LECTURER: ELAD HAZAN NAME: FERMI MA AND DANIEL SUO

1 Proving the Fundamental Theorem of Statistical Learning

In this section, we prove the following:

Theorem 1.1 (Fundamental Theorem of Statistical Learning)**.** *The hypothesis class $\mathcal{H}$ is learnable if and only if the VC-dimension of $\mathcal{H}$, denoted d, is less than ∞. If $\mathcal{H}$ is learnable, then the sample complexity is given by $m(\varepsilon, \delta) \sim \frac{d}{\varepsilon} \log \frac{1}{\varepsilon\delta}$.*

We refer to the Boolean mapping problem (i.e., learn some concept $C : \mathcal{X} \to \{0, 1\}$) throughout the proof. Recall the following definitions from previous lectures:

Definition 1.2 (VC-dimension)**.** Define a hypothesis class $\mathcal{H}$ as a class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. We say that the restriction of $\mathcal{H}$ to C, $\mathcal{H}_C$, is the set of functions from C to $\{0, 1\}$ we can derive from $\mathcal{H}$. In other words,

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) : h \in \mathcal{H}\}$$

or the set of vectors if we evaluate each of $h \in \mathcal{H}$ on each element in C. If $\mathcal{H}_C$ is the set of all functions from C to $\{0, 1\}$, we say that $\mathcal{H}$ *shatters* C. In a previous lecture, we showed that this equivalent to $|\mathcal{H}_C| = 2^{|C|}$ (i.e., have a hypothesis for all possible configurations of C, where each $c_i \in \{0, 1\}$.

The VC-dimension of a hypothesis class $\mathcal{H}$ is the largest set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter any size C, we say that the VC-dimension of $\mathcal{H}$ is infinity.

Definition 1.3 (Growth function $\tau_{\mathcal{H}}(m)$)**.** Define a function $\tau_{\mathcal{H}}(m) : \mathbb{N} \to \mathbb{N}$ for hypothesis class $\mathcal{H}$ as follows:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

The growth function $\tau_{\mathcal{H}}(m)$ counts the number of *different* mappings $C \to \{0, 1\}$ we can generate if we restrict $\mathcal{H}$ to C if $|C| = m$.

Notice that if VC-DIMENSION$(\mathcal{H}) = d$, then for all m less than d, we have $\tau_{\mathcal{H}}(m) = 2^m$ because $\mathcal{H}$ shatters all $C \subset \mathcal{X}$ of size less than or equal to d. Sauer's lemma, which we proved in the previous lecture, shows that if m is greater than d, $\tau_{\mathcal{H}}(m) = \mathcal{O}(m^d)$.

1.1 Overall strategy

1. We already proved one direction of Theorem 1.1 via the No-Free-Lunch theorem: if the VC-dimension of our hypothesis class $\mathcal{H}$ is infinite, then $\mathcal{H}$ is not learnable. In other words, if $\mathcal{H}$ is learnable, then VC-DIMENSION$(\mathcal{H})$ is finite.
2. To show that a hypothesis class $\mathcal{H}$ with finite VC-dimension is learnable, we first proved Sauer's Lemma. Then, we use the two-sample trick (because we cannot take infinitely-sized samples) to show

that the error on one sample cannot be substantially different from the error on the other, rather than rely on generalization arguments. We will make use of the following two events:

(a) Let A be the event that given a sample $S \sim \mathcal{D}^m$, there exists some $h \in \mathcal{H}$ such that the error on the sample $err_S(h)$ is 0 and the generalization error $err(h)$ is greater than some $\varepsilon > 0$. In other words:

$$\mathbf{Pr}[A] \triangleq \mathbf{Pr}[\exists h \in \mathcal{H} \text{ s.t. } err_S(h) = 0 \wedge err(h) > \varepsilon \mid S \sim \mathcal{D}^m]$$

We will prove that $\mathbf{Pr}[A] \leq \tau_{\mathcal{H}}(m) e^{-\frac{m\varepsilon}{2}}$.

(b) Let B be the event that given two samples, $S, S' \sim \mathcal{D}^m$, there exists some $h \in \mathcal{H}$ such that the error on the first sample, S is 0, and the error on the second sample S' is $\varepsilon/2$ for some $\varepsilon > 0$. In other words:

$$\mathbf{Pr}[B] \triangleq \mathbf{Pr}\left[\exists h \in \mathcal{H} \text{ s.t. } err_S(h) = 0 \wedge err_{S'}(h) > \frac{\varepsilon}{2} \mid S, S' \sim \mathcal{D}^m\right]$$

Claim 1.4. $\mathbf{Pr}[A] \leq 2\mathbf{Pr}[B]$.

Proof: By the law of total probability, we can write

$$\begin{aligned}\mathbf{Pr}[B] &= \mathbf{Pr}[B|A]\mathbf{Pr}[A] + \mathbf{Pr}[B|\neg A]\mathbf{Pr}[\neg A] \\ &\geq \mathbf{Pr}[B|A]\mathbf{Pr}[A]\end{aligned}$$

To prove the claim, it is sufficient to show $\mathbf{Pr}[B|A] \geq 1/2$. Let $S' = \{x_1, \ldots, x_m\} \sim \mathcal{D}^m$. Using the hypothesis $h \in \mathcal{H}$ defined for event A, we know that $err(h) > \varepsilon$ and $err_S(h) = 0$ by definition. Then, with

$$z_i \triangleq \begin{cases} 1 & loss(h, x_i) = 1 \\ 0 & o/w \end{cases}$$

and

$$\begin{aligned} Y &\triangleq \frac{1}{m}\sum_{i=1}^{m} z_i = err_{S'}(h) \\ \mathbb{E}Y &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} z_i\right] = \mathbb{E}[err_{S'}(h)] && \text{by definition of sample error} \\ &= err(h) > \varepsilon \end{aligned}$$

we see that $1 - \mathbf{Pr}[B|A] \leq \mathbf{Pr}[|Y - \mathbb{E}Y| > \frac{\varepsilon}{2}]$ because in order for $|Y - \mathbb{E}Y| > \frac{\varepsilon}{2}$, Y must deviate from its mean, the generalization error, by more than $\varepsilon/2$. Using the Chernoff bound, we see that

$$\mathbf{Pr}\left[|Y - \mathbb{E}Y| > \frac{\varepsilon}{2}\right] \leq 2e^{-m\frac{\varepsilon}{2}} \lll \frac{1}{2}$$

for any m we might choose (note that $m \sim \mathcal{O}(1/\varepsilon)$), thus completing the proof of Claim 1.4. □

Continuing our proof that $\mathbf{Pr}[A] \leq \tau_{\mathcal{H}}(m) e^{-\frac{m\varepsilon}{2}}$, we now seek to show $\mathbf{Pr}[B] \leq \tau_{\mathcal{H}}(2m) e^{-m\varepsilon/2}$. We employ the following symmetry argument that states that the probability of two samples S and S' being so different that some hypothesis performs much better on one versus the other is small.

Construct two new sets T and T' by randomly partitioning $S \cup S'$ into equal sets (note, this proof still works if $S \cap S' \neq \emptyset$, but need to use multisets; in practice, however, S and S' are nearly always disjoint because

the domain is very large). Now, define a distribution $\mathcal{T}$ over choices of T and T' and B_T as the event B, but with T and T' instead of S and S':

$$\mathbf{Pr}[B_T] \triangleq \mathbf{Pr}\left[\exists h \in \mathcal{H} \text{ s.t. } err_T(h) = 0 \wedge err_{T'}(h) > \frac{\varepsilon}{2} \mid T, T' \sim \mathcal{D}^m\right]$$

We claim that $\mathbf{Pr}_{S,S'}[B] = \mathop{\mathbb{E}}_{S,S'}\left[\mathop{\mathbf{Pr}}_{T,T'}[B_T|S,S']\right]$. We take this detour because it is much easier to analyze $\mathop{\mathbf{Pr}}_{T,T'}[B_T|S,S']$ (i.e., what is the probability that one set has all errors and the other has none).

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{T,T'}[B_T|S,S'] &= \mathop{\mathbf{Pr}}_{T,T'}\left[\exists h \in \mathcal{H} \text{ s.t. } err_T(h) = 0 \wedge err_{T'}(h) > \frac{\varepsilon}{2} \mid S, S' \sim \mathcal{D}^m\right] \\
&\leq |\mathcal{H}_{S\cup S'}| \max_h \mathop{\mathbf{Pr}}_{T,T'}\left[err_T(h) = 0 \wedge err_{T'}(h) > \frac{\varepsilon}{2}\right] && \text{by union bound} \\
&\leq \tau_{\mathcal{H}}(2m) \max_h \mathop{\mathbf{Pr}}_{T,T'}\left[err_T(h) = 0 \wedge err_{T'}(h) > \frac{\varepsilon}{2}\right]
\end{aligned}$$

If we let k be the number of errors that h makes on $S \cup S'$ and $k < \frac{\varepsilon m}{2}$, then there simply aren't enough errors to go around! In this case,

$$\mathop{\mathbf{Pr}}_{T,T'}[err_T(h) = 0 \wedge err_{T'}(h) > \varepsilon/2|S,S'] = 0$$

If $k \geq \frac{\varepsilon m}{2}$, then this probability is bounded above by 2^{-k} because all errors must land in T' so that $err_T(h) = 0$ (probability of k balls landing in the first m of $2m$ bins).

Collecting these results gives us

$$\mathbf{Pr}[A] \leq 2\mathbf{Pr}[B] \leq 2\mathbb{E}[\tau_{\mathcal{H}}(2m)2^{-k}],\ k \geq \frac{\varepsilon m}{2}$$

From this last expression, we have $2\tau_{\mathcal{H}}(2m)e^{-\varepsilon m/2}$, which we set to be less than δ for $m = \frac{\log \tau_{\mathcal{H}}(2m)}{\varepsilon} \log \frac{1}{\delta}$. If we use the fact that $\log \tau_{\mathcal{H}}(m) \sim d \log m$ (as given in Sauer's Lemma), then we have

$$\begin{aligned}
m &= 10\frac{d \log \tau_{\mathcal{H}}(2m)}{\varepsilon} \log \frac{1}{\delta} \\
&= \Omega\left(\frac{d}{\varepsilon} \log \frac{d}{\varepsilon\delta}\right)
\end{aligned}$$

A less naive approach can remove the d from within the log. This final step completes the proof of Theorem 1.1. □

2 Methods for bounding generalization error

So far, we have learned about two ways to bound generalization error. Today, we will learn about a third.

1. **VC-dimension**: the topic of discussion for the last two lectures. This the more general of the two approaches we have seen so far.
2. **Online2Batch**: this approach is less general because it requires a convex structure for the problem, but typically much more efficient.
3. **Rademacher complexity**: we will introduce this method today. Note that the computation of Rademacher complexity is NP-hard for some hypothesis classes.

Definition 2.1 (Rademacher variables)**.** Let $\boldsymbol{\sigma}$ be a vector whose elements are chosen independently and uniformly from $\{-1, +1\}$. That is, with probability $1/2$ a given element is either -1 or 1.

Definition 2.2 (Empirical Rademacher complexity)**.** Given a sample $S = \{x_1, \ldots, x_m\}$ chosen from $\mathcal{D}^m$, define the empirical Rademacher complexity $\hat{\mathfrak{R}}_S(\mathcal{H})$ as

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i loss(h(x_i))\right]$$

Definition 2.3 (Rademacher complexity)**.** For some $m \geq 1$, let the Rademacher complexity of $\mathcal{H}$ be the expectation of the empirical Rademacher over all samples S of size m drawn from some distribution $\mathcal{D}$.

$$\mathfrak{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathfrak{R}}_S(\mathcal{H})\right]$$

To give some intuition, we consider the value of $\hat{\mathfrak{R}}_S(\mathcal{H})$ when $\mathcal{H}$ shatters S. Because we have $h \in \mathcal{H}$ that can generate any mapping of S to $\{-1, +1\}^m$, select the one that maximizes the sum (i.e., -1 when σ_i is -1, 1 when σ_i is 1). This way, regardless of what we select for $\boldsymbol{\sigma}$, we have 1 inside the expectation. This measure captures the dimensionality of a hypothesis class very well because in a way, it is proportional to the VC-dimension. We can show that $\mathfrak{R}_m(\mathcal{H})$ is sort of bounded by the VC-dimension or $\tau_m(\mathcal{H})/m$ (it could be much smaller since the majority of the probability mass might not be over the shattered set).

Theorem 2.4. *With probability at least $1-\delta$, we have for all m and for all $h \in \mathcal{H}$,*

$$err(h) \leq err_S(h) + 2\mathfrak{R}_m(\mathcal{H}) + 3\sqrt{\frac{\log(1/\delta)}{m}}$$

This relation holds for agnostic learning as well since we do not assume realizability.

Proof: First, define a function $\Phi(S) = \sup_{h\in\mathcal{H}}\{err(h) - err_S(h)\}$. If S and S' differ on only one sample, $x_i \in S$ and $x_i' \in S'$, then

$$\begin{aligned}
|\Phi(S) - \Phi(S')| &= |\sup_{h\in\mathcal{H}}\{err(h) - err_S(h)\} - \sup_{h\in\mathcal{H}}\{err(h) - err_{S'}(h)\}| && \text{by definition}\\
&\leq |\sup_{h\in\mathcal{H}}\{err_{S'}(h) - err_S(h)\}| && \text{sub-additivity of supremum}\\
&= \left|\sup_{h\in\mathcal{H}}\left\{\frac{loss(h(x_i')) - loss(h(x_i))}{m}\right\}\right| && \text{only one sample different}\\
&= \frac{1}{m}
\end{aligned}$$

By McDiarmid's inequality, for any $\delta > 0$, with probability at least $1-\delta$, we get the following bounds:

$$|\Phi(S) - \mathbb{E}[\Phi(S)] < \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Therefore, to prove the theorem, we need only show that $\mathbb{E}[\Phi(S)] \leq 2\mathfrak{R}_m(\mathcal{H})$.

$$
\begin{aligned}
\mathbb{E}[\Phi(S)] &= \mathop{\mathbb{E}}_{S}[\sup_{h\in\mathcal{H}}\{err(h) - err_S(h)\}] && \text{by definition} \\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{h\in\mathcal{H}}\left\{\mathop{\mathbb{E}}_{S'}\left[err_{S'}(h) - err_S(h)\right]\right\}\right] && \text{expectation of i.i.d. sample error} \\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left\{err_{S'}(h) - err_S(h)\right\}\right] && \text{Jensen's and convexity of supremum} \\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\left[loss(h,x_i') - loss(h,x_i)\right]\right\}\right] && \text{by definition} \\
&= \mathop{\mathbb{E}}_{S,S',\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[loss(h,x_i') - loss(h,x_i)\right]\right\}\right] && \boldsymbol{\sigma} \text{ doesn't change } \mathbb{E} \\
&\leq \mathop{\mathbb{E}}_{S',\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[loss(h,x_i')\right]\right\}\right] \\
&\quad - \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[loss(h,x_i)\right]\right\}\right] && \text{sub-additivity of supremum} \\
&= 2\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[loss(h,x_i)\right]\right\}\right] && \sigma_i \text{ and } -\sigma_i \text{ distributed same way} \\
&= 2\mathfrak{R}_m(\mathcal{H})\ \square
\end{aligned}
$$

Using the Massart's Lemma, we have

$$\mathfrak{R}_m(\mathcal{H}) \leq \sqrt{\frac{2\log\tau_{\mathcal{H}}(m)}{m}}$$

Plugging this into Theorem 2.4, we get the desired bound. However, this is a much stronger claim because Rademacher complexity is a claim about averages whereas VC-dimension is a claim about worst cases.