**CSCI699: Theory of Machine Learning** Fall 2021

# Lecture 6: VC Theorem and Rademacher Complexity

*Instructor: Vatsal Sharan* *Scribe: Jesse Zhang*

### Today

- Finish VC Theorem proof
- Introduce Rademacher complexity

Last time, we introduced the Vapnik-Chervonenkis (VC) dimension and the VC Theorem. Today, we will finish its 4-step proof and introduce Rademacher complexity.

## 1 VC Theorem

**Theorem 1** (VC Theorem)**.** *Let $\mathcal{H}$ be a hypothesis class with* $\mathrm{VCdim}(\mathcal{H}) = d < \infty$*. Then there is an absolute constant $c > 0$ such that $\mathcal{H}$ has the uniform convergence property with,*

$$n_{\mathcal{H}}^{VC}(\epsilon, \delta) = c \cdot \frac{d \cdot \log(d/\epsilon) + \log(1/\epsilon)}{\epsilon^2}.$$

**Corollary 2.** *$\mathcal{H}$ is agnostically-PAC learnable with $\mathcal{O}\left(\frac{d \cdot \log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$ samples.*

*Proof.* Theorem 1 Proof Outline:

1) For any set $C \subseteq \mathcal{X}$, the effective size of the restriction of $\mathcal{H}$ on $C$, denoted $\mathcal{H}_C$, is approximately $|C|^d$ ($|\mathcal{H}_C| \approx |C|^d$).
2) This small "effective size" will be good for when using the union bound to get the VC theorem result.

**Definition 3** (Restriction)**.** *The restriction of class $\mathcal{H}$ to a set of examples $C = \{c_1, ..., c_n\} \in \mathcal{X}$ is a subset of $\{0, 1\}^{|C|}$ given by $\mathcal{H}_C = \{(h(c_1), ..., h(c_n)) : \forall h \in \mathcal{H}\}$.*

Step 1: Polynomial growth of $\mathcal{H}_C$.

We saw this in the last lecture, so here we will just re-state the definition of the growth function and Sauer's Lemma without re-proving it.

**Definition 4** (Growth function)**.** *The growth function of $\mathcal{H}$, $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as*

$$\tau_{\mathcal{H}} = \max_{C \subseteq \mathcal{X}, |C| = n} |\mathcal{H}_C|.$$

**Lemma 5** (Sauer's Lemma)**.** $\forall n$, $\mathrm{VCdim}(\mathcal{H}) = d$,

$$\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

For $n > d+1$, this implies:

$$\tau_{\mathcal{H}}(n) \leq \left(\frac{n \cdot e}{d}\right)^{d}.$$

Step 2: Symmetrization

**Lemma 6.** *For a class $\mathcal{H}$ with growth function $\tau_{\mathcal{H}}$,*

$$\mathbb{E}_{S \sim \mathcal{D}^n}\left[\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right|\right] \leq \sqrt{\frac{2 \cdot \log(2 \cdot \tau_{\mathcal{H}}(2n))}{n}}.$$

*Proof. (Lemma 6):* We will use the idea of symmetrization.

Let $S' = \{(x_i', y_i'), i \in [n]\}$ be a training set sample indentically distributed as $S$.

Note that $\mathbb{E}_{S'}\left[\hat{R}_{S'}(h)\right] = R(h)$.

Therefore,

$$\mathbb{E}_S\left[\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right|\right] = \mathbb{E}_S\left[\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{S'}\left[\hat{R}_{S'}(h)\right] - \hat{R}_S(h) \right|\right]. \tag{1}$$

Now, fix $S$.

**Claim 7.** $\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{S'}\left[\hat{R}_{S'}(h)\right] \right| \leq \mathbb{E}_{S'} \sup_{h \in \mathcal{H}} \left| \hat{R}_{S'}(h) \right|$.

*Proof. (Claim 7)*:

This follows from the fact that $|\cdot|$ is a convex function, and sup / max of convex functions is convex.

Therefore, $\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{S'}\left[\hat{R}_{S'}(h)\right] \right|$ is a convex function of $\hat{R}_{S'}(h)$.

By applying Jensen's inequality ($f(\mathbb{E}(X)) \leq \mathbb{E}(f(x))$ if $f$ convex), the claim follows.

□

Using Claim 7 and combining with Eq. 1 and pulling the expectation out, we have

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right|\right] &\leq \mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}} \left| \hat{R}_{S'}(h) - \hat{R}_S(h) \right|\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\}\right) \right|\right].
\end{aligned}$$

Now, let $\sigma_{1:n} = \{\sigma_1, ..., \sigma_n\}$ be independent Rademacher random variables, i.e. $\sim \text{Unif}(\{\pm 1\})$. Since $(x_i, y_i), (x'_i, y'_i)$ are i.i.d.,

$$\mathbb{1}\{h(x'_i) \neq y'_i\} - \mathbb{1}\{h(x_i) \neq y_i\} \sim \mathbb{1}\{h(x_i) \neq y_i\} - \mathbb{1}\{h(x'_i) \neq y'_i\}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \right] &\leq \mathbb{E}_{\sigma_{1:n}} \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left( \mathbb{1}\{h(x'_i) \neq y'_i\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right| \\
&= \mathbb{E}_{S,S'} \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left( \mathbb{1}\{h(x'_i) \neq y'_i\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|.
\end{aligned}$$

Now fix both $S, S'$ and let $C$ be the set of examples appearing in $S \cup S'$ (both of them). Note that $|C| \leq 2n$ as there can be some overlap between $S, S'$.

The key idea here is that we can replace the supremem over the (possibly infinite) set $\mathcal{H}$ by the maximum over the discrete restriction $\mathcal{H}_C$, as all possible labelings for all training examples from both $S, S'$ are included in $\mathcal{H}_C$. Thus,

$$\begin{aligned}
\mathbb{E}_S \left[ \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \right] &\leq \mathbb{E}_{S,S'} \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left( \mathbb{1}\{h(x'_i) \neq y'_i\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right| \\
&= \mathbb{E}_{S,S'} \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left( \mathbb{1}\{h(x'_i) \neq y'_i\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|.
\end{aligned}$$

Now denote $\theta_h = \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left( \mathbb{1}\{h(x'_i) \neq y'_i\} - \mathbb{1}\{h(x_i) \neq y_i\} \right)$. We shorten the above to

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \leq \mathbb{E}_{S,S'} \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h|. \tag{2}$$

Now we want to bound $\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h|$ in Eq. 2. To do this, we will prove a bound regarding the max of sub-Gaussian variables and also show that $\theta_h$ is sub-Gaussian.

**Lemma 8** (Max of sub-Gaussians). *If $(x_1, ..., x_m)$ are mean 0 $\sigma-$sub-Gaussians (not necessarily independent), then*

$$\mathbb{E} \max_i x_i \leq \sigma \sqrt{2 \log(m)}.$$

*Proof.* *(Lemma 8):*

$$
\begin{aligned}
\mathbb{E}\max_i x_i &= \frac{1}{\lambda}\log\exp\left(\lambda\mathbb{E}\left[\max_i x_i\right]\right) \quad \forall\lambda \\
&\leq \frac{1}{\lambda}\log\mathbb{E}\left[\exp(\lambda\max_i x_i)\right] \quad \text{(Jensen's)} \\
&\leq \frac{1}{\lambda}\log\mathbb{E}\left[\sum_{i=1}^{m}\exp(\lambda x_i)\right] \\
&= \frac{1}{\lambda}\log\left(\sum_{i=1}^{m}\mathbb{E}\left[\exp\left(\lambda x_i\right)\right]\right) \\
&\leq \frac{1}{\lambda}\log\left(\sum_{i=1}^{m}\exp(\frac{\lambda^2\sigma^2}{2})\right) \quad \text{(sub-Gaussian definition)} \\
&\leq \frac{\sigma}{\sqrt{2\log(m)}}\log\left(\sum_{i=1}^{m}\exp\left(\log m\right)\right) \quad \text{by setting } \lambda = \frac{\sqrt{2\log(m)}}{\sigma} \\
&= \sigma\sqrt{2\log m}.
\end{aligned}
$$

□

**Claim 9.** *$\theta_h$ is sub-Gaussian with parameter $\frac{1}{\sqrt{n}}, \mathbb{E}[\theta_h] = 0$.*

*Proof. (Claim 9)*:

Remember that $\theta_h = \sum_{i=1}^{n}\frac{\sigma_i}{n}\left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\}\right)$. Thus,

$$
\mathbb{E}\left[\theta_h\right] = \sum_{i=1}^{n}\frac{\mathbb{E}\left[\sigma_i\right]}{n}\left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\}\right) = 0 \quad \text{as } \mathbb{E}[\sigma_i] = 0.
$$

Now we show that $\theta_h$ is sub-Gaussian:

$$
\theta_h = \sum_{i=1}^{n}\underbrace{\frac{\sigma_i}{n}\left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\}\right)}_{\text{each term is sub-Gaussian with parameter } \frac{1}{n}}.
$$

The above is due to how Rademacher RV's are sub-Gaussian with parameter 1, and each $\sigma_i$ is being further multiplied by $\pm 1$ which does not change its sub-Gaussianaeity.

Therefore using the sub-Gaussian sum corrolary in lecture 4, $\theta_h$ is sub-Gaussian with parameter $\left(\sum_{i=1}^{n}\frac{1}{n^2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{n}}$.

□

Now we can finally bound $\mathbb{E}_{\sigma_{1:n}}\max_{h\in\mathcal{H}_C}|\theta_h|$ in Eq. 2.

**Claim 10.** $\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h| \leq \frac{1}{\sqrt{n}} \sqrt{2\log(2|\mathcal{H}_C|)}$

*Proof. (Claim 10)*:

$$\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h| = \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} \max\{\theta_h, -\theta_h\}.$$

Recall that if $\theta_h$ is sub-Gaussian then $-\theta_h$ is also sub-Gaussian. Thus we have the max over $2|\mathcal{H}_C|$ sub-Gaussian variables with the same parameter. Thus,

$$\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h| \leq \frac{1}{\sqrt{n}} \sqrt{2\log(2|\mathcal{H}_C|)},$$

by combining Claim 9 and Lemma 8.

□

In summary, we have now shown that the right hand side of Eq. 2, $\mathbb{E}_{S,S'}\left[\mathbb{E}_{\sigma_{1:n}} \max |\theta_h|\right]$ is bounded by $\sqrt{\frac{2 \cdot \log(2|\mathcal{H}_C|)}{n}}$. Thus, by plugging into Eq. 2,

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \leq \sqrt{\frac{2 \cdot \log(2|\mathcal{H}_C|)}{n}}. \tag{3}$$

Note that $|\mathcal{H}_C| \leq \tau_{\mathcal{H}}(2n)$ since $|C| \leq 2n$ and we can finally finish the proof of Lemma 6 by plugging in $\tau_{\mathcal{H}}(2n)$ for $|\mathcal{H}_C|$.

□

Step 3: McDiarmid's Inequality

Define

$$f(S) = \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right|.$$

Observe that $f(S)$ satisfies the bounded differences property with constant $\frac{1}{n}$ (changing $(x_i, y_i)$ can only change $\hat{R}_S(h)$ by $\frac{1}{n}$ for any $h \in \mathcal{H} \rightarrow$ the max also changes by at most $\frac{1}{n}$).

Using McDiarmid's, we get that

$$P\left[f(S) - \mathbb{E}\left[f(S)\right] > t\right] \leq 2\exp(-2nt^2).$$

If we choose $t = \sqrt{\frac{\log(2/\delta)}{2n}}$ to get the failure probability $\delta$, then with probability $1 - \delta$,

$$f(S) < \mathbb{E}\left[f(S)\right] + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Now plug in Lemma 1 to replace $\mathbb{E}\left[f(S)\right]$, replace $f(S)$, and we get that

$$\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| < \sqrt{\frac{2\log(2\tau_H(2n))}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{4}$$

Step 4: Finish the VC theorem proof

Using Sauer's lemma, for $n > d+1$, $\tau_{\mathcal{H}}(n) \leq \left(\frac{ne}{d}\right)^d$.

Plugging this into Eq. 4, with probability $(1-\delta)$,

$$\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \leq \sqrt{\frac{2 \cdot d \log(2ne/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{5}$$

Therefore, for $n \geq \mathcal{O}\left(\frac{d \log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$, the right hand side $\leq \epsilon$, implying the uniform convergence property for hypothesis classes $\mathcal{H}$ with finite $\mathrm{VCdim}(\mathcal{H}) = d$. □

Exercise: Show this explicity from Eq. 5.

## 2 Rademacher Complexity

Let us recall the proof of the VC theorem. We wanted to bound

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right|.$$

This quantity is called an "empirical process." Empirical process theory studies such quantities.

Let's use symmetrization to bound this empirical process (without the absolute values):

$$\begin{aligned}
\mathbb{E}_S \sup_{h \in \mathcal{H}} \left( R(h) - \hat{R}_S(h) \right) &\leq \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left( \hat{R}_{S'}(h) - \hat{R}_S(h) \right) \\
&= \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \left( \mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \\
&= \mathbb{E}_{\sigma_{1:n}} \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \\
&\leq \mathbb{E}_{S'} \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \mathbb{1}\{h(x_i') \neq y_i'\} \right) + \\
&\qquad \mathbb{E}_S \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (-\sigma_i) \left( \mathbb{1}\{h(x_i) \neq y_i\} \right) \\
&\leq 2\mathbb{E}_S \underbrace{\mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \mathbb{1}\{h(x_i) \neq y_i\} \right)}_{\text{Rademacher Complexity}} \quad \text{(i.i.d.)}.
\end{aligned}$$

As we will now see, we can bound the expected maximum error between training time and testing time of our hypothesis class by the "Rademacher Complexity."

**Definition 11** (Rademacher Complexity)**.** *Let $\mathcal{F}$ be a family of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Then the Rademacher Complexity $R(\mathcal{F})$ is defined as:*

$$R(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm 1\}^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(z_i)\right].$$

*More generally, given a (possibly infinite) set of vectors $A \subseteq \mathbb{R}^n$, the Rademacher Complexity $R(A)$ is defined as:*

$$R(A) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm 1\}^n}\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right].$$

<u>Intuition:</u> $R(\mathcal{F})$ captures how well the function class $\mathcal{F}$ can fit random noise as we're essentially measuring correlation between $f \in \mathcal{F}$ and a random vector $\sigma_{1:n}$. If $\mathcal{F}$ can fit random noise, then $\mathcal{F}$ will probably overfit on our training data, incurring high generalization error.

Next class we will see more of Rademacher Complexity and how to use it to bound such generalization errors.