Ok, we've analyzed that for a given $\mathcal{H}$

Uniform convergence

(Agnostic) PAC learnability

ERM being a (n Agnostic) PAC learner

VCdim $(\mathcal{H})$ is bounded

are equivalent.

ERM often isn't efficient, certainly in the agnostic setting..

But it is efficient for realizeable linear separators: $\mathcal{H}_{lin\text{-}d}$

Pick $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ s.t. $sgn(\langle w, x_i \rangle + b) = y_i \quad \forall (x_i, y_i) \in S$

$VCdim(\mathcal{H}_{lin\text{-}d}) = d+1$

→ LB

Shattereable set:

$x_0 = 0$
$x_1 = e_1$
$\vdots \quad \vdots$
$x_d \quad e_d$

For y_0 $\vdots$ y_d

Pick $b = y_0 \Rightarrow \langle w, x_0 \rangle + b = b = y_0$ ✓

$w_i = y_i - b \Rightarrow \langle w, x_i \rangle + b = \langle w, e_i \rangle + b = w_i + b = y_i$ ✓

→ UB

Take any set $|C| > d$ in $\mathbb{R}^d$.

WTS it cannot be shattered by $\mathcal{H}_{lin\text{-}d}$

[find a set of labels we can't induce]

[Let's ignore b and just argue about the bias-less ones]

Since $|C| > d+1$, C is not linearly $\perp$

$\Rightarrow \exists \vec{a} \neq 0$ s.t.

$\vec{a}\,\vec{x} = 0$

$P \triangleq \{i : a_i > 0\}$

$N \triangleq \{j : a_j < 0\}$

Neither can be empty. Why?

$\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$

$\vec{a} \neq \vec{0}$ so at least one is nonempty

[If on is $\neq \emptyset$, the other can't be...]

Suppose w induces labels $\mathrm{sgn}(a_i)$ for x_i

This means $a_i \cdot \langle w, x_i \rangle > 0$ $\forall i$.

$$0 < \sum_{i \in P} a_i \cdot \langle w, x_i \rangle = \langle \sum_{i \in P} a_i x_i , w \rangle = \langle \sum_{j \in N} |a_j| x_j , w \rangle = \sum_{j \in N} |a_j| \langle x_j, w \rangle < 0$$

since $\mathrm{sgn}\langle x_j, w \rangle = -1$

A contradiction!
So C can't be shattered by $\mathcal{H}_{lin-d}$

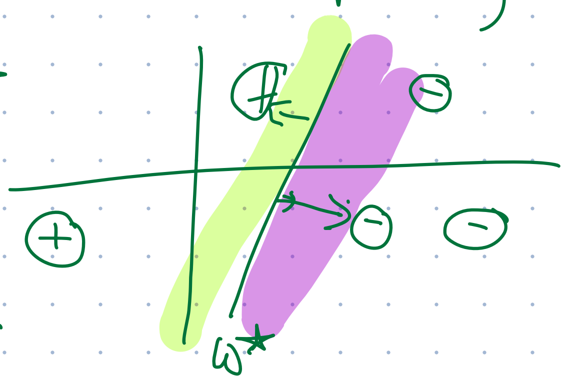
This means ERM on $m \geq C \cdot \frac{d + \log \frac{1}{\delta}}{\epsilon}$ samples PAC-learns linear realizable separators in $\mathbb{R}^d$.

[When not realizable, it's computationally hard :( ]

[Simple Alg: use Linear Programming! Other things also work...

If our data is not only realizable (separable) but also

Has a <u>margin</u>



min width of ▮ ▮ is the margin of w^*

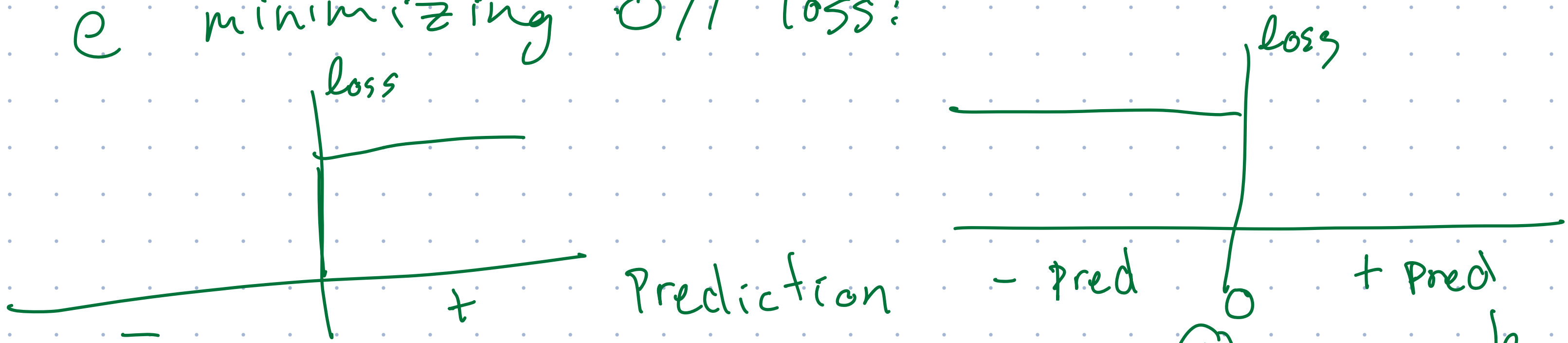
then we can learn with many fewer samples

$$w^* = \arg\min_w \|w\| \text{ s.t. } y_i \langle w, x_i \rangle \geq 1 \ \forall i \in [m]$$

.... $\Rightarrow$ & $\|x\| \leq R$

Finally, what is hard about non-realizable setting?

Our 2nd-fave alg, Gradient Descent, isn't good @ minimizing 0/1 loss:

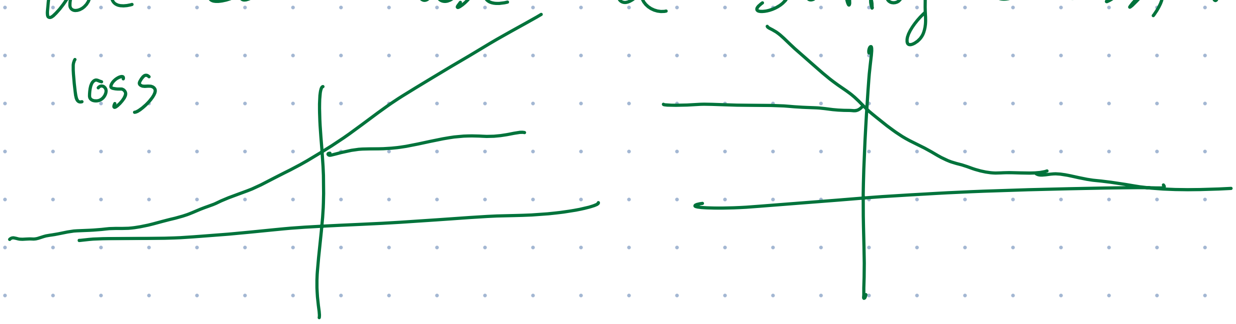


(–) example

(+) example

It's non convex,
non-smooth
doesn't have a gradient!

We can use a surrogate loss, like logistic



Convex, smooth,
has gradients

$\ell_w(x, y) = \log(1 + \exp(-y_i \langle w, x_i \rangle))$

So, $|L_D(\text{logistic}) - L_S(\text{logistic})|$ will be small (Uniform convergence!) but no guar that 0-1-L_S(logistic) will be small even if $\exists$ w w. small 0-1 L_S

What's another set of methods for arguing about good $L_D(h)$?

$$L_D(h) = |L_D(h) - L_S(h)| + L_S(h)?$$