

Learnability of Infinite Hypothesis Classes

Uri Shaham

May 31, 2026

1 The VC Dimension

In the case of finite $\mathcal{H}$, the term $\ln(|\mathcal{H}|)$ in the expression for $n(\epsilon, \delta)$ can be viewed as a measure for the richness of $\mathcal{H}$. Is there an equivalent measure of richness for infinite $\mathcal{H}$? The answer is yes, and is called “the VC dimension” of a hypothesis class, named after mathematicians Vladimir Vapnik and Alexey Chervonenkis. Before we define it formally, let us briefly present the intuition: we always have a finite training sample, say of n points. The set of all possible labelings of the training sample is of size 2^n . Thus, a richness measure for infinite hypothesis classes can be to count how many of these possible 2^n labeling can be obtained via hypotheses from the hypothesis class $\mathcal{H}$. This leads us to the definition of shattering, which will in turn lead to the one of the VC dimension.

Definition 1.1 (shattering). *A hypothesis class $\mathcal{H}$ shatters a finite set $C = \{x_1, \dots, x_n\} \subset \mathcal{X}$ if for any labels $y_1, \dots, y_n \in \{0, 1\}$ there exists $h \in \mathcal{H}$ such that $h(x_i) = y_i$ for all $i = 1, \dots, n$.*

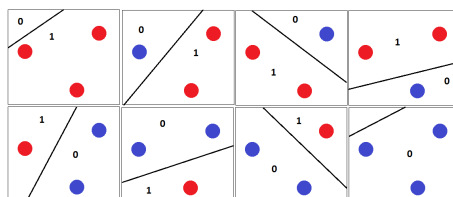
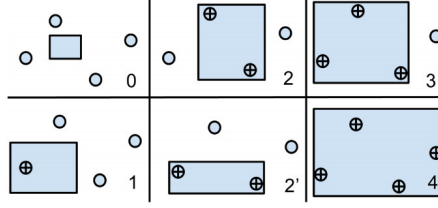


Figure 1: The set $\mathcal{H}$ of half-spaces shatters the set C of 3 points in $\mathcal{X} = \mathbb{R}^2$.

Definition 1.2 (VC dimension). *The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $VCdim(\mathcal{H})$ is the size of the largest set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.*

Examples:

- The VC dimension of the class of threshold functions over $\mathbb{R}$ is 1 (why?)
- The VC dimension of the class of intervals over $\mathbb{R}$ is 2 (why?)
- The VC dimension of the class of axis-aligned rectangles over $\mathbb{R}^2$ is 4 (why?)



Proving that rectangle concept space shatters at least 4

Figure 2: The VC dimension of axis-aligned rectangles in $\mathbb{R}^2$ is at least 4.

2 The Fundamental Theorem of Statistical Learning

The theorem states that hypothesis classes with finite VC dimension have the universal convergence property.

Theorem 2.1. *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0, 1\}$, such that $VCdim(\mathcal{H}) = d < \infty$. Then there exists a constant $c > 0$ such that for each $\epsilon, \delta \in (0, 1)$ there is $n = n(\epsilon, \delta)$ for which with probability $1 - \delta$ (over choice of S_n), we have*

$$\sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)| < c \sqrt{\frac{d}{n}}.$$

Corollary 2.2. *Under the conditions in the theorem, $\mathcal{H}$ has the uniform convergence property, and hence is PAC-learnable.*

In the two following subsections, we will prove the fundamental theorem of statistical learning. We will do so in two steps: (i) We will define the growth function as a quantifier of the effective size of a hypothesis class $\mathcal{H}$ and prove that finite VC dimension implies polynomial growth (Sauer lemma). (ii) We will prove that polynomial growth implies uniform convergence.

2.1 Proof of the Fundamental Theorem of Statistical Learning: Sauer Lemma

We have defined the richness of a hypothesis class via measuring the capability to yield different labelings on finite sets of n points. Clearly, if $d := VCdim(\mathcal{H}) \leq n$ is less than n , $\mathcal{H}$ attains all 2^n possible labelings. The growth function counts the possible labelings when n is larger than d .

Definition 2.3 (Growth Function). *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0, 1\}$, and let $C \subset \mathcal{X}$. Let $\mathcal{H}_C$ be the collection of all functions $h \in \mathcal{H}$, restricted to C . The growth function of $\mathcal{H}$ is*

$$\tau_{\mathcal{H}}(n) = \max_{C: |C|=n} |\mathcal{H}_C|.$$

A key to proving the fundamental theorem is the *Sauer Lemma*, which shows that if $d < \infty$, the growth function is polynomial (rather than exponential) in n . Put another way, if there is one labeling that $\mathcal{H}$ cannot overfit, there is at most a polynomial number of labelings that it can.

Lemma 2.4 (Sauer's lemma). *If $VCdim(\mathcal{H}) = d < \infty$, $\tau_{\mathcal{H}}$ grows **polynomially** in n . Specifically, for $n > d + 1$, $\tau_{\mathcal{H}}(n) \leq (en/d)^d$.*

Proof. First, observe that given a set C of size n , $|B \subseteq C : \mathcal{H} \text{ shatters } B| \leq \sum_{i=1}^d \binom{n}{i}$, and observe that $\sum_{i=1}^d \binom{n}{i}$ is polynomial in n (specifically, $\sum_{i=1}^d \binom{n}{i} < (en/d)^d$). Hence, it suffices to prove that for any set C , the size of the set $\mathcal{H}_C$ of possible labelings attained by $\mathcal{H}$ is less than $|B \subseteq C : \mathcal{H} \text{ shatters } B|$.

The proof goes by induction. The base case is $n = 1$. In this case, either $|\mathcal{H}_C| = 1$, and then $|B \subseteq C : \mathcal{H} \text{ shatters } B| = 1$ as well (by $B = \emptyset$), or $|\mathcal{H}_C| = 2$, in which case $|B \subseteq C : \mathcal{H} \text{ shatters } B| = 2$ as well (by $B = \emptyset, B = C$).

For the induction step, we assume that $|\mathcal{H}_C| = 1 \leq |B \subseteq C : \mathcal{H} \text{ shatters } B|$ for all C of size $k < n$, and prove for a set C of size n . Let $C = \{c_1, \dots, c_n\}$.

We will partition the set $\mathcal{H}_C$ to two disjoint sets $\mathcal{H}_C = \mathcal{H}_1 \cup \mathcal{H}_2$: for every $h \in \mathcal{H}_C$, if there exists $h' \in \mathcal{H}_C$ which agrees with h on $c_n, \dots, c_2$, and differs on c_1 , put h in $\mathcal{H}_1$, and otherwise put h in $\mathcal{H}_2$. Now,

- By the induction hypothesis, $|\mathcal{H}_2| \leq |B \subseteq \{c_2, \dots, c_n\} : \mathcal{H}_2 \text{ shatters } B|$.
- Clearly, every set $B \subseteq \{c_2, \dots, c_n\}$ that is shattered by $\mathcal{H}_2$, is also shattered by $\mathcal{H}$.
- In addition, by the induction hypothesis $|\mathcal{H}_1| \leq |B \subseteq \{c_2, \dots, c_n\} : B \cup \{c_1\} \text{ is shattered by } \mathcal{H}_1|$.
- Clearly, every set $B \subseteq C$ that is shattered by $\mathcal{H}_1$, is also shattered by $\mathcal{H}$.

Altogether this yields:

$$\begin{aligned} |\mathcal{H}_C| &= |\mathcal{H}_1| + |\mathcal{H}_2| \leq |B \subseteq C : c_1 \notin B, \mathcal{H} \text{ shatters } B| + |B \subseteq C : c_1 \in B, \mathcal{H} \text{ shatters } B| \\ &= |B \subseteq C : \mathcal{H} \text{ shatters } B|. \end{aligned}$$

□

2.2 Polynomial Growth implies Uniform Convergence

We will show that for any hypothesis class $\mathcal{H}$ with growth function $\tau_{\mathcal{H}}$, with probability at least $1 - \delta$ over the choice of training set $S_n \sim \mathcal{D}^n$, we have

$$|L_{S_n}(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta \sqrt{2n}}. \quad (1)$$

Using Sauer lemma and plugging $\tau_{\mathcal{H}}(n) \leq (en/d)^d$, we get

$$|L_{S_n}(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{d \log((2en/d))}}{\delta \sqrt{2n}},$$

which shows that uniform convergence holds ($|L_{S_n}(h) - L_{\mathcal{D}}(h)|$ is uniformly bounded for every $h \in \mathcal{H}$).

We will prove equation (1) by proving that

$$\mathbb{E}_{S_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)| \right] \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta \sqrt{2n}},$$

and applying Markov inequality (for nonnegative random variable X , $\Pr(X > \delta) \leq \frac{\mathbb{E}[X]}{\delta}$) on the random variable $X = \sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)|$.

Proof. Observe that for every $h \in \mathcal{H}$, we can write $L_D \mathcal{D}(h)$ as $\mathbb{E}_{S'_n \sim \mathcal{D}^n} L_{S'_n}(h)$, where $S' = \{z'_1, \dots, z'_n\}$ is an additional training sample of size n .

$$\begin{aligned}
\mathbb{E}_{S_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_D(h)| \right] &= \mathbb{E}_{S_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} \left| L_{S_n}(h) - \mathbb{E}_{S'_n \sim \mathcal{D}^n} L_{S'_n}(h) \right| \right] \\
&= \mathbb{E}_{S_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{S'_n \sim \mathcal{D}^n} [L_{S_n}(h) - L_{S'_n}(h)] \right| \right] \\
&\leq \mathbb{E}_{S_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} \mathbb{E}_{S'_n \sim \mathcal{D}^n} [|L_{S_n}(h) - L_{S'_n}(h)|] \right] \\
&\leq \mathbb{E}_{S_n, S'_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{S'_n}(h)| \right] \\
&= \mathbb{E}_{S_n, S'_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^n \ell(h, z_i) - \ell(h, z'_i) \right| \right]. \tag{2}
\end{aligned}$$

where the first inequality follows from Jensen's inequality (with the convex function $|\cdot|$), and the second inequality follows since supremum of expectations is less than or equal to the expectation of a supremum (this also can follow from Jensen, as the max function is convex).

Since all z_i and z'_i are sampled i.i.d, their notations are arbitrary and we can introduce *Rademacher* variables σ_i , distributed uniformly over $\{-1, 1\}$, and write the RHS of equation (2) as

$$\mathbb{E}_{S_n, S'_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\ell(h, z_i) - \ell(h, z'_i)) \right| \right],$$

for any $\sigma_i \in \{-1, 1\}$. Since this holds for any value of $\sigma_i \in \{-1, 1\}$, it also holds for the expectation, so we can write this as

$$\mathbb{E}_{S_n, S'_n \sim \mathcal{D}^n, \sigma_1, \dots, \sigma_n} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\ell(h, z_i) - \ell(h, z'_i)) \right| \right].$$

Next, we fix S_n, S'_n , and let $C = S_n \cup S'_n$. Then we can replace the supremum over $h \in \mathcal{H}$ by the maximum over $h \in \mathcal{H}_C$, i.e.,

$$\begin{aligned}
&\mathbb{E}_{\sigma_1, \dots, \sigma_n} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\ell(h, z_i) - \ell(h, z'_i)) \right| \right] \\
&= \mathbb{E}_{\sigma_1, \dots, \sigma_n} \left[\max_{h \in \mathcal{H}_C} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\ell(h, z_i) - \ell(h, z'_i)) \right| \right]. \tag{3}
\end{aligned}$$

Now fix h and denote $\theta_h = \frac{1}{n} \sum_{i=1}^n \sigma_i (\ell(h, z_i) - \ell(h, z'_i))$. Observe that θ_h is a sum of independent zero-mean random variables, each of which is bounded between $[-1, 1]$. Hence, we can apply Hoeffding inequality to get

$$\Pr(|\theta_h| > \rho) \leq 2 \exp(-2n\rho^2).$$

We next apply union bound (over all $h \in \mathcal{H}_C$ to write

$$\Pr\left(\max_{h \in \mathcal{H}_C} |\theta_h| > \rho\right) \leq 2 |\mathcal{H}_C| \exp(-2n\rho^2).$$

Lemma A4 in UML then gives

$$\Pr(\max_{h \in \mathcal{H}_C} |\theta_h|) \leq \frac{4 + \sqrt{\log \|\mathcal{H}_C\|}}{\sqrt{2n}}.$$

Lastly, recalling the definition of the growth function $\tau_{\mathcal{H}}$, we conclude that

$$\mathbb{E}_{S_n \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)| \right] \leq \frac{4 + \sqrt{\log \tau_{\mathcal{H}}(2n)}}{\sqrt{2n}}. \quad (4)$$

□

3 Reading

UML ch. 6