

# **SDS 384 11: Theoretical Statistics**

## **Lecture 12: Uniform Law of Large Numbers- Rademacher Complexity**

---

Purnamrita Sarkar  
Department of Statistics and Data Science  
The University of Texas at Austin

# Proof of the GC theorem

- We will work on a proof that can handle general function classes  $\mathcal{F}$  with bounded functions. WLOG let  $|f(X_i)| \leq 1$  for  $f \in \mathcal{F}$ .
- Recall that we want to bound  $\|\hat{P}_n - P\|_{\mathcal{F}}$
- The proof has three components:
  - Concentration inequality to bound  $\|\hat{P}_n - P\|_{\mathcal{F}} - E[\|\hat{P}_n - P\|_{\mathcal{F}}]$
  - Symmetrization to relate  $E[\|\hat{P}_n - P\|_{\mathcal{F}}]$  to Rademacher complexity
  - Bound this complexity using the effective “size” of the function class.

# Concentration

- First note that we cannot apply Hoeffding/Chernoff here.
- Let  $X := \{X_1, \dots, X_n\}$
- Let  $g(X) = \|\hat{P}_n - P\|_{\mathcal{F}}$ . Let  $Y$  be another sample  $\{Y_1, \dots, Y_n\}$ , where  $Y_i = X_i, \forall i \neq 1$ .
- Let  $f_1$  maximize  $g(X)$ , and  $f_2$  maximize  $g(Y)$
- 

$$\begin{aligned} g(X) - g(Y) &= \left| \frac{\sum_i f_1(X_i)}{n} - Ef_1[X_1] \right| - \left| \frac{\sum_i f_2(Y_i)}{n} - Ef_2[X_1] \right| \\ &\leq \left| \frac{\sum_i f_1(X_i)}{n} - Ef_1[X_1] \right| - \left| \frac{\sum_i f_1(Y_i)}{n} - Ef_1[X_1] \right| \\ &\leq \frac{2}{n} \end{aligned}$$

- Using McDiarmid's inequality, we get:

$$P(g(X) - E[g(X)] \geq \epsilon) \leq \exp(-\epsilon^2 n/2)$$

- So, with probability  $1 - \exp(-\epsilon^2 n/2)$ ,

$$\|\hat{P}_n - P\|_{\mathcal{F}} \leq E[\|\hat{P}_n - P\|_{\mathcal{F}}] + \epsilon.$$

- So, we need to bound  $E[\|\hat{P}_n - P\|_{\mathcal{F}}]$ .

# Symmetrization

- Consider an iid copy of  $X'$  of  $X$

$$\begin{aligned} E\|\hat{P}_n - P\|_{\mathcal{F}} &= E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i (f(X_i) - E[f(X_i)]) \right| \\ &= E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i (f(X_i) - E[f(X'_i)]) \right| \\ &= E_X \sup_{f \in \mathcal{F}} \left| E_{X'} \left[ \frac{1}{n} \sum_i (f(X_i) - f(X'_i)) \right] \right| \\ &\leq E_{X, X'} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i (f(X_i) - f(X'_i)) \right| \\ &= E_{X, X'} \|\hat{P}_n - \hat{P}'_n\|_{\mathcal{F}} \end{aligned}$$

## Symmetrize again

- Let  $\epsilon_i \in \{1, -1\}$ .
- Note that  $f(X_i) - f(X'_i)$  is symmetric
- For a symmetric random variable  $U$ , and a random variable  $\epsilon \in \{-1, 1\}$  (independent of  $R$ )

$$\begin{aligned}P(\epsilon R \leq t) &= P(R \leq t)P(\epsilon = 1) + P(R \geq -t)P(\epsilon = -1) \\ &= P(R \leq t)\end{aligned}$$

- Hence  $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i (f(X_i) - f(X'_i)) \right|$  and  $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i (f(X_i) - f(X'_i)) \right|$  have the same distribution, and expectation
- We will choose  $\epsilon_i$ 's uniformly, i.e. we will consider Rademacher random variables.

# Rademacher complexity

$$\begin{aligned} E\|\hat{P}_n - P\|_{\mathcal{F}} &\leq E_{X, X'} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i (f(X_i) - f(X'_i)) \right| \\ &= E_{X, X', \epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i (f(X_i) - f(X'_i)) \right| \\ &\leq E_{X, \epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right| + E_{X', \epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X'_i) \right| \\ &= 2E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right| =: 2\mathcal{R}_{\mathcal{F}} \end{aligned}$$

- $\mathcal{R}_{\mathcal{F}}$  is also called the Rademacher complexity of the function class.

# Why the Rademacher complexity?

- We have now shown that  $\|\hat{P}_n - P\|_{\mathcal{F}} \leq 2\mathcal{R}_{\mathcal{F}} + \epsilon$  with prob.  $1 - e^{-n\epsilon^2/2}$ .
- $\mathcal{R}_{\mathcal{F}}$  measures the maximum possible correlation (over all  $f \in \mathcal{F}$ ) between the vector  $(f(X_1), \dots, f(X_n))$  and the noise vector  $(\epsilon_1, \dots, \epsilon_n)$ .
- If a function class has some function which has a high correlation with a random noise vector, then we should not expect concentration.
- If  $\mathcal{R}_n$  is  $o(1)$  then the Borel Cantelli lemma gives  $\|\hat{P}_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0$ .



## Size of a function class $\mathcal{F}$

- Let  $\mathcal{F}(X) = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\}$
- $\mathcal{R}_{\mathcal{F}} = E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right| = E \left[ E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right| \middle| X_1, \dots, X_n \right]$
- In the next slide we will bound this using the cardinality of  $\mathcal{F}(X)$
- Note that when  $\mathcal{F}$  is the set of indicator functions  $1_{(-\infty, t]}$ ,

$$\begin{aligned} |\mathcal{F}(X)| &= |\{(f(X_{(1)}), \dots, f(X_{(n)})) : f \in \mathcal{F}\}| \\ &= |\{(1(X_{(1)} \leq t), \dots, 1(X_{(n)} \leq t)) : t \in \mathbb{R}\}| \\ &\leq n + 1 \quad (\text{HUH!!}) \end{aligned}$$

## Theorem

Let  $A \subseteq \mathbb{R}^n$ ,  $R = \max_{a \in A} \|a\|$ ,

$$E \sup_{a \in A} \langle \epsilon, a \rangle \leq \sqrt{2R^2 \log |A|}.$$

And,

$$E \sup_{a \in A} |\langle \epsilon, a \rangle| \leq \sqrt{2R^2 \log |2A|}.$$

**Proof.**

$$\begin{aligned}\exp \left( \lambda E \sup_{a \in A} \langle \epsilon, a \rangle \right) &\leq E \exp \left( \lambda \sup_{a \in A} \langle \epsilon, a \rangle \right) \\ &= E \sup_{a \in A} \exp (\lambda \langle \epsilon, a \rangle) \\ &\leq \sum_{a \in A} E \exp (\lambda \langle \epsilon, a \rangle) \\ (\langle \epsilon, a \rangle \sim \text{Subgaussian}(\|a\|_2^2)) \quad &\leq \sum_{a \in A} \exp \left( \frac{\lambda^2 \|a\|_2^2}{2} \right) \\ &\leq |A| \exp \left( \frac{\lambda^2 R^2}{2} \right)\end{aligned}$$

Take  $\lambda = 2 \log |A| / R^2$ .

□

## Proof.

If  $\mathcal{F}$  is the set of one sided indicator functions, then

$$\begin{aligned}\|\hat{P}_n - P\|_{\mathcal{F}} &\leq 2\mathcal{R}_{\mathcal{F}} + \epsilon = 2E[E[\sup_{f \in \mathcal{F}} \sum_i \epsilon_i f(X_i)/n] | X] + \epsilon \\ &\leq \sqrt{8R^2 \log(n+1)} + \epsilon \\ &\leq \sqrt{\frac{8 \log(n+1)}{n}} + \epsilon\end{aligned}$$

By Borel Cantelli,  $\|\hat{P}_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0$

□

