Notes for Week 2/16/19 - 2/22/19

Alden Green

February 22, 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 $t \geq 0$, Let $\mathbf{Z} = (z_1, \ldots, z_t)$, where for $i = 1, \ldots, t$, $z_i \sim \frac{\mathbb{P}+\mathbb{Q}}{2}$ are independent. Given $\mathbf{Z}$, for $i = 1, ..., t$ 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 conditional 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 $\mathbf{X} = \{x_1, \ldots, x_{N_X}\} := \{z_i : \ell_i = 1\}$ and similarly $\mathbf{Y} = \{y_1, \ldots, y_{N_Y}\} := \{y_i : \ell_i = -1\}$, where N_X and N_Y are of course random.

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 Total variation test

As in [2], define the *K-NN graph* $G_K = (V, E_K)$ to have vertex set $V = \{1, \ldots, t\}$ and edge set E_K which contains the pair (i, j) if and only if x_i is among the

K-nearest neighbors (with respect to Euclidean distance) of x_j, or vice versa. Let D_K denote the incidence matrix of G_K.

Define the *kNN-total variation* test statistic to be

$$T_{TV} = \sup_{\substack{\theta \in \mathbb{R}^t: \\ \mathcal{T}(C_{n,k})}} \left(\frac{1}{N_X} \sum_{i:(1_X)_i = 1} \theta_i - \frac{1}{N_Y} \sum_{j:(1_Y)_j=1} \theta_j \right) \tag{1}$$

where $\mathcal{T}(C_{n,k}) = \left\{ \theta : \|D_{G_K}\theta\|_1 \leq C_{n,k}, \|\theta\|_2 \leq C'_{n,k} \right\}$. Hereafter, take $\mathcal{D} = [0,1]^d$, and consider

$$\mathcal{H}_{lip}(L) = \left\{ f : [0,1]^d \to \mathbb{R}^+ : \int_{\mathcal{D}} f = 1, f \text{ } L\text{-piecewise Lipschitz, bounded above and below} \right\}$$

Definition 1.1 (Piecewise Lipschitz)**.** A function f is *L-piecewise lipschitz* over $[0,1]^d$ if there exists a set $\mathcal{S} \subset [0,1]^d$ such that

(a) $\nu(\mathcal{S}) = 0$

(b) There exist $C_{\mathcal{S}}, \epsilon_0$ such that $\mu\Big(\big(\mathcal{S}_\epsilon \cup (\partial \mathcal{D})_\epsilon \big) \cap [0,1]^d \Big) \leq C_{\mathcal{S}}\epsilon$ for all $0 < \epsilon \leq \epsilon_0$.

(c) For any z, z' in the same connected component of $[0,1]^d \setminus \big(\mathcal{S}_\epsilon \cup (\partial\mathcal{D})_\epsilon\big)$,

$$|g(z) - g(z')|_2 \leq L \, \|z - z'\|_2$$

Definition 1.2 (Bounded above and below)**.** A function $f : \mathcal{D} \to \mathbb{R}$ is *bounded above and below* if there exists $p_{\min}, p_{\max}$ such that

$$0 < p_{\min} < f(x) < p_{\max} < \infty \qquad (\forall x \in \mathcal{D})$$

Conjecture 1. *For τ =??? and $K \asymp \log^{1+2r}(n)$ for some $r \geq 0$, the test $\phi_{TV} = \{T_{TV} \geq \tau\}$ has worst-case risk*

$$R_\epsilon^{(t)}(\mathcal{H}_{lip}(L)) \leq 1/2$$

whenever $\epsilon \geq c_2 \log^\alpha m m^{-1/d}$ where $\alpha = 3r + 5/2 + (2r+1)/d$ and c_1 and c_2 are constants which depend only on (d, L).

Proof. Write

$$\left(\frac{1}{N_X} \sum_{i:(1_X)_i = 1} \theta_i - \frac{1}{N_Y} \sum_{j:(1_Y)_j=1} \theta_j \right) = \langle \theta, \frac{1_X}{N_X} - \frac{1_Y}{N_Y} \rangle$$

and let

$$\widehat{\theta} \in \operatorname*{argmax}_{\theta \in \mathbb{R}^t} \left\{ \langle \theta, \frac{1_X}{N_X} - \frac{1_Y}{N_Y} \rangle : \theta \in \mathcal{T}(C_{n,K}) \right\}$$

satisfy $T_{TV} = \langle \widehat{\theta}, \frac{1_X}{N_X} - \frac{1_Y}{N_Y} \rangle$.

Random denominators We will first account for the dependence on random denominators N_X and N_Y. For arbitrary $\theta \in \mathcal{T}(C_{n,K})$

$$\langle\theta, \frac{1_X}{N_X} - \frac{1_Y}{N_Y}\rangle = \frac{2}{t}\langle\theta, 1_X - 1_Y\rangle + 2\left(\frac{N_X - t/2}{(N_X)t}\right)\langle\theta, 1_X\rangle + 2\left(\frac{t/2 - N_Y}{(N_Y)t}\right)\langle\theta, 1_Y\rangle$$

A simple application of Holder's inequality, and standard concentration results for binomial random variables, will show the latter two terms to be $O_P\left(\frac{C'_{n,k}}{t}\right)$.

We turn to analyzing the first term. For $i = 1, \ldots, t$, introduce $\theta^\star$ defined by

$$(\theta^\star)_i = \left(\frac{f(z_i) - g(z_i)}{f(z_i) + g(z_i)}\right)$$

Letting $w = (1_X - 1_Y) - \theta^\star$, note that w is a vector of i.i.d bounded random variables with $\mathbb{E}(w) = 0$.

Type I error. Under the case $f = g$, we have $\theta^\star = 0$, and therefore

$$\langle\theta, 1_X - 1_Y\rangle = \langle\theta, w\rangle$$

Proceed using the empirical process bound of Lemma 1. $\square$

2 Bounding the empirical process

We collect here the bound on the empirical process from [2].

Lemma 1. *Let w be a vector of mean zero independent random variables with $w_\infty \leq 1$. Then, for any $\delta > 0$ such that $K \geq 3\log(n/\delta)$*

$$\sup_{\theta\in\mathcal{T}(C_{n,K})}\langle\theta, w\rangle \leq 2C_{n,K} + \sqrt{(1 + \frac{C(p_{\max},\delta)}{K})Kp_{\max}}\cdot\left(2\sqrt{2\log(e/\delta)}C'_{n,K} + 2C(d)\sqrt{\log(en/\delta)}C_{n,K}\right)$$

with probability at least $1 - 4\delta$.

Proof. Set

$$\kappa = \lceil\frac{3\sqrt{d}p_{\max}^{1/d}t^{1/d}}{2K^{1/d}}\rceil$$

and let

$$I = \left\{\left(\frac{k-1}{\kappa}, \frac{k}{\kappa}\right] : k \in [\kappa]^d\right\},\ M_{k,\kappa} = \#\left\{i \in [t] : z_i \in \left(\frac{k-1}{\kappa}, \frac{k}{\kappa}\right]\right\}$$

be a partition of $\mathcal{D}$ into cells, and the count in each cell, respectively. From [2] we have that with probability at least

$$1 - t\exp(-K/3) \geq 1 - \delta$$

the following statement holds:

$$\sup_{\theta \in \mathcal{T}(C_{n,K})} \leq 2\left\|w\right\|_\infty \left\|D_{G_K}\theta\right\|_1 + \max_k \sqrt{M_{k,\kappa}} \left(\left\|\Pi \widetilde{w}\right\|_2 \left\|\theta\right\|_2 + \left\|(D^\dagger)^T \widetilde{w}\right\|_\infty \left\|D_{G_K}\theta\right\|_1\right)$$

where Π is the projection onto the span of 1_{κ^d}, D is the incidence matrix of the grid graph κ^d, and $\widetilde{w}$ is defined by

$$(\widetilde{w})_k = \left[\max_{k \in [\kappa]^d} M_{k,\kappa}\right]^{-1/2} \sum_{l : z_l \in (\frac{k-1}{\kappa}, \frac{k}{\kappa}]} w_l$$

Take care of random denominator again.

[1] derive the following bounds, which hold with probability at least $1 - 2\delta$:

$$\left\|(D^\dagger)^T w\right\|_\infty \leq 2C(d)\sqrt{\log(en/\delta)}, \quad \left\|\Pi w\right\|_2 \leq 2\sqrt{2\log(e/\delta)}$$

Then, a simple concentration inequality for binomial random variables, along with a union bound, gives

$$\max_{k \in [\kappa]^d} \sqrt{M_{k,\kappa}} \leq \sqrt{(1 + \frac{C(p_{\max}, \delta)}{K}) \frac{t p_{\max}}{\kappa^d}}$$

with probability at least $1 - \delta$. The statement follows from our choice of κ. $\square$

3 Laplacian smooth test–Attempt 1

Let

$$\widehat{\theta}_{LS} = \sup_{\theta \in \mathcal{S}(C_{n,K}, C'_{n,K})} \langle \theta, \frac{1_X}{N_X} - \frac{1_Y}{N_Y} \rangle$$

where $\mathcal{S}(C_{n,K}, C'_{n,K}) = \left\{\theta \in \mathbb{R}^n : \left\|D_{G_K\theta}\right\|_2 \leq C_{n,K}, \left\|\theta\right\|_2 \leq C'_{n,K}\right\}$.

Random Denominator.

Lemma 2. *Let $\left\|\theta\right\|_2 \leq c$ and $\delta > 0$. Then,*

$$\left|\langle \theta, 1_X \rangle \left(\frac{1}{N_X} - \frac{2}{t}\right)\right| \vee \left|\langle \theta, 1_Y \rangle \left(\frac{1}{N_Y} - \frac{2}{t}\right)\right| \leq \frac{c\sqrt{\log(2/\delta)}}{t(1 - \log(2/\delta)/\sqrt{t})}$$

with probability at least $1 - \delta$.

Proof. By the Cauchy-Schwarz inequality,

$$\left|\langle \theta, 1_X \rangle\right| \leq \left\|\theta\right\|_2 \left\|1_X\right\|_2 \leq c\sqrt{N_X}.$$

Rearranging $1/N_X - 2/t$, we obtain

$$\left|\langle\theta, 1_X\rangle\left(\frac{1}{N_X} - \frac{2}{t}\right)\right| \leq 2\frac{c\sqrt{N_X}\left|N_X - \frac{2}{t}\right|}{tN_X} = \frac{c\left|N_X - \frac{2}{t}\right|}{t\sqrt{N_X}} \tag{2}$$

$N_X \sim Bin(t, 1/2)$, and so application of Hoeffding's inequality gives

$$\left|N_X - \frac{t}{2}\right| \leq \sqrt{t}\sqrt{\log(2/\delta)} \tag{3}$$

with probability at least $1-\delta$. Plugging this in to (2) yields the desired bound with respect to $1_X, N_X$. However, if (3) holds for N_X it holds for N_Y as well. All other steps hold for 1_Y, and therefore the desired bound holds with respect to $1_Y, N_Y$ as well. □

4 Laplacian smooth test: Attempt 2

Let

$$T_{LS} = \sup_{\theta \in \mathcal{S}(C_{n,K}, C'_{n,K})} \langle\theta, \frac{1_X}{N_X} - \frac{1_Y}{N_Y}\rangle$$

where $\mathcal{S}(C_{n,K}, C'_{n,K}) = \{\theta \in \mathbb{R}^n : \|D_{G_K}\theta\|_2 \leq C_{n,K}\}$. We can find a closed-form solution to this problem,

$$T_{LS} = C_{n,k} a^T L^\dagger a$$

where $a = (\frac{1_X}{N_X} - \frac{1_Y}{N_Y})$.

Type I error. To begin, we write $w = (1_X - 1_Y)\frac{2}{t}$, and rewrite

$$T_{LS} = C_{n,k} w^T L^\dagger w + C_{n,k}(w+\ell)^T L^\dagger(\ell - w)$$

We turn our attention to the second term, which we wish to show contributes negligibly to the overall sum. We have

$$\begin{aligned}(w+\ell)^T L^\dagger(\ell - w) &= (D(w+\ell))^T(D(\ell - w)) \\ &\leq \|D(w+\ell)\|_2 \|D(w-\ell)\|_2 \\ &\leq K\|w+\ell\|_2 \|w-\ell\|_2\end{aligned}$$

Then, based on Lemma make a 'random denominators' Lemma, with probability at least $1-\delta$,

$$\|w+\ell\|_2 \leq \frac{1}{\sqrt{t}}, \quad \|w-\ell\|_2 \leq 2\frac{\log(2/\delta)}{t\left(1 - \frac{\log(2/\delta)}{\sqrt{t}}\right)}$$

and as a result

$$(w+\ell)^T L^\dagger(\ell - w) \leq K \frac{\log(2/\delta)}{t^{3/2}\left(1 - \frac{\log(2/\delta)}{\sqrt{t}}\right)}$$

Now, we have that the entries of w are i.i.d random variables with mean 0 and absolute value of $2/t$.

5 Laplacian Smooth Test-Attempt 3

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_{G_r} denote the incidence matrix of G_r.

Define the *r-Laplacian smooth* test statistic to be

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

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

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

with associated incidence matrix D_{grid}.

Lemma 3 (Spectral similarity of r-graph to grid)**.** *Fix $r \geq 2\left(\frac{\log t}{t}\right)^{1/d} + (\frac{1}{t})^{1/d}$, and Let $\ell(t) = \sqrt{d}rt^{1/d} + 2\sqrt{d}(\log t)^{1/d}$. For any $\theta \in \mathbb{R}^t$, the following relations hold:*

$$\frac{\|D_{G_r}\theta\|_2}{\ell(t)} \leq \|D_{grid}\theta\|_2 \leq \|D_{G_r}\theta\|_2 \tag{4}$$

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

Proof. We begin by mapping the data $\mathbf{Z}$ to the grid points $[\kappa]^d$ in such a way that as little mass as possible is disturbed:

Lemma 4. *There exists a bijective mapping $T : \mathbf{Z} \to [\kappa]^d$ for $\kappa = t^{1/d}$ such that*

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

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

Hereafter, we assume there exists T such that Lemma 4 holds.

We first prove the second bound in (4). Consider grid points k k' connected in the grid graph. Then, there exist z_i and z_j such that $T(z_i) = k$ and $T(z_j) = k'$. By the triangle inequality,

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

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

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

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

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

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

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

Lemma 5 (Graph ordering)**.** *Fix $m \ge 0$. For vertices $V = \{1, \ldots, m\}$, we have*

1. $\frac{1}{m-1}P(1 \to m) \succeq G_{1,m}$
2. *If $A \succeq B$ and $C \succeq D$, then $A + B \succeq C + D$.*

□

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 3 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,

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

REFERENCES

[1] Jan-Christian Htter and Philippe Rigollet. Optimal rates for total variation denoising. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1115–1146, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[2] Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela M Witten. Adaptive non-parametric regression with the k-nn fused lasso. *arXiv preprint arXiv:1807.11641*, 2018.