Stats 300b: Theory of Statistics	Winter 2018

Lecture 16 – March 1

Lecturer: John Duchi *Scribe: Michael Hahn*

Warning: *these notes may contain factual errors*

Reading: Van der Vaart, Chapters 19.3 and 5.3

Outline:

- Applications of last lecture's theorem: Goodness-of-fit statistics
- Rates of Convergence for M-estimators based on nondifferen- tiable losses

1 Goodness-of-fit statistics

Let $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - \mathbb{P})$, viewed as a function on a function class $\mathcal{F}$: $\mathbb{G}_n F := \frac{1}{\sqrt{n}}(\sum_{i=1}^n f(x_i) - Pf)$.

Definition 1.1 (Donsker 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 $L^\infty(\mathcal{F})$.*

As discussed in the previous lecture, this limit is a Gaussian process.

Goodness-of-fit statistics address the testing problem when the null hypothesis is that the data comes from a given distribution: $H_0 : X \sim_{iid} \mathbb{P}$. We can use the theorem from the last lecture to show asymptotic properties of such tests.

Example: Kolmogorov-Smirnoff Test Define the *Kolmogorov-Smirnoff Test*: Let F be CDF of $X \in \mathbb{R}$, and let F_n be the empirical CDF. Define the test statistic

$$K_n := \sqrt{n}\|F_n - F\|_\infty = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

Corollary 1 (Corollary of Theorem from Last Lecture)**.** *Let K_n be the Kolmogorov-Smirnoff test statistic, and let $X_i \sim_{iid} \mathbb{P}$. Then $K_n \xrightarrow{d} \|\mathbb{G}_p\|_\infty$, where $\mathbb{G}_p$ is the limiting Gaussian process (Brownian bridge) with*

$$Cov(\mathbb{G}_t, \mathbb{G}_s) = F(s \wedge t) - F(t)F(s)$$

This limit is independent of the CDF F of $\mathbb{P}$ if F is continuous.

Proof For the function class $\mathcal{F} := \{1\{\cdot \le t\} : t \in \mathbb{R}\}$, the constant $F(x) := 1$ is an envelope function with a second moment, and $\int \sup_Q \sqrt{logN(\mathfrak{F}, L^2(Q), \|F\|_{L^2(Q)}\epsilon)} d\epsilon < \infty$ where Q runs over the finitely supported measures on X. So we can apply the theorem from last lecture and see that $\mathcal{F}$ is Donsker.

Applied to functions $f : \mathbb{R} \to \mathbb{R}$, the map $f \mapsto \sup_{t \in \mathbb{R}} |f(t)|$ is $\|\cdot\|_\infty$-continuous, so the continuous mapping theorem implies $K_n = \|\mathbb{G}_n\|_\infty \xrightarrow{d} \|\mathbb{G}_p\|_\infty$. To show independence from $\mathbb{P}$: Let $\lambda := Uniform([0,1])$. We have

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |\mathbb{G}_\lambda \circ F(t)| \ge \alpha\right) = \mathbb{P}\left(\sup_{u \in [0,1]} |\mathbb{G}_\lambda(u)| \ge \alpha\right)$$

Therefore, $\mathbb{G}_p =_d \mathbb{G}_\lambda \circ F$. □

Example: Cramér-von Mises statistic Define the Cramér-von Mises statistic:

$$C_n := n \int (F_n - F)^2 dF$$

Corollary 2. *$C_n \xrightarrow{d} \int \mathbb{G}_F^2 dF$. If F is continuous, then the limit is independent of F.*

Proof For $f \in L^\infty(\mathbb{R})$, the map $f \mapsto \int f^2 dF$ satisfies

$$|\int f^2 dF - \int g^2 dF| \leq \int |f-g||f+g| dF \leq \|f-g\|_\infty \|f+g\|_\infty$$

and is thus continuous in the supremum norm. Then $C_n \xrightarrow{d} \int \mathbb{G}_F^2 dF$ by the continuous mapping theorem.

Note that, if F is continuous, $\int \mathbb{G}_F^2(t) dF(t) = \int \mathbb{G}_\lambda^2(F(t)) dF(t) = \int_0^1 \mathbb{G}_\lambda^2(u) du$ by substituting $u = f(t)$. □

2 Rates of Convergence for M-estimators based on nondifferentiable losses

Example 1: The loss $\ell(\theta, x) := |\theta - x|$, $R(\theta) := \mathbb{E}[\ell(\theta, x)]$ is minimized by any median if X has a first moment. ♣

Example 2: The loss $\ell(\theta, x) := (1-\alpha)(\theta - x)_+ + \alpha(x - \theta)_+$, where $\alpha \in (0,1)$, $[t]_+ := max(t, 0)$, is minimized at the α-quantiles:

$$Q_{\mathbb{P}}(\alpha) := \inf\{\theta \in \mathbb{R} : \alpha \leq P(X \leq \theta)\}$$

♣

Goal Get analogues of classical conditions (via Taylor expansions) such that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{-1}{\sqrt{n}} I_{\theta_0}^{-1} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_P(1)$$

Step 1 Get an in-probability analogue of Taylor approximations even when ℓ is not differentiable. Suppose that $\ell_\theta : \mathcal{X} \to \mathbb{R}$ is locally-Lipschitz, i.e., for any θ_1, θ_2 in a neighborhood of the (fixed) pint θ_0

$$|\ell(\theta_1, x) - \ell(\theta_2, x)| \leq \dot{\ell}(x) \cdot \|\theta_1 - \theta_2\|$$

(for some $\dot{\ell} : \mathcal{X} \to X$). Moreover, assume that for $\mathbb{P}$-almost every $x \in \mathcal{X}$, $\theta \mapsto \ell(\theta, x)$ is differentiable at θ_0 with derivative $\dot{\ell}(\theta_0, x) = \frac{d}{d\theta}\ell(\theta, x)|_{\theta=\theta_0}$.
Example 3: For $\ell(\theta, x) := |\theta - x|$, having a density near the median θ_0 suffices. ♣

Lemma 3 (19.31 in Van der Vaart)**.** *If $P\dot{\ell}^2 < \infty$, then for all sequences $r_n \to \infty$, we have*

$$\sup_{h\in\mathbb{R}^d \|h\|\leq 1} \mathbb{G}_n(r_n(\ell_{\theta_0+\frac{h}{r_n}} - \ell_{\theta_0}) - h^T\dot{\ell}_{\theta_0}) \xrightarrow{p} 0$$

This says that, locally, we have accurate Taylor approximations.

Remark If this holds, then for any h_n (random or not) such that $h_n = O_p(1)$, we have

$$\mathbb{G}_n\left(r_n\left(\ell_{\theta_0+\frac{h_n}{r_n}} - \ell_{\theta_0}\right) - h_n^T\ell_{\theta_0}\right) \xrightarrow{p} 0$$

Proof We show that the process defined in the lemma has finite-dimensional convergence (FIDI) to 0, and is tight over $\|h\| \leq 1$. Define

$$e_n(h,x) := r_n\left(\ell_{\theta_0+\frac{h}{r_n}}(x) - \ell_{\theta_0}(x)\right) - h^T\ell_{\theta_0}(x)$$

We know $e_n(h,x) \to 0$ as $n \to \infty$ for $\mathbb{P}$-almost all $x \in \mathcal{X}$ by almost-everywhere-differentiability. Note also

$$r_n\left(\ell_{\theta_0+\frac{h}{r_n}}(x) - \ell_{\theta_0}(x)\right) \leq \dot{\ell}(x)\cdot\|h\|$$

for n large by assumption of local-Lipschitzness. Using that $\mathbb{E}[\mathbb{G}_n] = 0$, we get

$$Var(\mathbb{G}_n(e_n(h,x))) \leq \mathbb{E}\left[\left(r_n\left(\ell_{\theta_0+\frac{h}{r_n}}(x) - \ell_{\theta_0}(x)\right) - h^T\dot{\ell}_{\theta_0}(x)\right)^2\right] \to 0$$

as $n \to \infty$ by dominated convergence: A dominating function is $\left(\dot{\ell}(x)\|h\| + \|h\|\dot{\ell}(x)\right)^2$, which has finite expectation, since we assumed $P\dot{\ell}^2 < \infty$.

So if $Var(Z_n) \to 0$ and $\mathbb{E}[Z_n] = 0$, then $Z_n \xrightarrow{p} 0$, so

$$\mathbb{G}_n\left(r_n\left(\ell_{\theta_0+\frac{h}{r_n}}(x) - \ell_{\theta_0}(x) - h^T\dot{\ell}_{\theta_0}(x)\right)\right) \xrightarrow{p} 0$$

for each h.

Now we need to show tightness. For this, we look at the localized process around θ_0. We know that $h^T\dot{\ell}_{\theta_0}$ is tight, as $\sup_{\|h\|_2\leq 1} h^T\dot{\ell}_{\theta_0} = \|\dot{\ell}_{\theta_0}\|_2$, which has a second moment. Therefore, we only study $r_n\left(\ell_{\theta_0+\frac{h}{r_n}} - \ell_{\theta_0}\right)$ as h varies. Let

$$\mathcal{L}_\delta := \left\{\frac{1}{\delta}(\ell_\theta - \ell_{\theta_0}) : \|\theta - \theta_0\| \leq \delta\right\}$$

$\mathcal{L}_{1/r_n}$ is equal to $\{r_n(\ell_{\theta_0+\frac{h}{r_n}} - \ell_{\theta_0}) : \|h\| \leq 1\}$. Note that, for δ small: $\frac{1}{\delta}|\ell_\theta(x) - \ell_{\theta_0}(x)| \leq \dot{\ell}(x)\frac{\|\theta-\theta_0\|}{\delta} \leq \dot{\ell}(x)$ if $\|\theta - \theta_0\| \leq \delta$ by local-Lipschitzness. Then considering bracketing numbers for $\mathcal{L}_\delta$, we get:

$$\begin{aligned}
N_{[]}(\mathcal{L}_\delta, L^2(\mathbb{P}_n), \epsilon) &= N_{[]}(\{\ell_\theta - \ell_{\theta_0}\}_{\|\theta-\theta_0\|\leq\delta}, L^2(P_n), \delta\epsilon) \\
&\lesssim N(\{\theta : \|\theta - \theta_0\| \leq \delta\}, \|\cdot\|, \frac{\delta\epsilon}{2\sqrt{\mathbb{P}_n\dot{\ell}^2}}) \\
&\leq \left(1 + \frac{2\delta}{\delta\epsilon}\sqrt{\mathbb{P}_n\dot{\ell}^2}\right)^d = \left(1 + \frac{2\sqrt{\mathbb{P}_n\dot{\ell}^2}}{\epsilon}\right)^d
\end{aligned}$$

where d is the dimensionality of the parameter space, by our previous results on the covering numbers of norm balls, and the relationship between bracketing numbers and covering numbers. Therefore

$$\begin{aligned}\mathbb{E}[\sup_{f\in\mathcal{L}_\delta}|\mathbb{G}_n(f)|] &\le C\mathbb{E}[\int\sqrt{\log N_{[]}(\mathcal{L}_\delta, L^2(\mathbb{P}_n),\epsilon)}d\epsilon] \\ &\le C\mathbb{E}[\int_0^{\sqrt{\mathbb{P}_n\dot{\ell}^2}}\sqrt{d\log\left(1+\frac{\sqrt{\mathbb{P}_n\dot{\ell}^2}}{\epsilon}\right)}d\epsilon] \\ &\le C\sqrt{d}\cdot\mathbb{E}[\dot{\ell}^2]\end{aligned}$$

As we assumed $\mathbb{E}[\dot{\ell}^2] < \infty$, this shows that the expectations are uniformly bounded, thus the process

$$\left\{\mathbb{G}_n(r_n(\ell_{\theta_0+\frac{h}{r_n}} - \ell_{\theta_0}) - h^T\dot{\ell}_{\theta_0}) : \|h\| \le 1\right\}$$

is tight.

As we have FIDI to 0, and thus the whole process must converge to zero. □

With this differentiability result, we can get asymptotic normality of M-estimators with nondifferentiable losses.

Theorem 4 (Van der Vaart 5.23). *Let $\ell_\theta(x)$ locally-Lipschitz (as in the Lemma) near θ_0. Assume that $\theta\mapsto\ell_\theta(x)$ is differentiable at θ_0 with $\mathbb{P}$-probability 1. Define $R(\theta) := \mathbb{E}_{\mathbb{P}}[\ell_\theta(x)]$. Assume $R(\theta)$ is twice differentiable at θ_0 with $\nabla^2R(\theta_0)\succ 0$, where $\theta_0 := \operatorname{argmin}_\theta R(\theta)$.*

Let $\widehat{\theta}\xrightarrow{p}\theta_0$. Assume

$$R_n(\widehat{\theta}_n) \le \inf_\theta R_n(\theta) + o_p(\frac{1}{n})$$

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -(\nabla^2R(\theta_0))^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n\dot{\ell}_{\theta_0}(X_i) + o_p(1)$$

Proof By the lemma, for any $h_n = O_p(1)$, we have

$$\mathbb{G}_n(\sqrt{n}(\ell_{\theta_0+\frac{h_n}{\sqrt{n}}} - \ell_{\theta_0}) - h_n^T\dot{\ell}_{\theta_0}) = o_p(1)$$

Now we have

$$\mathbb{G}_n(\sqrt{n}(\ell_{\theta_0+\frac{h_n}{\sqrt{n}}} - \ell_{\theta_0}) - h_n^T\dot{\ell}_{\theta_0}) = n(\mathbb{P}_n\ell_{\theta_0+\frac{h_n}{\sqrt{n}}} - \mathbb{P}_n\ell_{\theta_0}) + n(\mathbb{P}\ell_{\theta_0} - \mathbb{P}\ell_{\theta_0+\frac{h_n}{\sqrt{n}}}) - h_n^T\mathbb{G}_n\dot{\ell}_{\theta_0}$$

Now by definition of R, we get

$$n(\mathbb{P}\ell_{\theta_0} - \mathbb{P}\ell_{\theta_0+\frac{h_n}{\sqrt{n}}}) = n(R(\theta_0) - R(\theta_0 + \frac{h_n}{\sqrt{n}}))) = -\frac{1}{2}h_n^T\nabla^2R(\theta_0)h_n + o_p(1)$$

as $n\to\infty$, where the second step holds because of our assumptions on the differentiability of $R(\theta)$. Note $\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_p(1)$ from Rates of Convergence and the quadratic growth of R around θ_0: $R(\theta)\ge R(\theta_0) + c\|\theta-\theta_0\|^2$ near θ_0.

Let $\widehat{h}_n := \sqrt{n}(\widehat{\theta}_n - \theta_0)$, and $\tilde{h}_n := -\nabla^2R(\theta_0^{-1})\mathbb{G}_n\dot{\ell}_{\theta_0}$.

The goal is to show $\widehat{h}_n = \tilde{h}_n + o_p(1)$. Both $\widehat{h}_n$ and $\tilde{h}_n$ are $O_p(1)$. Now substitute these into the empirical process:

$$n(\mathbb{P}_n\ell_{\theta_0+\frac{\widehat{h}_n}{\sqrt{n}}} - \mathbb{P}_n\ell_{\theta_0}) = \frac{1}{2}\widehat{h}_n^T\nabla^2 R(\theta_0)\widehat{h}_n + \widehat{h}_n\mathbb{G}_n\dot{\ell}_{\theta_0} + o_p(1)$$

On the left side, by the choice of $\widehat{h}_n$ and since $\widehat{\theta}_n$ is the empirical minimizer, we get:

$$n(\mathbb{P}_n\ell_{\theta_0+\frac{\widehat{h}_n}{\sqrt{n}}} - P_n\ell_{\theta_0}) \leq n(\mathbb{P}_n\ell_{\theta_0+\frac{\tilde{h}_n}{\sqrt{n}}} - \mathbb{P}_n\ell\theta_0) = \frac{-1}{2}(\mathbb{G}_n\dot{\ell}_{\theta_0})^T\nabla^2 R(\theta_0)^{-1}(\mathbb{G}_n(\ell_{\theta_0})) + o_p(1)$$

where the second step uses our Taylor approximation lemma and the definition of $\tilde{h}_n$. Substituting:

$$\widehat{h}_n^T\nabla^2 R(\theta_0)\widehat{h}_n + \widehat{h}_n^T\mathbb{G}_n\dot{\ell}_{\theta_0} \leq \frac{-1}{2}(\mathbb{G}_n\dot{\ell}_{\theta_0})^T\nabla^2 R(\theta_0)^{-1}(\mathbb{G}_n\dot{\ell}_{\theta_0}) + o_p(1)$$

Completing the square:

$$\frac{1}{2}\left(\widehat{h}_n + \nabla^2 R(\theta)^{-1}\mathbb{G}_n\dot{\ell}_\theta\right)^T \nabla^2 R(\theta_0)\cdot\left(\widehat{h}_n + \nabla^2 R(\theta_0)^{-1}\mathbb{G}_n\dot{\ell}_\theta\right) = o_p(1)$$

□