CSC 588: Machine learning theory **Spring 2022**

Lecture 11: Proof of the uniform convergence theorem for VC classes

Lecturer: Chicheng Zhang *Scribe: Minhang Zhou*

1 Three Lemmas used in the proof of Uniform Convergence

In the last lecture, we have seen some proof of the Uniform Convergence results via the following three lemmas. Lemma 1 helps us reduce the task of bounding something random to deterministic. Lemma 2 helps us reduce bounding the expectation of the maximum of a bunch of infinite collection of random variables to bounding the expectation of the maximum of finite collection of random variables. Lemma 3 helps us deal with the expectation of the maximum of a finite collection of random variables.

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

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

Lemma 2. (Symmetrization Lemma)

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

where

$$\mathrm{Rad}_n(f) = \mathbb{E}_{S \sim D^m} \mathrm{Rad}_S(f)$$

and

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

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

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

In the previous lecture, we used Massart's Finite Lemma to prove Lemma 3.

2 Proof of Massart's Finite Lemma

Lemma 4 (Massart's Finite Lemma). *If* $X_1, ..., X_N \sim$ are zero mean, σ^2-subgaussian, 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, by using Jensen's Inequality and subgaussian properties,

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

Note that this bound holds for all t, we can choose t that minimizes the right hand side to get the tightest bound. This is achieved when $t = \sqrt{\frac{2\ln N}{\sigma^2}}$. Thus, we have

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

3 Proof of Lemma 1

Lemma 5(McDiarmid's Lemma). If g is c-sensitive, $Z_1 ... Z_n$ are i.i.d from distribution D on V. Then:

$$P\left(|g(Z_1, ..., Z_n) - \mathbb{E}g(Z_1, ..., Z_n)| \ge \epsilon\right) \le 2\ \exp(\frac{-2\epsilon^2}{nc^2}),$$

In other words, with probability $1 - \delta'$:

$$|g(Z_1, ..., Z_n) - \mathbb{E}g(Z_1, ..., Z_n)| \le c\sqrt{\frac{n}{2}\ln(\frac{2}{\delta'})}$$

Def(sensitivity): g is c-sensitive if: for every $i \in \{1, \dots, n\}$, $z_1, ..., z_n, {z_i}' \in V$, it always holds that

$$|g(z_1, \dots, z_n) - g(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)| \le c.$$

Remarks:
(1) g can take value in an interval of size nc, but what this lemma says is that, when receiving iid inputs, g can "typically" take values in an interval of size $c\sqrt{n}$
(2)McDiarmid's Lemma implies Hoeffding's Inequality, as the mean function over a $V = [a, b]$ has sensitivity $c = \frac{b-a}{n}$.
(3)Example with large sensitivity constant c :

$$g(z_1, ..., z_n) = \text{Median}(z_1, ..., z_n)$$

Here, c can only be chosen as $b - a$, we can illustrate the idea by a simple example below:
Suppose we have $n = 99$ samples which include 49 $a's$ and 50 $b's$. If we change one input from b to a, then we will have 50 $a's$ and 49 $b's$. This would cost the median of the 99 samples changing from b to a. Therefore, the worst-case c can only choose a value which is as large as $b - a$ and is also independent of the sample size n.

Proof of Lemma 1. Let's examine the sensitivity parameter of

$$g(z_1, ..., z_n) = \sup_{f\in\mathcal{F}}(\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z))$$

Denote by $S = (z_1, ..., z_n), S^{(i)} = (z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_n)$, we would like to show that

$$|g(S) - g(S^{(i)})| \le \frac{1}{n} \tag{1}$$

The reason is as follows:

$$g(S) = \sup_{f\in\mathcal{F}} F(f) \quad F(f) = \mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)$$

$$g(S^{(i)}) = \sup_{f\in\mathcal{F}} G(f) \quad G(f) = \mathbb{E}_{S^{(i)}} f(Z) - \mathbb{E}_D f(Z)$$

Observe that, for $\forall f$,

$$|F(f) - G(f)| = |\frac{1}{n}(f(z_i) - f(z_i'))| \leq \frac{1}{n}$$

Now we can use the following fact to show Equation (1).
Fact: If for $\forall f$

$$|F(f) - G(f)| \leq \alpha$$

then

$$-\alpha \leq \sup_{f\in\mathcal{F}} F(f) - \sup_{f\in\mathcal{F}} G(f) \leq \alpha$$

Proof. We only show the upper bound; the lower bound can be shown symmetrically. Let

$$f_0 = \operatorname*{argmax}_{f\in\mathcal{F}} F(f)$$

$$\sup_{f\in\mathcal{F}} F(f) - \sup_{f\in\mathcal{F}} G(f) = F(f_0) - \sup_{f\in\mathcal{F}} G(f) \leq F(f_0) - G(f_0) \leq \alpha$$

Lemma 1 follows by taking the above g with:

$$c = \frac{1}{n}, \delta' = \frac{\delta}{2}$$

4 Partial Proof of Lemma 2

Step 1: Use double sampling lemma to reduce bounding the uniform deviation between empirical average and population average to bounding the uniform deviation between empirical average and another empirical average (over a fresh "validation set").

Lemma 1 (Double sampling lemma)**.**

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

It suffices to show: $\forall S$,

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

because by taking the expectation over S, we essentially get the double sampling lemma.
Fact: Suppose G is a random function that maps f to reals, then,

$$\sup_{f\in\mathcal{F}} \mathbb{E}[G(f)] \leq \mathbb{E}[\sup_{f\in\mathcal{F}} G(f)]$$

Proof. With the purpose of concluding the double sampling lemma, we pick

$$f_0 = \operatorname*{argmax}_{f\in\mathcal{F}} \mathbb{E}[G(f)]$$

$$\text{Since } G(f_0) \leq \sup_{f\in\mathcal{F}} G(f),\ \mathbb{E}[G(f_0)] \leq \mathbb{E}[\sup_{f\in\mathcal{F}} G(f)]$$

Step 2: introduce random signs:

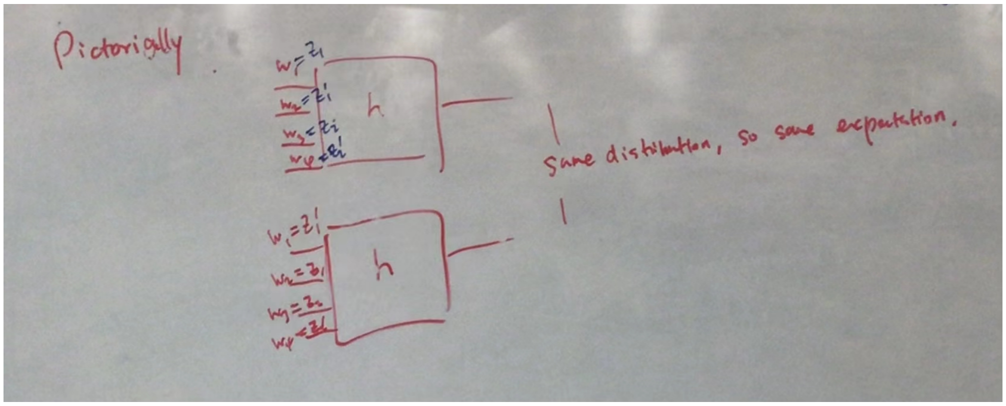

Figure 1: A simple example for Lemma 2

Lemma 2. *For any fixed $(\sigma_1, ..., \sigma_n) \in \{\pm 1\}$, we have:*

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

Therefore,

$$\frac{1}{n}\mathbb{E}_{\substack{S\sim D^n \\ S'\sim D^n}} \sup_{f\in\mathcal{F}}(\sum_{i=1}^{n}(f(z_i) - f(z_i'))) = \frac{1}{n}\mathbb{E}_{S,S',\sigma\sim U(\pm 1)^n} \sup_{f\in\mathcal{F}}(\sum_{i=1}^{n}(f(z_i) - f(z_i'))\,\sigma_i)$$

We leave the proof of the general lemma to the readers, and only illustrate the key idea via a simple example. Example: $n = 2, \sigma_1 = -1, \sigma_2 = +1$.
From equations above, we have:

$$LHS = \frac{1}{2}\mathbb{E}_{z_1,z_2,z_1',z_2'\sim D^4} \sup_{f\in\mathcal{F}}(f(z_1) - f(z_1') + f(z_2) - f(z_2')) = \mathbb{E}_{z_1,z_1',z_2,z_2'}\left[h(z_1, z_1', z_2, z_2')\right]$$

$$RHS = \frac{1}{2}\mathbb{E}_{z_1,z_2,z_1',z_2'\sim D^4} \sup_{f\in\mathcal{F}}(f(z_1') - f(z_1) + f(z_2) - f(z_2')) = \mathbb{E}_{z_1,z_1',z_2,z_2'}\left[h(z_1, z_1', z_2', z_2)\right]$$

We define:

$$h(w_1, w_2, w_3, w_4) = \sup_{f\in\mathcal{F}}(f(w_1) - f(w_2) + f(w_3) - f(w_4))$$

Also, note:

$$(z_1, z_1', z_2, z_2') \overset{d}{=} (z_1', z_1, z_2, z_2') \overset{d}{=} D^4,$$

where $\overset{d}{=}$ denotes equal in distribution.

Therefore,

$$\mathbb{E}_{z_1,z_1',z_2,z_2'}\left[h(z_1, z_1', z_2', z_2)\right] = \mathbb{E}_{z_1,z_1',z_2,z_2'}\left[h(z_1, z_1', z_2, z_2')\right] = \mathbb{E}_{w_1,w_2,w_3,w_4\sim D^4}\left[h(w_1, w_2, w_3, w_4)\right].$$

See Figure 1 for an illustration.

To be continued...