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

Alden Green **Sivaraman Balakrishnan** **Ryan J. Tibshirani**

Carnegie Mellon University

Abstract

In this paper we study the statistical properties of Laplacian smoothing, a graph-based approach to nonparametric regression. Under standard regularity conditions, we establish upper bounds on the error of the Laplacian smoothing estimator $\widehat{f}$, and a goodness-of-fit test also based on $\widehat{f}$. These upper bounds match the minimax optimal estimation and testing rates of convergence over the first-order Sobolev class $H^1(\mathcal{X})$, for $\mathcal{X} \subseteq \mathbb{R}^d$ and $1 \leq d < 4$; in the estimation problem, for $d = 4$, they are optimal modulo a $\log n$ factor. Additionally, we prove that Laplacian smoothing is manifold-adaptive: if $\mathcal{X} \subseteq \mathbb{R}^d$ is an m-dimensional manifold with $m < d$, then the error rate of Laplacian smoothing (in either estimation or testing) depends only on m, in the same way it would if $\mathcal{X}$ were a full-dimensional set in $\mathbb{R}^m$.

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

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

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. Let $G = (V, W)$ be a weighted undirected graph, where the vertices $V = \{1, \ldots, n\}$ are 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

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

Here L is the graph Laplacian matrix (defined formally in Section 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

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

encourages $\widehat{f}_i \approx \widehat{f}_j$ when $X_i \approx X_j$. Assuming (2) is a reasonable estimator of f_0, the statistic

$$\widehat{T} = \frac{1}{n} \|\widehat{f}\|_2^2 \tag{3}$$

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 [2008]), 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 (1),

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

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, these works 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. [2016, 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 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., (2) or (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 [2018] 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 $f^\top L f$ in problem (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.

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 (2) (setting aside computational considerations for the moment). After all, we could have instead solved the following variational problem:

$$\widetilde{f} = \operatorname*{argmin}_{f : \mathcal{X} \to \mathbb{R}} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2 + \rho \int_{\mathcal{X}} \|\nabla f(x)\|_2^2 \, dx, \quad (4)$$

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:

$$\widetilde{T} = \|\widetilde{f}\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \widetilde{f}(X_i)^2. \quad (5)$$

The penalty term in (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 (2) can be viewed as a discrete and noisy approximation to (4). At first blush, this suggests that Laplacian smoothing should at best inherit the statistical properties of (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

Dimension	Laplacian smoothing (2)	Thin-plate splines (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 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 2. Here, we use "1" to denote 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.

spline estimator (4) is not well-posed (to be explained shortly); Tables 1 and 2 summarize. As we establish in Theorems 1-5, 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 (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 (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

$$\|\widehat{f} - f_0\|_n^2 := \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{f}_i - f_0(X_i)\Big)^2. \qquad (6)$$

Dimension	Laplacian smoothing (3)	Thin-plate splines (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 2: Summary of testing rates over first-order Sobolev balls; black, red, and bold fonts are used as in Table 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 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 $\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}^2$ should remain small, but leave a formal analysis to future work.

In a different direction, problem (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

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

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 BACKGROUND

Before we present our main results in Section 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 $V = \{1, \ldots, n\}$, 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

$$L = D - W. \tag{7}$$

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 $0 = \lambda_1 \le \cdots \le \lambda_n$, 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 $j = 1, \ldots, d$, 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 $\nabla f = (D^1 f, \ldots, D^d f)$,

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

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

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

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

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

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 $M \ge n^{-1/2}$ is (e.g., Tsybakov [2008]):

$$\inf_{\widehat{f}} \sup_{f_0 \in H^1(\mathcal{X}, M)} \mathbb{E}\Big[\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}^2\Big] \asymp M^{2d/(2+d)} n^{-2/(2+d)}. \tag{8}$$

(Throughout we assume $M \ge n^{-1/2}$, 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

$$\mathbf{H}_0 : f_0 = f_0^\star, \quad \text{versus} \quad \mathbf{H}_a : f_0 \in \mathcal{F} \setminus \{f_0^\star\}. \tag{9}$$

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 $\mathbb{E}_0[\phi] \le \alpha$ 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 $b \ge 1$, 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,

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

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] See Ingster [1982, 1987], Ingster and Suslina [2012] for a more extended treatment of the minimax paradigm in nonparametric testing.

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

$$\epsilon^2\big(H^1(\mathcal{X}, M)\big) \asymp M^{2d/(4+d)} n^{-4/(4+d)}. \qquad (10)$$

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.

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.

(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 1, our main estimation result. Its proof, as with all proofs of results in this paper, can be found in the supplementary document.

Theorem 1. *Given i.i.d. draws (X_i, Y_i), $i = 1, \ldots, n$ from (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 (2) with $\rho = M^{-4/(2+d)}(nr^{d+2})^{-1} n^{-2/(2+d)}$ satisfies*

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

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 1 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 the supplement.
- 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 we show in the supplement, if f_0 is Lipschitz smooth and we extend $\widehat{f}$ to be piecewise constant over the Voronoi tessellation induced by $X_1, \ldots, X_n$, then the out-of-sample

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

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 2. *Under the assumptions of Theorem 1, 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 1. 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. [2016] 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 1, we study the (in-sample) mean squared error of the Laplacian smoothing estimator as the dimension d grows. Here $X_1, \ldots, X_n$ 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

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

We show in the supplement 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^2 is at least (a constant times) the squared critical radius given in (10). For this to hold, we will require a tighter range of scalings for the graph radius r.

(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 3, our main testing result.

Theorem 3. *Given i.i.d. draws (X_i, Y_i), $i = 1, \ldots, n$ from (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 (2), with $\rho = (nr^{d+2})^{-1}n^{-4/(4+d)}M^{-8/(4+d)}$, satisfies the following: for any $b \geq 1$, if*

$$\epsilon^2 \geq CM^{2d/(4+d)}n^{-4/(4+d)}\left(b^2 + b\sqrt{\frac{1}{\alpha}}\right), \tag{11}$$

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

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

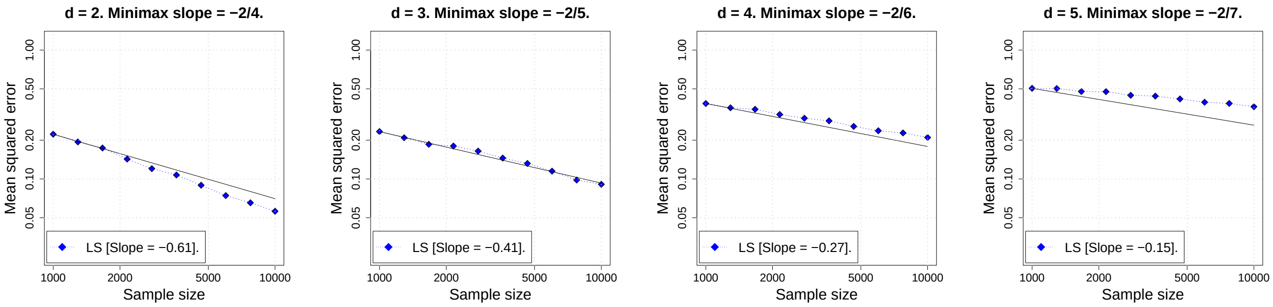

Figure 1: Mean squared error of Laplacian smoothing (LS) as a function of sample size n. Each plot is on the log-log scale, and the results are averaged over 5 repetitions, with Laplacian smoothing tuned for optimal average mean squared error. The black line shows the minimax rate (in slope only; the intercept is chosen to match the observed error).

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

$$\mathbb{E}_{f_0}\big[1 - \widehat{\varphi}\big] \leq \frac{C(1 + M^4)}{b^2}. \tag{12}$$

- 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 (2), (4) can be closely tied together: Bousquet et al. [2004] show that for $f \in C^2(\mathcal{X})$,

$$\lim \frac{1}{n^2 r^{d+2}} f^\top L f = \int_{\mathcal{X}} f(x) \cdot \Delta_P f(x) p(x)\, dx = \int_{\mathcal{X}} \|\nabla f(x)\|_2^2 p^2(x)\, dx. \tag{13}$$

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 smoothing estimator $\widehat{f}$ and the thin-plate spline estimator $\widetilde{f}$ are close (for one, note that (13) holds for $f \in C^2(\mathcal{X})$, whereas the optimization in (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 (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

[2] Assuming f satisfies e.g., Dirichlet boundary conditions.

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 (4). In this sense, any estimator resulting from solving (4) will clearly be inconsistent.

On the other hand, problem (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 (4) more tractable: for instance, one can replace the penalty $\int_{\mathcal{X}} \|\nabla f(x)\|_2^2 \, dx$ with a stricter choice like $\operatorname{ess\,sup}_{x \in \mathcal{X}} \|\nabla f(x)\|_2$, 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 $\widehat{f}$.

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 (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 $X_1, \ldots, X_n$). A standard calculation shows that

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

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

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

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., 2019, 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 1. *There are constants $N, C_2 > 0$ such that for $n \ge N$, $r \le c_0$, and $f \in H^1(\mathcal{X})$, with probability at least $1 - \delta$, it holds that*

$$f^\top L f \le \frac{C_2}{\delta} n^2 r^{d+2} |f|_{H^1(\mathcal{X})}^2. \qquad (14)$$

Lemma 2. *There are constants $N, C_1, C_3, c_1, c_3 > 0$ such that for $n \ge N$ and $C_0(\log n/n)^{1/d} \le r \le c_0$, with probability at least $1 - C_1 n \exp(-c_1 n r^d)$, it holds that*

$$c_3 A_{n,r}(k) \le \lambda_k \le C_3 A_{n,r}(k), \quad \text{for } 2 \le k \le n, \qquad (15)$$

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

Lemma 1 gives a direct upper bound on the bias term. Lemma 2 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 1-3.

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

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

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.

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 $X_1, \ldots, X_n$ 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, 2013] 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 $m \le d$. 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

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

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

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

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

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

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

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

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

Theorem 5. *As in Theorem 3, but where $\mathcal{X} \subseteq \mathbb{R}^d$ is a manifold with intrinsic dimension $m < 4$, $M \le M_{\max}(m)$, and the design distribution P obeys (P3), (P4). There are constants $N, C, c > 0$ such that for any $n \ge N$, and any r as in (R5), the Laplacian smoothing test $\widehat{\varphi}$ based on the estimator $\widehat{f}$ in (2), with $\rho = (nr^{m+2})^{-1}n^{-4/(4+m)}M^{-8/(4+m)}$, satisfies the following: for any $b \ge 1$, if*

$$\epsilon^2 \ge CM^{2m/(4+m)}n^{-4/(4+m)}\left(b^2 + b\sqrt{\frac{1}{\alpha}}\right), \qquad (16)$$

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

The proofs of Theorems 4 and 5 proceed in a similar manner to that of Theorems 1 and 3. The key difference is that in the manifold setting, the equations (14) and (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 4 and 5 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 (4) would have to be specially tailored to work in this setting, revealing another advantage of the discrete approach over the variational one.

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 $f^\top Lf$ 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 $H^s(\mathcal{X})$.

Acknowledgments

AG and RJT were supported by ONR grant N00014-20-1-2787. AG and SB were supported by NSF grants DMS-1713003 and CCF-1763734.

References

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

Sivaraman Balakrishnan, 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*, volume 26, 2013.

Mikhail Belkin. *Problems of Learning on Manifolds*. PhD thesis, University of Chicago, 2003.

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.

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 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.

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.

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

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.

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

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, 2019.

Peter J. Green and Bernard W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC Press, 1993.

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.

Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23(2):169–192, 1952.

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.

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.

Risi Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *International Conference on Machine Learning*, volume 19, 2002.

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.

Giovanni Leoni. *A first Course in Sobolev Spaces.* American Mathematical Society, 2017.

Meimei Liu, Zuofeng Shang, and Guang Cheng. Sharp theoretical analysis for nonparametric testing under random projection. In *Conference on Learning Theory*, volume 32, 2019.

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.

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, 2016.

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.

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.

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 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.

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

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, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.

Martin J Wainwright. *High-Dimensional Dtatistics: A Non-Asymptotic Biewpoint.* Cambridge University Press, 2019.

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.

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.

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.