Testing with Graphs

Alden Green

February 6, 2019

Two-sample non-parametric hypothesis-testing problem. For fixed integers $n_1 + n_2 = n$, let $\mathbf{X} = \{x_1, \ldots, x_{n_1}\} \subset \mathbb{R}^d$ and $\mathbf{Y} = \{y_1, \ldots, y_{n_2}\}$ 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 $\mathbf{Y}$.

Integral Probability Metric. Given $\mathcal{F}$ a class of real-valued bounded measurable functions on D, the *integral probability metric* between $\mathbb{P}$ and $\mathbb{Q}$ with respect to $\mathcal{F}$ is

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_D f d\mathbb{P} - \int_D f d\mathbb{Q} \right|$$

One such IPM that has not received close attention until surprisingly recently is the *(weighted) Sobolev IPM*. For $f \in L^2(\mathcal{D})$ and ρ a density function over $\mathcal{D}$, the (ρ^2-weighted) Sobolev 1,2 norm of f is given by

$$\|f\|_{1,2,\rho^2} := \begin{cases} \int_{\mathcal{D}} \|\nabla_x f(x)\|^2 \rho^2 dx, & f \in H^1(\mathcal{D}) \\ \infty, & f \in L^2(\mathcal{D}) \setminus H^1(\mathcal{D}) \end{cases}$$

where $H^1(\mathcal{D})$ is the Sobolev space of $L^2(\mathcal{D})$ functions with weak derivative $\nabla_x f(x) \in L^2(\mathcal{D})$.

Consider the unit ball of $\|\cdot\|_{1,2,\rho^2}$,

$$\mathcal{W}^{1,2}(\mathcal{D}, \rho^2) := \left\{ f : \|f\|_{1,2,\rho^2} \leq 1 \right\}.$$

The weighted Sobolev IPM is simply $\gamma_{\mathcal{W}^{1,2}(\mathcal{D},\rho^2)}(\mathbb{P}, \mathbb{Q})$.

Neighborhood graph. Let $G_{n,r_n} = (V, E, w)$ denote a weighted, undirected graph constructed from the samples $\mathbf{Z} = \{z_1, \ldots, z_n\} = (\mathbf{X}, \mathbf{Y})$ where $V = \{1, \ldots, n\}$, and $w_{uv} = K(z_u, z_v) := k(\frac{\|z_u - z_v\|}{\epsilon_n}) \geq 0$ for $u, v \in V$, and a particular kernel function k. Here $(u, v) \in E$ if and only if $w_{uv} > 0$.

Motivated by the integral probability metric, our test statistics will be of the form

$$\gamma_{\mathcal{F}_n}(\mathbb{P}_n, \mathbb{Q}_n) = \sup_{f_n \in \mathcal{F}_n} \left| \int_D f_n d\mathbb{P}_n - \int_D f_n d\mathbb{Q}_n \right|$$

where

$$\mathbb{P}_n := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{x_i}, \quad \mathbb{Q}_n := \frac{1}{n_2} \sum_{i=1}^{n_2} \delta_{y_i}$$

are the empirical distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\mathcal{F}_n$ is a class of functions $f_n : \mathbf{Z} \to \mathbb{R}$ exhibiting some regularity with respect to the neighborhood graph $G_{n,r}$.

Laplacian smoothing and Total Variation denoising. For convenience, we number the edges $E = (e_1, \ldots, e_m)$. We denote by $\mathbf{B} \in \mathbb{R}^{m \times n}$ the edge incidence matrix of $G_{n,r}$, which for kth edge $e_k = (u, v)$ has kth row $\mathbf{B}_k = (0, \ldots, -w_{uv}, \ldots, w_{uv}, \ldots, 0)$ with a $-w_{uv}$ in the uth location, and a w_{uv} in the vth location. The random (unnormalized) Laplacian matrix is then $\mathbf{L} = \mathbf{B}^T \mathbf{B}$. We also introduce a *label vector*, given by $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_n)$ with

$$\ell_k = \begin{cases} \frac{n}{n_1}, & z_k \in \mathbf{X} \\ -\frac{n}{n_2}, & z_k \in \mathbf{Y} \end{cases} \tag{1}$$

Our test statistics $T_1(\boldsymbol{\ell}; G_{n,r})$ and $T_2(\boldsymbol{\ell}; G_{n,r})$ are defined as follows:

$$T_1(\boldsymbol{\ell}; G_{n,r}) := \sup_{\mathbf{f} \in \mathbb{R}^n : \|\mathbf{B}\mathbf{f}\|_1 \leq C_{n,r}} \frac{1}{n} \sum_{k=1}^{n} \ell_k f_k$$

$$T_2(\boldsymbol{\ell}; G_{n,r}) := \sup_{\mathbf{f} \in \mathbb{R}^n : \|\mathbf{B}\mathbf{f}\|_2^2 \leq C_{n,r}} \frac{1}{n} \sum_{k=1}^{n} \ell_k f_k$$

where $C_{n,r} = \frac{\sigma_k}{n^2 r_n^{d+2}}$ and we write $\mathbf{f} = (f_1, \ldots, f_n)$. We note that, as promised, these satisfy the form

$$T_1(\boldsymbol{\ell}; G_{n,r}) = \sup_{f_n \in TV_n} \left| \int_D f_n d\mathbb{P}_n - \int_D f_n d\mathbb{Q}_n \right|$$

$$T_2(\boldsymbol{\ell}; G_{n,r}) = \sup_{f_n \in \mathcal{W}_n} \left| \int_D f_n d\mathbb{P}_n - \int_D f_n d\mathbb{Q}_n \right|$$

where $TV_n = \{\mathbf{f} : \|\mathbf{B}\mathbf{f}\|_1 \leq C_{n,r}\}$ and $\mathcal{W}_n = \{\mathbf{f} : \|\mathbf{B}\mathbf{f}\|_2 \leq C_{n,r}\}$.

1 Consistency under fixed alternative

Binomialized data model. For technical reasons, we would like $z_1, \ldots, z_n$ to be independent and identically distributed. We consider the following generative model, which we term the *binomialized data model*:

Fix $n \in \mathbb{N} > 0$, and $n_1 \sim \text{Bin}(n, 1/2),\ n_2 = n - n_1$. Then, let $x_1, \ldots, x_{n_1} \in \mathbb{R}^d$ be a sequence of i.i.d random points chosen according to $\mathbb{P}$, and $y_1, \ldots, y_{n_2} \in \mathbb{R}^d$ a separate sequence of i.i.d random points chosen according to $\mathbb{Q}$, with $x_j \perp y_k$ for all j, k. Fix $\widetilde{\mathbf{Z}} = (\widetilde{z}_1, \ldots, \widetilde{z}_n) := (x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2})$. Finally, for a permutation $\pi : [n] \to [n]$ chosen uniformly at random among all such permutations, let $\mathbf{Z} = (z_1, \ldots, z_n) = (\widetilde{z}_{\pi(1)}, \ldots, \widetilde{z}_{\pi(n)})$.

The label vector $\boldsymbol{\ell}$ remains defined as in (1) with respect to $\mathbf{Z}$. Note that now $z_i \overset{i.i.d}{\sim} \frac{\mathbb{P}}{2} + \frac{\mathbb{Q}}{2}$, as we desired, with density function $\mu(x) := \frac{p(x)+q(x)}{2}$.

Theorem 1 (Pointwise limit of Laplacian smooth test statistic.)**.** *Let $d \geq 2$ and let $\mathcal{D} \subset \mathbb{R}^d$ be an open, bounded, connected set with Lipschitz boundary. Let μ satisfy*

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

for some $0 < m \leq M$. Let (r_n) be a sequence of positive numbers converging to 0 and satisfying

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

Assume the kernel k satisfies conditions:

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

Then with probability one the following statement holds: For $(z_1, \ldots, z_n)$ chosen under the binomialized data model,

$$\lim_{n\to\infty} T_2(\boldsymbol{\ell}; G_{n,r_n}) = \gamma_{\mathcal{W}^{1,2}(\mathcal{D},\mu^2)}(\mathbb{P}, \mathbb{Q}).$$

2 Proofs

Fix $\mu_n = \frac{d\mathbb{P}_n + d\mathbb{Q}_n}{2}$. We will show a variational form of convergence of μ_n to μ.

2.1 Gamma convergence of constraint

Definition 2.1 (TL^2 convergence)**.** Denote by $\mathfrak{B}(\mathcal{D})$ the Borel σ-algebra of $\mathcal{D}$ and $\mathcal{P}(\mathcal{D})$ the set of all Borel probability measures on $\mathcal{D}$. Given a Borel map

$T : \mathcal{D} \to \mathcal{D}$, the *push-forward* of μ by T is given by

$$T_\star\mu(\mathcal{A}) := \mu(T^{-1}(\mathcal{A})), \quad \mathcal{A} \in \mathfrak{B}(\mathcal{D}).$$

Given $\widetilde{\mu} \in \mathcal{P}(\mathcal{D})$, we say that T is a *transportation map* between μ and $\widetilde{\mu}$ if $T_\star\mu = \widetilde{\mu}$. If for a sequence (T_n) of transportation maps

$$\int_{\mathcal{D}} |x - T_n(x)|^2 \, d\mu(x) \to 0, \text{ as } n \to \infty$$

we refer to the sequence as *stagnating.*

Take $f \in L^2(\mu)$ and a sequence (f_n) with $f_n \in L^2(\mu_n)$ for $n = 1, 2, \ldots$. If there exists a stagnating sequence of transportation maps (T_n) such that

$$\int_{\mathcal{D}} |f - T_n(f_n(x))|^2 \, d\mu(x) \to 0, \text{ as } n \to \infty$$

we say that (f_n) converges TL^2 to f, and write $f_n \overset{TL^2}{\to} f$.

We restate Theorem 1.4 of [1], changing notation to match the rest of this paper.

Theorem 2 (Theorem 1.4 of [1])**.** *Under the setup and conditions of Theorem 1, with probability one the following statements hold:*

- Liminf inequality: *For all $f \in L^2(\mu)$ and all sequences (f_n) with $f_n \in L^2(\mu_n)$ and $f_n \overset{TL^2}{\to} f$,*

$$\liminf_{n\to\infty} \frac{1}{n^2 r_n^{d+2}} \|Bf\|_2^2 \geq \sigma_k \|f\|_{1,2,\rho^2}$$

- Limsup inequality: *For all $f \in L^2(\mu)$, there exists a sequence (f_n) with $f_n \in L^2(\mu_n)$ and $f_n \overset{TL^2}{\to} f$ such that*

$$\limsup_{n\to\infty} \frac{1}{n^2 r_n^{d+2}} \|\mathbf{B}\mathbf{f}_n\|_2^2 \leq \sigma_k \|f\|_{1,2,\rho^2}$$

  where $\mathbf{f}_n = (f_n(z_1), \ldots, f_n(z_n))$.

- Compactness property: *Every sequence (f_n) with $f_n \in L^2(\mu_n)$ satisfying*

$$\sup_{n\in\mathbb{N}} \frac{1}{n^2 r_n^{d+2}} \|Bf\|_2^2 < \infty$$

  is precompact in TL^2, that is, every subsequence of (f_n) has a further subsequence which converges in the TL^2-sense to an element of $L^2(\mathcal{D})$.

2.2 Continuity of risk functional

Lemma 1. *With probability one the following statement holds: If a sequence (f_n) where $f_n : \{z_1, \ldots, z_n\} \to [-1,1]^n$ converges TL^2 to $f \in L^2(\mu)$, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \ell_k f_n(z_k) = \int_{\mathcal{D}} f d\mathbb{P}(x) - \int_{\mathcal{D}} f d\mathbb{Q}(x).$$

2.3 Proof of Theorem 1

Let $f^\star$ be the witness function for $\gamma_{\mathcal{W}^{1,2}(\mathcal{D},\mu^2)}(\mathbb{P},\mathbb{Q})$, meaning

$$\left|\int_{\mathcal{D}} f^\star d\mathbb{P} - \int_{\mathcal{D}} f^\star d\mathbb{Q}\right| = \gamma_{\mathcal{W}^{1,2}(\mathcal{D},\mu^2)}(\mathbb{P},\mathbb{Q})$$

where $f^\star \in \mathcal{W}^{1,2}(\mathcal{D},\mu^2)$ implies $f^\star \in L^2(\mu)$. By the limsup inequality in Theorem 2, there exists some $(f_n) \overset{TL^2}{\to} f$ with $f_n \in L^2(\mu_n)$ such that

$$\limsup_{n\to\infty} \frac{1}{n^2 r_n^{d+2}} \left\|\mathbf{B}\mathbf{f}_n\right\|_2^2 \le \sigma_k \left\|f\right\|_{1,2,\mu^2} \le \sigma_k. \tag{2}$$

From Lemma 1

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \ell_k f_n(z_k) = \int_{\mathcal{D}} f d\mathbb{P}(x) - \int_{\mathcal{D}} f d\mathbb{Q}(x) = \gamma_{\mathcal{W}^{1,2}(\mathcal{D},\mu^2)}.$$

and along with (2), this implies

$$\lim_{n\to\infty} T_2(\boldsymbol{\ell}; G_{n,r}) \le \gamma_{\mathcal{W}^{1,2}(\mathcal{D},\mu^2)}(\mathbb{P},\mathbb{Q}).$$

Let $\mathbf{L}^\dagger$ be the pseudoinverse of the Laplacian matrix $\mathbf{L}$. We introduce $f_n^\star = \boldsymbol{\ell}^T\mathbf{L}^\dagger\boldsymbol{\ell}$, which satisfies

$$\left|\int_{\mathcal{D}} f_n^\star d\mathbb{P}_n - \int_{\mathcal{D}} f_n^\star d\mathbb{Q}_n\right| = T_2(\boldsymbol{\ell}; G_{n,r})$$

Note that

$$\|\mathbf{B}\mathbf{f}_n^\star\|_2^2 \le C_n \Longrightarrow \frac{1}{n^2 r_n^{d+2}} \|\mathbf{B}\mathbf{f}_n^\star\|_2^2 \le \sigma_k < \infty$$

and so by the compactness property in Theorem 2, every subsequence of $f_n^\star$ has a further subsequence which is TL^2-convergent. For simplicity, and without loss of generality, let us work along a subsequence which is convergent (and call it $f_n^\star$) to $f \in L^2(\mu)$. Then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \ell_k f_n^\star(z_k) = \left|\int_{\mathcal{D}} f d\mathbb{P}(x) - \int_{\mathcal{D}} f d\mathbb{Q}(x)\right| \tag{3}$$

and by the liminf inequality of Theorem 2,

$$\liminf_{n\to\infty} \frac{1}{n^2 r_n^{d+2}} \left\|\mathbf{B}\mathbf{f}_n\right\|_2^2 \geq \sigma_k \left\|f\right\|_{1,2,\mu^2}$$

which implies $\|f\|_{1,2,\mu^2} \leq 1$. This, along with (3), implies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \ell_k f_n^\star(z_k) \geq \gamma_{\mathcal{W}^{1,2}(\mathcal{D},\mu^2)}(\mathbb{P},\mathbb{Q})$$

REFERENCES

[1] Nicolas Garcia Trillos and Dejan Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018.