Consistency, Uniform Convergence, and Learning: The VC Dimension

Finite $\mathcal{H}$: Quick Review

- **Any** consistent h has $\epsilon(h) \leq \epsilon$ as long as ✓

$$m \gtrsim \frac{1}{\epsilon} \log \frac{|\mathcal{H}|}{\delta}$$

- Even allows $\ell(h) \gtrsim \sqrt{m}$
- $\frac{1}{\epsilon} \to \frac{1}{\epsilon^2}$ buys us **uniform convergence**:

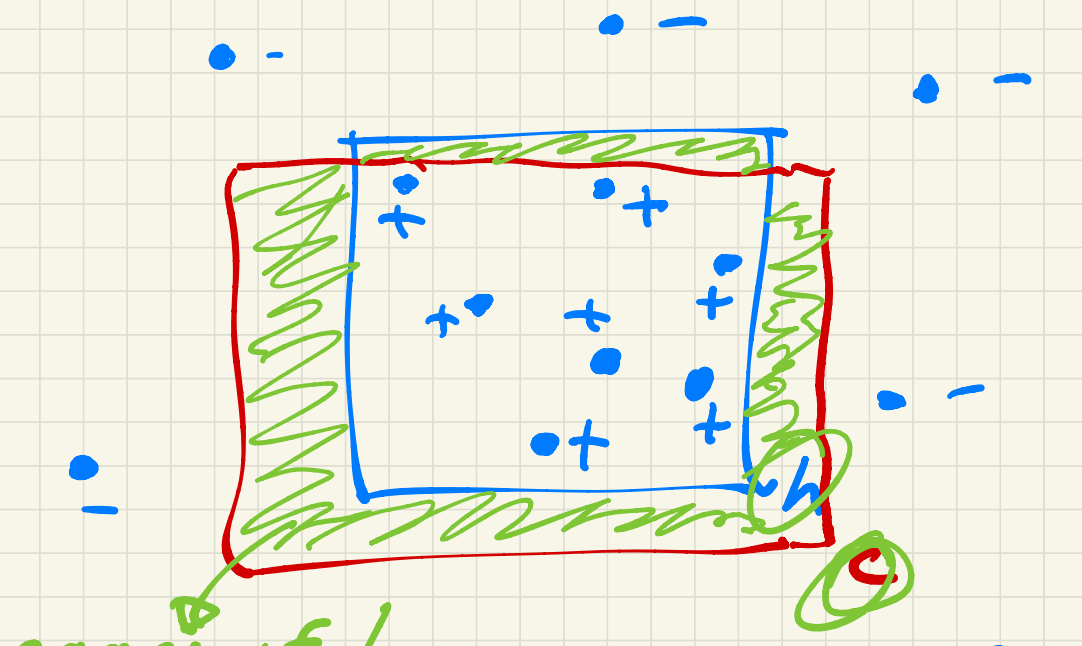
$$|\hat{\epsilon}_S(h) - \epsilon(h)| \leq \epsilon$$

$$\forall h \in \mathcal{H}$$

What happens when $|\mathcal{H}|$ is infinite?

What quantity should replace $\log|\mathcal{H}|$?

What's wrong with this:



If $\varepsilon(h) \geq \varepsilon$, then

prob we missed it

$\leq (1-\varepsilon)^m \leq \delta$

- Let $S = \{x_1, x_2, \ldots x_m\} \subseteq X$ be an ordered set (note: no labels)

Key definition #1:

$$\Pi_{\mathcal{H}}(S) \triangleq \{(h(x_1), h(x_2), \ldots h(x_m)) : h \in \mathcal{H}\}$$

$$\in \{0,1\}^m$$

Set of all labelings of S induced by $\mathcal{H}$.

Key definition #2:

Say that S is shattered by $\mathcal{H}$ if $\Pi_{\mathcal{H}}(S) = \{0,1\}^m$

vs. c_+

Key definition #3:

The VC dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Denote by $VC(\mathcal{H})$.

Quantification

To show $VC(\mathcal{H}) \geq d$,
we must find **some**
shattered S of size d.

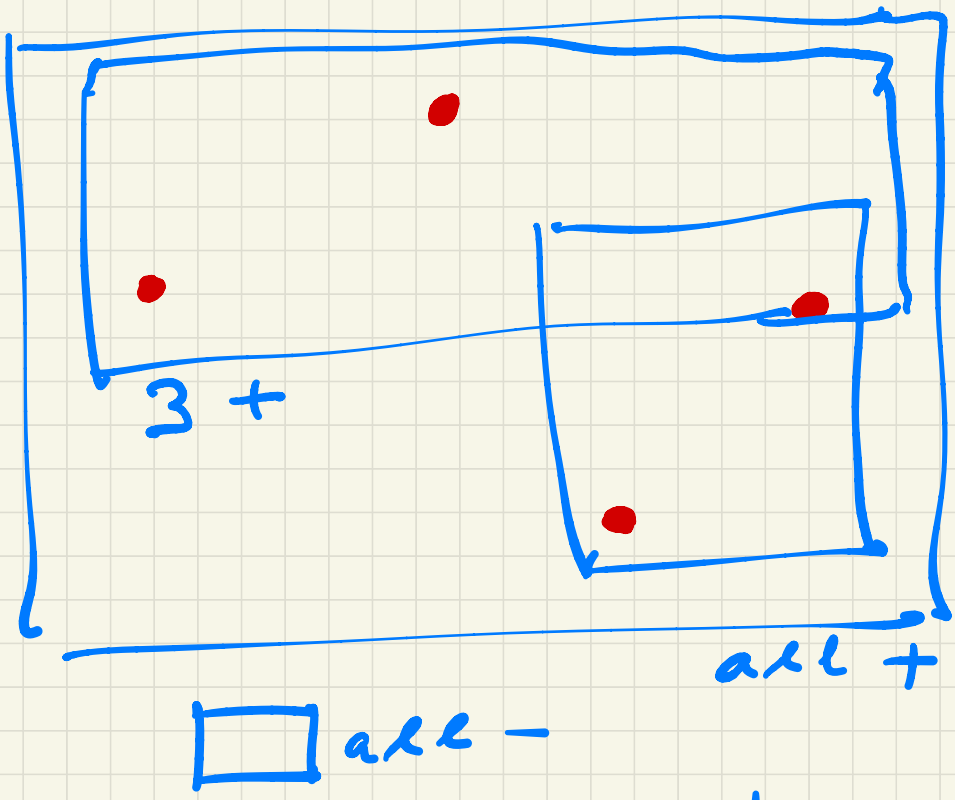
To show $VC(\mathcal{H}) \leq d$,
we must show that
no set of size $d+1$
is shattered.

Roadmap

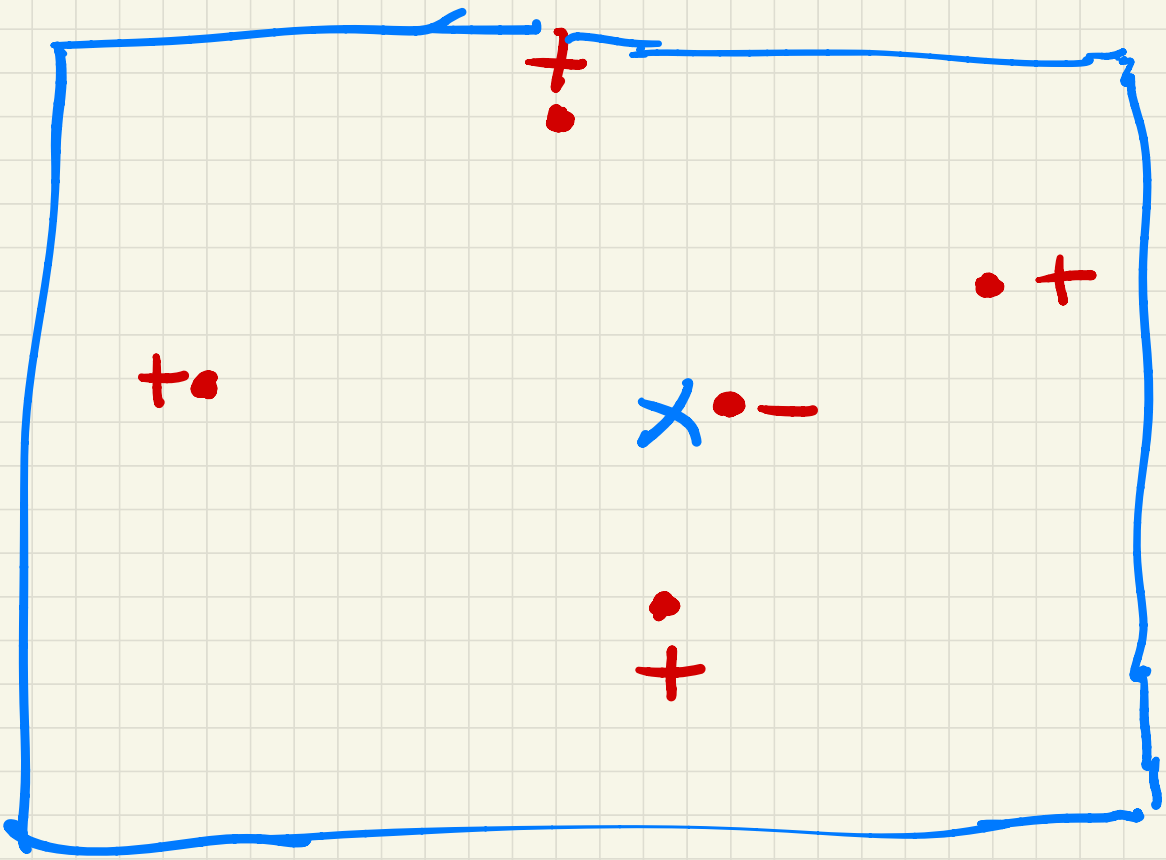
- We will see that $VC(\mathcal{H})$ replaces $\log|\mathcal{H}|$ in proving consistency $\Rightarrow$ PAC
- Warm-up with ex's of computing $VC(\mathcal{H})$
- Main result has both combinatorial and probabilistic components

Ex: Rectangles in $\mathbb{R}^2$

Can shatter these 4 points:



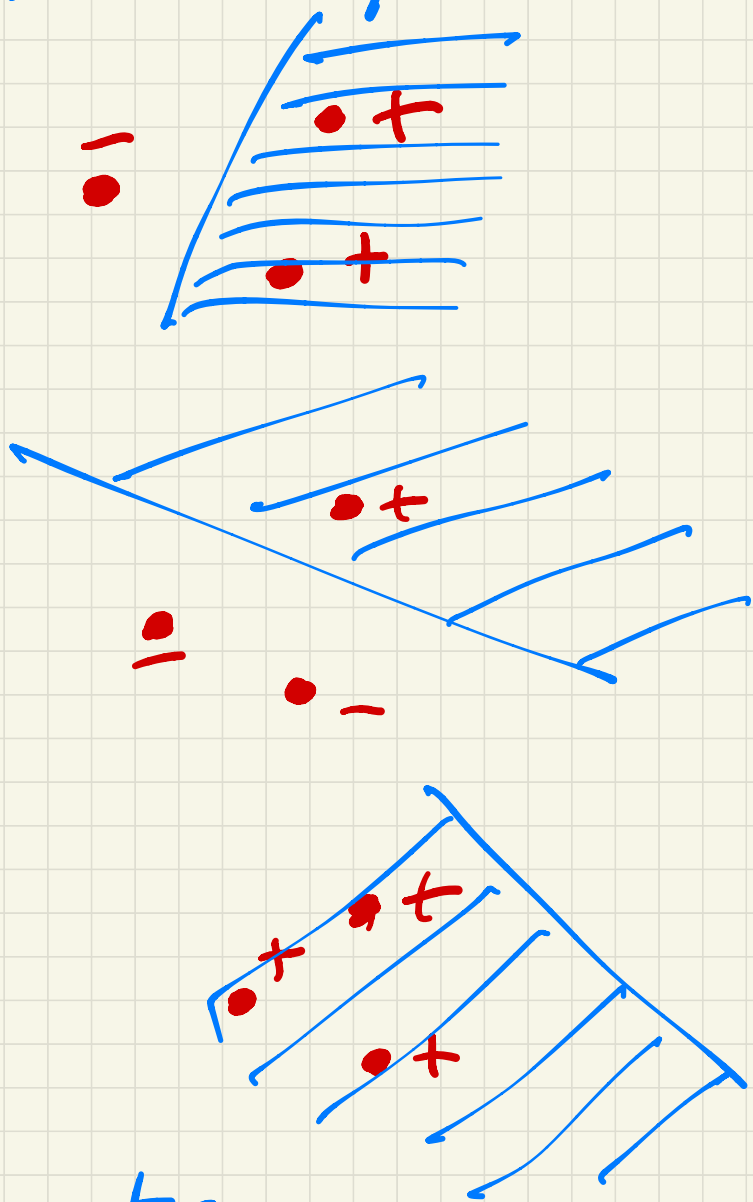
Can shatter no 5 points:



$\therefore$ VC dim = 4

Ex: Halfspaces in $\mathbb{R}^2$

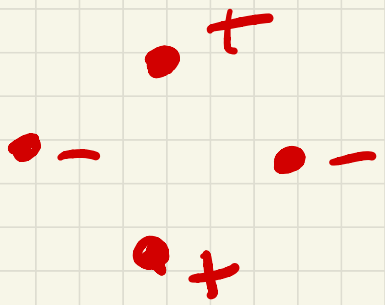
Can shatter 3 points.



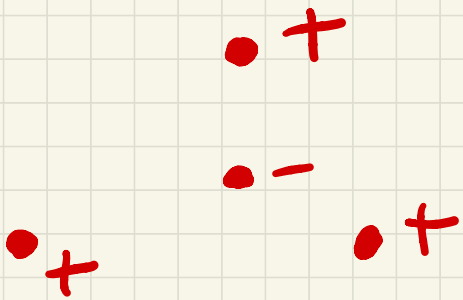
etc.

Can shatter no 4 points:

E.g.



and

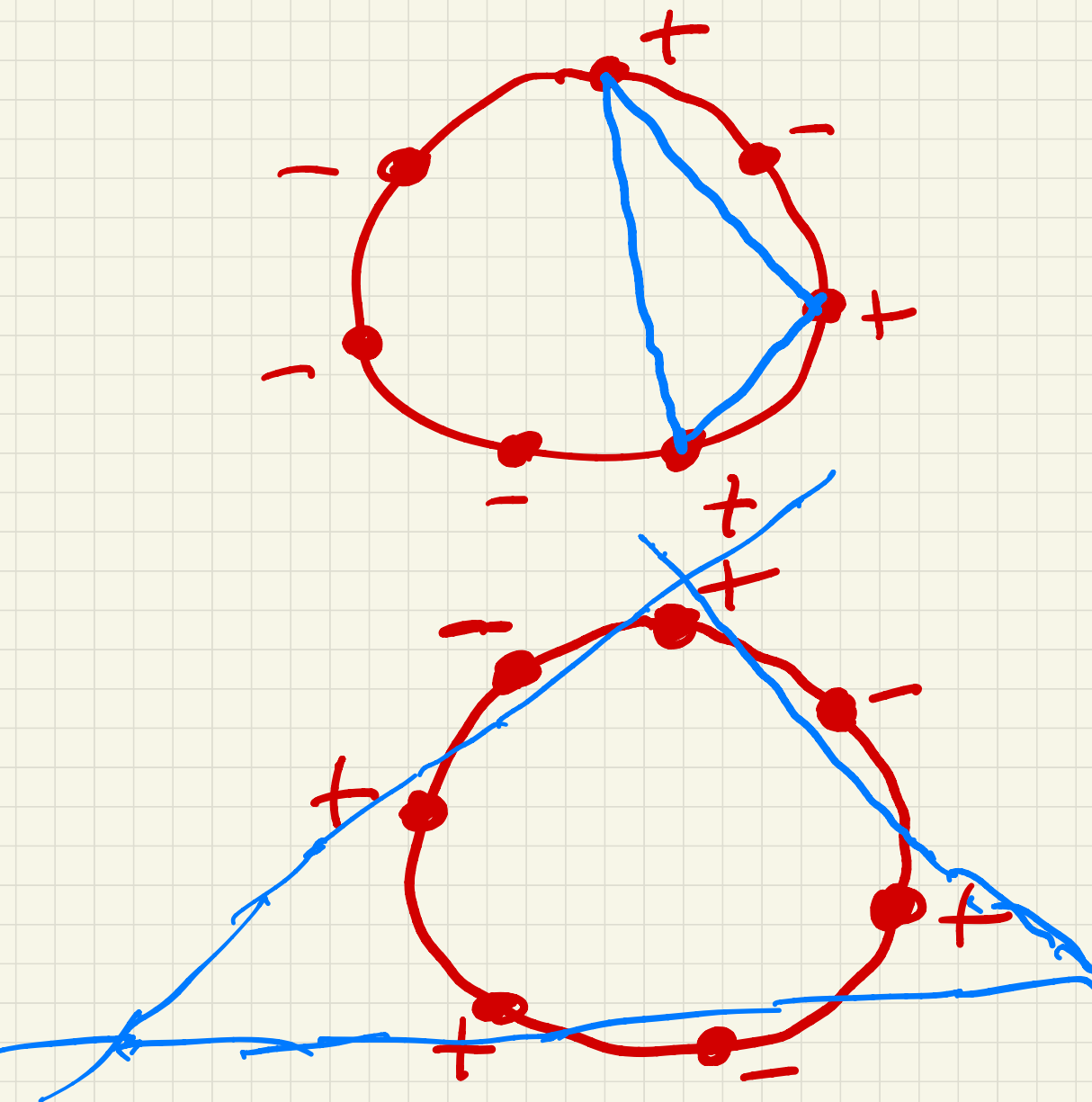


VC dim = 3

In $\mathbb{R}^d$: VC dim = $d+1$

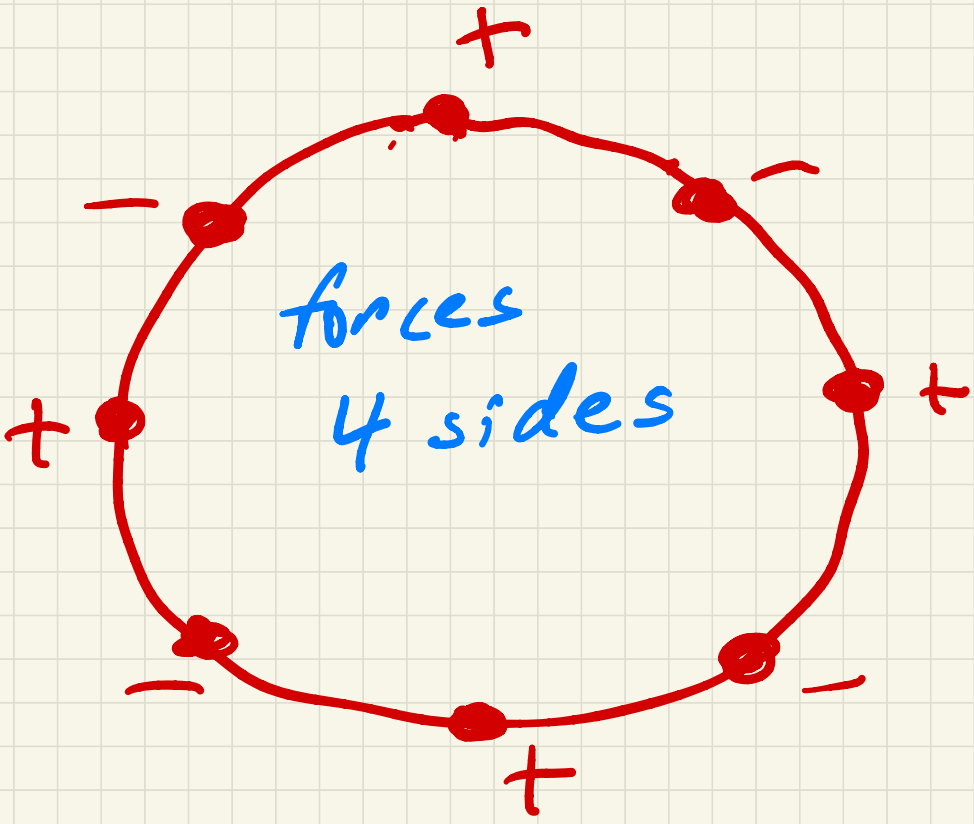
Ex: d-gons in $\mathbb{R}^2$

Can shatter $2d+1$ pts:



Can shatter no $2d+1$ pts:

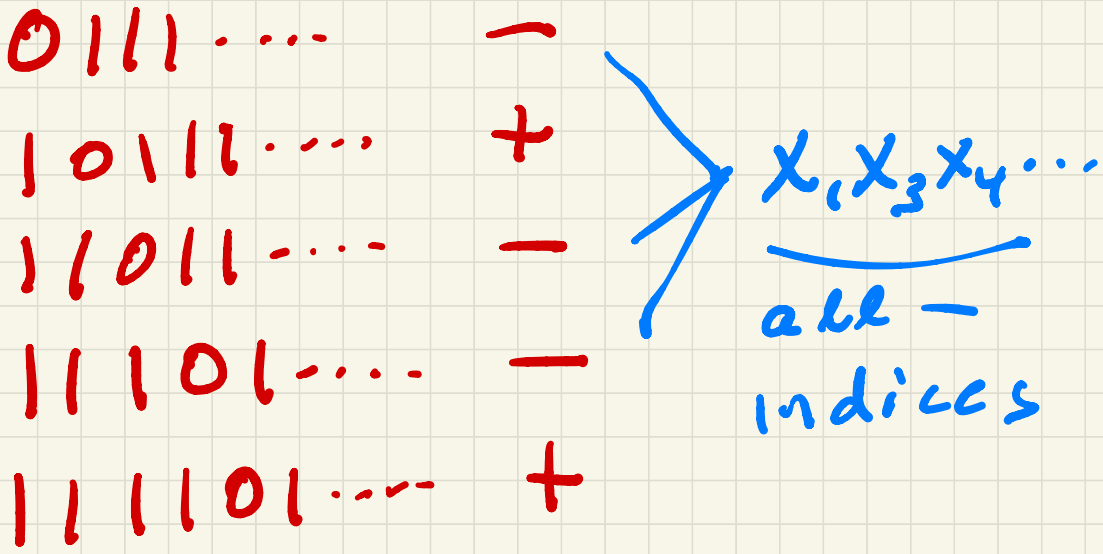
Worst case:



VC dim $= 2d+1$

Ex: Conjunctions over $\{0,1\}^n$

Can shatter these n pts:



Note: For any finite $\mathcal{H}$,

$$\text{VC dim} \leq \log_2 |\mathcal{H}|$$

since shattering d pts requires 2^d functions.

So VC dim conjunctions

$$\leq \log 3^n$$

$$\therefore \text{VC dim} = \Theta(n)$$

Recall

$\Pi_{\mathcal{H}}(S)$ = set of labelings of S induced by $\mathcal{H}$

Let's define

$$\Pi_{\mathcal{H}}(m) = \max_{S:|S|=m} \{|\Pi_{\mathcal{H}}(S)|\}$$

"growth function"

Then for $m \leq VC(\mathcal{H})$:

$$\Pi_{\mathcal{H}}(m) = 2^m \quad (\text{max})$$

What about $m > VC(\mathcal{H})$?

Amazing fact:

Let $d = VC(\mathcal{H})$. Then

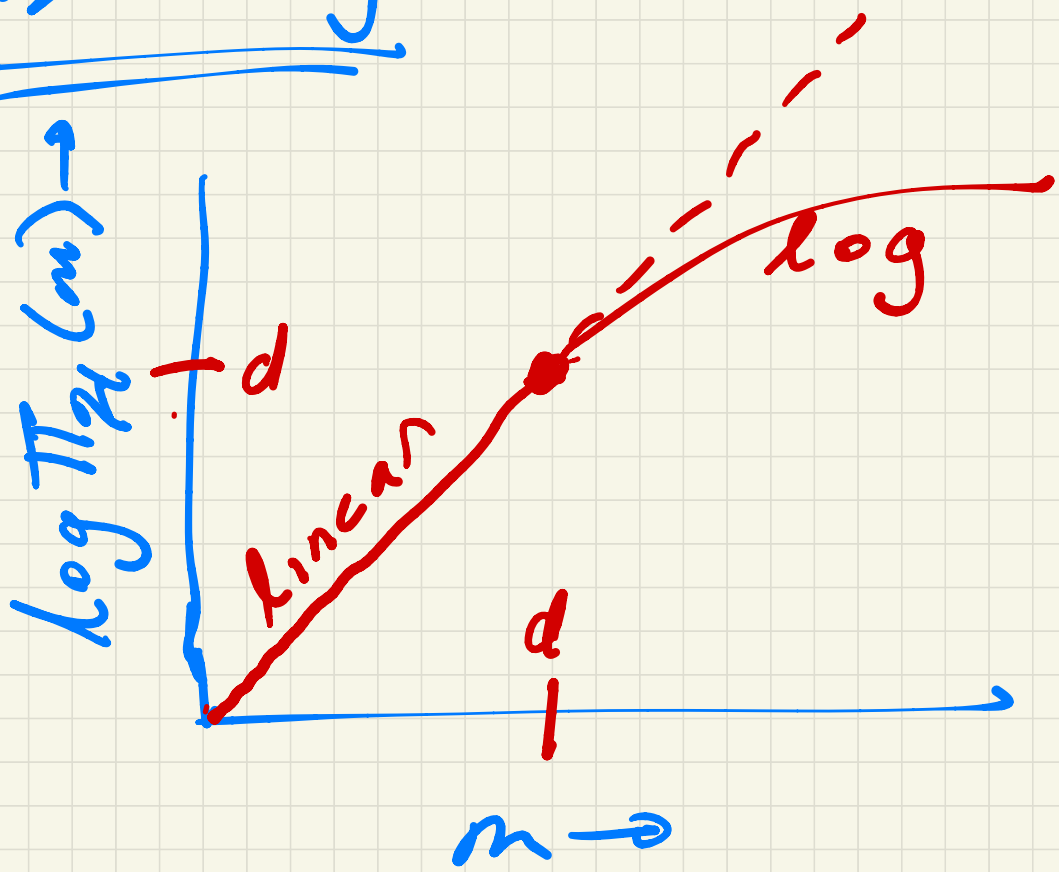
for $m > d$, $\Pi_{\mathcal{H}}(m) \sim m^d$.

$m \leq d$: exponential growth

$m > d$: polynomial growth

"Sauer's Lemma"

Visually:



Overall, $\Pi_H(m) = O(m^d)$

$(2^m \leq m^m)$

Proof outline

1. Define a function $\phi_d(m)$, show that $\pi_H(m) \leq \phi_d(m)$

2. Show that $\phi_d(m) = O(m^d)$

Define:

$$\Phi_d(m) = \Phi_d(m-1) + \Phi_{d-1}(m-1),$$

$$\Phi_0(m) = \Phi_d(0) = 1.$$

Lemma: If $VC(\mathcal{H}) = d$, then $\forall m$

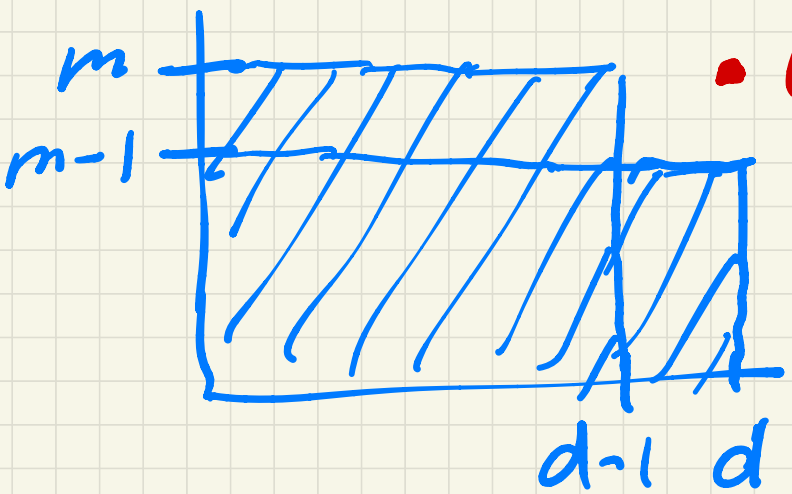
$$\Pi_{\mathcal{H}}(m) \leq \Phi_d(m)$$

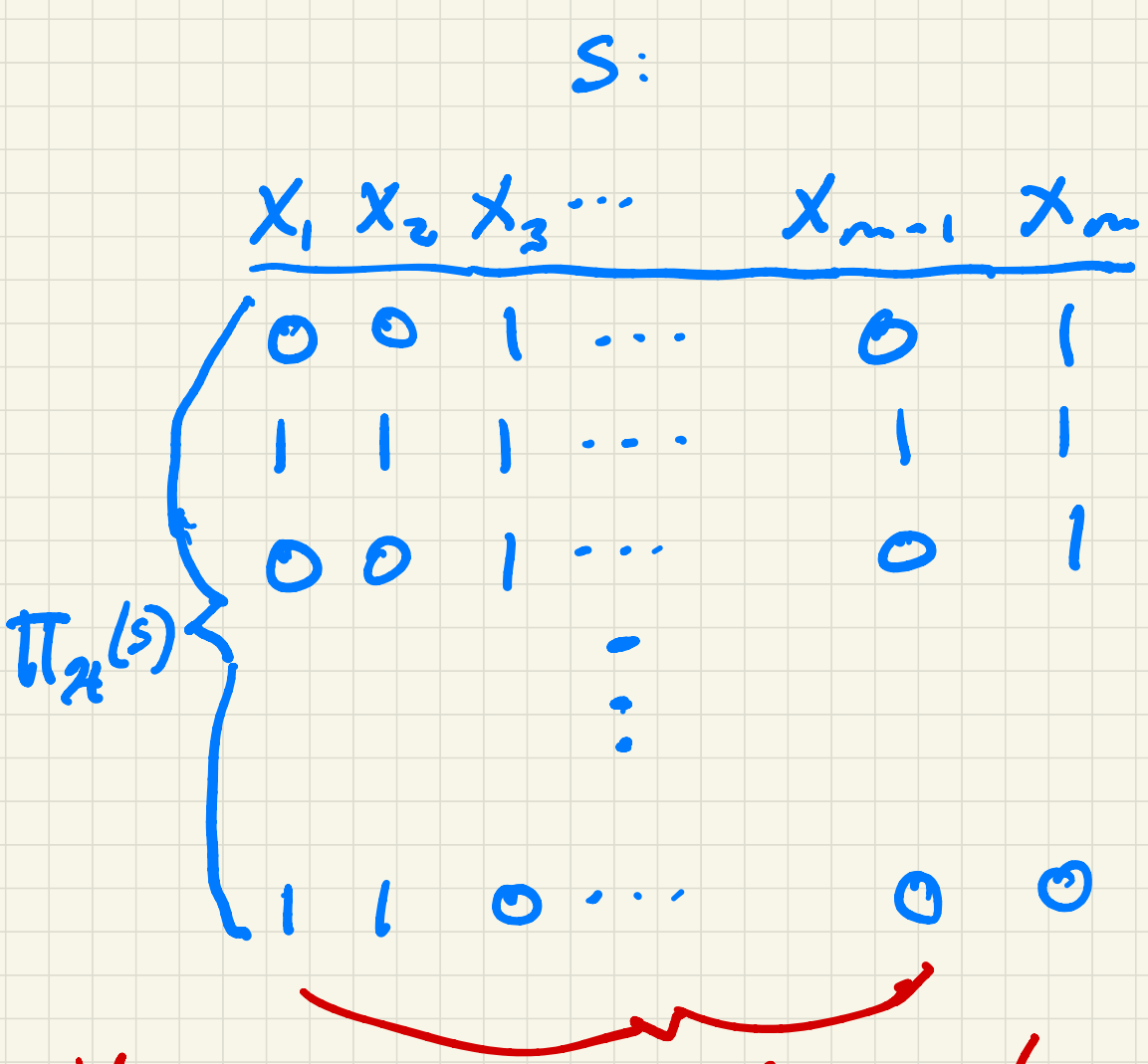
Let's prove by double induction on d & m.

Assume true for any

$d' \leq d$

$m' \leq m$ (one must be strict)





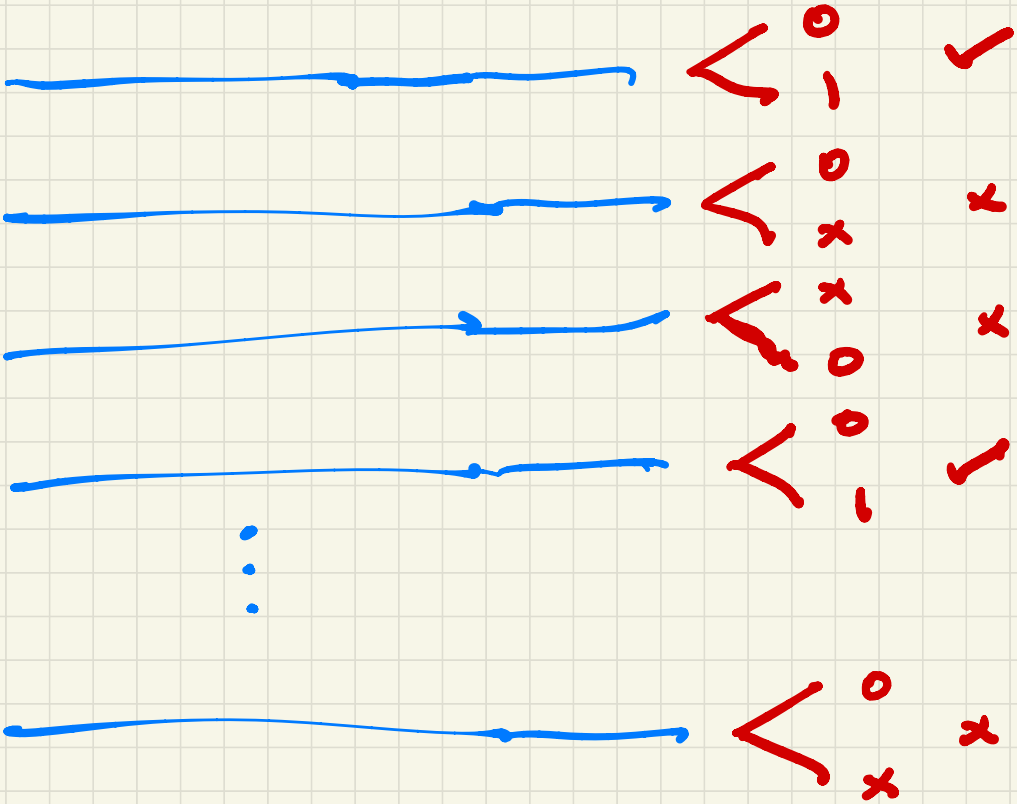
How many labelings here?

$\leq \Pi_{\mathcal{H}}(m-1) \leq \Phi_d(m-1)$

(induction)

labelings on $x_1, \ldots x_{m-1}$:

extension(s) on x_m:



Need to add count of ✓s.

But labelings of

$X' = x_1, x_2, \ldots x_{m-1}$

with a ✓ is just a fn. class (over X')

of VC dim $\leq d-1$.

Why?

$\therefore \#\checkmark s \leq \Phi_{d-1}(m-1)$

(induction again)

$\# \Pi_{\mathcal{H}}(S) \leq \Phi_d(m-1)$

$+ \Phi_{d-1}(m-1)$

$= \Phi_d(m)$

as desired.

Lemma $\phi_d(n) = \sum_{i=0}^{d} \binom{n}{i}$.

Pf $\phi_d(n) =$

$\phi_{d-1}(n) + \phi_{d-1}(n-1) =$

$\sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i}$

(induction)

hi $\to \binom{n-1}{d-1}$

$= \sum_{i=0}^{d} \binom{n-1}{i-1}$ Lo $\to \binom{n-1}{-1} \cong 0$

$= \sum_{i=0}^{d} \left[\binom{n-1}{i} + \binom{n-1}{i-1} \right] = \binom{n}{i}$

OK, that's it for combinatorial analysis of $\pi_{\mathcal{H}}(m)$.

Now for the probabilistic part.

- Let's fix all of these:
  target $c \in \mathcal{H}$, P, ε, δ

- Define error regions of $\mathcal{H}$ wrt c:

$$\mathcal{H} \Delta c \triangleq \{ \underbrace{h \Delta c}_{\text{viewed as sets of + ex's}} : h \in \mathcal{H} \}$$

As functions:

$$(h \Delta c)(x) = \begin{cases} 1 & \text{if } h(x) \neq c(x) \\ 0 & \text{else} \end{cases}$$

$$= h(x) \oplus c(x)$$

Note: $VC(\mathcal{H} \Delta c) = VC(\mathcal{H})$:

- If $h_1(x) \neq h_2(x)$,

  $h_1(x) \oplus c(x) \neq h_2(x) \oplus c(x)$

- labelings $h_1(s) \neq h_2(s)$

$\Rightarrow (h_1 \oplus c)(s) \neq (h_2 \oplus c)(s)$

$\therefore \forall s \; |\pi_{\mathcal{H}}(s)| = |\pi_{\mathcal{H} \Delta c}(s)|$

$\Rightarrow VC(\mathcal{H}) = VC(\mathcal{H} \Delta c)$

OK. Suppose some $h \in H$ has $\varepsilon(h) > \varepsilon$. We want to make sure S hits $h_{ac} \in H_{ac}$ at least once.

And we want this to happen for every h s.t. $\varepsilon(h) > \varepsilon$.

Call such S an ε-net and define

$$A(S) = \begin{cases} 1 & \text{if } S \text{ not an } \varepsilon\text{-net} \\ 0 & \text{else} \end{cases}$$

Goal: bound $P_S[A(S) = 1]$

Problem: Hard to tell if $A(S)=1$ by looking at S.

Solution: "two-sample trick"

$B(S,S')=1$ iff $\exists r \in H \Delta C$ s.t. $P[r] > \varepsilon$ and:

(S, S': $2m$ pts)

① $r \cap S = \phi$ $(\Rightarrow A(S)=1)$

② $|r \cap S'| \geq \frac{\varepsilon m}{2}$

Otherwise $B(S,S')=0$.

Claim: $\Pr[B] \geq \Pr[A] \cdot 1/2$

$B(S,S')=1 \Rightarrow A(S)=1$, so just need to lower bound

$\Pr_{S,S'}[B(S,S')=1 \mid A(S)=1]$

can fix missed r s.t. $P[r] > \varepsilon$

$\geq \Pr_{S'}[|r \cap S'| \geq \varepsilon m/2]$

$\geq 1/2$ (e.g. Chebyshev)

So $Pr[A] \leq 2 Pr[B]$

upper bound

Draw S, S' in 2 steps:

- draw $2m$ pts $T = S \cup S'$

At most $\Phi_d(2m)$ labelings!

- split $T \to S, S'$ randomly (random permutation)

Same as i.i.d. S, S' by

exchangability

$B(S,S')=1$ only if $\exists r \in H_{ac}$ s.t. r is hit $\ell \geq \varepsilon m/2$ times in T but all ℓ hits fall in S'

Prob. of above happening just wrt random permutation

$$\leq \frac{\binom{m}{\ell}}{\binom{2m}{\ell}}$$

$$\frac{\binom{n}{\ell}}{\binom{2n}{\ell}} = \frac{\frac{n!\,\ell!}{(n-\ell)!}}{\frac{(2n)!\,\ell!}{(2n-\ell)!}}$$

$$= \frac{n!}{(n-\ell)!}\,\frac{(2n-\ell)!}{(2n)!}$$

$$= \frac{n(n-1)(n-2)\cdots(n-\ell+1)}{(2n)(2n-1)\cdots(2n-\ell+1)}$$

$$= \frac{n}{2n}\cdot\frac{n-1}{2n-1}\cdot\frac{n-2}{2n-2}\cdots\frac{n-\ell+1}{2n-\ell+1}$$

$$\leq 2^{-\ell} \leq 2^{-\varepsilon n/2}$$

Wrapping up:

$$Pr_{S,S'}[B(S,S')=1] \leq \Phi_d(2m) \cdot 2^{-\varepsilon m/2}$$

$\Phi_d(2m)$: #labelings of T

$2^{-\varepsilon m/2}$: Pr[B] w.r.t permutation

$$\leq C_0 (2m)^d 2^{-\varepsilon m/2}$$

C_0 (Sauer)

$$= e^{C_1[d \log m - \varepsilon m]}$$

$Pr[A] \leq 2^x$, set $\leq \delta$ & solve...

Theorem Let $d = VC(\mathcal{H})$.

Then for

$$m \geq c_2\left(\frac{1}{\varepsilon}\log\frac{1}{\delta} + \frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\right)$$

with prob. $\geq 1-\delta$

any consistent $h \in \mathcal{H}$

has $\Sigma(h) \leq \varepsilon$.

More generally,

if $\frac{1}{\varepsilon} \to \frac{1}{\varepsilon^2}$,

have uniform convergence:

With prob. $\geq 1-\delta$, $\forall h \in \mathcal{H}$:

$$\left| \hat{\varepsilon}_S(h) - \varepsilon(h) \right| \leq \varepsilon.$$

(slightly different $B(S,S')$ & prob. analysis)

Lower Bound #1

- Let $d = VC(\mathcal{H})$
- Let $x_1, \ldots x_d$ be shattered
- $\forall \bar{v} \in \{0,1\}^d$ let P uniform

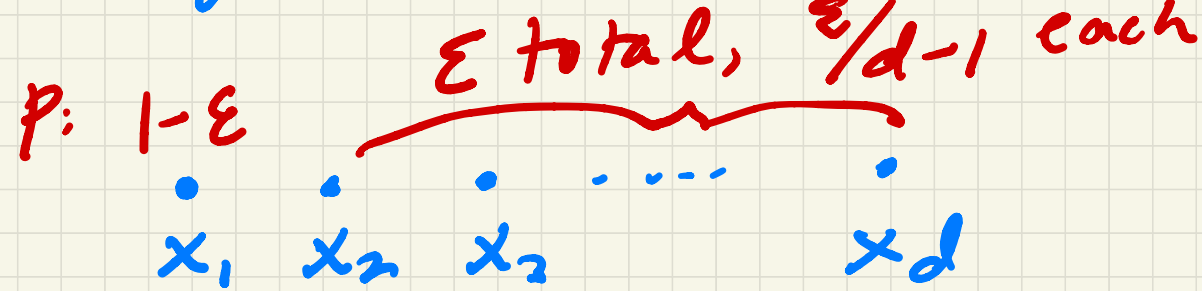
  $h_{\bar{v}}(x_i) = v_i \quad \forall 1 \le i \le d$
- Now choose target $c = h_{\bar{v}}$ for $\bar{v}$ random
- For $\varepsilon \le 1/4$, need $\Omega(d)$ examples - are just predicting coin flips!

Better lower bound?

E.g. $\Omega(d/\varepsilon)$ for any ε?

Lower Bound #2

- Same set-up as #1, but now we'll "give away" $c(x_1)$



- Still choose $c =$ random $h_{\bar{v}}$
- Now only see coin flip every $\sim 1/\varepsilon$ samples $\Rightarrow$ $\Omega(d/\varepsilon)$.

A Prescriptive Application of VC: Structural Risk Minimization

- Suppose we have not a single $\mathcal{H}$ but a **nested hierarchy**:

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \cdots \subset \mathcal{H}_d \subset \cdots$$

- E.g. neural networks of increasing depth/width
- For simplicity assume

$$VC(\mathcal{H}_d) = d$$

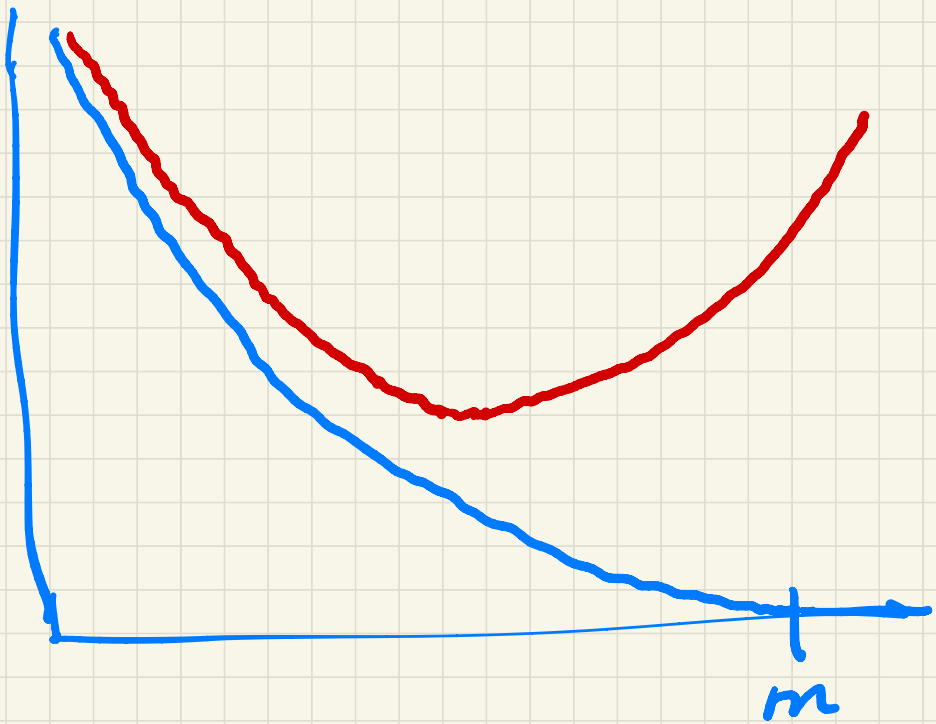
- Let $S = \{\langle x_i, y_i \rangle\}$ be of size m

VC theory says that for $\mathcal{H}_d$,

$$|\hat{\varepsilon}_S(h) - \varepsilon(h)| \lesssim \sqrt{\frac{d}{m}}$$

(ignoring log factors, fixing δ, spreading δ over the $\mathcal{H}$'s)

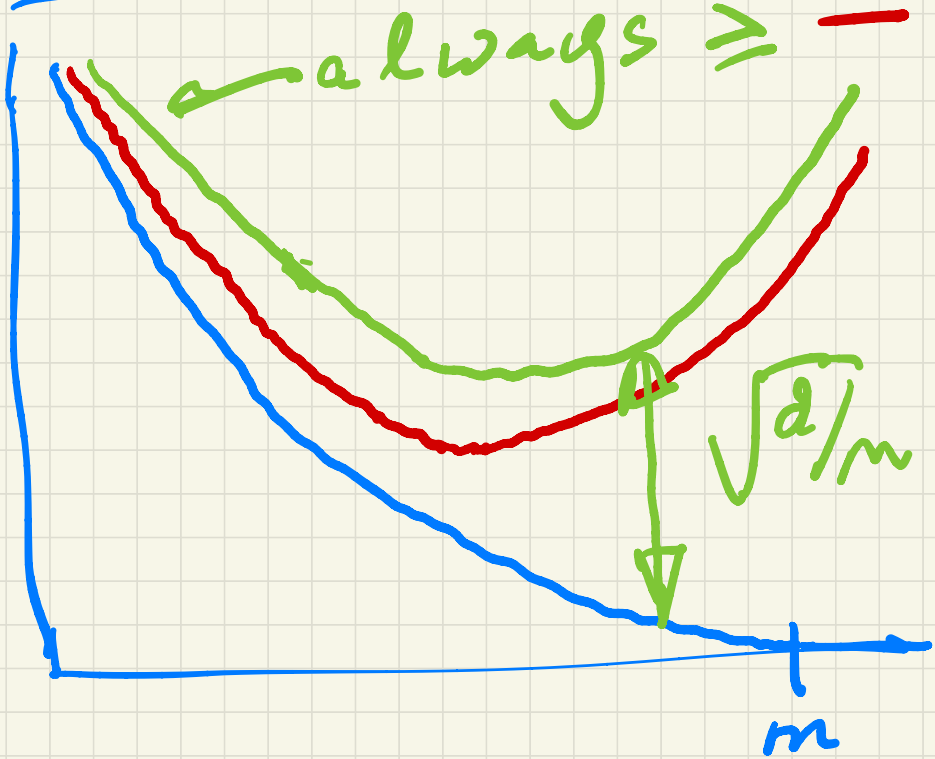
Classic overfitting cartoon:



—: optimal $\hat{\varepsilon}_S(h)$ in $\mathcal{H}_d$

—: true $\varepsilon(h)$ of $\hat{h}$

Classic overfitting cartoon:



—: optimal $\hat{\varepsilon}_S(h)$ in $\mathcal{H}_d$

—: true $\varepsilon(h)$ of $\hat{h}$

—: $\hat{\varepsilon}_S(h) + \sqrt{d/m}$

So choose:

$$d^* = \arg\min_d \left\{ \hat{\varepsilon}_S(h_d^*) + \sqrt{\frac{d}{m}} \right\}$$

$$h_d^* = \arg\min_{h \in \mathcal{H}_d} \left\{ \hat{\varepsilon}_S(h) \right\}$$

Then our true error

$$\leq \min_d \{\text{same}\}$$

Distribution-Dependent Improvements to VC Theory

- Replace $\Pi_{\mathcal{H}}(m)$ by $E_p[|\Pi_{\mathcal{H}}(s)|]$ VC Entropy
- Related: Rademacher complexity
- Replace $|\mathcal{H}|(1-\varepsilon)^m$ by $\sum_{\varepsilon_i} |\mathcal{H}_i|(1-\varepsilon_i)^m$

connections to stat mech

Data-Dependent Improvements to VC Theory

- Max-margin generalization (e.g. linear, SVMs, boosting)
- "PAC-Bayes"
- Small-norm generalization

Generalizations to other learning settings

- Models h in class $\mathcal{H}$
- Observations $z \sim P$ i.i.d.
- Loss function $\ell(h, z) \in \mathbb{R}$ also i.i.d

E.g. regression:

$z = \langle x, y \rangle$, $y \approx h(x) \in \mathbb{R}$

$\ell(h, \langle x, y \rangle) = (h(x) - y)^2$

(squared error)

or $\ell(h, \langle x, y \rangle) = |h(x) - y|$

(absolute err.)

or $\ell(h, \langle x, y \rangle) =$

$(h(x) - y)^2 + \alpha |h(x)|$

Distribution learning

- Models are distributions

$$p \in \mathcal{H}$$

- observations $z = x \sim p^*$
- loss $\ell(p, x) = \log \frac{1}{p(x)}$

log-loss $\Rightarrow$ MLE

Pretty much any model type, observations & loss function...

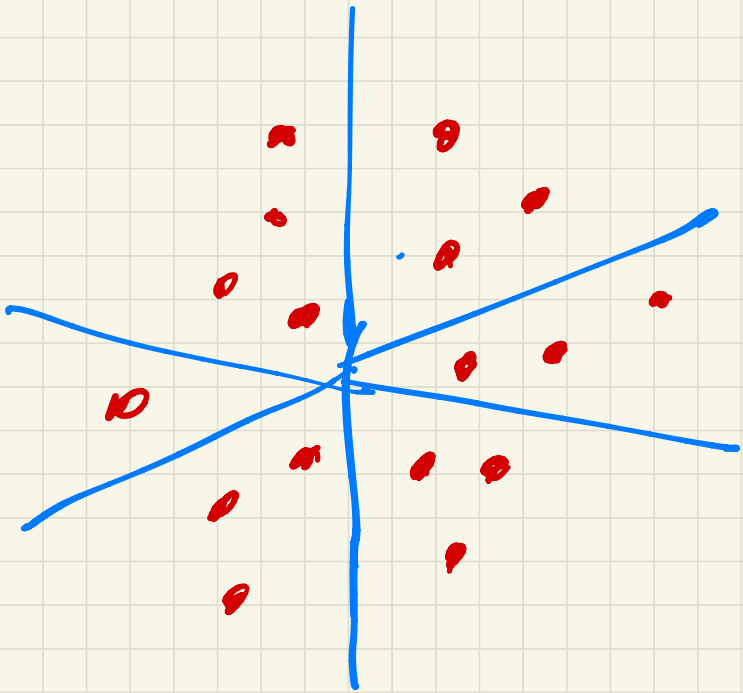
Now let $S = \{z_1, z_2, \ldots z_d\}$
and look at

$$\{\underbrace{\langle l(h,z_1), \ldots, l(h,z_d)\rangle}_{\in \mathbb{R}^d} : h \in \mathcal{H}\}$$

= generalization of
$\Pi_{\mathcal{H}}(S)$, now possibly
infinite!

Say $\mathcal{H}$ shatters S if
$\Pi_{\mathcal{H}}(S)$ is
"space filling"...

E.g. $\Pi_H(S)$ intersects all 2^d orthants of $\mathbb{R}^d$:



"fractal dimension"
"combinatorial dimension"

Moral: Uniform convergence is the norm, not the exception.