Notes for Week 2/23/19 - 2/29/19

Alden Green

May 15, 2019

Consider distributions $\mathbb{P}$ and $\mathbb{Q}$ supported on $\mathcal{D} \subset \mathbb{R}^d$ which are absolutely continuous with density functions f and g, respectively. For fixed $n \geq 0$, let $Z = (z_1, \ldots, z_n)$, where for $i = 1, \ldots, n$, $z_i \sim \frac{\mathbb{P}+\mathbb{Q}}{2}$ are independent. Given Z, for $i = 1, ..., n$ let

$$\ell_i = \begin{cases} 1 \text{ with probability } \frac{f(z_i)}{f(z_i)+g(z_i)} \\ -1 \text{ with probability } \frac{g(z_i)}{f(z_i)+g(z_i)} \end{cases}$$

be conditionally independent labels, and write

$$1_X = \begin{cases} 1, & l_i = 1 \\ 0, & \text{otherwise} \end{cases} \quad 1_Y = \begin{cases} 1, & l_i = -1 \\ 0, & \text{otherwise.} \end{cases}$$

We will write $X = \{x_1, \ldots, x_{N_X}\} := \{z_i : \ell_i = 1\}$ and similarly $Y = \{y_1, \ldots, y_{N_Y}\} := \{z_i : \ell_i = -1\}$, where N_X and N_Y are of course random but $N_X + N_Y = n$.

Our statistical goal is hypothesis testing: that is, we wish to construct a test function ϕ which differentiates between

$$\mathbb{H}_0 : f = g \text{ and } \mathbb{H}_1 : f \neq g.$$

For a given function class $\mathcal{H}$, some $\epsilon > 0$, and test function ϕ a Borel measurable function of the data with range $\{0, 1\}$, we evaluate the quality of the test using *worst-case risk*

$$R_\epsilon^{(t)}(\phi; \mathcal{H}) = \sup_{f \in \mathcal{H}} \mathbb{E}_{f,f}^{(t)}(\phi) + \sup_{\substack{f,g \in \mathcal{H} \\ \delta(f,g) \geq \epsilon}} \mathbb{E}_{f,g}^{(t)}(1 - \phi)$$

where

$$\delta^2(f, g) = \int_{\mathcal{D}} (f - g)^2 dx.$$

1 Laplacian smooth test statistic

For $r \geq 0$, define the *r-graph* $G_r = (V, E_r)$ to have vertex set $V = \{1, \ldots, t\}$ and edge set E_r which contains the pair (i, j) if and only if $\|z_i - z_j\|_2 \leq r$. Let D_r denote the incidence matrix of G_r.

For a critical radius $C_{n,r}$ to be determined later, define the *r-Laplacian Smooth* test statistic to be

$$T_{LS} = \sup_{\theta : \|D_r \theta\|_2 \leq C_{n,r}} \langle \theta, \frac{1_X}{N_X} - \frac{1_Y}{N_Y} \rangle$$

We would like to relate the graph G_r to a graph with a more easily accessible spectrum. For $\kappa = n^{1/d}$, consider the *grid graph*

$$G_{grid} = (V_{grid}, E_{grid}), \quad V_{grid} = \left\{ \frac{k}{\kappa} : k \in [\kappa]^d \right\}, \quad E_{grid} = \left\{ (k, k') : k, k' \in V_{grid}, \|k - k'\|_1 = \frac{1}{\kappa^d} \right\}$$

with associated incidence matrix D_{grid}.

Lemma 1 (Spectral similarity of r-graph to grid)**.** *Fix $r \geq 2C \left(\frac{\log n}{n}\right)^{1/d} + \left(\frac{1}{n}\right)^{1/d}$, where $C > 0$ is a universal constant, and let*

$$\sigma_{r,n} = d^{d+1/2} n^{2+1/d} \left(2C \left(\frac{\log n}{n} \right)^{1/d} + r \right)^{2d+1}$$

For any $\theta \in \mathbb{R}^n$, there exists a permutation $\Pi : \mathbb{R}^d \to \mathbb{R}^d$ such that the following relations hold:

$$\frac{\|D_{G_r}\theta\|_2}{\sigma_{r,n}} \leq \|D_{grid}(\Pi\theta)\|_2 \leq \|D_{G_r}\theta\|_2 \tag{1}$$

with probability at least $1 - n^{-\alpha}$ where $\alpha = c_1(\log n)^{1/2}$ for some constant $c_1 > 0$.

Lemma 1 relies heavily on theory regarding optimal transportation matchings between two sets of discrete points, this case Z and V_{grid}.

Lemma 2. *There exists a bijection $T : Z \to V_{\text{grid}}$ such that*

$$\max_{i \in [n]} \|T(z_i) - z_i\|_2 \leq C \left(\frac{\log n}{n} \right)^{1/d}$$

with probability at least $1 - n^{-\alpha}$, where $\alpha = c_1(\log n)^{1/2}$ and $c_1, C > 0$ are universal constants.

The upper bound of (1) follows easily.

Upper bound of (1). Assume there exists T such that Lemma 2 holds.

Let $k, k' \in [\kappa]^d$ satisfy $\frac{k}{\kappa}$ $\frac{k'}{\kappa}$ in the grid graph. There exist z_i and z_j such that $T(z_i) = \frac{k}{\kappa}$ and $T(z_j) = \frac{k'}{\kappa}$. By the triangle inequality,

$$\begin{aligned} \|z_i - z_j\|_2 &\leq \|T(z_i) - z_i\|_2 + \|T(z_i) - T(z_j)\|_2 + \|T(z_j) - z_j\|_2 \\ &\leq 2C \left(\frac{\log n}{n} \right)^{1/d} + \frac{1}{n^{1/d}} \end{aligned}$$

and so by our choice of r, $i \sim j$ in G_r. □

To show the lower bound of (1), we will make use of a technique from spectral graph theory known as Poincare's inequality.

Poincare inequality Let G and $\widetilde{G}$ be undirected, unweighted graphs over vertex set V, with edge sets E_G and $E_{\widetilde{G}}$, respectively. Let $\widetilde{\mathcal{P}}$ be the space of all paths over $E_{\widetilde{G}}$; that is, $\mathcal{P}$ consists of $\widetilde{P} \in \widetilde{\mathcal{P}}$ with

$$\widetilde{P} = (\widetilde{e}_1, \ldots, \widetilde{e}_m) \qquad (\widetilde{e}_i \in E_{\widetilde{G}})$$

for some integer $m \geq 1$.

Lemma 3 (Poincare inequality)**.** *Define a mapping $\gamma : E_G \to \mathcal{P}$ where for each $e = (\ell, \ell')$ in E_G*

$$\gamma(e) = ((\ell, u), \dots, (v, \ell'))$$

meaning e is mapped to a path which begins at ℓ and ends at ℓ'. Then

$$G \preceq \widetilde{G} \cdot \max_{e \in E_G} |\gamma(e)| \cdot b_\gamma$$

where b_γ is a bottleneck parameter given by

$$b_\gamma = \max_{\widetilde{e} \in E_{\widetilde{G}}} |\{e \in E : \widetilde{e} \in \gamma(e)\}|$$

Lemma 2 will allow us to construct such a mapping γ from E_r to E_{grid} and appropriately control parameters $\max_{e \in E_G} |\gamma(e)|$ and b_γ.

Lemma 4. *There exists a mapping $\gamma : E_r \to \mathcal{P}_{\text{grid}}$, the set of paths over G_{grid}, such that the following quantities are bounded:*

(i) Maximum path length.

$$\max_{e \in E_G} |\gamma(e)| \leq n^{1/d}\sqrt{d}\left(2C\left(\frac{\log n}{n}\right)^{1/d} + r\right)$$

(ii) Bottleneck.

$$b_\gamma \leq \left(n^{1/d}\sqrt{d}\left(2C\left(\frac{\log n}{n}\right)^{1/d} + r\right)\right)^{2d}$$

with probability at least $1 - n^{-\alpha}$ where $\alpha = c_1(\log n)^{1/2}$ and $C, c_1 > 0$ are universal constants.

Proof. Assume $i \sim j$ in the graph G_r. By a similar set of steps to the above, we have

$$\|T(z_i) - T(z_j)\|_2 \leq 2C\left(\frac{\log t}{t}\right)^{1/d} + r$$

As a result, using the simple relation $\|x\|_1 \leq \sqrt{d}\,\|x\|_2$ for any $x \in \mathbb{R}^d$, we have

$$\|T(z_i) - T(z_j)\|_1 \leq \sqrt{d}(2C\left(\frac{\log t}{t}\right)^{1/d} + r)$$

Since each edge in the grid graph is of length $n^{1/d}$, it is easy to see that there exists a path between $T(z_i)$ and $T(z_j)$ in G_{grid}, $P(T(Z_i) \to T(Z_j))$ with no more than

$$\frac{\sqrt{d}(2C\left(\frac{\log t}{t}\right)^{1/d} + r)}{t^{1/d}}$$

edges. The bound follows by Lemma **??**. ∎

2 Additional Theory and Proofs

2.1 Proof of Lemma 3

Lemma 5 (Poincare inequality for path graphs.)**.** *Fix $m \geq 0$. For vertices $V = \{1, \dots, m\}$ define the path $P(1 \to m) = ((1,2), (2,3), \dots, (m-1, m))$ and $G_{(1,m)}$ to be the graph consisting only of an edge between 1 and m. Then,*

$$(m-1) \cdot P(1 \to m) \succeq G_{(1,m)}$$

Proof of Lemma 3 Let $G_e = (V, \{e\})$ and $P_e = (V, \{\widetilde{e} : \widetilde{e} \in \gamma(e)\})$ be the graphs associated with e and $\gamma(e)$, respectively. By Lemma 5, we have

$$G_e \preceq |P_e|\, P_e$$

Summing over all $e \in E_G$, we obtain

$$\begin{aligned} G &\preceq \sum_{e \in E_G} |P_e|\, P_e \\ &\preceq \max_{e \in E_G} |\gamma(e)| \sum_{e \in E_G} P_e \\ &\preceq \max_{e \in E_G} |\gamma(e)|\, b_\gamma \cdot \widetilde{G} \end{aligned}$$

Decompose $\frac{1_X}{N_X} - \frac{1_Y}{N_Y} := \theta^\star + w$, where

$$(\theta^\star)_i := \frac{f(x) - g(x)}{f(x) + g(x)}$$

The upper bound in Lemma 1 allows us the following upper bound on the empirical process

$$\sup_{\theta : \|D_r \theta\|_2 \le C_{n,r}} \langle \theta, w \rangle \le \sup_{\theta : \|D_{grid} \theta\|_2 \le C_{n,r}} \langle \theta, w \rangle = C_{n,r} w^T L_{grid}^\dagger w$$

whereas the lower bound helps us with the approximation error term,

$$\sup_{\widetilde{\theta} : \|D_r \theta\|_2 \le C_{n,r}} \langle \widetilde{\theta}, \theta^\star \rangle \ge \sup_{\theta : \|D_{grid} \theta\|_2 \le C_{n,r}/\ell(n,r)} \langle \theta, \theta^\star \rangle \ge \frac{C_{n,r}}{\ell(n,r)} \theta^\star L_{grid}^\dagger \theta^\star$$