Stat210B: Theoretical Statistics **Lecture Date: February 22, 2007**

Lecture 11: Donsker Theorem

Lecturer: Michael I. Jordan *Scribe: Chris Haulk*

This lecture is devoted to the proof of the Donsker Theorem. We follow Pollard, Chapter 5.

1 Donsker Theorem

Theorem 1 (Donsker Theorem: Uniform case). *Let $\{\xi_i\}$ be a sequence of iid Uniform[0,1] random variables. Let*

$$U_n(t) = n^{-1/2}\sum_{i=1}^{n}[\{\xi_i \le t\} - t] \qquad \text{for} \qquad 0 \le t \le 1$$

be the empirical process associated with the $\{\xi_i\}$. Then

$$U_n \xrightarrow{d} U$$

where U is the brownian bridge on [0,1].

Proof. Recall that the brownian bridge U is a process on [0,1] with continuous sample paths, with $U(0) = U(1) = 0$, and with $\pi_S U$ being multivariate normal with mean zero and covariances $PU(s)U(t) = s(1-t)$ for $s \le t, s, t \in S$, for any finite subset $S \subset \mathbf{R}$.

By Theorem 3 of Pollard, Chapter 5, it is enough to check fidi convergence and asymptotic equicontinuity. First we treat fidi convergence, which is almost trivial:

$$PU_n(t) = 0$$

and if $s \le t$,

$$\begin{aligned}
PU_n(s)U_n(t) &= \frac{1}{n}\sum_{i,j\le n} P[(\{\xi_i \le s\} - s)(\{\xi_j \le t\} - t)] \\
&= \frac{1}{n}\sum_{i\le n} P[(\{\xi_i \le s\} - s)(\{\xi_i \le t\} - t)] \\
&= \frac{1}{n}\sum_{i}(P\{\xi_i \le s\} - tP\{\xi_i \le s\} - sP\{\xi_i \le t\}) + st \\
&= s(1-t)
\end{aligned}$$

So fidi convergence follows from the multivariate CLT. Recall that asymptotic equicontinuity means that to each $\epsilon > 0$ and each $\delta > 0$ there corresponds a grid $0 = t_0 < t_1 < ... < t_m = 1$ such that

$$\limsup_n P\left\{\max_i \sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\} < \epsilon$$

By the union bound,

$$P\left\{\max_i \sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\} \le \sum_{i=0}^{m-1} P\left\{\sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\}$$

For fixed $\epsilon, \delta > 0$, we will choose an equally-spaced grid $0 = t_0 < t_1 < ... < t_m = 1$ fine enough that

$$\sum_{i=0}^{m-1} P\left\{\sup_{J_i} |U_n(t) - U_n(t_i)| > \delta\right\} < \delta$$

By symmetry, all the summands are equal to the first, so the sum reduces to

$$mP\left\{\sup_{0\le t\le b} |U_n(t)| > \delta\right\}$$

where $b = m^{-1}$. We will show that this can be done so that the expression above goes to 0 as $n \uparrow \infty$. Let $\mathcal{E}_t = \sigma\{U_n(s) : 0 \le s \le t\}$. The number K of ξ_i that land in $[0, t]$ is $\mathcal{E}_t$-measurable. Conditional on $K = k$ data points falling in $[0, t]$, the remaining data points are uniform on $(t, 1]$. That is, on $\{U_n(t) = n^{-1/2}(k - nt)\}$, $U_n(b) - U_n(t)$ has a $n^{-1/2}(Bin(n - k, \theta) - n(b - t))$ distribution, where $\theta = (b - t)/(1 - t)$.

Apply Chebychev's inequality on the set $\{|U_n(t)| > \delta\} = \{|k - nt| > n^{1/2}\delta\}$:

$$\begin{aligned}
P\{|U_n(b) - U_n(t)| > \frac{1}{2}|U_n(t)| \mid \mathcal{E}_t\} &= P\{|Bin(n - k, \theta) - n(b - t)| > \frac{1}{2}|k - nt|\} \\
&\le 4\left[(n - k)\theta(1 - \theta) + [(n - k)\theta - n(b - t)]^2\right]/(k - nt)^2 \\
&\le 4n\theta/(k - nt)^2 + 4\theta^2 \\
&\le [4b/(1 - b)]/\delta^2 + 4b^2/(1 - b)^2 \\
&\le 1/2 \text{ for small enough } b
\end{aligned}$$

Now appeal to Lemma 7 from Pollard, Chapter 5:

Lemma 2. *Let $\{Z(t) : 0 \le t \le b\}$ be a process with cadlag paths and $Z(0) \equiv 0$. Furthermore suppose $Z(t)$ is $\mathcal{E}_t$-measurable for an increasing family of σ-fields $\{\mathcal{E}_t : 0 \le t \le b\}$. If*

$$P\left\{|Z(b) - Z(t)| \le \frac{1}{2}|Z(t)| \mid \mathcal{E}_t\right\} \ge \beta$$

on $\{|Z(t)| > \delta\}$, then

$$P\left\{\sup_{0\le t\le b} |Z(t)| > \delta\right\} \le \beta^{-1} P\{|Z(b)| > \frac{1}{2}\delta\}$$

Applying this lemma with $\beta = 1/2$ and m large enough, we get

$$mP\{\sup_{0\le t\le b} |U_n(t)| > \delta\} \le 2mP\{|U_n(b)| > \frac{1}{2}\delta\}$$

For fixed, large m, take $n \uparrow \infty$, then $U_n(b) \xrightarrow{d} N(0, b(1 - b))$ by the CLT, and the right-hand side of the expression above converges to (recall that $b = m^{-1}$)

$$\begin{aligned}
2mP(|N(0, b(1 - b))| > \frac{1}{2}\delta) &\le 2m[(b - b^2)^2(\frac{1}{2}\delta)^{-4}]P|N(0, 1)| \\
&\le 32m^{-1}\delta^{-4}P|N(0, 1)|^4 \\
&\le \epsilon \text{ for large enough } m. \qquad (1)
\end{aligned}$$

□

To extend this to distributions other than Uniform[0,1], let $\eta_i \stackrel{iid}{\sim} F$ where F is any distribution function, and set

$$E_n(r) = n^{-1/2}[F_n(r) - F(r)] = n^{-1/2}\sum_{i=1}^{n}[\{\eta_i \leq r\} - F(r)]$$

with $E_n(-\infty) = 0$, $E_n(\infty) = 1$. Note that $PE_n(r) = 0$ and if $s \leq t$, $PE_n(s)E_n(t) = F(s) - F(s)F(t)$.

Theorem 3 (Donsker Theorem). *The empirical processes* $\{E_n(r) : \infty \leq r \leq \infty\}$ *converge in distribution, as random elements of* $D[-\infty, \infty]$ *to a mean zero Gaussian process* E*, with* $PE(s_i)E(s_j) = F(s_i \wedge s_j) - F(s_i)F(s_j)$*.*

Proof. Define a map $H : D[0,1] \to D[-\infty, \infty]$ by $(Hx)(r) = x(F(r))$ for all $x \in D[0,1]$. H is (uniformly) continuous because

$$||Hx - Hy|| = \sup_r |x(F(r)) - y(F(r))| \leq \sup_t |x(t) - y(t)| = ||x - y||$$

By the uniform case of the Donsker theorem and the continuous mapping theorem, $HU_n \xrightarrow{d} HU$. Let Q be the quantile function associated with F; then $\xi_i \leq F(r)$ if and only if $Q(\xi_i) \leq r$. Thus

$$\begin{aligned}(HU_n)(r) &= n^{-1/2}\sum_{i=1}^{n}[\{\xi_i \leq F(r)\} - F(r)] \\ &= n^{-1/2}\sum_{i=1}^{n}[\{Q(\xi_i \leq r\} - F(r)] \\ &\stackrel{d}{=} E_n(r)\end{aligned}$$

where the final equality follows from noticing that $Q(\xi_i \leq r\} \stackrel{d}{=} \eta_i$. So $E_n \xrightarrow{d} HU$. Finally, HU is cadlag because U is continuous and F is cadlag, $P(HU)(s) = PU(F(s)) = 0$, and for any finite set of $s_i \in \mathbf{R}$, $P(HU)(s_i)(HU)(s_j) = PU(F(s_i))U(F(s_j)) = F(s_i) \wedge F(s_j) - F(s_i)F(s_j) = F(s_i \wedge s_j) - F(s_i)F(s_j)$.

□

References

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