Need $\Theta\left(\frac{\ln\frac{1}{\delta}}{\epsilon}\right)$ samples to learn $\hat{r}$ which has error $\leq \epsilon$ wp $1-\delta$.

Assume no pt masses in D

Pf: The errors our alg makes are only in the shaded region (realizable & smallest $\bigcirc$ containing all training +s).

$H_{Bad} = \{\hat{r} : D(\hat{r}, r^*) \geq \epsilon, \hat{r} < r^*\}$

$\hat{r} \to r_{max}$ is the max

$\Pr[\text{selecting } \hat{r} \in H_{BAD}]$

$\leq \Pr[\text{We see no samples} \in B(r^*, 0) \setminus B(r_{max}, 0)]$

$= (1-\epsilon)^m \leq e^{-\epsilon m}$

Recap: Finite H $\xrightarrow{\text{realizable}}$ $\Theta\left(\frac{\log\frac{|H|}{\delta}}{\epsilon}\right)$ - PAC learnable

Finite H $\xrightarrow{\text{Non}}$ $\Theta\left(\frac{\log\frac{|H|}{\delta}}{\epsilon^2}\right)$ - Agnostically PAC learnable

Others? Finiteness isn't necessary ($\uparrow$, circle example)

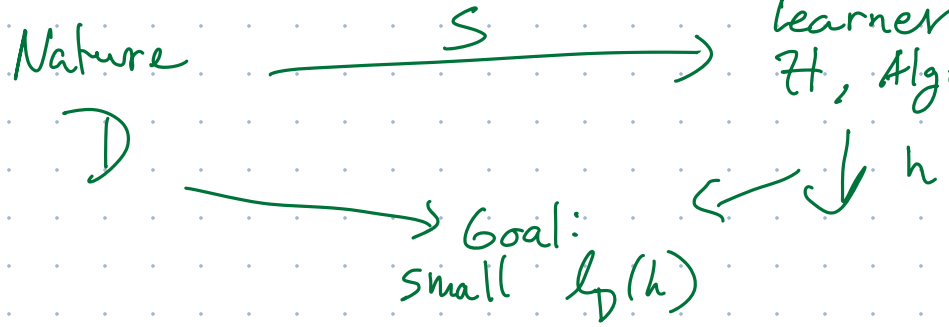
But we need info about D or H...

Our prior knowledge was that

- D had a ground truth circular decison bd
- H was the set of circular decison bds

Q: Was this (necessary)$^{\text{yes}}$ or ~~just what we used~~?

Learning Problem $P(X, Y)$ fixed.



Q: Can there be a universal alg that PAC learns every problem P?

Q': Is prior ~~k~~nowlege necessary for learning?

necessary for PAC

Parametric D
Realizable $\mathcal{H}$
Small $\min_{h \in \mathcal{H}} L_D(h)$

How much prior info should/can we assume?

A lot
- Learner knows D, h^* (Bayes opt) $\rightarrow \mathcal{H} = \{h^*\}$ achieves min risk
- Learner knows a class
  - Realizable finite $\mathcal{H}$ $\Rightarrow$ PAC $m \approx \frac{\log |\mathcal{H}|/\delta}{\epsilon}$
  - Some realizable inf $\mathcal{H}$ $\Rightarrow$ " "
  - Nonrealizable finite $\mathcal{H}$ $\rightarrow$ Agnostic PAC $m \approx \frac{\ln \mathcal{H}/\delta}{\epsilon^2}$
- Learner knows nothing:

Nothing

No free lunch thm:

Let Alg be any learning alg, $Y = \{0,1\}$, any X & $m \leq \frac{|X|}{2}$. Then $\exists$ D over $X \times Y$ s.t.

1. $\exists f : X \to Y$ w. $L_D(f) = 0$
2. $P_S(L_D(Alg(S)) \geq \frac{1}{8}) \geq \frac{1}{7}$

Pf idea Randomly label all of $\mathcal{X}$, make dist uniform over $\mathcal{X}$. S will contain $\leq \frac{|\mathcal{X}|}{2}$ samples, on the remaining $\frac{|\mathcal{X}|}{2}$ or $1/2$ of samples, can only guess it, error $1/2$ (Total error $\geq 1/4$)

But $f(x) = ?$, $L_D(f) = 0$.

Corollary With infinite $\mathcal{X}$ and $\mathcal{H}$ all fns $: \mathcal{X} \to \{0,1\}$ $\mathcal{H}$ is not PAC learnable.

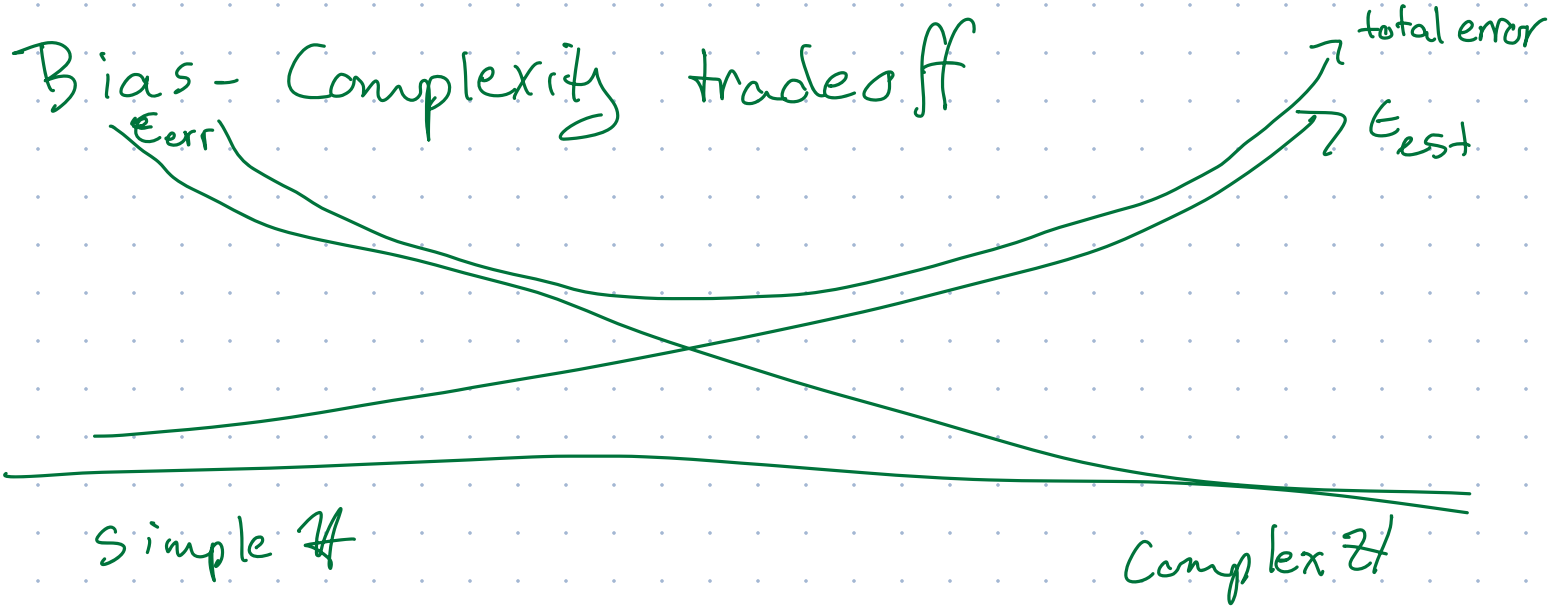
$$L_D(h_S) = \left(L_D(h_S) - \min_{h' \in \mathcal{H}} L_D(h')\right) + \min_{h' \in \mathcal{H}} L_D(h')$$

$L_D(h_S)$ → ERM

$\left(L_D(h_S) - \min_{h' \in \mathcal{H}} L_D(h')\right)$: ϵ_{est}, estimation err
- generally ↓ w m ↑ for Cawhile
- Depends on complexity of $\mathcal{H}$,

$\min_{h' \in \mathcal{H}} L_D(h')$: ϵ_{approx}, approxin.
- $\frac{1}{}$ of m
- fn of $D, \mathcal{H}$

Bias-Complexity tradeoff



What $\mathcal{H}$'s are PAC learnable?

→ Vlademir / Vapnik / Alexey / Chervonenkis 1970

Def (Restrict $\mathcal{H}$ to C)

$\mathcal{H}$ is a class of fns $\mathcal{X} \to \{0,1\}$ & $C = (c_1, \dots, c_m) \in \mathcal{X}$

The restriction of $\mathcal{H}$ to C is

$$\mathcal{H}_C \triangleq \{ \tilde{h} : C \to \{0,1\} \mid h \in \mathcal{H} \}$$

$$c_i \to h(c_i)$$

[h can be represented by vec $(h(c_1), \dots, h(c_m))$

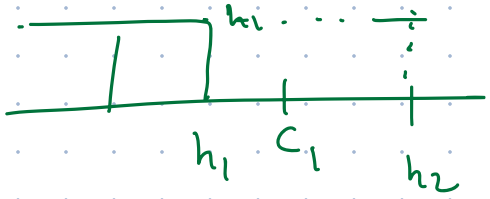
Def [Shattering]

$\mathcal{H}$ shatters $C \subseteq \mathcal{X}$ if $\mathcal{H}_C$ is the set of all fns from $C \to \{0,1\}$, $|\mathcal{H}_C| = 2^{|C|}$.

Ex $\mathcal{H} = \{ h_a : \mathbb{R} \to \mathbb{R} = \mathbb{I}_{(x \le a)} \}$ (all threshold fns)

$C = \{c_1\}$

$\mathcal{H}$ shatters C



Ex2 $C' = \{c_1, c_2\}$

$\mathcal{H}$ doesn't shattr C'

Can't get
$h(c_1) = 0$
$h(c_2) = 1$

Def | VC dimension

The VC dimension of $\mathcal{H}$, $VC(\mathcal{H})$, is the maximum-sized set C which is shattered by $\mathcal{H}$. (No larger sets shattreable by $\mathcal{H}$).

To Prove $VCdim(\mathcal{H}) = d$, • find a witness set $|C| = d$ which can be shattered
• & argue no $|C'| > d$ can be shattered.

Ex: Thresholds ↑

$VCdim \geq 1$

$= 1$ (NTS $\nexists$ a set of size 2 shattreable)

Ex Intervals $\mathcal{H} = \{ \mathbb{I}_{a \leq x \leq b} : a \leq b \in \mathbb{R} \}$

$VC(Intervals) = 2$

Ex: Axis align rectangle $H = \{ \mathbb{I}_{\substack{a_1 \leq x_1 \leq b_1 \\ a_2 \leq x_2 \leq b_2}} : a_1 \leq b_1, a_2 \leq b_2 \}$

$VC(Rect) = 4$

Parameters is often $\approx$ VCdim But not always!!!

Ex: Finite $\mathcal{H}$

Shattering requires $|\mathcal{H}| \geq 2^{|C|}$

$VC(\mathcal{H}) = |C| \leq \lg_2 |\mathcal{H}|$

Thm! Fund Thm of Statistical Learning

Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\{0,1\}$, ℓ the 0-1 loss fn. Then the following are equivalent:

(1) $\mathcal{H}$ has uniform convergence
(2) Any ERM rule sucessfully agnostically PAC learns $\mathcal{H}$
(3) $\mathcal{H}$ is agnostically PAC learnable
(4) $\mathcal{H}$ is PAC learnable
(5) Any ERM rule PAC learns $\mathcal{H}$
(6) $\mathcal{H}$ has finite VC dimension

$1 \to 2$ Last lec
$2 \to 3$ Yup
$3 \to 4$ Yup
$2 \to 5$ Yup
$2 \to 6$ ($\neg 6 \to \neg 2$ by NFL)

Remains to show $6 \Rightarrow 1$