Statistical Guarantees for Spectral Methods on Neighborhood Graphs

Alden Green

July 9th, 2021

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh PA, 15213

Thesis Committee:
Sivaraman Balakrishnan (Co-chair)
Ryan Tibshirani (Co-chair)
Alessandro Rinaldo
Dejan Slepčev (CMU Math Department)
Mikhail Belkin (UCSD)

Abstract

This thesis studies spectral methods on neighborhood graphs. These methods operate on point-cloud data. They form a neighborhood graph over this data, use the eigenvectors and eigenvalues of a graph Laplacian matrix to map each point to a set of data-dependent features, then run a simple algorithm which uses these features to perform a downstream task. Such methods are very general, and can be used to solve many different learning problems. They are also flexible, in that the feature mapping adapts to the structure of the data. However, this generality and flexibility also makes it challenging to understand the theoretical properties of spectral methods.

To understand these theoretical properties, we adopt the classical perspective of nonparametric statistics. We provide precise guarantees which establish that spectral methods on neighborhood graphs can effectively solve various statistical problems. Chapter 2 analyzes a local spectral clustering method, PPR clustering, and shows that it approximately recovers well-conditioned density clusters. Chapters 3 and 4 consider two methods for regression, Laplacian smoothing and Laplacian eigenmaps, and show that each method is minimax optimal under various data models. In Chapter 5, we discuss assorted other properties of these methods, which differentiate them from more classical nonparametric approaches.

Acknowledgments

To start at the beginning, I would like to thank my family: Laurin, Davis, Mom, and Papa. My family members are all incredibly bright and inquisitive; they are also incredibly argumentative. Growing up, I learned to only make statements I could justify, to discard those beliefs I could not, and to question everything else. On the other hand, I was also raised with the (unjustifiable) confidence that I could solve any problem if I thought hard enough about it. My parents also made sure I received a great education, and still lived a balanced life (at times over my own objections). I have them to thank for many of the qualities that served me well in obtaining my doctorate. At the same time, funnily enough we are not really a mathematical group by inclination or by training. So I also want to thank my family for being wholeheartedly supportive as I made the somewhat surprising choices that ended up leading to this thesis.

I would also like to thank various members of the Carnegie Mellon Statistics Department, which has been my academic home for the last six years. My first advisor, Cosma Shalizi, gave me my first problem to solve, showed me the ropes of research, and only slightly complained when I sent him a first draft with no citations and with the topic sentence "Networks matter." The members of the Carnegie Mellon Networkshop patiently listened to me repeatedly give the same basic talk, early versions of which were largely unintelligible, and gently pushed me to speak simply and clearly about complicated subject matter. Ale Rinaldo gave me essential professional and personal advice—for which I haven't thanked him enough—and ended up being one of my thesis committee members. I enjoyed talking about my research with Ale, and also with my (external to the CMU Stat Department) committee members Misha Belkin and Dejan Slepcev; I was fortunate to have a thesis committee representing a variety of intellectual communities (statistics, machine learning, computer science, math) and each of them brought a distinct perspective while making insightful comments about my work. I would especially like to thank my thesis advisors Siva Balakrishnan and Ryan Tibshirani. They taught me the intellectual joys of doing rigorous theoretical work (which I quickly appreciated), but also pushed me to view it as only one piece in the larger research puzzle (which I am still working on appreciating). Finally, I greatly enjoyed being a part of the community of PhD students in our department; it is one of my bigger regrets that I did not end up collaborating with other students on more of my work.

As any PhD student will tell you, doctoral work is a rewarding but lonely pursuit, and it has been particularly so over the past 16 months. I have been fortunate to have a group of friends who helped make sure it was a little less lonely (again, sometimes over my own objections). I enjoyed our game nights, getting dinners with Ilmun and Neil, going to the movies with Xiao Hui, and beating Kayla and Ben in egg toss. I am not always the most social person, and they deserve extra thanks for putting up with my occasional bouts of introversion. I would like to especially thank Abby, whose enthusiasm for life and adventure are an inspiration, and who has often served as my window to a wider world.

This document is the culmination of n years of work, where n equals (depending on how you look at it): three and a half (the time I've spent working on the topics of this thesis), six (the time I've spent in the Carnegie Mellon Statistics PhD program), ten (the time I've spent learning about statistics), or twenty-seven and three quarters (the time I've spent learning, period). Regardless of how you slice it, it represents a significant chunk of my life. Naturally, then, there are many people whose influence can be felt in this final product; that I have singled out only a strict, small subset of these above should not obscure this fact.

Contents

Chapter 1

Introduction

Spectral algorithms on neighborhood graphs are a powerful family of methods for learning from data in a flexible, nonparametric manner. They work as follows. Given n data points $X_1, \dots, X_n$, the algorithm first forms a *neighborhood graph* $G = (\{1, \dots, n\}, A)$, a weighted graph with weights A_{ij} measuring the proximity between points X_i and X_j. Then a *graph Laplacian* matrix $L \in \mathbb{R}^{n \times n}$ is constructed, which operates on vectors $u \in \mathbb{R}^n$ by taking sums of differences; for instance, the unweighted Laplacian $L = D - A$ (D being the diagonal degree matrix), or random walk Laplacian $L = I - D^{-1}A$. The spectrum (eigenvectors and eigenvalues) of this Laplacian matrix corresponds to a feature embedding, and a simple algorithm is then applied to this feature embedding, depending on the task at hand. For instance, if the goal is to cluster the points $X_1, \dots, X_n$, k-means clustering can be applied, whereas if the goal is to learn a relationship between $X_1, \dots, X_n$ and some observed responses $Y_1, \dots, Y_n$, one can run linear or ridge regression.

Spectral methods are intriguing because the spectrum of the graph Laplacian gives a data-dependent way to measure smoothness of functions f defined over the data $X_1, \dots, X_n$. Small eigenvalues of L correspond to smooth eigenvectors, and large eigenvalues to wiggly eigenvectors. Measuring and enforcing smoothness is one of the core principles underlying the theory and methods of nonparametric statistics. For this reason, it is very natural to study the properties spectral methods through such a nonparametric lens, and we adopt this perspective throughout this thesis.

Specifically, we consider three spectral methods—PPR clustering (Chapter 2), Laplacian smoothing (Chapter 3), and Laplacian eigenmaps (Chapter 4)—for three different learning tasks: clustering, estimation and hypothesis testing. We provide theoretical guarantees showing that when the data are randomly sampled, these methods approximate ideal ground-truth solutions as the number of samples n grows. In each case, we give finite-sample bounds on the error. In the latter two cases—meaning Laplacian smoothing and Laplacian eigenmaps for both estimation and testing—we show that the algorithms are optimal, in a minimax sense, over large classes of functions.

1.0.1 Why Spectral Methods on Neighborhood Graphs?

The primary contribution of this thesis is to show that spectral algorithms computed over neighborhood graphs are well suited to solve many classical problems in nonparametric statistics. It has long been understood that spectral methods (not using neighborhood graphs) are powerful candidates for such problems. Indeed, some fundamental results show that spectral algorithms are theoretically optimal for certain problems in nonparametric regression. However, these algorithms depend on the spectrum (eigenvalues and eigenfunctions) of complex continuum operators, which cannot typically be found in practice. In contrast, the spectrum of graph Laplacians can be easily computed, and serves as a practical approximation to the unknown spectrum of these aforementioned continuum operators. This thesis shows that, in certain senses,

methods that use the spectrum of a graph Laplacian inherit the strong theoretical guarantees of their continuum limits.

However, in each problem we consider there exist other (non-spectral) methods which are also well suited for the task at hand. This naturally leads to the question: why should one use spectral algorithms over neighborhood graphs? Or more precisely: do there exist *theoretical* (as opposed to empirical or practical) and *statistical* (as opposed to computational) reasons why one should prefer spectral algorithms over neighborhood graphs?

This question represents a high bar—one common theme to work on minimax theory for statistical problems is that there almost always exist multiple methods which are more or less optimal for any given problem—and we cannot give a conclusive answer. However, in Chapter 5 we do offer a few explanations about what graphs bring to the table that other methods cannot. Focusing on regression, we argue that spectral algorithms on graphs return estimates which are both *density adaptive* and, potentially, *spiky*. This is in contrast to more standard methods for nonparametric regression, which enforce a more uniform notion of smoothness. Both properties are intriguing, for separate reasons. Density adaptivity means spectral algorithms can provably outperform other estimators under certain structural assumptions on the relationship between design distribution and regression function. Spikiness, a property more pronounced in Laplacian smoothing than Laplacian eigenmaps, means that the former can have superior properties out-of-sample (that is, for prediction) than in-sample (that is, for estimation). This is similar to the phenomenon of statistically well-behaved interpolation, which of late has attracted tremendous interest in the statistics and machine learning communities.

Finally, a note on organization: each of Chapters 2-4 corresponds to a different paper, either published or currently in submission. These chapters are entirely self-contained, and each can be read separately from the other two. As such, we make no effort to standardize notation between them.

Chapter 2

Statistical Guarantees for Local Spectral Clustering on Neighborhood Graphs

In this paper, we consider the problem of clustering: splitting a given data set into groups that satisfy some notion of within-group similarity and between-group difference. Our particular focus is on spectral clustering methods, a family of powerful nonparametric clustering algorithms. Roughly speaking, a spectral algorithm first constructs a geometric graph G, where vertices correspond to samples, and edges correspond to proximities between samples. The algorithm then estimates a feature embedding based on a (suitable) Laplacian matrix of G, and applies a simple clustering technique (like k-means clustering) in the embedded feature space.

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], and then further developed by several others [Spielman and Teng, 2011, 2014, Andersen et al., 2006, Mahoney et al., 2012, Zhu et al., 2013].

Local spectral clustering techniques have been practically very successful [Leskovec et al., 2010, Andersen et al., 2012, Gleich and Seshadhri, 2012, Mahoney et al., 2012, Wu et al., 2012], which has led many authors to develop supporting theory [Spielman and Teng, 2013, Andersen and Peres, 2009, Gharan and Trevisan, 2012, Zhu et al., 2013] that gives worst-case guarantees on traditional graph-theoretic notions of cluster quality (such as normalized cut and conductance). In contrast, in this paper we adopt a classical statistical viewpoint, and examine what the output of local clustering on a data set reveals about the underlying density f of the samples. We establish conditions on f under which PPR, when appropriately tuned and initialized inside a candidate cluster $\mathcal{C} \subseteq \mathbb{R}^d$, will approximately recover this candidate cluster. We pay special attention to the case where $\mathcal{C}$ is a *density cluster* of f—defined as a connected component of the upper level set $\{x \in \mathbb{R}^d : f(x) \geq \lambda\}$ for some $\lambda > 0$—and show precisely how PPR accounts for both geometry and density in estimating a cluster.

Before giving a more detailed overview of our main results, we formally define PPR on a neighborhood graph, review some of the aforementioned worst-case guarantees, and introduce the population-level functionals that govern the behavior of local clustering in our statistical context.

2.0.1 PPR clustering

We start by reviewing the PPR clustering algorithm. Let $G = (V, E)$ be an undirected, unweighted, and connected graph. We denote by $A \in \mathbb{R}^{n \times n}$ the adjacency matrix of G, 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. The PPR vector $p_v = p(v, \alpha; G)$ is defined with respect to a given seed node $v \in V$ and a teleportation parameter $\alpha \in [0, 1]$, as the solution of the following linear system:

$$p_v = \alpha e_v + (1 - \alpha) p_v W, \tag{2.1}$$

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

In practice, exactly solving the system of equations (2.1) to compute the PPR vector may be too computationally expensive. To address this limitation, Andersen et al. [2006] introduced the *ε-approximate* PPR vector (aPPR), which we will denote by $p_v^{(\varepsilon)}$. We refer the curious reader to Andersen et al. [2006] for a formal algorithmic definition of the aPPR vector, and limit ourselves to highlighting a few salient points: the aPPR vector can be computed in order $\mathcal{O}(1/(\varepsilon\alpha))$ time, while satisfying the following uniform error bound:

$$\text{for all } u \in V, \quad p_v(u) - \varepsilon D_{uu} \leq p_v^{(\varepsilon)}(u) \leq p_v(u). \tag{2.2}$$

Once p_v or $p_v^{(\varepsilon)}$ is computed, the cluster estimate $\widehat{C}$ is chosen by taking a particular sweep cut. For a given level $\beta > 0$, the *β-sweep cut* of $p_v = (p_v(u))_{u \in V}$ is

$$S_{\beta,v} := \left\{ u \in V : \frac{p_v(u)}{D_{uu}} > \beta \right\}. \tag{2.3}$$

To determine $\widehat{C}$, one computes $S_{\beta,v}$ over all $\beta \in (L, U)$ (where the range (L, U) is user-specified), and then outputs the cluster estimate $\widehat{C} = S_{\beta^*,v}$ with minimum normalized cut. For a set $C \subseteq V$ with complement $C^c = V \backslash C$, the *cut* and *volume* are respectively,

$$\mathrm{cut}(C; G) := \sum_{u \in C} \sum_{v \in C^c} \mathbf{1}\{(u, v) \in E\}, \quad \mathrm{vol}(C; G) := \sum_{u \in C} \sum_{v \in V} \mathbf{1}\{(u, v) \in E\}, \tag{2.4}$$

and the *normalized cut* of C is

$$\Phi(C; G) := \frac{\mathrm{cut}(C; G)}{\min\{\mathrm{vol}(C; G), \mathrm{vol}(C^c; G)\}}. \tag{2.5}$$

2.0.2 Worst-case guarantees for PPR clustering

As mentioned, to date most analysis of local clustering has focused on worst-case guarantees, defined with respect to functionals of an a priori fixed graph $G = (V, E)$. For instance, Andersen et al. [2006] analyze the normalized cut of the cluster estimate $\widehat{C}$ output by PPR, showing that when PPR is appropriately seeded within a candidate cluster $C \subseteq V$, the normalized cut $\Phi(\widehat{C}; G)$ is upper bounded by (a constant times) $\sqrt{\Phi(C; G)}$. Zhu et al. [2013] build on this: they introduce a second functional, the *conductance* $\Psi(G)$, defined as

$$\Psi(G) := \min_{S \subseteq V} \Phi(S; G), \tag{2.6}$$

and show that if $\Phi(C; G)$ is much smaller than $\Psi(G[C])^2$—where $G[C] = (C, E \cap (C \times C))$ is the subgraph of G induced by C— then (in addition to having a small normalized cut) the cluster estimate $\widehat{C}$ approximately recovers C. Our own analysis builds on that of Zhu et al. [2013], and we give a more detailed summary of their results in Section 2.1. For now, we merely reiterate that the conclusions of Andersen et al. [2006], Zhu et al. [2013] cannot be straightforwardly applied to our statistical setting, where the input data are random samples $\{x_1, \ldots, x_n\}$ drawn from a distribution $\mathbb{P}$, the graph G is a random neighborhood graph formed by the user, and the candidate cluster is a set $\mathcal{C} \subset \mathbb{R}^d$.[1]

[1] Throughout, we use calligraphic notation to refer to subsets of $\mathbb{R}^d$.

2.0.3 PPR on a neighborhood graph

We now formally describe the statistical setting in which we operate, as well as the method we will study: PPR on a neighborhood graph. Let $X = \{x_1, \ldots, x_n\}$ be samples drawn i.i.d. from a distribution $\mathbb{P}$ on $\mathbb{R}^d$. We will assume throughout that $\mathbb{P}$ has a density f with respect to the Lebesgue measure ν on $\mathbb{R}^d$. 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 $i \neq j$ and $\|x_i - x_j\| \leq r$, where $\|\cdot\|$ is the Euclidean norm. Once the neighborhood graph $G_{n,r}$ is formed, the PPR vector p_v is then computed over $G_{n,r}$, with a resulting cluster estimate $\widehat{C} \subseteq X$. The precise PPR algorithm we analyze 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 estimate $\widehat{C} \subseteq V$.

1: Form the neighborhood graph $G_{n,r}$.
2: Compute the PPR vector $p_v = p(v, \alpha; G_{n,r})$ as in (2.1).
3: For $\beta \in (L, U)$, compute sweep cuts S_β as in (2.3).[2]
4: Return the cluster $\widehat{C} = S_{\beta^*}$, where

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

2.0.4 Cluster accuracy

We need a metric to assess the accuracy with which $\widehat{C}$ estimates the candidate cluster $\mathcal{C}$. One commonly used metric is the misclassification error, i.e. the size of the symmetric set difference between $\widehat{C}$ and the empirical cluster $\mathcal{C}[X] = \mathcal{C} \cap X$ [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 misclassified points according to their degree in $G_{n,r}$. To keep things simple, for a given set $S \subseteq X$ we write $\mathrm{vol}_{n,r}(S) := \mathrm{vol}(S; G_{n,r})$.

Definition 2.0.1. For an estimator $\widehat{C} \subseteq X$ and a set $\mathcal{C} \subseteq \mathbb{R}^d$, their symmetric set difference is

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

Furthermore, we denote the volume of the symmetric set difference by

$$\Delta(\widehat{C}, \mathcal{C}[X]) := \mathrm{vol}_{n,r}(\widehat{C} \vartriangle \mathcal{C}[X]).$$

2.0.5 Population normalized cut, conductance, and local spread

Next we define three population-level functionals of $\mathcal{C}$—the normalized cut $\Phi_{\mathbb{P},r}(\mathcal{C})$, conductance $\Psi_{\mathbb{P},r}(\mathcal{C})$, and local spread $s_{\mathbb{P},r}(\mathcal{C})$—which govern the volume of the symmetric set difference $\Delta(\widehat{C}, \mathcal{C}[X])$. Let the population-level *cut* of $\mathcal{C}$ be the expectation (up to a rescaling) of $\mathrm{cut}_{n,r}(\mathcal{C}[X]) := \mathrm{cut}(\mathcal{C}[X]; G_{n,r})$, and likewise let the population-level *volume* of $\mathcal{C}$ be the expectation (up to a rescaling) of $\mathrm{vol}_{n,r}(\mathcal{C}[X]) := \mathrm{vol}(\mathcal{C}[X]; G_{n,r})$; i.e. let

$$\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}) := \int_{\mathcal{C}} \int_{\mathcal{C}^c} \mathbf{1}\{\|x - y\| \leq r\}\, d\mathbb{P}(y)\, d\mathbb{P}(x), \quad \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}) := \int_{\mathcal{C}} \int_{\mathbb{R}^d} \mathbf{1}\{\|x - y\| \leq r\}\, d\mathbb{P}(y)\, d\mathbb{P}(x),$$

[2] Technically speaking, for each $\beta \in (L, U) \cap \{p_v(u)/D_{uu} : u \in V\}$.

where $\mathcal{C}^c := \mathbb{R}^d \setminus \mathcal{C}$. Also let $\deg_{\mathbb{P},r}(x) := \int_{\mathbb{R}^d} \mathbf{1}\{\|y - x\| \leq r\}\, d\mathbb{P}(y)$ to be the expected degree of x in $G_{n,r}$.

Definition 2.0.2 (Population normalized cut)**.** For a set $\mathcal{C} \subset \mathbb{R}^d$, distribution $\mathbb{P}$ and radius $r > 0$, the *population normalized cut* is

$$\Phi_{\mathbb{P},r}(\mathcal{C}) := \frac{\mathrm{cut}_{\mathbb{P},r}(\mathcal{C})}{\min\{\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}), \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^c)\}}. \tag{2.7}$$

Let $\widetilde{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot | x \in \mathcal{C})$ be the conditional distribution of x, i.e let $\widetilde{\mathbb{P}}(\mathcal{S}) = \widetilde{\mathbb{P}}(\mathcal{S} \cap \mathcal{C})/\widetilde{\mathbb{P}}(\mathcal{C})$ for measurable sets $\mathcal{S} \subseteq \mathbb{R}^d$.

Definition 2.0.3 (Population conductance)**.** For a set $\mathcal{C} \subset \mathbb{R}^d$, distribution $\mathbb{P}$ and radius $r > 0$, the *population conductance* is

$$\Psi_{\mathbb{P},r}(\mathcal{C}) = \inf_{\mathcal{S} \subseteq \mathcal{C}} \Phi_{\widetilde{\mathbb{P}},r}(\mathcal{S}). \tag{2.8}$$

Definition 2.0.4 (Population local spread)**.** For a set $\mathcal{C} \subset \mathbb{R}^d$, distribution $\mathbb{P}$ and radius $r > 0$, the *population local spread* is

$$s_{\mathbb{P},r}(\mathcal{C}) := \min_{x \in \mathcal{C}} \left\{ \frac{\left(\deg_{\widetilde{\mathbb{P}},r}(x)\right)^2}{\mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C})} \right\}, \tag{2.9}$$

It is quite natural that $\Phi_{\mathbb{P},r}(\mathcal{C})$ and $\Psi_{\mathbb{P},r}(\mathcal{C})$ should help quantify the role geometry plays in local spectral clustering. Indeed, by construction these functionals are quite obviously the population-level analogues of the empirical quantities $\Phi_{n,r}(\mathcal{C}[X]) := \Phi(\mathcal{C}[X]; G_{n,r})$ and $\Psi_{n,r}(\mathcal{C}[X]) := \Psi(G_{n,r}[\mathcal{C}[X]])$, and as we have already mentioned, these empirical quantities in turn suffice to upper bound the volume of the symmetric set difference. For this reason, similar population level functionals are used by [Shi et al., 2009, Schiebinger et al., 2015, García Trillos et al., 2019b] in the analysis of *global* spectral clustering in a statistical context. We will comment more on the relationship between these works and our own results in Section 2.0.7.

The role played by $s_{\mathbb{P},r}(\mathcal{C})$ is somewhat less obvious. For now, we mention only that it plays an essential part in obtaining tight bounds on the mixing time of a particular random walk that is closely related to the PPR vector, and defer further discussion until later in Section 2.1.

2.0.6 Main Results

We now informally state our two main upper bounds, regarding the recovery of a generic cluster $\mathcal{C}$, and a density cluster $\mathcal{C}_\lambda$. Theorem 1 informally summarizes the first of our main results (formally stated in Theorem 3) regarding the recovery of a generic cluster $\mathcal{C}$.

Theorem 1 (Informal)**.** *Let $\mathcal{C} \subset \mathbb{R}^d$ and $\mathbb{P}$ satisfy appropriate regularity conditions, and suppose Algorithm 1 is well-initialized with respect to $\mathcal{C}$. For all $n \in \mathbb{N}$ sufficiently large, with high probability it holds that*

$$\frac{\Delta(\widehat{C}, \mathcal{C}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}[X])} \leq c \cdot \Phi_{\mathbb{P},r}(\mathcal{C}) \cdot \left(\frac{\log(1/s_{\mathbb{P},r}(\mathcal{C}))}{\Psi_{\mathbb{P},r}(\mathcal{C})} \right)^2.$$

(In the above, and throughout, c stands for a universal constant that may change from line to line.) Put even more succinctly, we find that $\Delta(\widehat{C}, \mathcal{C}[X])$ is small when $\Phi_{\mathbb{P},r}(\mathcal{C})$ is small relative to $\left(\Psi_{\mathbb{P},r}(\mathcal{C})/\log(1/s_{\mathbb{P},r}(\mathcal{C}))\right)^2$. To the best of our knowledge, this gives the first population-level guarantees for local clustering in the nonparametric statistical context.

Theorem 2 informally summarizes the second of our main results (formally stated in Theorem 4) regarding the recovery of a λ-density cluster $\mathcal{C}_\lambda$ by PPR. For reasons that we explain later in Section 2.2, our cluster recovery statement will actually be with respect to the σ-thickened set $\mathcal{C}_{\lambda,\sigma} := \{x \in \mathbb{R}^d : \mathrm{dist}(x, \mathcal{C}_\lambda) < \sigma\}$,

for a given $\sigma > 0$. The upper bound we establish is a function of various parameters that measure the conditioning of both the density cluster $\mathcal{C}_{\lambda,\sigma}$ and density f for recovery by PPR. We assume that $\mathcal{C}_{\lambda,\sigma}$ is the image of a convex set $\mathcal{K}$ of finite diameter $\mathrm{diam}(\mathcal{K}) \leq \rho < \infty$ under an Lipschitz, measure-preserving mapping g, with Lipschitz constant L. We also assume that f is bounded away from 0 and ∞ on $\mathcal{C}_{\lambda,\sigma}$,

$$0 < \lambda_\sigma \leq f(x) \leq \Lambda_\sigma < \infty \quad \text{for all } x \in \mathcal{C}_{\lambda,\sigma},$$

while satisfying the following low-noise condition:

$$\inf_{y \in \mathcal{C}_{\lambda,\sigma}} f(y) - f(x) \geq \theta \cdot \mathrm{dist}(x, \mathcal{C}_{\lambda,\sigma})^\gamma \quad \text{for all } x \text{ such that } \mathrm{dist}(x, \mathcal{C}_{\lambda,\sigma}) \leq r.$$

(Here $\mathrm{dist}(x, \mathcal{C}) := \inf_{y \in \mathcal{C}} \|y - x\|$.)

Theorem 2 (Informal). *Let $\mathcal{C}_\lambda \subset \mathbb{R}^d$ be a λ-density cluster of a distribution $\mathbb{P}$ that satisfies appropriate regularity conditions, and suppose Algorithm 1 is well-initialized with respect to $\mathcal{C}_{\lambda,\sigma}$. For all $n \in \mathbb{N}$ sufficiently large, with high probability it holds that*

$$\frac{\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma})} \leq c \cdot d^4 \cdot \frac{L^2\rho^2}{\sigma r} \cdot \frac{\Lambda_\sigma^2 \lambda(\lambda - \theta \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma^4} \cdot \log^2\left(\frac{\Lambda_\sigma^{2/d} L\rho}{\lambda_\sigma^{2/d} 2r}\right). \tag{2.10}$$

Equation (2.10) reveals the separate roles played by geometry and density in the ability of PPR to recover a density cluster. The parameters L, ρ and σ capture whether $\mathcal{C}_{\lambda,\sigma}$ is well-conditioned (short and fat) or poorly-conditioned (long and thin) for recovery by PPR. Likewise, the parameters $\lambda_\sigma, \Lambda_\sigma, \gamma$ and θ measure whether f is well-conditioned (approximately uniform over the density cluster, and having thin tails outside the density cluster) or poorly conditioned (vice versa). Theorem 2 tells us that if the thickened density cluster $\mathcal{C}_{\lambda,\sigma}$ is well-conditioned—i.e. $L^2\rho^2/(\sigma r) \approx 1$—and the density f is well-conditioned near $\mathcal{C}_{\lambda,\sigma}$—i.e. $\Lambda_\sigma \approx \lambda \approx \lambda_\sigma$ and $\lambda - \theta r^\gamma/(\gamma+1)$ is much less than λ_σ—then PPR will approximately recover $\mathcal{C}_{\lambda,\sigma}$.

2.0.7 Related Work

We now summarize some related work (in addition to the background already given above), regarding the theory of spectral clustering, and of density cluster recovery.

Spectral clustering. In the stochastic block model (SBM), arguably one of the simplest models of network formation, edges between nodes independently occur with probability based on a latent community membership. In the SBM, the ability of spectral algorithms to perform clustering—or community detection—is well-understood, dating back to McSherry [2001] who gives conditions under which the entire community structure can be recovered. In more recent work, Rohe et al. [2011] upper bound the fraction of nodes misclassified by a spectral algorithm for the high-dimensional (large number of blocks) SBM, and Lei and Rinaldo [2015] extend these results to the sparse (low average degree) regime. Relatedly, Clauset et al. [2008], Balakrishnan et al. [2011], Li et al. [2018], analyze the misclassification rate when the block model exhibits some hierarchical structure. The framework we consider, in which nodes correspond to data points sampled from an underlying density, and edges between nodes are formed based on geometric proximity, is quite different than the SBM, and therefore so is our analysis.

In general, the study of spectral algorithms on neighborhood graphs has been focused on establishing asymptotic convergence of eigenvalues and eigenvectors of certain sample objects to the eigenvalues and eigenfunctions of corresponding limiting operators. Koltchinskii and Gine [2000] establish convergence of spectral projections of the adjacency matrix to a limiting integral operator, with similar results obtained using simplified proofs in Rosasco et al. [2010]. von Luxburg et al. [2008] studies convergence of eigenvectors of the Laplacian matrix for a neighborhood graph of fixed radius. Belkin and Niyogi [2007] and García Trillos and Slepčev [2018a] extend these results to the regime where the radius $r \to 0$ as $n \to \infty$.

These results are of fundamental importance. However, they remain silent on the following natural question: do the spectra of these continuum operators induce a partition of the sample space which is "good" in some sense? Shi et al. [2009], Schiebinger et al. [2015], García Trillos et al. [2019b], Hoffmann et al. [2019] address this question, showing that spectral algorithms will recover the latent labels in certain well-conditioned nonparametric mixture models. These works are probably the most similar to our own: the conditioning of these mixture models depend on population-level functionals resembling the population-level normalized cut and conductance introduced above, and the resulting bounds on the error of spectral clustering are comparable to those we establish in Theorem 3. However, these results focus on global rather than local methods, and impose global rather than local conditions on $\mathbb{P}$. Moreover, they do not explicitly consider recovery of density clusters, which is an important concern of our work. We comment further on the relationship between our results and these works after Theorem 3.

Density clustering. For a given threshold $\lambda \in (0, \infty)$, let $\mathbb{C}_f(\lambda)$ denote the connected components of the density upper level set $\{x \in \mathbb{R}^d : f(x) \geq \lambda\}$. In the density clustering problem, initiated by Hartigan [1975], the goal is to recover $\mathbb{C}_f(\lambda)$. By now, density clustering (and the related problem of level-set estimation) are quite well-understood. For instance, Polonik [1995], Rigollet and Vert [2009], Rinaldo and Wasserman [2010], Steinwart [2015] study density clustering under the symmetric set difference metric, Tsybakov [1997], Singh et al. [2009], Jiang [2017] describe minimax optimal level-set and cluster estimators under Hausdorff loss, and Hartigan [1981], Chaudhuri and Dasgupta [2010], Kpotufe and von Luxburg [2011], Balakrishnan et al. [2013a], Steinwart et al. [2017], Wang et al. [2019] consider consistent estimation of the cluster tree $\{\mathbb{C}_f(\lambda) : \lambda \in (0, \infty)\}$.

We emphasize that our goal is not to improve on these results, nor to offer a better algorithm for density clustering. Indeed, seen as a density clustering algorithm, PPR has none of the optimality guarantees found in the aforementioned works. Rather, we hope to better understand the implications of our general theory by applying it within an already well-studied framework. We should also note that since we study a local algorithm, our interest will be in a local version of the density clustering problem, where the goal is to recover a single density cluster $\mathcal{C}_\lambda \in \mathbb{C}_f(\lambda)$.

2.0.8 Organization

We now outline the rest of the paper.

- In Section 2.1, we derive bounds on the error of PPR as a function of sample normalized cut, conductance, and local spread. We then show that under certain conditions the sample normalized cut, conductance, and local spread are close to their population-level counterparts, with high probability for sufficient number of samples. As a result, we obtain an upper bound on $\Delta(\widehat{C}, \mathcal{C}[X])/\mathrm{vol}_{n,r}(\mathcal{C}[X])$ purely in terms of these population functionals (Theorem 4).
- In Section 2.2, we focus on the special case where the candidate cluster $\mathcal{C} = \mathcal{C}_\lambda$ is a λ-*density cluster*—that is, a connected component of the upper level set $\{x : f(x) \geq \lambda\}$. We derive bounds on the population normalized cut, conductance, and local spread of the density cluster, which depend on λ as well as some other natural parameters. This leads to an upper bound on the symmetric set difference between $\widehat{C}$ and the λ-density cluster. (Theorem 4).
- In Section 2.3, we prove a negative result: we give a hard distribution $\mathbb{P}$ with corresponding density cluster $\mathcal{C}_\lambda$ for which the symmetric set difference between $\widehat{C}$ and the λ-density cluster is provably large.
- In Section 2.4 we empirically investigate some of our conclusions, before ending with some discussion in Section 2.5.

2.1 Recovery of a generic cluster with PPR

In the main result (Theorem 3) of this section, we give a high probability upper bound on $\Delta(\widehat{C}, \mathcal{C}[X])$, in terms of the population normalized cut $\Phi_{\mathbb{P},r}(\mathcal{C})$ and conductance $\Psi_{\mathbb{P},r}(\mathcal{C})$. We build to this theorem slowly, giving new structural results in two distinct directions. First, we build on some previous work (mentioned in the introduction) to relate $\Delta(\widehat{C}, \mathcal{C}[X])$ to the sample normalized cut $\Phi_{n,r}(\mathcal{C}[X])$, conductance $\Psi_{n,r}(\mathcal{C}[X])$, and local spread $s_{n,r}(\mathcal{C}[X]) := s\big(G_{n,r}[\mathcal{C}[X]]\big)$. Second, we argue that when n is large, each of these graph functionals can be bounded by their population-level analogues.

2.1.1 PPR cluster recovery: the fixed graph case

When PPR is run on a fixed graph $G = (V, E)$ with the goal of recovering a candidate cluster $C \subset V$, Zhu et al. [2013] provide the sharpest known bounds on the volume of the symmetric set difference between the cluster estimate $\widehat{C}$ and candidate cluster C. Since these results will play a major part in our analysis, in Lemma 1 we restate them for the convenience of the reader.[3]

In their most general form, the results of Zhu et al. [2013] depend on the mixing time of a lazy random walk over the induced subgraph $G[C]$. The *mixing time* of a lazy random walk over a graph G is

$$\tau_\infty(G) := \min\left\{ t : \frac{\pi(u) - q_v^{(t)}(u)}{\pi(u)} \leq \frac{1}{4}, \text{ for all } u, v \in V \right\}; \tag{2.11}$$

here $q_v^{(t)} := e_v W^t$ is the distribution of a lazy random walk over G that originates at node v and runs for t steps, and $\pi := \lim_{t\to\infty} q_v^{(t)}$ is the limiting distribution of $q_v^{(t)}$.

Lemma 1 (Lemma 3.4 of Zhu et al. [2013])**.** *For a set $C \subseteq V$, suppose that*

$$\alpha \leq \min\left\{ \frac{1}{45}, \frac{1}{2\tau_\infty(G[C])} \right\}, \quad \beta \leq \frac{1}{5\mathrm{vol}(C; G)} \tag{2.12}$$

Then there exists a set $C^g \subset C$ with $\mathrm{vol}(C^g; G) \geq \frac{1}{2}\mathrm{vol}(C^g; G)$ such that for any $v \in C^g$, the sweep cut $S_{\beta,v}$ satisfies

$$\mathrm{vol}(S_{\beta,v} \vartriangle C; G) \leq 6\frac{\Phi(C; G)}{\alpha\beta}. \tag{2.13}$$

The upper bound in (2.13) does not obviously depend on the conductance $\Psi(G[C])$. However, as Zhu et al. [2013] point out, letting $\pi_{\min}(G) := \min_{u\in V}\{\pi(u)\}$, it follows from Cheeger's inequality [Chung, 1997] that

$$\tau_\infty(G) \leq \frac{\log(1/\pi_{\min}(G))}{\Psi(G)^2}. \tag{2.14}$$

Therefore, setting (for instance) $\alpha = \frac{\Psi(G[C])^2}{2\log(1/\pi_{\min}(G))}$ and $\widehat{C} = S_{\beta_0,v}$ for $\beta_0 = \frac{1}{5\mathrm{vol}(C;G)}$, we obtain from (2.13) that

$$\frac{\mathrm{vol}(C \vartriangle \widehat{C}; G)}{\mathrm{vol}(C; G)} \leq 60\frac{\Phi(C; G)\log\big(1/\pi_{\min}(G[C])\big)}{\Psi(G[C])^2}. \tag{2.15}$$

2.1.2 Improved bounds on mixing time

Having reviewed the conclusions of Zhu et al. [2013], we return now to our own setting, where the data is not a fixed graph G but instead random samples $\{x_1, \ldots, x_n\}$, and our goal is to recover a candidate

[3] Lemma 1 improves on Lemma 3.4 of Zhu et al. [2013] by some constant factors, and for completeness we prove Lemma 1 in the Appendix. Nevertheless, to be clear the essential idea of Lemma 1 is no different than that of Zhu et al. [2013], and we do not claim any novelty.

cluster $\mathcal{C} \subset \mathbb{R}^d$. Ideally, we would like to apply (2.15) with $C = \mathcal{C}[X]$ and $G = G_{n,r}$, replace $\Phi_{n,r}(\mathcal{C}[X])$ and $\Psi_{n,r}(\mathcal{C}[X])$ by $\Phi_{\mathbb{P},r}(\mathcal{C})$ and $\Psi_{\mathbb{P},r}(\mathcal{C})$ inside (2.15), and thereby obtain an upper bound on $\Delta(\widehat{C}; \mathcal{C}[X])$ that depends only on $\mathbb{P}$ and $\mathcal{C}$. Unfortunately, however, there is a catch: when the graph $G = G_{n,r}$ and the candidate cluster $C = \mathcal{C}[X]$, as $n \to \infty$ the sample normalized cut $\Phi_{n,r}(\mathcal{C}[X])$ and conductance $\Psi_{n,r}(\mathcal{C}[X])$ each converge to their population-level analogues, but $\pi_{\min}\big(G_{n,r}\big[\mathcal{C}[X]\big]\big) \asymp 1/n$.[4] Therefore the right hand side of (2.15) diverges at a $\log n$ rate, rendering (2.15) a vacuous upper bound whenever the number of samples is sufficiently large.

To fix this, in Proposition 1 we improve the upper bound on mixing time given in (2.14). Specifically, in (2.16) the "start penalty" of $\log(1/\pi_{\min}(G))$ is replaced by $\log(1/s(G))$, where $s(G)$ is the graph local spread, defined as

$$s(G) := d_{\min}(G) \cdot \pi_{\min}(G),$$

for $d_{\min}(G) = \min_{u \in V}\big\{\deg(u; G)\big\}$, and likewise $d_{\max}(G) = \max_{u \in V}\big\{\deg(u; G)\big\}$. Note that $s(G) \geq \pi_{\min}(G)$.

Proposition 1. *Assume $d_{\max}(G)/d_{\min}(G)^2 \leq 1/16$. Then,*

$$\tau_{\infty}(G) \leq \frac{13}{\ln(2)} \left(\frac{\ln\big(8/s(G)\big)}{\Psi(G)} \right)^2 \tag{2.16}$$

While Proposition 1 can be applied to *any* graph G (as long as the ratio of maximum degree to squared minimum degree is at most $1/16$), it is particularly useful when G is a geometric graph. In the case where $G = G_{n,r}\big[\mathcal{C}[X]\big]$ for a fixed radius r, the minimum degree $d_{\min}\big(G_{n,r}\big[\mathcal{C}[X]\big]\big) \asymp n$, and therefore $s\big(G_{n,r}\big[\mathcal{C}[X]\big]\big) \asymp 1$. We give a precise upper bound on $s\big(G_{n,r}\big[\mathcal{C}[X]\big]\big)$ in Proposition 2, which does not grow with n, and in combination with Proposition 1 this allows us to remove the unwanted $\log n$ factor from the upper bound in (2.15).

The local spread $s(G)$ plays an intuitive role in the analysis of mixing time. Indeed, in any graph G sufficiently small sets are expanders—that is, if a set $R \subseteq V$ has cardinality less than the minimum degree, the normalized cut $\Phi(R; G)$ will be much larger than the conductance $\Psi(G)$. As a consequence, a random walk over G will rapidly mix over all small sets R, and in our analysis of the mixing time we may therefore "pretend" that the random walk was given a warm start over a larger set S. The local spread $s(G)$ simply delineates small sets R from larger sets S. Of course, the proof of Proposition 1 requires a substantially more careful analysis, and—as with the proofs of all results in this paper—it is deferred to the appendix.

2.1.3 Sample-to-population results

In Propositions 2 and 3, we establish high probability bounds on the sample normalized cut, conductance, and local spread in terms of their population-level analogues. To establish these bounds, we impose the following regularity conditions on $\widetilde{\mathbb{P}}$ and $\mathcal{C}$.

(A1) The distribution $\widetilde{\mathbb{P}}$ has a density $\widetilde{f} : \mathcal{C} \to (0, \infty)$ with respect to Lebesgue measure. There exist $0 < f_{\min} \leq 1 \leq f_{\max} < \infty$ for which

$$(\forall x \in \mathcal{C}) \quad f_{\min} \leq \widetilde{f}(x) \leq f_{\max}$$

(A2) The candidate cluster $\mathcal{C} \subseteq \mathbb{R}^d$ is a bounded, connected, open set. If $d \geq 2$ then it has a Lipschitz boundary.

In what follows, we use $b_1, b_2, \ldots$ and $B_1, B_2, \ldots$ to refer to positive constants that may depend on $\mathbb{P}$, $\mathcal{C}$, and r, but do not depend on n or δ. We explicitly keep track of all constants in our proofs.

[4] For sequences (a_n) and (b_n), we say $a_n \asymp b_n$ if there exists a constant $c \geq 1$ such that $a_n/c \leq b_n \leq c a_n$ for all $n \in \mathbb{N}$.

Proposition 2. *Fix $\delta \in (0, 1/3)$. Suppose $\mathcal{C}$ and $\widetilde{\mathbb{P}}$ satisfy (A1) and (A2). Then each of the following statements hold.*

- *With probability at least $1 - 3\exp\{-b_1\delta^2 n\}$,*

$$\Phi_{n,r}(\mathcal{C}[X]) \leq (1 + 3\delta)\Phi_{\mathbb{P},r}(\mathcal{C}). \tag{2.17}$$

- *For any $n \in \mathbb{N}$ for which*

$$\frac{1}{n} \leq \delta \cdot \frac{2\mathbb{P}(\mathcal{C})}{3} \tag{2.18}$$

the following inequality holds with probability at least $1 - (n+2)\exp\{-b_2\delta^2 n\}$:

$$s_{n,r}(\mathcal{C}[X]) \geq (1 - 4\delta)s_{\mathbb{P},r}(\mathcal{C}). \tag{2.19}$$

Let $p_d := 1/2$ if $d = 1$, $p_d := 3/4$ if $d = 2$, and otherwise $p_d := 1/d$ for $d \geq 3$.

Proposition 3. *Fix $\delta \in (0, 1/2)$. Suppose $\widetilde{\mathbb{P}}$ and $\mathcal{C}$ satisfy (A1) and (A2). Then for any $n \in \mathbb{N}$ satisfying*

$$B_1 \frac{(\log n)^{p_d}}{\min\{n^{1/2}, n^{1/d}\}} \leq \delta, \tag{2.20}$$

the following inequality holds with probability at least $1 - B_2/n - (n+1)\exp\{-b_3 n\}$:

$$\Psi_{n,r}(\mathcal{C}[X]) \geq (1 - 2\delta)\Psi_{\mathbb{P},r}(\mathcal{C}). \tag{2.21}$$

A word on the proof techniques: the upper bound in (2.17) follows by applying Bernstein's inequality to control the deviations of $\mathrm{cut}_{n,r}(\mathcal{C}[X])$, $\mathrm{vol}_{n,r}(\mathcal{C}[X])$ and $\mathrm{vol}_{n,r}(\mathcal{C}^c[X])$ around their expectations (noting that each of these is an order-2 U-statistic). To prove the lower bound (2.19), we require a union bound to control the minimum degree $d_{\min}\big(G_{n,r}\big[\mathcal{C}[X]\big]\big)$, but otherwise the proof is similarly straightforward. On the other hand, the proof of (2.21) is considerably more complicated. Our proof relies on the recent results of [García Trillos and Slepcev, 2015], who upper bound the L^∞-optimal transport distance between the empirical measure $\mathbb{P}_n$ and $\mathbb{P}$. For further details, we refer to Appendix A.2.4, where we prove Proposition 3, as well as [García Trillos et al., 2016], who establish the asymptotic convergence of the sample conductance as $n \to \infty$ and $r \to 0$.

2.1.4 Cluster recovery

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 $\delta \in (0, 1/4)$ and a candidate cluster $\mathcal{C} \in \mathbb{R}^d$, we require that some of the tuning parameters of Algorithm 1 be chosen within specific ranges,

$$\begin{aligned} \alpha &\in \Big[(1-4\delta)^2, (1-2\delta)^2\Big) \cdot \frac{\alpha_{\mathbb{P},r}(\mathcal{C}, \delta)}{2} \\ (L, U) &\subseteq \Big(\frac{1}{5(1+2\delta)}, \frac{1}{5(1+\delta)}\Big) \cdot \frac{1}{n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C})}. \end{aligned} \tag{2.22}$$

where

$$\alpha_{\mathbb{P},r}(\mathcal{C}, \delta) := \frac{\ln(2)}{13} \cdot \frac{\Psi^2_{\mathbb{P},r}(\mathcal{C})}{\ln^2\Big(\frac{8}{(1-4\delta)s_{\mathbb{P},r}(\mathcal{C})}\Big)}. \tag{2.23}$$

Definition 2.1.1. If the input parameters to Algorithm 1 satisfy (2.22) for some $\mathcal{C} \subseteq \mathbb{R}^d$ and $\delta \in (0, 1/4)$, we say the algorithm is *δ-well-initialized* with respect to $\mathcal{C}$.

Of course, in practice it is not feasible to set tuning parameters based on the underlying (unknown) distribution $\mathbb{P}$ and candidate cluster $\mathcal{C}$. Typically, one runs PPR over some range of tuning parameter values and selects the cluster which has the smallest normalized cut.

By combining Lemma 1 and Propositions 1-3, we obtain an upper bound on $\Delta(\widehat{C}, \mathcal{C}[X])$ that depends solely on the distribution $\mathbb{P}$ and candidate cluster $\mathcal{C}$. To ease presentation, we introduce the *condition number*, defined for a given $\mathcal{C} \subseteq \mathbb{R}^d$ and $\delta \in (0, 1/4)$ as

$$\kappa_{\mathbb{P},r}(\mathcal{C}, \delta) := \frac{(1+3\delta)(1+2\delta)}{(1-4\delta)^2(1-\delta)} \cdot \frac{\Phi_{\mathbb{P},r}(\mathcal{C})}{\alpha_{\mathbb{P},r}(\mathcal{C}, \delta)}. \tag{2.24}$$

Theorem 3. *Fix $\delta \in (0, 1/4)$. Suppose $\widetilde{\mathbb{P}}$ and $\mathcal{C}$ satisfy (A1) and (A2). Then for any $n \in \mathbb{N}$ which satisfies (2.18), (2.20), and*

$$\frac{(1+\delta)}{(1-\delta)^4} \cdot B_3 \leq n, \tag{2.25}$$

the following statement holds with probability at least $1 - B_2/n - 4\exp\{-b_1\delta^2 n\} - (2n+2)\exp\{-b_2\delta^2 n\} - (n+1)\exp\{-b_3 n\}$: there exists a set $\mathcal{C}[X]^g \subseteq \mathcal{C}[X]$ of large volume, $\mathrm{vol}_{n,r}(\mathcal{C}[X]^g) \geq \mathrm{vol}_{n,r}(\mathcal{C}[X])/2$, such that if Algorithm 1 is δ-well-initialized with respect to $\mathcal{C}$, and run with any seed node $v \in \mathcal{C}[X]^g$, then the PPR estimated cluster $\widehat{C}$ satisfies

$$\frac{\Delta(\widehat{C}; \mathcal{C}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}[X])} \leq 60 \cdot \kappa_{\mathbb{P},r}(\mathcal{C}, \delta). \tag{2.26}$$

We now make some remarks:

- It is useful to compare Theorem 3 with what is already known regarding *global* spectral clustering in the context of nonparametric statistics. Schiebinger et al. [2015] consider the following variant of spectral clustering: first embed the data X into $\mathbb{R}^k$ using the bottom k eigenvectors of the degree-normalized Laplacian $I - D^{-1/2}AD^{-1/2}$, and then partition the embedded data into estimated clusters $\widehat{C}_1, \ldots, \widehat{C}_k$ using k-means clustering. They derive error bounds on the misclassification error that depend on a difficulty function $\varphi(\mathbb{P})$. In our context, where the goal is to successfully distinguish $\mathcal{C}$ and $\mathcal{C}^c$ and thus $k = 2$, this difficulty function is roughly

$$\varphi(\mathbb{P}) \approx \sqrt{\Phi_{\mathbb{P},r}(\mathcal{C})} \cdot \max\left\{ \frac{1}{\Psi_{\mathbb{P},r}(\mathcal{C})^2}; \frac{1}{\Psi_{\mathbb{P},r}(\mathcal{C}^c)^2} \right\}. \tag{2.27}$$

  We point out two ways in which (2.26) is a tighter bound than (2.27). First, (2.27) depends on $\Psi_{\mathbb{P},r}(\mathcal{C}^c)$ in addition to $\Psi_{\mathbb{P},r}(\mathcal{C})$, and is thus a useful bound only if $\mathcal{C}^c$ and $\mathcal{C}$ are both internally well-connected. In contrast (2.26) depends only on $\Psi_{\mathbb{P},r}(\mathcal{C})$, and is thus a useful bound if $\mathcal{C}$ has small conductance, regardless of the conductance of $\mathcal{C}^c$. This is intuitive: PPR is a local rather than global algorithm, and as such the analysis requires only local rather than global conditions. Second, (2.27) depends on $\sqrt{\Phi_{\mathbb{P},r}(\mathcal{C})}$ rather than $\Phi_{\mathbb{P},r}(\mathcal{C})$, and since $\Phi_{\mathbb{P},r}(\mathcal{C}) \leq 1$ this results in a weaker bound. [Schiebinger et al., 2015] provide experiments suggesting that the linear, rather than square-root, dependence is correct, and we theoretically confirm this in the local clustering setup. Of course, on the other hand (2.26) depends on $\log^2(1/s_{\mathbb{P},r}(\mathcal{C}))$, which is due to the locally-biased nature of the PPR algorithm, and does not appear in (2.27).

- Although Theorem 3 is stated with respect to the exact PPR vector p_v, for a sufficiently small choice of ε, the application of (2.2) within the proof of Theorem 3 leads to an analogous result which holds for the aPPR vector $p_v^{(\varepsilon)}$. We formally state and prove this fact in Appendix A.5.

2.2 Recovery of a density cluster with PPR

We now apply the general theory established in the last section to the special case where $\mathcal{C} = \mathcal{C}_\lambda$ is a λ-density cluster—that is, a connected component of the upper level set $\{x \in \mathbb{R}^d : f(x) \geq \lambda\}$. In Section 2.3, we also derive a lower bound, giving a "hard problem" for which PPR will provably fail to recover a density cluster. Together, these results can be summarized as follows: PPR recovers a density cluster $\mathcal{C}_\lambda$ if and only if both $\mathcal{C}_\lambda$ and f are well-conditioned, meaning that $\mathcal{C}_\lambda$ is not too long and thin, and that f is approximately uniform inside $\mathcal{C}_\lambda$ while satisfying a low-noise condition near its boundary.

2.2.1 Recovery of well-conditioned density clusters

All results on density clustering assume the density f satisfies some regularity conditions. A basic requirement is the need to avoid clusters which contain arbitrarily thin bridges or spikes, or more generally clusters which can be disconnected by removing a subset of (Lebesgue) measure 0, and thus may not be resolved by any finite number of samples. To rule out such problematic clusters, we follow the approach of Chaudhuri and Dasgupta [2010], who assume the density is lower bounded on a thickened version of $\mathcal{C}_\lambda$, defined as $\mathcal{C}_{\lambda,\sigma} := \{x \in \mathbb{R}^d : \mathrm{dist}(x, \mathcal{C}) < \sigma\}$ for a given $\sigma > 0$. The point is that regardless of the dimension of $\mathcal{C}_\lambda$, the set $\mathcal{C}_{\lambda,\sigma}$ is full dimensional. Under typical uniform continuity conditions, the requirement that the density be lower bounded over $\mathcal{C}_{\lambda,\sigma}$ will be satisfied. Such continuity conditions can be weakened (see for instance Rinaldo and Wasserman [2010], Steinwart [2015]) but we do not pursue the matter further.

In summary, our goal is to obtain upper bounds on $\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])$, for some fixed λ and $\sigma > 0$. We have already derived upper bounds on the symmetric set difference of $\widehat{C}$ and a generic cluster $\mathcal{C}$ that depend on some population-level functionals $\mathcal{C}$. What remains is to analyze these population-level functionals in the specific case where the candidate cluster is $\mathcal{C}_{\lambda,\sigma}$. To carry out this analysis, we will need to impose some conditions, and for the rest of this section we will assume the following.

(A3) *Bounded density within cluster:* There exist constants $0 < \lambda_\sigma < \Lambda_\sigma < \infty$ such that

$$\lambda_\sigma \leq \inf_{x \in \mathcal{C}_{\lambda,\sigma}} f(x) \leq \sup_{x \in \mathcal{C}_{\lambda,\sigma}} f(x) \leq \Lambda_\sigma.$$

(A4) *Low noise density:* There exist $\theta \in (0, \infty)$ and $\gamma \in [0, 1]$ such that for any $x \in \mathbb{R}^d$ with $0 < \mathrm{dist}(x, \mathcal{C}_{\lambda,\sigma}) \leq \sigma$,

$$\inf_{y \in \mathcal{C}_{\lambda,\sigma}} f(y) - f(x) \geq \theta \cdot \mathrm{dist}(x, \mathcal{C}_{\lambda,\sigma})^\gamma.$$

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

(A5) *Lipschitz embedding:* There exists $g : \mathbb{R}^d \to \mathbb{R}^d$, $\rho \in (0, \infty)$ and $L \in [1, \infty)$ such that

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

(b) $\det(\nabla g(x)) = 1$ for all $x \in \mathcal{K}$, 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 $\mathcal{C}_{\lambda,\sigma}$ is the image of a convex set with finite diameter under a measure preserving, Lipschitz transformation.

For convenience only, we will also make the following assumption.

(A6) *Bounded volume:* The volume of $\mathcal{C}_{\lambda,\sigma}$ is no more than half the total volume of $\mathbb{R}^d$:

$$\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) \leq \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}^c).$$

This assumption implies that the normalized cut of $\mathcal{C}_{\lambda,\sigma}$ will be equal to the ratio of $\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$ to $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$.

Normalized cut, conductance, and local spread of a density cluster. In Lemma 2, Proposition 4, and Proposition 5, we give bounds on the population-level local spread, normalized cut, and conductance of $\mathcal{C}_{\lambda,\sigma}$. These bounds depend on the various geometric parameters just introduced.

Lemma 2. *Assume $\mathcal{C}_{\lambda,\sigma}$ satisfies Assumptions (A3) and (A5) for some $\lambda_\sigma, \Lambda_\sigma, \rho$ and L. Then,*

$$s_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) \geq \frac{1}{4} \cdot \frac{\lambda_\sigma^2}{\Lambda_\sigma^2} \cdot \left(\frac{2r}{\rho}\right)^d \cdot \left(1 - \frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\right) \tag{2.28}$$

Proposition 4. *Assume $\mathcal{C}_{\lambda,\sigma}$ satisfies Assumptions (A3), (A4) and (A6) for some $\lambda_\sigma, \Lambda_\sigma, \theta$, and γ, and additionally that $0 < r \leq \frac{\sigma}{4d}$. Then,*

$$\Phi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) \leq \frac{16}{9} \cdot \frac{dr}{\sigma} \cdot \frac{\lambda\left(\lambda_\sigma - \theta\frac{r^\gamma}{\gamma+1}\right)}{\lambda_\sigma^2} \tag{2.29}$$

Proposition 5. *Assume $\mathcal{C}_{\lambda,\sigma}$ satisfies Assumptions (A3) and (A5) for some $\lambda_\sigma, \Lambda_\sigma, \rho$ and L. Then,*

$$\Psi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) \geq \left(1 - \frac{r}{4\rho L}\right) \cdot \left(1 - \frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\right)^2 \cdot \frac{\sqrt{2\pi}}{36} \cdot \frac{r}{\rho L\sqrt{d+2}} \cdot \frac{\lambda_\sigma^2}{\Lambda_\sigma^2} \tag{2.30}$$

Some remarks are in order:

- We prove Proposition 4 by separately upper bounding $\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$ and lower bounding the volume $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$. Of these two bounds, the trickier to prove is the upper bound on the cut, which involves carefully estimating the probability mass of thin tubes around the boundary of $\mathcal{C}_{\lambda,\sigma}$.
- Proposition 5 is proved in a completely different way. The proof relies heavily on bounds on the isoperimetric ratio of convex sets (as derived by e.g. Lovász and Simonovits [1990] or Dyer et al. [1991]), and thus the embedding assumption (A5) and Lipschitz parameter L play an important role in proving the upper bound in Proposition 5.
- There is some interdependence between L and σ, ρ, which might lead one to hope that (A5) is non-essential. However, it is not possible to eliminate condition (A5) without incurring an additional factor of at least $(\rho/\sigma)^d$ in (2.30), achieved, for instance, when $\mathcal{C}_{\lambda,\sigma}$ is a dumbbell-like set consisting of two balls of diameter ρ linked by a cylinder of radius σ. In contrast, (2.30) depends polynomially on d, and many reasonably shaped sets—such as star-shaped sets as well as half-moon shapes of the type we consider in Section 2.4—satisfy (A5) for reasonably small values of L [Abbasi-Yadkori, 2016, Abbasi-Yadkori et al., 2017].

Applying these results along with Theorem 3, we obtain an upper bound on $\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])$. In what follows, $C_{1,\delta}, C_{2,\delta}, \ldots$ are constants which may depend on δ, but not on n, $\mathbb{P}$ or $\mathcal{C}_{\lambda,\sigma}$, and which we keep track of in our proofs.

Theorem 4. *Let $C_\lambda \subseteq \mathbb{R}^d$ and $\delta \in (0, 1/4)$. Suppose that $C_{\lambda,\sigma}$ satisfies (A2)-(A6) for some $\lambda_\sigma, \Lambda_\sigma, \theta, \gamma, \rho$ and L, that $0 < r \leq \sigma/4d$, and that the sample size n satisfies the same conditions as in Theorem 4. Then*

with probability at least $1 - B_2/n - 4\exp\{-b_1\delta^2 n\} - (2n+2)\exp\{-b_2\delta^2 n\} - (n+1)\exp\{-b_3 n\}$*, the following statement holds: there exists a set* $\mathcal{C}_{\lambda,\sigma}[X]^g \subseteq \mathcal{C}_{\lambda,\sigma}[X]$ *of large volume,* $\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X]^g) \geq \mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])/2$*, such that if Algorithm 1 is* δ*-well-initialized with respect to* $\mathcal{C}_{\lambda,\sigma}$*, and run with any seed node* $v \in \mathcal{C}_{\lambda,\sigma}[X]^g$*, then the PPR estimated cluster* $\widehat{C}$ *satisfies*

$$\frac{\Delta(\widehat{C}; \mathcal{C}_{\lambda,\sigma}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}[X])} \leq C_{1,\delta} \cdot d^3(d+2) \cdot \frac{L^2\rho^2}{\sigma r} \cdot \frac{\Lambda_\sigma^2 \lambda(\lambda - \theta\frac{r^\gamma}{\gamma+1})}{\lambda_\sigma^4} \cdot \log^2\left(C_{2,\delta}^{1/d} \frac{\Lambda_\sigma^{2/d} L\rho}{\lambda_\sigma^{2/d} 2r}\right) \tag{2.31}$$

As we have previously summarized, Theorem 4 shows the separate roles played by geometry and density in the ability of PPR to recover a density cluster. Now, we make some other remarks:

- Observe that while the diameter ρ is absent from our upper bound on normalized cut in Proposition 4, it enters the ultimate bound in Theorem 4 through the conductance. This reflects (what may be regarded as) established wisdom regarding spectral partitioning algorithms more generally [Guattery and Miller, 1995, Hein and Bühler, 2010], but newly applied to the density clustering setting: if the diameter ρ is large, then PPR may fail to recover $\mathcal{C}_{\lambda,\sigma}[X]$ even when $\mathcal{C}_\lambda$ is sufficiently well-conditioned to ensure that $\mathcal{C}_{\lambda,\sigma}[X]$ has a small normalized cut in $G_{n,r}$. This will be supported by simulations in Section 2.4.2.
- Several modifications of global spectral clustering have been proposed with the intent of making such procedures essentially independent of the shape of the density cluster $\mathcal{C}_\lambda$. For instance, Arias-Castro [2009], Pelletier and Pudlo [2011] introduce a cleaning step to remove low-degree vertices, whereas Little et al. [2020] use a weighted geometric graph, where the weights are computed with respect to a density-dependent distance. The resulting procedures come with stronger density cluster recovery guarantees. However, the key ingredient in such procedures is the explicitly density-dependent part of the algorithm, and spectral clustering functions as more of a post-processing step. These methods are thus very different in spirit to PPR, which is a bona fide (local) spectral clustering algorithm.
- As mentioned in the discussion after Theorem 3, the population-level normalized cut and conductance also play a leading role in the analysis of global spectral clustering algorithms. It therefore seems likely that similar bounds to (2.31) would apply to the output of global spectral clustering methods as well, but formalizing this is outside the scope of our work.
- The symmetric set difference does not measure whether $\widehat{C}$ can (perfectly) distinguish any two distinct clusters $\mathcal{C}_\lambda, \mathcal{C}'_\lambda \in \mathbb{C}_f(\lambda)$. In Appendix A.5, we show that the PPR estimate $\widehat{C}$ can in fact distinguish two distinct clusters $\mathcal{C}_\lambda$ and $\mathcal{C}'_\lambda$, but the result holds only under relatively restrictive conditions.

2.3 Negative result

We now exhibit a hard case for density clustering using PPR, that is, a distribution $\mathbb{P}$ for which PPR is unlikely to recover a density cluster. Let $\mathcal{C}^{(0)}, \mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ be rectangles in $\mathbb{R}^2$,

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

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

$$\mathbb{P} = \frac{1-\epsilon}{2}\mathbb{P}_1 + \frac{1-\epsilon}{2}\mathbb{P}_2 + \frac{\epsilon}{2}\mathbb{P}_0,$$

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

$$f(x) = \frac{1}{\rho\sigma}\left(\frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}^{(1)}) + \frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}^{(2)}) + \epsilon\mathbf{1}(x \in \mathcal{C}^{(0)})\right), \tag{2.32}$$

so that for any $\epsilon < \lambda < (1-\epsilon)/2$, we have $\mathbb{C}_f(\lambda) = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}\}$. Figure 2.1 visualizes the density f for two different choices of ϵ, σ, ρ.

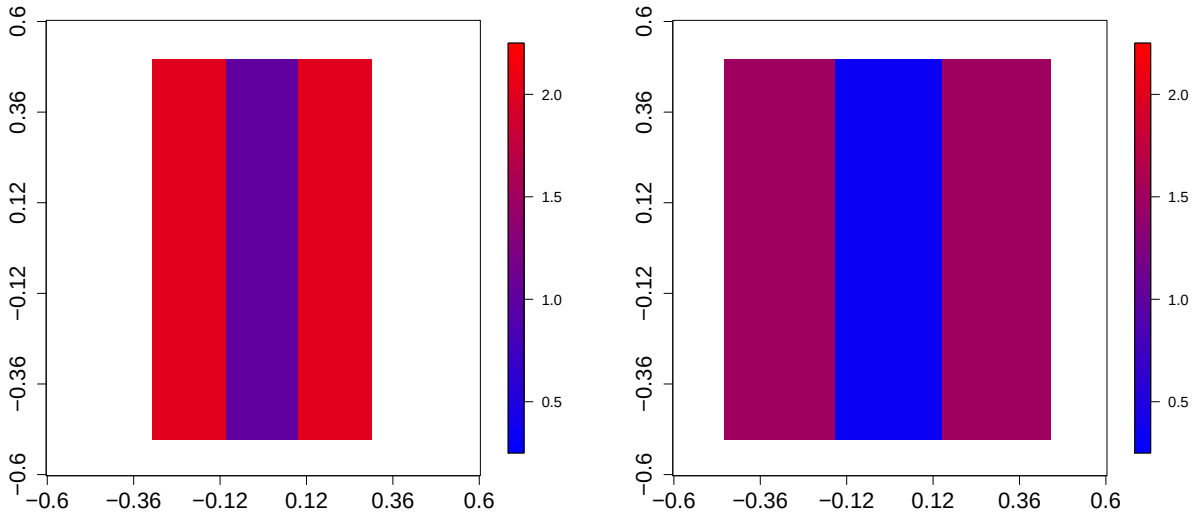


Figure 2.1: The density f in (2.32), for $\rho = 1$, and two different choices of ϵ and σ. Left: $\epsilon = 0.3$ and $\sigma = 0.1$; right: $\epsilon = 0.2$ and $\sigma = 0.2$.

2.3.1 Lower bound on symmetric set difference

As the following theorem demonstrates, even when Algorithm 1 is reasonably initialized, if the density cluster $\mathcal{C}^{(1)}$ is sufficiently geometrically ill-conditioned (in words, tall and thin) the cluster estimator $\widehat{C}$ will fail to recover $\mathcal{C}^{(1)}$. Let

$$\mathcal{L} = \{(x_1, x_2) \in \mathcal{X} : x_2 < 0\}. \tag{2.33}$$

In the following Theorem, $B_{1,\delta}$ and $B_{2,\delta}$ are constants which may depend on $\delta, \mathbb{P}, \mathcal{C}_{\lambda,\sigma}$ and r, but not on n.

Theorem 5. *Fix $\delta \in (0, 1/7)$. Assume the neighborhood graph radius $r < \sigma/4$, that*

$$\max\left\{B_{1,\delta} \cdot \frac{r}{\rho}, B_{2,\delta} \cdot \frac{1}{n}\right\} < \frac{1}{18} \quad \text{and} \quad n \geq 8\frac{(1+\delta)}{(1-\delta)}, \tag{2.34}$$

and that Algorithm 1 is initialized using inputs $\alpha = 36 \cdot \Phi_{n,r}(\mathcal{L}[X])$, and $(L, U) = (0, 1)$. Then the following statement holds with probability at least $1 - (B_4 + 2n + 10)\exp\{-n\delta^2 b_4\}$: there exists a set $\mathcal{C}[X]^g$ of large volume, $\mathrm{vol}_{n,r}(\mathcal{C}[X]^g \cap \mathcal{C}^{(1)}[X]) \geq \mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X]; G_{n,r})/8$, such that for any seed node $v \in \mathcal{C}[X]^g$, the PPR estimated cluster $\widehat{C}$ satisfies

$$\frac{\sigma\rho}{r^2 n^2} \cdot \mathrm{vol}_{n,r}(\widehat{C} \triangle \mathcal{C}^{(1)}[X]) \geq \frac{1-\delta}{2} - C_{3,\delta} \cdot \frac{\sqrt{\sigma/\rho}}{\epsilon^2} \cdot \sqrt{\log\left(C_{4,\delta} \cdot \frac{\rho\sigma}{\epsilon^2 r^2}\right)\frac{\sigma}{r}}, \tag{2.35}$$

We make a couple of remarks:

- Theorem 5 is stated with respect to a particular hard case, where the density clusters are rectangular subsets of $\mathbb{R}^2$. We chose this setting to make the theorem simple to state, and our results are generalizable to $\mathbb{R}^d$ and to non-rectangular clusters. Technically, the rectangles $\mathcal{C}^{(0)}, \mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ are not σ-expansions due to their sharp corners. To fix this, one can simply modify these sets to have appropriately rounded corners, and our lower bound arguments do not need to change significantly, subject to some additional bookkeeping. Thus we ignore this technicality in our subsequent discussion.

- Although we state our lower bound with respect to PPR run on a neighborhood graph, the conclusion is likely to hold for a much broader class of spectral clustering algorithms. In the proof of Theorem 5, we rely heavily on the fact that when ϵ^2 is sufficiently greater than σ/ρ, the normalized cut of $\mathcal{C}^{(1)}$ will be much larger than that of $\mathcal{L}$. In this case, not merely PPR but any algorithm that approximates the minimum normalized cut is unlikely to recover $\mathcal{C}^{(1)}$. In particular, local spectral clustering algorithms based on truncated random walks [Spielman and Teng, 2013], global spectral clustering algorithms [Shi and Malik, 2000], and p-Laplacian based spectral embeddings [Hein and Bühler, 2010] all have provable upper bounds on the normalized cut of cluster they output, and thus we expect that they would all fail to estimate $\mathcal{C}^{(1)}$.

2.3.2 Comparison between upper and lower bounds

To better digest the implications of Theorem 5, we translate the results of our upper bound in Theorem 4 to the density f given in (2.32). Observe that $\mathcal{C}^{(1)}$ satisfies each of the Assumptions (A3)–(A6):

(A1) The density $f(x) = \frac{1-\epsilon}{2\rho\sigma}$ for all $x \in \mathcal{C}^{(1)}$.

(A2) The density $f(x) = \frac{\epsilon}{\rho\sigma}$ for all x such that $0 < \mathrm{dist}(x, \mathcal{C}^{(1)}) \le \sigma$. Therefore for all such x,

$$\inf_{x' \in \mathcal{C}^{(1)}} f(x') - f(x) = \left\{ \frac{1-\epsilon}{2} - \epsilon \right\} \frac{1}{\rho\sigma},$$

which meets the decay requirement with exponent $\gamma = 0$.

(A3) The set $\mathcal{C}^{(1)}$ is itself convex, and has diameter ρ.

(A4) By symmetry, $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)}) = \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(2)})$, and therefore $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)}) \le \frac{1}{2}\mathrm{vol}_{\mathbb{P},r}(\mathbb{R}^d)$.

If the user-specified parameters are initialized according to (2.22), we may apply Theorem 4. This implies that there exists a set $\mathcal{C}^{(1)}[X] \subseteq \mathcal{C}^{(1)}$ with $\mathrm{vol}_{n,r}(\mathcal{C}[X]^g) \ge \frac{1}{2}\mathrm{vol}_{n,r}(\mathcal{C}[X])$ such that for any seed node $v \in \mathcal{C}^{(1)}[X]$, and for large enough n, the PPR estimated cluster $\widehat{C}$ satisfies with high probability

$$\frac{\mathrm{vol}_{n,r}(\widehat{C} \triangle \mathcal{C}^{(1)}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X])} \le 32 C_{1,\delta} \cdot \frac{\rho^2}{\sigma r} \cdot \frac{\epsilon}{1-\epsilon} \cdot \log^2\left(\sqrt{C_{2,\delta}} \frac{\rho}{2r} \right)$$

To facilitate comparisons between our upper and lower bounds set $r = \sigma/8$. Then the following statements each hold with high probability.

- If the user-specified parameters satisfy (2.22), and for some $a \ge 0$,

$$\frac{\epsilon}{1-\epsilon} \le \frac{a}{256 C_{1,\delta}} \left(\frac{\sigma}{\rho \log(\rho/\sigma\sqrt{C_{2,\delta}})} \right)^2,$$

then $\Delta(\widehat{C}, \mathcal{C}^{(1)}[X]) \le a \cdot \mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X])$.

- The population-level volume $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)}) \le (1-\epsilon)/2 \cdot \pi r^2/(\rho\sigma)$, and

$$\mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X]) \le (1+\delta) \cdot n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)}).$$

Therefore, if the user-specified parameters are as in Theorem 5, and

$$\epsilon \ge \sqrt{8 C_{3,\delta}} \left(\frac{\sigma}{\rho} \log\left(64 C_{4,\delta} \cdot \frac{\rho}{\epsilon^2 \sigma} \right) \right)^{1/4},$$

then $\Delta(\widehat{C}, \mathcal{C}^{(1)}[X]) \ge \frac{1}{8}\mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X])$.

Ignoring constants and log factors, we can summarize the above conclusions as follows: if ϵ is much less than $(\sigma/\rho)^2$, then PPR will approximately recover the density cluster $\mathcal{C}^{(1)}$, whereas if ϵ is much greater than $(\sigma/\rho)^{1/4}$ then PPR will fail to recover $\mathcal{C}^{(1)}$, even when reasonably initialized with a seed node $v \in \mathcal{C}^{(1)}$. Jointly, these upper and lower bounds give a relatively precise characterization of what it means for a density cluster to be well- or poorly-conditioned for recovery using PPR.[5]

Of course, it is not hard to show that in the example under consideration, classical plug-in density cluster estimators can consistently recover the σ-expansion $\mathcal{C}_{\lambda,\sigma}$ of a density cluster $\mathcal{C}_\lambda$, even if ϵ is large compared to σ/ρ. That PPR has trouble recovering density clusters here (where standard plug-in approaches do not) is not meant to be a knock on PPR. Rather, it simply reflects that while classical density clustering approaches are specifically designed to identify high-density regions regardless of their geometry, PPR relies on geometry as well as density when forming the output cluster.

2.4 Experiments

We provide numerical experiments to investigate the tightness of our theoretical results in Section 2.2, and compare the performance of PPR with a density clustering algorithm on the "two moons" dataset. We defer details of the experimental settings to Appendix A.6.

2.4.1 Validating theoretical bounds

We investigate the tightness of Lemma 2 and Propositions 4 and 5— i.e. the bounds on population functionals required for the eventual density cluster recovery result in Theorem 4—via simulation. Figure 2.2 compares our bounds on normalized cut, conductance, and local spread of a density cluster with the actual empirically-computed quantities, when samples are drawn from a mixture of uniform distributions over rectangular clusters. In the first row we vary the diameter ρ of the candidate cluster, in the second row we vary the width σ, and in the third row we vary the ratio $(\lambda - \theta)/\lambda$ of the density within and outside the cluster. In almost all cases, it is encouraging to see that our bounds track closely with their empirical counterparts, and are loose by roughly an order of magnitude at most. The one exception to this is the dependence of local spread on the width σ; this theoretical deficiency stems from a loose bound on the volume of sets with large aspect ratio (meaning ρ/σ is much greater than 1), but in any case the local spread contributes only log factors to the ultimate bound on cluster recovery. On the other hand, the looseness in each of these bounds will propagate to our eventual upper bound on $\Delta(\widehat{C}, \mathcal{C}_\sigma[X])/\mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X])$, which as a result is loose by several orders of magnitude.

2.4.2 Empirical behavior of PPR

In Figure 2.3, to drive home the implications of Sections 2.2 and 2.3, we compare the behavior of PPR and the density clustering algorithm of Chaudhuri and Dasgupta [2010] on the well-known "two moons" dataset (with added 2d Gaussian noise), considered a prototypical success story for spectral clustering algorithms. We also examine the cluster which minimizes the normalized cut; as we have discussed previously, this can be as a middle ground between the geometric sensitivity of PPR, and the geometric insensitivity of density clustering. The first column shows the empirical density clusters $\mathcal{C}_\lambda[X]$ and $\mathcal{C}'_\lambda[X]$ for a particular threshold λ of the density function; the second column shows the cluster recovered by PPR; the third column shows the global minimum normalized cut, computed according to the algorithm of Bresson et al. [2012]; and the last column shows a cut of the density cluster tree estimator of Chaudhuri and Dasgupta [2010]. We can see the degrading ability of PPR to recover density clusters as the two moons become less well-separated. Of particular interest is the fact that PPR fails to recover one of the moons even when normalized cut still succeeds in doing so. Additionally, we note that the Chaudhuri-Dasgupta algorithm succeeds even when

[5] It is worth pointing out that the above conclusions are reliant on specific (albeit reasonable) ranges and choices of input parameters, which in some instances differ between the upper and lower bounds. We suspect that our lower bound continues to hold even when choosing input parameters as dictated by our upper bound, but do not pursue the details.

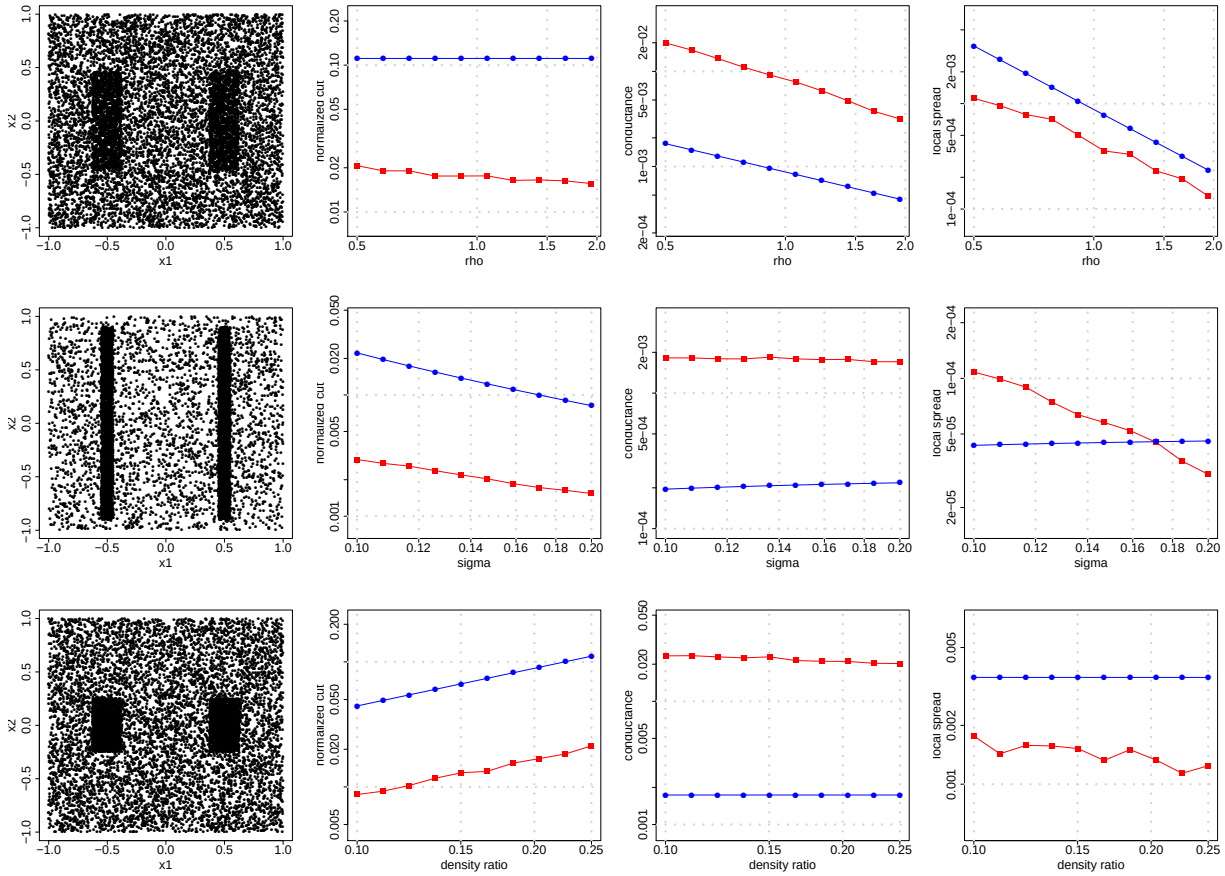


Figure 2.2: Empirical normalized cut, conductance, and local spread (in red), versus their theoretical bounds (in blue). In the first row we vary the diameter ρ, in the second row we vary the thickness σ, and in the third row we vary the density ratio $(\lambda - \theta)/\lambda$. The first column shows $n = 8000$ samples for three different parameter values.

both PPR and normalized cut fail. This supports one of our main messages, which is that PPR recovers only geometrically well-conditioned density clusters.

2.5 Discussion

In this work, we have analyzed the behavior of PPR in the classical setup of nonparametric statistics. We have shown how PPR depends on the distribution $\mathbb{P}$ through the population-level normalized cut, conductance, and local spread, and established upper bounds on the error with which PPR recovers an arbitrary candidate cluster $\mathcal{C} \subseteq \mathbb{R}^d$. In the particularly important case where $\mathcal{C} = \mathcal{C}_\lambda$ is a λ-density cluster, we have shown that PPR recovers $\mathcal{C}_\lambda$ if and only if both the density cluster and density are well-conditioned. We now conclude by summarizing a couple of interesting directions for future work.

Letting the radius of the neighborhood graph shrink, $r \to 0$ as $n \to \infty$, would be computationally attractive, as it would ensure that the graph $G_{n,r}$ is sparse. However, the bounds (2.26) and (2.31) will blow up as the radius r goes to 0, preventing us from making claims about the behavior of PPR in this regime. Although the restriction to a kernel function fixed in n is common in spectral clustering theory [von Luxburg et al., 2008, Schiebinger et al., 2015, Singer and Wu, 2017], recent works [Shi, 2015, Calder and García Trillos, 2019, García Trillos and Slepčev, 2018a, García Trillos et al., 2020, Yuan et al., 2020] have demonstrated

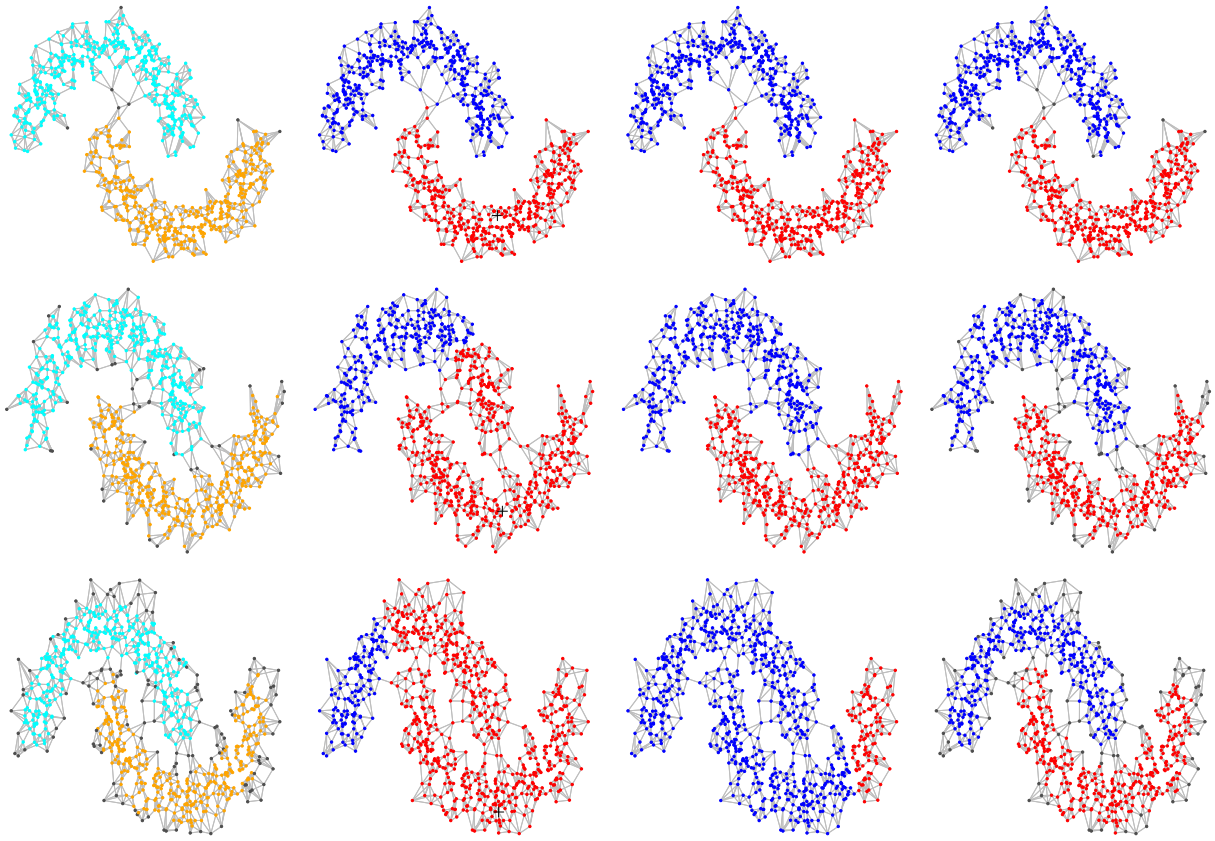


Figure 2.3: True density (column 1), PPR (column 2), minimum normalized cut (column 3) and estimated density (column 4) clusters for 3 different simulated data sets. Seed node for PPR denoted by a black cross.

that spectral methods have meaningful continuum limits when $r \to 0$ as $n \to \infty$, and given precise rates of convergence. García Trillos et al. [2019b] have applied these results to analyze global spectral clustering in the nonparametric mixture model, obtaining asymptotic upper bounds that do not depend on r; it seems plausible that similar bounds could be obtained for local spectral clustering with PPR, although the arguments would necessarily be quite different.

In another direction, it would be very useful to find reasonable conditions under which the ratio $\Delta(\widehat{C}, \mathcal{C}[X])/\mathrm{vol}_{n,r}(\mathcal{C}[X])$ would tend to 0 as $n \to \infty$. It seems likely that such a strong result would entail bounds on the L^∞-error of PPR. Although most results thus far derive bounds only on the L^1- or L^2-error of spectral clustering methods, some recent works [Dunson et al., 2020, Calder et al., 2020a] have established L^∞-bounds on the error with which the eigenvectors of a graph Laplacian matrix approximate the eigenvectors of a weighted Laplace-Beltrami operator. It is not clear whether the techniques used in these works can be applied to PPR.

Chapter 3

Minimax Optimal Regression over Sobolev Spaces via Laplacian Regularization on Neighborhood Graphs

3.1 Introduction

We adopt the standard nonparametric regression setup, where we observe samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are i.i.d. draws from the model

$$Y_i = f_0(X_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \tag{3.1}$$

where ε_i is independent of X_i. Our goal is to perform statistical inference on the unknown regression function f_0, by which we mean either *estimating* f_0 or *testing* whether $f_0 = 0$, i.e., whether there is any signal present.

Laplacian smoothing [Smola and Kondor, 2003] is a penalized least squares estimator, defined over a graph. Letting $G = (V, W)$ be a weighted undirected graph with vertices $V = \{1, \ldots, n\}$, associated with $\{X_1, \ldots, X_n\}$, and $W \in \mathbb{R}^{n \times n}$ is the (weighted) adjacency matrix of the graph. the Laplacian smoothing estimator $\widehat{f}$ is given by

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathbb{R}^n} \sum_{i=1}^{n} (Y_i - f_i)^2 + \rho \cdot f^\top \mathbf{L} f. \tag{3.2}$$

Here $\mathbf{L}$ is the graph Laplacian matrix (defined formally in Section 3.3), G is typically a geometric graph (such as a k-nearest-neighbor or neighborhood graph), $\rho \geq 0$ is a tuning parameter, and the penalty

$$f^\top \mathbf{L} f = \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} (f_i - f_j)^2$$

encourages $\widehat{f}_i \approx \widehat{f}_j$ when $X_i \approx X_j$. Assuming (3.2) is a reasonable estimator of f_0, the statistic

$$\widehat{T} = \frac{1}{n} \|\widehat{f}\|_2^2 \tag{3.3}$$

is in turn a natural test statistic to test if $f_0 = 0$.

Of course there are many methods for nonparametric regression (see, e.g., Györfi et al. [2006], Wasserman [2006], Tsybakov [2008b]), but Laplacian smoothing has its own set of advantages. For instance:

- *Computational ease.* Laplacian smoothing is fast, easy, and stable to compute. The estimate $\widehat{f}$ can be computed by solving a symmetric diagonally dominant linear system. There are by now various nearly-linear-time solvers for this problem (see e.g., the seminal papers of Spielman and Teng [2011, 2013, 2014], or the overview by Vishnoi [2012] and references therein).
- *Generality.* Laplacian smoothing is well-defined whenever one can associate a graph with observed responses. This generality lends itself to many different data modalities, e.g., text and image classification, as in Kondor and Lafferty [2002], Belkin and Niyogi [2003], Belkin et al. [2006].
- *Weak supervision.* Although we study Laplacian smoothing in the supervised problem setting (3.1), the method can be adapted to the semi-supervised or unsupervised settings, as in Zhu et al. [2003], Zhou et al. [2005], Nadler et al. [2009].

For these reasons, a body of work has emerged that analyzes the statistical properties of Laplacian smoothing, and graph-based methods more generally. Roughly speaking, this work can be divided into two categories, based on the perspective they adopt.

- *Fixed design perspective.* Here one treats the design points $X_1, \dots, X_n$ and the graph G as fixed, and carries out inference on $f_0(X_i)$, $i = 1, \dots, n$. In this problem setting, tight upper bounds have been derived on the error of various graph-based methods (e.g., Wang et al. [2016], Hütter and Rigollet [2016], Sadhanala et al. [2016a, 2017], Kirichenko and van Zanten [2017], Kirichenko et al. [2018]) and tests (e.g., Sharpnack and Singh [2010], Sharpnack et al. [2013a,b, 2015]), which certify that such procedures are *optimal* over "function" classes (in quotes because these classes really model the n-dimensional vector of evaluations). The upside of this work is its generality: in this setting G need not be a geometric graph, but in principle it could be any graph over $V = \{1, \dots, n\}$. The downside is that, in the context of nonparametric regression, it is arguably not as natural to think of the evaluations of f_0 as exhibiting smoothness over some fixed pre-defined graph G, and more natural to speak of the smoothness of the function f_0 itself.
- *Random design perspective.* Here one treats the design points $X_1, \dots, X_n$ as independent samples from some distribution P supported on a domain $\mathcal{X} \subseteq \mathbb{R}^d$. Inference is drawn on the regression function $f_0 : \mathcal{X} \to \mathbb{R}$, which is typically assumed to be smooth in some *continuum* sense, e.g., it possesses a first derivative bounded in L^∞ (Hölder) or L^2 (Sobolev) norm. To conduct graph-based inference, the user first builds a neighborhood graph over the random design points—so that W_{ij} is large when X_i and X_j are close in (say) Euclidean distance—and then computes e.g., (3.2) or (3.3). In this context, various graph-based procedures have been shown to be *consistent*: as $n \to \infty$, they converge to a continuum limit (see Belkin and Niyogi [2007], von Luxburg et al. [2008], García Trillos and Slepčev [2018b] among others). However, until recently such statements were not accompanied by error rates, and even so, such error rates as have been proved [Lee et al., 2016, García Trillos and Murray, 2020] are not optimal over continuum function spaces, such as Hölder or Sobolev classes.

The random design perspective bears a more natural connection with nonparametric regression (the focus in this paper), as it allows us to formulate smoothness based on f_0 itself (how it behaves as a continuum function, and not just its evaluations at the design points). In this paper, we will adopt the random design perspective, and seek to answer the following question:

> When we assume the regression function f_0 is smooth in a continuum sense, does Laplacian smoothing achieve optimal performance for estimation and goodness-of-fit testing?

This is no small question—arguably, it is *the* central question of nonparametric regression—and without an answer one cannot fully compare the statistical properties of Laplacian smoothing to alternative methods. It also seems difficult to answer: as we discuss next, there is a fundamental gap between the *discrete* smoothness

imposed by the penalty $f^\top \mathbf{L} f$ in problem (3.2) and the *continuum* smoothness assumed on f_0, and in order to obtain sharp upper bounds we will need to bridge this gap in a suitable sense.

3.2 Summary of Results

Advantages of the Discrete Approach. In light of the potential difficulty in bridging the gap between discrete and continuum notions of smoothness, it is worth asking whether there is any *statistical* advantage to solving a discrete problem such as (3.2) (setting aside computational considerations for the moment). After all, we could have instead solved the following variational problem:

$$\widetilde{f} = \underset{f:\mathcal{X}\to\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2 + \rho \int_{\mathcal{X}} \|\nabla f(x)\|_2^2 \, dx, \tag{3.4}$$

where the optimization is performed over all continuous functions f that have a weak derivative ∇f in $L^2(\mathcal{X})$. Analogously, for testing, we could use:

$$\widetilde{T} = \|\widetilde{f}\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \widetilde{f}(X_i)^2. \tag{3.5}$$

The penalty term in (3.4) leverages the assumption that f_0 has a smooth derivative in a seemingly natural way. Indeed, the estimator $\widetilde{f}$ and statistic $\widetilde{T}$ are well-known: for $d = 1$, $\widetilde{f}$ is the familiar *smoothing spline*, and for $d > 1$, it is a type of *thin-plate spline*. The statistical properties of smoothing and thin-plate splines are well-understood [van de Geer, 2000, Liu et al., 2019]. As we discuss later, the Laplacian smoothing problem (3.2) can be viewed as a discrete and noisy approximation to (3.4). At first blush, this suggests that Laplacian smoothing should at best inherit the statistical properties of (3.4), and at worst may have meaningfully larger error.

However, as we shall see the actual story is quite different: remarkably, Laplacian smoothing enjoys optimality properties even in settings where the thin-plate spline estimator (3.4) is not well-posed (to be explained shortly); Tables 3.1 and 3.2 summarize. As we establish in Theorems 6-10, when computed over an appropriately formed neighborhood graph, Laplacian smoothing estimators and tests are minimax optimal over first-order *continuum* Sobolev balls. This holds true either when $\mathcal{X} \subseteq \mathbb{R}^d$ is a full-dimensional domain and $d = 1, 2$, or 3, or when $\mathcal{X}$ is a manifold embedded in $\mathbb{R}^d$ of intrinsic dimension $m = 1, 2$, or 3. Additionally, the estimator $\widehat{f}$ is nearly minimax optimal (to within a $(\log n)^{1/3}$ factor) when $d = 4$ (or $m = 4$ in the manifold case).

By contrast, smoothing splines are optimal only when $d = 1$. When $d > 1$, the thin-plate spline estimator (3.4) is not even well-posed, in the following sense: for any $(X_1, Y_1), \ldots, (X_n, Y_n)$ and any $\delta > 0$, there exists (e.g., Green and Silverman [1993] give a construction using "bump" functions) a differentiable function f such that $f(X_i) = Y_i$, $i = 1, \ldots, n$, and

$$\int_{\mathcal{X}} \|\nabla f(x)\|_2^2 \leq \delta.$$

In other words, f achieves perfect (zero) data loss and arbitrarily small penalty in the problem (3.4). This will clearly not lead to a consistent estimator of f_0 across the design points (as it always yields Y_i at each X_i). In this light, our results when $d > 1$ favorably distinguish Laplacian smoothing from its natural variational analog.

Future Directions. To be clear, there is still much left to be investigated. For one, the Laplacian smoothing estimator $\widehat{f}$ is only defined at $X_1, \ldots, X_n$. In this work we study its in-sample mean squared error

$$\big\|\widehat{f} - f_0\big\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \Big(\widehat{f}_i - f_0(X_i)\Big)^2. \tag{3.6}$$

Dimension	Laplacian smoothing (3.2)	Thin-plate splines (3.4)
$d = 1$	$\boldsymbol{n^{-2/3}}$	$\boldsymbol{n^{-2/3}}$
$d = 2, 3$	$\boldsymbol{n^{-2/(2+d)}}$	1
$d = 4$	$\boldsymbol{n^{-1/3}}(\log n)^{1/3}$	1
$d \geq 5$	$(\log n/n)^{4/(3d)}$	1

Table 3.1: Summary of estimation rates over first-order Sobolev balls. Black font marks new results from this paper, red font marks previously-known results; bold font marks minimax optimal rates. Although we suppress it for simplicity, in all cases the dependence of the error rate on the radius of the Sobolev ball is also optimal. The rates for thin-plate splines with $d \geq 2$ assume the estimator $\widetilde{f}$ interpolates the responses, $\widetilde{f}(X_i) = Y_i$ for $i = 1, \ldots, n$; see the discussion in Section 3.2. Here, we use "1" to indicate inconsistency (error not converging to 0). Lastly, when $\mathcal{X}$ is an m-dimensional manifold embedded in $\mathbb{R}^d$, all Laplacian smoothing results hold with d replaced by m, without any change to the method itself.

Dimension	Laplacian smoothing (3.3)	Thin-plate splines (3.5)
$d = 1$	$\boldsymbol{n^{-4/5}}$	$\boldsymbol{n^{-4/5}}$
$d = 2, 3$	$\boldsymbol{n^{-4/(4+d)}}$	$n^{-1/2}$
$d \geq 4$	$\boldsymbol{n^{-1/2}}$	$\boldsymbol{n^{-1/2}}$

Table 3.2: Summary of testing rates over first-order Sobolev balls; black, red, and bold fonts are used as in Table 3.1. The rates for thin-plate splines with $d \geq 2$ assume the test statistic $\widetilde{T}$ is computed using an $\widetilde{f}$ that interpolates the responses, $\widetilde{f}(X_i) = Y_i$ for $i = 1, \ldots, n$. Rates for $d \geq 4$ assume that $f_0 \in L^4(\mathcal{X}, M)$. Lastly, when $\mathcal{X}$ is an m-dimensional manifold embedded in $\mathbb{R}^d$, all rates hold with d replaced by m.

In Section 3.4, we discuss how to extend $\widehat{f}$ to a function over all $\mathcal{X}$, in such a way that the out-of-sample mean squared error $\|\widehat{f} - f_0\|^2_{L^2(\mathcal{X})}$ should remain small, but leave a formal analysis to future work.

In a different direction, problem (3.4) is only a special, first-order case of thin-plate splines. In general, the kth order thin-plate spline estimator is defined as

$$\widetilde{f} = \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^d} \sum_{i=1}^n \big(Y_i - f(X_i)\big)^2 + \rho \sum_{|\alpha|=k} \int_{\mathcal{X}} \big(D^\alpha f(x)\big)^2 \, dx,$$

where for each multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$ we write $D^\alpha f(x) = \partial^k f / \partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$. This problem is in general well-posed whenever $2k > d$. In this regime, assuming that the kth order partial derivatives $D^\alpha f_0$ are all $L^2(\mathcal{X})$ bounded, the degree k thin-plate spline has error on the order of $n^{-2k/(2k+d)}$ [van de Geer, 2000], which is minimax rate-optimal for such functions. Of course, assuming f_0 has k bounded derivatives for some $2k > d$ is a very strong condition, but at present we do not know if (adaptations of) Laplacian smoothing on neighborhood graphs achieve these rates.

Notation. For an integer $p \geq 1$, we use $L^p(\mathcal{X})$ for the set of functions f such that

$$\|f\|^p_{L^p(\mathcal{X})} := \int_{\mathcal{X}} |f(x)|^p \, dx < \infty,$$

and $C^p(\mathcal{X})$ for the set of functions that are p times continuously differentiable. For sequences a_n, b_n, we write $a_n \lesssim b_n$ to mean $a_n \leq C b_n$ for a constant $C > 0$ and large enough n, and $a_n \asymp b_n$ to mean $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Lastly, we use $a \wedge b = \min\{a, b\}$.

3.3 Background

Before we present our main results in Section 3.4, we define neighborhood graph Laplacians, and review known minimax rates over first-order Sobolev spaces.

Neighborhood Graph Laplacians. In the graph-based approach to nonparametric regression, we first build a neighborhood graph $G_{n,r} = (V, W)$, for $V = \{1, \dots, n\}$, to capture the geometry of P (the design distribution) and $\mathcal{X}$ (the domain) in a suitable sense. The $n \times n$ weight matrix $W = (W_{ij})$ encodes proximity between pairs of design points; for a kernel function $K : [0, \infty) \to \mathbb{R}$ and radius $r > 0$, we have

$$W_{ij} = K\left(\frac{\|X_i - X_j\|_2}{r}\right),$$

with $\|\cdot\|_2$ denoting the ℓ_2 norm on $\mathbb{R}^d$. Defining D as the $n \times n$ diagonal matrix with entries $D_{ii} = \sum_{j=1}^n W_{ij}$, the graph Laplacian can then be written as

$$\mathbf{L} = D - W. \tag{3.7}$$

We use $L = \sum_{k=1}^n \lambda_k v_k v_k^\top$ for an eigendecomposition of L, and we always assume, by convention, ordered eigenvalues $0 = \lambda_1 \leq \cdots \leq \lambda_n$, and unit-norm eigenvectors.

Sobolev Spaces. We step away from graph-based methods for a moment, to briefly recall some classical results regarding minimax rates over Sobolev classes. We say that a function $f \in L^2(\mathcal{X})$ belongs to the *first-order Sobolev space* $H^1(\mathcal{X})$ if, for each $j = 1, \dots, d$, the weak partial derivative $D^j f$ exists and belongs to $L^2(\mathcal{X})$. For such functions $f \in H^1(\mathcal{X})$, the Sobolev seminorm $|f|_{H^1(\mathcal{X})}$ is the average size of the gradient $\nabla f = (D^1 f, \dots, D^d f)$,

$$|f|^2_{H^1(\mathcal{X})} := \int_{\mathcal{X}} \big\|\nabla f(x)\big\|_2^2 \, dx,$$

with corresponding Sobolev norm

$$\|f\|_{H^1(\mathcal{X})} := \|f\|_{L^2(\mathcal{X})} + |f|_{H^1(\mathcal{X})}.$$

The Sobolev ball $H^1(\mathcal{X}, M)$ for $M > 0$ is

$$H^1(\mathcal{X}, M) := \Big\{ f \in H^1(\mathcal{X}) : \|f\|^2_{H^1(\mathcal{X})} \leq M^2 \Big\}.$$

For further details regarding Sobolev spaces see, e.g., Evans [2010], Leoni [2017].

Minimax Rates. To carry out a minimax analysis of regression in Sobolev spaces, one must impose regularity conditions on the design distribution P. We shall assume the following.

(P1) P is supported on a domain $\mathcal{X} \subseteq \mathbb{R}^d$, which is an open, connected set with Lipschitz boundary.

(P2) P admits a density p such that

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

Additionally, p is Lipschitz on $\mathcal{X}$, with Lipschitz constant L_p.

Under conditions (P1), (P2), the minimax estimation rate over a Sobolev ball of radius $M \geq n^{-1/2}$ is (e.g., Tsybakov [2008b]):

$$\inf_{\widehat{f}} \sup_{f_0 \in H^1(\mathcal{X}, M)} \mathbb{E}\Big[\|\widehat{f} - f_0\|^2_{L^2(\mathcal{X})}\Big] \asymp M^{2d/(2+d)} n^{-2/(2+d)}. \tag{3.8}$$

(Throughout we assume $M \geq n^{-1/2}$, as otherwise the trivial estimator $\widehat{f} = 0$ achieves smaller error than the parametric rate n^{-1}, and the problem does not fit well within the nonparametric setup.)

As minimax rates in nonparametric hypothesis testing are (comparatively) less familiar than those in nonparametric estimation, we briefly summarize the main idea before stating the optimal error rate. In the goodness-of-fit testing problem, we ask for a test function—formally, a Borel measurable function ϕ taking values in $\{0, 1\}$—which can distinguish between the hypotheses

$$\mathbf{H}_0 : f_0 = f_0^\star, \quad \text{versus} \quad \mathbf{H}_a : f_0 \in \mathcal{F} \setminus \{f_0^\star\}. \tag{3.9}$$

Typically, the null hypothesis $f_0 = f_0^\star \in \mathcal{F}$ reflects the absence of interesting structure, and $\mathcal{F} \setminus \{f_0^\star\}$ is a set of smooth departures from this null. In this paper, as in Ingster and Sapatinas [2009], we focus on the problem of *signal detection* in Sobolev spaces, where $f_0^\star = 0$ and $\mathcal{F} = H^1(\mathcal{X}, M)$ is a first-order Sobolev ball. This is without loss of generality since our test statistic and its analysis are easily modified to handle the case when $f_0^\star$ is not 0, by simply subtracting $f_0^\star(X_i)$ from each observation Y_i.

The Type I error of a test ϕ is $\mathbb{E}_0[\phi]$, and if $\mathbb{E}_0[\phi] \leq \alpha$ for a given $\alpha \in (0, 1)$ we refer to ϕ as a level-α test. The worst-case risk of ϕ over $\mathcal{F}$ is

$$R_n(\phi, \mathcal{F}, \epsilon) := \sup\Big\{\mathbb{E}_{f_0}[1 - \phi] : f_0 \in \mathcal{F}, \|f_0\|_{L^2(\mathcal{X})} > \epsilon\Big\},$$

and for a given constant $b \geq 1$, the minimax critical radius $\epsilon(\mathcal{F})$ is the smallest value of ϵ such that some level-α test has worst-case risk of at most $1/b$. Formally,

$$\epsilon(\mathcal{F}) := \inf\Big\{\epsilon > 0 : \inf_{\phi} R_n(\phi, \mathcal{F}, \epsilon) \leq 1/b\Big\},$$

where in the above the infimum is over all level-α tests ϕ, and $\mathbb{E}_{f_0}[\cdot]$ is the expectation operator under the regression function f_0.[1]

The classical approach to hypothesis testing typically focuses on designing test statistics, and studying their (limiting) distribution in order to ensure control of the Type I error. In many cases the Type II error (or risk in our terminology) is not emphasized, or the risk of the test against fixed or directional alternatives (i.e. alternatives which deviate from the null in a fixed direction) is studied. In contrast, in the minimax paradigm the (uniform or worst-case) risk against a large collection of alternatives is the central focus. See Ingster [1982, 1987], Ingster and Suslina [2012], Arias-Castro et al. [2018], Balakrishnan and Wasserman [2019, 2018] for a more extended treatment of the minimax paradigm in nonparametric testing, and for a discussion of its advantages (and disadvantages) over other approaches to studying hypothesis tests.

Testing $f_0 = 0$ is an easier problem than estimating f_0, and hence the minimax testing critical radius over $H^1(\mathcal{X}, M)$ is smaller than the minimax estimation rate, for $1 \leq d < 4$ (see Ingster and Sapatinas [2009]):

$$\epsilon^2\big(H^1(\mathcal{X}, M)\big) \asymp M^{2d/(4+d)} n^{-4/(4+d)}. \tag{3.10}$$

When $d \geq 4$ the functions in $H^1(\mathcal{X})$ are very irregular; formally speaking $H^1(\mathcal{X})$ does not continuously embed into $L^4(\mathcal{X})$ when $d \geq 4$, and the minimax testing rates in this regime are unknown.

3.4 Minimax Optimality of Laplacian Smoothing

We now formalize the main conclusions of this paper: that Laplacian smoothing methods on neighborhood graphs are minimax rate-optimal over first-order continuum Sobolev classes. We will assume (P1), (P2) on P, and the following condition on the kernel K.

[1] Clearly, the minimax critical radius ϵ depends on α and b. However, we adopt the typical convention of treating $\alpha \in (0, 1)$ and $b \geq 1$ as small but fixed positive constants; hence they will not affect the testing error rates, and we suppress them notationally.

(K1) $K : [0, \infty) \to [0, \infty)$ is a nonincreasing function supported on $[0, 1]$, its restriction to $[0, 1]$ is Lipschitz, and $K(1) > 0$. Additionally, it is normalized so that

$$\int_{\mathbb{R}^d} K(\|z\|_2)\, dz = 1.$$

We assume $\sigma_K = \frac{1}{d} \int_{\mathbb{R}^d} \|x\|_2^2 K(\|x\|_2)\, dx < \infty$.

This is a mild condition: recall the choice of kernel is under the control of the user, and moreover (K1) covers many common kernel choices.

Estimation Error of Laplacian Smoothing. Under these conditions, the Laplacian smoothing estimator $\widehat{f}$ achieves an error rate that matches the minimax lower bound over $H^1(\mathcal{X}, M)$. This statement will hold whenever the graph $G_{n,r}$ is computed with radius r in the following range.

(R1) For constants $C_0, c_0 > 0$, the neighborhood graph radius r satisfies

$$C_0 \left(\frac{\log n}{n} \right)^{\frac{1}{d}} \leq r \leq c_0 \wedge M^{\frac{d-4}{4+2d}} n^{-\frac{3}{4+2d}}.$$

Next we state Theorem 6, our main estimation result. Its proof, as with all proofs of results in this paper, can be found in the appendix.

Theorem 6. *Given i.i.d. draws (X_i, Y_i), $i = 1, \ldots, n$ from (3.1), assume $f_0 \in H^1(\mathcal{X}, M)$ where $\mathcal{X} \subseteq \mathbb{R}^d$ has dimension $d < 4$ and $M \leq n^{1/d}$. Assume (P1), (P2) on the design distribution P, and assume the neighborhood graph $G_{n,r}$ is computed with a kernel K satisfying (K1). There are constants $N, C, C_1, c, c_1 > 0$ (not depending on f_0) such that for any $n \geq N$, and any radius r as in (R1), the Laplacian smoothing estimator $\widehat{f}$ in (3.2) with $\rho = M^{-4/(2+d)}(nr^{d+2})^{-1} n^{-2/(2+d)}$ satisfies*

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C}{\delta} M^{2d/(2+d)} n^{-2/(2+d)},$$

with probability at least $1 - \delta - C_1 n \exp(-c_1 n r^d) - \exp(-c(M^2 n)^{d/(2+d)})$.

To summarize: for $d = 1, 2$, or 3, with high probability, the Laplacian smoothing estimator $\widehat{f}$ has in-sample mean squared error that is within a constant factor of the minimax error. Some remarks:

- The first-order Sobolev space $H^1(\mathcal{X})$ does not continuously embed into $C^0(\mathcal{X})$ when $d > 1$ (in general, the kth order space $H^k(\mathcal{X})$ does not continuously embed into $C^0(\mathcal{X})$ except if $2k > d$). For this reason, one really cannot speak of pointwise evaluation of a Sobolev function $f_0 \in H^1(\mathcal{X})$ when $d > 1$ (as we do in Theorem 6 by defining our target of estimation to be $f_0(X_i)$, $i = 1, \ldots, n$). We can resolve this by appealing to what are known as *Lebesgue points*, as explained in Appendix B.1.
- The assumption $M \leq n^{1/d}$ ensures that the upper bound provided in the theorem is meaningful (i.e., ensures it is of at most a constant order).
- The lower bound on r imposed in condition (R1) is compatible with practice, where by far the most common choice of radius is the connectivity threshold $r \asymp (\log(n)/n)^{1/d}$, which makes $G_{n,r}$ as sparse as possible while still being connected, for maximum computational efficiency. The upper bound may seem a bit more mysterious—we need it for technical reasons to ensure that $\widehat{f}$ does not overfit, but we note that as a practical matter one rarely chooses r to be so large anyway.
- It is possible to extend $\widehat{f}$ to be defined on all of $\mathcal{X}$ and then evaluate the error of such an extension (as measured against f_0) in $L^2(\mathcal{X})$ norm. When $\widehat{f}$ and f_0 are suitably smooth, tools from empirical process theory (see e.g., Chapter 14 of Wainwright [2019]) or approximation theory (e.g., Section

15.5 of Johnstone [2011]) guarantee that the $L^2(\mathcal{X})$ error is not too much greater than its in-sample counterpart. In fact, as shown in Appendix B.7.1, if f_0 is Lipschitz smooth and we extend $\widehat{f}$ to be piecewise constant over the Voronoi tessellation induced by $X_1, \dots, X_n$, then the out-of-sample error $\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}$ is within a negligible factor of the in-sample error $\|\widehat{f} - f_0\|_n$. We leave analysis of the Sobolev case to future work.

- When f_0 is Lipschitz smooth, we can also replace the factor of δ in the high probability bound by a factor of δ^2/n, which is always smaller than δ when $\delta \in (0, 1)$.

When $d = 4$, our analysis results in an upper bound for the error of Laplacian smoothing that is within a $(\log n)^{1/3}$ factor of the minimax error rate. But when $d \geq 5$, our upper bounds do not match the minimax rates.

Theorem 7. *Under the assumptions of Theorem 6, if instead $\mathcal{X}$ has dimension $d = 4$, $r \asymp (\log n/n)^{1/4}$ and $\rho = M^{-2/3}(nr^6)^{-1}(\log n/n)^{1/3}$, then we obtain*

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C}{\delta} M^{4/3} \left(\frac{\log n}{n}\right)^{1/3},$$

with the same probability guarantee as in Theorem 6. If the dimension of $\mathcal{X}$ is $d \geq 5$, $r \asymp (\log n/n)^{1/d}$ and $\rho = M^{-2/3}(nr^{2+d})^{-1} n^{-4/(3d)}$, then

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C}{\delta} M^{4/3} \left(\frac{\log n}{n}\right)^{4/(3d)},$$

again with the same probability guarantee.

This mirrors the conclusions of Sadhanala et al. [2016a] who investigate estimation rates of Laplacian smoothing over the d-dimensional grid graph. These authors argue that their analysis is tight, and that it is likely the estimator, not the analysis, that is deficient when $d \geq 5$. Formalizing such a claim turns out to be harder in the random design setting than in the fixed design setting, and we leave it for future work.

However, we do investigate the matter empirically. In Figure 3.1, we study the (in-sample) mean squared error of the Laplacian smoothing estimator as the dimension d grows. Here $X_1, \dots, X_n$ are sampled uniformly over $\mathcal{X} = [-1, 1]^d$, and the regression function is taken as $f_0(x) \propto \Pi_{i=1}^d \cos(a\pi x_i)$, where $a = 2$ for $d = 2$, and $a = 1$ for $d \geq 3$. This regression function f_0 is quite smooth, and for $d = 2$ and $d = 3$ Laplacian smoothing appears to achieve or exceed the minimax rate. When $d = 4$, Laplacian smoothing appears modestly suboptimal; this fits with our theoretical upper bound, which includes a $(\log n)^{1/3}$ factor that plays a non-negligible role for these problem sizes ($n = 1000$ to $n = 10000$). On the other hand, when $d = 5$, Laplacian smoothing seems to be decidedly suboptimal.

Testing Error of Laplacian Smoothing. For a given $0 < \alpha < 1$, define a threshold $\widehat{t}_\alpha$ as

$$\widehat{t}_\alpha = \frac{1}{n} \sum_{k=1}^n \frac{1}{(\rho\lambda_k + 1)^2} + \frac{1}{n} \sqrt{\frac{2}{\alpha} \sum_{k=1}^n \frac{1}{(\rho\lambda_k + 1)^4}},$$

where we recall λ_k is the kth smallest eigenvalue of L. The Laplacian smoothing test is then simply

$$\widehat{\varphi} = \mathbf{1}\big\{\widehat{T} > \widehat{t}_\alpha\big\}.$$

We show in Appendix C.1 that $\widehat{f}$ is a level-α test. In the next theorem, we upper bound the worst-case risk $R_n(\widehat{\varphi}, H^1(\mathcal{X}, M), \epsilon)$ of $\widehat{\varphi}$, whenever ϵ is at least (a constant times) the critical radius given in (3.10). For this to hold, we will require a tighter range of scalings for the graph radius r.

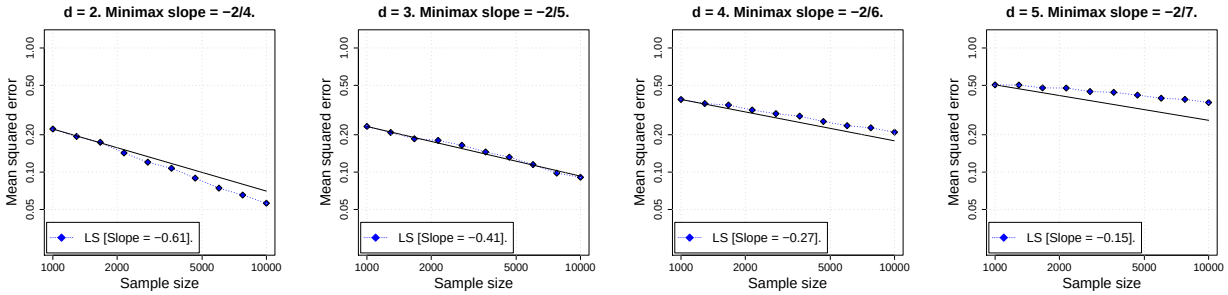


Figure 3.1: Mean squared error of Laplacian smoothing (LS) as a function of sample size n. Each plot is on the log-log scale, and the results are averaged over 5 repetitions, with Laplacian smoothing tuned for optimal average mean squared error. The black line shows the minimax rate (in slope only; the intercept is chosen to match the observed error).

(R2) For constants $C_0, c_0 > 0$, the neighborhood graph radius r satisfies

$$C_0\left(\frac{\log n}{n}\right)^{\frac{1}{d}} \leq r \leq c_0 \wedge M^{\frac{(d-8)}{8+2d}} n^{\frac{d-20}{32+8d}}.$$

We will also require that the radius of the Sobolev class not be too large. Precisely, we will require $M \leq M_{\max}(d)$, where we define

$$M_{\max}(d) := \begin{cases} n^{1/8} & d = 1 \\ n^{(4-d)/(4d)} & d \geq 2. \end{cases}$$

We now give Theorem 8, our main testing result.

Theorem 8. *Given i.i.d. draws (X_i, Y_i), $i = 1, \ldots, n$ from (3.1), assume $f_0 \in H^1(\mathcal{X}, M)$ where $\mathcal{X} \subseteq \mathbb{R}^d$ with $d < 4$, and $M \leq M_{\max}(d)$. Assume (P1), (P2) on the design distribution P, and assume $G_{n,r}$ is computed with a kernel K satisfying (K1). There exist constants $N, C, C_1, c_1 > 0$ such that for any $n \geq N$, and any radius r as in (R2), the Laplacian smoothing test $\widehat{\varphi}$ based on the estimator $\widehat{f}$ in (3.2), with $\rho = (nr^{d+2})^{-1} n^{-4/(4+d)} M^{-8/(4+d)}$, satisfies the following: for any $b \geq 1$, if*

$$\epsilon^2 \geq C M^{2d/(4+d)} n^{-4/(4+d)} \left(b^2 + b\sqrt{\frac{1}{\alpha}}\right), \tag{3.11}$$

then the worst-case risk satisfies the upper bound: $R_n(\widehat{\varphi}, H^1(\mathcal{X}, M), \epsilon) \leq C/b + C_1 n \exp(-c_1 n r^d)\big)$.

Some remarks:

- As mentioned earlier, Sobolev balls $H^1(\mathcal{X}, M)$ for $d \geq 4$ include quite irregular functions $f \notin L^4(\mathcal{X})$. Proving tight lower bounds in this case is nontrivial, and as far as we understand such an analysis remains outstanding. On the other hand, if we explicitly assume that $f_0 \in L^4(\mathcal{X}, M)$, then Guerre and Lavergne [2002] show that the testing problem is characterized by a dimension-free lower bound $\epsilon^2(L^4(\mathcal{X}, M)) \gtrsim n^{-1/2}$. Moreover, by setting $\rho = 0$ so that the resulting estimator $\widehat{f}$ interpolates the responses $Y_1, \ldots, Y_n$, the subsequent test $\widehat{\varphi}$ will achieve (up to constants) this lower bound. That is, for any $f_0 \in L^4(\mathcal{X}, M)$ such that $\|f_0\|^2_{L^2(\mathcal{X})} \geq C(b^2 + \sqrt{1/\alpha}) n^{-1/2}$, we have that $\mathbb{E}_0[\widehat{\varphi}] \leq \alpha$ and

$$\mathbb{E}_{f_0}\big[1 - \widehat{\varphi}\big] \leq \frac{C(1 + M^4)}{b^2}. \tag{3.12}$$

- To compute the data-dependent threshold $\widehat{t}_\alpha$, one must know all of the eigenvalues $\lambda_1, \ldots, \lambda_n$. Computing all these eigenvalues is far more expensive (cubic-time) than computing $\widehat{T}$ in the first place (nearly-linear-time). But in practice we would not recommend using $\widehat{t}_\alpha$ anyway, and would instead we make the standard recommendation to calibrate via a permutation test [Hoeffding, 1952]. Recent work Kim et al. [2020], has shown that in a variety of closely related settings, calibration of a test statistic via the permutation test often retains minimax-optimal power, and we expect similar results to hold for the Laplacian smoothing-based test statistic.

More Discussion of Variational Analog. With some results in hand, let us pause to offer some explanation of why Laplacian smoothing can be optimal in settings where thin-plate splines are not even consistent. First, we elaborate on why this difference in performance is so surprising. As mentioned previously, the penalties in (3.2), (3.4) can be closely tied together: Bousquet et al. [2004] show that for $f \in C^2(\mathcal{X})$,

$$\begin{aligned} \lim \frac{1}{n^2 r^{d+2}} f^\top L f &= \int_{\mathcal{X}} f(x) \cdot \Delta_P f(x) p(x)\, dx \\ &= \int_{\mathcal{X}} \|\nabla f(x)\|_2^2 p^2(x)\, dx. \end{aligned} \tag{3.13}$$

In the above, the limit is as $n \to \infty$ and $r \to 0$, Δ_P is the (weighted) Laplace-Beltrami operator

$$\Delta_P f := -\frac{1}{p} \mathrm{div}\big(p^2 \nabla f\big),$$

and the second equality follows using integration by parts.[2] To be clear, this argument does not formally imply that the Laplacian eigenmaps estimator $\widehat{f}$ and the thin-plate spline estimator $\widetilde{f}$ are close (for one, note that (3.13) holds for $f \in C^2(\mathcal{X})$, whereas the optimization in (3.4) considers a much broader set of continuous functions with weak derivatives in $L^2(\mathcal{X})$). But it does seem to suggest that the two estimators should behave somewhat similarly.

Of course, we know this is not the case: $\widehat{f}$ and $\widetilde{f}$ look very different when $d > 1$. What is driving this difference? The key point is that the discretization imposed by the graph $G_{n,r}$—which might seem problematic at first glance—turns out to be a blessing. The problem with (3.4) is that the class $H^1(\mathcal{X})$, which fundamentally underlies the criterion, is far "too big" for $d > 1$. This is meant in various related senses. By the Sobolev embedding theorem, for $d > 1$, the class $H^1(\mathcal{X})$ does not continuously embed into any Hölder space; and in fact it does not even continuously embed into $C^0(\mathcal{X})$. Thus we cannot really restrict the optimization to *continuous* and weakly differentiable functions, as we could when $d = 1$ (the smoothing spline case), without throwing out a substantial subset of functions in $H^1(\mathcal{X})$. Even among continuous and differentiable functions f, as we explained previously, we can use "bump" functions (as in Green and Silverman [1993]) to construct f that interpolates the pairs (X_i, Y_i), $i = 1, \ldots, n$ and achieves arbitrarily small penalty (and hence criterion) in (3.4). In this sense, any estimator resulting from solving (3.4) will clearly be inconsistent.

On the other hand, problem (3.2) is finite-dimensional. As a result $\widehat{f}$ has far less capacity to overfit than does $\widetilde{f}$, for any given sample size n. Discretization is not the only way to make the problem (3.4) more tractable: for instance, one can replace the penalty $\int_{\mathcal{X}} \|\nabla f(x)\|_2^2\, dx$ with a stricter choice like $\mathrm{ess\,sup}_{x \in \mathcal{X}} \|\nabla f(x)\|_2$, or conduct the optimization over some finite-dimensional linear subspace of $H^1(\mathcal{X})$ (i.e., use a sieve). While these solutions do improve the statistical properties of $\widetilde{f}$ for $d > 1$ (see e.g., Birgé and Massart [1993, 1998], van de Geer [2000]), Laplacian smoothing is generally speaking much simpler and more computationally friendly. In addition, the other approaches are usually specifically tailored to the domain $\mathcal{X}$, in stark contrast to $\widehat{f}$.

[2] Assuming f satisfies Neumann boundary conditions.

Overview of Analysis. The comparison with thin-plate splines highlights some surprising differences between $\widehat{f}$ and $\widetilde{f}$. Such differences also preclude us from analyzing $\widehat{f}$ by, say, using (3.13) to establish a coupling between $\widehat{f}$ and $\widetilde{f}$—we know this cannot work, because we would like to prove meaningful error bounds on $\widehat{f}$ in regimes where no such bounds exist for $\widetilde{f}$.

Instead we take a different approach, and directly analyze the error of $\widehat{f}$ and $\widehat{T}$ using a bias-variance decomposition (conditional on $X_1, \dots, X_n$). A standard calculation shows that

$$\|\widehat{f} - f_0\|_n^2 \leq \underbrace{\frac{2\rho}{n}(f_0^\top L f_0)}_{\text{bias}} + \underbrace{\frac{10}{n}\sum_{k=1}^{n} \frac{1}{(\rho\lambda_k + 1)^2}}_{\text{variance}},$$

and likewise that $\widehat{\varphi}$ has small risk whenever

$$\|f_0\|_n^2 \geq \underbrace{\frac{2\rho}{n}(f_0^\top L f_0)}_{\text{bias}} + \underbrace{\frac{2\sqrt{2/\alpha} + 2b}{n}\sqrt{\sum_{k=1}^{n} \frac{1}{(\rho\lambda_k + 1)^4}}}_{\text{variance}}.$$

The bias and variance terms are each functions of the random graph $G_{n,r}$, and hence are themselves random. To upper bound them, we build on some recent works [Burago et al., 2014, García Trillos et al., 2019a, Calder and García Trillos, 2019] regarding the consistency of neighborhood graphs to establish the following lemmas. These lemmas assume (P1), (P2) on the design distribution P, and (K1) on the kernel used to compute the neighborhood graph $G_{n,r}$.

Lemma 3. *There are constants $N, C_2 > 0$ such that for $n \geq N$, $r \leq c_0$, and $f \in H^1(\mathcal{X})$, with probability at least $1 - \delta$, it holds that*

$$f^\top L f \leq \frac{C_2}{\delta} n^2 r^{d+2} |f|^2_{H^1(\mathcal{X})}. \tag{3.14}$$

Lemma 4. *There are constants $N, C_1, C_3, c_1, c_3 > 0$ such that for $n \geq N$ and $C_0(\log n/n)^{1/d} \leq r \leq c_0$, with probability at least $1 - C_1 n \exp(-c_1 n r^d)$, it holds that*

$$c_3 A_{n,r}(k) \leq \lambda_k \leq C_3 A_{n,r}(k), \quad \text{for } 2 \leq k \leq n, \tag{3.15}$$

where $A_{n,r}(k) = \min\{n r^{d+2} k^{2/d}, n r^d\}$.

Lemma 3 gives a direct upper bound on the bias term. Lemma 4 leads to a sufficiently tight upper bound on the variance term whenever the radius r is sufficiently small; precisely, when r is upper bounded as in (R1) for estimation, or (R2) for testing. The parameter ρ is then chosen to minimize the sum of these upper bounds on bias and variance, as usual, and some straightforward calculations give Theorems 6-8.

It may be useful to give one more perspective on our approach. A common strategy in analyzing penalized least squares estimators is to assume two properties: first, that the regression function f_0 lies in (or near) a ball defined by the penalty operator; second, that this ball is reasonably small, e.g., as measured by metric entropy, or Rademacher complexity, etc. In contrast, in Laplacian smoothing, the penalty induces a ball

$$H^1(G_{n,r}, M) := \{f : f^\top L f \leq M^2\}$$

that is data-dependent and random, and so we do not have access to either of the aforementioned properties a priori, and instead, must prove they hold with high probability. In this sense, our analysis is different than the typical one in nonparametric regression.

3.5 Manifold Adaptivity

The minimax rates $n^{-2/(2+d)}$ and $n^{-4/(4+d)}$, in estimation and testing, suffer from the curse of dimensionality. However, in practice it can be often reasonable to assume a *manifold hypothesis*: that the data $X_1, \dots, X_n$ lie on a manifold $\mathcal{X}$ of $\mathbb{R}^d$ that has intrinsic dimension $m < d$. Under such an assumption, it is known [Bickel and Li, 2007, Arias-Castro et al., 2018] that the optimal rates over $H^1(\mathcal{X})$ are now $n^{-2/(2+m)}$ (for estimation) and $n^{-4/(4+m)}$ (for testing), which are much faster than the full-dimensional error rates when $m \ll d$.

On the other hand, a theory has been developed [Belkin, 2003, Belkin and Niyogi, 2008, Niyogi et al., 2008, Niyogi, 2013, Balakrishnan et al., 2012, 2013b] establishing that the neighborhood graph $G_{n,r}$ can "learn" the manifold $\mathcal{X}$ in various senses, so long as $\mathcal{X}$ is locally linear. We contribute to this line of work by showing that under the manifold hypothesis, Laplacian smoothing achieves the tighter minimax rates over $H^1(\mathcal{X})$.

Error Rates Assuming the Manifold Hypothesis. The conditions and results presented here will be largely similar to the previous ones, except with the ambient dimension d replaced by the intrinsic dimension m. For the remainder, we assume the following.

(P3) P is supported on a compact, connected, smooth manifold $\mathcal{X}$ embedded in $\mathbb{R}^d$, of dimension $m \leq d$. The manifold is without boundary and has positive reach [Federer, 1959].

(P4) P admits a density p with respect to the volume form of $\mathcal{X}$ such that

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

Additionally, p is Lipschitz on $\mathcal{X}$, with Lipschitz constant L_p.

Under the assumptions (P3), (P4), and (K1), and for a suitable range of r, the error bounds on the estimator $\widehat{f}$ and test $\widehat{\varphi}$ will depend on m instead of d.

(R4) For constants $C_0, c_0 > 0$, the neighborhood graph radius r satisfies

$$C_0 \left(\frac{\log n}{n} \right)^{\frac{1}{m}} \leq r \leq c_0 \wedge M^{\frac{(m-4)}{(4+2m)}} n^{\frac{-3}{(4+2m)}}.$$

Theorem 9. *As in Theorem 6, but where $\mathcal{X} \subseteq \mathbb{R}^d$ is a manifold with intrinsic dimension $m < 4$, the design distribution P obeys (P3), (P4), and $M \leq n^{1/m}$. There are constants $N, C, c > 0$ (not depending on f_0) such that for any $n \geq N$, and any r as in (R4), the Laplacian smoothing estimator $\widehat{f}$ in (3.2), with $L = L_{n,r}$ and $\rho = M^{-4/(2+m)}(nr^{m+2})^{-1} n^{-2/(2+m)}$, satisfies*

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C}{\delta} M^{2m/(2+m)} n^{-2/(2+m)},$$

with probability at least $1 - \delta - Cn\exp(-cnr^m) - \exp(-c(M^2 n)^{m/(2+m)})$.

In a similar vein, we obtain results for manifold adaptive testing under the following condition on the graph radius parameter.

(R5) For constants $C_0, c_0 > 0$, the neighborhood graph radius r satisfies

$$C_0 \left(\frac{\log n}{n} \right)^{\frac{1}{m}} \leq r \leq c_0 \wedge M^{\frac{(m-8)}{8+2m}} n^{\frac{m-20}{32+8m}}.$$

Theorem 10. *As in Theorem 8, but where $\mathcal{X} \subseteq \mathbb{R}^d$ is a manifold with intrinsic dimension $m < 4$, $M \leq M_{\max}(m)$, and the design distribution P obeys (P3), (P4). There are constants $N, C, c > 0$ such that for any $n \geq N$, and any r as in (R5), the Laplacian smoothing test $\widehat{\varphi}$ based on the estimator $\widehat{f}$ in (3.2), with $\rho = (nr^{m+2})^{-1} n^{-4/(4+m)} M^{-8/(4+m)}$, satisfies the following: for any $b \geq 1$, if*

$$\epsilon^2 \geq C M^{2m/(4+m)} n^{-4/(4+m)} \left(b^2 + b\sqrt{\frac{1}{\alpha}} \right), \tag{3.16}$$

then the worst-case risk satisfies the upper bound: $R_n(\widehat{\varphi}, H^1(\mathcal{X}, M), \epsilon) \leq C/b + Cn \exp(-cnr^m)$.

The proofs of Theorems 9 and 10 proceed in a similar manner to that of Theorems 6 and 8. The key difference is that in the manifold setting, the equations (3.14) and (3.15) used to upper bound bias and variance will hold with d replaced by m.

We emphasize that little about $\mathcal{X}$ need be known for Theorems 9 and 10 to hold. Indeed, all that is needed is the intrinsic dimension m, to properly tune r and ρ (from a theoretical point of view), and otherwise $\widehat{f}$ and $\widehat{\varphi}$ are computed without regard to $\mathcal{X}$. In contrast, the penalty in (3.4) would have to be specially tailored to work in this setting, revealing another advantage of the discrete approach over the variational one.

3.6 Discussion

We have shown that Laplacian smoothing, computed over a neighborhood graph, can be optimal for both estimation and goodness-of-fit testing over Sobolev spaces. There are many extensions worth pursuing, and several have already been mentioned. We conclude by mentioning a couple more. In practice, it is more common to use a k-nearest-neighbor (kNN) graph than a neighborhood graph, due to the guaranteed connectivity and sparsity of the former; we suspect that by building on the work of Calder and García Trillos [2019], one can show that our main results all hold under the kNN graph as well. In another direction, one can also generalize Laplacian smoothing by replacing the penalty $f^\top L f$ with $f^\top L^s f$, for an integer $s > 1$. The hope is that this would then achieve minimax optimal rates over the higher-order Sobolev class $H^s(\mathcal{X})$.

Chapter 4

Minimax Optimal Regression Over Sobolev Spaces via Laplacian Eigenmaps on Neighborhood Graphs

4.1 Introduction

Suppose we observe data $X_1, \ldots, X_n \in \mathbb{R}^d$, sampled independently from an unknown distribution P. As a replacement for P, we form a geometric graph G over the observed data, with vertices at $X_1, \ldots, X_n$ and weighted edges W_{ij} corresponding to proximity between samples X_i and X_j. Geometric graphs encode both local information and global information about P in an extremely general manner. For this reason they have been leveraged to conduct many different fundamental statistical tasks, such as clustering, manifold learning, semi-supervised learning, classification, and regression. Substantial theoretical progress has been made regarding the consistency of *graph-based learning*, that is, learning algorithms defined with respect to geometric graphs. This work sheds light on why such procedures work, by showing that they converge to interesting *continuum* limits as $n \to \infty$. However, thus far little has been said regarding the optimality of graph-based learning methods, even for classic statistical tasks.

In this paper we focus on the theoretical statistical properties of a particularly popular graph-based learning method for regression. We assume that in addition to the design points $X_1, \ldots, X_n$ one observes real-valued responses $Y_1, \ldots, Y_n$, and seeks to learn the unknown regression function $f_0(x) := E[Y|X = x]$. The specific graph-based method we study is *Laplacian eigenmaps*, first introduced by Belkin and Niyogi [2003], which projects the response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ onto the span of eigenvectors of a graph Laplacian. We focus on the the unnormalized graph Laplacian L, which is a difference operator acting on vectors $u \in \mathbb{R}^n$ as follows,

$$(Lf)_i = \sum_{j=1}^{n} (u_i - u_j) W_{ij}. \tag{4.1}$$

The graph Laplacian L is a discrete approximation to a weighted continuum Laplacian operator Δ_P, defined when P admits a differentiable density p as

$$\Delta_P f = -\frac{1}{p} \mathrm{div}(p^2 \nabla f). \tag{4.2}$$

The eigenvectors of L form an orthonormal basis of $\mathbb{R}^n$, and serve as estimates of eigenfunctions of Δ_P. Their corresponding eigenvalues are likewise estimates of eigenvalues of Δ_P, and give a notion of smoothness to each eigenvector—roughly speaking, the smaller the eigenvalue, the smoother the corresponding eigenvector.

We can therefore view Laplacian eigenmaps as a twist on a very classical approach to nonparametric regression: *spectral projection*, or more generally *orthogonal series*, regression. Classically, in orthogonal series regression one fixes a reference measure Q, takes an ordered orthonormal basis $\psi_1, \psi_2, \ldots$ of $L^2(Q)$, computes empirical Fourier coefficients $\langle \mathbf{Y}, \psi_k \rangle_n$, and uses the first few terms in the resulting Fourier series to construct an estimator. Regression by spectral projection is a special case of this general setup, where one takes ψ_k to be the kth eigenfunction of the continuum Laplacian operator Δ_Q. In contrast to this classical approach—in which the reference measure Q and resulting Laplacian Δ_Q are determined a priori—in Laplacian eigenmaps the eigenvectors of the graph Laplacian serve as the basis. These eigenvectors are data-dependent objects, and adapt to the geometry of the unknown design distribution P in a rich manner. For instance, they respect the cluster structure of P, meaning that if the density p has multiple connected components, the first few eigenvectors of L will (with high probability) be piecewise constant over each such component Von Luxburg [2007]. On the other hand if P is supported on some low-dimensional manifold, the graph Laplacian eigenvectors concentrate around the eigenfunctions of a manifold Laplace-Beltrami operator, and thus give a principled embedding of potentially high dimensional design points X_i into a lower dimensional space [Belkin and Niyogi, 2003].

Thus Laplacian eigenmaps is a data-dependent alternative to classical spectral projection estimators. This data-dependency is appealing. Classical spectral projection estimators possess attractive theoretical properties for nonparametric regression—in particular, they are minimax rate-optimal for regression over Hölder and Sobolev spaces Tsybakov [2008a], Johnstone [2011], Giné and Nickl [2016]—but in practice suffer from some serious drawbacks. A basic difficulty is that finding the eigenfunctions of Δ_Q is in general non-trivial, so that it may not be possible to compute the estimator in the first place. Moreover, spectral projection estimators make sense (and are minimax optimal) only if the reference measure Q is very close or equal to P, since otherwise the basis functions are orthogonalized with respect to the wrong measure. For these reasons, such estimators are typically proposed and studied under very restrictive conditions on the design points: for instance, that they are equally spaced fixed grid points, or that they are random but uniformly distributed on the unit cube $[0,1]^d$. There do exist fixes to the issues just raised, for instance using nonparametric least-squares. We will compare Laplacian eigenmaps to nonparametric least squares in more detail later, and for now merely point out that it fundamentally changes the estimator under consideration.

In contrast, Laplacian eigenmaps directly approximates a spectral projection method: it projects the responses onto an orthogonal basis (eigenvectors of the graph Laplacian) that approximates the smooth eigenfunctions of Δ_P. Laplacian eigenmaps is perfectly well-defined, and indeed straightforward to compute, when the design P is unknown. This includes the situation where P may be non-uniform, or even concentrated on a low-dimensional manifold. On the other hand, because graph Laplacian eigenvectors depend on the random design points $X_1, \ldots, X_n$ in complicated ways, it is substantially more difficult to analyze Laplacian eigenmaps than to analyze classical spectral projection methods. For this reason, the theoretical statistical properties of Laplacian eigenmaps, and in particular its optimality as a method for nonparametric regression with random design, remain poorly understood.

Our contributions. The primary contribution of our paper is to fill this theoretical gap, by answering the following question:

Is Laplacian eigenmaps an optimal method for nonparametric regression over Sobolev functions?

Broadly seeking, our answer is yes. We consider two different data models, one of which assumes the support $\mathcal{X}$ of the distribution P is a full-dimensional, flat Euclidean domain, and the other of which assumes that $\mathcal{X}$ is a manifold of small intrinsic dimension m. We show that when the regression function f_0 is smooth, in the Sobolev sense of having weak derivatives bounded in $L^2(\mathcal{X})$ norm, Laplacian eigenmaps methods

Smoothness order	Flat Euclidean (Model 4.2.1)	Manifold (Model 4.2.2)
$s \leq 3$	$\boldsymbol{n^{-2s/(2s+d)}}$	$\boldsymbol{n^{-2s/(2s+m)}}$
$s > 3$	$\boldsymbol{n^{-2s/(2s+d)}}$	$n^{-6/(6+m)}$

Table 4.1: Summary of estimation rates over Sobolev balls. Bold font marks minimax optimal rates. In each case, rates hold for all $d \in \mathbb{N}$ (under Model 4.2.1), and for all $m \in \mathbb{N}, 1 < m < d$ (under Model 4.2.2). Although we suppress it for simplicity, in all cases when the Laplacian eigenmaps estimator is optimal, the dependence of the error rate on the radius M of the Sobolev ball is also optimal, as long as $n^{-1/2} \lesssim M \lesssim n^{1/d}$.

Smoothness order	Dimension	Flat Euclidean (Model 4.2.1)	Manifold (Model 4.2.2)
$s = 1$	$\dim(\mathcal{X}) < 4$	$\boldsymbol{n^{-4s/(4s+d)}}$	$\boldsymbol{n^{-4s/(4s+m)}}$
	$\dim(\mathcal{X}) \geq 4$	$\boldsymbol{n^{-1/2}}$	$\boldsymbol{n^{-1/2}}$
$s = 2$ or 3	$\dim(\mathcal{X}) \leq 4$	$\boldsymbol{n^{-4s/(4s+d)}}$	$\boldsymbol{n^{-4s/(4s+m)}}$
	$4 < \dim(\mathcal{X}) < 4s$	$n^{-2s/(2(s-1)+d)}$	$n^{-2s/(2(s-1)+m)}$
	$\dim(\mathcal{X}) \geq 4s$	$\boldsymbol{n^{-1/2}}$	$\boldsymbol{n^{-1/2}}$
$s > 3$	$\dim(\mathcal{X}) \leq 4$	$\boldsymbol{n^{-4s/(4s+d)}}$	$n^{-12/(12+d)}$
	$4 < \dim(\mathcal{X}) < 4s$	$n^{-2s/(2(s-1)+d)}$	$n^{-6/(4+m)}$
	$\dim(\mathcal{X}) \geq 4s$	$\boldsymbol{n^{-1/2}}$	$\boldsymbol{n^{-1/2}}$

Table 4.2: Summary of Laplacian eigenmaps testing rates over Sobolev balls. Bold font marks minimax optimal rates. Rates when $d > 4s$ assume that $f_0 \in L^4(P, M)$. Although we suppress it for simplicity, in all cases when the Laplacian eigenmaps test is optimal, the dependence of the error rate on the radius M of the Sobolev ball is also optimal, as long as $n^{-1/2} \lesssim M \lesssim n^{1/d}$.

are statistically minimax optimal, for both estimation and goodness-of-fit testing. Our statements hold for different relations between the dimension d (or m) and number of derivatives s, depending on the problem (estimation or testing), and are summarized in Tables 4.1 and 4.2.

Related Work. There is an incredible amount of work on the properties of classical spectral projection methods, which go far beyond their optimality for nonparametric regression. We will not attempt to summarize this work, and instead refer to Wasserman [2006], Györfi et al. [2006], Tsybakov [2008a], Johnstone [2011], Giné and Nickl [2016] and the references therein.

More recent work has considered regression on a fixed graph, where one treats the design points $x_1, \ldots, x_n$ as vertices in a fixed graph G, and carries out inference with respect to the function evaluations $(f_0(x_i))_{i=1}^n$ By this point there exists a relatively mature theory describing this setting. Tight upper bounds have been established that certify the optimality of graph-based methods for both nonparametric estimation [Wang et al., 2016, Hütter and Rigollet, 2016, Sadhanala et al., 2016a, 2017, Kirichenko and van Zanten, 2017, Kirichenko et al., 2018]) and testing (e.g., Sharpnack and Singh [2010], Sharpnack et al. [2013a,b, 2015] over different "function" classes (in quotes because these classes really model the n-dimensional vector of evaluations) We call particular attention to Sadhanala et al. [2016a] and Sharpnack et al. [2015], who analyze the Laplacian eigenmaps estimator and test statistic, respectively. This setting is quite general, because the graph need not be a geometric graph defined on a vertex set which belongs to Euclidean space. On the other hand, in many situations it may be somewhat to unnatural to assume that the design points are a priori fixed, and that the regression function f_0 exhibits "smoothness" over this fixed design. Instead, it may be more reasonable to adopt the *random design* perspective that we work in, and assume that the regression function f_0 exhibits a more classical notion of smoothness.

However, as already mentioned, when the design points are random so too are the graph Laplacian eigenvectors, and grasping their properties is in general non-trivial. For this reason, there has not been much

analysis of random design nonparametric regression using Laplacian eigenmaps. Zhou and Srebro [2011] consider the Laplacian eigenmaps estimator, but in the semi-supervised setting, where one additionally observes unlabeled design points $X_{n+1}, \dots, X_{n+m}$. Their analysis assumed that for a fixed number of labeled samples n, the number of unlabeled samples m grows to infinity. In this case the eigenvectors of the graph Laplacian converge to eigenfunctions of a continuum Laplacian, and the analysis of the resulting estimator is identical to that of a classical spectral projection estimator. Lee et al. [2016] consider the *diffusion maps* estimator—which uses the eigenvectors of a different normalization of the graph Laplacian L—in both the supervised and semi-supervised setups. In the supervised case, they show that the diffusion maps estimator converges to the regression function as the sample size $n \to \infty$, but at a suboptimal rate. As far as we know, there has been no analysis of the test statistic $\widehat{T}$ in the random design framework which we study.

There has been much more analysis of the convergence properties of eigenvectors of random graph Laplacians to their continuum limits. Belkin and Niyogi [2007], von Luxburg et al. [2008], Singer and Wu [2017], García Trillos and Slepčev [2018a] show that eigenvalue-eigenvector pairs (λ, v) of a graph Laplacian converge to eigenvalue-eigenfunction pairs (λ, ψ) of a limiting differential or integral operator. Burago et al. [2014], Shi [2015], García Trillos et al. [2019a], Calder and García Trillos [2019], Cheng and Wu [2021] build on these works, by giving finite sample bounds, rates of convergence, and making statements uniform. These results justify our intuition that Laplacian eigenmaps, which projects the responses $\mathbf{Y}$ onto eigenvectors of a graph Laplacian, is approximating a classical spectral projection method, which projects $\mathbf{Y}$ onto eigenfunctions of the limiting differential operator Δ_P. In fact, more formally these results imply that for a fixed number of eigenvectors $K > 0$, the Laplacian eigenmaps estimator converges to its classical counterpart as $n \to \infty$. By taking $K \to \infty$ sufficiently slowly with n, we can even conclude that Laplacian eigenmaps will be a consistent estimator of f_0. Unfortunately, the resulting rates of convergence implied by this approach are severely suboptimal, compared to the minimax rates. Our analysis showing that Laplacian eigenmaps achieves minimax optimal rates of convergence will use some of the results mentioned above, but will overall proceed by a very different route than the one just considered.

Finally, we point out that there are other ways to use neighborhood graphs, and specifically graph Laplacians, for nonparametric regression. For instance, García Trillos and Murray [2020], Green et al. [2021] use the graph Laplacian to induce a penalty over functions $f : \{\mathbf{X}\} \to \mathbb{R}$. The *Laplacian smoothing* estimator $\widehat{f}_{\text{LS}}$ is obtained by minimizing the sum of this penalty with a data-fidelity term,

$$\widehat{f}_{\text{LS}} := \|Y - f\|_n^2 + \lambda \langle L_{n,\varepsilon} f, f \rangle_n.$$

Green et al. [2021] show that the resulting estimator is minimax optimal, but only for $s = 1$ and $d \leq 4$. In contrast, we show in this work that the Laplacian eigenmaps estimator is optimal for all s and d.

Organization. We now outline the structure of the rest of this paper. In Section 4.2, we formally define the regression problem and estimator we consider. We also give some background about minimax regression over Sobolev spaces, and recall the classical spectral projection estimators which achieve minimax rates of convergence. In Sections 4.3 and 4.4, we establish our main results regarding the optimality of Laplacian eigenmaps. Since thhe Laplacian eigenmaps estimator $\widehat{f}$ is defined only at the design points $X_1, \dots, X_n$, in Section 4.5 we propose a method for out-of-sample extension of $\widehat{f}$, and show that it has optimal out-of-sample error. In Section 4.6 we examine the empirical behavior of Laplacian eigenmaps, and compare it to some natural competitors for nonparametric regression. We conclude with some discussion of future work in Section 4.7.

Notation. We frequently refer to various classical function classes. For a domain $\mathcal{X}$ with volume form $d\mu$, we let $L^2(\mathcal{X})$ denote the set of functions f for which $\|f\|_{L^2(\mathcal{X})}^2 := \int f^2 \, d\mu < \infty$, and equip $L^2(\mathcal{X})$ with the norm $\|\cdot\|_{L^2(\mathcal{X})}$. We define $\langle f, g \rangle_P := \int fg \, dP$, and let $L^2(P)$ contain those functions f for which $\|f\|_P^2 := \langle f, f \rangle_P$ is finite. Finally, we let $L^2(P_n)$ consist of those "functions" $f : \{X_1, \dots, X_n\} \to \mathbb{R}$ for which the empirical norm $\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^n \big(f(X_i)\big)^2 < \infty$. When there is no chance of confusion, we will sometimes

associate functions in $L^2(P_n)$ with vectors in $\mathbb{R}^n$, and vice versa. We use $C^k(\mathcal{X})$ to refer to functions which are k times continuously differentiable in $\mathcal{X}$, either for some integer $k \geq 1$ or for $k = \infty$. We let $C_c^\infty(\mathcal{X})$ represent those functions in $C^\infty(\mathcal{X})$ which are compactly contained in $\mathcal{X}$. We write $\partial f / \partial r_i$ for the partial derivative of f in the ith standard coordinate of $\mathbb{R}^d$, and use the multi-index notation $D^\alpha f := \partial^{|\alpha|} f / \partial^{\alpha_1} x_1 \dots \partial^{\alpha_d} x_d$ for multi-indices $\alpha \in \mathbb{R}^m$.

We write $\|\cdot\|$ for Euclidean, and $d_{\mathcal{X}}(x', x)$ for the geodesic distance between points x and x' on a manifold $\mathcal{X}$. Then for a given $\delta > 0$, $B(x, \delta)$ is the radius-δ ball with respect to Euclidean distance, whereas $B_{\mathcal{X}}(x, \delta)$ is the radius-δ ball with respect to geodesic distance. Letting $T_x(\mathcal{X})$ be the tangent space at a point $x \in \mathcal{X}$, we write $B_m(v, \delta) \subset T_x(\mathcal{X})$ for the radius-δ ball centered at $v \in T_x(\mathcal{X})$.

For sequences (a_n) and (b_n), we use the asymptotic notation $a_n \lesssim b_n$ to mean that there exists a number C such that $a_n \leq C b_n$ for all n. We write $a_n \asymp b_n$ when $a_n \lesssim b_n$ and $b_n \lesssim a_n$. On the other hand we write $a_n = o(b_n)$ when $\lim a_n / b_n = 0$, and likewise $a_n = \omega(b_n)$ when $\lim a_n / b_n = \infty$. Finally $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

4.2 Setup and Background

In this section, we begin by giving a precise definition of our framework, and the Laplacian eigenmaps methods we study. We then review minimax rates for nonparametric regression over Sobolev spaces. We pay special attention to the classical spectral projection methods which achieve these rates, since such methods are closely connected to Laplacian eigenmaps.

4.2.1 Nonparametric regression with random design

We will operate in the usual setting of nonparametric regression with random design, in which we observe independent random samples $(X_1, Y_1), \dots, (X_n, Y_n)$. The design points $X_1, \dots, X_n$ are sampled from a distribution P with support $\mathcal{X} \subseteq \mathbb{R}^d$, and the responses follow the signal plus noise model

$$Y_i = f_0(X_i) + w_i, \tag{4.3}$$

with regression function $f_0 : \mathcal{X} \to \mathbb{R}$, and $w_i \sim N(0, 1)$ independent Gaussian noise. For simplicity we will assume throughout that the noise has unit-variance, but all of our results extend in a straightforward manner to the case where the variance is equal to a known positive value.

We now formulate two models, which differ in the assumed nature of the support $\mathcal{X}$ of the design distribution P: the *flat Euclidean* and *manifold* models.

Definition 4.2.1 (Flat Euclidean model)**.** The data $(X_1, Y_1), \dots, (X_n, Y_n)$ are sampled according to (4.3). The support $\mathcal{X}$ of the design distribution P is an open, connected, and bounded subset of $\mathbb{R}^d$, with Lipschitz boundary. The distribution P admits a Lipschitz density p with respect to the d-dimensional Lebesgue measure ν, which is bounded away from 0 and ∞,

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

In the following, we recall that the injectivity radius of a m-dimensional Riemannian manifold $\mathcal{X}$ is the maximum value of δ such that the exponential map $\exp_x : B_m(0, \delta) \subset T_x(\mathcal{X}) \to B_{\mathcal{X}}(x, \delta) \subset \mathcal{X}$ is a diffeomorphism for all $x \in \mathcal{X}$.

Definition 4.2.2 (Manifold model)**.** The data $(X_1, Y_1), \dots, (X_n, Y_n)$ are sampled according to (4.3). The support $\mathcal{X}$ of the design distribution P is a closed, connected, smooth and boundaryless Riemannian manifold embedded in $\mathbb{R}^d$, of intrinsic dimension $1 \leq m < d$. The injectivity radius of $\mathcal{X}$ is lower bounded by a positive constant $i_0 > 0$. The design distribution P admits a Lipschitz density p with respect to the volume form $d\mu$ induced by the Riemannian structure of $\mathcal{X}$, which is bounded away from 0 and ∞,

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

Finally, at various points we will have to assume that the density p also displays different types of higher-order regularity, beyond Lipschitz continuity. When such assumptions are necessary, we always will state them explicitly.

4.2.2 Laplacian eigenmaps

We now formally define the estimator and test statistic we study. Both are derived from eigenvectors of a graph Laplacian. For a positive, symmetric kernel $\eta : [0, \infty) \to [0, \infty)$, and a radius parameter $\varepsilon > 0$, let $G = ([n], W)$ be the neighborhood graph formed over the design points $\{X_1, \dots, X_n\}$, with a weighted edge $W_{ij} = \eta(\|X_i - X_j\|/\varepsilon)$ between vertices i and j. Then the *neighborhood graph Laplacian* $L_{n,\varepsilon} : \mathbb{R}^n \to \mathbb{R}$ is defined by its action on vectors $u \in \mathbb{R}^n$ as

$$\big(L_{n,\varepsilon} u\big)_i := \frac{1}{n\varepsilon^{2+\dim(\mathcal{X})}} \sum_{j=1}^{n} \big(u_i - u_j\big) \eta\bigg(\frac{\|X_i - X_j\|}{\varepsilon}\bigg). \tag{4.4}$$

(Here $\dim(\mathcal{X})$ stands for the dimension of $\mathcal{X}$. It is equal to d under the assumptions of Model 4.2.1, and equal to m under the assumptions of Model 4.2.2. The pre-factor $(n\varepsilon^{2+\dim(\mathcal{X})})^{-1}$ is purely for convenience in taking limits as $n \to \infty, \varepsilon \to 0$). Written in standard coordinates we have $(n\varepsilon^{\dim(\mathcal{X})+2}) \cdot L_{n,\varepsilon} = D - W$, where $D \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix, $D_{ii} = \sum_{i=1}^{n} W_{ij}$.

The graph Laplacian is a positive semi-definite matrix, and admits the eigendecomposition $L_{n,\varepsilon} = \sum_{k=1}^{n} \lambda_k v_k v_k^\top$, where for each $k = 1, \dots, n$ the eigenvalue-eigenvector pair (λ_k, v_k) satisfies

$$\frac{1}{n} L_{n,\varepsilon} v_k = \lambda_k v_k, \quad \|v_k\|_n^2 = 1.$$

We will assume without loss of generality that each eigenvalue λ of $L_{n,\varepsilon}$ has algebraic multiplicity 1, and so we can index the eigenpairs $(\lambda_1, v_1), \dots, (\lambda_n, v_n)$ in ascending order of eigenvalue, $0 = \lambda_1 < \ldots < \lambda_n$.

The Laplacian eigenmaps estimator $\widehat{f}$ simply projects the response vector $\mathbf{Y}$ onto the first K eigenvectors of $L_{n,\varepsilon}$: letting $V_K \in \mathbb{R}^{n \times K}$ be the matrix with columns $v_1, \dots, v_K$, we have that

$$\widehat{f} := \sum_{k=1}^{K} \langle Y, v_k \rangle_n v_k = \frac{1}{n} V_K V_K^\top Y. \tag{4.5}$$

Thus $\widehat{f}$ is equivalently a vector in $\mathbb{R}^n$, or a function in $L^2(P_n)$. If $\widehat{f}$ is a reasonable estimate of f_0, then the Laplacian eigenmaps test statistic

$$\widehat{T} := \|\widehat{f}\|_n^2 \tag{4.6}$$

is in turn a reasonable estimate of $\|f_0\|_P^2$, and can be used to distinguish whether or not $f_0 = 0$.

It may be helpful to comment briefly on the term "Laplacian eigenmaps", which we use a bit differently than is typical in the literature. Laplacian eigenmaps typically refers to an algorithm for embedding, which maps each design point $X_1, \dots, X_n$ to $\mathbb{R}^K$ according to $X_i \mapsto (v_{1,i}, \dots, v_{K,i})$. Viewing this embedding as a feature map, we can then interpret the estimator $\widehat{f}$ as the least-squares solution to a linear regression problem with responses $Y_1, \dots, Y_n$ and features $v_1, \dots, v_K$. Often, the Laplacian eigenmaps embedding is viewed as a tool for dimensionality reduction, wherein it is implicitly assumed that K is much smaller than d. We will neither explicitly nor implicitly take $K < d$; indeed, the embedding perspective is not particularly illuminating in what follows, and we do not henceforth make reference to it. Instead, we use "Laplacian eigenmaps" to directly refer to the estimator $\widehat{f}$ or test statistic $\widehat{T}$.

4.2.3 Sobolev Classes

We now review the definition of Sobolev classes, dividing our discussion into two cases—the flat Euclidean case and the manifold case, corresponding respectively to Models 4.2.1 and Model 4.2.2.

Flat Euclidean case. Take $\mathcal{X}$ to be a open, connected and bounded set with Lipschitz boundary, as in Model 4.2.1. Recall that for a given multiindex $\alpha \in \mathbb{N}^d$, a function f is *α-weakly differentiable* if there exists some $h \in L^1(\mathcal{X})$ such that

$$\int_{\mathcal{X}} hg = (-1)^{|\alpha|} \int_{\mathcal{X}} f D^\alpha g, \quad \text{for every } g \in C_c^\infty(\mathcal{X}).$$

If such a function h exists, it is the αth weak partial derivative of f, and denoted by $D^\alpha f := h$.

Definition 4.2.3 (Sobolev space on an open set)**.** For an integer $s \geq 1$, a function $f \in L^2(\mathcal{X})$ belongs to the Sobolev space $H^s(\mathcal{X})$ if for all $|\alpha| \leq s$, the weak derivatives $D^\alpha f$ exists and satisfy $D^\alpha f \in L^2(\mathcal{X})$. The jth order semi-norm for $f \in H^s(\mathcal{X})$ is $|f|_{H^j(\mathcal{X})} := \sum_{|j|=s} \|D^\alpha f\|_{L^2(\mathcal{X})}$, and the corresponding norm

$$\|f\|_{H^s(\mathcal{X})}^2 := \|f\|_{L^2(\mathcal{X})}^2 + \sum_{j=1}^{s} |f|_{H^j(\mathcal{X})}^2,$$

induces a ball

$$H^s(\mathcal{X}; M) := \big\{ f \in H^s(\mathcal{X}) : \|f\|_{H^s(\mathcal{X})} \leq M \big\}.$$

We note that $H^s(\mathcal{X})$ is the completion of $C^\infty(\mathcal{X})$ with respect to the $\|\cdot\|_{H^s(\mathcal{X})}$ norm, so that $C^\infty(\mathcal{X})$ is dense in $H^s(\mathcal{X})$.

Manifold case. There are several equivalent ways to define Sobolev spaces on a compact, smooth, m-dimensional Riemannian manifold embedded in $\mathbb{R}^d$. We will stick with a definition that parallels our setup in the flat Euclidean setting as much as possible. To do so, we first recall the notion of partial derivatives on a manifold, which are defined with respect to a local coordinate system. Letting $r_1, \ldots, r_m$ be the standard basis of $\mathbb{R}^m$, for a given chart (ϕ, U) (meaning an open set $U \subseteq \mathcal{X}$, and a smooth mapping $\phi : U \to \mathbb{R}^m$) we write $\phi =: (x_1, \ldots, x_m)$ in local coordinates, meaning $x_i = r_i \circ \phi$. Then the partial derivative $\partial f / \partial x_i$ of a function f with respect to x_i at $x \in U$ is

$$\frac{\partial f}{\partial x_i}(x) := \frac{\partial (f \circ \phi^{-1})}{\partial r_i}\big(\phi(x)\big).$$

The right hand side should be interpreted in the weak sense of derivative. As before, we use the multi-index notation $D^\alpha f := \partial^{|\alpha|} f / \partial^{\alpha_1} x_1 \ldots \partial^{\alpha_m} x_m$.

Definition 4.2.4 (Sobolev space on a manifold)**.** A function $f \in L^2(\mathcal{X})$ belongs to the Sobolev space $H^s(\mathcal{X})$ if for all $|\alpha| \leq s$, the weak derivatives $D^\alpha f$ exist and satisfy $D^\alpha f \in L^2(\mathcal{X})$. The jth order semi-norm $|f|_{H^j(\mathcal{X})}$, the norm $\|f\|_{H^s(\mathcal{X})}$, and the ball $H^s(\mathcal{X}; M)$ are all defined as in Defintion 4.2.3.

The partial derivatives $D^\alpha f$ will depend on the choice of local coordinates, and so will the resulting Sobolev norm $\|f\|_{H^s(\mathcal{X})}$. However, regardless of the choice of local coordinates the resulting norms will be equivalent[1] and so the ultimate Sobolev space $H^s(\mathcal{X})$ is independent of the choice of local coordinates.

Boundary conditions. In the flat Euclidean model (Model 4.2.1), in order to show that Laplacian eigenmaps is optimal over $H^s(\mathcal{X})$ for $s > 1$ we will need to assume that the regression function f_0 satisfies some boundary conditions. In particular, we will assume that f_0 is zero-trace.

[1] Recall that norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a space $\mathcal{F}$ are said to be equivalent if there exist constants c and C such that

$$c\|f\|_1 \leq \|f\|_2 \leq C\|f\|_2 \quad \text{for all} f \in \mathcal{F}.$$

Definition 4.2.5 (Zero-trace Sobolev space)**.** The *order-s zero-trace Sobolev space* $H_0^s(\mathcal{X})$ is the closure of $C_c^\infty(\mathcal{X})$ with respect to $\|\cdot\|_{H^s(\mathcal{X})}$ norm. That is, $f \in H_0^s(\mathcal{X})$ if $f \in H^s(\mathcal{X})$ and additionally there exists a sequence $f_1, f_2, \ldots$ of functions in $C_c^\infty(\mathcal{X})$ such that

$$\lim_{k\to\infty} \|f_k - f\|_{H^s(\mathcal{X})} = 0.$$

The normed ball $H_0^s(\mathcal{X}; M) := H_0^s(\mathcal{X}) \cap H^s(\mathcal{X}; M)$.

The zero-trace condition can be made more concrete when $f \in C^\infty(\mathcal{X})$, since we can then speak of the pointwise behavior of f and its derivatives. Letting $\partial/(\partial \mathbf{n})$ be the partial derivative operator in the direction of the vector $\mathbf{n}$ normal to the boundary of $\mathcal{X}$, then the zero-trace condition implies that $\partial^k f/\partial \mathbf{n}^k(x) = 0$ for each $k = 0, \ldots, s-1$, and for all $x \in \partial\mathcal{X}$.

We now explain why Laplacian eigenmaps should be optimal only when f_0 satisfies certain boundary conditions. Let $(\lambda_1(\Delta_P), \psi_1), (\lambda_2(\Delta_P), \psi_2), \ldots$ be the solutions to the weighted Laplacian eigenvector equation with Neumann boundary conditions,

$$\Delta_P \psi_k = \lambda_k(\Delta_P)\psi_k, \quad \frac{\partial}{\partial \mathbf{n}}\psi_k = 0 \ \text{ on } \partial\mathcal{X}. \tag{4.7}$$

As we have already alluded to, it is known [García Trillos and Slepčev, 2018a] that each graph Laplacian eigenpair (λ_k,v_k) converges to a corresponding solution ($\lambda_k(\Delta_P)$, ψ_k) of (4.7). Thus it is relevant to consider which Sobolev functions $f \in H^s(M)$ we can reconstruct using the eigenfunctions $\psi_1, \psi_2, \ldots$. To that end, we introduce the spectrally defined Sobolev space

$$\mathcal{H}^s(\mathcal{X}) := \left\{ f \in L^2(\mathcal{X}) : \sum_{k=1}^{\infty} \big[\langle f, \psi_k\rangle_P\big]^2 \cdot \big[\lambda_k(\Delta_P)\big]^s < \infty \right\}. \tag{4.8}$$

Under the conditions $p \in C^\infty(\mathcal{X})$ and $\partial\mathcal{X} \in C^{1,1}$, Dunlop et al. [2020] show the strict inclusion $\mathcal{H}^{2s}(\mathcal{X}) \subset H^{2s}(\mathcal{X})$. More precisely, they show that for any $s > 0$,

$$\mathcal{H}^{2s}(\mathcal{X}) = \left\{ f \in H^{2s}(\mathcal{X}) : \frac{\partial \Delta_P^r f}{\partial \mathbf{n}} = 0 \text{ on } \partial\mathcal{X}, \ \text{ for all } 0 \le r \le s-1 \right\}, \tag{4.9}$$

and for any $s \ge 0$, $\mathcal{H}^{2s+1}(\mathcal{X}) = \mathcal{H}^{2s}(\mathcal{X}) \cap H^{2s+1}(\mathcal{X})$. If P is uniform on $\mathcal{X}$, then (4.8) means that a Sobolev function $f \in H^s(\mathcal{X})$ additionally belongs to $\mathcal{H}^s(\mathcal{X})$ only if all its odd lower order derivatives vanish at the boundary $\partial\mathcal{X}$.

The bottom line is that the eigenvectors v_k of the graph Laplacian accurately approximate only those functions $f \in H^s(\mathcal{X})$ which satisfy some additional boundary conditions. Although the zero-trace boundary condition is more restrictive than (4.9)—since it also requires that derivatives of even-order be equal to 0—the point is that some kind of boundary condition will be necessary in order to obtain optimal rates. We will stick to the zero-trace condition, since it greatly eases some of the steps in our proofs.

On the other hand, in the manifold model (Model 4.2.2) the domain $\mathcal{X}$ is without boundary—precisely, every point $x \in \mathcal{X}$ has a neighborhood that is homeomorphic to an open set in $\mathbb{R}^m$, for instance the open ball $B(x, \delta)$ for any δ smaller than the injectivity radius i_0—and so boundary conditions are irrelevant.

4.2.4 Minimax Rates and Spectral Series Methods

We now review the minimax estimation and goodness-of-fit testing rates over Sobolev balls. We will pay special attention to certain classical spectral projection methods which achieve these rates. This is because, as we have already discussed, classical spectral projection methods are very related to Laplacian eigenmaps.

Estimation rates over Sobolev balls. In the estimation problem, we ask for an estimator—formally, a Borel measurable function $\widehat{f}$ that maps from $\mathcal{X}$ to $\mathbb{R}$—which is close to the regression function f_0 with respect to the squared norm $\|\widehat{f} - f_0\|_P^2$. Under Model 4.2.1, the minimax estimation rate over the order-s Sobolev ball is

$$\inf_{\widehat{f}} \sup_{f_0} \mathbb{E}\|\widehat{f} - f_0\|_P^2 \asymp M^2(M^2 n)^{-2s/(2s+d)}; \tag{4.10}$$

here infimum taken over all estimators $\widehat{f}$, and the supremum over all $f_0 \in H^1(\mathcal{X}; M)$ (first-order case) or $f_0 \in H_0^s(\mathcal{X}; M)$ (higher-order case), and we assume $n^{-1/2} \lesssim M$.

The lower bound in (4.10) is due to [Stone, 1980] (at least for the case of M constant, as is most typically considered). The upper bound can be certified by a particular spectral projection estimator $\widetilde{f}$,

$$\widetilde{f} := \sum_{k=1}^{K} \langle Y, \psi_k \rangle_n \psi_k, \tag{4.11}$$

where ψ_k are the eigenfunctions of Δ_P defined in (4.7). The optimality of spectral projection estimators over Sobolev type spaces is generally well-understood. For instance, see Tsybakov [2008a], Johnstone [2011], Giné and Nickl [2016], who work in the Gaussian sequence model and show that analogous estimators are optimal over Sobolev ellipsoids. However, we have not found an analysis of the specific estimator (4.11) under Model 4.2.1, and so for completeness we state the result in the following proposition.

Proposition 6. *Suppose Model 4.2.1, and additionally that $\partial\mathcal{X} \in C^{1,1}$, that $p \in C^\infty(\mathcal{X})$, and that $f_0 \in H^1(\mathcal{X}; M)$ (first-order case) or $f_0 \in H_0^s(\mathcal{X}; M)$ for some $s > 1$ (higher-order case). Then there exists a constant C which does not depend on f_0, M or n such that the following statement holds: if the spectral projection estimator $\widetilde{f}$ is computed with parameter $K = \lfloor M^2 n \rfloor^{d/(2s+d)}$, then*

$$\mathbb{E}\big[\|\widetilde{f} - f_0\|_P^2\big] \leq C \min\big\{M^2(M^2 n)^{-2s/(2s+d)}, M^2\big\}$$

The proof of Proposition 6, along with proofs of all of our results, can be found in the appendix. It is worth mentioning some aspects of the analysis here, because it sets the stage for the strategy we will use to analyze Laplacian eigenmaps.

In particular, three essential facts are needed to establish Proposition 6.

1. The continuous embedding of $H_0^s(\mathcal{X})$ into $\mathcal{H}^s(\mathcal{X})$—recall the latter is defined in (4.8)—which is a consequence of the zero-trace condition, the conditions on $\partial\mathcal{X}$ and p, and (4.9).
2. Weyl's Law, which gives the asymptotic scaling of eigenvalues $\lambda_k(\Delta_P) \asymp k^{2/d}$, and allows us to properly control the bias induced by spectral projection.
3. A local version of Weyl's law, which gives an estimate on $\sum_{k=1}^{K} \big(\psi_k(x)\big)^2$, and allows us to appropriately control the difference $\langle f_0, \psi_k \rangle_n - \langle f_0, \psi_k \rangle_P$ between the empirical and population Fourier coefficients.

With these facts in hand the proof of Proposition 6 follows from calculations standard to analysis of the Gaussian sequence model. Our analysis of Laplacian eigenmaps will depend on analogues to the first and second of these facts, with the space $\mathcal{H}^s(\mathcal{X})$ and eigenvalues $\lambda_k(\Delta_P)$ replaced by alternatives suitably defined with respect to the neighborhood graph Laplacian $L_{n,\varepsilon}$.

Goodness-of-fit testing rates over Sobolev balls. In the goodness-of-fit testing problem, we ask for a test function—formally, a Borel measurable function ϕ that takes values in $\{0, 1\}$— which can distinguish between the hypotheses

$$\mathbf{H}_0 : f_0 = f_0^*, \quad \text{versus} \quad \mathbf{H}_a : f_0 \in \mathcal{F} \setminus \{f_0^*\}. \tag{4.12}$$

Typically, the null hypothesis $f_0 = f_0^* \in \mathcal{F}$ reflects the absence of interesting structure, and $\mathcal{F} \setminus \{f_0^*\}$ is a set of smooth departures from this null. To fix ideas, as in Ingster and Sapatinas [2009] we focus on the problem of *signal detection* in Sobolev spaces, where $f_0^* = 0$ and $\mathcal{F} = \mathcal{H}^s(\mathcal{X}; M)$ is a Sobolev ball. This is without loss of generality since our test statistic and its analysis are easily modified to handle the case when f_0^* is not 0, by simply subtracting $f_0^*(X_i)$ from each observation Y_i.

The Type I error of a test ϕ is $\mathbb{E}_0[\phi]$, and if $\mathbb{E}_0[\phi] \leq a$ for a given $a \in (0,1)$ we refer to ϕ as a level-a test[2]. The worst-case risk of ϕ over $\mathcal{F}$ is

$$R_n(\phi, \mathcal{F}, \epsilon) := \sup\Big\{\mathbb{E}_{f_0}[1-\phi] : f_0 \in \mathcal{F}, \|f_0\|_P > \epsilon\Big\},$$

and for a given constant $b \in (0,1)$, the minimax critical radius $\epsilon_n(\mathcal{F})$ is the smallest value of ϵ such that some level-a test has worst-case risk of at most b. Formally,

$$\epsilon_n(\mathcal{F}) := \inf\Big\{\epsilon > 0 : \inf_{\phi} R_n(\phi, \mathcal{F}, \epsilon) \leq b\Big\},$$

where in the above the infimum is over all level-a tests ϕ, and $\mathbb{E}_{f_0}[\cdot]$ is the expectation operator under the regression function f_0.[3] We will refer to the rate at which the squared critical radius $\epsilon_n(\mathcal{F})^2$ tends to 0 as the minimax testing rate.

Testing whether a regression function f_0 is equal to 0 is an easier problem than estimating f_0, and so the minimax testing rate is much smaller than the minimax estimation rate. Ingster and Sapatinas [2009] give the minimax testing rate in the special case where $\mathcal{X} = [0,1]^d$ and P is uniform, and we restate their result in the terms and notation of our paper.

Theorem 11 (Theorem 1 of Ingster and Sapatinas [2009])**.** *Suppose Model 4.2.1, and additionally that $\mathcal{X} = [0,1]^d$, P is the uniform distribution on $[0,1]^d$, and $f_0 \in H_0^s(\mathcal{X})$. Then*

$$\epsilon_n^2\big(H_0^s(P;1)\big) \asymp n^{-4s/(4s+d)} \quad \textit{for } 1 \leq d < 4s. \tag{4.13}$$

The analysis used by Ingster and Sapatinas [2009] to show the upper bound in (4.13) relies on a similar trio of facts as used in the proof of Proposition 6. It is otherwise reminiscent of calculations made in the Gaussian sequence model, which can be found in Ingster and Suslina [2012]. This analysis can be straightforwardly adapted to handle design distributions P that satisfy the conditions of Model 4.2.1, or to handle the case where M is not 1. Finally, the test Ingster and Sapatinas [2009] use to certify the upper bound in (4.13) is implicitly a spectral projection method: the test statistic is the $L^2(P_n)$ norm of a particular spectral projection estimator.[4]

A major difference between testing and estimation over Sobolev spaces is the requirement that $4s > d$. When $4s \leq d$, the functions in $H^s(\mathcal{X})$ are very irregular. Crucially, $H^s(\mathcal{X})$ no longer continuously embeds into $L^4(\mathcal{X})$ when $4s \leq d$, and test statistics using the $L^2(P_n)$ norm are no longer guaranteed to have finite variance.[5] In fact, we have not seen any analysis which describes the minimax rate for nonparametric regression testing over Sobolev spaces in the $4s \leq d$ regime. However, if one explicitly assumes that $f_0 \in L^4(\mathcal{X})$, then the critical radius is characterized by the dimension-free rate $\epsilon_n^2(L^4(\mathcal{X}; M)) \asymp Mn^{-1/2}$.[6] As we discuss after our first main theorem regarding testing with Laplacian eigenmaps (Theorem 13), this rate is achievable by a test based on $\widehat{T}$.

[2] We reserve the more common symbol α for multi-indices, so as to avoid confusion

[3] Clearly, the minimax critical radius $\epsilon_n(\mathcal{F})$ depends on a and b. However, we adopt the typical convention of treating $\alpha, b \in (0,1)$ and as small but fixed positive constants; hence they will not affect the testing error rates, and we suppress them notationally.

[4] Ingster and Sapatinas [2009] project the responses onto a the span of trigonometric basis functions. Since P is uniform on the unit cube, such functions are eigenfunctions of Δ_P when the right boundary conditions are imposed.

[5] Note that this will not affect the analysis for estimation, because for estimation we only need to control the first two moments of f_0.

[6] As a sanity check note that this is strictly worse than the rate in (4.13). In other words, whenever $4s > d$, the embedding $H^s(\mathcal{X}) \subseteq L^4(\mathcal{X})$ never yields a tight upper bound.

Manifold setup. Under Model 4.2.2, both (4.10) and (4.13) continue to hold, but with the ambient dimension d replaced everywhere by the intrinsic dimension m. The estimator $\widetilde{f}$ and a test using the statistic $\widetilde{T}$—with $\psi_1, \psi_2, \ldots$ now the eigenfunctions of the manifold weighted Laplace-Beltrami operator Δ_P—achieve the optimal estimation and testing rates. This is because each of the three facts mentioned after Proposition 6 have analogues when the domain is a smooth manifold (See Hendriks [1990], who analyzes a spectral projection density estimator, for details.)

In-sample mean squared error. As mentioned in our introduction, roughly speaking one of our main conclusions is that the Laplacian eigenmaps estimator $\widehat{f}$ is minimax rate-optimal. It is worth being clear about what we do and do not mean by this statement. We do not mean that the estimator $\widehat{f}$ will match the upper bound given in (4.10), since such a statement does not make sense when the estimator is defined only at the random design points $X_1, \ldots, X_n$. Instead we will measure loss using the squared $L^2(P_n)$ error. In Section 4.5 we show that an extension of $\widehat{f}$ defined over all $\mathcal{X}$ has $L^2(P)$ error comparable to the $L^2(P_n)$ error of $\widehat{f}$. We also believe that in the random design setting we work in, simple arguments will imply that $L^2(P_n)$ risk has the same minimax rate of convergence as $L^2(P)$ risk. We sketch such an argument in Section 4.5, but do not further pursue the details.

Additionally, we will not actually measure accuracy using the expectation of the loss. Rather, we will give a high-probability bound on $\|\cdot\|_n^2$. For instance, when $f_0 \in H^1(\mathcal{X}; 1)$, we will show that with probability $1-\delta$ the loss $\|\widehat{f} - f_0\|_n^2 \leq C_\delta n^{-2/(2+d)}$, for a constant C_δ that depends on δ but not on f_0 or n. Thus we give an upper bound on the $(1-\delta)$th quantile of $\|\widehat{f} - f_0\|_n^2$, rather than an upper bound on its expectation. We explain the reason for this in Section 4.3. We also show that if f_0 is bounded in a larger norm—for instance, if it is Hölder rather than Sobolev smooth—then we can obtain bounds on the expected $L^2(P_n)$ loss.

There is one other subtlety introduced by the use of in-sample mean squared error. Technically speaking, elements $f \in H^s(\mathcal{X})$ are equivalence classes, defined only up to a set of measure zero. Thus one cannot speak of the pointwise evaluation $f_0(X_i)$, as we do by defining our target of estimation to be $f_0(X_i)$, $i = 1, \ldots, n$, until one selects *representatives*. When $s > d/2$, every element f of $H^s(\mathcal{X})$ admits a continuous version f^*, and as is standard we set this to be our favored representative. When $s \leq d/2$, some elements in $H^s(\mathcal{X})$ do not have any continuous version; however they admit a *quasi-continuous* version [Evans and Gariepy, 2015] known as the *precise representative*, and we use this representative. To be clear, however, it does not really matter which representative we choose. Since all versions agree except on a set of measure zero, and since P is absolutely continuous with respect to Lebesgue measure (in Model 4.2.1) or the volume form $d\mu$ (in Model 4.2.2), with probability 1 any two versions $g_0, h_0 \in f_0$ will satisfy $g_0(X_i) = h_0(X_i)$ for all $i = 1, \ldots, n$. The bottom line is that we can use the notation $f_0(X_i)$ without fear of ambiguity or confusion.

Finally, we note that for testing none of these comments are relevant. We will show that our test has small worst case risk whenever $\epsilon \gtrsim \epsilon_n(H_0^s(\mathcal{X}; M))$, thus establishing that it is a minimax optimal test in the usual sense.

4.3 Minimax Optimality of Laplacian Eigenmaps

As previously explained, Laplacian eigenmaps is a discrete and noisy approximation to a spectral projection method using the eigenfunctions of Δ_P. This is particularly useful when P is unknown, or when the eigenfunctions of Δ_P cannot be explicitly computed. Our goal is to show that Laplacian eigenmaps methods are rate-optimal, notwithstanding the potential extra error incurred by this approximation. In this section and the following one, we will see that this is indeed the case: the estimator $\widehat{f}$, and a test using the statistic $\widehat{T}$, achieve optimal estimation and goodness-of-fit testing rates over Sobolev classes.

In this section we will cover the flat Euclidean case, where we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to Model 4.2.1. We will divide our theorem statements based on whether we assume the regression function

f_0 belongs to the first order Sobolev class ($s = 1$) or a higher-order Sobolev class ($s > 1$), since the details of the two settings are somewhat different.

4.3.1 First-order Sobolev classes

We begin by assuming $f_0 \in H^1(\mathcal{X})$. We show that $\widehat{f}$ and a test based on $\widehat{T}$ are minimax optimal, for all values of d, and under no additional assumptions (beyond those of Model 4.2.1) on the data generating process, i.e. on either P or f_0.

Estimation. When the kernel η, graph radius ε, and number of eigenvectors K are chosen appropriately, we show in Theorem 12 that the estimator $\widehat{f}$ achieves the minimax rate over $H^1(\mathcal{X}; M)$, when error is measured in squared $L^2(P_n)$ norm.

(K1) The kernel function η is a nonincreasing function supported on $[0, 1]$. Its restriction to $[0, 1]$ is Lipschitz, and $\eta(1) > 0$. Additionally, it is normalized so that

$$\int_{\mathbb{R}^d} \eta(\|z\|)\, dz = 1.$$

and we assume $\sigma_\eta := \frac{1}{d}\int_{\mathbb{R}^d} \|x\|^2 \eta(\|x\|)\, dx < \infty$.

(P1) For constants c_0 and C_0, the graph radius ε and the number of eigenvectors K satisfy the following inequalities:

$$C_0\left(\frac{\log n}{n}\right)^{1/d} \le \varepsilon \le c_0 \min\{1, K^{-1/d}\}, \tag{4.14}$$

and

$$K = \min\Big\{\Big\lfloor (M^2 n)^{d/(2+d)} \Big\rfloor \vee 1, n\Big\}. \tag{4.15}$$

We comment on these assumptions after stating our first main theorem, regarding the estimation error of Laplacian eigenmaps.

Theorem 12. *Suppose Model 4.2.1, and additionally $f_0 \in H^1(\mathcal{X}, M)$. There are constants c, C and N (not depending on f_0, M or n), such that the following statement holds for all $n \ge N$ and any $\delta \in (0, 1)$: if the Laplacian eigenmaps estimator $\widehat{f}$ is computed with kernel η satisfying (K1), and parameters ε and K satisfying (P1), then*

$$\|\widehat{f} - f_0\|_n^2 \le C\Big(\frac{1}{\delta} M^2 (M^2 n)^{-2/(2+d)} \wedge 1\Big) \vee \frac{1}{n}, \tag{4.16}$$

with probability at least $1 - \delta - Cn\exp(-cn\varepsilon^d) - \exp(-K)$.

From (4.16) it follows immediately that with high probability $\|\widehat{f} - f_0\|_n^2 \lesssim M^2(M^2 n)^{-2/(2+d)}$ whenever $n^{-1/2} \lesssim M \lesssim n^{1/d}$.

Some other remarks:

- When $M = o(n^{-1/2})$, then computing Laplacian eigenmaps with $K = 1$ achieves the parametric rate $\|\widehat{f} - f_0\|_n^2 \lesssim n^{-1}$, and the zero-estimator $\widehat{f} = 0$ achieves the better rate $\|\widehat{f} - f_0\|_n^2 \lesssim M^2$. However, we do not know what the minimax rate is in this regime. On the other hand, when $M = \omega(n^{1/d})$, then computing Laplacian eigenmaps with $K = n$ achieves the rate $\|\widehat{f} - f_0\|_n^2 \lesssim 1$, which is better than the rate in (4.10). This is because we are evaluating error in $L^2(P_n)$ rather than $L^2(P)$. However, in truth these are edge cases, which do not fall neatly into the framework of nonparametric regression.

- The assumptions placed on the kernel function η are needed for technical reasons. They can likely be weakened, although we note that they are already fairly general. The lower bound on ε imposed by (4.14) is on the order of the connectivity threshold, the smallest length scale at which the resulting graph will still be connected with high probability. On the other hand, as we will see in Section 4.3.3, the upper bound on ε is needed to ensure that the graph eigenvalue λ_K is of at least the same order as the continuum eigenvalue $\lambda_K(\Delta_P)$. Finally, we choose $K = (M^2 n)^{2d/(2+d)}$ (when possible) to optimally trade-off bias and variance.
- We note that the ranges (4.14) and (4.15) depend on quantities, such as the dimension d and radius of the Sobolev ball M, which are usually unknown. In practice, one typically tunes hyper-parameters by sample-splitting or cross-validation. However, because the estimator $\widehat{f}$ is defined only in-sample, we cannot use such methods to select the graph radius ε, or number of eigenvectors K. We return to this issue in Section 4.5, when we propose an out-of-sample extension of $\widehat{f}$.
- The upper bound given in equation (4.16) holds with probability $1 - \delta - Cn\exp(-cn\varepsilon^d)$. Under the stronger assumption that $f_0 \in C^1(\mathcal{X}; M)$ we can replace the factor of δ by the sharper δ^2/n, which is less than δ because $\delta \in (0, 1)$. Then a routine calculation shows that the expected $L^2(P_n)$ loss is on the same order as (4.16), matching the minimax rate.

Testing. Consider the test $\varphi = \mathbf{1}\{\widehat{T} \geq t_a\}$, where t_a is the threshold

$$t_a := \frac{K}{n} + \frac{1}{n}\sqrt{\frac{2K}{a}}.$$

This choice of threshold t_a guarantees that φ is a level-a test. Moreover, when ε and K are chosen appropriately, the test φ has negligible Type II error against alternatives separated from the null by at least $\|f_0\|_P^2 \gtrsim M^2(M^2 n)^{-4/(4+d)}$, whenever $d < 4$.

(P2) The graph radius ε and the number of eigenvectors K satisfy (4.14). Additionally,

$$K = \min\left\{\left\lfloor (M^2 n)^{2d/(4+d)} \right\rfloor \vee 1, n\right\}. \tag{4.17}$$

Theorem 13. *Fix $a, b \in (0, 1)$. Suppose Model 4.2.1. Then $\mathbb{E}_0[\varphi] \leq a$, i.e φ is a level-a test. Suppose additionally $f_0 \in H^1(\mathcal{X}, M)$, and that $d < 4$. Then there exist constants C and N that do not depend on f_0, such that the following statement holds for all n larger than N: if the Laplacian eigenmaps test φ is computed with kernel η satisfying (K1), and parameters ε and K satisfying (P2), and if f_0 satisfies*

$$\|f_0\|_P^2 \geq C\left(\left(M^2(M^2 n)^{-4/(4+d)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{bn^{2/d}}\right) \vee \frac{1}{n}, \tag{4.18}$$

then $\mathbb{E}_{f_0}[1 - \phi] \leq b$.

Although (4.18) involves taking the maximum of several different terms, the important takeaway of Theorem 13 is that if $n^{-1/2} \lesssim M \lesssim n^{1/d}$, then φ has small worst-case risk as long as f_0 is separated from 0 by at least $M^2(M^2 n)^{-4/(4+d)}$. Note that unlike in the estimation setting—where we measured loss in $L^2(P_n)$ error—the separation in (4.18) is measured in $L^2(P)$ norm. Thus (4.18) implies that φ is a minimax rate-optimal test over $H^1(\mathcal{X}; M)$, in the usual sense.

Some other remarks:

- As mentioned previously, when $d \geq 4$ the first order Sobolev space $H^1(\mathcal{X})$ does not continuously embed into $L^4(\mathcal{X})$, and we do not know the optimal rates for regression testing over $H^1(\mathcal{X}, M)$. On the other hand, if we explicitly assume $f_0 \in L^4(\mathcal{X})$, then the Laplacian eigenmaps test with $K = n$, has small type II error whenever $\|f_0\|_P^2 \gtrsim Mn^{-1/2}$. Note that when $K = n$ the Laplacian eigenmaps test statistic is nothing but the squared $L^2(P_n)$ norm of $\mathbf{Y}$, $\widehat{T} = \|\mathbf{Y}\|_n^2$. See Green et al. [2021] for details.

- As in the estimation setting, the range of Sobolev ball radii $n^{-1/2} \lesssim M \lesssim n^{1/d}$ for which Theorem 13 implies that φ is a rate-optimal test covers all those cases for which the critical radius $\epsilon(H^1(\mathcal{X}; M))$ is both $\Omega(1/n)$ and $O(1)$.

4.3.2 Higher-order Sobolev classes

We now consider the situation where the regression function displays some higher-order regularity. For reasons already discussed, we also assume the zero-trace boundary condition, i.e. $f_0 \in H_0^s(\mathcal{X})$. We show that Laplacian eigenmaps methods continue to be optimal for all orders of s, as long as the design density is itself also sufficiently regular, $p \in C^{s-1}(\mathcal{X})$. In estimation, this is the case for any dimension d, whereas in testing it is the case only when $d \leq 4$.

Estimation. In order to show that $\widehat{f}$ is an optimal estimator over $H_0^s(\mathcal{X}; M)$, we will require that ε be meaningfully larger than the lower bound in (P1).

(P3) For constants c_0 and C_0, the graph radius ε and number of eigenvectors K satisfy

$$C_0 \max\left\{\left(\frac{\log}{n}\right)^{1/d}, (M^2 n)^{-1/(2(s-1)+d)}\right\} \leq \varepsilon \leq c_0 \min\{1, K^{-1/d}\} \tag{4.19}$$

and

$$K = \min\left\{\left\lfloor (M^2 n)^{d/(2s+d)} \right\rfloor \vee 1, n\right\}$$

Crucially, when n is sufficiently large the two conditions in (P3) are guaranteed to not be mutually exclusive. This is because so long as $M^2 = \omega(n^{-1})$ then $(M^2 n)^{-2/(2(s-1)+d)} = o((M^2 n)^{-2/(2s+d)})$, regardless of s and d.

Theorem 14. *Suppose Model 4.2.1, and additionally $f_0 \in H_0^s(\mathcal{X}, M)$ and $p \in C^{s-1}(\mathcal{X})$. There exist constants c, C and N that do not depend on f_0, such that the following statement holds all for all n larger than N and for any $\delta \in (0,1)$: if the Laplacian eigenmaps estimator $\widehat{f}$ is computed with kernel η satisfying (K1), and parameters ε and K satisfying (P3), then*

$$\|\widehat{f} - f_0\|_n^2 \leq C\left(\frac{1}{\delta} M^2 (M^2 n)^{-2s/(2s+d)} \wedge 1\right) \vee \frac{1}{n}, \tag{4.20}$$

with probability at least $1 - \delta - Cn\exp(-cn\varepsilon^d) - \exp(-K)$.

Theorem 14, in combination with Theorem 12, implies that in the flat Euclidean setting Laplacian eigenmaps is an in-sample minimax rate-optimal estimator over Sobolev classes, for all values of s and d. Some other remarks:

- We do not require that the regularity of the Sobolev space satisfy $s > d/2$, a condition often seen in the literature. In the sub-critical regime $s \leq d/2$, the Sobolev space $H^s(\mathcal{X})$ is quite irregular. It is not a Reproducing Kernel Hilbert Space (RKHS), nor does it continuously embed into $C^0(\mathcal{X})$, much less into any Hölder space. As a result, for certain versions of the nonparametric regression problem—e.g. when loss is measured in L^∞ norm, or when the design points $\{X_1, \ldots, X_n\}$ are assumed to be fixed—in a minimax sense even consistent estimation is not possible. Likewise, certain estimators are "off the table", most notably RKHS-based methods such as thin-plate splines of degree $k \leq d/2$. Nevertheless, for random design regression with error measured in $L^2(P)$-norm, the spectral projection estimator $\widetilde{f}$ defined in (4.11) obtains the "usual" minimax rates $n^{-2s/(2s+d)}$ for all values of s and d. Theorems 12 and 14 show that the same is true with respect to Laplacian eigenmaps, with error measured in $L^2(P_n)$-norm.
- The requirement $p \in C^{s-1}(\mathcal{X})$ is a strong condition, but is essential to showing that $\widehat{f}$ enjoys faster rates of convergence when $s > 1$. We explain why in Section 4.3.3, where we discuss our analysis.

Testing. The test φ can adapt to the higher-order smoothness of f_0, when ε and K are chosen correctly.

(P4) The graph radius ε and the number of eigenvectors K satisfy (4.19). Additionally,

$$K = \min\left\{\left\lfloor (M^2 n)^{2d/(4+d)} \right\rfloor \vee 1, n\right\}. \tag{4.21}$$

When $d \leq 4$, for any value of $s \in \mathbb{N}$ when n is sufficiently large it is possible to choose ε and K such that both (4.19) and (4.21) are satisfied, and our next theorem establishes that in this situation φ is an optimal test.

Theorem 15. *Fix $a, b \in (0,1)$. Suppose Model 4.2.1. Then $\mathbb{E}_0[\varphi] \leq a$, i.e φ is a level-a test. Suppose additionally $f_0 \in H_0^s(\mathcal{X}, M)$, that $p \in C^{s-1}(\mathcal{X})$, and that $d \leq 4$. Then there exist constants c, C and N that do not depend on f_0, such that the following statement holds for all $n \geq N$: if the Laplacian eigenmaps test φ is computed with kernel η satisfying (K1), and parameters ε and K satisfying (P4), and if f_0 satisfies*

$$\|f_0\|_P^2 \geq \frac{C}{b}\left(\left(M^2(M^2 n)^{-4s/(4s+d)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{b n^{2s/d}}\right) \vee \frac{1}{n}, \tag{4.22}$$

then $\mathbb{E}_{f_0}[1 - \phi] \leq b$.

Similarly to the first-order case, the main takeaway from Theorem 15 is that φ is a minimax optimal test over $H_0^s(\mathcal{X})$ when $n^{-1/2} \lesssim M^2 \lesssim n^{1/d}$. However, unlike the first-order case, when $4 < d < 4s$ the minimax testing rate over $H_0^s(\mathcal{X})$ is still on the order of $M^2(M^2 n)^{-4s/(4s+d)}$. Unfortunately, we can no longer claim that φ is an optimal test in this regime.

Theorem 16. *Under the same setup as Theorem 14, but with $4 < d < 4s$. If the Laplacian eigenmaps test φ is computed with kernel η satisfying (K1), number of eigenvectors K satisfying (4.21), and $\varepsilon = (M^2 n)^{-1/(2(s-1)+d)}$, and if*

$$\|f_0\|_P^2 \geq \frac{C}{b}\left(\left(M^2(M^2 n)^{-2s/(2(s-1)+d)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{b n^{2s/d}}\right) \vee \frac{1}{n}, \tag{4.23}$$

then $\mathbb{E}_{f_0}[1 - \phi] \leq b$.

Note that as a consequence of Theorem 14, if we choose $K = n^{-2s/(2s+d)}$ then φ must have small Type II error whenever $\|f_0\|_P^2 \gtrsim n^{-2s/(2s+d)}$. Theorem 16 shows that φ can achieve better, but still not optimal, rates. As a technical matter, the problem is that when $d > 4$ there do not exist any choices of ε and K which satisfy both (4.19) and (4.21), and as a result we cannot optimally balance (our upper bound on) testing bias and variance (defined momentarily in (4.25)). Although we suspect φ is truly suboptimal when $d > 4$, technically speaking (4.25) gives only an upper bound on testing bias, and thus we cannot rule out that the test φ is optimal for all $4 < d < 4s$. We leave the matter to future work.

That being said, it is somewhat remarkable that Laplacian eigenmaps *can* take advantage of higher-order smoothness, and especially surprising that it can do so in an optimal manner. The sharpest known results Cheng and Wu [2021] show that the graph Laplacian eigenvectors v_k converge to eigenfunctions ψ_k at a rate of $n^{-1/(4+d)}$. Naively applying these results, one can show that $\widehat{f}$ to $\widetilde{f}$, but only at a rate far slower than the optimal rates for regression. Of course when the index K increases with n, as is necessary to optimally balance bias and variance, the issue only gets worse. Clearly, as a method for regression, the rate of convergence of Laplacian eigenmaps is much better than the rate implied by (what is currently known about) the concentration of individual eigenvectors around their continuum limits.

4.3.3 Analysis

We now outline the high-level strategy we follow when proving each of Theorems 12-16. We analyze the estimation error of $\widehat{f}$, and the testing error of $\widehat{\varphi}$, by first conditioning on the design points $X_1, \dots, X_n$ and deriving *design-dependent* bias and variance terms. For estimation, we have that with probability at least $1 - \exp(-K)$,

$$\|\widehat{f} - f_0\|_n^2 \le \underbrace{\frac{\langle L^s f_0, f_0 \rangle_n}{\lambda_K^s}}_{\text{bias}} + \underbrace{\frac{5K}{n}}_{\text{variance}} . \tag{4.24}$$

For testing, we have that φ (which is a level-a test by construction) also has small Type II Error, $\mathbb{E}_{f_0}[1-\phi] \le b/2$, if

$$\|f_0\|_n^2 \ge \underbrace{\frac{\langle L^s f_0, f_0 \rangle_n}{\lambda_K^s}}_{\text{bias}} + \underbrace{32\frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]}_{\text{variance}} . \tag{4.25}$$

The quadratic form $\langle L_{n,\varepsilon}^s f_0, f_0 \rangle_n$, eigenvalue λ_K, and empirical squared norm $\|f_0\|_n^2$ are each random variables that depend the random design points $X_1, \dots, X_n$. We proceed to establish suitable upper and lower bounds on these quantities.

Estimates on graph quadratic forms. In Proposition 7 we restate an upper bound on the Dirichlet energy $\langle L_{n,\varepsilon} f, f \rangle_n$ from Green et al. [2021].

Proposition 7 (Lemma 1 of Green et al. [2021])**.** *Suppose Model 4.2.1, and additionally $f \in H^1(\mathcal{X})$. There exist constants c, C that do not depend on f or n such that the following statement holds for any $\delta \in (0,1)$: if η satisfies (K3) and $\varepsilon < c$, then*

$$\langle L_{n,\varepsilon} f, f \rangle_n \le \frac{C}{\delta} \|f\|_{H^1(\mathcal{X})}^2, \tag{4.26}$$

with probability at least $1 - \delta$.

Proposition 7 follows by upper bounding the expectation of $\langle L_{n,\varepsilon} f, f \rangle_n$, which is the Dirichlet energy $E_{P,\varepsilon}(f) := \langle L_{P,\varepsilon} f, f \rangle_P$, by (a constant times) the squared Sobolev norm $\|f\|_{H^1(\mathcal{X})}^2$.

In this work, we establish that an analogous bound holds for $\langle L_{n,\varepsilon}^s f_0, f_0 \rangle_n$ when $s > 1$. We call this quantity the order-s *graph Sobolev semi-norm*.

Proposition 8. *Suppose Model 4.2.1, and additionally that $f \in H_0^s(\mathcal{X})$ and $p \in C^{s-1}(\mathcal{X})$. Then there exist constants c and C that do not depend on f_0 or n such that the following statement holds for any $\delta \in (0,1)$: if η satisfies (K1) and $Cn^{-1/(2(s-1)+d)} < \varepsilon < c$, then*

$$\langle L_{n,\varepsilon}^s f, f \rangle_n \le \frac{C}{\delta} \|f\|_{H^s(\mathcal{X})}^2, \tag{4.27}$$

with probability at least $1 - \delta$.

We now summarize the techniques used to prove Proposition 8, which will help explain what role the conditions on f_0,p and ε play. To upper bound $\langle L_{n,\varepsilon}^s f, f \rangle_n$ in terms of $\|f\|_{H^s(\mathcal{X})}^2$, we introduce an intermediate quantity: the *order-s non-local Sobolev seminorm* $\langle L_{P,\varepsilon}^s f, f \rangle_P$. This seminorm is defined with respect to $L_{P,\varepsilon}$, which is a non-local approximation to Δ_P,

$$L_{P,\varepsilon} f(x) := \frac{1}{\varepsilon^{d+2}} \int_{\mathcal{X}} \big(f(z) - f(x)\big) \eta\left(\frac{\|z - x\|}{\varepsilon}\right) dP(x). \tag{4.28}$$

Then the proof of Proposition 8 proceeds according to the following steps.

- First we note that $\langle L^s_{n,\varepsilon} f, f\rangle_n$ is itself a biased estimate of the non-local seminorm $\langle L^s_{P,\varepsilon} f, f\rangle_P$. Specifically, $\langle L^s_{n,\varepsilon} f, f\rangle_n$ is a V-statistic, meaning it is the sum of an unbiased estimator of $\langle L^s_{P,\varepsilon} f, f\rangle_P$ (in other words, a U-statistic) plus some higher-order, pure bias terms. We show that these pure bias terms are negligible when $\varepsilon = \omega(n^{-1/(2(s-1)+d)})$.
- For x sufficiently in the interior of $\mathcal{X}$, we show that $L^j_{P,\varepsilon} f(x) \to \sigma_\eta^j \Delta^j_P f(x)$ as $\varepsilon \to 0$. Here $j = (s-1)/2$ when s is odd and $j = (s-2)/2$ when s is even. This step bears some resemblance to the analysis of the bias term in kernel smoothing, and requires that $p \in C^{s-1}(\mathcal{X})$.
- On the other hand for x sufficiently near the exterior of $\mathcal{X}$, $L^j_{P,\varepsilon} f(x)$ does not converge to $\Delta^j_P f(x)$. Instead, we use the zero-trace property of f to show that $L^j_{P,\varepsilon} f(x)$ is small.
- Finally, we combine the results of previous two steps to deduce an upper bound on $\langle L^s_{P,\varepsilon} f, f\rangle_P$ in terms of the squared Sobolev norm $\|f\|^2_{H^s(\mathcal{X})}$. Roughly speaking, when s is odd, $\langle L^s_{P,\varepsilon} f, f\rangle_P = E_{P,\varepsilon}(L^j_{P,\varepsilon} f) \approx \sigma_\eta^{2j} E_{P,\varepsilon}(\Delta^j_P f)$, whereas when s is even $\langle L^s_{P,\varepsilon} f, f\rangle_P = \|L_{P,\varepsilon} L^j_{P,\varepsilon} f\|^2_P \approx \sigma_\eta^{2j} \|L_{P,\varepsilon} \Delta_P f\|^2_P$. Reasoning in this way, we can translate estimates of $L^j_{P,\varepsilon} f$ into an upper bound on the order-s non-local Sobolev seminorm, even though $s > j$.

Together, these steps establish Proposition 8. It is worth pointing out that we do not try to show $L^s_{P,\varepsilon} f(x) \to \Delta^s_P f(x)$. This may seem like a natural first step towards a simple proof that $\langle L^s_{P,\varepsilon} f, f\rangle_P \to \langle \Delta^s_P f, f\rangle_P$. The problem is that Δ^s_P is an order-$2s$ differential operator, whereas we assume that f has only s bounded derivatives. Instead we go for the slightly more complicated approach outlined above.

Neighborhood graph eigenvalue. On the other hand, several recent works [Burago et al., 2014, García Trillos and Slepčev, 2018a, Calder and García Trillos, 2019] have analyzed the convergence of λ_k towards $\lambda_k(\Delta_P)$. They provide explicit bounds on the relative error $|\lambda_k - \lambda_k(\Delta_P)|/\lambda_k(\Delta_P)$, which show that the relative error is small for sufficiently large n and small ε. Crucially, the guarantees hold simultaneously for all $1 \le k \le K$ as long as $\lambda_K(\Delta_P) = O(\varepsilon^{-2})$. These results are actually stronger than are necessary to establish Theorems 12-15—in order to get rate-optimality, we need only show that for the relevant values of K, $\lambda_K/\lambda_K(P) = \Omega_P(1)$—but unfortunately they all assume P is supported on a manifold without boundary (i.e. they assume Model 4.2.2 rather than Model 4.2.1).

In the case where $\mathcal{X}$ is assumed to have a boundary, the graph Laplacian $L_{n,\varepsilon}$ is a reasonable approximation of the operator Δ_P at points $x \in \mathcal{X}$ for which $B(x,\varepsilon) \subseteq \mathcal{X}$. In contrast, at points x near the boundary of $\mathcal{X}$, the graph Laplacian is known to approximate a different operator altogether [Belkin et al., 2012]. This is reminiscent of the boundary effects present in the analysis of kernel smoothing. Thus proving convergence of λ_k to a continuum limit becomes a substantially more challenging problem when $\mathcal{X}$ has a boundary. Rather than establishing such a result, we will instead directly use Lemma 2 of Green et al. [2021], whose assumptions match our own, and who give a weaker bound on $\lambda_k/\lambda_k(\Delta_P)$ that will nevertheless suffice for our purposes.

Proposition 9 (Lemma 2 of Green et al. [2021])**.** *Suppose Model 4.2.1. Then there exist constants c and C such that the following statement holds: if η satisfies (K1) and $C(\log n/n)^{1/d} < \varepsilon < c$, then*

$$\lambda_k \ge c \cdot \min\left\{\lambda_k(\Delta_P), \frac{1}{\varepsilon^2}\right\} \quad \text{for all } 1 \le k \le n, \tag{4.29}$$

with probability at least $1 - Cn\exp\{-cn\varepsilon^d\}$.

Note immediately that $\lambda_0(\Delta_P) = \lambda_0 = 0$. Furthermore, Weyl's Law (Lemma 44) tells us that under Model 4.2.1, $k^{2/d} \lesssim \lambda_k(\Delta_P) \lesssim k^{2/d}$ for all $k \in \mathbb{N}, k > 1$. Combining these statements with (4.29), we conclude that $\lambda_K = \Omega_P(K^{2/d})$ so long as $K \lesssim \varepsilon^{-d}$.

Empirical norm. Finally, in order to show that φ has small Type II error whenever $\|f_0\|_P$ is greater than the critical radius given by (4.13), we require a lower bound on $\|f_0\|_n^2$ in terms of $\|f_0\|_P^2$. In Proposition 10 we establish that such a one-sided bound holds, whenever $\|f_0\|_P$ is sufficiently large.

Proposition 10. *Suppose Model 4.2.1, and additionally that $f \in H^s(\mathcal{X}, M)$ for some $s > d/4$. There exist constants c and C that do not depend on f_0 or n such that the following statement holds for any $\delta > 0$: if*

$$\|f\|_P \geq CM\left(\frac{1}{\delta n}\right)^{s/d} \tag{4.30}$$

then with probability at least $1 - \exp\{-(cn \wedge 1/\delta)\}$.

$$\|f\|_n^2 \geq \frac{1}{2}\|f_0\|_P^2. \tag{4.31}$$

To prove Proposition 10, we use the Gagliardo-Nirenberg interpolation inequality to control the 4th moment of f in terms of $\|f\|_P$ and $|f|_{H^s(\mathcal{X})}$, then invoke a one-sided Bernstein's inequality as in [Wainwright, 2019, Section 14.2]. Note carefully that the statement (4.31) is *not* a uniform guarantee over all $f \in H^s(\mathcal{X}; M)$, as such a statement cannot hold in the sub-critical regime ($2s \leq d$). Fortunately, a pointwise bound—meaning a bound that holds with high probability for a single $f \in H^s(\mathcal{X})$—is sufficient for our purposes.

Finally, invoking the bounds of Propositions 7-10 inside the bias-variance tradeoffs (4.24) and (4.25) and then choosing K to balance bias and variance (when possible), leads to the conclusions of Theorems 12-16.

4.3.4 Computational considerations

Recall that when $s = 1$, we have shown that Laplacian eigenmaps is optimal when $\varepsilon \asymp (\log n/n)^{1/d}$ is (up to a constant) as small as possible while still ensuring the graph G is connected. On the other hand, when $s > 1$, we can show Laplacian eigenmaps is optimal only when $\varepsilon = \omega(n^{-c})$ for some $c < 1/d$. For such a choice of ε, the average degree in G will grow polynomially in n as $n \to \infty$, and computing eigenvectors of the Laplacian of a graph will be more computationally intensive than if the graph were sparse. Thus Theorems 12 and 14 can be seen as revealing a tradeoff between statistical and computational efficiency; although to be clear, we have no theoretical evidence that Laplacian eigenmaps *fails* to adapt to higher-order smoothness when $\varepsilon \asymp (\log n/n)^{1/d}$—we simply cannot prove that it succeeds.

Suppose one does choose ε meaningfully larger than the connectivity threshold, as our theory requires when $s > 1$. We now discuss a procedure to efficiently compute an approximation to the Laplacian eigenmaps estimate, without changing the rate of convergence of the resulting estimator: *edge sparsification*. By now there exist various methods see (e.g., the seminal papers of Spielman and Teng [2011, 2013, 2014], or the overview by Vishnoi [2012] and references therein) to efficiently remove many edges from the graph G while only slightly perturbing the spectrum of the Laplacian. Specifically such algorithms take as input a parameter $\sigma \geq 1$, and return a sparser graph $\widetilde{G}$, $E(\widetilde{G}) \subseteq E(G)$, with a Laplacian $\widetilde{L}_{n,\varepsilon}$ satisfying

$$\frac{1}{\sigma} \cdot u^\top \widetilde{L}_{n,\varepsilon} u \leq u^\top L_{n,\varepsilon} u \leq \sigma \cdot u^\top \widetilde{L}_{n,\varepsilon} u \quad \text{for all } u \in \mathbb{R}^n.$$

Let $\widetilde{f}$ be the Laplacian eigenmaps estimator computed using the eigenvectors of the sparsified graph Laplacian $\widetilde{L}_{n,\varepsilon}$. Because $\widetilde{G}$ is sparser than G, it can be (much) faster to compute the eigenvectors of $\widetilde{L}_{n,\varepsilon}$ than the eigenvectors of $L_{n,\varepsilon}$, and consequently much faster to compute $\widetilde{f}$ than $\widehat{f}$. Statistically speaking, letting $\widetilde{\lambda}_k$ be the kth eigenvalue of $\widetilde{L}_{n,\varepsilon}$, we have that conditional on $X_1, \ldots, X_n$,

$$\|\widetilde{f} - f_0\|_n^2 \leq \frac{\langle \widetilde{L}_{n,\varepsilon}^s f_0, f_0\rangle_n}{\widetilde{\lambda}_{K+1}^s} + \frac{5K}{n} \leq \sigma^{2s} \frac{\langle \widetilde{L}_{n,\varepsilon}^s f_0, f_0\rangle_n}{\widetilde{\lambda}_{K+1}^s} + \frac{5K}{n},$$

with probability at least $1-\exp(-K)$. Consequently $\widetilde{f}$ has $L^2(P_n)$-error of at most σ^{2s} times our upper bound on the error of $\widehat{f}$, and for any choice of σ that is constant in n the estimator $\widetilde{f}$ will also be rate-optimal.

In fact the aforementioned edge sparsification algorithms are overkill for our needs. For one thing, they are designed to work when σ is much larger than 1, whereas in order for $\widetilde{f}$ to be rate-optimal setting σ to be any constant greater than 1, say $\sigma = 2$, is sufficient. Additionally, edge sparsification algorithms are traditionally designed to work in the worst-case, where no assumptions are made on the structure of the graph G. But the geometric graphs we consider in this paper exhibit a special structure, in which very roughly speaking no single edge is a bottleneck. As pointed out by Sadhanala et al. [2016b], in this special case there exist far simpler and faster methods for sparsification, which at least empirically seem to do the job.

4.4 Manifold Adaptivity

In this section we consider the manifold setting, where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed according to Model 4.2.2. A theory has been developed [Niyogi et al., 2008, Belkin, 2003, Belkin and Niyogi, 2008, Niyogi, 2013, Balakrishnan et al., 2012, 2013b] establishing that the neighborhood graph G can "learn" the manifold $\mathcal{X}$ in various senses, so long as $\mathcal{X}$ is locally linear. We build on this work by showing that when $f_0 \in H^s(\mathcal{X})$ and P is supported on a manifold, Laplacian eigenmaps achieve the sharper minimax estimation and testing rates reviewed in Section 4.2.4.

4.4.1 Laplacian eigenmaps error rates under the manifold hypothesis

Unlike in the flat-Euclidean case, since Model 4.2.2 assumes that $\mathcal{X}$ is boundaryless it is easy to deal with the first-order ($s = 1$) and higher-order ($s > 1$) cases all at once. A more important distinction between the results of this section and those of Section 4.3 is that we will establish Laplacian eigenmaps is optimal only when the regression function $f_0 \in H^s(\mathcal{X}; M)$ for $s \leq 3$. Otherwise, this section will proceed in a similar fashion to Section 4.3.2.

Estimation. To ensure that $\widehat{f}$ is an in-sample minimax rate-optimal estimator, we choose the kernel function η, graph radius ε and number of eigenvectors K as in (P3), except with ambient dimension d replaced by the intrinsic dimension m.

(P5) The kernel function η is a nonincreasing function supported on a subset of $[0, 1]$. Its restriction to $[0, 1]$ is Lipschitz, and $\eta(1/2) > 0$. Additionally, it is normalized so that

$$\int_{\mathbb{R}^m} \eta(\|z\|)\, dz = 1,$$

and we assume $\int_{\mathbb{R}^m} \|x\|^2 \eta(\|x\|)\, dx < \infty$.

(P6) For a constant C_0, the graph radius ε and number of eigenvectors K satisfy

$$C_0 \max\left\{\left(\frac{\log}{n}\right)^{1/m}, n^{-1/(2(s-1)+m)}\right\} \leq \varepsilon \leq \min\{i_0, K^{-1/m}\} \tag{4.32}$$

and

$$K = \min\left\{\left\lfloor (M^2 n)^{m/(2s+m)} \right\rfloor \wedge 1, n\right\}.$$

Theorem 17. *Suppose Model 4.2.2, and additionally $f_0 \in H^s(\mathcal{X}, M)$ and $p \in C^{s-1}(\mathcal{X})$ for $s \leq 3$. There exist constants c, C and N that do not depend on f_0, such that the following statement holds all for all n larger than N and for any $\delta \in (0, 1)$: if the Laplacian eigenmaps estimator $\widehat{f}$ is computed with kernel η satisfying (P5), and parameters ε and K satisfying (P6), then*

$$\|\widehat{f} - f_0\|_n^2 \leq C\Big(\frac{1}{\delta} M^2 (M^2 n)^{-2s/(2s+m)} \wedge 1\Big) \vee \frac{1}{n}, \tag{4.33}$$

with probability at least $1 - \delta - Cn\exp(-cn\varepsilon^m) - \exp(-K)$.

Testing. Likewise, to construct a minimax optimal test using $\widehat{T}$, we choose ε and K as in (P2), except with the ambient dimension d replaced by the intrinsic dimension m.

(P6) The graph radius ε and number of eigenvectors K satisfy (4.32). Additionally, the

$$K = \min\Big\{\Big\lfloor (M^2 n)^{2m/(4s+m)} \Big\rfloor \wedge 1, n\Big\}.$$

Theorem 18. *Fix $a, b \in (0,1)$. Suppose Model 4.2.2. Then $\mathbb{E}_0[\varphi] \leq a$, i.e φ is a level-a test. Suppose additionally $f_0 \in H^s(\mathcal{X}, M)$, that $p \in C^{s-1}(\mathcal{X})$, and that $s \leq 3$ and $m \leq 4$. Then there exist constants c, C and N that do not depend on f_0, such that the following statement holds for all n larger than N: if the Laplacian eigenmaps test φ is computed kernel η satisfying (P5), and parameters ε and K satisfying (P6), and if f_0 satisfies*

$$\|f_0\|_P^2 \geq \frac{C}{b}\left(\left(M^2(M^2 n)^{-4s/(4s+m)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{bn^{2s/d}}\right) \vee \frac{1}{n}, \tag{4.34}$$

then $\mathbb{E}_{f_0}[1 - \phi] \leq b$.

- The proofs of Theorems 17 and 18 follow very similarly to the full-dimensional setting. The difference is that when $\mathcal{X}$ is a manifold with intrinsic dimension m, we can prove analogous results to Propositions 7-9, but with the ambient dimension d replaced by the intrinsic dimension m.
- Unlike in the full-dimensional case, our upper bounds on the estimation and testing error of Laplacian eigenmaps match the minimax rate only when $s \leq 3$. Our upper bounds when $s \geq 4$ follow from the embedding $H^s(\mathcal{X}; M) \subset H^3(\mathcal{X}; M)$, i.e they match the rates we get just by assuming that the 3rd order derivative is bounded, and are clearly suboptimal.

  We now explain this discrepancy. At a high level, thinking of the graph G as an estimate of the manifold $\mathcal{X}$, we incur some error by using Euclidean distance rather than geodesic distance to form the edges of G. This is in contrast with the full-dimensional setting, where the Euclidean metric exactly coincides with the geodesic distance for all points $x, z \in \mathcal{X}$ that are sufficiently close to each other and far from the boundary of $\mathcal{X}$. This extra error incurred in the manifold setting by using the "wrong distance" dominates when $s \geq 4$.

  As this explanation suggests, by building G using the geodesic distance one could avoid this error, and might obtain superior rates of convergence. However this is not an option for us, as we assume $\mathcal{X}$—and in particular its geodesics—are unknown. Likewise, a classical spectral projection estimator, using eigenfunctions of the manifold Laplace-Beltrami operator, will achieve the minimax rate for all values of s and m; but this is undesirable for the same reason—we do not want to assume that $\mathcal{X}$ is known. It is not clear whether this gap between spectral projection and Laplacian eigenmaps estimators—or more generally, between estimators which assume the manifold is known, and those which do not—is real, or a product of loose upper bounds.
- Finally, when $m > 4$, we get an upper bound on testing error equivalent to that of Theorem 16, except with the ambient dimension d replaced by intrinsic dimension m.

Analysis. The high-level strategy used to prove Theorems 17 and 18 is the same as in the flat-Euclidean setting. More specifically, we will use precisely the same bias-variance decompositions (4.24) (for estimation) and (4.25) (for testing). The difference will be that our bounds on the graph Sobolev seminorm $\langle L_{n,\varepsilon}^s f_0, f_0\rangle_n$, graph eigenvalue λ_K, and empirical norm $\|f_0\|_n^2$ will now always depend on the intrinsic dimension m, rather than the ambient dimension d. The precise results we use are contained in Propositions 11-13.

Proposition 11. *Suppose Model 4.2.2, and additionally that $f_0 \in H^s(\mathcal{X}; M)$ and $p \in C^{s-1}(\mathcal{X})$ for $s = 1, 2$ or 3. Then there exist constants c_0, C_0 and C that do not depend on f_0, n or M such that the following statement holds for any $\delta \in (0,1)$: if η satisfies (P5) and $C_0 n^{-1/(2(s-1)+m)} < \varepsilon < c_0$, then*

$$\langle L_{n,\varepsilon}^s f, f \rangle_n \leq \frac{C}{\delta} \|f\|_{H^s(\mathcal{X})}^2, \tag{4.35}$$

with probability at least $1 - 2\delta$.

As discussed previously, when $\mathcal{X}$ is a domain without boundary and Δ_P is the manifold weighted Laplace-Beltrami operator, appropriate bounds on the graph eigenvalues λ_k have already been derived in [Burago et al., 2014, García Trillos et al., 2019a,b]. The precise result we need is a simple consequence of Theorem 2.4 of [Calder and García Trillos, 2019].

Proposition 12 (c.f Theorem 2.4 of [Calder and García Trillos, 2019]). *Suppose Model 4.2.2. Then there exist constants c and C such that the following statement holds: if η satisfies (P5) and $C(\log n/n)^{1/m} < \varepsilon < c$, then*

$$\lambda_k \geq c \cdot \min\Big\{\lambda_k(\Delta_P), \frac{1}{\varepsilon^2}\Big\} \quad \text{for all } 1 \leq k \leq n, \tag{4.36}$$

with probability at least $1 - Cn\exp\{-cn\varepsilon^d\}$.

(For the specific computation used to deduce Proposition 12 from Theorem 2.4 of [Calder and García Trillos, 2019], see Green et al. [2021].)

Finally, using a Gagliardo-Nirenberg inequality for functions on compact Riemmanian manifolds, we obtain a lower bound on empirical norm $\|f\|_n$ under the hypotheses of Model 4.2.2.

Proposition 13. *Suppose Model 4.2.2, and additionally that $f_0 \in H^s(\mathcal{X}, M)$ for some $s > m/4$. There exists a constant C that does not depend on f_0 such that the following statement holds for all $\delta > 0$: if*

$$\|f_0\|_P \geq \frac{CM}{\delta^{s/m}} n^{-s/m}, \tag{4.37}$$

then with probability at least $1 - \exp\{-(cn \wedge 1/\delta)\}$,

$$\|f_0\|_n^2 \geq \frac{1}{2}\|f_0\|_P^2. \tag{4.38}$$

4.5 Out-of-sample error

Sections 4.3 and 4.4 show that $\widehat{f}$ is a minimax optimal estimator over Sobolev spaces. However, as mentioned previously we have measured loss *in-sample*—that is, measured in $L^2(P_n)$ norm—whereas *out-of-sample* error—error measured in $L^2(P)$ norm—is the more typical metric in the random design setup.

Of course, the Laplacian eigenmaps estimator is only defined at the observed design points $X_1, \ldots, X_n$, and to measure its error in $L^2(P)$ norm we must first extend it to be defined over all of $\mathcal{X}$. We propose a simple method, kernel smoothing, to do the job. The method can applied to any estimator defined at the design points, including Laplacian eigenmaps, and we show that a smoothed version of our original estimator $\widehat{f}$ has optimal $L^2(P)$ error. For simplicity, in this section we will stick to the flat Euclidean setting, where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed according to Model 4.2.1.

Extension by kernel smoothing. We now formally define our approach to extension by kernel smoothing. For a kernel function $\psi(\cdot) : [0, \infty) \to (-\infty, +\infty)$, bandwidth $h > 0$, and a distribution Q, the *Nadaraya-Watson kernel smoother* $T_{h,Q}$ is given by

$$\big(T_{Q,h}f\big)(x) := \begin{cases} \dfrac{1}{d_{Q,h}(x)} \displaystyle\int_\Omega f(z)\psi\bigg(\frac{\|z-x\|}{h}\bigg)\, dQ(z), & \text{if } d_{Q,h}(x) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $d_{Q,h}(x) := \int_\Omega \psi\big(\|z-x\|/\varepsilon\big)\, dQ(z)$. For convenience, we will write $T_{\varepsilon,n}f(x) := T_{\varepsilon,P_n}f(x)$, and $d_{n,h}(x) := n \cdot d_{P_n,h}(x)$. We extend the Laplacian eigenmaps estimator by passing the kernel smoother $T_{h,n}$ over it, that is we consider the estimator $T_{h,n}\widehat{f}$, which is defined at every $x \in \mathcal{X}$ (indeed, at every $x \in \mathbb{R}^d$). Note that "extension" here is a slight abuse of nomenclature, since $T_{h,n}\widehat{f}(X_i)$ and $\widehat{f}_i$ may not agree in-sample.

Out-of-sample error of kernel smoothed Laplacian eigenmaps. In Lemma 5, we consider an arbitrary estimator $\check{f} \in L^2(P_n)$. We show that the out-of-sample error $\|T_{n,h}\check{f} - f_0\|_P^2$ can be upper bounded by three terms— (a constant times) the in-sample error $\|\check{f} - f_0\|_n^2$, and variance and bias terms that arise naturally in the analysis of kernel smoothing over noiseless data.. We shall assume the following conditions on ψ and h.

(K3) The kernel function ψ is supported on a subset of $[0, 1]$. Additionally, ψ is Lipschitz continuous on $[0, 1]$, and is normalized so that

$$\int_{-\infty}^{\infty} \psi(|z|)\, dz = 1.$$

(P7) For constants c_0 and C_0, the bandwidth parameter h satisfies

$$C_0\bigg(\frac{\log(1/h)}{n}\bigg)^{1/d} \le h \le c_0.$$

Lemma 5. *Suppose Model 4.2.1, and additionally that $\check{f} \in L^2(P_n)$, $f_0 \in H^1(\mathcal{X})$ and $p \in C^1(\mathcal{X})$. If the kernel smoothing estimator $T_{h,n}\check{f}$ is computed with kernel ψ satisfying (K3) and bandwidth h satisfying (P7), it holds that*

$$\|T_{n,h}\check{f} - f_0\|_P^2 \le C\bigg(\|\check{f} - f_0\|_n^2 + \frac{1}{\delta} \cdot \frac{h^2}{nh^d}|f|_{H^1(\mathcal{X})}^2 + \frac{1}{\delta}\|T_{h,P}f_0 - f_0\|_P^2\bigg), \tag{4.39}$$

with probability at least $1 - \delta - Ch^d \exp\{-Cnh^d\}$.

Notice that the variance term in the above is smaller than the typical variance term for kernel smoothing of noisy data, by a factor of h^2. On the other hand the bias term is typical. When ψ is an order-s kernel, a standard analysis shows that the $\|T_{h,P}f_0 - f_0\|_P^2 \lesssim \varepsilon^{2s}$.

(K4) The kernel function ψ is an order-s kernel, meaning that it satisfies

$$\int_{-\infty}^{\infty} \psi(|z|)\, dz = 1, \quad \int_{-\infty}^{\infty} z^j \psi(|z|)\, dz = 0 \ \text{ for } j = 1, \dots, s+d-2, \quad \text{and} \quad \int_{-\infty}^{\infty} z^{s+d-1}\psi(|z|)\, dz < \infty.$$

Choosing $h \asymp n^{-1/(2(s-1)+d)}$ balances the kernel smoothing bias and variance terms in (4.39), and implies that

$$\|T_{n,h}\check{f} - f_0\|_P^2 \le C\bigg(\|\check{f} - f_0\|_n^2 + \frac{1}{\delta}n^{-2s/(2(s-1)+d)}\bigg). \tag{4.40}$$

(4.40) tells us that the additional error incurred by passing a kernel smoother over an in-sample estimator $\check{f}$ is negligible compared to the minimax rate of estimation. Consequently, if $\check{f}$ converges at the minimax

rate in $L^2(P_n)$, then $T_{n,h}\check{f}$ will converge at the minimax rate in $L^2(P)$. It follows immediately from Theorem 12 (when $s = 1$) or Theorem 14 (when $s > 1$), that $T_{h,n}\widehat{f}$ achieves the optimal rate of convergence in $L^2(P)$.

Theorem 19. *Suppose Model 4.2.1. There exist constants c, C, and N that do not depend on f_0 or n such that each the following statements hold with probability at least $1 - \delta - Cn\exp\{-cn\varepsilon^d\} - Ch^d\exp\{-cnh^d\}$, for all $n \geq N$ and for any $\delta \in (0, 1)$.*

- *If $f_0 \in H^1(\mathcal{X}; M)$, the Laplacian eigenmaps estimator $\widehat{f}$ is computed with parameters ε and K that satisfy (P1), and the out-of-sample extension $T_{h,n}\widehat{f}$ is computed with bandwidth $h = n^{-1/d}$ and kernel ψ that satisfies (K3), then*

$$\|T_{h,n}\widehat{f} - f_0\|_P^2 \leq \frac{C}{\delta}M^2(M^2n)^{-2s/(2s+d)}.$$

- *If $f_0 \in H_0^s(\mathcal{X}; M)$ and $p \in C^{s-1}(\mathcal{X})$ for some $s \in \mathbb{N}, s > 1$, and the Laplacian eigenmaps estimator $\widehat{f}$ is computed with parameters ε and K that satisfy (P3), and the out-of-sample extension $T_{h,n}\widehat{f}$ is computed with bandwidth $h = n^{-1/(2(s-1)+d)}$ and kernel ψ that satisfies (K3) and (K4), then*

$$\|T_{h,n}\widehat{f} - f_0\|_P^2 \leq \frac{C}{\delta}M^2(M^2n)^{-2s/(2s+d)}.$$

Some remarks:

- Since $T_{n,h}\widehat{f}$ is defined out-of-sample, we can use sample splitting or cross validation methods to tune hyperparameters, which we could not do for the original estimator $\widehat{f}$. For instance, we can (i) split the sample into two halves, (ii) use the first half to compute $T_{h,n}\widehat{f}$ for various values of ε, h, and K, (iii) choose the optimal values of these three hyperparameters by minimizing error on the held out set. Practically speaking, cross-validation is one of the most common approaches to choosing hyperparameters. Theoretically, it is known that choosing hyper-parameters through sample splitting can result in estimators that optimally adapt to the order of regularity s. In other words, it leads to estimators that are rate-optimal (up to $\log n$ factors), even when s is unknown. Similar arguments should imply that $T_{h,n}\widehat{f}$ is adaptive in this sense when ε, h and K are chosen by sample splitting.
- There exist many other approaches to extending a function f using only evaluations $\{f(X_1), \ldots, f(X_n)\}$. We consider extension by kernel smoothing because it is a simple and statistically optimal procedure that does not require any knowledge of the domain $\mathcal{X}$ or distribution P—as we have argued, this latter property is one of the main selling points of Laplacian eigenmaps as a tool for nonparametric regression.

  We now comment on a few alternative methods. One such approach is minimum norm interpolation. Here one defines a normed space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ and then solves the optimization problem

$$\min_{u \in \mathcal{F}} \|u\|_{\mathcal{F}}, \quad \text{such that} \quad u(X_i) = f(X_i) \quad \text{for } i = 1, \ldots, n.$$

  A particularly popular version of this general approach takes $\mathcal{F}$ to be an RKHS Rieger and Zwicknagl [2010], Belkin [2018], which encompasses thin-plate spline interpolation (where $\mathcal{F} = H^s(\mathcal{X})$ for $s > d/2$) as a special case. Naturally, this approach works well when f is close to a function $u \in \mathcal{F}$ with reasonably small norm. This holds true when $f \in H^s(\mathcal{X}; M)$ and $s > d/2$, but as already discussed when $s \leq d/2$ the Sobolev space $H^s(\mathcal{X})$ is not an RKHS, and in fact when $s \leq d/2$ thin-plate spline interpolation is ill-posed [Green and Silverman, 1993]. Another, arguably simpler approach is to extend f to be piecewise constant on the Voronoi tessellation induced by $X_1, \ldots, X_n$, or equivalently to perform 1-nearest neighbors regression on f. However, this approach is theoretically optimal only when f_0 is Lipschitz, in contrast to the kernel smoothing method we propose and study.

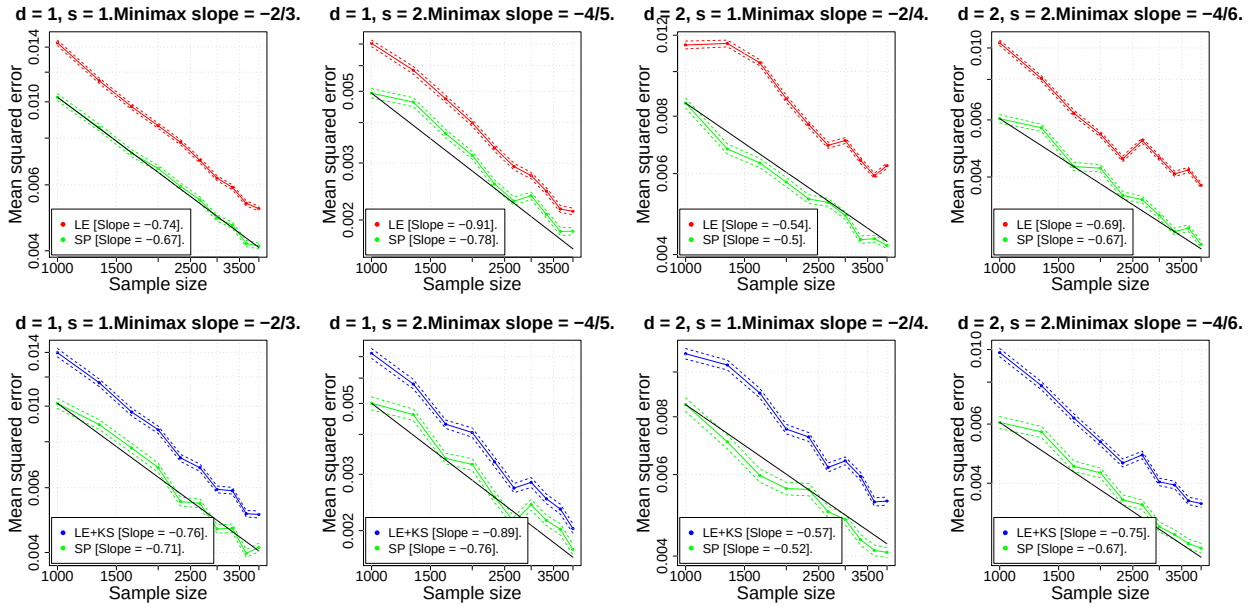


Figure 4.1: Mean squared error (mse) of Laplacian eigenmaps and spectral projection estimators. Top row: in-sample mse of Laplacian eigenmaps (`LE`) and a spectral projection estimator (`SP`) as a function of sample size n. Bottom row: out-of-sample mse of Laplacian eigenmaps plus kernel smoothing (`LE+KS`) and a spectral projection estimator. Each plot is on the log-log scale, and the results are averaged over 400 repetitions. All estimators are tuned for optimal average mse. The black line shows the minimax rate (in slope only; the intercept is chosen to match the observed error).

None of these three methods are intrinsically linked to Laplacian eigenmaps. This is in one sense a strength, since they can be used to extrapolate any estimator defined only in-sample. But it is also potentially a weakness. Each of these methods have their own approximation and estimation errors (bias and variance) which can be fundamentally different than those of Laplacian eigenmaps, and there is a danger that in extrapolating the Laplacian eigenmaps estimator in this way we are taking the "worst of both worlds". Our theory shows that when the data model is Model 4.2.1 and we perform extrapolation by kernel smoothing, this is not a problem, at least in a minimax sense.

4.6 Experiments

In this section we empirically demonstrate that Laplacian Eigenmaps is a reasonably good alternative to spectral projection, even when n is only moderately large. In order to compare the two methods, in our experiments we stick to simple settings where we can compute eigenfunctions of Δ_P, and thus the spectral projection estimator. Of course in general, it is not easy to compute these eigenfunctions: hence the appeal of Laplacian Eigenmaps.

In our first experiment, we compare the mean-squared error of Laplacian eigenmaps to that of its classical spectral projection counterpart. We vary the sample size from $n = 1000$ to $n = 4000$; sample n design points $X_1, \ldots, X_n$ from the uniform distribution on the cube $[-1, 1]^d$; and sample responses Y_i according to (4.3) with regression function $f_0 = M/\lambda_K^{s/2} \cdot \psi_K$ for $K \asymp n^{d/(2s+d)}$; the pre-factor $M/\lambda_K^{s/2}$ is chosen so that $|f_0|^2_{H^s(\mathcal{X})} = M^2$. In Figure 4.1 we show the in-sample mean-squared error of Laplacian eigenmaps and a classical spectral projection estimator as a function of n, for different dimensions d and order of smoothness s. We see that all estimators have mean-squared error converging to zero at roughly the minimax rate. We also see that the mean-squared error of Laplacian Eigenmaps gets closer to that of spectral projection as n

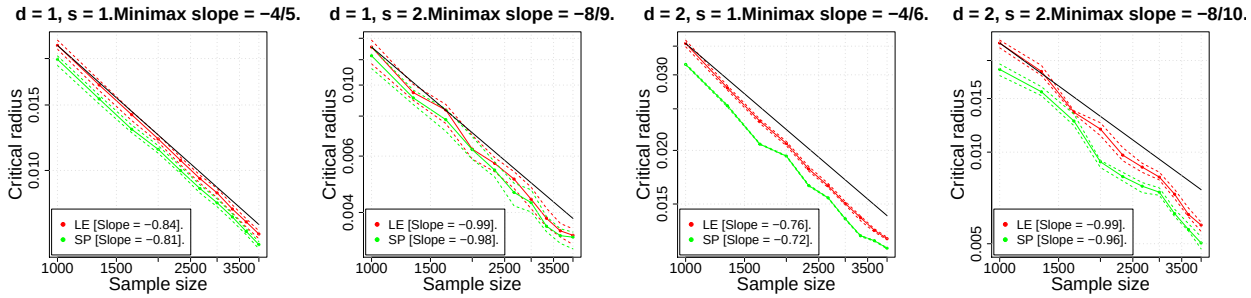


Figure 4.2: Worst-case testing risk Laplacian eigenmaps (LE) and spectral projection (SP) tests, as a function of sample size n. Plots are on the same scale as Figure 4.1, and black line shows the minimax rate. All tests are set to have .05 Type I error, and are calibrated by simulation under the null.

gets larger. The fact that spectral projection outperform Laplacian Eigenmaps

We also compare the error, over a held-out test set, of Laplacian eigenmaps plus kernel smoothing to the spectral projection estimator. The out-of-sample mean squared error of the two estimators is very similar to the in-sample mean-squared error. This supports our theoretical claim that the additional error incurred by kernel smoothing of Laplacian Eigenmaps is negligible.

In our second experiment, we compare tests using Laplacian eigenmaps and spectral projection test statistics. The setup, in terms of n and P, is the same as that of our first experiment. To empirically evaluate the critical radius $\epsilon_n(\phi, H^1(\mathcal{X}; M))$ of a test ϕ, we compute ϕ for each $f_0 \in \mathcal{F} \subset H^1(\mathcal{X}; M)$, where $\mathcal{F}$ is a discrete subset of $H^1(\mathcal{X}; M)$. For each of $b = 1, 2, \ldots, 100$, we compute ϵ_b, the smallest value of ϵ such that $R_n(\phi, \mathcal{F}, \epsilon_b) \geq b/100$. Then we take $\bar{\epsilon} = 1/100 \cdot \sum_{b=1}^{100} \epsilon_b$ to be our empirical measure of worst-case risk. In Figure 4.2, we see that the critical radii of both Laplacian eigenmaps and spectral projection tests are quite close to each other, and converge to 0 at roughly the minimax rate.

These experiments demonstrate that in terms of statistical error, Laplacian eigenmaps methods are reasonable replacements for spectral projection methods. Laplacian eigenmaps depends on two tuning parameters, and in our final experiment we investigate the importance of both, focusing now on estimation. In Figure 4.3, we see how the mean-squared error of Laplacian eigenmaps changes as each tuning parameter is varied. As suggested by our theory, properly choosing the number of eigenvectors K is crucial: the mean-squared error curves, as a function of K, always have a sharply defined minimum. On the other hand, as a function of the graph radius parameter ε the mean-squared error curve is much closer to flat. This squares completely with our theory, which requires that the number of eigenvectors K be much more carefully tuned that the graph radius ε.

We also plot the out-of-sample mean squared error of Laplacian eigenmaps plus kernel smoothing, as a function of its various tuning parameters (which include the bandwidth h as well as ε and K.) Here the relationship between theory and empirics is more nuanced. On the one hand, empirically it seems that the optimal choice of bandwidth parameter h is usually smaller than ε, as suggested by our theory. On the other hand, for Laplacian eigenmaps plus kernel smoothing we see that mean-squared error curves as a function of K are often quite close to their minima even when we choose many more eigenvectors than is optimal for Laplacian eigenmaps or spectral projection. This is not reflected in our theory, where we require that K be chosen in the same tight range as was required for Laplacian Eigenmaps to be optimal in-sample. However, it does make intuitive sense: extension by kernel smoothing further attenuates the noise, making the algorithm more forgiving to overfitting during the Laplacian Eigenmaps step.

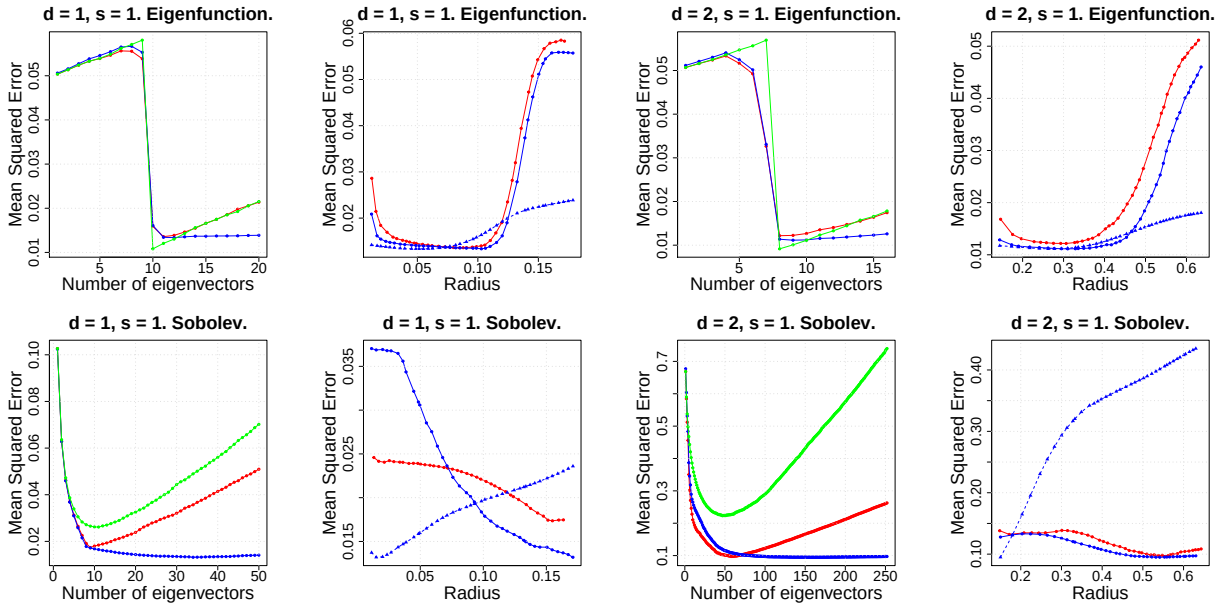


Figure 4.3: Mean squared error of Laplacian Eigenmaps (red), Laplacian Eigenmaps plus kernel smoothing (blue), and spectral projection (green) as a function of tuning parameters. Top row: the same regression function f_0 as used in Figure 4.1. Bottom row: the regression function $f_0 \propto \sum_k 1/\lambda_k^{1/2}\psi_k$. For all experiments, the sample size $n = 1000$, and the results are averaged over 200 repetitions. In each panel, all tuning parameters except the one being varied are set to their optimal values. For Laplacian Eigenmaps plus kernel smoothing, circular points and a solid line are used to denote the error as a function of the graph radius ε, whereas triangular points and a dashed line are used to denote the error as a function of the bandwidth h.

4.7 Future Work

We view our work can be viewed as a contribution both to the fields of nonparametric regression with series estimators, and to graph-based learning. We end our discussion by mentioning some open work in each of these directions.

Much is known about classical spectral projection methods beyond their rate optimality. For instance: such estimators and tests exhibit *sharp optimality*, meaning their risk is within a $(1 + o(1))$ factor of the optimal risk; they can adapt to unknown smoothness of the regression function; they can be used to estimate smooth functionals of the regression function; finally, they can be used to form confidence sets in $L^2(P)$. It would be interesting to see if Laplacian eigenmaps could replicate the performance of classical methods in any, or all, of these problems.

On the other hand, there are many variants of Laplacian eigenmaps worth considering. For instance, one can change the graph under consideration (e.g. by using the k-nearest neighbors), or the normalization of the graph Laplacian $L_{n,\varepsilon}$ (e.g. by using the symmetric normalized Laplacian). The former is practically useful, because it typically leads to connected graphs while always ensuring a given level of edge sparsity. In the latter, the graph Laplacian converges to a different limiting operator, which possesses different eigenvectors than Δ_P and thereby induces a different bias. We believe that under the setup we consider here, both methods will continue to be optimal.

Chapter 5

Discussion

In the previous two chapters, we separately considered two methods for estimation and testing: Laplacian smoothing (Chapter 3) and Laplacian eigenmaps (Chapter 4). In this chapter, we discuss the two methods jointly. We first compare some of their properties in Section 5.1. Then in Section 5.2, we compare both methods to some other classical methods for regression, which do not involve graphs. Each of these comparisons will be further illuminated by the equivalent kernel perspective, which we cover in Section 5.3. Throughout this section, we focus only on estimation. For ease of reading, as much as possible this chapter will be self-contained, meaning all quantities and notation are redefined, even if they have appeared in a prior chapter.

Setup. To that end, recall the setup of nonparametric regression given in Chapters 3 and 4. We observe design points $X_1, \dots, X_n$ independently sampled from a distribution P, with density p bounded away from 0 and ∞ on a domain $\mathcal{X} \subseteq \mathbb{R}^d$, $0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$. We assume the boundary $\partial\mathcal{X}$ is Lipschitz. Additionally we observe responses

$$Y_i = f_0(X_i) + w_i,$$

where $w_i \sim N(0,1)$ are independent Gaussian noise, and f_0 is the unknown regression function to be learned.

Both the Laplacian smoothing and Laplacian eigenmaps estimates are constructed using a neighborhood graph $G_{n,\varepsilon}$. To form this graph, we let $\eta : [0,\infty) \to [0,\infty)$ be a kernel that satisfies the hypotheses of (K1). For a radius parameter $\varepsilon > 0$, the graph $G_{n,\varepsilon} = (\{1,\dots,n\}, W)$ is formed over vertices $\{1,\dots,n\}$ corresponding to the design points $\{X_1,\dots,X_n\}$, and with a weighted edge $W_{ij} = \eta(\|X_i - X_j\|/\varepsilon)$ between vertices i and j. The neighborhood graph Laplacian $L_{n,\varepsilon} \in \mathbb{R}^{n\times n}$ is defined by its action on vectors $u \in \mathbb{R}^n$ as

$$\big(L_{n,\varepsilon}u\big)_i := \frac{1}{n\varepsilon^{d+2}} \sum_{j=1}^{n} \big(u_i - u_j\big)\eta\bigg(\frac{\|X_i - X_j\|}{\varepsilon}\bigg). \tag{5.1}$$

Equivalently, $L_{n,\varepsilon} = (n\varepsilon^{d+2})^{-1}(D - W)$, where D is the diagonal degree matrix with entries $D_{ii} = d_{n,\varepsilon}(X_i) = \sum_{j=1}^n W_{ij}$. We write $L_{n,\varepsilon} = \sum_{k=1}^n \lambda_k v_k v_k^\top$ for the spectral decomposition of $L_{n,\varepsilon}$, meaning

$$L_{n,\varepsilon} v_k = \lambda_k v_k, \quad v_k^\top v_k = 1.$$

Finally, we recall the two estimators analyzed in Chapters 3 and 4. The Laplacian smoothing estimator $\widehat{f}_{\text{LS}}$ is defined as

$$\widehat{f}_{\text{LS}} := (I + \rho L_{n,\varepsilon})^{-1}\mathbf{Y}, \tag{5.2}$$

and the Laplacian eigenmaps estimator $\widehat{f}_{\text{LE}}$ as

$$\widehat{f}_{\text{LE}} := \sum_{k=1}^{K} v_k v_k^\top \mathbf{Y}. \tag{5.3}$$

5.1 Comparison between Laplacian Smoothing and Laplacian Eigenmaps

We now compare Laplacian smoothing and Laplacian eigenmaps, along a few different axes. The first two—statistical and computational efficiency — are fairly standard. The last — regularity of the estimate — is somewhat less typical, and will reveal an interesting distinction between the two methods.

5.1.1 Statistical Efficiency

The main part of Chapters 3 and 4 deals with the statistical properties of Laplacian smoothing and Laplacian eigenmaps. We now very briefly summarize the conclusions of these chapters. In Chapter 3, we show that Laplacian smoothing is minimax rate-optimal over the Sobolev spaces $H^1(\mathcal{X})$ only when the dimension d satisfies $1 \leq d \leq 4$ (up to log factors when $d = 4$). In contrast, in Chapter 4 we show that Laplacian eigenmaps is minimax rate-optimal over $H^1(\mathcal{X})$ for all d. We also show that Laplacian eigenmaps can adapt to higher-order smoothness, i.e. it can be minimax rate-optimal over $H_0^s(\mathcal{X})$ when $s > 1$. Thus, the known statistical properties of Laplacian eigenmaps are much stronger than those of Laplacian smoothing.

5.1.2 Computational Efficiency

In this section we review several disparate lines of work regarding the computational properties of Laplacian smoothing and Laplacian eigenmaps. We begin by reviewing the fastest known algorithms for computing each solution, for a given choice of tuning parameters. Of course, typically we would like to compute solutions over a grid of tuning parameters, and we discuss how *graph sparsification* can make this process faster without substantially degrading the quality of the solution. In large part, the analysis of running time of these algorithms is *worst-case*, meaning it holds for any response vector $\mathbf{Y}$ and Laplacian matrix L. We conclude by returning to the setting of this thesis — where $L = L_{n,\varepsilon}$ is the Laplacian of a neighborhood graph, and $\mathbf{Y} = f_0 + \mathbf{w}$ is equal to a smooth signal plus independent noise — and mentioning some aspects of this problem which may lead to even faster computation.

In this setting, where the input data are design points $X_1, \ldots, X_n \in \mathbb{R}^d$, the first step of either Laplacian eigenmaps or Laplacian smoothing is forming the neighborhood graph $G_{n,\varepsilon}$. Naively, computing this graph takes $O(n^2)$ time (treating the dimension d as a constant), which will typically dominate the time necessary to solve either (5.2) or (5.3). A more clever approach—for instance, using a kd-tree—can speed this up to time $O(n \log n)$, but in practice this becomes very slow when d is even moderately large. However, forming the graph is an embarrassingly parallel operation, and can thus be quickly done in a distributed fashion. For this reason we ignore the complexity of forming $G_{n,\varepsilon}$ in our subsequent discussion, and treat the graph as having already been computed.

Algorithms for Laplacian smoothing and eigenmaps. Computing Laplacian smoothing as in (5.2) amounts to solving a single symmetric and diagonally dominant linear system of the form $Lf = \mathbf{Y}$, where L is a graph Laplacian.[1] A series of seminal works [Spielman and Teng, 2011, 2013, 2014] have shown that it possible to solve such systems in time nearly linear in the sparsity, that is, in the number of non-zero entries in L. We restate a formal result to this effect, from the review monograph of Vishnoi [2012].

[1] The Laplacian smoothing solution $\widehat{f} = \widehat{f}_{\text{LS}}$ satisfies $(I + \rho L_{n,\varepsilon})\widehat{f} = \mathbf{Y}$, not $L_{n,\varepsilon}\widehat{f} = \mathbf{Y}$. However, one can treat $(I + \rho L_{n,\varepsilon})$ as the Laplacian of a graph, in which all edges in $G_{n,\varepsilon}$ have been weighted by a factor of ρ, and a loop of weight $(n\varepsilon^{d+2})^{-1}$ has been added at every vertex.

Theorem 20 (Theorem 3.1 of Vishnoi [2012]). *There is an algorithm* `LSOLVE` *which takes as input a graph Laplacian L, a vector $\mathbf{Y}$, and an error parameter σ, and returns a vector f satisfying*

$$\|f - L^{\dagger}\mathbf{Y}\|_L \leq \sigma \|L^{\dagger}\mathbf{Y}\|_L \tag{5.4}$$

where $\|f\|_L := \sqrt{f^{\top} L f}$. The algorithm runs in time $\widetilde{O}(m \log(1/\sigma))$, where m is the number of non-zero entries in L.

(Here $\widetilde{O}(\cdot)$ hides factors of $\mathrm{poly}(\log n)$.). As discussed by Vishnoi [2012], the norms in (5.4) can be replaced by Euclidean norm without changing the computational complexity by more than $\log n$ factors.

We turn now to Laplacian eigenmaps. A naive approach to solving (5.3) involves first computing the eigenvectors $v_1, \ldots, v_K$, and then the inner products $\mathbf{Y}^{\top} v_1, \ldots, \mathbf{Y}^{\top} v_K$. However, the fastest known algorithms for computing eigenvectors $v_1, \ldots, v_K$ take time $\widetilde{O}(Km)$ [Musco and Musco, 2015].[2] Typically, in Laplacian eigenmaps one takes K to be polynomial in the number of vertices n, so that $\widetilde{O}(Km)$ is quite a bit larger than $\widetilde{O}(m)$.

More recent work [Frostig et al., 2016, Allen-Zhu and Li, 2017, Jin and Sidford, 2019] implies that computational cost of Laplacian eigenmaps can be independent of the number of eigenvectors K. The key insight towards achieving computational cost independent of K is that one need only compute the projection $\Pi_K = V_K V_K^{\top}$, and not each individual eigenvector v_k. Frostig et al. [2016] give an algorithm that finds a solution f satisfying $\|f - \Pi_K \mathbf{Y}\|_2^2 \leq \sigma \|\mathbf{Y}\|_2^2$ by solving a certain number of symmetric diagonally dominant linear systems. Loosely speaking, this result implies that any upper bound on the time needed to solve a symmetric diagonally dominant linear system, such as Theorem 20, can be translated into an upper bound on the time needed to compute Laplacian Eigenmaps. Unfortunately, the number of linear systems one needs to solve in order to approximately compute $\Pi_K \mathbf{Y}$ depends inversely on the spectral gap $(\lambda_{K+1} - \lambda_K)/\lambda_K$, and for neighborhood graphs the spectral gap is usually quite small. The subsequent work of Allen-Zhu and Li [2017], Jin and Sidford [2019] sharpen the dependence on the spectral gap, and the topic remains an area of active research.

Graph sparsification. The aforementioned discussion concerns solving (5.2) or (5.3) when the tuning parameters ρ or K are fixed. It is often desirable to compute candidate solutions for many values of these hyperparameters, and then choose among these candidates using some criterion, for instance by cross-validation. In this case *graph sparsification* can be helpful.

Graph sparsification is the process of producing a graph $H_{n,\varepsilon}$, also defined over the nodes $\{1, \ldots, n\}$ but with many fewer edges than $G_{n,\varepsilon}$. Once the sparse graph $H_{n,\varepsilon}$ is obtained, then one computes $\widetilde{f}_{\mathrm{LS}}$ or $\widetilde{f}_{\mathrm{LE}}$, defined as the solutions to (5.2) or (5.3) but computed with respect to the Laplacian $\widetilde{L}_{n,\varepsilon}$ of $H_{n,\varepsilon}$. As we have seen, the sparsity of the graph plays a key role in determining the time required to compute Laplacian smoothing or eigenmaps. On the other hand, it is important that $\widetilde{f}_{\mathrm{LS}}$ or $\widetilde{f}_{\mathrm{LE}}$ approximate $\widehat{f}_{\mathrm{LS}}$ or $\widehat{f}_{\mathrm{LE}}$, the solutions computed over the original graph $G_{n,\varepsilon}$. This produces a tradeoff: the sparser $H_{n,\varepsilon}$ is, the faster one can solve (5.2) or (5.3), but the worse the quality of approximation.

Spectral approximation is a particularly useful way to measure the quality with which $H_{n,\varepsilon}$ approximates $G_{n,\varepsilon}$. For some $\sigma > 0$, we say $H_{n,\varepsilon}$ is a $(1+\sigma)$-spectral approximation to $G_{n,\varepsilon}$ if

$$(1+\sigma)^{-1} u^{\top} \widetilde{L}_{n,\varepsilon} u \leq u^{\top} L_{n,\varepsilon} u \leq (1+\sigma) u^{\top} \widetilde{L}_{n,\varepsilon} u, \quad \text{for all } u \in \mathbb{R}^n.$$

Batson et al. [2012] show that for any graph G with n nodes, there exists a $(1+\sigma)$-spectral approximation H with $O(n/\sigma^2)$ edges. However, the algorithms supplied by Batson et al. [2012] to find such a graph H

[2] Technically speaking, these algorithms are designed to compute the eigenvectors associated with the largest eigenvalues of L, whereas the eigenvectors $v_1, \ldots, v_K$ needed for Laplacian eigenmaps are associated with the smallest eigenvalues of L. To fix this nit, we can just consider $d_{\max}(G_{n,\varepsilon})I - L_{n,\varepsilon}$.

are computationally infeasible. A more practicable approach [Spielman and Srivastava, 2011] is to sample $O(n \log n/\sigma^2)$ edges in a clever way, such that with high probability the resulting graph H is a $(1+\sigma)$-spectral approximation to G. The right choice of sampling probabilities are proportional to the effective resistance, and can be approximately computed in $\widetilde{O}(m)$ time, by solving $O(\log n)$ symmetric and diagonally dominant linear systems using the algorithm referred to in Theorem 20. For more details and alternative algorithms for sparsification, we refer to Batson et al. [2013], Sadhanala et al. [2016b].

If one wishes to solve (5.2) or (5.3) for only one choice of tuning parameter, the aforementioned method for sparsification provides no computational edge. As we have just reviewed, computing a sparse spectral approximation $H_{n,\varepsilon}$ takes $\widetilde{O}(m)$ time, which is the time it takes to solve solve (5.2) or (5.3) in the first place. On the other hand, sparsification is more computationally desirable when tuning over many values of hyperparameters. This is because the sparse graph $H_{n,\varepsilon}$ need be computed only once, and can then be used to solve (5.2) or (5.3) for all values of tuning parameters. Sadhanala et al. [2016b] provide some back-of-the-envelope calculations suggesting that sparsification is "worth it", from a computational perspective, whenever one wishes to seperately solve (5.2) for at least $O(\log^{3/2} n)$ different tuning parameters.

There has been relatively limited analysis of the statistical properties of graph sparsification. Here the chief question is: how small must σ be, in order for the solutions $\widetilde{f}_{\mathrm{LS}}$ and $\widetilde{f}_{\mathrm{LE}}$ to be sufficiently close to $\widehat{f}_{\mathrm{LS}}$ and $\widehat{f}_{\mathrm{LE}}$, respectively. Recall that the larger σ is, the sparser the graph $H_{n,\varepsilon}$ can be, and the more computational advantage achieved. Sadhanala et al. [2016b] give the only guarantees on the quality with which $\widetilde{f}_{\mathrm{LS}}$ approximates $\widehat{f}_{\mathrm{LS}}$, and to the best of our knowledge there has been no analysis of Laplacian eigenmaps over sparsified graphs. However, in the setting of nonparametric regression over neighborhood graphs, there are reasons to believe that we may take σ quite large without degrading statistical performance. We will discuss these reasons shortly.

(Even) faster algorithms in the setting of non-parametric regression. We now return to the setting of nonparametric regression over neighborhood graphs. In this setting, many of the computational times previously discussed are actually already quite fast. This is because the graph $G_{n,\varepsilon}$ is already quite sparse. For instance, suppose we take the graph radius $\varepsilon \asymp (\log n/n)^{1/d}$, so that with high probability $m \asymp n \log n$. We have shown that this choice of ε results in statistically optimal Laplacian smoothing and eigenmaps over first-order Sobolev classes. Computationally speaking, Theorem 20 shows that for this choice of ε and any choice of ρ, the Laplacian smoothing estimate can be computed in $\widetilde{O}(n)$ time, and the results of Frostig et al. [2016], Allen-Zhu and Li [2017], Jin and Sidford [2019] translate this into analogous upper bounds for computing Laplacian eigenmaps. In this case, there is no benefit to further sparsifying $G_{n,\varepsilon}$.

That being said, for higher order Sobolev classes, to show that Laplacian eigenmaps has minimax optimal statistical error we require $\varepsilon \gtrsim n^{-1/(2(s-1)+d)}$, which is much greater than $(\log n/n)^{1/d}$. For such a large choice of ε, it can potentially be much faster to first sparsify the graph and then compute the estimator, depending on how easy it is to sparsify $G_{n,\varepsilon}$, and how many edges are in the sparsified graph $H_{n,\varepsilon}$. In Section 4.3.4, we show that to the preserve minimax optimality of Laplacian eigenmaps, it suffices to let $H_{n,\varepsilon}$ be a relatively crude approximation of $G_{n,\varepsilon}$; for instance, taking $H_{n,\varepsilon}$ to be a 2-spectral approximation of $G_{n,\varepsilon}$ is fine. The results of Spielman and Srivastava [2011] imply that there exists a 2-spectral approximation of $G_{n,\varepsilon}$ with only $O(n)$ edges, which can be computed $\widetilde{O}(m)$ time.

Indeed, as discussed by Sadhanala et al. [2016b], it may be possible to find such a spectral approximation even more cheaply. Recall that the sparsification algorithm discussed in Spielman and Srivastava [2011] relies on sampling edges with probability proportional to their effective resistance, and that the bottleneck operation in graph sparsification was approximately computing these effective resistances. However, von Luxburg et al. [2014] have shown that in many geometric graphs $G_{n,\varepsilon}$, each edge (i, j) has an effective resistance converging to $1/d_{n,\varepsilon}(X_i) + 1/d_{n,\varepsilon}(X_j)$ (scaled by a constant) as $n \to \infty$. Suppose that in the algorithm of Spielman and Srivastava [2011], one were to use this theoretical limit in place of approximately computing effective resistances. This would reduce the time required to sparsify $G_{n,\varepsilon}$ to $\widetilde{O}(n)$, which is nearly linear in the number of nodes. It is not yet clear how to translate the bounds in von Luxburg et al. [2014] into a spectral

similarity guarantee for the resulting graph $H_{n,\varepsilon}$, but of course remember that we only need $H_{n,\varepsilon}$ to be a 2-spectral approximation to $G_{n,\varepsilon}$.

5.1.3 Regularity of Estimates

Now we compare the smoothness properties of Laplacian smoothing and Laplacian eigenmaps. Because each of these methods returns a vector $\widehat{f} \in \mathbb{R}^n$, we assess smoothness using the graph Sobolev seminorm.

Lemma 6. *Suppose $f_0 \in H^1(\mathcal{X})$, and $C_0(\log n/n)^{1/d} < \varepsilon < c_0$. Then there exist constants c, C, N which do not depend on f_0, such that the following statements hold.*

- *Let $\widehat{f}_{\mathrm{LE}}$ be given by (5.3). Suppose that additionally $\varepsilon \leq K^{-1/d}$. Then*

$$\mathbb{E}\Big[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n | X_1, \ldots, X_n\Big] \leq \bigg(\frac{K^{(2+d)/d}}{n} + \frac{\|f_0\|^2_{H^1(\mathcal{X})}}{\delta}\bigg), \tag{5.5}$$

  with probability at least $1 - \delta - Cn\exp\{-cn\varepsilon^d\}$.

- *Let $\widehat{f}_{\mathrm{LS}}$ be given by (5.2). Then*

$$\mathbb{E}\Big[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}}\rangle_n | X_1, \ldots, X_n\Big] \geq c\min\bigg\{\frac{\varepsilon^2}{\rho^2}, \frac{1}{\varepsilon^2}\bigg\}, \tag{5.6}$$

  with probability at least $1 - Cn\exp\{-cn\varepsilon^d\}$.

Correctly interpreted, Lemma 6 shows that $\widehat{f}_{\mathrm{LS}}$ is much less smooth than $\widehat{f}_{\mathrm{LE}}$, when both methods are properly tuned. To see this, recall that in Chapter 3, we show that when $\rho \asymp n^{-2/(2+d)}$, Laplacian smoothing is an optimal estimator over first-order Sobolev classes, when $1 \leq d \leq 4$. In Chapter 4, we show that when $K \asymp n^{d/(2+d)}$, Laplacian eigenmaps is an optimal estimator over first-order Sobolev classes, for all dimensions d. In both cases, these conclusions are valid when the graph radius $\varepsilon \asymp (\log(n)/n)^{1/d}$, and the function f_0 has unit norm in the first-order Sobolev space, $\|f_0\|^2_{H^1(\mathcal{X})} \leq 1$. For these choices of tuning parameters, we draw the following conclusions, each of which hold with probability at least $1 - \delta - Cn\exp\{-cn\varepsilon^d\}$.

- The first-order graph Sobolev seminorm of Laplacian eigenmaps is of a constant order, regardless of the dimension d:

$$\mathbb{E}\Big[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n | X_1, \ldots, X_n\Big] \lesssim \frac{1}{\delta}.$$

- Conversely, when $d > 1$, the first-order graph Sobolev seminorm of Laplacian smoothing is growing with n:

$$\mathbb{E}\Big[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}}\rangle_n | X_1, \ldots, X_n\Big] \gtrsim (\log n)^{2/d} \cdot n^{(2d-4)/(2+d)d}.$$

Thus when $d > 1$, Laplacian eigenmaps has a much smaller graph Sobolev seminorm than does Laplacian smoothing, when both methods are correctly tuned. Recall from Lemma 3 that the regression function f_0 also has bounded graph Sobolev seminorm, $\langle L_{n,\varepsilon}f_0, f_0\rangle_n \lesssim 1/\delta$ with probability at least $1 - \delta$. Thus Lemma 6 tells us that the solution $\widehat{f}_{\mathrm{LS}}$ to Laplacian smoothing is also much less smooth than the regression function f_0, except when $d = 1$. Interestingly, this means that in dimensions 2 and 3, Laplacian smoothing is an optimal estimator even though it has much larger graph Sobolev norm than f_0, and is thus an *improper* estimator.

Laplacian smoothing at an unlabeled point. The previous discussion suggests that the deficiencies of Laplacian smoothing as an estimator may be tied to its lack of smoothness. Now we show that remarkably, these deficiencies are in a certain sense purely an *in-sample* phenomenon. Suppose we compute the following variant of Laplacian smoothing,

$$\check{f}_{\mathrm{LS}} := \|\mathbf{Y} - f\|_{n-1}^2 + \rho \langle L_{n,\varepsilon} f, f \rangle_n. \tag{5.7}$$

To be perfectly clear, in (5.7) we are using the same penalty term as in (5.3), but in the loss term we do not measure loss at the nth data point, i.e. we ignore $(Y_n - f_n)^2$. We refer to $X_1, \dots, X_{n-1}$ as the labeled data, and X_n as the unlabeled data.

As before, suppose the regression function is first-order Sobolev smooth, $f_0 \in H^1(\mathcal{X})$. The following Theorem shows that the error of the estimator $\check{f}_{\mathrm{LS}}$ at the unlabeled point X_n is small—on the order of at most $n^{-2/(2+d)}$, which is the minimax optimal rate—for *all dimensions* d.

Theorem 21. *Suppose that the regression function $f_0 \in H^1(\mathcal{X}) \cap L^\infty(\mathcal{X})$, and that $\|f_0\|_{H^1(\mathcal{X})} \leq M$ and $\|f_0\|_{L^\infty(\mathcal{X})} \leq M$. If the estimator $\check{f}_{\mathrm{LS}}$ in (5.7) is computed with tuning parameters $0 \leq \rho \leq M^2 (M^2 n)^{-2/(2+d)}$ and $\varepsilon = n^{-1/(2+d)}$, then*

$$\big|(\check{f}_{\mathrm{LS}})_n - f_0(X_n)\big|^2 \leq C \frac{M^2}{\delta^2} (M^2 n)^{-2/(2+d)},$$

with probability at least $1 - C_1 \delta - C_2 \exp(-c_2 n \varepsilon^d)$.

Comparing Theorem 21 to Theorem 6, we see that when $d \geq 4$, the error of the estimator $\check{f}_{\mathrm{LS}}$ at the unlabeled point X_n is much smaller than (our upper bound on) the in-sample mean-squared error of $\widehat{f}_{\mathrm{LS}}$.[3] Moreover, one can show that $\|\check{f}_{\mathrm{LS}} - f_0\|_{n-1}^2$, the mean-squared error of $\check{f}_{\mathrm{LS}}$ over the labeled points $X_1, \dots, X_n$, is comparable to the in-sample mean-squared error $\|\check{f}_{\mathrm{LS}} - f_0\|_n^2$. Seemingly, $\check{f}_{\mathrm{LS}}$ has stronger properties at the one point (X_n) for which the response was discarded during training, than at the $n-1$ points $(X_1, \dots, X_{n-1})$ for which the responses were used during training. Put more simply, $\check{f}_{\mathrm{LS}}$ appears to be a better estimator at unlabeled data than it is at labeled data.

The proof of Theorem 21 sheds light on this phenomenon. It turns out that $\check{f}_{\mathrm{LS}}$ is performing two types of regularization. The first is the explicit regularization controlled by the parameter ρ, and affects the fit at all points $X_1, \dots, X_n$. The second is a kind of implicit regularization, and affects the fit only at the unlabeled point X_n; it is simply kernel smoothing, using the kernel η and bandwidth ε. That is, the fit $\check{f}_{\mathrm{LS}}$ at point X_n is given by

$$(\check{f}_{\mathrm{LS}})_n = \frac{1}{d_{n,\varepsilon}(X_n)} \sum_{i=1}^{n-1} (\check{f}_{\mathrm{LS}})_i \eta\bigg(\frac{\|X_i - X_n\|}{\varepsilon}\bigg), \tag{5.8}$$

where we recall the degree is $d_{n,\varepsilon}(X_n) = \sum_{i=1}^n \eta(\|X_i - X_n\|/\varepsilon)$. As it turns out, this second kind of regularization attenuates the noise without overly smoothing out the underlying trend, whereas we cannot prove that any choice of ρ has the same effect in-sample. This also explains why in Theorem 21 we allow ρ to take a wide range of values, but require $\varepsilon = n^{-1/(2+d)}$ to be precisely the right choice for optimal kernel smoothing.

We conclude by offering some joint interpretation to the results of this section. Lemma 6 establishes that $\widehat{f}_{\mathrm{LS}}$ has large in-sample graph Sobolev semi-norm, and it is not hard to show that $\check{f}_{\mathrm{LS}}$ inherits this property. On the other hand, in the proof of Theorem 21, we show that $\check{f}_{\mathrm{LS}}$ additionally attenuates noise at the out-of-sample point X_n, by passing a kernel smoother over the in-sample fitted values $(\check{f}_{\mathrm{LS}})_1, \dots, (\check{f}_{\mathrm{LS}})_{n-1}$. We believe (although we cannot yet show) that in a certain sense, these properties of Laplacian smoothing continue to hold in a *bona fide* semi-supervised setting, where there are many unlabeled points; for analysis of

[3] Of course, the hypothesis that $f_0 \in L^\infty(\mathcal{X}) \cap H^1(\mathcal{X})$, which we assume in Theorem 21, is meaningfully stronger than merely $f_0 \in H^1(\mathcal{X})$, which is what we assume in Theorem 6. We believe Theorem 21 should hold without assuming the regression function is bounded, but cannot prove it.

a related estimator in the noiseless case $\mathbf{Y} = f_0$, see Calder et al. [2020b]. If so, this would mean Laplacian smoothing is a spiky estimator—unsmooth at labeled points but smoother around unlabeled data—with better estimation properties at unlabeled points. The choice of the word "spiky" is not casual. As observed by various authors [Nadler et al., 2009, El Alaoui et al., 2016, Slepčev and Thorpe, 2017, Calder et al., 2020b], the solutions to problems which use graph-based regularizers often exhibit spikes at labeled data, even when they have well-posed continuum limits and appear smooth at unlabeled data. This phenomenon is related to, although distinct from, statistically well-behaved *interpolation*, which has garnered much recent interest in the statistical community. We investigate the structure of the Laplacian smoothing estimate further in Section 5.3.

5.2 Graph Laplacian methods and the cluster assumption

Chapters 3 and 4 show that Laplacian smoothing and Laplacian eigenmaps, respectively, are minimax optimal methods for nonparametric regression over certain Sobolev classes. These are not the only optimal methods. For instance, kernel smoothing and least squares using an appropriate set of basis functions as features are two other minimax optimal methods over these Sobolev classes. We now give an example where graph Laplacian methods are better than these two alternatives, in the sense of having (much) smaller risk. This is possible because Laplacian smoothing and Laplacian eigenmaps perform remarkably well when the regression function f_0 and design distribution P satisfy a *cluster assumption*: that is, when the regression function is (approximately) piecewise constant over high-density clusters of the design distribution P. On the other hand, kernel smoothing (with Euclidean distance) and least squares (using eigenfunctions of an unweighted Laplace operator) cannot take advantage of the cluster assumption. We call this property of graph Laplacian based estimators *density adaptivity*.

5.2.1 Setup

We begin by specifying a sequence of design densities and regression functions $\{(p^{(n)}, f_0^{(n)}) : n \in \mathbb{N}\}$. These distributions will all be chosen to satisfy the cluster assumption. To that end, we define two clusters $Q_1, Q_2 \subset \mathbb{R}$ using a cluster separation parameter r, as

$$Q_1 := [0, 1/2 - r], \quad Q_2 := [1/2 + r, 1],$$

and take the domain $\mathcal{X}^{(n)} := Q_1 \cup Q_2$. We then take the design density to be uniform over $\mathcal{X}^{(n)}$ and the regression function to be a piecewise constant function over Q_1 and Q_2 of height θ,

$$p^{(n)}(x) := \frac{1}{1-2r}\mathbf{1}\{x \in Q_1 \cup Q_2\}, \quad f_0^{(n)}(x) := \theta \cdot \Big(\mathbf{1}\{x \in Q_1\} - \mathbf{1}\{x \in Q_2\}\Big). \tag{5.9}$$

Thus $p^{(n)}$ and $f_0^{(n)}$ belong to a two-parameter family, where the parameters are the cluster separation r and height θ. Generally speaking, the smaller the separation r, and the larger the height θ, the more graph Laplacian methods will outperform both kernel smoothing and linear regression using eigenfunctions of the unweighted Laplace operator as features.

We have already defined Laplacian smoothing and Laplacian eigenmaps. Kernel smoothing and least squares using eigenfunction of an unweighted Laplace operator are defined in Chapter 4, but for completeness we review their definitions here. For a kernel function ψ and bandwidth parameter h, the kernel smoothing estimator $\widetilde{f}_{\mathrm{KS}}$ is defined at a point $x \in \mathcal{X}$ as

$$\widetilde{f}_{\mathrm{KS}}(x) := \begin{cases} 0, & \text{if } d_{n,h}(x) = 0, \\ \dfrac{1}{d_{n,h}(x)} \displaystyle\sum_{i=1}^{n} Y_i \psi\left(\frac{\|X_i - x\|}{h}\right), & \text{otherwise.} \end{cases} \tag{5.10}$$

Let $(\lambda_1, \phi_1), (\lambda_2, \phi_2), \ldots$ be eigenpairs of the unweighted Laplace operator Δ on $[0,1]$, meaning

$$\Delta\phi_k = \lambda_k\phi_k, \quad \|\phi_k\|_{L^2([0,1])} = 1, \quad \frac{d}{dx}\phi_k(0) = \frac{d}{dx}\phi_k(1) = 0. \tag{5.11}$$

In this case the eigenfunctions ϕ_k of Δ are simply cosine functions, with eigenvalues proportional to their squared frequency. Noting that $\phi_1(x) = 1$ and $\lambda_1 = 0$, for $k = 2, 3, \ldots$ we have

$$\phi_k(x) = \sqrt{2} \cdot \cos(2\pi k x), \quad \lambda_k(\Delta) = \pi^2 k^2.$$

The least squares estimator using $\phi_1, \ldots, \phi_K$ $(1 \le K \le n)$ eigenfunctions as features is simply[4]

$$\widetilde{f}_K := \operatorname*{argmin}_{f \in \mathrm{span}\{\phi_1, \ldots, \phi_K\}} \|Y - f\|_n^2 = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top Y. \tag{5.12}$$

Hereafter, we will refer to $\widetilde{f}_K$ as the *uniform least squares* estimator.

5.2.2 Upper bounds on risk of graph Laplacian methods

Now we are in a position to state our results. Each of Laplacian smoothing, Laplacian eigenmaps, and kernel smoothing depend in part on the choice of kernel. For simplicity, in our analysis we only consider the boxcar kernel,

$$\eta(z) = \psi(z) = \mathbf{1}\{z \le 1\}. \tag{5.13}$$

This is strictly for convenience, and the following results will also hold for any kernel that satisfies (K1).

Proposition 14. *Suppose $(X_1, Y_1), \ldots (X_n, Y_n)$ are sampled according to (5.9).*

- *Compute the Laplacian eigenmaps estimator $\widehat{f} = \widehat{f}_{LE}$ using a kernel η which satisfies (5.13), number of eigenvectors $K = 2$, and radius $\varepsilon = r/2$. Then,*

$$\mathbb{E}\Big[\|\widehat{f} - f_0^{(n)}\|_n^2\Big] \le \left(6\theta^2 + \frac{1}{n}\right) \cdot \frac{8}{r}\exp(-nr/8) + \frac{1}{n} \tag{5.14}$$

- *Compute the Laplacian smoothing estimator $\widehat{f} = \lim_{\rho\to\infty} \widehat{f}_{LS}$ using the same kernel η and radius ε. Then the same guarantee (5.14) holds.*

5.2.3 Lower bounds on risk of kernel smoothing and least squares

Proposition 15. *Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled according to (5.9). Suppose $(\log n)^2/n \le r \le c$, where c is a universal constant.*

- *Compute the kernel smoothing estimator $\widetilde{f} = \widetilde{f}_{\mathrm{KS}}$ as in (5.10), using a kernel ψ which satisfies (5.13). Then there exist universal constants $c, N > 0$ such that for all $n > N$,*

$$\inf_{h'>0} \mathbb{E}\Big[\|\widetilde{f} - f_0^{(n)}\|_n^2\Big] \ge c\min\left\{\frac{r^{-1}}{n}, \frac{\theta}{\sqrt{n}}\right\}. \tag{5.15}$$

- *Compute the least squares estimator $\widetilde{f} = \widetilde{f}_{\mathrm{SP}}$ as in (5.12). Then there exist universal constants $c, N > 0$ such that for all $n > N$,*

$$\inf_{1\le K\le n} \mathbb{E}\Big[\|\widetilde{f} - f_0^{(n)}\|_n^2\Big] \ge c\min\left\{\frac{r^{-1}}{n}, \frac{1}{\log(n)}, \frac{r^{-2/3}}{n}, \frac{\sqrt{\theta}}{n^{3/4}}\right\}. \tag{5.16}$$

[4]For convenience, we will assume $\Phi \in \mathbb{R}^{n\times K}$ is full rank. If this is not the case, the least squares estimator $\widetilde{f}_K$ is not uniquely defined, but any solution will equal $\mathbf{Y}$ in-sample, and will satisfy $\|\widetilde{f}_K - f_0\|_n^2 \ge 1/2$ with high probability.

Together, Propositions 14 and 15 illustrate that the risk of Laplacian eigenmaps and Laplacian smoothing can be dramatically smaller than that of kernel smoothing or uniform least squares. For instance, taking $\theta = n^{1/2}$ and $r = n^{-3/4}$, when appropriately tuned, $\widehat{f} = \widehat{f}_{\mathrm{LE}}$ or $\widehat{f} = \widehat{f}_{\mathrm{LS}}$ satisfy

$$\mathbb{E}\Big[\|\widehat{f} - f_0^{(n)}\|_n^2\Big] \leq C\bigg(n^{7/4}\exp(-n^{1/4}/8)) + \frac{1}{n}\bigg) \leq \frac{C}{n},$$

for a universal constant C and all n larger than some universal constant N, whereas for $\widetilde{f} = \widetilde{f}_{\mathrm{KS}}$,

$$\inf_{h'>0} \mathbb{E}\Big[\|\widetilde{f} - f_0^{(n)}\|_n^2\Big] \geq \frac{c}{n^{1/4}},$$

and for $\widetilde{f} = \widetilde{f}_{\mathrm{SP}}$,

$$\inf_{1\leq K\leq n} \mathbb{E}\Big[\|\widetilde{f} - f_0^{(n)}\|_n^2\Big] \geq \frac{c}{n^{1/2}}.$$

Other choices of θ and r lead to even more dramatic gaps between the risk of Laplacian-based estimators, and the risk of kernel smoothing and least squares. The overall takeaway is that under Model 5.9, estimators that use the graph Laplacian can converge to the true regression function $f_0^{(n)}$ at fast rates—parametric rates that do not depend on the L^2 norm of $f_0^{(n)}$—whereas other estimators, optimal for estimation over Sobolev spaces, converge to $f_0^{(n)}$ at slow rates—nonparametric rates that deteriorate as the L^2 norm of $f_0^{(n)}$ grows.

Some remarks:

- The lower bound on the in-sample risk of $\widetilde{f}_{\mathrm{KS}}$ given by (5.15) is larger than that of $\widetilde{f}_{\mathrm{SP}}$ given by (5.16). This does not mean that kernel smoothing exhibits less adaptivity to the cluster assumption than uniform least squares. Instead, we suspect it is due to looseness in our lower bounds: we are able to tightly control the bias of kernel smoothing, whereas we must use a potentially loose bound on the bias of uniform least squares. Experimentally, it appears that kernel smoothing usually outperforms uniform least squares, under various instantiations of the cluster assumption.

- The cluster assumption—in which the regression function is piecewise constant and p consists of multiple connected components—is a very strong assumption. The *low-density separation* condition is a related but weaker assumption, in which the regression function is assumed to be smoother (but not constant) in regions of higher density. This is a rather general hypothesis which can formalized in a number of different ways. For instance, one could insist that the regression function f_0 belong to a normed ball in a *weighted Sobolev space*, with semi-norm given by

  $$|f_0|_{H^s(P)} := \langle \Delta_P^s f_0, f_0 \rangle_P.$$

  Intuitively, when $\|f_0\|_{H^s(P)}$ is much smaller than $\|f_0\|_{H^s(\mathcal{X})}$, density-adaptive learners such as Laplacian eigenmaps or Laplacian smoothing should have the advantage on non-density adaptive linear smoothers, such as kernel smoothing or uniform least squares. Indeed, in the case of Model 5.9 we see that

  $$\|f_0^{(n)}\|_{H^s(P^{(n)})} = 0 \quad \text{for all } s \in \mathbb{N}, \text{ and all } r, \theta > 0,$$

  whereas $f_0^{(n)}$ does not even belong to the first-order Sobolev space $H^1([0,1])$. In words, this shows the cluster assumption is an extreme case of the low-density separation condition. Unfortunately, it is quite difficult to analyze graph-based estimators under the general low-density separation condition, without making strong assumptions on P.

- Finally, we note that either changing the graph or the normalization of the Laplacian fundamentally alters the type of density adaptivity displayed by graph-Laplacian-based estimators; see Hoffmann et al. [2019] for an extensive discussion.

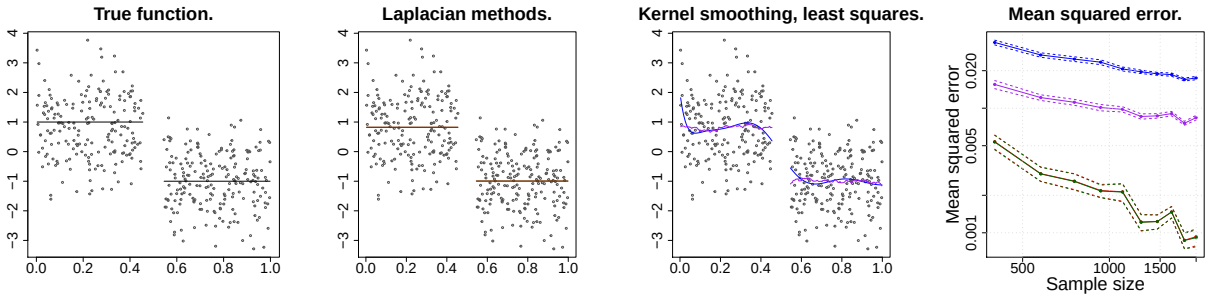


Figure 5.1: Estimation error of Laplacian eigenmaps (red), Laplacian smoothing (green), spectral projection with eigenfunctions of unweighted Laplace-Beltrami operator (blue), and kernel smoothing (purple) under Model 5.9. The leftmost plot shows the true regression function; the middle two plots show the fits of Laplacian eigenmaps and smoothing (left-middle), and of spectral projection and kernel smoothing (right-middle); the rightmost plot shows the mean squared error of all four methods as a function of sample size.

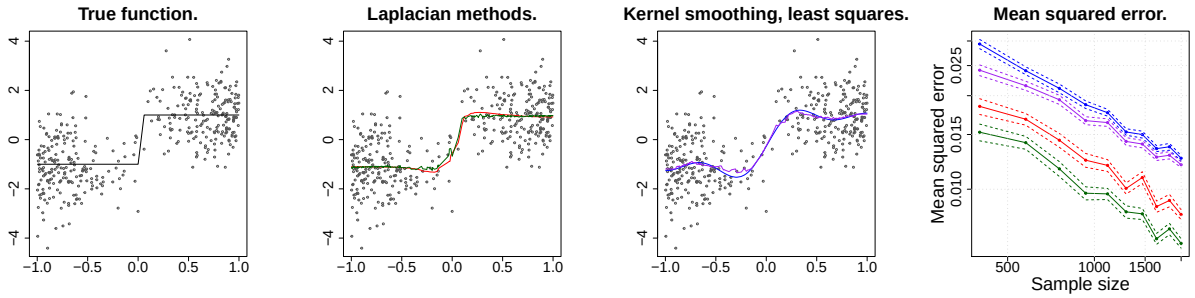


Figure 5.2: Same as Figure 5.1, but with p as in (5.17).

5.2.4 Experiments

We begin by verifying the practical relevance of Propositions 14 and 15 through simulation. We sample $n = 400, \dots, 2000$ points according to Model (5.9), with $r = (\log n)^2/(2n)$ and $\theta = 1$. We examine the empirical behavior of each of Laplacian smoothing, Laplacian eigenmaps, kernel smoothing, and uniform least squares. In Figure 5.1, we see that Laplacian smoothing and Laplacian eigenmaps indeed have smaller risk than kernel smoothing and uniform least squares, and moreover that the risk is decreasing at a faster rate. We also see that the Laplacian methods are able to perfectly recover the piecewise constant structure of f_0. On the other hand both kernel smoothing and least squares overfit, and least squares additionally displays boundary bias at points X_i near $1/2$.

In our second experiment, we keep the same regression function f_0 as in our first experiment. We change the design distribution P to be a mixture of two Gaussians truncated in $[-1, 1]$; specifically, P is the distribution with density

$$p(x) \propto \frac{1}{2}\Big(\varphi((x+1)/.4)) + \varphi((x-1)/.4)\Big) \cdot \mathbf{1}\{x \in [-1,1]\}, \tag{5.17}$$

with φ being the standard normal probability density function. Thus (p, f_0) no longer satisfy the cluster assumption, but rather only a weaker low-density separation condition. Even under these weaker condition, Laplacian eigenmaps and smoothing still handily outperform kernel smoothing and uniform least squares.

In our final experiment, we keep the same density p as in our second experiment, but change the regression function f_0. Instead of a stepfunction, we consider two alternatives, both of which are spatially inhomoge-

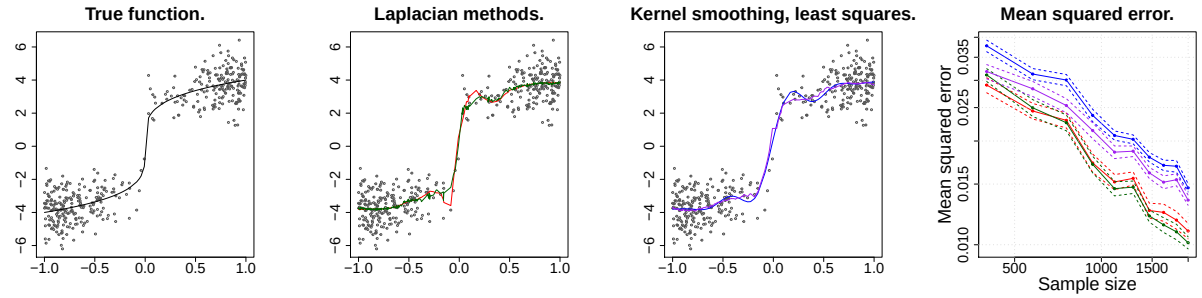


Figure 5.3: Same as Figure 5.2, but with $f_0(x) = x^{1/4}$.

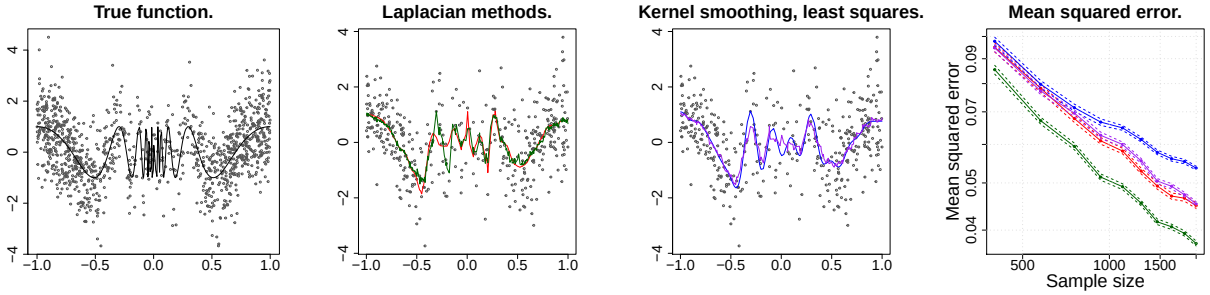


Figure 5.4: Same as Figure 5.2, but with $f_0(x) = \cos(4\pi|x|^{-1/3})$.

neous: the cubed root and Doppler functions,

$$f_0(x) = 4x^{1/4}, \quad \text{and } f_0(x) = \cos(4\pi|x|^{-1/3}).$$

When paired with the design distribution P, these functions satisfy the low density-separation condition, but to a lesser degree than the stepfunction. Unsurprisingly, the difference between the density adaptive graph-based estimators, and their non-density adaptive competitors, is more muted. Nevertheless, in Figures 5.3 and 5.4 we still see that graph-based estimators generally continue to have the edge.

5.3 Equivalent kernel perspective

Laplacian smoothing and Laplacian eigenmaps are both linear smoothers, meaning there exists a matrix $S \in \mathbb{R}^{n\times n}$ such that $\widehat{f} = SY$ for both $\widehat{f} = \widehat{f}_{\text{LS}}$ and $\widehat{f} = \widehat{f}_{\text{LE}}$. We can equivalently write this as

$$\widehat{f}(X_i) = \frac{1}{n}\sum_{j=1}^{n} g_{X_i}(X_j)Y_j \tag{5.18}$$

where for all $x \in \mathbb{R}^d$, $g_x : \mathbb{R}^d \to \mathbb{R}$. For a given estimator $\widehat{f}$, a function $g_x(\cdot)$ that satisfies (5.18) is the *equivalent kernel* of $\widehat{f}$.

Silverman [1984] studied the equivalent kernel of the smoothing spline estimator. We recall the definition of smoothing splines: for a given number of derivatives s,

$$\widetilde{f}_{\text{SS}} := \operatorname*{argmin}_{H^1(\mathbb{R})} \|Y - f\|_n^2 + \rho \int (f^{(S)}(x))^2\, dx. \tag{5.19}$$

Silverman [1984] showed that for any x, the equivalent kernel $g_x(\cdot)$ of $\widetilde{f}_{\text{SS}}$ asymptotically approaches $(p(x))^{-1}\cdot h\cdot\kappa(\|x-t\|/h)$, where the bandwidth $h = (\rho\cdot p(x))^{1/2s}$, and the kernel function κ is the fundamental solution

of the differential equation

$$(-1)^s \kappa^{(2s)} + \kappa = \delta_0. \tag{5.20}$$

The fundamental solution of (5.20) can be explicitly found for any value of s. For instance when $s = 1$, we have that the κ is the Laplace density function,

$$\kappa(x) = 1/2 \exp(-|x|).$$

These calculations show that, for smoothing splines, the regularization parameter ρ and density p determine the level of smoothening at a design point x. On the other hand, the order of derivative s reveals the shape of κ, and in turn shows the form of the ultimate estimator $\widehat{f}$. Together, these shed substantial light on the structure of smoothing spline estimates.

In this section we explore Laplacian smoothing from the equivalent kernel perspective. We begin in Section 5.3.1 by providing some heuristic calculations, which suggest that the equivalent kernel of Laplacian smoothing is close to the fundamental solution of a particular partial differential equation (PDE). In Section 5.3.2, we show how the bandwidth of this solution can be written as a function of the density p and regularization parameter ρ. Finally in Section 5.3.3, we show that the resulting kernel is not very smooth, particularly as the dimension d gets larger. Together these developments allow us to make a series of predictions regarding the nature of the equivalent kernel of Laplacian smoothing. In Section 5.3.5, we support these predictions with empirical evidence.

5.3.1 Discrete-to-continuum

Recall that the Laplacian smoothing estimator $\widehat{f}_{\text{LS}} = (I + \rho L_{n,\varepsilon})^{-1}\mathbf{Y}$. We fix a given point X_i, and let $\widehat{g} \in \mathbb{R}^n$ denote the equivalent kernel of Laplacian smoothing at $x = X_i$. Clearly, $\widehat{g}$ is the ith row of the smoother matrix $(I + \rho L_{n,\varepsilon})^{-1}$, multiplied by a factor of n, and so it satisfies the first-order condition

$$\frac{1}{n}(I + \rho L_{n,\varepsilon})\widehat{g} = \delta_i \tag{5.21}$$

where $\delta_i \in \mathbb{R}^n$ is the Kronecker delta, $(\delta_i)_j = \mathbf{1}\{i = j\}$.

Ideally, we would like to show that the solution $\widehat{g}$ to (5.21) is close to the fundamental solution g of the differential equation

$$\frac{1}{p(x)}(I + \rho\Delta_P)g = \delta_x, \tag{5.22}$$

where δ_x is the Dirac delta centered at $x = X_i$ and $\Delta_P := -1/p \cdot \operatorname{div}(p^2\nabla)$ is a weighted Laplace-Beltrami operator. If we could relate (5.21) to (5.22), then we could use the results of Silverman [1984], Wang et al. [2013] to determine the structure of g, and in turn the structure of $\widehat{g}$. Unfortunately, it is not at all straightforward to relate the solutions to (5.21) and (5.22).

A difficult calculation. Let us pause for a moment, to review why we might have expected the solutions $\widehat{g}$ and g to be close in the first place. It is well-known that the graph Laplacian $L_{n,\varepsilon}$ approaches the continuum Laplace operator, in the sense that

$$(L_{n,\varepsilon}u)_i \to \Delta_P u(X_i) \quad \text{as } n \to \infty, \varepsilon \to 0, n\varepsilon^{d+2} \to \infty$$

for each design point X_i in the interior of $\mathcal{X}$, and any function u sufficiently smooth, say $u \in C^3$. Suppose we knew that

$$(I + \rho L_{n,\varepsilon})\widehat{h} = u \quad \text{in } \mathbf{X}, \quad \text{and} \quad (I + \rho\Delta_P)h = u \quad \text{in } \mathcal{X}.$$

Then the following basic algebra relates $(I + \rho L_{n,\varepsilon})h$ and $(I + \rho L_{n,\varepsilon})\widehat{h}$,

$$(I + \rho L_{n,\varepsilon})h = (I + \rho\Delta_P)h + \rho(L_{n,\varepsilon} - \Delta_P)h = (I + \rho L_{n,\varepsilon})\widehat{h} + \rho(L_{n,\varepsilon} - \Delta_P)h. \tag{5.23}$$

By rearranging we obtain

$$(I + \rho L_{n,\varepsilon})(\widehat{h} - h) = \rho(L_{n,\varepsilon} - \Delta_P)h,$$

and taking the $L^2(P_n)$ norm of both sides then gives

$$\frac{1}{\rho^2}\|\widehat{h} - h\|_n^2 + \frac{2}{\rho}\langle L_{n,\varepsilon}(\widehat{h} - h), \widehat{h} - h\rangle_n + \|L_{n,\varepsilon}(\widehat{h} - h)\|_n^2 = \|(L_{n,\varepsilon} - \Delta_P)h\|_n^2. \tag{5.24}$$

In words, we have that the magnitude of the difference between the solutions $\widehat{h}$ and h, measured in a linear combination of various graph Sobolev semi-norms, is equal to the $L^2(P_n)$ norm of $(L_{n,\varepsilon} - \Delta_P)h$. As we have already mentioned, this latter quantity converges to 0 as $n \to \infty$ so long as h is sufficiently smooth.

Unfortunately, there are two reasons why these calculations do not apply to $\widehat{g}$ and g. The first is that, as we will see, the solution g to (5.22) does not possess enough regularity to imply $L_{n,\varepsilon}g \to \Delta_P g$. Even more fundamentally, the right hand sides of (5.21) and (5.22) are not equal; indeed the latter is not even a function. For this reason, is not possible to make sense of (5.23) when h and $\widehat{h}$ are replaced by g and $\widehat{g}$.

The bottom line is that it is very non-trivial to relate the equivalent kernel of Laplacian smoothing on a neighborhood graph to the fundamental solution of a relevant PDE. Instead, we will proceed by considering a pair of idealized geometric graphs in place of the neighborhood graph; these idealized graphs will have a very special structure which makes it easier to relate their equivalent kernels to the solution of a continuum PDE. We will draw several conclusions from these calculations, which we conjecture are not dependent on the special structure of the idealized graphs but instead apply more to more general classes of geometric graphs, including the neighborhood graphs otherwise considered in this work. Experimentally we will verify that each of these conclusions indeed appear to apply to neighborhood graphs.

Idealized Graph 1: Chain. Our first idealized graph is the (1d) chain graph. Let $x_i = (i - 1/2)/n$ for $i = 1, \dots, n$. (We abandon our usual convention of capitalizing design points to emphasize that we are dealing with a fixed design). Throughout this section, take P to be the uniform distribution on $[0, 1]$. Let $\bar{G}_n = (\{1, \dots, n\}, E)$ be the chain graph, meaning $E = \{(i, i+1) : i = 1, 2, \dots, n-1\}$, with $\bar{L}_n$ the Laplacian of the chain, meaning

$$(\bar{L}_n u)_i = \begin{cases} \big(u_i - u_{i+1} + u_i - u_{i-1}\big), & \text{if } i = 2, \dots, n-1 \\ u_1 - u_2, & \text{if } i = 1, \\ u_n - u_{n-1}, & \text{if } i = n. \end{cases} \tag{5.25}$$

We will show that the equivalent kernel $\bar{g}_n$ of Laplacian smoothing using $\bar{L}_n$ is close to the solution of a differential equation involving the unweighted Laplacian operator Δ on $[0, 1]$. We define $\bar{g}$ to be the solution to

$$\frac{1}{n}(I + n^2\rho\bar{L}_n)g = \delta_i, \quad \text{in } \mathbb{R}^n; \tag{5.26}$$

the pre-factor of n^2 puts the eigenvalues of $\bar{L}_n$ on the same scale as those of Δ. Taking $x = x_i$, let g be the solution to

$$(I + \rho\Delta)g = \delta_x, \quad \frac{d}{dx}g(0) = \frac{d}{dx}g(1) = 0. \tag{5.27}$$

Recall that the eigenpairs of Δ are $\phi_1(x) = 1, \lambda_1 = 0$, and then $\phi_k(x) = \sqrt{2} \cdot \cos(k\pi x), \lambda_k = \pi^2 k^2$ for $k \geq 2$. Let $g_n := \sum_{k=1}^n \langle g, \phi_k\rangle_P \phi_k$ be the projection of the equivalent kernel g onto the span of $\phi_1, \dots, \phi_n$ in $L^2([0, 1])$.

Proposition 16. *Let $\bar{g}$ be the solution to (5.26), let g be the solution to (5.27), and let $g_n = \sum_{k=1}^n \langle \phi_k, g\rangle_P \phi_k$. Then there exists a universal constant C such that*

$$\|\bar{L}_n^{-1}(\bar{g} - g_n)\|_n^2 + \rho\|\bar{g} - g_n\|_n^2 + \rho^2\langle \bar{L}_n(\bar{g} - g_n), \bar{g} - g_n\rangle_n \leq \frac{C}{n}. \tag{5.28}$$

Proposition (16) should be interpreted as saying that $\bar{g}$ and g_n are close in each of three norms: the $L^2(P_n)$ norm, the first-order graph Sobolev norm, and the first-order dual graph Sobolev norm. The strength of the estimate differs, in terms of the dependence on ρ, depending on the norm under consideration. We also note that (5.27) is the same as setting $s = 1$ in (5.20), except for the presence of ρ in the former equation (and ignoring boundary conditions). As we have already discussed, ρ influences the bandwidth but not the shape of the equivalent kernel. Thus Proposition 16 shows that away from boundary points, the equivalent kernel of Laplacian smoothing over the 1d chain has a similar shape to the equivalent kernel of a first-order smoothing spline (i.e. the Laplace density function), but only after the latter has been passed through a filter which preserves low-frequency components.

Idealized Graph 2: Tensor product of ring and complete graph. To discuss our second example, we begin by recalling the definition of the tensor product of two graphs.

Definition 5.3.1 (Tensor product of graphs)**.** For two weighted graphs $G = (\{1, \dots, M\}, W_G)$ and $H = (\{1, \dots, N\}, W_H)$, the *tensor product* $G \times H$ has vertex set $\big\{(i, j) : i \in \{1, \dots, M\}, j \in \{1, \dots, N\}\big\}$. In $G \times H$, the vertices (i, j) and (i', j') have an edge of weight $(W_G)_{ii'} \times (W_H)_{jj'}$.

Now we introduce our second idealized graph. For $0 < \varepsilon < 1$, let $M := 1/\varepsilon$ and $N := n\varepsilon$ be integers. Let R_M be the ring graph on $\{1, \dots, M\}$, i.e. the graph with edges $(1, 2), (2, 3), \dots, (M-1, M), (M, 1)$. Let K_N be the complete graph on $\{1, \dots, N\}$, with edges (i, j) for all $1 \le i \le j \le N$. The graph we will consider is $\widetilde{G}_{n,\varepsilon} := R_M \times K_N$. Intuitively, this graph is more similar to a neighborhood graph than is the chain, because microscopically—meaning at scale on the order of ε—it resembles the complete graph.

Let $\widetilde{L}_{n,\varepsilon} : \mathbb{R}^{M \times N} \to \mathbb{R}^{M \times N}$ be the Laplacian associated with the graph $\widetilde{G}_{n,\varepsilon}$. Take $\widehat{g} \in \mathbb{R}^{M \times N}$ to be the equivalent kernel of Laplacian smoothing on $\widetilde{G}_{n,\varepsilon}$, meaning the solution to

$$\frac{1}{n}\Big(I + \rho \frac{M^2}{N} \widetilde{L}_{n,\varepsilon}\Big) u = \delta_{ij}, \quad \text{in } \mathbb{R}^{M \times N}, \tag{5.29}$$

where $(\delta_{ij})_{k\ell} = \mathbf{1}\{i = k, j = \ell\} \in \mathbb{R}^{M \times N}$ is the Kronecker delta.

We will again compare the equivalent kernel of Laplacian smoothing to the solution g of a differential equation; in this case, the differential equation

$$(I + \rho\Delta) g = \delta_x, \quad g(0) = g(1), \quad \frac{d}{dx} g(0) = \frac{d}{dx} g(1). \tag{5.30}$$

where $x = x_i$. The eigenvalue and eigenvector pairs (λ_k, ψ_k) of (5.30) can be characterized as follows: for even $k \in \mathbb{N}$, $\lambda_k = \pi^2 k^2$ and $\psi_k(x) = \sqrt{2} \cdot \cos(2k\pi x)$, and for odd $k \in \mathbb{N}$, $\lambda_k = \pi^2 (k-1)^2$ and $\psi_k(x) = \sqrt{2} \cdot \sin(2(k-1)\pi x)$.

In Proposition 17, we show that $\widehat{g} = \widehat{g}_M + \widehat{g}_\perp$ is the sum of two terms. The first term $\widehat{g}_M$ is close to the $g_M = \sum_{k=1}^M \langle g, \phi_k \rangle_P \phi_k$, similar to Proposition 16. The second term $\widehat{g}_\perp$ is localized around point (i, j). In the following, let $\langle\!\langle g, h \rangle\!\rangle_n := n^{-1} \sum_{i=1}^M \sum_{j=1}^N g_{ij} h_{ij}$, and let $|\!|\!| g |\!|\!|_n^2 = \langle\!\langle g, g \rangle\!\rangle_n$.

Proposition 17. *Let $\widehat{g}$ be the solution to (5.29), and let g be the solution to (5.30). Then $\widehat{g} = \widehat{g}_M + \widehat{g}_\perp$, where*

$$\Big|\!\Big|\!\Big| \widetilde{L}_{n,\varepsilon}^{-1} (\widehat{g}_M - g_M) \Big|\!\Big|\!\Big|_n^2 + \rho \, |\!|\!| \widehat{g}_M - g_M |\!|\!|_n^2 + \rho^2 \langle\!\langle \widetilde{L}_{n,\varepsilon} (\widehat{g}_M - g_M), \widehat{g}_M - g_M \rangle\!\rangle_n \le \frac{C}{M}, \tag{5.31}$$

and

$$\widehat{g}^\perp = \frac{n}{(1 + 2\rho M^2)} \cdot \delta_i (\delta_j - 1/\sqrt{N})^\top. \tag{5.32}$$

The difference between $\widetilde{G}_{n,\varepsilon}$ and the chain $\bar{G}_n$ is the role played by ε. Proposition 17 reveals that changing ε trades off two kinds of smoothness in the resulting equivalent kernel $\widehat{g}$, as we now comment on.

- The first part of Proposition 17 is similar in spirit to Proposition (5.28). It says that $\widehat{g}_M$ lies close to g_M, which is the projection of the solution g to (5.27) onto the span of $\phi_1, \dots, \phi_M$ in $L^2(P)$. There is one importance difference, however. Unlike in Proposition 17, the continuum function g_M now depends on ε. As ε grows, M shrinks, and the continuum function g_M depends only on lower frequency functions.
- The second part of Proposition 17 shows that, as opposed to the chain $\bar{G}_n$, the equivalent kernel of $\widetilde{G}_{n,\varepsilon}$ has an additional component. This component $\widehat{g}^{\perp}$ is itself comprised of two separate terms: the first is the Kronecker delta $\delta_i \delta_j = \delta_{ij}$, and the second is the locally supported function $-\delta_i/\sqrt{N}$. These functions are scaled by a pre-factor of $n/(1 + 2\rho M^2) = n/(1 + 2\rho\varepsilon^{-2}) \approx 2n\varepsilon^2/\rho$, with the last approximation being accurate when ε^2 is much smaller than ρ.
- Thus, qualitatively speaking, we see that ε effects a tradeoff between different aspects of the equivalent kernel $\widehat{g}$. As ε grows, one component of the equivalent kernel, $\widehat{g}_M$, becomes smoother. On the other hand, the other component $\widehat{g}_{\perp}$ has a spike with height growing quadratically in ε.

While $\widetilde{G}_{n,\varepsilon}$ is not the neighborhood graph, previewing things to come, we shall see that the radius parameter ε of a neighborhood graph effects a similar tradeoff, between smoothness of the equivalent kernel and spikiness around x_{ij}.

5.3.2 Bandwidth of equivalent kernel

Propositions 16 and 17 show that by understanding the structure of a fundamental solution to (5.22), we can understand the structure of the equivalent kernel of Laplacian smoothing. In one dimension ($\mathcal{X} = [0, 1]$) equation (5.22) is a special case of the more general differential equation

$$g - \frac{\rho}{p}\{w'g' + wg''\} = \delta_x, \tag{5.33}$$

for $w : [0, 1] \to \mathbb{R}$ a twice-differentiable weight function, and $x \in (0, 1)$.

Wang et al. [2013] characterize the shape and bandwidth of the fundamental solution J to (5.33); rewriting their equation (8) in our notation,

$$J(t) = \rho^{-1/2} \varrho(t) Q'_\rho(t) \kappa(|Q_\rho(x) - Q_\rho(t)|/\rho^{1/2}),$$

where κ is the solution to (5.20) with $s = 1$—i.e. the Laplace density function—the function

$$Q_\rho(t) := \int_0^t \left(\frac{p(s)}{w(s)}\right)^{1/2} \left(1 + O(\rho^{1/2})\right) ds,$$

is an increasing function of t, and $\varrho(t)$ satisfies $\sup_t |\varrho(t)| = 1 + O(\rho^{1/2})$.

Taking $w = p^2$ in (5.33), we recover (5.22). In the limit as $\rho \to 0$, the fundamental solution g to (5.22) thus satisfies the following asymptotic equality.

Corollary 1. *Let g be the solution to (5.22). Suppose P has a twice differentiable density p bounded away from 0 and ∞ on $[0, 1]$. Then, for any $x \in (0, 1)$, and any $z \in \mathbb{R}$,*

$$\lim_{\rho \to 0} \rho^{1/2} p(x)^{1/2} \cdot g(x + \rho^{1/2} p(x)^{1/2} z) = \kappa(z).$$

Thus we see that the fundamental solution g to (5.22) corresponds asymptotically to a kernel with bandwidth $h = \rho^{1/2} p(x)^{1/2}$, and shape given by that of the Laplace density function. A few remarks:

- We note that the bandwidth h is proportional to $p(x)^{1/2}$, and will thus be smaller in low-density regions. This is in contrast to the equivalent kernel of a first-order smoothing spline, where the bandwidth is

instead proportional to $p(x)^{-1/2}$, and so is larger in low-density regions. Which is the better choice? It depends on the nature of the regression function f_0. Standard derivations establish that asymptotically, the optimal bandwidth for kernel smoothing should be proportional to $(p(x))^{-1/3}(f_0'(x))^{-2/3}$. Thus if $f_0'(x) = p(x)^{-5/4}$, Laplacian smoothing is preferable, whereas if $f_0'(x) = p(x)^{1/2}$ then the smoothing spline makes more sense.

- The shape of the equivalent kernel is the same as the equivalent kernel for smoothing splines. This is not necessarily the case when p violates the regularity conditions in Corollary 1; for instance, if p is a piecewise uniform distribution supported on multiple connected components of $[0, 1]$.
- The distance $|Q_\rho(x) - Q_\rho(t)|$ depends on the average of the inverse square-root density $p(s)^{-1/2}$ over all $x \le s \le t$. It distorts the uniform metric $|x - t|$ by "pulling" points x and t farther apart if $p(s)$ is on average small over $s \in [x, t]$. As $t \to x$, we have

$$\lim_{t \to x} \frac{|Q_\rho(x) - Q_\rho(t)|}{|t - x|} = \frac{1}{(p(x))^{1/2}}.$$

5.3.3 Shape of equivalent kernel

Finally, in this section we turn to the fundamental solution κ of the PDE

$$(I + \Delta)\kappa = \delta_0. \tag{5.34}$$

Our analyses in the previous two sections suggest that the solution to (5.34) partly determines the shape of the equivalent kernel of Laplacian smoothing. (Although to be clear, this is only a suggestion, since we analyze only a pair of special geometric graphs in the univariate setting.)

When $d = 1$, we have already seen that $\kappa(x) = 1/2 \exp(-|x|)$ is the Laplace density function. It follows that $\kappa \in H^1(\mathbb{R})$. More generally, we have that the fundamental solution to (5.34) is a *Bessel potential*, given by

$$\kappa(x) = \frac{1}{2^{d/2}} \int_0^\infty \frac{e^{-t - \frac{\|x\|^2}{4t}}}{t^{d/2}} \, dt, \quad \text{for } x \in \mathbb{R}^d.$$

(See Chapter 4.3 of Evans [2010].) In particular, as $\|x\| \to 0$

$$\kappa(x) = (1 + o(1)) \cdot \begin{cases} \dfrac{1}{2^{d-1}\pi^{d/2}} \ln \dfrac{1}{\|x\|}, & \text{if } d = 2, \\ \dfrac{\Gamma\left(\frac{d-2}{2}\right)}{4\pi} \dfrac{1}{\|x\|^{d-2}}, & \text{if } d \ge 3. \end{cases}$$

Thus κ has a singularity at the origin for all $d \ge 2$. Moreover, $\kappa \in H^1(\mathbb{R}^d)$ only if $d < 2$, and $\kappa \in L^2(\mathbb{R}^d)$ only if $d < 4$. This is easily seen by the fact that the Fourier transform of κ is

$$\big(\mathcal{F}(\kappa)\big)(y) = \frac{1}{1 + \|y\|^2}, \quad y \in \mathbb{R}^d.$$

Belkin et al. [2019] show that kernel smoothing of random data using a singular kernel can result in consistent and even optimal estimators in terms of $L^2(P)$ error. However their results require that the smoothing kernel itself belong to $L^2(\mathbb{R}^d)$. This gives another perspective on the failure of Laplacian smoothing when $d \ge 4$; it is equivalent to smoothing with a kernel that is not sufficiently regular.

5.3.4 Predictions based on theoretical findings

Based on the findings of Sections 5.3.1-5.3.3, we make a series of qualitative predictions regarding the equivalent kernel $\widehat{g}$ of Laplacian smoothing at a given design point X_i.

1. The equivalent kernel $\widehat{g}$ is close to the sum of two terms. The first term is approximately equal to the fundamental solution g to the PDE (5.22), after g has been passed through a spectral filter. The second term includes a spike at X_i.
2. The graph radius parameter ε trades off two types of smoothness, corresponding to the aforementioned two terms. As ε is taken larger, the aforementioned spectral filter more aggressively filters out the high-frequency components of g, but the aforementioned spike is larger.
3. The effective bandwidth h of $\widehat{g}$ depends on the regularization parameter ρ; as the regularization parameter ρ grows, so does h. The bandwidth also depends on the design density p evaluated at X_i; the larger $p(X_i)$, the larger the bandwidth.
4. As the dimension d grows, the shape of the continuum limit of the equivalent kernel becomes less regular.

Our theory does not conclusively establish that these predictions hold for the neighborhood graph (which is also why we do not summarize them in a more precise fashion). However, they do fit very nicely with some of our previous theoretical developments.

- Prediction 3 gives a complementary view on the density adaptivity of Laplacian smoothing, explored in Section 5.2.
- Predictions 2 and 4 explain why Laplacian smoothing is not optimal when the dimension is sufficiently large ($d \geq 4$). It corresponds to smoothing using a kernel with minimal regularity properties, and which additionally has a spike at X_i.
- Prediction 2 also explains the finding in Theorem 21 that Laplacian smoothing can be optimal at an unlabeled data point, when ε is taken to be sufficiently large. Taking ε larger makes the equivalent kernel more regular (i.e. it can be extended a function which has a smaller norm in $L^2(\mathbb{R}^d)$). It also increases the height of the spike, which degrades the quality of the estimate at labeled data, but has no effect at an unlabeled data point.

We now perform a series of experiments, which offer empirical validation of Predictions 1-3.

5.3.5 Experiments

We now empirically examine the equivalent kernel of Laplacian smoothing with a neighborhood graph, in the univariate setting $d = 1$. We consider two densities, the uniform density over $[-1, 1]$, and the truncated Gaussian mixture density given in (5.17). In both cases, we sample $n = 500$ points, form the neighborhood graph using a truncated Gaussian graph kernel $\eta(t) = \mathbf{1}|t| \leq 1 \cdot \exp(-s^2/2)$ and various graph radii ε. We then solve for the Laplacian smoothing smoother matrix $S = (I + \rho L_{n,\varepsilon})^{-1}$, and examine the ith row S_i for several different X_i.

Uniform density. We begin by examining the equivalent kernel of Laplacian smoothing when P is uniform. Figure 5.5 shows the equivalent kernel $\widehat{g}$, at each of $x \approx -.5$, $x \approx 0$, and $x \approx .5$, and for three different values of graph radius ε and smoothing parameter ρ.[5] We see the following:

- The bandwidth of the equivalent kernel is controlled by the regularization parameter ρ. As ρ decreases, the degrees of freedom of S increase, and the effective bandwidth becomes smaller.
- The equivalent kernel always has a spike at x. As ε increases, the height of this spike increases.
- Ignoring the spike, as in the 2nd row of Figure 5.5, we see that the equivalent kernel looks more regular as the graph radius ε increases.

These observations all accord with the predictions made in Section 5.3.4.

[5] To be precise, it shows the row S_i for the design point X_i closest to $-.5, 0$ and $.5$

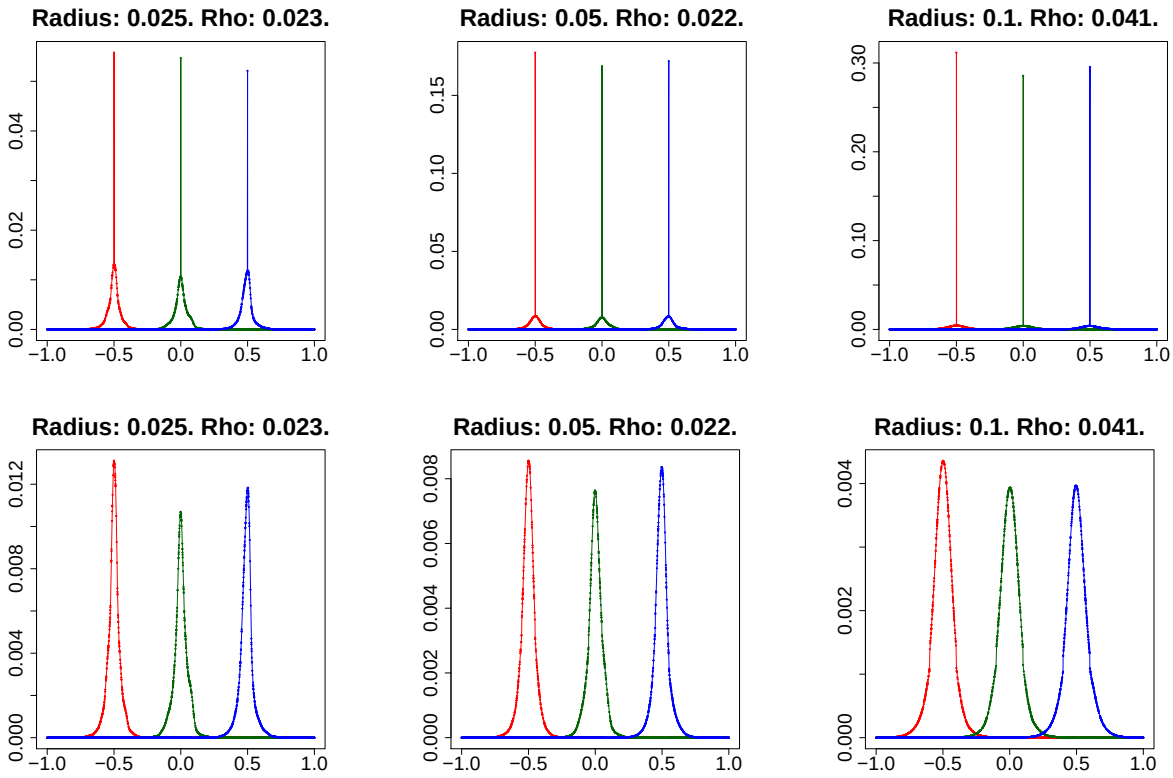


Figure 5.5: Top row: equivalent kernel of Laplacian smoothing with uniform design distribution. Bottom row: same as the top row, but ignoring the influence of S_{ii}.

Gaussian mixture density. Our second experiment examines the equivalent kernel of Laplacian smoothing under the same setup, but where the design density is the truncated Gaussian mixture distribution given in (5.17). Figure 5.6 shows the equivalent kernel at the same values of x as in Figure 5.5. Each of the observations from the first experiment also apply to this experiment. Additionally, we see empirically that the bandwidth of the equivalent kernel is smaller at points x where the density is smaller. Again, this is in accordance with our predictions.

References

Yasin Abbasi-Yadkori. Fast mixing random walks and regularity of incompressible vector fields. *arXiv preprint arXiv:1611.09252*, 2016.

Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, and Alan Malek. Hit-and-Run for Sampling and Planning in Non-Convex Spaces. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 888–895, 2017.

Zeyuan Allen-Zhu and Yuanzhi Li. Faster principal component regression and stable matrix chebyshev approximation. In *International Conference on Machine Learning*, pages 107–115. PMLR, 2017.

Reid Andersen and Yuval Peres. Finding sparse cuts locally using evolving sets. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 235–244, New York, NY, USA, 2009. ACM.

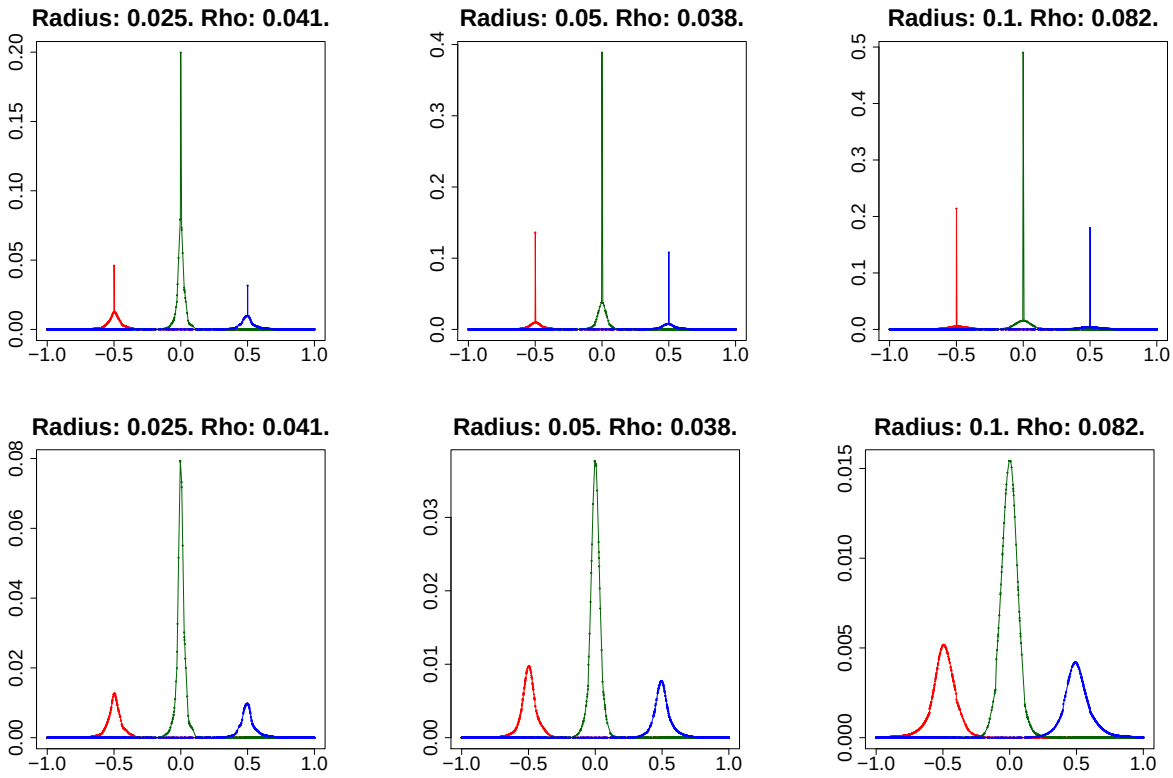


Figure 5.6: Top row: equivalent kernel of Laplacian smoothing with design distribution. Bottom row: same as the top row, but ignoring the influence of S_{ii}.

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.

Reid Andersen, David F Gleich, and Vahab Mirrokni. Overlapping clusters for distributed computation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 273–282. ACM, 2012.

Ery Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *arXiv preprint arXiv:0909.2353*, 2009.

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.

Thierry Aubin. *Nonlinear analysis on manifolds. Monge-Ampere equations*, volume 252. Springer Science & Business Media, 2012.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727 – 749, 2018.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Annals of Statistics*, 47(4):1893–1927, 2019.

Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011.

Sivaraman Balakrishnan, Alesandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *International Conference on Artificial Intelligence and Statistics*, volume 22, 2012.

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, 2013a. Curran Associates, Inc.

Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, volume 26, 2013b.

Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

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.

Mikhail Belkin. *Problems of Learning on Manifolds*. PhD thesis, University of Chicago, 2003.

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pages 1348–1361. PMLR, 2018.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Mikhail Belkin and Partha Niyogi. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Mikhail Belkin, Qichao Que, Yusu Wang, and Xueyuan Zhou. Toward understanding complex spaces: Graph laplacians on manifolds with singularities and boundaries. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 36.1–36.26, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems*, volume 54, pages 177–186. Institute of Mathematical Statistics, 2007.

Lucien Birgé. Model selection for density estimation with l2-loss. *arXiv preprint arXiv:0808.1416*, 2008.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure based regularization. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

Xavier Bresson, Thomas Laurent, David Uminsky, and James Brecht. Convergence and energy landscape for cheeger cut clustering. In *Advances in Neural Information Processing Systems 25*, pages 1385–1393, 2012.

Dmitri Burago, Sergei Ivanov, and Yaroslav Kurylev. A graph discretization of the Laplace-Beltrami operator. *Journal of Spectral Theory*, 4(4):675–714, 2014.

Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs. *arXiv preprint arXiv:1910.13476*, 2019.

Jeff Calder, Nicolas Garcia Trillos, and Marta Lewicka. Lipschitz regularity of graph laplacians on random data clouds. *arXiv preprint arXiv:2007.06679*, 2020a.

Jeff Calder, Dejan Slepčev, and Matthew Thorpe. Rates of convergence for laplacian semi-supervised learning with low labeling rates. *arXiv preprint arXiv:2006.02765*, 2020b.

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.

Xiuyuan Cheng and Nan Wu. Eigen-convergence of gaussian kernelized graph laplacian by manifold heat interpolation. *arXiv preprint arXiv:2101.09875*, 2021.

Fan RK Chung. *Spectral graph theory*. American Mathematical Soc., 1997.

Aaron Clauset, Cristopher Moore, and MEJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–102, 2008.

R. M. Dudley. Distances of probability measures and random variables. *Ann. Math. Statist.*, 39(5):1563–1572, 10 1968.

Matthew M Dunlop, Dejan Slepčev, Andrew M Stuart, and Matthew Thorpe. Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Applied and Computational Harmonic Analysis*, 49(2):655–697, 2020.

David B Dunson, Hau-Tieng Wu, and Nan Wu. Spectral convergence of graph laplacian and heat kernel reconstruction in l^∞ from random samples, 2020.

Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

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.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

Roy Frostig, Cameron Musco, Christopher Musco, and Aaron Sidford. Principal component projection without principal component analysis. In *International Conference on Machine Learning*, pages 2349–2357. PMLR, 2016.

Nicolás García Trillos and Ryan W. Murray. A maximum principle argument for the uniform convergence of graph Laplacian regressors. *SIAM Journal on Mathematics of Data Science*, 2(3):705–739, 2020.

Nicolas García Trillos and Dejan Slepcev. On the rate of convergence of empirical measures in infinity-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

Nicolás García Trillos and Dejan Slepčev. On the rate of convergence of empirical measures in infinity-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

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

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

Nicolás García Trillos, Dejan Slepčev, James Von Brecht, Thomas Laurent, and Xavier Bresson. Consistency of cheeger and ratio graph cuts. *Journal of Machine Learning Research*, 17(1):6268–6313, 2016.

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepcev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, 20:1–61, 2019a.

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

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph laplacian on random geometric graphs toward the laplace–beltrami operator. *Foundations of Computational Mathematics*, 20(4):827–887, 2020.

Shayan Oveis Gharan and Luca Trevisan. Approximating the expansion profile and almost optimal local graph clustering. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 187–196. IEEE, 2012.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

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

David F Gleich and C Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 597–605. ACM, 2012.

Alden Green, Sivaraman Balakrishnan, and Ryan Tibshirani. Minimax optimal regression over sobolev spaces via laplacian regularization on neighborhood graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 2602–2610. PMLR, 2021.

Peter J. Green and Bernard W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC Press, 1993.

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.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2006.

John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

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.

Harrie Hendriks. Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions. *The Annals of Statistics*, pages 832–849, 1990.

Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23(2):169–192, 1952.

Franca Hoffmann, Bamdad Hosseini, Assad A Oberai, and Andrew M Stuart. Spectral analysis of weighted laplacians arising in data clustering. *arXiv preprint arXiv:1909.06389*, 2019.

Lars Hörmander. *The analysis of linear partial differential operators III: Pseudo-differential operators*. Springer Science & Business Media, 2007.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, volume 29, 2016.

Yuri I. Ingster. Minimax nonparametric detection of signals in white Gaussian noise. *Problems in Information Transmission*, 18:130–140, 1982.

Yuri 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.

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

Yuri I. Ingster and Irina A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media, 2012.

Heinrich Jiang. Density level set estimation on manifolds with DBSCAN. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML'17*, pages 1684–1693, 2017.

Yujia Jin and Aaron Sidford. Principal component projection and regression in nearly linear time through asymmetric svrg. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. *Unpublished manuscript*, 2011.

Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*, 2020.

Alisa Kirichenko and Harry van Zanten. Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electronic Journal of Statistics*, 11(1):891–915, 2017.

Alisa Kirichenko, Harry van Zanten, et al. Minimax lower bounds for function estimation on graphs. *Electronic Journal of Statistics*, 12(1):651–666, 2018.

Vladimir Koltchinskii and Evarist Gine. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 02 2000.

Risi Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *International Conference on Machine Learning*, volume 19, 2002.

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.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Ann B. Lee, Rafael Izbicki, et al. A spectral series approach to high-dimensional nonparametric regression. *Electronic Journal of Statistics*, 10(1):423–463, 2016.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015.

Giovanni Leoni. *A first Course in Sobolev Spaces*. American Mathematical Society, 2017.

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.

Tianxi Li, Lihua Lei, Sharmodeep Bhattacharyya, Purnamrita Sarkar, Peter J Bickel, and Elizaveta Levina. Hierarchical community detection by recursive partitioning. *arXiv preprint arXiv:1810.01509*, 2018.

Anna V Little, Mauro Maggioni, and James M Murphy. Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *Journal of Machine Learning Research*, 21(6):1–66, 2020.

Meimei Liu, Zuofeng Shang, and Guang Cheng. Sharp theoretical analysis for nonparametric testing under random projection. In *Conference on Learning Theory*, volume 32, 2019.

László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of the 31st annual symposium on foundations of computer science (FOCS)*, pages 346–354. IEEE, 1990.

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.

Frank McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 529–537, 2001.

Ravi Montenegro. *Faster mixing by isoperimetric inequalities*. PhD thesis, Yale University, 2002.

Ben Morris and Yuval Peres. Evolving sets, mixing and heat kernel bounds. *Probability Theory and Related Fields*, 133(2):245–266, 2005.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. In *Neural Information Processing Systems*, volume 19, 2009.

Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(1):1229–1250, 2013.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

Bruno Pelletier and Pierre Pudlo. Operator norm convergence of spectral clustering on level sets. *Journal of Machine Learning Research*, 12(12):385–416, 2011.

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

Christian Rieger and Barbara Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32(1):103, 2010.

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

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *Ann. Statist.*, 38(5):2678–2722, 10 2010.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915, 08 2011.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

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*, volume 29, 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*, volume 30, 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.

James Sharpnack and Aarti Singh. Identifying graph-structured activation patterns in networks. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

James Sharpnack, Akshay Krishnamurthy, and Aarti Singh. Near-optimal anomaly detection in graphs using Lovasz extended scan statistic. In *Advances in Neural Information Processing Systems*, volume 26, 2013a.

James Sharpnack, Aarti Singh, and Akshay Krishnamurthy. Detecting activations over graphs using spanning tree wavelet bases. In *International Conference on Artificial Intelligence and Statistics*, volume 16, 2013b.

James Sharpnack, Alessandro Rinaldo, and Aarti Singh. Detecting anomalous activity on networks with the graph Fourier scan statistic. *IEEE Transactions on Signal Processing*, 64(2):364–379, 2015.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

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

Zuoqiang Shi. Convergence of laplacian spectra from random samples. *arXiv preprint arXiv:1507.00151*, 2015.

B. W. Silverman. Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*, 12 (3):898 – 916, 1984.

Amit Singer and Hau-Tieng Wu. Spectral convergence of the connection laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 2017.

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

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

Alexander J. Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines*, pages 144–158. Springer, 2003.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

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.

Ingo Steinwart. Fully adaptive density-based clustering. *Ann. Statist.*, 43(5):2132–2167, 10 2015.

Ingo Steinwart, Bharath K Sriperumbudur, and Philipp Thomann. Adaptive clustering using kernel density estimators. *arXiv preprint arXiv:1708.05254*, 2017.

Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

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, 2008a.

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

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Nisheeth K. Vishnoi. Laplacian solvers and their algorithmic applications. *Foundations and Trends in Theoretical Computer Science*, 8(1-2):1–141, 2012.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 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.

Martin J Wainwright. *High-Dimensional Dtatistics: A Non-Asymptotic Biewpoint*. Cambridge University Press, 2019.

Daren Wang, Xinyang Lu, and Alessandro Rinaldo. Dbscan: Optimal rates for density-based cluster estimation. *Journal of machine learning research*, 2019.

Xiao Wang, Pang Du, and Jinglai Shen. Smoothing splines with varying smoothing parameter. *Biometrika*, 100(4):955–970, 2013.

Yu-Xiang Wang, James Sharpnack, Alexander J. Smola, and Ryan J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

Xiao-Ming Wu, Zhenguo Li, Anthony M. So, John Wright, and Shih fu Chang. Learning with partially absorbing random walks. In *Advances in Neural Information Processing Systems 25*, pages 3077–3085. Curran Associates, Inc., 2012.

Amber Yuan, Jeff Calder, and Braxton Osting. A continuum limit for the pagerank algorithm. *arXiv preprint arXiv:2001.08973*, 2020.

Dengyong Zhou, Jiayuan Huang, and Bernhard Scholkopf. Learning from labeled and unlabeled data on a directed graph. In *International Conference on Machine Learning*, volume 22, 2005.

Xueyuan Zhou and Nathan Srebro. Error analysis of laplacian eigenmaps for semi-supervised learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 901–908. JMLR Workshop and Conference Proceedings, 2011.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, volume 20, 2003.

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.

Appendix A

Chapter 2 Appendix

The proofs of our major theorems largely consist of (at most) three modular parts.

1. **Fixed graph results.** Results which hold with respect to an arbitrary graph G, and are stated with respect to functionals (i.e. normalized cut, conductance, and local spread) of G;
2. **Sample-to-population results.** For the specific choice $G = G_{n,r}$, w e relate the aforementioned functionals to their population analogues.
3. **Bounds on population functionals.** (In the case of density clustering only.) When the candidate cluster is a λ-density cluster, we bound the population functionals by a function of λ, as well as the other relevant parameters introduced in Section 2.2.

Appendices A.1-A.3 will correspond to each of these three parts. In Appendix A.4, we will combine these parts to prove the major theorems of our main text, Theorems 3 and 4, as well as our negative result, Theorem 5. In Appendix A.5 we derive upper bounds for the aPPR vector, and show that under certain conditions the PPR vector can perfectly separate two density clusters. Finally, in Appendix A.6 we give relevant details regarding our experiments.

A.1 Fixed graph results

In this appendix, we give all results that hold with respect to an arbitrary graph G. For the convenience of the reader, we begin by reviewing some notation from the main text, and also introduce some new notation.

Notation. The graph $G = (V, E)$ is an undirected and connected but otherwise arbitrary graph, defined over vertices $V = \{1, \ldots, n\}$ with $m = |E|$ total edges. The adjacency matrix of G is A, the degree matrix is D, and the lazy random walk matrix over G is $W = (I + D^{-1}A)/2$. If the lazy random walk originates at a node v, the distribution of the lazy random walk $q_v^{(t)} := q(v, t; G)$ after t steps is $q_v^{(t)} := e_v W^t$, with stationary distribution $\pi := \pi(G) := \lim_{t\to\infty} q_v^{(t)}$ with entries $\pi(u) = \deg(u; G)/\mathrm{vol}(u; G)$.

For a starting distribution s (by distribution we mean a vector with non-negative entries), the PPR vector $p_s = p(s, \alpha; G)$ is the solution of

$$p_s = \alpha s + (1 - \alpha) p_s W. \tag{A.1}$$

When $s = e_v$, we write $p_v := p_{e_v}$. It is easy to check that $p_s = \alpha \sum_{t=0}^{\infty} (1 - \alpha)^t q_s^{(t)}$. Note that s need not be a probability distribution (i.e. its entries need not sum to 1) to make sense of (A.1).

Given a distribution q (for instance, $q = q_v^{(t)}$ for $t \in \mathbb{N}$, $q = p_v$, or $q = \pi$) and $\beta \in (0,1)$, the β-sweep cut of q is

$$S_\beta(q) = \left\{ u : \frac{q(u)}{\deg(u;G)} > \beta \right\};$$

in the special case where $q = p_v$ we write $S_{\beta,v}$ for $S_\beta(p_v)$. The argument of $S_\beta(\cdot)$ will usually be clear from context, in which case we will drop it and simply write S_β. For $j = 1, \ldots, n$, let β_j be the smallest value of $\beta \in (0,1)$ such that the sweep cut S_{β_j} contains at least j vertices. For notational ease, we will write $S_j := S_{\beta_j}$, and $S_0 = \emptyset$.

We now introduce the *Lovasz-Simonovits curve* $h_q(\cdot) : [0, 2m] \to [0,1]$ to measure the extent to which a distribution q is mixed. To do so, we first define a piecewise linear function $q[\cdot] : [0, 2m] \to [0,1]$. Letting $q(S) := \sum_{u \in S} q(u)$, we take $q[\mathrm{vol}(S_j)] = q(S_j)$ for each sweep cut S_j, and then extend $q[\cdot]$ by piecewise linear interpolation to be defined everywhere on its domain. Then the mixedness of q is measured by

$$h_q(k) := q[k] - \frac{k}{2m}.$$

The Lovasz-Simonovits curve is a non-negative function, with $h_q(0) = h_q(2m) = 0$. The stationary distribution π is mixed, i.e. $h_\pi(k) = 0$ for all $k \in [0, 2m]$. Finally, both $q[\cdot]$ and $h_q(\cdot)$ are concave functions, which will be an important fact later on.

The conductance of V is abbreviated as $\Psi(G) := \Psi(V;G)$, and likewise for the local spread $s(G) := s(V;G)$. Finally, for convenience we introduce the following functionals:

$$d_{\max}(C;G) := \max_{u \in C} \deg(u;G), \quad d_{\min}(C;G) := \min_{u \in C} \deg(u;G)$$
$$d_{\max}(G) := d_{\max}(V;G), \qquad d_{\min}(G) := d_{\min}(V;G)$$

We note that $d_{\min}(G)^2 \le d_{\min}(G) \cdot n \le \mathrm{vol}(G) \le n \cdot d_{\max}(G)$, and that for any $S \subseteq V$, $|S| \cdot d_{\min}(G) \le \mathrm{vol}(S;G)$ (where $|S|$ is the cardinality of S.)

Organization. In the following sections we establish: (Section A.1.1) an upper bound on the misclassification error of PPR in terms of α and $\Phi(C;G)$ (Lemma 1), and an analogous result for aPPR (Corollary 2); (Section A.1.2) a uniform bound on the perturbations of the PPR vector, to be used later in the proof of Theorem 26 (consistency of PPR); (Section A.1.3) upper bounds on the mixedness of $q_v^{(t)}$ (as a function of t) and p_v (as a function of α), which will be helpful in the proofs of Proposition 1 and Theorem 5; (Section A.1.4) an upper bound on $\tau_\infty(G)$ in terms of $\Psi(G)$ and $s(G)$ (Proposition 1); and (Section A.1.5) an upper bound on the normalized cut $\Phi(\widehat{C};G)$ in terms of $\Phi(C;G)$, to be used later in the proof of Theorem 5 (negative example).

A.1.1 Misclassification error of clustering with PPR and aPPR

For a candidate cluster $C \subseteq V$, we use the tilde-notation $\widetilde{G} = G[C]$ to refer to the subgraph of G induced by C. Similarly we write $\widetilde{q}_v^{(t)} := q(v,t;\widetilde{G})$ for the t-step distribution of the lazy random walk over $\widetilde{G}$, $\widetilde{\pi} = \pi(G[C])$ for the stationary distribution of $\widetilde{q}_v^{(t)}$ (we will always assume $G[C]$ is connected), and $\widetilde{p}_v := p(v,\alpha;\widetilde{G})$ for the PPR vector over $\widetilde{G}$.

Proof of Lemma 1. As mentioned in the main text, Lemma 1 is equivalent, up to constants, to Lemma 3.4 in Zhu et al. [2013], and the proof of Lemma 1 proceeds along very similar lines to the proof of that lemma. In fact, we directly use the following three inequalities, derived in that work:

- **(c.f. Lemma 3.2 of Zhu et al. [2013])** For any seed node $v \in C$, the PPR vector is lower bounded,

$$\widetilde{p}_v(u) \ge \frac{3}{4}\big(1 - \alpha \cdot \tau_\infty(\widetilde{G})\big) \cdot \widetilde{\pi}(u), \quad \text{for every } u \in C. \tag{A.2}$$

- **(c.f. Corollary 3.3 of Zhu et al. [2013])** For any seed node $v \in C$, there exists a so-called leakage distribution $\ell = \ell(v)$ such that $\mathrm{supp}(\ell) \subseteq C$, $\|\ell\|_1 \leq 2\Phi(C;G)/\alpha$, and

$$p_v(u) \geq \widetilde{p}_v(u) - \widetilde{p}_\ell(u), \quad \text{for every } u \in C. \tag{A.3}$$

- **(c.f. Lemma 3.1 of Zhu et al. [2013])** There exists a set $C^g \subset C$ with $\mathrm{vol}(C^g;G) \geq \frac{1}{2}\mathrm{vol}(C;G)$ such that for any seed node $v \in C^g$, the following inequality holds

$$p_v(C^c) \leq 2\frac{\Phi(C;G)}{\alpha}. \tag{A.4}$$

We use (A.2)-(A.4) to separately upper bound $\mathrm{vol}(S_{\beta,v} \setminus C;G)$, $\mathrm{vol}(C^{\mathrm{int}} \setminus S_{\beta,v};G)$ and $\mathrm{vol}(C^{\mathrm{bdry}} \setminus S_{\beta,v};G)$; here $C^{\mathrm{int}} \cup C^{\mathrm{bdry}} = C$ is a partition of C, with

$$C^{\mathrm{int}} := \Big\{u \in C : \deg(u;\widetilde{G}) > \big(1 - \alpha \cdot \beta \cdot \mathrm{vol}(C;G)\big)\deg(u;G)\Big\},$$

consisting of those vertices $u \in C$ with sufficient large degree in $\widetilde{G}$.

First we upper bound $\mathrm{vol}(S_{\beta,v} \setminus C;G)$. Observe that for any $u \in S_{\beta,v} \setminus C$, $p_v(u) > \beta \cdot \deg(u;G)$. Summing up over all such vertices, from (A.4) we conclude that

$$\mathrm{vol}(S_{\beta,v} \setminus C;G) \leq \frac{p_v(C^c)}{\beta} \leq 2\frac{\Phi(C;G)}{\beta \cdot \alpha}. \tag{A.5}$$

Next we upper bound $\mathrm{vol}(C^{\mathrm{int}} \setminus S_{\beta,v};G)$. From (A.2) and (A.3) we see that

$$p_v(u) \geq \frac{3}{4}\big(1 - \alpha \cdot \tau_\infty(\widetilde{G})\big) \cdot \widetilde{\pi}(u) - \widetilde{p}_\ell(u) \quad \text{for all } u \in C.$$

If additionally $u \notin S_{\beta,v}$ then $p_v(u) \leq \beta \deg(u;G)$, and for all such $u \in C \setminus S_{\beta,v}$,

$$\frac{3}{4}\big(1 - \alpha \cdot \tau_\infty(\widetilde{G})\big) \cdot \widetilde{\pi}(u) - \beta \deg(u;G) \leq \widetilde{p}_\ell(u). \tag{A.6}$$

On the other hand, for any $u \in C^{\mathrm{int}}$ it holds that

$$\widetilde{\pi}(u) = \frac{\deg(u;\widetilde{G})}{\mathrm{vol}(\widetilde{G})} \geq \frac{\deg(u;\widetilde{G})}{\mathrm{vol}(G)} \geq \frac{\big(1 - \alpha\beta\mathrm{vol}(C;G)\big)\deg(u;G)}{\mathrm{vol}(C;G)};$$

by plugging this in to (A.6) we obtain

$$\left(\frac{3\big(1 - \alpha\beta\mathrm{vol}(C;G)\big) \cdot \big(1 - \alpha\tau_\infty(\widetilde{G})\big)}{4\mathrm{vol}(C;G)} - \beta\right) \cdot \deg(u;G) \leq \widetilde{p}_\ell(u), \quad \text{for all } u \in C^{\mathrm{int}} \setminus S_{\beta,v};$$

and summing over all such u gives

$$\left(\frac{3\big(1 - \alpha\beta\mathrm{vol}(C;G)\big) \cdot \big(1 - \alpha\tau_\infty(\widetilde{G})\big)}{4\mathrm{vol}(C;G)} - \beta\right) \cdot \mathrm{vol}\big(C^{\mathrm{int}} \setminus S_{\beta,v};G\big) \leq \widetilde{p}_\ell\big(C^{\mathrm{int}} \setminus S_{\beta,v}\big) \leq 2\frac{\Phi(C;G)}{\alpha}.$$

The upper bounds on α and β in (2.12) imply

$$\left(\frac{3\big(1 - \alpha\beta\mathrm{vol}(C;G)\big) \cdot \big(1 - \alpha\tau_\infty(\widetilde{G})\big)}{4\mathrm{vol}(C;G)} - \beta\right) \geq \frac{2}{3}\beta,$$

and we conclude that

$$\mathrm{vol}(C^{\mathrm{int}} \setminus S_{\beta,v}; G) \leq \frac{3\Phi(C;G)}{\alpha\beta}. \tag{A.7}$$

Finally, we upper bound $\mathrm{vol}(C^{\mathrm{bdry}} \setminus S_{\beta,v}; G)$. Indeed, for any $u \in C^{\mathrm{bdry}}$,

$$\frac{1}{\mathrm{vol}(C;G)} \sum_{w \notin C} \mathbf{1}((u,w) \in E) \geq \alpha \cdot \beta \cdot \deg(u;G)$$

and summing over all such vertices yields

$$\mathrm{vol}(C^{\mathrm{bdry}}; G) \leq \frac{1}{\alpha\beta \mathrm{vol}(C;G)} \sum_{\substack{u \in C^{\mathrm{bdry}} \\ w \notin C}} \mathbf{1}((u,w) \in E) \leq \frac{\Phi(C;G)}{\alpha \cdot \beta}. \tag{A.8}$$

The claim follows upon summing the upper bounds in (A.5), (A.7) and (A.8).

If the cluster estimate $\widehat{C}$ is instead obtained by sweep cutting the aPPR vector $p_v^{(\varepsilon)}$, a similar upper bound on $\mathrm{vol}(\widehat{C} \vartriangle C)$ holds, provided that ε is sufficiently small.

Corollary 2. *For a set $C \subseteq V$, suppose that α, β satisfy (2.12), and additionally that*

$$\varepsilon \leq \frac{1}{25\mathrm{vol}(C;G)}. \tag{A.9}$$

Then there exists a set $C^g \subset C$ with $\mathrm{vol}(C^g; G) \geq \frac{1}{2}\mathrm{vol}(C^g; G)$ such that for any $v \in C^g$, the sweep cut $S_{\beta,v}$ of the aPPR vector $p_v^{(\varepsilon)}$ satisfies

$$\mathrm{vol}(S_{\beta,v} \vartriangle C; G) \leq 6\frac{\Phi(C;G)}{\alpha\beta}. \tag{A.10}$$

Proof of Corollary 2. Recall that the upper bound (2.13) on $\mathrm{vol}(\widehat{C} \vartriangle C; G)$ comes from combining the upper bounds on $\mathrm{vol}(\widehat{C} \setminus C; G)$, $\mathrm{vol}(C^{\mathrm{int}} \setminus \widehat{C}; G)$ and $\mathrm{vol}(C^{\mathrm{bdry}} \setminus \widehat{C}; G)$ in (A.5), (A.7) and (A.8). From the upper bound $p_v^{(\varepsilon)}(u) \leq p_v(u)$ for all $u \in V$, it is clear that both (A.5) and (A.8) continue to hold when the aPPR vector is used instead of the PPR vector.

It remains only to establish an upper bound on $\mathrm{vol}(C^{\mathrm{int}} \setminus \widehat{C}; G)$. For any $u \in C \setminus S_{\beta,v}$, from inequality (A.3) and the lower bound $p_v^{(\varepsilon)}(u) \geq p_v(u) - \varepsilon \deg(u;G)$ in (2.2) we deduce that

$$\frac{3}{4}\big(1 - \alpha \cdot \tau_\infty(\widetilde{G})\big) \cdot \widetilde{\pi}(u) - (\beta + \varepsilon)\deg(u;G) \leq \widetilde{p}_\ell(u). \tag{A.11}$$

Following the same steps as used in the proof of Lemma 1 yields the following inequality:

$$\left(\frac{3(1 - \alpha\beta\mathrm{vol}(C;G)) \cdot \big(1 - \alpha\tau_\infty(\widetilde{G})\big)}{4\mathrm{vol}(C;G)} - \beta\right) \cdot \mathrm{vol}\big(C^{\mathrm{int}} \setminus S_{\beta,v}; G\big) \leq \widetilde{p}_\ell\big(C^{\mathrm{int}} \setminus S_{\beta,v}\big) \leq 2\frac{\Phi(C;G)}{\alpha}.$$

The upper bounds on α, β in (2.12), and on ε in (A.9), imply that

$$\left(\frac{3(1 - \alpha\beta\mathrm{vol}(C;G)) \cdot \big(1 - \alpha\tau_\infty(\widetilde{G})\big)}{4\mathrm{vol}(C;G)} - \beta\right) \geq \frac{2}{3}\beta,$$

and we conclude that

$$\mathrm{vol}(C^{\mathrm{int}} \setminus S_{\beta,v}; G) \leq \frac{3\Phi(C;G)}{\alpha\beta}. \tag{A.12}$$

Summing the right hand sides of (A.3), (A.8), and (A.12) yields the claim.

A.1.2 Uniform bounds on PPR

As mentioned in our main text, in order to prove Theorem 26, we require a uniform bound on the PPR vector. Actually, we require two such bounds: for a candidate cluster $C \subseteq V$ and an alternative cluster $C' \subseteq V$, we require a lower bound on $p_v(u)$ for all $u \in C$, and an upper bound on $p_v(u')$ for all $u' \in C'$. In Lemma 7 we establish an upper bound that holds for all vertices u in the interior C_o of C, and a lower bound holds for all vertices u' in the interior of C'_o of C'; here

$$C_o = \Big\{u \in C : \deg(u, \widetilde{G}) = \deg(u; G)\Big\}, \quad \text{and} \quad C'_o = \Big\{u \in C' : \deg(u, G[C']) = \deg(u; G)\Big\},$$

and we remind the reader that $\widetilde{G} = G[C]$.

Lemma 7. *Let C and C' be disjoint subsets of V, and suppose that*

$$\alpha \leq \frac{1}{2\tau_\infty(\widetilde{G})}.$$

Then there exists a set $C^g \subseteq C$ with $\mathrm{vol}(C^g; G) \geq \mathrm{vol}(C; G)/2$ such that for any $v \in C^g$,

$$p_v(u) \geq \frac{3}{8}\widetilde{\pi}(u) - \frac{2\Phi(C;G)}{d_{\min}(\widetilde{G}) \cdot \alpha} \quad \textit{for all } u \in C_o \tag{A.13}$$

and

$$p_v(u') \leq \frac{2\Phi(C;G)}{d_{\min}(C';G) \cdot \alpha} \quad \textit{for all } u \in C'_o. \tag{A.14}$$

"Leakage" and "soakage" vectors. To prove Lemma 7, we will make use of the following explicit representation of the *leakage* distribution ℓ from (A.4), as well as an analogously defined *soakage* distribution s:

$$\begin{aligned} \ell^{(t)} &:= e_v(W\widetilde{I})^t(I - D^{-1}\widetilde{D}), & \ell &= \sum_{t=0}^{\infty}(1-\alpha)^t\ell^{(t)} \\ s^{(t)} &:= e_v(W\widetilde{I})^t W(I - \widetilde{I}), & s &= \sum_{t=0}^{\infty}(1-\alpha)^t s^{(t)}. \end{aligned} \tag{A.15}$$

In the above, $\widetilde{I} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $I_{uu} = 1$ if $u \in C$ and 0 otherwise, and $\widetilde{D}$ is the diagonal matrix with $\widetilde{D}_{uu} = \deg(u; \widetilde{G})$ if $u \in C$, and 0 otherwise.

These quantities admit a natural interpretation in terms of random walks. For $u \in C$, $\ell^{(t)}(u)$ is the probability that a lazy random walk over G originating at v stays within $\widetilde{G}$ for t steps, arriving at u on the tth step, and then "leaks out" of C on the $(t+1)$st step. On the other hand, for $u \notin C$, $s^{(t)}(u)$ is the probability that a lazy random walk over G originating at v stays within $\widetilde{G}$ for t steps and is then "soaked up" into u on the $(t+1)$st step. The vectors ℓ and s then give the total mass leaked and soaked, respectively, by the PPR vector.

Three properties of ℓ and s are worth pointing out. First, $\mathrm{supp}(\ell) \subseteq C \setminus C_o$ and $\mathrm{supp}(s) \subseteq V \setminus C$. Second, $\|\ell^{(t)}\|_1 = \|s^{(t)}\|_1$ for all $t \in \mathbb{N}$, and so $\|\ell\|_1 = \|s\|_1$. Third, for any $u \in V \setminus C$, $p_v(u) = p_s(u)$. The first two properties are immediate. The third property follows by the law of total probability, which implies that

$$q_v^{(\tau)}(u) = \sum_{t=0}^{\tau} q_{s^{(t)}}^{(\tau-t)}(u), \quad \text{for all } u \in V \setminus C.$$

or in terms of the PPR vector,

$$p_v(u) = \alpha\sum_{\tau=0}^{\infty}(1-\alpha)^\tau q_v^{(\tau)}(u) = \alpha\sum_{\tau=0}^{\infty}\sum_{t=0}^{\tau}(1-\alpha)^\tau q_{s^{(t)}}^{(\tau-t)}(u).$$

Substituting $\Delta = \tau + t$ and rearranging gives the claimed property, as

$$p_v(u) = \alpha \sum_{\tau=0}^{\infty} \sum_{t=0}^{\tau} (1-\alpha)^{\tau} q_{s^{(t)}}^{(\tau-t)}(u) = \sum_{\Delta=0}^{\infty} \sum_{t=0}^{\infty} (1-\alpha)^{\Delta+t} q_{s^{(t)}}^{(\Delta)}(u) = \alpha \sum_{\Delta=0}^{\infty} (1-\alpha)^{\Delta} q_s^{(\Delta)}(u) = p_s(u).$$

Proof of Lemma 7. We first show (A.13). From (A.4) and (A.3), we have that

$$p_v(u) \geq \frac{3}{4}\big(1 - \alpha \cdot \tau_\infty(\widetilde{G})\big) \cdot \widetilde{\pi}(u) - \widetilde{p}_\ell(u) \quad \text{for all } u \in C,$$

where ℓ has support $\mathrm{supp}(\ell) \subseteq C$ with $\|\ell\|_1 \leq 2\Phi(C;G)/\alpha$. Recalling that $u \in C_o$ implies that $u \notin \mathrm{supp}(C_o)$, as a consequence of (A.32),

$$\widetilde{p}_\ell(u) \leq \frac{\|\ell\|_1}{d_{\min}(\widetilde{G})} \quad \text{for all } u \in C_o,$$

establishing (A.13). The proof of (A.14) follows similarly:

$$p_v(u) = p_s(u) \overset{(i)}{\leq} \frac{\|s\|_1}{d_{\min}(C';G)} = \frac{\|\ell\|_1}{d_{\min}(C';G)}, \quad \text{for all } u \in C'_o,$$

where the presence of $d_{\min}(C';G)$ on the right hand side of (i) can be verified by inspecting (A.33).

A.1.3 Mixedness of lazy random walk and PPR vectors

In this subsection, we give upper bounds on $h^{(t)} := h_{q_v^{(t)}}$ and $h^{(\alpha)} := h_{p_v}$. Although similar bounds exist in the literature (see in particular Theorem 1.1 of [Lovász and Simonovits, 1990] and Theorem 3 of [Andersen et al., 2006]), we could not find precisely the results we needed, and so for completeness we state and prove these results ourselves.

Theorem 22. *For any $k \in [0, 2m]$, $t_0 \in \mathbb{N}$ and $t \geq t_0$,*

$$h^{(t)}(k) \leq \frac{1}{2^{t_0}} + \frac{d_{\max}(G)}{d_{\min}(G)^2} + \frac{m}{d_{\min}(G)^2}\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0}. \tag{A.16}$$

Theorem 23. *Let ϕ be any constant in $[0,1]$. Either the following bound holds for any $t \in \mathbb{N}$ and any $k \in [d_{\max}(G), 2m - d_{\max}(G)]$:*

$$h^{(\alpha)}(k) \leq \alpha t + \frac{2\alpha}{1+\alpha} + \frac{d_{\max}(G)}{d_{\min}(G)^2} + \frac{m}{d_{\min}(G)^2}\left(1 - \frac{\phi^2}{8}\right)^t,$$

or there exists some sweep cut S_j of p_v such that $\Phi(S_j;G) < \phi$.

The proofs of these upper bounds will be similar to each other (in places word-for-word alike), and will follow a similar approach and use similar notation to that of [Lovász and Simonovits, 1990, Andersen et al., 2006]. For $h : [0, 2m] \to [0, 1]$, $0 \leq K_0 \leq m$ and $k \in [K_0, 2m - K_0]$, define

$$L_{K_0}(k;h) = \frac{2m - K_0 - k}{2m - 2K_0} h(K_0) + \frac{k - K_0}{2m - 2K_0} h(2m - K_0)$$

to be the linear interpolant of $h(K_0)$ and $h(2m - K_0)$, and additionally let

$$C(K_0;h) := \max\left\{ \frac{h(k) - L_{K_0}(k;h)}{\sqrt{\bar{k}}} : K_0 \leq k \leq 2m - K_0 \right\}.$$

where we use the notation $\bar{k} := \min\{k, 2m-k\}$, and treat $0/0$ as equal to 1. Our first pair of Lemmas upper bound $h^{(t)}$ and $h^{(\alpha)}$ as a function of L_{K_0} and $C(K_0)$. Lemma 8 implies that if t is large relative to $\Psi(G)$, then $h^{(t)}(\cdot)$ must be small.

Lemma 8 (**c.f. Theorem 1.2 of [Lovász and Simonovits, 1990]**). *For any $K_0 \in [0, m]$, $k \in [K_0, 2m - K_0]$, $t_0 \in \mathbb{N}$ and $t \geq t_0$,*

$$h^{(t)}(k) \leq L_{K_0}(k; h^{(t_0)}) + C(K_0; h^{(t_0)})\sqrt{k} \cdot \left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0} \tag{A.17}$$

Lemma 9 implies that if the PPR random walk is not well mixed, then some sweep cut of p_v must have small normalized cut.

Lemma 9 (**c.f Theorem 3 of [Andersen et al., 2006]**). *Let $\phi \in [0, 1]$. Either the following bound holds for any $t \in \mathbb{N}$, any $K_0 \in [0, m]$, and any $k \in [K_0, 2m - K_0]$:*

$$h^{(\alpha)}(k) \leq \alpha t + L_{K_0}(k; h^{(\alpha)}) + C(K_0; h^{(\alpha)})\sqrt{k}\left(1 - \frac{\phi^2}{8}\right)^t \tag{A.18}$$

or else there exists some sweep cut S_j of p_v such that $\Phi(S_j; G) < \phi$.

In order to make use of these Lemmas, we require upper bounds on $L_{K_0}(\cdot, h)$ and $C(K_0; h)$, for each of $h = h^{(t_0)}$ and $h = h^{(\alpha)}$. Of course, trivially $L_{K_0}(k; h) \leq \max\{h(K_0); h(2m - K_0)\}$ for any $k \in [K_0, 2m - K_0]$. As it happens, this observation will lead to sufficient upper bounds on $L_{K_0}(k, h)$ for both $h = h^{(t_0)}$ (Lemma 10) and $h = h^{(\alpha)}$ (Lemma 11).

Lemma 10. *For any $t_0 \in \mathbb{N}$ and $K_0 \in [0, m]$, the following inequalities hold:*

$$h^{(t_0)}\big(2m - K_0\big) \leq \frac{K_0}{2m} \quad \text{and} \quad h^{(t_0)}\big(K_0\big) \leq \frac{K_0}{d_{\min}(G)^2} + \frac{1}{2^{t_0}}. \tag{A.19}$$

As a result, for any $k \in [K_0, 2m - K_0]$,

$$L_{K_0}(k; h^{(t_0)}) \leq \max\left\{\frac{K_0}{2m}, \frac{K_0}{d_{\min}(G)^2} + \frac{1}{2^{t_0}}\right\} = \frac{K_0}{d_{\min}(G)^2} + \frac{1}{2^{t_0}}. \tag{A.20}$$

Lemma 11. *For any $\alpha \in [0, 1]$ and $K_0 \in [0, m]$, the following inequalities hold:*

$$h^{(\alpha)}\big(2m - K_0\big) \leq \frac{K_0}{2m} \quad \text{and} \quad h^{(\alpha)}\big(K_0\big) \leq \frac{K_0}{d_{\min}(G)^2} + \frac{2\alpha}{1+\alpha}. \tag{A.21}$$

As a result, for any $k \in [K_0, 2m - K_0]$,

$$L_{K_0}(k; h^{(\alpha)}) \leq \max\left\{\frac{K_0}{2m}, \frac{K_0}{d_{\min}(G)^2} + \frac{2\alpha}{1+\alpha}\right\} = \frac{K_0}{d_{\min}(G)^2} + \frac{2\alpha}{1+\alpha}. \tag{A.22}$$

We next establish an upper bound on $C_{K_0}(k; h)$, which rests on the following key observation: since $h(k)$ is concave and $L_{K_0}(K_0; h) = h(K_0)$, it holds that

$$\frac{h(k) - L_{K_0}(k)}{\sqrt{k}} \leq \begin{cases} h'(K_0)\sqrt{k}, & k \leq m \\ -h'(2m - K_0)\sqrt{2m - k}, & k > m. \end{cases} \tag{A.23}$$

(Since h is not differentiable at $k = k_j$, here h' refers to the right derivative of h.)

Lemma 12 gives good estimates for $h'(K_0)$ and $h'(2m - K_0)$, which hold for both $h = h^{(t_0)}$ and $h = h^{(\alpha)}$, and result in an upper bound on $C(K_0; h)$. Both the statement and proof of this Lemma rely on the following explicit representation of the Lovasz-Simonovits curve $h_q(\cdot)$. Order the vertices $q(u_{(1)})/\deg(u_{(1)}; G) \geq$

$q(u_{(2)})/\deg(u_{(2)};G) \geq \cdots \geq q(u_{(n)})/\deg(u_{(n)};G)$. Then for each $j = 0,\ldots,n-1$, and for all $k \in [\mathrm{vol}(S_j), \mathrm{vol}(S_{j+1}))$, the function $h_q(k)$ satisfies

$$h_q(k) = \sum_{i=0}^{j} \big(q(u_{(i)}) - \pi(u_{(i)})\big) + \frac{\big(k - \mathrm{vol}(S_j;G)\big)}{\deg(u_{(j+1)};G)} \big(q(u_{(j+1)}) - \pi(u_{(j+1)})\big)\,. \tag{A.24}$$

Lemma 12. *The following statements hold for both $h = h^{(\alpha)}$ and $h = h^{(t_0)}$.*

- *Let $K_0 = k_1 = \deg(v;G)$ if $u_{(1)} = v$, and otherwise $K = 0$. Then*

$$h'\big(K_0\big) \leq \frac{1}{d_{\min}(G)^2}. \tag{A.25}$$

- *For all $K_0 \in [0,m]$,*

$$h'(2m - K_0) \geq -\frac{d_{\max}(G)}{d_{\min}(G)\cdot \mathrm{vol}(G)}. \tag{A.26}$$

As a result, letting $K_0 = \deg(v;G)$ if $u_{(1)} = v$, and otherwise letting $K_0 = 0$, we have

$$C(K_0, h) \leq \frac{\sqrt{m}}{d_{\min}(G)^2}.$$

Proof of Theorems 22 and 23

Proof of Theorem 22. Take $K_0 = 0$ if $u_{(1)} \neq v$, and otherwise take $K_0 = \deg(v;G)$. Combining Lemmas 8, 10 and 12, we obtain that for any $k \in [K_0, 2m - K_0]$,

$$\begin{aligned} h^{(t)}(k) &\leq \frac{1}{2^{t_0}} + \frac{K_0}{d_{\min}(G)^2} + \frac{\sqrt{m}}{d_{\min}(G)^2}\sqrt{\bar{k}}\Big(1 - \frac{\Psi^2(G)}{8}\Big)^{t-t_0} \\ &\leq \frac{1}{2^{t_0}} + \frac{d_{\max}(G)}{d_{\min}(G)^2} + \frac{m}{d_{\min}(G)^2}\Big(1 - \frac{\Psi^2(G)}{8}\Big)^{t-t_0}, \end{aligned}$$

where the second inequality follows since we have chosen $K_0 \leq d_{\max}(G)$, and since $\bar{k} \leq m$. If $K_0 = 0$, we are done.

Otherwise, we must still establish that (A.17) is a valid upper bound when $k \in [0, \deg(v;G)) \cup (2m - \deg(v;G), 2m]$. If $k \in [0, \deg(v;G))$ then

$$h^{(t)}(k) \overset{(A.29)}{\leq} h^{(t_0)}(k) \overset{(i)}{\leq} h^{(t_0)}(K_0) \overset{(A.19)}{\leq} \frac{K_0}{d_{\min}(G)^2} + \frac{1}{2^{t_0}}, \tag{A.27}$$

where (i) follows since $k \in [0, K_0]$, and $h^{(t_0)}$ is linear over $[0, K_0)$ with $h^{(t_0)}(0) = 0$ and $h^{(t_0)}(K_0) \geq 0$. For similar reasons,

$$h^{(t)}(k) \leq h^{(t_0)}(k) \leq h^{(t_0)}(2m - K_0) \leq \frac{\deg(v;G)}{2m}. \tag{A.28}$$

Since the ultimate upper bounds in (A.27) and (A.28) are each no greater than that of (A.17), the claim follows.

Proof of Theorem 23. The proof of Theorem 23 follows immediately from Lemmas 9, 11 and 12, taking $K_0 = 0$ if $u_{(1)} \neq v$ and otherwise $K_0 = \deg(v;G)$.

Proofs of Lemmas

In what follows, for a distribution q and vertices $u, w \in V$, we write $q(u,w) := q(u)/d(u) \cdot 1\{(u,w) \in E\}$, and similarly for a collection of dyads $\widetilde{E} \subseteq V \times V$ we write $q(\widetilde{E}) := \sum_{(u,w)\in\widetilde{E}} q(u,w)$.

Proof of Lemma 8. We will prove Lemma 8 by induction on t. In the base case $t = t_0$, observe that $C(K_0; h^{(t_0)}) \cdot \sqrt{\bar{k}} \geq h^{(t_0)}(k) - L_{K_0}(k; h^{(t_0)})$ for all $k \in [K_0, 2m - K_0]$, which implies

$$L_{K_0}(k; h^{(t_0)}) + C(K_0; h^{(t_0)}) \cdot \sqrt{\bar{k}} \geq h^{(t_0)}(k).$$

Now, we proceed with the inductive step, assuming that the inequality holds for $t_0, t_0 + 1, \ldots, t - 1$, and proving that it thus also holds for t. By the definition of L_{K_0}, the inequality (A.17) holds when $k = K_0$ or $k = 2m - K_0$. We will additionally show that (A.17) holds for every $k_j = \mathrm{vol}(S_j), j = 1, 2, \ldots, n$ such that $k_j \in [K_0, 2m - K_0]$. This suffices to show that the inequality (A.17) holds for all $k \in [K_0, 2m - K_0]$, since the right hand side of (A.17) is a concave function of k.

Now, we claim that for each k_j, it holds that

$$q_v^{(t)}[k_j] \leq \frac{1}{2}\Big(q_v^{(t-1)}[k_j - \bar{k}_j\Psi(G)] + q_v^{(t-1)}[k_j + \bar{k}_j\Psi(G)]\Big). \tag{A.29}$$

To establish this claim, we note that for any $u \in V$

$$q_v^{(t)}(u) = \frac{1}{2}q_v^{(t-1)}(u) + \frac{1}{2}\sum_{w\in V} q_v^{(t-1)}(w,u) = \frac{1}{2}\sum_{w\in V}\big(q_v^{(t-1)}(u,w) + q_v^{(t-1)}(w,u)\big),$$

and consequentially for any $S \subset V$,

$$\begin{aligned} q_v^{(t)}(S) &= \frac{1}{2}\big\{q_v^{(t-1)}(\mathrm{in}(S)) + q_v^{(t-1)}(\mathrm{out}(S))\big\} \\ &= \frac{1}{2}\big\{q_v^{(t-1)}\big(\mathrm{in}(S) \cup \mathrm{out}(S)\big) + q_v^{(t-1)}\big(\mathrm{in}(S) \cap \mathrm{out}(S)\big)\big\} \end{aligned}$$

where $\mathrm{in}(S) = \{(u,w) \in E : u \in S\}$ and $\mathrm{out}(S) = \{(w,u) \in E : w \in S\}$. We deduce that

$$\begin{aligned} q_v^{(t)}[k_j] = q_v^{(t)}(S_j) &= \frac{1}{2}\big\{q_v^{(t-1)}\big(\mathrm{in}(S_j) \cup \mathrm{out}(S_j)\big) + q_v^{(t-1)}\big(\mathrm{in}(S_j) \cap \mathrm{out}(S_j)\big)\big\} \\ &\leq \frac{1}{2}\big\{q_v^{(t-1)}\big[|\mathrm{in}(S_j) \cup \mathrm{out}(S_j)|\big] + q_v^{(t-1)}\big[|\mathrm{in}(S_j) \cap \mathrm{out}(S_j)|\big]\big\} \\ &= \frac{1}{2}\big\{q_v^{(t-1)}\big[k_j + \mathrm{cut}(S_j; G)\big] + q_v^{(t-1)}\big[k_j - \mathrm{cut}(S_j; G)\big]\big\} \\ &\leq \frac{1}{2}\big\{q_v^{(t-1)}\big[k_j + \bar{k}_j\Phi(S_j; G)\big] + q_v^{(t-1)}\big[k_j - \bar{k}_j\Phi(S_j; G)\big]\big\} \\ &\leq \frac{1}{2}\big\{q_v^{(t-1)}\big[k_j + \bar{k}_j\Psi(G)\big] + q_v^{(t-1)}\big[k_j - \bar{k}_j\Psi(G)\big]\big\}, \end{aligned}$$

establishing (A.29). The final two inequalities both follow from the concavity of $q_v^{(t)}[\cdot]$.

Subtracting $k_j/2m$ from both sides, we get

$$h^{(t)}(k_j) \leq \frac{1}{2}\big\{h^{(t-1)}\big(k_j + \bar{k}_j\Psi(G)\big) + h^{(t-1)}\big(k_j - \bar{k}_j\Psi(G)\big)\big\}. \tag{A.30}$$

At this point, we divide our analysis into cases.

Case 1. Assume $k_j - \Psi(G)\bar{k}_j$ and $k_j + 2\Psi(G)\bar{k}_j$ are both in $[K_0, 2m - K_0]$. We are therefore in a position to apply our inductive hypothesis to both terms on the right hand side of (A.30) and obtain the following:

$$\begin{aligned}
h^{(t)}(k_j) &\leq \frac{1}{2}\Big(L_{K_0}\big(k_j - \Psi(G)\bar{k}_j; h^{(t_0)}\big) + L_{K_0}\big(k_j + \Psi(G)\bar{k}_j; h^{(t_0)}\big)\Big) \;+ \\
&\quad \frac{1}{2}C\big(K_0; h^{(t_0)}\big) \cdot \Big(\sqrt{k_j - \Psi(G)\bar{k}_j} + \sqrt{k_j + \Psi(G)\bar{k}_j}\Big)\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0-1} \\
&= L_{K_0}(k; h^{(t_0)}) + \frac{1}{2}\left(C(K_0; h^{t_0})\big(\sqrt{k_j - \Psi(G)\bar{k}_j} + \sqrt{k_j + \Psi(G)\bar{k}_j}\big)\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0-1}\right) \\
&\leq L_{K_0}(k; h^{(t_0)}) + \frac{1}{2}\left(C(K_0; h^{(t_0)})\big(\sqrt{\bar{k}_j - \Psi(G)\bar{k}_j} + \sqrt{\bar{k}_j + \Psi(G)\bar{k}_j}\big)\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0-1}\right).
\end{aligned}$$

A Taylor expansion of $\sqrt{1+\Psi(G)}$ around $\Psi(G) = 0$ yields the following bound:

$$\sqrt{1+\Psi(G)} + \sqrt{1-\Psi(G)} \leq 2 - \frac{\Psi(G)^2}{4},$$

and therefore

$$\begin{aligned}
h^{(t)}(k_j) &\leq L_{K_0}(k; h^{(t_0)}) + \frac{C(K_0; h^{(t_0)})}{2} \cdot \sqrt{\bar{k}_j} \cdot \left(2 - \frac{\Psi(G)^2}{4}\right)\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-1} \\
&= L_{K_0}(k_j; h^{(t_0)}) + C(K_0; h^{(t_0)})\sqrt{\bar{k}_j}\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0}.
\end{aligned}$$

Case 2. Otherwise one of $k_j - 2\Psi(G)\bar{k}_j$ or $k_j + 2\Psi(G)\bar{k}_j$ is not in $[K_0, 2m - K_0]$. Without loss of generality assume $k_j < m$, so that (i) we have $k_j - 2\Psi(G)\bar{k}_j < K_0$ and (ii) $k_j + (k_j - K_0) \leq 2m - K_0$. We deduce the following:

$$\begin{aligned}
h^{(t)}(k_j) &\overset{\text{(i)}}{\leq} \frac{1}{2}\Big(h^{(t-1)}(K_0) + h^{(t-1)}\big(k_j + (k_j - K_0)\big)\Big) \\
&\overset{\text{(ii)}}{\leq} \frac{1}{2}\Big(h^{(t_0)}(K_0) + h^{(t)}\big(k_j + (k_j - K_0)\big)\Big) \\
&\overset{\text{(iii)}}{\leq} \frac{1}{2}\Big(L_{K_0}(K_0; h^{(t_0)}) + L_{K_0}(2k_j - K_0; h^{(t_0)}) + C(K_0; h^{(t_0)})\sqrt{2k_j - K_0}\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0-1}\Big) \\
&\leq L_{K_0}(k_j; h^{(t_0)}) + C(K_0; h^{(t_0)})\frac{\sqrt{2\bar{k}_j}}{2}\left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0-1} \\
&\leq L_{K_0}(k_j; h^{(t_0)}) + C(K_0; h^{(t_0)})\sqrt{\bar{k}_j} \cdot \left(1 - \frac{\Psi(G)^2}{8}\right)^{t-t_0}
\end{aligned}$$

where ((i)) follows from (A.30) and the concavity of $h^{(t-1)}$, we deduce (ii) from (A.30), which implies that $h^{(t)}(k) \leq h^{(t_0)}(k)$, and (iii) follows from applying the inductive hypothesis to $h^{(t-1)}(2k_j - K_0)$.

Proof (of Lemma 9). We will show that if $\Phi(S_j; g) \geq \phi$ for each $j = 1, \ldots, n$, then (A.18) holds for all t and any $k \in [K_0, 2m - K_0]$.

We proceed by induction on t. Our base case will be $t = 0$. Observe that $C(K_0; h^{(\alpha)}) \cdot \sqrt{\bar{k}} \geq h^{(\alpha)}(k) - L_{K_0}(k; h^{(\alpha)})$ for all $k \in [K_0, 2m - K_0]$, which implies

$$L_{K_0}(k; h^{(\alpha)}) + C(K_0; h^{(\alpha)}) \cdot \sqrt{\bar{k}} \geq h^{(\alpha)}(k).$$

Now, we proceed with the inductive step. By the definition of L_{K_0}, the inequality (A.18) holds when $k = K_0$ or $k = 2m - K_0$. We will additionally show that (A.18) holds for every $k_j = \mathrm{vol}(S_j), j = 1, 2, \ldots, n$ such that $k_j \in [K_0, 2m - K_0]$. This suffices to show that the inequality (A.18) holds for all $k \in [K_0, 2m - K_0]$, since the right hand side of (A.18) is a concave function of k.

By Lemma 5 of Andersen et al. [2006], we have that

$$\begin{aligned} p_v[k_j] &\leq \alpha + \frac{1}{2}\big(p_v[k_j - \mathrm{cut}(S_j; G)] + p_v[k_j + \mathrm{cut}(S_j; G)]\big) \\ &\leq \alpha + \frac{1}{2}\big(p_v[k_j - \Phi(S_j; G)\bar{k}_j] + p_v[k_j + \Phi(S_j; G)\bar{k}_j]\big) \\ &\leq \alpha + \frac{1}{2}\big(p_v[k_j - \phi\bar{k}_j] + p_v[k_j + \phi\bar{k}_j]\big) \end{aligned}$$

and subtracting $k_j/2m$ from both sides, we get

$$h^{(\alpha)}(k_j) \leq \alpha + \frac{1}{2}\big(h^{(\alpha)}(k_j - \phi\bar{k}_j) + h^{(\alpha)}(k_j + \phi\bar{k}_j)\big) \tag{A.31}$$

From this point, we divide our analysis into cases.

Case 1. Assume $k_j - 2\phi\bar{k}_j$ and $k_j + 2\phi\bar{k}_j$ are both in $[K_0, 2m - K_0]$. We are therefore in a position to apply our inductive hypothesis to (A.31), yielding

$$\begin{aligned} h^{(\alpha)}(k_j) \leq\ & \alpha + \\ & \alpha(t-1)\frac{1}{2}\left(L_{K_0}(k_j - \phi\bar{k}_j) + L_{K_0}(k_j + \phi\bar{k}_j) + C(K_0; h^{(\alpha)})\big(\sqrt{k_j - \phi\bar{k}_j} + \sqrt{k_j + \phi\bar{k}_j}\big)\left(1 - \frac{\phi^2}{8}\right)^{t-1}\right) \\ & \leq \alpha t + L_{K_0}(k; h^{(\alpha)}) + \frac{1}{2}\left(C(K_0; h^{(\alpha)})\big(\sqrt{k_j - \phi\bar{k}_j} + \sqrt{k_j + \phi\bar{k}_j}\big)\left(1 - \frac{\phi^2}{8}\right)^{t-1}\right) \\ & \leq \alpha t + L_{K_0}(k; h^{(\alpha)}) + \frac{1}{2}\left(C(K_0; h^{(\alpha)})\big(\sqrt{\bar{k}_j - \phi\bar{k}_j} + \sqrt{\bar{k}_j + \phi\bar{k}_j}\big)\left(1 - \frac{\phi^2}{8}\right)^{t-1}\right). \end{aligned}$$

and therefore

$$\begin{aligned} h^{(\alpha)}(k_j) &\leq \alpha t + L_{K_0}(k; h^{(\alpha)}) + \frac{C(K_0; h^{(\alpha)})}{2} \cdot \sqrt{\bar{k}_j} \cdot \left(2 - \frac{\phi^2}{4}\right)\left(1 - \frac{\phi^2}{8}\right)^{t-1} \\ &= \alpha t + L_{K_0}(k; h^{(\alpha)}) + C(K_0; h^{(\alpha)})\sqrt{\bar{k}_j}\left(1 - \frac{\phi^2}{8}\right)^{t}. \end{aligned}$$

Case 2. Otherwise one of $k_j - 2\phi\bar{k}_j$ or $k_j + 2\phi\bar{k}_j$ is not in $[K_0, 2m - K_0]$. Without loss of generality assume $k_j < m$, so that (i) we have $k_j - 2\phi\bar{k}_j < K_0$ and (ii) $k_j + (k_j - K_0) \leq 2m - K_0$. By the concavity of h, and applying the inductive hypothesis to $h^{(\alpha)}2k_j - K_0)$, we have

$$\begin{aligned} h^{(\alpha)}(k_j) &\leq \alpha + \frac{1}{2}\Big(h^{(\alpha)}(K_0) + h\big(k_j + (k_j - K_0)\big)\Big) \\ &\leq \alpha + \frac{\alpha(t-1)}{2} + \frac{1}{2}\Big(L_{K_0}(K_0; p^{\alpha}) + L_{K_0}(2k_j - K_0) + C(K_0; h^{(\alpha)})\sqrt{2k_j - K_0}\left(1 - \frac{\phi^2}{8}\right)^{t-1}\Big) \\ &\leq \alpha t + L_{K_0}(k_j) + C(K_0; h^{(\alpha)})\frac{\sqrt{2\bar{k}_j}}{2}\left(1 - \frac{\phi^2}{8}\right)^{t-1} \\ &\leq \alpha t + L_{K_0}(k_j) + C(K_0; h^{(\alpha)})\sqrt{\bar{k}_j} \cdot \left(1 - \frac{\phi^2}{8}\right)^{t} \end{aligned}$$

Proof of Lemma 10. We will prove that the inequalities of (A.19) hold at the knot points of $h^{(t_0)}$, whence they follow for all $K_0 \in [0, m]$.

We first prove the upper bound on $h^{(t_0)}(2m - K_0)$, when $2m - K_0 = k_j$ for some $j = 0, \dots, n-1$. Indeed, the following manipulations show the upper bound holds for $h_q(\cdot)$ regardless of the distribution q. Noting that $h_q(2m) = 0$, we have that,

$$h_q(k_j) = h_q(k_j) - h_q(2m) = \sum_{i=j+1}^{n} q(u_{(i)}) - \pi(u_{(i)}) \leq \sum_{i=j+1}^{n} \pi(u_{(i)}) = 1 - \frac{k_j}{2m} = \frac{K_0}{2m}.$$

In contrast, when $K_0 = k_j$ the upper bound on $h^{(t_0)}(\cdot)$ depends on the properties of $q = q_v^{(t_0)}$. In particular, we claim that for any $t \in \mathbb{N}$,

$$q_v^{(t)}(u) \leq \begin{cases} \dfrac{1}{d_{\min}(G)}, & \text{if } u \neq v \\ \dfrac{1}{d_{\min}(G)} + \dfrac{1}{2^t}, & \text{if } u = v. \end{cases} \tag{A.32}$$

This claim follows straightforwardly by induction. In the base case $t = 0$, the claim is obvious. If the claim holds true for a given $t \in \mathbb{N}$, then for $u \neq v$,

$$\begin{aligned} q_v^{(t+1)}(u) &= \frac{1}{2} \sum_{w \neq u} q_v^{(t)}(w, u) + \frac{1}{2} q_v^{(t)}(u) \\ &\leq \frac{1}{2d_{\min}(G)} \sum_{w \neq u} q_v^{(t)}(w) + \frac{1}{2d_{\min}(G)} \\ &\leq \frac{1}{d_{\min}(G)}, \end{aligned} \tag{A.33}$$

where the last inequality holds because $q_v^{(t)}$ is a probability distribution (i.e. the sum of its entries is equal to 1). Similarly, if $u = v$, then

$$\begin{aligned} q_v^{(t+1)}(v) &= \frac{1}{2} \sum_{w \neq v} q_v^{(t)}(w, v) + \frac{1}{2} q_v^{(t)}(v) \\ &\leq \frac{1}{2d_{\min}(G)} \sum_{w \neq u} q_v^{(t)}(w) + \frac{1}{2d_{\min(G)}} + \frac{1}{2^{t+1}} \\ &\leq \frac{1}{d_{\min}(G)} + \frac{1}{2^{t+1}}, \end{aligned}$$

and the claim (A.32) is shown. The upper bound on $h^{(t_0)}(K_0)$ for $K_0 = k_j$ follows straightforwardly:

$$h^{(t_0)}(K_0) \leq \sum_{i=0}^{j} q_v^{(t_0)}(u_{(j)}) \leq \frac{j}{d_{\min}(G)} + \frac{1}{2^{t_0}} \leq \frac{K_0}{d_{\min}(G)^2} + \frac{1}{2^{t_0}},$$

where the last inequality follows since $\mathrm{vol}(S) \geq |S| \cdot d_{\min}(G)$ for any set $S \subseteq V$.

Proof of Lemma 11. We have already established the first upper bound in (A.21), in the proof of Lemma 10. Then, noting that from (A.32),

$$p_v(u) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t q_v^{(t)}(u) \leq \begin{cases} \alpha \displaystyle\sum_{t=0}^{\infty} (1-\alpha)^t \Big(\frac{1}{d_{\min}(G)} + \frac{1}{2^t} \Big) = \frac{1}{d_{\min}(G)} + \frac{2\alpha}{1-\alpha} & \text{if } u = v \\ \alpha \displaystyle\sum_{t=0}^{\infty} (1-\alpha)^t \frac{1}{d_{\min}(G)} = \frac{1}{d_{\min}(G)} & \text{if } u \neq v, \end{cases} \tag{A.34}$$

the second upper bound in (A.21) follows similarly to the proof of the equivalent upper bound in Lemma 10.

Proof of Lemma 12. The result of the Lemma follows obviously from (A.23), once we show (A.25)-(A.26). We begin by showing (A.25). Inspecting the representation (A.24), we see that for any distribution q and knot point k_j, the right derivative of h_q can always be upper bounded,

$$h'_q(k_j) \leq \frac{q(u_{(j+1)})}{\deg(u_{(j+1)}; G)}.$$

We have chosen $K_0 = k_j$ so that $v \neq u_{(j+1)}$, and so (A.32) implies that $h'_q(k_j) \leq 1/(d_{\min}(G)^2)$, for either $q = q_v^{(t)}$ or $q = p_v$.

On the other hand, the inequality (A.26) follows immediately from the representation (A.24), since for any $K_0 \in [0, m]$, taking j so that $2m - K_0 \in [k_j, k_{j+1})$,

$$h'(2m - K_0) \geq -\frac{\pi(u_{(j+1)})}{\deg(u_{(j+1)}; G)} \geq -\frac{d_{\max}(G)}{d_{\min}(G) \cdot \mathrm{vol}(G)}.$$

A.1.4 Proof of Proposition 1

To prove Proposition 1, we will upgrade from an upper bound on the total variation distance between $q_v^{(t)}$ and π to the desired uniform upper bound. The *total variation distance* between distributions q and p is

$$\mathrm{TV}(q, p) := \frac{1}{2} \sum_{u \in v} \big| q(u) - p(u) \big|$$

It follows from the representation (A.24) that

$$\mathrm{TV}(q, \pi) = \max_{S \subseteq V} \Big\{ q(S) - \pi(S) \Big\} = \max_{j=1,\ldots,n} \Big\{ q(S_j) - \pi(S_j) \Big\} = \max_{k \in [0,2m]} h_q(k),$$

so that Theorem 22 gives an upper bound on $\mathrm{TV}(q_v^{(t)}, \pi)$. We can then use the following result to upgrade to a uniform upper bound.

Lemma 13. *For any $t \in \mathbb{N}$,*

$$\max_{u \in V} \Big\{ \frac{\pi(u) - q_v^{(t+1)}(u)}{\pi(u)} \Big\} \leq \frac{2 \cdot \mathrm{TV}(q_v^{(t)}, \pi)}{s(G)}.$$

The proof of Proposition 1 is then straightforward.

Proof of Proposition 1. Put $t_* = 8/(\Psi(G)^2) \ln(4/s(G)) + 4$. We will use Theorem 22 to show that $\mathrm{TV}(q_v^{(t_*)}, \pi) \leq 1/4$. This will in turn imply (Montenegro [2002] pg. 13) that for $\tau_* = t_* \log_2(8/s(G))$,

$$\mathrm{TV}(q_v^{(\tau_*)}, \pi) \leq \frac{1}{8} s(G),$$

and applying Lemma 13 gives

$$\max_{u \in V} \Big\{ \frac{\pi(u) - q_v^{(\tau_* + 1)}(u)}{\pi(u)} \Big\} \leq \frac{1}{4}.$$

Taking maximum over all $v \in V$, we conclude that $\tau_\infty(G) \leq \tau_* + 1$, which implies the claim of Proposition 1.

It remains to show that $\mathrm{TV}(q_v^{(t_*)}, \pi) \le 1/4$. Choosing $t_0 = 4$ in the statement of Theorem 22, we have that

$$\begin{aligned}\mathrm{TV}(q_v^{(t_*)}, \pi) &\le \frac{1}{16} + \frac{d_{\max}(G)}{d_{\min}(G)^2} + \frac{1}{2s(G)}\Big(1 - \frac{\Psi(G)^2}{8}\Big)^{t_*-4} \\ &\le \frac{1}{8} + \frac{1}{2s(G)}\Big(1 - \frac{\Psi(G)^2}{8}\Big)^{t_*-4} \\ &\le \frac{1}{8} + \frac{1}{2s(G)}\exp\Big(-\frac{\Psi(G)^2}{8}(t_* - 4)\Big) = \frac{1}{4},\end{aligned}$$

where the middle inequality follows by assumption.

Proof of Lemma 13. We proceed by induction. In the base case $t = 0$, we have that

$$\max_{u \in V}\Big\{\frac{\pi(u) - q_v^{(t+1)}(u)}{\pi(u)}\Big\} \le 1 \le 2(1 - \pi(v)) \le 2\frac{\mathrm{TV}(q_v^{(0)}, \pi)}{s(G)},$$

, where the second inequality follows since $\pi(v) \le d_{\max}(G)/(2m) \le d_{\max}(G)/d_{\min}(G)^2 \le 1/16$.

To prove the inductive step, the key observation is the following equivalence (see equation (16) of [Morris and Peres, 2005]):

$$\begin{aligned}\frac{\pi(u) - q_v^{(t+1)}(u)}{\pi(u)} &= \sum_{w \in V}\big(\pi(w) - q_v^{(t)}(w)\big) \cdot \Big(\frac{q_w^{(1)}(u) - \pi(u)}{\pi(u)}\Big) \\ &= \sum_{w \ne u}\big(\pi(w) - q_v^{(t)}(w)\big) \cdot \Big(\frac{q_w^{(1)}(u) - \pi(u)}{\pi(u)}\Big) + \big(\pi(u) - q_v^{(t)}(u)\big) \cdot \Big(\frac{q_u^{(1)}(u) - \pi(u)}{\pi(u)}\Big) \qquad \text{(A.35)}\end{aligned}$$

We separately upper bound each term on the right hand side of (A.35). The sum over all $w \ne u$ can be related to the TV distance between $q_v^{(t)}$ and π using Hölder's inequality,

$$\begin{aligned}\sum_{w \ne u}\big(\pi(w) - q_v^{(t)}(w)\big) \cdot \Big(\frac{q_w^{(1)}(u) - \pi(u)}{\pi(u)}\Big) &\le 2\mathrm{TV}(q_v^{(t)}, \pi) \cdot \max_{w \ne u}\Big|\frac{q_w^{(1)}(u) - \pi(u)}{\pi(u)}\Big| \\ &\le 2\mathrm{TV}(q_v^{(t)}, \pi) \cdot \max\Big\{1, \max_{w \ne u}\frac{q_w^{(1)}(u)}{\pi(u)}\Big\} \\ &\le 2\mathrm{TV}(q_v^{(t)}, \pi) \cdot \frac{m}{d_{\min}(G)^2} = \frac{\mathrm{TV}(q_v^{(t)}, \pi)}{s(G)}.\end{aligned}$$

On the other hand, the second term on the right hand side of (A.35) satisfies

$$\big(\pi(u) - q_v^{(t)}(u)\big) \cdot \Big(\frac{q_u^{(1)}(u) - \pi(u)}{\pi(u)}\Big) \le \big(\pi(u) - q_v^{(t)}(u)\big) \cdot \Big(\frac{1/2 - \pi(u)}{\pi(u)}\Big) \le \frac{\pi(u) - q_v^{(t)}(u)}{2\pi(u)},$$

so that we obtain the recurrence relation

$$\frac{\pi(u) - q_v^{(t+1)}(u)}{\pi(u)} \le \frac{\mathrm{TV}(q_v^{(t)}, \pi)}{s(G)} + \frac{\pi(u) - q_v^{(t)}(u)}{2\pi(u)}.$$

Then by the inductive hypothesis $\big(\pi(u) - q_v^{(t)}(u)\big)/\big(2\pi(u)\big) \le \mathrm{TV}(q_v^{(t-1)}, \pi)/s(G)$, and consequentially

$$\frac{\pi(u) - q_v^{(t+1)}(u)}{\pi(u)} \le \frac{\mathrm{TV}(q_v^{(t)}, \pi)}{s(G)} + \frac{\mathrm{TV}(q_v^{(t-1)}, \pi)}{s(G)} \le 2\frac{\mathrm{TV}(q_v^{(t)}, \pi)}{s(G)}.$$

This completes the proof of Lemma 13.

A.1.5 Spectral partitioning properties of PPR

The following theorem is the main result of Section A.1.5. It relates the normalized cut of the sweep sets $\Phi(S_\beta; G)$ to the normalized cut of a candidate cluster $C \subseteq V$, when p_v is properly initialized within C.

Theorem 24 (**c.f. Theorem 6 of Andersen et al. [2006]**). *Suppose that*

$$d_{\max}(G) \leq \mathrm{vol}(C; G) \leq \max\Big\{\frac{2}{3}\mathrm{vol}(G); \mathrm{vol}(G) - d_{\max}(G)\Big\} \tag{A.36}$$

and

$$\max\Big\{288\Phi(C;G) \cdot \ln\Big(\frac{36}{s(G)}\Big), 72\Phi(C;G) + \frac{d_{\max}(G)}{d_{\min}(G)^2}\Big\} < \frac{1}{18}. \tag{A.37}$$

Set $\alpha = 36 \cdot \Phi(C; G)$. The following statement holds: there exists a set $C^g \subseteq C$ of large volume, $\mathrm{vol}(C^g; G) \geq 5/6 \cdot \mathrm{vol}(C; G)$, such that for any $v \in C^g$, the minimum normalized cut of the sweep sets of p_v satisfies

$$\min_{\beta \in (0,1)} \Phi(S_{\beta,v}; G) < 72\sqrt{\Phi(C;G) \cdot \ln\Big(\frac{36}{s(G)}\Big)}. \tag{A.38}$$

A few remarks:

- Theorem 24 is similar to Theorem 6 of Andersen et al. [2006], but crucially the above bound depends on $\log\big(1/s(G)\big)$ rather than $\log m$. In the case where $d_{\min}(G)^2 \asymp \mathrm{vol}(G)$ and thus $s(G) \asymp 1$, this amounts to replacing a factor of $O(\log m)$ by a factor of $O(1)$, and therefore allows us to obtain meaningful results in the limit as $m \to \infty$.
- For simplicity, we have chosen to state Theorem 24 with respect to a specific choice of $\alpha = 36 \cdot \Phi(C; G)$, but if $\alpha \approx 36 \cdot \Phi(C; G)$ then the Theorem will still hold up to constant factors.

It follows from Markov's inequality (see Theorem 4 of Andersen et al. [2006]) that there exists a set $C^g \subseteq C$ of volume $\mathrm{vol}(C^g; G) \geq 5/6 \cdot \mathrm{vol}(C; G)$ such that for any $v \in C^g$,

$$p_v(C) \geq 1 - \frac{6\Phi(C;G)}{\alpha}. \tag{A.39}$$

The claim of Theorem 24 is a consequence of (A.39) along with Theorem 23, as we now demonstrate.

Proof of Theorem 24. From (A.39), the upper bound in (A.36), and the choice of $\alpha = 36 \cdot \Phi(C; G)$,

$$p_v(C) - \pi(C) \geq \frac{1}{3} - \frac{6\Phi(C;G)}{\alpha} = \frac{1}{6}. \tag{A.40}$$

Now, put

$$t_* = \frac{1}{648\Phi(C;G)}, \quad \phi_*^2 = \frac{8}{t_*} \cdot \ln\Big(\frac{36}{s(G)}\Big),$$

and note that by (A.37) $\phi_*^2 \in [0, 1]$. It therefore follows from (A.40) and Theorem 23 that either

$$\frac{1}{6} \leq p_v(C) - \pi(C) \leq \frac{1}{18} + 72\Phi(C;G) + \frac{d_{\max}(G)}{d_{\min}(G)^2} + \frac{1}{2s(G)} \cdot \Big(1 - \frac{\phi_*^2}{8}\Big)^{t_*}, \tag{A.41}$$

or $\min_{\beta \in (0,1)} \Phi(S_{\beta,v}; G) \leq \phi_*^2$. But by (A.37)

$$72\Phi(C;G) + \frac{d_{\max}(G)}{d_{\min}(G)^2} < \frac{1}{18},$$

and we have chosen ϕ_* precisely so that

$$\frac{1}{2s(G)} \cdot \left(1 - \frac{\phi_*^2}{8}\right)^{t_*} \leq \frac{1}{2s(G)} \exp\Big(-\frac{\phi_*^2 t_*}{8}\Big) \leq \frac{1}{18}.$$

Thus the inequality (A.41) cannot hold, and so it must be that $\min_{\beta\in(0,1)} \Phi(S_{\beta,v}; G) \leq \phi_*^2$. This is exactly the claim of the theorem.

A.2 Sample-to-population bounds

In this appendix, we prove Propositions 2 and 3, by establishing high-probability finite-sample bounds on various functionals of the random graph $G_{n,r}$: cut, volume, and normalized cut (A.2.2), minimum and maximum degree, and local spread (A.2.3), and conductance (A.2.4). To establish these results, we will use several different concentration inequalities, and we begin by reviewing these in (A.2.1). Throughout, we denote the empirical probability of a set $\mathcal{S} \subseteq \mathbb{R}^d$ as $\mathbb{P}_n(\mathcal{S}) = \sum_{i=1}^n \mathbf{1}\{x_i \in \mathcal{S}\}/n$, and the conditional (on being in $\mathcal{C}$) empirical probability as $\widetilde{\mathbb{P}}_n = \sum_{i=1}^n \mathbf{1}\{x_i \in (\mathcal{S}\cap\mathcal{C})\}/\widetilde{n}$, where $\widetilde{n} = |\mathcal{C}[X]|$ is the number of sample points that are in $\mathcal{C}$. For a probability measure $\mathbb{Q}$, we also write

$$d_{\min}(\mathbb{Q}) := \inf_{x\in\mathrm{supp}(\mathbb{Q})} \deg_{\mathbb{P},r}(x), \quad \text{and} \quad d_{\max}(\mathbb{Q}) := \sup_{x\in\mathrm{supp}(\mathbb{Q})} \deg_{\mathbb{P},r}(x). \tag{A.42}$$

A.2.1 Review: concentration inequalities

We use Bernstein's inequality to control the deviations of the empirical probability of $\mathcal{S}$.

Lemma 14 (Bernstein's Inequality.)**.** *Fix $\delta \in (0,1)$. For any measurable $\mathcal{S} \subseteq \mathbb{R}^d$, each of the inequalities,*

$$(1-\delta)\mathbb{P}(\mathcal{S}) \leq \mathbb{P}_n(\mathcal{S}) \quad \textit{and} \quad \mathbb{P}_n(\mathcal{S}) \leq (1+\delta)\mathbb{P}(\mathcal{S}),$$

hold with probability at least $1 - \exp\{-n\delta^2\mathbb{P}(\mathcal{S})/(2+2\delta)\} \leq 1 - \exp\{-n\delta^2\mathbb{P}(\mathcal{S})/4\}$.

Many graph functionals are order-2 U-statistics, and we use Bernstein's inequality to control the deviations of these functionals from their expectations. Recall that U_n is an order-2 U-statistic with kernel $\varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ if

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\neq i} \varphi(x_i, x_j).$$

We write $\|\varphi\|_\infty = \sup_{x,y} |\varphi(x,y)|$.

Lemma 15 (Bernstein's Inequality for Order-2 U-statistics.)**.** *Fix $\delta \in (0,1)$. Assume $\|\varphi\|_\infty \leq 1$. Then each of the inequalities,*

$$(1-\delta)\mathbb{E}U_n \leq U_n \quad \textit{and} \quad U_n \leq (1+\delta)\mathbb{E}U_n,$$

hold with probability at least $1 - \exp\{-n\delta^2\mathbb{E}U_n/(4+4\delta/3)\} \leq 1 - \exp\{-n\delta^2\mathbb{E}U_n/6\}$.

Finally, we use Lemma 16—a combination of Bernstein's inequality and a union bound—to upper and lower bound $d_{\max}(G_{n,r})$ and $d_{\min}(G_{n,r})$. For measurable sets $\mathcal{S}_1, \ldots, \mathcal{S}_M$, we denote $p_{\min} := \min_{m=1,\ldots,M} \mathbb{P}(\mathcal{A}_m)$, and likewise let $p_{\max} := \max_{m=1,\ldots,M} \mathbb{P}(\mathcal{A}_m)$

Lemma 16 (Bernstein's inequality + union bound.)**.** *Fix $\delta \in (0,1)$. For any measurable $\mathcal{S}_1, \ldots, \mathcal{S}_M \subseteq \mathbb{R}^d$, each of the inequalities*

$$(1-\delta)p_{\min} \leq \min_{m=1,\ldots,M} \mathbb{P}_n(\mathcal{A}_m), \quad \textit{and} \quad \max_{m=1,\ldots,M} \mathbb{P}_n(\mathcal{A}_m) \leq (1+\delta)p_{\max}$$

hold with probability at least $1 - M\exp\{-n\delta^2 p_{\min}/(2+2\delta)\} \geq 1 - M\exp\{-n\delta^2 p_{\min}/4\}$.

A.2.2 Sample-to-population: normalized cut

In this subsection we establish (2.17). For a set $\mathcal{S} \subseteq \mathbb{R}^d$, both $\mathrm{cut}_{n,r}(\mathcal{S}[X])$ and $\mathrm{vol}_{n,r}(\mathcal{S}[X])$ are order-2 U-statistics:

$$\mathrm{cut}_{n,r}(\mathcal{S}[X]) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}\{\|x_i - x_j\| \leq r\} \cdot \mathbf{1}\{x_i \in \mathcal{S}\} \cdot \mathbf{1}\{x_j \notin \mathcal{S}\},$$

and

$$\mathrm{vol}_{n,r}(\mathcal{S}[X]) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}\{\|x_i - x_j\| \leq r\} \cdot \mathbf{1}\{x_i \in \mathcal{S}\}.$$

Therefore with probability at least $1 - \exp\{-n\delta^2 \mathrm{cut}_{\mathbb{P},r}(\mathcal{S})/4\}$,

$$\frac{1}{n(n-1)} \mathrm{cut}_{n,r}(\mathcal{S}[X]) \leq (1+\delta)\mathrm{cut}_{\mathbb{P},r}(\mathcal{S}),$$

and likewise with probability at least $1 - \exp\{-n\delta^2 \mathrm{vol}_{\mathbb{P},r}(\mathcal{S})/4\} - \exp\{-n\delta^2 \mathrm{vol}_{\mathbb{P},r}(\mathcal{S}^c)/4\}$,

$$(1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{S}) \leq \frac{1}{n(n-1)} \mathrm{vol}_{n,r}(\mathcal{S}[X]), \quad \text{and} \quad (1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{S}^c) \leq \frac{1}{n(n-1)} \mathrm{vol}_{n,r}(\mathcal{S}^c[X])$$

Consequently, for any $\delta \in (0, 1/3)$,

$$\Phi_{n,r}(\mathcal{C}[X]) \leq \frac{1+\delta}{1-\delta} \cdot \frac{\mathrm{cut}_{\mathbb{P},r}(\mathcal{C})}{\min\{\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}), \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^c)\}} = \frac{1+\delta}{1-\delta} \cdot \Phi_{\mathbb{P},r}(\mathcal{C}) \leq (1+3\delta) \cdot \Phi_{\mathbb{P},r}(\mathcal{C})$$

with probability at least $1 - 3\exp\{-n\delta^2 \mathrm{cut}_{\mathbb{P},r}(\mathcal{C})/4\}$. This establishes (2.17) upon taking $b_1 := 3\mathrm{cut}_{\mathbb{P},r}(\mathcal{C})/4$.

A.2.3 Sample-to-population: local spread

In this subsection we establish (2.19). To ease the notational burden, let $\widetilde{G}_{n,r} := G_{n,r}\big[\mathcal{C}[X]\big]$. Conditional on $\widetilde{n}$, it follows from Lemma 16 that with probability at least $1 - \widetilde{n}\exp\{-(\widetilde{n}-1)\delta^2 d_{\min}(\widetilde{\mathbb{P}})/4\}$,

$$(1-\delta) \cdot d_{\min}(\widetilde{\mathbb{P}}) \leq \frac{1}{\widetilde{n}-1} d_{\min}(\widetilde{G}_{n,r}), \tag{A.43}$$

Likewise it follows from Lemma 15 that with probability at least $1 - \exp\{-\widetilde{n}\delta^2 \mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C})/6\}$,

$$(1-\delta) \cdot \mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C}) \leq \frac{1}{\widetilde{n}(\widetilde{n}-1)} \mathrm{vol}(\widetilde{G}_{n,r}).$$

Finally, it follows from Lemma 14 that with probability at least $1 - \exp\{-n\delta^2 \mathbb{P}(\mathcal{C})/4\}$

$$\widetilde{n} \geq (1-\delta) \cdot n \cdot \mathbb{P}(\mathcal{C}), \tag{A.44}$$

and therefore by (2.18), $(\widetilde{n}-1)/\widetilde{n} \geq 1 - \delta$. Consequently for any $\delta \in (0, 1/3)$,

$$s_{n,r}(\mathcal{C}[X]) = \frac{d_{\min}(\widetilde{G}_{n,r})^2}{\mathrm{vol}(\widetilde{G}_{n,r})} = \frac{\widetilde{n}-1}{\widetilde{n}} \cdot \frac{\frac{1}{(\widetilde{n}-1)^2} d_{\min}(\widetilde{G}_{n,r})^2}{\frac{1}{\widetilde{n}(\widetilde{n}-1)} \mathrm{vol}(\widetilde{G}_{n,r})} \geq \frac{(1-\delta)^3}{(1+\delta)} \cdot \frac{d_{\min}(\widetilde{\mathbb{P}})^2}{\mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C})} \geq (1-4\delta) \cdot s_{\mathbb{P},r}(\mathcal{C}).$$

with probability at least $1 - n\exp\{-n\mathbb{P}(\mathcal{C}) \cdot \delta^2 d_{\min}(\widetilde{\mathbb{P}})/9\} - \exp\{-n\mathbb{P}(\mathcal{C})\delta^2 \cdot \mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C})/14\} - \exp\{-n\delta^2\mathbb{P}(\mathcal{C})/4\}$. This establishes (2.19) upon taking $b_2 := \mathbb{P}(\mathcal{C}) \cdot d_{\min}(\widetilde{\mathbb{P}})/14$.

A.2.4 Sample-to-population: conductance

In this section we establish (2.21). As mentioned in our main text, the proof of (2.21) relies on a high-probability upper bound of the ∞-transportation distance between $\mathbb{P}$ and $\mathbb{P}_n$, from [García Trillos and Slepcev, 2015]. We begin by reviewing this upper bound, which we restate in Theorem 25. Subsequently in Proposition 18, we relate the ∞-transportation distance between two measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ to the difference of their conductances. Together these results will imply (2.21).

Review: ∞-transportation distance and transportation maps. We give a brief review of some of the main ideas regarding ∞-transportation distance, and transportation maps. This discussion is largely taken from [García Trillos and Slepcev, 2015, García Trillos et al., 2016], and the reader should consult these works for more detail.

For two measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ on a domain D, the *∞-transportation distance* $\Delta_\infty(\mathbb{Q}_1, \mathbb{Q}_2)$ is

$$\Delta_\infty(\mathbb{Q}_1, \mathbb{Q}_2) := \inf_\gamma \Big\{ \operatorname{esssup}_\gamma \big\{ |x-y| : (x,y) \in D \times D \big\} : \gamma \in \Gamma(\mathbb{Q}_1, \mathbb{Q}_2) \Big\}$$

where $\Gamma(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of all couplings of $\mathbb{Q}_1$ and $\mathbb{Q}_2$, that is the set of all probability measures on $D \times D$ for which the marginal distribution in the first variable is $\mathbb{Q}_1$, and the marginal distribution in the second variable is $\mathbb{Q}_2$.

Suppose $\mathbb{Q}_1$ is absolutely continuous with respect to the Lebesgue measure. Then $\Delta_\infty(\mathbb{Q}_1, \mathbb{Q}_2)$ can be more simply defined in terms of push-forward measures and transportation maps. For a Borel map $T : D \to D$, the *push-forward* of $\mathbb{Q}_1$ by T is $T_\sharp \mathbb{Q}_1$, defined for Borel sets U as

$$T_\sharp \mathbb{Q}_1(U) = \mathbb{Q}_1(T^{-1}(U)).$$

A *transportation map* from $\mathbb{Q}_1$ to $\mathbb{Q}_2$ is a Borel map T for which $T_\sharp \mathbb{Q}_1 = \mathbb{Q}_2$. Transportation maps satisfy two important properties. First, the transportation distance can be formulated in terms of transportation maps:

$$\Delta_\infty(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_T \|\mathrm{Id} - T\|_{L^\infty(\mathbb{Q}_1)}$$

where $\mathrm{Id} : D \to D$ is the identity mapping, and the infimum is over transportation maps T from $\mathbb{Q}_1$ to $\mathbb{Q}_2$. Second, they result in the following change of variables formula; if $T_\sharp \mathbb{Q}_1 = \mathbb{Q}_2$, then for any $g \in L^1(\mathbb{Q}_2)$,

$$\int g(y)\, d\mathbb{Q}_2(y) = \int g(T(x))\, d\mathbb{Q}_1(x). \tag{A.45}$$

∞-transportation distance between empirical and population measures. We now review the relevant upper bound on $\Delta_\infty(\mathbb{P}, \mathbb{P}_n)$, which holds under the following mild regularity conditions.

(A1) The distribution $\mathbb{P}$ has density $g : D \to (0, \infty)$ such that there exist $g_{\min} \le 1 \le g_{\max}$ for which

$$(\forall x \in D) \quad g_{\min} \le g(x) \le g_{\max}.$$

(A2) The distribution $\mathbb{P}$ is defined on a bounded, connected, open domain $D \subseteq \mathbb{R}^d$. If $d \ge 2$ then additionally D has Lipschitz boundary.

When $d = 1$, it follows from Proposition 6.2 of Dudley [1968] that $\Delta_\infty(\mathbb{P}, \mathbb{P}_n) \le B_5 \|F - F_n\|_\infty$ for some positive constant B_5, and in turn from the DKW inequality that

$$\Delta_\infty(\mathbb{P}, \mathbb{P}_n) \le B_5 \sqrt{\frac{\ln(2n/B_2)}{n}} \tag{A.46}$$

with probability at least $1 - B_2/n$.

When $d \ge 2$, García Trillos and Slepcev [2015] derive an upper bound on the transportation distance $\Delta_\infty(\mathbb{P}, \mathbb{P}_n)$.

Theorem 25 (Theorem 1.1 of García Trillos and Slepcev [2015]). *Suppose $\mathbb{P}$ satisfies (A1) and (A2). Then, there exists positive constants B_2 and B_5 that do not depend on n, such that with probability at least $1-B_2/n$:*

$$\Delta_\infty(\mathbb{P},\mathbb{P}_n) \leq B_5 \cdot \begin{cases} \dfrac{\ln(n)^{3/4}}{n^{1/2}}, & \text{if } d=2, \\ \dfrac{\ln(n)^{1/d}}{n^{1/d}}, & \text{if } d \geq 3. \end{cases}$$

Assuming the candidate cluster $\mathcal{C}$ and conditional distribution $\widetilde{\mathbb{P}}$ satisfy (A1) and (A2), then (A.46) ($d=1$) or Theorem 25 ($d \geq 2$) apply to $\Delta_\infty(\widetilde{\mathbb{P}},\widetilde{\mathbb{P}}_n)$; we will use these upper bounds on $\Delta_\infty(\widetilde{\mathbb{P}},\widetilde{\mathbb{P}}_n)$ to show (2.21).

Lower bound on conductance using transportation maps. Let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be probability measures, with $\mathbb{Q}_1$ absolutely continuous with respect to Lebesgue measure, and let T be a transportation map from $\mathbb{Q}_1$ to $\mathbb{Q}_2$. We write $\Delta_T(\mathbb{Q}_1,\mathbb{Q}_2) := \|\mathrm{Id}-T\|_{L^\infty(\mathbb{Q}_1)}$. To facilitate easy comparison between the conductances of two arbitrary distributions, let $\Psi_r(\mathbb{Q}) := \Psi_{\mathbb{Q},r}(\mathrm{supp}(\mathbb{Q}))$ for a distribution $\mathbb{Q}$. In the following Proposition, we lower bound $\Psi_r(\mathbb{Q}_2)$ by $\Psi_r(\mathbb{Q}_1)$, plus an error term that depends on $\Delta(\mathbb{Q}_1,\mathbb{Q}_2)$.

Proposition 18. *Let $\mathbb{Q}_1$ be a probability measure that admits a density g with respect to $\nu(\cdot)$, let $\mathbb{Q}_2$ be an arbitrary probability measure, and let T be a transportation map from $\mathbb{Q}_1$ to $\mathbb{Q}_2$. Suppose $\Delta_T(\mathbb{Q}_1,\mathbb{Q}_2) \leq r/(4(d-1))$. It follows that*

$$\Psi_r(\mathbb{Q}_2) \geq \Psi_r(\mathbb{Q}_1)\cdot\left(1-\frac{2B_6\Delta_T(\mathbb{Q}_1,\mathbb{Q}_2)}{\big(1-\Psi_r(\mathbb{Q}_1)\big)\cdot\big(d_{\min}(\mathbb{Q}_2)\big)^2}\right) - \frac{B_6\Delta_T(\mathbb{Q}_1,\mathbb{Q}_2)}{\big(1-\Psi_r(\mathbb{Q}_1)\big)\cdot\big(d_{\min}(\mathbb{Q}_2)\big)^2}, \tag{A.47}$$

where $B_6 := 2d\nu_d r^{d-1}\cdot\max_{x\in\mathbb{R}^d}\{g(x)\}$ is a positive constant that does not depend on $\mathbb{Q}_2$.

We note that the lower bound can also be stated with respect to the ∞-optimal transport distance $\Delta_\infty(\mathbb{Q}_1,\mathbb{Q}_2)$.

Proof of Proposition 18. Throughout this proof, we will write $\Delta_{12} = \Delta_T(\mathbb{Q}_1,\mathbb{Q}_2)$, and $\overline{\mathrm{vol}}_{\mathbb{Q},r}(\mathcal{R}) = \min\{\mathrm{vol}_{\mathbb{Q},r}(\mathcal{R}), \mathrm{vol}_{\mathbb{Q},r}(\mathcal{R}^c)\}$ for conciseness. Naturally, the proof of Proposition 18 involves using the transportation map T to relate $\mathrm{cut}_{\mathbb{Q}_2,r}(\cdot)$ to $\mathrm{cut}_{\mathbb{Q}_1,r}(\cdot)$, and likewise $\mathrm{vol}_{\mathbb{Q}_2,r}(\cdot)$ to $\mathrm{vol}_{\mathbb{Q}_1,r}(\cdot)$. Define the remainder term $R^{(\Delta)}_{\epsilon,\mathbb{Q}_1}(x) = \int \mathbf{1}\{\epsilon \leq \|x-y\| \leq \epsilon+\Delta\}\, d\mathbb{Q}_1(y)$ for any $\epsilon,\Delta>0$. Then for any set $\mathcal{S}\subseteq\mathrm{supp}(\mathbb{Q}_2)$, we have that

$$\begin{aligned}
\mathrm{cut}_{\mathbb{Q}_2,r}(\mathcal{S}) &= \iint \mathbf{1}\{\|x-y\|\leq r\}\cdot\mathbf{1}\{x\in\mathcal{S}\}\cdot\mathbf{1}\{y\in\mathcal{S}^c\}\, d\mathbb{Q}_2(y)\, d\mathbb{Q}_2(x) \\
&\overset{\text{(i)}}{=} \iint \mathbf{1}\{\|T(x)-T(y)\|\leq r\}\cdot\mathbf{1}\{x\in T^{-1}(\mathcal{S})\}\cdot\mathbf{1}\{y\in T^{-1}(\mathcal{S})^c\}\, d\mathbb{Q}_1(y)\, d\mathbb{Q}_1(x) \\
&\overset{\text{(ii)}}{\geq} \iint \mathbf{1}\{\|x-y\|\leq r-2\Delta_{12}\}\cdot\mathbf{1}\{x\in T^{-1}(\mathcal{S})\}\cdot\mathbf{1}\{y\in T^{-1}(\mathcal{S}^c)\}\, d\mathbb{Q}_1(y)\, d\mathbb{Q}_1(x) \\
&= \mathrm{cut}_{\mathbb{Q}_1,r}\big(T^{-1}(\mathcal{S})\big) - \int R^{(2\Delta_{12})}_{r-2\Delta_{12},\mathbb{Q}_1}(x)\, d\mathbb{Q}_1(x)
\end{aligned} \tag{A.48}$$

where (i) follows from the change of variables formula (A.45), and (ii) follows from the triangle inequality. Similar reasoning implies that

$$\mathrm{vol}_{\mathbb{Q}_2,r}(\mathcal{S}) \leq \mathrm{vol}_{\mathbb{Q}_1,r}\big(T^{-1}(\mathcal{S})\big) + \int R^{(2\Delta_{12})}_{r,\mathbb{Q}_1}(x)\, d\mathbb{Q}_1(x). \tag{A.49}$$

For any $x\in\mathbb{R}^d$, the remainder terms can be upper bounded: since $\Delta_{12}\geq 0$,

$$R^{(2\Delta_{12})}_{r-2\Delta_{12},\mathbb{Q}_1}(x) \leq \nu_d r^d\Big\{1-\Big(1-\frac{2\Delta_{12}}{r}\Big)^d\Big\}\cdot\max_{x\in\mathbb{R}^d}\{g(x)\} \leq \underbrace{2d\nu_d r^{d-1}\cdot\max_{x\in\mathbb{R}^d}\{g(x)\}}_{=B_6}\cdot\Delta_{12},$$

if $0 \le \Delta_{12} \le r/(4(d-1))$,

$$R^{(2\Delta_{12})}_{r,\mathbb{Q}_1}(x) \le \nu_d r^d \Big\{ \Big(1 + \frac{2\Delta_{12}}{r}\Big)^d - 1 \Big\} \cdot \max_{x \in \mathbb{R}^d} \{g(x)\} \le 2B_6 \cdot \Delta_{12}.$$

Plugging these bounds on the remainder terms back into (A.48) and (A.49) respectively, we see that

$$\begin{aligned}
\Phi_{\mathbb{Q}_2,r}(\mathcal{S}) &\ge \frac{\mathrm{cut}_{\mathbb{Q}_1,r}\big(T^{-1}(\mathcal{S})\big) - B_6\Delta_{12}}{\overline{\mathrm{vol}}_{\mathbb{Q}_1,r}(T^{-1}(\mathcal{S})) + 2B_6\Delta_{12}} \\
&= \Phi_{\mathbb{Q}_1,r}(T^{-1}(\mathcal{S})) \cdot \left(\frac{\overline{\mathrm{vol}}_{\mathbb{Q}_1,r}(T^{-1}(\mathcal{S}))}{\overline{\mathrm{vol}}_{\mathbb{Q}_1,r}(T^{-1}(\mathcal{S})) + 2B_6\Delta_{12}} \right) - \frac{B_6\Delta_{12}}{\overline{\mathrm{vol}}_{\mathbb{Q}_1,r}(T^{-1}(\mathcal{S})) + 2B_6\Delta_{12}} \\
&\overset{(A.49)}{\ge} \Phi_{\mathbb{Q}_1,r}(T^{-1}(\mathcal{S})) \cdot \left(\frac{\overline{\mathrm{vol}}_{\mathbb{Q}_2,r}(\mathcal{S}) - 2B_6\Delta_{12}}{\overline{\mathrm{vol}}_{\mathbb{Q}_2,r}(\mathcal{S})} \right) - \frac{B_6\Delta_{12}}{\overline{\mathrm{vol}}_{\mathbb{Q}_2,r}(\mathcal{S})}.
\end{aligned}$$

We would like to conclude by taking an infimum over $\mathcal{S}$ on both sides, but in order to ensure that the remainder term is small we must specially handle the case where $\overline{\mathrm{vol}}_{\mathbb{Q}_2,r}(\mathcal{S})$ is small. Let

$$\mathfrak{L}_r(\mathbb{Q}_1, \mathbb{Q}_2) = \big\{ \mathcal{S} \subseteq \mathrm{supp}(\mathbb{Q}_2) : \overline{\mathrm{vol}}_{\mathbb{Q}_2,r}(\mathcal{S}) \ge (1 - \Psi_r(\mathbb{Q}_1)) \cdot d_{\min}(\mathbb{Q}_2)^2 \big\}.$$

On the one hand, taking an infimum over all sets $\mathcal{S} \in \mathfrak{L}_r(\mathbb{Q}_1, \mathbb{Q}_2)$, we have that

$$\inf_{\mathcal{S}: \mathcal{S} \in \mathfrak{L}_r(\mathbb{Q}_1,\mathbb{Q}_2)} \Phi_{\mathbb{Q}_2,r}(\mathcal{S}) \ge \Psi_r(\mathbb{Q}_1) \cdot \left(1 - \frac{2B_6\Delta_{12}}{(1 - \Psi_r(\mathbb{Q}_1)) \cdot d_{\min}(\mathbb{Q}_2)^2}\right) - \frac{B_6\Delta_{12}}{(1 - \Psi_r(\mathbb{Q}_1)) \cdot d_{\min}(\mathbb{Q}_2)^2}$$

On the other hand, we claim that

$$\Phi_{r,\mathbb{Q}_2}(\mathcal{R}) \ge \Psi_r(\mathbb{Q}_1), \quad \text{for any } \mathcal{R} \notin \mathfrak{L}(\mathbb{Q}_1, \mathbb{Q}_2). \tag{A.50}$$

To derive (A.50), suppose that $\mathcal{R} \subseteq \mathrm{supp}(\mathbb{Q}_2)$ and $\mathcal{R} \notin \mathfrak{L}(\mathbb{Q}_1, \mathbb{Q}_2)$. Without loss of generality, we shall assume that $\mathrm{vol}_{\mathbb{Q}_2,r}(\mathcal{R}) \le (1 - \Psi_r(\mathbb{Q}_1)) \cdot d_{\min}(\mathbb{Q}_2)^2$ (otherwise we can work with respect to $\mathcal{R}^c$.) Then, for all $x \in \mathcal{R}$,

$$\int \mathbf{1}\{\|x - y\| \le r\} \cdot \mathbf{1}\{y \in \mathcal{R}^c\} \, d\mathbb{Q}_2(y) \ge \deg_{\mathbb{Q}_2,r}(x) - \mathbb{Q}_2(\mathcal{R}) \ge \deg_{\mathbb{Q}_2,r}(x) - \frac{\mathrm{vol}_{\mathbb{Q}_2,r}(\mathcal{R})}{d_{\min}(\mathbb{Q}_2)} \ge d_{\min}(\mathbb{Q}_2) \cdot \Psi_r(\mathbb{Q}_2),$$

whence integrating over all $x \in \mathcal{R}$ and dividing by $\mathrm{vol}_{\mathbb{Q}_2,r}(\mathcal{R})$ yields (A.50). This completes the proof of Proposition 18.

Putting the pieces together. First, we note that

$$\Psi_r(\widetilde{\mathbb{P}}) = \Psi_{\widetilde{\mathbb{P}},r}(\mathrm{supp}(\widetilde{\mathbb{P}})) = \Psi_{\mathbb{P},r}(\mathcal{C}), \quad \text{and} \quad \Psi_r(\widetilde{\mathbb{P}}_n) = \Psi_{\widetilde{\mathbb{P}}_n,r}(\mathrm{supp}(\widetilde{\mathbb{P}}_n)) = \Psi_{n,r}(\mathcal{C}[X]),$$

so that we may apply Proposition 18 to get a lower bound on $\Psi_{n,r}(\mathcal{C}[X])$ in terms of $\Psi_{\mathbb{P},r}(\mathcal{C})$, $\Delta_\infty(\widetilde{\mathbb{P}}, \widetilde{\mathbb{P}}_n)$, and $d_{\min}(\widetilde{\mathbb{P}}_n)$. We now use the bounds we have derived on transportation distance and minimum degree. From the derivations in Section A.2.4, it follows with probability at least $1 - (n+1)\exp\{-nb_2/16\}$ that,

$$d_{\min}(\widetilde{\mathbb{P}}_n) = \frac{1}{\widetilde{n}}\big(d_{\min}(\widetilde{G}_{n,r}) + 1\big) \ge \frac{1}{\sqrt{2}} d_{\min}(\widetilde{\mathbb{P}}).$$

On the other hand, taking

$$b_6 := \frac{1}{2B_6}\Psi_r(\widetilde{\mathbb{P}}) \cdot (1 - \Psi_r(\widetilde{\mathbb{P}})) \cdot d_{\min}(\widetilde{\mathbb{P}})^2, \quad \text{and} \quad B_1 := B_5\Big(\min\Big\{\frac{b_6}{2\Psi_r(\widetilde{\mathbb{P}})}, b_6, \frac{r}{4(d-1)}\Big\}\Big)^{-1},$$

by (A.46) (if $d = 1$) or Theorem 25 (if $d \geq 2$) along with (2.20), we have that

$$\Delta_\infty(\widetilde{\mathbb{P}}, \widetilde{\mathbb{P}}_n) \leq B_5 \frac{(\log n)^{p_d}}{\min\{n^{1/2}, n^{1/d}\}} \leq \min\Big\{\frac{b_6}{2\Psi_r(\widetilde{\mathbb{P}})}, b_6, \frac{r}{4(d-1)}\Big\} \cdot \delta$$

with probability at least $1 - B_2/n$. Therefore by Proposition 18,

$$\Psi_r(\widetilde{\mathbb{P}}_n) \geq \Psi_r(\widetilde{\mathbb{P}}) \cdot \left(1 - \frac{2B_6\Delta_\infty(\widetilde{\mathbb{P}}_n, \widetilde{\mathbb{P}})}{\big(1 - \Psi_r(\widetilde{\mathbb{P}})\big) \cdot \big(d_{\min}(\widetilde{\mathbb{P}}_n)\big)^2}\right) - \frac{B_6\Delta_\infty(\widetilde{\mathbb{P}}_n, \widetilde{\mathbb{P}}_N)}{\big(1 - \Psi_r(\widetilde{\mathbb{P}})\big) \cdot \big(d_{\min}(\widetilde{\mathbb{P}}_n)\big)^2} \geq \Psi_r(\widetilde{\mathbb{P}})(1 - 2\delta)$$

with probability at least $1 - B_2/n - (n+1)\exp\{-nb_2/16\}$, establishing (2.21) upon taking $b_3 := b_2/16$.

A.3 Population functionals for density clusters

In this appendix, we prove Lemma 2 (in Section A.3.4), Proposition 4 (in Section A.3.5), and Proposition 5 (in Section A.3.6), by establishing bounds on the population-level local spread, normalized cut, and conductance of a thickened density cluster $\mathcal{C}_{\lambda,\sigma}$. In these proofs, we make use of some estimates on the volume of spherical caps (in Section A.3.1); some isoperimetric inequalities (in Section A.3.2), and some reverse isoperimetric inequalities (given in Section A.3.3). Finally, in Section A.3.7, for the hard case distribution $\mathbb{P}$ defined in (2.32) and $\mathcal{L}$ defined in (2.33), establish bounds on the population-level normalized cut $\Phi_{\mathbb{P},r}(\mathcal{L})$ and local spread $s_{\mathbb{P},r}(\mathcal{X})$; these will be useful in the proof of Theorem 5. Throughout, we write $\nu_d := \nu(B(0,1))$ for the Lebesgue measure of a d-dimensional unit ball.

A.3.1 Balls, Spherical Caps, and Associated Estimates

In this section, we derive lower bounds on the volume of the intersection between two balls in $\mathbb{R}^d$, and the volume of a spherical cap. Results of this type are well-known, but since we could not find exactly the statements we desire, for completeness we also supply proofs. We use the notation $B(x,r)$ for a ball of radius r centered at $x \in \mathbb{R}^d$, and $\mathrm{cap}_r(h)$ for a spherical cap of height r and radius r. Recall that the Lebesgue measure of a spherical cap is

$$\nu\big(\mathrm{cap}_r(h)\big) = \frac{1}{2}\nu_d r^d I_{1-a}\left(\frac{d+1}{2}; \frac{1}{2}\right)$$

where $a = (r-h)^2/r^2$, and

$$I_{1-a}(z,w) = \frac{\Gamma(z+w)}{\Gamma(z)\Gamma(w)} \int_0^{1-a} u^{z-1}(1-u)^{w-1} du.$$

is the cumulative distribution function of a $\mathrm{Beta}(z,w)$ distribution, evaluated at $1-a$. (Here $\Gamma(\cdot)$ is the gamma function).

Lemma 17. *For any $x, y \in \mathbb{R}^d$ and $r > 0$, it holds that*

$$\nu\big(B(x,r) \cap B(y,r)\big) \geq \nu_d r^d \left(1 - \frac{\|x-y\|}{r}\sqrt{\frac{d+2}{2\pi}}\right). \tag{A.51}$$

For any $x, y \in \mathbb{R}^d$ and $r, \sigma > 0$ such that $\|x - y\| \leq \sigma$, it holds that,

$$\nu\big(B(x,r) \cap B(y,\sigma)\big) \geq \frac{1}{2}\nu_d r^d \left(1 - \frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\right). \tag{A.52}$$

Lemma 18. *For any $0 < h \leq r$, and $a = 1 - (2rh - h^2)/r^2$,*

$$\nu\big(\mathrm{cap}_r(h)\big) \geq \frac{1}{2}\nu_d r^d \big(1 - 2\sqrt{a} \cdot \sqrt{\frac{d+2}{2\pi}}\big)$$

An immediate implication of (A.52) is that for any $x \in \mathcal{C}_{\lambda,\sigma}$,

$$\nu\big(B(x,r) \cap \mathcal{C}_{\lambda,\sigma}\big) \geq \frac{1}{2}\nu_d r^d \left(1 - \frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\right). \tag{A.53}$$

Proof of Lemma 17. First, we prove (A.51). The intersection $B(x,r) \cap B(y,r)$ consists of two symmetric spherical caps, each with height $h = r - \frac{\|x-y\|}{2}$. As a result, by Lemma 18 we have

$$\nu\big(B(x,r) \cap B(y,r)\big) \geq \nu_d r^d \big(1 - 2\sqrt{a} \cdot \sqrt{\frac{d+2}{2\pi}}\big)$$

where $a = \|x-y\|^2/(4r^2)$, and the claim follows.

Next we prove (A.52). Assume that $\|x-y\| = \sigma$, as otherwise if $0 \leq \|x-y\| < \sigma$ the volume of the overlap will only be larger. Then $B(x,r) \cap B(y,\sigma)$ contains a spherical cap of radius r and height $h = r - \frac{r^2}{2\sigma}$, from Lemma 18 we deduce

$$\nu\big(B(x,r) \cap B(y,\sigma)\big) \geq \frac{1}{2}\nu_d r^d \left(1 - 2\sqrt{a} \cdot \sqrt{\frac{d+2}{2\pi}}\right)$$

for $a = (r-h)^2/r^2 = r^2/(4\sigma^2)$, and the claim follows.

Proof of Lemma 18. For any $0 \leq a \leq 1$, we have that

$$\int_0^{1-a} u^{(d-1)/2}(1-u)^{-1/2}du = \int_0^1 u^{(d-1)/2}(1-u)^{-1/2}du - \int_{1-a}^1 u^{(d-1)/2}(1-u)^{-1/2}du.$$

The first integral is simply

$$\int_0^1 u^{(d-1)/2}(1-u)^{-1/2}du = \frac{\Gamma\big(\frac{d+1}{2}\big)\Gamma\big(\frac{1}{2}\big)}{\Gamma\big(\frac{d}{2}+1\big)},$$

whereas for all $u \in [0,1]$ and $d \geq 1$, the second integral can be upper bounded as follows:

$$\int_{1-a}^1 u^{(d-1)/2}(1-u)^{-1/2}du \leq \int_{1-a}^1 (1-u)^{-1/2}du = \int_0^a u^{-1/2}du = 2\sqrt{a}.$$

As a result,

$$\nu\big(\mathrm{cap}_r(h)\big) \geq \frac{1}{2}\nu_d r^d \left(1 - 2\sqrt{a}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right) \geq \frac{1}{2}\nu_d r^d \left(1 - 2\sqrt{a} \cdot \sqrt{\frac{d+2}{2\pi}}\right).$$

A.3.2 Isoperimetric inequalities

Dyer et al. [1991] establish the following isoperimetric inequality for convex sets.

Lemma 19 (Isoperimetry of a convex set.)**.** *For any partition $(\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3)$ of a convex set $\mathcal{K} \subseteq \mathbb{R}^d$, it holds that*

$$\nu(\Omega_3) \geq 2\frac{\mathrm{dist}(\mathcal{R}_1, \mathcal{R}_2)}{\mathrm{diam}(\mathcal{K})} \min(\nu(\mathcal{R}_1), \nu(\mathcal{R}_2)).$$

Abbasi-Yadkori [2016] points out that if $\mathcal{S}$ is the image of a convex set under a Lipschitz measure-preserving mapping $g : \mathbb{R}^d \to \mathbb{R}^d$, a similar inequality can be obtained.

Corollary 3 (Isoperimetry of Lipschitz embeddings of convex sets.)**.** *Suppose $\mathcal{S}$ is the image of a convex set $\mathcal{K}$ under a mapping $g : \mathbb{R}^d \to \mathbb{R}^d$ such that*

$$\|g(x) - g(y)\| \le L \cdot \|x - y\|, \quad \text{for all } x, y \in \mathcal{K}, \text{ and } \quad \det(\nabla g(x)) = 1 \quad \text{for all } x \in \mathcal{K}.$$

Then for any partition $(\Omega_1, \Omega_2, \Omega_3)$ of $\mathcal{S}$,

$$\nu(\Omega_3) \ge 2\frac{\operatorname{dist}(\Omega_1, \Omega_2)}{\operatorname{diam}(\mathcal{K})L} \min(\nu(\Omega_1), \nu(\Omega_2)).$$

A.3.3 Reverse isoperimetric inequalities

For any set $\mathcal{C} \subseteq \mathbb{R}^d$ and $\sigma > 0$, let $\mathcal{C}_\sigma := \{x : \operatorname{dist}(x, \mathcal{C}) \le \sigma\}$. We begin with an upper bound on the volume of $\mathcal{C}_{\sigma+\delta}$ as compared to $\mathcal{C}_\sigma$.

Lemma 20. *For any bounded set $\mathcal{C} \subseteq \mathbb{R}^d$ and $\sigma, \delta > 0$, it holds that*

$$\nu(\mathcal{C}_{\sigma+\delta}) \le \nu(\mathcal{C}_\sigma) \cdot \Big(1 + \frac{\delta}{\sigma}\Big)^d. \tag{A.54}$$

Lemma 20 is a reverse isoperimetric inequality. To see this, note that if $\delta \le \sigma/d$ then $\big(1 + \delta/\sigma\big)^d \le 1 + d \cdot \delta/(\sigma - d\delta)$, and we deduce from (A.54) that

$$\nu(\mathcal{C}_{\sigma+\delta} \setminus \mathcal{C}_\sigma) = \nu(\mathcal{C}_{\sigma+\delta}) - \nu(\mathcal{C}_\delta) \le \frac{d\delta}{\sigma - d\delta} \cdot \nu(\mathcal{C}_\sigma). \tag{A.55}$$

We use (A.55) along with Assumption (A4) to derive a density-weighted reverse isoperimetric inequality.

Lemma 21. *Let $\mathcal{C}_{\lambda,\sigma}$ satisfy Assumption (A1) and (A4) for some θ, γ and λ_σ. Then for any $0 < r \le \sigma/d$, it holds that*

$$\mathbb{P}\big(\mathcal{C}_{\lambda,\sigma+r} \setminus \mathcal{C}_{\lambda,\sigma}\big) \le \Big(1 + \frac{dr}{\sigma - dr}\Big) \cdot \frac{dr}{\sigma} \cdot \Big(\lambda_\sigma - \theta\frac{r^\gamma}{\gamma+1}\Big) \cdot \nu(\mathcal{C}_{\lambda,\sigma}). \tag{A.56}$$

Proof of Lemma 20. Fix $\delta' > 0$, and take $\epsilon = \delta + \delta'$. We will show that

$$\nu(\mathcal{C}_{\sigma+\epsilon}) \le \nu(\mathcal{C}_\sigma) \cdot \Big(1 + \frac{\epsilon}{\sigma}\Big)^d, \tag{A.57}$$

whence taking a limit as $\delta' \to 0$ yields the claim.

To show (A.57), we need to construct a particular disjoint covering $\mathcal{A}_1(\sigma+\epsilon), \dots, \mathcal{A}_N(\sigma+\epsilon)$ of $\mathcal{C}_{\sigma+\delta}$. To do so, we first take a finite set of points $x_1, \dots, x_N$ such that the net $B(x_1, \sigma+\epsilon), \dots, B(x_N, \sigma+\epsilon)$ covers $\mathcal{C}_{\sigma+\delta}$. Note that such a covering exists for some finite $N = N(\epsilon)$ because $\mathcal{C}_{\sigma+\delta}$ is bounded, and the closure of $\mathcal{C}_{\sigma+\delta}$ is thus a compact subset of $\cup_{x\in\mathcal{C}} B(x, \sigma+\epsilon)$. Defining $\mathcal{A}_1(s), \dots, \mathcal{A}_N(s)$ for a given $s > 0$ to be

$$\mathcal{A}_1(s) := B(x_1, s), \quad \text{and} \quad \mathcal{A}_{j+1}(s) := B(x_{j+1}, s) \setminus \bigcup_{i=1}^{j} B(x_i, s) \quad \text{for } j = 1, \dots, N-1,$$

we have that $\mathcal{A}_1(\sigma+\epsilon), \dots, \mathcal{A}_N(\sigma+\epsilon)$ is a disjoint covering of $\mathcal{C}_{\sigma+\delta}$, and so $\nu(\mathcal{C}_{\sigma+\delta}) \le \sum_{j=1}^N \nu\big(\mathcal{A}_j(\sigma+\epsilon)\big)$.

We claim that for all $j = 1, \dots, N$, the function $s \mapsto \nu\big(\mathcal{A}_j(s)\big)/\nu\big(B(x_j, s)\big)$ is monotonically non-increasing in s. Once this claim is verified, it follows that

$$\nu(\mathcal{A}_j(\sigma+\epsilon)) = \nu\big(B(x_j, \sigma+\epsilon)\big) \cdot \frac{\nu\big(\mathcal{A}_j(\sigma+\epsilon)\big)}{\nu\big(B(x_j, \sigma+\epsilon)\big)} \le \Big(1+\frac{\epsilon}{\sigma}\Big)^d \cdot \nu\big(B(x_j, \sigma)\big) \cdot \frac{\nu\big(\mathcal{A}_j(\sigma)\big)}{\nu\big(B(x_j, \sigma)\big)} = \Big(1+\frac{\epsilon}{\sigma}\Big)^d \cdot \nu\big(\mathcal{A}_j(\sigma)$$

and summing over j, we see that

$$\nu(\mathcal{C}_{\sigma+\delta}) \le \sum_{j=1}^{N} \nu\big(\mathcal{A}_j(\sigma+\epsilon)\big) \le \Big(1+\frac{\epsilon}{\sigma}\Big)^d \cdot \sum_{j=1}^{N} \nu\big(\mathcal{A}_j(\sigma)\big) \le \Big(1+\frac{\epsilon}{\sigma}\Big)^d \nu(\mathcal{C}_\sigma).$$

The last inequality follows since $\mathcal{A}_1(\sigma), \dots, \mathcal{A}_N(\sigma)$ are disjoint subsets of the closure of $\mathcal{C}_\sigma$.

It remains to verify that $x \mapsto \nu\big(\mathcal{A}_j(s)\big)/\nu\big(B(x_j,s)\big)$ is monotonically non-increasing. For any $0 < s < t$ and $j = 1, \dots, N$, suppose $x \in \mathcal{A}_j(T) - \{x_j\}$, meaning $x \in B(0,t)$ and $x \notin B(x_i - x_j, t)$ for any $i = 1, \dots, j-1$. Thus $(s/t)x \in B(0,s)$, and

$$\|(s/t)x - (x_i - x_j)\| \ge \|x - (x_i - x_j)\| - \|x - (s/t)x\| > t - (1-s/t)\|x\| \ge t - (1-s/t)t = s,$$

or in other words $(s/t)x \notin B(x_i - x_j, s)$ for any $i = 1, \dots, j-1$. Consequently,

$$\Big(\mathcal{A}_j(t) - \{x_j\}\Big) \subset \frac{t}{s} \cdot \Big(\mathcal{A}_j(s) - \{x_j\}\Big),$$

and applying $\nu(\cdot)$ to both sides yields the claim.

Proof of Lemma 21. Fix $k \in \mathbb{N}$. To establish (A.56), we partition $\mathcal{C}_{\lambda,\sigma+r} \setminus \mathcal{C}_{\lambda,\sigma}$ into thin tubes $\mathcal{T}_1, \dots, \mathcal{T}_k$, with the jth tube $\mathcal{T}_j$ defined as $\mathcal{T}_j := \mathcal{C}_{\lambda,\sigma+jr/k} \setminus \mathcal{C}_{\lambda,\sigma+(j-1)r/k}$. We upper bound the Lebesgue measure of each tube $\mathcal{T}_j$ using (A.55):[1]

$$\nu(\mathcal{T}_j) \le \frac{dr/k}{\sigma - dr/k} \nu(\mathcal{C}_{\lambda,\sigma+(j-1)r/k}) \le \frac{dr/k}{\sigma - dr/k} \nu(\mathcal{C}_{\lambda,\sigma+r}) \le \Big(1 + \frac{dr}{\sigma - dr}\Big) \cdot \frac{dr/k}{\sigma - dr/k} \cdot \nu(\mathcal{C}_{\lambda,\sigma}),$$

and the maximum density within each tube using (A4):

$$\max_{x \in \mathcal{T}_j} f(x) \le \lambda_\sigma - \theta\Big(\frac{j-1}{k} r\Big)^\gamma;$$

combining these upper bounds, we see that

$$\mathbb{P}\big(\mathcal{C}_{\lambda,\sigma+r} \setminus \mathcal{C}_{\lambda,\sigma}\big) = \sum_{j=1}^{k} \mathbb{P}(\mathcal{T}_j) \le \Big(1 + \frac{dr}{\sigma - dr}\Big) \cdot \frac{dr/k}{\sigma - dr/k} \cdot \nu(\mathcal{C}_{\lambda,\sigma}) \cdot \Big(\sum_{j=0}^{k-1} \lambda_\sigma - \theta r^\gamma \Big(\frac{j}{k}\Big)^\gamma\Big). \tag{A.58}$$

Treating the sum in the previous expression as a Riemann sum of a non-increasing function evaluated at $0, \dots, k-1$ gives the upper bound

$$\sum_{j=0}^{k-1} \lambda_\sigma - \theta r^\gamma \Big(\frac{j}{k}\Big)^\gamma \le \lambda_\sigma + \int_0^{k-1} \Big(\lambda_\sigma - \theta r^\gamma \Big(\frac{x}{k}\Big)^\gamma\Big)\, dx \le k\lambda_\sigma + (k-1)\frac{\theta r^\gamma}{\gamma+1}\Big(\frac{k-1}{k}\Big)^\gamma,$$

and plugging back in to (A.58), we obtain

$$\mathbb{P}\big(\mathcal{C}_{\lambda,\sigma+r} \setminus \mathcal{C}_{\lambda,\sigma}\big) \le \Big(1 + \frac{dr}{\sigma - dr}\Big) \cdot \frac{dr}{\sigma - dr/k} \nu(\mathcal{C}_{\lambda,\sigma}) \cdot \Big(\lambda - \frac{\theta r^\gamma}{\gamma+1} \cdot \Big(\frac{k-1}{k}\Big)^{\gamma+1}\Big).$$

The above inequality holds for any $k \in \mathbb{N}$, and taking the limit of the right hand side as $k \to \infty$ yields the claim.

[1] Note that $\mathcal{C}$ must be bounded, since the density $f(x) \ge \lambda_\sigma$ for all $x \in \mathcal{C}$.

A.3.4 Proof of Lemma 2

The population-level local spread of $\mathcal{C}_{\lambda,\sigma}$ is

$$s_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) = \frac{\big(d_{\min}(\widetilde{\mathbb{P}})\big)^2}{\mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C}_{\lambda},\sigma)}$$

where we recall that $\widetilde{\mathbb{P}}(\mathcal{S}) = \frac{\mathbb{P}(\mathcal{S}\cap\mathcal{C}_{\lambda,\sigma})}{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})}$ for Borel sets $\mathcal{S}$, and $d_{\min}(\widetilde{\mathbb{P}}) := \min_{x\in\mathcal{C}_{\lambda,\sigma}}\{\deg_{\widetilde{\mathbb{P}},r}(x)\}^2$. To lower bound $s_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$, we first lower bound $d_{\min}(\widetilde{\mathbb{P}})$, and then upper bound $\mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C}_{\lambda},\sigma)$. Using the lower bound $f(x) \geq \lambda_\sigma$ for all $x \in \mathcal{C}_{\lambda,\sigma}$ stipulated in (A3), we deduce that

$$\begin{aligned}
d_{\min}(\widetilde{\mathbb{P}}) &= \min_{x\in\mathcal{C}_{\lambda,\sigma}}\Big\{\int \mathbf{1}\{\|x-y\|\leq r\}\,d\widetilde{\mathbb{P}}(y)\Big\} \\
&\geq \frac{\lambda_\sigma}{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})}\cdot\min_{x\in\mathcal{C}_{\lambda,\sigma}}\Big\{\int_{\mathcal{C}_{\lambda,\sigma}}\mathbf{1}\{\|x-y\|\leq r\}\,dy\Big\} \\
&\geq \frac{\lambda_\sigma}{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})}\cdot\frac{1}{2}\nu_d r^d\cdot\Big(1-\frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\Big),
\end{aligned}$$

where the final inequality follows from Lemma 17.

On the other hand, using the upper bound $f(x) \leq \Lambda_\sigma$ for all $x \in \mathcal{C}_{\lambda,\sigma}$, we deduce that

$$\begin{aligned}
\mathrm{vol}_{\widetilde{\mathbb{P}},r}(\mathcal{C}_{\lambda,\sigma}) &= \iint \mathbf{1}\{\|x-y\|\leq r\}\,d\widetilde{\mathbb{P}}(y)\,d\widetilde{\mathbb{P}}(x) \\
&\leq \frac{\Lambda_\sigma^2}{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})^2}\cdot\int_{\mathcal{C}_{\lambda,\sigma}}\int_{\mathcal{C}_{\lambda,\sigma}}\mathbf{1}\{\|x-y\|\leq r\}\,dy\,dx \\
&\leq \frac{\Lambda_\sigma^2}{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})^2}\cdot\nu_d r^d\cdot\nu(\mathcal{C}_{\lambda,\sigma}) \\
&\leq \frac{\Lambda_\sigma^2}{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})^2}\cdot\nu_d^2 r^d\cdot\Big(\frac{\rho}{2}\Big)^d;
\end{aligned}$$

the final inequality follows from (A5), which implies that $\nu(\mathcal{C}_{\lambda,\sigma}) = \nu(\mathcal{K}) \leq \nu_d(\rho/2)^d$. The claim of Lemma 2 follows.

A.3.5 Proof of Proposition 4

By Assumption (A6), we have that $\Phi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) = \mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})/\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$, and to prove Proposition 4 we must therefore upper bound $\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$ and lower bound $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$.

Let $\mathcal{C}_{\lambda,\sigma+r} = \{x : \mathrm{dist}(x,\mathcal{C}_\lambda) \leq \sigma + r\}$. We upper bound $\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$ in terms of the probability mass of $\mathcal{C}_{\lambda,\sigma+r}\setminus\mathcal{C}_{\lambda,\sigma}$:

$$\begin{aligned}
\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) &= \iint \mathbf{1}\{\|x-y\|\leq r\}\cdot\mathbf{1}\{x\in\mathcal{C}_{\lambda,\sigma}\}\cdot\mathbf{1}\{y\notin\mathcal{C}_{\lambda,\sigma}\}\,d\mathbb{P}(y)\,d\mathbb{P}(x) \\
&\leq \iint \mathbf{1}\{\|x-y\|\leq r\}\cdot\mathbf{1}\{x\in\mathcal{C}_{\lambda,\sigma}\}\cdot\mathbf{1}\{y\in\mathcal{C}_{\lambda,\sigma+r}\setminus\mathcal{C}_{\lambda,\sigma}\}\,d\mathbb{P}(y)\,d\mathbb{P}(x) \\
&\leq \lambda\nu_d r^d\cdot\mathbb{P}\big(\mathcal{C}_{\lambda,\sigma+r}\setminus\mathcal{C}_{\lambda,\sigma}\big).
\end{aligned}$$

On the other hand, using the lower bound $f(x) \geq \lambda_\sigma$ for all $x \in \mathcal{C}_{\lambda,\sigma}$, we lower bound $\mathrm{cut}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})$ in terms

of the Lebesgue measure of $\mathcal{C}_{\lambda,\sigma}$:

$$\begin{aligned}\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) &= \iint \mathbf{1}\{\|x-y\| \le r\} \cdot \mathbf{1}\{x \in \mathcal{C}_{\lambda,\sigma}\}\, d\mathbb{P}(y)\, d\mathbb{P}(x) \\ &\ge \lambda_\sigma^2 \cdot \iint \mathbf{1}\{\|x-y\| \le r\} \cdot \mathbf{1}\{x, y \in \mathcal{C}_{\lambda,\sigma}\}\, dy\, dy \\ &\ge \lambda_\sigma^2 \cdot \frac{1}{2}\nu_d r^d \cdot \Big(1 - \frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\Big) \cdot \nu(\mathcal{C}_{\lambda,\sigma}).\end{aligned}$$

The claim of Proposition 4 follows upon using Lemma 21 to upper bound $\mathbb{P}\big(\mathcal{C}_{\lambda,\sigma+r} \setminus \mathcal{C}_{\lambda,\sigma}\big)$.

A.3.6 Proof of Proposition 5

The following Lemma lower bounds the population-level uniform conductance $\Psi_{\nu,r}(\mathcal{C}_{\lambda,\sigma})$.

Lemma 22. *Suppose $\mathcal{C}_{\lambda,\sigma}$ satisfies Assumption (A5) with respect to some $\rho \in (0,\infty)$ and $L \in [1,\infty)$. For any $0 < r \le \sigma \cdot \sqrt{2\pi/(d+2)}$, it holds that*

$$\Psi_{\nu,r}(\mathcal{C}_{\lambda,\sigma}) \ge \Big(1 - \frac{r}{4\rho L}\Big) \cdot \Big(1 - \frac{r}{\sigma}\sqrt{\frac{d+2}{2\pi}}\Big)^2 \cdot \frac{\sqrt{2\pi}}{36} \cdot \frac{r}{\rho L\sqrt{d+2}}. \tag{A.59}$$

Noting that $\Psi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}) \ge \Psi_{\nu,r}(\mathcal{C}_{\lambda,\sigma}) \cdot \lambda_\sigma^2/\Lambda_\sigma^2$, Proposition 5 follows from (A.59).

Proof of Lemma 22. For ease of notation, throughout this proof we write $\widetilde{\nu}$ for the uniform probability measure over $\mathcal{C}_{\lambda,\sigma}$, put $\ell\nu_d r^d := \min_{x \in \mathcal{C}_{\lambda,\sigma}} \nu(B(x,r) \cap \mathcal{C}_\sigma)$ and $a := r/(2\rho L)$.

Let $\mathcal{S}$ be an arbitrary measurable subset of $\mathcal{C}_{\lambda,\sigma}$, and let $\mathcal{R} = \mathcal{C}_{\lambda,\sigma} \setminus \mathcal{S}$. For a given $\delta \in (0,1)$, let the δ-interior of $\mathcal{S}$ be

$$\mathcal{S}^\delta := \{x \in \mathcal{S} : \nu\big(B(x,r) \cap \mathcal{R}\big) \le \ell\delta\nu_d r^d\};$$

define $\mathcal{R}^\delta$ likewise, and let $\mathcal{B}^\delta = \mathcal{C}_{\lambda,\sigma} \setminus (\mathcal{S}^\delta \cup \mathcal{R}^\delta)$ consist of the remaining boundary points. As is standard (see for example Dyer et al. [1991], Lovász and Simonovits [1990]), the proof of Lemma 22 uses several inequalities to lower bound the normalized cut $\Phi_{\widetilde{\nu},r}(\mathcal{S})$.

- **Bounds on cut and volume.** We can lower bound $\mathrm{cut}_{\widetilde{\nu},r}(\mathcal{S})$ as follows:

$$\begin{aligned}\nu(\mathcal{C}_{\lambda,\sigma})^2 \cdot \mathrm{cut}_{\widetilde{\nu},r}(\mathcal{S}) &= \int_{\mathcal{S}}\int_{\mathcal{R}} \mathbf{1}(\|x-y\| \le r)\, dy\, dx \\ &= \frac{1}{2}\Big(\int_{\mathcal{S}}\int_{\mathcal{R}} \mathbf{1}(\|x-y\| \le r)\, dy\, dx + \int_{\mathcal{R}}\int_{\mathcal{S}} \mathbf{1}(\|x-y\| \le r)\, dy\, dx\Big) \\ &\ge \frac{1}{2}\delta\ell\nu_d r^d \cdot \nu(\mathcal{B}^\delta).\end{aligned}$$

  We can upper bound $\mathrm{vol}_{\widetilde{\nu},r}(\mathcal{S})$ as follows:

$$\nu(\mathcal{C}_{\lambda,\sigma})^2 \cdot \mathrm{vol}_{\widetilde{\nu},r}(\mathcal{S}) = \int_{\mathcal{C}_{\lambda,\sigma}}\int_{\mathcal{S}} \mathbf{1}(\|x-y\| \le r)\, dy\, dx \le \nu_d r^d \nu(\mathcal{S})$$

  and likewise for $\mathrm{vol}_{\widetilde{\nu},r}(\mathcal{S})$. Therefore,

$$\Phi_{\widetilde{\nu},r}(\mathcal{S}) \ge \frac{\delta\ell \cdot \nu(\mathcal{B}^\delta)}{2 \cdot \min\{\nu(\mathcal{S}), \nu(\mathcal{R})\}}. \tag{A.60}$$

- **Isoperimetric inequality.** Applying Corollary 3, we have that

$$\nu(\mathcal{B}^\delta) \geq \frac{2 \cdot \operatorname{dist}(\mathcal{S}^\delta, \mathcal{R}^\delta)}{\rho L} \cdot \min\big\{\nu(\mathcal{S}^\delta), \nu(\mathcal{R}^\delta)\big\}. \tag{A.61}$$

- **Lebesgue measure of δ-interiors.** Suppose $\nu(\mathcal{S}^\delta) \leq (1-a) \cdot \nu(\mathcal{S})$ or $\nu(\mathcal{R}^\delta) \leq (1-a) \cdot \nu(\mathcal{R})$. Then $\nu(\mathcal{B}^\delta) \geq a \cdot \min\{\nu(\mathcal{S}), \nu(\mathcal{R})\}$, and combined with (A.60) we have that $\Phi_{\widetilde{\nu},r}(\mathcal{S}) \geq \delta a \ell/2$. Otherwise,

$$\min\big\{\nu(\mathcal{S}^\delta), \nu(\mathcal{R}^\delta)\big\} \geq (1-a) \cdot \min\big\{\nu(\mathcal{S}), \nu(\mathcal{R})\big\}. \tag{A.62}$$

- **Distance between δ-interiors.** For any $x \in \mathcal{S}^\delta$ and $y \in \mathcal{R}^\delta$, we have that

$$\begin{aligned}
\nu\big(B(x,r) \cap B(y,r)\big) &= \nu\big(B(x,r) \cap B(y,r) \cap \mathcal{R}\big) + \nu\big(B(x,r) \cap B(y,r) \cap \mathcal{S}\big) + \nu\big(B(x,r) \cap B(y,r) \cap \mathcal{C}_{\lambda,\sigma}^c\big) \\
&\leq \nu\big(B(x,r) \cap \mathcal{R}\big) + \nu\big(B(y,r) \cap \mathcal{S}\big) + \nu\big(B(x,r) \cap \mathcal{C}_{\lambda,\sigma}^c\big) \\
&\leq \big(2\ell\delta + (1-\ell)\big) \cdot \nu_d r^d.
\end{aligned}$$

  It follows from (A.51) that

$$\|x-y\| \geq \frac{r}{\nu_d r^d} \cdot \Big(\nu_d r^d - \nu\big(B(x,r) \cap B(y,r)\big)\Big) \cdot \sqrt{\frac{2\pi}{d+2}} \geq r \cdot \ell \cdot (1-2\delta) \cdot \sqrt{\frac{2\pi}{d+2}},$$

  and taking the infimum over all $x \in \mathcal{S}^\delta$ and $y \in \mathcal{R}^\delta$, we have

$$\operatorname{dist}(\mathcal{S}^\delta, \mathcal{R}^\delta) \geq r \cdot \ell \cdot (1-2\delta) \cdot \sqrt{\frac{2\pi}{d+2}}. \tag{A.63}$$

Combining (A.60)-(A.63) and taking $\delta = 1/3$ implies that

$$\Phi_{\widetilde{\nu},r}(\mathcal{S}) \geq \min\left\{(1-a) \cdot \frac{r}{\rho L} \cdot \frac{\ell^2}{9} \cdot \sqrt{\frac{2\pi}{d+2}}, \frac{a\ell}{6}\right\}$$

and the claim follows from (A.53), which implies that $\ell \geq 1/2 \cdot (1 - r/\sigma)\sqrt{2\pi/(d+2)}$.

A.3.7 Population functionals, hard case

Let $\mathbb{P}$ be the hard case distribution over rectangular domain $\mathcal{X}$, defined as in (2.32), and $\mathcal{L}$ the lower half of $\mathcal{X}$. Suppose $r \in (0, \sigma/2)$. Then the population normalized cut $\Phi_{\mathbb{P},r}(\mathcal{L})$ is upper bounded,

$$\Phi_{\mathbb{P},r}(\mathcal{L}) \leq \frac{8}{3} \cdot \frac{r}{\rho}. \tag{A.64}$$

and the population local spread $s_{\mathbb{P},r}(\mathcal{X})$ is lower bounded,

$$s_{\mathbb{P},r}(\mathcal{X}) \geq \frac{\pi r^2 \epsilon^2}{2\rho\sigma} \tag{A.65}$$

Proof of (A.64). Noting that $\operatorname{vol}_{\mathbb{P},r}(\mathcal{L}) = \operatorname{vol}_{\mathbb{P},r}(\mathcal{X} \setminus \mathcal{L})$, it suffices to upper bound $\operatorname{cut}_{\mathbb{P},r}(\mathcal{L})$ and lower bound $\operatorname{vol}_{\mathbb{P},r}(\mathcal{L})$. Note that for any $x = (x_1, x_2) \in \mathcal{L}$, if $x_2 \leq -r$ the ball $B(x,r)$ and the set $\mathcal{X} \setminus \mathcal{L}$ are disjoint. As a result,

$$\operatorname{cut}_{\mathbb{P},r}(\mathcal{L}) \leq \mathbb{P}\Big(\big\{x \in \mathcal{X} : x_2 \in (-r, 0)\big\}\Big) \cdot d_{\max}(\mathbb{P}) \leq \frac{r}{2\rho} \cdot \frac{\pi r^2}{2\sigma\rho}.$$

On the other hand, noting that $\deg_{\mathbb{P},r}(x) \geq \frac{\pi r^2}{2\sigma\rho}$ for all $x \in \mathcal{C}^{(1)}$ such that $\mathrm{dist}(x, \partial\mathcal{C}^{(1)}) > r$, we have

$$\begin{aligned}\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) &\geq \mathbb{P}\Big(\big\{x \in \mathcal{C}^{(1)} \cap \mathcal{L} : \mathrm{dist}(x, \partial\mathcal{C}^{(1)}) > r\big\}\Big) \cdot \frac{\pi r^2}{2\sigma\rho} \\ &= \frac{(\sigma - 2r)(\rho - r)}{2\sigma\rho} \cdot \frac{\pi r^2}{2\sigma\rho} \\ &\geq \frac{3}{16} \cdot \frac{\pi r^2}{2\sigma\rho}\end{aligned}$$

where the last inequality follows since $r \leq \frac{1}{4}\sigma \leq \frac{1}{4}\rho$.

Proof of (A.65)**.** The statement follows since

$$d_{\min}(\mathbb{P}) \geq \frac{\pi r^2}{2} \cdot \min_{x \in \mathcal{X}} f(x) = \frac{\pi r^2}{2} \cdot \frac{\epsilon}{\rho\sigma},$$

and

$$\mathrm{vol}_{\mathbb{P},r}(\mathcal{X}) \leq d_{\max}(\mathbb{P}) \leq \frac{\pi r^2}{2\sigma\rho}.$$

A.4 Proof of Major Theorems

We now prove the three major theorems of our paper: Theorem 3 (in Section A.4.1), Theorem 4, and Theorem 5. Throughout, we use the notation $\widetilde{n} = |\mathcal{C}[X]|$ and $\widetilde{G}_{n,r} = G_{n,r}\big[\mathcal{C}[X]\big]$ as defined above.

A.4.1 Proof of Theorem 3

We begin by recalling some probabilistic estimates needed for the proof of Theorem 3, along with the probability with which they hold.

Probabilistic estimates. Throughout the proof of Theorem 3, we will assume (i) that the inequalities (2.17)-(2.21) are satisfied; (ii) that the volume of $\mathcal{C}[X]$ is upper and lower bounded,

$$(1-\delta) \cdot \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}) \leq \frac{1}{n(n-1)} \mathrm{vol}_{n,r}(\mathcal{C}[X]) \leq (1+\delta) \cdot \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}); \tag{A.66}$$

(iii) that the number of sample points in $\mathcal{C}$ is lower bounded,

$$\widetilde{n} \geq (1-\delta) \cdot n \cdot \mathbb{P}(\mathcal{C}) \overset{(2.18)}{\Longrightarrow} \widetilde{n} - 1 \geq (1-\delta)^2 \cdot n \cdot \mathbb{P}(\mathcal{C}); \tag{A.67}$$

and finally (iv) that the minimum and maximum degree of $\widetilde{G}_{n,r}$ are lower and upper bounded respectively,

$$\frac{1}{\widetilde{n}-1} d_{\min}(\widetilde{G}_{n,r}) \geq (1-\delta) \cdot d_{\min}(\widetilde{\mathbb{P}}), \quad \text{and} \quad \frac{1}{\widetilde{n}-1} d_{\max}(\widetilde{G}_{n,r}) \leq (1+\delta) \cdot d_{\max}(\widetilde{\mathbb{P}}). \tag{A.68}$$

By Propositions 2 and 3, and Lemmas 14-16, these inequalities are satisfied with probability at least $1 - B_2/n - 4\exp\{-b_1\delta^2 n\} - (2n+2)\exp\{-b_2\delta^2 n\} - (n+1)\exp\{-b_3 n\}$.

Proof of Theorem 3. We use Lemma 1 to upper bound $\Delta(\widehat{C}, \mathcal{C}[X])$. In order to do so, we must verify that the tuning parameters α and (L, U) satisfy the condition (2.12) of this lemma, i.e. that $\alpha \leq 1/\big(2\tau_\infty(\widetilde{G}_{n,r})\big)$

and $U \leq 1/\big(5\mathrm{vol}_{n,r}(\mathcal{C}[X])\big)$. In order to verify the upper bound on α, we will use Proposition 18 to upper bound $\tau_\infty(\widetilde{G}_{n,r})$, which we may validly apply because

$$\frac{d_{\max}(\widetilde{G}_{n,r})}{\big(d_{\min}(\widetilde{G}_{n,r})\big)^2} \leq \frac{(1+\delta)}{(1-\delta)^2} \cdot \frac{d_{\max}(\widetilde{\mathbb{P}})}{(\widetilde{n}-1)\cdot\big(d_{\min}(\widetilde{\mathbb{P}})\big)^2} \leq \frac{(1+\delta)}{(1-\delta)^4} \cdot \frac{d_{\max}(\widetilde{\mathbb{P}})}{n\mathbb{P}(\mathcal{C})\big(d_{\min}(\widetilde{\mathbb{P}})\big)^2} \leq \frac{1}{16}.$$

The last inequality in the above follows by taking $B_3 := 16 \cdot d_{\max}(\widetilde{\mathbb{P}})/\big(d_{\min}(\widetilde{\mathbb{P}})\big)^2$ in (2.25).

Therefore by Proposition 18, along with inequalities (2.19) and (2.21) and the initialization conditions (2.22) and (2.23), we have that $\alpha \leq 1/45 \wedge 1/\big(2\tau_\infty(\widetilde{G}_{n,r})\big)$. On the other hand, by the upper bound on $\mathrm{vol}_{n,r}(\mathcal{C}[X])$ given in (A.66) and the initialization condition (2.22), we have that $U \leq 1/\big(5\mathrm{vol}_{n,r}(\mathcal{C}[X])\big)$. In summary, we have confirmed that the condition (2.12) is satisfied.

Invoking Lemma 1, we conclude that there exists a set $\mathcal{C}[X]^g \subset \mathcal{C}[X]$ of volume at least $\mathrm{vol}_{n,r}(\mathcal{C}[X]^g) \geq \mathrm{vol}_{n,r}(\mathcal{C}[X])/2$, such that for any $\beta \in (L, U)$,

$$\mathrm{vol}_{n,r}(S_{\beta,v} \vartriangle \mathcal{C}[X]) \leq 60 \cdot \frac{\Phi_{n,r}(\mathcal{C}[X])}{\alpha L} \leq 60\frac{(1+2\delta)}{(1-4\delta)^2} \cdot \frac{\Phi_{n,r}(\mathcal{C}[X])}{\alpha_{\mathbb{P},r}(\mathcal{C},\delta)} \cdot n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C})$$

Noting that $\widehat{C} = S_{\beta,v}$ for some $\beta \in (L, U)$, the claimed upper bound (2.26) on $\Delta(\widehat{C}, \mathcal{C}[X])$ then follows from the upper bound (2.17) on $\Phi_{n,r}(\mathcal{C}[X])$ and the upper bound on $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C})$ in (A.66).

A.4.2 Proof of Theorem 4

From Theorem 3, we have with probability $1 - B_2/n - 4\exp\{-b_1\delta^2 n\} - (2n+2)\exp\{-b_2\delta^2 n\} - (n+1)\exp\{-b_3 n\}$, there exists a set $\mathcal{C}_{\lambda,\sigma}[X]^g \subset \mathcal{C}_{\lambda,\sigma}[X]$ of volume at least $\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X]^g) \geq \mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])/2$, such that

$$\begin{aligned}\frac{\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])} &\leq 60 \cdot \frac{(1+3\delta)(1+2\delta)}{(1-4\delta)^2(1-2\delta)} \cdot \frac{\Phi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})}{\alpha_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma},\delta)} \\ &\leq \frac{780}{\ln(2)} \cdot \frac{(1+3\delta)(1+2\delta)}{(1-4\delta)^2(1-2\delta)} \cdot \frac{\Phi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})}{\Psi_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})^2} \cdot \ln^2\Big(\frac{8}{(1-3\delta)s_{\mathbb{P},r}(\mathcal{C})}\Big)\end{aligned}$$

The claimed upper bound (2.31) on $\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])$ then follows from the bounds (2.28)-(2.30) on the population-level local spread, normalized cut, and conductance of $\mathcal{C}_{\lambda,\sigma}$, noting that the condition $r \leq \sigma/(4d)$ implies that $(1 - r/(4\rho L)) \geq 1 - 1/16$ and $1 - r/\sigma \cdot \sqrt{(d+2)/(2\pi)} \geq 1 - 1/\sqrt{32}$, and taking

$$C_{1,\delta} := \frac{898560}{\pi \cdot \ln(2) \cdot (1-1/16)^2 \cdot (1-1/\sqrt{32})^4} \cdot \frac{(1+3\delta)(1+2\delta)}{(1-4\delta)^2(1-2\delta)}, \quad \text{and} \quad C_{2,\delta} := \frac{36}{(1-1/\sqrt{32})} \cdot \frac{1}{1-3\delta}.$$

A.4.3 Proof of Theorem 5

We start by defining some constants, to make our proof statements easier to digest. Put

$$\begin{aligned} C_{3,\delta} &:= \frac{72(1+\delta)}{(1-\delta)}\sqrt{8/3+8\delta}, & C_{4,\delta} &:= \frac{72}{(1-3\delta)\pi}, \\ B_{1,\delta} &:= 768 \cdot (1+3\delta) \cdot \ln\Big(C_{4,\delta}\frac{\rho\sigma}{r^2\epsilon^2}\Big), & B_{2,\delta} &:= \frac{(1+\delta)^2}{(1-\delta)^2} \cdot \frac{\rho\sigma}{r^2\epsilon^2} \\ B_4 &:= 1 + \frac{12\sigma\rho}{r^2} + \frac{2\rho}{r}, & b_4 &:= b_8 \wedge \mathrm{cut}_{\mathbb{P},r}(\mathcal{L}) \wedge d_{\min}(\mathbb{P})/14 \wedge \mathrm{vol}_{\mathbb{P},r}(\mathcal{L} \cap \mathcal{C}^{(1)}), \\ b_8 &:= \mathrm{vol}_{\mathbb{P},r}(\mathcal{X})/4 \wedge \frac{\epsilon r^2}{4\rho\sigma} \wedge \frac{\pi r^3}{8\sigma\rho^2}. \end{aligned}$$

To prove Theorem 5, we use Theorem 24, Proposition 2 and (A.64) to show that the cluster estimate $\widehat{C}$ must have a small normalized cut. On the other hand, in Lemma 23 we establish that any set $Z \subseteq X$ which is close to $\mathcal{C}^{(1)}[X]$—meaning $\mathrm{vol}_{n,r}(Z \vartriangle \mathcal{C}^{(1)}[X])$ is small—has a large normalized cut.

Lemma 23. *Fix $\delta \in (0,1)$. With probability at least $1 - B_4 \exp\{-n\delta^2 b_8\}$, the following statement holds:*

$$\Phi_{n,r}(Z) \geq \frac{(1-\delta)^2}{2(1+\delta)\pi}\left(1 - 2\frac{\sigma\rho}{(1-\delta)r^2n^2}\mathrm{vol}_{n,r}(Z \vartriangle \mathcal{C}^{(1)}[X])\right)\frac{\epsilon^2 r}{\sigma}, \quad \textit{for all } Z \subseteq X. \tag{A.69}$$

We therefore conclude that $\mathrm{vol}_{n,r}(\widehat{C} \vartriangle \mathcal{C}^{(1)}[X])$ must be large. In the remainder of this, we detail the probabilistic estimates used in the proof of Theorem 5, and then give a formal proof of Theorem 5 and then of Lemma 23.

Probabilistic estimates. In addition to (A.69), we will assume (i) that the graph normalized cut of $\mathcal{L}$ and local spread of $\mathcal{X}$ are respectively upper and lower bounded,

$$\Phi_{n,r}(\mathcal{L}[X]) \leq (1+3\delta)\cdot\Phi_{\mathbb{P},r}(\mathcal{L}), \quad \text{and} \quad s_{n,r}(X) \geq (1-3\delta)\cdot s_{\mathbb{P},r}(\mathcal{X});$$

(ii) that the graph volume of $\mathcal{L}$ is upper and lower bounded,

$$(1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) \leq \frac{1}{n(n-1)}\mathrm{vol}_{n,r}(\mathcal{L}[X]) \leq (1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L});$$

(iii) that the graph volumes of $\mathcal{L}\cap\mathcal{C}^{(1)}$ and $\mathcal{C}^{(1)}$ are respectively lower and upper bounded,

$$(1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}\cap\mathcal{C}^{(1)}) \leq \frac{1}{n(n-1)}\mathrm{vol}_{n,r}(\mathcal{L}[X]\cap\mathcal{C}^{(1)}[X]) \quad \text{and} \quad \frac{1}{n(n-1)}\mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X]) \leq (1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)}); \tag{A.70}$$

(iv) that the graph volume of $\mathcal{X}$ is lower bounded,

$$\frac{1}{n(n-1)}\mathrm{vol}_{n,r}(X) \geq (1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{X});$$

and finally (v) that the maximum degree of $G_{n,r}$ is upper bounded,

$$\frac{1}{n-1}d_{\max}(G_{n,r}) \leq (1+\delta)d_{\max}(\mathbb{P}).$$

It follows from Lemma 23, Propositions 2 and 3, and Lemmas 14 and 16 that these estimates are together satisfied with probability at least $1 - B_4\exp\{-n\delta^2 b_8\} - 3\exp\{-n\delta^2\mathrm{cut}_{\mathbb{P},r}(\mathcal{L})\} - (2n+2)\exp\{-n\delta^2\cdot d_{\min}(\mathbb{P})/14\} - 5\exp\{-n\delta^2\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}\cap\mathcal{C}^{(1)})\} \geq 1 - (B_4+2n+10)\exp\{-n\delta^2 b_4\}$.

Proof of Theorem 5. As mentioned, we would like to use Theorem 24 to upper bound $\Phi_{n,r}(\widehat{C})$, and so we first verify that the conditions of Theorem 24 are met. In particular, we have each of the following.

- Recall that $n \geq 8\cdot(1+\delta)/(1-\delta)$ (2.34) and that $\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) \geq 3/16\cdot\pi r^2/(2\sigma\rho)$ (as shown in the proof of (A.64)). It is additionally clear that $d_{\max}(\mathbb{P}) \leq \pi r^2/(2\rho\sigma)$, and consequently,

$$d_{\max}(G_{n,r}) \leq (n-1)\cdot(1+\delta)d_{\max}(\mathbb{P}) \leq \frac{1}{3}n^2(1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) \leq \frac{1}{3}\mathrm{vol}_{n,r}(\mathcal{L}[X]). \tag{A.71}$$

  Therefore the lower bound in condition (A.36) is satisfied.

- Note that $\delta \in (0, 1/7)$ implies $(1-\delta)/(1+\delta) > 3/4$, and additionally that $\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) \le \mathrm{vol}_{\mathbb{P},r}(\mathcal{X})/2$. It follows that

$$\mathrm{vol}_{n,r}(X) \ge n(n-1)(1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{X}) \ge 2n(n-1)(1-\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) \ge 2\frac{(1-\delta)}{(1+\delta)}\mathrm{vol}_{n,r}(\mathcal{L}[X]) \ge \frac{3}{2}\mathrm{vol}_{n,r}(\mathcal{L}[X]). \tag{A.72}$$

  Therefore the upper bound in condition (A.36) is satisfied.

- By (A.64), the normalized cut of $\mathcal{L}$ satisfies the following lower bound,

$$\Phi_{n,r}(\mathcal{L}[X]) \le (1+3\delta)\cdot\Phi_{\mathbb{P},r}(\mathcal{L}) \le (8/3+8\delta)\cdot\frac{r}{\rho}, \tag{A.73}$$

  and by (A.65) the local spread of $\mathcal{X}$ satisfies the following upper bound,

$$s_{n,r}(X) \ge (1-3\delta)\cdot s_{\mathbb{P},r}(\mathcal{X}) \ge (1-3\delta)\cdot\frac{\pi r^2}{2\rho\sigma}. \tag{A.74}$$

  The constants $B_{1,\delta}$ and $B_{2,\delta}$ in assumption (2.34) are chosen so that condition (A.37) is satisfied.

As a result, we may apply Theorem 24, and deduce the following: there exists a set $\mathcal{L}[X]^g \subset \mathcal{L}$ of large volume, $\mathrm{vol}_{n,r}(\mathcal{L}[X]^g) \ge 5/6\cdot\mathrm{vol}_{n,r}(\mathcal{L}[X])$, such that for any seed node $v \in \mathcal{L}[X]^g$, the normalized cut of the PPR cluster estimate $\Phi_{n,r}(\widehat{C})$ satisfies the following upper bound:

$$\Phi_{n,r}(\widehat{C}) < 72\sqrt{\Phi_{n,r}(\mathcal{L}[X])\cdot\ln\Big(\frac{36}{s_{n,r}(X)}\Big)} \le 72\sqrt{(8/3+8\delta)\cdot\frac{r}{\rho}\cdot\ln\Big(\frac{72\rho\sigma}{(1-3\delta)\pi r^2\epsilon^2}\Big)}.$$

Combined with Lemma 23, this implies

$$\frac{(1-\delta)^2}{2(1+\delta)\pi}\left(1-2\frac{\sigma\rho}{(1-\delta)r^2n^2}\mathrm{vol}_{n,r}(\widehat{C}\,\triangle\,\mathcal{C}^{(1)}[X])\right)\frac{\epsilon^2 r}{\sigma} \le 72\sqrt{(8/3+8\delta)\cdot\frac{r}{\rho}\cdot\ln\Big(\frac{72\rho\sigma}{(1-3\delta)\pi r^2\epsilon^2}\Big)}, \tag{A.75}$$

and solving for $\mathrm{vol}_{n,r}(\widehat{C}\,\triangle\,\mathcal{C}^{(1)}[X])$ yields (2.35).

We conclude by observing that the set $\mathcal{L}[X]^g$ must have significant overlap with $\mathcal{C}^{(1)}[X]$. In particular,

$$\begin{aligned}
\mathrm{vol}_{n,r}(\mathcal{L}[X]^g \cap \mathcal{C}^{(1)}[X]) &\ge \mathrm{vol}_{n,r}\big((\mathcal{L}\cap\mathcal{C}^{(1)})[X]\big) - \frac{1}{6}\mathrm{vol}_{n,r}\big(\mathcal{L}[X]\big) \\
&\overset{\text{(i)}}{\ge} n(n-1)\cdot\left((1-\delta)-\frac{1}{2}(1+\delta)\right)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}\cap\mathcal{C}^{(1)}) \\
&\overset{(ii)}{\ge} n(n-1)\cdot\frac{1}{7}\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}\cap\mathcal{C}^{(1)}) \\
&\overset{(iii)}{\ge} \frac{1}{8}\mathrm{vol}_{n,r}(\mathcal{L}\cap\mathcal{C}^{(1)})
\end{aligned}$$

where in (i) we have used $\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) \le 3\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)})$, and in (ii) and (iii) we have used $\delta \in (0, 1/7)$.

Proof of Lemma 23. To lower bound the normalized cut $\Phi_{n,r}(Z)$, it suffices to lower bound $\mathrm{cut}_{n,r}(Z)$ and upper bound $\mathrm{vol}_{n,r}(Z)$. A crude upper bound on the volume is simply

$$\mathrm{vol}_{n,r}(Z) \le \mathrm{vol}_{n,r}(G_{n,r}) \overset{(i)}{\le} (1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{X})n(n-1) \le (1+\delta)\frac{\pi r^2}{\rho\sigma}n^2 \tag{A.76}$$

where by Lemma 15, inequality (i) holds with probability at least $1-\exp\{-n\delta^2\mathrm{vol}_{\mathbb{P},r}(\mathcal{X})/4\}$. This crude upper bound will suffice for our purposes.

We turn to lower bounding $\mathrm{cut}_{n,r}(Z)$, which is considerably more involved. We will approximate the cut of Z by discretizing the space $\mathcal{X}$ into bins, relate the cut of Z to the boundary of the binned set $\overline{Z}$, and then lower bound the size of the boundary of $\overline{Z}$.

Let (k_1, k_2) for $k_1 \in \left[\frac{6\sigma}{r}\right], k_2 \in \left[\frac{2\rho}{r}\right]$ be the upper right corner of the cube

$$Q_{(k_1,k_2)} = \left[-\frac{3\sigma}{2} + \frac{(k_1-1)}{2}r, -\frac{3\sigma}{2} + \frac{k_1}{2}r\right] \times \left[-\frac{\rho}{2} + \frac{(k_2-1)}{2}r, -\frac{\rho}{2} + \frac{k_2}{2}r\right]$$

and let $\overline{Q} = \left\{Q_{(k_1,k_2)} : k_1 \in \left[\frac{6\sigma}{r}\right], k_2 \in \left[\frac{2\rho}{r}\right]\right\}$ be the collection of such cubes. For a set $Z \subset X$ we define the binned set $\overline{Z} \subset \overline{Q}$ as follows

$$\overline{Z} := \left\{Q \in \overline{Q} : \mathbb{P}_n(Z \cap Q) \geq \frac{1}{2}\mathbb{P}_n(Q)\right\},$$

and we let

$$\partial\overline{Z} := \left\{Q_{(k_1,k_2)} \in \overline{Z} : \exists(\ell_1, \ell_2) \in \left[\frac{3\sigma}{r}\right] \times \left[\frac{\rho}{r}\right] \text{ such that } Q_{(\ell_1,\ell_2)} \notin \bar{Z}, \|k - \ell\|_1 = 1\right\}.$$

be the boundary set of $\overline{Z}$ in $\overline{Q}$. Intuitively, every point $x_i \in Z$ in the boundary set of $\overline{Z}$ will have many edges to $X \setminus Z$. Formally, letting $Q_{\min} := \min_{Q \in \overline{Q}} \mathbb{P}_n(Q)$, we have

$$\mathrm{cut}_{n,r}(Z) \geq \mathrm{cut}_{n,r}(Z \cap \{x_i \in \overline{Z}\}) \geq \frac{1}{4}\left|\partial\overline{Z}\right| Q_{\min}^2, \tag{A.77}$$

where the last inequality follows since for every cube $Q_k \in \partial\overline{Z}$, there exists a cube $Q_\ell \notin \overline{Z}$ such that $\|i - j\|_1 \leq 1$, and since each cube has side length $r/2$, this implies that for every $x_i \in Q_k$ and $x_j \in Q_\ell$ the edge (x_i, x_j) belongs to $G_{n,r}$.

Now we move on lower bounding the size of the boundary $\left|\partial\bar{Z}\right|$. To do so, we divide $\mathcal{X}$ into slices horizontally. Let $R_k = \left\{(x_1, x_2) \in \mathcal{X} : x_2 \in \left[-\frac{\rho}{2} + \frac{(k-1)}{2}r, -\frac{\rho}{2} + \frac{k}{2}r\right]\right\}$ be the kth horizontal slice, and $\bar{R}_k = \left\{Q_{(k_1,k)} \in \overline{Q} : k_1 \in [\frac{6\sigma}{r}]\right\}$ be the binned version of R_k. For each k, either

1. $\bar{R}_k \cap \bar{Z} = \emptyset$, in which case

$$\mathrm{vol}_{n,r}\Big(\big(Z \triangle \mathcal{C}^{(1)}[X]\big) \cap R_k\Big) \geq \frac{1}{2}\mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X] \cap R_k), \quad \text{or}$$

2. $\bar{R}_k \cap \bar{Z} = \bar{R}_k$, in which case

$$\mathrm{vol}_{n,r}\Big(\big(Z \triangle \mathcal{C}[X]\big) \cap R_k\Big) \geq \frac{1}{2}\mathrm{vol}_{n,r}(\mathcal{C}^{(2)}[X] \cap R_k), \quad \text{or}$$

3. $\bar{R}_k \cap \partial\bar{Z} \neq \emptyset$.

Let $N(R)$ be the number of slices for which $\bar{R}_k \cap \partial\bar{Z} \neq \emptyset$. By the cases elucidated above, letting

$$R_{\min} := \min_k \Big\{\mathrm{vol}_{n,r}(\mathcal{C}^{(1)}[X] \cap R_k) \wedge \mathrm{vol}_{n,r}(\mathcal{C}^{(2)}[X] \cap R_k)\Big\}$$

we obtain the following lower bound on the volume of the symmetric set difference,

$$\mathrm{vol}_{n,r}(Z \triangle \mathcal{C}^{(1)}[X]) \geq \frac{1}{2}R_{\min}\left[\frac{2\rho}{r} - N(R)\right] \Longleftrightarrow N(R) \geq 2\Big(\frac{\rho}{r} - \frac{\mathrm{vol}_{n,r}(Z \triangle \mathcal{C}^{(1)}[X])}{R_{\min}}\Big) \tag{A.78}$$

Finally note that $\left|\partial \bar{Z}\right| \geq N(R)$. Therefore combining (A.77) and (A.78), we have that

$$\begin{aligned}\mathrm{cut}_{n,r}(Z) &\geq \frac{1}{4} N(R) Q_{\min}^2 \\ &\geq \frac{1}{2}\left(\frac{\rho}{r} - \frac{\mathrm{vol}_{n,r}(Z \triangle \mathcal{C}^{(1)}[X])}{R_{\min}}\right) Q_{\min}^2 \end{aligned} \tag{A.79}$$

for all $Z \subset X$.

It remains to lower bound the random quantities $R_{\min}$ and $Q_{\min}$. To do so, we first lower bound the expected probability of any cell Q,

$$\min_{Q \in \overline{Q}} \mathbb{P}(Q) \geq \frac{\epsilon r^2}{\rho \sigma}. \tag{A.80}$$

and the expected volume of $\mathcal{C}^{(1)}[X] \cap R_k$ and $\mathcal{C}^{(2)}[X] \cap R_k$,

$$\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(1)} \cap R_k) = \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}^{(2)} \cap R_k) \geq \frac{\pi r^3}{2\sigma\rho^2} \quad \text{for all } k. \tag{A.81}$$

Since $Q_{\min}$ and $R_{\min}$ are obtained by taking the minimum of functionals over a fixed number of sets in n, they concentrate tightly around their means. Specifically, note that the total number of cubes is $\left|\overline{Q}\right| = \frac{12\sigma\rho}{r^2}$, and the total number of horizontal slices is $\frac{2\rho}{r}$. Along with (A.80) and (A.81), by Lemma 16

$$Q_{\min} \geq (1-\delta)\frac{\epsilon r^2}{\rho\sigma} \quad \text{and} \quad R_{\min} \geq (1-\delta)\frac{\pi r^3}{2\sigma\rho^2},$$

with probability at least $1 - \frac{12\sigma\rho}{r^2}\exp\left\{-\frac{n\delta^2\epsilon r^2 n}{4\rho\sigma}\right\} - \frac{2\rho}{r}\exp\left\{-\frac{n\delta^2\pi r^3}{8\sigma\rho^2}\right\}$. Combining these lower bounds with (A.76) and (A.79), we obtain

$$\Phi_{n,r}(Z) \geq \frac{(1-\delta)^2}{2(1+\delta)\pi}\left(1 - 2\frac{\sigma\rho}{(1-\delta)r^2n^2}\mathrm{vol}_{n,r}(Z \triangle \mathcal{C}^{(1)}[X])\right)\frac{\epsilon^2 r}{\sigma},$$

for all $Z \subseteq X$.

A.5 Additional results: aPPR and Consistency of PPR

In this appendix, we prove two additional results regarding PPR remarked upon in our main text. In Section A.5.1, we show that clustering using the aPPR vector satisfies an equivalent guarantee to Theorem 3. In Section A.5.2, we show that the PPR vector can perfectly distinguish two distinct density clusters $\mathcal{C}_\lambda, \mathcal{C}'_\lambda$.

A.5.1 Generic cluster recovery with aPPR

Our formal claim regarding cluster recovery with aPPR is contained in Corollary 4.

Corollary 4. *Consider instead of Algorithm 1 using the approximate PPR vector from Andersen et al. [2006] satisfying (2.2), and forming the corresponding cluster estimate $\widehat{C}$ in the same manner. Then provided we take*

$$\varepsilon = \frac{1}{25(1+\delta)n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C})}, \tag{A.82}$$

under the assumptions of Theorem 3 the upper bound on symmetric set difference in (2.26) still holds.

Proof of Corollary 4. Note that the choice of ε in (A.82) implies $\varepsilon \leq 1/\big(25\mathrm{vol}_{n,r}(\mathcal{C}[X])\big)$ with probability at least $1-\exp\big\{-n\delta^2\mathrm{vol}_{\mathbb{P},r}(\mathcal{C})\big\}$. The proof of Corollary 4 is then identical to that of Theorem 3, except one uses Corollary 2 rather than Lemma 1 to relate the symmetric set difference to the graph normalized cut and mixing time.

A.5.2 Perfectly distinguishing two density clusters

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

Definition A.5.1. For an estimator $\widehat{C} \subseteq X$ and distinct clusters $\mathcal{C}_\lambda, \mathcal{C}'_\lambda \in \mathbb{C}_f(\lambda)$, we say $\widehat{C}$ *separates* $\mathcal{C}_\lambda$ from $\mathcal{C}'_\lambda$ if

$$\mathcal{C}_\lambda[X] \subseteq \widehat{C} \quad \text{and} \quad \widehat{C} \cap \mathcal{C}'_\lambda[X] = \emptyset. \tag{A.83}$$

The bound on symmetric set difference (2.31) does not imply (A.83), which requires a uniform bound over the PPR vector p_v. As an example, suppose that we were able to show that for all $\mathcal{C}' \in \mathbb{C}_f(\lambda), \mathcal{C}' \neq \mathcal{C}$, and each $u \in \mathcal{C}, u' \in \mathcal{C}'$,

$$\frac{p_v(u')}{\deg(u';G)} \leq \frac{1}{10} \cdot \frac{1}{n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})} < \frac{1}{5} \cdot \frac{1}{n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})} \leq \frac{p_v(u)}{\deg(u;G)}. \tag{A.84}$$

Then, any (L,U) satisfying (2.22) and any sweep cut S_β for $\beta \in (L,U)$ would fulfill both conditions laid out in (A.83). In Theorem 26, we show that a sufficiently small upper bound on $\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])$ ensures that with high probability the uniform bound (A.84) is satisfied, and hence implies that $\widehat{C}$ will separate $\mathcal{C}_\lambda$ from $\mathcal{C}'_\lambda$. In what follows, put

$$c_{1,\delta} := \frac{(1-\delta)^5}{4} \cdot \min\Big\{\Big(\frac{3}{8(1+\delta)} - \frac{1}{5}\Big), \frac{1}{10}\Big\}$$

and note that if $\delta \in (0, 7/8)$ then $c_{1,\delta} > 0$. (In fact, we will have to take $\delta \in (0, 1/4)$ in order to use Propositions 2 and 3). Additionally, denote $\mathbb{P}'$ for the conditional distribution of a sample point given that it falls in $\mathcal{C}'_{\lambda,\sigma}$, i.e. $\mathbb{P}'(\mathcal{S}) := \mathbb{P}(\mathcal{S} \cap \mathcal{C}'_{\lambda,\sigma})/\mathbb{P}(\mathcal{C}'_{\lambda,\sigma})$, and $G'_{n,r} := G_{n,r}[\mathcal{C}'_{\lambda,\sigma}]$ for the subgraph of $G_{n,r}$ induced by $\mathcal{C}'_{\lambda,\sigma}$.

Theorem 26. *For any $\delta \in (0, 1/4)$ any $n \in \mathbb{N}$ such that*

$$\frac{1}{n} \leq \delta \cdot \frac{4\mathbb{P}(\mathcal{C}'_{\lambda,\sigma})}{3} \tag{A.85}$$

and otherwise under the same conditions as Theorem 3, the following statement holds with probability at least $1 - B_2/n - 4\exp\{-b_1\delta^2 n\} - (n+2)\exp\{-b_3 n\} - 3(n+3)\exp\{-b_7\delta^2 n\}$: there exists a set $\mathcal{C}_{\lambda,\sigma}[X]^g \subseteq \mathcal{C}_{\lambda,\sigma}[X]$ of large volume, $\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X]^g) \geq \mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])/2$, such that if Algorithm 1 is δ-well-initialized and run with any seed node $v \in \mathcal{C}_{\lambda,\sigma}[X]^g$, and moreover

$$\kappa_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}, \delta) \leq c_{1,\delta} \cdot \frac{\min\big\{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})^2 \cdot d_{\min}(\widetilde{\mathbb{P}})^2, \mathbb{P}(\mathcal{C}'_{\lambda,\sigma})^2 \cdot d_{\min}(\mathbb{P}')^2\big\}}{\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})} \tag{A.86}$$

then the PPR estimated cluster $\widehat{C}$ satisfies (A.83).

Before we prove Theorem 26, we make a few brief remarks:

- In one sense, Theroem 26 is a strong result: if the density clusters $\mathcal{C}_\lambda, \mathcal{C}'_\lambda$ satisfies the requirement (A.86), and we are willing to ignore the behavior of the algorithm in low-density regions, Theorem 26 guarantees that PPR will *perfectly distinguish* the candidate cluster $\mathcal{C}_\lambda$ from $\mathcal{C}'_\lambda$.

- On the other hand, unfortunately the requirement (A.86) is rather restrictive. Suppose the density cluster $\mathcal{C}_{\lambda,\sigma}$ satisfies (A3). Then from the following chain of inequalities,

$$\frac{\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])}{\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])} \overset{\text{(Thm. 3)}}{\leq} \kappa_{\mathbb{P},r}(\mathcal{C}, \delta) \overset{\text{(A.86)}}{\leq} c_{1,\delta} \cdot \frac{\mathbb{P}(\mathcal{C}_{\lambda,\sigma})^2 \cdot d_{\min}(\widetilde{\mathbb{P}})^2}{\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})} \overset{\text{(A3)}}{\leq} c_{1,\delta} \cdot \frac{\Lambda_\sigma}{\lambda_\sigma} \nu_d r^d,$$

  we see that in order for (A.86) to be met, it is necessary that $\Delta(\widehat{C}, \mathcal{C}_{\lambda,\sigma}[X])/\mathrm{vol}_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])$ be on the order of r^d. In plain terms, we are able to recover a density cluster $\mathcal{C}_\lambda$ in the strong sense of (A.83) only when we can guarantee the volume of the symmetric set difference will be very small. This strong condition is the price we pay in order to obtain the uniform bound in (A.84).

- The proof of Theorem 26 relies heavily on Lemma 7. This lemma—or more accurately, the equation (A.32) used in the proof of the lemma—can be thought of as a smoothness result for the PPR vector, showing that the mass of $p_v(\cdot)$ cannot be overly concentrated at any one vertex $u \in V$. However, (A.32) is a somewhat crude bound. By plugging a stronger result on the smoothness of $p_v(\cdot)$ in to the proof of Lemma 7, we could improve the uniform bounds of the lemma, and in turn show that the conclusion of Theorem 26 holds under weaker conditions than (A.86).

Now we recall some probabilistic estimates before proceeding to the proof of Theorem 26.

Probabilistic estimates. As in the proof of Theorem 3, we will assume that the inequalities (2.17)-(2.21) and (A.66)-(A.68) are satisfied. We will additionally assume that

$$n' \geq (1-\delta) \cdot n \cdot \mathbb{P}(\mathcal{C}'_{\lambda,\sigma}) \overset{\text{(A.85)}}{\Longrightarrow} n' - 1 \geq (1-\delta)^2 \cdot n \cdot \mathbb{P}(\mathcal{C}'_{\lambda,\sigma}) \tag{A.87}$$

and that

$$\frac{1}{n'-1} d_{\min}(G'_{n,r}) \geq (1-\delta) \cdot d_{\min}(\mathbb{P}'). \tag{A.88}$$

By Propositions 2-3 and Lemmas 14-16, these inequalities hold with probability at least $1 - B_2/n - 4\exp\{-b_1\delta^2 n\} - (n+1)\exp\{-b_3 n\} - (3n+3)\exp\{-b_7\delta^2 n\}$, taking $b_7 := b_2 \wedge \mathbb{P}(\mathcal{C}'_{\lambda,\sigma}) \cdot d_{\min}(\mathbb{P}')/9$.

Proof of Theorem 26. We have already verified in the proof of Theorem 3 that $\alpha \leq 1/(2\tau_\infty(\widetilde{G}_{n,r}))$, and we may therefore apply Lemma 7, which gives an upper bound on $p_v(u)$ for all $u \in \mathcal{C}_{\lambda,\sigma}[X]_o$ and a lower bound on $p_v(u')$ for all $p_v(u')$ for all $u' \in \mathcal{C}'_{\lambda,\sigma}[X]_o$. These bounds are useful because $r \leq \sigma$, which implies that $\mathcal{C}_\lambda[X] \subseteq \mathcal{C}_{\lambda,\sigma}[X]_o$ and likewise that $C'_\lambda[X] \subseteq \mathcal{C}'_{\lambda,\sigma}[X]_o$. We will show that these bounds in turn imply (A.84), from which the claim of the theorem follows.

We begin with the lower bound in (A.84). From (in order) Lemma 7, our assorted probabilistic estimates, and the assumed lower bound (A.86) on $\kappa_{\mathbb{P},r}(\mathcal{C}, \delta)$, we have that for all $u \in \mathcal{C}_\lambda[X]$,

$$\begin{aligned}
\frac{p_v(u)}{\deg_{n,r}(u)} &\geq \frac{3}{8\mathrm{vol}_{n,r}(\mathcal{C}[X])} - 2\frac{\Phi_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])}{d_{\min}(\widetilde{G}_{n,r})^2\alpha} \\
&\geq \frac{1}{n(n-1)}\left(\frac{3}{8(1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})} - 4 \cdot \frac{n(n-1)}{(\widetilde{n}-1)^2} \cdot \frac{\kappa_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}, \delta)}{(1-\delta)d_{\min}(\widetilde{\mathbb{P}})^2}\right) \\
&\geq \frac{1}{n(n-1)}\left(\frac{3}{8(1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})} - 4 \cdot \frac{\kappa_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma}, \delta)}{(1-\delta)^5\mathbb{P}(\mathcal{C})^2 d_{\min}(\widetilde{\mathbb{P}})^2}\right) \\
&\geq \frac{1}{5n^2\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})}.
\end{aligned}$$

An equivalent derivation implies the upper bound in (A.84): for all $u' \in \mathcal{C}'_\lambda[X]$,

$$\begin{aligned}\frac{p_v(u')}{\deg_{n,r}(u')} &\leq 2\frac{\Phi_{n,r}(\mathcal{C}_{\lambda,\sigma}[X])}{d_{\min}(G'_{n,r})^2\alpha} \\ &\leq 4\cdot\frac{1}{(n'-1)^2}\cdot\frac{\kappa(\mathcal{C}_{\lambda,\sigma},\delta)}{(1-\delta)d_{\min}(\mathbb{P}')^2} \\ &\leq 4\frac{\kappa(\mathcal{C}_{\lambda,\sigma},\delta)}{n^2(1-\delta)^5\mathbb{P}(\mathcal{C}'_{\lambda,\sigma})^2 d_{\min}(\mathbb{P}')^2} \leq \frac{1}{10n^2\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_{\lambda,\sigma})},\end{aligned}$$

completing the proof of Theorem 26.

A.6 Experimental Details

Finally, we detail the settings of our experiments, and include an additional figure.

A.6.1 Experimental settings for Figure 2.2

Let $\mathcal{R}_{\sigma,\rho} = [-\sigma/2,\sigma/2]\times[-\rho/2,\rho/2]$ be the two-dimensional rectangle of width σ and height ρ, centered at the origin. We sample $n = 8000$ points according to the density function $f_{\rho,\sigma,\lambda}$, defined over domain $\mathcal{X} = [-1,1]^2$ and parameterized by ρ,σ and λ as follows:

$$f_{\rho,\sigma,\lambda}(x) := \begin{cases}\lambda, & \text{if } x \in R_{\sigma,\rho} - (-.5,0) \text{ or } x \in R_{\sigma,\rho} + (-.5,0) \\ \dfrac{4-2\lambda\rho\sigma}{1-2\rho\sigma}, & \text{if } x \in \mathcal{X},\ x \notin R_{\sigma,\rho} - (-.5,0) \text{ and } x \notin R_{\sigma,\rho} + (-.5,0).\end{cases} \tag{A.89}$$

Then $\theta := \lambda - \frac{4-2\lambda\rho\sigma}{1-2\rho\sigma}$ measures the difference in density between the density clusters and the rest of the domain. The first column displays $n = 8000$ points sampled from three different parameterizations of $f_{\rho,\sigma,\lambda}$:

$\rho = .913,$	$\sigma = .25,$	$(\lambda-\theta)/\lambda = .25$	(top panel)
$\rho = .25,$	$\sigma =,$	$(\lambda-\theta)/\lambda = .05$	(middle panel)
$\rho = .5,$	$\sigma = .25,$	$(\lambda-\theta)/\lambda = .12$	(bottom panel.)

In each of the first, second, and third rows, we fix two parameters and vary the third. In the first row, we fix $\sigma = .25$, $(\lambda-\theta)/\lambda = .25$, and vary ρ from .25 to 2. In the second row, we fix $\rho = 1.8$, $(\lambda-\theta)/\lambda = .05$, and vary σ from .1 to .2 In the third row, we fix $\rho = .5$, $\sigma = .25$ and vary $(\lambda-\theta)/\lambda$ from .1 to .25. In the first and third rows, we take $r = \sigma/8$; in the second row, where we vary σ, we take $r = .1/8$.

A.6.2 Experimental settings for Figure 2.3

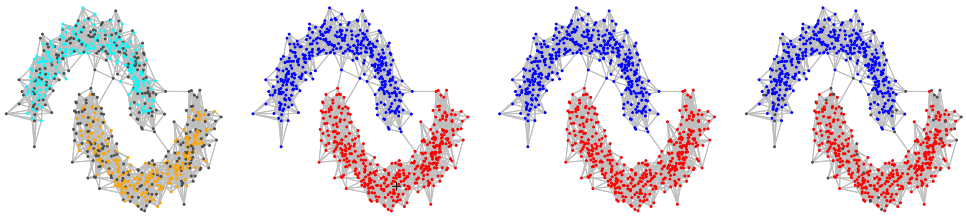
To form each of the three rows in Figure 2.3, $n = 800$ points are independently sampled following a 'two moons plus Gaussian noise model'. Formally, the (respective) generative models for the data are

$$Z \sim \mathrm{Bern}(1/2), \theta \sim \mathrm{Unif}(0,\pi) \tag{A.90}$$

$$X(Z,\theta) = \begin{cases}\mu_1 + (r\cos(\theta), r\sin(\theta)) + \sigma\epsilon, & \text{if } Z = 1 \\ \mu_2 + (r\cos(\theta), -r\sin(\theta)) + \sigma\epsilon, & \text{if } Z = 0\end{cases} \tag{A.91}$$

where

$$\mu_1 = (-.5,0),\ \mu_2 = (0,0),\ \epsilon \sim N(0,I_2) \quad \text{(row 1)}$$
$$\mu_1 = (-.5,-.07),\ \mu_2 = (0,.07),\ \epsilon \sim N(0,I_2) \quad \text{(row 2)}$$
$$\mu_1 = (-.5,-.125),\ \mu_2 = (0,.125),\ \epsilon \sim N(0,I_2) \quad \text{(row 3)}$$



(a) (b) (c) (d)

Figure A.1: True density (column 1), PPR (column 2), minimum normalized cut (column 3) and estimated density (column 4) clusters for two-moons with 10 dimensional noise. Seed node for PPR denoted by a black cross.

for I_d the $d \times d$ identity matrix. In all cases $\sigma = .07$. In each case λ is taken as small as possible such that there exist exactly two distinct density clusters, which we call $\mathcal{C}_\lambda$ and $\mathcal{C}'_\lambda$; r is taken as small as possible so that each vertex has at least 2 neighbors. The first column consists of the empirical density clusters $\mathcal{C}_\lambda[X]$ and $\mathcal{C}'_\lambda[X]$ for a particular threshold λ of the density function; the second column shows the PPR plus minimum normalized sweep cut cluster, with hyperparameter α and all sweep cuts considered; the third column shows the global minimum normalized cut, computed according to the algorithm of Bresson et al. [2012]; and the last column shows a cut of the density cluster tree estimator of Chaudhuri and Dasgupta [2010].

Performance of PPR with high-dimensional noise. Figure A.1 is similar to Figure 2.3 of the main text, but with parameters

$$\mu_1 = (-.5, -.025),\ \mu_2 = (0, .025),\ \epsilon \sim N(0, I_{10}).$$

The gray dots in (a) (as in the left-hand column of Figure 2.3 in the main text) represent observations in low-density regions. While the PPR sweep cut (b) has relatively high symmetric set difference with the chosen density cut, it still recovers separates $C_\lambda[X]$ and $\mathcal{C}'_\lambda[X]$, in the sense of Definition A.5.1.

Appendix B

Chapter 3 Appendix

B.1 Preliminaries

In the appendix, we provide complete proofs of all results. Our main theorems (Theorems 6-10) all follow the same general proof strategy of first establishing bounds in the fixed-design setup. In Section B.2, we establish (estimation or testing) error bounds which hold for any graph G; these bounds are stated with respect to (functionals of) the graph G, and allow us to upper bound the error of $\widehat{f}$ and $\widehat{\varphi}$ conditional on the design $\{X_1, \ldots, X_n\} = \{x_1, \ldots, x_n\}$. In Sections B.3, B.4, B.5, and B.6 we develop all the necessary probabilistic estimates on these functionals, for the particular random neighborhood graph $G = G_{n,r}$. It is in these sections where we invoke our various assumptions on the distribution P and regression function f_0. In Section B.7, we prove our main theorems and some other results. In Section B.8, we state a few concentration bounds that we use repeatedly in our proofs.

Pointwise evaluation of Sobolev functions. First, however, as promised in our main text we clarify what is meant by pointwise evaluation of the regression function f_0. Strictly speaking, each $f \in H^1(X)$ is really an equivalence class, defined only up to sets of Lebesgue measure 0. In order to make sense of the evaluation $x \mapsto f(x)$, one must therefore pick a representative $f^\star \in f$. When $d = 1$, this is resolved in a standard way—since $H^1(\mathcal{X})$ embeds continuously into $C^0(\mathcal{X})$, there exists a continuous version of every $f \in H^1(\mathcal{X})$, and we take this continuous version as the representative $f^\star$. On the other hand, when $d \geq 2$, the Sobolev space $H^1(\mathcal{X})$ does not continuously embed into $C^0(\mathcal{X})$, and we must choose representatives in a different manner. In this case we let $f^\star$ be the precise representative [Evans and Gariepy, 2015], defined pointwise at points $x \in \mathcal{X}$ as

$$f^\star(x) = \begin{cases} \lim_{\varepsilon \to 0} \dfrac{1}{\nu(B(x,\varepsilon))} \displaystyle\int_{B(x,\varepsilon)} f(z)dz, & \text{if the limit exists,} \\ 0, & \text{otherwise.} \end{cases}$$

Note that when $d = 1$, the precise representative of any $f \in H^1(\mathcal{X})$ is continuous.

Now we explain why the particular choice of representative is not crucial, using the notion of a Lebesgue point. Recall that for a locally Lebesgue integrable function f, a given point $x \in \mathcal{X}$ is a *Lebesgue point* of f if the limit of $1/(\nu(B(x,\varepsilon))) \int_{B(x,\varepsilon)} f(x)dx$ as $\varepsilon \to 0$ exists, and satisfies

$$\lim_{\varepsilon \to 0} \frac{1}{\nu\big(B(x,\varepsilon)\big)} \int_{B(x,\varepsilon)} f(x)dx = f(x).$$

Let E denote the set of Lebesgue points of f. By the Lebesgue differentiation theorem [Evans and Gariepy, 2015], if $f \in L^1(\mathcal{X})$ then almost every $x \in \mathcal{X}$ is a Lebesgue point, $\nu(\mathcal{X} \setminus E) = 0$. Since $f_0 \in H^1(\mathcal{X}) \subseteq L^1(\mathcal{X})$, we can conclude that any function $g_0 \in f_0$ disagrees with the precise representative $f_0^\star$ only on a set of Lebesgue measure 0. Moreover, since we always assume the design distribution P has a continuous density, with probability 1 it holds that $g_0(X_i) = f_0^\star(X_i)$ for all $i = 1, \dots, n$. This justifies the notation $f_0(X_i)$ used in the main text.

B.2 Graph-dependent error bounds

In this section, we adopt the fixed design perspective; or equivalently, condition on $X_i = x_i$ for $i = 1, \dots, n$. Let $G = \big([n], W\big)$ be a fixed graph on $\{1, \dots, n\}$ with Laplacian matrix $\mathbf{L} = D - W$. The randomness thus all comes from the responses

$$Y_i = f_0(x_i) + \varepsilon_i \tag{B.1}$$

where the noise variables ε_i are independent $N(0, 1)$. In the rest of this section, we will mildly abuse notation and write $f_0 = (f_0(x_1), \dots, f_0(x_n)) \in \mathbb{R}^n$. We will also write $\mathbf{Y} = (Y_1, \dots, Y_n)$.

Recall (3.2) and (3.3): the Laplacian smoothing estimator of f_0 on G is

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathbb{R}^n} \bigg\{ \sum_{i=1}^n (Y_i - f_i)^2 + \rho \cdot f^\top \mathbf{L} f \bigg\} = (\rho \mathbf{L} + I)^{-1} \mathbf{Y}.$$

and the Laplacian smoothing test statistic is

$$\widehat{T} := \frac{1}{n} \|\widehat{f}\|_2^2.$$

We note that in this section, many of the derivations involved in upper bounding the estimation error of $\widehat{f}$ are similar to those of Sadhanala et al. [2016a], with the difference being that we seek bounds in high probability rather than in expectation. We keep the work here self-contained for purposes of completeness.

B.2.1 Error bounds for linear smoothers

Let $S \in \mathbb{R}^{n \times n}$ be a fixed square, symmetric matrix, and let

$$\breve{f} := SY$$

be a linear estimator of f_0. In Lemma 24 we upper bound the error $\frac{1}{n}\|\breve{f} - f_0\|_2^2$ as a function of the eigenvalues of S. Let $\boldsymbol{\lambda}(S) = (\lambda_1(S), \dots, \lambda_n(S)) \in \mathbb{R}^n$ denote these eigenvalues, and let $v_k(S)$ denote the corresponding unit-norm eigenvectors, so that $S = \sum_{k=1}^n \lambda_k(S) \cdot v_k(S) v_k(S)^\top$. Denote $Z_k = v_k(S)^\top \varepsilon$, and observe that $\mathbf{Z} = (Z_1, \dots, Z_n) \sim N(0, I)$.

Lemma 24. *Let $\breve{f} = SY$ for a square, symmetric matrix, $S \in \mathbb{R}^{n \times n}$. Then*

$$\mathbb{P}_{f_0}\bigg(\frac{1}{n}\big\|\breve{f} - f_0\big\|_2^2 \geq \frac{10}{n}\big\|\boldsymbol{\lambda}(S)\big\|_2^2 + \frac{2}{n}\big\|(S - I) f_0\big\|_2^2 \bigg) \leq 1 - \exp\Big(-\big\|\boldsymbol{\lambda}(S)\big\|_2^2\Big)$$

Here we have written $\mathbb{P}_{f_0}(\cdot)$ for the probability law under the regression "function" $f_0 \in \mathbb{R}^n$.

In Lemma 25, we upper bound the error of a test involving the statistic $\|\breve{f}\|_2^2 = \mathbf{Y}^\top S^2 \mathbf{Y}$. We will require that S be a *contraction*, meaning that it has operator norm no greater than 1, $\|Sv\|_2 \leq \|v\|_2$ for all $v \in \mathbb{R}^n$.

Lemma 25. *Let $\check{T} = \mathbf{Y}^\top S^2 \mathbf{Y}$ for a square, symmetric matrix $S \in \mathbb{R}^{n\times n}$. Suppose S is a contraction. Define the threshold $\check{t}_\alpha$ to be*

$$\check{t}_\alpha := \|\boldsymbol{\lambda}(S)\|_2^2 + \sqrt{\frac{2}{\alpha}}\|\boldsymbol{\lambda}(S)\|_4^2. \tag{B.2}$$

It holds that:

- ***Type I error.***

$$\mathbb{P}_0\big(\check{T} > \check{t}_\alpha\big) \le \alpha. \tag{B.3}$$

- ***Type II error.*** *Under the further assumption*

$$f_0^\top S^2 f_0 \ge \Big(2\sqrt{\frac{2}{\alpha}} + 2b\Big) \cdot \|\boldsymbol{\lambda}(S)\|_4^2, \tag{B.4}$$

then

$$\mathbb{P}_{f_0}\big(\check{T} \le \check{t}_\alpha\big) \le \frac{1}{b^2} + \frac{16}{b\|\boldsymbol{\lambda}(S)\|_4^2}. \tag{B.5}$$

Proof of Lemma 24. The expectation $\mathbb{E}_{f_0}[\check{f}] = Sf_0$, and by the triangle inequality,

$$\begin{aligned}\frac{1}{n}\big\|\check{f} - f_0\big\|_2^2 &\le \frac{2}{n}\Big(\big\|\check{f} - \mathbb{E}_{f_0}[\check{f}]\big\|_2^2 + \big\|\mathbb{E}_{f_0}[\check{f}] - f_0\big\|_2^2\Big)\\ &= \frac{2}{n}\Big(\big\|S\varepsilon\big\|_2^2 + \big\|(S-I)f_0\big\|_2^2\Big).\end{aligned}$$

Writing $\|S\varepsilon\|_2^2 = \sum_{k=1}^n \lambda_k(S)^2 Z_k^2$, the claim follows from the result of Laurent and Massart [2000] on concentration of χ^2-random variables, which for completeness we restate in Lemma 37. To be explicit, taking $t = \|\boldsymbol{\lambda}(S)\|_2^2$ in Lemma 37 completes the proof of Lemma 24.

Proof of Lemma 25. We compute the mean and variance of T as a function of f_0, then apply Chebyshev's inequality.

Mean. We make use of the eigendecomposition $S = \sum_{k=1}^n \lambda_k(S) \cdot v_k(S)v_k(S)^\top$ to obtain

$$\begin{aligned}\check{T} &= f_0^\top S^2 f_0 + 2f_0^\top S^2\varepsilon + \varepsilon^\top S^2\varepsilon\\ &= f_0^\top S^2 f_0 + 2f_0^\top S^2\varepsilon + \sum_{k=1}^n \big(\lambda_k(S)\big)^2(\varepsilon^\top v_k(S))^2\\ &= f_0^\top S^2 f_0 + 2f_0^\top S^2\varepsilon + \sum_{k=1}^n \big(\lambda_k(S)\big)^2 Z_k^2,\end{aligned} \tag{B.6}$$

implying

$$\mathbb{E}_{f_0}\big[\check{T}\big] = f_0^\top S^2 f_0 + \sum_{k=1}^n \big(\lambda_k(S)\big)^2. \tag{B.7}$$

Variance. We start from (B.6). Recalling that $\mathrm{Var}(Z_k^2) = 2$, it follows from the Cauchy-Schwarz inequality that

$$\mathrm{Var}_{f_0}\big[\check{T}\big] \le 8f_0^\top S^4 f_0 + 4\sum_{k=1}^n \big(\lambda_k(S)\big)^4. \tag{B.8}$$

Bounding Type I and Type II error. The upper bound (B.3) on Type I error follows immediately from (B.7), (B.8), and Chebyshev's inequality.

We now establish the upper bound (B.5) on Type II error. From assumption (B.4), we see that $f_0^\top S^2 f_0^\top - \check{t}_\alpha \leq 0$. As a result,

$$\begin{aligned}\mathbb{P}_{f_0}\Big(\check{T} \leq \check{t}_\alpha\Big) &= \mathbb{P}_{f_0}\Big(\check{T} - \mathbb{E}_{f_0}\big[\check{T}\big] \leq \check{t}_\alpha - \mathbb{E}_{f_0}\big[\check{T}\big]\Big) \\ &\leq \mathbb{P}_{f_0}\Big(\Big|\check{T} - \mathbb{E}_{f_0}\big[\check{T}\big]\Big| \geq \Big|\check{t}_\alpha - \mathbb{E}_{f_0}\big[\check{T}\big]\Big|\Big) \\ &\leq \frac{\mathrm{Var}_{f_0}\big[\check{T}\big]}{\big(\check{t}_\alpha - \mathbb{E}_{f_0}\big[\check{T}\big]\big)^2},\end{aligned}$$

where the last line follows from Chebyshev's inequality. Plugging in the expressions (B.7) and (B.8) for the mean and variance of $\check{T}$, as well as the definition of $\check{t}_\alpha$ in (B.2), we obtain that

$$\mathbb{P}_{f_0}\Big(\check{T} \leq \check{t}_\alpha\Big) \leq \frac{4\|\boldsymbol{\lambda}(S)\|_4^4}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2} + \frac{8 f_0^\top S^4 f_0}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2}. \tag{B.9}$$

We now use the assumed lower bound $f_0^\top S^2 f_0 \geq (2\sqrt{2/\alpha} + 2b)\|\boldsymbol{\lambda}(S)\|_4^2$ to separately upper bound each of the two terms on the right hand side of (B.9). It follows immediately that

$$\frac{4\|\boldsymbol{\lambda}(S)\|_4^4}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2} \leq \frac{1}{b^2}, \tag{B.10}$$

giving a sufficient upper bound on the first term. Now we upper bound the second term,

$$\frac{8 f_0^\top S^4 f_0}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2} \leq \frac{32 f_0^\top S^4 f_0}{\big(f_0^\top S^2 f_0\big)^2} \leq \frac{16}{b\|\boldsymbol{\lambda}(S)\|_4^2}\frac{f_0^\top S^4 f_0}{f_0^\top S^2 f_0} \leq \frac{16}{b\|\boldsymbol{\lambda}(S)\|_4^2}, \tag{B.11}$$

where the final inequality is satisfied because S is a contraction. Plugging (B.10) and (B.11) back into (B.9) then gives the desired result.

B.2.2 Analysis of Laplacian smoothing

Upper bounds on the mean squared error of $\widehat{f}$, and Type I and Type II error of $\widehat{T}$, follow from setting $S = (\rho L + I)^{-1}$ in Lemmas 24 and 25. We give these results in Lemma 26 and 27, and prove them immediately. Recall that $\lambda_1, \ldots, \lambda_n$ are the n eigenvalues of $\mathbf{L}$ (sorted in ascending order).

Lemma 26. *For any $\rho > 0$,*

$$\frac{1}{n}\big\|\widehat{f} - f_0\big\|_2^2 \leq \frac{2\rho}{n}\big(f_0^\top \mathbf{L} f_0\big) + \frac{10}{n}\sum_{k=1}^n \frac{1}{\big(\rho\lambda_k + 1\big)^2}, \tag{B.12}$$

with probability at least $1 - \exp\Big(-\sum_{k=1}^n \big(\rho\lambda_k + 1\big)^{-2}\Big)$.

Recall that

$$\widehat{t}_\alpha := \frac{1}{n}\sum_{k=1}^n \frac{1}{\big(\rho\lambda_k + 1\big)^2} + \frac{1}{n}\sqrt{\frac{2}{\alpha}\sum_{k=1}^n \frac{1}{\big(\rho\lambda_k + 1\big)^4}}.$$

Lemma 27. *For any $\rho > 0$ and any $b \geq 1$, it holds that:*

- ***Type I error.***

$$\mathbb{P}_0\Big(\widehat{T} > \widehat{t}_\alpha\Big) \leq \alpha. \tag{B.13}$$

- ***Type II error.*** *If*

$$\frac{1}{n}\|f_0\|_2^2 \geq \frac{2\rho}{n}\big(f_0^\top \mathbf{L} f_0\big) + \frac{2\sqrt{2/\alpha}+2b}{n}\left(\sum_{k=1}^n \frac{1}{(\rho\lambda_k+1)^4}\right)^{1/2}, \tag{B.14}$$

then

$$\mathbb{P}_{f_0}\Big(\widehat{T}(G) \leq \widehat{t}_\alpha\Big) \leq \frac{1}{b^2} + \frac{16}{b}\left(\sum_{k=1}^n \frac{1}{(\rho\lambda_k+1)^4}\right)^{-1/2}. \tag{B.15}$$

Proof of Lemma 26. Let $\widehat{S} = (I+\rho\mathbf{L})^{-1}$, the estimator $\widehat{f} = \widehat{S}Y$, and

$$\big\|\boldsymbol{\lambda}(\widehat{S})\big\|_2^2 = \sum_{k=1}^n \frac{1}{\big(1+\rho\lambda_k\big)^2}.$$

We deduce the following upper bound on the bias term,

$$\begin{aligned}
\big\|(\widehat{S}-I)f_0\big\|_2^2 &= f_0^\top \mathbf{L}^{1/2}\mathbf{L}^{-1/2}\big(\widehat{S}-I\big)^2\mathbf{L}^{-1/2}\mathbf{L}^{1/2}f_0 \\
&\leq f_0^\top \mathbf{L} f_0 \cdot \lambda_n\Big(\mathbf{L}^{-1/2}\big(\widehat{S}-I\big)^2\mathbf{L}^{-1/2}\Big) \\
&= f_0^\top \mathbf{L} f_0 \cdot \max_{k\in[n]}\left\{\frac{1}{\lambda_k}\Big(1-\frac{1}{\rho\lambda_k+1}\Big)^2\right\} \\
&\leq f_0^\top \mathbf{L} f_0 \cdot \rho.
\end{aligned}$$

In the above, we have written $\mathbf{L}^{-1/2}$ for the square root of the pseudoinverse of $\mathbf{L}$, the maximum is over all indices k such that $\lambda_k > 0$, and the last inequality follows from the basic algebraic identity $1-1/(1+x)^2 \leq 2x$ for any $x > 0$. The claim of the Lemma then follows from Lemma 24.

Proof of Lemma 27. Let $\widehat{S} := (I+\rho L)^{-1}$, so that $\widehat{T} = \frac{1}{n}\mathbf{Y}^\top \widehat{S}^2\mathbf{Y}$. Note that $\widehat{S}$ is a contraction, so that we may invoke Lemma 25. The bound on Type I error (B.13) follows immediately from (B.3). To establish the bound on Type II error, we must lower bound $f_0^\top \widehat{S}^2 f_0$. We first note that by assumption (B.14),

$$\begin{aligned}
f_0^\top \widehat{S}^2 f_0 &= \big\|f_0\big\|_2^2 - f_0^\top (I-\widehat{S}^2) f_0 \\
&\geq 2\rho\big(f_0^\top \mathbf{L} f_0\big) - f_0^\top \big(I-\widehat{S}^2\big) f_0 + \Big(2\sqrt{\frac{2}{\alpha}}+2b\Big)\cdot\left(\sum_{k=1}^n \frac{1}{(\rho\lambda_k+1)^4}\right)^{-1/2}.
\end{aligned}$$

Upper bounding $f_0^\top \big(I-\widehat{S}^2\big) f_0$ as follows:

$$\begin{aligned}
f_0^\top \Big(I-\widehat{S}^2\Big) f_0 &= f_0^\top \mathbf{L}^{1/2}\mathbf{L}^{-1/2}\Big(I-\widehat{S}^2\Big)\mathbf{L}^{-1/2}\mathbf{L}^{1/2} f_0 \\
&\leq f_0^\top \mathbf{L} f_0 \cdot \lambda_n\Big(\mathbf{L}^{-1/2}\Big(I-\widehat{S}^2\Big)\mathbf{L}^{-1/2}\Big) \\
&= f_0^\top \mathbf{L} f_0 \cdot \max_k\left\{\frac{1}{\lambda_k}\Big(1-\frac{1}{(\rho\lambda_k+1)^2}\Big)\right\} \\
&\leq f_0^\top \mathbf{L} f_0 \cdot 2\rho,
\end{aligned}$$

—where in the above the maximum is over all indices k such that $\lambda_k > 0$—we deduce that

$$f_0^\top \widehat{S}^2 f_0 \geq \Big(2\sqrt{\frac{2}{\alpha}}+2b\Big)\cdot\left(\sum_{k=1}^n \frac{1}{(\rho\lambda_k+1)^4}\right)^{-1/2}.$$

The upper bound on Type II error (B.15) then follows from Lemma 25.

B.3 Neighborhood graph Sobolev semi-norm

In this section, we prove Lemma 3, which states an upper bound on $f^\top L f$ that holds when f is bounded in Sobolev norm. We also establish stronger bounds in the case when f has a bounded Lipschitz constant; this latter result justifies one of our remarks after Theorem 6.

Throughout this proof, we will assume that $f \in H^1(\mathcal{X})$ has zero-mean, meaning $\int_{\mathcal{X}} f(x)\,dx = 0$. This is without loss of generality—assuming for the moment that (3.14) holds for zero-mean functions, for any $f \in H^1(\mathcal{X})$, taking $a = \int_{\mathcal{X}} f(x)\,dx$ and $g = f - a$, we have that

$$f^\top L f = g^\top L g \leq \frac{C_2}{\delta} n^2 r^{d+2} |g|^2_{H^1(\mathcal{X})} = \frac{C_2}{\delta} n^2 r^{d+2} |f|^2_{H^1(\mathcal{X})}.$$

Now, for any zero-mean function $f \in H^1(\mathcal{X})$ it follows by the Poincare inequality (see Section 5.8, Theorem 1 of Evans [2010]) that $\|f\|^2_{H^1(\mathcal{X})} \leq C_8 |f|^2_{H^1(\mathcal{X})}$, for some constant C_8 that does not depend on f. Therefore, to prove Lemma 3, it suffices to show that

$$\mathbb{E}\Big[f^\top L f\Big] \leq C n^2 r^{d+2} \|f\|^2_{H^1(\mathcal{X})},$$

since the high-probability upper bound then follows immediately by Markov's inequality. (Recall that L is positive semi-definite, and therefore $f^\top L f$ is a non-negative random variable).

Since

$$f^\top L f = \frac{1}{2} \sum_{i,j=1}^{n} \big(f(X_i) - f(X_j)\big)^2 W_{ij},$$

it follows that

$$\mathbb{E}\Big[f^\top L f\Big] = \frac{n(n-1)}{2} \mathbb{E}\bigg[\Big(f(X') - f(X)\Big)^2 K\bigg(\frac{\|X' - X\|}{r}\bigg)\bigg], \tag{B.16}$$

where X and X' are random variables independently drawn from P.

Now, take Ω to be an arbitrary bounded open set such that $B(x, c_0) \subseteq \Omega$ for all $x \in \mathcal{X}$. For the remainder of this proof, we will assume that (i) $f \in H^1(\Omega)$ and additionally (ii) $\|f\|_{H^1(\Omega)} \leq C_5 \|f\|_{H^1(\mathcal{X})}$ for a constant C_5 that does not depend on f. This is without loss of generality, since by Theorem 1 in Chapter 5.4 of Evans [2010] there exists an extension operator $E : H^1(\mathcal{X}) \to H^1(\Omega)$ for which the extension Ef satisfies both (i) and (ii). Additionally, we will assume $f \in C^\infty(\Omega)$. Again, this is without loss of generality, as $C^\infty(\Omega)$ is dense in $H^1(\Omega)$ and the expectation on the right hand side of (B.16) is continuous in $H^1(\Omega)$. The reason for dealing with a smooth extension $f \in C^\infty(\Omega)$ is so that we can make sense of the following equality for any x and x' in $\mathcal{X}$:

$$f(x') - f(x) = \int_0^1 \nabla f\big(x + t(x' - x)\big)^\top (x' - x)\,dt. \tag{B.17}$$

Obviously

$$\mathbb{E}\bigg[\big(f(X') - f(X)\big)^2 K\bigg(\frac{\|X' - X\|}{r}\bigg)\bigg] \leq p^2_{\max} \int_{\mathcal{X}} \int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 K\bigg(\frac{\|x' - x\|}{r}\bigg)\,dx'\,dx, \tag{B.18}$$

so that it remains now to bound the double integral. Replacing difference by integrated derivative as in

(B.17), we obtain

$$\begin{aligned}\int_{\mathcal{X}}\int_{\mathcal{X}}\big(f(x')-f(x)\big)^2 K\Big(\frac{\|x'-x\|}{r}\Big)\,dx'\,dx &= \int_{\mathcal{X}}\int_{\mathcal{X}}\Big[\int_0^1 \nabla f\big(x+t(x'-x)\big)^\top (x'-x)\,dt\Big]^2 K\Big(\frac{\|x'-x\|}{r}\Big)\,dx'\,dx\\ &\overset{(i)}{\le} \int_{\mathcal{X}}\int_{\mathcal{X}}\int_0^1\Big[\nabla f\big(x+t(x'-x)\big)^\top (x'-x)\Big]^2 K\Big(\frac{\|x'-x\|}{r}\Big)\,dt\,dx'\,dx\\ &\overset{(ii)}{\le} r^{d+2}\int_{\mathcal{X}}\int_{B(0,1)}\int_0^1\Big[\nabla f\big(x+trz\big)^\top z\Big]^2 K\big(\|z\|\big)\,dt\,dz\,dx\\ &\overset{(iii)}{\le} r^{d+2}\int_{\Omega}\int_{B(0,1)}\int_0^1\Big[\nabla f\big(\widetilde{x}\big)^\top z\Big]^2 K\big(\|z\|\big)\,dt\,dz\,d\widetilde{x}, \end{aligned}\tag{B.19}$$

where (i) follows by Jensen's inequality, (ii) follows by substituting $z=(x'-x)/r$ and (K1), and (iii) by exchanging integrals, substituting $\widetilde{x}=x+trz$, and noting that $x\in\mathcal{X}$ implies that $\widetilde{x}\in\Omega$.

Now, writing $\big(\nabla f(\widetilde{x})^\top z\big)^2=\big(\sum_{i=1}^d z_i f^{(e_i)}(x)\big)^2$, expanding the square and integrating, we have that for any $\widetilde{x}\in\mathcal{X}$,

$$\begin{aligned}\int_{B(0,1)}\Big[\nabla f\big(\widetilde{x}\big)^\top z\Big]^2 K\big(\|z\|\big)\,dz &= \sum_{i,j=1}^d f^{(e_i)}(\widetilde{x})f^{(e_j)}(\widetilde{x})\int_{\mathbb{R}^d} z_i z_j K(\|z\|)\,dz\\ &= \sum_{i=1}^d \big(f^{(e_i)}(\widetilde{x})\big)^2\int_{B(0,1)} z_i^2 K\big(\|z\|\big)\,dz\\ &= \sigma_K\|\nabla f(\widetilde{x})\|^2,\end{aligned}$$

where the last equality follows from the rotational symmetry of $K(\|z\|)$. Plugging back into (B.19), we obtain

$$\int_{\mathcal{X}}\int_{\mathcal{X}}\big(f(x')-f(x)\big)^2 K\Big(\frac{\|x'-x\|}{r}\Big)\,dx'\,dx \le r^{d+2}\sigma_K\|f\|^2_{H^1(\Omega)} \le C_5 r^{d+2}\sigma_K\|f\|^2_{H^1(\mathcal{X})},$$

proving the claim of Lemma 3 upon taking $C_2 := C_8C_5\sigma_K p_{\max}^2$ in the statement of the lemma.

B.3.1 Stronger bounds under Lipschitz assumption

Suppose f satisfies $|f(x')-f(x)|\le M\|x-x\|$ for all $x,x'\in\mathcal{X}$. Then we can strengthen the high probability bound in Lemma 3 from $1-\delta$ to $1-\delta^2/n$, at the cost of only a constant factor in the upper bound on $f^\top Lf$.

Proposition 19. *Let $r\ge C_0(\log n/n)^{1/d}$. For any f such that $|f(x')-f(x)|\le M\|x-x\|$ for all $x,x'\in\mathcal{X}$, with probability at least $1-C\delta^2/n$ it holds that*

$$f^\top Lf \le \Big(\frac{1}{\delta}+C_2\Big)n^2 r^{d+2}M^2.$$

Proof of Proposition 19. We will prove Proposition 19 using Chebyshev's inequality, so the key step is to upper bound the variance of $f^\top Lf$. Putting $\Delta_{ij}:=K(\|X_i-X_j\|/r)\cdot(f(X_i)-f(X_j))^2$, we can write the variance of $f^\top Lf$ as a sum of covariances,

$$\mathrm{Var}\big[f^\top Lf\big]=\frac{1}{4}\sum_{i,j=1}^n\sum_{\ell,m=1}^n \mathrm{Cov}\big[\Delta_{ij},\Delta_{\ell m}\big].$$

Clearly $\mathrm{Cov}\big[\Delta_{ij},\Delta_{\ell m}\big]$ depends on the cardinality of $I:=\{i,j,k,\ell\}$; we divide into cases, and upper bound the covariance in each case.

$|I| = 4$. In this case Δ_{ij} and $\Delta_{\ell m}$ are independent, and $\mathrm{Cov}\big[\Delta_{ij}, \Delta_{\ell m}\big] = 0$.

$|I| = 3$. Taking $i = \ell$ without loss of generality, and noting that the expectation of Δ_{ij} and Δ_{im} is non-negative, we have

$$\begin{aligned}
\mathrm{Cov}\big[\Delta_{ij}, \Delta_{im}\big] &\leq \mathbb{E}\big[\Delta_{ij}\Delta_{im}\big] \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(z) - f(x)\big)^2 \big(f(x') - f(x)\big)^2 K\left(\frac{\|x' - x\|}{r}\right) K\left(\frac{\|z - x\|}{r}\right) p(z)p(x')p(x)\, dz\, dx'\, dx \\
&\leq p_{\max}^3 M^4 r^4 \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} K\left(\frac{\|x' - x\|}{r}\right) K\left(\frac{\|z - x\|}{r}\right) dz\, dx'\, dx \\
&\leq p_{\max}^3 M^4 r^{4+2d}.
\end{aligned}$$

$|I| = 2$. Taking $i = \ell$ and $j = m$ without loss of generality, we have

$$\begin{aligned}
\mathrm{Var}\big[\Delta_{ij}\big] &\leq \mathbb{E}\big[\Delta_{ij}^2\big] \\
&\leq \int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^4 \left[K\left(\frac{\|x' - x\|}{r}\right)\right]^2 p(x')p(x)\, dx'\, dx \\
&\leq p_{\max}^2 M^4 r^4 K(0) \int_{\mathcal{X}}\int_{\mathcal{X}} K\left(\frac{\|x' - x\|}{r}\right) dx'\, dx \\
&\leq p_{\max}^2 M^4 r^{4+d} K(0).
\end{aligned}$$

$|I| = 1$. In this case $\Delta_{ij} = \Delta_{\ell m} = 0$.

Therefore

$$\mathrm{Var}\big[f^\top L f\big] \leq n^3 p_{\max}^3 M^4 r^{4+2d} + n^2 p_{\max}^2 M^4 r^{4+d} K(0) \leq C M^4 n^3 r^{4+2d},$$

where the latter inequality follows since $nr^d \gg 1$. For any $\delta > 0$, it follows from Chebyshev's inequality that

$$\mathbb{P}\left(\left|f^\top L f - \mathbb{E}\big[f^\top L f\big]\right| \geq \frac{M^2}{\delta} n^2 r^{d+2}\right) \leq C\frac{\delta^2}{n},$$

and since we have already upper bounded $\mathbb{E}\big[f^\top L f\big] \leq C_2 M^2 n^2 r^{d+2}$, the proposition follows.

Note that the bound on $\mathrm{Var}[\Delta_{ij}]$ follows as long as we can control $\|\nabla f\|_{L^4(\mathcal{X})}$; this implies the Lipschitz assumption—which gives us control of $\|\nabla f\|_{L^\infty(\mathcal{X})}$—can be weakened. However, the Sobolev assumption—which gives us control only over $\|\nabla f\|_{L^2(\mathcal{X})}$—will not do the job.

B.4 Bounds on neighborhood graph eigenvalues

In this section, we prove Lemma 4, following the lead of Burago et al. [2014], García Trillos et al. [2019a], Calder and García Trillos [2019], who establish similar results with respect to a manifold without boundary. To prove this lemma, in Theorem 27 we give estimates on the difference between eigenvalues of the graph Laplacian L and eigenvalues of the weighted Laplace-Beltrami operator Δ_P. We recall Δ_P is defined as

$$\Delta_P f(x) := -\frac{1}{p(x)} \mathrm{div}\big(p^2 \nabla f\big)(x).$$

To avoid confusion, in this section we write $\lambda_k(G_{n,r})$ for the kth smallest eigenvalue of the graph Laplacian matrix L and $\lambda_k(\Delta_P)$ for the kth smallest eigenvalue of Δ_P [1]. Some other notation: throughout this section,

[1] Under the assumptions (P1) and (P2), the operator Δ_P has a discrete spectrum; see García Trillos and Slepčev [2018a] for more details.

we will write $A, A_0, A_1, \ldots$ and $a, a_0, a_1, \ldots$ for constants which may depend on $\mathcal{X}$, d, K, and p, but do not depend on n; we keep track of all such constants explicitly in our proofs. We let L_K denote the Lipschitz constant of the kernel K. Finally, for notational ease we set θ and $\widetilde{\delta}$ to be the following (small) positive numbers:

$$\widetilde{\delta} := \max\left\{ n^{-1/d}, \min\left\{ \frac{1}{2^{d+3}A_0}, \frac{1}{A_3}, \frac{K(1)}{8L_K A_0}, \frac{1}{8\max\{A_1, A\}c_0} \right\} r \right\}, \quad \text{and} \quad \theta := \frac{1}{8\max\{A_1, A\}}. \tag{B.20}$$

We note that each of $\widetilde{\delta}, \theta$ and $\widetilde{\delta}/r$ are of at most constant order.

Theorem 27. *For any $\ell \in \mathbb{N}$ such that*

$$1 - A\left(r\sqrt{\lambda_\ell(\Delta_P)} + \theta + \widetilde{\delta} \right) \geq \frac{1}{2} \tag{B.21}$$

with probability at least $1 - A_0 n \exp(-a_0 n \theta^2 \widetilde{\delta}^d)$, it holds that

$$a\lambda_k(G_{n,r}) \leq nr^{d+2}\lambda_k(\Delta_P) \leq A\lambda_k(G_{n,r}), \quad \text{for all } 1 \leq k \leq \ell \tag{B.22}$$

Before moving forward to the proofs of Lemma 4 and Theorem 27, it is worth being clear about the differences between Theorem 27 and the results of Burago et al. [2014], García Trillos et al. [2019a], Calder and García Trillos [2019]. First of all, the reason we cannot directly use the results of these works in the proof of Lemma 4 is that they all assume the domain $\mathcal{X}$ is without boundary, whereas for our results in Section 3.4 we instead assume $\mathcal{X}$ has a (Lipschitz smooth) boundary. Fortunately, in this setting the high-level strategy shared by Burago et al. [2014], García Trillos et al. [2019a], Calder and García Trillos [2019] can still be used—indeed we follow it closely, as we summarize in Section B.4.1. However, many calculations need to be redone, in order to account for points x which are on or sufficiently close to the boundary of $\mathcal{X}$. For completeness and ease of reading, we provide a self-contained proof of Theorem 27, but we comment where appropriate on connections between the technical results we use in this proof, and those derived in Burago et al. [2014], García Trillos et al. [2019a], Calder and García Trillos [2019].

On the other hand, we should also point out that unlike the results of Burago et al. [2014], García Trillos et al. [2019a], Calder and García Trillos [2019], Theorem 27 does not imply that $\lambda_k(G_{n,r})$ is a consistent estimate of $\lambda_k(\Delta_P)$, i.e. it does not imply that $|(nr^{d+2})^{-1}\lambda_k(G_{n,r}) - \lambda_k(\Delta_P)| \to 0$ as $n \to \infty, r \to 0$. The key difficulty in proving consistency when $\mathcal{X}$ has a boundary can be summarized as follows: while at points $x \in \mathcal{X}$ satisfying $B(x, r) \subseteq \mathcal{X}$, the graph Laplacian L is a reasonable approximation of the operator Δ_P, at points x near the boundary L is known to approximate a different operator altogether [Belkin et al., 2012]. This is reminiscent of the boundary effects present in the analysis of kernel smoothing. We believe a more subtle analysis might imply convergence of eigenvalues in this setting. However, the conclusion of Theorem 27—that $\lambda_k(G_{n,r})/(nr^{d+2}\lambda_k(\Delta_P))$ is bounded above and below by constants that do not depend on k—suffices for our purposes.

The bulk of the remainder of this section is devoted to the proof of Theorem 27. First, however, we show that under our regularity conditions on p and $\mathcal{X}$, Lemma 4 is a simple consequence of Theorem 27. The link between the two is Weyl's Law.

Proposition 20 (Weyl's Law)**.** *Suppose the density p and the domain $\mathcal{X}$ satisfy (P1) and (P2). Then there exist constants a_2 and A_2 such that*

$$a_2 k^{2/d} \leq \lambda_k(\Delta_P) \leq A_2 k^{2/d} \quad \text{for all } k \in \mathbb{N}, k > 1. \tag{B.23}$$

See Lemma 28 of Dunlop et al. [2020] for a proof that (P1) and (P2) imply Weyl's Law.

Proof of Lemma 4. Put

$$\ell_\star = \left\lfloor \left(\frac{\big(1/(2A) - (\theta + \widetilde{\delta})\big)}{r A_2^{1/2}} \right)^d \right\rfloor.$$

Let us verify that $\lambda_{\ell_\star}(\Delta_P)$ satisfies the condition (B.21) of Theorem 27. Setting $c_0 := 1/(2^{1/d} 4 A_2^{1/2})$, the assumed upper bound on the radius $r \leq c_0$ guarantees that $\ell_\star \geq 2$. Therefore, by Proposition 20 we have that

$$\sqrt{\lambda_{\ell_\star}(\Delta_P)} \leq A_2^{1/2} \ell_\star^{1/d} \leq \frac{1}{r}\left(\frac{1}{2A} - (\theta + \widetilde{\delta}) \right).$$

Rearranging the above inequality shows that condition (B.21) is satisfied.

It is therefore the case that the inequalities in (B.22) hold with probability at least $1 - A_0 n \exp(-a_0 n \theta^2 \widetilde{\delta}^d)$. Together, (B.22) and (B.23) imply the following bounds on the graph Laplacian eigenvalues:

$$\frac{a}{A_2} n r^{d+2} k^{2/d} \leq \lambda_k(G_{n,r}) \leq \frac{A}{a_2} n r^{d+2} k^{2/d} \quad \text{for all } 2 \leq k \leq \ell_\star.$$

It remains to bound $\lambda_k(G_{n,r})$ for those indices k which are greater than $\ell_\star$. On the one hand, since the eigenvalues are sorted in ascending order, we can use the lower bound on $\lambda_{\ell_\star}(G_{n,r})$ that we have just derived:

$$\lambda_k(G_{n,r}) \geq \lambda_{\ell_\star}(G_{n,r}) \geq \frac{a_2}{A} n r^{d+2} \ell_\star^{2/d} \geq \frac{a_2}{64 A^3 A_2} n r^d.$$

On the other hand, for any graph G the maximum eigenvalue of the Laplacian is upper bounded by twice the maximum degree [Chung, 1997]. Writing $D_{\max}(G_{n,r})$ for the maximum degree of $G_{n,r}$, it is thus a consequence of Lemma 40 that

$$\lambda_k(G_{n,r}) \leq 2 D_{\max}(G_{n,r}) \leq 4 p_{\max} n r^d,$$

with probability at least $1 - 2n \exp\Big(-n r^d p_{\min}/(3K(0)^2)\Big)$. In sum, we have shown that with probability at least $1 - A_0 n \exp(-a_0 n \theta^2 \widetilde{\delta}^d) - 2n \exp\Big(-n r^d p_{\min}/(3K(0)^2)\Big)$,

$$\min\left\{ \frac{a_2}{A} n r^{d+2} k^{2/d}, \frac{a_2}{A^3 64 A_3} n r^d \right\} \leq \lambda_k(G_{n,r}) \leq \min\left\{ \frac{A_2}{a} n r^{2+d} k^{2/d}, 4 p_{\max} n r^d \right\} \quad \text{for all } 2 \leq k \leq n.$$

Lemma 4 then follows upon setting

$$C_1 := \max\{2A_0, 4\}, \qquad c_1 := \min\left\{ \frac{p_{\min}}{3K(0)^2}, \frac{\theta^2 \widetilde{\delta}}{r} \right\}$$

$$C_3 := \max\left\{ \frac{A_2}{a}, 4 p_{\max} \right\}, \qquad c_3 := \min\left\{ \frac{a_2}{A}, \frac{a_2}{A^3 64 A_3} \right\}.$$

in the statement of that Lemma.

B.4.1 Proof of Theorem 27

In this section we prove Theorem 27, following closely the approach of Burago et al. [2014], García Trillos et al. [2019a], Calder and García Trillos [2019]. As in these works, we relate $\lambda_k(\Delta_P)$ and $\lambda_k(G_{n,r})$ by means of the Dirichlet energies

$$b_r(u) := \frac{1}{n^2 r^{d+2}} u^\top L u$$

and

$$D_2(f) := \begin{cases} \int_{\mathcal{X}} \|\nabla f(x)\|^2 p^2(x)\,dx & \text{if } f \in H^1(\mathcal{X}) \\ \infty & \text{otherwise,} \end{cases}$$

Let us pause briefly to motivate the relevance of $b_r(u)$ and $D_2(f)$. In the following discussion, recall that for a function $u : \{X_1, \dots, X_n\} \to \mathbb{R}$, the empirical norm is defined as $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n (u(X_i))^2$, and the class $L^2(P_n)$ consists of those $u \in \mathbb{R}^n$ for which $\|u\|_n < \infty$. Similarly, for a function $f : \mathcal{X} \to \mathbb{R}$, the $L^2(P)$ norm of f is

$$\|f\|_P^2 := \int_{\mathcal{X}} \big|f(x)\big|^2 p(x)\,dx,$$

and the class $L^2(P)$ consists of those f for which $\|f\|_P < \infty$. Now, suppose one could show the following two results:

(1) an upper bound of $b_r(u)$ by $D_2\big(\mathcal{I}(u)\big)$ for an appropriate choice of interpolating map $\mathcal{I} : L^2(P_n) \to L^2(\mathcal{X})$, and vice versa an upper bound of $D_2(f)$ by $b_r(\mathcal{P}(f))$ for an appropriate choice of discretization map $\mathcal{P} : L^2(\mathcal{X}) \to L^2(P_n)$,

(2) that $\mathcal{I}$ and $\mathcal{P}$ were near-isometries, meaning $\|\mathcal{I}(u)\|_P \approx \|u\|_n$ and $\|\mathcal{P}(f)\|_P \approx \|f\|_n$.

Then, by using the variational characterization of eigenvalues $\lambda_k(\Delta_P)$ and $\lambda_k(G_{n,r})$—i.e. the Courant-Fischer Theorem—one could obtain estimates on the error $\big|nr^{d+2}\lambda_k(\Delta_P) - \lambda_k(G_{n,r})\big|$.

We will momentarily define particular maps $\widetilde{\mathcal{I}}$ and $\widetilde{\mathcal{P}}$, and establish that they satisfy both (1) and (2). In order to define these maps, we must first introduce a particular probability measure $\widetilde{P}_n$ that, with high probability, is close in transportation distance to both P_n and P. This estimate on the transportation distance—which we now give—will be the workhorse that allows us to relate b_r to D_2, and $\|\cdot\|_n$ to $\|\cdot\|_P$.

Transportation distance between P_n and P. For a measure μ defined on $\mathcal{X}$ and map $T : \mathcal{X} \to \mathcal{X}$, let $T_\sharp\mu$ denote the *push-forward* of μ by T, i.e the measure for which

$$\big(T_\sharp\mu\big)(U) := \mu\big(T^{-1}(U)\big)$$

for any Borel subset $U \subseteq \mathcal{X}$. Suppose $T_\sharp\mu = P_n$; then the map T is referred to as transportation map between μ and P_n. The ∞-transportation distance between μ and P_n is then

$$d_\infty(\mu, P_n) := \inf_{T : T_\sharp\mu = P_n} \|T - \mathrm{Id}\|_{L^\infty(\mu)} \tag{B.24}$$

where $\mathrm{Id}(x) = x$ is the identity mapping.

Calder and García Trillos [2019] take $\mathcal{X}$ to be a smooth submanifold of $\mathbb{R}^d$ without boundary, i.e. they assume $\mathcal{X}$ satisfies (P3). In this setting, they exhibit an absolutely continuous measure $\widetilde{P}_n$ with density $\widetilde{p}_n$ that with high probability is close to P_n in transportation distance, and for which $\|p - \widetilde{p}_n\|_{L^\infty}$ is also small. In Proposition 21, we adapt this result to the setting of full-dimensional manifolds with boundary.

Proposition 21. *Suppose $\mathcal{X}$ satisfies (P1) and p satisfies (P2). Then with probability at least $1 - A_0 n \exp\{-a_0 n\theta^2\widetilde{\delta}^d\}$, the following statement holds: there exists a probability measure $\widetilde{P}_n$ with density $\widetilde{p}_n$ such that:*

$$d_\infty(\widetilde{P}_n, P_n) \le A_0\widetilde{\delta} \tag{B.25}$$

and

$$\|\widetilde{p}_n - p\|_\infty \le A_0\big(\widetilde{\delta} + \theta\big). \tag{B.26}$$

For the rest of this section, we let $\widetilde{P}_n$ be a probability measure with density $\widetilde{p}_n$, that satisfies the conclusions of Proposition 21. Additionally we denote by $\widetilde{T}_n$ an *optimal transport map* between $\widetilde{P}_n$ and P_n, meaning a transportation map which achieves the infimum in (B.24). Finally, we write $U_1, \dots, U_n$ for the preimages of $X_1, \dots, X_n$ under $\widetilde{T}_n$, meaning $U_i = \widetilde{T}_n^{-1}(X_i)$.

Interpolation and discretization maps. The discretization map $\widetilde{\mathcal{P}} : L^2(\mathcal{X}) \to L^2(P_n)$ is given by averaging over the cells $U_1, \dots, U_n$,

$$\big(\widetilde{\mathcal{P}} f\big)(X_i) := n \cdot \int_{U_i} f(x) \widetilde{p}_n(x)\, dx.$$

On the other hand, the interpolation map $\widetilde{\mathcal{I}} : L^2(P_n) \to L^2(\mathcal{X})$ is defined as $\widetilde{\mathcal{I}} u := \Lambda_{r - 2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^\star u)$. Here, $\widetilde{\mathcal{P}}^\star = u \circ \widetilde{T}$ is the adjoint of $\widetilde{\mathcal{P}}$, i.e.

$$\big(\widetilde{\mathcal{P}}^\star u\big)(x) = \sum_{j=1}^n u(x_i) \mathbf{1}\{x \in U_i\}$$

and $\Lambda_{r - 2A_0\widetilde{\delta}}$ is a kernel smoothing operator, defined with respect to a carefully chosen kernel ψ. To be precise, for any $h > 0$,

$$\Lambda_h(f) := \frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \eta_h(x', x) f(x')\, dx', \quad \eta_h(x', x) := \psi\bigg(\frac{\|x' - x\|}{r}\bigg)$$

where $\psi(t) := (1/\sigma_K) \int_t^\infty s K(s)\, ds$ and $\tau_h(x) := (1/h^d) \int_{\mathcal{X}} \eta_h(x', x)\, dx'$ is a normalizing constant.

Propositions 22 and 23 establish our claims regarding $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{I}}$: first, that they approximately preserve the Dirichlet energies b_r and D_2, and second that they are near-isometries for functions $u \in L^2(P_n)$ (or $f \in L^2(P)$) of small Dirichlet energy $b_r(u)$ (or $D_2(f)$).

Proposition 22 (**cf. Proposition 4.1 of Calder and García Trillos [2019]**). *With probability at least* $1 - A_0 n \exp(-a_0 n \theta^2 \widetilde{\delta}^d)$, *we have the following.*

(1) For every $u \in L^2(P_n)$,

$$\sigma_K D_2(\widetilde{\mathcal{I}} u) \le A_8 \Big(1 + A_1(\theta + \widetilde{\delta})\Big) \cdot \bigg(1 + A_3 \frac{\widetilde{\delta}}{r}\bigg) b_r(u). \tag{B.27}$$

(2) For every $f \in L^2(\mathcal{X})$,

$$b_r(\widetilde{\mathcal{P}} f) \le \Big(1 + A_1(\theta + \widetilde{\delta})\Big) \cdot \Big(1 + A_9 \frac{\widetilde{\delta}}{r}\Big) \cdot \Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big) \cdot \sigma_K D_2(f). \tag{B.28}$$

Proposition 23 (**cf. Proposition 4.2 of Calder and García Trillos [2019]**). *With probability at least* $1 - A_0 n \exp(-a_0 n \theta^2 \widetilde{\delta}^d)$, *we have the following.*

(1) For every $f \in L^2(\mathcal{X})$,

$$\Big| \|f\|_P^2 - \|\widetilde{\mathcal{P}} f\|_n^2 \Big| \le A_5 r \|f\|_P \sqrt{D_2(f)} + A_1\big(\theta + \widetilde{\delta}\big) \|f\|_P^2. \tag{B.29}$$

(2) For every $u \in L^2(P_n)$,

$$\Big| \|\widetilde{\mathcal{I}} u\|_P^2 - \|u\|_n^2 \Big| \le A_6 r \|u\|_n \sqrt{b_r(u)} + A_7\big(\theta + \widetilde{\delta}\big) \|u\|_n^2. \tag{B.30}$$

We will devote most of the rest of this section to the proofs of Propositions 21, 22, and 23. First, however, we use these propositions to prove Theorem 27.

Proof of Theorem 27. Throughout this proof, we assume that inequalities (B.27)-(B.30) are satisfied. We take A and a to be positive constants such that

$$\frac{1}{a} \geq 2\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\Big(1 + A_9\frac{\widetilde{\delta}}{r}\Big)\Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big), \quad \text{and} \quad A \geq \max\left\{A_1, A_5, \frac{1}{\sqrt{a}}A_6, A_7, 2A_8\left(1 + A_1(\theta + \widetilde{\delta})\right)\left(1 + A_3\frac{\widetilde{\delta}}{r}\right)\right\}.$$

Let k be any number in $1, \ldots, \ell$. We start with the upper bound in (B.22), proceeding as in Proposition 4.4 of Burago et al. [2014]. Let $f_1, \ldots, f_k$ denote the first k eigenfunctions of Δ_P and set $W := \operatorname{span}\{f_1, \ldots, f_k\}$, so that by the Courant-Fischer principle $D_2(f) \leq \lambda_k(\Delta_P)\|f\|_P^2$ for every $f \in W$. As a result, by Part (1) of Proposition 23 we have that for any $f \in W$,

$$\big\|\widetilde{\mathcal{P}} f\big\|_n^2 \geq \left(1 - A_5 r\sqrt{\lambda_k(\Delta_P)} - A_1(\theta + \widetilde{\delta})\right)\|f\|_P^2 \geq \frac{1}{2}\|f\|_P^2,$$

where the second inequality follows by assumption (B.21).

Therefore $\widetilde{\mathcal{P}}$ is injective over W, and $\widetilde{\mathcal{P}}W$ has dimension ℓ. This means we can invoke the Courant-Fischer Theorem, along with Proposition 22, and conclude that

$$\begin{aligned}
\frac{\lambda_k(G_{n,r})}{nr^{d+2}} &\leq \max_{\substack{u \in \widetilde{\mathcal{P}}W \\ u \neq 0}} \frac{b_r(u)}{\|u\|_n^2} \\
&= \max_{\substack{f \in W \\ f \neq 0}} \frac{b_r(\widetilde{\mathcal{P}} f)}{\big\|\widetilde{\mathcal{P}} f\big\|_n^2} \\
&\leq 2\Big(1 + A_1(\theta + \widetilde{\delta})\Big) \cdot \Big(1 + A_9\frac{\widetilde{\delta}}{r}\Big) \cdot \Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big)\sigma_K \lambda_k(\Delta_P),
\end{aligned}$$

establishing the lower bound in (B.22).

The upper bound follows from essentially parallel reasoning. Recalling that $v_1, \ldots, v_k$ denote the first k eigenvectors of L, set $U := \operatorname{span}\{v_1, \ldots, v_k\}$, so that $nr^{d+2} b_r(u) \leq \lambda_k(G_{n,r})\|u\|_n^2$. By Proposition 23, Part (2), we have that for every $u \in U$,

$$\begin{aligned}
\big\|\widetilde{\mathcal{I}} u\big\|_P^2 &\geq \|u\|_n^2 - A_6 r\|u\|_n\sqrt{b_r(u)} - A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2 \\
&\geq \|u\|_n^2 - A_6 r\|u\|_n^2\sqrt{\frac{\lambda_k(G_{n,r})}{nr^{d+2}}} - A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2 \\
&\geq \|u\|_n^2 - A_6 r\|u\|_n^2\sqrt{\frac{1}{a}\lambda_k(\Delta_P)} - A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2 \\
&\geq \frac{1}{2}\|u\|_n^2,
\end{aligned}$$

where the second to last inequality follows from the lower bound $a\lambda_k(G_{n,r}) \leq nr^{d+2}\lambda_k(\Delta_P)$ that we just derived, and the last inequality from assumption (B.21).

Therefore $\widetilde{\mathcal{I}}$ is injective over U, $\widetilde{\mathcal{I}}U$ has dimension k, and by Proposition 22 we conclude that

$$\begin{aligned}
\lambda_k(\Delta_P) &\leq \max_{u \in U} \frac{D_2(\widetilde{\mathcal{I}} u)}{\|u\|_P^2} \\
&\leq 2A_8\left(1 + A_1(\theta + \widetilde{\delta})\right)\left(1 + A_3\frac{\widetilde{\delta}}{r}\right)\max_{u \in U}\frac{b_r(u)}{\|u\|_n^2} \\
&\leq 2A_8\left(1 + A_1(\theta + \widetilde{\delta})\right)\left(1 + A_3\frac{\widetilde{\delta}}{r}\right)\frac{\lambda_k(G_{n,r})}{nr^{d+2}},
\end{aligned}$$

establishing the upper bound in (B.22).

Organization of this section. The rest of this section will be devoted to proving Propositions 21, 22 and 23. To prove the latter two propositions, it will help to introduce the intermediate energies

$$\widetilde{E}_r(f,\eta,V) := \frac{1}{r^{d+2}}\int_V\int_{\mathcal{X}}\big(f(x')-f(x)\big)^2\eta\bigg(\frac{\|x'-x\|}{r}\bigg)\widetilde{p}_n(x')\widetilde{p}_n(x)\,dx'\,dx$$

and

$$E_r(f,\eta,V) := \frac{1}{r^{d+2}}\int_V\int_{\mathcal{X}}\big(f(x')-f(x)\big)^2\eta\bigg(\frac{\|x'-x\|}{r}\bigg)p(x')p(x)\,dx'\,dx.$$

Here $\eta : [0,\infty) \to [0,\infty)$ is an arbitrary kernel, and $V \subseteq \mathcal{X}$ is a measurable set. We will abbreviate $\widetilde{E}_r(f,\eta,\mathcal{X})$ as $\widetilde{E}_r(f,\eta)$ and $\widetilde{E}_r(f,K) = \widetilde{E}_r(f)$ (and likewise with E_r.)

The proof of Proposition 21 is given in Section B.4.2. In Section B.4.3, we establish relationships between the (non-random) functionals $E_r(f)$ and $D_2(f)$, as well as providing estimates on some assorted integrals. In Section B.4.4, we establish relationships between the stochastic functionals $\widetilde{E}_r(f)$ and $E_r(f)$, between $\widetilde{E}_r\big(\widetilde{\mathcal{I}}(u)\big)$ and $b_r\big(u\big)$, and between $\widetilde{E}_r\big(f\big)$ and $b_r\big(\widetilde{\mathcal{P}}f\big)$. Finally, in Section B.4.5 we use these various relationships to prove Propositions 22 and 23.

B.4.2 Proof of Proposition 21

We start by defining the density $\widetilde{p}_n$, which will be piecewise constant over a particular partition $\mathcal{Q}$ of $\mathcal{X}$. Specifically, for each Q in $\mathcal{Q}$ and every $x \in Q$, we set

$$\widetilde{p}_n(x) := \frac{P_n(Q)}{\mathrm{vol}(Q)}, \tag{B.31}$$

where $\mathrm{vol}(\cdot)$ denotes the Lebesgue measure. Then $\widetilde{P}_n(U) = \int_U \widetilde{p}_n(x)\,dx$.

We now construct the partition $\mathcal{Q}$, in progressive degrees of generality on the domain $\mathcal{X}$.

- In the special case of the unit cube $\mathcal{X} = (0,1)^d$, the partition will simply be a collection of cubes,

  $$\mathcal{Q} = \Big\{Q_k : k \in [\widetilde{\delta}^{-1}]^d\Big\},$$

  where $Q_k = \widetilde{\delta}\Big([k_1-1,k_1]\otimes\cdots\otimes[k_d-1,k_d]\Big)$ and we assume without loss of generality that $\widetilde{\delta}^{-1} \in \mathbb{N}$.

- If $\mathcal{X}$ is an open, connected set with smooth boundary, then by Proposition 3.2 of García Trillos and Slepčev [2015], there exist a finite number $N(\mathcal{X}) \in \mathbb{N}$ of disjoint polytopes which cover $\mathcal{X}$. Moreover, letting U_j denote the intersection of the jth of these polytopes with $\bar{\mathcal{X}}$, this proposition establishes that for each j there exists a bi-Lipschitz homeomorphism $\Phi_j : U_j \to [0,1]^d$. We take the collection

  $$\mathcal{Q} = \Big\{\Phi_j^{-1}(Q_k) : j = 1,\dots,N(\mathcal{X}) \;\text{ and }\; k \in [\widetilde{\delta}^{-1}]^d\Big\}$$

  to be our partition. Denote by L_Φ the maximum of the bi-Lipschitz constants of $\Phi_1,\dots,\Phi_{N(\mathcal{X})}$.

- Finally, in the general case where $\mathcal{X}$ is an open, connected set with Lipschitz boundary, then there exists a bi-Lipschitz homeomorphism Ψ between $\mathcal{X}$ and a smooth, open, connected set with Lipschitz boundary. Letting Φ_j and $\widetilde{Q}_{j,k}$ be as before, we take the collection

  $$\mathcal{Q} = \Big\{\widetilde{Q}_{j,k} = \Big(\Psi^{-1}\circ\Phi_j^{-1}\Big)(Q_k) : j = 1,\dots,N(\mathcal{X}) \;\text{ and }\; k \in [\widetilde{\delta}^{-1}]^d\Big\}$$

  to be our partition. Denote by L_Ψ the bi-Lipschitz constant of Ψ.

Let us record a few facts which hold for all $\widetilde{Q}_{j,k} \in \mathcal{Q}$, and which follow from the bi-Lipschitz properties of Φ_j and Ψ: first that

$$\operatorname{diam}(\widetilde{Q}_{j,k}) \leq L_\Psi Ł_\Phi \widetilde{\delta}, \tag{B.32}$$

and second that

$$\operatorname{vol}(\widetilde{Q}_{j,k}) \geq \left(\frac{1}{L_\Psi L_\Phi}\right)^d \widetilde{\delta}^d. \tag{B.33}$$

We now use these facts to show that $\widetilde{P}_n$ satisfies the claims of Proposition 21. On the one hand for every $Q \in \mathcal{Q}$, letting $N(Q)$ denote the number of design points $\{X_1, \dots, X_k\}$ which fall in Q, we have

$$\widetilde{P}_n(Q) = \int_Q \widetilde{p}_n(x)\,dx = P_n(Q) = \frac{N(Q)}{n}.$$

Moreover, ignoring those cells for which $N(Q) = 0$ (since $\widetilde{P}_n(Q) = 0$ for such Q, and so they do not contribute to the essential supremum in (B.24)), appropriately dividing each remaining cell $Q \in \mathcal{Q}$ into $N(Q)$ subsets $S_1, \dots, S_{N(Q)}$ of equal volume, and mapping each S_ℓ to a different design point $X_i \in Q$, we can exhibit a transport map T from $\widetilde{P}_n$ to P_n for which

$$\|T - \mathrm{Id}\|_{L^\infty(\widetilde{P}_n)} \leq \max_{Q \in \mathcal{Q}} \operatorname{diam}(Q) \leq L_\Psi Ł_\Phi \widetilde{\delta}.$$

On the other hand, applying the triangle inequality we have that for $x \in \widetilde{Q}_{j,k}$

$$|\widetilde{p}_n(x) - p(x)| \leq \left|\frac{P_n(\widetilde{Q}_{j,k}) - P(\widetilde{Q}_{j,k})}{\operatorname{vol}(\widetilde{Q}_{j,k})}\right| + \frac{1}{\operatorname{vol}(\widetilde{Q}_{j,k})}\int_{\widetilde{Q}_{j,k}} |p(x') - p(x)|\,dx,$$

and using the Lipschitz property of p we find that

$$\|\widetilde{p}_n - p\|_{L^\infty} \leq \max_{j,k}\left|\frac{P_n(\widetilde{Q}_{j,k}) - P(\widetilde{Q}_{j,k})}{\operatorname{vol}(\widetilde{Q}_{j,k})}\right| + L_p L_\Phi L_\Psi \widetilde{\delta}. \tag{B.34}$$

From Hoeffding's inequality and a union bound, we obtain that

$$\begin{aligned}\mathbb{P}\Big(\big|P_n(\widetilde{Q}) - P(\widetilde{Q})\big| \leq \theta P(\widetilde{Q}) \quad \forall \widetilde{Q} \in \mathcal{Q}\Big) &\geq 1 - 2\sharp(\mathcal{Q}) \cdot \exp\left\{-\frac{\theta^2 n \min\{P(\widetilde{Q})\}}{3}\right\} \\ &\geq 1 - \frac{2N(\mathcal{X})}{\widetilde{\delta}^d} \cdot \exp\left\{-\frac{\theta^2 n p_{\min} \widetilde{\delta}^d}{3\big(L_\Psi L_\Phi\big)^d}\right\}.\end{aligned}$$

Noting that by assumption $P(\widetilde{Q}) \leq p_{\max}\operatorname{vol}(\widetilde{Q})$ and $\widetilde{\delta}^{-d} \leq n$, the claim follows upon plugging back into (B.34), and setting

$$a_0 := \frac{1}{3\big(L_\Psi L_\Phi\big)^d} \quad \text{and} \quad A_0 := \max\Big\{2N(\mathcal{X}), L_p L_\Psi L_\Phi, L_\Psi L_\Phi\Big\}$$

in the statement of the proposition.

B.4.3 Non-random functionals and integrals

Let us start by making the following observation, which we make use of repeatedly in this section. Let $\eta : [0, \infty) \to [0, \infty)$ be an otherwise arbitrary function. As a consequence of (P1), there exist constants c_0 and a_3 which depend on $\mathcal{X}$, such that for any $0 < \varepsilon \leq c_0$ it holds that

$$\int_{B(x,\varepsilon)\cap\mathcal{X}} \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' \geq a_3 \cdot \int_{B(x,\varepsilon)} \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' \tag{B.35}$$

As a special case: when $\eta(x) = 1$, this implies $\mathrm{vol}\big(B(x,\varepsilon) \cap \mathcal{X}\big) \geq a_3 \nu_d \varepsilon^d$ for any $0 < \varepsilon \leq c_0$.

We have already upper bounded $E_r(f)$ by (a constant times) $D_2(f)$ in the proof of Lemma 3. In Lemma 28, we establish the reverse inequality.

Lemma 28 (cf. Lemma 9 of García Trillos et al. [2019a], Lemma 5.5 of Burago et al. [2014]). *For any $f \in L^2(\mathcal{X})$, and any $0 < h \leq c_0$, it holds that*

$$\sigma_K D_2(\Lambda_h f) \leq A_8 E_h(f).$$

To prove Lemma 28, we require upper and lower bounds on $\tau_h(x)$, as well as an upper bound on the gradient of τ_h. The lower bound here—$\tau_h(x) \geq a_3$—is quite a bit a looser than what can be shown when $\mathcal{X}$ has no boundary. The same is the case regarding the upper bound of the size of the gradient $\|\nabla \tau_h(x)\|$. However, the bounds as stated here will be sufficient for our purposes.

Lemma 29. *For any $0 < h \leq c_0$, for all $x \in \mathcal{X}$ it holds that*

$$a_3 \leq \tau_h(x) \leq 1.$$

and

$$\|\nabla \tau_h(x)\| \leq \frac{1}{\sqrt{d\sigma_K} h}.$$

Finally, to prove part (2) of Proposition 23, we require Lemma 30, which gives an estimate on the error $\Lambda_h f - f$ in $\|\cdot\|_P^2$ norm.

Lemma 30 (c.f Lemma 8 of García Trillos et al. [2019a], Lemma 5.4 of Burago et al. [2014]). *For any $0 < h \leq c_0$,*

$$\big\|\Lambda_h f\big\|_P^2 \leq \frac{p_{\max}}{a_3 p_{\min}} \big\|f\big\|_P^2, \tag{B.36}$$

and

$$\big\|\Lambda_h f - f\big\|_P^2 \leq \frac{1}{a_3 \sigma_K p_{\min}} h^2 E_h(f), \tag{B.37}$$

for all $f \in L^2(\mathcal{X})$.

Proof of Lemma 28. For any $a \in \mathbb{R}$, $\Lambda_h f$ satisfies the identity

$$\Lambda_h f(x) = a + \frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \eta_h(x', x)\big(f(x') - a\big)\, dx',$$

and by differentiating with respect to x, we obtain

$$\big(\nabla \Lambda_h f\big)(x) = \frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \big(\nabla \eta_h(x', \cdot)\big)(x)\big(f(x') - a\big)\, dx' + \nabla\left(\frac{1}{\tau_h}\right)(x) \cdot \frac{1}{h^d} \int_{\mathcal{X}} \eta_h(x', x)\big(f(x') - a\big)\, dx'$$

Plugging in $a = f(x)$, we get $\nabla \Lambda_h f(x) = J_1(x)/\tau_h(x) + J_2(x)$ for

$$J_1(x) := \frac{1}{h^d} \int_{\mathcal{X}} \big(\nabla \eta_h(x', \cdot)\big)(x)\big(f(x') - f(x)\big)\, dx', \quad J_2(x) := \nabla\left(\frac{1}{\tau_h}\right)(x) \cdot \frac{1}{h^d} \int_{\mathcal{X}} \eta_h(x', x)\big(f(x') - f(x)\big)\, dx'.$$

To upper bound $\big\|J_1(x)\big\|^2$, we first compute the gradient of $\eta_h(x', \cdot)$,

$$\begin{aligned}\big(\nabla \eta_h(x', \cdot)\big)(x) &= \frac{1}{h} \psi'\left(\frac{\|x' - x\|}{h}\right) \frac{(x - x')}{\|x' - x\|} \\ &= \frac{1}{\sigma_K h^2} K\left(\frac{\|x' - x\|}{h}\right)(x' - x),\end{aligned}$$

and additionally note that $\|J_1(x)\|^2 = \sup_w \big(\langle J_1(x), w\rangle\big)^2$ where the supremum is over unit norm vector. Taking w to be a unit norm vector which achieves this supremum, we have that

$$\begin{aligned}\big\|J_1(x)\big\|^2 &= \frac{1}{\sigma_K^2 h^{4+2d}}\left[\int_{\mathcal{X}} \big(f(x') - f(x)\big) K\Big(\frac{\|x'-x\|}{h}\Big)(x'-x)^\top w\, dx'\right]^2 \\ &\le \frac{1}{\sigma_K^2 h^{4+2d}}\left[\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big((x'-x)^\top w\big)^2\, dx'\right]\left[\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big(f(x') - f(x)\big)^2\, dx'\right].\end{aligned}$$

By a change of variables, we obtain

$$\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big((x'-x)^\top w\big)^2\, dx' \le h^{d+2}\int_{\mathcal{X}} K\big(\|z\|\big)\big(z^\top w\big)^2\, dz \le \sigma_K h^{d+2},$$

with the resulting upper bound

$$\big\|J_1(x)\big\|^2 \le \frac{1}{\sigma_K h^{2+d}}\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big(f(x') - f(x)\big)^2\, dx'.$$

To upper bound $\big\|J_2(x)\big\|^2$, we use the Cauchy-Schwarz inequality along with the observation $\eta_h(x', x) \le \frac{1}{\sigma_K} K\big(\|x'-x\|/h\big)$ to deduce

$$\begin{aligned}\big\|J_2(x)\big\|^2 &\le \Big\|\nabla\Big(\frac{1}{\tau_h}\Big)(x)\Big\|^2 \frac{1}{h^{2d}}\left[\int_{\mathcal{X}} \eta_h(x', x)\, dx'\right]\cdot\left[\int_{\mathcal{X}} \eta_h(x', x)\big(f(x') - f(x)\big)^2\, dx'\right] \\ &= \Big\|\nabla\Big(\frac{1}{\tau_h}\Big)(x)\Big\|^2 \frac{\tau_h(x)}{h^d}\int_{\mathcal{X}} \eta_h(x', x)\big(f(x') - f(x)\big)^2\, dx' \\ &\le \Big\|\nabla\Big(\frac{1}{\tau_h}\Big)(x)\Big\|^2 \frac{\tau_h(x)}{\sigma_K h^d}\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big(f(x') - f(x)\big)^2\, dx' \\ &\le \frac{1}{d a_3^2 \sigma_K^2 h^{2+d}}\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big(f(x') - f(x)\big)^2\, dx',\end{aligned}$$

where the last inequality follows from the estimates on τ_h and $\nabla\tau_h$ provided in Lemma 29. Combining our bounds on $\big\|J_1(x)\big\|^2$ and $\big\|J_2(x)\big\|^2$ along with the lower bound on $\tau_h(x)$ in Lemma 29 and integrating over $\mathcal{X}$, we have

$$\begin{aligned}\sigma_K D_2(\Lambda_h f) &= \sigma_K \int_{\mathcal{X}} \Big\|\Big(\nabla\Lambda_h f)(x)\Big\|^2 p^2(x)\, dx \\ &\le 2\sigma_K \int_{\mathcal{X}}\left(\frac{\big\|J_1(x)\big\|^2}{\tau_h^2(x)} + \big\|J_2(x)\big\|^2\right) p^2(x)\, dx \\ &\le \left(\frac{1}{a_3^2} + \frac{1}{d a_3^2 \sigma_K}\right)\frac{2}{h^{d+2}}\int_{\mathcal{X}}\int_{\mathcal{X}} K\Big(\frac{\|x'-x\|}{h}\Big)\big(f(x') - f(x)\big)^2 p^2(x)\, dx'\, dx \\ &\le 2\Big(1 + \frac{L_p h}{p_{\min}}\Big)\left(\frac{1}{a_3^2} + \frac{1}{d a_3^2 \sigma_K}\right) E_h(f),\end{aligned}$$

and taking $A_8 := 2\Big(1 + \frac{L_p c_0}{p_{\min}}\Big)\left(\frac{1}{a_3^2} + \frac{1}{d a_3^2 \sigma_K}\right)$ completes the proof of Lemma 28.

Proof of Lemma 29. We first establish our estimates of $\tau_h(x)$, and then upper bound $\|\nabla\tau_h(x)\|$. Using (B.35), we have that

$$\begin{aligned}\tau_h(x) &= \frac{1}{h^d}\int_{\mathcal{X}\cap B(x,h)} \psi\bigg(\frac{\|x'-x\|}{h}\bigg)\,dx' \\ &\geq \frac{a_3}{h^d}\int_{B(x,h)} \psi\bigg(\frac{\|x'-x\|}{h}\bigg)\,dx' \\ &= a_3\int_{B(0,1)} \psi(\|z\|)\,dz,\end{aligned}$$

and it follows from similar reasoning that $\tau_h(x) \leq \int_{B(0,1)}\psi(\|z\|)\,dz$.

We will now show that $\int_{B(0,1)}\psi(\|z\|)\,dz = 1$, from which we derive the estimates $a_3 \leq \tau_h(x) \leq 1$. To see the identity, note that on the one hand, by converting to polar coordinates and integrating by parts we obtain

$$\int_{B(0,1)}\psi\big(\|z\|\big)\,dz = d\nu_d\int_0^1 \psi(t)t^{d-1}\,dt = -\nu_d\int_0^1\psi'(t)t^d\,dt = \frac{\nu_d}{\sigma_K}\int_0^1 t^{d+1}K(t)\,dt;$$

on the other hand, again converting to polar coordinates, we have

$$\sigma_K = \frac{1}{d}\int_{\mathbb{R}^d}\|x\|^2K(\|x\|)\,dx = \nu_d\int_0^1 t^{d+1}K(t)\,dt,$$

and so $\int_{B(0,1)}\psi(\|z\|)\,dz = 1$.

Now we upper bound $\|\nabla\tau_h(x)\|^2$. Exchanging derivative and integral, we have

$$\nabla\tau_h(x) = \frac{1}{h^d}\int_{\mathcal{X}}\big(\nabla\eta_h(x',\cdot)\big)(x)\,dx' = \frac{1}{\sigma_K h^{d+2}}\int_{\mathcal{X}} K\bigg(\frac{\|x'-x\|}{h}\bigg)(x'-x)\,dx',$$

whence by the Cauchy-Schwarz inequality,

$$\|\nabla\tau_h(x)\|^2 \leq \frac{1}{\sigma_K^2h^{2d+4}}\bigg[\int_{\mathcal{X}}K\bigg(\frac{\|x'-x\|}{h}\bigg)\,dx'\bigg]\bigg[\int_{\mathcal{X}}K\bigg(\frac{\|x'-x\|}{h}\bigg)\|x'-x\|^2\,dx',\bigg] \leq \frac{1}{d\sigma_K h^2},$$

concluding the proof of Lemma 29.

We remark that while $\nabla\tau(x) = 0$ when $B(x,r)\in\mathcal{X}$, near the boundary the upper bound we derived by using Cauchy-Schwarz appears tight.

Proof of Lemma 30. By Jensen's inequality and Lemma 29,

$$\begin{aligned}\Big|\Lambda_h f(x)\Big|^2 &\leq \frac{1}{h^d\tau_h(x)}\int_{\mathcal{X}}\eta_h(x',x)\big[f(x')\big]^2\,dx' \\ &\leq \frac{1}{a_3h^dp_{\min}}\int_{\mathcal{X}}\eta_h(x',x)\big[f(x')\big]^2p(x')\,dx'.\end{aligned}$$

Then, integrating over x, and recalling that $\int_{B(0,1)} \psi(\|z\|) = 1$ as shown in the proof of Lemma 29, we have

$$\begin{aligned}
\|\Lambda_h f\|_P^2 &\le \frac{1}{a_3 h^d p_{\min}} \int_{\mathcal{X}} \int_{\mathcal{X}} \eta_h(x', x) \big[f(x')\big]^2 p(x') p(x)\, dx'\, dx \\
&\le \frac{p_{\max}}{a_3 h^d p_{\min}} \int_{\mathcal{X}} \big[f(x')\big]^2 p(x') \left(\int_{\mathcal{X}} \eta_h(x', x)\, dx \right) dx' \\
&\le \frac{p_{\max}}{a_3 p_{\min}} \int_{\mathcal{X}} \big[f(x')\big]^2 p(x') \left(\int_{B(0,1)} \psi(\|z\|)\, dz \right) dx' \\
&= \frac{p_{\max}}{a_3 p_{\min}} \|f\|_P^2.
\end{aligned}$$

To establish (B.37), noting that $\Lambda_h a = a$ for any $a \in \mathbb{R}$, we have that

$$\begin{aligned}
\big|\Lambda_r f(x) - f(x)\big|^2 &= \left[\frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \eta_h(x', x) \big(f(x') - f(x)\big)\, dx' \right]^2 \\
&\le \frac{1}{h^{2d} \tau_h^2(x)} \left[\int_{\mathcal{X}} \eta_h(x', x)\, dx' \right] \cdot \left[\int_{\mathcal{X}} \eta_h(x', x) \big(f(x') - f(x)\big)^2\, dx' \right] \\
&= \frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \eta_h(x', x) \big(f(x') - f(x)\big)^2\, dx'. \\
&\le \frac{1}{h^d \tau_h(x) p_{\min}} \int_{\mathcal{X}} \eta_h(x', x) \big(f(x') - f(x)\big)^2 p(x')\, dx'.
\end{aligned}$$

From here, we can use the lower bound $\tau_h(x) \ge a_3$ stated in Lemma 29, as well as the upper bound $\eta_h(x', x) \le (1/\sigma_K) K(\|x' - x\|/h)$, to deduce

$$\big|\Lambda_r f(x) - f(x)\big|^2 \le \frac{1}{h^d a_3 \sigma_K p_{\min}} \int_{\mathcal{X}} K\left(\frac{\|x' - x\|}{h} \right) \big(f(x') - f(x)\big)^2 p(x')\, dx'.$$

Then integrating over $\mathcal{X}$ with respect to p yields (B.37).

B.4.4 Random functionals

We will use Lemma 31 in the proof of Proposition 23.

Lemma 31 (cf. Lemma 3.4 of Burago et al. [2014]). *Let $U \subseteq \mathcal{X}$ be a measurable subset such that $\mathrm{vol}(U) > 0$, and $\mathrm{diam}(U) \le 2A_0\widetilde{\delta}$. Then, letting $a = (\widetilde{P}_n(U))^{-1} \cdot \int_U f(x) \widetilde{p}_n(x)\, dx$ be the average of f over U, it holds that*

$$\int_U \Big| f(x) - a \Big|^2 \widetilde{p}_n(x)\, dx \le A_3 r^2 \widetilde{E}_r(f, U).$$

Now we relate $\widetilde{E}_r(f)$ and $E_r(f)$. Some standard calculations show that for $A_1 := 3A_0/p_{\min}$,

$$\big(1 - A_1(\theta + \widetilde{\delta})\big) E_r(f) \le \widetilde{E}_r(f) \le \big(1 + A_1(\theta + \widetilde{\delta})\big) E_r(f), \tag{B.38}$$

as well as implying that the norms $\|f\|_P$ and $\|f\|_n$ satisfy

$$\big(1 - A_1(\theta + \widetilde{\delta})\big) \|f\|_P^2 \le \|f\|_{\widetilde{P}_n}^2 \le \big(1 + A_1(\theta + \widetilde{\delta})\big) \|f\|_P^2. \tag{B.39}$$

Lemma 32 relates the graph Sobolev semi-norm $b_r(\widetilde{\mathcal{P}} f)$ to the non-local energy $\widetilde{E}_r(f)$.

Lemma 32 (**cf. Lemma 13 of García Trillos et al. [2019a], Lemma 4.3 of Burago et al. [2014]**). *For any $f \in L^2(\mathcal{X})$,*

$$b_r(\widetilde{\mathcal{P}} f) \leq \Big(1 + A_9 \frac{\widetilde{\delta}}{r}\Big) \widetilde{E}_{r+2A_0\widetilde{\delta}}(f).$$

In Lemma 33, we establish the reverse of Lemma 32.

Lemma 33 (**cf. Lemma 14 of García Trillos et al. [2019a]**). *For any $u \in L^2(P_n)$,*

$$\widetilde{E}_{r-2A_0\widetilde{\delta}}\big(\widetilde{\mathcal{P}}^\star u\big) \leq \Big(1 + A_3 \frac{\widetilde{\delta}}{r}\Big) b_r(u).$$

Proof of Lemma 31. A symmetrization argument implies that

$$\int_U \Big|f(x) - a\Big|^2 \widetilde{p}_n(x)\, dx = \frac{1}{2\widetilde{P}_n(U)} \int_U \int_U \big|f(x') - f(x)\big|^2 \widetilde{p}_n(x') \widetilde{p}_n(x)\, dx'\, dx \tag{B.40}$$

Now, since x' and x belong to U, we have that $\|x' - x\| \leq 2A_0\widetilde{\delta}$. Set $V = B(x, r) \cap B(x', r)$, and note that $B(x, r - 2A_0\widetilde{\delta}) \subseteq V$. Moreover, $r - 2A_0\widetilde{\delta} \leq r \leq c_0$ by assumption. Therefore by (B.35),

$$\mathrm{vol}\big(V \cap \mathcal{X}\big) \geq \mathrm{vol}\big(B(x, r - 2A_0\widetilde{\delta}) \cap \mathcal{X}\big) \geq a_3 \nu_d (r - 2A_0\widetilde{\delta})^d \geq \frac{a_3 \nu_d}{2^d} r^d$$

where the last inequality follows since $\widetilde{\delta} \leq \frac{1}{4A_0} r$. Using the triangle inequality

$$\big|f(x') - f(x)\big|^2 \leq 2\big(\big|f(x') - f(z)\big|^2 + \big|f(z) - f(x)\big|^2\big)$$

we have that for any x and x' in U,

$$\begin{aligned}
\big|f(x') - f(x)\big|^2 &\leq \frac{2}{\mathrm{vol}(V \cap \mathcal{X})} \int_{V \cap \mathcal{X}} \big|f(x') - f(z)\big|^2 + \big|f(z) - f(x)\big|^2\, dz \\
&\leq \frac{2^{d+1}}{a_3 \nu_d r^d} \int_{V \cap \mathcal{X}} \big|f(x') - f(z)\big|^2 + \big|f(z) - f(x)\big|^2\, dz \\
&\leq \frac{2^{d+2}}{K(1) a_3 \nu_d r^d p_{\min}} \Big(F(x') + F(x)\Big),
\end{aligned} \tag{B.41}$$

where in the last inequality we set

$$F(x) := \int_{\mathcal{X}} K\Big(\frac{\|z - x\|}{r}\Big) \big(f(z) - f(x)\big)^2 \widetilde{p}_n(x)\, dx,$$

and use the facts that $\widetilde{p}_n(x) \geq p_{\min}/2$, that $K(\|z - x\|/r) \geq K(1)$ for all $z \in B(x, r)$.

Plugging the upper bound (B.41) back into (B.40), we have that

$$\begin{aligned}
\int_U \Big|f(x) - a\Big|^2 \widetilde{p}_n(x)\, dx &\leq \frac{2^{d+2}}{K(1) a_3 \nu_d r^d} \int_U F(x) \widetilde{p}_n(x)\, dx \\
&= \frac{2^{d+2}}{K(1) a_3 \nu_d} r^2 \widetilde{E}_r(f, U),
\end{aligned}$$

and Lemma 31 follows by taking $A_3 := 2^{d+2}/(K(1) a_3 \nu_d)$.

Proof of Lemma 32. Recalling that $\big(\widetilde{\mathcal{P}}f\big)(X_i) = n \cdot \int_{U_i} f(x)\widetilde{p}_n(x)\,dx$, by Jensen's inequality,

$$\Big(\big(\widetilde{\mathcal{P}}f\big)(X_i) - \big(\widetilde{\mathcal{P}}f\big)(X_j)\Big)^2 \le n^2 \cdot \int_{U_i}\int_{U_j} \big(f(x') - f(x)\big)^2 \widetilde{p}_n(x')\widetilde{p}_n(x)\,dx'\,dx.$$

Additionally, the non-increasing and Lipschitz properties of K imply that for any $x \in U_i$ and $x' \in U_j$,

$$K\bigg(\frac{\|X_i - X_j\|}{r}\bigg) \le K\bigg(\frac{\big(\|x' - x\| - 2A_0\widetilde{\delta}\big)_+}{r}\bigg) \le K\bigg(\frac{\|x' - x\|}{r + 2A_0\widetilde{\delta}}\bigg) + \frac{2L_K A_0\widetilde{\delta}}{r}\mathbf{1}\Big\{\|x' - x\| \le r + 2A_0\widetilde{\delta}\Big\}.$$

As a result, the graph Dirichlet energy is upper bounded as follows:

$$\begin{aligned}
b_r(\widetilde{\mathcal{P}}f) &= \frac{1}{n^2 r^{d+2}} \sum_{i,j=1}^n \Big(\big(\widetilde{\mathcal{P}}f\big)(X_i) - \big(\widetilde{P}f\big)(X_j)\Big)^2 K\bigg(\frac{\|X_i - X_j\|}{r}\bigg) \\
&\le \frac{1}{r^{d+2}} \sum_{i,j=1}^n \int_{U_i}\int_{U_j} \big(f(x') - f(x)\big)^2 \widetilde{p}_n(x')\widetilde{p}_n(x) K\bigg(\frac{\|X_i - X_j\|}{r}\bigg)\,dx'\,dx \\
&\le \frac{1}{r^{d+2}} \sum_{i,j=1}^n \int_{U_i}\int_{U_j} \big(f(x') - f(x)\big)^2 \widetilde{p}_n(x')\widetilde{p}_n(x)\bigg[K\bigg(\frac{\|x' - x\|}{r + 2A_0\widetilde{\delta}}\bigg) + \frac{2L_K A_0\widetilde{\delta}}{r}\mathbf{1}\Big\{\|x' - x\| \le r + 2\widetilde{\delta}\Big\}\bigg]\,dx'\,dx \\
&= \Big(1 + 2A_0\frac{\widetilde{\delta}}{r}\Big)^{d+2}\bigg[\widetilde{E}_{r+2A_0\widetilde{\delta}}(f) + \frac{2L_K A_0\widetilde{\delta}}{r}\widetilde{E}_{r+2A_0\widetilde{\delta}}(f; \mathbf{1}_{[0,1]})\bigg],
\end{aligned}$$

for $\mathbf{1}_{[0,1]}(t) = \mathbf{1}\{0 \le t \le 1\}$. But by assumption $\widetilde{E}_{r+2A_0\widetilde{\delta}}(f; \mathbf{1}_{[0,1]}) \le 1/(K(1))\widetilde{E}_{r+2A_0\widetilde{\delta}}(f)$, and so we obtain

$$b_r(\widetilde{\mathcal{P}}f) \le \Big(1 + 2A_0\frac{\widetilde{\delta}}{r}\Big)^{d+2}\Big(1 + \frac{2L_K A_0\widetilde{\delta}}{rK(1)}\Big)\widetilde{E}_{r+2A_0\widetilde{\delta}}(f);$$

the Lemma follows upon choosing $A_9 := A_0(2^{d+4} + \frac{4L_K}{K(1)})$.

Proof of Lemma 33. For brevity, we write $\widetilde{r} := r - 2A_0\widetilde{\delta}$. We begin by expanding the energy $\widetilde{E}_{\widetilde{r}}\big(\widetilde{\mathcal{P}}^\star u\big)$ as a double sum of double integrals,

$$\widetilde{E}_{\widetilde{r}}\big(\widetilde{\mathcal{P}}^\star u\big) = \frac{1}{\widetilde{r}^{d+2}} \sum_{i=1}^n \sum_{j=1}^n \int_{U_i}\int_{U_j} \Big(u(X_i) - u(X_j)\Big)^2 K\bigg(\frac{\|x' - x\|}{\widetilde{r}}\bigg)\widetilde{p}_n(x')\widetilde{p}_n(x)\,dx'\,dx.$$

We next use the Lipschitz property of the kernel K—in particular that for $x \in U_i$ and $x' \in U_j$,

$$K\bigg(\frac{\|x' - x\|}{\widetilde{r}}\bigg) \le K\bigg(\frac{\|X_i - X_j\|}{r}\bigg) + \frac{2A_0 L_K\widetilde{\delta}}{\widetilde{r}} \cdot \mathbf{1}\bigg\{\frac{\|x' - x\|}{\widetilde{r}} \le 1\bigg\},$$

—to conclude that

$$\begin{aligned}
\widetilde{E}_{\widetilde{r}}\big(\widetilde{\mathcal{P}}^\star u\big) &\le \frac{1}{n^2\widetilde{r}^{d+2}} \sum_{i=1}^n \sum_{j=1}^n \Big(u(X_i) - u(X_j)\Big)^2 K\bigg(\frac{\|X_i - X_j\|}{r}\bigg) + \frac{2A_0 L_K\widetilde{\delta}}{\widetilde{r}}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u, \mathbf{1}_{[0,1]}) \\
&\le \Big(1 + 2^{d+2}A_0\frac{\widetilde{\delta}}{r}\Big) b_r(u) + \frac{2A_0 L_K\widetilde{\delta}}{\widetilde{r}}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u, \mathbf{1}_{[0,1]}) \\
&\le \Big(1 + 2^{d+2}A_0\frac{\widetilde{\delta}}{r}\Big) b_r(u) + \frac{4A_0 L_K\widetilde{\delta}}{K(1)r}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u).
\end{aligned}$$

In other words,

$$\begin{aligned}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^{\star}u) &\le \Big(1-\frac{4A_0L_K\widetilde{\delta}}{K(1)r}\Big)^{-1}\Big(1+2^{d+2}A_0\frac{\widetilde{\delta}}{r}\Big)b_r(u)\\ &\le \Big(1+\frac{\widetilde{\delta}}{r}\Big(\frac{8A_0L_K}{K(1)}+2^{d+3}\Big)\Big)b_r(u),\end{aligned}$$

where the second inequality follows from the algebraic identities $(1-t)^{-1}\le(1+2t)$ for any $0<t<1/2$ and $(1+s)(1+t)<1+2s+t$ for any $0<t<1$ and $s>0$. The Lemma follows upon choosing $A_3 := \frac{8A_0L_K}{K(1)}+2^{d+3}$.

B.4.5 Proof of Propositions 22 and 23

Proof of Proposition 22. Part (1) of Proposition 22 follows from

$$\begin{aligned}\sigma_K D_2(\Lambda_{r-2A_0\widetilde{\delta}}\widetilde{\mathcal{P}}^{\star}u) &\overset{(i)}{\le} A_8E_{r-2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^{\star}u)\\ &\overset{(ii)}{\le} A_8\Big(1+A_1(\theta+\widetilde{\delta})\Big)\widetilde{E}_{r-2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^{\star}u)\\ &\overset{(iii)}{\le} A_8\Big(1+A_1(\theta+\widetilde{\delta})\Big)\cdot\Big(1+A_3\frac{\widetilde{\delta}}{r}\Big)b_r(u),\end{aligned}$$

where (i) follows from Lemma 28, (ii) follows from (B.38), and (iii) follows from Lemma 33.

Part (2) of Proposition 22 follows from

$$\begin{aligned}b_r(\widetilde{\mathcal{P}}f) &\overset{(iv)}{\le} \Big(1+A_9\frac{\widetilde{\delta}}{r}\Big)\widetilde{E}_{r+2A_0\widetilde{\delta}}(f)\\ &\overset{(v)}{\le} \Big(1+A_1(\theta+\widetilde{\delta})\Big)\Big(1+A_9\frac{\widetilde{\delta}}{r}\Big)E_{r+2A_0\widetilde{\delta}}(f)\\ &\overset{(vi)}{\le} \Big(1+A_1(\theta+\widetilde{\delta})\Big)\cdot\Big(1+A_9\frac{\widetilde{\delta}}{r}\Big)\cdot\Big(\frac{C_5p_{\max}^2}{p_{\min}^2}\Big)\cdot\sigma_K D_2(f),\end{aligned}$$

where (iv) follows from Lemma 32, (v) follows from (B.38), and (vi) follows from the proof of Lemma 3.

Proof of Proposition 23. *Proof of (1).* We begin by upper bounding $\|\widetilde{\mathcal{P}}f\|_n$. By the Cauchy-Schwarz inequality and the bound on $\|\widetilde{p}_n-p\|_\infty$ in (B.26),

$$\begin{aligned}\Big|\widetilde{\mathcal{P}}f(X_i)\Big|^2 &= n^2\Big|\int_{U_i}f(x)\widetilde{p}_n(x)\,dx\Big|^2\\ &\le n\int_{U_i}\big|f(x)\big|^2\widetilde{p}_n(x)\,dx\\ &\le n\Big(1+A_1(\theta+\widetilde{\delta})\Big)\Big[\int_{U_i}\big|f(x)\big|^2p(x)\,dx+A_1(\theta+\widetilde{\delta})\int_{U_i}\big|f(x)\big|^2p(x)\,dx\Big],\end{aligned}$$

and summing over $i=1,\dots,n$, we obtain

$$\|\widetilde{\mathcal{P}}f\|_n^2 \le \Big(1+A_1(\theta+\widetilde{\delta})\Big)\|f\|_P^2. \tag{B.42}$$

Now, noticing that $\|\widetilde{\mathcal{P}}f\|_n = \|\widetilde{P}^\star \widetilde{P} f\|_{\widetilde{P}_n}$, we can use the upper bound (B.42) to show that

$$\begin{aligned}
\left|\|\widetilde{\mathcal{P}}f\|_n^2 - \|f\|_P^2\right| &\le \left|\|\widetilde{\mathcal{P}}f\|_n^2 - \|f\|_{\widetilde{P}_n}^2\right| + \left|\|f\|_{\widetilde{P}_n}^2 - \|f\|_P^2\right| \\
&\overset{(i)}{\le} \left|\|\widetilde{\mathcal{P}}f\|_n^2 - \|f\|_{\widetilde{P}_n}^2\right| + A_1(\theta+\widetilde{\delta})\|f\|_P^2 \qquad \text{(B.43)} \\
&\overset{(ii)}{\le} 2\sqrt{1+A_1(\theta+\widetilde{\delta})}\left|\|\widetilde{\mathcal{P}}f\|_n - \|f\|_{\widetilde{P}_n}\right| \cdot \|f\|_P + A_1(\theta+\widetilde{\delta})\|f\|_P^2 \\
&\le 2\sqrt{1+A_1(\theta+\widetilde{\delta})}\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\|_{\widetilde{P}_n} \cdot \|f\|_P + A_1(\theta+\widetilde{\delta})\|f\|_P^2, \qquad \text{(B.44)}
\end{aligned}$$

where (i) follows from (B.39) and (ii) follows from (B.39) and (B.42).

It remains to upper bound $\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\|_{\widetilde{P}_n}^2$. Noting that $\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f$ is piecewise constant over the cells U_i, we have

$$\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\|_{\widetilde{P}_n}^2 = \sum_{i=1}^n \int_{U_i} \left(f(x) - n\cdot\int_{U_i} f(x')\widetilde{p}_n(x')\,dx'\right)^2 \widetilde{p}_n(x)\,dx.$$

From Lemma 31, we have that for each $i = 1,\dots,n$,

$$\int_{U_i} \left(f(x) - n\cdot\int_{U_i} f(x')\widetilde{p}_n(x')\,dx'\right)^2 \widetilde{p}_n(x)\,dx \le A_3 r^2 \widetilde{E}_r(f, U_i).$$

Summing up over i on both sides of the inequality gives

$$\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\|_{\widetilde{P}_n}^2 \le A_3 r^2 \widetilde{E}_r(f,\mathcal{X}) \le A_3\Big(1+A_1(\theta+\widetilde{\delta})\Big)\cdot\Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big)\cdot\sigma_K r^2 D_2(f),$$

where the latter inequality follows from the proof of Proposition 22, Part (2). Then Proposition 23, Part (1) follows by plugging this inequality into (B.44) and taking

$$A_5 := 2\sqrt{A_3}\Big(1+A_1(\theta+\widetilde{\delta})\Big)\Big(\frac{\sqrt{C_5}p_{\max}}{p_{\min}}\Big)\cdot\sqrt{\sigma_K}.$$

Proof of (2). By the triangle inequality and (B.39),

$$\begin{aligned}
\left|\|\widetilde{\mathcal{I}}u\|_P^2 - \|u\|_n^2\right| &\le \left|\|\widetilde{\mathcal{I}}u\|_P^2 - \|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2\right| + \left|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 - \|u\|_n^2\right| \\
&\le A_1(\theta+\widetilde{\delta})\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 + \left|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 - \|u\|_n^2\right| \\
&= A_1(\theta+\widetilde{\delta})\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 + \Big(\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} + \|u\|_n\Big)\cdot\left|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} - \|u\|_n\right| \qquad \text{(B.45)}
\end{aligned}$$

To upper bound the second term in the above expression, we first note that $\|u\|_n = \|\widetilde{\mathcal{P}}^\star u\|_{\widetilde{P}_n}$, and thus

$$\begin{aligned}
\left|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} - \|u\|_n\right| &= \left|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} - \|\widetilde{\mathcal{P}}^\star u\|_{\widetilde{P}_n}\right| \\
&\overset{(iii)}{\le} \|\Lambda_{\widetilde{r}}\widetilde{\mathcal{P}}^\star u - \widetilde{\mathcal{P}}^\star u\|_{\widetilde{P}_n} \\
&\overset{(iv)}{\le} \widetilde{r}\sqrt{\frac{1}{a_3\sigma_K p_{\min}} E_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u)} \\
&\overset{(v)}{\le} \widetilde{r}\sqrt{\frac{1+A_1(\theta+\widetilde{\delta})}{a_3\sigma_K p_{\min}}\Big(1+A_3\frac{\widetilde{\delta}}{r}\Big) b_r(u)}, \qquad \text{(B.46)}
\end{aligned}$$

where (iii) follows by the triangle inequality, (iv) follows from Lemma 30, and (v) follows from (B.38) and Lemma 33. On the other hand, by (B.39) and Lemma 30,

$$\begin{aligned}\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 &\leq \Big(1 + A_1(\theta + \widetilde{\delta})\Big)\|\widetilde{\mathcal{I}}u\|_P^2 \\ &\leq \frac{p_{\max}}{a_3 p_{\min}} \cdot \Big(1 + A_1(\theta + \widetilde{\delta})\Big)\|\widetilde{\mathcal{P}}^\star u\|_P^2 \\ &\leq \frac{p_{\max}}{a_3 p_{\min}} \cdot \Big(1 + A_1(\theta + \widetilde{\delta})\Big)^2\|\widetilde{\mathcal{P}}^\star u\|_{\widetilde{P}_n}^2 \\ &= \frac{p_{\max}}{a_3 p_{\min}} \cdot \Big(1 + A_1(\theta + \widetilde{\delta})\Big)^2\|u\|_n^2.\end{aligned}$$

Plugging this estimate along with (B.46) back into (B.45), we obtain part (2) of Proposition 23, upon choosing

$$A_6 := \left(3\sqrt{\frac{2p_{\max}}{p_{\min}}} + 1\right)\sqrt{\frac{4}{a_3\sigma_K p_{\min}}}, \quad A_7 := 4A_1\frac{p_{\max}}{a_3 p_{\min}}.$$

B.5 Bound on the empirical norm

In Lemma 34, we lower bound $\|f_0\|_n^2$ by (a constant times) the $L^2(\mathcal{X})$ norm of f.

Lemma 34. *Fix $\delta \in (0,1)$ Suppose P satisfies (P2). If $f \in H^1(\mathcal{X}, M)$ is lower bounded in $L^2(\mathcal{X})$ norm,*

$$\|f\|_{L^2(\mathcal{X})} \geq \frac{C_6 M}{\delta} \cdot \max\Big\{n^{-1/2}, n^{-1/d}\Big\}. \tag{B.47}$$

Then with probability at least $1 - 5\delta$,

$$\|f\|_n^2 \geq \delta \cdot \mathbb{E}\Big[\|f\|_n^2\Big]. \tag{B.48}$$

Proof of Lemma 34. In this proof, we will find it more convenient to deal with the parameterization $b = 1/\delta$. To establish (B.48), it is sufficient to show that

$$\mathbb{E}\big[\|f\|_n^4\big] \leq \left(1 + \frac{1}{b^2}\right) \cdot \big(\mathbb{E}\big[\|f\|_n^2\big]\big)^2;$$

then (B.48) follows from the Paley-Zygmund inequality (Lemma 38). Since $p \leq p_{\max}$ is uniformly bounded, we can relate $\mathbb{E}\big[\|f\|_n^4\big]$ to the $L^4(\mathcal{X})$-norm,

$$\mathbb{E}\big[\|f\|_n^4\big] = \frac{(n-1)}{n}\Big(\mathbb{E}\Big[\|f\|_n^2\Big]\Big)^2 + \frac{\mathbb{E}\Big[\big(f(X_1)\big)^4\Big]}{n} \leq \Big(\mathbb{E}\Big[\|f\|_n^2\Big]\Big)^2 + p_{\max}\frac{\|f\|_{L^4(\mathcal{X})}^4}{n}.$$

We will use the Sobolev inequalities as a tool to show that $\|f\|_{L^4(\mathcal{X})}^4/n \leq \big(\mathbb{E}[\|f\|_n^2]\big)^2/(b^2 p_{\max})$, whence the claim of the Lemma is shown. The nature of the inequalities we use depend on the value of d. In particular, we will use the following relationships between norms: for any $f \in H^1(\mathcal{X}; M)$,

$$\left.\begin{array}{ll} \sup_{x\in\mathcal{X}}|f(x)|, & d = 1 \\ \|f\|_{L^q(\mathcal{X})}, & d = 2, \text{ for all } 0 < q < \infty \\ \|f\|_{L^q(\mathcal{X})}, & d \geq 3, \text{ for all } 0 < q \leq 2d/(d-2) \end{array}\right\} \leq C_7 \cdot M.$$

(See Theorem 6 in Section 5.6.3 of Evans [2010] for a complete statement and proof of the various Sobolev inequalities.)

As a result, we divide our analysis into three cases: (i) the case where $d < 2$, (ii) the case where $d > 2$, and (iii) the borderline case $d = 2$.

Case 1: $d < 2$. The $L^4(\mathcal{X})$-norm of f can be bounded in terms of the $L^2(\mathcal{X})$ norm,

$$\|f\|_{L^4(\mathcal{X})}^4 \leq \left(\sup_{x\in\mathcal{X}} |f(x)|\right)^2 \cdot \int_{\mathcal{X}} [f(x)]^2\, dx \leq C_7^2 M^2 \cdot \|f\|_{L^2(X)}^2.$$

Since by assumption

$$\|f\|_{L^2(\mathcal{X})}^2 \geq C_6^2 \cdot b^2 \cdot M^2 \cdot \frac{1}{n},$$

we have

$$p_{\max}\frac{\|f\|_{L^4(\mathcal{X})}^4}{n} \leq C_7^2 M^2 p_{\max} \cdot \frac{\|f\|_{L^2(\mathcal{X})}^2}{n} \leq \frac{C_7 p_{\max}}{C_6^2 b^2}\|f\|_{L^2(\mathcal{X})}^4 \leq \frac{\left(\mathbb{E}\left[\|f\|_n^2\right]\right)^2}{b^2},$$

where the last inequality follows by taking $C_6 \geq C_7\sqrt{p_{\max}/p_{\min}}$.

Case 2: $d > 2$. Let $\theta = 2 - d/2$ and $q = 2d/(d-2)$. Noting that $4 = 2\theta + (1-\theta)q$, Lyapunov's inequality implies

$$\|f\|_{L^4(\mathcal{X})}^4 \leq \|f\|_{L^2(\mathcal{X})}^{2\theta} \cdot \|f\|_{L^q(\mathcal{X})}^{(1-\theta)q} \leq \|f\|_{L^2(\mathcal{X})}^4 \cdot \left(\frac{C_7\|f\|_{H^1(\mathcal{X})}}{\|f\|_{L^2(\mathcal{X})}}\right)^d.$$

By assumption, $\|f\|_{L^2(\mathcal{X})} \geq C_6 b \|f\|_{H^1(\mathcal{X})} n^{-1/d}$, and therefore

$$p_{\max}\frac{\|f\|_{L^4(\mathcal{X})}^4}{n} \leq \|f\|_{L^2(\mathcal{X})}^4 p_{\max} \cdot \left(\frac{C_7\|f\|_{H^1(\mathcal{X})}}{n^{1/d}\|f\|_{L^2(\mathcal{X})}}\right)^d \leq \frac{C_7^d p_{\max}\|f\|_{L^2(\mathcal{X})}^4}{C_6^d b^d} \leq \frac{\left(\mathbb{E}\left[\|f\|_n^2\right]\right)^2}{b^2}.$$

where the last inequality follows by taking $C_6 \geq C_7(p_{\max}/p_{\min})^{1/d}$, and keeping in mind that $d > 2$ and $b \geq 1$.

Case 3: $d = 2$. Fix $t \in (1/2, 1)$, and suppose that

$$\|f\|_{L^2(\mathcal{X})} \geq \frac{C_6 M}{\delta} \cdot n^{-t/2}. \tag{B.49}$$

Putting $q = 2/(1-t)$, we have that $\|f\|_{L^q(\mathcal{X})} \leq C_7 \cdot M$, and it follows from derivations similar to those in Case 2 that $\|f\|_{L^4(\mathcal{X})}^4/n \leq \left(\mathbb{E}[\|f\|_n^2]\right)^2/(b^2 p_{\max})$ when $C_6 \geq C_7\sqrt{p_{\max}/p_{\min}}$.

Now, suppose $f \in L^4(\mathcal{X})$ satisfies (B.49) only when $t = 1$. For each $k = 1, 2, \ldots$ let $f_k := n^{1/(2k)} f$, so that each f_k satisfies (B.49) with respect to $t = 1 - 1/k$. Clearly $\|f_k - f\|_{L^4(\mathcal{X})} \to 0$ as $k \to \infty$, and therefore

$$\frac{1}{n}\|f\|_{L^4(\mathcal{X})}^4 = \frac{1}{n}\lim_{k\to\infty}\|f_k\|_{L^4(\mathcal{X})}^4 \leq \frac{1}{b^2 p_{\max}}\lim_{k\to\infty}\left(\mathbb{E}[\|f_k\|_n^2]\right)^2 = \frac{1}{b^2 p_{\max}}\left(\mathbb{E}[\|f\|_n^2]\right)^2.$$

This establishes the claim when $d = 2$, and completes the proof of Lemma 34.

B.6 Graph functionals under the manifold hypothesis

In this section, we restate a few results of García Trillos et al. [2019a], Calder and García Trillos [2019], which are analogous to Lemmas 3 and 4 but cover the case where $\mathcal{X}$ is an m-dimensional submanifold without boundary. As such, the results in this section will hold under the assumption (P3). We refer to García Trillos et al. [2019a], Calder and García Trillos [2019] for the proofs of these results.

Proposition 24 follows from Lemma 5 of García Trillos et al. [2019a] and Markov's inequality.

Proposition 24. *For any $f \in H^1(\mathcal{X})$, with probability at least $1-\delta$,*

$$f^\top L f \leq \frac{C}{\delta} n^2 r^{m+2} |f|^2_{H^1(\mathcal{X})}.$$

In Proposition 25, it is assumed that r, $\widetilde{\delta}$ and θ satisfy the following smallness conditions.

(S1)
$$n^{-1/m} < \widetilde{\delta} \leq \frac{1}{4} r \quad \text{and} \quad C(\theta + \widetilde{\delta}) \leq \frac{1}{2} p_{\min} \quad \text{and} \quad C_4 \big(\log(n)/n\big)^{1/m} \leq r \leq \min\{c_4, 1\}.$$

Proposition 25 (c.f Theorem 2.4 of Calder and García Trillos [2019]). *With probability at least $1 - Cn\exp(-cn\theta^2\widetilde{\delta}^m)$, the following statement holds. For any $k \in \mathbb{N}$ such that*

$$\sqrt{\lambda_k(\Delta_P)} r + C(\theta + \widetilde{\delta}) \leq \frac{1}{2},$$

it holds that

$$nr^{m+2}\lambda_k(\Delta_P)\left(1 - C\left(r(\sqrt{\lambda_k(\Delta_P)}+1) + \frac{\widetilde{\delta}}{r} + \theta\right)\right) \leq \lambda_k(G_{n,r}) \leq nr^{m+2}\lambda_k(\Delta_P)\left(1 + C\left(r(\sqrt{\lambda_k(\Delta_P)}+1) + \frac{\widetilde{\delta}}{r} + \theta\right)\right).$$

Proposition 26 follows from Lemma 3.1 of Calder and García Trillos [2019], along with a union bound.

Proposition 26. *With probability at least $1 - 2Cn\exp(-cp_{\max} n r^m)$, it holds that*

$$D_{\max}(G_{n,r}) \leq Cnr^m.$$

Finally, we note that a Weyl's Law holds for Riemmanian manifolds without boundary, i.e.

$$\lambda_k(\Delta_P) \asymp k^{2/m}.$$

Put $B_{n,r}(k) := \min\{nr^{m+2}k^{2/m}, nr^m\}$. Following parallel steps to the proof of Lemma 4, one can derive from Propositions 25 and 26, and Weyl's Law, that with probability at least $1 - Cn\exp(-cnr^m)$,

$$cB_{n,r}(k) \leq \lambda_k \leq CB_{n,r}(k), \quad \text{for all } 2 \leq k \leq n. \tag{B.50}$$

B.7 Proofs of main results

We are now in a position to prove Theorems 6-10, as well as a few other claims from our main text. In Section B.7.1 we prove all of our results regarding estimation and in Section B.7.2 we prove all of our results regarding testing; in Section B.7.3, Lemmas 35 and 36, we provide some useful estimates on a particular pair of sums that appear repeatedly in our proofs. Throughout, it will be convenient for us to deal with the normalization $\widetilde{\rho} := \rho n r^{d+2}$. We note that in each of our Theorems, the prescribed choice of ρ will always result in $\widetilde{\rho} \leq 1$.

B.7.1 Proof of estimation results

Proof of Theorem 6. We have shown that the inequalities (3.14) and (3.15) are satisfied with probability at least $1 - \delta - C_1 n\exp(-c_1 n r^d)$, and throughout this proof we take as granted that both of these inequalities hold.

Now, set $\widetilde{\rho} = M^{-4/(2+d)} n^{-2/(2+d)}$ as prescribed in Theorem 6, and note that $\widetilde{\rho}^{-d/2} \leq n$ is implied by the assumption $M \leq n^{1/d}$. Therefore from (3.15) and Lemma 35, it follows that

$$\sum_{k=1}^{n} \left(\frac{1}{\rho \lambda_k + 1} \right)^2 \geq 1 + \frac{1}{C_3^2} \sum_{k=2}^{n} \left(\frac{1}{\widetilde{\rho} k^{2/d} + 1} \right)^2 \geq \frac{1}{8 C_3^2} \widetilde{\rho}^{-d/2}.$$

As a result, by Lemma 26 along with (3.14) and (3.15), with probability at least $1 - \delta - C_1 n \exp(-c_1 n r^d) - \exp(-\widetilde{\rho}^{-d/2}/8C_3^2)$ it holds that,

$$\begin{aligned} \|\widehat{f} - f_0\|_n^2 &\leq \frac{C_2}{\delta} \widetilde{\rho} M^2 + \frac{10}{n} + \frac{10}{n} \sum_{k=2}^{n} \left(\frac{1}{c_3 \widetilde{\rho} \min\{k^{2/d}, r^{-2}\} + 1} \right)^2 \\ &\leq \frac{C_2}{\delta} \widetilde{\rho} M^2 + \frac{10}{n} + \frac{10}{n c_3^2} \sum_{k=2}^{n} \left(\frac{1}{\widetilde{\rho} k^{2/d} + 1} \right)^2 + \frac{10 r^4}{c_3^2 \widetilde{\rho}^2}. \end{aligned} \tag{B.51}$$

The first term on the right hand side of (B.51) is a bias term, while the second, third, and fourth terms each contribute to the variance. Of these, under our assumptions the third term dominates, as we show momentarily. First, we use Lemma 35 to get an upper bound on this variance term,

$$\sum_{k=2}^{n} \left(\frac{1}{\widetilde{\rho} k^{2/d} + 1} \right)^2 \leq 4 \widetilde{\rho}^{-d/2}.$$

Then plugging this upper bound back into (B.51), we have that

$$\begin{aligned} \|\widehat{f} - f_0\|_n^2 &\leq \frac{C_2}{\delta} \widetilde{\rho} M^2 + \frac{10}{n} + \frac{40 \widetilde{\rho}^{-d/2}}{c_3^2 n} + \frac{10 r^4}{c_3^2 \widetilde{\rho}^2} \\ &= \left(\frac{C_2}{\delta} + \frac{40}{c_3^2} \right) M^{2d/(2+d)} n^{-2/(2+d)} + \frac{10}{n} + \frac{10}{c_3^2} r^4 M^{8/(2+d)} n^{4/(2+d)} \\ &\leq \left(\frac{C_2}{\delta} + \frac{50}{c_3^2} \right) M^{2d/(2+d)} n^{-2/(2+d)}, \end{aligned}$$

with the last inequality following from (R1) and the assumption $M \geq n^{-1/2}$. This completes the proof of Theorem 6.

Proof of Theorem 7. We first establish that $\widehat{f}$ achieves nearly-optimal rates when $d = 4$, and then establish the claimed sub-optimal rates when $d > 4$.

Nearly-optimal rates when $d = 4$.

Continuing on from (B.51), from Lemma 35 we have that

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C_2}{\delta} \widetilde{\rho} M^2 + \frac{10}{n} + \frac{10}{n c_3^2 \widetilde{\rho}^2} + \frac{10 \log n}{n c_3^2 \widetilde{\rho}^2} + \frac{10 r^4}{c_3^2 \widetilde{\rho}^2}.$$

Setting $r = (C_0 \log(n)/n)^{1/4}$, we obtain

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C_2}{\delta} \widetilde{\rho} M^2 + \frac{10}{n} + \frac{10}{n c_3^2 \widetilde{\rho}^2} + \frac{10 \log n}{n c_3^2 \widetilde{\rho}^2} + \frac{10 C_0 \log n}{n c_3^2 \widetilde{\rho}^2},$$

and choosing $\widetilde{\rho} = M^{-2/3} (\log n / n)^{1/3}$ yields

$$\|\widehat{f} - f_0\|_n^2 \leq \left(\frac{C_2}{\delta} + \frac{20}{c_3^2} + \frac{10 C_0}{c_3^2} \right) M^{4/3} \left(\frac{\log n}{n} \right)^{1/3} + \frac{10}{n}.$$

Suboptimal rates when $d > 4$.

Once again continuing on from (B.51), from Lemma 35 we have that

$$\|\widehat{f} - f_0\|_n^2 \le \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{nc_3^2\widetilde{\rho}^{d/2}} + \frac{10}{n^{4/d}\widetilde{\rho}^2 c_3^2} + \frac{10r^4}{\widetilde{\rho}^2 c_3^2}.$$

Setting $r = (C_0 \log n/n)^{1/d}$, we obtain

$$\|\widehat{f} - f_0\|_n^2 \le \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{n\widetilde{\rho}^{d/2}c_3^2} + \frac{10}{n^{4/d}\widetilde{\rho}^2 c_3^2} + \frac{10C_0^{4/d}(\log n)^{4/d}}{n^{4/d}\widetilde{\rho}^2 c_3^2},$$

and choosing $\widetilde{\rho} = M^{-2/3}n^{-4/(3d)}$ yields

$$\|\widehat{f} - f_0\|_n^2 \le \left(\frac{C_2}{\delta} + \frac{10}{c_3^2} + \frac{10C_0^{4/d}}{c_3^2}\right)M^{4/3}\left(\frac{\log n}{n^{1/3}}\right)^{4/d} + \frac{10}{c_3^{d/2}}M^{d/3}n^{-1/3} + \frac{10}{n}.$$

Bounds on $L^2(\mathcal{X})$ error under Lipschitz assumption. Let $V_1, \dots, V_n$ denote the Voronoi tesselation of $\mathcal{X}$ with respect to $X_1, \dots, X_n$. Extend $\widehat{f}$ over $\mathcal{X}$ by taking it piecewise constant over the Voronoi cells, i.e.

$$\widehat{f}(x) := \sum_{i=1}^n \widehat{f}_i \cdot \mathbf{1}\{x \in V_i\}.$$

Note that we are abusing notation slightly by also using $\widehat{f}$ to refer to this extension.

In Proposition 27, we establish that the out-of-sample error $\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}$ will not be too much larger than the in-sample error $\|\widehat{f} - f_0\|_n$.

Proposition 27. *Suppose f_0 satisfies $|f_0(x') - f_0(x)| \le M\|x' - x\|$ for all $x', x \in \mathcal{X}$. Then for all n sufficiently large, with probability at least $1 - \delta$ it holds that*

$$\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}^2 \le C\log(1/\delta)\left(\log(n) \cdot \|\widehat{f} - f_0\|_n^2 + M^2\left(\frac{\log n}{n}\right)^{2/d}\right).$$

Note that $n^{-2/d} \ll n^{-2/(2+d)}$. Therefore Proposition 27 together with Theorem 6 implies that with high probability, $\widehat{f}$ achieves the nearly-optimal (up to a factor of $\log n$) estimation rates out-of-sample error—that is, $\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}^2 \le C\log(n)M^{2d/(2+d)}n^{-2/(2+d)}$—as long as $M \le Cn^{1/d}$. This justifies one of our remarks after Theorem 6.

Proof of Proposition 27. Suppose $x \in V_i$, so that we can upper bound the pointwise squared error $|\widehat{f}(x) - f(x)|^2$ using the triangle inequality:

$$\big(\widehat{f}(x) - f_0(x)\big)^2 = \big(\widehat{f}(X_i) - f_0(x)\big)^2 \le 2\big(\widehat{f}(X_i) - f_0(X_i)\big)^2 + 2\big(f_0(X_i) - f_0(x)\big)^2.$$

Integrating both sides of the inequality, we have

$$\begin{aligned}\int_{\mathcal{X}}\big(\widehat{f}(x) - f_0(x)\big)^2\,dx &\le 2\sum_{i=1}^n \int_{V_i}\Big(\widehat{f}(X_i) - f_0(X_i)\Big)^2\,dx + 2\sum_{i=1}^n\int_{V_i}\Big(f_0(X_i) - f_0(x)\Big)^2\,dx \\ &= 2\sum_{i=1}^n \mathrm{vol}(V_i)\Big(\widehat{f}(X_i) - f_0(X_i)\Big)^2 + 2\sum_{i=1}^n\int_{V_i}\Big(f_0(X_i) - f_0(x)\Big)^2\,dx,\end{aligned}$$

and so by invoking the Lipschitz property of f_0, we obtain

$$\|\widehat{f} - f\|_{L^2(\mathcal{X})}^2 \leq 2\sum_{i=1}^n \mathrm{vol}(V_i)\Big(\widehat{f}(X_i) - f_0(X_i)\Big)^2 + 2M^2\sum_{i=1}^n \Big(\mathrm{diam}(V_i)\Big)^2. \tag{B.52}$$

Here we have written $\mathrm{diam}(V)$ for the diameter of a set V.

Now we will use some results of Chaudhuri and Dasgupta [2010] regarding uniform concentration of empirical counts, to upper bound $\mathrm{diam}(V_i)$ Set

$$\varepsilon_n := \left(\frac{2C_o \log(1/\delta) d \log n}{\nu_d p_{\min} a_3 n}\right)^{1/d},$$

where C_o is a constant given in Lemma 16 of Chaudhuri and Dasgupta [2010]. Note that for n sufficiently large, $\varepsilon_n \leq c_0$, and therefore by (B.35) we have that for every $x \in \mathcal{X}$, $P(B(x, \varepsilon_n)) \geq 2C_o \log(1/\delta) d \frac{\log n}{n}$. Consequently, by Lemma 16 of Chaudhuri and Dasgupta [2010] it holds that with probability at least $1 - \delta$,

$$\text{for all } x \in \mathcal{X}, \quad B(x, \varepsilon_n) \cap \{X_1, \ldots, X_n\} \neq \emptyset. \tag{B.53}$$

But if (B.53) is true, it must also be true that for each $i = 1, \ldots, n$ and for every $x \in V_i$, the distance $\|x - X_i\| \leq \varepsilon_n$. Thus by the triangle inequality, $\max_{i=1,\ldots,n} \mathrm{diam}(V_i) \leq 2\varepsilon_n$. Plugging back in to (B.52), and using the upper bound volume $\mathrm{vol}(V_i) \leq \nu_d\big(\mathrm{diam}(V_i)\big)^d$, we obtain the desired upper bound on $\|\widehat{f} - f\|_{L^2(\mathcal{X})}^2$.

Proof of Theorem 9. The proof of Theorem 9 follows exactly the same steps as the proof of Theorem 6, replacing the references to Lemma 3 and 4 by references to Proposition 24 and (B.50), and the ambient dimension d by the intrinsic dimension m.

B.7.2 Proofs of testing results

Proof of Theorem 8. Let $\delta = 1/b$. Recall that we have shown that the inequalities (3.14) and (3.15) are satisfied with probability at least $1 - 1/b - C_1 n \exp(-c_1 n r^d)$, and throughout this proof we take as granted that both of these inequalities hold.

Now, we would like to invoke Lemma 27, and in order to do so, we must show that the inequality (B.14) is satisfied with respect to $G = G_{n,r}$. First, we upper bound the right hand side of this inequality. Setting $\widetilde{\rho} = M^{-8/(4+d)} n^{-4/(4+d)}$ as prescribed by Theorem 8, it follows from (3.14) and (3.15) that

$$\begin{aligned}
\frac{2\rho}{n}(f_0^\top \mathbf{L} f_0) + \frac{2\sqrt{2/\alpha} + 2b}{n}\left(\sum_{k=1}^n \frac{1}{(\rho\lambda_k + 1)^4}\right)^{1/2} &\leq C_2 b\widetilde{\rho} M^2 + \frac{2\sqrt{2/\alpha} + 2b}{n}\left[1 + \frac{1}{c_3^2}\left(\sum_{k=2}^n \frac{1}{(\widetilde{\rho} k^{2/d} + 1)^4}\right)^{1/2} + \frac{r^4 n^{1/2}}{c_3^2 \widetilde{\rho}^2}\right] \\
&\leq C_2 b\widetilde{\rho} M^2 + \frac{2\sqrt{2/\alpha} + 2b}{n}\left(1 + \frac{\sqrt{2}}{c_3^2}\widetilde{\rho}^{-d/4} + \frac{r^4 n^{1/2}}{c_3^2 \widetilde{\rho}^2}\right) \\
&\leq \left(C_2 + 2 + \frac{2\sqrt{2}}{c_3^2} + \frac{2}{c_3^2}\right) \cdot \left(\sqrt{\frac{2}{\alpha}} + b\right) \cdot M^{2d/(4+d)} n^{-4/(4+d)}.
\end{aligned}$$

The second inequality in the above is justified by Lemma 36, keeping in mind that $M \leq M_{\max}(d)$ implies that $\widetilde{\rho}^{-d/2} \leq n$. The third inequality follows from the upper bound on r assumed in (R2) as well as the fact that $M \geq n^{-1/2}$.

Next we lower bound the left hand side of the inequality (B.14)—i.e. we lower bound the empirical norm $\|f_0\|_n^2$—using Lemma 34. Recall that by assumption, $M \leq M_{\max}(d)$. Therefore, taking $C \geq C_6$ in (3.11) implies that the lower bound on $\|f\|_{L^2(\mathcal{X})}$ in (B.47) is satisfied. As a result, it follows from (B.48) that

$$\|f\|_n^2 \geq \frac{\mathbb{E}[\|f\|_n^2]}{b} \geq \frac{p_{\min}}{b}\|f\|_{L^2(\mathcal{X})}^2 \geq C\left(\sqrt{\frac{1}{\alpha}} + b\right) M^{2d/(4+d)} n^{-4/(4+d)},$$

with probability at least $1 - 5/b$. Taking $C \geq C_2 + 2 + (2\sqrt{2})/c_3^2 + 2/c_3^2$ in (3.11) thus implies (B.14), and we may therefore use Lemma 27 to upper bound the type II error the Laplacian smoothing test $\widehat{\varphi}$. Observe that by (3.15) and the lower bound in Lemma 36,

$$\sum_{k=1}^{n}\left(\frac{1}{\rho\lambda_k + 1}\right)^4 \geq 1 + \frac{1}{C_3^4}\sum_{k=2}^{n}\left(\frac{1}{\widetilde{\rho}k^{2/d} + 1}\right)^4 \geq \frac{1}{32C_3^4}\widetilde{\rho}^{-d/2}.$$

We conclude that

$$\begin{aligned}\mathbb{P}_{f_0}(\widehat{T} \leq \widehat{t}_\alpha) &\leq \frac{6}{b} + \frac{1}{b^2} + \frac{16}{b}\left(\sum_{k=1}^{n}\frac{1}{(\rho\lambda_k + 1)^4}\right)^{-1/2} + C_1 n\exp(-c_1 n r^d) \\ &\leq \frac{7}{b} + \frac{64\sqrt{2}}{b}C_3^2\widetilde{\rho}^{d/4} + C_1 n\exp(-c_1 n r^d),\end{aligned}$$

establishing the claim of Theorem 8.

Proof of Theorem 10. The proof of Theorem 10 follows exactly the same steps as the proof of Theorem 8, replacing the references to Lemma 3 and 4 by references to Propositions 24 and (B.50), and the ambient dimension d by the intrinsic dimension m.

Proof of (3.12). When $\rho = 0$, the Laplacian smoother $\widehat{f} = \mathbf{Y}$, the test statistic $\widehat{T} = \frac{1}{n}\|\mathbf{Y}\|_2^2$, and the threshold $\widehat{t}_\alpha = 1 + n^{-1/2}\sqrt{2/\alpha}$. The expectation of $\widehat{T}$ is

$$\mathbb{E}\big[\widehat{T}\big] = \mathbb{E}\big[f_0^2(X)\big] + 1 \geq p_{\min}\|f_0\|_{L^2(\mathcal{X})}^2 + 1.$$

When $f_0 \in L^4(\mathcal{X}, M)$, the variance can be upper bounded

$$\mathrm{Var}\big[\widehat{T}\big] \leq \frac{1}{n}\Big(3 + p_{\max}M^4 + p_{\max}\|f_0\|_{L^2(\mathcal{X})}^2\Big).$$

Now, let us assume that

$$\|f_0\|_{L^2(X)}^2 \geq \frac{2\sqrt{2/\alpha} + 2b}{p_{\min}}n^{-1/2},$$

so that $E[\widehat{T}] - \widehat{t}_\alpha \geq E[f_0^2(X)]/2$. Hence, by Chebyshev's inequality

$$\begin{aligned}\mathbb{P}_{f_0}\Big(\widehat{T} \leq \widehat{t}_\alpha\Big) &\leq 4\frac{\mathrm{Var}_{f_0}\big[\widehat{T}\big]}{\mathbb{E}[f_0^2(X)]^2} \\ &\leq \frac{4}{n}\cdot\frac{3 + p_{\max}\big(M^4 + \|f_0\|_{L^2(\mathcal{X})}^2\big)}{p_{\min}^2\|f_0\|_{L^2(\mathcal{X})}^4} \\ &\leq \frac{1}{b^2}\Big(3 + \frac{4bp_{\max}}{p_{\min}n^{1/2}} + p_{\max}M^4\Big).\end{aligned}$$

B.7.3 Two convenient estimates

The following Lemmas provides convenient upper and lower bounds on our estimation variance term (Lemma 35) and testing variance term (Lemma 36).

Lemma 35. *For any $t > 0$ such that $1 \leq t^{-d/2} \leq n$,*

$$\frac{1}{8}t^{-d/2} - 1 \leq \sum_{k=2}^{n}\left(\frac{1}{tk^{2/d} + 1}\right)^2 \leq t^{-d/2} + \begin{cases} 3t^{-d/2}, & \text{if } d < 4 \\ \frac{1}{t^2}\log n, & \text{if } d = 4 \\ \frac{1}{t^2}n^{1-4/d}, & \text{if } d > 4. \end{cases}$$

Lemma 36. *Suppose $d \le 4$. Then for any $t > 0$ such that $1 \le t^{-d/2} \le n$,*

$$\frac{1}{32} t^{-d/2} - 1 \le \sum_{k=2}^{n} \left(\frac{1}{tk^{2/d} + 1} \right)^4 \le 2t^{-d/2}.$$

Proof of Lemma 35. We begin by proving the upper bounds. Treating the sum over k as a Riemann sum of a non-increasing function, we have that

$$\sum_{k=2}^{n} \left(\frac{1}{tk^{2/d} + 1} \right)^2 \le \int_1^n \left(\frac{1}{tx^{2/d} + 1} \right)^2 dx \le t^{-d/2} + \int_{t^{-d/2}}^{n} \left(\frac{1}{tx^{2/d} + 1} \right)^2 dx \le t^{-d/2} + \frac{1}{t^2} \int_{t^{-d/2}}^{n} x^{-4/d} \, dx.$$

The various upper bounds (for $d < 4$, $d = 4$, and $d > 4$) then follow upon computing the integral.

For the lower bound, we simply recognize that for each $k = 2, \dots, n$ such that $k \le \lfloor t^{-d/2} \rfloor$, it holds that $1/(tk^{2/d} + 1)^2 \ge 1/4$, and there are at least $\min\left\{ \lfloor t^{-d/2} \rfloor - 1, n - 1 \right\} > \frac{1}{2} t^{-d/2} - 1$ such values of k.

Proof of Lemma 36. The upper bound follows similarly to that of Lemma 35:

$$\sum_{k=1}^{n} \left(\frac{1}{tk^{2/d} + 1} \right)^4 \le t^{-d/2} + \frac{1}{t^4} \sum_{k=t^{-d/2}+1}^{n} \frac{1}{k^{8/d}} \le t^{-d/2} + \frac{1}{t^4} \int_{t^{-d/2}}^{n} x^{-8/d} \, dx \le 2t^{-d/2}.$$

The lower bound follows from the same logic as we used to derive the lower bound in Lemma 35.

B.8 Concentration inequalities

Lemma 37. *Let $\xi_1, \dots, \xi_N$ be independent $N(0,1)$ random variables, and let $U := \sum_{k=1}^{N} a_k(\xi_k^2 - 1)$. Then for any $t > 0$,*

$$\mathbb{P}\Big[U \ge 2\|a\|_2 \sqrt{t} + 2\|a\|_\infty t \Big] \le \exp(-t).$$

In particular if $a_k = 1$ for each $k = 1, \dots, N$, then

$$\mathbb{P}\Big[U \ge 2\sqrt{Nt} + 2t \Big] \le \exp(-t).$$

The proof of Lemma 34 relies on (a variant of) the Paley-Zygmund Inequality.

Lemma 38. *Let f satisfy the following moment inequality for some $b \ge 1$:*

$$\mathbb{E}\big[\|f\|_n^4\big] \le \left(1 + \frac{1}{b^2}\right) \cdot \Big(\mathbb{E}\big[\|f\|_n^2\big]\Big)^2. \tag{B.54}$$

Then,

$$\mathbb{P}\left[\|f\|_n^2 \ge \frac{1}{b} \mathbb{E}\big[\|f\|_n^2\big] \right] \ge 1 - \frac{5}{b}. \tag{B.55}$$

Proof. Let Z be a non-negative random variable such that $\mathbb{E}(Z^q) < \infty$. The Paley-Zygmund inequality says that for all $0 \le \lambda \le 1$,

$$\mathbb{P}(Z > \lambda \mathbb{E}(Z^p)) \ge \left[(1 - \lambda^p) \frac{\mathbb{E}(Z^p)}{(\mathbb{E}(Z^q))^{p/q}} \right]^{\frac{q}{q-p}}. \tag{B.56}$$

Applying (B.56) with $Z = \|f\|_n^2$, $p = 1$, $q = 2$ and $\lambda = \frac{1}{b}$, by assumption (B.54) we have

$$\mathbb{P}\Big(\|f\|_n^2 > \frac{1}{b}\mathbb{E}[\|f\|_n^2]\Big) \geq \Big(1 - \frac{1}{b}\Big)^2 \cdot \frac{\big(\mathbb{E}[\|f\|_n^2]\big)^2}{\mathbb{E}[\|f\|_n^4]} \geq \frac{\Big(1 - \frac{2}{b}\Big)}{\Big(1 + \frac{1}{b^2}\Big)} \geq 1 - \frac{5}{b}.$$

□

Let $Z_1, \ldots, Z_n$ be independently distributed and bounded random variables, such that $\mathbb{E}[Z_i] = \mu_i$. Let $S_n = Z_1 + \ldots + Z_n$ and $\mu = \mu_1 + \ldots + \mu_n$. The multiplicative form of Hoeffding's inequality gives sharp bounds when $\mu \ll 1$.

Lemma 39 (Hoeffding's Inequality, multiplicative form)**.** *Suppose Z_i are independent random variables, which satisfy $Z_i \in [0, B]$ for $i = 1, \ldots, n$. For any $0 < \delta < 1$, it holds that*

$$\mathbb{P}\Big(\Big|S_n - \mu\Big| \geq \delta\mu\Big) \leq 2\exp\Big(-\frac{\delta^2\mu}{3B^2}\Big).$$

We use Lemma 39, along with properties of the kernel K and density p, to upper bound the maximum degree in our neighborhood graph, which we denote by $D_{\max}(G_{n,r}) := \max_{i=1,\ldots,n} D_{ii}$.

Lemma 40. *Under the conditions of Lemma 4,*

$$D_{\max}(G_{n,r}) \leq 2p_{\max}nr^d,$$

with probability at least $1 - 2n\exp\Big(-nr^d a_3 p_{\min}/(3[K(0)]^2)\Big)$.

Proof of Lemma 40. Fix $x \in \mathcal{X}$, and set

$$D_{n,r}(x) := \sum_{i=1}^n K\Big(\frac{\|X_i - x\|}{r}\Big);$$

note that $D_{n,r}(X_i)$ is just the degree of X_i in $G_{n,r}$. By Hoeffding's inequality

$$\mathbb{P}\Big(\Big|D_{n,r}(x) - \mathbb{E}\big[D_{n,r}(x)\big]\Big| \geq \delta\mathbb{E}\big[D_{n,r}(x)\big]\Big) \leq 2\exp\Big(-\frac{\delta^2\mathbb{E}\big[D_{n,r}(x)\big]}{3[K(0)]^2}\Big). \tag{B.57}$$

Now we lower bound $\mathbb{E}[D_{n,r}(x)]$ using the boundedness of the density p, and the fact that $\mathcal{X}$ has Lipschitz boundary:

$$\begin{aligned}
\mathbb{E}\big[D_{n,r}(x)\big] &= n\int_{\mathcal{X}} K\Big(\frac{\|x' - x\|}{r}\Big)p(x)\,dx \\
&\geq np_{\min}\int_{\mathcal{X}} K\Big(\frac{\|x' - x\|}{r}\Big)\,dx \\
&\geq np_{\min}a_3\int_{\mathcal{X}} K\Big(\frac{\|x' - x\|}{r}\Big)\,dx \\
&\geq nr^d p_{\min},
\end{aligned}$$

with the second inequality following from (B.35), and the final inequality from the normalization $\int_{\mathbb{R}^d} K(\|z\|)\,dz = 1$. Similar derivations yield the upper bound

$$\mathbb{E}\big[D_{n,r}(x)\big] \leq nr^d p_{\max},$$

and plugging these bounds in to (B.57), we determine that

$$\mathbb{P}\Big(D_{n,r}(x) \geq (1+\delta)nr^d p_{\max}\Big) \leq 2\exp\Big(-\frac{\delta^2 nr^d a_0 p_{\min}}{3[K(0)]^2}\Big).$$

Applying a union bound, we get that

$$\mathbb{P}\Big(\max_{i=1,\ldots,n} D_{n,r}(X_i) \geq (1+\delta)nr^d p_{\max}\Big) \leq 2n\exp\Big(-\frac{\delta^2 nr^d a_0 p_{\min}}{3[K(0)]^2}\Big),$$

and taking $\delta = 1$ gives the claimed upper bound.

Appendix C

Chapter 4 Appendix

C.1 Graph-dependent error bounds

In this section, we adopt the fixed design perspective; or equivalently, condition on $X_i = x_i$ for $i = 1, \dots, n$. Let $G = \big([n], W\big)$ be a fixed graph on $\{1, \dots, n\}$ with Laplacian matrix $L = \sum_{k=1}^n \lambda_k v_k v_k^\top$; the eigenvectors have unit empirical norm, $\|v_k\|_n^2 = 1$. The randomness thus all comes from the responses

$$Y_i = f_0(x_i) + w_i \tag{C.1}$$

where the noise variables w_i are independent $N(0, 1)$. In the rest of this section, we will mildly abuse notation and write $f_0 = (f_0(x_1), \dots, f_0(x_n)) \in \mathbb{R}^n$. We will also write $\mathbf{Y} = (Y_1, \dots, Y_n)$.

C.1.1 Upper bound on Estimation Error of Laplacian Eigenmaps

Lemma 41. *For any integer $s > 0$, and any integer $0 \le K \le n$, the Laplacian eigenmaps estimator $\widehat{f}$ of (4.5) satisfies*

$$\|\widehat{f} - f_0\|_n^2 \le \frac{\langle L^s f_0, f_0 \rangle_n}{\lambda_{K+1}^s} + \frac{5K}{n}; \tag{C.2}$$

this is guaranteed if $K = 0$, and otherwise holds with probability at least $1 - \exp(-K)$ if $1 \le K \le n$.

Proof (of Lemma 41). By the triangle inequality,

$$\|\widehat{f} - f_0\|_n^2 \le 2\Big(\|\mathbb{E}\widehat{f} - f_0\|_n^2 + \|\widehat{f} - \mathbb{E}\widehat{f}\|_n^2\Big). \tag{C.3}$$

The first term in (C.3) (approximation error) is non-random, since the design is fixed. The expectation $\mathbb{E}\widehat{f} = \sum_{k=1}^K \langle v_k, f_0 \rangle_n v_k$, so that

$$\|\mathbb{E}\widehat{f} - f_0\|_n^2 = \Big\| \sum_{k=K+1}^n \langle v_k, f_0 \rangle_n v_k \Big\|_n^2 = \sum_{k=K+1}^n \langle v_k, f_0 \rangle_n^2.$$

In the above, the last equality relies on the fact that v_k are orthonormal in $L^2(P_n)$. Using the fact that the eigenvalues are in increasing order, we obtain

$$\sum_{k=K+1}^n \langle v_k, f_0 \rangle_n^2 \le \frac{1}{\lambda_{K+1}^s} \sum_{k=K+1}^n \lambda_k^s \langle v_k, f_0 \rangle_n^2 \le \frac{\langle L^s f_0, f_0 \rangle_n}{\lambda_{K+1}^s}.$$

If $K = 0$, $\widehat{f} = \mathbb{E}\widehat{f} = 0$, and the second term in (C.3) is 0. Otherwise the second in (C.3) (estimation error) is random. Observe that $\langle v_k, \varepsilon\rangle_n \stackrel{d}{=} Z_k/\sqrt{n}$, where $(Z_1, \dots, Z_n) \sim N(0, I_{n\times n})$. Again using the orthonormality of the eigenvectors v_k, we have

$$\|\widehat{f} - \mathbb{E}\widehat{f}\|_n^2 = \sum_{k=1}^{K} \langle v_k, \varepsilon\rangle_n^2 \stackrel{d}{=} \frac{1}{n}\sum_{k=1}^{K} Z_k^2.$$

Thus $\|\widehat{f} - \mathbb{E}\widehat{f}\|_n^2$ is equal to $1/n$ times a χ^2 distribution with K degrees of freedom. Consequently, it follows from a result of [Laurent and Massart, 2000] that

$$\mathbb{P}\left(\|\widehat{f} - \mathbb{E}\widehat{f}\|_n^2 \geq \frac{K}{n} + 2\frac{\sqrt{K}}{n}\sqrt{t} + \frac{2t}{n}\right) \leq \exp(-t).$$

Setting $t = K$ completes the proof of the lemma.

C.1.2 Upper bound on Testing Error of Laplacian Eigenmaps

Let $\widehat{T} = \sum_{k=1}^{K} \langle \mathbf{Y}, v_k\rangle_n^2$, and let $\varphi = \mathbf{1}\{\widehat{T} \geq t_a\}$. In the following Lemma, we upper bound the Type I and Type II error of the test φ.

Lemma 42. *Suppose we observe $(Y_1, x_1), \dots, (Y_n, x_n)$ according to (C.1).*

- *If $f_0 = 0$, then $\mathbb{E}_0[\varphi] \leq a$.*
- *Suppose $f_0 \neq 0$ satisfies*

$$\|f_0\|_n^2 \geq \frac{\langle L^s f_0, f_0\rangle_n}{\lambda_{K+1}^s} + \frac{\sqrt{2K}}{n}\left[2\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right], \tag{C.4}$$

for some $s \in \mathbb{N}\setminus\{0\}$. Then $\mathbb{E}_{f_0}[1-\phi] \leq b$.

Proof (of Lemma 42). We first compute the expectation and variance of $\widehat{T}$, then apply Chebyshev's inequality to upper bound the Type I and Type II error.

Expectation. Recall that $\widehat{T} = \sum_{k=1}^{K}\langle Y, v_k\rangle_n^2$. Expanding the square gives

$$\mathbb{E}[\widehat{T}] = \sum_{k=1}^{K}\mathbb{E}[\langle Y, v_k\rangle_n^2] = \sum_{k=1}^{K}\langle f_0, v_k\rangle_n^2 + \mathbb{E}[2\langle f_0, v_k\rangle_n\langle \varepsilon, v_k\rangle_n + \langle \varepsilon, v_k\rangle_n^2] = \frac{K}{n} + \sum_{k=1}^{K}\langle f_0, v_k\rangle_n^2.$$

Thus $\mathbb{E}[\widehat{T}] - t_a = \sum_{k=1}^{K}\langle f_0, v_k\rangle_n^2 - \sqrt{2K}/n \cdot \sqrt{1/a}$. Furthermore, it is a consequence of (C.4) that

$$\sum_{k=1}^{K}\langle f_0, v_k\rangle_n^2 - \frac{\sqrt{2K}}{n}\sqrt{1/a} \geq \|f_0\|_n^2 - \frac{\langle L^s f_0, f_0\rangle_n}{\lambda_{K+1}^s} - \frac{\sqrt{2K}}{n}\sqrt{1/a} \geq \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right]. \tag{C.5}$$

Variance. Recall from the proof of Lemma 41 that $\langle \varepsilon, v_k\rangle_n \stackrel{d}{=} Z_k/\sqrt{n}$ for $(Z_1, \dots, Z_n) \sim N(0, I_{n\times n})$. Expanding the square, and recalling that $\mathrm{Cov}[Z, Z^2] = 0$ for Gaussian random variables, we have that

$$\mathrm{Var}\left[\langle \mathbf{Y}, v_k\rangle_n^2\right] = \mathrm{Var}\left[\frac{2}{n}\langle f_0, v_k\rangle_n Z_k + \frac{2}{n^2}Z_k^2\right] = \frac{4\langle f_0, v_k\rangle_n^2}{n} + \frac{2}{n^2}.$$

Moreover, since $\mathrm{Cov}[Z_k^2, Z_\ell^2] = 0$ for each $k = 1, \dots, K$, we see that

$$\mathrm{Var}\left[\widehat{T}\right] = \sum_{k=1}^{K}\mathrm{Var}\left[\langle \mathbf{Y}, v_k\rangle_n^2\right] = \frac{2K}{n^2} + \sum_{k=1}^{K}\frac{4\langle f_0, v_k\rangle_n^2}{n}.$$

Bounds on Type I and Type II error. The upper bound on Type I error follows immediately from Chebyshev's inequality.

The upper bound on Type II error also follows from Chebyshev's inequality. We observe that (C.4) implies $\mathbb{E}_{f_0}[\widehat{T}] = t_a$, and apply Chebyshev's inequality to deduce

$$\mathbb{P}_{f_0}\big(\widehat{T} < t_a\big) \le \mathbb{P}_{f_0}\Big(|\widehat{T} - \mathbb{E}_{f_0}[\widehat{T}]|^2 > |\mathbb{E}_{f_0}[\widehat{T}] - t_a|^2\Big) \le \frac{\mathrm{Var}\big[\widehat{T}\big]}{\big[\mathbb{E}_{f_0}[\widehat{T}] - t_a\big]^2} = \frac{2K/n^2 + 4/n\sum_{k=1}^K \langle f_0, v_k\rangle_n^2}{\big[\mathbb{E}_{f_0}[\widehat{T}] - t_a\big]^2}.$$

Thus we have upper bounded the Type II error by the sum of two terms, each of which are no more than $1/(2b)$, as we now show. For the first term, after noting that (C.5) implies $\mathbb{E}_{f_0}[\widehat{T}] - t_a \ge \sqrt{2K}/n \cdot \sqrt{2/b}$, the upper bound follows:

$$\frac{2K/n^2}{\big[\mathbb{E}_{f_0}[\widehat{T}] - t_a\big]^2} \le \frac{b}{2}.$$

On the other hand, for the second term we use (C.5) in two ways: first to conclude that $\mathbb{E}_{f_0}[\widehat{T}] - t_a \ge 1/2 \cdot \sum_{k=1}^K \langle f_0, v_k\rangle_n^2$, and second to obtain

$$\frac{4\sum_{k=1}^K \langle f_0, v_k\rangle_n^2}{n\big[\mathbb{E}_{f_0}[\widehat{T}] - t_a\big]^2} \le \frac{4\sum_{k=1}^K \langle f_0, v_k\rangle_n^2}{n\big(\sum_{k=1}^K \langle f_0, v_k\rangle_n^2/2\big)^2} \le \frac{16}{n\sum_{k=1}^K \langle f_0, v_k\rangle_n^2} \le \frac{b}{2}.$$

C.2 Analysis of Spectral Series Estimator

In this section we prove Proposition 6. As mentioned in our main text, this proof relies on three facts: (i) a continuous embedding of $H_0^s(\mathcal{X})$ into $\mathcal{H}^s(\mathcal{X})$, (ii) the Weyl's Law asymptotic scaling $\lambda_k(\Delta_P) \asymp k^{2/d}$, and (iii) a local Weyl's Law scaling $\sum_{k=1}^K \big(\psi_k(x)\big)^2 \lesssim K$. (Recall that $(\lambda_k(\Delta_P), \psi_k)$ are defined to be solutions to (4.7)). We record the precise results we need in the following three lemmas.

The first result follows immediately from Lemma 17 of Dunlop et al. [2020]. Recall the spectral Sobolev class $\mathcal{H}^s(\mathcal{X})$ defined in (4.8). Equip $\mathcal{H}^s(\mathcal{X})$ with the norm

$$\|f\|_{\mathcal{H}^s(\mathcal{X})}^2 := \sum_{k=1}^\infty \big[\langle f, \psi_k\rangle_P\big]^2 \cdot [\lambda_k(\Delta_P)]^2.$$

Lemma 43 (cf. Lemma 17 of Dunlop et al. [2020])**.** *Suppose Model 4.2.1, and additionally that $p \in C^\infty(\mathcal{X})$ and $\partial\mathcal{X} \in C^{1,1}$. Then for any $s \in \mathbb{N}$, we have $H_0^s(\mathcal{X}) \subseteq \mathcal{H}^s(\mathcal{X})$. Moreover, there exists a constant C such that for any $f \in H_0^s(\mathcal{X})$ we have*

$$\|f\|_{\mathcal{H}^s(\mathcal{X})} \le C\|f\|_{H^s(\mathcal{X})}. \tag{C.6}$$

The particular version of Weyl's Law we use is also due to Dunlop et al. [2020].

Lemma 44 (cf. Lemma 28 of Dunlop et al. [2020])**.** *Suppose Model 4.2.1. Then there exist positive constants c and C such that for all $k \ge 2$,*

$$ck^{2/d} \le \lambda_k(\Delta_P) \le Ck^{2/d}. \tag{C.7}$$

As we discuss in our main text, the asymptotic scaling $\lambda_k(\Delta_P) \asymp k^{2/d}$ also plays a key role in our analysis of Laplacian eigenmaps. We point out that although Lemma 28 of Dunlop et al. [2020] assumes $p \in C^\infty(\mathcal{X})$ and $\partial\mathcal{X} \in C^{1,1}$, the proof uses only the hypotheses of Model 4.2.1: that p is bounded away from 0 and ∞, and that $\partial\mathcal{X}$ is Lipschitz.

Finally, we need a local version of Weyl's Law, due to Hörmander [2007].

Lemma 45 (cf. Theorem 17.5.3 of Hörmander [2007]). *Suppose Model 4.2.1, and additionally that $p \in C^\infty(\mathcal{X})$ and $\partial\mathcal{X} \in C^{1,1}$. Then there exists a positive constant C such that for all $K \in \mathbb{N} \setminus 0$,*

$$\sup_{x \in \mathcal{X}} \left\{ \sum_{k=1}^{K} \big(\psi_k(x)\big)^2 \right\} \leq CK. \tag{C.8}$$

Translated into our notation, Theorem 17.5.3 of Hörmander [2007] says that

$$\sup_{x \in \mathcal{X}} \left\{ \sum_{k : \lambda_k(\Delta_P) \leq \lambda_K(\Delta_P)} \big(\psi_k(x)\big)^2 \right\} \leq C\big[\lambda_K(\Delta_P)\big]^{d/2},$$

and from here (C.8) follows from (C.7). Hörmander [2007] proves the result only when $\partial\mathcal{X} \in C^\infty$, but the proof goes through unchanged when $\partial\mathcal{X} \in C^1$, since $\partial\mathcal{X} \in C^1$ is sufficient to apply the relevant Sobolev embedding theorem (Lemma 17.5.2 of Hörmander [2007]) that is the key to proving Lemma 45.

We note that the assumptions $p \in C^\infty(\mathcal{X})$ and $\partial\mathcal{X} \in C^{1,1}$ used in Lemmas 43 and 45 can likely be removed. This would allow us to remove the same assumptions in the statement of Proposition 6. Since this does not change the main point of the proposition, we do not pursue the details further.

Proof (of Proposition 6). We decompose risk into squared bias and variance,

$$\mathbb{E}\|\widetilde{f} - f_0\|_P^2 = \mathbb{E}\|\mathbb{E}[\widetilde{f}] - f_0\|_P^2 + \mathbb{E}\|\widetilde{f} - \mathbb{E}[\widetilde{f}]\|_P^2. \tag{C.9}$$

Then some standard arguments (which we give below) give the following upper bound:

$$\mathbb{E}\|\widetilde{f} - f_0\|_P^2 \leq \frac{\|f_0\|_{\mathcal{H}^s(\mathcal{X})}^2}{\big[\lambda_{K+1}(\Delta_P)\big]^s} + \frac{K}{n} + \frac{1}{n}\mathbb{E}\Big[(f_0(X))^2 \cdot \sum_{k=1}^{K} (\psi_k(X))^2\Big]. \tag{C.10}$$

The claim of the proposition follows from (C.10) and Lemmas 43-45, upon taking $K = \big\lfloor (M^2 n)^{d/(2s+d)} \big\rfloor$. (Here we use the convention $\sum_{k=1}^{0} c_k = 0$ for any sequence (c_k).) □

The upper bound (C.10) is to be compared with (4.24), which gives upper bounds on the (design-dependent) squared bias and variance of Laplacian eigenmaps. We see (4.24) has terms analogous to the first two terms on the right hand side of (C.10), but not the third. The relevance of this third term to L^2 estimation problems is discussed by Birgé [2008].

Proof (of (C.10)). Since $\{\psi_k\}$ are an orthonormal basis of $L^2(P)$, and $f_0 \in H_0^s(\mathcal{X}) \subseteq L^2(P)$, we have $f_0 = \sum_{k=1}^{\infty} \langle f_0, \psi_k \rangle_P \psi_k$ in $L^2(P)$. This allows us to expand the squared bias in terms of squared Fourier coefficients of f_0, and gives the following upper bound,

$$\|f_0 - \mathbb{E}\widetilde{f}\|_P^2 = \sum_{k=K+1}^{\infty} \langle f_0, \psi_k \rangle^2 \leq \frac{1}{\lambda_{K+1}(\Delta_P)^s} \sum_{k=K+1}^{\infty} \lambda_{k+1}(\Delta_P)^s \langle f_0, \psi_k \rangle^2 \leq \frac{\|f_0\|_{\mathcal{H}^s}}{\big[\lambda_{K+1}(\Delta_P)\big]^s}.$$

On the other hand, the variance term can be written as the sum of the variance of each empirical Fourier coefficient,

$$\mathbb{E}\|\widetilde{f} - \mathbb{E}[\widetilde{f}]\|_P^2 = \sum_{k=1}^{K} \mathrm{Var}\Big[\langle \mathbf{Y}, \psi_k \rangle_n\Big] \tag{C.11}$$

Then from the law of total variance,

$$\mathrm{Var}\big[\langle \mathbf{Y}, \psi_k \rangle_n\big] = \mathrm{Var}\Big[\mathbb{E}[\langle Y, \psi_k \rangle_n | \mathbf{X}\Big] + \mathbb{E}\Big[\mathrm{Var}[\langle Y, \psi_k \rangle_n | \mathbf{X}\Big] = \mathrm{Var}\Big[\langle f_0, \psi_k \rangle_n\Big] + \frac{1}{n}\mathbb{E}\Big[\|\psi_k\|_n^2\Big].$$

Finally, $\mathbb{E}\|\psi_k\|_n^2 = \|\psi_k\|_P^2 = 1$,

$$\mathrm{Var}\Big[\langle f_0, \psi_k\rangle_n\Big] = \frac{1}{n}\mathrm{Var}\Big[f_0(X)\psi_k(X)\Big] \leq \frac{1}{n}\mathbb{E}\Big[\Big(f_0(X)\psi_k(X)\Big)^2\Big],$$

and plugging back into (C.11) yields the claim.

C.3 Graph Sobolev semi-norm, flat Euclidean domain

In this section we prove Proposition 8. The proposition will follow from several intermediate results.

1. In Section C.3.1, we show that

$$\langle L_{n,\varepsilon}^s f, f\rangle_n \leq \frac{1}{\delta}\langle L_{P,\varepsilon}^s f, f\rangle_P + \frac{C\varepsilon^2}{\delta n \varepsilon^{2+d}} M^2. \tag{C.12}$$

   with probability at least $1 - 2\delta$.

   We term the first term on the right hand side the *non-local Sobolev semi-norm*, as it is a kernelized approximation to the Sobolev semi-norm $\langle \Delta_P^s f, f\rangle_P$. The second term on the right hand side is a pure bias term, which as we will see is negligible compared to the non-local Sobolev semi-norm as long as $\varepsilon \ll n^{-1/(2(s-1+d))}$.

2. In Section C.3.2, we show that when x is sufficiently in the interior of $\mathcal{X}$, then $L_{P,\varepsilon}^k f(x)$ is a good approximation to $\Delta_P^k f(x)$, as long as $f \in H^s(\mathcal{X})$ and $p \in C^{s-1}(\mathcal{X})$ for some $s \geq 2k+1$.

3. In Section C.3.3, we show that when x is sufficiently near the boundary of $\mathcal{X}$, then $L_{P,\varepsilon}^k f(x)$ is close to 0, as long as $f \in H_0^s(\mathcal{X})$ for some $s > 2k$.

4. In Section C.3.4, we use the results of the preceding two sections to show that if $f \in H_0^s(\mathcal{X}; M)$ and $p \in C^{s-1}(\mathcal{X})$, there exists a constant C which does not depend on f such that

$$\langle L_{P,\varepsilon}^s f, f\rangle_P \leq CM^2. \tag{C.13}$$

Finally, in Section C.3.5 we provide some assorted estimates used in Sections C.3.1.

Proof (of Proposition 8). Proposition 8 follows immediately from (C.12) and (C.13). □

One note regarding notation: suppose a function $g \in H^\ell(U)$, where $\ell \in \mathbb{N}$ and U is an open set. Let V be another open set, compactly contained within U. Then we will use the notation $g \in H^\ell(V)$ to mean that the restriction $g|_V$ of g to V belongs to $H^\ell(V)$.

C.3.1 Decomposition of graph Sobolev semi-norm

In Lemma 46, we decompose the graph Sobolev semi-norm (a V-statistic) into an unbiased estimate of the non-local Sobolev semi-norm (a U-statistic), and a pure bias term. We establish that the pure bias term will be small (in expectation) relative to the U-statistic whenever ε is sufficiently small.

Lemma 46. *For any $f \in L^2(\mathcal{X})$, the graph Sobolev semi-norm satisfies*

$$\langle L_{n,\varepsilon}^s f, f\rangle_n = U_{n,\varepsilon}^{(s)}(f) + B_{n,\varepsilon}^{(s)}(f), \tag{C.14}$$

such that $\mathbb{E}[U_{n,\varepsilon}^{(s)}(f)] = (n-s-1)!/n! \cdot \langle L_{P,\varepsilon}^s f, f\rangle_P$. If additionally $f \in H^1(\mathcal{X}; M)$ and $\varepsilon \geq n^{-1/d}$, then the bias term $B_{n,\varepsilon}^{(s)}(f)$ satisfies

$$\mathbb{E}\big[|B_{n,\varepsilon}^{(s)}(f)|\big] \leq \frac{C\varepsilon^2}{\delta n \varepsilon^{2+d}} M^2. \tag{C.15}$$

Then C.12 follows immediately from Lemma 46, by Markov's inequality.

Proof (of Lemma 46). We begin by introducing some notation. We will use bold notation $\mathbf{j} = (j_1, \dots, j_s)$ for a vector of indices where $j_i \in [n]$ for each i. We write $[n]^s$ for the collection of all such vectors, and $(n)^s$ for the subset of such vectors with no repeated indices. Finally, we write $D_i f$ for a kernelized difference operator,

$$D_i f(x) := \big(f(x) - f(X_i)\big)\eta\bigg(\frac{\|X_i - x\|}{\varepsilon}\bigg),$$

and we let $D_{\mathbf{j}} f(x) := \big(D_{j_1} \circ \cdots \circ D_{j_s} f\big)(x)$.

With this notation in hand, it is easy to represent $\langle L^s_{n,\varepsilon} f, f\rangle_n$ as the sum of a U-statistic and a bias term,

$$\begin{aligned}\langle L^s_{n,\varepsilon} f, f\rangle_n &= \frac{1}{n}\sum_{i=1}^n L^s_{n,\varepsilon} f(X_i)\cdot f(X_i) \\ &= \underbrace{\frac{1}{n^{s+1}\varepsilon^{s(d+2)}} \sum_{i\mathbf{j}\in(n)^{s+1}} D_{\mathbf{j}} f(X_i)\cdot f(X_i)}_{=:U^{(s)}_{n,\varepsilon}(f)} + \underbrace{\frac{1}{n^{s+1}\varepsilon^{s(d+2)}} \sum_{\substack{i\mathbf{j}\in \\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}} f(X_i)\cdot f(X_i)}_{=:B^{(s)}_{n,\varepsilon}(f)}\end{aligned}$$

When the indices of $i\mathbf{j}$ are all distinct, it follows straightforwardly from the law of iterated expectation that

$$\mathbb{E}[D_{\mathbf{j}} f(X_i)\cdot f(X_i)] = \varepsilon^{s(d+2)}\mathbb{E}[L^s_{P,\varepsilon} f(X_i)\cdot f(X_i)] = \langle L^s_{P,\varepsilon} f, f\rangle_P,$$

which in turn implies $\mathbb{E}[U^{(s)}_{n,\varepsilon}(f)] = (n-s-1)!/n!\cdot\langle L^s_{P,\varepsilon} f, f\rangle_P$.

It remains to show (C.15). By adding and subtracting $f(X_{\mathbf{j}_1})$, we obtain by symmetry that

$$\sum_{\substack{i\mathbf{j}\in \\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}} f(X_i)\cdot f(X_i) = \frac{1}{2}\cdot \sum_{\substack{i\mathbf{j}\in \\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}} f(X_i)\cdot\big(f(X_i) - f(X_{\mathbf{j}_1})\big),$$

and consequently

$$\mathbb{E}\Big[\sum_{\substack{i\mathbf{j}\in \\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}} f(X_i)\cdot f(X_i)\Big] \le \frac{1}{2}\cdot \sum_{\substack{i\mathbf{j}\in \\ [n]^{s+1}\setminus(n)^{s+1}}} \mathbb{E}\Big[\big|D_{\mathbf{j}} f(X_i)\big|\cdot\big|f(X_i) - f(X_{\mathbf{j}_1})\big|\Big].$$

In Lemma 51, we show that if $f \in H^1(\mathcal{X}; M)$, then for any $i\mathbf{j} \in [n]^{s+1}$ which contains a total of $k+1$ distinct indices,

$$\mathbb{E}\Big[\big|D_{\mathbf{j}} f(X_i)\big|\cdot\big|f(X_i) - f(X_{\mathbf{j}_1})\big|\Big] \le C_1\varepsilon^{2+kd}M^2.$$

This shows us that the expectation of $|B^s_{n,\varepsilon}(f)|$ can bounded from above by the sum over several different terms, as follows:

$$\begin{aligned}\mathbb{E}\Big[|B^s_{n,\varepsilon}(f)|\Big] &\le C_1\frac{\varepsilon^2}{n\varepsilon^{2s}}M^2 \sum_{\substack{i\mathbf{j}\in \\ [n]^{s+1}\setminus(n)^{s+1}}} \frac{1}{(n\varepsilon^d)^s}\varepsilon^{(|i\mathbf{j}|-1)d} \\ &\le C_1\frac{\varepsilon^2}{n\varepsilon^{2s}}M^2\sum_{k=1}^{s-1}\frac{(n\varepsilon^d)^k}{(n\varepsilon^d)^s}n.\end{aligned}$$

Finally, we note that by assumption $n\varepsilon^d \ge 1$, so that in the above sum the factor of $(n\varepsilon^d)^k$ is largest when $k = s-1$. We conclude that

$$\mathbb{E}\Big[|B^s_{n,\varepsilon}(f)|\Big] \le C_1(s-1)\frac{\varepsilon^2}{n\varepsilon^{2s+d}}M^2,$$

which is the desired result.

C.3.2 Approximation error of non-local Laplacian

In this section, we establish the convergence $L^k_{P,\varepsilon} f \to \sigma_\eta^k \Delta_P^k f$ as $\varepsilon \to 0$. More precisely, we give an upper bound on the squared difference between $L^k_{P,\varepsilon} f$ and $\sigma_\eta^k \Delta_P^k f$ as a function of ε. The bound holds for all $x \in \mathcal{X}_{k\varepsilon}$, and $f \in H^s(\mathcal{X})$, as long as $s \geq 2k+1$.

Lemma 47. *Assume Model 4.2.1. Let $s \in \mathbb{N} \setminus \{0,1\}$, suppose that $f \in H^s(\mathcal{X}; M)$, and if $s > 1$ suppose that $p \in C^{s-1}(\mathcal{X})$. Let $L_{P,\varepsilon}$ be define with respect to a kernel η that satisfies (K1). Then there exist constants C_1 and C_2 that do not depend on f, such that each of the following statements hold.*

- *If s is odd and $k = (s-1)/2$, then*

$$\|L^k_{P,\varepsilon} f - \Delta_P^k f\|_{L^2(\mathcal{X}_{k\varepsilon})} \leq C_1 M \varepsilon \tag{C.16}$$

- *If s is even and $k = (s-2)/2$, then*

$$\|L^k_{P,\varepsilon} f - \Delta_P^k f\|_{L^2(\mathcal{X}_{k\varepsilon})} \leq C_2 M \varepsilon^2. \tag{C.17}$$

We remark that when $k = 1$ and $f \in C^3(\mathcal{X})$ or $C^4(\mathcal{X})$, statements of this kind are well known, and indeed stronger results—with $L^\infty(\mathcal{X})$ norm replacing $L^2(\mathcal{X})$ norm—hold. When dealing with the iterated Laplacian, and functions f which are regular only in the Sobolev sense, the proof is somewhat more lengthy, but the spirit of the result is largely the same.

Proof (of Lemma 47). Throughout this proof, we shall assume that f and p are smooth functions, meaning they belong to $C^\infty(\mathcal{X})$. This is without loss of generality, since $C^\infty(\mathcal{X})$ is dense in both $H^s(\mathcal{X})$ and $C^{s-1}(\mathcal{X})$, and since both sides of the inequalities (C.16) and (C.17) are continuous with respect to $\|\cdot\|_{H^s(\mathcal{X})}$ and $\|\cdot\|_{C^{s-1}(\mathcal{X})}$ norms.

We will actually prove a more general set of statements than contained in Lemma 47, more general in the sense that they give estimates for all k, rather than simply the particular choices of k given above. In particular, we will prove that the following two statements hold for any $s \in \mathbb{N}$ and any $k \in \mathbb{N} \setminus \{0\}$.

- If $k \geq s/2$, then for every $x \in \mathcal{X}_{k\varepsilon}$,

$$L^k_{P,\varepsilon} f(x) = g_s(x) \varepsilon^{s-2k} \tag{C.18}$$

  for a function g_s that satisfies

$$\|g_s\|_{L^2(\mathcal{X}_{k\varepsilon})} \leq C \|p\|^k_{C^q(\mathcal{X})} M \tag{C.19}$$

  where $q = 1$ if $s = 0$ or $s = 1$, and otherwise $q = s-1$.

- If $k < s/2$, then for every $x \in \mathcal{X}_{k\varepsilon}$,

$$L^k_{P,\varepsilon} f(x) = \sigma_\eta^k \cdot \Delta_P^k f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - k} g_{2(j+k)}(x) \varepsilon^{2j} + g_s(x) \varepsilon^{s-2k}. \tag{C.20}$$

  for functions g_j that satisfy

$$\|g_j\|_{H^{s-j}(\mathcal{X}_{k\varepsilon})} \leq C \|p\|^k_{C^{s-1}(\mathcal{X})} M. \tag{C.21}$$

In the statement above, recall that $H^0(\mathcal{X}_{k\varepsilon}) = L^2(\mathcal{X}_{k\varepsilon})$. Additionally, note that we may speak of the pointwise behavior of derivatives of f because we have assumed that f is a smooth function. Observe that (C.16) follows upon taking $k = \lfloor (s-1)/2 \rfloor$ in (C.20), whence we have

$$\big(L^k_{P,\varepsilon} f(x) - \sigma_\eta^k \Delta_P^k f(x)\big)^2 = \varepsilon^2 \big(g_s(x)\big)^2$$

for some $g_s \in L^2(\mathcal{X}_{k\varepsilon}, C \cdot M \cdot \|p\|_{C^{s-1}(\mathcal{X})})$, and integrating over $\mathcal{X}_{k\varepsilon}$ gives the desired result. (C.17) follows from (C.20) in an identical fashion.

It thus remains establish (C.20), and (C.18) which is an important part of proving (C.20). We will do so by induction on k. Note that throughout, we will let g_j refer to functions which may change from line to line, but which always satisfy (C.21).

Proof of (C.18) and (C.20), base case.

We begin with the base case, where $k = 1$. Again, we point out that although desired result is known when $s = 3$ or $s = 4$, and f is regular in the Hölder sense, we require estimates for all $s \in \mathbb{N}$ when f is regular in the Sobolev sense.

When $s = 0$, the inequality (C.18) is implied by Lemma 49. When $s \geq 1$, we proceed using Taylor expansion. For any $x \in \mathcal{X}_\varepsilon$, we have that $B(x, \varepsilon) \subseteq \mathcal{X}$. Thus for any $x' \in B(x, \varepsilon)$, we may take an order s Taylor expansion of f around $x' = x$, and an order q Taylor expansion of p around $x' = x$, where $q = 1$ if $s = 1$, and otherwise $q = s - 1$. (See Section C.8.2 for a review of the notation we use for Taylor expansions, as well as some properties that we make use of shortly.) This allows us to express $L_{P,\varepsilon} f(x)$ as the sum of three terms,

$$\begin{aligned} L_{P,\varepsilon} f(x) = \frac{1}{\varepsilon^{d+2}} \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{q-1} \frac{1}{j_1! j_2!} \int_{\mathcal{X}} \big(d_x^{j_1} f\big)(x' - x)\big(d_x^{j_2} p\big)(x' - x)\eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' \quad + \\ \frac{1}{\varepsilon^{d+2}} \sum_{j=1}^{s-1} \frac{1}{j!} \int_{\mathcal{X}} \big(d_x^{j} f\big)(x' - x) r_{x'}^q(x; p)\eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' \quad + \\ \frac{1}{\varepsilon^{d+2}} \int_{\mathcal{X}} r_{x'}^j(x; f)\eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x'). \end{aligned}$$

Here we have adopted the convention that $\sum_{j=1}^{0} = 0$.

Changing variables to $z = (x' - x)/\varepsilon$, we can rewrite the above expression as

$$\begin{aligned} L_{P,\varepsilon} f(x) = \frac{1}{\varepsilon^2} \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{q-1} \frac{\varepsilon^{j_1+j_2}}{j_1! j_2!} \int d_x^{j_1} f(z) d_x^{j_2} p(z) \eta\big(\|z\|\big)\, dz \quad + \\ \frac{1}{\varepsilon^2} \sum_{j=1}^{s-1} \frac{\varepsilon^j}{j!} \int d_x^j f(z) r_{zh+x}^q(x; p)\eta\big(\|z\|\big)\, dz \quad + \\ \frac{1}{\varepsilon^2} \int r_{zh+x}^j(x; f)\eta\big(\|z\|\big) p(zh + x)\, dz \\ := G_1(x) + G_2(x) + G_3(x). \end{aligned}$$

We now separately consider each of $G_1(x), G_2(x)$ and $G_3(x)$. We will establish that if $s = 1$ or $s = 2$, then $G_1(x) = 0$, and otherwise if $s \geq 3$ that

$$G_1(x) = \sigma_\eta \Delta_P f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - 1} g_{2(j+1)}(x)\varepsilon^{2j} + g_s(x)\varepsilon^{s-2}.$$

On the other hand, we will establish that if $s = 1$ then $G_2(x) = 0$, and otherwise for $s \geq 2$

$$\|G_2\|_{L^2(\mathcal{X}_\varepsilon)} \leq C\varepsilon^{s-2} M \|p\|_{C^{s-1}(\mathcal{X})}; \tag{C.22}$$

this same estimate will hold for G_3 for all $s \geq 1$. Together these will imply (C.18) and (C.20).

Estimate on $G_1(x)$. If $s = 1$, then $s - 1 = 0$, and so $G_1(x) = 0$. We may therefore suppose $s \geq 2$. Recall that

$$G_1(x) = \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{q-1} \frac{\varepsilon^{j_1+j_2-2}}{j_1! j_2!} \underbrace{\int_{B(0,1)} d_x^{j_1} f(z) d_x^{j_2} p(z) \eta(\|z\|)\, dz}_{:= g_{j_1,j_2}(x)} \tag{C.23}$$

The nature of $g_{j_1,j_2}(x)$ depends on the sum $j_1 + j_2$. Since $d_x^{j_1} f d_x^{j_2}$ is an order $j_1 + j_2$ (multivariate) monomial, we have (see Section C.8.2) that whenever $j_1 + j_2$ is odd,

$$g_{j_1,j_2}(x) = \int_{\mathcal{X}} d_x^{j_1} f(z) d_x^{j_2} p(z) \eta(\|z\|)\, dz = 0.$$

In particular this is the case when $j_1 = 1$ and $j_2 = 0$. Thus when $s = 2$, $G_1(x) = g_{1,0}(x) = 0$. On the other hand if $s \geq 3$, then the lowest order terms in (C.23) are those where $j_1 + j_2 = 2$, so that either $j_1 = 1$ and $j_2 = 1$, or $j_1 = 2$ and $j_2 = 0$. We have that

$$\begin{aligned}
g_{1,1}(x) + \frac{1}{2} g_{2,0}(x) &= \int_{\mathcal{X}} d_x^1 f(z) d_x^1 p(z) \eta(\|z\|)\, dz + \frac{p(x)}{2} \int_{\mathcal{X}} d_x^2 f(z) \eta(\|z\|)\, dz \\
&= \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} D^{e_{i_1}} f(x) D^{e_{i_2}} p(x) \int_{\mathcal{X}} z^{e_{i_1}+e_{i_2}} \eta(\|z\|)\, dz + \frac{p(x)}{2} \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} D^{e_{i_2}+e_{i_2}} f(x) \int_{\mathcal{X}} z^{e_{i_1}+e_{i_2}} \eta(\|z\|)\, dz \\
&= \sum_{i=1}^{d} D^{e_i} f(x) D^{e_i} p(x) \int_{\mathcal{X}} z^2 \eta(\|z\|)\, dz + \frac{p(x)}{2} \sum_{i=1}^{d} D^{2e_i} f(x) \int_{\mathcal{X}} z^2 \eta(\|z\|)\, dz \\
&= \sigma_\eta \Delta_P f(x),
\end{aligned}$$

which is the leading term order term. Now it remains only to deal with the higher-order terms, where $j_1 + j_2 > 2$, and where it suffices to show that each function g_{j_1,j_2} satisfies (C.21) for $j = \min\{j_1 + j_2 - 2, s - 2\}$. It is helpful to write g_{j_1,j_2} using multi-index notation,

$$g_{j_1,j_2}(x) = \sum_{|\alpha_1|=j_1} \sum_{|\alpha_2|=j_2} D^{\alpha_1} f(x) D^{\alpha_2} p(x) \int_{B(0,1)} z^{\alpha_1+\alpha_2} \eta(\|z\|)\, dz,$$

where we note that $|\int_{B(0,1)} z^{\alpha_1+\alpha_2} \eta(\|z\|)\, dz| < \infty$ for all α_1, α_2, by the assumption that η is Lipschitz on its support. Finally, by Hölder's inequality we have that

$$\begin{aligned}
\|D^{\alpha_1} f D^{\alpha_2} p\|_{H^{s-(j+2)}(\mathcal{X})} &\leq \|D^{\alpha_1} f\|_{H^{s-(j+2)}(\mathcal{X})} \|D^{\alpha_2} p\|_{C^{s-(j+2)}(\mathcal{X})} \\
&\leq \|D^{\alpha_1} f\|_{H^{s-j_1}(\mathcal{X})} \|D^{\alpha_2} p\|_{C^{s-(j_2+1)}(\mathcal{X})} \\
&\leq M \cdot \|p\|_{C^{s-1}(\mathcal{X})},
\end{aligned}$$

and summing over all $|\alpha_1| = j_1$ and $|\alpha_2| = j_2$ establishes that g_{j_1,j_2} satisfies (C.21).

Estimate on $G_2(x)$. Note immediately that $G_2(x) = 0$ if $s = 1$. Otherwise if $s \geq 2$, then $q = s - 1$. Recalling that $|r_{x+z\varepsilon}^{s-1}(x; p)| \leq C\varepsilon^{s-1} \|p\|_{C^{s-1}(\mathcal{X})}$ for any $z \in B(0,1)$, and that $d_x^j f(\cdot)$ is a j-homogeneous function, we have that

$$\begin{aligned}
|G_2(x)| &\leq \sum_{j=1}^{s-1} \frac{\varepsilon^{j-2}}{j!} \int_{B(0,1)} \left| \left(d_x^j f\right)(z) \right| \cdot |r_{x+z\varepsilon}^{s-1}(x; p)| \cdot \eta(\|z\|)\, dz \\
&\leq C\varepsilon^{s-2} \|p\|_{C^{s-1}(\mathcal{X})} \sum_{j=1}^{s-1} \frac{1}{j!} \int_{B(0,1)} \left| \left(d_x^j f\right)(z) \right| \cdot \eta(\|z\|)\, dz.
\end{aligned} \tag{C.24}$$

Furthermore, for each $j = 1, \dots, s-1$ convolution of $d_x^j f$ with η only decreases the $L^2(\mathcal{X}_\varepsilon)$ norm, meaning

$$\begin{aligned}\int_{\mathcal{X}_\varepsilon} \left(\int_{B(0,1)} \left| \big(d_x^j f\big)(z) \right| \cdot \eta(\|z\|)\, dz \right)^2 dx &\le \int_{\mathcal{X}_\varepsilon} \left(\int_{B(0,1)} \left| \big(d_x^j f\big)(z) \right|^2 \eta(\|z\|)\, dz \right) \cdot \left(\int_{B(0,1)} \eta(\|z\|)\, dz \right) dx \\ &\le \int_{B(0,1)} \int_{\mathcal{X}_\varepsilon} \Big[\big(d^j f\big)(x) \Big]^2 \eta(\|z\|)\, dx\, dz \\ &\le \|d^j f\|^2_{L^2(\mathcal{X}_\varepsilon)}.\end{aligned} \tag{C.25}$$

In the above, we have used both that $|d_x^j f(z)| \le |d^j f(x)|$ for all $z \in B(0,1)$, and that the kernel is normalized so that $\int \eta(\|z\|)\, dz = 1$. Combining this with (C.24), we conclude that

$$\begin{aligned}\int_{\mathcal{X}_\varepsilon} |G_2(x)|^2\, dx &\le C\Big(\varepsilon^{s-2} \|p\|_{C^{s-1}(\mathcal{X})}\Big)^2 \sum_{j=1}^{s-1} \int_{\mathcal{X}_\varepsilon} \left(\frac{1}{j!} \int_{B(0,1)} \left| \big(d_x^j f\big)(z) \right| \cdot \Big| \eta(\|z\|) \Big|\, dz \right)^2 dx \\ &\le C\Big(\varepsilon^{s-2} \|p\|_{C^{s-1}(\mathcal{X})}\Big)^2 \sum_{j=1}^{s-1} \|d^j u\|^2_{L^2(\mathcal{X}_\varepsilon)},\end{aligned}$$

establishing the desired estimate.

Estimate on $G_3(x)$. Applying the Cauchy-Schwarz inequality, we deduce a pointwise upper bound on $|G_3(x)|^2$,

$$\begin{aligned}|G_3(x)|^2 &\le \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \cdot \left(\int_{B(0,1)} \big| r^s_{x+\varepsilon z}(x;u) \big|^2 \eta(\|z\|)\, dz \right) \cdot \left(\int_{B(0,1)} \eta(\|z\|)\, dz \right) \\ &\le \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \int_{B(0,1)} \big| r^s_{x+\varepsilon z}(x;u) \big|^2 \eta(\|z\|)\, dz.\end{aligned}$$

Applying this pointwise over all $x \in \mathcal{X}_\varepsilon$ and integrating, we obtain

$$\begin{aligned}\int_{\mathcal{X}_\varepsilon} |G_3(x)|^2\, dx &\le \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \int_{\mathcal{X}_\varepsilon} \int_{B(0,1)} \big| r^s_{x+\varepsilon z}(x;f) \big|^2 \eta(\|z\|)\, dz\, dx \\ &= \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \int_{B(0,1)} \int_{\mathcal{X}_\varepsilon} \big| r^s_{x+\varepsilon z}(x;f) \big|^2 \eta(\|z\|)\, dx\, dz \\ &\le \left(\frac{p_{\max} \varepsilon^s}{\varepsilon^2} \right)^2 \|d^s f\|^2_{L^2(\mathcal{X}_\varepsilon)},\end{aligned}$$

with the last inequality following from (C.74). Noting that $p_{\max} = \|p\|_{C^0(\mathcal{X})} \le \|p\|_{C^{s-1}(\mathcal{X})}$, we see that this is a sufficient bound on $\|G_3\|_{L^2(\mathcal{X}_\varepsilon)}$.

Proof of (C.18) and (C.20), induction step. We now assume that (C.18) and (C.20) hold for all order up to some k, and show that they then hold for order $k+1$ as well. The proof is relatively straightforward, once we introduce a bit of notation. Namely, for any $\ell, j \in \mathbb{N}$ such that $1 \le j \le \ell \le$, we will use g_j^ℓ to refer to a function satisfying

$$\|g_j^\ell\|_{H^{\ell-j}(\mathcal{X}_{(k+1)\varepsilon})} \le C \|p\|^{k+1}_{C^q(\mathcal{X})} M. \tag{C.26}$$

Note that $g_j^\ell(x) = g_{(s-\ell)+j}(x)$, so that $g_j^s(x) = g_j(x)$. As before, the functions g_j^ℓ may change from line to line, but will always satisfy (C.26). We immediately illustrate the purpose of this notation. Suppose $g \in H^\ell(\mathcal{X}_{k\varepsilon}; C\|p\|^k_{C^q(\mathcal{X})} M)$ for some $\ell \le s$. If $\ell \le 2$, then by the inductive hypothesis, it follows that for any $x \in \mathcal{X}_{(k+1)\varepsilon}$

$$L_{P,\varepsilon} g(x) = g_\ell^\ell(x) \varepsilon^{\ell-2}. \tag{C.27}$$

On the other hand if $2 < \ell \leq s$, then by the inductive hypothesis, it follows that for any $x \in \mathcal{X}_{(k+1)\varepsilon}$,

$$L_{P,\varepsilon} g(x) = \sigma_\eta \Delta_P g(x) + \sum_{j=1}^{\lfloor (\ell-1)/2 \rfloor - 1} g^\ell_{2j+2}(x)\varepsilon^{2j} + g^\ell_\ell(x)\varepsilon^{\ell-2}. \tag{C.28}$$

Proof of (C.18). If $s \leq 2(k+1)$, then by the inductive hypothesis it follows that for all $x \in \mathcal{X}_{k\varepsilon}$, we have $L^k_{P,\varepsilon} f(x) = g_s(x) \cdot \varepsilon^{s-2k}$, for some $g_s \in L^2(\mathcal{X}_{k\varepsilon}, C\|p\|^k_{C^{s-1}(\mathcal{X})} M)$. Note that we may know more about $L^k_P f(x)$ than simply that it is bounded in L^2-norm, but a bound in L^2-norm suffices. In particular, from such a bound along with (C.27) we deduce that for any $x \in \mathcal{X}_{(k+1)\varepsilon}$,

$$L^{k+1}_{P,\varepsilon} f(x) = \big(L_{P,\varepsilon} \circ L^k_{P,\varepsilon} f\big)(x) = L_{P,\varepsilon} g_s(x) \varepsilon^{s-2k} = g^s_s(x)\varepsilon^{s-2(k+1)}, \tag{C.29}$$

establishing (C.18).

Proof of (C.20). If $s > 2(k+1)$, then by the inductive hypothesis we have that for all $x \in \mathcal{X}_{k\varepsilon}$,

$$L^k_{P,\varepsilon} f(x) = \sigma^k_\eta \Delta^k_P f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - k} g_{2(j+k)}(x)\varepsilon^{2j} + g_s(x)\varepsilon^{s-2k}.$$

Thus for any $x \in \mathcal{X}_{(k+1)\varepsilon}$,

$$L^{k+1}_{P,\varepsilon} f(x) = \big(L_{P,\varepsilon} \circ L^k_{P,\varepsilon} f\big)(x) = \sigma^k_\eta L_{P,\varepsilon} \Delta^k_P f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - k} L_{P,\varepsilon} g_{2(j+k)}(x)\varepsilon^{2j} + L_{P,\varepsilon} g_s(x)\varepsilon^{s-2k}$$

There are three terms on the right hand side of this equality, and we now analyze each separately.

1. Noting that $\Delta^k_P f \in H^{s-2k}(\mathcal{X}; C\|p\|^k_{C^{s-1}(\mathcal{X})} M)$, we use (C.28) to derive that

$$\begin{aligned} L_{P,\varepsilon} \Delta^k_P f(x) &= \sigma_\eta \Delta^{k+1}_P f(x) + \sum_{j=1}^{(s-2k-1)/2-} g^{s-2k}_{2j+2}(x)\varepsilon^{2j} + g^{s-2k}_{s-2k}(x)\varepsilon^{s-2k-2} \\ &= \sigma_\eta \Delta^{k+1}_P f(x) + \sum_{j=1}^{(s-1)/2-(k+1)} g_{2(k+1+j)}(x)\varepsilon^{2j} + g_s(x)\varepsilon^{s-2(k+1)}, \end{aligned} \tag{C.30}$$

   where in the second equality we have simply used the fact $g^\ell_j(x) = g_{(s-\ell)+j}(x)$ to rewrite the equation.

2. Suppose $j < \lfloor (s-1)/2 \rfloor - k$. Then we use (C.28) to derive that

$$\begin{aligned} L_{P,\varepsilon} g_{2(j+k)}(x) &= \sigma_\eta \Delta_P g_{2(j+k)}(x) + \sum_{i=1}^{\lfloor (s-2j-2k-1)/2 \rfloor - 1} g^{s-2(j+k)}_{2(i+1)}(x)\varepsilon^{2i} + g^{s-2(j+k)}_{s-2(j+k)}(x)\varepsilon^{s-2(j+k+1)} \\ &= g_{2(j+k+1)}(x) + \sum_{i=1}^{\lfloor (s-1)/2 \rfloor - (j+k+1)} g_{2(i+j+k+1)}(x)\varepsilon^{2i} + g_s(x)\varepsilon^{s-2(j+k+1)}, \end{aligned}$$

   where in the second equality we have again used $g^\ell_j(x) = g_{(s-\ell)+j}(x)$, and also written $\sigma_\eta \Delta_P f = g^{s-2(j+k)}_2 = g_{2(j+k+1)}$, since the particular dependence on the Laplacian Δ_P will not matter. From here, multiplying by ε^{2j}, we conclude that

$$\begin{aligned} \varepsilon^{2j} L_{P,\varepsilon} g_{2(j+k)}(x) &= g_{2(j+k+1)}(x)\varepsilon^{2j} + \sum_{i=1}^{\lfloor (s-1)/2 \rfloor - (j+k+1)} g_{2(i+j+k+1)}(x)\varepsilon^{2(i+j)} + g_s(x)\varepsilon^{s-2(k+1)} \\ &= g_{2(j+k+1)}(x)\varepsilon^{2j} + \sum_{m=1}^{\lfloor (s-1)/2 \rfloor - (k+1)} g_{2(m+k+1)}(x)\varepsilon^{2m} + g_s(x)\varepsilon^{s-2(k+1)}, \end{aligned} \tag{C.31}$$

with the second equality following upon changing variables to $m = i + j$.

On the other hand if $j = \lfloor (s-1)/2 \rfloor - k$, then the calculation is much simpler,

$$\varepsilon^{2j} L_{P,\varepsilon} g_{2(j+k)}(x) = g_{s-2(j+k)}^{s-2(j+k)}(x) \varepsilon^{2j} \varepsilon^{s-2(j+k)-2} = g_s(x) \varepsilon^{s-2(k+1)}. \tag{C.32}$$

3. Finally, it follows immediately from (C.28) that

$$L_{P,\varepsilon} g_s(x) \varepsilon^{s-2k} = g_s(x) \varepsilon^{s-2(k+1)}. \tag{C.33}$$

Plugging (C.30)-(C.33) back into (C.29) proves the claim.

C.3.3 Boundary behavior of non-local Laplacian

In Lemma 48, we establish that if f is Sobolev smooth of order $s > 2k$ and zero-trace, then near the boundary of $\mathcal{X}$ the non-local Laplacian $L_{P,\varepsilon}^k f$ is close to 0 in the L^2-sense.

Lemma 48. *Assume Model 4.2.1. Let $s, k \in \mathbb{N}$. Suppose that $f \in H_0^s(\mathcal{X}; M)$. Then there exist numbers $c, C > 0$ that do not depend on M, such that for all $\varepsilon < c$,*

$$\|L_{P,\varepsilon}^k f\|_{L^2(\partial_{k\varepsilon}\mathcal{X})}^2 \leq C \varepsilon^{2(s-2k)} M^2.$$

Proof (of Lemma 48) Applying Lemma 49, we have that

$$\|L_{P,\varepsilon}^k f\|_{L^2(\partial_{k\varepsilon}(\mathcal{X}))}^2 \leq \frac{(C p_{\max})^2}{\varepsilon^4} \|L_{P,\varepsilon}^{k-1} f\|_{L^2(\partial_{k\varepsilon}(\mathcal{X}))}^2 \leq \cdots \leq \frac{(C p_{\max})^2}{\varepsilon^{4k}} \|f\|_{L^2(\partial_{k\varepsilon}(\mathcal{X}))}^2$$

Thus it remains to show that for all $\varepsilon < c$,

$$\|f\|_{L^2(\partial_{k\varepsilon}(\mathcal{X}))}^2 = \int_{\partial_{k\varepsilon}(\mathcal{X})} \big(f(x)\big)^2 \, dx \leq C_1 \varepsilon^{2s} \|f\|_{H^s(\mathcal{X})}^2. \tag{C.34}$$

We will build to (C.34) by a series of intermediate steps, following the same rough structure as the proof of Theorem 18.1 in Leoni [2017]. For simplicity, we will take $k = 1$; the exact same proof applies to the general case upon assuming $\varepsilon < c/k$.

Step 1: Local Patch. To begin, we assume that for some $c_0 > 0$ and a Lipschitz mapping $\phi : \mathbb{R}^{d-1} \to [-c_0, c_0]$, we have that $f \in C_c^\infty(U_\phi(c_0))$, where

$$U_\phi(c_0) = \Big\{ y \in Q(0, c_0) : \phi(y_{-d}) \leq y_d \Big\},$$

and here $Q(0, c_0)$ is the d-dimensional cube of side length c_0, centered at 0. We will show that for all $0 < \varepsilon < c_0$, and for the tubular neighborhood $V_\phi(\varepsilon) = \{y \in Q(0, c_0) : \phi(y_{-d}) \leq y_d \leq \phi(y_{-d}) + \varepsilon\}$, we have that

$$\int_{V_\phi(\varepsilon)} |f(x)|^2 \, dx \leq C \varepsilon^{2s} \|f\|_{H^s(U_\phi(c_0))}^2.$$

For a given $y = (y', y_d) \in V_\phi(\varepsilon)$, let $y_0 = (y', \phi(y'))$. Taking the Taylor expansion of $f(y)$ around $y = y_0$ because u is compactly supported in V_ϕ it follows that,

$$f(y) = f(y_0) + \sum_{j=1}^{s-1} \frac{1}{j!} D^{j e_d} f(y_0) \big(y_d - \phi(y')\big)^j + \frac{1}{(s-1)!} \int_{\phi(y')}^{y_d} (1-t)^{s-1} D^{s e_d} f(y', z) \big(y_d - z\big)^{s-1} \, dz \Longrightarrow$$

$$|f(y)| \leq C \varepsilon^{s-1} \int_{\phi(y')}^{y_d} \big| D^{s e_d} f(y', z) \big| \, dz.$$

Consequently, by squaring both sides and applying Cauchy-Schwarz, we have that

$$|f(y)|^2 \le C\varepsilon^{2(s-1)}\left(\int_{\phi(y')}^{y_d} \left|D^{se_d} f(y',z)\right| dz\right)^2 \le C\varepsilon^{2s-1}\int_{\phi(y')}^{y_d} \left|D^{se_d} f(y',z)\right|^2 dz.$$

Applying this bound for each $y \in V_\phi(\varepsilon)$, and then integrating, we obtain

$$\begin{aligned}\int_{V_\phi(\varepsilon)} |f(y)|^2\, dy &\le \int_{Q_{d-1}(c_0)} \int_{\phi(y')}^{\phi(y')+\varepsilon} |f(y',y_d)|^2\, dy_d\, dy' \\ &\le C\varepsilon^{2s-1}\int_{Q_{d-1}(c_0)} \int_{\phi(y')}^{\phi(y')+\varepsilon} \int_{\phi(y')}^{y_d} \left|D^{se_d} f(y',z)\right|^2 dz\, dy_d\, dy' \end{aligned} \tag{C.35}$$

where we have written $Q_{d-1}(0,c_0)$ for the $d-1$ dimensional cube of side length c_0, centered at 0. Exchanging the order of the inner two integrals then gives

$$\begin{aligned}\int_{\phi(y')}^{\phi(y')+\varepsilon} \int_{\phi(y')}^{y_d} \left|D^{se_d} f(y',z)\right|^2 dz\, dy_d &= \int_{\phi(y')}^{\phi(y')+\varepsilon} \int_{z}^{\varepsilon} \left|D^{se_d} f(y',z)\right|^2 dy_d\, dz \\ &\le C\varepsilon \int_{\phi(y')}^{\phi(y')+\varepsilon} \left|D^{se_d} f(y',z)\right|^2 dz \\ &\le C\varepsilon \int_{\phi(y')}^{c_0} \left|D^{se_d} f(y',z)\right|^2 dz.\end{aligned}$$

Finally, plugging back into (C.35), we conclude that

$$\int_{V_\phi(\varepsilon)} |f(y)|^2\, dy \le C\varepsilon^{2s} \int_{Q_{d-1}(0,c_0)} \int_{\phi(y')}^{c_0} \left|D^{se_d} f(y',z)\right|^2 dz\, dy' \le C\varepsilon^{2s}|u|^2_{H^s(U_\phi(c_0))}.$$

Step 2: Rigid motion of local patch. Now, suppose that at a point $x_0 \in \partial\mathcal{X}$, there exists a rigid motion $T : \mathbb{R}^d \to \mathbb{R}^d$ for which $T(x_0) = 0$, and a number C_0 such that for all $\varepsilon \cdot C_0 \le c_0$,

$$T\big(Q_T(x_0,c_0) \cap \partial_\varepsilon \mathcal{X}\big) \subseteq V_\phi\big(C_0\varepsilon\big) \quad \text{and} \quad T\big(Q_T(x_0,c_0) \cap \mathcal{X}\big) = U_\phi(c_0).$$

Here $Q_T(x_0,c_0))$ is a (not necessarily coordinate-axis-aligned) cube of side length c_0), centered at x_0. Define $v(y) := f(T^{-1}(y))$ for $y \in U_\phi(c_0)$. If $u \in C_c^\infty(\mathcal{X})$, then $v \in C_c^\infty(U_\phi(c_0))$, and moreover $\|v\|^2_{H^s(U_\phi(c_0))} = \|f\|^2_{H^s(Q_T(x_0,c_0)\cap\mathcal{X})}$. Therefore, using the upper bound that we derived in Step 1,

$$\int_{V_\phi(C_0\cdot\varepsilon)} |v(y)|^2\, dy \le C\varepsilon^{2s}\|v\|^2_{H^s(U_\phi(c_0))},$$

we conclude that

$$\begin{aligned}\int_{Q_T(x_0,c_0)\cap\partial_\varepsilon\mathcal{X}} |f(x)|^2\, dx &= \int_{T(Q_T(x_0,c_0))\cap\partial_\varepsilon\mathcal{X})} |v(y)|^2\, dy \\ &\le \int_{V_\phi(C_0\cdot\varepsilon)} |v(y)|^2\, dy \\ &\le C\varepsilon^{2s}\|v\|^2_{H^s(U_\phi(c_0))} = C\varepsilon^{2s}\|f\|^2_{H^s(Q_T(x_0,c_0))\cap\mathcal{X})} \le C\varepsilon^{2s}\|f\|^2_{H^s(\mathcal{X})}.\end{aligned}$$

Step 3: Lipschitz domain. Finally, we deal with the case where $\mathcal{X}$ is assumed to be an open, bounded subset of $\mathbb{R}^d$, with Lipschitz boundary. In this case, at every $x_0 \in \partial\mathcal{X}$, there exists a rigid motion $T_{x_0} : \mathbb{R}^d \to \mathbb{R}^d$

such that $T_{x_0}(x_0) = 0$, a number $c_0(x_0)$, a Lipschitz function $\phi_{x_0} : \mathbb{R}^{d-1} \to [-c_0, c_0]$, and a number $C_0(x_0)$, such that for all $\varepsilon \cdot C_0(x_0) \leq c_0(x_0)$,

$$T\big(Q_T(x_0, c_0(x_0)) \cap \partial_\varepsilon \mathcal{X}\big) \subseteq V_\phi\big(C_0(x_0) \cdot \varepsilon\big) \quad \text{and} \quad T\big(Q_T(x_0, c_0(x_0)) \cap \mathcal{X}\big) = U_\phi(c_0(x_0)).$$

Therefore for every $x_0 \in \partial\mathcal{X}$, it follows from the previous step that

$$\int_{Q_{T_{x_0}}(x_0, c_0(x_0)) \cap \partial_\varepsilon \mathcal{X}} |f(x)|^2 \, dx \leq C(x_0)\varepsilon^{2s}\|f\|^2_{H^s(\mathcal{X})},$$

where on the right hand side $C(x_0)$ is a constant that may depend on x_0, but not on u or ε.

We conclude by taking a collection of cubes that covers $\partial_\varepsilon \mathcal{X}$ for all ϵ sufficiently small. First, we note that by a compactness argument there exists a finite subset of the collection of cubes $\{Q_{T_{x_0}}(x_0, c_0(x_0)/2) : x_0 \in \partial\mathcal{X}\}$ which covers $\partial\mathcal{X}$, say $Q_{T_{x_1}}(x_1, c_0(x_1)/2), \ldots, Q_{T_{x_N}}(x_N, c_0(x_N)/2)$. Then, for any $\varepsilon \leq \min_{i=1,\ldots,N} c_0(x_i)/2$, it follows from the triangle inequality that

$$\partial_\varepsilon \mathcal{X} \subseteq \bigcup_{i=1}^N Q_{T_{x_i}}(x_i, c_0(x_i)).$$

As a result,

$$\int_{\partial_\varepsilon \mathcal{X}} |f(x)|^2 \leq \sum_{i=1}^N \int_{Q_{T_{x_i}}(x_i, c_0(x_i)) \cap \partial_\varepsilon(\mathcal{X})} |f(x)|^2 \leq \varepsilon^{2s}\|f\|^2_{H^s(\mathcal{X})} \sum_{i=1}^N C_0(x_i),$$

which proves the claim of (C.34).

C.3.4 Estimate of non-local Sobolev seminorm

Now, we use the results of the preceding two sections to prove (C.13). We will divide our analysis in two cases, depending on whether s is odd or even, but before we do this we state some facts that will be applicable to both cases. First, we recall that $L_{P,\varepsilon}$ is self-adjoint in $L^2(P)$, meaning $\langle L_{P,\varepsilon} f, g\rangle_P = \langle f, L_{P,\varepsilon} g\rangle_P$ for all $f, g \in L^2(P)$. We also recall the definition of the Dirichlet energy $E_{P,\varepsilon}(f; \mathcal{X})$,

$$\langle L_{P,\varepsilon} f, f\rangle_P = \frac{1}{\varepsilon^{d+2}} \int_{\mathcal{X}} \int_{\mathcal{X}} \big(f(x) - f(x')\big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x') \, dP(x) =: E_{P,\varepsilon}(f; \mathcal{X}). \tag{C.36}$$

Finally, we recall a result of Green et al. [2021]: there exist constants c_0 and C_0 which do not depend on M, such that for all $\varepsilon < c_0$ and for any $f \in H^1(\mathcal{X}; M)$,

$$E_{P,\varepsilon}(f; \mathcal{X}) \leq C_0 M^2. \tag{C.37}$$

Case 1: s odd. Suppose s is odd, so that $s \geq 3$. Taking $k = (s-1)/2$, we use the self-adjointness of $L_{P,\varepsilon}$ to relate the non-local semi-norm $\langle L^s_{P,\varepsilon} f, f\rangle_P$ to a non-local Dirichlet energy,

$$\langle L^s_{P,\varepsilon} f, f\rangle_P = \langle L^{k+1}_{P,\varepsilon} f, L^k_{P,\varepsilon} f\rangle_P = E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathcal{X}).$$

We now separate this energy into integrals over $\mathcal{X}_{k\varepsilon}$ and $\partial_{k\varepsilon}(\mathcal{X})$,

$$\begin{aligned} E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathcal{X}) = \frac{1}{\varepsilon^{d+2}} \Bigg\{ &\int_{\mathcal{X}_{k\varepsilon}} \int_{\mathcal{X}_{k\varepsilon}} \big(L^k_{P,\varepsilon} f(x) - L^k_{P,\varepsilon} f(x')\big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x') \, dP(x) \\ &+ \int_{\partial_{k\varepsilon}\mathcal{X}} \int_{\partial_{k\varepsilon}\mathcal{X}} \big(L^k_{P,\varepsilon} f(x) - L^k_{P,\varepsilon} f(x')\big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x') \, dP(x) \Bigg\} \\ := E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathcal{X}_{k\varepsilon}) &+ E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \partial_{k\varepsilon}\mathcal{X}) \end{aligned} \tag{C.38}$$

and upper bound each energy separately. For the first term, we add and substract $\sigma_\eta^k \Delta_P^k f(x)$ and $\sigma_\eta^k \Delta_P^k f(x')$ within the integrand, then use the triangle inequality and the symmetric role played by x and x' to deduce that

$$E_{P,\varepsilon}(L_{P,\varepsilon}^k f; \mathcal{X}_{k\varepsilon}) \leq 3\sigma_\eta^{2k} E_{P,\varepsilon}(\Delta_P^k f; \mathcal{X}_{k\varepsilon}) + \frac{2}{\varepsilon^{d+2}} \int_{\mathcal{X}_{k\varepsilon}} \int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}^k f(x) - \sigma_\eta^k \Delta_P^k f(x)\right)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dP(x')\, dP(x). \tag{C.39}$$

Noticing that $\Delta_P^k f \in H^1(\mathcal{X}; \|p\|_{C^{s-1}(\mathcal{X})}^k M)$, we use (C.37) to conclude that $E_{P,\varepsilon}(\Delta_P^k f; \mathcal{X}_{k\varepsilon}) \leq C_0 M^2$. On the other hand, it follows from Assumption (K1) and (C.16) that

$$\begin{aligned}\frac{2}{\varepsilon^{d+2}} \int_{\mathcal{X}_{k\varepsilon}} \int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}^k f(x) - \sigma_\eta^k \Delta_P^k f(x)\right)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dP(x')\, dP(x) &\leq \frac{2p_{\max}}{\varepsilon^2} \int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}^k f(x) - \sigma_\eta^k \Delta_P^k f(x)\right)^2 dP(x) \\ &\leq C_1 M^2.\end{aligned}$$

Plugging these two bounds into (C.39) gives the desired upper bound on $E_{P,\varepsilon}(L_{P,\varepsilon}^k; \mathcal{X}_{k\varepsilon})$.

For the second term in (C.38), we apply Lemmas 50 and 48 and conclude that,

$$E_{P,\varepsilon}(L_{P,\varepsilon}^k f; \partial_{k\varepsilon}\mathcal{X}) \leq \frac{4p_{\max}^2}{\varepsilon^2} \|L_{P,\varepsilon}^k f\|_{L^2(\partial_{k\varepsilon}\mathcal{X})} \leq CM^2.$$

Case 2: s even. If $s \in \mathbb{N}$ is even, $s \geq 2$, then letting $k = (s-2)/2$, the self-adjointness of $L_{P,\varepsilon}$ implies

$$\langle L_{P,\varepsilon}^s f, f\rangle_P = \|L_{P,\varepsilon}^{k+1} f\|_P^2.$$

As in the first case, we divide the integral up into the interior region $\mathcal{X}_{k\varepsilon}$ and the boundary region $\partial_{k\varepsilon}\mathcal{X}$,

$$\|L_{P,\varepsilon}^{k+1} f\|_P^2 \leq p_{\max} \|L_{P,\varepsilon}^{k+1} f\|_{L^2(\mathcal{X})}^2 \leq p_{\max}\left\{\int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 dP(x) + \int_{\partial_{k\varepsilon}\mathcal{X}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 dP(x)\right\}, \tag{C.40}$$

and upper bound each term separately. For the first term, adding and subtracting $\sigma_\eta^k \Delta_P^k f(x)$ gives

$$\begin{aligned}\int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 dP(x) &\leq 2\int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon} \Delta_P^k f(x)\right)^2 dP(x) + 2\int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}\left(L_{P,\varepsilon}^k f - \sigma_\eta \Delta_P^k f\right)(x)\right)^2 dP(x) \\ &\overset{(i)}{\leq} CM^2 + 2\int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}\left(L_{P,\varepsilon}^k f - \sigma_\eta \Delta_P^k f\right)(x)\right)^2 dP(x) \\ &\overset{(ii)}{\leq} CM^2 + \frac{Cp_{\max}^2}{\varepsilon^2} \|L_{P,\varepsilon}^k f - \sigma_\eta \Delta_P^k f\|_{L^2(\mathcal{X}_{k\varepsilon})}^2 \\ &\overset{(iii)}{\leq} CM^2,\end{aligned}$$

with (i) following from (C.18) since $\Delta_P^k f \in H^2(\mathcal{X}; M\|p\|_{C^{s-1}(\mathcal{X})}^l)$, (ii) following from Lemma 49, and (iii) following from (C.17).

Then Lemma 48 shows that the second term in (C.40) satisfies

$$\int_{\partial_{k\varepsilon}\mathcal{X}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 dP(x) \leq CM^2.$$

C.3.5 Assorted integrals

Lemma 49. *Assume Model 4.2.1. Suppose $f \in L^2(U; M)$ for a Borel set $U \subseteq \mathcal{X}$, and let $L_{P,\varepsilon}$ be defined with respect to a kernel η that satisfies (K1). Then there exists a constant C which does not depend on f or M such that*

$$\|L_{P,\varepsilon} f\|_{L^2(U)} \leq \frac{2p_{\max}}{\varepsilon^2} \|f\|_{L^2(U)} \tag{C.41}$$

Lemma 50. *Assume Model 4.2.1. Suppose $f \in L^2(U;M)$ for a Borel set $U \subseteq \mathcal{X}$, and let $L_{P,\varepsilon}$ be defined with respect to a kernel η that satisfies (K1). Then there exists a constant C which does not depend on f or M such that*

$$E_{P,\varepsilon}(f;U) \leq \frac{4p_{\max}^2}{\varepsilon^2}\|f\|_{L^2(U)}^2 \tag{C.42}$$

Lemma 51. *Assume Model 4.2.1. Suppose $f \in H^1(\mathcal{X};M)$, and let $D_i f$ be defined with respect to a kernel η that satisfies (K1). Then there exists a constant C which does not depend on f or M, such that for any $i \in [n]$ and $\mathbf{j} \in [n]^s$,*

$$\mathbb{E}\Big[|D_{\mathbf{j}}f(X_i)| \cdot |f(X_i) - f(X_{\mathbf{j}_1})|\Big] \leq C\varepsilon^{2+dk}M^2,$$

where $k+1$ is the number of distinct indices in $i\mathbf{j}$.

Proof (of Lemma 49). We fix a version of $f \in L^2(U)$, so that we may speak of its pointwise values.

At a given point $x \in U$, we can upper bound $|L_{P,\varepsilon}f(x)|^2$ using the Cauchy-Schwarz inequality as follows,

$$\begin{aligned}
|L_{P,\varepsilon}f(x)|^2 &\leq \left(\frac{p_{\max}}{\varepsilon^{2+d}}\right)^2 \left(\int_U \big(|f(x')| + |f(x)|\big)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'\right)^2 \\
&\leq \left(\frac{p_{\max}}{\varepsilon^{2+d}}\right)^2 \left(\int_U \big(|f(x')| + |f(x)|\big)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx' \cdot \int \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'\right) \\
&= \frac{p_{\max}^2}{\varepsilon^{4+d}} \int_U \big(|f(x')| + |f(x)|\big)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'.
\end{aligned}$$

The equality follows by the assumption $\int_{\mathbb{R}^d} \eta(\|z\|)\,dx = 1$ in (K1). Integrating over all $x \in U$, it follows from the triangle inequality that

$$\begin{aligned}
\|L_{P,\varepsilon}\|_{L^2(U)}^2 &\leq \frac{2p_{\max}^2}{\varepsilon^{4+d}} \int_U \int_U \big(|f(x')|^2 + |f(x)|^2\big) \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'\,dx \\
&\leq \frac{2p_{\max}^2}{\varepsilon^{4+d}} \int_U \int_U \big(|f(x')|^2 + |f(x)|^2\big) \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'\,dx.
\end{aligned} \tag{C.43}$$

Finally, using Fubini's Theorem we determine that

$$\int_U \int_U \big(|f(x')|^2 + |f(x)|^2\big) \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'\,dx = 2\int_U \int_U |f(x)|^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx \leq 2\varepsilon^d \int_U |f(x)|^2\,dx = 2\varepsilon^d \|f\|_{L^2(U)}^2, \tag{C.44}$$

and by combining (C.43) and (C.44) we conclude that

$$\|L_{P,\varepsilon}\|_{L^2(U)}^2 \leq \frac{4p_{\max}^2}{\varepsilon^4}\|f\|_{L^2(U)}^2.$$

Proof (of Lemma 50). We have

$$E_{P,\varepsilon}(f) = \frac{1}{\varepsilon^{2+d}} \int_U \int_U \big(f(x) - f(x')\big)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dP(x')\,dP(x) \leq \frac{2p_{\max}^2}{\varepsilon^{2+d}} \int_U \int_U \big(|f(x)|^2 + |f(x')|^2\big) \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx'\,dx,$$

and the claim follows from (C.44).

Proof (of Lemma 51). Let $G_{n,\varepsilon}[X_{i\mathbf{j}}]$ be the subgraph induced by vertices $X_i, X_{\mathbf{j}_1}, \ldots, X_{\mathbf{j}_s}$. We make two observations. First, in order for $|D_{\mathbf{j}} f(X_i)| \cdot |f(X_i) - f(X_j)|$ to be non-zero, it must be the case that the subgraph $G_{n,\varepsilon}[X_{i\mathbf{j}}]$ is connected. Second, noting that for any indices i and j,

$$|D_{ij} f(x)| \leq \Big(|D_j f(X_i)| + |D_j f(x)|\Big) \|\eta\|_\infty,$$

a straightforward inductive argument implies that

$$|D_{\mathbf{j}} f(X_i)| \leq s\|\eta\|_\infty^s \sum_{j \in i\mathbf{j}} |D_{\mathbf{j}_s} f(X_j)|.$$

Combining these two observations, we reduce the task to upper bounding the product of two (first-order) differences,

$$\begin{aligned}
\mathbb{E}\Big[|D_{\mathbf{j}} f(X_i)||f(X_i) - f(X_{\mathbf{j}_1})|\Big] &= \mathbb{E}\Big[|D_{\mathbf{j}} f(X_i)||f(X_i) - f(X_{\mathbf{j}_1})| \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big] \\
&\leq s\|\eta\|_\infty^s \sum_{j \in i\mathbf{j}} \mathbb{E}\Big[|D_{\mathbf{j}_s} f(X_j)| \cdot |f(X_i) - f(X_{\mathbf{j}_1})| \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big] \\
&\leq s\|\eta\|_\infty^s \sum_{j \in i\mathbf{j}} \mathbb{E}\Big[|f(X_j) - f(X_{\mathbf{j}_s})| \cdot |f(X_i) - f(X_{\mathbf{j}_1})| \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big]
\end{aligned}$$

Next, from the Cauchy-Schwarz inequality we have that for any $j \in \mathbf{j}$,

$$\begin{aligned}
&\mathbb{E}\Big[|f(X_j) - f(X_{\mathbf{j}_s})| \cdot |f(X_i) - f(X_{\mathbf{j}_1})| \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big] \\
&\quad \leq \sqrt{\mathbb{E}\Big[|f(X_j) - f(X_{\mathbf{j}_s})|^2 \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big]} \cdot \sqrt{\mathbb{E}\Big[|f(X_j) - f(X_{\mathbf{j}_s})|^2 \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big]} \\
&\quad = \mathbb{E}\Big[|f(X_j) - f(X_i)|^2 \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big],
\end{aligned}$$

with the equality following since each X_i are identically distributed. Marginalizing out the contribution of all indices in $\mathbf{j}$ not equal to i or j gives

$$\begin{aligned}
\mathbb{E}\Big[|f(X_j) - f(X_i)|^2 \cdot \mathbf{1}\big\{G_{n,\varepsilon}[X_{i\mathbf{j}}] \text{ is connected.}\big\}\Big] &\leq \big((s+1)p_{\max}\nu_d\varepsilon^d\big)^{|i\mathbf{j}\setminus\{j\cup i\}|} \cdot \mathbb{E}\Big[|f(X_j) - f(X_i)|^2 \mathbf{1}\{\|X_i - X_j\| \leq \varepsilon\}\Big] \\
&\leq \big((s+1)p_{\max}\nu_d\varepsilon^d\big)^{|i\mathbf{j}\setminus\{j\cup i\}|} \cdot p_{\max}^2 \nu_d \varepsilon^{2+d} M^2 \qquad \text{(C.45)}
\end{aligned}$$

with the second inequality following from the proof of Lemma 1 in Green et al. [2021]. Finally, we notice that $|i\mathbf{j} \setminus \{i \cup j\}| + 1 = k$, so that (C.45) gives the desired result.

C.4 Graph Sobolev semi-norm, manifold domain

In this section we prove Proposition 11. Note that when $s = 1$, the upper bound (4.35) follows immediately from Lemma 53 and Markov's inequality.

On the other hand when $s = 2$ or $s = 3$, we prove Proposition 11 by first establishing some intermediate results, many of which are analogous to results we have already shown in the flat Euclidean case. Indeed, in some ways the proof will be simpler in the manifold setting than in the flat Euclidean case: there is no boundary, and we do not need to analyze the iterated nonlocal Laplacian $L_{P,\varepsilon}^j$ for $j > 1$.

That being said, as mentioned in our main text, in the manifold setting there is some extra error induced by using Euclidean rather than geodesic distance. We upper bound this error by comparing $L_{P,\varepsilon}$ to an alternative

nonlocal Laplacian $\widetilde{L}_{P,\varepsilon}$, which is defined with respect to geodesic distance. Precisely, let $d_{\mathcal{X}}(x, x')$ denote the geodesic distance between $x, x' \in \mathcal{X}$, and define

$$\widetilde{L}_{P,\varepsilon} f(x) := \int_{\mathcal{X}} \big(f(x') - f(x)\big) \eta\bigg(\frac{d_{\mathcal{X}}(x', x)}{\varepsilon}\bigg) p(x')\, dx'.$$

We show the following results, each of which hold under the same assumptions as Proposition 11.

- In Section C.4.1 we show that the graph Sobolev seminorm $\langle L^s_{n,\varepsilon} f, f\rangle_n$ is upper bounded by the sum of a nonlocal seminorm and a pure bias term: specifically, with probability at least $1 - 2\delta$,

$$\langle L^s_{n,\varepsilon} f, f\rangle_n \leq \frac{\langle L^s_{P,\varepsilon} f, f\rangle_P}{\delta} + C_1 \frac{\varepsilon^2}{n\varepsilon^{2s+m}} M^2. \tag{C.46}$$

  This upper bound is essentially the same as (C.12), but with the intrinsic dimension m taking the place of the ambient dimension d. The pure bias term will be of at most constant order when $\varepsilon \gtrsim n^{-1/(2(s-1)+m)}$.

- In Section C.4.2, we show that the error incurred by using the "wrong" metric is negligible. Precisely, we find that

$$\|L_{P,\varepsilon} f - \widetilde{L}_{P,\varepsilon} f\|^2_{L^2(\mathcal{X})} \leq C_2 \varepsilon^2 |f|^2_{H^1(\mathcal{X})}. \tag{C.47}$$

- In Section C.4.3, we analyze the approximation error of $\widetilde{L}_{P,\varepsilon}$. We show that when $f \in H^2(\mathcal{X})$ and $p \in C^1(\mathcal{X})$,

$$\|\widetilde{L}_{P,\varepsilon} f\|^2_{L^2(\mathcal{X})} \leq C_3 \|f\|^2_{H^2(\mathcal{X})}, \tag{C.48}$$

  whereas if $f \in H^3(\mathcal{X})$ and $p \in C^2(\mathcal{X})$,

$$\|\widetilde{L}_{P,\varepsilon} f - \sigma_\eta \Delta_P f\|^2_{L^2(\mathcal{X})} \leq C_3 \varepsilon^2 \|f\|^2_{H^3(\mathcal{X})}. \tag{C.49}$$

- In Section C.4.4, we use the results of the preceding two sections to show that if $f \in H^s(\mathcal{X})$ and $p \in C^{s-1}(\mathcal{X})$, then

$$\langle L^s_{P,\varepsilon} f, f\rangle_P \leq C_4 \|f\|^2_{H^s(\mathcal{X})}. \tag{C.50}$$

- In Section C.4.5 we state some technical results used in the previous sections.

We point out that when f is Hölder smooth, results analogous to (C.49) have been established in Calder and García Trillos [2019]. When f is Sobolev smooth, our analysis (which relies heavily on Taylor expansions) is largely similar, except that the remainder term in the relevant Taylor expansion will be bounded in $L^2(\mathcal{X})$ norm rather than $L^\infty(\mathcal{X})$ norm. This is analogous to the situation in the flat Euclidean model.

Proof (of Proposition 11). Follows immediately from (C.46) and (C.50). □

C.4.1 Decomposition of graph Sobolev seminorm

The proof of (C.46) is identical to the proof of (C.12), except substituting the intrinsic dimension m for ambient dimension d, and using Lemma 55 rather than Lemma 51.

C.4.2 Error due to Euclidean Distance

In this section, we prove (C.47). By applying Cauchy-Schwarz we obtain an upper bound on $|L_{P,\varepsilon}f(x) - \widetilde{L}_{P,\varepsilon}f(x)|^2$:

$$\begin{aligned}\big[L_{P,\varepsilon}f(x) - \widetilde{L}_{P,\varepsilon}f(x)\big]^2 &\le \frac{p_{\max}^2}{\varepsilon^{2(2+m)}}\int_{\mathcal{X}}\big[f(x')-f(x)\big]^2\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)-\eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right|d\mu(x') \\ &\quad\cdot\int_{\mathcal{X}}\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)-\eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right|d\mu(x') \\ &= \frac{1}{\varepsilon^{2(2+m)}}A_1(x)\cdot A_2(x) \qquad\text{(C.51)}\end{aligned}$$

Thus we have upper bounded $|L_{P,\varepsilon}f(x) - \widetilde{L}_{P,\varepsilon}f(x)|^2$ by the product of two terms, each of which we now suitably bound.

To do so, we will use the following estimates, from Proposition 4 of García Trillos et al. [2019a]: letting R denote the reach of $\mathcal{X}$, for all $\|x'-x\| \le R/2$,

$$\|x'-x\| \le d_{\mathcal{X}}(x',x) \le \|x'-x\| + \frac{8}{R^2}\|x'-x\|^3. \qquad\text{(C.52)}$$

Upper bound on $A_1(x)$. From here forward we will assume $\varepsilon < R/2$. Consequently $\eta(\|x'-x\|/\varepsilon) \ge \eta(d_{\mathcal{X}}(x',x)/\varepsilon)$. Furthermore, letting L_η denote the Lipschitz constant of η, and setting $\widetilde{\varepsilon} := (1+27\varepsilon^2/R^2)\varepsilon$ we have that

$$\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)-\eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right| \le \frac{L_\eta 8\varepsilon^2}{R^2}\cdot\mathbf{1}\big\{d_{\mathcal{X}}(x',x)\le\varepsilon\big\} + \|\eta\|_\infty\cdot 1\{\varepsilon < d_{\mathcal{X}}(x',x)\le\widetilde{\varepsilon}\}.$$

Thus,

$$A_1(x) \le \frac{8L_\eta\varepsilon^2}{R^2}\int_{\mathcal{X}}\big[f(x')-f(x)\big]^2\mathbf{1}\{\|x'-x\|\le\varepsilon\}\,d\mu(x') + \|\eta\|_\infty\int_{\mathcal{X}}\big[f(x')-f(x)\big]^2\mathbf{1}\big\{\varepsilon < d_{\mathcal{X}}(x',x)\le\widetilde{\varepsilon}\big\}\,d\mu(x')$$

Integrating over $\mathcal{X}$, we conclude from Lemma 54 and Lemma 3.3 of [Burago et al., 2014] and that

$$\int_{\mathcal{X}}A_1(x)\,d\mu(x) \le \frac{8L_\eta\nu_m\varepsilon^2}{R^2(m+2)}\Big(1+CmKR^2\Big)\varepsilon^{m+2}|f|^2_{H^1(\mathcal{X})} + C\|\eta\|_\infty\varepsilon^{m+4}|f|^2_{H^1(\mathcal{X})} =: C_5\varepsilon^{m+4}|f|_{H^1(\mathcal{X})}.$$

Upper bound on $A_2(x)$. Integrating over $x'\in\mathcal{X}$, we see that

$$\begin{aligned}\int_{\mathcal{X}}\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)-\eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right|d\mu(x') &\le \frac{8L_\eta\varepsilon^2}{R^2}\int_{\mathcal{X}}\mathbf{1}\big\{d_{\mathcal{X}}(x',x)\big\}\,d\mu(x') + p_{\max}\|\eta\|_\infty\int_{\mathcal{X}}\mathbf{1}\big\{\varepsilon < d_{\mathcal{X}}(x',x)\le\widetilde{\varepsilon}\big\}\,d\mu(x') \\ &= \frac{8L_\eta\varepsilon^2}{R^2}\cdot\mu\big(B(x,\varepsilon)\big) + p_{\max}\|\eta\|_\infty\Big[\mu\big(B(x,\widetilde{\varepsilon})\big)-\mu\big(B(x,\varepsilon)\big)\Big]. \qquad\text{(C.53)}\end{aligned}$$

Equation (1.36) in García Trillos et al. [2019a] states that

$$\big|\mu(B_{\mathcal{X}}(x,\varepsilon)) - \omega_m\varepsilon^m\big| \le CmK\varepsilon^{m+2},$$

where K is an upper bound on the sectional curvature of $\mathcal{X}$. Plugging this back into (C.53), we conclude that

$$\begin{aligned}\int_{\mathcal{X}}\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)-\eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right|d\mu(x') &\le \frac{8L_\eta\varepsilon^2}{R^2}\Big[\omega_m\varepsilon^m + CmK\varepsilon^{m+2}\Big] + \|\eta\|_\infty\Big[\omega_m(\widetilde{\varepsilon}^m-\varepsilon^m) + 2CmK\varepsilon^{m+2}\Big] \\ &\le \frac{8L_\eta\varepsilon^2}{R^2}\Big[\omega_m\varepsilon^m + R^2CmK\varepsilon^m\Big] + \|\eta\|_\infty\varepsilon^{m+2}\Big[\frac{27\omega_m}{R^2}+2CmK\Big] \\ &=: C_6\varepsilon^{m+2}.\end{aligned}$$

Putting together the pieces. Plugging our upper bounds on $A_1(x)$ and $A_2(x)$ back into (C.51), we deduce that

$$\begin{aligned}\|\widetilde{L}_{P,\varepsilon}f - L_{P,\varepsilon}f\|^2_{L^2(\mathcal{X})} &\leq \frac{1}{\varepsilon^{2(2+m)}}\int_{\mathcal{X}} A_1(x)\cdot A_2(x)\,d\mu(x)\\ &\leq \frac{C_6}{\varepsilon^{(2+m)}}\int_{\mathcal{X}} A_1(x)\,d\mu(x)\\ &\leq C_5C_6\varepsilon^2|f|^2_{H^1(\mathcal{X})},\end{aligned}$$

thus proving the claimed result.

C.4.3 Approximation Error of non-local Laplacian

Fix $x \in \mathcal{X}$. We begin with a pointwise estimate of $\widetilde{L}_{P,\varepsilon}f$, facilitated by expressing $w(v) = f(\exp_x(v))$ and $q(v) = p(\exp_x(v))$ in normal coordinates, as in [Calder and García Trillos, 2019]. Letting $J_x(v)$ be the Jacobian of the exponential map $\exp_x : B(0,\varepsilon) \subseteq T_x(\mathcal{X}) \to B_{\mathcal{X}}(x,\varepsilon)$ evaluated at $v \in T_x(\mathcal{X})$, we have

$$\begin{aligned}\widetilde{L}_{P,\varepsilon}f(x) &= \frac{1}{\varepsilon^{m+2}}\int_{\mathcal{X}}\big(f(x') - f(x)\big)\eta\bigg(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\bigg)\,dP(x')\\ &= \frac{1}{\varepsilon^{m+2}}\int_{B(0,\varepsilon)\subset T_x(\mathcal{X})}\big(w(v) - w(0)\big)\eta\bigg(\frac{\|v\|}{\varepsilon}\bigg)J_x(v)q(v)\,dv\\ &= \frac{1}{\varepsilon^2}\bigg\{\int_{B(0,1)}\big(w(\varepsilon v) - w(0)\big)\eta(\|v\|)q(\varepsilon v)\,dv + \int_{B(0,1)}\big(w(\varepsilon v) - w(0)\big)\eta(\|v\|)q(\varepsilon v)\big(J_x(\varepsilon v) - 1\big)\,dv\bigg\}\\ &= A_1(x) + A_2(x)\end{aligned}$$

Note that w and q have the same smoothness properties as f and p. Moreover, arguing exactly as we did in the flat Euclidean case, we can show that when $f \in H^2(\mathcal{X})$ and $p \in C^1(\mathcal{X})$, then

$$\|A_1\|^2_{L^2(\mathcal{X})} \leq C\|f\|^2_{H^2(\mathcal{X})}$$

whereas if $f \in H^3(\mathcal{X})$ and $p \in C^2(\mathcal{X})$ then

$$\|A_1 - \sigma_\eta\Delta_P f\|^2_{L^2(\mathcal{X})} \leq C\|f\|^2_{H^3(\mathcal{X})}\varepsilon^2.$$

Therefore it remains only to upper bound A_2 in $L^2(\mathcal{X})$ norm. To do so, we recall (1.34) of García Trillos et al. [2019a]: for any $\varepsilon < i_0$ and all $x \in \mathcal{X}$, the Jacobian $J_x(v)$ satisfies the upper bound

$$|J_x(v) - 1| \leq CmK\varepsilon^2, \quad \text{for all} \;\; v \in B(0,\varepsilon) \subseteq T_x(\mathcal{X}).$$

Combining this estimate with the Cauchy-Schwarz inequality, we conclude that

$$\begin{aligned}\|A_2\|^2_{L^2(\mathcal{X})} &\leq Cm^2K^2\bigg[\int_{B(0,1)}\big(w(\varepsilon v) - w(0)\big)^2\eta(\|v\|)q(\varepsilon v)\,dv\bigg]\cdot\bigg[\int_{B(0,1)}\eta(\|v\|)q(\varepsilon v)\,dv\bigg]\\ &\leq Cm^2K^2\sigma_\eta(1 + L_q\varepsilon)\int_{B(0,1)}\big(w(\varepsilon v) - w(0)\big)^2\eta(\|v\|)q(\varepsilon v)\,dv\\ &\leq Cm^2K^2\sigma_\eta^2(1 + L_q\varepsilon)p_{\max}\varepsilon^2|f|^2_{H^1(\mathcal{X})},\end{aligned}$$

with the final inequality following from (3.2) of Burago et al. [2014]. Combining our estimates on A_1 and A_2 yields the claim.

C.4.4 Estimate of non-local Sobolev seminorm

In this subsection we establish that the upper bound (C.50) holds when $f \in H^s(\mathcal{X})$ and $p \in C^{s-1}(\mathcal{X})$. We first consider $s = 2$, and then $s = 3$.

Case 1: $s = 2$. When $s = 2$, the triangle inequality implies that

$$\langle L_{P,\varepsilon}^s f, f\rangle_P \le 2p_{\max}\Big(\|L_{P,\varepsilon}f - \widetilde{L}_{P,\varepsilon}\|_{L^2(\mathcal{X})}^2 + \|\widetilde{L}_{P,\varepsilon}f\|_{L^2(\mathcal{X})}^2\Big)$$

The first term on the right hand side is upper bounded in (C.47), and the second term is upper bounded in (C.48). Together these estimates imply the claim.

Case 2: $s = 3$. When $s = 3$, the triangle inequality implies that

$$\langle L_{P,\varepsilon}^s f, f\rangle_P = E_{P,\varepsilon}(L_{P,\varepsilon}f; \mathcal{X}) \le 3\Big(E_{P,\varepsilon}(L_{P,\varepsilon}f - \widetilde{L}_{P,\varepsilon}f; \mathcal{X}) + E_{P,\varepsilon}(\widetilde{L}_{P,\varepsilon}f - \sigma_\eta \Delta_P f; \mathcal{X}) + \sigma_\eta^2 E_{P,\varepsilon}(\Delta_P f; \mathcal{X})\Big)$$

We now upper bound each of the three terms on the right hand side of the above inequality. First, we note that by Lemma 52 and (C.47),

$$E_{P,\varepsilon}(L_{P,\varepsilon}f - \widetilde{L}_{P,\varepsilon}f; \mathcal{X}) \le \frac{C}{\varepsilon^2}\|L_{P,\varepsilon}f - \widetilde{L}_{P,\varepsilon}f\|_{L^2(\mathcal{X})}^2 \le C|f|_{H^1(\mathcal{X})}^2.$$

An equivalent upper bound on $E_{P,\varepsilon}(\widetilde{L}_{P,\varepsilon}f - \sigma_\eta \Delta_P f; \mathcal{X})$ follows from Lemma 52 and (C.49). Finally, we notice that $f \in H^3(\mathcal{X})$ and $p \in C^2(\mathcal{X})$ implies $\Delta_P f \in H^1(\mathcal{X})$, and furthermore $|\Delta_P f|_{H^1(\mathcal{X})} \le \|p\|_{C^2(\mathcal{X})} \cdot \|f\|_{H^3(\mathcal{X})}$. We conclude from Lemma 53 that

$$E_{P,\varepsilon}(\Delta_P f; \mathcal{X}) \le C|\Delta_P f|_{H^1(\mathcal{X})}^2 \le C\|f\|_{H^3(\mathcal{X})}^2,$$

where in the final inequality we have absorbed $\|p\|_{C^2(\mathcal{X})}$ into the constant C. Together, these upper bounds prove the claim.

C.4.5 Integrals

Recall the Dirichlet energy $E_{P,\varepsilon}(f; \mathcal{X}) = \langle L_{P,\varepsilon}f, f\rangle_P$, defined in (C.36). Now we establish some estimates on $E_{P,\varepsilon}(f; \mathcal{X})$ under Model 4.2.2, and under various assumptions regarding the regularity of f.

Lemma 52. *Suppose Model 4.2.2, and additionally that $f \in L^2(\mathcal{X})$. Then there exists a constant C such that*

$$E_{P,\varepsilon}(f; \mathcal{X}) \le \frac{C}{\varepsilon^2}\|f\|_{L^2(\mathcal{X})}^2. \tag{C.54}$$

Lemma 53. *Suppose Model 4.2.2, and additionally that $f \in H^1(\mathcal{X})$. Then there exist constants c and C which do not depend on f such that for any $0 < \varepsilon < c$,*

$$E_{P,\varepsilon}(f; \mathcal{X}) \le C|f|_{H^1(\mathcal{X})}^2. \tag{C.55}$$

We use Lemma 54 to help upper bound the error incurred by using $\|\cdot\|$ rather than $d_{\mathcal{X}}(\cdot,\cdot)$. Recall the notation $\widetilde{\varepsilon} = (1 + 27\varepsilon^2/R^2)\varepsilon$, where R is the reach of $\mathcal{X}$.

Lemma 54. *Suppose Model 4.2.2, and additionally that $f \in H^1(\mathcal{X})$. There exist constants c and C such that for any $\varepsilon < c$,*

$$\int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 \mathbf{1}\{\varepsilon < d_{\mathcal{X}}(x', x) \le \widetilde{\varepsilon}\}\, d\mu(x')\, d\mu(x) \le C\varepsilon^{4+m}\|f\|_{H^1(\mathcal{X})}^2 \tag{C.56}$$

Finally, we use Lemma 55 to show that the pure bias component of $\langle L_n^s f, f_{,\rangle}\, n$ is small in expectation. This is analogous to Lemma 51, except assuming Model 4.2.2 rather than Model 4.2.1.

Lemma 55. *Assume Model 4.2.2. Suppose $f \in H^1(\mathcal{X})$, and let $D_i f$ be defined with respect to a kernel η that satisfies (P5). Then there exists a constant C which does not depend on f or n, such that for any $i \in [n]$ and $\mathbf{j} \in [n]^s$,*

$$\mathbb{E}\Big[|D_{\mathbf{j}} f(X_i)| \cdot |f(X_i) - f(X_{\mathbf{j}_1})|\Big] \leq C\varepsilon^{2+mk} \cdot \|f\|^2_{H^1(\mathcal{X})},$$

where $k+1$ is the number of distinct indices in $i\mathbf{j}$.

Proof (of Lemmas 52 and 53). Define the non-local energy $\widetilde{E}_{P,\varepsilon}$ with respect to geodesic distance,

$$\widetilde{E}_{P,\varepsilon}(f;\mathcal{X}) := \langle \widetilde{L}_{P,\varepsilon} f, f\rangle_P = \int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 \eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right) dP(x')\, dP(x).$$

From the lower bound in (C.52), it follows that $E_{P,\varepsilon}(f;X) \leq \widetilde{E}_{P,\varepsilon}(f;\mathcal{X})$, and from the upper bounds $p(x) \leq p_{\max}$ and $\eta(|x|) \leq \|\eta\|_\infty \cdot \mathbf{1}\{x \in [-1,1]\}$ we further have

$$\widetilde{E}_{P,\varepsilon}(f;\mathcal{X}) \leq p^2_{\max}\|\eta\|_\infty \cdot \int_{\mathcal{X}}\int_{B_{\mathcal{X}}(\varepsilon)} \big(f(x') - f(x)\big)^2 \, d\mu(x')\, d\mu(x).$$

The estimates (C.54) and (C.55) then respectively follow from (3.1) and Lemma 3.3 of Burago et al. [2014].

Proof (of Lemma 54). Following exactly the steps of the proof of Lemma 3.3 of Burago et al. [2014], but replacing all references to a ball of radius r by references to the set difference between balls of radius $\widetilde{\varepsilon}$ and ε, we obtain that

$$\int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 \mathbf{1}\{\varepsilon < d_{\mathcal{X}}(x',x) \leq \widetilde{\varepsilon}\}\, d\mu(x')\, d\mu(x) \leq (1 + CmK\varepsilon^2) \cdot \int_{\mathcal{X}}\int_{B_m(0,\widetilde{\varepsilon})} |d^1_x f(v)|^2 \, dv\, d\mu(x).$$

From (2.7) of Burago et al. [2014], we further have

$$\int_{\mathcal{X}}\int_{B_m(0,\widetilde{\varepsilon})} |d^1_x f(v)|^2 \, dv\, d\mu(x) = \frac{\nu_m}{2+m}(\widetilde{\varepsilon}^{2+m} - \varepsilon^{2+m}) \int_{\mathcal{X}} |d^1_x f|^2 \, d\mu(x) = 27\frac{\nu_m}{(2+m)R^2}\varepsilon^{4+m}\|d^1 f\|^2_{L^2(\mathcal{X})}.$$

Recalling that $\|d^1 f\|^2_{L^2(\mathcal{X})} \leq \|f\|^2_{H^1(\mathcal{X})}$, we see that this implies the claim of Lemma 54.

Proof (of Lemma 55). The proof of Lemma 55 is identical to the proof of Lemma 51, upon substituting the ambient dimension m for the intrinsic dimension d, and using Lemma 53 rather than Lemma 50 to establish (C.45).

C.5 Lower bound on empirical norm

In this Section we prove Proposition 10 (in Section C.5.1). We also prove an analogous result when $\mathcal{X}$ is a manifold as in Model 4.2.2 (in Section C.5.2).

C.5.1 Proof of Proposition 10

In this section we establish Proposition 10. As mentioned, the proof of this Proposition follows from the Gagliardo-Nirenberg interpolation inequality, and a one-sided Bernstein's inequality (Lemma 60).

Lemma 56 (Gagliardo-Nirenberg inequality)**.** *Suppose Model 4.2.1, and that $f \in H^s(\mathcal{X})$ for some $s \geq d/4$. Then there exist constants C_1 and C_2 that do not depend on f, such that*

$$\|f\|_{L^4(\mathcal{X})} \leq C_1 |f|^{d/4s}_{H^s(\mathcal{X})} \|f\|^{1-d/(4s)}_{L^2(\mathcal{X})} + C_2\|f\|_{L^2(\mathcal{X})} \tag{C.57}$$

Proof (of Proposition 10). Rearranging (C.57) and raising both sides to the 4th power, we see that

$$\frac{\mathbb{E}[f^4(X)]}{\|f\|_P^4} \leq C\left(\frac{\|f\|_{L^4(\mathcal{X})}}{\|f\|_{L^2(\mathcal{X})}}\right)^4 \leq C_1\left(\frac{|f|_{H^s(\mathcal{X})}}{\|f\|_{L^2(\mathcal{X})}}\right)^{d/s} + C_2,$$

here the constants C_1, C_2 are not the same as in (C.57). Therefore taking the constant C in assumption (4.30) to be sufficiently large relative to C_1 and C_2, we have that

$$C_1\left(\frac{|f|_{H^s(\mathcal{X})}}{\|f\|_{L^2(\mathcal{X})}}\right)^{d/s} \leq \frac{\delta n}{64},$$

and consequently

$$\frac{\mathbb{E}[f^4(X)]}{\|f\|_P^4} \leq \frac{\delta n}{8} + 8C_2^3.$$

The claim then follows from Lemma 60, upon taking $c = 1/(64C_2^3)$ in the statement of Proposition 10.

C.5.2 Proof of Proposition 13

The proof of Proposition 13 follows exactly the same steps as the proof of Proposition 10, upon replacing Lemma 56 by Lemma 57.

Lemma 57 ((c.f Theorem 3.70 of Aubin [2012])). *Suppose Model 4.2.2, and that $f \in H^s(\mathcal{X})$ for some $s \geq m/4$. Then there exist constants C_1 and C_2 that do not depend on f, such that*

$$\|f\|_{L^4(\mathcal{X})} \leq C_1|f|_{H^s(\mathcal{X})}^{m/4s}\|f\|_{L^2(\mathcal{X})}^{1-m/(4s)} + C_2\|f\|_{L^2(\mathcal{X})}. \tag{C.58}$$

C.6 Proof of Main Results

C.6.1 Estimation Results

Proof of Theorem 12. We condition on the event that the design points $X_1, \ldots, X_n$ satisfy

$$\langle L_{n,\varepsilon} f_0, f_0\rangle_n \leq \frac{C}{\delta}M^2 \quad \text{and} \quad \lambda_k \geq \min\{\lambda_k(\Delta_P), \varepsilon^{-2}\} \ \text{ for all } 2 \leq k \leq n. \tag{C.59}$$

Note that by Propositions 7 and 9, these statements are both satisfied with probability at least $1 - \delta - Cn\exp\{-cn\varepsilon^d\}$.

Conditional on (C.59), we have from Lemma 41 that for any $0 \leq K \leq n$,

$$\|\widehat{f} - f_0\|_n^2 \leq C\left\{\frac{M^2}{\delta(\lambda_{K+1}(\Delta_P) \wedge \varepsilon^{-2})} + \frac{K}{n}\right\},$$

either deterministically (when $K = 0$), or with probability at least $1 - \exp(-K)$ (when $K \geq 1$). Further, from the bounds $\varepsilon \leq c_0 K^{-1/d}$ (Assumption (P1)) and $\lambda_{K+1}(\Delta_P) \geq c(K+1)^{2/d}$ (Weyl's Law) we can simply the above expression to the following,

$$\|\widehat{f} - f_0\|_n^2 \leq C\left\{\frac{M^2}{\delta}(K+1)^{-2/d} + \frac{K}{n}\right\}. \tag{C.60}$$

We now upper bound the right hand side of (C.60), based on the value of K chosen in (P1). When possible we choose $K = \lfloor M^2 n\rfloor^{d/(2+d)}$ to balance bias and variance, in which case (C.60) implies

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C}{\delta}M^2(M^2 n)^{-2/(2+d)}.$$

If $M^2 < n^{-1}$, then we take $K = 1$, and from (C.60) we get

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{C}{n\delta}.$$

Finally if $M > n^{1/d}$, we take $K = n$. In this case, we note that $\widehat{f}(X_i) = Y_i$ for all $i = 1, \ldots, n$, and it immediately follows that

$$\|\widehat{f} - f_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n w_i^2 \leq 5,$$

with probability at least $1 - \exp(-n)$. Combining these three separate cases yields the conclusion of Theorem 12.

Proof of Theorem 14. Follows identically to the proof of Theorem 12, except substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 8 rather than Proposition 7 and Assumption (P3) rather than Assumption (P1).

Proof of Theorem 17. Follows identically to the proof of Theorem 12, substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 11 rather than Proposition 7, Proposition 12 rather than Proposition 9, and Assumption (P6) rather than Assumption (P2).

C.6.2 Testing Results

Proof of Theorem 13. We have already upper bounded the Type I error of φ in Lemma 42, and it remains to upper bound the Type II error. To do so, we condition on the event that the design points $X_1, \ldots, X_n$ satisfy,

$$\langle L_{n,\varepsilon} f_0, f_0 \rangle_n \leq \frac{C}{\delta} M^2, \quad \text{and} \quad \lambda_k \geq \min\{\lambda_k(\Delta_P), \varepsilon^{-2}\} \ \text{ for all } 2 \leq k \leq n, \tag{C.61}$$

as well as that

$$\|f_0\|_n^2 \geq \frac{1}{2}\|f_0\|_P^2. \tag{C.62}$$

Note that by Propositions 7 and 9, both statements in (C.61) are satisfied with probability at least $1 - \delta - Cn\exp\{-cn\varepsilon^d\}$. Additionally, by Proposition 10 and the assumption in (4.18) that $\|f_0\|_P^2 \geq CM^2/(bn^{2/d})$, the one-sided inequality (C.62) follows with probability at least $1 - \exp\{-(cn \wedge 1/b)\}$. Setting $\delta = b/3$ and taking $n \geq N$ to be sufficiently large, the bottom line is that both (C.61) and (C.62) are together satisfied with probability at least $1 - b/2$.

Now, to complete the proof of Theorem 13, we would like to invoke Lemma 42, and conclude that conditional on $X_1, \ldots, X_n$ satisfying (C.61) and (C.62), our test φ will equal 1 with probability at least $1 - b/2$. To use Lemma 42, we will need to establish that (C.4) is satisfied, which we now show.

On the one hand, we have that the right hand side of (C.4) is upper bounded,

$$\begin{aligned}\frac{\langle L_{n,\varepsilon} f_0, f_0 \rangle_n}{\lambda_{K+1}} + \frac{\sqrt{2K}}{n}\left[2\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right] &\leq C\left(\frac{M^2}{b\min\{\lambda_{K+1}(\Delta_P), \varepsilon^{-2}\}} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right) \\ &\leq C\left(\frac{M^2}{b}K^{-2/d} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right)\end{aligned}$$

with the second inequality following by the assumption $\varepsilon \leq K^{-1/d}$ and Weyl's Law. On the other hand, we have that $\|f_0\|_n^2 \geq \|f_0\|_P^2/2$. Consequently, to prove Theorem 13, it remains only to verify that

$$\|f_0\|_P^2 \geq C\left(\frac{M^2}{b}K^{-2/d} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right). \tag{C.63}$$

As in the estimation case, we can further upper bound the right hand side of (C.63), depending on the value of K chosen in (P2). The classical case is $K = (M^2 n)^{d/(2+d)}$, in which case (C.63) is satisfied as long as

$$\|f_0\|_P^2 \geq C M^2 (M^2 n)^{-4/(4+d)} \left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]$$

If $M^2 < n^{-1}$, then we take $K = 1$, and (C.63) is satisfied whenever

$$\|f_0\|_P^2 \geq \frac{C}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right].$$

Finally if $M > n^{1/d}$, we take $K = n$, and (C.63) is satisfied if

$$\|f_0\|_P^2 \geq C\left(\frac{M^2}{n^{2/d} b} + n^{-1/2}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right).$$

We conclude by observing that (4.18) implies each of these three inequalities, and thus implies (C.63).

Proof of Theorem 15. Follows identically to the proof of Theorem 12, except substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 8 rather than Proposition 7 and Assumption (P4) rather than Assumption (P2).

Proof of Theorem 18. Follows identically to the proof of Theorem 12, except substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 11 rather than Proposition 7, Proposition 12 rather than Proposition 9, Proposition 13 rather than Proposition 10, and Assumption (P6) rather than Assumption (P2).

Proof of Theorem 16. Note that our choices of K and ε ensure that (C.61) (with $L_{n,\varepsilon}^s$ replacing $L_{n,\varepsilon}$) and (C.62) are satisfied with probability at least $1 - b/2$. Proceeding as in the proof of Theorem 13, we upper bound the right hand side of (C.4),

$$\begin{aligned}\frac{\langle L_{n,\varepsilon} f_0, f_0\rangle_n}{\lambda_{K+1}} + \frac{\sqrt{2K}}{n}\left[2\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right] &\leq C\left(\frac{M^2}{b \min\{\lambda_{K+1}(\Delta_P), \varepsilon^{-2}\}} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right) \\ &\leq C\left(\frac{M^2}{b}\varepsilon^2 + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right).\end{aligned}$$

Unlike in the proof of Theorem 13, we note that in this case $\varepsilon^2 \leq C\lambda_K(\Delta_P)$ rather than vice versa. From here, proceeding as in the proof of Theorem 13 gives the claimed result.

C.7 Analysis of kernel smoothing

In this section we prove Lemma 5 (in Section C.7.2) and Lemma 58 (in Section C.7.3). We begin with some preliminary estimates in Section C.7.1, which will ease the subsequent analysis.

C.7.1 Some preliminary estimates

In certain parts the analysis of this section will overlap with Section C.3, where we upper bounded the non-local graph-Sobolev seminorm of a function f in terms of the Sobolev norm of f. To see why this should be, note that for an function f and point $x \in \mathcal{X}$, we have

$$T_{P,h} f(x) - f(x) = \frac{1}{d_{Q,h}(x)} \int (f(x') - f(x)) \psi\left(\frac{\|x' - x\|}{h}\right) dQ(x') = \frac{h^{d+2}}{d_{P,h}(x)} L_{P,h} f(x).$$

This expression reflects the known fact that the bias operator of kernel smoothing is equal to the non-local Laplacian, up to a rescaling by the population degree functional $d_{P,h}(x)$. In the second equality, we are using the notation $L_{P,h}f(x)$ exactly as defined in (4.28), but with the kernel ψ instead of η. Note that ψ satisfies all the same assumptions as η, except that of positivity; when ψ is a higher-order kernel it may take negative values.

Now we provide a lower bound on $d_{P,h}(x)$ that holds uniformly over all $x \in \mathcal{X}$. Recall that by assumption the density p is Lipschitz. Letting L_p denote the Lipschitz constant of p, we have that

$$\begin{aligned} d_{P,h}(x) &= \int \psi\bigg(\frac{\|x'-x\|}{h}\bigg) p(x')\,dx' \\ &= h^d \int \psi(\|z\|) p(hz+x) \mathbf{1}\Big\{hz+x \in \mathcal{X}\Big\}\,dz \\ &\geq h^d p(x) \int \psi(\|z\|) \mathbf{1}\Big\{hz+x \in \mathcal{X}\Big\}\,dz - L_p h^{d+1} \|\psi\|_\infty \nu_d. \end{aligned}$$

Since by assumption $\mathcal{X}$ has Lipschitz boundary, setting c_0 to be a sufficiently small constant in (P7), we can further deduce that $\int \psi(\|z\|)\mathbf{1}\{hz+x \in \mathcal{X}\}\,dz \geq 1/3$, and consequently that

$$d_{P,h}(x) \geq \frac{p(x)}{3} h^d \geq \frac{p_{\min}}{3} h^d \quad \text{for all } x \in \mathcal{X}. \tag{C.64}$$

C.7.2 Proof of Lemma 5

To begin with, we apply the triangle inequality to upper bound $\|T_{n,h}\check{f} - f_0\|_P$ by the sum of two terms,

$$\|T_{n,h}\check{f} - f_0\|_P \leq \|T_{n,h}(\check{f} - f_0)\|_P + \|T_{n,h}f_0 - f_0\|_P. \tag{C.65}$$

We proceed by separately upper bounding each term on the right hand side of (C.65). We will show that

$$\|T_{n,h}(\check{f} - f_0)\|_P^2 \leq C\|\check{f} - f_0\|_n^2 \tag{C.66}$$

and that

$$\|T_{n,h}f_0 - f_0\|_P^2 \leq \frac{C}{\delta} \cdot \frac{h^2}{nh^d} |f|_{H^1(\mathcal{X})}^2 + \frac{C}{\delta}\|T_{h,P}f_0 - f_0\|_P^2, \tag{C.67}$$

each with probability at least $1 - C\exp(-cnh^d)$. Together these will imply the claim.

Proof of (C.66). Fix $x \in \mathcal{X}$. By the Cauchy-Schwarz inequality we have

$$\begin{aligned} \Big[T_{n,h}(\check{f} - f_0)(x)\Big]^2 &= \bigg[\frac{1}{d_{n,h}(x)^2} \int \psi\bigg(\frac{\|x'-x\|}{h}\bigg) \cdot \big(\check{f}(x') - f_0(x')\big)\,dP_n(x')\bigg]^2 \\ &\leq \bigg[\frac{1}{d_{n,h(x)}^2} \int \bigg|\psi\bigg(\frac{\|x'-x\|}{h}\bigg)\bigg|\,dP_n(x')\bigg] \cdot \bigg[\int \bigg|\psi\bigg(\frac{\|x'-x\|}{h}\bigg)\bigg| \cdot \big(\check{f}(x') - f_0(x')\big)^2\,dP_n(x')\bigg] \\ &= \frac{d_{n,h}^+(x)}{|d_{n,h}(x)|} \cdot \frac{1}{|d_{n,h}(x)|} \bigg[\int \bigg|\psi\bigg(\frac{\|x'-x\|}{h}\bigg)\bigg| \cdot \big(\check{f}(x') - f_0(x')\big)^2\,dP_n(x')\bigg]. \end{aligned}$$

In the last line all we have done is written $d_{n,h}^+(x)$ for the degree functional computed with respect to the kernel $|\psi|$, recalling that ψ may take negative values so $d_{n,h}^+(x)$ may not be equal to $d_{n,h}(x)$.

Now we integrate over $x \in \mathcal{X}$ to get

$$\begin{aligned}
\|T_{n,h}(\check{f} - f_0)\|_P^2 &= \int \Big[T_{n,h}\big(\check{f} - f_0\big)(x)\Big]^2 \, dP(x) \\
&\le \int\int \frac{d_{n,h}^+(x)}{|d_{n,h}(x)|} \cdot \frac{1}{|d_{n,h}(x)|} \left|\psi\Big(\frac{\|x' - x\|}{h}\Big)\right| \cdot \big(\check{f}(x') - f_0(x')\big)^2 \, dP_n(x') \, dP(x) \\
&\le \sup_{x \in \mathcal{X}} \frac{d_{n,h}^+(x)}{|d_{n,h}(x)|} \cdot \int\int \frac{1}{|d_{n,h}(x)|} \left|\psi\Big(\frac{\|x' - x\|}{h}\Big)\right| \cdot \big(\check{f}(x') - f_0(x')\big)^2 \, dP(x) \, dP_n(x') \\
&\le \sup_{x \in \mathcal{X}} \frac{d_{n,h}^+(x) d_{P,h}^+(x)}{|d_{n,h}(x)|^2} \cdot \|\check{f} - f_0\|_n^2. \qquad \text{(C.68)}
\end{aligned}$$

Thus we have reduced the problem to showing that the various degree functionals $d_{n,h}^+, d_{P,h}^+$ and $d_{n,h}$ all put similar weight on a given point x. We use (C.73), which gives a uniform multiplicative bound on deviations of the empirical degree around its mean, to conclude that with probability at least $1 - C\exp\{-cnh^d\}$,

$$d_{n,h}(x) \ge \frac{1}{2} d_{P,h}(x) \quad \text{and} \quad d_{n,h}^+(x) \le \frac{3}{2} d_{P,h}^+(x) \quad \text{for all } x \in \mathcal{X}.$$

Therefore,

$$\sup_{x \in \mathcal{X}} \frac{d_{n,h}^+(x) d_{P,h}^+(x)}{|d_{n,h}(x)|^2} \le 6 \cdot \sup_{x \in \mathcal{X}} \frac{|d_{P,h}^+(x)|^2}{|d_{P,h}(x)|^2} \le 36 \left(\frac{\|\psi\|_\infty p_{\max} \nu_d}{p_{\min}}\right)^2$$

with the second inequality following from (C.64). Plugging this back into (C.68) gives the claim.

Proof of (C.67). At a given point $x \in \mathcal{X}$, we have

$$\begin{aligned}
T_{n,h} f_0(x) - f_0(x) &= \frac{1}{d_{n,h}(x)} \sum_{i=1}^n \big(f_0(X_i) - f_0(x)\big) \psi\Big(\frac{\|X_i - x\|}{h}\Big) \\
&= \frac{d_{P,h}(x)}{d_{n,h}(x)} \cdot \frac{1}{n d_{P,h}(x)} \sum_{i=1}^n \big(f_0(X_i) - f_0(x)\big) \psi\Big(\frac{\|X_i - x\|}{h}\Big).
\end{aligned}$$

Thus,

$$\Big[T_{n,h} f_0(x) - f_0(x)\Big]^2 = \left[\frac{d_{P,h}(x)}{d_{n,h}(x)}\right]^2 \cdot \underbrace{\left[\frac{1}{n d_{P,h}(x)} \sum_{i=1}^n \big(f_0(X_i) - f_0(x)\big) \psi\Big(\frac{\|X_i - x\|}{h}\Big)\right]^2}_{:= \widetilde{L}_{n,h} f_0(x)} \qquad \text{(C.69)}$$

In the proof of (C.66) we have already given an upper bound on the ratio of population to empirical degree, which implies that

$$\sup_{x \in \mathcal{X}} \left[\frac{d_{P,h}(x)}{d_{n,h}(x)}\right]^2 \le 4,$$

with probability at least $1 - C\exp\{-cnh^d\}$. On the other hand, we note that the second term in the product in (C.69) has expectation

$$\mathbb{E}\Big[\widetilde{L}_{n,h} f_0(x)\Big] = T_{P,h} f_0(x) - f_0(x),$$

and variance

$$\mathrm{Var}\Big[\widetilde{L}_{n,h} f_0(x)\Big] \le \frac{1}{n(d_{P,h}(x))^2} \mathbb{E}\left[(f_0(X) - f_0(x))^2 \cdot \left|\psi\Big(\frac{\|X - x\|}{h}\Big)\right|^2\right].$$

Integrating with respect to P gives

$$\begin{aligned}\mathbb{E}\bigg[\int \Big(\widetilde{L}_{n,h} f_0(x)\Big)^2 \, dP(x)\bigg] &= \int \mathbb{E}\Big[\Big(\widetilde{L}_{n,h} f_0(x)\Big)^2\Big] \, dP(x) \\ &\leq \|T_{P,h} f_0 - f_0\|_P^2 + \frac{1}{n} \int\int \frac{1}{\big(d_{P,h}(x)\big)^2} \big(f_0(x') - f_0(x)\big)^2 \cdot \bigg|\psi\bigg(\frac{\|x' - x\|}{h}\bigg)\bigg|^2 \, dP(x') \, dP(x) \\ &\leq \|T_{P,h} f_0 - f_0\|_P^2 + \frac{3h^2}{p_{\min} n} \widetilde{E}_{P,h}(f_0; \psi^2).\end{aligned}$$

In the final inequality we have used the lower bound on $d_{P,h}(x)$ from (C.64), and written $E_{P,h}(f_0; \psi^2)$ for the non-local Dirichlet energy defined with respect to the kernel ψ^2.

Putting the pieces together, we conclude that

$$\begin{aligned}\|T_{n,h} f_0(x) - f_0(x)\|_P^2 &= \int \big(T_{n,h} f_0(x) - f_0(x)\big)^2 \, dP(x) \\ &\leq \sup_{x \in \mathcal{X}} \bigg[\frac{d_{P,h}(x)}{d_{n,h}(x)}\bigg]^2 \cdot \int \Big(\widetilde{L}_{n,h} f_0(x)\Big)^2 \, dP(x) \\ &\overset{(i)}{\leq} 4 \frac{\|T_{P,h} f_0 - f_0\|_P^2}{\delta} + \frac{12h^2}{\delta p_{\min} n h^d} E_{P,h}(f_0; \psi^2) \\ &\overset{(ii)}{\leq} 4 \frac{\|T_{P,h} f_0 - f_0\|_P^2}{\delta} + \frac{Ch^2}{\delta p_{\min} n h^d} |f_0|_{H^1(\mathcal{X})}^2,\end{aligned}$$

with probability at least $1 - \delta - C\exp(-cnh^d)$. In (i) we have used Markov's inequality, and in (ii) we have applied the estimate (C.37) to the non-local Dirichlet energy $E_{P,h}(f_0; \psi^2)$. This establishes (C.67).

C.7.3 Kernel smoothing bias

Lemma 58 gives the necessary upper bounds on the bias of kernel smoothing.

Lemma 58. *Suppose Model 4.2.1, and that the kernel smoothing operator $T_{P,h}$ is computed with respect to a kernel η that satisfies (K1).*

- *If $f_0 \in H^1(\mathcal{X})$, then there exists a constant C which does not depend on f_0 such that*

$$\|T_{P,h} f_0 - f_0\|_P^2 \leq Ch^2 |f|_{H^1(\mathcal{X})}^2.$$

- *If $f_0 \in H_0^s(\mathcal{X})$, $p \in C^{s-1}(\mathcal{X})$, and η satisfies (K4), then there exists a constant C which does not depend on f_0 such that*

$$\|T_{P,h} f_0 - f_0\|_P^2 \leq Ch^{2s} |f|_{H^s(\mathcal{X})}^2.$$

We separately prove the first-order ($s = 1$) and higher-order ($s > 1$) parts of Lemma 58. In both cases, the proof will rely heavily on results already established regarding the non-local Laplacian $L_{P,h}$ and non-local Dirichlet energy $E_{P,h}$, which we recall are given for a kernel function $\mathcal{K}$ by

$$L_{P,h} f(x) = \frac{1}{h^{d+2}} \int \big(f(x') - f(x)\big) \mathcal{K}\bigg(\frac{\|x' - x\|}{h}\bigg) \, dP(x'),$$

and $E_{P,h}(f; \mathcal{K}) = \langle L_{P,h} f, f \rangle_P$, respectively.

Proof of Lemma 58, $s = 1$. Using the conclusions from Section C.7.1, we have that

$$\|T_{P,h}f - f\|_P^2 \le \frac{9h^4}{p_{\min}^2} \int \big[L_{P,h}f(x)\big]^2 \, dP(x). \tag{C.70}$$

By the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
\int \big[L_{P,h}f(x)\big]^2 \, dP(x) &= \frac{1}{h^{2d+4}} \int \bigg[\int \big(f(x') - f(x)\big)\psi\Big(\frac{\|x' - x\|}{h}\Big) \, dP(x')\bigg]^2 \, dP(x) \\
&\le \frac{C}{h^{d+4}} \int\int \big(f(x') - f(x)\big)^2 \cdot \bigg|\psi\Big(\frac{\|x' - x\|}{h}\Big)\bigg| \, dP(x') \, dP(x) \\
&= \frac{C}{h^2} E_{P,h}(f; |\psi|).
\end{aligned}$$

Applying the estimate (C.37) to the non-local Dirichlet energy $E_{P,h}(f; |\psi|)$ and plugging back into (C.70) gives the claimed result.

Proof of Lemma 58, $s > 1$. Proceeding from (C.70), we separate the integral into the portion sufficiently in the interior of $\mathcal{X}$ and that near the boundary, obtaining

$$\|T_{P,h}f - f\|_P^2 \le \frac{9p_{\max}h^4}{p_{\min}^2}\Big(\|L_{P,h}f\|_{L^2(\mathcal{X}_h)}^2 + \|L_{P,h}f\|_{L^2(\partial_h(\mathcal{X}))}^2\Big). \tag{C.71}$$

In Lemma 48, we established a sufficient upper bound on the second term,

$$\|L_{P,h}f\|_{L^2(\partial_h(\mathcal{X}))}^2 \le Ch^{2(s-2)}\|f\|_{H^s(\mathcal{X})}^2.$$

Thus it remains to upper bound the first term. Here we recall that at a given $x \in \mathcal{X}_h$, we can write

$$\begin{aligned}
L_{P,h}f(x) = \frac{1}{h^2} \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{q-1} \frac{h^{j_1+j_2}}{j_1!j_2!} \int d_x^{j_1} f(z) d_x^{j_2} p(z) \psi\big(\|z\|\big) \, dz \quad + \\
\frac{1}{h^2} \sum_{j=1}^{s-1} \frac{h^j}{j!} \int d_x^j f(z) r_{zh+x}^q(x; p) \psi\big(\|z\|\big) \, dz \quad + \\
\frac{1}{h^2} \int r_{zh+x}^j(x; f) \psi\big(\|z\|\big) p(zh + x) \, dz \\
= G_1(x) + G_2(x) + G_3(x).
\end{aligned}$$

(Here $q = s - 1$.)

We have already given sufficient upper bounds on $\|G_j\|_{L^2(\mathcal{X}_h)}$ for $j = 2, 3$ in (C.22). Thus it remains only to upper bound $\|G_1\|_{L^2(\mathcal{X}_h)}$. Recall the expansion of G_1 from (C.23),

$$G_1(x) = \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{q-1} \frac{h^{j_1+j_2-2}}{j_1!j_2!} \underbrace{\int_{B(0,1)} d_x^{j_1} f(z) d_x^{j_2} p(z) \eta(\|z\|) \, dz}_{:= g_{j_1,j_2}(x)}.$$

Noting that $d_x^{j_1} \cdot d_x^{j_2}$ is a degree-$(j_1 + j_2)$ multivariate polynomial, and recalling that ψ is an order-s kernel, we have that

$$\int g_{j_1,j_2}(z)\psi\big(\|z\|\big) \, dz = 0, \quad \text{for all } j_1, j_2 \text{ such that } j_1 + j_2 < s.$$

Otherwise, derivations similar to those used in the proof of Lemma 47 imply that

$$\|g_{j_1,j_2}\|_{L^2(\mathcal{X}_h)} \leq C\|f\|_{H^s(\mathcal{X})}\|p\|_{C^{s-1}(\mathcal{X})}, \quad \text{for all } j_1, j_2 \text{ such that } j_1 + j_2 \geq s,$$

from which it follows that

$$\|G_1\|^2_{L^2(\mathcal{X}_h)} \leq Ch^{2(s-2)}\|f\|_{H^s(\mathcal{X})}\|p\|_{C^{s-1}(\mathcal{X})}.$$

Together these upper bounds on $\|G_j\|_{L^2(\mathcal{X}_j)}$ for $j = 1, 2, 3$ imply that

$$\|L_{P,h}f\|^2_{\mathcal{X}_h} \leq Ch^{2(s-2)}\|f\|^2_{H^s(\mathcal{X})},$$

and plugging this back into (C.71) yields the claim.

C.8 Miscellaneous

Here we give assorted helpful Lemmas used at various points in the above proofs. We also review notation and relevant facts regarding Taylor expansion.

C.8.1 Concentration Inequalities

Lemma 59 controls the deviation of a chi-squared random variable. It is from Laurent and Massart [2000].

Lemma 59. *Let $\xi_1, \ldots, \xi_N$ be independent $N(0,1)$ random variables, and let $U := \sum_{k=1}^N a_k(\xi_k^2 - 1)$. Then for any $t > 0$,*

$$\mathbb{P}\Big[U \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\Big] \leq \exp(-t).$$

In particular if $a_k = 1$ for each $k = 1, \ldots, N$, then

$$\mathbb{P}\Big[U \geq 2\sqrt{Nt} + 2t\Big] \leq \exp(-t).$$

Lemma 60 is an immediate consequence of the one-sided Bernstein's inequality (14.23) in Wainwright [2019].

Lemma 60 (One-sided Bernstein's inequality)**.** *Let $X, X_1, \ldots, X_n \sim P$, and f satisfy $\mathbb{E}[f^4(X)] < \infty$. Then*

$$\|f\|_n^2 \geq \frac{1}{2}\|f\|_P^2,$$

with probability at least $1 - \exp\big(-n/8 \cdot \|f\|_P^4/\mathbb{E}[f^4(X)]\big)$.

In the proof of Lemma 5, we require uniform control of the empirical degree functional $d_{n,h}(x)$ over all $x \in \mathcal{X}$. Such a result is available to us because the kernel ψ is Lipschitz on its support, so that the class of functions $\{\psi((x - \cdot)/h) : h \in \mathbb{R}^d\}$ has finite VC dimension. The precise estimate we use is due to Giné and Guillou [2002].

Lemma 61 (Uniform bound for empirical degree.)**.** *Suppose Model 4.2.1. For a kernel ψ satisfying (K1) and bandwidth h satisfying (P7), there exist constants c, C, c_1 and C_1 which do not depend on h or n such that*

$$\mathbb{P}\bigg(\sup_{x\in\mathbb{R}^d} n \cdot \Big|d_{n,h}(x) - d_{P,h}(x)\Big| > t\bigg) \leq C\exp\bigg(-c\frac{t^2}{nh^d}\bigg),$$

for any $t \in \mathbb{R}$ satisfying

$$C_1\sqrt{nh^d\log(1/h)} \leq t \leq c_1 nh^d. \tag{C.72}$$

Now we translate Lemma 61 into a multiplicative bound, which will be more useful for our purposes. Recall the lower bound on $d_{P,h}(x)$ given in (C.64). By setting C_0 to be a sufficiently large constant in (P7), we can ensure that choosing $t = nh^d p_{\min}/6$ satisfies both the inequalities (C.72). For this choice of t it follows from (C.64) and Lemma 61 that

$$\sup_{x\in\mathbb{R}^d}\left|\frac{d_{n,h}(x)-d_{P,h}(x)}{d_{P,h}(x)}\right| \le \sup_{x\in\mathbb{R}^d}\left|\frac{d_{n,h}(x)-d_{P,h}(x)}{2t/n}\right| \le \frac{1}{2} \tag{C.73}$$

with probability at least $1-C\exp(-cnh^d)$. This is the form of the result we use in the proof of Lemma 5.

C.8.2 Taylor expansion

We begin with some notation that allows us to concisely expresss derivatives. For a given $z \in \mathbb{R}^d$ and s-times differentiable function $f : \mathcal{X} \to \mathbb{R}$, we denote $\big(d_x^s f\big)(z) := \sum_{|\alpha|=s} D^\alpha f(x) z^\alpha$. We also write $d^s f := \sum_{|\alpha|=j} D^\alpha f$. We point out that in the first-order case $d_x^1 f$ is the differential of f at $x \in \mathcal{X}$, while $d^1 f$ is the divergence of f.

Let u be a function which is s times continuously differentiable at all $x \in \mathcal{X}$, for $k \in \mathbb{N}\setminus\{0\}$. Suppose that for some $h > 0$, $x \in \mathcal{X}_h$ and $x' \in B(x,h)$. We write the order-s Taylor expansion of $u(x')$ around $x' = x$ as

$$u(x') = u(x) + \sum_{j=1}^{s-1}\frac{1}{j!}\big(d_x^j u\big)(x'-x) + r_{x'}^s(x;u)$$

For notational convenience we have adopted the convention that $\sum_{j=1}^{0} a_j = 0$. Thus $\big(d_x^j f\big)(z)$ is a degree-j polynomial—and so a j-homogeneous function—in z, meaning for any $t \in \mathbb{R}$,

$$\big(d_x^j f\big)(tz) = t^j \cdot \big(d_x^j f\big)(z).$$

The remainder term $r_{x'}$ is given by

$$r_{x'}^s(x;f) = \frac{1}{(j-1)!}\int_0^1 (1-t)^{j-1}\big(d_{x+t(x'-x)}^s f\big)(x'-x)\,dt,$$

where we point out that the integral makes sense because $x + t(x'-x) \in B(x,h) \subseteq \mathcal{X}$. We now give estimates on the remainder term in both sup-norm and $L^2(\mathcal{X}_h)$ norm, each of which hold for any $z \in B(0,1)$. In sup-norm, we have that

$$\sup_{x\in\mathcal{X}_h}|r_{x+hz}^j(x;f)| \le Ch^j\|f\|_{C^j(\mathcal{X})},$$

whereas in $L^2(\mathcal{X}_h)$ norm we have,

$$\int_{\mathcal{X}_h}\big|r_{x+thz}^j(x;f)\big|^2\,dx \le h^{2j}\int_{\mathcal{X}_h}\int_0^1 |d_{x+thz}^j f(z)|^2\,dt\,dx \le h^{2j}\|d^j f\|_{L^2(\mathcal{X})}^2. \tag{C.74}$$

In the last inequality

Finally, we recall some facts regarding the interaction between smoothing kernels and polynomials. Let $q_j(z)$ be an arbitrary degree-j (multivariate) polynomial. If η is a radially symmetric kernel and j is odd, then by symmetry it follows that

$$\int_{B(0,1)} q_j(z)\eta(\|z\|)\,dz = 0.$$

On the other hand, if ψ is an order-s kernel for some $s > j$, then by converting to polar coordinates we can verify that

$$\int_{B(0,1)} q_j(z)\eta(\|z\|)\,dz = 0.$$

Appendix D

Chapter 5 Appendix

D.1 Proof of Lemma 6

Let $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|X_1, \dots, X_n]$ denote the expectation operator conditional on $X_1, \dots, X_n$. First we prove the desired upper bound on $\mathbb{E}_n[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n]$, and the lower bound on $\mathbb{E}_n[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}}\rangle_n]$.

Upper bounds. We beginn by giving an upper bound on the expected graph Sobolev semi-norm $\mathbb{E}_n[\langle L^s_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n]$, in terms of functionals of the graph $G_{n,\varepsilon}$. This bound will hold for any $s \in \mathbb{N}$, and the choice $s = 1$ will correspond to $\mathbb{E}_n[\langle L_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n]$. Then we will give estimates on these graph functionals, which will hold with high probability, and will imply (5.5).

Recall that $\mathbf{Y} = f_0 + \mathbf{w}$ and the spectral decomposition $L_{n,\varepsilon} = \sum_{k=1}^n \lambda_k v_k v_k^\top$. From these decompositions, we can rewrite the graph Sobolev seminorm of $\widehat{f}_{\mathrm{LE}}$ in terms of that of f_0 and $\mathbf{w}$:

$$\langle L^s_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n = \sum_{k=1}^K \lambda_k \langle v_k, \mathbf{Y}\rangle_n^2 = \sum_{k=1}^K \lambda_k^s \Big(\langle v_k, f_0\rangle_n^2 + \langle v_k, \mathbf{w}\rangle_n^2 + 2\langle v_k, f_0\rangle_n \langle v_k, \mathbf{w}\rangle_n \Big)$$

Conditional on $X_1, \dots, X_n$, the graph Sobolev seminorm of the signal f_0 is deterministic, and satisfies the upper bound $\sum_{k=1}^n \lambda_k^s \langle v_k, f_0\rangle_n^2 \le \langle L^s_{n,\varepsilon} f_0, f_0\rangle_n^2$. On the other hand, the graph seminorm of the noise $\mathbf{w}$ is random, with expectation given by

$$\mathbb{E}_n\Big[\sum_{k=1}^K \lambda_k^s \langle v_k, \mathbf{w}\rangle_n^2\Big] = \sum_{k=1}^K \lambda_k^s \cdot \mathbb{E}_n\big[\langle v_k, \mathbf{w}\rangle_n^2\big] = \frac{1}{n}\sum_{k=1}^K \lambda_k^s \le \frac{K}{n}\lambda_K^s.$$

Finally, since the noise variables w_i are zero-mean, $\mathbb{E}[\langle v_k, \mathbf{w}\rangle_n] = 0$. Thus, as promised, we can upper bound the conditional expectation of graph semi-norm of $\widehat{f}_{\mathrm{LE}}$ in terms of graph functionals,

$$\langle L^s_{n,\varepsilon}\widehat{f}_{\mathrm{LE}}, \widehat{f}_{\mathrm{LE}}\rangle_n \le \Big(\langle L^s_{n,\varepsilon} f_0, f_0\rangle_n + \frac{K}{n}\lambda_K^s \Big). \tag{D.1}$$

It remains to upper bound these graph functionals. To do so, we recall some estimates from Chapter 3:[1]

[1] Note that in the setup of Chapter 5, the Laplacian $L_{n,\varepsilon}$ is defined with a pre-factor of $(n\epsilon^{d+2})^{-1}$, which is not the case in the setup of Chapter 3. Multiplying the estimates of Chapter 3 by this prefactor gives (D.2) and (D.3).

- *Lemma 3.* There exist constants c, C, N such that if $n \geq N$ and $\varepsilon < c$,

$$\langle L_{n,\varepsilon} f_0, f_0 \rangle_n \leq C \frac{|f_0|^2_{H^1(\mathcal{X})}}{\delta} \tag{D.2}$$

  with probability at least $1 - \delta$.

- *Lemma 4.* There exist constants $N, C, c > 0$ such that if $n \geq N$ and $C(\log n/n)^{1/d} \leq \varepsilon \leq c$, then

$$c \min\{k^{2/d}, \varepsilon^{-2}\} \leq \lambda_k \leq C \min\{k^{2/d}, \varepsilon^{-2}\}, \quad \text{for all } 2 \leq k \leq n, \tag{D.3}$$

  with probability at least $1 - Cn\exp(-cn\varepsilon^d)$.

Plugging (D.2)-(D.3) back into (D.1) gives (5.5).

Lower bounds. Again we bound $\mathbb{E}_n[\langle L^s_{n,\varepsilon} \widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}} \rangle_n]$ by graph functionals. Using similar derivations to the lower bound for Laplacian eigenmaps, we find that

$$\mathbb{E}_n[\langle L^s_{n,\varepsilon} \widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}} \rangle_n] = \sum_{k=1}^n \frac{\lambda_k^s}{(1+\rho\lambda_k)^2} \mathbb{E}_n[\langle \mathbf{Y}, v_k \rangle_n^2] = \sum_{k=1}^n \frac{\lambda_k^s}{(1+\rho\lambda_k)^2} \Big(\langle v_k, f_0 \rangle_n^2 + \mathbb{E}_n[\langle v_k, \mathbf{w} \rangle_n^2] \Big).$$

Recalling that $\mathbb{E}_n[\langle v_k, \mathbf{w} \rangle_n^2] = 1/n$, and noticing that clearly $\langle v_k, f_0 \rangle_n^2 \geq 0$, we have

$$\mathbb{E}_n[\langle L^s_{n,\varepsilon} \widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}} \rangle_n] \geq \frac{1}{n} \sum_{k=1}^n \frac{\lambda_k^s}{(1+\rho\lambda_k)^2} \geq \frac{1}{2} \min\Big\{ \lambda^s_{n/2}, \frac{\lambda^s_{n/2}}{\rho^2 \lambda_n^2} \Big\}. \tag{D.4}$$

Plugging the lower bounds on λ_k from (D.3) into (D.4), we conclude that with probability at least $1 - Cn\exp(-cn\varepsilon^d)$,

$$\mathbb{E}_n \Big[\langle L^s_{n,\varepsilon} \widehat{f}_{\mathrm{LS}}, \widehat{f}_{\mathrm{LS}} \rangle_n \Big] \geq c \min\Big\{ n^{2s/d} \wedge \varepsilon^{-2s}, \frac{1}{\rho^2} \big(n^{(2s-4)/d} \vee \varepsilon^{4-4s} \big) \Big\} \geq c\varepsilon^{-2s} \min\Big\{ 1, \frac{\varepsilon^4}{\rho^2} \Big\};$$

the last inequality follows because $C(\log n/n)^{1/d} \leq \varepsilon$. Taking $s = 1$ gives (5.6).

We remark that (D.4) is potentially loose, although it is tight enough to imply (5.6). The potential looseness is because we have ignored the contributions of $\langle f_0, v_k \rangle_n^2$ for $k = 1, \ldots, n$.

D.2 Laplacian Regularization Out-of-Sample

In this section we build to the proof of Theorem 21. We begin by giving a structural result for $\check{f}_{\mathrm{LS}}$ in Proposition 28, which represents $\check{f}_{\mathrm{LS}}$ as a particular two-stage estimator. The first stage returns an estimate at the design points $X_1, \ldots, X_{n-1}$, which can be cast as the solution obtained by smoothing with respect to the weighted average of two different Laplacian matrices. The second stage then extends this estimate to X_n by simple kernel smoothing, with the same kernel η and radius ε used to construct the graph $G_{n,\varepsilon}$.

To explicitly write the first-stage estimate, we must first define the two Laplacian matrices used in its construction. Let $G_{n-1,\varepsilon}$ be the neighborhood graph defined over $X_1, \ldots, X_{n-1}$, with corresponding Laplacian matrix $L_{n-1,\varepsilon}$ acting on vectors $u \in \mathbb{R}^{n-1}$ by

$$\big(L_{n-1,\varepsilon} u\big)_i := \frac{1}{(n-1)\varepsilon^{d+2}} \sum_{j=1}^{n-1} (u_i - u_j) \cdot \eta\Big(\frac{\|X_i - X_j\|}{\varepsilon} \Big)$$

We now introduce a second graph, $G^{(n)}_{n-1,\varepsilon} = (\{1,\dots,n-1\}, W^{(n)})$, which is defined over the same set of vertices as $G_{n-1,\varepsilon}$, but with a different adjacency matrix $W^{(n)} \in \mathbb{R}^{(n-1)\times(n-1)}$ defined by its entries

$$W^{(n)}_{ij} = \frac{1}{d_{n-1,\varepsilon}(X_n)} \eta\left(\frac{\|X_i - X_n\|}{\varepsilon}\right) \cdot \eta\left(\frac{\|X_j - X_n\|}{\varepsilon}\right) \tag{D.5}$$

where $d_{n-1,\varepsilon}(x) = \sum_{i=1}^{n-1} \eta(\|X_i - x\|/\varepsilon)$ is the degree functional. We see that points i and j are connected in $G^{(n)}_{n-1,\varepsilon}$ only if they both within distance ε of X_n. Let $L^{(n)}_{n-1,\varepsilon}$ be a Laplacian matrix associated with $G^{(n)}_{n-1,\varepsilon}$, defined by its action on vectors $u \in \mathbb{R}^{n-1}$ as

$$(L^{(n)}_{n-1,\varepsilon} u)_i := \frac{1}{\varepsilon^{d+2}} \sum_{j=1}^{n-1} (u_i - u_j) W^{(n)}_{ij}. \tag{D.6}$$

We observe that for sufficiently regular functions f—say $f \in C^1(\mathcal{X})$—we have that $\langle L^{(n)}_{(n-1),\varepsilon} f, f\rangle_{n-1} = O(1)$ as $n \to \infty$, confirming that the pre-factor of ε^{d+2} is the "right" scaling of $L^{(n)}_{(n-1),\varepsilon}$.

Proposition 28. *Recall $\check{f}_{\mathrm{LS}}$ is the solution to (5.7).*

- *At the design points $X_1,\dots,X_{n-1}$, the solution $\check{f}_{\mathrm{LS}}$ is given by*

$$(\check{f}_{\mathrm{LS}})_{1:(n-1)} := \big((\check{f}_{\mathrm{LS}})_1, \dots, (\check{f}_{\mathrm{LS}})_{n-1}\big) = \left(I_{n-1} + \frac{\rho(n-1)}{n}\left[\frac{n-1}{n} L_{n,\varepsilon} + \frac{1}{n} L^{(n)}_{n,\varepsilon}\right]\right)^{-1} \mathbf{Y}. \tag{D.7}$$

- *At the design point X_n, the solution $(\check{f}_{\mathrm{LS}})_n$ is given by*

$$(\check{f}_{\mathrm{LS}})_n = \big(T_{n,\varepsilon}(\check{f}_{\mathrm{LS}})_{1:(n-1)}\big)(X_n) := \frac{1}{d_{n-1,\varepsilon}(X_n)} \sum_{i=1}^{n-1} (\check{f}_{\mathrm{LS}})_i \cdot \eta(\|X_i - X_n\|/\varepsilon). \tag{D.8}$$

The proof of Proposition 28 is a matter of some standard linear algebra. We proceed to prove Theorem 21 given Proposition 28, and then return to prove Proposition 28.

D.2.1 Proof of Theorem 21

We begin by mapping out the high-level strategy we will follow to prove Theorem 21.

Strategy. We will use the decomposition in Proposition 28 to upper bound the squared loss $\big|(\check{f}_{\mathrm{LS}})_n - f_0(X_n)\big|^2$. Let S_n and S_{n-1} denote the smoother matrices such that $(\check{f}_{\mathrm{LS}})_{1:(n-1)} = \check{S}_{n-1}\mathbf{Y}$ and $\widehat{f}_{n-1} = S_{n-1}\mathbf{Y}$; explicitly

$$\check{S}_{n-1} := \left(I_{n-1} + \rho\left[\frac{(n-1)}{n} L_{n-1,\varepsilon} + \frac{1}{n} L^{(n)}_{n-1,\varepsilon}\right]\right)^{-1}, \quad \text{and} \quad S_{n-1} := \left(I_n + \rho L_{n-1,\varepsilon}\right)^{-1}.$$

Now we consider the following decomposition

$$\begin{aligned}
&(\check{f}_{\mathrm{LS}})_n - f_0(X_n) = \\
&\quad T_{n,\varepsilon}\check{S}_{n-1}\mathbf{Y}(X_n) - T_{n,\varepsilon}\check{S}_{n-1} f_0(X_n) + T_{n,\varepsilon}\check{S}_{n-1} f_0(X_n) - T_{n,\varepsilon} S_{n-1} f_0(X_n) + T_{n,\varepsilon} S_{n-1} f_0(X_n) - T_{n,\varepsilon} f_0(X_n) + \\
&\quad T_{n,\varepsilon} f_0(X_n) - f_0(X_n) \\
&= \underbrace{T_{n,\varepsilon}\check{S}_{n-1}\mathbf{w}(X_n)}_{\text{Term 1}} + \underbrace{T_{n,\varepsilon}[(\check{S}_{n-1} - S_{n-1}) f_0](X_n)}_{\text{Term 2}} + \underbrace{T_{n,\varepsilon} S_{n-1} f_0(X_n) - T_{n,\varepsilon} f_0(X_n)}_{\text{Term 3}} + \underbrace{T_{n,\varepsilon} f_0(X_n) - f_0(X_n)}_{\text{Term 4}}.
\end{aligned}$$

The first of these terms is the contribution of the noise $\mathbf{w} = (w_1, \dots, w_n)$ in the responses $\mathbf{Y}$. The second of these terms represents the difference between smoothing with $\check{S}_{n-1}$ and smoothing with S_{n-1}. The third of these terms is the bias due to Laplacian regularization. The fourth term is the error inherent to kernel smoothing of noiseless samples. Now we proceed to upper bound the error induced by each of these terms. We will find that

$$\big|(\check{f}_{\mathrm{LS}})_n - f_0(X_n)\big|^2 \le 4\Big(\frac{C_1}{\delta n \varepsilon^d} + C_2 \rho^2 \|f_0\|^2_{L^\infty(\mathcal{X})} + \frac{C_3 \rho}{\delta^2}|f_0|^2_{H^1(\mathcal{X})} + \frac{C_4}{\delta}\cdot \varepsilon^2 |f_0|^2_{H^1(\mathcal{X})}\Big)$$

with probability at least $1 - 4\delta - C\exp(-cn\varepsilon^d)$. From here, plugging in $\varepsilon = n^{-1/(2+d)}$ and $0 \le \rho \le n^{-2/(2+d)}$ proves the claim.

Term 1: Noise contribution. Let $\mathbb{E}_n$ denote the expectation conditional on the design points $X_1, \dots, X_n$. For any $i, j \in \{1, \dots, n\}$, we have

$$\mathbb{E}_n\big[(\check{S}_{n-1}\mathbf{w})_i \cdot (\check{S}_{n-1}\mathbf{w})_j\big] = (\check{S}^2_{n-1})_{ij}.$$

It is evident that $\check{S}_{n-1}$ is a contraction in $L^2(\mathbb{R}^n)$, meaning that

$$\|\check{S}_{n-1}\|_{\mathrm{op}} := \sup_{\substack{u \in \mathbb{R}^n \\ u \neq 0}} \frac{u^\top \check{S}_{n-1} u}{u^\top u} \le 1. \tag{D.9}$$

Let $K_{(x)} = \big(\eta(\|X_1 - x\|/\varepsilon), \dots, \eta(\|X_n - x\|/\varepsilon)\big) \in \mathbb{R}^n$ represent the vector of evaluations of the kernel η at a point $x \in \mathcal{X}$. From (D.9), we have

$$\begin{aligned}
\mathbb{E}_n\Big[\big(T_{n,\varepsilon}\check{S}_{n-1}\mathbf{w}(X_n)\big)^2\Big] &= \frac{1}{\big(d_{n,\varepsilon}(X_n)\big)^2}\sum_{i,j=1}^n \mathbb{E}_n\big[(\check{S}_{n-1}\mathbf{w})_i \cdot (\check{S}_{n-1}\mathbf{w})_j\big] \\
&= \frac{1}{\big(d_{n,\varepsilon}(X_n)\big)^2} K_{(X_n)}^\top \check{S}^2_{n-1} K_{(X_n)} \\
&\le \frac{1}{\big(d_{n,\varepsilon}(X_n)\big)^2}\|\check{S}_{n-1}\|^2_{\mathrm{op}} \cdot \big(K_{(X_n)}\big)^\top K_{(X_n)} \\
&\le \frac{1}{\big(d_{n,\varepsilon}(X_n)\big)^2} \cdot \big(K_{(X_n)}\big)^\top K_{(X_n)} \\
&\le \frac{\|\eta\|_{L^\infty(\mathbb{R})}}{d_{n,\varepsilon}(X_n)};
\end{aligned} \tag{D.10}$$

to obtain the last estimate we have used Holder's inequality to deduce

$$\big(K_{(X_n)}\big)^\top K_{(X_n)} = \sum_{i=1}^n \Big[\eta\Big(\frac{\|X_i - X_n\|}{\varepsilon}\Big)\Big]^2 \le \|\eta\|_{L^\infty(\mathbb{R})} \cdot d_{n,\varepsilon}(X_n).$$

Using Lemma 61, we can show that there exists a constant N such that for all $n \ge N$, with probability at least $1 - C\exp\big(-cn\varepsilon^d\big)$,

$$\sup_{x \in \mathcal{X}}\Big|\frac{d_{n,\varepsilon}(x)}{n} - d_{P,\varepsilon}(x; K)\Big| \le \frac{1}{2}\cdot \inf_{x \in \mathcal{X}} d_{P,\varepsilon}(x), \tag{D.11}$$

where $d_{P,\varepsilon}(x) = \mathbb{E}[d_{n,\varepsilon}(x)] = \int \eta(\|z - x\|/\varepsilon)\, dP(x)$. By assumption, the density $p(x) \ge p_{\min}$ for all $x \in \mathcal{X}$ and $\mathcal{X}$ has a C^1 boundary $\partial\mathcal{X}$. Thus there exists a positive constant $c_0 > 0$ such that $d_{P,\varepsilon}(x) \ge (1/3)p_{\min}\varepsilon^d$ for all $0 < \varepsilon < c_0$. Combined with (D.10) and (D.11), this implies

$$\mathbb{E}_n\Big[\big(T_{n,\varepsilon}\check{S}_{n-1}\mathbf{w}(X_n)\big)^2\Big] \le 12\frac{\|\eta\|_{L^\infty(\mathbb{R})}}{np_{\min}\varepsilon^d}.$$

Finally, using Markov's inequality, we conclude that with probability at least $1 - \delta - C\exp\bigl(-cn\varepsilon^d\bigr)$

$$\bigl(T_{n,\varepsilon}\check{S}_{n-1}w(X_n)\bigr)^2 \leq \frac{12}{\delta}\frac{\|\eta\|_{L^\infty(\mathbb{R})}}{np_{\min}\varepsilon^d} = C_1\frac{1}{\delta n\varepsilon^d}.$$

Term 2: Perturbation. From the Cauchy-Schwarz inequality, we have that for any $u \in \mathbb{R}^n$,

$$|T_{n,\varepsilon}u(X_n)|^2 = \left|\frac{1}{d_{n,\varepsilon}(X_n)}\sum_{i=1}^n u_i\eta\biggl(\frac{\|x_i - X_n\|}{\varepsilon}\biggr)\right|^2 \leq \frac{1}{d_{n,\varepsilon}(X_n)}\sum_{i=1}^n |u_i|^2\eta\biggl(\frac{\|x_i - X_n\|}{\varepsilon}\biggr) \leq \|u\|_\infty^2,$$

where in the last inequality we write $\|u\|_\infty := \max_{i=1,\dots,n}|u_i|$ for the $L^\infty(\mathbb{R}^n)$ norm. Thus to upper bound Term 2 we focus our attention on upper bounding $\|(\check{S}_{n-1} - S_{n-1})f_0\|_\infty$.

To begin with, we note that

$$(\check{S}_{n-1} - S_{n-1})f_0 = \check{S}_{n-1}(S_{n-1}^{-1} - \check{S}_{n-1}^{-1})S_{n-1}f_0 = \frac{\rho}{n}\check{S}_{n-1}\Bigl(L_{n-1,\varepsilon} - L_{n-1,\varepsilon}^{(n)}\Bigr)S_{n-1}f_0$$

and consequently

$$\begin{aligned}
\|(\check{S}_{n-1} - S_{n-1})f_0\|_\infty &\leq \frac{\rho}{n}\|\check{S}_{n-1}\|_\infty\Bigl(\|L_{n-1,\varepsilon}S_{n-1}f_0\|_\infty + \|L_{n-1,\varepsilon}^{(n)}S_{n-1}f_0\|_\infty\Bigr)\\
&\leq \frac{\rho}{n}\|\check{S}_{n-1}\|_\infty\Bigl(\|L_{n-1,\varepsilon}\|_\infty + \|L_{n-1,\varepsilon}^{(n)}\|_\infty\Bigr)\|S_{n-1}\|_\infty\|f_0\|_\infty\\
&\leq \frac{\rho}{n}\|\check{S}_{n-1}\|_\infty\Bigl(\|L_{n-1,\varepsilon}\|_\infty + \|L_{n-1,\varepsilon}^{(n)}\|_\infty\Bigr)\|S_{n-1}\|_\infty\|f_0\|_{L^\infty(\mathcal{X})} \qquad \text{(D.12)}
\end{aligned}$$

Here we have used that with probability 1, $\|f\|_\infty \leq \|f\|_{L^\infty(\mathcal{X})}$, and we recall that the ∞-operator norm of a matrix is $\|A\|_\infty = \max\{\|Au\|_\infty : \|u\|_\infty = 1\}$. It remains to give estimates on each of $\|\check{S}_{n-1}\|_\infty, \|S_{n-1}\|_\infty, \|L_{n-1,\varepsilon}\|_\infty$ and $\|L_{n-1,\varepsilon}^{(n)}\|_\infty$.

First, we observe that $\check{S}_{n-1}$ and S_{n-1} are contractions in $L^\infty(\mathbb{R}^n)$, meaning that $\|\check{S}_{n-1}\|_\infty, \|S_{n-1}\|_\infty \leq 1$. We establish that S_{n-1} is a contraction, and the exact same reasoning will hold with respect to $\check{S}_{n-1}$ as well. For any $u \in \mathbb{R}^n$, assuming without loss of generality that $u_1 = \|u\|_\infty$, we have that

$$(S_{n-1}^{-1}u)_1 = u_1 + (\rho L_{n,\varepsilon}u)_1 \geq u_1,$$

which implies that $\|S_{n-1}^{-1}u\|_\infty \geq \|u\|_\infty$. Thus for any $u \in \mathbb{R}^n$, we also have that

$$\|u\|_\infty = \|S_{n-1}^{-1}S_{n-1}u\|_\infty \geq \|S_{n-1}u\|_\infty,$$

demonstrating that S_{n-1} is a contraction.

On the other hand, for any $v \in \mathbb{R}^n$ such that $\|v\|_\infty = 1$,

$$\begin{aligned}
\|L_{n-1,\varepsilon}v\|_\infty &= \max_{i=1,\dots,n}\left|\frac{1}{n\varepsilon^{d+2}}\sum_{j=1}^n (v_i - v_j)\eta\biggl(\frac{\|x_i - x_j\|}{\varepsilon}\biggr)\right|\\
&\leq \max_{i=1,\dots,n}\frac{2}{n\varepsilon^{d+2}}\sum_{j=1}^n \eta\biggl(\frac{\|x_i - x_j\|}{\varepsilon}\biggr)\\
&\leq \frac{2}{\varepsilon^{d+2}},
\end{aligned}$$

and

$$\begin{aligned}\|L^{(n)}_{n-1,\varepsilon} v\|_\infty &= \max_{i=1,\dots,n} \left| \frac{1}{\varepsilon^{d+2}} \sum_{j=1}^n (v - v) W^{(u)}_{n,ij} \right| \\ &\leq \max_{i=1,\dots,n} \frac{2}{\varepsilon^{d+1}} \sum_{j=1}^n |W^{(u)}_{n,ij}| \\ &\leq \frac{2}{\varepsilon^{d+2}} \|\eta\|_\infty,\end{aligned}$$

so that $\|L_{n-1,\varepsilon}\|_\infty, \|L^{(n)}_{n-1,\varepsilon}\|_\infty \leq C\varepsilon^{-(d+2)}$. Plugging these estimates back into (D.12), we conclude that

$$\big(T_{n,\varepsilon}(\check{S}_{n-1} - S_{n-1}) f_0(X_n)\big)^2 \leq \Big(\|(\check{S}_{n-1} - S_{n-1}) f_0\|_\infty\Big)^2 \leq \left(\frac{C\rho \|f_0\|_{L^\infty(\mathcal{X})}}{n\varepsilon^{d+2}}\right)^2 = C_2 \rho^2 \|f_0\|^2_{L^\infty(\mathcal{X})}.$$

Term 3: Bias. We begin by taking an expectation of Term 3 conditional on the labeled design points $X_1, \dots, X_{n-1}$ (which we denote by $\mathbb{E}_{n-1}$):

$$\mathbb{E}_{n-1}\Big[\Big(T_{n,\varepsilon} S_{n-1} f_0(X_n) - T_{n,\varepsilon} f_0(X_n)\Big)^2\Big] = \|T_{n,\varepsilon} S_{n-1} f_0 - T_{n,\varepsilon} f_0\|_P^2.$$

In (C.66), we have shown that with probability at least $1 - C\exp(-cn\varepsilon^d)$

$$\|T_{n,\varepsilon}(u - f_0)\|_P^2 \leq C\|u - f_0\|_n^2, \quad \text{for all } u \in \mathbb{R}^n. \tag{D.13}$$

Now we can use upper bounds on $\|S_{n-1} f_0 - f_0\|_n^2$, the in-sample squared-bias of Laplacian smoothing, derived in Chapter 3. Combining the proof of Lemma 26 and the statement of Lemma 3, we have that with probability at least $1 - \delta$,

$$\|S_{n-1} f_0 - f_0\|_n^2 \leq C\rho \cdot \langle L_{n,\varepsilon} f_0, f_0 \rangle_n \leq \frac{C\rho}{\delta} |f_0|^2_{H^1(\mathcal{X})}. \tag{D.14}$$

Combining (D.13) and (D.14), we have that

$$\mathbb{E}_{n-1}\Big[\Big(T_{n,\varepsilon} S_{n-1} f_0(X_n) - T_{n,\varepsilon} f_0(X_n)\Big)^2\Big] \leq \frac{C_3 \rho}{\delta} |f_0|^2_{H^1(\mathcal{X})}$$

with probability at least $1 - \delta$. Thus it follows from Lemma 62 that with probability at least $1 - 2\delta$,

$$\Big(T_{n,\varepsilon} S_{n-1} f_0(X_n) - T_{n,\varepsilon} f_0(X_n)\Big)^2 \leq \frac{C_3 \rho}{\delta^2} |f_0|^2_{H^1(\mathcal{X})}.$$

Lemma 62. *Let $Z \geq 0$ and X be random variables, and let $0 \leq E[Z|X] \leq a$ with probability at least $1 - \delta$. Then*

$$\mathbb{P}\left(Z \geq \frac{a}{\delta}\right) \leq 2\delta.$$

Proof. By the law of iterated expectation,

$$\mathbb{P}(Z \geq \frac{a}{\delta}) = \mathbb{E}[\mathbb{P}(Z \geq \frac{a}{\delta}|X)] = \mathbb{E}[\mathbb{P}(Z \geq \frac{a}{\delta}|X) \wedge 1].$$

Since $E[Z|X] \leq a$ with probability at least $1 - \delta$,

$$\begin{aligned}\mathbb{E}[\mathbb{P}(Z \geq \frac{a}{\delta}|X) \wedge 1] &= \mathbb{E}\Big[\big(\mathbb{P}(Z \geq \frac{a}{\delta}|X) \wedge 1\big) \cdot (\mathbf{1}\{E[Z|X] \leq a\} + \mathbf{1}\{E[Z|X] \geq a\})\Big] \\ &\leq \mathbb{E}\Big[\big(\mathbb{P}(Z \geq \frac{a}{\delta}|X) \cdot (\mathbf{1}\{E[Z|X] \leq a\} + \mathbb{E}\Big[\mathbf{1}\{E[Z|X] \geq a\}\big)\Big] \\ &\leq 2\delta,\end{aligned}$$

with the last upper bound following from (conditional) Markov's inequality. □

Term 4: Noiseless kernel smoothing. We have already derived upper bounds on the error of noiseless kernel smoothing in Chapter 4. From (C.67) and Lemma 58, we have that

$$\mathbb{E}_{n-1}\Big[\Big(T_{n,\varepsilon}f_0(X_n) - f_0(X_n)\Big)^2\Big] = \|T_{n,\varepsilon}f_0 - f_0\|_P^2 \leq \frac{C}{\delta}\cdot\left(\frac{\varepsilon^2}{n\varepsilon^d} + \varepsilon^2\right)|f|_{H^1(\mathcal{X})}^2 \leq \frac{C_4}{\delta}\cdot\varepsilon^2|f|_{H^1(\mathcal{X})}^2$$

with probability at least $1 - \delta - C\exp(-cn\varepsilon^d)$.

D.2.2 Proof of Proposition 28

In this proof, for convenience we write $\check{f}$ for $\check{f}_{\mathrm{LS}}$. We begin by giving a closed-form expression for $\check{f}$. This is stated with respect to the restriction matrix $R_{n-1} \in \mathbb{R}^{n-1\times n}$ and extension-by-zero matrix $E_n \in \mathbb{R}^{n\times n-1}$; the former maps a vector $u \in \mathbb{R}^n$ to $R_{n-1}u = (u_1, \ldots, u_{n-1})$, and the latter maps $u \in \mathbb{R}^{n-1}$ to $E_n u = (u, 0)$. It is not hard to verify that

$$\check{f} = (\rho L_{n,\varepsilon} + E_n R_{n-1})^{-1} E_n \mathbf{Y}. \tag{D.15}$$

Next, we establish that $\check{f}_{1:(n-1)}$ satisfies (D.7). This is easily done using block matrix inversion. Writing

$$L_{n,\varepsilon} = \begin{pmatrix} L_{11} & L_{12} \\ L_{22} & L_{21} \end{pmatrix}, \quad E_n R_{n-1} = \begin{pmatrix} I_{n-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad E_n \mathbf{Y} = (\mathbf{Y}, 0)$$

we immediately have

$$\check{f} = \big(\rho(L_{11} - L_{12}L_{22}^{-1}L_{21}) + I_{n-1}\big)^{-1}\mathbf{Y}. \tag{D.16}$$

Note that $L_{11} = \frac{n-1}{n}L_{n-1,\varepsilon} + \frac{1}{n\varepsilon^{d+2}}\mathrm{diag}(K_{(X_n)})$—where $\mathrm{diag}(K_{(X_n)}) \in \mathbb{R}^{n-1\times n-1}$ denotes the diagonal matrix with entries $\mathrm{diag}(K_{(X_n)})_{ii} = (K_(X_n))_i$—and furthermore that

$$L_{12}L_{22}^{-1}L_{21} = \frac{1}{n\varepsilon^{d+2}d_{n-1,\varepsilon}(X_n)}K_{(X_n)}K_{(X_n)}^\top.$$

Consequently,

$$L_{11} - L_{12}L_{22}^{-1}L_{21} = \frac{n-1}{n}L_{n-1,\varepsilon} + \frac{1}{n}L_{n-1,\varepsilon}^{(n)},$$

and plugging this back into (D.16) yields (D.7).

On the other hand, we have that at the design point X_n,

$$\check{f}_n = \operatorname*{argmin}_a \sum_{i=1}^{n-1}(\check{f}_i - a)^2\eta\left(\frac{\|X_i - X_n\|}{\varepsilon}\right) = \frac{1}{\sum_{i=1}^{n-1}\eta(\|X_i - X_n\|/\varepsilon)}\sum_{i=1}^{n-1}\check{f}_i\eta\left(\frac{\|X_i - X_n\|}{\varepsilon}\right), \tag{D.17}$$

verifying (D.8).

D.3 Proof of Proposition 14

First, we prove (5.14), the upper bound on the in-sample risk of Laplacian eigenmaps. After that we explain why the corresponding upper bound on the in-sample risk of Laplacian smoothing follows. The proof of (5.14) will be made easier by two useful lemmas, which we stated in Section D.3.1.

We begin by showing that, with high probability, the eigenvectors v_1, v_2 respect the cluster structure of $p^{(n)}$. Denote $u_1 = (\mathbf{1}\{X_i \in Q_1\})_{i\in[n]}$, and likewise $u_2 = (\mathbf{1}\{X_i \in Q_2\})_{i\in[n]}$. We make the following two observations:

1. Because $\varepsilon < r$ and the kernel η is compactly supported on $[0, 1]$, for each $X_i \in Q_1$ and $X_j \in Q_2$, it must be the case that $\eta(\|X_i - X_j\|/\varepsilon) = 0$.

2. Using an elementary concentration argument (stated in Lemma 63) and the triangle inequality, we deduce that with probability at least $1 - 4/\varepsilon \exp(-n\varepsilon/4)$ there exists a path in $G_{n,\varepsilon}$ between each $X_i, X_j \in Q_1$, and likewise between each $X_i, X_j \in Q_2$.

Together these observations imply that with high probability the neighborhood graph $G_{n,\varepsilon}$ consists of exactly two connected components: one consisting of all design points $X_i \in Q_1$, and the other consisting of all design points $X_i \in Q_2$. In other words,

$$\mathbb{P}\Big(\operatorname{span}\{v_1, v_2\} = \operatorname{span}\{u_1, u_2\}\Big) \geq 1 - 4/\varepsilon \exp(-n\varepsilon/4). \tag{D.18}$$

Let us condition on the "good" event $\mathcal{E}$ that the design points $X_1, \dots, X_n$ satisfy (D.21), and therefore that $\operatorname{span}\{v_1, v_2\} = \operatorname{span}\{u_1, u_2\}$. Consider the empirical mean $\bar{Y}_Q := \frac{1}{\sharp\{Q \cup \mathbf{X}\}} \sum_{i: X_i \in Q} Y_i$. Since $\operatorname{span}\{v_1, v_2\} = \operatorname{span}\{u_1, u_2\}$, the estimator $\widehat{f} = \widehat{f}_{\mathrm{LE}}$ will be piecewise constant on Q_1 and Q_2, and in fact we have that

$$\widehat{f} = \bar{Y}_{Q_1} u_1 + \bar{Y}_{Q_2} u_2. \tag{D.19}$$

Therefore conditional on $\mathcal{E}$,

$$\|\widehat{f} - f_0^{(n)}\|_n^2 = P_n(Q_1) \cdot (\bar{Y}_{Q_1} - \theta)^2 + P_n(Q_2) \cdot (\bar{Y}_{Q_2} + \theta)^2$$

and consequently,

$$\mathbb{E}\Big[\|\widehat{f} - f_0^{(n)}\|_n^2 \Big| \mathcal{E}\Big] = \mathbb{E}\Big[\mathbb{E}\big[\|\widehat{f} - f_0^{(n)}\|_n^2 \big| X_1, \dots, X_n\big]\Big| \mathcal{E}\Big] = \frac{1}{n}. \tag{D.20}$$

Now we derive a crude upper bound on $\|\widehat{f} - f_0^{(n)}\|_n$ that will suffice to control the error conditional on $\mathcal{E}^c$. We observe that the empirical norm of $\widehat{f}$ is bounded,

$$\|\widehat{f}\|_n^2 \leq \frac{2}{n} \sum_{i=1}^n \langle Y, v_1\rangle_n^2 v_{1,i}^2 + \langle \mathbf{Y}, v_2\rangle_n^2 v_{2,i}^2 \leq 2\big(\langle \mathbf{Y}, v_1\rangle_n^2 + \langle \mathbf{Y}, v_2\rangle_n^2\big) \leq 4\|\mathbf{Y}\|_n^2.$$

Noting that $\mathbb{E}[\|\mathbf{Y}\|_n^2 | X_1, \dots, X_n] = \|f_0\|_n^2 + 1/n = \theta^2 + 1/n$, we conclude that

$$\mathbb{E}\Big[\|\widehat{f} - f_0\|_n^2 \cdot \mathbf{1}\{\mathcal{E}^c\}\Big] \leq \mathbb{E}\Big[\Big(2\|f_0\|_n^2 + 4(\theta^2 + 1/n) \cdot \mathbf{1}\{\mathcal{E}^c\}\Big] \leq (6\theta^2 + n^{-1}) \cdot 4\varepsilon^{-1} \exp(-n\varepsilon/4).$$

Combining this with (D.20) implies (5.14).

As for Laplacian smoothing, it suffices to note that $\lim_{\rho \to \infty} \widehat{f}_{\mathrm{LS}}$ will also satisfy (D.19). Thus all of the subsequent calculations used to upper bound the mean squared error of $\widehat{f} = \widehat{f}_{\mathrm{LE}}$ will also apply to $\widehat{f} = \lim_{\rho \to \infty} \widehat{f}_{\mathrm{LS}}$.

D.3.1 A Useful Lemma

We introduce some notation: for a positive number m and each $i = 0, \dots, m-1$, let

$$Q_{i1} = [i/m, (i+1)/m] \cdot (1/2 - r), \quad Q_{i2} = 1/2 + [i/m, (i+1)/m] \cdot (1/2 - r).$$

Lemma 63. *Suppose $(X_1, Y_1), \dots (X_n, Y_n)$ are sampled according to (5.9), and suppose $r \leq 1/4$. We have that*

$$\mathbb{P}\Big(\sharp\{Q_{ij} \cup \mathbf{X}\} > 0 \ \text{ for all } i = 1, \dots, m-1 \text{ and } j = 1, 2\Big) \geq 1 - 2m \exp\{-n/2m\}. \tag{D.21}$$

Proof (of Lemma 63). For each Q_{ij}, we have that $P(Q_{ij}) = (1/2 - r)/m \geq 1/(2m)$. Therefore

$$\mathbb{P}\big(\sharp\{Q_{ij} \cup \mathbf{X}\} = 0\big) = (1 - 1/(2m))^n \leq \exp\{-n/2m\}.$$

By a union bound,

$$\mathbb{P}\Big(\sharp\{Q_{ij} \cup \mathbf{X}\} = 0 \ \text{ for any } i = 1, \ldots, m-1 \text{ and } j = 1, 2\Big) \leq 2m \exp\{-n/2m\}. \quad \square$$

Let $\varepsilon = 2/m$. Note that by construction, (D.21) implies that any points x and x' in adjacent intervals Q_{ij} and $Q_{i'j}$ must be connected in $G_{n,\varepsilon}$. Likewise, it implies that for $h = 1/m$ the degree $d_{n,h}(x) > 0$ for every $x \in Q_1 \cup Q_2$.

D.4 Proof of Proposition 15

First we show (5.15), then (5.16).

D.4.1 Proof of (5.15)

A standard argument using the law of iterated expectation implies the following lower bound on the pointwise risk in terms of squared-bias and variance-like quantities,

$$\mathbb{E}\Big[\Big(\widetilde{f}(X_i) - f_0(X_i)\Big)^2 | X_i = x\Big] \geq \frac{(n-1)}{n} \mathbb{E}\Big[\Big(f_0(X) - f_0(x)\Big)^2 | X \in B(x, h')\Big] + \mathbb{E}\Big[\frac{1}{d_{n,h'}(x)}\Big].$$

The variance term can be lower bounded quite simply for any $x \in \mathcal{X}^{(n)}$; noting that $\sup_x p^{(n)}(x) < 2$ and $\nu(B(x, h') \cap \mathcal{X}^{(n)}) \leq 2h'$, it follows by Jensen's inequality that

$$\mathbb{E}\Big[\frac{1}{d_{n,h'}(x)}\Big] \geq \frac{1}{\mathbb{E}[d_{n,h'}(x)]} \geq \frac{1}{4nh'}.$$

On the other hand the squared bias term is quite large for x close to $1/2$. Precisely, if $h' \geq 4r$ then a simple calculation implies

$$\mathbb{E}[(f_0(X) - f_0(x))^2 | X \in B(x, h')] \geq \frac{\theta^2}{8} \quad \text{for all } x \in [(1 - h'/2)_+, 1/2 - r].$$

Combining these lower bounds on variance and squared bias terms and summing over $X_1, \ldots, X_n$, we arrive at the following: if $h' \leq 4r$, then

$$\mathbb{E}\Big[\|\widetilde{f} - f_0^{(n)}\|_n^2\Big] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}\Big[\Big(\widetilde{f}(X_i) - f_0(X_i)\Big)^2 | X_i\Big]\Big] \geq \frac{1}{16rn},$$

whereas if $h' > 4r$ then

$$\begin{aligned}
\mathbb{E}\Big[\|\widetilde{f} - f_0^{(n)}\|_n^2\Big] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}\Big[\Big(\widetilde{f}(X_i) - f_0(X_i)\Big)^2 | X_i\Big]\Big] \\
&\geq \frac{1}{4nh'} + \frac{\theta^2}{8} \frac{(n-1)}{n} P^{(n)}\Big([(1 - h'/2)_+, 1/2 - r]\Big) \\
&\geq \frac{1}{4nh'} + \frac{\theta^2 h'}{64}.
\end{aligned}$$

In the latter case, setting the derivative equal to 0 shows that the right hand side is always at least $\theta/\sqrt{64n}$, and taking the minimum over the two cases then yields (5.15).

D.4.2 Proof of (5.16)

We begin by decomposing the risk into conditional bias and variance terms. Let $\mathbb{E}_n = \mathbb{E}[\cdot|X_1, \dots, X_n]$ denote expectation conditional on the design points $X_1, \dots, X_n$. Then by the law of iterated expectation, and the fact that $\mathbb{E}_n[w] = 0$,

$$\mathbb{E}\big[\|\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\big] = \mathbb{E}\big[\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\big] + \mathbb{E}\big[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\big].$$

We separately lower bound the expected conditional squared bias and variance terms. To anticipate what is to come: we will show that the expected conditional variance is equal to K/n; on the other hand we will show that the expected conditional squared bias is lower bounded,

$$\mathbb{E}\big[\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\big] = \frac{K}{n} \quad \text{and} \quad \mathbb{E}\big[\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\big] \geq \frac{\theta^2}{2601\pi^2 K^3}, \tag{D.22}$$

with the lower bound holding so long as $K \leq \min\{1/(16r), n/(8\log(8n)), (\sqrt{160}\pi/r)^{2/3}\}$. If K is larger than this, then the expected conditional variance is lower bounded,

$$\mathbb{E}\big[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\big] \geq \min\bigg\{\frac{1}{16rn}, \frac{1}{8\log(8n)}, \frac{(\sqrt{160}\pi)^{2/3}}{r^{2/3}n}\bigg\} \tag{D.23}$$

Otherwise (D.22) implies that the in-sample risk is always at least

$$\mathbb{E}\big[\|\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\big] \geq \frac{\theta^2}{2601\pi^2 K^3} + \frac{K}{n} \geq 2\frac{\theta^{1/2}}{n^{3/4}}\frac{1}{(2601\pi^2)^{1/4}}.$$

Along with (D.23), this implies the claim. It remains to show the bounds on conditional bias and variance.

Conditional variance. The expected conditional variance is exactly equal to K/n, a standard fact that is verified by the following calculations: first,

$$\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2 = \|\Phi(\Phi^\top\Phi)^{-1}\Phi^\top w\|_n^2 = \frac{1}{n}w^\top\Phi(\Phi^\top\Phi)^{-1}\Phi^\top w;$$

thus standard properties of the Gaussian distribution and the trace trick imply

$$\mathbb{E}_n\big[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\big] = \frac{1}{n}\mathrm{tr}(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top) = \frac{K}{n};$$

and finally by the law of iterated expectation and the independence of the noise $(w_1, \dots, w_n)$ and the design points $X_1, \dots, X_n$,

$$\mathbb{E}\Big[\mathbb{E}_n\big[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\big]\Big] = K/n.$$

Conditional bias. It takes more work to lower bound the conditional bias. We will first upper bound the Lipschitz constant of $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$ in terms of the empirical norm $\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n$. Then we will use this upper bound to argue that either $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$ has empirical norm much larger than that of f_0, or $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$ is a smooth function, in the sense of having a small Lipschitz constant. In the former case, the triangle inequality will then imply that $\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n$ must be large. In the latter case, the smoothness of $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$ will imply that $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$ must be far from f_0 at many points X_i close to $x = 1/2$.

The following Lemma gives our upper bound on the Lipschitz constant of $\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n$. Here we treat $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} = \sum_{k=1}^K \widetilde{\beta}_k\phi_k$ as a function defined at all $x \in [0,1]$ by extending it in the canonical way. As a function over $[0,1]$, clearly $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} \in C^\infty([0,1])$. Let $\Sigma \in \mathbb{R}^{K\times K}$ be the covariance matrix of $(\phi_1, \dots, \phi_K)$, i.e. the matrix with entries $\Sigma_{k\ell} = \langle\phi_k, \phi_{\ell,}\rangle_{P^{(n)}}$. Let $\widehat{\Sigma} := (\Phi^\top\Phi)/n$ be the empirical covariance matrix. Let $I_K \in \mathbb{R}^{K\times K}$ be the identity matrix.

Lemma 64 (Lipschitz regularity of $\mathbb{E}_n \widetilde{f}_{\mathrm{SP}}$.)**.** *Let* $\widetilde{f}_n = \mathbb{E}_n \widetilde{f}_{\mathrm{SP}}$. *Then*

$$\|\widetilde{f}_n\|^2_{C^1(\mathcal{X})} \le \pi^2 \frac{K^3 \cdot \|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}}}{(1 - \|I_K - \Sigma\|_F)} \cdot \|\widetilde{f}_n\|_n^2. \tag{D.24}$$

Moreover, suppose $K \le 1/(16r)$ *and* $r \le (1 - 2^{-1/2})/2$.

- **(Matrix perturbation)** *Then*

$$\|\Sigma - I_K\|_F \le \frac{1}{2}. \tag{D.25}$$

- **(Matrix concentration, cf. Hsu et al. [2012])** *If additionally* $n \ge 8K\log(K/\delta)$ *for some* $\delta \in (0, 1/2)$, *then with probability at least* $1 - 2\delta$,

$$\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}} \le 5. \tag{D.26}$$

Therefore, if $K \le \min\{1/(16r), n/(8\log(K/\delta))\}$, *then with probability at least* $1 - 2\delta$,

$$\|\widetilde{f}_n\|^2_{C^1(\mathcal{X})} \le 10\pi^2 K^3 \|\widetilde{f}_n\|_n^2. \tag{D.27}$$

We defer the proof of Lemma 64 until after we complete the proof of (5.16).

Now, if $\|\widetilde{f}_n\|_n^2 \ge \frac{3}{2}\|f_0\|_n^2$, then by the triangle inequality

$$\|\widetilde{f}_n - f_0\|_n \ge \|\widetilde{f}_n\|_n - \|f_0\|_n \ge \sqrt{\frac{3}{2}} \cdot \|f_0\|_n = \sqrt{\frac{3}{2}} \cdot \theta.$$

Otherwise $\|\widetilde{f}_n\|_n^2 \ge \frac{3}{2}\|f_0\|_n^2$. In this case, we show that $|\widetilde{f}_n(X_i) - f_0(X_i)|$ must be large (on the order of θ) for many points X_i which are close to $x = 1/2$. Let us suppose without loss of generality that $\widetilde{f}_n(1/2) \le \theta/2$ and consider points $X_i \in Q_1$ close to $x = 1/2$; otherwise if $\widetilde{f}_n(1/2) > \theta/2$ we could obtain the exact same bound by considering $X_i \in Q_2$. For each point $X_i \in Q_1$, by Lemma 64 we have that with probability at least $1 - 2\delta$,

$$|\widetilde{f}_n(X_i) - \widetilde{f}_n(1/2)| \le CK^{3/2}\|\widetilde{f}_n\|_n \cdot |X_i - 1/2| \le \sqrt{10}\pi K^{3/2}\theta \cdot |X_i - 1/2|.$$

Since $\widetilde{f}_n(1/2) \le \theta/2$ and $f_0(X_i) = \theta/2$ for all $X_i \in Q_1$ it follows that

$$|\widetilde{f}_n(X_i) - f_0(X_i)| \ge \theta - \sqrt{10}\pi K^{3/2}\theta \cdot |X_i - 1/2|,$$

and consequently

$$|\widetilde{f}_n(X_i) - f_0(X_i)| \ge \theta/2, \quad \text{for any } X_i \in Q_1 \text{ such that } |X_i - 1/2| \le 1/(\sqrt{40}\pi K^{3/2}).$$

This yields a lower bound on $\|\widetilde{f}_n - f_0\|_n$; letting $Q_K := \left[\frac{1}{2} - \frac{1}{\sqrt{40}\pi K^{3/2}}, \frac{1}{2} - r\right]$, we have that

$$\|\widetilde{f}_n - f_0\|_n \ge \frac{\theta}{2} \cdot P_n\Big(Q_k\Big).$$

Then as long as $K^{-3/2} \ge \sqrt{160}\pi r$, from the multiplicative form of Hoeffding's inequality (Lemma 15)

$$P^{(n)}(Q_K) \ge \frac{1}{\sqrt{160}\pi K^{3/2}} \ge 2r \Longrightarrow \mathbb{P}\left(P_n(Q_K) \ge \frac{1}{\sqrt{640}\pi K^{3/2}}\right) \ge 1 - \exp(-nr/4) \ge 1 - \frac{4}{n^2}.$$

Putting the pieces together, we conclude that if $K \le \min\{1/(8r), n/(8\log(K/\delta)), (\sqrt{160}\pi/r)^{2/3}\}$, then

$$\|\widetilde{f}_n - f_0\|_n \ge \frac{\theta}{51\pi K^{3/2}},$$

with probability at least $1 - 2\delta - 4n^2$. Taking $\delta = 1/8$ then implies the claim.

Proof of Lemma 64. *Proof of (D.24).* Recall that $\widetilde{f}_n = \sum_{k=1}^K \widetilde{\beta}_k \phi_k$. Exchanging sum with derivative, we have that

$$\frac{d}{dx}\widetilde{f}_n(x) = -\pi \sum_{k=1}^K \widetilde{\beta}_k k \sin(k\pi x).$$

Thus taking absolute value and applying the Cauchy-Schwarz inequality gives

$$|\widetilde{f}_n'(x)|^2 \le \pi^2 K^2 \sum_{k=1}^K \big(\sin(k\pi x)\big)^2 \|\beta\|_2^2 \le \pi^2 K^3 \|\beta\|_2^2.$$

On the other hand, we can also relate the empirical norm $\|\widetilde{f}_n\|_n^2$ to the ℓ^2 norm of β. Specifically,

$$\|\widetilde{f}_n\|_n^2 = \beta^\top \widehat{\Sigma}\beta \ge \frac{\|\beta\|_2^2}{\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}}} \ge \frac{\|\beta\|_2^2}{\|\Sigma^{-1}\|_{\mathrm{op}} \cdot \|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}}} = \frac{\|\beta\|_2^2 \|\Sigma\|_{\mathrm{op}}}{\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}}}$$

Rearranging, we see that

$$\sup_{x\in[0,1]} |\widetilde{f}_n'(x)|^2 \le \frac{\pi^2 K^3}{\|\Sigma\|_{\mathrm{op}}} \|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}} \le \frac{\pi^2 K^3}{1 - \|I_K - \Sigma\|_F} \|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}}$$

with the latter inequality following since $\|\Sigma\|_{\mathrm{op}} \ge \|I_K\|_{\mathrm{op}} - \|I_K - \Sigma\|_{\mathrm{op}} \ge 1 - \|I_K - \Sigma\|_F$.

Proof of (D.25). We will show that for all $1 \le k < \ell \le K$,

$$(1 - \langle \phi_k, \phi_k \rangle_{P^{(n)}})^2 \le 32r^2, \quad \text{and} \quad |\langle \phi_k, \phi_\ell \rangle_{P^{(n)}}| \le 64r^2. \tag{D.28}$$

This implies $\|I - \Sigma\|_F^2 \le 32K^2r^2$, so that $\|I - \Sigma\|_F \le 1/2$ so long as $K \le 1/(16r)$.

The proof of (D.28) follows from computing some standard integrals. We separate the computation based on whether $k = 1$ or $k > 1$.

Case 1: $k = 1$. When $k = 1$, $\langle \phi_1, \phi_1 \rangle_{P^{(n)}} = 1$ and $(1 - \langle \phi_1, \phi_1 \rangle_{P^{(n)}})^2 = 0$. Additionally, by symbolic integration we find that

$$\langle \phi_k, \phi_\ell \rangle_{P^{(n)}} = \frac{-2\sqrt{2}}{(1-2r)} \cdot \frac{\cos(\ell\pi/2)\sin(\ell\pi r)}{\ell\pi},$$

and therefore

$$\big[\langle \phi_k, \phi_\ell \rangle_{P^{(n)}}\big]^2 \le \frac{8}{(1-2r)^2} \cdot \left(\frac{\sin(\ell\pi r)}{\ell\pi}\right)^2 \le \frac{8}{(1-2r)^2} r^2 \le 16r^2,$$

where in the second-to-last inequality follows because $\sin(x)/x \le 1$, and the last inequality follows by our assumed upper bound on r.

Case 2: $k > 1$. When $k > 1$,

$$\langle \phi_k, \phi_k \rangle_{P^{(n)}} = 1 - \frac{2}{(1-2r)} \frac{\cos(k\pi)\sin(2k\pi r)}{k\pi} \Longrightarrow \big[1 - \langle \phi_k, \phi_k \rangle_{P^{(n)}}\big]^2 \le \frac{4}{(1-2r)^2} \cdot \left(\frac{\sin(2k\pi r)}{k\pi}\right)^2 \le 32r^2.$$

Similarly,

$$\langle \phi_k, \phi_\ell \rangle_{P^{(n)}} = -\frac{4}{(1-2r)} \left[\frac{\cos((k+\ell)\pi)\sin((k+\ell)\pi r)}{(k+\ell)\pi} + \frac{\cos((k-\ell)\pi)\sin((k-\ell)\pi r)}{(k-\ell)\pi}\right]$$

and therefore

$$\left[\langle \phi_k, \phi_\ell \rangle_{P^{(n)}}\right]^2 \le \frac{16}{(1-2r)^2}\left(\left[\frac{\sin((k+\ell)\pi r)}{(k+\ell)\pi}\right]^2 + \left[\frac{\sin((k-\ell)\pi r)}{(k-\ell)\pi}\right]^2\right) \le 64r^2.$$

Proof of (D.26) Denote $\Phi(x) = (\phi_1, \dots, \phi_K(x)) \in \mathbb{R}^K$ for any $x \in [0,1]$. Then for any $x \in [0,1]$,

$$\|\Sigma^{-1/2}\Phi(x)\| \le \|\Sigma^{-1}\|_{\text{op}}^{1/2}\|\Phi(x)\|_2 \le \|\Sigma^{-1}\|_{\text{op}}^{1/2}\sqrt{2K} \le 2\sqrt{K}$$

with the second-to-last inequality following from (D.25), and the last inequality following since $|\phi_k(x)| \le \sqrt{2}$ for all k. Thus $\|\Sigma^{-1/2}\Phi(x)\|/\sqrt{K} \le 2$, and (D.26) follows from Theorem 1 of Hsu et al. [2012].

Proof of (D.27). Follows immediately.

D.5 Proof of Proposition 16

To begin with, we show that

$$\frac{1}{n}(I + \rho\Delta)g_n = \delta_i = \frac{1}{n}(I + \rho n^2 \bar{L}_n)\bar{g}, \quad \text{in } \mathbb{R}^n, \tag{D.29}$$

and therefore

$$(I + \rho\bar{L}_n)(g_n - \bar{g}) = \rho(\bar{L}_n - \Delta)g_n, \quad \text{in } \mathbb{R}^n. \tag{D.30}$$

To show (D.29), we must recall a series of facts.

1. **Fact 1**. The kth eigenpair (λ_k, v_k) of $n^2\bar{L}_n$ is given by

$$\lambda_k = 2n^2(1 - \cos(\pi k/n), \quad v_{k,i} = \cos(\pi k i/n - \pi k/2n) = \phi_k(x_i), \quad \text{for } 1 \le k \le n.$$

2. **Fact 2**. The eigenfunctions ϕ_k of Δ each have unit $L^2(P_n)$ norm,

$$\|\phi_k\|_n^2 = 1, \quad \text{for } 1 \le k \le n.$$

   Thus $\phi_1, \dots, \phi_n$ form an orthonormal basis of $\mathbb{R}^n$, with respect to the inner product $\langle \cdot, \cdot \rangle_n$. For any $u = \sum_{k=1}^n a_k\phi_k, v = \sum_{k=1}^n b_k\phi_k$, it follows that

$$\langle u, v \rangle_n = \sum_{k=1}^n a_k b_k = \langle u, v \rangle_P.$$

3. **Fact 3**. The function $(I + \rho\Delta)g_n$ is in the span of $\phi_1, \dots, \phi_n$. Therefore for any $u = \sum_{k=1}^n a_k\phi_k$, we have

$$\langle (I + \rho\Delta)g_n, u \rangle_n = \langle (I + \rho\Delta)g_n, u \rangle_P = \langle (I + \rho\Delta)g, u \rangle_P = u(x_i),$$

   the second equality following because $(I + \rho\Delta)g_n$ and u belong to the span of $\phi_1, \dots, \phi_n$.

Since $\phi_1, \dots, \phi_n$ form an orthonormal basis of $\mathbb{R}^n$, Fact 3 implies $\langle (I+\rho\Delta)g_n, v \rangle_n = v_i$ for all vectors $v \in \mathbb{R}^n$, which is equivalent to (D.29).

Proceeding from (D.30), multiplying by $\bar{L}_n^{-1/2}$, squaring and averaging each side gives

$$\|(\bar{L}_n^{-1/2} + \rho\bar{L}_n^{1/2})(\bar{g} - g)\|_n^2 = \rho^2\|\bar{L}_n^{-1/2}(\bar{L}_n - \Delta)g_n\|_n^2.$$

The left hand side of this equality is exactly the left hand side of (5.28). Thus it remains only to upper bound the right hand side. To do this, we first observe that because ϕ_k are eigenfunctions of both $\bar{L}_n$ and Δ,

$$\bar{L}_n^{-1/2}(\bar{L}_n - \Delta)g_n = \sum_{k=1}^{n} \langle g, \phi_k \rangle_P \bar{L}_n^{-1/2}(\bar{L}_n - \Delta)\phi_k = \sum_{k=2}^{n} \langle g, \phi_k \rangle_P \frac{\lambda_k - \lambda_k(\Delta_P)}{\lambda_k^{1/2}} \phi_k$$

Since $\phi_1, \dots, \phi_n$ are orthonormal with respect to $\langle \cdot, \cdot \rangle_n$, in squared norm we have

$$\begin{aligned} \|\bar{L}_n^{-1/2}(\bar{L}_n - \Delta)g_n\|_n^2 &= \sum_{k=2}^{n} \langle g, \phi_k \rangle_P^2 \cdot \left(\frac{\lambda_k - \lambda_k(\Delta_P)}{\lambda_k^{1/2}} \right)^2 \\ &= \sum_{k=2}^{n} \frac{\big(\phi_k(x_i)\big)^2}{(1 + \rho\lambda_k(\Delta_P))^2} \cdot \left(\frac{\lambda_k - \lambda_k(\Delta_P)}{\lambda_k^{1/2}} \right)^2 \\ &\le \sum_{k=2}^{n} \frac{1}{(1 + \rho\lambda_k(\Delta_P))^2} \cdot \left(\frac{\lambda_k - \lambda_k(\Delta_P)}{\lambda_k^{1/2}} \right)^2; \end{aligned} \tag{D.31}$$

the latter equality follows from the spectral representation $g(x) = \sum_k \phi_n(x_i)\phi_n(x)/(1 + \rho\lambda_k(\Delta_P))$.

Taking a Taylor expansion of $\lambda_k = 2n^2\big(1 - \cos(k\pi/n)\big)$ around $k = 0$, we obtain the following estimates,

$$|\lambda_k - \lambda_k(\Delta_P)| \le C\frac{k^4}{n^2}, \quad \text{and} \quad \lambda_k^{1/2} \ge ck, \quad \text{for all } 2 \le k \le n.$$

Plugging these estimates back into (D.31) gives the upper bound

$$\|\bar{L}_n^{-1/2}(\bar{L}_n - \Delta)g_n\|_n^2 \le C\frac{1}{\rho^2 n^4} \sum_{k=2}^{n} k^2 = C\frac{1}{\rho^2 n},$$

which proves the claim.

D.6 Proof of Proposition 17

(TODO)