Stat210B: Theoretical Statistics **Lecture Date: February 01, 2007**

Lecture 4

Lecturer: Michael I. Jordan *Scribe: Sriram Sankararaman*

Empirical Process theory allows us to prove uniform convergence laws of various kinds. One of the ways to start Empirical Process theory is from the Glivenko-Cantelli theorem. Recall the Glivenko-Cantelli theorem.

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\xi_i \leq t\}} \quad (1)$$

$$F(t) = \mathbb{P}\{\xi \leq t\} \quad (2)$$

We would like to show that $\sup_t |F_n(t) - F(t)| \xrightarrow{P} 0$. The proof makes use of the compactness of the class of indicator functions on the real line to break this class into bins and bound the oscillations in each bin. This leads to the question of whether the same idea can be generalized to other function classes.

1 Empirical Process Theory

Denote $\sup_t |\cdot|$ by $||\cdot||$. To bound the difference $||F_n(t) - F(t)||$, we compare two independent copies of the empirical quantity - $F_n(t)$ and $F'_n(t)$. A symmetrization lemma is used to bound the former in terms of the latter.

1.1 First Symmetrization

Lemma 1. *(Pollard, 1984, Section II.8, p. 14) Let $Z(t)$ and $Z'(t)$ be independent stochastic processes. Suppose that $\exists \alpha, \beta > 0$ such that $\mathbb{P}\{|Z'(t)| \leq \alpha\} \geq \beta, \quad \forall t$. Then*

$$\mathbb{P}\{\sup_t |Z(t)| > \epsilon\} \leq \beta^{-1}\mathbb{P}\{\sup_t |Z(t) - Z'(t)| > \epsilon - \alpha\} \quad (3)$$

An application of Lemma 1 can be seen by setting $Z(t) = F_n(t) - F(t)$ and $Z'(t) = F'_n(t) - F(t)$.

Proof. Suppose that the event $\{\sup_t |Z(t)| > \epsilon\}$ occurs. Choose $\tau \ni |Z(\tau)| > \epsilon$. Note that τ is a random variable. By definition of τ,

$$\mathbb{P}\{\sup_t |Z(t)| > \epsilon\} \leq \mathbb{P}\{|Z(\tau)| > \epsilon\} \quad (4)$$

From the independence of Z and Z', we have

$$\mathbb{P}\{|Z'(t)| < \alpha | Z\} \geq \beta \quad (5)$$

Suppose that both $\{|Z(\tau)| > \epsilon\}$ and $\{|Z'(\tau)| \leq \alpha\}$ occur. Then we have

$$\{|Z(\tau) - Z'(\tau)| \geq \epsilon - \alpha\} \quad (6)$$

Also

$$\begin{aligned}\beta\{|Z(\tau)| > \epsilon\} &\leq \mathbb{P}\{|Z'(\tau)| \leq \alpha, |Z(\tau)| > \epsilon | Z\} \\ &\leq \mathbb{P}\{|Z'(\tau)| > \alpha, |Z(\tau)| > \epsilon\} \qquad (7)\end{aligned}$$

Here $\{|Z(\tau)| > \epsilon\}$ is an indicator function on the event $\{|Z(\tau)| > \epsilon\}$. The inequality 7 uses the independence of Z and Z'. From Equation 6

$$\begin{aligned}\beta\{|Z(\tau)| > \epsilon\} &\leq \mathbb{P}\{|Z'(\tau) - Z(\tau)| \geq \epsilon - \alpha\} \\ &\leq \mathbb{P}\{\sup_t |Z(t) - Z'(t)| \geq \epsilon - \alpha\} \qquad (8)\end{aligned}$$

The proof follows from Equations 8 and 4. □

1.1.1 Example

$$U_n(\omega, t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} (\{\xi_i(\omega) \leq t\} - t)$$

where $\xi_i \overset{iid}{\sim} Unif(0, 1)$.

For fixed value of t,

$$U_n \sim \frac{Bin(n, t)}{n^{\frac{1}{2}}} - \frac{t}{n^{\frac{1}{2}}}$$

$$\begin{aligned}\mathbb{P}\{|F_n(t) - F(t)| > \frac{\epsilon}{2}\} &\leq \frac{4}{\epsilon^2} E(F_n(t) - F(t))^2 \\ &= \frac{4}{\epsilon^2} E(\frac{1}{n} \sum_i \{\xi \leq t\} - F(t))^2 \\ &= \frac{4}{n\epsilon^2} E(\{\xi \leq t\} - F(t))^2 \\ &= \frac{4F(t)(1 - F(t))}{n\epsilon^2} \\ &\leq \frac{1}{n\epsilon^2} \\ &= \frac{1}{2} \quad \text{for} \quad n \geq \frac{2}{\epsilon^2}\end{aligned}$$

1.2 Second Symmetrization

The second symmetrization lemma allows us to replace the difference $F_n - F_n'$ with a single empirical quantity consisting of n observations. We can further bound the latter so that the bound is independent of the data ξ.

Define Rademacher variables $\{\sigma_i\} \overset{iid}{\in} \{-1, +1\}$. For any choice of $\{\sigma_i\}$, the distribution of $(\{\xi_i \leq t\} - \{\xi_i' \leq t\})$ is equal to the distribution of $\sigma_i(\{\xi_i \leq t\} - \{\xi_i' \leq t\})$. We change notation here so that $P_n = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\xi_i \leq t\}}$. P_n' is defined similarly.

Lemma 2. *Pollard (1984, II.8,p. 15)* $\mathbb{P}\{||P_n - P'_n|| > \frac{\epsilon}{2}\} \leq 2\mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}| \geq \frac{\epsilon}{4}\}$

Proof.

$$\begin{aligned}
&\mathbb{P}\{||P_n - P'_n|| > \frac{\epsilon}{2}\} \\
&\qquad = \mathbb{P}\{\frac{1}{n}\underset{t}{sup}|\sum_i \sigma_i(\{\xi_i \leq t\} - \{\xi'_i \leq t\})| \geq \frac{\epsilon}{2}\} \\
&\qquad \leq \mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}| \geq \frac{\epsilon}{4}\} + \mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi'_i \leq t\}| \geq \frac{\epsilon}{4}\} \qquad (9) \\
&\qquad = 2\mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}| \geq \frac{\epsilon}{4}\}
\end{aligned}$$

□

Inequality 9 was derived using the equivalence of the two random quantities and the triangle inequality.

1.3 Hoeffding bound for independent RVs

We state here the Hoeffding bound which we use to bound the quantity $\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}$. Consider n independent RVs $\{Y_i\}$s so that $EY_i = 0$ and $a_i \leq Y_i \leq b_i$.

Theorem 3 (Hoeffding Bound). $\mathbb{P}\{\sum_{i=1}^n Y_i > \eta\} \leq 2e^{-\frac{2\eta^2}{\sum_i (b_i - a_i)^2}}$

The proof proceeds by considering the random variable $e^{s\sum_i Y_i}$ where s is a free parameter. Using Markov's inequality,

$$\begin{aligned}
\mathbb{P}\{e^{s\sum_i Y_i} > e^{s\eta}\} &\leq \frac{Ee^{s\sum_i Y_i}}{e^{s\eta}} \\
&\leq \frac{\prod_i Ee^{sY_i}}{e^{s\eta}}
\end{aligned}$$

Minimizing s gives the necessary bound.

References

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.