STAT 583: Advanced Theory of Statistical Inference **Spring 2018**

Lecture 1: Big Picture (Addendum)

Lecturer: Fang Han *Apr 02*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Lecturer.*

1.1 First proof of GC Theorem

We consider

$$\mathcal{G} := \{g_t : x \to \mathbb{1}_{(-\infty,t]}(x)\},$$

and can rewrite

$$\sup_t |F_n(t) - F(t)| = \sup_{g\in\mathcal{G}} |E_n g - Eg|.$$

We are now ready to rigorously prove the GC Theorem:

$$E \sup_t |F_n(t) - F(t)| \le 2\sqrt{\frac{2\log 2(n+1)}{n}}.$$

Proof. The proof is twofold. First is a classic symmetrization (we will repeatedly rediscover this trick everywhere). Secondly, we will employ the celebrated Massart's finite class lemma.

Step I. Let $Z_1, \ldots, Z_n$ be an independent copy of $X_1, \ldots, X_n$. We then have

$$\begin{aligned}
E\sup_{g\in\mathcal{G}} |E_n g - Eg| =& E\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^n g(X_i) - \frac{1}{n}Eg(Z_i)| \\
=& E\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^n \{g(X_i) - E[g(Z_i)|X_1,\ldots,X_n]\}| \\
=& E\sup_{g\in\mathcal{G}} |E(\frac{1}{n}\sum_{i=1}^n \{g(X_i) - g(Z_i)\}|X_1,\ldots,X_n)| \\
\le& EE[\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^n (g(X_i) - g(Z_i))||X_1,\ldots,X_n] \\
=& E\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^n (g(X_i) - g(Z_i))|.
\end{aligned}$$

We then employ the Rademacher sequence $\epsilon_1, \ldots, \epsilon_n$, where $\epsilon_i \in \{-1, 1\}$ is symmetric around 0. We then have

$$E\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^n (g(X_i) - g(Z_i))| = E\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(g(X_i) - g(Z_i))\right| \le 2E\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right|,$$

where the last inequality is via the triangle inequality.

Step II. Secondly, we employ Massart's finite class lemma.

Lemma 1 (Massart's finite class lemma)**.** *Let $A \subset \mathbb{R}^n$ with $|A| < \infty$ and $R := \max_{a \in A} \|a\|_2$. We then have*

$$E \max_{a \in A} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right) \leq \frac{R\sqrt{2 \log |A|}}{n}$$

and

$$E \max_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| \leq \frac{R\sqrt{2 \log 2|A|}}{n}$$

Proof. Define $Z_a := \sum_{i=1}^{n} \epsilon_i a_i$. We have

$$\exp(tE \max_{a \in A} Z_a) \leq E \exp(t \max_{a \in A} Z_a) = E \max_{a \in A} \exp(tZ_a) \leq E \sum_{a \in A} \exp(tZ_a).$$

Using Hoeffding's inequality, we have

$$E \sum_{a \in A} \exp(tZ_a) \leq \sum_{a \in A} \exp(t^2 \sum_{i=1}^{n} a_i^2/2) \leq \sum_{a \in A} \exp(t^2 R^2/2) = |A| \exp(t^2 R^2/2).$$

Accoridngly, we have

$$E \max_{a \in A} Z_a \leq \inf_{t > 0} \left(\frac{\log |A|}{t} + \frac{tR^2}{2} \right).$$

Setting $t = \sqrt{2 \log |A| / R^2}$, we have the desired bound for the first term. The second inequality comes from enriching the class A to $\{A, -A\}$. □

Now applying Massart's lemma, we have

$$E \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| = EE \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) | X_1, \ldots, X_n \right| \leq \sqrt{\frac{2 \log 2(n+1)}{n}}.$$

This completes the proof. □

1.2 Second proof of GC Theorem

Theorem 2 (Glivenko-Cantalli)**.** *Suppose $X_1, \ldots, X_n$ be n i.i.d. random variables, and*

$$F_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t) = \frac{1}{n} \sum_{i=1}^{n} Z_i(t) \quad \text{and} \quad F(t) := P(X_i \leq t)$$

are the empirical and population cdf. We then have

$$E \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \leq \frac{4}{\sqrt{n}}.$$

Proof. Using the standard symmetrization argument, we have

$$E \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| = E \sup_{t \in \mathbb{R}} |\frac{1}{n} \sum_{i=1}^{n} Z_i(t) - EZ_i(t)| \leq E \sup_{t \in \mathbb{R}} |\frac{1}{n} \sum_{i=1}^{n} (Z_i(t) - \widetilde{Z}_i(t))| = E \sup_{t \in \mathbb{R}} |\frac{1}{n} \sum_{i=1}^{n} \epsilon_i (Z_i(t) - \widetilde{Z}_i(t))|,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Bernoulli random variables of $P(\epsilon_1 = 1) = P(\epsilon_1 = -1) = 1/2$. We hence have

$$E \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \leq \frac{2}{n} E \sup_{t \in \mathbb{R}} |\sum_{i=1}^{n} \epsilon_i \mathbb{1}(X_i \leq t)| \leq \frac{2}{n} E \max_{\ell \leq n} |\sum_{k=1}^{\ell} \epsilon_k| \leq \frac{2}{n} (E \max_{\ell \leq n} |\sum_{k=1}^{\ell} \epsilon_k|^2)^{1/2}.$$

We then employ Doob's L_p maximal inequality.

Theorem 3 (Doob's L_p maximal inequality, Theorem (4.3) of Chapter 4.4 in D2005)**.** *If Y_n is a martingale, then for any $p \in (1, \infty)$,*

$$E \max_{m \leq n} |Y_m|^p \leq \left(\frac{p}{p-1}\right)^p E(|Y_n|^p).$$

which yields

$$E \max_{\ell \leq n} |\sum_{k=1}^{\ell} \epsilon_k|^2 \leq 4E|\sum_{k=1}^{n} \epsilon_k|^2 = 4n,$$

and hence completes the proof. □