Consistency, Uniform Convergence, and Learning: The VC Dimension

Finite $\mathcal{H}$: Quick Review

- Any consistent h has $\varepsilon(h) \le \varepsilon$ as long as

$$\boxed{m \sim \frac{1}{\varepsilon} \log \frac{|\mathcal{H}|}{m}}$$ ✓

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

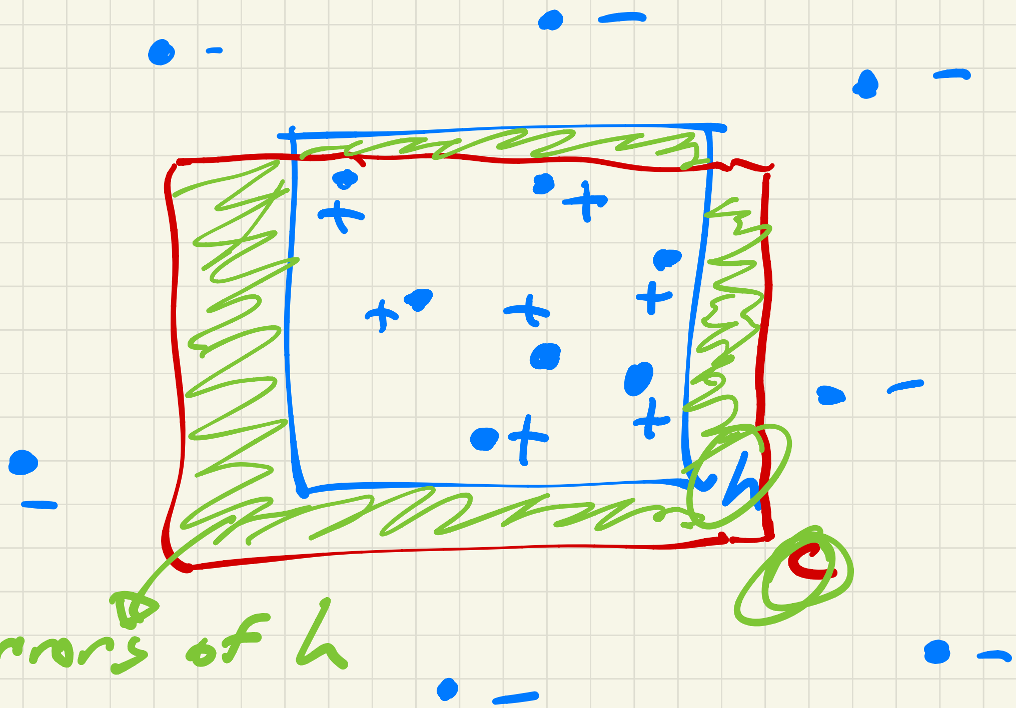
$$|\hat{\varepsilon}_S(h) - \varepsilon(h)| \le \varepsilon$$

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

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

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

What's wrong with this:



If $\mathcal{E}(h) \geq \varepsilon$, 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:

$$\Pi_{\mathcal{H}}(S) \triangleq \{\underbrace{\langle h(x_1), h(x_2), \ldots h(x_m)\rangle}_{\in \{0,1\}^m} : h \in \mathcal{H}\}$$

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. $\subseteq$

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

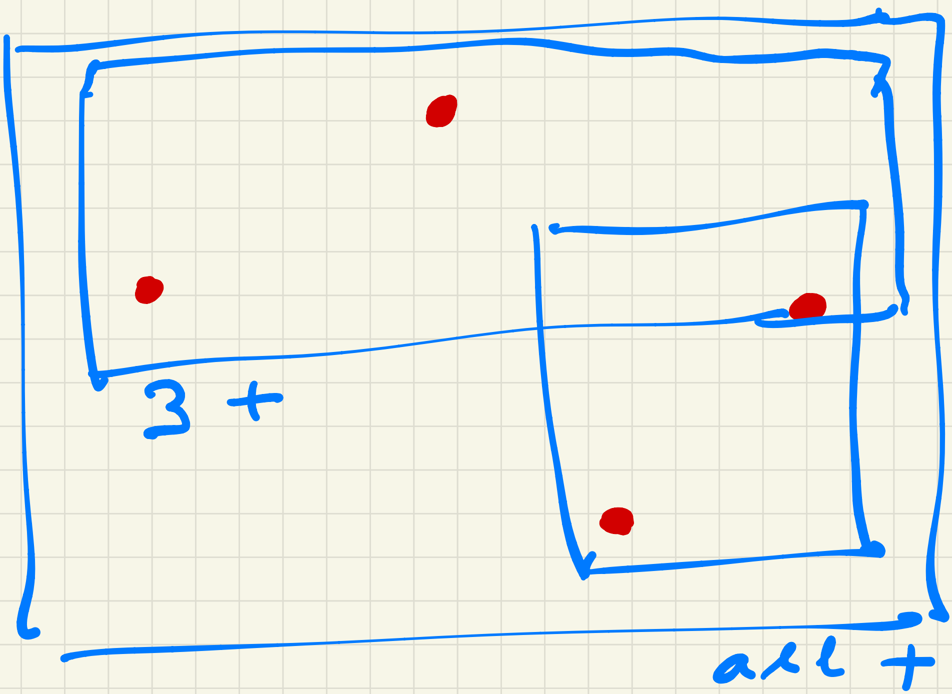
To show $VC(\mathcal{H}) \leq d$,
we must show that
$\forall$ no set of size $d+1$
is shattered.

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 $\mathbb{R}^2$

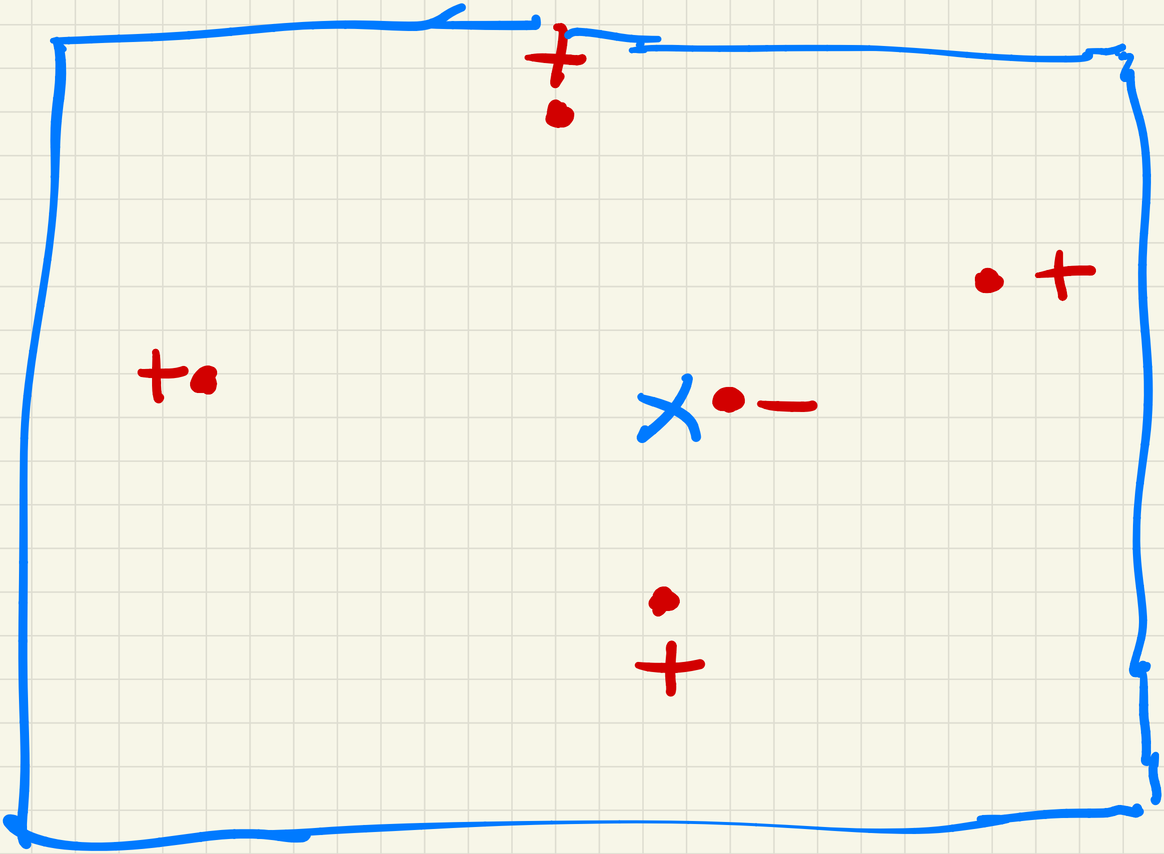
Can shatter these 4 points:



$\square$ all $-$

etc.

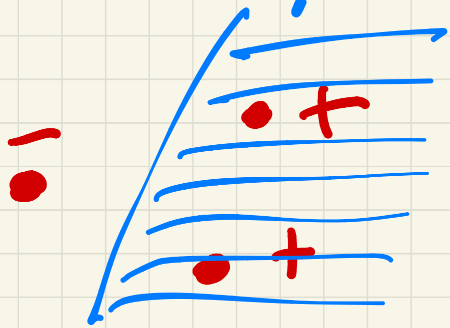
Can shatter no 5 points:

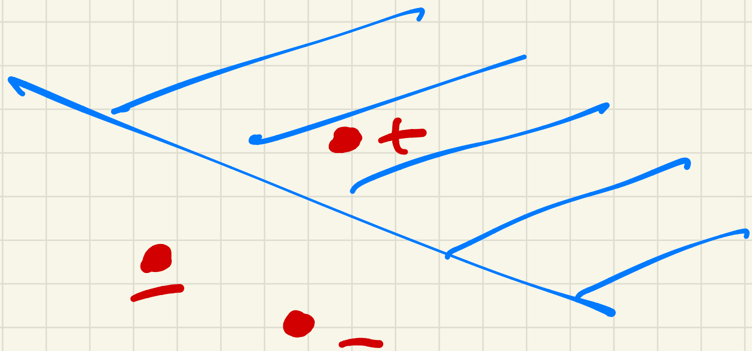


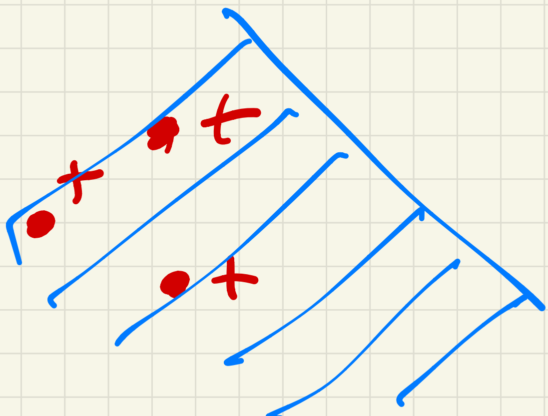
$\therefore$ VC dim = 4

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

Can shatter 3 points.



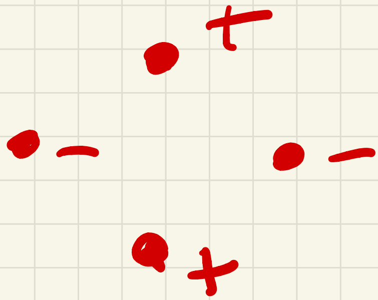




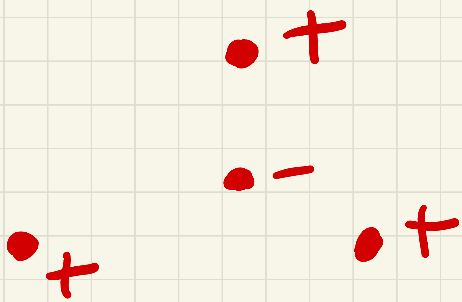
etc.

Can shatter no 4 points:

E.g.



and

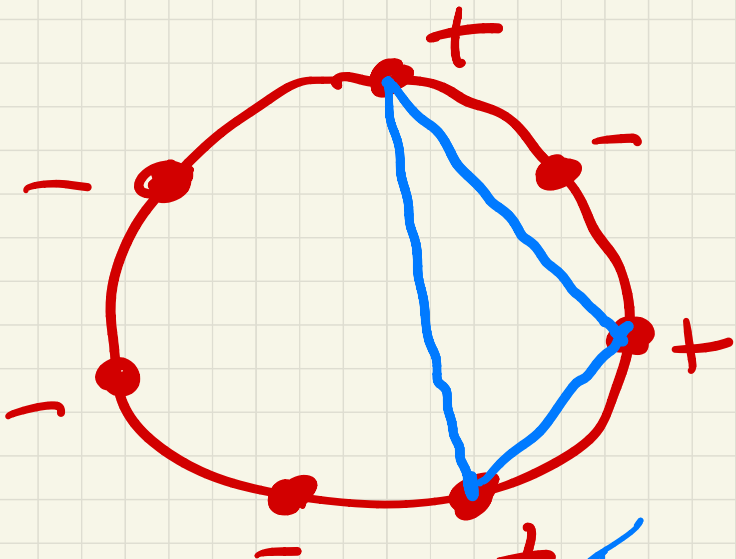


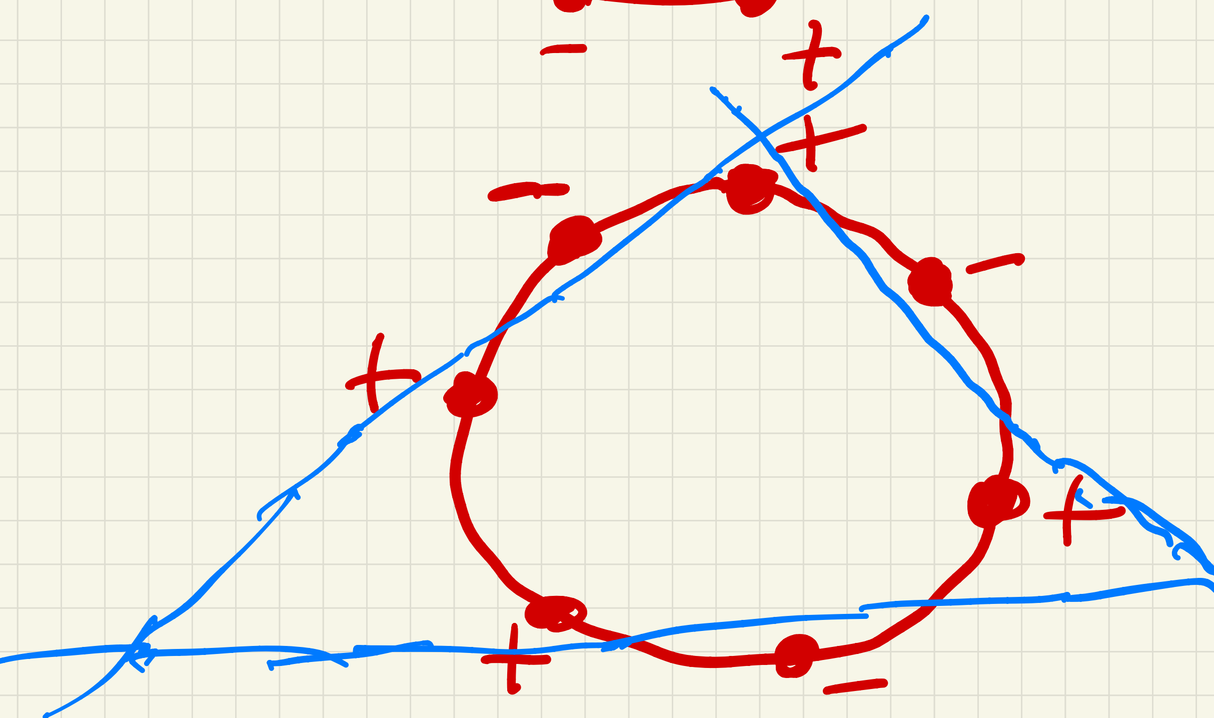
VC dim = 3

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

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

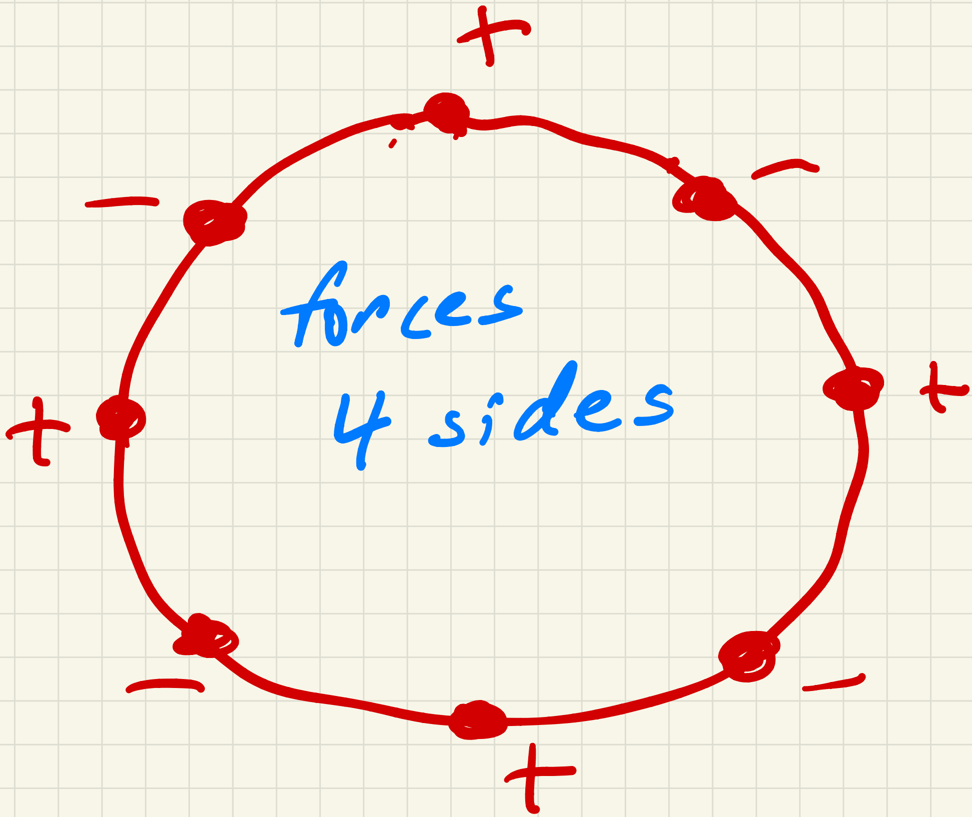
Can shatter $2d+1$ pts:





Can shatter no $2d+1$ pts:

Worst case:



VC dim $= 2d+1$

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

Can shatter these n pts:

$$\left.\begin{array}{ll} 0111\cdots & - \\ 10111\cdots & + \\ 11011\cdots & - \\ 11101\cdots & - \end{array}\right\} \frac{x_1 x_3 x_4 \cdots}{\text{all } - \text{ indices}}$$

$$\begin{array}{ll} 111101\cdots & + \end{array}$$

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

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

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

Amazing fact:

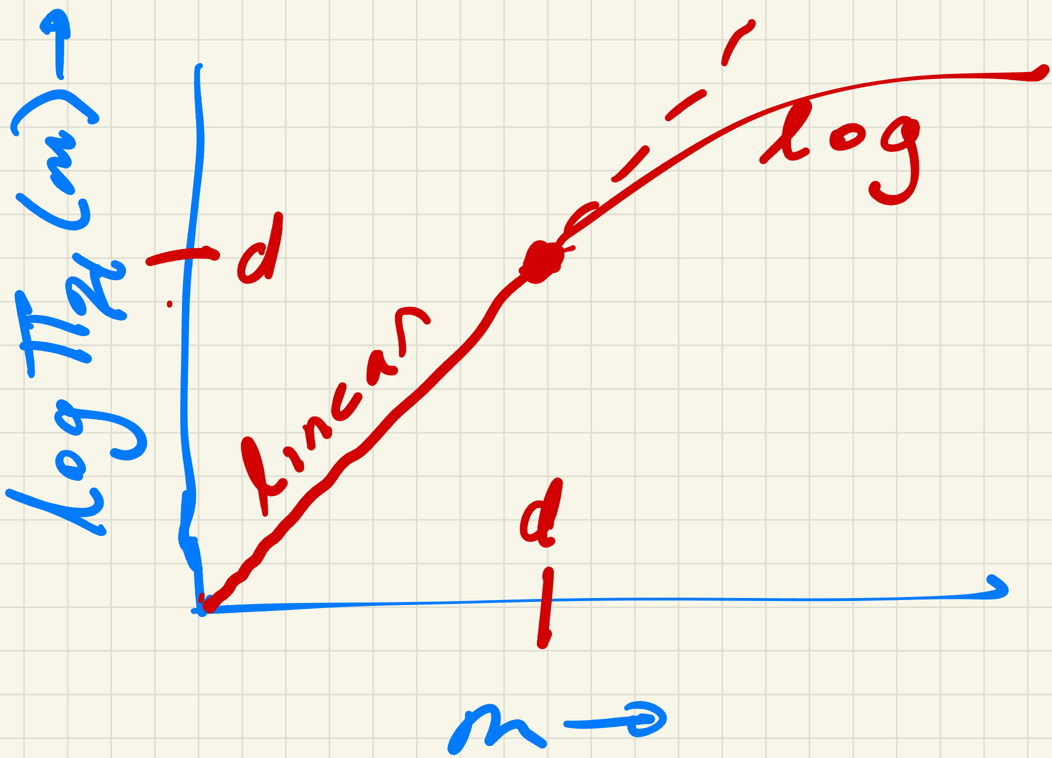
Let $d = VC(\mathcal{H})$. Then for $m > d$, $\Pi_{\mathcal{H}}(m) \sim m^d$.

$m \leq d$: exponential growth

$m > d$: polynomial growth

"Sauer's Lemma"

Visually:



Overall, $\Pi_{\mathcal{H}}(m) = O(m^d)$

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

Proof outline

1. Define a function $\Phi_d(m)$, show that $\Pi_{\mathcal{H}}(m) \le \Phi_d(m)$

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

Define:

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

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

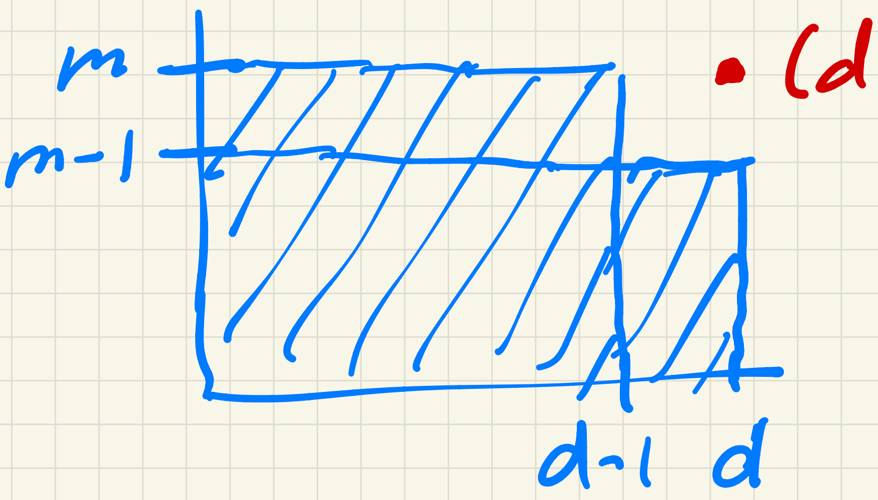
Lemma: If $VC(\mathcal{H}) = d$, then $\forall m$

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

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

Assume true for any

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



S:

$$\Pi_{\mathcal{H}}(S)\left\{\begin{array}{cccccc}
x_1 & x_2 & x_3 & \cdots & x_{m-1} & x_m \\
\hline
0 & 0 & 1 & \cdots & 0 & 1 \\
1 & 1 & 1 & \cdots & 1 & 1 \\
0 & 0 & 1 & \cdots & 0 & 1 \\
 & & & \vdots & & \\
1 & 1 & 0 & \cdots & 0 & 0
\end{array}\right.$$

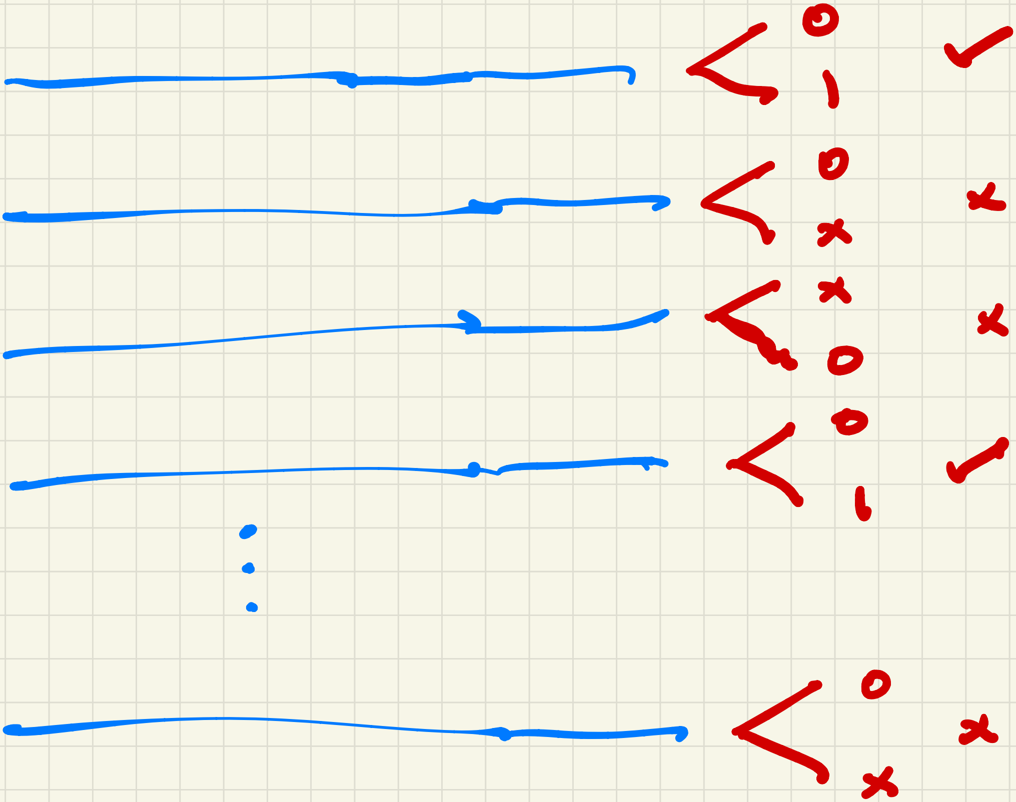
How many labelings here?

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

(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 $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.

Pf $\Phi_d(m) =$

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

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

(induction)

$$= \sum_{i=0}^{d} \binom{m-1}{i-1} \quad \begin{matrix} \xrightarrow{\text{hi}} \binom{m-1}{d-1} \\ \xrightarrow{\text{lo}} \binom{m-1}{-1} \triangleq 0 \end{matrix}$$

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

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:

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

As <u>functions</u>:

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

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

<u>Note</u>: $VC(\mathcal{H} \Delta c) = VC(\mathcal{H})$:

- 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 $h \in \mathcal{H}$ has $\mathcal{E}(h) > \varepsilon$. We want to make sure S hits $h \triangle c \in \mathcal{H} \triangle c$ at least once.

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

Call such S an ε-net and define

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

<u>Goal</u>: bound $\Pr_S[A(S) = 1]$

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

Solution: "two-sample trick"

$$\underbrace{B(S,S')}_{2m \text{ pts}} = 1 \text{ iff } \exists r \in \mathcal{H} \Delta C \text{ s.t. } P[r] > \varepsilon$$

and:

1. $r \cap S = \emptyset \quad (\Rightarrow A(S)=1)$

2. $|r \cap S'| \geq \varepsilon m / 2$

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

$$\Pr_{S,S'}[B(S,S')=1 \mid \underbrace{A(S)=1}_{\substack{\text{can fix missed } r \\ \text{s.t. } P[r] > \varepsilon}}]$$

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

$$\geq 1/2 \quad \text{(e.g. Chebyshev)}$$

So $\Pr[A] \leq 2 \underbrace{\Pr[B]}_{\text{upper bound}}$

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 $\exists r \in \mathcal{H}\Delta c$ s.t. r is hit $l \geq \varepsilon m/2$ times in T but **all l hits** fall in S'

Prob. of above happening

just wrt random permutation

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

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

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

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

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

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

Wrapping up:

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

$$\underbrace{\Phi_d(2m)}_{\text{\# labelings of } T} \cdot \underbrace{2^{-\varepsilon m/2}}_{\Pr[B] \text{ wrt permutation}}$$

$$\le \underbrace{C_0(2m)^d}_{\text{(Sauer)}} 2^{-\varepsilon m/2}$$

$$= e^{c_1[d\log m - \varepsilon m]}$$

$\Pr[A] \le 2\times$, set $\le \delta$ & 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 $\varepsilon(h) \leq \varepsilon$.

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 $\underbrace{x_1, \ldots x_d}_{P \text{ uniform}}$ be shattered
- $\forall\, \bar{v} \in \{0,1\}^d$ let
  $h_{\bar{v}}(x_i) = v_i \quad \forall\, 1 \le i \le d$
- 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" $c(x_1)$

$$P: \; \underset{x_1}{1-\varepsilon} \quad \overbrace{x_2 \quad x_2 \quad \cdots \quad x_d}^{\varepsilon \text{ total}, \; \varepsilon/d-1 \text{ each}}$$

- 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

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

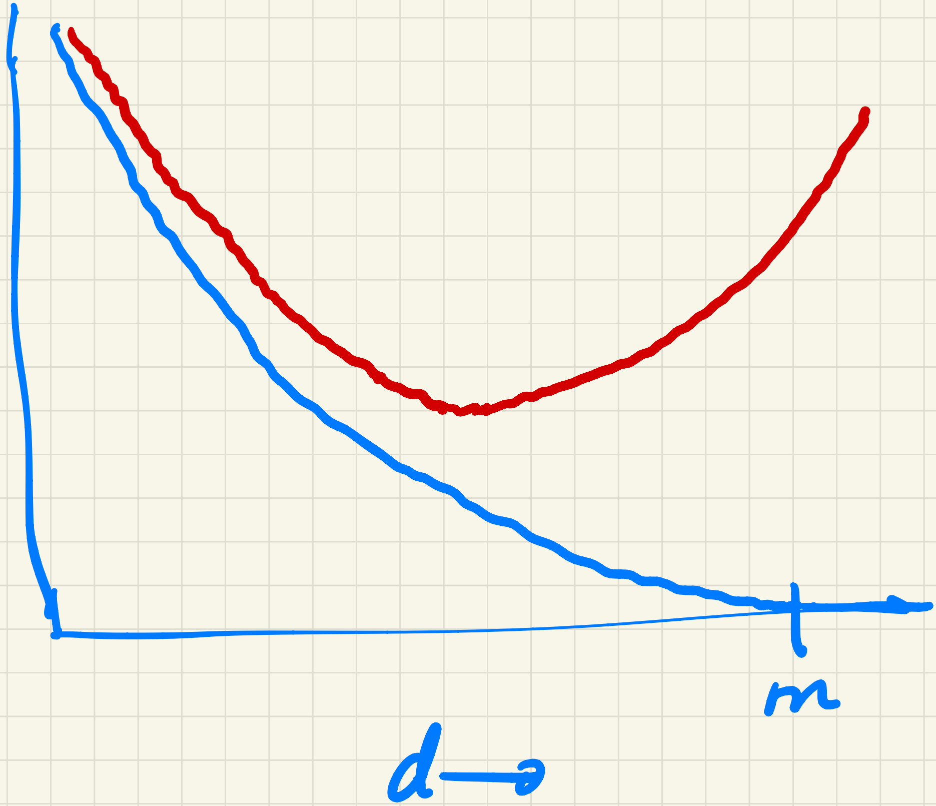
- Let $S = \{\langle x_i, y_i \rangle\}$ be of size m

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

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

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

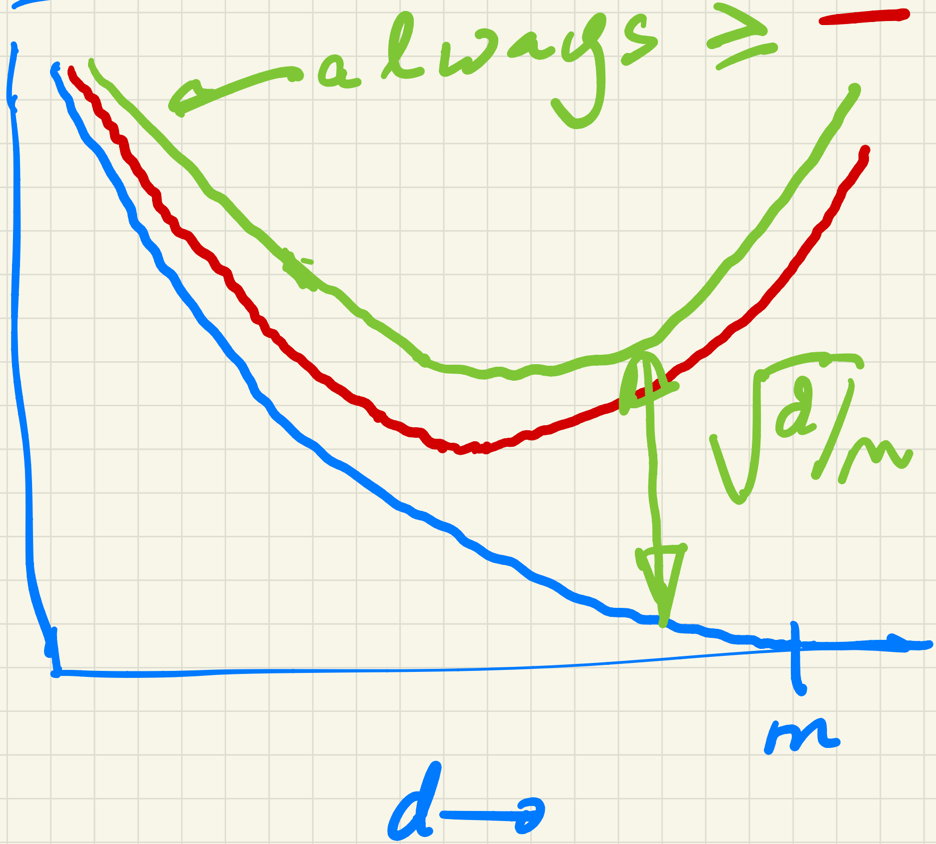
Classic overfitting cartoon:



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

—: true $\mathcal{E}(h)$ of

Classic overfitting cartoon:



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

— : true $\varepsilon(h)$ of

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

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:

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

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

(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...

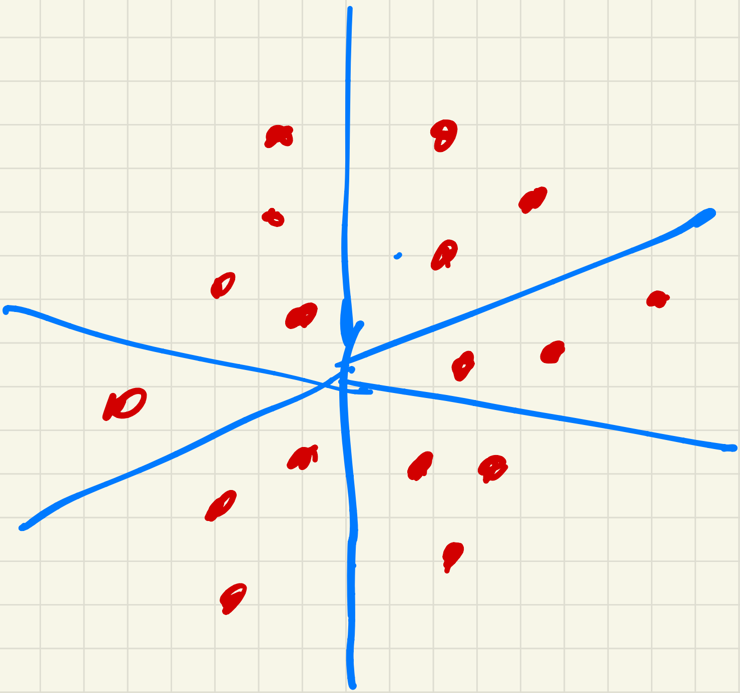
Now let $S = \{z_1, z_2, \dots z_d\}$ and look at

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

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

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

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



"fractal dimension"

"combinatorial dimension"

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