Graph Testing Notes

Alden Green

December 18, 2018

1 Setup

Data model. We are given two distributions, P and Q, with the ability to sample from either one. Our goal is to test the hypothesis $H_0 : P = Q$ vs. the alternative $H_a : P \neq Q$.

Under the **binomial data model**, our sampling procedure is to draw i.i.d Rademacher labels $L_i \in \{1, -1\}$ for $i \in \{1, \ldots, N\}$, and then sample $Z_i \sim P$ if $L_i = 1$ and $Z_i \sim Q$ otherwise. Define 1_X to be the length-N indicator vector for $L_i = 1$

$$1_X[i] = \begin{cases} 1, L_i = 1 \\ 0 \text{ otherwise} \end{cases}$$

and similarly for 1_Y

$$1_Y[j] = \begin{cases} 1, L_i = -1 \\ 0 \text{ otherwise} \end{cases}$$

and define $a = \frac{1_X}{N/2} - \frac{1_Y}{N/2}$.

Under the **fixed label data model** we use the same data generating process as above, except fix $\mathcal{L}_X = \{1, \ldots, N/2\}$ and $\mathcal{L}_Y = \{N/2, \ldots, N\}$. Say that $L_i = 1$ for $i \in \mathcal{L}_X$ and $L_i = -1$ for $i \in \mathcal{L}_Y$, and call $\left\{X_1, \ldots, X_{|\mathcal{L}_X|}\right\} = \{Z_i : i \in \mathcal{L}_X\}$ and likewise for Y.

Graph. Form an $N \times N$ Gram matrix A, where $A_{ij} = K(Z_i, Z_j)$ for **kernel function** K. Let $G = (V, E)$ with $V = \{Z_1, \ldots, Z_n\}$ and $E = \{A_{ij} : 1 \leq i < j \leq n\}$. Take $L = D - A$ to be the (unnormalized) **Laplacian matrix** of A (where D is the diagonal degree matrix with $D_{ii} = \sum_{j \in [n+m]} A_{ij}$). Denote by B the $M \times N$ **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.

Resistance distances. There are many distances one can define over nodes in a graph. The **resistance distance between nodes u and v, R_{uv},** is defined as

$$R_{uv} = (e_u - e_v)^T L^\dagger (e_u - e_v).$$

Test statistics. We begin by defining our **laplacian smooth** test statistic.

$$T_2 = \left(\max_{\theta : \|B\theta\|_2 \leq 1} a^T \theta \right)^2 = a^T L^\dagger a.$$

(Bhattacharya 2018) defines a general notion of 2-sample **graph-based test statistics**

$$T_{\mathcal{G}} = \frac{1}{N^2} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} A_{ij}$$

Although he develops theory for this statistic in the context of kNN and minimum spanning tree graphs, we will at present consider it for the complete weighted similarity graph defined by A above. Then, we can write

$$T_{\mathcal{G}} = a^T L a.$$

Finally, define $\mathcal{H}$ to be a **reproducing kernel Hilbert space** with K the associated kernel. Let $\mathcal{F}$ be the unit ball of $\mathcal{H}$, and let the evaluation of $f \in \mathcal{F}$ at the sample points $Z_1, \ldots, Z_N$ be denoted by $\mathbf{f} = f(Z_1, \ldots, Z_N)$. Then, the statistic MMD_b of (Gretton 2012) can be written as

$$T_{\mathcal{K}} = \sup_{f \in \mathcal{F}} a^T \mathbf{f} = a^T K a.$$

Distances between probability measures. We will need distances between probability measures for two different purposes. The first is that they are self-evidently useful in analyzing limiting distributions of statistics (in particular in this case, our test statistics).

For a function f, define its **Lipschitz norm** $\|f\|_L$ to be

$$\inf \left\{ K : |f(x) - f(y)| \leq K \|x - y\| \right\}.$$

Define the **Wasserstein distance** between two measures μ and ν to be

$$\mathcal{W}(\mu, \nu) := \sup \left\{ \left| \int h \, d\mu - \int h \, d\nu \right| : h \text{ Lipschitz, with } \|h\|_L \leq 1 \right\}.$$

If the measures μ and ν have corresponding cumulative distribution functions F_μ and F_ν then we can define the **Kolmogorov-Smirnov distance** to be

$$\|F_\mu - F_\nu\|_\infty := \sup_t |F_\mu(t) - F_\nu(t)|.$$

The second reason we will use distances between probability measures is that they themselves make for good test statistics!

An **integral probability metric** (IPM) with respect to a function class $\mathcal{F}$ is defined

$$\sup_{f\in\mathcal{F}} \mathbb{E}\left[f(X)\right] - \mathbb{E}\left[f(Y)\right]$$

for $X \sim P$, $Y \sim Q$.

Hereafter, we will assume P and Q are absolutely continuous with respect to Lebesgue measure, with density functions p and q, respectively. Denote the **mixture density** by $\mu = \frac{p+q}{2}$.

Denote the **gradient** of a function f by ∇_x. Then we can define the **Sobolev semi-norm** and **dot product**, $\|f\|_{W_0^{1,2}(\mathcal{X},\mu^2)}$ and $\langle f,g\rangle_{W_0^{1,2}(\mathcal{X},\mu^2)}$, by

$$\langle f,g\rangle_{W_0^{1,2}(\mathcal{X},\mu)} = \int_{\mathcal{X}} \langle \nabla_x f(x), \nabla_x g(x)\rangle_{\mathbb{R}^d} \mu^2(x), \quad \|f\|_{W_0^{1,2}(\mathcal{X},\mu)} = \sqrt{\int_{\mathcal{X}} \left\|\nabla_x f(x)\right\|^2 \mu^2(x)dx}$$

Let the **Sobolev space**, $W^{1,2}(\mathcal{X},\mu^2)$, be

$$W^{1,2}(\mathcal{X},\mu^2) = \left\{ f:\mathcal{X}\to\mathbb{R}, \int_{\mathcal{X}} \left\|\nabla_x f(x)\right\|^2 \mu^2(x)dx < \infty \right\}.$$

and denote by $W_0^{1,2}(\mathcal{X},\mu^2)$ the restriction of $W^{1,2}(\mathcal{X},\mu^2)$ to functions which vanish at the boundary of $\mathcal{X}$. Note that $\|f\|_{W_0^{1,2}(\mathcal{X},\mu^2)}$ defines a semi-norm over $W_0^{1,2}(\mathcal{X},\mu^2)$. Finally, let $B_W(\mathcal{X},\mu^2)$ be the **unit ball** of $W_0^{1,2}(\mathcal{X},\mu^2)$, meaning

$$B_W(\mathcal{X},\mu^2) = \left\{ f \in W_0^{1,2}(\mathcal{X},\mu^2) : \|f\|_{W_0^{1,2}(\mathcal{X},\mu^2)} \le 1 \right\}$$

Now we can define the **Sobolev IPM**, $\mathcal{S}_{\mu^2}(P,Q)$ It is simply an IPM where the function class is the Sobolev unit ball with respect to μ^2.

$$\mathcal{S}_{\mu^2}(P,Q) \overset{\text{def}}{=} \sup_{f\in B_W} \left\{ \mathbb{E}\left[f(X)\right] - \mathbb{E}\left[f(Y)\right] \right\}$$

Holder functions. We will show that the Laplacian constraint $\|B\theta\|_2 \le 1$ is very similar to the constraint $f_\theta \in B_W(X,\mu^2)$ for the right choice of K, over all Holder functions.

For mapping $f : \mathbb{R}^d \to \mathbb{R}$ and β a positive integer, we say f is a β**-Holder function** if there exists $C > 0$ such that for all $x, y \in \mathcal{X}$

$$\left| f^{(\beta-1)}(x) - f^{(\beta-1)}(y) \right| \le K \left\|x-y\right\|$$

Roughly speaking, this means the functions have bounded β partial derivatives.

2 Conjectures

Conjectures 1 and 2 will be needed for Theorem 2.

Conjecture 1. There exists a sequence of scaling factors $(\rho_n)_{n=1}^{\infty}$ such that the **spectral measure** μ_n of $\rho_n L^\dagger$ converges weakly in probability

$$\mu_n(\rho_n L^\dagger) \xrightarrow{*} \nu_\infty.$$

where $V \sim \nu_\infty$ and $V_n \sim \mu_n$ are bounded almost surely for all n by some constant C.

Conjecture 2. For all $\epsilon > 0$, there exists N such that

$$\mathbb{P}\left(\max_{i\in[n]} \frac{1}{n}\left(\{\rho_n L^\dagger\}^2\right)_{ii} \le \epsilon\right) \ge 1-\epsilon$$

for all $n \ge N$.

3 DESIRED RESULTS

Theorem 1. For bandwidth parameter $h > 0$ and decreasing function $k(\cdot,\cdot)$, write

$$K(Z_i, Z_j) = \frac{1}{h^m} k(\|Z_i - Z_j\|^2 / h^2).$$

For Sobolev IPM $\mathcal{S}_{\mu^2}(P,Q)$ as defined above,

$$\sqrt{T_2} \xrightarrow{p} \mathcal{S}_{\mu^2}(P,Q)$$

Proof attempt of Proposition 1. Recall that, for incidence matrix B,

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

We expand $\left|\sqrt{T_2} - \mathcal{S}_{\mu^2}(P,Q)\right|$,

$$\begin{aligned}\left|\sqrt{T_2} - \mathcal{S}_{\mu^2}(P,Q)\right| \le &\left|\max_{\theta:\|B\theta\|_2\le 1}\left\{a^T\theta\right\} - \sup_{f\in B_W(\mathcal{X},\mu^2)}\left\{\mathbb{P}_n(f) - \mathbb{Q}_n(f)\right\}\right| \\ &+ \left|\sup_{f\in B_W(\mathcal{X},\mu^2)}\left\{\mathbb{P}_n(f) - \mathbb{Q}_n(f)\right\} - \sup_{f\in B_W(\mathcal{X},\mu^2)}\left\{\mathbb{P}(f) - \mathbb{Q}(f)\right\}\right|\end{aligned} \tag{1}$$

(The following statement would hold only if Proposition 4 held over $B_W(\mathcal{X},\mu^2)$, rather than over $B_W([0,1],\lambda)$ for λ Lebesgue measure.)

By Proposition 4, the second term in the summand on the right hand side of (1) is $o_P(1)$.

(The following statement would hold only if Proposition 5 were uniform over $B_W(\mathcal{X}, \mu^2)$ rather than over the class of α-Holder functions $\mathcal{F}_\alpha$)

Then, Proposition 5 implies that for any $\epsilon > 0$, there exists N such that for $n \geq N$,

$$\sup_{f \in B_W(\mathcal{X}, \mu^2)} \{\mathbb{P}_n(f) - \mathbb{Q}_n(f)\} - \max_{\theta: \|B\theta\|_2 \leq 1} \left\{a^T \theta\right\} \leq \epsilon$$

with high probability.

To complete the proof, we will have to show that for any $\epsilon > 0$, there exists N such that for $n \geq N$,

$$\max_{\theta: \|B\theta\|_2 \leq 1} \left\{a^T \theta\right\} - \sup_{f \in B_W(\mathcal{X}, \mu^2)} \{\mathbb{P}_n(f) - \mathbb{Q}_n(f)\} \leq \epsilon$$

with high probability. □

4 Results

Expectation of two-sample test statistics. The expectation of the above statistics is potentially a good way to understand their large sample behavior, as quadratic forms often satisfy laws of large numbers assuming the matrices are well-conditioned.

Proposition 1. Draw Z and a under the binomial data model, and assume both P and Q are absolutely continuous with respect to Lebesgue measure over Euclidean space $\mathbb{R}^d$. Write $h_0(x) = p(x) - q(x)$ with empirical analogue $\mathbf{h_0} = (h_0(Z_1), \ldots, h_0(Z_n))$. Then

$$\begin{aligned}\mathbb{E}\left[T_{\mathcal{G}}\right] = &\int\int K(\|\mathbf{x}-\mathbf{y}\|)\left[p(\mathbf{x})+q(\mathbf{x})\right]\left[p(\mathbf{y})+q(\mathbf{y})\right]d\mathbf{x}d\mathbf{y} \\ &- \frac{N}{N-1}\int\int K(\|\mathbf{x}-\mathbf{y}\|)\left[h_0(\mathbf{x})\right]^2 \frac{p(\mathbf{y})+q(\mathbf{y})}{p(\mathbf{x})+q(\mathbf{x})}d\mathbf{x}d\mathbf{y} \\ &+ \frac{N}{N-1}\int\int K(\|\mathbf{x}-\mathbf{y}\|)\left[h_0(\mathbf{x})-h_0(\mathbf{y})\right]^2\left[p(\mathbf{y})+q(\mathbf{y})\right]\left[p(\mathbf{x})+q(\mathbf{x})\right]d\mathbf{x}d\mathbf{y}.\end{aligned} \tag{2}$$

Note that even under the null hypothesis, where $h_0 = 0$,

$$\mathbb{E}\left[T_{\mathcal{G}}\right] = \int\int K(\|\mathbf{x}-\mathbf{y}\|)\left[p(\mathbf{x})+q(\mathbf{x})\right]\left[p(\mathbf{y})+q(\mathbf{y})\right]d\mathbf{x}d\mathbf{y}$$

which is not distribution-free, unlike in the case of the kNN graph.

Another interesting consequence of Proposition 1 comes when we take $K(\mathbf{x}, \mathbf{y}) = K(\frac{\|\mathbf{x}-\mathbf{y}\|}{t})$ and let $t \to 0$.

Proposition 2. If p and q are Lipschitz continuous functions with bounded hessians, and K is a continuous function on R^+ such that $x^{2+d}K(x) \in L_2$, then under the same setup as in Proposition 1,

$$\frac{N-1}{N}\lim_{t\to 0}\frac{1}{t^d}\mathbb{E}\left[T_{\mathcal{G}}\right] = \int h_0(\mathbf{x})^2 d\mathbf{x} + \int (p(\mathbf{x}) + q(\mathbf{x}))^2 d\mathbf{x} \tag{3}$$

We turn now to the expectation of the Laplacian smooth statistic.

Proposition 3. Under the fixed label data model

$$\mathbb{E}\left[a^T L^\dagger a\right] = \mathbb{E}\left[R_{X_1 Y_1}\right] - \frac{N-1}{2N}\mathbb{E}\left[R_{X_1,X_2}\right] - \frac{N-1}{2N}\mathbb{E}\left[R_{Y_1,Y_2}\right] \tag{4}$$

Note that, although the resistance distances are between only two nodes, in each case the expectation is over the entire (random) graph G.

Asymptotic null distribution for T_2. We can compute an asymptotic null distribution for T_2, although its formulation depends on the eigenvalues of the matrix $L^\dagger$ which themselves are not obvious.

Theorem 2. Denote the scaled version of the Laplacian smooth test statistic

$$W_n = \sqrt{\frac{N^4}{32 \cdot \mathrm{tr}((L^\dagger)^2)}}\left(T_2^2 - \frac{\mathrm{tr}(L^\dagger)}{4N^2}\right).$$

If Conjectures 1 and 2 hold,

$$\lim_{n\to\infty}\sup_t \left|\mathbb{P}\left(W_n \le t\right) - \Phi(t)\right| = 0.$$

To prove Theorem 2, we will need the following calculations of moments under H_0.

Lemma 1. Under H_0, the conditional expectation $\mathbb{E}\left[T_2|Z\right] = \frac{\mathrm{tr}(L^\dagger)}{N^2}$.

Lemma 2. Under H_0, the conditional variance $\mathrm{Var}\left(T_2|Z\right) = \frac{32\mathrm{tr}[(L^\dagger)^2]}{N^4}$.

5 Supplemental Results

Empirical process over Sobolev classes. The following theorem is a stand-in; it handles only functions with domain on the unit interval, and is stated specifically with respect to Lebesgue measure.

Proposition 4. Let $\mathcal{F}$ be the set of all absolutely continuous functions $f : [0,1] \to \mathbb{R}$ such that $\|f\|_\infty \le 1$ such that $\int (f'(x))^2 dx \le 1$. Then, there exists a constant K such that for every $\epsilon > 0$,

$$\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le K \left(\frac{1}{\epsilon}\right).$$

Thus, the class $\mathcal{F}$ is P-Donsker (and P-Glivenko-Cantelli) for all P.

Regularization functionals. When taking the supremum over functions which satisfy $\|B\theta\|_2 \le 1$, we will argue that this constraint is well-behaved in the limit, i.e. that it converges to the **regularization functional** $\|\cdot\|_{W_0^{1,2}(\mathcal{X},\mu^2)}$. Proposition 5 makes this convergence uniform over the set of 3-Holder functions (essentially functions with bounded $3rd$ derivative). Proposition 6 makes this convergence only pointwise, but merely requires that f have bounded 2nd derivative.

Proposition 5. Let $\mathcal{F}_\alpha$ be a unit ball in the space of α-Holder functions, and define $k(\cdot,\cdot)$ as in Theorem 1. For function $f \in \mathcal{F}_\alpha$, denote f evaluated on the data, $\mathbf{f} = (f(Z_1), \ldots, f(Z_N))$. Then, there exists a constant c depending only on k such that for $\alpha \ge 3$ and a sequence $(h_n) \to 0$ such that

$$\sup_{f \in \mathcal{F}_\alpha} \left| \|B\mathbf{f}_2\| - \|f\|_{W_0^{1,2}(\mathcal{X},\mu^2)} \right| \xrightarrow{p} 0$$

Proposition 6 (Bousquet 04)**.** If p and q are Lipschitz continuous functions with bounded hessians, and K is a continuous function on R^+ such that $x^{2+d}K(x) \in L_2$, then

$$\begin{aligned} &\lim_{t \to 0} \frac{d}{Ct^{d+2}} \int K(\|\mathbf{x}-\mathbf{y}\|/t)(h_0(\mathbf{x}) - h_0(\mathbf{y}))^2 (p(\mathbf{x}) + q(\mathbf{x}))(p(\mathbf{y}) + q(\mathbf{y})) d\mathbf{x} d\mathbf{y} \\ &= \int \|\nabla h_0(\mathbf{x})\|^2 (p(\mathbf{x}) + q(\mathbf{x}))^2 d\mathbf{x} \end{aligned} \tag{5}$$

Lemma 3 (von Luxburg 12)**.** Assume P and Q are absolutely continuous with respect to Lebesgue measure on Euclidean space $\mathbb{R}^d$, with density functions p and q, respectively. Let $K(x,y) = \frac{1}{(2\pi\sigma^2)}^{d/2} \exp -\frac{\|x-y\|^2}{2\sigma^2}$.

Under some regularity assumptions on p and q, if $n \to \infty$, $\sigma \to 0$, and $n\sigma^{d+2}/\log(n) \to \infty$, then

$$nR_{XY} \to \frac{2}{p(X) + q(X)} + \frac{2}{p(Y) + q(Y)} \text{ almost surely}$$

with equivalent statements holding for X_1, X_2 and Y_1, Y_2.

Central limit theorem for quadratic forms.

Theorem 3 (Chatterjee 08)**.** Let $a = (a_1, \dots, a_n)$ be i.i.d random variables with with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. For some fixed real valued symmetric matrix $M = (M_{ij})_{1 \le i,j \le n}$, define

$$W = a^T M a.$$

with μ denoting the law of $(W - EW)/\sqrt{\mathrm{Var}(W)}$.

Then, letting $\mathcal{G}$ be the standard Gaussian measure

$$\mathcal{W}(\mu, \mathcal{G}) \le \left(\frac{\mathrm{tr}(M^4)}{\mathrm{tr}(M^2)^2}\right)^{1/2} + \left(\frac{5 \max_i (M_{ii})^2}{\mathrm{tr}(M^2)}\right)^{1/2}. \tag{6}$$

Translating from Wasserstein to Kolmogorov distance.

Lemma 4 (Wasserstein to Kolmogorov distance)**.** For any probability measures μ, ν with corresponding cdfs F_μ and F_ν and any $\epsilon' > 0$, there exists some $\epsilon > 0$ such that

$$\mathcal{W}(\mu, \nu) < \epsilon \implies \sup_t |F_\mu(t) - F_\nu(t)| \le \epsilon'.$$

6 Proofs

Proof of Proposition 1. Throughout, we will use the fact that $a_i | Z_i \sim \mathrm{Rademacher}(\frac{p(Z_i)}{p(Z_i)+q(Z_i)})$, which is easily seen by an application of Bayes rule.

Begin by rewriting

$$a^T L a = (\mathbf{h_0} + a - \mathbf{h_0})^T L (\mathbf{h_0} + a - \mathbf{h_0}) := (\mathbf{h_0} + \epsilon)^T L (\mathbf{h_0} + \epsilon).$$

Expanding the quadratic form yields

$$a^T L a = \mathbf{h_0}^T L \mathbf{h_0} + \epsilon^T L \epsilon + 2 \mathbf{h_0}^T L \epsilon.$$

Going from back to front, we have that the first term has expectation 0, because

$$\mathbb{E}[L_{ij} h_0(Z_i) \epsilon_j] = \mathbb{E}[L_{ij} h_0(Z_i) \mathbb{E}[\epsilon_j | Z]] = \mathbb{E}[L_{ij} h_0(Z_i) 0] = 0.$$

For the middle term, only the diagonal terms have non-zero expectation.

$$\begin{aligned}\mathbb{E}\left[\epsilon^T L\epsilon\right] &= \sum_{i,j=1}^{N} \mathbb{E}\left[L_{ij}\mathbb{E}\left[\epsilon_j\epsilon_i|Z\right]\right] \\ &\overset{(i)}{=} \sum_{1\le i<j\le N} \mathbb{E}\left[L_{ij}\mathbb{E}\left[\epsilon_j|Z\right]\mathbb{E}\left[\epsilon_i|Z\right]\right] + \sum_{i=1}^{n} \mathbb{E}\left[L_{ii}^2\mathbb{E}\left[\epsilon_i^2|Z\right]\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[L_{ii}^2\mathbb{E}\left[\epsilon_i^2|Z\right]\right].\end{aligned}$$

where (i) follows from the conditional independence relation $a_i \perp\!\!\!\perp a_j|Z$.

Then

$$\mathbb{E}\left[\epsilon_i^2|Z\right] = \mathbb{E}\left[(a(Z_i) - h_0(Z_i))^2|Z\right] = \mathrm{Var}\left(a(Z_i)|Z_i\right) = \frac{4}{N^2}\left(\frac{4p(Z_i)q(Z_i)}{(p(Z_i)+q(Z_i))^2}\right)$$

and plugging this in, we have

$$\begin{aligned}\mathbb{E}\left[\epsilon^T L\epsilon\right] &= \frac{16}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[L_{ii}\left(\frac{p(Z_i)q(Z_i)}{(p(Z_i)+q(Z_i))^2}\right)\right] \\ &= \frac{16}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\sum_{j\neq i}K(\|Z_i - Z_j\|)\left(\frac{p(Z_i)q(Z_i)}{(p(Z_i)+q(Z_i))^2}\right)\right] \\ &= \frac{4N}{N-1}\int\int K(\|\mathbf{x}-\mathbf{y}\|)\left[p(\mathbf{x})q(\mathbf{x})\right]\frac{p(\mathbf{y})+q(\mathbf{y})}{p(\mathbf{x})+q(\mathbf{x})}d\mathbf{x}d\mathbf{y}\end{aligned}$$

Using the relation $\frac{(a+b)^2-(a-b)^2}{4} = ab$ yields the 1st and 2nd integrals of (2). The 3rd integral is exactly $\mathbb{E}\left[h_0^T L h_0\right]$.

□

Proof of Proposition 2. Write $\mathbf{h} = \frac{\mathbf{x}-\mathbf{y}}{t}$. Via Taylor expansion, we can write

$$\begin{aligned}&\int K\left(\frac{\|\mathbf{x}-\mathbf{y}\|}{t}\right)(p(\mathbf{y})+q(\mathbf{y}))d\mathbf{y} \\ &\overset{(i)}{=} \int K(\|\mathbf{h}\|)(p(\mathbf{x})+q(\mathbf{x})+\mathcal{O}(t\,\|h\|))t^d d\mathbf{h} \\ &\overset{(ii)}{=} (p(\mathbf{x})+q(\mathbf{x}))+\mathcal{O}(t^{d+1})\end{aligned}$$

where (i) follows from the Lipschitz continuity of p and q, and (ii) follows from the integrability condition on K.

Applying this to the 2nd and 3rd integrals of (2) yields the two integrals of (3). The 3rd integral is $\mathcal{O}(t^{d+1})$ by Lemma 6. □

Proof. First, we rewrite T_2, using the fact that $a = \frac{2}{N}\left(\sum_{i\in\mathcal{L}_X} e_i - \sum_{i\in\mathcal{L}_Y} e_i\right)$.

$$a^T L^\dagger a = \frac{4}{N^2}\left(\sum_{i,j\in\mathcal{L}_X} e_i L^\dagger e_j + \sum_{i,j\in\mathcal{L}_Y} e_i L^\dagger e_j - 2\sum_{i\in\mathcal{L}_X, j\in\mathcal{L}_Y} e_i L^\dagger e_j\right)$$

Via this expression, we see that in the above summations

- For $i = j$, $e_i^T L^\dagger e_i$ appears exactly once.
- For $i \neq j$ and $i, j \in \mathcal{L}_X$ or $i, j \in \mathcal{L}_Y$, $e_i^T L^\dagger e_j$ appears exactly twice.
- For $i \in \mathcal{L}_X$, $j \in \mathcal{L}_Y$, $-e_i^T L^\dagger e_j$ appears exactly twice.

Now, consider the expression

$$\sum_{u\in\mathcal{L}_X, v\in\mathcal{L}_Y} R_{uv} - \sum_{u<v\in\mathcal{L}_Y} R_{uv} - \sum_{u<v\in\mathcal{L}_X} R_{uv}.$$

Going from bottom to top, we have

- When $i \in \mathcal{L}_X$ and $j \in \mathcal{L}_Y$, R_{ij} will contribute $-2e_i L^\dagger e_j$. No other R_{uv} will contribute anything to this term.
- When $i < j \in \mathcal{L}_X$ or $i < j \in \mathcal{L}_Y$, the term R_{ij} in the 2nd or 3rd sum will appear exactly once and will contribute $2e_i L^\dagger e_j$. No other R_{uv} will contribute anything to this term.
- When $i = j \in \mathcal{L}_X$, $-R_{ik}$ will contribute $-e_i L^\dagger e_i$ for each $k \neq i \in \mathcal{L}_X$, and will contribute $e_i L^\dagger e_i$ for each $k \in \mathcal{L}_Y$. The total contribution will be $(|\mathcal{L}_Y| - |\mathcal{L}_X| + 1)(e_i L^\dagger e_i) = e_i L^\dagger e_i$. The same reasoning holds for $i = j \in \mathcal{L}_Y$.

All contributions from all R_{uv} can be put into one of the three proceeding categories. Therefore,

$$a^T L^\dagger a = \frac{4}{N^2}\left(\sum_{u\in\mathcal{L}_X, v\in\mathcal{L}_Y} R_{uv} - \sum_{u<v\in\mathcal{L}_Y} R_{uv} - \sum_{u<v\in\mathcal{L}_X} R_{uv}\right)$$

(4) follows from taking expectation and noting that X_i and X_j are identically distributed for all i and j.

□

Proof of Theorem 2. We will proceed by

1. Conditioning on the high-probability outcome that the Laplacian converges to a limiting object in the right sense.

2. Showing that, under such convergence of the Laplacian, both terms in Theorem 3 grow small with n.

3. Converting from Wasserstein distance to Kolmogorov distance.

Step 1. Fix $\epsilon > 0$. Throughout, let P_Z denote the distribution of Z, and likewise P_a denote the distribution of a.

For $V_n \sim \nu_n(\rho_n L^\dagger)$, and $V \sim \nu_\infty$ let

$$A_n = \left\{ z \in \mathbb{R}^n : |EV_n^p - EV^p| \leq \epsilon \text{ for } p = 1, 2, 4 \right\} \bigcup \left\{ z \in \mathbb{R}^n : \max_{i \in [n]} \frac{1}{n} \left(\{\rho_n L^\dagger\}^2 \right)_{ii} \leq \epsilon \right\}.$$

It is not hard to see that our Conjectures 1 and 2 imply A_n will eventually have high probability.

$$\begin{aligned}
\mathbb{P}(A_n) &\geq \mathbb{P}\left(\left\{ z \in \mathbb{R}^n : |EV_n^p - EV^p| \leq \epsilon \right\} \right) + \mathbb{P}\left(\left\{ z \in \mathbb{R}^n : \max_{i \in [n]} \frac{1}{n} \left(\{\rho_n L^\dagger\}^2 \right)_{ii} \leq \epsilon \right\} \right) \\
&\overset{(i)}{\geq} 1 - 2\epsilon \text{ for all } n \geq N. \qquad (7)
\end{aligned}$$

where (i) follows from Conjecture 2 (for the second term), and Conjecture 1 (for the first term).

Writing $W_n := W_n(z, a)$ to emphasize that it is a function of z and a, we have by Tonelli's theorem that

$$\begin{aligned}
\sup_t |\mathbb{P}(W_n \leq t) - \Phi(t)| &\overset{(i)}{=} \sup_t \left| \int_{\mathbb{R}^N} \left(\int_{\{-1,1\}^N} 1(W_n(z, a) \leq t) dP_a \right) dP_z - \Phi(t) \right| \\
&= \sup_t \left| \int_{\mathbb{R}^N} \left(\int_{\{-1,1\}^N} 1(W_n(z, a) \leq t) dP_a \right) - \Phi(t) dP_z \right| \\
&\leq \int_{\mathbb{R}^N} \sup_t \left| \left(\int_{\{-1,1\}^N} 1(W_n(z, a) \leq t) dP_a \right) - \Phi(t) \right| dP_z \\
&\overset{(ii)}{\leq} \int_{A_n} \sup_t \left| \left(\int_{\{-1,1\}^N} 1(W_n(z, a) \leq t) dP_a \right) - \Phi(t) \right| dP_z + 2\epsilon
\end{aligned} \qquad (8)$$

where (i) follows from Tonelli's theorem and (ii) from (7).

Step 2. Denote as

$$F_{a|z}(z, t) := \left(\int_{\{-1,1\}^N} 1(W_n(z, a) \leq t) dP_a \right)$$

and note that for any z this defines a measure over the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, which we will call $\mu_{a|Z}(z)$.

We wish to upper bound $\mathcal{W}(\mu_{a|Z}(z), \mathcal{G})$. To do so, we will compute upper bounds for each present in (6). For the first term, we have

$$\begin{aligned}\frac{\operatorname{tr}(\{L^\dagger\}^4)}{\operatorname{tr}(\{L^\dagger\}^2)^2} &= \frac{1}{n}\frac{\frac{1}{n}\operatorname{tr}(\rho_n^4\{L^\dagger\}^4)}{\frac{1}{n^2}\rho_n^4\operatorname{tr}(\{L^\dagger\}^2)^2} \\ &\le \frac{1}{n}\frac{\mathbb{E}\left[V^4\right]+\epsilon}{\mathbb{E}\left[V^2\right]^2-\epsilon}.\end{aligned}$$

For the second term, we have

$$\begin{aligned}\frac{\max_i(\{L^\dagger\}^2)_{ii}}{\operatorname{tr}(\{L^\dagger\}^2)} &= \frac{\frac{\rho_n^2}{n}(\{L^\dagger\}^2)_{ii}}{\frac{\rho_n^2}{n}\operatorname{tr}(\{L^\dagger\}^2)} \\ &\le \frac{\epsilon}{\mathbb{E}\left[V^2\right]-\epsilon}.\end{aligned}$$

By Theorem 3 we therefore have

$$\mathcal{W}(\mu_{a|Z}(z), \mathcal{G}) \le \frac{1}{n}\frac{\mathbb{E}\left[V^4\right]+\epsilon}{\mathbb{E}\left[V^2\right]^2-\epsilon} + \left(\frac{\epsilon}{\mathbb{E}\left[V^2\right]-\epsilon}\right)^{1/2}. \tag{9}$$

Step 3. Note that the right hand side of (9) converges to 0 with ϵ. Therefore, for any ϵ sufficiently small, by (9) and Lemma 4 we have

$$\left\|F_{Z|a} - \Phi\right\|_\infty \le \epsilon'.$$

Combined with (8) we have

$$\sup_t \left|\mathbb{P}\left(()\, W_n \le t\right) - \Phi(t)\right| \le 2\epsilon + \epsilon'.$$

for all $n \ge n_0$.

□

Proof of Lemma 1.

$$\begin{aligned}\mathbb{E}\left[T^2|Z\right] &= \mathbb{E}\left[a^T L^\dagger a|Z\right] \\ &\overset{(i)}{=} \sum_{i=1}^N\sum_{j=1}^N \mathbb{E}\left[a_i a_j\right] L^\dagger_{ij} \\ &= \sum_{i=1}^N \frac{1}{4N^2} L^\dagger_{ii} \\ &= \frac{\operatorname{tr}(L^\dagger)}{4N^2}.\end{aligned} \tag{10}$$

where (i) comes from the independence of Z and a under H_0. □

Proof of Lemma 2. First, we re-arrange T_2.

$$\begin{aligned}
T_2 &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j L^{\dagger}_{ij} \\
&= 2\sum_{i\le j} a_i a_j L^{\dagger}_{ij} - \frac{4}{N^2}\sum_{i=1}^{N} L^{\dagger}_{ii}.
\end{aligned}$$

Therefore, for $R_i \overset{i.i.d}{\sim}$ Rademacher(1/2),

$$\begin{aligned}
\mathrm{Var}\left(T_2|Z\right) &= 4\mathrm{Var}\left(\sum_{i\le j} a_i a_j L^{\dagger}_{ij}|Z\right) \\
&= \frac{64}{N^4}\mathrm{Var}\left(\sum_{i\le j} R_i R_j L^{\dagger}_{ij}|Z\right) \\
&= \frac{32}{N^4}\mathrm{tr}[(L^{\dagger})^2].
\end{aligned}$$

$\square$