Notes for Week 2/1/18 - 2/8/18

Alden Green

February 7, 2019

For fixed integers $n_1+n_2 = n$, let $\mathbf{X} = \{x_1, \dots, x_{n_1}\} \subset \mathbb{R}^d$ and $\mathbf{Y} = \{y_1, \dots, y_{n_2}\}$ be sampled i.i.d from distributions $\mathbb{P}$ and $\mathbb{Q}$ with density functions p and q, respectively, both with support on $D \subset \mathbb{R}^d$. Our statistical problem is testing the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ vs. the alternative $H_1 : \mathbb{P} \neq \mathbb{Q}$, where our knowledge of $\mathbb{P}$ and $\mathbb{Q}$ come from the samples $\mathbf{X}$ and $\mathbf{Y}$.

Recall the *Laplacian smooth* and *total variation smooth* test statistics

$$T_1(\boldsymbol{\ell}; G_{n,r}) = \sup_{\theta : \|\mathbf{B}\theta\|_1 \leq C_{n,r}} \left|\boldsymbol{\ell}^T \theta\right|$$
$$T_2(\boldsymbol{\ell}; G_{n,r}) = \sup_{\theta : \|\mathbf{B}\theta\|_2 \leq C_{n,r}} \left|\boldsymbol{\ell}^T \theta\right|$$

where $\mathbf{B}$ is the incidence matrix of the r-neighborhood graph and $C_{n,r} = \frac{\sigma_k}{n^2 r_n^{d+2}}$.

Theorem 1 is the type of theorem we are looking for.

Theorem 1. *Under assumptions,*

$$\sqrt{n} T_2(\boldsymbol{\ell}; G_{n,r}) \rightsquigarrow ???$$

under the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$.

The rest of this document details the strategy for proving this convergence.

1 Quantization

To ease proofs, we will assume $\mathcal{D} = [0,1]^d$.

Construct $G_{lat} = (V_{lat}, E_{lat})$ a lattice graph with equal side lengths in $[0,1]^d$, where

$$V_{lat} = P_{lat}(N) := \left\{ \left(\frac{i_1}{N} - \frac{1}{2N}, \cdots, \frac{i_d}{N} - \frac{1}{2N} \right) : i_1, \dots, i_d \in \{1, \dots, N\} \right\}$$

$(z, z') \in E_{lat}$ if and only if $\|z - z'\| \leq \frac{1}{N}$

where z and $z' \in P_{lat}(N)$.

Denoting $I = P_{lat}$, we define

$$P_I(x) = \text{argmin}\left\{\|x - z'\|_\infty, z' \in P_{lat}(N)\right\}$$

Then, let $C(z) = \left\{x \in [0,1]^d : z = P_I(x)\right\}$ be the collection of cells associated with the mesh $P_{lat}(N)$, noting that $\{C(z) : z \in P_{lat}(N)\}$ defines a partition over $[0,1]^d$.

Introduce $\bar{f}$.

2 High-Level Proof Strategy for Theorem 1

(i) *Quantization:* Consider the function class

$$\mathcal{W}_n = \left\{\frac{\overline{f}}{\|Bf\|_2} : f \in \mathcal{W}^{1,2}(\mathcal{D}, \rho^2)\right\}.$$

Under assumptions,

$$\sup_{\theta:\|B\theta\|_2 \le 1} \left|\ell^T\theta\right| - \sup_{\widetilde{f} \in \mathcal{W}_n} \left|\mathbb{P}_n\widetilde{f} - \mathbb{Q}_n\widetilde{f}\right| = o_{\mathbb{P}}(n^{-1/2})$$

(ii) *Donsker-convergence of empirical process:* Write

$$\widetilde{f} = \frac{\overline{f}}{\|Bf\|_2}.$$

Under assumptions,

$$\left\{\mathbb{G}_{\mathbb{P}_n}\widetilde{f} : f \in \mathcal{W}^{1,2}(\mathcal{D}, \rho^2)\right\} \rightsquigarrow G_{\mathbb{P}}$$

and

$$\left\{\mathbb{G}_{\mathbb{Q}_n}\widetilde{f} : f \in \mathcal{W}^{1,2}(\mathcal{D}, \rho^2)\right\} \rightsquigarrow G_{\mathbb{Q}},$$

where $G_{\mathbb{P}}$ is a tight Gaussian process with support on $\mathcal{W}^{1,2}(\mathcal{D}, \rho^2)$ and for measures P_n and P, $\mathbb{G}_{P_n,P} = \sqrt{n}(P_n - P)$ and we suppress notational dependence on P when obvious from context.

(iii) *Continuous mapping:* Under the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$, for any $\widetilde{f} \in \mathcal{W}_n$,

$$\sqrt{n}(\mathbb{P}_n\widetilde{f} - \mathbb{Q}_n\widetilde{f}) = (\mathbb{G}_{\mathbb{P}_n} - \mathbb{G}_{\mathbb{Q}_n})\widetilde{f}$$

Thus, by i), the independence of $\mathbb{P}_n$ and $\mathbb{Q}_n$, and the continuous mapping theorem,

$$\left\{\sqrt{n}(\mathbb{P}_n\widetilde{f} - \mathbb{Q}_n\widetilde{f}), f \in \mathcal{W}^{1,2}(\mathcal{D}, \rho^2)\right\} \rightsquigarrow G_{\mathbb{P}} - G'_{\mathbb{P}}$$

where $G_{\mathbb{P}}$ and $G'_{\mathbb{P}}$ are i.i.d. Gaussian processes.

By the continuous mapping theorem again

$$\sup_{f\in\mathcal{W}^{1,2}(\mathcal{D},\rho^2)}\left|\sqrt{n}(\mathbb{P}_n\widetilde{f}-\mathbb{Q}_n\widetilde{f})\right| \rightsquigarrow \sup_{f\in\mathcal{W}^{1,2}(\mathcal{D},\rho^2)}|(G_{\mathbb{P}}-G'_{\mathbb{P}})f|$$

3 Proof Strategy for *Quantization*

4 Proof Strategy for ii)

We want to exhibit a function class

$$\overline{\mathcal{W}}_n = \left\{f_{n,t} : t \in \mathcal{W}^{1,2}(\mathcal{D},\rho^2)\right\}$$

where $f_{n,t} : \mathcal{D} \to \mathbb{R}$ such that:

(a) There exists a bijection $\phi : \mathcal{W}_n \to \overline{\mathcal{W}}_n$ with

$$\sqrt{n}\mathbb{P}_n(f_n - \phi(f_n)) \rightsquigarrow 0.$$

(b) $\overline{\mathcal{W}}_n$ is totally bounded.

(c) For every sequence $\delta_n \downarrow 0$,

$$\sup_{\|s-t\|_{1,2,\rho^2}\le\delta_n}\mathbb{P}(f_{n,s}-f_{n,t})^2 \to 0$$

(d) There exists a sequence of envelope functions F_n satisfying the Lindeberg condition

$$\mathbb{P}F_n^2 = \mathcal{O}(1)$$
$$\mathbb{P}F_n^2\mathbf{1}F_n > \epsilon\sqrt{n} \to 0 \qquad \text{(for any } \epsilon > 0)$$

(e) The bracketing integral

$$\int_0^\delta \sqrt{\log N_{[]}(\epsilon,\overline{W}_n,L_2(\mathbb{P}))}$$

converges to 0 for any $\delta_n \downarrow 0$.

Theorem 2. *Let $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ be a class of measurable functions indexed by a totally bounded semimetric space (T,ρ) satisfying*

$$\sup_{\rho(s,t)<\delta_n} P(f_{n,s}-f_{n,t})^2 \to 0, \quad \textit{every } \delta_n \downarrow 0$$

and with envelope function F_n satisfying the Lindeberg condition

$$PF_n^2 = \mathcal{O}(1)$$
$$PF_n^2 \mathbf{1}\left\{F_n > \epsilon\sqrt{n}\right\} \to 0, \quad \textit{for every } \epsilon > 0.$$

If $J_{[]}(\delta_n, \mathcal{F}_n, L_2(P)) \to 0$ for every $\delta_n \downarrow 0$, then the sequence $\{\mathbb{G}_n f_{n,t}, t \in T\}$ converges in distribution to a tight Gaussian process, provided the sequence of covariance functions

$$Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$$

converges pointwise on $T \times T$.