

The Fundamental Theorem of PAC Learning

Deepak Rajendraprasad
IIT Palakkad

March 26, 2021

1 PAC Learning

Definition 1 (Hypothesis Class). A *hypothesis class* H over a domain X is a collection of binary functions over X . That is, $H \subseteq \{h : X \rightarrow \{+1, -1\}\}$. Since a binary function h can be uniquely represented by the set $h^{-1}(+1)$, H can also be represented as a collection of subsets of X .

Definition 2 (PAC learnable). A hypothesis class H over a domain X is called *PAC learnable* if, for every $\epsilon, \delta > 0$, there is an $n = n(H, \epsilon, \delta)$ such that for every binary function f on X and every probability distribution D on X ,

$$P_{S \sim D^n} [\exists h \in H : (E_S(h, f) = 0) \wedge (E_D(h, f) > \epsilon)] \leq \delta. \quad (1)$$

If so, then the function $t_H : (0, 1)^2 \rightarrow \mathbb{N}$, where $t_H(\epsilon, \delta)$ is the smallest n which satisfies Inequality 1 is called the *sample complexity* of H .

In Definition 2, ϵ is called the *accuracy parameter*, δ the *confidence parameter*, f the *true labelling*, and D is called the *sampling distribution*. The *error region* $h\Delta f = \{x \in X : h(x) \neq f(x)\}$ and $E_S(h, f) = |H \cap (h\Delta f)| / |H|$ is the *in-sample error* or *training error* while $E_D(h, f) = P_D(h\Delta f)$ is the *out-of-sample error* or *true error*. In Equation 1, $S \sim D^n$ denotes the random experiment of sampling n points from X independently according to the distribution D . The *event highlighted in red* is an event associated with this random experiment and hence *is a collection of n -sized subsets of X* . Let us call an n -sized subset S of X *potentially misleading* if a hypothesis h with true error more than ϵ can perfectly match with f inside of S . That is, S is potentially misleading if $\exists h \in H : (E_S(h, f) = 0) \wedge (E_D(h, f) > \epsilon)$. Hence “ $\exists h \in H : (E_S(h, f) = 0) \wedge (E_D(h, f) > \epsilon)$ ” is nothing but the collection of potentially misleading training sets. The same event can also be expressed as $\bigcup_{h \in H_\epsilon} A_h$ where $H_\epsilon = \{h \in H : E_D(h, f) > \epsilon\}$ is the set of hypothesis with a true error larger than ϵ and $A_h = \{S \in X^n : E_S(h, f) = 0\}$ is the collection of training data sets which results in zero in-sample error for h .

It is important to note that while the sample complexity is allowed to depend on the accuracy and confidence expected, it has to be independent of f and D . From an application engineer's perspective, she chooses ϵ and δ and hence those are known to her. The independence on f and D are blessings to her. Note that f is unknown to the engineer and estimating D may be a laborious and error-prone statistical exercise. The dependence of the sample complexity on H itself is something we wish could be avoided. But it turns out that such a “universal sample complexity” does not exist. In fact, we will prove that there are many hypothesis classes which are not PAC learnable. There is another big blessing for the engineer that is not so evident. Since we guarantee that with high probability *every* 0-error hypothesis on the training data

has true error at most ϵ , she is free to choose any 0-error separator for the training data. Hence as far as the error guarantees are considered, it doesn't matter which algorithm is used to find the separator on the training data. So she can pick her favourite algorithm.

There are two minor variants of PAC learnability. If we can establish Equation 1 only for true labellings f which are contained in H , then it is called PAC learnability under *realisability assumption*. If the true labelling f can be any binary function on X (as in Definition 2), then it is called *agnostically* PAC learnable. In this note, without either of the adjectives, PAC learnability will refer to the agnostic variant. *Theist PAC* would have been a more appropriate name for the weaker variant ;).

Definition 3 (VC-dimension). A hypothesis class H over a domain X is said to *shatter* a set $S \subset X$ if every binary function over S can be obtained as $h|_S$ (h restricted to S) for an $h \in H$. The *VC-dimension* of H is the size of a largest set that is shattered by H .

Many interesting hypothesis classes have finite VC-dimension. Half-spaces in \mathbb{R}^d has VC-dimension $d + 1$, intervals in \mathbb{R} has VC-dimension 2 and axis-aligned rectangles in \mathbb{R}^2 has VC-dimension 4. On the other hand, convex polygons in \mathbb{R}^2 have infinite VC-dimension. Proving all of these, except the case of half spaces in higher dimensions, are cute geometric exercises. The proof that the no set of $d + 2$ points in \mathbb{R}^d can be shattered by half spaces is an interesting application of Radon's theorem.

Theorem 4 (**The Fundamental Theorem of PAC Learning - Part I**). *A hypothesis class is PAC learnable if and only if its VC-dimension is finite. Moreover, Equation 1 is satisfied whenever*

$$n \geq \frac{4}{\epsilon} \left(2d \log_2 \frac{8e}{\epsilon} + \log_2 \frac{2}{\delta} \right), \quad (2)$$

where d is the VC-dimension of the hypothesis class.

Proof sketch. The “if” part will be established in two steps. In the first step, we show that if the VC-dimension of a hypothesis class is some finite number d , then its growth function g_H (Definition 7) is $O(n^d)$ (at most polynomial). This is called Sauer's Lemma. Secondly we show that the probability of sampling a misleading set of size n is at most $2g_H(2n)2^{-\epsilon n/4}$. Since a polynomial in n (g_H) is multiplied by a negative exponential in n , for large enough n , the product will fall below δ as needed. A careful substitution will establish the quantitative guarantees. The “only if” part TODO.

Before we begin the proof of the 2-step “if” part, we do a warm up exercise with finite hypothesis classes.

1.1 Warm up. Finite hypothesis classes are PAC learnable

Definition 5 (PAC instance). For a hypothesis class H over a domain X , a *PAC instance* for H is a 4-tuple (f, D, ϵ, δ) , where f is a binary function over X ; D a probability distribution over X ; and $\epsilon, \delta \in (0, 1)$. They are called as the true labelling, the sampling distribution, the accuracy parameter, and the confidence parameter, respectively.

Theorem 6. *Every finite hypothesis class H is PAC learnable. Moreover, Equation 1 is satisfied whenever*

$$n \geq \frac{1}{\epsilon} \left(\ln |H| + \ln \frac{1}{\delta} \right). \quad (3)$$

Proof. Let X be the domain of H and (f, D, ϵ, δ) be an arbitrary PAC instance for H . Pick any n satisfying Inequality 3. Let $H_\epsilon = \{h \in H : E_D(h, f) > \epsilon\}$ denote the subset of hypotheses which have true error more than ϵ .

Now consider the random experiment of picking a set S of n samples from X independently according to D (which we denote as $S \sim D^n$). For each hypothesis $h \in H$, we define the event

$$\begin{aligned} A_h &= \{S : E_S(h, f) = 0\} \\ &= \{S : S \cap (h\Delta f) = \emptyset\} \end{aligned}$$

That is, A_h consists of all those n -element subsets of S which are perfectly labelled (separated) by the hypothesis h . We argue that it is quite unlikely that the sampled S can be perfectly labelled by a hypothesis with large true error. More precisely, we want to show

$$P_{S \sim D^n} \left[\bigcup_{h \in H_\epsilon} A_h \right] \leq \delta. \quad (4)$$

Notice that Equation 4 is same as Equation 1. For a hypothesis $h \in H$, the event A_h occurs when S contains no point from the region of disagreement $h\Delta f$. Hence

$$\begin{aligned} P_{S \sim D^n}[A_h] &= (1 - P_D(h\Delta f))^n \\ &= (1 - E_D(h, f))^n. \end{aligned}$$

Union bound then gives

$$\begin{aligned} P_{S \sim D^n} \left[\bigcup_{h \in H_\epsilon} A_h \right] &\leq \sum_{h \in H_\epsilon} (1 - E_D(h, f))^n \\ &< |H_\epsilon| (1 - \epsilon)^n && \forall h \in H_\epsilon, E_D(h, f) > \epsilon \\ &\leq |H_\epsilon| e^{-\epsilon n} && \forall x \in \mathbb{R}, 1 + x \leq e^x \\ &\leq |H| e^{-\epsilon n} && H_\epsilon \subseteq H \\ &\leq \delta && n \geq \frac{1}{\epsilon} \left(\ln |H| + \ln \frac{1}{\delta} \right). \end{aligned}$$

Hence H is PAC learnable with sample complexity $\frac{1}{\epsilon} (\ln |H| + \ln \frac{1}{\delta})$. \square

1.2 Step 1/2. Hypothesis classes with finite VC-dimension have polynomially bounded growth functions

Definition 7 (Growth Function). The *growth function* $g_H : \mathbb{N} \rightarrow \mathbb{N}$ of H is defined by $g_H(n) = \max |\{h|_S : h \in H\}|$, where the maximisation is over all n -sized subsets S of X .

Notice that if the VC-dimension of a hypothesis class is d , then for every $n \leq d$, $g_H(n) = 2^n$, while for every $n > d$, $g_H(n) < 2^n$. Hence VC-dimension of a hypothesis class H can also be defined as the largest n for which $g_H(n) = 2^n$.

Lemma 8 (Sauer's Lemma). *If a hypothesis class H has a finite VC-dimension d , then*

$$g_H(n) \leq \sum_{i=0}^d \binom{n}{i} = \binom{n}{\leq d}. \quad (5)$$

Proof. Let $g(d, n) = \max g_H(n)$, where the maximisation is over all hypothesis classes H with VC-dimension d . We will show by a double induction argument that $g(d, n) \leq \binom{n}{\leq d}$. The base cases are easy but interesting to verify. If H contains at least two functions h_1 and h_2 , then the singleton set $\{x\}$, where x is a point in X where f_1 and f_2 differ, is shattered by

H . Hence any hypothesis class with VC-dimension 0 contains at most one function and hence $g(0, n) \leq 1 = \binom{n}{\leq 0}$ for all $n \in \mathbb{N}$. Also since $g(d, n) \leq 2^n$, we see that $\forall d \geq 1, g(d, 1) \leq 2 = \binom{1}{\leq d}$. Let's fix some $d \geq 1$ and $n \geq 2$ and assume that the bound is true for $(d-1, n-1)$ and $(d, n-1)$.

Let H be any hypothesis class over a domain X of VC-dimension d . For $S \subset X$, let $H|_S = \{h|_S : h \in H\}$ and let $g_H(S) = |H|_S|$. Let $S_n = \{x_1, \dots, x_n\} \subset X$ be a set such that $g_H(S_n) = \max\{g_H(S) : S \in \binom{X}{n}\} = g_H(n)$. Two functions $h_1, h_2 \in H|_{S_n}$ is said to form a *pair* if $\forall i \in [n-1], h_1(x_i) = h_2(x_i)$ but $h_1(x_n) \neq h_2(x_n)$. A function $h \in H|_{S_n}$ without a pair in $H|_{S_n}$ is called *single*. Let α be the number of singles and β be the number of pairs in $H|_{S_n}$ so that $|H|_{S_n}| = \alpha + 2\beta$.

Let $S_{n-1} = S_n \setminus \{x_n\}$. Since every pair in $H|_{S_n}$ collapses to the same function when restricted to S_{n-1} , we can see that $|H|_{S_{n-1}}| = \alpha + \beta$. Hence, $\alpha + \beta \leq g(d, n-1)$. Let H' denote the subset of $H|_{S_{n-1}}$ obtained by the collapses of pairs in $H|_{S_n}$. If we consider H' as a hypothesis class over S_{n-1} , then its VC-dimension is at most $d-1$. This is because if any d sized subset T of S_{n-1} is shattered by H' , then the $(d+1)$ -sized set $T \cup \{x_n\}$ is shattered by $H|_{S_n}$ and hence H , contradicting the assumption that VC-dimension of H is d . Hence $\beta = |H'| \leq g(d-1, n-1)$.

Putting it all together, we get $g_H(n) \leq \alpha + 2\beta \leq g(d, n-1) + g(d-1, n-1)$. Since H was an arbitrary hypothesis class of VC-dimension d , we conclude $g(d, n) \leq g(d, n-1) + g(d-1, n-1)$. The lemma now follows from the easy combinatorial identity $\binom{n-1}{\leq d} + \binom{n-1}{\leq d-1} = \binom{n}{\leq d}$. \square

Lemma 8 was discovered thrice in three different contexts. By Norbert Sauer in 1972 while studying the combinatorics of set systems. By Saharon Shelah, also in 1972, while studying model theory. By the Russian Mathematicians Vladimir Vapnik and Alexey Chervonenkis in 1971 while studying statistics. The diversity of these contexts is taken as an indication of the fundamental nature of the lemma. A much shorter proof of the lemma can be obtained using Pajor's Lemma stated below. The proof of Pajor's lemma and the proof that it implies Sauer's lemma are left as two challenging exercises.

Lemma 9 (Pajor's Lemma). *Any hypothesis class H shatters at least $|H|$ different sets.*

1.3 Step 2/2. Hypothesis classes with finite VC-dimension are PAC learnable

Lemma 10 (Multiplicative Chernoff Bound). *Let $Z = Z_1 + \dots + Z_n$ be a sum of n independent Bernoulli random variables. Then*

$$\begin{aligned} P(Z \leq (1 - \delta)\mu) &\leq e^{-\frac{\delta^2 \mu}{2}}, & 0 \leq \delta \leq 1, \\ P(Z \geq (1 + \delta)\mu) &\leq e^{-\frac{\delta^2 \mu}{2 + \delta}}, & 0 \leq \delta, \end{aligned}$$

where $\mu = EZ$. In particular,

$$\begin{aligned} P\left(Z \leq \frac{1}{2}\mu\right) &\leq e^{-\frac{1}{8}\mu} \\ &< \frac{1}{2}, & \text{if } \mu \geq 8. \end{aligned} \tag{6}$$

Lemma 11. *Let D be any probability distribution over a set X and $\epsilon \in (0, 1)$. Let Y be n points, $n \geq 8/\epsilon$, sampled from X according to D , i.e., $Y \sim D^n$. For any subset $R \subset X$ with $P_D(R) \geq \epsilon$, with probability at least $1/2$, Y contains more than $\epsilon n/2$ points from R .*

Proof. This is a direct application of Inequality 6. Take Z_i as the indicator random variable for the event that the i -th sample in Y is from R . Hence $Z = |Y \cap R|$ and $\mu = EZ = nP_D(R) \geq \epsilon n$. \square

Lemma 12 (“Smartest Idea”). Let H be any hypothesis class over a domain X . Let (f, D, ϵ, δ) be a PAC instance for H and $n \geq 8/\epsilon$. Then

$$P_{S \sim D^n} [\exists h \in H : (E_S(h, f) = 0) \wedge (E_D(h, f) > \epsilon)] \leq 2P_{S, S' \sim D^n} [\exists h \in H : (E_S(h, f) = 0) \wedge (E_{S'}(h, f) > \epsilon/2)].$$

Proof. Let A be the **first** and B be the **second** event. Event A occurs if there is a hypothesis $h \in H$ with zero in-sample error in S and more than ϵ true error. Pick such an h and let $R = h\Delta f$. By Lemma 11, S' has at least half chance of containing more than $\epsilon n/2$ points from R . Hence with probability at least half, $E_{S'}(h, f) > \epsilon/2$. That is $P[B/A] \geq 1/2$. Hence $P[B] \geq P[B \cap A] = P[A]P[B/A] \geq P[A]/2$ as claimed. \square

Why do I call Lemma 12 the smartest? Consider the random experiment of sampling two sets S and S' independently according to D^n . For each hypothesis $h \in H$, we define the event

$$B_h = \{S, S' : (E_S(h, f) = 0) \wedge (E_{S'}(h, f) > \epsilon/2)\}.$$

The key observation is that for any two hypothesis h_1 and h_2 which result in the same restriction on the set $S \cup S'$, the events B_{h_1} and B_{h_2} are the same. Hence the seemingly infinite family $\{B_h : h \in H\}$ of events actually has only $|H_{|S \cup S'}| \leq g_H(2n)$ events. And hence we can mimic the proof of Theorem 6 to prove our main result (Theorem 4).

Firstly, notice that the random experiment of sampling n -points independently from D is the same as sampling m points ($m \geq n$) independently from D and then choosing any n of these m as long as the choice of the n points out of m is not dependent on the samples. This choice of n points out of m may be random or deterministic, but should be independent of the samples chosen. Hence choosing two sets S and S' independently according to D^n is equivalent to the experiment of sampling a set T according to D^{2n} and then partitioning it into two equal sized subsets S and S' .

In our case, we will do the partitioning of T into S and S' uniformly at random so that all the $\binom{2n}{n}$ n -sized subsets of T has equal chance of being chosen as S . Then, for any $h \in H$, the event B_h occurs if T contains at least $\epsilon n/2$ points from $h\Delta f$ and all of them end up in S' during a uniform equipartition of T . Let $k = |T \cap (h\Delta f)|$. Then

$$\begin{aligned} P_{S, S' \sim D^n}[B_h] &= \binom{2n-k}{n} / \binom{2n}{n} \\ &= \frac{n(n-1) \cdots (n-k+1)}{2n(2n-1) \cdots (2n-k+1)} \\ &\leq (1/2)^k \\ &\leq (1/2)^{\epsilon n/2} \end{aligned}$$

Since there are only $|H_{|S \cup S'}| \leq g_H(2n)$ distinct events among $\{B_h : h \in H\}$, by union bound

$$\begin{aligned} P_{S, S' \sim D^n} [\exists h \in H : (E_S(h, f) = 0) \wedge (E_{S'}(h, f) > \epsilon/2)] &= P_{S, S' \sim D^n} [\exists h \in H : B_h] \\ &\leq g_H(2n) 2^{-\epsilon n/2}. \end{aligned} \tag{7}$$

Equation 7 and Lemma 12 gives us the following result.

Lemma 13. Let H be a hypothesis class with growth function g_H and let (f, D, ϵ, δ) be a PAC instance for H . Then Equation 1 is satisfied whenever the number n of training samples satisfies $n \geq 8/\epsilon$ and

$$g_H(2n) 2^{-\epsilon n/2} \leq \delta/2. \tag{8}$$

Since the growth function of any hypothesis class with a finite VC dimension is bounded by a polynomial (Lemma 8), it should be clear that Equation 8 will be eventually satisfied for a large enough n since no polynomial growth can asymptotically beat an exponential decay. The rest of this section is only to quantify the same. Please be warned that the steps are a bit cumbersome. You are only expected to verify it carefully once and then add it to your baggage of beliefs. Interestingly, while everyone familiar with order notation will agree that e^x will eventually be larger than x^k for any fixed k , it is not that commonly known as to when (from which x) will that happen!¹

Lemma 14 (Exponential beats polynomial). *For every $a \geq 1$, $b \geq 0$, if $x \geq 4a \ln(2a) + 2b$, then $x \geq a \ln x + b$. Moreover if $a \geq 2$, then $x \geq 4a \log_2(2a) + 2b$ implies $x \geq a \log_2(x) + b$.*

Proof. For small x , i.e, when $a \ln x \leq b$, the hypothesis $x \geq 4a \ln(2a) + 2b$ directly gives us $x \geq 2b \geq a \ln x + b$ as required. Hence we can assume that $a \ln x > b$. In this case, it suffices to prove that $x \geq 4a \ln(2a)$ implies $x \geq 2a \ln x$.

Consider the function $f : (0, \infty) \rightarrow \mathbb{R}$ defined by $f(x) = x - 2a \ln x$. Then $f'(x) = 1 - 2a/x$ which is positive for all $x > 2a$. Hence $f(x)$ is increasing in $(2a, \infty)$. Hence it suffices for our purposes to show that $f(4a \ln(2a)) \geq 0$.

$$\begin{aligned} f(4a \ln(2a)) &= 4a \ln(2a) - 2a \ln(4a \ln(2a)) \\ &> 4a \ln(2a) - 2a \ln(4a^2) & (\forall x \in (0, \infty), \ln x \leq x/e < x/2) \\ &= 0. \end{aligned}$$

One can prove that $\forall x \in (0, \infty), \ln x \leq x/e$ by drawing their graphs. The line $y = x/e$ will be the tangent to the curve $y = \ln x$ at $(e, 1)$.

The proof for the binary logarithm is similar. We will use the fact that $\log_2(x) \leq x/2$ for all $x \in [4, \infty)$. This fact can be proved by the monotonicity of the function $f(x) = x/2 - \log_2(x)$ for $x > 2 \log_2(e) \approx 2.9$ and the evaluation $f(4)$. \square

We now complete the proof of one direction in Theorem 4.

Lemma 15. *Let H be a hypothesis class with VC dimension d and let (f, D, ϵ, δ) be a PAC instance for H . Then if we pick*

$$n \geq \frac{8d}{\epsilon} \log_2 \frac{8e}{\epsilon} + \frac{4}{\epsilon} \log_2 \frac{2}{\delta} \quad (9)$$

samples, Equation 1 is satisfied.

Proof. By Lemma 8 $g_H(2n) \leq \binom{2n}{\leq d}$, which is at most $(2en/d)^d$ by standard binomial bounds. By Lemma 13, Inequality 1 is satisfied whenever $n \geq 8/\epsilon$ and $g_H(2n)2^{-\epsilon n/2} \leq \delta/2$. Inequality 9 already ensures the former. For the latter, it suffices to have $(2en/d)^d 2^{-\epsilon n/2} \leq \delta/2$.

$$\begin{aligned} (2en/d)^d 2^{-\epsilon n/2} \leq \delta/2 &\iff d \log \frac{2en}{d} - \frac{\epsilon n}{2} \leq \log \frac{\delta}{2} && \log \text{ denotes } \log_2 \\ &\iff n \geq \frac{2d}{\epsilon} \log \frac{2en}{d} + \frac{2}{\epsilon} \log \frac{2}{\delta} && \text{rearrangement} \\ &\iff \frac{2en}{d} \geq \frac{4e}{\epsilon} \log \frac{2en}{d} + \frac{4e}{\epsilon d} \log \frac{2}{\delta} && \text{scale with } \frac{2e}{d} \\ &\iff m \geq \frac{4e}{\epsilon} \log m + \frac{4e}{\epsilon d} \log \frac{2}{\delta} && m = \frac{2en}{d} \\ &\iff m \geq \frac{16e}{\epsilon} \log \frac{8e}{\epsilon} + \frac{8e}{\epsilon d} \log \frac{2}{\delta} && \text{Lemma 14} \\ &\iff n \geq \frac{8d}{\epsilon} \log \frac{8e}{\epsilon} + \frac{4}{\epsilon} \log \frac{2}{\delta} && \text{scale with } \frac{d}{2e} \end{aligned}$$

\square

¹The answer is between $k \ln k$ and $2k \ln k$; closer to the former than the latter.

1.4 Converse. Hypothesis classes with infinite VC-dimension are not PAC learnable