Notes for the week 12/4 - 12/10

Alden Green

December 11, 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** $K : \mathcal{X} \times \mathcal{X} \to [0, \infty)$. 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 $N \times N^2$ **incidence matrix** of A, where the ith column of $B = B_i$ has 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

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

Holder condition We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is **α-Holder continuous** when

$$|f(x) - f(y)| \leq \|x - y\|^\alpha .$$

We will require this condition so that degrees in geometric graphs are well-behaved in the limit.

2 Desiderata

- Let K be a **uniform kernel of radius** ϵ, meaning

  $$K(x, y) = I(\|x - y\| \leq \epsilon).$$

  Assume P and Q have densities p and q with respect to Lebesgue measure. Say that for some $\alpha > 0$, p and q are α-holder continuous. For the graph G corresponding to the matrix A, with accompanying resistance distances, we wish to upper bound

  $$\left| N\epsilon^d \mathbb{E}\left[R_{XY}\right] - \mathbb{E}\left[\frac{2}{p(X)+q(X)} + \frac{2}{p(Y)+q(Y)}\right]\right|$$

3 Supplemental Results

Lemma 1 follows from an application of a discrete version of Poincare's inequality. See (von Luxburg 12) for proof and details.

Lemma 1. For some $\widetilde{N}_{\max}, \widetilde{N}_{\min}, d_{\max}, d_{\min}$, for all $i \neq j$

$$\left| R_{ij} - \left(\frac{1}{d_i} + \frac{1}{d_j}\right)\right| \leq 2a_1 \frac{1}{N\epsilon^{d+2}} \left(\frac{d_{\max}^2}{d_{\min}^3} \cdot (1 + 2\frac{\widetilde{N}_{\max}^2}{\widetilde{N}_{\min}^2})\right)$$

where $a_1 = \left(\frac{d\sqrt{d+3}}{L_{\min}}\right)^{d+1}$.

Lemmas **??**

Lemma 2. Denote

$$\mu_{\max} := N\epsilon^d \nu_d (p_{\max} + q_{\max})/2, \quad \mu_{\min} := N\epsilon^d \nu_d (p_{\min} + q_{\min})/2\beta$$

and let $a_2 = \left(\frac{L_{\min}}{L_{\max}}\right)^d \frac{\nu_d}{2^d (d+3)^{d/2}}$, $a_3 = \frac{\sqrt{d+1}}{L_{\min}^d}$.

For $\widetilde{N}_{\max}, \widetilde{N}_{\min}, d_{\max}, d_{\min}$ as in Lemma 1, the following bounds hold

$$\mathbb{P}\left(\widetilde{N}_{\max} \geq (1+z)\mu_{\max}\right) \leq \frac{a_3}{\epsilon^d} \cdot \exp(-z^2 \mu_{\max}/3)$$
$$\mathbb{P}\left(\widetilde{N}_{\min} \leq a_2(1-z)\mu_{\min}\right) \leq \frac{a_3}{\epsilon^d} \cdot \exp(-z^2 a_2 \mu_{\min}/3)$$
$$\mathbb{P}\left(d_{\max} \geq (1+z)\mu_{\max}\right) \leq n \cdot \exp(-z^2 \mu_{\max}/3)$$
$$\mathbb{P}\left(d_{\min} \leq (1-z)\mu_{\min}\right) \leq n \cdot \exp(-z^2 \mu_{\min}/3)$$

Lemma 3. For random variable X satisfying

$$\mathbb{P}\left(X \leq (1-z)\mu_n\right) \leq \exp(-z^2\mu_n/3 + \log n)$$

the inverse moment $\mathbb{E}\left[\frac{1}{(1+X)^k}\right]$, $k > 0$, satisfies for any $z < 1$

$$\mathbb{E}\left[\frac{1}{(1+X)^k}\right] \leq \exp(-z^2\mu_n/3 + \log n) + \frac{1}{(1+\mu_n(1-z))^k}$$

Similarly, for random variable Y satisfying

$$\mathbb{P}\left(Y \geq (1+z)\mu_n\right) \leq \exp(-z^2\mu_n/3 + c_n)$$

the moment $\mathbb{E}\left[(1+Y)^k\right]$, $k > 0$, satisfies for any $z > 0$

$$\mathbb{E}\left[(1+Y)^k\right] \leq \frac{2n}{}$$

4 Proofs

Begin by expanding

$$\begin{aligned}
&\left| N\epsilon^d \mathbb{E}\left[R_{XY}\right] - \mathbb{E}\left[\frac{2}{p(X)+q(X)} + \frac{2}{p(Y)+q(Y)}\right] \right| \\
&\quad = N\epsilon^d \left| \mathbb{E}\left[R_{XY}\right] - \mathbb{E}\left[\frac{1}{d(X)} + \frac{1}{d(Y)}\right] \right| \\
&\quad + N\epsilon^d \left| \mathbb{E}\left[\frac{1}{d(X)} - \frac{1}{N\mathbb{P}\left(B(X,\epsilon)\right)} + \frac{1}{d(Y)} - \frac{1}{N\mathbb{P}\left(B(Y,\epsilon)\right)}\right] \right| \\
&\quad + \left| \mathbb{E}\left[\frac{\epsilon^d}{\mathbb{P}\left(B(X,\epsilon)\right)} - \frac{2}{p(X)+q(X)}\right] + \mathbb{E}\left[\frac{\epsilon^d}{\mathbb{P}\left(B(Y,\epsilon)\right)} - \frac{2}{p(Y)+q(Y)}\right] \right|
\end{aligned} \tag{1}$$

We will bound the summands on the right side of (1) from last to first.

Third term. For the last term, we begin by rewriting

$$\left|\frac{\epsilon^d}{\mathbb{P}\left(B(X,\epsilon)\right)} - \frac{2}{p(X)+q(X)}\right| \le \left|\frac{\epsilon^d(p(X)+q(X)) - 2\mathbb{P}\left(B(X,\epsilon)\right)}{\mathbb{P}\left(B(X,\epsilon)\right)[p(X)+q(X)]}\right|$$

Then, we can bound the numerator using the fact we have required the densities p and q be Holder continuous, so

$$\begin{aligned}
[p(X)+q(X)]\epsilon^d - 2\mathbb{P}\left(B(X,\epsilon)\right) &= \int_{B(X,\epsilon)} [p(\mathbf{x}) - p(\mathbf{z})]d\mathbf{z} + \int_{B(X,\epsilon)} [q(\mathbf{x}) - q(\mathbf{z})]d\mathbf{z} \\
&\le \int_{B(X,\epsilon)} 2\left\|x-y\right\|^{\alpha} d\mathbf{z} \\
&\le 2\epsilon^{\alpha+d}.
\end{aligned}$$

We can lower bound the denominator using the lower bound on our densities

$$\mathbb{P}\left(B(X,\epsilon)\right)[p(X)+q(X)] \ge \epsilon^d(p_{\min}+q_{\min})^2/2$$

and therefore

$$\frac{\epsilon^d}{\mathbb{P}\left(B(X,\epsilon)\right)} - \frac{2}{p(X)+q(X)} \le \frac{4\epsilon^{\alpha}}{(p_{\min}+q_{\min})^2}.$$

The same bound holds for the corresponding term with Y instead of X.

Second term. To bound the second term, we will upper and lower bound $\mathbb{E}\left[\frac{1}{d(X)}\right]$ by something close to $\mathbb{E}\left[\frac{1}{N\mathbb{P}(B(X,\epsilon))}\right]$.

The lower bound

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{d(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{d(X)}\middle|X\right]\right] \\
&\ge \mathbb{E}\left[\frac{1}{1+(N-1)\mathbb{P}\left(B(X,\epsilon)\right)}\right]
\end{aligned}$$

follows from Jensen's inequality.

For the upper bound, note that the distribution of $d(X)$, conditional on X, is

$1 + \text{Binomial}(N-1, \mathbb{P}(B(X,\epsilon)))$. Then, letting $q = \mathbb{P}(B(X,\epsilon))$

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{d(X)}\middle| X\right] &= \sum_{k=0}^{N-1} \frac{1}{k+1} \binom{N-1}{k} q^k (1-q)^{N-1-k} \\
&= \frac{1}{Nq} \sum_{k=0}^{N-1} \binom{N-1}{k+1} q^{k+1} (1-q)^{N-1-k} \\
&\leq \frac{1}{Nq} \sum_{k=0}^{N} \binom{N}{k} q^k (1-q)^{N-k} \\
&= \frac{1}{Nq} \left(q + (1-q)\right)^N = \frac{1}{Nq}.
\end{aligned}$$

Combining this with the above, we have

$$\begin{aligned}
N\epsilon^d \left| \mathbb{E}\left[\frac{1}{d(X)} - \frac{1}{N\mathbb{P}(B(X,\epsilon))}\right]\right| &\leq N\epsilon^d \left| \mathbb{E}\left[\frac{1}{1+(N-1)\mathbb{P}(B(X,\epsilon))}\right] - \mathbb{E}\left[\frac{1}{N\mathbb{P}(()\, B(X,\epsilon))}\right]\right| \\
&\leq N\epsilon^d \left| \mathbb{E}\left[\frac{1}{N^2\mathbb{P}(B(X,\epsilon))^2}\right]\right|.
\end{aligned}$$

with a corresponding bound holding for Y.

First term. We begin by reducing the first term to a product of moments and inverse moments of maxima and minima of binomials.

$$\begin{aligned}
N\epsilon^d \left| \mathbb{E}[R_{XY}] - \mathbb{E}\left[\frac{1}{d(X)} + \frac{1}{d(Y)}\right]\right| &\overset{(i)}{\leq} \frac{2a_1}{\epsilon^2} \mathbb{E}\left[\frac{d_{\max}^2}{d_{\min}^3} \cdot (1 + 2\frac{\widetilde{N}_{\max}}{\widetilde{N}_{\min}})\right] \\
&\overset{(ii)}{\leq} \frac{2a_1}{\epsilon^2} \left(2\mathbb{E}\left[d_{\max}^8\right] \cdot \mathbb{E}\left[d_{\min}^{12}\right] \cdot \mathbb{E}\left[\widetilde{N}_{\max}^8\right] \cdot \mathbb{E}\left[\frac{1}{\widetilde{N}_{\min}^8}\right]\right)^{1/4} \\
&\quad + \frac{2a_1}{\epsilon^2} \left(\mathbb{E}\left[d_{\max}^4\right] \cdot \mathbb{E}\left[d_{\min}^6\right]\right)^{1/4}
\end{aligned}$$

where (i) follows from Lemma 1 and (ii) from repeated applications of Holder's inequality.