CSCI699: Theory of Machine Learning Fall 2021

Lecture 7

Instructor: Vatsal Sharan *Scribes: Ali Omrani, Ta-Yang*

1 Rademacher complexity

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

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

This quantity is called an **empirical process**. Empirical process theory studies such quantities. Let us try and expand this further, using the symmetrization idea we've seen before.

$$\begin{aligned}
\mathbb{E}_s \sup_{h\in\mathcal{H}} R(h) - \hat{R}_S(h) &\le \mathbb{E}_{S,S'} \sup_{h\in\mathcal{H}} \left(\hat{R}_{S'}(h) - \hat{R}_S(h)\right) \\
&= \mathbb{E}_{S,S'} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \left(1\{h(x_i') \ne y_i'\} - 1\{h(x_i) \ne y_i\}\right) \\
&= \mathbb{E}_{\sigma_{1:n}}\mathbb{E}_{S,S'} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(1\{h(x_i') \ne y_i'\} - 1\{h(x_i) \ne y_i\}\right) \\
&\le \mathbb{E}_{S}\mathbb{E}_{\sigma_{1:n}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i 1\{h(x_i') \ne y_i'\} \\
&+ \mathbb{E}_{S}\mathbb{E}_{\sigma_{1:n}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} (-\sigma_i) 1\{h(x_i) \ne y_i\} \\
\implies \mathbb{E}_S \sup_{h\in\mathcal{H}} R(h) - \hat{R}_S(h) &\le 2\mathbb{E}_{S}\mathbb{E}_{\sigma_{1:n}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i 1\{h(x_i) \ne y_i\}.
\end{aligned}$$

The quantity on the right hand side is what we will call the Rademacher complexity. We define this formally now. Let

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

Definition (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*

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

More generally, given a set of vectors $A \subset \mathbb{R}^n$, the Rademacher complexity $R(A)$ is defined as

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

Intuition:

- $R(\mathcal{F})$ captures how well can $\mathcal{F}$ fit random noise $\rightarrow$ if $\mathcal{F}$ can fit random noise, $\mathcal{F}$ will probably overfit.

Geometric Picture

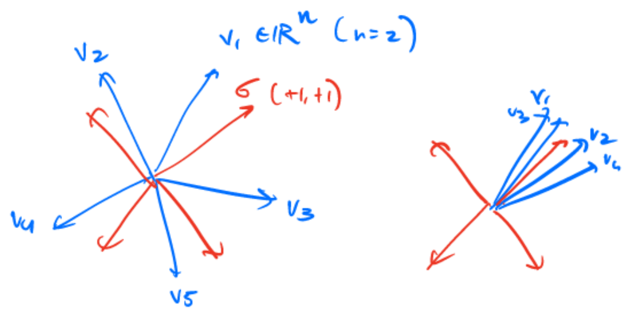

Figure 1: In expectation over $\sigma \sim \{\pm 1\}^n$, what is the max inner product we can get with σ? For the figre on the left the set of vectors points in very different directions, so for every σ there is some vector v_i which has good inner product with σ. This is not the case in the figure on the right.

1.1 How do we use Rademacher complexity?

- $S = \{(x_i, y_i), i \in [n]\}$
- $\mathcal{H}$: function from $\mathcal{X} \rightarrow \mathcal{Y}$.
- $\mathcal{H} \circ S = \{h(x_1), \ldots, h(x_n) : h \in \mathcal{H}\}$
- $\ell(h(x), y)$: 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 1 (Symmetrization with Rademacher)**.**

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

Proof.

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

Therefore we get that,

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

□

Theorem 2 (Excess risk bounds using Rademacher)**.** *Assume that for all z and $h\in\mathcal{H}$ we have that $|\ell(h,z)|\leq C$. Then with probability at least $(1-\delta)$ over $S\sim\mathcal{D}^n$,*

(1)

$$\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h)) \leq 2\mathbb{E}_{S'}R(\ell\circ\mathcal{H}\circ S') + c\sqrt{\frac{2\log(1/\delta)}{n}}$$

(2)

$$\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h)) \leq 2R(\ell\circ\mathcal{H}\circ S) + 3c\sqrt{\frac{2\log(2/\delta)}{n}}$$

(3) For any $h^*\in\mathcal{H}$,

$$R(ERM_{\mathcal{H}}(S)) - R(h^*) \leq 2R(\ell\circ\mathcal{H}\circ S) + 4c\sqrt{\frac{2\log(4/\delta)}{n}}.$$

(in particular, this holds for $h^* = \arg\min_{h\in\mathcal{H}} R(h)$)

Proof. We will keep using McDiarmid's inequality throughout the proof.

(1) Note that $\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h))$ satisfies the bounded differences property with constant $\frac{2c}{n}$. (changing any (x_i,y_i) changes the loss by at most $\frac{2c}{n}$).

$\therefore$ Using McDiarmid's

$$\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h)) \leq \mathbb{E}(\sup_{h\in\mathcal{H}}(R(h)-\hat{R}_S(h))) + \epsilon$$

with probability

$$1 - \exp\left(\frac{-2\epsilon^2}{n(2c/n)^2}\right) = 1 - \underbrace{\exp\left(-\frac{n\epsilon^2}{2c^2}\right)}_{\delta}.$$

We choose $\epsilon = c\sqrt{\dfrac{2\log(1/\delta)}{n}}$ to set the error probability to be δ. Therefore we get that with probability $1-\delta$,

$$\sup_{h\in\mathcal{H}}(R(h) - \hat{R}_S(h)) \le \mathbb{E}\sup_{h\in\mathcal{H}}(R(h) - \hat{R}_S(h)) + c\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Now use Lemma 1 (Symmetrization with Rademacher), and result follows.

(2) Note that

$$R(\ell\circ\mathcal{H}\circ S) = \mathbb{E}_{\sigma_{1:n}}\left(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(h,z_i)\right)$$

also satisfies bounded differences with constant $2c/n$ (swapping σ_i by σ_i' changes the value by $\le 2c/n$). With probability $1-\delta$,

$$R(\ell\circ\mathcal{H}\circ S) \ge \mathbb{E}_{S'}(R(\ell\circ\mathcal{H}\circ S')) - c\sqrt{\frac{2\log(1/\delta)}{n}}.$$

So

$$\mathbb{E}_{S'}(R(\ell\circ\mathcal{H}\circ S')) \le R(\ell\circ\mathcal{H}\circ S) + c\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Now set $\delta = \delta'/2$, with probability $1 - \dfrac{\delta'}{2}$,

$$\mathbb{E}_{S'}(R(\ell\circ\mathcal{H}\circ S')) \le R(\ell\circ\mathcal{H}\circ S) + c\sqrt{\frac{2\log(2/\delta')}{n}},$$

$$\sup_{h\in\mathcal{H}}(R(h) - \hat{R}_S(h)) \le \mathbb{E}_{S'}(R(\ell\circ\mathcal{H}\circ S')) + c\sqrt{\frac{2\log(2/\delta)}{n}} \text{ (from part (1)).}$$

The result now follows by doing a union bound and combining the above results. With probability $1-\delta$,

$$\sup_{h\in\mathcal{H}}(R(h) - \hat{R}_S(h)) \le 2R(\ell\circ\mathcal{H}\circ S) + 3c\sqrt{\frac{2\log(2/\delta)}{n}}.$$

(3) Let $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$.

$$R(h_S) - R(h^*) = \underbrace{R(h_S) - \hat{R}_S(h_S)}_{\text{bounded by part (2)}} + \underbrace{\hat{R}_S(h_S) - \hat{R}_S(h^*)}_{\le 0} + \underbrace{\hat{R}_S(h^*) - R(h^*)}_{\text{Hoeffding's}}.$$

With probability $1-\delta/2$,

$$\hat{R}_S(h^*) - R(h^*) \le c\sqrt{\frac{2\log(2/\delta)}{n}}.$$

$$\therefore R(h_S) - R(h^*) \le 2R(\ell\circ\mathcal{H}\circ S) + 4c\sqrt{\frac{2\log(4/\delta)}{n}}.$$

□

Takeaways

- **Could be much better than VC bound**: Rademacher complexity takes the data distribution into account, and could give tighter bounds than worst case VC bounds.
- **Data-dependent bound**. (3) in our theorem is a data-dependent bound, we use a training set S both for learning a hypothesis from $\mathcal{H}$, and for estimating its generalization error (we can check if we are overfitting)

1.2 Rademacher calculus

Claim 3 (Translation and Scaling)**.** *Let $A' = \{\rho a + v, a \in A\}$. Then $R(A') = \rho R(A)$.*

Proof. Exercise. □

Can also show that

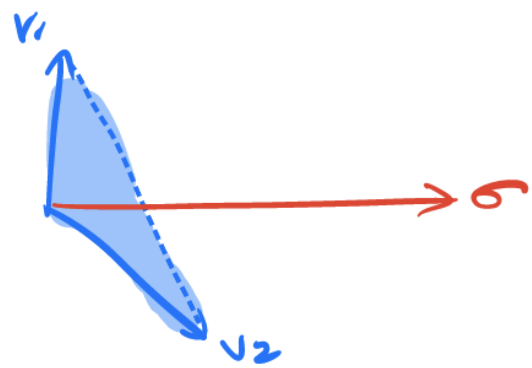

Figure 2: R({convex hull of A}) = R(A).

Lemma 4 (Massart Lemma)**.** *Let $A = \{v_1, \ldots, v_m\}$ be a finite set of vectors in $\mathbb{R}^n$. Let $\overline{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$. Then*

$$R(A) \le \max_i \|v_i - \overline{v}\|_2 \frac{\sqrt{2\log m}}{n}$$

Proof. Exercise. Hint: First, by translation invariance, we can take $\overline{v} = 0$ without loss of generality. Then use the max of sub-Gaussian result from last time. □

Note: This gives a bound for finite hypothesis classes.

Lemma 5 (Contraction lemma)**.** *For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function i.e. $|\phi_i(x) - \phi_i(y)| \le \rho|x - y| \;\forall\; x, y \in \mathbb{R}$. For any $a \in \mathbb{R}^n$ define $\phi(a) \in \mathbb{R}^n$ as*

$$\phi(a) = (\phi_1((a)_1), \ldots, \phi_n((a)_n)).$$

For a set A, let $\phi \circ A = \{\phi(a) : a \in A\}$. Then

$$R(\phi \circ A) \leq \rho R(A).$$

Proof. Refer to book. □

1.3 Rademacher complexity of linear classes

- $\mathcal{H}_1 = \{h_w(x) = \langle w, x\rangle\} : \|w\|_1 \leq B_1\}$
- $\mathcal{H}_2 = \{h_w(x) = \langle w, x\rangle : \|w\|_2 \leq B_2\}$

Lemma 6 (ℓ_2 bounded linear predictor)**.** *Let $S = (x_1, \ldots, x_n)$. Define*

$$H_2 \circ S = \{(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle) : \|w\|_2 \leq B_2\}$$

Then

$$R(\mathcal{H}_2 \circ S) \leq \frac{B_2 \max_i \|x_i\|_2}{\sqrt{n}}$$

Proof. By Cauchy-Schwartz: $\langle w, v\rangle \leq \|w\|_2 \|v\|_2$.

$$\begin{aligned}
\therefore nR(H_2 \circ S) &= \mathbb{E}_\sigma \left[\sup_{\mathcal{H}_2 \circ S} \sum_{i=1}^n \sigma_i a_i \right] \\
&= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_2 \leq B_2}, \sum_{i=1}^n \sigma_i \langle w, x_i\rangle \right] \\
&= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_2 \leq B_2} \langle w, \sum_{i=1}^n \sigma_i x_i\rangle \right] \\
&\leq B_2 \cdot \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2 \right]. \qquad (1)
\end{aligned}$$

Using Jensen's,

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2 \right] = \mathbb{E}_\sigma \left[\left(\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right)^{1/2} \right] \leq \left(\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right] \right)^{1/2} \qquad (2)$$

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right] = \mathbb{E}_\sigma \left[\sum_{i,j} \sigma_i \sigma_j \langle x_i, x_j\rangle \right]$$

Since σ_i are independent,

$$\mathbb{E}_\sigma[\sigma_i, \sigma_j] = 0 \quad \forall\, i \neq j$$

$$\implies \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right] = \sum_{i=1}^n \|x_i\|_2^2 \leq n \max_i \|x_i\|_2^2. \qquad (3)$$

The proof follows by combining (1), (2) and (3). □

Lemma 7 (ℓ_1 bounded linear model)**.** *Let* $S = (x_1, \ldots, x_n)$ *where* $x_i \in \mathbb{R}^d$ $\forall i \in [n]$ *Then*

$$R(H_1 \circ S) \leq B_1 \max_i \|x_i\|_\infty \sqrt{\frac{2\log(2d)}{n}}$$

Proof. By Holder's inequality $\langle w, v\rangle = \|w\|_1 \|v\|_\infty$. Therefore,

$$\begin{aligned}
nR(H_1 \circ S) &= \mathbb{E}_\sigma \left[\sup_{a \in H_1 \circ S} \sum_{i=1}^n \sigma_i a_i \right] \\
&= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_1 \leq B_1} \sum_{i=1}^n \sigma_i \langle w_i, x_i \rangle \right] \\
&= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_1 \leq B_1} \langle w, \sum_{i=1}^n \sigma_i x_i \rangle \right] \\
&\leq B_1 \cdot \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_\infty \right] \\
&= B_1 \mathbb{E}_\sigma \left[\max_{j \in [d]} \left| \sum_{i=1}^n \sigma_i (x_i)_j \right| \right].
\end{aligned}$$

Note that each term $\sigma_i(x_i)_j$ is $|(x_i)_j|$ sub-Gaussian. Since $|(x_i)_j| \leq \max_i \|x_i\|_\infty$, each term $\sigma_i(x_i)_j$ is $\max_i \|x_i\|_\infty$ sub-Gaussian. The sum $\sum_{i=1}^n \sigma_i(x_i)_j$ is sub-Gaussian with parameter

$$\left(\sum_{i=1}^n \left(\max_i \|x_i\|_\infty \right)^2 \right)^{1/2} \leq \sqrt{n} \cdot \max_i \|x_i\|_\infty .$$

By bound for max of sub-Gaussian, including negations to take care of the absolute value function we have

$$nR(H_1 \circ S) \leq B_1 \sqrt{n} \max_i \|x_i\|_\infty \sqrt{2\log(2d)}.$$

$\square$