Consistency, Uniform Convergence, and Learning: The VC Dimension
Finite H: Quick Review

- Any consistent \( h \) has
\[
\mathbb{P}(\|h\|_\infty \leq 3) \leq \frac{1}{2} \log \frac{121}{m}
\]
as long as
\[
m \geq \frac{1}{\epsilon} \log \frac{121}{m}
\]
- Even allows \( \|h\|_n \sqrt{m} \)
- \( \frac{1}{3} \to \frac{1}{3^2} \) buys us uniform convergence:
\[
\frac{1}{|A|} \leq \frac{1}{|h(x)|} \leq 3 \quad \forall x \in H
\]
What happens when $1^\infty$ is infinite?

What quantity should replace $\log 1^\infty$?
What's wrong with this:

If \( E(h) \leq Z E \), then it

\[ \sum \text{errors of } h \]

If \( 3 \leq (y_j)^3 \), then prob we missed it

\[ \leq (3-1)^m \]
Let \( S = \{x_1, x_2, \ldots, x_m\} \subseteq X \) be an ordered set (note: no labels).

**Key definition #1:**

\[
\Pi_H(S) = \{ (h(x_1), h(x_2), \ldots, h(x_m)) : h \in H \}
\]

Set of all labelings of \( S \) induced by \( H \).
Key definition #2: Say that $S$ is shattered by $H$ if $\Pi(S) = \{0,1\}^m$.

Key definition #3: The VC dimension of $H$ is the size of the largest set shattered by $H$. Denote by $\text{VC}(H)$. 
Quantification

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

- We will see that \( \text{VC}(H) \) replaces \( \log |H| \) in proving consistency \( \Rightarrow \) PAC

- Warm-up with ex's of computing \( \text{VC}(H) \)

- Main result has both combinatorial and probabilistic components
Ex: Rectangles in $\mathbb{R}^2$

Can shatter these 4 points:

```
3 +

all -
```

etc.
Can shatter no 5 points:

\[ \begin{align*}
T & \cdot f. \\
+ & \cdot + \\
+ & \cdot - \\
+ & \cdot + \\
- & \cdot + \\
& \vdots \\
\therefore \text{VC dim} & = 4
\end{align*} \]
Ex: Halfspaces in $\mathbb{R}^2$

Can shatter 3 points.

e tc.
Can shatter no 4 points:

E.g.

\[ \begin{array}{c}
\cdot + \\
\cdot - \\
\cdot + \\
\cdot - \\
\end{array} \]

and

\[ \begin{array}{c}
\cdot + \\
\cdot - \\
\cdot + \\
\cdot - \\
\end{array} \]

\[ \text{VC dim} = 3 \]

\[ \text{ln } \mathbb{R}^d; \text{ VC dim} = d+1 \]
Ex: d-gons in $\mathbb{R}^2$

Can shatter $2d+1$ pts:
Can shatter no $2d+1$ pts:

Worst case:

forces
4 sides

$\text{VC dim} = 2d+1$
Ex: Conjunctions over $\mathbb{F}_0,1^{13n}$

Can shatter these $n$ pts:

\[
\begin{align*}
0111 \ldots & - \\
10111 \ldots & + \\
11011 \ldots & - \\
11101 \ldots & - \\
111101 \ldots & +
\end{align*}
\Rightarrow \frac{X_1X_3X_4 \ldots}{\text{all - indices}}
\]
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
\( \Pi_H(s) = \text{set of labelings of } S \text{ induced by } H \)
Let's define
\[
\Pi_H(m) = \max \left\{ \left| \Pi_H(s) \right| : s : |s| = m \right\}
\]
"growth function"
Then for \( m \leq \text{VC}(H) \):
\[
\Pi_H(m) = 2^m \ (\text{max})
\]
What about \( m > \text{VC}(H) \)?
Amazing fact:

Let $d = \text{VC}(\mathcal{H})$. Then for $m > d$, $\Pi_{2d}(m) = m^d$.

$m \leq d$: exponential growth

$m > d$: polynomial growth

“Sauer’s Lemma”
Visually:

\[ \log \mathcal{PM}(m) \]

Overall, \( \mathcal{T}(m) = O(m^d) \) (\( 2^m \leq m^m \))
Proof outline

1. Define a function $\phi_d(m)$, show that $\Pi_{2X}(m) \leq \phi_d(m)$

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 \( \nu(c) = d \), then for

\[ \Pi_{\nu} \langle m \rangle \leq \phi_d \langle m \rangle \]

Let's prove by double induction on \( d \) and \( m \).
Assume true for any
\[ d' \leq d \quad (\text{one must}) \]
\[ m' \leq m \quad (\text{be strict}) \]
How many labelings here?

\[ \leq \prod_{k} (m-1) \leq \Phi_d (m-1) \] (induction)
labelings on $x_1, \ldots, x_{m-1}$:

extension(s) on $x_m$:

Need to add count of $\checkmark$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 ≤ $d-1$. Why?
\[ \therefore \; \text{arcs} \leq \Phi_{d-1}(m-1) \]

\text{(induction again)}

\[ \Pi_{\mathcal{N}}(s) \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) \]

\[ = \Phi_d(m) \]

\[ \text{as desired.} \]
Lemma \( \phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \).

Proof \( \phi_d(m) = \phi_{d-1}(m) + \phi_{d-1}(m-1) = \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \) (induction)

= \sum_{i=0}^{d} \binom{m-1}{i-1} + \binom{m-1}{0} \Rightarrow \binom{m-1}{d-1} \equiv 0 \Rightarrow \binom{m}{i} = \binom{m-1}{i-1} + \binom{m-1}{i}
OK, that's it for combinatorial analysis of $\Pi_2^1(m)$. Now for the probabilistic part.
- Let's fix all of these: target c ∈ H, P, E, 8
- Define error regions of H wrt c:

\[ H \Delta c = \{ h : h \in H \} \]

viewed as sets of + ex’s

As functions:

\[ (h \Delta c)(x) = \begin{cases} 1 & \text{if } h(x) \neq c(x) \\ 0 & \text{else} \end{cases} = h(x) \oplus c(x) \]
\textbf{Note:} \( \text{VC}(\mathcal{H}_{\Delta c}) = \text{VC}(\mathcal{H}) \):

- If \( h_1(x) \neq h_2(x) \),
  \( h_1(x) \odot c(x) \neq h_2(x) \odot c(x) \)
- labelings \( h_1(s) \neq h_2(s) \)

\[ \Rightarrow (h_1 \odot c)(s) \neq (h_2 \odot c)(s) \]
\[ \Rightarrow \forall S \left| \prod_{x \in S} h(x) \right| = \left| \prod_{x \in S} h_{\Delta c}(x) \right| \]

\[ \Rightarrow \text{VC}(\mathcal{H}) = \text{VC}(\mathcal{H}_{\Delta c}) \]
OK. Suppose some \( h \in \mathcal{H} \) has \( \exists(h) > \varepsilon \). We want to make sure \( S \) hits\n\( h \) at least once. And we want this to happen for every \( h \) s.t. \( \exists(h) > \varepsilon \).

Call such \( S \) an \( \varepsilon \)-net and define
\[
A(s) = \begin{cases} 
1 & \text{if } S \text{ not an } \varepsilon \text{-net} \\
0 & \text{else}
\end{cases}
\]

Goal: bound \( Pr_s[A(s) = 1] \)
Problem: Hard to tell if $A(s)=1$ by looking at $S$.

Solution: “two-sample trick”

$$B(s, s') = 1 \text{ iff } \exists c \in H_0 \cap 2m \text{ p's s.t. } P[R] > \varepsilon$$

and:

1. $r \cap s = \emptyset \implies A(s)=1$

2. $|r \cap s'| \geq \frac{3m}{2}$

Otherwise $B(s, s') = 0$. 
Claim: \( \Pr[B] \geq \Pr[A] \cdot \frac{1}{2} \)

\( B(s, s') = 1 \Rightarrow A(s) = 1 \), so just need to lower bound \( \Pr_{s, s'}[B(s, s') = 1 \mid A(s) = 1] \)

Can fix missed runs s.t. \( \Pr[r] > \varepsilon \)

\[ \geq \Pr_{s, s'}[1 \leq s] \geq \varepsilon \omega(n) \]

\[ \geq \frac{1}{2} \quad (\text{e.g. Chebyshev}) \]
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 $T \rightarrow S, S'$ randomly (random permutation)

Same as i.i.d. $S, S'$ by exchangability
$B(s,s') = 1$ only if there's
s.t. $r$ is hit $l \geq 3m/2$
times in $T$ but all $l$
hits fall in $S'$

\[ \leq \frac{\binom{m}{e}}{\binom{2m}{e}} \]

Prob. of above happening

just wrt random permutation
\[
\binom{m}{l^c} = \frac{m! \cdot l!}{(m-l)!} \frac{(2m)! \cdot l!}{(2m-l)!} \\
= \frac{m!}{(m-l)!} \frac{(2m-l)!}{(2m)!} \\
= \frac{m(m-1)(m-2)\ldots (m-l+1)}{(2m)(2m-1)\ldots (2m-l+1)} \\
= \frac{m}{2m} \cdot \frac{m-1}{2m-1} \cdot \frac{m-2}{2m-2} \ldots \frac{m-l+1}{2m-l+1} \\
\leq 2^{-l} \leq 2^{-3m/2}
\]
Wrapping up:

\[ \Pr_{s,s', t}[B(s,s') = 1] \leq \Phi_d(2m) \cdot 2^{-\varepsilon m/2} \]

#lossings of T

\[ \leq C_0(2m) \cdot d \cdot 2^{-\varepsilon m/2} \]

(5au2o)

= \exp \left[ d \log m - \varepsilon m \right]

\[ \Pr[A] \leq 2, \ \text{sct} \leq \delta \ \text{and solve...} \]
Theorem. Let \( d = \text{vc}(\mathcal{H}) \).

Then for

\[
m \geq c_2 \left( \frac{1}{3} \log \frac{1}{\delta} + \frac{d}{3} \log \frac{1}{\delta} \right)
\]

with prob. \( \geq 1 - \delta \)

any consistent \( h \in \mathcal{H} \)

has \( \exists \mathbf{y} \in \mathbb{R}^d \)
More generally, if \( \frac{1}{3} \to \frac{1}{3^2} \), have uniform convergence:

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

\[
| \hat{\mathbb{E}}_n(y^n) - \mathbb{E}y^n | \leq \varepsilon.
\]

(slightly different \( B(\mathcal{S}, \mathcal{S}') \) & prob. analysis)
LowerBound #1

• Let $d = \text{VC}(\mathcal{H})$
• Let $x_1, \ldots, x_d$ be shattered
• For $\nu \in \{0, 1\}^d$ let $P \overset{\text{uniform}}{\leftarrow} \nu$
  
  $h_\nu(x_i) = v_i \quad \forall 1 \leq i \leq d$

• Now choose target $C = h_\nu$ for $\sqrt{\nu}$ random

• For $\varepsilon \leq \frac{1}{4}$, need $\Omega(d)$ examples – are just predicting coin flips!
Better lower bound?

E.g. $\sqrt{\frac{d}{\varepsilon}}$ for any $\varepsilon$?
Lower Bound #2

- Same set-up as #1, but now we'll "give away" \( c(x_i) \) each.

\[ P: 1 - \frac{1}{3^E + \frac{1}{d-1}} \]

- Still choose \( c = \) random \( h \sqrt{v} \)
- Now only see coin flip every \( \approx \frac{1}{3} \) samples \( \Rightarrow N(d/3). \)
A Prescriptive Application of VC: Structural Risk Minimization
Suppose we have not a single $H$ but a nested hierarchy:

$$H_1 \subset H_2 \subset H_3 \ldots \subset H_d \subset \ldots$$

- E.g. neural networks of increasing depth/width
- For simplicity assume $\text{VC}(H_d) = d$
- Let $S = \{(x_i, y_i)\}$ be of size $m$
VC theory says that that in $H_d$, $\frac{1}{\sqrt{m}} |\mathbb{E}\{(y^3 - y^3)_{y \in \mathcal{H}}\}|\leq \sqrt{a \log m}$.

(ignoring log factors, fixing $d$, spreading $\delta$ over the $H$'s)
Classic overfitting cartoon:

- $d$: optimal $\hat{\mathcal{E}}_s(h)$ in $\mathcal{H}_d$
- $\cdot$: true $\mathcal{E}(h)$ of
Classic overfitting cartoon:

\[ d \rightarrow \]

\[ \text{always} \geq - \]

\[ \frac{d}{m} \]

\[ \text{optimal } \hat{\epsilon}_s(h) \text{ in } \mathcal{H}_d \]

\[ \text{true } \epsilon(h) \text{ of } \]

\[ \hat{\epsilon}_s(h) + \sqrt{\frac{d}{m}} \]
So choose:

\[
d^* = \arg \min_{d} \left\{ \sum_{s} \left( h_d^* + \sqrt{\frac{d}{m}} \right) \right\}
\]

\[
h_d^* = \arg \min_{h \in H_d} \left\{ \sum_{s} \hat{E}_s(h) \right\}
\]

Then our true error

\[
\leq \min_{d} \left\{ \sum_{s} \text{samec} \right\}
\]
Distribution-Dependent Improvements to VC Theory
• Replace $\Pi_{m} (m)$ by $E P[1 \Pi_{m} (s)]$ VC Entropy

• Related: Rademacher complexity

• Replace $12 \ell (1-3)^{m}$ by $\sum_{3 \ell } 12 \ell (1-3)^{m}$

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 $l(h, z) \in \mathbb{R}$ also i.i.d.

**E.g. regression:**

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

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

(squared error)

or

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

(absolute error)

or

$$l(h, \langle x, y \rangle) = (h(x) - y)^2 + \alpha |h(x)|$$
Distribution learning

- Models are distributions $P \in \mathcal{H}$
- Observations $Z = X - P^*$
- Loss $l(P, X) = \log \frac{1}{p(x)}$

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

$$\{ \langle h(z_1), \ldots, h(z_d) \rangle : h \in \mathcal{H} \}$$

$\mathcal{H}$ is the generalization of $\Pi_{x}(S)$, now possibly in finite!

Say $\mathcal{H}$ shatters $S$ if $\Pi_{x}(S)$ is “space filling”...
E.g. $\mathbb{P}_H(s)$ intersects all 2d orthants of $\mathbb{R}^d$.

"fractal dimension"

"combinatorial dimension"
Moral: Uniform convergence is the norm, not the exception.