Stats 300b: Theory of Statistics **Winter 2017**

Lecture 15 – February 28

Lecturer: John Duchi *Scribe: Fang Cai*

Warning: *these notes may contain factual errors*

Reading: Van der Vaart Chapter 18 and 19.

Outline:

- Uniform laws/limits
  - Chaining and equicontinuity (entropy integrals)
  - Uniform CLTs and Brownian Bridge
  - Donsker classes (uniform CLTs)
- Moduli of continuity and convergence rate

1 Uniform laws/limits

Recall from last lecture

Theorem 1 (Uniform limits via stochastic equicontinuity)**.** *Stochastic process $(X_{n,t})_{t\in T}$ converges in distribution to $(X_t)_{t\in T}$ in $L_\infty(T)$ iff*

(i) Finite dimensional convergence (FIDI):

$$(X_{n,t_1}, \ldots, X_{n,t_k}) \xrightarrow{d} (X_{t_i}, \ldots, X_{t_k}).$$

(ii) Equicontinuity: $\forall \epsilon > 0, \eta > 0, \exists$ partition $T_1, \ldots, T_m$ of T s.t.

$$\limsup_{n\to\infty} \mathbb{P}(\max_i \sup_{s,t\in T_i} |X_{n,s} - X_{n,t}| \geq \epsilon) \leq \eta.$$

Usually (i) is free by CLT but game is in showing (ii), i.e. moduli of continuity of $X_n \in L_\infty(T)$

Corollary 2 (Uniform limits via entropy integral)**.** *Suppose (T, d) is a totally bounded metric space with $\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}(\sup_{d(s,t)\leq\delta} |X_{n,s} - X_{n,t}| \geq \epsilon) = 0$ and FIDI of X_n to $X \in L_\infty(T)$. Then $(X_{n,t})_{t\in T} \xrightarrow{d} (X_t)_{t\in T}$ in $L_\infty(T)$.*

Proof Let $\epsilon > 0$. Choose partition $\{T_i\}_{i=1}^m$ of T such that $\mathrm{diam}(T_i) \leq \delta$. Then

$$\max_{i\leq m} \sup_{s,t\in T_i} |X_{n,s} - X_{n,t}| \leq \sup_{d(s,t)\leq\delta} |X_{n,s} - X_{n,t}|,$$

so the process is stochastic equicontinuous. □

Definition 1.1 (Donster Class)**.** *A collection of functions $\mathcal{F}$ is $\mathbb{P}$-Donsker if the process $(\sqrt{n}(\mathbb{P}_n - \mathbb{P})f)_{f\in\mathcal{F}}$ converges to a tight limit in $L_\infty(\mathcal{F})$.*

This limit is a Gaussian process because $\sqrt{n}(\mathbb{P}_n - \mathbb{P})f \xrightarrow{d} \mathsf{N}(0, \mathrm{Var}_p(f)))$ by CLT, so FIDI gives that limit process, denoted by $\mathbb{G}_p \in L_\infty(\mathcal{F})$ (i.e $\mathbb{G}_p f \in \mathbb{R}$).

$$\mathbb{E}[\mathbb{G}_p f] = 0$$
$$\mathbb{E}[\mathbb{G}_p f \mathbb{G}_p g] = \underset{p}{\mathrm{Cov}}(f,g) = \mathbb{P}fg - \mathbb{P}f\,\mathbb{P}g$$

Idea: If $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - \mathbb{P})$, where $X_i \overset{\text{iid}}{\sim} \mathbb{P}$, $\mathbb{G}_n \in L_\infty(\mathcal{F})$. Then

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - \mathbb{P})f = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}_p[f(X_i)])$$

and

$$(\mathbb{G}_p f_1, \dots, \mathbb{G}_p f_k) \sim \mathsf{N}\left(0,\, [\underset{p}{\mathrm{Cov}}(f_i, f_j)]_{i,j=1}^k\right)$$

Example 1 ($\mathbb{P}$-Brownian bridge)**:** $F(t) = \mathbb{P}(X \le t)$. Let $\mathcal{F} = \{\mathbf{1}\{\cdot \le t\}, t \in \mathbb{R}\}$. Then $\sqrt{n}(F_n(\cdot) - F(\cdot)) \xrightarrow{d} \mathbb{G}_p$ in $L_\infty(\mathbb{R})$. Note that

$$\mathbb{E}[\mathbf{1}\{X \le t\}\mathbf{1}\{X \le s\}] = F(s\wedge t) \implies \mathrm{Cov}(\mathbf{1}\{X \le t\}, \mathbf{1}\{X \le s\}) = F(s\wedge t) - F(S)F(t).$$

Furthermore, if $\mathbb{P} = \mathrm{Uniform}([0,1])$, get $\mathbb{G}_p$ with $\mathrm{Cov}(\mathbb{G}_p(t), \mathbb{G}_p(s)) = s\wedge t - st$ and $\mathbb{G}_P(t) \sim \mathsf{N}(0, t(1-t))$. ♣

Theorem 3 (Uniform CLT via entropy integral)**.** *Let $\mathcal{F}$ be a collection of functions. There exists an envelop function $B : X \to \mathbb{R}_+$, with*

$$\int_0^\infty \sup_Q \sqrt{\log N(\mathcal{F},\, L_2(Q),\, \|B\|_{L_2(Q)}\,\epsilon)}\, d\epsilon < \infty$$

where the supremum is take over all discrete probability measure. If $\mathbb{P}B^2 < \infty$, then $\mathcal{F}$ is $\mathbb{P}$-Donsker.

Sketch of Proof Let $\mathcal{F}_\delta := f - g : \|f-g\|_{L_2(\mathbb{P})} \le \delta,\, f, g \in \mathcal{F}$. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ and $\mathbb{P}_n^o = \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbf{1}\{X_i\}$, then

$$\mathbb{P}\left(\sup_{\|f-g\|\le\delta} |\mathbb{G}_n(f-g)| \le \epsilon \ge \epsilon\right) \le \frac{\mathbb{E}\left[\sup_{\|f-g\|\le\delta} |\sqrt{n}(\mathbb{P}_n - \mathbb{P})(f-g)|\right]}{\epsilon}$$
$$\le \frac{2}{\epsilon}\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta} \sqrt{n}|\mathbb{P}_n^o f|\right]$$

by symmetrization. By Dudley's integral,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta} \sqrt{n}|\mathbb{P}_n^o f|\right] \le C\mathbb{E}\left[\int_0^\infty \sqrt{\log N(\mathcal{F}_\delta, L_2(\mathbb{P}_n), \epsilon)}\, d\epsilon\right]$$

If we set $\theta_n = \sup_{f \in \mathcal{F}_\delta} \|f\|_{L_2(\mathbb{P}_n)}$ and $b_n = \|B\|_{L_2(\mathbb{P}_n)}$, then the above

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}_\delta} \sqrt{n}|\mathbb{P}_n^o f|\right] \leq & C\mathbb{E}\left[\int_0^{\theta_n} \sqrt{\log N(\mathcal{F}_\delta, L_2(\mathbb{P}_n), \epsilon)}\, d\epsilon\right] \\
\leq & Cb_n\mathbb{E}\left[\int_0^{\theta_n/b_n} \sqrt{\log N(\mathcal{F}_\delta, L_2(\mathbb{P}_n), b_n\epsilon)}\, d\epsilon\right]
\end{aligned}$$

Note that $N(\mathcal{F}_\delta, L_2(Q), \epsilon) \leq N^2(\mathcal{F}, L_2(Q), \epsilon)$. (Consider all pairs $(f-g)$ in $\mathcal{F}$) so

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}_\delta} \sqrt{n}|\mathbb{P}_n^o f|\right] &\leq Cb_n\mathbb{E}\left[\int_0^{\theta_n/b_n} \sqrt{\log N(\mathcal{F}, L_2(\mathbb{P}_n), b_n\epsilon)}\, d\epsilon\right] \\
&\leq Cb_n\mathbb{E}\left[\int_0^{\theta_n/b_n} \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|B\|_Q\, \epsilon)}\, d\epsilon\right]
\end{aligned}$$

By the our entropy integral condition, we have

$$\int_0^a \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|B\|_Q\, \epsilon)}\, d\epsilon \to 0 \quad \text{as} \quad a \downarrow 0$$

Knowing that $\mathcal{F}$ is Glivenko-Canteli, get $\theta_n \xrightarrow{p} 0$ as $\delta \downarrow 0$. Note that $b_n \xrightarrow{p} \mathbb{E}\|B\|_{L_2(P)} > 0$ as $n \to \infty$. Again using our entropy integral condition, we have

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{\|f-g\| \leq \delta} |\mathbb{G}_n(f-g)| \leq \epsilon \geq \epsilon\right) \\
\leq &\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \frac{Cb_n}{\epsilon}\mathbb{E}\left[\int_0^{\theta_n/b_n} \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|B\|_Q\, \epsilon)}\, d\epsilon\right] = 0.
\end{aligned}$$

Then $\mathcal{F}$ is $\mathbb{P}$-Donsker follows from Corollary 2. □

2 M-Estimator Revisited

We will use all of our uniform laws, CLTs to get

(i) consistency result

(ii) rate of convergence using moduli of continuity.

Definition 2.1. *Criterion function* $m_\theta : \mathcal{X} \to \mathbb{R}$. *Let* $M_n(\theta) := \mathbb{P}_n m_\theta$ *and* $M(\theta) = \mathbb{P}m_\theta$. *Then* $\widehat{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta}$ *is the* ***M-estimator***.

Criterion function $\psi_\theta : \mathcal{X} \to \mathbb{R}$. *Let* $\Psi_n(\theta) := \mathbb{P}_n \psi_\theta$ *and* $\Psi(\theta) = \mathbb{P}\psi_\theta$. *Then* $\widehat{\theta}_n$ *s.t.* $\Psi_n(\widehat{\theta}_n) = 0$ *is the* ***Z-estimator***. *(Often* $\psi_\theta(x) = \nabla_\theta m_\theta(x)$ *)*

2.1 Consistency

Theorem 4 (Argmax consistency theorem)**.** *Suppose we have*

1. *Uniform law of large number(ULLN):*

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$$

2. *Separation:*

$$\sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0)$$

for all $\epsilon > 0$.

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

Proof □

Fix $\epsilon > 0$. Let $\eta(\epsilon) := M(\theta_0) - \sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta)$. Then

$$d(\widehat{\theta}_n, \theta_0) \geq \epsilon \implies M(\theta_0) \geq M(\widehat{\theta}_n) + \eta(\epsilon)$$

but then

$$\begin{aligned} M_n(\theta_0) \geq & M(\theta_0) - o_p(1) \\ \geq & M(\widehat{\theta}_n) + \eta(\epsilon) - o_p(1) \\ \geq & M_n(\widehat{\theta}_n) + \eta(\epsilon) - o_p(1) \end{aligned}$$

by ULLN. Hence,

$$\mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \geq \epsilon\right) \leq \mathbb{P}\left(M_n(\theta_0) \geq M_n(\widehat{\theta}_n) + \eta(\epsilon) - o_p(1)\right) \to 0.$$

2.2 Rate of convergence via moduli of continuity

Idea: if $M(\theta))$ shrinks quickly way from θ_0, but $(M_n(\theta) - M(\theta)) - (M_n(\theta_0)) - M(\theta_0))$ are well behaved (i.e. have small error), then the M-estimator $\widehat{\theta}_n = \operatorname{argmax}_\theta M_n(\theta)$ cannot be bad.

Proof of non-random case:

Suppose

$$M(\theta_0) \geq M(\theta) + \|\theta - \theta_0\|^2,$$

which is reasonable by Taylor expansion $\nabla M(\theta_0) = 0, \nabla^2 M(\theta_0) < 0$. Suppose we have local error guarantee,

$$\sup_{\|\theta - \theta_0\| \leq \delta} |(M_n(\theta) - M(\theta)) - (M_n(\theta_0) - M(\theta_0))| \leq \frac{\phi(\delta)}{\sqrt{n}}$$

for some ϕ with $\phi(\delta) \to 0$ as $\phi \to 0$, i.e. modulus of continuity near θ_0.

Let $\widehat{\theta}_n = \operatorname{argmax}_\theta M_n(\theta)$ and assume consistency $\widehat{\theta}_n \xrightarrow{p} \theta_0$. Then

$$\begin{aligned} M_n(\widehat{\theta}_n) \geq & M_n(\theta_0) \\ = & M_n(\theta_0) - M(\theta_0) + M(\widehat{\theta}_n) + M(\theta_0) - M(\widehat{\theta}_n) \\ \geq & M_n(\theta_0) - M(\theta_0) + M(\widehat{\theta}_n) + \left\|\widehat{\theta}_n - \theta_0\right\|^2 \end{aligned}$$

i.e.

$$
\begin{aligned}
0 \geq & -\left((M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n)) - (M_n(\theta_0) - M(\theta_0))\right) + \left\|\widehat{\theta}_n - \theta_0\right\|^2 \\
\geq & \frac{-\phi\left(\left\|\widehat{\theta}_n - \theta_0\right\|\right)}{\sqrt{n}} + \left\|\widehat{\theta}_n - \theta_0\right\|^2.
\end{aligned}
$$

Rearranging gives

$$
\left\|\widehat{\theta}_n - \theta_0\right\|^2 \leq \frac{\phi\left(\left\|\widehat{\theta}_n - \theta_0\right\|\right)}{\sqrt{n}}. \tag{1}
$$

Let $r_n^2 \geq \frac{\phi\left(\frac{1}{r_n}\right)}{\sqrt{n}}$. If $\left\|\widehat{\theta}_n - \theta_0\right\| > \frac{1}{r_n}$, then contradicts (1) , so we have

$$
\left\|\widehat{\theta}_n - \theta_0\right\| \leq \frac{1}{r_n}.
$$

□

Example: Let $\phi(\delta) = \delta$. Then $\left\|\widehat{\theta}_n - \theta_0\right\|^2 \leq \frac{\|\widehat{\theta}_n - \theta_0\|}{\sqrt{n}}$. Then $\left\|\widehat{\theta}_n - \theta_0\right\| \leq \frac{1}{\sqrt{n}}$. ♣