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

Unit 4

Agnostic PAC, Uniform Convergence and the VC Dimension

1 Introduction

In the previous lecture we presented the PAC definition of learning. We had three goals in mind: (i) learning with finite computational and sample complexity; (ii) modeling uncertainty or noise; and (iii) using absence of evidence as evidence of absence. We saw that at least for finite hypothesis classes, the PAC definition achieves the first goal. But the PAC definition does not achieve the second goal, because it makes the strong assumption that there exists a target function f such that the labels are a deterministic function of the instance, $y = f(x)$.

Furthermore, we saw that it is necessary also to assume that $f \in \mathcal{H}$ for some restricted class of function $\mathcal{H}$ (for instance, if $\mathcal{H}$ is finite then PAC learning is possible). This is problematic because even if there exists such a class $\mathcal{H}$, the nature of unknown systems is that we often won't know what $\mathcal{H}$ is.

Our goals for this lecture and the next are the following:

- Show how to relax the assumption that the target function f satisfies $f \in \mathcal{H}$ (while still guaranteeing that learning is possible with finite sample complexities).
- Show how to relax the assumption that a target function exists, i.e., that the label y is a deterministic function of x (while still maintaining finite sample complexity).
- Continue characterizing which hypothesis classes $\mathcal{H}$ are learnable and with what complexity.

2 The Agnostic PAC Definition

Agnostic PAC leaning is a relaxation of PAC learning in which we are *agnostic*, in the sense that we attempt to make as few assumptions as possible about the unknown system. The main idea is that, to enable us to weaken our assumptions, we will also weaken the requirements that the output hypothesis h chosen by the learner is required to meet. Instead of aiming for h to be good in an *absolute* sense, we will only aim only for h to be good in a *competitive* (relative) sense. We present this idea in two steps.

Step 1: Removing the assumption $f \in \mathcal{H}$. Recall that a successful PAC learner should output a hypothesis h such that with high probability, h is good in the absolute sense $L_{\mathcal{D},f}(h) \leq \varepsilon$. In contrast, consider the following competitive objective: with high probability, $\forall h' \in \mathcal{H} :\ L_{\mathcal{D},f}(h) \leq L_{\mathcal{D},f}(h') + \varepsilon$. Using the notation $L_{\mathcal{D},f}(\mathcal{H}) = \inf_{h' \in \mathcal{H}} L_{\mathcal{D},f}(h')$, this can also be rephrased as saying that with high probability,

$$L_{\mathcal{D},f}(h) \leq L_{\mathcal{D},f}(\mathcal{H}) + \varepsilon. \tag{1}$$

This means that h is almost as good as any hypothesis in $\mathcal{H}$. This is a conditional guarantee. We do not *assume* that $f \in \mathcal{H}$. Instead, we say that *if* $f \in \mathcal{H}$, then $L_{\mathcal{D},f}(h) \leq \varepsilon$; otherwise, *if*

f is α-close to $\mathcal{H}$ in the sense that $\exists h' \in \mathcal{H} :\ \mathbb{P}_{x\sim\mathcal{D}}[h'(x) \neq f(x)] \leq \alpha$, then $L_{\mathcal{D},f}(h) \leq \alpha + \varepsilon$; however, if f is far from $\mathcal{H}$ then nothing meaningful is guaranteed.

Step 2: Removing the assumption that labels are deterministic. Using the competitive objective, we can go further and eliminate the assumption that a target function f exists at all. Instead of saying that y is a deterministic function of x, namely $y = f(x)$ for some target function f, we can instead say that y is a random variable that depends on x. Namely, there is a conditional probability function $\mathbb{P}[y \mid x]$. Another way to say this, is that there exists a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathbb{P}[(x,y)] = \mathbb{P}[y \mid x]\,\mathbb{P}[x]$.

Because we are no longer assuming that there exists a target function, we will redefine the loss function as follows.

▶ **Definition 1.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$.*

(a) *Let $h :\ \mathcal{X} \to \mathcal{Y}$ be a hypothesis. For any function $\ell : \mathcal{Y}^2 \to \mathbb{R}$, the loss function corresponding to ℓ is*

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h(x), y\right)\right].$$

In particular, the 0-1 loss of h with respect to $\mathcal{D}$ is $\mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$.

(b) *Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. The loss of $\mathcal{H}$ with respect to $\mathcal{D}$ is*

$$L_{\mathcal{D}}(\mathcal{H}) = \inf\{L_{\mathcal{D}}(h) :\ h \in \mathcal{H}\}.$$

This quantity is also called the approximation error of $\mathcal{H}$ with respect to $\mathcal{D}$.

▶ Remark 2. Formally, for the loss function to be well defined, we must restrict our discussion to functions ℓ and hypothesis functions $h : \mathcal{X} \to \mathcal{Y}$ such that the composed function $(x,y) \mapsto \ell\left(h(x), y\right)$ is integrable. For example, instead of making statements concerning "the set of all functions $f : \mathcal{X} \to \mathcal{Y}$," one should more acculturate consider "the set of all functions $f : \mathcal{X} \to \mathcal{Y}$ such that the loss function $L_{\mathcal{D}}(f)$ exists." However, we will generally not be very precise about matters concerning measure theory and integration. ⌟

Using this definition we can the competitive objective as $L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(\mathcal{H}) + \varepsilon$, and this makes sense even if we do not assume that a target function exists (note that this differs from Eq. (1)). We only require that *if* there exists a hypothesis $h' \in \mathcal{H}$ that has loss α, *then* the learner's output h will satisfy $L_{\mathcal{D}}(h) \leq \alpha + \varepsilon$.

These steps yield the notion of *agnostic PAC Learning*, which is the main definition of learning in this course.

▶ **Definition 3** (Agnostic PAC Learning, Vapnik & Chervonenkis [6, 7], Haussler [2])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets, let $\mathcal{F}$ be the set of all functions $\mathcal{X} \to \mathcal{Y}$, and let $\mathcal{H} \subseteq \mathcal{F}$ be a class of functions. We say that a (possibly randomized) algorithm A is an agnostic PAC learner for $\mathcal{H}$ if there exists a sample complexity function $m :\ (0,1)^2 \to \mathbb{N}$ such that for every precision parameter $\varepsilon \in (0,1)$, every confidence parameter $\delta \in (0,1)$, and every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if A receives as input the parameters ε and δ and a sample S of size $m = m(\varepsilon,\delta)$ such that $S = \Big((x_1,y_1),\ldots,(x_m,y_m)\Big)$ where for each $i \in [m]$, the pair (x_i,y_i) is sampled independently from $\mathcal{D}$, then A halts and outputs a hypothesis $h \in \mathcal{F}$ that with probability at least $1-\delta$ (over the sample S and the randomness of A) satisfies*

$$L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(\mathcal{H}) + \varepsilon.$$

We say that a class $\mathcal{H} \subseteq \mathcal{F}$ is agnostic PAC learnable if there exists an algorithm that is an agnostic PAC learner for $\mathcal{H}$.

▶ Remark 4.

- Every agnostic PAC learner for $\mathcal{H}$ is in particular also a PAC learner for $\mathcal{H}$. Hence, the notion of agnostic PAC is stronger (harder to satisfy) than regular PAC. This is despite the fact that the competitive objective in agnostic PAC is weaker (easier to satisfy) than the absolute objective in PAC (namely, it is possible that h does not satisfy $L_{\mathcal{D}}(h) \leq \varepsilon$ but does satisfy $L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(\mathcal{H}) + \varepsilon$).
- The "agnostic" setting is actually not completely agnostic, because it still makes two non-trivial assumptions: that there exists a distribution that generates the sample (the sample is not arbitrary), and that the loss to be minimized is also measured according to the same distribution that generates the sample. These assumptions are very minimal and natural. Nonetheless, later on in the course we will see the setting of *online learning*, which employs the same idea of a competitive objective to discard even of these minimal assumptions.

2.1 Learning Scenarios

PAC learning and agnostic PAC will serve as our main definitions of learning for most of the course. Now that we have fully introduced them, it makes sense to reflect what real-world scenarios are captured by these definitions. Following are three such types of real-world scenarios.

1. **Binary classification.** The label space $\mathcal{Y}$ has cardinality 2, for instance $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{1, -1\}$, etc. Examples include:
   - Classifying an email as spam or ham (not spam).
   - Deciding whether a photograph depicts a cat.
   - Given input to a fingerprint reader, decide whether to unlock a device or not (whether the fingerprint belongs to the owner of the device or not).
   - Given data from a medical examination (e.g., an MRI scan), diagnose whether a patient has a specific medical condition or not (e.g., cancer).
   - In a vehicle, decide whether or not to activate the emergency brake assist (EBA) system.
   - Given a text message, decide whether it is intended humorously or literally.

   The 0-1 loss is a natural loss function for most cases of binary classification.
2. **Multi-class classification.** The label space $\mathcal{Y}$ is discrete and has more than two elements. Examples include:
   - Given a text, identify what language it is written in.
   - Given a photograph of an object, identify the object ("chair", "cat", etc.).
   - Given the current state in a game of chess, output a good next move.
   - Structured learning tasks, in which the required label has a nontrivial internal structure. For example, given a sentence in English, produce a syntactical parse tree of the sentence.
   - Ranking: given a set of objects, order the objects according to some criteria. For instance, a search engine is given a query and a set of documents, and is required to order the documents according to their relevance to the query. The label space is the set of possible permutations of the documents.

   The 0-1 loss can be a natural loss function for the first two examples above, but the remaining examples might be better served by a richer loss function that distinguishes between different levels of correctness.

3. **Regression.** The label space $\mathcal{Y}$ is $\mathbb{R}$, or some other large metric space, and the objective is to output a label that is close to the correct label with respect to the metric. Examples include:
   - Given sensor input from a self-driving car, try to identify the distance to an object ahead (e.g., a pedestrian).
   - Given current weather conditions, predict the temperature tomorrow.
   - Given recent sport statistics, predict the probability that a specific sports team will win an upcoming match.

   There are a variety of natural loss functions for regression, including (for real-valued labels) square loss $\ell(y, y') = (y - y')^2$, absolute error loss $\ell(y, y') = |y - y'|$, hinge loss, cross entropy loss, etc.

All of the above are instances of *supervised learning*, meaning that the learner first receives a training set with labeled instances, and then is required to predict the labels of previously unseen instances. There are also many important learning scenarios that are not instances of supervised learning and are not covered by the PAC and agnostic PAC models. These include reinforcement learning, online learning, transductive learning, semi-supervised learning, active learning, and learning with queries, as well as unsupervised learning scenarios such as clustering and dimensionality reduction. We will touch on some of these settings in later stages of the course.

3 Error Decomposition: The Bias–Complexity Tradeoff

The Bayes optimal error is the best (minimal) loss that can be achieved when predicting labels for a specific distribution, and it is determined by how noisy the labels are. Formally:

▶ **Definition 5.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets, assume $|\mathcal{Y}| < \infty$, let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and let $L_{\mathcal{D}}$ denote the* 0-1 *loss.*

(a) *The Bayes optimal error for $\mathcal{D}$ is*

$$L^*_{\mathcal{D}} = \inf \left\{ L_{\mathcal{D}}(f) \;\middle|\; f : \mathcal{X} \to \mathcal{Y} \right\}.$$

(b) *The Bayes optimal classifier for $\mathcal{D}$ is*

$$f^*(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \mathbb{P}_{(X,Y)\sim\mathcal{D}} \left[Y = y \mid X = x \right].$$

(c) *The noise of $\mathcal{D}$ at instance $x \in \mathcal{X}$ is*

$$\operatorname{noise}(x) = 1 - \max_{y \in \mathcal{Y}} \mathbb{P}_{(X,Y)\sim\mathcal{D}} \left[Y = y \mid X = x \right].$$

(d) *The noise of $\mathcal{D}$ is*

$$\operatorname{noise}(\mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}_x} \left[\operatorname{noise}(x) \right],$$

where $\mathcal{D}_x$ is the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$.

▶ **Fact 6.** $\operatorname{noise}(\mathcal{D}) = L^*_{\mathcal{D}} = L_{\mathcal{D}}(f^*)$.

Assume an agnostic PAC learning algorithm outputs hypothesis from a class $\mathcal{H}$ of functions $\mathcal{X} \to \mathcal{Y}$. Let $h \in \mathcal{H}$, and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. $L_{\mathcal{D}}(h) - L^*_{\mathcal{D}}$ is a measure of

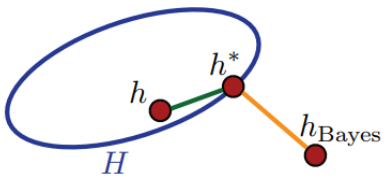


Figure 1 Illustration of the estimation error (in green) and approximation error (in orange). Here, it is assumed that there exists a best-in-class hypothesis, that is h^* such that $L_{\mathcal{D}}(h^*) = L_{\mathcal{D}}(\mathcal{H})$. Source: Mohri, Rostamizadeh and Talwalkar [4].

how well h compares to the best possible function f^* (which will typically not be a member of $\mathcal{H}$). We can always decompose this loss as follows.

$$L_{\mathcal{D}}(h) - L^*_{\mathcal{D}} = \underbrace{L_{\mathcal{D}}(h) - L_{\mathcal{D}}(\mathcal{H})}_{\text{estimation error}} + \underbrace{L_{\mathcal{D}}(\mathcal{H}) - L^*_{\mathcal{D}}}_{\text{approximation error}}.$$

There is a tradeoff between these two terms. Assume we make the class $\mathcal{H}$ larger and more complex, so that the algorithm has more output functions to choose from. Then

- The approximation error will typically become smaller, because $\mathcal{H}$ will gain functions that are closer to f^*. However,
- The estimation error will become larger, because finding a good hypothesis in $\mathcal{H}$ is harder when $\mathcal{H}$ is larger.

Hence, when designing a learning algorithm, we need to carefully consider which class of functions the algorithm should use. This decision will typically also be informed by any prior knowledge available about the learning problem.

4 Agnostic PAC Learning via Uniform Convergence

We develop the notion of *uniform convergence*, using the following lemma as an example. It generalizes the result for finite classes to the agnostic setting.

▶ **Lemma 7.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be (finite or infinite) nonempty sets, and let $\mathcal{H}$ be a finite subset of the functions $\mathcal{X} \to \mathcal{Y}$. Then $\mathcal{H}$ is agnostic PAC learnable with sample complexity* $m(\varepsilon, \delta) = \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$.

▶ Remark 8. Note that the dependence on $1/\varepsilon$ is quadratic, unlike in the PAC case.

We previously saw a proof that finite classes are PAC learnable. That outline of the proof was:

1. For any single hypothesis $h \in \mathcal{H}$, we can efficiently estimate whether h is ε-good or not using a few samples. (Taking $\frac{\ln(1/\delta)}{\varepsilon}$ samples suffices to check whether $L_{\mathcal{D}}(h) \leq \varepsilon$ with confidence $1 - \delta$.)
2. Because $\mathcal{H}$ is finite, we can use a union bound in order to estimate this for all $h \in \mathcal{H}$ simultaneously. (Taking $\frac{\ln(|\mathcal{H}|/\delta)}{\varepsilon}$ samples suffices to check this for all $h \in \mathcal{H}$.)

To prove Lemma 7, we will follow the same outline, using a notion of *uniform convergence*. Recall that in calculus, their is a notion of uniform convergence for sequences of functions, which contrasts with pointwise convergence:

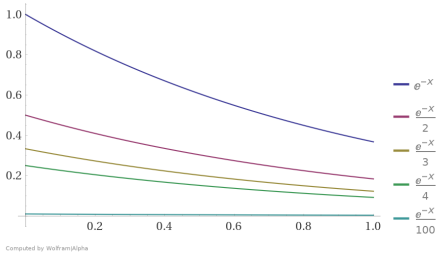
▶ **Definition 9.** *Let Ω be a set, let $f_1, f_2, f_3 \ldots$ be an infinite sequence of functions $\Omega \to \mathbb{R}$, and let $f^* : \Omega \to \mathbb{R}$ be a function.*

1. *We say that* $\underline{(f_n)_{n\in\mathbb{N}} \text{ convergence pointwise to } f^*}$ *if*

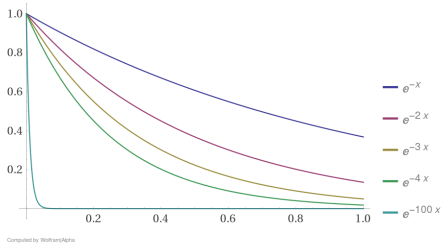
$$(\forall \varepsilon > 0)\ (\forall x \in \Omega)\ (\exists N \in \mathbb{N})\ (\forall n \geq N):\ |f_n(x) - f^*(x)| \leq \varepsilon.$$

2. *We say that* $\underline{(f_n)_{n\in\mathbb{N}} \text{ convergence uniformly to } f^*}$ *if*

$$(\forall \varepsilon > 0)\ (\exists N \in \mathbb{N})\ (\forall x \in \Omega)\ (\forall n \geq N):\ |f_n(x) - f^*(x)| \leq \varepsilon.$$



(a) $f_n = e^{-x}/n$ converges to $f^* \equiv 0$ pointwise and uniformly in $(0, 1)$.



(b) $f_n = e^{-nx}$ converges to $f^* \equiv 0$ pointwise but *not* uniformly in $(0, 1)$.

Figure 2 Pointwise vs. uniform convergence for sequences of real functions.

Similarly, we introduce the following definition which concerns the convergence of the empirical loss functions L_{S_m} to the population loss $L_{\mathcal{D}}$, where L_{S_m} is the empirical loss for a sample of size m.

▶ **Definition 10.** *Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. We say that $\underline{\mathcal{H} \text{ satisfies the uniform convergence property}}$ if*

$$(\forall \varepsilon, \delta \in (0,1))\ (\exists M \in \mathbb{N})\ (\forall \text{ distribution } \mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y})\ (\forall h \in \mathcal{H})\ (\forall m \geq M):$$
$$\mathbb{P}_{S_m \sim \mathcal{D}^m}\left[|L_{S_m}(h) - L_{\mathcal{D}}(h)| > \varepsilon\right] \leq \delta. \tag{2}$$

Let $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ be a function. We say that $\underline{\mathcal{H} \text{ satisfies uniform convergence with sample complexity } m_{\mathcal{H}}^{\mathrm{UC}}}$ if taking $M = m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ satisfies (2) above.[1]

The following lemma states that uniform convergence is sufficient for agnostic learnability.

▶ **Lemma 11.** *Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. If $\mathcal{H}$ satisfies the uniform convergence property with sample complexity $m_{\mathcal{H}}^{\mathrm{UC}}$, then $\mathrm{ERM}_{\mathcal{H}}$ is an agnostic PAC learner for $\mathcal{H}$ with sample complexity $m(\varepsilon, \delta) = m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon/2, \delta)$.*

Proof. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, fix $\varepsilon, \delta \in (0,1)$, and let $h = \mathrm{ERM}_{\mathcal{H}}(S)$ where $S \sim \mathcal{D}^m$ and $m = m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon/2, \delta)$. From uniform convergence, with probability at least $1 - \delta$,

$$\forall h' \in \mathcal{H}:\ |L_S(h') - L_{\mathcal{D}}(h')| \leq \varepsilon/2. \tag{3}$$

In that case, for every $h^* \in \mathcal{H}$,

$$\begin{aligned}
L_{\mathcal{D}}(h) &\leq L_S(h) + \varepsilon/2 && \text{(from uniform convergence (3))}\\
&\leq L_S(h^*) + \varepsilon/2 && (h \text{ is an ERM hypothesis})\\
&\leq L_{\mathcal{D}}(h^*) + \varepsilon/2 + \varepsilon/2 && \text{(from uniform convergence (3))}\\
&= L_{\mathcal{D}}(h^*) + \varepsilon. && \blacktriangleleft
\end{aligned}$$

[1] Namely, $(\forall \varepsilon, \delta \in (0,1))\ (\forall \mathcal{D})\ (\forall h \in \mathcal{H})\ (\forall m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)):\ \mathbb{P}_{S_m \sim \mathcal{D}^m}\left[|L_{S_m}(h) - L_{\mathcal{D}}(h)| > \varepsilon\right] \leq \delta$.

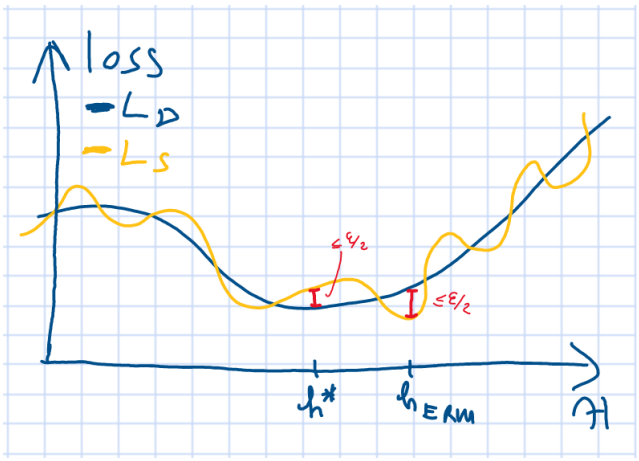


Figure 3 Uniform convergence implies agnostic PAC learnability (Lemma 11). The empirical loss is $\frac{\varepsilon}{2}$-close to the true loss, and therefore the ERM algorithm performs well.

To prove Lemma 7, we will show that finite classes satisfy uniform convergence.

▷ Claim 12. Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. If $\mathcal{H}$ is finite, then $\mathcal{H}$ satisfies uniform convergence with sample complexity $m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta) = \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$. ⌟

Proof of Claim 12. Let $h \in \mathcal{H}$. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and parameters $\varepsilon, \delta \in (0,1)$. Let $S \sim \mathcal{D}^m$ for $m = m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ as in the statement. Fix $h \in \mathcal{H}$. To complete the proof it suffices to show that $\mathbb{P}\left[|L_S(h) - L_{\mathcal{D}}| > \varepsilon\right] \leq \delta/|\mathcal{H}|$, and then the result follows from a union bound over all $h \in \mathcal{H}$.

Let $S = \Big((x_1, y_1), \ldots, (x_m, y_m)\Big)$. Note that $\{\mathbb{1}(h(x_i) \neq y_i)\}_{i \in [m]}$ are i.i.d. random variables with support in $\{0,1\}$ and mean $L_{\mathcal{D}}(h)$. Hence, Hoeffding's inequality yields

$$\begin{aligned}
\mathbb{P}_{S \sim \mathcal{D}^m}\left[|L_S(h) - L_{\mathcal{D}}| > \varepsilon\right] &= \mathbb{P}_S\left[\left|\frac{1}{m}\sum_{i \in [m]} \mathbb{1}(h(x_i) \neq y_i) - L_{\mathcal{D}}(h)\right| > \varepsilon\right] \\
&\leq 2\exp\left(2m\varepsilon^2\right) && \text{(Hoeffding's inequality)} \\
&\leq \delta/|\mathcal{H}|, && \left(m = \left\lceil \tfrac{\ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil\right)
\end{aligned}$$

as desired. ◀

As a corollary, we obtain Lemma 7.

Proof of Lemma 7. From Claim 12, $\mathcal{H}$ satisfies uniform convergence with sample complexity $m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta) = \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$. From Lemma 11, $\mathrm{ERM}_{\mathcal{H}}$ is an agnostic PAC learner for $\mathcal{H}$ with sample complexity $\left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$. ◀

5 The VC Dimension: Uniform Convergence for Infinite Classes

So far, we have seen that the set of all functions over an infinite domain is not PAC learnable (and therefore not agnostic PAC learnable), while finite classes of functions are agnostic PAC learnable (and therefore also PAC learnable). But what about classes in between these two extremes?

First, we observe that some infinite classes over infinite domains are agnostic PAC learnable, and therefore being finite is a sufficient condition for learnability, but is not a necessary condition.

▶ **Lemma 13.** *Let*

$$\mathcal{T} = \left\{ f_t : \mathbb{R} \to \{0,1\} \;\middle|\; \forall x \in \mathbb{R} : \; f_t(x) = \mathbb{1}(x \geq t) \right\}$$

be the set of threshold functions (boolean monotone increasing functions over the domain $\mathbb{R}$). Then $\mathrm{ERM}_{\mathcal{T}}$ is a PAC learner for $\mathcal{T}$ with sample complexity $m(\varepsilon, \delta) = \left\lceil \frac{\ln(2/\delta)}{\varepsilon} \right\rceil$.

Proof. We prove the lemma under the assumption that the distribution $\mathcal{D}$ over the domain $\mathbb{R}$ has a continuous cumulative distribution function (CDF), denoted $F_{\mathcal{D}}$.[2] The proof for the case with point masses is similar but requires a bit more care.

Fix $\varepsilon, \delta \in (0,1)$, a distribution $\mathcal{D}$ over $\mathbb{R}$, and a target function $f_{t^*} \in \mathcal{T}$. Let $S \sim (\mathcal{D}, f_{t^*})^m$ and $h = \mathrm{ERM}_{\mathcal{T}}(S)$. We need to show that $\mathbb{P}_S \left[L_{\mathcal{D}}(h) > \varepsilon \right] < \delta$.

Let $t_h \in \mathbb{R}$ be such that $h = f_{t_h}$. Let B be the set of points that are misclassified by h. Namely,

$$B = \begin{cases} [\min\{t_h, t^*\}, \max\{t_h, t^*\}) & t_h \neq t^* \\ \varnothing & t_h = t^* \end{cases}$$

Note that $L_{\mathcal{D}}(h) = \mathcal{D}(B)$, so

$$\begin{aligned} \mathbb{P}\left[L_{\mathcal{D}}(h) > \varepsilon \right] &= \mathbb{P}\left[\mathcal{D}(B) > \varepsilon \right] \\ &= \mathbb{P}_S \left[t_h < t^* \;\wedge\; \mathcal{D}(B) > \varepsilon \right] + \mathbb{P}_S \left[t_h > t^* \;\wedge\; \mathcal{D}(B) > \varepsilon \right]. \end{aligned} \tag{4}$$

Hence, it suffices to show that each of the two summands in Eq. (4) is bounded by $\delta/2$. We prove that $\mathbb{P}_S \left[t_h < t^* \;\wedge\; \mathcal{D}(B) > \varepsilon \right] \leq \delta/2$. Note that if $F_{\mathcal{D}}(t^*) \leq \varepsilon$ then we are done, because $\mathcal{D}(B) \leq \mathcal{D}((-\infty, t^*)) = F_{\mathcal{D}}(t^*) \leq \varepsilon$. Otherwise, $F_{\mathcal{D}}(t^*) > \varepsilon$, $\lim_{t \to -\infty} F_{\mathcal{D}}(t) = 0$, and $F_{\mathcal{D}}$ is continuous, and so by the intermediate value theorem there exists $t_\varepsilon \in (-\infty, t^*)$ such that $F_{\mathcal{D}}(t_\varepsilon) = F_{\mathcal{D}}(t^*) - \varepsilon$. Namely, the interval $I_\varepsilon = (t_\varepsilon, t^*)$ satisfies $\mathcal{D}(I_\varepsilon) = F_{\mathcal{D}}(t^*) - F_{\mathcal{D}}(t_\varepsilon) = \varepsilon$.

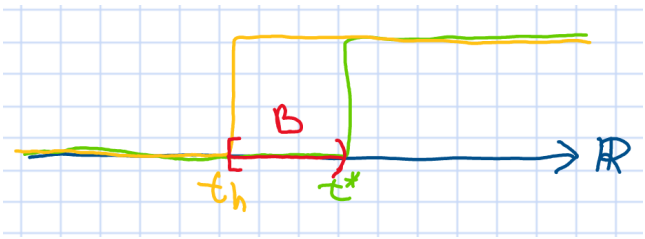
Let $S = \Big((x_i, y_i) \Big)_{i \in [m]}$, and observe that if $t_h < t^*$ and $\exists i \in [m]$ such that $x_i \in I_\varepsilon$, then $t_\varepsilon < x_i < t_h < t^*$, and this implies $\mathcal{D}(B) = \mathcal{D}([t_h, t^*)) \leq \mathcal{D}(I_\varepsilon) = \varepsilon$. Hence,

$$\begin{aligned} \mathbb{P}_{S \sim \mathcal{D}^m} \left[t_h < t^* \;\wedge\; \mathcal{D}(B) > \varepsilon \right] &\leq \mathbb{P}_S \left[\forall i \in [m] : \; x_i \notin I_\varepsilon \right] \\ &= \prod_{i \in [m]} \mathbb{P}_S \left[x_i \notin I_\varepsilon \right] && (x_i\text{'s are independent}) \\ &= (1 - \varepsilon)^m && (\mathcal{D}(I_\varepsilon) = \varepsilon) \\ &\leq e^{-\varepsilon m} \leq \delta/2. && \left(m = \left\lceil \tfrac{\ln(2/\delta)}{\varepsilon} \right\rceil\right) \end{aligned}$$

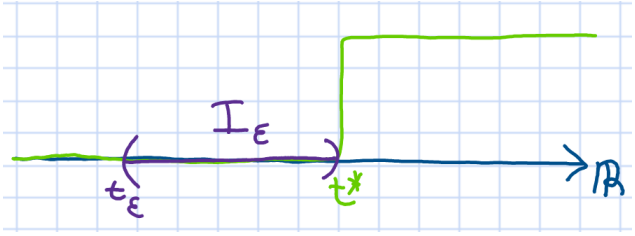
This proves the bound for the left summand in Eq. (4). The proof for the right summand is analogous. ◀

Towards developing a condition that precisely characterizes which classes are PAC or agnostic PAC learnable, we start with a slightly more general version of Corollary 12 in Unit 2, which stated that the class of all functions from an infinite domain is not learnable.

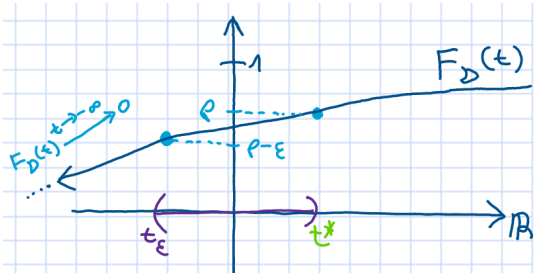
[2] Namely, $\mathcal{D}(x) = 0$ for all $x \in \mathbb{R}$.



(a) The interval B contains the points that are misclassified by h.



(b) The interval I_ϵ has measure ε, and t^* is its right endpoint.



(c) t_ε, the left endpoint of I_ε, exists by an application of the intermediate value theorem to the CDF of $\mathcal{D}$, which is assumed to be continuous.

Figure 4 Details of the proof of Lemma 13.

This followed from Corollary 11 in Unit 2, which implied that PAC learning the class of all functions from a set $\mathcal{X}$ of cardinality n to $\{0,1\}$ with parameters $\varepsilon, \delta < \frac{1}{8}$ is not possible with less than $n/2$ samples. This implied Corollary 12 in Unit 2 because for any $n \in \mathbb{N}$, the set of all functions from an infinite domain contains within it a a set of all functions from a subset of the domain of cardinality n, and so PAC learning requires at least $n/2$ samples for all $n \in \mathbb{N}$ – namely, a non-finite number of samples. We can therefore reformulate Corollaries 11 and 12 in the following more general way.

▶ **Definition 14.** *Let $\mathcal{X}$ be a nonempty set, $\mathcal{Y} = \{0,1\}$, $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$, $f \in \mathcal{H}$, and let $X \subseteq \mathcal{X}$.*

1. *The restriction of the function f to the subset X, is the function $f|_X : X \to \mathcal{Y}$ such that $\forall x \in X : f|_X(x) = f(x)$.*
2. *The restriction of the class $\mathcal{H}$ to the subset X, denoted $\mathcal{H}|_X$, is the set of functions $\{f|_X : f \in \mathcal{H}\}$.*
3. *We say that $\mathcal{H}$ shatters X if the restriction of $\mathcal{H}$ to X is the set of all functions $X \to \mathcal{Y}$, namely, if $\left|\mathcal{H}|_X\right| = 2^{|X|}$.*

▷ Claim 15. Let $\mathcal{X}$ be a set, let $\mathcal{Y} = \{0,1\}$, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$.

1. Assume $\mathcal{H}$ shatters some subset $X \subseteq \mathcal{X}$ such that $|X| = n$. Then any algorithm that PAC learns $\mathcal{H}$ with parameters $\varepsilon, \delta < 1/8$ must use a sample of size at least $n/2$.

2. Assume that for any $n \in \mathbb{N}$, there exists $X \subseteq \mathcal{X}$ such that $|X| = n$ and $\mathcal{H}$ shatters X. Then $\mathcal{H}$ is not PAC learnable.

Proof. For Part 1, assume that A is an algorithm that PAC learns $\mathcal{H}$ with parameters $\varepsilon, \delta < 1/8$. Then in particular, A PAC learns $\mathcal{H}$ with parameters $\varepsilon, \delta < 1/8$ for any unknown distribution $\mathcal{D}$ such that $\mathcal{D}(X) = 1$. This is equivalent to A PAC learning the class $\mathcal{H}|_X$ of functions $X \to \mathcal{Y}$ with parameters $\varepsilon, \delta < 1/8$. Because $\mathcal{H}$ shatters X, $\mathcal{H}|_X$ is the set of all functions $X \to \mathcal{Y}$, and so from Corollary 11 in Lecture 2, A must use a samples of size at least $n/2$.

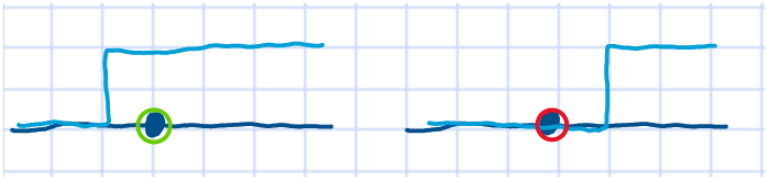
Part 2 follows from Part 1 as follows. Assume for contradiction that there exists an algorithm A that PAC learns $\mathcal{H}$ with parameters $\varepsilon, \delta < 1/8$ and uses m samples. From the assumption, there exists a set $X \subseteq \mathcal{X}$ such that $\mathcal{H}$ shatters X and $|X| > 2m$. Then from Part 1, A must use strictly more than m samples, a contradiction. ◀

This motivates the following definition.

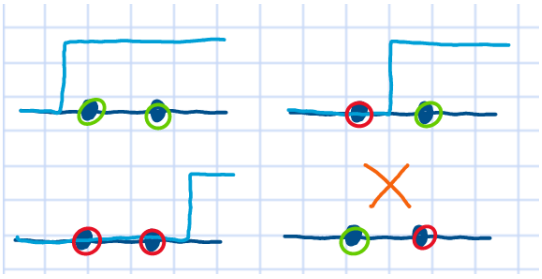
▶ **Definition 16** (Vapnik & Chervonenkis [6]). *Let $\mathcal{X}$ be a nonempty set, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{H}$ is*

$$\mathsf{VC}(\mathcal{H}) = \sup \{|X| : \ X \subseteq \mathcal{X} \ \wedge \ \mathcal{H} \ \text{ shatters } X\}.$$

Namely, $\mathsf{VC}(\mathcal{H})$ is the largest number $n \in \mathbb{N}$ such that $\mathcal{H}$ shatters a set of cardinality n, and if $\mathcal{H}$ shatters sets of unbounded size then $\mathsf{VC}(\mathcal{H}) = \infty$.



(a) $\mathcal{T}$ can shatter a point.



(b) $\mathcal{T}$ cannot shatter any set of two points.

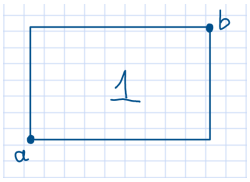
Figure 5 The VC dimension of monotone increasing thresholds is 1.

▶ **Example 17.** *Let $\mathcal{T} = \left\{ f_t : \mathbb{R} \to \{0, 1\} \;\middle|\; \forall x \in \mathbb{R} : \ f_t(x) = \mathbb{1}(x \geq t) \right\}$ be the class of thresholds (as in Lemma 13). Then $\mathsf{VC}(\mathcal{T}) = 1$.*

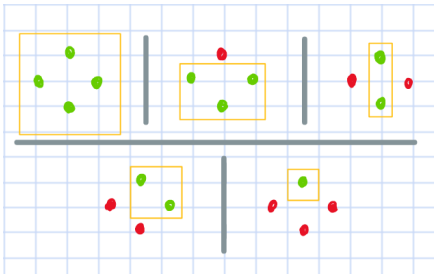
To see this, we need to show two things:

- *There exists a set $A \subseteq \mathbb{R}$ such that $|A| = 1$ and $\mathcal{T}$ shatters A. Indeed, let $A = \{0\}$, and note that $\left|\mathcal{T}\big|_A\right| = 2 = 2^{|A|}$ because $f_{-1}(0) = 1$ and $f_1(0) = 0$. Hence $\mathcal{T}$ shatters A.*

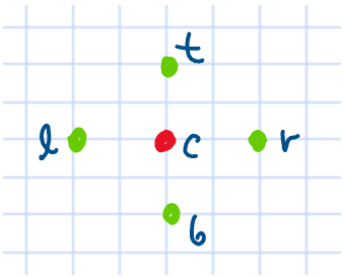
- *For any set $A \subseteq \mathbb{R}$ such that $|A| > 1$, $\mathcal{T}$ does not shatter A. It suffices to show that no set of size 2 is shattered, because if a set of size $n > 2$ is shattered then in particular every subset of size 2 of that set is also shattered. Fix a set $A = \{x, y\}$ with $x < y$. There does not exist $f \in \mathcal{T}$ such that $f(x) = 1$ and also $f(y) = 0$. Hence, A is not shattered.*



(a) An example of an axis-aligned rectangle $f_{a,b} \in \mathcal{R}$.



(b) There exists a set of 4 points in $\mathbb{R}^2$ that is shattered by $\mathcal{R}$.



(c) There does not exist a set of 5 points in $\mathbb{R}^2$ that is shattered by $\mathcal{R}$.

Figure 6 The VC dimension of axis-aligned rectangles is 4.

▶ **Example 18.** *An axis-aligned rectangle in $\mathbb{R}^2$ is a function $f_{a,b} : \mathbb{R}^2 \to \{0, 1\}$ with $a = (a_x, a_y) \in \mathbb{R}^2$ and $b = (b_x, b_y) \in \mathbb{R}^2$ such that*

$$f_{a,b}(x, y) = \mathbb{1}(a_x \leq x \leq b_x \ \wedge \ a_y \leq y \leq b_y).$$

Let $\mathcal{R}$ be the class of all axis-aligned rectangles, $\mathcal{R} = \{f_{a,b} : \ a, b \in \mathbb{R}^2\}$. Then $\mathsf{VC}(\mathcal{R}) = 4$.

To see this, we need to show two things:

- *There exists a set $A \subseteq \mathbb{R}$ such that $|A| = 4$ and $\mathcal{R}$ shatters A. Indeed, the set $A = \{(1, 0), (-1, 0), (0, 1), (0, -1)\}$ is shattered.*
- *For any set $A \subseteq \mathbb{R}$ such that $|A| > 4$, $\mathcal{R}$ does not shatter A. Fix a set A with $|A| \geq 5$. We claim that A is not shattered. Let ℓ, r, t, b denote the leftmost, rightmost, topmost and bottommost members of A (if more than one members of A are equally extreme in some direction, choose one of them arbitrarily). Let $c \in A \setminus \{\ell, r, t, b\}$ (c exists because $|A| \geq 5$). To see that A is not shattered is suffices to observe that for any*

rectangle $f_{z,z'}$, if we assume that $f_{z,z'}(\ell) = f_{z,z'}(r) = f_{z,z'}(t) = f_{z,z'}(b) = 1$, then $f_{z,z'}(c) = 1$. This observation is true because the assumption implies $z_x \leq \ell_x \leq c_x \leq r_x \leq z'_x$, and similarly $z_y \leq c_y \leq z'_y$, and therefore $f_{z,z'}(c) = 1$.

▶ **Example 19.** *Let $\mathcal{H}$ be a class of functions with $|\mathcal{H}| < \infty$. Then $\mathsf{VC}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$. Indeed, let A be a shattered subset of the domain. Then*

$$2^{|A|} = \left|\mathcal{H}|_A\right| \leq |\mathcal{H}| \implies |A| \leq \log_2(|\mathcal{H}|).$$

So $\mathcal{H}$ can shatter sets of cardinality at most $\log_2(|\mathcal{H}|)$.

The above inequality is tight when $\mathcal{H}$ is the set of all boolean functions over a finite domain. Notice that $\mathsf{VC}(\mathcal{H})$ can be much smaller than $\log_2(|\mathcal{H}|)$. For instance, let $\mathcal{H} \subseteq \mathcal{T}$ where $\mathcal{T}$ is as in Example 17 and $|\mathcal{H}| < \infty$. Then $\mathsf{VC}(\mathcal{H}) = 1$.

6 Discussion

In this unit, we have introduced the agnostic PAC models, and the concepts of uniform convergence and VC dimension. In the next unit we will see how to combine these concepts to prove the fundamental theorem of PAC learning, which states that the VC dimension completely characterizes the learnability of classes in the PAC and agnostic PAC models.

6.1 Bibliographic Notes

The VC dimension and many related results are due to Vapnik and Chervonenkis [6, 7]. See also [8]. The formalization of agnostic PAC learning is due to Haussler [2].

Good expositions of the material in this unit are available in [5, Chapters 3–6], [4, Chapters 2–3], as well as [1, Chapters 5–7] and [3, Chapter 3].

References

1 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`.

2 David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992. `doi:10.1016/0890-5401(92)90010-D`.

3 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`.

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

5 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`.

6 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`.

7 Vladimir N. Vapnik and Alexey Ya. Chervonenkis. Theory of pattern recognition (in Russian), 1974. German translation: Theorie der Zeichenerkennung, Akademie-Verlag, Berlin (1979).

8 Vladimir Naumovich Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2nd edition, 2000.