CSCI699: Machine Learning Theory Fall 2023

Lecture 3: VC Theorem, Rademacher Complexity, Stability

Instructor: Vatsal Sharan

These lecture notes are based on an initial version scribed by Berk Tinaz, Jesse Zhang and Ali Omrani.

We begin with the notion of **shattering**.

Definition 1 (Restriction & Shattering)**.** *The **restriction** of a hypothesis class $\mathcal{H}$ to a set of examples $C = \{c_1, \ldots, c_n\} \in \mathcal{X}$ is a subset of $\{0,1\}^{|C|}$, given by $\mathcal{H}_C = \{(h(c_1), \ldots, h(c_n)), \forall h \in \mathcal{H}\}$. We say that $\mathcal{H}$ **shatters** C if $|\mathcal{H}_C| = 2^{|C|}$.*

Basically, shattering says that all possible labelings are realized when we use $\mathcal{H}$ to label the set C.

Corollary 2 (of No Free-lunch Theorem)**.** *Let $\mathcal{H}$ be a hypothesis class and assume there exists a set $C \subseteq \mathcal{X}$ of size $2n$ such that $\mathcal{H}$ shatters C. Then, $\exists$ a distribution D over $\mathcal{X} \times \{0,1\}$ and a predictor $h^* \in \mathcal{H}$ such that $R(h^*) = 0$, but for any learning algorithm A, $\mathbb{P}_{S \sim D^n}[R(A(S)) \geq 1/8] \geq 1/7$.*

In short, if $\mathcal{H}$ shatters a set of size $2n$ then one cannot learn with just n examples. Can we do something if C is such that $|\mathcal{H}_C| \ll 2^{|C|}$?

Idea: For any distribution supported on C, the real hypothesis space under consideration is actually $\mathcal{H}_C$. Moreover, because of the construction, $\mathcal{H}_C$ is finite. Therefore, if $|\mathcal{H}_C|$ is small, then maybe one can learn.

Definition 3 (VC Dimension)**.** *The VC dimension of a hypothesis class $\mathcal{H}$, denoted by* $\text{VCdim}(\mathcal{H})$ *is the size of the largest set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrary size, then* $\text{VCdim}(\mathcal{H}) = \infty$.

How to that $\text{VCdim}(\mathcal{H}) = d$:

1. Verify that there exists some set C of size d that can be shattered by $\mathcal{H}$.
2. Verify that no set of size $d+1$ is shattered by $\mathcal{H}$.

Examples

- Example 1 (Threshold functions): Let $x = [0,1]$, $\mathcal{H} = \{h_\delta(x) = \mathbb{1}(x \geq \delta), \delta \in [0,1]\}$. $\mathcal{H}$ are set of thresholds in $\mathbb{R}$.

  Claim: $\text{VCdim}(\mathcal{H}) = 1$.

  To verify the claim, we will use the 2-step approach depicted above. As a first step, we will check if there is a set C of size 1 that can be shattered by $\mathcal{H}$. Select any point x, e.g. $x = 1/3$. We can see that for $\delta \leq 1/3$, $h_\delta(x) = 1$ and similarly for $\delta > 1/3$, $h_\delta(x) = 0$. Hence, all possible labeling are realized for $|C| = 1$. For visualization, refer to the first row of Figure 1.

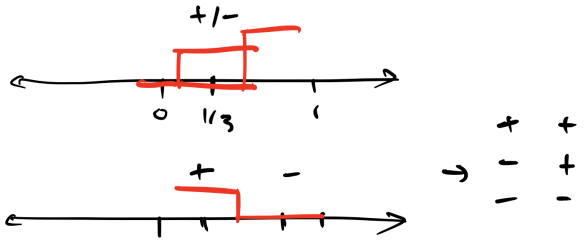

Figure 1: Setup in the first row is used to show that $\mathcal{H}$ can shatter a set C of size 1. Setup in the second row is used to show that $\text{VCdim}(\mathcal{H}) < 2$. To realize all possible labelings for $|C| = 2$, one requires reverse thresholds.

Now we have to check that $\mathcal{H}$ can't shatter any set C with $|C| = 2$. To see this, pick two points $x_1 = a$ and $x_2 = b$ such that $a, b \in [0, 1]$ and without loss of generality (w.l.o.g.) assume $a < b$. Then, for $\delta \leq a$ we have $h_\delta(a) = h_\delta(b) = 1$. For $a < \delta \leq b$ we have $h_\delta(a) = 0, h_\delta(b) = 1$ and for $b < \delta$ we have $h_\delta(a) = h_\delta(b) = 0$. However, notice that with this hypothesis class $\mathcal{H}$, we can't get the labeling $h_\delta(a) = 1, h_\delta(b) = 0$ for any $\delta \in [0, 1]$ (which requires a reverse threshold as can be seen in second row of Figure 1). Therefore, $\mathcal{H}$ does not shatter C with $|C| = 2$. Hence we are done.

If we also allow reverse thresholds, i.e. $\mathbb{1}(x < \delta)$, then we can show that $\text{VCdim}(\mathcal{H}) = 2$.

- Example 2 (Axis-aligned rectangles): Let $\mathcal{X} = \mathbb{R}^2$ and define,

$$\mathcal{H}_{a_1,a_2,b_1,b_2}(x_1, x_2) = \mathbb{1}\left(a_1 \leq x_1 \leq b_1 \;\&\; a_2 \leq x_2 \leq b_2\right)$$

Claim: $\text{VCdim}(\mathcal{H}) = 4$.

Similar to the previous example, let us first show that there is a set C of size 4 that can be shattered by $\mathcal{H}$. Consider the points in the first row of Figure 2 (points organized in diamond shape). Notice that we can enclose any subset of these points with a rectangle. Therefore, all labelings can be realized with $\mathcal{H}$.

To see that $\mathcal{H}$ cannot shatter any set C with size 5, consider the case in the second row of Figure 2. Pick any 5 points and label the left-most point c_1, the right-most point c_2, the bottom-most point c_3, and the top-most point c_4. The last point c_5 can be anywhere in the tightest rectangle fitted to the first 4 points. We would like to label first 4 points 1 but the last point 0. For the first 4 points to be labeled 1, $\mathcal{H}$ must enclose them with the rectangle. However, due to construction, c_5 must also be in that rectangle which means it can't be labeled 0. Therefore, desired labeling cannot be realized. Hence, $\text{VCdim}(\mathcal{H}) < 5$ which proves the claim.

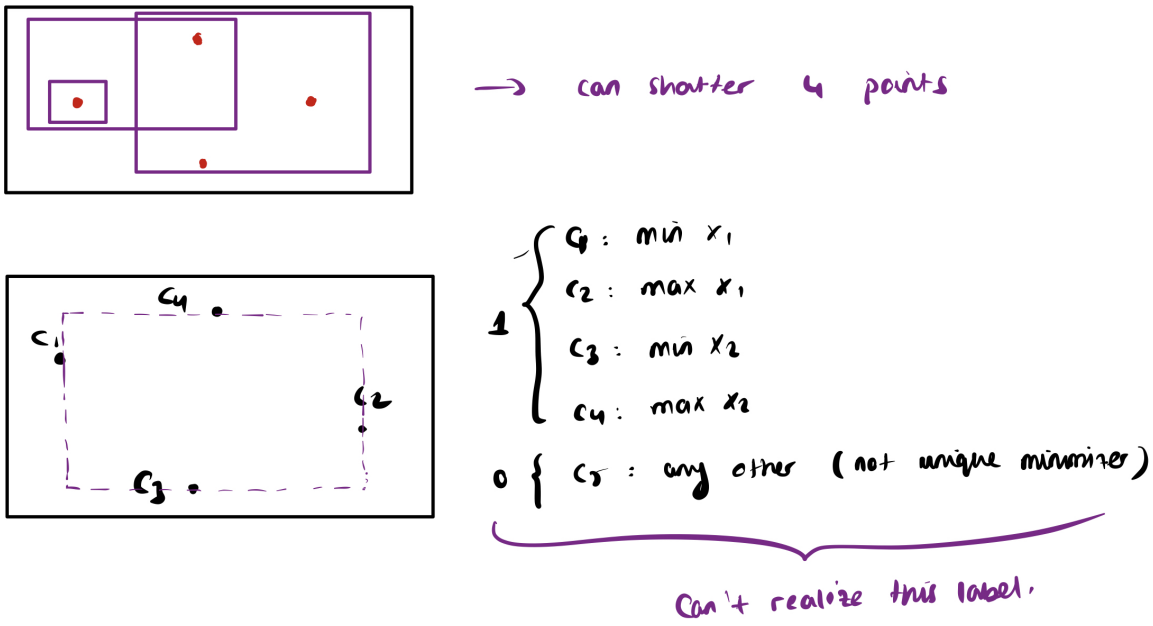

Figure 2: Setup in the first row is used to show that $\mathcal{H}$ can shatter a set C of size 4. Setup in the second row is used to show that $\text{VCdim}(\mathcal{H}) < 5$.

- Example 3 (Finite classes): For any finite hypothesis class $\mathcal{H}$, we have $\text{VCdim}(\mathcal{H}) \leq \log(|\mathcal{H}|)$. This is because for any set C, $|\mathcal{H}_C| \leq |\mathcal{H}|$. Therefore, if $2^{|C|} > |\mathcal{H}|$, then we cannot shatter C.

1 VC Theorem

Theorem 4 (VC Theorem)**.** *Let $\mathcal{H}$ be a hypothesis class with $\text{VCdim}(\mathcal{H}) = d < \infty$. Then there is an absolute constant $c > 0$ such that $\mathcal{H}$ has uniform convergence property with,*

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

Corollary 5. *$\mathcal{H}$ is agnostic-PAC learnable with $\mathcal{O}\left(\frac{d \cdot \log(d/\epsilon) + \log(1/\epsilon)}{\epsilon^2}\right)$ samples.*

Note:

(1) It is also possible to show that $n_{\mathcal{H}}^{VC}(\epsilon, \delta) \leq c \cdot \frac{d + \log(1/\delta)}{\epsilon^2}$. For $d = \log(|\mathcal{H}|)$, this bound reduces to $c \cdot \frac{\log(|\mathcal{H}|/\delta)}{\epsilon^2}$ which is the same as the $\mathcal{O}\left(\frac{\log(|\mathcal{H}|/\delta)}{\epsilon^2}\right)$ sample complexity that we derived earlier for agnostic-PAC learning.

(2) The result above is for binary classification with 0/1 loss. There are also some characterizations known beyond the 0/1 loss.

Proof Outline:

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

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

Definition 6 (Growth function)**.** *The growth function of $\mathcal{H}$, $T_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as*

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

If $\text{VCdim}(\mathcal{H}) = d$, then $T_{\mathcal{H}}(n) = 2^n, \forall n \leq d$. Sauer's Lemma gives a good upper bound $\forall n > d$. The key takeaway is that the number of possible labellings goes from being exponentially large in the number of datapoints n to only being polynomialls large in n.

Lemma 7 (Sauer's Lemma)**.** $\forall n, \text{VCdim}(\mathcal{H}) = d$,

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

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

$$T_{\mathcal{H}}(n) \leq \left(\frac{n \cdot e}{d}\right)^d \quad \textbf{\textit{(exponential to polynomial regime)}}$$

Proof. We will instead show a stronger inequality. For any $C = \{c_1, \ldots, c_n\}$ & any $\mathcal{H}$,

$$|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|. \tag{1}$$

This is sufficient since if $\text{VCdim}(\mathcal{H}) = d$, $\mathcal{H}$ cannot shatter any set B of size $|B| > d$. There are $\binom{n}{i}$ subsets of size i, hence, we will get our bound.

We will prove (1) by induction.

Base Step ($n = 1$): We have either,

1) $|\mathcal{H}_C| = 2^0 = 1$. Then, $LHS = RHS$ in (1) since one labeling shatters $\{\emptyset\}$.
2) $|\mathcal{H}_C| = 2^1 = 2$. Then, again $LHS = RHS$ as two labelings shatter $\{\{\emptyset\}, \{c_1\}\}$.

Induction Step: Assume that (1) holds for all sets of size $k < n$. Let $C = \{c_1, \ldots, c_n\}$ & $C' = \{c_2, \ldots, c_n\}$. Define,

$$Y_0 = \{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \text{ or } (1, y_2, \ldots, y_n) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \text{ and } (1, y_2, \ldots, y_n) \in \mathcal{H}_C\}.$$

Claim: $|\mathcal{H}_C| = |Y_0| + |Y_1|$. This is true because, $(y_2, \ldots, y_n)$ is counted once in Y_0, but counted again in Y_1 if it can be shattered.

By the induction hypothesis we get,

$$|Y_0| \leq \left|\{B \subseteq C^{'} : \mathcal{H} \text{ shatters } B\}\right| = |\{B \subseteq C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}|.$$

For Y_1, define $\mathcal{H}^{'} \subseteq \mathcal{H}$ to be:

$$\mathcal{H}^{'} = \left\{h \in \mathcal{H}, \exists h^{'} \in \mathcal{H} \text{ such that } ((1 - h^{'}(c_1), h^{'}(c_2), \ldots, h^{'}(c_n)) = (h(c_1), h(c_2), \ldots, h(c_n))\right\}$$

In words, $\mathcal{H}^{'}$ is the set of hypothesis h which have the property that the hypothesis that agrees with h everywhere in C except c_1 is also in $\mathcal{H}$.

Note:

1) If $\mathcal{H}^{'}$ shatters $B \subseteq C^{'}$ then it also shatters $B \cup \{c_1\}$.
2) $Y_1 = \mathcal{H}^{'}_{C'}$

Then,

$$\begin{aligned}
|Y_1| = |\mathcal{H}^{'}_{C'}| &\leq \left|\{B \subseteq C^{'} : \mathcal{H}^{'} \text{ shatters } B\}\right| \quad \text{(By induction hypothesis (1))} \\
&= \left|\left\{B \subseteq C^{'} : \mathcal{H}^{'} \text{ shatters } B \cup \{c_1\}\right\}\right| \\
&= \left|\left\{B \subseteq C : c_1 \in B \text{ and } \mathcal{H}^{'} \text{ shatters } B\right\}\right| \\
&\leq |\{B \subseteq C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}|
\end{aligned}$$

From previous claim:

$$\begin{aligned}
|\mathcal{H}_C| &= |Y_0| + |Y_1| \\
&\leq |\{B \subseteq C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| \\
&= |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|
\end{aligned}$$

which completes our proof. ■

Step 2: Symmetrization

In this step we will get a bound on the expected deviation of the empirical and true risks.

Lemma 8. *For a class $\mathcal{H}$ with growth function $\tau_{\mathcal{H}}$,*

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

Note that with this expectation bound, we can use Markov's inequality to get a high probability statement such as:

$$\Pr\left[\sup_{h\in\mathcal{H}}|R(h)-\hat{R}_S(h)|>t\right]\leq\frac{\sqrt{\frac{2\cdot\log(2\cdot T_{\mathcal{H}}(2n))}{n}}}{t}$$

However in the next step we will use McDiarmid's inequality to get a better bound than what Markov's provides. But we first need to prove the expectation bound (Lemma 8)

Proof. *(Lemma 8):* We will use the idea of **symmetrization**. Symmetrization means introducing an identical copy of a random variable to help with analysis.

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,

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right|\right]=\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'}\left[\hat{R}_{S'}(h)\right]-\hat{R}_S(h)\right|\right]. \tag{2}$$

For now we will fix S and work with S'.

Claim 9. $\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'}\left[\hat{R}_{S'}(h)\right]\right|\leq\mathbb{E}_{S'}\sup_{h\in\mathcal{H}}\left|\hat{R}_{S'}(h)\right|$.

Proof. *(Claim 9)*: 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 9 and combining with Eq. 2 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. $\sim$ Unif($\{\pm1\}$).

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,

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

Now fix both S, S' and let C be the set of examples appearing in $S \cup S'$ (both of them). Note that $|C| \le 2n$ 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,

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

Let the random variable θ_h be $\theta_h = \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}\{h(x_i')\neq y_i'\}-\mathbb{1}\{h(x_i)\neq y_i\}\right)$. Note that S and S' are fixed here, and the randomness in θ_h only comes from the randomness in the σ_i. With this notation, we can shorten the above to

$$\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right| \le \mathbb{E}_{S,S'}\mathbb{E}_{\sigma_{1:n}}\max_{h\in\mathcal{H}_C}|\theta_h|. \tag{3}$$

Now we want to bound $\mathbb{E}_{\sigma_{1:n}}\max_{h\in\mathcal{H}_C}|\theta_h|$ in Eq. 3. To do this, we will prove a bound regarding the max of sub-Gaussian variables, and then show that θ_h is sub-Gaussian.

Lemma 10 (Max of sub-Gaussians)**.** *If $(x_1, ..., x_m)$ are mean 0 and sub-Gaussian with parameter λ (they need not be independent), then*

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

As the statement of Lemma 10 says, in contrast to previous concentration bounds this one does not require the random variables involved in the bound to be independent. We will see why that is the case in the proof, but the intuition is that when we are looking at upper bounding the maximum of a set of random variables, the case where they are independent is actually the worst-case. If the random variables are independent then their maximum can only be larger with a higher probability (since the maximum only cares about just one of the random variables being large).

Proof. *(Lemma 10)*:

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

■

Claim 11. *θ_h is sub-Gaussian with parameter $\frac{1}{\sqrt{n}}, \mathbb{E}[\theta_h] = 0$.*

Proof. *(Claim 11)*: 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 θ_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 because Rademacher RV's are sub-Gaussian with parameter 1, and each σ_i is multiplied by ± 1, which does not change its sub-Gaussianaeity.

Therefore using the result for sums of sub-Gaussian random variables from the previous lecture, θ_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. 3.

Claim 12. $\mathbb{E}_{\sigma_{1:n}} \max_{h\in\mathcal{H}_C} |\theta_h| \le \frac{1}{\sqrt{n}}\sqrt{2\log(2|\mathcal{H}_C|)}$

Proof. *(Claim 12)*:

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

Recall that if θ_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 11 and Lemma 10. ■

In summary, we have now shown that the right hand side of Eq. 3, $\mathbb{E}_{S,S'} \left[\mathbb{E}_{\sigma_{1:n}} \max |\theta_h|\right]$ is bounded by $\sqrt{\frac{2 \cdot \log(2|\mathcal{H}_C|)}{n}}$. Thus, by plugging into Eq. 3,

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \leq \sqrt{\frac{2 \cdot \log(2|\mathcal{H}_C|)}{n}}. \tag{4}$$

Note that $|\mathcal{H}_C| \leq \tau_{\mathcal{H}}(2n)$ since $|C| \leq 2n$ and we can finally finish the proof of Lemma 8 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 $1/n$ (changing (x_i, y_i) can only change $\hat{R}_S(h)$ by $1/n$ for any $h \in \mathcal{H}$, therefore the max also changes by at most $1/n$).

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 δ, then with probability $1 - \delta$,

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

Now plug in Lemma 2 to replace $\mathbb{E}\left[f(S)\right]$, replace $f(S)$, and we get that

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

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. 5, with probability $(1-\delta)$,

$$\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right| \le \sqrt{\frac{2\cdot d\log(2ne/d)}{n}}+\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{6}$$

Therefore, for $n \ge \mathcal{O}\left(\frac{d\log(d/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ the right hand side $\le \epsilon$, showing the uniform convergence property for a hypothesis classes $\mathcal{H}$ with finite $\text{VCdim}(\mathcal{H}) = d$. *Exercise: Show this explicity from Eq. 6.*

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):

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

The quantity in paranthesis above is known as the "Rademacher Complexity", and we will later see that it can be used to bound generalization error. We begin by first defining this formally. Let

- $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$
- $\mathcal{F}$: function class $\mathcal{Z}\to\mathbb{R}$
- $\mathcal{D}$: distribution over $\mathcal{Z}$

Definition 13 (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 $RC(\mathcal{F})$ is defined as:*

$$RC(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(z_i)\right].$$

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

$$RC(A) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right].$$

Intuition: $RC(\mathcal{F})$ 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.

Geometric Picture

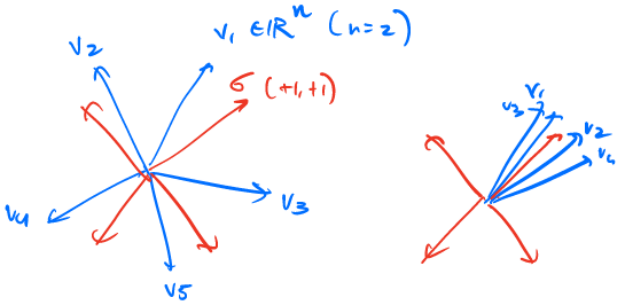

Figure 3: In expectation over $\sigma \sim \{\pm1\}^n$, what is the max inner product we can get with σ? For the figure on the left the set of vectors points in very different directions, so for every σ there is some vector v_i which has good inner product with σ. This is not the case in the figure on the right.

2.1 How do we use Rademacher complexity?

- $S = \{(x_i, y_i), i \in [n]\}$
- $\mathcal{H}$: function from $\mathcal{X} \to \mathcal{Y}$.
- $\mathcal{H} \circ S = \{h(x_1), \ldots, h(x_n) : h \in \mathcal{H}\}$
- $\ell(h(x), y)$: instead of writing $\ell(h(x), y)$ we can write $\ell(h, z) = \ell(h(x), y)$ where $z = (x, y)$
- $\ell \circ \mathcal{H} \circ S = \{(\ell(h, z_i), i \in [n]) : h \in \mathcal{H}\}$
  For example if $\mathcal{H} = \{h_1, h_2, h_3\}$

$$\ell \circ \mathcal{H} \circ S = \{(\ell(h_1, z_1), \ldots, \ell(h_1, z_n)), \ell(h_2, z_1), \ldots, \ell(h_2, z_n)), (\ell(h_3, z_1), \ldots \ell(h_3, z_n))\}$$

Lemma 14 (Symmetrization with Rademacher)**.**

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h))\leq 2\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}RC(\ell\circ\mathcal{H}\circ S)$$

Proof. The proof follows from the same argument that we used to motivate the definition of Rademacher complexity.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h)) &\leq \mathop{\mathbb{E}}_{S,S'}\sup_{h\in\mathcal{H}}\frac{1}{n}\left(\sum_{i=1}^{n}(\ell(h,z_i)-\ell(h,z_i'))\right)\\
&= \mathop{\mathbb{E}}_{S,S',\sigma_{1:n}}\sup_{h\in\mathcal{H}}\frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i(\ell(h,z_i)-\ell(h,z_i'))\right)\\
&\leq \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\sigma_{1:n}}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(h,z_i)\\
&+ \mathop{\mathbb{E}}_{S'}\mathop{\mathbb{E}}_{\sigma_{1:n}}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\ell(h,z_i').
\end{aligned}$$

Therefore we get that,

$$\mathop{\mathbb{E}}_{S}\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h))\leq 2\mathop{\mathbb{E}}_{S,\sigma_{1:n}}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(h,z_i)=2\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}RC(\ell\circ\mathcal{H}\circ S).$$

■

3 Further reading

You can read Chapter 6 of [1] for the VC theorem. The chapter does a Markov's inequality though instead of McDiardmid's to get the high probability bound, hence it is looser. The bound we show in this lecture (and also a bound for the realizable case which has a $1/\epsilon$ dependence instead of a $1/\epsilon^2$ dependence) is in Chapter 28 of the book. Getting the right bound which does not have an additional factor of $\log(d/\epsilon)$ requires using a more advanced technique for showing concentration bounds called *chaining*, and was shown in [2]. Rademacher complexity is Chapter 26 of the book.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[2] Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.