CSC 588: Machine learning theory **Spring 2021**

Lecture 9: Proof of the uniform convergence theorem for VC classes

Lecturer: Chicheng Zhang *Scribe: Yinan Li*

1 Three Lemmas used in the proof of Uniform Convergence

In the last lecture, we have seen the proof of the Uniform Convergence via the following three Lemmas.

Lemma 1. *With probability* $1-\delta/2$

$$\sup_{f\in\mathcal{F}} \mathbb{E}_S[f(z)] - \mathbb{E}_{\mathcal{D}}[f(z)] \le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_S[f(z)] - \mathbb{E}_{\mathcal{D}}[f(z)]\right] + \sqrt{\frac{\ln(4/\delta)}{2n}}$$

Lemma 2.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_S[f(Z)] - \mathbb{E}_{\mathcal{D}}[f(Z)]\right] \le 2\,\mathrm{Rad}_n(\mathcal{F})$$

where

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}_{S\sim D^n}\,\mathrm{Rad}_S(\mathcal{F})$$

is the expectation of empirical Rademacher Complexity, and

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{n}\cdot \mathbb{E}_{\sigma\sim U(\pm 1)^n}[\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(Z_i)\sigma_i].$$

Lemma 3. *For any set* S *of size* n

$$\mathrm{Rad}_S(\mathcal{F}) \le \sqrt{\frac{2\ln\left(S(\mathcal{F},n)\right)}{n}}$$

Lemma 1 reduces the data dependent concentration quality to the distribution dependent concentration quality. Lemma 2 further bounds this distribution dependent quality by Rademacher complexity. Note that it reduces bounding the supremum of a possibly infinite collection to the supremum of a finite collection.

In this lecture, we prove these Lemmas.

2 Proof of Lemma 1

Proof.

Lemma 4 (McDiarmid's Lemma)**.** *g is c-sensitive, $X_1,\dots,X_n$ are i.i.d from distribution $\mathcal{D}$ on V. Then with probability* $1-\delta'$,

$$|g(X_1,\dots,X_n) - \mathbb{E}g(X_1,\dots,X_n)| \le c\cdot\sqrt{\frac{n}{2}\ln(\frac{2}{\delta'})}.$$

Define $g(x_1,\dots,x_n) = \sup_{f\in\mathcal{F}}\left[\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)\right]$, where $S=\{x_1,\dots,x_n\}$.

We show that g is c-sensitive with $c=\frac{1}{n}$. Denote by $F(f) = \mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)$, which is actually the inner part of $g(x_1,\dots,x_i,\dots,x_n)$, denote by $G(f)$ the corresponding inner part of $g(x_1,\dots,x_i',\dots,x_n)$. It is easy to see that $F(f)-G(f)\le\frac{1}{n}$.

Suppose $F(f_0) = \sup_{f\in\mathcal{F}} F(f)$, then

$$F(f_0) \le G(f_0) + \frac{1}{n} \le \sup_{f\in\mathcal{F}} G(f) + \frac{1}{n}$$

which implies that g is $\frac{1}{n}$-sensitive.

Now we are ready to apply McDiarmid's Lemma with $\delta' = \delta/2$, which gives us

$$g(X_1,\dots,X_n) \le \mathbb{E}g(X_1,\dots,X_n) + \sqrt{\frac{\ln(4/\delta)}{2n}}.$$

□

3 Proof of Lemma 2

Proof. **Step 1:** Symmetrization (double sampling trick). This step is to show that

$$\mathbb{E}_{S\sim D^n} \sup_{f\in\mathcal{F}} [\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)] \le \mathbb{E}_{S,S'\sim D^n} \sup_{f\in\mathcal{F}} [\mathbb{E}_S f(Z) - \mathbb{E}_{S'} f(Z)]$$

(Think of S and S' as training and validation dataset of the same size, so the RHS is evaluating the difference between the train error and test error, and taking the maximum over functions. RHS is already the supremum over a finite collection of size at most 2^{2n}, furthermore, it is at most $S(\mathcal{F}, 2n)$.)

Proof. It suffices to show that

$$\sup_{f\in\mathcal{F}} [\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)] \le \mathbb{E}_{S'\sim D^n} \sup_{f\in\mathcal{F}} [\mathbb{E}_S f(Z) - \mathbb{E}_{S'} f(Z)]$$

for all realizations of S. To see why this is true, note that the LHS is deterministic, and $\mathbb{E}_D f(Z) = \mathbb{E}_{S'\sim D^n}\mathbb{E}_{S'} f(Z)$, as well as the following observation.

Claim 5. *Suppose G is a random function, then*

$$\sup_{f\in\mathcal{F}} \mathbb{E}G(f) \le \mathbb{E} \sup_{f\in\mathcal{F}} G(f)$$

Proof. Suppose $f_0 = \text{argmax}_{f\in\mathcal{F}} \mathbb{E}G(f)$. Then we have

$$\mathbb{E}G(f_0) \le \mathbb{E} \sup_{f\in\mathcal{F}} G(f)$$

□

Back to the proof of Step 1, taking expectation with $S \sim D^n$, we get the symmetrization Lemma.

□

Step 2: Introducing random signs.

Claim 6.

$$\frac{1}{n}\mathbb{E}_{S,S'\sim D^n} \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n f(Z_i) - f(Z_i')\right] = \frac{1}{n}\mathbb{E}_{S,S'\sim D^n} \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n (f(Z_i) - f(Z_i'))\sigma_i\right]$$

for all $\sigma_1,\dots,\sigma_n \in \{\pm1\}$. Furthermore,

$$\frac{1}{n}\mathbb{E}_{S,S'\sim D^n} \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n f(Z_i) - f(Z_i')\right] = \frac{1}{n}\mathbb{E}_{S,S'\sim D^n,\sigma\sim U(\pm1)^n} \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n (f(Z_i) - f(Z_i'))\sigma_i\right]$$

(Recall that $U(\pm1)$ is the uniform distribution over $\{-1,+1\}$, also known as the Rademacher distribution.)

Remark 7. *Chicheng notes: The second part of the claim above was only briefly touched upon in the lecture. However, it is very important, because we are going to upper bound it further in the next step.*

Example 1. *Suppose $n = 2$. $\sigma_1 = -1$, $\sigma_2 = +1$.*

$$\mathbb{E}_{Z_1,Z_2.Z_1',Z_2'\sim D^4} \sup_{f\in\mathcal{F}}(f(Z_1) - f(Z_1') + f(Z_2) - f(Z_2')) = \mathbb{E}_{Z_1,Z_2.Z_1',Z_2'\sim D^4} \sup_{f\in\mathcal{F}}(f(Z_1') - f(Z_1) + f(Z_2) - f(Z_2'))$$

this is because (Z_1, Z_1', Z_2, Z_2') has the same distribution as (Z_1', Z_1, Z_2, Z_2').

Step 3: This step is to show that

$$\frac{1}{n}\mathbb{E}_{S,S'\sim D^n,\sigma\sim U(\pm 1)^n} \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n (f(Z_i) - f(Z_i'))\sigma_i\right] \leq 2\,\mathrm{Rad}_n(\mathcal{F}).$$

Proof. By the fact that $\sup_f(A(f) + B(f)) \leq \sup_f A(f) + \sup_f B(f)$,

$$\begin{aligned}
LHS =& \frac{1}{n}\mathbb{E}_{S,S'\sim D^n,\sigma\sim U(\pm 1)^n} \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n f(Z_i)\sigma_i - \sum_{i=1}^n f(Z_i')\sigma_i\right] \\
\leq& \frac{1}{n}(\mathbb{E}_{S,\sigma} \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(Z_i)\sigma_i + \mathbb{E}_{S',\sigma} \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(Z_i')(-\sigma_i)) \\
\leq& 2\mathbb{E}_{S,\sigma} \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(Z_i)\sigma_i \\
\leq& 2\,\mathrm{Rad}_n(\mathcal{F}).
\end{aligned}$$

□

Combining Step 1-3,

$$\mathbb{E} \sup_{f\in\mathcal{F}} [\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)] \leq 2\,\mathrm{Rad}_n(\mathcal{F})$$

□

4 Proof of Lemma 3

Proof. For all $(b_1, \ldots, b_n) \in \Pi_{\mathcal{F}}(S)$, there exists an f from $\mathcal{F}$, such that it achieves this labeling. Denote by $\mathcal{F}_S$ the set of representatives f's picked in above way, $|\mathcal{F}_S| \leq S(\mathcal{F}, n)$. Therefore.

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(Z_i)\sigma_i$$

Lemma 8 (Massart's Finite Lemma)**.** *Suppose $X_1, \ldots, X_N$ are zero mean, σ^2-subgaussion, then*

$$\mathbb{E}[\max_{i=1}^N X_i] \leq \sigma\sqrt{2\ln N}$$

Proof. For $\forall t > 0$,

$$\max_i X_i \leq \frac{\ln(\sum_{i=1}^N e^{tx_i})}{t}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\max_i X_i \leq & \frac{\mathbb{E}\ln(\sum_{i=1}^N e^{tx_i})}{t} \\ \leq & \frac{\ln(\mathbb{E}\sum_{i=1}^N e^{tx_i})}{t} \\ \leq & \frac{\ln N}{t} + \frac{\sigma^2 t}{2}\end{aligned}$$

where these inequalities are by taking expectation, Jensen's Inequality and subgaussian properties, respectively. Now we are free to choose t to minimize the RHS. By picking $t = \sqrt{\frac{2\ln N}{\sigma^2}}$,

$$\mathbb{E}[\max_i X_i] \leq \sigma\sqrt{2\ln N}$$

□

This Lemma is tight up to constant. Consider $X_1, \ldots, X_N \sim N(0, \sigma^2)$, for a fixed N and any i, with probability at least $\frac{1}{N}$, $X_i \geq \sigma\sqrt{\ln N}$. Because there are N trials, the expectation of numbers of X_i's that lie on the right of $\sigma\sqrt{\ln N}$ is greater than 1. This concludes that $\mathbb{E}[\max_{i=1}^N X_i] \geq c\sigma\sqrt{\ln N}$ for some constant c.

Now back to the proof of Lemma 3. Define $X_f = \sum_{i=1}^n f(Z_i)\sigma_i$ for any $f \in \mathcal{F}$. X_f is zero mean and σ^2-subgaussion with $\sigma^2 = n$. Applying Massart's Lemma, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}} X_f \leq \sqrt{n \cdot 2\ln S(\mathcal{F}, n)},$$

which gives

$$\text{Rad}_S(\mathcal{F}) \leq \sqrt{\frac{2\ln\left(S(\mathcal{F}, n)\right)}{n}}.$$

□