Notes for Week of 12/24/18 - 12/31/18

Alden Green

January 28, 2019

1 SETUP

Data model. We are given two unknown distributions, $\mathbb{P}$ and $\mathbb{Q}$, both supported on $D \subset \mathbb{R}^d$, with continuous density functions p and q, respectively. We have the capacity to sample from either distribution. Our goal is to test the nonparametric hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ vs. the alternative $H_1 : \mathbb{P} \neq \mathbb{Q}$.

Under the *binomial data model*, we sample data $\{z_1, \ldots, z_n\}$ as follows: for $i = 1, \ldots, n$, we draw an independent Rademacher label $\ell_i \in \{1, -1\}$, $\Pr(\ell_i = 1) = \Pr(\ell_i = -1) = 1/2$. Then, if $\ell_i = 1$ we sample $z_i \sim \mathbb{P}$, whereas if $\ell_i = -1$ we sample $z_i \sim \mathbb{Q}$. Define $\mathbf{1}_X$ to be the length n indicator vector for $\ell_i = 1$

$$\mathbf{1}_X(i) = \begin{cases} 1, \ell_i = 1 \\ 0 \text{ otherwise} \end{cases}$$

and similarly for $\mathbf{1}_Y$

$$\mathbf{1}_Y(i) = \begin{cases} 1, \ell_i = -1 \\ 0 \text{ otherwise} \end{cases}$$

Denote the number of positive labels $N = \sum_{i=1}^n \mathbf{1}_X(i)$, for M the number of negative labels we have $n = N + M$. Then the *normalized label vector* a is given by $a = \frac{1}{N}\mathbf{1}_X - \frac{1}{M}\mathbf{1}_Y$.

Graph. Heuristically, our graph-based test will attempt to identify whether samples from the same distribution are on average more similar than samples from different distributions. As a result, we introduce K, a **kernel function** which measures similarity, and make the following assumptions on K

- $K : [0, \infty) \to [0, \infty)$ is non-increasing.
- The integral $\int_0^\infty K(r) r^d dr$ is finite.

and define

$$K_\epsilon(z) = \frac{1}{\epsilon^d} K\left(\frac{z}{\epsilon}\right)$$

Given a sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$, form the **$\epsilon$-radius neighborhood graph** $G_{n,r} = (V_n, E_n)$ with $V_n = \{z_1, \ldots, z_n\}$ and $E_n = \{(i,j) : K_\epsilon(\|z_i - z_j\|) > 0\}$. Let A be the adjacency matrix associated with $G_{n,r}$. Take $L_n = D - A$ to be the (unnormalized) **Laplacian matrix** of A (where D is the diagonal degree matrix with $D_{ii} = \sum_{j\in[n]} A_{ij}$). Denote by B the $|E| \times n$ **incidence matrix** of A, where we denote the ith row of B as B_i and set B_i to have entry A_{ij} in position i, $-A_{ij}$ in position j, and 0 everywhere else.

Test Statistic. For a given neighborhood graph $G_{n,r}$, let A be the adjacency matrix associated with $G_{n,r}$. Take $L_n = D - A$ to be the (unnormalized) **Laplacian matrix** of A (where D is the diagonal degree matrix with $D_{ii} = \sum_{j\in[n]} A_{ij}$). Denote by B the $|E| \times n$ **incidence matrix** of A, where we denote the ith row of B as B_i and set B_i to have entry A_{ij} in position i, $-A_{ij}$ in position j, and 0 everywhere else. Now, we can define our **Laplacian smooth** test statistic

$$T_2 := \left(\max_{\theta:\|B\theta\|_2 \le C_n} a^T\theta\right)^2 \tag{1}$$

for some sequence of positive numbers $\{C_n\}_{n\in\mathbb{N}} \geq 0$.

Empirical Risk Minimization. To introduce the continuous limit of T_2, it will be useful to slightly recast the variational problem of (1) as an empirical risk minimization problem. (Really, all we are doing is introducing some new notation.) Let ν_n be the **empirical measure** induced by $\{z_1, \ldots, z_n\}$

$$\nu_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$$

Then, for any mapping $u_n : \{z_1, \ldots, z_n\} \to \mathbb{R}$ such that $u_n \in L^2(\nu_n)$, let the **empirical risk functional** $R_n(u_n)$ be given by

$$R_n(u_n) \stackrel{\text{def}}{=} -\sum_{i=1}^N u_n(z_i)\widetilde{\ell}_n(z_i) \tag{2}$$

where $\widetilde{\ell}_n : \{z_1, \ldots, z_n\} \to \{0,1\}$ is the **normalized label function** defined by $\widetilde{\ell}_n(z_i) := a_i$.

To relate the risk functional of (2) to the variational problem of (1), we introduce the **constrained empirical risk functional**, $R_n^{(con)}(u_n)$, defined by

$$R_n^{(con)}(u_n) := \begin{cases} R_n(u_n), & \text{if } \mathcal{E}_n^2(u_n) \le 1 \\ \infty, & \text{otherwise} \end{cases}$$

where $\mathcal{E}_n^2(u_n)$ is the **Laplacian regularization functional** given by

$$\mathcal{E}_n^2(u_n) \stackrel{\text{def}}{=} \frac{1}{n^2\epsilon_n^{d+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{\|z_i - z_j\|}{\epsilon_n}\right)(u_n(z_i) - u_n(z_j))^2.$$

Note that for B the incidence matrix of the ϵ-radius neighborhood graph G_{n,ϵ_n} as defined above, and for $\theta = \{u_1(z_i), \ldots, u_n(z_i)\}$, we have

$$N^2\epsilon_N^{d+1}\mathcal{E}_N^2(u_N) = \|B\theta\|_2\,, \quad \mathcal{E}_N^2(u_N) \leq 1 \Leftrightarrow \|B\theta\|_2 \leq N^2\epsilon_N^{d+1}$$

and $a^T\theta = -R_n(u_n)$. As a result, letting $C_n = N^2\epsilon_N^{d+1}$, we have that for

$$u_n^\star := \operatorname*{argmin}_{u_n \in L^2(\nu_n)} R_n^{(con)}(u_n) \tag{3}$$

the following relation holds:

$$T_2^{1/2} = R_n(u_n^\star)$$

Continuum limit. As the preceding manipulations make clear, the statistic T_2 can be seen as a constrained minimization problem with constraint enforced by a regularization functional over the neighborhood graph $G_{n,\epsilon}$. It is well known that, for an appropriate schedule of $\{\epsilon_n\}_{n\in\mathbb{N}}$ and data generated from a density satisfying certain regularity conditions, such regularization functionals are well behaved in the limit.

Let $\nu = \frac{p+q}{2}$. For $u \in L^2(\nu)$, define the **continuous risk functional** $R(u)$ via

$$R(u) = -\int_D u(x)\big(p(x) - q(x)\big)dx$$

the **weighted L^2 regularization functional**

$$\mathcal{E}_\infty^2(u) = \int_D \|\nabla u(x)\|^2 \mu^2(x)dx$$

and the **constrained continuous risk functional** $R^{(con)}(u)$ as

$$R^{(con)}(u) = \begin{cases} R(u), \text{if } \mathcal{E}_\infty^2(u) \leq 1 \\ \infty, \text{otherwise} \end{cases}$$

Let $u^\star$ be defined analogously to $u_n^\star$,

$$u^\star = \operatorname*{argmin}_{u \in L^2(\nu)} R^{(con)}(u) \tag{4}$$

2 RESULTS

Theorem 1. *Consider a sequence* $\{\epsilon_n\}_{n\in\mathbb{N}} \to 0$ *satisfying*

$$\left(\frac{\log n}{n}\right)^{1/d} = o(\epsilon_n).$$

Then, for $u_n^\star$ *satisfying (3) and likewise* $u^\star$ *satisfying (4), with probability one:*

$$R_n(u_n^\star) \to R(u^\star) \tag{5}$$

Proof of Theorem 1. We know $\mathcal{E}_\infty^2(u^\star) \le 1$ (otherwise $R^{(con)}(0) = 0 \le R^{(con)}(u^\star)$)

By Lemma 2, we have that there exists some $u_n \overset{TL^2}{\to} u^\star$ such that

$$\limsup_{n\to\infty} \mathcal{E}_n^2(u_n) \le \mathcal{E}_\infty^2(u^\star) \le 1$$

and therefore by Lemma 1

$$\limsup_{n\to\infty} R_n(u_n) = R(u^\star)$$

Of course, we do not know that $\mathcal{E}_n^2(u_n) \le 1$, and so we do not know that the $R_n^{(con)}(u_n) < \infty$, even in the limit. However, taking $u_n' = u_n \cdot \left(\max\{1, \mathcal{E}_n^2(u_n)\}\right)^{-1}$, we have $\mathcal{E}_n^2(u_n') \le 1$. Moreover,

$$\begin{aligned} \lim_{n\to\infty} R_n(u_n') &= \lim_{n\to\infty} \frac{R_n(u_n)}{\left(\max\{1, \mathcal{E}_n^2(u_n)\}\right)} \\ &= R(u^\star) \end{aligned} \tag{6}$$

where the latter equality follows from the continuous mapping theorem. Since

$$R_n(u_n^\star) \le R_n^{(con)}(u_n^\star) \le R_n^{(con)}(u_n') = R_n(u_n')$$

we have

$$\lim_{n\to\infty} R_n(u_n^\star) \le R(u^\star).$$

Finally, the above reasoning implies

$$\lim_{n\to\infty} R_n^{(con)}(u_n^\star) \le R(u^\star) < \infty.$$

As a result, clearly

$$\limsup_{n\to\infty} \mathcal{E}_n^2(u_n^\star) \le 1 < \infty$$

and so by Theorem 2, we have that every subsequence of $u_n^\star$ is TL^2 convergent. As a result,

$$\liminf_{n\to\infty} \mathcal{E}_n^2(u_n^\star) \ge \inf_{u\in L^2(\nu)} R(u) = R(u^\star)$$

and so we have shown

$$\lim_{n\to\infty} R_n(u_n^\star) = R(u^\star).$$

$\square$

2.1 Technical results.

Theorem 2 (Garcia-Trillos 17)**.** *Let $d \geq 2$ and let $\mathcal{D} \subset \mathbb{R}^d$ be an open, bounded, connected set with Lipschitz boundary. Let μ be a probability measure on $\mathcal{D}$ with continuous density ρ, satisfying*

$$m \leq \rho(x) \leq M \qquad (\forall x \in D)$$

for some $0 < m \leq M$. Let $z_1, \ldots, z_n$ be a sequence of i.i.d random points chosen according to μ. Let (ϵ_n) be a sequence of positive numbers converging to 0 and satisfying

$$\lim_{n\to\infty} \frac{(\log n)^{3/4}}{n^{1/2}} \frac{1}{\epsilon_n} = 0 \quad \textit{if } d = 2$$
$$\lim_{n\to\infty} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{\epsilon_n} = 0 \quad \textit{if } d \geq 3$$

Assume the kernel K satisfies conditions:

$$K(0) > 0 \textit{and } K \textit{ is continuous at } 0. \qquad (\mathbf{K1})$$
$$K \textit{is non-increasing.} \qquad (\mathbf{K2})$$
$$\textit{The integral } \int_0^\infty K(r) r^{d+1} dr \textit{ is finite.} \qquad (\mathbf{K3})$$

Then, with probability one, the following statement holds:

$$\mathcal{E}_n^2(u_n) \xrightarrow{\Gamma} \mathcal{E}_\infty^2(u)$$

in the TL^2 sense.

Moreover, every sequence (u_n) with $u_n \in L^2(\mu_n)$ for which

$$\sup_{n\in\mathbb{N}} \|u_n\|_{\mu_n} < \infty$$
$$\sup_{n\in\mathbb{N}} \mathcal{E}_n^2(u_n) < \infty$$

is pre-compact in TL^2.

Lemma 1 is very similar to Proposition 2.7 in cite (Garcia-Trillos 16).

Lemma 1. *With probability one the following statement holds: Let $\{u_n\}_{n\in\mathbb{N}}$ be a sequence of $[-1,1]$-valued functions, with $u_n \in L^1(\nu_n)$. If $u_n \overset{TL^1}{\to} u$ as $n \to \infty$, then*

$$\lim_{n\to\infty} R_n(u_n) = R(u).$$

Lemma 2 is needed to prove Lemma 1. It is very similar to a result from cite (Garcia-Trillos 16). To understand Lemma 2, we must introduce a notion of weak convergence of functions.

Definition 2.1. Given a sequence of functions $g_n \in L^1(\nu)$, and $g \in L^1(\nu)$, we say g_n **converges weakly to** g, $g_n \rightharpoonup g$ if for all $L^\infty(\nu)$,

$$\lim_{n\to\infty} \int_D g_n(x) f(x) d\nu(x) = \int_D g(x) f(x) d\nu(x)$$

Given a sequence $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(\nu_n)$, we say that $\{u_n\}_{n\in\mathbb{N}}$ **converges weakly** to $u \in L^1(\nu)$, $u_n \rightharpoonup u$, if the sequence of functions $\{u_n \circ T_n\} \in L^1(\nu)$ converges weakly to u, for T_n stagnating transportation maps.

Lemma 2. *For the label function $\ell_n : \{z_1, \dots, z_n\} \to \{-1, 1\}$ defined by*

$$\ell_n(z_i) := \ell_i, \quad i \in 1, \dots, n \tag{7}$$

with probability one $\ell_n \rightharpoonup \frac{p-q}{2\mu}$.

Lemma 3. *The Laplacian regularization functional $\mathcal{E}_n^2(u_n)$ satisfies the following three properties,*

3 PROOFS

Proof of Lemma 1. We begin by removing the effect of the random normalization in $\widetilde{\ell}_n(z_i)$ via

$$-R_n(u_n) = \frac{2}{n}\sum_{i=1}^n u_n(z_i)\ell_n(z_i) + \sum_{i=1}^n \left(\frac{2}{n}\ell_n(z_i) - \widetilde{\ell}_n(z_i)\right) u_n(z_i)$$

First, we show the second term converges to zero with probability one,

$$\begin{aligned} \left|\sum_{i=1}^n \left(\frac{2}{n}\ell_n(z_i) - \widetilde{\ell}_n(z_i)\right) u_n(z_i)\right| &\le \sum_{i=1}^n \left|\frac{2}{n}\ell_n(z_i) - \widetilde{\ell}_n(z_i)\right| |u_n(z_i)| \\ &\le \left(\left|\frac{1}{N} - \frac{1}{n/2}\right| + \left|\frac{1}{M} - \frac{1}{n/2}\right|\right) \sum_{i=1}^n |u_n(z_i)| \\ &\le \left(\left|\frac{N - n/2}{N}\right| + \left|\frac{M - n/2}{M}\right|\right) \cdot \frac{2}{n}\sum_{i=1}^n |u_n(z_i)| \end{aligned}$$

Then, the TL^1 convergence of u_n to $u \in L^1(\nu)$ implies

$$\frac{2}{n}\sum_{i=1}^n |u_n(z_i)| \xrightarrow{n} 2\,\|u\|_{L^1(\nu)}$$

and by standard concentration results of binomial random variables, with probability one

$$\left|\frac{N - n/2}{N}\right|, \left|\frac{M - n/2}{M}\right| \xrightarrow{n} 0,$$

and so with probability one

$$\left|\sum_{i=1}^{n}\left(\frac{2}{n}\ell_n(z_i)-\widetilde{\ell}_n(z_i)\right)u_n(z_i)\right| \xrightarrow{n} 0.$$

Now, we rewrite the first term in the summand on the right hand side using transportation maps,

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}u_n(z_i)\ell_n(z_i) &= 2\int_D u_n(z)\ell_n(z)d\nu_n(z)\\ &\overset{(i)}{=} 2\int_D \big(u_n\circ T_n(z)\big)\big(\ell_n\circ T_n(z)\big)d\nu(z)\\ &= 2\int_D \big(u_n\circ T_n(z)-u(z)\big)\big(\ell_n\circ T_n(z)\big)d\nu(z)+2\int_D \big(u(z)\big)\big(\ell_n\circ T_n(z)\big)d\nu(z)\end{aligned}$$

with (i) following from the change of variables formula

$$\int_D f(T(x))d\theta(x) = \int_D f(z)dT_\sharp\theta x$$

where $f : D \to \mathbb{R}$ is an arbitrary Borel function, θ a Borel measure, and $T_\sharp\theta$ the push-forward measure of θ.

Now, the first term converges to zero,

$$\left|2\int_D \big(u_n\circ T_n(z)-u(z)\big)\big(\ell_n\circ T_n(z)\big)d\nu(z)\right| \le 2\int_D |u_n\circ T_n(z)-u(z)|\,d\nu(z) \xrightarrow{n} 0.$$

by the boundedness of ℓ_n and the TL^1 convergence of u_n to u.

By Lemma 2, with probability one the 2nd term converges to (negative of) the risk functional,

$$2\int_D \big(u(z)\big)\big(\ell_n\circ T_n(z)\big)d\nu(z) \xrightarrow{n} \int_D u(z)\left(\frac{p(z)-q(z)}{\mu(z)}\right)d\nu(z) = -R(u).$$

□

Proof of Lemma 1. Lemma 2 is essentially a restatement of Lemma 2.5 from cite (Garcia-Trillos 16). For completeness purposes, we restate that result here.

Lemma 4. *Let $f(z) = \mathbb{E}(\ell|Z = z)$ be the conditional expectation of ℓ given $Z = z$. For the label function $\ell_n : \{x_1, \ldots, x_n\} \to \{-1, 1\}$ defined as in (7), with probability one $\ell_n \rightharpoonup f$.*

Given Lemma 4, all that is needed to show Lemma 2 is that $f = \frac{p-q}{2\mu}$. This follows from a simple application of Bayes Rule. □