**47-837: Statistical Foundations of OR** Lecture 2

# Statistical Learning Theory

Lecturer: Karan Singh

Our goal in this lecture is to introduce Probably Approximate Correct (PAC) learning and build up to *the* central result in learning theory, namely, that learnability for binary classification is exactly characterized by the VC dimension of the underlying hypothesis class. We will see a couple of applications of this result: the DKW inequality, and decision-theoretic learning of quantiles. At the end, we will also pave an alternative path to learning that relies on generalization through algorithmic stability.

But before we introduce the PAC framework, we pay our debt to concentration inequalities for nonlinear functions, which will be essential in deriving the results promised earlier.

## Contents



## 1. Martingales

Concentration inequalities for nonlinear functions of independent variables boils down to concentration of linear forms of dependent, but controlled random variables. We will make this concrete in the next section. With this motivation, we introduce *martingales* to capture sequences of such dependent random variables.

**Definition 1** A sequence of random variables $\{X_i\}_{i\geq 1}$ is a martingale if for all $n$, we have

$$\mathbb{E}[X_{n+1}|\ X_n, X_{n-1}, ..., X_1] = X_n.$$

More generally, a sequence of random variables $\{X_i\}_{i\geq 1}$ forms a martingale with respect to another sequence $\{Y_i\}_{i\geq 1}$ if for all $n$, we have

$$\mathbb{E}[X_{n+1}|\ Y_n, Y_{n-1}, ..., Y_1] = X_n,$$

where $X_n$ is measurable in (*read as*: completely determined given) $Y_1, ...Y_n$. The associated sequence $\{X_{i+1} - X_i\}_{i\geq 1}$ is called a *martingale difference sequence*.

As an example of martingale, consider prefix sums of independent zero-mean random variables, for example, $X_n = \sum_{i\leq n} Y_i$, where $Y_n \sim \{\pm 1\}$ are independent, which model random walks in one dimension. We can also have non-sum-like sequences, for example, $X_{n+1} = (1 + Y_{n+1}\sin(X_n))X_n$, where $Y_i$'s remain as defined previously.

Historically, the term *martingales* originates from a certain class of French betting strategies; one can still find people puzzled about these on Youtube. The setup is as follows: at any stage one can bet any amount of choice on a fair random coin toss, receiving twice the initial amount upon success, and nothing on failure. Clearly, there is no way to predict a fair coin toss, and hence, one should not expect to make money in this circumstance. But consider the following strategy:

"Starting with a \$1 bet in the first round, double the bet upon losing, and quit when you win."

Since $2^{n+1} - 1 = 1 + 2 + ... + 2^n$, it is easy to observe that upon winning, one makes up all the money lost in the previous rounds and gains an extra dollar, ending up in the green. Winning in the long run happens almost surely, and hence, this specious argument seems to guarantee a small profit. Of course, one might run out of money this way, but consider having an infinite purse. Even then, we can observe that the return $S_n$ at end of round $n$ is distributed as

$$S_n = \begin{cases} 1 & \text{with probability } 1 - \frac{1}{2^n} \\ -(2^n - 1) & \text{with probability } \frac{1}{2^n} \end{cases}$$

which on expectation is zero. In fact, $S_n$ forms a martingale sequence. (Verify this!) The almost surety of winning happens at the cost of cataclasmic losses with exponential small probability.

## 2. Bounded Differences Inequality

We generalize the subgaussian character of sums of independent random variables to martingales.

**Lemma 2** If $\Delta_{j+1}|X_{1:j} = x_{1:j} \sim \text{SubGaussian}(\sigma_j^2)$ for all $x_{1:j}$ and $j$, and $\{\Delta_n\}$ forms a martingale difference sequence with respect to $\{X_n\}$, then $\sum_{i=1}^n \Delta_i \sim \text{SubGaussian}\left(\sum_{i=1}^n \sigma_i^2\right)$.

*Proof.* The proof follows by repeated applications of $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, while noting that $\Delta_{1:j}$ is measurable in $X_{1:j}$.

$$\begin{aligned}\mathbb{E}\left[e^{t\left(\sum_{i=1}^{n}(\Delta_i-\mathbb{E}\Delta_i)\right)}\right] &= \mathbb{E}_{X_{1:n-1}}\left[\mathbb{E}_{X_n}\left[e^{t\left(\sum_{i=1}^{n}(\Delta_i-\mathbb{E}\Delta_i)\right)}|X_{1:n-1}\right]\right] \\ &= \mathbb{E}_{X_{1:n-1}}\left[\mathbb{E}_{X_n}\left[e^{t(\Delta_n-\mathbb{E}\Delta_n)}|X_{1:n-1}\right]e^{t\left(\sum_{i=1}^{n-1}(\Delta_i-\mathbb{E}\Delta_i)\right)}\right] \\ &\leq e^{-t^2\sigma_n^2/2}\mathbb{E}\left[e^{t\left(\sum_{i=1}^{n-1}(\Delta_i-\mathbb{E}\Delta_i)\right)}\right] = \ldots = e^{t^2\sum_{i=1}^{n}\sigma_i^2/2}\end{aligned}$$

□

Our main result in this section is that any nonlinear function $f$ with independent random variables as arguments concentrates to its mean, as long as it can not changed a lot by tweaking a single argument in isolation. One unfortunate aspect of this result, although it will suffice for us in this lecture, is that the sensitivities to coordinates must be measured in a worst-case sense, that is, by fixing the other coordinates to their worst configuration. Time permitting, in later lectures, we will fix this and also extend the result to Lipschitz functions.

**Theorem 3 (McDiarmid's Inequality)** Consider a $n$-variate function $f : \mathcal{X}^n \to \mathbb{R}$. Define

$$\delta_i(x_{1:n}) = \max_{x\in\mathcal{X}} f(x_{1:i}, x, x_{i+1:n}) - \min_{y\in\mathcal{X}} f(x_{1:i}, y, x_{i+1:n}).$$

Then, for any $t > 0$ and $c_i \geq \|\delta_i\|_\infty := \max_{x_{1:n}} \delta_i(x_{1:n})$, we have that

$$\Pr(|f(X_1, \ldots X_n) - \mathbb{E}f(X_1, \ldots X_n)| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}}.$$

*Proof.* We begin by noting that $f(X_{1:n}) - \mathbb{E}f(X_{1:n})$ can be decomposed as

$$\begin{aligned}f(X_{1:n}) - \mathbb{E}f(X_{1:n}) &= f(X_{1:n}) - \mathbb{E}_{X_n}[f(X_{1:n})|X_{1:n-1}] + \mathbb{E}_{X_n}[f(X_{1:n})|X_{1:n-1}] - \mathbb{E}[f(X_{1:n})] \\ &= \sum_{i=1}^{n} \mathbb{E}[f(X_{1:n})|X_{1:i}] - \mathbb{E}[f(X_{1:n})|X_{1:i-1}].\end{aligned}$$

Clearly, $\Delta_i := \mathbb{E}[f(X_{1:n})|X_{1:i}] - \mathbb{E}[f(X_{1:n})|X_{1:i-1}]$ forms a martingale difference sequence with respect to $\{X_n\}$, since $\Delta_i$ is measurable in $X_{1:i}$ and

$$\mathbb{E}\big[\Delta_{j+1}|X_{1:j}\big] = \mathbb{E}_{X_{j+1}}\big[\mathbb{E}\big[f(X_{1:n})|X_{1:j+1}\big]\big] - \mathbb{E}\big[f(X_{1:n})|X_{1:j}\big] = 0.$$

It is plain to see that $\Delta_i = \mathbb{E}_{Y_i, X_{i+1:n}}[f(X_{1:n}) - f(X_{1:i-1}, Y_i, X_{i+1:n})|X_{1:i}] \leq c_i$, and hence it is subgaussian with variance proxy $c_i^2/4$, by Hoeffding's lemma. Thus we have fulfilled all the requirements of the previous lemma, and hence $f(X_{1:n})$ is subgaussian with variance proxy $\left(\sum_{i=1}^{n} c_i^2\right)/4$. The tail bound immediately follows from this observation. □

## 2.1. Application: Max Cut

As our first example, consider the $G(n, 1/2)$ family of random graphs. This is a distribution over all undirected (simple) graphs over $n$ vertices where each pair of distinct vertices is connected by an edge with probability $1/2$, independently of the other pairs. We are interested in figuring out the size of the maximum cut, that is, the the maximum number of edges that cross any partition of the vertex set, with probability 0.99. Treating the presence of edges as independent Bernoulli

variables, any balanced cut, one with nearly equal number of vertices on both sides, has $\frac{n^2}{4} \times \frac{1}{2} = \frac{n^2}{8}$ edges on expectation, and is subgaussian with variance proxy $\frac{n^2}{4} \times \frac{1}{2} \times \left(1 - \frac{1}{2}\right) = \frac{n^2}{16}$. Hence, since there are $2^n$ possible cuts in total, a blind application of the maximal inequality gives the size of the maximum cut to be $\frac{n^2}{8} \pm O\left(\sqrt{\frac{n^2}{16} \log 2^n}\right) = \frac{n^2}{8} \pm O\left(n^{3/2}\right)$.

While the maximal inequality gives the correct *expected* size of the maxcut as $\frac{n^2}{8} + Cn^{3/2}$, for some universal constant $C$, using McDiarmid, we will see that the fluctuations due to randomness are just $\pm O(n)$ in size. To see this, think of the maximum cut as a function of $\binom{n}{2}$ indicator variables of individual edges, which go up (or down) by at most one while adding (or removing) an edge, that is $c_i = 1$. As a consequence, we get the maximum cut lies in $\frac{n^2}{8} + Cn^{3/2} \pm O(n)$.

### 2.2. Application: Bin Packing

As a second example, consider $n$ items of independent random sizes $\{X_i\}$ in the range $[0, 1]$. Let $Y_n$ be the minimum number of unit-sized bins required to pack these, where we cannot split an item between two bins. Imagine, for example, $X_i \sim \text{Unif}[0, 1]$, in which case $\mathbb{E}[Y_n] \geq \left\lceil \sum_{i=1}^n X_i \right\rceil = n/2$. Again, $Y_n$ can go up (or down) by at most one by increasing (or decreasing) the size of a single item. Hence, $Y_n$ lies in $\mathbb{E}[Y_n] \pm O(\sqrt{n})$ with probability 0.99. Thus, although $Y_n$ on any specific day involves a NP-hard problem, over provisioning boxes by a vanishingly small fraction fulfills the demand with high probability on any day without knowledge of the realized item sizes.

## 3. PAC Learning

Let us first concretely define the learning task. We will imagine that there is a feature space $\mathcal{X}$ and label space $\mathcal{Y}$, and on top of this is a data-generating distribution $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$. A loss function $l : \mathcal{Y}^2 \to \mathbb{R}$ assigns a loss to the prediction $\hat{y} \in \mathcal{Y}$ when the correct label is $y$ as $l(\hat{y}, y)$. In this lecture, we will deal with binary classification, where $\mathcal{Y} = \{0, 1\}$ and $l(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y}$, often termed the zero-one loss. Given this, we can define the population error of any classifier $h$ to be

$$\text{err}_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h(x), y)] = \Pr_{(x,y)\sim\mathcal{D}}(h(x) \neq y).$$

Our first result is a negative one. Note that a random classifier has an error of $1/2$. In words, the proposition states that no learning algorithm can have a significantly better error without observing a constant fraction of all the data points, even if a perfect classifier exists. If $\mathcal{X} = \{0, 1\}^d$, this sample requirement for nontrivial error scales as $2^d$.

Note that if the latter requirement of a perfect classifier is dropped, then we can take $\mathcal{D}$ to be the uniform distribution on $\mathcal{X}$ augmented with $\Pr(Y = 1|X = x) \sim \text{Be}(1/2)$ for all $x \in \mathcal{X}$, for which even the best classifier can do no better than half on error. But this is a setting in which knowing $\mathcal{D}$ beforehand confers no advantage; hence, this does not capture a failure of learnability.

**Proposition 4 (No Free Lunch)** Consider any finite $\mathcal{X}$, and any learning algorithm that upon observing $m$ samples produces a classifier $h_{\mathcal{A}}$. Then, there exists a distribution $\mathcal{D}$ supported on $\mathcal{X} \times \{0, 1\}$ such that $\mathbb{E}[\text{err}_{\mathcal{D}}(h_{\mathcal{A}})] \geq \frac{1}{2}\left(1 - \frac{m}{|\mathcal{X}|}\right)$ while $\min_{f^* \in [0,1]^{\mathcal{X}}} \text{err}_{\mathcal{D}}(f^*) = 0$.

*Proof.* Our proof essentially works via *Yao's minimax lemma*, although we will not call it by name. Instead of constructing a single distribution, we construct a distribution of distributions $\mathcal{F}$ as follows. For any $\boldsymbol{y} \in \{0,1\}^{\mathcal{X}}$, let $\mathcal{D}_{\boldsymbol{y}}$ be the distribution with uniform distribution on $\mathcal{X}$, where each $x \in \mathcal{X}$ has a deterministic label $\boldsymbol{y}(x)$. Clearly, $\boldsymbol{y}$ itself is perfect classifier for $\mathcal{D}_{\boldsymbol{y}}$. Let $\mathcal{F}$ be a uniform distribution over $\left\{\mathcal{D}_{\boldsymbol{y}} : \boldsymbol{y} \in \{0,1\}^{\mathcal{X}}\right\}$. Now, we observe for any algorithm $\mathcal{A}$ that

$$\begin{aligned}
\max_{\boldsymbol{y}\in\{0,1\}^{\mathcal{X}}} \mathbb{E}_{S^m\sim\mathcal{D}_{\boldsymbol{y}}}\left[\mathrm{err}_{\mathcal{D}_{\boldsymbol{y}}}(h_{\mathcal{A}})\right] &\geq \mathbb{E}_{\mathcal{D}_{\boldsymbol{y}}\sim\mathcal{F}}\mathbb{E}_{S^m\sim\mathcal{D}_{\boldsymbol{y}}}\left[\mathrm{err}_{\mathcal{D}_{\boldsymbol{y}}}(h_{\mathcal{A}})\right] \\
&= \mathbb{E}_{X^m\sim\mathrm{Unif}(\mathcal{X})}\mathbb{E}_{\boldsymbol{y}\sim\mathrm{Unif}(\{0,1\}^{\mathcal{X}})}\left[\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\mathbf{1}_{h_{\mathcal{A}}(x)\neq\boldsymbol{y}(x)}\right] \\
&\geq \mathbb{E}_{X^m\sim\mathrm{Unif}(\mathcal{X})}\mathbb{E}_{\boldsymbol{y}\sim\mathrm{Unif}(\{0,1\}^{\mathcal{X}})}\left[\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}-X^m}\mathbf{1}_{h_{\mathcal{A}}(x)\neq\boldsymbol{y}(x)}\right] \\
&= \mathbb{E}_{S^m}\left[\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}-X^m}\mathbb{E}_{y\sim\mathrm{Unif}(\{0,1\})}\mathbf{1}_{h_{\mathcal{A}}(x)\neq\boldsymbol{y}(x)}\right] \geq \frac{|\mathcal{X}|-m}{2|\mathcal{X}|},
\end{aligned}$$

where in the first equality, we use a *double sampling* argument, namely that sampling $\boldsymbol{y}$ uniformly randomly and then choosing $m$ samples from $\mathcal{D}_{\boldsymbol{y}}$ can also be seen as sampling $m$ feature vectors from the uniform distribution over $\mathcal{X}$ and them choosing $\boldsymbol{y}$ uniformly randomly. In the second equality, we use the fact that since components of $\boldsymbol{y}$ are independent, even conditioned on $S^m$, $\boldsymbol{y}(x)$ is a uniformly random binary label for all $x \notin X^m$, on which any predict rule makes error with probability $1/2$. Finally, we note that due to sampling with replacement, $X^m$ captures at most $m$ distinct elements from $\mathcal{X}$. □

In the face of this impossibility, there can be two responses. The more obvious of these developments is to limit the class of distributions under consideration, so $\mathcal{D}$ can no longer be arbitrary. One can hope that for *nice* and *natural* distributions such impossibilities do not arise. For a long as time, this was the only approach to learning, as embodied in classical (especially parametric) statistics. In this vein, one assumes that the true data-generating distribution $\mathcal{D}^*$ belongs to some known class $\{\mathcal{D}_1, ...\mathcal{D}_n\}$, finite here for simplicity of illustration. As more samples are gathered from $\mathcal{D}^*$, one can *identify* the true distribution, at least in a functional sense. A severe disadvantage with this approach is that if $\mathcal{D}^*$ happens to lie outside our considered class, it is unclear how a learning algorithm of this sort performs, or if it converges, or if it does, does the convergent distribution yield a reasonable classifier for the true distribution.

The other recourse, and perhaps *the* defining choice in learning theory, is to redefine the notion of success and seek a different sort of learning guarantee. Instead of limiting $\mathcal{D}^*$, we choose a limited hypothesis class $\mathcal{H} = \{h_1, ...h_n\}$. Instead of succeeding in absolute terms, success of an (agnostic) learning algorithm lies in ensuring that the classifier it produces is almost as good as the best hypothesis in $\mathcal{H}$. The advantage is immediate: by choosing $\mathcal{H}$ to be the set of Bayes-optimal classifiers for $\{\mathcal{D}_1, ...\mathcal{D}_n\}$, one can ensure that the classifier produced is as good as the best classifier for the true data-generating distribution $\mathcal{D}^*$, if $\mathcal{D}^*$ lies in the aforementioned set, thus recovering the classical statistical guarantee. On the other hand, robustness to the latter assumption is built in, insofar that, even if all our models of $\mathcal{D}^*$ were wrong, we still retain performance as good as the best hypothesis in $\mathcal{H}$; thus this approach degrades the right way against mis-specification.

This way of thinking in terms of a relative error guarantee is something even experts in other (classical) fields find hard to accept – although the acceptance is growing by the day – perhaps because hypothesis classes embody solution concepts and do not produce an explicit mechanistic description of how the data was generated. But to the extent one cares about minimizing the loss, this approach cannot be beat.

We begin with the definition of realizable PAC learning that makes the above setting concrete, but also does not quite deliver on what was promised. The *realizability* assumption here is that there exists a perfect classifier in the hypothesis class $\mathcal{H}$, or in other words, the learning guarantee only extends distributions $\mathcal{D}$ for which this condition holds.

**Definition 5** A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is realizable PAC learnable if there exists a sample complexity $m : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{A}$, which for any $\varepsilon, \delta > 0$ and distribution $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$, upon taking $m(\varepsilon, \delta)$ samples produces a classifier $h_{\mathcal{A}} : \mathcal{X} \to \{0,1\}$ such that with probability at least $1 - \delta$, we have

$$\mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}}) \leq \varepsilon,$$

as long as there exists an $h^* \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h^*) = 0$.

The more general model, but one which is also computationally challenging, is agnostic PAC learning, which forgoes the realizability assumption, and gives a relative error guarantee, instead of an absolute one. The sample requirement here is also generally higher, as we will see.

**Definition 6** A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable if there exists a sample complexity $m : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{A}$, which for any $\varepsilon, \delta > 0$ and distribution $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$, upon taking $m(\varepsilon, \delta)$ samples produces a classifier $h_{\mathcal{A}} : \mathcal{X} \to \{0,1\}$ such that with probability at least $1 - \delta$, we have

$$\mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}}) \leq \min_{h^* \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h^*) + \varepsilon.$$

## 4. Finite Classes

As a warm-up, in this section, we consider the task of learning finite classes $\mathcal{H}$. Along the way, we will develop the framework of learning infinite class (not all of are learnable!), which is our ultimate goal. For any $m$-sized sample set $S \subseteq \mathcal{X} \times \mathcal{Y}$, let $\mathrm{err}_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h(x_i), y_i)$ be the empirical error of the hypothesis $h$. We begin with realizable learning.

**Theorem 7** Any finite hypothesis class $\mathcal{H}$ is realizable PAC learnable with

$$m(\varepsilon, \delta) = O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}\right) \text{ samples.}$$

*Proof.* Certifying PAC learnability requires specifying a learning algorithm. We choose the most obvious one, namely, pick $h_{\mathcal{A}} \in \mathcal{H}$ arbitrarily as long as $\mathrm{err}_S(h_{\mathcal{A}}) = 0$. Due to realizability,

generically, such a choice always exists, else our claim is vacuous. What does failure to learn mean? Define $\mathcal{H}_{\text{Bad}} = \{h \in \mathcal{H} : \text{err}_{\mathcal{D}}(h) > \varepsilon\}$. Now, failure is synonymous with $h_{\mathcal{A}}$ in $\mathcal{H}_{\text{Bad}}$, which only happens if there is a hypothesis $h$ in $\mathcal{H}_{\text{Bad}}$ with $\text{err}_S(h) = 0$. Now fix any $h \in \mathcal{H}_{\text{Bad}}$. We have

$$\Pr(\text{err}_S(h) = 0) = \prod_{i=1}^{m} \Pr(h(x_i) = y_i) = (1-\varepsilon)^m,$$

$$\Pr(\text{err}_{\mathcal{D}}(h_A)) \leq \sum_{h \in \mathcal{H}_{\text{Bad}}} \Pr(\text{err}_S(h) = 0) = |\mathcal{H}_{\text{Bad}}|(1-\varepsilon)^m \leq |\mathcal{H}|e^{-\varepsilon m},$$

where we use the inequality $1 + x \leq e^x$ for all $x$, concluding the claim. □

For the agnostic case, we introduce the concept of uniform convergence, which requires that with enough samples, the maximum difference between the population error and the sample error across all hypothesis in the class can be made arbitrarily small. This means that the performance on the sample set transfers to the population. We will reuse this notion for infinite hypothesis classes.

**Definition 8** A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ exhibits uniform convergence with sample complexity $m_{\text{UC}} : (0,1)^2 \to \mathbb{N}$ if, for any $\varepsilon, \delta > 0$, upon taking $m_{\text{UC}}(\varepsilon, \delta)$ samples from any distribution $\mathcal{D}$, supported over $\mathcal{X} \times \mathcal{Y}$, to form $S$, we have with probability at least $1 - \delta$ that

$$\max_{h \in \mathcal{H}} |\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| \leq \varepsilon.$$

We will now see that uniform convergence immediately implies agnostic PAC learnability.

**Theorem 9** If a hypothesis class $\mathcal{H}$ exhibits uniform convergence, then it is agnostic PAC learnable with sample complexity $m(\varepsilon, \delta) = m_{\text{UC}}(\frac{\varepsilon}{2}, \delta)$

*Proof.* Again, we start with the learning algorithm, which picks $h_{\mathcal{A}} \in \arg\min_{h \in \mathcal{H}} \text{err}_S(h)$ arbitrarily. Let $h^* \in \arg\min_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h)$. By uniform convergence, with probability $1 - \delta$, we have

$$\text{err}_{\mathcal{D}}(h_{\mathcal{A}}) \leq \text{err}_{\mathcal{S}}(h_{\mathcal{A}}) + \varepsilon \leq \text{err}_{\mathcal{S}}(h^*) + \varepsilon \leq \text{err}_{\mathcal{D}}(h^*) + 2\varepsilon,$$

completing the proof. □

While in general uniform convergence arguments require some care, for finite classes, uniform convergence follows essentially by a union bound.

**Theorem 10** Any finite hypothesis class $\mathcal{H}$ exhibits uniform convergence with

$$m_{\text{UC}}(\varepsilon, \delta) = O\left(\frac{\log|\mathcal{H}|/\delta}{\varepsilon^2}\right) \text{ samples.}$$

*Proof.* Since $\mathbb{E}_S \text{err}_S(h) = \text{err}_{\mathcal{D}}(h)$ for any fixed hypothesis $h$, we observe that

$$\Pr\left(\max_{h \in \mathcal{H}} |\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| > \varepsilon\right) \leq \sum_{h \in \mathcal{H}} \Pr(|\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| > \varepsilon) \leq 2|\mathcal{H}|e^{-2n\varepsilon^2},$$

where the last inequality follows from the tail bound for averages from the last lecture. □

Combining the previous two results, we get the following corollary concerning the agnostic learning of finite hypothesis classes.

**Corollary 10.1** Any finite hypothesis class $\mathcal{H}$ is agnostic PAC learnable with

$$m(\varepsilon, \delta) = O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right) \text{ samples.}$$

# 5. VC Dimension

Now, we are ready to extend learnability to infinite hypothesis classes. Not all hypothesis classes are learnable. Hence, a key question is to find out when learning is possible for infinite classes, by coming up with an appropriate notion of *size* for hypothesis classes.

**Definition 11** For $C = \{x_1, ...x_m\} \subseteq \mathcal{X}$ of finite size and hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, let

$$\mathcal{H}_C = \{(h(x_1), ...h(x_m)) : h \in \mathcal{H}\}$$

be the set of labelings $\mathcal{H}$ induces on $C$. The VC dimension $\mathrm{VC}(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the size of the largest set with $|\mathcal{H}_C| = 2^{|C|}$, in other words, where all possible labelings are realized by $\mathcal{H}$.

To state this explicitly, to establish that $\mathrm{VC}(\mathcal{H}) = d$ for a class $\mathcal{H}$, we must establish that there exists at least *one* set $C$ of size $d$ where all possible labelings of $C$ are realized, thus, the VC dimension is at least $d$, and further that *all* larger sets have at least one unrealizable labeling, implying the VC dimension is strictly less than $d + 1$.

Let us look at a few examples.

1. For $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\mathbf{1}_{x \le a} : a \in \mathbb{R}\}$ has VC dimension one. Clearly, on $C = \{0\}$, this class realizes a positive label by choose $a = 1$ and a negative label by choosing $a = -1$. Furthermore, for any two point set $C = \{a, b\}$ where $a \le b$ generically, a negative label on $a$ and a positive label on $b$ are not realizable simultaneously. Similarly, $\mathrm{VC}(\{\mathbf{1}_{x \ge a} : a \in \mathbb{R}\}) = 1$.
2. Take $\mathcal{X} = \mathbb{R}$. The VC dimension of $\mathcal{H} = \{\mathbf{1}_{a \le x \le b} : a, b \in \mathbb{R}\}$ is two. It is easy to see that there exists a two-point set, e.g., $\{0, 1\}$, where all possible labelings are realized. For any three-point set $\{a, b, c\}$ with $a \le b \le c$, a negative label in the middle and positive labels at extremities is unrealizable.
3. Take $\mathcal{X} = \mathbb{R}^2$. The VC dimension of all (closed) axis-aligned rectangles is four. Note that not all four point sets, e.g., $\{(0,0), (0,1), (1,0), (1,1)\}$, can be assigned arbitrary labelings, but that's okay, because we just need to show *one* four-point set that can be labeled in all possible ways. Such a set exists, e.g., $\{(1,0), (0,1), (-1,0), (0,-1)\}$. Arguing that every five-point set has an unrealizable labeling is trickier. The cleanest argument here is that the smallest axis-aligned rectangle containing any five-point set is also the smallest axis-aligned

rectangle for some four-point subset of the original set. Hence, the point in the "middle" can not be assigned a label independently of the other four.

# 6. Learning VC Classes

At this point, it might not be obvious why VC dimension is the correct notion of *size* for learning. The theorem on sufficiency will shed some light on this, but far more obvious is the fact that finite VC dimension is necessary for learning.

**Corollary 4.1** For any hypothesis class $\mathcal{H}$ with infinite VC dimension, and learning algorithm $\mathcal{A}$ that produces the classifier $h_{\mathcal{A}}$ after seeing $m$ samples, for any $\varepsilon > 0$, there exists a distribution $\mathcal{D}$ such that $\mathbb{E}[\text{err}_{\mathcal{D}}(h_{\mathcal{A}})] \geq \frac{1}{2} - \varepsilon$ and $\min_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h) = 0$. Thus, a finite VC dimension is necessary for realizable (and hence also agnostic) PAC learning.

*Proof.* Since the VC dimension of $\mathcal{H}$ is unbounded, we can find arbitrarily large subsets $C$ of $\mathcal{X}$ on which $\mathcal{H}$ can realize all possible labelings. We apply Proposition 4 to such a set of size $m/\varepsilon$. □

In fact, from a quick examination of the proof of Proposition 4, we can also see that the sample complexity of realizable and agnostic learning must scale as $\Omega(\text{VC}(\mathcal{H}))$, although the argument does not by itself imply a correct, that is, tight, bound on other parameters.

Now, we will prove that a finite VC dimension is sufficient for learning. We will focus on the agnostic case, which also implies realizable learnability. In fact, we will prove a near-optimal sample complexity bound for agnostic PAC learning. This proof is interesting to the extent that it will involve all little probabilistic tools that we have developed so far, ranging from symmetrization to maximal inequalities and McDiarmid's inequality.

**Theorem 12** Any hypothesis class $\mathcal{H}$ with a finite VC dimension $d$ exhibits uniform convergence with sample complexity

$$m_{\text{UC}}(\varepsilon, \delta) = O\left(\frac{d + \log 1/\delta}{\varepsilon^2}\right).$$

*Proof.* In fact, we will prove a bound of $O((d \log d/\varepsilon + \log 1/\delta)/\varepsilon^2)$ on the sample complexity, which is worse by logarithmic factors, because proving the tight bound, although very much in reach of the present course, is slightly painful.

Fix any hypothesis class $\mathcal{H}$ with VC dimension $d$, and a random sample set $S$ of size $m$ drawn from $\mathcal{D}$. Notice that $S \mapsto \max_{h \in \mathcal{H}} |\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)|$ goes up (or down) by at most one on changing a sample. Hence, by Theorem 3, we get that

$$\Pr\left(\left|\max_{h \in \mathcal{H}} |\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| - \mathbb{E}_S\left[\max_{h \in \mathcal{H}} |\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)|\right]\right| \geq t\right) \leq 2e^{-2mt^2}.$$

To make the right side at most $\delta$, $m$ needs to be at least $\log(2/\delta)/\varepsilon^2$. This explains the second term in the sample complexity.

**Lemma 13** For a sample set of size $m$ and a class of VC dimension $d$, we have that

$$\mathbb{E}_S\left[\max_{h\in\mathcal{H}}|\mathrm{err}_{\mathcal{D}}(h)-\mathrm{err}_S(h)|\right]\leq O\left(\sqrt{\frac{d\log 2m}{m}}\right).$$

To conclude the claim and note the necessity of the first term in the sample complexity, we use the maximal inequality above that specifically deals with VC classes. □

*Proof of Lemma 13.* On the face of it, Lemma 13 looks exactly like the standard maximal inequality. After all, $\mathrm{err}_S(h)-\mathrm{err}_{\mathcal{D}}(h)$ for any fixed $h$ is exactly zero-mean and, being an average over $m$ independent samples, subgaussian with variance proxy $1/m$. The catch is that the standard maximal inequality scales here as $\sqrt{\log|\mathcal{H}|}$, which is just a fancy way of recovering our finite hypothesis results. With some manipulation, we will show the *effective* size of $\mathcal{H}$, one that matters here anyway, is $O(m^d)$, using the lemma below.

**Lemma 14 (Sauer-Shelah-Perles Lemma)** For any hypothesis class $\mathcal{H}$ with VC dimension $d$ and a set $C$ of size $m$ on the same feature space, we have that

$$|\mathcal{H}_C|\leq\sum_{k=0}^{d}\binom{m}{k}.$$

In particular, if $m$ is at least $d$, then $|\mathcal{H}_C|\leq O(m^d)$.

The main trick is to use symmetrization. Observe the following sequence of inequalities.

$$\begin{aligned}
&\mathbb{E}_S\left[\max_{h\in\mathcal{H}}|\mathrm{err}_{\mathcal{D}}(h)-\mathrm{err}_S(h)|\right]\\
&\underset{1}{=}\mathbb{E}_S\left[\max_{h\in\mathcal{H}}|\mathbb{E}_{S'}\mathrm{err}_{S'}(h)-\mathrm{err}_S(h)|\right]\\
&\underset{2}{\leq}\mathbb{E}_{S,S'}\left[\max_{h\in\mathcal{H}}|\mathrm{err}_{S'}(h)-\mathrm{err}_S(h)|\right]\\
&=\mathbb{E}_{S,S'}\left[\max_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}(l(h(x_i'),y_i')-l(h(x_i),y_i))\right|\right]\\
&\underset{3}{=}\mathbb{E}_{S,S'}\left[\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\max_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i(l(h(x_i'),y_i')-l(h(x_i),y_i))\right|\,\middle|\,S,S'\right]\right]\\
&\underset{4}{=}\mathbb{E}_{S,S'}\left[\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\max_{h\in\mathcal{H}_{S\cup S'}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i(l(h(x_i'),y_i')-l(h(x_i),y_i))\right|\,\middle|\,S,S'\right]\right]\\
&\underset{5}{\leq}C\sqrt{\frac{\log\max_{S\subseteq\mathcal{X},|S|\leq 2m}|\mathcal{H}_S|}{m}}\\
&\underset{6}{\leq}O\left(\sqrt{\frac{d\log m}{m}}\right)
\end{aligned}$$

Here (1) introduces $S'$ as $m$ samples, chosen independently from $S$, and (2) follows from Jensen's inequality, while noting that $x \mapsto |x|$ is a convex function, and taking a maximum over a family of functions preserves convexity. Step (3) follows from noting that for any $S, S'$ exchanging the corresponding samples between these at any index $i$ results in a new pair of sets that are equiprobable. The key step is (4), where we narrow $\mathcal{H}$ to $\mathcal{H}_{S \cup S'}$, since the quantity being maximized only depends on the samples via the signs realized by $\mathcal{H}$ on the same set. Step (5) follows by the standard maximal inequality, having fixed $S, S'$. In step (6), we apply Lemma 14.

Thus, we have the desired claim. But it is worth taking a moment to appreciate why all the machinations above were needed. Clearly, replacing the population error by the sample error on newly sampled points was instrumental in ultimately narrowing $\mathcal{H}$ to $\mathcal{H}_{S \cup S'}$, by saying that only the signs realized on some $2m$ points matter. But, one might also be tempted to directly use the maximal inequality at the end of step (2), without introducing Rademacher random variables. This idea fails. While even at the end of step (2) we could have narrowed $\mathcal{H}$ to $\mathcal{H}_{S \cup S'}$, now the index set being maximized over is stochastic, while an implicit promise in the maximal inequality is that the index set is deterministic, or at lease independent of other sources on randomness. Had we conditioned on $S, S'$ at this stage, that would have fixed the stochastic index set, but made the remaining quantity deterministic too, losing the subgaussian character. □

Before proving Lemma 14, let us take a second to review its implication. It says for any class with bounded VC dimension, the number of labelings grows polynomially in the size of the set. If on the other hand, the VC dimension is infinite, then, by definition, there are sets of any needed size where the number of labelings are exponentially many. The lemma rules out all possibilities in the middle, that is, those associated with a superpolynomial but subexponential growth for all sets simultaneously. This is also why VC dimension sharply characterizes learnability. Each class is either learnable, to a nontrivial degree, with a constant number of samples, or unlearnable altogether. There is no in-between.

*Proof of Lemma 14.* The proof proceeds by induction on $m + d$. The base cases can be verified separately. Fix a hypothesis class $\mathcal{H}$ with VC dimension $d$ and a set $C = \{x_1, ...x_m\} \subseteq \mathcal{X}$. Let $C' = C - \{x_1\}$ and $\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H}, h'(x_1) = 1 - h(x_1)\}$. Now, we claim that

$$|\mathcal{H}_C| = |\mathcal{H}_{C'}| + |\mathcal{H}'_{C'}|.$$

In words, all labelings of $C'$ have either a unique extension to $C$ under $\mathcal{H}$, in which case they are counted in the first term, or admit two extensions, with both $\pm$ signs on $x_1$, to $C$, in which case they are accounted for once in the first term and again in the second term.

Let us construct another hypothesis class $\mathcal{H}''$ which is the same as $\mathcal{H}'$ except that $x_1$ is not in its domain. (If this makes you uneasy, it is also fine to assign all of $\mathcal{H}''$ an arbitrary label on $x_1$.) Since $C'$ does not contain $x_1$, $\mathcal{H}'_{C'}$ and $\mathcal{H}''_{C'}$ are identical. Further, we claim that the VC dimension of $\mathcal{H}''$ is at most $d - 1$, because if all labelings of a set $D$ are attained by $\mathcal{H}''$, then $\mathcal{H}$ attains all labelings on $C \cup \{x_1\}$. We then apply the inductive hypothesis on $\mathcal{H}_{C'}$ and $\mathcal{H}''_{C'}$ to get

$$|\mathcal{H}_C| \leq \sum_{k=0}^{d} \binom{m-1}{k} + \sum_{k=0}^{d-1} \binom{m-1}{k} = \binom{m-1}{0} + \sum_{k=1}^{d} \left( \binom{m-1}{k} + \binom{m-1}{k-1} \right) = \sum_{k=0}^{d} \binom{m}{k},$$

where we use the identity that $\binom{m}{k} = \binom{m}{k-1} + \binom{m-1}{k-1}$. The following display completes the final clause, as long as $m \geq d$.

$$|\mathcal{H}_C| \leq \left(\frac{m}{d}\right)^d \sum_{k=0}^{d} \binom{m}{k} \left(\frac{d}{m}\right)^k \leq \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{me}{d}\right)^d$$

□

Using Theorem 9, uniform convergence of finite VC classes implies agnostic PAC learnability.

**Corollary 14.1** Any hypothesis class $\mathcal{H}$ with a finite VC dimension $d$ is agnostic PAC learnable with sample complexity

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

Finally, we remark the sample complexity derived above can be improved to scale as $1/\varepsilon$ in the realizable case. The proof of this result utilizes a nice double sampling argument.

## 6.1. Application: DKW Inequality

Consider the task of estimating the CDF $F$ of a continuous real-valued random variable given $m$ independent samples. A natural choice of the estimator is

$$\hat{F}(t) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{x_i \leq t}$$

Using subgaussian tail bounds, for any fixed $t$, $\Pr\left(|F(t) - \hat{F}(t)| \geq t\right) \leq 2e^{-2mt^2}$. The Dvoeretzky-Keifer-Wolfowitz inequality says the same bound holds uniformly, without needing a union bound.

**Theorem 15 (DKW Inequality)** $\Pr\left(\sup_{t \in \mathbb{R}} |F(t) - \hat{F}(t)| \geq t\right) \leq 2e^{-2mt^2}$

The proof is a simple application of Theorem 12 – this is after all a statement about uniform convergence – if one were willing to ignore the precise constants, since the VC dimension of the indicator variables of $\{(-\infty, a] : a \in \mathbb{R}\}$ is one. In terms of history, the DKW inequality predates VC theory by a good couple of decades.

## 6.2. Application: Learning Quantiles

Consider an inventory replenishment problem for nondurable goods, typically termed the newsvendor problem. At the end of each day, a store owner owners orders $a$ goods to delivered the next morning. The stochastic demand $X \in [0, 1]$ is realized the next day. The resultant loss is

$$l(X, a) = \rho[X - a]_+ + (1 - \rho)[a - X]_+,$$

which imposes unequal penalties for over and under meeting the demand. Here $[X - a]^+ = \max\{X - a, 0\}$. We are interested in pick an action that minimizes the expected cost $\mathbb{E}_X[l(X, a)]$. Since $a \mapsto l(X, a)$ is convex, the subgradient is

$$\frac{d}{da}\mathbb{E}_X[l(X,a)] = \mathbb{E}_X[\rho \mathbf{1}_{X>a} - (1-\rho)\mathbf{1}_{X\le a}] = \rho \Pr(X > a) - (1-\rho)\Pr(X \le a),$$

and hence the optimal choice $a$ is the bottom $\rho$-quantile of $\mathcal{D}$.

What can we do if $\mathcal{D}$ is unknown and instead we observe $m$ samples from $\mathcal{D}$? By the DKW inequality, we can choose $\hat{a}$ to be the bottom $\rho$-quantile of observed samples, and guarantee that $F(\hat{a}) = \rho \pm O(1/\sqrt{m})$ with probability 0.99. An exercise in integration by parts gives (Verify!)

$$\mathbb{E}_X[l(X,a)] = \rho \int_a^1 \Pr(X \ge x)dx + (1-\rho)\int_{-1}^a \Pr(X \le x)dx.$$

Now, we can see how good $\hat{a}$ is by observing

$$\mathbb{E}_X[l(X,\hat{a})] - \mathbb{E}_X[l(X,a^*)] = \int_{\hat{a}}^{a^*} (\rho - F(x))dx \le O\left(\frac{1}{\sqrt{m}}\right).$$

# 7. Algorithmic Stability

The plan here was to carve an alternative path to generalization and learning via (uniform) stability of the learning algorithm. Instead of focusing on the characteristics of the hypothesis class, this is an algorithm-centric approach. Being short on time, we will instead consider a simple example that deals with Leave-One-Out (LOO) stability and proves generalization in expectation.

**Definition 16** The Leave-One-Out (LOO) stability $\Delta(\mathcal{A}, S^{m+1})$ of a learning algorithm $\mathcal{A}$ with respect to a sample $S = \{(x_i, y_i)\}_{i\in[m+1]}$ of size $m+1$ is defined as

$$\Delta(\mathcal{A}, S) = \frac{1}{m+1}\sum_{i=1}^{m+1}\Big(l\Big(h_{\mathcal{A}_i}(x_i), y_i\Big) - l(h_{\mathcal{A}}(x_i), y_i)\Big),$$

where the algorithm $\mathcal{A}$ receives $S^{m+1}$ as its input and $\mathcal{A}_i$ receives all samples but $(x_i, y_i)$.

The following is a funny sort of in-expectation generalization guarantee that compares the population error of an algorithm that is given $m$ samples to the in-sample error on $m+1$ samples, instead of $m$ as one would expect. Nevertheless, this will suffice for our application. In fact, for reasonable learning algorithms, for example, if one chooses a hypothesis with the smallest error on training data, this upper bounds the usual generalization error in expectation, up to a small $1/m$ additive term (Verify this!).

**Theorem 17** For any distribution $\mathcal{D}$, we have that

$$\mathbb{E}_{S^m}[\mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}'})] = \mathbb{E}_{S^{m+1}}[\mathrm{err}_{S^{m+1}}(h_{\mathcal{A}})] + \mathbb{E}_{S^{m+1}}[\Delta(\mathcal{A}, S^{m+1})],$$

where $\mathcal{A}'$ and $\mathcal{A}$ receive $S^m$ and $S^{m+1}$ as their inputs, respectively.

*Proof.* The proof is a simple consequence of the definition.

$$\begin{aligned}
\mathbb{E}_{S^{m+1}}[\Delta(\mathcal{A}, S^{m+1})] &= \frac{1}{m+1}\sum_{i=1}^{m+1}\mathbb{E}_{S^{m+1}}\left[l\big(h_{\mathcal{A}_i}(x_i), y_i\big)\right] - \mathbb{E}_{S^{m+1}}\left[\frac{1}{m+1}\sum_{i=1}^{m+1} l(h_{\mathcal{A}}(x_i), y_i)\right] \\
&= \frac{1}{m+1}\sum_{i=1}^{m+1}\mathbb{E}_{S^m}\mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h_{\mathcal{A}'}(x), y)] - \mathbb{E}_{S^{m+1}}[\mathrm{err}_{S^{m+1}}(h_{\mathcal{A}})] \\
&= \mathbb{E}_{S^m}[\mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}'})] - \mathbb{E}_{S^{m+1}}[\mathrm{err}_{S^{m+1}}(h_{\mathcal{A}})]
\end{aligned}$$

□

As an application, consider the task of learning $d$-dimensional axis-aligned rectangles in the realizable case. Our learning algorithm in this case simply outputs the smallest axis-aligned rectangle containing all positive examples. The empirical error given any number of samples is zero. We will soon see that the worst-case LOO stability over any $m+1$ example is at most $\frac{2d}{m+1}$. Hence, this algorithm given $m$ samples has a population error of at most $\frac{2d}{m+1}$, matching the sample complexity of $O(d/\varepsilon)$ we would get from the VC approach. To see this bound on the LOO stability, note that the smallest rectangle is supported by at least one sample on every one of its $2d$ facets, and a facet shifts only if all samples supporting it are deleted, and thus, $\mathcal{A}_i$ and $\mathcal{A}$ are identical on all except at most 4 indices.

# 8. References


1. Primary reference: Chapters 2–6 and 28 in Shalev-Shwartz and Ben-David.
2. Alternative: Chapters 2 and 3 in Mohri, Rostamizadeh, Talwalkar.
3. Alternative: Chapter 5 in Blum, Hopcroft, Kannan.
4. Bounded Differences Inequality: Chapters 5 and 6 in Dubhashi and Panconesi.