Thesis Proposal

Alden Green

November 17, 2019

Abstract

In the classical data analysis setting, where we observe i.i.d. point-cloud data sampled from an unknown distribution P, there has been recent interest in the design and analysis of procedures which involve the geometry of P. Broadly speaking, these procedures either estimate geometric properties such as high-density regions of P, or leverage the geometry of P in some downstream statistical learning task. One way to implicitly adapt to the geometry of P is through the use of graph-based spectral methods. The underlying intuition is that when one appropriately constructs a neighborhood graph, the resulting graph Laplacian encodes a great deal of information about the geometry of P; for example, one can view its top eigenvectors as (discrete versions of) functions which are smooth with respect to P. For this reason, there has been extensive analysis on consistency properties of graph Laplacians, that is, in what ways and with what error they approximate continuum-level operators.

In this thesis we will take a different approach, and analyze the effectiveness with which spectral algorithms perform traditional statistical learning tasks such as density clustering and hypothesis testing. In the density clustering problem, we consider a local version of spectral clustering using the Personalized PageRank (PPR) algorithm. We show that when a density cluster satisfies a natural set of geometric conditions PPR will provably recover it with high probability; we also derive lower bounds conversely showing that PPR will fail to recover geometrically poorly conditioned density clusters. In the hypothesis testing context, in some preliminary work we define a goodness-of-fit test using spectral-based test statistics formed over a neighborhood graph, and prove that it achieves minimax optimal testing rates against a certain class of smooth alternatives. We propose various extensions of this test targeted for wider classes of alternatives, and for the two-sample problem.

1 Background.

TODO: This is why you are going to read.

TODO: This is what you are going to read.

1.1 Neighborhood graphs and their spectra.

Let $X = \{x_1, \dots, x_n\}$ be a sample drawn i.i.d. from a distribution $\mathbb{P}$ on $\mathbb{R}^d$, with density f. For a radius $r > 0$, we define $G_{n,r} = (V, E)$ to be the *r-neighborhood graph* of X, an unweighted, undirected graph with vertices $V = X$, and an edge $(x_i, x_j) \in E$ if and only if $\|x_i - x_j\| \leq r$, where $\|\cdot\|$ is the Euclidean norm. We denote by $A \in \mathbb{R}^{n \times n}$ the adjacency matrix, with entries $A_{uv} = 1$ if $(u, v) \in E$ and 0 otherwise. We also denote by D the diagonal degree matrix, with entries $D_{uu} := \sum_{v \in V} A_{uv}$, and by I the $n \times n$ identity matrix.

Spectral graph theory involves the study of the eigenvalues and eigenvectors of any of following Laplacian matrices:

$$\text{combinatorial: } L = D - A, \quad \text{random walk: } L_{\text{rw}} = I - D^{-1}A, \quad \text{normalized: } L_{\text{n}} = D^{1/2} L_{\text{rw}} D^{-1/2}.$$

and is broadly applicable to fields as diverse as the theory of randomized algorithms and differential geometry Chung. When specifically considering Laplacians of neighborhood graphs such as $G_{n,r}$, one well-developed line of work concerns the asymptotic convergence properties of the various graph Laplacians to associated continuum Laplacian operators.

- Define the Laplacian operator.
- Introduce the study of Laplacians on point-cloud data, manifold learning, and give references as to their background. Discuss the study of the convergence of spectra Belkin, Coifman, Garcia-Trillos and Slepcev 1, Garcia-Trillos and Slepcev Spec, von Luxburg, Hein + von Luxburg, Lafon,Koltchinskii and Gine, Gine and Koltchinskii 1, pointwise convergence of operators Belkin thesis, Lafon thesis, Hein + Audibert, Singer, Belkin + Niyogi08, and convergence of smoothness functionals Bousquet + Chapelle, Hein07.
- Discuss the study of other functionals, such as total variation and normalized cut, on graphs (Garcia-Trillos and Slepcev TV, Arias-Castro, Maier + von Luxburg)
-

Why should I care about the asymptotic convergence of graph Laplacians?

- The convergence of graph Laplacians, and their spectra, to underlying continuum operators suggests they may be used to perform a variety of desirable tasks in statistical learning.
- Dimension reduction: under the hypothesis that the data lie on, or are concentrated near, a manifold $\mathcal{M} \subset \mathbb{R}^d$ of dimension $m < d$, the eigenvectors $v_1, \ldots, v_m$, can be used to embed $x_i \mapsto (v_{i,1}, \ldots v_{i,m})$, reducing the dimension from d to m without losing the geometric structure of $\mathbb{P}$. This algorithm is known as Laplacian Eigenmaps (Belkin).
- Clustering: under the hypothesis that the distribution $\mathbb{P}$ is supported on disjoint connected components $\mathcal{C}_1, \ldots, \mathcal{C}_k$, the span of the smallest k eigenvectors of the Laplacian operator L_n will approximately equal the span of the functions $\phi_j(x) = \mathbf{1}\{x \in \mathcal{C}_j\}, j = 1, \ldots, k$. (Belkin + Arias-Castro)
- Semi-supervised and supervised learning: Discuss the study of Laplacians in supervised learning (Bousquet + Chapelle, Lee + Izbicki, GarciaTrillos + Murray, Padilla) Discuss the study of Laplacians in semi-supervised learning (Belkin + Niyogi, Zhou + Belkin, Slepcev and Thorpe).

In this thesis proposal, we pick up a few of these threads, showing how spectral methods can be used to perform the statistical tasks of density clustering and nonparametric hypothesis testing. Before doing so, we provide some background on each of these areas.

1.2 Clustering.

Clustering involves splitting a given data set into groups that satisfy some notion of within-group similarity and between-group difference. There are many ways to define different types of clusters; we review a few that are especially related to our work.

1.2.1 Density clustering.

Let $X = \{x_1, \ldots, x_n\}$ consist of i.i.d. samples drawn from a distribution $\mathbb{P}$ on $\mathbb{R}^d$, with density f. The *density clusters* of f, defined as the connected components of the upper level set $\{x \in \mathbb{R}^d : f(x) \geq \lambda\}$ for some $\lambda > 0$ are a central object of interest in the statistical clustering literature, dating back to the work of Hartigan [1981], who shows that the classical single-linkage hierarchical clustering algorithm is guaranteed to recover density clusters only when $d = 1$. Polonik [1995], Rigollet and Vert [2009] study density clustering under the symmetric set difference metric, Tsybakov [1997], Singh et al. [2009] describe minimax optimal level-set estimators under Hausdorff loss and Hartigan [1981], Chaudhuri and Dasgupta [2010], Balakrishnan et al. [2013], Kpotufe and von Luxburg [2011] consider consistent estimation of the cluster tree. None of

these works, however, study the ability of spectral algorithms to recover density clusters, which will be our focus.

1.2.2 Nonparametric mixture model.

The results in Section 1.1 detailing with spectral convergence of the graph Laplacian operator have straightforward consequences for the consistency of spectral clustering (for instance, see von Luxburg et al. [2008], Trillos and Slepcev [2018]). However, the behavior of the spectra of these continuum operators can in general be hard to grasp. Therefore, further work relating this spectra to the geometry of the distribution $\mathbb{P}$ is of interest. In this spirit, Shi et al. [2009], Schiebinger et al. [2015], Trillos et al. [2019], Arias-Castro [2011] examine the ability of spectral algorithms to recover the latent labels in certain geometrically well-conditioned nonparametric mixture models $\mathbb{P} = \sum_{i=1}^{k} \pi_i \mathbb{P}_i$. These results focus on global rather than local methods, and thus impose global rather than local conditions on the nature of the density. Moreover, they do not in general guarantee recovery of density clusters, which is the focus in our work. Perhaps most importantly, these works rely on general cluster saliency conditions, which implicitly depend on many distinct geometric aspects of the cluster $\mathcal{C}$ under consideration. We will make this dependence more explicit, and in doing so expose the role each geometric condition plays in the clustering problem.

1.2.3 Worst-case analysis.

An altogether different approach is to treat the graph $G = (V, E)$ as fixed, rather than as a random geometric graph formed over data as is the case when $G = G_{n,r}$. Now, the quality of the clustering algorithm necessarily depends on the extent to which G contains natural clusters. To quantify the clusterability of a given subset $S \subseteq V$, let the normalized cut $\Phi(S; G)$ be given by

$$\Phi(S; G) = \frac{\text{cut}(S; G)}{\min\{\text{vol}(S; G), \text{vol}(S^c; G)\}}, \quad \text{cut}(S; G) := \sum_{u \in S} \sum_{w \in S^c} A_{uv}, \quad \text{vol}(S; G) = \sum_{u \in S} D_{uu}.$$

and the conductance $\Psi(S; G)$ be

$$\Psi(S; G) = \min_{A \subset S} \Phi(A; G[S])$$

where $G[S] = (S, E_S), E_S = E \cap (S \times S)$ is the subgraph of G induced by S. The normalized cut and conductance functionals measure the external connectivity of S to the rest of G, and the internal connectivity of $G[S]$, respectively. Kannan et al. [2004] assume there exists a partition $S_1 \cup \ldots \cup S_k$ such that $\Phi(S_j) \leq \phi$ and $\Psi(S_j) \geq \psi$ for all j. They analyze an algorithm which recursively splits the graph G based on its eigenvectors, resulting in clusters $C_1 \cup \ldots \cup C_k = V$, and prove upper bounds on the normalized cuts $\Phi(C_j)$ and conductances $\Psi(C_j)$ in terms of ϕ and ψ. Spielman and Teng [2013], Andersen et al. [2006] consider the local clustering problem, where assumptions are made only with respect to some small subset $S \subseteq V$. They each consider random walk based algorithms which cluster using locally-biased spectra, and derive upper bounds on the normalized cut of the outputted cluster $\Phi(C)$ in terms of $\Phi(S)$. Zhu et al. [2013] extends this analysis, showing that improvements in the quality of a cluster estimate are possible when $\Psi(S) > \Phi(S)$. We rely on the analysis of Zhu et al. [2013] in our work; however, our setting is quite different, as we work with respect to the random graph $G_{n,r}$ rather than a fixed G, and seek to derive bounds which depend on the distribution $\mathbb{P}$ rather than on Φ and Ψ, which in our context are random functionals.

1.3 Nonparametric hypothesis testing.

In hypothesis testing, the goal is distinguish whether our data are generated from some null model – generally a simple model displaying a lack of meaningful structure – or an alternative model. *Regression* and *density* testing are two related but distinct hypothesis testing problems. We will formally introduce each problem and provide some relevant background. Additionally, we will cover some common classes of test statistics used for nonparametric hypothesis testing.

1.3.1 Regression testing.

In addition to observing i.i.d samples $x_1, \ldots, x_n$ from a distribution $\mathbb{P}$ with support $\mathcal{X} := [0,1]^d \subset \mathbb{R}^d$[1], suppose we observe $Y = \{y_1, \ldots, y_n\}$ according to the following regression model:

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1) \tag{1}$$

We wish to distinguish

$$\mathbf{H}_0 : f = f_0 := 0 \quad \text{vs} \quad \mathbf{H}_\text{a} : f \neq f_0$$

We will evaluate our performance using worst-case risk: for a given function class $\mathcal{H}$ and test function $\phi : \mathbb{R}^n \to \{0,1\}$, let

$$\mathcal{R}(\phi; \mathcal{H}) = \mathbb{E}_{f=f_0}(\phi) + \sup_{f \in \mathcal{H}, f \neq f_0} \mathbb{E}_f(1-\phi).$$

Without further assumptions on $\mathcal{H}$ no test achieves meaningful power uniformly over all $f \in \mathcal{H}$ Janssen [2000]; to make the hypothesis testing question well-posed we will therefore restrict the function class $\mathcal{H}$ in two standard ways. One, when considering type II error we will remove a radius-ϵ ball in the L_2 norm around f_0 and consider only the remaining functions

$$\mathcal{H}_\epsilon := \{f \in \mathcal{H} : \|f - f_0\|_2 > \epsilon\}.$$

Otherwise we can always pick a function $f \neq f_0$ in $\mathcal{H}$ which is indistinguishable from f_0 given a finite number of samples. Two, we will insist that the functions in $\mathcal{H}$ display some degree of regularity or smoothness; otherwise it may not be possible to consistently estimate the functional $\|f - f_0\|_2$.

Let $\alpha \in (0,1)$ be a user input specifying a tolerated level of testing error. A typical way to characterize the performance of a test ϕ is through the critical radius $\epsilon_n(\phi; \mathcal{H})$, defined as

$$\epsilon_n(\phi; \mathcal{H}) = \inf\{\epsilon : \mathcal{R}(\phi; \mathcal{H}_\epsilon) \leq \alpha\}.$$

We can then measure the difficulty of a given problem through the minimax critical radius $\epsilon_n^\star$, formally

$$\epsilon_n^\star(\mathcal{H}) = \inf\left\{\epsilon : \inf_\phi \mathcal{R}(\phi; \mathcal{H}_\epsilon) \leq \alpha\right\}.$$

where the infimum is over all Borel measurable functions $\phi : \mathbb{R}^n \to \{0,1\}$. We say a test ϕ is minimax optimal over a function class $\mathcal{H}$ if $\epsilon_n(\phi; \mathcal{H}) \asymp \epsilon_n^\star(\mathcal{H})$, and we call the rate at which $\epsilon_n^\star(\mathcal{H}) \to 0$ as a function of n the *minimax testing rate* over $\mathcal{H}$. To concisely state rates, we will use order notation, in which we write $a_n = O(b_n)$ when $a_n \leq c b_n$ for some constant c and all $n \in \mathbb{N}$, $a_n = o(b_n)$ when $a_n / b_n \to 0$ as $n \to \infty$, and $a_n \asymp b_n$ when $c b_n \leq a_n \leq C b_n$ for constants c and C and all $n \in \mathbb{N}$.

Testing over Sobolev classes. The first type of function class $\mathcal{H}$ we consider will be balls in Sobolev spaces. We say a Borel measurable function $f : \mathcal{X} \to \mathbb{R}$ is in the space $\mathcal{L}^p(\mathcal{X})$ for $1 \leq p < \infty$ if $\int_\mathcal{X} |f(x)|^p \, dx < \infty$. Then, the Sobolev space $W_d^{s,2}(\mathcal{X})$ consists of all functions $f \in \mathcal{L}^1(\mathcal{X})$ such that for each $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $|\alpha| := \sum_{i=1}^d \alpha_i \leq s$, the weak derivative $D^\alpha f$(See Evans [2010] for a formal definition) belongs to $\mathcal{L}^p(\mathcal{X})$. The Sobolev $\{s, 2\}$ norm is then

$$\|f\|^2_{W_d^{s,2}(\mathcal{X})} = \sum_{|\alpha| \leq s} \int_\mathcal{X} |D^\alpha f|^2 \, dx$$

and for a given $L > 0$, the corresponding ball is $W_d^{s,2}(\mathcal{X}; L) = \left\{f : \|f\|_{W^{s,2}(\mathcal{X})} \leq L\right\}$. In this work, we will treat L as a fixed, known constant.

[1]For the purposes of this proposal, whenever we discuss hypothesis testing we will assume $\mathcal{X} = [0,1]^d$, i.e. the unit cube with dimension equal to that of the ambient space $\mathbb{R}^d$.

Minimax testing rates for the Sobolev regression testing problem in one dimension were first derived by Ingster [1982] for the idealized *white noise* regression model:

$$dY(x) = f(x)\,dx + \frac{\sigma}{\sqrt{n}}\,dW(x) \tag{2}$$

where $dW(x)$ denotes a Gaussian white noise process on $\mathcal{X} \subset \mathbb{R}$ (defined formally in Giné and Nickl [2016] Section 1.2) and $\sigma > 0$ is a known constant. In this case, the minimax test rate is $\epsilon_n^\star(\mathcal{W}_d^{s,2}(\mathcal{X}, L)) \asymp n^{-2s/(4s+1)}$ for all values of $s > 0$. Ingster and Suslina [2005] extend this to the multivariate setting where $\mathcal{X} \subseteq \mathbb{R}^d$, and show that $\epsilon_n^\star(\mathcal{W}_d^{s,2}(\mathcal{X}, L)) \asymp n^{-2s/(4s+d)}$, again for all $d \geq 1, s > 0$. The case where we observe data according to the generative model (1) was treated in Ingster and Sapatinas [2009], wherein it is shown that the minimax rate remains the same $\epsilon_n^\star(\mathcal{W}_d^{s,2}(\mathcal{X}, L)) \asymp n^{-2s/(4s+d)}$ so long as $4s > d$. When $4s < d$, if we replace the assumption $f \in W_d^{s,2}(\mathcal{X}, L)$ by the assumption $\left(\int_{\mathcal{X}} f^4(x)\,dx\right)^{1/4}$, then Guerre and Lavergne [2002] show that a test ϕ based on the statistic $\frac{1}{n}\sum_{i=1}^n y_i^2$ will have testing error on the order of $n^{-1/4}$) . Since $\mathcal{W}^{s,2}$ does not compactly embed into $\mathcal{L}^4$ when $4s < d$ this result does not apply to Sobolev spaces in this regime; indeed it seems likely that uniformly consistent testing over these spaces is not possible, as the function f^2 may no longer have finite variance.

1.3.2 Density testing.

In the two-sample density testing problem, we observe independent samples $Z = z_1, \ldots, z_N \sim \mathbb{P}$ and $Y = y_1, \ldots, y_M \sim \mathbb{Q}$, where $\mathbb{P}$ and $\mathbb{Q}$ are distributions over $\mathbb{R}^d$ with densities p and q, respectively. Our goal is to distinguish the hypotheses

$$\mathbf{H}_0 : \mathbb{P} = \mathbb{Q} \quad \text{vs} \quad \mathbf{H}_\text{a} : \mathbb{P} \neq \mathbb{Q}$$

and we again evaluate our performance using worst-case risk; letting $\phi : \mathbb{R}^{N+M} \to \{0, 1\}$,

$$\mathcal{R}(\phi; \mathcal{H}) = \inf_{p \in \mathcal{H}} \mathbb{E}_{p,p}(\phi) + \sup_{p \neq q, p, q \in \mathcal{H}} \mathbb{E}_{p,q}(1 - \phi),$$

and the critical radius ϵ_n. (Although we overload notation by having $\mathcal{R}$ and ϵ_n refer to the worst-case risk and critical radius in both the regression and density testing problems, it will always be clear from context which problem we are referring to.)

Density testing in Holder spaces. Unlike in the regression testing context, where f is assumed to belong to a Sobolev space, most of the study of density testing seems to have dealt with the case where $\mathcal{H}$ is Holder space. For a given $s > 0$, the sth Holder norm is given by

$$\|f\|_{C_d^s(\mathcal{X})} := \sum_{|\alpha| \leq s} \|D^\alpha f\|_\infty + \sum_{|\alpha| = s} \sup_{x, y \in \mathcal{X}} \frac{|D^\alpha f(y) - D^\alpha f(x)|}{\|x - y\|_2}$$

and the sth Holder space $C_d^s(\mathcal{X})$ consists of all functions which are s times continuously differentiable with finite s Holder norm. Denote the Holder unit ball by $C_d^s(\mathcal{X}, L) = \left\{ f \in C_d^s(\mathcal{X}) : \|f\|_{C_d^s(\mathcal{X})} \leq L \right\}$.

Ingster [1987] considers the 1d version of the problem and shows that the minimax critical radius for the Holder balls $C_d^s(\mathcal{X}; L)$ is $\epsilon_n^\star(C^s(\mathcal{X}; L)) \asymp n^{-2s/(4s+1)}$. This means that the Holder space $C_d^s(\mathcal{X})$ and the smallest Sobolev space it compactly embeds into, $W_d^{s,2}(\mathcal{X})$, have the same minimax testing rates in the density and regression testing problems, respectively. Arias-Castro et al. [2018] examines the multivariate version of the problem and derives the rate $\epsilon_n^\star(C_d^s(\mathcal{X}; L)) \asymp n^{-2s/(4s+d)}$. In both cases, the stated rates hold for all $s > 0$, and in the second case for all $d \geq 1$.

Using asymptotic equivalence results Reiß [2008], Nussbaum [1996], Brown and Low [1996] between the density and regression setups, and sampling and Gaussian white noise models, we can infer some results about minimax rates over Sobolev classes in the former from those in the latter. However, to the best of our knowledge, the density testing problem over Sobolev spaces has not been directly studied.

1.3.3 Background on nonparametric test statistics.

The χ^2-test is a classical two-sample test, in which the domain $\mathcal{X}$ is partitioned into bins $B_1, \ldots, B_\kappa$ where $\kappa \in \mathbb{N}$ is a tuning parameter. Letting $\mathbb{P}_N$ be the empirical distribution of Z, and likewise for $\mathbb{Q}_M$ and Y, the resulting statistic can be expressed as

$$T_{\chi^2} := \sum_{k=1}^{\kappa} \left(\frac{\mathbb{P}_N(B_k) - \mathbb{Q}_M(B_k)}{\mathbb{P}_N(B_k) + \mathbb{Q}_M(B_k)} \right)^2. \tag{3}$$

The upper bounds in the density testing problem discussed in the previous section are derived using an unnormalized version of T_{χ^2}. In addition to the χ^2-test, there are several classes of test statistics used to perform nonparametric hypothesis testing which are worthy of our attention.

Edge-counting. In the two-sample density testing problem, Schilling [1986], Bhattacharya [2015] analyze edge-counting statistics, where a kNN graph $G_{n,k}$ is formed over all the samples $X = (z_1, \ldots, z_N, y_1, \ldots, y_M)$ (where $n = M + N$ is the total number of samples) and the statistic considered is then

$$T_{\text{edge}} := \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbf{1}(z_i \sim y_j \text{ in } G_{n,k}). \tag{4}$$

While there are some similarities between T_{edge} and the statistic T_{spec} we propose and analyze later in Section 3.1, they are quite distinct. For one, they depend on different constructions of the neighborhood graph; perhaps more fundamentally, they use the spectrum of the resulting neighborhood graphs in different ways. Preliminary results (see Figure 3.2) show that these differences in construction lead to empirically different behavior over a range of testing problems.

Truncated-series. In the regression testing problem, Ingster and Sapatinas [2009] upper bound the minimax testing rate by considering a truncated series test statistic, where the data is projected onto a subspace spanned by (evaluations of) low-frequency harmonics, and the test statistic is the empirical norm of this projection. Let $\{\phi_j\}$ be the standard Fourier basis in $L_2([0,1])$, meaning

$$\phi_0(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(2\pi j u) \text{ for each } j \in \mathbb{Z}_-, \quad \phi_j(x) = -\sqrt{2}\sin(2\pi j u) \text{ for each } j \in \mathbb{Z}_+ :$$

The test statistic which Ingster and Sapatinas [2009] analyze is then

$$T_{\text{Harm}} := \frac{1}{n} \sum_{\ell, -\ell \in [\kappa]^d} \left(\sum_{i=1}^{n} \phi_\ell(x_i) y_i \right)^2, \quad \text{where } \phi_\ell(x) = \prod_{i=1}^{d} \phi_{\ell_i}(x_i) \tag{5}$$

for $\ell \in \mathbb{Z}^d$, $x \in \mathbb{R}^d$, and $\kappa \in \mathbb{N}$ a tuning parameter. The test statistic T_{spec} which we propose and analyze in Section 3.1 is closely related to T_{Harm}, the crucial difference being that instead of the Fourier basis $\{\phi_\ell\}$, which are eigenfunctions of the Laplacian operator over $[0,1]^d$, we use eigenvectors $\{v_k\}$ of the graph Laplacian matrix.

Integral Probability Metrics (IPMs). Much recent attention Müller [1997], Sriperumbudur et al. [2009], Gretton et al. [2012], Mroueh et al. [2017] has been paid to IPMs, which are statistics of the form

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{Q}_n f|$$

used to perform two-sample density testing. (Here, $\mathcal{F}$ is appropriately chosen set of functions, for example $\mathcal{F}$ might be the unit ball in a Holder or Sobolev space). Of course, for any graph $G = (V, E)$, the corresponding

incidence matrix B induces a notion of smoothness for vectors $\theta \in \mathbb{R}^{|V|}$. Formally for some $s, p \in \mathbb{N}$, let the smoothness functional $S_{s,p}(\theta) = \|B^{(s)}\theta\|_p$ where

$$B^{(s)} = \begin{cases} L^{s/2}, & \text{when } s \text{ is even} \\ BL^{(s-1)/2}, & \text{when } s \text{ is odd.} \end{cases}$$

Then the classes $\Theta_{s,p} := \{\theta \in \mathbb{R}^n : S_{s,p}(\theta) \leq L\}$ can be viewed as possessing various types of smoothness, analogous to function classes such as Sobolev or Holder classes. It is therefore natural to consider IPMs of the form

$$\sup_{\theta \in \Theta_{s,p}} |\mathbb{P}_N \theta - \mathbb{Q}_M \theta| \,;$$

and we will propose a pair of such statistics in Section 3.3.

2 Density clustering with PPR.

In our density clustering work, we examine the ability of Personalized PageRank (PPR) – a local spectral algorithm– to recover density clusters. First, we motivate and formally define the PPR clustering algorithm we subsequently analyze.

2.1 PPR clustering.

When applied to geometric graphs built from a large number of samples, global spectral clustering methods can be computationally cumbersome and insensitive to the local geometry of the underlying distribution [Leskovec et al., 2010, Mahoney et al., 2012]. This has led to increased interest in *local* spectral clustering algorithms, which leverage locally-biased spectra computed using random walks around some user-specified seed node. A popular local clustering algorithm is the Personalized PageRank (PPR) algorithm, first introduced by Haveliwala [2003], then further developed by several others [Spielman and Teng, 2011, 2014, Andersen et al., 2006, Mahoney et al., 2012, Zhu et al., 2013].

The PPR vector $p_v = p(v, \alpha; G_{n,r})$, based on a seed node $v \in V$ and a teleportation parameter $\alpha \in [0, 1]$, is defined to be the solution of the following linear system:

$$p_v = \alpha e_v + (1 - \alpha) p W, \tag{6}$$

where $W = \frac{1}{2}(I + D^{-1}A)$ is the lazy random walk matrix over $G_{n,r}$ and e_v is the indicator vector for node v (that has a 1 in the v-th position and 0 elsewhere).

2.1.1 Sweep cut.

We employ the sweep cut method to obtain a local cluster from the PPR embedding p_v. In more detail, for a level $\beta > 0$, we define a *β-sweep cut* of $p_v = (p_v(u))_{u \in V}$ as

$$S_{\beta,v} := \left\{ u \in V : \frac{p_v(u)}{D_{uu}} > \beta \right\}. \tag{7}$$

We will use the normalized cut metric to determine which sweep cut S_β is the best cluster estimate. For a set $S \subseteq V$ with complement $S^c = V \backslash S$, we define $\mathrm{cut}(S; G_{n,r}) := \sum_{u \in S, v \in S^c} A_{uv}$, and $\mathrm{vol}(S; G_{n,r}) := \sum_{u \in S} D_{uu}$. We define the *normalized cut* of S as

$$\Phi(S; G_{n,r}) := \frac{\mathrm{cut}(S; G_{n,r})}{\min\{\mathrm{vol}(S; G_{n,r}), \mathrm{vol}(S^c; G_{n,r})\}}. \tag{8}$$

Having computed sweep cuts S_β over a range $\beta \in (L, U)$ (where the range (L, U) is also a user-specified parameter), we output the cluster estimate $\widehat{C} = S_{\beta^*}$ that has minimum normalized cut. For concreteness, the PPR algorithm is summarized in Algorithm 1.

Algorithm 1 PPR on a neighborhood graph

Input: data $X = \{x_1, \ldots, x_n\}$, radius $r > 0$, teleportation parameter $\alpha \in [0, 1]$, seed $v \in X$, sweep cut range (L, U).

Output: cluster $\widehat{C} \subseteq V$.

1: Form the neighborhood graph $G_{n,r}$.
2: Compute the PPR vector $p_v = p(v, \alpha; G_{n,r})$ as in (6).
3: For $\beta \in (L, U)$ compute sweep cuts S_β as in (7).
4: Return the cluster $\widehat{C} = S_{\beta^*}$, where

$$\beta^* = \operatorname*{argmin}_{\beta \in (L,U)} \Phi(S_\beta; G_{n,r}).$$

2.2 Estimation of density clusters

Let $\mathbb{C}_f(\lambda)$ denote the connected components of the density upper level set $\{x \in \mathbb{R}^d : f(x) > \lambda\}$. For a given density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we call $\mathcal{C}[X] = \mathcal{C} \cap X$ the *empirical density cluster*. The size of the symmetric set difference between estimated and empirical cluster is a commonly used metric to quantify cluster estimation error [Korostelev and Tsybakov, 1993, Polonik, 1995, Rigollet and Vert, 2009]. We will consider a related metric, the volume of the symmetric set difference, which weights points according to their degree in $G_{n,r}$.

Definition 2.1. For an estimator $\widehat{C} \subseteq X$ and a set $\mathcal{S} \subseteq \mathbb{R}^d$, the symmetric set difference is

$$\widehat{C} \triangle \mathcal{S}[X] := \big(\widehat{C} \setminus \mathcal{S}[X]\big) \cup \big(\mathcal{S}[X] \setminus \widehat{C}\big).$$

We will measure the discrepancy between $\widehat{C}$ and $\mathcal{S}[X]$ by the volume of the symmetric set difference,

$$\Delta(\widehat{C}, \mathcal{S}[X]) := \mathrm{vol}(\widehat{C} \triangle \mathcal{S}[X]; G_{n,r})$$

However, the symmetric set difference does not measure whether $\widehat{C}$ can distinguish any two distinct clusters $\mathcal{C}, \mathcal{C}' \in \mathbb{C}_f(\lambda)$. We therefore also study a second notion of cluster estimation, first introduced by Hartigan [1981], and defined asymptotically.

Definition 2.2 (Consistent density cluster estimation)**.** For an estimator $\widehat{C} \subseteq X$ and cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we say $\widehat{C}$ is a consistent estimator of $\mathcal{C}$ if for all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$ with $\mathcal{C} \neq \mathcal{C}'$, the following holds as $n \to \infty$:

$$\mathcal{C}[X] \subseteq \widehat{C} \quad \text{and} \quad \widehat{C} \cap \mathcal{C}'[X] = \emptyset, \tag{9}$$

with probability tending to 1.

Consistent cluster recovery roughly ensures that, for a given density threshold λ, the estimated cluster $\widehat{C}$ contains all points in a true density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, and simultaneously does not contain any points in any other density cluster $\mathcal{C}' \in \mathbb{C}_f(\lambda)$.

With these definitions in place, our broad goal will be to understand the extent to which the PPR algorithm is able to recover a cluster which either guarantees low symmetric set difference to a true density cluster, or which consistently estimates a true density cluster.

2.3 Geometric Conditioning of a Density Cluster.

At a high level, for PPR to be successful, the underlying density cluster must be geometrically well-conditioned. A basic requirement is that we need to avoid clusters which contain arbitrarily thin bridges

or spikes. As in the work of Chaudhuri and Dasgupta [2010], we consider a thickened version of $\mathcal{C} \in \mathbb{C}_f(\lambda)$ defined as $\mathcal{C}_\sigma := \{x \in \mathbb{R}^d : \mathrm{dist}(x, \mathcal{C}) \leq \sigma\}$, which we call the σ-*expansion* of $\mathcal{C}$. Here $\mathrm{dist}(x, \mathcal{C}) := \inf_{y \in \mathcal{C}} \|y - x\|$. We now list our conditions on $\mathcal{C}_\sigma$.

(A1) *Bounded density within cluster:* There exist constants $0 < \lambda_\sigma < \Lambda_\sigma < \infty$ such that $\lambda_\sigma \leq \inf_{x \in \mathcal{C}_\sigma} f(x) \leq \sup_{x \in \mathcal{C}_\sigma} f(x) \leq \Lambda_\sigma$.

(A2) *Low noise density:* There exists $c_0 > 0$ and $\gamma \in [0, 1]$ such that for any $x \in \mathbb{R}^d$ with $0 < \mathrm{dist}(x, \mathcal{C}_\sigma) \leq \sigma$,

$$\inf_{x' \in \mathcal{C}_\sigma} f(x') - f(x) \geq c_0 \mathrm{dist}(x, \mathcal{C}_\sigma)^\gamma.$$

Roughly, this assumption ensures that the density decays sufficiently quickly as we move away from the target cluster $\mathcal{C}_\sigma$, and is a standard assumption in the level-set estimation literature (see for instance Singh et al. [2009]).

(A3) *Lipschitz embedding:* There exists $g : \mathbb{R}^d \to \mathbb{R}^d$ with the following properties:

(a) we have $\mathcal{C}_\sigma = g(\mathcal{K})$, for a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ with $\mathrm{diam}(\mathcal{K}) = \sup_{x,y \in \mathcal{K}} \|x - y\| =: \mathrm{diam} < \infty$

(b) $\det(\nabla g(x)) = 1$ for all $x \in \mathcal{C}_\sigma$, where $\nabla g(x)$ is the Jacobian of g evaluated at x and

(c) for some $L \geq 1$,

$$\|g(x) - g(y)\| \leq L\|x - y\| \text{ for all } x, y \in \mathcal{K}.$$

Succinctly, we assume that $\mathcal{C}_\sigma$ is the image of a convex set with finite diameter under a measure preserving, Lipschitz transformation.

(A4) *Bounded volume:* For a set $\mathcal{S} \subseteq \mathbb{R}^d$, define the $\mathbb{P}$-weighted volume of $\mathcal{S}$ to be

$$\mathrm{vol}_{\mathbb{P},r}(\mathcal{S}) := \int_{\mathcal{S}} \mathbb{P}(B(x, r))\, d\mathbb{P}(x). \tag{10}$$

where $B(x, r)$ is the closed ball of radius r centered at x.

We assume that the neighborhood graph radius $0 < r \leq \sigma/2d$ is chosen such that

$$\mathrm{vol}_{\mathbb{P},r}(\mathcal{S}) \leq \frac{1}{2}\mathrm{vol}_{\mathbb{P},r}(\mathbb{R}^d).$$

2.3.1 Condition number

We now introduce the condition number $\kappa(\mathcal{C})$. As we will see in Theorems 1 and 2, the smaller $\kappa(\mathcal{C})$ is, the more success PPR will have in recovering the target cluster $\mathcal{C}$.

Definition 2.3 (Well-conditioned density clusters)**.** For $\lambda > 0$ and $\mathcal{C} \in \mathbb{C}_f(\lambda)$, let $\mathcal{C}$ satisfy (A1)–(A4) for some $\sigma > 0$. Then, for universal constants $c_1, c_2, c_3 > 0$ to be specified later, we set

$$\Phi_u(\mathcal{C}) := c_1 r \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma}, \ \tau_u(\mathcal{C}) := c_2 \frac{\Lambda_\sigma^4 d^3 \rho^2 L^2}{\lambda_\sigma^4 r^2} \log^2\left(\frac{\Lambda_\sigma}{\lambda_\sigma^2 r}\right) + c_3, \tag{11}$$

and letting $\kappa(\mathcal{C}) := \Phi_u(\mathcal{C}) \cdot \tau_u(\mathcal{C})$, we call $\mathcal{C}$ a κ-*well-conditioned density cluster*.

Remark: The population-level quantities $\Phi_u(\mathcal{C})$ and $\tau_u(\mathcal{C})$ measure the internal and external connectivity, respectively, of the thickened density cluster $\mathcal{C}_\sigma$. In particular, as we will see in Section 2.7, we will derive that $\Phi_u(\mathcal{C})$ is an upper bound on the normalized cut $\Phi(\mathcal{C}_\sigma[X]; G_{n,r})$, and that $\tau_u(\mathcal{C})$ is an upper bound on the mixing time $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$ (defined later in (20)). These bounds on graph functionals translate to estimates of PPR cluster quality via the results we mentioned in Section 1.2.3.

2.3.2 Well-initialized algorithm

As is typical in the local clustering literature, our algorithmic results will be stated with respect to specific ranges of each of the user-specified parameters. In particular, for a well-conditioned density cluster $\mathcal{C}$, we require that some of the tuning parameters of Algorithm 1 are chosen to fall within specific ranges,

$$\begin{aligned} 0 < r \leq \frac{\sigma}{2d}, \quad \alpha \in [1/10, 1/9) \cdot \frac{1}{\tau_u(\theta)}, \\ (L, U) \subseteq (1/50, 1/5) \cdot \frac{1}{2\binom{n}{2}\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_\sigma)}. \end{aligned} \tag{12}$$

Definition 2.4. If the input parameters to Algorithm 1 satisfy (12) for some well-conditioned density cluster $\mathcal{C}$, we say the algorithm is *well-initialized*.

In practice it is not feasible to set hyperparameters based on the underlying (unknown) density f. Typically, one tunes PPR over a range of hyperparameters and selects the cluster which has the smallest normalized cut.

2.4 Cluster Recovery in Symmetric Set Difference

We now present a bound on the volume of the symmetric set difference between the estimated cluster $\widehat{C}$ and empirical cluster $\mathcal{C}_\sigma[X]$. In this theorem, and hereafter, we let $c, c_i > 0$ represent universal constants, and $c(\mathbb{P}, r), c_i(\mathbb{P}, r) > 0$ represent constants which may depend on $\mathbb{P}, \lambda$ and r but not on the sample size n; where possible, we keep track of all constants in our proofs. This bound is stated with respect to the volume $\mathrm{vol}(\mathcal{C}_\sigma[X]; G_{n,r})$, which we will shorten to $\mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X])$ to keep things simple.

Theorem 1. *Fix $\lambda > 0$ and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a κ-well-conditioned density cluster for some $\sigma > 0$. If Algorithm 1 is* well-initialized *with respect to $\mathcal{C}$, then there exists a set $\mathcal{C}_\sigma[X]^g \subseteq \mathcal{C}_\sigma[X]$ of large volume with $\mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X]^g) \geq \frac{1}{2}\mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X])$ such that the following statement holds: let Algorithm 1 be run with any seed node $v \in \mathcal{C}_\sigma[X]^g$. Then, for any n large enough so that*

$$n \geq c_1(\mathbb{P}, r) \cdot (\log n)^{\max\{\frac{3}{d}, 1\}}$$

the volume of the symmetric set difference is upper bounded

$$\Delta(\mathcal{C}_\sigma[X], \widehat{C}) \leq c \cdot \kappa(\mathcal{C}) \cdot \mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X]), \tag{13}$$

with probability at least $1 - \frac{c_2(\mathbb{P},r)}{n}$.

This result establishes that the volume of the symmetric set difference $\Delta(\mathcal{C}_\sigma[X], \widehat{C})$ is upper-bounded by a quantity proportional to the difficulty of the clustering problem, as measured by the condition number $\kappa(\mathcal{C})$. The primary technical difficulty involved in proving Theorem 1 comes from upper bounding the random graph functionals $\Phi(\mathcal{C}_\sigma[X]; G_{n,r})$ and $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$; once this is accomplished, the result is a straightforward implication of Zhu et al. [2013].

2.5 Consistent Cluster Recovery

Theorem 2 implies that when $\kappa(\mathcal{C})$ is sufficiently small, $\widehat{C}$ will be a consistent density cluster estimator. We will need one additional regularity condition, to preclude arbitrarily low degree vertices for points $x \in \mathcal{C}'[X]$.

(A5) *Bounded density in other clusters:* Letting σ, λ_σ be as in (A1), for each $\mathcal{C}' \in \mathbb{C}_f(\lambda)$ and for all $x \in \mathcal{C}'_\sigma$, $\lambda_\sigma \leq f(x)$.

Next we give our main result on consistent cluster recovery by PPR.

Theorem 2. *Fix $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a κ-well-conditioned density cluster for some $\sigma > 0$, and additionally assume f satisfies (A5). If Algorithm 1 is well-initialized with respect to $\mathcal{C}$ and if*

$$\kappa(\mathcal{C}) \leq c\frac{(\lambda_\sigma r^d \nu_d)^2}{\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_\sigma)}, \tag{14}$$

then the following statement holds: there exists a set $\mathcal{C}_\sigma[X]^g \subseteq \mathcal{C}_\sigma[X]$ of large volume with $\mathrm{vol}(\mathcal{C}_\sigma[X]^g; G_{n,r}) \geq \frac{1}{2}\mathrm{vol}(\mathcal{C}_\sigma[X]; G_{n,r})$ such that if Algorithm 1 is run with any seed node $v \in \mathcal{C}_\sigma[X]^g$, then for any n large enough so that

$$n \geq c_1(\mathbb{P}, r) \cdot (\log n)^{\max\{\frac{3}{d}, 1\}}$$

$\widehat{C}$ satisfies (9) with respect to $\mathcal{C}$ with probability at least $1 - \frac{c_2(\mathbb{P},r)}{n}$. Therefore, $\widehat{C}$ is a consistent estimator of $\mathcal{C}$ in the sense of Definition 2.2.

2.6 Lower bound.

To show a lower bound for density clustering using PPR, we exhibit a hard case: that is, a distribution $\mathbb{P}$ for which PPR is unlikely to recover a density cluster. Let $\mathcal{C}_\sigma^{(0)}$, $\mathcal{C}_\sigma^{(1)}$, and $\mathcal{C}_\sigma^{(2)}$ be rectangles in $\mathbb{R}^2$,

$$\mathcal{C}_\sigma^{(0)} = \left[-\frac{\sigma}{2}, \frac{\sigma}{2}\right] \times \left[-\frac{\rho}{2}, \frac{\rho}{2}\right], \quad \mathcal{C}_\sigma^{(1)} = \mathcal{C}_\sigma^{(0)} - \{(\sigma, 0)\}, \quad \mathcal{C}_\sigma^{(2)} = \mathcal{C}_\sigma^{(0)} + \{(\sigma, 0)\} \qquad (0 < \sigma < \rho)$$

and let $\mathbb{P}$ be the mixture distribution over $\mathcal{X} = \mathcal{C}_\sigma^{(0)} \cup \mathcal{C}_\sigma^{(1)} \cup \mathcal{C}_\sigma^{(2)}$ given by

$$\mathbb{P} = \frac{1-\epsilon}{2}\Psi_1 + \frac{1-\epsilon}{2}\Psi_2 + \frac{\epsilon}{2}\Psi_0,$$

where Ψ_m is the uniform distribution over $\mathcal{C}_\sigma^{(m)}$ for $m = 0, 1, 2$. The density function f of $\mathbb{P}$ is simply

$$f(x) = \frac{1}{\rho\sigma}\left(\frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_\sigma^{(1)}) + \frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_\sigma^{(2)}) + \frac{\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_\sigma^{(0)})\right) \tag{15}$$

so that for any $\epsilon < \lambda < (1-\epsilon)/2$, $\mathbb{C}_f(\lambda) = \{\mathcal{C}_\sigma^{(1)}, \mathcal{C}_\sigma^{(2)}\}$. As the following theorem demonstrates, even when Algorithm 1 is reasonably initialized, if the density cluster $\mathcal{C}_\sigma^{(1)}$ is sufficiently geometrically ill-conditioned the cluster estimator $\widehat{C}$ will fail to recover $\mathcal{C}_\sigma^{(1)}$. Let

$$\mathcal{L} = \{(x_1, x_2) \in \mathcal{X} : x_2 < 0\}. \tag{16}$$

Theorem 3. *Suppose $r < \frac{1}{40}\rho \wedge \frac{1}{4}\sigma$, $\alpha = 65\Phi_{\mathbb{P}}(\mathcal{L})$, and $(L, U) = (0, 1)$ are inputs to Algorithm 1. Then, for any*

$$n \geq \max\left\{\frac{64}{\epsilon^2\rho\sigma\pi r^2}, \frac{8}{\epsilon}\right\} \tag{17}$$

the following statement holds: there exists a set $C^g \subset X$ with $\mathrm{vol}_{n,r}(C^g \cap \mathcal{C}_\sigma^{(1)}[X]) \geq \frac{1}{10}\mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(1)}[X])$ such that for any seed node $v \in C^g$, the estimator $\widehat{C}$ computed by Algorithm 1 has symmetric set difference with $\mathcal{C}_\sigma^{(1)}[X]$ of volume at least

$$\frac{\sigma\rho}{r^2}\frac{\mathrm{vol}_{n,r}(\widehat{C} \triangle \mathcal{C}_\sigma^{(1)}[X])}{n^2} \geq \frac{1}{4} - c\frac{\sqrt{\frac{\sigma}{\rho}}}{\epsilon^2}\sqrt{\log\left(\frac{\rho\sigma}{\epsilon^2 r^2}\right)\frac{\sigma}{r}} \tag{18}$$

with probability at least $1 - c_1(\mathbb{P}, r)n \exp\{-c_2(\mathbb{P}, r)n\}$. *Consequently, if*

$$\epsilon^2 > \frac{c}{8}\sqrt{\frac{\sigma}{\rho}} \cdot \sqrt{\log\left(\frac{\rho\sigma}{\epsilon^2 r^2}\right)\frac{\sigma}{r}}$$

then with high probability $\frac{\sigma\rho}{r^2}\frac{\mathrm{vol}_{n,r}(\widehat{C} \triangle \mathcal{C}_\sigma^{(1)}[X])}{n^2}$ *is at least* $1/8$.

The overall takeaway of Theorems 1 and 3 is summarized by Figure 1, which examines the behavior of PPR clustering over a density concentrated on two semicircles (the "two-moons" example). The takeaway is that even in a prototypical example where spectral clustering is designed to work well, PPR will recover high-density clusters (i.e. the moons) only when these clusters are geometrically compact. This is in contrast to traditional plug-in density cluster estimators, which are far more robust to the geometry of $\mathbb{P}$. Figure 1 shows that our theory applies even at moderate sample sizes (in this case $n = 800$ samples).

2.7 Analysis Overview

One of the primary technical contributions of our work is showing that the geometric conditions (A1)–(A4) translate to meaningful bounds on the normalized cut and mixing time of $\mathcal{C}_\sigma[X]$ in $G_{n,r}$.

2.7.1 Upper Bound on the Normalized Cut

We start with a finite sample upper bound on the normalized cut (8) of $\mathcal{C}_\sigma[X]$. For simplicity, we write $\Phi_{n,r}(\mathcal{C}_\sigma[X]) := \Phi(\mathcal{C}_\sigma[X]; G_{n,r})$.

Theorem 4. *Fix* $\lambda > 0$, *and assume* $\mathcal{C} \in \mathbb{C}_f(\lambda)$ *satisfies Assumptions (A1)–(A2), (A4) for some* $\sigma > 0$. *Then*

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[X])}{r} \leq c_1 \frac{d}{\sigma}\frac{\lambda}{\lambda_\sigma}\frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} \tag{19}$$

with probability at least $1 - 3\exp\{-nc_3(\mathbb{P}, r)\}$.

Remark: Observe that the diameter ρ is absent from Theorem 4, in contrast to the condition number $\kappa(\mathcal{C})$, which worsens (increases) as ρ increases. This reflects established wisdom regarding spectral partitioning algorithms more generally [Guattery and Miller, 1995, Hein and Bühler, 2010], albeit newly applied to the density clustering setting. It suggests that if the diameter ρ is large, PPR may fail to recover $\mathcal{C}_\sigma[X]$ even when $\mathcal{C}$ is sufficiently well-conditioned to ensure $\mathcal{C}_\sigma[X]$ has a small normalized cut in $G_{n,r}$.

2.7.2 Upper Bounds on the Mixing Time

For $S \subseteq V$, denote by $G[S] = (S, E_S)$ the subgraph induced by S (where the edges are $E_S = E \cap (S \times S)$). Let W_S be the (lazy) random walk matrix over $G[S]$, and write

$$q_v^{(t)}(u) = e_v W_S^t e_u$$

for the t-step transition probability of the lazy random walk over $G[S]$ originating at $v \in V$. Also write $\pi = (\pi(u))_{u \in S}$ for the stationary distribution of this random walk.(As W_S is the transition matrix of a lazy random walk, it is well-known that a unique stationary distribution exists and is given by $\pi(u) = \deg(u; G[S])/\mathrm{vol}(S; G[S])$, where we write $\deg(u; G[S]) = \sum_{w \in S} \mathbf{1}((u, w) \in E_S)$ for the degree of u in $G[S]$.) We define the *mixing time* of $G[S]$ as

$$\tau_\infty(G[S]) = \min\left\{t : \frac{\pi(u) - q_v^{(t)}(u)}{\pi(u)} \leq \frac{1}{4}, \text{ for } u, v \in V\right\}. \tag{20}$$

Next, we give an asymptotic (in the number of samples n) upper bound on $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$.

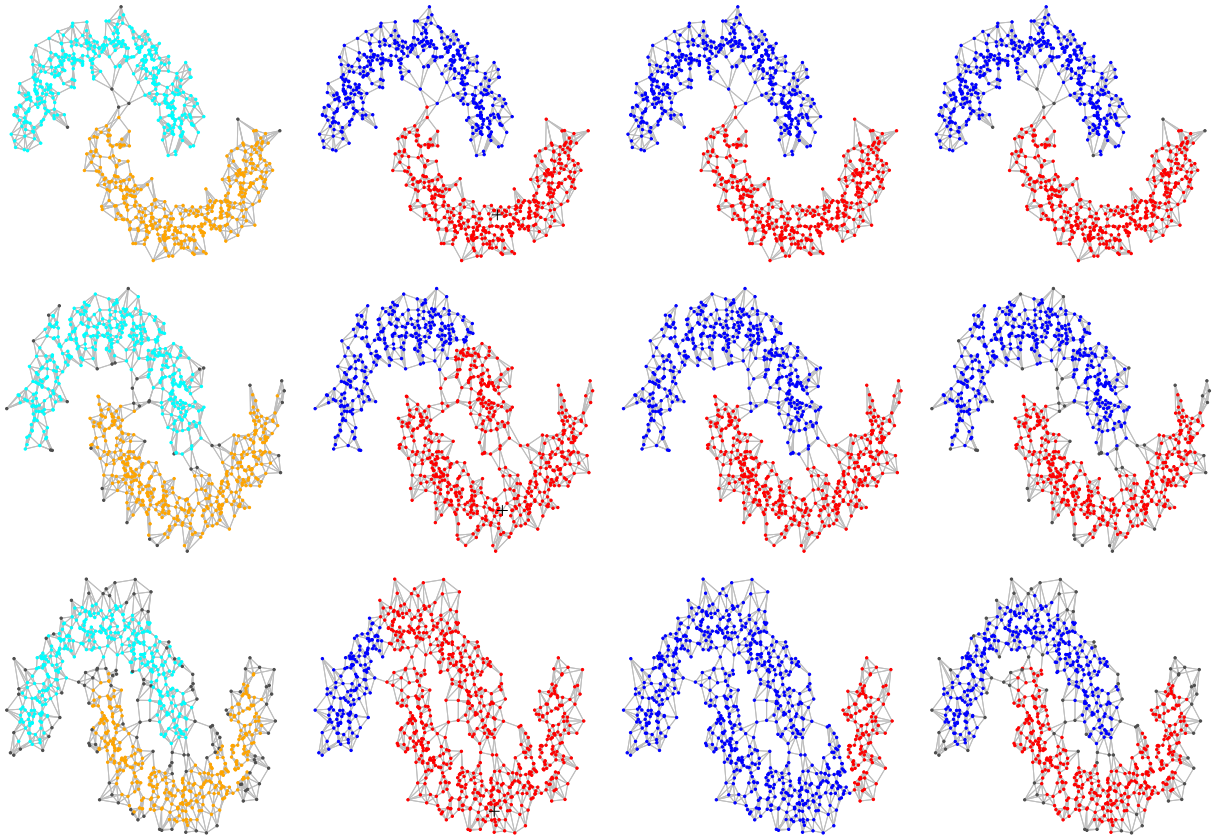


Figure 1: *True density (column 1), PPR (column 2), normalized cut (column 3) and estimated density (column 4) clusters for 3 different simulated data sets. Seed node for PPR denoted by a black cross.*

Theorem 5. *Fix $\lambda > 0$, and assume that $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfies Assumptions (A1) and (A3) for some $\sigma > 0$. Additionally assume that the radius $0 < r < \sigma/2\sqrt{d}$. Then the following statement holds: for any*

$$n \geq c_1(\mathbb{P}, r) \cdot (\log n)^{\max\{\frac{3}{d}, 1\}}$$

the mixing time $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$ satisfies

$$\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]]) \leq c_2 \frac{\Lambda_\sigma^4 d^3 \rho^2 L^2}{\lambda_\sigma^4 r^2} \log^2\left(\frac{\Lambda_\sigma}{\lambda_\sigma^2 r}\right) + c_3 \tag{21}$$

with probability at least $1 - \frac{c_4(\mathbb{P},r)}{n}$.

The proof of Theorem 5 relies heavily on analogous mixing time bounds developed for the mixing time of a continuous-space "ball walk" over convex sets. To the best of our knowledge, our result is the first bound on the mixing time of random walks over neighborhood graphs that is independent of n, the number of vertices.

3 Hypothesis testing with neighborhood graphs.

We define test statistics for both the regression and density testing setups. In the former, we show in some preliminary work that our proposed test statistic is minimax optimal over the Sobolev ball $W^{1,2}(\mathcal{X}; 1)$. In

the latter, we empirically demonstrate that our proposed approach has merit. We conclude this section by discussing some practical and computational considerations in this problem. Throughout, we detail the work we intend to complete for a final thesis.

3.1 Regression testing using graph spectral projections.

Recalling the regression model (1), let VSV^T be the spectral decomposition of the Laplacian matrix L of the neighborhood graph $G_{n,r}$. Inspired by (5), we introduce the following truncated-series test statistic:

$$T_{\text{spec}} := \frac{1}{n}\sum_{k=0}^{\kappa}\left(\sum_{i=1} v_i y_i\right)^2 \tag{22}$$

where the only difference between T_{harm} and T_{spec} is that the eigenvectors $\{v_i\}$ are now playing the role of the harmonic functions $\{\phi_\ell\}$. In Theorem 6 we show that under some mild regularity conditions on $\mathbb{P}$, the test $\phi_{\text{spec}} := \mathbf{1}\{T_{\text{spec}} \geq \tau\}$ is, up to log factors, a minimax optimal test over the Sobolev ball $W^{1,2}(\mathcal{X}; 1)$. To conveniently state our results we introduce the notation

$$h(\beta, d) = (1/d + \beta)(d/2 + s)(2d/(4s + d))$$

Theorem 6. *Let $b \geq 1, \beta > 0$ be fixed constants and let $d < 4$. Suppose that $\mathbb{P}$ is an absolutely continuous probability measure over $\mathcal{X} = [0,1]^d$ with density function p bounded above and below by constants, i.e*

$$0 < p_{min} < p(x) < p_{max} < \infty, \quad \text{for all } x \in \mathcal{X}.$$

Then there exists a constant $c_1(d, L, p_{\max}, b, \beta)$ which is independent of the sample size n such that the following statement holds: Construct the test statistic T_{spec} with parameter choices $r \asymp n^{-1/d}$ and $\kappa \asymp n^{2d/(4+d)}$, and additionally let the threshold $\tau \asymp \frac{\kappa}{n} + b\sqrt{\frac{\kappa}{n^2}}$. Then, for every ϵ satisfying

$$\epsilon^2 \geq c_1(d, L, p_{\max}, b, \beta) \cdot b \cdot n^{-4/(4+d)} (\log n)^{h(\beta,d)} \tag{23}$$

the worst-case risk is lower bounded

$$\mathcal{R}(\phi_{\text{spec}}; \mathcal{W}^{1,2}(\mathcal{X}; L)) \leq \left(\frac{2}{b^2} + \frac{2}{b\sqrt{\kappa}}\right) + o(1). \tag{24}$$

Therefore, $\epsilon_n^2(\phi_{\text{spec}}, \mathcal{W}^{1,2}(\mathcal{X}; L)) \asymp n^{-4/(4+d)} (\log n)^{h(\beta,d)}$.

A comparison of Theorem 6 with the rates derived by Ingster and Sapatinas [2009] for (5) demonstrates that the additional error incurred by using estimated eigenfunctions of the Laplacian, rather than directly using the Fourier series, results in the critical radius widening by only at most a factor of $(\log n)^{h(\beta,d)/2}$ from the minimax critical radius over $W^{1,2}(\mathcal{X}; 1)$. This means our test based on T_{spec} is almost *as good as*, in a minimax sense, the best possible test over $W^{1,2}(\mathcal{X}; 1)$. As mentioned previously, the problem when $d \geq 4$ requires different assumptions on f and is characterized by a different minimax rate.

3.1.1 Weighted Sobolev norms.

However, we would like to argue that in certain cases, T_{spec} is in fact *better* than competitors such as T_{harm}. While we do not yet have theoretical results along these lines, intuitively T_{spec} should be particularly sensitive to functions f which have smooth deviations from $f_0 = 0$ in high-density regions, but are potentially much wigglier in low-density regions. We therefore intend to (i) alter our function class $\mathcal{H}$ to be defined as the unit ball in a weighted Sobolev norm; precisely

$$\mathcal{H}_{\mathbb{P}} = \left\{ f \in L^2(\mathcal{X}) : \sum_{|\alpha| \leq s} \int \int_{\mathcal{X}} |D^\alpha f(x)|^2 \, p^2(x) \, dx \right\}$$

(ii) analyze the performance of T_{spec} over $\mathcal{H}_{\mathbb{P}}$, and (iii) provide example distributions $\mathbb{P}$ where T_{spec} outperforms reasonable competitors over $\mathcal{H}_{\mathbb{P}}$ (See Figure 3.2 for some empirical examples in the two-sample density testing case).

3.1.2 Higher-order derivative classes.

Theorem 6 covers only the $\{1,2\}$-Sobolev ball. In order to extend to higher-order Sobolev classes (i.e. the Sobolev $\{s,2\}$-balls where $s \geq 2$), we will likely need to consider kernels $K_r(u,v) = K\left(\frac{\|u-v\|}{r}\right)$ for which

$$\int x^i K(x)\, dx = 0, \quad \text{i} = 1,\ldots,\text{s} \tag{25}$$

as is typical in nonparametric smoothing problems Tsybakov [2008]. Choosing kernels which satisfy (25) as opposed to $K_r(x,y) = \mathbf{1}(\|x-y\| \leq r)$ allows for tighter control on the quadratic functional

$$\int\int (f(x)-f(y))^2 K_r(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y);$$

this control is critical to obtain minimax rates, and otherwise the analysis when $s = 1$ versus $s > 1$ is not substantially different. We therefore intend to show that with an appropriate choice of kernel K, along with tuning parameters κ and r, the test statistics T_{spec} and $T^{(2)}_{\text{spec}}$ (the latter of which we define in the following section) result in minimax optimal tests over all Sobolev $\{s,2\}$-balls, in the regression and density testing problems respectively.

3.2 Density testing using graph spectral projections.

It is straightforward to adapt the test statistic T_{spec} to the two-sample testing problem. Concatenate the samples in $X = (z_1,\ldots,z_N,y_1,\ldots,y_M)$, and let $a = (\underbrace{N^{-1},\ldots,N^{-1}}_{\text{length } N}, \underbrace{-M^{-1},\ldots,-M^{-1}}_{\text{length } M})$ encode whether a sample in X comes from Z or Y. Letting $n = N+M$, we define the variant $T^{(2)}_{\text{spec}}$ of T_{spec} to be

$$T^{(2)}_{\text{spec}} := \frac{1}{n}\sum_{k=0}^{\kappa}\left(\sum_{i=1}^{n} v_i a_i\right)^2 \tag{26}$$

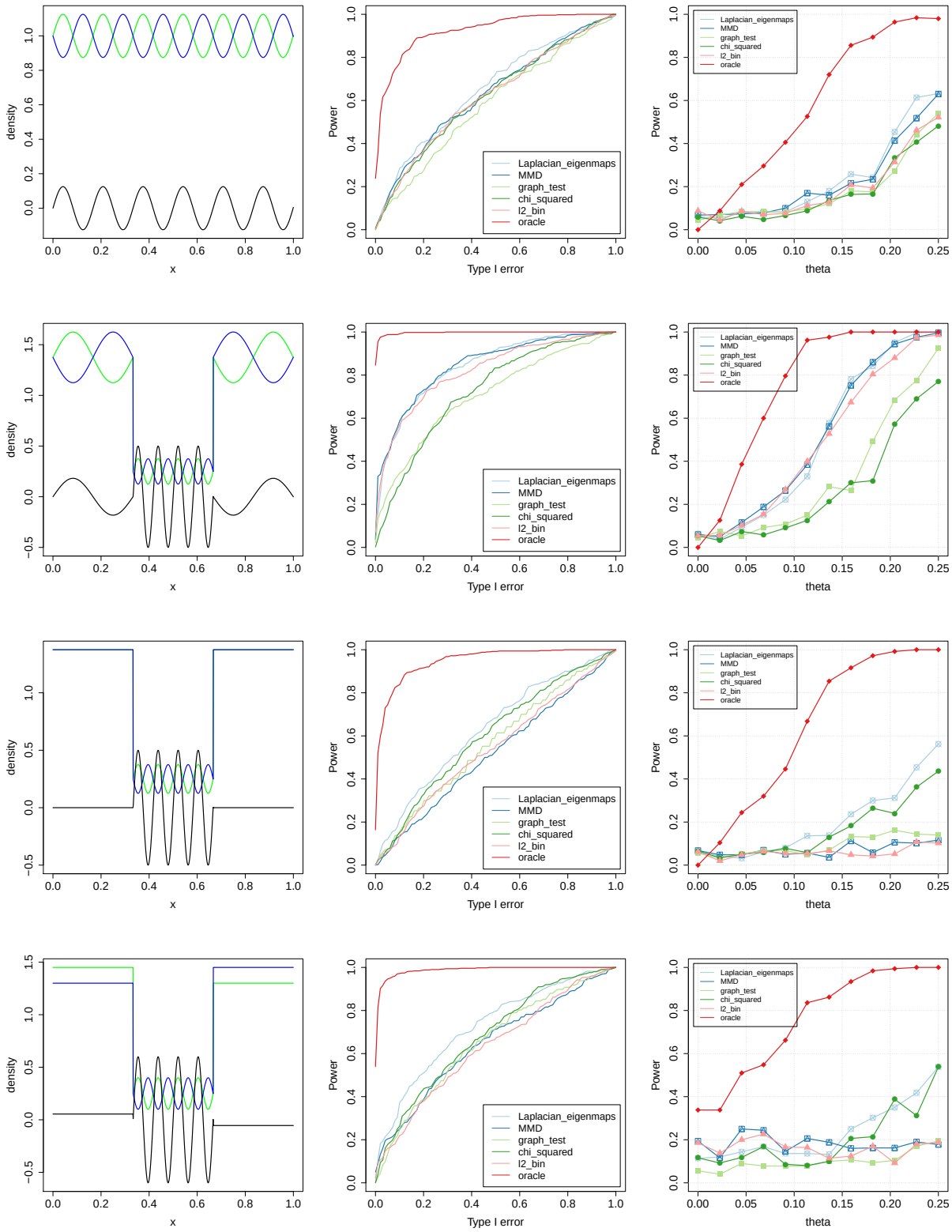
where as before $v_1,\ldots,v_n$ are eigenvectors of the Laplacian $L = L(X,r)$.

While we do not yet have theoretical results characterizing the behavior of $T^{(2)}_{\text{spec}}$, Figure 3.2 shows some instances in which $T^{(2)}_{\text{spec}}$ outperforms a range of common nonparametric tests. We see that the examples support our intuition; $T^{(2)}_{\text{spec}}$ is especially strong when the difference in densities $p-q$ oscillates at a higher frequency in regions where the mixture density $\frac{1}{2}(p+q)$ is small.

We intend to (i) derive upper bounds on the worst-case risk of the test $\phi^{(2)}_{\text{spec}} = \mathbf{1}(T^{(2)}_{\text{spec}} \leq \tau)$ over the class $C^s_d(\mathcal{X};L)$, and prove that it is minimax optimal, (ii) investigate the minimax rate for two-sample density testing over the Sobolev class $\mathcal{W}^{s,2}_d(\mathcal{X};L)$, and (iii) prove an analogue to Theorem 6 establish that $\phi^{(2)}_{\text{spec}}$ remains minimax optimal over the larger classes $\mathcal{W}^{s,2}_d(\mathcal{X};L)$.

3.2.1 Density testing for irregular alternatives.

As mentioned in Section 1.3.2, the regression and density testing models over e.g. Holder spaces are formally asymptotically equivalent; however, these asymptotic equivalence statements hold only when $2s > d$. To illuminate some differences between the density and regression testing problems when the regression function f or difference in densities $p-q$ is considered to be irregular (meaning $2s < d$), it is helpful to consider the


density
x
Power
Type I error
theta
Laplacian_eigenmaps
MMD
graph_test
chi_squared
l2_bin
oracle

following generative model: suppose we observe pairs (x_i, ℓ_i) for $i = 1, \ldots, (n+m)$, where as usual $x_i \sim \mathbb{P}$ are independent, and additionally ℓ_i are conditionally independent given x_i with distribution

$$\ell_i = \begin{cases} 1, & \text{with probability } \frac{p(x)}{p(x)+q(x)} \\ -1, & \text{with probability } \frac{q(x)}{p(x)+q(x)}. \end{cases}$$

We can relate this to the density testing problem by applying Bayes rule, whence we derive that the conditional distribution of $X_i|\ell_i = 1$ is $\mathbb{P}$, and similarly the conditional distribution of $X_i|\ell_i = -1$ is $\mathbb{Q}$. On the other hand, since $\ell_i = f(x_i) + w_i$ where $f = \frac{p-q}{p+q}$ and w_i is a mean-zero error term, this resembles in some respects the regression testing setup. However, w_i is now heteroskedastic with variance depending on the unknown densities p and q. The fact that the variance of w_i is unknown is crucial, and as previously mentioned it implies that there will no longer be a elbow in rates at $d = 4$. Indeed, as previously stated the minimax critical radius for the two-sample density testing problem over $C_d^1(\mathcal{X}; L)$ is known to be $\epsilon_n^\star(C_d^1(\mathcal{X}; L)) \asymp n^{-2/(4+d)}$ for all d, including $d \geq 4$.

When $d \geq 4$, choosing the hyperparameters for $T_{\text{spec}}^{(2)}$ becomes more subtle. Statistics such as T_{harm} – which take empirical evaluations of an orthogonal basis in $L^2(\mathcal{X})$–continue to be minimax optimal over $C_d^1(\mathcal{X}; L)$ when $\kappa \asymp n^{2d/(4+d)}$. However, since this choice implies $\kappa > n$ the statistic $T_{\text{spec}}^{(2)}$ – which takes an orthogonal basis in $L^2(\mathbb{P}_n)$ – reduces to simply $T_{\text{spec}}^{(2)} = 1$ regardless of $\mathbb{P}_n$ and $\mathbb{Q}_n$. Intuition from chi-squared tests–where the optimal choice of bin-width is on the order of $n^{-2/(4+d)}$, less than $n^{-1/d}$ whenever $d > 4$–suggests that we should modify our test statistic by choosing r to be shrinking at a rate faster than $n^{-\frac{1}{d}}$. Of course at this connectivity scale with high probability the graph $G_{n,r}$ is disconnected Penrose [1999], and our analysis will necessarily be quite different than before. We intend to pursue such a modification and the resulting analysis.

3.3 IPMs on neighborhood graphs.

We define a pair of graph-based test statistics on the neighborhood graph $G_{n,r}$, which we cast as IPMs over different "function" classes $\Theta \subset \mathbb{R}^n$.

3.3.1 A graph-based Sobolev IPM.

We term our first test statistic the *graph Sobolev* IPM, since the function class Θ over which we take the supremum will resemble a continuous Sobolev space. Formally we let

$$T_{\text{sob}} := \sup_{\theta \in \mathbb{R}^n : \|B^{(s)}\theta\|_2 + \lambda\|\theta\|_2 \leq 1} \left|a^T\theta\right| = \sqrt{a^T(L^s + \lambda I)^{-1}a}, \quad \lambda > 0, \tag{27}$$

where we recall the notation $a = \left(N^{-1}, \ldots, N^{-1}, -M^{-1}, \ldots, -M^{-1}\right)$. (We could easily adapt this statistic to the regression testing case by replacing a with y, although we would lose the analogy to an IPM).

Remark: Two aspects of the statistic T_{sob} are worthy of comment. The first regards the role played by the parameter λ, which is at first glance somewhat mysterious. It may seem more natural to consider the statistic

$$\sup_{\theta \in \mathbb{R}^n : \|B\theta\|_2 \leq 1} \left|a^T\theta\right|,$$

so that the function class is defined only with respect to the semi-norm $\|B\theta\|_2$. However, it can be shown that this function class is too rich, and our test statistic will overfit to noise; hence we restrict it in the manner of (27), a modification similar to that of Arbel et al. [2018] who observe that it improves empirical performance. Second, we note that in general, an analogous IPM in the continuum Sobolev space is a poor candidate for testing. Specifically, we have that when $2s < d$,

$$\sup_{f \in L^2(\mathcal{X}) : \|f\|_{W^{s,2}(\mathcal{X})} \leq L} |\mathbb{P}_n f - \mathbb{Q}_n f| = \infty \tag{28}$$

regardless of $\mathbb{P}_n$ and $\mathbb{Q}_n$, achieved by taking spike functions at data. (This phenomenon has been observed in the context of semi-supervised learning in e.g. Nadler et al. [2009], Zhou and Belkin [2011], Slepcev and Thorpe [2017], El Alaoui et al. [2016].) However, by replacing the continuous Sobolev norm in (28) with a discrete Sobolev norm in (27), we have obviated this concern.

Based on the success of Sobolev IPMs in real-world problems Mroueh et al. [2017], Arbel et al. [2018] and the favorable properties of analogous estimators when performing regression on a grid Sadhanala et al. [2016a], we believe testing with T_{sob} is a reasonable idea. We intend to investigate the minimax properties of such a test with respect to the function classes $C^s(\mathcal{X})$ and $W^{s,2}(\mathcal{X})$.

3.3.2 A graph-based Total Variation (TV) IPM.

Our second IPM test statistic will resemble the first, except we replace the graph Sobolev norm by a graph total variation norm. We hence call this statistic the *graph TV* IPM, defined as

$$T_{\text{TV}} := \sup_{\theta \in \mathbb{R}^n : \|B\theta\|_1 + \lambda \|\theta\|_2 \leq 1} \left| a^T \theta \right|.$$

To motivate the graph TV IPM, we give some background on testing in TV spaces.

Testing in TV spaces. To one extent or another, functions in the Sobolev and Holder balls exhibit homogeneous smoothness. In reality we may not wish to restrict our attention to such functions. This is in part the motivation behind considering functions of bounded variation. Formally, for a function $f \in L^1(\mathcal{X})$ the *total variation* semi-norm of f is Evans and Gariepy [2015]

$$TV(f; \mathcal{X}) := \sup \left\{ \int_U f \operatorname{div} \psi \, dx : \psi \in C_c^1(\mathcal{X}; \mathbb{R}^d), |\psi| \leq 1 \right\};$$

and we write $BV_d(\mathcal{X})$ for the subset of functions $f \in L^1(\mathcal{X})$ which have bounded norm

$$\|f\|_{BV_d(\mathcal{X})} := \|f\|_\infty + TV(f; \mathcal{X}).$$

In the one-dimensional case, some work has been done on nonparametric testing over the Besov spaces $B^s_{p,q}(\mathcal{X})$ (for a formal definition see Section 4.3 of Giné and Nickl [2016]), which we can relate to bounded variation spaces via the embeddings

$$B^1_{11}(\mathcal{X}) \subseteq BV_1(\mathcal{X}) \subseteq B^1_{1\infty}(\mathcal{X}). \tag{29}$$

Lepski and Spokoiny [1999] study the regression testing problem, with data observed according to the white noise model as in (2). They derive among other things that for the Besov balls $B^s_{1,q}(\mathcal{X}; L)$, the minimax critical radius $\epsilon_n^\star(B^s_{1,q}(\mathcal{X}; L)) \asymp n^{-2s'/(4s'+1)}$, where $s' = s - \frac{1}{4}$ and the result holds for any $s > 1, q \geq 1$. Ingster and Suslina [2000] improve on this by showing the same result holds whenever $s > 1/2$. This along with (29) implies that $\epsilon_n^\star(BV_1(\mathcal{X})) \asymp n^{-\frac{3}{8}}$.

When $d \geq 2$ minimax testing rates over the Besov and bounded variation spaces are still unknown. However, study of the equivalent estimation problem suggests the answers may be substantially different from the 1d problem. Sadhanala et al. [2016a] consider a discretized version of this problem – where observations are made on a discrete grid and the regression function is assumed to lie within a discrete bounded variation class – and show that the minimax estimation rate in empirical norm is on the order of $n^{-1/d}$ for $d \geq 2$. This is slower than minimax estimation rate over the Sobolev space $\mathcal{W}^{1,2}_d$, the largest Sobolev space inscribed within BV_d. Delyon and Juditsky [1996], Kerkyacharian et al. [2008], Lepski [2015], Ruiz et al. [2018] all consider the regression white noise model (2) and show, in progressive degrees of generality, that the minimax rates over the Besov spaces $B^s_{p,q}$ exhibit a phase transition at the boundary $2s = d$. When $2s > d$, the minimax

estimation rate (with loss measured in squared $\mathcal{L}_2$ norm) is the standard $n^{-2s/(2s+d)}$; on the other hand when $2s < d$, the minimax rate is $n^{-s/d}$ which is much slower. When $s = 1$, this phase transition occurs at $d = 2$, and so in a minimax sense the univariate and multivariate estimation problems over BV_d are revealed to be quite different. Sadhanala et al. [2017] show that this same phenomenon carries over to the grid, and the discrete TV classes defined upon it.

We intend to investigate whether such a phase transition (from $d = 1$ to $d \geq 2$) exists in the testing setting as well, by deriving upper and lower bounds for the critical radius $\epsilon_n(BV_d(\mathcal{X}; L))$, where $BV_d(\mathcal{X}; L) = \big\{f \in BV_d(\mathcal{X}) : \|f\|_{BV_d(\mathcal{X})} \leq L\big\}$. Our upper bound will be based on the test $\phi_{TV} := \mathbf{1}\{T_{\text{TV}} \geq \tau\}$.

3.3.3 Computational considerations.

Although our primary focus has been on the statistical properties of various tests we have defined, there are also some computational considerations which merit our attention.

When n is large and the graph $G_{n,r}$ is relatively dense, exactly computing the eigenvectors of L (as in (22)) or the matrix inverse $(L + \lambda I)^{-1}$ (as in (27)) may be computationally expensive. Recently, there has been remarkable progress made (e.g. Spielman and Teng [2011], Batson et al. [2013], Spielman and Teng [2014], Cohen et al. [2014]) in efficiently computing *ϵ-spectral sparsifiers* of G: that is, graphs $\widetilde{G}$ with Laplacian matrices $\widetilde{L}$ which satisfy:

$$(1 - \epsilon)x^T\widetilde{L}x \leq x^T Lx \leq (1 + \epsilon)x^T\widetilde{L}x, \quad \text{for some } 0 \leq \epsilon < 1, \text{ and all } x \in \mathbb{R}^n. \tag{30}$$

and have many fewer edges than G (hence the term "sparsifier".) Since $\widetilde{L}$ is a sparse matrix, we can compute its eigenvectors $\{\widetilde{v}\}$ or the inverse $(\widetilde{L} + \lambda I)^{-1}$ efficiently. Additionally (30) should imply that the resulting test statistics behave similarly to those defined with respect to L. As yet limited work von Luxburg et al. [2014], Sadhanala et al. [2016b] has been done to rigorously analyze the performance of spectral sparsifiers in statistical learning problems, despite their obvious promise. We plan to address this gap by providing a theoretical characterization of the performance of graph statistics on sparsified graphs in the regression and density testing problems.

REFERENCES

Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 475–486, 2006.

Michael Arbel, Dougal J. Sutherland, and Arthur Gretton. On gradient regularizers for mmd gans. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6701–6711, 2018.

Ery Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *Information Theory, IEEE Transactions on*, 57:1692 – 1706, 04 2011.

Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471, 2018.

Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems 26*, pages 2679–2687, USA, 2013. Curran Associates, Inc.

Joshua Batson, Daniel Spielman, Nikhil Srivastava, and Shang-Hua Teng. Spectral sparsification of graphs: Theory and algorithms. *Communications of the ACM*, 56:87–94, 08 2013.

Bhaswar B Bhattacharya. Two-sample tests based on geometric graphs: Asymptotic distribution and detection thresholds. *arXiv preprint arXiv:1512.00384*, 2015.

Lawrence D. Brown and Mark G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 12 1996.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23*, pages 343–351. Curran Associates, Inc., 2010.

Michael B. Cohen, Rasmus Kyng, Gary L. Miller, Jakub W. Pachocki, Richard Peng, Anup B. Rao, and Shen Chen Xu. Solving sdd linear systems in nearly mlog1/2n time. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 343–352, 2014.

Bernard Delyon and Anatoli Juditsky. On minimax wavelet estimators. *Applied and Computational Harmonic Analysis*, 3(3):215–228, 1996.

Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of\ell_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, 2010.

Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. Chapman and Hall/CRC, 2015.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, March 2012.

Stephen Guattery and Gary L Miller. On the performance of spectral graph partitioning methods. In *SODA*, volume 95, pages 233–242, 1995.

Emmanuel Guerre and Pascal Lavergne. Optimal minimax rates for nonparametric specification testing in regression models. *Econometric Theory*, 18(5):1139–1171, 2002.

John A. Hartigan. Consistency of single-linkage for high-density clusters. *Journal of the American Statistical Association*, 1981.

Taher H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

Matthias Hein and Thomas Bühler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca. In *Advances in Neural Information Processing Systems 23*, pages 847–855, 2010.

Yu. I. Ingster. Minimax nonparametric detection of signals in white gaussian noise. *Problems Inform. Transmission*, 18:130–140, 1982.

Yu. I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

Yu. I. Ingster and T. Sapatinas. Minimax goodness-of-fit testing in multivariate nonparametric regression. *Mathematical Methods of Statistics*, 18(3):241–269, 2009.

Yu I. Ingster and Irena Suslina. Minimax nonparametric hypothesis testing for ellipsoids and besov bodies. *ESAIM: Mathematical Modelling and Numerical Analysis*, 4:53–135, 2000.

Yu. I. Ingster and Irena Suslina. On estimation and detection of smooth function of many variables. *Math. Methods Statist.*, 14:299–331, 2005.

Arnold Janssen. Global power functions of goodness of fit tests. *Ann. Statist.*, 28(1):239–253, 02 2000.

Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, May 2004.

Gerard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. sparse case. *Siam Journal on Theory of Probability and its Applications*, 52:58–77, 03 2008.

Aleksandr P. Korostelev and Alexandre B. Tsybakov. *Minimax theory of image reconstruction.* Springer, 1993.

Samory Kpotufe and Ulrike von Luxburg. Pruning nearest neighbor cluster trees. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 225–232, 2011.

Oleg Lepski. Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.*, 43 (3):1178–1242, 06 2015.

Oleg V. Lepski and Vladimir G. Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 04 1999.

Jure Leskovec, Kevin J. Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Michael W. Mahoney, Lorenzo Orecchia, and Nisheeth K. Vishnoi. A local spectral method for graphs: with applications to improving graph partitions and exploring data graphs locally. *Journal of Machine Learning Research*, 13:2339–2365, 2012.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 1330–1338, 2009.

Michael Nussbaum. Asymptotic equivalence of density estimation and gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 12 1996.

Mathew D. Penrose. On k-connectivity for a geometric random graph. *Random Struct. Algorithms*, 15(2): 145–164, 9 1999.

Wolfgang Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Ann. Statist.*, 23(3):855–881, 1995.

Markus Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982, 08 2008.

Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15 (4):1154–1178, 2009.

Miguel del Álamo Ruiz, Housen Li, and Axel Munk. Frame-constrained total variation regularization for white noise regression. *arXiv preprint arXiv:1807.02038*, 2018.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, pages 3513–3521, 2016a.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James L Sharpnack, and Ryan J Tibshirani. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, pages 5800–5810, 2017.

Veeru Sadhanala, Yu-Xiang Wang, and Ryan Tibshirani. Graph sparsification approaches for laplacian smoothing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 1250–1259, 2016b.

Geoffrey Schiebinger, Martin J. Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. *Ann. Statist.*, 43(2):819–846, 04 2015.

Mark F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.*, 37(6B):3960–3984, 12 2009.

Aarti Singh, Clayton Scott, and Robert Nowak. Adaptive hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782, 10 2009.

Dejan Slepcev and Matthew Thorpe. Analysis of p-laplacian regularization in semi-supervised learning. *SIAM Journal on Mathematical Analysis*, 51:2085–2120, 2017.

Daniel A Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Daniel A Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.

Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35 (3):835–885, 2014.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Nicolas Garcia Trillos, Franca Hoffmann, and Bamdad Hosseini. Geometric structure of graph laplacian embeddings. *arXiv preprint arXiv:1901.10651*, 2019.

Nicolás García Trillos and Dejan Slepcev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018.

Alexandre B Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 1997.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 04 2008.

Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15:1751–1798, 2014.

Xueyuan Zhou and Mikhail Belkin. Semi-supervised learning by higher order regularization. *Journal of Machine Learning Research - Proceedings Track*, 15:892–900, 01 2011.

Zeyuan Allen Zhu, Silvio Lattanzi, and Vahab S Mirrokni. A local algorithm for finding well-connected clusters. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, pages 396–404, 2013.