UC Berkeley **Jonathan Shafer**
CS 294-220, Spring 2021

Unit 5

The Fundamental Theorem of PAC Learning

1 Introduction

In Unit 4 we saw Claim 15, which states a lower bound of $\Omega(\mathsf{VC}(\mathcal{H}))$ on the sample complexity of PAC learning $\mathcal{H}$, and Lemma 11, which states an upper bound on the sample complexity of agnostic PAC learning that follows from uniform convergence. Our main result on characterizing learnability and sample complexity will tie together the notions of uniform convergence and VC dimension, showing that these two bounds are tight: every class of finite VC dimension satisfies uniform convergence with sample complexity $O(\mathsf{VC}(\mathcal{H}))$. Hence, a class is PAC and agnostic PAC learnable if and only if it has finite VC dimension, and if the VC dimension is finite then the sample complexity of learning is $\Theta(\mathsf{VC}(\mathcal{H}))$.

2 The Growth Function and Sauer's Lemma

The growth function measures how rich a class $\mathcal{H}$ is on finite subsets of the domain.

▶ **Definition 1.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. The growth function of $\mathcal{H}$ is a function $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ given by*

$$\tau_{\mathcal{H}}(m) = \sup\left\{\left|\mathcal{H}|_A\right| : \ A \subseteq \mathcal{X} \ \wedge \ |A| = m\right\}.$$

Notice that the growth function and the VC dimension are related by

$$\mathsf{VC}(\mathcal{H}) = \sup\{m \in \mathbb{N} : \ \tau_{\mathcal{H}}(m) = 2^m\}.$$

The following combinatorial lemma describes how the growth function behaves more generally, and its relation to the VC dimension. Conceptually, it has two phases: below the VC dimension, $\tau_{\mathcal{H}}$ grows exponentially; above the VC dimension, $\tau_{\mathcal{H}}$ grows at most polynomially.

▶ **Lemma 2** (Sauer)**.** *Let $\mathcal{X}$ be a set and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ such that $\mathsf{VC}(\mathcal{H}) = d < \infty$. Then for natural numbers m,*

$$\forall m \le d : \ \tau_{\mathcal{H}}(m) = 2^m,$$
$$\forall m > d : \ \tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} \le \left(\frac{em}{d}\right)^d \overset{(\star)}{\le} m^d, \tag{1}$$

where $(\star)$ holds whenever $d > e$.

The lemma is a consequence of the following claim.

▷ Claim 3. Let $\mathcal{X}$ be a set and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ such that $\mathsf{VC}(\mathcal{H}) = d < \infty$. Then for any $A \subseteq \mathcal{X}$,

$$\left\{B \subseteq A \;\middle|\; \exists h \ \in \mathcal{H} : \ B = A \cap h^{-1}(1)\right\} \le \left|\left\{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B\right\}\right|,$$

where $h^{-1}(1) = \left\{x \in \mathcal{X} \;\middle|\; h(x) = 1\right\}$.

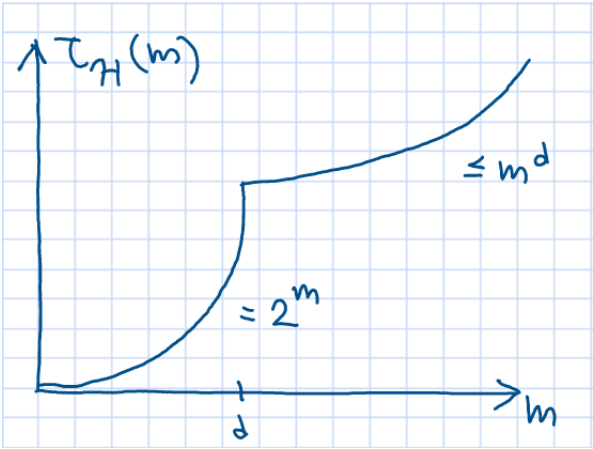

Figure 1 Sauer's lemma says that the growth function $\tau_{\mathcal{H}}$ has two distinct phases, which are determined by the VC dimension of the class, $d = \mathsf{VC}(\mathcal{H})$. For inputs $m \leq d$, it precisely equals 2^m. For inputs $m > d$, it grows at most polynomially.

Proof of Lemma 2. If $m \leq d$ then $\tau_{\mathcal{H}}(m) = 2^m$ by the definition of $\tau_{\mathcal{H}}$ and the VC dimension.

For the case $m > d$, fix $m \in \mathbb{N}$ such that $d < m \leq |\mathcal{X}|$. For any $A \subseteq \mathcal{X}$ such that $|A| = m$,

$$\begin{aligned}
\left|\mathcal{H}|_A\right| &= \left|\left\{B \subseteq A \;\middle|\; \exists h \in \mathcal{H} :\ B = A \cap h^{-1}(1)\right\}\right| && \text{(def. of } \mathcal{H}|_A\text{)} \\
&\leq \left|\left\{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B\right\}\right| && \text{(Claim 3)} \\
&= \left|\left\{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B \ \wedge\ |B| \leq d\right\}\right| && (\mathsf{VC}(\mathcal{H}) = d) \\
&\leq \left|\left\{B \subseteq A \;\middle|\; |B| \leq d\right\}\right| \\
&= \sum_{i=0}^{d} \binom{m}{i}.
\end{aligned}$$

Hence, $\tau_{\mathcal{H}}(m) = \max_{A \subseteq \mathcal{X},\, |A|=m} \left|\mathcal{H}|_A\right| \leq \sum_{i=0}^{d} \binom{m}{i}$. This completes the proof of the first inequality in Eq. (1).

For the second inequality in Eq. (1),

$$\begin{aligned}
\sum_{i=0}^{d} \binom{m}{i} &\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} && ((m/d)^{d-i} \geq 1 \text{ for all } i \leq d) \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d} \left(1 + \left(\frac{d}{m}\right)\right)^{m} && \text{(binomial formula)} \\
&\leq \left(\frac{m}{d}\right)^{d} e^{d} = \left(\frac{em}{d}\right)^{d}. && (\forall x \in \mathbb{R} :\ 1 + x \leq e^{x})
\end{aligned}$$

◀

Proof of Claim 3. We proceed by induction on $n = |A|$. For the base case $n = 0$,

$$\left\{B \subseteq A \;\middle|\; \exists h \in \mathcal{H} :\ B = A \cap h^{-1}(1)\right\} = \{\varnothing\} = \{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B\}.$$

For the induction step, we assume the claim holds for all sets $A \subseteq \mathcal{X}$ of cardinality n, and prove that it holds for all sets of cardinality $n+1$. Fix a set $A \subseteq \mathcal{X}$ with $|A| = n+1$, fix $a \in A$, and let $A' = A \setminus \{a\}$. Define

$$H_1' = \left\{ g : A' \to \{0,1\} \;\middle|\; \exists h \in \mathcal{H} :\ g = h|_{A'} \right\} = \mathcal{H}|_{A'},$$

$$H_2' = \left\{ g : A' \to \{0,1\} \;\middle|\; \exists h_1, h_2 \in \mathcal{H} :\ g = h_1|_{A'} = h_2|_{A'} \ \wedge\ h_1(a) \neq h_2(a) \right\} \subseteq \mathcal{H}|_{A'},$$

$$H_2 = \left\{ g : A \to \{0,1\} \;\middle|\; \exists h_1, h_2 \in \mathcal{H} :\ g|_{A'} = h_1|_{A'} = h_2|_{A'} \ \wedge\ h_1(a) \neq h_2(a) \right\} \subseteq \mathcal{H}|_{A}.$$

Namely, $H_2' = H_2|_{A'}$.

Observe that

$$\begin{aligned}
\left|\mathcal{H}|_A\right| &= \left|\left\{ g : A \to \{0,1\} \;\middle|\; \exists h \in \mathcal{H} :\ g = h|_A \right\}\right| \\
&= \left|\left\{ g : A' \to \{0,1\} \;\middle|\; \exists h \in \mathcal{H} :\ g = h|_{A'} \ \wedge\ h(a) = 0 \right\}\right| \\
&\quad + \left|\left\{ g : A' \to \{0,1\} \;\middle|\; \exists h \in \mathcal{H} :\ g = h|_{A'} \ \wedge\ h(a) = 1 \right\}\right| \\
&= \left|\left\{ g : A' \to \{0,1\} \;\middle|\; \exists h \in \mathcal{H} :\ g = h|_{A'} \right\}\right| && (2) \\
&\quad + \left|\left\{ g : A' \to \{0,1\} \;\middle|\; \exists h_1, h_2 \in \mathcal{H} :\ g = h_1|_{A'} = h_2|_{A'} \ \wedge\ h_1(a) = 0 \ \wedge\ h_2(a) = 1 \right\}\right| \\
&= |H_1'| + |H_2'|. && (3)
\end{aligned}$$

Eq. (2) holds because we can count the restrictions of $\mathcal{H}$ to A by counting the number of restrictions of $\mathcal{H}$ to A', where we count twice any restriction that has both possible extensions $h(a) = 0$ and $h(a) = 1$.

Consider each term separately.

$$\begin{aligned}
|H_1'| &= \left|\left\{ g : A' \to \{0,1\} \;\middle|\; \exists h \in \mathcal{H} :\ g = h|_{A'} \right\}\right| \\
&= \left|\left\{ B \subseteq A' \;\middle|\; \exists h \in \mathcal{H} :\ B = A \cap h^{-1}(1) \right\}\right| \\
&\leq \left|\left\{ B \subseteq A' \;\middle|\; \mathcal{H} \text{ shatters } B \right\}\right| && \text{(induction hypothesis)} \\
&= \left|\left\{ B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B \ \wedge\ a \notin B \right\}\right|. && (4)
\end{aligned}$$

$$\begin{aligned}
|H_2'| &= \left|H_2|_{A'}\right| \\
&= \left|\left\{ B \subseteq A' \;\middle|\; \exists h \in H_2 :\ B = A \cap h^{-1}(1) \right\}\right| \\
&\leq \left|\left\{ B \subseteq A' \;\middle|\; H_2 \text{ shatters } B \right\}\right| && \text{(induction hypothesis)} && (5) \\
&= \left|\left\{ B \subseteq A \;\middle|\; H_2 \text{ shatters } B \ \wedge\ a \in B \right\}\right| && && (6) \\
&\leq \left|\left\{ B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B \ \wedge\ a \in B \right\}\right|. && \text{(because } H_2 \subseteq \mathcal{H}|_A\text{)} && (7)
\end{aligned}$$

To understand the equality in line (6), notice (by the definition of H_2) that H_2 shatters $B \subseteq A'$ iff H_2 shatters $B \cup \{a\}$. Hence $B \mapsto B \cup \{a\}$ is a bijection from the set in line (5) to the set in line (6).

Combining Eq. (3), (4) and (7) yields

$$\begin{aligned}\left|\mathcal{H}|_A\right| &\leq \left|\left\{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B \;\wedge\; a \notin B\right\}\right| + \left|\left\{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B \;\wedge\; a \in B\right\}\right| \\ &= \left|\left\{B \subseteq A \;\middle|\; \mathcal{H} \text{ shatters } B\right\}\right|,\end{aligned}$$

as desired. ◀

3 Uniform Convergence for Classes of Finite VC Dimension

▶ **Definition 4.** *A set system is a pair $(\Omega, \mathcal{R})$ such that Ω is a nonempty set and $\mathcal{R}$ is a set of subsets of Ω.*

▶ **Definition 5.** *Let $(\Omega, \mathcal{R})$ be a set system, let $\mathcal{D}$ be a distribution over Ω, and let*

$$S = (z_1, \ldots, z_m) \in \Omega^m$$

for some $m \in \mathbb{N}$. Let $\varepsilon > 0$.

(i) *We say that S is an ε-net for $(\Omega, \mathcal{R})$ with respect to distribution $\mathcal{D}$ if*

$$\forall R \in \mathcal{R} : \; \mathcal{D}(R) > \varepsilon \implies S \cap R \neq \varnothing.$$

(ii) *We say that S is an ε-representative sample for $(\Omega, \mathcal{R})$ with respect to distribution $\mathcal{D}$ if*

$$\forall R \in \mathcal{R} : \; \left|\frac{|S \cap R|}{m} - \mathcal{D}(R)\right| \leq \varepsilon.$$

The notation $S \cap R$ above denotes $\{i \in [m] : \; z_i \in R\}$.

Observe that every ε-representative sample for $(\Omega, \mathcal{R})$ is also an ε-net for for $(\Omega, \mathcal{R})$, but not vice versa. Similarly, we can define ε-nets and ε-representative samples for classes of binary functions.

▶ **Definition 6.** *Let $\mathcal{X}$ be a nonempty set, let $\mathcal{H}$ be a set of function $\mathcal{X} \to \{0,1\}$, let $\Omega = \mathcal{X} \times \{0,1\}$, let $S = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in \Omega^m$, and let $\mathcal{D}$ be a distribution over Ω. For each $h \in \mathcal{H}$, let*

$$R_h = \big\{(x, y) \in \Omega : \; y \neq h(x)\big\},$$

and let $\mathcal{R} = \{R_h : \; h \in \mathcal{H}\}$. We say that S is an ε-net (ε-representative sample) for class $\mathcal{H}$ with respect to distribution $\mathcal{D}$ if it is an ε-net (ε-representative sample) for $(\Omega, \mathcal{R})$ with respect to $\mathcal{D}$.

In other words, the 0-1 loss satisfies that:

- S is an ε-net for $\mathcal{H}$ with respect to $\mathcal{D}$ if

$$\forall h \in \mathcal{H} : \; L_{\mathcal{D}}(h) > \varepsilon \implies L_S(h) > 0.$$

- S is an ε-representative sample for $\mathcal{H}$ with respect to $\mathcal{D}$ if

$$\forall h \in \mathcal{H} : \; |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

▶ Remark 7. $\mathcal{H}$ has the uniform convergence property if and only if for any $\varepsilon, \delta \in (0,1)$ there exists $m \in \mathbb{N}$ such that for any distribution $\mathcal{D}$, $\mathbb{P}_{S \sim \mathcal{D}^m}\left[S \text{ is an } \varepsilon\text{-representative sample}\right] \geq 1 - \delta$.

▶ **Theorem 8.** *There exists a function*

$$m(d, \varepsilon, \delta) = O\left(\frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon}\right)$$

such that the following holds. For any $\varepsilon, \delta \in (0,1)$, nonempty sets $\mathcal{X}$ and $\mathcal{Y}$, distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and class $\mathcal{H}$ of functions $\mathcal{X} \to \mathcal{Y}$, if $\mathsf{VC}(\mathcal{H}) = d < \infty$ and $S \sim \mathcal{D}^m$ such that $m \geq m(d, \varepsilon, \delta)$, then with probability at least $1 - \delta$, S is an ε-net for $\mathcal{H}$ with respect to the distribution $\mathcal{D}$.

▶ **Theorem 9.** *There exists a function*

$$m(d, \varepsilon, \delta) = O\left(\frac{d \ln(d/\varepsilon) + \ln(1/\delta)}{\varepsilon^2}\right)$$

such that the following holds. For any $\varepsilon, \delta \in (0,1)$, nonempty sets $\mathcal{X}$ and $\mathcal{Y}$, distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, class $\mathcal{H}$ of functions $\mathcal{X} \to \mathcal{Y}$, and loss function $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to [0,1]$, if $\mathsf{VC}(\mathcal{H}) = d < \infty$ and $S \sim \mathcal{D}^m$ such that $m \geq m(d, \varepsilon, \delta)$, then:

- **(a)** $\mathbb{P}\left[\exists h \in \mathcal{H} : \ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\right] \leq 4\tau_{\mathcal{H}}(2m) \exp\left(-\frac{\varepsilon^2 m}{8}\right)$.
- **(b)** *With probability at least $1 - \delta$, S is an ε-representative sample for $\mathcal{H}$ with respect to distribution $\mathcal{D}$ and loss function ℓ.*

▶ Remark 10. The proofs of the two theorems are very similar and contain the same ideas. We present the proof of Theorem 9, which is slightly more involved, and leave proving Theorem 8 as an exercise.

Proof of Theorem 9. Let $S = (z_1, \ldots, z_m)$, and let $S' \sim \mathcal{D}^m$, $S' = (z'_1, \ldots, z'_m)$ be an additional sample taken independently of S.[1] For any $h \in \mathcal{H}$, let $\Delta_S(h) = |L_S(h) - L_{\mathcal{D}}(h)|$. Consider the following two events.

$$E_1 = \{\exists h \in \mathcal{H} : \ \Delta_S(h) > \varepsilon\},$$
$$E_2 = \{\exists h \in \mathcal{H} : \ \Delta_S(h) > \varepsilon \ \wedge \ \Delta_{S'}(h) \leq \varepsilon/2\}.$$

We need to show that $\mathbb{P}[E_1] \leq \delta$. The proof is partitioned into three claims.

Claim I: $\mathbb{P}[E_1] \leq 2\mathbb{P}[E_2]$.

Intuitively, the idea is that for a fixed $h \in \mathcal{H}$, each of $L_S(h)$ and $L_{S'}(h)$ is a good estimate of $L_{\mathcal{D}}(h)$, and they are independent. Hence, even if we fix a particular h such that $L_S(h)$ is a bad estimate, we can still expect that $L_{S'}(h)$ will be a good estimate.

Note that $\mathbb{P}[E_2] \geq \mathbb{P}[E_1 \cap E_2] = \mathbb{P}[E_2 \mid E_1]\,\mathbb{P}[E_1]$. Hence, to prove Claim I it suffices to show that $\mathbb{P}[E_2 \mid E_1] \geq 1/2$. Seeing as E_2 is a subset of E_1,

$$\begin{aligned}\mathbb{P}[E_2 \mid E_1] &= \mathbb{P}\left[\exists h \in \mathcal{H} : \ \Delta_S(h) > \varepsilon \ \wedge \ \Delta_{S'}(h) \leq \varepsilon/2 \ \middle| \ \exists g \in \mathcal{H} : \ \Delta_S(g) > \varepsilon\right] \\ &\geq \mathbb{P}\left[\Delta_{S'}(g) \leq \varepsilon/2 \ \middle| \ \exists g \in \mathcal{H} : \ \Delta_S(g) > \varepsilon\right]. \end{aligned} \tag{8}$$

[1] S' is sampled purely to aid our analysis in the proof, it is not necessary to actually take any additional samples beyond S to obtain an ε-representative sample.

Notice that for any $g \in \mathcal{H}$, $L_{S'}(g) - L_{\mathcal{D}}(g) = \frac{1}{m}\sum_{i\in[m]}(\ell(g, z'_i) - L_{\mathcal{D}}(g))$, and furthermore $Z_i = \ell(g, z'_i) - L_{\mathcal{D}}(g)$ for $i \in [m]$ are i.i.d. random variables with mean 0 and support in $[-1, 1]$. So for any fixed $g \in \mathcal{H}$, Hoeffding's inequality implies,

$$\mathbb{P}\left[\Delta_{S'}(g) > \varepsilon/2\right] = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i\in[m]} Z_i\right| > \frac{\varepsilon}{2}\right] \leq 2\exp\left(-\frac{\varepsilon^2 m}{8}\right) \leq \frac{1}{2}, \tag{9}$$

where the last inequality holds for m as in the statement of the theorem.[2] This holds also when conditioning on the event $\Delta_S(g) > \varepsilon$, because $S' \perp S$. Combining (8) and (9) implies $\mathbb{P}\left[E_2 \mid E_1\right] \geq 1/2$, concluding the proof of Claim I.

Claim II: $\mathbb{P}\left[E_2\right] \leq \tau_{\mathcal{H}}(2m) \cdot 2\exp\left(-\frac{\varepsilon^2 m}{8}\right)$.

Intuitively, seeing as $L_S(h)$ and $L_{S'}(h)$ are both good estimates of the same value $L_{\mathcal{D}}(h)$, the probability that they be markedly different for a particular h is small. Let $S_x, S'_x \subseteq \mathcal{X}$ be the set of domain elements that appear in S and S' respectively. A key idea in the proof is that even though $\mathcal{H}$ is an infinite class, the event in which $L_S(h)$ and $L_{S'}(h)$ are very different for a particular h is an event that concerns only how h behaves on the set $X = S_x \cup S'_x$, which is a finite subset of the domain (and this event does *not* depend on how h and $\mathcal{D}$ behave outside of X). Hence, we can restrict our attention to the projections $h|_X \in \mathcal{H}|_X$ instead of considering functions $h \in \mathcal{H}$. For any particular restricted function $h|_X$, the probability that $L_S(h)$ and $L_{S'}(h)$ are very different vanishes exponentially. Seeing as $\mathcal{H}|_X$ is a finite set of functions ($\left|\mathcal{H}|_X\right| \leq \tau_{\mathcal{H}}(2m)$ because $|X| \leq 2m$), we can apply the union bound, and this yields the inequality.

To make this argument formal, there are two issues we need to consider. The first issue is that in order to apply a union bound using the fact that $\mathcal{H}|_X$ is finite, we need the set X to be fixed (if the class $\mathcal{H}|_X$ is itself a random variable, we cannot apply a union bound). We solve this by generating our samples via a two-step process: (1) a vector Z of $2m$ i.i.d. samples is chosen from $\mathcal{D}$; (2) Z is partitioned into two vectors S and S' of length m. Thus, for each fixed value $X = Z_x = S_x \cup S'_x$, the set $\mathcal{H}|_X$ is finite and fixed, and we can apply a union bound separately for each value of Z (using the law of total probability).

The second issue is that for each $h \in \mathcal{H}|_X$, we want to use Hoeffding's inequality to bound the probability that $|L_S(h) - L_{S'}(h)|$ is large. To do so, we need to present this quantity as an average of independent random variables $Q_i = \ell(h, z_i) - \ell(h, z'_i)$. To ensure that $Q_1, \ldots, Q_m$ are independent, we specify that in the two-step process above, Z is partitioned into S and S' in a specific manner as follows. Denote $Z = (a_1, \ldots, a_m, b_1, \ldots, b_m)$. For each $i \in [m]$, with probability $\frac{1}{2}$, we set $z_i = a_i$ and $z'_i = b_i$, and with probability $\frac{1}{2}$ we make the opposite assignment, namely $z_i = b_i$ and $z'_i = a_i$. Thus, for each fixed value Z, the variables Q_i are independent, and furthermore $\mathbb{E}\left[Q_i\right] = 0$ for all $i \in [m]$.[3] Observe that sampling (S, S') using this two-step process produces the same joint distribution as sampling $S \sim \mathcal{D}^m$ and $S' \sim \mathcal{D}^m$ independently. This technique of "mixing" or "swapping" the samples between S and S' is known as *symmetrization*.

[2] In fact, using Chebyshev's inequality would suffice here.

[3] In contrast, notice that if for example we chose instead to assign the members of Z to S and S' by choosing an assignment uniformly from the set of $2m!$ possible assignments, then the Q_i's would not be independent.

Putting this all together,

$$\begin{aligned}
\mathbb{P}_{S,S'}[E_2] &= \mathbb{P}[\exists h \in \mathcal{H} :\ \Delta_S(h) > \varepsilon \ \wedge\ \Delta_{S'}(h) \le \varepsilon/2] \\
&\le \mathbb{P}[\exists h \in \mathcal{H} :\ |L_S(h) - L_{S'}(h)| \ge \varepsilon/2] && \text{(inverse triangle inequality)} \\
&= \mathbb{P}[\exists h \in \mathcal{H}|_X :\ |L_S(h) - L_{S'}(h)| \ge \varepsilon/2] && \text{(event depends only on the projection of } h \text{ to } X = S_x \cup S'_x) \\
&= \mathop{\mathbb{E}}_{Z}\left[\mathbb{P}_{S,S'}\left[\exists h \in \mathcal{H}|_X :\ |L_S(h) - L_{S'}(h)| \ge \varepsilon/2 \;\middle|\; Z\right]\right] && \text{(law of total probability)}
\end{aligned}$$

For any fixed value of Z,

$$\begin{aligned}
&\mathbb{P}_{S,S'}\left[\exists h \in \mathcal{H}|_X :\ |L_S(h) - L_{S'}(h)| \ge \varepsilon/2 \;\middle|\; Z\right] \\
&\le \sum_{h \in \mathcal{H}|_X} \mathbb{P}\left[|L_S(h) - L_{S'}(h)| \ge \varepsilon/2 \;\middle|\; Z\right] && \text{(union bound, } \mathcal{H}|_X \text{ is finite)} \\
&= \sum_{h \in \mathcal{H}|_X} \mathbb{P}\left[\left|\frac{1}{m}\sum_{i \in m} \ell(h, z_i) - \ell(h, z'_i)\right| \ge \varepsilon/2 \;\middle|\; Z\right] \\
&= \sum_{h \in \mathcal{H}|_X} \mathbb{P}\left[\left|\frac{1}{m}\sum_{i \in m} Q_i\right| \ge \varepsilon/2 \;\middle|\; Z\right] && (Q_i = \ell(h, z_i) - \ell(h, z'_i)) \\
&= \sum_{h \in \mathcal{H}|_X} 2\exp\left(-\frac{\varepsilon^2 m}{8}\right) && \text{(Hoeffding's: } Q_i \text{ are independent,} \\
& && \quad \mathbb{E}[Q_i] = 0,\ Q_i \in [-1, 1]) \\
&\le \tau_{\mathcal{H}}(2m) \cdot 2\exp\left(-\frac{\varepsilon^2 m}{8}\right). && (\left|\mathcal{H}|_X\right| \le \tau_{\mathcal{H}}(2m))
\end{aligned}$$

This establishes Claim II. Combining the two claims yields $\mathbb{P}[E_1] \le 4\tau_{\mathcal{H}}(2m)\exp\left(-\frac{\varepsilon^2 m}{8}\right)$, completing the proof of Item (a).

To prove Item (b), the main idea is that by Sauer's lemma $\tau_{\mathcal{H}}(2m) \le \left(\frac{2em}{d}\right)^d \le (2em)^d$, so the number of possible projections $h|_X$ only grows polynomially in m, while Hoeffding's inequality above showed that the probability of the bad event for a particular $h|_X$ vanishes exponentially in m. Namely,

$$\mathbb{P}[E_1] \le 4\left(\frac{2em}{d}\right)^d \exp\left(-\frac{\varepsilon^2 m}{8}\right). \tag{10}$$

Seeing as the exponential factor dominates the polynomial factor, $\mathbb{P}[E_1] \xrightarrow{m\to\infty} 0$. Numerically, we need to show that taking $m(d, \varepsilon, \delta)$ as in the statement suffices to ensure that $\mathbb{P}[E_1] \le \delta$.

Claim III: For $m(d, \varepsilon, \delta)$ as in the statement, $\mathbb{P}[E_1] \le \delta$.

It suffices to show that for m as in the statement, the expression in Eq. (10) is upper

bounded by δ. Rearranging this requirement yields:

$$\begin{aligned} 4\left(\frac{2em}{d}\right)^d \exp\left(-\frac{\varepsilon^2 m}{8}\right) \leq \delta \quad &\iff \quad d\ln\left(\frac{2em}{d}\right) - \frac{\varepsilon^2 m}{8} \leq \ln\left(\frac{\delta}{4}\right) \\ &\iff \quad m \geq \frac{8}{\varepsilon^2}\left(d\ln\left(\frac{2em}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right) \\ &\iff \quad m \geq \frac{8d}{\varepsilon^2}\ln(m) + \frac{8}{\varepsilon^2}\left(d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right) \\ &\iff \quad m \geq a\ln(m) + b, \end{aligned} \tag{11}$$

where $a = \frac{8d}{\varepsilon^2}$ and $b = \frac{8}{\varepsilon^2}\left(d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right)$. Consider two cases:

- Case I: $b \leq 0$. To complete the proof it suffices to show that $m \geq a\ln(m)$. By Claim 16 below, a sufficient condition for this to hold is that

  $$m \geq 2a\ln(a) = \frac{16d}{\varepsilon^2}\ln\left(\frac{8d}{\varepsilon^2}\right),$$

  which is satisfied for $m \geq m(d,\varepsilon,\delta)$ as in the statement.
- Case II: $b > 0$. By Claim 19 below, a sufficient condition for Eq. (11) to hold is that

  $$m \geq 4a\ln(2a) + 2b.$$

  Furthermore, because $b > 0$, it suffices to take m such that

  $$\begin{aligned} m &\geq 4a\ln(2a) + 4b \\ &= \frac{32d}{\varepsilon^2}\ln\left(\frac{16d}{\varepsilon^2}\right) + \frac{32}{\varepsilon^2}\left(d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right) \\ &= \frac{32d}{\varepsilon^2}\ln\left(\frac{16d}{\varepsilon^2}\cdot\frac{2e}{d}\right) + \frac{32}{\varepsilon^2}\ln\left(\frac{4}{\delta}\right) \\ &= \frac{32d}{\varepsilon^2}\ln\left(\frac{32e}{\varepsilon^2}\right) + \frac{32}{\varepsilon^2}\ln\left(\frac{4}{\delta}\right), \end{aligned}$$

  which is satisfied for $m \geq m(d,\varepsilon,\delta)$ as in the statement. ◀

▶ Remark 11.

- The constants for the bound that appear in the proof are not tight.
- For the case $b > 0$, we showed a stronger bound of $m(d,\varepsilon,\delta) = O\left(\frac{d\ln(1/\varepsilon)+\ln(1/\delta)}{\varepsilon^2}\right)$.

▶ **Exercise 12.** *Prove Theorem 8.*

4 The Fundamental Theorem of PAC Learning

We have shown that classes with finite VC dimension have the uniform convergence property, and so they are agnostic PAC learnable (and therefore PAC learnable). Together with the lower bound of Claim 15 from Unit 4, this yields the following important characterization of learnability.

▶ **Theorem 13** (Fundamental Theorem of PAC Learning – Qualitative Version)**.** *Let $\mathcal{X}$ be a nonempty set, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$. The following conditions are equivalent:*

1. $\mathsf{VC}(\mathcal{H}) < \infty$.

2. *$\mathcal{H}$ has the uniform convergence property.*
3. *Every* $\mathrm{ERM}_{\mathcal{H}}$ *algorithm is an agnostic PAC learner for* $\mathcal{H}$.
4. *$\mathcal{H}$ is agnostic PAC learnable.*
5. *Every* $\mathrm{ERM}_{\mathcal{H}}$ *algorithm is a PAC learner for* $\mathcal{H}$.
6. *$\mathcal{H}$ is PAC learnable.*

Proof of Theorem 13. We show the following implications (see Figure 2):

- $1 \Rightarrow 2$. This follows from Theorem 9, together with Remark 7.
- $2 \Rightarrow 3$. This follows from Lemma 11 in Unit 4.
- $3 \Rightarrow 4 \Rightarrow 6$ and $3 \Rightarrow 5 \Rightarrow 6$. These implications are immediate from the definitions of PAC and agnostic PAC learning.
- $6 \Rightarrow 1$. This is the contrapositive of Claim 15 in Unit 4. ◀

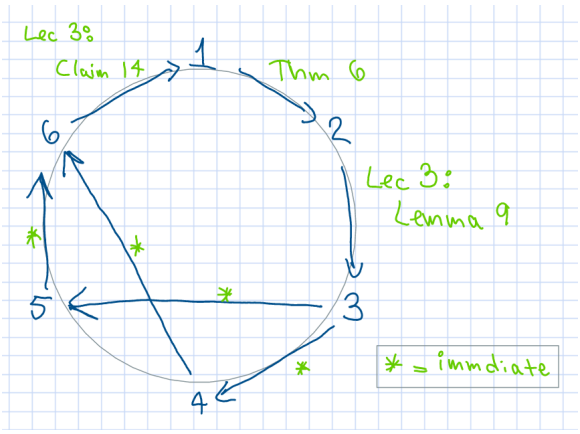

Figure 2 Proof outline for Theorem 13.

Furthermore, it is possible to give quantitative bounds on the sample complexity for classes with finite VC dimension.

▶ **Theorem 14** (Fundamental Theorem of PAC Learning – Quantitative Version)**.** *Let $\mathcal{X}$ be a nonempty set, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$, and let $d = \mathsf{VC}(\mathcal{H})$. Assume $d = \mathsf{VC}(\mathcal{H}) < \infty$. Then there exist constants $c_0, c_1, c_2 > 0$ such that:*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$c_0 \cdot \frac{d + \ln(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta) \leq c_1 \cdot \frac{d + \ln(1/\delta)}{\varepsilon^2}.$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$c_0 \cdot \frac{d + \ln(1/\delta)}{\varepsilon^2} \leq m(\varepsilon, \delta) \leq c_1 \cdot \frac{d + \ln(1/\delta)}{\varepsilon^2}.$$

3. *$\mathcal{H}$ is PAC learnable with sample complexity*

$$c_0 \cdot \frac{d + \ln(1/\delta)}{\varepsilon} \leq m(\varepsilon, \delta) \leq c_1 \cdot \frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon} + c_2.$$

The upper bounds in Theorem 14 for the the realizable and agnostic cases are related to Theorems 8 and 9 respectively. More specifically, Item 1 is similar to Theorem 9, and Item 2 follows from Item 1 because uniform convergence implies agnostic PAC learnability

(by Lemma 11 in Unit 4). Similarly, for the realizable case, if the sample is an ε-net with probability $1 - \delta$ then any $\mathrm{ERM}_{\mathcal{H}}$ algorithm is a PAC learner, and therefore Item 3 follows from Theorem 8.

However, employing Theorem 9 to prove Items 1 and 2 in the manner just discussed yields upper bounds of

$$c \cdot \frac{d \ln (d/\varepsilon) + \ln (1/\delta)}{\varepsilon^2},$$

which is not tight. This expression is a factor of $\ln (d/\varepsilon)$ larger than in the statement of Theorem 14. Usually this would not make a big difference in practice, but nonetheless, in upcoming units we will present an analysis using Rademacher complexity and covering numbers, which provide a different perspective on this theorem and yield the tighter bounds.

For the lower bounds in Theorem 14, we have already seen that if ε and δ are constants that are less than $1/8$ then Claim 15 in Unit 4 implies a lower bound of $\Omega(d)$ on the sample complexity. However, we will not prove the dependence of the lower bound on ε and δ that is stated in the theorem. A proof is available in [10, Chapter 28]

5 Discussion

This unit tied together many of the notions we have seen so far in the course, creating one unified theory of PAC learning. In a sentence, this can be summarized as follows: The VC dimension determines the general outline of the growth function, which in turn determines whether a class satisfies uniform convergence, which is equivalent to agnostic PAC learning, which implies PAC learning, which (by the no free lunch theorems) implies a finite VC dimension.

In the next two units we will see a different perspective on these issues, employing Rademacher complexity, covering numbers, and chaining.

5.1 Bibliographic Notes

Theorem 8 is due to Haussler and Welzl [3]. See also [4, 8] for related results and a tighter version of the theorem. Theorem 9 is due to Vapnik and Chervonenkis [12]. Sauer's Lemma (Lemma 2) was proved independently by Sauer [9] and by Shelah [11] (who gives credit to Micha Perles). A slightly weaker version of the lemma was proved shortly before that by Vapnik and Chervonenkis [12]. An alternative proof using linear algebra was given by Peter Frankl and Janos Pach [7].

A good expositions of the material in this unit is available in [10, Chapters 6 and 28], as well as in [2, Chapters 7–8], [5, Chapter 3], [13, Chapter 1], [1, Section 13.4], and [6, Chapter 3].

References

1 Noga Alon and Joel H. Spencer. *The Probabilistic Method, Second Edition*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2000.

2 Bruce Hajek and Maxim Raginsky. *ECE 543: Statistical Learning Theory*. University of Illinois at Urbana-Champaign, 2018. URL: `https://web.archive.org/web/20210213043003/http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf`.

3 David Haussler and Emo Welzl. epsilon-nets and simplex range queries. *Discret. Comput. Geom.*, 2:127–151, 1987. `doi:10.1007/BF02187876`.

4 János Komlós, János Pach, and Gerhard J. Woeginger. Almost tight bounds for epsilon-nets. *Discret. Comput. Geom.*, 7:163–173, 1992. doi:10.1007/BF02187833.

5 Percy Liang. *CS229T/STAT231: Statistical Learning Theory*. Stanford University, 2016. URL: https://web.archive.org/web/20210114222356/https://web.stanford.edu/class/cs229t/notes.pdf.

6 Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.

7 János Pach and Pankaj K. Agarwal. *Combinatorial geometry*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1995.

8 János Pach and Gerhard J. Woeginger. Some new bounds for epsilon-nets. In Raimund Seidel, editor, *Proceedings of the Sixth Annual Symposium on Computational Geometry, Berkeley, CA, USA, June 6-8, 1990*, pages 10–15. ACM, 1990. doi:10.1145/98524.98529.

9 Norbert Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13(1):145–147, 1972. doi:10.1016/0097-3165(72)90019-2.

10 Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning – From Theory to Algorithms*. Cambridge University Press, 2014. URL: http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

11 Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

12 Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi:https://doi.org/10.1137/1116025.

13 Michael M. Wolf. Mathematical foundations of supervised learning, 2020. URL: https://web.archive.org/web/20210114220017/https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2020S/ML_notes_main.pdf.

A Useful Inequalities for Logarithms

▷ Claim 15. For all $x > 0$: $x > 2\ln(x)$.

Proof. Let $f(x) = x - 2\ln(x)$. The claim holds because f is positive for all $x > 0$. To see this, observe that f has a single global minimum at $x = 2$, because $f'(x) = 1 - \frac{2}{x}$, which is negative for $x \in (0, 2)$ and positive for $x \in (2, \infty)$, and note that $f(2) = 2(1 - \ln(2)) > 0$. ◀

▷ Claim 16. For all $x, a > 0$: $x \geq 2a\ln(a) \implies x \geq a\ln(x)$.

Proof. Let $f(x) = x - a\ln(x)$. We need to show that $x \geq 2a\ln(a) \implies f(x) \geq 0$.

First, observe that f has a single global minimum at $x = a$, because $f'(x) = 1 - \frac{a}{x}$, which is negative for $x \in (0, a)$ and positive for $x \in (a, \infty)$.

Second, note that if $a \in (0, e)$ then $f(x) \geq 0$ for all $x > 0$. Indeed, because f has a minimum at $x = a$, it suffices to show that $f(a) \geq 0$ whenever $a \in (0, e)$:

$$f(a) \geq 0 \iff a(1 - \ln(a)) \geq 0 \iff \ln(a) \leq 1 \iff a \in (0, e).$$

Thus, it suffices to prove the claim for $a \geq e$.

Finally, assume $a \geq e$ and $x \geq 2a\ln(a)$, we show that $f(x) \geq 0$. Seeing as $2a\ln(a) \geq a$ and $f'(x)$ is positive in (a, ∞), it suffices to show that $f(2a\ln(a)) \geq 0$. Indeed,

$$\begin{aligned} f(2a\ln(a)) \geq 0 &\iff 2a\ln(a) \geq a\ln(2a\ln(a)) \\ &\iff a^2 \geq 2a\ln(a) \\ &\iff a \geq 2\ln(a). \end{aligned}$$

The last inequality holds for all $a > 0$ by Claim 15. ◀

By contrapositive, this implies the following corollary.

▷ Claim 17. For all $x, a > 0$: $x < a\ln(x) \implies x < 2a\ln(a)$.

The following claim is a very similar, with a different proof.

▷ Claim 18. For all $x, a > 0$: $x \leq a\ln(x) \implies x \leq \frac{1}{e-1} \cdot ea\ln(ea)$. Furthermore, if the first inequality is strict, then so is the second one.

Proof. From monotonicity and concavity of the logarithm function,

$$\ln(x) \leq \ln(x + ea) \leq \ln(ea) + \ln'(ea)x = \ln(ea) + \frac{x}{ea}.$$

Plugging this into the assumption yields

$$x \leq a\ln(x) \leq a\ln(ea) + \frac{x}{e},$$

and this implies

$$x \leq \frac{1}{1 - \frac{1}{e}} \cdot a\ln(ea) = \frac{1}{e-1} \cdot ea\ln(ea).$$

If the first inequality in the statement is strict, then so is this final inequality. ◀

▷ Claim 19. For all $a \geq 1$ and $b, x > 0$: $x \geq 4a\ln(2a) + 2b \implies x \geq a\ln(x) + b$.

Proof. Assume that $x \geq 4a\ln(2a) + 2b$. Because $a \geq 1$:

$$x \geq 4a\ln(2a) + 2b \geq 2b. \tag{12}$$

From Claim 16 and the fact that $b > 0$:

$$x \geq 4a\ln(2a) + 2b \geq 4a\ln(2a) \implies x \geq 2a\ln(x). \tag{13}$$

Combining Eq. (12) and (13),

$$x \geq 2 \cdot \max\{a\ln(x), b\} \geq a\ln(x) + b.$$ ◀