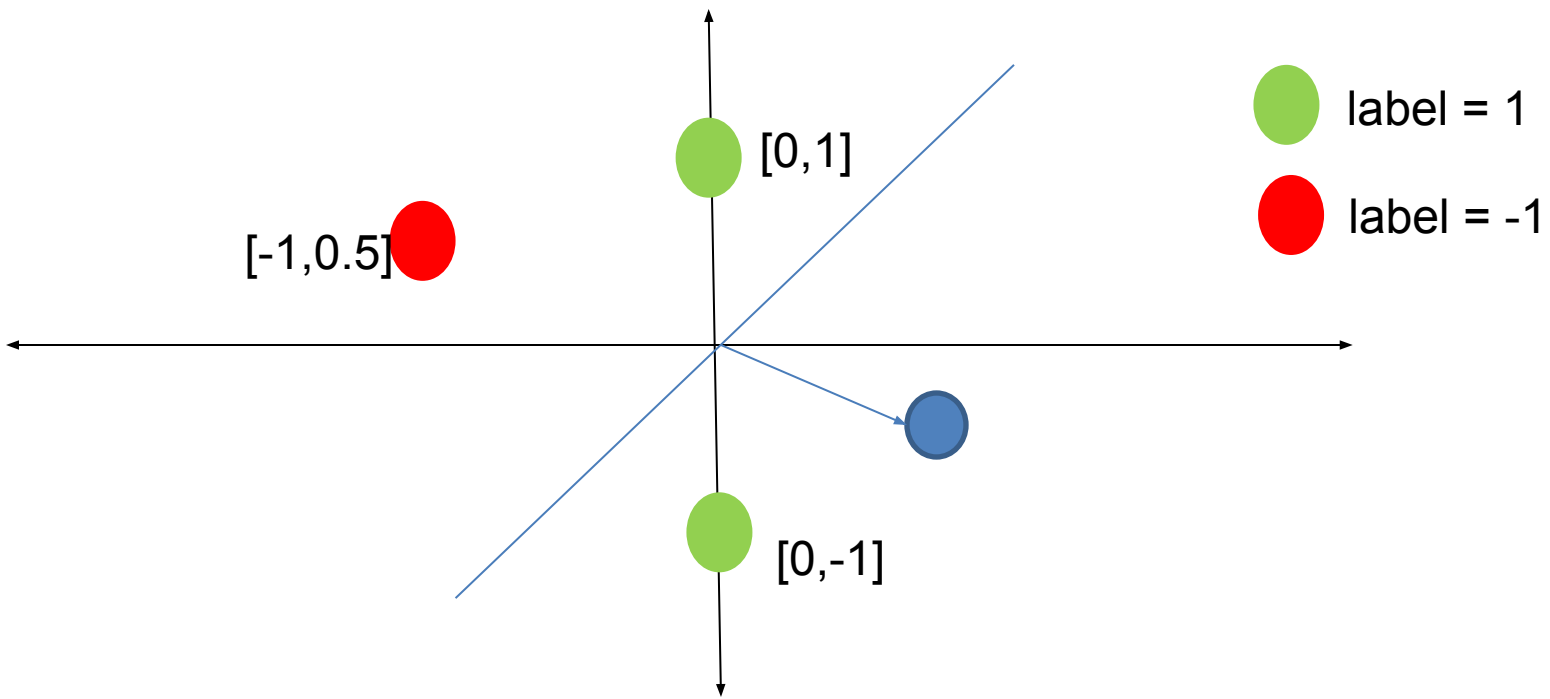
Any w in the positive x-axis linearly separates the data.
The dataset is linearly separable.

Let’s see what perceptron learns from this data!



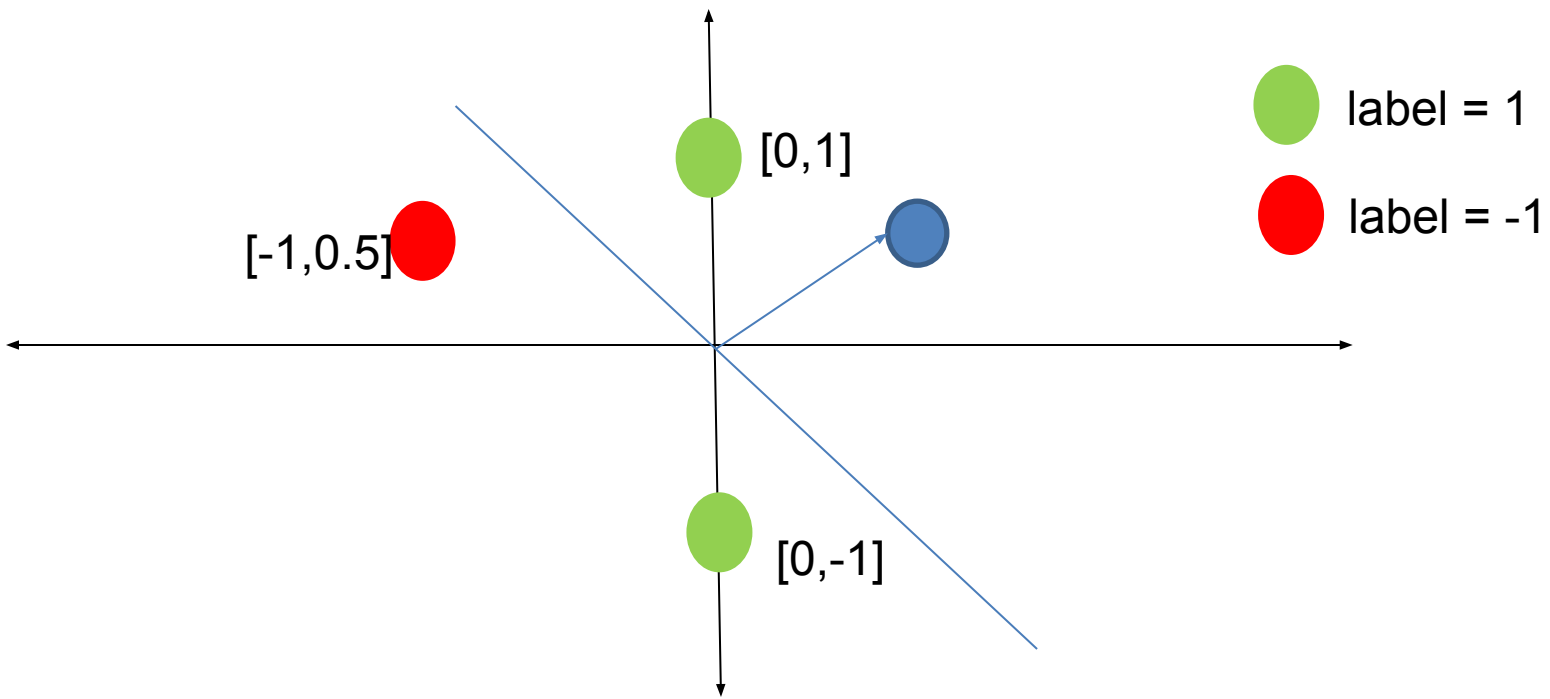
$w^0 = [0\ 0]$

	Predicted label	True label
[0 1]	1	1
[0 -1]	1	1
[-1 0.5]	1	-1



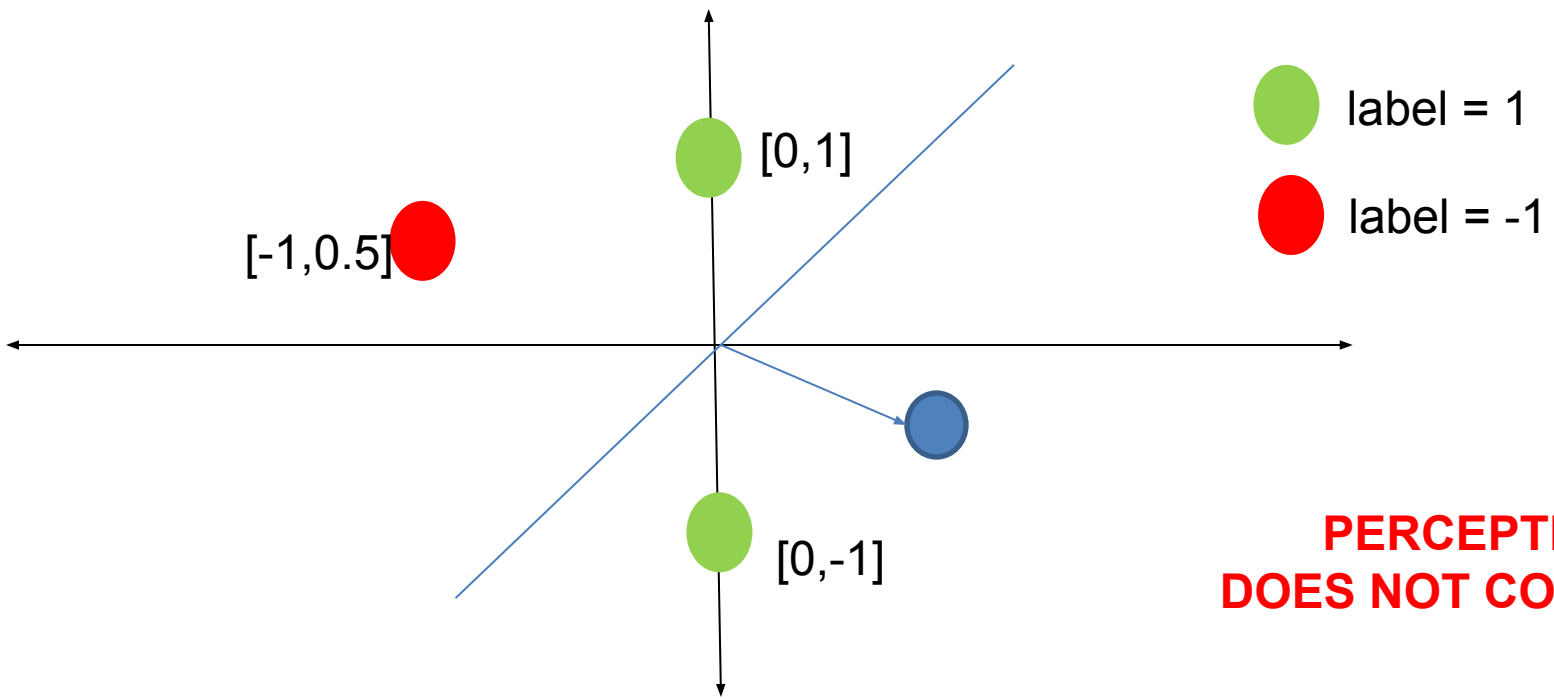
w^1= [1 -0.5]

	Predicted label	True label
[0 1]	-1	1
[0 -1]	1	1
[-1 0.5]	-1	-1



w^2= [1 0.5]

	Predicted label	True label
[0 1]	1	1
[0 -1]	-1	1
[-1 0.5]	-1	-1



PERCEPTRON DOES NOT CONVERGE

w^3= [1 -0.5]

	Predicted label	True label
[0 1]	-1	1
[0 -1]	1	1
[-1 0.5]	-1	-1

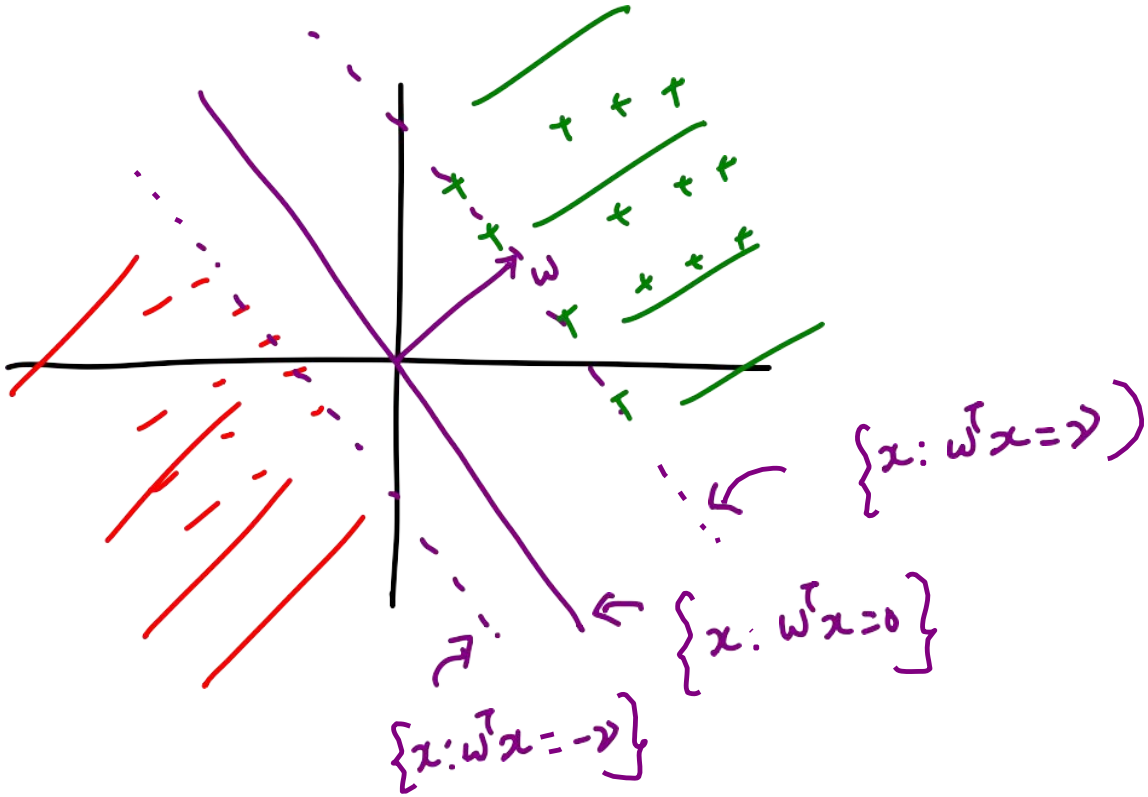
Issue

- optimal $w^* = \begin{bmatrix} c \\ 0 \end{bmatrix}$ $c > 0$ has datapoints that lie on the linear separator.

- If we assume this isn't the case, will perceptron converge?

ASSUMPTIONS

- LINEAR SEPERABILITY with γ - MARGIN



A Dataset $D = \{(x_1, y_1) \dots, (x_n, y_n)$ is L.S

with γ-margin if $\exists\, w^*$ s.t

$$(w^{*T} x_i)\, y_i \geq \gamma \quad \forall i \quad \text{for some } \gamma > 0$$

PERCEPTRON

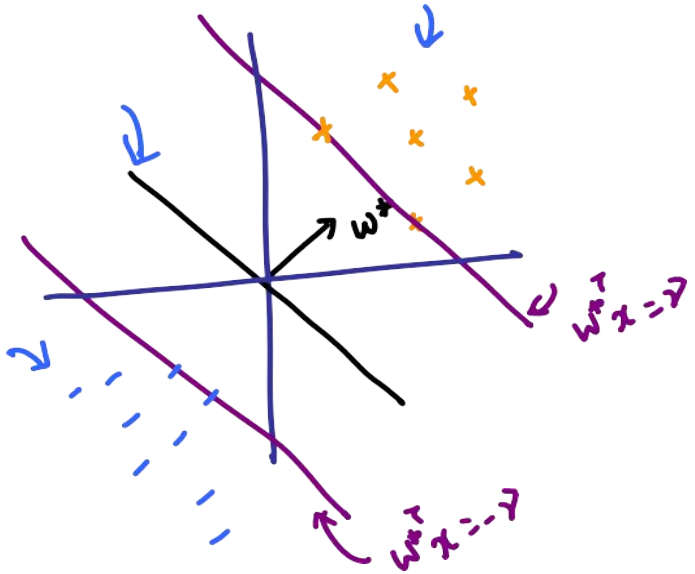
$$w^{t+1} = w^{t} + x_i y_i$$

ASSUMPTIONS

1. **Linear separability with γ margin**

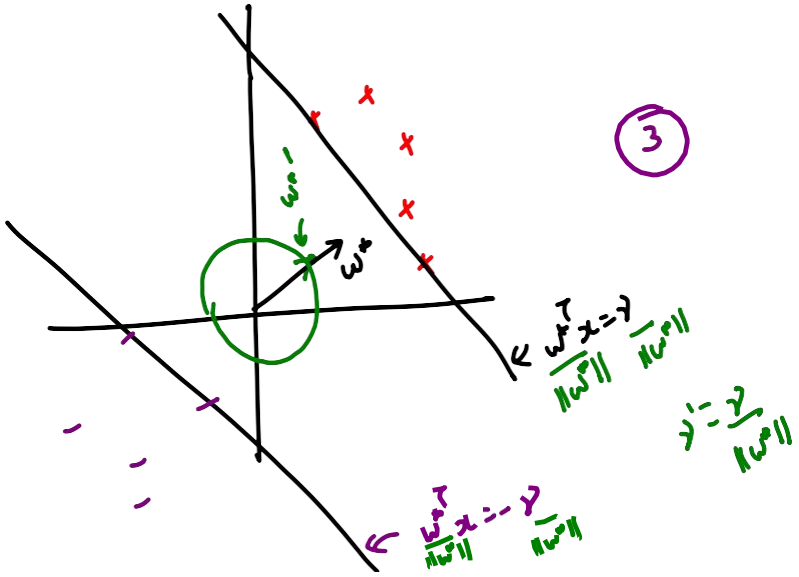
A dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is L.S with γ margin

if $\exists\, w^*$ s.t $(w^{*T} x_i) y_i \geq \gamma \quad \forall i$ for some $\gamma > 0$



(2) RADIUS ASSUMPTION

$$\forall i \in D, \quad \|x_i\|_2 \leq R \quad \text{for some } R > 0$$



(3) Without loss of generality

$$\|w^*\| = 1$$

ANALYSIS OF "mistakes" of Perceptron.

- observe that an update happens only when a "mistake" occurs.
- Say the current guess = w_t and a mistake happens w.r.t (x, y)

$$w_{t+1} = w_t + x \cdot y$$

$$\|w_{t+1}\|^2 = \|w_t + x \cdot y\|^2$$

$$= (w_k + x \cdot y)^T (w_k + x \cdot y)$$

$$\|w_{k+1}\|^2 = \|w_k\|^2 + \underbrace{2(w_k^T x) \cdot y}_{\leq 0 \text{ because [mistake]}} + \underbrace{(x^T x) \cdot y^2}_{\|x\|^2 \, 1 \leq R^2}$$

$$\leq \|w_k\|^2 + R^2 \leq \left(\|w_{k-1}\|^2 + R^2\right) + R^2$$

$$\|w_0\|^2 + kR^2$$

Inductively

$$\boxed{\|w_{k+1}\|^2 \leq k \cdot R^2} \quad \text{——— (1)}$$

$$w_{k+1}^T w^* = (w_k + x \cdot y)^T w^*$$

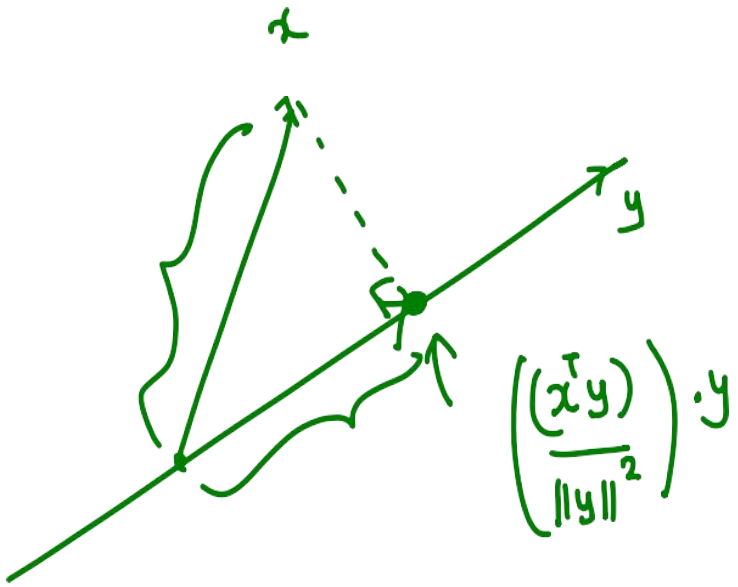
$$= w_k^T w^* + (w^{*T} x)\, y$$

$$\geq \gamma$$

$$\geq \left(w_{k-1}^T w^* + \gamma\right) + \gamma$$

$$\vdots$$

$$\boxed{w_{k+1}^T w^* \geq k \cdot \gamma}$$



$$\|x\|^2 \geq \left\| \left(\frac{x^T y}{\|y\|^2} \right) y \right\|^2$$

$$\geq \frac{(x^T y)^2}{\|y\|^4} \cdot \|y\|^2$$

$$\Rightarrow (x^T y)^2 \leq \|x\|^2 \|y\|^2$$

$$x^T y \leq \|x\| \|y\|$$

(C.S) Cauchy Schwartz inequality $\longrightarrow$

From before

$$w_{k+1}^T w^* \geq k\cdot\gamma$$

$$\|w_{k+1}\|^2 \|w^*\|^2 \geq \left(w_{k+1}^T w^*\right)^2 \geq k^2\gamma^2$$

$$\uparrow \text{C.S}$$

$$\|w_{k+1}\|^2 \geq k^2\gamma^2 \qquad -(2)$$

Combining ① & ②.

$$k^2\gamma^2 \leq \|w_{k+1}\|^2 \leq kR^2$$

② ↳ ↳ ①

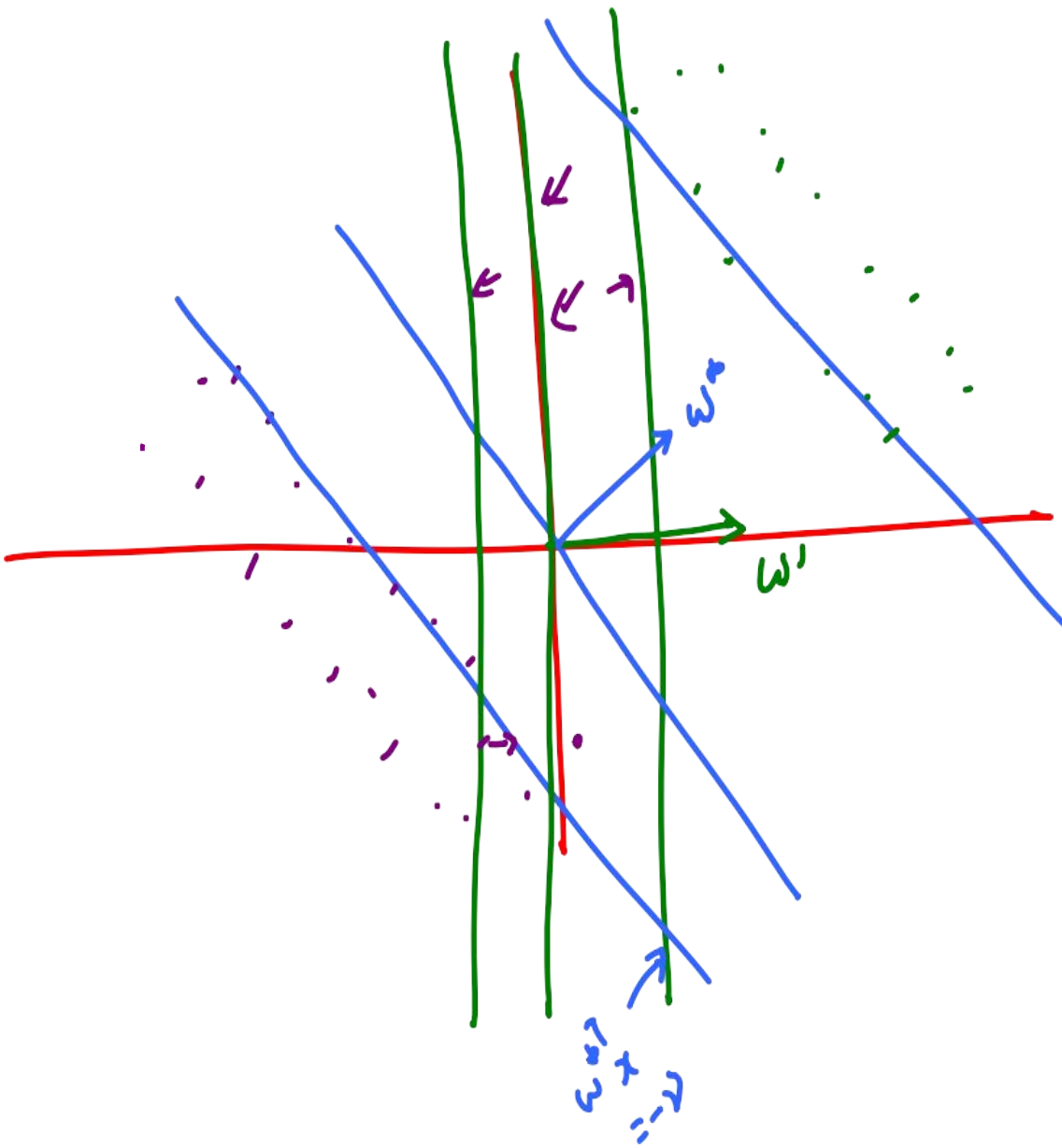
$$\Rightarrow \quad k^2\gamma^2 \leq kR^2$$

$$\boxed{k \leq \frac{R^2}{\gamma^2}}$$

RADIUS-MARGIN BOUND.

$\Rightarrow$ # mistakes of Perceptron is bounded

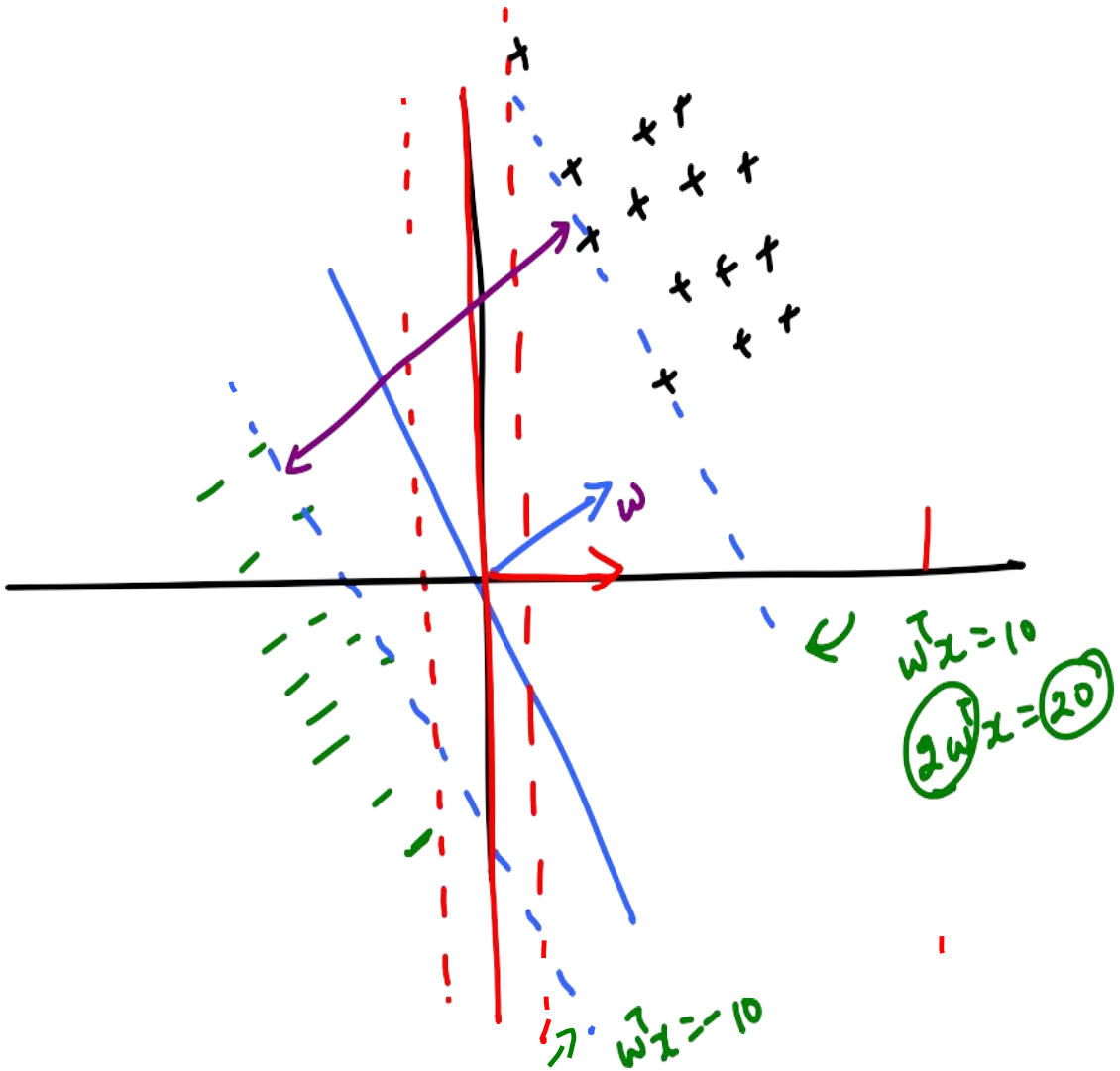
$\Rightarrow$ Perceptron converges!



Perceptron's # mistakes depends on w^*.

But it might output w'.

Goal: To come up with algorithms that maximize margin



Blue is better than Red.

Goal: Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

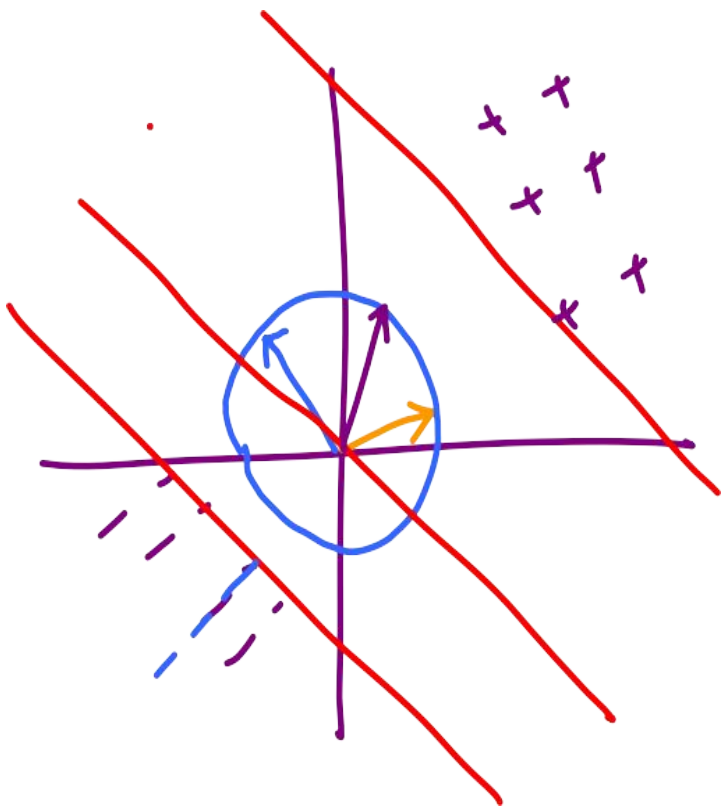
find a w with maximum width s.t all points in dataset D are classified correctly

$$\max_{w, \gamma} \quad \gamma$$

$$\text{s.t} \quad (w^T x_i) y_i \geq \gamma \quad \forall i$$

Issue:

Can scale w arbitrarily.

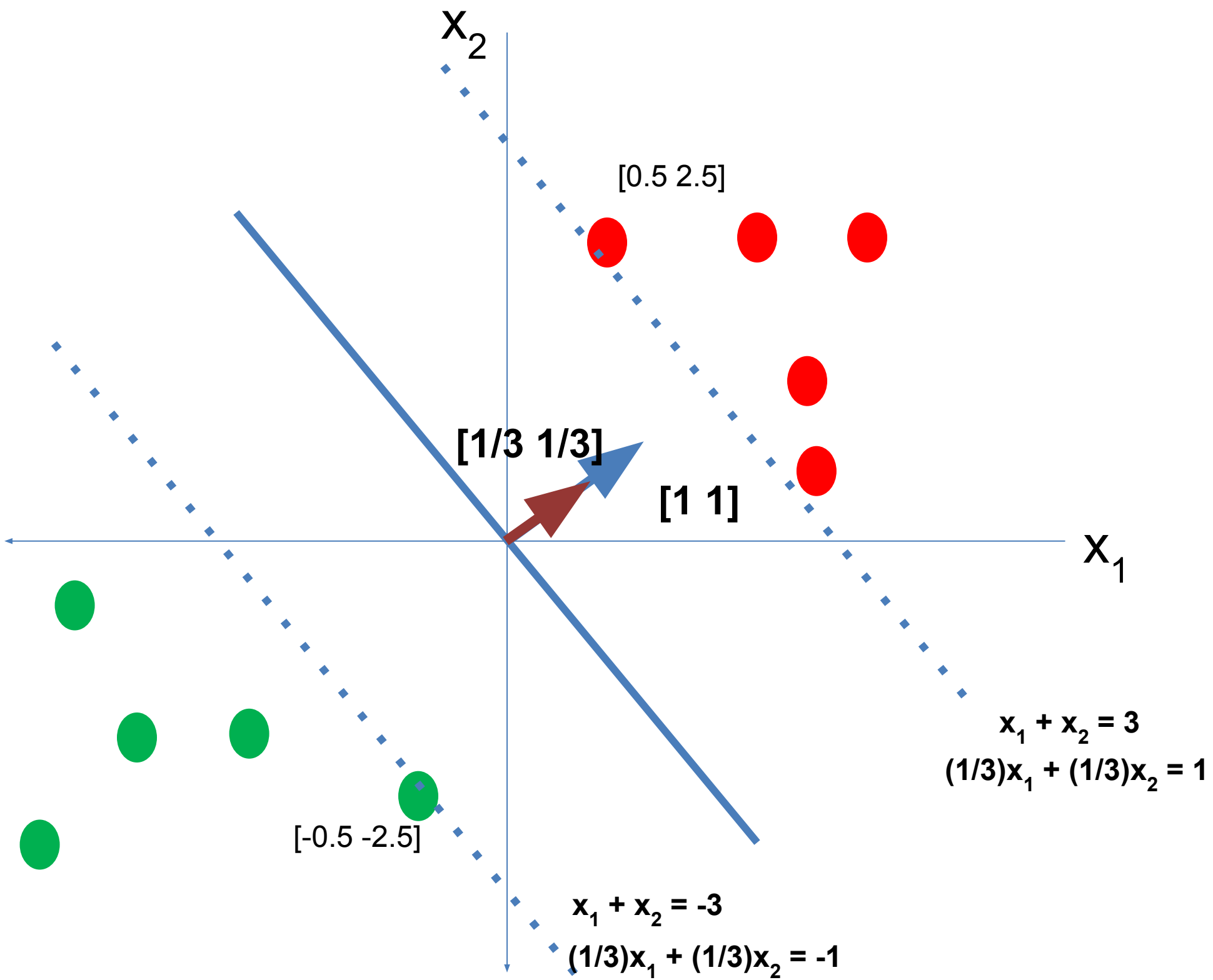


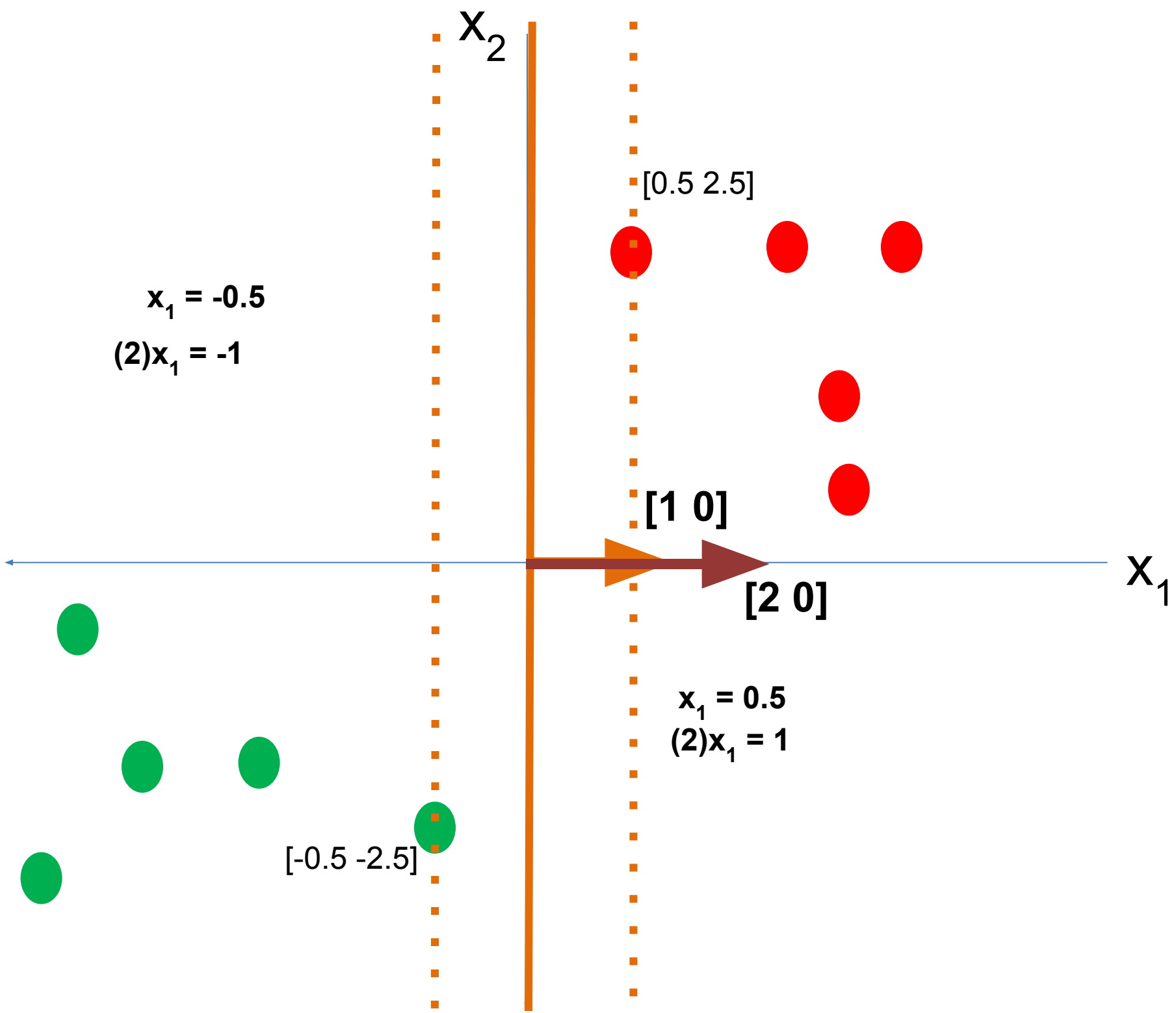
Possible fix

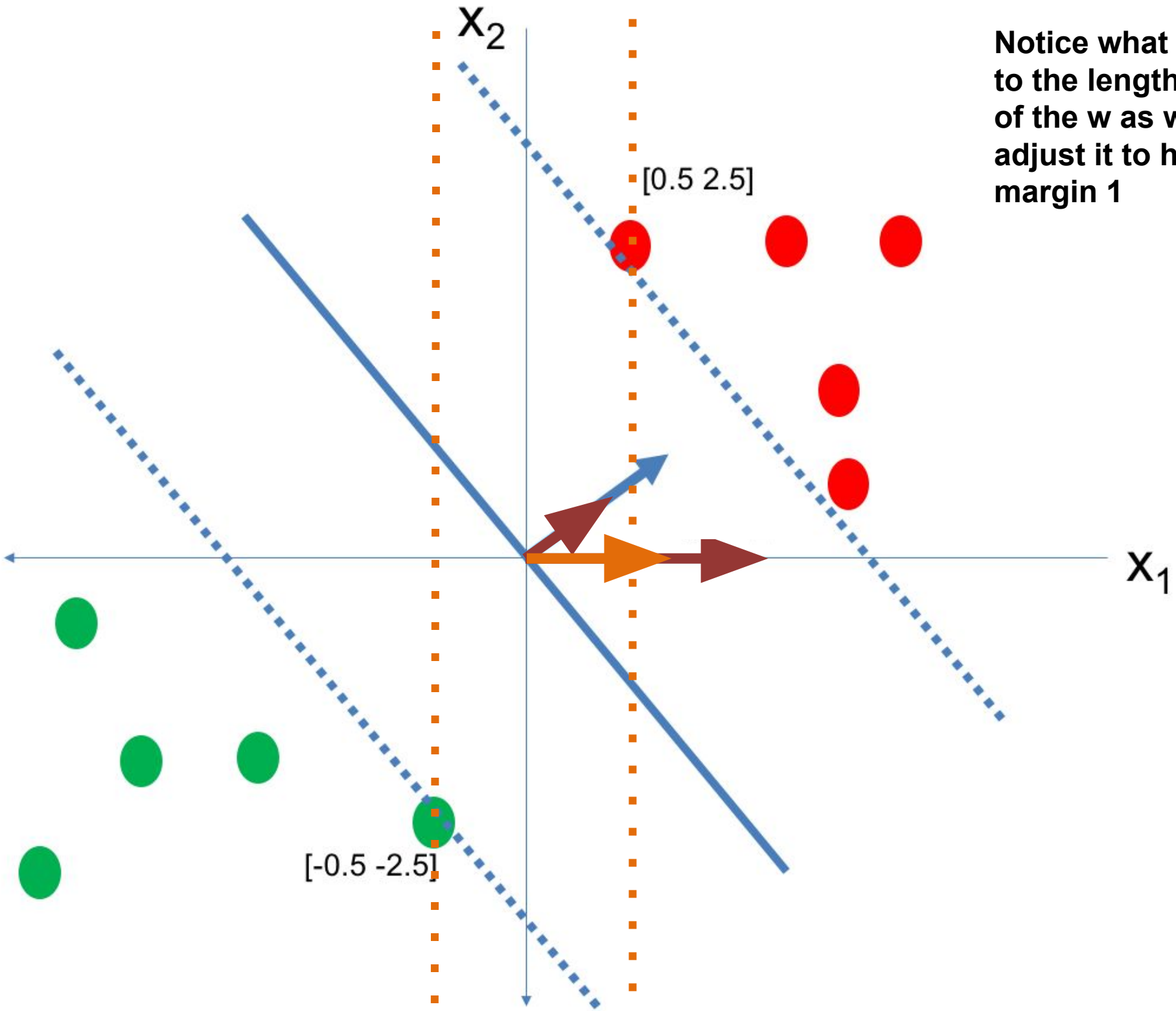
$$\max_{w,\gamma} \quad \gamma$$

$$(w^T x_i) y_i \geq \gamma$$

$$\|w\|^2 = 1$$


x_2
[0.5 2.5]
[1/3 1/3]
[1 1]
x_1
$x_1 + x_2 = 3$
$(1/3)x_1 + (1/3)x_2 = 1$
[-0.5 -2.5]
$x_1 + x_2 = -3$
$(1/3)x_1 + (1/3)x_2 = -1$


x_2
[0.5 2.5]
$x_1 = -0.5$
$(2)x_1 = -1$
[1 0]
x_1
[2 0]
$x_1 = 0.5$
$(2)x_1 = 1$
[-0.5 -2.5]



Notice what happens to the lengths of the w as we adjust it to have margin 1

OBSERVATIONS

-> Once a direction is fixed,
the width between the margin lines
is fixed

-> If the width is large, then the w that
achieves margin 1 in that direction
has smaller length

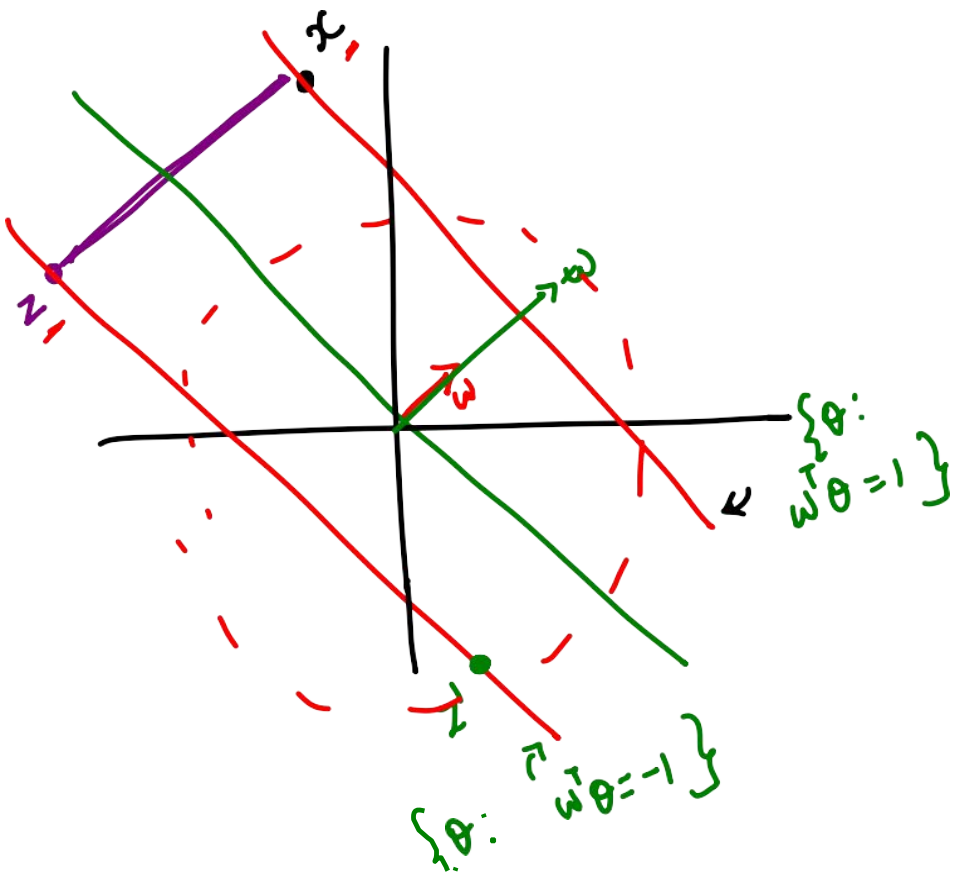
-> If the width is small, then the w that
achieves margin 1 in that direction
has larger length

-> In general, width(w) seems to be
inversely proportional to length(w)

$\max_w$ width(w)

s.t $(w^T x_i) y_i \geq 1 \quad \forall i$

What is width(w) ?



$$\min_{z:} \quad \frac{1}{2}\|x - z\|^2$$

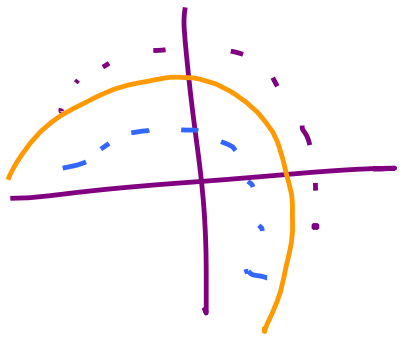
$$\text{s.t} \quad w^T x = 1, \quad w^T z = -1$$

[Exercise]

$$\text{width}(w) = \frac{2}{\|w\|^2}$$

$$\max_{w} \frac{2}{\|w\|^2}$$

$$\text{s.t} \quad \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption

$$\min_{w} \frac{1}{2}\|w\|^2$$

$$\text{s.t} \quad \forall i \quad (w^T x_i) y_i \geq 1$$

- Non-linear structure?

DETOUR

$$\min_w f(w)$$

$$g(w) \leq 0$$

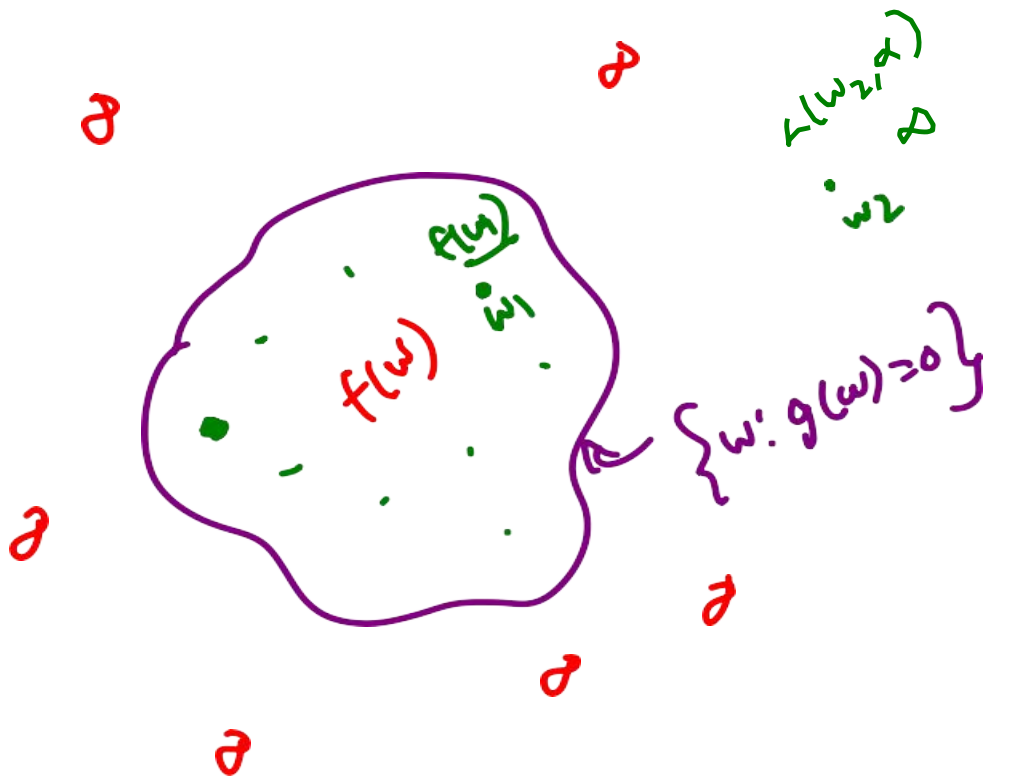
$$\mathcal{L}(w, \alpha) = f(w) + \alpha \cdot g(w)$$

Fix some w.

Consider

$$\max_{\alpha \geq 0} \; L(w, \alpha)$$

$$= \max_{\alpha \geq 0} \; f(w) + \alpha\, g(w)$$



$$\begin{cases} \infty & \text{if } g(w) > 0 \\ f(w) & \text{if } g(w) \leq 0 \end{cases}$$

$$\min_{w} \underbrace{\left[\max_{\alpha \geq 0} \mathcal{L}(w, \alpha) \right]}_{B(w)}$$

equivalent

$$\equiv \min_{w} f(w) \quad \text{s.t } g(w) \leq 0.$$

So far

Max-margin formulation

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t} \quad (w^T x_i) y_i \geq 1 \quad +i$$

— (A)

Detour

$$\min_w \quad f(w)$$

$$\text{s.t} \quad g(w) \leq 0$$

$\Leftrightarrow$

$$\min_w \left[\max_{\alpha \geq 0} \left[f(w) + \alpha\, g(w) \right] \right]$$

- Can we swap min and max?

Multiple Constraints

→ Same idea

$$\min_w f(w) \quad \text{s.t} \quad g_i(w) \le 0 \quad \forall i = 1 \dots k$$

$$\equiv \min_w \left[\max_{\{\alpha_1, \dots \alpha_k \ge 0\}} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) \right] \right]$$

$$\equiv \text{Strong duality for convex } f, g_i$$

$$\max_{\alpha_1, \dots, \alpha_k \ge 0} \min_w f(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

$$\min_{w} \quad \underbrace{\frac{1}{2}\|w\|^2}_{f(w)}$$

$$\text{s.t} \quad \underbrace{(w^T x_i) y_i \geq 1}_{1-(w^T x_i) y_i \leq 0} \quad \forall i$$

$$g_i(w) = 1 - (w^T x_i) y_i$$

$$\mathcal{L}(w, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

$\alpha \in \mathbb{R}^n$

$$\min_{w} \left[\max_{\alpha \geq 0} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

$\alpha \geq 0 \rightarrow \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \geq 0$

$$\Big\Vert$$

$$\max_{\alpha \geq 0} \left[\min_{w} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Fix $\alpha \geq 0$

$$\min_{w} \left[\frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i \right) \right]$$

Grad w.r.t w

$$w^* + \sum_{i=1}^{n} -\alpha_i x_i y_i = 0$$

$$w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

$x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$, α_i: Fixed Choice

In matrix notation

$$w^* = X Y \alpha$$

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}_{d \times n} \quad Y = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_n \end{bmatrix}_{n \times n} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}_{n \times 1}$$

Substituting $\overset{\text{soln}}{=}$ back in the objective.

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - w^T x_i) y_i)$$

$$= \frac{1}{2} \underline{w^T w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i (w^T x_i) y_i$$

$$\frac{1}{2}(xy\alpha)^T(xy\alpha) + \alpha^T 1 - \sum_{i=1}^{n}(xy\alpha)^T x_i\, y_i\, \alpha_i$$

$\alpha^T 1$: $\begin{bmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{bmatrix}\begin{bmatrix}1 \\ \vdots \\ 1\end{bmatrix}$

On Simplification [please do this]

$$\alpha^T 1 - \frac{1}{2}(xy\alpha)^T(xy\alpha)$$

can be KERNELIZED!

DUAL PROBLEM

$$\max \quad \alpha^T 1 - \frac{1}{2}\alpha^T y^T x^T x\, y\, \alpha$$

$x^T x$: $n \times n$.

$\alpha \geq 0$: easy constraints

Solving in n instead of d.

Revisiting The Lagrangian

$$\min_{w} \left[\max_{\alpha \geq 0} f(w) + \alpha g(w) \right] \equiv \max_{\alpha \geq 0} \left[\min_{w} f(w) + \alpha g(w) \right]$$

PRIMAL

DUAL

w^*

α^*

$$\max_{\alpha \geq 0} f(w^*) + \alpha g(w^*) = \min_{w} f(w) + \alpha^* g(w)$$

$$f(w^*) = f(w) + \alpha^* g(w')$$

$$f(w^*) \leq f(w^*) + \alpha^* g(w^*)$$

$$\Rightarrow \boxed{\alpha^* g(w^*) \geq 0}$$

But we know $\alpha^* g(w^*) \leq 0$

$$\Rightarrow \boxed{\alpha^* g(w^*) = 0} \rightarrow \text{COMPLEMENTARY SLACKNESS}$$

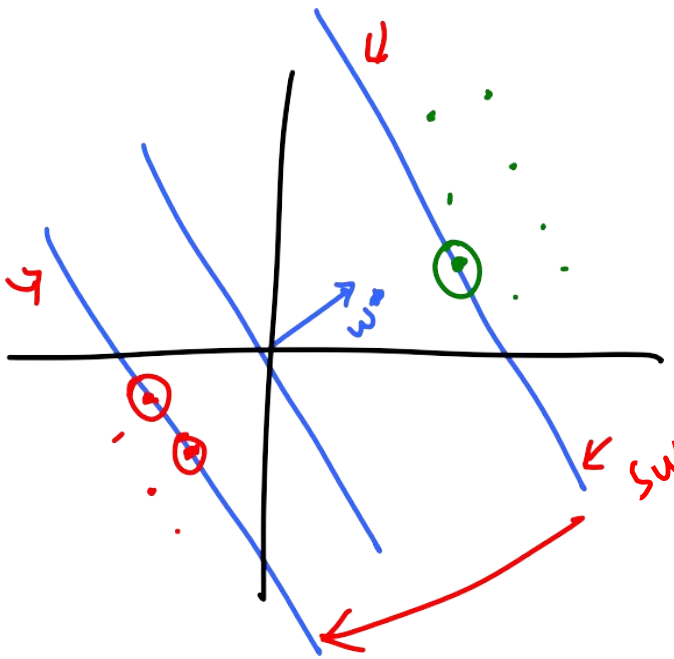
For multiple constraints

$$\alpha_i^* g_i(w^*) = 0 \quad \forall i$$

For our problem

$$\boxed{\alpha_i^* \left(1 - (w^{*T} x_i) y_i\right) = 0 \quad \forall i}$$

$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i} \leftarrow$$

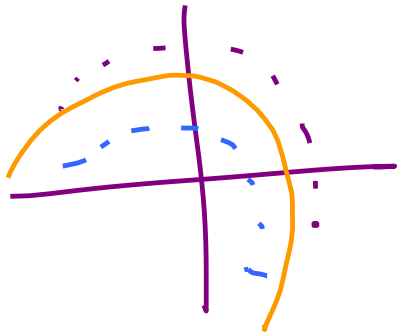


If $\alpha_i^* > 0 \Rightarrow 1 - (w^{*T} x_i) y_i = 0$

$\Rightarrow (w^{*T} x_i) y_i = 1$

$$\min_{w} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption

- Non-linear structure?